MSA

Daffodil International University

Department of Software Engineering

Course code:- DS 331

Course Title:- Introduction to Data Science & Data Management

Final Exam; Summer-2021.

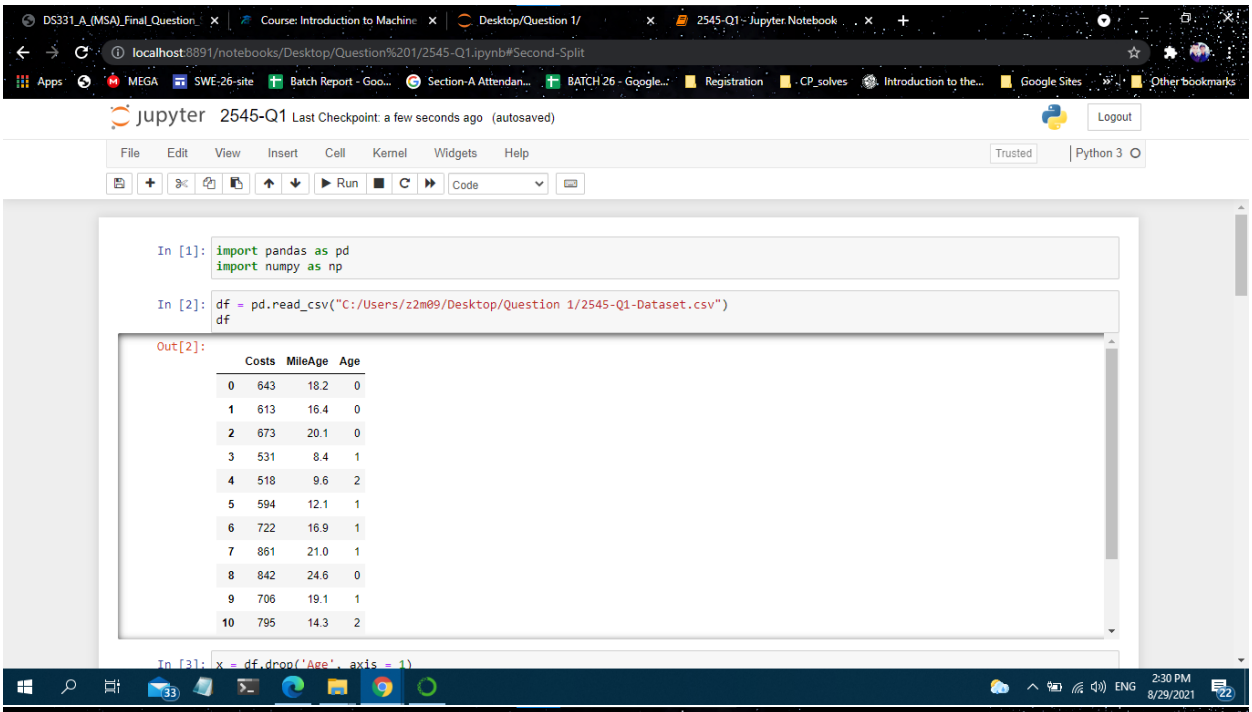
Student Id:- 182-35-2545

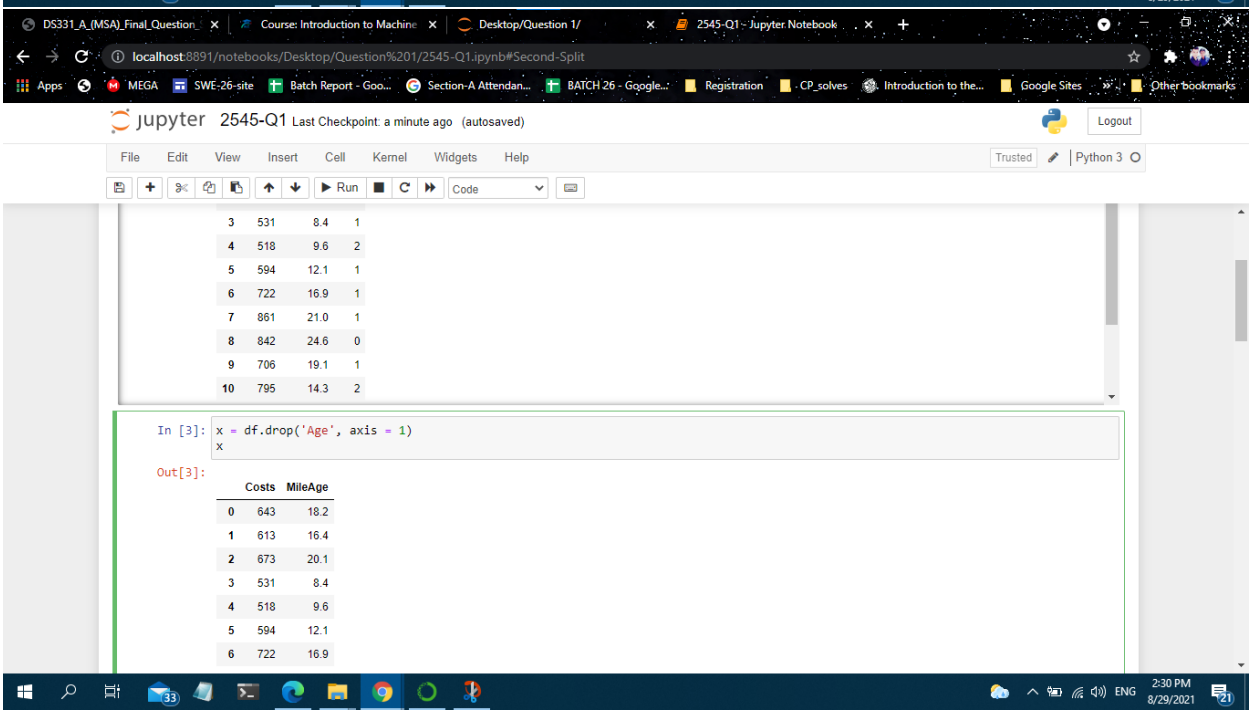
Section:- A

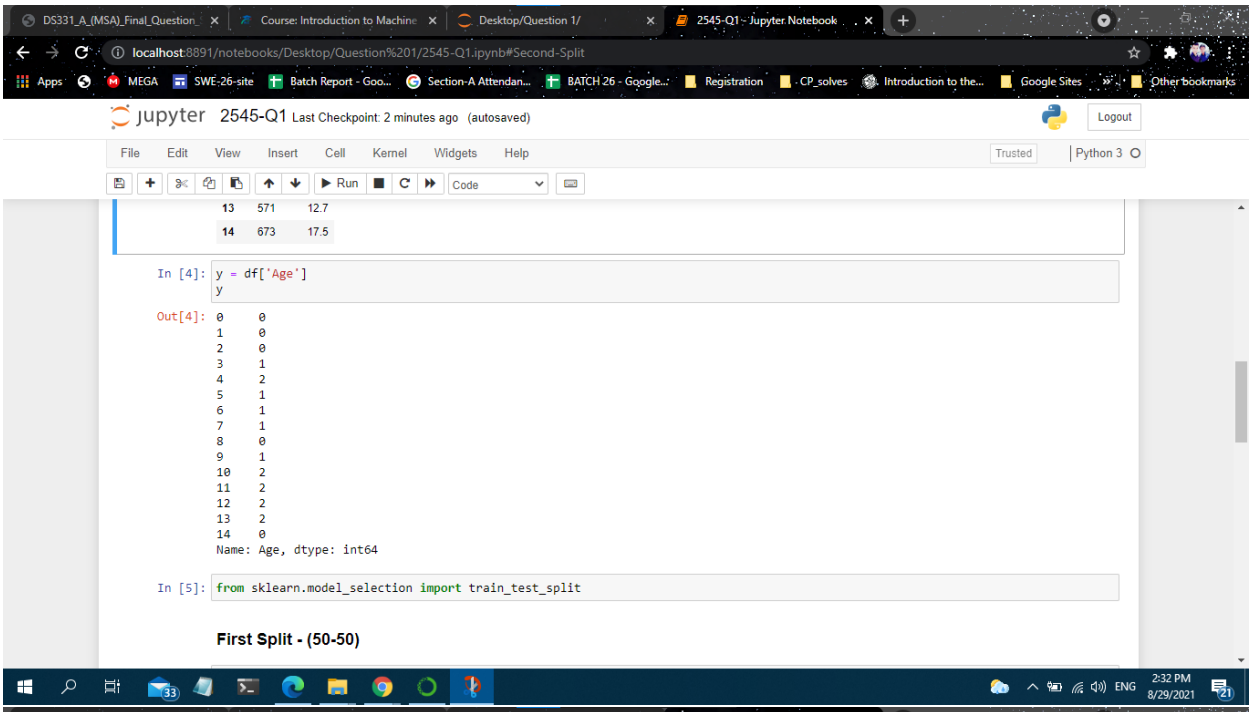
Campus:- MC

Batch:- 26

Date:- 29/08/2021.


DS331_A_(MSA)_Final_Question_
Course: Introduction to Machine
Desktop/Question 1/
2545-Q1 - Jupyter Notebook
localhost:8891/notebooks/Desktop/Question%201/2545-Q1.ipynb#Second-Split
jupyter 2545-Q1 Last Checkpoint: a few seconds ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
In [1]:
import pandas as pd
import numpy as np
In [2]:
df = pd.read_csv("C:/Users/z2m09/Desktop/Question 1/2545-Q1-Dataset.csv")
df
Out[2]:
Costs MileAge Age
0 643 18.2 0
1 613 16.4 0
2 673 20.1 0
3 531 8.4 1
4 518 9.6 2
5 594 12.1 1
6 722 16.9 1
7 861 21.0 1
8 842 24.6 0
9 706 19.1 1
10 795 14.3 2
In [3]:
x = df.drop('Age', axis = 1)
2:30 PM
8/29/2021


DS331_A_(MSA)_Final_Question_
Course: Introduction to Machine
Desktop/Question 1/
2545-Q1 - Jupyter Notebook
localhost:8891/notebooks/Desktop/Question%201/2545-Q1.ipynb#Second-Split
jupyter 2545-Q1 Last Checkpoint: a minute ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
3 531 8.4 1
4 518 9.6 2
5 594 12.1 1
6 722 16.9 1
7 861 21.0 1
8 842 24.6 0
9 706 19.1 1
10 795 14.3 2
In [3]:
x = df.drop('Age', axis = 1)
x
Out[3]:
Costs MileAge
0 643 18.2
1 613 16.4
2 673 20.1
3 531 8.4
4 518 9.6
5 594 12.1
6 722 16.9
2:30 PM
8/29/2021



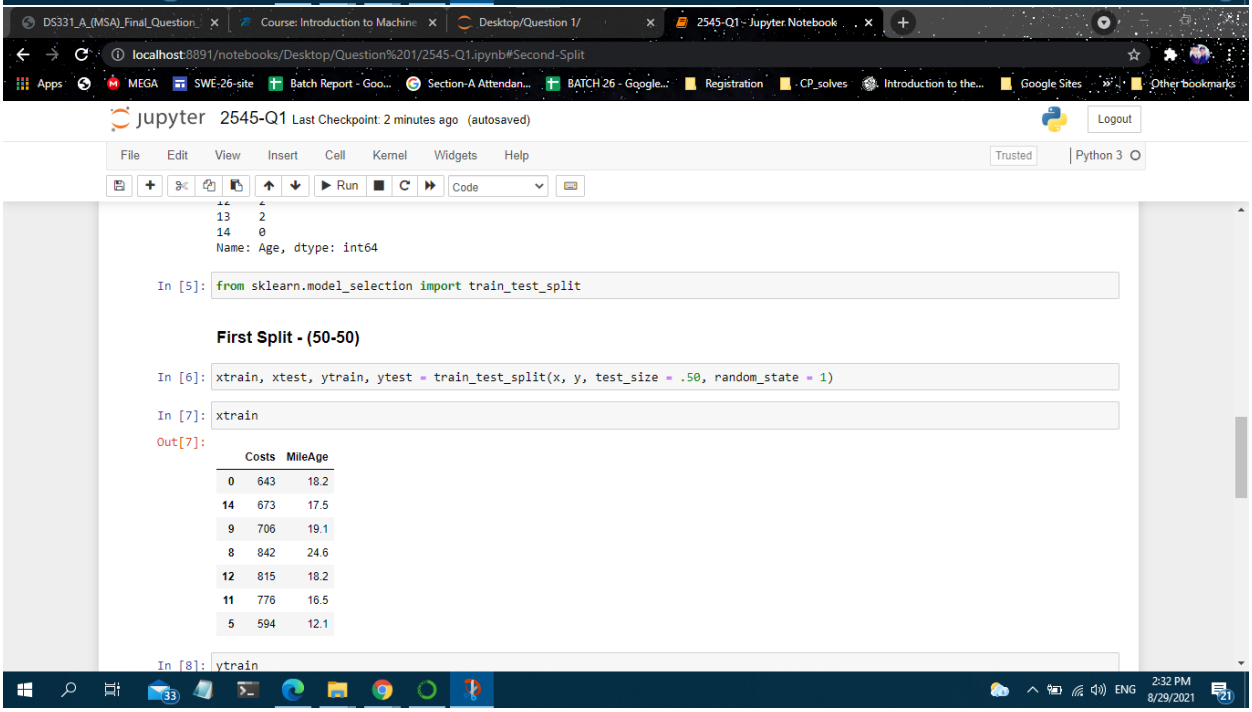
13	571	12.7
14	673	17.5

In [4]:
```
y = df['Age']
y
```

Out[4]:
```
0     0
1     0
2     0
3     1
4     2
5     1
6     1
7     1
8     0
9     1
10    2
11    2
12    2
13    2
14    0
Name: Age, dtype: int64
```

In [5]:
```
from sklearn.model_selection import train_test_split
```

First Split - (50-50)



13	2	
14	0	

```
Name: Age, dtype: int64
```

In [5]:
```
from sklearn.model_selection import train_test_split
```

First Split - (50-50)

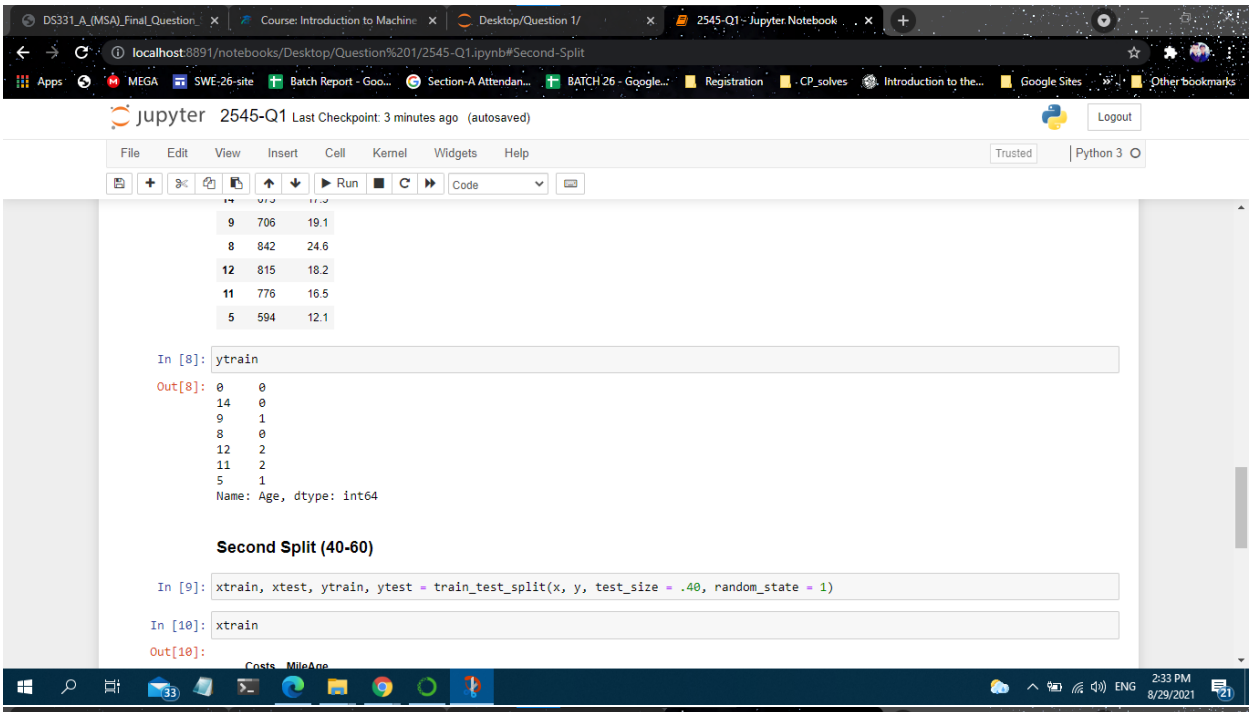
In [6]:
```
xtrain, xtest, ytrain, ytest = train_test_split(x, y, test_size = .50, random_state = 1)
```

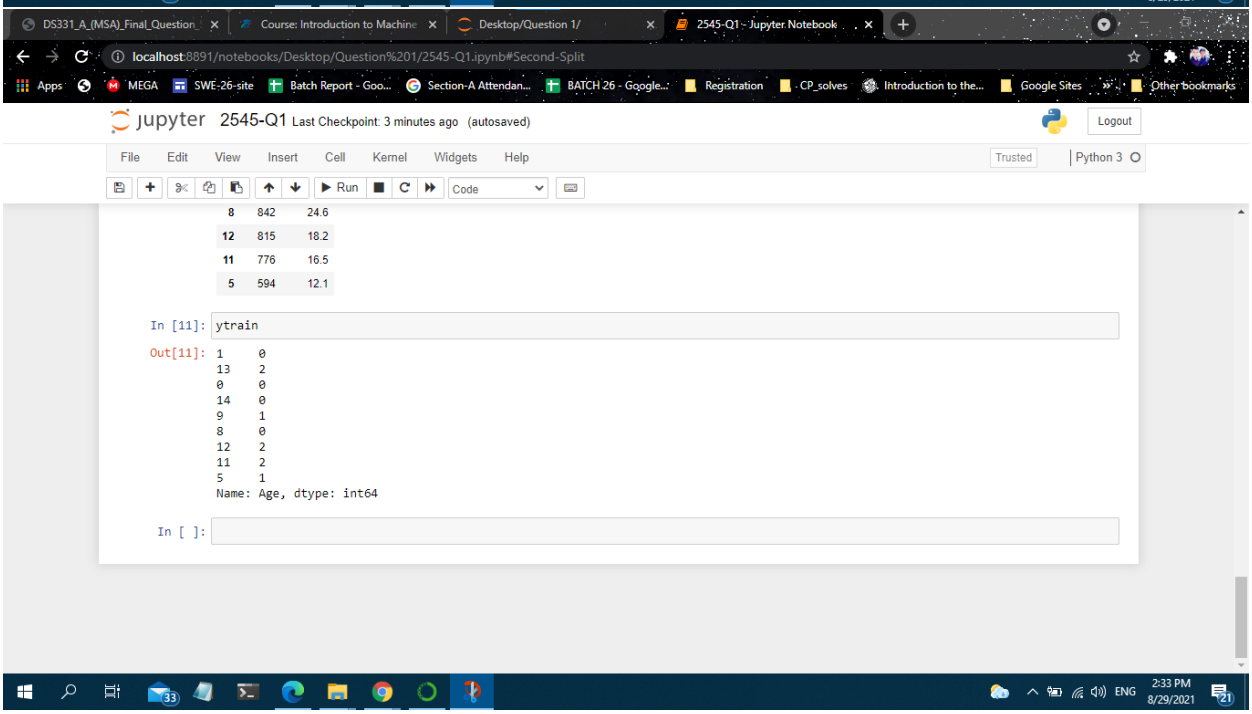
In [7]:
```
xtrain
```

Out[7]:

	Costs	MileAge
0	643	18.2
14	673	17.5
9	706	19.1
8	842	24.6
12	815	18.2
11	776	16.5
5	594	12.1

In [8]:
```
ytrain
```


DS331_A_(MSA)_Final_Question_
Course: Introduction to Machine
Desktop/Question 1/
2545-Q1 - Jupyter Notebook
localhost:8891/notebooks/Desktop/Question%201/2545-Q1.ipynb#Second-Split
jupyter 2545-Q1 Last Checkpoint: 3 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
9 706 19.1
8 842 24.6
12 815 18.2
11 776 16.5
5 594 12.1
In [8]: ytrain
Out[8]: 0 0
14 0
9 1
8 0
12 2
11 2
5 1
Name: Age, dtype: int64
Second Split (40-60)
In [9]: xtrain, xtest, ytrain, ytest = train_test_split(x, y, test_size = .40, random_state = 1)
In [10]: xtrain
Out[10]:
Costs MileAge
2:33 PM
8/29/2021


DS331_A_(MSA)_Final_Question_
Course: Introduction to Machine
Desktop/Question 1/
2545-Q1 - Jupyter Notebook
localhost:8891/notebooks/Desktop/Question%201/2545-Q1.ipynb#Second-Split
jupyter 2545-Q1 Last Checkpoint: 3 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
8 842 24.6
12 815 18.2
11 776 16.5
5 594 12.1
In [11]: ytrain
Out[11]: 1 0
13 2
0 0
14 0
9 1
8 0
12 2
11 2
5 1
Name: Age, dtype: int64
In [ ]:
2:33 PM
8/29/2021

2.1

2545

Ans. to the Ques⇒02

From the given picture, we can see that there multiple variable we can use to predict. So, multivarient linear regression is the algorithm I will use to find out the final mark. The formula ⇒

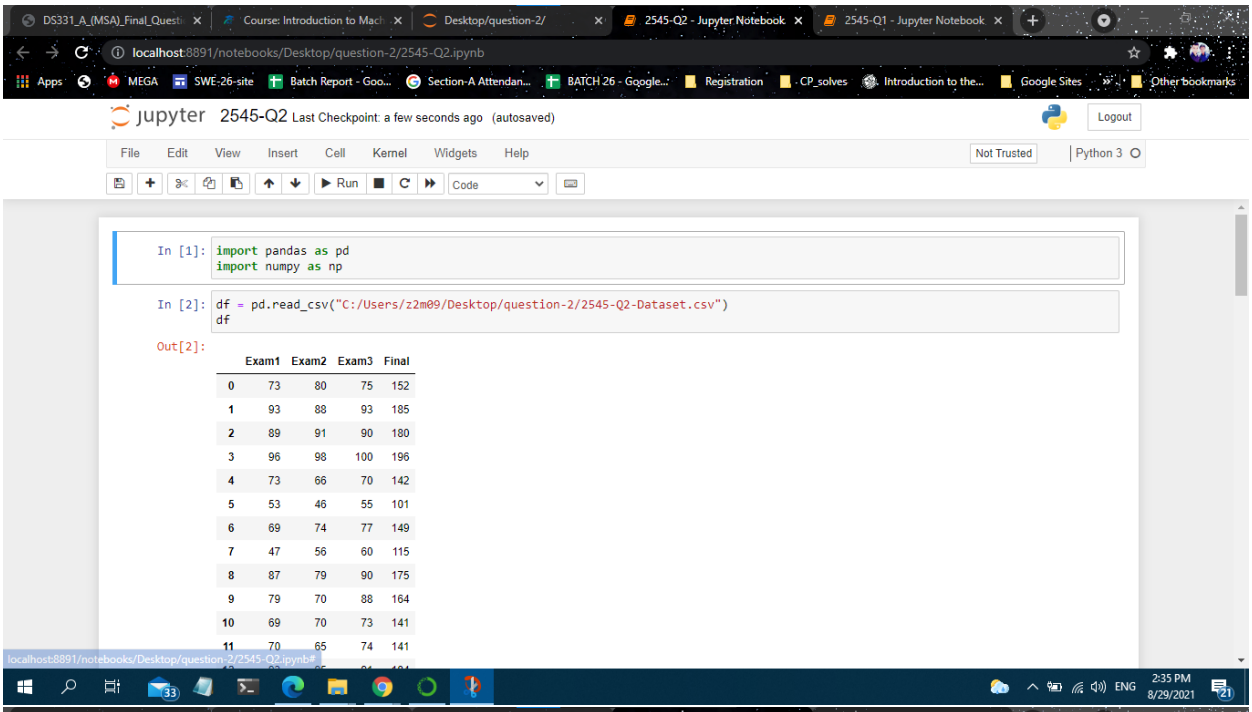
$$y = m_1x_1 + m_2x_2 + m_3x_3 + \dots + c.$$

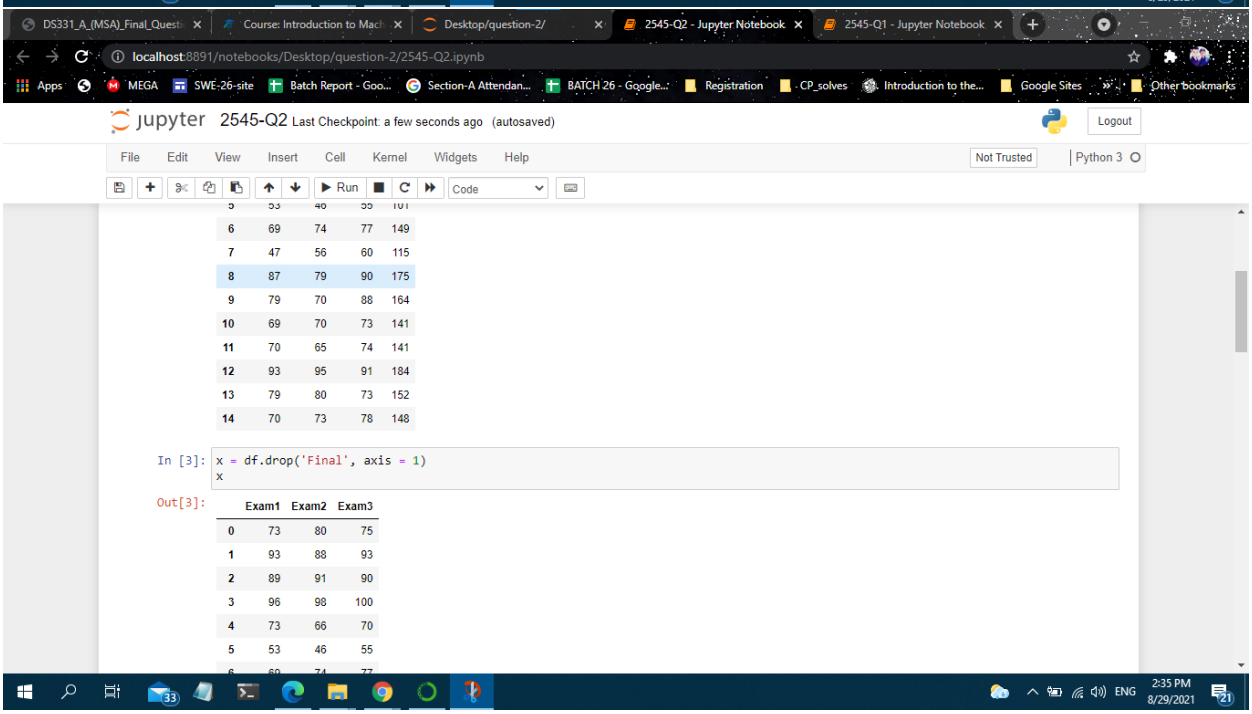
-this is the equation of multivarient linear regression, where more than one independent variable are present and only one dependent variable.

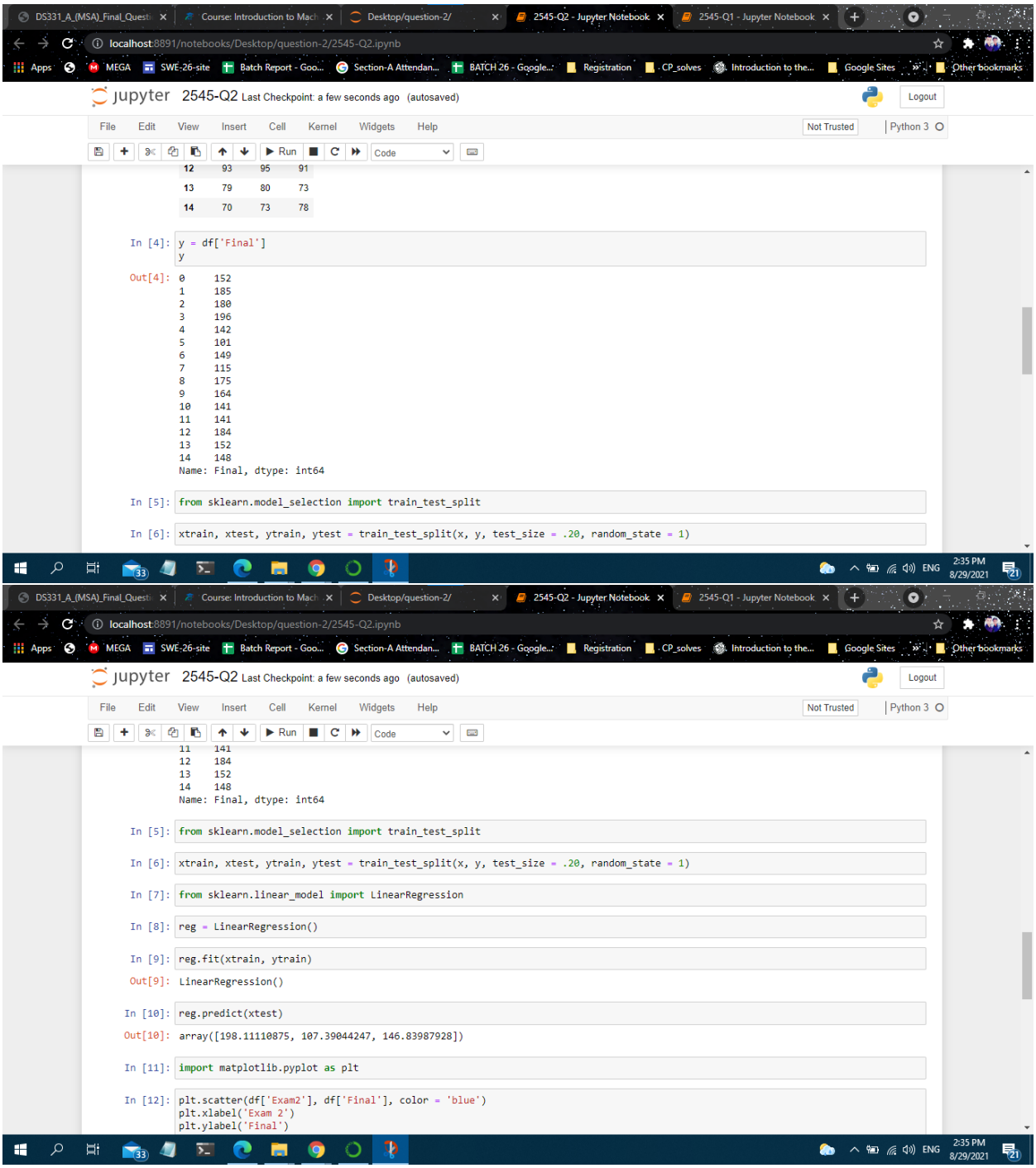
Now, Hypothesis equation ⇒

$$h_\theta(x) = \theta_0 + \theta_1x_1 + \theta_2x_2 + \theta_3x_3 + \dots$$

As there were multiple features, so the hypothesis is dependent on multiple variables as well.


DS331_A_(MSA)_Final_Questi
Course: Introduction to Mach
Desktop/question-2/
2545-Q2 - Jupyter Notebook
2545-Q1 - Jupyter Notebook
localhost:8891/notebooks/Desktop/question-2/2545-Q2.ipynb
Apps
MEGA
SWE-26-site
Batch Report - Goo...
Section-A Attendan...
BATCH 26 - Google...
Registration
CP_solves
Introduction to the...
Google Sites
Other bookmarks
jupyter 2545-Q2 Last Checkpoint: a few seconds ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3
Run
Code
In [1]:
import pandas as pd
import numpy as np
In [2]:
df = pd.read_csv("C:/Users/z2m09/Desktop/question-2/2545-Q2-Dataset.csv")
df
Out[2]:
Exam1 Exam2 Exam3 Final
0 73 80 75 152
1 93 88 93 185
2 89 91 90 180
3 96 98 100 196
4 73 66 70 142
5 53 46 55 101
6 69 74 77 149
7 47 56 60 115
8 87 79 90 175
9 79 70 88 164
10 69 70 73 141
11 70 65 74 141
localhost:8891/notebooks/Desktop/question-2/2545-Q2.ipynb#
ENG
2:35 PM
8/29/2021


DS331_A_(MSA)_Final_Questi
Course: Introduction to Mach
Desktop/question-2/
2545-Q2 - Jupyter Notebook
2545-Q1 - Jupyter Notebook
localhost:8891/notebooks/Desktop/question-2/2545-Q2.ipynb
Apps
MEGA
SWE-26-site
Batch Report - Goo...
Section-A Attendan...
BATCH 26 - Google...
Registration
CP_solves
Introduction to the...
Google Sites
Other bookmarks
jupyter 2545-Q2 Last Checkpoint: a few seconds ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3
Run
Code
6 69 74 77 149
7 47 56 60 115
8 87 79 90 175
9 79 70 88 164
10 69 70 73 141
11 70 65 74 141
12 93 95 91 184
13 79 80 73 152
14 70 73 78 148
In [3]:
x = df.drop('Final', axis = 1)
x
Out[3]:
Exam1 Exam2 Exam3
0 73 80 75
1 93 88 93
2 89 91 90
3 96 98 100
4 73 66 70
5 53 46 55
ENG
2:35 PM
8/29/2021



jupyter 2545-Q2 Last Checkpoint: a few seconds ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3

Run Code

12	93	95	91
13	79	80	73
14	70	73	78

In [4]:
```
y = df['Final']
y
```

Out[4]:
```
0     152
1     185
2     180
3     196
4     142
5     101
6     149
7     115
8     175
9     164
10    141
11    141
12    184
13    152
14    148
Name: Final, dtype: int64
```

In [5]:
```
from sklearn.model_selection import train_test_split
```

In [6]:
```
xtrain, xtest, ytrain, ytest = train_test_split(x, y, test_size = .20, random_state = 1)
```

jupyter 2545-Q2 Last Checkpoint: a few seconds ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3

Run Code

```
11    141
12    184
13    152
14    148
Name: Final, dtype: int64
```

In [5]:
```
from sklearn.model_selection import train_test_split
```

In [6]:
```
xtrain, xtest, ytrain, ytest = train_test_split(x, y, test_size = .20, random_state = 1)
```

In [7]:
```
from sklearn.linear_model import LinearRegression
```

In [8]:
```
reg = LinearRegression()
```

In [9]:
```
reg.fit(xtrain, ytrain)
```

Out[9]:
```
LinearRegression()
```

In [10]:
```
reg.predict(xtest)
```

Out[10]:
```
array([198.11110875, 107.39044247, 146.83987928])
```

In [11]:
```
import matplotlib.pyplot as plt
```

In [12]:
```
plt.scatter(df['Exam2'], df['Final'], color = 'blue')
plt.xlabel('Exam 2')
plt.ylabel('Final')
```

DS331_A_(MSA)_Final_Questi | Course: Introduction to Mach | Desktop/question-2/ | 2545-Q2 - Jupyter Notebook | 2545-Q1 - Jupyter Notebook

localhost:8891/notebooks/Desktop/question-2/2545-Q2.ipynb

Apps | MEGA | SWE-26-site | Batch Report - Goo... | Section-A Attendan... | BATCH 26 - Google... | Registration | CP_solves | Introduction to the... | Google Sites | Other bookmarks

jupyter 2545-Q2 Last Checkpoint: a minute ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3

Run Code

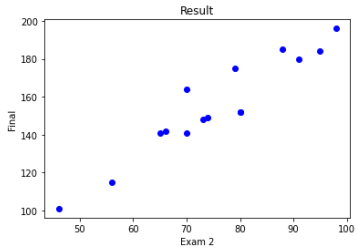
In [10]:
```
reg.predict(xtest)
```

Out[10]:
```
array([198.11110875, 107.39044247, 146.83987928])
```

In [11]:
```
import matplotlib.pyplot as plt
```

In [12]:
```
plt.scatter(df['Exam2'], df['Final'], color = 'blue')
plt.xlabel('Exam 2')
plt.ylabel('Final')
plt.title('Result')
```

Out[12]:
```
Text(0.5, 1.0, 'Result')
```



2:35 PM 8/29/2021

DS331_A_(MSA)_Final_Questi | Course: Introduction to Mach | Desktop/question-2/ | 2545-Q2 - Jupyter Notebook | 2545-Q1 - Jupyter Notebook

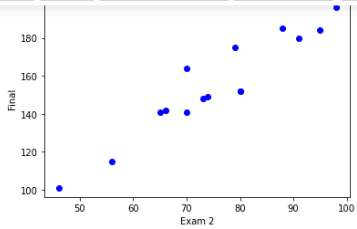
localhost:8891/notebooks/Desktop/question-2/2545-Q2.ipynb

Apps | MEGA | SWE-26-site | Batch Report - Goo... | Section-A Attendan... | BATCH 26 - Google... | Registration | CP_solves | Introduction to the... | Google Sites | Other bookmarks

jupyter 2545-Q2 Last Checkpoint: 2 minutes ago (unsaved changes) Logout

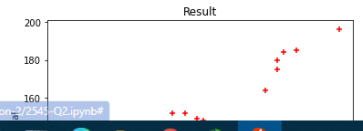
File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3

Run Code



In [13]:
```
plt.scatter(df['Exam3'], df['Final'], color = 'red', marker = '+')
plt.xlabel('Exam 3')
plt.ylabel('Final')
plt.title('Result')
```

Out[13]:
```
Text(0.5, 1.0, 'Result')
```



localhost:8891/notebooks/Desktop/question-2/2545-Q2.ipynb#

2:37 PM 8/29/2021

DS331_A_(MSA)_Final_Questi | Course: Introduction to Mach | Desktop/question-2/ | 2545-Q2 - Jupyter Notebook | 2545-Q1 - Jupyter Notebook

localhost:8891/notebooks/Desktop/question-2/2545-Q2.ipynb

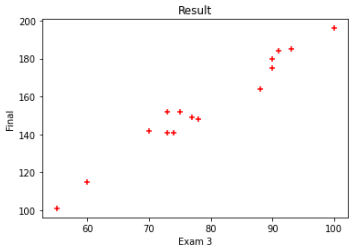
Apps | MEGA | SWE-26-site | Batch Report - Goo... | Section-A Attendan... | BATCH 26 - Google... | Registration | CP_solves | Introduction to the... | Google Sites | Other bookmarks

jupyter 2545-Q2 Last Checkpoint: 2 minutes ago (unsaved changes) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted | Python 3

Run Code

Out[13]: Text(0.5, 1.0, 'Result')



In [ ]:

ENG 2:37 PM 8/29/2021

3.1

2545

Ans. of the Ques ⇒03

In our question, there are three diagrams given showing different types of statistical fitting. If we analyze we will see ⇒

1) The first graph, the model fit on the dataset is too simple to explain the variance. It doesn't or barely shows any accuracy. In a downward straight-line slope, there are data points on ~~both~~ sides of it. So, this graph shows Under-fitting, which is high bias & low variance.

2) The second graph, the model fit on this dataset is appropiate. As the curved line touches a lot of the data points and all the

point are either very close to the curve on touching the curve. So, it is an ideal model for the dataset, showing appropiate fitting.

3) The third/last graph, the model fit on this dataset is like force-fitting it too good to be true. It shows an example of Overfitting, which is a very complex model. The data here gets trained with so much of data, it starts to learning from the noise & inaccurate data entries. Thus, it loses the accuracy needed. So, it is Overfitting - which is low bias and high variance -

3.3

2545

Techniques to reduce overfitting ⇒

1. Increase training data.
2. Reduce model complexity
3. Early stopping during the training phase.
4. Ridge regularization & lasso regularization.
5. Use dropout for neural networks to tackle ~~love~~ overfitting.

Techniques to reduce underfitting ⇒

1. Increase model complexity
2. Increase number of features, performing feature engineering
3. Remove noise from the data.
4. Increase the number of epochs or increase the duration of training to get better results.

3.4 2545

Confusion Matrix:- As the name suggests it gives us a matrix as output & describes the complete performance of the model. It is an evaluation metrics used for accuracy measurement of a model. Example:-

n= 150	Predicted: NO	Predicted:- Yes
Actual:- NO	35	15
Actual:- Yes	10	90

This confusion matrix works with 4 important terms ⇒

True Positives:- The cases in which we predicted "Yes" and the actual output was also "Yes."

Example Predicted:- Rain will fall, actual:- Rain falls.

True Negatives:- The cases in which we predicted "NO" and the actual output was also "NO." Example:- Rain won't fall, actual:- Rain didn't fall.

3.6

2545

False Positives:- The cases in which we predicted 'Yes' but the actual output was "No." Example - Rain will fall, Actual - It didn't fall.

False Negatives:- The cases in which we predicted "NO" but the actual output was "Yes". Example:- Rain won't fall, Actual - It did fall.

Measuring accuracy by confusion matrix =>

- Accuracy = $\frac{\text{True positive + true negative}}{\text{total sample}}$.

4.1

2545

Ans. to the ques. => 04

After evaluating the given picture we can say that it is a Gradient descent algorithm. In the picture the graph is showing a hyperbola where there is a starting point and a final value ; goal. It refers to gradient descent algorithm, which follows the equation $y = x^2$, so the curve of this algorithm becomes a hyperbola - just like the given graph.

Benefits of Gradient descent algorithm =>

1. It can find an optimize way to reach the goal.

2. It can work with any arbitary objective function.

2545

3. This algorithm uses less memory and saves time by minimizing functions -

4. It helps to find the values of function's parameters. (co-efficients).

5. This algorithm minimizes a cost function as much as possible.

Now, the given function :- $(x+7)^2$

$x_0 = 4$

Learning rate, $\alpha = 0.05$

(as my Id last digit = 5)

$$\frac{dy}{dx} = \frac{d}{dx}(x+7)^2 = 2(x+7)$$

4.3

2545

Iteration - 1 :-

$$x_1 = x_0 - \alpha * \frac{dy}{dx}$$

$$= 4 - [0.05 * \{2(4+7)\}]$$

$$= 2.9$$

Iteration - 2 :-

$$x_2 = x_1 - \alpha * \frac{dy}{dx}$$

$$= 2.9 - [0.05 * \{2*(2.9+7)\}]$$

$$= 1.91$$

Iteration - 3 :-

$$x_3 = x_2 - \alpha * \frac{dy}{dx}$$

$$= 1.91 - [0.05 * \{2*(1.91+7)\}]$$

$$= 1.019$$

Iteration - 4 :-

$$x_4 = x_3 - \alpha * \frac{dy}{dx}$$

$$= 1.019 - [0.05 * \{2*(1.019+7)\}]$$

$$= 0.2171$$