Assignment-based Subjective Questions

Question 1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Ans:

From the dataset provided 'day.csv' there are 7 categorical variables which are finally selected for initial model development out of which 2 are converted to dummy variables as follows:

1. yr
2. mnth
3. holiday
4. weekday
5. workingday
6. season: (Converted to dummy variables)
    ○ spring
    ○ summer
    ○ winter
    ○ fall(dropped)
7. weathersit: (Converted to dummy variables)
    ○ weather-t1(dropped)
    ○ weather-t2
    ○ weather-t3

Based on analysis from categorical variables we can infer following observations:

(1) yr:

As per observation from distribution with respect to target variable, 'yr' (year 2019=1, year 2018=0)category shows as follows:
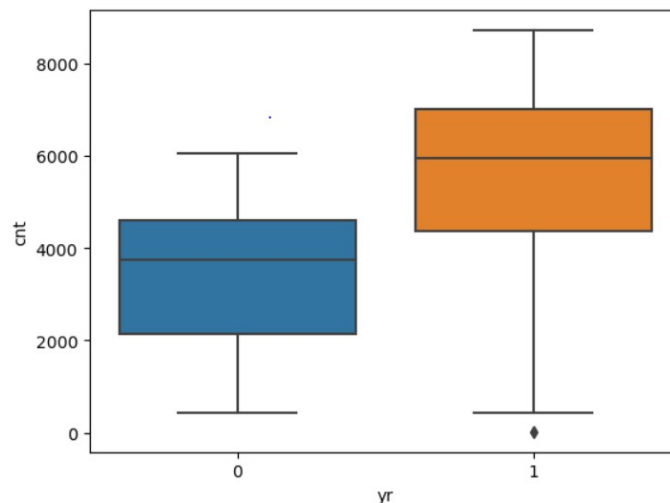
- The more number of BikeShaing bookings are happened in year 2019(1) compared with year 2018(0).
- The trend shows that BikeSharing bookings are increasing every year due to popularity and usage.

(2) mnth:

As per observation from distribution with respect to target variable, 'mnth' category shows as follows:


Bookings vs day of month
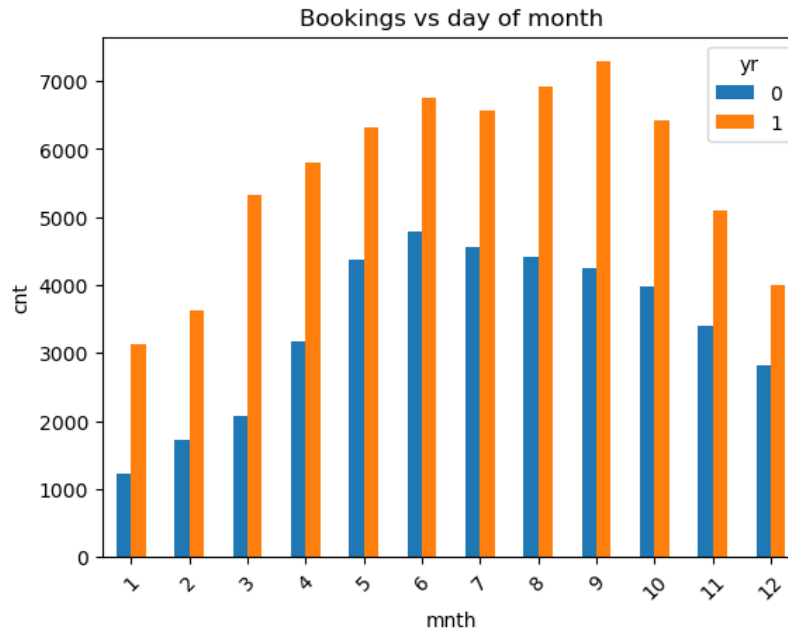
From the above figure we can infer following points with respect to dependent variable:
- More number of BikeSharing bookings happened in September-2019 month, June-2018 month. This might be due to **lock down** in 2019.
- And there is an increasing pattern of bookings from January to mid of year and decrease in bookings from mid to end of year. This might be due to people going for **vacation** mostly in months from October to March.
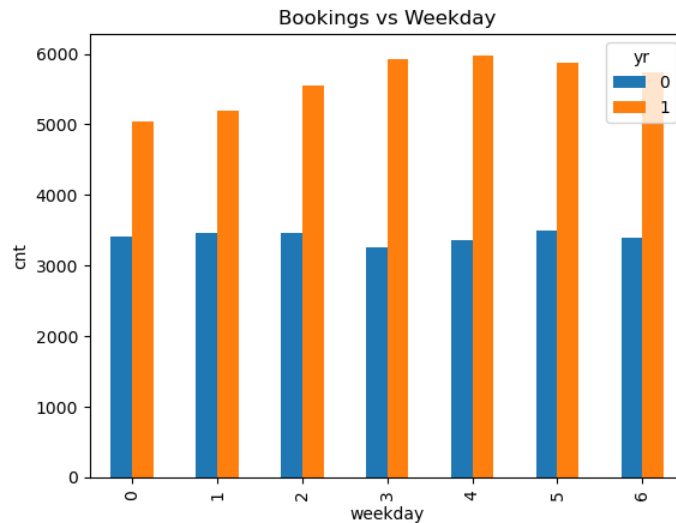
(3) holiday:

As per observation from distribution with respect to target variable, 'holiday' category shows as follows:

Bookings vs Holiday
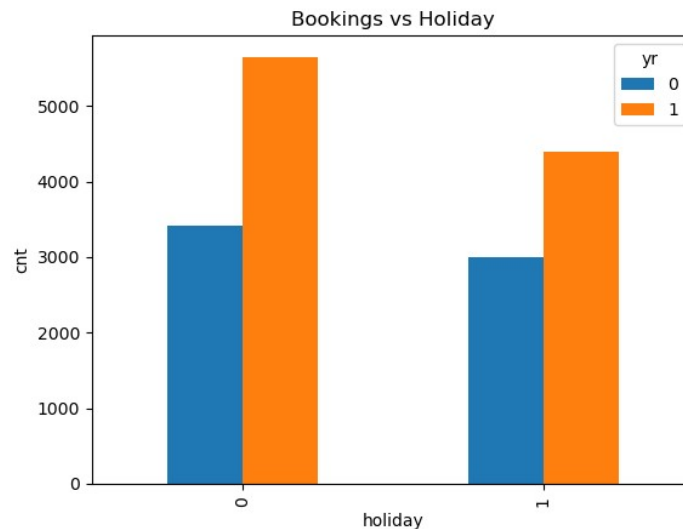
- From the above plot we can infer that many people are using shared bikes on non-holiday, might be using bike sharing for office commute.
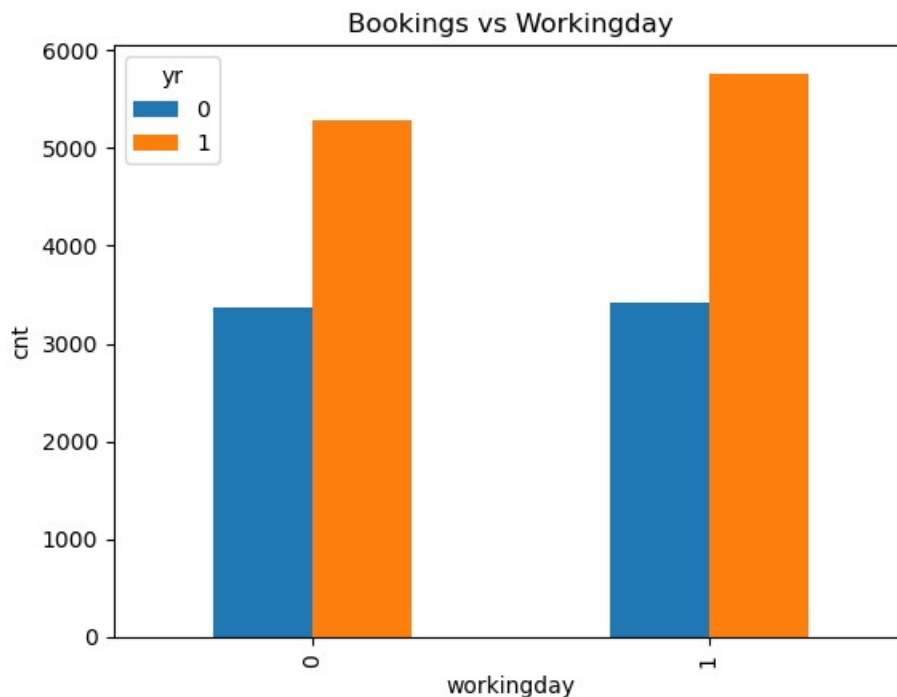
(4) weekday:

As per observation from distribution with respect to target variable, 'weekday' category shows as follows:

From the above plot we can observe that most bike sharing bookings are happened in mid of the week in 2019 and marginally less bookings in mid-week in 2018. this is because of hybrid work model in 2019.
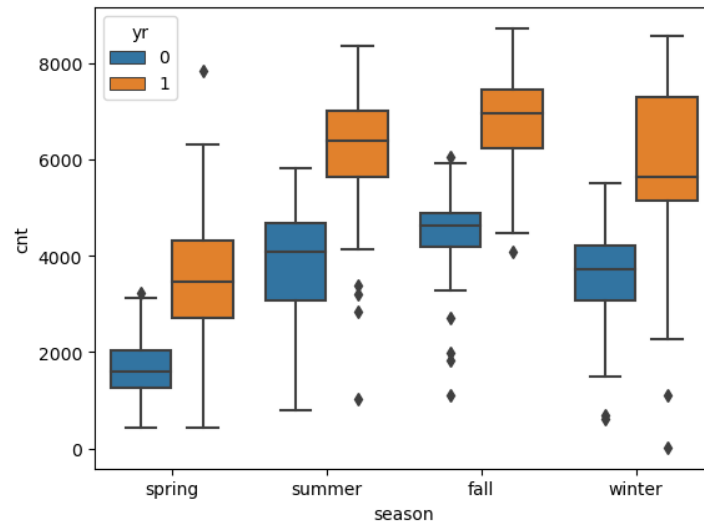
(5)workingday:

As per observation from distribution with respect to target variable, 'workingday' category shows as follows:

From the above plot we can observe that there is small increase in bookings on workingdays compared with non-working days.

(6) season:

As per observation from distribution with respect to target variable, 'season' category shows as follows:

- From the distribution we can observe that many bike bookings are happened during fall season.
- And during spring less number of bookings happened, in year 2018 its much lesser compared with year 2019.

(7) weathersit:

As per observation from distribution with respect to target variable, 'weathersit' category shows as follows:
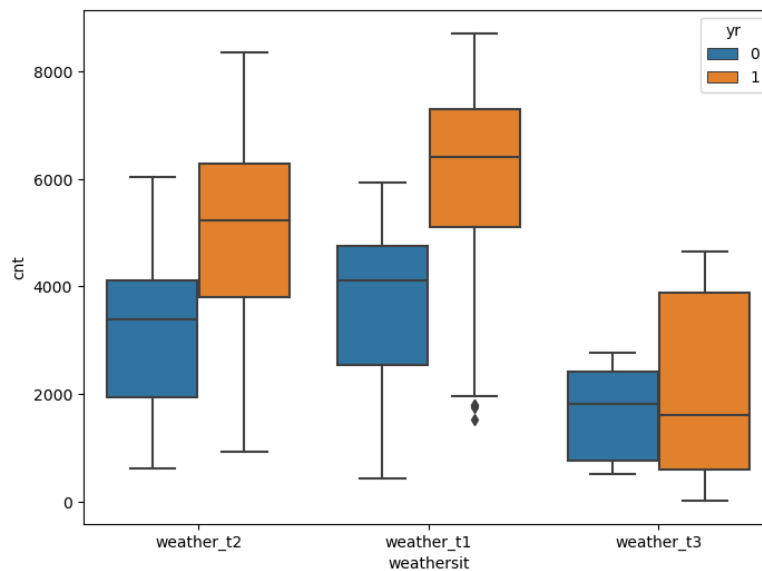
The variable weathersit has 4 types of weather conditions as follows.

weather_t1: Clear, Few clouds, Partly cloudy, Partly cloudy

weather_t2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

weather_t3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

weather_t4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

But from the data we have observed that there is no data available on weather type – 4

and from the observation from distribution:

- Less number of bookings happened where there is ab-normal weather conditions like weather_t3.
- Most bookings are happened when there is a clear weather.

Question 2: Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Answer:

In general dummy variable are created for independent, unordered relative categorical variables. So creating dummy variable for all categories will lead to multicollinearity between each individual categories.

Also if there are 'n' categories in variable 'n-1' categories will represent/explain/pridict n_{th} category.

For example in the case of categorical variable season:

Fall	Winter	Summer	Spring
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

In this case if season dummy values for all winter, summer, spring are '0' this represents season 'Fall'. So we don't need to explicitly define the dummy variable for Fall. As fallows.

0	0	0	→ fall
1	0	0	→ winter
0	1	0	→ summer
0	0	1	→ spring

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

Answer:

The numeric continuous variable **atemp** (feeling temperature)has the highest correlation with target variable.

From the pair plot both temp and atemp have almost same correlation but the correlation value for 'atemp' is **0.631059** more compared with 'temp'.

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

There are 3 steps to validate assumptions of linear regression after building model with training set.

1) **Linear relation ship between independent variables and predicted dependent variable**. By visualizing independent variables along with predicted dependent variable on training will be used for validating this assumption. For the BikeSharing assignment this assumption is validated using following figure.

Data points VS y-pred
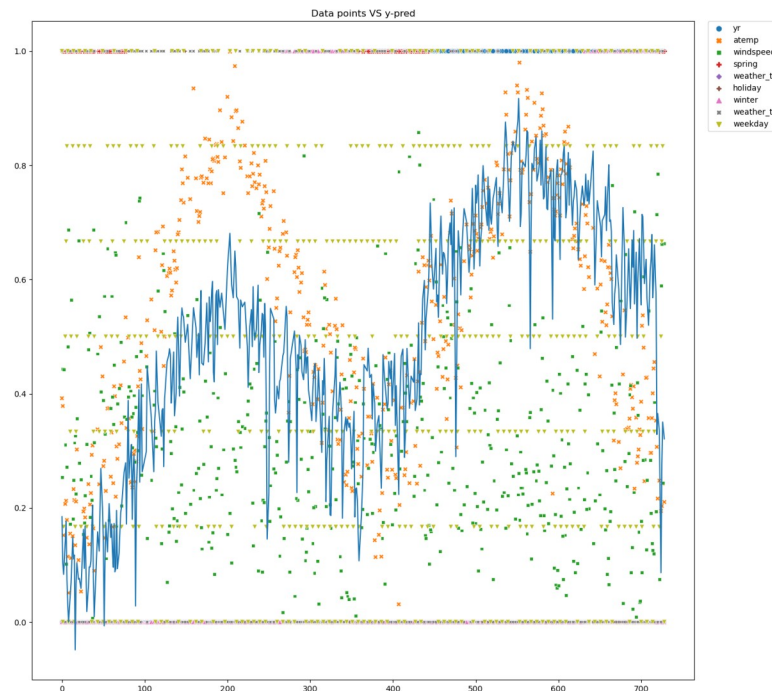
From the above figure we can observe that predicted dependent variable (blue line) and independent variables data points, there is a significant linear relation between independent and dependent target variable from train set.

2) **Residual/error terms distribution follow normal** and centered at 0 (mean=0). By visualizing distribution of error terms we have validated this assumption as shown in figure below.
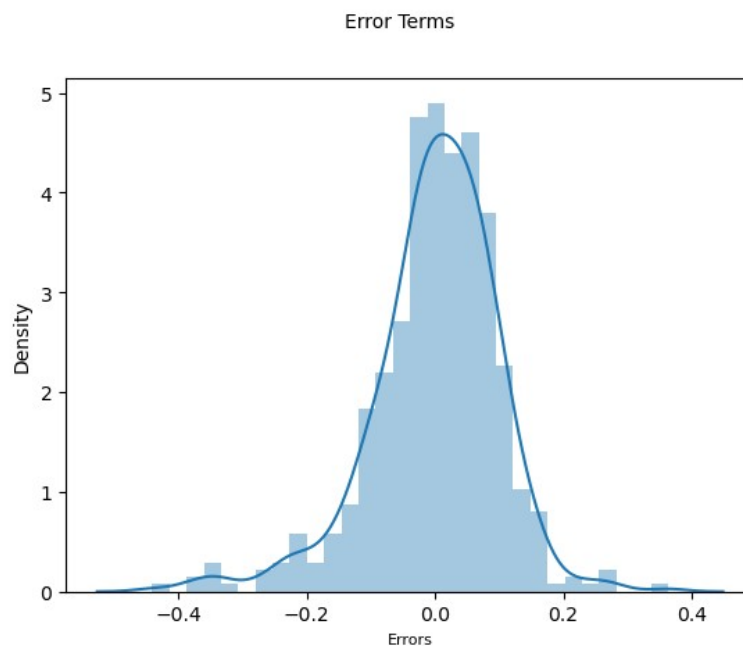
Error Terms

3) **The error terms should not be dependent on one another.** We can validate this by finding residuals on train plotting both predicted dependent variable and residuals as follows.
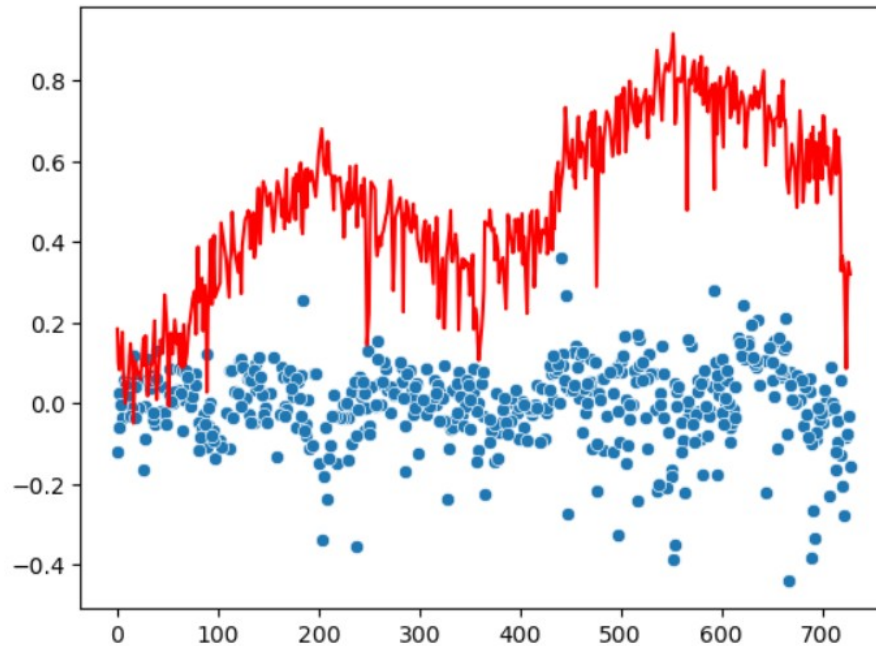
Correlation between residuals and y_train_pred: 0.0

From the above plot red=predictions on train set, blue= residuals , we can observe that there is no correlation between them. Also we can see the correlation value between then is '0'.

4) **Error terms have constant variance (homoscedasticity).**
To validate this assumption we can observe the residual scatter plot, if most of the residuals are ranging at -0.2 to 0.2. And when we compute the variance of residuals we got value 0.009 this is shows that all residuals are approximately same.

```
# Assumption 3: Error terms have constant variance:
print("Varience of error terms :", np.var((y_train - y_train_pred)))
```

Varience of error terms : 0.009907198510584716

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

Answer:

To know which of the feature are significantly contributing to predict dependent variable we need to observe regression coefficients. In our care the coefficient are as follows.

```
==================================================
                 coef    std err          t      P
--------------------------------------------------
const          0.1811      0.026      6.841      0
yr             0.2436      0.009     27.016      0
atemp          0.4664      0.031     14.985      0
windspeed     -0.1102      0.027     -4.068      0
spring        -0.1288      0.016     -7.883      0
weather_t3    -0.2627      0.031     -8.423      0
holiday       -0.0594      0.027     -2.163      0
winter         0.0479      0.013      3.639      0
weather_t2    -0.0775      0.009     -8.182      0
weekday        0.0549      0.013      4.113      0
==================================================
Omnibus:                     79.694    Durbin-Wat
Prob(Omnibus):                0.000    Jarque-Ber
Skew:                        -0.821    Prob(JB):
Kurtosis:                     5.449    Cond. No.
==================================================
```

As per linear regression equation we can write predicted dependent variable as follows:

y_pred = 0.1811 + (yr * 0.1811) + (atemp * 0.4664) – (windspeed * 0.1102) – (spring * 0.1288)
          - (weahter_t3 * 0.2627) - (holiday * 0.0594) + (winter * 0.0479)
          - (weather_t2 * 0.0775) + (weekday * 0.0549)

So from the above equation we can list top tree features which are significantly explains the demand of shared bikes are as follows.

1. **atemp** (positive demand
2. weather_t3(**Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds**)   (negative demand)
3. **yr** (positive demand)

Here yr is an independent catagorical variable so we can also consider next highest contributing feature **season-spring**(negative demand)

General Subjective Questions:

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

Linear regression is a supervised machine learning algorithm. This algorithm is built based on a simple line formula. This algorithm mainly used for predicting continuous values using one or many independent features or variables by building a statistical model between variables. This algorithm assumes linear relation between input independent and output dependent variables so it is named as linear regression.

Linear regression can be used when we have a dataset with continuous variables and trying to predict a continuous dependent variable using multiple independent variables.

A simple linear regression with one independent variable can be described as follows:

$$Y = \beta0 + \beta1*X + \varepsilon$$

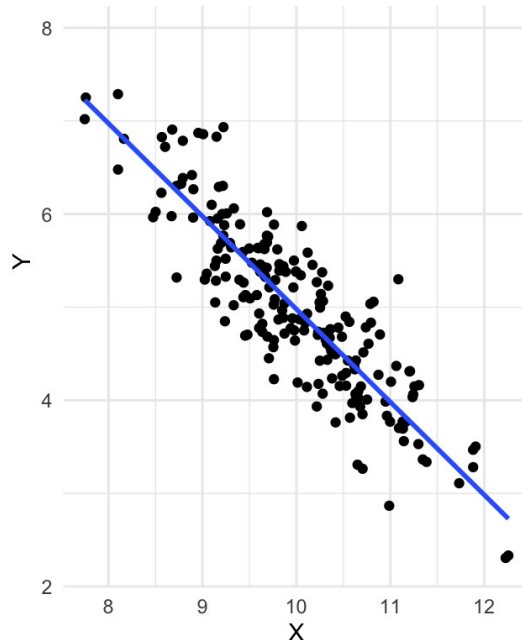Where Y= Predicted dependent variable

      X= input independent variable.

      $\beta1$ = slope of the regression line.

      $\beta0$ = intercept (base value of Y when X=0).

      ε = error term.

Following figure describes the simple linear regression model.

The main objective of the linear regression is to find the optimal fitting line that minimizes the squared error terms between actual expected observed values and predicted Y(Y_{pred}) values. This wll be done by minimizing **residual sum of squares** or the **mean square error**.

From the equation above we know that the X value is the input and available for predicting Y value, but to estimate Y value we need additional parameters $\beta 0$, $\beta 1$. These parameters can be estimated using Ordinary least squares ,method by minimizing the sum of squared errors.

Linear regression has some assumptions to be satisfied in order to achieve best fit model as follows.
- The residuals (actual Y – predicted Y) for all data points should be normally distributed with center 0, mean value(0)
- The independent variables have the linear correlation with dependent variable/predicted Y values.
- The variance of the each residual at each data point should be constant this is call as Homoscedasticity.
- Observations/dependent variable values are independent of each other.

The above shown equation is used when there is a one independent variable, but in realtime application there will be more number of independent variables used to estimate dependent variable. Process of predicting Y value with multiple independent variables using linear regression is also called as multiple linear regression. And the representation of dependent variable is as follows.

$$Y = \beta 0 + \beta 1*X1 + \beta 2*X2 + ... + \beta n*Xn + \varepsilon$$

Where $\beta 1$, $\beta 2$ ….. βn denotes regression coefficients for independent variables X1,X2…..Xn

In case of multiple linear regression model need to optimize n number of coefficients using ordinary least squares method. This will be more complex process to achive using normal matrix operations. So to simplify the process of optimization we use **gradient descent** method.
The functionality of the gradient descent method can be described using following equation.

$$W_i= W_{i-1} - \alpha \cdot \nabla J(W_{i-1})$$

Where W_i is the updated coefficient/parameter
$W{i-1}$ is the previous parameters
 α is learning rate
$\nabla J(W_{i-1})$ gradient of the objective function.

parameters will update in the opposite direction of the gradient to minimize the objective function

Once the parameters are optimized and model has been build the next step is to evaluate the model. We can interpret the coefficients (βn) to understand the relationship between the independent and dependent variables.

Following factors can be used to evaluate the model:
- R-squared : Denotes proportion of the variance in the dependent variable that is predictable from the independent variables.
- Mean squared error (MSE): Measures the average of the squares of the errors terms.
- Root mean squared error (RMSE): The square root of the MSE, which provides an interpretable measure of the average magnitude of the residuals.