Assignment-based Subjective Questions

1:From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

The year box plots indicates that more bikes are rent during 2019.

The season box plots indicates that more bikes are rent during fall season.

The working day and holiday box plots indicate that more bikes are rent during normal working days than on weekends or holidays.

The month box plots indicates that more bikes are rent during September month.

The weekday box plots indicates that more bikes are rent during Saturday.

The weathers it box plots indicates that more bikes are rent during Clear, Few clouds, Partly cloudy weather.

2:Why is it important to use drop first=True during dummy variable creation?

It helps in reducing the extra column created during dummy variable creation.

So that it reduces the correlations created among dummy variables

3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

By looking at the pair plot temp variable has the highest (0.63) correlation with target variable 'cnt'.

4: How did you validate the assumptions of Linear Regression after building the model on the training set?

The assumption in Linear regression can best be tested with scatter plots, below are two examples depict two cases, where either no and little linearity is present.

Secondly, the linear regression analysis requires all variables to be multivariate normal. This assumption can best be checked with a histogram or a Q-Q-Plot

5: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Top 3 features contributing significantly towards the demands of share bikes are:

weathersit_Light_Snow(Negative correlation).

yr_2019(Positive correlation).

temp(Positive correlation).

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression: It is machine learning algorithm that comes below supervised learning.

It is the method to predict the dependent variable (y) based on the given independent variable.

So, regression finds a linear relationship between x (input) and y (output).

Linear Regression are two types

1.Simple Linear Regression

2.Multiple Linear Regression

1.Simple Linear Regression $Y=\beta o+\beta 1\ X +e$

where Y: output or target variable

X: input/dependent variable

βo: Intercept when X =0

β1: Slope of X =change in Y/Change in X

e: error

2. Multiple Linear Regression: it's simple as its name, to elucidate the connection between the target variable and two or more explanatory variables.

Multiple linear regression is used to do any kind of predictive analysis as there is more than one explanatory variable.

$Y=\beta o+\beta 1\ X1 +\beta 2\ X2+.....+\beta n\ Xn +e$

2. Explain the Anscombe's quartet in detail.

This comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. The statistician Francis Anscombe in 1973 had demonstrated both the importance of graphing data before analysing it and the effect of outliers on statistical properties.

Simple understanding:

Those 4 sets of 11 data-points are given below.

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

After that, the council analysed them using only descriptive statistics and found the mean, standard deviation, and correlation between x and y.

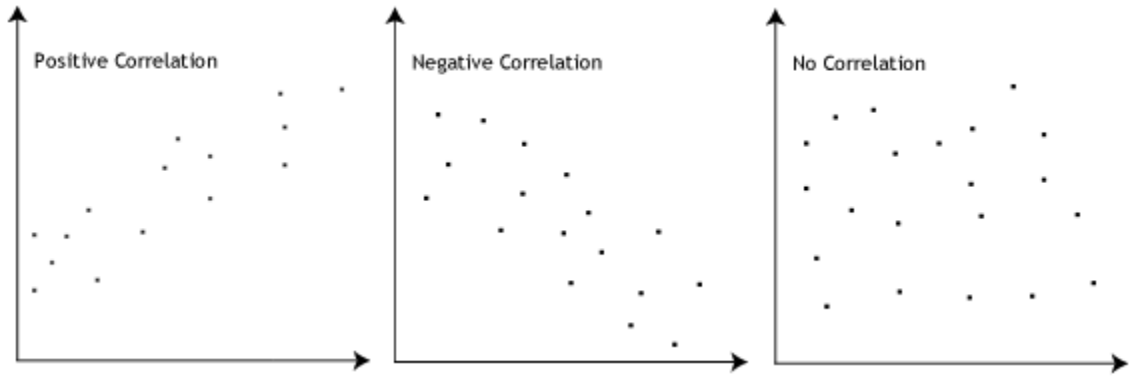
3:

What is Pearson's R?

The Pearson correlation coefficient Known as Pearson's r is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance.It has range of value between −1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope
- r = -1 means the data is perfectly linear with a negative slope
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association



Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Here,

- =correlation coefficient
- =values of the x-variable in a sample
- =mean of the values of the x-variable
- =values of the y-variable in a sample
- =mean of the values of the y-variable

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

This is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

*It is important to note that **scaling just affects the coefficients** and none of the other parameters like **t-statistic, F-statistic, p-values, R-squared**, etc.*

Normalization/Min-Max Scaling:

- *It brings all of the data in the range of 0 and 1. **sklearn.preprocessing.MinMaxScaler** helps to implement normalization in python.*

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

- *Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ**) zero and standard deviation one (**σ**).*

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

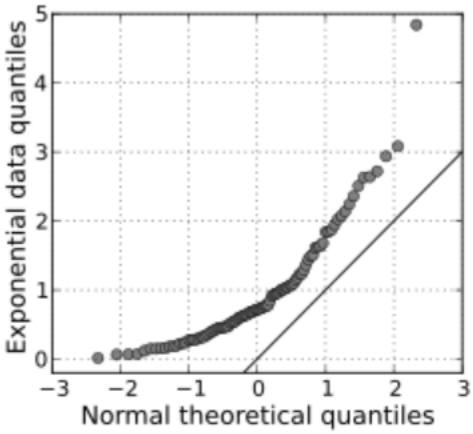
- ***sklearn.preprocessing.scale** helps to implement standardization in python.*
- *One disadvantage of normalization over standardization is that it **loses** some information in the data, especially about **outliers**.*

5.You might have observed that sometimes the value of VIF is infinite. Why does this happen?

- If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
- An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

6.What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.