Best Used Cars for Car Sharing Business

Team 5

Umutbek Abdimanan uulu - MSCS, City University of Seattle

abdimananuuluumutbe@cityuniversity.edu

Asia Shvets - MSCS, City University of Seattle

shvetsasia@cityuniversity.edu

Osama Khan - MSCS, City University of Seattle

[khanosama@cityuniversity.edu](mailto:khanosama@cityuniversity.edu)

**Abstract**

This data acquisition and analytics project aims to clean, pre-process, wrangle and explore a dataset from Kaggle which is used to determine the best used cars for a car-sharing business based on various factors such as reliability, safety ratings, and affordability. The project involves the acquisition of a dataset of used cars and their attributes, followed by data cleaning and processing to remove any inconsistencies, errors, or missing values. Exploratory data analysis will be performed to gain insights into the data and identify any patterns or trends. Data visualization techniques will be used to present the findings in a clear and concise manner. The project will help car-sharing businesses make informed decisions on the types of cars to purchase for their fleet, which will ultimately lead to better customer satisfaction and profitability.

**Keywords:** Data acquisition, Data pre-processing, Data exploration, Data visualization, Car sharing business, Used cars.

1. **INTRODUCTION**

The purpose of this project is to highlight how data visualization affects informed decision making and ease of access to data that might have trends that are not apparent briefly. The first step of the project will include cleaning of unprocessed datasets from Kaggle. The preprocessed data acquired from Kaggle will be processed according to the projects needs and this processed will later be used to produce visual representation of trends and display information in an organized fashion that will make it easy to understand to not only anyone who uses a computer proficiently but to anyone who needs it. In this case customers and used car owner company or personnel.

Tools used in the scope of this project will be discussed later as the project progresses.

A brief background to the problem that the team intends to address in the project is to make it more convenient for customers and business to search for cars to purchase and sell. A business like a rental company can contract with these car dealerships and if they have cars available that are to be sold. This would reduce the time to search a car for a customer plus customers would have a guarantee that rental company serves as an excellent place to purchase a used vehicle due to their comprehensive insurance coverage and meticulous post-rental care. The product of this project has the capability to merge the interest of two industries (Car rental companies and Used car dealerships) who can collaborate effectively by sharing their data and other resources.

Within the scope of this project, our team focused on preparing the data to establish a relationship between the information required for the sale, purchase, and rental of these cars. The parties involved in this case would be a used car company and a rental company.

To achieve this goal, our team followed a systematic approach that began with data gathering from sources such as Kaggle. The subsequent step involved data preprocessing, cleaning, exploration, and wrangling, although not necessarily in that specific order. Lastly, a visualization tool was utilized to import the data and visually analyze different aspects of each dataset independently. This analytical process aimed to provide valuable insights and diagnostics that could facilitate the visualization of the data. Using this tool, we intend to make it possible to come up with well-informed decisions regarding the integration of both datasets, but also offer individual business owners’ insights into their performance in terms of reliability, safety ratings, and affordability.

1. **LITERATURE REVIEW**

Describe related problems and where your problem is situated. What is significant about your problem compared to existing ones? Two papers or book chapters are required per student.

1. **DATA**

For this project the team selected two datasets:

* used\_cars\_dataset (*Used Car Price Prediction,* n.d.)
* Cornell\_car\_rentals\_dataset (*Cornell Car Rental Dataset,* n.d.).

Both of them were located in Kaggle datasets, an online community platform for data scientists and machine learning enthusiasts.

Datasets contain information about names of the cars, year of production, fuel type, engine, transmission, price, mileage and other characteristics of the cars.

The two selected datasets have the same base as they both are related to cars but differ in their application and so they have different entities. The reason for selecting these similar datasets get cleared when we think of both industries as an intwined business model.

1. **DATA PREPROCESSING**

Data preprocessing is a part of data mining that takes raw data and processes it so that it is formatted in such a way that the computer can understand it. Let's say the data in the form of pictures or scattered or unorganized (numbers and alphabets or characters have not significant distinction). This step would deal with these irregularities. This is also important to understand that if the data is disorganized or messed up in any way then it will yield inefficient results. “Garbage-in garbage-out" is a terminology used to better describe this scenario. (*What Is Data Preprocessing & What Are the Steps Involved?*, 2021).

From this stage onwards the process in pretty much simple and straight forward. The team used Visual Studio to preprocess the datasets individually. The steps included in the preprocessing within the scope of this course are explained in the following figures.

1) Importing *pandas* library. Pandas is a library in python that has functions that are used for cleaning, manipulating and exploring datasets.



Figure 1: Importing pandas module

2) In figure 2 *df* represents a data frame that reads the ‘used\_cars.csv’ file and save it in memory to be worked on.



Figure 2: Importing dataset

3) In Figure 3 *df.head()* shows the first few rows of *df*. *df.info()* function is used to show further information about the dataset. For example, the column names and data types. The output of this is shown in Figure 4.

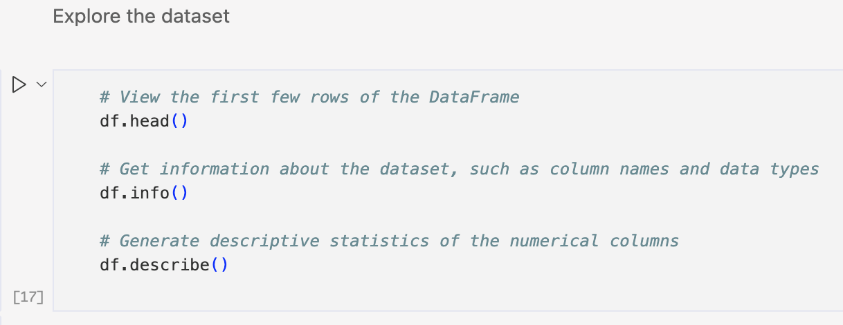


Figure 3: Exploring dataset

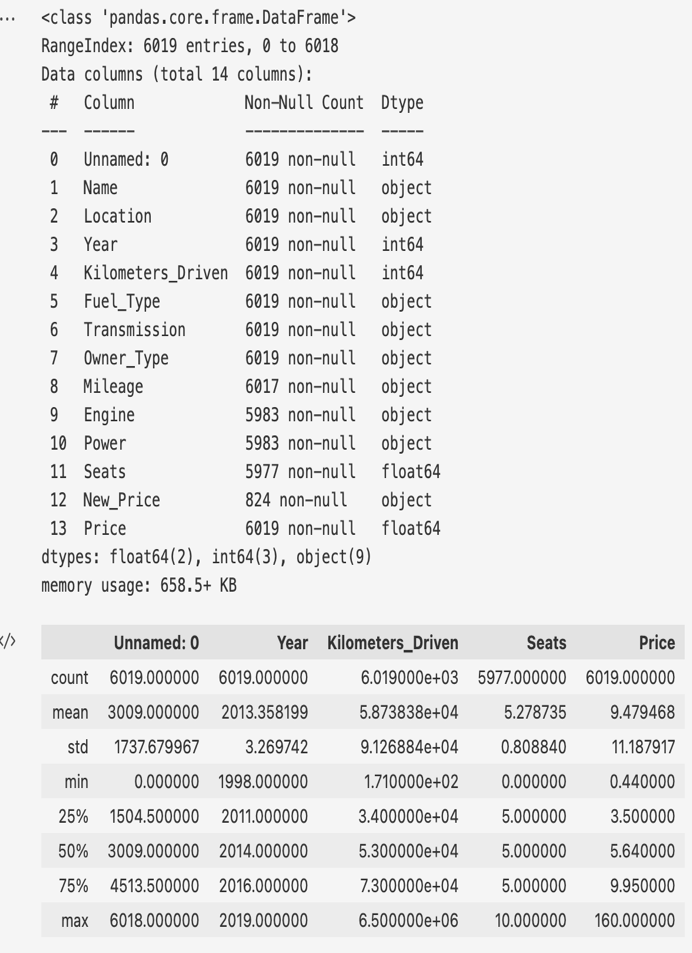


Figure 4: Output for *df.head, df.info* and *df.describe*

4) There are always some missing values and anomalies that need treating. For this we used functions like *df.isnull().sum()* and more as shown in figure 5 and figure 6. The treated output is shown in figure 7 using *df* as a display command.

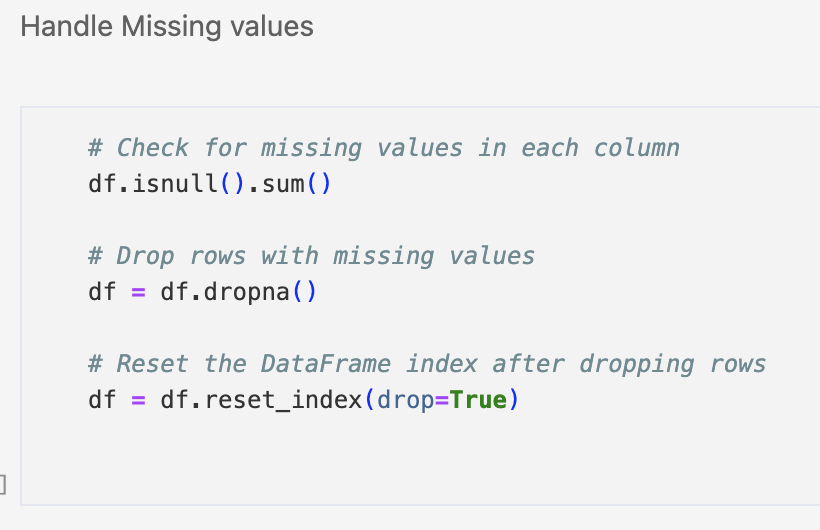


Figure 5: Handling missing values

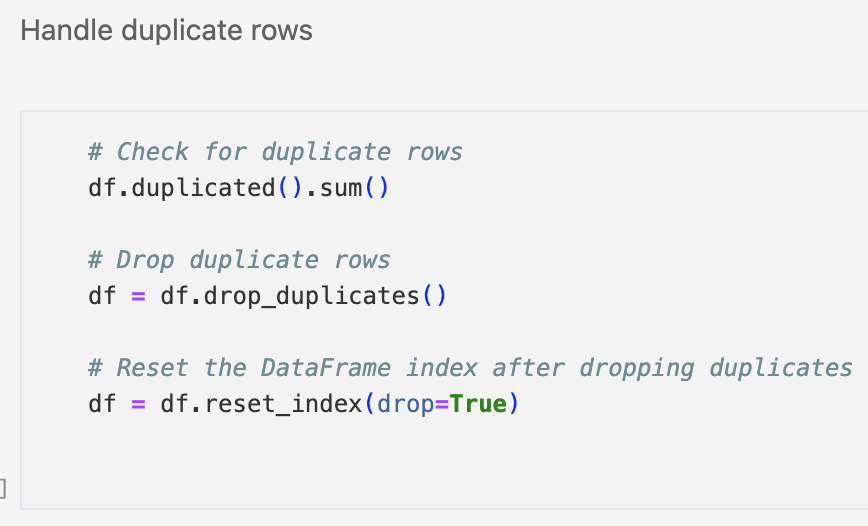


Figure 6: Handling duplicated rows

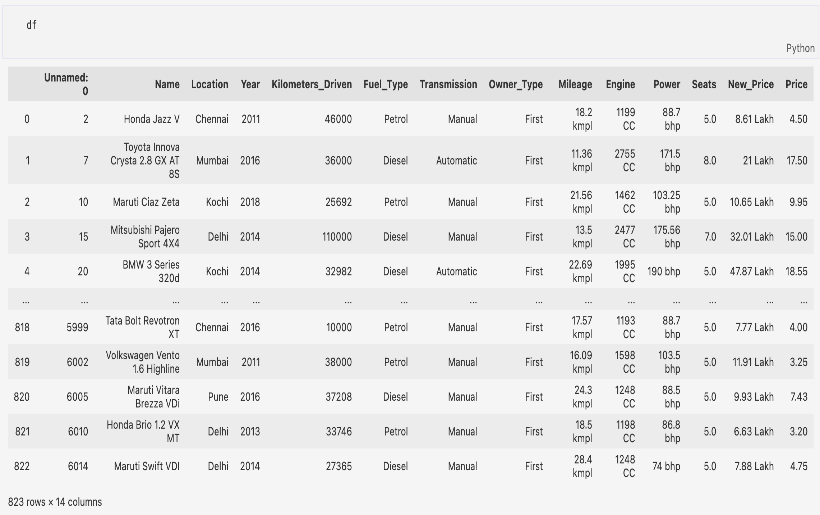


Figure 7: Output of the data after handling duplicates

5) On Figure 8 we show how the data type for ‘Mileage’, ‘Engine’ and ‘Power’ columns is fixed. They should be float instead of object. This could be achieved by removing the units (kmpl, cc, bhp).

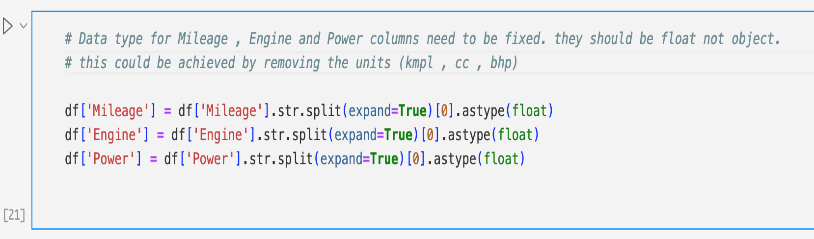


Figure 8: Fixing data type for three columns.

The output is shown in Figure 9.

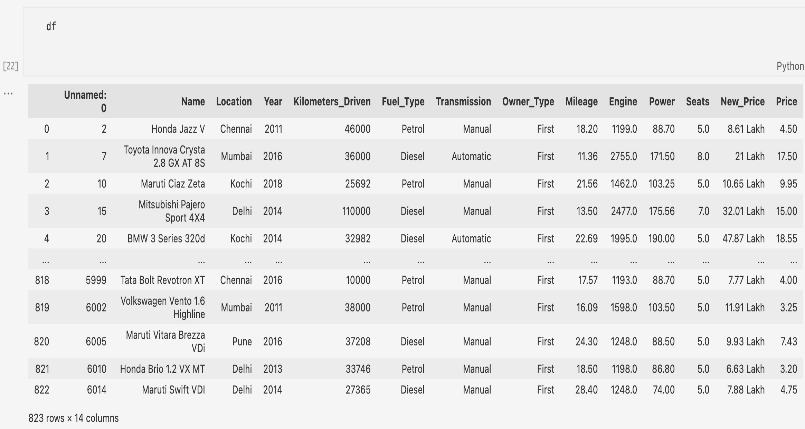


Figure 9: Output of the data after fixing the data types

6) Two columns were dropped as they are not needed for analysis (Figure 10). The output is displayed in Figure 11.

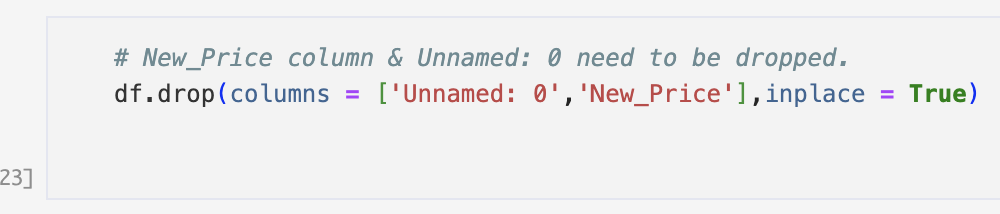


Figure 10: Dropping unneeded columns

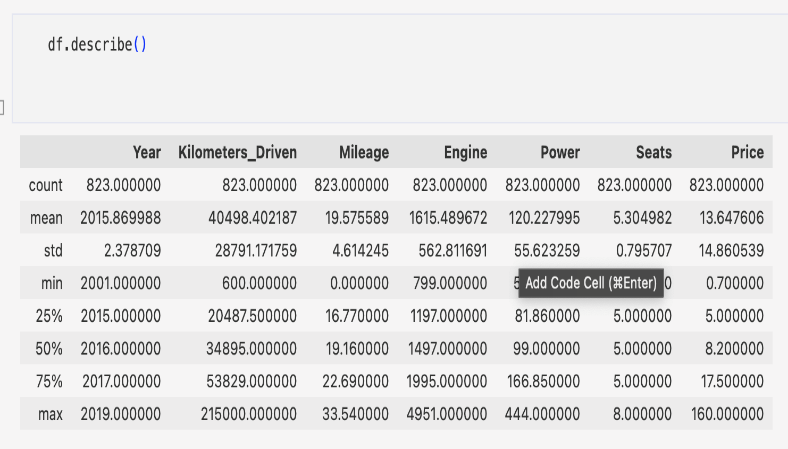


Figure 11: Output after dropping unnecessary columns

7) The final outcome of data preparation step is shown in Figure 12.



Figure 12: Final view of the data

8) Last step is to write the object (data frame) to a comma-separated values (csv) file. This file will be further used in Power BI tool for visualization.



Figure 13: Writing the data frame to a csv file

Both the datasets are treated to the same process of data preprocessing. For the sake of length of the report both .py files are shared along with the project and one is briefly explained in the scope of this report.

1. **DATA VISUALIZAION TECHNIQUES**

As mentioned earlier in the report, Microsoft Power BI is used as a visualizing and data analysis tool. Further details will be explained in the following submission.

As mentioned earlier in the report, Microsoft Power BI is used as a visualizing and data analysis tool. The data prepared initially in the preprocessing step and how it appears in Power BI is shown in figure 14. This is the first step in which the data processed data is loaded in Power BI.

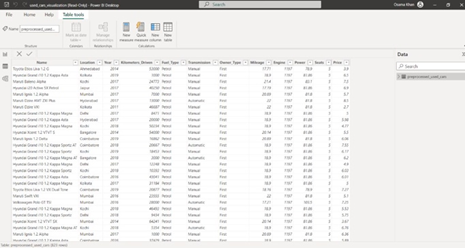


Figure 14:

Rather than going through the list of used cars one by one to find trends in the data, we produced dependencies. These include the number of used cars (count) available in a Year, on one Location, what type of transmission they have, and which fuel these cars run on. Other than this the crucial factors that might affect the price of these car are the cars themselves (name of cars), number of kilometers the car is driven so far and the number of seats they have. To summarize we have tried to find trends for the availability of used cars and their prices based on the data available.

The following figures summarize general trends that are consistent in the dataset.

**Count by Year**

The graph in figure 15 shows an increase in the number of used cars available as the years pass and shows a decrease.

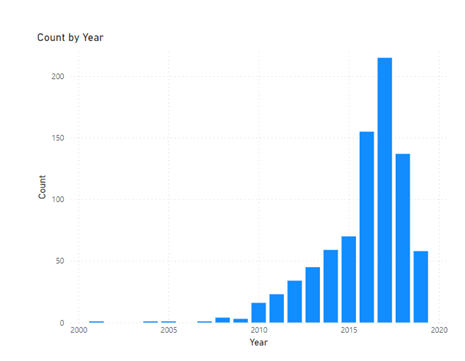


Figure 15:

**Count by Location**

How the number of used cars available varies by location is shown in fig-16.

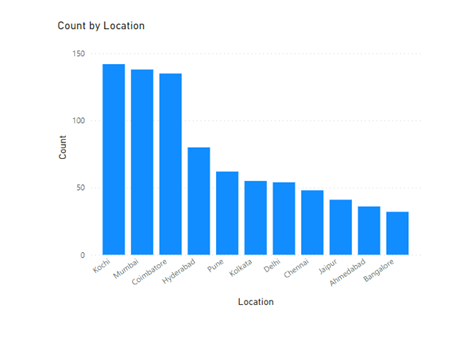


Figure 16:

**Count by Transmission**

The following figure - 17 is a pie chart which distinguishes the number of cars available depending on the type of transmission they have.

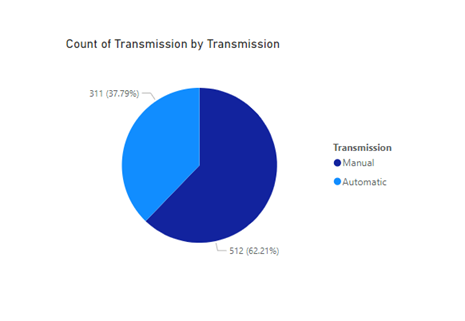


Figure 17:

**Count by Fuel Type**

Again, we used a pie chart to separate out cars based on the type of fuel they run on. Fig-18.

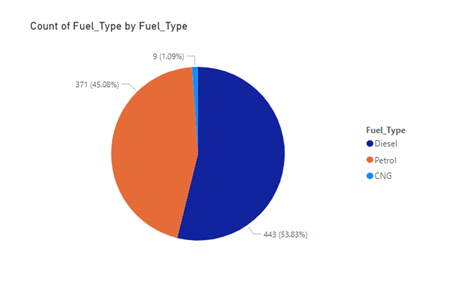


Figure 18:

The purpose of using pie chart is that there are categories that the distinction is being based upon. So, a pie chart gives a better representation of the data.

**Pricing**

An important is pricing of cars and all the effort revolves around providing the best rate to customers and readily available prices and that best matches customer preferences. To address this, the team produced graphical representations of the sum of prices depending on the factors available in the dataset. Name, Kilometers driven and, the number of seating capacity the car has.

**Pricing by number of seats**

The information in the following figures Fig-19, Fig-20, Fig-21 is straight forward and understandable.

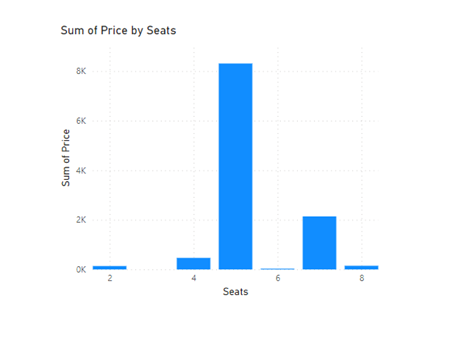


Figure 19:

**Pricing by Name**

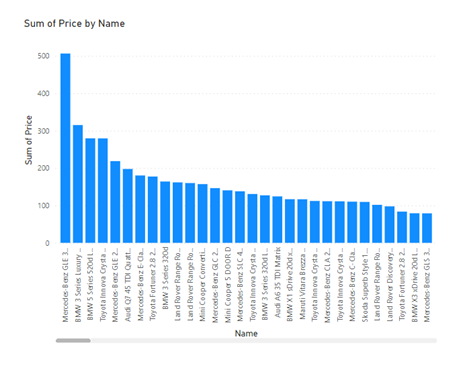


Figure 20:

**Pricing by Kilometers Driven**

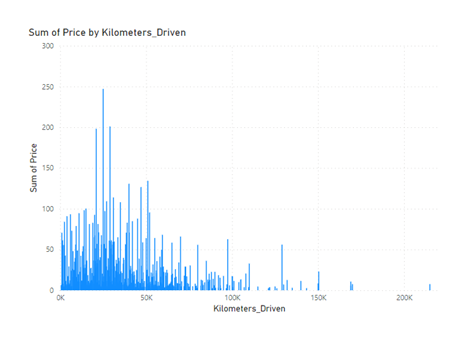


Figure 21:

The interesting part about using Power BI that is not clear in the figures is that it can adjust trends and information on graph. For example, a customer comes in and asks about the availability of cars in Delhi. Power BI enables us with the capability to do that.

1. **CONCLUSION**
2. **REFERENCES**

*Used Car Price Prediction*. (n.d.). Used Car Price Prediction | Kaggle. <https://www.kaggle.com/code/iabhishekmaurya/used-car-price-prediction>

*Cornell Car Rental Dataset*. (n.d.). Cornell Car Rental Dataset | Kaggle. https:///datasets/kushleshkumar/cornell-car-rental-dataset

*Build rich reports and share Insights that drive results | Power BI*. (n.d.). Build Rich Reports and Share Insights That Drive Results | Power BI. <https://powerbi.microsoft.com/en-us/landing/free-account/?ef_id=_k_EAIaIQobChMI6cC2uNyD_wIV7QKtBh398QOkEAAYASAAEgI2oPD_BwE_k_&OCID=AIDcmm80atqgos_SEM__k_EAIaIQobChMI6cC2uNyD_wIV7QKtBh398QOkEAAYASAAEgI2oPD_BwE_k_&gclid=EAIaIQobChMI6cC2uNyD_wIV7QKtBh398QOkEAAYASAAEgI2oPD_BwE>

*What Is Data Preprocessing & What Are The Steps Involved?* (2021, May 24). MonkeyLearn Blog. <https://monkeylearn.com/blog/data-preprocessing/>

1. **WORKLOAD ASSIGNMENT**

A screenshot of a calendar

Description automatically generated with medium confidence

1. **GITHUB ADDRESS**

GitHub link with Python project and Power BI: <https://github.com/Umutbek/ds522_group_project>