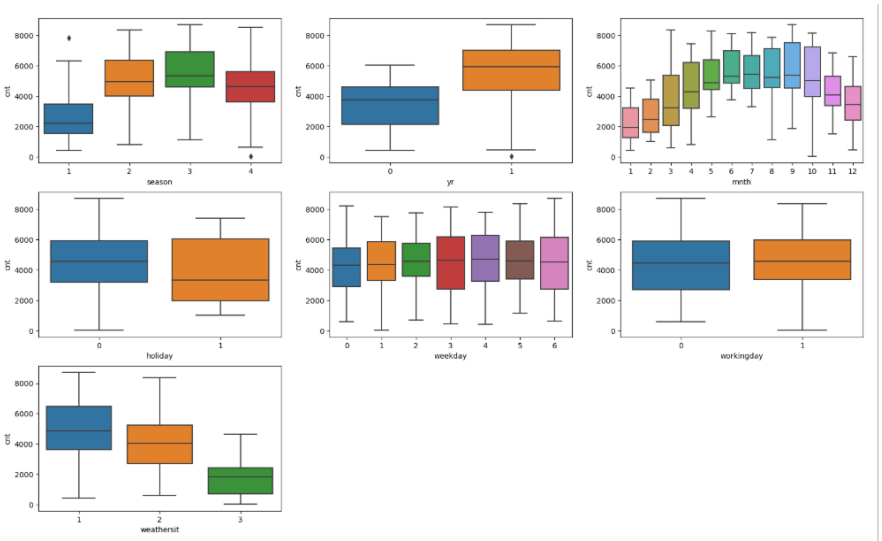
Assignment-based Subjective Questions

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)

There are few categorical variables in the dataset – season, yr, month, holiday, weekday, workingday, weathersit. These categorical variables have a major effect on the dependent variable cnt.



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 to use as it helps in reducing the extra column created during dummy variable creation.

Let's say we have 5 levels, drop_first=True will drop first level and 4 dummy variable will be created.

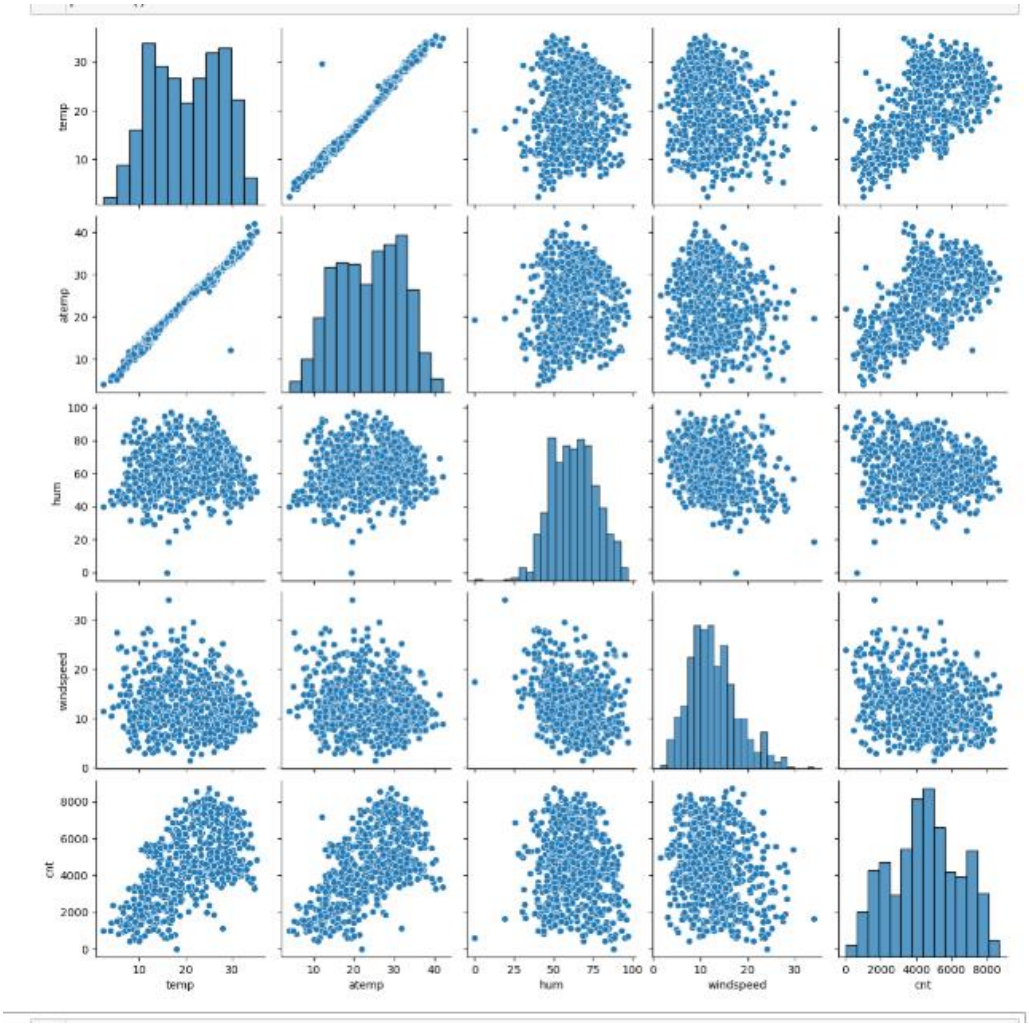
Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

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

There are few numerical variables in the dataset – temp, atemp, hum, windspeed, cnt.

'temp' and 'atemp' variable have the highest correlation with 'cnt' target variable as compared to other variables as seen in the pair-plot chart.



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)

Assumptions to validate after building the Linear Regression Model on the training set-

a. Normality of error terms- Error terms should be normally distributed
b. Multicollinearity- There should be insignificant multicollinearity among variables.
c. Linearity- Relationship between the independent and dependent variables should be linear.
d. Homoscedasticity of Residuals- Spread of the errors should be relatively uniform, regardless of the value of the predictor.
e. Independence of Errors- 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 features contributing significantly towards explaining the demand of the shared bikes are-

a. Temp

b. Season
c. Humidity

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 machine learning algorithm. It is a Predictive model used for finding linear relationship between dependent and one or more independent variable.
It is used for regression problems. Linear Regression is used to find the line that best fits the data points on the plot.

Two types of Linear Regression-

1. Simple Linear Regression:

Simple Linear Regression is having only one input/independent variable
It determine the value of one dependent variable from value of one given independent variable.
Y= mX + c
Here, Y is dependent variable
X is independent variable
c is y-intercept
m is coefficient

2. Multiple Linear Regression-

Multiple Linear Regression is having twi or more input/independent variable
It determine the value of one dependent variable from value of two or more independent variable.
Y=m1X1 + m2X2 + m3X3+….+mnXn + c
Here, , Y is dependent variable
X1, X2, X3..Xn are independent variable
c is y-intercept
m1,m2,m3…mn are coefficient

Assumptions of Linear Regression-
1. Linearity
2. Independence
3. No Multicolinearity
4. Normality
5. Homoscadesticity >

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, comprising four datasets with nearly identical summary statistics, underscores the limitations of relying solely on numerical metrics.

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 datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

Purpose of Anscombe's Quartet

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone. >

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's correlation is also referred to as Pearson's R, the Pearson product-moment correlation coefficient or bivariate correlation. It is a statistic that measures the linear correlation between two 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 two variables. A value greater than 0 indicates a positive association i.e., as the value of one variable increases, value of other variable increases.

A value less than 0 indicates negative association i.e., as the value of one variable increases, value of other variable decreases. >

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 dataset in a fixed range. Feature Scaling is done to handle highly varying magnitudes or values. If feature scaling is not done, then a machine learning algorithm tends to weight greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Normalized Scaling-

a. Minimum and Maximum value of features are used for scaling.
b. It is used when features are of different scales
c. Scales values between [0,1] or [-1,1].
d. It is really affected by outliers.
e. Scikit-learn provides a transformer called MinMaxScaler for normalization.

Standardized Scaling-

a. Mean and standard deviation is used for scaling.
b. It is used when we want to ensure zero mean and unit standard deviation.
c. It is not bounded to a any range.
d. It is much less affected by outliers.
e. Scikit-learn provides a transformer called StndardScaler for standardization.

>

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)

<
When multicollinearity is perfect, the VIF tends to infinity. A large value of VIF indicates that there is a correlation between the variables.
If the VIF is 5, it means that the variance of the model coefficient is inflated by a factor of 5 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 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)

<

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

Use of q-q plot-

A q-q plot is a plot of quantiles of the first dataset against the quantiles of the second dataset.

30% or 0.3 quantile is the point at which 30% if the data fall below and 70% fall above that value.

Importance of q-q plot-

1. Assessing Normality of Residuals-
   Linear regression assumes that the residuals (differences between observed and predicted values) are normally distributed.
   A Q-Q plot helps visualize if the residuals follow a normal distribution. In a Q-Q plot:
   The x-axis represents the theoretical quantiles from a normal distribution.
   The y-axis represents the observed quantiles from the data.
   If the residuals are normally distributed, the points should lie approximately along a straight 45-degree line.
2. Identifying Deviations from Normality
   Systematic deviations from the straight line in the Q-Q plot can indicate:
   Heavy tails (leptokurtosis): Points curve away at both ends, suggesting extreme residuals.
   Light tails : Points are closer to the line than expected.
   Skewness: Points systematically deviate upward or downward from the line.
   These deviations can signal issues with the model or the data, such as: Outliers.
   Non-linear relationships.
   Incorrectly specified models. >