

ICP 1

Edit

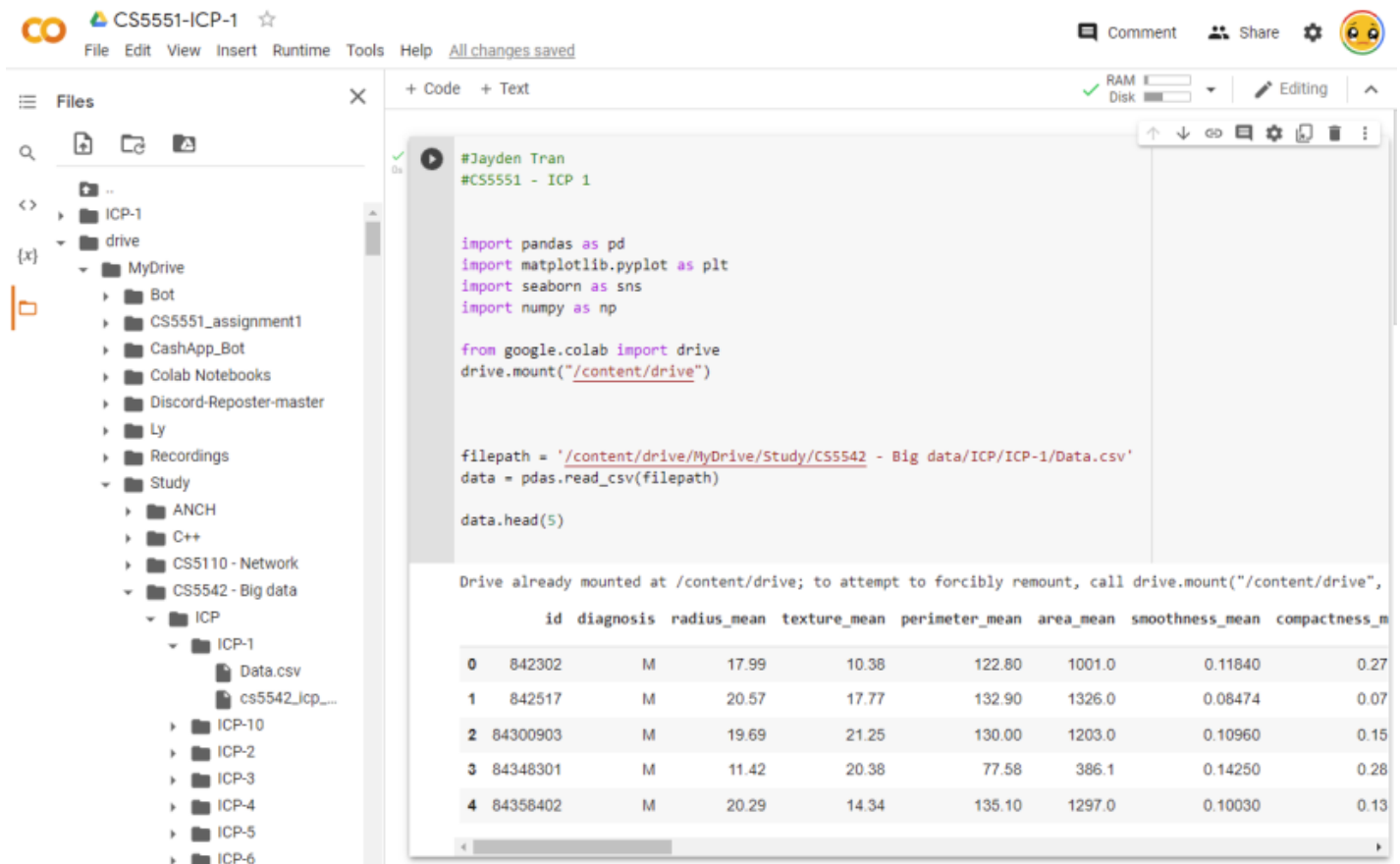
New Page

[Jump to bottom](#)

JaydenT4864 edited this page 39 minutes ago · 4 revisions

Mount the google drive to Google Colab notebook and read the data using python Pandas library.

Start to import libraries and load the dataset. Print out the data to see the format of dataset.



The screenshot shows a Google Colab notebook interface. The left sidebar displays the file explorer with a tree structure: `..`, `ICP-1`, `drive`, `MyDrive`, `Bot`, `CS5551_assignment1`, `CashApp_Bot`, `Colab Notebooks`, `Discord-Reposter-master`, `Ly`, `Recordings`, `Study`, `ANCH`, `C++`, `CS5110 - Network`, `CS5542 - Big data`, `ICP`, `ICP-1`, `Data.csv`, `cs5542_icp...`, `ICP-10`, `ICP-2`, `ICP-3`, `ICP-4`, `ICP-5`, and `ICP-6`.

The main code editor contains the following Python code:

```
#Jayden Tran
#CS5551 - ICP 1

import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np

from google.colab import drive
drive.mount("/content/drive")

filepath = '/content/drive/MyDrive/Study/CS5542 - Big data/ICP/ICP-1/Data.csv'
data = pd.read_csv(filepath)

data.head(5)
```

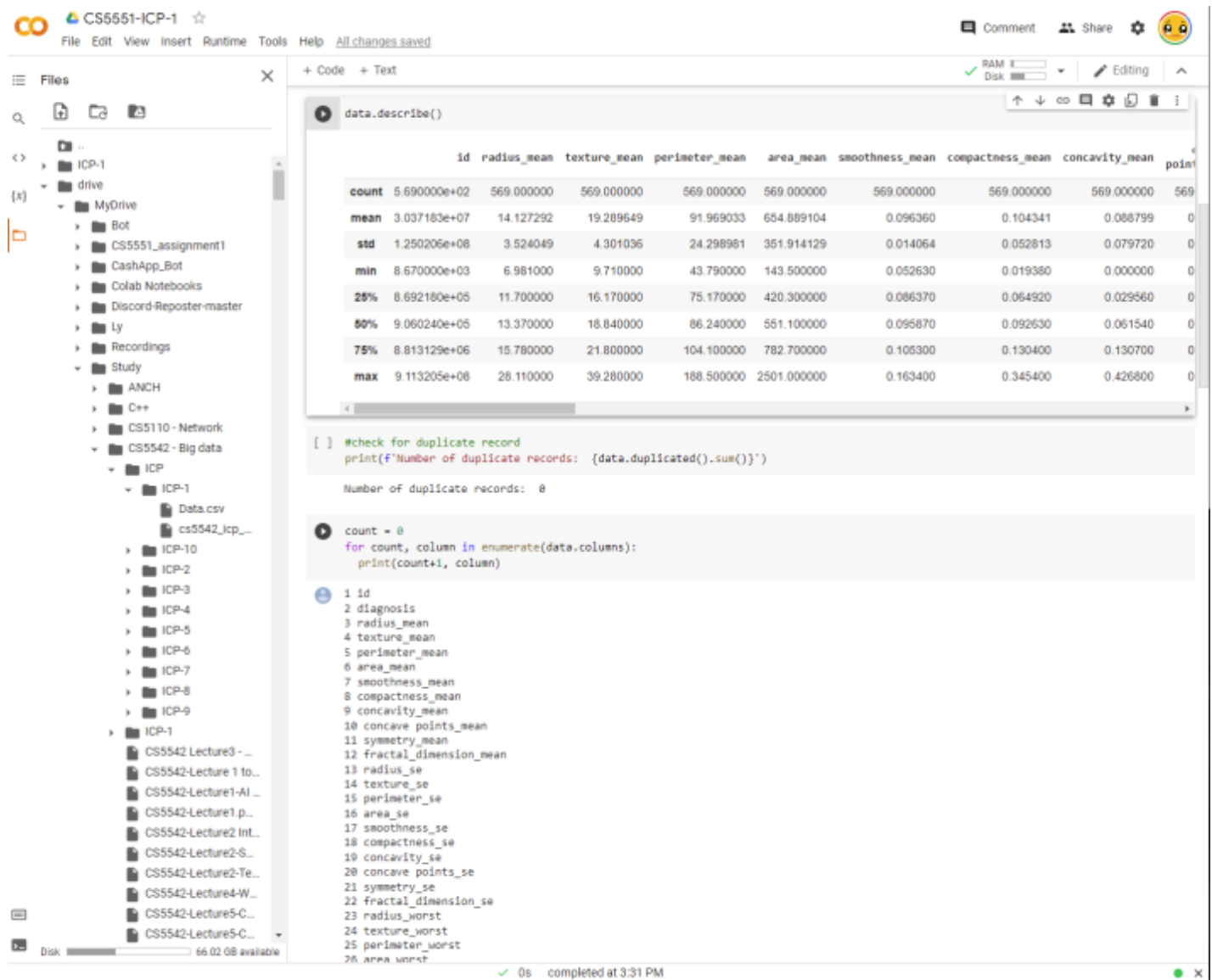
Below the code, a message states: "Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount(\"/content/drive\", force_remount=True)".

The output of the code shows the first five rows of the CSV file:

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_m
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13

Perform 3 data analysis tasks on the data. Analysis tasks are the task that give you insight about data. It could be visualization tasks, class distribution in the data, or anything that provides more information about the data.

Review the data describe and find duplicate records



The screenshot shows a Jupyter Notebook interface with a file explorer on the left and a code editor on the right. The file explorer shows a project structure with folders like 'ICP-1', 'CS5542-Big data', and 'ICP-10'. The code editor contains the following code:

```
data.describe()
```

	id	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	points
count	5.690000e+02	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569
mean	3.037183e+07	14.127292	19.289649	91.969033	654.889104	0.096360	0.104341	0.088799	0
std	1.250206e+08	3.524049	4.301036	24.298981	351.914129	0.014064	0.052813	0.079720	0
min	8.670000e+03	6.981000	9.710000	43.790000	143.500000	0.052630	0.019380	0.000000	0
25%	8.692180e+05	11.700000	16.170000	75.170000	420.300000	0.086370	0.064920	0.029560	0
50%	9.060240e+05	13.370000	18.840000	86.240000	551.100000	0.095870	0.092630	0.061540	0
75%	8.813129e+06	15.780000	21.800000	104.100000	782.700000	0.105300	0.130400	0.130700	0
max	9.113205e+08	28.110000	39.280000	188.500000	2501.000000	0.163400	0.345400	0.426800	0

```
[ ] #check for duplicate record
print(f'Number of duplicate records: {data.duplicated().sum()}')
```

Number of duplicate records: 0

```
count = 0
for count, column in enumerate(data.columns):
    print(count+1, column)
```

```
1 id
2 diagnosis
3 radius_mean
4 texture_mean
5 perimeter_mean
6 area_mean
7 smoothness_mean
8 compactness_mean
9 concavity_mean
10 concave points_mean
11 symmetry_mean
12 fractal_dimension_mean
13 radius_se
14 texture_se
15 perimeter_se
16 area_se
17 smoothness_se
18 compactness_se
19 concavity_se
20 concave points_se
21 symmetry_se
22 fractal_dimension_se
23 radius_worst
24 texture_worst
25 perimeter_worst
26 area_worst
```

0% completed at 3:31 PM

Find isnull values

CS5551-ICP-1

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

RAM Disk

Editing

Files

ICP-1

drive

MyDrive

Bot

CS5551_assignment1

CashApp_Bot

Colab Notebooks

Discord-Repoter-master

Ly

Recordings

Study

ANCH

C++

CS5110 - Network

CS5542 - Big data

ICP

ICP-1

ICP-10

ICP-2

ICP-3

ICP-4

ICP-5

ICP-6

ICP-7

ICP-8

ICP-9

CS5542-Lecture3 - ...

CS5542-Lecture1 to...

CS5542-Lecture1-Al ...

CS5542-Lecture1.p...

CS5542-Lecture2.Int...

CS5542-Lecture2-S...

CS5542-Lecture2-Te...

CS5542-Lecture4-W...

CS5542-Lecture5-C...

CS5542-Lecture5-C...

Disk 66.02 GB available

+ Code + Text

```

24 texture_worst
25 perimeter_worst
26 area_worst
27 smoothness_worst
28 compactness_worst
29 concavity_worst
30 concave_points_worst
31 symmetry_worst
32 fractal_dimension_worst
33 Unnamed: 32

```

#check the null values in the data

```
data.isnull().sum()
```

```

id                0
diagnosis         0
radius_mean      0
texture_mean     0
perimeter_mean  0
area_mean       0
smoothness_mean 0
compactness_mean 0
concavity_mean  0
concave_points_mean 0
symmetry_mean   0
fractal_dimension_mean 0
radius_se       0
texture_se      0
perimeter_se    0
area_se        0
smoothness_se  0
compactness_se  0
concavity_se    0
concave_points_se 0
symmetry_se     0
fractal_dimension_se 0
radius_worst    0
texture_worst   0
perimeter_worst 0
area_worst      0
smoothness_worst 0
compactness_worst 0
concavity_worst 0
concave_points_worst 0
symmetry_worst  0
fractal_dimension_worst 0
Unnamed: 32     560
dtype: int64

```

[11] #drop the unnamed columns and count again

```
data.drop(columns = ['Unnamed: 32'], axis = 0, inplace = True)
data.isnull().sum()
```

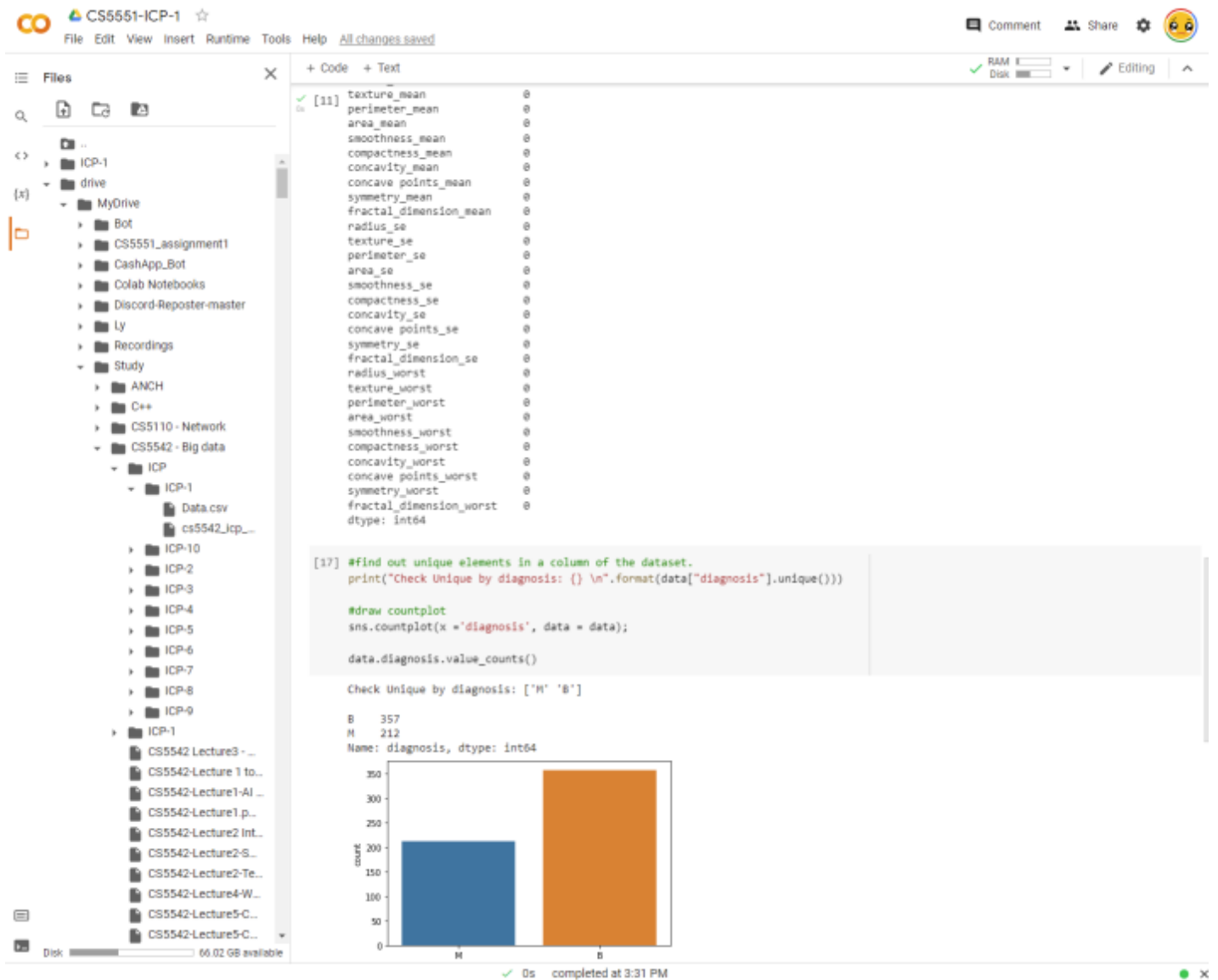
```

id                0
diagnosis         0
radius_mean      0
texture_mean     0
perimeter_mean  0

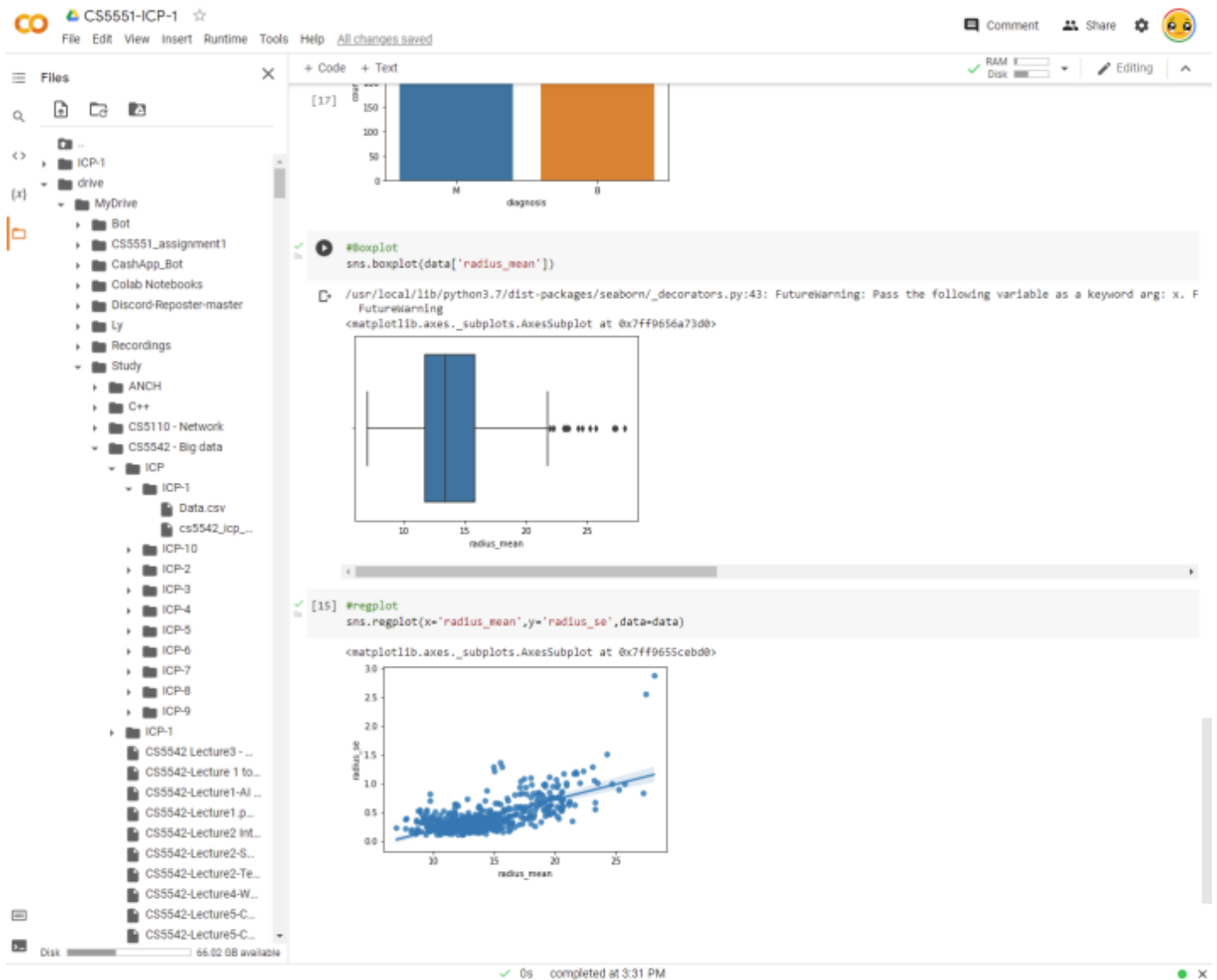
```

0s completed at 3:31 PM

Find unique element in a column of the dataset, and print out the plot



Boxplot and regplot



Conclusion

I learned how to sync and analyze the data from google colab. The challenge I have faced was mounting google drive with google collab

+ Add a custom footer

▼ Pages 11

Find a Page...

► [Home](#)

▼ [ICP 1](#)

Mount the google drive to Google Colab notebook and read the data using python Pandas library.

Conclusion

▸ [ICP 10](#)

▸ [ICP 2](#)

▸ [ICP 3](#)

▸ [ICP 4](#)

▸ [ICP 5](#)

▸ [ICP 6](#)

▸ [ICP 7](#)

▸ [ICP 8](#)

▸ [ICP 9](#)

+ Add a custom sidebar

Clone this wiki locally

<https://github.com/JaydenT4864/CS5542---Big-Data.wiki.git>

