Cliente “X”
ONG

[Logo do Cliente]

brasil .monks

Análise de Dados

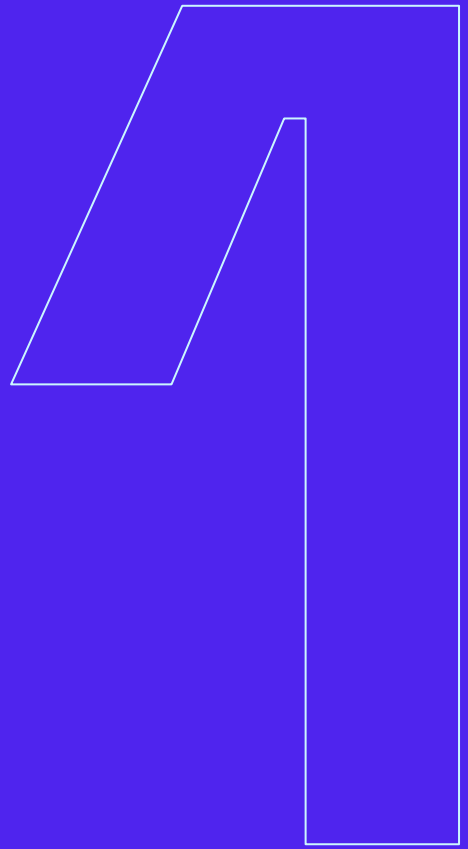
Links dos Materiais:

GitHub - Todo o Case está no GitHub - SQL + Análise de Dados.

Arquivos em Anexo foram enviados por e-mail também.

Última etapa do projeto (Aqui o trabalho poderia ser repassado para DA, por exemplo, mas optei por fazer uma dash simples - Bem simples mesmo, pois há espaço para melhorias - o Intuito é mostrar as integrações de dados no DataViz e que as etapas deram certo). LINK da Dash (existem 3 abas na Dash, uma por Estado).

Índice



Introdução

Introdução

1 – Dados da data: 01/2015 até 12/2021 (por mês)

2 – Bibliotecas utilizadas:

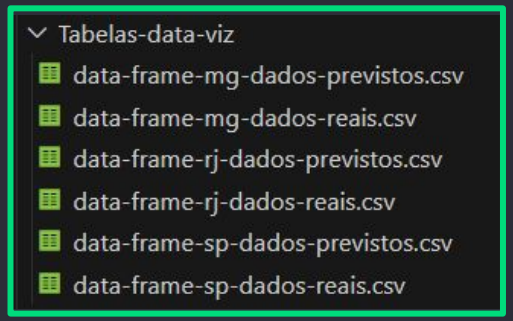
pandas - Para DataFrame
datetime - Para o tratamento de dados de data
os - Para leitura, download e upload de arquivos
matplotlib - Para plot de gráficos (dataviz)
statsmodels - Para a regressão linear
numpy - Numpy trabalha em conjunto com o statsmodels no presente trabalho
sklearn - Biblioteca para regressão linear e calcular algumas métricas

3 – Cohorts considerados:

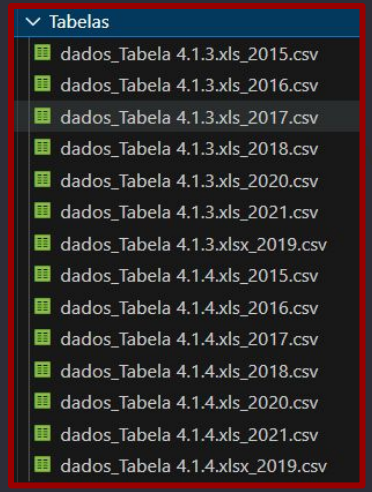
Por Região e Estado
Por Gênero
Séries Temporais (por Mês e Ano)

4 – Fonte de Dados e Ferramentas:

Google Drive - Download de .csv (fonte inicial dos dados)
VSCode (Python - ipynb) + bibliotecas Python



Planilhas de Output



Planilhas de Input

Métodos

Métodos

1 – Ler todos os Arquivo baixados (Extract) - E transportá-los em um único Data Frame em Python (Pandas);

2 – Transformar Dados e Guardá-los (Transform e Load) para aplicar a análise - Criação das Camadas Bronze, Silver e Gold para Backup de Dados;

3 – Elaborar a Regressão Linear (Análise feita por Estados):
SP - Teve uma maior análise de dados para comprovar algumas hipóteses, logo em seguida vimos os Estados de MG e RJ para fazer uma previsão de número de casamentos homoafetivos.
Foram retirados os Outliers;
Foi feita uma análise de precisão do modelo de regressão linear averiguamos os erros com base em uma Base de Dados real x dados previsto. (Validar modelo);
Depois de validar o modelo, criamos um range de cenários por Estado:

- Previsão dos dados em um cenário Pessimista
- Previsão dos dados em um cenário Otimista
- Previsão dos dados em um cenário Realista

4 – Plots de Gráficos por Estado e Alguns Histogramas:
Plots 2-D por entre data x número de Casamentos com a Regressão Linear aplicada;
Histogramas de Cohorts - Para tirar insights.

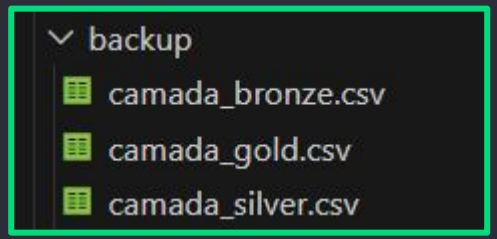
5 - Output do modelo preditivo - Planilhas .csv

Leitura do Data Frame

```python
# Inicializo o DataFrame principal em Pandas:

df = pd.DataFrame()
dfs = [] # Vetor fora do Pandas para receber appends


caminho = 'Tabelas/'
lista_arquivos = os.listdir(caminho)

# 'For' para varrer toda a pasta de planilhas mockadas:
for arquivo in lista_arquivos:
    df_aux = pd.read_csv(caminho+str(arquivo))
    dfs.append(df_aux)

df = pd.concat(dfs, ignore_index=True) # Concatenação de dataframes

df.to_csv('backup/camada_bronze.csv') # transformar Arquivo em .csv para camada Bronze

df
```

✓ 0.1s

	ano	uf	genero	mes	numero
0	2019	Rondônia	Masculino	Janeiro	0
1	2019	Rondônia	Masculino	Fevereiro	4
2	2019	Rondônia	Masculino	Março	1
3	2019	Rondônia	Masculino	Abril	0
4	2019	Rondônia	Masculino	Maio	0
...	...	...	...	...	...



Tratamento dos Dados

Métodos - Esboço do Projeto

1 – Ler todos os Arquivo baixados (Extract);

2 – Transformar Dados e Guardá-los (Transform e Load) para aplicar a análise:

Camada Bronze - Dados Brutos recebidos do Extract;

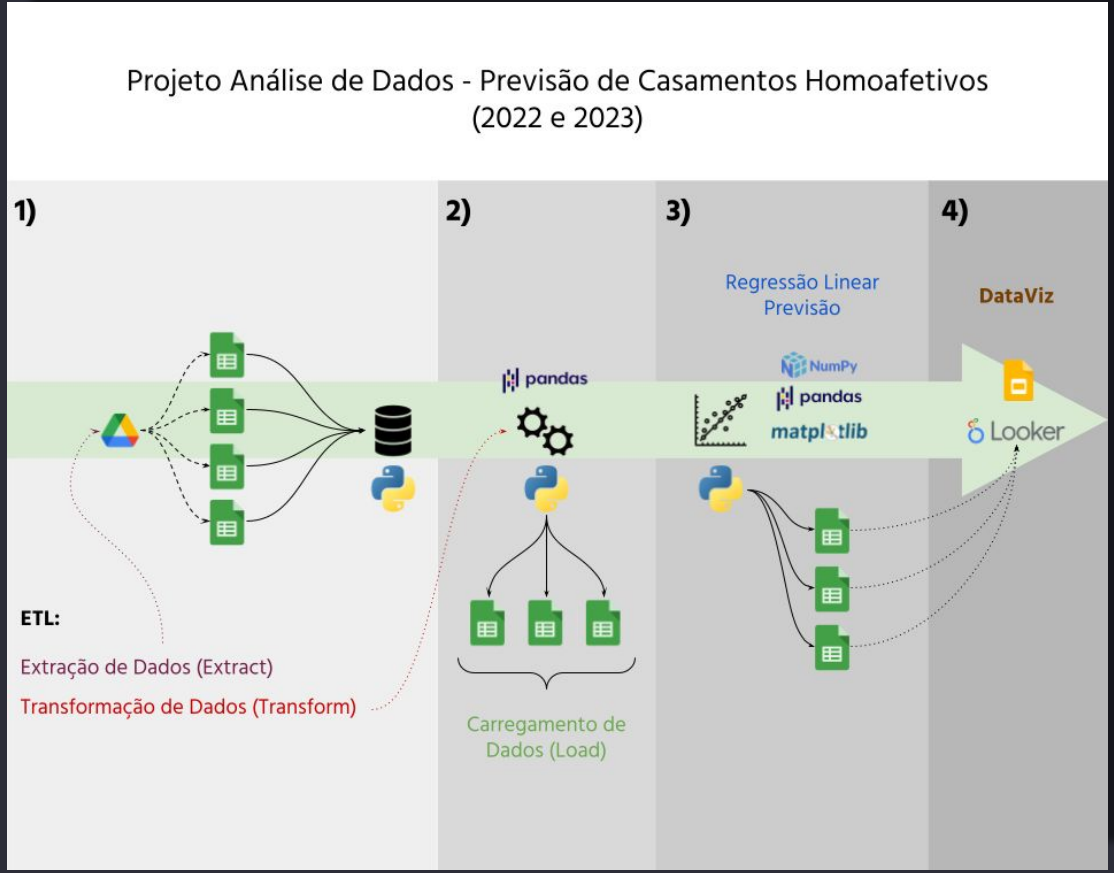
Camada Silver - Dados com tipagem correta (data correta + tipagem das colunas de número) e remoção dos valores NaN e Duplicatas;

Camada Gold - Dados prontos para serem visualizados - inserção de algumas colunas pertinentes para a análise - Exemplo inserção de Regiões (Sudeste, Nordeste, Norte, etc.)

3 – Elaborar a Regressão Linear + Output do modelo preditivo - Planilhas .csv;

4 – Plots de Gráficos por Estado e Alguns Histogramas - Vincular com a presente Apresentação e Looker (Dash);

5 – Conclusão com base nos Dados e Visão de Mercado.



Métodos - Esboço do Projeto

Regressão Linear
Previsão

NumPy

pandas

matplotlib

Falando um pouco mais sobre a Regressão Linear…

3 – Elaborar a Regressão Linear + Output do modelo preditivo - Planilhas .csv:

Foi feita usando a biblioteca Stasmodel;

Antes de aplicar a regressão linear - Usou-se o método de Tukey para retirar os Outliers pertinentes;

Para SP: Foi feito os Cohorts de Estado e Gênero para verificar se era possível somar as previsões fazendo para os gêneros separados e comparar com uma base que continha ambos os gêneros:

- **Regressão para ambos os Gêneros - SP;**
- **Regressão para Feminino - SP;**
- **Regressão para Masculino - SP;**
- **Verificação dos resultados (Masculino e Feminino) x (Cenário de Ambos) - SP.**

Para MG: Foi feito os Cohorts de Estado - Ambos os Gêneros (pois foi constatado que não há diferenças fazer regressão para gêneros separados em SP);

Para RJ: Foi feito os Cohorts de Estado - Ambos os Gêneros (pois foi constatado que não há diferenças fazer regressão para gêneros separados em SP).

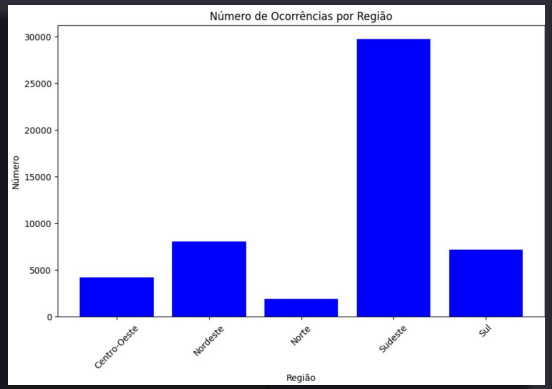
Resultados

3) - Explorando o problema do Cliente com Análise de Dados

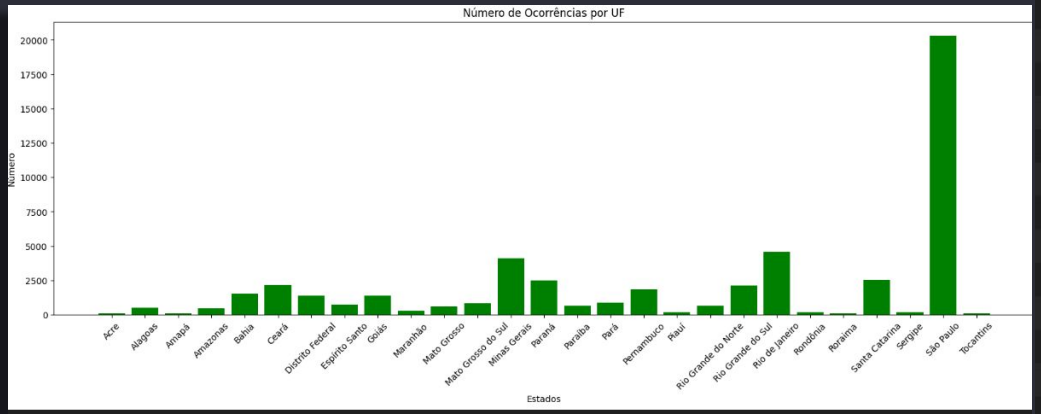
Queremos entender quais insights conseguimos tirar dessa base, e também qual a previsão de casamentos homoafetivos para os próximos 2 anos nos 3 estados com maior número de casamentos no último ano. Seria possível realizar essa análise?

Resultados - Insights Iniciais

Antes de responder a pergunta, vamos a alguns Insights...



regiao	numero
Centro-Oeste	4192
Nordeste	8070
Norte	1898
Sudeste	29714
Sul	7192

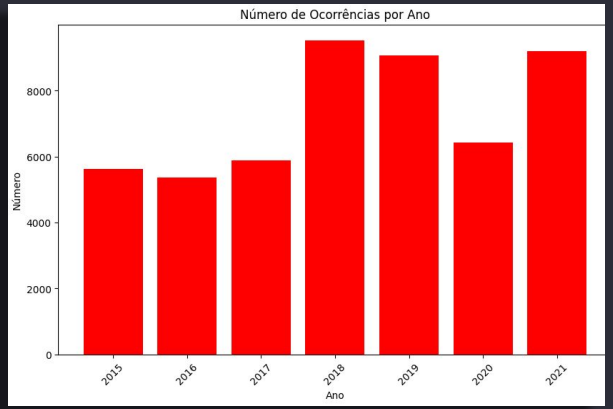


uf	numero
São Paulo	20285
Rio de Janeiro	4584
Minas Gerais	4087
Santa Catarina	2544
Paraná	2500
Ceará	2166
Rio Grande do Sul	2148
Pernambuco	1855
Bahia	1531
Distrito Federal	1378
Goiás	1367
Pará	882
Mato Grosso do Sul	819
Espírito Santo	758
Paraíba	656
Rio Grande do Norte	637
Mato Grosso	628
Alagoas	512
Amazonas	454
Maranhão	306
Rondônia	209
Sergipe	208
Piauí	199
Tocantins	102
Roraima	89
Amapá	81
Acre	81

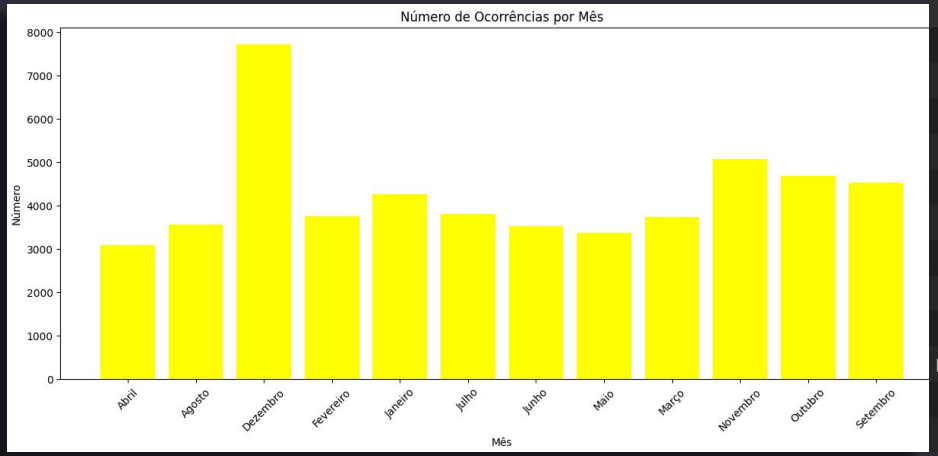
Recorte por Regiões e UF's - Dos Anos de 2015 até 2021
Histogramas do Número de Casamentos por Localidades
Ambos os Gêneros

Resultados - Insights Iniciais

Antes de responder a pergunta, vamos a alguns Insights...



ano	numero
2015	5614
2016	5354
2017	5887
2018	9520
2019	9056
2020	6433
2021	9202

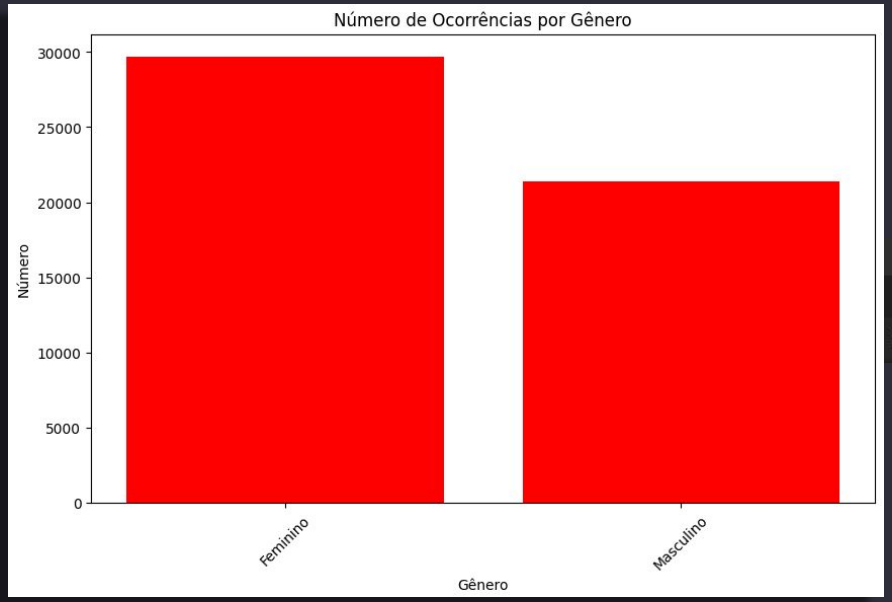


mes	numero
Abril	3077
Agosto	3566
Dezembro	7716
Fevereiro	3753
Janeiro	4252
Julho	3806
Junho	3519
Maio	3371
Março	3727
Novembro	5062
Outubro	4689
Setembro	4528

Recorte por Mês e Ano - De todas as Regiões do Brasil
Histogramas do Número de Casamentos por Tempo
Ambos os Gêneros

Resultados - Insights Iniciais

Antes de responder a pergunta, vamos a alguns Insights...



genero	numero
Feminino	29693
Masculino	21373

Recorte por Gênero - De todas as Regiões do Brasil

Histogramas do Número de Casamentos por Gênero

Das datas de 01/2015 até 12/2021

3) - Explorando o problema do Cliente com Análise de Dados

Queremos entender quais insights conseguimos tirar dessa base, e também qual a previsão de casamentos homoafetivos para os próximos 2 anos nos 3 estados com maior número de casamentos no último ano. Seria possível realizar essa análise?

Resultados - Insights

Voltando a pergunta do Cliente...

Filtros pertinentes e achar os 3 Estados com maior número de casamentos no último ano:

```
df_insight = df[(df['ano'] == '2021')] # Filtra todo df no último ano que é 2021
df_insight = df_insight.groupby('uf')['numero'].sum().reset_index() # Agrupa por Estado (uf)
df_insight = df_insight.sort_values(by='numero', ascending=False) # Ordena de forma decrescente pelo número somado
df_insight
```

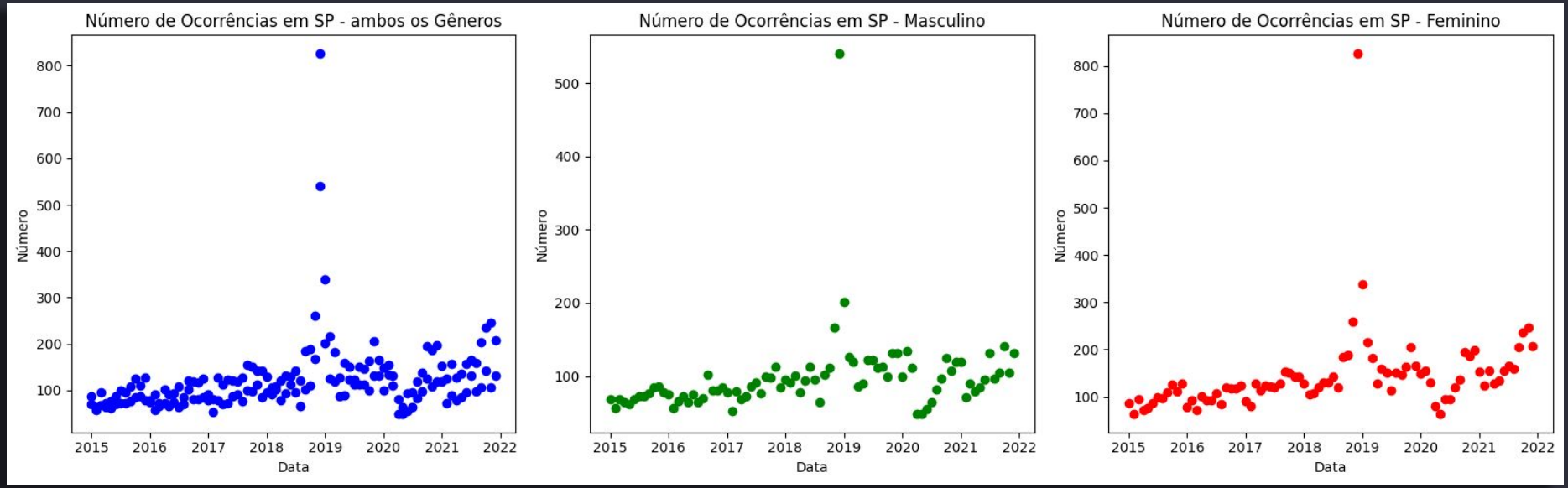
	uf	numero
25	São Paulo	3319
12	Minas Gerais	815
20	Rio de Janeiro	734

Portanto, os Estados a serem considerados com mais número de Casamentos no último ano são: SP, MG e RJ respectivamente!

Análise para SP

Resultados - SP

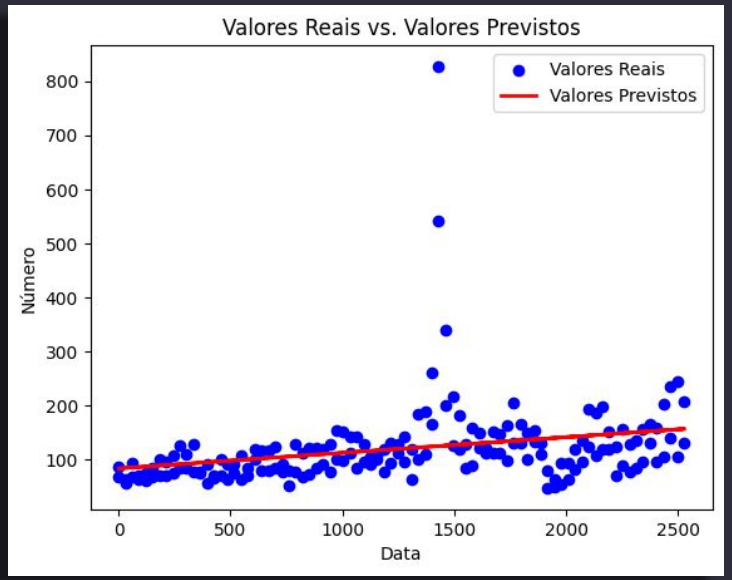
Visão somente de São Paulo - 01/2015 até 12/2021



Em São Paulo nossa análise é mais profunda, pois queremos entender se a regressão linear para ambos os gêneros é semelhante à soma da regressão linear para gêneros distintos (Fem. e Masc.)

Resultados - SP

Para analisarmos a fundo São Paulo, devemos fazer a regressão linear para **ambos** os gêneros e considerar a análise **com Outliers** e **sem Outliers:**

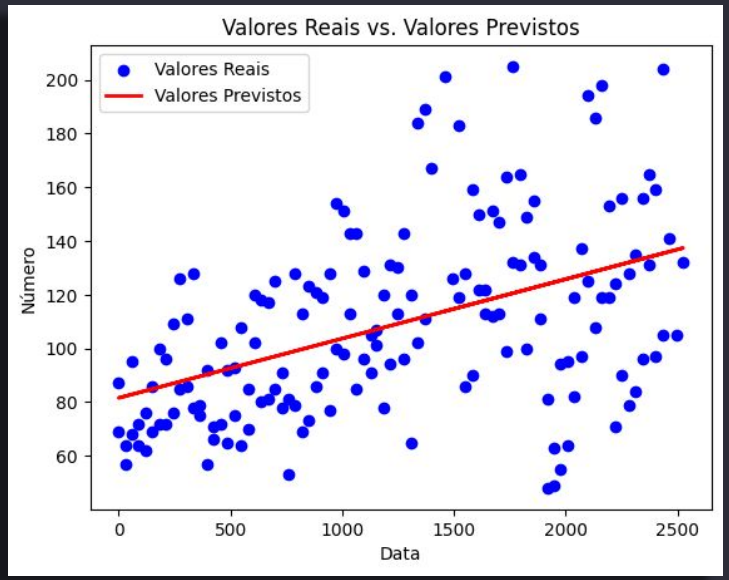


Aplicação do Método de Tukey

Remoção dos Outliers

R^2 com Outlier = 0.076

R^2 sem Outlier = 0.215



OLS Regression Results

Dep. Variable:	numero	R-squared:	0.076
Model:	OLS	Adj. R-squared:	0.070
Method:	Least Squares	F-statistic:	13.63
Date:	Sat, 25 May 2024	Prob (F-statistic):	0.000302
Time:	16:04:17	Log-Likelihood:	-962.50
No. Observations:	168	AIC:	1929.
Df Residuals:	166	BIC:	1935.
Df Model:	1		
Covariance Type:	nonrobust		

Temos um R^2 superior sem outliers >> Portanto nossa reta se ajusta melhor aos dados sem Outliers

OLS Regression Results

Dep. Variable:	numero	R-squared:	0.215
Model:	OLS	Adj. R-squared:	0.210
Method:	Least Squares	F-statistic:	43.22
Date:	Sat, 25 May 2024	Prob (F-statistic):	6.78e-10
Time:	16:04:17	Log-Likelihood:	-777.03
No. Observations:	160	AIC:	1558.
Df Residuals:	158	BIC:	1564.
Df Model:	1		
Covariance Type:	nonrobust		

Resultados - SP

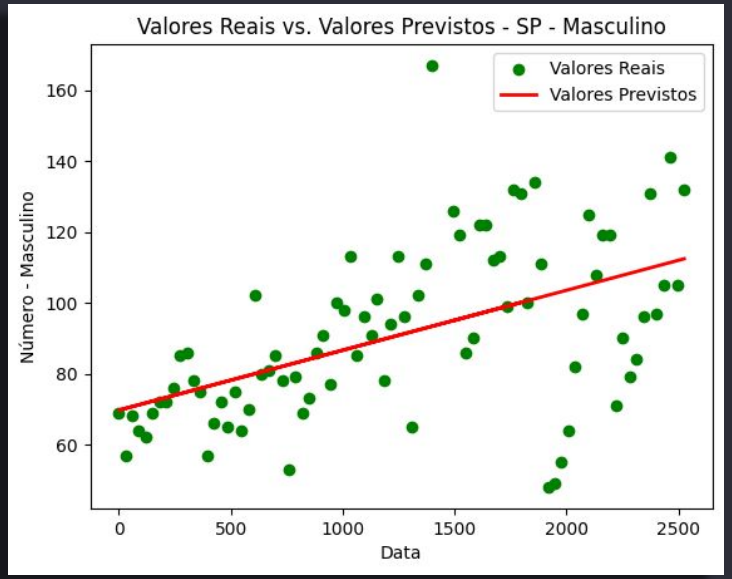
A Biblioteca Stasmodel nos permite achar a equação da reta que melhor se ajusta nos dados, sendo assim, podemos obter a função f(x) para o Estado de SP - Ambos os Gêneros - Sem Outliers (Outliers removidos):

Neste caso a Eq. da Reta é: $y = 81.55 + 0.02x$

Sendo assim, podemos aplicar a equação para dados “futuros” se o modelo foi validado antes

Resultados - SP

Próximo passo é fazer a regressão linear para os Gêneros de forma separada (Ainda do Estado de SP) - Obter as equações da reta:

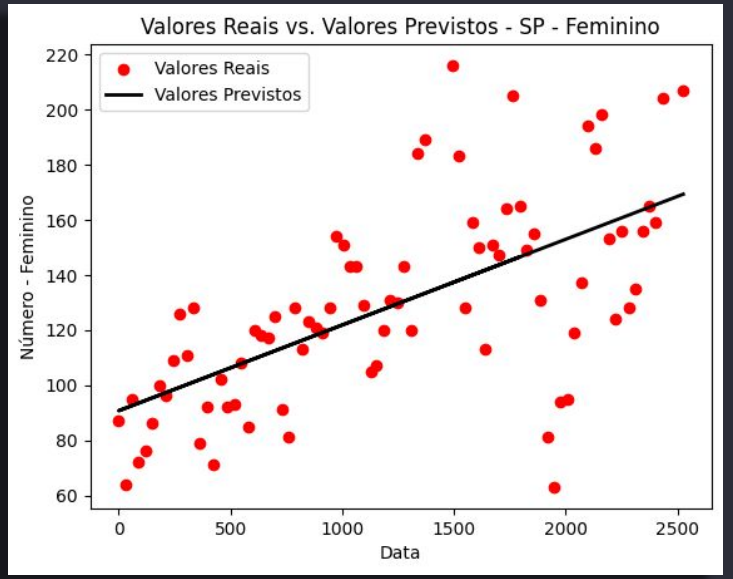


Equação da reta Masculino - SP:

$y = 69.69 + 0.02*x$

Equação da reta Feminino - SP:

$y = 90.67 + 0.03*x$



OLS Regression Results

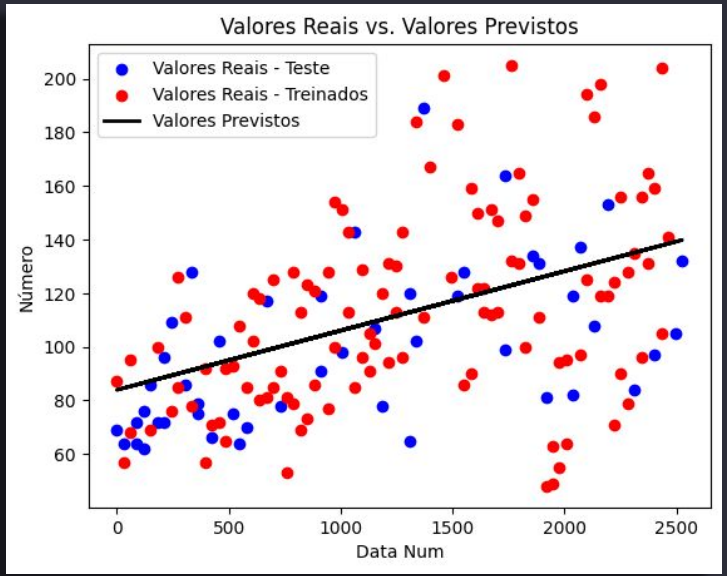
Dep. Variable:	numero	R-squared:	0.274
Model:	OLS	Adj. R-squared:	0.265
Method:	Least Squares	F-statistic:	30.26
Date:	Sat, 25 May 2024	Prob (F-statistic):	4.40e-07
Time:	16:04:17	Log-Likelihood:	-364.16
No. Observations:	82	AIC:	732.3
Df Residuals:	80	BIC:	737.1
Df Model:	1		
Covariance Type:	nonrobust		

OLS Regression Results

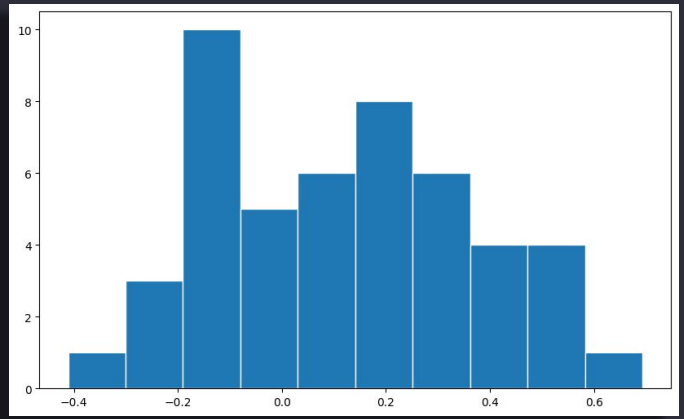
Dep. Variable:	numero	R-squared:	0.394
Model:	OLS	Adj. R-squared:	0.387
Method:	Least Squares	F-statistic:	50.15
Date:	Sat, 25 May 2024	Prob (F-statistic):	5.82e-10
Time:	16:04:18	Log-Likelihood:	-376.21
No. Observations:	79	AIC:	756.4
Df Residuals:	77	BIC:	761.2
Df Model:	1		
Covariance Type:	nonrobust		

Resultados - SP

Próximo passo agora é verificar a assertividade do **Modelo x Dados Reais:**
Separando os dados totais reais em **70% de grupo Treinados** e 30% em Teste para vermos a assertividade do modelo perante aos 30% dos dados separados de forma aleatória. Aqui a ideia é fazer a regressão linear com 70% e verificar os acertos dos dados **30% do grupo Teste**:



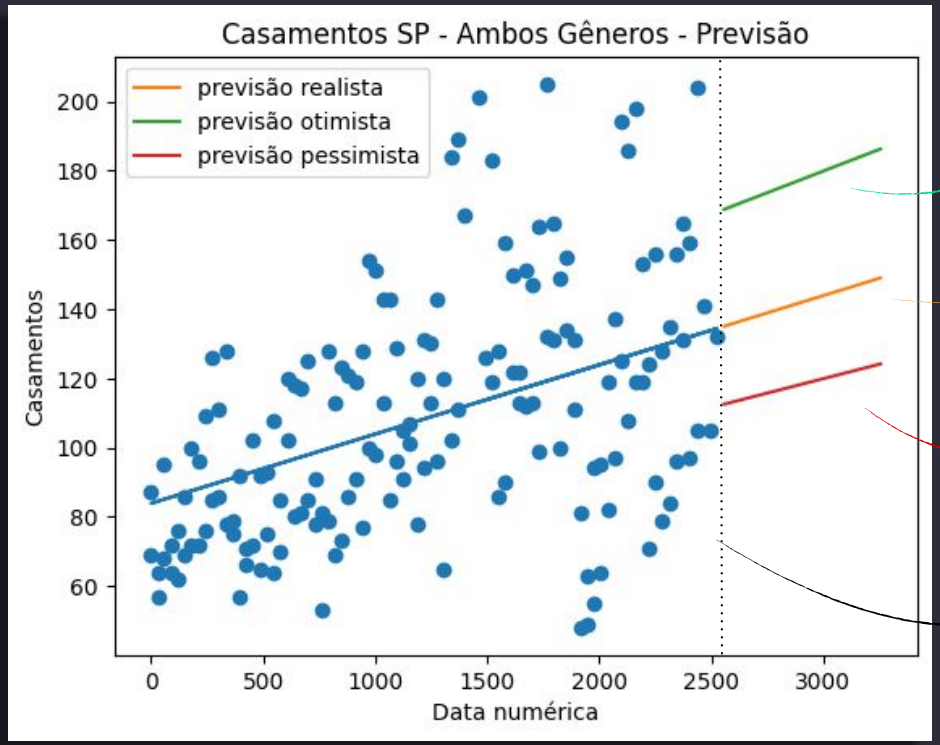
Verificar a distribuição dos Erros do grupo Teste com os Dados Reais (30% da base de SP)



Histograma dos Erros do Modelo X Dados Reais - Maioria dos erros são compreendidos entre **-20% e + 20%** (logo, podemos **criar cenários!**)

Resultados - SP

Resultados Finais da Previsão do Número de Casamentos Homoafetivos em SP - Para **Ambos** os Gêneros - Com Cenários Otimista, Pessimista e Realista com base nos erros do modelo:



+20 Erro do Modelo

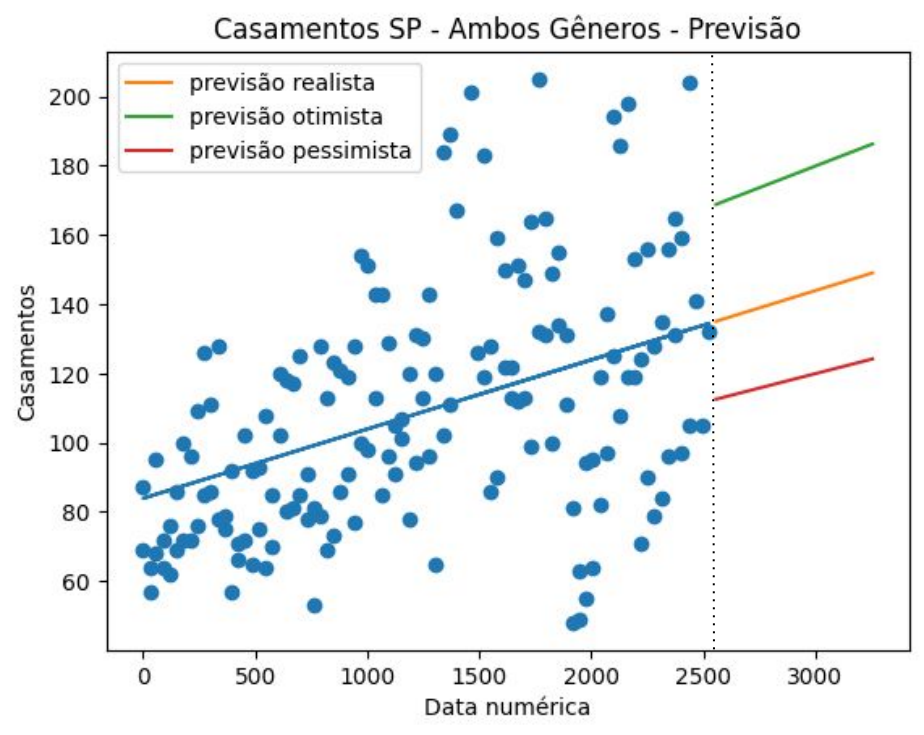
Cenário realista

-20% Erro do modelo

Zona pontilhada - É a zona de dados do Futuro (Previsão)

Resultados - SP

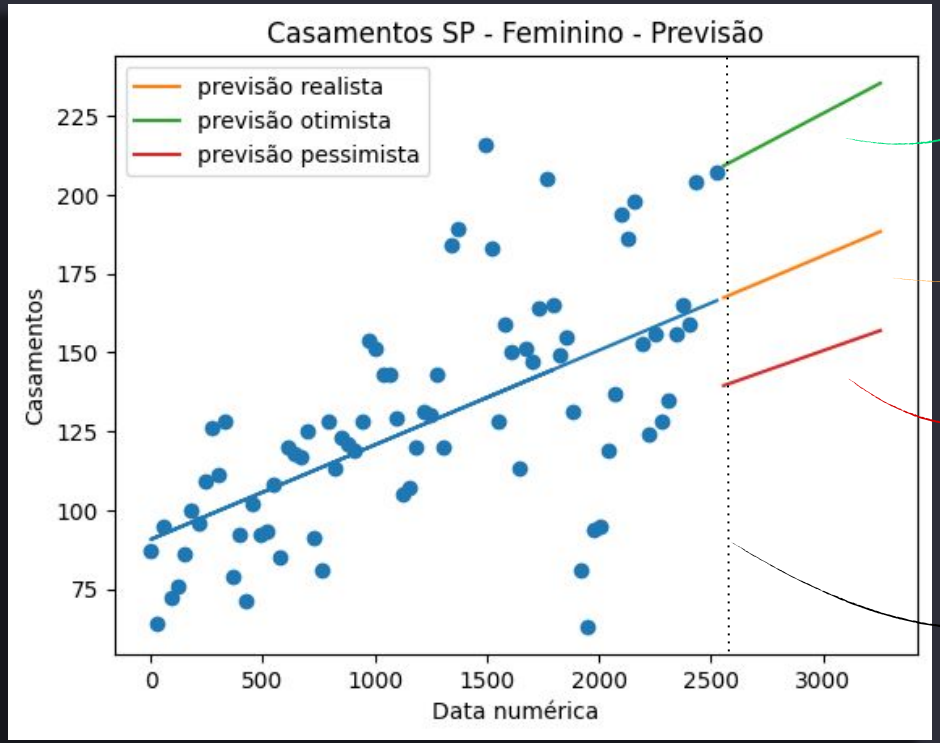
Resultados Finais da Previsão do Número de Casamentos Homoafetivos em SP - Para **Ambos** os Gêneros - Com Cenários Otimista, Pessimista e Realista com base nos erros do modelo:




```
Soma dos Casamentos para 2022 e 2023 no cenário Otimista:
8521
Soma dos Casamentos para 2022 e 2023 no cenário Realista:
6816
Soma dos Casamentos para 2022 e 2023 no cenário Pessimista:
5680
```

Resultados - SP

Os mesmo passos anteriores foram feitos para Feminino - SP:



+20 Erro do Modelo

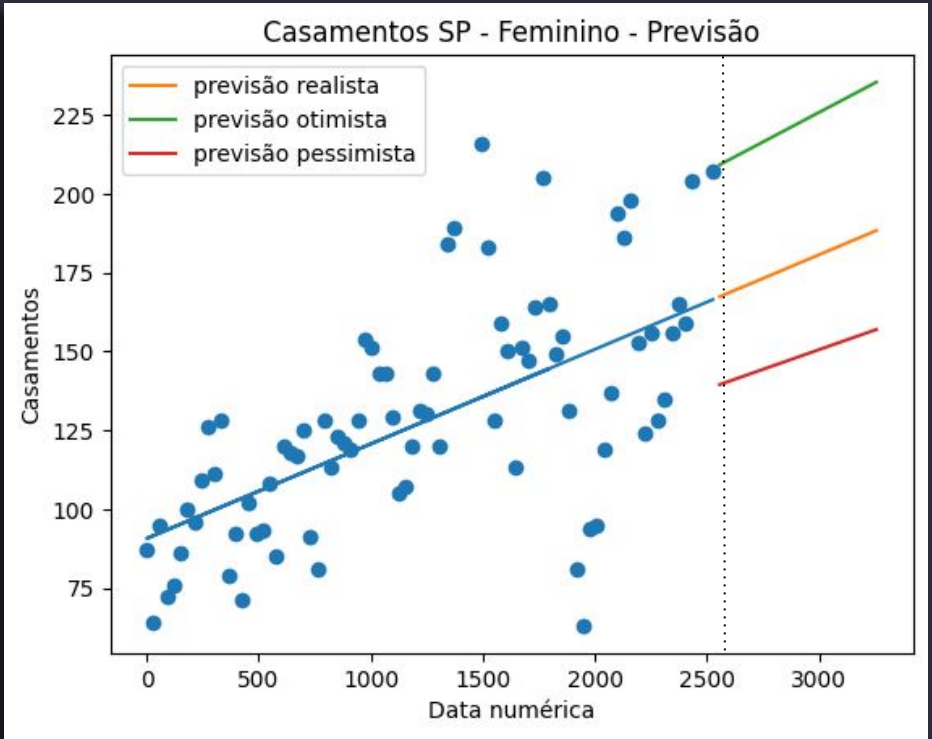
Cenário realista

-20% Erro do modelo

Zona pontilhada - É a zona de dados do Futuro (Previsão)

Resultados - SP

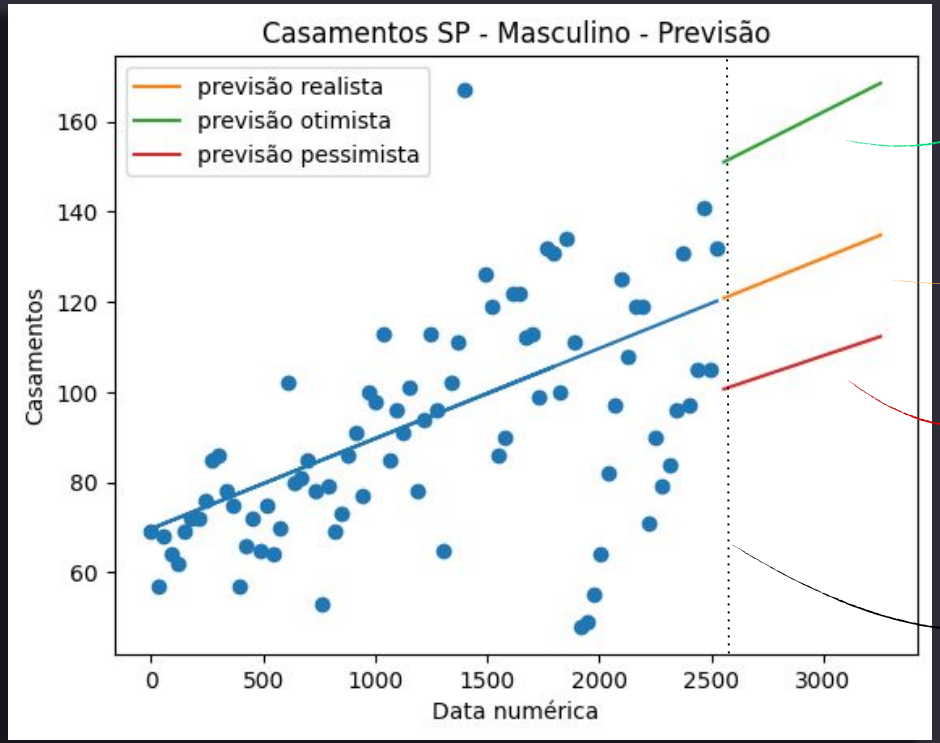
Os mesmo passos anteriores foram feitos para Feminino - SP:




```
Soma dos Casamentos para 2021 e 2022 no cenário Otimista - SP - Feminino:
5335
Soma dos Casamentos para 2021 e 2022 no cenário Realista - SP - Feminino:
4268
Soma dos Casamentos para 2021 e 2022 no cenário Pessimista - SP - Feminino:
3557
```

Resultados - SP

Os mesmo passos anteriores foram feitos para Masculio - SP:



+20 Erro do Modelo

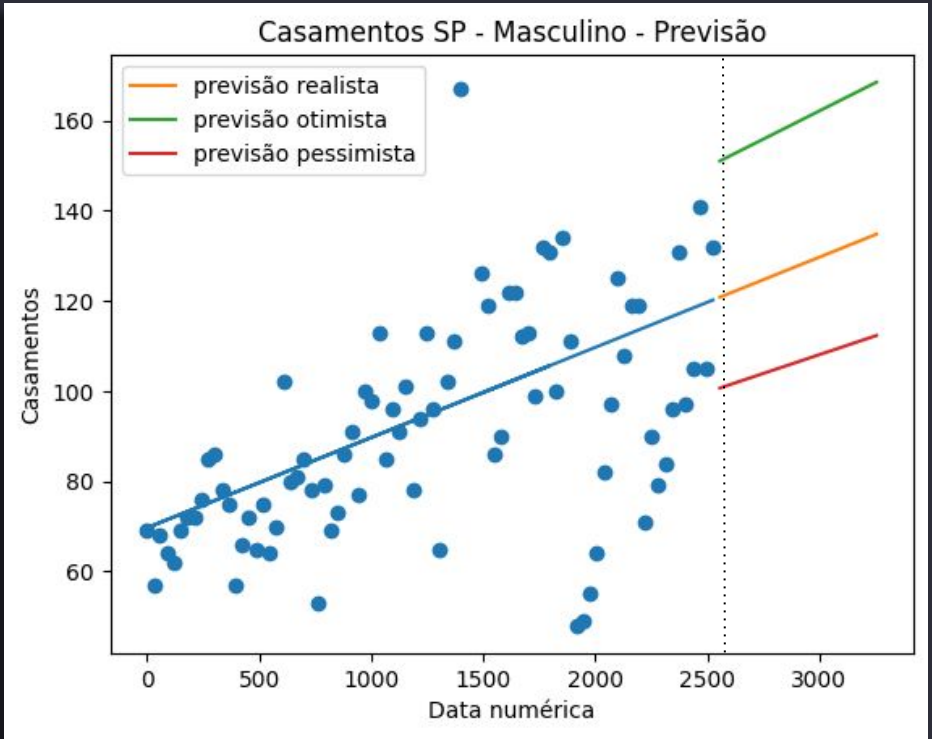
Cenário realista

-20% Erro do modelo

Zona pontilhada - É a zona de dados do Futuro (Previsão)

Resultados - SP

Os mesmo passos anteriores foram feitos para Masculio - SP:




```
Soma dos Casamentos para 2022 e 2023 no cenário Otimista - SP - Masculino:
3834
Soma dos Casamentos para 2022 e 2023 no cenário Realista - SP - Masculino:
3067
Soma dos Casamentos para 2022 e 2023 no cenário Pessimista - SP - Masculino:
2556
```

Resultados - SP

Analisando os resultados:

```
Soma dos Casamentos para 2022 e 2023 no cenário Otimista:
8521
Soma dos Casamentos para 2022 e 2023 no cenário Realista:
6816
Soma dos Casamentos para 2022 e 2023 no cenário Pessimista:
5680
```

```
Soma dos Casamentos para 2021 e 2022 no cenário Otimista - SP - Feminino:
5335
Soma dos Casamentos para 2021 e 2022 no cenário Realista - SP - Feminino:
4268
Soma dos Casamentos para 2021 e 2022 no cenário Pessimista - SP - Feminino:
3557
```

```
Soma dos Casamentos para 2022 e 2023 no cenário Otimista - SP - Masculino:
3834
Soma dos Casamentos para 2022 e 2023 no cenário Realista - SP - Masculino:
3067
Soma dos Casamentos para 2022 e 2023 no cenário Pessimista - SP - Masculino:
2556
```

Foi demonstrado acima que podemos fazer a previsão para ambos os sexos e multiplicar por 2 no final, no caso não precisaríamos fazer cohort por gênero e depois somar os resultados, pois a soma final tem uma diferença de 7% (entre fazer a regressão linear para os dois gêneros (juntos) e multiplicar por 2 ou fazer a regressão linear para os gêneros de forma separada e depois somar os resultados).

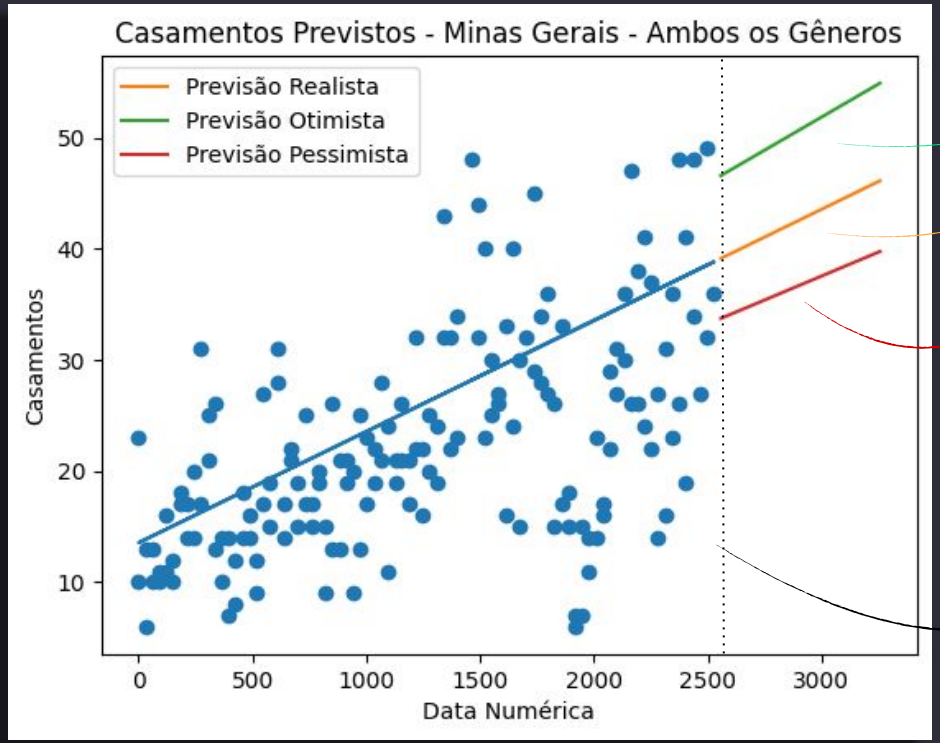
```
Soma dos Casamentos para 2022 e 2023 no cenário Realista - Ambos os Gêneros - SP:
6816
Soma dos Casamentos para 2022 e 2023 no cenário Realista: Soma das previsões de Gênero Masculino e Feminino - Separados - SP:
7335
```

Diferença percentual das duas somas: **-7.07%**

Análise para MG

Resultados - MG

Resultados Finais da Previsão do Número de Casamentos Homoafetivos em MG - Para **Ambos** os Gêneros - Com Cenários Otimista, Pessimista e Realista com base nos erros do modelo: **Os mesmos passos feitos em SP foram feitos em MG, portanto podemos pular os métodos e irmos direto para o Resultado Final**



+16 Erro do Modelo

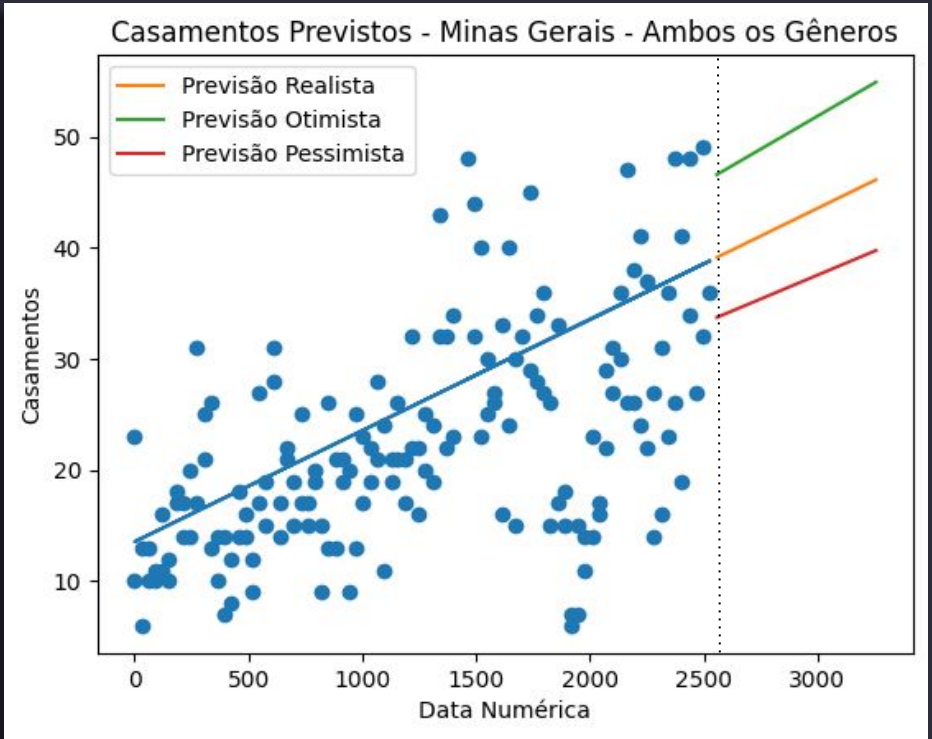
Cenário realista

-16% Erro do modelo

Zona pontilhada - É a zona de dados do Futuro (Previsão)

Resultados - MG

Resultados Finais da Previsão do Número de Casamentos Homoafetivos em MG - Para **Ambos** os Gêneros - Com Cenários Otimista, Pessimista e Realista com base nos erros do modelo: **Os mesmos passos feitos em SP foram feitos em MG, portanto podemos pular os métodos e irmos direto para o Resultado Final**

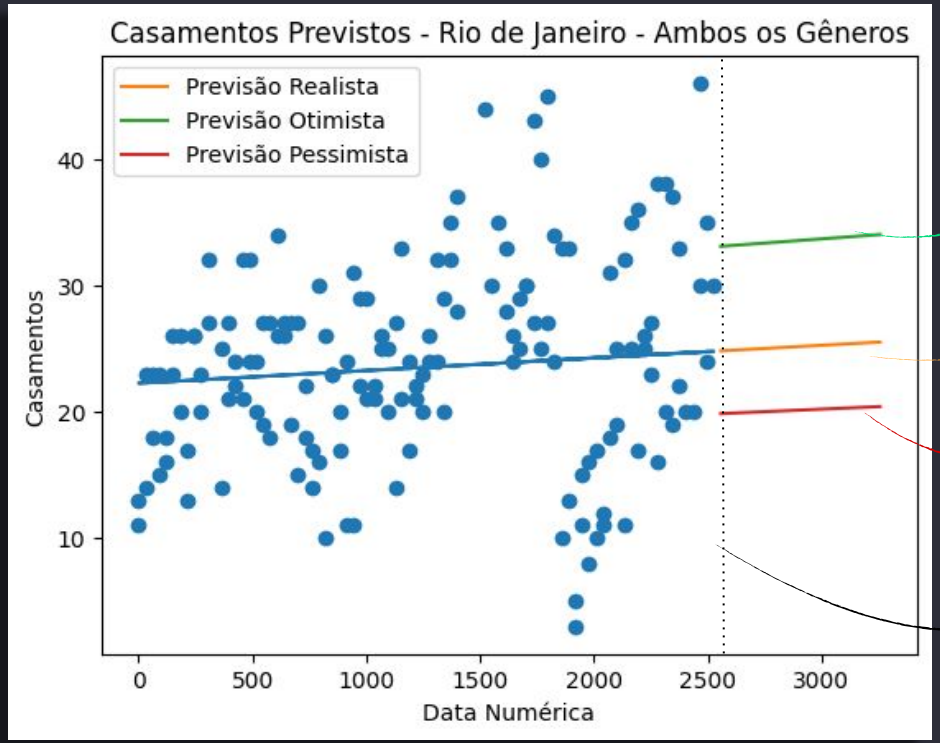



```
Soma dos Casamentos para 2022 e 2023 no cenário Otimista:
2435.0
Soma dos Casamentos para 2022 e 2023 no cenário Realista:
2045.0
Soma dos Casamentos para 2022 e 2023 no cenário Pessimista:
1763.0
```

Análise para RJ

Resultados - RJ

Resultados Finais da Previsão do Número de Casamentos Homoafetivos em RJ - Para **Ambos** os Gêneros - Com Cenários Otimista, Pessimista e Realista com base nos erros do modelo: **Os mesmos passos feitos em SP foram feitos em RJ, portanto podemos pular os métodos e irmos direto para o Resultado Final**



+25 Erro do Modelo

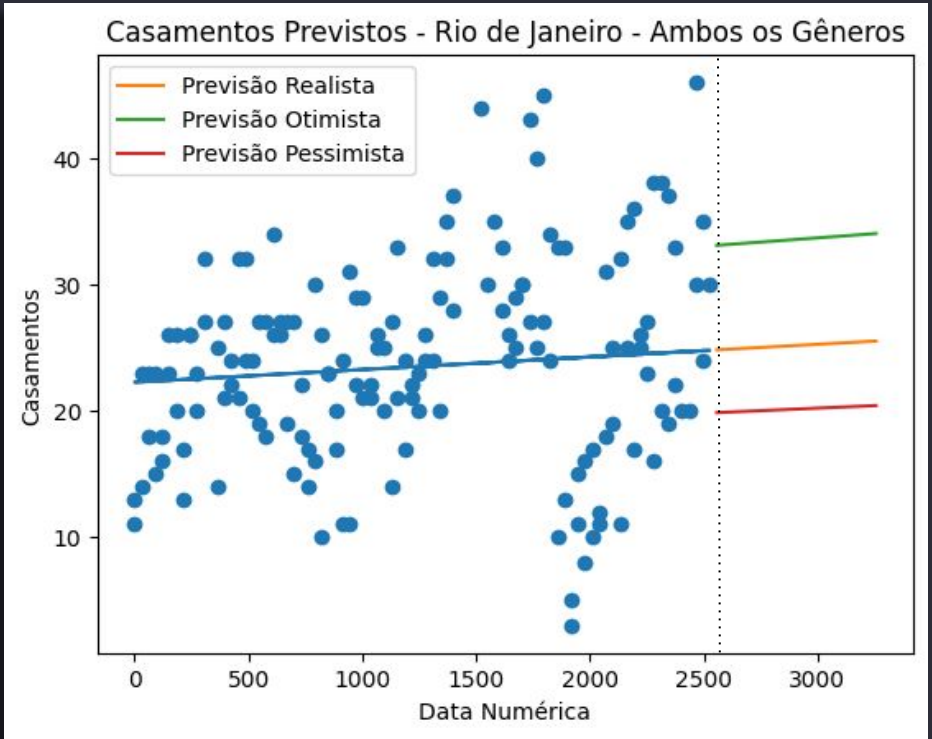
Cenário realista

-25% Erro do modelo

Zona pontilhada - É a zona de dados do Futuro (Previsão)

Resultados - RJ

Resultados Finais da Previsão do Número de Casamentos Homoafetivos em RJ - Para **Ambos** os Gêneros - Com Cenários Otimista, Pessimista e Realista com base nos erros do modelo: **Os mesmos passos feitos em SP foram feitos em RJ, portanto podemos pular os métodos e irmos direto para o Resultado Final**




```
Soma dos Casamentos para 2022 e 2023 no cenário Otimista:
1611.0
Soma dos Casamentos para 2022 e 2023 no cenário Realista:
1208.0
Soma dos Casamentos para 2022 e 2023 no cenário Pessimista:
966.0
```

Conclusão

Conclusão

1 – Foi constatado que o número de previsões das regressões lineares fazendo o cohort por gênero é o mesmo que fazer a regressão com ambos os sexos e multiplicar por 2: - 7% de erro! Isto tudo é analisado por Estado.

2 – Insights:

Nota-se uma tendência de aumento do número de casamentos homoafetivos para ambos os gêneros em SP e MG, logo a tendência é termos uma crescente nos próximos 2 anos (2022 e 2023).

Em RJ nota-se uma tendência de estagnação no número de casamentos ao longo dos próximos anos.

Em SP, mais pessoas do gênero Feminino se casaram e mais pessoas deste mesmo gênero irão casar.

3 – Referências:

Método de Backup por Camadas - Aprendizado em MBA Data Engineering - Santander's Coders 2023 - Ver referências no GitHub
Método de Tukey
Outliers
Stasmodel
Pandas

Obrigado!

brasil .monks