**Report on Car Prices : Insights from Diverse Automotive Data**

**Abstract**

The car dataset analysis aims to deepen the insights gained from data by incorporating secondary data, advanced analytics, and additional research methodologies. This report dives deeper into the factors impacting car prices while employing sophisticated data analysis tools to draw more comprehensive conclusions.

**Introduction**

The automotive industry's dynamism necessitates an in-depth understanding of the factors influencing car prices. This project builds upon the foundation laid in Project #1, aiming to augment the analysis with secondary data, advanced statistical techniques, and a refined approach to unravel the complexities governing car pricing dynamics.

**Research Objectives**

1. To augment the understanding of factors influencing car prices.
2. To employ advanced statistical methods for more robust insights.
3. To draw refined conclusions facilitating informed decision-making in the automotive sector.

**Methodology**

1. **Data Collection and Integration**: Incorporating additional datasets, market trends, and economic indicators related to the automotive industry to supplement the existing dataset.
2. **Data Cleaning and Preparation**: Extensive data cleaning to rectify inconsistencies, handle missing values, and ensure compatibility among integrated datasets.
3. **Exploratory Data Analysis (EDA):** Utilizing descriptive statistics, correlation analysis, and advanced visualization techniques (scatter plots, heatmaps, etc.) to identify patterns and relationships within the extended dataset.
4. **Advanced Analytics**: Implementing regression analysis, machine learning models (such as decision trees or random forests), and predictive modeling to forecast car prices based on a multitude of variables.
5. **Hypothesis Testing**: Formulating and testing hypotheses regarding the impact of specific features (e.g., engine type, market category) on car prices.

Summary of Dataset:

The dataset contains 11914 entries and 16 columns, including both numerical and categorical data.

A black text with black letters

Description automatically generated with medium confidence

Columns like Engine Fuel Type, Engine HP, Engine Cylinders, Number of Doors, and Market Category contain missing values.

A screenshot of a computer

Description automatically generated

The MSRP (Manufacturer's Suggested Retail Price) ranges widely from $2,000 to over $2 million.

Outliers were identified in various columns such as Engine Cylinders, Engine HP, MSRP, Year, city mpg, and highway MPG using the Interquartile Range (IQR) method.

Information related to Dataset like datatypes

A screenshot of a computer

Description automatically generated

**The Statistics of the dataset are as follows :**

A screenshot of a computer screen

Description automatically generated

**Missing Values and NaN Values of Data are as follows :**

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated A list of text on a white background

Description automatically generated

**VISUALISATIONS:**

Numerical vs Numerical

**A collage of blue and white lines

Description automatically generated**

# Numerical vs Categorical:

##### Barplot: Transmission Type vs MSRP

**A graph of different colored bars

Description automatically generated**

##### **Boxplot: Engine Fuel Type vs MSRP for outliner Analysis**

**A graph of fuel type

Description automatically generated**

# Categorical vs Categorical:

# Countplot: Vehicle Style by Drivetrain

**A graph of a vehicle style

Description automatically generated**

# Numerical and Categorical vs Categorical:

##### Barplot: Engine Cylinders & Vehicle Size by Market Category

**A chart of engine cyclonics

Description automatically generated with medium confidence**

# Two Numerical vs One Categorical:

##### Scatterplot: Engine HP, Highway MPG by Vehicle Style

**A graph of a number of cars

Description automatically generated with medium confidence**

# CORRELATION HEATMAP

**A screenshot of a graph

Description automatically generated**

**Test Results**:

1. **Correlation Analysis:**

* There's a moderate positive correlation (0.66) between Engine HP and MSRP, indicating that higher engine horsepower tends to be associated with higher prices.

A close-up of numbers

Description automatically generated

1. **ANOVA Test:**

* ANOVA test results suggest a significant variation in MSRP among different Vehicle Styles (F-statistic: 92.56, p-value: 5.61e-271), indicating that the Vehicle Style significantly affects the MSRP.

1. **T-tests:**

* T-tests comparing MSRP for different Engine Fuel Types reveal significant differences in MSRP for specific fuel types (e.g., premium unleaded (required), regular unleaded, flex-fuel, etc.).

A screenshot of a computer

Description automatically generated

1. **Chi-square Test:**

The Chi-square test indicates a strong association between Vehicle Style and Market Category (Chi-square value: 27100.05, p-value: 0.0), suggesting that the Vehicle Style is not independent of the Market Category.

**Fitting the OLS Simple Linear Regression Model :**

**A screenshot of a computer

Description automatically generated**

**Conclusion** :

The dataset, which includes 16 columns and 11914 items, reveals important factors influencing car prices. In particular, missing data in columns like Engine HP and Engine Fuel Type should be closely examined. The broad MSRP range—from $2,000 to more than $2 million—highlights the variance of prices. Overall conclusions may be impacted by anomalies in engine cylinders, horsepower, MSRP, year, city MPG, and highway MPG.

Important tests found strong relationships between price and engine horsepower (0.66 correlation), vehicle style and MSRP changes, and different engine fuel types and pricing. The market category and vehicle style are interdependent, as demonstrated by the Chi-square test. Car pricing methods depend on an understanding of these influences, which include engine horsepower, vehicle style, and engine fuel type. Robust assessments therefore require the analysis of outliers.