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:

There are 5 categorical variable in our data set, namely Season, Month, Holiday, Weather, and Weekday. Every category has some effect on the target variable which is the count of users. Among Season,
• it was found that July-Aug-September resulted in highest user count.
• Similarly, clear weather resulted in more demand
• working day affected the overall demand.
• Year 1 resulted in more booking than year 0.
• Thursday, Friday, Saturday resulted in more number of bookings.

---

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

ANS:

It reduces the correlation among dummy variables by dropping the extra creation columns during dummy variable creation.

---

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

Ans: Temp is the most correlated variable with target variable.

---

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

Ans: Assumptions validated:
- Error terms should be normally distributed.
- Insignificant multi-collinearity which was evident from the summary analysis and dropping variables with VIF > 5

- No visible pattern in residual values.
- Variables were linear by plotting their graphs.

---

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

Ans: Top 3 features that contributed to demand were *Temp, September and January-Feb-March quarter.*

General Subjective Questions

---

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

Ans:

Linear regression is defined as mathematical model which analyses the relationship between a dependent variable and a set of independent variable.

Mathematically, it is written as an equation of straight line

Y = mX + C where

Y is the dependent variable.
X is the independent variable
m is the slope of the line
C is the intercept.

This equation is a case of simple regression. In case of multiple regression, the number of independent variable affecting the dependent variable is greater than 1.

If m> 0, it is called a positive linear regression.
If m < 0, it is called a negative linear regression.

The straight line equation is derived by reducing the slope so that the overall sum of errors is minimised for a set of m.

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

Ans:

Anscombe's quartet consists of four datasets with very similar summary statistics but very different when plotted.

Anscombe's quartet

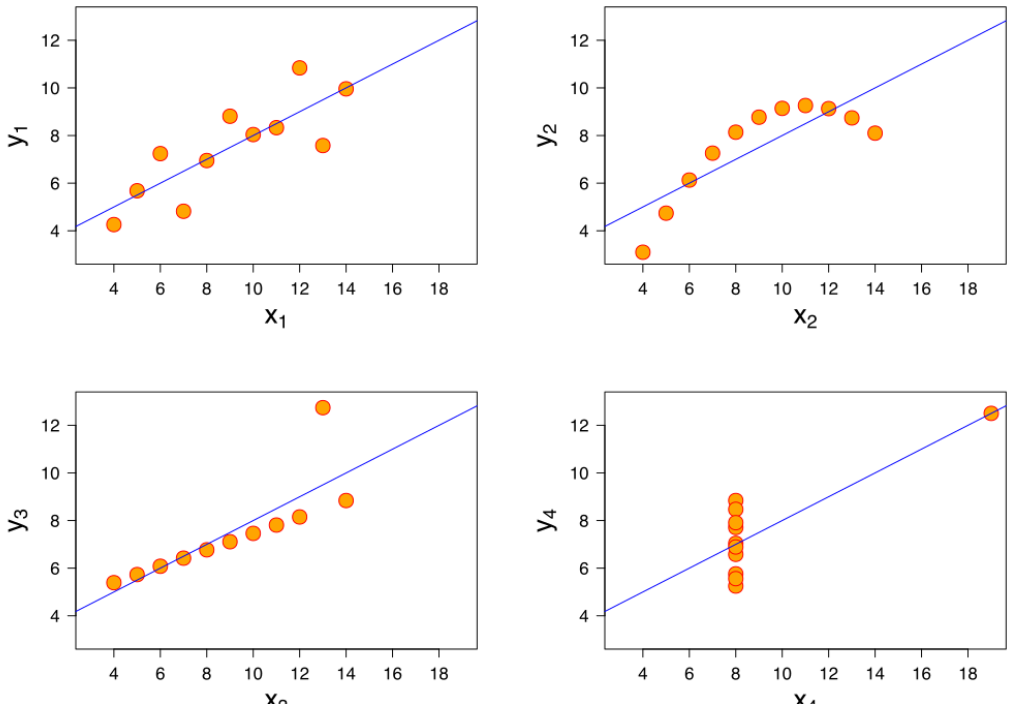
I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Pic from Wikipedia

For all four datasets the summary statistics is similar:

Property	Value	Accuracy
Mean of *x*	9	exact
Sample variance of *x* : s_{x}^2	11	exact
Mean of *y*	7.50	to 2 decimal places
Sample variance of *y* : s_{y}^2	4.125	±0.003
Correlation between *x* and *y*	0.816	to 3 decimal places
Linear regression line	*y* = 3.00 +	to 2 and 3 decimal
Coefficient of determination of the	0.67	to 2 decimal places

However, when plotted, the plots will be



Pic from Wikipedia

---

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

Ans: It is a measure of linear correlation between two sets of data. It is also known as Pearson product moment correlation coefficient (PPMCC). It is ration between covariance of two variables and the product of their standard deviations.

So, P(x,y) = COV(X,Y)/STD.DEV(x).STD.DEV(y)

The value of Pearson R can be between -1 and 1 as it is essentially a normalised measurement of covariance.

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

Ans:

Feature Scaling is a technique to standardise the independent features present in the data in a fixed range. 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, higher and consider smaller values as the lower values, regardless of the unit of the values.

Example: If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km 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 same magnitudes and thus, tackle this issue.

Normalised scaling	*Standardised scaling*
Min-Max value for features are used for scaling	Mean and Std. Deviation is used for scaling.
Used when features are in different scale	Used when we want to ensure 0 mean and unit std. deviation.
Affected by outliers	Not affected by outliers
Scales between [0,1] or [-1,1]	Not bounded by any scale.

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 between variables, VIF will be infinite. R2 in such case become 1 and the denominator becomes 0. As a result 1/(1-R2) becomes infinity.

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 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 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.