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. **Year**: seeing a huge growth in the no. of users in 2019 (2047742) when compared with 2018 (1243103) as visible from plot
    b. **Season**: fall and summer account for maximum number of users availing services as compared with winter and spring (as snow and rain tend be difficult for bike rides in general)
    c. **Month**: Aug, Jun, Sep, Aug attracts maximum people where the harsh winter days (Jan, Feb, Dec) see less people
    d. **Holiday**: working days attract more people towards the service as compared to holidays
    e. **Weekday**: one interesting insight here is that people using bike services on weekends are maximum. Implying that maximum people generally run household errands or fun activities on weekends and thus only explanation for these high numbers on weekends
    f. **Working day**: on comparison w.r.t. 5 working days and 2 non-working days, for obvious reasons, people using for weekdays would come out to be higher
    g. **Weather situation**: favourable weather attracts more people to use bike services as visible from the plot (clear skies has maximum people enjoying bike services as compared with cloudy and light rain or snow. Heavy rain and snow times have no user availing bike services as for obvious reasons

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

    a. Given n number of categories for any variable, the number of dummy variables which can represent the accurate information for all levels within that variable are (n-1)
    b. Drop_first, hence, helps us to drop the extra column which gets created while creating dummies, as the getdummies function creates dummy variables for all levels within that variable
    c. Taking weather column from assignment problem, Dummy_winter can be described using rest 3 dummies available as (0,0,0). So, there is scope for deletion of one extra column)

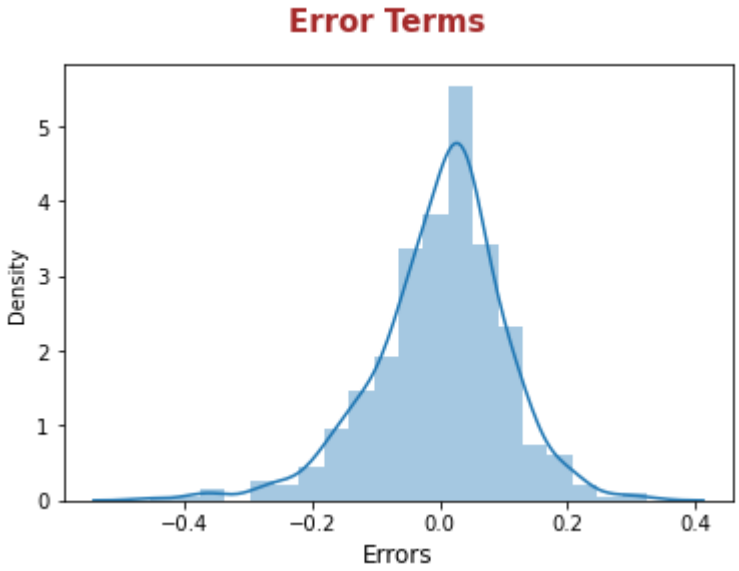
Weather	Dummy_Spring	Dummy_Summer	Dummy_Fall	Dummy_Winter
Spring	1	0	0	0
Summer	0	1	0	0
Fall	0	0	1	0
Winter	0	0	0	1

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

    a. Registered column indicates highest correlation w.r.t. cnt (0.945411)
    b. But we have not considered registered and casual variables for modelling as addition of both these columns give us value of cnt. So remove this bias, we exclude these two columns
    c. Now, temperature column shows highest correlation w.r.t cnt (0.627044) post removing those columns.

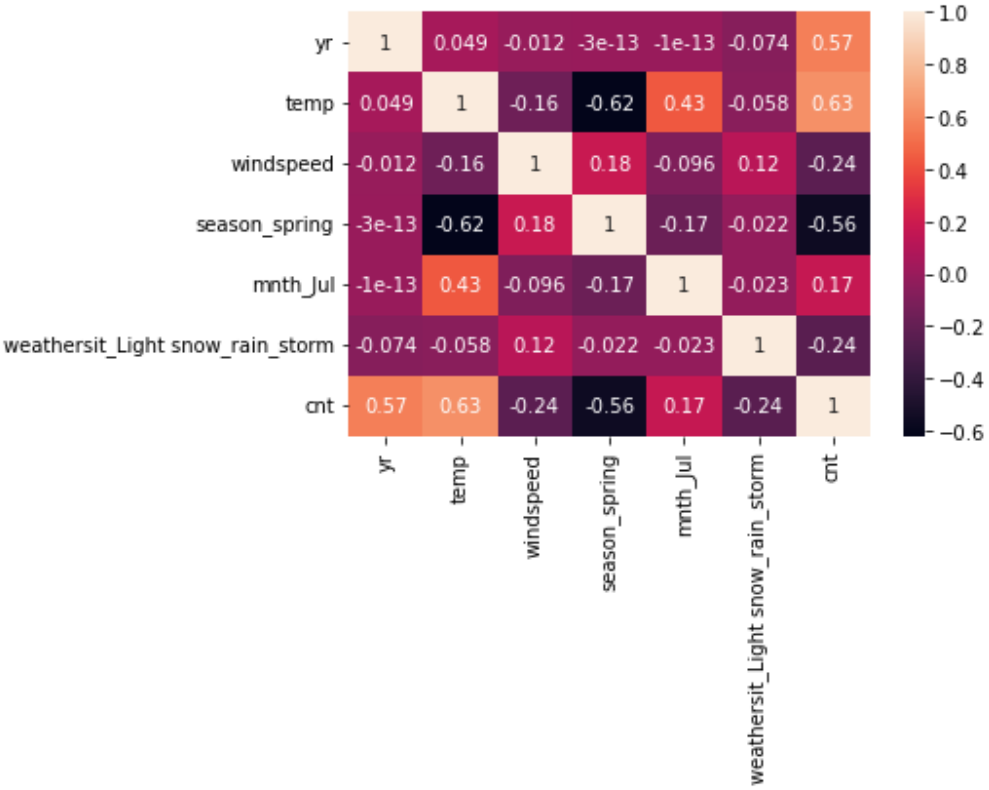
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

   a. Error terms should be normally distributed in order to make inferences from the model
   b. The error terms distribution after final model are:



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

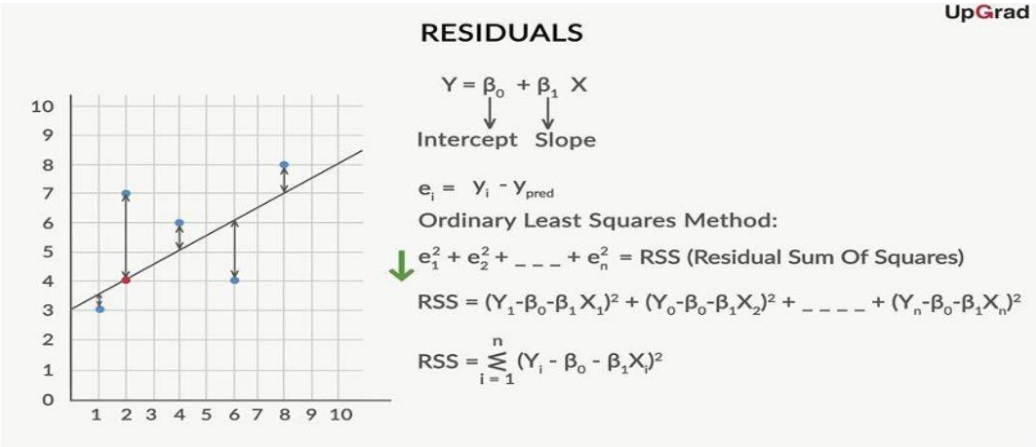
   a. Temperature, Year (both with high positive corr.) and Spring season (strong negative corr.) appear to be showing most significant among all



General Subjective Questions

1. Explain the linear regression algorithm in detail.

- The fundamental aim for regression algorithm is to make prediction just as like the other models. Here only that the nature of the variable to be predicted in continuous.
- This falls under supervised learning category where the output labels are available in dataset
- We have two broad types of linear regression:
  - **Simple Linear Regression:**
  - Here we use a single feature to make prediction.
  - And we aim to pass a straight line through the dataset which is capable of explaining or predicting most of the outcomes for independent feature
  - Since, it is a straight line, the equation holds true for this case
    Y=mx + c, where m is slope and c is intercept
  - We aim to minimize the error of all y values obtained from this equation w.r.t. to the actual values. (attaching sample fig from the lecture notes below)



  - **Multiple Linear Regression:**
  - Here we use multiple features in order to make predictions on dataset
  - It has been observed that with rise in number of features included to make predictions, the accuracy of model increases up to a certain degree.
  - But with this feat, there comes some areas where extra attention is needed, as we do not want our model to overfit or learn exactly the data fed to it as it will fail to make decent predictions.

2. Explain the Anscombe's quartet in detail.

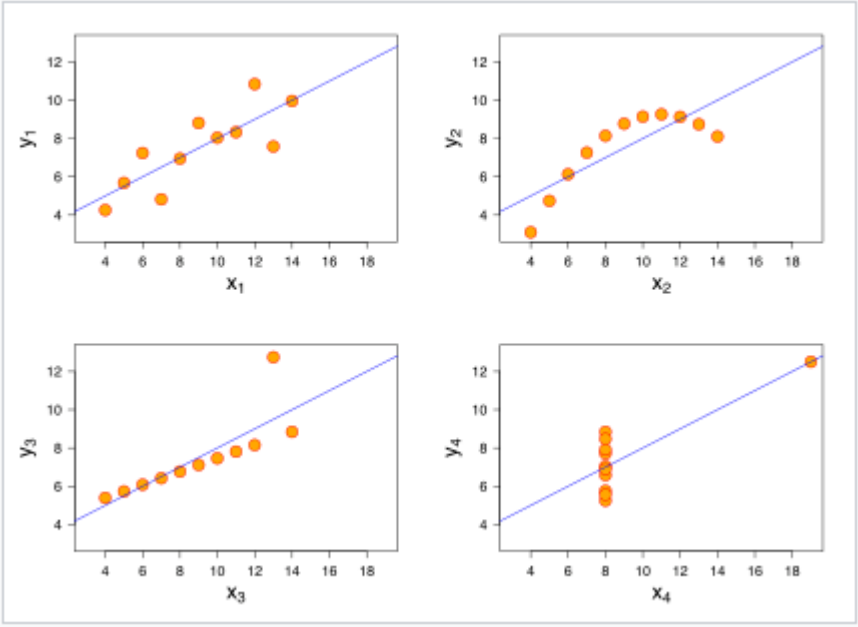
a. statistician Francis Anscombe identified 4 set of X and y variables which had similar statistical properties. But when plotted on graph, they had very different shapes or nature of plots.

b. The statistical properties for those set of data (from internet) is as:

Summary

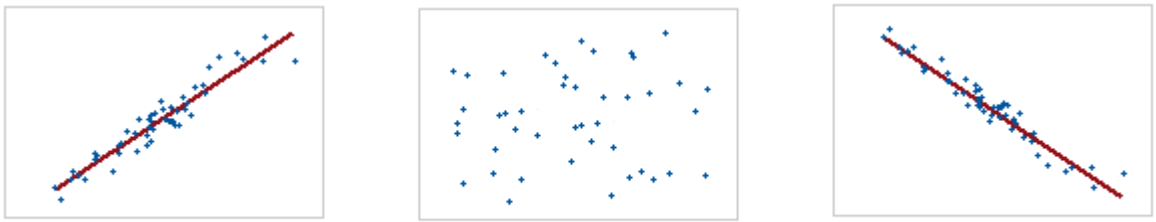
Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817

c. But their scatter plots have different shapes as below (from Wikipedia)



3. What is Pearson's R?

a. The Pearson's R or Pearson Correlation Coefficient, is used to measure the correlation between two variables.
b. Corr() function from Pandas library returns these correlation coefficients.
c. This can be used for numerical variables only and not categorical.
d. The values for Pearson's R range from -1 to 1
e. Positive value for Pearson's R implies that one variable increases with increase in another variable. And negative values imply that the other variable decreases with increase in one variable



Positive correlation | no correlation | Negative correlation

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

a. We usually have different independent features on different scales. Eg. In the assignment problem, we had different scales for temperature, humidity and windspeed features. The process of transforming all these features on the same scale, bringing in values between 0 and 1, is called scaling.
b. If we don't perform scaling, the coefficients returned for the model would be absurd and difficult to interpret. And it improves model performance too.
c. Standardized Scaling is implemented to bring the mean values to zero and standard deviation of zero. StandardScaler transformer from Scikit-Learn is used to implement this technique.
d. Normalized scaling technique is used to bring the underlying values in the range [0,1]. This is also called MinMax Scaling. And MinMaxScaler from Scikit-Learn is used to implement this technique.

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

   a. Variance Inflation Factor (VIF): Sometimes, one variable might not completely explain some other variable but can be achieved with some combination of variables. VIF is designed to exactly serve this purpose. The formulation of VIF is given below:

   VIF = 1/(1-R2)

   b. If VIF = Infinity => R2 = 1
      - Perfect correlation between features

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

   a. Q-Q(quantile-quantile) plots are used to graphically analyse and compare two probability distributions by plotting their quantiles against each other.
   b. These plots are used to find the type of distribution for a random variable whether it be a Gaussian Distribution, Uniform Distribution, Exponential Distribution, etc.
   c. These plots can also be used to determine whether different sets of data are coming from same distribution.