UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"
FACULDADE DE CIÊNCIAS - CAMPUS BAURU
DEPARTAMENTO DE COMPUTAÇÃO
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

JOÃO PAULO SCHIAVON

ESTUDO DE CIÊNCIA DE DADOS APLICADA A DADOS ABERTOS GOVERNAMENTAIS

BAURU
Junho/2019

JOÃO PAULO SCHIAVON

ESTUDO DE CIÊNCIA DE DADOS APLICADA A DADOS ABERTOS GOVERNAMENTAIS

Trabalho de Conclusão de Curso do Curso de Ciência da Computação da Universidade Estadual Paulista "Júlio de Mesquita Filho", Faculdade de Ciências, Campus Bauru.
Orientador: Prof. Dr. João Pedro Albino

BAURU
Junho/2019

João Paulo Schiavon Estudo de Ciência de Dados aplicada a dados abertos governamentais/ João Paulo Schiavon. – Bauru, Junho/2019- 28 p. : il. (algumas color.) ; 30 cm.
Orientador: Prof. Dr. João Pedro Albino
Trabalho de Conclusão de Curso – Universidade Estadual Paulista "Júlio de Mesquita Filho"
Faculdade de Ciências
Ciência da Computação, Junho/2019.
1. Ciência de dados 2. Análise de dados 3. Dados Abertos 4. Grandes dados

João Paulo Schiavon

Estudo de Ciência de Dados aplicada a dados abertos governamentais

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

Banca Examinadora

Prof. Dr. João Pedro Albino
Orientador
Universidade Estadual Paulista "Júlio de Mesquita Filho"
Faculdade de Ciências
Departamento de Ciência da Computação

Profa. Dra. Simone das Graças Domingues Prado
Universidade Estadual Paulista "Júlio de Mesquita Filho"
Faculdade de Ciências
Departamento de Ciência da Computação

Dr. Humberto Ferasoli Filho
Universidade Estadual Paulista "Júlio de Mesquita Filho"
Faculdade de Ciências
Departamento de Ciência da Computação

Bauru, _____ de _____ de _____.

Este trabalho é dedicado a minha família e a minha república Havana, pelo apoio incondicional durante toda minha trajetória.

Agradecimentos

Gostaria de começar agradecendo aos meus pais Ângela e Victor, meu irmão João Victor, pela contribuição na essência do ser humano que eu sou hoje, pois esta formação acadêmica existe, é por conta da ajuda de vocês.

Agora que eu já agradeci minha família de sangue, gostaria de agradecer a família que Bauru me fez encontrar, minha república Havana. Obrigado por serem meu ponto de refúgio para todas as vezes que eu pensei em desistir (não foram poucas vezes), e se hoje eu sou capaz de completar esse ciclo, acredito que é por conta da força que vem de vocês.

"Havana quer dizer família"

Queria agradecer a Atlética da Unesp de Bauru pelo ano de 2018, por ter sido uma das experiências mais alucinantes da minha vida. Obrigado por me fazer recuperar todo o amor ao esporte.

"Eu sou Bauru até o último segundo!"

E gostaria de agradecer uma pessoa recém chegada, mas que sem ela eu não estaria escrevendo esse trabalho agora. Obrigado Conterrânea, por todo carinho e compreensão em um momento da minha vida que eu não espera ser compreendido.

Por fim, gostaria de agradecer aos meus professores por contribuírem para minha formação acadêmica.

"Se eu vi mais longe, foi por estar sobre ombros de gigantes"

"Números tem uma importante historia para contar. Eles dependem de você para dar a eles uma voz." – Stephen Few

Resumo

Com a massiva quantidade de dados produzida pela interatividade dos usuários proporcionada pela internet, surgem oportunidades para quem conseguir controlar tais dados e gerar conhecimento útil necessário para questões sociais e econômicas para a sociedade. Dito isto, pode-se observar que grandes empresas e governos vêm realizando trabalhos para oferecer dados abertos online relevantes para que a comunidade possa atuar com os dados compartilhados e agilizar a descoberta de novos conhecimentos, realizando análises mais precisas e aprofundando a interpretação dos dados disponíveis. O presente trabalho objetiva demonstrar de modo prático como a disponibilização de dados governamentais de maneira estruturada pode gerar conhecimentos úteis para a sociedade como um todo ,e para tal, o trabalho usará a base de dados disponível pelo projeto Operação Serenata de Amor (Operação Serenata de amor, 2019) de recibos de reembolsos da cota financeira para exercício parlamentar dos parlamentares da Câmara de deputados do Brasil; e analisará se os reembolsos tem caráter suspeito, assim como feito no projeto OSA, buscando gerar conhecimento útil com tal análise.
Palavras-chave: Grandes dados, Dados abertos, Ciência de dados, Análise de dados.

Abstract

With the production of massive quantity of data by the interaction between internet users, opportunities arrive for those who can control this data and could bring insights about economics and social questions for the society. Said that, it can be observed that big enterprises and governments are working to offer online open data to the community use that shared data and speed up the discovery of new knowledge, doing more accurate analysis and deeper interpreting with the available data. The actual work have the objective to demonstrate practically how availability of government data in a structured way could make insights for the society as a whole, and for such, this work will use the data base available for the project Operação Serenata de Amor, for exploring reimbursement of the financial quota for the parliamentary in Brazil and analyze if they are suspect or not, seeking out to generate insights with that analysis.
Keywords: Big Data, Open Data, Data Science, Data Analytics.

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

OSA	Operação Serenata de Amor
LAI	Lei de Acesso à Informação
AED	Análise Exploratória de dados
DAG	Dados Abertos Governamentais

Sumário

1 Introdução

Desde sua criação, a *World Wide Web* impactou de maneira radical o modo como as informações são compartilhados, seu impacto atenuou as dificuldades de acesso e disseminação de conhecimentos globalmente (BIZER; HEATH; BERNERS-LEE, 2009). Atualmente, tem se tornado mais cotidiano e rotineiro o uso intenso de tecnologias de informação e de comunicação, que acompanhadas do crescimento da *World Wide Web*, culminam em uma quantidade massiva de dados (BANDEIRA et al., 2014), que em sua maioria são dados em diversos formatos e normalmente com a limitação de serem adequados apenas para o consumo humano (WOOD et al., 2014). Dificultando assim, o consumo, acesso e interpretação das informações de forma automática por máquinas e especialistas.

No Brasil, o acesso às tecnologias de informação tem aumentado ano após ano, como constatado pela (CETIC, 2017). No ano de 2008 apenas 18% dos domicílios que participaram da pesquisa tinham acesso à internet, já no ano de 2017 esse número subiu para 61%, um demonstrativo bem expressivo de como a presença da internet tem sido cada vez mais marcante na vida dos brasileiros. Com esse número expressivo de domicílios que possuem internet, surgem novas possibilidades de relacionamento entre o governo e a sociedade civil, propiciando crescimento na democracia participativa e contribuindo para a promoção de direitos da cidadania (VAZ; RIBEIRO; MATHEUS, 2013). Um desses direitos é o direito ao controle social do governo, como dito por (VAZ; RIBEIRO; MATHEUS, 2013), faz referência ao "direito dos cidadãos de acompanharem as ações dos agentes públicos e das organizações governamentais", esse direito tem como objetivo facilitar que os cidadãos e suas entidades representativas se sustentem nas práticas de transparência pública e consigam acompanhar iniciativas governamentais e fazer uma fiscalização mais rígida sobre os órgãos públicos.

> "Atualmente, são investidos diversos recursos para a melhoria da comunicação entre o Governo e os cidadãos. Cada vez mais as tecnologias de informação estão sendo utilizadas para aprimorar a qualidade dos produtos e serviços oferecidos pelo Estado à população, que busca uma maior participação nas práticas de gestão pública"(RIBEIRO C. Y VIEIRA PEREIRA, 2015).

Assim, nasce a Lei de Acesso à Informação (LAI), uma lei brasileira baseada no conceito de dados abertos, que conforme diz a ONG *Open Knowleadge International*, os dados aberto são definidos como dados que podem ser reutilizados, usados de graça e redistribuídos livremente por qualquer pessoa, sendo necessário a atribuição da fonte e compartilhamento das regras da mesma apenas para a sua utilização. A aprovação da LAI em 2017 representou um significativo progresso na área de transparência governamental, e a possibilidade de que qualquer cidadão possa ter acesso a informações governamentais, de maneira rápida e descomplicada (SILVA; EIRÃO;

CAVALCANTE, 2014), demonstrando que o governo reconhece que o acesso a informações orçamentárias é fundamental para a transparência do setor público e que a transparência tem um papel importante para aumentar sua efetividade e *accountability* (BEGHIN; ZIGONI, 2014).

Junto da conectividade que a internet representa no dia a dia e com a expansão do número de usuários interligados, a produção de dados se torna cada vez maior. A internet propicia o aumento da interatividade, pois a comunicação entre seus usuários é horizontal, o que reduz drasticamente os custos para organizar e acessar informações (SHIRKY, 2008), assim, para que a internet e sua quantidade massiva de dados se tornem uma ferramenta relevante na vida das pessoas é preciso ter controle dos dados gerados por ela. Então, tem se tornado cada vez mais comum em Economia da Informação, órgãos públicos e privados, que buscando a transparência dos processos e do seu desempenho, publiquem seus dados relevantes online para os compartilhar com o público, para assim a comunidade ajudar com soluções e inferências dos processos internos nos grandes órgãos (ECONOMIST, 2010).

Ser capaz de trabalhar com grandes quantidades de dados derivados de muitas localidades e com formatos variados é uma das habilidades mais desejadas na última década (DAVENPORT; PATIL, 2012). Entretanto, se achar pessoas com habilidades que consigam tratar grandes quantidades de dados já se demonstra difícil, encontrar dados que estejam devidamente estruturados é ainda mais difícil. Resultando assim na diminuição da agilidade e eficácia da produção e extração de conhecimento útil necessário para lidar com questões sociais e econômicas para a sociedade do século 21. Afim de mudar essa situação, diversas empresas e governos vem realizando esforços para oferecer dados e tecnologias *web* que possam agilizar a geração de novos conhecimentos, usando de conceitos como *Web Semântica* (BERNERS-LEE et al., 2001) e *Linked Data* (BIZER; HEATH; BERNERS-LEE, 2009), que são técnicas de desenvolvimento de *software* desenvolvidas para que o uso de dados na *web* seja efetivo (ISOTANI; BITTENCOURT, 2015).

Dito isto, o objetivo deste trabalho, é o de exemplificar de maneira prática como a disponibilização dos dados governamentais de maneira estruturada adequada pode gerar conhecimentos úteis para a sociedade como um todo. Para tal, será usado a linguagem de programação *Python* e o ambiente de desenvolvimento de código aberto *Jupyter Notebook*, o qual permite o desenvolvimento interativo, ajudando assim o melhor estudo dos dados. Para a análise de dados abertos, o trabalho utilizou a base de dados disponível pelo projeto Operação Serenata de Amor (Operação Serenata de amor, 2019) de recibos de reembolsos da cota financeira para exercício parlamentar dos parlamentares da Câmara de deputados do Brasil; e analisará se os reembolsos tem caráter suspeito, assim como no projeto OSA. Parte do objetivo do trabalho é proporcionar maior familiaridade com os temas "grandes dados", dados abertos, tratamento e visualização gráfica de informação, com vistas a criar valor em tais dados aberto, como em (ALBINO, 2016).

1.1 Objetivos Gerais

Realizar uma Análise exploratória de dados para identificar na base de dados do projeto OSA de recibos de reembolso da cota para exercício da atividade parlamentar da câmara de deputados, quais desses recibos tem caráter suspeito.

1.2 Objetivos Específicos

- Estudar técnicas para identificar gastos suspeitos e identificar qual técnica utilizar.
- Estudar o projeto Operação Serenata de Amor para ter como base um projeto real.
- Fazer a captação, limpeza e transformação dos dados.
- Gerar visualizações gráficas das informações.

1.3 Organização da Monografia

Esta monografia está organizada da seguinte forma.

O Capítulo 1 contém a fundamentação teórica.

O Capítulo 2 descreve o processo de desenvolvimento e as ferramentas usadas na análise exploratória.

E por fim, no Capítulo 3, a conclusão finaliza com os resultados obtidos e sugestões de pesquisas futuras para melhorar o tema abordado pela monografia.

2 Fundamentação Teórica

2.1 Ciência de Dados

Ciência de Dados descreve uma área em ascensão de trabalho relacionada a coleta, processamento, limpeza, análise e visualização de dados, com a intenção de gerar um produto ou conhecimento útil para tomadas de decisão. Apesar do nome Ciência de dados parecer se conectar mais fortemente com áreas como Ciência da Computação e Sistemas de Informação, muitos tipos diferentes de habilidades, incluindo habilidades não matemáticas, são necessárias (STANTON, 2013). Por exemplo, a visão de negócio, é uma habilidade muito desejada dentro do campo da Ciência de Dados.

No entanto, apesar do grande crescimento nos últimos anos e até ser considerada uma das carreiras mais promissoras da atualidade (DAVENPORT; PATIL, 2012), a área já é estudada há pelo menos 30 anos. Porém, o que impedia que a área se popularizasse eram os preços altíssimos de processamento computacional e armazenamento de dados. Com o barateamento dessas tecnologias, a Ciência de Dados se tornou uma ferramenta poderosa para tomada de decisão, sendo decisiva para que empresas de menor porte possam brigar de igual para igual com grande conglomerados (RACONTEUR, 2015).

A Ciência de Dados tem como pretensão a tomada de decisão baseada em dados, para isso diversas competências são necessárias. Por isso, pode se dizer que a Ciência de Dados é uma "caixa de ferramentas", onde se encontram habilidades de modelagem de dados, para manipular os dados não estruturados e em larga quantidade; o aprendizado de máquina para utilizar o reconhecimento de padrões e realizar abstrações profundas; a matemática e estatística para desenvolver modelos e distribuições sólidas; e a análise de dados para conseguir identificar tendências, tanto de mercado como de comportamento para ajudar assim na tomada de decisão.

A grande maioria dos serviços que são consumidos em larga escala e utilizam de algum tipo de estudo de dados para melhorar seus serviços, como sistemas de recomendações, sistemas de busca na internet, serviços financeiros, propagandas online, entre outros muitos serviços.

2.2 Dados Abertos Governamentais

Começando pelo conceito de Dados Abertos, sua definição segundo a ONG *Open Knowledge Internacional*, é a de que os dados são abertos quando qualquer pessoa pode livremente acessá-los, utilizá-los, modificá-los e compartilhá-los para qualquer finalidade, estando sujeito a, no máximo, a exigências que visem preservar sua proveniência e sua abertura (Open

Knowledge Foundation, 2019).

Para os Dados Abertos Governamentais (DAG), foram desenvolvidos alguns princípios que os dados devem seguir para que possam ser considerados DAG. Em 2007, um grupo de estudiosos chamado *OpenGovData* formularam os 8 princípios que os dados devem seguir para ser considerados DAG (Open Gov Data, 2007), são eles:

1. **Completos**: Todos os dados públicos são disponibilizados. Dados são informações eletronicamente gravadas, incluindo, mas não se limitando a, documentos, bancos de dados, transcrições e gravações audiovisuais. Dados públicos são dados que não estão sujeitos a limitações válidas de privacidade, segurança ou controle de acesso, reguladas por estatutos.

2. **Primários**: Os dados são publicados na forma coletada na fonte, com a mais fina granularidade possível, e não de forma agregada ou transformada.

3. **Atuais**: Os dados são disponibilizados o quão rapidamente seja necessário para preservar o seu valor.

4. **Acessíveis**: Os dados são disponibilizados para o público mais amplo possível e para os propósitos mais variados possíveis.

5. **Processáveis por máquina**: Os dados são razoavelmente estruturados para possibilitar o seu processamento automatizado.

6. **Acesso não discriminatório**: Os dados estão disponíveis a todos, sem que seja necessária identificação ou registro.

7. **Formatos não proprietários**: Os dados estão disponíveis em um formato sobre o qual nenhum ente tenha controle exclusivo.

8. **Licenças livres**: Os dados não estão sujeitos a restrições por regulações de direitos autorais, marcas, patentes ou segredo industrial. Restrições razoáveis de privacidade, segurança e controle de acesso podem ser permitidas na forma regulada por estatutos.

Completando os 8 princípios, em 2009, foram publicas as três leis dos DAG, por (EAVES, 2009):

1. Se o dado não pode ser encontrado e indexado na *Web*, ele não existe.

2. Se não estiver aberto e disponível em formato compreensível por máquina, ele não pode ser reaproveitado.

3. Se algum dispositivo legal não permitir sua replicação, ele não é útil.

Dados abertos Governamentais tem cada vez mais possibilitado que pessoas e organizações consigam usar informações públicas ou governamentais para apoiar suas decisões, ou mesmo

a criação de soluções tecnológicas. Com essa leitura de oportunidade de negócio nos dados públicos quem muito se beneficia é a sociedade. Os benefícios da adoção dos DAG em questões como controle social e transparência governamental são, ao menos em tese, evidentes (CARLOS; MAIA; RICARDO, 2010).

2.3 Cota para Exercício da Atividade Parlamentar

Na Câmara dos Deputados, além de salário, benefícios próprios da CLT e adicionais (como auxílio-moradia), parlamentares brasileiros têm à sua disposição um orçamento entre R$ 30 mil e R$ 45 mil mensais para despesas vinculadas ao exercício da função parlamentar (RODRIGUES; FONTES, 2018). Esse orçamento é a Cota para Exercício da Atividade Parlamentar (CEAP, ou popularmente conhecido "Cotão"), nele estão inclusos pagamentos e ressarcimentos das despesas de exercício dos parlamentares. O Cotão é divido em algumas categorias, são elas: despesas com passagens aéreas, fretamento de aeronaves, alimentação do parlamentar, combustíveis e lubrificantes, consultorias, divulgação do mandato, aluguel e demais despesas de escritórios políticos, assinatura de publicações e serviços de TV e internet, contratação de serviços de segurança e cota postal e telefônica (RODRIGUES; FONTES, 2018).

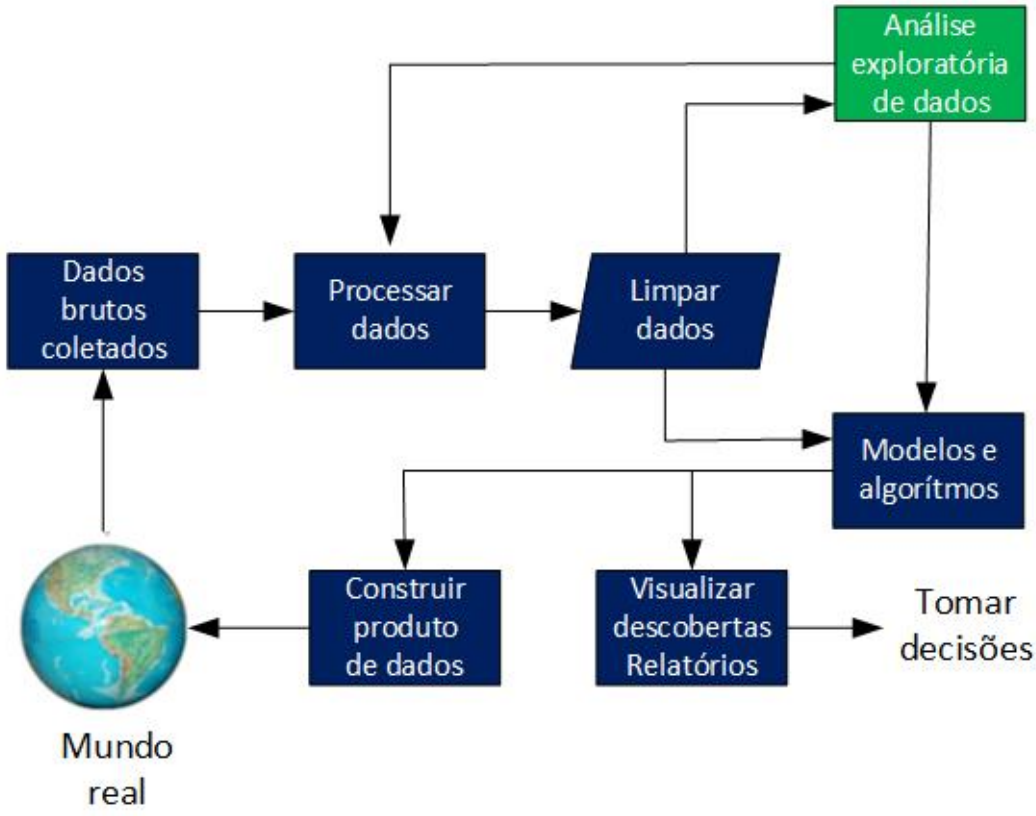
No ano de 2011, no Brasil foi aprovada a Lei no.12.527, que objetiva a divulgação dos gastos de órgãos públicos da União, para qualquer cidadão que deseje ter acesso a tais dados. A "Lei da transparência"como é conhecida, efetivamente disponibiliza as contas públicas em *websites*. O Cotão por exemplo, está disponível no endereço <https://www.camara.leg.br/transparencia/gastos-parlamentares> . Apesar da lei representar um grande marco quando se trata de transparência governamental no Brasil, grande parte dos dados contidos não seguem os padrões dos Dados Abertos Governamentais, tornando o ato de fiscalização de gastos parlamentares uma tarefa exaustiva.

3 Desenvolvimento

Para guiar o desenvolvimento da análise de dados, foi aplicado o processo descrito por (SCHUTT, 2014), representado na Figura 1. Para realizar as etapas do processo, utilizou-se a linguagem *Python*, suas bibliotecas derivadas e ambiente de desenvolvimento interativo *Jupyter Notebook*.

A primeira etapa do processo foi o de olhar o Mundo Real, e definir qual problema se quer resolver, para identificar exatamente o que se quer resolver. A partir disso, é possível decidir quais questões o estudo irá abordar.

Figura 1 – Processo de Ciência de Dados.



Fonte: (SCHUTT, 2014)

Neste cenário, buscasse abordar o problema de que muitos reembolsos de parlamentares são lançados diariamente e precisam ser analisados um a um manualmente, para identificar se é um reembolso válido. Dessa maneira, procurou-se automatizar o processo de lançamento e verificação dos reembolsos. Para que o escopo do projeto seja mais acurado utilizou-se os reembolsos categorizados como "Fornecimento de alimentação dos parlamentares"do ano de 2009 até 2019.

Com o problema definido, partiu-se para uma parte importante voltada à solução da problemática: quais dados serão utilizados para basear a solução. Além de identificar quais dados serão necessários: para encontrar um meio de realizar a aquisição dos dados. Neste trabalho, os dados foram coletados por meio de um *script Scrapy*, a extração foi feita na ferramenta da OSA, o Jarbas (Jarbas, 2019).

Os dados foram compilados no formato de planilha (.csv), pois esse é um formato adequado e viável para se processar os dados brutos usando *Python*. Preparou-se o ambiente de desenvolvimento, *Jupyter Notebook*, no nosso gerenciador de pacotes *Python*, o *Anaconda*. Usando essa arquitetura de desenvolvimento, é possível organizar um ambiente de desenvolvimento *Python* individualmente para cada projeto.

Com o ambiente devidamente configurado, realizou-se o carregamento dos dados, utilizando a biblioteca *Pandas*, exibindo como é o formato dos dados do arquivo, que segundo o método *shape*, conclui-se que os dados têm 229892 linhas e 6 colunas. Exibe-se assim, os 5 primeiros dados inseridos na planilha, como na Figura 2.

Explicando o cabeçalho da planilha, temos:

- **Congress_person_name** : Nome do parlamentar que está registrado no reembolso
- **Value_meal** : O valor do reembolso
- **Year** : O ano que foi realizado o reembolso
- **Subquota_translated** : A categoria do reembolso
- **Supplier_info** : Informações sobre o fornecedor do serviço

Em seguida do carregamento das informações fica a cargo do cientista de dados, verificar e avaliar que os dados estão armazenados de forma precisa, tomando o cuidado para que não contenha dados que possam corromper a análise (ALBINO, 2016). Logo, antes de começar a análise é importante chegar se existem linhas duplicadas, ou valores definidos como zero, além de preparar os dados para que a máquina possa entender os dados.

Tabela 1 – Detalhes estatísticos do modelo.

Métricas	**valores**
Média	65.897423
Desvio padrão	110.105499
Mínimo	0
Máximo	7097.000000

Com os dados limpos já é possível começar a análise exploratória dos dados (AED), onde se começa a produção do modelo. Para iniciar a AED, utilizando o comando *describe*

Figura 2 – Carregamento e visualização dos dados.

```
In [4]: training = pd.read_csv('../data/congressPersons.csv')
```

Verificando qual o formato que os dados inseridos

```
In [5]: training.shape
Out[5]: (229892, 6)
```

```
In [6]: training.drop(['page'],1).head(5)
```

Out[6]:

	congress_person_name	value_meal	year	subquota_translated	supplier_info
0	CAPITÃO AUGUSTO	R$ 55,00	2019	Fornecimento de alimentação do parlamentar	WRC RESTAURANTES LTDA
1	PAULÃO	R$ 69,33	2019	Fornecimento de alimentação do parlamentar	26.058.791/0001-02
2	PASTOR EURICO	R$ 66,49	2019	Fornecimento de alimentação do parlamentar	RESTAURANTE ROMA LTDA
3	RONALDO LESSA	R$ 62,20	2019	Fornecimento de alimentação do parlamentar	01.222.256/0001-06
4	JORGE SOLLA	R$ 65,78	2019	Fornecimento de alimentação do parlamentar	SERVICO NACIONAL DE APRENDIZAGEM COMERCIAL SENAC

Fonte: Elaborado pelo autor

da biblioteca *pandas*, retira-se algumas métricas estatísticas dos dados para tentar visualizar como estão dispostos. Esta análise está representada na Tabela 1. Na Tabela 2, constata-se como estão distribuídos os reembolso em faixas de valores, melhor visualizado no histograma da Figura 3.

Tabela 2 – Distribuição dos valores de refeição por faixa de valores.

Faixa	valores
(0,100]	189837
(100, 150]	24846
(150, 200]	9892
(200,300]	3619
(300,500]	694
(500,1000]	547
(1000, 8000]	454

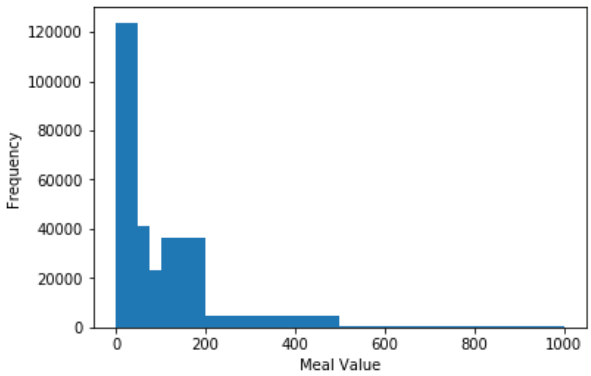
Na Figura 4, pode-se observar um gráfico com os parlamentares que possuem mais reembolsos referentes a alimentação nos últimos 10 anos.

Finalmente, agora que já há o conhecimento de como os dados estão dispostos, utilizou-se um método para detectar valores que sejam atípicos ou suspeitos. Para detecta-los, buscou-se valores isolados (ou em inglês *outliers*). E o método utilizado foi o *zscore*, onde calculasse o *zscore* utilizando a formula :

$$z = \frac{x - \bar{x}}{S}, \tag{3.1}$$

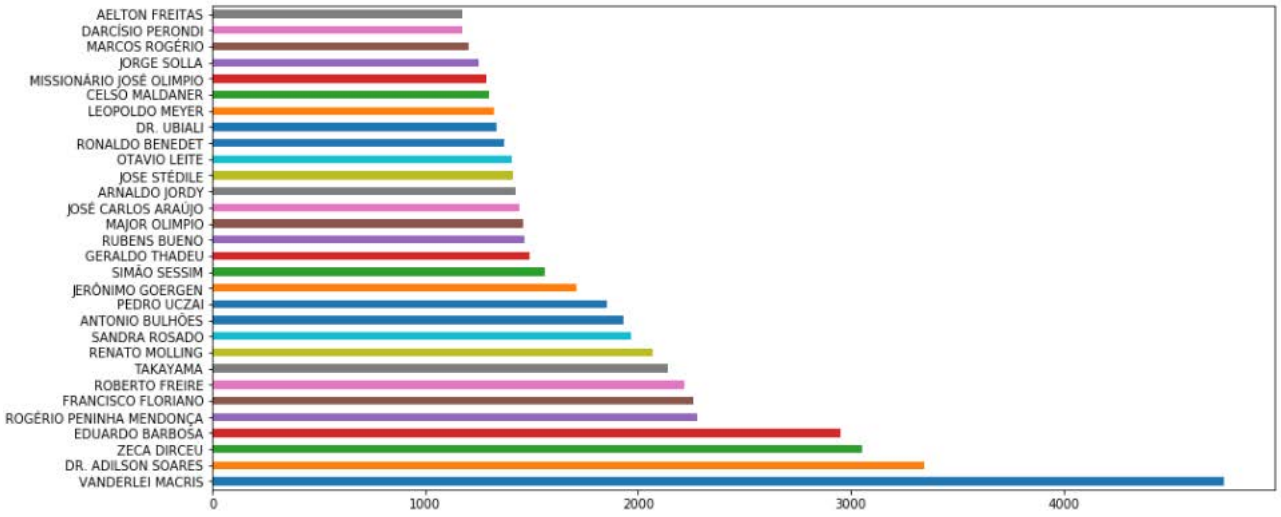
Onde: $\bar{x}$ é a média da amostra e S é o desvio padrão da amostra.

Figura 3 – Histograma dos reembolsos de refeições até 1000 reais.



Fonte: Elaborado pelo autor

Figura 4 – Gráfico de barras mostrando mais os 10 parlamentares com mais reembolsos referente a refeições no últimos 10 anos.



Fonte: Elaborado pelo autor

Então é calculado o *zscore* com base na entrada de dado. Aqueles que tiverem o valor maior que o limite estabelecido serão classificados como valores isolados. Com isso construiu-se a Tabela 3 com algumas variações de valores para o limite e a porcentagem de dados classificados como isolados.

Tabela 3 – Tabela da porcentagem de valores isolados por limites.

Limite	Porcentagem de valores isolados	Número de valores isolados
4	**0.43%**	**996**
3	**0.53%**	**1234**
2	**0.83%**	**1928**
1	**4.12%**	**9484**
0.5	**17.89%**	**41139**

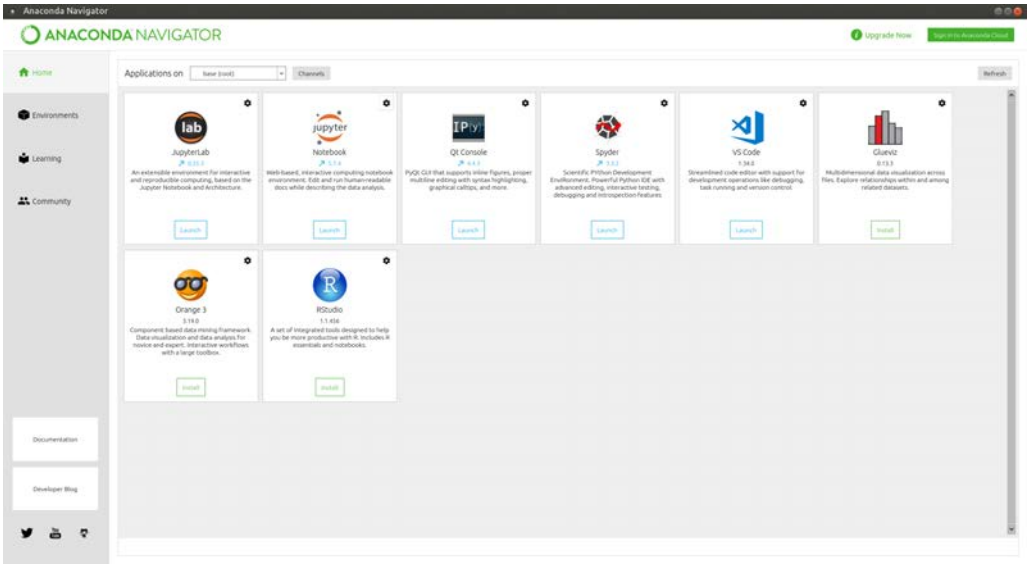
3.1 Ferramentas

3.1.1 Gerenciador de pacotes *Anaconda*

Anaconda é uma ferramenta de gerenciamento de pacotes e *deployment* para as linguagens *Python* e *R*, a sua distribuição é gratuidade e *open-source*. Focado para a computação científica, a *Anaconda* é usada em projetos de ciência de dados, aprendizado de máquina, análise preditiva e projetos científicos em geral.

Para seu uso a *Anaconda* possui uma aplicação de Interface gráfica de usuário (em inglês *Graphical User Interface*), chamada *Anaconda Navigator*, onde é possível encontrar aplicações como : *Jupyter Notebook*, *Rstudio*, *Visual Studio Code*, entre outras. Existe ainda a possibilidade de usar a *Anaconda* em nuvem e armazenar seus projetos em ambientes na nuvem.

Figura 5 – Página inicial do *Anaconda Navigator*.



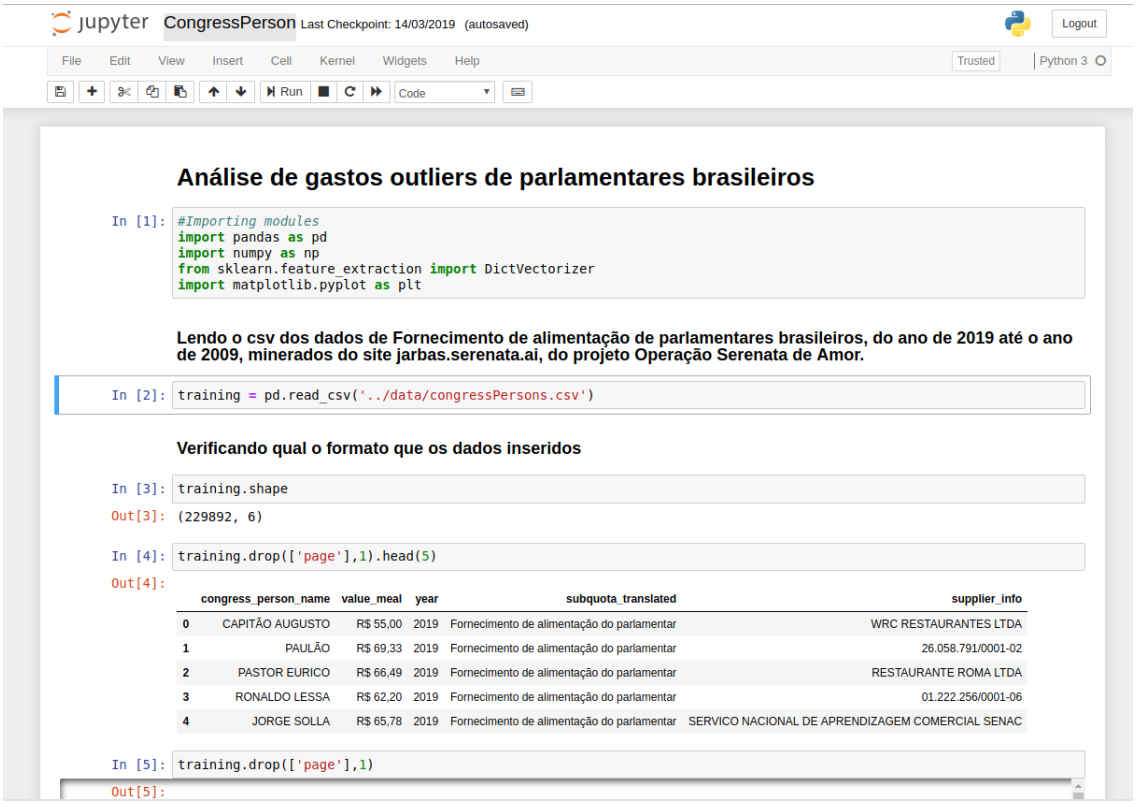
Fonte: Elaborado pelo autor

3.1.2 Ambiente de desenvolvimento *Jupyter Notebook*

O projeto *Jupyter* tem como ambição "desenvolver software de código aberto, padrões abertos e serviços para computação interativa em dezenas de linguagens de programação"(Projeto Jupyter, 2019). O *Jupyter notebook* é uma aplicação *web* que permite criar e compartilhar documentos que contém códigos, texto, gráficos, equações de modo interativos. Por permitir que você execute, ou re-execute blocos de código, esse mecanismo faz com que seu uso seja excelente para quem está realizando estudos como modelagens estatísticas, modelos de aprendizagem de máquina, simulações numérica, no geral a computação científica.

O *Jupyter notebook* no seu inicio tinha a ambição de suportar três linguagens: *Python*, *Julia* e *R*. Por isso o nome *Jupyter*, no entanto, nos dias atuais mais de 40 linguagens são suportadas, além de possuir integração com as mais famosas ferramentas usadas para manipular Grandes dados, como *Apache Spark*, *TensorFlow*, *Scala*, entre outras.

Figura 6 – Interface do *Jupyter Notebook*.



Fonte: Elaborado pelo autor

3.1.3 *Scrapy*

Scrapy é um *framework open source* para realizar a extração de dados em páginas *web*, de maneira rápida e simples. Para realizar a extração de dados com o *Scrapy* se utiliza de "aranhas"que "rastejam"pela página *web* desejada, seguindo um conjunto de instruções.

3.1.4 *Pandas*

Pandas é uma biblioteca *open source* usada para a análise de dados, além de fácil de usar o *Pandas* tem uma alta performance para análise de grandes documentos. Com essa biblioteca é possível manipular os dados com indexação integrada, ler/escrever dados entre dados que estão na memória e diferentes formatos de arquivos, fazer a remodelagem de conjuntos de dados, entre outras muitas funções para manipulação de dados.

3.1.5 *Numpy*

Numpy é um pacote *open source* e considerado fundamental para a computação científica, pois contém uma ampla coleção de funções matemáticas em matrizes e vetores. Por trabalhar com diversos arranjos de dados, o *Numpy* é facilmente integrado com uma grande variedade de base de dados, o que permite que as integrações sejam ocorram perfeitamente e prontamente.

3.1.6 *Matplotlib*

Matplotlib é uma biblioteca *open source* de plotagem muito versátil, sendo possível plotar desde as mais simples figuras, até figuras complexas e interativas. Assim agradando usuário que querem fazer coisas fáceis como plotar histogramas, gráficos de barras, gráficos de dispersão e também agradando usuários mais experientes que gostam de ter controle de todos os estilos das linhas, as propriedades dos eixos ou as propriedades das fontes.

4 Conclusão

O presente trabalho apresenta ferramentas e processos que tem como objetivo mostrar como tecnologias atuais podem ajudar a gerar conhecimento útil no referente a uso de Dados Abertos Governamentais (DAG) disponibilizado por diversas esferas dos organismos do governo.

Neste projeto a expectativa era incitar o começo da discussão de como as tecnologias atuais podem facilitar processos antigos, como por exemplo, o de verificar a veracidade de recibos de reembolsos. Apesar da abordagem simples sobre o processo de ciência de dados, fica evidente como existe um grande espaço para pesquisa e desenvolvimento da área dos processos governamentais, ainda não desbravados na área da computação.

Além disso, o projeto se utilizou apenas de ferramentas de código aberto mostrando que as ferramentas para o desenvolvimento de projetos de ciência de dados estão disponíveis pela e para a comunidade.

O projeto todo está disponível no *Github*, sob licença *MIT*, com a intenção de incentivar outros pesquisadores a melhorar o modelo, ou até mesmo criar o seu próprio modelo, fomentando assim a comunidade de colaboradores no processo de ciência de dados.

Por se tratar de um trabalho inicial, o número de estudos futuros acaba por ser muito vasto, visto que o modelo do trabalho tem muito espaço para crescimento e desenvolvimento.

Logo, conclui-se que é possível auxiliar processos governamentais através do uso da DAG para apoiar tomadas de decisões governamentais, além de reforçar o potencial dos dados abertos sobre as políticas públicas de governo.

Referências

ALBINO, J. Uma abordagem para criação de valor em dados abertos para pequenas e médias empresas utilizando o ecossistema r. In: . [S.l.: s.n.], 2016.

BANDEIRA, J.; ALCANTARA, W.; SOBRINHO, A.; ÁVILA, T.; BITTENCOURT, I.; ISOTANI, S. Dados abertos conectados. In: _____. [S.l.: s.n.], 2014. ISBN 244708210000.

BEGHIN, N.; ZIGONI, C. Avaliando os websites de transparência orçamentária nacionais e sub-nacionais e medindo impactos de dados abertos sobre direitos humanos no brasil. Brasilia, Brazil, 2014.

BERNERS-LEE; TIM; HENDLER; JAMES; LASSILA, O. The semantic web: A new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *ScientificAmerican.com*, 05 2001.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked data - the story so far. *International Journal on Semantic Web and Information Systems*, IGI Global, v. 5, n. 3, p. 1–22, jul. 2009. Acessado em: 25/05/2019. Disponível em: <https://doi.org/10.4018/jswis.2009081901>.

CARLOS, V. J.; MAIA, R. M.; RICARDO, M. Dados governamentais abertos e seus impactos sobre os conceitos e práticas de transparência no brasil. *Cadernos ppg-au/ufba*, v. 9, n. 1, 2010.

CETIC. *Pesquisa TIC Domicílios*. 2017. Acessado em: 05-05-2019. Disponível em: <https://cetic.br/pesquisa/domicilios/analises>.

DAVENPORT, T. H.; PATIL, D. Data scientist. *Harvard business review*, v. 90, n. 5, p. 70–76, 2012.

EAVES, D. The three laws of open government data. *Eaves. ca*, v. 30, p. 8, 2009.

ECONOMIST, T. *MS Windows NT Data, data everywhere*. 2010. Acessado em: 25/05/2019. Disponível em: <http://www.economist.com/node/15557443>.

ISOTANI, S.; BITTENCOURT, I. *Dados Abertos Conectados: em Busca da Web do Conhecimento*. [S.l.: s.n.], 2015. ISBN 978-85-7522-449-6.

Jarbas. *Documentação*. 2019. Acessado em: 25/05/2019. Disponível em: <https://jarbas.serenata.ai/>.

Open Gov Data. *Open Government Data Definition: The 8 Principles of Open Government Data*. 2007. Acessado em: 25/05/2019. Disponível em: <https://opengovdata.io/2014/8-principles/>.

Open Knowledge Foundation. *Documentação*. 2019. Acessado em: 25/05/2019. Disponível em: <https://okfn.org/>.

Operação Serenata de amor. *Documentação*. 2019. Acessado em: 25/05/2019. Disponível em: <https://serenata.ai/>.

Projeto Jupyter. *Documentação*. 2019. Acessado em: 25/05/2019. Disponível em: <https://jupyter.org/>.

RACONTEUR. *Smaller firms think big data*. 2015. Disponível em: <<https://www.raconteur.net/finance/smaller-firms-think-big-data>>. Acessado em: 25/05/2019.

RIBEIRO C. Y VIEIRA PEREIRA, D. S. A publicação de dados governamentais abertos: proposta de revisão da classe sobre previdência social do vocabulário controlado do governo eletrônico. *Transinformação, [en linea] 27(1)*, p. 73–82, 2015. Acessado em: 25/05/2019. Disponível em: <https://www.redalyc.org/articulo.oa?id=384351519008>.

RODRIGUES, J.; FONTES, C. Estudo de caso "operação serenata de amor": a análise de big data no combate à festa dos gastos públicos. In: . [S.l.: s.n.], 2018.

SCHUTT, R. Doing data science: Straight talk from the frontline. 2014.

SHIRKY, C. *Here comes everybody: The power of organizing without organizations*. [S.l.]: Penguin, 2008.

SILVA, T. E. d.; EIRÃO, T. G.; CAVALCANTE, R. Lei de acesso à informação na câmara dos deputados: um ano de funcionamento do serviço de informação ao cidadão. 2014.

STANTON, J. M. Introduction to data science. 2013.

VAZ, J. C.; RIBEIRO, M. M.; MATHEUS, R. Desafios para a governança eletrônica e dados governamentais abertos em governos locais. In: *WTRANS13-Workshop de Transparência em Sistemas*. [S.l.: s.n.], 2013.

WOOD, D.; ZAIDMAN, M.; RUTH, L.; HAUSENBLAS, M. *Linked Data: Structured Data on the Web*. [S.l.]: Manning Publications, 2014. ISBN 9781617290398.