Assignment 02

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)

   The categorical variables in the dataset that helps us decide about the demand count of bikes are - season, mnth, weekday, weathersit, year, holiday, workingday

   All these categorical variable are initially studied as boxplot in the python notebook and the results convey the effect of these variables on the dependent variable

   - **season** - We see that spring season sees the least demand count, and Fall accounts to most sales (highest 75% quantile). Fall and Summer almost go neck to neck in the demand created on bikes
   - **mnth** - Similar to the season, the mnth where the highest demand is seen is mostly in June - November. We already are familiar with this, because we know that the highest count happened in the seasons Summer and Fall. Hence, month provides similar effect on the count as seasons
   - **weekday** - As long as the effect of weekday on the dependent variable goes, all seems to have similar medians : which makes us come to an inference that people create demand on working days as well as weekend almost equally. This is also a similar case in working day variable
   - **weathersit** - Weathersit is a good affecting factor to the count. It seems likely that the light rain under weather situation is correlated to the demand count. People are tending to go less on bike during light rain condition
   - **yr** - As far as the year goes, 2019 gives a overall demand created picture to the business, and is better than the demand count in 2018

- **holiday** - There is not much of a difference in the level of demand here. Except that all of the datapoints are located a little higher for the non-holiday day, and for holiday, the datapoints are a bit spread, and the median is lower than that of a non- holiday

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

   In the process of creating dummy variables, if the number of levels is n, then n-1 dummy variables is sufficient for representing them.

   For example, consider the number of levels is 3

   And, the variables are A, B and C

   While making the dummy values for all the 3 variables, we get

   A = 1 0 0 (dummification can be thought of as a sensor signalling whenever the variable is A, and is silent for the other two, hence we get this)

   B = 0 1 0 (Similarly, firing for B)

   C = 0 0 1 (for C)

   Now, we have use 3 variables for the dummy variable creation

   But, there's a possibility of using 2 variables

   Converting all the values to 0 0 can also be thought of as A, and the other two variable value remains the same

   A = 0 0 , **B** = 1 0 , **C** = 0 1

   The whole scenario of dummy creation is now created in two variable. It is computationally simpler on the algorithm being used and is logically good enough.

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

   Looking at the pair-plot among the numerical variables, it seems like temp and atemp have the highest correlation with the target variable. Since, atemp and temp already have higher correlation with each other, temp is dropped in my model.

   So, **atemp** has the highest correlation with the target variable.

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

   While doing linear regression and assessing the implications on the target variable, we need to assume the following things -

   - The relationship between X and Y has to be linear
   - Error terms are normally distributed with a mean = 0
   - Error terms are independent of each other
   - Error terms have constant variance (homoscedasticity):

   The steps taken are -

   1. It is difficult to check whether there is a **linear relationship** between X and y visually since it involves as much dimension as the number of independent variables. However, getting a good R^2 value or adjusted, tells us that this is a linear model.
   2. To check if the error terms are **normally distributed** (which is infact, one of the major assumptions of linear regression), the histogram plot of the error terms (y_train - y_train_pred) was done, and the mean was found to be at 0
   3. A regplot was done of the error vs the y_train giving us how the error looks. We have to visualize that the error doesn't follow a timeseries, or the previous error here and are random. This tells us that the error terms are **independent** of each other
   4. A regplot was done for the error vs the y_train_pred, gives us a plot of error which seems to have equal variance throughout the graph existance. This proves us that the error terms have constant variance or **homoscedasticity**.

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

   Based on the final model and the coefficient of the variables given by building the model, the top 3 features actually contributing significantly and explaining the demand for the shared bikes are -

   - **atemp** (the felt temperature)
   - **hum** (The humidity of the region)

- **yr** (Either 2018 or 2019) [General demand increase in bikes]

Next feature which is equally relevent as the other features

- Light Rain (Weather situation variable - Special mention)

General Subjective Questions

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

   Linear Regression is a linear (straight line) representation of a relationship between the independent and the dependent variable.

   The representation is a linear equation that combines a specific set of input values (denoted by x) the solution to which is the predicted output for the same set of input values (denoted by y). The values of the output variable (y) is numeric but the set of values for input (x) can be both numeric as well as categorical.

   The linear equation assigns a scale factor to each input value or column, called a coefficient and represented by the capital Greek letter Beta (B). One additional coefficient is also added, giving the line an additional degree of freedom and is often called the intercept or the bias coefficient.

   For example, in a simple regression problem (a single x and a single y), the form of the model would be:

   y = B0 + B1*x

   The reason why intercept is required is for example, if price (y) is depedent upon the demand (x), this scenario doesn't suggest that the price of a material is going to be 0 if the demand is null. So, intercept here acts like a base price, which remains there even when there is no demand.

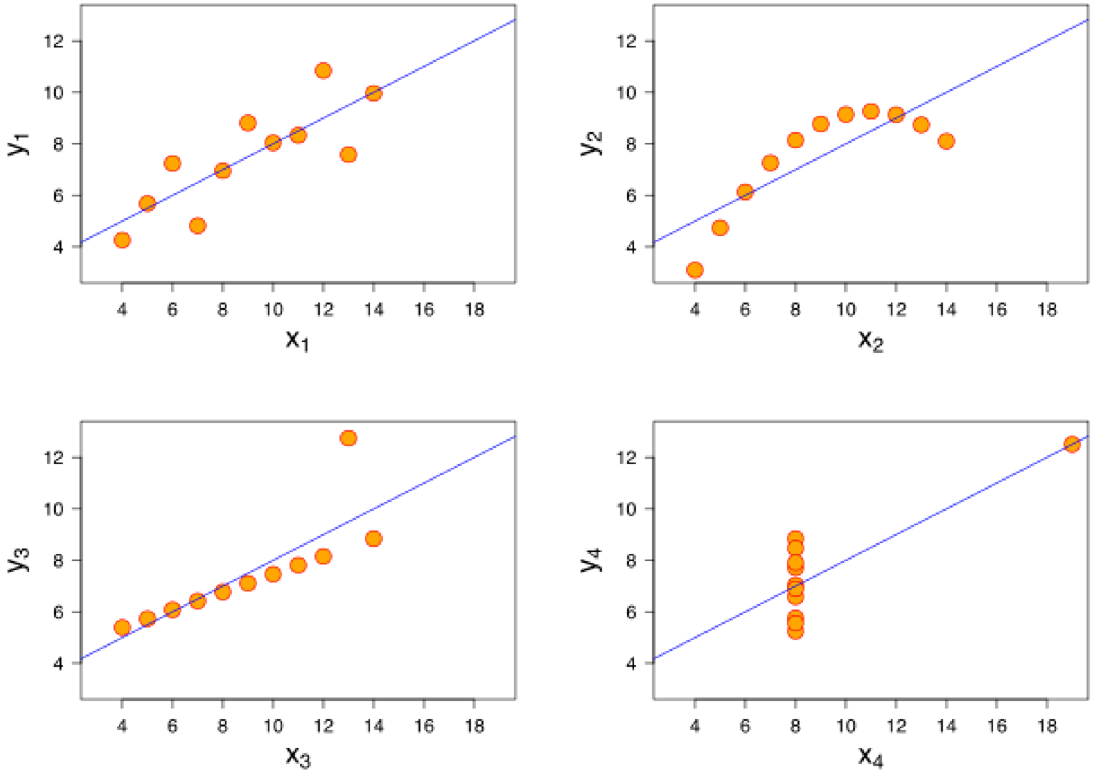
   In higher dimensions (which is referred to as multiple linear regression, in contrast to the previously discussed simple linear regression) when we have more than one input (x), the line is called a plane or a hyper-plane. The number of coefficients thus increases with the number of input values, thus denoting its significance to the target variable (y)

   When a coefficient becomes zero, it effectively removes the influence of the input variable on the model and therefore from the prediction made from the model (0 * x = 0). This becomes relevant if you look at regularization methods that change the learning algorithm to reduce the complexity of regression models by putting pressure on the absolute size of the coefficients, driving some to zero.

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

Anscombe's Quartet was developed by statistician named Francis Anscombe. It comprises four datasets, each containing eleven (x,y) pairs. The important thing to note here about these datasets is that they have a similar descriptive statistics. But when they are visualized, each graph tells a new visual story irrespective of their identical summary statistics.

Visualization may not be as precise as statistics, but it provides a unique view onto data that can make it much easier to discover interesting structures than numerical methods. Visualization also provides the context necessary to make better choices and to be more careful when fitting models or in the pre-processing stage. Anscombe's Quartet is a case in point, showing that four datasets that have identical statistical properties can indeed be very different.



All four of these data sets have the same variance in x, variance in y, mean of x, mean of y, and linear regression. But, as you can clearly tell, they are all quite different from one
another. Anscombe's Quartet shows that multiple data sets with many similar statistical properties can still be vastly different from one another when seen through visual plots.

Additionally, Anscombe's Quartet warns of the dangers of outliers in data sets. Think about it: if the bottom two graphs didn't have that one point that strayed so far from all the other points, their statistical properties would no longer be identical to the two top graphs. In fact, their statistical properties would more accurately resemble the lines that the graphs seem to depict.

It is obvious that the top right graph really shouldn't be analyzed with a linear regression at all because it's a curvature. Conversely, the top left graph probably must be analyzed using a linear regression because it's a scatter plot that moves in a linearly manner. These observations demonstrate the value in graphing your data before analyzing it.

Anscombe's Quartet is a important demonstration which shows us that **visualization** is a pre-processing step which must be included in the process before model building to throw in some accurate results from the model.

3. What is Pearson’s R? (3 marks)

   Pearson correlation coefficient is a measure of the strength of a linear association between two variables and is denoted by R.

   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.

   The meaning of these values of correlation is -

   - A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.

- A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decreases in (almost) perfect correlation with speed.
- Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.

Formula for Pearson's R is -

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

The assumptions that Pearson's R coefficient -

- For the Pearson's R correlation, both variables should be **normally distributed**.
- There should be no significant **outliers**. We all know what outliers are but we don't know the effect of outliers on Pearson's correlation coefficient, R. This coefficient is very sensitive to outliers, can have a very large effect on the line of best fit. This also means including outliers in your analysis dataset, can lead to misleading results.
- Each variable should be **continuous** i.e. interval or ratios for example time, height, age etc. must be equal.
- Scatter plots will help you tell whether the variables have a linear relationship. If the data points have a straight line, then the data satisfies the linearity assumption. If the data you have is not linearly related you might have to run a non-parametric .
- The observations are **paired observations**. That is, for every observation of the independent variable, there must be a corresponding observation of the

dependent variable

- **Homoscedascity** - Homoscedascity simply refers to ‘equal variances’ of errors as well as dependent variable. A scatter-plot makes it easy to check for this. If the points lie equally on both sides of the line of best fit, then the data is homoscedastic. As a bonus — the opposite of homoscedascity is heteroscedascity which refers to refers to the circumstance in which the variability of a variable is unequal across the range of values of a second variable that predicts it.

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

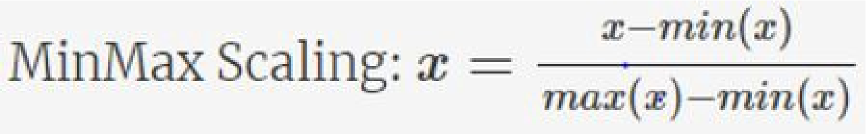
Scaling is a step of **Data Pre-Processing** which is applied to independent variables to normalize or standardize the data within a particular range. It also helps in speeding up the calculations in an algorithm, and helps in creating a common range for all the variables present

Many a times, the data set contains features strongly varying in magnitudes, units or range. If scaling is not done in the pre-processing step, then algorithm only takes magnitude into consideration and not units hence leading to incorrect modelling. Also, the beta (significance of a variable) is uneven and the highest magnitude variable, even though most significant might get lesser beta.

It is important to note that **scaling just affects the coefficients** and none of the other parameters involved in the modelling.

Normalization/Min-Max Scaling:

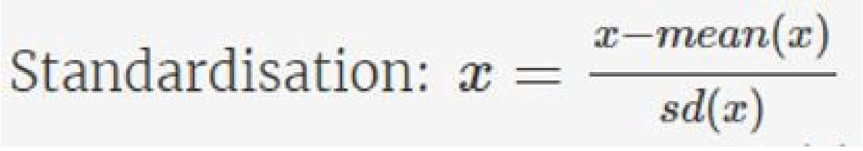
- It brings all of the data in the range of 0 and 1. **sklearn.preprocessing.MinMaxScaler** helps to implement normalization in python.
- Normalization can also be used to scale categorical variables as their values are always 0 and 1. Normalization can be safely used on all the variables overall.



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

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?
(3 marks)

VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables.

In VIF, each feature is regression against all other features. If R^2 is more which means this feature is correlated with other features. [0]

- VIF = 1 / (1 – R^2)
- When R2 reaches 1, VIF reaches infinity

VIF infinite thus means **perfect correlation** that exists between two independent variable. There might be a possibility of 2 variables being extremely similar in their way of affecting the target variable. For example, one can be a student's CGPA and one variable can be his percentage. This means that these two independent variable essentially tells the same thing, i.e the student's performace, and this is highly collinear and has an infinite VIF. Thus, understanding VIF and how collinear two variables are, hold a very good significance with respect to the model building and its accuracy.

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

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us understand if a set of data
came from some theoretical distribution such as a Normal,
exponential or Uniform distribution. Also, it helps to determine if two
data sets originated from the populations with a common distribution.

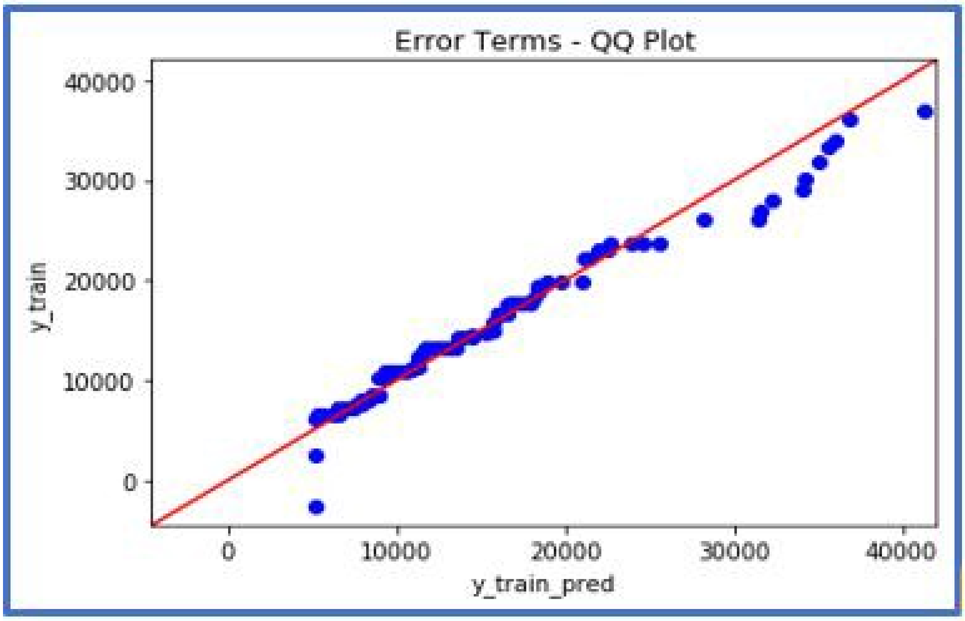
This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot
that both the data sets are from populations with same distributions.

Interpretation:

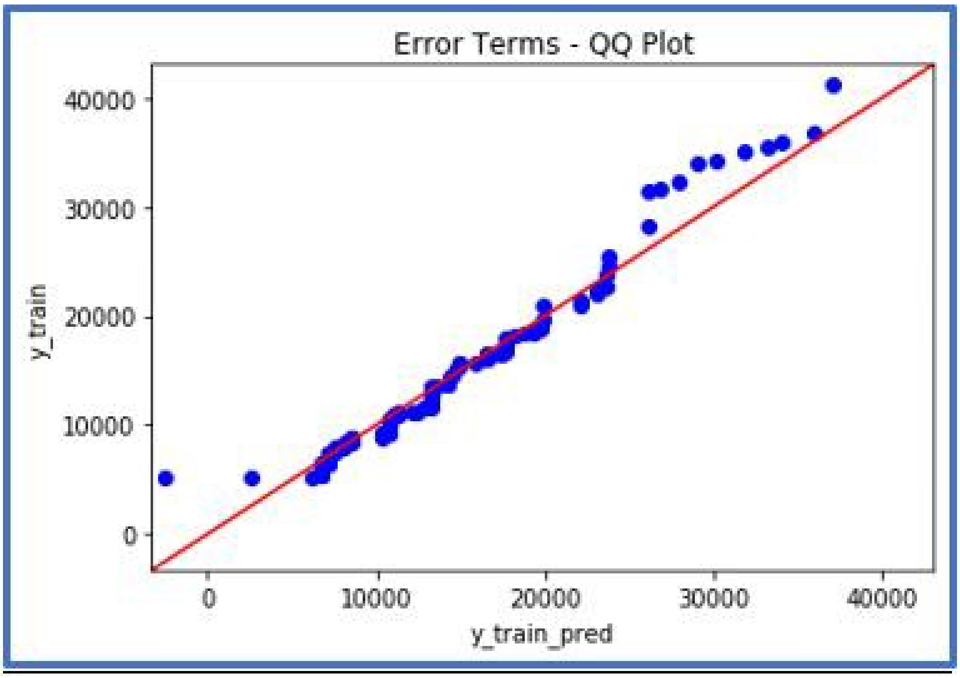
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

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) **X-values > Y-values:** If x-quantiles are higher than the y-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