Vehicle Sales Data Analysis Report

Project Overview

The goal of this project was to analyze trends in the automotive market using the Vehicle Sales Data. By leveraging Python for data cleaning and exploratory data analysis (EDA) and Power BI for dashboard visualizations, we explored pricing fluctuations, vehicle conditions, and mileage, among other factors. This report documents the insights and methodology used in this project.

Dataset Description

The dataset contains various attributes related to vehicle sales. Below is a brief explanation of each column:

- **saledate**: The date when the vehicle was sold. This column was converted from text to DateTime format for analysis.
- **sellingprice**: The price at which the vehicle was sold.
- **odometer**: The mileage of the vehicle at the time of sale.
- **make**: The brand or manufacturer of the vehicle (e.g., Toyota, Ford).
- **body**: The type of vehicle body (e.g., Sedan, SUV, Pickup Truck).
- **state**: The state in which the vehicle was sold.
- **condition**: A numeric value representing the vehicle's condition. This was binned into categories: Excellent, Good, Average, Poor, and Very Poor.
- **year**: The manufacturing year of the vehicle.
- **model**: The specific model name of the vehicle.

Data Cleaning in Python

Steps Taken

1. Converted the **saledate** column from text to DateTime format.
2. Created a new column, **Condition Category**, by binning the numeric condition values into predefined categories (Excellent, Good, Average, Poor, Very Poor).
3. Removed blank and null values across all columns.

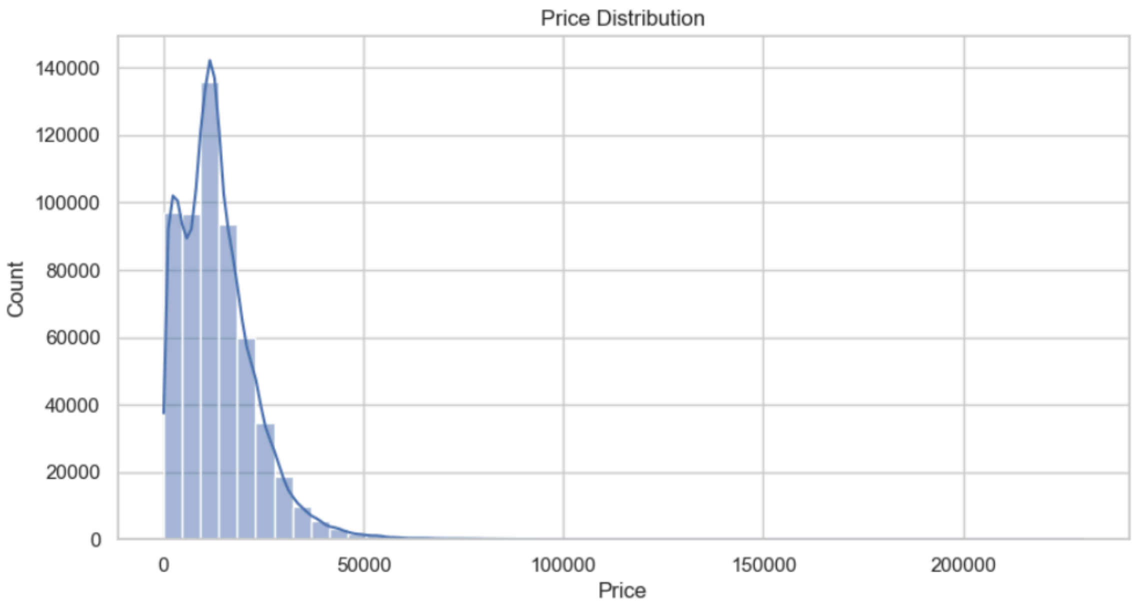
Python Code for Data Cleaning

Below are some of the Python scripts used for cleaning the data:

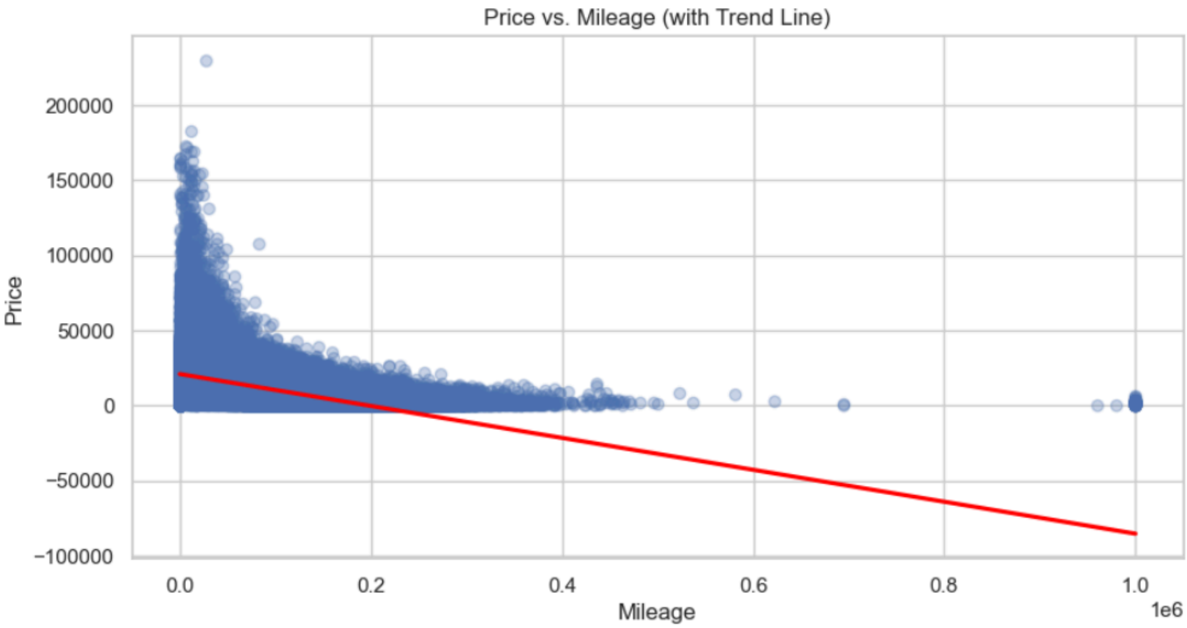
```
import pandas as pd

# Load dataset
data = pd.read_csv('vehicle_sales_data/car_prices.csv')

# Convert 'saledate' to datetime
data['saledate'] = pd.to_datetime(data['saledate'], errors='coerce')

# Bin 'condition' into categories
def categorize_condition(value):
    if value >= 4.5:
        return 'Excellent'
    elif value >= 3.5:
        return 'Good'
    elif value >= 2.5:
        return 'Average'
    elif value >= 1.5:
        return 'Poor'
    else:
        return 'Very Poor'

data['Condition Category'] = data['condition'].apply(categorize_condition)

# Remove null values
data = data.dropna()

# Save cleaned data
data.to_csv('cleaned_vehicle_sales_data.csv', index=False)
```

Exploratory Data Analysis (EDA)

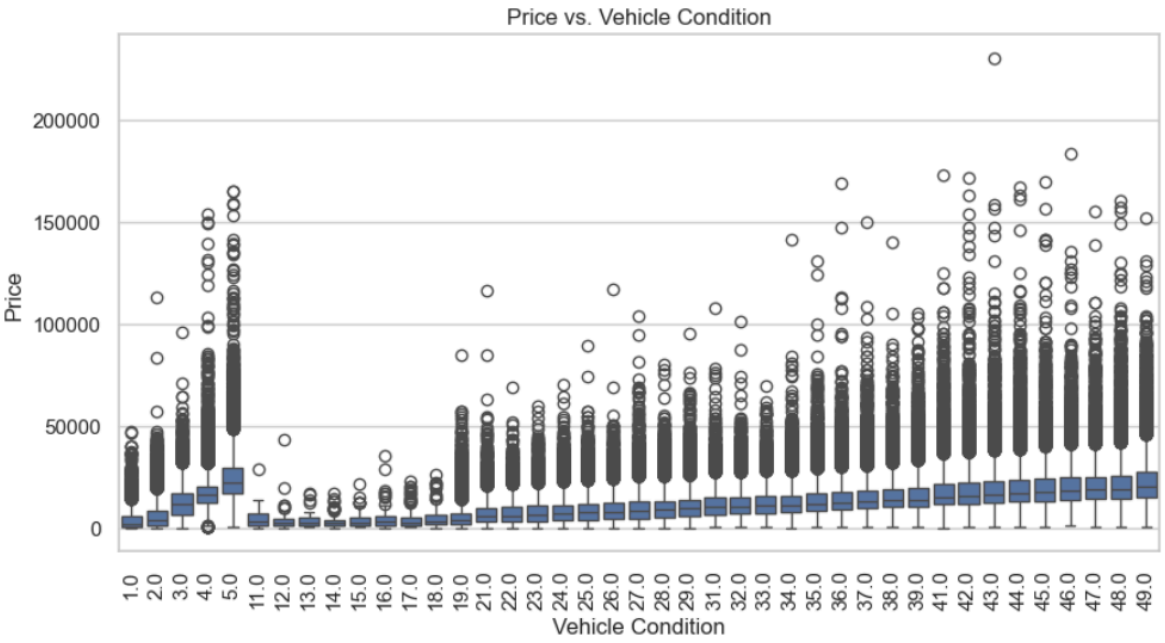
Key Findings



1. **Price Distribution**:
    - ○ The selling prices ranged from a few hundred dollars to over $200,000.
    - ○ Most vehicles were sold within the $5,000 to $20,000 range.



2. **Odometer and Price Relationship**:
    - ○ Vehicles with lower mileage generally sold for higher prices, highlighting the importance of mileage as a key determinant of value.

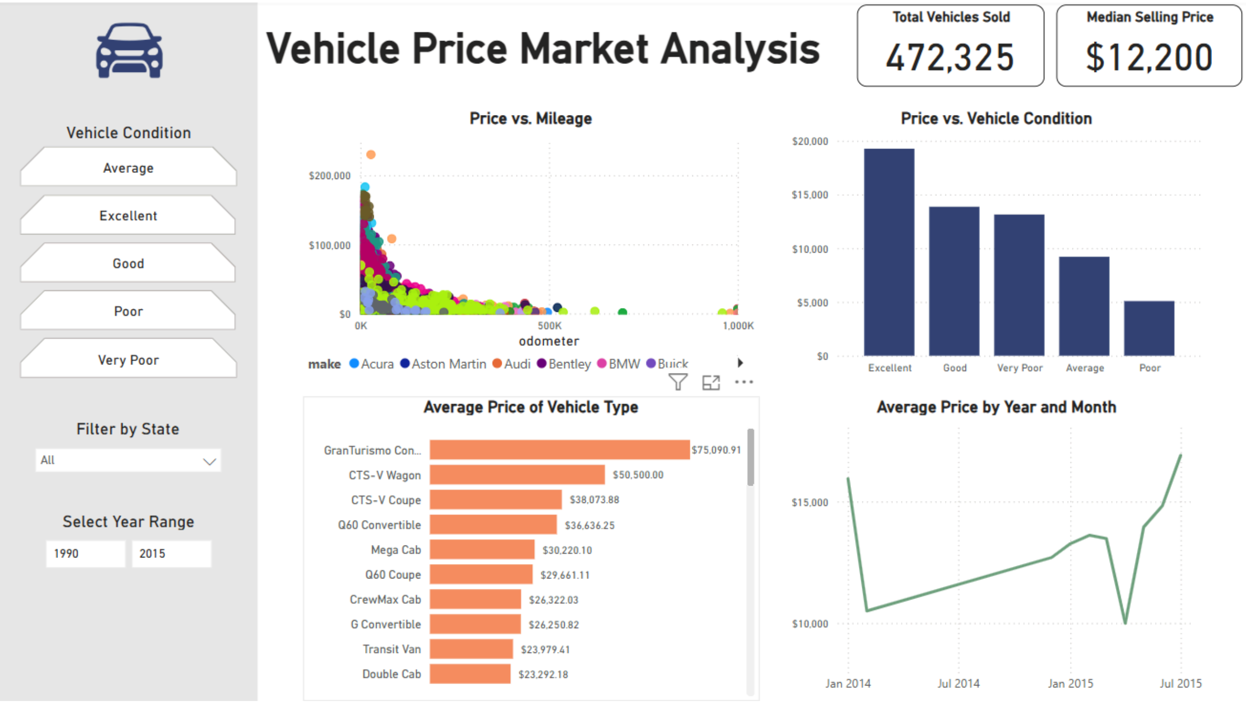


3. **Condition and Price Relationship**:
   - Vehicles in "Excellent" condition had significantly higher average selling prices compared to those in "Poor" or "Very Poor" conditions.
4. **Year and Price Trends**:
   - Newer vehicles (manufactured in recent years) commanded higher prices.
5. **Vehicle Brands**:
   - Luxury brands such as Aston Martin and Bentley consistently had higher average prices compared to mass-market brands like Ford and Toyota.

Python Code for EDA

```
import matplotlib.pyplot as plt
import seaborn as sns

# Price distribution
plt.figure(figsize=(10, 6))
sns.histplot(data['sellingprice'], bins=50, kde=True)
plt.title('Price Distribution')
plt.xlabel('Selling Price')
plt.ylabel('Frequency')
plt.show()

# Scatterplot: Price vs. Mileage
plt.figure(figsize=(10, 6))
sns.scatterplot(x=data['odometer'], y=data['sellingprice'], hue=data['Condition Category'])
```

```
plt.title('Price vs. Mileage')
plt.xlabel('Odometer')
plt.ylabel('Selling Price')
plt.legend(title='Condition Category')
plt.show()

# Average price by condition
avg_price_by_condition = data.groupby('Condition Category')['sellingprice'].mean()
print(avg_price_by_condition)
```



Total Vehicles Sold 472,325

Median Selling Price $12,200

Filter by State: All

Select Year Range: 1990 2015

Power BI Dashboard

Visualizations and Insights

1. **Price vs. Mileage (Scatterplot)**:
   - A negative correlation was observed, indicating that vehicles with lower mileage tend to have higher prices.
2. **Price vs. Vehicle Condition (Column Chart)**:
   - Vehicles in "Excellent" condition had the highest average prices, while those in "Poor" condition had the lowest.

3. **Average Price of Vehicle Type (Horizontal Bar Chart)**:
    - The GranTurismo Convertible was the most expensive vehicle type, with an average price exceeding $75,000.
4. **Average Price by Year and Month (Line Chart)**:
    - A noticeable seasonal trend was observed, with prices peaking during mid-year periods.
5. **Price Distribution & Outliers (Box Plot)**:
    - Most vehicles fell within the $5,000 to $20,000 range, but outliers such as luxury and sports vehicles were identified with prices exceeding $200,000.
6. **Total Vehicles Sold & Median Selling Price (Cards)**:
    - Total Vehicles Sold: 472,325
    - Median Selling Price: $12,200

Slicers

- **Vehicle Condition**: Filter by condition categories (Excellent, Good, Average, Poor, Very Poor).
- **State**: Narrow down results by the state of sale.
- **Year Range**: Select a range of manufacturing years.

Conclusion

The analysis provided valuable insights into the automotive market, showcasing the impact of factors like condition, mileage, and brand on vehicle prices. The findings can be used by dealerships to optimize pricing strategies and inventory management.

Future work includes implementing predictive models to forecast vehicle prices based on these insights.

Appendix

For further details, please refer to the full Python scripts and Power BI dashboard file included in the project repository.