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: - Based on the analysis of the categorical variables from the dataset, we can make the following inferences about their potential effects on the dependent variable, which is "cnt" (the count of bike rentals):

- **Season**: The negative correlation coefficient of -55.809 between "season" and "cnt" is unusual and suggests that there may be a data issue or a strong outlier. In general, we would expect "season" to have some influence on bike rentals. For example, bike rentals might increase during the summer and fall seasons when the weather is pleasant, and people are more likely to ride bikes. However, further investigation is needed to resolve the unexpected correlation value.
- **Year (Yr):** The moderate positive correlation coefficient of 0.57 between "year" and "cnt" suggests that there is a relationship between the year and bike rentals. This might indicate that bike rentals increased from 2018 to 2019, which could be due to various factors like increased popularity or improved bike-sharing infrastructure.
- **Month (Mnth)**: The negative correlation coefficient of -0.3431 between "mnth" and "cnt" suggests a moderate negative relationship between the month and bike rentals. This implies that bike rentals tend to be lower in certain months compared to others. It's possible that factors like weather conditions, holidays, or seasonality influence this relationship.
- **Weekday Day:** The very weak negative correlation coefficient of -0.0282 between "weekday day" and "cnt" suggests that the day of the week has a minimal impact on bike rentals. In other words, variations in bike rentals across different weekdays are not significant based on this analysis.
- **Holiday:** A correlation coefficient of 0.00030 for the "holiday" variable suggests a very weak or negligible relationship between the "holiday" variable and bike rentals ("cnt"). In other words, whether a day is a holiday or not does not appear to have a substantial impact on bike rental counts in this dataset.
- **Workingday:** A correlation coefficient of 0.048 for the "workingday" variable suggests a very weak positive relationship between "workingday" and bike rentals ("cnt"). This indicates that there is a slight tendency for bike rentals to be slightly higher on days classified as "workingdays" (where "workingday" equals 1) compared to non-working days (weekends and holidays) when "workingday" equals 0.
- **Weathersit:** A correlation coefficient of 0.07 for the "weathersit" variable suggests a very weak positive relationship between weather conditions ("weathersit") and bike rentals ("cnt"). This indicates that there may be a slight tendency for bike rentals to be slightly higher on days with better weather conditions (lower "weathersit" values) compared to days with worse weather conditions (higher "weathersit" values).

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

Ans: - using **drop_first=True w**hen creating dummy variables is a common practice to avoid multicollinearity, enhance the interpretability of regression models, and reduce dimensionality. It simplifies the model without losing important information about the categorical variable's effect on the dependent variable.

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

Ans: - In a pair plot among numerical variables, you can identify the variable with the highest correlation with the target variable (in this case, "cnt") by looking at the scatterplots and observing which variable's scatterplot has the most apparent linear relationship with "cnt." The variable that shows the steepest or most consistent linear trend in its scatterplot with "cnt" is likely to have the highest correlation.

From the correlation coefficients you provided earlier:

- Temp (Temperature) has a correlation of 0.63 with "cnt."
- Atemp (Adjusted Temperature) also has a correlation of 0.63 with "cnt."
- Hum (Humidity) has a correlation of -0.099 with "cnt."
- Windspeed has a correlation of -0.24 with "cnt."

Based on these correlation coefficients, both "Temp" and "Atemp" have the highest correlation with "cnt," with a value of 0.63. Therefore, both "Temp" and "Atemp" are strong candidates for having the highest correlation with the target variable "cnt."

To confirm which one has the highest correlation, you may want to calculate the exact correlation coefficients between "cnt" and "Temp" and between "cnt" and "Atemp" and compare the numerical values.

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

Ans: - Validating the assumption of Linear Regression Model:

- Linear Relationship: in notebook step 7 is Model Validation. The first step of Model validation is that linear relationship plot represents the relationship between the model and the predictor variables. As we can see, linearity is well preserved.
- Homoscedasticity: All the predictor variables have VIF value less than 5. So, we can consider that there is insignificant multicollinearity among the predictor variables.
- Independence of residuals: Autocorrelation refers to the fact that observations' errors are correlated. To verify that the observations are not auto-correlated, we can use the Durbin-Watson test. The test will output values between 0 and 4. The closer it is to 2, the less autocorrelation there is between the various variables. On this dataset based on The Durbin-Watson value for Final Model lr10 is 1.9244 so There is almost no autocorrelation.
- Normality of Errors: Based on the histogram, we can conclude that error terms are following a normal distribution.

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

Ans: - Based on the final model here the top 3 feature contributing significantly towards explaining the demand of the shared bikes.

- Temperature (Temp) A coefficient value of '0.4283' indicated that a temperature has significant impact on bike rentals

- windspeed A coefficient value of '-0.2054' indicated that the light snow and rain deters people from renting out bikes
- Year (yr) A coefficient value of '0.2413' indicated that a year wise the rental numbers are increasing

It is recommended to give utmost importance to these three variables while planning to achieve maximum bike rental booking. As high temperature and good weather positively impacts bike rentals, it is recommended that bike availability and promotions to be increased during summer months to further increase bike rentals.

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Linear regression is a supervised machine learning algorithm used for predicting a continuous target variable (also called the dependent variable) based on one or more input features (independent variables). It assumes a linear relationship between the input features and the target variable, meaning it tries to find the best-fit straight line that describes this relationship.

Here's a detailed explanation of the linear regression algorithm:

1. Basic Assumptions:

- Linearity: Linear regression assumes that there is a linear relationship between the independent variables (features) and the dependent variable (target). This means that changes in the features have a constant effect on the target.
- Independence: It assumes that the errors (the differences between the actual and predicted values) are independent of each other.
- Homoscedasticity: The variance of the errors should be constant across all levels of the independent variables. This means that the spread of errors should be roughly the same for all values of the features.
- Normality: Linear regression assumes that the errors are normally distributed. This is important for making statistical inferences and hypothesis testing.

2. Simple Linear Regression vs. Multiple Linear Regression:

- In simple linear regression, there is only one independent variable, whereas in multiple linear regression, there are multiple independent variables. The general form of a simple linear regression equation is:

$$Y = b_0 + b_1 * X + \epsilon$$

where:

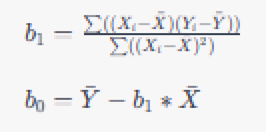
- Y is the dependent variable.
- X is the independent variable.
- b_0 is the intercept (the value of Y when X is 0).
- b_1 is the slope (the change in Y for a one-unit change in X).
- ϵ is the error term.

3. Objective of Linear Regression:

- The goal of linear regression is to find the values of b_0 and b_1 that minimize the sum of the squared differences between the actual target values and the predicted values. This is known as the Least Squares method.

4. Training the Model:

- To find the values of b_0 and b_1, the algorithm uses a training dataset consisting of known values of the independent and dependent variables.
- It calculates the mean of X and Y and then computes b_1 (the slope) and b_0 (the intercept) using the following formulas:



$$b_1 = \frac{\sum((X_i - \bar{X})(Y_i - \bar{Y}))}{\sum((X_i - \bar{X})^2)}$$

$$b_0 = \bar{Y} - b_1 * \bar{X}$$

where $\bar{X}$ is the mean of X and $\bar{Y}$ is the mean of Y.

5. Making Predictions:

- Once the model is trained, it can be used to make predictions on new data.
- Given a new value of X, you can plug it into the regression equation to predict the corresponding value of Y.

6. Model Evaluation:

- Common metrics for evaluating the performance of a linear regression model include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2) which measures the goodness of fit.

7. Assumptions Checking and Model Interpretation:

- It's important to validate the assumptions of linear regression and check if they hold true for the data. If not, it might be necessary to consider alternative models.
- Linear regression also allows for interpretation of the relationship between independent and dependent variables through the coefficients (b_0 and b_1).

8. Regularization (Optional):

- In some cases, regularization techniques like Ridge or Lasso regression are applied to prevent overfitting and improve the model's generalization.

Q2. Explain the Anscombe’s quartet in detail.

Ans: Anscombe's quartet is a famous dataset in statistics that consists of four sets of data points. What makes Anscombe's quartet particularly interesting is that these four datasets have nearly identical simple descriptive statistics, including means, variances, and correlation coefficients, but they are visually and behaviorally quite different. It was created by the statistician Francis Anscombe in 1973 to emphasize the importance of visualizing data and not relying solely on summary statistics. The quartet serves as a compelling example of why data visualization is crucial in understanding and interpreting data.

Here are the details of Anscombe's quartet:

Dataset 1:

- x-values: [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]
- y-values: [8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68]

Dataset 2:

- x-values: [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]
- y-values: [9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74]

Dataset 3:

- x-values: [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]
- y-values: [7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73]

Dataset 4:

- x-values: [8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8]
- y-values: [6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89]

Now, let's discuss the key observations from Anscombe's quartet:

- Descriptive Statistics: When you calculate summary statistics for each of the four datasets (mean, variance, correlation), you will find that they are nearly identical for all of them. For example, the means and variances of x and y are very close in all four cases.
- Visual Differences: Despite the similar summary statistics, when you plot these datasets, you'll notice significant visual differences. Each dataset has a unique pattern. Some are linear, some are curved, and some have outliers.
- Implication: Anscombe's quartet illustrates that relying solely on summary statistics can be misleading. Even though the statistical properties of these datasets are alike, the underlying data distribution and relationships between variables can be fundamentally different.
- Importance of Data Visualization: The quartet underscores the importance of data visualization in data analysis. By visualizing data, you can uncover patterns, outliers, and relationships that might not be apparent from summary statistics alone. It also serves as a reminder that graphical exploration should precede statistical analysis.

Q3. What is Pearson's R?

- Pearson's correlation coefficient, often denoted as "r" or "Pearson's r," is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It was developed by Karl Pearson in the early 20th century and is widely used in statistics to assess the degree of association between two variables.

Pearson's correlation coefficient can take values between -1 and 1, where:

- r = 1: Indicates a perfect positive linear relationship. As one variable increases, the other also increases, and the relationship is perfectly linear.
- r = -1: Indicates a perfect negative linear relationship. As one variable increases, the other decreases, and the relationship is perfectly linear but in the opposite direction.
- r = 0: Indicates no linear relationship between the variables. There is no linear pattern between the variables.

The formula for calculating Pearson's correlation coefficient "r" between two variables, X and Y, with n data points, is as follows:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Where:

- Xi and Yi are individual data points.
- X¯ and Y¯ are the means of X and Y, respectively.

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

Ans: Scaling in the context of data preprocessing refers to the process of transforming the numerical values of features (variables) in a dataset to a specific range or distribution. The goal of scaling is to make the data more suitable for machine learning algorithms and statistical analysis. Scaling is performed to ensure that the features contribute equally to model training and prevent certain features from dominating others due to differences in their scales. There are two common types of scaling: normalized scaling and standardized scaling.

1. Normalized Scaling (Min-Max Scaling):

Normalized scaling, also known as Min-Max scaling, transforms the values of a feature into a specific range, typically between 0 and 1. It is done using the following formula for each feature:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Where:

- X is the original feature value.
- Xmin is the minimum value of the feature in the dataset.
- Xmax is the maximum value of the feature in the dataset.

2. Standardized Scaling (Z-Score Scaling):

Standardized scaling, also known as Z-score scaling or zero-mean scaling, transforms the values of a feature to have a mean (average) of 0 and a standard deviation of 1. It is done using the following formula for each feature:

$$x' = \frac{x - \bar{x}}{\sigma}$$

Where:

- X is the original feature value.
- μ is the mean of the feature in the dataset.
- σ is the standard deviation of the feature in the dataset.

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

Ans: The Variance Inflation Factor (VIF) is a measure used in regression analysis to assess multicollinearity among predictor variables. It quantifies how much the variance of the estimated coefficients in a regression model is increased due to the presence of multicollinearity. Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated with each other.

The formula for calculating the VIF for a specific predictor variable in a multiple regression model is as follows:

$$\text{VIF} = \frac{1}{1 - R_i^2}$$

Where:

- VIFi is the VIF for the i-th predictor variable.
- Ri2 is the coefficient of determination (R-squared) when the i-th predictor is regressed against all the other predictor variables in the model.

Q6. 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 in statistics and data analysis to assess whether a dataset follows a particular theoretical distribution, such as the normal distribution. It helps you visually compare the quantiles of your dataset (empirical quantiles) against the quantiles of a theoretical distribution (expected quantiles). The primary purpose of a Q-Q plot is to check if the data deviates from the expected distribution and identify departures from normality or other distributions.

Here's how a Q-Q plot works and its importance in the context of linear regression:

How to Create a Q-Q Plot:

1. Sort the Data: First, you sort your dataset in ascending order.
2. Calculate Expected Quantiles: Calculate the expected quantiles for your dataset based on the theoretical distribution you are interested in. For example, if you want to check for normality, you'd calculate the quantiles that correspond to a standard normal distribution (mean = 0, standard deviation = 1).
3. Plot the Data: On the Q-Q plot, you place the expected quantiles on the x-axis and the corresponding quantiles from your dataset on the y-axis. Each point on the plot represents a quantile pair.

Interpreting a Q-Q Plot:

- If the data closely follows the theoretical distribution, the points on the Q-Q plot will lie along a straight line (usually a 45-degree diagonal line) from the bottom left to the top right. This indicates a good fit to the theoretical distribution.
- Departures from the straight line suggest deviations from the expected distribution. For example:
    - Points curving upward indicate that the data has heavier tails than the expected distribution.
    - Points curving downward suggest that the data has lighter tails than the expected distribution.
    - S-shaped deviations suggest non-linear departures.

Importance of Q-Q Plot in Linear Regression:

In the context of linear regression, Q-Q plots are important for several reasons:

- Assumption Checking: Linear regression models often assume that the residuals (the differences between observed and predicted values) are normally distributed. A Q-Q plot of the residuals helps assess whether this assumption holds. If the points on the Q-Q plot deviate significantly from the straight line, it indicates that the residuals may not be normally distributed.
- Model Validation: Q-Q plots are useful in model validation. They help you check whether the errors (residuals) from your regression model are normally distributed. Non-normality of residuals can affect the validity of statistical tests and confidence intervals associated with the model.

- Outlier Detection: Q-Q plots can reveal the presence of outliers in the dataset. Outliers often deviate from the expected distribution and can be spotted as points far from the straight line in the Q-Q plot.
- Data Transformation: If the Q-Q plot suggests that the data deviates significantly from the expected distribution, it may be necessary to consider data transformations or robust regression techniques to address the departure from normality.