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: We can see from analysis that demand of bikes is influenced by following categorical variable :

- Season- demand decreases in spring.
- Month- demand increases in the months of March, May, June, July, August, September, October
- Weathersit- demand decreases in mist_cloudy and Light rain_Light snow_Thunderstorm
- Weekday- demand increases on Thursday
- Year- demand was high during 2019

Therefore final recommendation for the company is that demand is high in the months of March, May, June, July, August, September, October

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

Answer : It reduces the extra column created during the creation of dummy variables. Hence it reduces the correlations created among dummy variables.

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

Answer: atemp has highest correlation.

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

- Linear relationship

  Linear relation between dependent variable and independent variable(using scatter plot).

  Both the train and the test results shows linear dependency.

- Homoscedasticity

  Homoscedasticity means that the residuals have constant variance no matter the level of the dependent variable or in other words error term is the same across all values of the independent variables.

  A scatter plot of residual values vs. predicted values is a good way to check for homoscedasticity and we found that there is no specific pattern in the distribution.

- Absence of multicolinearity

Multicollinearity refers to the fact that two or more independent variables are highly correlated (or even redundant in the extreme case).

Pairplots and heat maps are used to indentify multicolinearity-atemp showed high correlation which was dropped in our model.

It is also verified through VIF(Variance inflation factor), VIF greater than 5 indicates high multicolinearity ,in this model we have not considers VIF greater than 3.

- Independence of residuals

  Autocorrelation occurs when the residual errors are dependent on each other and it reduces model's accuracy.

  Autocorrelation can be tested with the help of Durbin-Watson test. This statistic will always be between 0 and 4. The closer to 0 the statistic, the more evidence for positive serial correlation. The closer to 4, the more evidence for negative serial correlation.

  Model 25 has the value of Durbin-Watson test as 2. A value of 2.0 means there is no autocorrelation detected in the sample.

- Normality of errors

  If the residuals are not normally distributed, Ordinary Least Squares (OLS), and thus the regression, may become biased.

  From the graph it is clear that errors are normally distributed.

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

- Month- demand increases in the months of March, May, June, July, August, September, October
- Weathersit- demand decreases in mist_cloudy and Light rain_Light snow_Thunderstorm
- Weekday- demand increases on Thursday

General Subjective Questions

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

Answer : Linear regression helps us to determine relationship between the independent and dependent variables using a straight line. It is a method of finding the best straight fitting line to the given data.

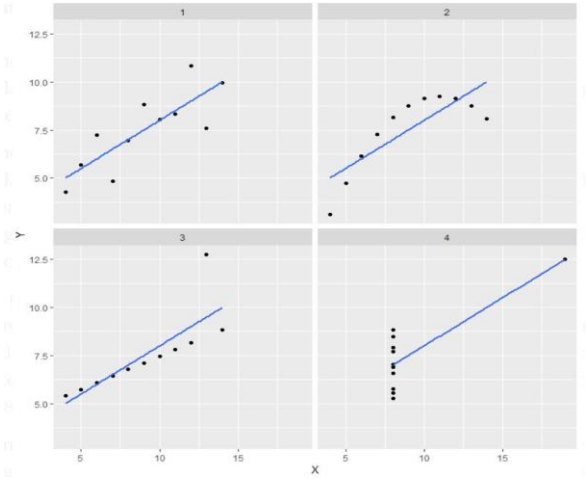
Simple linear regression**:** attempts to explain the relationship between a dependent variable and an independent variable using a straight line.

Multiple linear regression**:** explain the relationship between two or more independent variable and a dependent variable and an using a straight line. It is required when one variable is not sufficient to create a good model and make predictions.

- The independent variable is also known as the predictor variable.
- The dependent variables are also known as output variable

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

Answer : Anscombe's quartet comprises four data that appear similar when using typical summary statistics, yet have very different distributions and appear very different when graphed.



Explanation of this output**:**

- In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

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

Answer: It is a common coefficient used in linear regression to measure the strength between the two.

Results always lies between -1 to 1, -1 indicates a negative linear relationship between variables, and r of 0 indicates no linear relationship between variables, and an r of 1 indicates a perfect positive linear relationship between variables.

Assumptions

- Variables should be normally distributed
- There should be no significant outliers
- Two variables should have linear relationship
- Each variables should be continuous
- The observations are paired observations
- Homoscedasticity

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

Answer: It is a data pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Usually collected data set contains variables with varying 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.

In normalized scaling values range between 1 and 0 (it is also known as Min MaxScaling) whereas in Standardized scaling values are centered around mean with a unit of standard deviation.

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

Answer: VIF is a indicator of muticolinearity /correlation, it calculates how well one independent variable is explained by all the other independent variables combined.
If VIF = infinity , then there is perfect correlation.
Thumb rule of VIF

- >10 variable is eliminated .i.e variable is dropped.
- >5 variable is worth inspecting
- <5 no need to eliminate the variable.

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

Answer : Quantile-Quantile (Q-Q) plot, is a graphical tool to assess if a set of data came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

It is used in linear regression when we have training and test data set obtained separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions. If the all the points are on or close to straight line inclined at 45 degree from x axis then it has similar distribution.

Advantages

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