Prepared by: Ankita Maheshwari

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)**
   Answer:
   I have plotted the categorical variables with the target variables on boxplot and has inferred following effect on target:

- Season 3, fall has highest demand for rental bikes
- I see that demand for next year has grown
- 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 demand is more. During the year end and beginning

2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**
   Answer:
   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 which will create multicollinearity issue.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**
   Answer:
   The feature “temp” has highest correlation. It is very well linearly related with target “cnt”

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

   Answer:
   I have checked the following assumptions:

- Error terms are normally distributed with mean 0.
- Error Terms do not follow any pattern.
- Multicollinearity check using VIF(s).
- Linearity Check.
- Ensured the overfitting by looking the R2 value and Adjusted R2.

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:
   Features “temp”, “yr” and “season” are highly related with target column, so these are top contributing features in model building

Prepared by: Ankita Maheshwari

General Subjective Questions

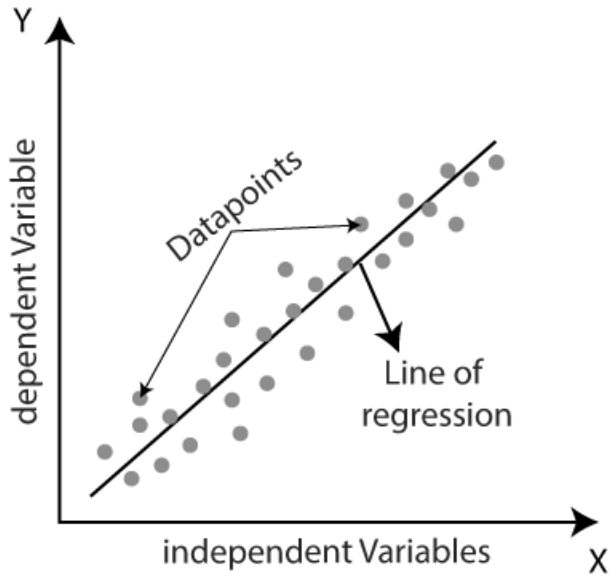
1. **Explain the linear regression algorithm in detail. (4 marks)**

Answer: 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 behaviour 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.

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.

Linear regression is used to predict a quantitative response Y from the predictor variable X. The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:



Mathematically, we can represent a linear regression as: **y= B0+B1x+ ε**

Y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)
B0= intercept of the line
B1 = Linear regression coefficient
ε = random error

There are 2 types of Linear Regression:

- **Simple Linear Regression**: If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.
- **Multiple Linear Regression**: If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Finding the best fit line:
When working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized. The best fit line will have the least error.

The different values for weights or the coefficient of lines (B0, B1) gives a different line of regression, so we need to calculate the best values for B0 and B1 to find the best fit line, so to calculate this we use cost function.

Cost function-
The different values for weights or coefficient of lines (B0, B1) gives the different line of regression, and the cost function is used to estimate the values of the coefficient for the best fit line.
Cost function optimizes the regression coefficients or weights. It measures how a linear regression model is performing. We can use the cost function to find the accuracy of the mapping function, which maps the input variable to the output variable. This mapping function is also known as Hypothesis function.
For Linear Regression, we use the Mean Squared Error (MSE) cost function, which is the average of squared error occurred between the predicted values and actual values.

Residuals: The distance between the actual value and predicted values is called residual. If the observed points are far from the regression line, then the residual will be high, and so cost function will high. If the scatter points are close to the regression line, then the residual will be small and hence the cost function.

Gradient Descent:
Gradient descent is used to minimize the MSE by calculating the gradient of the cost function. A regression model uses gradient descent to update the coefficients of the line by reducing the cost function. It is done by a random selection of values of coefficient and then iteratively update the values to reach the minimum cost function.

Assumptions of Linear Regression:
Below are some important assumptions of Linear Regression. These are some formal checks while building a Linear Regression model, which ensures to get the best possible result from the given dataset.

Linear relationship between the features and target:
Linear regression assumes the linear relationship between the dependent and independent variables.

- Small or no multicollinearity between the features:

Multicollinearity means high correlation between the independent variables. Due to multicollinearity, it may be difficult to find the true relationship between the predictors and target variables. Or we can say, it is difficult to determine which predictor variable is affecting the target variable and which is not. So, the model assumes either little or no multicollinearity between the features or independent variables.

- Homoscedasticity Assumption:
  Homoscedasticity is a situation when the error term is the same for all the values of independent variables. With homoscedasticity, there should be no clear pattern distribution of data in the scatter plot.
- Normal distribution of error terms:
  Linear regression assumes that the error term should follow the normal distribution pattern. If error terms are not normally distributed, then confidence intervals will become either too wide or too narrow, which may cause difficulties in finding coefficients.
  It can be checked using the q-q plot. If the plot shows a straight line without any deviation, which means the error is normally distributed.
- No autocorrelations:
  The linear regression model assumes no autocorrelation in error terms. If there will be any correlation in the error term, then it will drastically reduce the accuracy of the model. Autocorrelation usually occurs if there is a dependency between residual errors.

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.

These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another. They have very different distributions and appear differently when plotted on scatter plots. Each dataset consists of eleven (x,y) points.

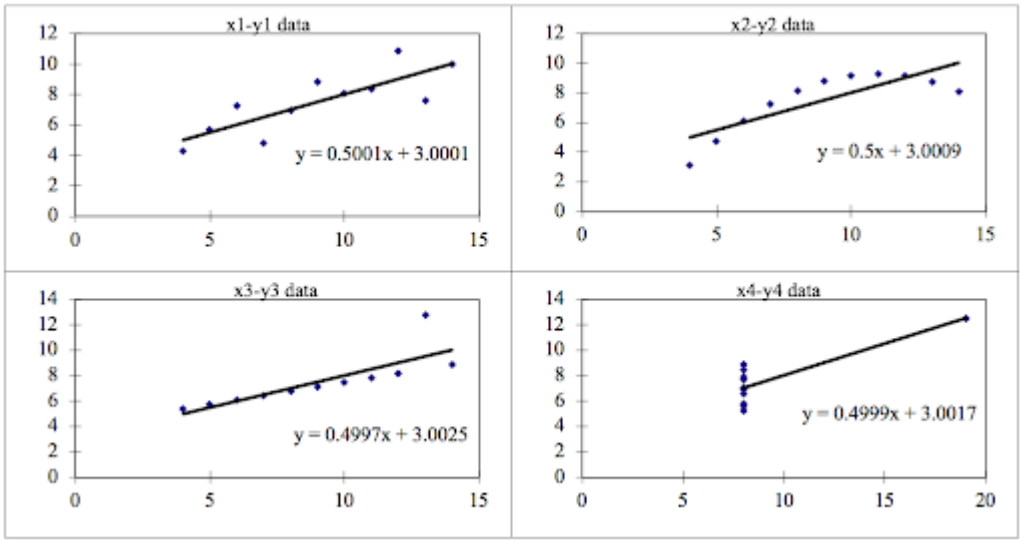
We can define these four plots as follows:

Anscombe's Data									
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89

The statistical information for these four data sets is approximately similar. We can compute them as follows:

Anscombe's Data								
Observation	x1	y1	x2	y2	x3	y3	x4	y4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.1	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.1	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89
			Summary Statistics					
N	11	11	11	11	11	11	11	11
mean	9.00	7.50	9.00	7.500909	9.00	7.50	9.00	7.50
SD	3.16	1.94	3.16	1.94	3.16	1.94	3.16	1.94
r	0.82		0.82		0.82		0.82	

However, when these models are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm, as you can see below:



We can describe the four data sets as:

- Dataset 1: this fits the linear regression model well.
- Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.
- Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model
- Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model.

As you can see, Anscombe's quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

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

Answer:

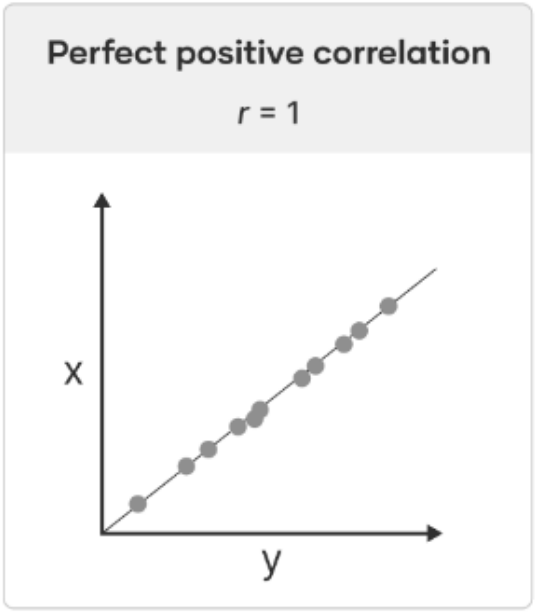
The Pearson correlation coefficient (r) is the most widely used correlation coefficient and is known by many names:

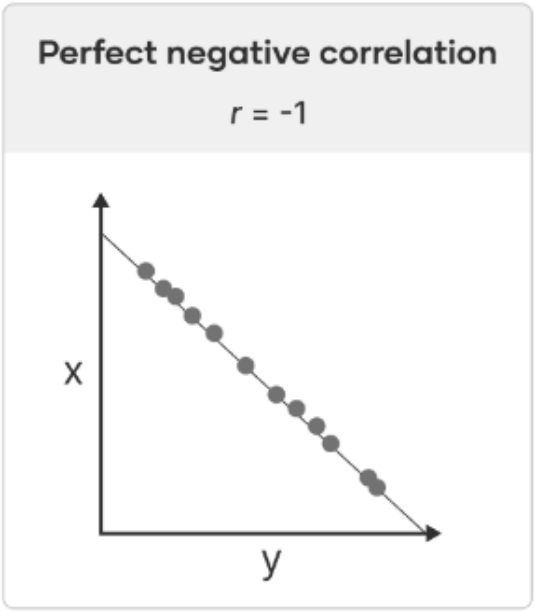
- Pearson's r
- Bivariate correlation
- Pearson product-moment correlation coefficient (PPMCC)
- The correlation coefficient

The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables. If the variables tend to go up and down together, the correlation coefficient will be positive.

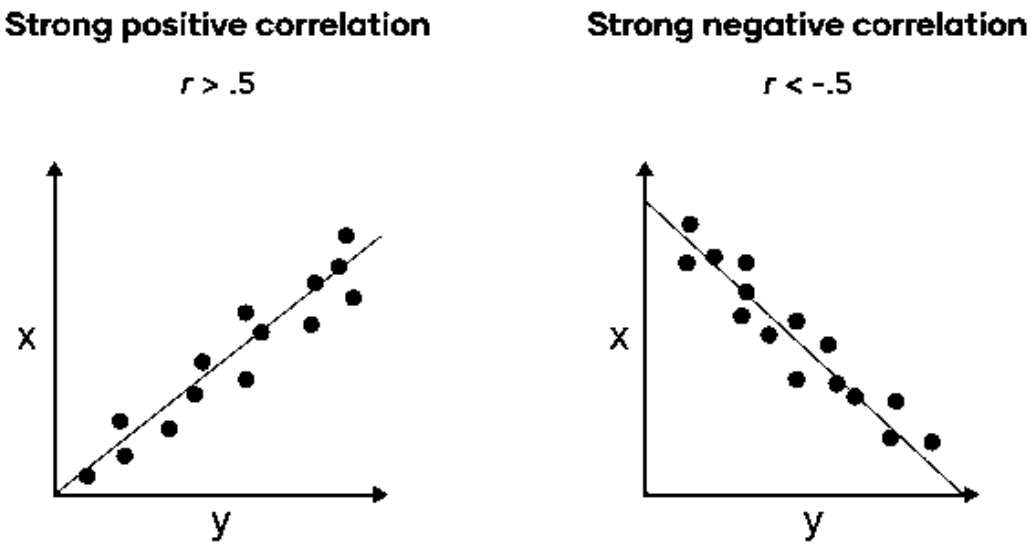
Pearson's r always between -1 and 1.

- When r is 1 or –1, all the points fall exactly on the line of best fit:

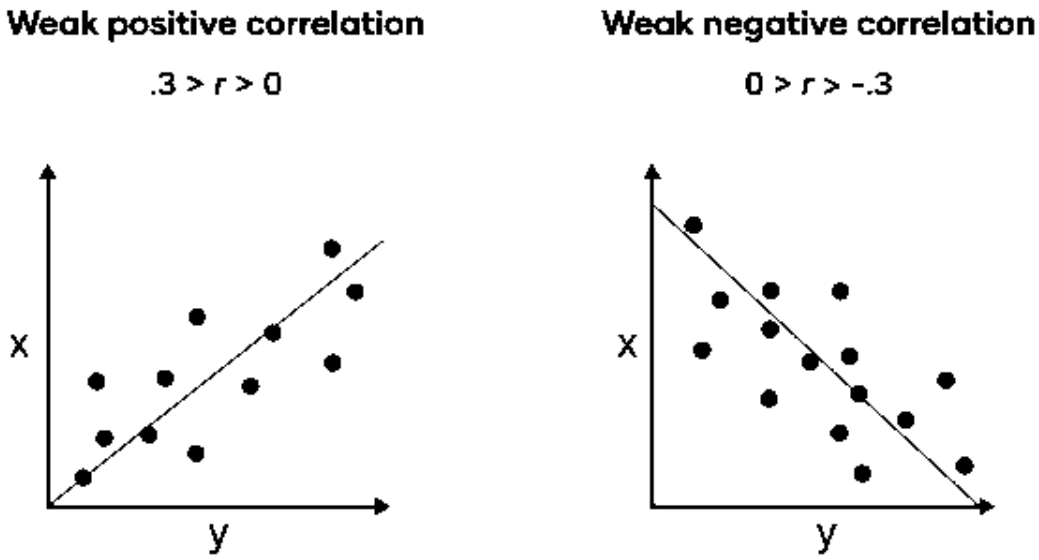




- When r is greater than .5 or less than –.5, the points are close to the line of best fit:



- When r is between 0 and .3 or between 0 and –.3, the points are far from the line of best fit:



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

   Answer:

   Scaling is a method to normalize the range of independent variables. It is performed to bring all the independent variables on a same scale in regression. If Scaling is not done, then regression algorithm will consider greater values as higher and smaller values as lower values.

   It is important to note that scaling just affects the coefficients and none of the parameters like t-statistics, F-statistics, values, R-squared, etc.

Example: Weight of a device = 500 grams, and weight of another device is 5 kg. In this example machine learning algorithm will consider 500 as greater value which is not the case. And it will do wrong prediction. Machine Learning algorithm works on numbers not units. So, before regression on a dataset it is a necessary step to perform.

Scaling can be performed in two ways:

a) Normalization/MinMax Scaling: It scale a variable in range 0 and 1.

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

b) Standardization: It transforms data to have a mean of 0 and standard deviation of 1

$$x = \frac{x - mean(x)}{sd(x)}$$

sklearn.preprocessing.scale helps to implement standardization in python.

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

Answer:

A variance inflation factor (VIF) provides a measure of multicollinearity among the independent variables in a multiple regression model. Detecting multicollinearity is important because while multicollinearity does not reduce the explanatory power of the model, it does reduce the statistical significance of the independent variables.

A large VIF on an independent variable indicates a highly collinear relationship to the other variables that should be considered or adjusted for in the structure of the model and selection of independent variables.

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

Where Ri2 = Unadjusted coefficient of determination for regressing the ith independent variable on the remaining ones.

When there is a perfect relationship then Ri2 = 1 and then VIF = Infinity, means if a variable is expressed exactly by a linear combination of another variable, then it is said that VIF is infinite

6. What is 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 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**.**

Few advantages:

- It can be used with sample sizes 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.

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 behaviour

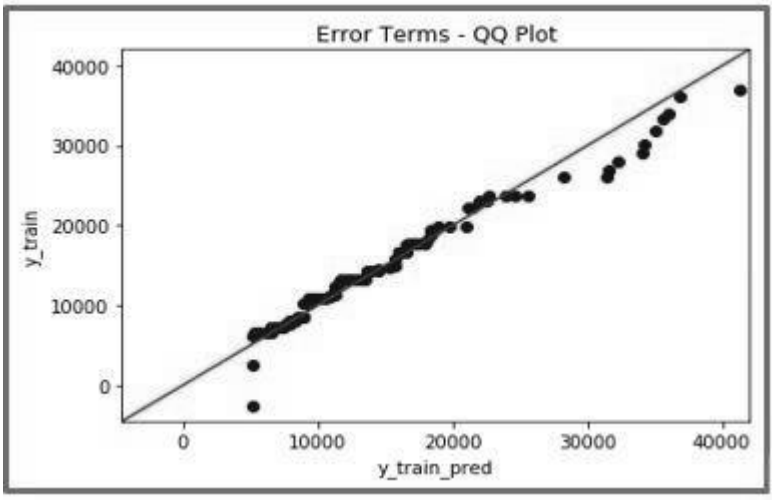
Interpretation:

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

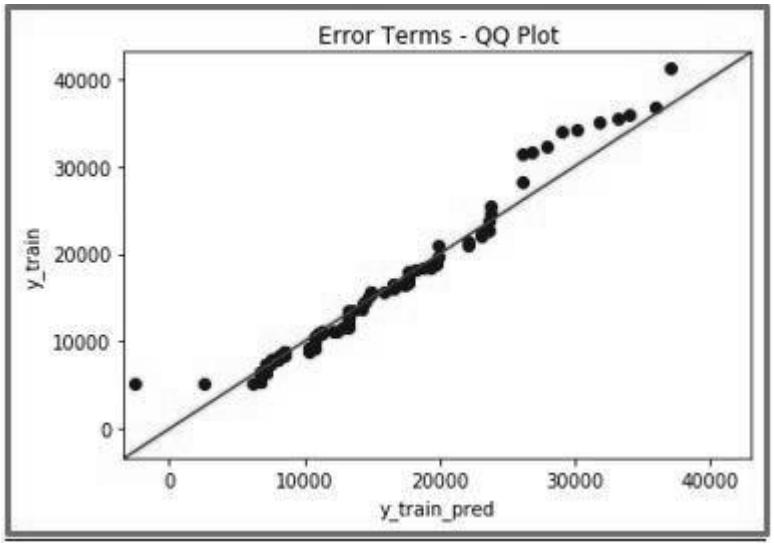
Below are the possible interpretations for two data sets.

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

Importance of Q-Q plot in Linear Regression:

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.