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)**

   The following inferences are drawn about the effect on categorical variables on target variable,

   - Season:3: fall has highest demand for rental bikes
   - Demand is continuously growing each month till June. September month has highest demand. After September, demand is decreasing
   - When there is a holiday, demand has decreased.
   - Weekday is not giving clear picture about demand.
   - The clear weathershit has highest demand
   - During September, bike sharing is more. During the year end and beginning, it is less, could be due to extreme weather conditions.

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

   drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. If we do not drop one of the dummy variables created from a categorical variable then it becomes redundant with dataset as we will have constant variable(intercept) which will create multicollinearity issue.

   For example, there are four categories (spring, winter, summer, fall) defined within a feature (season). When the season is column is mapped into four dummy variables, when one variable is not mapped to spring, winter, summer, then it is obvious to be fall. On assigning **drop_first = True**, it will reduce this column 'fall', keeping remaining three columns (spring, winter, summer).

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

   'temp' variable has the highest correlation with the target variable.

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

   i. Multi-collinearity: After selecting the features using RFE and Manual method, VIF Is checked for them. And it is observed that none of them has a VIF greater than 5, hence there is no significant multi-collinearity among the independent variables used for predictions.

   ii. Normality of Error Terms: Plotted frequency distribution error terms, and it is visible that the residuals are normally distributed around mean = 0.

   iii. Homoscedasticity of Error Terms: On plotting the variance of the error terms, it is visible that variance is almost constant.

   iv. Independent Error terms: By plotting the residuals, it is observed that there is no pattern or trend among themselves. This shows they are independent of each other.

   v. Linear relationship: From the final regression equation, it is observed that target variable is linearly related to independent variable.

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 contributing significantly towards the demand of shared bikes are '"holiday", "temp" and season "hum".

General Subjective Questions

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

   Linear Regression Algorithm is a machine learning algorithm based on supervised learning. Linear regression is a part of regression analysis. Regression analysis is a technique of predictive modelling that helps you to find out the relationship between Input and the target variable.

   Linear regression is one of the very basic forms of machine learning where we train a model to predict the behavior of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

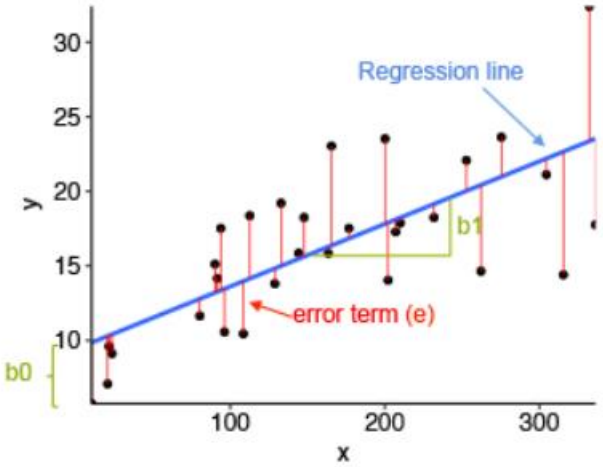
   Example for that can be let's say you are running a sales promotion and expecting a certain number of count of customers to be increased now what you can do is you can look the previous promotions and plot if over on the chart when you run it and then try to see whether there is an increment into the number of customers whenever you rate the promotions and with the help of the previous historical data you try to figure it out or you try to estimate what will be the count or what will be the estimated count for my current promotion this will give you an idea to do the planning in a much better way about how many numbers of stalls maybe you need or how many increase number of employees you need to serve the customer. Here the idea is to estimate the future value based on the historical data by learning the behavior or patterns from the historical data.

   In some cases, the value will be linearly upward that means whenever X is increasing Y is also increasing or vice versa that means they have a correlation or there will be a linear downward relationship.

   One example for that could be that the police department is running a campaign to reduce the number of robberies; in this case, the graph will be linearly downward.

   Linear regression is used to predict a quantitative response Y from the predictor variable X.

   Mathematically, we can write a simple linear regression equation as follow y ~ b0 + b1*x Where y is the predicted variable (dependent variable), b1 is slope of the line, x is independent variable, b0 is intercept(constant). It is cost function which helps to find the best possible value for m and c which in turn provide the best fit line for the data points.
   Here, x and y are two variables on the regression line.
   b1 = Slope of the line.
   b0 = y-intercept of the line.
   x = Independent variable from dataset
   y = Dependent variable from dataset

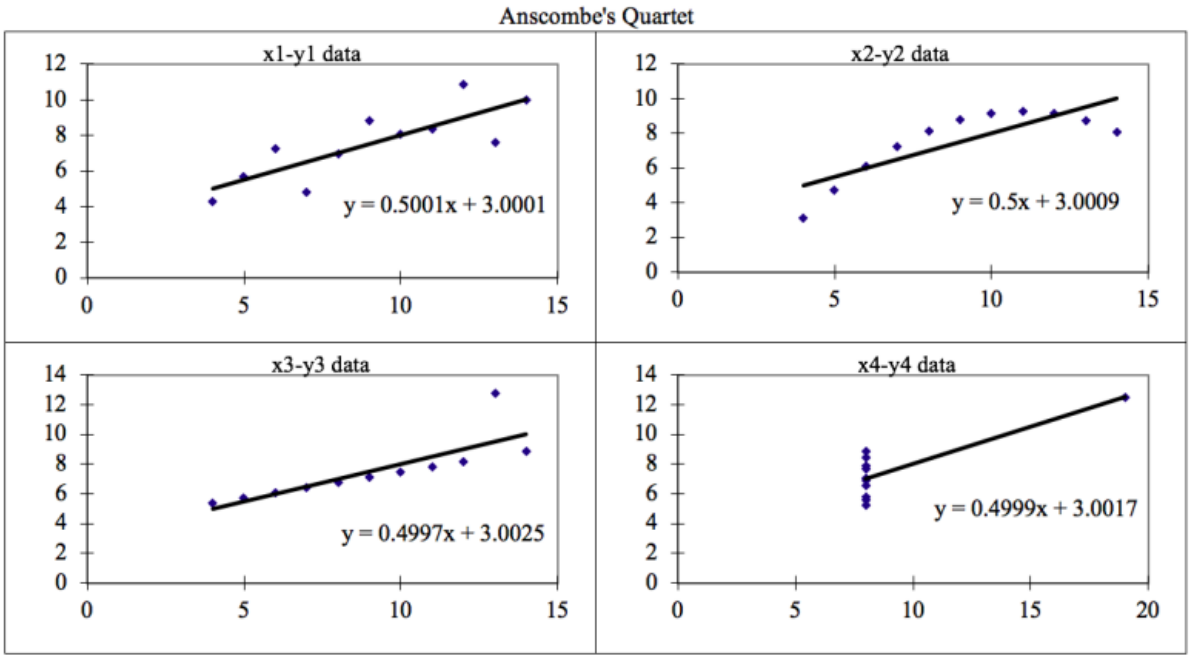


Linear Regression has following assumptions in order to establish the relationship between predictor and target variables:

1. Linearity: Target variables are linearly related to predictor variables.
2. Independence of residual terms: There should not be any dependency between residual terms. If the residual is dependent on other residual, it gives rise to autocorrelation.
3. Normality of error terms: The residuals are normally distributed with mean around zero.
4. Homoscedasticity: The variance of error terms is constant.
5. No Multicollinearity: There should not be high correlation between independent variables.

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. Each dataset consists of eleven (x,y) points.



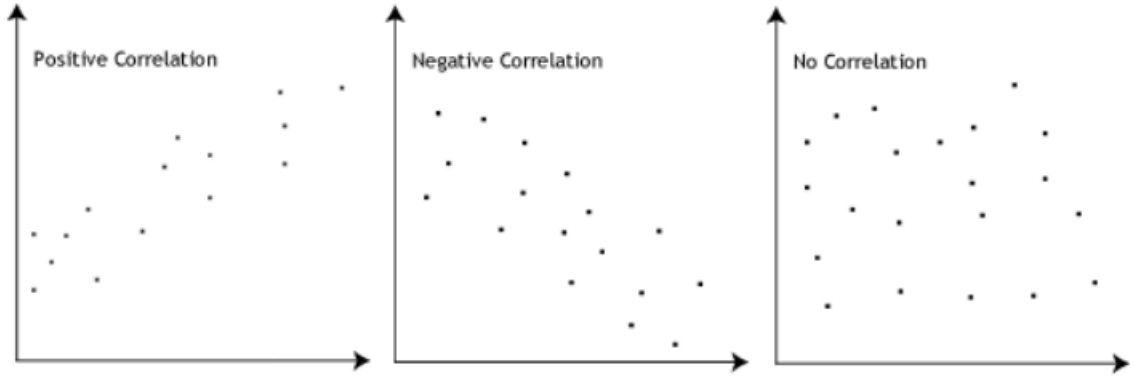
The four datasets can be described as:

a. **Dataset 1:** this fits the linear regression model pretty well.

b. **Dataset 2:** this could not fit linear regression model on the data quite well as the data is non-linear.
c. **Dataset 3:** shows the outliers involved in the dataset which cannot be handled by linear regression model
d. **Dataset 4:** shows the outliers involved in the dataset which cannot be handled by linear regression model

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

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.
Pearson's r measures the strength of the linear relationship between two variables.
Pearson's r always between -1 and 1.
If data lie on a perfect straight line with negative slope, then r = -1.



Positive correlation indicates the both the variable increase and decrease together. Negative correlation indicates the one the variable increase and the other variable decrease and vice versa.

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

Answer: Scaling is a technique to standardize the independent variables present in the data within a fixed range for a comparable scale. It is performed to make it easy for the Machine Learning algorithm to learn from the data and understand the problem because it may be impacted by the magnitude of the independent variables. There are two types of scaling techniques – MinMax Scaling and Standardized Scaling.

MinMax Scaling:
Through this technique, the data points are scaled to a range between 0 to 1. Here each data point is normalized by subtracting the minimum value and dividing by the difference between max and min.

Standardized Scaling:
Each of the data points are normalized by subtracting the mean and dividing the difference with standard deviation.

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

VIF stands for Variance Inflation Factor, which shows the correlation effect between the variables. An infinite value of VIF indicates there is perfect correlation between the variables. Infinite VIF depicts that R-square is 1, which means there is over fitting of the model.
Since VIF = 1/(1 – R_Square).

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 It is used for determining if two data sets come from populations with a common distribution.

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. Whether the Distributions is Gaussian, Uniform, Exponential or even Pareto distribution, it can be found out.

Few advantages:

a) It can be used with sample sizes also

b) 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.

It is used to check following scenarios:

If two data sets —

i. come from populations with a common distribution
ii. have common location and scale
iii. Have similar distributional shapes
iv. Have similar tail behavior

Interpretation: A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

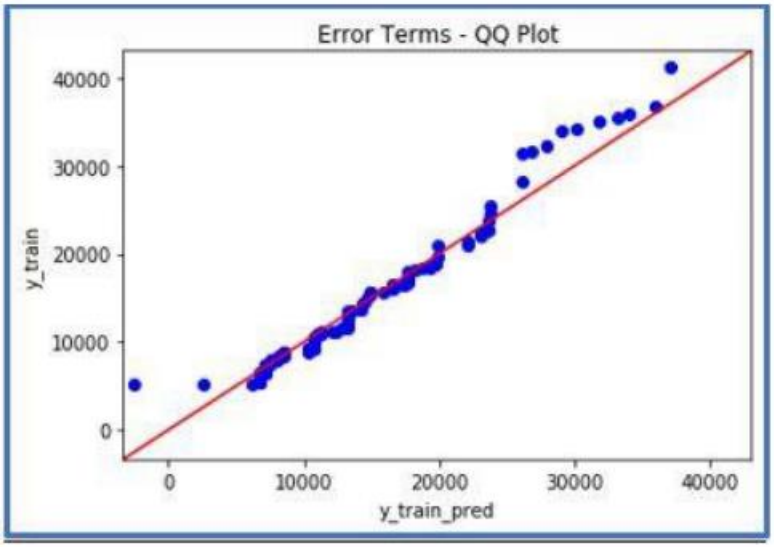
Below are the possible interpretations for two data sets.

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.

c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.

d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis


Error Terms - QQ Plot
40000
30000
20000
10000
0
y_train
0
10000
20000
30000
40000
y_train_pred