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)

From the analysis of the categorical variables in the BoomBikes dataset, we can infer the following effects on the dependent variable (cnt, the total number of bike rentals):

Season - The spring season shows the lowest demand, while fall and summer have the highest demand.People prefer biking in pleasant weather conditions. Spring may have lower demand due to unpredictable weather or transitions from winter.

Month - September has the highest bike demand, while January and winter months show significantly lower demand. September likely benefits from pleasant weather, while winter months see a decline due to cold and snow, making biking less appealing.

Weekday- Weekends (Saturday & Sunday) have lower demand than weekdays. This indicates that a significant portion of the bike-sharing demand comes from commuters who use the service for work or school during weekdays.

Year - The demand has increased from year 0 (2011) to year 1 (2012). This suggests that bike-sharing is becoming more popular over time, likely due to increased adoption and awareness.

Holiday - Bike demand is lower on holidays. Since many people use the service for commuting, holidays reduce the number of regular users.

Working Day - Higher demand on working days, lower on non-working days. Supports the hypothesis that the service is mostly used for daily commuting rather than leisure.

Weather Situation (Weathersit) - Clear and partly cloudy weather has the highest demand, while snowy, rainy, and humid conditions show a sharp decline. Bad weather discourages outdoor activities, including biking.

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)

By using drop_first=True, we drop one category as a reference, keeping only n-1 dummy columns. The dropped category is implicitly represented when all other dummies are 0.

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)

From the pair-plot among the numerical variables (temp, atemp, humidity, windspeed, and cnt), the variable that has the highest correlation with the target variable (cnt) is temp (temperature).

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)

After building the linear regression model on the training set, I validated the key assumptions of linear regression using various statistical techniques and visualizations.

1. Linearity - Used pair plot and scatter plot to check the Linear relationship
2. Multicollinearity - Calculated VIF and dropped variables with strong corr values
3. Homoscedasticity - Used Residual vs Predicted plots
4. Normality of Residuals - Histogram and QQ plots were used to compare normality
5. Model Performance Evaluation - R Square score was calculated to check for variance

---

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 the final model, following is my analysis

1. Temperature (atemp) - Strongly impact on demand for bike renting
2. Year (year) - Strong impact where demand increases year on year
3. Humidity (humidity) - it has strong negative impact on demand for bike renting

---

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)

Linear regression is a **statistical** method used to model and analyze the linear relationship between a dependent variable and one or more independent variables. This relationship implies that changes in the independent variables (increase or decrease) are associated with proportional changes in the dependent variable. As a foundational algorithm in supervised learning, linear regression predicts outcomes based on labeled training data.

Mathematical Representation
The relationship between variables is expressed by the equation:

Mathematical Representation

The relationship between variables is expressed by the equation:

$$Y = mX + c$$

Where:

- **Y**: Dependent variable (the outcome we aim to predict or explain).
- **X**: Independent variable (the input used for predictions).
- **m**: Slope of the regression line, indicating the change in Y per unit change in X.
- **c**: Y-intercept, representing the value of Y when $X = 0$.

Types of Linear Relationships

A **positive linear relationship** occurs when both the independent and dependent variables move in the same direction. For example, if (X) increases and (Y) also increases proportionally, the relationship is positive.

This model is widely used in fields like economics, biology, and social sciences to predict trends, analyze correlations, and inform decision-making.

A **negative Linear** relationship occur if independent increases and dependent variable decreases. For example, if (X) increases and (Y) decrease proportionally, the relationship is negative.

Linear regression is categorized into two primary types:

1. **Simple Linear Regression**:
   - Uses one independent variable to predict a dependent variable.
   - Example: Predicting house prices based solely on square footage.
2. **Multiple Linear Regression**:
   - Uses two or more independent variables to predict a dependent variable.
   - Example: Predicting house prices using square footage, location, and number of bedrooms.

Assumptions in Linear Regression

For reliable results, linear regression relies on the following assumptions about the data:

1. Minimal Multi-collinearity:
   - Independent variables should not be highly correlated with each other.
   - Why? Overlapping information between features can skew the model's interpretation of their individual effects.

2. No Auto-correlation:
   - Residuals (errors) should not correlate with each other, especially in time-series data.
   - Example: In stock price prediction, today's error should not predict tomorrow's error.

3. Linear Relationship:
   - The connection between dependent and independent variables must be linear.
   - Non-linear relationships may require transformations (e.g., log, square root) before modeling.

4. Normality of Residuals:

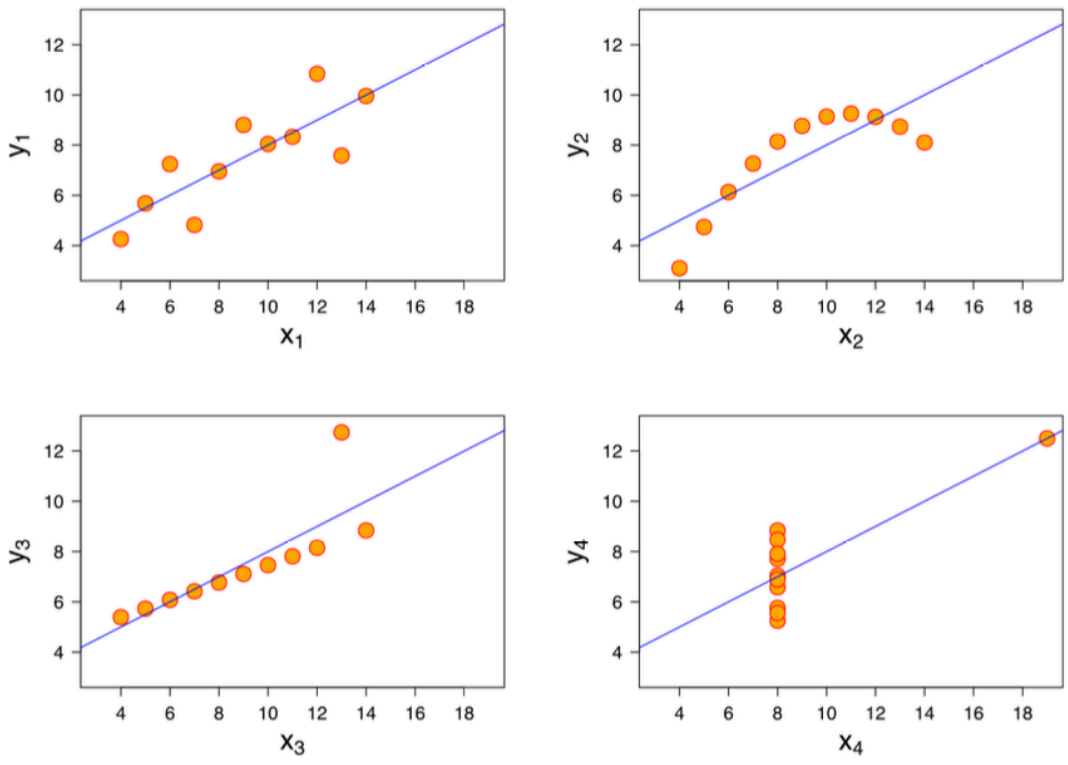
- Residuals should follow a normal distribution (bell-shaped curve).
- Ensures reliable confidence intervals and hypothesis tests.

5. Homoscedasticity:
- Residuals should have constant variance across all predicted values.
- Violation (heteroscedasticity) often appears as a “fan-shaped” pattern in residual plots.

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, created by statistician Francis Anscombe in 1973, highlights why visualizing data is essential before applying statistical models. The quartet consists of four datasets that share nearly identical summary statistics (e.g., mean, variance, correlation). However, when plotted graphically, they reveal entirely different patterns, proving that numerical summaries alone can be misleading.

Anscombe's quartet is mainly a group of four data sets which are nearly identical in simple descriptive statistics, but there are some strange behaviours in the dataset. They have very different distributions and appear differently when plotted on scatter plots.
Each dataset consists of eleven (x,y) points.



Observations from above data sets:
Data Set 1: fits the linear regression model well

Data Set 2: cannot fit the linear regression model because the data is non-linear
Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model
Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model
We observed that Anscombe's quartet helps us to understand the importance of data visualizations to build a well-fit model.

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)

The Pearson's Correlation Coefficient (r) is a statistical measure of the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0. Pearson's R cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.
r = 1 means strong positive relationship. If one increases, the other also increases
r = -1 means strong negative relationship. If one increases, the other decreases
r = 0 means there is no linear relationship
Formual for Pearson's Correlation Coefficient (r)

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2} \times \sqrt{\sum(Y_i - \bar{Y})^2}}$$

- r = Pearson's correlation coefficient
- X_i = Individual values of the first variable (X)
- Y_i = Individual values of the second variable (Y)
- $\bar{X}$ = Mean of X
- $\bar{Y}$ = Mean of Y
- $\sum$ = Summation notation

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 is nothing but putting the numerical feature values in the same range. Scaling is very important because few variables may have values on a different scale (smaller/higher) compared to other variables. Scaling also helps in speeding up the calculation in an algorithm.
Scaling is performed to bring all the numerical features in the same range. If some feature having value in range of thousands or lakhs while other in the range on tens or hundreds, for example, if salary column has values ranging from 25k to 10L and no of years experienced is in the range 1-25, so the model will take magnitude in account and not the units resulting in wrong or incorrect modelling. Hence, it is very important to scale all numerical feature in the same range before building the model.

There are many types of scaling, but normalized and standardized scaling are popular and widely used.

Normalized scaling: Scaling which makes all numerical feature lie in the range of 0 and 1. One disadvantage of normalization is that it losses some information in the date such as outliers.

Standardized scaling (Z-score Scaling): Standardized scaling replaces the values by their Z scores. It brings all
the data into a standard normal distribution which has mean (μ) 0 and standard deviation (σ) 1.

Min-Max Scaling is used in our project for normalizing temperature, humidity, and windspeed.

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)

Variance Inflation Factor (VIF) measures multicollinearity. An infinite VIF suggests **perfect collinearity**, meaning one independent variable is a perfect linear combination of others. In such cases, that variable should be removed.

VIF tells how much an independent variable is correlated with other independent variables. It detects the multicollinearity in the OLS regression analysis.
VIF below 5 is a good VIF. And VIF above 10 shows high correlation and should be removed.
If there is a perfect correlation between two independent features, then the VIF is infinite.

VIF is defined as:VIF = 1 / (1-R^2)
So, if r2 is 1 then the denominator becomes 0 and hence the VIF i.e., 1/0 is infinite. This means that variable is fully explained by some other variable in the model and hence does not make any sense to keep this feature in the model.

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)

Quantile-Quantile or Q-Q plot is a plot of the quantiles of the theoretical set against the quantiles of the sample set. It helps us understand if a sample comes from a known distribution such as normal distribution. In Regression, we use Q-Q plot to check if the data in the sample is normally distributed. Plotting the first data set's quantiles along the x-axis and plotting the second data set's quantiles along the y-axis is how the plot is constructed. If two distributions being compared are similar, the points on Q-Q plot will lie approx. on the line y=x

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions

Q-Q plot helps us to determine if two population are of the same distribution, if residuals follow normal distribution and if there is any skewness in the distribution.
If date sets, we are comparing are of the same type of distribution type, then plot would be a straight line.
