Assignment Based Questions Answers

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

Ans- Count('cnt') is highly (positively) correlated with 'casual' and 'registered' and further it is high with 'atemp'. We can clearly understand the high positive correlation of count with 'registered' and 'casual' as both of them together add up to represent count.

- Count is negatively correlated to 'windspeed' (-0.24 approximately). This gives us an impression that the shared bikes demand will be somewhat less on windy days as compared to normal days.
- correlation with "humidity" and "cnt" is almost negligible (- 0.09).

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

Ans- It is important to use **drop_first = True** while creating the dummy variables. It helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. Hence if we have categorical variables with N distinct values, then we need only N-1 columns to represent the dummy variables.

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

Ans- 'casual' seems highest correlated to ‘`cnt’` which is 0.95 the most, then 'casual' which is 0.67 and ‘temp’ as 0.64 .

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

Ans- After getting OLS regression results we can see Standard Errors assume that the covariance matrix of the errors is correctly specified. The model seems to be doing a good job. The new model built on the selected features doesn't show much dip in the accuracy in comparison to the model which was built on all the features. It has gone from **84.5%** to **84.4%**. All the VIF values and p-values seem to be in the permissible range now. Also the `Adjusted R-squared` value has dropped from `84.5%` with **28 variables** to just `79.3%` using **7 variables**. This model is explaining most of the variance without being too complex.

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

Ans- We can conclude from our model that below three features are the most influential features for BIke Rentals:

- Temperature : with coefficient `0.42`
- Weather C [3-Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds]: with coefficient `0.24`
- Year: with coefficient `0.23`

<u>General Subjective Questions Answers</u>

Q1. Explain Linear regression algorithm in detail

Ans. Linear regression is a fundamental statistical and machine learning algorithm used for modeling the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables. In simple linear regression, there is one independent variable, while in multiple linear regression, there are multiple independent variables.

Simple Linear Regression:

1. Model Representation:

The relationship between the independent variable x and the dependent variable y is represented as:

$y=\beta0+\beta1x+\varepsilon$

Where: *y is the dependent variable (what we want to predict).*

x is the independent variable (what we use to make predictions).

β0 is the intercept (the value of y when x is zero).

β1 is the slope (the change in y for a unit change in x).

ε is the error term, representing the difference between the observed and predicted values.

2. **Objective:**

The goal of linear regression is to find the best-fitting line (or hyperplane in multiple linear regression) that minimises the sum of squared differences between the observed and predicted values. This is known as minimising the cost function or loss function.

3. Cost Function:

The most commonly used cost function in linear regression is the Mean Squared Error (MSE):

$MSE=n1\sum i=1n(yi-y\hat{}i)2$

Where: *n is the number of data points. y^i is the predicted value of the dependent variable. yi is the actual value of the dependent variable.*

4. Parameter Estimation (Ordinary Least Squares):

The coefficients β0 and β1 are estimated using the method of Ordinary Least Squares

(OLS). This involves finding the values of β0 and β1 that minimize the MSE.

5. Fitting the Model:
This involves finding the values of β0 and β1 that minimize the MSE. This can be done analytically using mathematical techniques like matrix algebra.

6. Making Predictions:
Once the model is trained, you can use it to make predictions on new or unseen data.
Multiple Linear Regression:
In multiple linear regression, the model is extended to handle multiple independent variables:

$y=\beta0+\beta1x1+\beta2x2+\ldots+\beta nxn+\varepsilon$

Where x1,x2,…,xn are the independent variables, and β1,β2,…,βn are their respective coefficients. The cost function, parameter estimation, and model fitting are extended to accommodate multiple variables.
Assumptions of Linear Regression:
Linearity: The relationship between the independent and dependent variables is linear.

Independence of Errors: The errors (ε) are independent of each other.

Homoscedasticity: The variance of errors is constant across all levels of the independent variable(s).
Normality of Errors: The errors are normally distributed.
No Multicollinearity: The independent variables are not highly correlated with each other.
Evaluation of Linear Regression Models:
Common metrics for evaluating linear regression models include R-squared (R2), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).
Remember, linear regression is a powerful tool, but it makes certain assumptions about the data. It's important to validate these assumptions and, if necessary, consider other modeling techniques if the assumptions are not met.

Q2. Explain the Anscombe's quartet in detail.
Ans- Anscombe's quartet is a set of four small datasets that have nearly identical simple descriptive statistics, yet they have very different distributions and appear quite different when graphed. This set of datasets was created by the statistician Francis Anscombe in 1973 to emphasise the importance of graphing data before analysing it and to demonstrate the effect of outliers on statistical properties.

Here are the details of the four datasets in Anscombe's quartet:

Dataset I:
- **Descriptive Statistics**:
  - Mean of \(x\): 9.0
  - Mean of \(y\): 7.5
  - Linear Regression Line: \(y = 3.0 + 0.5x\)

  - Correlation Coefficient: 0.816

- **Distribution**:
  - The data points roughly follow a linear pattern.
  - There is a clear relationship between \(x\) and \(y\).

- **Graphical Representation**:
  - When plotted, it forms a fairly straight line with some minor deviation.

Dataset II:
- **Descriptive Statistics**:
  - Mean of \(x\): 9.0
  - Mean of \(y\): 7.5
  - Linear Regression Line: \(y = 3.0 + 0.5x\)
  - Correlation Coefficient: 0.816

- **Distribution**:
  - Unlike Dataset I, the points form a clear non-linear pattern (more of a quadratic shape).

- **Graphical Representation**:
  - When plotted, it resembles a quadratic curve.

Dataset III:
- **Descriptive Statistics**:
  - Mean of \(x\): 9.0
  - Mean of \(y\): 7.5
  - Linear Regression Line: \(y = 3.0 + 0.5x\)
  - Correlation Coefficient: 0.816

- **Distribution**:
  - The data points are mostly in a straight line, except for one outlier.

- **Graphical Representation**:
  - When plotted, it forms a line with an outlier.

Dataset IV:
- **Descriptive Statistics**:
  - Mean of \(x\): 9.0
  - Mean of \(y\): 7.5
  - Linear Regression Line: \(y = 3.0 + 0.5x\)
  - Correlation Coefficient: 0.817

- **Distribution**:
  - The data points are separated into two distinct clusters.

- **Graphical Representation**:
  - When plotted, it shows two clusters of data points.

Key Points:

1. Despite having identical summary statistics, these datasets are drastically different when plotted.

2. This illustrates the importance of visualising data before drawing conclusions.

3. It also highlights the influence that outliers can have on statistical properties.

4. It emphasises that numerical summary statistics alone may not capture the complete story of the data.

In summary, Anscombe's quartet serves as a powerful reminder that visual exploration of data is crucial in understanding its underlying patterns and making meaningful interpretations.

Q3. What is Pearson’s R?

Ans- Pearson's r, often referred to simply as the correlation coefficient or Pearson correlation coefficient, is a
statistic that measures the strength and direction of a linear relationship between two continuous variables.

In the simple linear regression the R - squared will get its square of R or Pearson’s R. Suppose, for example

```
Corrs = np.Corrcoef(x_train,y_train)
Corrs = 0.9321
```

Then R - squared will be the square of Corrs.
R - squared = 0.81

Here are some key points about Pearson's r:

Range: The value of r ranges from -1 to 1
Sensitive to Outliers: It can be sensitive to outliers, meaning that extreme data points can have a notable impact on the correlation coefficient.
Assumption: Pearson's r assumes a linear relationship between the variables. If the relationship is non-linear, Pearson's r may not accurately reflect the association.
Sensitive to Outliers: It can be sensitive to outliers, meaning that extreme data points can have a notable impact on the correlation coefficient.
No Causation: Correlation does not imply causation. Even if two variables are highly correlated, it does not mean that changes in one variable cause changes in the other.
Calculation: The formula for Pearson's r involves calculating the covariance between the two variables and dividing it by the product of their standard deviations.

$r = (\Sigma((X - \bar{X})(Y - \bar{Y}))) / (n * \sigma X * \sigma Y)$

Where, X and Y are individual data points, $\bar{X}$ and $\bar{Y}$ are the means of X and Y, n is the number of data points, σX and σY are the standard deviations of X and Y

Pearson's correlation coefficient is widely used in statistics and research to assess the strength and direction of relationships between variables. However, it's important to remember that correlation does **not imply causation, and other statistical methods may be needed to establish causation.**

Q4. What is scaling? Why is scaling performed? What is the difference between normalised scaling and standardised scaling?
Ans- **Scaling** is a preprocessing step in data preparation where you transform the features of your dataset so that they have similar scales or distributions. This is important because many machine learning algorithms are sensitive to the scale of the input features.

Here are the key reasons why scaling is performed:

1. **Equalizes the Impact of Features**: Scaling ensures that all features contribute equally to the computations in the model. Without scaling, features with larger magnitudes can dominate the learning process.

2. **Improves Convergence**: Algorithms that rely on numerical optimization techniques (like gradient descent) converge faster when features are on a similar scale. This means the algorithm reaches an optimal solution more quickly.

3. **Facilitates Interpretation**: When features are on similar scales, it becomes easier to interpret the importance of each feature in the model.

4. **Regularization Methods**: Some regularization methods (e.g., L1 and L2 regularization) assume that all features are on a similar scale. Scaling ensures that regularization is applied fairly to all features.

Normalized Scaling (Min-Max Scaling):

- In normalized scaling, also known as min-max scaling, the values of the features are scaled to a fixed range, usually [0, 1].

- The formula for min-max scaling is:

\[X_{\text{normalized}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}\]

- This ensures that the minimum value of the feature is mapped to 0, and the maximum value is mapped to 1.

Standardized Scaling (Z-score Scaling):

- Standardized scaling, also known as z-score scaling or zero-mean scaling, transforms the features so that they have a mean of 0 and a standard deviation of 1.

- The formula for standardization is:

  \[X_{\text{standardized}} = \frac{X - \mu}{\sigma}\]

  where \(\mu\) is the mean and \(\sigma\) is the standard deviation of the feature.

- Standardized features have a mean of 0 and a standard deviation of 1, which makes them adhere to a standard normal distribution.

Differences:

1. **Range of Values**:
   - Normalized scaling scales features to a fixed range (e.g., [0, 1]).
   - Standardized scaling centers the features around 0 with a standard deviation of 1.

2. **Impact on Outliers**:
   - Normalized scaling is sensitive to outliers because it depends on the range of the data.
   - Standardized scaling is less affected by outliers because it is based on the mean and standard deviation.

3. **Interpretability**:
   - Normalized scaling maintains the original interpretability of the data (i.e., you can still understand the values in terms of the original units).
   - Standardized scaling creates features with a mean of 0 and standard deviation of 1, which might not have a direct interpretable meaning.

In practice, the choice between normalized and standardized scaling depends on the specific requirements of the algorithm and the nature of the data.

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 to assess multicollinearity in a regression analysis. It quantifies how much the variance of an estimated regression coefficient increases if your predictors are correlated.

VIF is calculated for each predictor variable by regressing it against all the other predictor variables in the model. If the VIF value for a particular predictor is too high, it suggests that this predictor may be highly correlated with the other predictors in the model.

A VIF value can become infinite if there is perfect multicollinearity in the model. Perfect multicollinearity occurs when one or more predictor variables can be perfectly predicted from the others. This means that there is a linear relationship among the predictors that can be expressed as an exact mathematical formula.

Here are some common scenarios that lead to infinite VIF:

1. **Duplicate Variables**: If you accidentally include two identical or nearly identical variables in the model, this can lead to perfect multicollinearity. For example, using temperature in both Celsius and Fahrenheit in the same model would cause this issue.

2. **Linear Dependency**: If one or more predictor variables can be expressed as exact linear combinations of the others, this leads to perfect multicollinearity.

3. **Dummy Variable Trap**: If you include dummy variables for all levels of a categorical variable without dropping one level (known as the dummy variable trap), it can lead to perfect multicollinearity.

In cases of infinite VIF, it means that there is a linear dependency among the predictor variables that is exact, which makes it impossible to estimate the individual coefficients. This is a serious issue because it renders the regression model invalid.

To address this problem, you'll need to identify and address the source of multicollinearity. This may involve removing one of the correlated variables, redefining the variables, or using techniques like principal component analysis (PCA) to reduce dimensionality.

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, or quantile-quantile plot, is a graphical tool used to assess the similarity in distributions between two datasets. It is particularly useful in statistics for comparing the distribution of sample data to a theoretical distribution, like the normal distribution.

Here's how a Q-Q plot works:

1. **Sorting**: Both datasets (the sample data and the theoretical distribution) are sorted in ascending order.

2. **Quantiles**: The quantiles (percentiles) of the sample data are plotted against the corresponding quantiles of the theoretical distribution.

3. **Diagonal Line**: A diagonal reference line is added to the plot, which represents perfect alignment between the two distributions.

If the points in the Q-Q plot closely follow the diagonal line, it suggests that the distributions are similar. Deviations from the diagonal line indicate differences in distribution.

Use and Importance in Linear Regression:

1. **Assumption Checking**:
   - In linear regression, it is assumed that the residuals (the differences between the observed and predicted values) are normally distributed. A Q-Q plot of the residuals helps to assess this assumption.

2. **Identifying Departures from Normality**:
   - If the Q-Q plot deviates significantly from the diagonal line, it suggests that the residuals do not follow a normal distribution. This indicates a violation of the assumption of normality.

3. **Detecting Outliers**:
   - Outliers can affect the normality of residuals. A Q-Q plot can help identify if there are extreme values in the data.

4. **Model Validity**:
   - Ensuring that the residuals are normally distributed is crucial for the validity of statistical inferences made from the regression model, such as confidence intervals and hypothesis tests.

5. **Choosing Transformations**:
   - If the Q-Q plot reveals non-normality, it might be necessary to consider data transformations or non-parametric regression techniques.

6. **Comparing Different Models**:
   - When comparing different models, you can use Q-Q plots to evaluate the normality of residuals for each model and choose the one that best satisfies the assumptions.

In summary, the Q-Q plot is a valuable diagnostic tool for assessing the normality assumption in linear regression. It provides a visual indication of whether the residuals follow a normal distribution and helps in identifying potential issues that need to be addressed for a valid regression analysis.