Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

- The demand of bike is almost similar throughout the weekdays.
- Bike demand in the fall is the highest followed by the summer on a close range
- Bike demand is lowest in the spring ○ Bike demand in year 2019 is higher as compared to 2018
- Bike demand is high in the months from May to October
- Bike demand is high for the weather Clear, Few clouds, Partly cloudy, Partly cloudy while it is low for the weather Light Snow, Light Rain + Thunderstorm + Scattered qclouds, Light Rain + Scattered clouds

---

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

- It is mandatory to drop unwanted, redundant and extra columns while building a Model
- It is helpful in reducing collinearity among variables
- And using k-1 dummy variables we can determine values for k columns/variables
- For example, furnished, semi-furnished and un-furnished. We can only take 2 variables as furnished will be 1-0, semi-furnished will be 0-1, so we don't need unfurnished as we know 0-0 will indicate un-furnished

---

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer: <Your answer for Question 3 goes below this line> (Do not edit)

- temp and atemp both have same correlation with target variable of 0.63 which is the highest among all numerical variables.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 4 goes below this line> (Do not edit)

- We validate the assumptions of Linear Regression by plotting a distplot of the residuals and analysing to see if it's a normal distribution i.e. mean = 0
- Low VIF i.e. Less Multi-collinearity between features

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

- temp (Temperature)
- weathersit 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
- yr (Year)

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

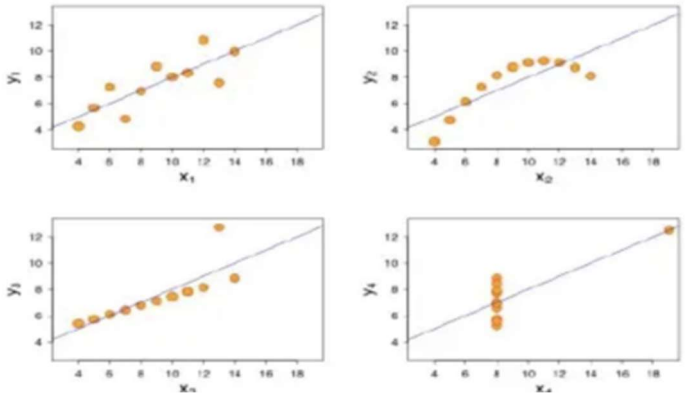
- Linear regression is a supervised machine learning method. It is a form of regression, where the target variable is continuous. It estimates the relationship between a target variable and one or more predictor variables
- It finds a linear equation that best describes the correlation of the explanatory variables with the dependent variable. This is achieved by fitting a line to the data using least squares
- The Equation of linear Regression is y= m1x1 + m2x2 + m3x3 + ……. + m(n) x(n) + c where y is target variable and x1, x2 , x3 …… xn are predictor variables
- We have two unknowns, m, and c, and we need to choose those values of m and c, which provides us with the minimum error
- We need to get the best fit line which is the line that has the minimum error
- When the error is calculated using the sum of squared error, this type of regression is known as OLS, i.e., Ordinary Least Squared Error Regression
- Error function is explained by 'e = - y', and error depends on the values of 'm' and 'c'. Our aim is to build an algorithm which can minimize the error
- And in order to do so we use cost function of Linear Regression, Which is: J (mi, c)= (1/2n)Σ(yi – yp)^2 where yi and yp are expected values and predicted values ○ Cost function measures the performance of a Machine Learning model for given data

---

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

- Anscombe's quartet comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph
- The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When

plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line



- Top-Left appears to be simple linear relationship
- Top-Right shows non-linear relationship, correlation coefficient is irrelevant
- Bottom-Left have linear relationship but have a regression line (outliers)
- Bottom-Right isn't linear but got adjusted due to outliers
- Hence, It is better to visualize the data and remove outliers beforehand

---

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

- Pearson's R measures the strength of association of two variables
- It is the covariance of the two variables divided by the product of their standard deviations
- It has a value ranging from +1 to -1
- +1 means a positive linear correlation, meaning that if one variable increases then the other also increases and vice versa
- 0 means no correlation, meaning that the increase/decrease in one variable doesn't affect the other
- -1 means a negative linear correlation, meaning that if one variable increases then the other also decreases and vice versa

---

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

- Scaling is a geometric change that linearly enlarges or reduces things. A property of objects or rules known as scale invariance is that they remain unchanged when scales of length, energy, or other variables are multiplied by a common factor
- Scaling performed because it is a data pre-processing procedure used to normalize data

within a specific range by applying it to independent variables. Additionally, it aids in accelerating algorithmic calculations. The majority of the time, the obtained data set includes characteristics that vary greatly in magnitudes, units, and range

- The difference between normalized scaling and standardized scaling is that the values of a normalized dataset will always fall between 0 and 1. A standardized dataset will have a mean of 0 and a standard deviation of 1, but the maximum and minimum values are not constrained by any specified upper or lower bounds.
- Formula of Normalized scaling:

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

- Formula of Standardized scaling:

$$x = \frac{x - mean(x)}{sd(x)}$$

-
-

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

- The value of VIF is calculated by the below formula

$$VIF_1 = \frac{1}{1 - R^2}$$

-
- If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

- Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly 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
- Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.
- It is used to check following scenarios: ○ If two data sets come from populations with a common distribution
- If two data sets have common location and scale
- If two data sets have similar distributional shapes ○ If two data sets have similar tail

behavior