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? (3 marks)**

ANS – From the analysis of categorical variables in the dataset, several inferences can be drawn regarding their effect on the dependent variable (demand for Boombikes):

1. Season:
    - The demand for shared bikes appears to vary across different seasons.
    - Typically, higher demand is observed during seasons like spring and summer, while lower demand may occur during fall and winter.
    - This suggests that seasonal variations play a significant role in influencing bike rental demand, with weather conditions and outdoor activities likely contributing to fluctuations in demand.
2. Year (yr):
    - The year variable indicates the year of data collection, with values representing 2018 and 2019.
    - There may be an increasing trend in bike rental demand over the years due to factors such as the popularity of bike-sharing systems, changes in consumer behavior, and improvements in infrastructure.
    - The year variable may serve as an important predictor of bike rental demand, as it captures temporal trends and shifts in demand patterns over time.
3. Month (mnth):
    - The month variable represents the month of the year, ranging from 1 to 12.
    - Bike rental demand may exhibit seasonal patterns across different months, influenced by factors such as weather conditions, holidays, and cultural events.
    - Certain months, such as those corresponding to summer vacation periods or festive seasons, may experience higher demand compared to others.
4. Holiday:
    - The holiday variable indicates whether the day is a holiday or not.
    - Holidays may have a significant impact on bike rental demand, with increased demand observed during holidays as people engage in leisure activities, sightseeing, or recreational outings.
    - Conversely, lower demand may be expected on non-holiday days, particularly weekdays when individuals are occupied with work or routine activities.
5. Workingday:
    - The workingday variable identifies whether the day is a working day (1) or not (0).

- Bike rental demand may vary depending on whether it is a working day or not, with higher demand typically observed on weekends and non-working days when individuals have more leisure time and opportunities for outdoor activities.
- Conversely, lower demand may occur on working days, particularly during weekdays when people are engaged in work or other commitments.

Overall, categorical variables such as season, year, month, holiday, and working day provide valuable insights into the factors influencing bike rental demand. These variables capture temporal patterns, seasonal variations, and socio-cultural factors that affect the utilization of shared bikes, thereby serving as important predictors in the regression model for predicting bike rental demand.

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

ANS – In this assignment, we are dealing with categorical variables in linear multiple regression models. In multiple linear regression models, it is important to use because:

i. When we create dummy variable from categorical variable, multicollinearity can be avoided by dropping the first level (category). Multicollinearity occurs when two or more predictor variables are highly correlated.
ii. Also, we can reduce the number of dummy variables and complexity of the model by dropping the first level.

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

ANS – The 'registered' variable has the highest correlation with the target variable 'cnt'.

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

ANS - After building the linear regression model on the training set, it is essential to validate the assumptions of linear regression to ensure that the model's predictions are reliable and accurate. The key assumptions of linear regression include:

1. Linearity: The relationship between the independent variables and the dependent variable should be linear.
2. Independence of residuals: The residuals (the differences between observed and predicted values) should be independent of each other.
3. Homoscedasticity: The variance of the residuals should be constant across all levels of the independent variables.
4. Normality of residuals: The residuals should follow a normal distribution.

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

ANS - To identify the top features contributing significantly to explaining the demand for shared bikes based on the final model, we typically rely on the coefficients or importance scores assigned to each feature by the regression model. The features with higher coefficients or importance scores are considered more influential in explaining the variation in the target variable (demand for shared bikes).

Here's how we can determine the top features based on the final model:

1. Coefficient Analysis: If using linear regression, examine the coefficients assigned to each feature in the model. Higher magnitude coefficients indicate stronger associations with the target variable. The top features are those with the highest absolute coefficient values.
2. Feature Importance: If using other models like decision trees, random forests, or gradient boosting, check the feature importance scores provided by the model. These scores represent the contribution of each feature to the model's predictive performance. The top features are those with the highest importance scores.
3. Statistical Significance: Consider the statistical significance of each feature, usually measured by p-values. Features with low p-values (typically below a predefined threshold, e.g., 0.05) are considered statistically significant and contribute significantly to the model.
4. Domain Knowledge: Incorporate domain knowledge and context to interpret the results. Certain features may be intuitively expected to have a strong influence on bike demand based on prior knowledge or understanding of the domain.

Once the top features are identified using one or more of the above methods, they can be ranked accordingly. Typically, the top 3 features contributing significantly to explaining the demand for shared bikes would be those with the highest coefficients, importance scores, or statistical significance.

It's important to note that the specific method used to identify top features may vary depending on the model type, dataset characteristics, and analytical goals. Additionally, interpretation of feature importance should be done cautiously, considering potential biases, confounding factors, and model assumptions.

General Subjective Questions

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

ANS - Linear regression is a statistical technique used for modeling the relationship between a dependant variable and one or more independent variables. The goal of linear regression is to find the best-fitting linear equation that describes the relationship between the independent variables and the dependent variable.

i. Linear Equation: Linear regression assumes that the relationship between the independent variables (denoted as X) and the dependent variable (denoted as Y) can be represented by a linear equation of the form:

$Y=\beta 0+\beta 1*X1+\beta 2*X2+...+\beta n*Xn+\epsilon$

Where:

- Y is the dependent variable.
- $\beta 0$ is the intercept (the value of Y when all independent variables are 0).
- $\beta 1,\beta 2,...,\beta n$ are the coefficients (also called slopes) that represent the change in Y for a one-unit change in the corresponding independent variable.
- $X1,X2,...,Xn$ are the independent variables.
- ϵ is the error term, representing the difference between the observed and predicted values of Y. It accounts for factors not included in the model.

ii. Fitting the Model: The coefficients $\beta 0,\beta 1,...,\beta n$ are estimated from the training data using a method such as Ordinary Least Squares (OLS). The goal is to minimize the sum of the squared differences between the observed values of Y and the values predicted by the linear equation.

iii. Ordinary Least Squares (OLS): In OLS, the model parameters are estimated by minimizing the residual sum of squares (RSS), which is the sum of the squared differences between the observed and predicted values of the dependent variable. This is achieved by calculating the derivatives of the RSS with respect to each parameter and setting them to zero to find the optimal values.

iv. Model Evaluation: Once the model is trained, it is evaluated using various metrics such as R-squared, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), etc., to assess how well it fits the data and makes accurate predictions.

v. Making Predictions: After the model is trained and evaluated, it can be used to make predictions on new or unseen data. Given values of the independent variables, the model predicts the value of the dependent variable using the learned coefficients.

vi. Assumptions of Linear Regression:

- Linearity: The relationship between the independent and dependent variables is linear.
- Independence of Errors: The errors (residuals) should be independent of each other.
- Homoscedasticity: The variance of the errors should be constant across all values of the independent variables.
- The errors are normally distributed around the true regression line.

Overall, linear regression is a versatile and widely used technique for modelling and predicting continuous outcomes based on the relationship between independent and dependent variables.

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

ANS - Anscombe's quartet is a set of four datasets that have nearly identical descriptive statistics, including means, variances, correlations, and linear regression coefficients, but vastly different visual appearances when plotted. This quartet was introduced by the statistician Francis Anscombe in 1973 to illustrate the importance of visualizing data and the limitations of relying solely on summary statistics.

The four datasets in Anscombe's quartet have the following characteristics:

1. Dataset I:
    - Consists of linearly related data points.
    - Follows a simple linear relationship with a slope of approximately 0.5 and an intercept of around 3.
    - The relationship between the variables is apparent when plotted, and a linear regression model fits the data well.
2. Dataset II:
    - Consists of data points with a non-linear relationship.
    - Despite having the same mean and variance for both variables as Dataset I, the relationship is quadratic rather than linear.
    - A linear regression model would not accurately capture the relationship between the variables, highlighting the importance of considering alternative functional forms.
3. Dataset III:
    - Consists of mostly linearly related data points, with one outlier that significantly influences the correlation coefficient and linear regression line.
    - The outlier has a substantial effect on the regression model, demonstrating the sensitivity of these statistics to extreme values.
4. Dataset IV:
    - Consists of data points with a perfect linear relationship except for one outlier.
    - The presence of the outlier leads to a high correlation coefficient and a linear regression line that does not accurately represent the underlying relationship between the variables.

The key takeaway from Anscombe's quartet is that summary statistics alone, such as means, variances, and correlations, may not fully capture the complexity of a dataset.

Visualizing the data through plots and graphs is essential for understanding the relationships between variables, identifying patterns, and detecting outliers or unusual observations.

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

ANS - Pearson's correlation coefficient, often denoted as *r*, is a measure of the strength and direction of the linear relationship between two continuous variables. It quantifies the degree to which two variables are linearly related to each other.

Pearson's *r* can take on values between -1 and 1:

- *r*=1: Perfect positive correlation. This means that as one variable increases, the other variable also increases in a linear fashion.
- *r*=−1: Perfect negative correlation. This means that as one variable increases, the other variable decreases in a linear fashion.
- *r*=0: No linear correlation. This means that there is no linear relationship between the two variables.

The formula for Pearson's correlation coefficient between two variables *X* and *Y* is:

$r=\sum i=1n(Xi-X¯)2\cdot\sum i=1n(Yi-Y¯)2\sum i=1n(Xi-X¯)(Yi-Y¯)$

Where:

- *Xi* and *Yi* are individual data points for variables *X* and *Y*, respectively.
- $X¯$ and $Y¯$ are the means of variables *X* and *Y*, respectively.
- *n* is the number of data points.

Pearson's correlation coefficient is widely used in statistics and data analysis to assess the relationship between variables. However, it assumes that the relationship between the variables is linear and may not capture non-linear relationships. Additionally, it is sensitive to outliers and can be influenced by them. Therefore, it is essential to complement Pearson's *r* with other measures and visualizations when analysing relationships between variables.

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

ANS - Scaling is the process of transforming numerical features in a dataset to a common scale without distorting the differences in the range of values. It involves adjusting the values of the features so that they fall within a specific range or distribution. Scaling is performed to ensure that the features contribute equally to the analysis and prevent features with large magnitudes from dominating the algorithm's learning process.

Reasons for scaling:

1. Equal Weightage: Scaling ensures that all features contribute equally to the analysis, preventing features with larger magnitudes from dominating the algorithm's learning process.
2. Faster Convergence: Scaling can help algorithms converge faster during the training process, especially for gradient-based optimization algorithms like gradient descent.
3. Better Performance: Some machine learning algorithms, such as k-nearest neighbors (KNN) and support vector machines (SVM), are sensitive to the scale of features. Scaling can improve the performance and accuracy of these algorithms.
4. Regularization: Scaling is essential when using regularization techniques like L1 and L2 regularization, as it ensures that the regularization penalty is applied uniformly across all features.

Difference between normalized scaling and standardized scaling:

1. Normalized Scaling (Min-Max Scaling):
    - Range: Scales the features to a fixed range, typically between 0 and 1.
    - Formula: $x\text{scaled}=\max(x)-\min(x)x-\min(x)$
    - Effect: Preserves the original distribution of the data but shifts and rescales it so that the minimum value becomes 0 and the maximum value becomes 1.
    - Use Cases: Useful when the features have a known minimum and maximum value and the distribution is approximately Gaussian.
2. Standardized Scaling (Z-score normalization):
    - Mean and Standard Deviation: Scales the features to have a mean of 0 and a standard deviation of 1.
    - Formula: $x\text{scaled}=\text{std}(x)x-\text{mean}(x)$
    - Effect: Centers the data around 0 and adjusts the spread of the data so that it has a standard deviation of 1.
    - Use Cases: Suitable when the distribution of the data is Gaussian-like and the standard deviation is not too small.

In summary, normalized scaling (Min-Max scaling) transforms features to a fixed range (usually 0 to 1), while standardized scaling (Z-score normalization) standardizes features to have a mean of 0 and a standard deviation of 1. The choice between the two depends on the distribution of the data and the requirements of the machine learning algorithm being used.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**

ANS - The occurrence of infinite values for the Variance Inflation Factor (VIF) typically arises due to perfect multicollinearity among predictor variables. Perfect multicollinearity occurs when one or more independent variables in a regression model are perfectly linearly related to each other, meaning that they can be expressed as exact linear combinations of each other.

When perfect multicollinearity exists:

1. Matrix Inversion: In regression analysis, the calculation of VIF involves computing the inverse of the correlation matrix of the predictor variables. However, if perfect multicollinearity exists, the correlation matrix becomes singular (i.e., non-invertible), leading to numerical instability in the computation of VIF.
2. Division by Zero: The formula for calculating VIF involves dividing by the variance of each predictor variable. If a variable is perfectly collinear with other variables, its variance becomes zero. As a result, division by zero occurs when calculating VIF, leading to infinite values.
3. Perfect Redundancy: Perfect multicollinearity implies that one or more predictor variables contain redundant information, rendering them redundant in the regression model. Therefore, the VIF for these variables cannot be computed as they do not provide additional information beyond what is already captured by other variables in the model.

In practical terms, encountering infinite values for VIF serves as a warning sign that perfect multicollinearity exists among predictor variables. It indicates that the regression coefficients are not uniquely identifiable, and the model estimation becomes problematic. In such cases, it is necessary to address multicollinearity issues by either removing redundant variables, transforming variables, or using regularization techniques to stabilize the model estimation process and obtain reliable coefficient estimates.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**

ANS - A Q-Q plot, short for quantile-quantile plot, is a graphical tool used to assess whether a set of data follows a specific probability distribution, such as the normal distribution. It compares the quantiles of the data against the quantiles of a theoretical distribution, typically the standard normal distribution (mean = 0, standard deviation = 1), on the same scale.

How does a Q-Q plot work?

1. The data is sorted in ascending order.
2. The sorted data points are plotted against the quantiles of the theoretical distribution.

3. If the data closely follows the theoretical distribution, the points will fall approximately along a straight line.

Use and Importance of Q-Q Plot in Linear Regression:

1. Normality Assumption: In linear regression, one of the key assumptions is that the residuals (the differences between observed and predicted values) are normally distributed. Q-Q plots are commonly used to assess the normality of residuals. If the residuals follow a normal distribution, the points on the Q-Q plot will fall along a straight line. Deviations from this line indicate departures from normality.
2. Detecting Outliers: Q-Q plots can help detect outliers in the data. Outliers are data points that deviate significantly from the expected pattern. In a Q-Q plot, outliers may appear as points that deviate from the straight line, suggesting that they do not follow the same distribution as the rest of the data.
3. Model Fit Assessment: Q-Q plots can also be used to assess the fit of the linear regression model. If the residuals follow a normal distribution, it indicates that the model adequately captures the variation in the data. Deviations from normality may suggest that the model does not fully explain the variability in the data or that there are additional factors influencing the outcome.
4. Model Improvement If the Q-Q plot reveals departures from normality, it may indicate areas for model improvement. For example, transforming the response variable or including additional predictor variables may help improve the model's fit and address any violations of the normality assumption.
5. In summary, Q-Q plots are valuable tools in linear regression analysis for assessing the normality of residuals, detecting outliers, evaluating model fit, and identifying areas for model improvement. They provide visual insights into the distributional properties of the data and help ensure the validity and reliability of the regression analysis results.