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 my analysis considering 'cnt' as the target variable, below are my inferences:

1) Fall season has more number of bookings
2) Bookings has been significantly increased in the year 2019 when compared to 2018
3) Bookings are more in the month of May, Jun, Jul, Aug, Sep, Oct.
4) Rentals are higher on working days compared to non-working days.
5) Clear weather is associated with higher bike rentals.

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 during dummy variable creation is important to avoid perfect multicollinearity and ensure that the regression model can be estimated and interpreted correctly. i.e., incase if we have 'n' categories, we need to create 'n-1' dummy variables to avoid perfect multicollinearity

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 and **atemp** has the highest 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)

Linear regression model can be validated on the basis of :

There is a linear relationship between X and Y:

- ➔ X and Y should display some sort of a linear relationship; otherwise, there is no use of fitting a linear model between them.

Error terms are normally distributed with mean zero(not X, Y):

- ➔ There is no problem if the error terms are not normally distributed if you just wish to fit a line and not make any further interpretations.
- ➔ The assumption of normality is made, as it has been observed that the error terms generally follow a normal distribution with mean equal to zero in most cases.

Error terms are independent of each other:

➔ The error terms should not be dependent on one

Error terms have constant variance (homoscedasticity):

➔ The variance should not increase (or decrease) as the error values change.
Also, the variance should not follow any pattern as the error terms change.

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)

The top 3 features are :

a) Temperature
b) Season
c) Year

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 aims to model the relationship between a dependent variable (also called the target variable) and one or more independent variables. If the independent variable is only one we call it as linear regression and it is mathematically represented as :

$Y = mX+c$

m is slope
X is independent variable
c is intercept

If there are more than one input variable i.e., if there are more than one independent variables we call it as multi linear regression.

The equation is

$Y = \beta0+\beta1x1+\beta2x2+\cdots+\beta nxn$

So the regression line depends on the independent variables. A regression line can be positive regression line or negative regression line

---

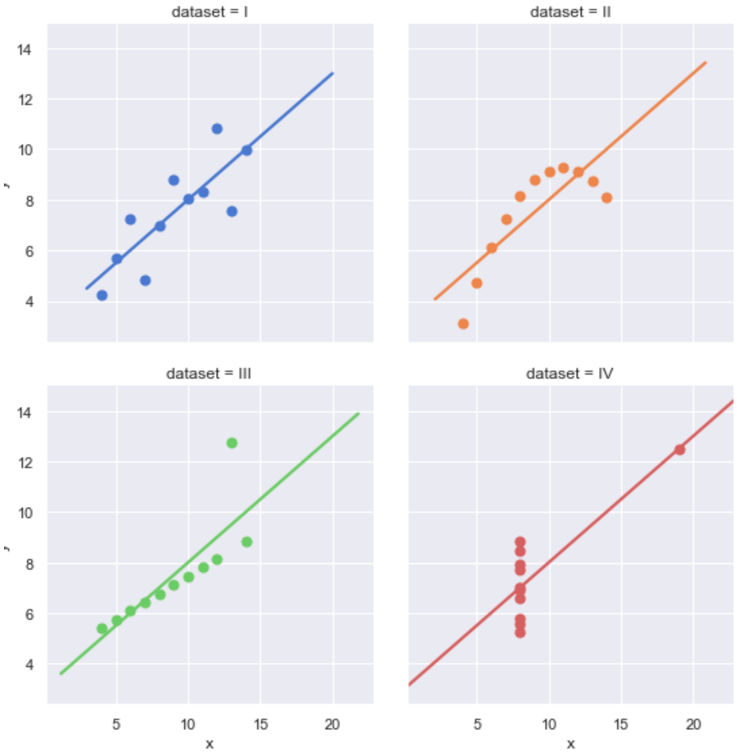
Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Anscombe's Quartet is a famous set of four datasets that were created by the statistician Francis Anscombe to demonstrate the importance of visualizing data before drawing conclusions.

These datasets have nearly identical simple descriptive statistics (mean, variance, correlation, etc.), their visualizations reveal very different patterns and relationships between the variables.

The quartet reinforces the idea that data analysis is not just about computing numbers, but about interpreting those numbers in the context of the data's structure, distribution, and behavior.

Example plot of Anscombe's Quatret is :



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, also known as Pearson's correlation coefficient or simply correlation coefficient, is a statistical measure that quantifies the linear relationship between two variables. It ranges from -1 to 1 and is used to assess the strength and direction of the relationship between two continuous variables.

Formula is : **$r = (n * \Sigma(xy) - (\Sigma x)(\Sigma y)) / \sqrt{([n * \Sigma(x^2) - (\Sigma x)^2] * [n * \Sigma(y^2) - (\Sigma y)^2])}$.**

We can make the following interpretations from r values:

If r=1: A perfect positive linear relationship.
If r=−1: A perfect negative linear relationship
If r=0: No linear relationship
If 0 < r <1: A positive linear relationship, but the relationship is not perfect.
If −1<r<0: A negative linear relationship, but the relationship is not perfect.

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 refers to the process of transforming the features of a dataset into a specific range or distribution.

There are two major methods to scale the variables

- ➔ standardisation : Standardisation basically brings all of the data into a standard normal distribution with mean zero and standard deviation one.
- ➔ MinMax scaling: brings all of the data in the range of 0 and 1. The formulae in the background used for each of these methods are as given below:

•Standardisation: x=(x−mean(x))/sd(x)
•MinMax Scaling: x=(x−min(x))/(max(x)−min(x))

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 calculates how well one independent variable is explained by all the other independent variables combined

The VIF becomes infinite when there is perfect multicollinearity (i.e., one variable can be exactly predicted from others), perfect correlation between variables, or a lack of variability in the data

---

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)

The Q-Q plot compares the quantiles (percentiles) of the observed data with the quantiles of a specified theoretical distribution.

By visually inspecting the plot, we can determine whether the residuals follow a normal distribution and identify potential problems such as outliers.

This helps ensure that the regression model is appropriate and that the results are reliable.