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?**
   For analysis of the categorical variables I have used boxplots. There are 7 categorical variables present in the dataset: 'season', 'yr', 'mnth', 'holiday', 'weekday', 'workingday', 'weathersit' . The following points canbe inferred from this variables:

   - Season 3 saw the highest booking with more than 5000 median bookings followed by season 2 and season 4.
   - Month 4,5,6,7,8,9,10,11 have median over 4000 bookings per month.
   - Weathersit 1 has the highest bookings while weathersit 3 has the lowest.There were no bookings in weathersit 4.
   - There were more bookings on a non-holiday i.e. working day as compared to holiday and this feature will also play a major role in model building.
   - Booking during week days are following a similar trend and have median between 4000 and 5000.
   - There was a significant increase in bookings in 2019.

2. **Why is it important to use drop_first=True during dummy variable creation?**
   There are two major reasons to use drop_first=True:
   a. It helps in reducing the extra columns while creating dummy variable. Too many variable makes it harder for a model to fit and hence might lead to overfitting.
   b. Also if you don't drop one column while creating dummy variable then you will get a redundant feature which will lead to multicollinearity between the dummy variables.

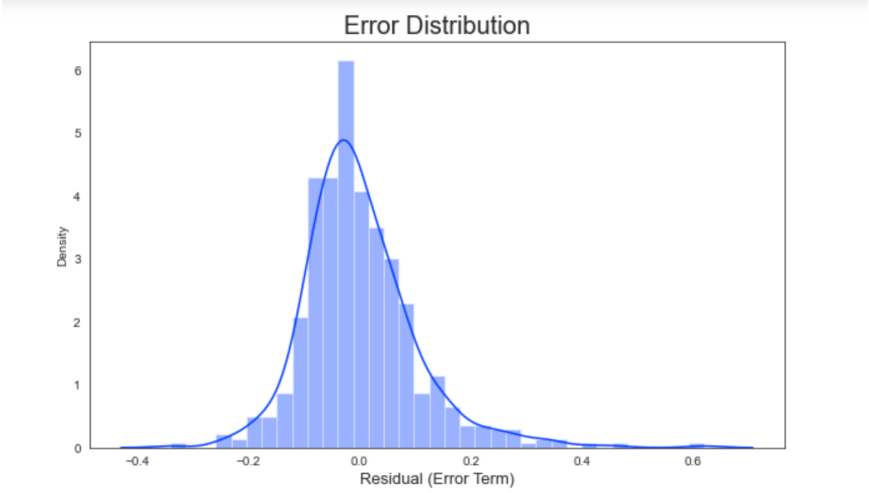
   Ex: Let say you have variable for result of coin toss. Now you are trying to create dummy variables of this variable, if you just create one, lets say HEAD and the values are 0 and 1. This completely refers that it's a tail if the value of HEAD is 0. Hence its highly recommended to create dummy variable of size (n-1) for a categorical variable with n levels.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   There is not one but two numerical variables that are highly correlated with the target variable: temp and atemp.

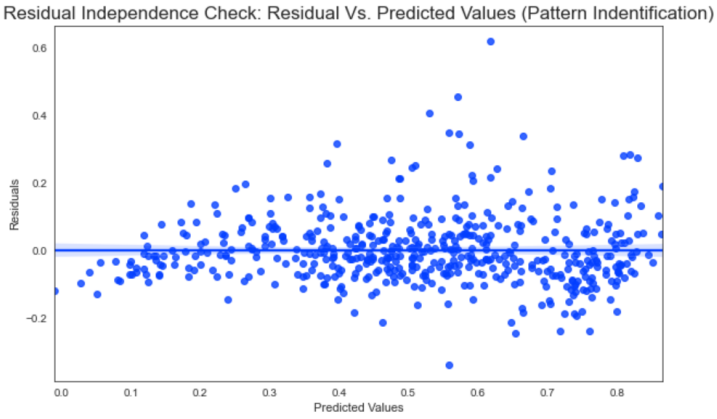
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

   I have used the following validations for checking the assumptions of Linear Regression:

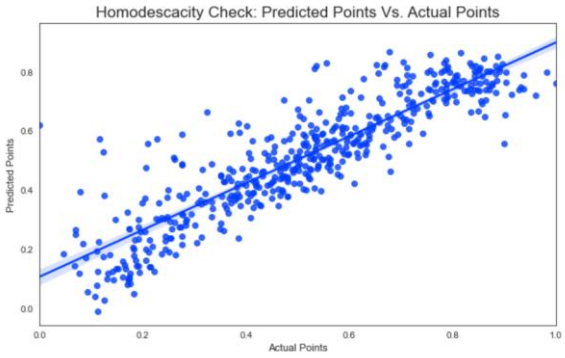
   1. Normal distribution of residuals.



   2. Independence of Residuals.



   3. Homodescacity Check.



4. Absence of multicollinearity.

	Features	VIF
0	temp	5.09
1	windspeed	4.54
2	season_summer	2.21
3	season_spring	2.08
4	yr	2.06
5	season_winter	1.76
6	mnth_Jul	1.58
7	weathersit_Fine	1.52
8	mnth_Sep	1.33
9	holiday	1.04

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

The top 3 features are:

1. Yr: coef= 0.239
2. Temp: coef=0.503
3. Season_winter: coef=0.073

General Subjective Questions

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

Linear Regression Algorithm is a type of supervised machine learning methodology which attempts to explain the linear relationship between a dependent(Y) and an independent(X) variable using a straight line. In linear regression we try to fit the best line to build a predictive model.

It is used to predict the dependent/target variable which is of continuous type using the independent variable/variables(the data can be continuous or categorical). As the model creates a straight line, the equation for the model is derived from the equation of straight line: mx+c.

Linear Regression algorithms are classified into two types:

a. Simple Linear Regression: We use simple linear regression when we have only one predictor/independent variable. The equation of SLR is given as:

$$y = \alpha + \beta x$$

Here, y= dependent variable
a(alpha) = intercept [also represented as Bo]
b(beta) = slope coefficient[also represented as B1]
x = independent/predictor variable

b. Multiple Linear Regression: We use multivariate linear regression when have multiple/morethan one predictors. The equation for MLR is given as:

$$y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_n x_n$$

Here, y= dependent variable
bo = intercept
b1 = slope coefficient for x1
x = independent/predictor variable
bn = slope coefficient for xn

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet contains four data set that have indentical statistical property like standard deviation, mean etc but when plotted all the four have different distribution. Each dataset consists of 11 points as X and y. the main purpose of devising this technique was to demonstrate the importance of plotting data while analyzing and also effect of outliers on the on statistical properties. It was constructed by Francis John "Frank" Anscombe in 1973.

The four datasets looked like following;

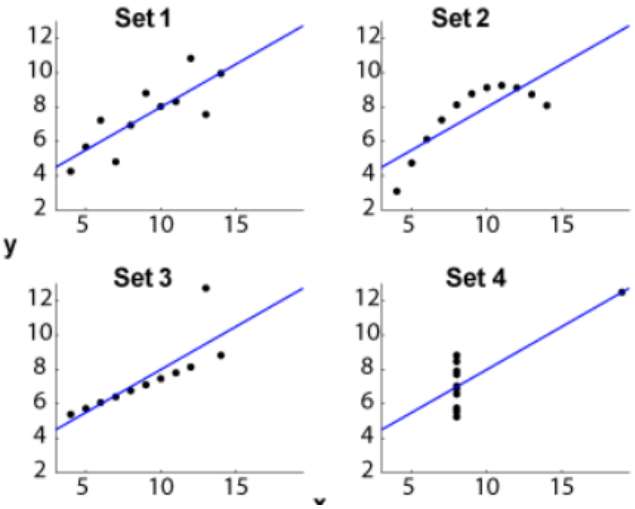
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

The descriptive statistics of the above data:

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

We can see that mean standard deviation and correlation of X and y are same for all the datasets. Now we will look at the graph to check the difference.



Now lets interpret the plots:

Set 1: There is a linear relationship between X and y

Set 2: There is a non-linear relationship between X and y

Set 3: There is almost a perfect linear relation but there is also one outlier.

Set 4: The major proportion of data is outliers and a linear model cannot handle this.

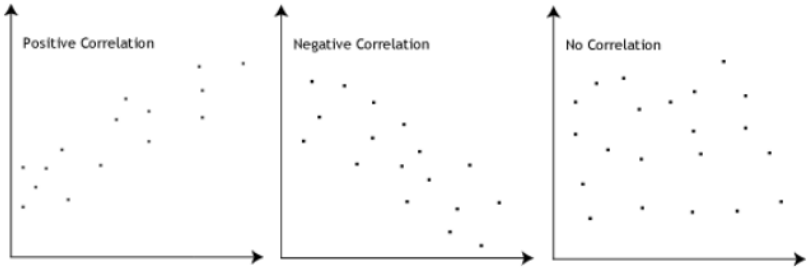
3. What is Pearson's R?

A correlation coefficient is a measure of the strength of a linear association between two variables. Pearsons R is Pearson's Coefficient which is a correlation coefficient that is majorly used while modeling a linear regression algorithm. Its is also known as Pearson Product-Moment Correlation(PPMC).

- The Pearson's coefficient ranges within -1 and +1.

- If coefficient is lesser than 0 then it indicates a negative association.
- If coefficient is 0 it represents there is no correlation.
- If coefficient is more than 0 then it indicates a positive association.

A graphical representation:



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

   Feature Scaling is a technique to normalize the independent/predictor features i.e. bringing them to a scale that there wont be any major gap which our model might misinterpret. Its performed on the data during data preprocessing or preparation step.

   It is very common for data to have a lot of varying magnitude of features like age, weight, etc. There are machine learning models that calculate the Euclidean distance between the data points. Now then there is a lot of variance in data, the distance calculations performed by the algorithm will not be as desired. So to get accurate prediction/results from models we need to convert the data into a range where there is least variance. This process of conversion is called Scaling and now we know why its required.

   Normalized Scaling: Normalized scaling or normalization is a way of scaling that rescales the data between 0 and 1. Its is also known as MinMax Scaling. We should use this kind of scaling when we know that the data is not normally distributed.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

   Standardized Scaling: Standard scaling or standardization the values are centered around the mean with a unit standard deviation. We should use this kind of scaling when our data is normally distributed.

$$X' = \frac{X - \mu}{\sigma}$$

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

   VIF or Variance Inflation Factor is the measure of multicollinearity between the independent features/ predictors of a multiple linear regression. The VIF for a feature is calculated as the ratio of of the

overall model variane to the variance of a model that includes only that single independent variable. This calculation is done for all the independent variables.

$$\text{VIF} = \frac{1}{1 - R_i^2}$$

High VIF indicates higher level of multicollinearity, so an infinite VIF indicates perfect multicollinearity. In the equation above, R2i is the r-squared of the ith feature. If this variable is explained well by other independent variables then it will have perfect correlation for which its r2 value will be 1. Putting this in equation, it will make the denominator 0 i.e. VIF = 1/0, hence the VIF is infinity.

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

To understand Q-Q plot, first we need to know what are quantiles. A quantile is a fraction where all values fall below that quantile. It can be interpreted well with a boxplots. Foe example, the median is one quantile which divides the whole data in two halfs. The data below 50% falls under median and above 50% falls above the median. Q-Q plots(Quantile-Quantile plots) are used to plot two quantiles against each other to check if the data comes from the same distribution.

- Q-Q plots are used to find the type of distribution for a random variable.
- Q-Q plots are used to find skewness of the distribution.
- These are also used to find Kurtosis(type of tail the data has).
- These are also used to determine the quantity of data that we need for analysis and modeling.