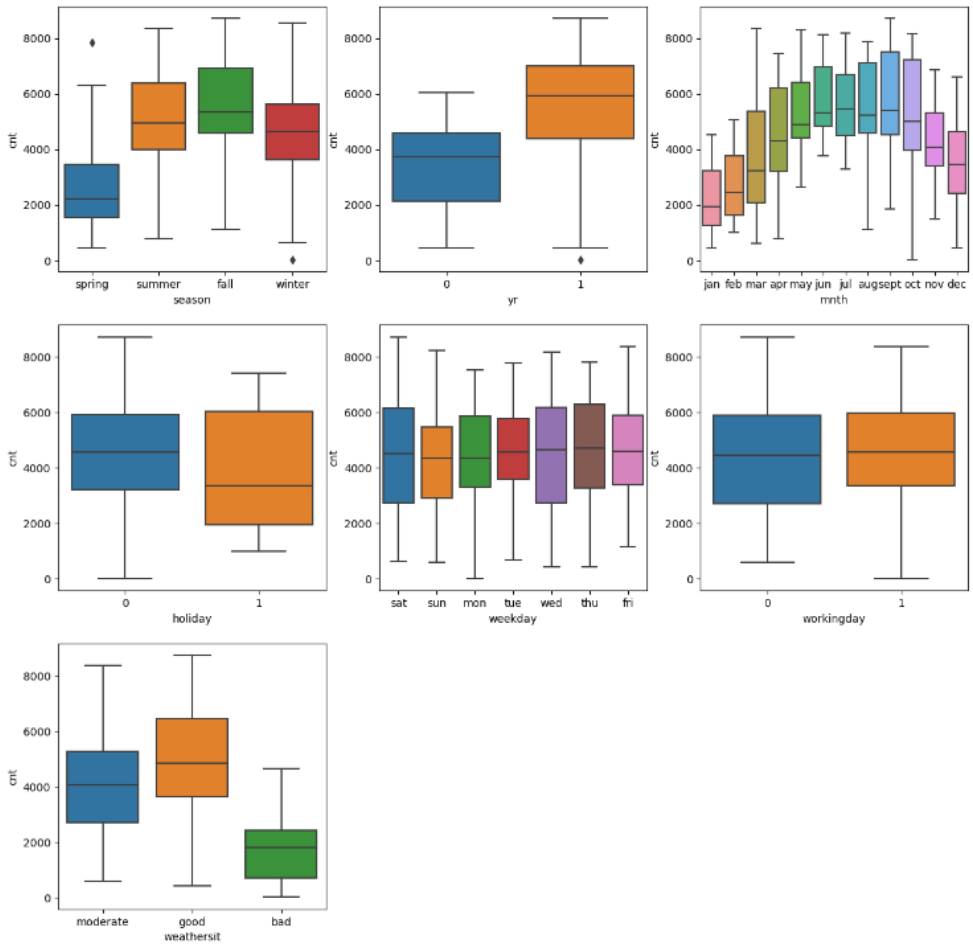
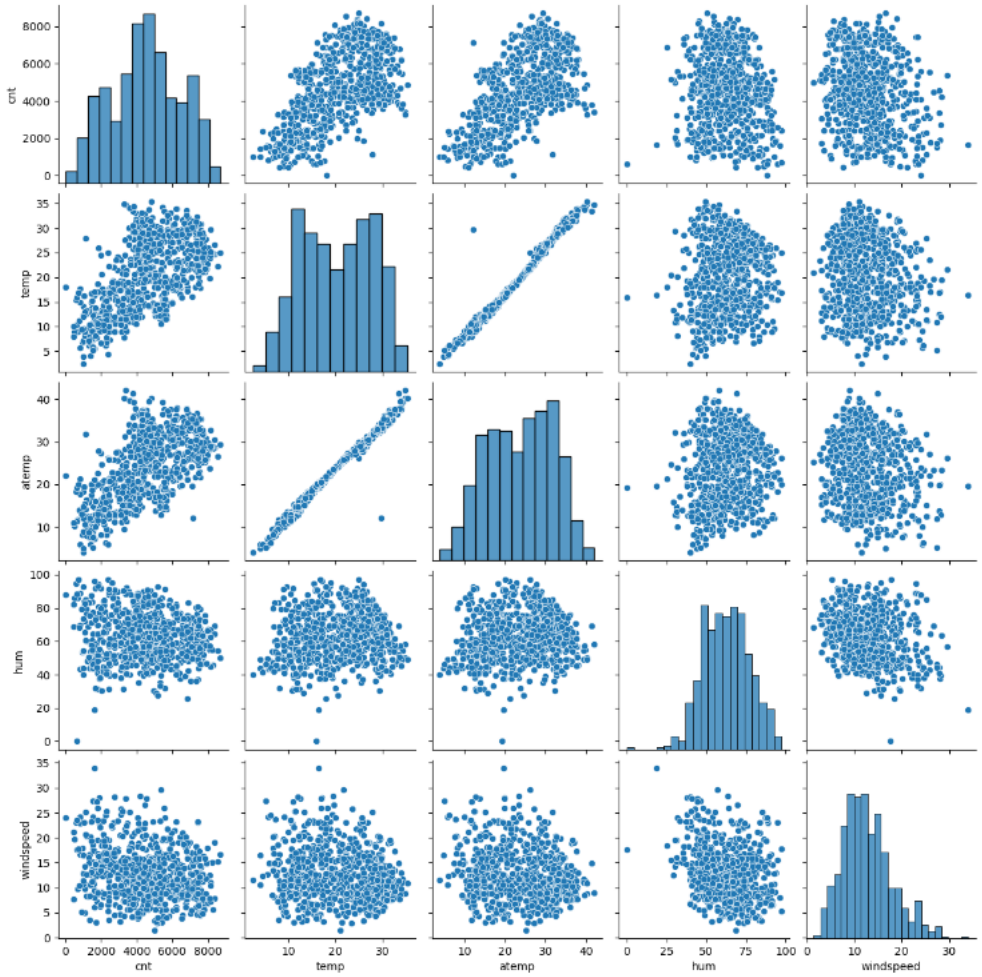
Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

    - These categorical variables have a major effect on the dependent variable 'cnt'. The below fig shows the correlation among the same



2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**
    - The intention behind the dummy variable is that for a categorical variable with 'n' levels, you create 'n-1' new columns each indicating whether that level exists or not using a zero or one. Hence drop_first=True is used so that the resultant can match up n-1 levels. Hence it reduces the correlation among the dummy variables.
        - Eg: If there are 3 levels, the drop_first will drop the first column.
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**



- temp and atem has the highest corelation with the target variable cnt
- temp and atemp are highly co-related with each other

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**
    - Linear Regression models are validated based on Linearity,No auto-correlation,Normality of error,Homoscedasticity, Multicollinearity.
5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**
    - Top 3 features that has significant impact towards explaining the demand of the shared bikes are temperature,year and season

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**

    - Linear regression is a statistical method for modeling the relationship between a dependent variable and one or more independent variables. In simple linear regression, we try to find the best-fit straight line that explains the relationship between two variables, while in multiple linear regression, we use more than one independent variable to explain the dependent variable.

    - The basic idea behind linear regression is to find a line of best fit that minimizes the sum of the squared differences between the predicted values and the actual values.

The equation of a straight line is y = mx + b, where y is the dependent variable, x is the independent variable, m is the slope of the line, and b is the y-intercept.

- To find the line of best fit, we need to find the values of m and b that minimize the sum of squared errors (SSE), which is the sum of the squared differences between the predicted values and the actual values. This can be done using the method of least squares.
- Here are the steps to perform linear regression:

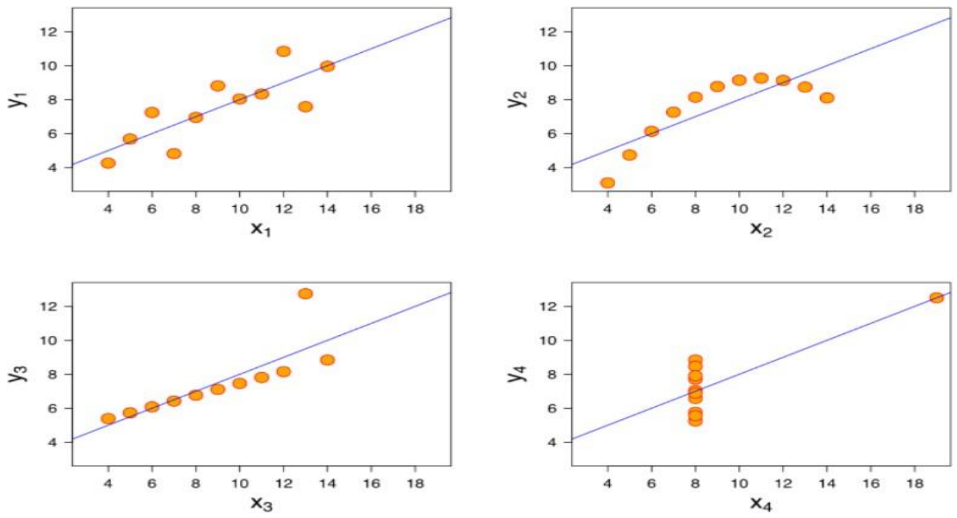
    1. Collect and organize the data: Gather the data you want to analyze and organize it into a table or spreadsheet with columns for the dependent variable and one or more independent variables.
    2. Plot the data: Create a scatter plot to visualize the relationship between the dependent and independent variables. This will help you to determine if there is a linear relationship between the variables.
    3. Calculate the correlation coefficient: Calculate the correlation coefficient between the dependent and independent variables to determine the strength and direction of the relationship. If the correlation coefficient is close to 1 or -1, it indicates a strong positive or negative linear relationship.
    4. Choose the model: Choose the appropriate model for the data. If there is only one independent variable, use simple linear regression. If there are multiple independent variables, use multiple linear regression.
    5. Estimate the parameters: Use the method of least squares to estimate the values of the parameters that minimize the sum of squared errors. For simple linear regression, these are the slope and y-intercept. For multiple linear regression, these are the coefficients for each independent variable and the y-intercept.
    6. Evaluate the model: Evaluate the model by calculating the coefficient of determination (R^2), which is the proportion of the variance in the dependent variable that is explained by the independent variables. A high R^2 value indicates a good fit.
    7. Make predictions: Once you have a good fit for the data, you can use the model to make predictions about new data. Simply plug the values of the independent variables into the equation of the line and solve for the dependent variable.

- In summary, linear regression is a statistical method for modeling the relationship between a dependent variable and one or more independent variables. It is used to find the line of best fit that minimizes the sum of squared errors, and can be used to make predictions about new data.

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots. It was constructed to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.There are these four data set plots which have nearly same statistical observations,

which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.



1. 1 st data set fits linear regression model as it seems to be linear relationship between X and y
2. 2 nd data set does not show a linear relationship between X and Y , which means it does not fit the linear regression model.
3. 3 rd data set shows some outliers present in the dataset which can't be handled by a linear regression model.
4. 4 th data set has a high leverage point means it produces a high correlation coeff.

- Its conclusion is that regression algorithms can be fooled so, it's important to data visualization before build machine learning model.

3. What is Pearson's R? (3 marks)

- In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variable

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling means you're transforming your data so that it fits within a specific scale. It is one type of data pre-processing step where we will fit data in specific scale and speed up the calculations in an algorithm. Collected data contains features varying in magnitudes, units and range. If scaling is not performed than algorithm tends to weigh high values magnitudes and ignore other parameters which will result in incorrect modeling.

Difference between Normalizing Scaling and Standardize Scaling:

- In normalized scaling minimum and maximum value of features being used whereas in Standardize scaling mean and standard deviation is used for scaling.
- Normalized scaling is used when features are of different scales whereas

standardized scaling is used to ensure zero mean and unit standard deviation.

- Normalized scaling scales values between (0,1) or (-1,1) whereas standardized scaling is not having or is not bounded in a certain range.
- Normalized scaling is affected by outliers whereas standardized scaling is not having any effect by outliers.
- Normalized scaling is used when we don't know about the distribution whereas standardized scaling is used when distribution is normal.
- Normalized scaling is called as scaling normalization whereas standardized scaling is called as Z Score Normalization.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**

- VIF(VarianceInflationFactor) basically helps explain the relationship of one independent variable with all the other independent variables. The formulation of VIF is given below: A VIF value of greater than 10 is definitely high, a VIF of greater than 5 should also not be ignored and inspected appropriately.
- A very high VIF value shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**

Q–Q plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data possibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.

QQ plot can also be used to determine whether or not two distributions are similar or not. If they are quite similar you can expect the QQ plot to be more linear. The linearity assumption can best be tested with scatter plots. Secondly, the linear regression analysis requires all variables to be multivariate normal. This assumption can best be checked with a histogram or a Q-Q-Plot.

Importance of QQ Plot in Linear Regression:

In Linear Regression when we have a train and test dataset then we can create Q-Q plot by which we can confirm that both the data train and test data set are from the population with the same distribution or not.

Advantages:

- It can be used with sample size also
- Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot

Q-Q plot use on two datasets to check

- If both datasets came from population with common distribution
- If both datasets have common location and common scale
- If both datasets have similar type of distribution shape

If both datasets have tail behavior