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? (3marks)**

   ANSWER:
   From the Dataset the Categorical variables are:
   Season, Year, Month, Holiday, Weekday, Working day, Weather Situation.

   To visualise the data a Boxplot was used. The inference acquired from the plots are:

   - **Season:** From the Boxplot the information we get is that during the Fall season there has been a greater number of booking and the season which attracted the least booking is Spring.
   - **Year:** The year 2019 has attracted a greater number of bookings than the previous year.
   - **Month:** The count of users has been significantly higher during the months of May, June, July, august, September
   - **Holiday:** The booking has found to be lower on a holiday
   - **Weekday:** Friday, Saturday, Sunday seems to have greater number of booking when compared to beginning of the week.
   - **Working day:** It does not affect the dependent variable much
   - **Weather Situation:** Clear and misty weather situation has attracted a greater number of users, during heavy rain or storm the has been no users.

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

ANSWER:

It is important to use drop_first=True during dummy variable creation as it reduces the correlations created among dummy variables. it helps in reducing the extra column created during dummy variable creation. If we do not use drop_first = True, then n dummy variables will be created, and these predictors (n dummy variables) are themselves correlated which is known as multicollinearity and it, in turn, leads to Dummy Variable Trap.

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

ANSWER:

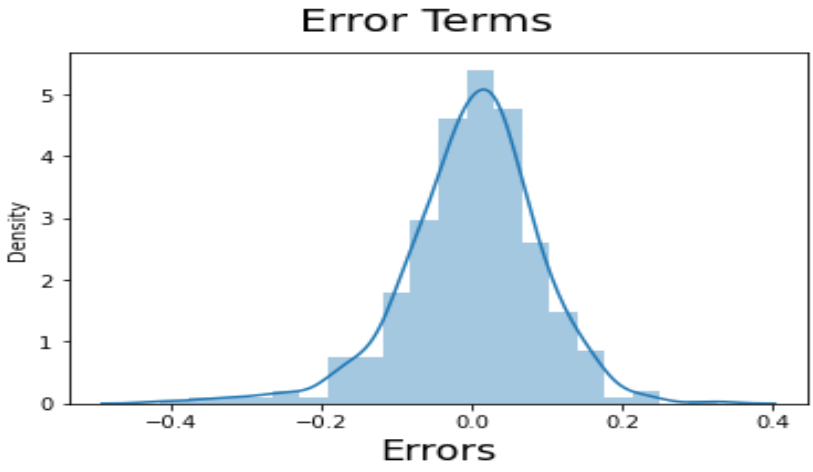
The numerical variable "temp" has the highest correlation with the target variable.

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

ANSWER:

I have validated the assumption of Linear Regression Model based on below 5 assumptions –

- **Normality of error terms** - Error terms should be normally distributed



- **Multicollinearity check** - There should be insignificant multicollinearity among variables.
- **Linear relationship validation** - Linearity should be visible among variables
- **Homoscedasticity** - There should be no visible pattern in residual values.
- **Independence of residuals** - No auto-correlation

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

ANSWER:

TEMPERATURE (temp): It has a coefficient of 0.5471. So, a unit increase in the temp variable increases the number of users by 0.5471 units.

LIGHT_SNOWRAIN: It has a coefficient of -0.2892. So, a unit decrease in the temp variable decreases the number of users by 0.2892 units.

YEAR: It has a coefficient of 0.2327. So, a unit increase in the temp variable increases the number of users by 0.2327 units.

General Subjective Questions

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

   Linear regression is a type of supervised machine learning algorithm.
   It analyses and models a Linear relationship between a dependent variable and one or many independent variables. The ultimate goal of the algorithm is to find that best linear equation which can predict the linear the dependent variable based on the given independent variable.

   The linear Regression model follows the equation of:

   Y = MX + C

   Here,

   Y is the dependent variable
   X is the independent variable
   M is the Slope of the line
   C is the Constant

   In a supervised learning algorithm, one of the most important tasks is Regression (Continuous data). The model tries to learn from the previous recodes of X and Y values, then a function is developed through this learning process. The function is then used to predict and unknown Y from known values of X. The Regression finds the **Best-Fit** line among various possible outcomes.

   Linear regression can be divided in Two types:

   1. **Simple Linear Regression:** It is used when only one independent variable is used to predict the dependent variable.
   2. **Multivariate Linear Regression:** It is called MLR when more than one independent variable is used to predict the dependent variable.

   There are few assumptions which are made about the data by the Linear regression model:

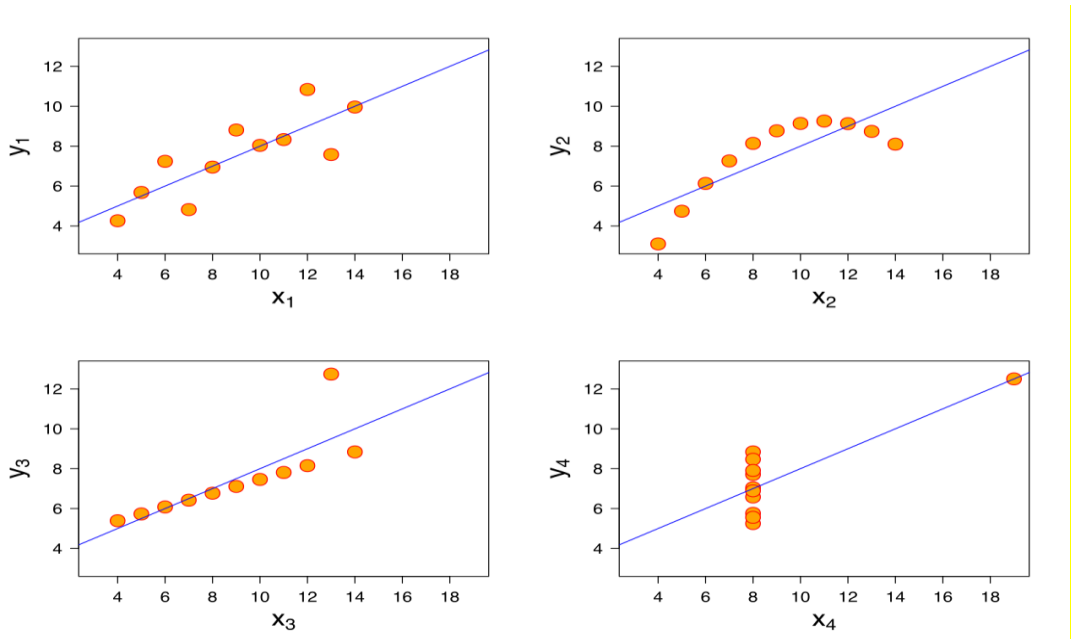
   1.Normality of Errors

2.multi-Collinearity
3.Homoscedasticity
4.Linear relationship between variables.

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

ANSWER:

Anscombe's quartet was developed by Francis Anscombe in the year 1973. It was to demonstrate both the importance of graphing data when analysing it, and the effect of outliers and other influential observations on statistical properties. It includes four datasets and they have almost nearly identical simple statistical features, and yet have very different distributions. When they are plotted on a graph, they appear very different.



Description:

From the Graph,

All the plots show the same regression line, but when we visual

Graph 1(Top-left): The Scatter-plot appears to be Simple linear regression

Graph 2(Top-right): It is not normally distributed

Graph 3(Bottom-left): The distribution of the plot seems to be linear, but due to one outlier the regression line is getting offset which otherwise should have been a different line.

Graph 4(Bottom-Right): It displays that when one High-Leverage Point is enough to produce a high correlation coefficient, even though other data point the graph do not indicate any relationship between the variables.

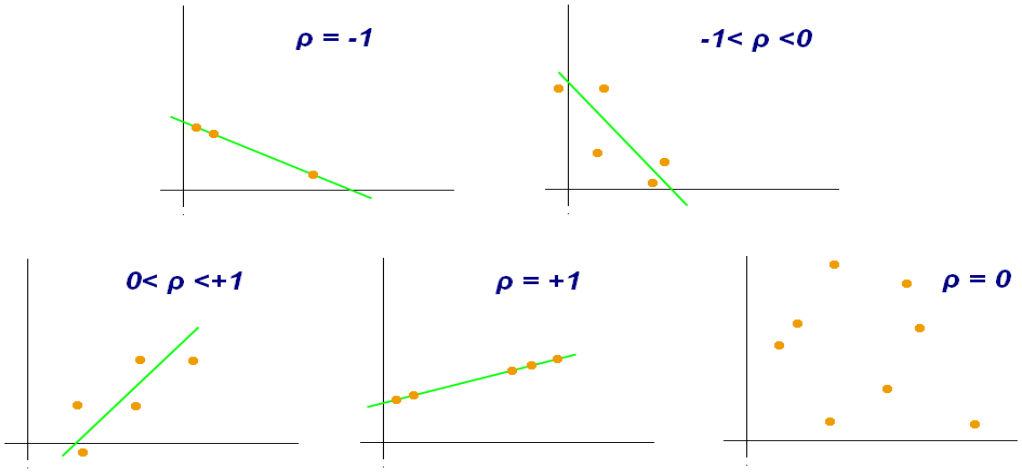
All the plots show the same regression line, but when we visualise them, they convey a unique information for each dataset.

3. What is Pearson's R? (3Marks)

ANSWER:

The **Pearson correlation coefficient (*r*)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

It is a Linear correlation between two datasets.



Pearson correlation coefficient (*r*)	Correlation type	Interpretation
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the **same direction**.
0	No correlation	There is **no relationship** between the variables.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the **opposite direction**.

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

ANSWERS:

Feature scaling is the process of normalizing the range of features in a dataset. It is performed during the data pre-processing to handle highly varying magnitudes or values in the dataset. For machine learning models to interpret these features on the same scale, feature Scaling is done.

If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Normalization is a suitable choice when the data's distribution does not match a Gaussian distribution. Outliers in the data will be impacted by normalization because it needs a wide range to function correctly. Highest and Lowest values of features are used for scaling in Normalised scaling.

When the data has a Gaussian distribution, standardization scaling in the machine learning model is useful but not always. In contrast to Normalization, Standardization does not always have a bounding range, thereby it is much less affected by Outliers.

Scales for normalization fall between [0,1] and [-1,1]. Standardization does not have any range restrictions. When the algorithms do not make any assumptions about the distribution of the data, Normalization is considered. When the data distribution is known then, standardization is applied.

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

ANSWER:

Variance Inflation Factor (VIF) is a measure of the amount of multicollinearity in regression analysis. Multi-Collinearity occurs when multiple independent variables are correlated with each other.

Due to this, the regression results get adversely affected. Variance Inflation factor (VIF) can be used to estimate to what extent the variance of a regression coefficient is inflated due to multi-collinearity.

If the value if VIF is large then, it suggests that there is high correlation between the variables. **IF there exists a perfect correlation between variables, the VIF will be equal to Infinity.**

To solve this issue of multi-collinearity, it is better to drop the variables from the dataset which are having high VIF values, as a result to reduce the multicollinearity.

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

where:

R_i^2 = Unadjusted coefficient of determination for regressing the ith independent variable on the remaining ones

So, If R^2 =1 in case of perfect correlation, then the VIF value will become equated to Infinity.

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

ANSWER:

The Quantile-Quantile plot or Q-Q plot is a Scatter Plot created by plotting 2 different quantiles against each other. It helps in determining if two datasets originate from populations with a common distribution.

Use of Q-Q Plot:

The Quantile-Quantile plot is used for the following purpose:

- Determine whether two samples are from the same population.
- Whether two samples have the same tail
- Whether two samples have the same distribution shape.
- Whether two samples have common location behaviour

Importance:

When there are two data set, it is better to analyse if the assumption of common distribution is true. If two data set do differ the it is also important to get some understanding of their differences.