1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
Ans: The categorical variables available in the assignment are “season”, “workingday”, “weathersit”, “weekday”, “yr”, “holiday”, and “mnth”.

1: “season” –

- Based on the data available, the most favourable seasons for biking are summer and fall.
- Spring has significant low consumption ratio

2. “weathersit” –

-Most favourable weather condition is the clean/few clouds days.

3. “weekday” –

- If we consider “cnt” column we do not find any significant pattern with the weekday.

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

Ans: Using one-hot encoding the dummy variables are created to cover the range of values of categorical variable. Each dummy variable have 1 and 0 values. 1 is used to depict the presence and 0 for absence of the respective category. This means if the category variable has 3 categories, there will be 3 dummy variables.
The drop_first = True is used while creating dummy variables to drop the base/reference category. The reason for this is to avoid the multi-collinearity getting added into the model if all dummy variables are included. The reference category can be easily deduced where 0 is present in a single row for all the other dummy variables of a particular category.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Ans: temp of 0.63

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

Ans: Validated the assumptions of linear regression by checking the VIF, error distribution of residuals and linear relationship between the dependent variable and a feature variable.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Ans: holiday, yr, windspeed

6. Explain the linear regression algorithm in detail.

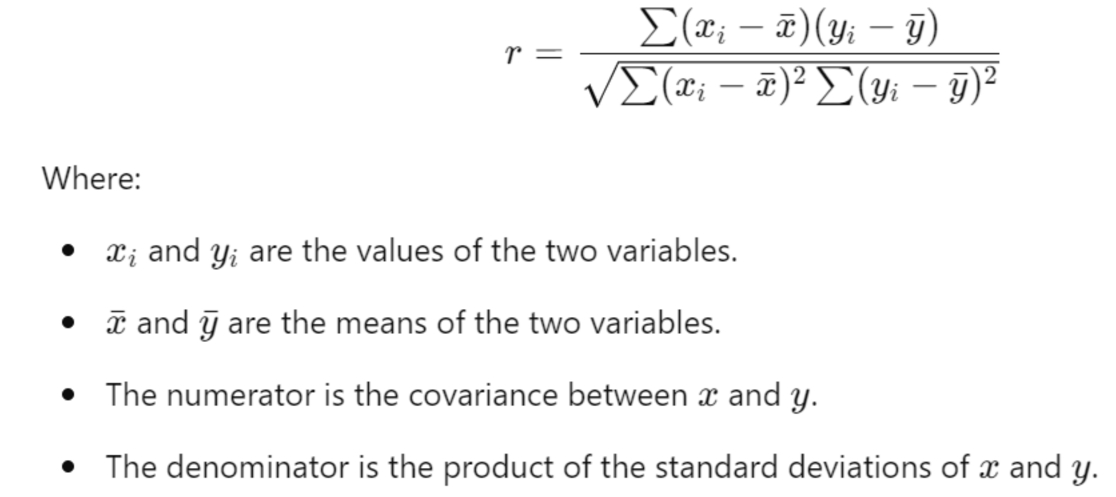
Ans: Linear regression is the method of finding the best linear relationship within the independent variables and dependent variables.
Page 5 of 8
• The algorithm uses the best fitting line to map the association between independent variables with dependent variable.
• There are 2 types of linear regression algorithms
o Simple Linear Regression – Single independent variable is used.
▪ □ = □0 + □1□ is the line equation used for SLR.
o Multiple Linear Regression – Multiple independent variables are used.
▪ □ = □0 + □1□1 + ⋯ + □□□□+ ∈ is the line equation for MLR.
o □0 = □□□□□ □□ □ℎ□ □ □ℎ□□ □ = 0 (□ □□□□□□□□□)
o □1, □2, … , □□ = □□□□□ □□ □ℎ□ □□□□□□□□.
• Cost functions – The cost functions helps to identify the best possible values for the □0, □1, □2, …, □□ which helps to predict the probability of the target variable. The minimization approach is used to reduce the cost functions to get the best fitting line to predict the dependent variable. There are 2 types of cost function minimization approaches – Unconstrained and constrained.
o Sum of squared function is used as a cost function to identify the best fit line. The cost functions are usually represented as
▪ The straight-line equation is □ = □0 + □1□
▪ The prediction line equation would be □□□□□ = □0 + □1□□ and the actual Y is as Yi.
▪ □□□ □ℎ□ □□□□ □□□□□□□□ □□□□ □□ □(□1, □0) = ∑(□□ − □1□□ − □0)2
o The unconstrained minimization are solved using 2 methods
▪ Closed form
▪ Gradient descent
• While finding the best fit line we encounter that there are errors while mapping the actual values to the line. These errors are nothing but the residuals. To minimize the error squares OLS (Ordinary least square) is used.
o □□ = □□ − □□□□□ is provides the error for each of the data point.
o OLS is used to minimize the total e2 which is called as Residual sum of squares.
o RSS = = ∑i=1 n (□□ − □□□□□)2
• Ordinary Lease Squares method is used to minimize Residual Sum of Squares and estimate beta coefficients.

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

Ans: Anscombe's quartet consists of four data sets that have nearly identical simple descriptive statistics but have
very different distributions and appear very different when presented graphically. Each dataset consists of eleven
points. The primary purpose of Anscombe’s quartet is to illustrate the importance of looking at a set of data
graphically before beginning the analysis process as the statistics merely does not give the an accurate
representation of two datasets being compared.

3. What is Pearson’s R? (3 marks)

Ans: **Pearson’s r**, also known as **Pearson’s correlation coefficient**, is a measure of the strength and direction of the linear relationship between two continuous variables. It was developed by Karl Pearson, and it is one of the most commonly used methods for understanding how two variables are related.



$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Where:

- x_i and y_i are the values of the two variables.
- $\bar{x}$ and $\bar{y}$ are the means of the two variables.
- The numerator is the covariance between x and y.
- The denominator is the product of the standard deviations of x and y.

Magnitude Interpretation:

- **0.0 to 0.1**: Very weak or no linear relationship.
- **0.1 to 0.3**: Weak linear relationship.
- **0.3 to 0.5**: Moderate linear relationship.
- **0.5 to 0.7**: Strong linear relationship.
- **0.7 to 1.0**: Very strong linear relationship.

Positive vs. Negative:

- **Positive r**: When one variable increases, the other tends to increase.
- **Negative r**: When one variable increases, the other tends to decrease.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Ans :
What - The scaling is the data preparation step for regression model. The scaling normalizes these varied datatypes to a particular data range.
• Why – Most of the times the feature data is collected at public domains where the interpretation of variables and units of those variables are kept open collect as much as possible. This results in to the high variance in units and ranges of data. If scaling is not done on these data sets, then the chances of processing the data without the appropriate unit conversion are high. Also the higher the range then higher the possibility that the coefficients are impaired to compare the dependent variable variance. The scaling only affects the coefficients. The prediction and precision of prediction stays unaffected after scaling.
• Normalization/Min-Max scaling – The Min max scaling normalizes the data within the range of 0 and 1. The Min max scaling helps to normalize the outliers as well.

□□□□□□□□□□□□□: □ = □ − min(□) / max(□) − min(□)

• Standardization converges all the data points into a standard normal distribution where mean is 0 and standard deviation is 1.

□□□□□□□□□□□□□□□: □ = □ − □□□□(□) /□□(□)

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

Ans:

□□□ = 1 /(1 − □2)

The VIF formula clearly signifies when the VIF will be infinite. If the R2 is 1 then the VIF is infinite. The reason for R2 to be 1 is that there is a perfect correlation between 2 independent variables.

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

Ans: Q-Q plots are the quantile-quantile plots. It is a graphical tool to assess the 2 data sets are from common distribution. The theoretical distributions could be of type normal, exponential or uniform. The Q-Q plots are useful in the linear regression to identify the train data set and test data set are from the populations with same distributions. This is another method to check the normal distribution of the data sets in a straight line with patterns explained below • Interpretations o Similar distribution: If all the data points of quantile are lying around the straight line at an angle of 45 degree from x-axis. o Y values < X values: If y-values quantiles are lower than x-values quantiles. o X values < Y values: If x-values quantiles are lower than y-values quantiles. o Different distributions – If all the data points are lying away from the straight line. • Advantages o Distribution aspects like loc, scale shifts, symmetry changes and the outliers all can be daintified from the single plot. o The plot has a provision to mention the sample size as well.