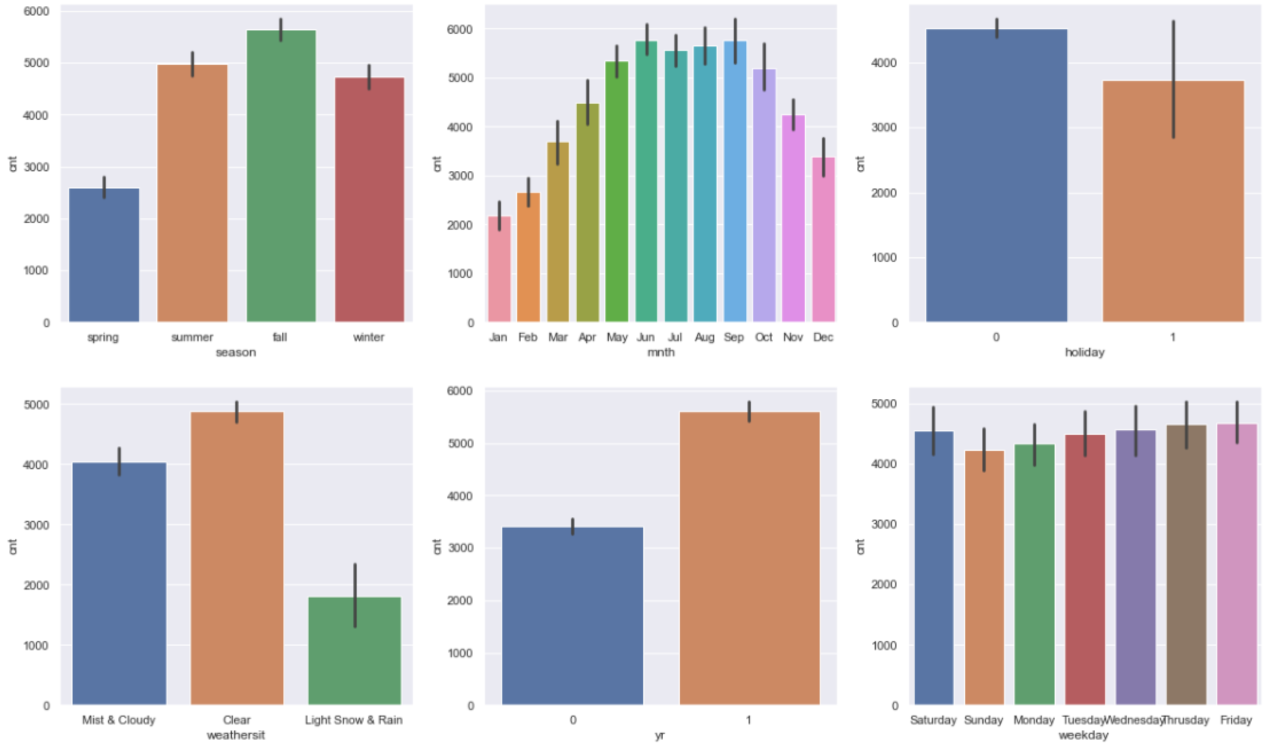
BIKE SHARING DEMAND

Submitted By: - Aarushi Gupta

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?**
   ‘season’, ‘mnth’, ‘holiday’, ‘weathersit’, ‘yr’ and ‘weekday’ are categorical variables from the dataset. These variables were visualized using bar plots where x axis represents the categories and y axis as bike counts.
   - **Season**: Maximum bike counts is for fall season and spring season has least number of bikes counts with summer and winters being intermediate.

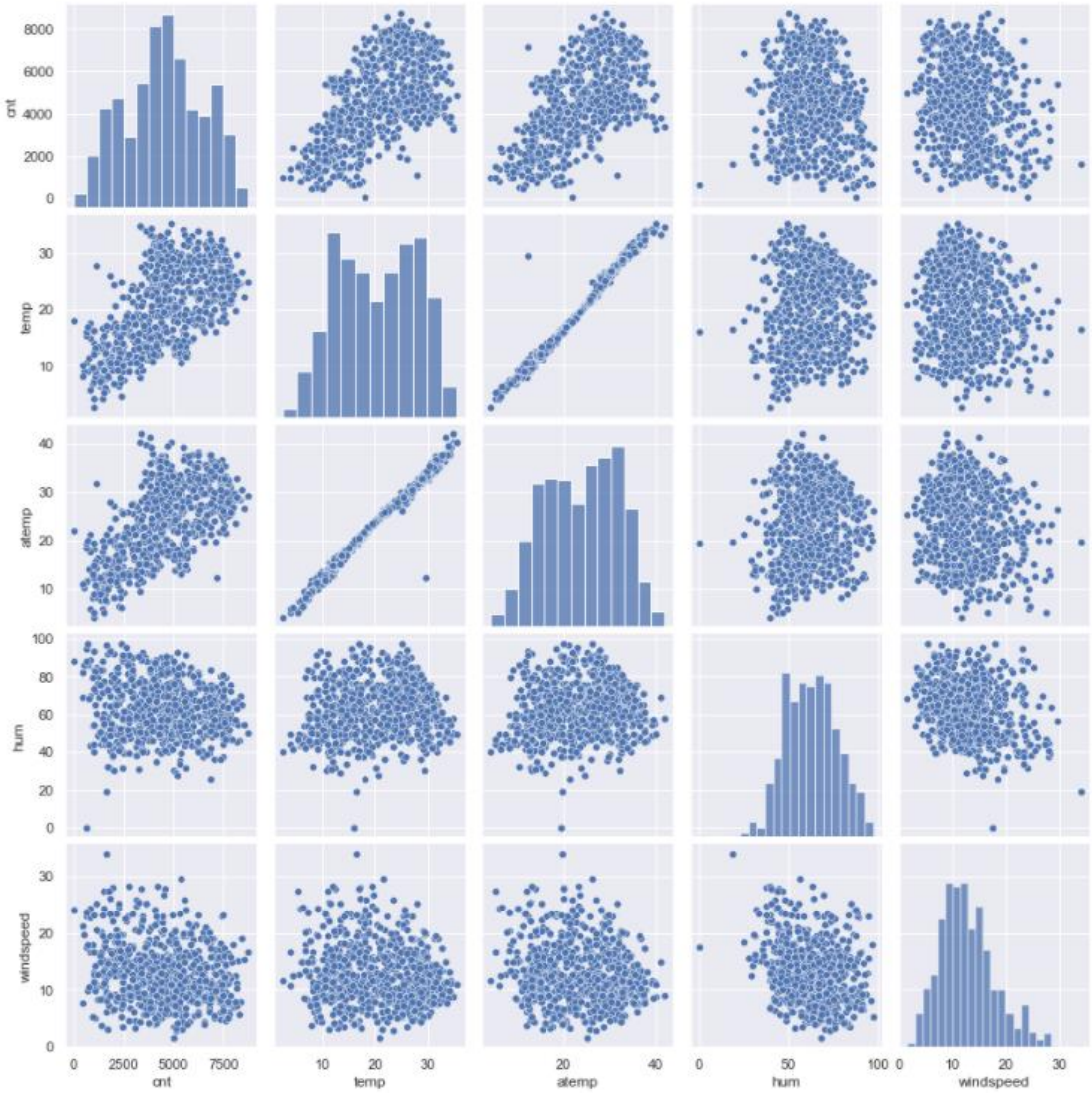


   - **Mnth:** June and September have maximum bike counts, January and December has least bike counts since there is heavy snowfall during these months.
   - **Holiday**: No holiday days have a greater number of bikes counts since people might need bike to travel to work, markets during no holiday days.
   - **Weathersit:** Clear weather situation has maximum number of bikes counts whereas It is risky to hire bike during mist, cloudy, light snow and rain weather situations.
   - **Yr:** Year 2019 has a greater number of bikes counts than the previous year.
   - **Weekday**: Sunday has least bike counts since people prefer to stay home during weekends and Monday has large number of bikes counts since people hire bike to travel to offices or work-related travelling.

2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**
   drop_first = True helps in reducing the extra column created during dummy variable creation. Thus, it helps in reducing the correlations (redundancy) among dummy variables. Dropping first categorical variable is possible because if your every dummy variable is 0 which automatically means that first value would have been 1. Thus, it helps in dealing with multicollinearity.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

   As can be seen from below pair-plot, ‘temp’ and ‘atemp’ has highest correlation (almost linear) with the target variable (here ‘cnt’).

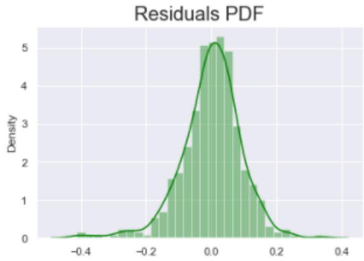


4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

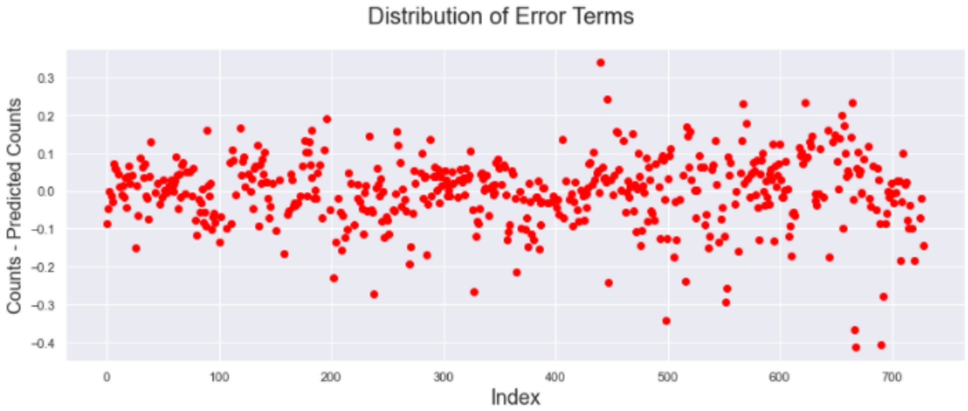
   a. There is a linear and additive relationship between independent variables (such as temp, yr, weathersit, weekday, mnth) and dependent variables(cnt) given by the following equation of a straight line.

      cnt = 0.132812 + (yr × 0.233324) + (temp × 0.522373) − (windspeed × 0.151594) + (season_winter × 0.138459) - (season_summer × 0.102020) + (mnth_Sep × 0.110481) + (mnth_Aug × 0.052873)- (weathersit_Light Snow and Rain × 0.282943) − (weathersit_Mist & Cloudy × 0.081710) - (weekday_Sunday × 0.045317)

b. This relationship suggests that a change in dependent variables due to one unit change in independent variables in constant. The error terms are normally distributed and centered around 0 (mean = 0). We validated this assumption using distribution plot of seaborn known as **residual analysis**. (residue = y_actual- y_test).



c. The error terms (residues) should show constant variance (homoscedasticity). If there is a visible pattern shown then it is known as heteroscedasticity.



5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

The top 3 features contributing significantly towards explaining the demand of the shared bikes are:

- '***temp'*** with coefficient of (0.522373).
- ***'weathersit_Light Snow and Rain'*** with coefficient of (- 0.282943).
- ***'yr'*** with coefficient of (0.233324)

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**

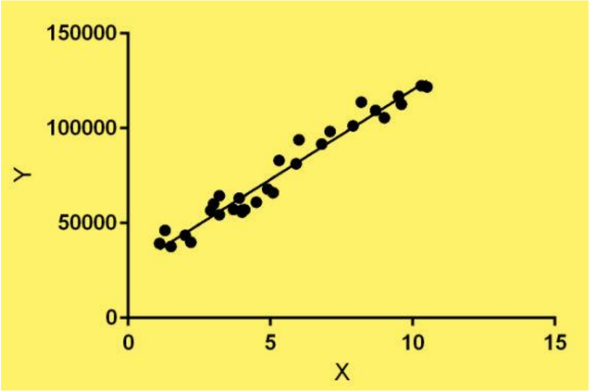
Linear regression is a simple statistical regression method used for predictive analysis of continuous variables and shows the relationship between the continuous variables by fitting the best fit straight line through the plots. It shows the linear relationship between the independent variable(or features) (X-axis) and the dependent variable (predictor variable)(Y-axis), consequently called linear regression.

If there is a single input variable (x), such linear regression is known as ***Simple linear regression***.

If there is more than one input variable, such linear regression is known as ***multiple linear regression*.**

Both simple and multiple linear regression model aims to give a sloped best fit straight line describing the relationship within the variables.

The graph presents the linear relationship between the dependent variable on y axis and independent variables on x axis. When the value of x (**independent variable**) increases, the value of y (**dependent variable**) is likewise increasing. The line passing through most of the points is referred to as the best fit straight line. Based on the given data points, we try to plot a line that fits the points in the best possible way.



The line can be modelled based on the linear equation shown below.

$$y = a_0 + a_1 * x$$

The motive of the linear regression algorithm is to find the best values for a_0 and a_1.

In Multiple Linear Regression, the target variable(Y) is a linear combination of multiple predictor (independent) variables $x_1, x_2, x_3, ..., x_n$. Since it is an enhancement of Simple Linear Regression, so the same is applied for the multiple linear regression equation, the equation becomes:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots$$

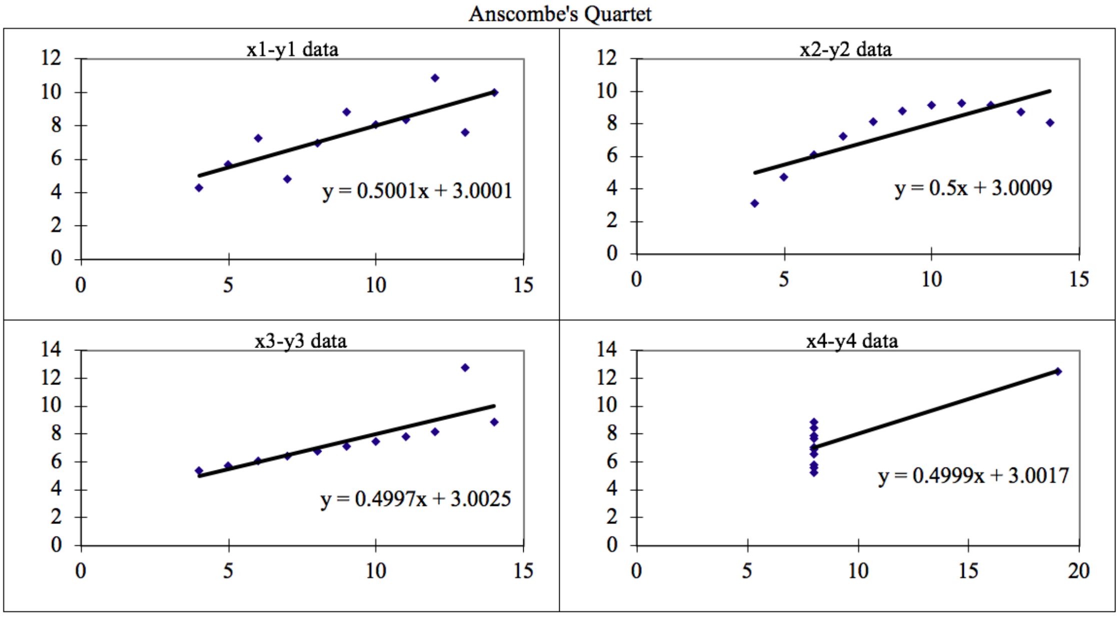
Y= Output/Response variable

$\beta_0, \beta_1, \beta_2, \beta_3, \beta_n$....= Coefficients of the model.

X_1, X_2, X_3, X_4,...= Various Independent/feature variable

2. **Explain the Anscombe's quartet in detail. (3 marks)**

Anscombe's Quartet is defined as a group of four datasets which might be nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the linear regression model if built. They have very different distributions and appear differently when plotted on scatter plots.



It was developed to emphasize both the importance of plotting data before analysing and the effect of outliers and other influential observations on statistical properties.

The four datasets can be described as:
Dataset 1: this *fits* the linear regression model pretty well.
Dataset 2: this *could not fit* linear regression model on the data quite well as the data is non-linear.
Dataset 3: shows the *outliers* involved in the dataset which *cannot be handled* by linear regression model
Dataset 4: shows the *outliers* involved in the dataset which *cannot be handled* by linear regression model

3. **What is Pearson's R? (3 marks)**

In Statistics, the Pearson's Correlation Coefficient is also referred to as **Pearson's r,** the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

There are certain requirements for Pearson's Correlation Coefficient:

- Scale of measurement should be interval or ratio
- Variables should be approximately normally distributed
- The association should be linear
- There should be no outliers in the data

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

where,
N = the number of pairs of scores
Σxy = the sum of the products of paired scores
Σx = the sum of x scores
Σy = the sum of y scores
Σx2 = the sum of squared x scores
Σy2 = the sum of squared y scores

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**

Scaling (or Feature Scaling) means to normalize the range of independent variables. It is generally performed during data pre-processing. Scaling of the data makes it easy for the model to learn as well as understand the problem within the given range of features.

Normalization- It is a scaling technique in which the data is scaled such that they end up being in the range of 0 and 1. MinMaxScaler() method in python can be used for normalization.

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

where X_{min} and X_{max} are the minimum and maximum values of the respective feature

Standardization – It is a scaling technique where values are centered around 0 with the standard deviation of 1. In this case, values are not restricted to the particular range. StandardScaler() method in python is used for standardization.

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

where μ is the mean and σ is the standard deviation for that particular feature.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.

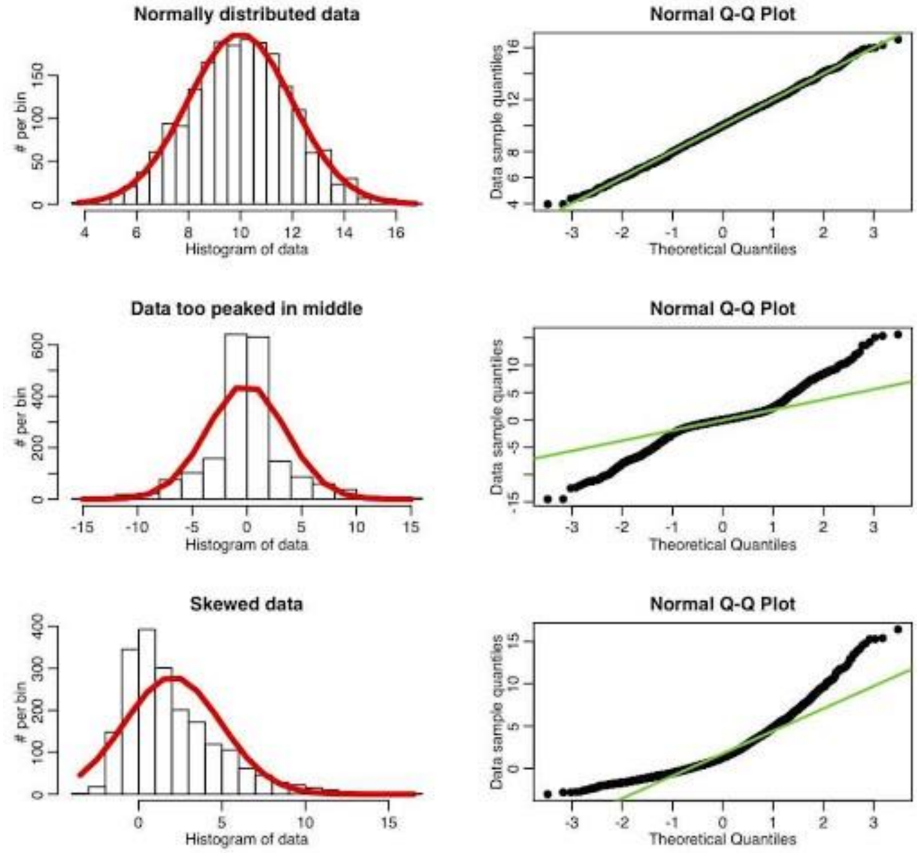
$$VIF_i = \frac{1}{1 - R_i^2}$$

From the above formula it can be seen that when value of R-squared is 1, then the VIF becomes infinity. This case happens when the linear regression line passes through each and every data points of the given dataset.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**

Q-Q (quantile-quantile) plot is a probability plot which is used to find the type of distribution for a random variable such as whether the distribution is Uniform, Gaussian, Exponential or even Pareto distribution. In simple words, it plots the quantile of sample distribution against quantiles of theoretical distributions.

It is a scatter plot created by plotting two sets of quantiles against one another. If both the quantiles follow the same distribution, it can be seen the points forming a line that is roughly straight.


Normally distributed data
# per bin
Histogram of data
Normal Q-Q Plot
Data sample quantiles
Theoretical Quantiles
Data too peaked in middle
# per bin
Histogram of data
Normal Q-Q Plot
Data sample quantiles
Theoretical Quantiles
Skewed data
# per bin
Histogram of data
Normal Q-Q Plot
Data sample quantiles
Theoretical Quantiles