

**CAMPUS
NITERÓI**

ANÁLISE DE DADOS DE ASSALTOS DA CIDADE DO RIO DE JANEIRO

**Discentes
CAUÃ RAMOS COSTA GOMES
GUSTAVO LEAL PARDINI**

**Professora orientadora
SIMONE GAMA**

2025

NITERÓI/RJ

Sumário

1. DIAGNÓSTICO E TEORIZAÇÃO.....	3
1.1. Problemática e/ou problemas identificados.....	4
1.2. Objetivos a serem alcançados	4
2. PLANEJAMENTO E DESENVOLVIMENTO DO PROJETO	4
2.1. Detalhamento técnico do projeto	4
3. ENCERRAMENTO DO PROJETO	5
3.1. Relato de Experiência Individual.....	5
3.1.1 RESULTADOS E DISCUSSÃO	5

1. DIAGNÓSTICO E TEORIZAÇÃO

1.1. Problemática e/ou problemas identificados

O Rio de Janeiro é conhecido por residentes e não residentes como uma das cidades mais perigosas do Brasil para se viver, não só tendo uma das maiores taxas de roubo a cada 100mil habitantes, como também não mostrando uma perspectiva de melhora com o passar dos anos.

1.2. Objetivos a serem alcançados

- Informar a população sobre o número de roubos ao decorrer dos anos.
- Levantar questionamentos sobre o porquê da situação atual.

Com uma maior conscientização da população tende-se a adotar hábitos para combater tais problemas, um planejamento acerca do que pode ser feito para mitigá-los são discutidas, e se tornando um tema de maior destaque no debate público o governo pode dar mais atenção, e implementar soluções para reduzir tal situação.

2. PLANEJAMENTO E DESENVOLVIMENTO DO PROJETO

2.1. Detalhamento técnico do projeto

Análise de Big Data a respeito dos dados de denúncias de roubos no município do RJ entre 2003 e 2023, para obter informações como padrões, taxa de crescimento, e fazer previsões, com a finalidade de informar o povo.

- Primeiro foi criada a sessão do Spark.
- Logo em seguida se iniciou a limpeza dos dados da tabela, renomeando as colunas, removendo espaços vazios, convertendo valores para inteiro, e selecionando os dados utilizados.
- A terceira etapa é a vetorização da coluna "Ano", para ser usada com Pyspark.
- Na sequência é feito o treinamento da regressão linear usando "Ano_Vector" como entrada, e selecionando as colunas que serão utilizadas para o treinamento do modelo.

- Depois disso ocorre a impressão dos dados do coeficiente angular e do intercepto (b) para então converter para o pandas.
- Com essas informações é feito o gráfico para se observar os resultados desejados.
- Na etapa final é feita a análise do desvio padrão e da média para então formar o gráfico de barras.

Para a realização do projeto foram utilizados:

- Tabela provinda do site Data.Rio
- Word para formatação do roteiro de extensão
- Prezi para a criação de slide de apoio
- Colab como IDE
- Linguagem de programação Python
- Bibliotecas Python – Pandas, Pyspark, Matplotlib
- ChatGPT para correções e auxílio no código
- Github para armazenamento do código

3. ENCERRAMENTO DO PROJETO

3.1. Relato de Experiência Individual

Cauã Ramos – apesar de já ter tido contato anterior com Python

anteriormente, meu nível de conhecimento ainda era muito básico, com esse trabalho além do conteúdo aprendido na sala de aula eu tive que ir atrás de videoaulas online, fazer pesquisas e praticar bastante, vários temas foram considerados antes de serem descartados, e vários contratempos ocorreram, mas acredito que consegui aprofundar meus conhecimentos, e que cumprimos com o objetivo estabelecido.

Gustavo Leal - No início, tive algumas dificuldades para entender como aplicar as bibliotecas na prática, mas com o passar das aulas, vídeos e pesquisas, consegui compreender melhor como estruturar o código e interpretar os resultados. Trabalhar com dados reais foi uma experiência interessante, apesar dos desafios e ajustes durante o processo, acredito que o aprendizado foi muito significativo e contribuiu para meu desenvolvimento pessoal.

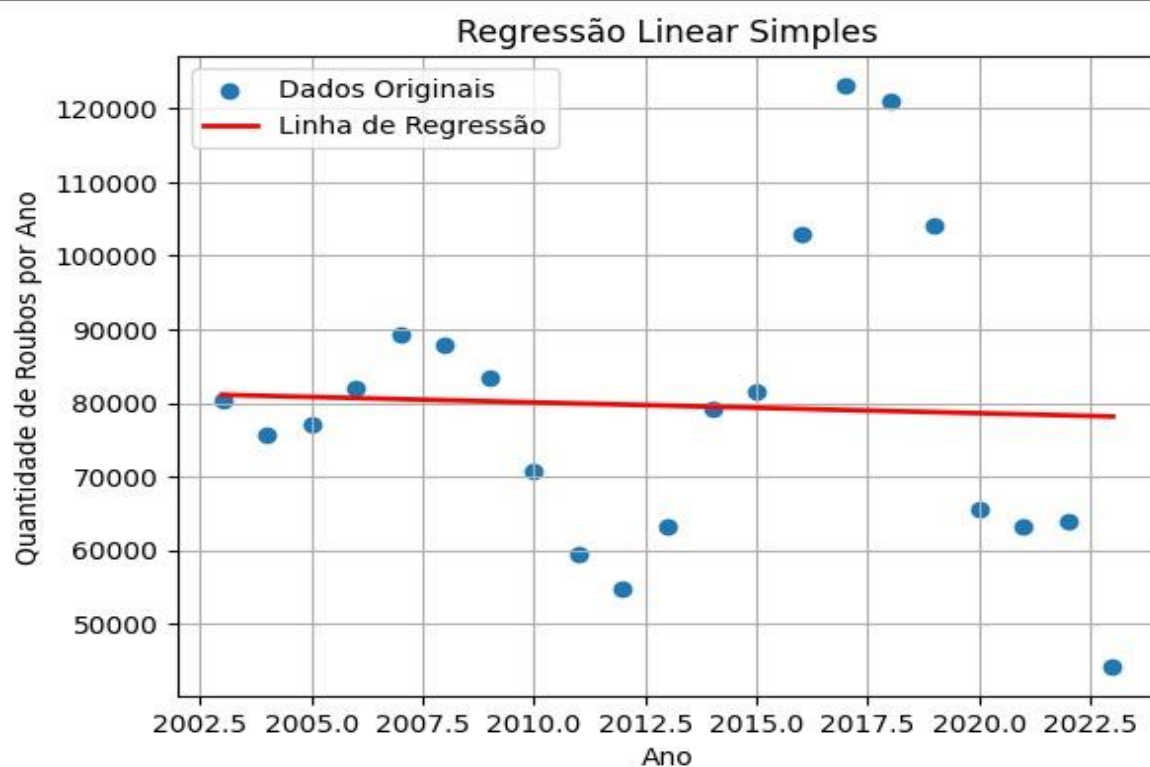
3.1.1 RESULTADOS E DISCUSSÃO

Tabela de Dados

	Ano	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez	Total
2	2003	6909	7130	6579	6739	7374	6343	6354	6262	6604	6863	6423.0	6832.0	80412
3	2004	6630	6168	7252	6716	6694	6449	6433	5520	5699	5853	5931.0	6289.0	75634
4	2005	6044	6158	7035	6880	7031	6579	6290	6495	5846	6312	6167.0	6147.0	76984
5	2006	5952	6301	6494	7109	7504	6667	6596	6671	6744	7488	7393.0	6999.0	81918
6	2007	7161	6658	8417	7532	8008	7417	6768	7457	7246	7539	7379.0	7611.0	89193
7	2008	7813	7201	7137	7011	7112	6961	7502	7406	7313	8079	7070.0	7301.0	87906
8	2009	7512	7558	8272	7361	7724	7160	6908	6426	5996	6064	6158.0	6190.0	83329
9	2010	6199	6114	7190	6405	6253	5829	5791	5617	5607	5671	5165.0	5011.0	70852
10	2011	5221	5126	5437	5055	5044	4751	4959	5041	4963	4711	4474.0	4674.0	59456
11	2012	4505	4966	5078	4810	5087	4556	4519	4526	4046	4380	4111.0	4165.0	54749
12	2013	4482	4549	4897	4852	5230	4838	5009	5638	5450	6026	5925.0	6233.0	63129
13	2014	6859	6539	7017	6796	6677	6084	6558	5970	6084	6760	6846.0	7014.0	79204
14	2015	7740	7304	7341	6712	6793	6181	6156	6100	6410	6567	6856.0	7457.0	81617
15	2016	8191	8523	8124	7759	8240	8446	8233	8117	8571	9503	9908.0	9399.0	103014
16	2017	7240	5423	11976	12087	11801	11148	11811	11110	10271	10115	10249.0	9895.0	123126
17	2018	11192	10820	11076	10188	10441	9779	9306	9677	9466	10017	9987.0	9162.0	121111
18	2019	9806	9536	10413	9626	9570	7781	8461	8361	7216	8102	7757.0	7510.0	104139
19	2020	8410	7916	6537	4158	3747	3920	5167	4729	4761	5055	5502.0	5575.0	65477
20	2021	6111	6007	5913	5182	5214	5021	5355	5123	5153	4753	4576.0	4721.0	63129
21	2022	5015	4811	5599	5432	5257	5272	5383	5340	5379	5394	5546.0	5477.0	63905
22	2023	4552	4369	5119	4565	4981	4193	3676	4157	4113	4501	NaN	NaN	44226

A tabela apresenta registros anuais de roubos no município do Rio de Janeiro entre **2003 e 2023**.

Os valores estão organizados por mês e, ao final de cada linha, há o **Total anual** de ocorrências, que é o dado principal utilizado nas análises estatísticas do projeto.

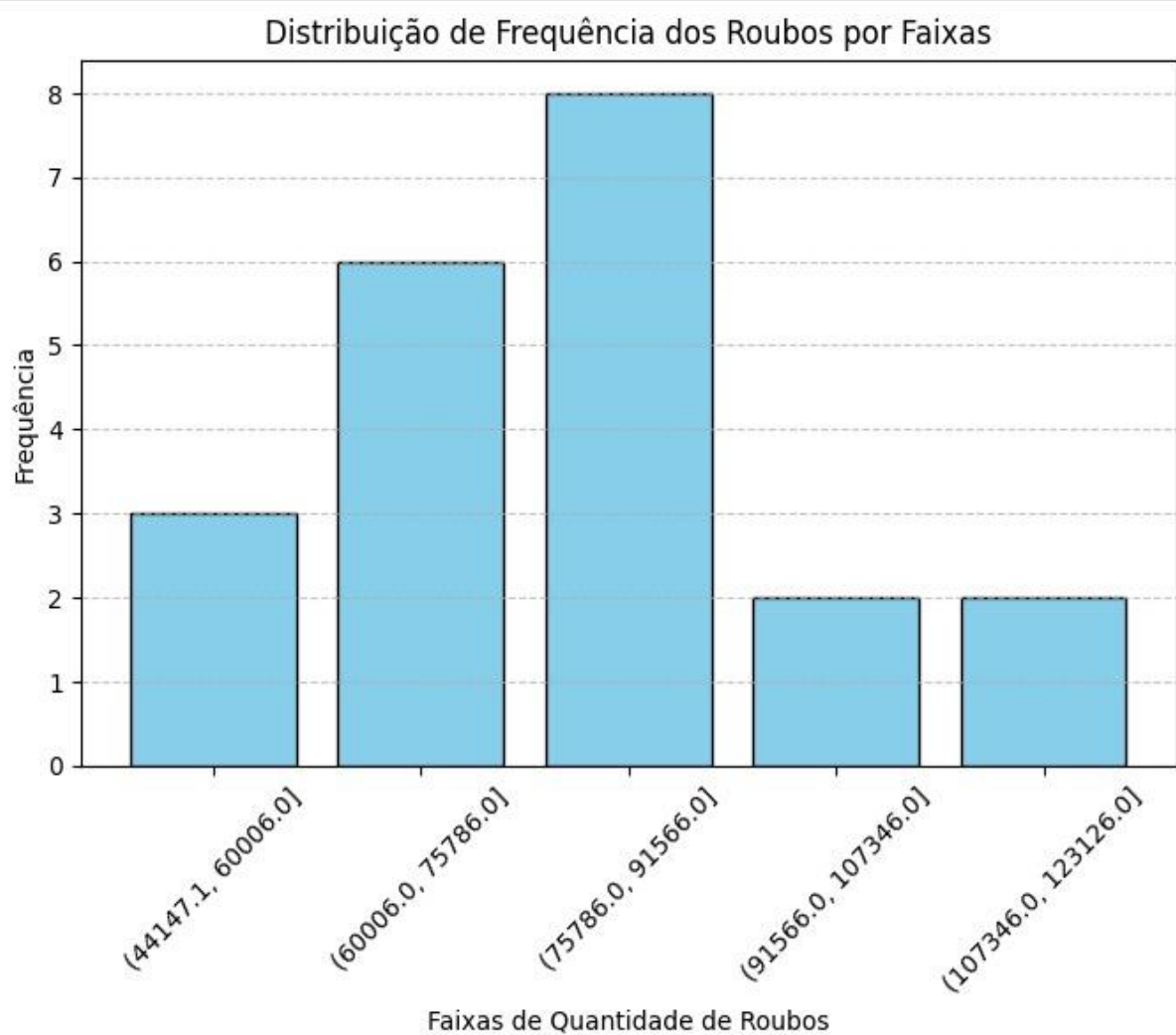


A regressão linear foi utilizada para analisar a relação entre o **tempo (Ano)** e a **quantidade total de roubos (Total2)**.

Variáveis utilizadas

- **Variável Independente (X):** Ano
- **Variável Dependente (Y):** Quantidade anual de roubos

O Objetivo é verificar se existe uma **tendência crescente ou decrescente** ao longo dos anos.



A distribuição de frequência foi calculada para **categorizar os totais de roubos por ano** em faixas e contar quantos anos se enquadram em cada intervalo.

A distribuição mostra:

- A maioria dos anos ficou na faixa **intermediária** (entre 75 mil e 91 mil roubos).
- Poucos anos apresentaram **valores extremamente altos** (acima de 107 mil).
- Poucos anos apresentaram valores **muito baixos** (abaixo de 60 mil).

Link da base de dados utilizada:

<https://www.data.rio/documents/4fb0316436ad474f96bb0079e2256e07/about>