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?

Answer - We have identified the seasons (spring, summer, fall, winter), weather situation (clear, misty, rainy, snowy), and months (jan, feb, mar, apr, may, jun, jul, aug, sep, oct, nov, dec) as the categorical variables. Based on the analysis done we can see all these specific categorical variables has a significant impact on the dependent variable. E.g., when we introduced the 'fall' season along 'yr', 'workingday', and 'windspeed' the adjusted r-squared dropped significantly and reached to **~50%**. Similarly, when we added the weather situation 'clear' alongside 'atemp', 'yr', and 'workingday' we saw a significant jump in the adjusted r-squared which reached to **~72%.**

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

Answer - In a multiple linear regression model, using **drop_first=True** during dummy variable creation is important to avoid multicollinearity. When we create dummy variables for a categorical variable with k categories, get_dummies typically create **k-1** dummy variables by default (**drop_first=False**). These **k-1** variables represent all but one category of the original variable. By setting **drop_first=True**, we can drop one of the dummy variables created by get_dummies. This ensures that the remaining dummy variables are independent of each other and do not contain redundant information.

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

Answer - Looking at the pair plot among all the numerical variables, the temp and atemp numerical variables has the highest correlation with the target variable. Although registered variable has even a higher correlation to the target, we have not considered it for the modeling exercise because it is part of the target of variable itself i.e., **casual + registered = total count**.

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

Answer - I validated the following linear regression assumptions after the building the model on the training set.

1. **Linear Relationship** - The core assumption is that the relationship between the independent and dependent variables is linear. Used scatterplot to check the visual linear patterns.

2. **Homoscedasticity** - Used the variance of errors (how spread out the errors are) to see if they are constant across all levels of the feature variables.

3. **No Multicollinearity** - I measured VIFs for each model to see if the feature variables are correlated with each other. Multicollinearity can make it difficult to isolate the effects of individual variables on the target variable, leading to unreliable coefficient estimates and inflated standard errors.

4. **Independence of Errors** - I plotted line graphs to see the differences between the actual and predicted values.

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

Answer - Instead of a single final model, I have produced 3 models that explains the demand of the shared bikes:

- **Model #8** - that uses the 'atemp','yr','workingday', and 'clear' feature variables to determine the target. This was arrived using the forward-pass technique where we kept adding a new feature variable on every iteration to get to an optimal model. We can see the Adjusted R-Squared on the training data is **72%** and on test data is **70%**
- **Model #13** - that uses the 'yr', 'workingday', 'feb', 'mar', 'apr', 'may', 'jun', 'jul', 'aug', 'sep', 'oct', 'nov', 'misty', and 'snowy' feature variables to determine the target. This was arrived using the backward-pass technique where we started with all the feature variables and then on every iteration removed one feature variable at a time, to get to an optimal model. We can see the Adjusted R-Squared on the training data is **75%** and on test data is **76%**
- **Model #18** - that uses 'yr', 'windspeed', 'spring', 'fall', 'aug', 'sep', 'clear', and 'snowy' feature variables to determine the target. This was arrived using the RFE technique to get to an optimal model. We can see the Adjusted R-Squared on the training data is **75%** and on test data is **71%**

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer - Linear regression is a fundamental statistical method used to model the relationship between a dependent variable (what we are trying to predict) and one or more independent variables (what we are using to make the prediction). It assumes a linear relationship between the variables. Here is a detailed breakdown of the algorithm:

- **Data Representation** - Our data is typically represented in a tabular format, like a pandas DataFrame in Python. Each row represents an observation (data point), and each column represents a variable. The dependent variable (y) is the variable we are trying to predict. The independent variables (X) are the variables we are using to make the prediction.

- **Model Formulation**: The core of linear regression lies in the linear equation: $y = b0 + b1*x1 + b2*x2 + ... + bn*xn + \varepsilon$, where:
    - y: Dependent variable
    - x1, x2, ..., xn: Independent variables
    - b0: Intercept (the y-axis value where the regression line crosses)
    - b1, b2, ..., bn: Coefficients (slopes) for each independent variable
    - ε: Error term (represents the difference between the actual y value and the predicted y value)
- **The Goal: Minimizing the Error**: The goal is to find the values of the coefficients (b0, b1, ..., bn) that minimize the sum of squared errors (SSE). $SSE = \Sigma (yi - \hat{y}i)^2$.
- **Least Squares Method**: The most common method to find the optimal coefficients is the least squares method. It minimizes the SSE. Solving the resulting system of equations yields the optimal coefficients.
- **Model Fitting**: Once we have the optimal coefficients, we can plug them back into the linear equation to get the fitted regression line. This line represents the predicted values of the dependent variable based on the independent variables.
- **Model Evaluation**: Several metrics can be used to evaluate the performance of our linear regression model:
    - **R-squared**: Represents the proportion of variance in the dependent variable explained by the model.
    - **Mean Squared Error (MSE)**: Average squared difference between actual and predicted values. (Lower is better)
    - **Adjusted R-squared**: Considers the number of independent variables to avoid overfitting.
    - **Coefficient p-values**: Assess the statistical significance of each coefficient.
- **Prediction**: Once we have a trained model, we can use it to predict the dependent variable for new observations (data points not included in the training data). We simply plug the new independent variable values into the fitted equation to get the predicted y value.

2. Explain the Anscombe's quartet in detail.

Answer - According to Wikipedia, Anscombe's quartet is a collection of four unrelated datasets created by statistician **Francis Anscombe in 1973**. The purpose was to highlight the importance of data visualization in exploratory data analysis (EDA) and to demonstrate that summary statistics alone can be misleading. This is what it describes:

1. **The Data**: Each dataset consists of 11 data points with identical:
    - Mean (x): 9
    - Mean (y): 7.5
    - Standard Deviation (x): 3.16
    - Standard Deviation (y): 1.94
    - Correlation Coefficient (~): 0.8 (approximate linear relationship)
2. **The Catch**: Despite these identical summary statistics, the data points in each dataset are distributed very differently.
3. **The Importance**: By visualizing the data (scatter plots), we can see the differences between the datasets:
    - Dataset 1: A clear linear relationship exists.

    - Dataset 2: A curved, non-linear relationship is evident.
    - Dataset 3: A tight linear relationship with one outlier.
    - Dataset 4: A cluster of data points with no clear relationship.
4. **The Lesson**: Summary statistics like mean, standard deviation, and correlation coefficient can provide a high-level overview of the data, but they can't capture the full picture. Visualization is crucial to understand the underlying distribution, identify potential outliers, and assess the nature of the relationship between variables.
5. **Real-World Implications**: Relying solely on summary statistics can lead to misinterpretations and inaccurate conclusions. Data visualization helps us to:
    - Explore data patterns and identify anomalies.
    - Choose appropriate statistical tests or models.
    - Gain a deeper understanding of the relationships between variables.

Anscombe's quartet reminds us that data visualization is a vital step in any data analysis process. It helps us move beyond the limitations of summary statistics and uncover the true story hidden within the data.

3. What is Pearson's R?

Answer - Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that indicates the strength and direction of the linear relationship between two continuous variables. It is denoted by the symbol r. Pearson's R can range from -1 to +1.

- +1: Indicates a perfect positive linear relationship. As one variable increases, the other variable consistently increases proportionally.
- -1: Indicates a perfect negative linear relationship. As one variable increases, the other variable consistently decreases proportionally.
- 0: Indicates no linear relationship between the variables. The changes in one variable are unrelated to the changes in the other.
- Values between 0 and +1 or 0 and -1 indicate a positive or negative correlation, respectively, but the strength of the relationship weakens as the value gets closer to 0.

The absolute value of Pearson's R (ignoring the sign) tells the strength of the correlation:

- Closer to 1: Stronger linear relationship (positive or negative)
- Closer to 0: Weaker or no linear relationship

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

Answer - Scaling refers to the process of transforming the values of features in our data to a common range. This is often done before training a model to improve its performance and stability. It is performed to:

- **Improve Model Performance**: The machine learning algorithms rely on the distances between data points to make predictions. Features with vastly different scales can skew

these distances, leading to suboptimal model performance. Scaling puts all features on a more balanced footing.
- **Preventing Numerical Issues**: Some algorithms are sensitive to the scale of features. Scaling can prevent numerical problems that might arise during calculations.

The two common type of scaling used are:

- **Normalization**: This technique scales the features to a specific range, typically between 0 and 1 (min-max scaling) or -1 and 1. It is determined by the formula `Scaled value = (original value - min_value) / (max_value - min_value)`
- **Standardization**: This technique scales the features by subtracting the mean and then dividing by the standard deviation. It is determined by the formula `Scaled value = (original value - mean) / standard_deviation`

The key differences between these two scaling algorithms are:

- Normalization uses a range between 0 and 1, whereas Standardization uses Mean=0 and Std. Dev. = 1
- Normalization is not centered around mean, whereas Standardization is centered around mean.
- Normalization is less sensitive, whereas Standardization more sensitive

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

Answer - The formula for VIF involves dividing `1 / (1 - R^2)`, where `R^2` is the coefficient of determination for a regression model where one independent variable is regressed on all the other independent variables.

- In perfect multicollinearity, R^2 becomes exactly 1. This is because one variable can be perfectly predicted by the others, leading to a perfect fit (R^2 = 1).
- Dividing 1 by (1 - 1) results in an indeterminate form, which is typically represented as infinity.

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

Answer - A Q-Q plot, also known as a Quantile-Quantile plot, is a graphical tool used to compare the quantiles (distribution) of two datasets. In linear regression, it is primarily used to assess whether the errors (residuals) of our model follow a normal distribution, which is one of the key assumptions of linear regression. Here is how it is used:

- **Data Preparation**: We plot the quantiles of the residuals from our linear regression model on the x-axis. The quantiles of a distribution (usually a normal distribution) are plotted on the y-axis.

- **Interpretation** - If the residuals follow a normal distribution, the points on the Q-Q plot will fall along a straight diagonal line. This indicates that the errors are randomly scattered and do not exhibit any specific pattern. Deviations from the straight line suggest potential violations of normal distribution assumption.

There are several reasons why it is important in linear regression:

- Reliable Hypothesis Tests: Many statistical tests used in linear regression, such as t-tests for coefficient significance, rely on the normality assumption. Violating this assumption can lead to unreliable p-values.
- Confidence Intervals: Confidence intervals around the regression coefficients also depend on the normality of errors.