Assignment-based Subjective Questions

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

- ***Below are the potential independent variables which affect the dependent variable (cnt) and the booking of the bike changes according to these fields.***

•Windspeed

- Holiday

•Year 2019

- Weathersit
- Season
- ***The summary shows of R2 0.75 - suggests that your model is explaining a significant portion of the variance in the target variable.***
- ***None of the p value is >0.05 so the model suggests that the observed results are unlikely to have occurred by random chance alone, and we can reject the null hypothesis***

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

   When creating dummy variables, especially for categorical variables with more than two categories, the *drop_first=True* ***parameter is used in order to prevent multicollinearity and to simplify the interpretation of the regression coefficients***

   By excluding one dummy variable it reduces the complexity of the regression model.

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

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

Model Fit Statistics: ***R-squared on the train set provides insight into how well the model fits the data. In our case the R squared does matches the test case R squared which uses as a validation criteria to test the linear regression in the test case***

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

   1. ***year2019: This variable has the highest absolute t-statistic of 25.992, indicating strong statistical significance.***

2. ***fall: The variable "fall" has an absolute t-statistic of 23.975, also indicating strong statistical significance.***
3. ***summer: The variable "summer" has an absolute t-statistic of 20.368, making it the third most significant contributor.***

General Subjective Questions

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

This is a supervised machine learning algorithm which does analyses the train model and do a prediction on test model by finding best fitted the suitable line.

This algorithm helps in creating best-fitting line that minimizes the difference between the predicted and actual values which known as residual analysis

Equation to thesimple linear algorithm y= $\beta_0+\beta_1 \cdot x+\varepsilon$

Y=Predicted/targeted variable

x-slope

e-constant

multiple linear - $y=\beta_0+\beta_1 \cdot x_1+\beta_2 \cdot x_2+...+\beta_p \cdot x_p+\varepsilon$

model evaluation-R2- It measures the proportion of the variance in the target variable that can be explained by the predictors. Higher R2 indicates a better fit.

Assumption-

- ***Error terms are normally distributed***
- ***They are independent of each other***
- ***They are centered around 0 mean area***

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

is a set of four small datasets that have nearly identical simple descriptive statistics, yet appear very different when graphed.

Despite having similar summary statistics, the datasets have significantly different properties when plotted.

Dataset-1

This dataset exhibits a linear relationship between x and y

Dataset-2

This dataset appears to follow a non-linear, but strong, relationship

Dataset-3

This dataset appears to follow a linear relationship except for one extreme outlier.

Dataset-4

This dataset has no apparent relationship between x and y

Summary-
summary statistics for each dataset are nearly identical
Mean and variance of x and y
Correlation coefficient between x and y

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

It is a statistic that measures the linear relationship between two continuous variables. It quantifies the strength and direction of a linear association between two sets of data. Pearson's R is a widely used statistical measure in various fields, including statistics, science, social science, and economics. The Pearson's R value ranges between -1 and 1,

- ***If R is positive, it suggests a positive linear relationship,***
- ***If R is negative, it suggests a negative linear relationship,***

Assumptions-relationship between the two variables is linear, and it is sensitive to outliers.

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

Scaling in the context of data preprocessing refers to the process of transforming the values of variables to a specific range or distribution. So it would be easy to draw inference on a normalized data set .

Min-max- Also known as min-max scaling or min-max normalization, rescaling is the simplest method and consists in rescaling the range of features to scale the range in [0, 1] or [−1, 1]. Selecting the target range depends on the nature of the data.

standardized - Standardization entails scaling data to fit a standard normal distribution. A standard normal distribution is defined as a distribution with a mean of 0 and a standard deviation of 1.

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

Variance inflation factor- It does measure multicollinearity among independent variables. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated with each other.. Which also makes the VIF to go on a infinite scale

(3 marks)

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

It is a graphical tool used in statistics to assess whether a dataset follows a particular theoretical distribution, such as the normal distribution. Each point in the Q-Q plot represents how close the dataset's quantiles are to the expected quantiles of the theoretical distribution.

(3 marks)