Report

DATA PREPARATION

1)TYPOS

Typos is a mistake made in typed or printed text.There were many typing errors in dataset like vol00112ov' instead of volvo in attribute(make). These errors are removed using pandas's replace function.

Method used for replacing typos:
dataframe.replace()

Command used for correcting volvo in make attribute:

```
df['aspiration'].replace(['turrrrbo'], ['turbo'], inplace = True)
```

Other Examples of typos in dataset are :

- peugot   instead of Peugeot in make attribute
- alfa-romero   instead of Alfa-romeo in make attribute
- turrrrbo   instead of turbo in aspiration attribute

2)EXTRA WHITESPACE

Extrawhite space in dataset occurs before or after the string like ' volvo ' and are replaced using
the method provided by pandas library.

Method used for removing extra whitespace : **Pandas Series.str.strip()**

Command used in code: df['make'] = df['make'].str.strip()

Other examples: 'std ', 'four ' and many more

3)MISSING VALUES:

Missing values are replaced using the mean if does not effect the result of our goal like horsepower attribute has two missing values. Following are the attribute with respective no of missing values:

normalized-losses	: 47
stroke	: 4
bore	: 4
horsepower	: 2
peak-rpm	: 2

```
price              : 4
```

Command used:

```
df['horsepower'].fillna(df['horsepower'].mean(axis=0),inplace=True)
```

Above method is for interval values but for nominal or ordinal value mean cannot be applied so mode is applied, for example
In num of door attribute.

Command used for nominal attribute:

```
df['num-of-doors'].mode()
df['num-of-doors'] = df['num-of-doors'].replace(np.nan, 'four')
```

4)IMPOSSIBLE VALUES:

Impossible values in the dataset can be guessed from the range of that attribute which was given alongside the dataset, for example symboling attribute varies from -3 to 3 but values with 4 are impossible values.

Command used for removing impossible values in symboling attribute

```
df.drop(df.loc[df['symboling']==4].index, inplace=True)
```

5)
Changing datatype of attributes

Datatype of attribute having ordinal data are changed using pd.categorical function. For example in symboling attribute's data was int and so with this command it changed into categorical.

```
df['symboling'] = pd.Categorical(df['symboling'], categories = [-2,   -1,0 ,1,2,  3], ordered = True)
```

6) Outliers

There are few outlier like in compression-ratio, wheel base and in some other columns as it does have impact on result I'm analysis so there are left as it is.

Data Exploration

subsection 1

chart 1: for attribute with numerical value

Below chart is a histogram of horsepower attribute. This attribute is chosen to conduct further analysis with other attribute and to show that most of cars have horse power between 50 to 125

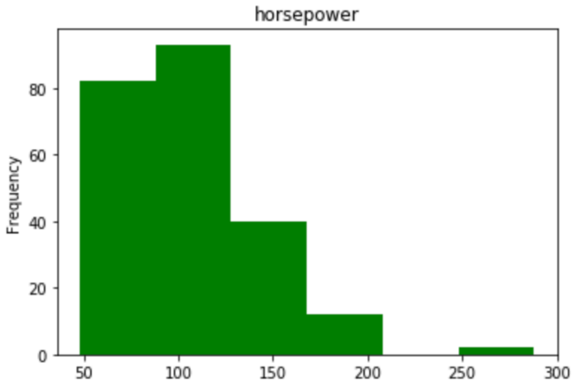


chart 2: for attribute with nominal value

Pie chart is chosen to analyze how many doors in a car are popular.

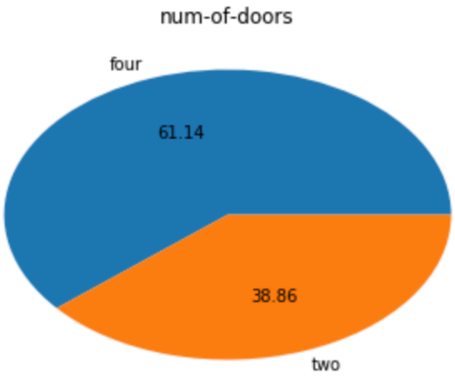
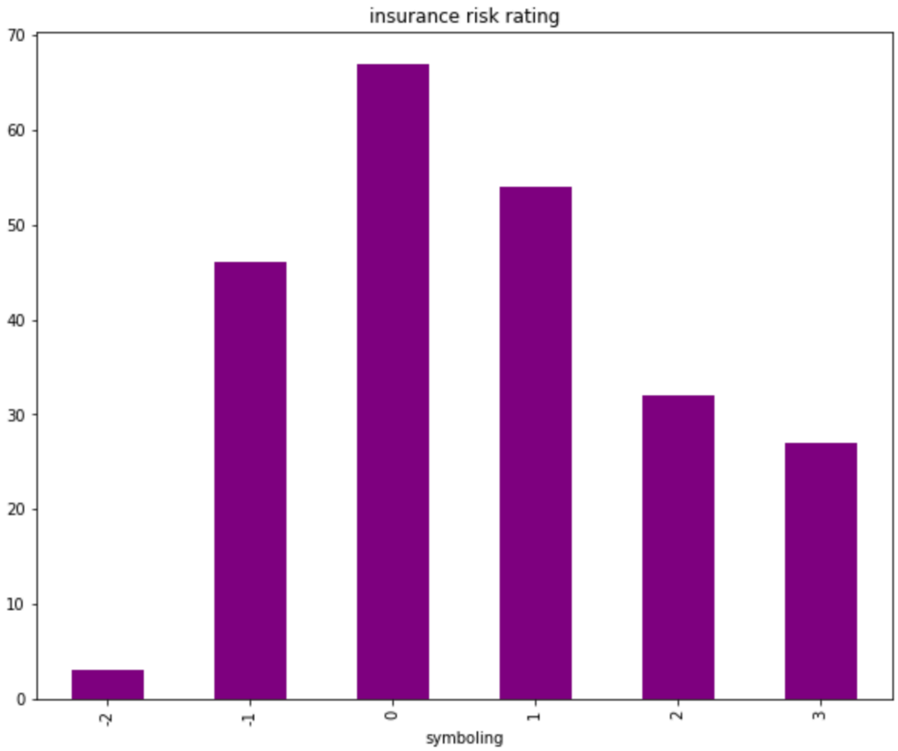


chart 3: for attribute with ordinal value

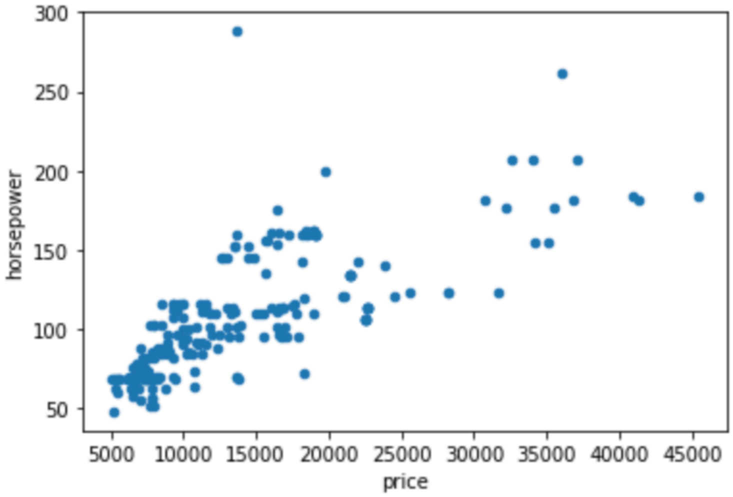
This bar chart show that very safe vehicle with higher insurance rating are less compared to vehicle which are high in number.



subsection 2:

subsection 2.1

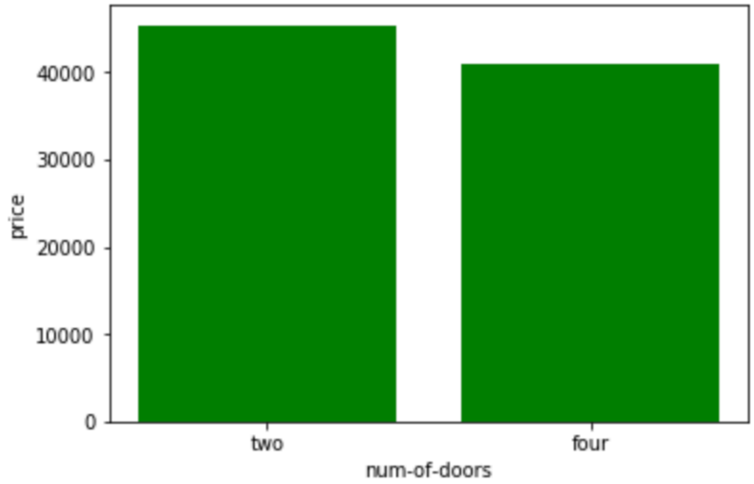
Plausible hypothesis is that price of car is increase with increase in hypothesis. Scatter plot clearly show that price of car is increases with horsepower



subsection 2.2

Hypothesis : There is a relationship between num of doors and price of car

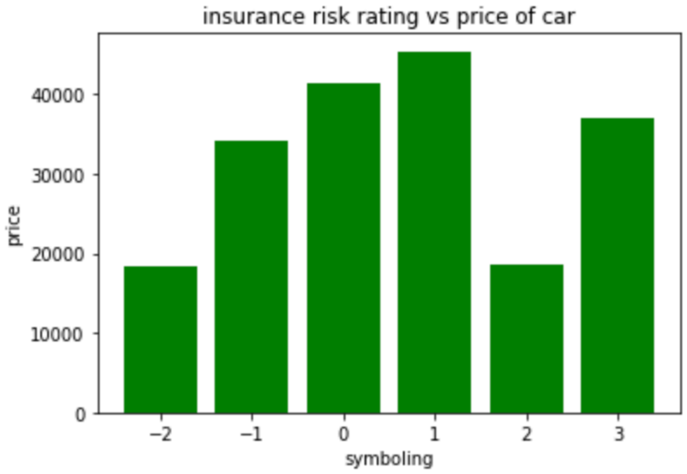
Its a plausible hypnosis as it is clearly shown in the chart that car with two doors are expensive



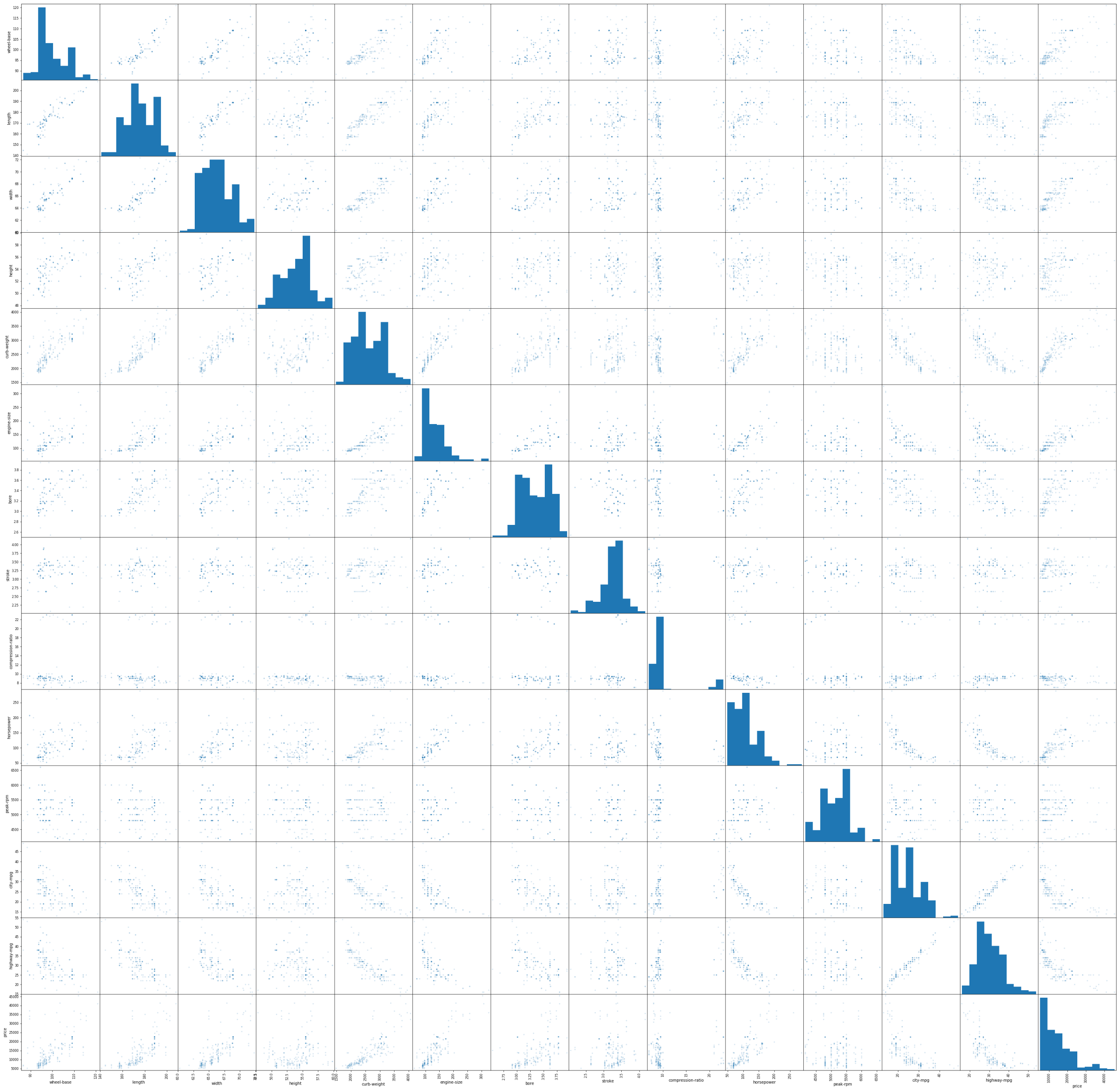
subsection 2.3

Hypothesis : There is a relationship between symboling and price of car

Its is a plausible hypnosis as it is clearly shown in the chart that there is a relationship between Safety of car and price expect at symboling(value =2). It clearly show that as price increases risk factor also increases except at 2.



Section 3



- Vehicle Mileage decrease as increase in Horsepower , engine-size, Curb Weight
- As horsepower increase the engine size increases
- Curbweight increases with the increase in Engine Size