BIKE SHARING CASE STUDY

Question Answers

SEPTEMBER 14, 2022

IIIT B

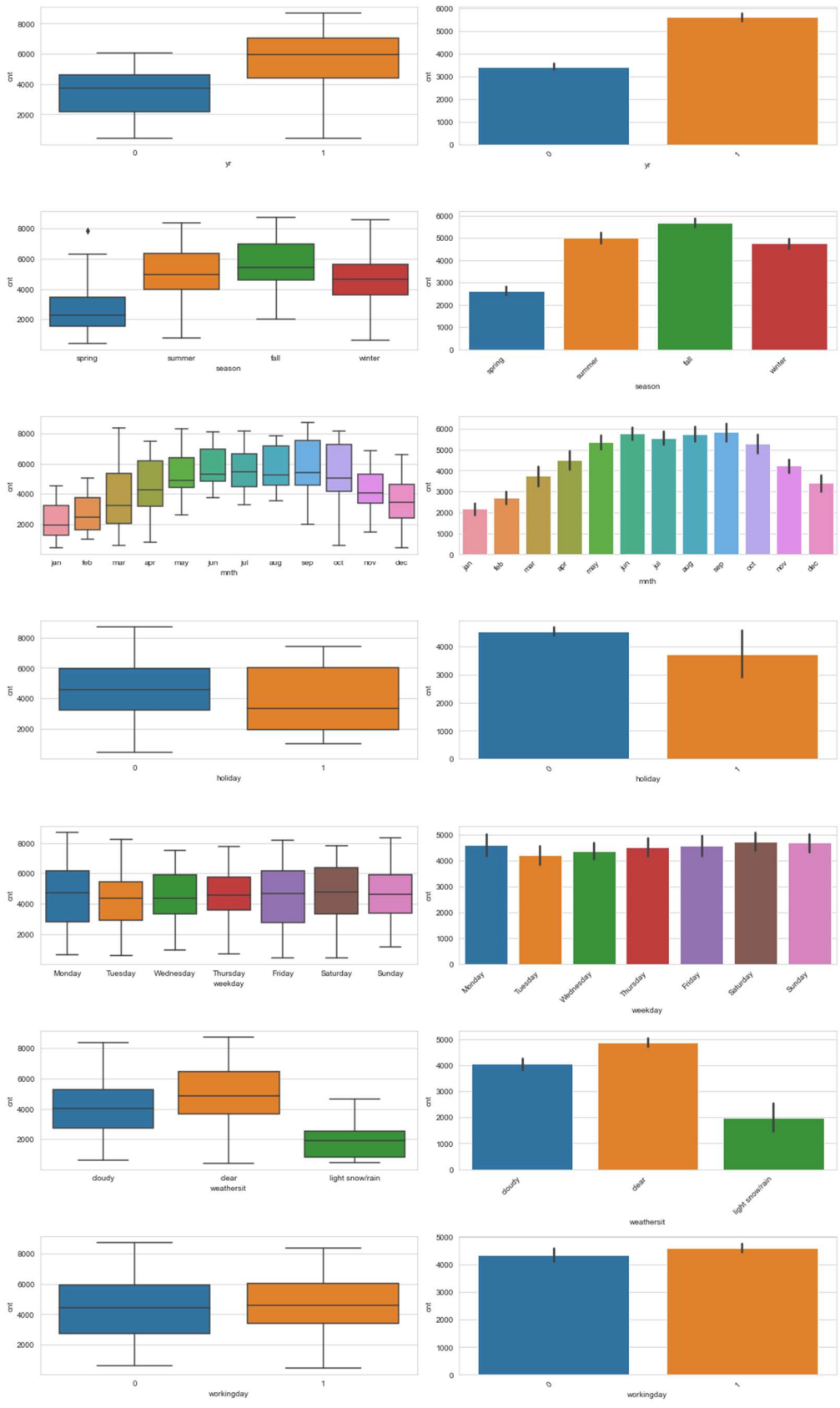
Contents

Assignment-based Subjective

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
   Answer:
   The bike sharing data has the following categorical columns:
   1. yr
   2. season
   3. mnth
   4. holiday
   5. weekday
   6. weathersit
   7. workingday

By plotting these variables, we can infer the following:

1. The highest impact on bike sharing is of year(yr) which could be the result of them gaining popularity over the years.
2. The most popular season is fall and least is spring.
3. The demand increases through the year and dips at year end.
4. The demand is more on days which are not a holiday
5. The demand is high on Misty and Cloudy, followed by Light snow/Rain and least during heavy Snow/Rain.


cnt
yr
0
1
season
spring
summer
fall
winter
mnth
jan
feb
mar
apr
may
jun
jul
aug
sep
oct
nov
dec
holiday
weekday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
weathersit
cloudy
clear
light snow/rain
workingday

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

   Answer:
   "drop_first=True" ultimately, helps us get rid of multicollinearity issue that will arise as we need only n-1 variables to interpret n variables. The pandas function "get_dummies" function is nothing but one-hot encoding which lets us define a column for each category. However, for any categorical column having n unique values, n-1 columns are enough to deduce the category.
   For example, for a car classification, having categories "Hatchback", "Sedan" and "SUV", a value 1 identifies it.

Hatchback	Sedan	SUV
1	0	0
0	1	0
0	0	1

The above data can be easily interpreted instead using the below:

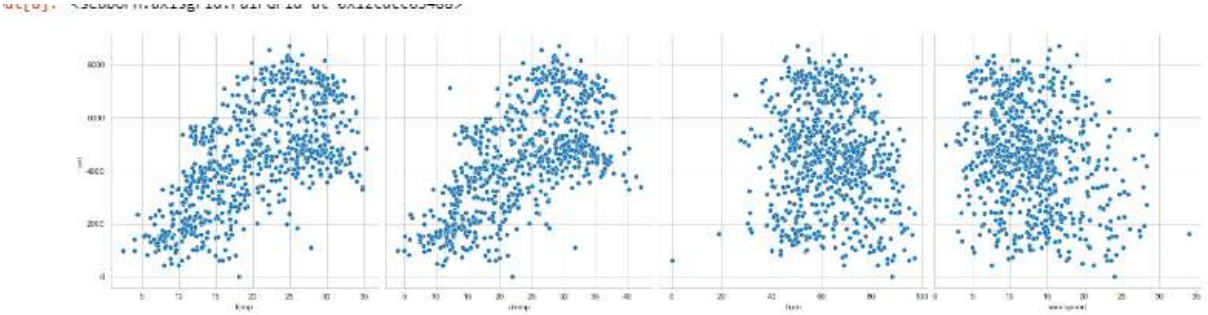
Sedan	SUV
0	0
1	0
0	1

Where Sedan=0 and SUV=0 , represents a Hatchback.

With drop-first=True , we are reducing the column by one, aiding the OLS prediction as well.

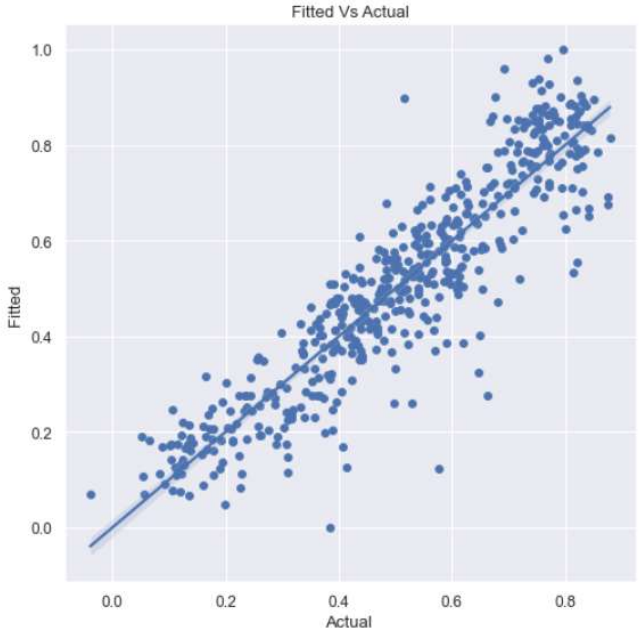
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
   Answer:

   Based on pairplot, we observe "temp" and "atemp" have the highest co-relation with target variable. Considering the outliers, "temp" has the highest.

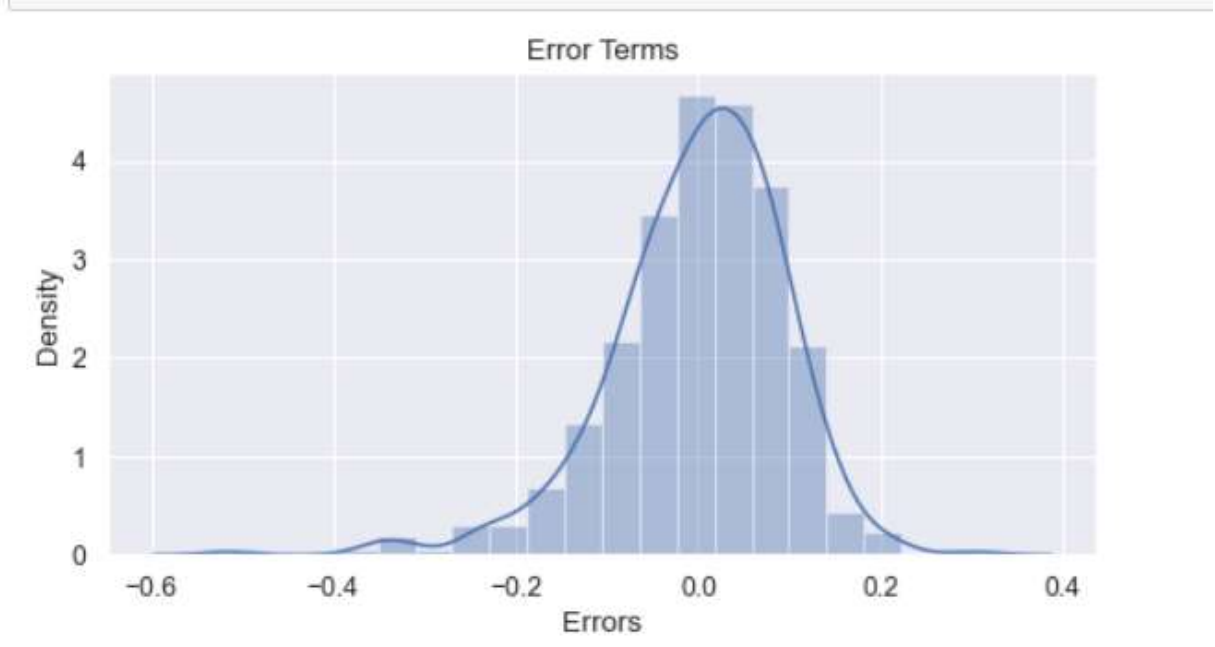


4. How did you validate the assumptions of Linear Regression after building the model on the raining set?
   Answer:
   The following are the assumptions of Linear Regression that need to be validated:
   1. There is a linear relationship between target and dependent variables (X and Y).

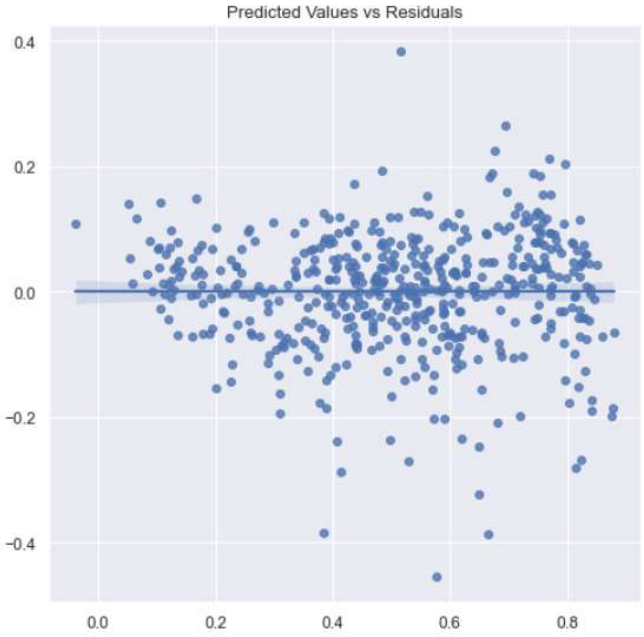
A plot of the predicted value versus the actual value follows a linear trend, supporting the assumption of linearity.



2. Error terms are *normally distributed* with mean zero(not X, Y):
The below plot shows that mean is about zero. Also, the calculated mean square value is about -1.3259545966650644e-15 which can be approximated to zero.



3. Error terms are independent of each other:
   The below graph shows the error terms do not follow a pattern and are independent of each other.



4. Error terms have constant variance (homoscedasticity):
   As can be seen in the diagram above(Point 3), the residuals are have equal or almost equal variance across the regression line.

   The final model has a r2-square value of train data is 0.828422243178494 while r2-square value of test data is 0.8075460660066102 and very low F-statistic. Also, VIF is below 5 for all variables.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   Answer:
   The final model coefficients indicate the most weighted variables are:
   1. Temperature(temp)
   2. Year (yr)
   3. Weather light now/Rainy (affecting negatively)

```
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.828
Model:                            OLS   Adj. R-squared:                  0.825
Method:                 Least Squares   F-statistic:                     268.2
Date:                Wed, 14 Sep 2022   Prob (F-statistic):          4.69e-185
Time:                        21:12:29   Log-Likelihood:                488.00
No. Observations:                 510   AIC:                            -956.0
Df Residuals:                     500   BIC:                            -913.7
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.2909      0.019     15.136      0.000       0.253       0.329
yr             0.2357      0.008     28.102      0.000       0.219       0.252
temp           0.4017      0.026     15.280      0.000       0.350       0.453
windspeed     -0.1499      0.025     -5.931      0.000      -0.200      -0.100
spring        -0.1360      0.013    -10.718      0.000      -0.161      -0.111
cloudy        -0.0810      0.009     -9.074      0.000      -0.099      -0.063
light snow_rain -0.2885    0.025    -11.369      0.000      -0.338      -0.239
jul           -0.0669      0.018     -3.769      0.000      -0.102      -0.032
oct            0.0567      0.016      3.646      0.000       0.026       0.087
sep            0.0606      0.016      3.750      0.000       0.029       0.092
==============================================================================
Omnibus:                       75.897   Durbin-Watson:                   1.955
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              199.999
Skew:                          -0.742   Prob(JB):                     3.72e-44
Kurtosis:                       5.686   Cond. No.                         11.2
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Answer:
   Regression in statistics mean to find co-relation between mean value of one variable (normally, the dependent variable) and corresponding value of other variables (independent variables). Establishing a linear relationship between the dependent and independent variables is termed as linear regression.
   Linear Regression is achieved using supervised machine learning algorithm and is known as linear regression algorithm and is simply defined as below:

   y = m*x +c
   where y is the dependent or target variable
   m is the slope coefficient

c is the intercept
x is the independent variable

This is simple linear algorithm. We can have multiple independent variables to predict value of the dependent variables in which case we use the multiple linear regression algorithm:

y = m1 * x1 + m2*x2+ m3* x3+….. +c
where y is the dependent or target variable
m1, m2, m3.. mn are the slope coefficients of n independent variables
c is the intercept
x1,x2,x3…xn are the values of the n independent variables.

Based on above algorithm, we try to determine the coefficients values such that the difference between the predicted value (obtained from the equation above) and the actual value is minimum.

2. Explain the Anscombe's quartet in detail.
   Answer:

   Anscombe's quartet is great way to demonstrate the importance of graphical visualization in the process of data analysis. Via graphical representation of 4 different datasets with identical statistics, she proved how the data has very different distribution and can lead to different inferences.

   Statistics for dataset Anscombe from sns:

```
In [12]: import numpy as np
         import pandas as pd
         pd.pivot_table(df,values=['x','y'],index='dataset',aggfunc=np.mean)
```

Out[12]:

dataset	x	y
I	9.0	7.500909
II	9.0	7.500909
III	9.0	7.500000
IV	9.0	7.500909

```
In [21]: import numpy as np
         import pandas as pd
         pd.pivot_table(df,values=['x','y'],index='dataset',aggfunc=np.var)
```

Out[21]:

dataset	x	y
I	11.0	4.127269
II	11.0	4.127629
III	11.0	4.122620
IV	11.0	4.123249

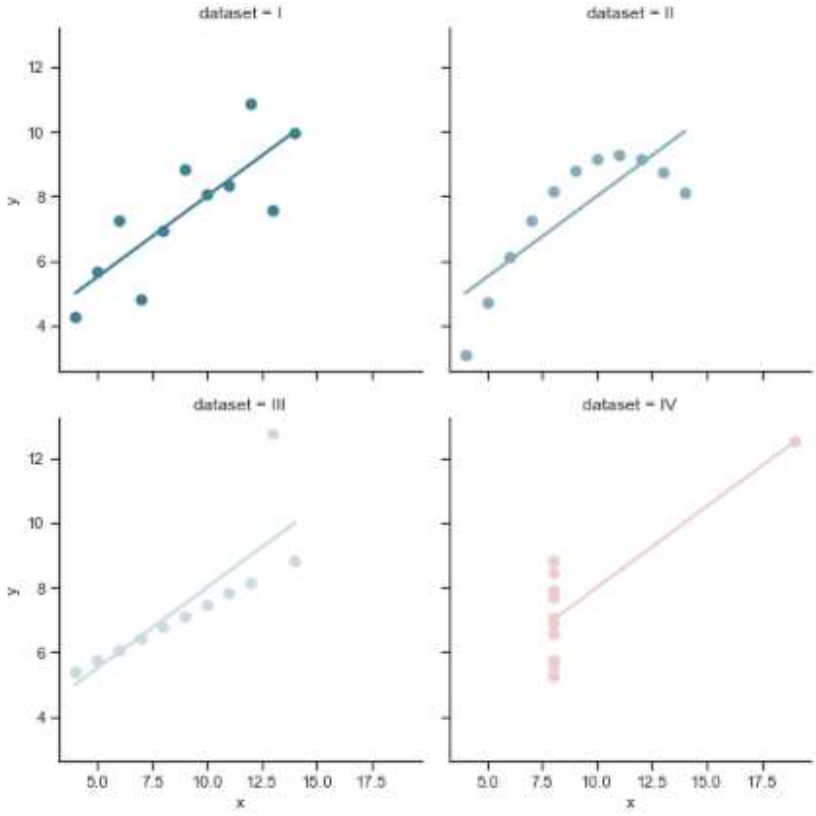
Graphical representation of the same data:

```
In [4]: import seaborn as sns
sns.set_theme(style="ticks")

# Load the example dataset for Anscombe's quartet
df = sns.load_dataset("anscombe")

# Show the results of a linear regression within each dataset
sns.lmplot(
    data=df, x="x", y="y", col="dataset", hue="dataset",
    col_wrap=2, palette=sns.diverging_palette(220, 10), ci=None,
    height=4, scatter_kws={"s": 50, "alpha": 1}
)
```

```
Out[4]: <seaborn.axisgrid.FacetGrid at 0x2cb173812c8>
```



The above figures, demonstrate, even if statistics of the for datasets is similar, they are very different.

	Dataset-I	Dataset-II	Dataset-III	Dataset-IV
Mean x	9.0	9.0	9.0	9.0
Mean y	7.5	7.5	7.5	7.5
Standard variance of x	11	11	11	11
Standard variance of y	4.127	4.127	4.122	4.123
Relationship between x and y	Linear	Non-linear	Linear but outlier	0 slope and one outlier

3. What is Pearson's R? (3 marks)
   Answer:

   Pearson's r is the most commonly used type of correlation coefficient used in linear regression and is used to determine the strength of relationship(linear) between two variables. It is obtained by multiplying the covariance of two variables and diving them with product of their standard deviations.

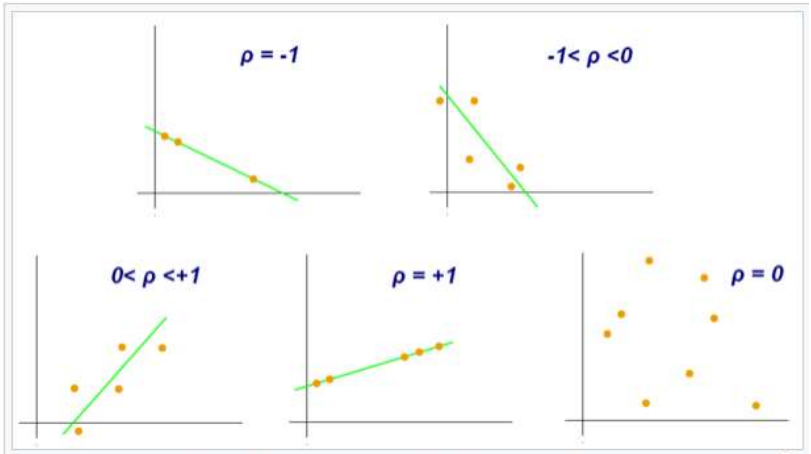
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

   where:

   - n is sample size
   - x_i, y_i are the individual sample points indexed with i
   - $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$

   Reference: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

   The value of the coefficient lies between -1 and 1 with -1 indicating a perfect negative co-relation and 1 indicating a perfect positive correlation and 0 indicating no correlation.



Examples of scatter diagrams with different values of correlation coefficient (ρ)

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
   Answer:

A sample data can compromise of various features having dimensions very different from each other. Scaling is a simple technique used to bring these values to a common scale so that they can be compared. This could be via diminishing large values or magnifying very small values. Two popular methods used for scaling are:

1. Min-Max Scaler:
   It is a simple method of scaling the range of features to scale down to the range between -1 and 1.

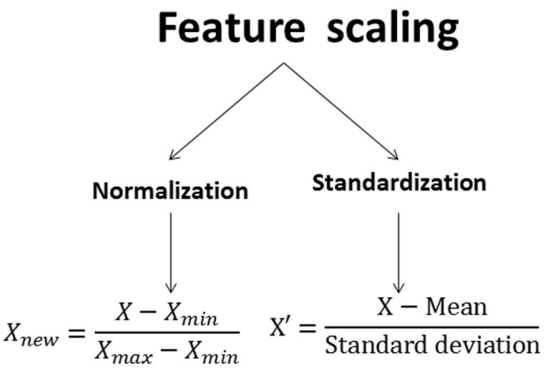
$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

   Where x is the normal value and x` is the normalized value.

2. Standardization:
   Simple method based on mean value of data.

$$z = \frac{x_i - \mu}{\sigma}$$

   Where xi is each data, μ is the mean and σ is the standard deviation.



5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
   Answer:

   VIF is the Variance Inflation Factor is the quantification of how much the variance is inflated in multiple regression due to multicollinearity in an ordinary least square regression analysis.

$$VIF = \frac{1}{1 - R^2}$$

Where R^2 is the coefficient of determination.

If R-square is 1, then the formula will result in infinite VIF (1/ (1-1)). A R-Squared of infinite is possible when there is a perfect correlation between two independent variables. Infinite VIF indicates that the variable can be expressed exactly by a linear combination of other variables. Hence are multicollinear. This can be solved my simply dropping the variable as it will be factored in by other variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
   Answer:
   Q-Q stands for Quantile-Quantile plot and is a graph to assess if two datasets belong to same populations. This can be used in linear regression to assess goodness of the fit.

   Q-Q plot can be very beneficial for the following:
   1. Do the datasets belong to populations with a common distribution?
   2. Do the datasets have a have common location and scale?
   3. The datasets have similar distributional shapes- uniform, exponential, normal?
   4. Datasets have similar tail behavior?

   A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

   Below are the possible interpretations for two data sets.

   a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

   b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.

   c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.

   d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis