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)

ANSWER1

1. Season 3 (FALL) had better rental business among all seasons and season-2&4 were better than season1

2. In 2019 in every category business seems progressed a lot better than in 2018

3. In mid of the year, May-October more users have taken bike rentals

4. From plots, we can see Clear weather was best for many riders roaming

5. Holidays are best for touring as a result more bike rentals were taken by public

6. There is not much difference between bike rentals booked on working and non-working days.

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

ANSWER2

1. When we create dummy variables for k types in a category, it is enough to create k-1 dummy variables. Because if all k-1 variables are assigned ‘0’ then it is obvious to understand the category for that column is the k variable which is dropped
2. So we use dropfirst=True to drop the first dummy variable column and make the table look simple

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

Answer3:

TEMP and Atemp both have highest correlation with target variable ‘cnt’

Temp-0.64

Atemp-0.65

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

Answer4:

1. Checked the linear relationship exists between dependent and independent variables
2. From the graph, we check the distribution of error terms. It is normally distributed
3. From VIF, we ensured there is no multicollinearity and it is within acceptable range of VIF
4. Homoscedasticity, there is constant variance

5. Error terms are independent. There are no visible patterns

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

ANSWER5:

From the equation, contributing features are

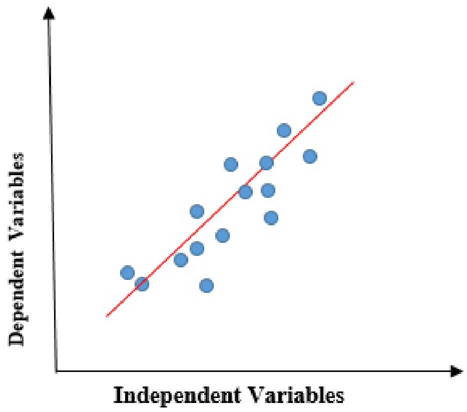
1. Temp
2. Year
3. Weathersit_3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)

GENERAL SUBJECTIVE QUESTIONS

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

ANSWER1:

Linear regression is one of the machine learning techniques that analyses the strength of linear relationships between dependent and independent variables. The linear regression model gives a sloped straight line describing the relationship within the variables.



Mathematical Relation

Y= a0 +a1 X

a0= intercept of line

a1= linear regression coefficient

X= Independent variable

There are two types of regression models

1. Simple Linear Regression: it accounts the effect of one independent variable the for prediction of the dependent variable

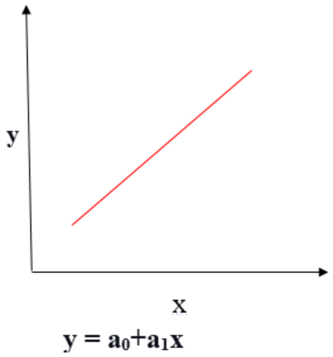
2. Multiple Linear regression: It accounts a more than one independent variable on dependent variable for improving predicting power.

Assumptions of the Linear regression model are :

1. The linear relationship exists between dependent and independent variables
2. The distribution of error terms is normally distributed
3. There is no multicollinearity
4. Homoscedasticity, there is constant variance in the spread of data
5. Error terms are independent. There are no visible patterns in the error terms data

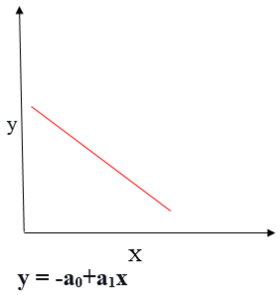
Positive Linear Relationship:

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



Negative Linear relationship

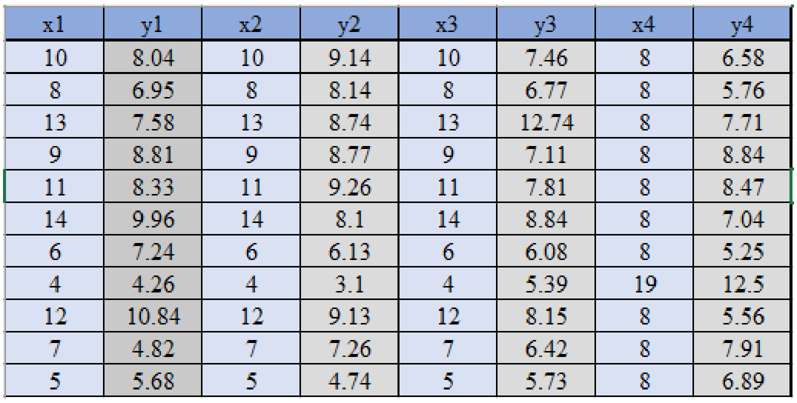
If the dependent variable decreases as the independent variable progress on X-axis, then such a relationship is termed a Negative linear relationship.



The aim of the linear regression algorithm is to get the best values for a0 and a1 to find the best fit line. It is achieved through the ordinary least squares method. It optimizes the regression coefficients or weights and measures how a linear regression model is performing.

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

Anscombe’s Quartet is that the modal example to demonstrate the importance of information visualization which was developed by the statistician Francis Anscombe in 1973 to suggest both the importance of plotting data before analyzing it with statistical properties. It comprises of 4 data-set and every data-set consists of 11 (x,y) points. the fundamental thing to research about these data-sets is that all of them share the identical descriptive statistics(mean, variance, variance etc) but different graphical representation. Each graph plot shows the various behavior no matter statistical analysis.



x1	y1	x2	y2	x3	y3	x4	y4
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

They all share the same descriptive statistics(mean, variance, standard deviation etc) but different graphical representation.

Average Value of x = 9

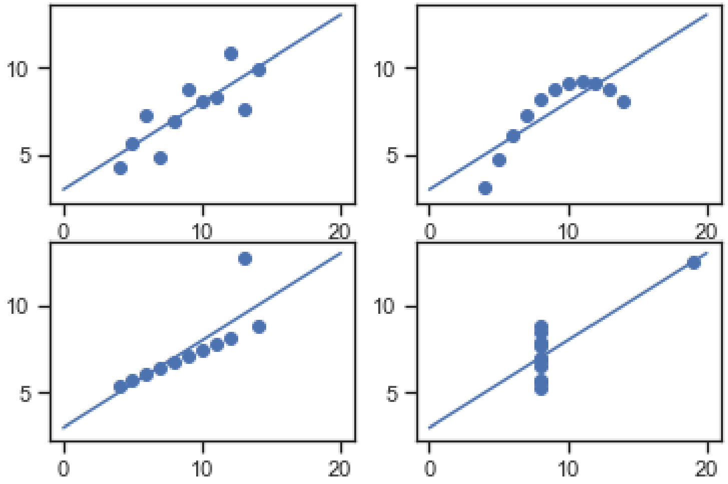
Average Value of y = 7.50

Variance of x = 11

Variance of y =4.12

Correlation Coefficient = 0.816

Linear Regression Equation : y = 0.5 x + 3



Graph-I shows a linear relationship with some variance.

Graph II shows a curve shape but doesn't show a linear relationship, it might be a quadratic

Graph III looks like a tight linear relationship between x and y and have one large outlier.

Graph IV looks like the value of x remains constant, except for one outlier as well.

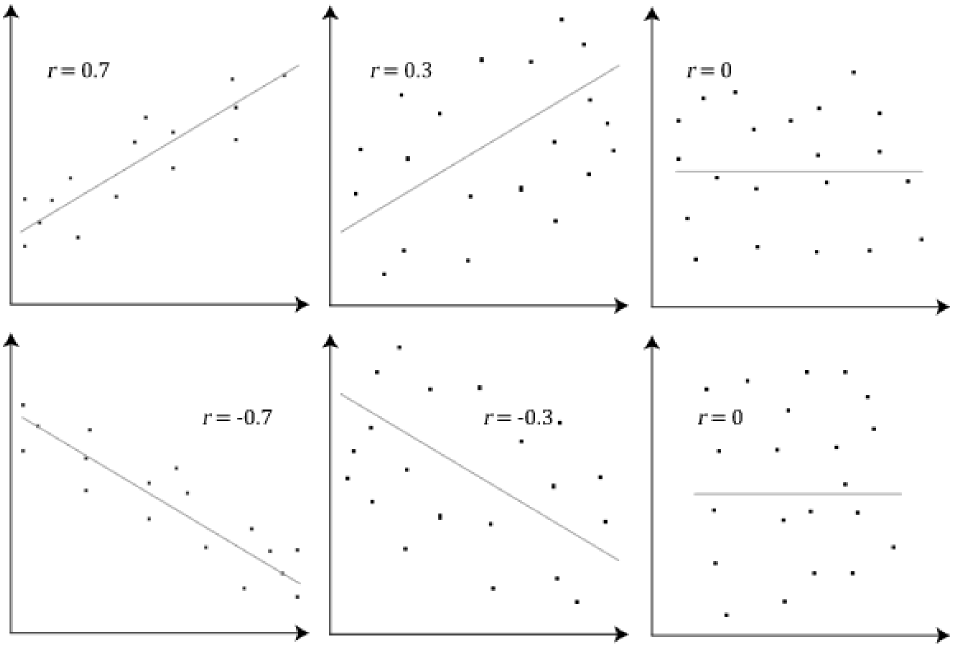
3. What is Pearson's R? (3 marks)

ANSWER 3:

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to travel up and down together, the correlation is going to be positive. If the variables tend to travel up and down opposite with low values of 1 variable related to high values of the opposite, the coefficient of correlation is negative. If r = 1 means the data is perfectly linear with a positive slope. if r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions).

If r = 0 means there is no linear association

If r > 0 < 5 means there is a weak association



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

Scaling is a data Pre-Processing that is applied to independent variables to normalize the data within a particular range to ease the computation

Most of the time, the collected data set contains features highly varying in magnitudes, units, and ranges. If scaling is not done then the algorithm only takes magnitude into account and not units. This will result in forming a bad linear model and interpretation of coefficients become difficult. To solve this issue, we should do scaling so that all variables come to the same level of magnitude.

Normalization:

It brings all the data within ranges 0 to 1

sklearn. preprocessing.MinMaxScaler is used for performing normalization in python.

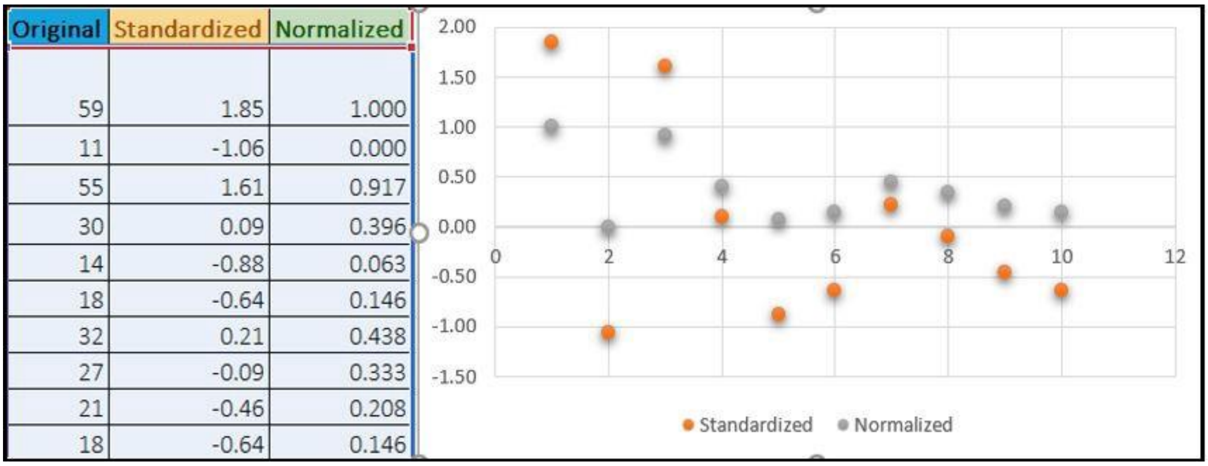
$$X^{'} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Standardization Scaling:

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

$$X^{'} = \frac{X - \mu}{\sigma}$$

Below depicts the Standardized and Normalized scaling on original values.



Original	Standardized	Normalized
59	1.85	1.000
11	-1.06	0.000
55	1.61	0.917
30	0.09	0.396
14	-0.88	0.063
18	-0.64	0.146
32	0.21	0.438
27	-0.09	0.333
21	-0.46	0.208
18	-0.64	0.146

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

ANSWER5:

VIF shows infinite If there is a perfect correlation between two independent variables. In the case of perfect correlation, we do get R2 =1, which results in 1/(1-R2) to be infinity. To solve this issue We need to eliminate terms that are duplicates or do not add value to explain the variation in the model. For example, if your inputs are measuring the weight in kgs and lbs then just keep one of these variables in the model and drop the other one, Dropping the term with a large value of VIF will hopefully, fix the VIF for the remaining terms.

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

ANSWER6:

Quantile-Quantile (Q-Q) plot, is a graphical tool for assessing whether 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. For example, if you are testing if the distribution of age of employees in your team is normally distributed, you are comparing the quantiles of your team members' age vs quantile from a normally distributed curve. If two quantiles are sampled from the same distribution, they should roughly fall in a straight line.

Since this is a visual tool for comparison, results can also be quite subjective but nonetheless useful in the understanding underlying distribution of a variable(s)

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.