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 –
   I had created plot to get the understanding of data and get pattern analysis from it –

   1- In year 2019 - more bikes were booked as compared to year 2018.
   2- From March the bikes booking graph increases and by start of Q4 it declines.
   3- Summer & fall are the peak seasons for bikers, also holidays are when bookings are more.
   4- Also weathersit has high impact on bookings - Mist & Clean weather are peak times.
   5- Weekday/working day has less impact on the booking - there's less likely chance to figure booking pattern based on days.
   6- Booking are more when the weather is pleasant.

2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**
   Ans-
   drop_first=True drops the first column during dummy variable creation
   drop_first feature was very well used when we worked on housing data and we had a column representing furnished, unfurnished and semi-furnished, in such cases two variable can represent all 3 types. For example let's use 0 and 1
   00-fully furnished
   01-semi-furnished
   10-unfurnished

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

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**
   Ans-
   1- Using error plot
   2- Scatter plot between y_test and y_test_pred
   3- R2_score

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-year_2019
   3-season_winter
   4-month_Sep

General Subjective Questions

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

Ans-

Linear regression analysis is used to predict the value of a variable based on value of another variable. The variable you want to predict is called dependent variable and the variable which is used to predict value or the known variable is known as independent variable.

This form of analysis estimate the coefficient of the linear equation involving one or more independent variable that best predict value of dependent variable.

Linear regression fits a straight line on surface which minimise discrepancies between actual and predicted variables.

Formula – $Y = MX + C$ or $Y=\beta_0+\beta_1X$

Where
Y – dependent variable
M/ β_1 – slope
X / β_0 – independent variable
C – bias (intercept on y variable)

Linear regression is also known as statistical way of measuring the relationship between variable.

Residual – For a data point difference between actual and predicted values is called residual and can also be denoted as error.

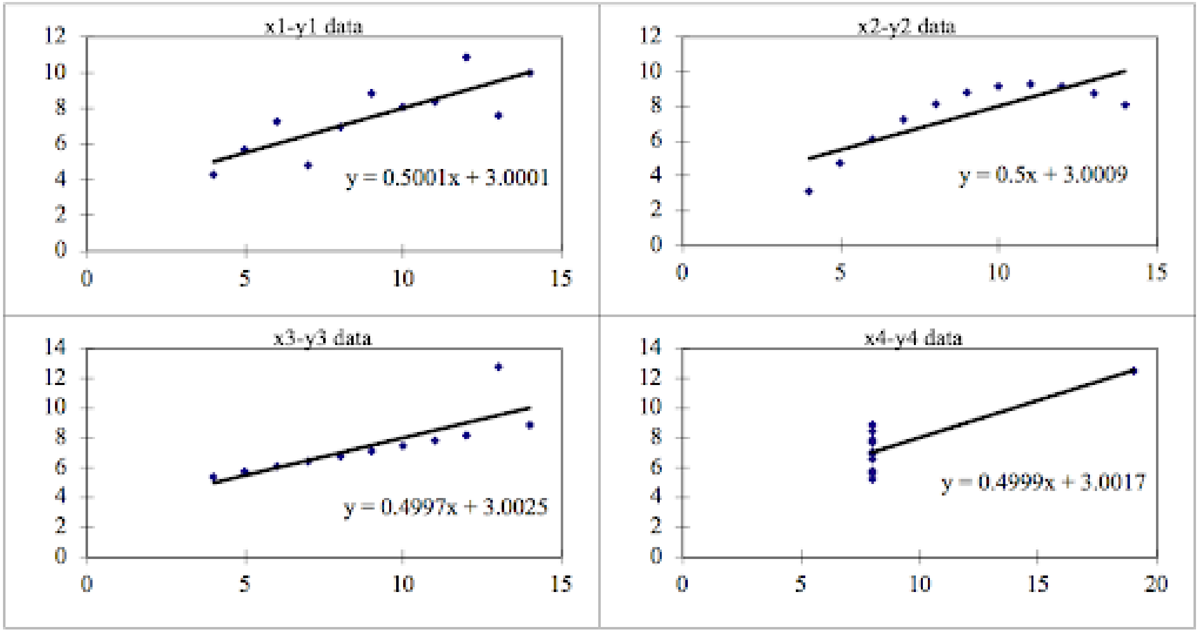
Formula – ei = yi -ypred

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

Ans-

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties.

Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.



3. What is Pearson's R ?

Ans- Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

The Pearson's correlation coefficient varies between -1 and +1 where:
r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
r = 0 means there is no linear association

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

Ans-

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

Normalization is good to use when the distribution of data does not follow a Gaussian distribution. It can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors. It is used when features are of different scale. Gets impacted due to outlier. Min and Max values are used for scaling.

Standardization can be helpful in cases where the data follows a Gaussian distribution. Though this does not have to be necessarily true. Since standardization does not have a bounding range, so, even if there are outliers in the data, they will not be affected by standardization. It is used when we expect mean to be zero. It has no effect due to outlier. Mean and SD are used for scaling.

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

   Ans-

   If the correlation between variable is very true then the VIF is observed to be infinite. If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.

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

   Ans-

   Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if 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.

   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.

   Few advantages:

   a) It can be used with sample sizes also

   b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

   It is used to check following scenarios:

   If two data sets —

   i. come from populations with a common distribution

   ii. have common location and scale

   iii. have similar distributional shapes

   iv. have similar tail behaviour