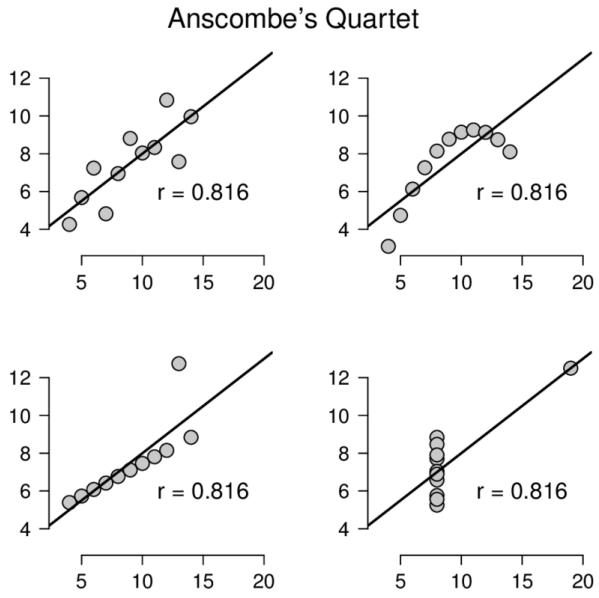
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
    a. Compared to other seasons, fall season has more bookings
    b. There is no relationship w.r.t weekdays.
    c. During snow rain less bookings are seen
    d. When there is clear weather, bookings are more
    e. Working day and holiday doesn't affect bookings.
    f. 2019 seen increase in booking compared to 2018.
2. Why is it important to use drop_first=True during dummy variable creation?
    a. When we are creating dummy variable, out of n variables only n-1 variable will be significant as the nth variable can be deduced from remaining n-1 variable. TO reduce multicollinearity, I is good to eliminate one dummy variable
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
    a. Temp is the only variable showing clear relationship with the target variable
4. How did you validate the assumptions of Linear Regression after building the model on the training set?
    a. Residual Plot pattern- no visible relationship
    b. Errors are normally distributed- normal distribution is observed
    c. Heatmap of model independent variables- no relationship
    d. Linearity between y train and y predict- R^2 0.77
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
    a. Year, Temp & Light Snow_rain

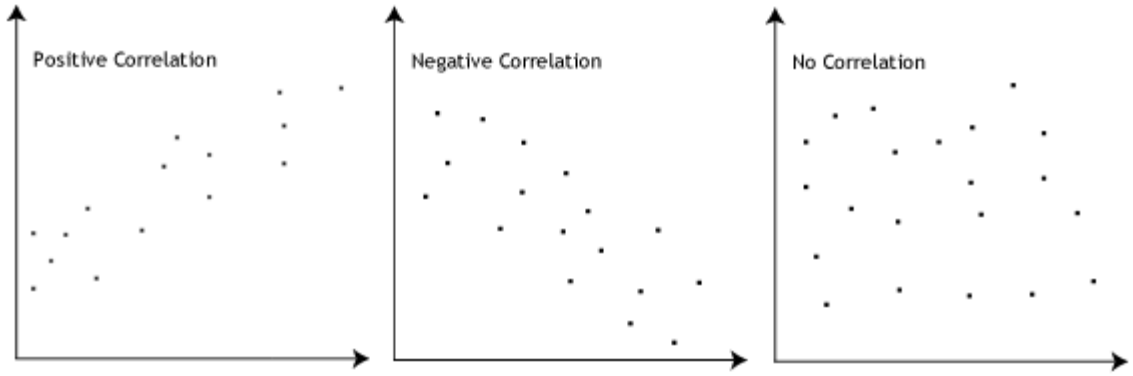
General Subjective Questions

1. Explain the linear regression algorithm in detail.
    - Linear Regression alogithm is a statistical and ML methodology to find relationship between target variables and independent variables
    - Y= a1x1+a2x2+……+b is the general form of the Linear Regression
    - As the equation is Linear in nature it is called as Linear Regression
    - a1,a2,a3 are the coefficients for the independent variables
    - x1,x2,x3 are the independent variables
    - b is the constant
    - Linear Regression mainly defined using 2 metrics : R^2 and Correlation Metric
    - Once a Model is built, using the model, we can predict values for the unknown dataset

2. Explain the Anscombe’s quartet in detail.
    a. It states on the importance of plotting the graph in addition to verification of metrics
    b. Though metrics are same, in real the predicted graphs wont explain actual population
    c. Anscombe Quartet contains 4 dataset (x,y) pair, explaining this phenomena
    d. These all dataset has mean of x is 9, mean of y is 7.50 for each dataset
    e. For all data set Correlation is 0.816
    f. In below pic, we can see clearly all data set behave in different way visually



3. What is Pearson's R?
    a. Pearson correlation is a measure of linear relationship associated with two variables
    b. It is denoted by r.
    c. It takes value from -1 to 1
    d. Positive r indicates (x increases , y increases)
    e. Negative r indicates (x increases , y increases)
    f. Zero r indicates (x increases , y acts random)



4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
    a. Scaling is used to standardise the values in the dataset within certain range. By doing so all the values are treated with same weightage during modelling. ML errors are eliminated.
    b. Normalised Scaling – Min max values are used for scaling. 0 to 1 is typically used. MimMaxScaler is the ML method.
    c. Standardised Scaling- Mean and Stadard Deviation is used for scaling. This is used when dealing with distribution of dataset
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
    a. VIF says relationship between Independent variables. If VIF is infinite it means the variables is 100% correlated. R^2 goes to 1, VIF goes to infinity

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression
    a. The quantile-quantile plot is a graphical method for determining whether two samples of data came from the same population or not
    b. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set
    c. Use: Determine whether two samples are from the same population
    d. Importance: Normal Error in Regression can be verified using QQ Plot