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?**

Ans- There are some categorical variables such as season, month ,year, weathersit, weekday and working day. These variables have a major effect on dependent variable ‘cnt’.

- Fall has the highest mean and it can be expected as it is best season to take rental bike followed by summer.
- People tend to rent more bikes on non-holidays compared to holidays because possibly the have some other plans on holidays.
- Median bike rentals are increasing over the years, with 2019 having a higher median than 2018. This could be due to the growing popularity of bike rentals and increased environmental awareness.
- Clear weather is most optimal for bike renting, as the temperature is optimal, humidity is less, and the temperature is less.
- People prefer rental bikes more in the month of September and October.

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

Ans: - It is important to use drop_first = True during dummy variable creation because by doing this, one one level of each categorical variable is automatically dropped, which serves as a reference level. This means that the dropped level is represented implicitly by the absence of all other levels. This helps to prevent multicollinearity and redundancy because the dropped level can be inferred from the other levels. A variable with n levels can be represented by n-1 dummy variable .

For eg. If a variable has 4 levels so the drop_first will drop one column so the remaining three column will represent all the 4 levels.

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

Ans: temp and atemp have highest correlation with the target variable ‘cnt’ as compared to other variables.

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

Ans: The assumptions of Linear Regression are validated based on the normality of error with mean 0, multicollinearity and homoscedasticity .

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

Ans: The top 3 features contributing significantly towards explaining the demand on the shared bikes are temp, windspeed and year.

General Subjective Questions

1. **Explain the linear regression algorithm in detail**

Ans: Linear regression is a fundamental supervised learning algorithm used for predictive analysis which tells us the relationship between a dependent variable (target) and one or more independent variables (predictors) by fitting a linear equation to the observed data. The goal is to find the best-fitting line (or hyperplane) that minimizes the sum of squared differences between the observed values and the predicted values by the model.

The detailed algorithms are:

1. Assumption:
    - Linearity: The relationship between the independent and dependent variables is linear.
    - Independence: The observations are independent of each other.
    - Homoscedasticity: The variance of the errors is constant across all levels of the independent variables.
    - Normality: The errors are normally distributed with a mean zero.
2. Model: The linear regression model is represented by the equation:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon$$

Where y is the dependent variable,

$X_1, X_2, ... X_n$ are the independent variables,

β_0 is the intercept (the value of Y when X = 0),

$\beta_1, \beta_2, ... \beta_n$ are the coefficients of $X_1, X_2, ... X_n$

3 . Error Term: The difference between the predicted values and the actual values is called the error or residual. The goal is to minimize the sum of squared errors (SSE) or the mean squared error (MSE) to find the best-fit line.

4. Parameter Estimation: - The parameters β□, β□, β□, ..., βn are estimated using a technique called ordinary least squares (OLS). OLS finds the values of these parameters that minimize the sum of squared errors.

5. Model Evaluation: - R-squared (coefficient of determination): It measures the proportion of the variance in the dependent variable that can be explained by the independent variables. - Mean squared error (MSE): It measures the average squared difference between the predicted and actual values. - Root mean squared error (RMSE): It is the square root of MSE and provides a measure in the same units as the dependent variable.

6. Making Predictions: - After the model is trained and evaluated, it can be used to make predictions on new or unseen data. Given the values of the independent variables, the model predicts the corresponding value of the dependent variable using the equation: Y = β□ + β□X□ + β□X□ + ... + βnXn.

7. Extensions and Variations: - Linear regression can be extended to handle nonlinear relationships by including polynomial terms or transforming the independent variables. - Regularized regression techniques like ridge regression and lasso regression can be used to prevent overfitting and improve model performance.

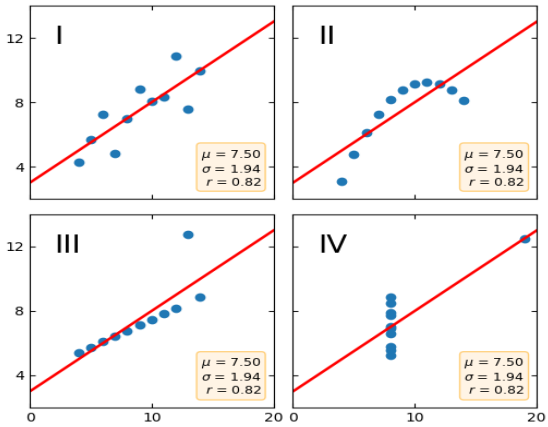
2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet consist of four data sets that have the same descriptive values such as mean, variance ,R-squared, correlation and linear regression but having different representations when we scatter plots on a graph.

The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

The property of all the four data sets are-

i. The mean of x is 9 for all the dataset
ii. The sample variance of x is 11 for all the dataset.
iii. The mean of y is 7.50 for all the dataset.
iv. The sample variance of x is 4.125 for all the dataset.
v. Correlation coefficient between x and y is 0.816, which shows there is strong linear relationship between x and y.



Observation:

- Data-set I — consists of a set of (x,y) points that represent a linear relationship with some variance.
- Data-set II — shows a curve shape but doesn't show a linear relationship .
- Data-set III — looks like a tight linear relationship between *x* and *y*, except for one large outlier.
- Data-set IV — looks like the value of *x* remains constant, except for one outlier as well.

Data-sets which are identical over a number of statistical properties, yet produce dissimilar graphs, are frequently used to illustrate the importance of graphical representations when exploring data.

3. What is Pearson's R?

Ans: Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It assesses how closely the data

points in a scatter plot follow a straight line pattern. The Pearson correlation coefficient, denoted as r, ranges from -1 to +1. The value of r indicates the following:

- r = +1: A perfect positive linear relationship, where all data points lie exactly on a straight line with a positive slope.
- r = -1: A perfect negative linear relationship, where all data points lie exactly on a straight line with a negative slope.
- r ≈ 0: No linear relationship, implying that the data points are not well described by a straight line. However, it does not necessarily mean there is no relationship at all. Non-linear relationships or other forms of association might exist.

Pearson's R is calculated as the covariance of the two variables divided by the product of their standard deviations. It is sensitive to outliers and assumes that the relationship between the variables is linear.

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

Ans: Scaling is a step of data pre-processing which is applied to independent variables to normalize the data within a particular range, basically in between 0 and 1.

It is of two types: Minmax Scaler and Standaradization

Scaling is crucial in preprocessing data for machine learning models, especially when features have different magnitudes, units, or ranges. By scaling the features, you bring them to a similar scale, ensuring that the algorithm doesn't prioritize one feature over another based solely on its magnitude. Scaling just

affect the coefficients and none of the other parameters like t-statistic, F-statistics, p-value, R-squared.

Difference between Normalizing scaling and standardized scaling:

- In normalized scaling, minimum and maximum values of features being used for scaling whereas in standardize scaling, mean and standard deviation is used for scaling.
- Normalization scales the data to a fixed range (e.g., [0, 1]), while standardization centres the data around 0 with a standard deviation of 1.
- Standardization is less sensitive to outliers compared to normalization, as it uses the mean and standard deviation which are less affected by extreme values.
- Standardized data is more easily interpretable since it represents the number of standard deviations away from the mean.
- Normalization preserves the shape of the original distribution but scales it to a smaller range, while standardization preserves the shape but shifts it to have a mean of 0.

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

Ans: VIF stands for Variance Inflation Factor, which is a measure used to assess multicollinearity in regression analysis. It quantifies how much the variance of the estimated regression coefficients is increased due to collinearity among the predictor variables. The VIF for a particular predictor variable is calculated by regressing that variable against all the other predictor variables and obtaining the R-squared value.

The formula for VIF is:

$$VIF = 1 / (1 – R2)$$

Yes, a VIF value of infinity can occur in situations where there is a perfect correlation between two independent variables or we can say perfect multicollinearity exist between predictors variables in a regression model. Perfect multicollinearity means that one or more predictors in a regression model can be exactly predicted from others using a linear combination of those predictors. When this happens, the VIF for the predictor(s) involved becomes infinity because the estimated coefficient for that predictor is not unique and can take and can take on any value.

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 (Quantile-Quantile plot) is a graphical tool used to assess the distributional similarity between a sample of data and a theoretical distribution. It is commonly employed in statistical analysis, including linear regression, to evaluate whether the assumption of normality holds for a dataset. Here's an explanation of the use and importance of a Q-Q plot in linear regression:

1. Assessing Normality: - Linear regression often assumes that the residuals (the differences between observed and predicted values) follow a normal distribution.

- A Q-Q plot allows us to visually compare the quantiles of the residuals (or any variable of interest) against the quantiles of a theoretical normal distribution.

- If the points on the Q-Q plot closely follow a straight line, it suggests that the residuals are normally distributed.

2. Identifying Departures from Normality:

- If the points on the Q-Q plot deviate significantly from a straight line, it indicates departures from normality.

– If the points curve upward or downward, it suggests heavy-tailed or skewed distributions, respectively.

2. Identifying Departures from Normality:

   - If the points on the Q-Q plot deviate significantly from a straight line, it indicates departures from normality. –

   If the points curve upward or downward, it suggests heavy-tailed or skewed distributions, respectively. –

   Outliers or extreme values in the data may cause deviations from the expected straight line pattern.

3. Model Assumption Validation:

   - A Q-Q plot helps to validate one of the key assumptions of linear regression, namely the normality of residuals.

   - Violations of normality can affect the validity of statistical inference, hypothesis tests, and confidence intervals associated with the regression model.

   - By visually inspecting the Q-Q plot, we can determine if the assumption of normality holds or if additional steps, such as data transformations, are needed.

   4. Diagnostic Tool:

   - Q-Q plots are diagnostic tools that provide insights into the distributional characteristics of the data.

   - In linear regression, examining the Q-Q plot of the residuals can help identify potential issues, such as heteroscedasticity (unequal variances) or nonlinearity.

   5. Model Improvement:

   - If deviations from normality are observed in the Q-Q plot, it indicates areas for potential model improvement.

   - Data transformations or the use of robust regression techniques may be considered to address non-normality in the residual.