Statistics for Data Science (*UE19CS203*)

ANALYSIS OF A USED CARS DATASET:

ABSTRACT:

We performed a statistical analysis on a used car dataset with over 2000 rows and 15 attributes. The main objective of our study was to identify the factors which positively and negatively affected the sale price of used cars in the United States. First, we cleaned the data, removed the duplicate values, converted all the categorical values to a single case and imputed some of the missing data using mode. We then performed an exploratory data analysis for which we plotted a box plot, identified the outliers in the price features and removed them. We also visualized our data and the relationship between its features using graphs such as heat maps and bar charts. Next, we normalized the numerical columns and plotted histograms to visualize their distribution. We then standardized the numerical columns to get a mean 0 and variance 1. Finally, we performed a hypothesis test on one of the features and concluded our analysis.

INTRODUCTION:

Our problem statement is to identify the factors that positively and negatively affect the sales price of a used car in the United States. These factors include the miles travelled by the car, its fuel type, year of manufacture and so on. Our study focuses on the relationship of the price of the car with these variables and the magnitude of this corelation. This project involves – data pre-processing, data-cleaning, data-preparation, data-analysis, and hypothesis-testing.

DATASET:

The input dataset is a XLS file, obtained from www.kaggle.com and it comes from Craigslist which contains a list of all the entries made over the years of used vehicles on sale in the United States. It contains over 2000 records, with 25 features that describe the car being put up on sale. These include categorical data, numeric discrete data and numeric continuous data.
The primary features relevant to our study include:

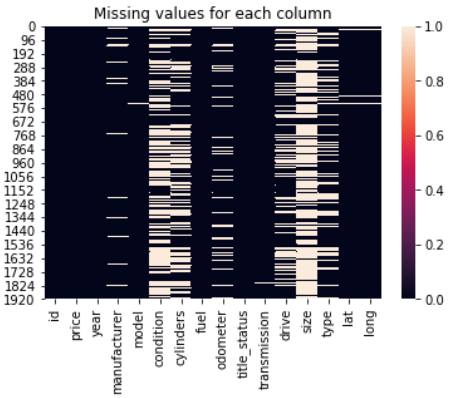
- id
- entry price (price)
- entry year (year)
- manufacturer of vehicle (manufacture)
- model of the vehicle (model)
- condition of the vehicle (condition)
- no. of cylinders (cylinders)
- fuel type (fuel)
- miles travelled by the vehicle (odometer)
- title status of the vehicle (title_status)
- transmission of vehicle (transmission)

- type of drive (drive)
- size of the vehicle (size)
- generic type of vehicle (type)
- latitude (lat) and longitude (lat) of the listing.

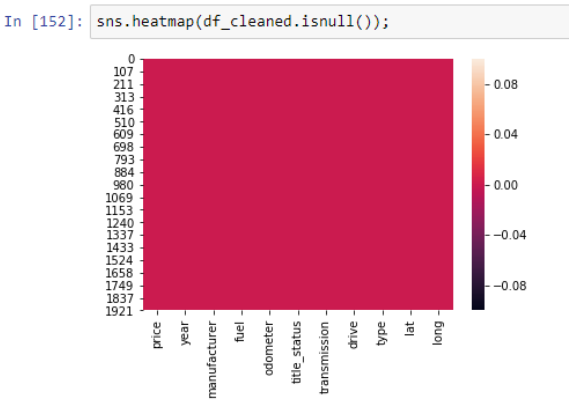
PREPROCESSING OR DATA CLEANING:

The input dataset was analysed for its inaccuracies, incompleteness, redundancies, and/or validity. First, we dropped off all the columns that were not significant to our study. Then we identified the number of missing values in each column and plotted a heatmap to visualize the same.
Following this, we defined a function to remove columns that had more than 40% of their data missing. On identifying that all of the cars were put up on sale between the years 1900 and 2020, we removed all values outside this range. Then we replaced all the missing values in the categorical columns with their respective modes. The remaining missing values in the dataset were dropped as they formed only upto 3% of the dataset. Next, we changed all the values in the categorical column “type” to one case as there were irregularities in the entries. We then plotted a boxplot to identify all the outliers in the “price” column, following which we defined a function to remove these outliers using the upper, lower and inter-quartile range. This data cleaning process was crucial to our analysis as our dataset had a number of missing values and inconsistent entries which may lead us to a wrong result. By removing outliers, imputing the missing categorical variables and dropping the remaining values our dataset became more efficient to work with.

BEFORE DATA CLEANING



AFTER DATA CLEANING

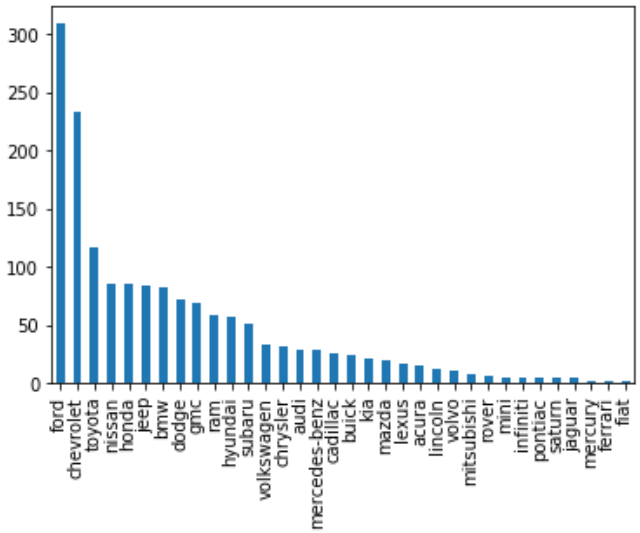


EXPLORATORY DATA ANALYSIS:

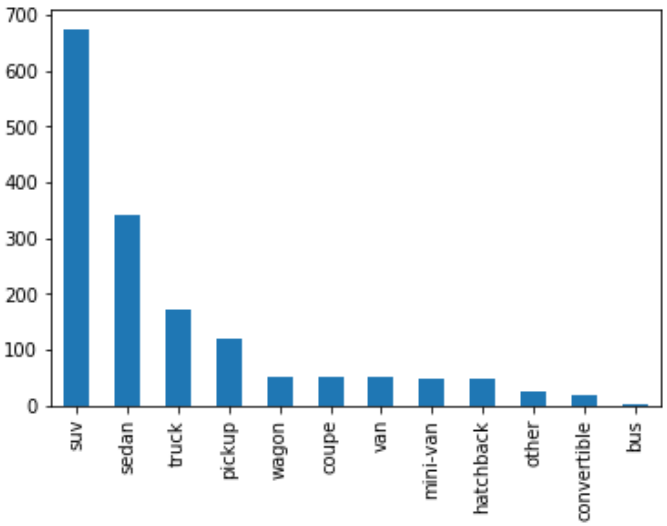
Using the bar plot, we identified that the cars manufactured by Ford are the highest in number to be put up on sale after use, and those manufactured by Fiat the lowest.

We also identify that the type of used vehicle that was put up on sale the most is suv's, followed by sedans and trucks.

```
In [169]: df_cleaned['manufacturer'].value_counts().plot(kind='bar')
Out[169]: <matplotlib.axes._subplots.AxesSubplot at 0x1a185c6cd0>
```

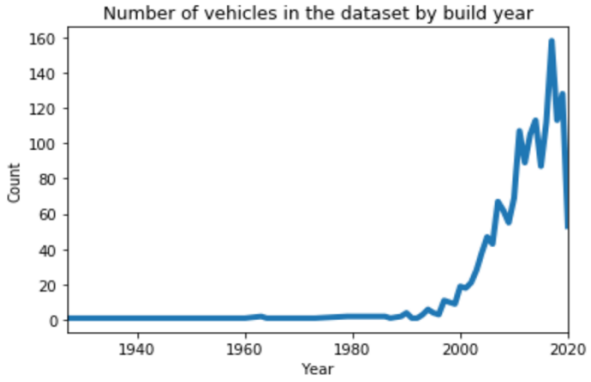



```
In [170]: df_cleaned['type'].value_counts().plot(kind='bar')
Out[170]: <matplotlib.axes._subplots.AxesSubplot at 0x1a17c5ac50>
```

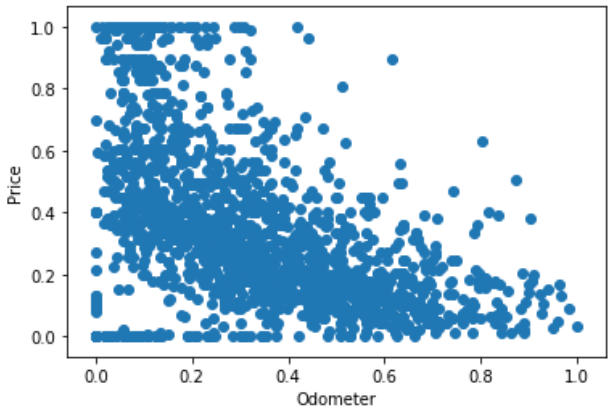


From this graph we can infer that most of the cars were put up for sale between the years 2010 and 2020.

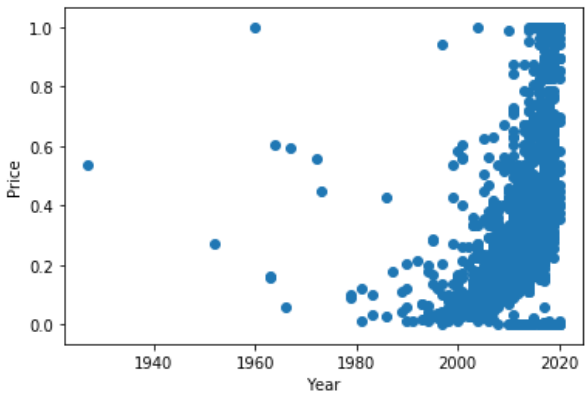
```
In [39]: df_cleaned[df_cleaned.year>=1900].year.value_counts().sort_index().plot(lw = 4)
         plt.title("Number of vehicles in the dataset by build year")
         plt.xlabel("Year")
         plt.ylabel("Count")
         plt.show()
```



Next, we plotted a scatter plot of price against odometer and observed that as the miles driven by the car increased, the sales price of it decreased.

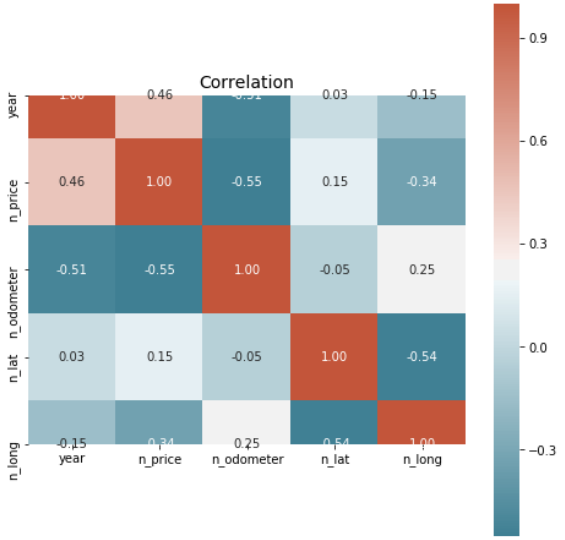


We also plotted a scatter plot of price against entry year and found that the used car sales price was higher for cars that were put up for sale in the more recent years, as compared to those put up years ago.



CORRELATION:

The correlation coefficient measures how strong a linear relationship exists between two numeric variables x and y.

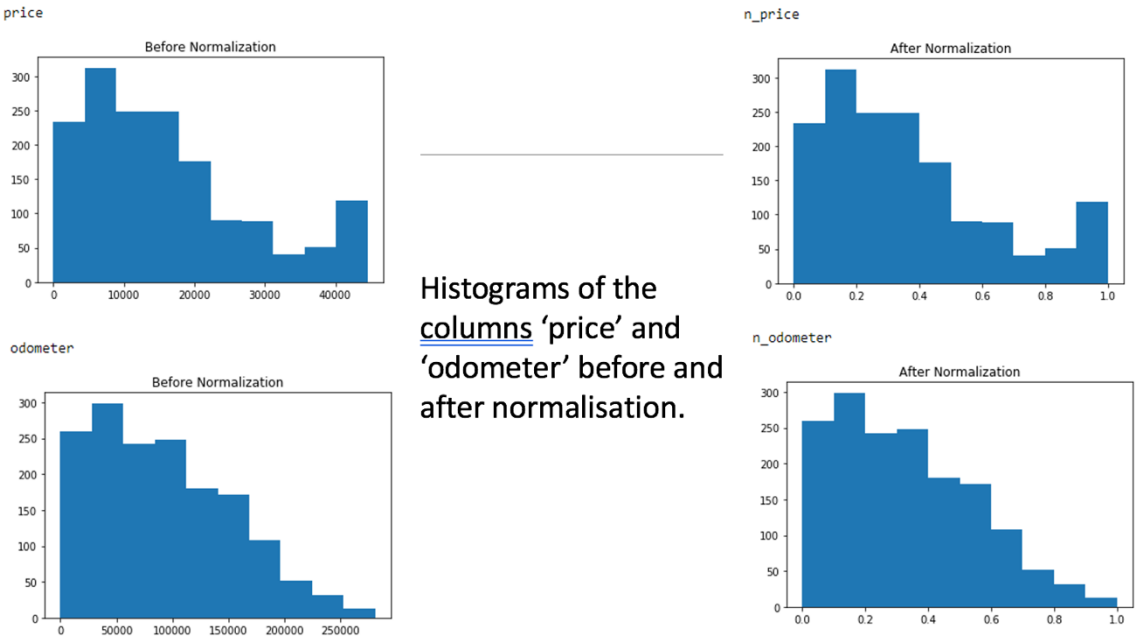


On plotting a heat map for all the numerical columns, we can infer that:

- Price is positively correlated to year with a correlation factor of 0.46. Therefore, as the year of entry of the car is more recent (higher), its sales price will be higher.
- Price is negatively correlated to odometer with a correlation factor of -0.55. Therefore, as the number of miles driven by the car increases, its sale price decreases.

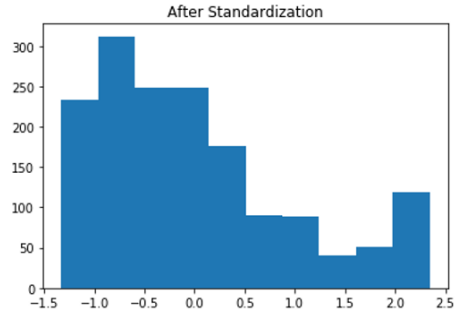
NORMALIZATION AND STANDARDIZATION:

We normalized all the numerical columns other than the "year" column to values between 0 and 1. Normalization is essential because we want to convert the values measured on different scales to one common scale, in order to make better and more accurate comparisons. We also standardized the numerical columns to have a mean of 0 and variance of 1.

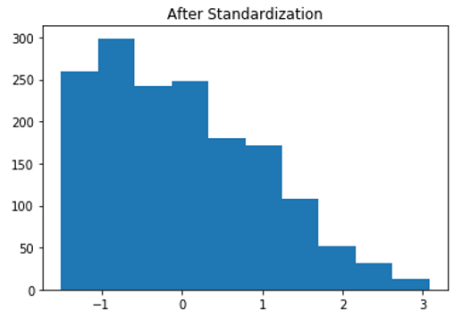


After standardization:

price



odometer



HYPOTHESIS TESTING:

We performed hypothesis testing using the following hypothesis:
We assume that the sales price of used suv's is low. We assume that the average sales price of used suv's is not more than $16110.35.

Null Hypothesis: H_0 : $\mu <= 16110.35$

Alternate Hypothesis: H_1 : $\mu > 16110.35$

Significance level: 0.05
We took a sample of 300 observations from the dataset, for *type*= "suv".
The sample mean is computed using the mean of *price* of this sample.

In [44]:
```
from scipy.stats import norm
def one_sided_hypo(sample_mean, pop_mean, pop_std, n, alpha):
    actual_z = abs(norm.ppf(alpha))
    hypo_z = (sample_mean - pop_mean) / (pop_std/np.sqrt(n))
    print('actual z value :', actual_z)
    print('hypothesis z value :', hypo_z, '\n')
    if hypo_z >= actual_z:
        return True
    else:
        return False

df_x = df_cleaned.loc[df_cleaned['type'] == 'suv']
alpha = 0.05
sample = df_x['price'].sample(300)
sample_mean = sample.mean()
df_x = df_cleaned.loc[df_cleaned['type'] == 'suv']
pop_mean = df_cleaned['price'].mean()
n =  300
pop_std = df_cleaned['price'].std()

print('H0 : μ <=', pop_mean)
print('H1 : μ >', pop_mean)
print('alpha value is :', alpha, '\n')
```

```
reject = one_sided_hypo(sample_mean, pop_mean, pop_std, n, alpha)
if reject:
    print('Reject NULL hypothesis')
else:
    print('Cannot reject NULL hypothesis')
```

```
H0 : μ <= 16110.347663551402
H1 : μ > 16110.347663551402
alpha value is : 0.05

actual z value : 1.6448536269514729
hypothesis z value : -1.1788614469018992

Cannot reject NULL hypothesis
```

On performing the z hypothesis test, we obtain the actual z value as 1.64 and the hypothesis z value as -1.18.
Since the obtained hypothesis z value is less than the actual z value, we cannot reject the Null hypothesis, and therefore we conclude that **it is plausible that the sales price of used suv's is low.**

RESULTS AND DISCUSSION:

The results of the exploratory data analysis is as follows:
Used cars that were put up on sale on Craigslist in the United States between the years 1900 and 2020 had their entry price mainly dependant on two of their features:

- The year in which the car was put up on sale.
- The odometer or the miles travelled by the car before it was put up on sale.

Through our analysis, we conclude that the entry or sales price of a used car is **positively correlated** to its entry year, as cars put up in more recent years will be of higher value than of those put up years ago.
With regards to odometer, we conclude that the entry price of a car is **negatively correlated** to its odometer or the number of miles travelled by the car, as a car that has been used less will be of higher value.
We also observed that cars manufactured by Ford are the highest to be put up on sale in the U.S after use, and those manufactured by Fiat the lowest. The type of car that was mostly put up were suv's, followed by sedans and most of the entries were made between the years 2010 and 2020. With respect to the hypothesis test we conducted, we concluded that it is plausible that the sales price of used suv's is not more than $16110.35, as we could not reject the null hypothesis from our z-hypothesis test.