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)

The categorical variables that would affect the ctn are :

Season (Summer and Fall) – It has been found that there has been a positive correlation between ctn and season

Summer Season – Positive correlation of 0.13

Fall Season – Postive correlation of 0.37

More Bikes are rented during the summer / Fall Season

The second categorical variable that has a influed on CTN is:

Weathersit :

(Mist + Cloudy Weather) – There is a negative correlation of -0.18.

(Light Snow / Light Rain) – There is a negative correlation of -0.23.

This indicates that in cloud weather / light snow / light rain conditions the demand for bikes goes down.

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

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.

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

Temp and atemp have a positive correlation of 0.64 and 0.65 which is the highest correlation with the target variable ctn.

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

After building the training set the assumptions of Linear Regression were validated with the following:

!) Check the value of R2 squared – R2Squared should be greater than 80% for the selected variables

2) Check on the p coefficient for the vaiables and ensure that the p coefficient is less than 0.05

3) Check on the Adjusted R Square and ensure that it is greater than 80%

4) Calcualte the VIF and ensure that the VIF for each variable in the linear regression model is less than 5

5) Check if the error terms are normally distributed

6) Check the model with test data that was selected (df_test)

7) Plot between ytest and ypredict values. You should have a scatter plot with correlation between the data points.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

The top three features that contribute towards the demand of bikes are:

a) Temp – Temperature (tmp has a 0.64 positive correlation with cnt). As per the scatter plot, higher the temperate the demand for bikes increases
b) Year – Year has a 0.59 positive correlation with cnt. The demand for bike has increased in 2018 and 2019.
c) Season – The summer and fall season has an increase in demand for bikes

d) Humidity also has a positive correlation with cnt – higher humidity or higher temperature has seen an increase in demand of bikes.

General Subjective Questions

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

Linear regression is a statistical technique that models the relationship between a dependent variable and an independent variable using a linear equation.
Linear regression predicts the value of a dependent variable by using a known independent variable. One example would be income and expenses data – the relation between income and expenses.
There are two kinds of linear regression

a) Simple Linear Regression – There is only one explanatory variable
b) Multiple Linear Regression – There are more than one explanatory variable

The equation for a linear regression line is

$y=b+mx$

b is the y-intercept and mis the slope. The slope and y-intercept together allow to calculate any point on the linear regression line.

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

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

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.

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

The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable.

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y.

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

Scaling, or feature scaling, refers to the process of transforming the values of variables to a specific range. This is often done to ensure that all variables have a comparable impact on the regression model. Scaling can help prevent certain variables from dominating the model due to their larger magnitude

Normalization is a technique used to scale numerical data in the range of 0 to 1. This technique is useful when the distribution of the data is not known or when the data is not normally distributed. On the other hand, standardization is a technique used to transform data into a standard normal distribution.

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

The Variance Inflation Factor (VIF) is infinite when there is perfect correlation between variables, meaning the regressor is equal to a linear combination of other regressors:

If the VIF values for three variables are infinite, it means these variables are perfectly or near-perfectly multicollinear. This means these variables can predict each other with a high degree of accuracy

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

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. It helps to determine if two data sets come from populations with a common distribution.

It is used to check following scenarios:
If two data sets have :

- come from populations with a common distribution
- have common location and scale
- have similar distributional shapes
- have similar tail behaviour

The advantage of Q-Q Plots are :

- It can be used with sample sizes
- 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.