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? (3 marks)

   ANS: Categorical variables are: season, jan, sept, mon,weat_light, weat_mist in final model. The final equation for cnt is:

   cnt = B0+(-1725.69)*spring+(-114.35)*jan+(484.90)*sept+(436.36)*mon+(-2256.69)*weat_light + (-564.79)*weat_mist.

   Spring, Jan, weat_light, weat_mist is negatively correlated with cnt dependent variable. And Sept, Monday is positively correlated with cnt.

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

   ANS: drop_first=True is important to use because it helps to reduce one extra column which we don't required during dummy variable creation. Hence reduces the correlation between created variables.
   Example, if we have two values in gender column i.e. Female and Male. Female represents 1 then Male would be 0. We don't need to create Male as dummy variable column. If indicator column has 0 then it is obvious Male. We need to use n-1 Dummy variable creation columns with n-levels.

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

   ANS: temp i.e. 0.63

4. How did you validate the assumptions of Linear Regression after building the model on the training set?(3 marks)

   ANS: 1. R2score **2.** p-value of all variables <0.05 **3.** VIF of all variables < 5 **4.** Prob(F-stats)

   As much as r2 will be greater than model will be good as in our case its near to 80% which represent good model. The p- value of all variables should be less than <0.05. VIF i.e. variance inflation factor of all the variables should be less than 5. Because if its greater that 5 than shows multicolinearity between independent variables. Probability of F stats should be near to 0.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?(2 marks)

   ANS: temp, yr, spring. Temperature has very high correlation with cnt variable i.e. 0.63. Similarly yr(Year) has high correlation i.e. 0.57 and spring (season) has 0.56.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

   ANS: Linear Regression: Linear regression is one of the very basic forms of machine learning where we train a model to predict the behavior of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

   Mathematically, Linear regression :

   y = mx + c where m is slope and c is intercept.

   Step1: Load the important libraries i.e. pandas, matplotlib and Statsmodel to show statistic of data.

   Step2: Load the data and check the shape and info of the data

   Step3: Visualize the relationship between dependent variable (Y-axis) and independent variables (X-axis) with scatter plot or heatmap for numeric variables and box plot for categorical variables.

   Step4: Break the data into Train and Test model. And create model with OLS method i.e. least square method of statsmodel library to check the coefficient of the models.

   Step5: We will get the value of m and c from Step4 which we will use to predict the values of Test model.

   Step6: Plot the least square line in between store values and predict variables. There relation should be linear.

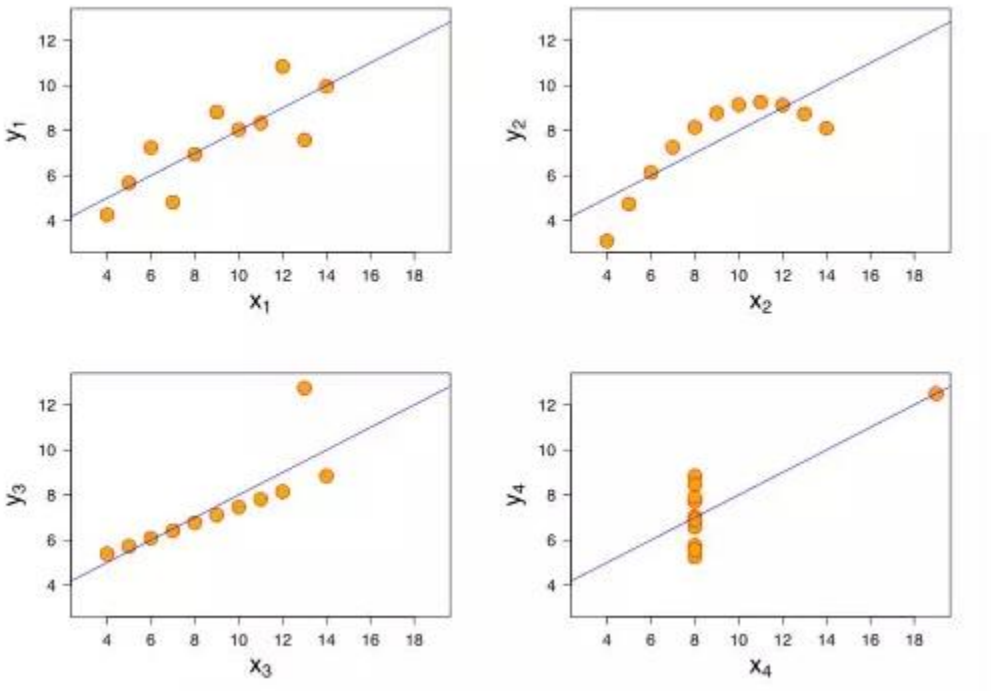
   Step7: Check the P-values, VIF and r2 score of final model.

2. Explain the Anscombe's quartet in detail. (3 marks)

   ANS: Anscombe's quartet comprises four datasets that have nearly identical simple Statistical properties yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties. Those 4 sets of 11-points given below:

Anscombe's quartet

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



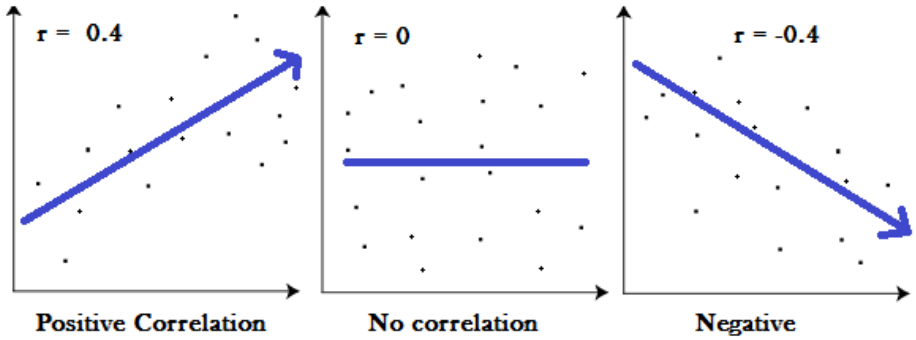
Explanation of this output:

- In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

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

ANS: Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, where:

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.



Pearson's correlation coefficient formula:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,]\,[\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

Example: Scientists in India wanted to know if there was a relationship between how weedy rice populations are different genetically. The goal was to find out the evolutionary potential of the rice. Pearson's correlation between the two groups was analyzed. It showed a positive Pearson Product Moment correlation of between 0.783 and 0.895 for weedy rice populations. This figure is quite high, which suggested a fairly strong relationship.

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

ANS: **Scaling:** It 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.

Need of scaling: 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.

Normalized Scaling: It brings all the data in range of 0 to 1.

Sklearn.preprocessing.MinMaxScaler is used in python to implement mimamx scaler.

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

Disadvantage: Loss of information especially outliers.

- **Standardized 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**).

  Sklearn.preprocessing. StandardScaler is used in python to implement standardized scaler.

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

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

   (3 marks)

   ANS: This is because of 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 (which show an infinite VIF as well). After that check VIF again to get the VIF of all variables. We should not drop multiple columns once. Drop columns one by one.

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

   (3 marks)

ANS: **Quantile-Quantile (Q-Q)** plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.
This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Advantages:

a) It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:

If two data sets —

- come from populations with a common distribution

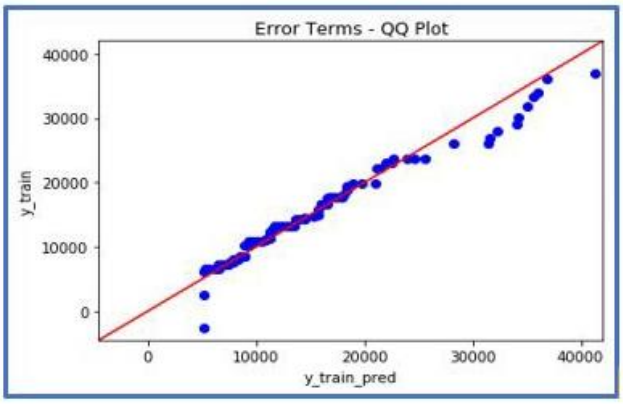
- have common location and scale
- have similar distributional shapes
- have similar tail behavior

Interpretation:

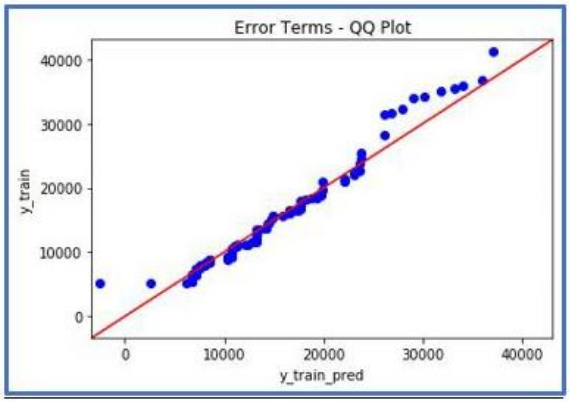
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.Below are the possible interpretations for two data sets.

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

Python Library: statsmodels.api provide qqplot and qqplot_2samples