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)

1. Bookings are typically high during the fall season, followed closely by the summer season. Generally, bookings peak from April to November compared to the average. Clear weather also correlates with increased bookings. On holidays, there is a consistent trend of high bookings, which cluster around the higher side of the average.

2. In 2019, there were more bookings than in 2018, with significant increases across all months. While demand was particularly high in May, June, and July in 2018, the peak months in 2019 shifted to June through September. This trend of increased bookings during clear weather continued from 2018 to 2019, with a notable rise in numbers.

3. Additionally, bookings tend to increase as the weekend approaches, peaking on Saturdays. Conversely, bookings are generally lower on non-holidays, which makes sense since many people prefer to spend time at home with family during these times.

4. Overall, while there isn't much difference in bookings on weekdays, 2019 saw an increase compared to 2018. This upward trend suggests that the company may have penetrated the market more effectively in 2019.

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)

Using **drop_first = True** is important because it reduces the number of dummy variables created, which helps maintain the correlation among variables and preserves their independence. This is crucial when building a linear regression model.

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)
The variable "temp" exhibits the strongest correlation with the target variable.

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)

A linear regression model must satisfy the following conditions:

1. **Normality of Error Terms**: The error terms should follow a normal distribution with a mean of 0.
2. **Multicollinearity Check**: All variables (features) should be independent of one another.
3. **Linear Relationship**: There should be a clear linear relationship among the variables.
4. **Homoscedasticity**: The residual values should not exhibit any visible patterns.
5. **Independence of Residuals**: There should be no autocorrelation among the residuals.

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)

Top 3 are:

1. Temp
2. Year
3. Light_snowrain

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 a statistical method used to model the relationship between one or more independent variables (predictors) and a dependent variable (outcome). It is widely used for predictive modeling and inference in various fields such as economics, biology, engineering, and social sciences.

Key Concepts:

Dependent Variable: Variable that Is dependent on other variables. Normally it is the target variable which we want to detect

Independent Variable: Variables that are not dependent on each other variables. We can call them features which are used to predict the target variable

Linear Relationship: Dependent variable must have linear relationship with independent variable. Then only Linear regression can be applied

Mathematical Formula:

$$Y=c + mX$$

Y: Target variable
c: constant known as Y intercept
m: coefficient of X, represent the slope between Y and X.
X: independent variable on which Y depends and used to predict value of Y.

Estimation of Coefficients

The coefficients are estimated using a method called **Ordinary Least Squares (OLS)**, which aims to minimize the sum of the squares of the residuals (the differences between observed and predicted values).

Assumptions:

Linearity: The relationship between dependent and independent variables is linear**.**

Independence: Observations are independent of each other.

Homoscedasticity: The variance of residuals is constant across all levels of X.

Normality: The residuals of the model should be approximately normally distributed

---

Model Evaluation

To evaluate the performance of a linear regression model, several metrics can be used:

- **R-squared**: Represents the proportion of the variance for the dependent variable that is explained by the independent variable(s). Ranging from 0 to 1, a higher value indicates a better fit.
- **Adjusted R-squared**: Adjusts R-squared for the number of predictors in the model; useful for comparing models with different numbers of independent variables.
- **P-values**: Used to determine the significance of each coefficient.
- **Mean Squared Error (MSE)**: The average of the squares of the errors, used to measure how close the predictions are to the actual outcomes.

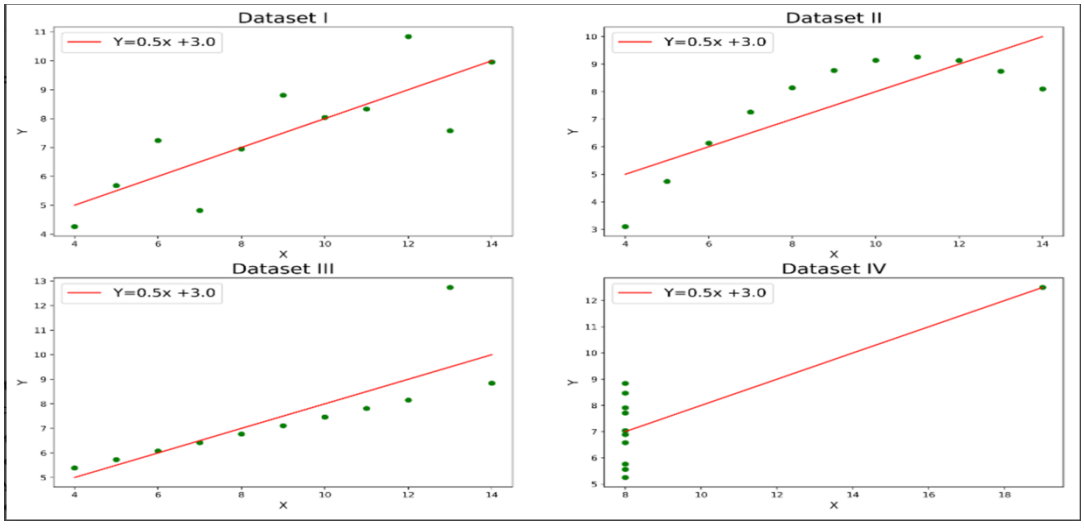
Question 7. Explain the Anscombe's quartet in detail. (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading. The four dataset contains 11 x-y pairs of data. These data have same statistical summary but when plotted graphs each data set has unique connection between X and Y.



From above plots it is observed that even though data set has same statistic values when they plotted conclusions are completely different

1. Data set 1 has cleat linear relationship and thus good fit
2. Data set 2 doesn't have linear relationship .
3. Data set 3 has problem with outlier.
4. Data set 4 has shown that one outlier is enough to show high correlation coefficient

---

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 coefficient is a correlation coefficient that quantifies the relationship between two variables measured on the same interval or ratio scale. It assesses the strength of the association between two continuous variables.

The value of the Pearson coefficient ranges from -1 to +1. A value of 0 indicates no correlation, while +1 and -1 represent strong positive and negative correlations, respectively. A positive value signifies a positive relationship between the variables, whereas a negative value indicates a negative 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 is a technique used to standardize the independent variables in the dataset. This preprocessing step is essential before building the model on the training set. The purpose of scaling is to reduce the range of values of different independent variables, which can impact model performance. By doing so, it mitigates the influence of variables with larger value ranges on the model. After scaling, all variables will have the same minimum and maximum values.

Normalized Scaling	**Standardized Scaling**
Minimum and Maximum values are used for scaling	Mean and standard deviation are used for scaling
It is used when features are of different scale	It is used when we want to ensure zero mean and unit standard deviation
Min and max are [0,1] or [-1,1]	There is no absolute min and max
It is affected by outliers	It is not effected by outliers

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)

VIF is used to calculate if there is variable or feature that is dependent on other variables. Higher the VIF higher the correlation of variable on other variables

$$VIF=1/(1-R2)$$

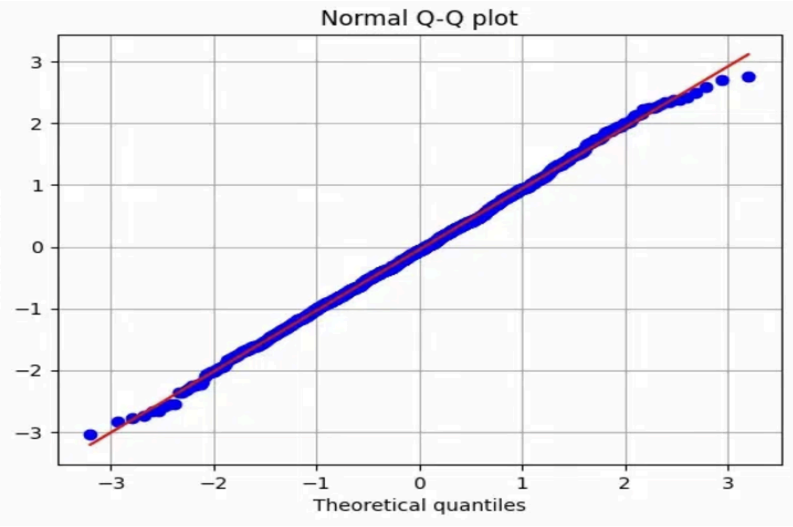
If VIF is infinite this means R2 is 1, which indicates strong correlation between two independent variables. When this is observed we can immediately drop the any one of the variables.

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)

QQ plot is called Quantile-Quantile plot. QQ plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.



Q-Q plot

QQ plot is used to validate if training data set and testing data set are having same distribution or not. Which will help determine statistical nature of the training data not much changed from testing data so model can be applied.
