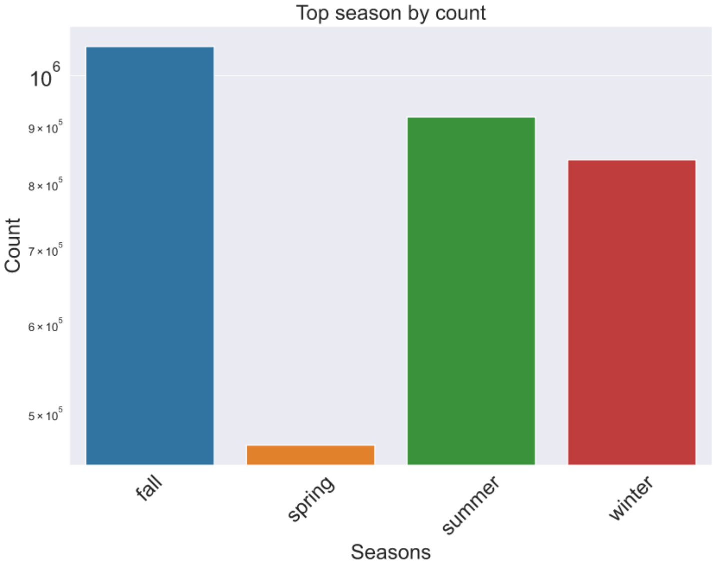
Assignment-based Subjective Questions

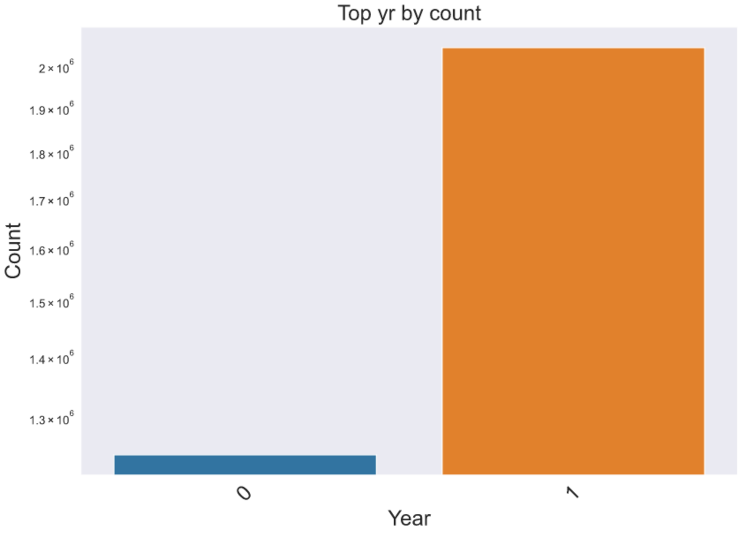
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

a) Seasons:



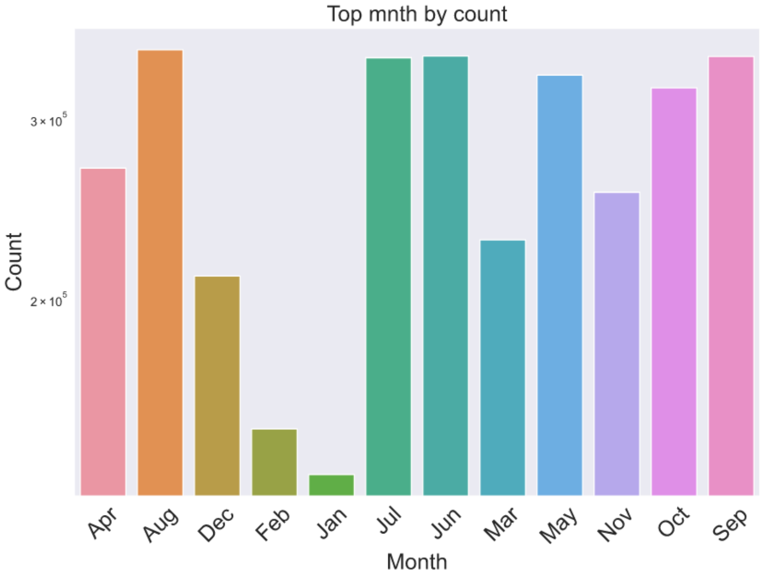
As seen from the plot, most of the Bike bookings take place during the Fall and Summer.

b) Year:



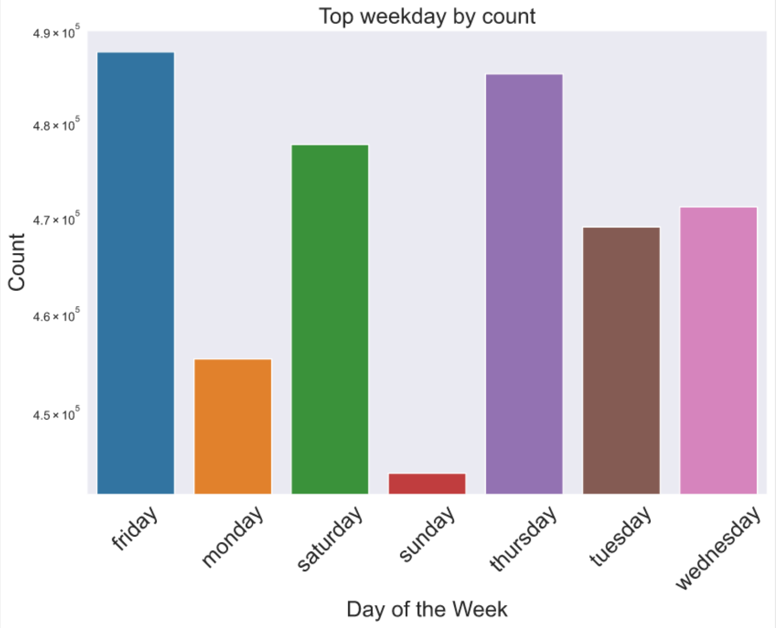
As the Year is increasing the bike demand also increases.

c) Month:



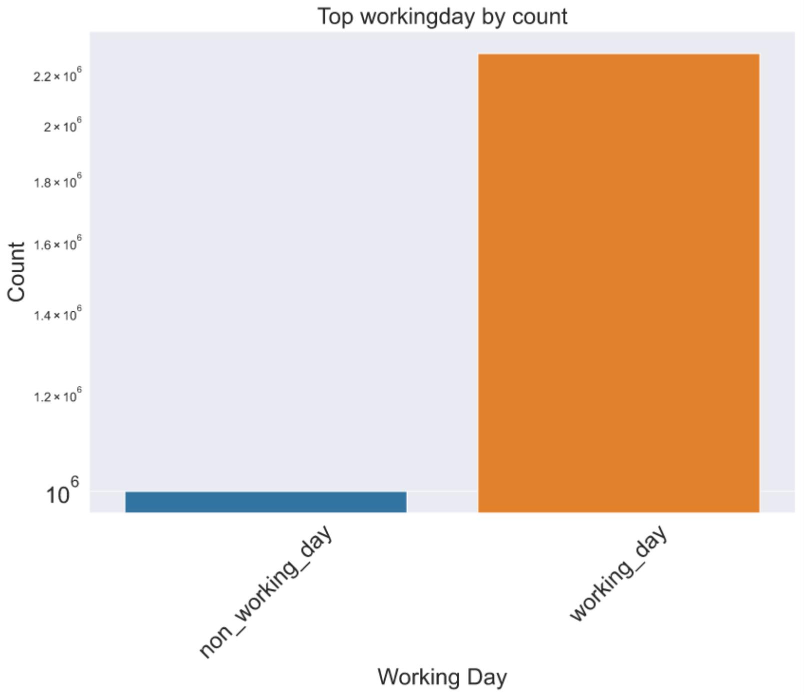
As seen previously the Fall months September - December have the highest bookings.

d) Weekday:



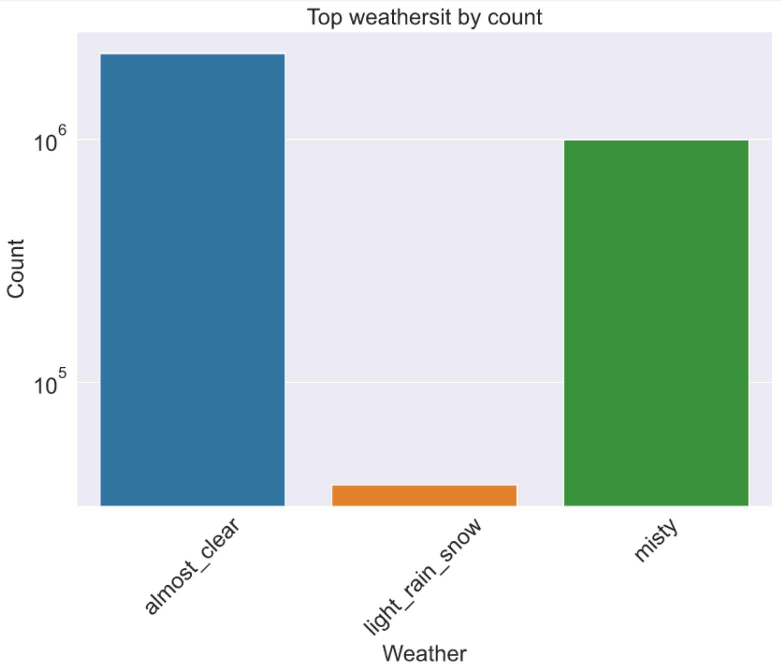
Fridays see the most bookings followed by Thursdays and Sundays being the least booked day.

e) Working day:



The most bookings occur only on Working Days and not on holidays or Weekends.

f) Weather:



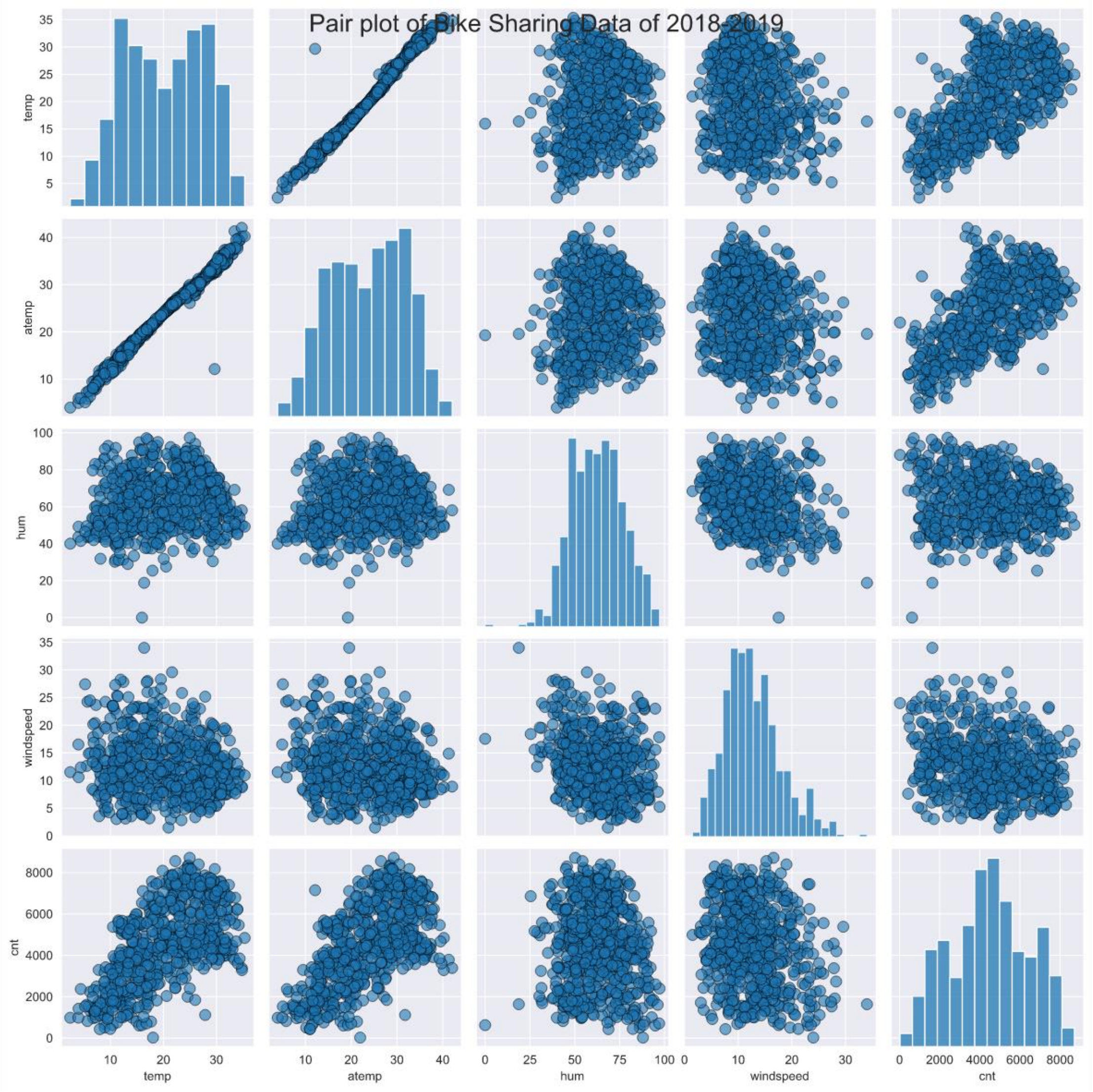
People tend to book bikes on Clear Days rather than on Rainy/ Snow fall days.

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

If we keep all dummy variables as is, we risk the possibility of multicollinearity between the dummy variables and keeping all categories is redundant. This would take a toll on the model performance.

So, if we have three categories (Bangalore, Mumbai, Delhi) for a feature. Then after performing pd.get_dummies on the feature and on dropping Bangalore we would get two columns for Mumbai and Delhi. So, now if Mumbai and Delhi are 0, it would simply mean the value is Bangalore.

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



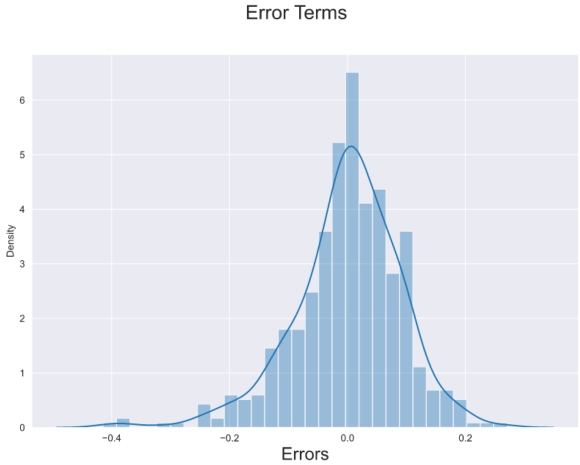
Looking at the pairplot we can easily find that the column temp/ atemp has the highest correlation (atemp a little more correlated) with the target variable ‘cnt’.

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

The Assumptions of linear regression are as follows:

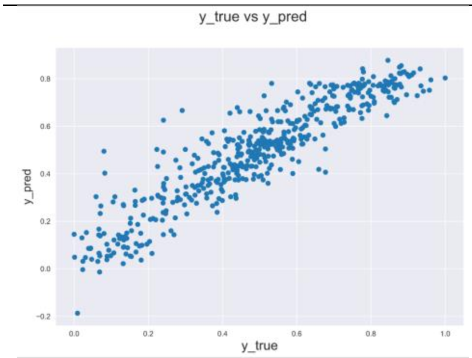
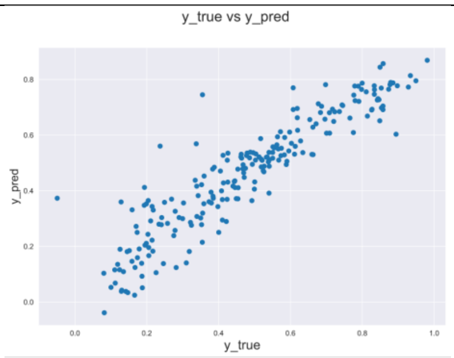
a) Linear relationship between X and Y – From the pairplot we saw that some of the features have a linear relationship with the Target.

b) Error terms are normally distributed (not X, Y) –



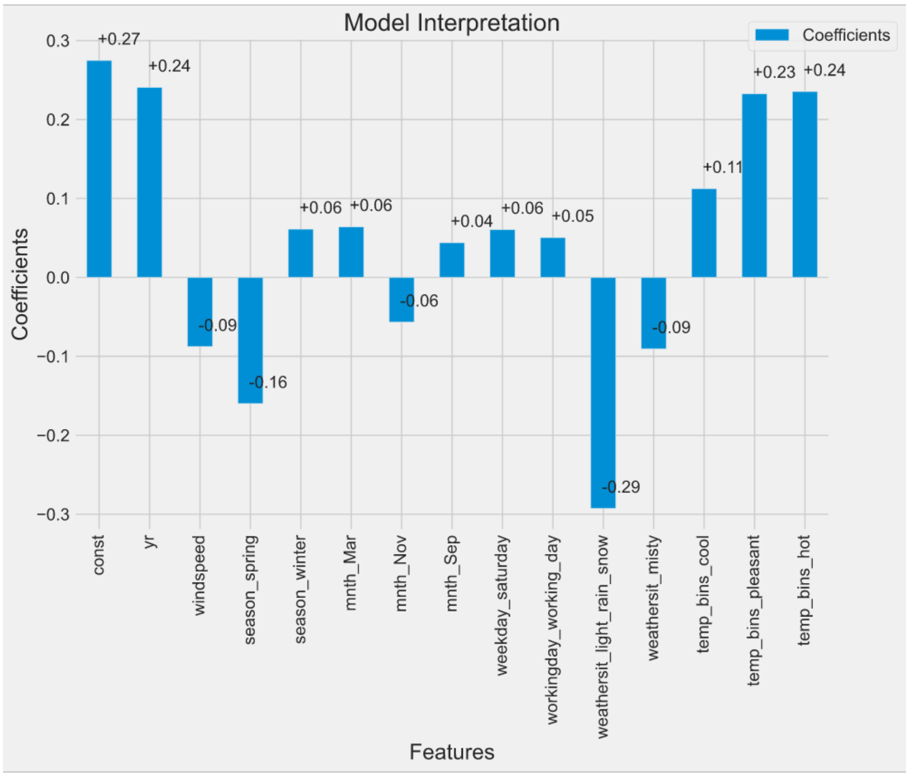
After building the model on training set, I plotted the distribution of error terms and as it is visible the error terms are centered around 0 and are normally distributed which is evident from the bell-shaped curve.

c) Error terms are independent of each other and Error terms have constant variance (homoscedasticity):

Training Data Error Terms Variance	Testing Data Error Terms Variance
	

The above plots help us ascertain the assumption of Homoscedasticity i.e. constant variance of error terms.

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



The above plot gives us an idea about the coefficient for each of the final list of predictors for our model.

Some of the features that have a greater positive impact on the target variable (Count of bookings) are **yr**, **temp_bins_hot**, **temp_bins_pleasant**.

Some of the features that have a negative impact on the target variable (Count of bookings)
are **weathersit_light_rain_snow**, **season_spring**, **windspeed**.

If we must translate the above model interpretation in a much more understandable form, we can easily
say that with the increase in year there will be an increasing demand for bikes. The demand will also be comparatively more
during days with pleasant temperatures than on colder days.

During days when there is light shower or snow, demand will be low and it would also be the case during the spring season,
as it is a well know vacation period at universities and schools in the US

General Subjective Questions

1. Explain the linear regression algorithm in detail.

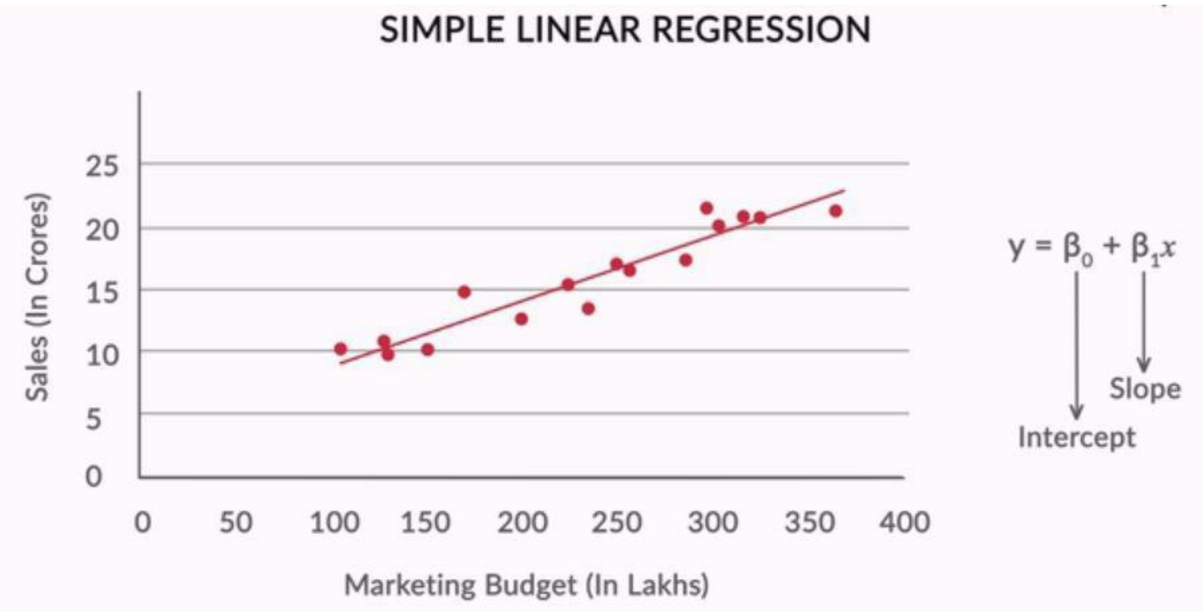
In simple terms, Linear Regression is a form of predictive modelling technique which tells us the relationship between the dependent (target variable) and independent variables (predictors).

It generally classified into two types:

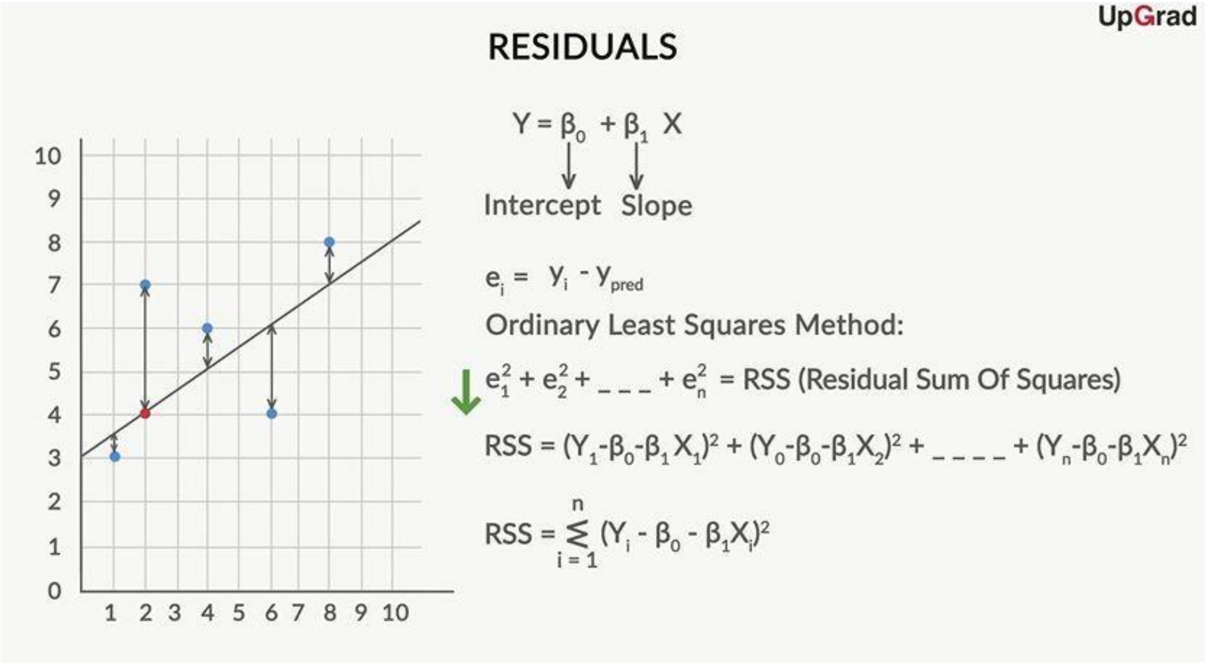
1. **Simple Linear Regression:** It Explains the relationship between a dependent variable and one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points and the straight line passing through these points is called the best fit line.

   The general equation of the line is (y) = B0+B1X

   Where B0 -> Intercept and B1 -> Slope.



The best fit line is derived by minimizing the RSS (Residual Sum of Squares) which is the cost function in this case. Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable:



RESIDUALS

$$Y = \beta_0 + \beta_1 X$$

Intercept Slope

$$e_i = y_i - y_{pred}$$

Ordinary Least Squares Method:

$$e_1^2 + e_2^2 + \_\_\_ + e_n^2 = \text{RSS (Residual Sum Of Squares)}$$

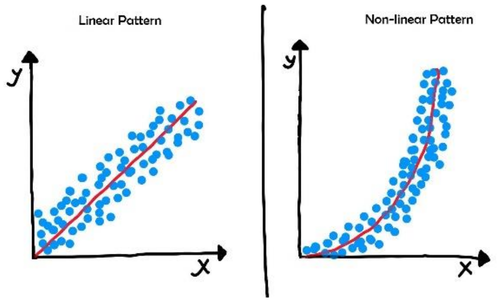
$$RSS = (Y_1-\beta_0-\beta_1X_1)^2 + (Y_0-\beta_0-\beta_1X_2)^2 + \_\_\_\_ + (Y_n-\beta_0-\beta_1X_n)^2$$

$$RSS = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

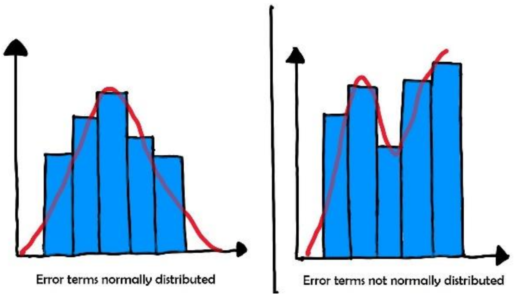
One important metric to judge the model fit is:
R-squared (Coefficient of Determination): The R-squared value tells the extent of the fit, i.e. how well the straight line describes the variance in the data. Its value ranges from 0 to 1, with the value 1 being the best fit and the value 0 showcasing the worst.

Assumptions of a simple linear regression model are:

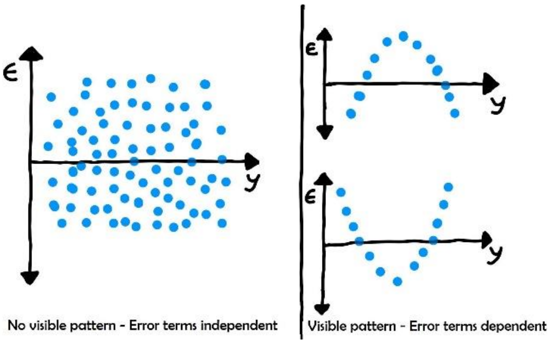
a. Linear relationship between X and Y



b. Error terms are normally distributed (not X, Y)



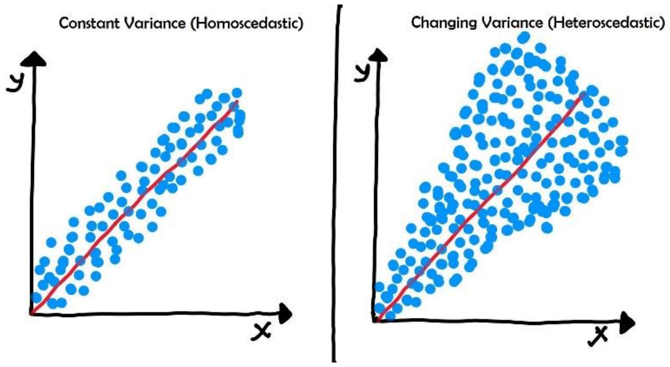
c. Error terms are independent of each other



No visible pattern - Error terms independent

Visible pattern - Error terms dependent

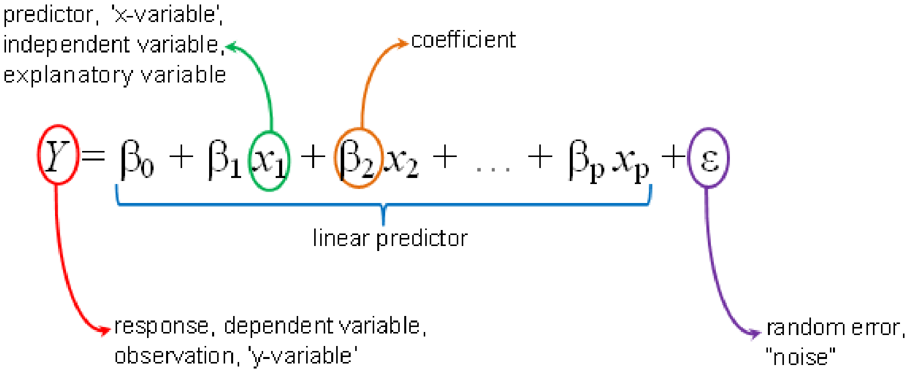
d. Error terms have constant variance (homoscedasticity)



2. **Multiple Linear Regression:** It explains the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X.

   Multiple linear regression is also like simple linear regression with the small change that instead of having beta for just one variable, we will now have betas for all the variables used.

   The general equation for a MLR is given below:


predictor, 'x-variable',
independent variable,
explanatory variable
coefficient
$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$
linear predictor
response, dependent variable,
observation, 'y-variable'
random error,
"noise"

Similarities to Simple Linear Regression:

a. The model now fits a 'hyperplane' instead of a line
b. Coefficients still obtained by minimizing the sum of squared error (Least squares criterion)
c. For inference, the assumptions of Zero mean, independent, normally distributed error terms that have a constant variance from Simple Linear Regression still hold.

Considerations:

a. **Overfitting**: When we add more and more variables, the model might end up memorizing all the data points in the training set. This will cause major problems with the generalization on unseen data.
b. Multicollinearity: A model that has been built using several independent variables, some of these variables might be interrelated, i.e. some of these variables might completely explain some other independent variable in the model due to which the presence of that variable in the model is redundant. This affects the model interpretation and coefficients of variables.

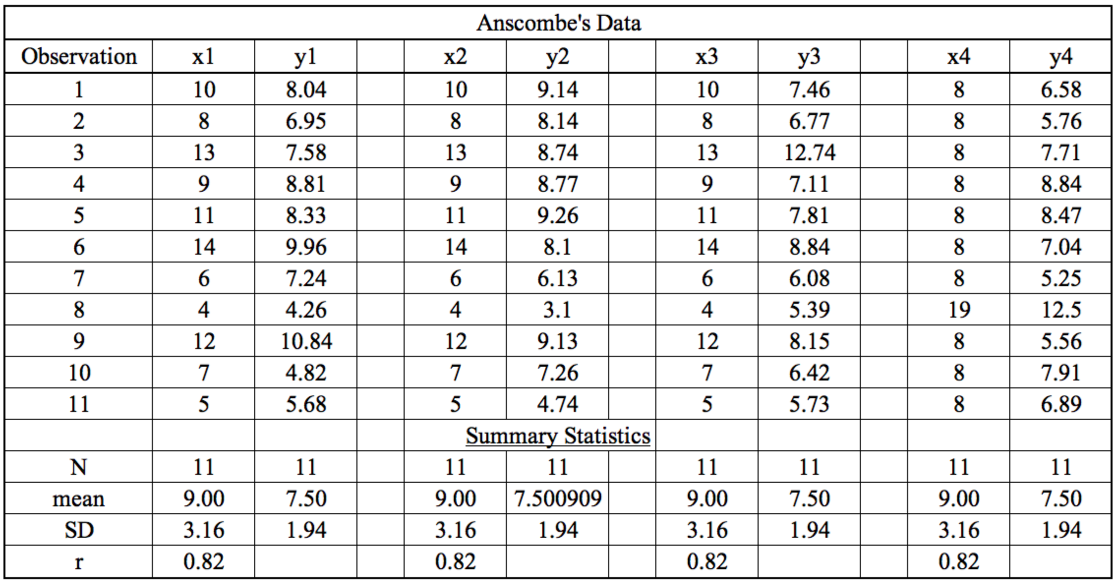
2. Explain the Anscombe's quartet in detail.

> *Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.*
> *— Wikipedia*

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties. There are these four data set plots that have nearly the same statistical observations, which provides the same statistical information that involves variance and mean of all x,y points in all four datasets.

The data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc.

Anscombe’s data with the Summary Statistics:

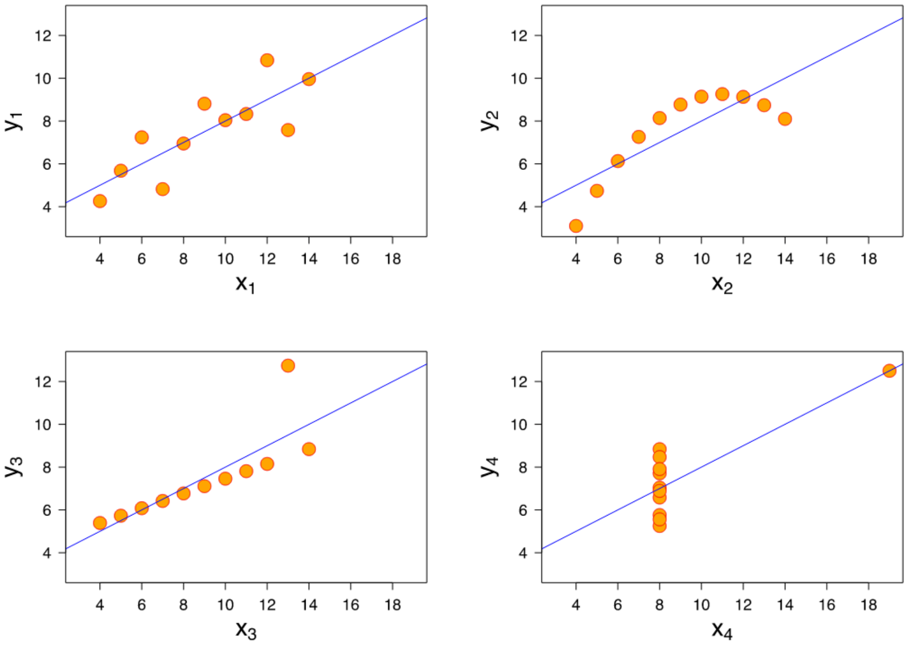


Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
					Summary Statistics						
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

On looking at the summary statistics we can see that:

- The average x value is 9 for each dataset
- The average y value is 7.50 for each dataset
- The correlation between x and y is 0.816 for each dataset
- The standard deviation for x is 31.6 and y is 1.94 for each dataset
- A linear regression (line of best fit) for each dataset follows the equation y = 0.5x + 3

But if we plot each of the 4 datasets along with the best fit line, it produces different stories.



Dataset -1 has points that somewhat have a linear relationship with some variance.
Dataset -2 does not follow a linear relationship
Dataset -3 has an almost perfect linear relationship between x and y.
Dataset -4 x is constant but for one data point.

This behavior was not visible in the summary statistics.
So, it is important to plot the data to visualize patterns before building a model for taking business decisions.

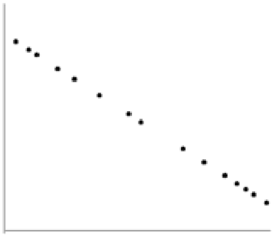
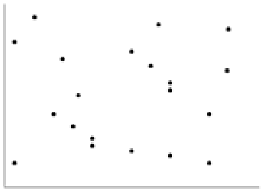
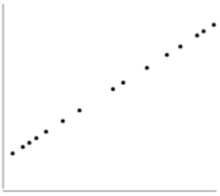
3. What is Pearson's R?

Pearson's correlation coefficient (r) is a technique to investigate the relationship i.e. the measure of the strength of association between two continuous variables.

We can observe the relationship between two continuous variables with a scatter plot to check for linearity. The closer the points on the plot are to a straight line, the stronger is the correlation between the variables.

The value for Pearson's Correlation Coefficient ranges between -1 to +1

Below are the examples of scatter points and the Correlation value.

r = -1 The points lie on a straight line but have a negative slope.	
r = 0 No Correlation or no relation between variables	
r = +1 The points lie on a straight line with a positive slope	

A positive correlation indicates that both variables increase or decrease together, whereas a negative correlation indicates that as one variable increases, so the other decreases, and vice versa.

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

When we have a lot of numerical features or independent variables while building a model, they might be on different scales. It will lead to a model with very weird coefficients that might be difficult to interpret.

We perform scaling mainly because of two reasons:

a. Ease of interpretation
b. Faster convergence for gradient descent methods

The two most popular techniques for scaling are:

1. **Normalized Scaling:** The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

2. **Standardized Scaling:** The variables are scaled in such a way that their mean is zero and the standard deviation is one.

$$x = \frac{x - mean(x)}{sd(x)}$$

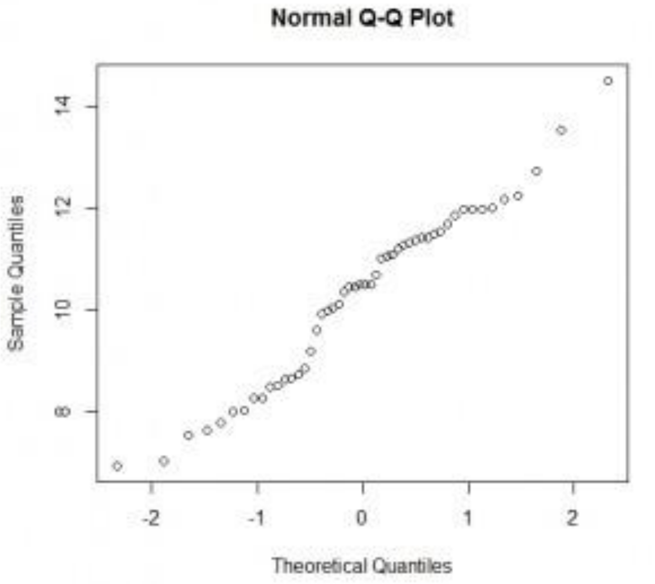
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

The Variance Inflation Factor (VIF) is a measure of colinearity among predictor variables within a multiple regression. It is calculated by taking the ratio of the variance of all a given model's betas divided by the variance of a single beta if it were fit alone.

The higher the value of VIF, the greater is the correlation with other variables.
Now, if the value of VIF approaches infinity, then it can be concluded that there is a perfect correlation among the predictor variables.
It means that the corresponding feature can be expressed exactly by a linear combination of other variables.

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

The Q-Q or Quantile – Quantile plots helps us assess if a dataset came from the same distribution or not. If both sets of quantiles came from the same distribution, we should see the points forming a roughly straight line.

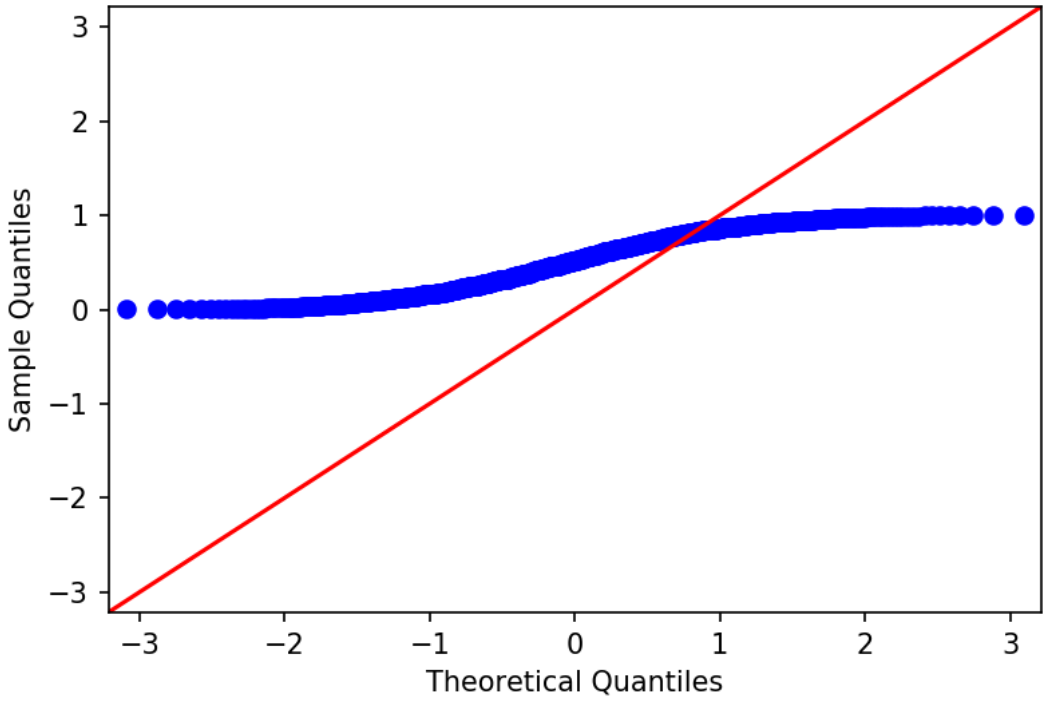


A Q–Q plot is generally a more powerful approach to do this than the common technique of comparing histograms of the two samples, but requires more skill to interpret.

The points plotted in a Q–Q plot are always non-decreasing when viewed from left to right. If the two distributions being compared are identical, the Q–Q plot follows the 45°-line y = x. If the two distributions agree after linearly transforming the values in one of the distributions, then the Q–Q plot follows some line, but not necessarily the line y = x.

Steps for creating a Q-Q Plot in python:

```
import random

# Generate some uniformly distributed random variables

random_uniform = [random.random() for i in range(1000)]

# Create QQ plot

sm.qqplot(np.array(random_uniform), line='45')

plt.show()
```


Sample Quantiles
Theoretical Quantiles