TECNOLOGIA BARREIRO
ESCOLA SUPERIOR POLITÉCNICO SETÚBAL

Análise do Mercado Imobiliário em Washington

Unidade Curricular: Data Mining

Licenciatura em Bioinformática

Docentes da UC:

Ana Mendes & Luís Pereira

2023/2024

Data: 12/06/2024

Joana Caetano 202200646

Lucas Terlica 202200239

Lara Evangelista 202200655

Índice

Índice Figuras

Índice de Tabelas

Apresentação do Caso de Estudo

Introdução

O mercado imobiliário é uma área complexa e dinâmica, onde diversos fatores influenciam os preços das propriedades. Analisar esses fatores é essencial para tomar decisões informadas, seja para compra, venda ou investimento. Este caso de estudo foca-se na análise de um dataset detalhado de preços de moradias em Washington, disponível no Kaggle. Através de técnicas de mineração de dados, o nosso objetivo é extrair insights valiosos que ajudem a entender as dinâmicas do mercado imobiliário.

A mineração de dados permite transformar grandes volumes de dados brutos em informações úteis, revelando padrões e relações ocultas que podem não ser evidentes à primeira vista. No contexto do mercado imobiliário, essa análise pode destacar quais características de uma propriedade têm maior impacto no seu preço de venda, identificar tendências de mercado e prever preços futuros com maior precisão.

Descrição do Dataset

O dataset contém as seguintes colunas, cada uma representando uma característica específica das moradias:

id: Identificação única de cada propriedade.

date: Data da venda do imóvel.

price: Preço de venda da casa.

bedrooms: Número de quartos na casa.

bathrooms: Número de casas de banho na casa.

sqft_living: Área útil da casa em pés quadrados.

sqft_lot: Tamanho do lote em pés quadrados.

floors: Número de andares da casa.

waterfront: Indica se a casa possui vista para a água (1) ou não (0).

view: Índice de qualidade da vista da casa.

condition: Condição atual da casa, com base numa avaliação.

grade: Qualidade da construção e design da casa.

sqft_above: Área em pés quadrados do espaço acima do solo.

sqft_basement: Área em pés quadrados do sótão.

yr_built: Ano de construção da casa.

yr_renovated: Ano da última renovação da casa.

zipcode: Código postal da localização da casa.

lat: Latitude da localização da casa.

long: Longitude da localização da casa.

sqft_living15: Área útil dos 15 vizinhos mais próximos em pés quadrados.

sqft_lot15: Tamanho do lote dos 15 vizinhos mais próximos em pés quadrados.

Estas colunas fornecem uma visão abrangente das características das propriedades, permitindo análises detalhadas e a criação de modelos preditivos para entender e prever os preços das moradias.

Objetivos

O principal objetivo deste trabalho é realizar um estudo detalhado do mercado imobiliário de Washington, DC, utilizando um conjunto abrangente de técnicas e ferramentas aprendidas nas aulas. Para enriquecer e contextualizar a análise, incorporámos dados do mercado imobiliário da Califórnia como ponto de comparação. Esta abordagem permite-nos identificar e entender os principais fatores que influenciam os preços das propriedades em Washington, ao mesmo tempo que comparamos estas dinâmicas com as observadas na Califórnia.

Especificamente, pretendemos:

- Identificar as características das propriedades que mais impactam os seus preços de venda em Washington;
- Analisar as tendências de preços ao longo do tempo;
- Investigar como eventos históricos e económicos específicos afetaram o mercado imobiliário em ambas as regiões;
- Examinar a influência de fatores ambientais e geográficos, como proximidade ao mar ou áreas urbanas centrais, sobre a valorização imobiliária;
- Analisar a relação entre a qualidade das construções (classificação "grade") e os preços das propriedades;
- Comparar e contrastar as influências socioeconómicas e demográficas nos mercados imobiliários de Washington e Califórnia;
- Estudar a disponibilidade e distribuição de diferentes tipos de propriedades (por número de quartos e casas de banho) e como isso varia entre as duas regiões.

Utilizando técnicas de mineração de dados, transformações de variáveis e modelos preditivos, procuramos aplicar ao máximo as competências e conhecimentos adquiridos ao longo do curso. Através desta análise comparativa, esperamos fornecer insights valiosos que possam apoiar decisões informadas de compra, venda e investimento no setor imobiliário.

Desenvolvimento

Transformação de Tipos de Variáveis e Criação de uma Nova Coluna

A transformação de dados é uma etapa crucial no processo de análise e mineração de dados. A qualidade e a precisão das análises dependem diretamente da forma como os dados são preparados. Em seguida, será apresentada a importância de alterar os tipos das variáveis e a criação de novas colunas.

Alteração dos Tipos de Variáveis

Transformar os tipos de variáveis significa converter os dados de um tipo para outro, como de texto para número, ou de data/hora para texto. Esta transformação é essencial por várias razões:

Precisão Analítica: Certos tipos de análise requerem que os dados estejam em formatos específicos. Por exemplo, análises estatísticas e cálculos matemáticos precisam que os números estejam no formato numérico.

Consistência: A homogeneidade dos dados é fundamental para evitar erros durante a análise. Dados inconsistentes podem levar a resultados imprecisos ou falhas nos algoritmos de mineração de dados.

Eficiência de Processamento: Tipos de dados adequados permitem que as ferramentas de análise processem as informações de maneira mais rápida e eficiente.

Transformações Realizadas

id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade
1873100390	02/03/2015 00:00:00	719000	4	2.5	2570	7173	2	0	0	3	8
2078500320	20/06/2014 00:00:00	605000	4	2.5	2620	7553	2	0	0	3	8
5416510140	10/07/2014 00:00:00	360000	4	2.5	2380	5000	2	0	0	3	8
9545230140	25/07/2014 00:00:00	597750	4	2.5	2310	9624	2	0	0	3	8
1873100060	29/08/2014 00:00:00	693000	4	2.5	2460	4425	2	0	0	3	8

Figura 1. Tabela Original do Dataset Escolhido

sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15
2570	0	2005	0	98052	47.7073	-122.11	2630	6026
2620	0	1996	0	98056	47.5301	-122.18	2620	11884
2380	0	2005	0	98038	47.3608	-122.036	2420	5000
2310	0	1984	0	98027	47.5386	-122.053	1940	9636
2460	0	2006	0	98052	47.7048	-122.109	2990	5659

Figura 2. Tabela Original do Dataset Escolhido

id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade
1873100390	20150302T000000	$719 000,00	4	2,5	2570	7173	2	0	0	3	8
2078500320	20140620T000000	$605 000,00	4	2,5	2620	7553	2	0	0	3	8
5416510140	20140710T000000	$360 000,00	4	2,5	2380	5000	2	0	0	3	8
9545230140	20140725T000000	$597 750,00	4	2,5	2310	9624	2	0	0	3	8
1873100060	20140829T000000	$693 000,00	4	2,5	2460	4425	2	0	0	3	8

Figura. 3 Tabela Transformada do Dataset Escolhido

sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15	IntervaloPreco
2570	0	2005	0	98052	47.7073	-122.11	2630	6026	500K-750K
2620	0	1996	0	98056	47.5301	-122.18	2620	11884	500K-750K
2380	0	2005	0	98038	47.3608	-122.036	2420	5000	250K-500K
2310	0	1984	0	98027	47.5386	-122.053	1940	9636	500K-750K
2460	0	2006	0	98052	47.7048	-122.109	2990	5659	500K-750K

Figura. 4 Tabela Transformada do Dataset Escolhido

Tabela 1. Tabela com as Alterações de Tipo Realizadas no Dataset Escolhido

Coluna	***Original***	***Transformada***
date	Data/hora	Texto
price	Texto	Número decimal
bathrooms	Texto	Número decimal
floors	Texto	Número decimal
IntervaloPreco	*NA*	Texto

Conforme detalhado na Tabela 2.:

- A coluna "*date*" foi convertida de Data/hora para Texto. É necessário quando se quer analisar componentes específicos da data (como o mês ou o ano) de forma mais direta.
- A coluna "*price*" foi transformada de Texto para Número decimal, permitindo cálculos precisos de preços médios, totais e outras análises financeiras.
- As colunas "*bathrooms*" e *"floors"* foram transformadas de Texto para Número decimal, possibilitando análises quantitativas mais robustas.

Criação de uma Nova Coluna

A criação de novas colunas é muitas vezes necessária para enriquecer a análise dos dados. Novas colunas podem ser derivadas das existentes para:

Agrupar Dados: Facilitar a agregação de dados, como categorizar preços em faixas (IntervaloPreco).

Calcular Novas Métricas: Criar métricas derivadas que podem ser mais informativas que os dados brutos originais.

Ajudar na Visualização: Melhorar a visualização dos dados em relatórios e dashboards, proporcionando uma compreensão mais clara e imediata dos dados.

```
IntervaloPreco = SWITCH(
  TRUE(),
  Housing[price] <= 250000, "0-250K",
  Housing[price] <= 500000, "250K-500K",
  Housing[price] <= 750000, "500K-750K",
  Housing[price] <= 1000000, "750K-1000K",
  Housing[price] > 1000000, ">1000K"
)

```

Figura 5. Código para a Criação da Coluna “IntervaloPreco”

A coluna *“IntervaloPreco”* foi adicionada para categorizar os preços em diferentes faixas. Esta categorização facilita a análise de padrões de preços e permite a criação de visualizações que destacam as distribuições de preços de maneira clara e compreensível.

Incorporação de um Segundo Dataset

Inicialmente, utilizamos um dataset focado no mercado imobiliário de Washington, DC. No entanto, para obter uma visão comparativa mais completa e aprofundada entre diferentes regiões, decidimos incorporar um novo dataset que inclui informações sobre o mercado imobiliário da Califórnia.

Razões para a Utilização do Segundo Dataset

Amplitudes Regionais:

- Cobertura Geográfica Ampliada: Ao incluir dados de ambas as regiões, Washington e Califórnia, podemos comparar e contrastar mercados imobiliários distintos, proporcionando uma análise mais robusta e abrangente das tendências e variáveis que influenciam os preços das propriedades em diferentes contextos geográficos.

- Diversidade de Mercado: Washington e Califórnia apresentam diferentes características socioeconómicas e demográficas. Incorporar dados de ambas as regiões permite identificar como esses fatores distintos afetam o mercado imobiliário em cada área.

Comparação e Contraste:

- Análise Comparativa: Comparar mercados diferentes como Washington e Califórnia permite identificar padrões comuns e divergentes, revelando insights sobre fatores específicos que afetam cada mercado de maneira única.

- Identificação de Tendências Gerais e Específicas: A análise comparativa ajuda a distinguir entre tendências que são gerais para o mercado imobiliário como um todo e aquelas que são específicas a uma região particular.

Enriquecimento da Análise:

- Maior Volume de Dados: A inclusão de um novo dataset aumenta o volume de dados disponíveis para análise, melhorando a robustez estatística e a precisão dos modelos preditivos.

Transformações Realizadas:

Column1	Column2	Column3	Column4	Column5	Column6
$599,999	3bed	2bath	2,168sqft	8,712sqft lot	135 Hillside Pl, Jackson, CA 95642
$735,000	3bed	2bath	1,583sqft	5,227sqft lot	54960 Avenida Rubio, La Quinta, CA 92253
$495,000	3bed	2bath	1,516sqft	0.85acre lot	17652 Cindy Ln, Grass Valley, CA 95945
$319,000	3bed	2bath	1,757sqft	0.32acre lot	20800 Melville Dr, California City, CA 93505
$398,000	3bed	2bath	1,315sqft	6,098sqft lot	477 S Maidstone St, Banning, CA 92220

Figura 6. Tabela Original do Segundo Dataset

price	bedrooms	bathrooms	sqft	sqft_lot	address
499000	3	2	1664	8,195sqft lot	34 E 15th St, Antioch, CA 94509
568000	3	2	1160	7,000sqft lot	215 Allan Ave, Rohnert Park, CA 94928
499500	3	2	1224	5,663sqft lot	7472 Freda Ave, Riverside, CA 92504
499999	3	2	1980	0.46acre lot	19075 Bay Meadows Dr, Apple Valley, CA 92308
284000	3	2	1661	0.41acre lot	2643 Evans Rd, Wofford Heights, CA 93285

Figura 7. Tabela Transformada do Segundo Dataset

Tabela 2. Tabela com as Alterações de Tipo Realizadas no Segundo Dataset

Coluna	***Original***	***Transformada***
price	Texto	Número decimal
bedrooms	Texto	Número inteiro
bathrooms	Texto	Número decimal
sqft	Texto	Número Inteiro

Para realizar as alterações de tipo observadas na Tabela 2., tivemos que:

- Alterar o nome de cada coluna;
- Remover símbolos como "$" e "+";
- Remover vírgulas, palavras como "bed", "bath" e "sqft";
- Remover todas as linhas que tinham "studio" na coluna do número de quartos devido ao facto de não se aplicarem ao nosso estudo;
- Remover todas as linhas que tinham, por exemplo, "$144,000$5k" na coluna do preço.

Aqui estão as razões para as alterações de Tipo na Tabela 2.

Coluna "*price*": Texto para Número Decimal

Precisão nas Análises Financeiras: O preço de venda das propriedades é uma variável numérica que deve ser tratada como tal para permitir cálculos precisos de médias, somas, variâncias, entre outros indicadores financeiros.

Consistência de Dados: Manter a consistência do tipo de dado facilita a manipulação e análise dos dados, reduzindo erros de processamento.

Coluna "*bedrooms*": Texto para Número inteiro

Análises Quantitativas: O número de quartos é uma variável discreta que deve ser tratada como um número inteiro para possibilitar análises quantitativas robustas, como distribuições e contagens.

Facilidade de Interpretação: Representar os quartos como um número inteiro torna mais intuitivo interpretar e visualizar a quantidade de quartos nas propriedades.

Coluna "*bathrooms*": Texto para Número decimal

Precisão nos Detalhes: O número de casas de banho pode incluir frações (por exemplo, 1,5 casas de banho), o que justifica a necessidade de um tipo de dado decimal para capturar essa precisão.

Cálculos e Análises: Transformar o número de casas de banho num número decimal permite a realização de cálculos precisos, como médias e distribuições, e facilita a integração dessa variável em modelos preditivos.

Coluna "*sqft*": Texto para Número inteiro

Cálculos de Espaço: Representar a área útil como um número inteiro permite cálculos de áreas médias, totais e outras métricas relacionadas ao espaço, essenciais para a análise do valor e funcionalidade das propriedades.

Outputs Criados

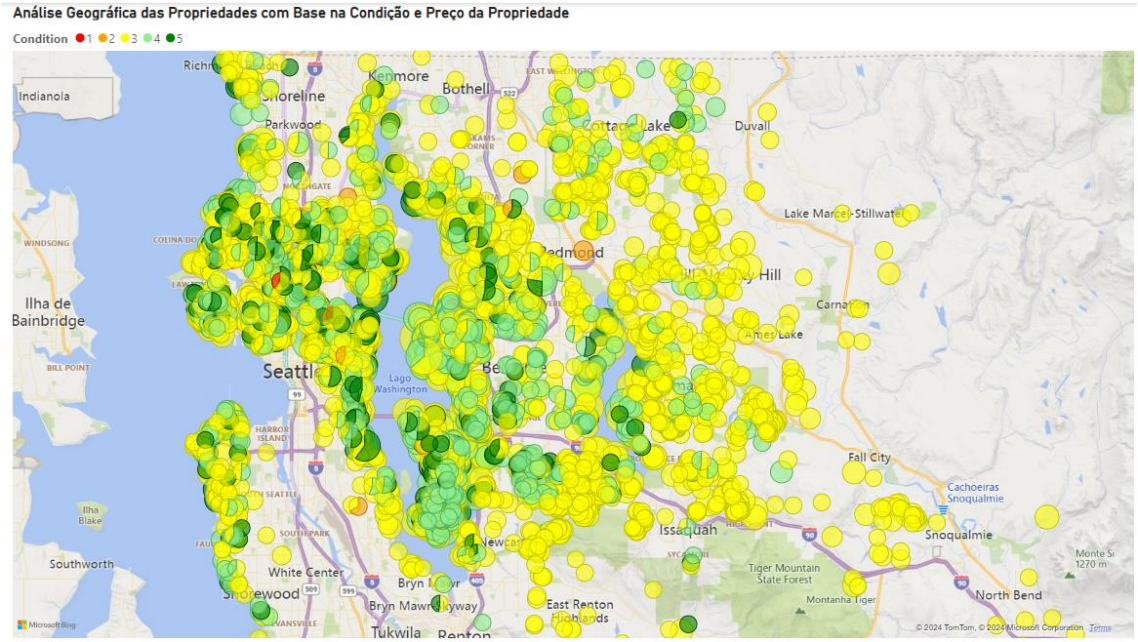

Figura. 8 Análise Geográfica das Propriedades com Base na Condição e Preço da Propriedade

O objetivo do mapa da Figura 8. é fornecer uma visualização geográfica das propriedades na região de Washington utilizando duas dimensões importantes: a condição das casas e o preço das mesmas. Este mapa interativo permite uma análise visual clara e intuitiva, facilitando a identificação de padrões e tendências no mercado imobiliário da área.

A cor dos círculos representa a condição da casa, variando entre 1 (vermelho), indicando a pior condição, e 5 (verde), que indica a melhor condição.

O tamanho dos círculos representa o preço das casas. Círculos maiores indicam preços mais altos, enquanto círculos menores indicam preços mais baixos.

Conclusões retiradas

Identificação de Áreas com Melhores Condições: As casas localizadas próximas ao mar tendem a estar em melhores condições do que as casas localizadas mais no interior.

Correlação entre Condição e Preço: Não se observa uma correlação clara entre o preço das casas e a sua condição. Tanto as casas em boas condições (círculos verdes) quanto em

condições medianas ou ruins (círculos amarelos e vermelhos) apresentam uma variedade de preços. Ou seja, o preço das propriedades não é significativamente influenciado apenas pela condição da casa.

Distribuição Geográfica de Preços: As casas que se encontram ao longo da linha costeira tendem a ser mais caras do que as que não se encontram.

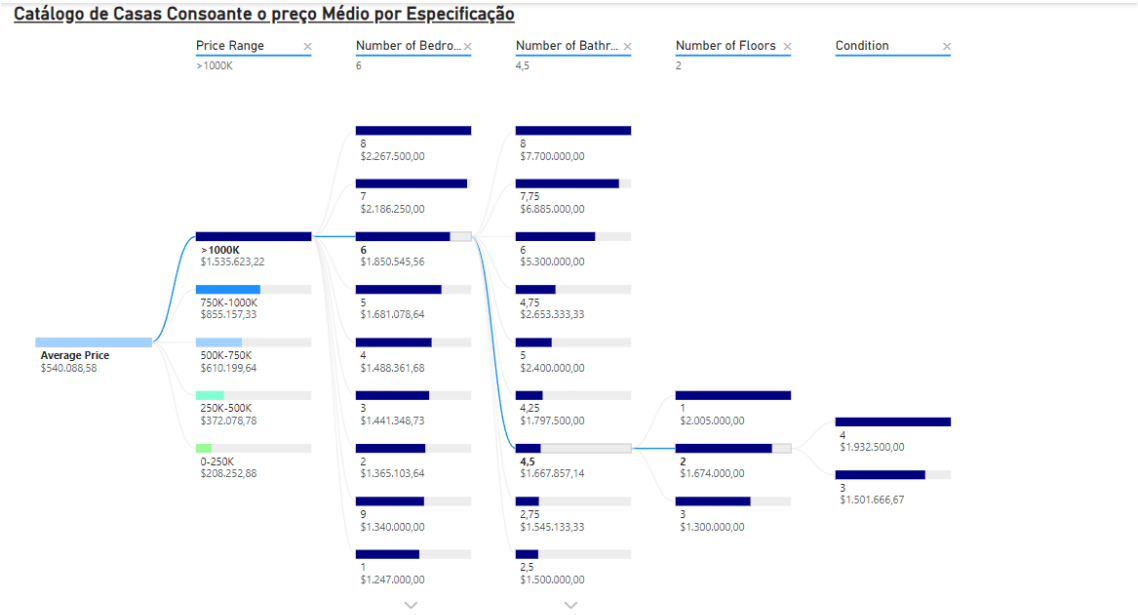

Figura 9. Árvore de Decomposição Consoante o Preço Médio por Especificação

O gráfico apresentado na Figura 9. visa catalogar as casas de acordo com diversas especificações, mostrando como cada uma influencia o preço médio das propriedades, e mostrar o preço médio para cada intervalo.

As casas são categorizadas em cinco intervalos de preço: 0-250k, 250-500k, 500-750k, 750-1000k e superior a 1000k. As cores das barras nos intervalos de preço proporcionam uma visualização clara dos preços médios, com tons mais escuros indicando preços mais altos.

Insight para Decisões de Compra e Venda:

Compradores: Podem usar estas informações para ajustar as suas expectativas e entender melhor o mercado. Por exemplo, se estiverem à procura de casas com um determinado número de quartos, podem ver como isso afeta o preço.

Vendedores: Podem definir preços competitivos e realistas para as suas propriedades, considerando como cada especificação influencia o valor.

Investidores: Podem identificar oportunidades de investimento, entendendo quais as características que são mais valorizadas no mercado.

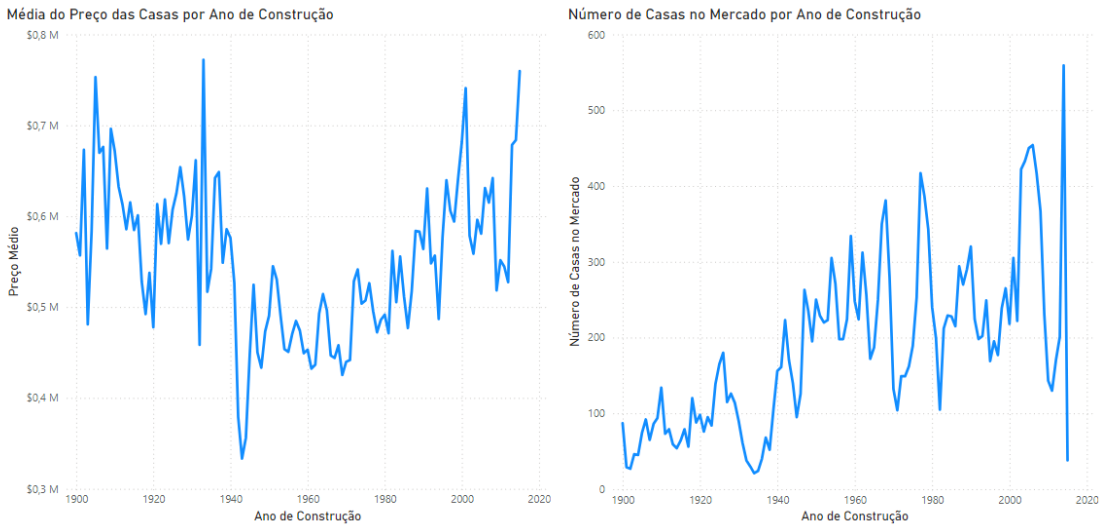

Figura 10. Gráficos Lineares da Análise do Preço Médio e Número de Casas por Ano de Construção

Os gráficos apresentados na Figura 10. têm como principal objetivo fornecer uma análise detalhada do preço médio das casas e do número de casas disponíveis no mercado, segmentados por ano de construção. Esta visualização ajuda a identificar tendências históricas no mercado imobiliário e a entender como diferentes períodos históricos influenciaram os preços das propriedades e a disponibilidade de novas construções.

Conclusões Retiradas

Média do Preço das Casas por Ano de Construção:

- Flutuações no Início do Século XX: Os preços médios mostram grandes flutuações nas primeiras décadas do século XX, refletindo a instabilidade económica e os impactos da Grande Depressão (1929).
- Queda Durante a Segunda Guerra Mundial (1939-1945): Há uma queda significativa no preço médio das casas durante a Segunda Guerra Mundial, possivelmente devido ao redirecionamento de recursos para o esforço de guerra e à redução da construção civil.
- Crescimento Pós-Guerra: A partir dos anos 1950, observa-se um aumento gradual no preço médio das casas, impulsionado pelo boom económico do pós-guerra e o crescimento suburbano.
- Aumento Recente: Nas décadas mais recentes, especialmente a partir dos anos 2000, os preços médios das casas têm aumentado significativamente, refletindo a valorização do mercado imobiliário e a demanda crescente por habitação em áreas urbanas.

Número de Casas no Mercado por Ano de Construção:

- Baixa Construção no Início do Século XX: O número de casas construídas no início do século XX é relativamente baixo, refletindo a menor urbanização e crescimento populacional da época.
- Aumento Durante os Anos 1940 e 1950: Há um aumento no número de casas construídas após a Segunda Guerra Mundial, durante o boom económico e a expansão suburbana.
- Flutuações nas Décadas Seguintes: O número de novas construções oscila ao longo das décadas, influenciado por recessões económicas, políticas habitacionais e mudanças demográficas.
- Pico Recentes: Nas últimas décadas, observa-se um aumento no número de casas no mercado, refletindo a urbanização contínua e a demanda por novas habitações.

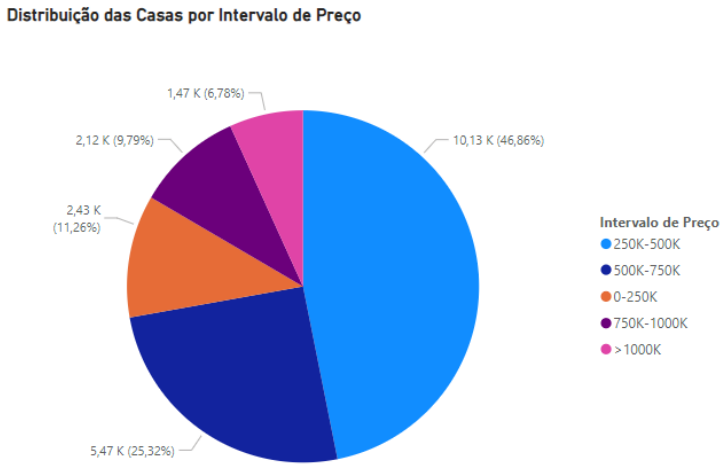

Figura 11. Gráfico Circular da Distribuição das Casas por Intervalo de Preço feito no PowerBI

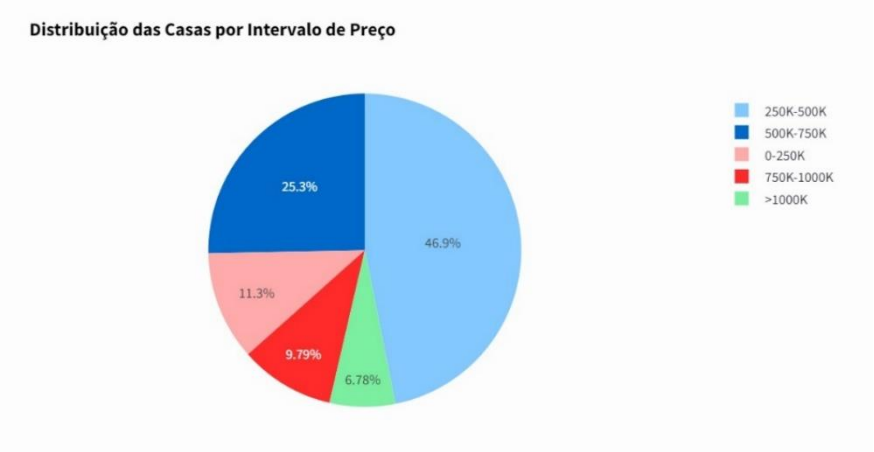

Figura 12. Gráfico Circular da Distribuição das Casas por Intervalo de Preço feito em Python

```python
import streamlit as st
import pandas as pd
import plotly.express as px

file_path = 'TransformedHousing.csv'  # Update this path
housing_data = pd.read_csv(file_path, delimiter=';')

price_intervals = housing_data['IntervaloPreco'].value_counts().reset_index()
price_intervals.columns = ['IntervaloPreco', 'Count']

fig = px.pie(price_intervals, values='Count', names='IntervaloPreco',
             title='Distribuição das Casas por Intervalo de Preço',
             hole=0)

st.title('Distribuição das Casas por Intervalo de Preço')
st.plotly_chart(fig)
```

Figura 13. Script do Gráfico da Figura 12.

OS gráficos circulares apresentados na Figuras 11. e na Figura 12., um criado no Power BI e outro em Python, têm como principal objetivo mostrar a distribuição das casas por diferentes intervalos de preço. Esta visualização permite uma análise clara e imediata da concentração de propriedades em cada faixa de preço, oferecendo insights valiosos.

Conclusões Retiradas

Identificação de Faixas de Preço Dominantes:

A maioria das casas está no intervalo de preço entre 250K e 500K, representando aproximadamente 46,9% das propriedades.

A segunda maior concentração está no intervalo de preço entre 500K e 750K, representando 25,3% do número total de casas.

Análise de Outras Faixas de Preço:

Os intervalos de preço mais altos (>1000K) e mais baixas (0-250K) têm menos casas, refletindo uma menor concentração de propriedades.

O intervalo de 750K-1000K também tem uma presença menor comparada aos dois intervalos principais, indicando uma menor quantidade de propriedades nessa categoria de preço.

Insight para Tomada de Decisões

Para Compradores: Ajuda a identificar quais faixas de preço têm mais opções disponíveis, facilitando a escolha de acordo com o orçamento.

Para Vendedores: Fornece uma visão sobre a competitividade em diferentes faixas de preço, ajudando a definir estratégias de venda.

Para Investidores: Identifica segmentos de mercado com maior ou menor concentração de propriedades, auxiliando na tomada de decisões de investimento.

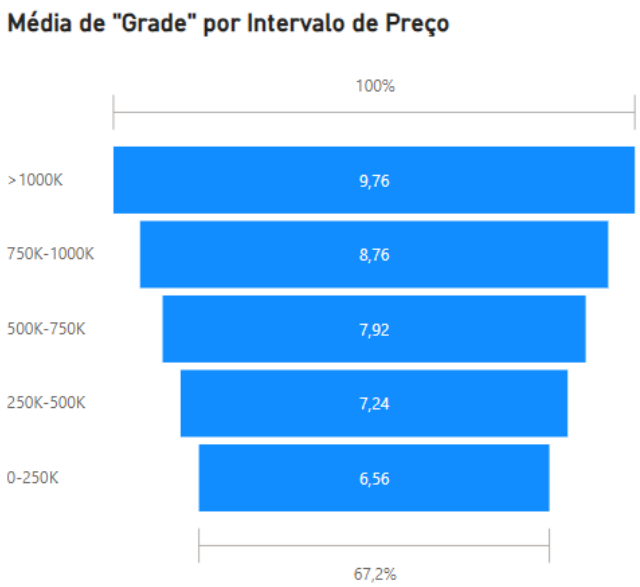

Figura 14. Gráfico em Funil da Média de Grade por Intervalo de Preço

O gráfico apresentado na Figura 14. tem como objetivo mostrar a média da classificação ("grade") das propriedades em diferentes intervalos de preço. A classificação ou "grade" é uma medida da qualidade da construção e design da propriedade, onde valores mais altos indicam uma qualidade superior. Este gráfico é fundamental para entender como a qualidade das propriedades varia conforme o preço e oferece uma visão clara e imediata de como o mercado valoriza a qualidade em diferentes faixas de preço.

Conclusões Retiradas

Relação entre Preço e Qualidade:

O gráfico mostra uma correlação positiva clara entre o preço das propriedades e a sua média de grade. Propriedades mais caras (>1000K) têm uma média de grade significativamente mais alta (9,76) em comparação com as propriedades mais baratas (0-250K) que têm uma média de grade de 6,56, como esperado. Ou seja, à medida que o intervalo de preço aumenta, a média de grade também aumenta consistentemente.

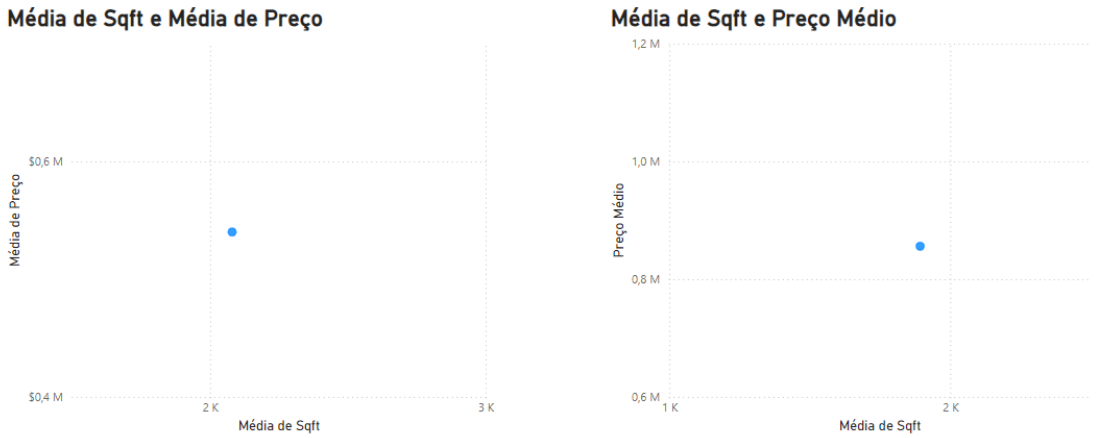

Figura 15. Gráficos de Dispersão da Média

Os gráficos apresentados na Figura 15. têm como objetivo mostrar a relação entre a média de área útil (em pés quadrados, "sqft") e a média de preço das propriedades em duas regiões distintas: Washington e Califórnia. Estes gráficos de dispersão permitem uma visualização clara de como a área útil de uma casa influencia o seu preço médio em cada região.

Conclusões Retiradas

Relação entre Tamanho e Preço em Washington:

Média de Sqft	2079,90
Média de Preço	$540.088,58

Figura 16. Valores do Ponto de Dispersão do Gráfico de Washington

Como observado na Figura 16., o ponto de dispersão revela que a média de área útil das casas é em torno de 2.080 sqft, com um preço médio próximo de $540.000.

Relação entre Tamanho e Preço na Califórnia:

Média de Sqft	1892,75
Preço Médio	855.991,99

Figura 17. Valores do Ponto de Dispersão do Gráfico na Califórnia

Como observado na Figura 17, a média de área útil das casas na Califórnia é em torno de 1.890 sqft, com um preço médio de aproximadamente $855.000.

Comparação entre Regiões:

Comparando os dois gráficos, fica evidente que as propriedades na Califórnia são geralmente mais caras do que em Washington, mesmo para áreas úteis semelhantes. O preço médio por 1000 sqft é cerca de $260.000 em Washington e $450.000 na Califórnia, ou seja, o preço médio por sqft na Califórnia é 73% mais elevado do que em Washington.

Os preços das propriedades na Califórnia são geralmente mais altos devido a uma combinação de fatores. Primeiro, a elevada procura por habitação, impulsionada por oportunidades de emprego, clima atraente e um estilo de vida desejável, supera a oferta limitada de novas casas. Em segundo lugar, a forte economia da Califórnia, com indústrias de alta renda como tecnologia, entretenimento e turismo, atrai trabalhadores bem remunerados que podem pagar preços mais elevados por moradia.

Além disso, as rigorosas regulamentações de construção e zoneamento limitam a quantidade de novas construções e aumentam os custos de desenvolvimento, contribuindo para os preços elevados. A atratividade geográfica e climática, com belas praias, montanhas e um clima ameno, torna a região ainda mais desejável para viver, aumentando a procura por habitação. O custo de vida na Califórnia também é significativamente mais alto do que em muitas outras partes dos Estados Unidos, o que inclui não apenas habitação, mas também impostos, alimentação e serviços.

Outro fator é o investimento internacional, que muitas vezes inflaciona os preços das propriedades, pois compradores estrangeiros pagam prémios elevados. Finalmente, a infraestrutura bem desenvolvida e os serviços públicos de alta qualidade, incluindo boas escolas, universidades de prestígio e instalações de saúde avançadas, tornam a Califórnia uma localização atraente para famílias e profissionais. Estes fatores combinados resultam em preços de propriedades significativamente mais altos na Califórnia em comparação com Washington e outras regiões.

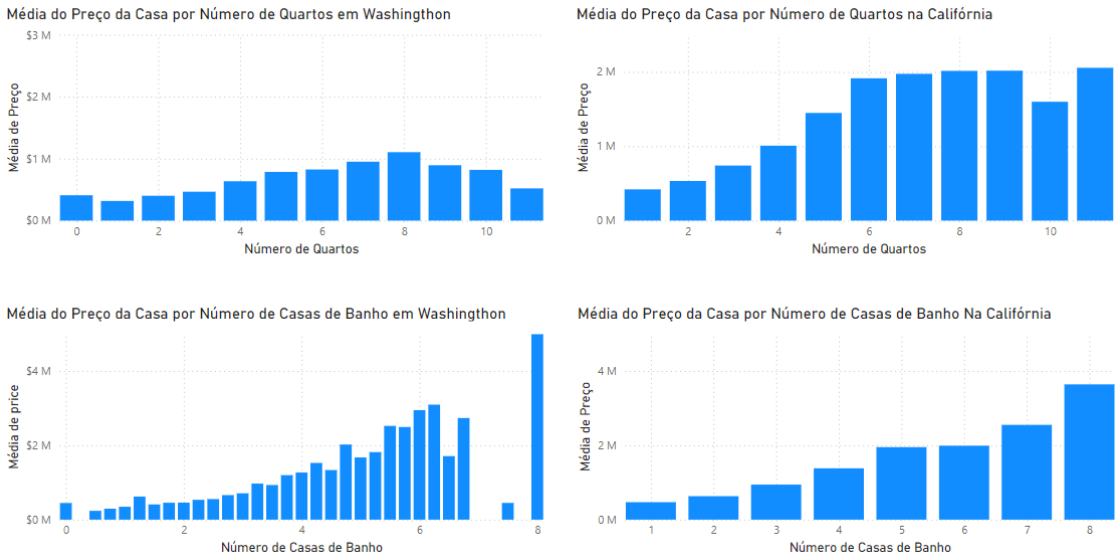

Figura 18. Gráficos de Colunas da Média do Preço das Casas por Número de Quartos e Casas de Banho em Washington e Califórnia

Os gráficos apresentados na Figura 18. têm como objetivo mostrar a relação entre o número de quartos e casas de banho das propriedades e o preço médio das mesmas, comparando duas regiões distintas: Washington e Califórnia. Esta visualização permite uma análise clara de como essas características específicas influenciam os preços das propriedades em cada região.

Conclusões Retiradas

Relação entre Número de Quartos e Preço:

Em Washington, o preço médio das casas aumenta com o número de quartos até cerca de 8 quartos, onde se estabiliza ou diminui ligeiramente.

Na Califórnia, o preço médio também aumenta com o número de quartos, mas a tendência é mais consistente, com os preços mais altos observados para casas com 8 a 10 quartos.

Relação entre Número de Casas de Banho e Preço:

Em Washington, por norma, há um aumento no preço médio com o aumento do número de casas de banho. Também se observou um valor muito elevado no preço médio das casas com 8 casas de banho em comparação aos outros valores.

Na Califórnia, o preço médio das casas também aumenta com o número de casas de banho, com uma tendência mais linear e preços significativamente mais altos para casas com 7 a 8 casas de banho.

Comparação entre Regiões:

Comparando as duas regiões, as propriedades na Califórnia têm preços médios mais altos do que em Washington para o mesmo número de quartos e casas de banho. Este padrão reforça a perceção de que o mercado imobiliário da Califórnia é mais caro.

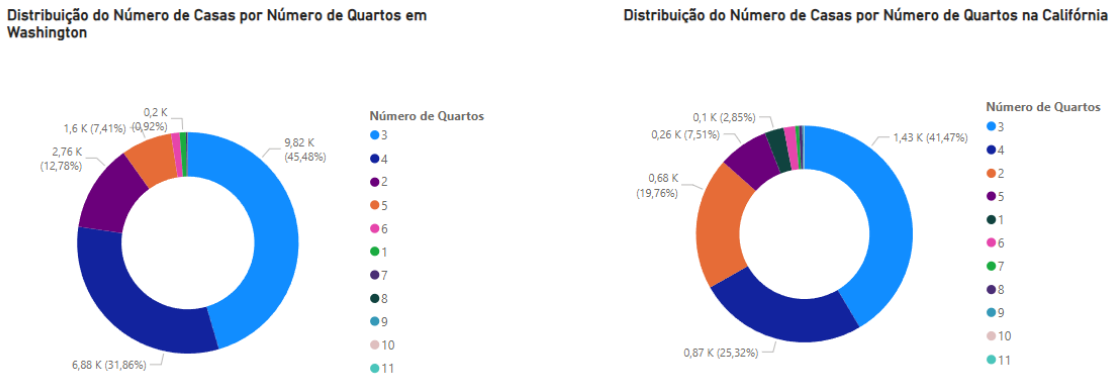

Figura 19. Gráficos em Anel da Distribuição Do Número de Casas por Número de Quartos em Washington e na Califórnia

Os gráficos apresentados na Figura 19. têm como objetivo comparar a distribuição do número de casas com diferentes quantidades de quartos nas regiões de Washington e Califórnia.

Conclusões

Distribuição em Washington:

- Grande parte das casas em Washington tem 3 quartos (45,48%), seguidas por casas com 4 quartos (31,86%).
- As casas com 2, 5 e 6 quartos têm uma menor representação no mercado.
- Casas com mais de 6 quartos representam uma fração mínima do total, indicando uma menor disponibilidade de propriedades de grande porte.

Distribuição na Califórnia:

- Grande parte das casas na Califórnia apresentam 3 (41,47%) e 4 quartos (25,32%).
- As casas com 2 quartos também têm uma representação significativa (19,76%).
- Assim como em Washington, as casas com mais de 6 quartos são menos comuns, mas há uma presença ligeiramente maior de casas com 5 quartos (7,51%).

Comparação entre Regiões:

Ambas as regiões mostram uma predominância de casas com 3 e 4 quartos, refletindo uma demanda similar por essas configurações familiares comuns.

A Califórnia tem uma maior percentagem de casas com 2 quartos em comparação com Washington, refletindo uma maior demanda por unidades menores devido aos altos preços imobiliários na região.

Washington tem uma percentagem ligeiramente maior de casas com 4 quartos em comparação com a Califórnia.

Dificuldades Sentidas e Como Foram Ultrapassadas

A análise do mercado imobiliário é uma tarefa desafiadora que envolve a manipulação e interpretação de grandes volumes de dados. Durante a realização do trabalho, enfrentamos algumas dificuldades tais como:

1. Qualidade e Consistência dos Dados

Dificuldade:

Os dados brutos frequentemente apresentavam inconsistências, como valores ausentes, formatos incorretos (números armazenados como texto), e símbolos indesejados (por exemplo, "$", "+" ou descrições textuais).

Solução:

Limpeza de Dados: Aplicação de técnicas de limpeza de dados, incluindo a remoção de símbolos indesejados e conversão de tipos de dados (de texto para numérico, por exemplo).

2. Integração de Dados de Múltiplas Fontes

Dificuldade:

A integração de datasets distintos (Washington e Califórnia) apresentou desafios na unificação dos dados devido a diferenças nas estruturas e nomenclaturas das colunas.

Solução:

Normalização: Harmonização dos datasets, unificando os nomes das colunas e assegurando que todas as colunas tivessem o mesmo formato.

Documentação: Criação de documentação detalhada sobre os processos de transformação e integração para garantir a reprodutibilidade e compreensão dos passos realizados.

Conclusão

A análise do mercado imobiliário de Washington DC permitiu uma compreensão aprofundada das variáveis que influenciam os preços das propriedades na região. Utilizando técnicas de mineração de dados, conseguimos transformar grandes volumes de dados em informações valiosas, revelando padrões e relações importantes. Através da análise geográfica, gráficos lineares, gráficos de dispersão e outras visualizações de dados, identificámos várias tendências e insights cruciais.

A criação de uma nova coluna, "IntervaloPreco", e a transformação de tipos de variáveis permitiram uma análise mais detalhada e precisa. A categorização dos preços em diferentes intervalos ajudou a visualizar a distribuição das propriedades por intervalos de preço, facilitando o estudo da concentração de propriedades, por exemplo.

A incorporação de um dataset adicional da Califórnia possibilitou uma análise comparativa, destacando diferenças regionais significativas. Esta comparação entre Washington e Califórnia permitiu contextualizar melhor as dinâmicas do mercado imobiliário de Washington.

Concluindo, este estudo destacou as variáveis mais impactantes nos preços das propriedades em Washington DC. A análise comparativa com a Califórnia proporcionou um contexto adicional, mas o foco principal permaneceu em Washington, oferecendo insights valiosos para decisões informadas de compra, venda e investimento no mercado imobiliário da região. As competências desenvolvidas e os conhecimentos adquiridos ao longo deste trabalho demonstram a importância da análise de dados no entendimento de mercados complexos e dinâmicos.

Referências

Brar, S. (n.d.). Housing price dataset. Kaggle. Disponível em https://www.kaggle.com/datasets/sukhmandeepsinghbrar/housing-price-dataset

Ibriee, E. (n.d.). USA California for Sale Properties. Kaggle. Disponível em https://www.kaggle.com/datasets/ibriiee/usa-california-for-sale-properties

Harvard Business School. (n.d.). Real Estate Prices During the Roaring Twenties and the Great Depression. Harvard Business School. Disponível em https://hbs.edu

Library of Congress. (n.d.). Overview | Progressive Era to New Era, 1900-1929 | U.S. History Primary Source Timeline | Classroom Materials at the Library of Congress. Library of Congress. Disponível em https://loc.gov

Library of Congress. (n.d.). Overview | The Post War United States, 1945-1968 | U.S. History Primary Source Timeline | Classroom Materials at the Library of Congress. Library of Congress. Disponível em https://loc.gov

Pruitt, S. (2020, August 5). The Post-War Economic Boom After WWII. ThoughtCo. Disponível em https://thoughtco.com/post-war-economic-boom-after-wwii-4056829

Legislative Analyst's Office. (2015, March 17). California's High Housing Costs: Causes and Consequences. Disponível em https://lao.ca.gov/reports/2015/finance/housing-costs/housing-costs.aspx

Flex. (n.d.). 9 Reasons Why California Is So Expensive - Flex | Pay Rent On Your Own Schedule. Flex. Disponível em https://getflex.com/blog/why-california-is-so-expensive