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)

A: Inference of of the categorical variables effect on the dependent variable :

1. Number of bike rentals has increased in 2019 compared to 2018 which is a good sign for Boombikes as bike renting trend is showing up curve
2. Among weekdays Friday has highest number of bikes rented
3. Working days have 60% more people renting than non-working days or holidays. Interestingly the rent count is more on a wednesday when it is a holiday, compared to other weekdays. Hence if it is a holiday or working day affects the bike rental count significantly.
4. Among months - August/September has most bike rental count hinting at high rentals during Clear weather
5. Among seasons, Fall season has highest number of bikes rented because of the temp(pleasent weather) and Clear weather
6. Bikes were rented maximum when temperature was between 9-12 in all seasons. So company should boost their service when temperature is more than 9 degree celsius regardless of season

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

A: When we have a categorical variable with say 'n' levels, the idea of dummy variable creation is to build 'n-1' variables, indicating the levels.
Hence we first drop one of the columns, it reduces the correlations among dummy variables being created thereby addressing multicolinearity.

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

A: temp and atemp have 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)

A: Performed

- **Residual analysis of traning data :** by plotting histogram of error terms ,they should be normally distributed
- **Linearity :** scatter plot between actual and predicted values, the graph should show similar linear pattern
- **Multicolinearity :** Check for correlations using heatmap and remove highly correlated variables, p-value of variables and the Variance Inflation Factor (VIF)

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

1. temp(temperature)
2. yr(year)
3. season , since september month has good postive coefficient,which is beginning of fall season in US

General Subjective Questions

1. Explain the linear regression algorithm in detail (4 marks)
A: Linear Regression is the supervised machine learning model building method, in which the model finds the best fit linear line between the independent and dependent variable
i.e it finds the linear relationship between the dependent(y) and independent variable(x).

Linear Regression is of two types:
Simple Linear Regression is where only one independent variable is present and the model has to find the linear relationship of it with the dependent variable

Equation of Simple Linear Regression, where bo is the intercept, b1 is coefficient or slope, x is the independent variable and y is the dependent variable.

$$y = b_o + b_1 x$$

Multiple Linear Regression there are more than one independent variables for the model to find the relationship.

Equation of Multiple Linear Regression, where bo is the intercept,
b1,b2,b3,b4...bn are coefficients
x1,x2,x3,x4...xn are independent variables
y is dependent or target variables

$$y = b_o + b_1 x_1 + b_2 x_2 + b_3 x_3 \ldots . + b_n x_n$$

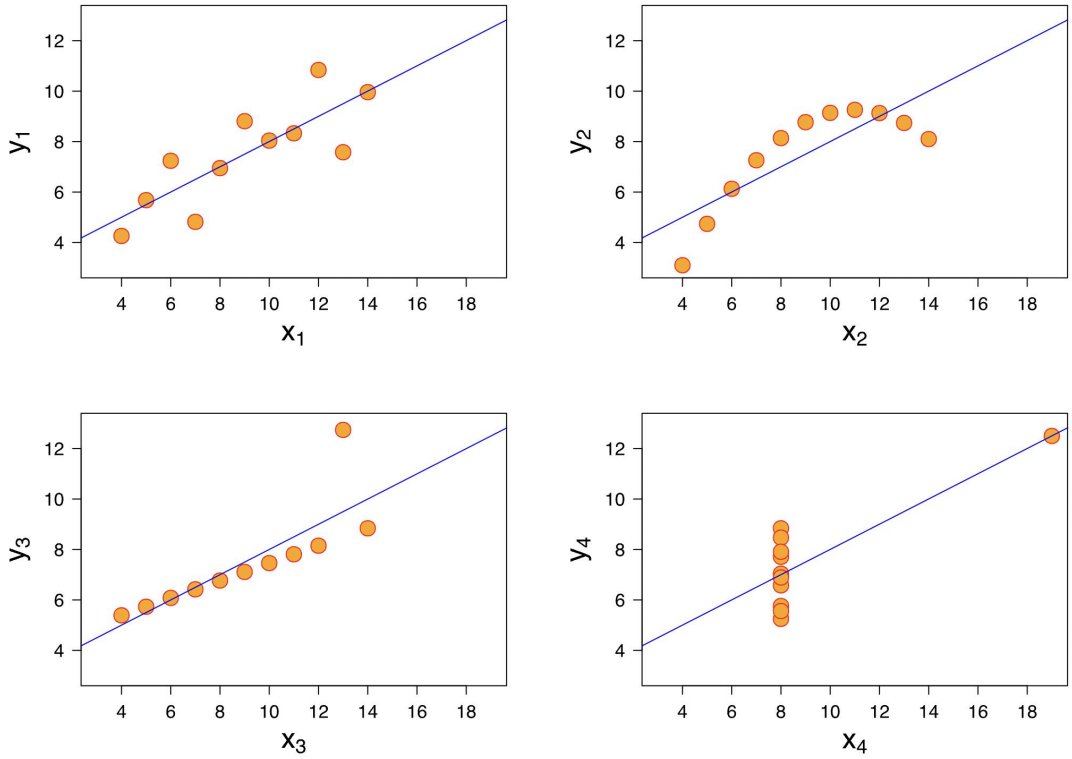
A Linear Regression model's main aim is to find the best fit linear line and the optimal values of intercept and coefficients such that the error is minimized.

Error is the difference between the actual value and Predicted value and the goal is to reduce this difference.

These are also few assumptions of linear regression:
1. Error terms should be normally distributed
2. Multicolinearitiy should be handled
3. Linearity exists between dependent and independent variables
4. Homoscedasticity in Error terms

2. Explain the Anscombe's quartet in detail. (3 marks)
A: Anscombe's quartet is when the data samples apper to be same when looking at the values statistically while the graphs of the data samples followcompletely different pattern.

As shown in the figure below



The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets

3. What is Pearson's R? (3 marks)
A: Pearson correlation coefficient measures linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations. It is represented by **r** : measures the strength of the correlation.

It is important to normalized measurement of the covariance, such that the result always has a value between –1 and 1.
-1 indicates a perfect negative linear correlation, 1 indicates a perfect positive linear correlation, and 0 indicates no linear correlation.

It is used along with p-value (probability that correlation coefficients are zero null hypothesis)

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
A: Scaling is a method used to normalize the range of independent variables or features of data.
It is extremely important to rescale the variables so that they have a comparable scale.
If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients.
This might become very annoying at the time of model evaluation. So it is advised to use standardization or normalization so that the units of the coefficients obtained are all on the same scale.

Normalized scaling : It is a scaling method that reduces duplication in which the features are scaled and moved between 0 and 1. Using the formula
normalisation : (x - xmin) / (xmax-xmin)
It is used when feature distribution is unclear

Standardized scaling : is a method for rescaling the values that meet the characteristics of the standard normal distribution while being similar to normalizing.
In thisfeature variable's mean is 0 and standard deviation is 1
standardization: (x- mu) / sigma

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
A: If there is perfect correlation, then VIF = infinity. If two or more independent variables have an exact linear relationship between them then we have perfect multicollinearity. Hence this leads to infinte VIF values if more independent are involved are correlated.

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

A: Q-Q plots are also known as Quantile-Quantile plots, plot the quantiles of a sample distribution against quantiles of a theoretical distribution. It helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.

In linear regression Q-Q plot can be used to check assumptions of LR like
- if error terms are distributed normally
- also to determine the linearity analysis of residulas