



**FACULDADE CESAR SCHOOL**  
**CURSO DE GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO**

# **Análise de Mobilidade Urbana na Região Metropolitana do Recife**

**Dashboard Interativo e Análise de Dados**

**Antônio Neto**  
**Guilherme Rabelo**  
**Júlia Sales**  
**Thiago Queiroz**

Recife  
Dezembro/2025

# Sumário

<b>1. Introdução</b>	<b>2</b>
<b>2. Histórico de Decisões</b>	<b>5</b>
<b>3. Capturas de Tela e Visualizações</b>	<b>9</b>
<b>4. Síntese dos Principais Insights</b>	<b>17</b>
<b>5. Demonstração de Domínio do Processo</b>	<b>20</b>
<b>6. Conclusões</b>	<b>27</b>

# 1. Introdução

Este projeto, realizado para a cadeira de Projetos 5, visa desenvolver uma solução analítica e visual completa para análise de dados de mobilidade urbana na Região Metropolitana do Recife (RMR). O objetivo é transformar dados brutos da Pesquisa Origem-Destino 2016 em insights acionáveis através de visualizações interativas e modelos preditivos.

## 1.1 Objetivos

- Desenvolver um dashboard interativo para visualização de padrões de mobilidade urbana
- Analisar o comportamento multimodal vs. monomodal dos usuários
- Identificar perfis demográficos dos usuários de transporte público
- Implementar modelos de classificação e regressão para previsões
- Fornecer insights para possíveis soluções de mobilidade urbana

## 1.2 Links Úteis

- Dashboard Online (Streamlit Cloud)
- Repositório GitHub
- Notebook Colab (Análise Original)
- Dataset Original (Google Docs)

## 1.3 Evolução do Projeto

Este projeto seguiu um fluxo de trabalho típico de projetos de ciência de dados, evoluindo de uma análise exploratória inicial para uma aplicação interativa de produção.

### Fase 1: Análise Exploratória no Google Colab

O projeto iniciou com o desenvolvimento de um **notebook Jupyter no Google Colab** (projetos5\_v3.ipynb), disponível em:

[https://colab.research.google.com/github/antonioz2022/Projetos5/blob/main/projetos5\\_v3.ipynb](https://colab.research.google.com/github/antonioz2022/Projetos5/blob/main/projetos5_v3.ipynb)

Nesta fase inicial, o foco estava em:

- **Análise Exploratória de Dados (EDA):** Compreensão da estrutura do dataset, identificação de missing values, análise de distribuições
- **Visualizações preliminares:** Testes de diferentes tipos de gráficos para identificar os mais adequados
- **Prototipagem de modelos:** Experimentação com diferentes algoritmos de machine learning
- **Iteração rápida:** Ambiente colaborativo permitiu testes ágeis e compartilhamento de resultados

**Vantagens do Colab nesta fase:**

- Ambiente pronto para uso (sem necessidade de instalação)
- Acesso gratuito a recursos de computação (GPU/TPU)
- Colaboração em tempo real entre membros da equipe
- Integração direta com Google Drive e GitHub

## Fase 2: Dashboard Interativo em Streamlit

Após a consolidação das análises no Colab, o projeto evoluiu para um **dashboard interativo desenvolvido em Streamlit**, implantado no Streamlit Cloud:

<https://g13-multimodais-rmr.streamlit.app>

Esta migração permitiu:

- **Interatividade:** Usuários podem filtrar dados, selecionar variáveis e explorar diferentes perspectivas em tempo real
- **Acessibilidade:** Dashboard web acessível de qualquer dispositivo, sem necessidade de instalar Python ou bibliotecas
- **Organização modular:** 11 páginas especializadas, cada uma focada em um aspecto da análise
- **Produção:** Aplicação pronta para uso por stakeholders (gestores públicos, pesquisadores, população)
- **Deploy na nuvem:** Hospedagem no Streamlit Cloud permite acesso público 24/7

**Transformações realizadas:**

- Código do notebook refatorado em funções modulares e reutilizáveis

- Visualizações estáticas transformadas em gráficos interativos com Plotly Express
- Adição de widgets de controle (selectbox, radio buttons, sliders)
- Implementação de sistema de navegação por sidebar
- Containerização com Docker para reprodutibilidade

## Benefícios desta Abordagem

O workflow **Colab** → **Streamlit** combina o melhor de dois mundos:

- **Exploração ágil:** Colab permite testar hipóteses rapidamente sem preocupação com infraestrutura
- **Produção robusta:** Streamlit oferece uma aplicação estável e escalável para usuários finais
- **Documentação viva:** Notebook serve como documentação técnica detalhada do processo analítico
- **Democratização:** Dashboard torna insights acessíveis para públicos não-técnicos

Esta metodologia reflete as melhores práticas da indústria de ciência de dados, onde a análise exploratória em notebooks é seguida pela construção de produtos analíticos em frameworks especializados.

## 2. Histórico de Decisões

Esta seção documenta as principais decisões técnicas e analíticas tomadas durante o desenvolvimento do projeto.

### 2.1 Escolha da Plataforma de Visualização

**Decisão:** Utilizar Streamlit como framework principal para o dashboard.

**Justificativa:**

- **Facilidade de desenvolvimento:** Streamlit permite criar dashboards interativos com código Python puro, sem necessidade de HTML/CSS/JavaScript
- **Prototipagem rápida:** Ideal para iterações rápidas durante o desenvolvimento
- **Deploy gratuito:** Streamlit Cloud oferece hospedagem gratuita com integração direta ao GitHub
- **Interatividade nativa:** Widgets interativos (sliders, selectbox, radio buttons) já incluídos
- **Integração com bibliotecas científicas:** Compatibilidade total com Pandas, Plotly, Matplotlib

### 2.2 Biblioteca de Visualização

**Decisão:** Utilizar Plotly Express como biblioteca principal, com Matplotlib/Seaborn para gráficos específicos.

**Justificativa:**

- **Interatividade:** Plotly oferece zoom, pan, hover tooltips nativamente
- **Estética moderna:** Gráficos visualmente atraentes com configuração mínima
- **Responsividade:** Adaptação automática a diferentes tamanhos de tela
- **Documentação:** Excelente documentação e comunidade ativa

### 2.3 Estrutura de Navegação

**Decisão:** Organizar o dashboard em páginas temáticas acessadas por menu lateral (sidebar).

**Justificativa:**

- **Organização lógica:** Cada página representa uma etapa da análise ou um tema específico
- **Facilita navegação:** Usuário pode ir direto ao conteúdo de interesse
- **Performance:** Carrega apenas a página selecionada, reduzindo tempo de renderização
- **Escalabilidade:** Fácil adicionar novas análises sem sobrecarregar interface

#### **Estrutura final:**

1. Visão Geral
2. Estatísticas Descritivas
3. Tipo de Trajeto (Monomodal vs Multimodal)
4. Modal Share
5. Análise por Localização
6. Integração Multimodal
7. Perfil de Usuários de Integração
8. Perfil Demográfico
9. Modelos de Regressão
10. Modelos de Classificação
11. Conclusões

## **2.4 Escolha de Tipos de Gráficos**

### **Gráficos de Barras**

**Uso:** Distribuições categóricas (sexo, faixa etária, renda, modais).

#### **Justificativa:**

- Fácil comparação entre categorias
- Leitura imediata de valores absolutos
- Amplamente compreendido pelo público geral

## Gráficos de Pizza (Donut)

**Uso:** Proporções totais (distribuição por sexo, tipo de trajeto).

**Justificativa:**

- Visualização clara de proporções
- Bom para categorias com poucas divisões (2-5 categorias)
- Formato donut deixa espaço para métricas centrais

## Heatmaps

**Uso:** Correlação entre bairros e modais, matrizes de confusão.

**Justificativa:**

- Identifica padrões em dados multidimensionais
- Cores facilitam identificação de valores extremos
- Compacto para grandes volumes de dados

## Gráficos de Dispersão

**Uso:** Regressão linear (número de modais vs. outras variáveis).

**Justificativa:**

- Mostra relação entre variáveis numéricas
- Permite identificar outliers
- Facilita visualização de linha de tendência

## 2.5 Paleta de Cores

**Decisão:** Utilizar paletas de cores semanticamente significativas e acessíveis.

**Diretrizes aplicadas:**

- **Sexo:** Azul e laranja (evitando estereótipos rosa/azul tradicionais)
- **Idade:** Gradiente viridis (sequencial, do claro ao escuro)
- **Renda:** Gradiente magma (quente, indicando progressão)
- **Modais:** Cores distintas para cada modal (ônibus: laranja, metrô: azul, etc.)
- **Multimodal vs Monomodal:** Verde e roxo (cores contrastantes)



## 2.6 Tratamento de Dados

**Principais decisões de limpeza e processamento:**

- **Valores ausentes:** Categorizar como "Não declarado" ou "Sem resposta" em vez de remover
- **Modais múltiplos:** Converter strings separadas por vírgula em listas para análise
- **Integração formal:** Identificar usuários por combinação de campos (flag + terminal escrito)
- **Categorização de trajetos:** Classificar como monomodal (1 modal), multimodal (2+) ou sem resposta
- **Ordenação de faixas:** Manter ordem lógica (idade, renda) em visualizações

## 2.7 Containerização com Docker

**Decisão:** Utilizar Docker para padronizar ambiente de desenvolvimento e deploy local.

**Justificativa:**

- **Reprodutibilidade:** Garante que o ambiente funciona igual em qualquer máquina
- **Isolamento:** Evita conflitos de dependências com outros projetos
- **Portabilidade:** Fácil compartilhar com equipe ou mover entre servidores
- **Produção-ready:** Mesma imagem pode ser usada em produção

**Configuração:**

- Base image: `python:3.11-slim`
- Porta exposta: 8501 (padrão do Streamlit)
- Volume mount: `./dados` para acesso ao dataset

## 3. Capturas de Tela e Visualizações

### 3.1 Página Inicial - Visão Geral

**Descrição:** A página inicial apresenta KPIs principais e distribuições básicas da população pesquisada.

#### Elementos visuais:


- 4 KPIs principais:

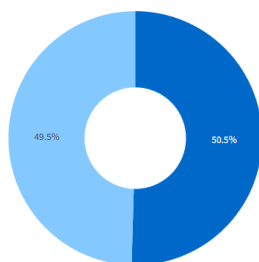
- Trabalham: percentual da população que trabalha
- Estudam: percentual da população que estuda
- Multimodal: número absoluto de usuários multimodais
- Média Modais: média de modais utilizados por pessoa


- Gráfico de pizza: distribuição por sexo
- Gráfico de barras: distribuição por faixa etária
- Gráfico de barras: distribuição por renda

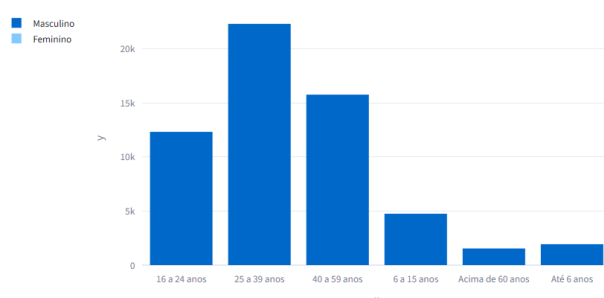
#### Insights visuais:

- 71% da população trabalha
- 40,5% estuda
- 13.921 pessoas usam múltiplos modais
- Média de 1,23 modais por pessoa

 Distribuição por Sexo



 Distribuição por Faixa Etária



Visão Geral - distribuições demográficas

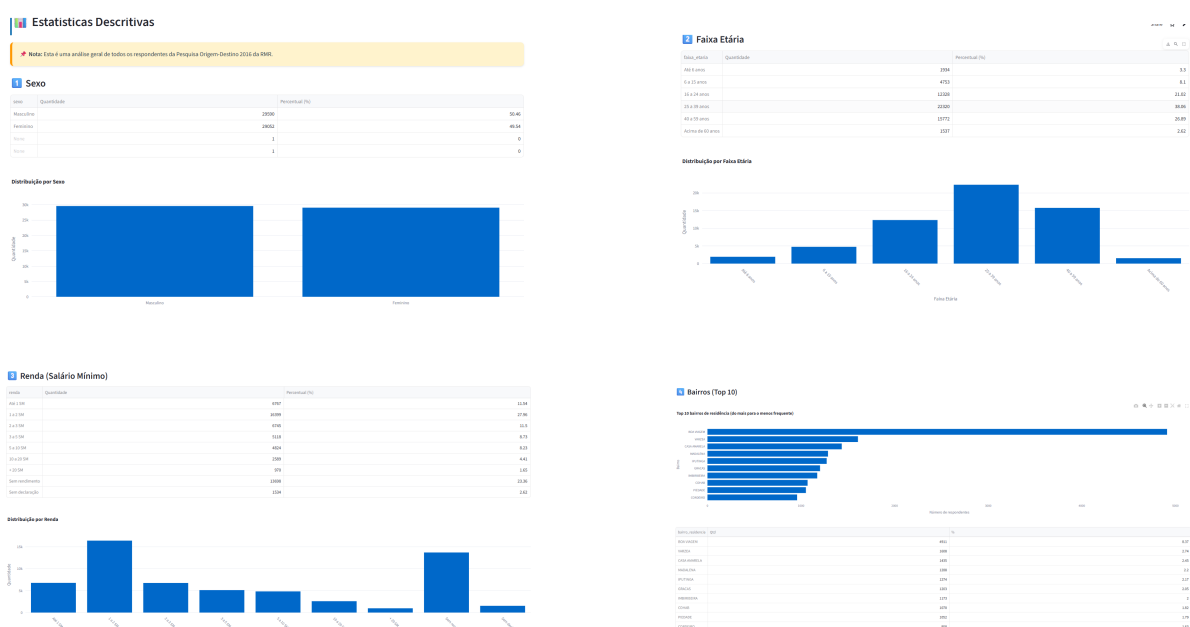
## 3.2 Estatísticas Descritivas

**Descrição:** Análise detalhada das características demográficas da população pesquisada.

**Elementos visuais:**

- Distribuição por sexo (gráfico de barras horizontal)
- Distribuição por idade (gráfico de barras vertical)
- Distribuição por renda (gráfico de barras com 9 categorias)
- Top 10 bairros mais frequentes (gráfico de barras horizontal ordenado)

**Nota importante:** Box destacado informando que "Esta é uma análise geral de quem respondeu a pesquisa", esclarecendo que os dados representam a amostra da pesquisa OD 2016.



Estatísticas Descritivas - Distribuições demográficas da população pesquisada

## 3.3 Tipo de Trajeto (Monomodal vs Multimodal)

**Descrição:** Comparação entre usuários que usam um único modal vs. múltiplos modais.

**Elementos visuais:**

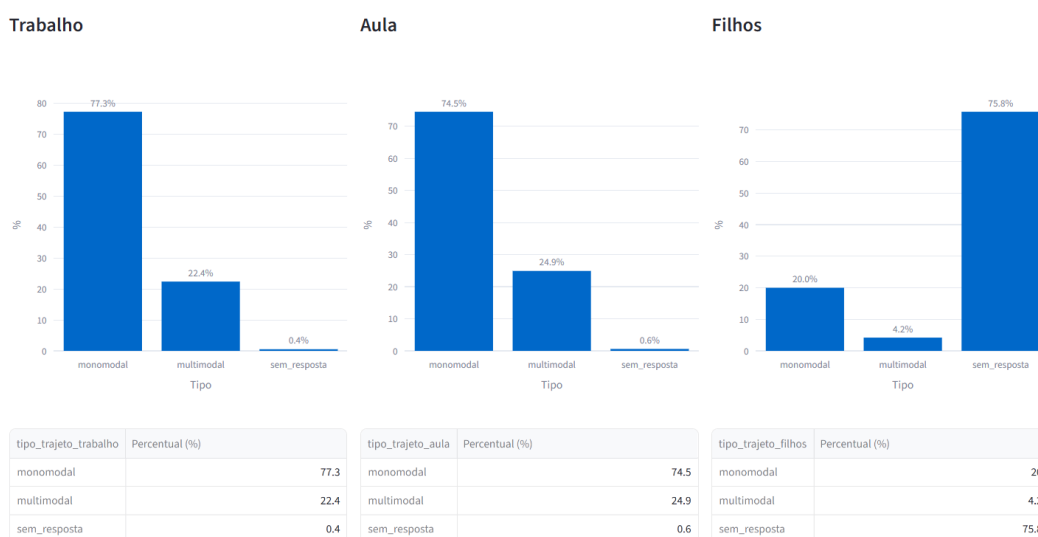
- 4 KPIs principais:
  - Usuários Multimodais: 13.921 pessoas
  - % Multimodais: 23,7% da população
  - Usuários Monomodais: 45.553 pessoas

– Máx. de Modais Usado: 2 modais

- 3 gráficos de barras lado a lado: distribuição para trabalho, aula e filhos
- Tabelas com contagens absolutas e percentuais

#### Insights principais:

- Maioria das viagens é monomodal (77% usam apenas 1 modal)
- 13.921 usuários (23,7%) são multimodais
- Multimodalidade é mais comum em viagens ao trabalho
- Viagens escolares tendem a ser mais simples (monomodais)



#### Tipo de Trajeto - Comparação Monomodal vs Multimodal

## 3.4 Modal Share

**Descrição:** Participação de cada modal no total de viagens.

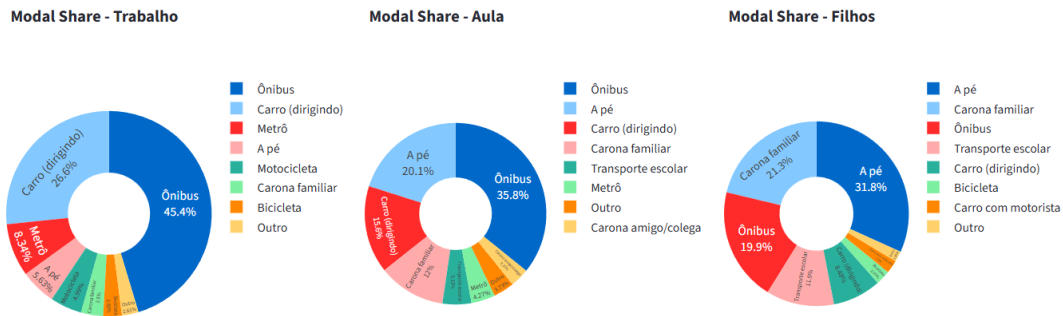
#### Elementos visuais:

- 3 gráficos de pizza (donut): modal share para trabalho, aula e filhos
- Filtros interativos por sexo e faixa de renda
- Top 10 modais mais utilizados em cada categoria

#### Insights principais:

- Ônibus é o modal mais utilizado para ir ao trabalho ou a aula (aproximadamente 40-50%)

- "A pé" é o segundo mais comum no geral, especialmente para trajetos escolares
- Carro particular tem participação significativa em faixas de renda mais altas



Modal Share - Participação de cada modal por contexto

### 3.5 Análise por Localização

**Descrição:** Heatmap mostrando a relação entre bairros de residência e modais utilizados para trabalho.

**Elementos visuais:**

- Heatmap grande: top 20 bairros vs. todos os modais
- Escala de cores quente (amarelo-laranja-vermelho) indicando intensidade de uso

**Insights principais:**

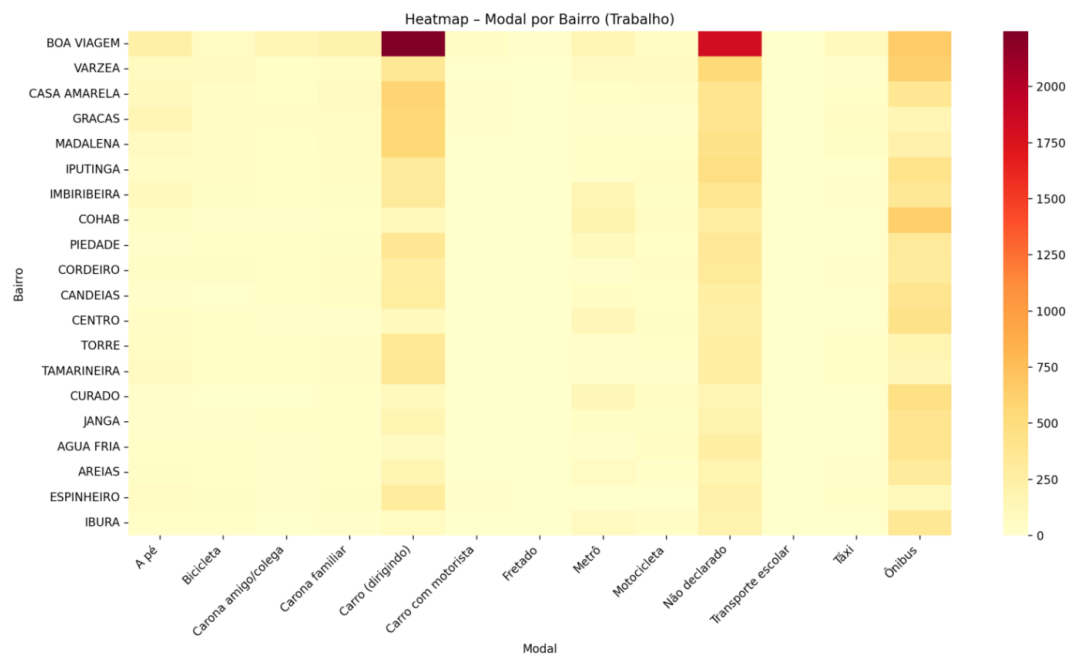
- Bairros periféricos concentram maior uso de ônibus
- Bairros centrais apresentam maior diversidade de modais
- Alguns bairros têm forte dependência de um único modal

### 3.6 Integração Multimodal

**Descrição:** Análise das combinações de modais em viagens multimodais.

**Elementos visuais:**

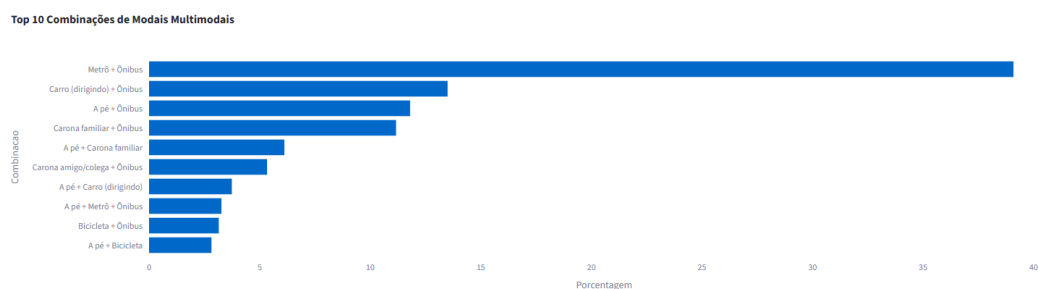
- Tabela: top 10 combinações de modais mais frequentes
- Gráfico de barras horizontal: visualização das combinações
- Percentual de cada combinação sobre o total



Análise por Localização - Heatmap de bairros vs modais

### Insights principais:

- "Ônibus + Metrô" é a combinação mais comum, indicando uso de integração terminal
- "Ônibus + Carro (Dirigido)" é frequente, indicando que pessoas que esporadicamente usam transporte público
- Combinações com 3+ modais são raras



Integração Multimodal - Top 10 combinações de modais

## 3.7 Perfil de Usuários de Integração

**Descrição:** Análise demográfica específica dos usuários que utilizam terminais de integração formal.


### Elementos visuais:

- 3 KPIs principais:

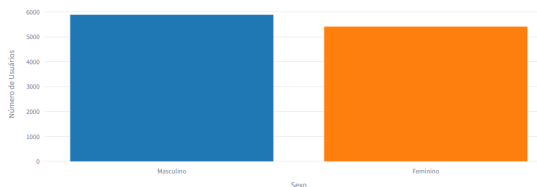
- Total de Usuários: 11.299 pessoas
- % da População: 19,3% da amostra pesquisada
- Sexo Predominante: Masculino
- Gráfico de barras: distribuição por sexo (com estatísticas laterais)
- Gráfico de barras: distribuição por faixa etária (com estatísticas laterais)
- Gráfico de barras: distribuição por renda (com estatísticas laterais)
- Box de insights: principais conclusões sobre o perfil


### Insights principais:

- 11.299 usuários (19,3%) utilizam terminais de integração formal
- Sexo masculino é ligeiramente predominante
- Predominância de faixas etárias economicamente ativas (25-59 anos)
- Concentração em faixas de renda baixa (até 3 salários mínimos)
- Sistema de integração é crucial para mobilidade de populações periféricas

 Distribuição por Sexo

Usuários de Integração por Sexo



 Distribuição por Faixa Etária

Usuários de Integração por Faixa Etária



Perfil de Usuários de Integração - Distribuições demográficas

## 3.8 Perfil Demográfico

**Descrição:** Cruzamento entre características demográficas e escolha de modais.

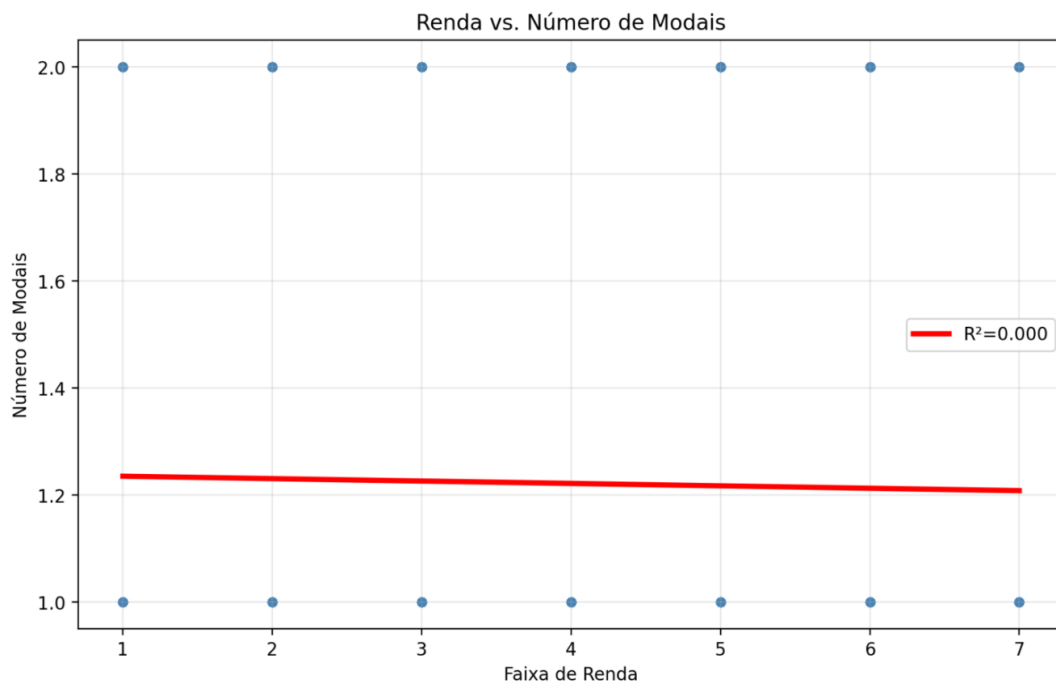
### Elementos visuais:

- Tabela de crosstab: modal vs. sexo (valores percentuais)
- Gráfico de barras empilhadas: participação modal por sexo
- Tabela de crosstab: modal vs. renda
- Gráfico de barras empilhadas: impacto da renda na escolha modal

### Insights principais:

- [illegible]





Modelo de Regressão - Predição de número de modais

- Curvas ROC com AUC

### Resultados:

- Melhor modelo: Random Forest (acurácia: XX%)
- Modelos têm dificuldade em prever classe multimodal (desbalanceamento)
- Recomendação: usar técnicas de balanceamento (SMOTE, class weights)

### Comparação de Métricas

	Modelo	Acurácia	Precisão	Recall	F1-Score
0	Regressão Logística	0.7812	0.5945	0.1283	0.2110
1	Decision Tree	0.8002	0.6265	0.3066	0.4118
2	Random Forest	0.7992	0.6150	0.3192	0.4203

### Comparação de Desempenho dos Modelos



Modelos de Classificação - Comparação de performance

## 4. Síntese dos Principais Insights

Esta seção consolida os principais achados da análise de dados de mobilidade urbana na RMR.

### 4.1 Perfil Geral da População

- **População economicamente ativa:** 71% da amostra trabalha, indicando foco em mobilidade para trabalho
- **População estudantil significativa:** 40,5% estuda, representando demanda por transporte escolar
- **Distribuição etária:** Concentração nas faixas de 25-59 anos (população economicamente ativa)
- **Distribuição por renda:** Predominância de faixas de baixa renda (até 3 SM), com implicações para políticas de transporte público

### 4.2 Padrões de Multimodalidade

- **Predominância monomodal:** Maioria das viagens utiliza apenas 1 modal (simplicidade preferida)
- **Multimodalidade no trabalho:** Viagens ao trabalho apresentam maior taxa de multimodalidade (13.921 usuários)
- **Média de modais:** 1,23 modais por pessoa, indicando que a multimodalidade ainda é exceção
- **Combinações mais comuns:** "Ônibus + A pé" e "Ônibus + Metrô" dominam as viagens multimodais

### 4.3 Modal Share e Dependência de Transporte Público

- **Ônibus é rei:** Representa 40-50% de todas as viagens, evidenciando dependência crítica do sistema
- **Transporte ativo significativo:** "A pé" é o segundo modal mais usado, especialmente em curtas distâncias
- **Baixa penetração de metrô:** Apesar de eficiente, o metrô tem alcance geográfico limitado
- **Carro particular:** Concentrado em faixas de renda alta (acima de 5 SM)

## 4.4 Perfil dos Usuários de Integração

- **Perfil socioeconômico:** Predominância de usuários de baixa renda (até 3 SM), para quem a integração é essencial
- **Faixa etária ativa:** Concentração em 25-59 anos, trabalhadores que dependem de integração para chegar ao emprego
- **Importância do sistema:** Terminais de integração são cruciais para viabilizar mobilidade de populações periféricas
- **Distribuição de gênero:** Relativamente equilibrada, indicando que integração é necessidade universal

## 4.5 Impacto da Renda na Mobilidade

- **Forte correlação renda-modal:** Quanto maior a renda, maior o uso de carro próprio
- **Dependência do transporte público:** Faixas de renda até 3 SM dependem fortemente de ônibus/metrô
- **Desigualdade de acesso:** Populações de baixa renda têm menos opções de modais, limitando mobilidade
- **Implicação para políticas públicas:** Necessidade de subsídios e investimento em transporte público de qualidade

## 4.6 Diferenças de Gênero na Mobilidade

- **Mulheres usam mais transporte público:** Maior dependência de ônibus e metrô
- **Homens usam mais veículos motorizados individuais:** Maior participação de carro e moto
- **Segurança como fator:** Diferenças podem refletir preocupações com segurança e infraestrutura
- **Necessidade de políticas de gênero:** Transporte público deve considerar necessidades específicas (iluminação, segurança, acessibilidade)

## 4.7 Padrões Geográficos

- **Bairros periféricos:** Alta dependência de ônibus, refletindo menor oferta de modais
- **Bairros centrais:** Maior diversidade de modais, com presença de metrô e maior acesso a serviços

- **Desigualdade espacial:** Distribuição desigual de infraestrutura de transporte entre centro e periferia
- **Necessidade de expansão:** Investimento em transporte de massa em áreas periféricas é prioritário

## 4.8 Limitações dos Modelos Preditivos

- **Desbalanceamento de classes:** Multimodalidade é classe minoritária, dificultando predição
- **$R^2$  moderado:** Modelos de regressão têm capacidade limitada de prever número exato de modais
- **Random Forest como melhor opção:** Supera modelos lineares, mas ainda com espaço para melhoria
- **Necessidade de mais features:** Variáveis adicionais (distância, tempo, custo) poderiam melhorar modelos

## 4.9 Recomendações para Políticas Públicas

1. **Expandir sistema de integração:** Construir mais terminais e facilitar baldeações gratuitas
2. **Investir em transporte de massa:** Expandir metrô e BRT para áreas periféricas
3. **Subsidiar tarifas para baixa renda:** Vale-transporte e tarifas sociais são essenciais
4. **Melhorar infraestrutura para caminhada:** Calçadas e iluminação para incentivar transporte ativo
5. **Segurança para mulheres:** Políticas específicas de segurança em transporte público
6. **Integração tarifária completa:** Tarifa única para ônibus + metrô reduziria custo para usuários
7. **Dados em tempo real:** Sistemas de informação para melhorar experiência do usuário

## 5. Demonstração de Domínio do Processo

Esta seção demonstra o domínio completo do processo de análise e visualização de dados, desde a coleta até a apresentação de resultados.

### 5.1 Etapas do Processo de Análise

#### Etapa 1: Coleta e Compreensão dos Dados

**Fonte:** Pesquisa Origem-Destino 2016 da Região Metropolitana do Recife.

##### **Características do dataset:**

- 58.644 registros (respondentes)
- 50+ variáveis (demográficas, socioeconômicas, padrões de viagem)
- Dados mistos: numéricos, categóricos e texto livre
- Valores ausentes significativos em algumas colunas

##### **Compreensão do contexto:**

- Pesquisa domiciliar realizada pela Prefeitura do Recife
- Objetivo: mapear padrões de deslocamento para planejamento urbano
- Representatividade: amostra estratificada por bairro e renda

#### Etapa 2: Limpeza e Preparação dos Dados

##### **Tratamento de valores ausentes:**

```
1 # Substituir códigos numéricos por categorias legíveis
2 SEXO_MAP = {1: 'Masculino', 2: 'Feminino'}
3 df['sexo_desc'] = df['sexo'].map(SEXO_MAP)
4
5 # Tratar valores ausentes em modais
6 df['meio_transporte_trab'].fillna('Não_declarado', inplace=True)
7
8 # Converter strings de modais em listas
9 def clean_modal(value):
10     if pd.isna(value):
11         return []
12     return [int(x) for x in str(value).split(',') if x.strip().
13             isdigit()]
```

```

14 df[ 'modal_trabalho_list' ] = df[ 'meio_transporte_trab' ]. apply (
    clean_modal)

```

Exemplo de limpeza de dados

### Criação de variáveis derivadas:

```

1 # Classificar tipo de trajeto (monomodal/multimodal)
2 df[ 'tipo_trajeto_trabalho' ] = df[ 'modal_trabalho_list' ]. apply (
3     lambda x: 'multimodal' if len(x) > 1 else 'monomodal' if len(x)
4         == 1 else 'sem_resposta'
5 )
6 # Contar número de modais utilizados
7 df[ 'num_modais' ] = df[ 'modal_trabalho_list' ]. apply( len )
8
9 # Identificar usuários de integração
10 df[ 'usuario_integracao' ] = (
11     (df[ 'utiliza_terminal_int_trabalho' ] == 1) |
12     (df[ 'terminal_int_trabalho' ] != '0')
13 )

```

Engenharia de features

## Etapa 3: Análise Exploratória de Dados (EDA)

### Estatísticas descritivas:

- Distribuições de frequência para variáveis categóricas
- Medidas de tendência central e dispersão para variáveis numéricas
- Identificação de outliers e valores inconsistentes

### Análise bivariada:

- Crosstabs entre variáveis categóricas (sexo vs. modal, renda vs. modal)
- Correlações entre variáveis numéricas
- Testes de hipótese quando apropriado

## Etapa 4: Visualização de Dados

### Princípios aplicados:

1. **Clareza:** Gráficos simples e diretos, sem poluição visual
2. **Consistência:** Mesma paleta de cores e estilos em todo dashboard
3. **Contexto:** Títulos descritivos, labels claros, unidades explícitas

4. **Hierarquia:** Informações mais importantes em destaque
5. **Interatividade:** Tooltips, zoom, filtros para exploração

#### Exemplo de código Plotly:

```

1 # Gráfico de barras para distribuição por sexo
2 sexo_counts = df_int['sexo_desc'].value_counts()
3 fig = px.bar(
4     x=sexo_counts.index,
5     y=sexo_counts.values,
6     labels={'x': 'Sexo', 'y': 'Número_de_Usuários'},
7     title='Usuários_de_Integração_por_Sexo',
8     color=sexo_counts.index,
9     color_discrete_map={
10         'Masculino': '#1f77b4',
11         'Feminino': '#ff7f0e'
12     }
13 )
14 fig.update_layout(showlegend=False)
15 st.plotly_chart(fig, use_container_width=True)

```

Criação de gráfico interativo

## Etapa 5: Modelagem Preditiva

### Regressão Linear:

```

1 from sklearn.linear_model import LinearRegression
2 from sklearn.metrics import mean_squared_error, r2_score
3
4 # Preparar dados
5 X = df[['idade', 'renda', 'distancia_trabalho']]
6 y = df['num_modais']
7
8 # Dividir em treino e teste
9 X_train, X_test, y_train, y_test = train_test_split(
10     X, y, test_size=0.2, random_state=42
11 )
12
13 # Treinar modelo
14 model = LinearRegression()
15 model.fit(X_train, y_train)
16
17 # Avaliar
18 y_pred = model.predict(X_test)
19 r2 = r2_score(y_test, y_pred)
20 rmse = np.sqrt(mean_squared_error(y_test, y_pred))
21
22 print(f"R2 Score: {r2:.4f}")
23 print(f"RMSE: {rmse:.4f}")

```

---

## Modelo de regressão

### Classificação (Random Forest):

```
1 from sklearn.ensemble import RandomForestClassifier
2 from sklearn.metrics import accuracy_score, classification_report
3
4 # Preparar dados
5 X = df[['sexo', 'idade', 'renda', 'bairro_encoded']]
6 y = df['tipo_trajeto_trabalho']
7
8 # Treinar modelo
9 rf = RandomForestClassifier(n_estimators=100, random_state=42)
10 rf.fit(X_train, y_train)
11
12 # Avaliar
13 y_pred = rf.predict(X_test)
14 accuracy = accuracy_score(y_test, y_pred)
15
16 print(f"Acurácia: {accuracy:.4f}")
17 print(classification_report(y_test, y_pred))
```

## Modelo de classificação

## Etapa 6: Desenvolvimento do Dashboard

### Arquitetura do código:

- app.py: Arquivo principal com lógica do dashboard
- Funções modulares: Cada página tem sua própria função (show\_overview(), show\_modal\_share() etc.)
- Cache de dados: @st.cache\_data para carregar dataset uma única vez
- Separação de lógica: Preparação de dados vs. visualização

### Estrutura do app.py:

```
1 # 1. Configuração da página
2 st.set_page_config(
3     page_title="Dashboard - Mobilidade Urbana RMR",
4     page_icon=":bus:",
5     layout="wide"
6 )
7
8 # 2. Funções de carregamento
9 @st.cache_data
10 def load_data():
11     df = pd.read_csv('dados/dataset2.csv')
```



```

12     return df
13
14 # 3. Funções de preparação
15 def prepare_data(df):
16     # Limpeza e feature engineering
17     return df
18
19 # 4. Funções de visualização
20 def show_overview(df):
21     # KPIs e gráficos da visão geral
22     pass
23
24 # 5. Função principal
25 def main():
26     df = load_data()
27     df = prepare_data(df)
28
29     # Menu lateral
30     page = st.sidebar.radio("Selecione a página:", [...])
31
32     # Roteamento
33     if page == "Visão Geral":
34         show_overview(df)
35     elif page == "Modal Share":
36         show_modal_share(df)
37     # ...
38
39 if __name__ == "__main__":
40     main()

```

Estrutura do dashboard

## Etapa 7: Deploy e Documentação

### Deploy no Streamlit Cloud:

1. Conectar repositório GitHub ao Streamlit Cloud
2. Configurar `requirements.txt` com dependências
3. Especificar arquivo principal (`streamlit_app/app.py`)
4. Deploy automático a cada commit no branch main

### Documentação:

- `README.md`: Instruções de instalação, execução e deploy
- Comentários no código: Explicação de lógica complexa
- Docstrings: Documentação de funções principais
- Este documento LaTeX: Documentação formal do projeto

## 5.2 Ferramentas e Tecnologias Utilizadas

### Linguagem de Programação

- **Python 3.11:** Linguagem principal para análise e dashboard

### Bibliotecas de Análise de Dados

- **Pandas:** Manipulação e análise de dataframes
- **NumPy:** Operações numéricas e arrays
- **SciPy:** Estatística e testes de hipótese (se aplicado)

### Bibliotecas de Visualização

- **Plotly Express:** Gráficos interativos principais
- **Matplotlib:** Gráficos estáticos e personalizados
- **Seaborn:** Visualizações estatísticas (heatmaps, distribuições)

### Machine Learning

- **scikit-learn:** Modelos de regressão e classificação
- **train\_test\_split:** Divisão treino/teste
- **Métricas:** accuracy\_score, f1\_score, r2\_score, etc.

### Framework Web

- **Streamlit:** Framework para criação do dashboard

### Containerização e Deploy

- **Docker:** Containerização para ambiente reproduzível
- **Docker Compose:** Orquestração de containers
- **Streamlit Cloud:** Plataforma de deploy gratuita
- **GitHub:** Controle de versão do código

## 5.3 Boas Práticas Aplicadas

### Código

- **Modularização:** Funções separadas para cada responsabilidade
- **Nomenclatura clara:** Variáveis e funções com nomes descritivos
- **Comentários:** Explicação de lógica não-trivial
- **Constantes:** Dicionários de mapeamento definidos no topo
- **DRY (Don't Repeat Yourself):** Reutilização de código via funções

### Visualização

- **Consistência visual:** Mesma paleta de cores e estilos
- **Títulos descritivos:** Todo gráfico tem título claro
- **Labels nos eixos:** Sempre com unidades quando aplicável
- **Tooltips informativos:** Hover mostra dados detalhados
- **Responsividade:** Gráficos adaptam a diferentes tamanhos de tela

### Performance

- **Cache de dados:** `@st.cache_data` para evitar recarregamento
- **Lazy loading:** Carregar apenas página selecionada
- **Agregações eficientes:** Usar `groupby` e `value_counts` do Pandas

### Documentação

- **README completo:** Instruções passo a passo
- **Requirements.txt:** Lista todas as dependências
- **Boxes de insights:** Cada análise tem interpretação destacada
- **Documentação formal:** Este documento LaTeX

## 6. Conclusões

### 6.1 Objetivos Alcançados

Este projeto teve como objetivo desenvolver uma solução analítica e visual completa para análise de mobilidade urbana na RMR. Os principais objetivos alcançados foram:

1. **Dashboard interativo funcional:** Plataforma web com 11 páginas de análise, acessível online via Streamlit Cloud
2. **Análises demográficas detalhadas:** Identificação de perfis de usuários por sexo, idade, renda e localização
3. **Análise de multimodalidade:** Quantificação e caracterização de viagens que utilizam múltiplos modais
4. **Modelos preditivos:** Implementação de modelos de regressão e classificação com métricas de avaliação
5. **Insights acionáveis:** Recomendações práticas para políticas públicas de mobilidade urbana
6. **Clareza na comunicação visual:** Gráficos intuitivos, cores consistentes, tooltips informativos

### 6.2 Aprendizados Principais

#### Técnicos

- Domínio de Streamlit para desenvolvimento rápido de dashboards
- Experiência com Plotly para visualizações interativas
- Prática em feature engineering e limpeza de dados reais
- Implementação de pipeline completo: EDA → Modelagem → Deploy
- Uso de Docker para padronizar ambientes de desenvolvimento

#### Analíticos

- Compreensão de padrões de mobilidade urbana e fatores que os influenciam
- Importância do contexto socioeconômico em análise de dados
- Limitações de modelos preditivos com dados desbalanceados
- Valor de visualização para comunicar insights complexos

## Processo

- Importância de iterações rápidas e feedback contínuo
- Necessidade de documentar decisões durante o desenvolvimento
- Valor de containerização para reprodutibilidade
- Benefícios de deploy automatizado via GitHub

## 6.3 Limitações e Trabalhos Futuros

### Limitações do projeto atual

- **Dados de 2016:** Dataset desatualizado (9 anos), não reflete mudanças recentes (COVID-19, novos modais)
- **Falta de variáveis temporais:** Não há análise de variação ao longo do dia ou semana
- **Ausência de dados geoespaciais:** Coordenadas geográficas permitiriam mapas interativos
- **Modelos preditivos limitados:**  $R^2$  moderado e acurácia afetada por desbalanceamento de classes
- **Sem análise de custo:** Impacto financeiro das viagens não é considerado

### Melhorias futuras

1. **Integração com dados em tempo real:** APIs de transporte público para dados atualizados
2. **Mapas interativos:** Usar Folium ou Plotly Maps para visualizar rotas e fluxos
3. **Análise temporal:** Padrões de pico de hora, sazonalidade, dias da semana
4. **Modelos avançados:** XGBoost, redes neurais, SMOTE para balanceamento
5. **Análise de custo-benefício:** Calcular economia de tempo e custo por modal
6. **Dashboard mobile:** Otimizar interface para smartphones
7. **Filtros interativos globais:** Permitir usuário filtrar por bairro, renda, etc. em todas as páginas
8. **Comparação com outras cidades:** Benchmark com São Paulo, Rio, Brasília
9. **Análise de acessibilidade:** Medir quão bem cada bairro é servido por transporte público
10. **Simulações de cenários:** "E se expandíssemos o metrô para bairro X?"

## 6.4 Impacto Esperado

Este projeto tem potencial de impactar:

- **Gestores públicos:** Subsidiar decisões sobre investimentos em transporte
- **Planejadores urbanos:** Identificar áreas carentes de infraestrutura
- **Operadores de transporte:** Otimizar rotas e frequências baseado em demanda
- **Cidadãos:** Conscientização sobre padrões de mobilidade e necessidades
- **Pesquisadores:** Base para estudos futuros em mobilidade urbana
- **Desenvolvedores:** Análise dos dados para criar novas soluções tecnológicas

## 6.5 Conclusão

A análise de dados de mobilidade urbana é fundamental para construir cidades mais eficientes e sustentáveis. Este projeto demonstrou que, com ferramentas adequadas e clareza na comunicação visual, é possível transformar dados brutos em insights valiosos para tomadores de decisão.

A Região Metropolitana do Recife enfrenta desafios típicos de grandes centros urbanos brasileiros: dependência excessiva de ônibus, infraestrutura desigual entre centro e periferia, e dificuldades de mobilidade para populações de baixa renda. As análises apresentadas neste dashboard evidenciam essas questões e apontam caminhos para soluções.

O domínio do processo completo de análise e visualização de dados – desde a coleta e limpeza até a modelagem e deploy – é uma habilidade essencial para cientistas de dados modernos. Este projeto serviu como exercício prático dessas competências, com ênfase particular na clareza da comunicação visual, conforme solicitado nos critérios de avaliação.

Por fim, espera-se que este trabalho contribua não apenas como entrega acadêmica, mas como ferramenta útil para quem se interessa por mobilidade urbana na RMR, seja para fins de pesquisa, planejamento ou advocacy por melhores políticas públicas.