11 April 2019

ASSIGNMENT I

PRACTICAL DATA SCIENCE - COSC 2670

A. DATA PREPARATION

1. POTENTIAL ISSUES/ERRORS

a) Typos/Data Entry Error

Most errors of this type are easy to fix with "replace()"

For example:

```
automobile['symboling'].replace(4, -3, inplace=True)
```

b) Extra Whitespaces

Whitespaces tend to be hard to detect but cause errors like other redundant characters would.

A String function that will remove the leading and trailing whitespaces and can be apply to whole dataframe once.

For example:

```
def strip_obj(col):
    if col.dtypes == object:
        return (col.astype(str)
                   .str.strip()
                   .replace({'nan': np.nan}))
    return col

automobile = automobile.apply(strip_obj, axis=0)
```

c) Lowercase/Uppercase

Can be fix using "lower()" or "upper"

For example:

```
automobile['make'] = automobile['make'].str.lower()
```

d) Impossible values and sanity checks

Sanity checks are another valuable type of data check where you check the value against physically or theoretically impossible values such as people taller than 3 meters or someone with an age of 299 years.

e) Missing values

Missing values aren't necessarily wrong, but you still need to handle them separately.

There are many ways to deal with missing values like omitting the value, impute a static value such as 0 or mean, impute a value from an estimated or theoretical distribution.

For example:

```
automobile['stroke'].fillna(automobile['stroke'].mean(axis=0),inplace=True)
```

f) Outliers

An outlier is an observation that seems to be distant from other observations;

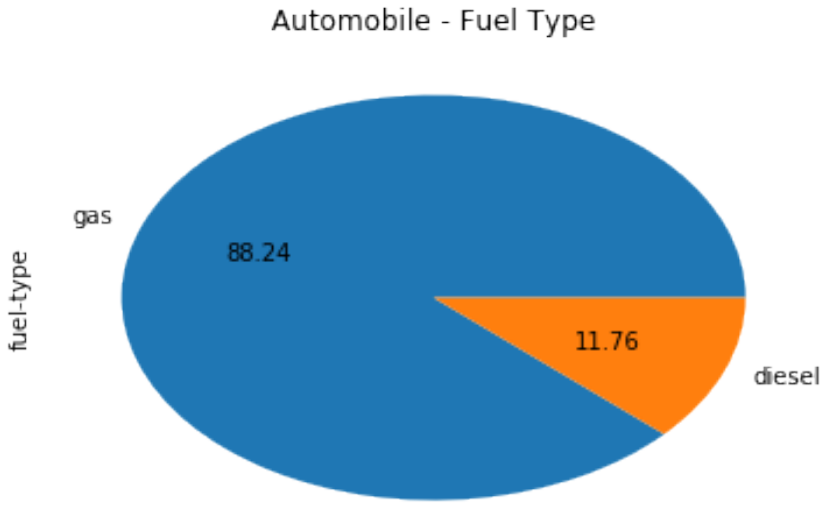
More specifically, one observation that follows a different logic or generative process than the other observations.

The easiest way to find outliers is to user a plot or a table with the minimum or maximum value. Outliers can be removed or replaced with mean values.

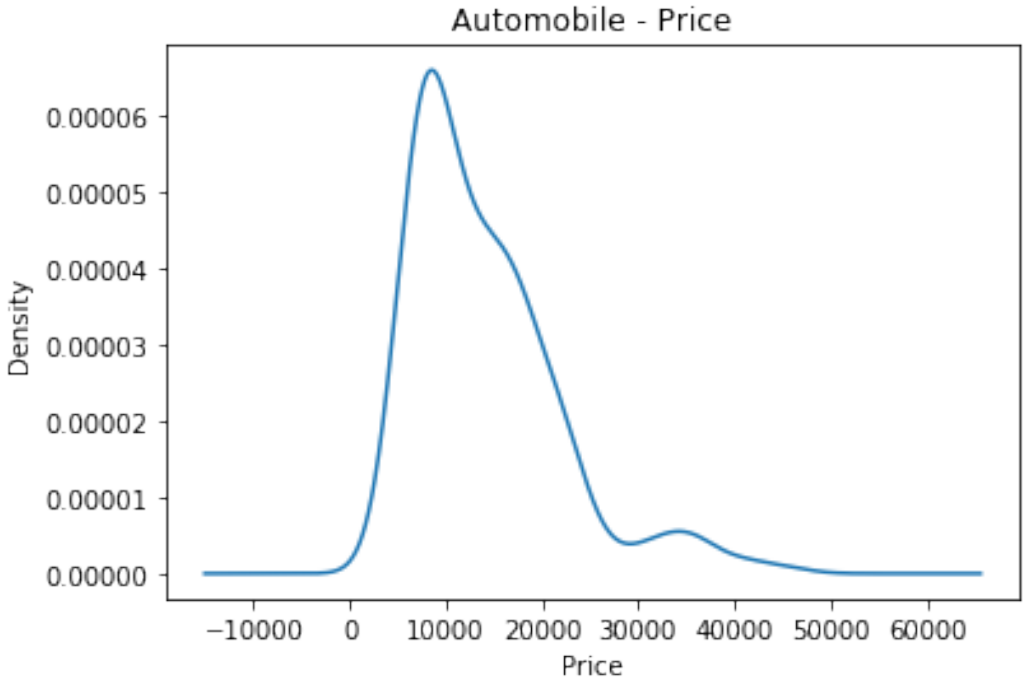
B. DATA EXPLORATION

1. DATA VISUALIZATION

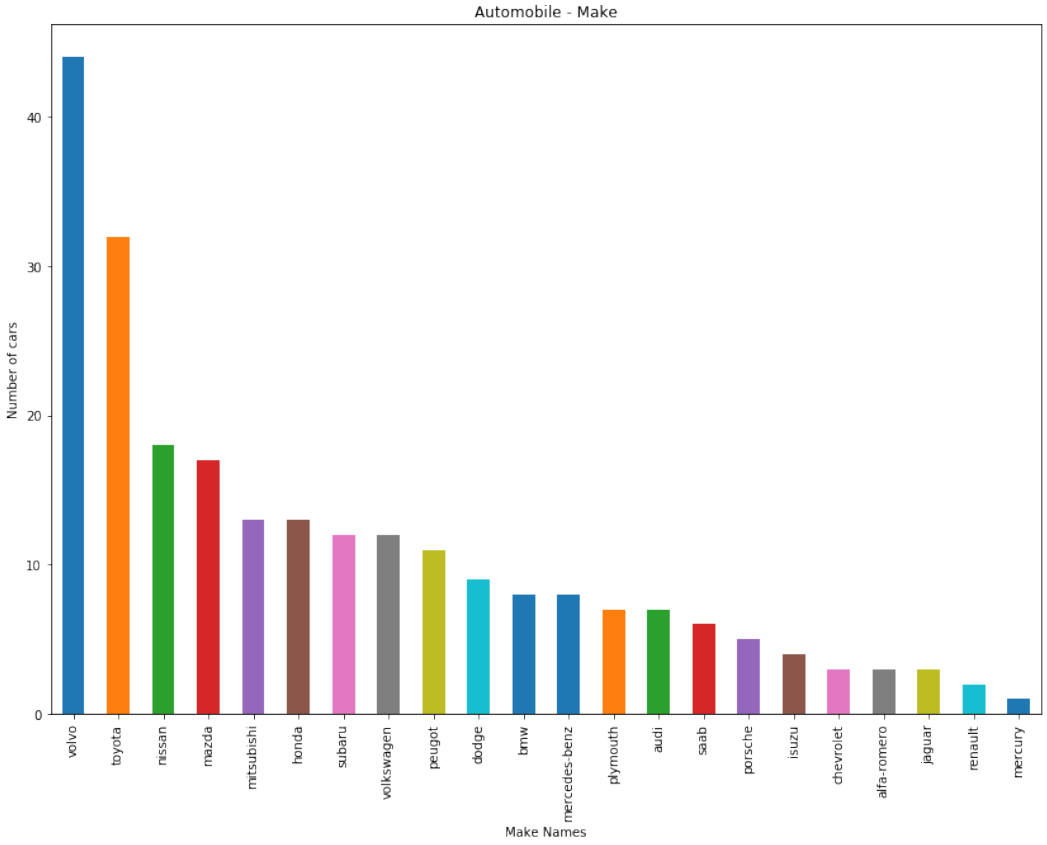
Pie chart is chosen because there are only two components: gas and diesel and the percentages are significantly different. Almost 90% of automobile use gas.



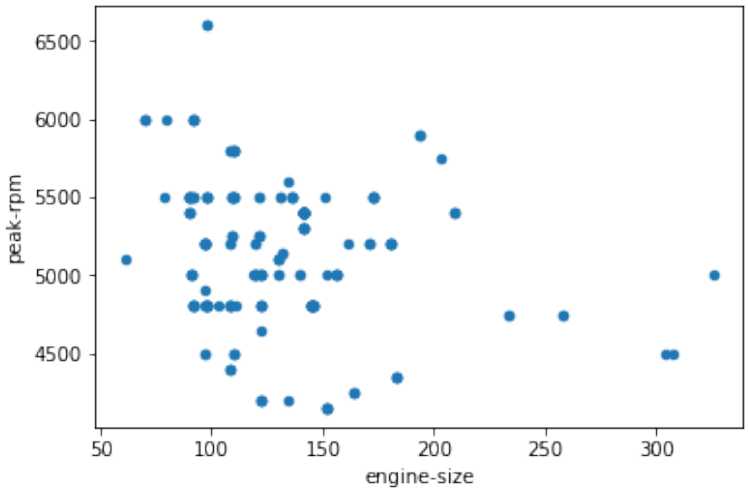
Density chart is chosen because it is good for showing the distribution of data over a range. Most of automobile price ranges between $8000 and $16000.



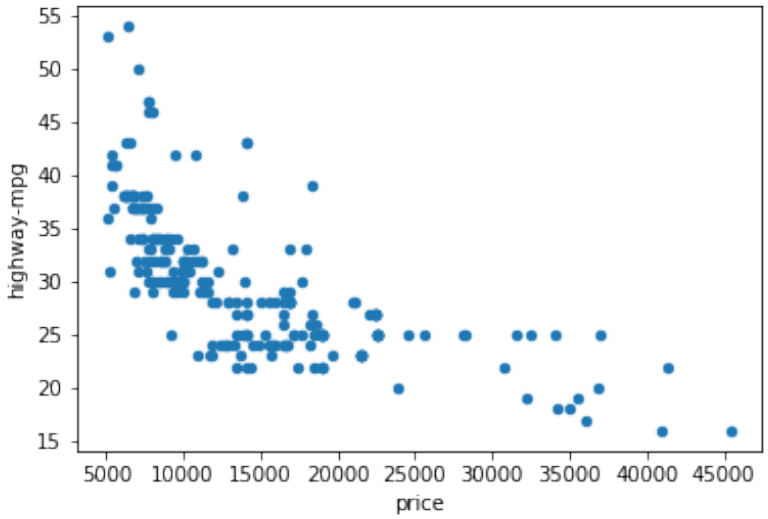
The bar chart is chosen to visualise the data because the number of categories is big and there are many similar values.



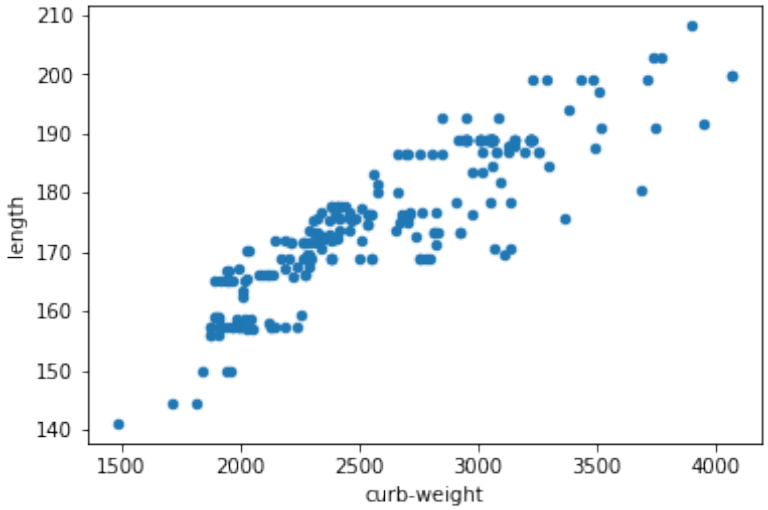
2. DATA ANALYSING



From the plot we can see there is no obervable relationship between peak-rpm and engine-size.

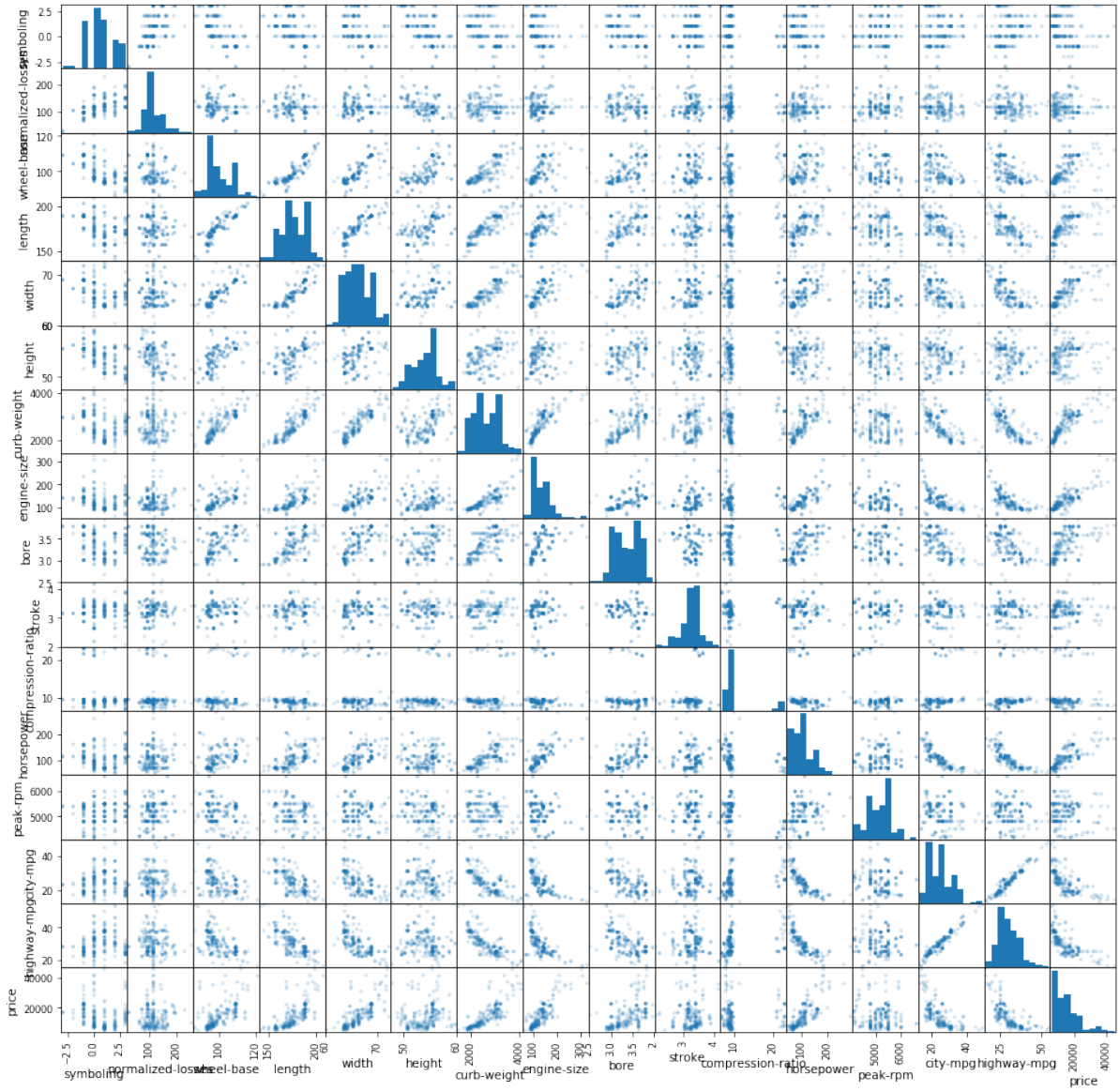


From the plot we can see the there is a relationship between price and miles per gallon for highway drive. It seems like expensive automobiles are not as energy efficient as cheap ones. The more expensive, the more gallon needed for the same length of highway.



From the plot we can see there is a visible relationship between the curb weight and length of the automobiles. Curb weights and vehicle length are proportionate.

3. SCATTER MATIXES


symboling
normalized-losses
wheel-base
length
width
height
curb-weight
engine-size
bore
stroke
compression-ratio
horsepower
peak-rpm
city-mpg
highway-mpg
price