TASK

Exploratory Data Analysis on the Automobile Data Set

Visit our website

Introduction

Summary of the data set

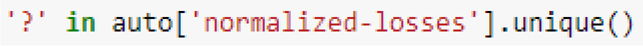
DATA CLEANING

So, obviously before any cleaning can be done, we must load the data into a data frame so we can start manipulating it. Then we must understand the data.

So, to achieve the above I did a couple of things. First, before anything I set the display column to max so, when I printed the df I can see all the columns and not just a few columns. Next, I checked the shape of the data to see how many rows and columns we are dealing with. The dataset has 205 rows and 26 columns. Next, I made use of the describe function to get an understand of the dataset over all and how the variables relate to each other. Next, we check for the dreaded null values. With the 'isnull' and 'sum' functions to sum up all the null values per column. This returned no null values. Next, I called the info function on the dataset to see what kinds of data types each column is.
Finally, I used the head function to print the first five lines of the data. This leads us onto the next heading.

MISSING DATA

The head function shows us that there are '?'s in the data. This gives us insight into why seemingly numerical columns where of dtype object. This means that there was a combination of string and integer values in the column. This limits our functionality when working with the data. So, I wrote a little script to check for '?'s in a column of the data set.

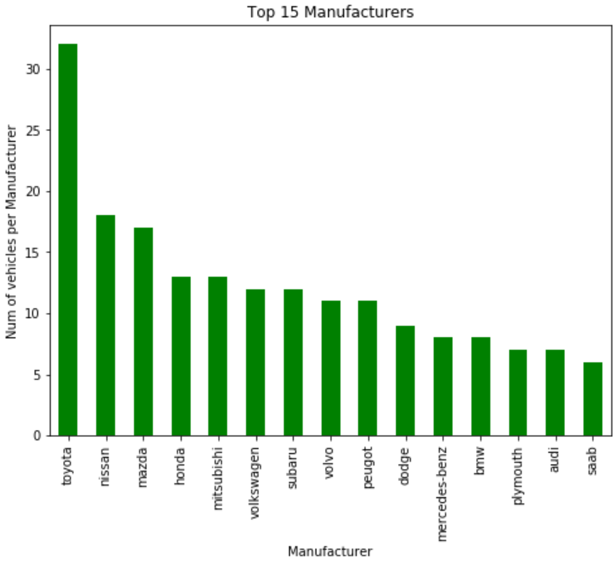


```
'?' in auto['normalized-losses'].unique()
```

The script can be seen above. This allows me to preform a quick check on a column to see if there are '?'s in that specific column. If 'True' is returned the column in question contains '?'s.
So, with that in mind I ran the script on the normalized-losses column and true was returned. For this column it made sense to replace the '?'s with the average. So, I isolated all the columns that did not contain a '?' and calculated the mean. Then I replaced all occurrences of a '?' with the mean. And converted the data type of the column to int64.
The same steps where taken for the horsepower column and the price columns. Checking for questions marks, calculating the mean, then replacing all occurrences of a '?' with the mean and changing the data type to int64.
For bore, stroke and peak rpm anther route was taken. I casted the data types for these columns to int64 and replaced an occurrence of a '?' value with a null value. I took this step because inputting the mean for these columns does not make sense. For example, peak-rpm varies greatly between cars as a multitude of factors have an influence on this variable, so I preferred to leave these columns with null values.
Finally, for this section I ran a data type check to see if all the columns were in the desired format.

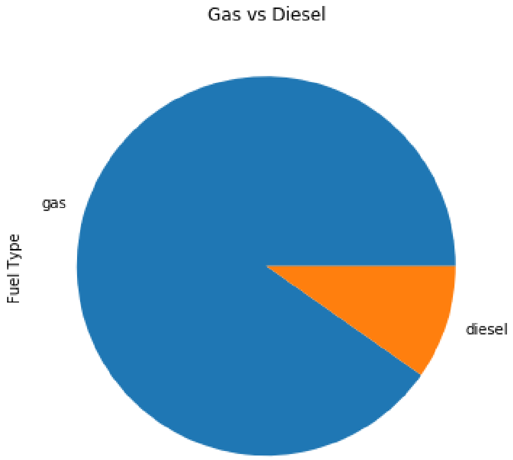
DATA STORIES AND VISUALIZATIONS

For this section I created a few visualisations with the intention of answering a question. The first question I had was what the most common manufacturer in the dataset was? This can be answered with a simple bar graph.



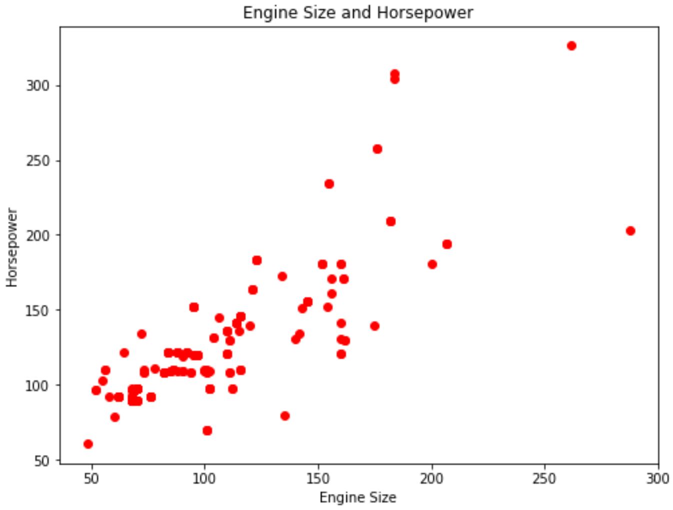
From the visualisation above we can see that Toyota is the by far the most common manufacturer in the dataset followed by Nissan and then Mazda.

Next, what is the most common petrol type in the dataset. Gas or diesel? For this visualisation I chose a pie chart is it best illustrates the share of the two petrol types in the data.



From the pie chart above we can see that standard gas is by far the most common fuel type in the data.

The next visualisation is to illustrate whether there is a correlation between engine size and horsepower.

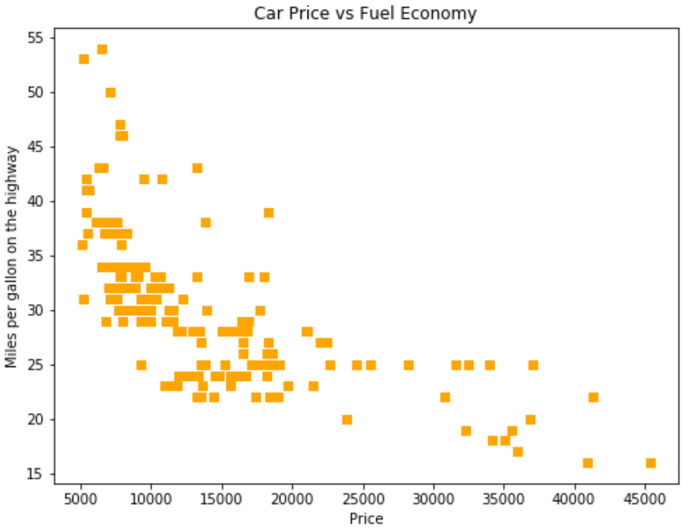


The visualisation shows that there is a clear and positive correlation between horsepower and engine size. Which makes sense.

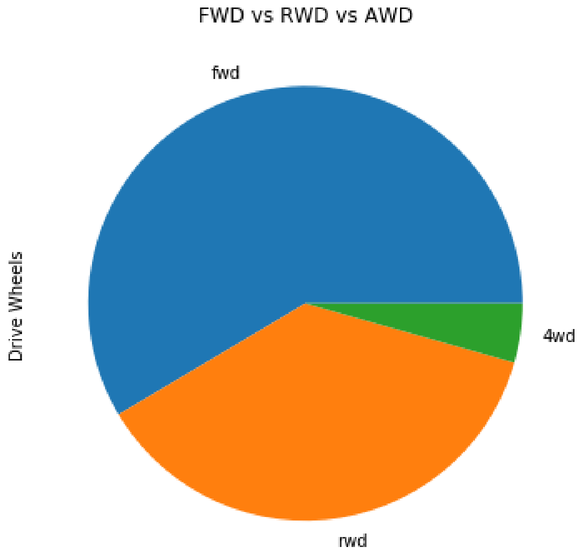
I chose to create the below visualisation because I wanted to see if there was a relationship between the price of a car and fuel efficiency. The findings are quite interesting as you can see from the figure below the trend seems to be that the cheaper a car is the more fuel efficient the car is.

I believe this to be the case because, manufacturers know when developing a car before it is put into production how much they are going to charge for the car. So, if it is a more budget orientated car the manufacturers know that the people buying this car cannot or will not want to spend a lot of money on petrol. Meaning if manufacturers want to be competitive in this section of the market (the biggest section of the market) they need to create cheap and fuel-efficient vehicles.

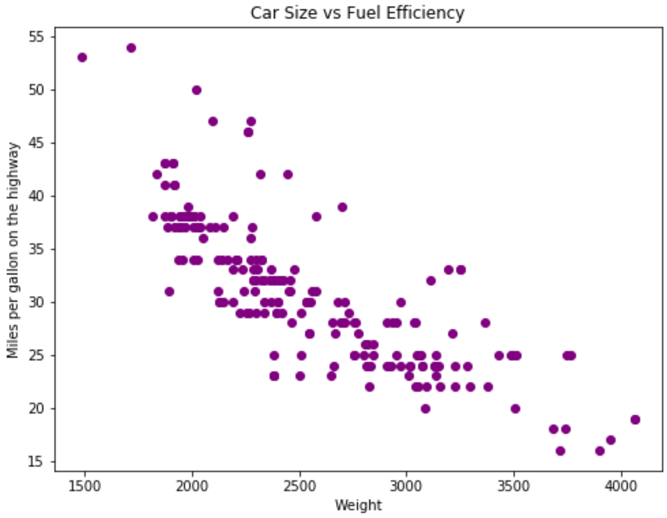
The inverse is true for expensive cars. The target market for more expensive vehicles obviously have more money to spend on petrol as they have more money to spend on a car. However, other factors must also be considered such as the size of the vehicle. Bigger vehicles use more petrol (more on that later). Additionally, more expensive cars may be faster and have more horsepower and therefore command a higher price and use more petrol.



The next visualisation is to illustrate the split between the different drive wheels of all cars in the database. I chose a pie graph for this visualisation as pie charts are the best way to show the split of certain values in a whole dataset.



As we can see from the figure above forward wheel drive is the most common by far. I believe this to be the case because forward wheel drive cars are the cheapest to manufacture. Forward wheel drive is then followed by rear wheel drive and then all wheel drive.



The figure above clearly shows a negative correlation between weight and miles per gallon on the highway. This is because the heavier a car is, the more energy is required to get the car to move. Thus, more fuel is consumed leading to higher fuel consumption. In addition, bigger cars have larger engines as again they require more energy to move and a larger engine obviously uses more fuel. These two factors result in the trend in the scatter plot.

THIS REPORT WAS WRITTEN BY : Matthew van der Westhuizen