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)

From the model built, following were the interpretations with respect to their dependent variable

1. In the data we had demand for year of 2018 and 2019 and by this, we were able to see the increase in demand for year 2019

2. Demand will be high during the working days and demand goes down during holidays

3. Demand will be high whenever there is an increase in temperature

4. During seasons of summer and winter, demand will be higher and demand goes down during spring season

5. Demand goes down during higher windspeed

6. In terms of months, september will have higher demand and july will have lower demand

7. During moderate weather, demand goes down

---

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

It is important to use drop_first = 'True' to drop reduntant dummy variable as we know for a categorical variable with n levels, we need only n-1 dummy variables

---

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

Temp variable has highest correlation with target variable of cnt

---

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

To validate the assumptions of linear regression, need to do residual analysis where we can check the assumption of whether the residuals(error terms) are normally distributed and its mean is centered around zero.

---

5. Based on the final model, which are the top 3 features contributing significantly towards

explaining the demand of the shared bikes? (2 marks)

Top 3 features contributing towards the demand of shared bikes are decided based on the coefficients of the independent variables and as per the formula, temperature (positively correlated), year (positively correlated - for every increasing year, demand is expected to grow) and windspeed (negatively correlated)

General Subjective Questions

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

Linear regression uses linear equation y = mx+c to predict values for continuous target variables

Linear regression assumes that target variable is linearly dependent on its independent variables and error terms should be normally distributed with its mean centered at zero and error terms should have constant variance.

For every unit increase in dependent variable, target variable is expected to increase or decrease based on its coefficient

In case of including multiple independent variables in prediction of target variable , y is calculated as follows

y= m1x1+m2x2+...+mpxp+constant

On including multiple variables, any dependent variable should be linearly dependent on the target variable on assumption that other dependent variables remain constant. Because of this condition, we should make sure that dependent variable correlated with other dependent variables should not be considered . Including correlated variables will swing the coefficients and it will be difficult to interpret the results but does not affect the prediction.

To deal with multi collinearity, check using pair-plot and vif to eliminate correlated variables

---

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

Anscombe’s quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

It demonstrates the importance of visualizing data and to show that summary statistics alone can be misleading.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.

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

In statistics, the Pearson correlation coefficient (PCC) is a correlation coefficient that measures linear correlation between two sets of data.

It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationships or correlations

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

Scaling is bringing all the values of the variables in same unit but in real time, variables are expected to have different range of values

Scaling is performed to bring all the variable values within a range so that it is easy to interpret the coefficients of variables from the model built.

normalized scaling is (x-xmin)/(xmax-xmin) and the range of values will be between 0 and 1 while,

Standardized scaling is (x-mean)/standard deviation) and the range is the not fixed here.

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

(3 marks)

The Variance Inflation Factor (VIF) can be infinite when there is perfect correlation between variables, or when one variable can be expressed as a linear combination of other variables

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

(3 marks)

A QQ plot is a scatterplot created by plotting two sets of quantiles against one another.

A Q-Q plot can be used in regression models to check some of the assumptions that are required for valid inference.

For example, you can use a Q-Q plot to check if the residuals of the model are normally distributed, which is an assumption for many parametric tests and confidence intervals. You can also use a Q-Q plot to check if the residuals have a constant variance, which is an assumption for the homoscedasticity of the model. To do this, you need to create a Q-Q plot for the residuals of the model and compare them with the normal distribution.