**Course Work Project Description and Rubric**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Semester** | **202420** | | **Division** | | CIS |
| **Assessment title in Syllabus** | **Project** | | **Program** | | **IT and IS** |
| **1** |  | |  | |  |
| **Course Code** | **CIS 2423** | | | | |
| **Course Title** | **Programming for Data Analytics** | | | | |
| **CLOs** | **All CLOs** | | **Accreditation Body** | | **CAA & CIPS** |
| **Course Instructor** |  | | **CRN** | |  |
| **Assessment Weight** | **40%** | | **Submission Date** | | **Week 14** |
| **For Group Work submissions, an additional individual assessment will be conducted.**  **Grades for the students in one group will vary based on the individual performance in the additional assessment.** | | | | | |
|  | | | | | |
| **Student Declaration**:  **Academic Integrity Statement**  In accordance with the HCT Academic Integrity Policy  • Students are required to refrain from all forms of academic integrity breaches as defined and explained by HCT.  • A student found guilty of having committed acts of academic integrity breach(es) will be subject to the relevant sanctions as outlined by HCT.  إفادة النزاهة الأكاديمية  **وفقًا لسياسة كليات التقنية العليا للنزاهة الأكاديمية**  **• على الطلبة الإلتزام بلوائح وقواعد النزاهة الأكاديمية، كما هو مبيّن وموضح في السياسات والإجراءات الخاصة بكليات التقنية العليا.**  **• في حالة ارتكاب الطالب أي شكل من أشكال الإخلال بالنزاهة الأكاديمية، سيتعرض الى العقوبات الموضحة في السياسات ذات الصلة.**  This assignment is entirely my own work except where I have duly acknowledged other sources in the text and listed those sources at the end of the assignment.  I have not previously submitted this work to the HCT, or any other entity. I understand that I may be orally examined on my submission.  **Student (s) Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**  **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | | | | | |
|  | | | | | |
| **Student Name(s):** |  |  | |  | |
| **Student HCT ID(s):** | H00 | H00 | | H00 | |

**For Examiner’s Use Only**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Group (50%)** | | | | | **Individual (50%)** |  |  |
| **C**LO | **1** | **2** | **3** | **4** | **Report Formatting** | **Oral Defense** | **Total** | **%** |
| **Marks Allocated** | 10 | 10 | 42 | 26 | 12 | **50** | **100** | **4**0 |
| **Marks Obtained** |  |  |  |  |  |  |  |  |

# Deliverable 1: Purpose & Design Report

1. **Define the purpose of data analysis for the chosen dataset.**

This dataset is related to the car prices, where we can perform the analysis to predict the car price, and we can understand which attributes increase the car price and which one’s negative effect on the side to decrease the price of the car. So the main goal of the analysis is the dataset to understand the factors that affect car price and create a model that helps us to predict the accurate car price based on its features. Our analysis helps the busy of the car and also the seller of the car. Our model will inform them of the actual price of the car so that they can decide whether they will buy/sell the car or not. The analysis of this dataset also helps the businessmen who are running car showroom businesses and buy the used cars or companies that provide loans for used cars. They can use our model to get an accurate estimation of car prices.

1. **Identify and Justify the type of programming used for data analysis.**

We select the Python language to use for this data analysis. Python has easy syntax. We just focus more on the analysis of data as compared to how to write code for logic. Python language has many libraries for data analysis such as Pandas, Numpy, Matplotlib, and many more to analysis of data. It easily fetches data from our dataset file, and it can handle large datasets, so we can easily perform analysis on it.

1. **Identify the type and purpose of the machine learning algorithm to be implemented for the chosen dataset.**

To analyze this data, we can apply both supervised and unsupervised learning algorithms.

**Supervised Learning:**

In the regression model, we can apply linear regression and also non-linear algorithms for the price prediction of cars.

In the classification model, we can use logistics regression, K-nearest neighbors, naïve bayes, and decision tree algorithms to categorize the car prices or condition of a car.

**Un-supervised Learning:**

In unsupervised learning, we can use K-Mean and Hierarchical clustering to identify the car groups or car price decreasing ratio according to car model or any other attribute etc.

We can evaluate each algorithm based on accuracy. We can identify the algorithms which perform well on our dataset.

1. **Identify and Justify the independent and dependent variables for the chosen dataset.**

**Dependent Variable**:

**Price:**

Price is the target variable. The value of this variable we want to predict that’s why its depends on other variables.

**Independent Variables:**

In our dataset, we have the following independent variables.

**Brand:**

Band is used to identify the price of the car according to car type. Each brand has its rates.

**Model:**

Car model also affects the car price. The latest model car price is high, and the old model car price is low. So the car price also depends on the car model.

**Year:**

The year variable shows the age of the car. Older cars are cheap.

**Engine-Size:**

The large engine car indicates high performance and cost.

**Fuel-type:**

Different fuel types have different cost effects. Nowadays, people prefer electric and hybrid cars; that’s why electric car prices are high compared to petrol and diesel.

**Transmission:**

Transmission is also affects on price. Automatic car price is high as compared to manual cars.

**Mileage:**

Which car has high mileage has low value.

**Doors:**

Number of door indicate the car type. Its sedan, hash back or SUV.

**Owner\_Count:**

The owner count show how many peoples used this car. Low owner count shows the high values.

# Deliverable 2: Data Descriptive Analysis

1. **Justify why you want to perform the descriptive analysis for the chosen dataset.**

The descriptive analysis is the essential part of understanding any dataset. Our dataset is related to car prices, so descriptive analysis can allow us to identify and understand the following things, which justify the reason why we need to perform the descriptive analysis.

* It can help us to identify the trends and patterns in car price, age, mileage, and other numerical features.
* It helps to understand the key variables of the dataset. Our data has multiple features, such as car brand, model, year, mileage, fuel type, transmission, and price, so we can apply descriptive statistics like mean, median and mode, etc, that summarize the variables and highlight central tendencies and variability.
* The descriptive analysis can reveal the relationship between price, age, mileage, and brand. How car price changes according to other variables. It can show which brand of car is the most expensive. It also shows transmission and fuel type effects on price.
* Through descriptive analysis, we can find the most expensive and cheapest cars.
* It also has visualization tools like box plots, bar charts etc., that show the data distribution in the dataset. we can easily understand price distributions through visualization, which helps the car dealers to make decisions in inventory management and purchasing cars.

We just highlighted some of the reasons why we needed the descriptive analysis, but through that, we easily understood the importance of descriptive analysis.

1. **Create a script to develop a Python function for descriptive statistics. The input for the function should be the sample and the field to perform the descriptive statistics.**

We created the following function:

def descriptive\_stats(sample, field):

stats = {

'Count': sample[field].count(),

'Mean': sample[field].mean(),

'Standard Deviation': sample[field].std(),

'Minimum': sample[field].min(),

'25th Percentile': sample[field].quantile(0.25),

'Median': sample[field].median(),

'75th Percentile': sample[field].quantile(0.75),

'Maximum': sample[field].max()

}

return pd.Series(stats)

1. **Create a program to random sampling of size 150 and find the descriptive statistics for the dependent variable from the sample [Apply the descriptive function which you created].**

We create the following program that find required output with help of last part function:

random\_sample = df.sample(n=150, random\_state=42)

random\_stats = descriptive\_stats(random\_sample, 'Price')

print("Random Sample Descriptive Statistics:\n", random\_stats)



1. **Create a script for systematic sampling by giving certain conditions and finding the desc stat for the dependent variable from the sample [Apply the descriptive function which you created].**

#Filter condition we used Mileage < 100000

filtered\_df = df[df['Mileage'] < 100000]

#Systematic sampling on filtered data

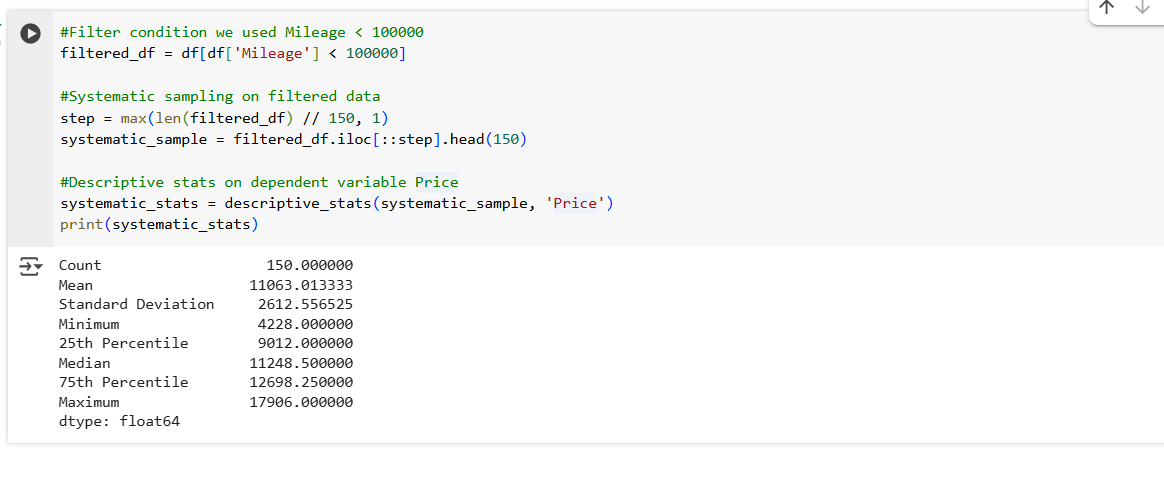
step = max(len(filtered\_df) // 150, 1)

systematic\_sample = filtered\_df.iloc[::step].head(150)

#Descriptive stats on dependent variable Price

systematic\_stats = descriptive\_stats(systematic\_sample, 'Price')

print(systematic\_stats)



1. **Create a detailed descriptive statistics report about the dependent variable of the chosen dataset.**

The price is the dependent variable in our dataset which represent the resale car price. Below code and output show the summery of detail statistical analysis of price on entire dataset.

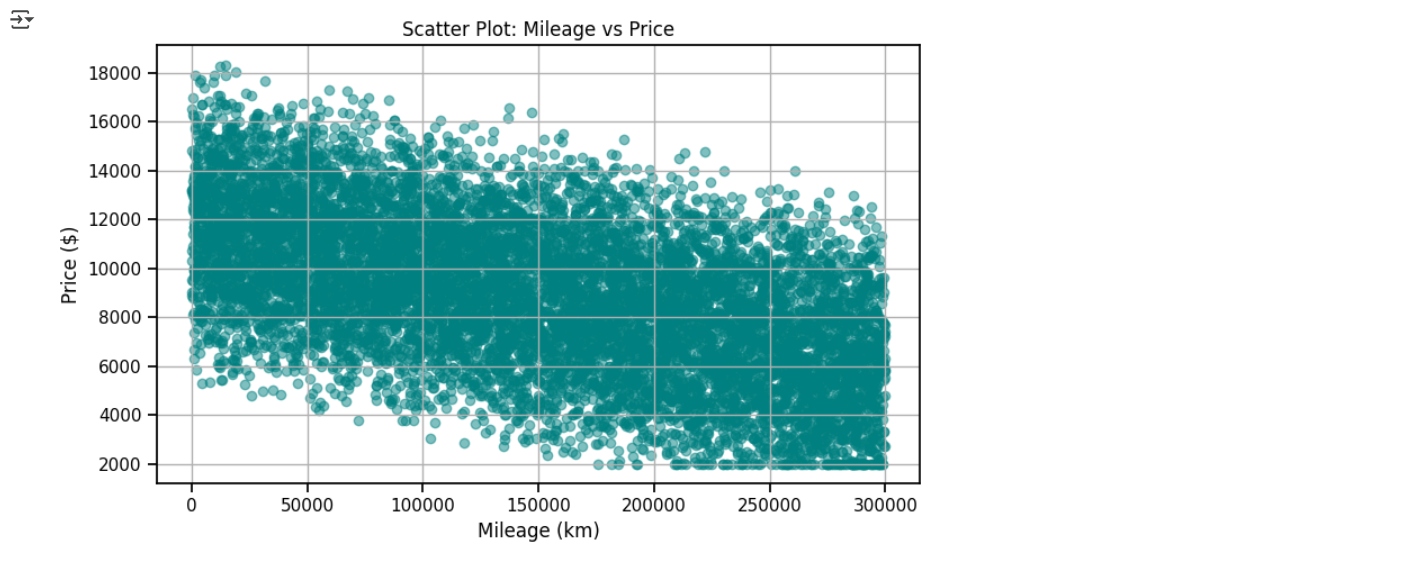


Here the mean is 9953.96 and median is 8858.50 that show the proper distribution. We can see the arrange of car prices from maximum and minimum values.

1. **Visualize the dependent variable by the Graph/Chart of the following using Python Program:**
   1. **Scatter plot**

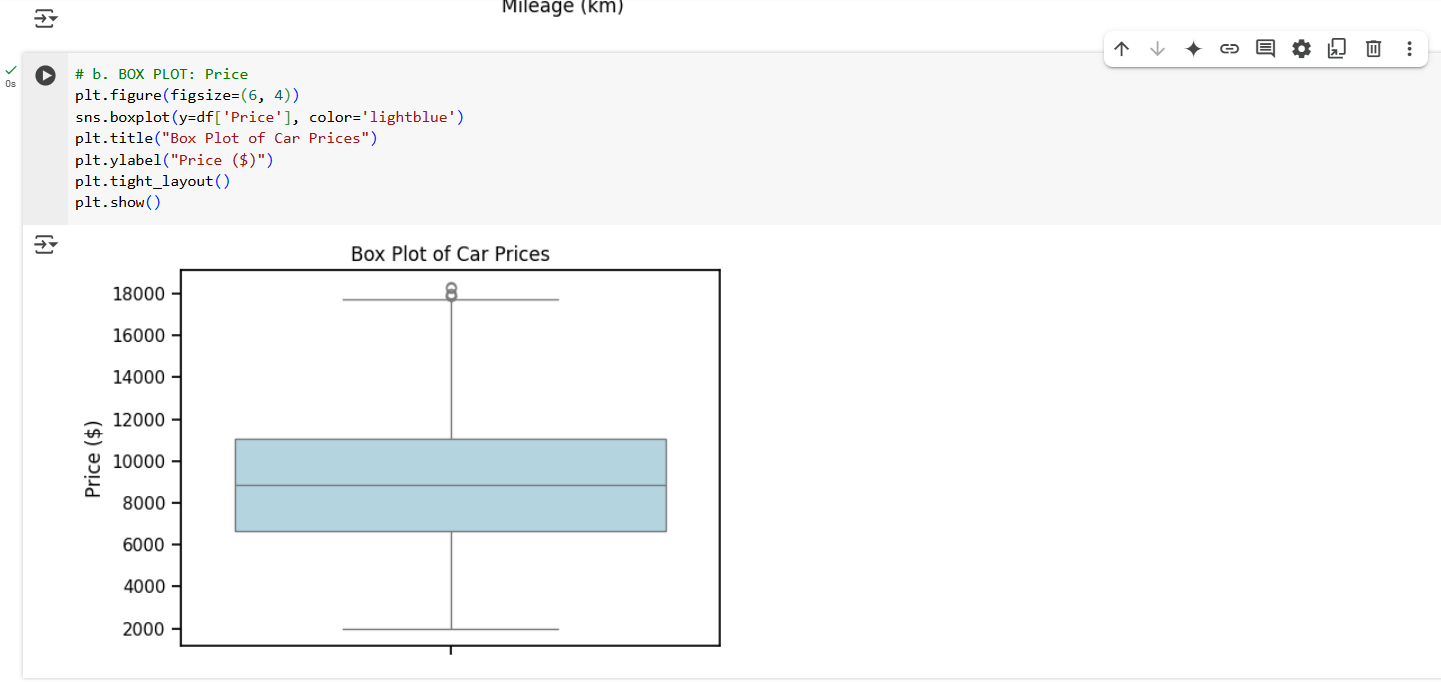
We represent how mileage effects on car price. We display it on scatter plot.





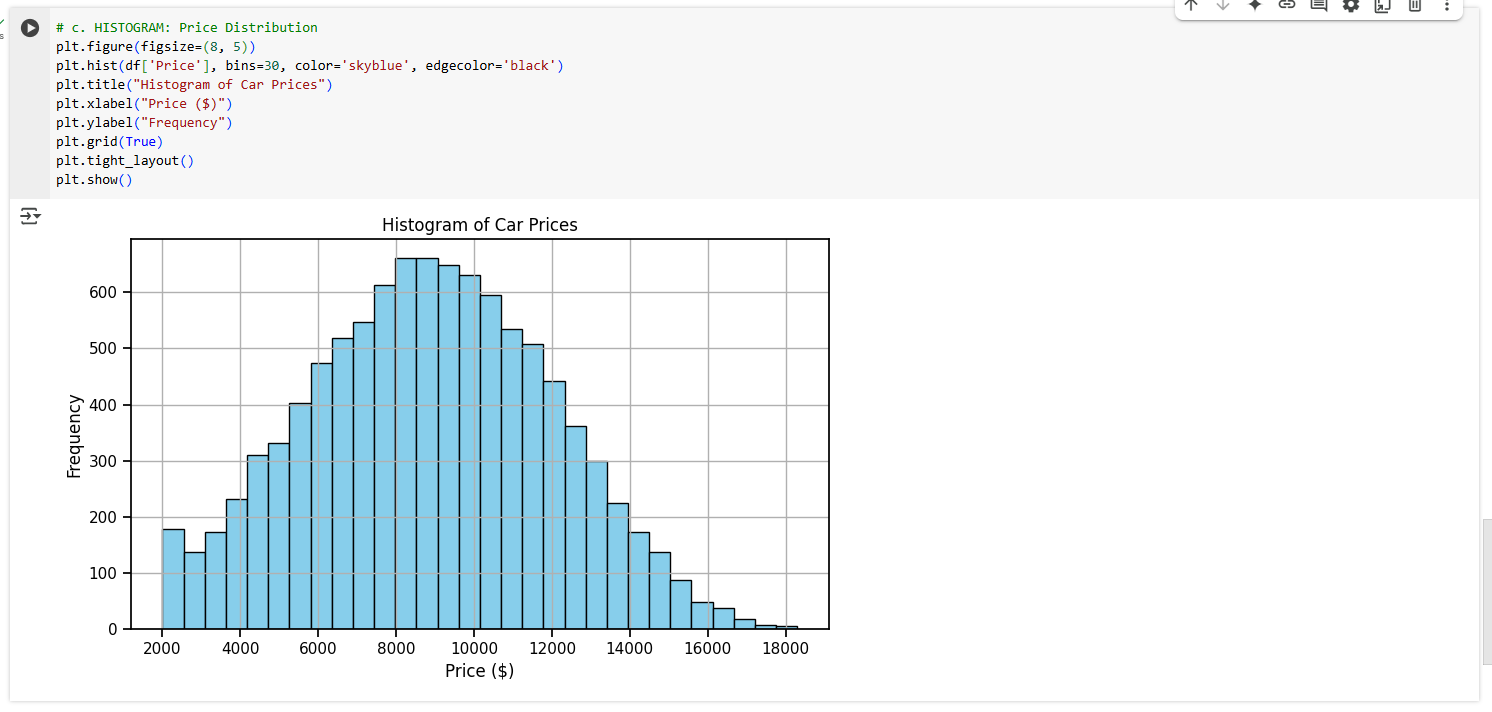
* 1. **Box Plot**

In box plot we represent the distribution of price.



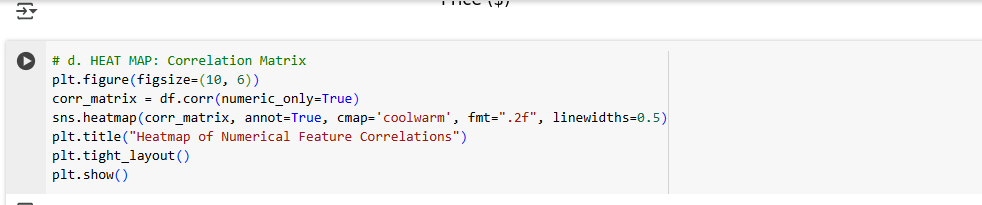
* 1. **Histogram**

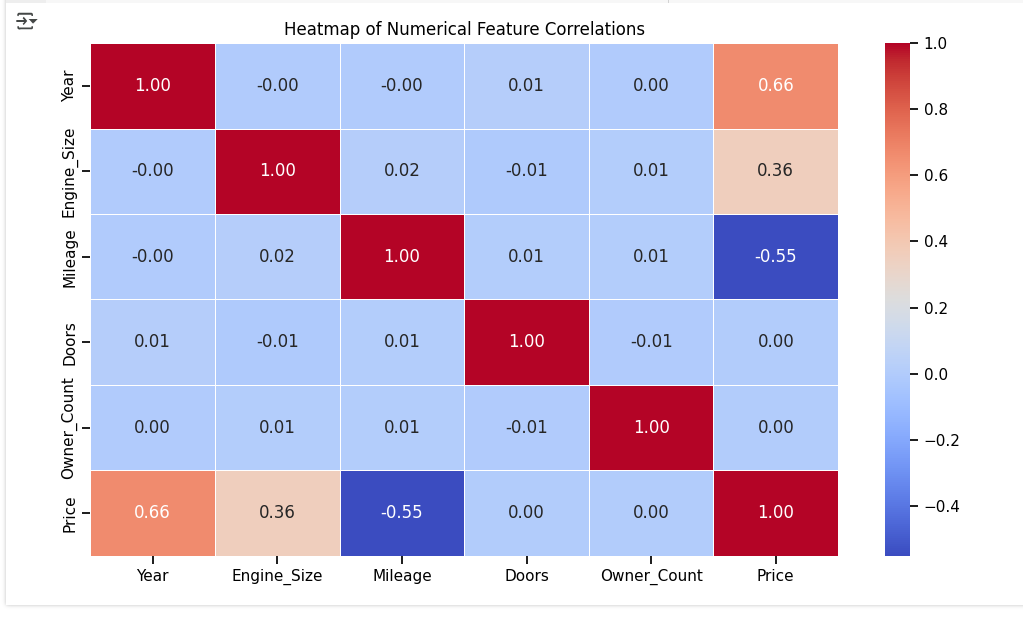
In a histogram we represent the frequency of different prices.



* 1. **Heat Map**

Heat map show the all correlation coefficients between all numerical values.





1. **Perform the hypothesis test to find the correlation (Pearson and Spearman for numerical variable and chi-square test for categorical variable) between the independent variable and the dependent variable.**

We have price dependent variable and more than one independent we have in this dataset. To perform the Pearson and Spearman tests on variables, we set the following hypothesis.

We find the relationship between price dependent variable and mileage independent variable.

**Goal:** Determine if there is a statistical connection between car price with its mileage.

**Null Hypothesis (H₀):**

There is no correlation between Mileage and Price.

**Alternative Hypothesis (H₁):**

There is a correlation between Mileage and Price.



Here Pearson Correlation Coefficient: -0.5512

Correlation Coefficient: -0.5512

Both have p-value: less than 0.001

That means the chance of a random pattern is very low. So we can say there is very strong negative relationship between mileage and price. If the mileage is increased, than car’s price will drop accord to mileage.

Now, we want to apply the chi-square test. We need the categorical independent variable to apply this test. We use fuel type to check whether it effects on price or not. We need to convert the price into price group (low and high), then we can apply this test.

**Goal:** Test whether fuel-type effects car price or not.

**Null Hypothesis (H₀):**

There is no association between Fuel Type and Price Group.

**Alternative Hypothesis (H₁):**

There is an association between Fuel Type and Price Group.



We convert the price into low, high groups, then apply the test where we found the P-value is less than 0.001, so we reject the null hypothesis and we can say there is a statistically significant relationship between Fuel Type and Price.

1. **Assess the performance of the dependent variable to know whether the sample is representative of the normal population by a one-sample t-test.**

**Goal:** We want to check whether the sample mean of car prices is statistically different from the dataset’s known means or not.

**Null Hypothesis (H₀):**

The average car price in the sample is equal to the known dataset average.

**Alternative Hypothesis (H₁):**

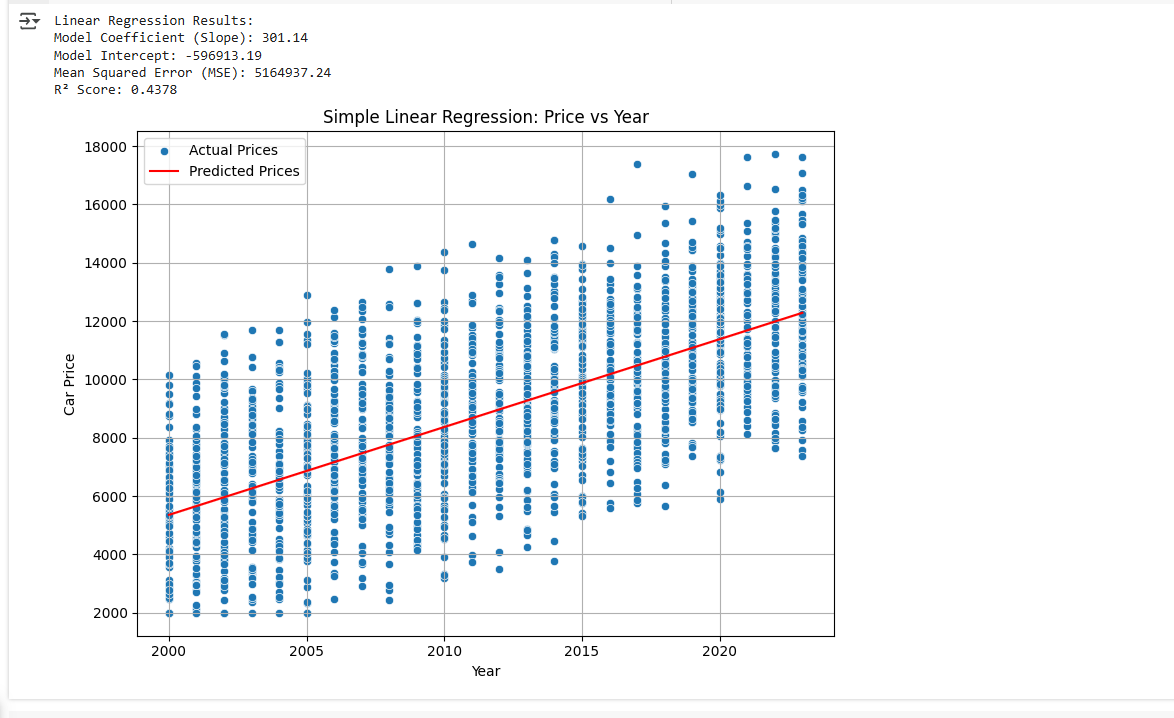
The average car price in the sample is not equal to the known dataset average.



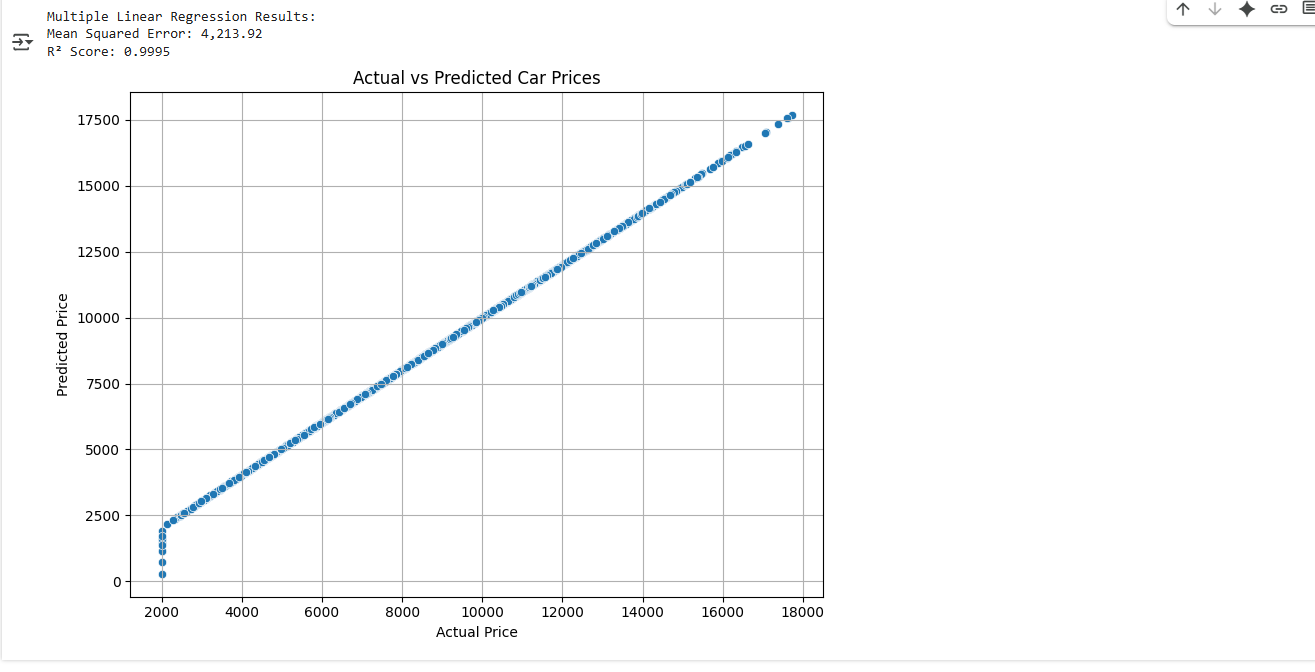
Here the P-values =0.4242 which is much greater than 0.05 that means we failed to reject null hypothesis. So we can say that the sample of 150 car prices is statistically represent the full dataset.

# Deliverable 3: Classification and Clustering Algorithm implementations

1. **Build, Train, Develop and Evaluate using Simple Regression for chosen dataset.**



1. **Develop a script to forecast the value of the dependent variable from all the relevant independent variables using Multiple Linear Regression.**



1. **Predict the value of the dependent variable from the different classifier such as Logistic Regression, KNN, Naïve-Bayes and Decision Tree.**



1. **Evaluate the performance of each model using confusion matrix and accuracy and identify the best fit classifier for the chosen dataset.**

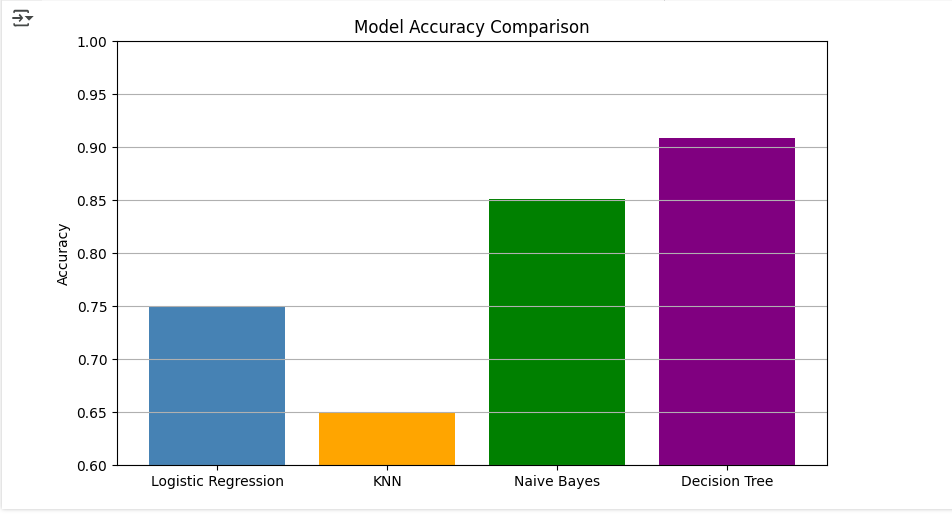
As we see in the last part implementation, we get the following confusion matrix information:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **TP** | **TN** | **FP** | **FN** |
| Logistic Regression | 767 | 734 | 264 | 235 |
| KNN | 673 | 628 | 358 | 341 |
| Naïve Bayes | 879 | 823 | 152 | 146 |
| Decision Tree | 937 | 881 | 94 | 88 |

The accuracies information are:

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| Logistic Regression | 0.7505 |
| KNN | 0.6505 |
| Naïve Bayes | 0.8510 |
| Decision Tree | 0.9090 |

The visualization comparison of algorithms accuracies are:

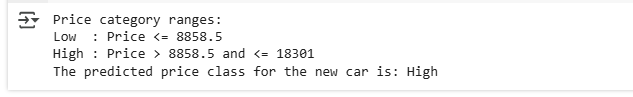


Based on all evolutions, the decision tree has the highest accuracy, almost 90.90%, and the confusion matrix also shows the balance between TP and TN. So the decision tree classifier is the most suitable and best for car price prediction.

1. **Predict the dependent variable by using best-fit classifier.**

To predict the price, we select following information to predicts its price:

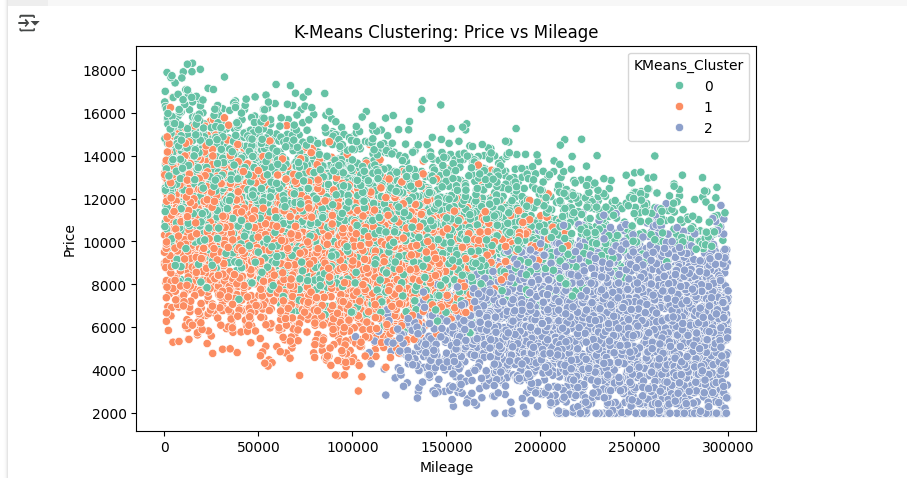
Brand: Toyota, Model: Corolla, Year: 2018, Engine\_Size: 1.6, Mileage: 45000, Doors: 4, Owner\_Count: 2, Fuel\_Type: Petrol, and Transmission: Manual. So we got the following output:



1. **Perform the cluster analysis such as K-means and Horizontal for any field from the chosen dataset.**

We can apply the cluster analysis on numerical data so we select only numerical field from dataset such as mileage, engine\_size and price. We create clusters groups based on these fields.

Its create three groups output diagram are below:



To test the clustering algorithm, we create a car with Mileage: 45000, and Engine\_Size: 1.6. When we test it on our algorithm, it gives the following output:



So the algorithm predicts its cluster 2 and price is 6,129.41.

1. **Explain the strategy for improving the system after viewing the cluster diagram.**

After implementing clustering in visualization, we recognize that it creates three clusters from our dataset. The first cluster 0 represents the high-mileage and low-price cars, second cluster 1 represents the medium mileage and medium price, and the last cluster 2 shows the premium high-quality, low-mileage cars. Everyone can use this to find the price prediction of any car, but as we see in the diagram the clusters some points are overlapping it shows that some features are not good to make proper separation between clusters. But when we test the prediction, it gives a good prediction on unseen data. So we decided to implement Hierarchical Clustering with dendrogram to visualize the grouping. We use it to display the data in the group of tree structure, which can help us to separate the three clusters. The hierarchical clustering builds the different levels of each cluster that provide the grantee to show each cluster separately. When we implement this on our dataset, it gives us a good results. Output are below:

