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 high in fall season followed by summer season
2. bookings are normally high in months between [apr - nov], compared to average booking
3. booking is high when weather is clear
4. When day is holiday there is consistency of bookings. means bookings are usually high and spread around mean of bookings at higher side.
5. 2019 has more booking than 2018
6. compared to 2018 in 2019 in all month's bookings increased a lot. but demand seems to be high in May, Jun, July in 2018 and June to September in 2019
7. bookings are high in clear weather. same followed from 2018 to 2019 but increase in bookings a lot.
8. bookings are increasing as we are reaching weekend and through the weekend and peeking at Saturday.
9. When its not holiday, booking seems to be less in number which seems reasonable as on holidays, people may want to spend time at home and enjoy with family.
10. don't see much difference with working day, but bookings increased in 2019 compared to 2018.
11. obviously, bookings increased from 2018 to 2019. this may suggest the penetration of market by company 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)

drop_first = True is important as it helps create less number of dummy variables when creating them. Which helps correlation among variables and maintains independency of variables which is important when building Linear regression

---

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)

"temp" has the highest correlation with 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)

Linear Regression model must have following conditions

1. Normality of Error terms: Error follow normal distribution with mean 0

2. Multicollinearity Check: All variables[features] should be independent of each other.
3. Linear Relationship: Linearity should be visible among the variables
4. Homoscedasticity: there should be no visible pattern among residual values
5. Independent of Residuals: No auto correlation

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**.**

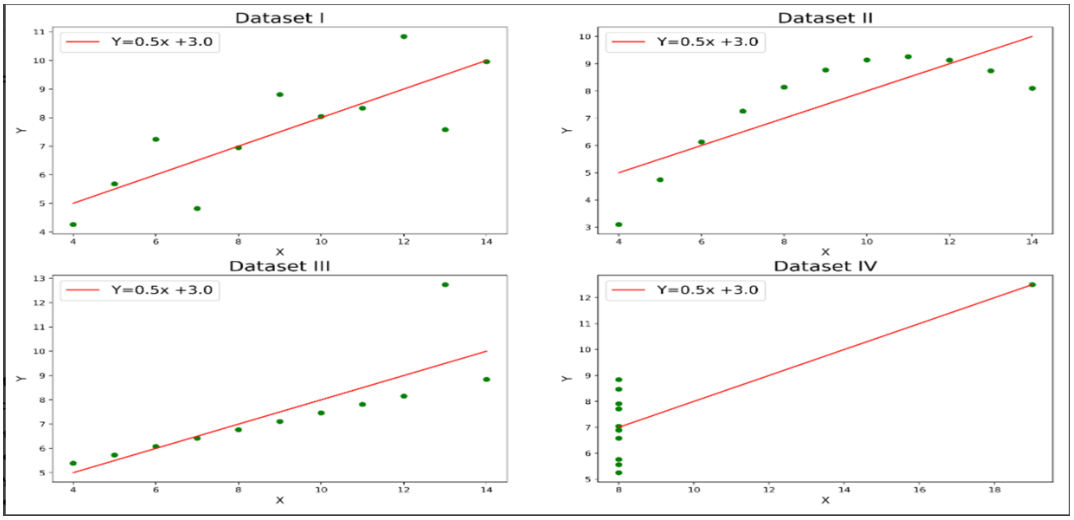
Question 7. Explain the Anscombe's quartet in detail. (Do not edit)

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 type of correlation coefficient that represents the relationship between two variables that are measured on the same interval or ratio scale. The Pearson coefficient is a measure of the strength of the association between two continuous variables.

The value of Pearson coefficient varies from -1 to +1. Value of 0 means there is no correlation 1 and -1 being the strong correlation. If the value is positive, it means it is positive relationship between variables. Otherwise, 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 the technique used to standardize the independent variables present in the data. It is data preprocessing step before building the model on training set. This is done to reduce the range of values of different independent variables having effect on model building. This will eliminate the variable which has larger range of values having effect on the model. After performing the scaling all variables will have same min and max.

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)

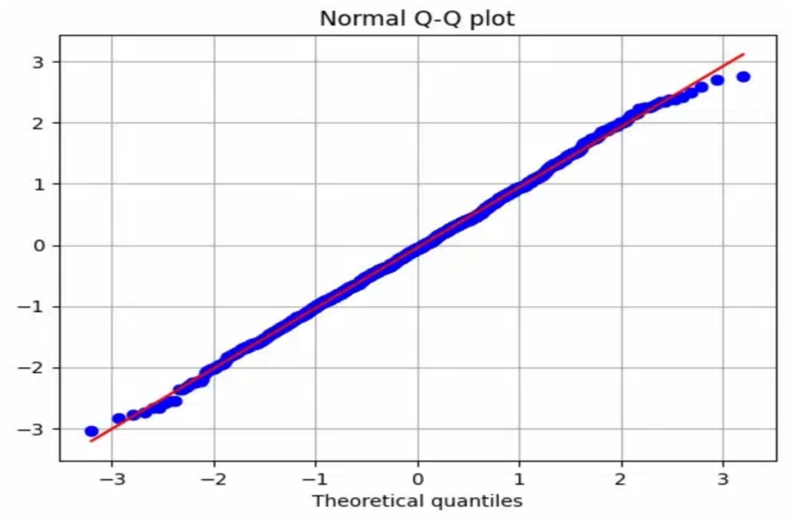
If VIF is infinite this means R2 is 1, which indicates strong correlation between two independent variables. When this is observed we can immediately drop the any one of the variables.

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.