

**TASK**

**Exploratory Data Analysis on the Automobile Data Set**

[](http://www.hyperiondev.com/portal/)

**Introduction**

Summary of the data set

The data that’s contained within the automobile.csv are details of each make in regards to performance, attributes and the pricing of each car.

**DATA CLEANING**

# SUMMARY OF THE METHODS AND VISUALIZATIONS DONE DURING DATA CLEANING

The following methods were done:

* Creation of the dataframe of the automobile.txt file.
* Specifying the columns that exist.
* Removal of the normalized-losses column as it contained redundant data.
* Converted the datatypes of five 5 columns to int
* Exploring the dataset asking the following questions
  + The car makes that currently exist within the dataset
  + The type of aspiration the dataset currently contains
  + the fuel type for each make
  + The price of each make
  + 5 Most expensive cars
  + 5 cheapest cars
  + Comparison of Each makes MPG
  + the number of body-styles within the dataset

The visualisations that were created using both seaborn and plot or matplotlib:

Bar graphs:

* The amount of makes within the dataset
* Aspiration type
* 7 Most expensive Cars
* City vs Highway MPG for each make

Pie Charts:

* Percentage of body-types within the dataset
* Percentage of Aspirations within the dataset

MISSING DATA

# ANY MISSING DATA? HOW DID YOU HANDLE IT

Yes. There weren’t any Nan values, however, from the normalized-losses column there were question marks which didn’t say anything about the series or the relationship within the rows. So I made the decision to drop the column since a lot of the data was unknown data.

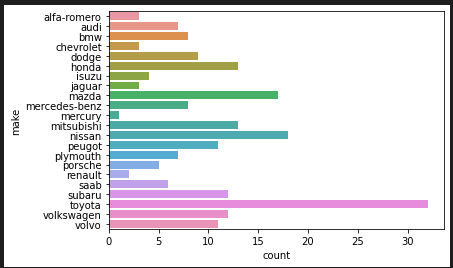
DATA STORIES AND VISUALIZATIONS

# THIS IS THE BULK OF THIS PROJECT. EXTRACT STORIES AND ASSUMPTIONS BASED ON VISUALIZATIONS OF THE DATA  
  
**Exploration of the dataset**

The following below are the investigations I have done within this dataset. Each screenshot displayed will answer a question that I asked in order to understand the type of data that exists within the automobile dataset.

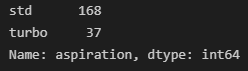
**What is the total number of each car make within the dataset?**

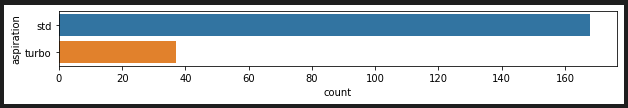


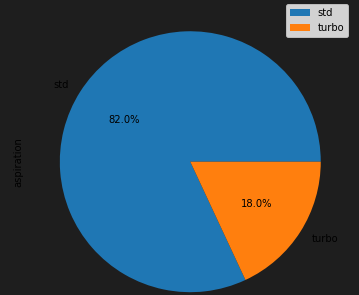


From both the value count and the visualization above, we can say that the Toyota seems to be the most common car within the dataset and the Mercury the least common car within the dataset.

**How many cars are either std or turb aspirations?**

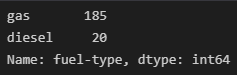


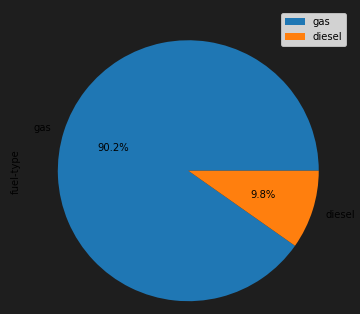




From the visualizations above, STD seems to be the most used aspiration within the dataset.

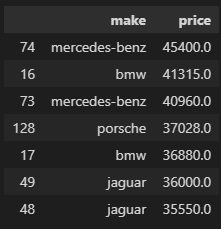
**The fuel type of each car?**

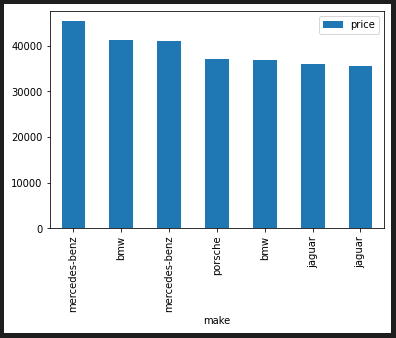




From this, I can confirm that gas is the most used most fuel type within all the makes in the dataset.

**What are the 7 most expensive cars?**

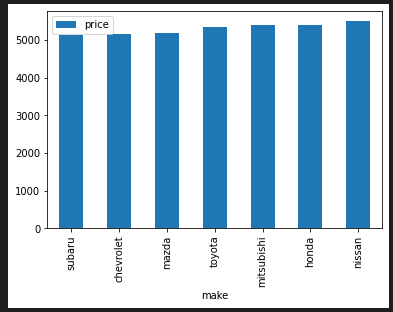




I’ll simply point out that Mercedes is the most expensive car and also the top 7 most expensive cars come at a cost above 35550.00. Not sure which currency is being used.

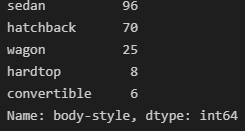
**What are the 7 cheapest cars?**

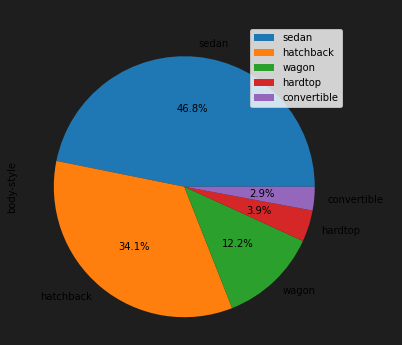




Subaru is the cheapest car, with the 7 cheapest cars being under 5500.00.

**The number of body types within the dataset?**

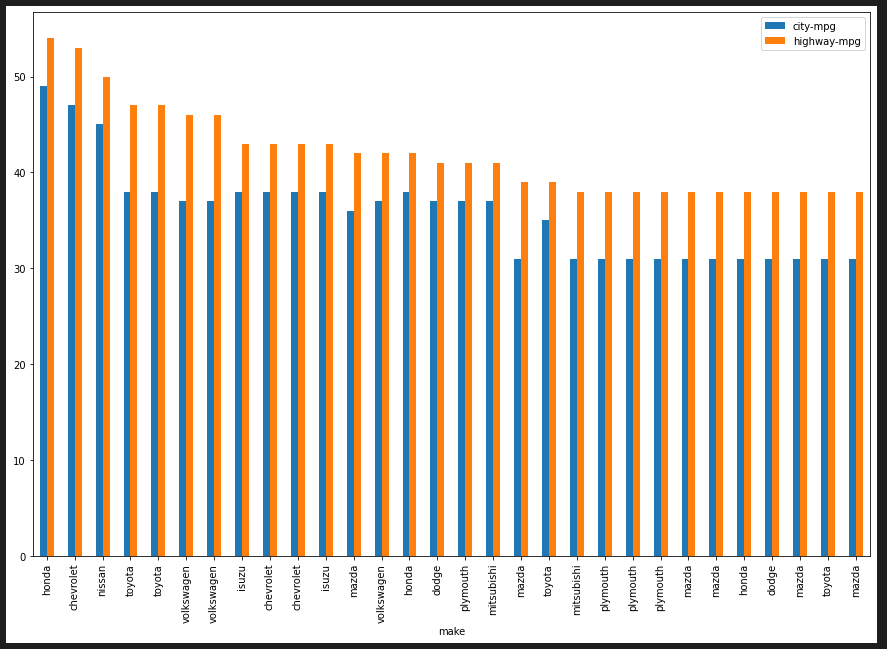




There are five body types within the dataset, making sedan the most common body type within the dataset.

**What is the comparison between city-mpg and highway-mpg of 30 cars with the highest to lowest highway-mpg?**





Within the 30 cars that have been displayed, Honda seems to have the highest mpg for both city and highway. However, not all cars would have with a high highway-mpg would have higher city-mpg in comparison with all the other makes.

# ENSURE THIS DOCUMENT IS NEAT AND CAN BE ADDED IN YOUR PORTFOLIO

**THIS REPORT WAS WRITTEN BY : Naledi Motau**

