1.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)

Ans: From the analysis of the categorical columns using the boxplot and bar plot. We observed the following things.

- From the season vs count bar plot we can conclude that the number of user count is highest in Fall season
- In the yr vs count bar plot we can conclude that in year 2019 the bookings are more than in the year 2018
- From the month vs count bar plot we can see that the as the we are going from jan to june the sales is increasing and then gradually decreasing till december
- When it's not holiday, booking are less in number which seems reasonable as on holidays, people may want to spend time at home and enjoy with family.
- Booking are same either on working day or non-working day

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

Ans: drop_first = True is important to use, it basically help us in reducing the extra column created during dummy variable creation. Hence it decreases the correlations created among dummy variables.

Syntax - drop_first: bool, default False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level.

Let's say we have 2 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not A, then It is obvious B. So, we do not need 2nd variable to identify the B.

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

Ans: 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)

Ans: We used 5 assumptions to evaluate the linear regression model

- Normality of error terms

  Error term should be distributed normally

- Multi Colinearity

  There should be insignificant multicollinearity between variables.

- Homoscedasticity

  There should be no visible pattern in the residual variable

- **i**ndependence of residuals

  No auto-correlation

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

Ans: 1.Temp

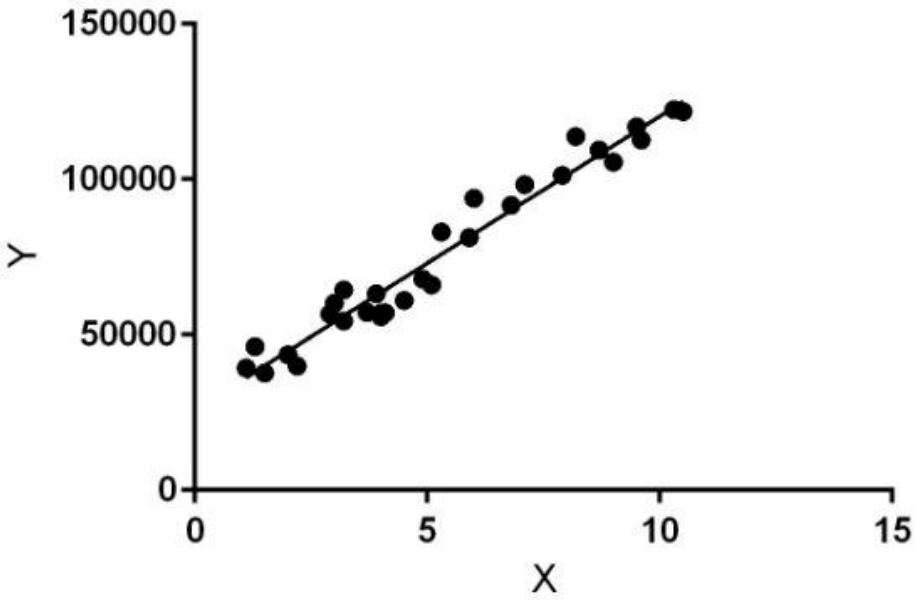
2.Winter

3. Sep

General Subjective Questions

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

linear regression is a linear approach for modelling the relationship between a scaler response and one or more explanatory variables (also known as dependent and independent variable).The case of only one exploratory variable is called simple linear regression . Whereas if we are doing it for multiple variable it is called multi linear regression model.Different linear regression differs based on the kind of relationship between dependent and independent variables.



The linear regression perform task by taking dependent variable as Y and than predicting the value of Y based on the independent variable X.

The mathematical representation can be represented using the following equation

Y = mX+ C

Y : Y is the dependent variable whose value will be predicted

m: m is the slope of the line which represent the relationship between effect of x on y

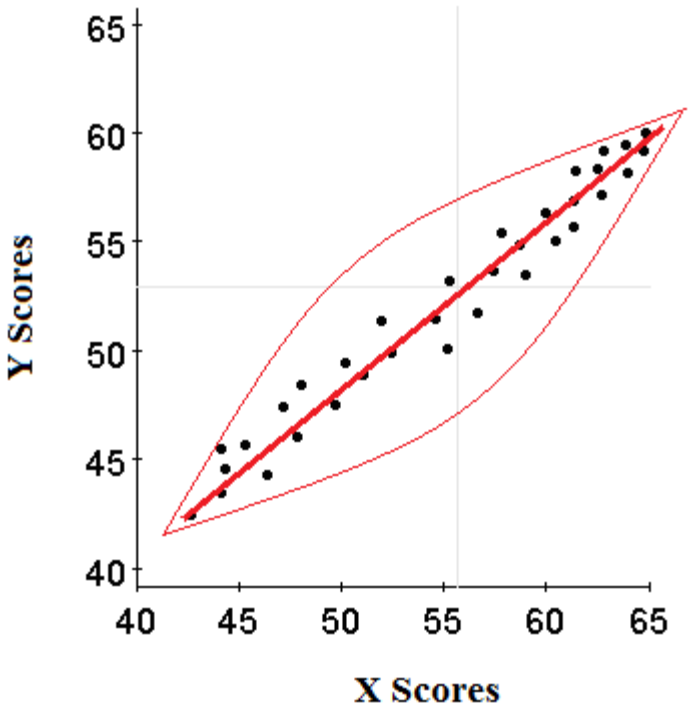
X: X is the independent variable using which the value if y is predicted

C: C is the constant. I.e when X= 0 , Y = c.

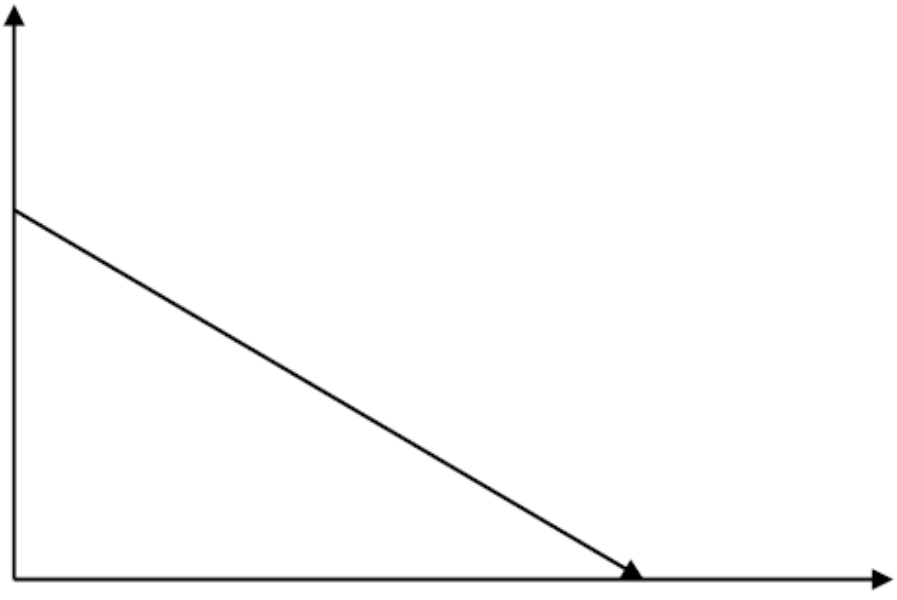
Once we find the best m and c values, we get the best fit line.

Now the linear regression relationship can be positiive or it can be negative both.

- A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph −



- If the slope is negative, then there is a negative linear relationship, i.e., as one increases the other variable decreases. It can be understood with the help of following graph –



Negative Linear Relationship

Assumptions

- Multi-collinearity

As per Linear regression model there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.

- Auto-correlation –

Another assumption of Linear regression model is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.

- Relationship between variables

Linear regression model says that the relationship between response and feature variables must be linear

- Normality of error terms

Error terms should be normally distributed

- Homoscedasticity

There should be no visible pattern in residual values.

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

Ans:

Anscombe's Quartet is the modal example to showcase the importance of data visualization which was actually created by the statistician ***Francis Anscombe*** in 1973 in order to showcase both the importance of plotting data prior to analyzing it with statistical properties. It constituted of four data-set where each data-set consists of 11 (x,y) data points. The most basic thing to analyze about these data-sets is that they all share the same descriptive statistics(mean, variance, standard deviation etc) but have different graphical representation. Each graph plot shows the different behavior in spite of 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

Apply the statistical formula on the above data-set,

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

But, statistical analysis of these 4 data-sets are very much similar. But when we are plotting the four data-sets across the x & y coordinate plane, we will get the following results & each pictorial view represent the different behavior.

• Dataset I appears to have clean and well-fitting linear models. • Dataset II is not distributed normally.

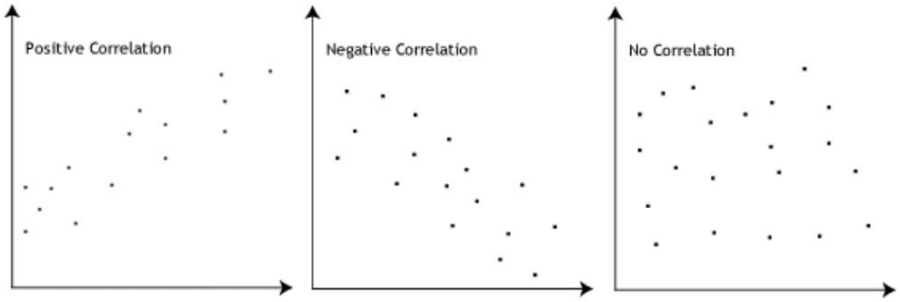
• In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.

• 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)

Ans:

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables are going up and down simultaneously, the correlation coefficient will be +ve. 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 -ve. The Pearson correlation coefficient, r, can take the range of values ranging from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value > 0 indicates a positive association; that is, as the value of one variable will ncreases, so does the value of the other variable. A value < 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)**

Scaling is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

For example:

If an algorithm is not using the feature scaling method, then it can consider the value 100 MB to be greater than 1 GB

but that's actually not true and, in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to the same magnitudes and thus, tackle this issue.

Normal Scaling	Standardized scaling
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is really affected by outliers.	It is much less affected by outliers.
Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

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

Ans:

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Ans:

The quantile-quantile (q-q) plot is a graphical technique to identify 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 1st data set against the quantiles of the 2nd 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 tests.