Roberto Williams Batista
Udacity
Machine Learning Nanodegree
22 de maio de 2018
Versão: 1.7

Projeto Final

1. Definição

1.1 Visão geral do projeto

O projeto em questão trata de uma análise descritiva de dados anualizados sobre gastos de verbas parlamentares para exercício da função que compreende os anos de 2015, 2016 e 2017. Após a análise estatística descritiva será realizada a predição de gastos para o ano de 2018 de três deputados que mais gastaram no ano de 2017.

1.2 Histórico

Tendo em vista a repercussão de mal uso de cotas parlamentares[1] no momento político do Brasil surgiu a demanda de maior transparência quanto ao gastos destas e outras verbas públicas. Um exemplo escandaloso do mal uso destas verbas é compra de reportagens por deputados[2] identificada pelo Projeto Marco Zero[3].

Com intuito de produzir uma análise de dados e predição para colaborar com este movimento, decidiu-se pela analise do CEAP[4], e posteriormente publicá-lo nas mídias sociais para conscientização da população em geral, assim como incentivar outras iniciativas de monitoramento de gastos públicos.

1.2 Descrição do problema

A câmara dos deputados federais fornece dados históricos de gastos do CEAP organizados anualmente. Infelizmente não disponibilizam também um **estudo de análise estatística** destes gastos, assim como a **previsão de comportamento de gastos** do CEAP para o ano vigente baseado em dados históricos. Para resolver este problema estaremos utilizando dados públicos de gastos parlamentares e realizando:

- Análise Estatística Descritiva dos gastos no ano de 2015, 2016 e 2017.
  - Será utilizado gráficos estatísticos.
  - Dados estatísticos.
- Predição de gastos para o ano de 2018 utilizando-se dos dados do triênio dos top 3 deputados de 2017.
  - Serão testados alguns algoritmos de machine *learning* supervisionados, de regressão linear, disponíveis na biblioteca Scikit Learn.
  - A performance identificada será crucial para definir o mais adequado para predição.

Conjuntos de dados e entradas

Serão utilizados os *datasets* dos gastos consolidados para os anos de 2015, 2016 e 2017, disponíveis no site abaixo:

https://dadosabertos.camara.leg.br/swagger/api.html

Os dados estão organizados de forma estruturada em três arquivos de formato CSV com 28 *features* (verificar tabela 1) :

1. Ano-2015.csv (376.862 linhas, 29 colunas e 77.5MB)
2. Ano-2016.csv (358.238 linhas, 29 colunas e 74.1MB)
3. Ano-2017.csv (341.222 linhas, 29 colunas e 71.2MB)

Foram utilizadas para análise dos dados as seguintes ferramentas:

1. Jupyter Notebook.
2. Numbers.
3. Linguagem e bibliotecas Python.

---

[1] https://medium.com/data-science-brigade/precisamos-falar-sobre-a-cota-parlamentar-c58a73392148

[2] https://theintercept.com/2018/01/16/deputados-usam-cota-parlamentar-para-comprar-reportagens/

[3] http://www.marcozero.info/

[4] Cota para exercício parlamentar.

Descrição da solução

A solução é composta por três fases:

1. **Limpeza de dados.**

   A Limpeza de Dados realiza uma análise inicial dos *datasets* a serem utilizados. Esta análise auxilia o entendimento de possíveis problemas no processo de importação dos *datasets*, assim como as ações necessária para resolve-las. Também foram identificados problemas relacionados com formatação.

   Nesta fase o objetivo é modelar o *dataset* original criando um *dataset* adequado para as fases de análise exploratória e predição.

2. **Exploração de dados.**

   A exploração de dados busca encontrar na análise descritiva informações a respeito dos gastos em questão, gerando gráficos e resultados de cálculos relevantes para o entendimento.

3. **Predição.**

   A fase da Predição tem como objetivo predizer os gastos para o ano de 2018 de três deputados de maior gasto em 2017.

Métricas de avaliação

A solução poderá ser avaliada através do cumprimento da seguintes métricas:

a. Criação de gráficos de estatística descritiva e a análise de sua representação e potenciais descobertas.

b. Aplicação de algoritmo de aprendizagem supervisionada de regressão para predição de gastos para 2018 dos top 3 deputados de maior gastos em 2017 e análise de seus resultados. Será utilizado R2 *Score* (*coefficient of determination*) para determinação do score do algoritmo frente ao problema.

   I. Apesar dos três *datasets* originais terem juntos mais de um milhão de linhas e 29 *features*, os dados úteis para a nossa análise terão em média 30 linhas úteis e doze *features*, dependendo dos tipos de gastos. Estes dados representam os valores totalizados de cada tipo de gasto, para cada um dos três anos.

2. Análise

2.1 Exploração de dados

A exploração de dados é uma parte muito importante do projeto, assim como trabalhosa. Foi necessário adequar os datastes originais para que fosse possível a análise estatística descritiva e a predição de gastos para 2018 dos top 3 deputados de 2017.

Bibliotecas

Foram utilizadas as seguintes bibliotecas:

- Pandas
- Numpy
- OS

Passos

Foram realizadas os seguintes passos para realizar a limpeza de dados.

1. **Aquisição de *dataset*.**
    a. O *dataset* é disponibilizado pelo web site https://dadosabertos.camara.leg.br/swagger/api.html, e fazem parte do programa de transparência das contas publicas do governo Federal.
    b. No endereço acima foi realizado download dos três *datasets* de gastos CEAP dos anos de 2015, 2016 e 2017.
    c. Como a interface API estava em testes até o momento do início do projeto, decidiu-se realizar a obtenção dos dados manualmente em favor da replicabilidade do estudo por outras pessoas.
    d. Por motivo de tamanho os *datasets* originais não podem ser colocados no Github, na qual a conta gratuita tem um limite de 25 MB de tamanho para arquivos.
    e. Para importação utilizou-se a função da biblioteca Pandas, read_csv.
2. **Exploração e ajustes iniciais dos três datasets.**
    a. Inicialmente utilizou-se o *Numbers* para abrir o arquivo 'Ano-2015.csv' e fazer uma avaliação inicial da estrutura de dados, assim como identificar potenciais problemas na importação.

Imagem 1 - Parte 1 das primeiras *features* do *dataset.*

txNomeParlamentar	idecadastro	nuCarteiraParlamentar	nuLegislatura	sgUF	sgPartido	codLegislatura	numSubCota	txtDescricao	numEspecificacaoSubCota
ABEL MESQUITA JR.	178957	1	2015	RR	DEM	55	1	MANUTENÇÃO DE ESCRITÓRIO DE APOIO À ATIVIDADE PARLAMENTAR	0
ABEL MESQUITA JR.	178957	1	2015	RR	DEM	55	1	MANUTENÇÃO DE ESCRITÓRIO DE APOIO À ATIVIDADE PARLAMENTAR	0
ABEL MESQUITA JR.	178957	1	2015	RR	DEM	55	1	MANUTENÇÃO DE ESCRITÓRIO DE APOIO À ATIVIDADE PARLAMENTAR	0
ABEL MESQUITA JR.	178957	1	2015	RR	DEM	55	1	MANUTENÇÃO DE ESCRITÓRIO DE APOIO À ATIVIDADE PARLAMENTAR	0
ABEL MESQUITA JR.	178957	1	2015	RR	DEM	55	1	MANUTENÇÃO DE ESCRITÓRIO DE APOIO À ATIVIDADE PARLAMENTAR	0
ABEL MESQUITA JR.	178957	1	2015	RR	DEM	55	3	COMBUSTÍVEIS E LUBRIFICANTES.	1
ABEL MESQUITA JR.	178957	1	2015	RR	DEM	55	3	COMBUSTÍVEIS E LUBRIFICANTES.	1
ABEL MESQUITA JR.	178957	1	2015	RR	DEM	55	3	COMBUSTÍVEIS E LUBRIFICANTES.	1
ABEL MESQUITA JR.	178957	1	2015	RR	DEM	55	3	COMBUSTÍVEIS E LUBRIFICANTES.	1
ABEL MESQUITA JR.	178957	1	2015	RR	DEM	55	3	COMBUSTÍVEIS E LUBRIFICANTES.	1

Imagem 2 - Parte 2 das primeiras *features* do *dataset.*

txtDescricaoEspecificacao	txtFornecedor	txtCNPJCPF	txtNumero	indTipoDocumento	datEmissao	vlrDocumento	vlrGlosa	vlrLiquido	numMes	numAno
	COMPANHIA DE AGUAS E ESGOTOS DE RORAIMA	5939467000115	0010100910378000	0	2015-11-14 00:00:00	165,65	0	165,65	11	2015
	COMPANHIA DE AGUAS E ESGOTOS DE RORAIMA	5939467000115	0010100910378000	0	2015-12-10 00:00:00	59,48	0	59,48	12	2015
	ELETROBRAS DISTRIBUIÇÃO RORAIMA	2341470000144	103091	0	2015-11-27 00:00:00	130,95	0	130,95	11	2015
	ELETROBRAS DISTRIBUIÇÃO RORAIMA	2341470000144	103400	0	2015-12-30 00:00:00	196,53	3,47	193,06	12	2015
	PAPELARIA ABC Com. e Ind. LTDA.	540252000103	42320	0	2015-02-23 00:00:00	310,25	0	310,25	2	2015
Veículos Automotores	Auto Posto Aeroporto Ltda	8202116000115	477642	0	2015-06-30 00:00:00	32	0	32	6	2015
Veículos Automotores	Auto Posto Aeroporto Ltda	8202116000115	506884	0	2015-08-07 00:00:00	50	0	50	8	2015
Veículos Automotores	AUTO POSTO AEROPORTO LTDA.	8202116000115	528751	0	2015-09-08 00:00:00	50	0	50	9	2015
Veículos Automotores	AUTO POSTO AEROPORTO LTDA.	8202116000115	544551	0	2015-09-24 00:00:00	75	0	75	9	2015
Veículos Automotores	AUTO POSTO CHAVES LTDA	746278000102	815253	0	2015-05-06 00:00:00	170,02	0	170,02	4	2015

Imagem 3 - Parte 3 das primeiras *features* do *dataset.*

txtPassageiro	txtTrecho	numLote	numRessarcimento	vlrRestituicao	nuDeputadoId	ideDocumento
		1255355	5294	0	3074	5886345
		1255361	5294	0	3074	5886361
		1255355	5294	0	3074	5886341
		1268867	5370	0	3074	5928783
		1168538	4966	0	3074	5608486
		1220276	5131	0	3074	5772352
		1220275	5131	0	3074	5772246
		1229908	5173	0	3074	5804438
		1229908	5173	0	3074	5804312
		1192297	5036	0	3074	5682898

3. Importação dos datasets

a. Foi-se realizada a importação dos datasets e verificou-se o primeiros desafios a serem solucionados.

I. ***txtCNPJCPF***: Durante a importação dos elementos relacionados ao *feature txtCNPJCPF* foi tratado pela função como integer, passando assim por uma transformação à notação científica.

II. **Delimitador de separação de dados**: Como os dados dos *datasets* contém notações financeiras brasileiras a presença da vírgula como marcador decimal está presente, obrigando o documento utilizar como separador o ponto e vírgula (';') ao invés da vírgula (','). Caso isto não fosse assim estruturada não conseguiríamos importar corretamente os dados, já que toda vírgula do valor monetário indicaria o fim de um dado e o início de um novo.

III. **Demais valores financeiros**: Os itens relacionados a valores financeiros foram importados como objeto, preservando então o conteúdo, por este motivo serão avaliados sua transformação do tipo de dados após a definição das colunas a serem selecionadas para a análise.

4. Análise inicial dos dados importados

a. Foram realizadas as seguintes análises:

I. Estatística descritiva através da função describe.

	idecadastro	nuCarteiraParlamentar	nuLegislatura	codLegislatura	numSubCota	numEspecificacaoSubCota	indTipoDocumento	numMes	numAno
count	376,142.00	376,142.00	376,863.00	376,142.00	376,863.00	376,863.00	376,863.00	376,863.00	376,863.00
mean	138,323.59	299.90	2,011.08	54.98	327.84	0.20	0.19	6.78	2,015.00
std	42,826.91	154.18	88.05	0.13	453.69	0.41	0.40	3.26	0.00
min	3,151.00	1.00	0.00	54.00	1.00	0.00	0.00	1.00	2,015.00
25%	74,752.00	178.00	2,015.00	55.00	3.00	0.00	0.00	4.00	2,015.00
50%	160,536.00	306.00	2,015.00	55.00	13.00	0.00	0.00	7.00	2,015.00
75%	178,860.00	443.00	2,015.00	55.00	999.00	0.00	0.00	10.00	2,015.00
max	193,069.00	674.00	2,015.00	55.00	999.00	4.00	3.00	12.00	2,015.00

Nesta análise preliminar foram identificadas algumas características interessantes:

i. Há *features* com quantidades totais maiores e menores, indicando a presença de NaN no *dataset*.

ii. A presença de NaN pode ser preocupante caso ocorra em *features* escolhidas para análise, principalmente se relacionadas ao valor dos gastos. Neste último caso pode-se requerer a exclusão da linha ou, se for importante, a regressão destes valores.

iii. Na coluna numMes e NumAno, sabia de antemão que será uma *feature* utilizada, percebe-se através dos valores de min, max e percentis (25%, 50% e 75%), que não devem ter valores errados presentes nestes features, já que fornecem valores esperados.

iv. Foi identificado através dos valores de max e min, que há uma grande amplitude de valores nas features. Isto indica, caso a feature com alta amplitude seja usada, que há potencial utilização de pré-processamento, tal como normalização ou "logaritimização".

v. Existem diferentes distribuições de dados identificados. Comentários mais específicos serão feitos após a seleção dos *features*.

II. Verificação de tipo de dados para cada *feature*.

```
1 df2015.dtypes
```

```
txNomeParlamentar              object
idecadastro                   float64
nuCarteiraParlamentar         float64
nuLegislatura                   int64
sgUF                           object
sgPartido                      object
codLegislatura                float64
numSubCota                      int64
txtDescricao                   object
numEspecificacaoSubCota         int64
txtDescricaoEspecificacao      object
txtFornecedor                  object
txtCNPJCPF                     object
txtNumero                      object
indTipoDocumento                int64
datEmissao                     object
vlrDocumento                   object
vlrGlosa                       object
vlrLiquido                     object
numMes                          int64
numAno                          int64
numParcela                      int64
txtPassageiro                  object
txtTrecho                      object
numLote                         int64
numRessarcimento              float64
vlrRestituicao                float64
nuDeputadoId                    int64
ideDocumento                    int64
dtype: object
```

i. Podemos identificar alguns pontos interessantes a respeito de *features* que de antemão já é sabido a sua importância e utilização.

ii. TxNomeParlamentar, conforme Anexo 1 deste documento, diz respeito ao nome do parlamentar foi importado com *data type object*, assim como *sgUF* e *sgPartido*. O que é uma boa indicação de que não estes potenciais *features* não precisarão de alteração de *data type*.

iii. O *feature* txtCNPJCPF, conforme foi importado corretamente, sendo considerado object, não sofrendo qualquer alteração do seu conteúdo.

iv. O *feature* vlrDocumento aparece como object, que irá requerer posterior alteração em seu tipo de dados.

III. Verificação das dimensões do dado tabular através da função *shape*.

```
1 print ('Ano de 2015:', df2015.shape, 'Ano de 2016:', df2016.shape , 'Ano de 2017:', df2017.shape)
```

```
Ano de 2015: (376863, 29) Ano de 2016: (358239, 29) Ano de 2017: (341223, 29)
```

v. As dimensões do dados são de em média 350 mil linhas e vinte e nove colunas (*features*).

vi. O valor de colunas e seus nomes são consistentes. Não há *features* faltando ou a mais.

IV. Visualização das primeiras linhas através da função *head*.

Imagem - Parte 1/3 do resultado da função *head.*

```
'Ano de 2015:'
```

	txNomeParlamentar	idecadastro	nuCarteiraParlamentar	nuLegislatura	sgUF	sgPartido	codLegislatura	numSubCota	txtDescricao
0	ABEL MESQUITA JR.	178,957.00	1.00	2015	RR	DEM	55.00	1	MANUTENÇÃO DE ESCRITÓRIO DE APOIO À ATIVIDADE ...

1 rows × 29 columns

Imagem - Parte 2/3 do resultado da função *head.*

'Ano de 2015:'

numEspecificacaoSubCota	...	numMes	numAno	numParcela	txtPassageiro	txtTrecho	numLote	numRessarcimento	vlrRestituicao	nuDeputadoId
0	...	11	2015	0	NaN	NaN	1255355	5,294.00	0.00	3074

Imagem - Parte 3/3 do resultado da função *head.*

ideDocumento
5886345

i. A primeira linha do arquivo de 2015 nos mostra já a indicação de dados faltantes no *dataset*, assim como a estrutura destes dados.

b. Estas analises de repetiram para cada *dataset* importado.

5. Consolidação dos dados tubulares.

a. Após a análise preliminar realizou-se a concatenação dos três *datasets*.

b. Definiu-se as colunas a serem utilizadas na análise.

I. Permanceram para análise:

Tabela 1 - Lista dos *features* a serem utilizados na análise.

#	Número da Coluna	Elemento de Dado
1	1	txNomeParlamentar
2	5	sgUF
3	6	sgPartido
4	9	txtDescricao
5	17	vlrDocumento
6	20	numMes
7	21	numAno

II. Dados completos são fornecidos pela tabela no Anexo 1 deste documento lista todas as *features* do *dataset*, descreve seu significado e indica através da cor azul os *features* escolhidos para análise.

III. É importante ressaltar que não foi necessário a utilização de todas as colunas definidas durante o processo de planejamento do projeto. O manuseio com os dados permitiu uma decisão precisa neste tema.

6. Alteração de notação financeira brasileira para americana

a. Na seleção de colunas ficou apenas a *feature vlrDocumento* com valores financeiros a ser corrigida e transformada.

b. Foi necessário utilizar do método *str.replace* para trocar a vírgula como indicador de centavos pelo ponto.

7. Verificação de valores

a. A coluna *numAno* detém valores relacionados ao ano na qual foi realizada a despesa. Por este motivos verificou-se se os valores indicados nela diziam respeito apenas aos anos 2015, 2016 e 2017, assim como algum erro de digitação.

8. Alterando os nome originais das colunas

a. Realizou-se a escolha de novos nomes para as *features*, trazendo significado mais objetivo e conciso para que no momento de criação de gráficos fique mais inteligível e compreensível.

b. A tabela 2, abaixo, descreve o nome original e o novo nome:

Tabela 2 - Tabela de relação de nome de *features* originais e seus respectivos nomes novos.

txNomeParlamentar => Nome

sgUF .....=> UF

sgPartido .....=> Partido

txtDescricao .....=> Tipo

vlrDocumento .....=> Valor

numAno .....=> Ano

numMes .....=> Mes

9. Padronização de descrição.

a. Alterou-se a descrição das entradas da coluna Tipo para uniformização e simplificação.

b. A alteração segue o modelo da tabela 3.

Tabela 3 - Tabela de nomes originais e novos nomes.

#	NOME ORIGINAL	NOVO NOME
1	MANUTENÇÃO DE ESCRITÓRIO DE APOIO À ATIVIDADE PARLAMENTAR	Escritório
2	COMBUSTÍVEIS E LUBRIFICANTES.	C&L
3	CONSULTORIAS PESQUISAS E TRABALHOS TÉCNICOS.	Consultorias
4	DIVULGAÇÃO DA ATIVIDADE PARLAMENTAR.	Divulgação
5	SERVIÇO DE SEGURANÇA PRESTADO POR EMPRESA ESPECIALIZADA.	Segurança
6	PASSAGENS AÉREAS	Passagens Aéreas
7	TELEFONIA	Telefonia
8	SERVIÇOS POSTAIS	Correios
9	FORNECIMENTO DE ALIMENTAÇÃO DO PARLAMENTAR	Alimentação
10	LOCAÇÃO OU FRETAMENTO DE VEÍCULOS AUTOMOTORES	Locação de Carro
11	LOCAÇÃO OU FRETAMENTO DE EMBARCAÇÕES PEDÁGIO E ESTACIONAMENTO	Locação de Embarcação
12	SERVIÇO DE TÁXI, PEDÁGIO E ESTACIONAMENTO	Taxi
13	Emissão Bilhete Aéreo	Passagens Aéreas
14	HOSPEDAGEM EXCETO DO PARLAMENTAR NO DISTRITO FEDERAL.	Hospedagem

10. Remoção de valores negativos

a. Verificou-se que a *feature* Valor possui alguns itens negativos. Verificou-se que se tratava de certa compensação contábil de gastos, tendo em vista que não foi encontrado descritivo sobre este tipo de lançamento contábil todos estes valores foram removidos.

11. Identificação de colunas com NAN

a. Foram identificadas através da função *isnull* a falta de certos valores nas *features UF* e *Partido*, foram investigadas. Foi identificado que no *dataset* estão presentes gastos relacionados com a liderança partidária dentro da câmara dos deputados. Estas lideranças são multi-partidárias, causando assim a ausência da indicação de partido. Da mesma forma a indicação de *UF* (Unidade Federal) é ausente, já que são lideranças da federação e não um estado específico.

b. Além das colunas acima citadas, foi encontrado a ausência de valores para a *feature Tipo*. Foi identifica que o relatório de despesas do deputado não descriminou o gasto.

c. Para resolução deste problema foi decidido que haveria a inserção de valores para preservação destes dados potencialmente significativos. A inserção seguiu a tabela 4 abaixo:

Tabela 4 - Tabela com inserção de dados para posições NaN.

#	Feature	Valor original	Valor inserido	Comentários
1	UF	NaN	N/A	Não Aplicável
2	Partido	NaN	N/A	Não Aplicável
3	Tipo	NaN	S/D	Sem Descrição

12. Verificação final

a. Verificou-se o *dataframe* modelado para identificar potenciais erros como valores negativos, *features* com nome errado, entre outras possibilidades.

13. Exportação de Dataset

a. Foi realizada a exportação do conteúdo do pandas *dataframe* para um arquivo CSV, através da função pandas *to_csv*, definindo separador como ';' e *enconding* com *latin1*.

b. Nome do arquivo: *df_trienio_limpo.csv*

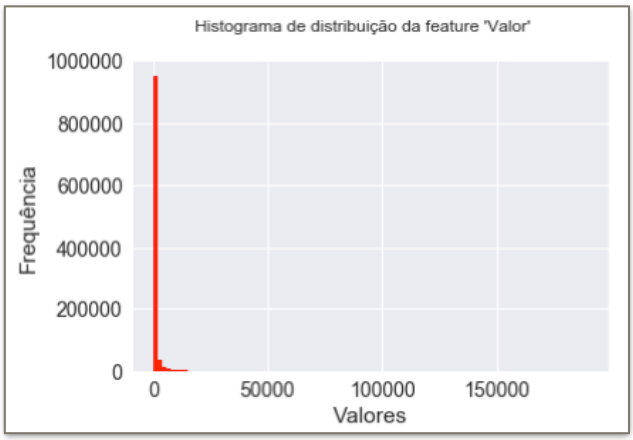
Visualização exploratória

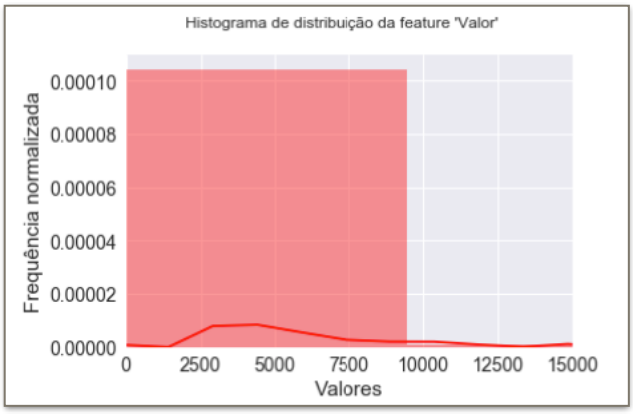
Visualização exploratória é importante para visualizarmos conceitos complexos para entendermos numericamente. Seguem as visualizações realizadas e suas explicações.

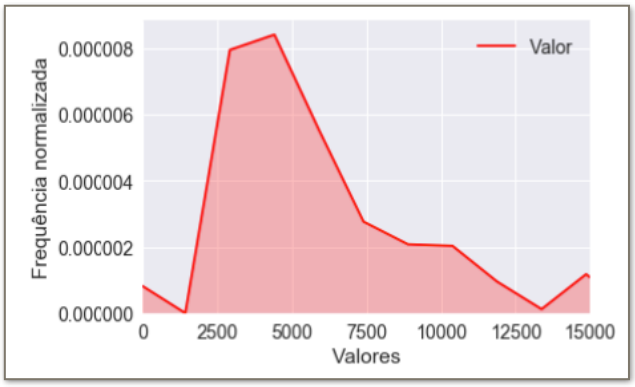
Visualização do *Dataset*

a. Distribuição

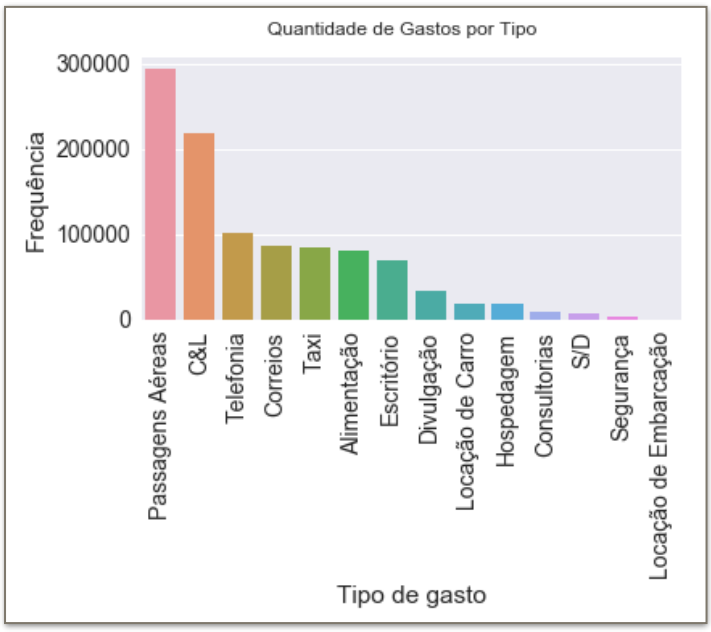
Utilizou-se de histograma para verificar a distribuição dos dados. Identificou-se que os dados tem uma distribuição right skewed. Isto se deve a uma concentração muito grande de valores pequenos nas despesas registradas.
Esta concentração já alerta para a necessidade de normalização ou aplicação de 'logaritimização' sobre o label no momento da aplicação de algoritmos de regressão linear.







b. Quantidade de gastos por tipo

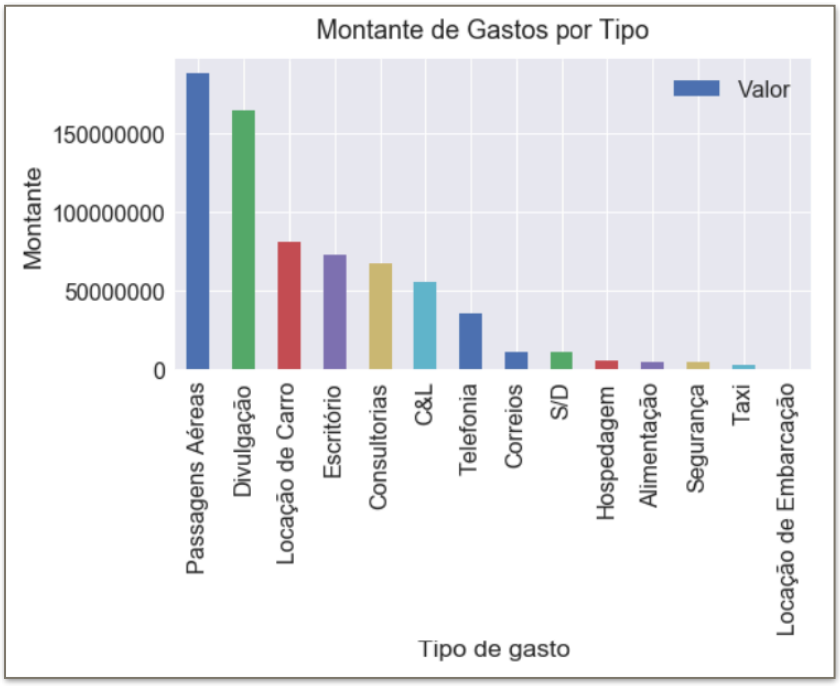


O gráfico acima expressa a quantidade de despesas submetidas para registro de despesas, podemos identificar as três maiores despesas: passagens aéreas, combustíveis e lubrificantes (C&L) e telefonia.

O que chama atenção é que despesas sem descrição (S/D) estão a frente de Segurança e Locação de Embarcação em sua quantidade de lançamentos contábeis. Relação entre si:

- As despesas de Passagens Aéreas representam aproximadamente três vezes o gasto de telefonia, e uma vez e meia os gastos de combustíveis e lubrificantes (C&L).
- As despesas que estão em posição 4 a 7, juntas representam o primeiro tipo de maior despesa.
- Chama muito a atenção que os deputados tem mais pedido de reembolso de passagens aéreas que reembolso de uso de Taxi, pedágio e estacionamento, ou mesmo outros meios de transporte.

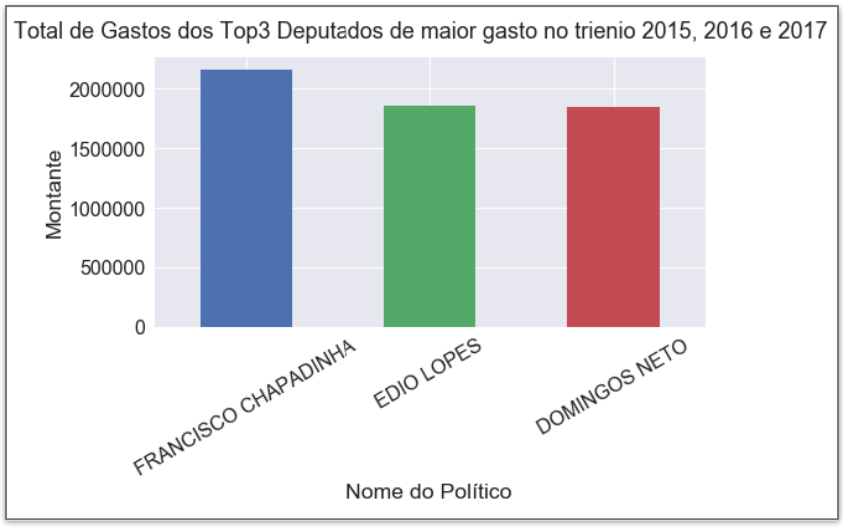
c. Montante de gastos por tipo



O gráfico acima expressa o 'Montante de gastos por tipo de despesa'. Pode-se notar que o primeiro lugar em montante, também está em primeiro no gráfico de quantidade de gastos, sendo um item com muito peso no total de despesas.
Relação entre si:

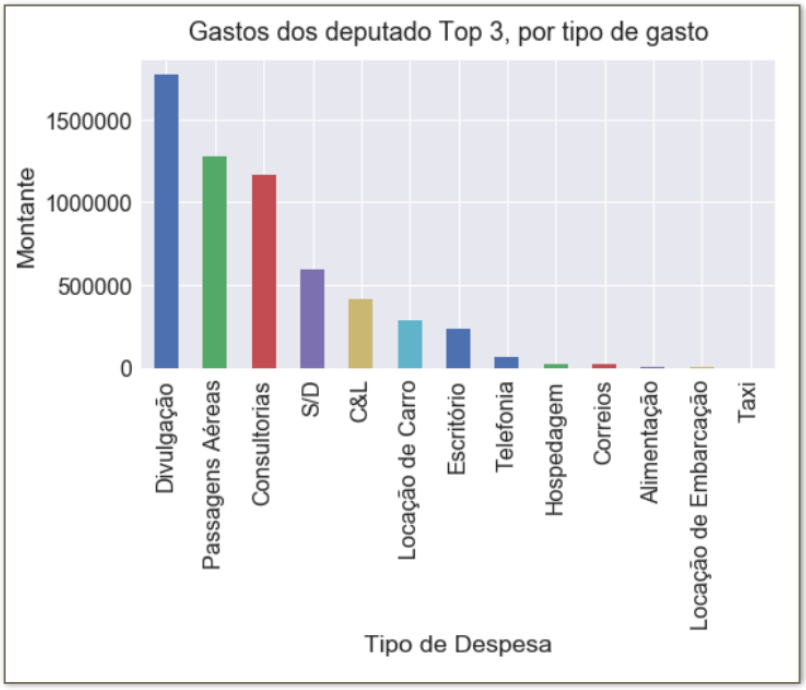
- O gasto com despesas relacionadas a passagens aéreas, como esperado, tem o maior montante acumulado. Esta despesa é seguida muito de perto pela despesa com divulgação.
- A despesa com passagens aéreas definitivamente sobrepuja todas as demais despesas relacionadas a transporte.

d. Top 3 Deputados no triênio (2015, 2016 e 2017)



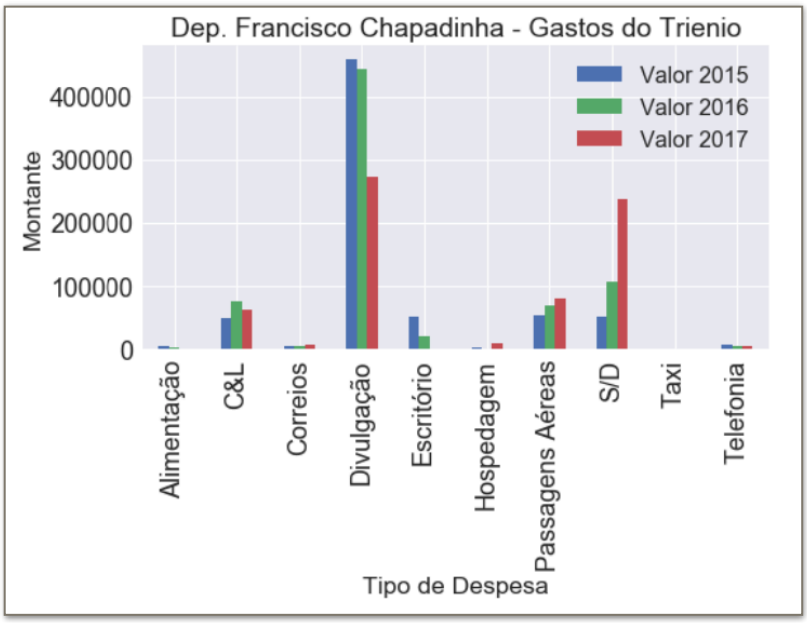
O gráfico acima lista os top 3 deputados em ranking de gastos totais no triênio analisado.
Édio Lopes e Domingos Neto tem os gastos totais com valores muito próximos.

e. Gastos dos deputado Top 3, por tipo de gasto



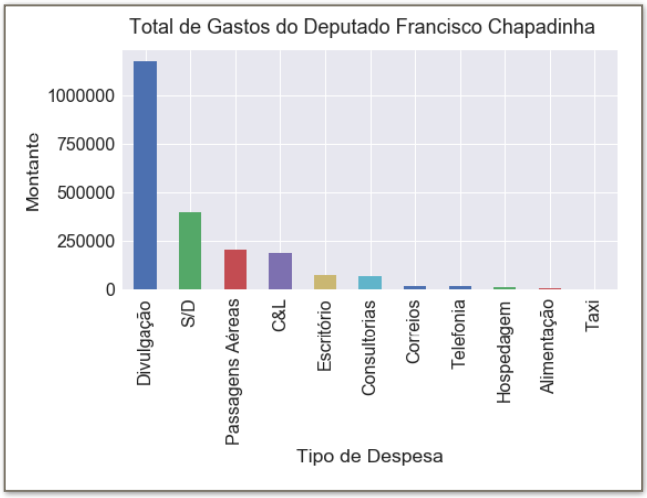
O gráfico acima ilustra os tipos de despesas realizados.
Relação entre si:
O gasto tem um comportamento distinto do apresentado na análise geral de gastos. Passagens aéreas que representavam notoriamente o maior gasto no *dataset*, neste caso ficou em terceiro lugar comm menos da metade do primeiro colocado: divulgação.

f. Dep. Francisco Chapadinha - Gastos do triênio

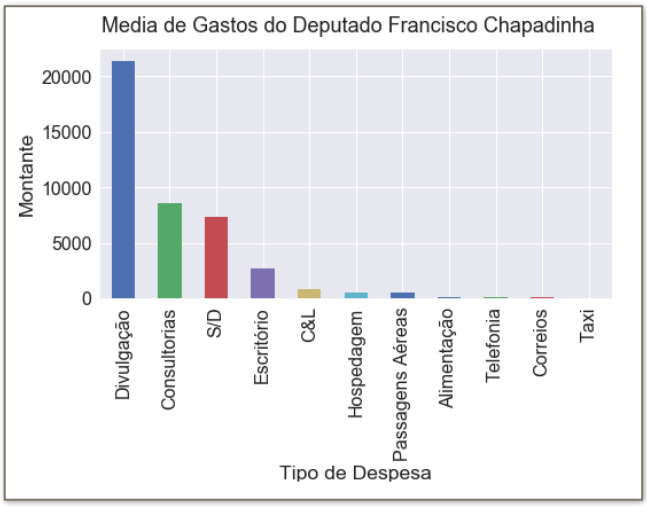


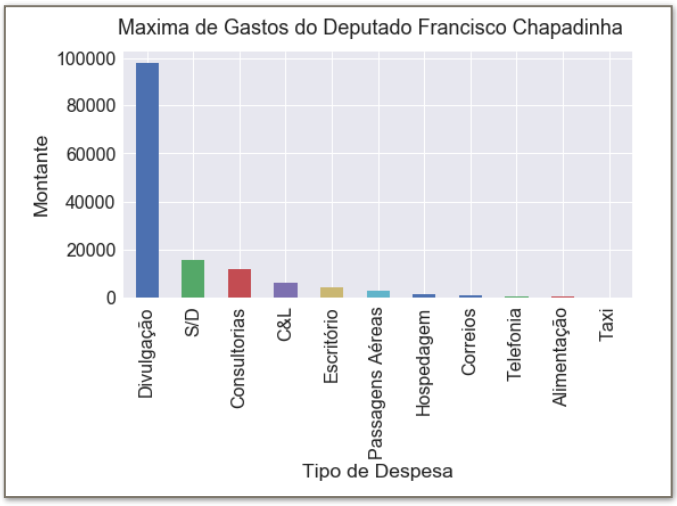
O gráfico acima expressa uma comparação de gastos por tipo de despesa do Dep. Francisco Chapadinha.
Há uma queda significativa de valores relacionados a divulgação, o aumento suave dos gastos com passagens aéreas, e o aumento de gastos S/D (sem descrição).
É possível que queda com gastos de divulgação estejam distorcidos em detrimento de lançamento de parte destes gastos em S/D.

g. Dep. Francisco Chapadinha - Total, Média e Máxima de Gastos

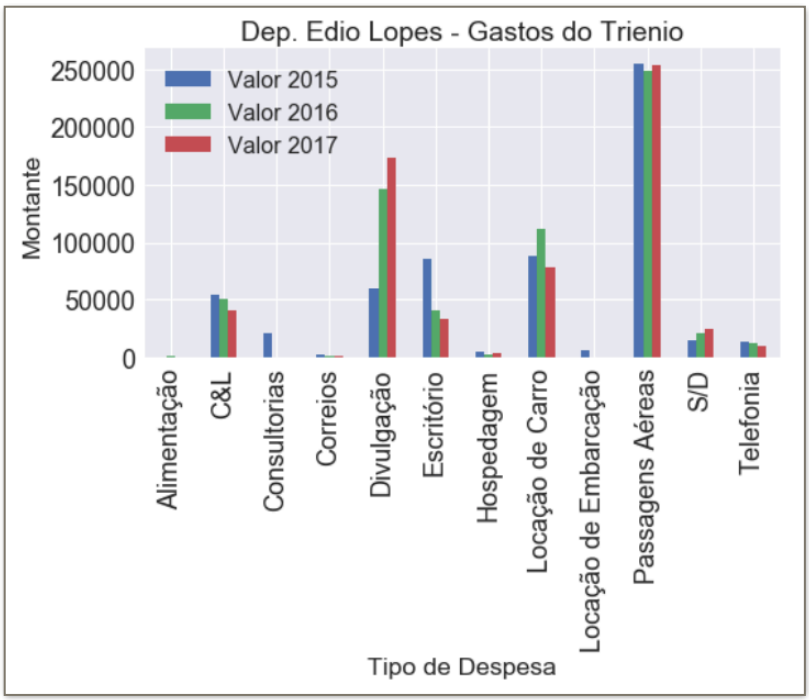


Os gráficos ao lado expressam a média e máxima de gastos por tipo de despesa do dep. Francisco Chapadinha.
A concentração de gastos está focada no tipo Divulgação/
Comentários a respeito de S/D:
O gráfico com a indicação de valor máximo de despesa tem em sua segunda posição gastos sem descrição (S/D), o que do ponto de vista de análise prejudica em muito o entendimento.
S/D ocupa a terceira posição de gastos médios.
Gastos S/D (sem descrição) dificultam muito a análise de gastos dos deputados.





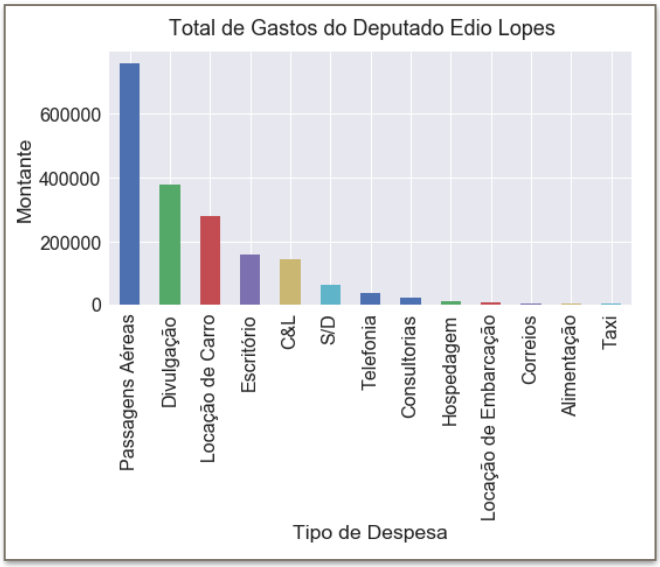
h. Dep. Edio Lopes - Gastos do triênio



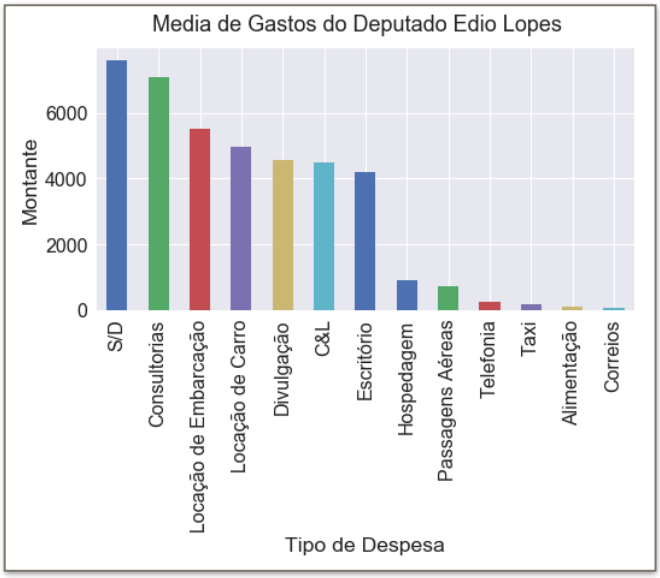
O gráfico acima expressa uma comparação de gastos por tipo de despesa do Dep. Edio Lopes dentro do triênio. O que chama atenção neste gráfico é a distribuição, mesmo que pequena, em outras despesas. Os gastos com veículo auto-motor e passagens aéreas são crescentes.

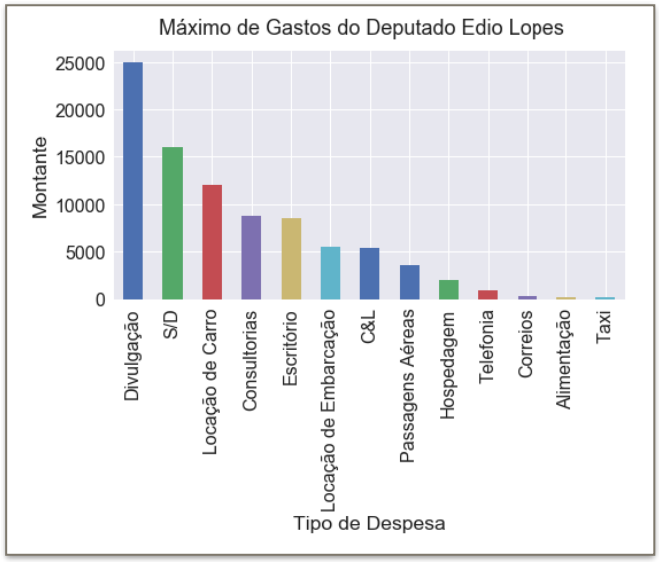
Gastos com C&L, escritório e hospedagem tem comportamentos de aumento discreto e similares. Gastos com divulgação tiveram um pico no ano de 2016.

i. Dep. Edio Lopes - Total, Média e Máxima de Gastos

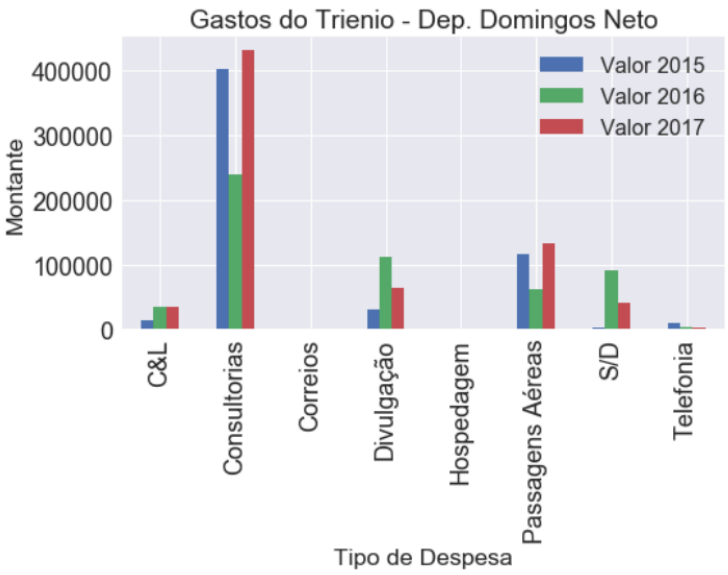


Os gráficos ao lado expressam a média e máxima de gastos por tipo de despesa do dep. Edio Lopes.
Os gastos do deputado ficaram melhor distribuídas em ambos gráficos, se comparada com os outros deputados top 3.
Mais uma vez a despesa sem descrição ocupa grande destaque. S/D ocupa primeiro lugar na média de gastos e segundo lugar entre os gastos mais altos.
Novamente impossibilitando uma análise mais acurada de gastos.



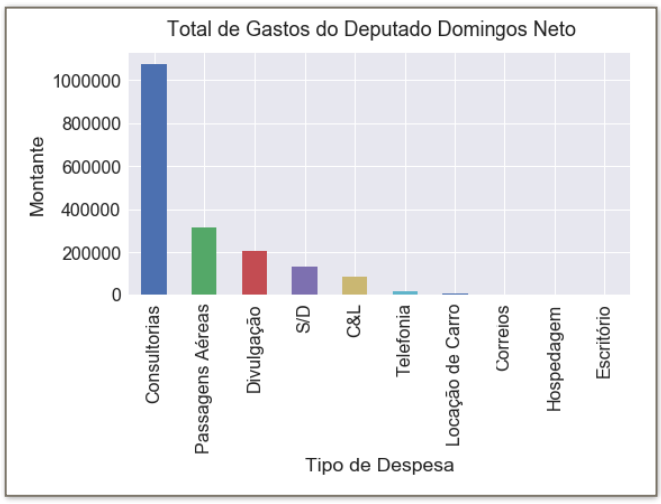


j. Dep. Domingos Neto - Gastos do triênio

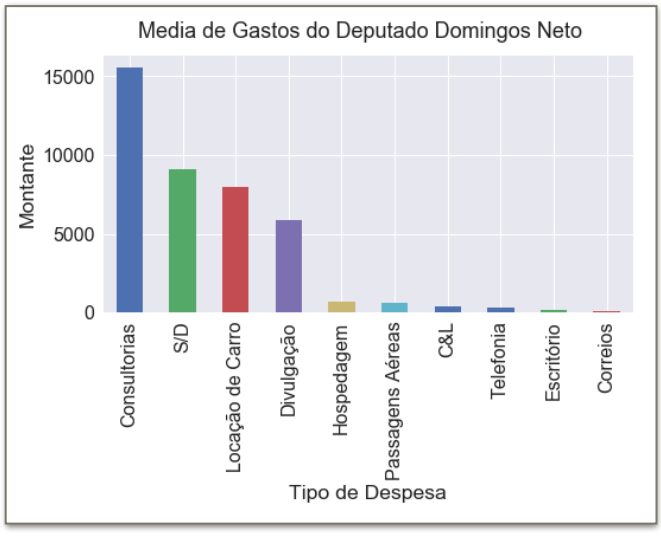


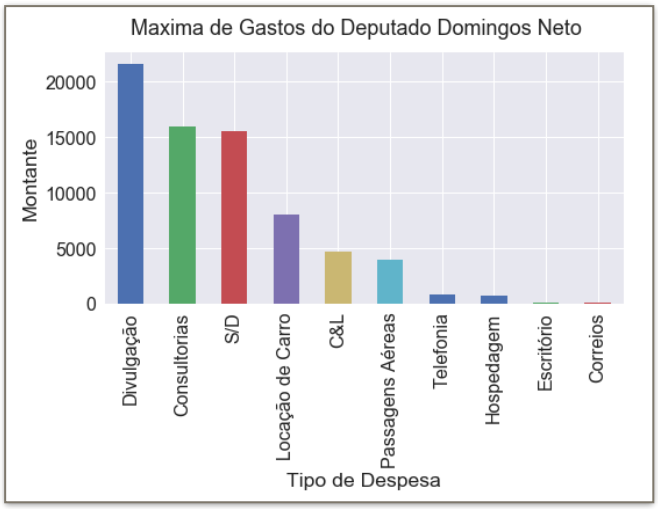
O gráfico acima expressa uma comparação de gastos por tipo de despesa do Dep. Domingos Neto.
Há uma significativa oscilação de valores de gastos na qual os gastos com consultorias e passagens aéreas tiveram o mesmo comportamento, assim como divulgação e S/D tiveram comportamento similares.

k. Dep. Domingos Neto - Total, Média e Máxima de Gastos



Os gráficos ao lado expressam a média e máxima de gastos por tipo de despesa do dep. Edio Lopes.
Os gastos do deputado ficaram melhor distribuídas em ambos gráficos, se comparada com os outros deputados top 3.
Mais uma vez a despesa sem descrição ocupa grande destaque. S/D ocupa primeiro lugar na média de gastos e segundo lugar entre os gastos mais altos. Novamente impossibilitando uma análise mais acurada de gastos.





Algoritmos e técnicas

Organização

Para manter a organização do trabalho, cada fase foi realizada em diferentes Jupyter Notebooks. Esta prática proporcionou o controle de revisão das atividades e sua navegação pelo notebook.

São estes os notebooks gerados:

- '1 - DATA CLEANING_V2.ipynb'
- '2 - EXPLORATION_V4.1.ipynb'
- '3 - PREDICTION_DOMINGOS_NETO_V6.ipynb'
- '4 - PREDICTION_FCO_CHAPADINHA_V6.ipynb'
- '5 - PREDICTION_ROCHA_V6.ipynb'

Limpeza de dados

Foram utilizados diversas funções e métodos das bibliotecas *Numpy, Pandas, OS, seaborn e matlibplot,* para realizar a limpeza dos *datasets* importados.

Além das bibliotecas foram utilizadas métodos para a verificação a cada operação com *dataframes,* como por exemplo a execução consecutivas de funções *head, describe* e *shape.* Esta sequência impediu o monitoramento contra alterações indesejadas.

Predição

Inicialmente no planejamento do projeto foi estimado a utilização de algoritmos de regressão linear abaixo:

- Support Vector Machine.
- K-Nearest Neighbors.
- Linear regression.
- Bayesian regression.

Tendo em vista a baixa dimensionalidade do *dataset* final para predição, sua característica de distribuição right skewed e alta amplitude, foram testadas os algoritmos abaixo listados, tendo em vista a preferência a simplicidade de sua aplicação. Abaixo estão detalhados os algoritmos e seus prós e contras mais relevantes a este projeto.

PróS e Contras dos Algoritmos

#	MODELO	PRÓS	CONTRAS
1	Linear Regression	★Algoritmo de rápida execução. ★Não requer tuning ★Altamente interpretável. ★Fácil entendimento de seu funcionamento.	★Tende a ter baixa precisão. ★Presume uma relação completamente linear do feature com o label. ★Pode tender a overfitting.
2	Lasso Regression	★Tende a ter menor overfitting que outros algoritmos de regressão linear. ★Costuma ter boa aderência para aplicações diversas de RL. ★Tem boa performance em caso da quantidade de features ser maior que a de linhas.	★Costuma ter um risco maior de produzir um modelo que não tenha sentido. ★Tende a ignorar features de menor importância. ★Funciona bem com features esparsas, típicas de dummies featues,
3	Lars Lasso Regression	★Performa bem com datasets pequenos. ★Algoritmo de rápida execução. ★Não requer tuning ★Altamente interpretável.	★Tende a ter baixa precisão. ★Funciona bem com features esparsas, típicas de dummies featues,
4	Ridge Regression	★Performa bem com datasets pequenos. ★Algoritmo de rápida execução. ★Não requer tuning ★Altamente interpretável.	★Tende a ter baixa precisão. ★Pode tender a overfitting. ★Sofre com outiers.
5	Bayesian Ridge	★Performa bem com datasets pequenos. ★Algoritmo de rápida execução. ★Não requer tuning ★Altamente interpretável.	★Tende a ter baixa precisão. ★Pode tender a overfitting. ★Sofre com outiers.

Entre estes algoritmos percebeu-se que reagiam diferentemente para com o pré-processamento dos dados. Como por exemplo a aplicação de função logarítmica, na qual alguns algoritmos tinham diferenciadas performances, ou até mesmo não conseguiam generalizar com os dados de treino fornecidas.

Durante o pré-processo dos dados foi necessário a utilização da função *get_dummies* para poder transformar os dados object da *feature* tipo em valores numéricos. Sendo entre as possibilidades possíveis na mais simples e mostrou-se adequada ao tipo de dados.

1.5 Benchmark

Este projeto tem como benchmark iniciativas que analisam através de estatística descritiva os gastos de verbas parlamentares de senadores e deputados. Por outro lado, a predição de através de regressão linear utiliza como referência projeto de predição de preços.

Estatística descritiva

Serão utilizadas como referência as soluções disponíveis nos web sites abaixo:

a. **Precisamos falar sobre a Cota Parlamentar** - https://medium.com/data-science-brigade/precisamos-falar-sobre-a-cota-parlamentar-c58a73392148

   - Esta publicação realiza uma análise estatística dos gastos parlamentares e encontram relações interessantes entre gastos realizados. Esta é a principal referência a ser utilizada no projeto no tocante a estatística descritiva.

b. **Monitora, Brasil** - https://monitorabrasil.org

   - Este web site oferece a somatória de gastos dos deputados e também o ranking de maiores gastos dentre todos os deputados. É uma referência adicional ao projeto com alguns dados estatísticos descritivos.

c. **AKAN** - Acompanhamento de gastos de deputados - http://visualizemobile.github.io/akan/

   - Este aplicativo também pode ser utilizado como referência complementar no projeto.

d. **Operação Serenata de Amor** - https://serenata.ai

   - Este site tem referências estatísticas complementares e algumas abordagens que podem ser utilizadas neste projeto.

Predição de valor

e. Regressão linear: para as atividades relacionadas a regressão será usada o projeto de predição de valor de imóveis da cidade de Boston criado durante o curso e o projeto abaixo:

https://www.coursera.org/learn/ml-foundations/lecture/2HrHv/learning-a-simple-regression-model-to-predict-house-prices-from-house-size

3. METODOLOGIA

3.1 Pré-processamento de dados

Foi necessário realizar algumas transformações nos dados para que pudesse ser realizada a utilização da biblioteca *Scikit Learn*.

Primeiramente o algoritmo de regressão da *Scikit* não possui tratamento para valores não numéricos (*category*). O que impactou diretamente no *feature Tipo*, que diz respeito a tipos de despesas realizadas. Para resolver esta questão foi estudado os tipo de encobre disponíveis para serem aplicados, como por exemplo o *encoding* para valores ordinais. Foi escolhido por ser mais adequado utilizar (OHE) *dummies* pois as categorias não eram ordinais em sua essência.

A distribuição dos dados é *right skewed,* com uma altíssima concentração de valores baixos e alguns valores cem vezes maiores. Como esta alta amplitude é real devido a características intrínsecas do *dataset* e não deve ser tratada como *outliers*, definiu-se que a aplicação de função logarítmica seria mais adequado. Um problema ocorreu com o dataset do Dep. Rocha. No qual a utilização de random para separar os dataset X e y prejudicou a utilização de quaisquer modelos preditivos, como pode ser visto na coluna amarela. A solução foi deixar a opção random da função como None.

Após transforma os dados *labels* (y) e realizar alguns testes com os algoritmos percebeu-se que era importante verificar qual é o comportamento dos algoritmos com o y logaritmo e com o y original. Surpreendentemente houve algoritmos com melhores performances com o dado transformado e outros com os dados originais.

Esta operação é muito simples e requer a recuperação dos dados através da aplicação de função exponencial.

3.2 Implementação

DEPUTADO				FRANC.	ROCHA	DOMIN.	ROCHA
#	MODEL	Logarítmo	Type	Score	Score	Score	Score
1	Linear Regression	Não	Train	0.90	0.97	0.90	0.94
2	Linear Regression	Não	Test	0.71	0.85	0.80	-3.26
3	Linear Regression	Sim	Train	0.89	0.89	0.92	0.94
4	Linear Regression	Sim	Test	0.61	0.88	0.94	-0.98
5	Lasso Regression	Não	Train	0.90	0.97	0.90	0.94
6	Lasso Regression	Não	Test	0.71	0.88	0.82	-0.90
7	Lasso Regression	Sim	Train	0.0	0	0.0	0.0
8	Lasso Regression	Sim	Test	-0.33	-0.04	-0.05	-0.68
9	Lars Lasso Regression	Não	Train	0.90	0.97	0.90	0.94
10	Lars Lasso Regression	Não	Test	0.71	0.88	0.82	-1.28
11	Lars Lasso Regression	Sim	Train	0.67	0.23	0.68	0.47
12	Lars Lasso Regression	Sim	Test	-0.02	0.20	0.68	-0.37
13	Ridge Regression	Não	Train	0.88	0.88	0.87	0.92
14	Ridge Regression	Não	Test	0.63	0.70	0.66	-2.04
15	Ridge Regression	Sim	Train	0.86	0.84	0.87	0.91
16	Ridge Regression	Sim	Test	0.73	0.82	0.83	-0.38
17	Bayesian Ridge	Não	Train	3.12 e-10	0.0	5.68	0.0
18	Bayesian Ridge	Não	Test	-0.56	-0.65	-0.20	-15.81
19	Bayesian Ridge	Sim	Train	0.88	0.87	0.91	0.94
20	Bayesian Ridge	Sim	Test	0.72	0.86	0.91	-0.77

Para aplicação dos algoritmos de Machine Learning para predição de gastos de 2018 foram utilizados dados transformado por função logarítmica ou não, com intuito de verificar a performance do algoritmo nos dois casos. Pode-se ver nas indicações verdes os melhores resultados comparativos entre o mesmo algoritmo, com o uso de logaritmo ou não.

Percebeu-se durante a análise dos top 3 deputados de 2017 e do *dataset* como um todo, que os *datasets* dos deputados tem muitas diferenças quanto ao perfil de gastos, alterando assim em muito as *features dummies* para a predição. Estas variações impedem a escolha de um único algoritmo para tratar a analise de cada deputado, sendo necessário investigar qual algoritmo é mais adequado.

Para a organização dos resultados das predições foi criada uma única função para criar uma tabela com os dados calculados, chamada *tabela_resultado*.

3.3 Refinamento

Para realização das predições de regressão linear foram escolhidas algoritmos simples para aplicar em um *dataset* de predição um tanto pequeno. Durante o processo houve uma maior preocupação com o pré-processamento de dados para que as características de grande amplitude de dados conseguisse ser mitigada para um melhor resultado. Inicialmente houveram algumas versões intermediárias para exploração de dados e testes de algoritmos para entender a melhor abordagem.

O estudo inicial fez com que dois dos algoritmos definidos inicialmente fossem trocados por mais três

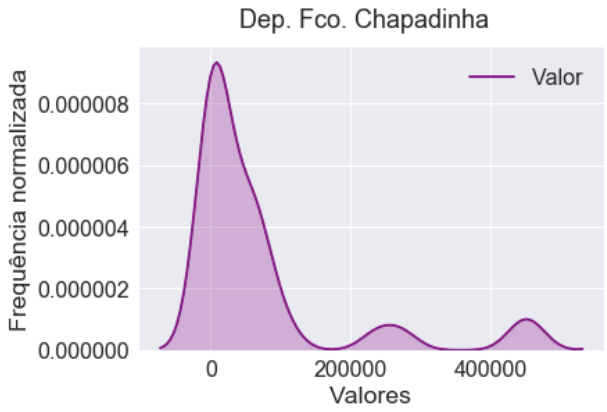
#	Algoritmos definidos inicialmente	Algoritmos utilizados
1	**Support Vector Machine.**	**Linear Regression**
2	K-Nearest Neighbors.	Lasso Regression
3	**Linear regression.**	**Lars Lasso Regression**
4	Bayesian regression.	Ridge Regression
5	**-**	**Bayesin Ridge Regression**

4. RESULTADOS

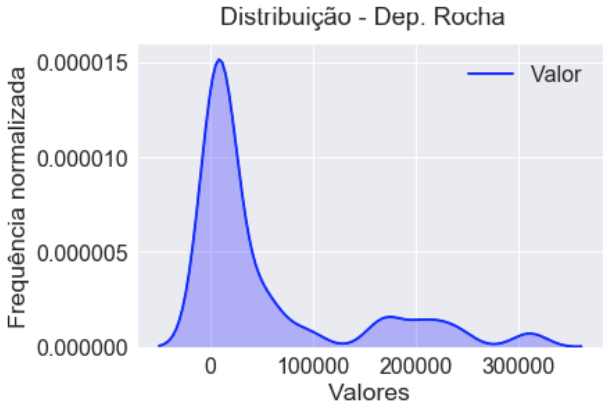
4.1 Modelo de avaliação e validação

Como mencionado anteriormente, percebeu-se que não há como definir um único algoritmo de Machine Learning para predizer com eficiência os gastos totais por tipo para o ano de 2018. Ao menos para os algoritmos de Machine Learning utilizados neste projeto.
Isto se deve ao fato de que não se pode generalizar entre os datasets individuais dos deputados, já que os perfis de gastos são diferentes como mostra o gráfico abaixo.

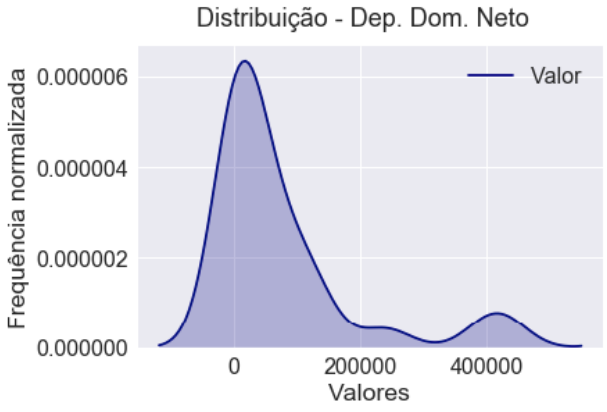
	Ano	Valor
count	31.00	31.00
mean	2,016.00	69,666.77
std	0.82	120,495.97
min	2,015.00	17.50
25%	2,015.00	4,441.03
50%	2,016.00	12,000.00
75%	2,017.00	65,434.91
max	2,017.00	459,300.00

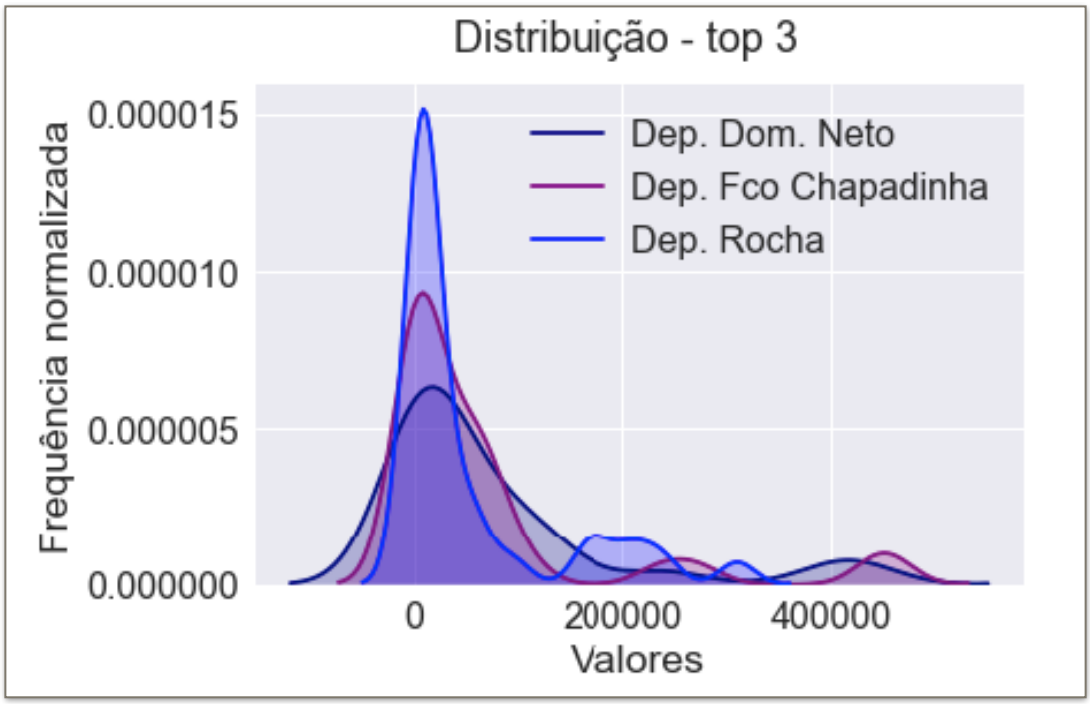


	Ano	Valor
count	33.00	33.00
mean	2,016.00	54,764.79
std	0.83	83,640.38
min	2,015.00	600.00
25%	2,015.00	7,207.45
50%	2,016.00	11,800.00
75%	2,017.00	50,002.27
max	2,017.00	310,900.00



	Ano	Valor
count	24.00	24.00
mean	2,015.96	77,000.54
std	0.81	119,900.64
min	2,015.00	165.98
25%	2,015.00	2,557.35
50%	2,016.00	33,353.68
75%	2,017.00	97,585.00
max	2,017.00	432,000.00





Entende-se que seria possível que o objetivo fosse realizar a predição de gastos totais de toda a base juntamente, o que não nos interessa neste projeto.
Para os top 3 deputados de 2017 foram escolhidos os seguintes algoritmos com seus respectivos ranking por resultado para *datasets* de treinamento e pré-processamentos:

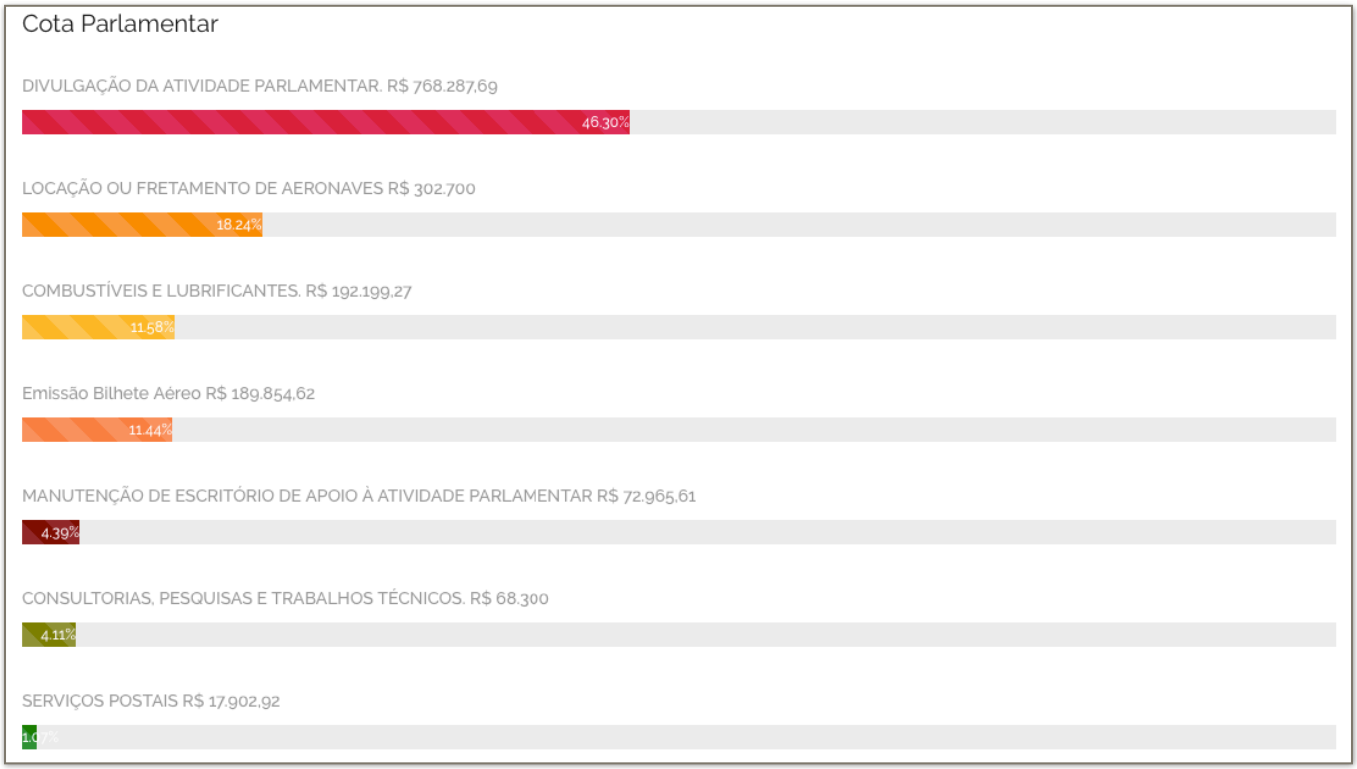
Ranking de Algoritmos por Deputados

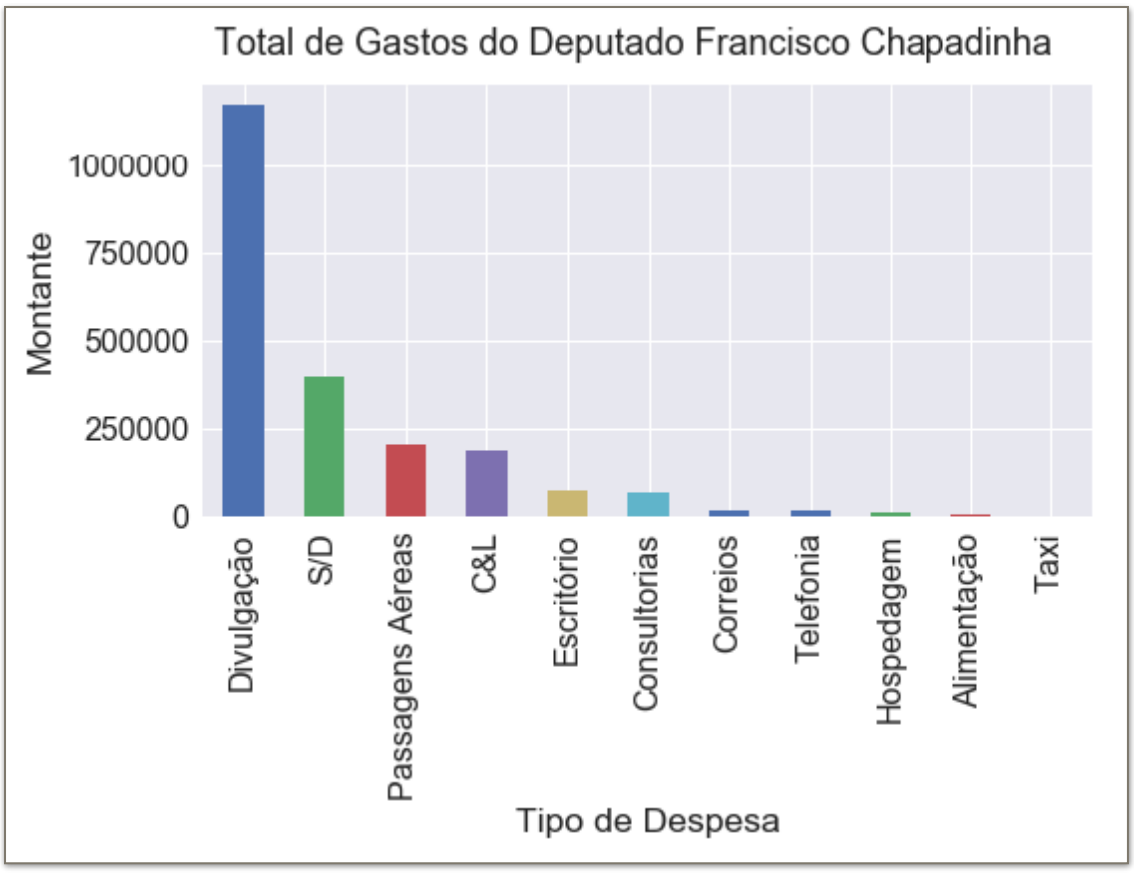
			Ranking de Algoritmos		
#	Modelo	Pre-processamento Logaritimico	Francisco Chapadinha	Rocha	Domingos Neto
1	Linear Regression	Dummie Features	3	-	-
2	Linear Regression	Dummie Features/ Log	-	-	1
3	Lasso Regression	Dummie Features	3	-	
4	Lasso Regression	Dummie Features/ Log	X	-	X
5	Lars Lasso Regression	Dummie Features	3	-	4
6	Lars Lasso Regression	Dummie Features/ Log	X	-	-
7	Ridge Regression	Dummie Features	-	-	-
8	Ridge Regression	Dummie Features/ Log	1	-	3
9	Bayesin Ridge Regression	Dummie Features	X	-	X
10	Bayesin Ridge Regression	Dummie Features/ Log	2	-	2

LEGENDA: ***N***: Número ordinal / '**-**' : Desclassificado / '**X**': Não utilizável.

4.1 Modelo de avaliação, validação e justificativa

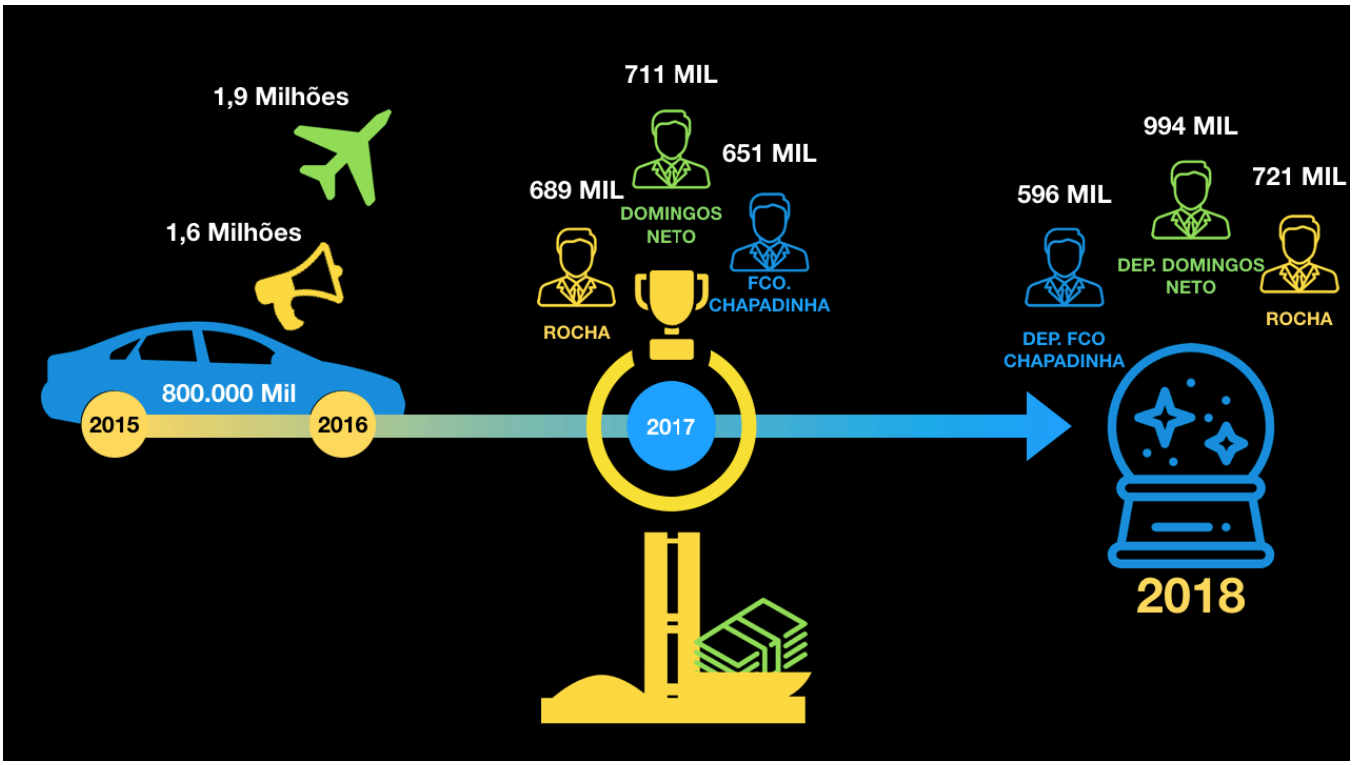
O benchmark disponível não possui cálculos preditos de gastos anuais, possuem acompanhamento de gastos atuais dos deputados. Esta nova abordagem aumenta o potencial de contribuição para a comunidade.
Abaixo tem um exemplo do site Monitora Brasil e de do gráfico realizado neste projeto com valores totais.





5. Conclusão

5.1 Forma livre de visualização



5.2 Reflexão

De uma forma geral o projeto foi bem interessante em vários aspectos. Abaixo são listados o principais:

Dataset

Apesar do dataset original conter mais de um milhão de linhas, ao realizar o estudo de limpeza destes dados para realizar a predição anualizada de três do deputados, verificou-se que o dataset “útil” passa a ser muito pequeno, ao redor de 30 linhas, podendo não conseguir generalizar o modelo, sofrendo inclusive overfit.

Este dataset de trabalho é significativamente diferente para cada deputado, gerando uma série de problemas, como pré-processamento para tratar a alta amplitude dos dados e sua alta concentração (right skewed). O impacto foi percebido com clareza quando os modelos de predição ótimos para um dataset de deputado não tinha a mesma aderência em outro dataset de outro deputado.

Tendo em vista o pequeno dataset ‘útil' o resultado pode ser considerado satisfatório como estudo de tendência de gastos por tipo de gasto.

5.3 Melhorias

Uma série de melhorias podem ser feitas. Como as listadas abaixo:

Dataset

Pode-se enriquecer o dataset com dados adicionais de notas fiscais no arquivo governamental, porém demandará muito trabalho em um longo período de tempo para transformar estes dados de imagem em dados eletrônicos. Salvo utilização de ML para imagem.

Algoritmos

Percebeu-se que é necessário realizar testes com outros algoritmos existentes para regressão linear, como por exemplo rede neural do Scikit Learn.

Os algoritmos utilizados também podem passar por uma série adicional de testes para identificar diferentes possibilidades.

Pose-se também utilizar outros algoritmos que não façam parte da biblioteca Scikit Learn.

Comentário geral

Acredita-se que um modelo muito melhor que o presente trabalho pode ser desenvolvido com as melhorias apontadas acima.

6. ANEXOS

6.1 Tabela de features

A tabela abaixo é fornecida pelo site de transparência da câmara, podendo ser acessado através do link http://www2.camara.leg.br/transparencia/cota-para-exercicio-da-atividade-parlamentar/explicacoes-sobre-o-formato-dos-arquivos-xml

#	Uso	Elemento de Dado	Definição do Dado
1	S	**txNomeParlamentar**	Nome adotado pelo Parlamentar ao tomar posse do seu mandato.
2	N	**ideCadastro**	Número que identifica unicamente um deputado federal na CD.
3	N	**nuCarteiraParlamentar**	Documento usado para identificar um deputado federal na CD.
4	N	**nuLegislatura**	Legislatura: Período de quatro anos coincidente com o mandato parlamentar dos Deputados Federais.
5	S	**sgUF**	No contexto da cota CEAP, representa a unidade da federação pela qual o deputado foi eleito e é utilizada para definir o valor da cota a que o deputado tem.
6	S	**sgPartido**	O seu conteúdo representa a sigla de um partido. Definição de partido: é uma organização formada por pessoas com interesse ou ideologia comuns, que se associam com o fim de assumir o poder para implantar um programa de governo.
7	N	**codLegislatura**	Legislatura: Período de quatro anos coincidente com o mandato parlamentar dos Deputados Federais.
8	N	**numSubCota**	No contexto da Cota CEAP, o conteúdo deste dado representa o código do Tipo de Despesa referente à despesa realizada pelo deputado e comprovada por meio da emissão de um documento fiscal, a qual é debitada na cota do deputado.
9	S	**txtDescricao**	O seu conteúdo é a descrição do Tipo de Despesa relativo à despesa em questão.
10	N	**numEspecificacaoSubCota**	No contexto da Cota CEAP, há despesas cujo Tipo de Despesa necessita ter uma especificação mais detalhada (por exemplo, “Combustível”).
11	N	**txtDescricaoEspecificacao**	Representa a descrição especificação mais detalhada de um referido Tipo de Despesa.
12	N	**txtFornecedor**	O conteúdo deste dado representa o nome do fornecedor do produto ou serviço presente no documento fiscal
13	N	**txtCNPJCPF**	O conteúdo deste dado representa o CNPJ ou o CPF do emitente do documento fiscal, quando se tratar do uso da cota em razão do reembolso despesas comprovadas pela emissão de documentos fiscais.
14	N	**txtNumero**	O conteúdo deste dado representa o número de face do documento fiscal emitido ou o número do documento que deu causa à despesa debitada na cota do deputado.
15	N	**indTipoDocumento**	Este dado representa o tipo de documento do fiscal – 0 (Zero), para Nota Fiscal; 1 (um), para Recibo; e 2, para Despesa no Exterior.
16	N	**datEmissao**	O conteúdo deste dado é a data de emissão do documento fiscal ou a data do documento que tenha dado causa à despesa.
17	S	**vlrDocumento**	O seu conteúdo é o valor de face do documento fiscal ou o valor do documento que deu causa à despesa.
18	N	**vlrGlosa**	O seu conteúdo representa o valor da glosa do documento fiscal que incidirá sobre o Valor do Documento, ou o valor da glosa do documento que deu causa à despesa.
19	N	**vlrLiquido**	O seu conteúdo representa o valor líquido do documento fiscal ou do documento que deu causa à despesa e será calculado pela diferença entre o Valor do Documento e o Valor da Glosa.
20	S	**numMes**	O seu conteúdo representa o Mês da competência financeira do documento fiscal ou do documento que deu causa à despesa.
21	S	**numAno**	O seu conteúdo representa o Ano da competência financeira do documento fiscal ou do documento que deu causa à despesa.
22	N	**numParcela**	O seu conteúdo representa o número da parcela do documento fiscal. Ocorre quando o documento tem de ser reembolsado de forma parcelada.
23	N	**txtPassageiro**	O conteúdo deste dado representa o nome do passageiro, quando o documento que deu causa à despesa se tratar de emissão de bilhete aéreo.
24	N	**txtTrecho**	O conteúdo deste dado representa o trecho da viagem, quando o documento que deu causa à despesa se tratar de emissão de bilhete aéreo.
25	N	**numLote**	No contexto da Cota CEAP, o Número do Lote representa uma capa de lote que agrupa os documentos que serão entregues à Câmara para serem ressarcidos.
26	N	**numRessarcimento**	No contexto da Cota CEAP, o Número do Ressarcimento indica o ressarcimento do qual o documento fez parte por ocasião do processamento do seu reembolso.
27	N	**vlrRestituicao**	O seu conteúdo representa o valor restituído do documento fiscal que incidirá sobre o Valor do Documento.
28	N	**nuDeputadoId**	Número que identifica um Parlamentar ou Liderança na Transparência da Cota para Exercício da Atividade Parlamentar.