

UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"

FACULDADE DE CIÊNCIAS - CAMPUS BAURU

DEPARTAMENTO DE COMPUTAÇÃO

BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

BRUNO BELLUZZO

**BUSINESS INTELLIGENCE: DATA SCIENCE APLICADO À
TOMADA DE DECISÕES EM CINEMA DE PEQUENO PORTE**

BAURU

Novembro/2019

BRUNO BELLUZZO

**BUSINESS INTELLIGENCE: DATA SCIENCE APLICADO À
TOMADA DE DECISÕES EM CINEMA DE PEQUENO PORTE**

Trabalho de Conclusão de Curso do Curso
de Bacharelado em Ciência da Computação
da Universidade Estadual Paulista “Júlio
de Mesquita Filho”, Faculdade de Ciências,
Campus Bauru.

Orientador: Prof. Assoc. Dr. João Pedro Albino

BAURU

Novembro/2019

Bruno Belluzzo Business Intelligence: Data Science aplicado à tomada de decisões em cinema de pequeno porte/ Bruno Belluzzo. – Bauru, Novembro/2019-53 p. : il. (algumas color.) ; 30 cm.

Orientador: Prof. Assoc. Dr. João Pedro Albino

Trabalho de Conclusão de Curso – Universidade Estadual Paulista “Júlio de Mesquita Filho”

Faculdade de Ciências

Bacharelado em Ciência da Computação, Novembro/2019.

1. Data Science 2. Business Intelligence 3. Smart Solutions

Bruno Belluzzo

Business Intelligence: Data Science aplicado à tomada de decisões em cinema de pequeno porte

Trabalho de Conclusão de Curso do Curso de Bacharelado em Ciência da Computação da Universidade Estadual Paulista "Júlio de Mesquita Filho", Faculdade de Ciências, Campus Bauru.

Banca Examinadora

Prof. Assoc. Dr. João Pedro Albino

Orientador

Universidade Estadual Paulista "Júlio de Mesquita Filho"

Faculdade de Ciências

Departamento de Computação

Profa. Dra. Simone das Graças Domingues Prado

Universidade Estadual Paulista "Júlio de Mesquita Filho"

Faculdade de Ciências

Departamento de Computação

Prof. Dr. Kelton Augusto Pontara da Costa

Universidade Estadual Paulista "Júlio de Mesquita Filho"

Faculdade de Ciências

Departamento de Computação

Bauru, _____ de _____ de _____.

Agradecimentos

Gostaria de agradecer a todos que contribuíram para o meu crescimento pessoal e profissional ao longo desses anos, talvez sem essas pessoas eu não teria conseguido chegar onde cheguei.

Aos meus amigos, que sempre se colocaram a disposição para me ajudar e são parte fundamental dessa jornada.

Aos meus professores, pela competência e por todo conhecimento que pude adquirir com eles.

Aos meus pais, que não mediram esforços para que eu pudesse realizar meus sonhos e sempre me apoiaram incondicionalmente.

The world is one big data problem.

Andrew McAfee

Resumo

Data Science, *Inteligência Artificial*, *Machine Learning*, *Deep Learning*, entre outras tecnologias emergentes, vêm ganhando força tanto no contexto acadêmico, quanto no empresarial. Empresas e negócios que dominam e utilizam tais tecnologias em seu empreendimento conseguem alavancar de maneira considerável seus resultados. Essas ferramentas podem ser de muita ajuda quando inseridas em negócios de entretenimento e lazer.

O uso de *Ciência de Dados* e suas ferramentas de *Business Intelligence*, que permitem obter dados e analisá-los antes da tomada de decisão pode ser importante para os negócios. Nesta pesquisa, utilizou-se dessa estratégia para compreender as preferências dos clientes de um cinema de pequeno porte, e, por meio dos dados coletados e das análises realizadas, procurou-se gerar soluções oportunas para auxiliar os executivos da organização a melhorar a lucratividade do negócio.

Com as análises efetuadas nesta pesquisa, os executivos agora possuem informações para tomar decisões sobre o futuro do cinema. Foi possível observar que os resultados atingiram um nível de satisfação que agradou aos executivos do negócio, pois os mesmos puderam obter uma visão do perfil dos clientes e também do empreendimento.

Palavras-chave: Ciência de dados, business intelligence, soluções inteligentes.

Abstract

Data Science, Artificial Intelligence, Machine Learning, Deep Learning, among other emerging technologies in recent times have gained strength in the academic context, but not only. Businesses that own the domain and use these technologies in their business can significantly leverage their earnings.

These tools can be very helpful when engaged in entertainment businesses. Dealing with people's satisfaction, pleasing everyone, or at least the majority, and ensuring the best possible customer experience while maximizing profits can be a complex task.

The use of data science in the corporate environment is called Business Intelligence, and analyzing them before making a decision can be crucial for business. In this research, we will use this strategy to understand the preferences of a small cinema customers, and, on the basis of the collected data and analysis, generate intelligent solutions to help executives improve business profits.

Keywords: Data Science , business intelligence, smart solutions.

Lista de figuras

Figura 1 – Características de <i>Data Science</i>	13
Figura 2 – Data Mining	15
Figura 3 – Processo de Data Mining	16
Figura 4 – Crescimento da linguagem Python	17
Figura 5 – Tempo de execução lista x NumPy Array	18
Figura 6 – Criação de matrizes com NumPy	19
Figura 7 – Multiplicação de matrizes com NumPy	19
Figura 8 – Cálculo de determinante com NumPy	19
Figura 9 – Exemplo Pandas Series	20
Figura 10 – Exemplo Pandas Data Frame	21
Figura 11 – Função <i>describe()</i> <i>Data Frame</i>	21
Figura 12 – Amostragem dos dados por condição	22
Figura 13 – Exemplo gráfico de barras com Matplotlib	23
Figura 14 – Exemplo gráfico de setores com Matplotlib	23
Figura 15 – Exemplo gráfico de linha com Matplotlib	24
Figura 16 – Exemplo gráfico de barras com Seaborn	24
Figura 17 – Exemplo gráfico de linha com Seaborn	25
Figura 18 – Preparação inicial dos dados	28
Figura 19 – Separação das respostas	28
Figura 20 – Motivos para nunca terem frequentado	30
Figura 21 – Renda de quem não frequenta	31
Figura 22 – Distribuição da idade dos frequentadores	32
Figura 23 – Distribuição de sexo dos frequentadores	33
Figura 24 – Distribuição de renda dos frequentadores	33
Figura 25 – Gêneros de filmes preferidos	34
Figura 26 – Gêneros de filmes evitados	35
Figura 27 – Frequência dos clientes	36
Figura 28 – Motivos da frequência dos clientes	36
Figura 29 – Motivos da frequência dos clientes filtrada	37
Figura 30 – Função <i>Counter()</i>	37
Figura 31 – Tratativas de respostas livres	38
Figura 32 – O que encoraja a assistir um filme	38
Figura 33 – Melhores dias para ir ao cinema	39
Figura 34 – Companhias	40
Figura 35 – Áudio preferido dos clientes	41
Figura 36 – Horários preferidos dos clientes	42

Figura 37 – Considera preço do ingresso justo	43
Figura 38 – Sugestões do preço do ingresso	44
Figura 39 – Gasto considerado justo por sessão	45
Figura 40 – Benefícios para fiéis incentivaria ir ao cinema	46
Figura 41 – Margem de lucro por produto	47
Figura 42 – Margem de lucro X Lucro líquido	48
Figura 43 – Total de ingressos vendidos por semana	50
Figura 44 – Lucro bruto por semana	51

Lista de abreviaturas e siglas

BI	Business Intelligence
DS	Data Science
DM	Data Mining
IA	Inteligência Artificial

Sumário

1	INTRODUÇÃO	12
2	FUNDAMENTAÇÃO TEÓRICA	13
2.1	Data Science	13
2.2	Business Intelligence	14
2.3	Data Mining	15
3	METODOLOGIA	17
3.1	BIBLIOTECAS PARA DATA SCIENCE	17
3.1.1	Manipulação dos dados	17
3.1.1.1	NumPy	18
3.1.1.2	Pandas	20
3.1.1.2.1	Series	20
3.1.1.2.2	Data Frame	21
3.1.2	Visualização dos dados	22
3.1.2.1	Matplotlib	22
3.1.2.2	Seaborn	24
3.2	Formulário	25
3.3	Base de dados do cinema	25
4	DESENVOLVIMENTO DA PESQUISA	27
4.1	Coleta e preparação dos dados	27
4.2	Análise e Validação	29
4.2.1	Análise de quem nunca frequentou o cinema	29
4.2.2	Análise de quem frequenta o cinema	31
4.2.3	Dados financeiros	46
5	RESULTADOS	49
6	CONCLUSÃO	51
	REFERÊNCIAS	53

1 INTRODUÇÃO

Data Science ou Ciência de Dados é uma área que já existe há mais de 30 anos, mas vem ganhando destaque nos últimos anos, devido ao *Big Data*. O desenvolvimento de áreas como *machine learning* reforçam o crescimento e a importância de *Data Science* mas não é apenas neste ramo que este campo é bastante válido, sendo que é cada vez mais comum empresas se beneficiarem do estudo para a tomada de decisões, de forma a alavancar os seus crescimentos (D. P. SILVEIRA, 2016).

Os mais diferentes tipos de negócios podem (e devem) fazer uso dos benefícios de *data science* para alavancar seus negócios, e não seria diferente com empresas do ramo do entretenimento, mais especificamente, trabalhado nesse estudo, do ramo cinematográfico.

Não é novidade que a indústria cinematográfica mundial gera uma receita bilionária todos os anos. No Brasil, números levantados pela Motion Picture Association na América Latina (MPA-AL) em parceria com o Sindicato Interestadual da Indústria do Audiovisual, mostram que a indústria cinematográfica injeta anualmente mais de R\$19 bilhões na economia brasileira.

Apesar do grande impacto na economia mundial e do sucesso que são os cinemas ao redor do mundo, com o crescimento de serviços de vídeo online as pessoas vão deixando cada vez mais de ir ao cinema, pois "uma hora ou outra" o filme estará disponível *online*. Não foi a toa que a sétima edição da CinemaCon, o maior evento de exibição cinematográfica dos Estados Unidos, em 2017, teve como tema informal o futuro do negócio da exibição cinematográfica na era do *streaming*. Muitos estúdios já possuem versões *online* de seus canais, contudo é preciso balancear as coisas com os donos das salas de cinema, que sempre foram fieis aliados dos grandes estúdios cinematográficos.

Em meio a toda essa revolução cinematográfica, cinemas de pequeno porte, que possuem uma única sala para exibição de filmes, sofrem com a disponibilidade de cópias, pois os estúdios e suas distribuidoras dão preferência aos grandes cinemas, que irão gerar uma maior receita a eles. Com isso em mente e buscando entender as limitações desses pequenos cinemas, técnicas de *data science* podem ser utilizadas para maximizarem os lucros e melhorarem a satisfação dos clientes.

2 FUNDAMENTAÇÃO TEÓRICA

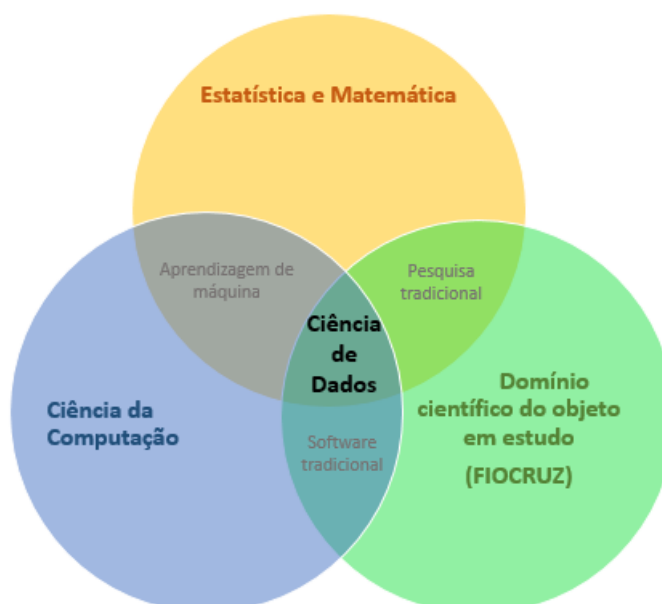
Neste capítulo apresentam-se conceitos e estudos levantados sobre áreas em que este trabalho está inserido, como por exemplo, *Business Intelligence*, *Data Science* e *Data Mining*.

2.1 Data Science

Data Science é uma área extremamente multidisciplinar, que envolve conceitos de estatística, computação, matemática e conhecimento de negócio. É uma ciência que visa estudar as informações, seu processo de captura, transformação, geração e, posteriormente, análise de dados. *Data Science* busca, a partir de uma quantidade grande e pesada de dados, gerar conhecimentos e informações relevantes para tomar decisões e fazer previsões, e não simplesmente interpretar os números. (L. COELHO, 2018)

A Figura 1 ilustra as habilidades e técnicas que constituem a área conhecida como Data Science (DS):

Figura 1 – Características de *Data Science*



Fonte: <https://bigdata.iciet.fiocruz.br/ciencia-de-dados-aplicada-saude>; Acesso em nov. 2019.

Para este trabalho, no que diz respeito a Ciência da Computação, foram utilizadas a linguagem Python e suas bibliotecas para se trabalhar com os dados e analisar as informações estatísticas e matemáticas, como por exemplo: Matplotlib e Seaborn para a visualização dos dados; e Pandas e Numpy para a manipulação dos dados. Essas ferramentas são apresentadas na Seção 3.

Não menos importante para se trabalhar com DS, o conhecimento do domínio em que irá se utilizar a abordagem da análise dos dados é fundamental. Conhecer e compreender as necessidades do negócio em que será feito o estudo é a chave principal para alcançar bons resultados no final da aplicação.

2.2 Business Intelligence

Business Intelligence (BI) é uma técnica que organiza dados coletados dentro de uma empresa e utiliza *softwares* para simplificar as informações de acordo com os interesses dos executivos. Com base nos estudos e nas análises feitas sobre as informações levantadas, os diretores serão capazes de tomar as melhores decisões para o negócio.

"BI é utilizado especialmente para a gestão da empresa, trabalhando dados do passado para traçar as tendências do futuro, fazendo previsões de vendas, por exemplo, a partir do comportamento do banco de dados de seus próprios clientes."(CANALCOMSTOR, 2018)

O conhecimento obtido a partir do uso de DS é fundamental para apoiar o processo de tomada de decisão. A informação gerada pelas aplicações informáticas disponibiliza aos gestores um conjunto de indicadores sobre o negócio, que lhe dão indicações do que aconteceu no passado e lhe permitem traçar cenários para o futuro. (SANTOS; RAMOS, 2006)

Obter resultados satisfatórios não depende exclusivamente da quantidade de dados gerados ou encontrar maneiras de gerar mais dados, é necessário saber estudar e trabalhar com as informações disponíveis, e é nesse contexto que entra *Data Science*. (D. P. SILVEIRA, 2016)

O modo com que as empresas gerenciam as informações e realizam tomadas de decisões pode influenciar consideravelmente o sucesso da mesma. "As empresas que possuem ferramentas de BI e a utilizam em seus processos durante uma tomada de decisão, apresentam vantagem, para se posicionarem a frente de uma nova oportunidade no mercado". Alcançar a fidelidade de seus clientes depende de jogadas de *marketing* e tomadas de decisões precisas e estratégicas. Tomar a decisão correta, com base nos dados levantados, mantém as empresas ativas e lucrativas em seus mercados. Porém não tem sido fácil para as empresas tomarem decisões corretas e analisar a causa raiz de seus problemas, segundo o Sebrae-Sp (2013) 27% das empresas paulistanas fecham suas portas em seu primeiro ano de atividade. (SILVA; TERRA, 2015, p. 11)

Como diz o INSTITUTO ATLÂNTICO (2018), os dados são o novo petróleo e o grande recurso desta década. Nos últimos anos, o mundo tem gerado uma grande quantidade de informação, e as companhias têm interesse nesse recurso como uma forma de aprimorar seus serviços. Os mais diferentes tipos de negócios podem (e devem) fazer uso dos benefícios da DS para alavancar seus negócios, e não seria diferente com empresas do ramo do entretenimento.

2.3 Data Mining

O *Data Mining* (DM) descende fundamentalmente de 3 linhagens: estatística clássica, Inteligência Artificial (IA) e Machine Learning (ML) (DEV MEDIA, 2011). Talvez a definição mais importante tenha sido elaborada por Usama Fayyad (FAYYAD; PIATETSKY-SHAPIO; SMYTH, 1996):

"...o processo não-trivial de identificar, em dados, padrões válidos, novos, potencialmente úteis e ultimamente compreensíveis"

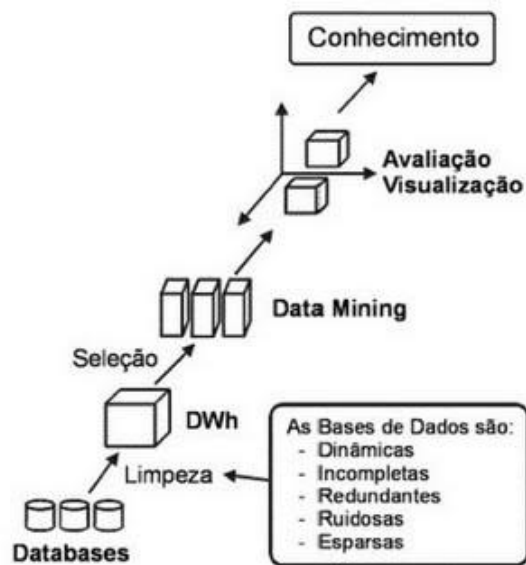
Figura 2 – Data Mining



Fonte: https://www.sas.com/en_sg/insights/analytics/data-mining.html; Acesso em jul. 2019.

Esse processo pode ser feito por diversos algoritmos que processam os dados e encontram esses "padrões válidos, novos e valiosos", mas como dito por Navega(2002) é necessário ressaltar que apesar dos algoritmos atuais serem capazes de descobrir padrões "válidos e novos", ainda não tem uma solução eficaz para determinar padrões valiosos. Por essa razão, ainda requer uma interação muito forte com analistas humanos, que são, em última instância, os principais responsáveis pela determinação do valor dos padrões encontrados. Além disso, a condução da exploração de dados é também tarefa fundamentalmente confiada a analistas humanos, já que cada caso em que se deseja realizar DM se busca diferentes resultados finais.

Figura 3 – Processo de Data Mining



Fonte: <https://www.devmedia.com.br/conceitos-e-tecnicas-sobre-data-mining/19342>; Acesso em jul. 2019.

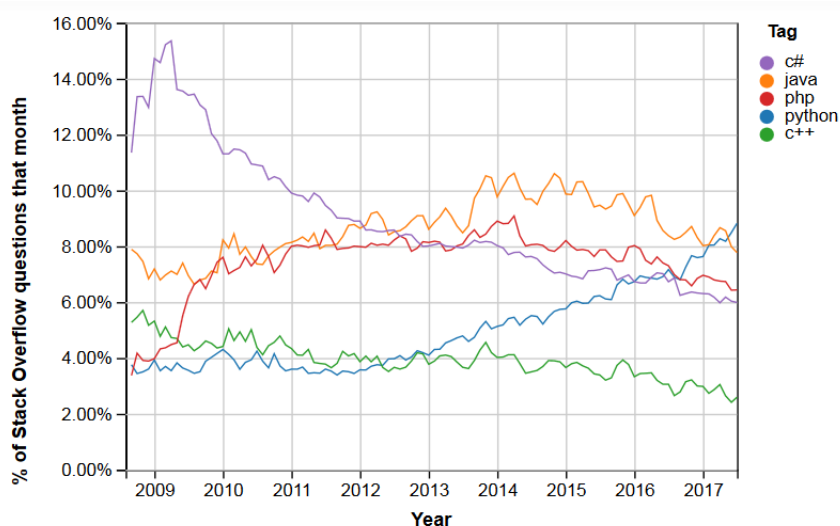
A partir de dados brutos efetua-se uma limpeza (consistência, preenchimento de informações, remoção de ruído e redundâncias, etc). Disto nascem os repositórios organizados (Data Marts e Data Warehouses), que já são úteis de diversas maneiras, porém é a partir da análise feita por um analista, utilizando visualização gráfica, o processo é refinado e conduzido até que valiosos padrões apareçam. É um processo hierarquico, onde os dados começam de forma volumosa e terminam em um ponto relativamente concentrado, mas muito valioso.

3 METODOLOGIA

3.1 BIBLIOTECAS PARA DATA SCIENCE

Diversas linguagens de programação são utilizadas em data science, como por exemplo Java, R, SAS, SQL, Excel e Python, mas é essa última que vem se tornando a favorita entre os cientistas de dados. Como mostra a Figura 4, que mostra a porcentagem de questões no Stack Overflow sobre algumas linguagens de programação, vemos a ascensão do Python em relação as outras linguagens de programação mais tradicionais, como C# e Java, mostrando uma comunidade extremamente ativa.

Figura 4 – Crescimento da linguagem Python



Fonte: <https://medium.com/tendências-digitais/python-engolindo-o-mercado-6872769800b2>; Acesso em jul. 2019.

3.1.1 Manipulação dos dados

A primeira fase no processo de Análise de Dados é a fase de Preparação com o objetivo de realizar o tratamento dos dados, para que se consiga uma boa visualização dos mesmos sem precisar recorrer a ferramentas mais específicas, como softwares de planilhas de texto, de estatística e banco de dados relacional (SIEGEL, 2018). Para essa abordagem, a linguagem Python possui duas bibliotecas que vem ganhando notoriedade no cenário de data science: NumPy e Pandas.

3.1.1.1 NumPy

O Numpy é a junção de duas ferramentas (*Numeric* e *NumArray*), que possuem o propósito de efetuar cálculos, realizar análises estatísticas e matemáticas bem como manipulações numéricas a partir de matrizes n-dimensionais, porém essa abordagem de matrizes n-dimensionais possui problemas de performance e funcionalidade. Buscando unificar as duas ferramentas para se aproveitar das melhores características de cada uma, Travis Oliphant, juntamente com uma comunidade de desenvolvedores, publicaram em 2006 a primeira versão do NumPy.

Menor consumo de memória para armazenamento dos dados, menor tempo de execução de instruções e a otimização de performance em relação as funcionalidades nativas do Python, como por exemplo listas, colaborou para o crescimento e adoção da ferramenta pela comunidade científica. A Figura 5 faz uma comparação entre o tempo de execução para se calcular o seno de 10.000 valores contidos em uma lista do Python e um ndarray do NumPy. Para a lista, o tempo de execução foi 2.52ms, enquanto para o ndarray foi de 0.247ms, ou seja, quase 10 vezes mais rápido.

Figura 5 – Tempo de execução lista x NumPy Array

```
In [1]: #Importando a biblioteca que contém funções matemáticas do Python
import math

#Importando Numpy
import numpy as np

elementos = 10000

#Criando a lista utilizando Python
lista_python = [(x) for x in range(elementos)]

#Criando a lista utilizando Numpy
np_array = np.arange(elementos)

print('Tempo utilizando lista:')

%timeit [math.sin(x) for x in lista_python]

print('\nTempo utilizando NumPy Array:')

%timeit np.sin(np_array)

Tempo utilizando lista:
2.52 ms ± 88.8 µs per loop (mean ± std. dev. of 7 runs, 100 loops each)

Tempo utilizando NumPy Array:
247 µs ± 8.45 µs per loop (mean ± std. dev. of 7 runs, 1000 loops each)
```

Fonte: Elaborado pelo autor

Como pode ser visto na Figura 6, é possível efetuar a criação de diversas matrizes utilizando o NumPy, desde a criação informando os números manualmente até lendo arquivos contendo os dados. Além disso, existem algumas funções que facilitam a criação de matrizes especiais, como matriz nula, matriz identidade e matriz com números aleatórios.

Figura 6 – Criação de matrizes com NumPy

```

In [1]: import numpy as np

In [9]: #Criando matriz nula de dimensão 6 x 6
Matriz_nula = np.zeros([6,6])
print("Matriz Nula: \n",Matriz_nula)

#Criando matriz identidade de dimensão 5x5
Matriz_iden = np.eye(5)
print("\nMatriz Identidade:\n", Matriz_iden)

#Criando matriz com números randômicos de 0 a 5 - 5 x 5
Matriz_random = np.random.randint(5, size = (5,5))
print("\nMatriz Randômica:\n", Matriz_random)

Matriz Nula:
[[0. 0. 0. 0. 0. 0.]
 [0. 0. 0. 0. 0. 0.]
 [0. 0. 0. 0. 0. 0.]
 [0. 0. 0. 0. 0. 0.]
 [0. 0. 0. 0. 0. 0.]
 [0. 0. 0. 0. 0. 0.]]

Matriz Identidade:
[[1. 0. 0. 0. 0.]
 [0. 1. 0. 0. 0.]
 [0. 0. 1. 0. 0.]
 [0. 0. 0. 1. 0.]
 [0. 0. 0. 0. 1.]]

Matriz Randômica:
[[0 1 4 1 2]
 [4 4 0 0 4]
 [1 3 1 1 0]
 [2 2 3 1 1]
 [4 0 4 1 4]]

```

Fonte: Elaborado pelo autor

O NumPy também possui recursos matemáticos para se trabalhar com as matrizes. Ele possibilita efetuar operações de álgebra linear com bastante simplicidade e eficiência, como multiplicação de matrizes, normalização, cálculo de determinantes, inversão e transposição de matrizes. As Figuras 7 e 8 mostram algumas dessas funcionalidades.

Figura 7 – Multiplicação de matrizes com NumPy

```

In [1]: import numpy as np

In [2]: M1 = np.random.randint(5, size=(3,3))
M2 = np.random.randint(5, size=(3,3))

In [3]: print(M1)
print('-----')
print(M2)

[[2 1 2]
 [4 2 3]
 [1 1 3]]
-----
[[4 2 0]
 [3 2 1]
 [2 4 1]]

In [4]: M1*M2

Out[4]: array([[ 8,  2,  0],
               [12,  4,  3],
               [ 2,  4,  3]])

```

Fonte: Elaborado pelo autor

Figura 8 – Cálculo de determinante com NumPy

```

In [1]: import numpy as np

In [2]: M1 = np.random.randint(5, size=(4,4))

In [3]: print(M1)
print('Determinante:', round(np.linalg.det(M1),3))

[[2 1 0 1]
 [4 2 2 2]
 [4 4 1 2]
 [4 2 4 0]]
Determinante: 16.0

```

Fonte: Elaborado pelo autor

3.1.1.2 Pandas

No que diz respeito a manipulação e análise de dados estruturados, a biblioteca Pandas é a mais completa em Python. Devido sua capacidade de lidar com grandes massas de dados e facilidade de uso, sem perder a eficiência, a ferramenta Pandas ganhou grande notoriedade, pois não é necessário utilizar outros softwares de planilhas de texto, de estatísticas e banco de dados.

Cálculos de desvio padrão, média, mediana, quartil, soma, valores mínimos e máximos, entre outros, podem ser feitos através de funções simples da biblioteca. Entretanto, uma das características mais interessantes do Pandas é a facilidade da manipulação dos dados, podendo ser aplicadas funções apenas para um conjunto dos dados, agrupar os dados baseado em algum critério, separar determinados objetos em grupos, entre outros. As duas principais estruturas de dados do Pandas para realizar as ações descritas são as *Series* (1 dimensão) e o *Data Frame* (2 dimensões).

3.1.1.2.1 Series

Series é um *array* unidimensional contendo um vetor de dados e um vetor de índices, que são os rótulos dos dados. Na Figura 9 é criado um *Series* contendo nome de alunos como rótulo, e as notas dos mesmos sendo os dados, além de exibir algumas estatísticas sobre as notas.

Figura 9 – Exemplo Pandas Series

```
In [1]: import pandas as pd

In [2]: #Criação da Series das notas dos alunos
notas = pd.Series([10, 8, 7, 9, 3], index = ['Pedro', 'Paulo', 'Caio', 'Marcelo', 'Vitor'])

In [3]: notas
Out[3]: Pedro      10
        Paulo      8
        Caio       7
        Marcelo    9
        Vitor      3
        dtype: int64

In [4]: #Selecionando apenas alguns rótulos
notas[['Vitor', 'Marcelo']]
Out[4]: Vitor      3
        Marcelo    9
        dtype: int64

In [5]: #Estatísticas sobre os valores da Series
notas.describe()
Out[5]: count      5.000000
        mean       7.400000
        std        2.701851
        min        3.000000
        25%        7.000000
        50%        8.000000
        75%        9.000000
        max        10.000000
        dtype: float64
```

Fonte: Elaborado pelo autor

3.1.1.2.2 Data Frame

Data Frame é uma estrutura de dados bidimensional, em formato de tabela, que contém várias colunas com seus valores nas linhas. É uma matriz onde cada coluna pode conter diferentes tipos de dados, como número inteiro, número decimal, texto, entre outros. Seguindo com a ideia da Figura 9, a Figura 10 mostra um Data Frame contendo mais informações sobre os alunos.

Figura 10 – Exemplo Pandas Data Frame

```
In [1]: import pandas as pd

In [2]: #Criando o Data Frame
df = pd.DataFrame(columns=['Aluno', 'Nota P1', 'Nota P2', 'Frequência'],
                  data=[["Vitor", 10, 5, 70], ["Pedro", 10, 10, 100], ["Marcelo", 6, 3, 60]])

In [3]: #Exibindo o Data Frame
df
Out[3]:
```

	Aluno	Nota P1	Nota P2	Frequência
0	Vitor	10	5	70
1	Pedro	10	10	100
2	Marcelo	6	3	60

Fonte: Elaborado pelo autor

Com a função *describe()* é possível fazer um mapeamento matemático e estatísticos dos dados numéricos do *Data Frame*. É possível descobrir a quantidade de registros (*count*), média (*mean*), desvio padrão (*std*), valores mínimos e máximos (*min*, *max*) e quartis (Q1: 25%, Q2: 50%, Q3: 75%). A Figura 11 mostra a aplicação dessa função utilizando o *Data Frame* anterior.

Figura 11 – Função *describe()* *Data Frame*

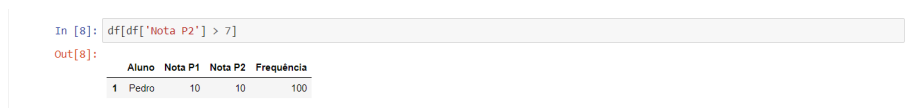
```
In [4]: df.describe()
Out[4]:
```

	Nota P1	Nota P2	Frequência
count	3.000000	3.000000	3.000000
mean	8.666667	6.000000	76.666667
std	2.309401	3.605551	20.816660
min	6.000000	3.000000	60.000000
25%	8.000000	4.000000	65.000000
50%	10.000000	5.000000	70.000000
75%	10.000000	7.500000	85.000000
max	10.000000	10.000000	100.000000

Fonte: Elaborado pelo autor

É possível filtrar os dados mostrados por meio de uma condição determinada em alguma coluna do *Data Frame*, como por exemplo na Figura 12, onde é exibido apenas as linhas onde o dado da coluna "Nota P2" seja maior que 7. Esse tipo de amostragem condicional dos dados é extremamente importante para compreender algumas relações entre as colunas.

Figura 12 – Amostragem dos dados por condição



The image shows a Jupyter Notebook interface. The input cell contains the code `df[df['Nota P2'] > 7]`. The output cell displays a table with the following data:

	Aluno	Nota P1	Nota P2	Frequência
1	Pedro	10	10	100

Fonte: Elaborado pelo autor

3.1.2 Visualização dos dados

É essencial em *Data Science* saber analisar os dados e perceber as relações entre eles. A linguagem Python possui duas bibliotecas para facilitar a visualização dos dados, para assim conseguir, de forma mais intuitiva, analisar alguns padrões nas respostas e compreender melhor como as amostras estão disitribuidas.

3.1.2.1 Matplotlib

O Matplotlib surgiu como uma alternativa ao MATLAB para ser usado com Python. Com a proposta de gerar gráficos de maneira simples, utilizando uma linguagem de fácil entendimento além de funcionar em diferentes ambientes e sistemas operacionais, John D. Hunter lançou em 2003 a primeira versão do Matplotlib.

Na Análise de Dados, é preciso entender como os dados estão distribuídos, realizar uma avaliação do conjunto de dados, efetuar comparações entre as variáveis, identificar valores discrepantes e obter conhecimentos para uma tomada de decisão mais adequada.

Quando se trata de BI, a representação dos dados por meio de gráficos de simples compreensão é importante para que os executivos da empresa consigam compreender como o cientista de dados identificou as soluções inteligentes que ele propos para serem implementadas.

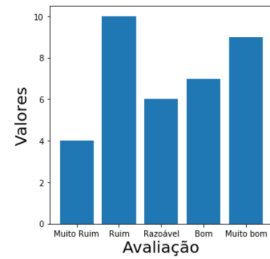
As Figuras 13, 14 e 15 exemplificam três tipos de gráficos gerados pelo Matplotlib que serão utilizados futuramente neste projeto.

Figura 13 – Exemplo gráfico de barras com Matplotlib

```
In [9]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

In [10]: valores = np.array([4,10,6,7,9])
labels = ['Muito Ruim', 'Ruim', 'Razoável', 'Bom', 'Muito bom']

In [11]: plt.figure(figsize=(5,5))
plt.bar(labels, valores)
plt.ylabel('Valores', fontsize=20)
plt.xlabel('Avaliação', fontsize=20)
plt.show()
```



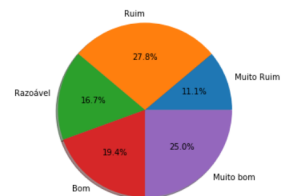
Fonte: Elaborado pelo autor

Figura 14 – Exemplo gráfico de setores com Matplotlib

```
In [9]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

In [10]: valores = np.array([4,10,6,7,9])
labels = ['Muito Ruim', 'Ruim', 'Razoável', 'Bom', 'Muito bom']

In [14]: plt.figure(figsize=(5,5))
plt.pie(valores, labels=labels, shadow=True, autopct='%1.1f%%')
plt.show()
```



Fonte: Elaborado pelo autor

Figura 15 – Exemplo gráfico de linha com Matplotlib

```

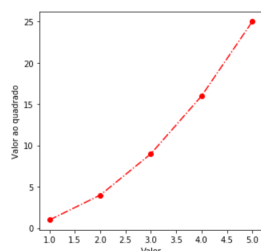
In [9]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

In [28]: x = np.arange(1, 6)
y = x**2

In [29]: y
Out[29]: array([ 1,  4,  9, 16, 25], dtype=int32)

In [35]: plt.figure(figsize=(5,5))
plt.plot(x,y, linestyle='-.', color='red', marker='o')
plt.ylabel('Valor ao quadrado')
plt.xlabel('Valor')
plt.show()

```



Fonte: Elaborado pelo autor

3.1.2.2 Seaborn

O Seaborn é uma biblioteca de visualização de dados *Python* baseada no matplotlib. Ele fornece uma interface de alto nível para desenhar gráficos estatísticos atraentes e informativos (SHARMA, 2018).

Como afirma Michael Waskom, no site oficial do Seaborn: "Se o Matplotlib tenta tornar as coisas fáceis, fáceis, e coisas difíceis possíveis, o Seaborn tenta tornar um conjunto bem definido de coisas difíceis fáceis também.". E é exatamente por esse motivo que torna essa biblioteca a favorita para visualização de dados estatísticos em *Python*.

As Figuras 17 e 16 demonstram como criar alguns gráficos com Seaborn:

Figura 16 – Exemplo gráfico de barras com Seaborn

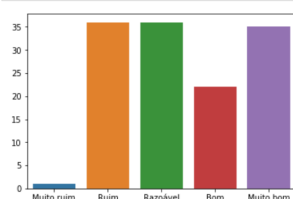
```

import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

labels = ['Muito ruim', 'Ruim', 'Razoável', 'Bom', 'Muito bom']
values = np.random.randint(1,50,5)

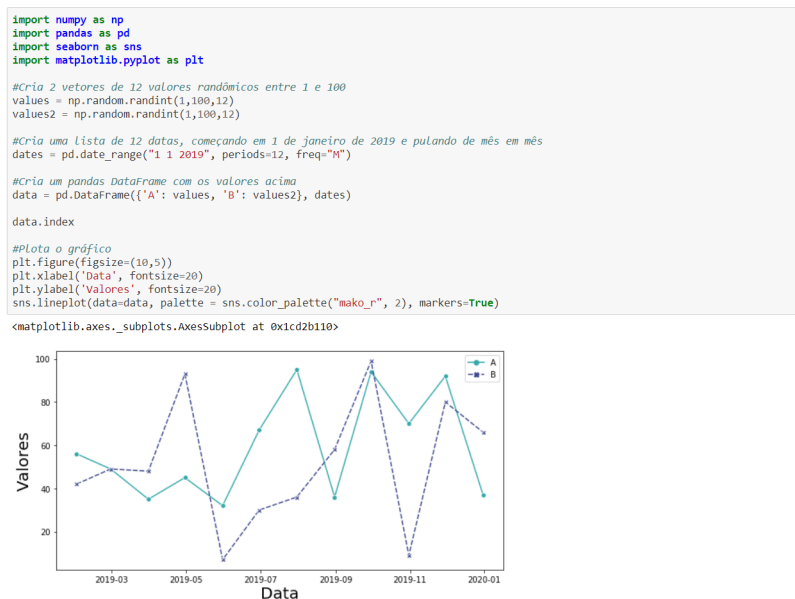
sns.barplot(labels, values)
plt.show()

```



Fonte: Elaborado pelo autor

Figura 17 – Exemplo gráfico de linha com Seaborn



Fonte: Elaborado pelo autor

3.2 Formulário

Nesta pesquisa uma das formas utilizada para obtenção de insumos para aplicar data science foi realizada por meio de um formulário com várias perguntas de múltipla escolha e resposta livre o qual foi disponibilizado aos clientes do cinema para responderem.

Para preparação do formulário, foi extremamente importante ao cientista de dados obter um conhecimento sobre o negócio e compreender as expectativas do cliente quanto aos resultados que ele pretendia com as análises. Levantar as questões que se encaixariam no formulário dependeu de um trabalho conjunto entre o analista e os executivos da empresa.

Com os principais pontos que queriam saber em relação aos clientes e ao cinema levantado, foi utilizado o Google Forms, por ser uma ferramenta simples de ser manipulada e intuitiva, além de poder posteriormente gerar uma planilha contendo todas as respostas, que pode ser importada para dentro de um Pandas Dataframe. As perguntas foram divididas visando dois grupos de pessoas, sendo o primeiro pessoas que já frequentaram o cinema e outro as que nunca assistiram a um filme no local.

3.3 Base de dados do cinema

Para apoiar as tomadas de decisão e compreender melhor as questões financeiras da empresa, foi feito um backup do programa que é utilizado para realizar as vendas tanto de

produtos quanto de ingressos. Esses dados foram importantes uma vez que, ao fazer um comparativo entre diferentes períodos do cinema, é possível compreender em quais situações houve um maior faturamento e tentar reproduzir as condições que haviam naquele momento para atingir esses ganhos.

É possível descer em detalhes por produto e identificar oportunidades de negócio, seja em possibilidades de promoções, filtragem de produtos a serem vendidos, compreender os produtos que mais deram lucro, os que tem maior margem de lucro e vice-versa, entre outras informação.

Idealmente os dados deveriam estar estruturados de forma que pudessem ser manipulados, entretanto para este caso específico não era possível fazer um backup da base, era possível apenas exportar um .jpeg contendo as informações entre um determinado período de tempo, posteriormente foi necessário passar esses dados um a um para uma planilha a mão.

4 DESENVOLVIMENTO DA PESQUISA

4.1 Coleta e preparação dos dados

Com as bibliotecas e a metodologia definidas, o próximo passo foi a coletar dos dados. Vale frizar novamente que é extremamente importante o conhecimento do domínio, se reunir com os executivos da empresa e levantar os objetivos que se pretende alcançar com os dados que serão levantados, para que o conjunto de dados tenham uma grande representatividade.

No caso de um cinema, a ideia é fazer com que mais pessoas frequentem o estabelecimento ao mesmo tempo que os lucros cresçam em igual proporção, é preciso encontrar um ponto de equilíbrio entre preço dos produtos/ingressos e frequência dos clientes, afim de maximizar os ganhos. Quanto mais dados forem coletados, mais insumo será gerado para se tomar soluções inteligentes.

Para atingir este ponto, é necessário compreender as preferências das pessoas que frequentam o cinema e suas limitações, além de ter opiniões sobre diferentes aspectos do empreendimento e, com esses dados, elaborar as melhores decisões a serem tomadas.

Para a coleta desses dados, duas estratégias foram elaboradas: a primeira consistiu em um formulário disponibilizado ao público contendo perguntas do interesse dos executivos, como por exemplo, gêneros de filme preferidos, o que costumam consumir no cinema, os dias preferidos para assistir, os horários preferidos, com quem costumam ir ao cinema, entre outras perguntas; a segunda foi coletar o histórico de vendas tanto de produtos quanto de ingressos desde a reabertura do cinema.

O formulário era diferente para pessoas que já frequentaram o cinema e pessoas que nunca o fizeram, para assim ter uma amostragem de dados bem abrangente. Ao final da pesquisa, obteve-se um total de 447 respostas, o que pode ser considerado um número aceitável, capaz de generalizar os diferentes clientes que frequentam, ou deixam de frequentar, o cinema.

Para levantar informações sobre as vendas da empresa, foi necessário realizar um backup da base de dados do software que ela utilizava e estruturar esses dados de forma manipulável(dataframe), pois era possível apenas exportar um .jpeg sobre algumas informações limitadas das vendas. Além disso, informações de negócio foram levantadas diretamente com os executivos, como por exemplo o custo por unidade de cada produto que é vendido no cinema.

Com os dados em mão, a próxima etapa em um processo de *Data Science* é a manipulação dos dados e, como apresentado na Seção 3.1.1, será utilizado o NumPy e o Pandas para tal.

O primeiro passo é ler o arquivo .xlsx gerado pelo Google Forms para dentro de um

Pandas Dataframe, para assim conseguir manipular esses dados de uma forma mais simples. Um problema dessa abordagem é o fato de que os nomes das colunas do Dataframe correspondem as perguntas que foram elaboradas no formulário, o que dificulta a filtragem dos dados por coluna, portanto é extremamente importante renomear o nome das colunas para facilitar a manipulação, além de poder remover as primeiras duas colunas, que são a data da resposta e o nome de quem respondeu, que no caso não irão agregar valor a pesquisa. A Figura 18 mostra como proceder para realizar estas operações.

Figura 18 – Preparação inicial dos dados

```
In [1]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from collections import Counter
import regex as re

In [2]: df = pd.read_excel('../Pesquisa Cine Belluzzo.xlsx')

In [3]: df = df.iloc[:, 2:]

In [4]: new_columns = ['Idade',
                        'Sexo',
                        'RendaMensal',
                        'JaAssistiu',
                        'GenerosPreferidos',
                        'GenerosEvitados',
                        'frequencia',
                        'MotivoFrequencia',
                        'EncorajaCinema',
                        'MelhoresDias',
                        'Companhia',
                        'PrecoAtualIngresso',
                        'PrecoIngressoIlaAdequado',
                        'TotalGasto',
                        'CostumaConsumir',
                        'Dublado/Legendado',
                        'MelhoresHorarios',
                        'BeneficiosFieis',
                        'NotaFilmesExibidos',
                        'NotaPrecoProdutos',
                        'NotaDivulgacao',
                        'NotaPromocoesCombos',
                        'NotaAtendimento',
                        'NotaSessao',
                        'NotaGerai',
                        'Sugestoes1',
                        'PorqueIaofrequentou',
                        'VontadeDecomecar',
                        'Sugestoes2'
                        ]

In [5]: df.columns = new_columns
```

Fonte: Elaborado pelo autor

Com as colunas renomeadas, foi separado o dataframe em dois, um com as respostas das pessoas que já frequentaram o cinema alguma vez e outro com as respostas das pessoas que nunca foram ao cinema. Fazendo isso, não é necessário filtrar os dados toda vez que formos realizar uma análise de alguma resposta. A filtragem condicional sobre o *dataframe* pode ser utilizado para tal finalidade e a função *isshape* pode ser usada para ver a estrutura de cada *dataframe*, como mostra a Figura 19.

Figura 19 – Separação das respostas

```
In [6]: df_freq = df[df['JaAssistiu'] == 'Sim']
df_nao_freq = df[df['JaAssistiu'] == 'Não']

In [8]: print(df.shape)
print(df_freq.shape)
print(df_nao_freq.shape)

(447, 29)
(394, 29)
(53, 29)
```

Fonte: Elaborado pelo autor

Nota-se que 394 pessoas que responderam ao formulário já frequentaram ou frequentam o cinema, enquanto 53 nunca foram a uma sessão. Esse tratamento é importante uma vez que as perguntas são diferentes dependendo desta primeira resposta e é interessante avaliar separadamente estes dois agrupamentos.

No caso dos dados extraídos da base de dados do software de vendas, foi necessário criar outras informações para poder extrair *insights* dos dados, uma vez que o programa disponibilizava apenas quantas unidades foram vendidas de cada produto bem como o total dessas vendas. Com base nessas informações, foi feito o cálculo do preço unitário por produto e, somada à informação dos custos dos produtos levantados com os executivos, uma coluna com as informações de lucro líquido e margem de lucro por produto também foi criada.

4.2 Análise e Validação

Com os agrupamentos feitos, foram analisadas as respostas em busca de insumos para apresentar aos executivos da empresa para que decisões inteligentes possam ser tomadas. Primeiramente houve uma análise das respostas de pessoas que nunca frequentaram o cinema e posteriormente as que já frequentaram.

Em seguida iremos trabalhar com os dados financeiros levantados para auxiliar e expor aos diretores do cinema os pontos que considerarmos relevantes para criar soluções em cima dos produtos oferecidos pelo empreendimento.

4.2.1 Análise de quem nunca frequentou o cinema

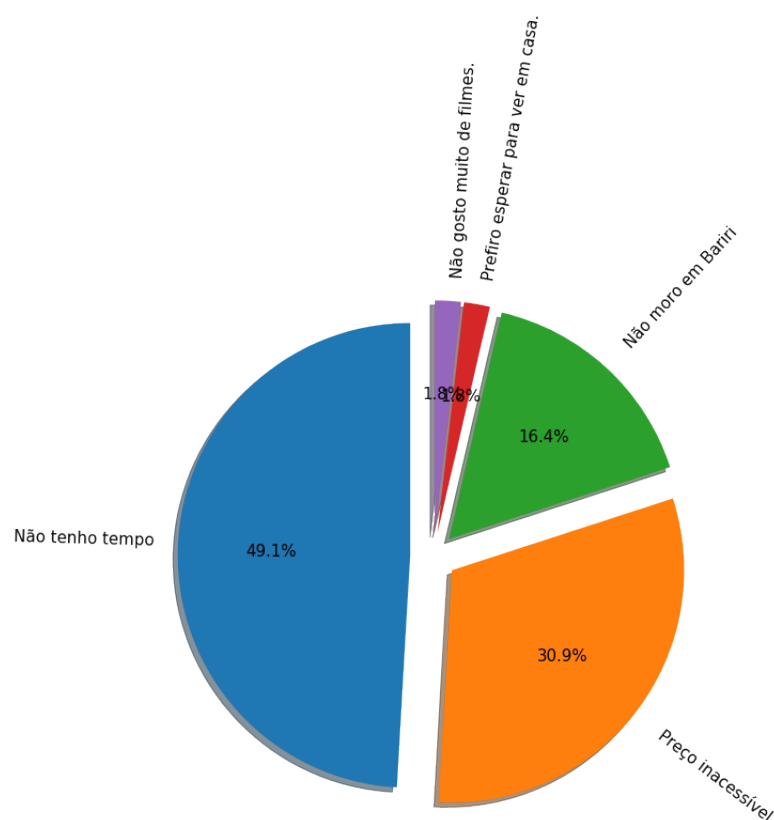
Para fazer com que novas pessoas frequentem o negócio, é importante entender os motivos para cada um nunca ter o feito, e assim conseguir compreender essas questões através dos dados levantados pela pesquisa. Cada uma das perguntas será trabalhada de forma que no final possa ser gerado um gráfico de simples compreensão para os executivos da empresa. Para começar, ao analisar a distribuição de idade e do sexo das pessoas que se encaixam nessa categoria, verificou-se que 52,8% possuem entre 15-24 anos e 22,6% entre 25-34 anos, quanto ao sexo, 75,5% são mulheres e 24,5% homens.

A primeira resposta a ser analisa é o motivo das pessoas não frequentarem o cinema. Nessa pergunta a resposta das pessoas é dissertativa, ou seja, não é múltipla escolha, o que dificulta um pouco o processo de agrupamento e geração de insumos. É necessário entender os principais tipos de respostas e procurar uma palavra chave que classifique a resposta em determinado grupo.

Com essa técnica, é possível agrupar as respostas e levantar insumos para que os executivos da empresa analisem. A Figura 20 mostra os principais motivos para as pessoas

nunca terem assistido um filme no cinema.

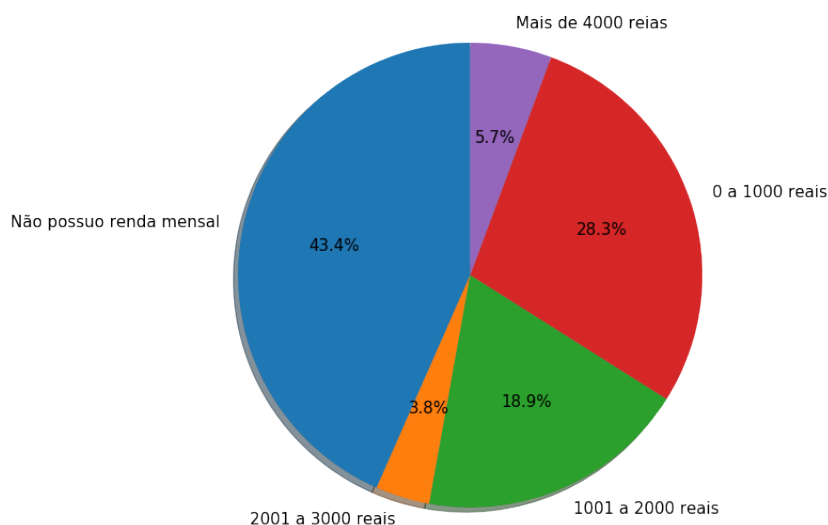
Figura 20 – Motivos para nunca terem frequentado



Fonte: Elaborado pelo autor

Pode-se notar que os dois principais motivos são tempo e dinheiro, o que nos direciona para analisar as rendas das pessoas que responderam o formulário e nunca visitaram o estabelecimento, mostrado na Figura 21. Essa pergunta era de múltipla escolha, o que facilita o agrupamento dos dados.

Figura 21 – Renda de quem não frequenta



Fonte: Elaborado pelo autor

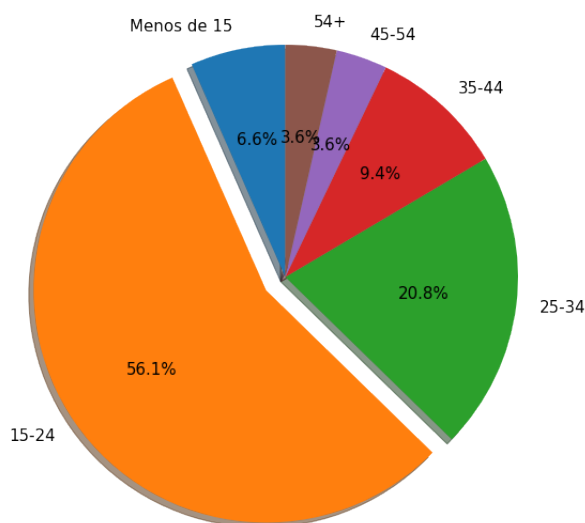
Nota-se que 71,7% das pessoas possuem renda mensal igual ou inferior a 1000 reais, o que comprova um dos principais motivos das pessoas não irem ao cinema ser financeiro. Quanto a falta de tempo, é muito difícil analisar, uma vez que cada pessoa possui uma rotina muito diferente da outra, e a falta de tempo são por motivos diversos.

4.2.2 Análise de quem frequenta o cinema

Continuando com a análise das respostas, a próxima análise será em cima das respostas daqueles que frequentam ou já frequentaram o cinema, que corresponde a uma amostragem mais valiosa para a empresa, já que representa as preferências dos clientes atuais.

Os primeiros estudos feitos diz respeito a informações pessoais das pessoas, que são idade, sexo e renda. Como pode-se verificar na Figura 22 mais da metade dos clientes possuem entre 15 e 24 anos, o que direciona a tomar decisões como escolhas de filmes visando este público alvo, tentando abranger um público mais adulto ao mesmo tempo, de 25 a 34 anos.

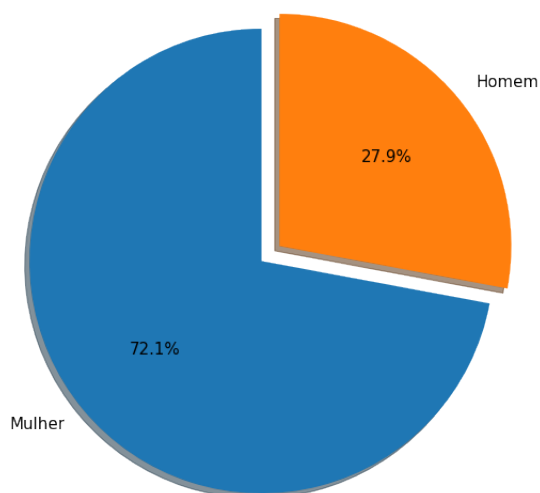
Figura 22 – Distribuição da idade dos frequentadores



Fonte: Elaborado pelo autor

Já no que diz respeito ao sexo, surgiu um *insight* interessante aos empresários, uma vez que a Figura 23 mostra que a grande maioria (72,1%) das pessoas que responderam o formulário são do sexo feminino. Vale ressaltar que não necessariamente a maioria do público seja feminino, pode ser que mulheres utilizem mais redes sociais, que foi onde o formulário foi disponibilizado, mas de qualquer forma essa análise sugere sim que há um grande público feminino interessado em assistir filmes no cinema, o que indica que uma possível promoção para mulheres possa atrair ainda mais esse segmento.

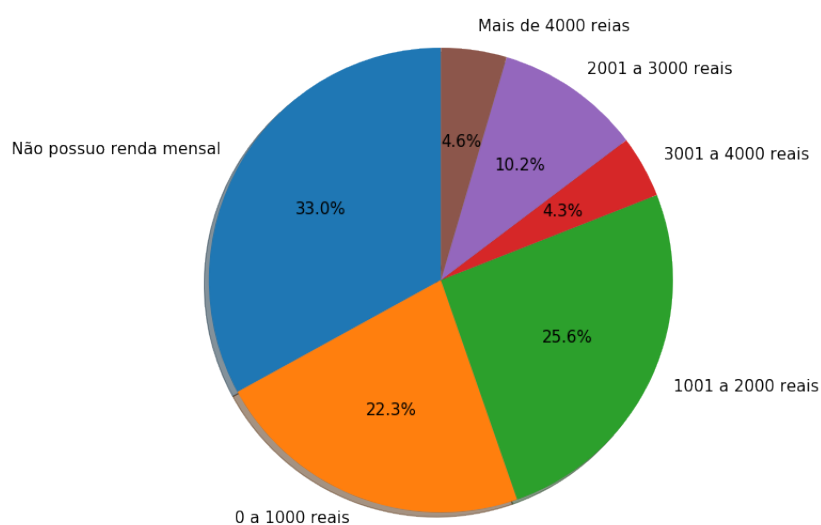
Figura 23 – Distribuição de sexo dos frequentadores



Fonte: Elaborado pelo autor

Quanto a renda do público do cinema apresentada na Figura 24, vale fazer um comparativo entre as pessoas que não frequentam o cinema, pois é nítida a diferença da renda entre os dois grupos, apesar de que ainda nota-se uma predominância de pessoas que possuem menos de R\$1000,00 mensais (55,2%) e menos de R\$2000,00 (80,9%) mensais, o que evidência que as pessoas interessadas no cinema não possuem uma renda muito alta.

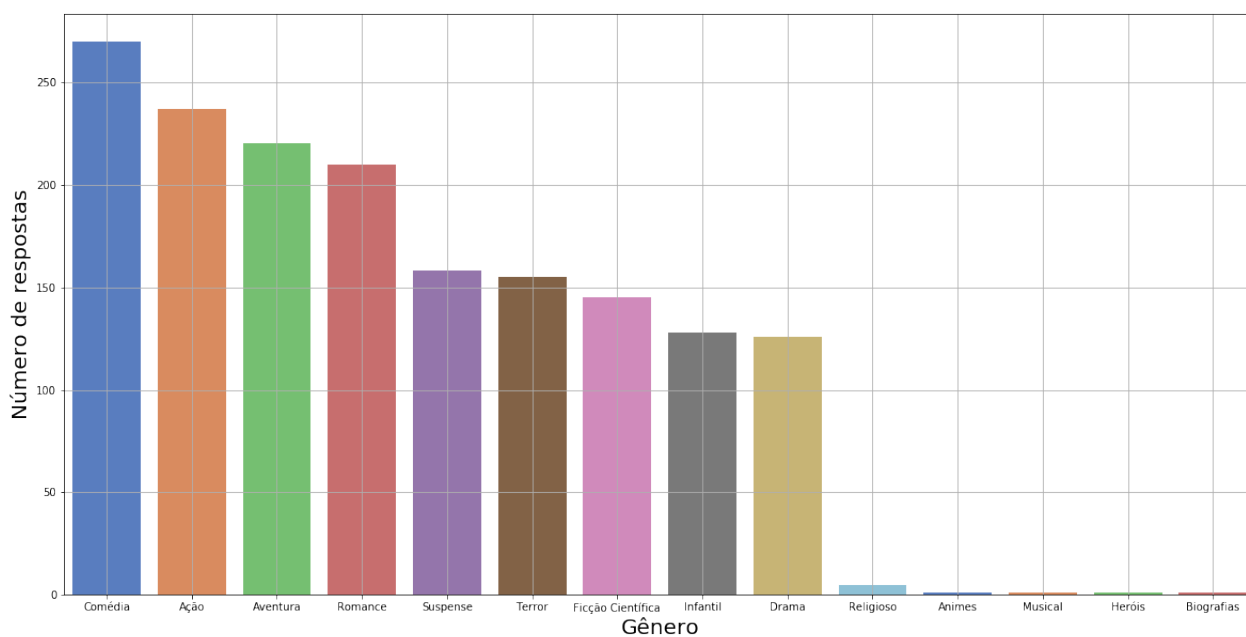
Figura 24 – Distribuição de renda dos frequentadores



Fonte: Elaborado pelo autor

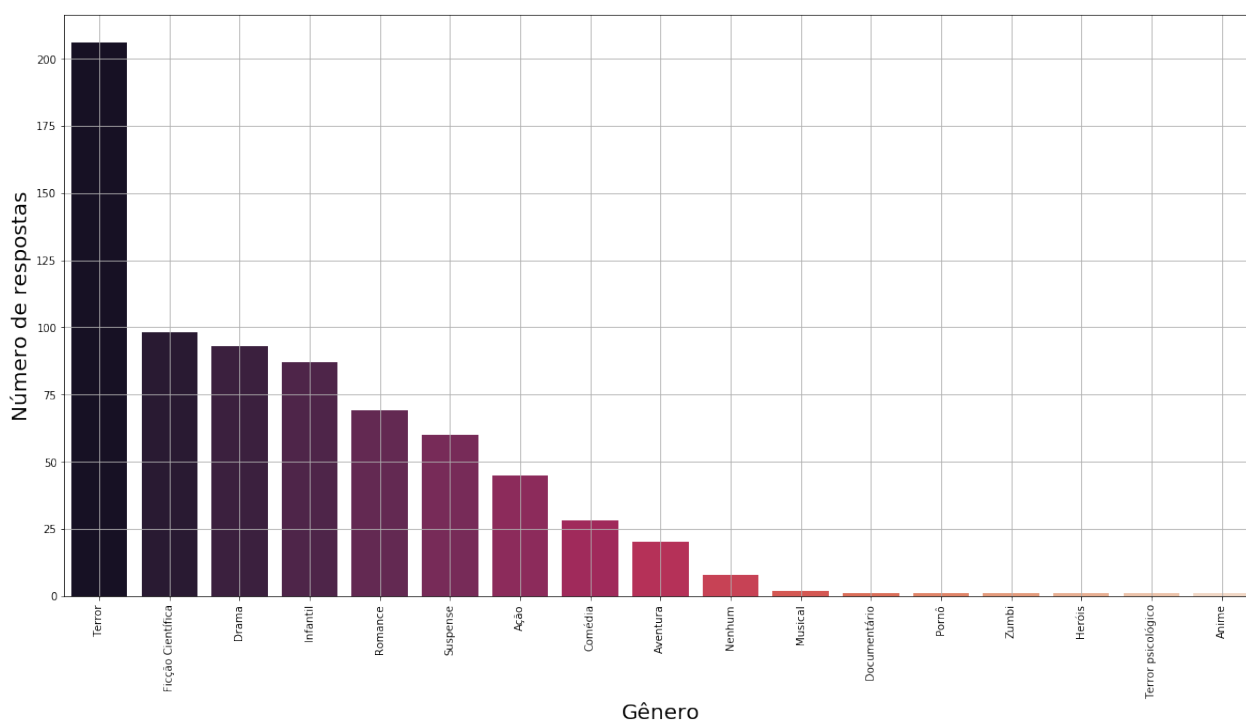
Analisando agora os dados sobre os gêneros de filmes preferidos e evitados pelo coletivo. Esta informação é de extrema relevância aos executivos, uma vez que, por ser um cinema de pequeno porte, existem algumas restrições quanto a exibição de filmes. Por possuir apenas uma sala, há um limite de exibir apenas dois filmes ao dia, o que torna crítica a escolha do que será exibido, uma vez que se uma escolha errada for feita, a semana inteira pode ser perdida, além de prejudicar o prazo para exibições de outras obras.

Figura 25 – Gêneros de filmes preferidos



Fonte: Elaborado pelo autor

Figura 26 – Gêneros de filmes evitados



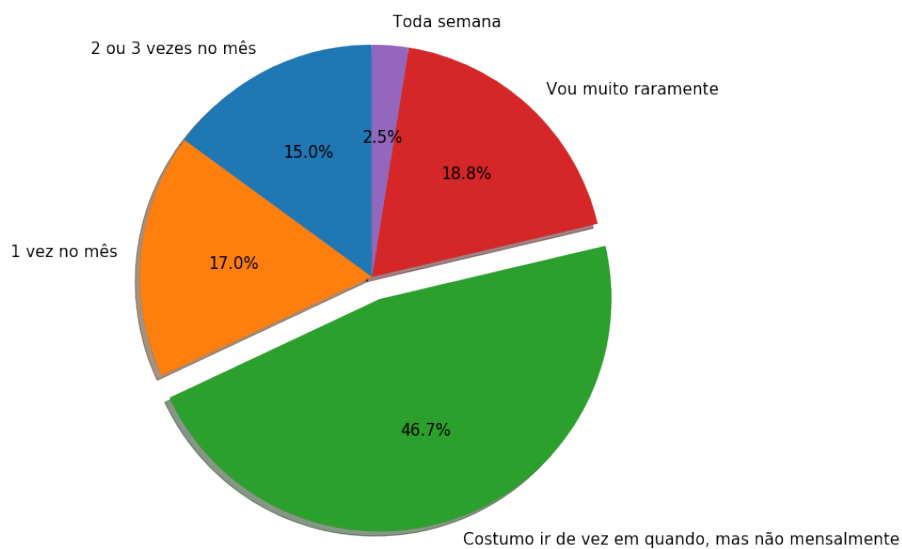
Fonte: Elaborado pelo autor

Na Figura 25 nota-se um equilíbrio entre os gêneros preferidos, sem haver algum que se sobressaia demasiadamente sobre os demais, apesar de possuir quatro que se destacam, que são comédia, ação, aventura e romance. Por outro lado, pelo fato do formulário possuir a alternativo 'Outro' alguns estilos possuem pouquíssimos votos, até por possuírem outros temas em que podiam ser encaixados, como o caso do gênero "heróis" que poderia ser classificado como ação ou aventura.

Em relação aos gêneros evitados, apresentados na Figura 26, é possível identificar uma forte rejeição aos filmes de "terror", o que auxilia a pensar bem antes de decidir por exibir um filme de terror, tendo em vista a grande repulsa pelos frequentadores perante este estilo. Vale pontuar também alguns gêneros que possuem uma quantidade de pessoas que gostam e evitam semelhantes, que são os casos de ficção científica, drama e infantil.

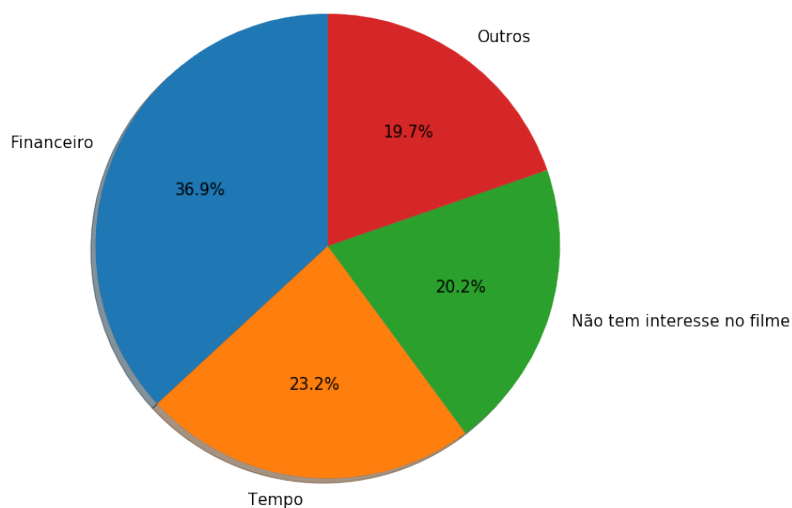
É importante também entender a frequência com que os clientes costumam ir ao cinema mensalmente para posteriormente entender os motivos para não ir mais vezes. Unindo as informações que os gráficos das Figuras 27 e 28 revelam, é possível notar que a grande maioria do público não vai frequentemente ao estabelecimento, principalmente por motivos financeiros e de tempo, seguidos por falta de interesse no filme exibido.

Figura 27 – Frequência dos clientes



Fonte: Elaborado pelo autor

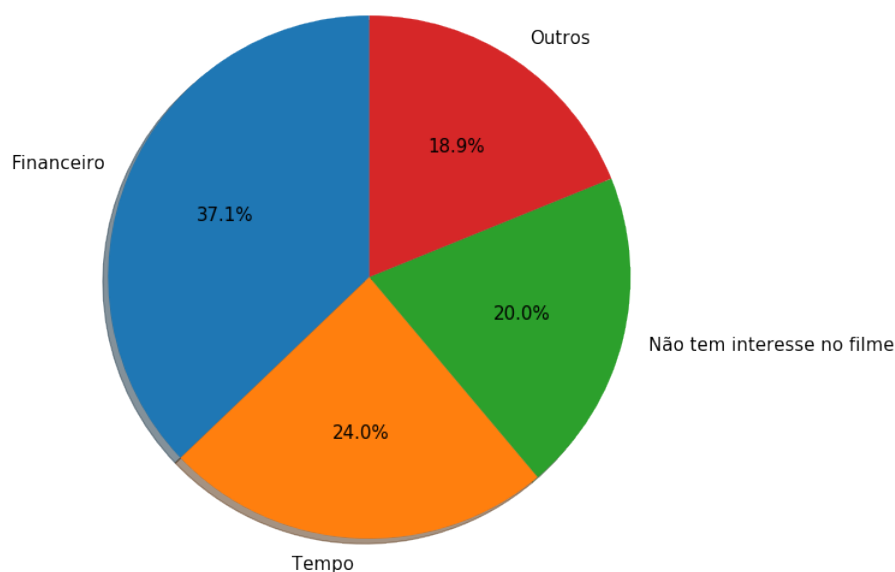
Figura 28 – Motivos da frequência dos clientes



Fonte: Elaborado pelo autor

Para compreender melhor esta distribuição, foi feita uma análise condicional filtrando o Data Frame apenas por pessoas que responderam ir de vez em quando, mas não mensalmente e as que vão muito raramente, exibida na Figura 29.

Figura 29 – Motivos da frequência dos clientes filtrada



Fonte: Elaborado pelo autor

Ficou claro que o principal motivo para que essas pessoas não sejam muito ativas nas sessões é por conta de falta de dinheiro ou de tempo.

Vale ressaltar nessa análise a estratégia que foi realizada para efetuar o agrupamento das respostas dos motivos das frequências, tendo em vista que as pessoas podiam escrever qualquer coisa. Essa estratégia foi utilizada em todos os campos que continham resposta livre e havia a necessidade do agrupamento para melhor compreensão dos dados. Em um primeiro momento foi visualizado um geral das respostas e as mais comuns utilizando a biblioteca *collections* do Python e sua função *Counter*, como mostrado na Figura 30, buscando entender os diferentes grupos que poderiam ser criados e também extrair palavras chaves para classificar determinada frase em um destes grupos.

Figura 30 – Função *Counter()*

```
In [9]: Counter(list(df_freq[df_freq['MotivoFrequencia'] != 'nan']['MotivoFrequencia'])).most_common()

Out[9]: [('Falta de tempo', 17),
('Falta de dinheiro', 10),
('Dinheiro', 7),
('Falta de dinheiro ', 4),
('Sem dinheiro ', 4),
('Financeiro', 3),
('Falta de tempo ', 3),
('Falta de grana', 3),
('Dinheiro ', 3),
('Sem dinheiro', 2),
('Não tenho dinheiro ', 2),
('Sem tempo', 2),
('Tempo', 2),
('Tempo e dinheiro', 2),
('A mãe não deixar ', 1),
('Poucos filmes me interessam, além da falta de companhia. ', 1),
('Tempo e dinheiro que não sobra. :/', 1),
('Falta de transporte e dinheiro ', 1),
('Outros programas', 1),
('Não tenho dinheiro', 1)]
```

Fonte: Elaborado pelo autor

Em seguida, foi realizada uma *list comprehension* para separar as tuplas em listas distintas e foi utilizado um *loop* que iteravam sobre essas estruturas buscando classificar as respostas. Caso o motivo não se enquadrasse em nenhum agrupamento, ela seria classificada como "Outro".

Figura 31 – Tratativas de respostas livres

```
mots = [c[0] for c in counter]
cnt = [c[1] for c in counter]

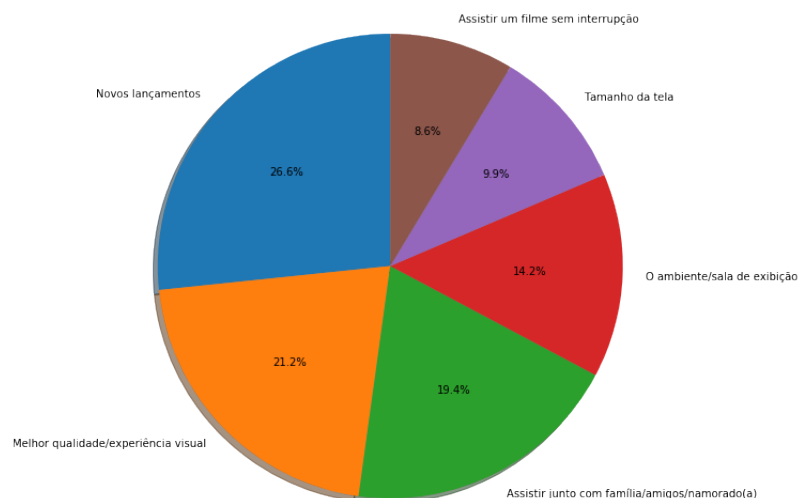
falta_tempo = 0
falta_dinheiro = 0
filme_desagrado = 0
filmes_array = []
for i, motivo in enumerate(mots):
    if 'tempo' in motivo.lower() or 'disponibilidade' in motivo.lower():
        falta_tempo += cnt[i]
    if 'dinheiro' in motivo.lower() or 'financeiro' in motivo.lower() or 'financeiros' in motivo.lower() or 'grana' in motivo.lower():
        falta_dinheiro += cnt[i]
    if 'filme' in motivo.lower() or 'filmes' in motivo.lower():
        filme_desagrado += cnt[i]
        filmes_array.append(motivo)

outros = cnt.sum() - falta_dinheiro - falta_tempo - filme_desagrado
```

Fonte: Elaborado pelo autor

Definida e explicada a estratégia, foram realizadas as análises referentes ao formulário. A próxima questão trata dos motivos que encorajam uma pessoa a ir no cinema, essa questão foi bem aceita pelos executivos para entenderem quais os pontos principais em seu negócio e o que o público vê como positivo. A Figura 32 mostra uma grande distribuição entre as alternativas, com um destaque para novos lançamentos, melhor qualidade/experiência visual e assistir com alguma companhia, o que sugere algumas observações como sempre tentar exibir lançamentos, dedicar recursos a melhorar qualidade visual e criar promoções visando família, amigos e casais.

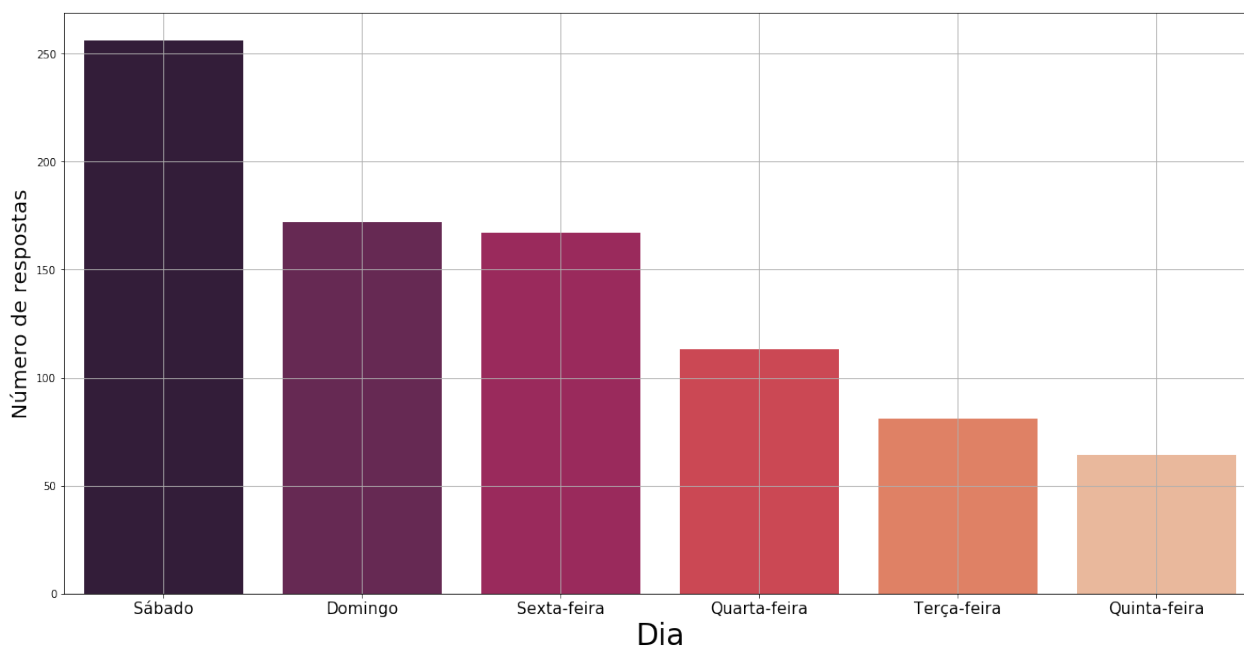
Figura 32 – O que encoraja a assistir um filme



Fonte: Elaborado pelo autor

Outras duas questões que apontam algumas soluções inteligentes referem-se aos dias da semana que o público considera mais adequado ir a uma sessão e com quem eles costumam ir. A Figura 33 mostra que a maioria do público prefere o sábado, seguido do domingo e sexta-feira, enquanto quinta-feira é o dia menos preferido das pessoas, talvez caiba alguma promoção de quinta-feira para atrair mais o público.

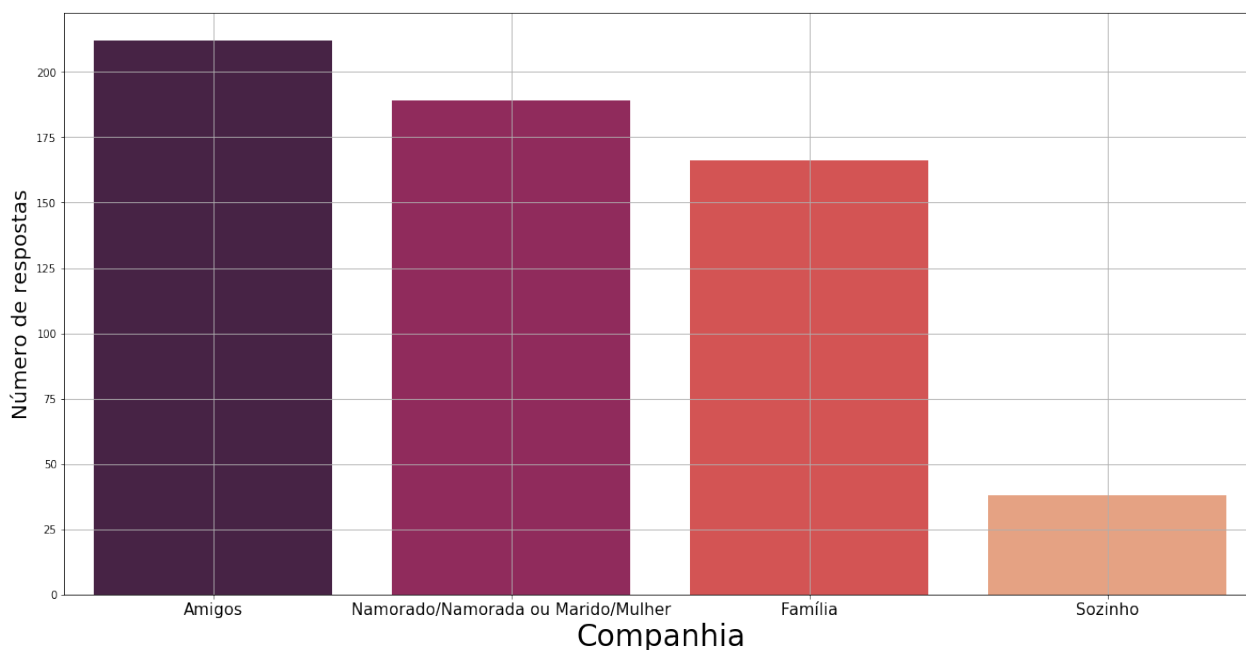
Figura 33 – Melhores dias para ir ao cinema



Fonte: Elaborado pelo autor

Já o gráfico da Figura 34 mostra com quem as pessoas costumam assistir a um filme, mostrando que pouca gente costuma ir sozinho, dando preferência para ir com amigos, namorado(a) ou marido(a) e família, respectivamente. É interessante essa análise para levantar sugestões de combos ou promoções para esses agrupamentos buscando atraí-los cada vez mais.

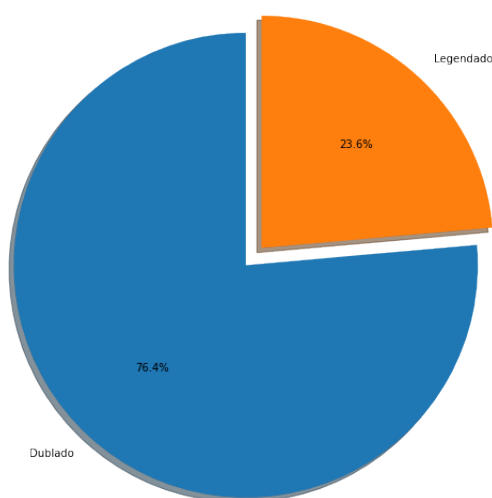
Figura 34 – Companhias



Fonte: Elaborado pelo autor

Somada a essas análises, outras duas perguntas pertinentes aos donos dizem respeito aos horários preferidos dos clientes e se preferem sessão dublada ou legendada. Sempre surgiam dúvidas em relação a como distribuir as sessões entre legendadas e dubladas entre os executivos, e a Figura 35 consolidou a necessidade de haver mais sessões dubladas que legendadas, tendo em vista a preferência esmagadora por esse tipo de áudio. Outra mudança que ocorreu com base nesse estudo combinado com a Figura 33 foi alterar a sessão legendada de sábado para domingo, já que não fazia sentido o dia preferido do público ir contra o tipo de áudio predileto.

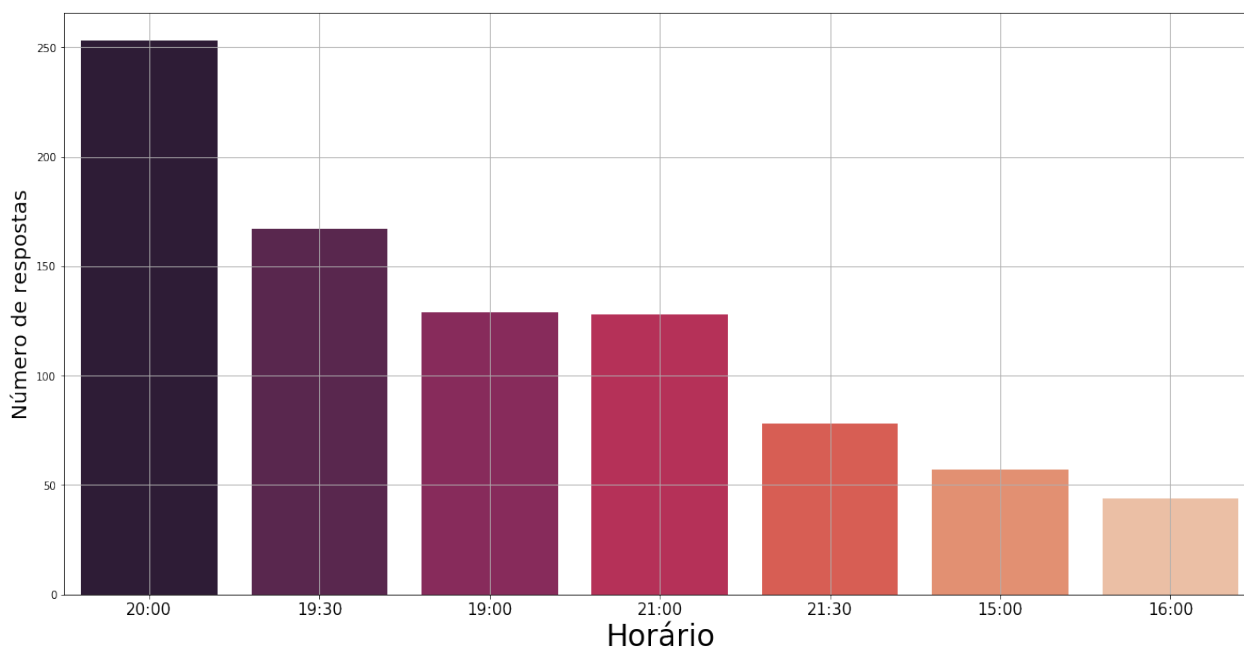
Figura 35 – Áudio preferido dos clientes



Fonte: Elaborado pelo autor

Em relação aos horários preferidos das pessoas, as opções eram limitadas devido as condições impostas pelo estabelecimento, onde os executivos possuíam algumas restrições quanto a horários muito tarde da noite e de manhã. A Figura 36 mostra que eles estavam acertando nesse quesito, pois sempre houve sessões às 20:00. Infelizmente por possuir apenas uma sala, o cinema não consegue realizar sessões às 20:00 e às 19:30, 19:00 ou 21:00, o que limita as opções de horários e combinações.

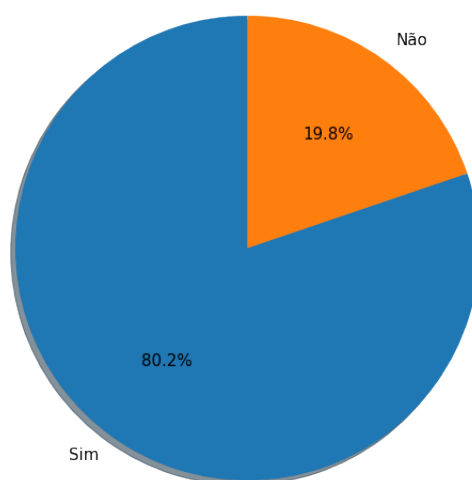
Figura 36 – Horários preferidos dos clientes



Fonte: Elaborado pelo autor

Em seguida foi feita a análise de dados referentes ao ingresso do cinema, buscando um preço mais adequado que se encaixe ao perfil do público e maximizar os ganhos, podendo utilizar as respostas de renda mensal para auxiliar na tomada dessa decisão. A Figura 37 mostra que a maioria do público (80,2%) considera o preço do ingresso justo, o que sugere que se for necessária alguma alteração e redução de preços, o ingresso não deve ser o foco para esse momento, até porque, segundo os executivos, a alteração do preço do ingresso sugere muito mais burocracias do que apenas aumentar ou diminuir o preço, é necessário entrar em acordo com as distribuidoras.

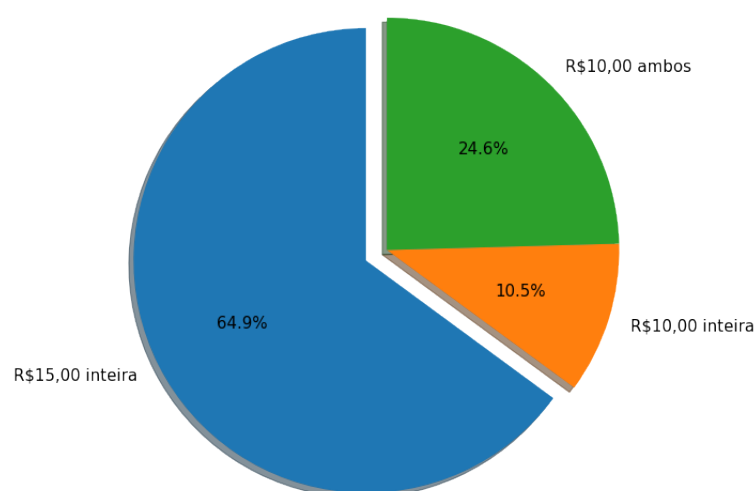
Figura 37 – Considera preço do ingresso justo



Fonte: Elaborado pelo autor

Seria interessante também saber que preço do ingresso os clientes considerariam mais adequado e se seria possível atender essa sugestão. Para analisar a essa pergunta, foi necessário utilizar novamente a estratégia das figuras 30 e 31, pois a resposta era um texto livre. Na Figura 38 é possível visualizar as principais sugestões das pessoas, onde 64,9% consideram R\$15,00 a inteira um bom preço. Essas respostas foram levadas aos executivos que optaram por não abaixar o preço do ingresso, tendo em vista a Figura 37 e todas as dificuldades que envolvem a mudança do preço do ingresso.

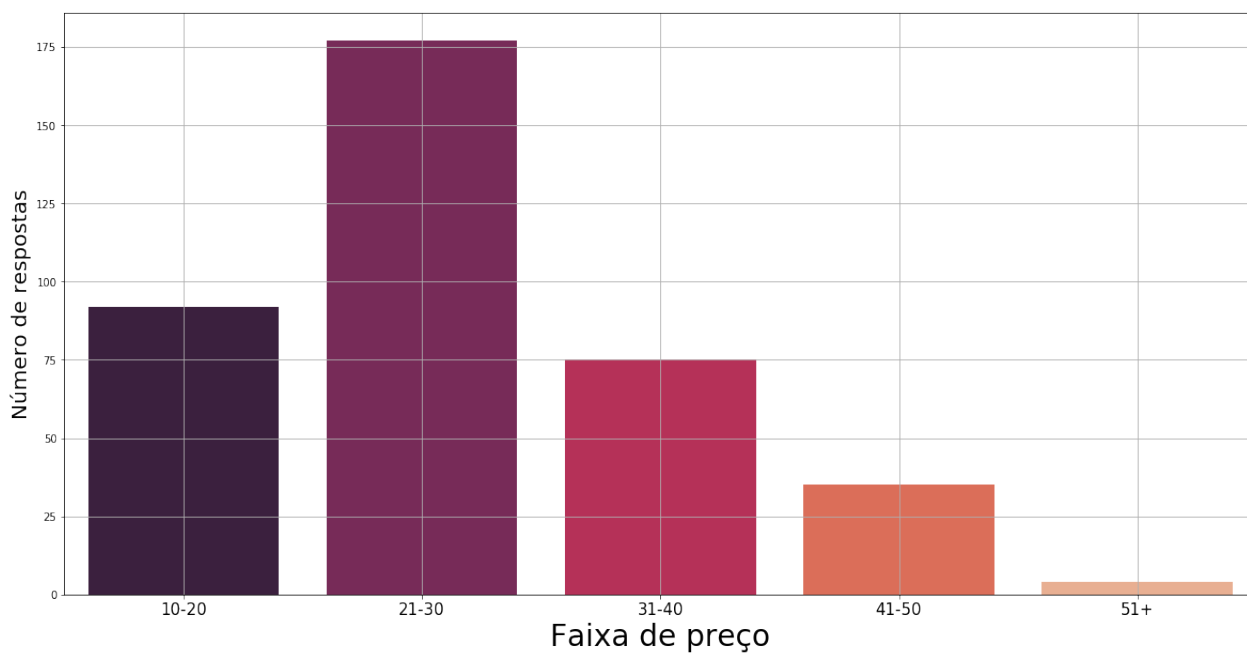
Figura 38 – Sugestões do preço do ingresso



Fonte: Elaborado pelo autor

Tendo a noção de que o ingresso não teria seu preço abaixado, outra questão que pode auxiliar a tomar soluções inteligentes quanto a promoções e combos buscando atrair mais pessoas ao cinema é a Figura 39 onde as pessoas responderam o quanto elas consideram adequado gastar em uma sessão por pessoa. Nota-se que a maioria considera justo gastar entre 21 e 30 reais, entretanto, se for considerar uma pessoa que paga a inteira do ingresso (20 reais) sobra 10 reais nessa faixa para gastar com produtos, enquanto uma pessoa que paga 10 reais do ingresso tem 20 reais de folga para gastar com consumíveis, portanto é interessante fazer uma avaliação dessa resposta por idade.

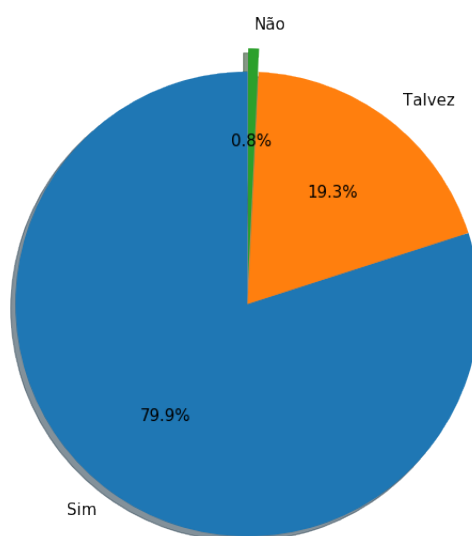
Figura 39 – Gasto considerado justo por sessão



Fonte: Elaborado pelo autor

Os donos do cinema queriam ter conhecimento também sobre como o público enxergaria a possibilidade de haver benefícios para clientes fiéis e assim elaborar algum plano de fidelidade. A resposta da pergunta agradou aos executivos que começaram a pensar sobre levar essa idéia adiante e traçar estratégias para implementar essa solução, pois como mostra a Figura 40, 79,9% das pessoas acreditam que iriam com mais frequência ao cinema caso houvesse esse benefício, 19,3% talvez fossem mais vezes e apenas 0,8% acreditam que isso não as afetaria.

Figura 40 – Benefícios para fiéis incentivaria ir ao cinema



Fonte: Elaborado pelo autor

Para fechar a análise do formulário, algumas questões foram feitas para que os clientes votassem de 0 a 5 alguns pontos em relação ao cinema, então foi feita uma média da pontuação. As perguntas com as respectivas notas são:

- Nível de satisfação em relação aos filmes exibidos: 4.45
- Preço dos produtos oferecidos para consumo: 3.48
- Divulgação dos filmes: 4.3
- Promoções e combos oferecidos: 3.8
- Atendimento no cinema: 4.51
- Sessão no geral (conforto da poltrona, qualidade de som e imagem, etc): 4.31
- Nota geral para o cinema: 4.47

As notas apresentam uma média muito boa, o que agradou aos responsáveis e aqueles que colaboram com o desenvolvimento do cinema. O maior ponto de atenção deveria ser mesmo a criação de novas promoções e combos com base na pesquisa e nos dados apresentados acima.

4.2.3 Dados financeiros

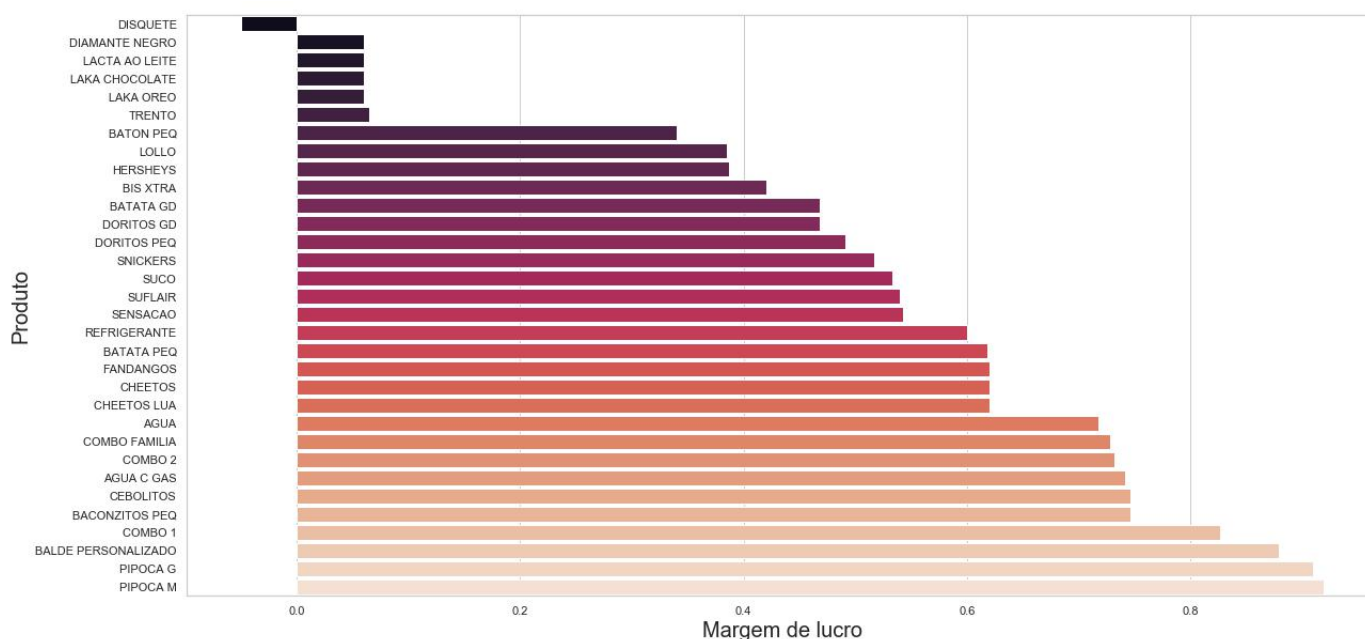
Depois de explorar o formulário, outro levantamento de dados foi realizado utilizando o programa de vendas do cinema. Porém, alguns problemas ocorreram para manipular os dados

do software em questão, que foram a falta de dados estruturados, sejam eles em SQL ou uma planilha, pois era possível fazer a coleta apenas de um arquivo .png contendo um relatório de vendas com base em um determinado intervalo de tempo, e outro problema foi o fato de dados anteriores a 14 de março de 2019 serem perdidos, portanto as análises feitas a seguir são a partir dessa data.

Buscou-se inicialmente estudar os produtos que eram oferecidos para consumo e verificar quais poderiam ser explorados para possíveis soluções, então foi realizada a coleta de um relatório de vendas desde a reabertura do cinema, que ocorreu dia 18 de dezembro de 2018, até meados de julho de 2019, então esses dados foram estruturados para ser possível manipulá-los, onde foi extraído informações como o lucro bruto e o preço unitário por produto. Em um segundo momento, foi pedido aos executivos que levantassem os custos para cada um, para que assim pudesse verificar o lucro líquido e a margem de lucro.

Com esses dados estruturados, foi realizado o processo da mineração, que, apesar de possuir poucas análises possíveis a serem feitas, ainda foram extraídas informações valiosas referentes a informações financeiras do negócio. A Figura 41 mostra a margem de lucro por produto, onde chama a atenção o fato do Disquete ter uma margem de lucro negativa, conhecimento esse que os executivos não possuíam. É possível notar também dois extremos do gráfico, alguns itens que possuem margem de lucro baixíssimas, enquanto outros possuem ótima margem de lucro, que são opções que envolvem pipoca, seja ela própria ou algum combo.

Figura 41 – Margem de lucro por produto



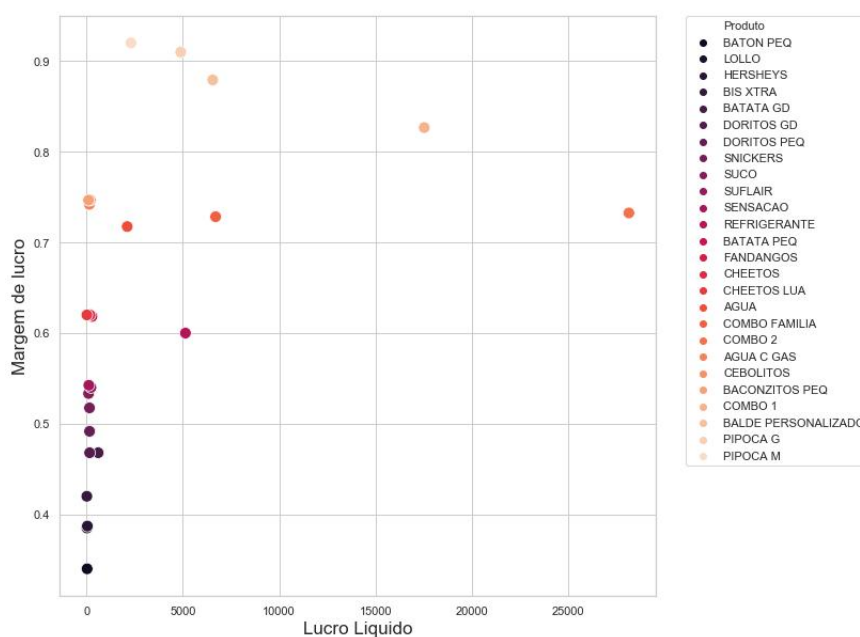
Fonte: Elaborado pelo autor

Entretanto, é importante destacar que uma boa margem de lucro não significa ne-

cessariamente grandes retornos financeiros ao negócio, é necessário avaliar também o lucro líquido daquele produto para assim determinar quais devem ser as soluções tomadas e atacar os melhores itens. Com base nesse pensamento, a Figura 42 deixa explícito a importância dessa análise, onde pode-se notar alguns pontos interessantes, como o fato da Pipoca M, apesar de ter a maior margem de lucro, não teve um grande lucro líquido, por outro lado o Combo 2 não possui uma margem de lucro muito grande, mas é o item que mais deu retorno financeiro até a data do levantamento dos dados.

Uma informação relevante que pôde ser extraída desse gráfico é a comparação entre o Balde Personalizado e o Combo Família, pois os dois possuem praticamente o mesmo lucro líquido, porém a margem de lucro do balde é maior que a do combo, apesar disso o balde não possui nenhuma promoção ou combo em cima dele, o que poderia ser explorado pelos executivos, uma vez que o produto é muito bem aceito pelo público.

Figura 42 – Margem de lucro X Lucro líquido



Fonte: Elaborado pelo autor

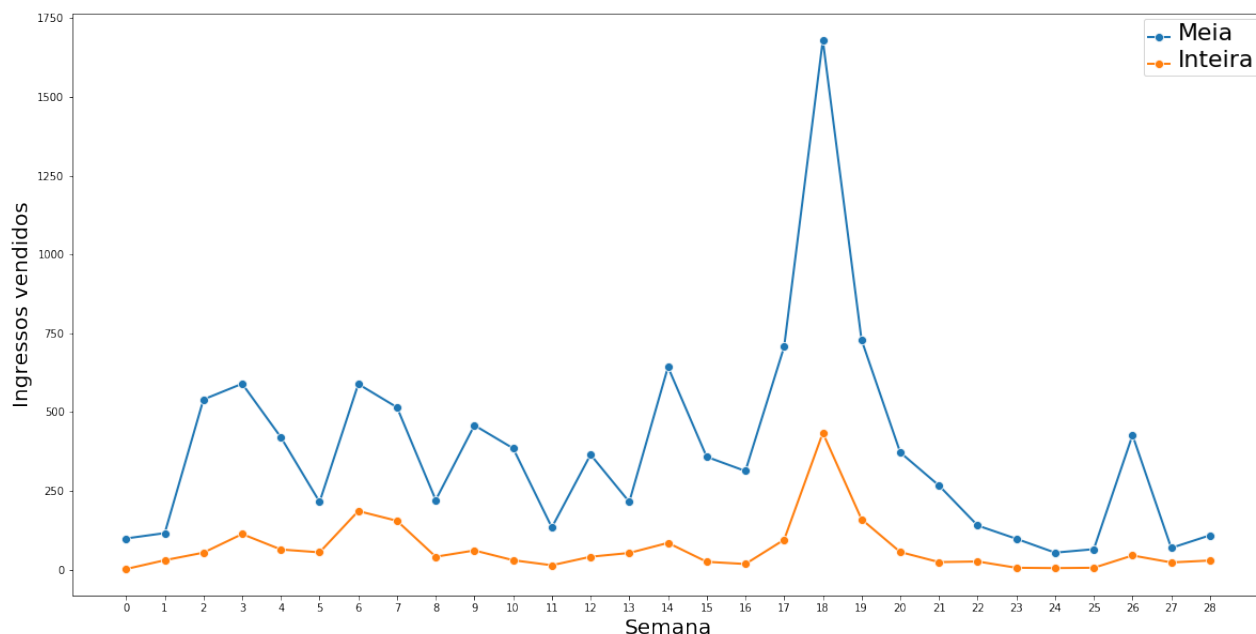
5 RESULTADOS

BI pode ser traduzido como inteligência de negócios, ou inteligência empresarial. Isto significa que é um método que visa ajudar as empresas a tomar as decisões inteligentes, mediante dados e informações. É papel do cientista de dados levar essas informações processadas para os executivos da empresa para que eles decidam o que é melhor para o futuro dos negócios.

Com os resultados das análises apresentadas aos executivos do cinema, vários pontos foram discutidos para serem colocados em prática. Questões que envolviam precificação eram mais sensíveis, pois para alterar o preço do ingresso, é necessário passar por toda uma burocracia, entretanto, como foi mostrado na Figura 37, as pessoas, em sua grande maioria, estão de acordo com o preço do ingresso. Porém a nota para combos e promoções mostrou que era necessário tomar alguma atitude quanto a precificação dos produtos ou criar novas soluções, então várias estratégias foram discutidas, como por exemplo criar uma promoção toda quinta-feira (dia menos movimentado) para as famílias, (vide a análise da Figura 34), criar mais combos, utilizando o Balde de Pipoca em um deles tendo em vista a margem de lucro desse produto, bem como seu lucro líquido. Foi levantada a possibilidade de criar um dia de promoção para mulheres, pois, segundo a Figura 23 a maioria do público é mulher, promoção essa que existia em outras épocas do cinema.

No fim, os executivos fizeram poucas mudanças quanto a precificação do cinema, entretanto tomaram várias decisões baseadas nas análises dos dados e quiseram por a prova algumas dessas análises. Utilizando do programa de venda, foi feita uma coleta de dados semana por semana do cinema, para ter uma visão mais detalhada dos lucros e do público que frequentou o local (por semana). Infelizmente, como foi dito na sessão 4.2.3, os dados anteriores a 14 de março foram perdidos. A Figura 43 mostra a quantidade de ingressos vendidos, separados por meia e inteira, onde consegue-se extrair algumas informações, como o fato de serem vendidos muito mais ingressos meia entrada do que inteiras, e o gráfico seguir um sequência de picos e vales.

Figura 43 – Total de ingressos vendidos por semana



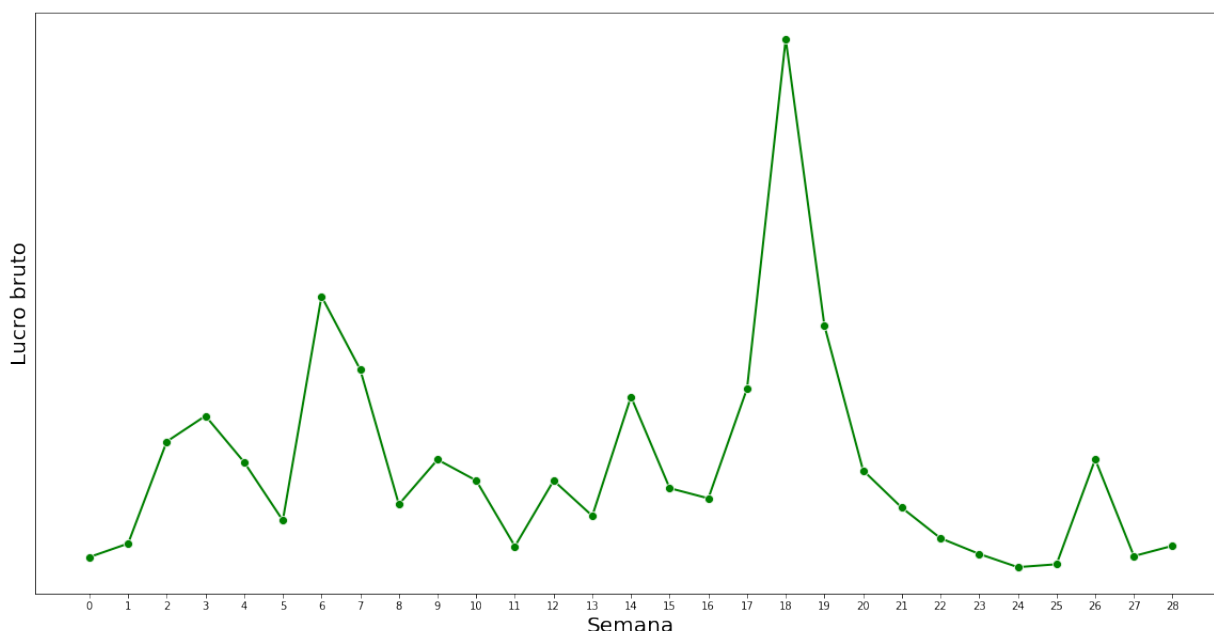
Fonte: Elaborado pelo autor

Com uma análise mais detalhada, tendo conhecimento dos filmes que foram exibidos em cada semana, verifica-se que filmes para adolescentes e crianças fazem um grande sucesso no cinema. Os principais sucessos, onde nota-se um pico de público, foram "Shazam!", "Os Vingadores: Ultimato", "Detetive Pikachu", "Toy Story 4" e "O Rei Leão" (semanas 3, 6, 7, 9, 10, 14, 18 e 19), outros picos foram de filmes sucessos de bilheteria pelo mundo, como são os casos de "Annabelle 3" que foi exibido ao mesmo tempo que "Turma da Mônica: Laços" (semana 17), outro filme infantil, e "IT 2" (semana 26), filme de terror extremamente famoso. Ponto de atenção principalmente para "O Rei Leão", que superou todos os filmes com sobras, tanto na venda de meia-entrada quanto inteira. Vale ressaltar também que, por ser um cinema localizado em uma cidade pequena, não é válido exibir o mesmo filme por três semanas ou mais, sendo a terceira semana válida somente para filmes que estão fazendo grande sucesso, tendo em vista por exemplo o caso dos Vingadores, que teve um público de 775 e 670 ingressos nas primeiras duas semanas, respectivamente, e 262 na terceira.

Por outro lado, os vales no gráfico em sua grande maioria são filmes para o público mais adulto, que são os casos de "John Wick 3", "Kardec" e "Velozes e Furiosos: Hobbs & Shaw" (semanas 11, 13, 22 e 23). Uma observação interessante é que, o vale que sucede "O Rei Leão", são de filmes de terror que não possuem tanto sucesso, que são "Brinquedo Assassino" e "Histórias Assustadoras Para Contar no Escuro" (semanas 24 e 25), o que comprova que filmes de terror, segundo a Figura 26, são realmente evitados pelo público. O mesmo pode-se dizer de "IT 2", que apesar da primeira semana ter dado um ótimo público com a ajuda de "Meu Amigo Enzo", não conseguiu se sustentar sozinho na segunda semana.

Para fechar as análises, a Figura 44 mostra o lucro bruto por semana, incluindo os ganhos tanto de ingressos quanto de produtos. Pode-se notar que ele segue uma distribuição muito semelhante a vendas de ingressos, mas dois pontos em questão chamam a atenção, que são as semanas 3 e 6. Ao olhar novamente a Figura 43 nota-se que estas duas semanas possuem uma quantidade de ingressos meia-entrada muito parecidos, entretanto pelo gráfico abaixo notamos um renda da semana 6 muito superior a semana 3. Isso ocorre porque de uma para a outra, há uma pequena diferença de ingressos inteira vendidos, que fez com que alavancasse os ganhos de uma semana para a outra. Essa análise reforça a ideia de que é importante criar soluções e incentivar as pessoas irem em família ao cinema, que incentivem os pais irem junto com as crianças, pois isso faz com que o lucro do cinema aumente consideravelmente.

Figura 44 – Lucro bruto por semana



Fonte: Elaborado pelo autor

6 CONCLUSÃO

Com as análises feitas, os executivos possuem as informações necessárias para tomar as decisões no futuro. Cabe ao cientista de dados coletar, manipular e analisar os dados, minerar possíveis soluções com base nas análises e saber repassar esse conhecimento aos donos do negócio de forma simples e bem estruturada.

Apesar do programa responsável por salvar as informações financeiras do cinema ser obsoleto e não ter sido possível fazer um *backup* adequado desses dados, ainda foi atingido um

nível de satisfação adequado que agradou os executivos do negócio, pois eles obtiveram uma visão do perfil dos clientes e também do empreendimento que eles não tinham anteriormente. Ainda que não tenha sido posto em prática muitas mudanças durante o desenvolvimento da pesquisa, o *feedback* foi extremamente positivo, tanto que foi pedido para que seja utilizado data science em um negócio de comunicação deles com a mesma estratégia.

Por fim, seria interessante acompanhar o desenvolvimento e o planejamento futuro do cinema, como eles irão se comportar diante dos insumos levantados, até que ponto isso influenciará na tomada de decisões. Acompanhar os lucros e a satisfação da clientela após essas mudanças seria o cenário ideal para validar que o uso de BI em uma empresa de entretenimento.

Referências

- CANALCOMSTOR. *QUAIS AS DIFERENÇAS ENTRE BUSINESS INTELLIGENCE E DATA SCIENCE?* 2018. Disponível em: <<https://blogbrasil.comstor.com/quais-as-diferencas-entre-business-intelligence-e-data-science>>. Acesso em: 14 mai. 2019.
- D. P. SILVEIRA. *O que é Data Science?* 2016. Disponível em: <<https://www.oficinadanet.com.br/post/16919-o-que-e-data-science>>. Acesso em: 17 mar. 2019.
- DEVMEDIA. *Conceitos e Técnicas sobre Data Mining*. 2011. Disponível em: <<https://www.devmedia.com.br/conceitos-e-tecnicas-sobre-data-mining/19342>>. Acesso em: 16 mai. 2019.
- FAYYAD, U.; PIATETSKY-SHAPIO, G.; SMYTH, P. The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, ACM, v. 39, n. 11, p. 27–34, 1996.
- L. COELHO. *CIÊNCIA DE DADOS: O QUE É, CONCEITO E DEFINIÇÃO*. 2018. Disponível em: <<https://www.cetax.com.br/blog/data-science-ou-ciencia-de-dados/>>. Acesso em: 14 mai. 2019.
- SANTOS, M. Y.; RAMOS, I. *Business Intelligence: Tecnologias da informação na gestão de conhecimento*. [S.l.]: FCA-Editora de Informática, Lda, 2006.
- SHARMA, M. Data visualization using seaborn. 2018. Disponível em: <<https://towardsdatascience.com/data-visualization-using-seaborn-fc24db95a850>>. Acesso em: 16 mai. 2019.
- SIEGEL, I. F. *Linguagem python e suas aplicações em ciência de dados*. Niterói, 2018.
- SILVA, V. C. L.; TERRA, L. A. A. Business intelligence como fator decisivo na competitividade empresarial: Uma análise a partir de multicaseos. *Revista Inteligência Competitiva*, v. 5, n. 1, p. 1–13, 2015.