Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans- Upon the bivariate analysis of the categorical variables on the dataset using boxplots the observations are as follows:

There is high number of bike using customers when

1. it is the fall season

2. in the year 2019

3. not on a holiday

4. marginally higher on a monday

5. slightly more on a working day

6. when the weather is clear

Q2. Why is it important to use drop_first=True during dummy variable creation?

Ans- It is important to use drop_first=True, as it helps to create the dummy variables for n-1 values.

Essentially while creating the dummy values it helps to avoid redundancy in data and for more clear interpretation of the values in the model creation.

Eg- if there are three values:

	Apple	Mango	Strawberry
Apple	1	0	0
Mango	0	1	0
Strawberry	0	0	1

We can drop the last column – strawberry here, it would still make sense for strawberry- 00

Apple-10, mango 01.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans- The variables temp and atemp have highest correlation with cnt.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans- I have validated the assumptions by

1. firstly predicting the values for y_train_pred, then calculated an error term as = y_train-y_train_pred. Then plotted this on the distplot to check If the error terms follow a normal distribution or not and if the mean =0.
2. Plotted a graph between the y_train_pred and y_train and observed a linear relation as it should be.
3. From the same graph we can see that the variations of the datapoints are not that widespread.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans- Year, temperature, windspeed

General Subjective Questions

Q1. Explain Linear regression algorithm in detail.

Ans- Linear regression algorithm is a supervised machine learning model. It is a predictive model, it takes input as continuous data and builds a relationship between the dependent and independent variables. This relationship is built by trying to fit in a best straight line that would pass through most of the points associated with x and y coordinates(x- predictor variable, independent variable, and y is the dependent variable, target variable).

It follows the form: y= mx+c

c represents the intercept (the value of (y) when (x) is zero)

m represents the slope of the line (how much (y) changes for a unit change in (x)

There are 2 types of linear regression:

1. Simple linear regression- when there is one independent variable/ feature
2. Multiple linear regression- when there are more than one independent variables/ features

Assumptions of Linear Regression:

1. Linear relationship: The relationship between (x) and (y) is linear.
2. Independence: Residuals (errors) are independent of each other.
3. Homoscedasticity: Residuals have constant variance.
4. Normality: Residuals follow a normal distribution.

The cost function here that we try to reduce is the square of the errors i.e, y_pred-y_actual.

When we reduce this error, the model will learn to predict the values closer to the datapoints, but we should keep in mind that the model shouldn't overfit.

Q2. Explain the Anscombe's quartet in detail.

Ans- Anscombe's quartet is a set of four datasets that share almost identical simple descriptive statistics, but when plotted on graph, the show a different pattern and distribution of the data points. These datasets were constructed based on two main points:

1. Graphing data matters: Anscombe wanted to demonstrate the importance of visualizing data alongside numerical calculations. While statistics provide valuable insights, graphs reveal nuances that numbers alone might miss.
2. Influence of Outliers: The quartet highlights how outliers and other influential observations can significantly impact statistical properties.

Characteristics of the dataset:

Dataset I: Appears as a simple linear relationship, suitable for linear regression.

Dataset II: Shows a clear relationship but not linear; Pearson correlation is not relevant.

Dataset III: Has a linear relationship but should use a different regression line (robust regression would be better).

Dataset IV: Demonstrates how a high-leverage point can inflate the correlation coefficient, even when other data points show no clear relationship.

Statistics for all 4 datasets are as follows:

Mean of x: 9 (exact)

Sample variance of x: 11 (exact)

Mean of y: 7.50 (rounded to 2 decimal places)

Sample variance of y: 4.125 (±0.003)

Correlation between x and y: 0.816 (rounded to 3 decimal places)

Linear regression line: (y = 3.00 + 0.500x) (coefficients rounded to 2 and 3 decimal places, respectively)

Coefficient of determination (R-squared): 0.67 (rounded to 2 decimal places)

Q3. What is Pearson's R?

Ans- Pearson's R is a statistical measure that quantifies the linear relationship between two continuous variables. It checks how closely the data points are present in a scatter plot and if they follow a straight line.

Pearson's (r) ranges from -1 to 1.

An (r) value of 1 indicates a perfect positive linear relationship (as one variable increases, the other increases proportionally).

An (r) value of -1 indicates a perfect negative linear relationship (as one variable increases, the other decreases proportionally).

An (r) value of 0 suggests no linear relationship (variables are not correlated).

If the (r) value is closer to 1 or -1 it means that there exists a strong linear relationship, else if it is a value closer to 0, it means that there is no linear relationship.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans- Scaling is a process by which we try to bring all the columns having vast ranges to fit into a similar range so that we can easily interpret the values and run and train the model easily.

If the values are at all different scales, then if we run the linear regression model we may get a false constant and coefficient and also the p-values may get affected.

Difference between normalized scaling(minmax) and standardizing:

Normalization:

1. The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.
2. Scales values between ([0, 1]) or ([-1, 1]).
3. Highly affected by outliers.
4. Scikit-Learn provides MinMaxScaler
5. Formula: x= (x-mean(x))/sd(x)

Standardization:

1. The variables are scaled in such a way that their mean is zero and standard deviation is one.
2. Not bounded to a specific range.
3. Much less affected by outliers.
4. Scikit-Learn provides StandardScaler.
5. Formula: x= (x-min(x))/(max(x)-min(x))

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans- Reasons for infinite VIF could be perfect collinearity.

Perfect collinearity means that one predictor can be expressed as a linear combination of other predictors.

For example, if you have two identical columns or one column that is a constant multiple of another, the VIF will be infinite.

We can overcome this issue by removing the redundant variables, and by combining few variables like a derived metrics based on the domain knowledge.

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

Ans- A Q-Q plot, short for quantile-quantile plot, is a graphical tool used to assess the conformity between an empirical distribution (observed data) and a theoretical distribution (usually a known distribution, such as the normal distribution).

The primary purpose of a Q-Q plot is to visually compare how well the observed data aligns with the expected distribution.

Importance in Linear Regression:

Q-Q plots play a crucial role in linear regression for several reasons:

1. Normality of Residuals:

In linear regression, we assume that the residuals (differences between observed and predicted values) follow a normal distribution.

By creating a Q-Q plot of the residuals, we can visually check if this assumption holds.

If the residuals align well with the 45-degree line, it suggests that the errors are approximately normally distributed.

2. Detecting Non-Normality:

If the Q-Q plot deviates significantly from the line, it indicates non-normality of residuals.

Departures from linearity may suggest heavy tails, skewness, or other distributional issues.

3. Model Validity:

A linear regression model's validity relies on the normality assumption.

If the residuals are not normally distributed, confidence intervals, hypothesis tests, and other statistical inferences may be affected.