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:

   We have done analysis on categorical columns using the boxplot .Below are the few points we can infer from the visualization –

   - 1.Bike hire numbers are maximum during fall
     2. 2019 year sees a boom in bike hirers
     3. September gets maximum bike hirers and January least
     4.There is not much difference in bike hires number, even if there is a holiday
     5.Good weather accounts for hire bike numbers

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

   Answer:

   drop_first = True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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' 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)**

   Answer:

   I have validated the assumption of Linear Regression Model based on below 5 assumptions -

   - Normality of error terms - Error terms should be normally distributed
   - Multicollinearity check – There should be insignificant multicollinearity among variables. I Have checked the VIF values , making sure all are below 5
   - Linear relationship validation - Linearity should be visible among variables

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 contributing significantly are-

   - temp
   - bad weather
   - year

General Subjective Questions

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

Answer:

It is the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables. If there is Linear relationship between variables then if the value of one or more independent variables will change simultaneously it will change the value of dependent variables. And change will be same, if there is positive change in independent variable then in dependent variable also there will be positive change and vice-versa.

Mathematically we can say that-

$$Y = mX + c$$

Here,

Y is the dependent variable we are trying to predict.

X is the independent variable we are using to make predictions.

m is the slope of the regression line

c is a constant, known as the Y-intercept. If X = 0, Y=c

Linear relationship can be positive or negative -

- Positive Linear Relationship:
  - Both independent and dependent variable increases.
- Negative Linear relationship:
  - Independent variable increases and dependent variable decreases.

Linear regression is of the following two types –

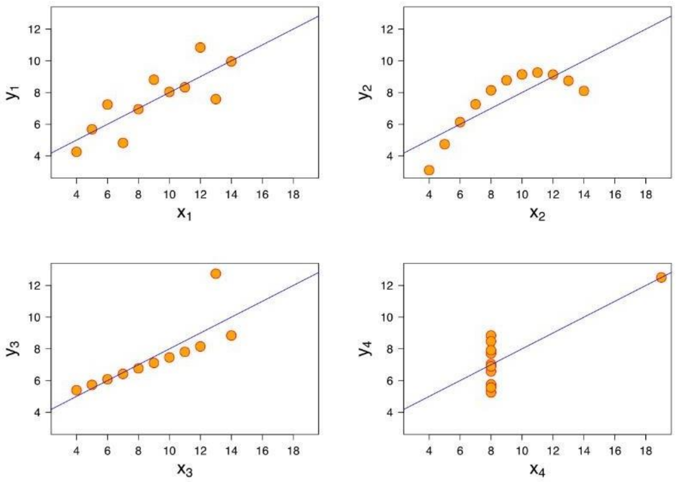
1. Simple Linear Regression- Where only one variable is involved
2. Multiple Linear Regression- Where we judge the model on various variables.

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

Answer:

Anscombe's Quartet is a group of datasets(x,y) that have the same mean, standard deviation, but they are qualitatively different. It was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. All these datasets have similar descriptive statistics. Since they are qualitatively different, so when we plot them on graph, each graph is different in its own way.

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:



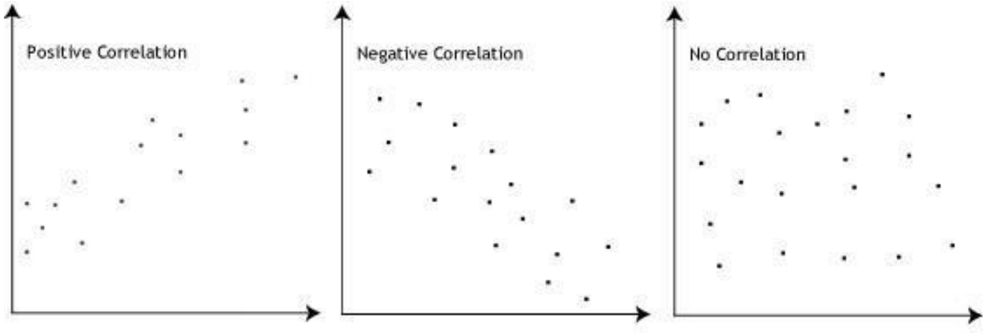
- Only First Dataset is well fitted linear model
- In Dataset III the distribution is linear, but there is outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

3.What is Pearson’s R? (3 marks)

Answer:

Pearson's r is a tells us about 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. Its value ranges between -1 to +1.

r = 1 means the data is perfectly linear with a positive slope
r = -1 means the data is perfectly linear with a negative slope
r = 0 means there is no linear association.
This is shown in the diagram below:



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

Answer:

Scaling is done due to ease of interpretation. If scaling is not done then model will not be able to calculate which value is higher. Model will calculate higher value on the basis of number and not considering its unit

Example**:** A model can consider 100paise more than 10rs.

Two major methods are :

Standardisation and MinMax Scaling(Normalisation)

S.NO.	Min-Max Scaling	Standardized scaling
1.	Data will be compromised b/w 0 and 1	Data is centered around 0, but mean and all points will not be only in range 0 and 1
2.	It handles outliers	It do not handles outliers
3.	x-xmin/xmax-xmin	x-mu/sigma

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

Answer:

VIF - the variance inflation factor -If VIF is high , that means that variable is related to other variables. The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity. (VIF) =1/ (1-R_1^2).
VIF = infinity, means there is perfect correlation. VIF>5 should be ignored, and we should immediately drop that feature from our model, if it even has high p-value

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.

3.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 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 tests.

3. **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 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 tests.