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:

   Created box plot for category column w.r.t cnt (demand of the shared bikes) variable.

   Following inferences have been made based on visualization:

   1. Cnt is high in summer and fall season.
   2. Cnt is more in 2019 as compared to 2018.
   3. Cnt is more in month June to October.
   4. More people prefer to rent bike on working days and during holiday people prefer to stay home.
   5. Bike rent out most when the sky is clear or partly clouds.

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

   Answer:

   Setting drop_first = **True** causes get_dummies to exclude the dummy variable for the first category of the variable you're operating on.

   When you have a categorical variable with K mutually exclusive categories, you actually only need K – 1 new dummy variables to encode the same information. This is because if all of the existing dummy variables equal 0, then we know that the value should be 1 for the remaining dummy variable.

   Example,

   if region_North == 0, and region_South == 0, and region_West == 0, then region_East must equal 1. This is implied by the existing 3 dummy variables, so we don't need the 4th.

   The extra dummy variable literally contains redundant information. So, it's a common convention to drop the dummy variable for the first level of the categorical variable that you're encoding.

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

Answer:

'temp' variable has the highest correlation with the target variable (demand of the shared bikes).

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

Answer:

1. Linear relationship validation: Linearity should be visible among variables.
2. Normality of error terms: Error terms should be normally distributed. Error terms are centered on zero.
3. Multicolinearity: There should be insignificant multicollinearity among variables. All the values of VIF are less than 5. So the model is having no multicolinearity.
4. Homoscedasticity: No visible pattern observed. Model should be distributed over the line.
5. Error term have constant variance or constant standard deviation.

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:
All the positive coefficients like,
i. temp
ii. Season-Winter
iii. Month- September

indicate that, an increase in these values will lead to an increase in the value of cnt (demand of the shared bikes).

General Subjective Questions

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

Answer:

Definition:

Linear regression is an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events.

Mathematically, we can write a linear regression equation as:

$$y = a + bx$$

Where a and b given by the formulas:

$$b(slobe) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a(intercept) = \frac{n\sum y - b(\sum x)}{n}$$

Here, x and y are two variables on the regression line.

b = Slope of the line

a = y-intercept of the line

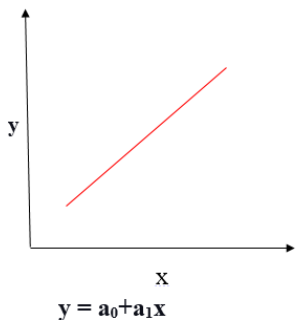
x = Independent variable from dataset

y = Dependent variable from dataset

A regression line can be a Positive Linear Relationship or a Negative Linear Relationship.

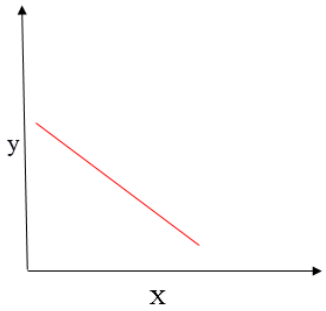
Positive Linear Relationship:

If the dependent variable expands on the Y-axis and the independent variable progress on X-axis, then such a relationship is termed a Positive linear relationship.



Negative Linear Relationship:

If the dependent variable decreases on the Y-axis and the independent variable increases on the X-axis, such a relationship is called a negative linear relationship.



$y = -a_0 + a_1 x$

The goal of the linear regression algorithm is to get the best values for a0 and a1 to find the best fit line. The best fit line should have the least error means the error between predicted values and actual values should be minimized.

Cost function:

The cost function helps to figure out the best possible values for a0 and a1, which provides the best fit line for the data points. Cost function optimizes the regression coefficients or weights and measures how a linear regression model is performing. The cost function is used to find the accuracy of the **mapping function** that maps the input variable to the output variable. This map ping function is also known as **the Hypothesis function**. In Linear Regression, **Mean Squared Error (MSE)** cost function is used, which is the average of squared error that occurred between the predicted values and actual values.

By simple linear equation y=mx+b we can calculate MSE as:

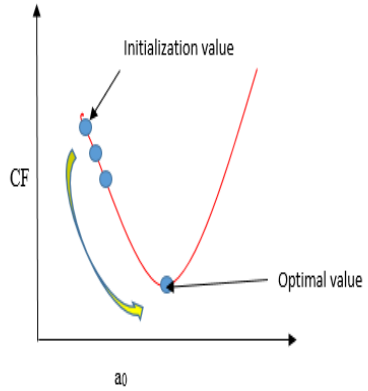
$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - (mx_i + b))^2$$

Let's y = actual values, y_i = predicted values

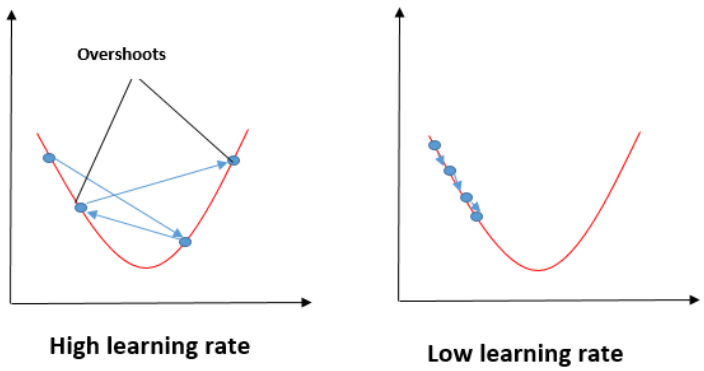
Using the MSE function, we will change the values of a0 and a1 such that the MSE value settles at the minima. Model parameters *xi, b* (a_0, a_1) can be manipulated to minimize the cost function. These parameters can be determined using the gradient descent method so that the cost function value is minimum.

Gradient descent:

Gradient descent is a method of updating a0 and a1 to minimize the cost function (MSE). A regression model uses gradient descent to update the coefficients of the line (a0, a1 => xi, b) by reducing the cost function by a random selection of coefficient values and then iteratively update the values to reach the minimum cost function.



Imagine a pit in the shape of U. You are standing at the topmost point in the pit, and your objective is to reach the bottom of the pit. There is a treasure, and you can only take a discrete number of steps to reach the bottom. If you decide to take one footstep at a time, you would eventually get to the bottom of the pit but, this would take a longer time. If you choose to take longer steps each time, you may get to sooner but, there is a chance that you could overshoot the bottom of the pit and not near the bottom. In the gradient descent algorithm, the number of steps you take is the learning rate, and this decides how fast the algorithm converges to the minima.



To update a_0 and a_1, we take gradients from the cost function. To find these gradients, we take partial derivatives for a_0 and a_1.

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

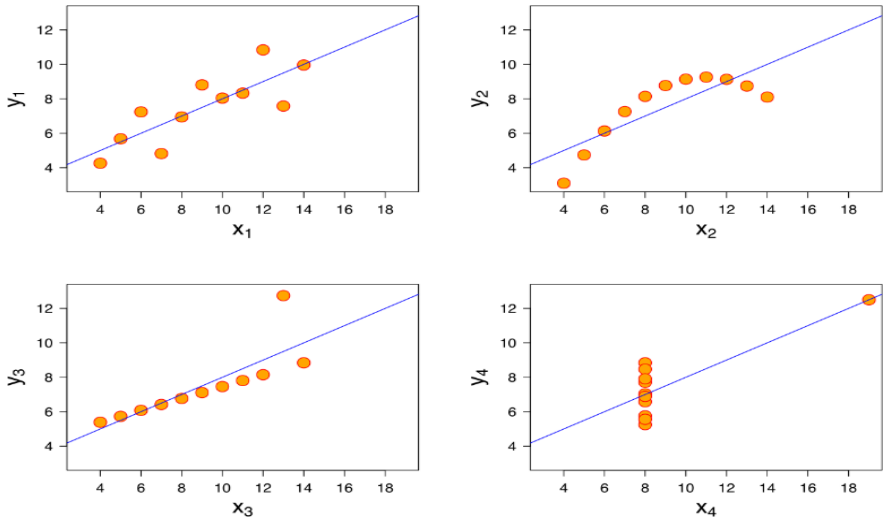
Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

The summary statistics show that the means and the variances were identical for x and y across the groups:

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset
- The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset.

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:



- Dataset I appears to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

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

Answer:

Definition:

Pearson's r is a numerical summary of the strength of the linear association between the variables.

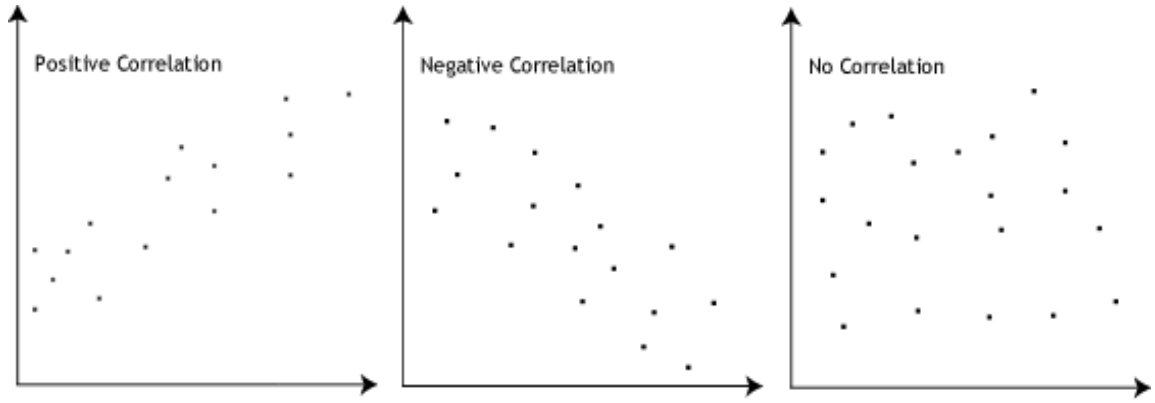
If the variables tend to go up and down together, the correlation coefficient will be positive.

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.

i. A value of 0 indicates that there is no association between the two variables.
ii. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable.
iii. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.

This is shown in the diagram below:



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

Answer:

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm. Most of the times, collected data set 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 have to do scaling to bring all the variables to the same level of magnitude.

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

S.NO.	Normalized scaling	Standardized scaling
1.	Scaling is done by the highest and the lowest values.	Mean and standard deviation is used for scaling.
2.	It is applied when the features are of separate scales.	It is applied when we verify zero mean and unit standard deviation.
3.	Scales range from 0 to 1	Not bounded
4.	Affected by outliers	Less affected by outliers
5.	It is applied when we are not sure about the data distribution	It is used when the data is Gaussian or normally distributed
6.	It is also known as Scaling Normalization	It is also known as Z-Score

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

Answer:

The value of VIF is calculated by the below formula:

$$VIF_i = \frac{1}{1-R_i^2}$$

Where, 'i' refers to the i^{th} variable.

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

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

Answer:

The Quantile-Quantile (Q-Q) plot:

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

Use of Q-Q plot:

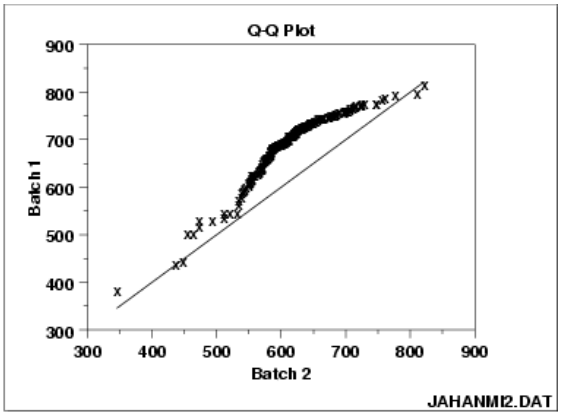
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q plot:

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample test.

Example:

The q-q plot is similar to a probability plot. For a probability plot, the quantiles for one of the data samples are replaced with the quantiles of a theoretical distribution.



These 2 batches do not appear to have come from populations with a common distribution.

The batch 1 values are significantly higher than the corresponding batch 2 values.

The differences are increasing from values 525 to 625. Then the values for the 2 batches get closer again.