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)**
   - **season**, **weathersit**, **holiday**, **mnth**, **yr**, **weekday** and **workingay** were the variables in the dataset that were categorical and boxplots were plotted to visualize and find inferences from these variables. The inferences about the effect of these variables on the dependent variable (**cnt**) was as follows:-
     - **Season -** The boxplot showed that Spring is the season that had the lowest count for bikes (**cnt**) shared amongst all seasons. Fall had the highest value of **cnt,** then followed by Summer and then winter
     - **Weathersit** - There were no bikes shared when the weather situation was like 'Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog'. When the weathersit was 'Clear, Partly Cloudy', the highest number of shared bikes were used.
     - **Holiday -** The bike sharing count for Holidays were lower as compared to non-Holidays
     - **Mnth –** The Summer months showed a rise in the number of bike rentals. January and February were the only months with lower number or bike rentals.
     - **Yr** - The bike sharing count increased significantly in 2019 as compared to 2018
     - **Weekday** - The median value of count of shared bikes for all the days in the week were almost similar
     - **Workingday** -The maximum and minimum number of Bike rentals on Non-working days are slightly higher than Working days, but their median value was almost similar.

2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**
   - **drop_first=True** is important to use during dummy variable creation because it helps in removing the extra variables that get created during making dummy variables. It also helps in reducing correlation between dummy variables.
     - If we have 3 values in a variable, namely value_1, value_2, value_3 and we apply function to create dummy variables on this variable, the function will by default create 3 columns for these 3 values. But we do not need all these 3 columns for prediction of each of these 3 values. If a value is not value_1 and not value_2, it will then be value_3.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**
   - It is seen that `atemp` has the highest correlation with the target variable (`cnt`)

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**
   - After building the Linear Regression model, the model was validated as follows:
     - Checking the distribution of error terms (Residuals) – The error terms were normally distributed and centred around the mean 0.
     - Checking for Multicollinearity : The VIF scores for all the variables in the final model were below 5. Also, a correlation heatmap was plotted no high correlation among independent variables were seen.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**
   - The Top 3 features that are contributing significantly towards explaining the demand of the shared bikes are :
     - temp : 0.4785 (coefficient)
     - yr : 0.2347 (coefficient)
     - season_winter : 0.0960 (coefficient)

General Subjective Questions

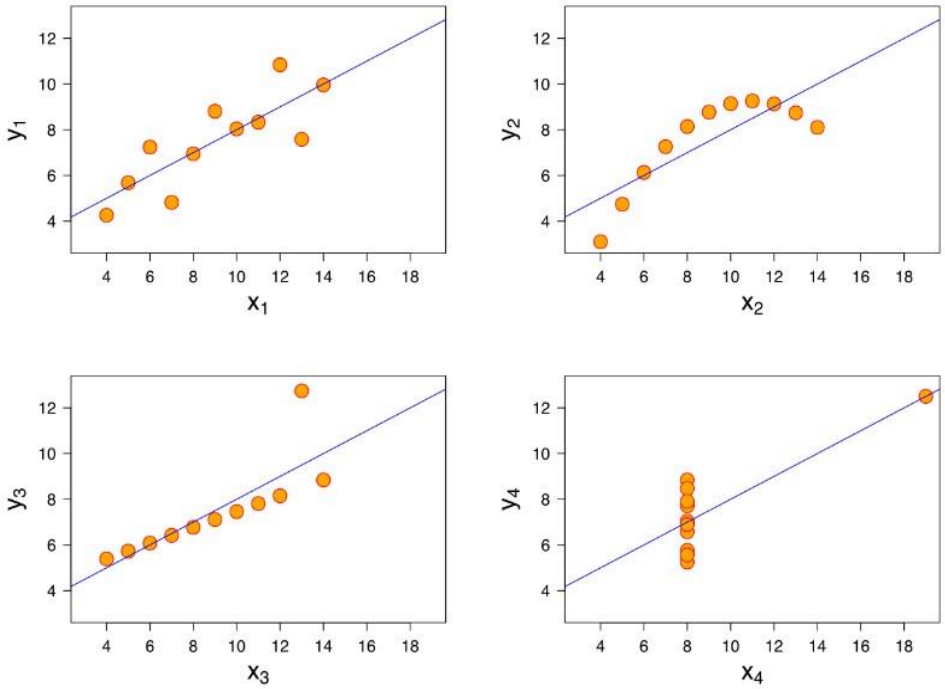
1. **Explain the linear regression algorithm in detail. (4 marks)**
   - Linear Regression can be defined as an algorithm that describes the linear relationship of a dependent feature with 1 or more independent features. The dependent feature is continuous in nature because Linear Regression is used to predict continuous value. It explains the effect of the independent features on the dependent feature. This basically means that it explains how will the dependant feature value change given per unit change in an independent feature. The Linear Regression model fits line to the dataset, which follows a linear equation and that line is used for predicting values. The line is called Best Fit Line and it is fitted by minimising the sum of squares between the line and the actual data values. The equation of Linear Regression is as follows:
     - $y = mx + c$
       - where, y = output or the dependent feature, m = slope, x = independent feature, c= intercept

   - Also, there are 2 types of Linear Regression models:
     1. **Simple Linear Regression :** Simple Linear Regression means that there is 1 dependent variable and 1 independent variable

2. **Multiple Linear Regression :** Multiple Linear Regression means that there are more than 1 dependent variables and 1 independent variable

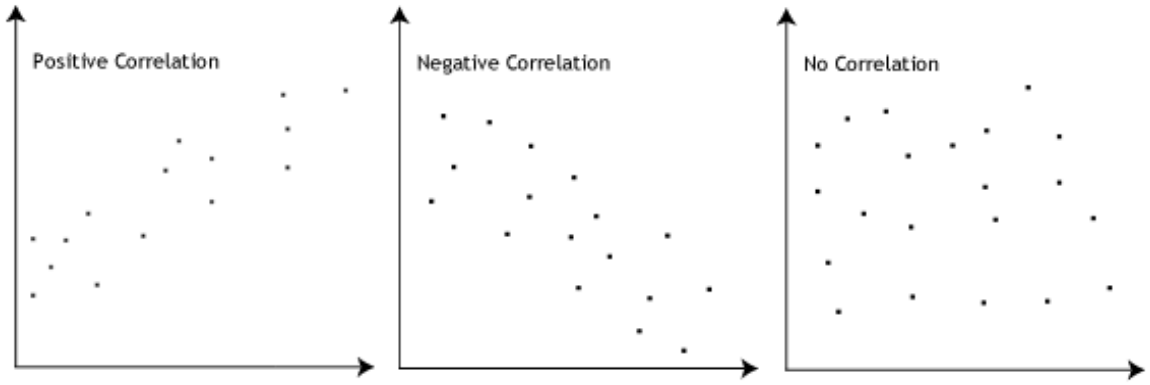
- Assumptions of Linear Regression:
  - **Linear Regression Model fits a hyperplane to the data values**
  - **There is a linear relationship between X and Y**
  - **Error terms are normally distributed with mean zero (not X, Y)**
  - **Error terms are independent of each other and there should be very little or no correlation between independent features**
  - **Error terms have constant variance (homoscedasticity)**

2. **Explain the Anscombe's quartet in detail. (3 marks)**
   - Anscombe's Quartet was developed by a statistician named Francis Anscombe. It has four datasets. Each of these four datasets consists of 11 points. These datasets have a simple similar descriptive statistics. But, the graphs look very different when they are plotted and compared with each other.
   - The statistics for all the four datasets are similar. The statistics are as follows:
     - Mean of x is 9
     - Variance of x is 11
     - Mean of y is 7.50
     - Variance of y is 4.13
     - The correlation between x and y is 0.816
     - The equation for the best fit line is : $y = 3 + 0.5x$
     - The R-squared value is 0.67
   - We plot scatterplots from these dataset and then we notice that the best fit line is the same for all of these 4 datasets but the datasets look completely different after plotting.



- The first scatter plot indicates a simple linear relationship between x and y
- The second scatter plot does not follow a normal distribution and the relationship between x and y is not linear
- In the third scatter plot, we can see that there is a linear relationship but the regression line is affected by an outlier and the line should be different
- The fourth scatter plot shows that a high correlation coefficient can be created by the effect of an outlier

3. **What is Pearson's R? (3 marks)**
   - Pearson's R is the measurement of the linear association between the variables. It shows the strength of the linear relationship between two variables. Pearson's R value lies between -1 and 1.
   - If the correlation coefficient is 1, it indicates a strong positive linear relationship between the variables. This means that values from one variable increase with increase in the values of the other variable
   - If the correlation coefficient is -1, it indicates a strong negative linear relationship between the variables. This means that values from one variable decrease with increase in the values of the other variable
   - If the correlation coefficient is 0, it indicates that there is no linear relationship between the variables



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

- Scaling is a technique that is being used to normalize the variables in the of dataset. Scaling compresses the values of the all the variables of the dataset in to a specific range. If features are not scale, then the machine learning algorithm gibes more importance or more weight to features with larger values and lower importance or lower weight to features with low values. This confuses the model and makes the model predict wrongly.
- The common types of scaling techniques are :
    - **Normalization**: It uses minimum and maximum values for scaling the values of a variable and is known a Min-Max scaling. It is used when features are of different scales. It outputs values that ranges between 0 and 1.
    - **Standardization**: In this technique, the values are centred around the mean with a standard deviation of 1. The mean of the variable becomes 0 and the standard deviation becomes 1. It outputs values between -1 and 1 with values normally distributed around mean 0.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
(3 marks)

- VIF is infinity when there is a perfect correlation.
- VIF is infinite means that the variable can be represented by a linear combination of other variables, which can lead to multicollinearity
- Higher the VIF value, higher there is a chance of being a perfect correlation between the variables.
- In the case of perfect correlation, we get R-squared value =1, which lead to 1/(1-R-squared) infinity.

- To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
- If the variable has a VIF score of 2, this would mean that the variance of the model coefficient is affected by a factor of 2 because of multicollinearity.

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 are plots in which two quantiles are plotted against each other.
- The full form of Q Q plots is Quantile Quantile plots
- A quantile is a fraction and below that fraction, specific amount of values lie.
- Q Q plots are used to check if the two datasets that are plotted are from the same distribution or not

- Use of Q Q plots in Linear Regression:
    - To see if both the samples are from the same population.
    - If both the samples have similar tail
    - If both the samples have similar distribution shape.
    - If two samples have common location behaviour.
    - Q Q plots are used to plot quantiles of both the datasets against each other. Quantiles are where percentile of data lies below the given quantile value. Here, 25th quantile means that 25 percent of data lie below the 25th quantile and the rest 75 percent of data lies above the 25th quantile
    - If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

- Importance of Q-Q plot in Linear Regression:
    - If there are two sample data, it is usually expected for the data to have a similar distribution. In that case, the location and the scale estimators can pool both data sets for calculating a common location and scale
    - But, if the samples are different, the differences can be analysed as well to find insights.
    - The Q Q plot can provide more understanding into the nature of the differences than the analytical methods like the chi-square and Kolmogorov-Smirnov 2-sample tests.