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)

Answer:

- Demand is dull during Jan, Feb and is gradually increased (more demand) till sep then it falls down again
- Demand for bikes are more in Fall, Summer, Winter rather than Sprint seasons
- From the above bar graph, the bikes are due to demand during the weekdays rather than the holidays
- There is no much variance on the week days or week ends
- Working Day: Business is operating almost same/similar working days and in nonworking days as well
- As mentioned above during the segregation and replacing the categorial values, there is no data for one of the category i.e. Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog.
- Demand of bikes are high during clear_fewClouds and reduces during mist_cloudy and LightSnow_LightRain situations
- Demand is almost same on all weekdays/weekends

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

Answer:

- drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.
- Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi_furnished, then It is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished.

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

Answer:

- Temp

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

Answer:

1. Error term should show the normal distribution with mean at Zero
2. No Patters are identified when we plot the residuals

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:

The following are the top 3 features contributing significantly explaining the demand of the shared bikes:

- Temp with co-efficient value of 0.492
- weathersit_LightSnow_LightRain with co-efficient value of -0.285
- and yr with co-efficient value of 0.234

General Subjective Questions

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

Answer:

- Linear regression is a statistical method that is used for predictive analysis.
- Linear regression makes predictions for continuous/real or numeric variables on a provided dataset
- Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression.
- This algorithm helps us to find how the value of the dependent variable is changing according to the value of the independent variable.

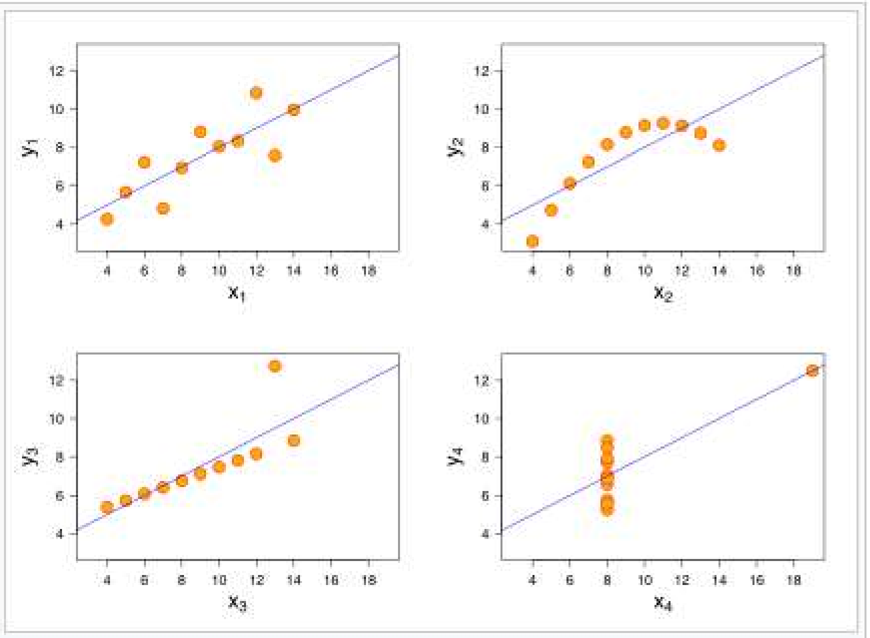
2. Explain the Anscombe's quartet in detail. (3 marks)

Answer:

Anscombe's quartet says 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.

To prove this, person named **Anscombe's** created 4 datasets where nearly all the datasets have identical simple descriptive statistics(same mean, SD and so on..)

And Anscombe quartet says that when these datasets are plotted, they considerably vary



<u>**So the theory is : Always plot your data before deriving any conclusions**</u>

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

<u>**Answer:**</u>

This is another method of measuring the linear correlation, It helps us in understanding the strength and the direction of the relationship(positive direction or negative direction) between two variables

The range of pearson correlation will be always between -1 and 1.

Pearson Correlation Coefficient (r)	Correlation type
Between 0 and 1	Positive Correlation
0	No correlation
Between 0 and -1	Negative correlation

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

<u>**Answer:**</u>

Scaling is a method used to normalize the range of independent variables or features of the data. It is also known as data normalization and is generally performed during the data preprocessing step.

Most of the times, the 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.

Normalization	**Standardization**
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
Scales values between [0, 1] or [-1, 1].	No specific range is defined
Used when we have outliers and we want to keep the outliers	Used when there is no use outliers

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

Answer:

Variance Inflation Factor: it is calculated as 1/(1-R2), to get VIF has infinity the value of R2 should be 1.

Meaning that there is perfect co-relation between two independent variables or perfect multicollinearity.

In this case we can drop one such feature

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

Answer:

Q-Q Plots are plots of two quantiles against each other. Quantile is a fraction where certain values fall below than quantile. The purpose of Q-Q quantile is to find out whether two data sets have same distribution.

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.