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:

- Fall Season seems to have attracted more booking as compared to other season also in each season the booking count has increased exceptionally from 2018 to 2019.
- Most of booking done in the month between may to oct and we can see that it also increasing exceptionally from 2018 to 2019.
- It seems clear whether attract more people.
- Thursday Friday and Saturday attract more people hence booking is more in these days.
- People like to book a bike on non-holiday it may be possible people like spend more time with family on holiday rather than going out. But this trend decreases from year 2018 to 2019.
- There is more booking done in 2019 compared to 2018.

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

ANSWER:

- We need to use drop_first=True during dummy variable creation is avoid extra column created and these dummy variables are highly correlated to each other so there is chance we will end up increasing multicollinearity.
- For e.g.:

  Let say column as 3 values A B and C so dummy variables will create using drop_first = True will be for B and C column where A represents as 00, B represent as 10 and C will represent as 01. We need to created column for A.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation

with the target variable? (1 mark)

ANSWER:

temp and atemp have the highest correlation with target variable around 0.63.

4. How did you validate the assumptions of Linear Regression after building the model on the

training set? (3 marks)

ANSWER:

1. Target Should be Numerical in nature.

2. Normality of error terms

- Error terms should be normally distributed (Using QQ plot).

3. Multicollinearity check
- There should be medium to no multicollinearity among variables (Condition number should be less than 100).

4. Linear relationship validation
- Linearity should be visible among variables.

5. Homoscedasticity
- There should be no visible pattern in residual values.
- Error term should have constant variance (Prob (Omnibus) should be close to 0)

6. Independence of residuals
- No autocorrelation (Durbin Watson number should be between 1.5 to 2.5)

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:

Top 3 features significantly towards explaining the demand of the shared bikes are:

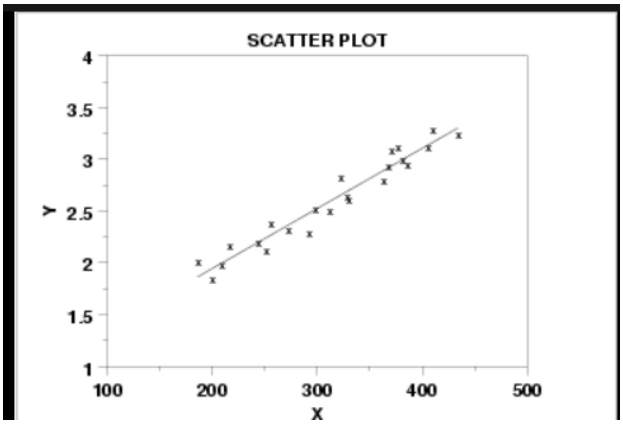
- Temp
- Year
- Winter

General Subjective Questions

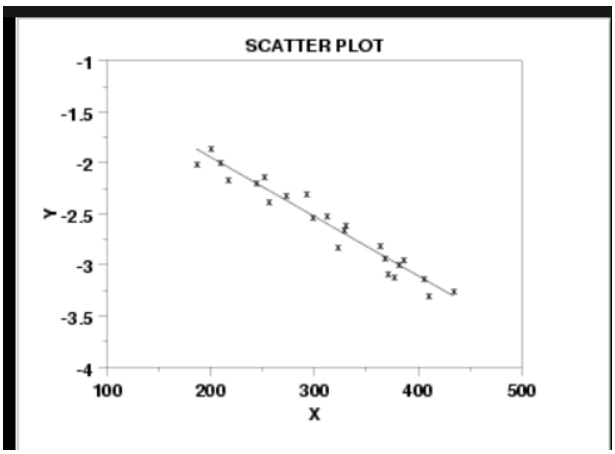
1. Explain the linear regression algorithm in detail. (4 marks)
   ANSWER:
   Linear Regression is defined as statistical model analyses the linear relationship between dependent variable and given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease) but linearly (like a straight line).
   Mathematically the relationship can be represented with the help of following equation.
   $Y = mX + c$
   Which nothing but equation of straight line.
   Here Y is dependent variable which we need to predict and X is set of independent variables which will help in predicting y.
   m is the slope of the regression line which represents the effect X has on Y (per unit change in Y).
   c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

   Furthermore, the linear relationship can be positive or negative in nature as explained below−

Positive Linear Relationship: A linear relationship will be called positive if independent increases then dependent variable also increases.



Negative Linear relationship: A linear relationship will be called positive if independent increases and dependent variable decreases.



- Linear regression is of the following two types:

1. **Simple Linear Regression**
2. **Multiple Linear Regression**

Assumptions -
The following are some assumptions about dataset that is made by Linear Regression model:

1. Target should be numeric(continuous) in nature.
2. Multi-collinearity –

- Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency between themselves.

3. Auto-correlation –

- Another assumption Linear regression model assumes is that there is very little or no autocorrelation in the data. Basically, auto-correlation occurs when there is dependency between residual errors. Durbin Watson number can be used to check autocorrelation.

4. Relationship between variables
- Linear regression model assumes that the relationship between target and feature variables must be linear.
5. Normality of error terms –
- Error terms should be normally distributed.
6. Homoscedasticity –
- There should be no visible pattern (funnel shape in scatter plot) in residual values or we can say the error term should have constant variable.

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

Answer:

Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises a set of four dataset, each containing eleven (x, y) pairs, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph. Each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths.
But things change completely .

	x1	x2	x3	x4	y1	y2	y3	y4
0	10	10	10	8	8.04	9.14	7.46	6.58
1	8	8	8	8	6.95	8.14	6.77	5.76
2	13	13	13	8	7.58	8.74	12.74	7.71
3	9	9	9	8	8.81	8.77	7.11	8.84
4	11	11	11	8	8.33	9.26	7.81	8.47
5	14	14	14	8	9.96	8.10	8.84	7.04
6	6	6	6	8	7.24	6.13	6.08	5.25
7	4	4	4	19	4.26	3.10	5.39	12.50
8	12	12	12	8	10.84	9.13	8.15	5.56
9	7	7	7	8	4.82	7.26	6.42	7.91
10	5	5	5	8	5.68	4.74	5.73	6.89

	I	II	III	IV
Mean_x	9.000000	9.000000	9.000000	9.000000
Variance_x	11.000000	11.000000	11.000000	11.000000
Mean_y	7.500909	7.500909	7.500000	7.500909
Variance_y	4.127269	4.127629	4.122620	4.123249
Correlation	0.816421	0.816237	0.816287	0.816521
Linear Regression slope	0.500091	0.500000	0.499727	0.499909
Linear Regression intercept	3.000091	3.000909	3.002455	3.001727

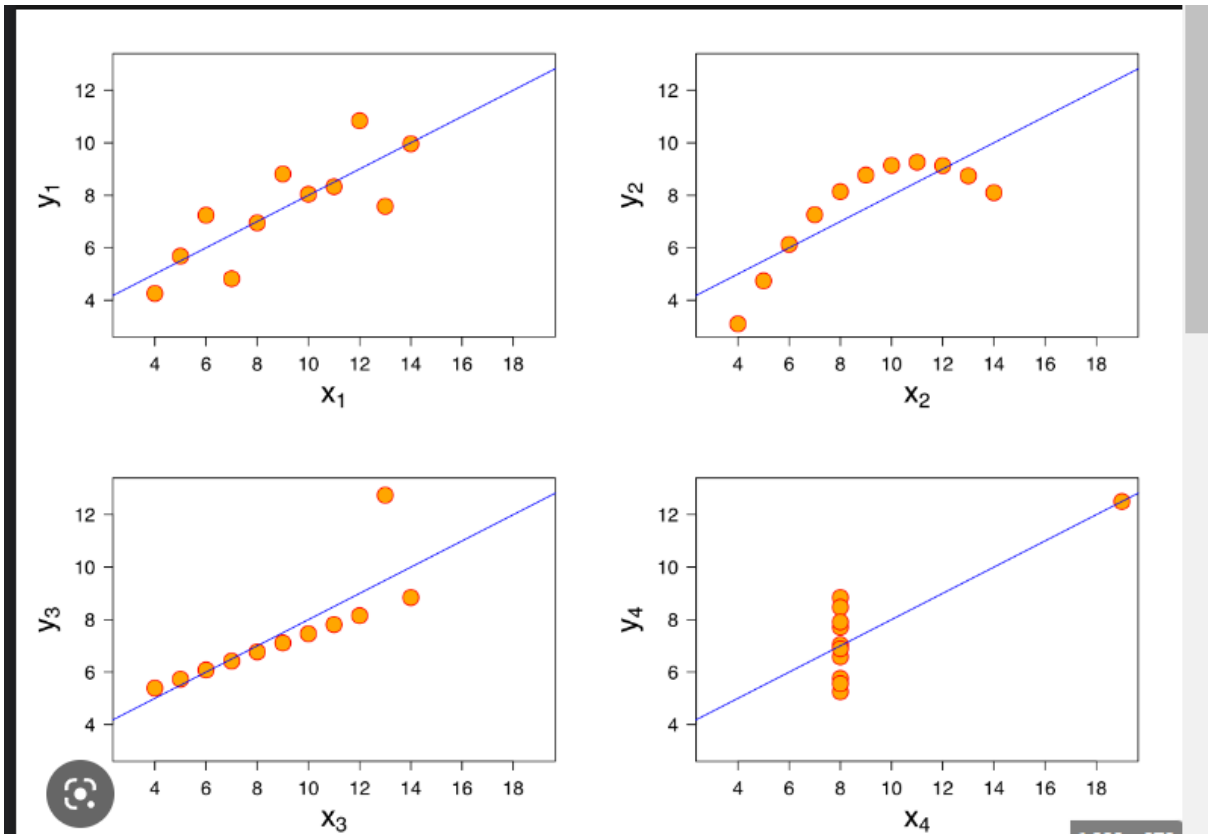
The summary statistics show that the means and the variances were identical for x and y across the groups:

• 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(approx.) 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 and Each graph tells a different story irrespective of their similar summary statistics:



• Dataset I appear 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.

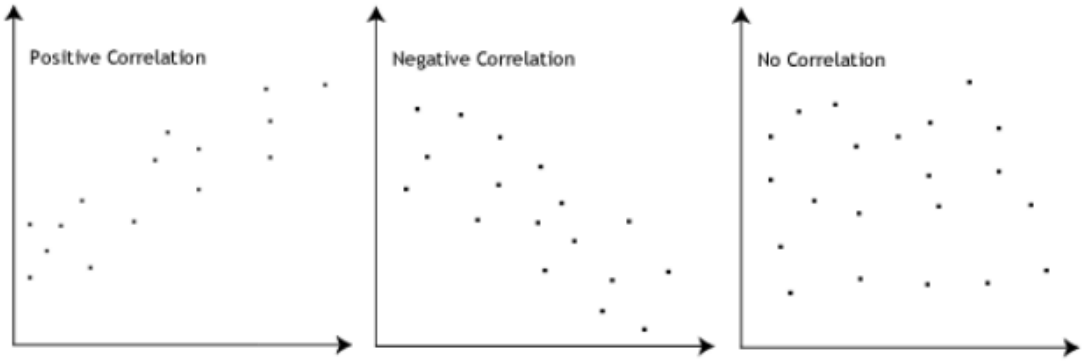
• Dataset IV shows that one outlier is enough to produce a high correlation coefficient. This quartet emphasizes the importance of visualization in Data A.

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:

Pearson's r is a numerical summary of strength of linear association between the two variables.
If increase one variable result in increase of another variable, then we can say that two variables are positively correlated. If increasing one variable result in decreasing other variables, then these variables are negative correlated. Coefficient of Pearson's r values lies between -1 to 1.
If any two variables have Coefficient of Pearson's r as 1 means they are highly positive correlated. If any two variables have Coefficient of Pearson's r as -1 means they are highly negative correlated and if any two variables have Coefficient of Pearson's r as 0 means they are not correlated.



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

Answer:

Feature scaling is very important in machine learning especially all the algorithms which used distance based approach. It is important steps in data pre-preparation to bring down each variable in similar scale. It 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 as higher and consider smaller values as the lower values, regardless of the unit of the values. Due to this model performance will impact.

For eg: Let say we have two variable salary and age. Range of salary is between 10000 to 500000 and range of age is 18-70. Here we can see that range of salary is much greater than age so there is chance if we don't scale both the column then model will affect more because of salary rather than because of age which may lead wrong prediction for target column.

Standardization Scaling (StandardScaler)	**Normalization Scaling (MinMaxScaler)**
It used when data is following normal distribution.	It used when data is not following normal distribution.
Mean and standard deviation is used for scaling.	It used minimum and maximum value for scaling.
It scales the values such that standard deviation become 1 and mean become 0 of the scaled variables.	It scales the value such that range of scaled value is between 0 and 1 for variable.
It is much less affected by outliers	It is really affected by outliers.

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

(3 marks)

Answer:

If VIF value is infinite, it means that there is perfect correlation. A Large value indicate that there is high correlation between the variables. If VIF is 3 means that this means that the variance of the model coefficient is inflated by a factor of 3 due to the presence of multicollinearity. It is also means that there is redundancy in independent features.
In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2)
Infinity. Since
VIF = 1/ (1-R2)
To solve this issue, we need to drop one of variable which is involved in multicollinearity. Ideally VIF value should be less than 5 but sometimes it's better to be as low as 2 or 3.

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 plot is a graphical method for determining whether two samples of data came from the same population or not.

By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.4 (or 40%) quantile is the point at which 40% percent of the data fall below and 60% 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.

Importance of qq plot:

When we have two data sample, so we want to know if assumption of common distribution is valid/justified or not. If so the 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 between them. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and 2-sample tests.