**Canva Link: https://www.canva.com/design/DAGMYwMt12c/VnbCP84QLQjpbofxECYe4g/edit?utm\_content=DAGMYwMt12c&utm\_campaign=designshare&utm\_medium=link2&utm\_source=sharebutton**

**READMEFIRST:** You don’t need to add any info here anymore. The documentation that we will submit is a powerpoint presentation (see the link above). Kindly proceed right away on adding info and designing in the **Canva ppt**. Look for your names beside each section, you are responsible for whatever’s assigned to you there. If you wanna switch, if the distribution seems unfair, or need to ask help, just drop a message in our gc.

**aug 6 deadline**

**DATAPRE - Final Project**

Group 2 Members:

* Abiño, Renz Wendell
* Abot, Nikolas Benjamin
* Co, Matthew Benedict
* Detablan, John Daniel

**Predicting Car Resale Prices**

1. **Topic description [NIKI]**

Brand new cars are expensive that most people cannot afford and that is why there are buy and sell businesses that provide an alternative for people to own a car at a cheaper price. However, business owners can sometimes make mistakes in buying cars that are pricier than what it is really worth. That is why we decided that the data set we web scraped could be used as an analysis in predicting car resale prices using its components. Car resale price prediction basically refers to the estimation of the prices of cars that have already been used. While the global automotive industry is growing, predicting car prices is very important because more stakeholders are becoming interested in buying used cars, especially individual buyers and sellers, dealerships, as well as insurance companies. Accuracy is very crucial in predicting car prices since it indicates how reliable the prediction is, and how well the model can classify or predict the results.

1. **Business Application [NIKI]**

From studies, \_\_\_\_\_\_, it was mentioned that these are the features of a car used that provided promising results that measures their model’s goodness of fit. Unfortunately, we were only able to get 8/13 of the said car features they’ve used in predicting car resale prices for our **primary dataset**. Nonetheless, the goal of our dataset is to be used in model prediction where it would determine the ranking of the car components that contribute significantly in estimating its price. That way, business owners or even individuals can prioritize looking for the car parts when they buy or sell. The potential usage of our dataset could be translated to information to safeguard themselves from scams or get the most profit from the information.

As for our first external dataset (brand new prices of cars), we decided that the price of brand new cars (SRP) as a variable can be a valuable addition to the second-hand cars where SRP can be used to compare resale price (car depreciation).   
  
For our second external dataset, the number of car service centers in Metro Manila per car make/brand could be a deciding factor because it is important to consider the convenience of car parts’ availability for maintenance. It is known that every vehicle requires maintenance from time-to-time that includes replacement parts. So, the number of car service centers variable can be used to assess the car parts’ availability.

1. **Data Preparation Process Documentation**
2. **Data Collection [NIKI]**

We decided to gather data using web scraping from two car buy and sell websites for our primary datasets which are AutoDeal and Zigwheels. Note that the data for the primary dataset consists of only used cars. Initially, we only web scraped from AutoDeal because most of the websites we have tried required doing advanced techniques in web scraping, but we’ve only achieved under 1000 rows of data. So, we decided to do another data collection from Zigwheels using Selenium in order to achieve the minimum rows of data for this project. To give an overview of the columns that we collected from both websites, see the table below.

| Variables Reference | AutoDeal | Zigwheels |
| --- | --- | --- |
| brand | ✓ | ✓ |
| model | ✓ | ✓ |
| price | ✓ | ✓ |
| vehicle age | ✓ | ✓ |
| Kilometers driven | ✓ | ✓ |
| seller type |  |  |
| fuel type | ✓ | ✓ |
| transmission type | ✓ | ✓ |
| mileage |  |  |
| engine |  |  |
| engine size | ✓ | ✓ |
| max power |  |  |
| seats |  |  |
| Total | 8/13 | 8/13 |

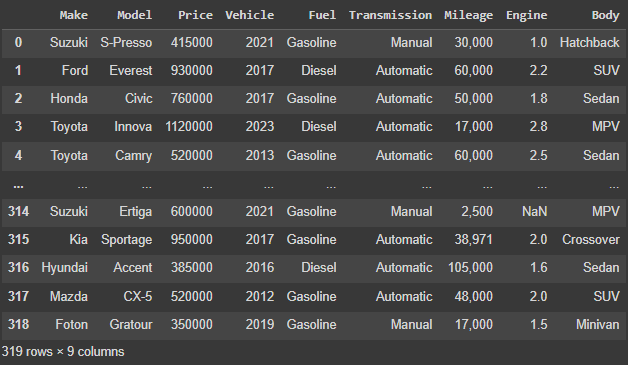
**Table 1. Summary of Columns Collected**

The Variables Reference consists of the columns from \_\_\_\_ that provided significant results in terms of model fitness

The detailed explanation in the procedure of our data collection from each website will be explained in the following subsections below.

1. **Primary Dataset Web Scraping**
   * 1. **AutoDeal Used Cars [WENDELL]**

Upon observing and seeing patterns from AutoDeal, we noticed that after clicking the filter “Used Cars”, the website has a page number in the url that iterates upon clicking onto the succeeding pages. With that, the for loop was utilized to ensure that after collecting all the cars and its data from each page, it moves on to the next. After getting all the *cards* from the cars in each page, we took the necessary car features and proceeded to contain them in pandas dataframe, and saved as csv (See Image 1. AutoDeal Dataset).

**Image 1. AutoDeal Dataset**

* + 1. **Zigwheels Used Cars** [Daniel]
  1. **External Dataset**
     1. **AutoDeal Brand New Cars Web Scraping [WENDELL]**

We decided to make use of the price of brand new cars from AutoDeal to merge it with the primary dataset so that we have an additional column called SRP (price of a brand new car) that could provide a valuable use in comparison to the resale price from the used cars dataset. To make the data collection faster, we looked for the unique *Makes* and stored it in a list, then iterated through it to only web scrape the *Makes* that are in the primary dataset.

uniqueMakes = primary['Make'].unique()

print(uniqueMakes)

i = 1

for make in uniqueMakes:

print(f"{i}. {make}")

i += 1

**Code 1. uniqueMakes**

**card\_listings = []**

**for i in range(11, len(uniqueMakes)):**

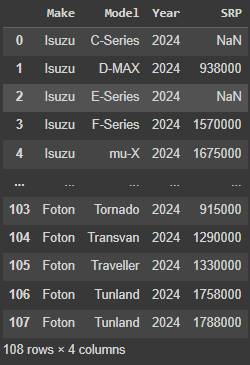
**print(uniqueMakes[i])**

**parent\_url = f"https://www.autodeal.com.ph/cars/{uniqueMakes[i]}"**

**print(parent\_url)**

**Code 2. uniqueMakes Filter**

There was a problem during web scraping that stopped the code because one car had no listed price that caused an error. A simple solution we made is that we web scraped first from Suzuki to Kia, then Isuzu to Foton to avoid the error.



**Image 2. Split Web Scraping**

Then, we concatenated the two to finally get the first external dataset and saved as a csv file.



**Image 3. First External Dataset**

* + 1. **Zigwheels Car Service Centers [WENDELL]**

Another possible factor for car buyers is the availability of the parts for their car to maintain. Cars need replacement parts as they are used, and not all car *Makes* have the same quantity of available parts. It could be a factor in deciding which car *Make* to choose based on the number of service centers available to buy car parts. So, we used Zigwheels Car Service Centers Search Engine to find the number of Service Centers in Metro Manila and listed them manually in an excel sheet named make\_mshops.

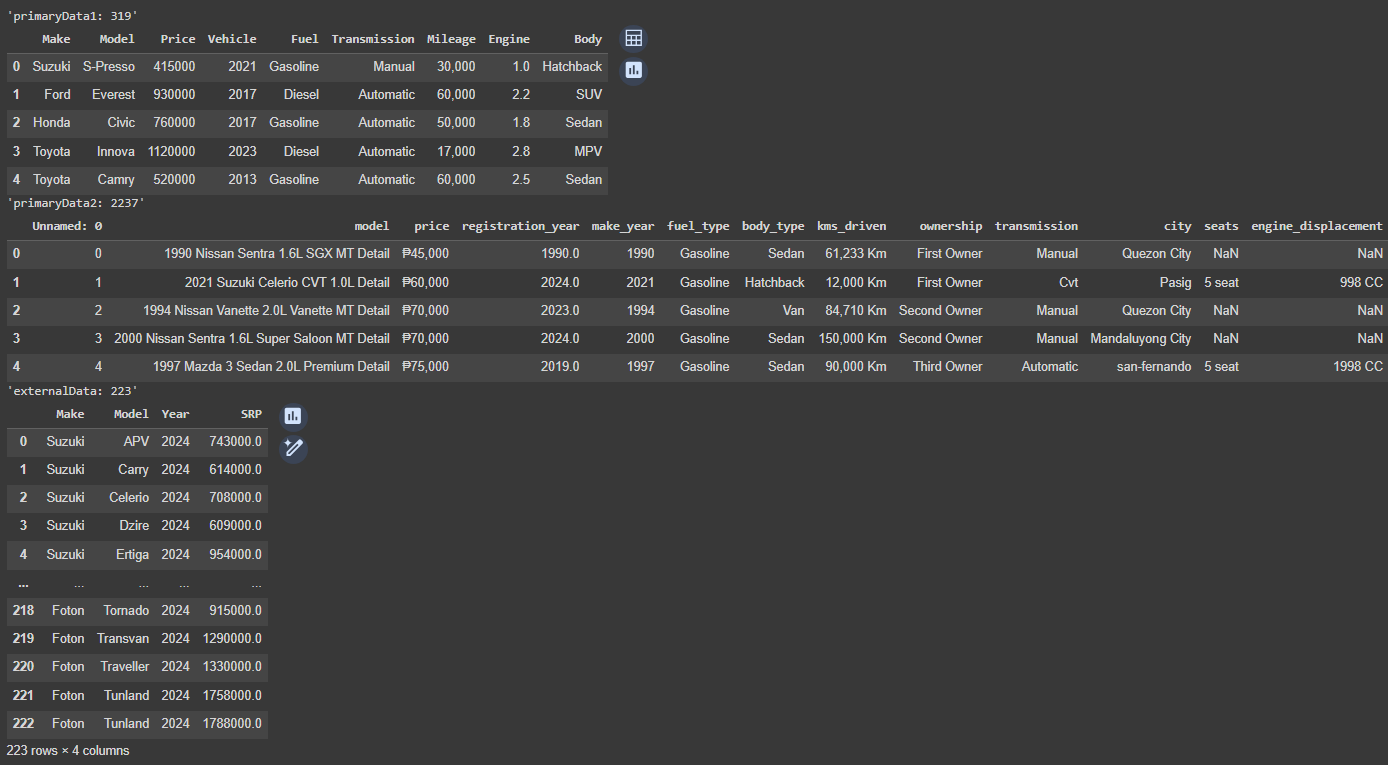
| make | no\_sc\_mmnl |
| --- | --- |
| Suzuki | 9 |
| Ford | 2 |
| Honda | 4 |
| Toyota | 9 |
| Hyundai | 6 |
| Mitsubishi | 4 |
| Mazda | 1 |
| Geely | 4 |
| Nissan | 2 |
| Chevrolet | 4 |
| Isuzu | 6 |
| Subaru | 6 |
| BMW | 4 |
| Mercedes-Benz | 1 |
| Kia | 6 |
| MG | 4 |
| MINI | 1 |
| Chery | 2 |
| Lexus | 1 |
| Foton | 0 |
| Volkswagen | 0 |
| Audi | 1 |
| Jeep | 0 |

**Table 2. make\_mshops**

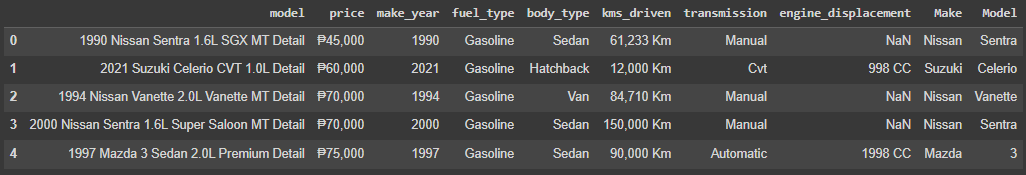
1. **Data Processing**

**Part 1. Concatenation and Rearrangement [WENDELL]**

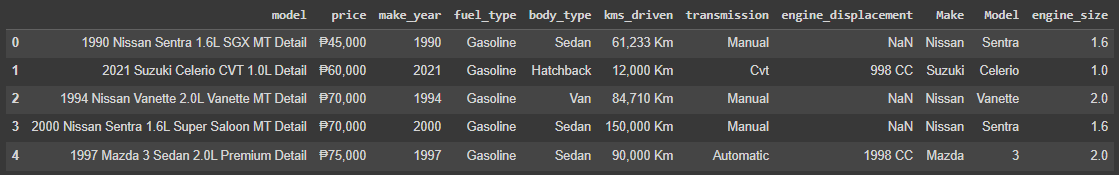
**.** We imported primaryData1 (autodeal.csv), primaryData2 (zigwheels.csv), and externalData (autodealExternal.csv or brand new cars dataset).

**Image 4. primaryData1, primaryData2, and externalData**

Then started with primaryData2 cleaning and dropped the unnecessary columns of primaryData2 to concatenate it with primaryData1. We then proceeded to split a column to make, and model.

**Image 5. Splitting model into Make and Model**

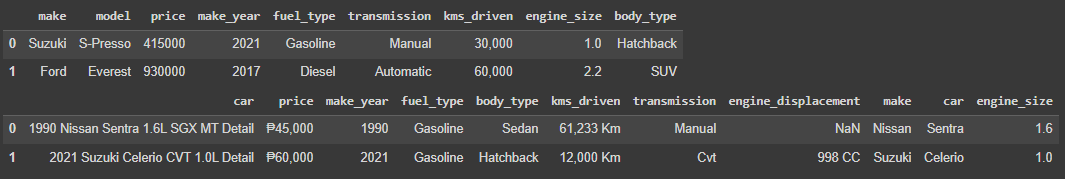
Then, we added a new column called *engine* in primaryDataset 2 by either extracting the engine from the column *model* or from *engine\_displacement.*

**Image 6. engine\_size column addition**

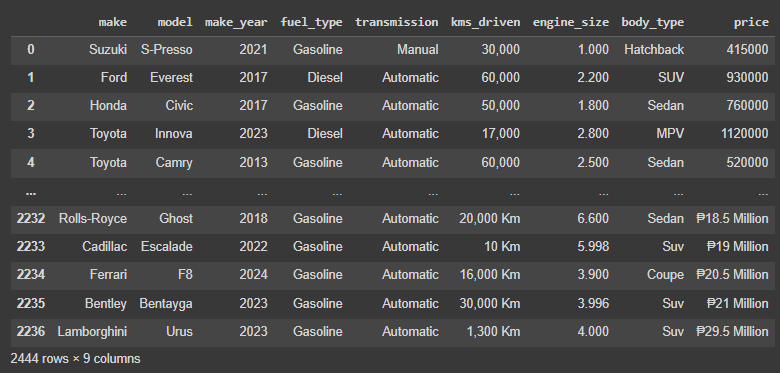
Then, we dropped the columns that are null in the column *engine\_size* in primaryDataset2 since there are only 112 rows, and filling techniques cannot be applied since the dataset we have is not based on time.

**Image 7. engine\_size row dropping**

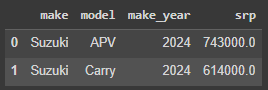
Then, renaming of columns was done for both primary datasets in preparation for the concatenation of the two later.

**Image 8. Columns renaming**

After, we dropped the unnecessary columns in primaryDataset2 to match the columns of primaryDataset1, rearranged the columns for both primary datasets, and then concatenated them.

**Image 9. Primary datasets concatenation**

For the external dataset, we renamed and rearranged the dataset for uniformity.

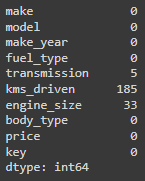


**Image 10. External dataset rearranging and renaming**

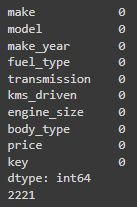
Afterwards, a new column called key for the concatenated dataset and external dataset was created by combining the columns *make* and *model* for merging in the next part of the data processing. Then, we removed duplicates in the externalDataset and kept the first copy because that could be a problem in the merging procedure. To explain why duplicates were not removed in the primary dataset, it is because there could be multiple car listings of the same car but differ in other car features e.g., kms\_driven, price, etc. Then finally exported two datasets: primaryData.csv, and externalData.csv.

**Part 2. Data Cleaning [WENDELL]**

**.** We imported primaryData.csv, and externalData.csv, and proceeded to check the number of nulls per column in the primary dataset first.

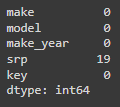
**Image 11. Null count per column in the primaryDataset**

We then proceed to drop nulls in the columns *kms\_driven*, *transmission*, and *engine\_size* because no other data cleaning techniques taught in class besides dropping rows are reasonable to do. Now, that leaves us having zero nulls for the primary dataset.



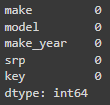
**Image 12. Summary of the number of nulls in the primaryDataset**

Next is the procedure of data cleaning of externalData. Below is the summary of the number of nulls in the external dataset.



**Image 12. Null count per column in the externalDataset**

Dropping all the rows of *srp* was necessary since those nulls are the cars in AutoDeal that do not have a current price listing from the website. Now, we have no nulls anymore for the externalDataset.



**Image 13. Summary of the number of nulls in the externalDataset**

Finally, saving these two cleaned datasets and exporting them as: cleanedPrimaryData.csv, and cleanedExternalData.csv.

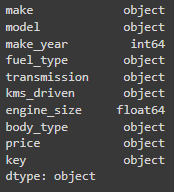
**Part 3. Data Transformation [WENDELL]**

**.** We used cleanedPrimaryData.csv, and cleanedExternalData.csv from part 2.

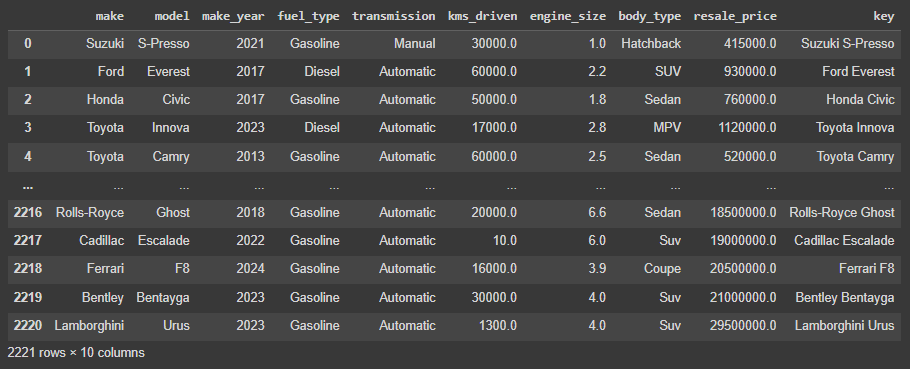


**Image 14. Cleaned primary and external dataset**

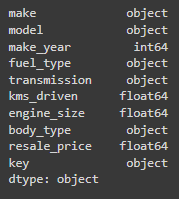
We proceeded with the data transformation of the primary dataset first. Below is the summary of the data types of primaryData.

**Image 15. Data types of primaryData**

We then converted the column name *price* into *resale\_price* then transformed it from an object to float, then kms\_driven (truncated “km”), and engine\_size (rounded it by two decimals).



**Image 16. Data types transformation**



**Image 17. Transformed Data types of primaryData**

As for the external dataset, no transformation was needed and so we then proceeded with exporting the two datasets: transformedCleanedPrimaryData.csv, and transformedCleanedExternalData.csv.

**Part 4. Merge [WENDELL]**

**.** In this portion of the data processing, we merged transformedCleanedPrimaryData.csv, and transformedCleanedExternalData.csv by inner join, retaining all the columns from the primary dataset, and only getting the columns *srp,* and of course the column *key.*

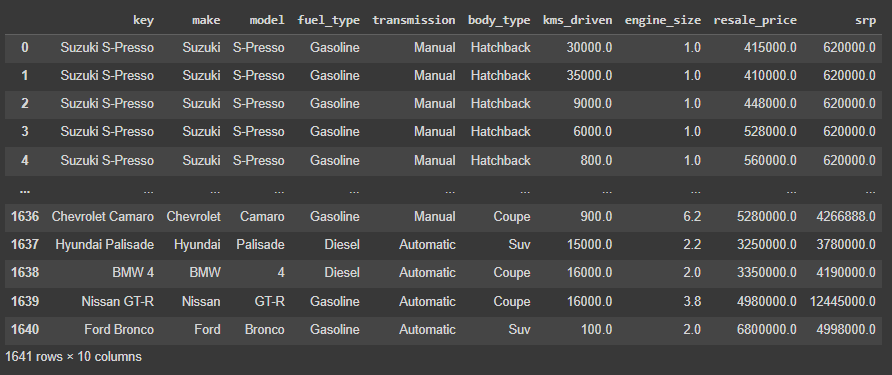
*mergedDataset = pd.merge(primaryData, externalData[["srp", "key"]], on="key")*

*display(len(primaryData))*

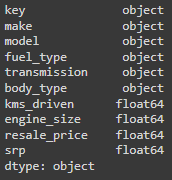
*display(mergedDataset)*

*display(len(mergedDataset))*

**Code 3. Inner join of primary and external dataset**

**Image 18. Merged dataset**

Afterwards, rearranging was done for the newly merged dataset to separate the data categorical variables, and numerical variables.

****

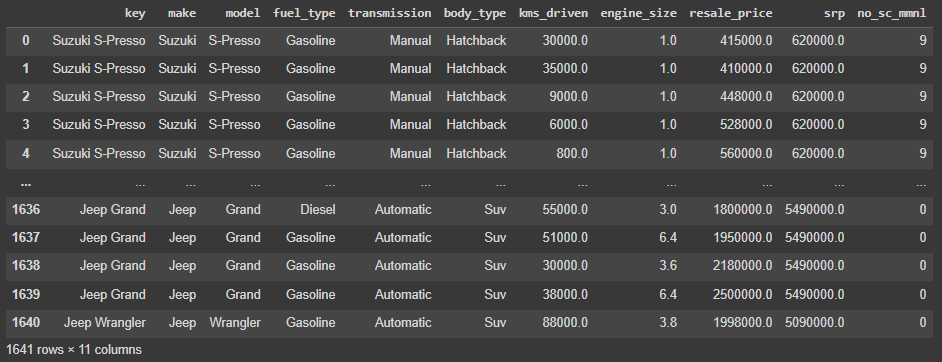
**Image 19. Merged dataset data types**

Now, with the second external dataset to merge it with, make\_mshops.csv was imported into this part’s notebook and used inner join for two datasets: mergedDataset (the primary dataset, and the first external dataset), and make\_mshops. We kept all the columns for the first merged dataset, and took the columns *no\_sc\_mmnl* (number of car service centers in Metro Manila), and *make* as the key for merging.

mergedDataset2 = pd.merge(mergedDataset, externalDataset2[["make", "no\_sc\_mmnl"]], on="make")

display(mergedDataset2)

display(len(mergedDataset))

**Code 4. Inner join of the first merged and the second external dataset**

**Image 20. Final merged dataset**

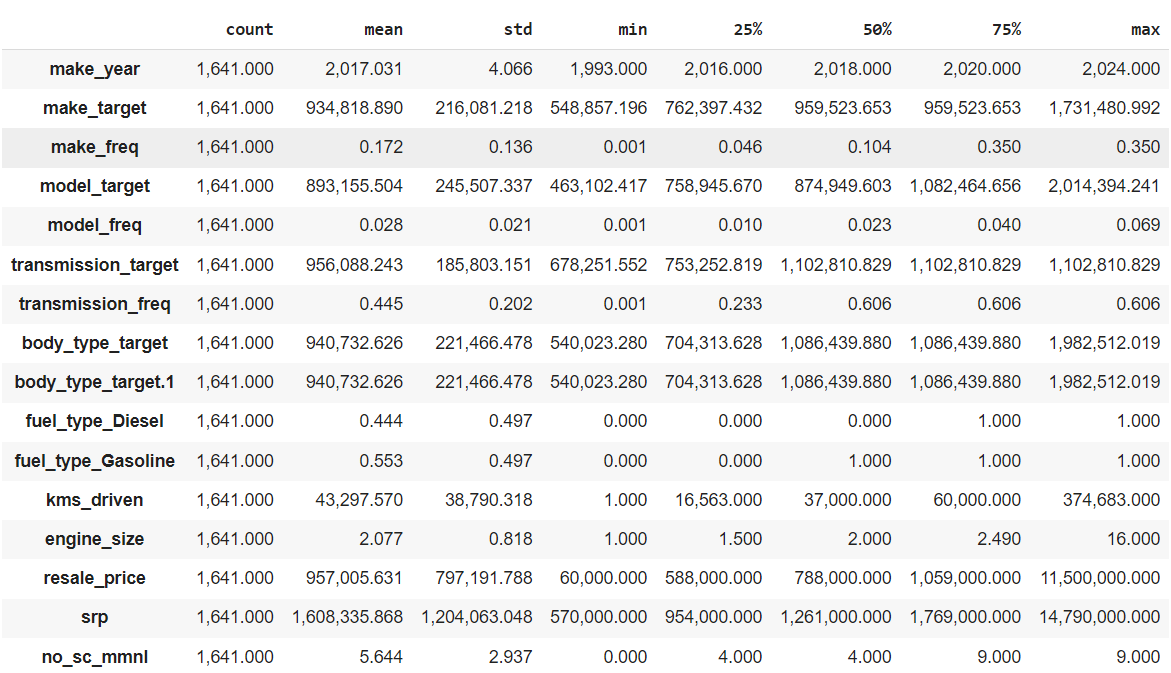
Finally exporting the final merged dataset in the part as mergedDataset.csv.

**Part 5. Categorical Profiling and Encoding Categorical Variables [DANIEL]**

1. **Exploratory Data Analysis [MATTHEW]**

* **Statistics Summary**

There are 1641 cars. Based on the statistics summary, years range from 1993-2024 for both used and brand new cars. 25% of the cars were created from 2020 onwards. 55.3% of cars are running on gasoline while the rest are running on diesel. Cars can be driven on an average of around 43000 km. Car prices range from 60K to 11.5 M. 25% of the cars can be brought to 9 service centers in Metro Manila, which is the maximum value of the said variable.

****

* **Univariate Analysis**

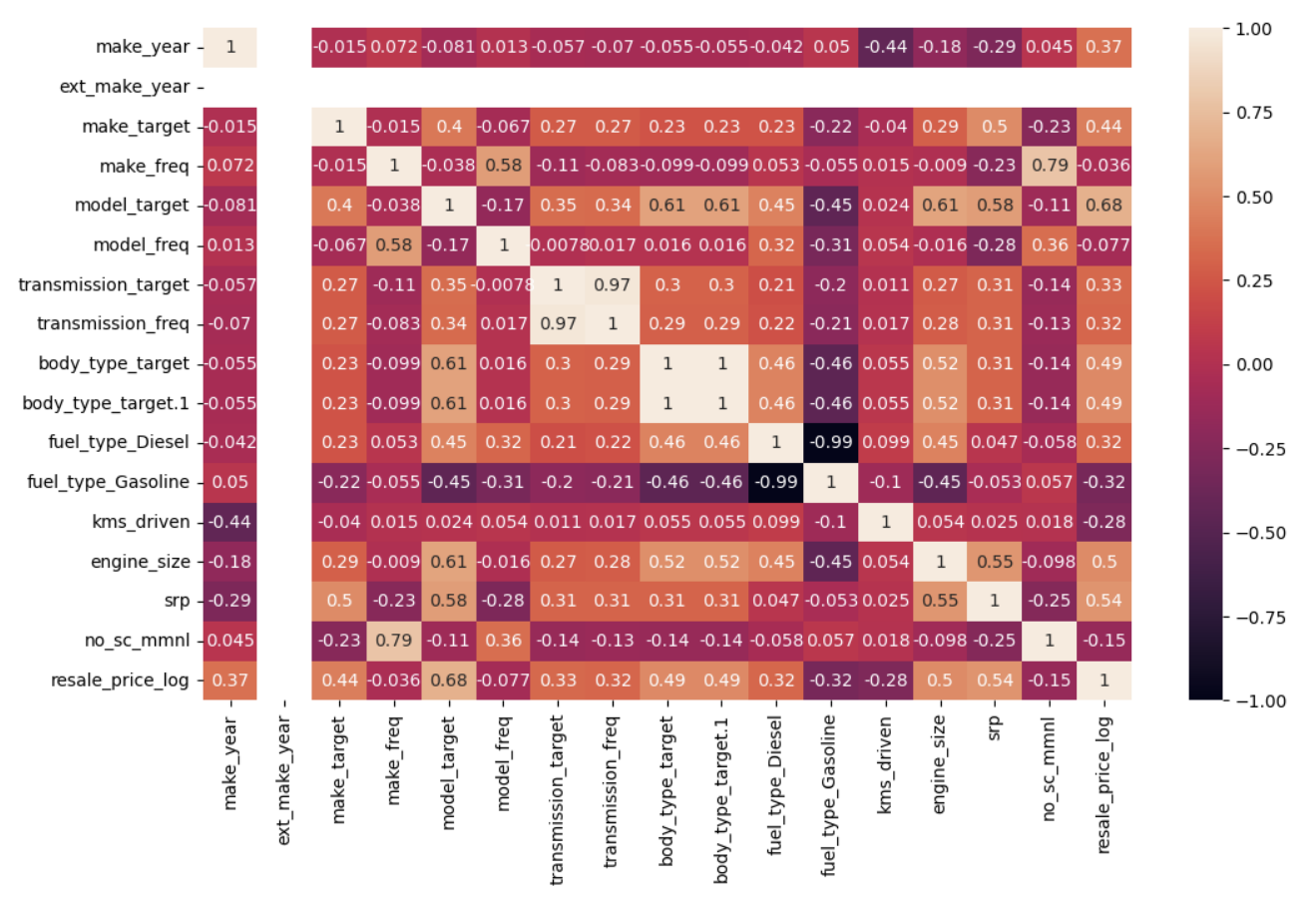
Numerical variables are visualized into histograms and boxplots, and some of the variables are shown to have high skewness. There is no further visualization since there are no categorical variables in the dataset. Since the resale price is highly skewed, log transformation was used in normalizing the variables.

* **Bivariate Analysis**

A pair plot was used to visualize the relationship between two variables from the dataset.

* **Multivariate Analysis**

A heat map was used to show the correlation between two variables from the dataset.

****

Make year and kilometers-Driven are very weakly correlated to most of the variables from the dataset. Resale price has a positive correlation to diesel cars and a negative correlation to gasoline cars. Model frequency does not create much impact on body type, engine size, and transmission (both target and frequency encoding). However, model target has a fairly strong positive correlation to body type, engine size, and the price. One fuel type has a very strong negative correlation to one other fuel type (-0.99). Make frequency has a strong positive correlation to the number of service centers in Metro Manila. SRP are fairly strongly positive in correlation to make target, model target, engine size, and price.

1. **References**

**Put references here**