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?
   Answer:
   By analysing categorical column, we can infer the effect of categories on the dependent variable:

- Most of the bikes were rented during May, June, July, Aug, Sep. This trend increased in the beginning of the year and then started decreasing as we approach end of the year.
- Fall season had the highest number of booking followed by summer.
- There was a sharp rise in booking count from 2018 to 2019.
- Clear weather is suitable for renting, which seems to be obvious.
- Bike rentals seem to remain almost the same every day of the week.
- Booking are nearly same on both working and non-working day.

2. Why is it important to use drop_first=True during dummy variable creation?
   Answer:
   drop_first=True is used to reduce the extra column created during dummy variable creation. It also reduces the correlations created among dummy variables. A variable with n levels can be represented by n-1 dummy variables.
   Example: Suppose you have 3 columns under size category (small, medium, large). drop_first=True will create dummy variable for medium and large with small as reference category. Without using drop_first=True, all the three variables will have dummy variable leading to multicollinearity.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   Answer:
   'temp' variable 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?
   Answer:
   The assumptions of Linear Regression can be validated by:
   - Normality of error terms
     - Error terms should be normally distributed.
   - Homoscedasticity
     - Error terms must have constant variance.
   - Linear relationship validation
     - Linearity should be visible among variable
   - Multicollinearity check
     - There should be insignificant multicollinearity among variables.
   - 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?

Answer:
Temperature, year and Light_snowrain are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

General Subjective Questions

1) Explain the linear regression algorithm in detail.

Answer:

Linear regression is a form of predictive modelling technique that tells relationship between the dependent variable and independent variable. Since linear regression is linear relationship it means it shows how the dependent variable will change if one or more independent change. The dependent variable is also known as target variable and independent variable is also known as predictor variable.

Mathematically the relationship can be represented with the help of the following equation:

$Y = mX + c$

Here Y is the dependent variable which we are trying to predict
X is the dependent variable
m here is the slope of regression
c is the Y-intercept.

Linear regression can be of two types;

- Simple Linear regression
- Multiple Linear regression

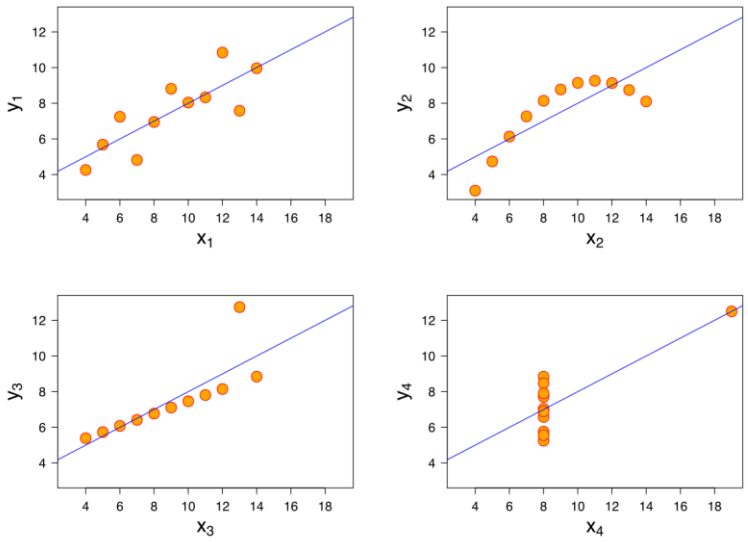
There are some assumptions about the dataset that are made by Linear regression:

- Normality of error terms
  - Error terms should be normally distributed.
- Homoscedasticity
  - Error terms must have constant variance.
- Linear relationship validation
  - Linearity should be visible among variable
- Multi-collinearity check
  - There should be insignificant multi-collinearity among variables.
- Independence of residuals
  - No auto-correlation

2) Explain the Anscombe's quartet in detail.

Answer:

Anscombe's quartet comprises four datasets that have nearly identical simple, descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x, y) points. Anscombe's quartet was developed by statistician Francis Anscombe.



- ✓ Dataset I appears to have clean data and well-optimized models.
- ✓ Dataset II is not distributed normally. It is not linear
- ✓ Dataset III have linear distribution but the calculated regression is affected by an outlier.
- ✓ Dataset IV shows that one outlier is sufficient to create high correlation coefficient.

The quartet is used to emphasize the importance of visualizing the dataset. It reveals many patterns and provide a clear understanding about the dataset.

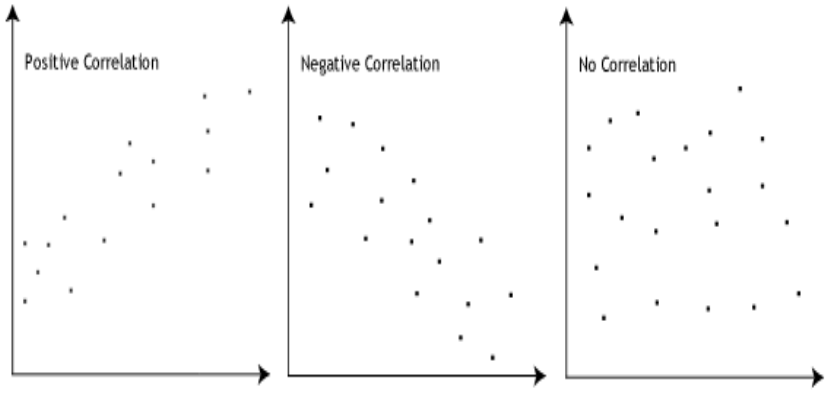
3) What is Pearson's R?

Answer:

Pearson's R also known as Pearson correlation coefficient, is a static that measures the strength and direction of the linear relationship between two variables.

It ranges from -1 to 1.

- 1 indicates a perfect linear relationship.
- -1 represents a perfect negative linear relationship.
- 0 indicated no linear relationship.



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

Answer:

Scaling is the process of normalizing the independent features in the data to a fixed range. It is one of the pre-processing step where the data scaled to a specific range to speed up the calculations. It is performed during model-building process. If scaling is not performed, then the algorithm tends to weigh high values magnitudes and ignore other parameters. Hence it results in incorrect modelling.

Normalised Scaling	Standardised Scaling
Maximum and Minimum features are used for scaling.	Mean and standard deviation is used for scaling.
It is used when the features are of different scales.	It is used to ensure zero mean and standard deviation.
It is affected by outliers.	It has no effect of outliers.
It scales between [0,1] to [-1,1].	It is not bounded in a certain range.
It is used when we don't know about distribution.	It is used when the distribution is normal.
Scikit-Learn provides a transformer called MinMaxScaler for normalization.	Scikit-Learn provides a transformer called StandardScaler for normalization.

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

Answer:

Variance inflation factor(VIF) is a measure of amount of multi-collinearity in a set of multiple regression variables.

The formula for VIF is:

$$VIF = 1/(1-R^2)$$

If there is perfect correlation, then VIF is infinite. In case of perfect correlation, the $R^2=1$, which leads to $1/(1-R^2)$ infinity. To solve this, we drop one of the variables from dataset causing the perfect multi-collinearity. We should drop the columns having VIF greater than 5. A high VIF shows a perfect correlation between two independent variables.

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

Answer:

Q-Q plot is a probability plot. It is graphical method used for comparing two distributions by plotting their quantiles against each other. It compares the quantiles of the data to the quantiles of the reference distribution.

Use of Q-Q plot:

Q-Q plot is used to visually assess whether a dataset is likely to have come from a particular distribution. Q-Q plot is used to determine whether or not two distributions are similar or not. If they are similar, Q-Q plot is more to be linear.

Moreover, it is also used for:

- Check assumptions: A Q-Q plot is used to check if a dataset is normally distributed or if two sets of data have the same distribution.
- Identify outliers: It is used to spot data points that are different from expected distribution.
- Compare distributions: It is helpful in comparing the distribution of a sample to a theoretical distribution
- Understand differences: It assess in understanding the differences between two sets of data.

Importance of Q-Q plot:

A Q-Q plot is a visual way to check for distributional assumptions. It is helpful in comparing quantiles of the sample data with quantiles of the theoretical distribution. In Linear Regression when we have a train and test dataset then we create Q-Q plot by which we can confirm that both the data train and test data set are from the population with the same distribution or not.

Advantages:

- It can be used with sample size also.
- Many distributional aspects like shifts in location, shifts in scale, changes in symmetry and presence of outliers can all be detected from Q-Q plot.

Q-Q plot can also be used on two datasets to check:

- If both datasets came from population with common distribution.
- If both datasets have similar type of distribution shape.
- If both datasets have common location and common scale.
- If both datasets have tail behaviour.

Q-Q plot provide clear, efficient and adaptable method for evaluating the distribution of data. They enable quick identification of outliers, confirmation of distribution assumptions and analysing critical tail behaviour hence making them a significant tool for data analysis.