1 Assignment-based Subjective Questions

1.1 From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

- There are following categorical variable in the dataset:
    1. season
    2. workingday
    3. weathersit
    4. weekday
    5. yr
    6. holiday
    7. mnth

1. season –
    - Most favourable seasons for biking are summer and fall.
    - Spring has significant low consumption ratio
    - ***Higher targets can be planned in summer and fall with strategic advertising***

2. workingday –
    - Working day represents weekday and weekend/holiday information.
    - The registered users are renting bikes on working days whereas casual users prefer the bikes on non-working days.
    - ***Registered and casual users' identity and relevant strategy for working and not working days shall help to increase the numbers***

3. weathersit –
    - Most favourable weather condition is the clean/few clouds days.
    - ***Registered users count is comparatively high even on the light rainy days, so the assumption can be drawn that the bikes are being used for daily commute to the workplace.***

4. weekday –
    - If we consider "cnt" column we do not find any significant pattern with the weekday.
    - **However if the relation is plotted with "registered" users, we observe that bike usage is higher on working days. And with "casual" users it vice versa.**

5. yr –
    - ***Increased from 2018 to 2019***

6. holiday –
    - Holiday consumption of bikes if compared within "registered" and "casual" users then the observation is "casual" users are using bikes more on holiday.

7. mnth -
    - The bike rental ratio is higher for June, July, August, September and October months.

1.2 Why is it important to use drop_first=True during dummy variable creation?

- The drop_first = True is used while creating dummy variables to drop the base/reference category
- The reason for this is to avoid the multi-collinearity getting added into the model if all dummy variables are included
- The reference category can be easily deduced where 0 is present in a single row for all the other dummy variables of a particular category.
- ***Hence drop_first=True helps in reducing the extra column created during the dummy variable creation and hence avoid redundancy of any kind***

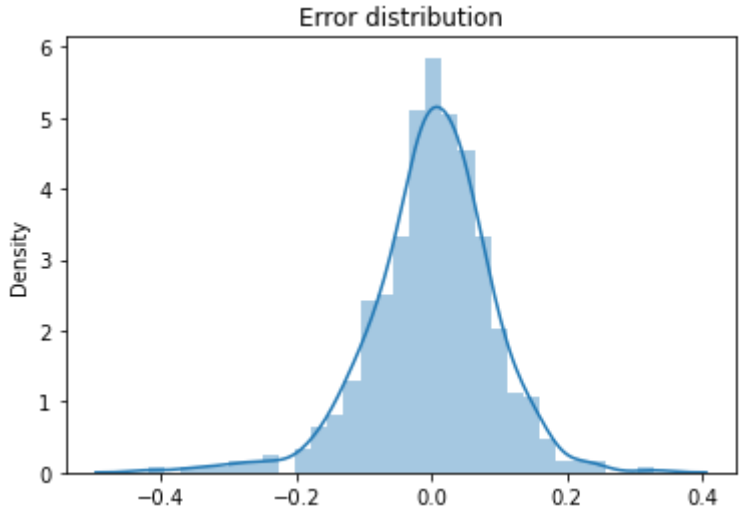
1.3 Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

The variable **"temp"** has the highest correlation with the target variable i.e. 0.63

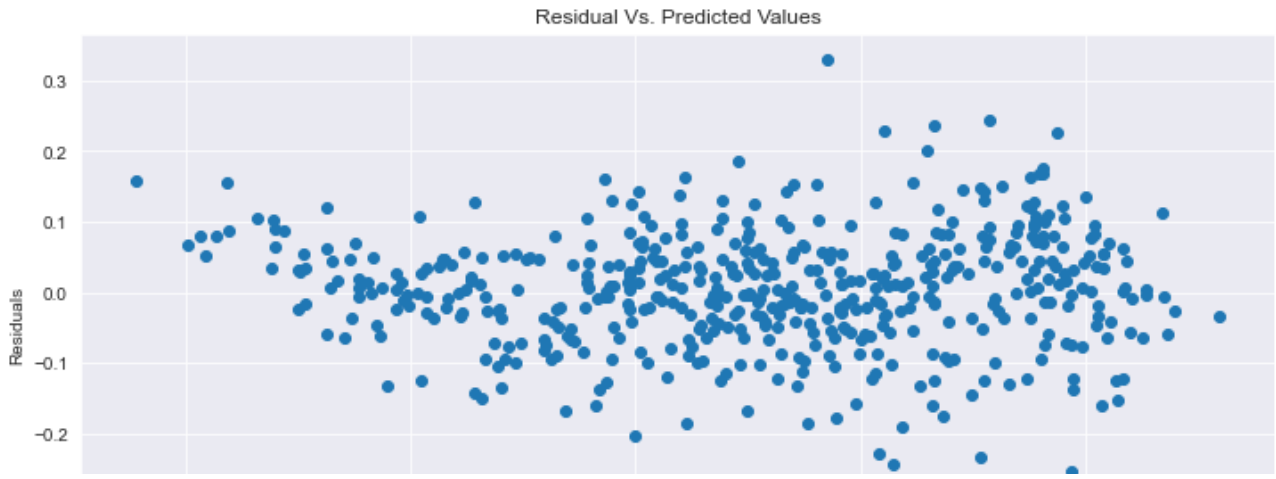
1.4 How did you validate the assumptions of Linear Regression after building the model on the training set?

- Validated the assumptions of linear regression by checking the VIF, error distribution of residuals and linear relationship between the dependent variable and a feature variable

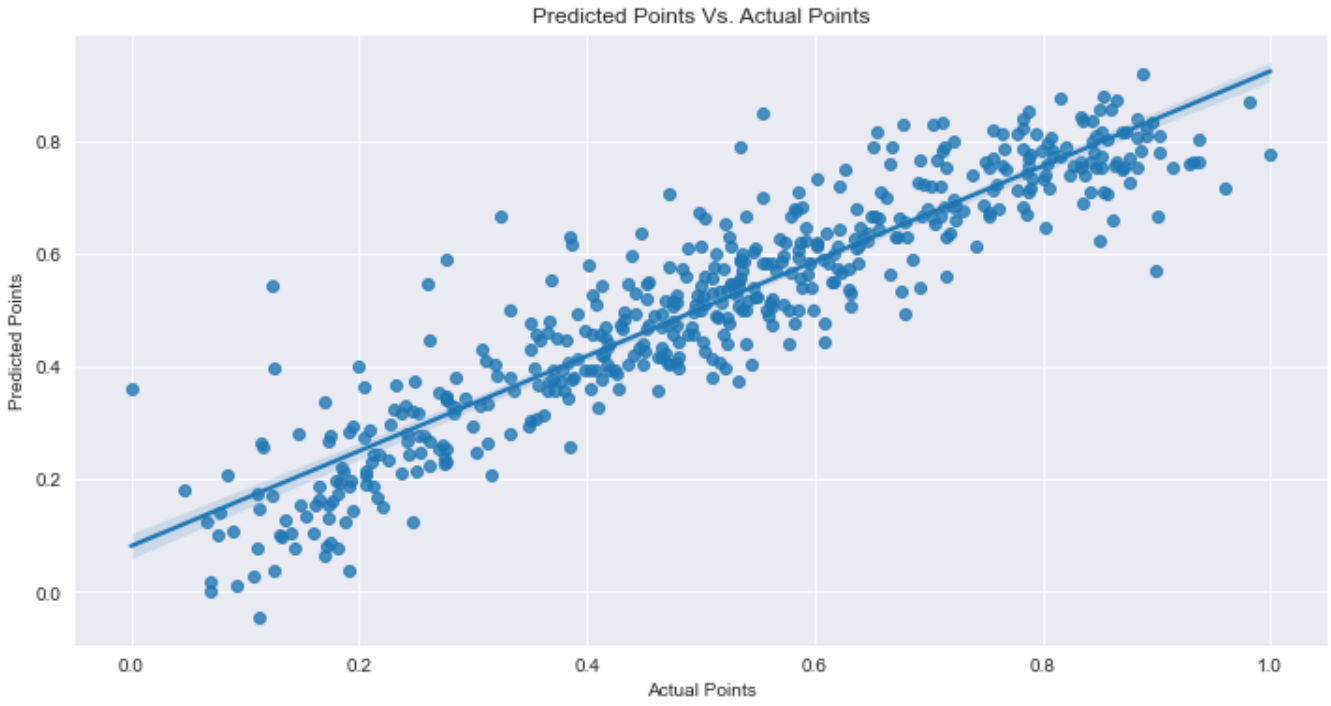
- Looking For Normality :



- Looking For Patterns in the Residuals



- Looking for Constant Variance



1.5 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 contributing significantly towards the demand of the shared bikes are:
  1. temperature
  2. year
  3. holiday

2 General Subjective Questions

2.1 Explain the linear regression algorithm in detail.

1. Linear Regression is a ***Supervised machine Learning algorithm***

2. It is Predictive model used for finding **linear relationship** between dependent and one or more independent variable

3. It is used for ***regression problems:*** When the target column is with Distribution values (contains large number of distinct values) :

    - Dependent Variable >> Continuous Data
    - Independent Variable >> Continuous / Discrete

4. Linear regression is used to find the line that **best fits** the data points on the plot

Two types of Linear Regression:

1. Simple Linear Regression:
    A. Simple linear regression is having ***only one input variable***
    B. Simple Linear Regression determine the value of one dependent variable from value of one given independent variable
    C. Y = mX + c
        - Here, Y is dependent variable
        - X is independent variable
        - c is y-intercept
        - m is gradient/coefficient/slope
2. Multiple Linear Regression:
    A. Multiple Linear regression is having ***two or more input variables***
    B. Multiple Linear Regression determine the value of one dependent variable from value of two or more independent variable
    C. Y = m1X1+m2X2+m3X3…mnXn + c
        - Here, Y is a dependent variable
        - X1, X2, X3, …Xn are independent variables
        - c is Y-intercept
        - m1,m2,m3,…mn are coefficient

- Assumptions Of linear Regression
    1. Linearity
    2. Independence
    3. No Multicolinearity
    4. Normality
    5. Homoscadesticity

2.2 Explain the Anscombe’s quartet in detail.

- Francis Anscombe realized in 1973 that only statistical measures are not good enough to depict the data sets. He created several data sets all with several identical statistical properties to illustrate the fact.

- Anscombe's quartet consists of four data sets that have nearly identical simple descriptive statistics but have very different distributions and appear very different when presented graphically.

- Each dataset consists of eleven points.

- The primary purpose of Anscombe's quartet is to illustrate the importance of looking at a set of data graphically before beginning the analysis process as the statistics merely does not give the an accurate representation of two datasets being compared.

2.3 What is Pearson's R?

- Pearson's Correlation Coefficient is used to establish a linear relationship between two quantities.
- It gives an indication of the measure of strength between two variables and the value of the coefficient can be between -1 and +1.
- **R Value : -1 to +1:**
  - Strongly correlation:
    - R > 0.7
    - R < - 0.7
  - Weakly correaltion
    - R = -0.3 to + 0.3 (0.2, 0.1, 0, -0.1, -0.2)
  - No Relation:
    - R = 0 means no relationship

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where:

N	=	number of pairs of scores
Σxy	=	sum of the products of paired scores
Σx	=	sum of x scores
Σy	=	sum of y scores
Σx^2	=	sum of squared x scores
Σy^2	=	sum of squared y scores

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

- Scaling is a technique performed in pre-processing during building a machine learning model to standardize the independent feature variables in the dataset in a fixed range.
- The dataset could have several features which are highly ranging between high magnitudes and units. If there is no scaling performed on this data, it leads to incorrect modelling as there will be some mismatch in the units of all the features involved in the model.
- The difference between normalization and standardization is that while normalization brings all the data points in a range between 0 and 1, standardization replaces the values with their Z scores.
- Normalization/Min-Max scaling – The Min max scaling normalizes the data within the range of 0 and 1. The Min max scaling helps to normalize the outliers as well.
  - MinMaxScaling:x=(x-min(x))/(max(x)-min(x))
- Standardization converges all the data points into a standard normal distribution where mean is 0 and standard deviation is 1.
  - Standardization:x=(x-mean(x))/sd(x)

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

- VIF=1/(1-R^2)
- The VIF formula clearly signifies when the VIF will be infinite :If the R2 is 1 then the VIF is infinite. The reason for R2 to be 1 is that there is a perfect correlation between 2 independent variables.
- The value of VIF is infinite when there is a perfect correlation between the two independent variables.
- This concept suggests that is there is a problem of multi-collinearity and one of these variables need to be dropped in order to define a working model for regression

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

- The quantile-quantile (Q-Q) plot are used to plot quantiles of a sample distribution with a theoretical distribution to determine if any dataset concerned follows any distribution such as normal, uniform or exponential distribution.
- It helps us determine if two datasets follow the same kind of distribution.
- It also helps to find out if the errors in dataset are normal in nature or not.
- Interpretations
  - Similar distribution: If all the data points of quantile are lying around the straight line at an angle of 45 degree from x-axis.
  - Y values < X values: If y-values quantiles are lower than x-values quantiles.
  - X values < Y values: If x-values quantiles are lower than y-values quantiles.
  - Different distributions – If all the data points are lying away from the straight line.
- Advantages
  - Distribution aspects like loc, scale shifts, symmetry changes and the outliers all can be daintified from the single plot.
  - The plot has a provision to mention the sample size as well.