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)

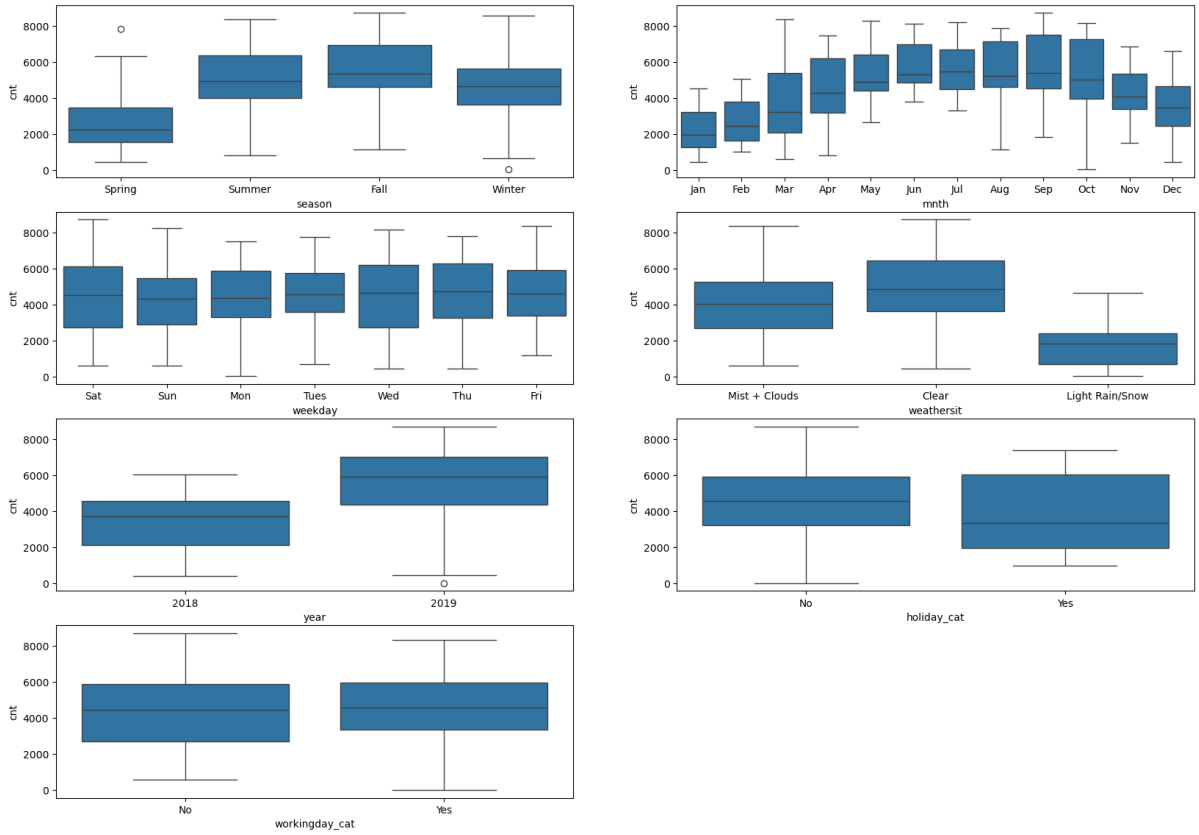
Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

The dependent variable is cnt (count of total rental bikes). We see that cnt is higher when:

1. weather is clear
2. season is Fall or Summer
3. months from May to Oct
4. increases in 2019 compared to 2018

In case of weekday, we find that the mean is almost similar (approx. 4,500) but the variance is more on Wednesday and Saturday.



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)

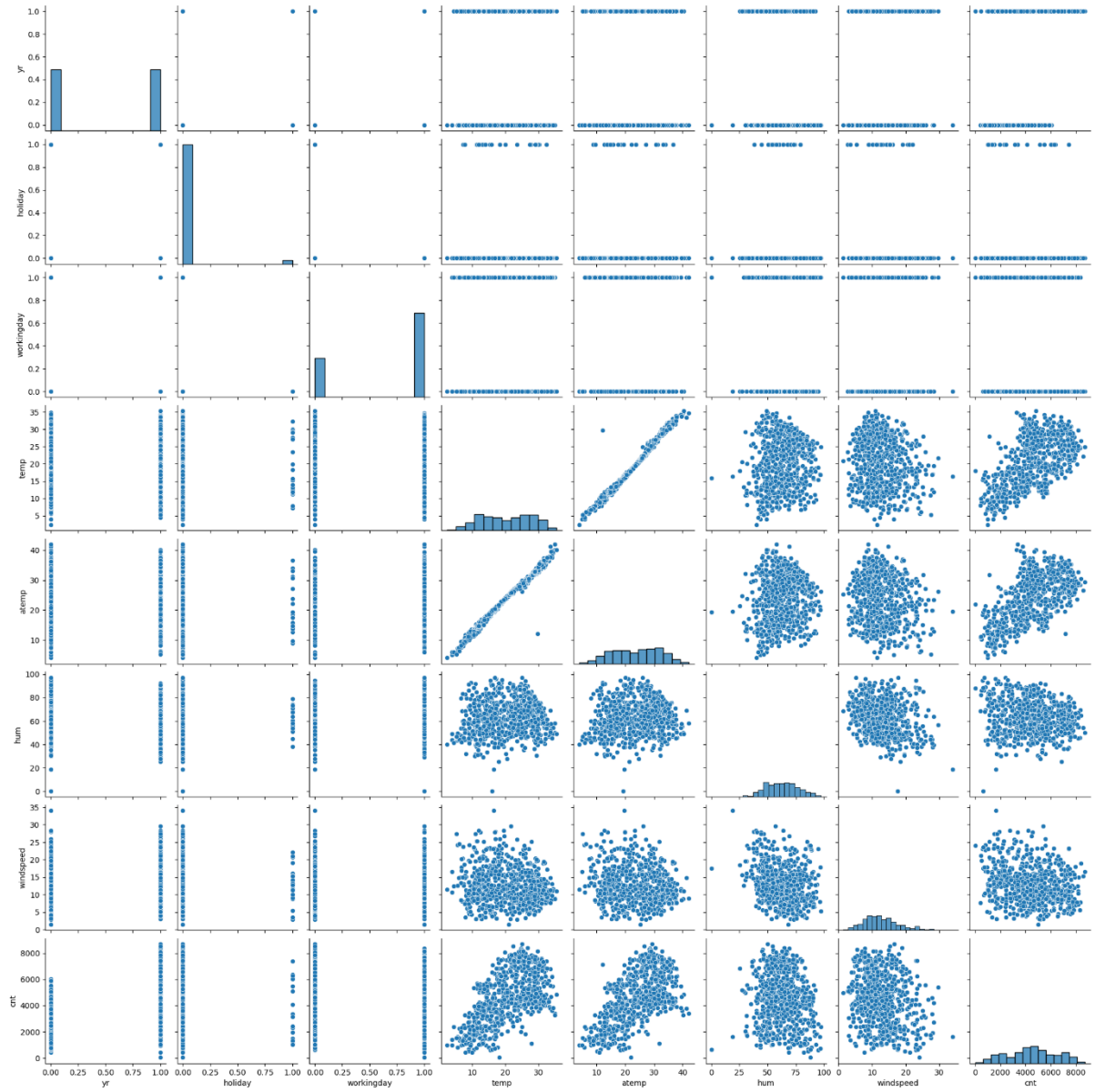
By using drop_first=True, we reduce the number of predictors by dropping the first dummy variable. It is important to use drop_first=True during dummy variable creation to prevent multicollinearity. Multicollinearity (where predictor variables are highly correlated) can lead to unstable models and that is something we want to avoid.

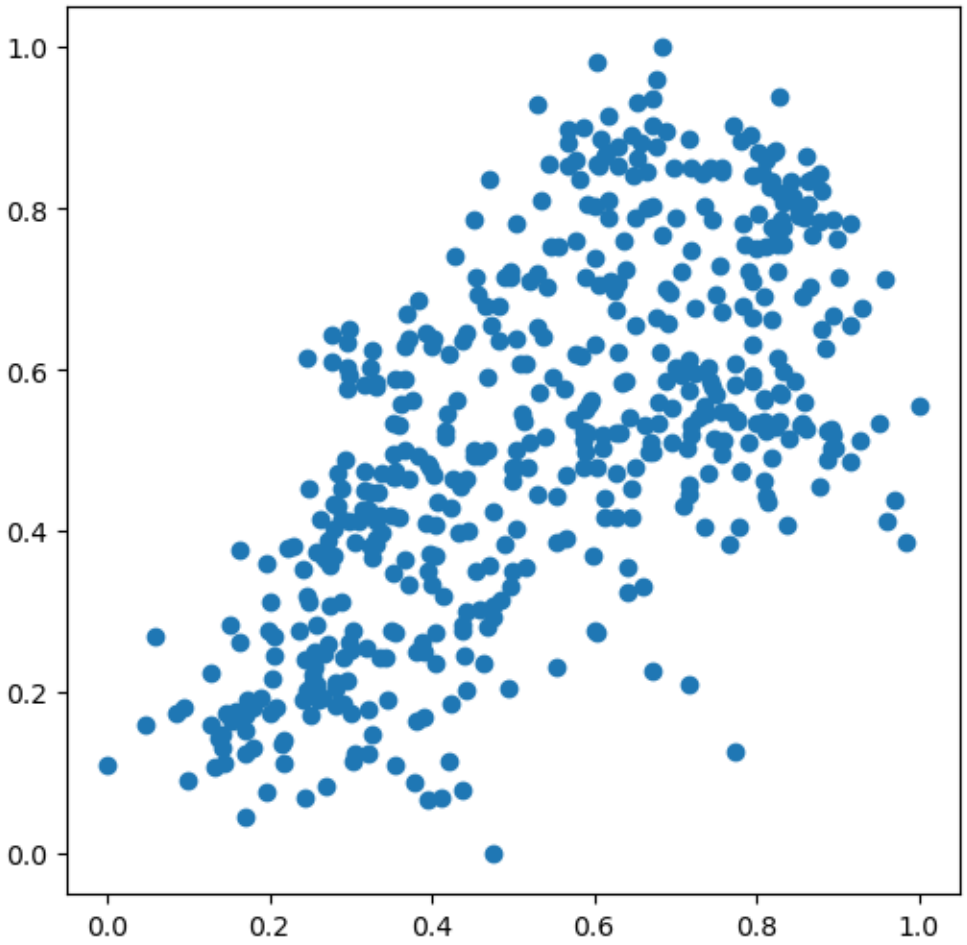
Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

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 the target variable (cnt). atemp (which itself is highly correlated to temp) also has high correlation with cnt with but as the question asks one variable, I'll answer temp.





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)

I validated the No multicollinearity assumption of Linear Regression, by checking the VIF between the independent variables. It is less than 5 for all of them.
I validated the normality assumption of Linear Regression by plotting the distribution of the error terms. It was a normal distribution with a mean value of 0.
There was no Homoscedasticity and no autocorrelation as well.

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 contributing to the demand of the shared bikes are:

1. temp – what is the temperature. As the temperature increases, so does the demand.
2. Light Rain/Snow – If there is light rain or light snow. It had a negative coefficient, so as

the rain or snow increased, the demand went down.

3. Yr – which year is it. The demand was higher in 2019 compared to 2018.

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 supervised learning algorithm which aims to model a dependent variable, y as a function of some independent variables, X_i, by finding a line that best fits the data. In general, we assume y to be numeric and each of X_i can be numeric or categorical.

The equation for linear regression is: $y = \beta_0+\beta_1X_1+\beta_2X_2+...+\beta_iX_i$

In the above equation:
y is the dependent variable that we are trying to predict.
X_i are the independent variables, which our model uses to predict y.
β_i are the coefficients of the regression model.

Assumptions for Linear Regression Model
The various assumptions underlying linear regression are as under:
Linearity: The deterministic component of the regression model is a linear function of the separate predictors.
Independence of Errors: The errors from our model are independent.
Homoscedasticity: The errors from our model have equal variance.
Normality of Errors: The errors from our model are normally distributed.

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 group of 4 data sets that are almost identical in terms of their descriptive statistics, but are very different once these data sets are plotted with a regression model. All four of them have very different distributions so they look completely different from one another when they are visualized on scatter plots. They are provided below (image found by googling)

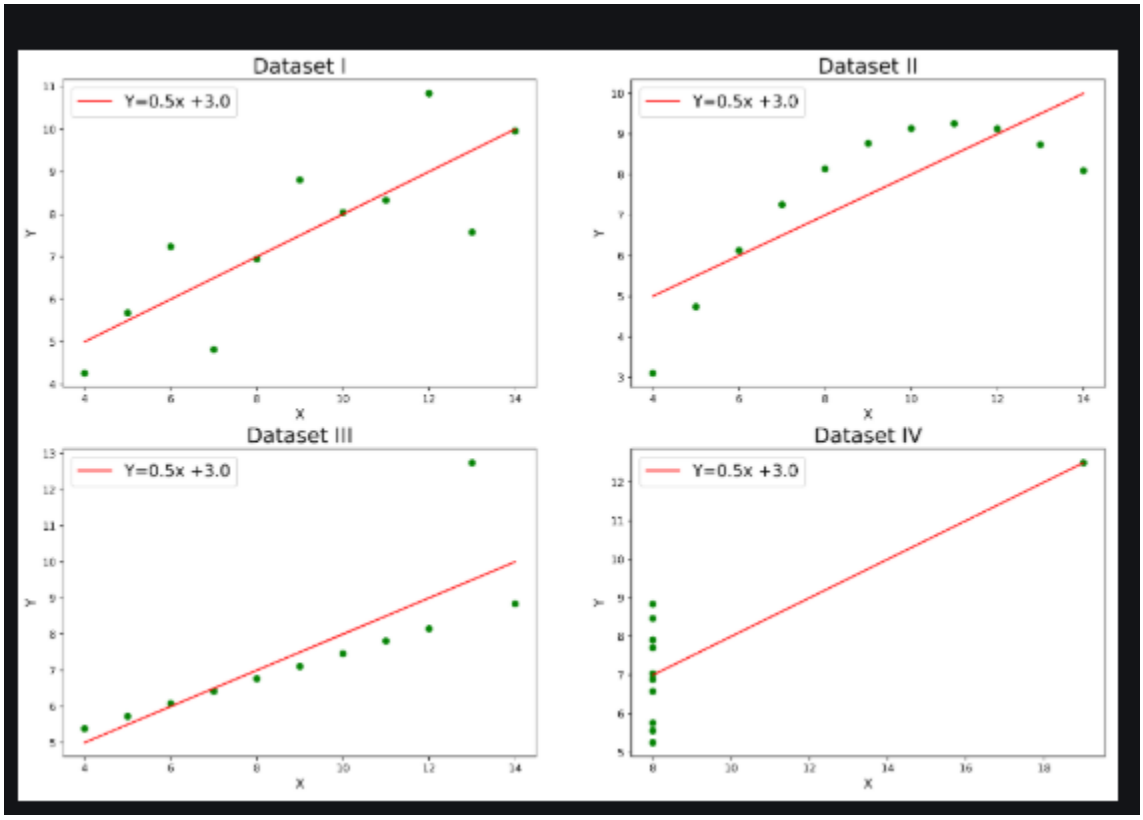
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

The descriptive statistics for them are:

	I	II	III	IV
Mean_x	9.000000	9.000000	9.000000	9.000000
Variance_x	11.000000	11.000000	11.000000	11.000000
Mean_y	7.500909	7.500909	7.500000	7.500909
Variance_y	4.127269	4.127629	4.122620	4.123249
Correlation	0.816421	0.816237	0.816287	0.816521
Linear Regression slope	0.500091	0.500000	0.499727	0.499909
Linear Regression intercept	3.000091	3.000909	3.002455	3.001727

Looking at the summary statistics may lead one to believe they are essentially the same data

The scatter plot along with the linear regression model for these is as shown below:



Basically, Anscombe's quartet demonstrates the necessity of combining statistical analysis with graphical exploration for good data interpretation. The descriptive statistics of datasets may seem to be similar; it is the accompanying visualizations which reveal distinct patterns.

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 correlation coefficient, or Pearson's R, is a metric for assessing linear relationships between two variables. It has a value between –1 to 1, which indicates both the magnitude and direction of the correlation.

If the value is 1, it indicates a perfect positive linear relationship, meaning an increase in one variable corresponds to an increase in the other. A value of -1 indicates a perfect negative linear relationship, meaning an increase in one variable corresponds to a decrease in the other. A value of 0 indicates there is no linear relationship between the two variables.

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 to standardize the independent features present in the data, which is performed during the data preparation or pre-processing step. If scaling is not done, the Linear regression algorithm will return higher coefficient for variables with higher value and lower coefficients for variables with as lower value. This may distort the model or our understanding of the model.

In normalized scaling, we first find the mean, minimum and maximum values for the variable / column. Then we subtract the mean value from each value and divide it by the difference between the maximum and the minimum value. This method is still prone to outliers as we are using the maximum and the minimum.

In Standarized scaling, we first find the mean and standard deviation of the variable / column. Then we subtract the mean from each value and divide the result by the standard deviation. This provides data with a mean of 0 and a standard deviation of 1. This method is better for handling 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)

The value of VIF can become infinite if there is prefect correlation between one independent variable and a linear combination of the other independent variables in the regression model, which basically implies that this particular independent variable can be perfectly predicted using the other 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)

A Q-Q plot or quantile-quantile plot is a scatterplot which is used to compare two sets of data to determine if they come from the same distribution. They are used in linear regression to validate the normality assumption about distribution of error terms.
