# Linear Regression Assignment

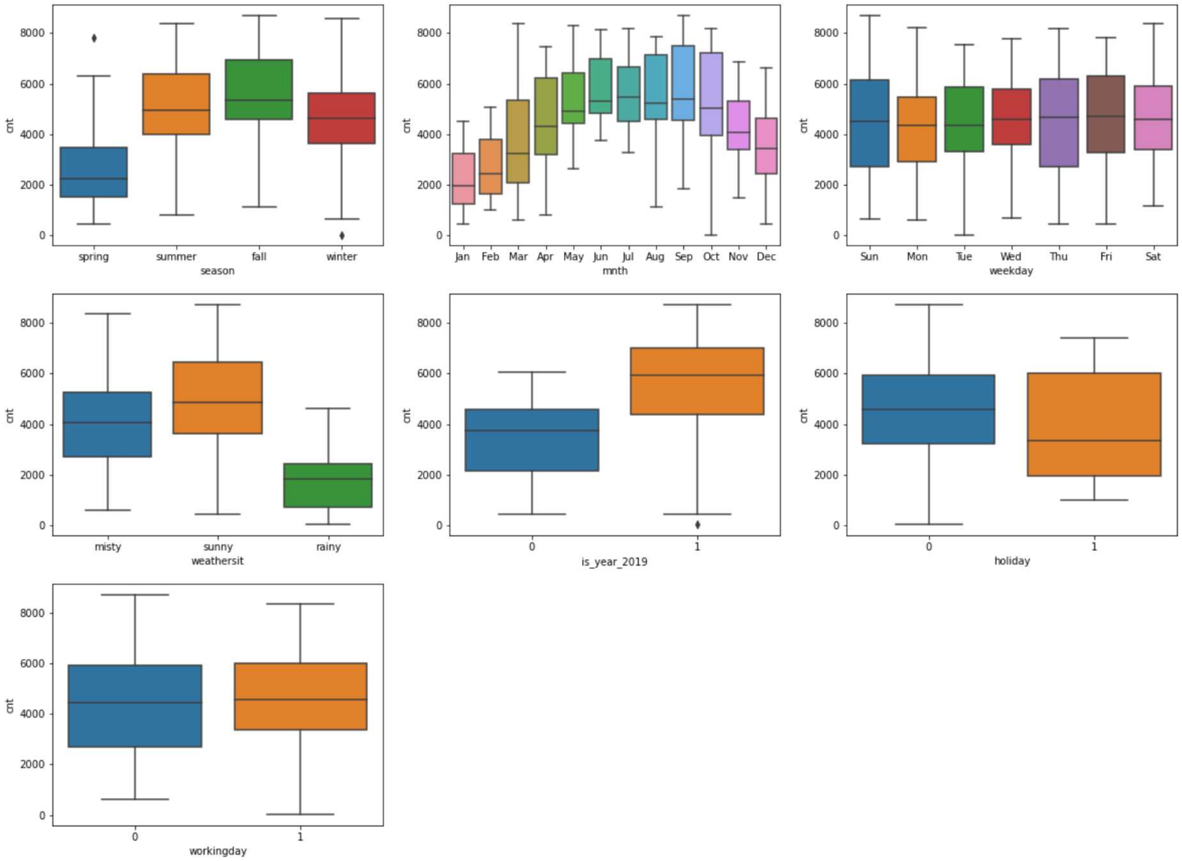
Submitted by Ayush Mandowara

## Assignment-based Subjective Questions

### Q1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

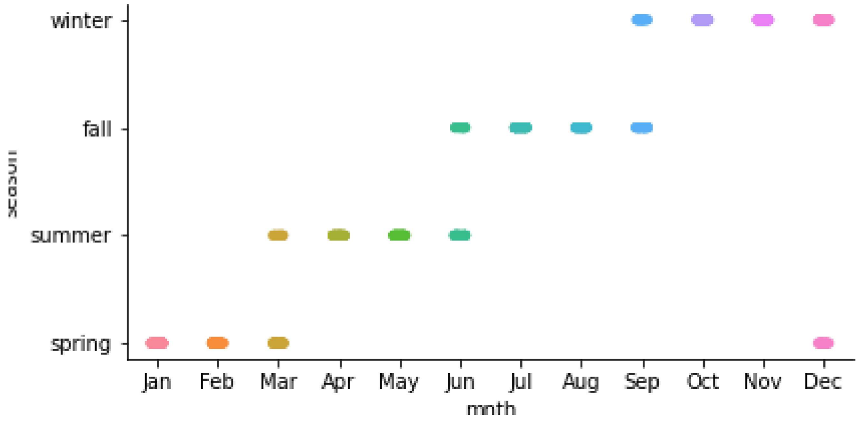
The following is the categorical variables in the dataset:

> season, mnth, weekday, weathersit, is_year_2019, holiday, workingday



The following observations were made on the categorical data in the EDA step:

- `workingday` medians are similar for off/working hence it is possibly low impact variable
- all `weekdays` have somewhat similar medians.
    - Sunday has the largest upper fence.
    - IQR for Wednesday is the smallest, with values ranging from ~3500 to ~5500
    - Tuesday has the lowest lower fence.
- Summer and Fall are the best `seasons` for bike rentals while Spring is the worst
- holiday, year, month and weather have significant variations and hence could be good predictors
    - `holiday` : the 25th-50th quantile is much higher when it is not a holiday, implying that people generally rent a bike when it is not a holiday.
    - `year` : the 75th quantile for year 2018, is the 25the quantile for year 2019 (approximately), hence we can say that demand for bike rentals is increasing significantly year by year
    - `month` : May, June, August, September, October have large upper boxes
        - the month and season variables are aligning correctly in terms of demand



- `weather` : sunny weather is the best, while rainy weather is the worst for bike rentals

Looking at the final model parameters, we can say that most of our analysis holds true.

| Top Parameters | Coefficient |
|---|---|
| temp | 0.528 |
| is_year_2019 | 0.230 |
| weathersit_rainy | -0.188 |
| windspeed | -0.181 |
| hum | -0.159 |
| season_winter | 0.100 |
| mnth_Sep | 0.082 |
| holiday | -0.059 |

| Top Parameters | Coefficient |
|---|---|
| weathersit_sunny | 0.058 |
| season_spring | -0.055 |
| mnth_Jul | -0.055 |
| weekday_Sun | 0.053 |
| season_summer | 0.053 |
| workingday | 0.043 |

---

**Q2: Why is it important to use drop_first=True during dummy variable creation?**

In general, whenever we have categorical variables with k-levels, it can be represented by k-1 variables.
Example:

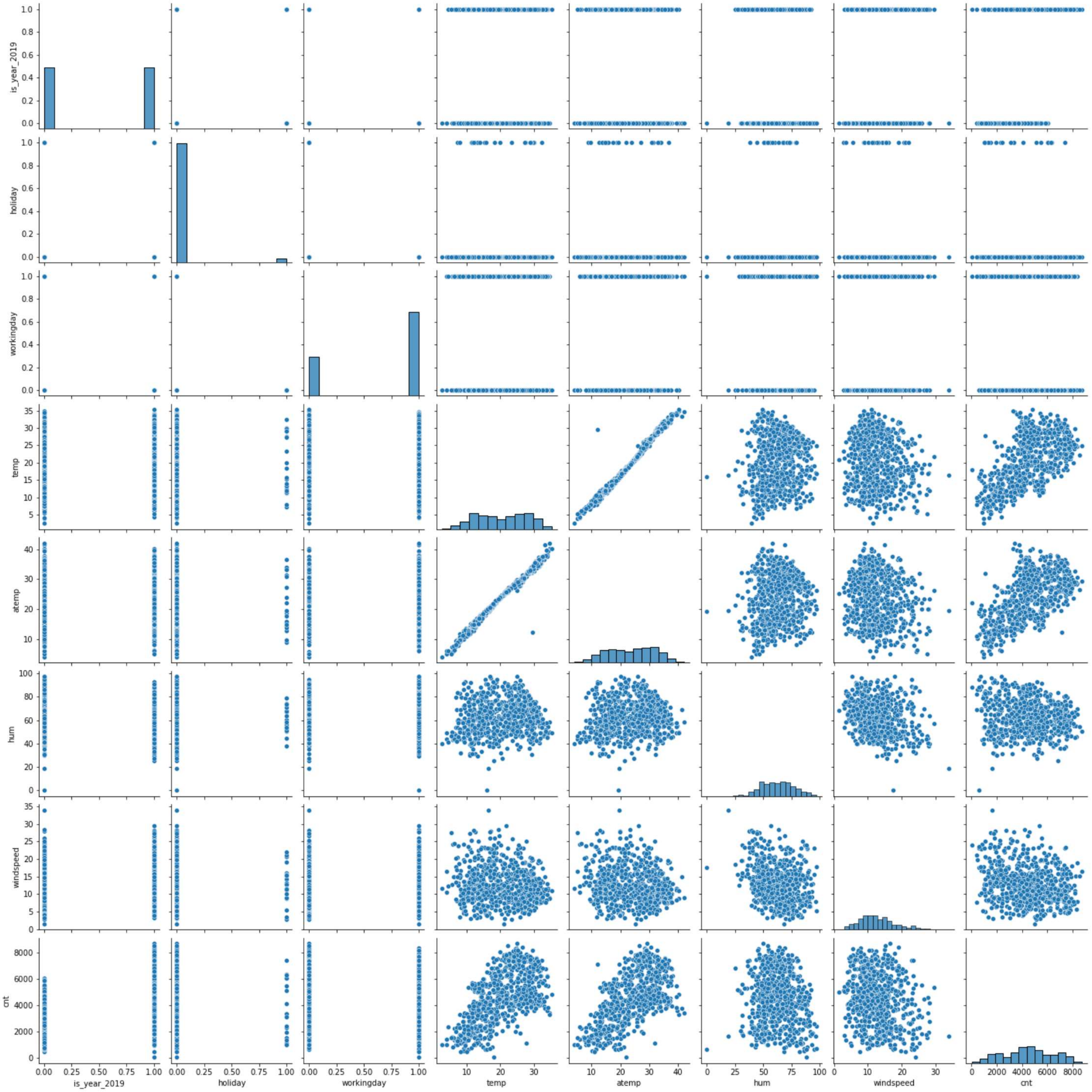
| Value | Indicator Variable |
|---|---|
| Gender | **Girl** |
| Boy | 0 |
| Girl | 1 |

| Value | Indicator | Variable |
|---|---|---|
| Season | **Sunny** | **Rainy** |
| Windy | 0 | 0 |
| Rainy | 1 | 0 |
| Sunny | 0 | 0 |

- When we use `pd.get_dummies()` it gives k dummy variables for k levels.
- Using `drop_first=True` ensures that one column (first one) is reduced, giving us k-1 levels, from which the data can be interpreted just as well.
- Moreover, it also helps in reducing the correlation among the dummy variables created.

---

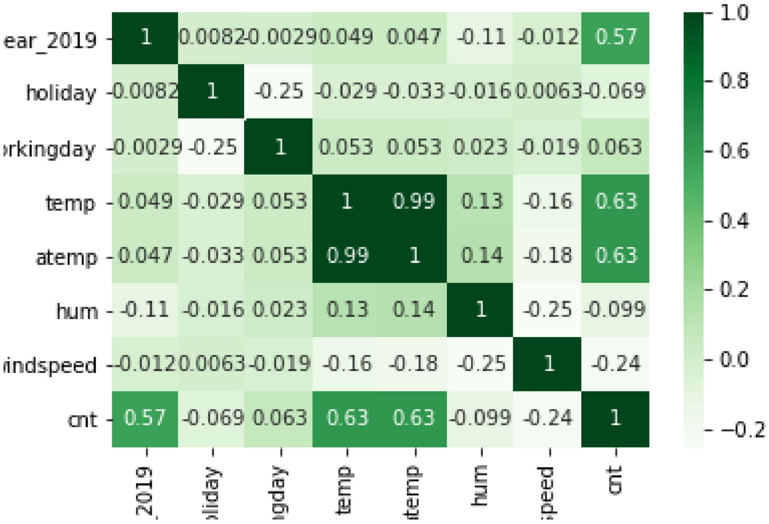
**Q3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**



Looking at the pair plot, we can see that `temperature` is the most highly correlated variable with the target variable.
It is a positive relation, i.e., higher the temperature, higher the number of bike rentals

The same was also confirmed with

- correlation matrix (highest correlation i.e. 0.63)



- when the model was built (highest coefficient i.e 0.528)

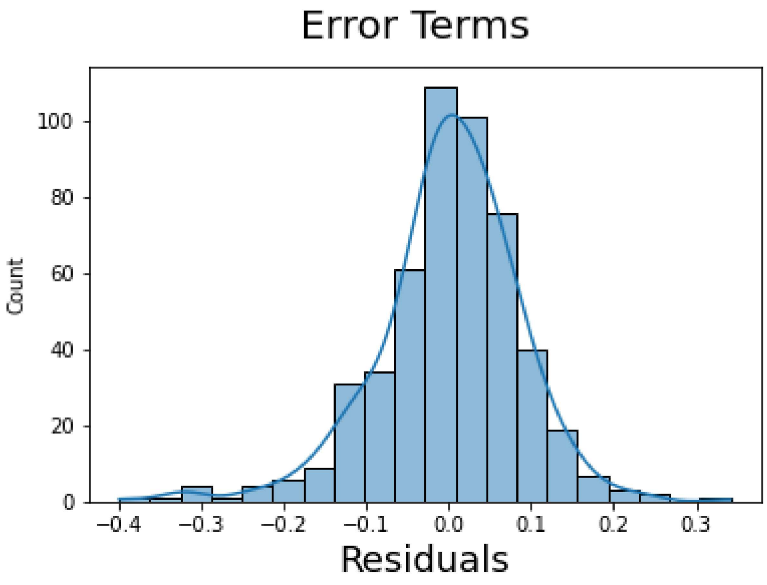
---

**Q4: How did you validate the assumptions of Linear Regression after building the model on the training set?**

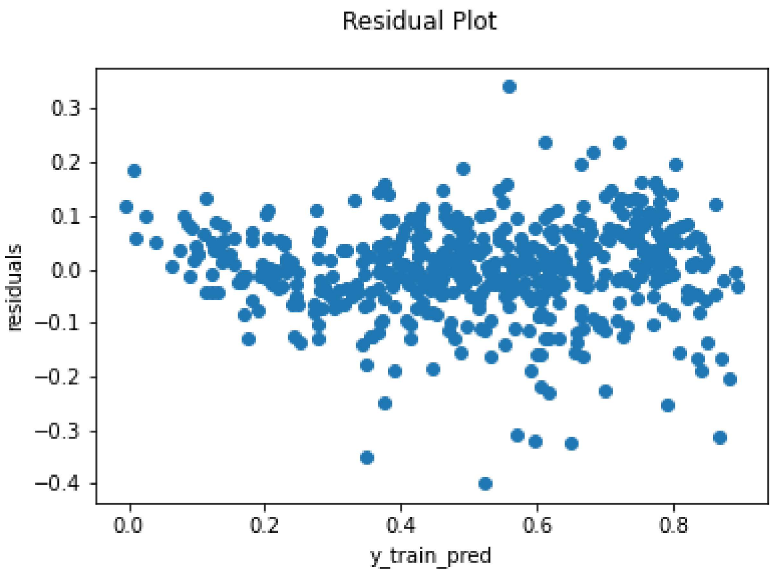
The assumptions were validated by

- plotting histogram for residuals to check normality and mean
- plotting scatter plot between residuals and fitted line
- Check for multicollinearity among independent variables using pair plot on X_train (i.e. training dataset - target)
- Check for linearity between independent x and dependent y using pair plot on X_train + target

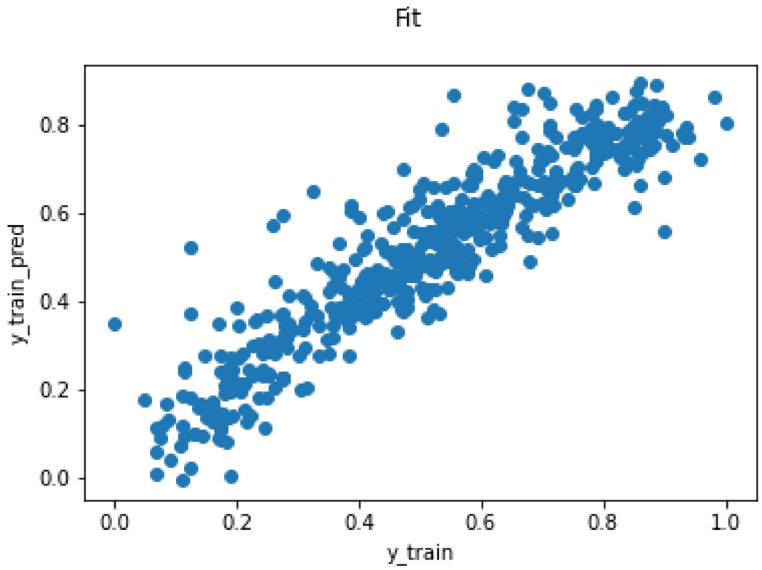
Moreover, To check the fit of the data, we plot the scatter plot between predicted values on training set and actual values on training set.



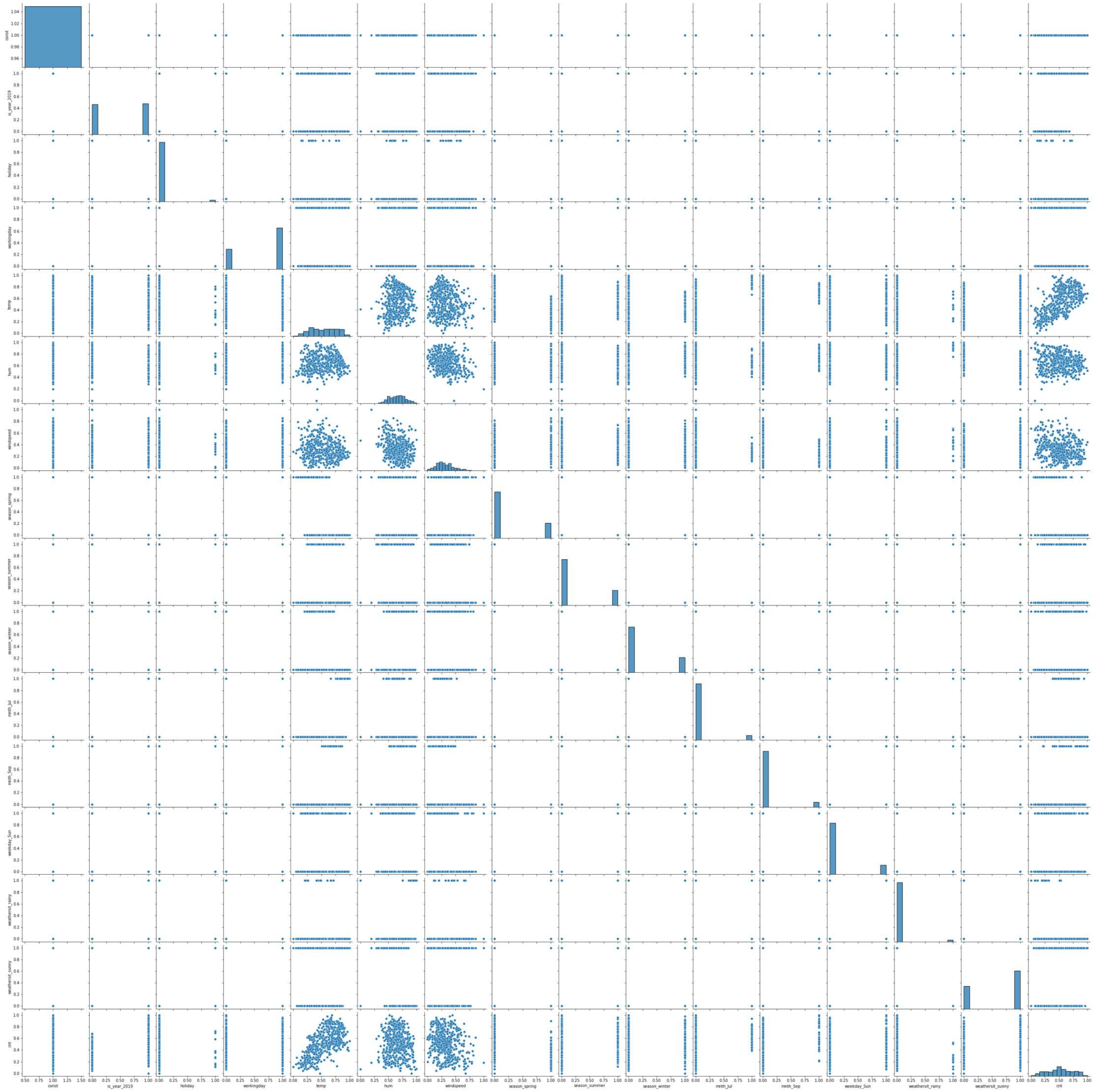
- Residuals are normally distributed, and have a mean value of zero



- Residuals have a constant variance



- y_train and y_train_pred are tightly related, which means model fits well on training set



- No multicollinearity among independent X variables
- Some independent X variables have linear relationship with dependent Y

---

### Q5: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Looking at the final model parameters:

| Top Parameters | Coefficient |
|---|---|

| Top Parameters | Coefficient |
|---|---|
| temp | 0.528 |
| is_year_2019 | 0.230 |
| weathersit_rainy | -0.188 |
| windspeed | -0.181 |
| hum | -0.159 |
| season_winter | 0.100 |
| mnth_Sep | 0.082 |
| holiday | -0.059 |
| weathersit_sunny | 0.058 |
| season_spring | -0.055 |
| mnth_Jul | -0.055 |
| weekday_Sun | 0.053 |
| season_summer | 0.053 |
| workingday | 0.043 |

The top 3 indicators are:

- temperature:
  - Positively correlated: higher the temperature, higher the rentals
  - Impact: coefficient is 0.528 hence compared with other variables, this is the dominating feature
- year
  - Positively correlated: demand is increasing year by year
  - Impact: coefficient is 0.230
- weathersit_rainy: (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)
  - negatively correlated: if the weather is rainy, then bike rentals will be less
  - Impact: coefficient is -0.188

If we do not want to consider year, as we can’t have the same year twice, windspeed is the next variable with highest coefficient

- negatively correlated: if the windspeed is high, then bike rentals will be less

- Impact: coefficient is -0.181

---

# General Subjective Questions

**Q1: Explain the linear regression algorithm in detail.**

Linear Regression is a Supervised Learning Algorithm in which the target variable is assummed to be dependent on the Independent Variable(s) linearly. The model is built such that the target variable (y) can be explained using the equation of a straight line.

Straight Line Equation: $y = mx + c$; where

- `m` is the slope: for one unit change in x, how much will y change
- `c` is the intercept: value of y, when x is 0

## Assumptions

For linear regression, some assumptions have to be valiadated on the dataset,

1. The Independent and Dependent Variables are linearly related
2. Assumptions on Error Terms
   - Error terms are normally distributed
   - Error terms are independent of each other
   - Error terms have constant variance (homoscedasticity)

Note: There are no assumptions on distributions on X and Y

## There are two types of Linear Regresssion Models

### Simple Linear Regresssion

- only one independent variable
- $y = \beta_0 + \beta_1 X_1$; where $\beta_0$ is the intercept and $\beta_1$ is the slope

- Interpretation: How much does y change, on one unit change in X

### Multiple Linear Regression

- more than one independent variable
- $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$
- Interpretation: How much does y change, per unit change in one of the independent variables when all other predictors are held constant

#### Additional Considerations:

- `Overfitting` : Increasing the number of variables may end up having a problem where the model learns the training dataset too well. If that happens, it will to generalize well on the test data and hence the real data
- `Multicollinearity` : The independent variables may have a relationship amongst themselves. This will cause problems with interpeting the model parameters.

### Idea

The idea here to find the best fit line, which can represent the variability in the data to a certain degree and is able to reasonably predict the target varaible on unseen data.

## Best Fit Line

To find the best fit line, we try to minimize the Residual sum of the squares (RSS). This process is formally known as `Ordinary Least Squres` Method. Residuals are defined as the difference between the y-coordinates of the actual data and the y-coordinates of the predicted data i.e. $e_i =$ Measured Value - Predicted Value $= y_i - y_i pred$.

## Analyzing Model

### Residual Sum of Squares

RSS: $\sum_{i=1}^{N}(e_i)^2 = \sum_{i=1}^{N}(y_i - y_i pred)^2$.

### Total Sum of Squares (TSS):

It is calculated by subtracting $y_{actual} - y_{mean}$ value for each of the data points and taking a sum of it. Formula: $\sum_{i=1}^{N}(y_i - \bar{y})^2$

### $R^2$

Since RSS is not a relative term with units of $y^2$, it is difficult to compare models that have different units, hence, we use:

$R^2 = \frac{variability\ in\ Y\ explained\ by\ model}{Total\ variability\ in\ Y} = \frac{Explained\ Sum\ of\ Squares}{Total\ Sum\ of\ Squares} = \frac{ESS}{TSS} = \frac{TSS-RSS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{Residual\ Sum\ of\ Squares}{Total\ Sum\ of\ Squares}$.

### Adjusted $R^2$

In general, Higher the $R^2$ better the model. However, since, adding more variables will either keep the $R^2$ same or increase it, we use another term which is Adjusted $R^2$ which penalizes the model based on the number of parameters that are present.
$Adj\ R^2 = 1 - \frac{(1-R^2)(N-1)}{N-p-1}$; where n = sample size, p = number of predictors

## Hypothesis Testing in Linear Regression

To test the significance of $\beta_1$, we can perform hypothesis testing. Here, the null hypothesis ($H_0$) will be that $\beta_1$ is 0
$H_0 : \beta_1 = 0$
$H_a : \beta_1 \neq 0$

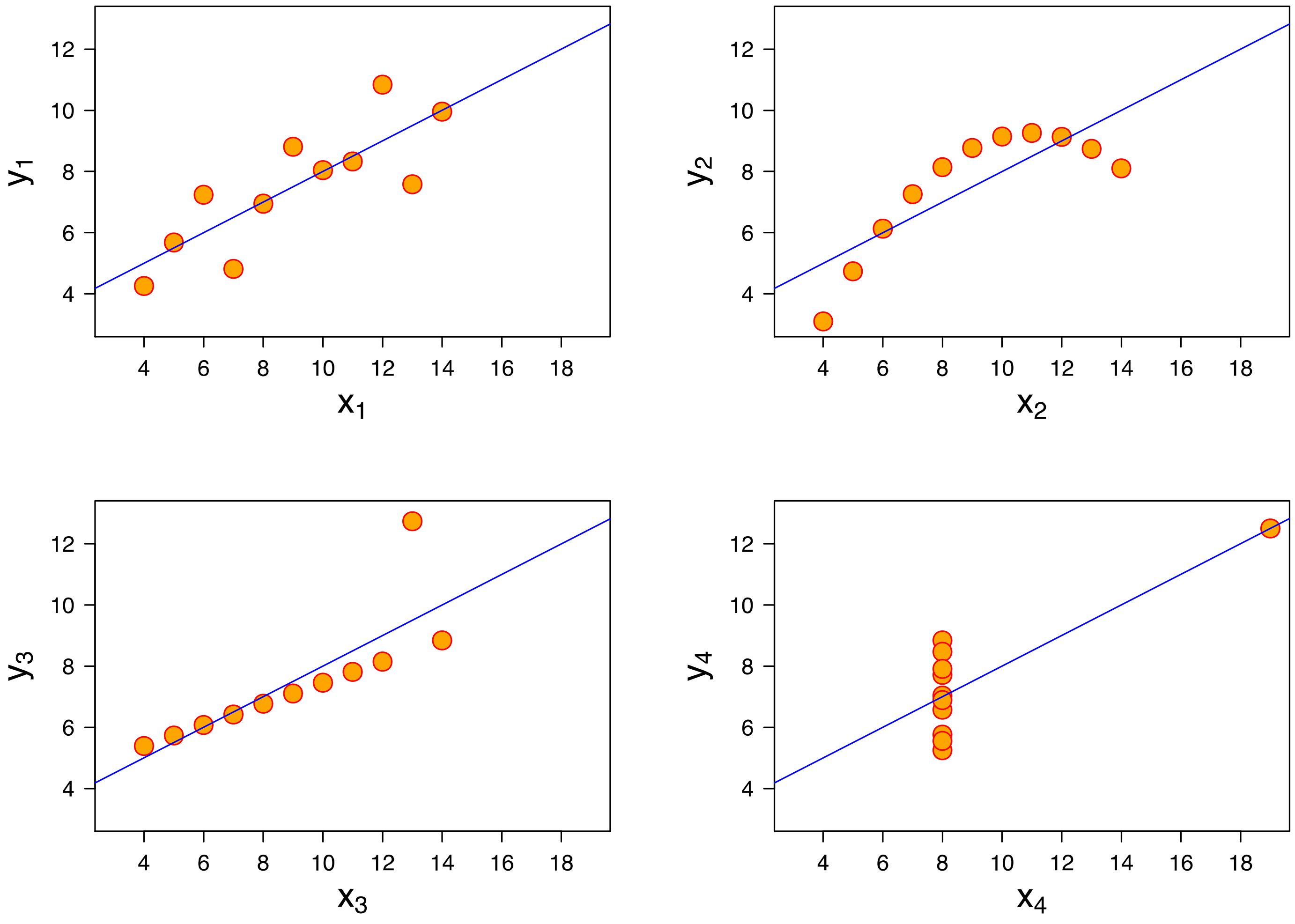
- If we fail to reject the Null hypothesis, $\beta_1$ and thus, the independent variable it is associated with is insignificant in the prediction of the dependent variable.
- The t-statistic along with the t-distribution table is used to determine the p-value of the coefficient.
- t-score for $\beta_1$ is given as $\frac{\beta_1}{Standard\ Error(\beta_1)}$

---

#### Q2: Explain the Anscombe’s quartet in detail.

Anscombe’s Qurtet is a combination of four datasets such that each of the datasets has nearly identicial descriptive statistics (such as variance and mean), and yet the data for all the four set is distributed quite differently. This difference is clearly visible, when the four datasets are plotted.

It was constructed by a statistician by the name of `Francis anscombe` in 1973 with the main purpose of illustrating the importance of visualizing data. It also helps us understand the effect of outliers in the dataset.



- Source: By Anscombe.svg: Schutz(label using subscripts): Avenue - Anscombe.svg, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=9838454

In the above image we can see that data is distributed quite differently in all four datasets.

- First one is a linear relation
- the second one is parabolic
- the third is linear, but it fits much better on a straight line than first one overall
- in the fourth one, most data points have the same x value, and one value which is far away

The striking thing is that all these datasets have the same variance, mean and standard deviation. This clearly demonstrates what effect outliers can have on the data and how if we don't visualize our data, we might infer incorrect information.

### Q3: What is Pearson's R?

Pearson's R is the correlation coefficient between any two variables i.e. it gives us information about bivariate linear correlation. It shows us how two variables are related to each other. In other words, what is the impact of changing one variable over the other other one. It is the ratio of covariance of two variables and the product of their standard deviation.

The value for the correlation coefficient can range from -1 to 1.

- Positive Correlation: Increasing x will increase y
- Negative Correlation: Increasing x will decrease y

Correlation can thus be of three types:

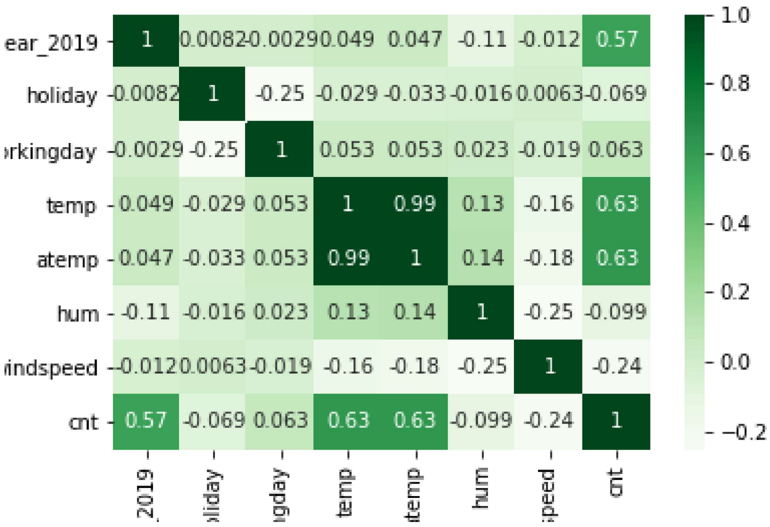
- Highly Correlated:
  - The absolute value of correlation coefficient is closer to 1
  - Changing one variable, changes the other one significantly
- Weakly Correlated:
  - The absolute value of correlation coefficient is closer to 0
  - Changing one variable does not change the other one significantly
- Zero Correlation:
  - The absolute value of correlation coefficient is 0
  - Changing one variable does not change the other one at all

While using Python, one can find out the correlation of all numeric variables with one another using

```
df.corr()
```

One can also plot the correlation using a heatmap for understanding the numbers visually using

```
sns.heatmap(df.corr())
```



The linear regression model's $R^2$ is actually the square of Pearson's $R$ itself, when the dataset has been normalized.

---

**Q4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

Scaling is the process of normalizing the range of independent variables or features of a dataset so that all variables can be compared with each other without variables that have a larger base value, don't end up having the smallest coefficients. If scaling is not performed, interpreting the coefficients becomes impractical.

Use of Scaling:

1. Ease of interpretation
2. Faster convergence for gradient descent methods

Example:

| Age | Salary |
|---|---|
| 20 | 10000 |
| 25 | 25000 |

If we don't normalize the data, the model may assume the columns with larger values have larger impact or in other words it will take the magnitude in account and not the units hence the model might be incorrect.

**Effect of scaling**

| Statistic | Change |
|---|---|
| p-values | No |
| Model Accuracy | No |
| F-statistic | No |
| R-squared | No |
| Coefficients | Yes |

## MinMaxScaling

- also called Normalization
- get the values between 0 and 1

- $X_{changed} = \frac{X - X_{min}}{X_{max} - X_{min}}$
- Generally Min-Max Scaling is preferred when there are outliers in the data

## Standardization

- also called Z-score Normalization
- subtracting the mean and dividing by standard deviation such that the variable is centered at zero and the standard deviation is 1
- Formula: $X_{changed} = \frac{X - \mu}{\sigma}$

---

### Q5: You might have observed that sometimes the value of VIF is infinite. Why does this happen?

VIF or Variance Inflation Factor indicates how well an independent variable can be explained with the help of all the other indepdent variables.

Formula: $VIF_i = \frac{1}{1 - R_i^2}$ where ‘i’ refers to the i-th variable which is being represented as a linear combination of rest of the independent variables.

To get infinity as the output of VIF, the denominator must become zero

$$1 - R_i^2 = 0 \implies R_i^2 = 1$$

Now, $R_i^2$ is the $R^2$ value for variable i with respect to all other variables in the dataset.

- We know that $R^2$ is the measure of how well the model fits.
- To get $R^2 = 1$ the model must have a perfect correlation.

In other words, we can say that $R_i$ can be explained perfectly well by all other variables in the dataset, which is a case of multicollinearity.

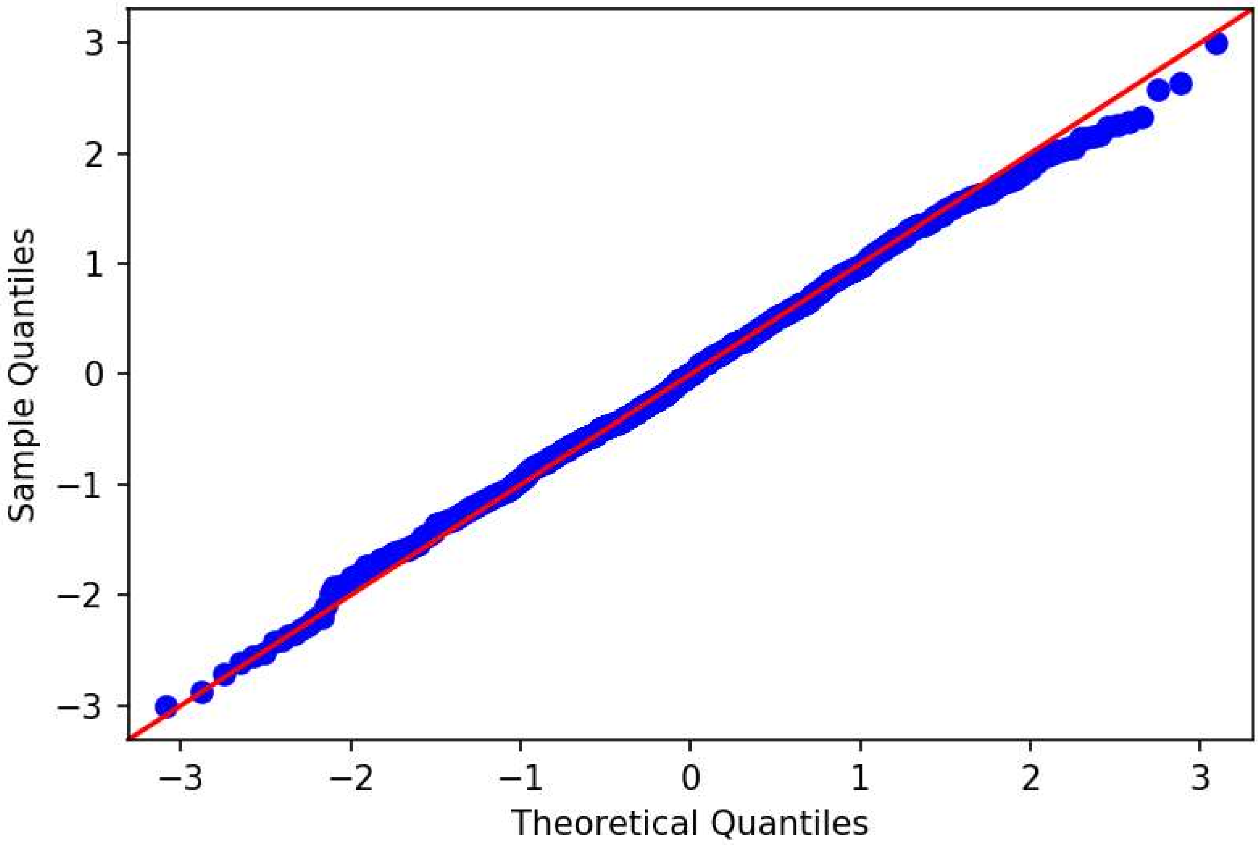
If we have a variable which has infinite VIF, we should drop the variable which is causing the multicollinearity.

---

### Q6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Q-Q plots are a quick visual way to asses whether a distribution is similar to a theoretical distribution such as Normal, Exponential or Uniform distribution. It can summarize for us whether the distributions of two variables are similar or not with respect to the locations.
These can be plotted with the help of statsmodel easily.

```
import statsmodel.api as sm
sm.qqplot(data_set, line='45')
```



If the data falls nicely on the $45^\circ$ red line, we can say that the distribution is normal.

The Q-Q here means quantile-quantile, i.e., qq plot computes the quantiles of our dataset against the desired distribution.
Quantiles break our data into buckets which are proportioned equally.

For Linear Regression,

- it can help us identify if our training data and test data came from the same dataset or not
- since we can check against theoretical distribution such as Normal Distribution, we can plot residuals using q-q plot to check whether they are normally distributed to validate the assumption of linear regression

---

Author’s Note: This assignment/document was original written in markdown with the occasional of LaTeX to integrate mathematical equations.