Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 1 goes below this line> (Do not edit)

Based on the analysis I have done, categorical variables have a significant influence on the dependent variable ("cnt") i.e. bike rental demand. Below are the highlights of few such categorical variables:

a. **Year ("year"):** Rentals increased significantly from 2018 to 2019, as indicated by the positive coefficient of "year". This could reflect increased demand over time, potentially due to growing popularity, expanded bike-sharing infrastructure, or marketing efforts.
b. **Season ("season"):** "Summer" and "Fall" seasons showed increase in demand whereas "Spring" and "Winter" demonstrated a decline.
c. **Weather Situation("weathersit"):** Misty or cloudy days ("weathersit" category level 2) reduced rentals compared to clear days ("weathersit" category level 1). Bad weather conditions (light snow or rain i.e. "weathersit" category level 3) further reduced rentals.
d. Then there are some minor impacts month ("Sep") i.e. September month or Fall showed increase in demand.

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 2 goes below this line> (Do not edit)

For categorical variables with "n" levels, we only need to create dummy variables for "n-1" levels as we can explain the left-out variable with the values of other variables. This also helps avoid multicollinearity issue. Multicollinearity caused by redundant dummy variables can lead to instability in the estimated coefficients, causing large variances and making the model sensitive to small changes in data. Dropping one dummy variable helps ensure that each feature contributes uniquely, making the model's estimates more stable and reliable.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)

Temperature ("temp") has the highest correlation with the target variable of bike rental demand ("cnt"). The Pearson's r correlation value is 0.63. I dropped the Temperature Feel variable (" atemp") due to multicollinearity issue (with Temperature ("temp") variable) which also showed a similar correlation.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

I used the following methods to validate the assumptions of Linear Regression:

- **Linearity**:
  - I checked for linear relationships between predictors/independent variables and the target variable ("cnt") during Exploratory Data Analysis (EDA) in section 6.2 of the Jupyter notebook "bike-sharing-model-abhishek-mukherjee.ipynb" using a correlation heatmap and scatter plots ("temp" and "atemp" variables showed promising linear relationship). Linear relationships support the assumption that predictors have a linear association with the target.
- **Normal Distribution of Residuals**:
  - After fitting the model, I plotted a seaborn distplot of the residuals (differences between actual and predicted values) in section 12 of the Jupyter notebook "bike-sharing-model-abhishek-mukherjee.ipynb". The residuals/error terms showed a normal distribution.
- **Homoscedasticity**:
  - I created a scatter plot of residuals vs. predicted values in section 12 of the Jupyter notebook "bike-sharing-model-abhishek-mukherjee.ipynb". For homoscedasticity, the residuals should show constant variance across predicted values, with no clear pattern. The residual plot did not show any systematic pattern, indicating that the assumption of homoscedasticity was met.
- **Multicollinearity**:
  - I assessed multicollinearity before modelling for the "temp" and "atemp" variables which showed similar patterns and were highly correlated with each other in the heatmap visual (correlation of 0.99) in EDA section 6.2 of the Jupyter notebook. Then I used Variance Inflation Factor (VIF) after the initial model fitting. I also dropped the "const" variable which had a high value of ~9.87 indicating multicollinearity issues.
- **Independence of Errors**:
  - The Durbin-Watson statistic (reported in the OLS summary) was close to 2, suggesting that residuals are uncorrelated (no significant autocorrelation), which is ideal for linear regression.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

Based on my final model (model 6 in manual feature selection post RFE) which had six variables below are the top 3 predictors/features:

a. **Spring Season:** The variable "spring" has a coefficient/weightage of -0.8644. The bike demand will significantly fall in spring and the company needs to invest in some marketing strategies to attract more people during spring. "Summer" and "Fall" seasons drives demands up.
b. **Weather Conditions:** Weathersit variables like "Light Snow" (coefficient/weightage of -1.4208) and "Mist" (coefficient/weightage of -0.4554) drives demands down.
c. **Year:** This variable "year" has a coefficient/weightage of 0.9081. Year on year demand is increasing. Covid-19 may have temporarily stifled demand but as the pandemic is gone now (or is less serious) the demand is expected to rise again. People are also becoming more climate conscious and prefer bikes for shorter transits and for health benefits.

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Based on what I have learned in the course so far, Linear regression is a predictive modeling technique that establishes a linear relationship between a dependent (target) variable "y" and one or more independent (predictor) variables "X". The primary goal of linear regression is to predict the target variable by fitting a linear equation, which can be represented in two main forms:

a. Simple Linear Regression: This form involves only one predictor variable and models "y" as:
$y = \beta_0 + \beta_1 X_1 + \varepsilon$

Where β_0 is the intercept (the expected value of y when X=0), β_1 is the slope (indicating the rate of change in y for a unit change in X), and ϵ represents random error (accounting for variance not explained by X).

b. Multiple Linear Regression: This form extends the model to multiple predictors, represented as:
$y = \beta_0 + \beta_1 X_1 + \beta_1 X_1 \ldots\ldots + \beta_n X_n + \varepsilon$
where each β represents the partial effect of each X on y, holding other variables constant.

Main Goal: The primary objective of linear regression is to estimate the coefficients β values that minimize the residual sum of squares (errors), or RSS. This error-minimizing process, known as Ordinary Least Squares (OLS), finds the line of best fit that minimizes the vertical distances between observed values and the predicted line.

Key Assumptions:

a. Linearity: There is a linear relationship between the predictors and the outcome/target variable.
b. Independence of Errors: Observations are independent, and error terms are uncorrelated.
c. Homoscedasticity: The variance of residuals is constant across all levels of the independent variables.
d. Normality of Errors: Residuals should be normally distributed, especially for inference.
e. Low Multicollinearity: Predictor variables should not be too highly correlated with each other (VIF or Variance Inflation Factor values should be less than 5).

The resulting model can be used to interpret the effect of each predictor on "y" or the target variable and make predictions on new data. Coefficients represent the expected change in "y" for a unit increase in "X", holding other variables constant, allowing insights into the influence of each predictor.

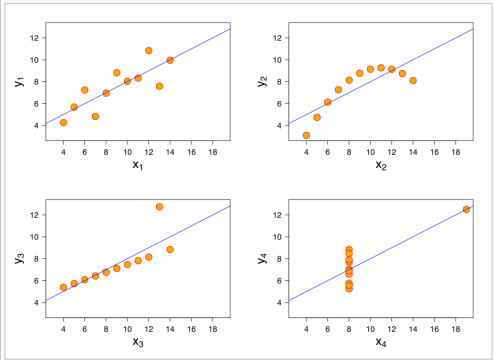
Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Anscombe's Quartet is a set of four datasets that demonstrate the importance of data visualization and caution against relying solely on summary statistics. Each dataset in the quartet has nearly identical statistical properties:

- Mean of X and y values
- Variance of X and y
- Correlation coefficient between X and y
- Linear regression line fitting $y = \beta_0 + \beta_1 X_1 + \varepsilon$.

Despite these identical summary statistics, visualizing each dataset reveals different patterns, highlighting the importance of graphical analysis through data visualization.
When plotted, the datasets reveal drastically different relationships:

- Dataset I: Shows a simple linear relationship between X and y.
- Dataset II: Exhibits a clear non-linear relationship.
- Dataset III: Displays a linear relationship with a single outlier significantly influencing the regression line.
- Dataset IV: Shows a vertical line with a single outlier.



Source: https://en.wikipedia.org/wiki/Anscombe%27s_quartet
Following are the key takeaways:

- Summary statistics can be misleading. Anscombe's quartet highlights how datasets with identical statistical properties can have vastly different distributions and underlying relationships.
- Graphical visualization is crucial. Plotting the data reveals patterns, outliers, and trends that are hidden by summary statistics alone.
- Data analysis or EDA should be comprehensive. A thorough analysis should involve both numerical summaries and graphical visualization to gain a complete understanding of the data.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Pearson's R, or the Pearson Correlation Coefficient, is a statistic used to measure the strength and direction of a linear relationship between two continuous variables, often labeled X and y.
Key characteristics of Pearson's R are as follows:

- **Range:** It ranges from -1 to +1.
  - **+1:** Indicates a perfect positive linear relationship (as one variable increases, the other increases proportionally).
  - **-1:** Indicates a perfect negative linear relationship (as one variable increases, the other decreases proportionally).
  - **0:** Indicates no linear relationship between the variables.

- **Magnitude:** The absolute value of r indicates the strength of the relationship. A value closer to 1 signifies a stronger relationship, while a value closer to 0 signifies a weaker relationship.
- **Linearity:** Pearson's R specifically measures the linear association. It may not accurately capture the relationship if the variables have a non-linear association (e.g., curved relationship).

When interpreting Pearson's R, we need to consider both the sign and magnitude. For example, an r of 0.8 suggests a strong positive linear relationship, while an r of -0.2 indicates a weak negative linear relationship.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Scaling, also known as feature scaling, is a crucial data preprocessing step in machine learning modelling techniques as we have learned during the Linear Regression course.
It involves transforming the values of numeric features in a dataset to a common scale. This is done without distorting the differences in the ranges of values or losing information.
Scaling is performed because of following reasons

- ML Algorithm performance: Machine learning algorithms perform better or converge faster when features are on a similar scale. We don't want the model to be influenced or dominated by features with larger values, so it's best to bring all the features at the same scale
- Gradient Descent: Linear Regression uses the gradient descent algorithm as an optimization technique to minimize errors. Having features on same scale can speed up the optimization and help the algorithm converge faster.
- Interpretability: Scaled data allows for better interpretation and comparison between features when building models or creating visualizations.

Normalization vs. Standardization
Here are the differences:

a. **Normalization or Min-Max Scaling:** It rescales the data between the range of 0 and 1
Its formula is $X_{norm} = (X - X_{min}) / (X_{max} - X_{min})$
It is sensitive to outliers which may impact the min and max values. It is also bounded within a specific range and is useful for non-Gaussian distributions.

b. **Standardization or Z-Score Scaling:** Standardization rescales data to have a mean of 0 and a standard deviation of 1.
Its formula is $X_{std} = (X - \mu) / \sigma$, where μ is the mean, and σ is the standard deviation.
It centers data around zero with unit variance, ideal for normally distributed data or when both positive and negative values are meaningful in the analysis.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

A Variance Inflation Factor (VIF) value becomes infinite (or extremely large) when a feature in the dataset is a perfect linear combination of one or more other features, resulting in a phenomenon

called perfect multicollinearity. This situation typically arises due to redundancy in predictor variables.
The formula for VIF is:
$VIF_i = 1 / 1 - R_i^2$
where R_i^2 is the value from regressing the i^{th} feature against all other features.
If there is perfect multicollinearity (one feature is an exact linear combination of others),
then $R_i^2 = 1$
Substituting 1 into the VIF formula:
$VIF_i = 1 / (1 - 1) = \infty$

Some of the common causes of this can be:

- When all dummy variables of a categorical feature are included, one dummy is redundant because it can be derived from the others. This causes perfect multicollinearity.
- Including highly similar or derived features (like a feature and its squared term) can lead to extreme or infinite VIF.
- In certain datasets, features may need transformation (e.g., removing redundant features) to avoid dependencies.

What we have learned during the Linear Regression module is that when we derive dummy variables for a categorical feature with "n" levels, we need to only use "n-1" features and drop the first one using option drop_first=True in the pd.get_dummies() function.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
(Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, or exponential.
For Linear Regression we have used Normal/Gaussian distribution as the theoretical distribution for checking the error terms or residuals distribution alignment with it. If the error terms had normal distribution with a mean centered around 0 and a standard deviation of 1, then the modelling process is on the right path.
