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 demand of bike is less in the month of spring when compared with other seasons
- The demand bike increased in the year 2019 when compared with year 2018.
- Month Jun to Sep is the period when bike demand is high. The Month Jan is the lowest demand month.
- Bike demand is less in holidays in comparison to not being holiday.
- The demand of bike is almost similar throughout the weekdays.
- There is no significant change in bike demand with working day and non working day.
- The bike demand is high when weather is clear and Few clouds however demand is less in case of Light snow and light rainfall. We do not have any data for Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog , so we can not derive any conclusion. May be the company is not operating on those days or there is no demand of bike.

2. Why is it important to use drop_first=True during dummy variable creation?
Ans- drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence 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?
Ans- 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?

Ans-
i) One of the most important assumptions is that a linear relationship is said to exist between the dependent and the independent variables. If you try to fit a linear relationship in a non-linear data set, the proposed algorithm won't capture the trend as a linear graph, resulting in an inefficient model.
The simple way to determine if this assumption is met or not is by creating a scatter plot x vs y. If the data points fall on a straight line in the graph, there is a linear relationship between the dependent and the independent variables, and the assumption holds.
ii) No auto-correlation or independence: The residuals (error terms) are independent of each other. In other words, there is no correlation between the consecutive error terms of the time series data. The presence of correlation in the error terms drastically reduces the accuracy of the model. If the error terms are correlated, the estimated standard error tries to deflate the true standard error.
Conduct a Durbin-Watson (DW) statistic test. The values should fall between 0-4. If DW=2, no auto-correlation; if DW lies between 0 and 2, it means that there exists a positive correlation. If DW lies between 2 and 4, it means there is a negative correlation. Another method is to plot a graph against residuals vs time and see patterns in residual values.

iii) No Multicollinearity: The independent variables shouldn't be correlated. If multicollinearity exists between the independent variables, it is challenging to predict the outcome of the model. In other words, it is unclear which independent variables explain the dependent variable.
Use a scatter plot to visualize the correlation between the variables. Another way is to determine the VIF (Variance Inflation Factor). VIF<=4 implies no multicollinearity, whereas VIF>=10 implies serious multicollinearity.

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

Ans- The 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.
Ans- 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.

Example for that can be let's say you are running a sales promotion and expecting a certain number of count of customers to be increased now what you can do is you can look the previous promotions and plot if over on the chart when you run it and then try to see whether there is an increment into the number of customers whenever you rate the promotions and with the help of the previous historical data you try to figure it out or you try to estimate what will be the count or what will be the estimated count for my current promotion this will give you an idea to do the planning in a much better way about how many numbers of stalls maybe you need or how many increase number of employees you need to serve the customer. Here the idea is to estimate the future value based on the historical data by learning the behaviour or patterns from the historical data.

In some cases, the value will be linearly upward that means whenever X is increasing Y is also increasing or vice versa that means they have a correlation or there will be a linear downward relationship.

One example for that could be that the police department is running a campaign to reduce the number of robberies, in this case, the graph will be linearly downward.

Linear regression is used to predict a quantitative response Y from the predictor variable X.

Mathematically, we can write a linear regression equation as:

what-is-linear-regression-2
Where a and b given by the formulas:

what-is-linear-regression-1
Here, x and y are two variables on the regression line.

b = Slope of the line.

a = y-intercept of the line.

x = Independent variable from dataset

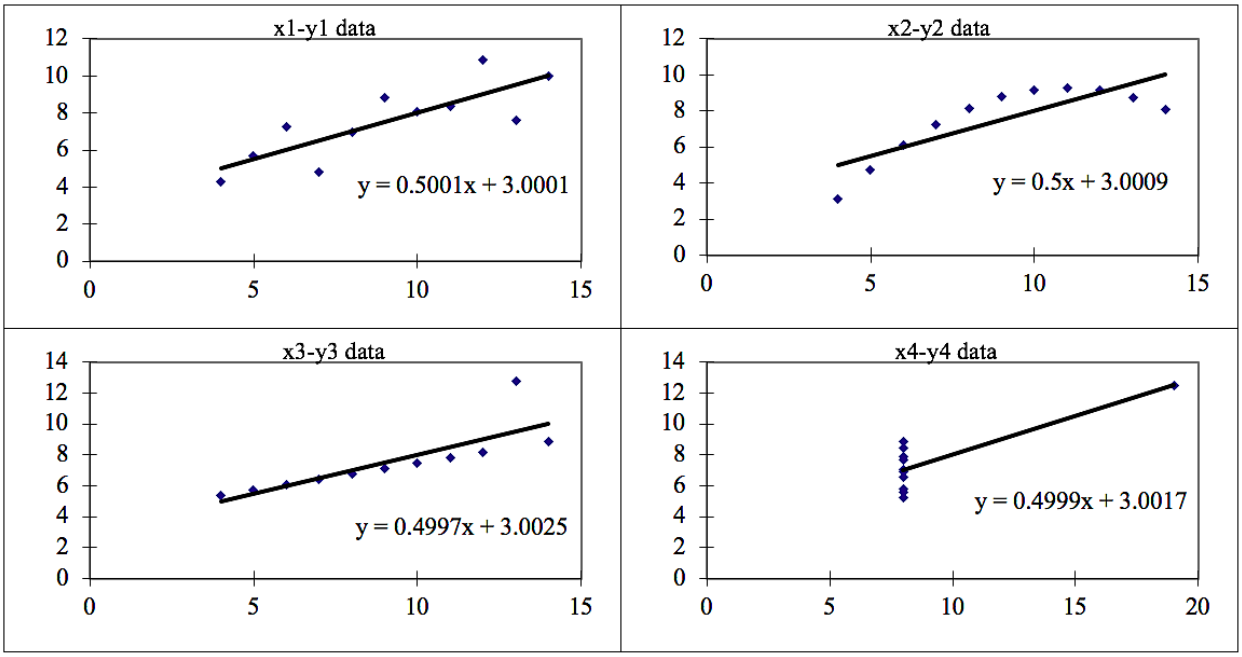
y = Dependent variable from dataset

Use Cases of Linear Regression:

Prediction of trends and Sales targets – To predict how the industry is performing or how many sales targets the industry may achieve in the future.
Price Prediction – Using regression to predict the change in price of stock or product.
Risk Management- Using regression to the analysis of Risk Management in the financial and insurance sector.

2. Explain the Anscombe's quartet in detail.
Ans- Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.
but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.
This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets. These four plots can be defined as follows:



The four datasets can be described as:
Dataset 1: this fits the linear regression model pretty well.
Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.
Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model

Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

We have described the four datasets that were intentionally created to describe the importance of data visualization and how any regression algorithm can be fooled by the same. Hence, all the important features in the dataset must be visualized before implementing any machine learning algorithm on them which will help to make a good fit model.

3. What is Pearson's R?
Ans- The Pearson coefficient correlation has a high statistical significance. It looks at the relationship between two variables. It seeks to draw a line through the data of two variables to show their relationship. The relationship of the variables is measured with the help of the Pearson correlation coefficient calculator. This linear relationship can be positive or negative.
For example:
i) Positive linear relationship: In most cases, universally, the income of a person increases as his/her age increases.
ii) Negative linear relationship: If the vehicle increases its speed, the time taken to travel decreases, and vice versa.

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

Where:

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

Σx2 = the sum of squared x scores

Σy2 = the sum of squared y scores

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

Ans- Scaling 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.

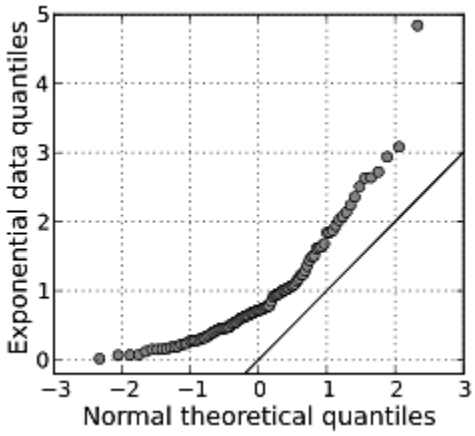
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Ans-
If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. 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 (which show an infinite VIF as well).

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

Ans- 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.

A Q Q plot showing the 45 degree 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.