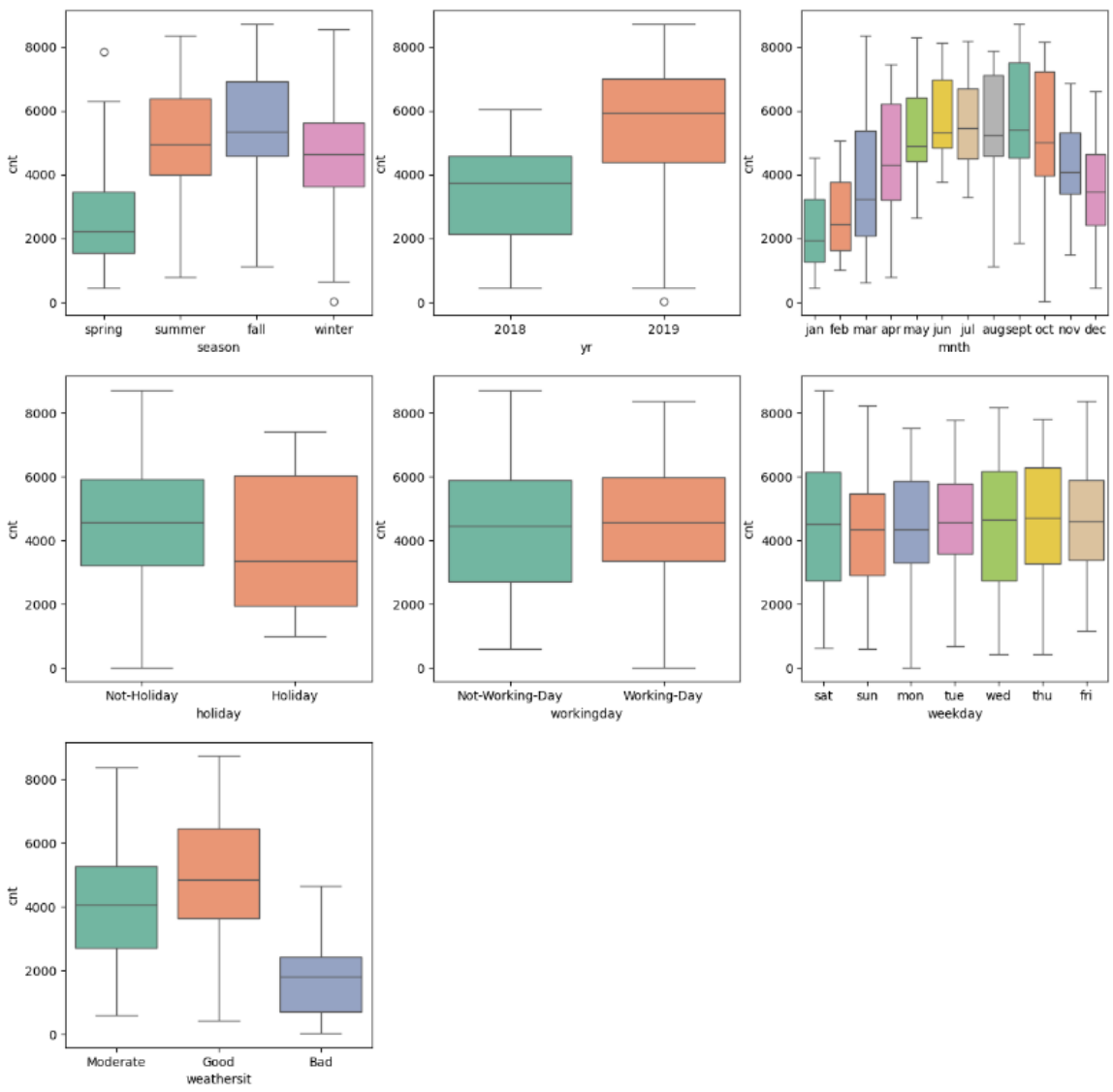
Assignment-based Subjective Question

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

   Answer:



   The insight of categorical variables effect on dependent variables is as follows,

1. Fall season had highest demand on rental bikes.
2. Compared to 2018, 2019 had highest demand. This shows there is a Growth ahead.
3. September month had highest demand for rental bikes. The demand is less in the on start and end of the year.
4. Rental bikes are demand on holidays and on not working days.
5. Weather had an impact on demand. its high on good weather.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

   Answer : drop_first=True option in pd.get_dummies is used to eliminate the dummy variable when categorical variables are transformed to binary variables. For example, if a column Season had 1 , 2 , 3 and 4 , pd.get_dummies would create three columns as season_2,season_3 and season_4. Where the first category 1 will be dropped. This is mainly to avoid multicollinearity. This help us create a cleaner set of dummy variables which can lead to better model performance.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

   Answer: Here the target variable is cnt and temp and atemp are highly correlated to target variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

   Answer:

   1. Residual analysis:
      - Errors are normally distributed here with mean 0
      - Actual and predicted follows the same pattern.
      - Error terms are randomly distributed, and no pattern can be drawn. Also they seems to be independent.

   2. R-Squared value:
      - R-squared value of test data which is 0.800 is very close to the train data which is 0.823. This consistency suggests that the model performs similarly on both the training set and the unseen test set.
      - The value of test data and train data are close which also indicates there is no overfitting.
   3. Plotting y_test and y_test_pred
      - This shows a linear relationship which confirms that the model correctly fitted
      - Prediction of test data is very close to actuals

   4. Validating Error terms of y_test & y_predicted

      - The Error terms are randomly scattered, and error terms are around 0 which indicates the model is performing well.
      - There is no outliers where the model is robust and will not overly influenced by extreme values.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer : The Top 3 features contributing significantly towards explaining the demand of the shared bikes are

1. temp
2. weathersit
3. yr

General Subjective Questions

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

Answer: Linear regression is a statistical regression algorithm used for predictive analysis and shows the relationship between one or more independent variables (predictors) and a dependent variable by fitting a linear equation to observed data. Linear regression shows independent variable which is on x-axis and dependent variable on y-axis. If there is only one input variable at x-axis such linear regression is called **Simple Linear regression.** If there is more than one input variable at x-axis it is called **Multiple Linear regression**. The Linear regression gives a slopped straight line describing the relationship between variables.

The Relationship in Linear regression can be represented as,

$$y=\beta_0 + \beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n+ \epsilon$$

where,

y = Dependent variable .

β_0 = y-intercept.

$\beta_1, \beta_2, \beta_3 .. \beta_n$ = Coefficients of independent variables.

$x_1, x_2, x_3 \ldots x_n$ = Independent variables (predictors)

ϵ = Error terms (The difference between predicted and actual errors)

The objective of Linear regression is to find the coefficients and minimize the difference between predicted and actual values of the dependent variable. This is done through OLS (Ordinary Least Square) method which will minimize the squared residuals (errors).

OLS is calculated as follows,

Residuals: The difference between the actual value y and the predicted value y^ (i.e., $y^= \beta_0 + \beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n$).

Sum of Squared Residuals (SSR)

$SSR = \sum(i = 1)^n (y - y^)^2$

Where n is the number of observations.

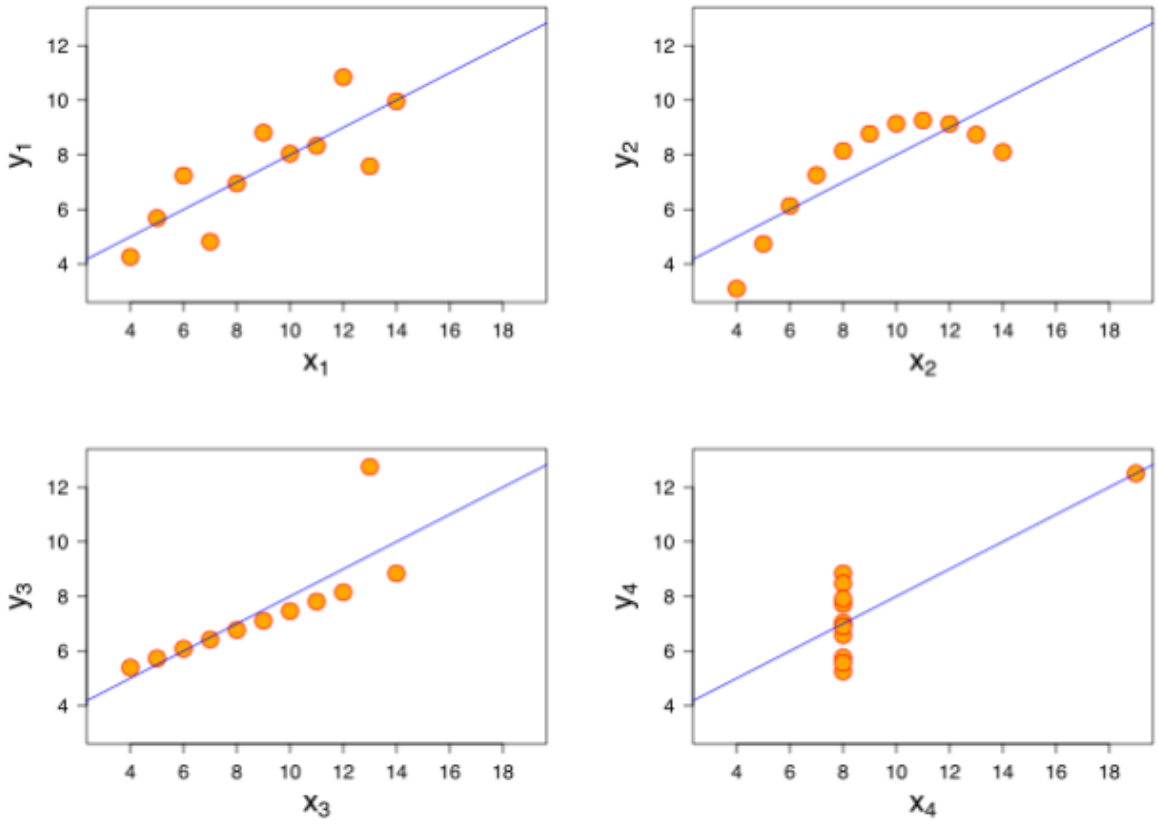
The OLS estimates are found by minimising the SSR.

For Linear regression to provide reliable estimates, few assumptions need to be met, and they are as follows,

1. **Linearity**: The relationship between independent to dependent should be Linear.
2. Independence: The observations are independent of each other.
3. **Homoscedasticity**: The residuals have constant variance at every level of the independent variable
4. **Normality**: The Residuals should be Normal.

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

Answer: 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. This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.



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

Answer: Pearson correlation coefficient (PCC), also referred to as Pearson's r or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations thus, it is essentially a normalized measurement of the covariance, such that the result is always has a value between −1 and 1.

The Pearson correlation coefficient r formula is as follows

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where ,

r = Correlation coefficient

X_i = Values of the x variable in the sample.

$\bar{x}$ = mean of the values of the x variable.

y_i = values of the y variable in the sample.

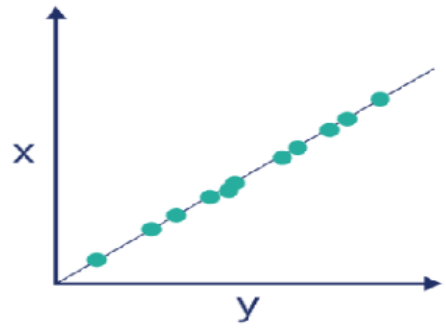
$\bar{y}$ = mean of the values of the y variable.

The Pearson's correlation coefficient varies between -1 and +1 where

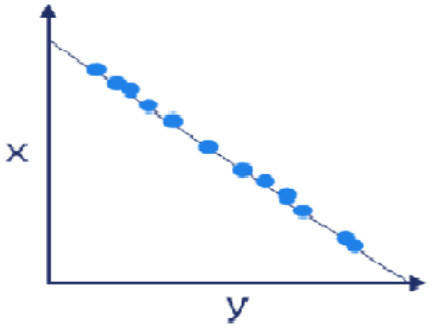
- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear correlation.

- r > .5 means there is a strong positive correlation.
- r < -.5 means there is a strong negative correlation.

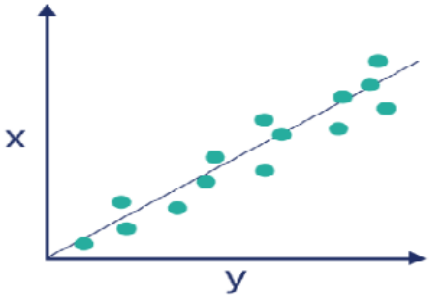
- r is between 0 to .3 there is a weak positive correlation.
- r is between 0 to -.3 there is a weak negative correlation.



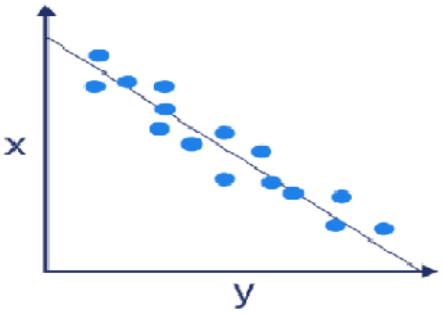
r = 1 positive correlation



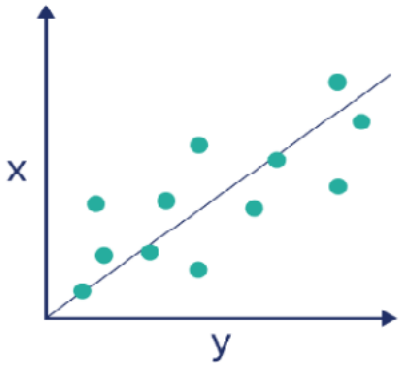
r = -1 negative correlation



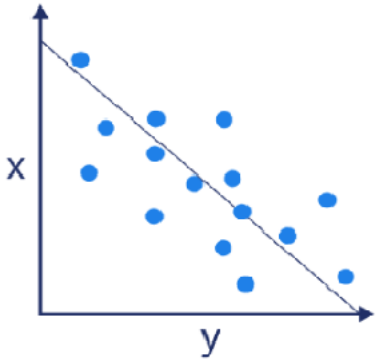
r>.5 Strong positive correlation



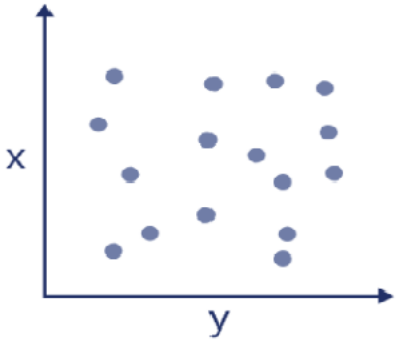
r< -.5 Strong negative correlation



.3 > r > 0 Weak positive correlation



0 > r > -.3 Weak negative correlation



r =0 no correlation

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

Answer: Scaling is a data preprocessing technique used to transform features in a dataset so that all variables are in a common scale. It is important because machine learning algorithms perform better when features are in same scales. When we collect data set it contains features highly varying in magnitudes, units, and range. If scaling is not done, then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we must do scaling to bring all the variables to the same level of magnitude.

Normalization/Min-Max Scaling: It brings all the data in the range of 0 and 1. This method is useful when you want to preserve the relationships between the data points while ensuring that all values are within a specific range.

Formula: **MInMaxScaling : x = (x-min(x)) / (max(x) – min(x))**

Standardization Scaling: Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

Formula: **Standardisation x = (x-mean(x)) / sd(x)**

The difference between Normalization and Standard scaling is Normalization is preferred when the dataset has no clear outliers and is bounded whereas Standardisation is typically better for data with outliers or when distribution is not uniform.

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

If Variance Inflation factor (VIF) value is infinite, this means that there perfect multicollinearity between two independent variables or the predictors. If this perfect multicollinearity happened, then R2 =1 which lead to 1/(1-R2) which is infinite. To avoid this, we need to drop one of the variable in the dataset which causes multicollinearity. Addressing infinite VIF is crucial for ensuring the stability of your regression model.

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

Answer: Quantile-Quantile (Q-Q) plot is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution. This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions. Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile.

The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45-degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line. If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.