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?**
   Ans: by Analysing of the categorical variables by using box plot of all categorical variables using lib seaborn, below points infer from visualization
   ->from the year 2018 to 2019 booking of bike increased.
   ->most of the booking take place in the month of May> June > July > Aug >sept> oct>t then is started decreasing slowly from Nov to April.
   ->there is no much difference between working day and holiday.
   ->clear observe that weather sit, during the clear cloud and good climate status takes booking more then light rain or mist cloudy.

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

   Ans: To delete extra column in the dummy variables & it takes part in correlation

   df = pd.get_dummies(new_df, drop_first=True)

   it deletes n-1 column from n variables

   for example if we have three data set X,Y,Z after creating dummy

   we can interpret by , Y if both X,Y are not in the case then automatically known

   that's Z

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   Ans: temp have 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?**
   Ans: By validating with
   - Normality of error terms that is error terms should be normally distributed
   - By checking vif and p-value
   - Multicollinearity check
   - Homoscedasticity
   - Independence of residuals
5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   Ans: temp , Yr , weathersit

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

   Ans: Linear regression comes under supervised learning methods,

   There are two types of linear regression

   - 1. Simple Linear Regression
   - 2. Multiple Linear Regression

   1. Simple linear Regression

      The most elementary type of regression model is the simple linear regression which explains the relationship between a dependent variable and one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points.

      - Regression Line

        $Y = \beta_o + \beta_1.X$

      - Best fit line
      - Strength of Linear Regression Model
        1. R^2 or Coefficient of Determination
        2. Residual Standard Error (RSE)

   2. Multiple Linear Regression

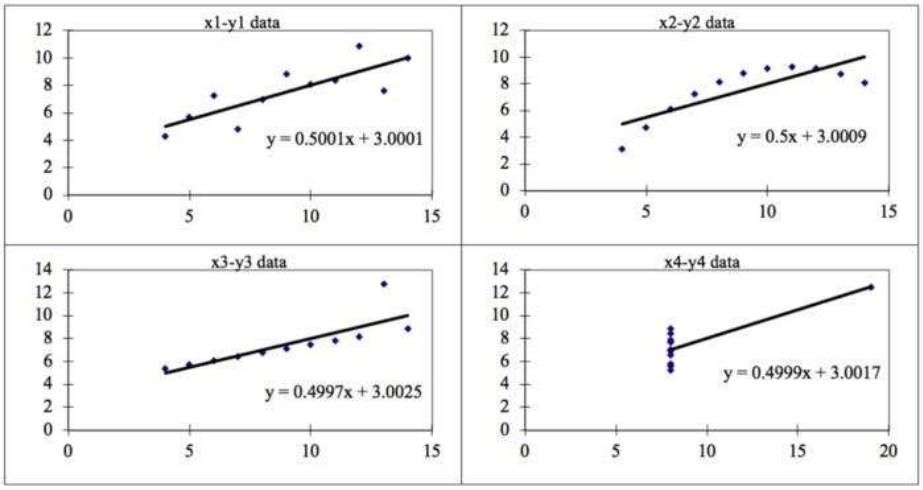
      Model building with MLR

      Assumptions:

      Multi-collinearity , auto collinearity ,normality of error terms , homosceadasticity

2. **Explain the Anscombe's quartet in detail.**

   Ans: Anscombe's quartet has 4 dataset have 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 outliers the statistical properties.



By the graph describe as:

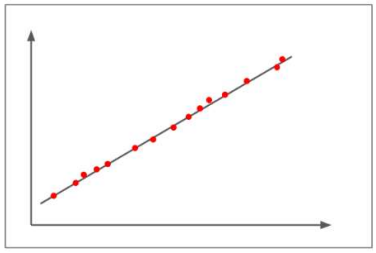
Dataset 1: best fit in linear relationship with some variance

Dataset 2: fits a neat curve but doesn't follow a linear relationship

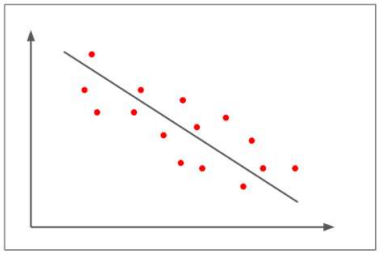
Dataset 3: looks like tight linear relationship between x and y, expect one outlier

Dataset 4: x is constant expect one Outliers.

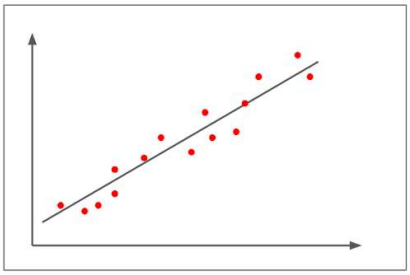
3. **What is Pearson's R?**
   Ans: Pearson correlation coefficient or Pearson's correlation or Pearson's r is defined is defined in statistics as the measurement of the strength of the relationship between two variables and their association with each other.
   Large Positive correlation:



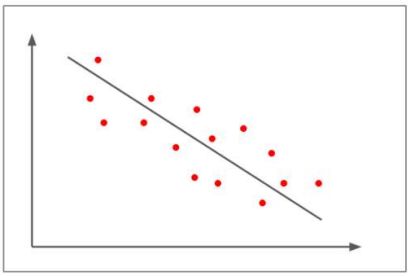
Small negative correlation:



Medium positive correlation:



Weak / no correlation:



4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

   Ans: scaling is a method of normalize the range of independent variables. It is performed to bring all the independent variables on a same scale in regression. if scaler is not done, then regression algorithm will consider greater value as higher and smaller values as lower values.

   it is important to scaling before performing model, 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 units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to same level of magnitude.

   - **Normalized scaling**
   - **Standardization scaling**
     1. **In normalization scaling**

        -Maximum and maximum value of features are used for scaling

        -scales between [0,1][-1,1]

        - min max scaler

     2.Standardization scaling

     - mean and standard deviation is used for scaling.

     - it is not bounded to certain range.

     - standard Scaler

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

Ans: if there is a perfect correlation, then VIF = inf. 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 in 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 this 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.**

   Ans: A q-q plot is a plot the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction of point below the given value.
   That is, the 0.3 quantile is the point at which 30% of the data fall below and
   70% fall above the value. A 45-degree reference line is also plotted. If the two sets come from 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