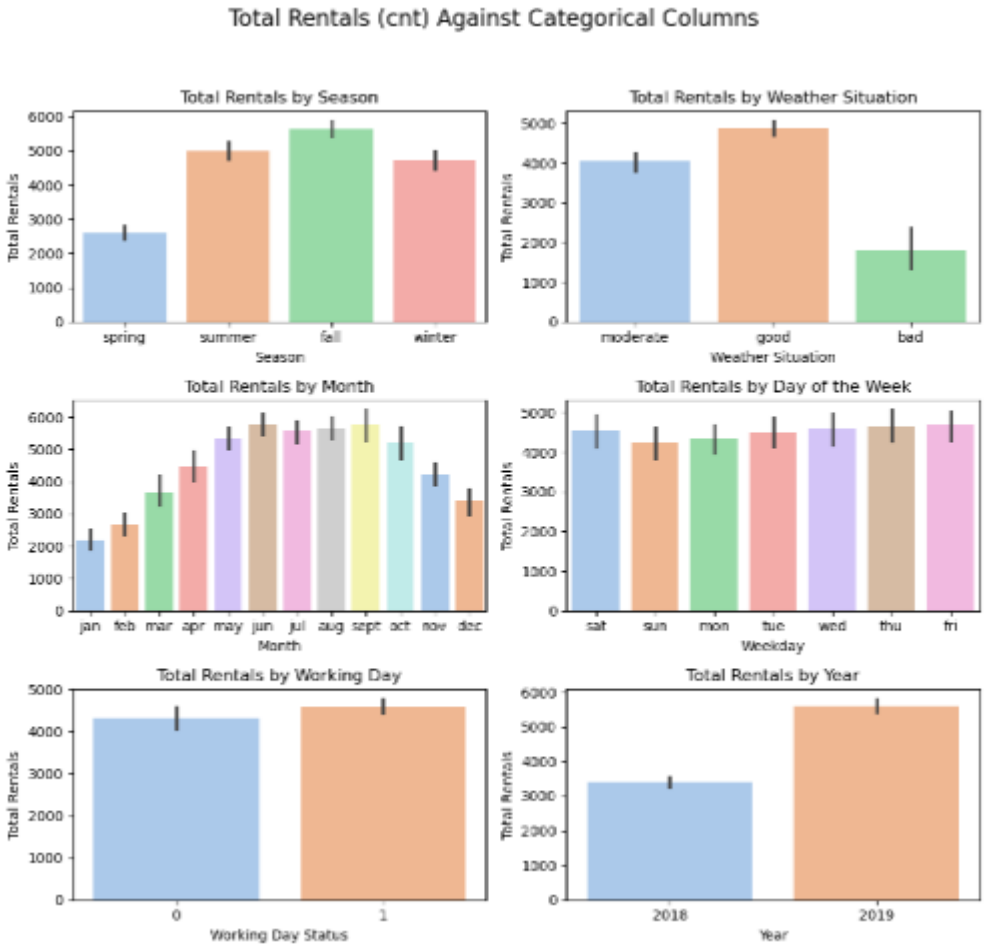
Assignment-based Subjective Questions

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)

We have few categorical variables as 'season', 'weathersit', 'month', 'weekday', 'workingday' and 'year'. From my analysis, these variables have a major effect on the dependent variable 'cnt', the same has been represented in the correlation in the barplots.



Based on the above visualization, following points could be inferred:

1) Fall season got more booking than other seasons
2) Following months June, August and September are the months with highest bookings
3) January month got the least bookings
4) Get more booking on good weather.
5) Not much difference in booking on working/non-working day.
6) Bookings increased drastically in the year 2019 compared to 2018

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)

When we are creating a dummy variable, an extra column is created. By using **drop_first=True,** you drop one extra column created during dummy variable creation. This ensures the remaining variables are independent and there is no dependency.

Example: If you have a categorical feature like 'season' with values as 'spring', 'summer', 'fall', 'winter' and you create dummy variable, you will end up with four dummy variables, for each season. As knowing the values of three variables is enough to infer fourth, this created redundancy in the model.

However, when you use **drop_first=True,** you drop one of the dummy variable ensuring remaining variables are not redundant.

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)

Variables 'temp' and 'atemp' have highest correlation in with the target variable 'cnt'.

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)

Following methods are used to validate the assumptions of Linear regression after building the model on training set:
Normally of error
Multicollinearity check
Homoscedasticity
Independence of residual

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)

Our final model indicates following top 3 features contributing significantly towards demand of shared bikes [1] temp [2] year [3] season_winter

---

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)

<Your answer for Question 6 goes here>
Linear regression is foundation algorithm in machine learning for modeling the relationship between a dependent variable (target variable) and one or more independent variables (predictors or features). The goal of linear regression is to fit a linear relationship between the predictors and the target variable.
Linear regression shows linear relationship which means it finds the value of the dependent variable is changing according to the value of independent variables.

There are broadly two kinds of liner regression:
[1] Single linear regression: Involves only one independent variable.
[2] Multiple linear regression: Involves two or more independent variables.

The objective to linear regression is to find the best-fitting line, that minimizes the error between the predicted values and the actual values of the target variable.

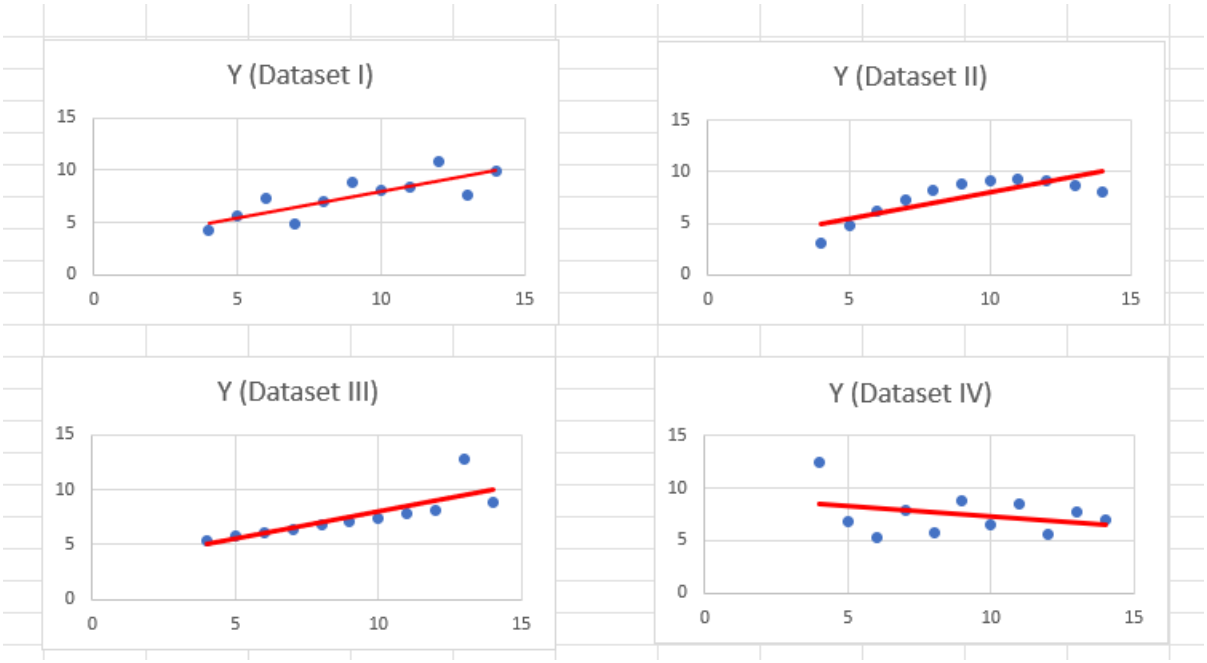
Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>
Anscombe's quartet created by the statistician 'Francis Anscombe' in 1973 to demonstrate the importance of graphical analysis and the potential pitfalls of relying solely on summary statistics like mean and variance.
The quartet consists of four datasets, each containing eleven (x, y) pairs. These datasets have nearly identical simple descriptive statistics, but they exhibit very different underlying distributions and relationships when they are graphed.
Example:

X	Y (Dataset I)	Y (Dataset II)	Y (Dataset III)	Y (Dataset IV)
10	8.04	9.14	7.46	6.58
8	6.95	8.14	6.77	5.76
13	7.58	8.74	12.74	7.71
9	8.81	8.77	7.11	8.84
11	8.33	9.26	7.81	8.47
14	9.96	8.1	8.84	7.04
6	7.24	6.13	6.08	5.25
4	4.26	3.1	5.39	12.5
12	10.84	9.13	8.15	5.56
7	4.82	7.26	6.42	7.91
5	5.68	4.74	5.73	6.89

All four datasets share identical:
Mean of X is 9 and Y is 7.50 for each dataset
Variance of X is 11 and Y is 4.13
Correlation (r = 0.816)

However, the data distributions and relationships between X and Y differ drastically when visualized.



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)

<Your answer for Question 8 goes here>
Pearson's R is a numerical summary of the strength of the linear association between two variables. It quantifies the strength and direction of the linear relationship between two continuous variables.
The value of Pearson's R ranges from -1 to +1
Value of +1 indicates a perfect positive linear relationship. As one variable increases, the other variable also increases in same proportion.
Value of 0 indicates there is no linear relationship. The variables do not have a linear correlation.
Value of -1 indicates a perfect negative linear relationship. As one variable increases, the other variable decreases in same proportion.

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)

<Your answer for Question 9 goes here>
Scaling is the process of adjusting the range of features in your dataset so that they are on the same scale. It is data processing step in machine learning and statistics used to standardize the

value of numerical features. Scaling makes sure that all features contribute equally to the model's learning process.

Scaling needs to be performed, for following reasons:
[1] Improve model performance.
[2] Avoid bias due to differing units.
[3] Prepares data for models that assumes specific scale.

Difference between normalized scaling and standardized scaling.
[1] Normalization transforms the features so that they are within a specific range, usually between 0 and 1, or -1 and +1. Standardization transforms the data to have a mean of 0 and a standard deviation of 1.
[2] Normalization rescales the data to a specific range [0 to 1, -1 to +1] and is useful when you need bounded data. Standardization centers data around 0 with standard deviation of 1 and is best suited for data that follows normal distribution.
[3] Normalization can be affected if there are outliers in the dataset due to the range can be drastically altered by extreme values. Standardization does not bound the data within a specific range, which could be problematic if you need to scale feature to a particular interval.

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)

<Your answer for Question 10 goes here>
VIF (Variance Inflation Factor) helps explain the relationship of one independent variable with all other independent variable. A very high VIF values shows a perfect correlation between two independent variables.
When we observe VIF [Variance Inflation Factor] of a feature is infinite, it indicates issue related to multicollinearity in your dataset.

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)

<Your answer for Question 11 goes here>
A Q-Q plot is a graphical tool used to assess whether a dataset follows a specified theoretical distribution, as normal distribution. In Q-Q plot, the quantiles of the data are plotted against the quantile of the expected distribution.
Use of Q-Q plot in linear regression: In linear regression, the residuals are assumed to be normally distributed. A Q-Q plot is a diagnostic tool to check whether the residual follow this assumption. By examining a Q-Q plot you can assess whether your regression model's residual key assumptions, and if not, take appropriate steps to addresses any issue.