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)**

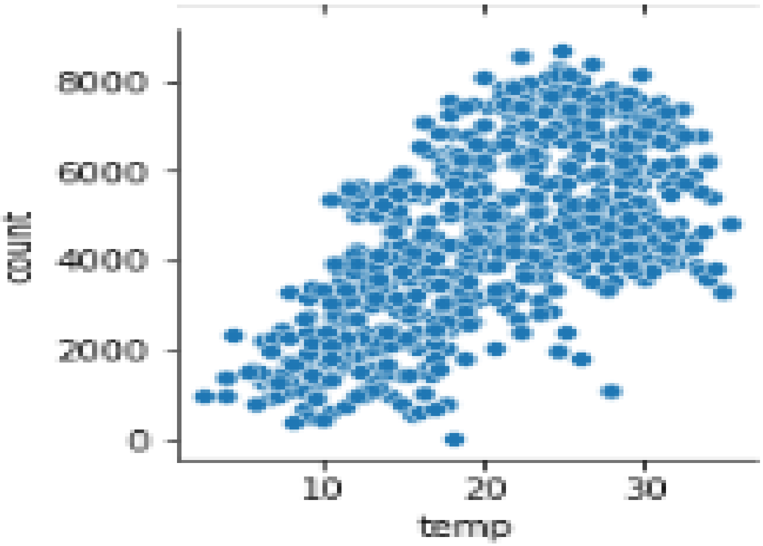
   Before developing the model, the categorial variables when plotted in boxplots the below effect is observed on the dependent variable:

- Bike Rental counts are more during the Fall season and then in summer and very less during the spring.
- Bike Rental counts are more in the year 2019 compared to 2018.
- Bike Rental counts are more in clear or partly cloudy weather situations and very less during snow and rainy conditions.
- Bike Rental counts are more on Saturday, Wednesday followed by Thursday, Friday.

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

   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? (1 mark)**
   Temperature seems to have the highest correlation with the target variables.



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

- Normality of Residuals – This assumption of Linear Regression is that the residuals follow a normally distribution. Once the residuals is obtained from our model we test it either using a histogram or a QQ plot.
- No or little multicollinearity -This assumption of linear regression is that there is no or little multicollinearity between independent variables. The test for multicollinearity is done using VIF.
- All Independent variables are uncorrelated with the error terms - This serves to check whether there is a correlation between any of the independent variables and the error terms. If this happens, it is likely that we have a case of a mis specified model.
- Observations of the error term are uncorrelated with each other

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

count = 0.2333 X Year + 0.0568 X workingday + 0.4918 X temp - 0.0647 X spring + 0.0516 X summer + 0.0984 X winter - 0.3051 X Light Snow - 0.08 X (Mist + Cloudy) + 0.0647 X Sat + 0.0915 X Sep
Temp, Light Snow, winter

General Subjective Questions

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

Linear regression is one of the forms of machine learning where we train a model to predict the behaviour of our data based on some variables. In the case of linear regression as the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated. Mathematically, we can write a linear regression equation as:

$$y = a + bx$$

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

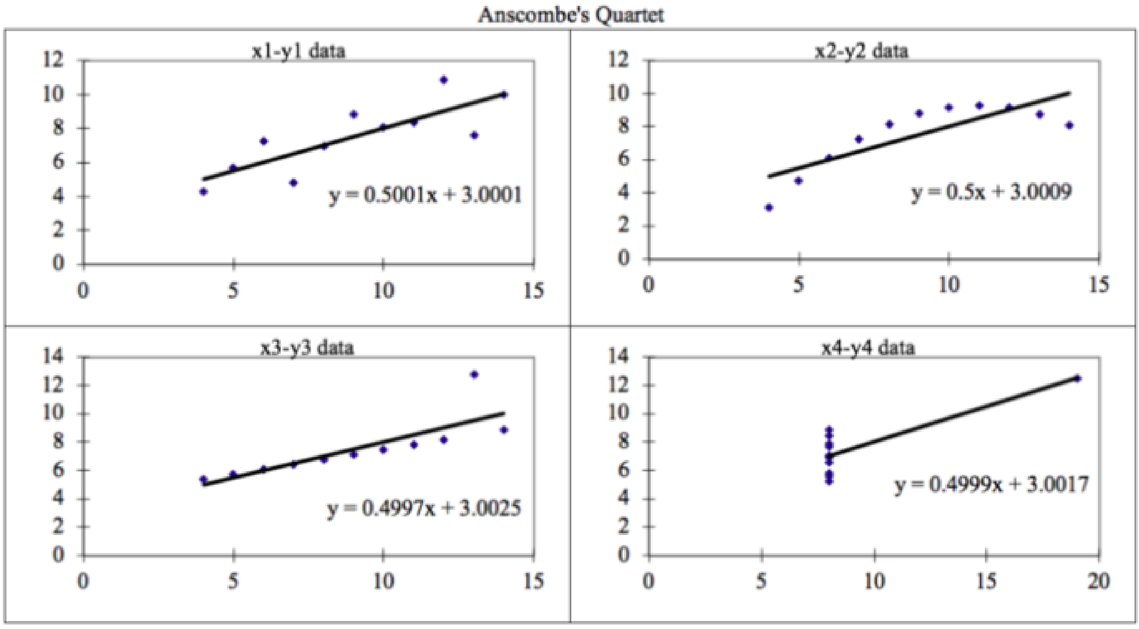
In simple words, linear regression means fitting the best fit line between independent and target variables with the least mean square error.

Before implementing linear regression, we should check whether the data is following these assumptions:

- Data should be linear
- No Multicollinearity
- No auto-correlation
- Homoskedasticity should be there

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

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, 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 highlights the importance of plotting data to confirm the validity of the model fit.



3. What is Pearson’s R? (3 marks)

In statistics, the Pearson correlation coefficient also known as Pearson's r, is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation.

Pearson's correlation coefficient cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

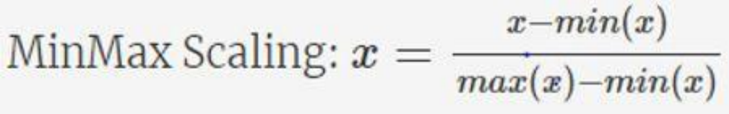
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling is a step of data Pre-Processing which is applied to independent variables to normalize 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? (3 marks)

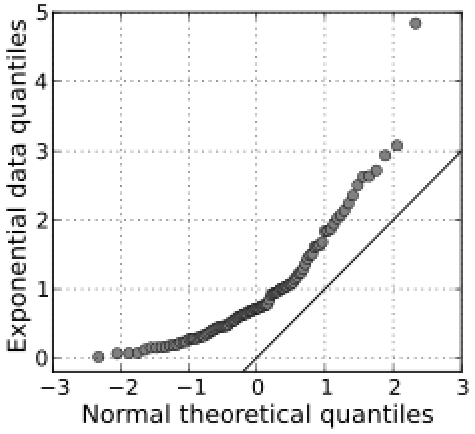
If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect

correlation, we get R^2 =1, which lead to 1/(1-R^2) 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. (3 marks)**

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.