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 following are the categorical variables:

1. season
2. weathersit
3. holiday
4. mnth
5. yr
6. Weekday

The following inferences can be drawn:

- Bike sharing is the least in the spring season.
- The count variable is less during the holidays.
- The demand does not give a clear picture of whether it is a working day or a holiday.
- The demand increases for good weathersit drastically.
- Demand for rental bikes is increasing till the month of June, then there is a fallback of demand. For the month of September, demand is highest for the year and then demand for rental bikes again decreases.
- Demand for bike rent decreases in the year beginning and end. This may be due to bad weather conditions. Also, the number of rentals is more in 2019 than in 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)

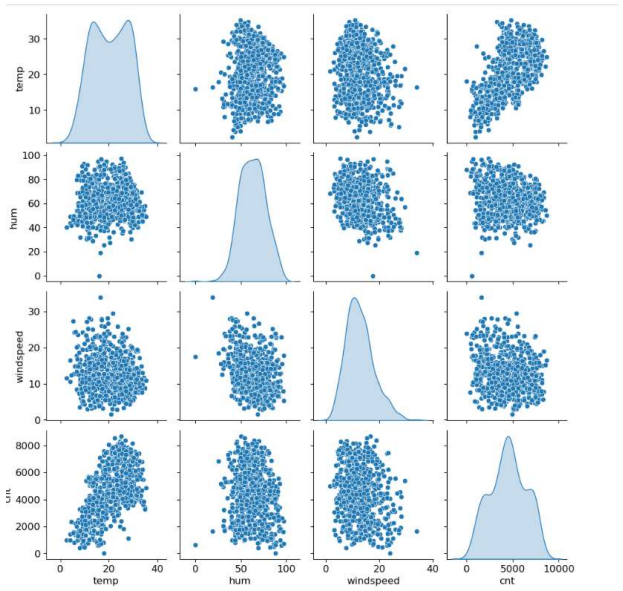
It reduces the extra variables created during the creation of the dummy variables. This implies it also reduces the correlation created between the dummy variables.

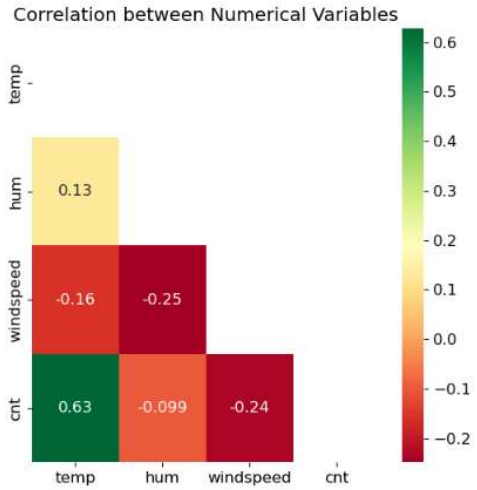
Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer: <Your answer for Question 3 goes below this line> (Do not edit)

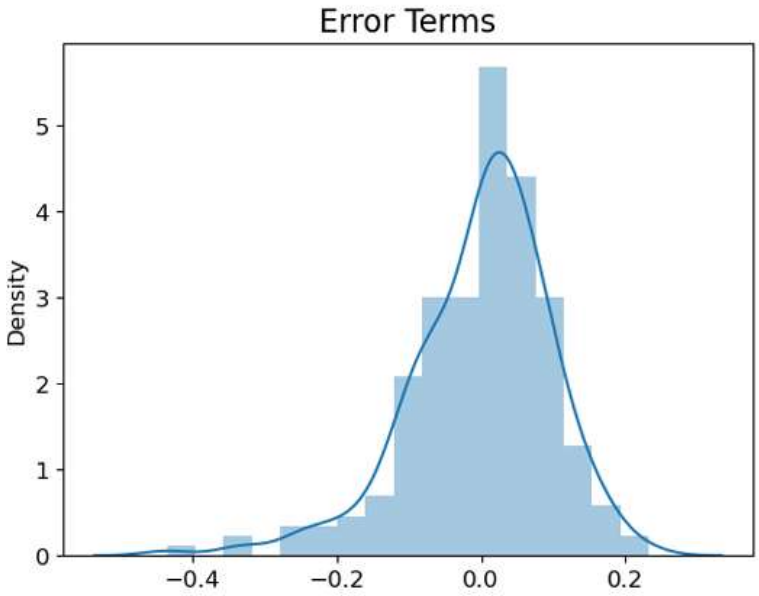
temp variable was having the highest correlation with the target variable ‘cnt’ (0.63).





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)

To validate the assumptions, we should verify the residual to follow a normal distribution and mean = 0. Please find below the chart which verifies the normal distribution along with mean = 0



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)

Following are the top 3 features contributing significantly to the demand for shared bikes in our final model:

1. temp is directly proportional with the coefficient of 0.552
2. yr is directly proportional with the coefficient of 0.256
3. weathersit_LightSnow_LightRain is inversely proportional with the coefficient of -0.264

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 in Machine Learning is a statistical regression method which can be used to predict the analysis and visualize the relationship between the continuous variables.

It is based on the equation: y = mx + c
where, m = gradient c is the intercept of the y-axis Regression tries to find the best-fit line between the dependent and the predicted variables with minimal error.

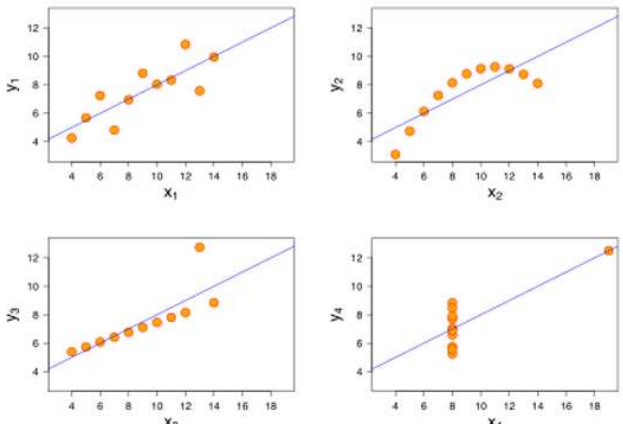
It shows the linear relationship between the dependent variable (y-axis) and the independent variable (x axis).

It can be broadly divided into two types:
1. Simple Linear Regression – When the dependent variable is predicted using only one independent variable.
2. Multiple Linear Regression – When the dependent variable is predicted using multiple independent variables.

Some use cases: We can use linear regression to predict the following:
- Predict the sales target for a company.
- Predict the scores of a student.
- Predict the change in stock price etc.

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 is a group of four datasets which are identical statistical features but different distribution and looks different when we try to plot in a scatter plot. It was developed by Francis Anscombe. It helps us illustrate the importance of plotting the graph before analyzing the model. The four types of charts are:



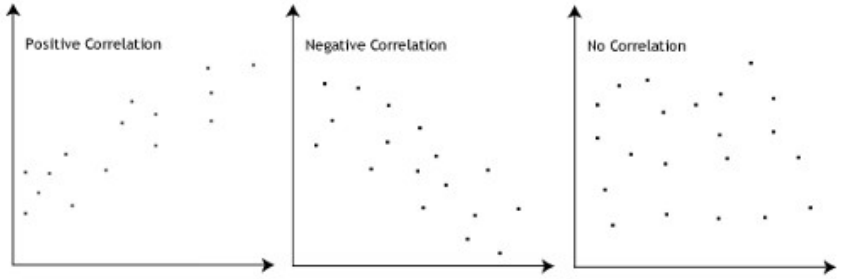
1. One chart appears to be having a simple linear relationship.
2. The second chart could not fit the linear regression model and shows it as non-linear.
3. The third chart shows the outliers involved.
4. The fourth chart also shows the outliers involved with one high leverage point, which produces a 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)

<Your answer for Question 8 goes here>

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



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 a method which is used to normalize or standardize some independent variables of the data set. Scaling is performed at the pre-processing stage so that we can deal with the varying values in the entire dataset. Else if the units of the values are different and not standardized then it tends to give higher values for higher numbers and lower values for lower numbers.

Normalized scaling brings all the data in the range of 0 and 1. Minmaxscaler helps implement normalized scaling whereas standardized scaling replaces the values with z scores. One disadvantage of normalized scaling is it misses out on the outliers as it ranges from 0 to 1.

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>

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

When the value of VIF is infinite, it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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>

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

Use of Q-Q plot:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q plot:

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.
