Write-up	Correctness of Program	Documentation of Program	Viva	Timely Completion	Total	Dated Sign of Subject Teacher
4	4	4	4	4	20	

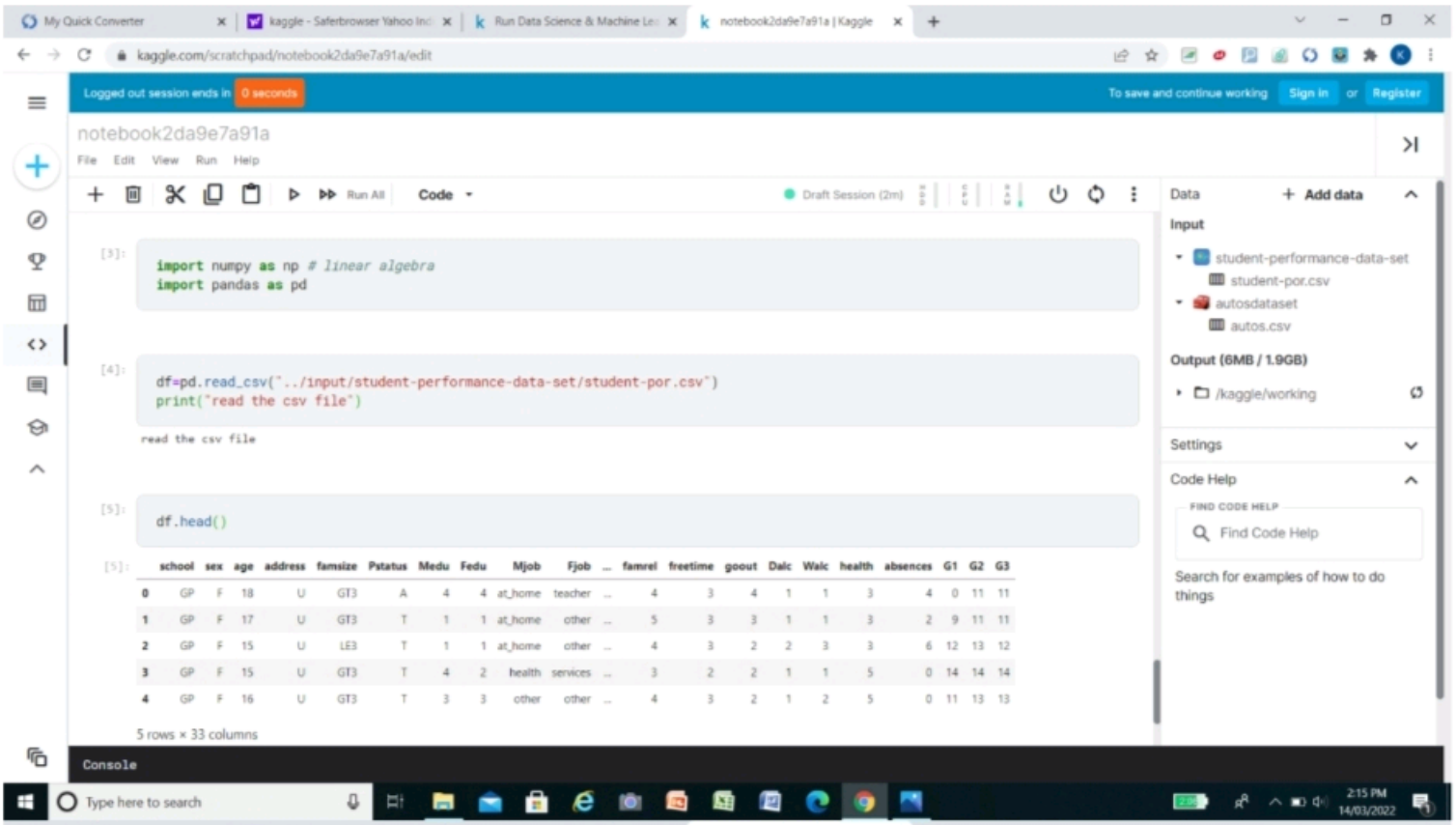
Group A
Assignment No: 1

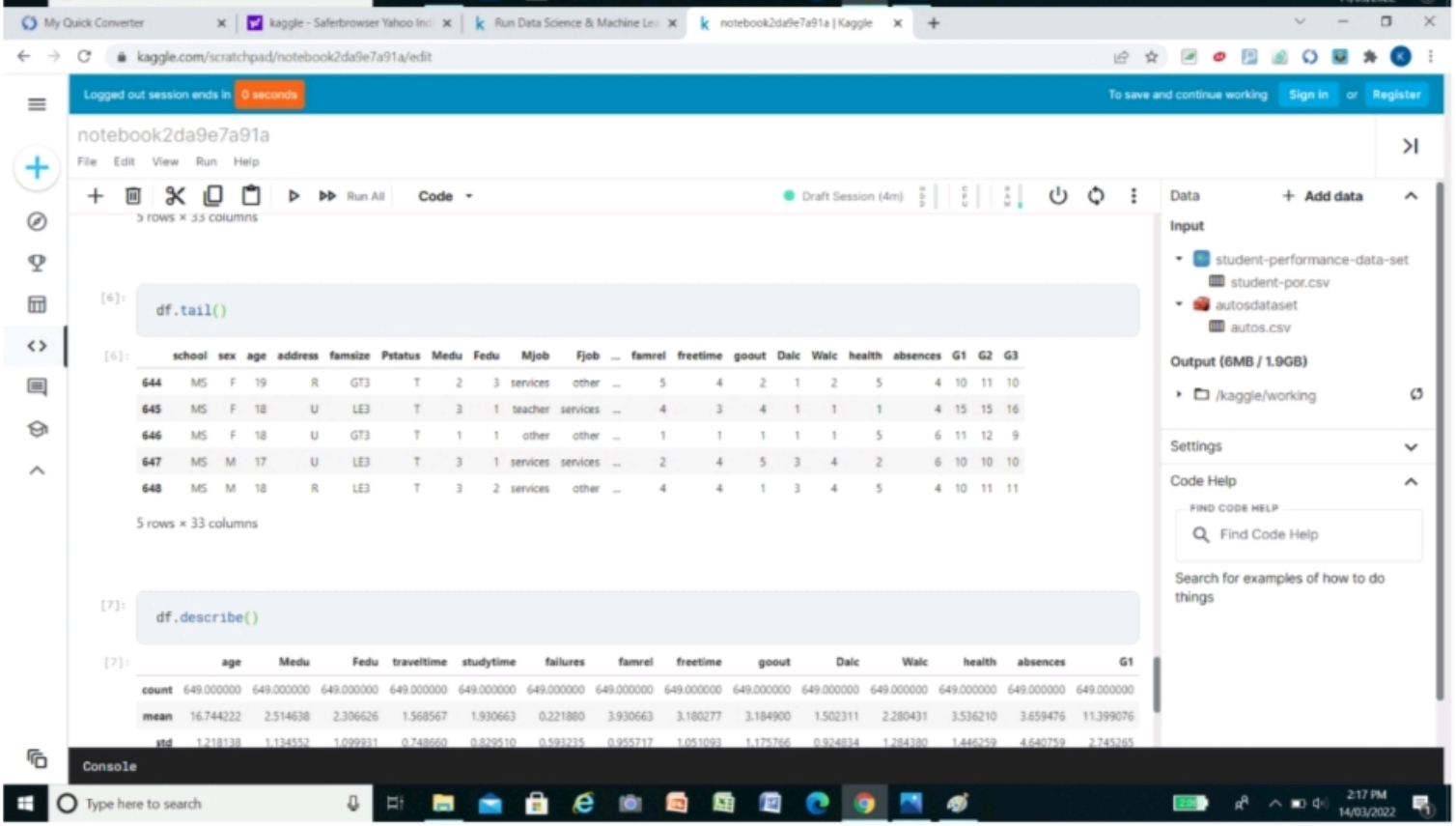
Title of the Assignment: Data Wrangling, I

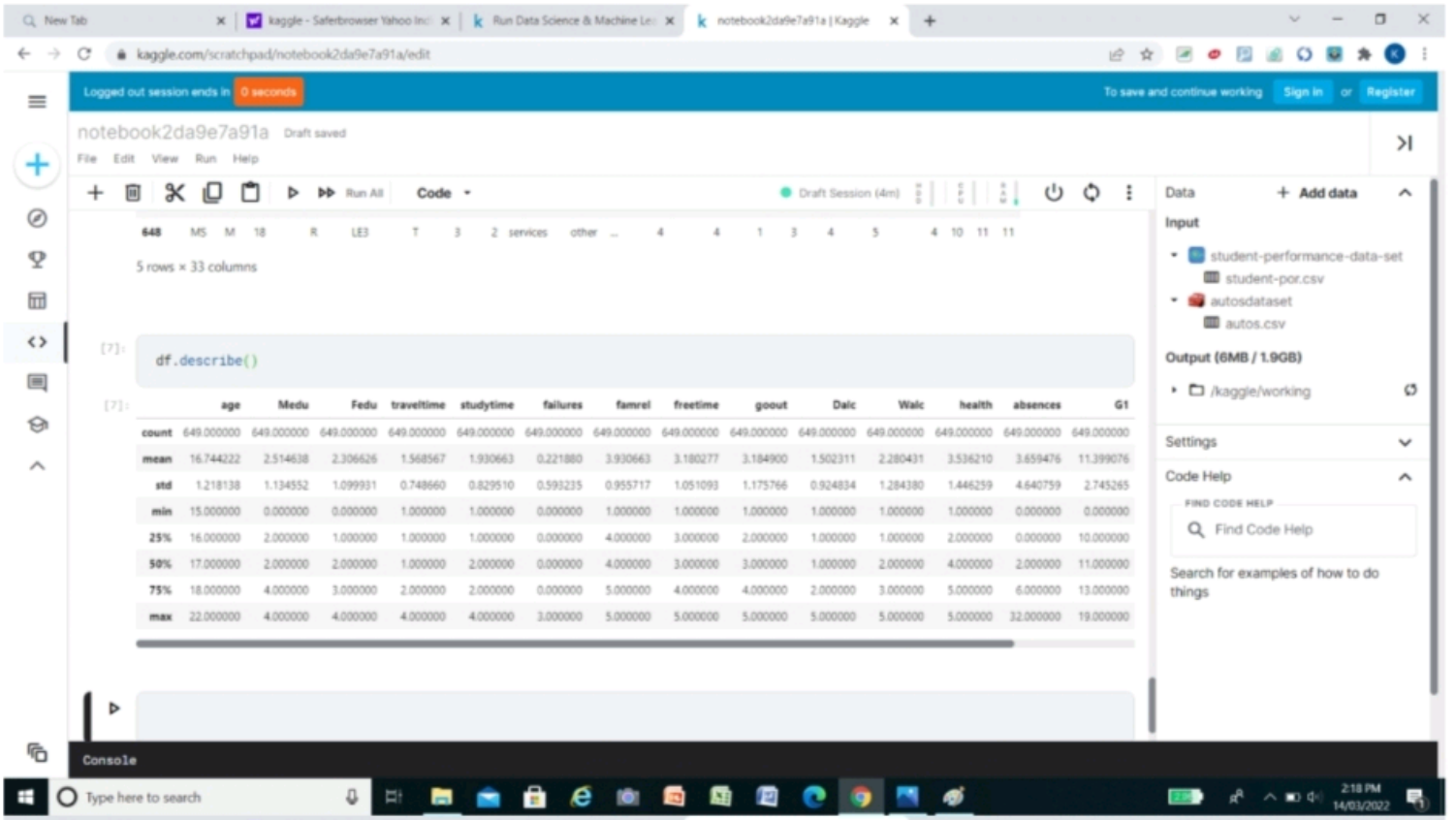
Perform the following operations using Python on any open source dataset (e.g., data.csv) Import all the required Python Libraries.

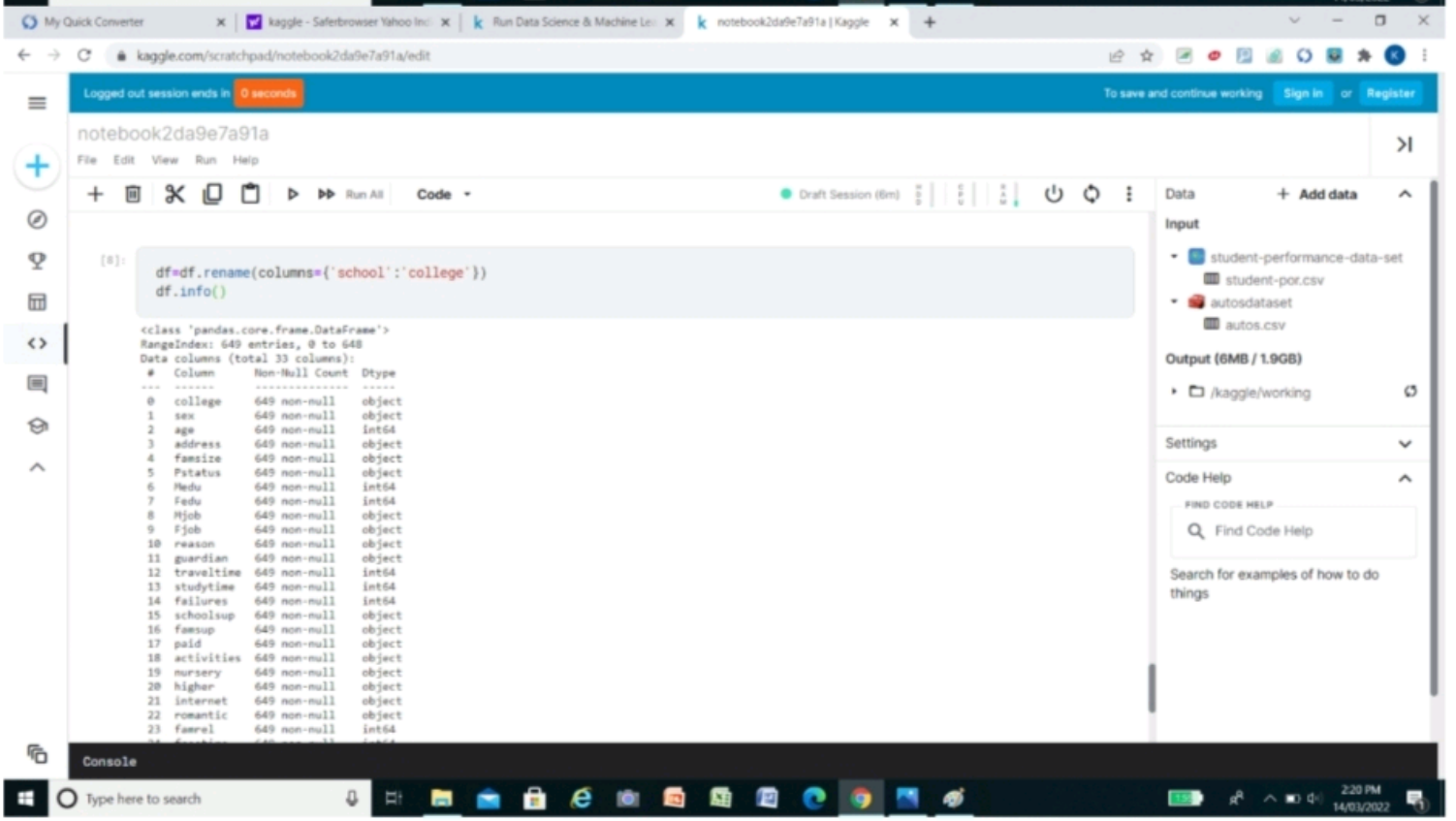
1. Locate open source data from the web (e.g. https://www.kaggle.com).
2. Provide a clear description of the data and its source (i.e., URL of the web site).
3. Load the Dataset into the pandas data frame.
4. Data Preprocessing: check for missing values in the data using pandas insult(), describe() function to get some initial statistics. Provide variable descriptions. Types of variables etc. Check the dimensions of the data frame.
5. Data Formatting and Data Normalization: Summarize the types of variables by checking the data types (i.e., character, numeric, integer, factor, and logical) of the variables in the data set. If variables are not in the correct data type, apply proper type conversions.
6. Turn categorical variables into quantitative variables in Python.

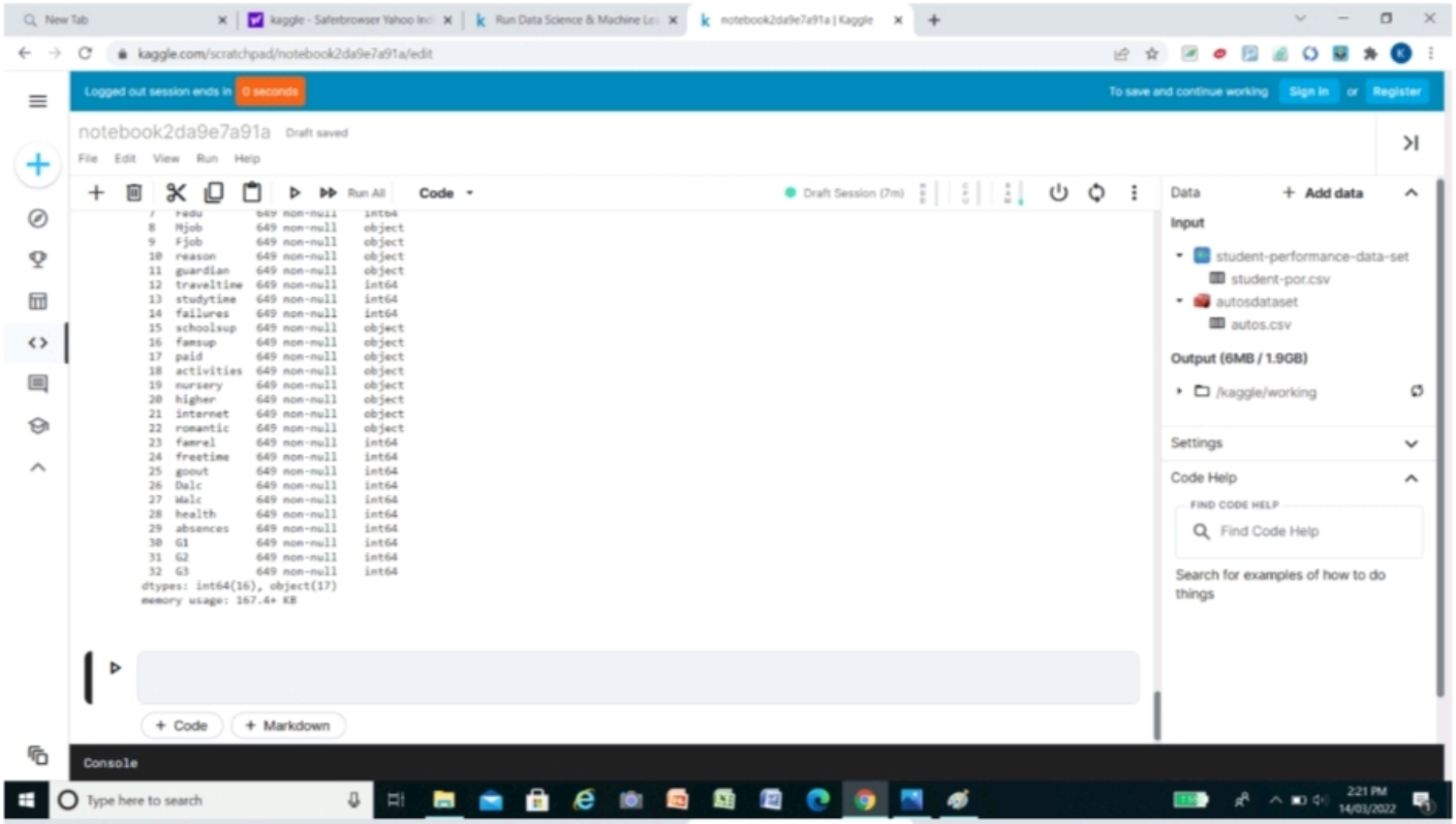
Output:

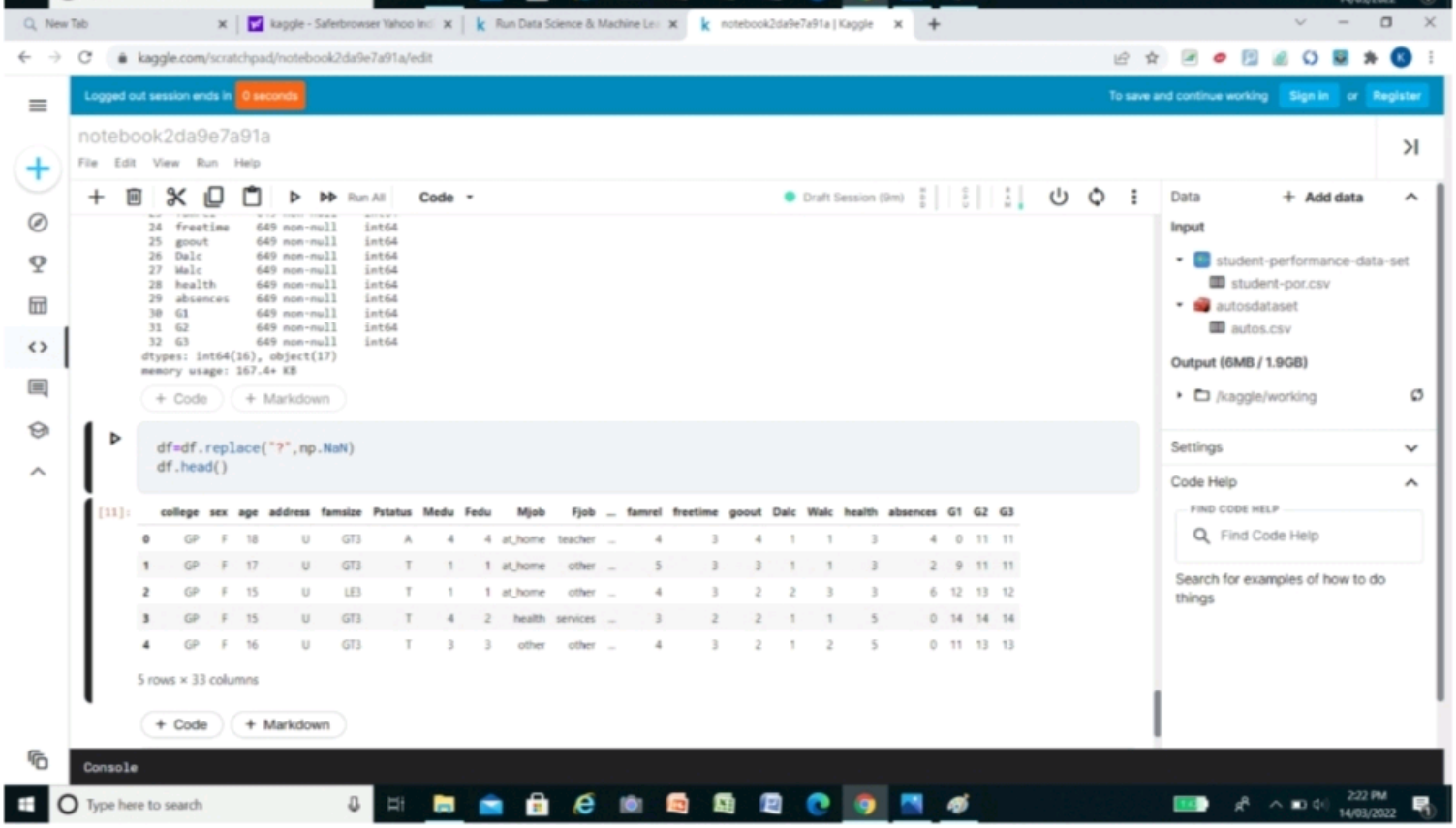

My Quick Converter
kaggle - Saferbrowser Yahoo Inc
Run Data Science & Machine Le
notebook2da9e7a91a | Kaggle
kaggle.com/scratchpad/notebook2da9e7a91a/edit
Logged out session ends in 0 seconds
To save and continue working Sign in or Register
notebook2da9e7a91a
File Edit View Run Help
Run All
Code
Draft Session (2m)
Data
Add data
Input
student-performance-data-set
student-por.csv
autosdataset
autos.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
[3]:
import numpy as np # linear algebra
import pandas as pd
[4]:
df=pd.read_csv("../input/student-performance-data-set/student-por.csv")
print("read the csv file")
read the csv file
[5]:
df.head()
[5]:
school sex age address famsize Pstatus Medu Fedu Mjob Fjob ... famrel freetime goout Dalc Walc health absences G1 G2 G3
0 GP F 18 U GT3 A 4 4 at_home teacher ... 4 3 4 1 1 3 4 0 11 11
1 GP F 17 U GT3 T 1 1 at_home other ... 5 3 3 1 1 3 2 9 11 11
2 GP F 15 U LE3 T 1 1 at_home other ... 4 3 2 2 3 3 6 12 13 12
3 GP F 15 U GT3 T 4 2 health services ... 3 2 2 1 1 5 0 14 14 14
4 GP F 16 U GT3 T 3 3 other other ... 4 3 2 1 2 5 0 11 13 13
5 rows × 33 columns
Console
Type here to search
2:15 PM
14/03/2022

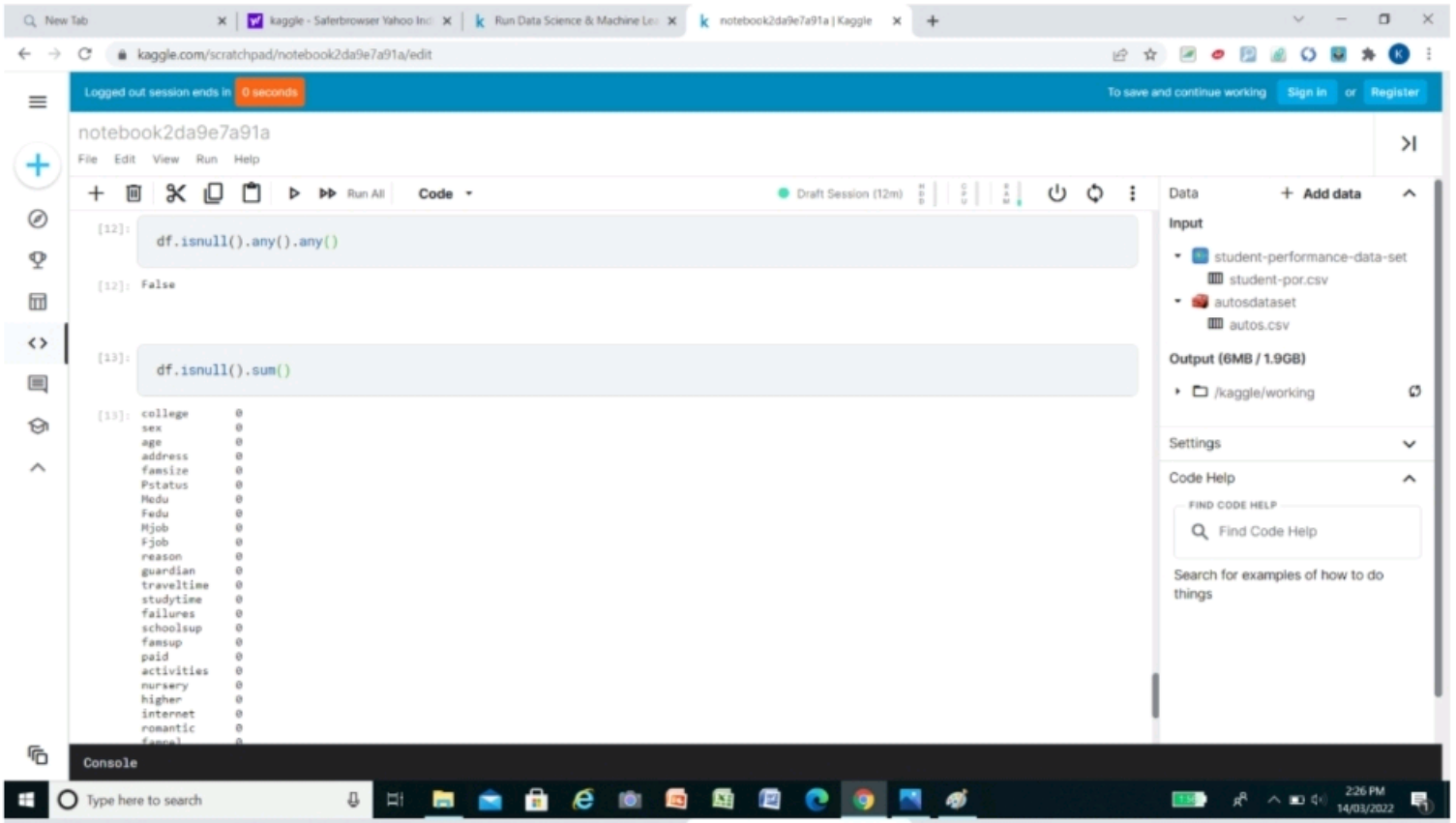

My Quick Converter
kaggle - Saferbrowser Yahoo Inc
Run Data Science & Machine Le
notebook2da9e7a91a | Kaggle
kaggle.com/scratchpad/notebook2da9e7a91a/edit
Logged out session ends in 0 seconds
To save and continue working Sign in or Register
notebook2da9e7a91a
File Edit View Run Help
Run All
Code
Draft Session (4m)
5 rows × 33 columns
Data
Add data
Input
student-performance-data-set
student-por.csv
autosdataset
autos.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
[6]:
df.tail()
[6]:
school sex age address famsize Pstatus Medu Fedu Mjob Fjob ... famrel freetime goout Dalc Walc health absences G1 G2 G3
644 MS F 19 R GT3 T 2 3 services other ... 5 4 2 1 2 5 4 10 11 10
645 MS F 18 U LE3 T 3 1 teacher services ... 4 3 4 1 1 1 4 15 15 16
646 MS F 18 U GT3 T 1 1 other other ... 1 1 1 1 1 5 6 11 12 9
647 MS M 17 U LE3 T 3 1 services services ... 2 4 5 3 4 2 6 10 10 10
648 MS M 18 R LE3 T 3 2 services other ... 4 4 1 3 4 5 4 10 11 11
5 rows × 33 columns
[7]:
df.describe()
[7]:
age Medu Fedu traveltime studytime failures famrel freetime goout Dalc Walc health absences G1
count 649.000000 649.000000 649.000000 649.000000 649.000000 649.000000 649.000000 649.000000 649.000000 649.000000 649.000000 649.000000 649.000000 649.000000
mean 16.744222 2.514638 2.306626 1.568567 1.930663 0.221880 3.930663 3.180277 3.184900 1.502311 2.280431 3.536210 3.659476 11.399076
std 1.218138 1.134552 1.099931 0.748660 0.829510 0.593235 0.955717 1.051093 1.175766 0.924834 1.284380 1.446259 4.640759 2.745265
Console
Type here to search
2:17 PM
14/03/2022

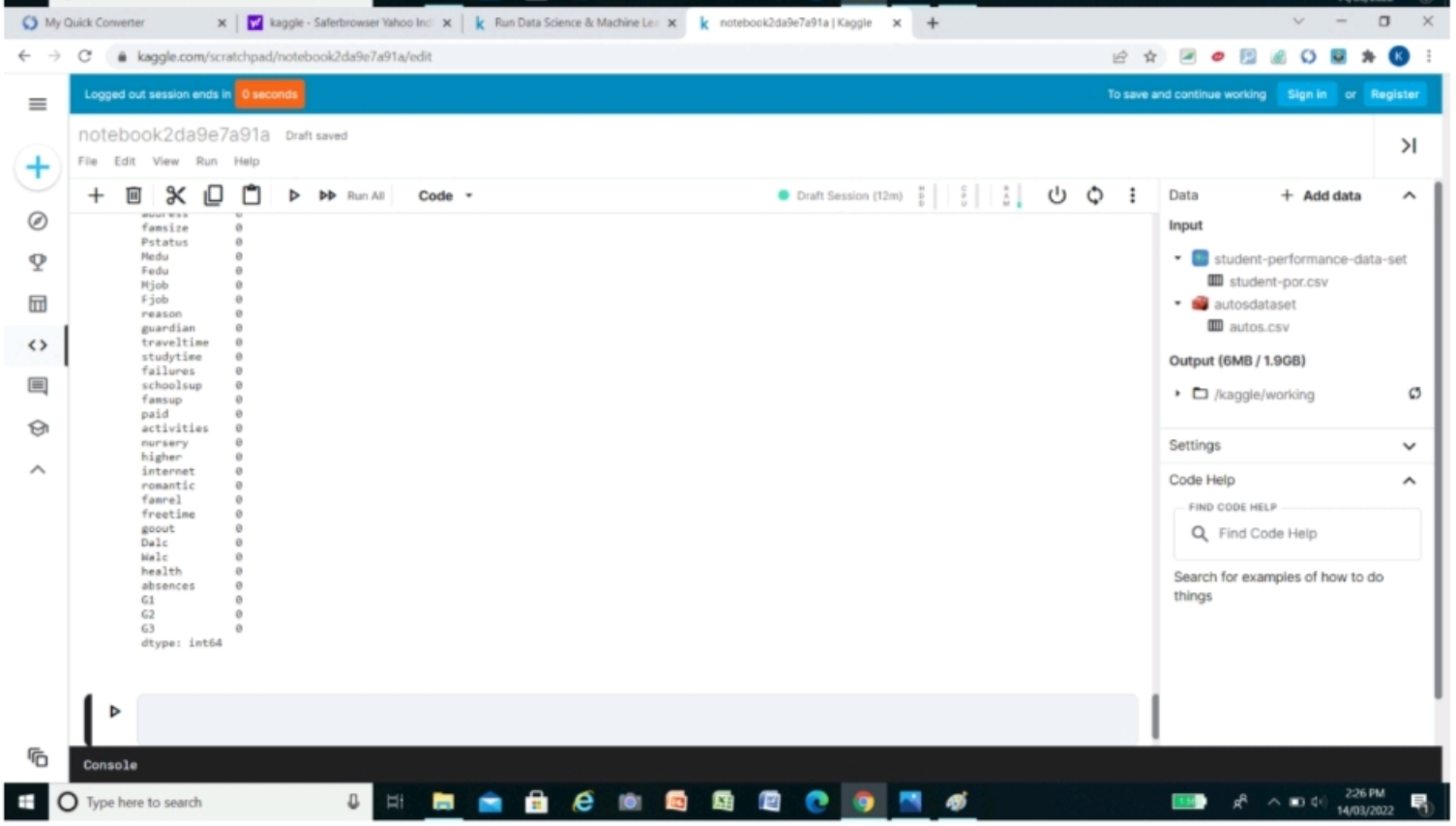

kaggle.com/scratchpad/notebook2da9e7a91a/edit
Logged out session ends in 0 seconds
To save and continue working Sign in or Register
notebook2da9e7a91a Draft saved
File Edit View Run Help
Run All Code Draft Session (4m)
Data Add data
Input
student-performance-data-set
student-por.csv
autosdataset
autos.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
648 MS M 18 R LE3 T 3 2 services other ... 4 4 1 3 4 5 4 10 11 11
5 rows × 33 columns
[7]: df.describe()
[7]:
age Medu Fedu traveltime studytime failures famrel freetime goout Dalc Walc health absences G1
count 649.000000 649.000000 649.000000 649.000000 649.000000 649.000000 649.000000 649.000000 649.000000 649.000000 649.000000 649.000000 649.000000 649.000000
mean 16.744222 2.514638 2.306626 1.568567 1.930663 0.221880 3.930663 3.180277 3.184900 1.502311 2.280431 3.536210 3.659476 11.399076
std 1.218138 1.134552 1.099931 0.748660 0.829510 0.593235 0.955717 1.051093 1.175766 0.924834 1.284380 1.446259 4.640759 2.745265
min 15.000000 0.000000 0.000000 1.000000 1.000000 0.000000 1.000000 1.000000 1.000000 1.000000 1.000000 1.000000 0.000000 0.000000
25% 16.000000 2.000000 1.000000 1.000000 1.000000 0.000000 4.000000 3.000000 2.000000 1.000000 1.000000 2.000000 0.000000 10.000000
50% 17.000000 2.000000 2.000000 1.000000 2.000000 0.000000 4.000000 3.000000 3.000000 1.000000 2.000000 4.000000 2.000000 11.000000
75% 18.000000 4.000000 3.000000 2.000000 2.000000 0.000000 5.000000 4.000000 4.000000 2.000000 3.000000 5.000000 6.000000 13.000000
max 22.000000 4.000000 4.000000 4.000000 4.000000 3.000000 5.000000 5.000000 5.000000 5.000000 5.000000 5.000000 32.000000 19.000000
Console
Type here to search
2:18 PM 14/03/2022

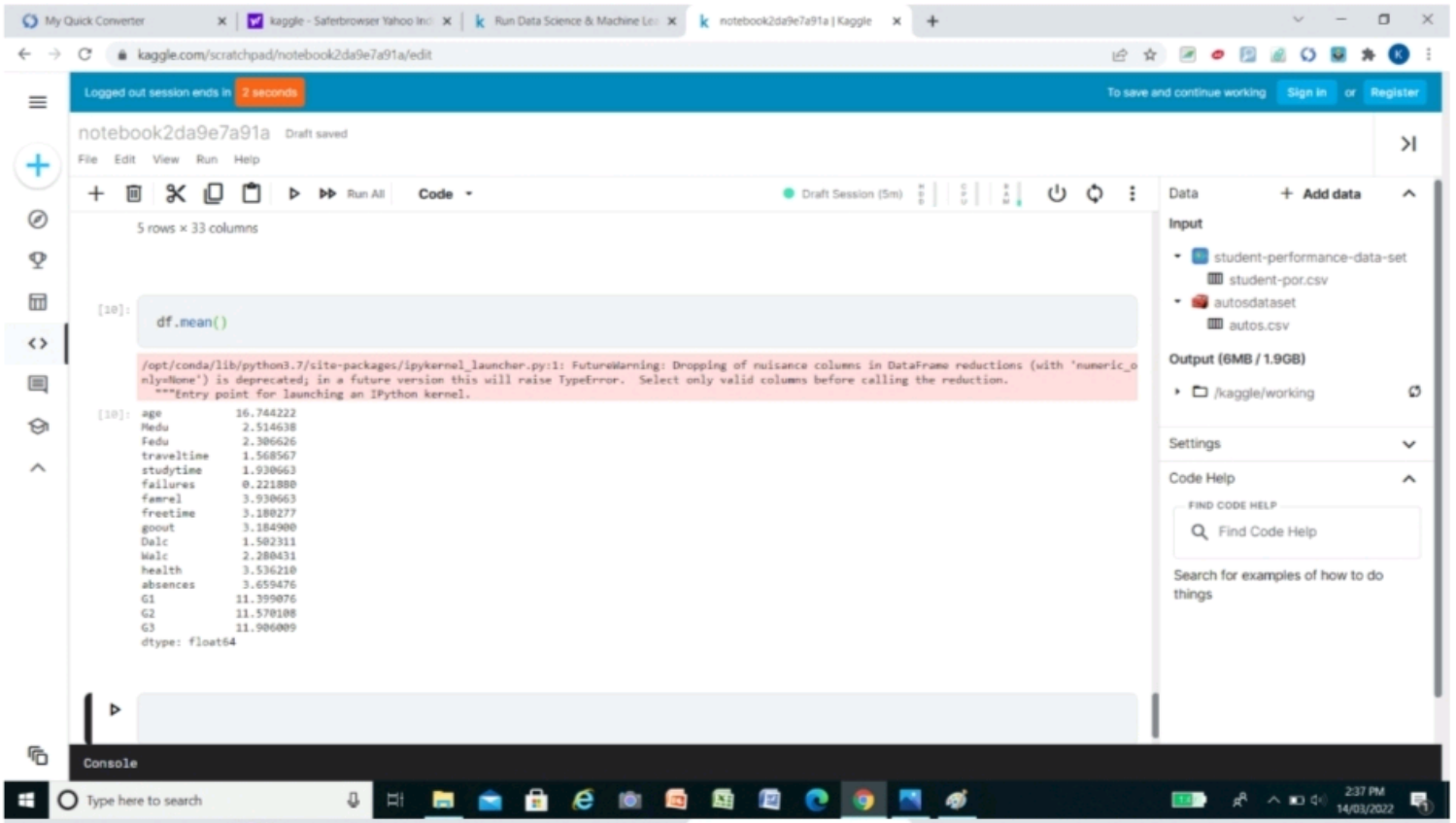

kaggle.com/scratchpad/notebook2da9e7a91a/edit
Logged out session ends in 0 seconds
To save and continue working Sign in or Register
notebook2da9e7a91a
File Edit View Run Help
Run All Code Draft Session (6m)
Data Add data
Input
student-performance-data-set
student-por.csv
autosdataset
autos.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
[8]:
df=df.rename(columns={'school':'college'})
df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 649 entries, 0 to 648
Data columns (total 33 columns):
Column Non-Null Count Dtype
0 college 649 non-null object
1 sex 649 non-null object
2 age 649 non-null int64
3 address 649 non-null object
4 famsize 649 non-null object
5 Pstatus 649 non-null object
6 Medu 649 non-null int64
7 Fedu 649 non-null int64
8 Mjob 649 non-null object
9 Fjob 649 non-null object
10 reason 649 non-null object
11 guardian 649 non-null object
12 traveltime 649 non-null int64
13 studytime 649 non-null int64
14 failures 649 non-null int64
15 schoolsup 649 non-null object
16 famsup 649 non-null object
17 paid 649 non-null object
18 activities 649 non-null object
19 nursery 649 non-null object
20 higher 649 non-null object
21 internet 649 non-null object
22 romantic 649 non-null object
23 famrel 649 non-null int64
Console
Type here to search
2:20 PM 14/03/2022

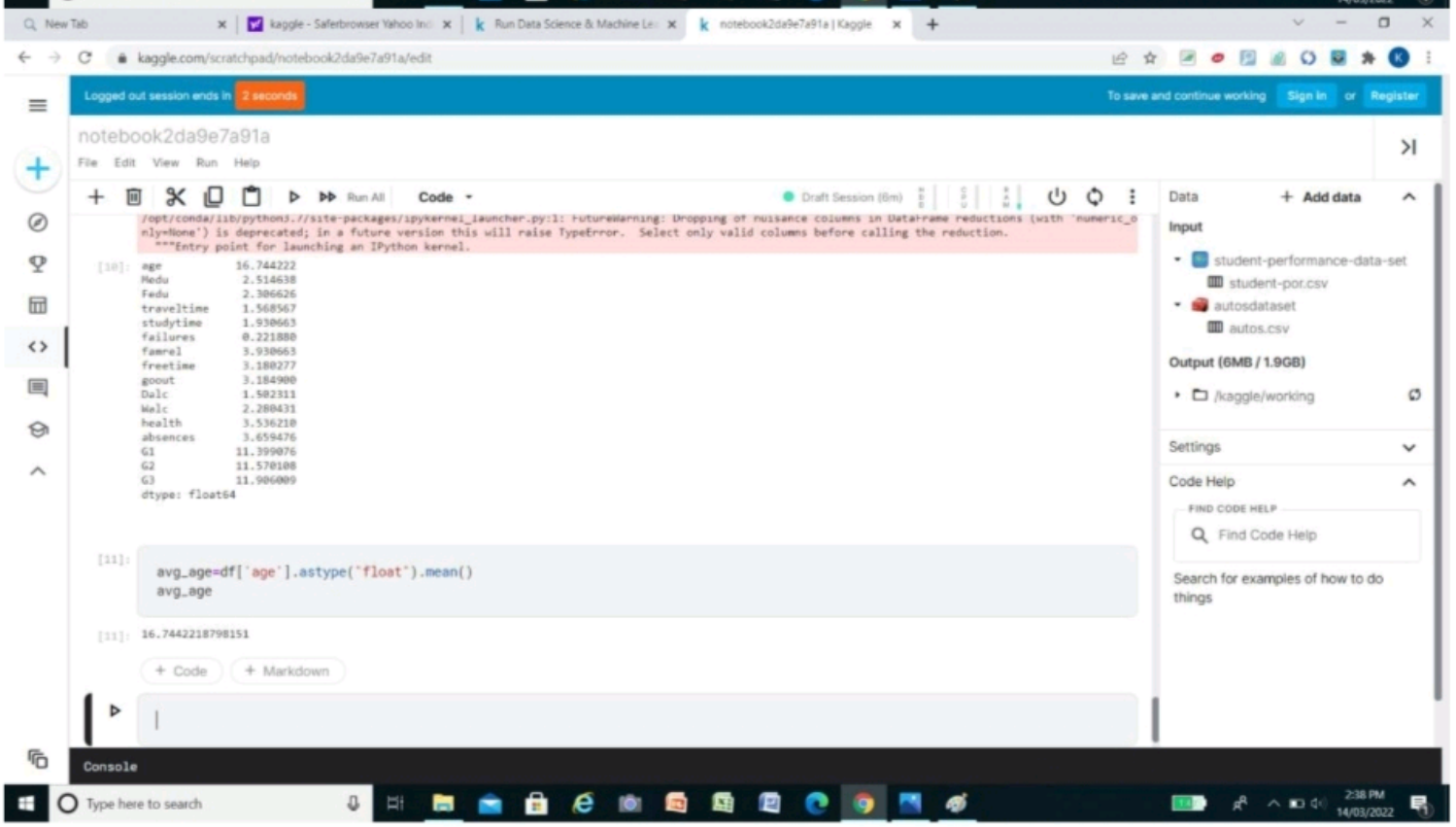

notebook2da9e7a91a | Kaggle
kaggle.com/scratchpad/notebook2da9e7a91a/edit
Logged out session ends in 0 seconds
To save and continue working Sign In or Register
notebook2da9e7a91a Draft saved
File Edit View Run Help
Run All Code Draft Session (7m)
7 Fedu 649 non-null int64
8 Mjob 649 non-null object
9 Fjob 649 non-null object
10 reason 649 non-null object
11 guardian 649 non-null object
12 traveltime 649 non-null int64
13 studytime 649 non-null int64
14 failures 649 non-null int64
15 schoolsup 649 non-null object
16 famsup 649 non-null object
17 paid 649 non-null object
18 activities 649 non-null object
19 nursery 649 non-null object
20 higher 649 non-null object
21 internet 649 non-null object
22 romantic 649 non-null object
23 famrel 649 non-null int64
24 freetime 649 non-null int64
25 goout 649 non-null int64
26 Dalc 649 non-null int64
27 Walc 649 non-null int64
28 health 649 non-null int64
29 absences 649 non-null int64
30 G1 649 non-null int64
31 G2 649 non-null int64
32 G3 649 non-null int64
dtypes: int64(16), object(17)
memory usage: 167.4+ KB
+ Code + Markdown
Console
Data + Add data
Input
student-performance-data-set
student-por.csv
autosdataset
autos.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Type here to search
2:21 PM 14/03/2022

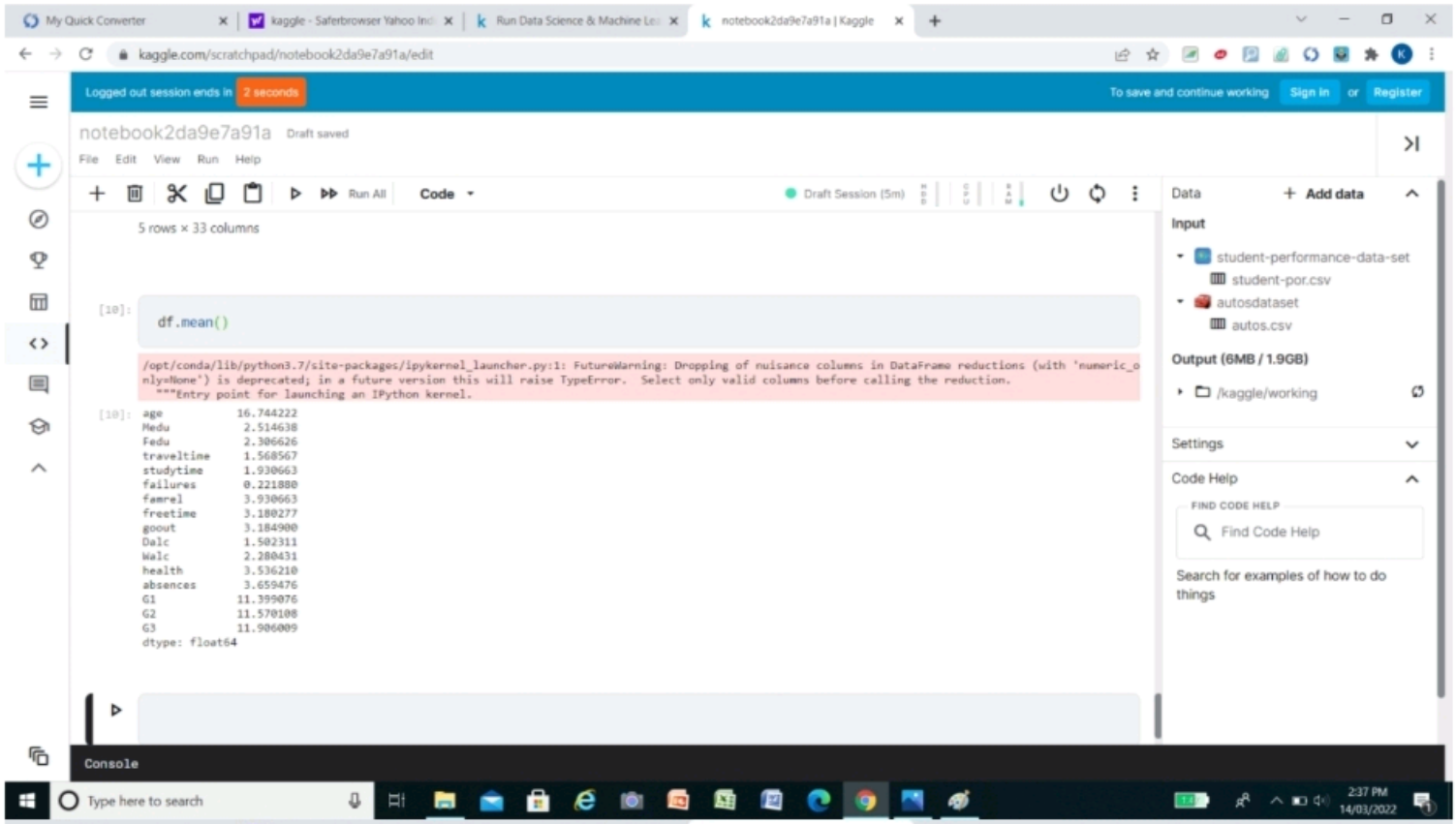

notebook2da9e7a91a | Kaggle
kaggle.com/scratchpad/notebook2da9e7a91a/edit
Logged out session ends in 0 seconds
To save and continue working Sign In or Register
notebook2da9e7a91a
File Edit View Run Help
Run All Code Draft Session (9m)
24 freetime 649 non-null int64
25 goout 649 non-null int64
26 Dalc 649 non-null int64
27 Walc 649 non-null int64
28 health 649 non-null int64
29 absences 649 non-null int64
30 G1 649 non-null int64
31 G2 649 non-null int64
32 G3 649 non-null int64
dtypes: int64(16), object(17)
memory usage: 167.4+ KB
+ Code + Markdown
df=df.replace("?",np.NaN)
df.head()
[11]:
college sex age address famsize Pstatus Medu Fedu Mjob Fjob ... famrel freetime goout Dalc Walc health absences G1 G2 G3
0 GP F 18 U GT3 A 4 4 at_home teacher ... 4 3 4 1 1 3 4 0 11 11
1 GP F 17 U GT3 T 1 1 at_home other ... 5 3 3 1 1 3 2 9 11 11
2 GP F 15 U LE3 T 1 1 at_home other ... 4 3 2 2 3 3 6 12 13 12
3 GP F 15 U GT3 T 4 2 health services ... 3 2 2 1 1 5 0 14 14 14
4 GP F 16 U GT3 T 3 3 other other ... 4 3 2 1 2 5 0 11 13 13
5 rows × 33 columns
+ Code + Markdown
Console
Data + Add data
Input
student-performance-data-set
student-por.csv
autosdataset
autos.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Type here to search
2:22 PM 14/03/2022

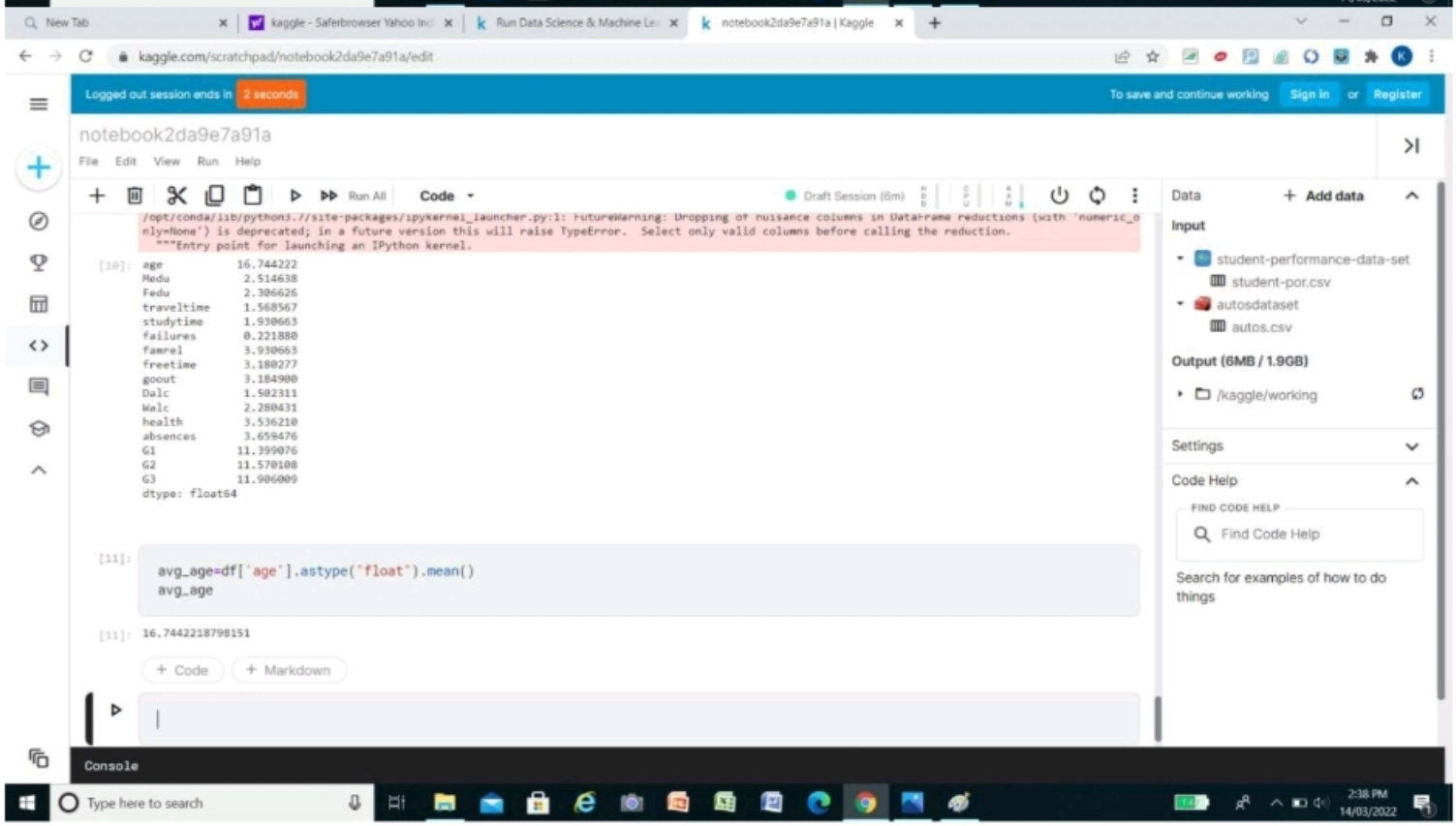

kaggle.com/scratchpad/notebook2da9e7a91a/edit
Logged out session ends in 0 seconds
To save and continue working Sign In or Register
notebook2da9e7a91a
File Edit View Run Help
Run All
Code
Draft Session (12m)
[12]: df.isnull().any().any()
[12]: False
[13]: df.isnull().sum()
[13]: college 0
sex 0
age 0
address 0
famsize 0
Pstatus 0
Medu 0
Fedu 0
Mjob 0
Fjob 0
reason 0
guardian 0
traveltime 0
studytime 0
failures 0
schoolsup 0
famsup 0
paid 0
activities 0
nursery 0
higher 0
internet 0
romantic 0
Data
Add data
Input
student-performance-data-set
student-por.csv
autosdataset
autos.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Console
Type here to search
2:26 PM
14/03/2022

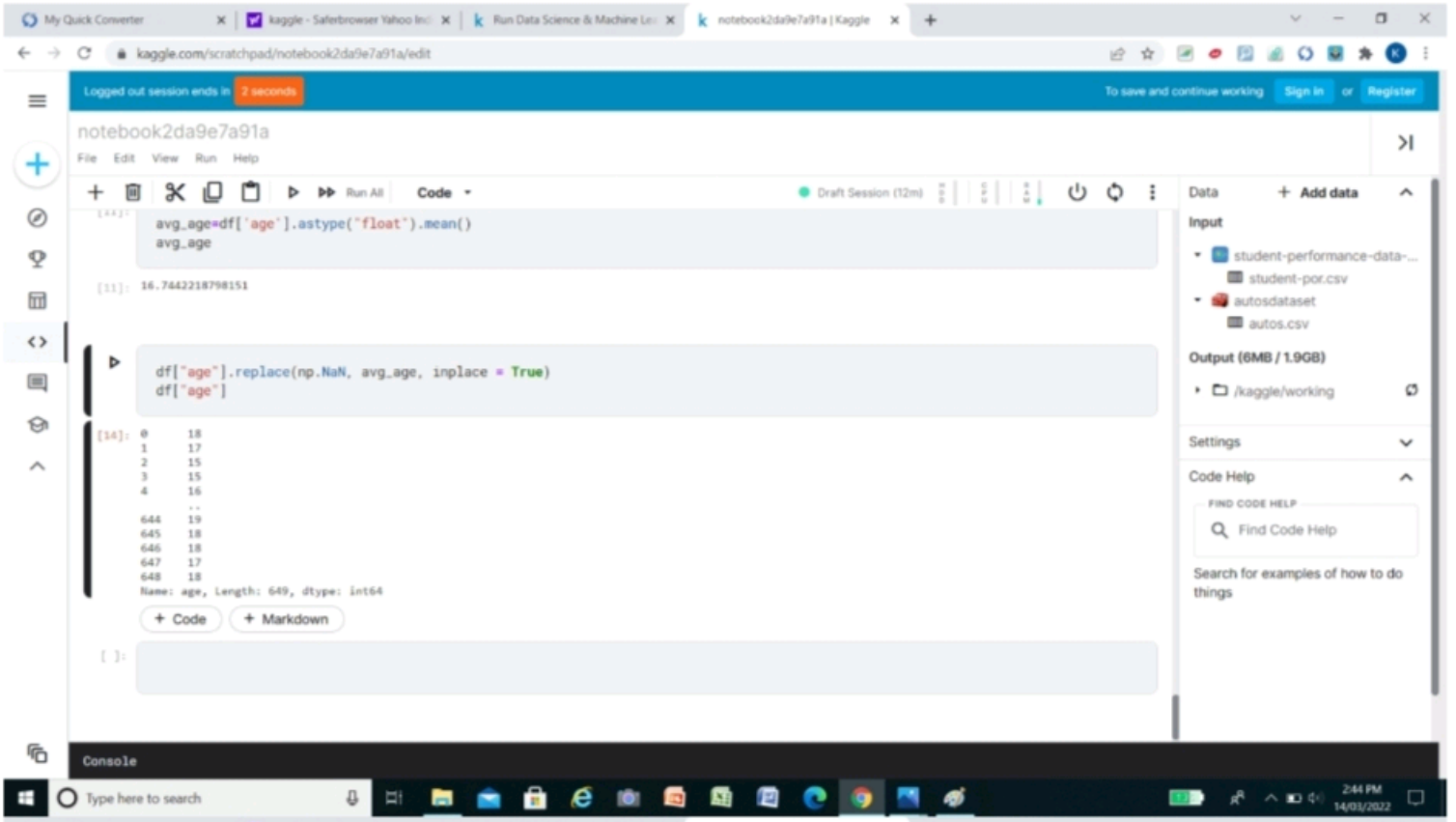

kaggle.com/scratchpad/notebook2da9e7a91a/edit
Logged out session ends in 0 seconds
To save and continue working Sign In or Register
notebook2da9e7a91a Draft saved
File Edit View Run Help
Run All
Code
Draft Session (12m)
famsize 0
Pstatus 0
Medu 0
Fedu 0
Mjob 0
Fjob 0
reason 0
guardian 0
traveltime 0
studytime 0
failures 0
schoolsup 0
famsup 0
paid 0
activities 0
nursery 0
higher 0
internet 0
romantic 0
famrel 0
freetime 0
goout 0
Dalc 0
Walc 0
health 0
absences 0
G1 0
G2 0
G3 0
dtype: int64
Data
Add data
Input
student-performance-data-set
student-por.csv
autosdataset
autos.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Console
Type here to search
2:26 PM
14/03/2022

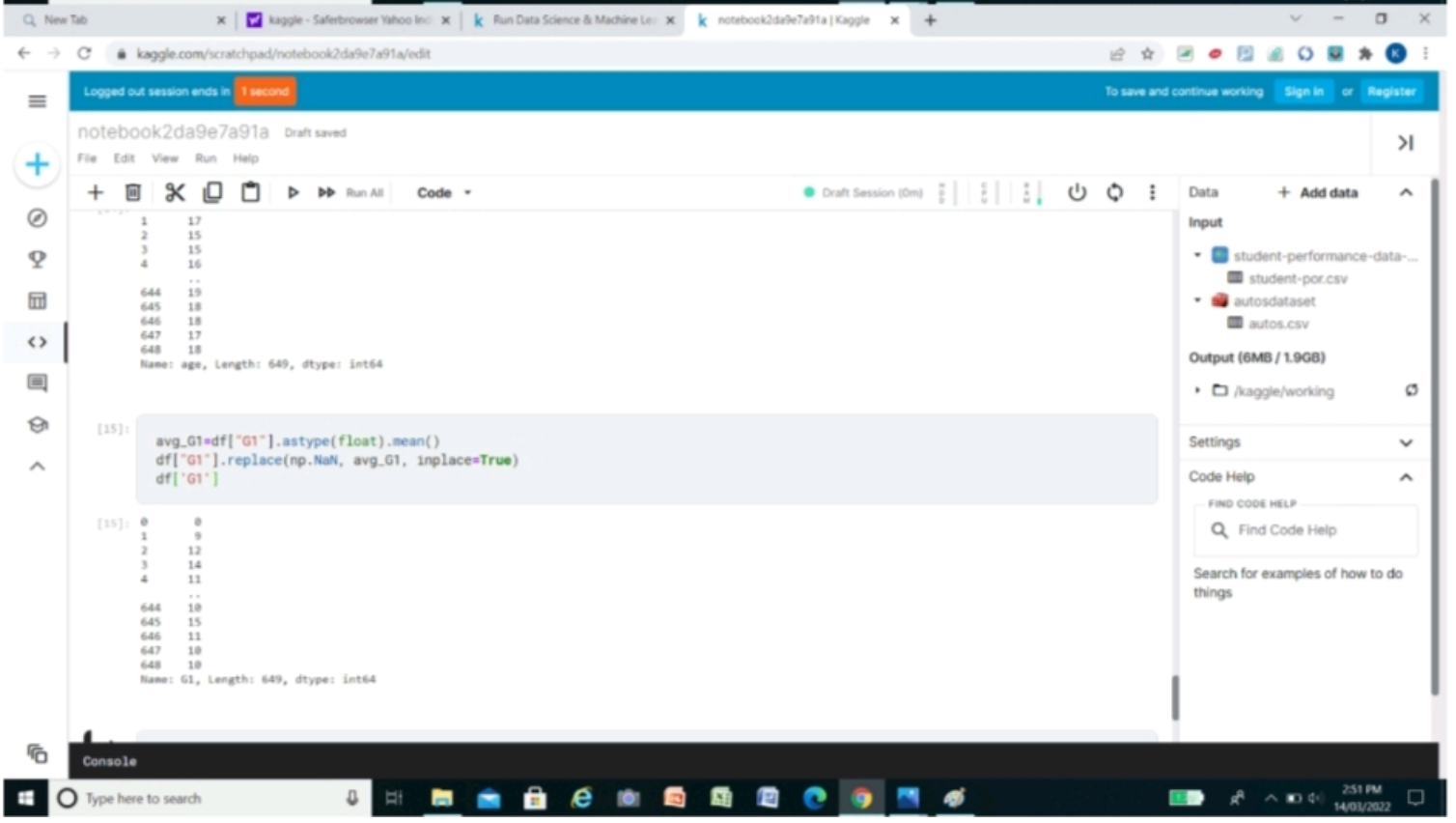

My Quick Converter
kaggle - Saferbrowser Yahoo Inc
Run Data Science & Machine Le
notebook2da9e7a91a | Kaggle
kaggle.com/scratchpad/notebook2da9e7a91a/edit
Logged out session ends in 2 seconds
To save and continue working Sign In or Register
notebook2da9e7a91a Draft saved
File Edit View Run Help
Run All Code Draft Session (5m)
5 rows × 33 columns
[10]: df.mean()
/opt/conda/lib/python3.7/site-packages/ipykernel_launcher.py:1: FutureWarning: Dropping of nuisance columns in DataFrame reductions (with 'numeric_only=None') is deprecated; in a future version this will raise TypeError. Select only valid columns before calling the reduction.
"""Entry point for launching an IPython kernel.
[10]:
age 16.744222
Medu 2.514638
Fedu 2.306626
traveltime 1.568567
studytime 1.930663
failures 0.221880
famrel 3.930663
freetime 3.180277
goout 3.184900
Dalc 1.502311
Walc 2.280431
health 3.536210
absences 3.659476
G1 11.399076
G2 11.570108
G3 11.906009
dtype: float64
Data + Add data
Input
student-performance-data-set
student-por.csv
autosdataset
autos.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Console
Type here to search
2:37 PM
14/03/2022

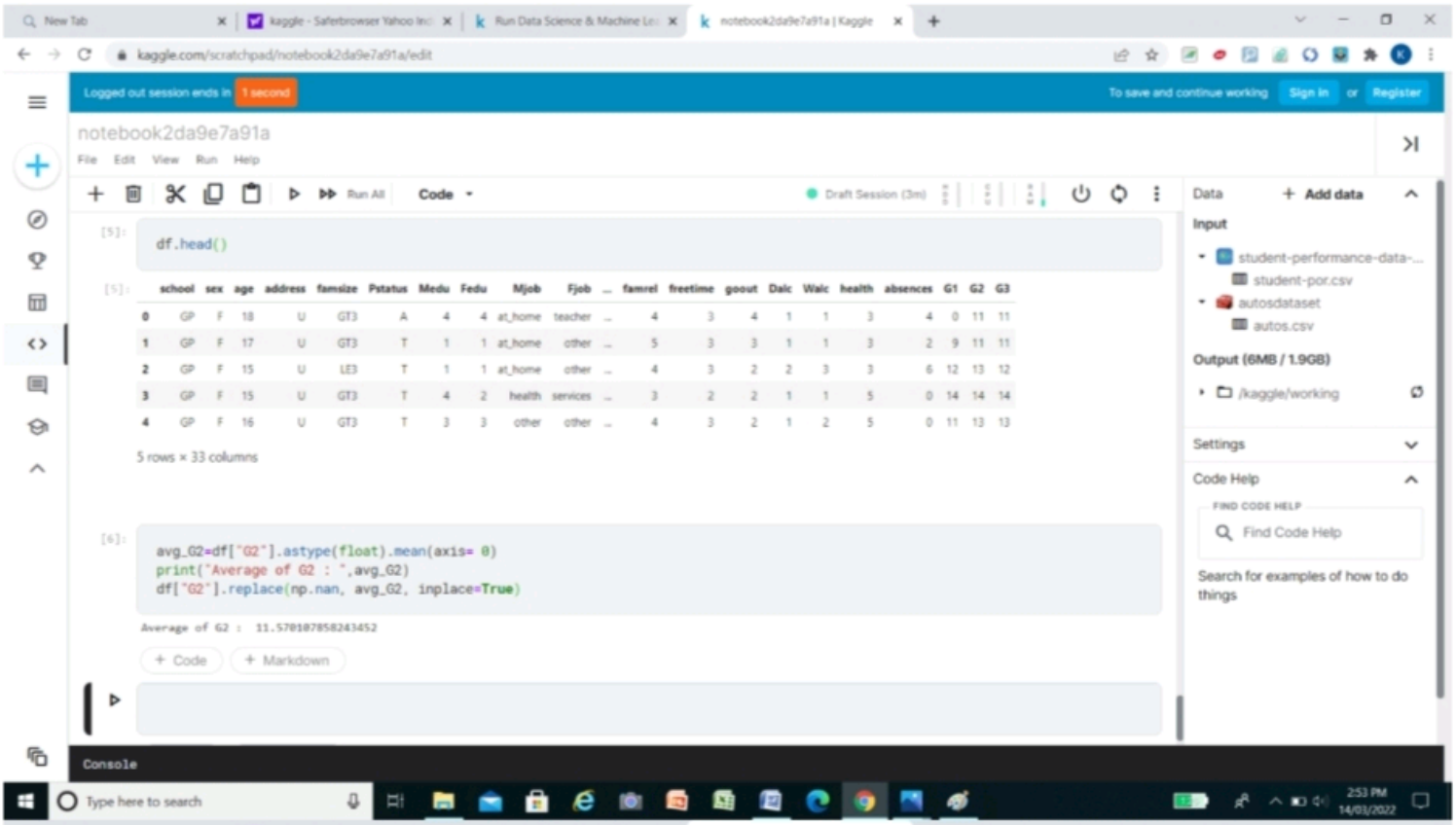

New Tab
kaggle - Saferbrowser Yahoo Inc
Run Data Science & Machine Le
notebook2da9e7a91a | Kaggle
kaggle.com/scratchpad/notebook2da9e7a91a/edit
Logged out session ends in 2 seconds
To save and continue working Sign In or Register
notebook2da9e7a91a
File Edit View Run Help
Run All Code Draft Session (6m)
/opt/conda/lib/python3.7/site-packages/ipykernel_launcher.py:1: FutureWarning: Dropping of nuisance columns in DataFrame reductions (with 'numeric_only=None') is deprecated; in a future version this will raise TypeError. Select only valid columns before calling the reduction.
"""Entry point for launching an IPython kernel.
[10]:
age 16.744222
Medu 2.514638
Fedu 2.306626
traveltime 1.568567
studytime 1.930663
failures 0.221880
famrel 3.930663
freetime 3.180277
goout 3.184900
Dalc 1.502311
Walc 2.280431
health 3.536210
absences 3.659476
G1 11.399076
G2 11.570108
G3 11.906009
dtype: float64
[11]: avg_age=df['age'].astype("float").mean()
avg_age
[11]: 16.7442218798151
+ Code + Markdown
Data + Add data
Input
student-performance-data-set
student-por.csv
autosdataset
autos.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Console
Type here to search
2:38 PM
14/03/2022

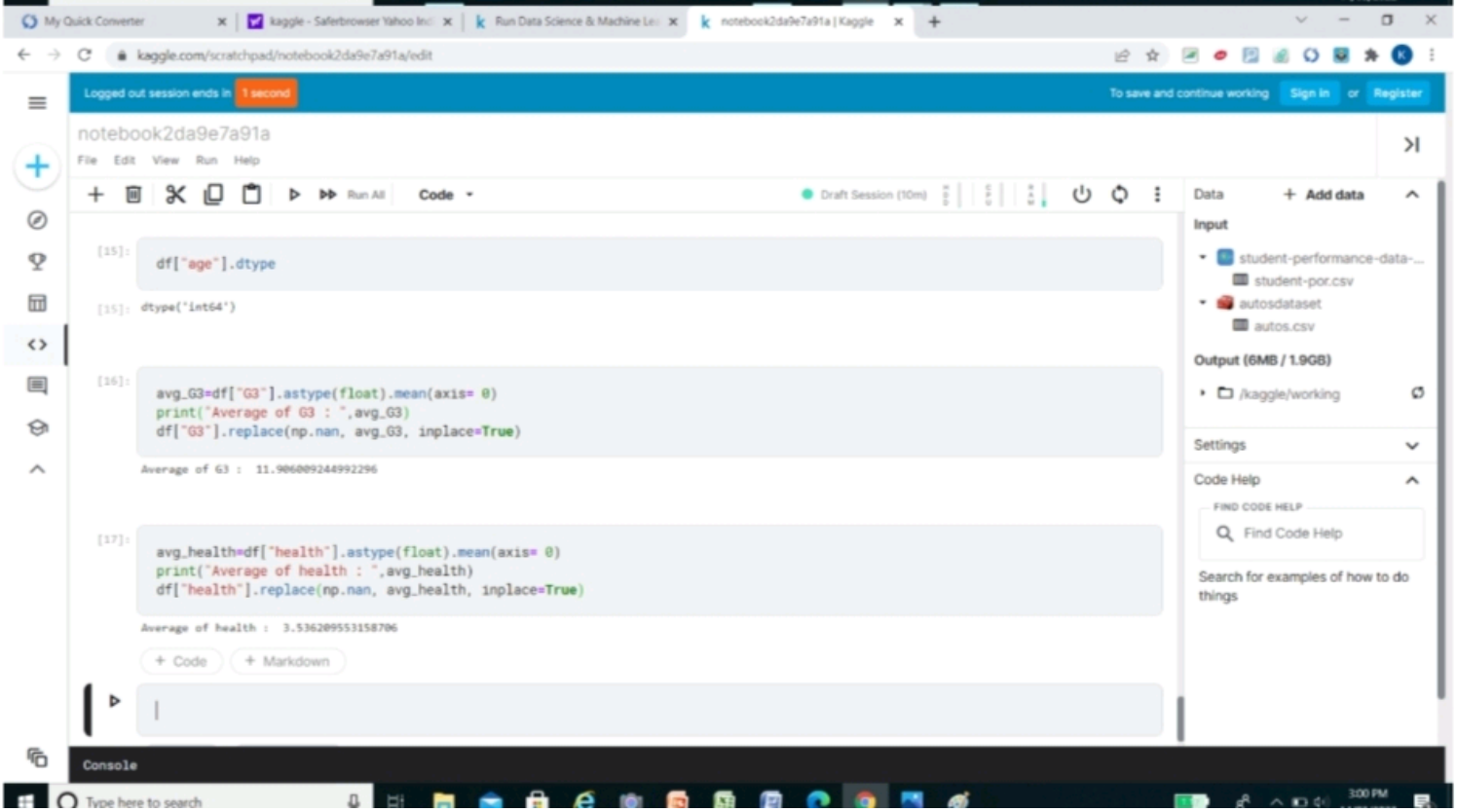

My Quick Converter
kaggle - Saferbrowser Yahoo Inc
Run Data Science & Machine Le
notebook2da9e7a91a | Kaggle
kaggle.com/scratchpad/notebook2da9e7a91a/edit
Logged out session ends in 2 seconds
To save and continue working Sign in or Register
notebook2da9e7a91a Draft saved
File Edit View Run Help
Run All Code Draft Session (5m)
5 rows × 33 columns
[10]: df.mean()
/opt/conda/lib/python3.7/site-packages/ipykernel_launcher.py:1: FutureWarning: Dropping of nuisance columns in DataFrame reductions (with 'numeric_only=None') is deprecated; in a future version this will raise TypeError. Select only valid columns before calling the reduction.
"""Entry point for launching an IPython kernel.
[10]:
age 16.744222
Medu 2.514638
Fedu 2.306626
traveltime 1.568567
studytime 1.930663
failures 0.221880
famrel 3.930663
freetime 3.180277
goout 3.184900
Dalc 1.502311
Walc 2.280431
health 3.536210
absences 3.659476
G1 11.399076
G2 11.570108
G3 11.906009
dtype: float64
Data + Add data
Input
student-performance-data-set
student-por.csv
autosdataset
autos.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Console
Type here to search
2:37 PM
14/03/2022

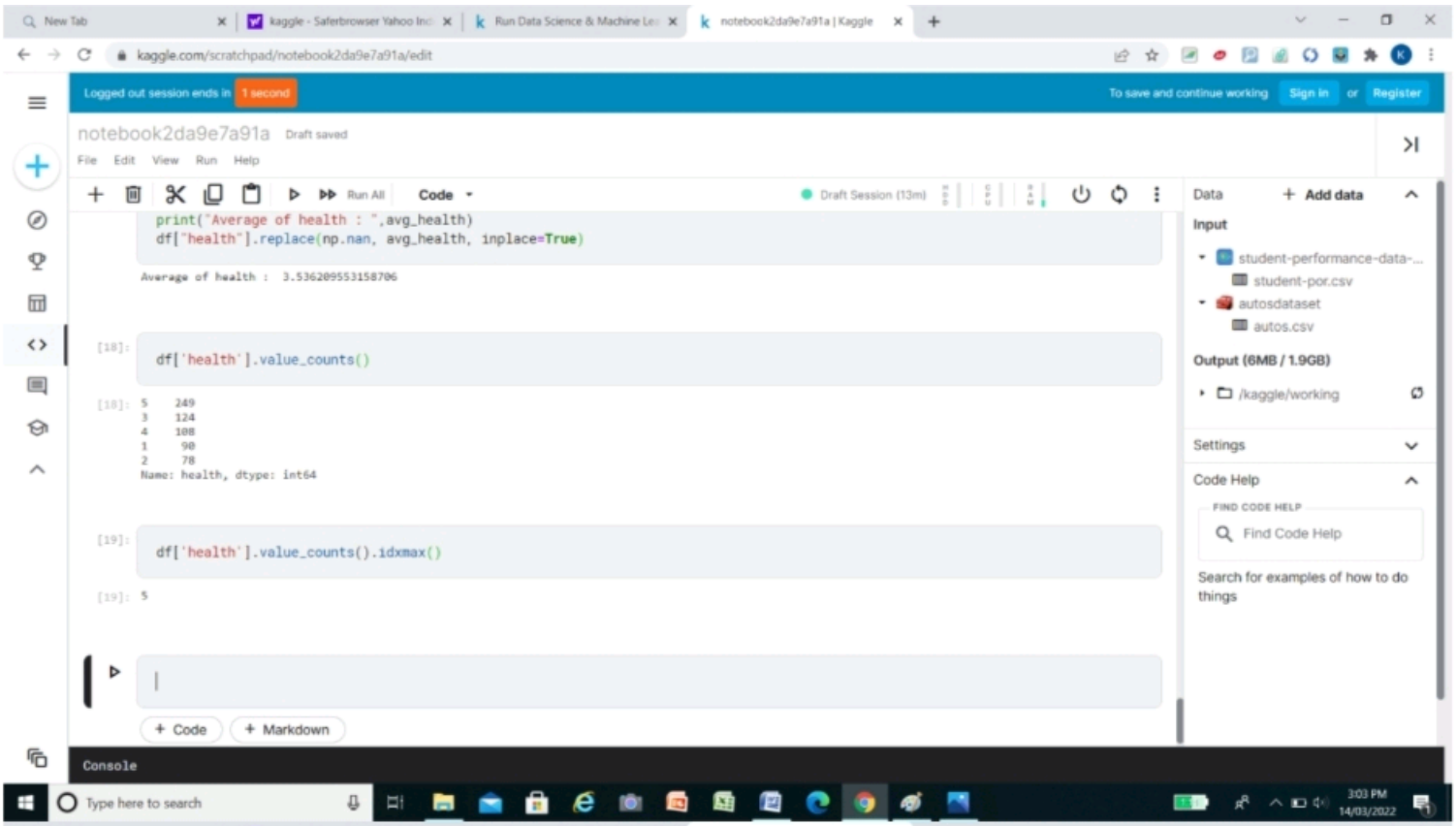

New Tab
kaggle - Saferbrowser Yahoo Inc
Run Data Science & Machine Le
notebook2da9e7a91a | Kaggle
kaggle.com/scratchpad/notebook2da9e7a91a/edit
Logged out session ends in 2 seconds
To save and continue working Sign in or Register
notebook2da9e7a91a
File Edit View Run Help
Run All Code Draft Session (6m)
/opt/conda/lib/python3.7/site-packages/ipykernel_launcher.py:1: FutureWarning: Dropping of nuisance columns in DataFrame reductions (with 'numeric_only=None') is deprecated; in a future version this will raise TypeError. Select only valid columns before calling the reduction.
"""Entry point for launching an IPython kernel.
[10]:
age 16.744222
Medu 2.514638
Fedu 2.306626
traveltime 1.568567
studytime 1.930663
failures 0.221880
famrel 3.930663
freetime 3.180277
goout 3.184900
Dalc 1.502311
Walc 2.280431
health 3.536210
absences 3.659476
G1 11.399076
G2 11.570108
G3 11.906009
dtype: float64
[11]:
avg_age=df['age'].astype("float").mean()
avg_age
[11]: 16.7442218798151
+ Code + Markdown
Data + Add data
Input
student-performance-data-set
student-por.csv
autosdataset
autos.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Console
Type here to search
2:38 PM
14/03/2022

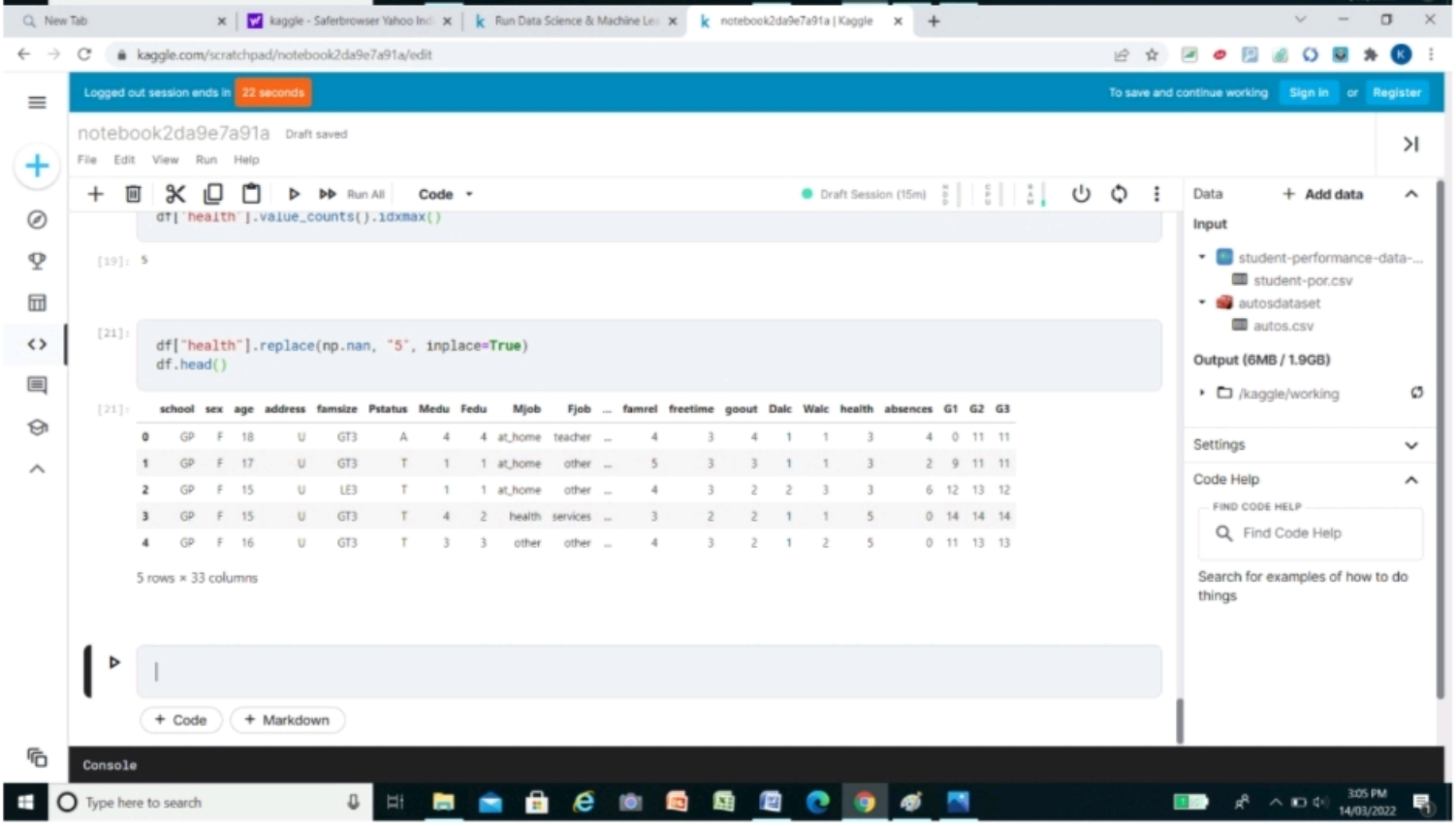

notebook2da9e7a91a | Kaggle
kaggle.com/scratchpad/notebook2da9e7a91a/edit
Logged out session ends in 2 seconds
To save and continue working Sign in or Register
notebook2da9e7a91a
File Edit View Run Help
Run All Code
Draft Session (12m)
avg_age=df['age'].astype("float").mean()
avg_age
[11]: 16.7442218798151
df["age"].replace(np.NaN, avg_age, inplace = True)
df["age"]
[14]: 0 18
1 17
2 15
3 15
4 16
..
644 19
645 18
646 18
647 17
648 18
Name: age, Length: 649, dtype: int64
+ Code + Markdown
Data + Add data
Input
student-performance-data-...
student-por.csv
autosdataset
autos.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Console
Type here to search
2:44 PM
14/03/2022

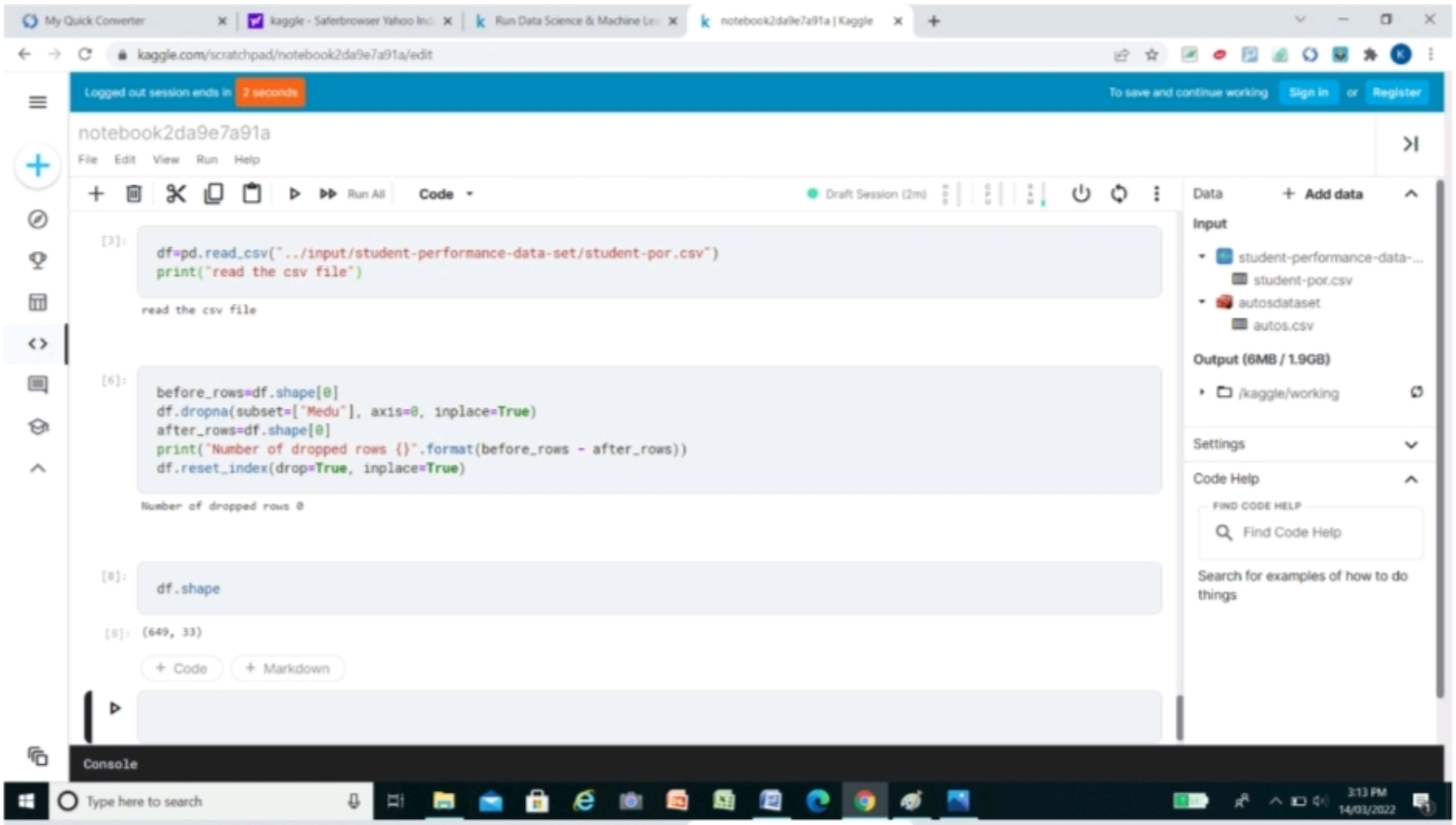

notebook2da9e7a91a | Kaggle
kaggle.com/scratchpad/notebook2da9e7a91a/edit
Logged out session ends in 1 second
To save and continue working Sign in or Register
notebook2da9e7a91a Draft saved
File Edit View Run Help
Run All Code
Draft Session (0m)
1 17
2 15
3 15
4 16
..
644 19
645 18
646 18
647 17
648 18
Name: age, Length: 649, dtype: int64
avg_G1=df["G1"].astype(float).mean()
df["G1"].replace(np.NaN, avg_G1, inplace=True)
df['G1']
[15]: 0 0
1 9
2 12
3 14
4 11
..
644 10
645 15
646 11
647 10
648 10
Name: G1, Length: 649, dtype: int64
Data + Add data
Input
student-performance-data-...
student-por.csv
autosdataset
autos.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Console
Type here to search
2:51 PM
14/03/2022

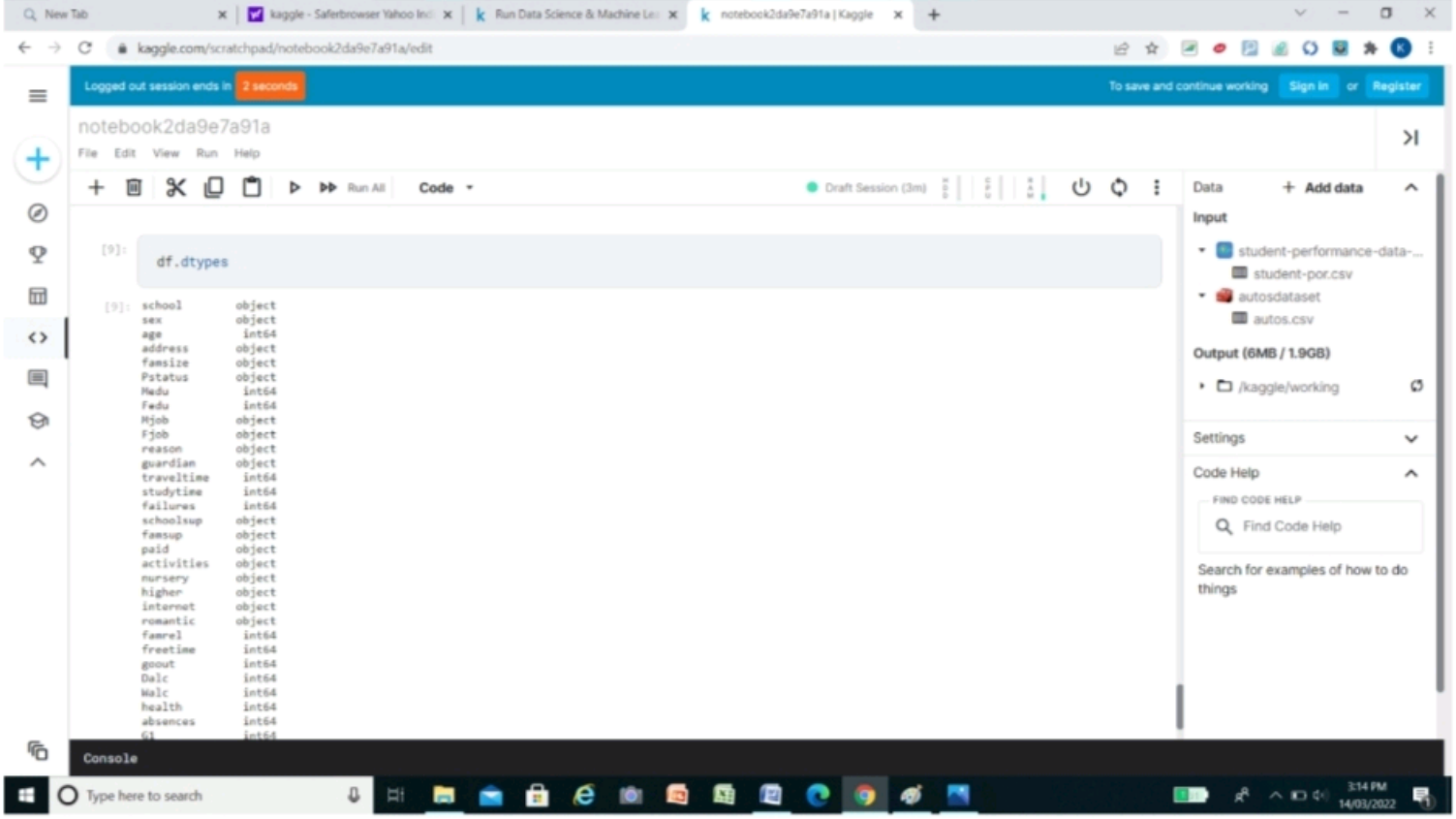

notebook2da9e7a91a
File Edit View Run Help
Logged out session ends in 1 second
To save and continue working Sign In or Register
Run All
Code
Draft Session (3m)
Data
Add data
Input
student-performance-data-...
student-por.csv
autosdataset
autos.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
[5]:
df.head()
[5]:
school sex age address famsize Pstatus Medu Fedu Mjob Fjob ... famrel freetime goout Dalc Walc health absences G1 G2 G3
0 GP F 18 U GT3 A 4 4 at_home teacher ... 4 3 4 1 1 3 4 0 11 11
1 GP F 17 U GT3 T 1 1 at_home other ... 5 3 3 1 1 3 2 9 11 11
2 GP F 15 U LE3 T 1 1 at_home other ... 4 3 2 2 3 3 6 12 13 12
3 GP F 15 U GT3 T 4 2 health services ... 3 2 2 1 1 5 0 14 14 14
4 GP F 16 U GT3 T 3 3 other other ... 4 3 2 1 2 5 0 11 13 13
5 rows × 33 columns
[6]:
avg_G2=df["G2"].astype(float).mean(axis= 0)
print("Average of G2 : ",avg_G2)
df["G2"].replace(np.nan, avg_G2, inplace=True)
Average of G2 : 11.570107858243452
+ Code
+ Markdown
Console
Type here to search
2:53 PM
14/03/2022

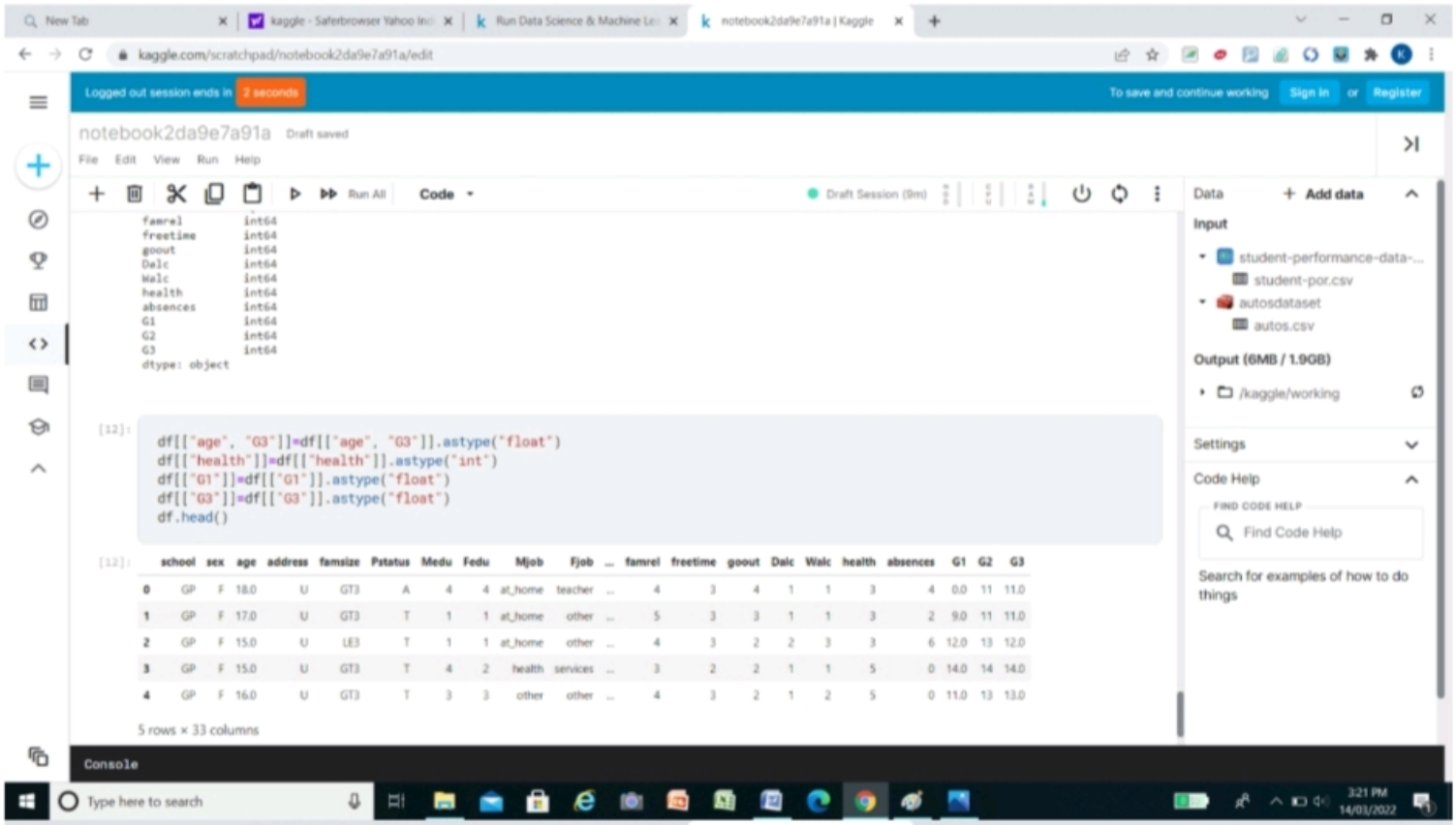

notebook2da9e7a91a
File Edit View Run Help
Logged out session ends in 1 second
To save and continue working Sign In or Register
Run All
Code
Draft Session (10m)
Data
Add data
Input
student-performance-data-...
student-por.csv
autosdataset
autos.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
[15]:
df["age"].dtype
[15]:
dtype('int64')
[16]:
avg_G3=df["G3"].astype(float).mean(axis= 0)
print("Average of G3 : ",avg_G3)
df["G3"].replace(np.nan, avg_G3, inplace=True)
Average of G3 : 11.906009244992296
[17]:
avg_health=df["health"].astype(float).mean(axis= 0)
print("Average of health : ",avg_health)
df["health"].replace(np.nan, avg_health, inplace=True)
Average of health : 3.536209553158706
+ Code
+ Markdown
Console
Type here to search
3:00 PM

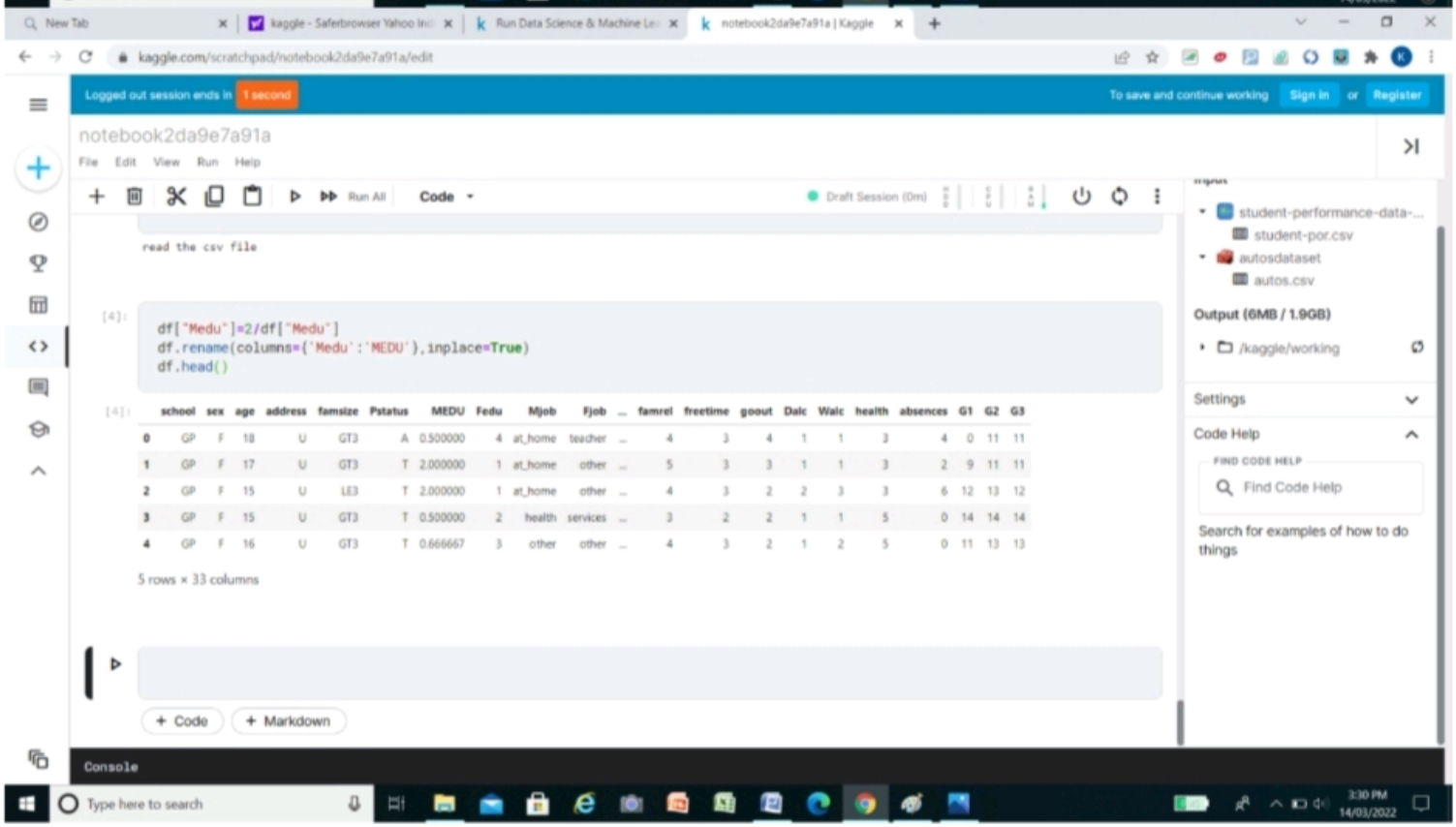

notebook2da9e7a91a | Kaggle
kaggle.com/scratchpad/notebook2da9e7a91a/edit
Logged out session ends in 1 second
To save and continue working Sign in or Register
notebook2da9e7a91a Draft saved
File Edit View Run Help
Run All Code
Draft Session (13m)
print("Average of health : ",avg_health)
df["health"].replace(np.nan, avg_health, inplace=True)
Average of health : 3.536209553158706
[18]: df['health'].value_counts()
[18]: 5 249
3 124
4 108
1 90
2 78
Name: health, dtype: int64
[19]: df['health'].value_counts().idxmax()
[19]: 5
+ Code + Markdown
Console
Data + Add data
Input
student-performance-data-...
student-por.csv
autosdataset
autos.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Type here to search
3:03 PM
14/03/2022

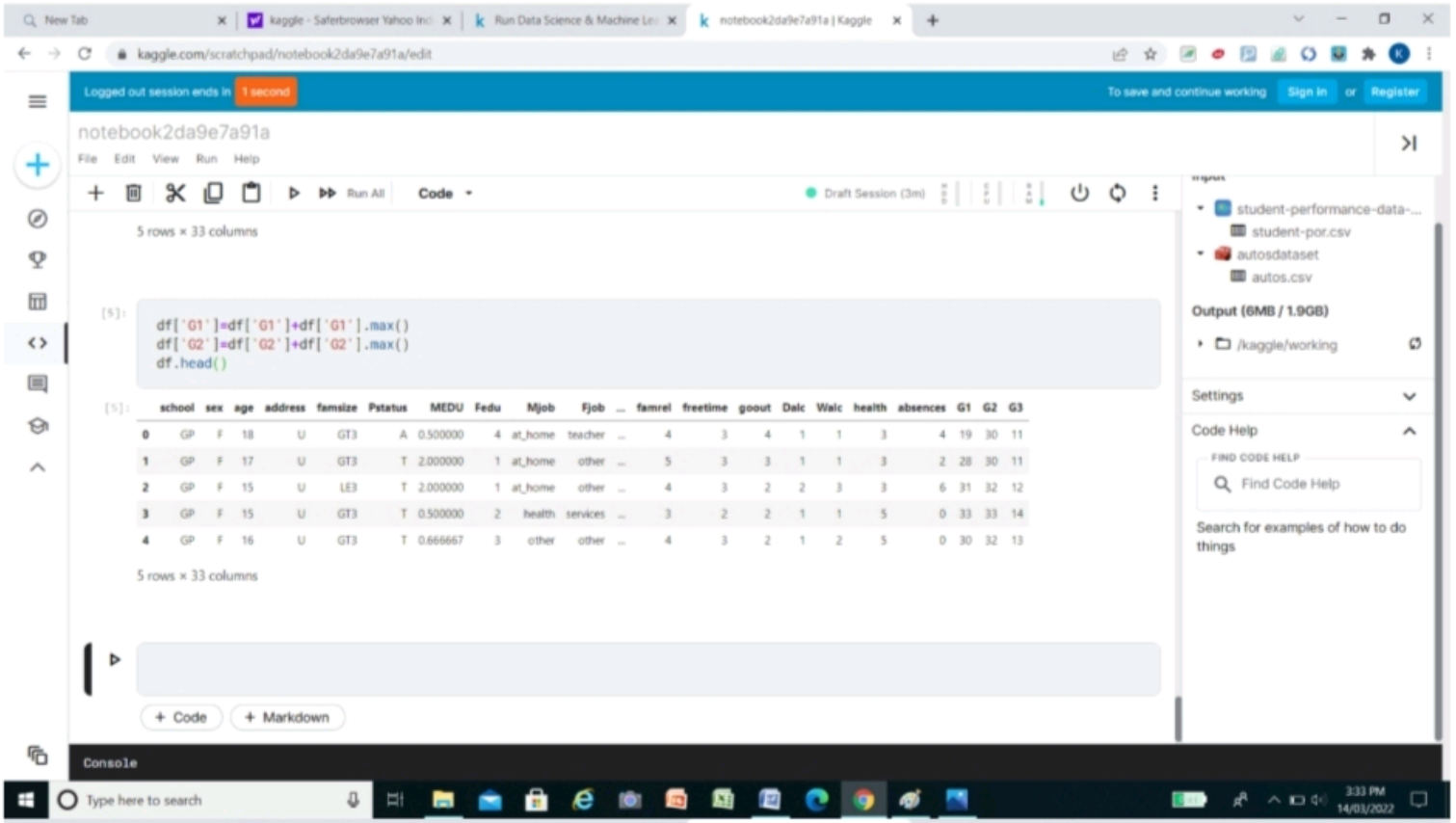

notebook2da9e7a91a | Kaggle
kaggle.com/scratchpad/notebook2da9e7a91a/edit
Logged out session ends in 22 seconds
To save and continue working Sign in or Register
notebook2da9e7a91a Draft saved
File Edit View Run Help
Run All Code
Draft Session (15m)
df['health'].value_counts().idxmax()
[19]: 5
[21]: df["health"].replace(np.nan, "5", inplace=True)
df.head()
school sex age address famsize Pstatus Medu Fedu Mjob Fjob ... famrel freetime goout Dalc Walc health absences G1 G2 G3
0 GP F 18 U GT3 A 4 4 at_home teacher ... 4 3 4 1 1 3 4 0 11 11
1 GP F 17 U GT3 T 1 1 at_home other ... 5 3 3 1 1 3 2 9 11 11
2 GP F 15 U LE3 T 1 1 at_home other ... 4 3 2 2 3 3 6 12 13 12
3 GP F 15 U GT3 T 4 2 health services ... 3 2 2 1 1 5 0 14 14 14
4 GP F 16 U GT3 T 3 3 other other ... 4 3 2 1 2 5 0 11 13 13
5 rows × 33 columns
+ Code + Markdown
Console
Data + Add data
Input
student-performance-data-...
student-por.csv
autosdataset
autos.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Type here to search
3:05 PM
14/03/2022

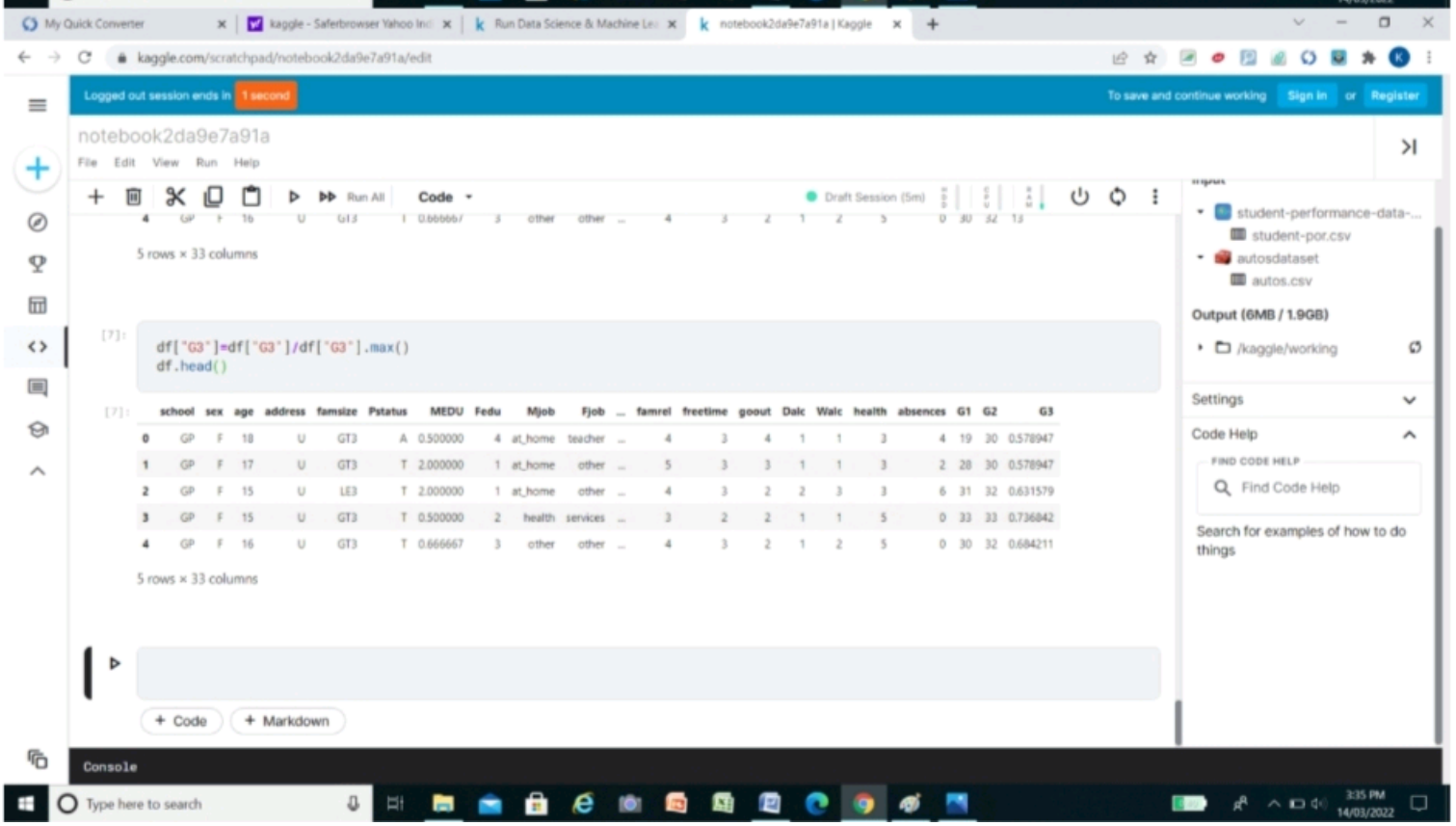

My Quick Converter
kaggle - Saferbrowser Yahoo Ind
Run Data Science & Machine Lea
notebook2da9e7a91a | Kaggle
kaggle.com/scratchpad/notebook2da9e7a91a/edit
Logged out session ends in 2 seconds
To save and continue working Sign In or Register
notebook2da9e7a91a
File Edit View Run Help
Run All
Code
Draft Session (2m)
Data
Add data
Input
student-performance-data-...
student-por.csv
autosdataset
autos.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
[3]:
df=pd.read_csv("../input/student-performance-data-set/student-por.csv")
print("read the csv file")
read the csv file
[6]:
before_rows=df.shape[0]
df.dropna(subset=["Medu"], axis=0, inplace=True)
after_rows=df.shape[0]
print("Number of dropped rows {}".format(before_rows - after_rows))
df.reset_index(drop=True, inplace=True)
Number of dropped rows 0
[8]:
df.shape
[8]: (649, 33)
+ Code
+ Markdown
Console
Type here to search
3:13 PM
14/03/2022

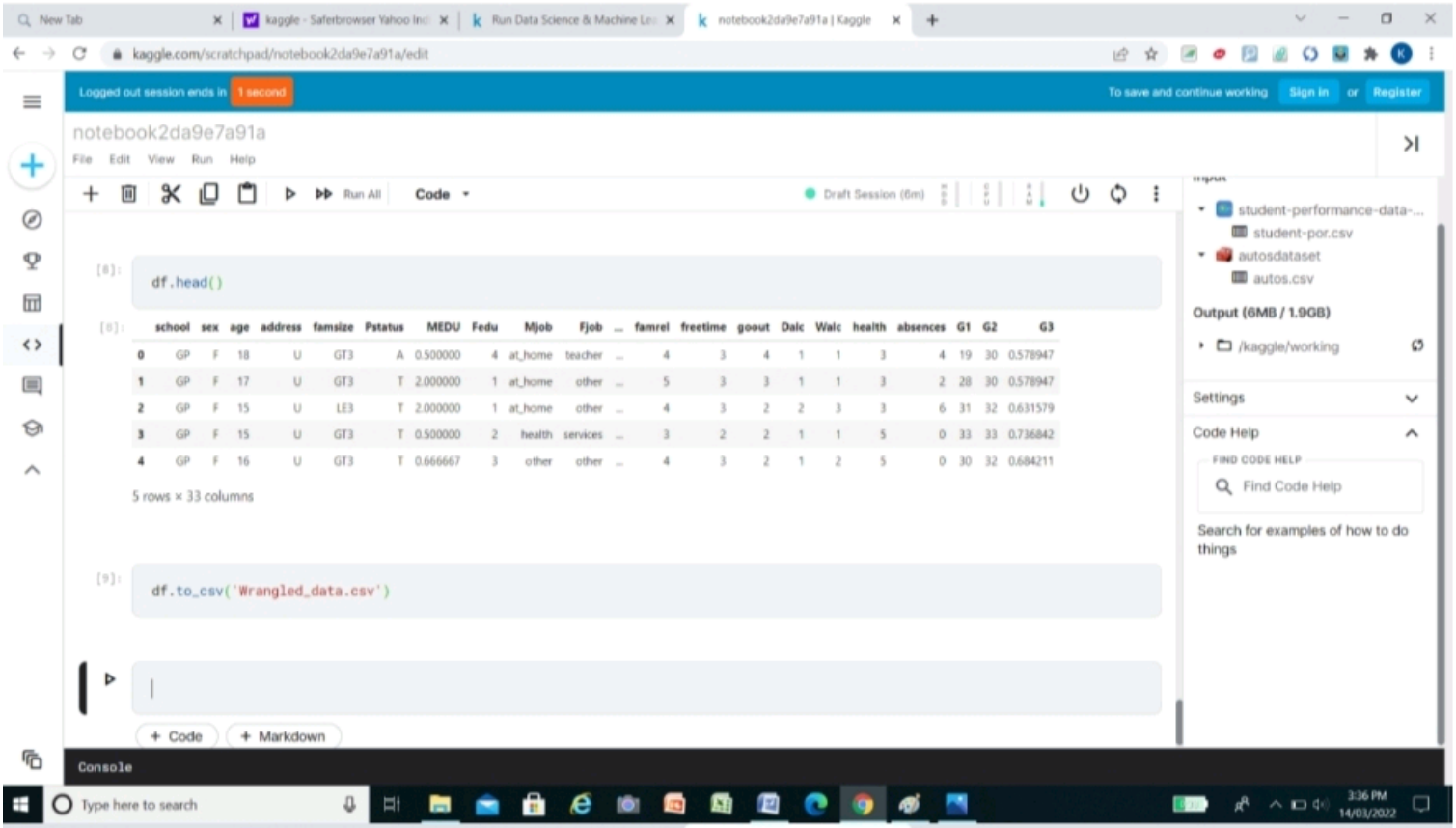

New Tab
kaggle - Saferbrowser Yahoo Ind
Run Data Science & Machine Lea
notebook2da9e7a91a | Kaggle
kaggle.com/scratchpad/notebook2da9e7a91a/edit
Logged out session ends in 2 seconds
To save and continue working Sign In or Register
notebook2da9e7a91a
File Edit View Run Help
Run All
Code
Draft Session (3m)
Data
Add data
Input
student-performance-data-...
student-por.csv
autosdataset
autos.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
[9]:
df.dtypes
[9]:
school object
sex object
age int64
address object
famsize object
Pstatus object
Medu int64
Fedu int64
Mjob object
Fjob object
reason object
guardian object
traveltime int64
studytime int64
failures int64
schoolsup object
famsup object
paid object
activities object
nursery object
higher object
internet object
romantic object
famrel int64
freetime int64
goout int64
Dalc int64
Walc int64
health int64
absences int64
G1 int64
Console
Type here to search
3:14 PM
14/03/2022

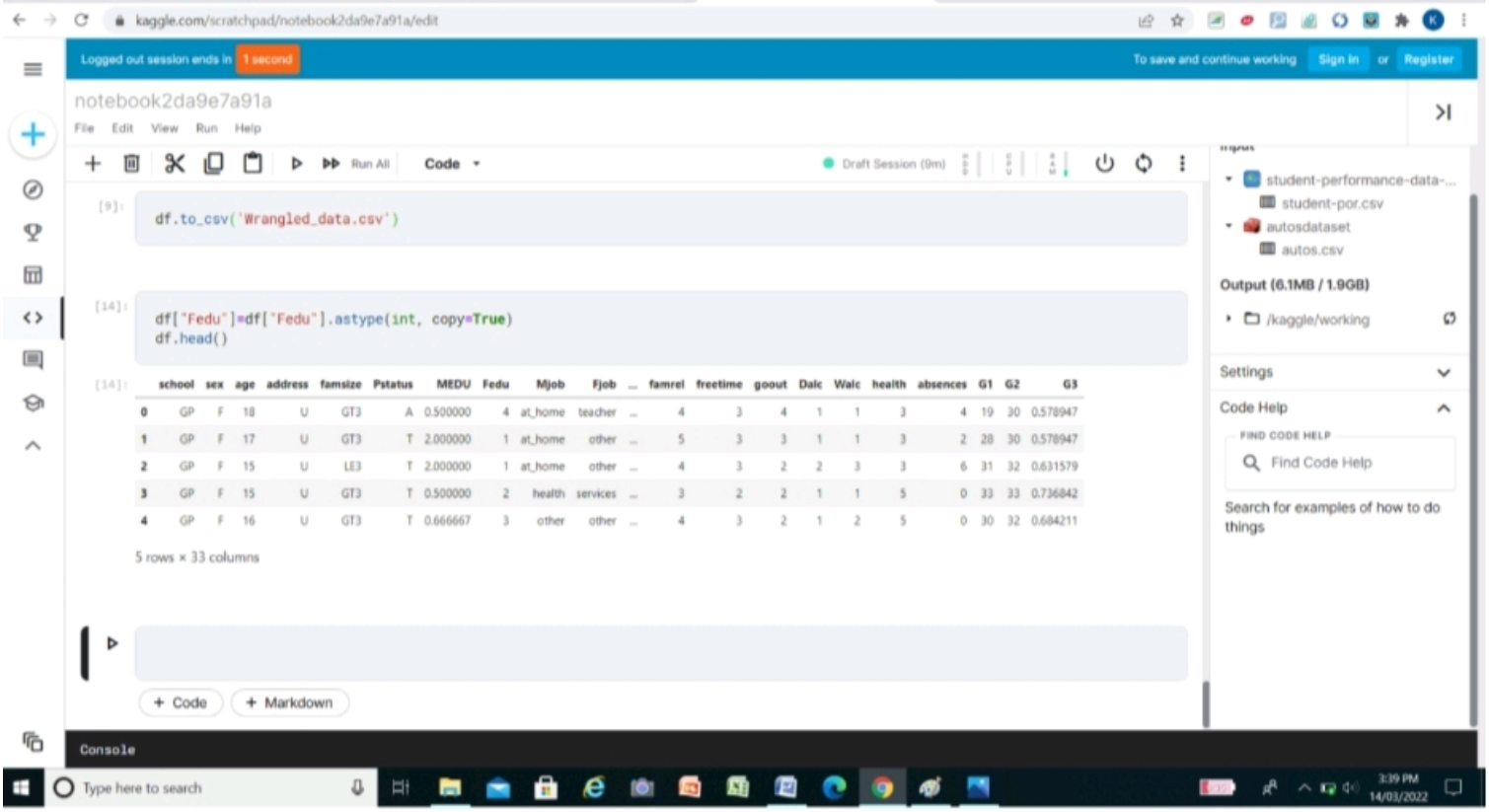

notebook2da9e7a91a | Kaggle
kaggle.com/scratchpad/notebook2da9e7a91a/edit
Logged out session ends in 2 seconds
To save and continue working Sign in or Register
notebook2da9e7a91a Draft saved
File Edit View Run Help
Run All Code Draft Session (0m)
famrel int64
freetime int64
goout int64
Dalc int64
Walc int64
health int64
absences int64
G1 int64
G2 int64
G3 int64
dtype: object
[12]:
df[["age", "G3"]]=df[["age", "G3"]].astype("float")
df[["health"]]=df[["health"]].astype("int")
df[["G1"]]=df[["G1"]].astype("float")
df[["G3"]]=df[["G3"]].astype("float")
df.head()
[12]:
school sex age address famsize Pstatus Medu Fedu Mjob Fjob ... famrel freetime goout Dalc Walc health absences G1 G2 G3
0 GP F 18.0 U GT3 A 4 4 at_home teacher ... 4 3 4 1 1 3 4 0.0 11 11.0
1 GP F 17.0 U GT3 T 1 1 at_home other ... 5 3 3 1 1 3 2 9.0 11 11.0
2 GP F 15.0 U LE3 T 1 1 at_home other ... 4 3 2 2 3 3 6 12.0 13 12.0
3 GP F 15.0 U GT3 T 4 2 health services ... 3 2 2 1 1 5 0 14.0 14 14.0
4 GP F 16.0 U GT3 T 3 3 other other ... 4 3 2 1 2 5 0 11.0 13 13.0
5 rows × 33 columns
Data Add data
Input
student-performance-data-...
student-por.csv
autosdataset
autos.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Console
Type here to search
3:21 PM 14/03/2022


notebook2da9e7a91a | Kaggle
kaggle.com/scratchpad/notebook2da9e7a91a/edit
Logged out session ends in 1 second
To save and continue working Sign in or Register
notebook2da9e7a91a
File Edit View Run Help
Run All Code Draft Session (0m)
read the csv file
[4]:
df["Medu"]=2/df["Medu"]
df.rename(columns={'Medu':'MEDU'},inplace=True)
df.head()
[4]:
school sex age address famsize Pstatus MEDU Fedu Mjob Fjob ... famrel freetime goout Dalc Walc health absences G1 G2 G3
0 GP F 18 U GT3 A 0.500000 4 at_home teacher ... 4 3 4 1 1 3 4 0 11 11
1 GP F 17 U GT3 T 2.000000 1 at_home other ... 5 3 3 1 1 3 2 9 11 11
2 GP F 15 U LE3 T 2.000000 1 at_home other ... 4 3 2 2 3 3 6 12 13 12
3 GP F 15 U GT3 T 0.500000 2 health services ... 3 2 2 1 1 5 0 14 14 14
4 GP F 16 U GT3 T 0.666667 3 other other ... 4 3 2 1 2 5 0 11 13 13
5 rows × 33 columns
+ Code + Markdown
Data Add data
Input
student-performance-data-...
student-por.csv
autosdataset
autos.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Console
Type here to search
3:30 PM 14/03/2022


kaggle.com/scratchpad/notebook2da9e7a91a/edit
Logged out session ends in 1 second
To save and continue working Sign In or Register
notebook2da9e7a91a
File Edit View Run Help
Run All
Code
Draft Session (3m)
5 rows × 33 columns
[5]:
df['G1']=df['G1']+df['G1'].max()
df['G2']=df['G2']+df['G2'].max()
df.head()
[5]:
school sex age address famsize Pstatus MEDU Fedu Mjob Fjob ... famrel freetime goout Dalc Walc health absences G1 G2 G3
0 GP F 18 U GT3 A 0.500000 4 at_home teacher ... 4 3 4 1 1 3 4 19 30 11
1 GP F 17 U GT3 T 2.000000 1 at_home other ... 5 3 3 1 1 3 2 28 30 11
2 GP F 15 U LE3 T 2.000000 1 at_home other ... 4 3 2 2 3 3 6 31 32 12
3 GP F 15 U GT3 T 0.500000 2 health services ... 3 2 2 1 1 5 0 33 33 14
4 GP F 16 U GT3 T 0.666667 3 other other ... 4 3 2 1 2 5 0 30 32 13
5 rows × 33 columns
+ Code
+ Markdown
Console
student-performance-data-...
student-por.csv
autosdataset
autos.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Type here to search
3:33 PM
14/03/2022


kaggle.com/scratchpad/notebook2da9e7a91a/edit
Logged out session ends in 1 second
To save and continue working Sign In or Register
notebook2da9e7a91a
File Edit View Run Help
Run All
Code
Draft Session (5m)
5 rows × 33 columns
[7]:
df["G3"]=df["G3"]/df["G3"].max()
df.head()
[7]:
school sex age address famsize Pstatus MEDU Fedu Mjob Fjob ... famrel freetime goout Dalc Walc health absences G1 G2 G3
0 GP F 18 U GT3 A 0.500000 4 at_home teacher ... 4 3 4 1 1 3 4 19 30 0.578947
1 GP F 17 U GT3 T 2.000000 1 at_home other ... 5 3 3 1 1 3 2 28 30 0.578947
2 GP F 15 U LE3 T 2.000000 1 at_home other ... 4 3 2 2 3 3 6 31 32 0.631579
3 GP F 15 U GT3 T 0.500000 2 health services ... 3 2 2 1 1 5 0 33 33 0.736842
4 GP F 16 U GT3 T 0.666667 3 other other ... 4 3 2 1 2 5 0 30 32 0.684211
5 rows × 33 columns
+ Code
+ Markdown
Console
student-performance-data-...
student-por.csv
autosdataset
autos.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Type here to search
3:35 PM
14/03/2022


notebook2da9e7a91a
File Edit View Run Help
Run All
Code
Draft Session (6m)
[8]:
df.head()
[8]:
school sex age address famsize Pstatus MEDU Fedu Mjob Fjob ... famrel freetime goout Dalc Walc health absences G1 G2 G3
0 GP F 18 U GT3 A 0.500000 4 at_home teacher ... 4 3 4 1 1 3 4 19 30 0.578947
1 GP F 17 U GT3 T 2.000000 1 at_home other ... 5 3 3 1 1 3 2 28 30 0.578947
2 GP F 15 U LE3 T 2.000000 1 at_home other ... 4 3 2 2 3 3 6 31 32 0.631579
3 GP F 15 U GT3 T 0.500000 2 health services ... 3 2 2 1 1 5 0 33 33 0.736842
4 GP F 16 U GT3 T 0.666667 3 other other ... 4 3 2 1 2 5 0 30 32 0.684211
5 rows × 33 columns
[9]:
df.to_csv('Wrangled_data.csv')
+ Code + Markdown
Console
student-performance-data-...
student-por.csv
autosdataset
autos.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things


notebook2da9e7a91a
File Edit View Run Help
Run All
Code
Draft Session (9m)
[9]:
df.to_csv('Wrangled_data.csv')
[14]:
df["Fedu"]=df["Fedu"].astype(int, copy=True)
df.head()
[14]:
school sex age address famsize Pstatus MEDU Fedu Mjob Fjob ... famrel freetime goout Dalc Walc health absences G1 G2 G3
0 GP F 18 U GT3 A 0.500000 4 at_home teacher ... 4 3 4 1 1 3 4 19 30 0.578947
1 GP F 17 U GT3 T 2.000000 1 at_home other ... 5 3 3 1 1 3 2 28 30 0.578947
2 GP F 15 U LE3 T 2.000000 1 at_home other ... 4 3 2 2 3 3 6 31 32 0.631579
3 GP F 15 U GT3 T 0.500000 2 health services ... 3 2 2 1 1 5 0 33 33 0.736842
4 GP F 16 U GT3 T 0.666667 3 other other ... 4 3 2 1 2 5 0 30 32 0.684211
5 rows × 33 columns
+ Code + Markdown
Console
student-performance-data-...
student-por.csv
autosdataset
autos.csv
Output (6.1MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things

New Tab | kaggle - Saferbrowser Yahoo Ind | Run Data Science & Machine Le | notebook2da9e7a91a | Kaggle

kaggle.com/scratchpad/notebook2da9e7a91a/edit

Logged out session ends in 1 second — To save and continue working Sign In or Register

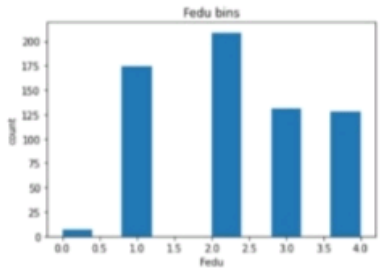
notebook2da9e7a91a Draft saved

File Edit View Run Help

Run All | Code | Draft Session (13m)

[15]:
```
%matplotlib inline
import matplotlib.pyplot as plt
plt.hist(df["Fedu"])
plt.xlabel("Fedu")
plt.ylabel("count")
plt.title("Fedu bins")
```

[15]:
```
Text(0.5, 1.0, 'Fedu bins')
```



student-performance-data-...
student-por.csv
autosdataset
autos.csv

Output (6.1MB / 1.9GB)

/kaggle/working

Settings

Code Help

FIND CODE HELP

Find Code Help

Search for examples of how to do things

Console

Type here to search — 3:43 PM 14/03/2022

New Tab | kaggle - Saferbrowser Yahoo Ind | Run Data Science & Machine Le | notebook2da9e7a91a | Kaggle

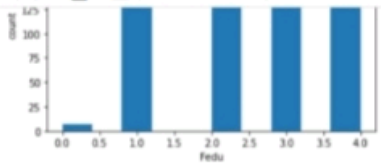
kaggle.com/scratchpad/notebook2da9e7a91a/edit

Logged out session ends in 2 seconds — To save and continue working Sign In or Register

notebook2da9e7a91a

File Edit View Run Help

Run All | Code | Draft Session (0m)



[16]:
```
bins = np.linspace(min(df["Fedu"]),max(df["Fedu"]),4)
bins
```

[16]:
```
array([0.        , 1.33333333, 2.66666667, 4.        ])
```

[1]:
```
group_names = ['Low', 'Medium', 'High']
```

+ Code + Markdown

student-performance-data-...
student-por.csv
autosdataset
autos.csv

Output (6MB / 1.9GB)

/kaggle/working

Settings

Code Help

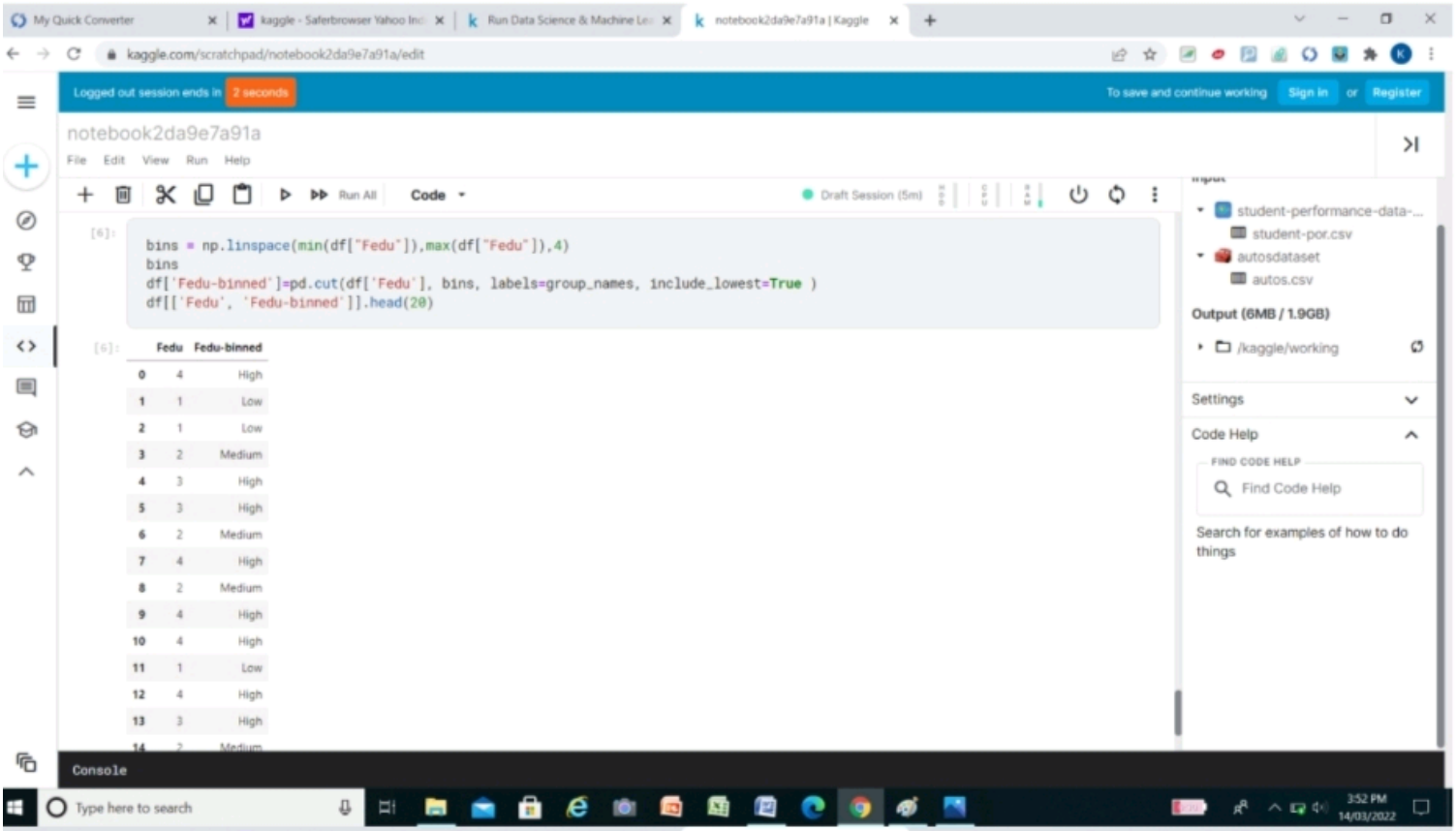
FIND CODE HELP

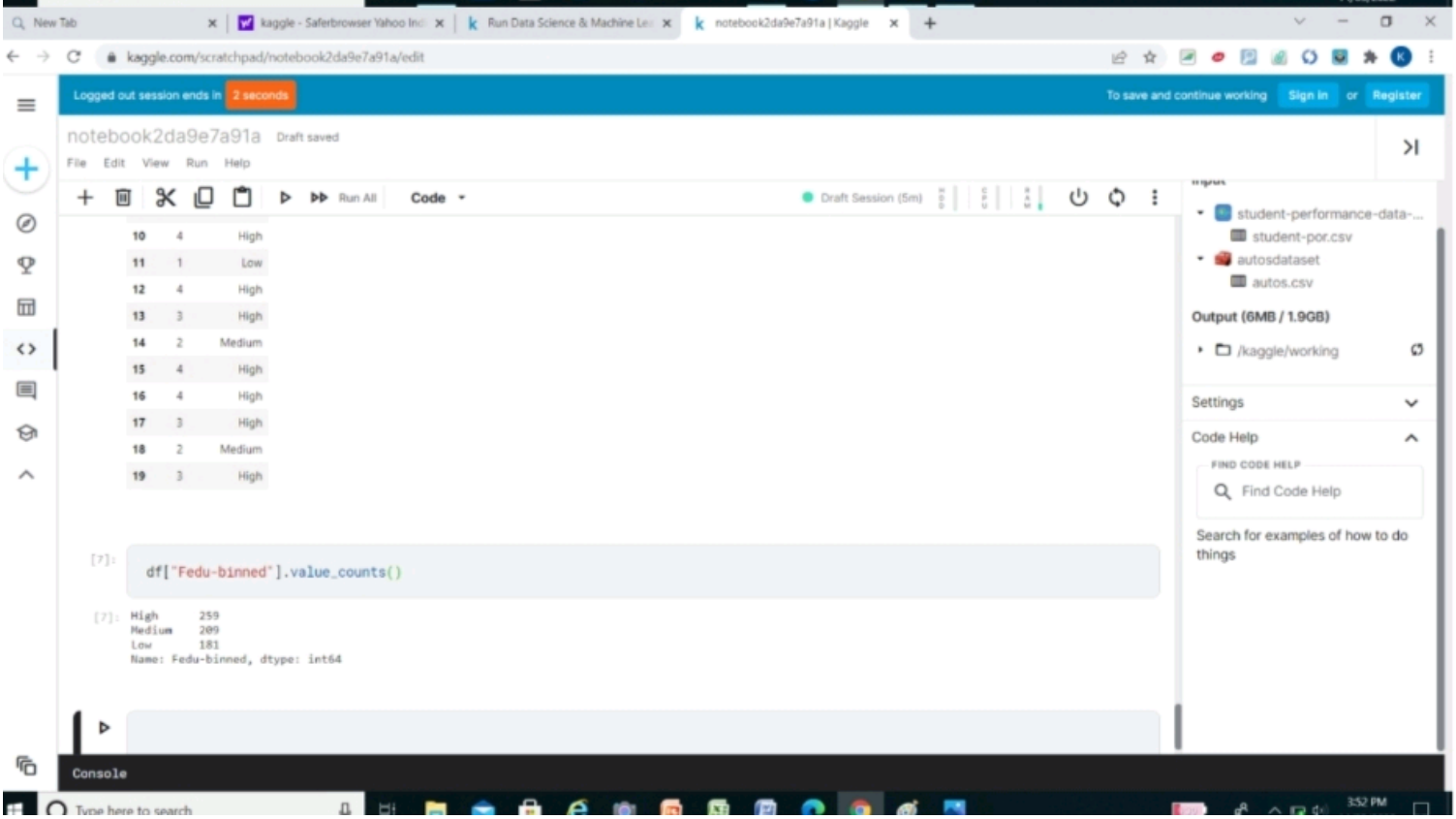
Find Code Help

Search for examples of how to do things

Console

Type here to search — 3:47 PM 14/03/2022


kaggle.com/scratchpad/notebook2da9e7a91a/edit
Logged out session ends in 2 seconds
To save and continue working Sign in or Register
notebook2da9e7a91a
File Edit View Run Help
Run All
Code
Draft Session (5m)
[6]:
bins = np.linspace(min(df["Fedu"]),max(df["Fedu"]),4)
bins
df['Fedu-binned']=pd.cut(df['Fedu'], bins, labels=group_names, include_lowest=True)
df[['Fedu', 'Fedu-binned']].head(20)
[6]:
Fedu Fedu-binned
0 4 High
1 1 Low
2 1 Low
3 2 Medium
4 3 High
5 3 High
6 2 Medium
7 4 High
8 2 Medium
9 4 High
10 4 High
11 1 Low
12 4 High
13 3 High
14 2 Medium
student-performance-data-...
student-por.csv
autosdataset
autos.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Console
Type here to search
3:52 PM
14/03/2022


kaggle.com/scratchpad/notebook2da9e7a91a/edit
Logged out session ends in 2 seconds
To save and continue working Sign in or Register
notebook2da9e7a91a Draft saved
File Edit View Run Help
Run All
Code
Draft Session (5m)
10 4 High
11 1 Low
12 4 High
13 3 High
14 2 Medium
15 4 High
16 4 High
17 3 High
18 2 Medium
19 3 High
[7]:
df["Fedu-binned"].value_counts()
[7]:
High 259
Medium 209
Low 181
Name: Fedu-binned, dtype: int64
student-performance-data-...
student-por.csv
autosdataset
autos.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Console
Type here to search
3:52 PM

My Quick Converter | kaggle - Saferbrowser Yahoo Inc | Run Data Science & Machine Le | notebook2da9e7a91a | Kaggle

kaggle.com/scratchpad/notebook2da9e7a91a/edit

Logged out session ends in 2 seconds — To save and continue working Sign in or Register

notebook2da9e7a91a Draft saved

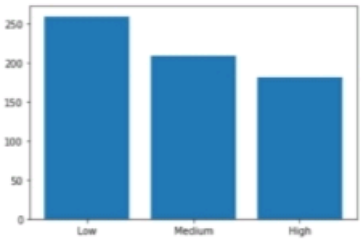
File Edit View Run Help

Run All | Markdown | Draft Session (8m)

```
Name: Fedu-binned, dtype: int64
```

[8]:
```
%matplotlib inline
import matplotlib.pyplot as plt
plt.bar(group_names, df["Fedu-binned"].value_counts())
```

[8]:
```
<BarContainer object of 3 artists>
```



%

+ Code + Markdown

Console

student-performance-data-...
student-por.csv
autosdataset
autos.csv

Output (6MB / 1.9GB)

/kaggle/working

Settings

Code Help

FIND CODE HELP

Find Code Help

Search for examples of how to do things

Type here to search — 3:55 PM 14/03/2022

My Quick Converter | kaggle - Saferbrowser Yahoo Inc | Run Data Science & Machine Le | notebook2da9e7a91a | Kaggle

kaggle.com/scratchpad/notebook2da9e7a91a/edit

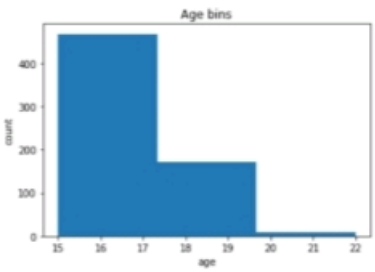
Logged out session ends in 2 seconds — To save and continue working Sign in or Register

notebook2da9e7a91a

File Edit View Run Help

Run All | Markdown | Draft Session (11m)

[9]:
```
plt.hist(df["age"], bins = 3)

plt.xlabel("age")
plt.ylabel("count")
plt.title("Age bins")
plt.show()
```



%

Console

student-performance-data-...
student-por.csv
autosdataset
autos.csv

Output (6MB / 1.9GB)

/kaggle/working

Settings

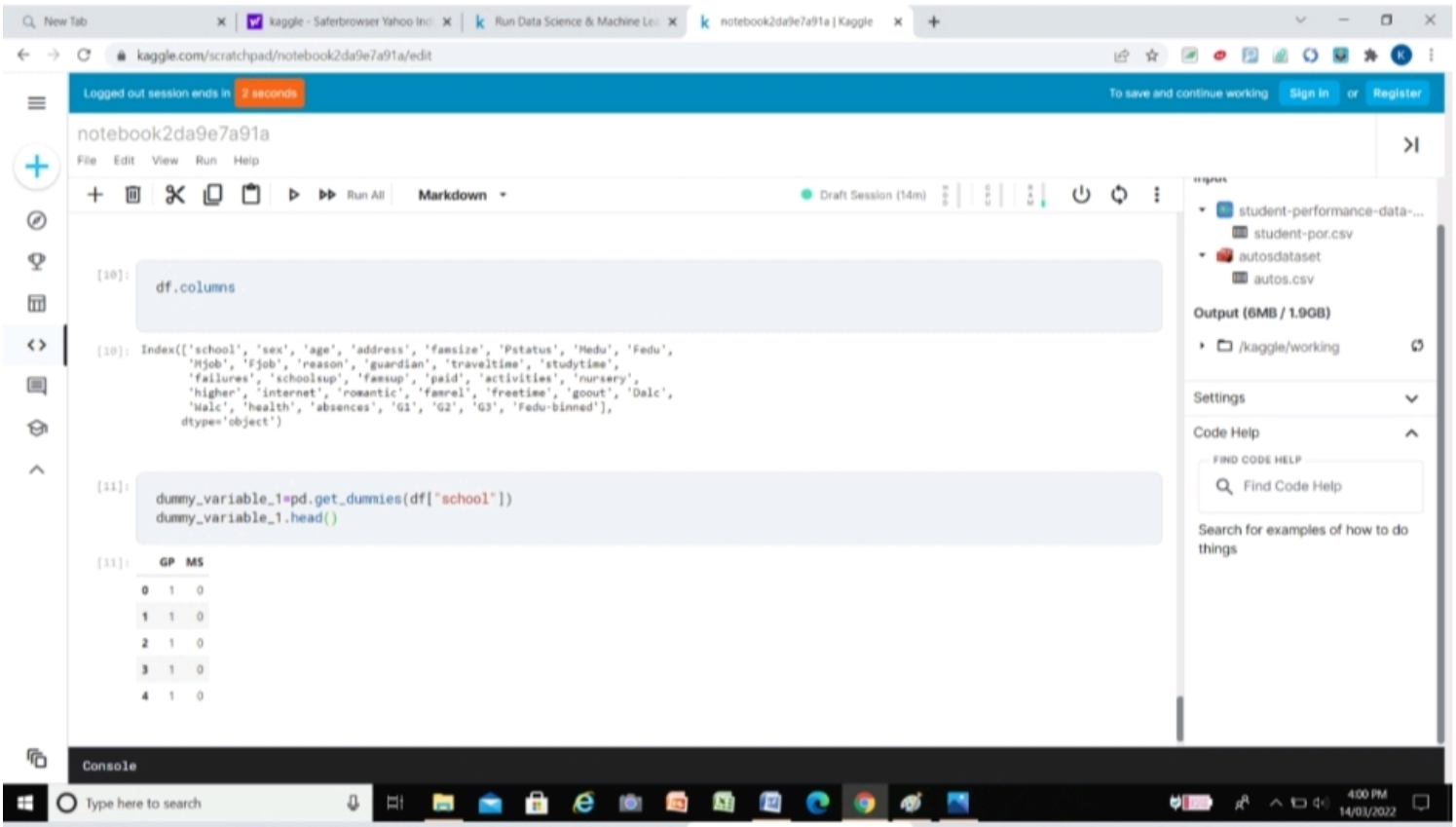
Code Help

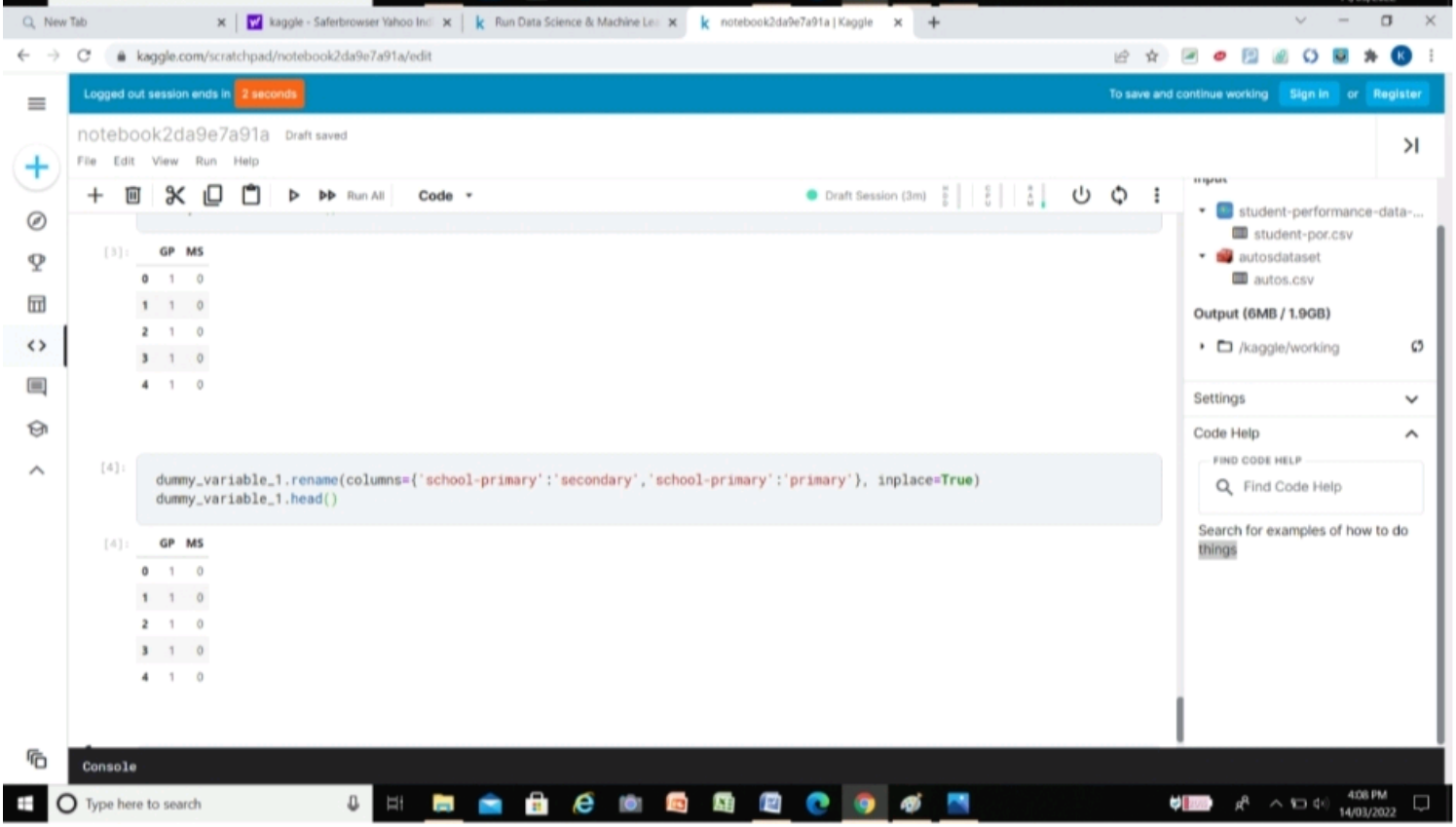
FIND CODE HELP

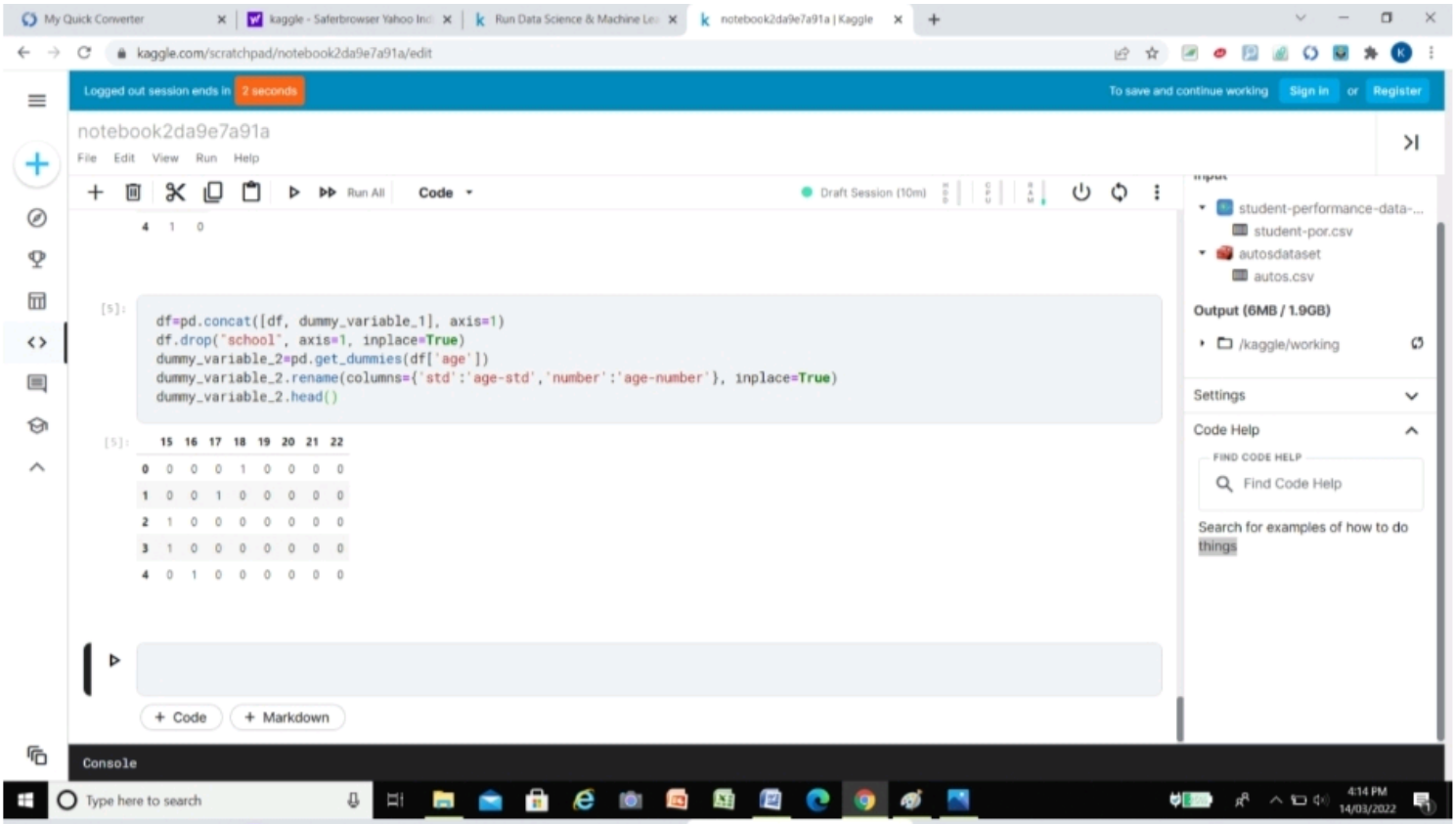
Find Code Help

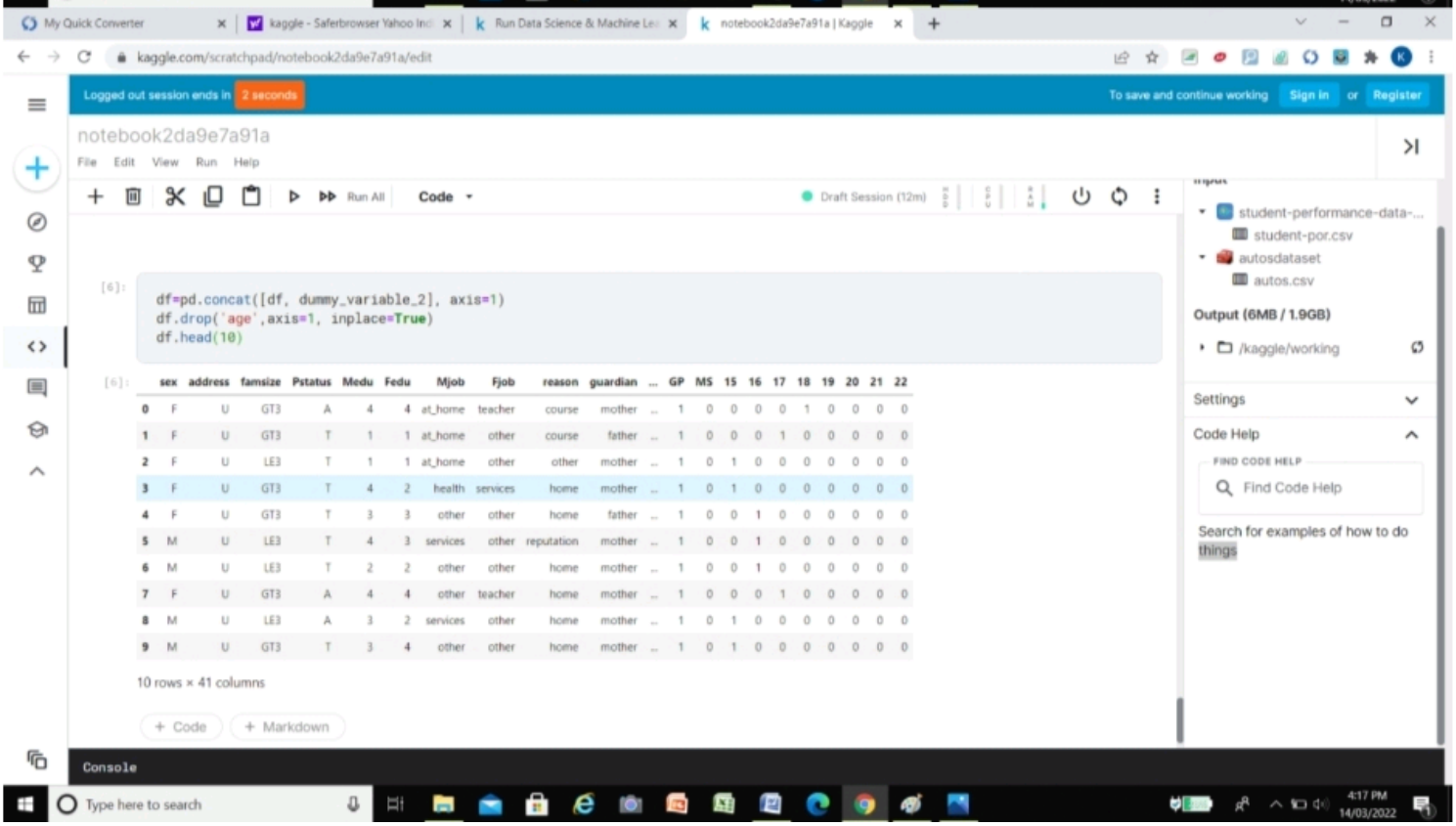
Search for examples of how to do things

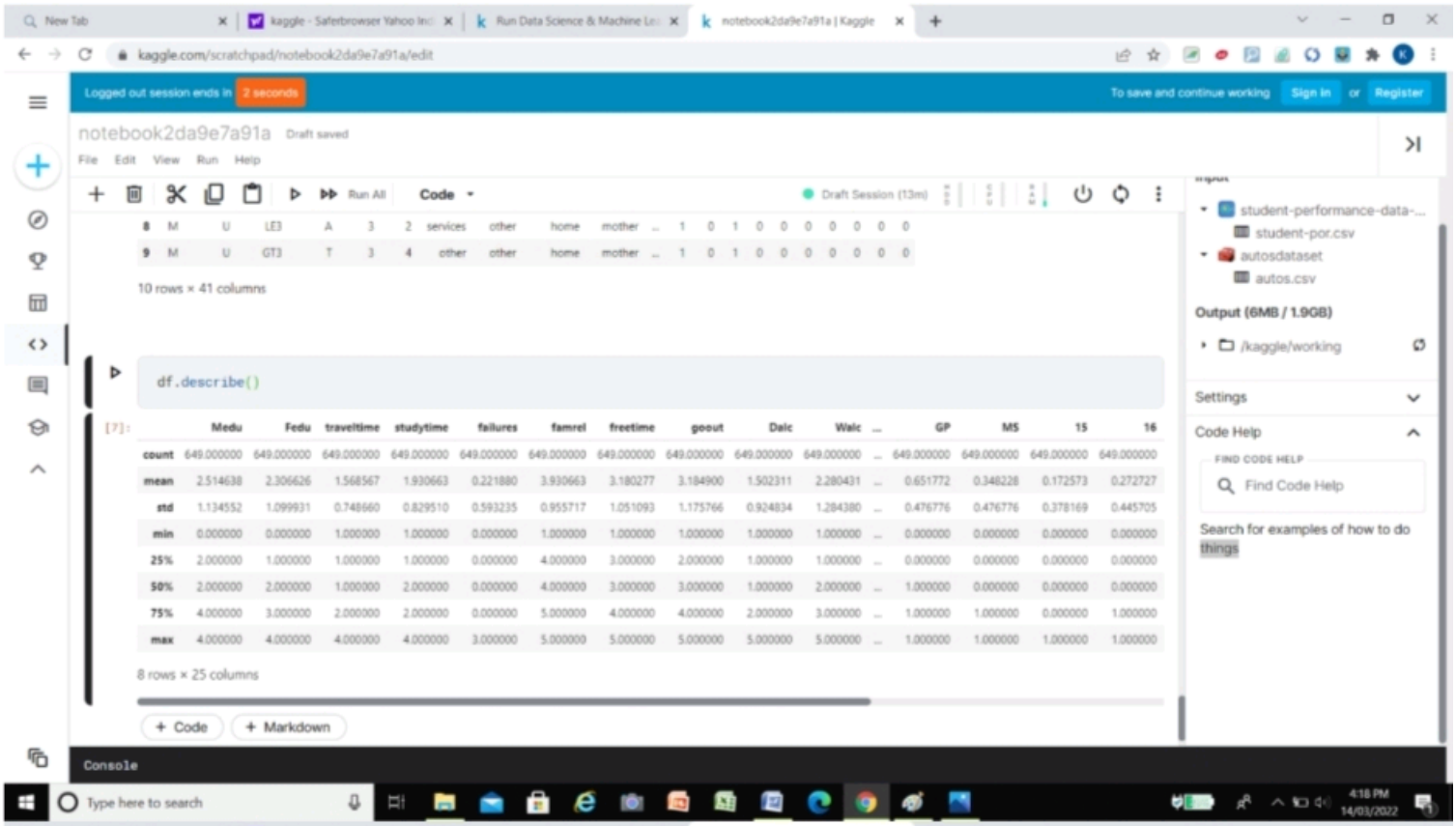
Type here to search — 3:58 PM 14/03/2022

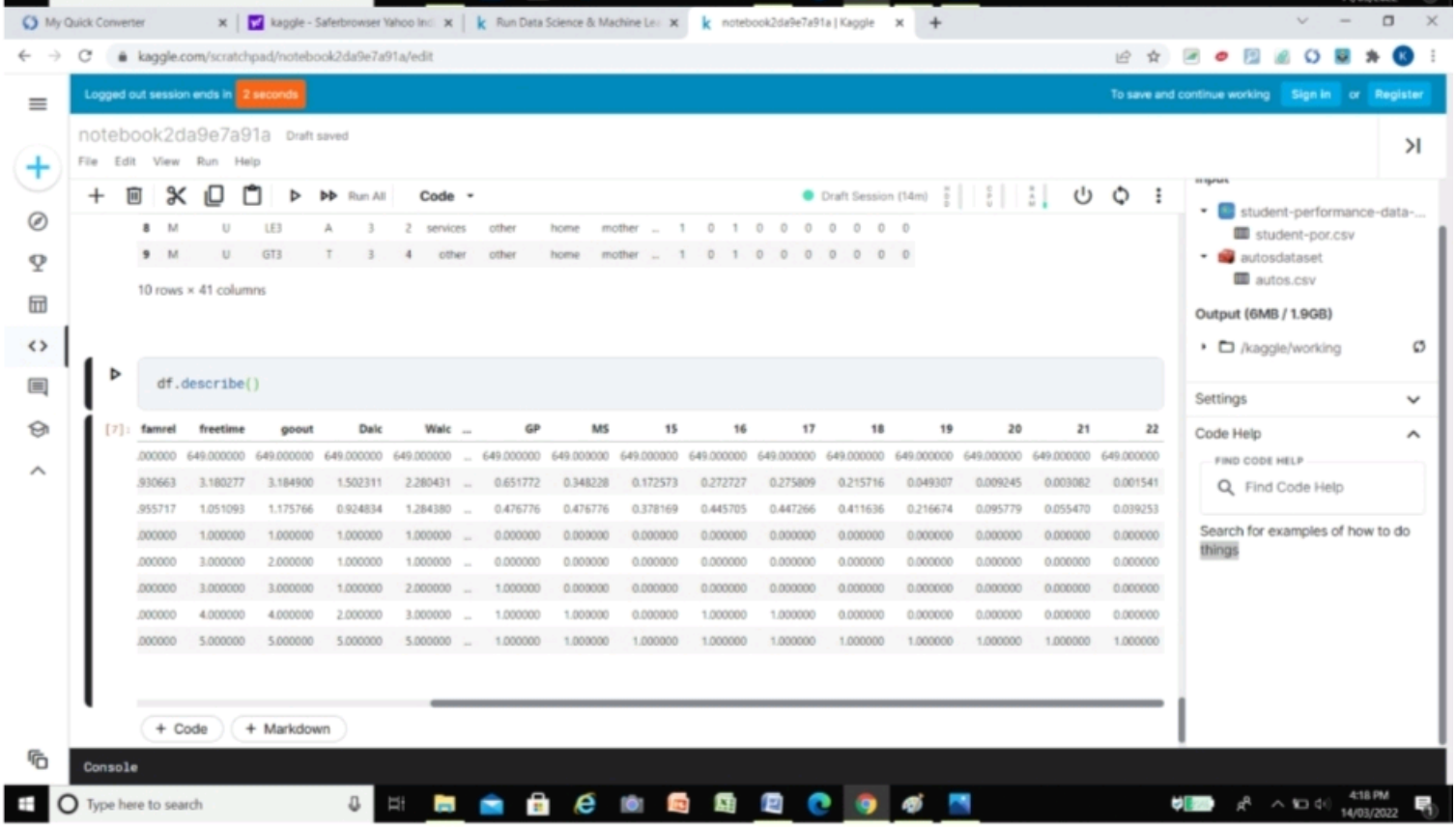

kaggle.com/scratchpad/notebook2da9e7a91a/edit
Logged out session ends in 2 seconds
To save and continue working Sign In or Register
notebook2da9e7a91a
File Edit View Run Help
Run All Markdown
Draft Session (14m)
[10]: df.columns
[10]: Index(['school', 'sex', 'age', 'address', 'famsize', 'Pstatus', 'Medu', 'Fedu',
'Mjob', 'Fjob', 'reason', 'guardian', 'traveltime', 'studytime',
'failures', 'schoolsup', 'famsup', 'paid', 'activities', 'nursery',
'higher', 'internet', 'romantic', 'famrel', 'freetime', 'goout', 'Dalc',
'Walc', 'health', 'absences', 'G1', 'G2', 'G3', 'Fedu-binned'],
dtype='object')
[11]: dummy_variable_1=pd.get_dummies(df["school"])
dummy_variable_1.head()
[11]:
GP MS
0 1 0
1 1 0
2 1 0
3 1 0
4 1 0
student-performance-data-...
student-por.csv
autosdataset
autos.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Console
Type here to search
4:00 PM
14/03/2022


kaggle.com/scratchpad/notebook2da9e7a91a/edit
Logged out session ends in 2 seconds
To save and continue working Sign In or Register
notebook2da9e7a91a Draft saved
File Edit View Run Help
Run All Code
Draft Session (3m)
[3]:
GP MS
0 1 0
1 1 0
2 1 0
3 1 0
4 1 0
[4]: dummy_variable_1.rename(columns={'school-primary':'secondary','school-primary':'primary'}, inplace=True)
dummy_variable_1.head()
[4]:
GP MS
0 1 0
1 1 0
2 1 0
3 1 0
4 1 0
student-performance-data-...
student-por.csv
autosdataset
autos.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Console
Type here to search
4:08 PM
14/03/2022


My Quick Converter
kaggle - Saferbrowser Yahoo Inc
Run Data Science & Machine Le
notebook2da9e7a91a | Kaggle
kaggle.com/scratchpad/notebook2da9e7a91a/edit
Logged out session ends in 2 seconds
To save and continue working Sign In or Register
notebook2da9e7a91a
File Edit View Run Help
Run All
Code
Draft Session (10m)
4 1 0
[5]:
df=pd.concat([df, dummy_variable_1], axis=1)
df.drop("school", axis=1, inplace=True)
dummy_variable_2=pd.get_dummies(df['age'])
dummy_variable_2.rename(columns={'std':'age-std','number':'age-number'}, inplace=True)
dummy_variable_2.head()
[5]:
15 16 17 18 19 20 21 22
0 0 0 0 1 0 0 0 0
1 0 0 1 0 0 0 0 0
2 1 0 0 0 0 0 0 0
3 1 0 0 0 0 0 0 0
4 0 1 0 0 0 0 0 0
+ Code
+ Markdown
Console
student-performance-data-...
student-por.csv
autosdataset
autos.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Type here to search
4:14 PM
14/03/2022


My Quick Converter
kaggle - Saferbrowser Yahoo Inc
Run Data Science & Machine Le
notebook2da9e7a91a | Kaggle
kaggle.com/scratchpad/notebook2da9e7a91a/edit
Logged out session ends in 2 seconds
To save and continue working Sign In or Register
notebook2da9e7a91a
File Edit View Run Help
Run All
Code
Draft Session (12m)
[6]:
df=pd.concat([df, dummy_variable_2], axis=1)
df.drop('age',axis=1, inplace=True)
df.head(10)
[6]:
sex address famsize Pstatus Medu Fedu Mjob Fjob reason guardian ... GP MS 15 16 17 18 19 20 21 22
0 F U GT3 A 4 4 at_home teacher course mother ... 1 0 0 0 0 1 0 0 0 0
1 F U GT3 T 1 1 at_home other course father ... 1 0 0 0 1 0 0 0 0 0
2 F U LE3 T 1 1 at_home other other mother ... 1 0 1 0 0 0 0 0 0 0
3 F U GT3 T 4 2 health services home mother ... 1 0 1 0 0 0 0 0 0 0
4 F U GT3 T 3 3 other other home father ... 1 0 0 1 0 0 0 0 0 0
5 M U LE3 T 4 3 services other reputation mother ... 1 0 0 1 0 0 0 0 0 0
6 M U LE3 T 2 2 other other home mother ... 1 0 0 1 0 0 0 0 0 0
7 F U GT3 A 4 4 other teacher home mother ... 1 0 0 0 1 0 0 0 0 0
8 M U LE3 A 3 2 services other home mother ... 1 0 1 0 0 0 0 0 0 0
9 M U GT3 T 3 4 other other home mother ... 1 0 1 0 0 0 0 0 0 0
10 rows × 41 columns
+ Code
+ Markdown
Console
student-performance-data-...
student-por.csv
autosdataset
autos.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Type here to search
4:17 PM
14/03/2022


kaggle.com/scratchpad/notebook2da9e7a91a/edit
Logged out session ends in 2 seconds
To save and continue working Sign In or Register
notebook2da9e7a91a Draft saved
File Edit View Run Help
Run All
Code
Draft Session (13m)
8 M U LE3 A 3 2 services other home mother ... 1 0 1 0 0 0 0 0 0 0
9 M U GT3 T 3 4 other other home mother ... 1 0 1 0 0 0 0 0 0 0
10 rows × 41 columns
df.describe()
[7]:
Medu Fedu traveltime studytime failures famrel freetime goout Dalc Walc ... GP MS 15 16
count 649.000000 649.000000 649.000000 649.000000 649.000000 649.000000 649.000000 649.000000 649.000000 649.000000 ... 649.000000 649.000000 649.000000 649.000000
mean 2.514638 2.306626 1.568567 1.930663 0.221880 3.930663 3.180277 3.184900 1.502311 2.280431 ... 0.651772 0.348228 0.172573 0.272727
std 1.134552 1.099931 0.748660 0.829510 0.593235 0.955717 1.051093 1.175766 0.924834 1.284380 ... 0.476776 0.476776 0.378169 0.445705
min 0.000000 0.000000 1.000000 1.000000 0.000000 1.000000 1.000000 1.000000 1.000000 1.000000 ... 0.000000 0.000000 0.000000 0.000000
25% 2.000000 1.000000 1.000000 1.000000 0.000000 4.000000 3.000000 2.000000 1.000000 1.000000 ... 0.000000 0.000000 0.000000 0.000000
50% 2.000000 2.000000 1.000000 2.000000 0.000000 4.000000 3.000000 3.000000 1.000000 2.000000 ... 1.000000 0.000000 0.000000 0.000000
75% 4.000000 3.000000 2.000000 2.000000 0.000000 5.000000 4.000000 4.000000 2.000000 3.000000 ... 1.000000 1.000000 0.000000 1.000000
max 4.000000 4.000000 4.000000 4.000000 3.000000 5.000000 5.000000 5.000000 5.000000 5.000000 ... 1.000000 1.000000 1.000000 1.000000
8 rows × 25 columns
+ Code + Markdown
Console
student-performance-data-...
student-por.csv
autosdataset
autos.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Type here to search
4:18 PM
14/03/2022


kaggle.com/scratchpad/notebook2da9e7a91a/edit
Logged out session ends in 2 seconds
To save and continue working Sign In or Register
notebook2da9e7a91a Draft saved
File Edit View Run Help
Run All
Code
Draft Session (14m)
8 M U LE3 A 3 2 services other home mother ... 1 0 1 0 0 0 0 0 0 0
9 M U GT3 T 3 4 other other home mother ... 1 0 1 0 0 0 0 0 0 0
10 rows × 41 columns
df.describe()
[7]:
famrel freetime goout Dalc Walc ... GP MS 15 16 17 18 19 20 21 22
.000000 649.000000 649.000000 649.000000 649.000000 ... 649.000000 649.000000 649.000000 649.000000 649.000000 649.000000 649.000000 649.000000 649.000000 649.000000
.930663 3.180277 3.184900 1.502311 2.280431 ... 0.651772 0.348228 0.172573 0.272727 0.275809 0.215716 0.049307 0.009245 0.003082 0.001541
.955717 1.051093 1.175766 0.924834 1.284380 ... 0.476776 0.476776 0.378169 0.445705 0.447266 0.411636 0.216674 0.095779 0.055470 0.039253
.000000 1.000000 1.000000 1.000000 1.000000 ... 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000
.000000 3.000000 2.000000 1.000000 1.000000 ... 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000
.000000 3.000000 3.000000 1.000000 2.000000 ... 1.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000
.000000 4.000000 4.000000 2.000000 3.000000 ... 1.000000 1.000000 0.000000 1.000000 1.000000 0.000000 0.000000 0.000000 0.000000 0.000000
.000000 5.000000 5.000000 5.000000 5.000000 ... 1.000000 1.000000 1.000000 1.000000 1.000000 1.000000 1.000000 1.000000 1.000000 1.000000
+ Code + Markdown
Console
student-performance-data-...
student-por.csv
autosdataset
autos.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Type here to search
4:18 PM
14/03/2022

Write-up	Correctness of Program	Documentation of Program	Viva	Timely Completion	Total	Dated Sign of Subject Teacher
4	4	4	4	4	20	

Group A

Assignment No: 2

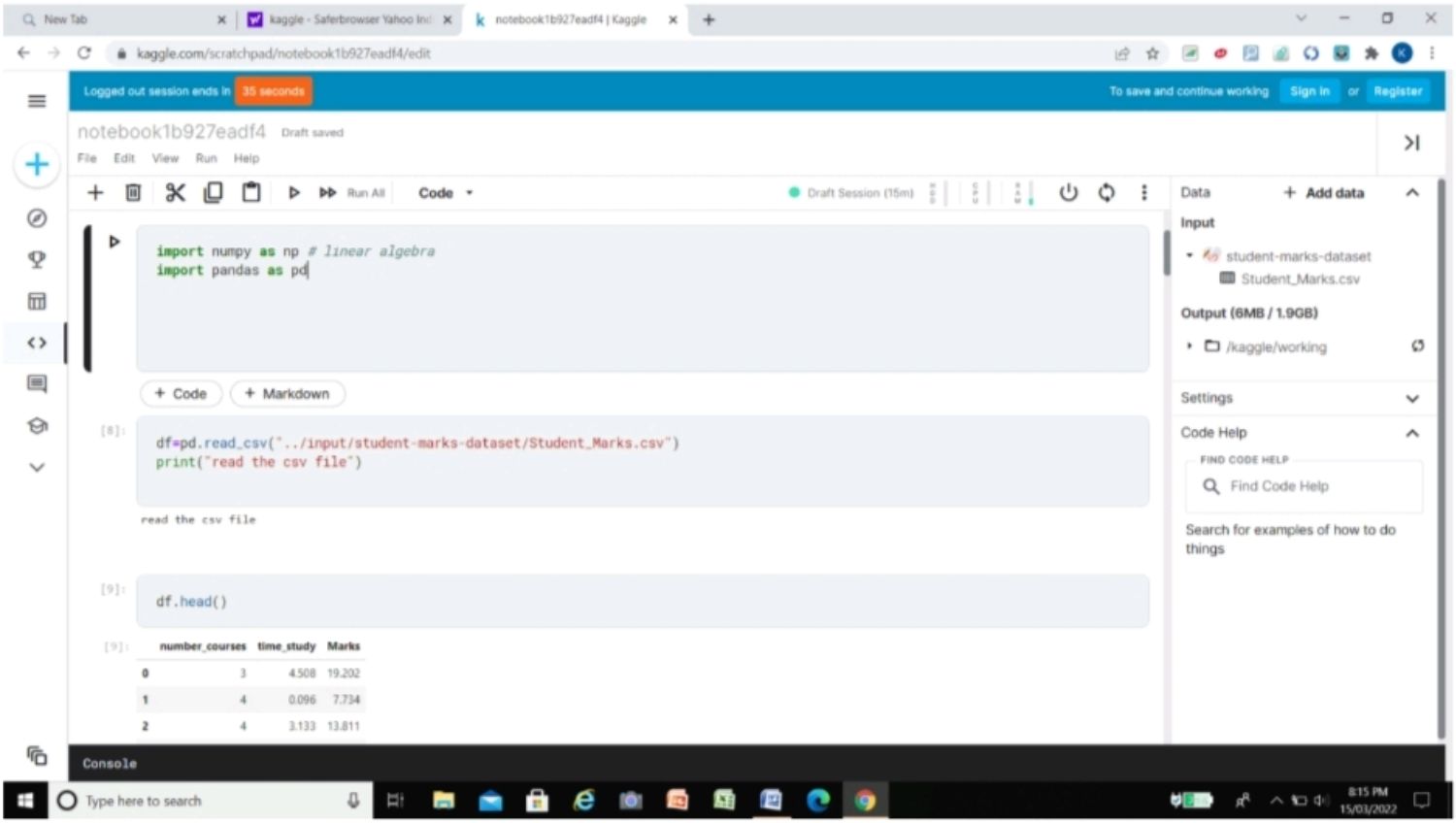
Title of the Assignment: Data Wrangling, II

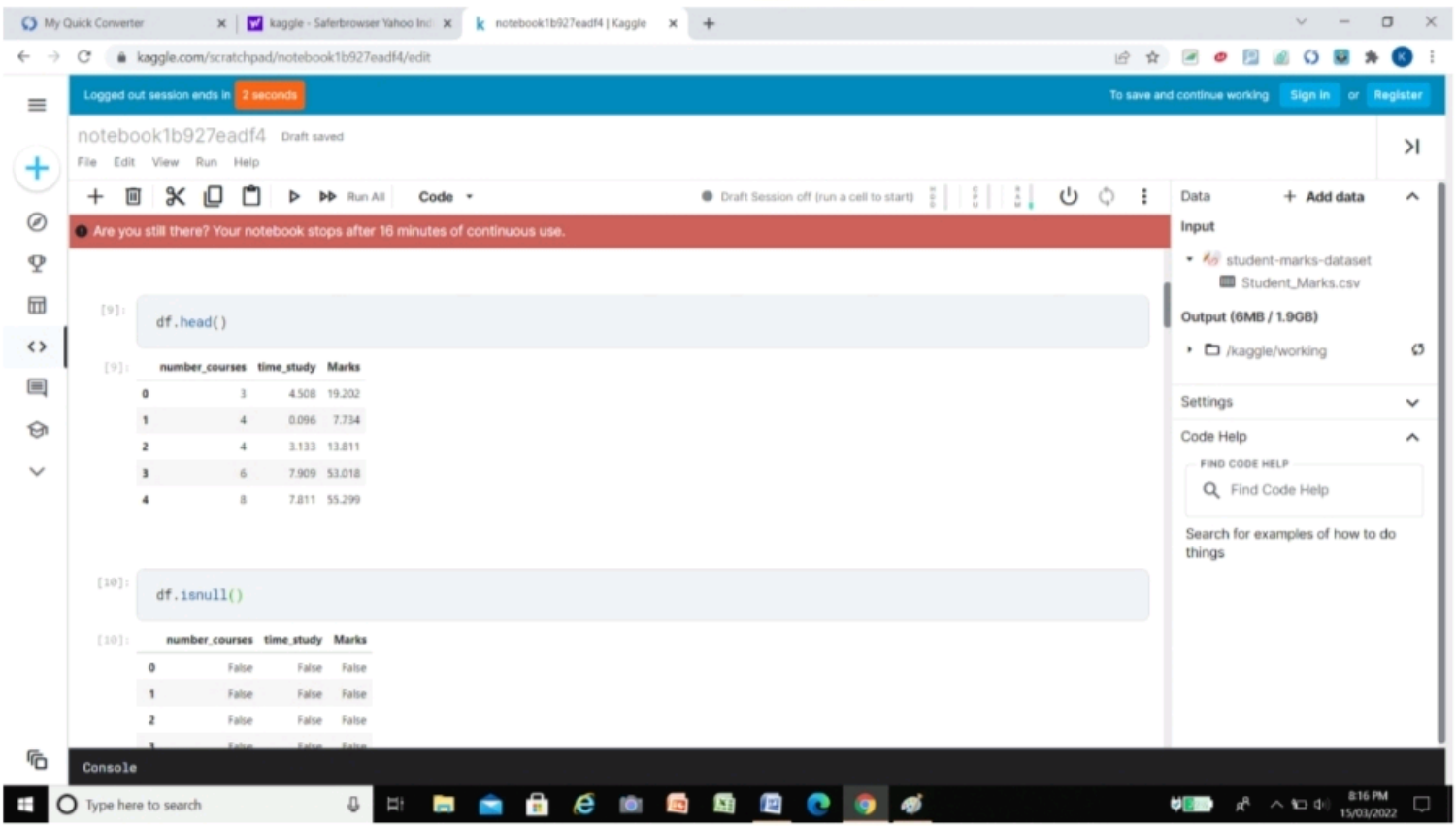
Create an "Academic performance" dataset of students and perform the following operations using Python.

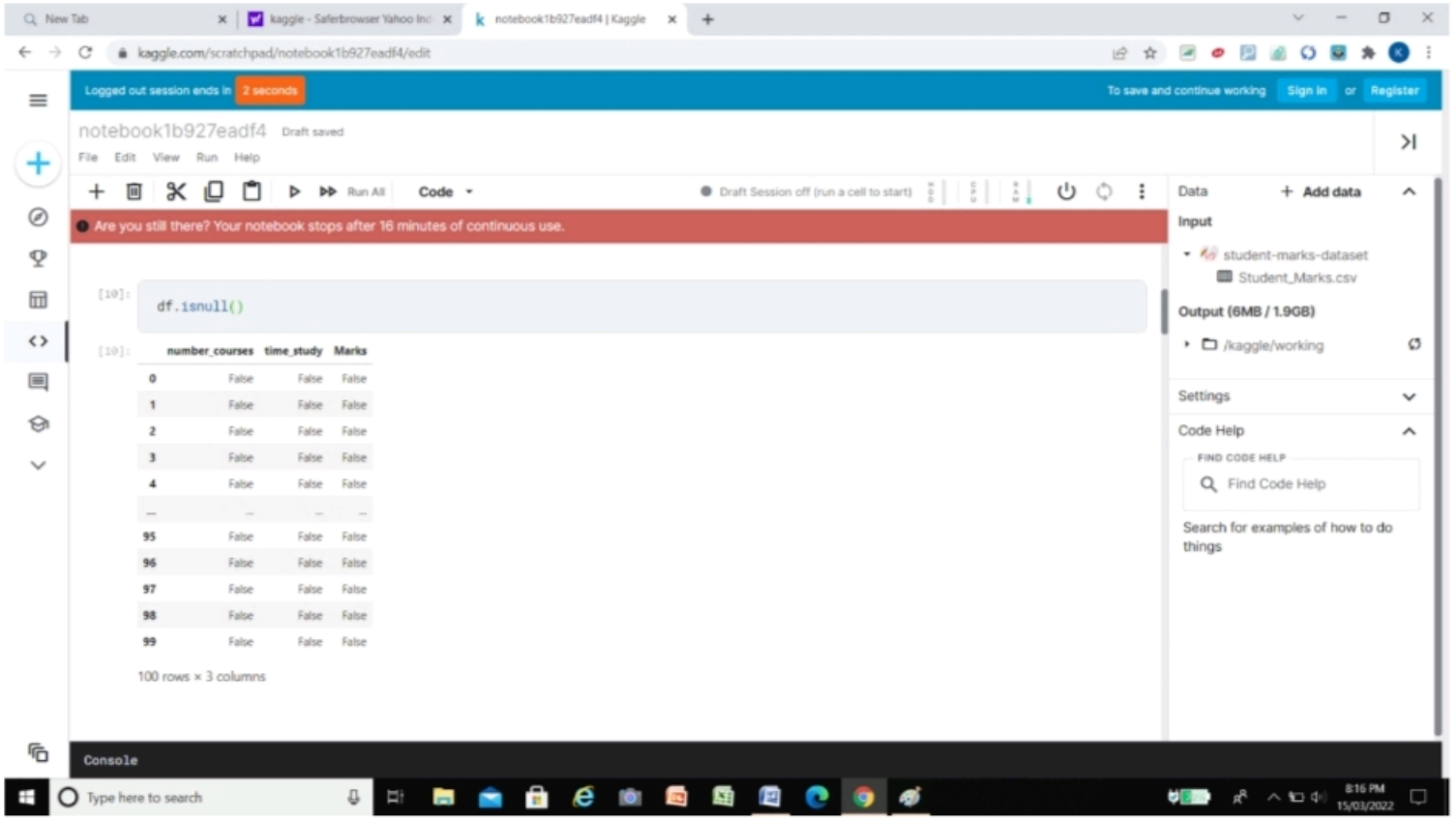
1. Scan all variables for missing values and inconsistencies. If there are missing values and/or inconsistencies, use any of the suitable techniques to deal with them.
2. Scan all numeric variables for outliers. If there are outliers, use any of the suitable techniques to deal with them.
3. Apply data transformations on at least one of the variables. The purpose of this transformation should be one of the following reasons: to change the scale for better understanding of the variable, to convert a non-linear relation into a linear one, or to decrease the skewness and convert the distribution into a normal distribution.

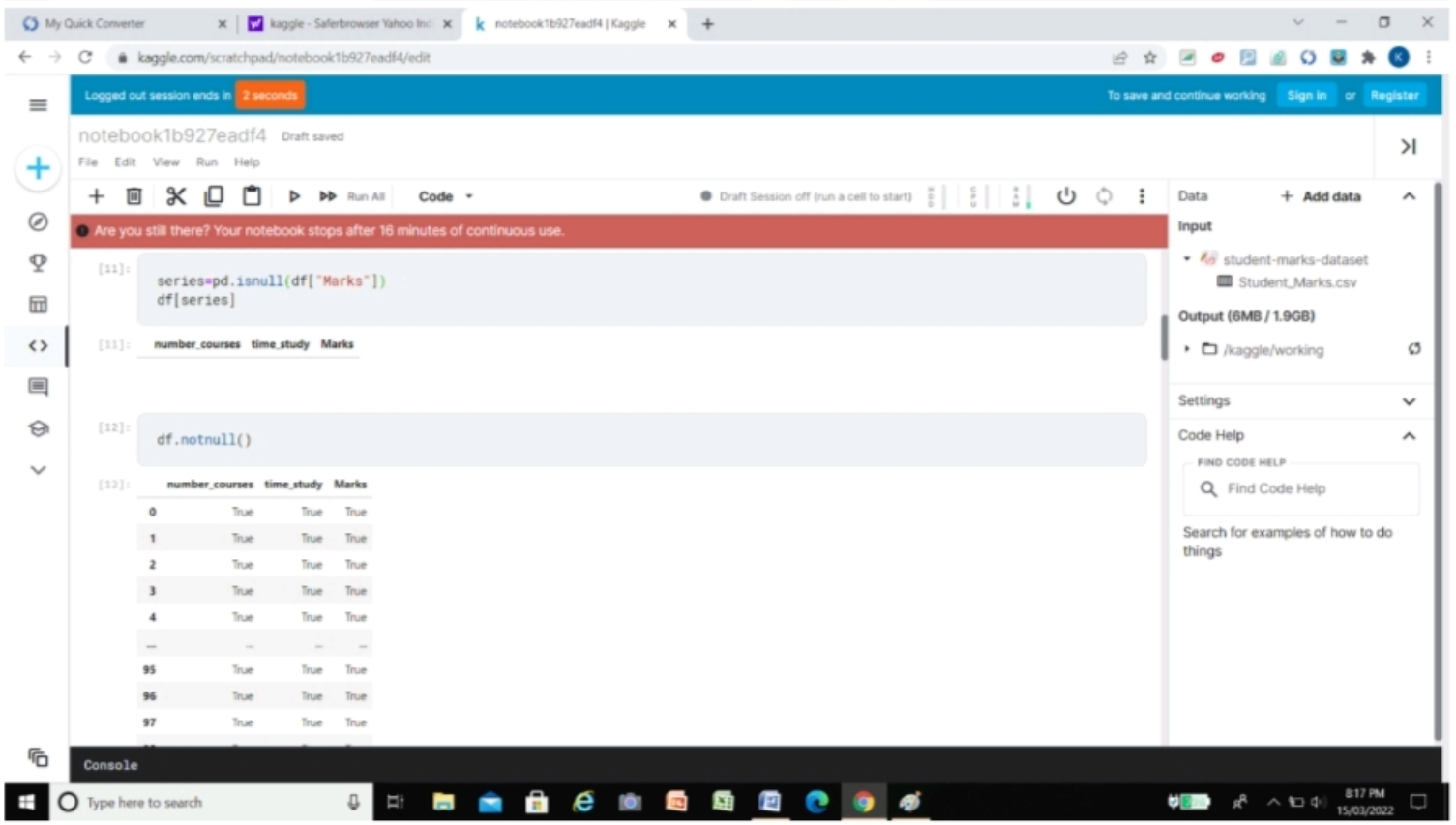
   Reason and document your approach properly.

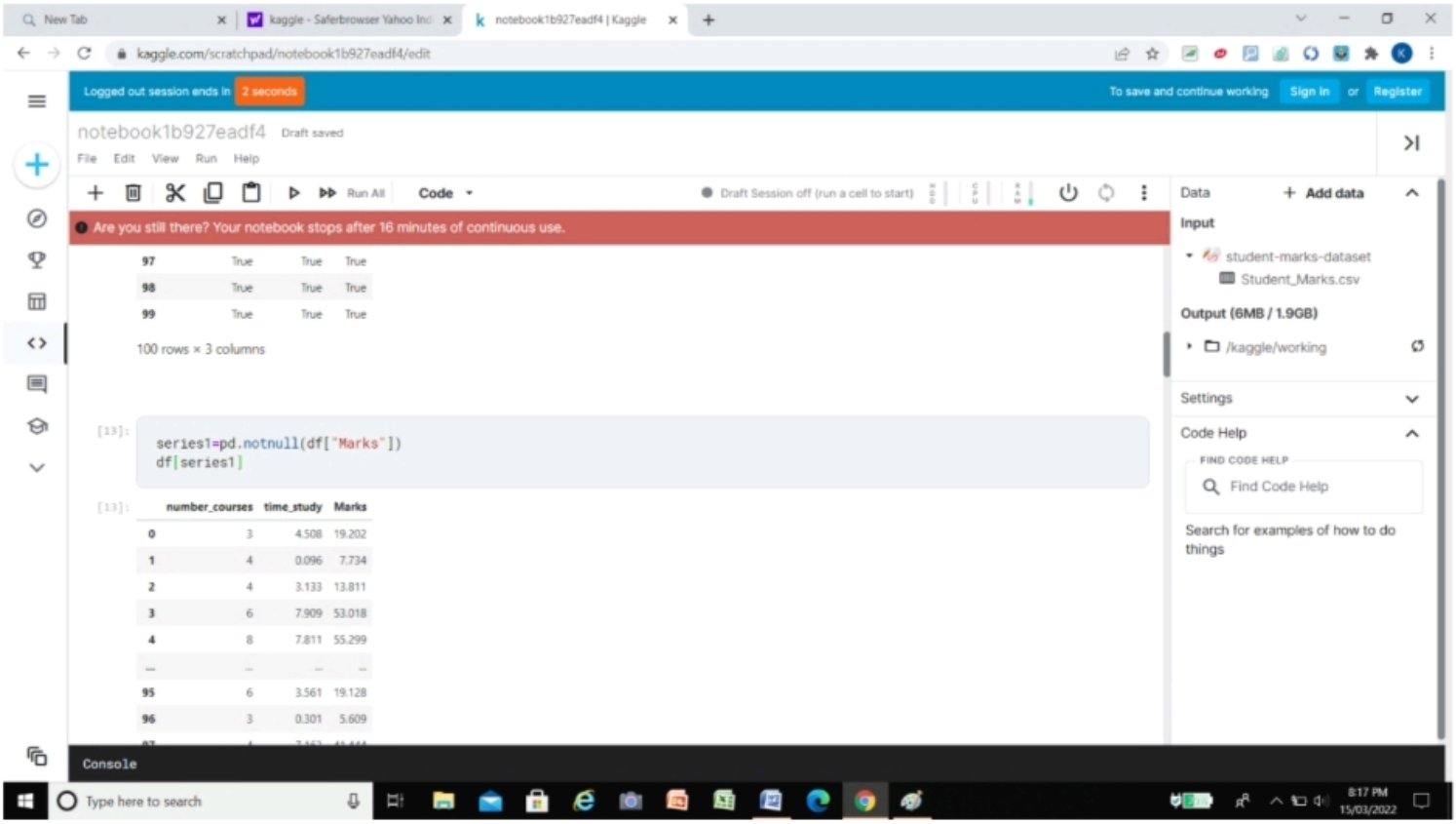
Output:

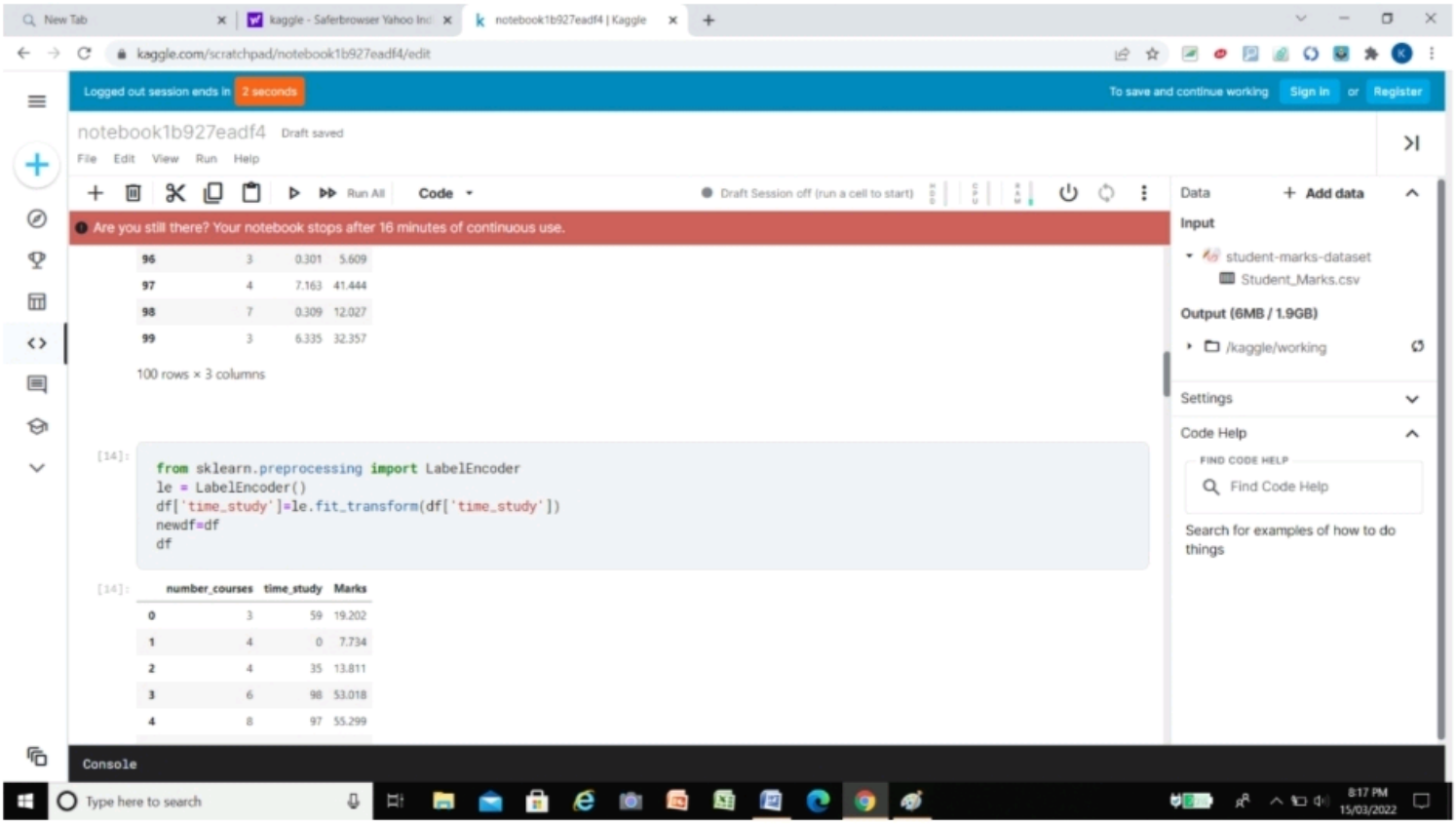

kaggle.com/scratchpad/notebook1b927eadf4/edit
Logged out session ends in 35 seconds
To save and continue working Sign In or Register
notebook1b927eadf4 Draft saved
File Edit View Run Help
Run All Code Draft Session (15m)
import numpy as np # linear algebra
import pandas as pd
+ Code + Markdown
[8]:
df=pd.read_csv("../input/student-marks-dataset/Student_Marks.csv")
print("read the csv file")
read the csv file
[9]:
df.head()
[9]:
number_courses time_study Marks
0 3 4.508 19.202
1 4 0.096 7.734
2 4 3.133 13.811
Console
Data + Add data
Input
student-marks-dataset
Student_Marks.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Type here to search
8:15 PM
15/03/2022

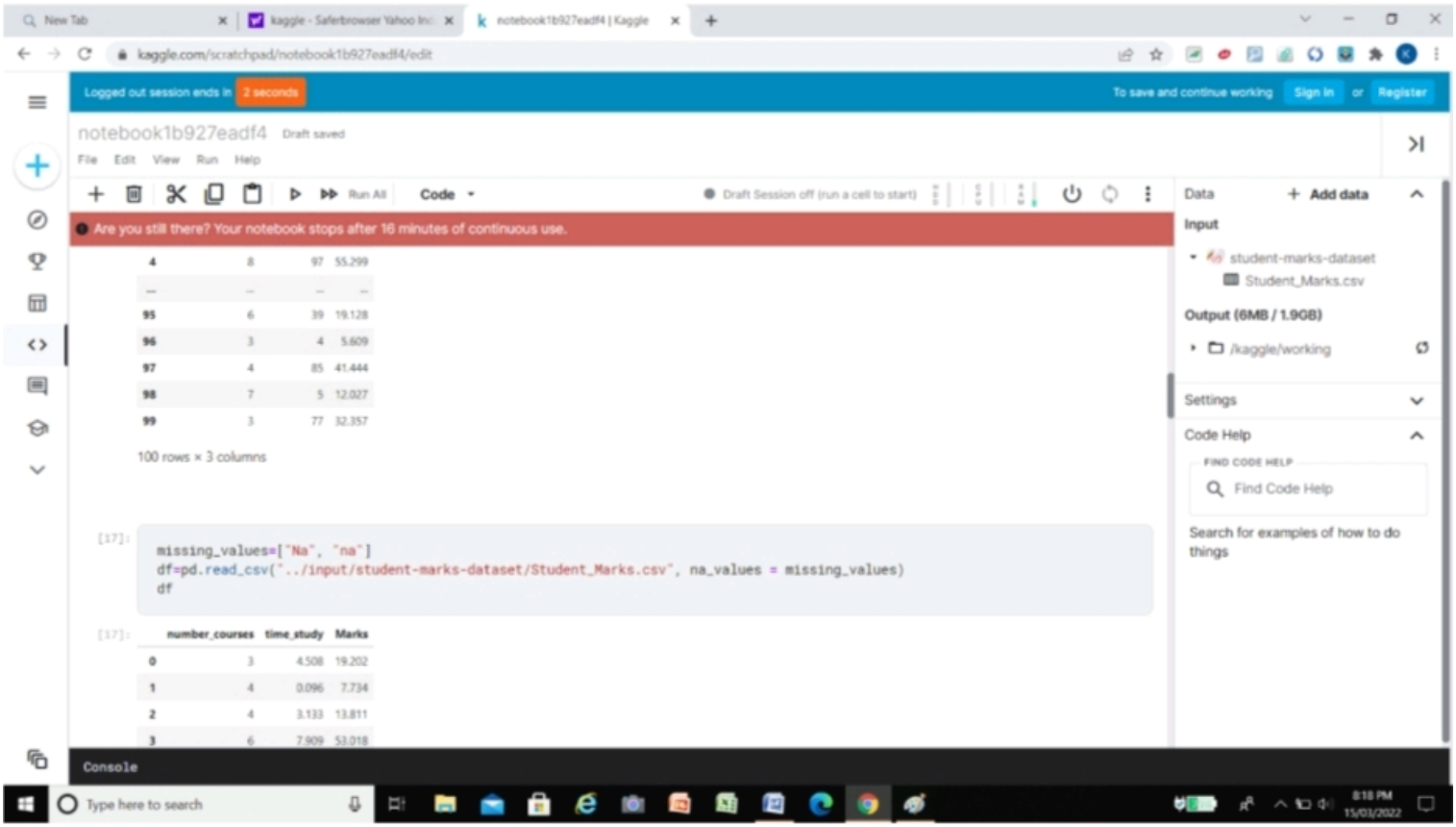

kaggle.com/scratchpad/notebook1b927eadf4/edit
Logged out session ends in 2 seconds
To save and continue working Sign In or Register
notebook1b927eadf4 Draft saved
File Edit View Run Help
Run All Code Draft Session off (run a cell to start)
Are you still there? Your notebook stops after 16 minutes of continuous use.
[9]:
df.head()
[9]:
number_courses time_study Marks
0 3 4.508 19.202
1 4 0.096 7.734
2 4 3.133 13.811
3 6 7.909 53.018
4 8 7.811 55.299
[10]:
df.isnull()
[10]:
number_courses time_study Marks
0 False False False
1 False False False
2 False False False
Console
Data + Add data
Input
student-marks-dataset
Student_Marks.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Type here to search
8:16 PM
15/03/2022

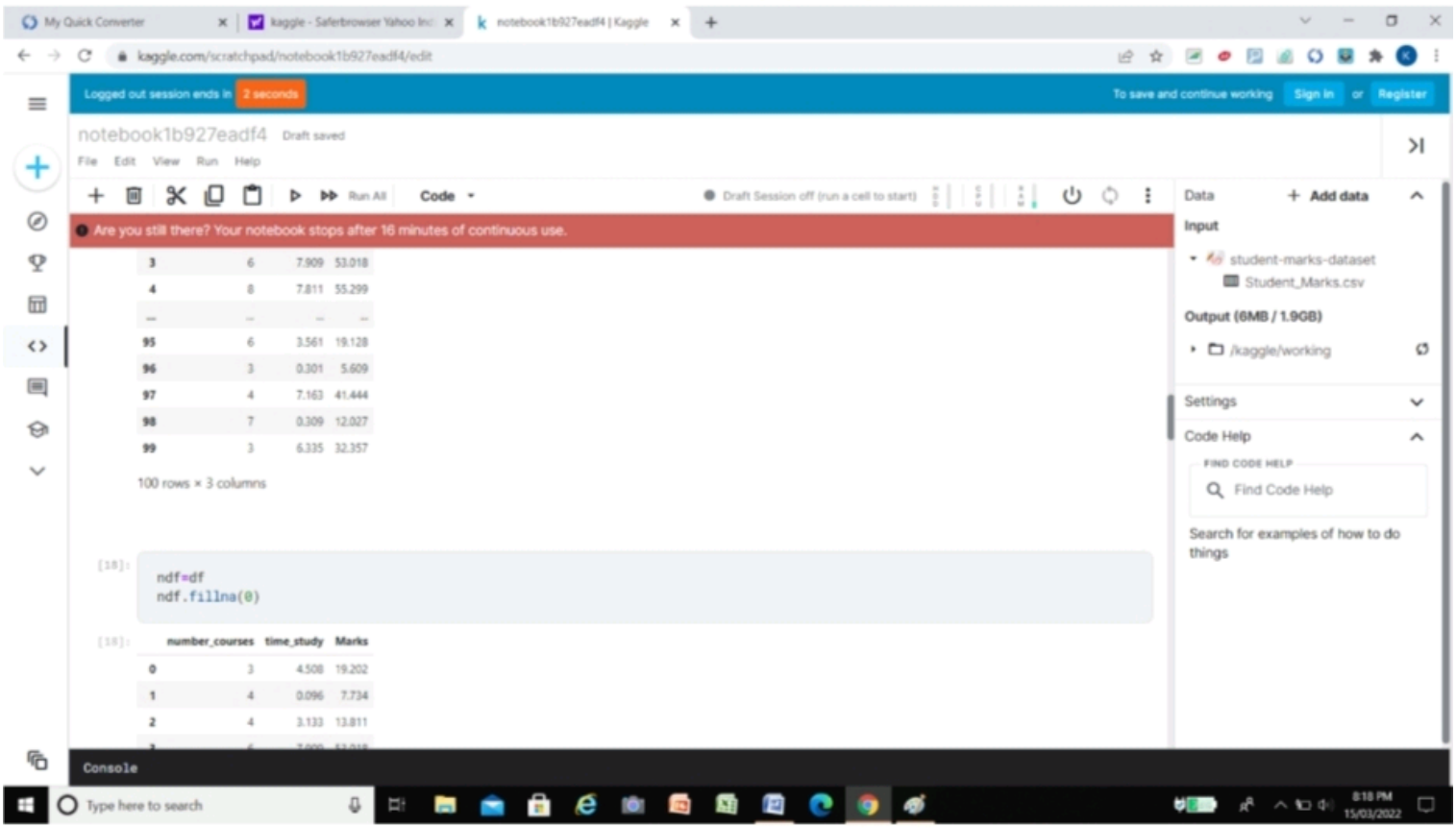

kaggle.com/scratchpad/notebook1b927eadf4/edit
Logged out session ends in 2 seconds
To save and continue working Sign In or Register
notebook1b927eadf4 Draft saved
File Edit View Run Help
Run All Code Draft Session off (run a cell to start)
Are you still there? Your notebook stops after 16 minutes of continuous use.
[10]: df.isnull()
[10]:
number_courses time_study Marks
0 False False False
1 False False False
2 False False False
3 False False False
4 False False False
95 False False False
96 False False False
97 False False False
98 False False False
99 False False False
100 rows × 3 columns
Data Add data
Input
student-marks-dataset
Student_Marks.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Console
Type here to search
8:16 PM
15/03/2022

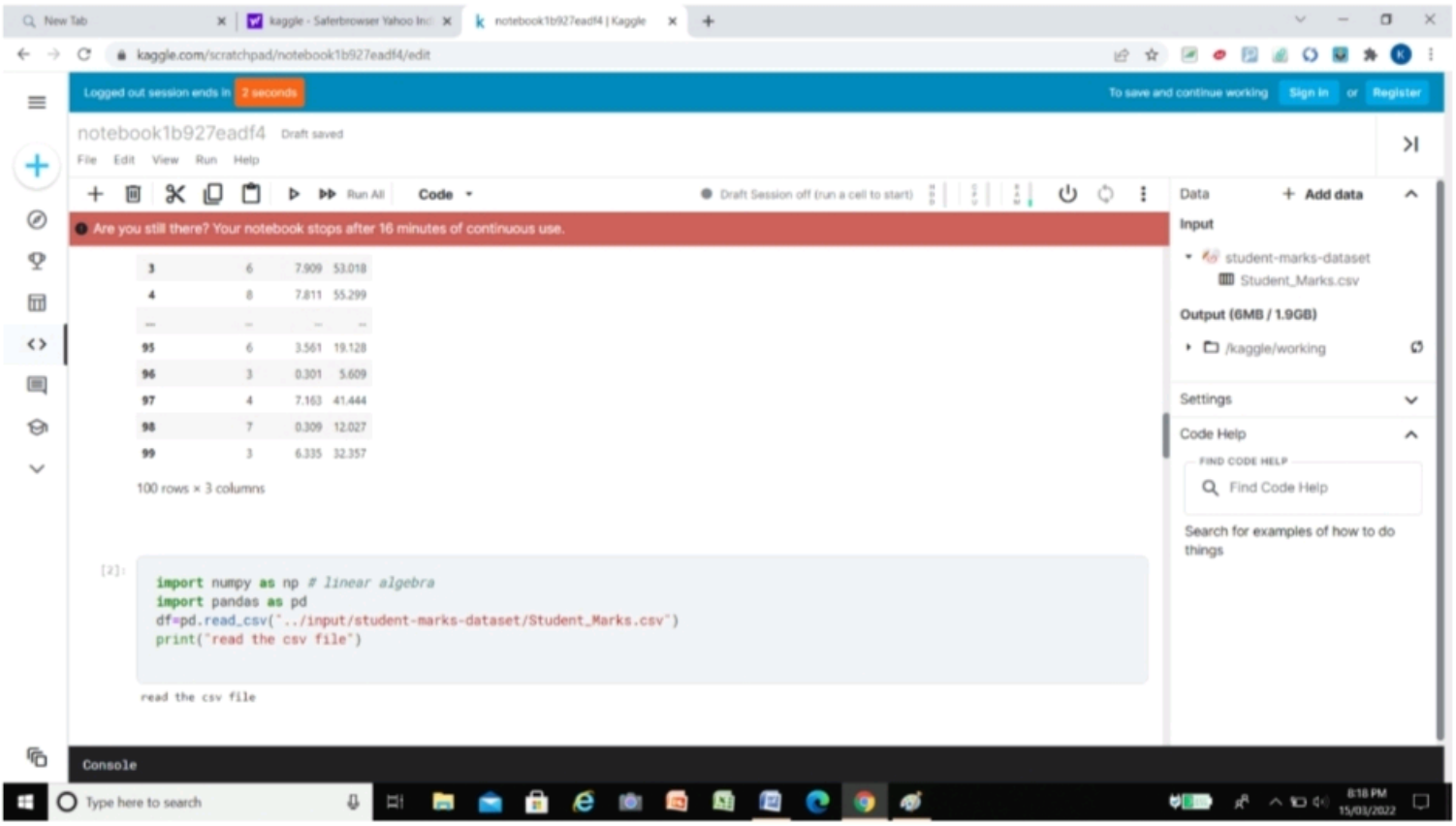

kaggle.com/scratchpad/notebook1b927eadf4/edit
Logged out session ends in 2 seconds
To save and continue working Sign In or Register
notebook1b927eadf4 Draft saved
File Edit View Run Help
Run All Code Draft Session off (run a cell to start)
Are you still there? Your notebook stops after 16 minutes of continuous use.
[11]: series=pd.isnull(df["Marks"])
df[series]
[11]:
number_courses time_study Marks
[12]: df.notnull()
[12]:
number_courses time_study Marks
0 True True True
1 True True True
2 True True True
3 True True True
4 True True True
95 True True True
96 True True True
97 True True True
Data Add data
Input
student-marks-dataset
Student_Marks.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Console
Type here to search
8:17 PM
15/03/2022

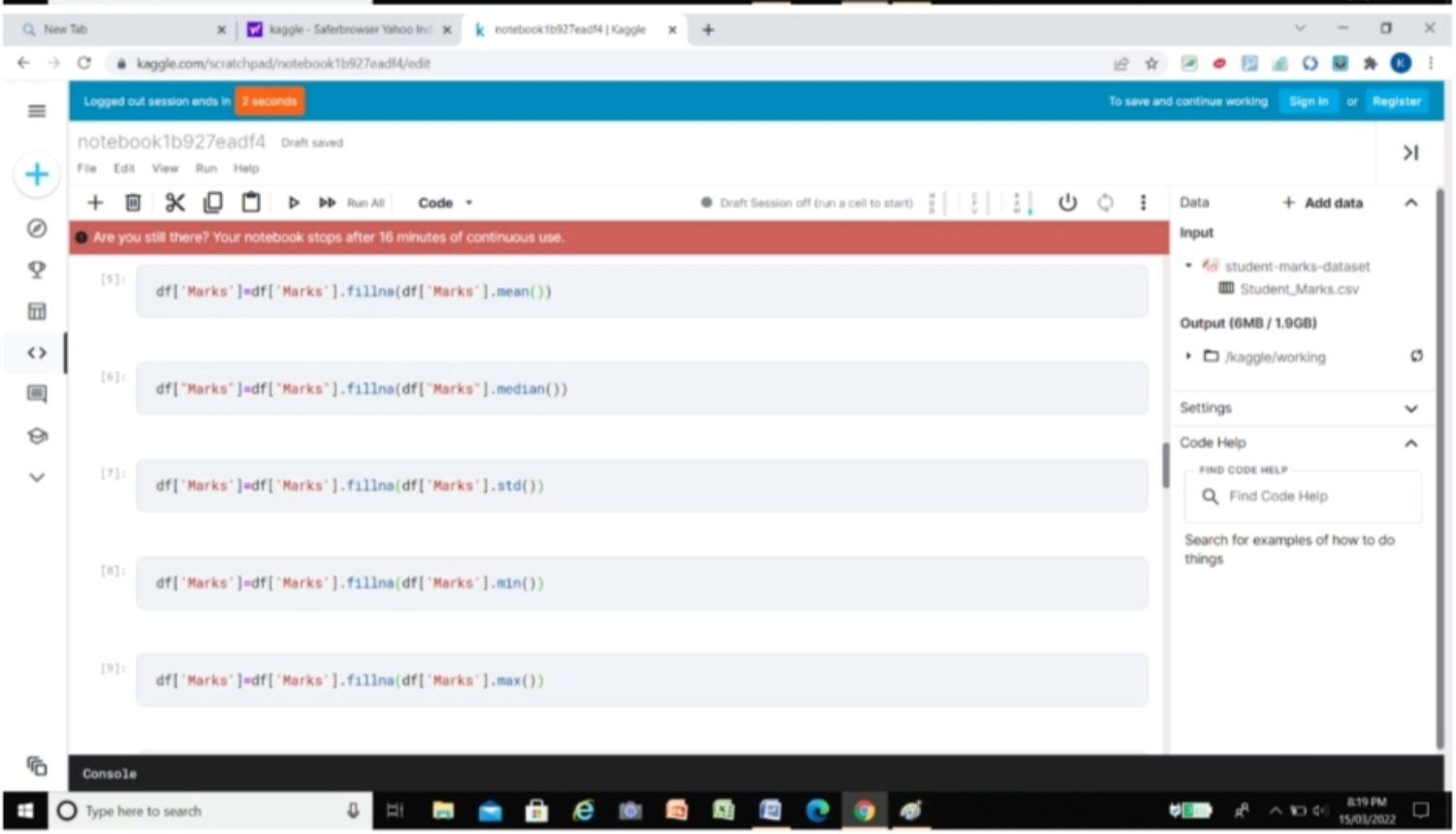

New Tab
kaggle - Saferbrowser Yahoo Ind
notebook1b927eadf4 | Kaggle
kaggle.com/scratchpad/notebook1b927eadf4/edit
Logged out session ends in 2 seconds
To save and continue working Sign in or Register
notebook1b927eadf4 Draft saved
File Edit View Run Help
Run All
Code
Draft Session off (run a cell to start)
Are you still there? Your notebook stops after 16 minutes of continuous use.
97 True True True
98 True True True
99 True True True
100 rows × 3 columns
[13]:
series1=pd.notnull(df["Marks"])
df[series1]
[13]:
number_courses time_study Marks
0 3 4.508 19.202
1 4 0.096 7.734
2 4 3.133 13.811
3 6 7.909 53.018
4 8 7.811 55.299
95 6 3.561 19.128
96 3 0.301 5.609
Data
Add data
Input
student-marks-dataset
Student_Marks.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Console
Type here to search
8:17 PM
15/03/2022

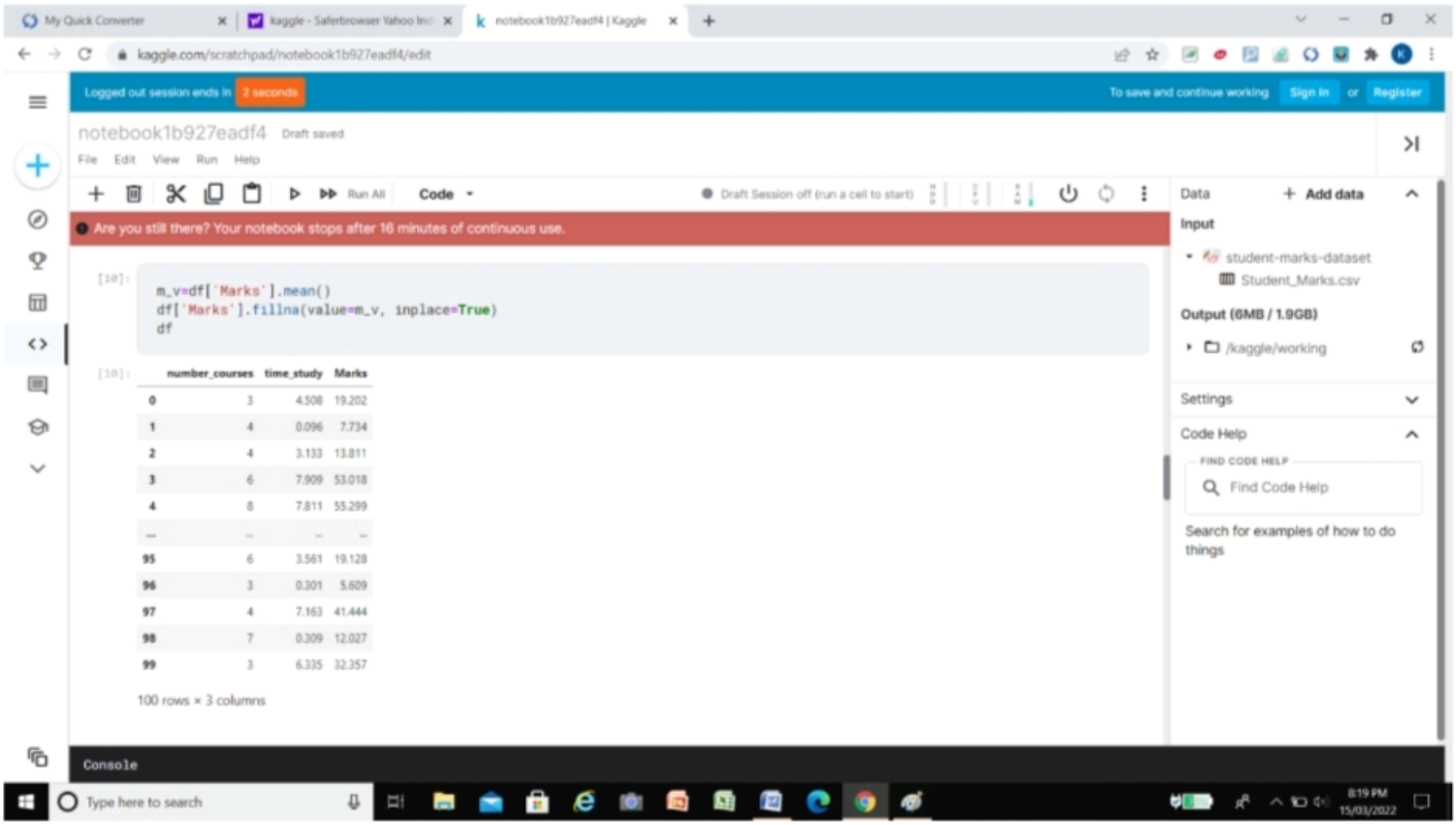

New Tab
kaggle - Saferbrowser Yahoo Ind
notebook1b927eadf4 | Kaggle
kaggle.com/scratchpad/notebook1b927eadf4/edit
Logged out session ends in 2 seconds
To save and continue working Sign in or Register
notebook1b927eadf4 Draft saved
File Edit View Run Help
Run All
Code
Draft Session off (run a cell to start)
Are you still there? Your notebook stops after 16 minutes of continuous use.
96 3 0.301 5.609
97 4 7.163 41.444
98 7 0.309 12.027
99 3 6.335 32.357
100 rows × 3 columns
[14]:
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
df['time_study']=le.fit_transform(df['time_study'])
newdf=df
df
[14]:
number_courses time_study Marks
0 3 59 19.202
1 4 0 7.734
2 4 35 13.811
3 6 98 53.018
4 8 97 55.299
Data
Add data
Input
student-marks-dataset
Student_Marks.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Console
Type here to search
8:17 PM
15/03/2022

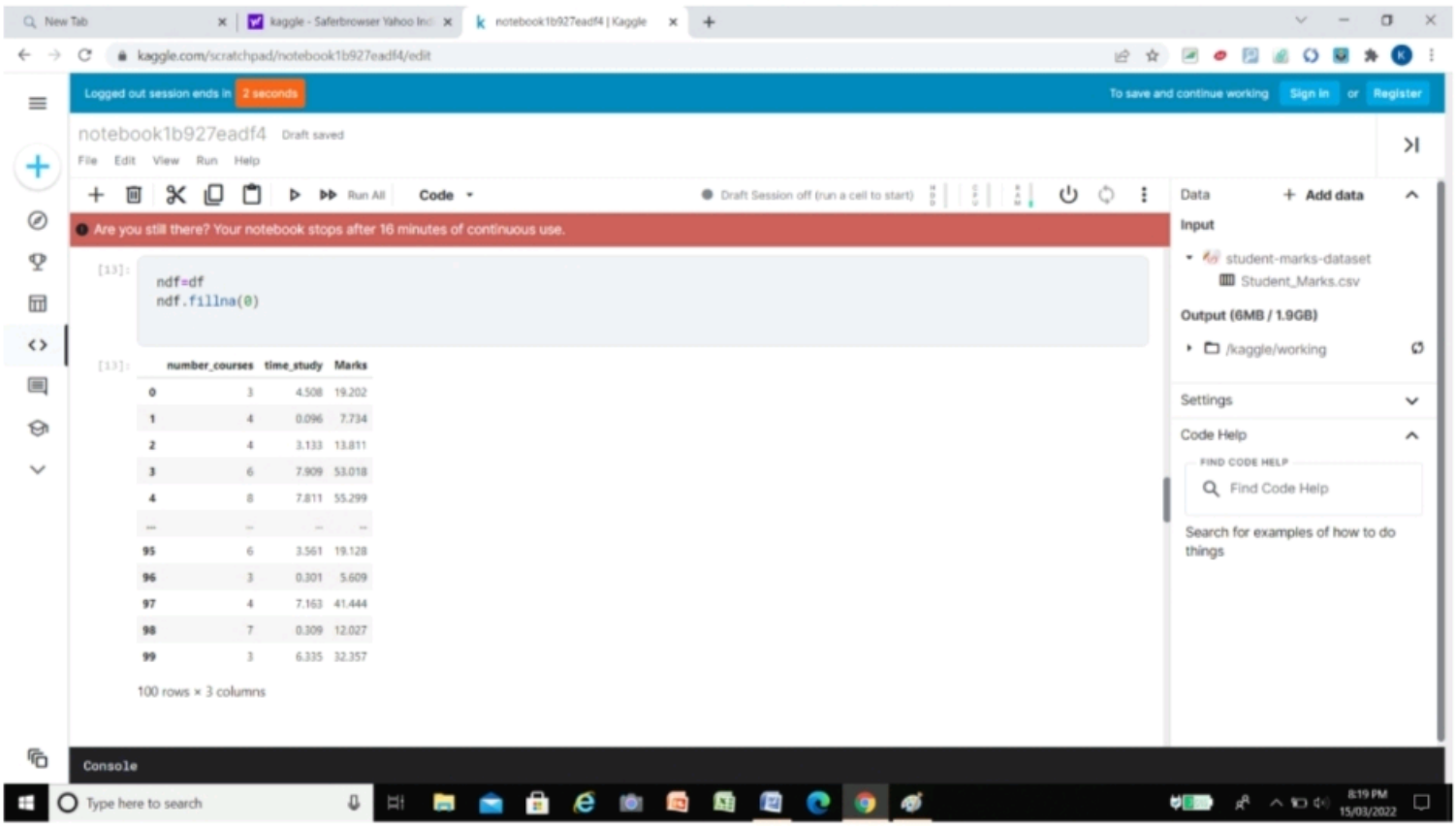

New Tab
kaggle - Saferbrowser Yahoo Inc
notebook1b927eadf4 | Kaggle
kaggle.com/scratchpad/notebook1b927eadf4/edit
Logged out session ends in 2 seconds
To save and continue working Sign in or Register
notebook1b927eadf4 Draft saved
File Edit View Run Help
Run All
Code
Draft Session off (run a cell to start)
Are you still there? Your notebook stops after 16 minutes of continuous use.
4 8 97 55.299
95 6 39 19.128
96 3 4 5.609
97 4 85 41.444
98 7 5 12.027
99 3 77 32.357
100 rows × 3 columns
[17]:
missing_values=["Na", "na"]
df=pd.read_csv("../input/student-marks-dataset/Student_Marks.csv", na_values = missing_values)
df
[17]:
number_courses time_study Marks
0 3 4.508 19.202
1 4 0.096 7.734
2 4 3.133 13.811
3 6 7.909 53.018
Console
Data
Add data
Input
student-marks-dataset
Student_Marks.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Type here to search
8:18 PM
15/03/2022

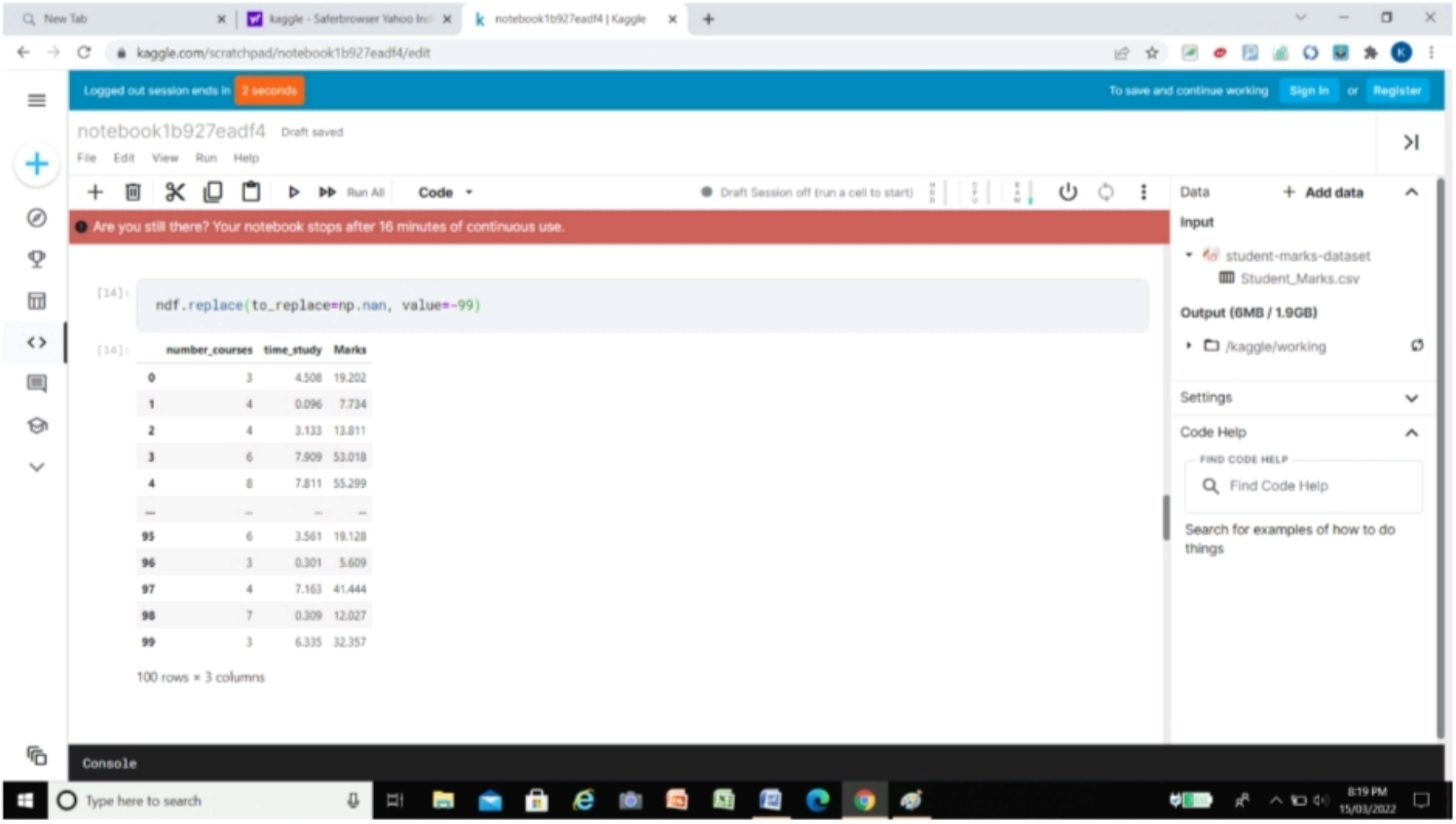

My Quick Converter
kaggle - Saferbrowser Yahoo Inc
notebook1b927eadf4 | Kaggle
kaggle.com/scratchpad/notebook1b927eadf4/edit
Logged out session ends in 2 seconds
To save and continue working Sign in or Register
notebook1b927eadf4 Draft saved
File Edit View Run Help
Run All
Code
Draft Session off (run a cell to start)
Are you still there? Your notebook stops after 16 minutes of continuous use.
3 6 7.909 53.018
4 8 7.811 55.299
95 6 3.561 19.128
96 3 0.301 5.609
97 4 7.163 41.444
98 7 0.309 12.027
99 3 6.335 32.357
100 rows × 3 columns
[18]:
ndf=df
ndf.fillna(0)
[18]:
number_courses time_study Marks
0 3 4.508 19.202
1 4 0.096 7.734
2 4 3.133 13.811
Console
Data
Add data
Input
student-marks-dataset
Student_Marks.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Type here to search
8:18 PM
15/03/2022


notebook1b927eadf4 | Kaggle
kaggle.com/scratchpad/notebook1b927eadf4/edit
Logged out session ends in 2 seconds
To save and continue working Sign In or Register
notebook1b927eadf4 Draft saved
File Edit View Run Help
Run All Code Draft Session off (run a cell to start)
Are you still there? Your notebook stops after 16 minutes of continuous use.
3 6 7.909 53.018
4 8 7.811 55.299
95 6 3.561 19.128
96 3 0.301 5.609
97 4 7.163 41.444
98 7 0.309 12.027
99 3 6.335 32.357
100 rows × 3 columns
[2]:
import numpy as np # linear algebra
import pandas as pd
df=pd.read_csv("../input/student-marks-dataset/Student_Marks.csv")
print("read the csv file")
read the csv file
Data Add data
Input
student-marks-dataset
Student_Marks.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Console
Type here to search
8:18 PM
15/03/2022


notebook1b927eadf4 | Kaggle
kaggle.com/scratchpad/notebook1b927eadf4/edit
Logged out session ends in 2 seconds
To save and continue working Sign In or Register
notebook1b927eadf4 Draft saved
File Edit View Run Help
Run All Code Draft Session off (run a cell to start)
Are you still there? Your notebook stops after 16 minutes of continuous use.
[5]:
df['Marks']=df['Marks'].fillna(df['Marks'].mean())
[6]:
df["Marks"]=df["Marks"].fillna(df["Marks"].median())
[7]:
df['Marks']=df['Marks'].fillna(df['Marks'].std())
[8]:
df['Marks']=df['Marks'].fillna(df['Marks'].min())
[9]:
df['Marks']=df['Marks'].fillna(df['Marks'].max())
Data Add data
Input
student-marks-dataset
Student_Marks.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Console
Type here to search
8:19 PM
15/03/2022


My Quick Converter
kaggle - Saferbrowser Yahoo Inc
notebook1b927eadf4 | Kaggle
kaggle.com/scratchpad/notebook1b927eadf4/edit
Logged out session ends in 2 seconds
To save and continue working Sign In or Register
notebook1b927eadf4 Draft saved
File Edit View Run Help
Run All Code Draft Session off (run a cell to start)
Are you still there? Your notebook stops after 16 minutes of continuous use.
[10]:
m_v=df['Marks'].mean()
df['Marks'].fillna(value=m_v, inplace=True)
df
[10]:
number_courses time_study Marks
0 3 4.508 19.202
1 4 0.096 7.734
2 4 3.133 13.811
3 6 7.909 53.018
4 8 7.811 55.299
... ...
95 6 3.561 19.128
96 3 0.301 5.609
97 4 7.163 41.444
98 7 0.309 12.027
99 3 6.335 32.357
100 rows × 3 columns
Data Add data
Input
student-marks-dataset
Student_Marks.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Console
Type here to search
8:19 PM
15/03/2022


New Tab
kaggle - Saferbrowser Yahoo Inc
notebook1b927eadf4 | Kaggle
kaggle.com/scratchpad/notebook1b927eadf4/edit
Logged out session ends in 2 seconds
To save and continue working Sign In or Register
notebook1b927eadf4 Draft saved
File Edit View Run Help
Run All Code Draft Session off (run a cell to start)
Are you still there? Your notebook stops after 16 minutes of continuous use.
[13]:
ndf=df
ndf.fillna(0)
[13]:
number_courses time_study Marks
0 3 4.508 19.202
1 4 0.096 7.734
2 4 3.133 13.811
3 6 7.909 53.018
4 8 7.811 55.299
... ...
95 6 3.561 19.128
96 3 0.301 5.609
97 4 7.163 41.444
98 7 0.309 12.027
99 3 6.335 32.357
100 rows × 3 columns
Data Add data
Input
student-marks-dataset
Student_Marks.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Console
Type here to search
8:19 PM
15/03/2022

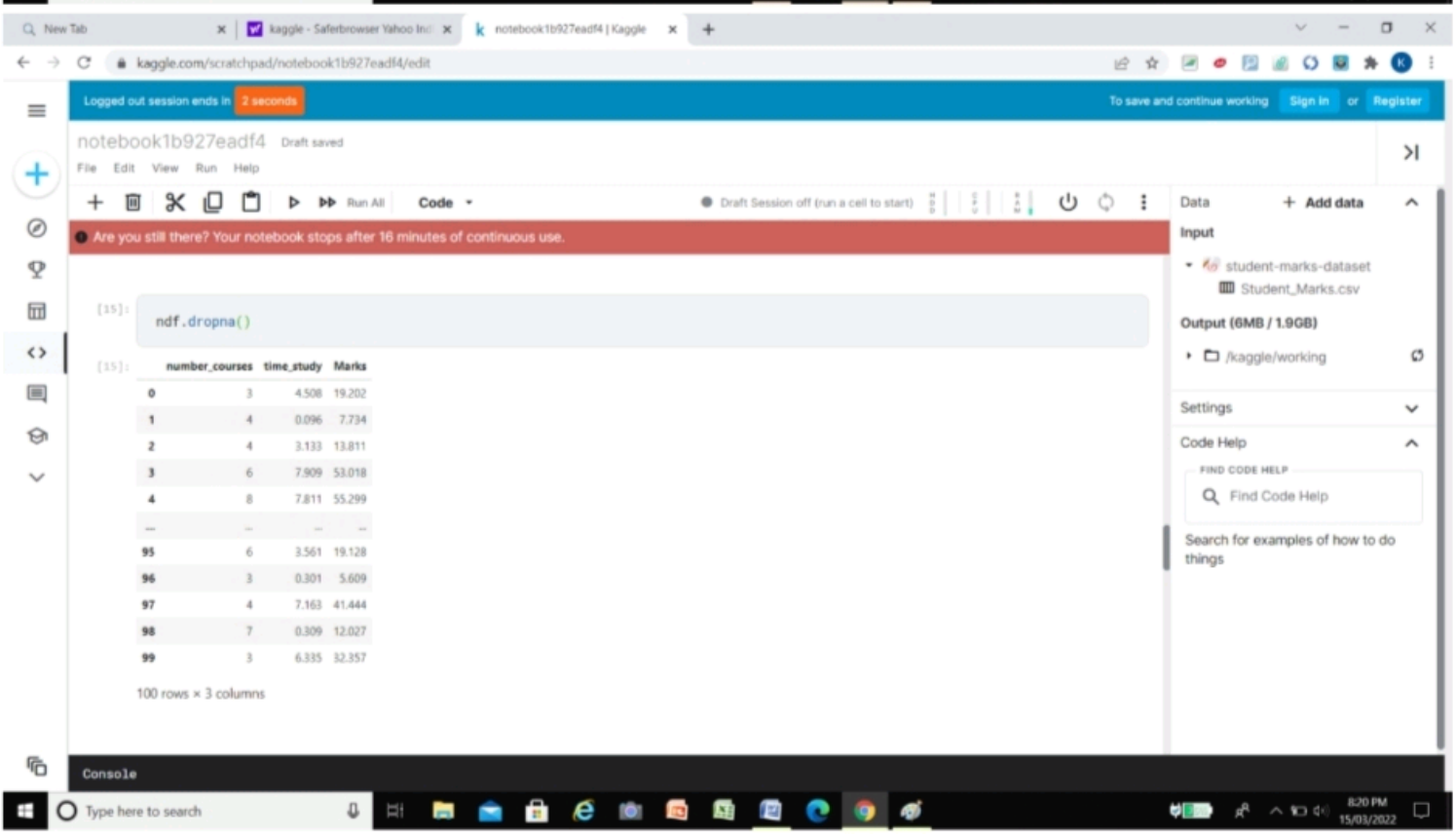


[14]:
```
ndf.replace(to_replace=np.nan, value=-99)
```

[14]:

	number_courses	time_study	Marks
0	3	4.508	19.202
1	4	0.096	7.734
2	4	3.133	13.811
3	6	7.909	53.018
4	8	7.811	55.299
...	...	...	...
95	6	3.561	19.128
96	3	0.301	5.609
97	4	7.163	41.444
98	7	0.309	12.027
99	3	6.335	32.357

100 rows × 3 columns

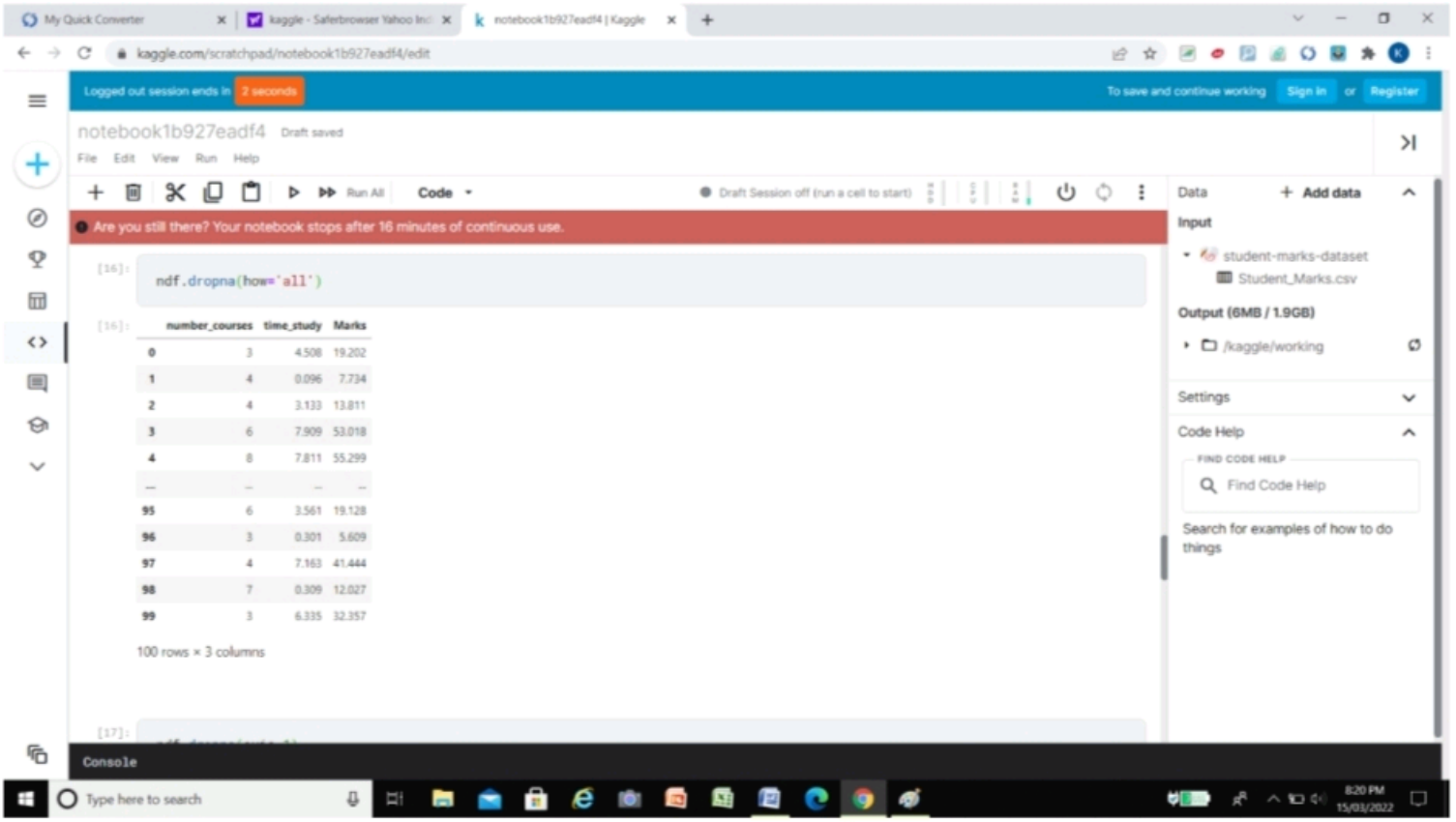


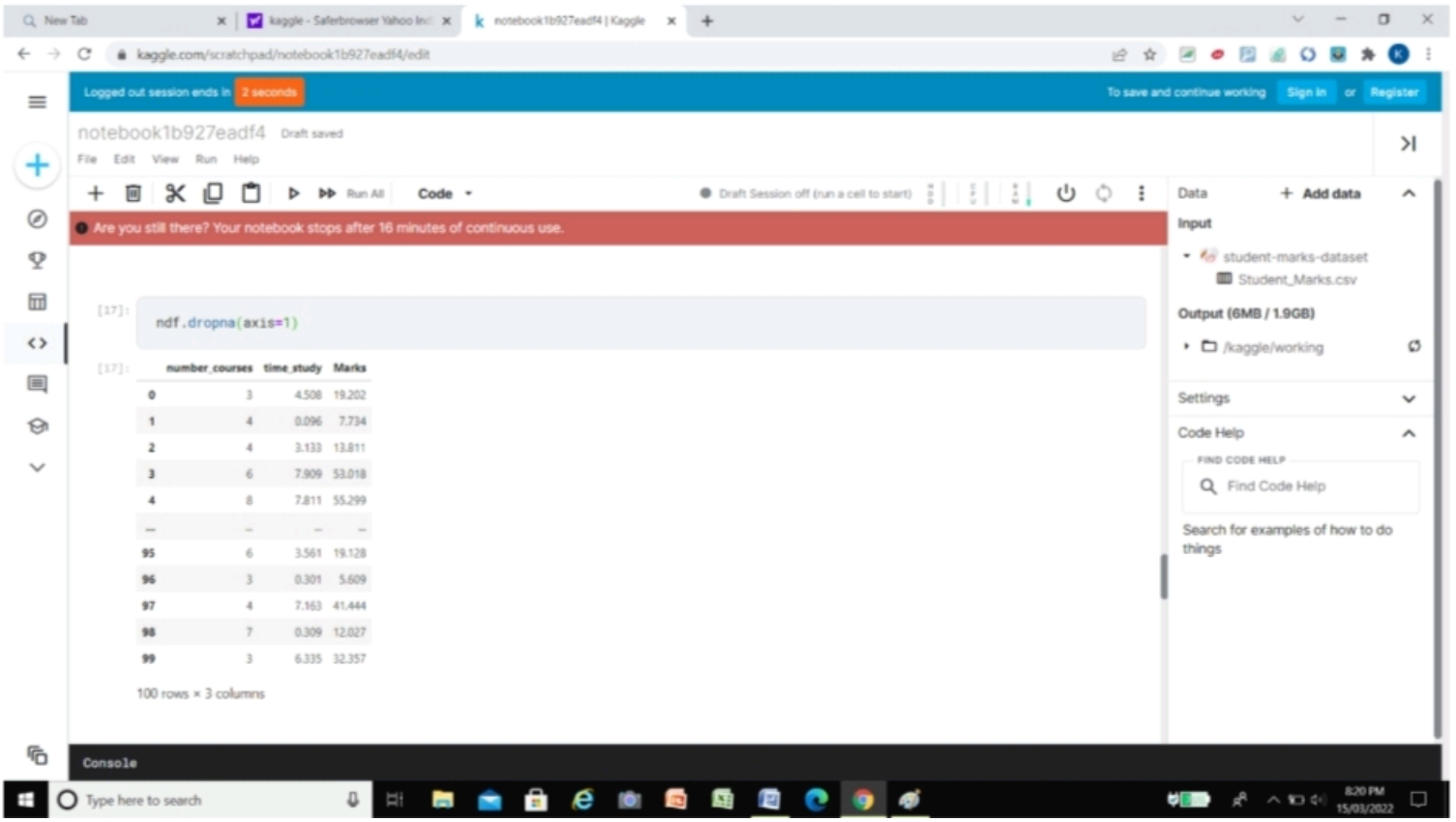
[15]:
```
ndf.dropna()
```

[15]:

	number_courses	time_study	Marks
0	3	4.508	19.202
1	4	0.096	7.734
2	4	3.133	13.811
3	6	7.909	53.018
4	8	7.811	55.299
...	...	...	...
95	6	3.561	19.128
96	3	0.301	5.609
97	4	7.163	41.444
98	7	0.309	12.027
99	3	6.335	32.357

100 rows × 3 columns


My Quick Converter
kaggle - Saferbrowser Yahoo Inc
notebook1b927eadf4 | Kaggle
kaggle.com/scratchpad/notebook1b927eadf4/edit
Logged out session ends in 2 seconds
To save and continue working Sign in or Register
notebook1b927eadf4 Draft saved
File Edit View Run Help
Run All
Code
Draft Session off (run a cell to start)
Are you still there? Your notebook stops after 16 minutes of continuous use.
[16]: ndf.dropna(how='all')
[16]:
number_courses time_study Marks
0 3 4.508 19.202
1 4 0.096 7.734
2 4 3.133 13.811
3 6 7.909 53.018
4 8 7.811 55.299
95 6 3.561 19.128
96 3 0.301 5.609
97 4 7.163 41.444
98 7 0.309 12.027
99 3 6.335 32.357
100 rows × 3 columns
[17]:
Console
Data
Add data
Input
student-marks-dataset
Student_Marks.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Type here to search
8:20 PM
15/03/2022


New Tab
kaggle - Saferbrowser Yahoo Inc
notebook1b927eadf4 | Kaggle
kaggle.com/scratchpad/notebook1b927eadf4/edit
Logged out session ends in 2 seconds
To save and continue working Sign in or Register
notebook1b927eadf4 Draft saved
File Edit View Run Help
Run All
Code
Draft Session off (run a cell to start)
Are you still there? Your notebook stops after 16 minutes of continuous use.
[17]: ndf.dropna(axis=1)
[17]:
number_courses time_study Marks
0 3 4.508 19.202
1 4 0.096 7.734
2 4 3.133 13.811
3 6 7.909 53.018
4 8 7.811 55.299
95 6 3.561 19.128
96 3 0.301 5.609
97 4 7.163 41.444
98 7 0.309 12.027
99 3 6.335 32.357
100 rows × 3 columns
Console
Data
Add data
Input
student-marks-dataset
Student_Marks.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Type here to search
8:20 PM
15/03/2022

My Quick Converter | kaggle - Saferbrowser Yahoo Inc | notebook1b927eadf4 | Kaggle

kaggle.com/scratchpad/notebook1b927eadf4/edit

Logged out session ends in 2 seconds — To save and continue working Sign In or Register

notebook1b927eadf4 Draft saved

File Edit View Run Help

Run All | Code | Draft Session off (run a cell to start)

Are you still there? Your notebook stops after 16 minutes of continuous use.

[18]:
```
new_data=ndf.dropna(axis=0, how='any')
new_data
```

[18]:

	number_courses	time_study	Marks
0	3	4.508	19.202
1	4	0.096	7.734
2	4	3.133	13.811
3	6	7.909	53.018
4	8	7.811	55.299
...	...	...	...
95	6	3.561	19.128
96	3	0.301	5.609
97	4	7.163	41.444
98	7	0.309	12.027
99	3	6.335	32.357

100 rows × 3 columns

Data | + Add data

Input

student-marks-dataset

Student_Marks.csv

Output (6MB / 1.9GB)

/kaggle/working

Settings

Code Help

FIND CODE HELP

Find Code Help

Search for examples of how to do things

Console

Type here to search

8:21 PM 15/03/2022

My Quick Converter | kaggle - Saferbrowser Yahoo Inc | notebook1b927eadf4 | Kaggle

kaggle.com/scratchpad/notebook1b927eadf4/edit

Logged out session ends in 2 seconds — To save and continue working Sign In or Register

notebook1b927eadf4 Draft saved

File Edit View Run Help

Run All | Code | Draft Session off (run a cell to start)

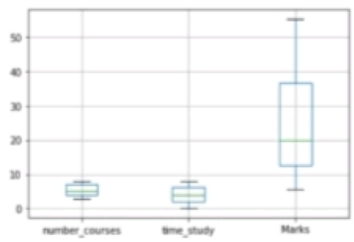
Are you still there? Your notebook stops after 16 minutes of continuous use.

100 rows × 3 columns

[19]:
```
col=['number_courses','time_study', 'Marks']
df.boxplot(col)
```

[19]:

```
<AxesSubplot:>
```



[2]:
```
import numpy as np # linear algebra
```

Data | + Add data

Input

student-marks-dataset

Student_Marks.csv

Output (6MB / 1.9GB)

/kaggle/working

Settings

Code Help

FIND CODE HELP

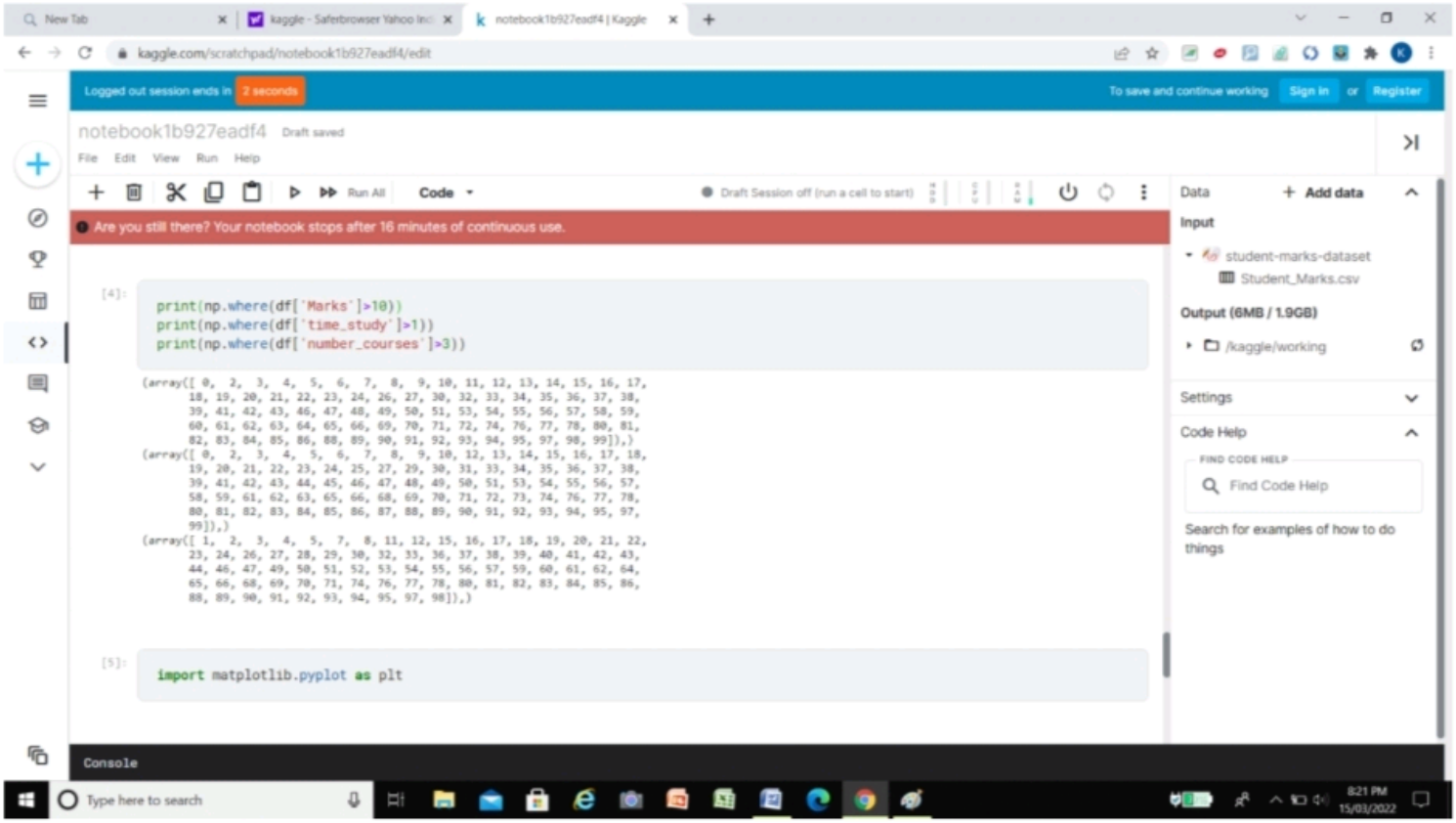
Find Code Help

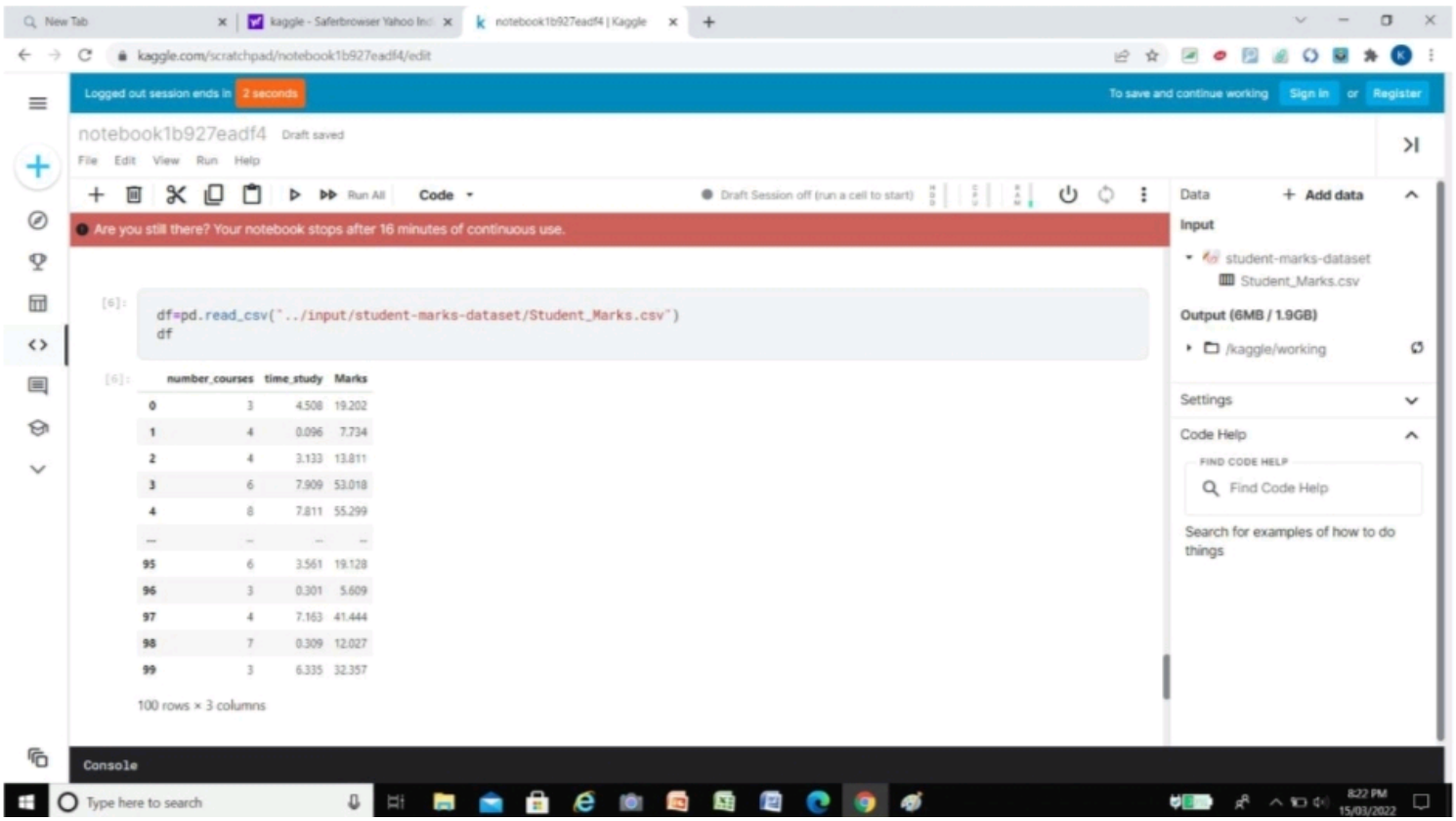
Search for examples of how to do things

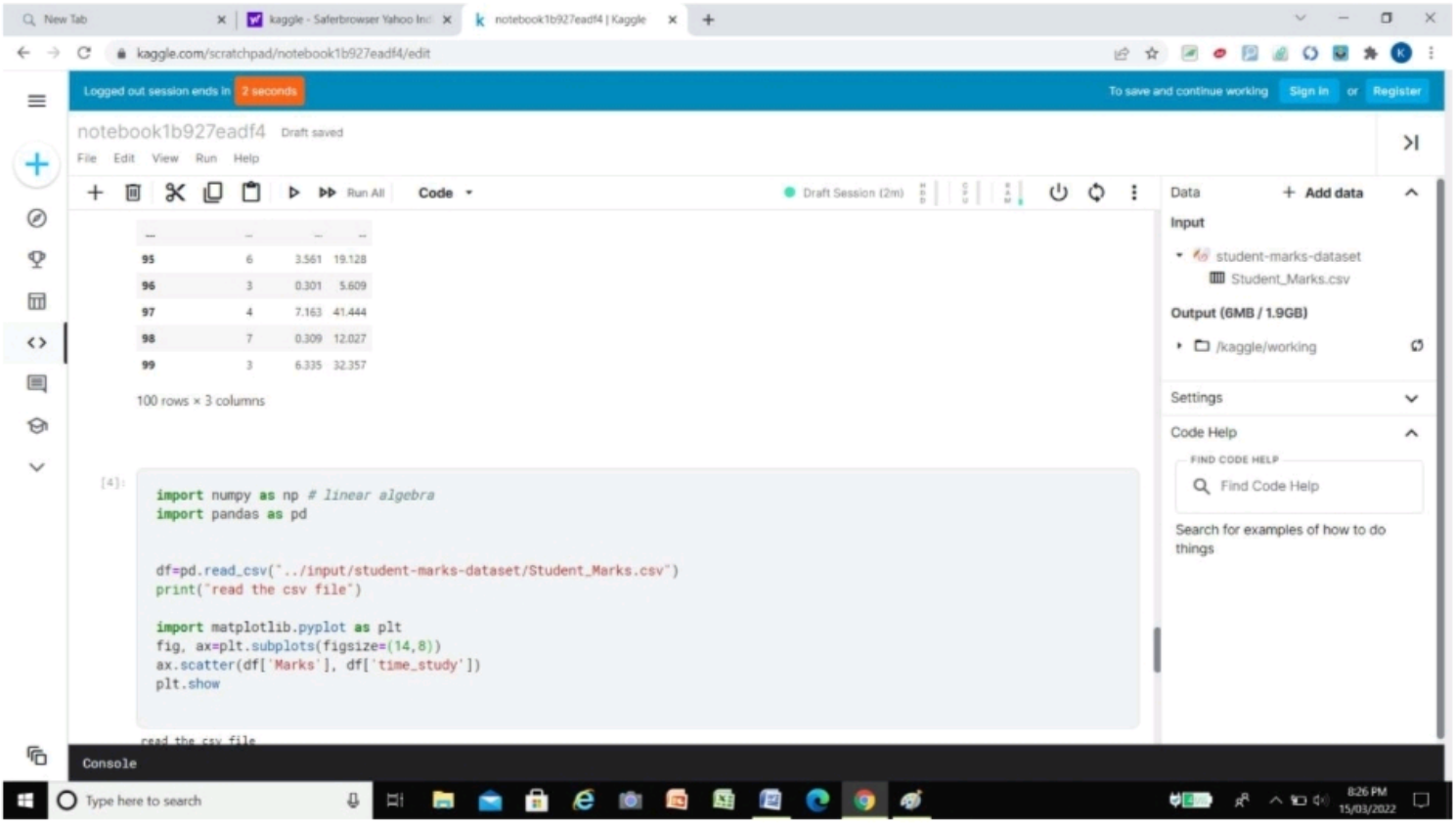
Console

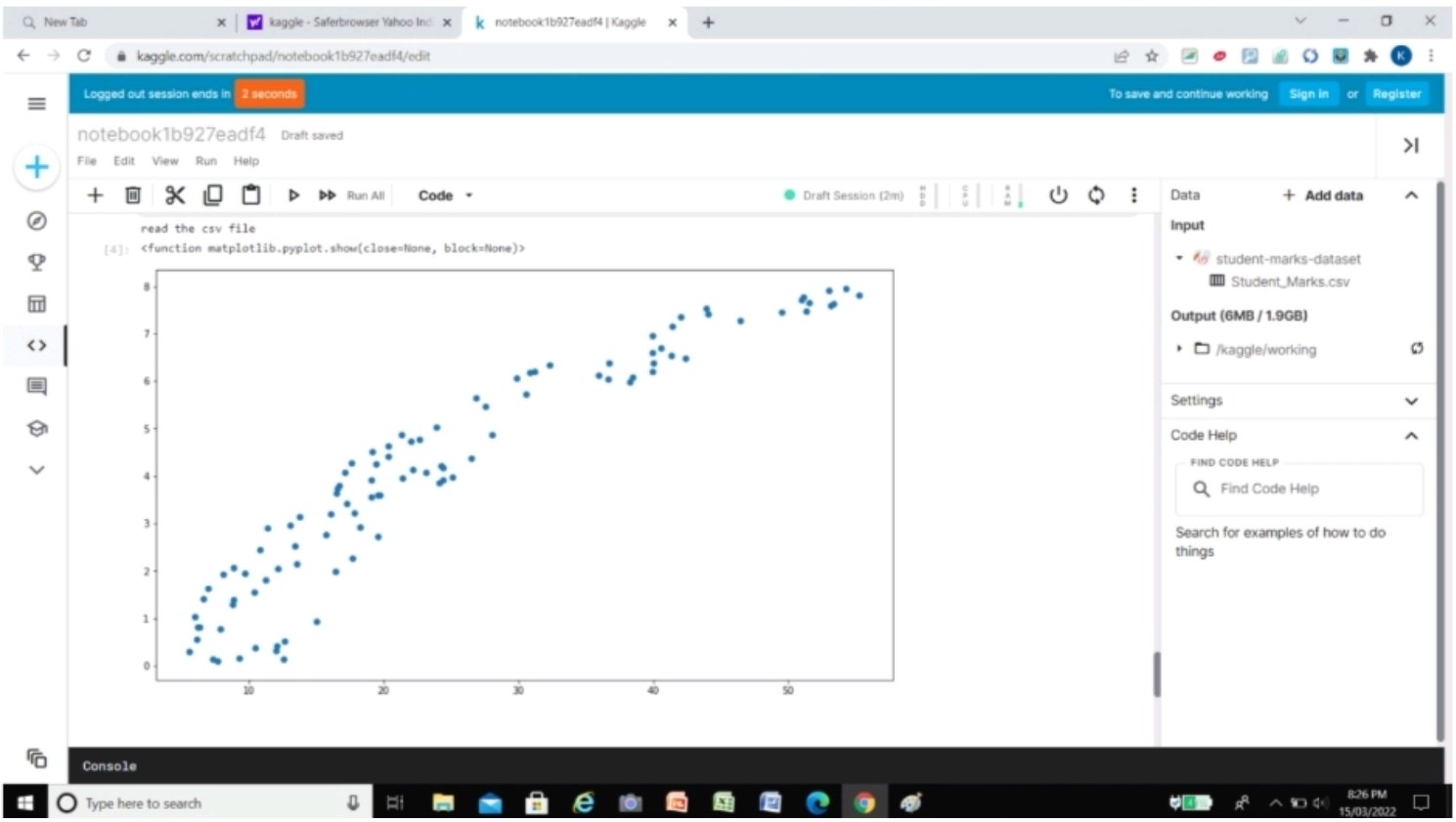
Type here to search

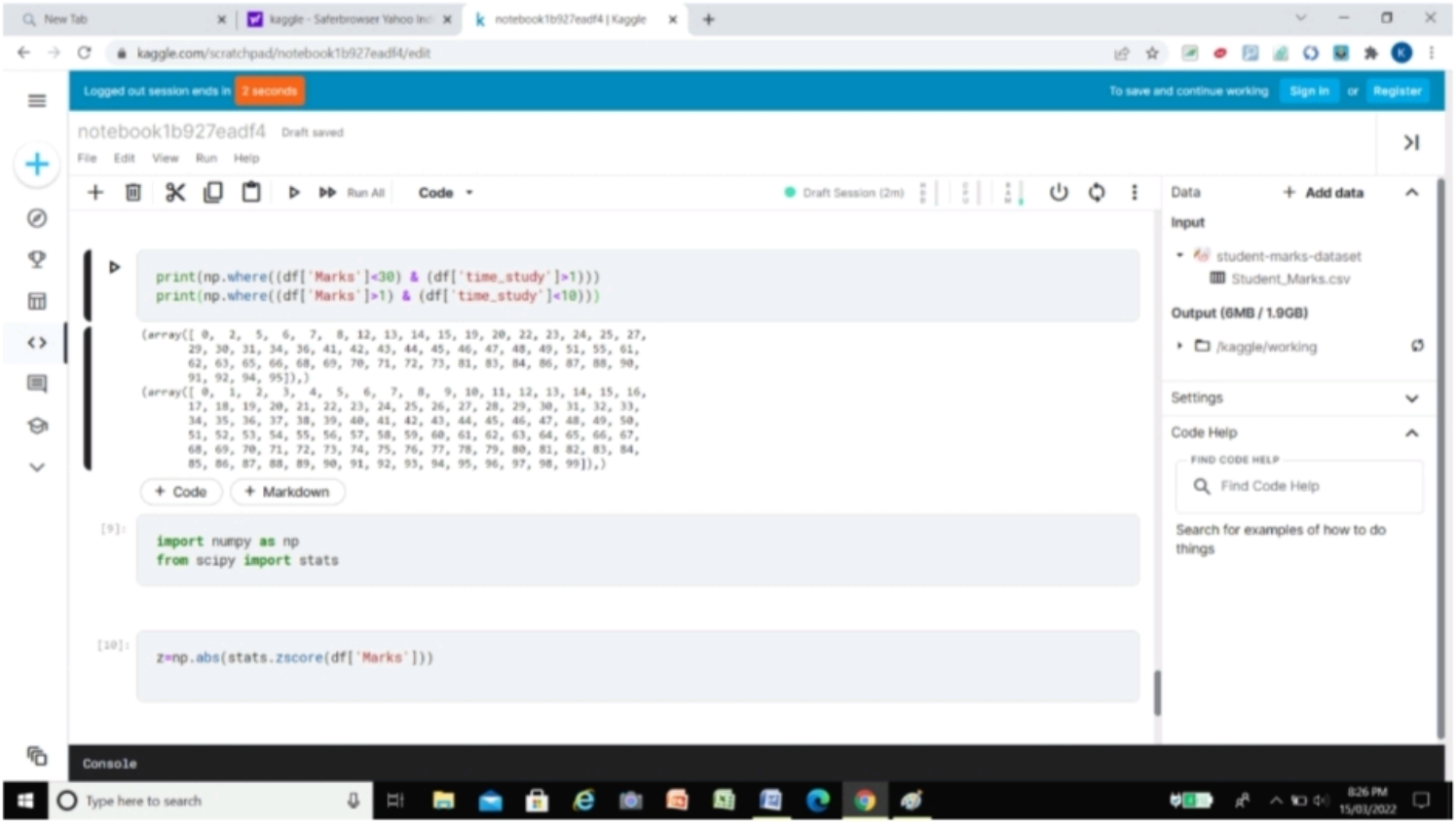
8:21 PM 15/03/2022

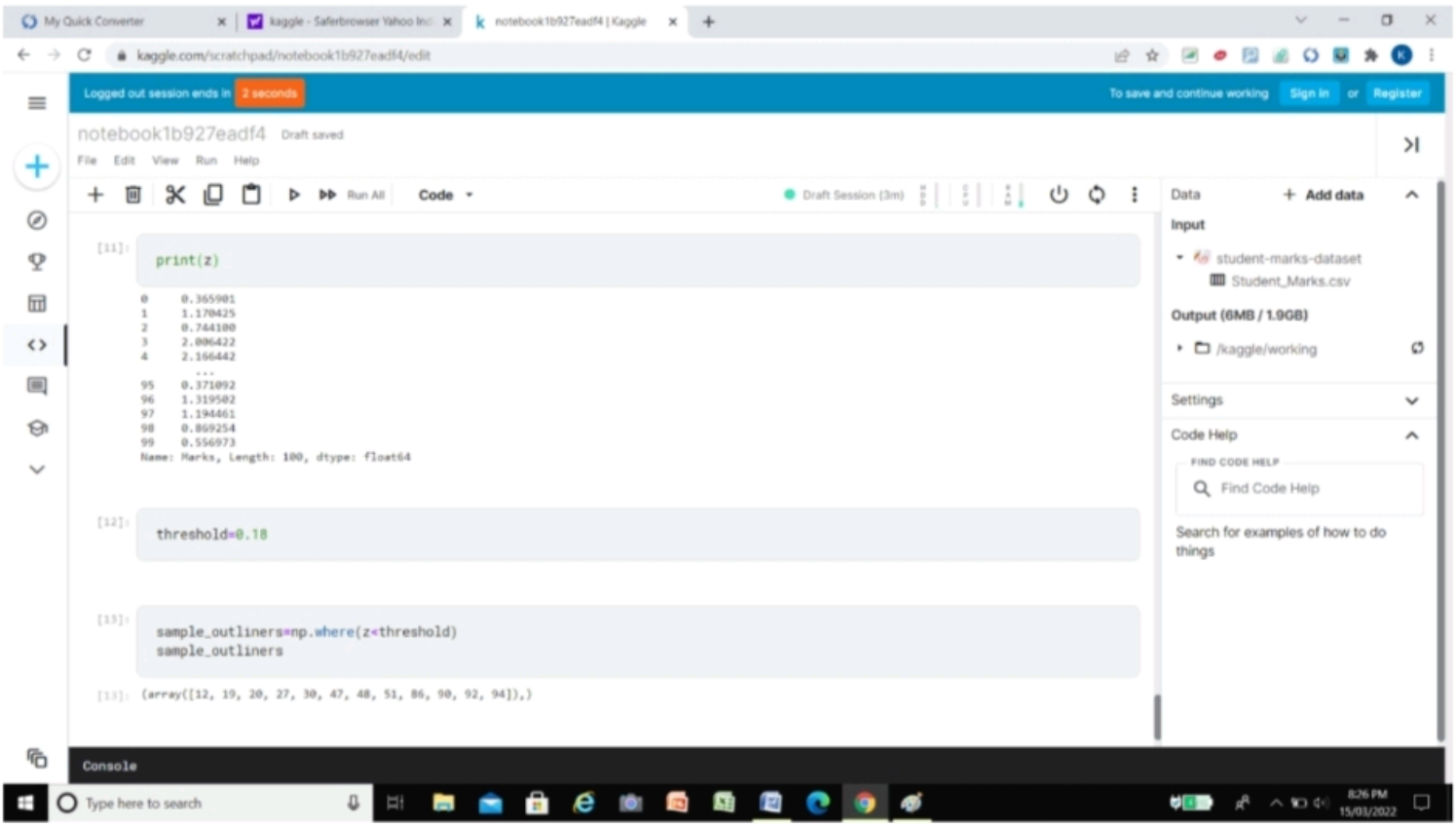

kaggle.com/scratchpad/notebook1b927eadf4/edit
Logged out session ends in 2 seconds
To save and continue working Sign In or Register
notebook1b927eadf4 Draft saved
File Edit View Run Help
Run All Code Draft Session off (run a cell to start)
Are you still there? Your notebook stops after 16 minutes of continuous use.
[4]:
print(np.where(df['Marks']>10))
print(np.where(df['time_study']>1))
print(np.where(df['number_courses']>3))
(array([0, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,
18, 19, 20, 21, 22, 23, 24, 26, 27, 30, 32, 33, 34, 35, 36, 37, 38,
39, 41, 42, 43, 46, 47, 48, 49, 50, 51, 53, 54, 55, 56, 57, 58, 59,
60, 61, 62, 63, 64, 65, 66, 69, 70, 71, 72, 74, 76, 77, 78, 80, 81,
82, 83, 84, 85, 86, 88, 89, 90, 91, 92, 93, 94, 95, 97, 98, 99]),)
(array([0, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 13, 14, 15, 16, 17, 18,
19, 20, 21, 22, 23, 24, 25, 27, 29, 30, 31, 33, 34, 35, 36, 37, 38,
39, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 53, 54, 55, 56, 57,
58, 59, 61, 62, 63, 65, 66, 68, 69, 70, 71, 72, 73, 74, 76, 77, 78,
80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 97,
99]),)
(array([1, 2, 3, 4, 5, 7, 8, 11, 12, 15, 16, 17, 18, 19, 20, 21, 22,
23, 24, 26, 27, 28, 29, 30, 32, 33, 36, 37, 38, 39, 40, 41, 42, 43,
44, 46, 47, 49, 50, 51, 52, 53, 54, 55, 56, 57, 59, 60, 61, 62, 64,
65, 66, 68, 69, 70, 71, 74, 76, 77, 78, 80, 81, 82, 83, 84, 85, 86,
88, 89, 90, 91, 92, 93, 94, 95, 97, 98]),)
[5]:
import matplotlib.pyplot as plt
Console
Data Add data
Input
student-marks-dataset
Student_Marks.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Type here to search
8:21 PM
15/03/2022

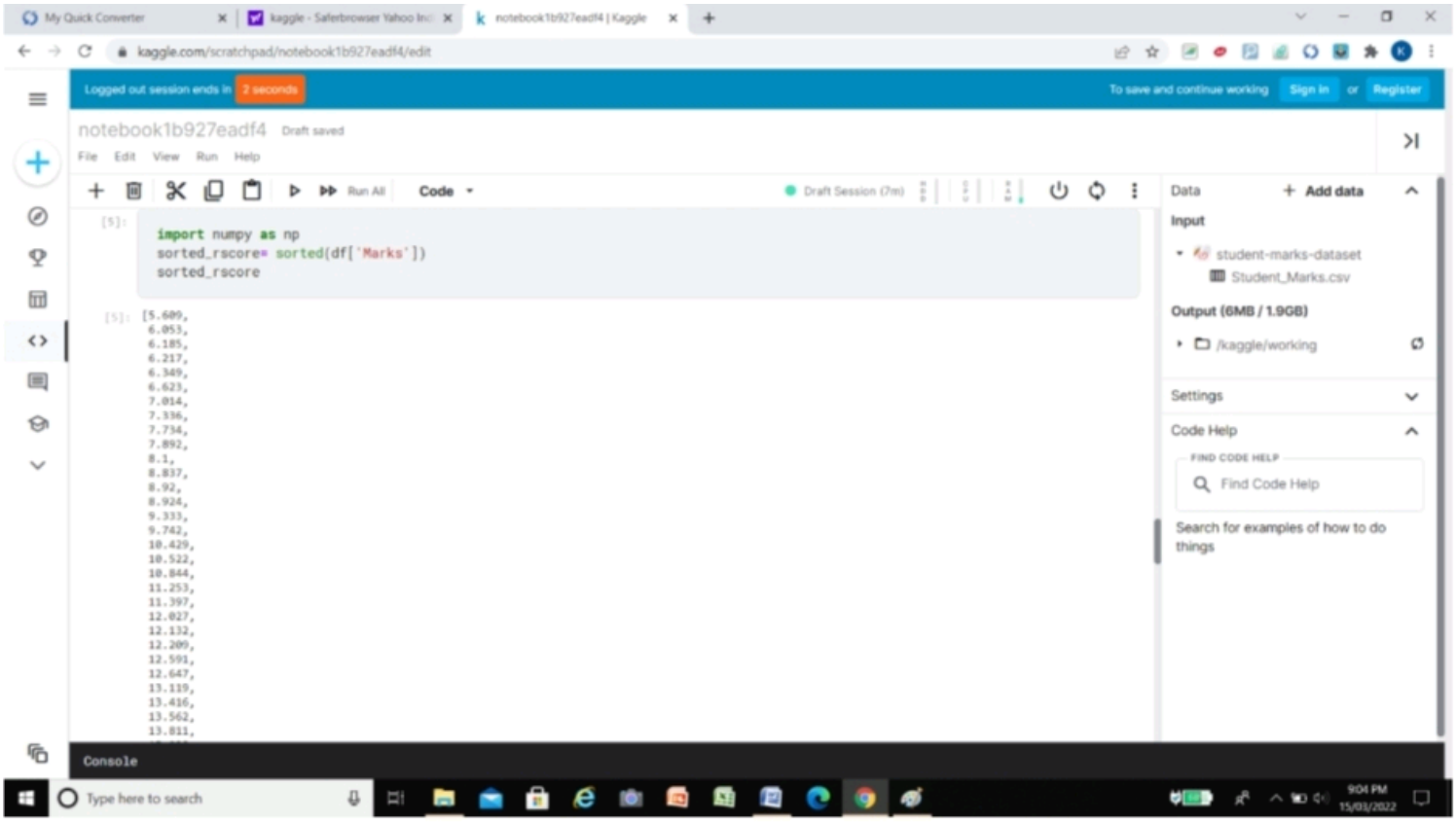

kaggle.com/scratchpad/notebook1b927eadf4/edit
Logged out session ends in 2 seconds
To save and continue working Sign In or Register
notebook1b927eadf4 Draft saved
File Edit View Run Help
Run All Code Draft Session off (run a cell to start)
Are you still there? Your notebook stops after 16 minutes of continuous use.
[6]:
df=pd.read_csv("../input/student-marks-dataset/Student_Marks.csv")
df
[6]:
number_courses time_study Marks
0 3 4.508 19.202
1 4 0.096 7.734
2 4 3.133 13.811
3 6 7.909 53.018
4 8 7.811 55.299
... ...
95 6 3.561 19.128
96 3 0.301 5.609
97 4 7.163 41.444
98 7 0.309 12.027
99 3 6.335 32.357
100 rows × 3 columns
Console
Data Add data
Input
student-marks-dataset
Student_Marks.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Type here to search
8:22 PM
15/03/2022

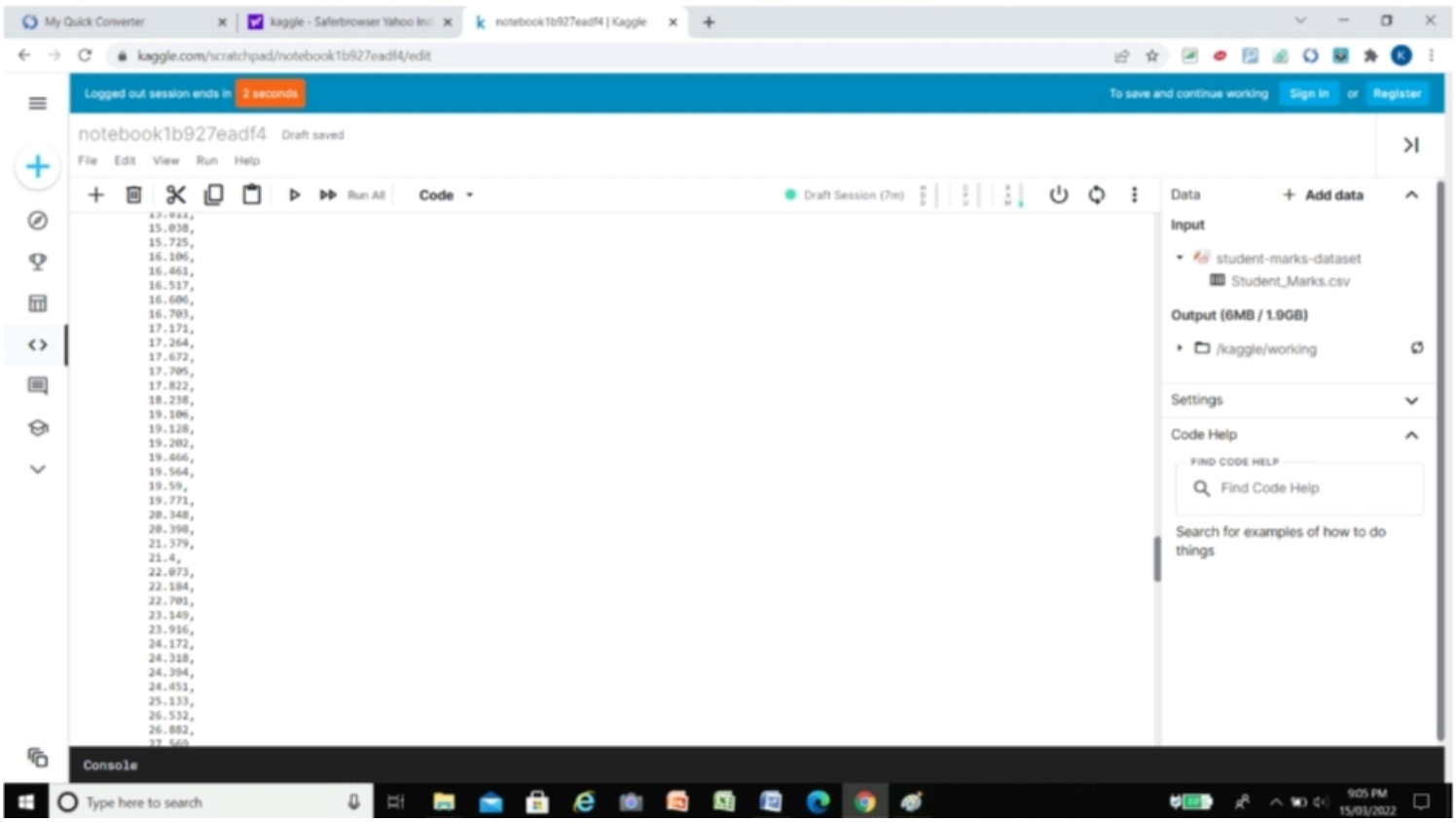

notebook1b927eadf4 | Kaggle
kaggle.com/scratchpad/notebook1b927eadf4/edit
Logged out session ends in 2 seconds
To save and continue working Sign In or Register
notebook1b927eadf4 Draft saved
File Edit View Run Help
Run All Code Draft Session (2m)
95 6 3.561 19.128
96 3 0.301 5.609
97 4 7.163 41.444
98 7 0.309 12.027
99 3 6.335 32.357
100 rows × 3 columns
[4]:
import numpy as np # linear algebra
import pandas as pd
df=pd.read_csv("../input/student-marks-dataset/Student_Marks.csv")
print("read the csv file")
import matplotlib.pyplot as plt
fig, ax=plt.subplots(figsize=(14,8))
ax.scatter(df['Marks'], df['time_study'])
plt.show
read the csv file
Console
Data Add data
Input
student-marks-dataset
Student_Marks.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Type here to search
8:26 PM
15/03/2022

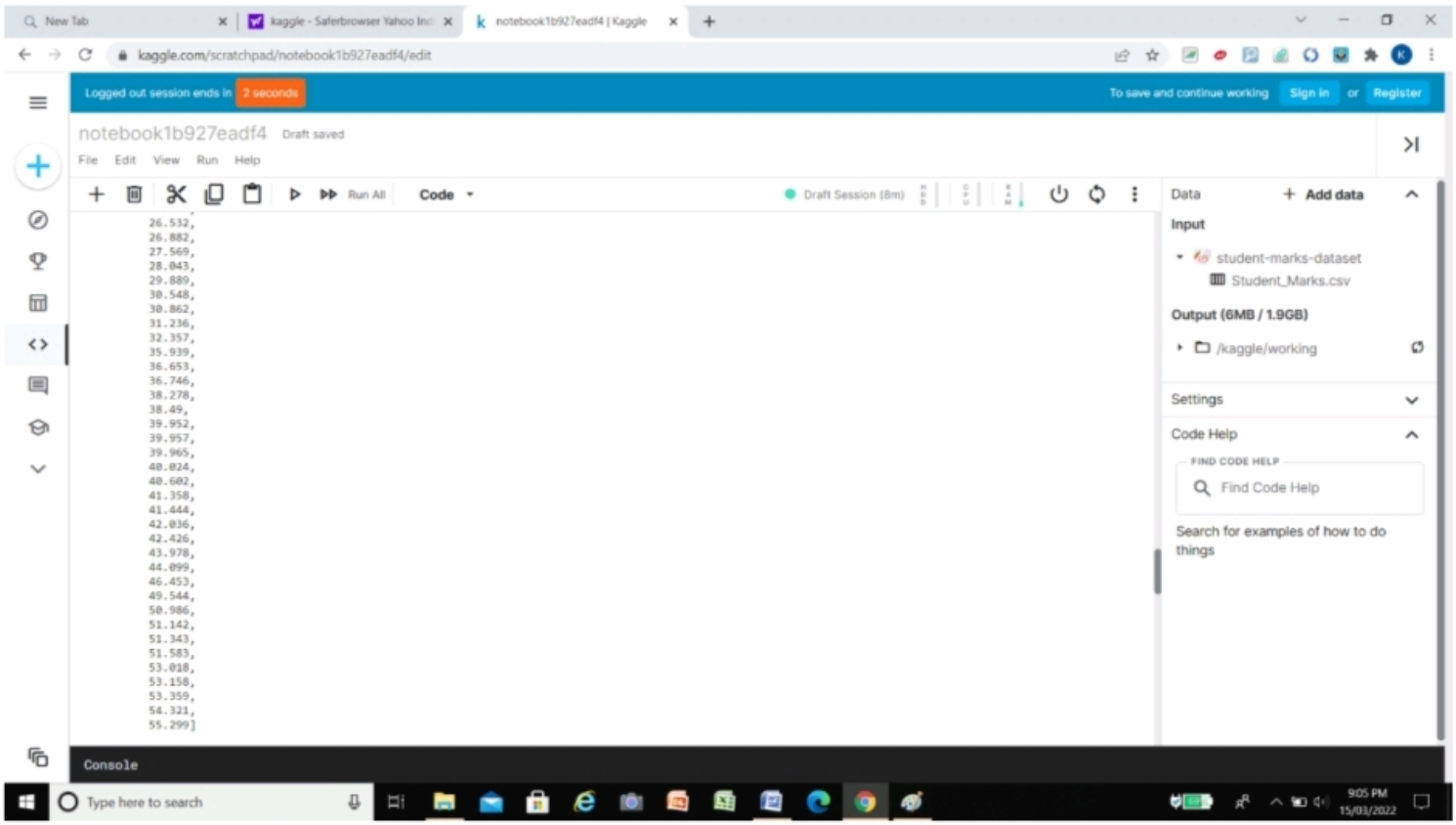

notebook1b927eadf4 | Kaggle
kaggle.com/scratchpad/notebook1b927eadf4/edit
Logged out session ends in 2 seconds
To save and continue working Sign In or Register
notebook1b927eadf4 Draft saved
File Edit View Run Help
Run All Code Draft Session (2m)
read the csv file
[4]: <function matplotlib.pyplot.show(close=None, block=None)>
Console
Data Add data
Input
student-marks-dataset
Student_Marks.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Type here to search
8:26 PM
15/03/2022

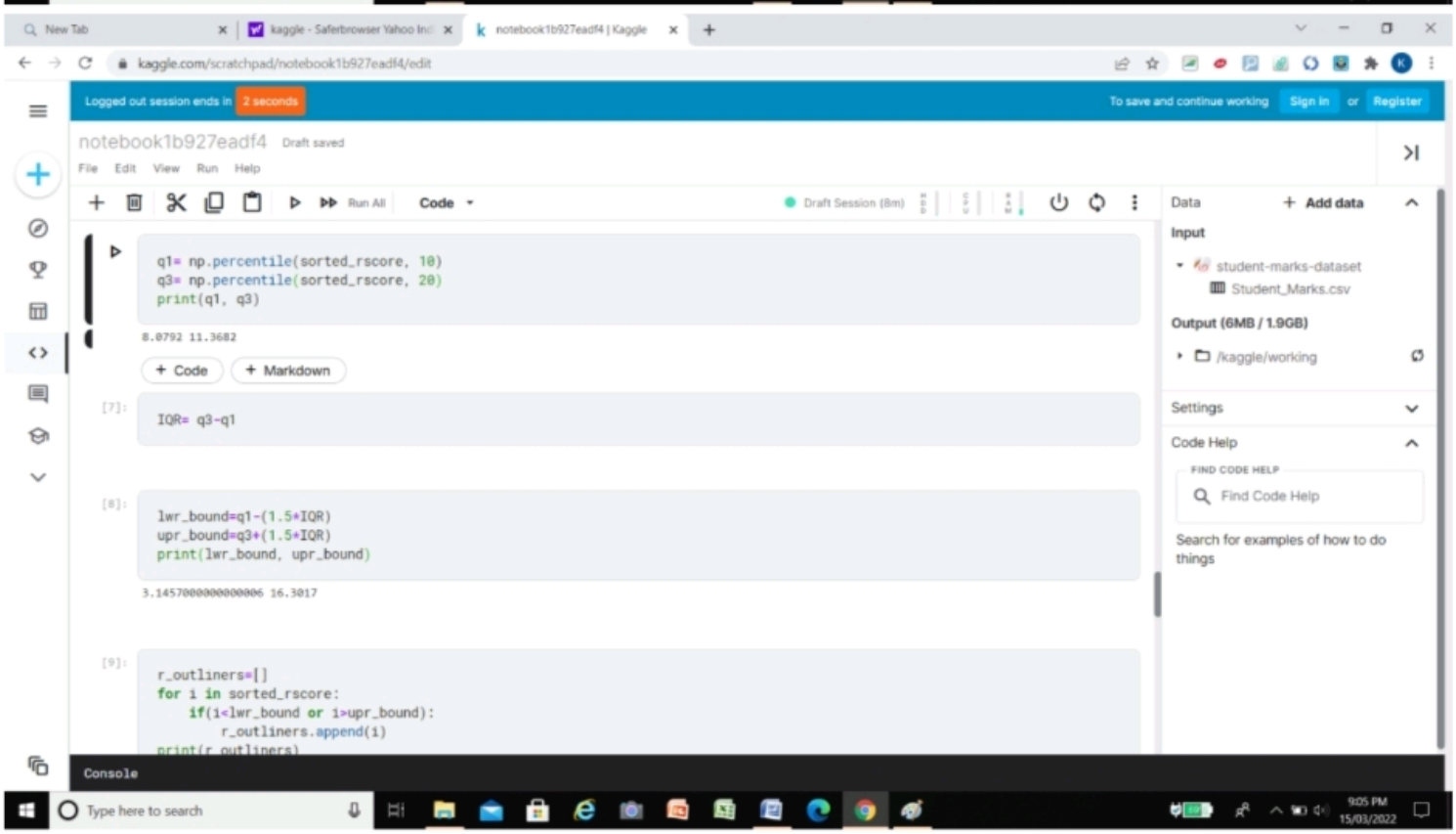

notebook1b927eadf4 | Kaggle
kaggle.com/scratchpad/notebook1b927eadf4/edit
Logged out session ends in 2 seconds
To save and continue working Sign In or Register
notebook1b927eadf4 Draft saved
File Edit View Run Help
Run All Code Draft Session (2m)
print(np.where((df['Marks']<30) & (df['time_study']>1)))
print(np.where((df['Marks']>1) & (df['time_study']<10)))
(array([0, 2, 5, 6, 7, 8, 12, 13, 14, 15, 19, 20, 22, 23, 24, 25, 27,
29, 30, 31, 34, 36, 41, 42, 43, 44, 45, 46, 47, 48, 49, 51, 55, 61,
62, 63, 65, 66, 68, 69, 70, 71, 72, 73, 81, 83, 84, 86, 87, 88, 90,
91, 92, 94, 95]),)
(array([0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,
17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33,
34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50,
51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67,
68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84,
85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99]),)
+ Code + Markdown
[9]: import numpy as np
from scipy import stats
[10]: z=np.abs(stats.zscore(df['Marks']))
Console
Data + Add data
Input
student-marks-dataset
Student_Marks.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Type here to search
8:26 PM
15/03/2022

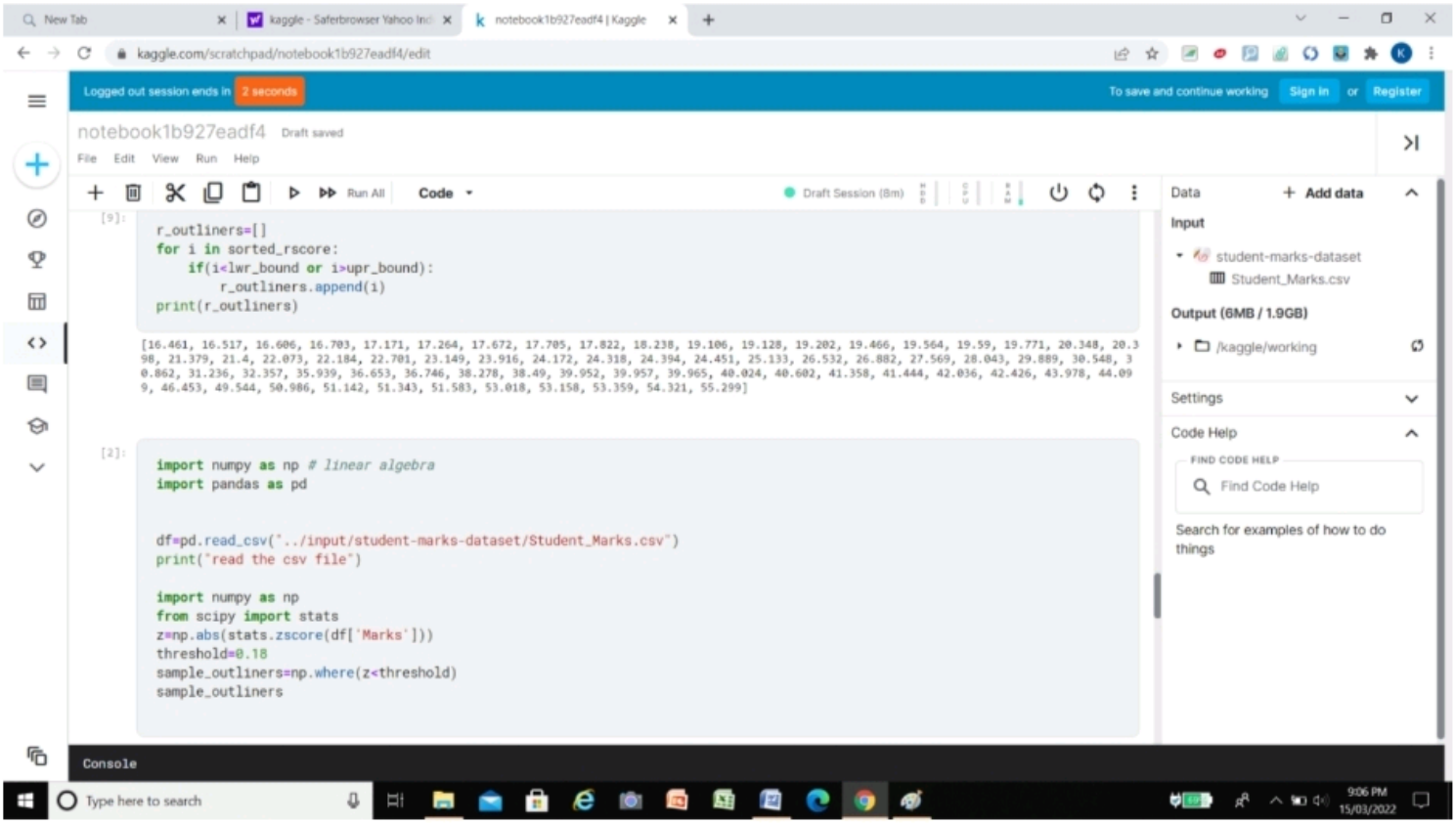

My Quick Converter | kaggle - Saferbrowser Yahoo Inc | notebook1b927eadf4 | Kaggle
kaggle.com/scratchpad/notebook1b927eadf4/edit
Logged out session ends in 2 seconds
To save and continue working Sign In or Register
notebook1b927eadf4 Draft saved
File Edit View Run Help
Run All Code Draft Session (3m)
[11]: print(z)
0 0.365901
1 1.170425
2 0.744100
3 2.006422
4 2.166442
...
95 0.371092
96 1.319502
97 1.194461
98 0.869254
99 0.556973
Name: Marks, Length: 100, dtype: float64
[12]: threshold=0.18
[13]: sample_outliners=np.where(z<threshold)
sample_outliners
[13]: (array([12, 19, 20, 27, 30, 47, 48, 51, 86, 90, 92, 94]),)
Console
Data + Add data
Input
student-marks-dataset
Student_Marks.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Type here to search
8:26 PM
15/03/2022

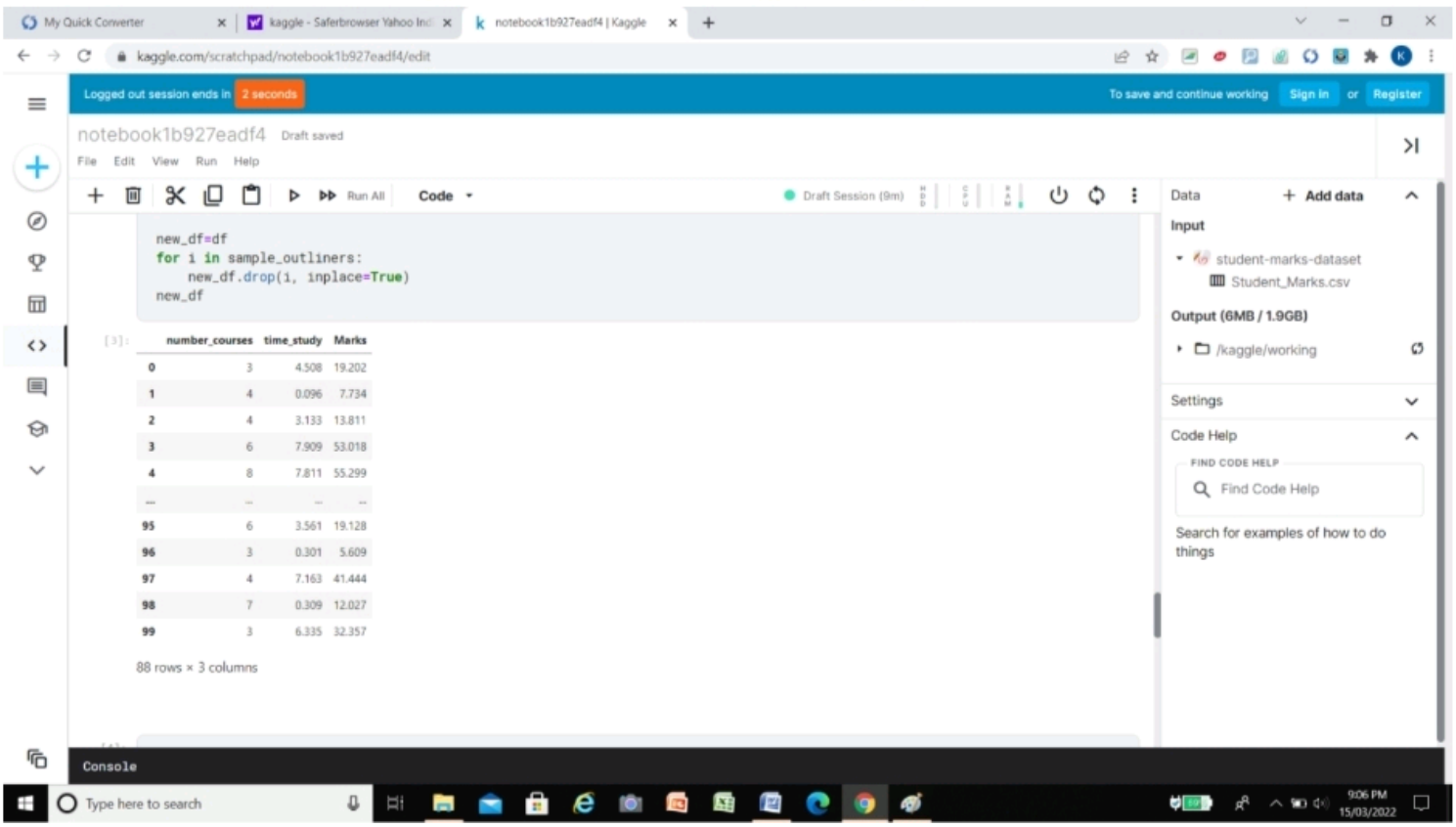

My Quick Converter
kaggle - Saferbrowser Yahoo Inc
notebook1b927eadf4 | Kaggle
kaggle.com/scratchpad/notebook1b927eadf4/edit
Logged out session ends in 2 seconds
To save and continue working Sign In or Register
notebook1b927eadf4 Draft saved
File Edit View Run Help
Run All Code Draft Session (7m)
[5]:
import numpy as np
sorted_rscore= sorted(df['Marks'])
sorted_rscore
[5]: [5.609,
6.053,
6.185,
6.217,
6.349,
6.623,
7.014,
7.336,
7.734,
7.892,
8.1,
8.837,
8.92,
8.924,
9.333,
9.742,
10.429,
10.522,
10.844,
11.253,
11.397,
12.027,
12.132,
12.209,
12.591,
12.647,
13.119,
13.416,
13.562,
13.811,
Data Add data
Input
student-marks-dataset
Student_Marks.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Console
Type here to search
9:04 PM
15/03/2022

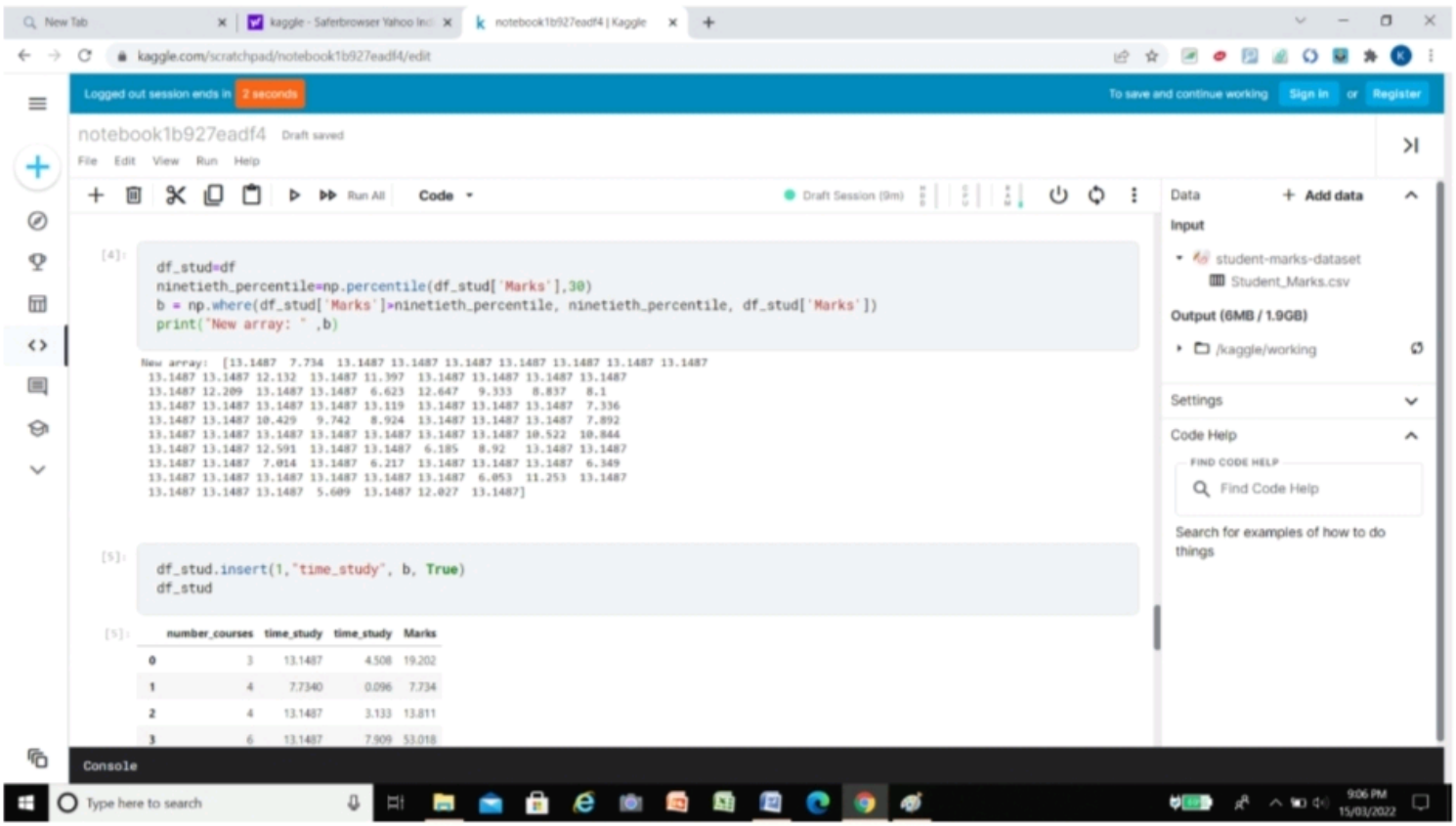

My Quick Converter
kaggle - Saferbrowser Yahoo Inc
notebook1b927eadf4 | Kaggle
kaggle.com/scratchpad/notebook1b927eadf4/edit
Logged out session ends in 2 seconds
To save and continue working Sign In or Register
notebook1b927eadf4 Draft saved
File Edit View Run Help
Run All Code Draft Session (7m)
15.038,
15.725,
16.106,
16.461,
16.517,
16.606,
16.703,
17.171,
17.264,
17.672,
17.705,
17.822,
18.238,
19.106,
19.128,
19.202,
19.466,
19.564,
19.59,
19.771,
20.348,
20.398,
21.379,
21.4,
22.073,
22.184,
22.701,
23.149,
23.936,
24.172,
24.318,
24.394,
24.451,
25.133,
26.532,
26.882,
Data Add data
Input
student-marks-dataset
Student_Marks.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Console
Type here to search
9:05 PM
15/03/2022

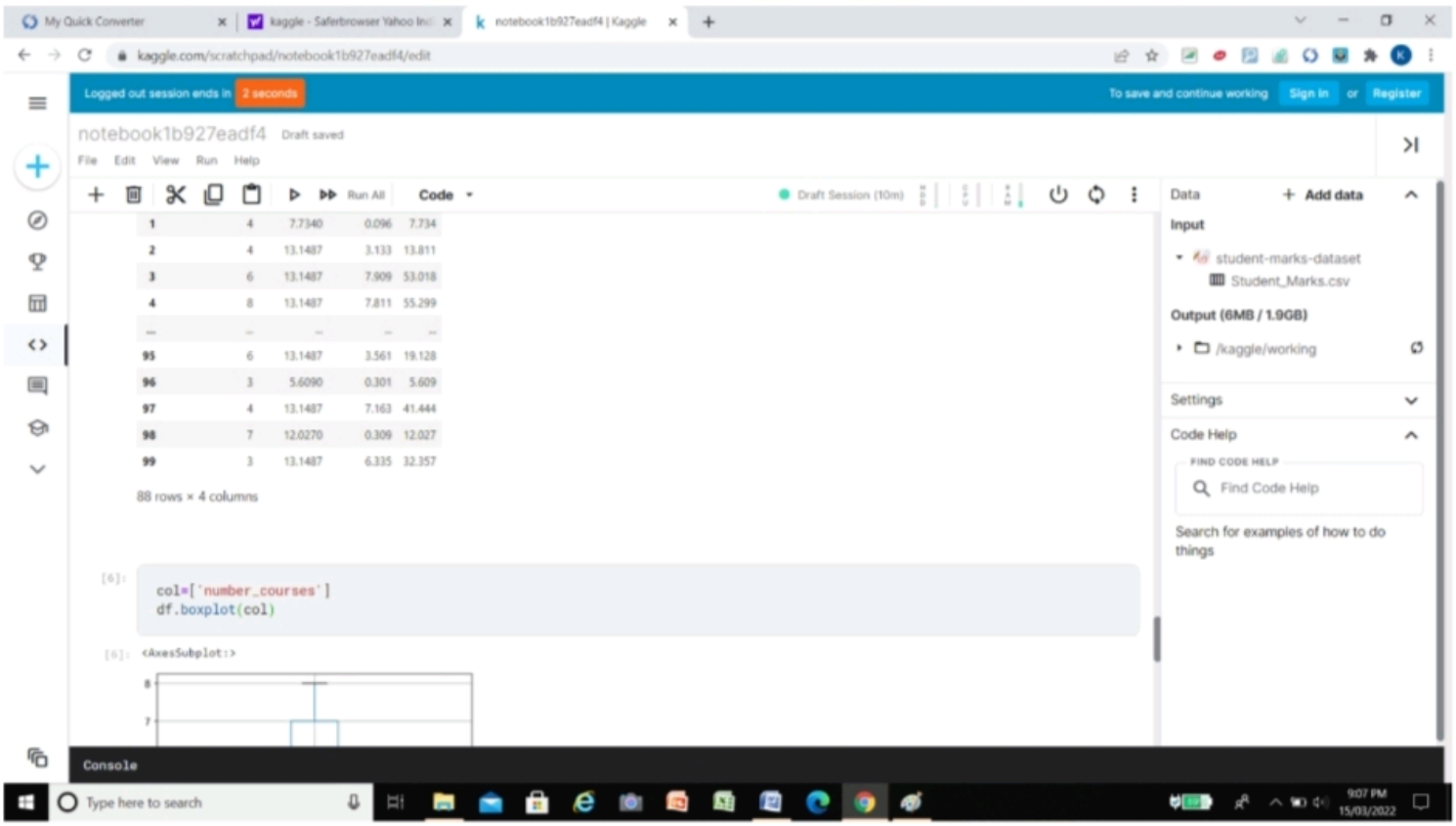

New Tab
kaggle - Saferbrowser Yahoo Inc
notebook1b927eadf4 | Kaggle
kaggle.com/scratchpad/notebook1b927eadf4/edit
Logged out session ends in 2 seconds
To save and continue working Sign In or Register
notebook1b927eadf4 Draft saved
File Edit View Run Help
Run All
Code
Draft Session (8m)
26.532,
26.882,
27.569,
28.043,
29.889,
30.548,
30.862,
31.236,
32.357,
35.939,
36.653,
36.746,
38.278,
38.49,
39.952,
39.957,
39.965,
40.024,
40.602,
41.358,
41.444,
42.036,
42.426,
43.978,
44.099,
46.453,
49.544,
50.986,
51.142,
51.343,
51.583,
53.018,
53.158,
53.359,
54.321,
55.299]
Data
Add data
Input
student-marks-dataset
Student_Marks.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Console
Type here to search
9:05 PM
15/03/2022


New Tab
kaggle - Saferbrowser Yahoo Inc
notebook1b927eadf4 | Kaggle
kaggle.com/scratchpad/notebook1b927eadf4/edit
Logged out session ends in 2 seconds
To save and continue working Sign In or Register
notebook1b927eadf4 Draft saved
File Edit View Run Help
Run All
Code
Draft Session (8m)
q1= np.percentile(sorted_rscore, 10)
q3= np.percentile(sorted_rscore, 20)
print(q1, q3)
8.0792 11.3682
+ Code
+ Markdown
[7]:
IQR= q3-q1
[8]:
lwr_bound=q1-(1.5*IQR)
upr_bound=q3+(1.5*IQR)
print(lwr_bound, upr_bound)
3.1457000000000006 16.3017
[9]:
r_outliners=[]
for i in sorted_rscore:
if(i<lwr_bound or i>upr_bound):
r_outliners.append(i)
print(r_outliners)
Data
Add data
Input
student-marks-dataset
Student_Marks.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Console
Type here to search
9:05 PM
15/03/2022


notebook1b927eadf4 | Kaggle
kaggle.com/scratchpad/notebook1b927eadf4/edit
Logged out session ends in 2 seconds
To save and continue working Sign In or Register
notebook1b927eadf4 Draft saved
File Edit View Run Help
Run All Code Draft Session (8m)
[9]:
r_outliners=[]
for i in sorted_rscore:
if(i<lwr_bound or i>upr_bound):
r_outliners.append(i)
print(r_outliners)
[16.461, 16.517, 16.606, 16.703, 17.171, 17.264, 17.672, 17.705, 17.822, 18.238, 19.106, 19.128, 19.202, 19.466, 19.564, 19.59, 19.771, 20.348, 20.398, 21.379, 21.4, 22.073, 22.184, 22.701, 23.149, 23.916, 24.172, 24.318, 24.394, 24.451, 25.133, 26.532, 26.882, 27.569, 28.043, 29.889, 30.548, 30.862, 31.236, 32.357, 35.939, 36.653, 36.746, 38.278, 38.49, 39.952, 39.957, 39.965, 40.024, 40.602, 41.358, 41.444, 42.036, 42.426, 43.978, 44.099, 46.453, 49.544, 50.986, 51.142, 51.343, 51.583, 53.018, 53.158, 53.359, 54.321, 55.299]
[2]:
import numpy as np # linear algebra
import pandas as pd
df=pd.read_csv("../input/student-marks-dataset/Student_Marks.csv")
print("read the csv file")
import numpy as np
from scipy import stats
z=np.abs(stats.zscore(df['Marks']))
threshold=0.18
sample_outliners=np.where(z<threshold)
sample_outliners
Data + Add data
Input
student-marks-dataset
Student_Marks.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Console
Type here to search
9:06 PM
15/03/2022


My Quick Converter
kaggle - Saferbrowser Yahoo Inc
notebook1b927eadf4 | Kaggle
kaggle.com/scratchpad/notebook1b927eadf4/edit
Logged out session ends in 2 seconds
To save and continue working Sign In or Register
notebook1b927eadf4 Draft saved
File Edit View Run Help
Run All Code Draft Session (9m)
new_df=df
for i in sample_outliners:
new_df.drop(i, inplace=True)
new_df
[3]:
number_courses time_study Marks
0 3 4.508 19.202
1 4 0.096 7.734
2 4 3.133 13.811
3 6 7.909 53.018
4 8 7.811 55.299
95 6 3.561 19.128
96 3 0.301 5.609
97 4 7.163 41.444
98 7 0.309 12.027
99 3 6.335 32.357
88 rows × 3 columns
Data + Add data
Input
student-marks-dataset
Student_Marks.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Console
Type here to search
9:06 PM
15/03/2022


notebook1b927eadf4 Draft saved
File Edit View Run Help
Run All Code
Draft Session (9m)
Data Add data
Input
student-marks-dataset
Student_Marks.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
[4]:
df_stud=df
ninetieth_percentile=np.percentile(df_stud['Marks'],30)
b = np.where(df_stud['Marks']>ninetieth_percentile, ninetieth_percentile, df_stud['Marks'])
print("New array: " ,b)
New array: [13.1487 7.734 13.1487 13.1487 13.1487 13.1487 13.1487 13.1487 13.1487 13.1487
13.1487 13.1487 12.132 13.1487 11.397 13.1487 13.1487 13.1487 13.1487
13.1487 12.209 13.1487 13.1487 6.623 12.647 9.333 8.837 8.1
13.1487 13.1487 13.1487 13.1487 13.119 13.1487 13.1487 13.1487 7.336
13.1487 13.1487 10.429 9.742 8.924 13.1487 13.1487 13.1487 7.892
13.1487 13.1487 13.1487 13.1487 13.1487 13.1487 13.1487 10.522 10.844
13.1487 13.1487 12.591 13.1487 13.1487 6.185 8.92 13.1487 13.1487
13.1487 13.1487 7.014 13.1487 6.217 13.1487 13.1487 13.1487 6.349
13.1487 13.1487 13.1487 13.1487 13.1487 13.1487 6.053 11.253 13.1487
13.1487 13.1487 13.1487 5.609 13.1487 12.027 13.1487]
[5]:
df_stud.insert(1,"time_study", b, True)
df_stud
[5]:
number_courses time_study time_study Marks
0 3 13.1487 4.508 19.202
1 4 7.7340 0.096 7.734
2 4 13.1487 3.133 13.811
3 6 13.1487 7.909 53.018
Console
Type here to search
9:06 PM
15/03/2022


notebook1b927eadf4 Draft saved
File Edit View Run Help
Run All Code
Draft Session (10m)
Data Add data
Input
student-marks-dataset
Student_Marks.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
1 4 7.7340 0.096 7.734
2 4 13.1487 3.133 13.811
3 6 13.1487 7.909 53.018
4 8 13.1487 7.811 55.299
... ...
95 6 13.1487 3.561 19.128
96 3 5.6090 0.301 5.609
97 4 13.1487 7.163 41.444
98 7 12.0270 0.309 12.027
99 3 13.1487 6.335 32.357
88 rows × 4 columns
[6]:
col=['number_courses']
df.boxplot(col)
[6]:
<AxesSubplot:>
8
7
Console
Type here to search
9:07 PM
15/03/2022

My Quick Converter | kaggle - Saferbrowser Yahoo Inc | notebook1b927eadf4 | Kaggle

kaggle.com/scratchpad/notebook1b927eadf4/edit

Logged out session ends in 2 seconds — To save and continue working Sign In or Register

notebook1b927eadf4 Draft saved

File Edit View Run Help

Run All | Code | Draft Session (10m)

Data | + Add data

Input

student-marks-dataset

Student_Marks.csv

Output (6MB / 1.9GB)

/kaggle/working

Settings

Code Help

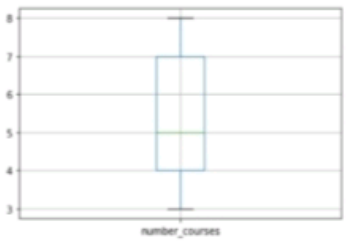
FIND CODE HELP

Find Code Help

Search for examples of how to do things

[6]:
```
col=['number_courses']
df.boxplot(col)
```

[6]:
```
<AxesSubplot:>
```



[8]:
```
import numpy as np
sorted_rscore= sorted(df['Marks'])
sorted_rscore
median=np.median(sorted_rscore)
median
```

Console

Type here to search

9:07 PM 15/03/2022

My Quick Converter | kaggle - Saferbrowser Yahoo Inc | notebook1b927eadf4 | Kaggle

kaggle.com/scratchpad/notebook1b927eadf4/edit

Logged out session ends in 2 seconds — To save and continue working Sign In or Register

notebook1b927eadf4 Draft saved

File Edit View Run Help

Run All | Code | Draft Session (10m)

Data | + Add data

Input

student-marks-dataset

Student_Marks.csv

Output (6MB / 1.9GB)

/kaggle/working

Settings

Code Help

FIND CODE HELP

Find Code Help

Search for examples of how to do things

[8]:
```
19.117
```

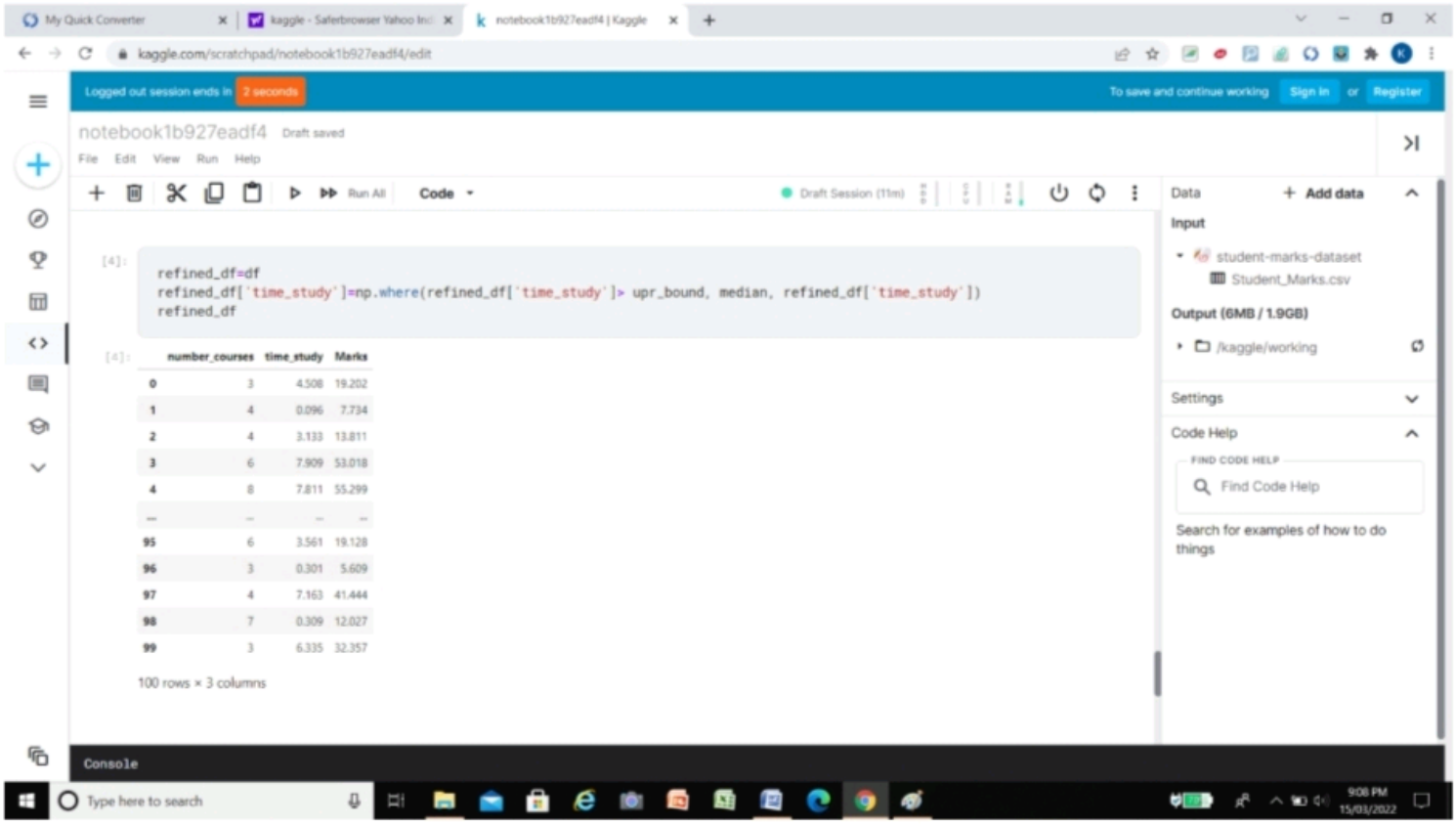
[3]:
```
import numpy as np # linear algebra
import pandas as pd


df=pd.read_csv("../input/student-marks-dataset/Student_Marks.csv")
print("read the csv file")
import numpy as np
sorted_rscore= sorted(df['Marks'])
sorted_rscore
median=np.median(sorted_rscore)
median
q1= np.percentile(sorted_rscore, 10)
q3= np.percentile(sorted_rscore, 20)
print(q1, q3)
IQR= q3-q1
lwr_bound=q1-(1.5*IQR)
upr_bound=q3+(1.5*IQR)
print(lwr_bound, upr_bound)
```

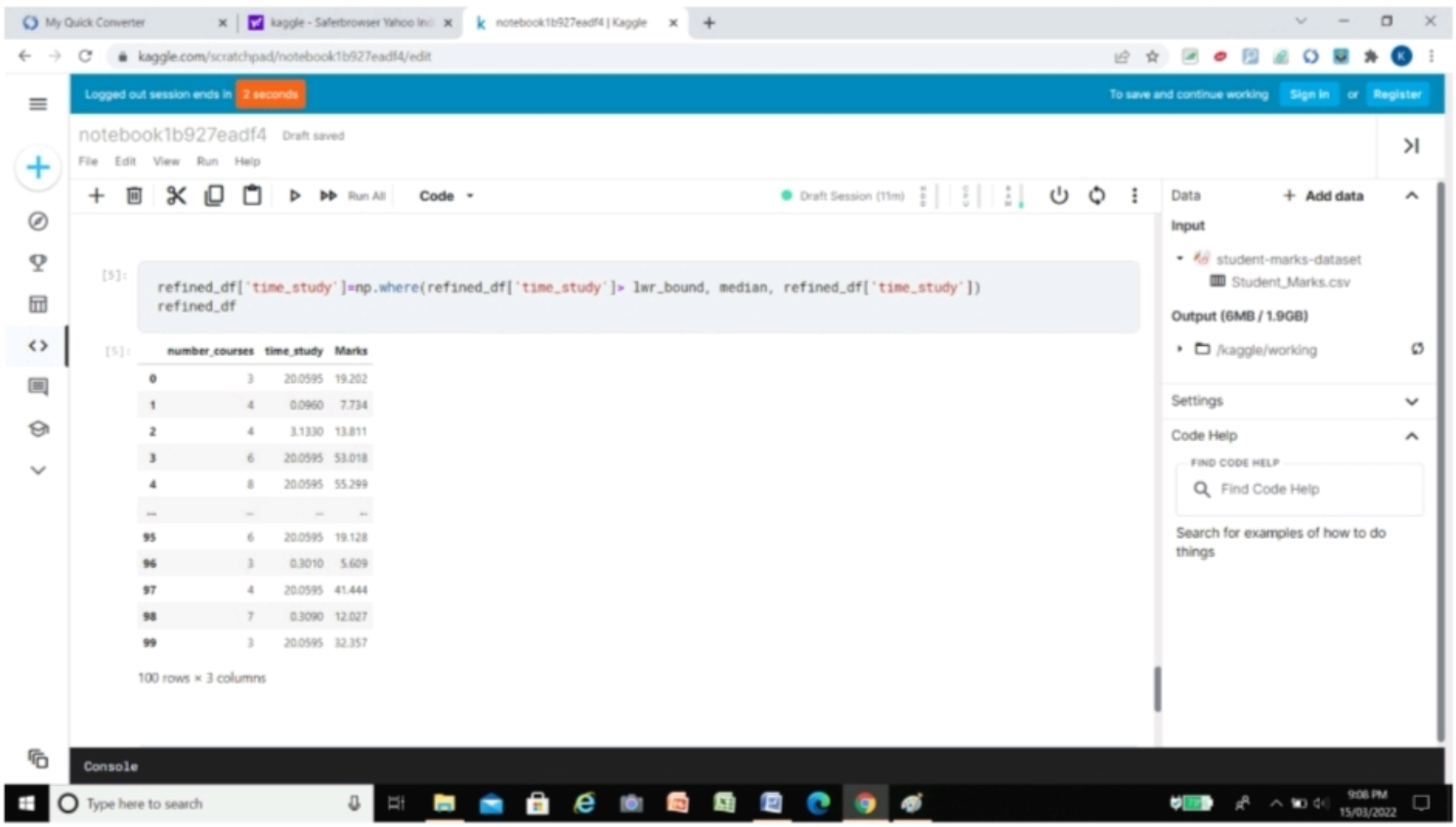
```
read the csv file
8.0792 11.3682
3.1457000000000006 16.3017
```

Console

Type here to search

9:08 PM 15/03/2022


My Quick Converter
kaggle - Saferbrowser Yahoo Inc
notebook1b927eadf4 | Kaggle
kaggle.com/scratchpad/notebook1b927eadf4/edit
Logged out session ends in 2 seconds
To save and continue working Sign In or Register
notebook1b927eadf4 Draft saved
File Edit View Run Help
Run All
Code
Draft Session (11m)
Data
Add data
Input
student-marks-dataset
Student_Marks.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
[4]:
refined_df=df
refined_df['time_study']=np.where(refined_df['time_study']> upr_bound, median, refined_df['time_study'])
refined_df
[4]:
number_courses time_study Marks
0 3 4.508 19.202
1 4 0.096 7.734
2 4 3.133 13.811
3 6 7.909 53.018
4 8 7.811 55.299
95 6 3.561 19.128
96 3 0.301 5.609
97 4 7.163 41.444
98 7 0.309 12.027
99 3 6.335 32.357
100 rows × 3 columns
Console
Type here to search
9:08 PM
15/03/2022


My Quick Converter
kaggle - Saferbrowser Yahoo Inc
notebook1b927eadf4 | Kaggle
kaggle.com/scratchpad/notebook1b927eadf4/edit
Logged out session ends in 2 seconds
To save and continue working Sign In or Register
notebook1b927eadf4 Draft saved
File Edit View Run Help
Run All
Code
Draft Session (11m)
Data
Add data
Input
student-marks-dataset
Student_Marks.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
[5]:
refined_df['time_study']=np.where(refined_df['time_study']> lwr_bound, median, refined_df['time_study'])
refined_df
[5]:
number_courses time_study Marks
0 3 20.0595 19.202
1 4 0.0960 7.734
2 4 3.1330 13.811
3 6 20.0595 53.018
4 8 20.0595 55.299
95 6 20.0595 19.128
96 3 0.3010 5.609
97 4 20.0595 41.444
98 7 0.3090 12.027
99 3 20.0595 32.357
100 rows × 3 columns
Console
Type here to search
9:08 PM
15/03/2022

New Tab | kaggle - Saferbrowser Yahoo Inc | notebook1b927eadf4 | Kaggle

kaggle.com/scratchpad/notebook1b927eadf4/edit

Logged out session ends in 2 seconds — To save and continue working Sign In or Register

notebook1b927eadf4 Draft saved

File Edit View Run Help

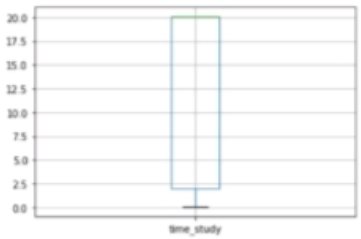
Run All | Code | Draft Session (11m)

99	3	20.0595	32.357

100 rows × 3 columns

[6]:
```
col=['time_study']
refined_df.boxplot(col)
```

[6]:
```
<AxesSubplot:>
```



[8]:

Console

Data | + Add data

Input

student-marks-dataset

Student_Marks.csv

Output (6MB / 1.9GB)

/kaggle/working

Settings

Code Help

FIND CODE HELP

Find Code Help

Search for examples of how to do things

Type here to search

9:08 PM 15/03/2022

New Tab | kaggle - Saferbrowser Yahoo Inc | notebook1b927eadf4 | Kaggle

kaggle.com/scratchpad/notebook1b927eadf4/edit

Logged out session ends in 2 seconds — To save and continue working Sign In or Register

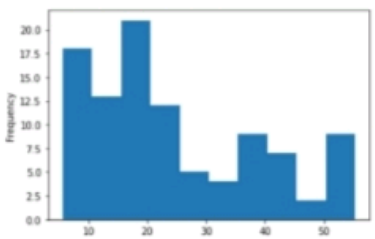
notebook1b927eadf4 Draft saved

File Edit View Run Help

Run All | Code | Draft Session (11m)

[8]:
```
import matplotlib.pyplot as plt
new_df=df
new_df['Marks'].plot(kind='hist')
```

[8]:
```
<AxesSubplot:ylabel='Frequency'>
```



[10]:
```
df['log_math']=np.log10(df['Marks'])
df['log_math'].plot(kind='hist')
```

Console

Data | + Add data

Input

student-marks-dataset

Student_Marks.csv

Output (6MB / 1.9GB)

/kaggle/working

Settings

Code Help

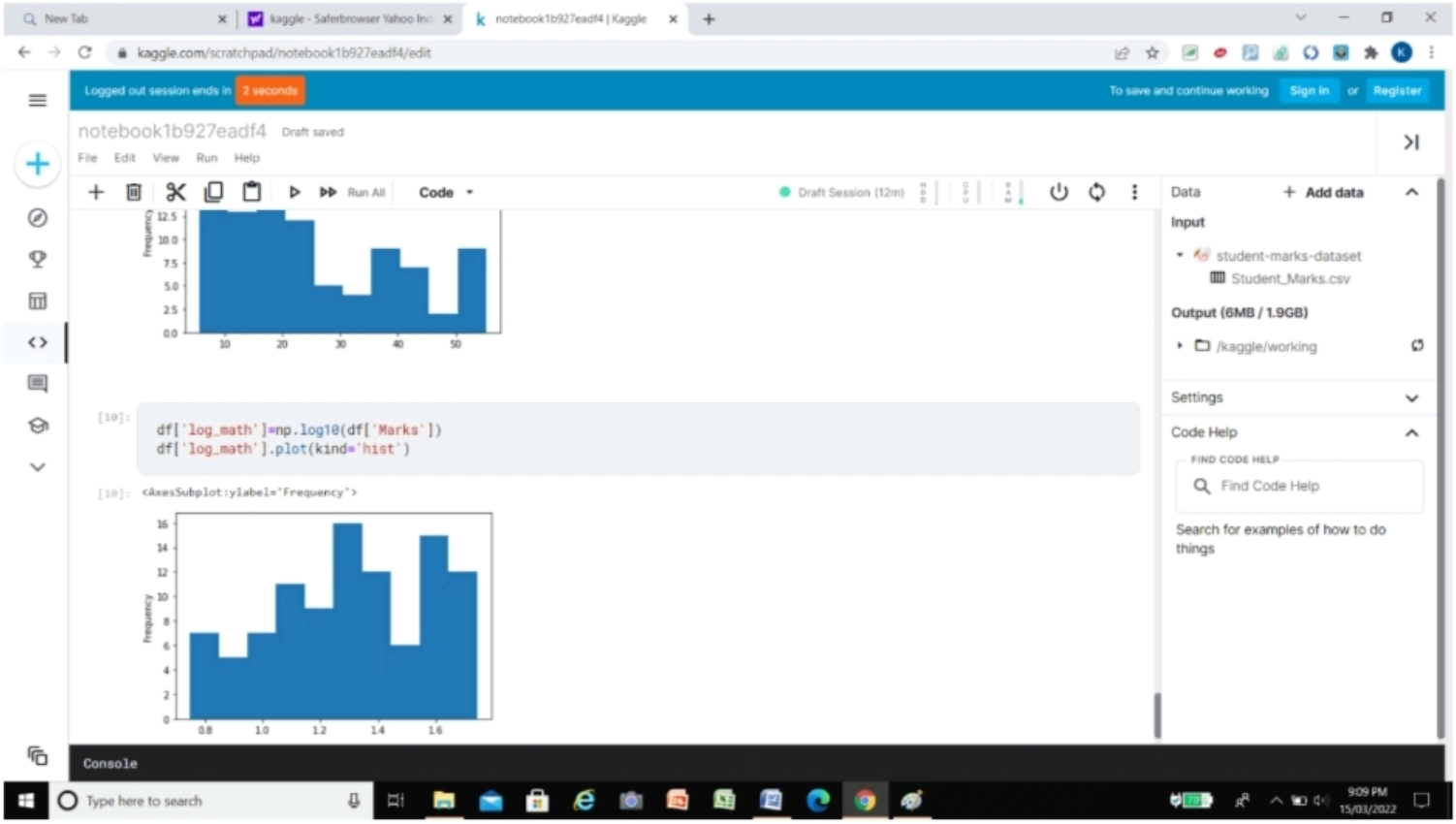
FIND CODE HELP

Find Code Help

Search for examples of how to do things

Type here to search

9:09 PM 15/03/2022


New Tab
kaggle - Saferbrowser Yahoo Ind
notebook1b927eadf4 | Kaggle
kaggle.com/scratchpad/notebook1b927eadf4/edit
Logged out session ends in 2 seconds
To save and continue working Sign in or Register
notebook1b927eadf4 Draft saved
File Edit View Run Help
Run All
Code
Draft Session (12m)
Frequency
12.5
10.0
7.5
5.0
2.5
0.0
10
20
30
40
50
[10]:
df['log_math']=np.log10(df['Marks'])
df['log_math'].plot(kind='hist')
[10]: <AxesSubplot:ylabel='Frequency'>
Frequency
16
14
12
10
8
6
4
2
0
0.8
1.0
1.2
1.4
1.6
Data
Add data
Input
student-marks-dataset
Student_Marks.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Console
Type here to search
9:09 PM
15/03/2022

Write-up	Correctness of Program	Documentation of Program	Viva	Timely Completion	Total	Dated Sign of Subject Teacher
4	4	4	4	4	20	

Group A

Assignment No. 3

Title of the Assignment: Descriptive Statistics - Measures of Central Tendency and variability

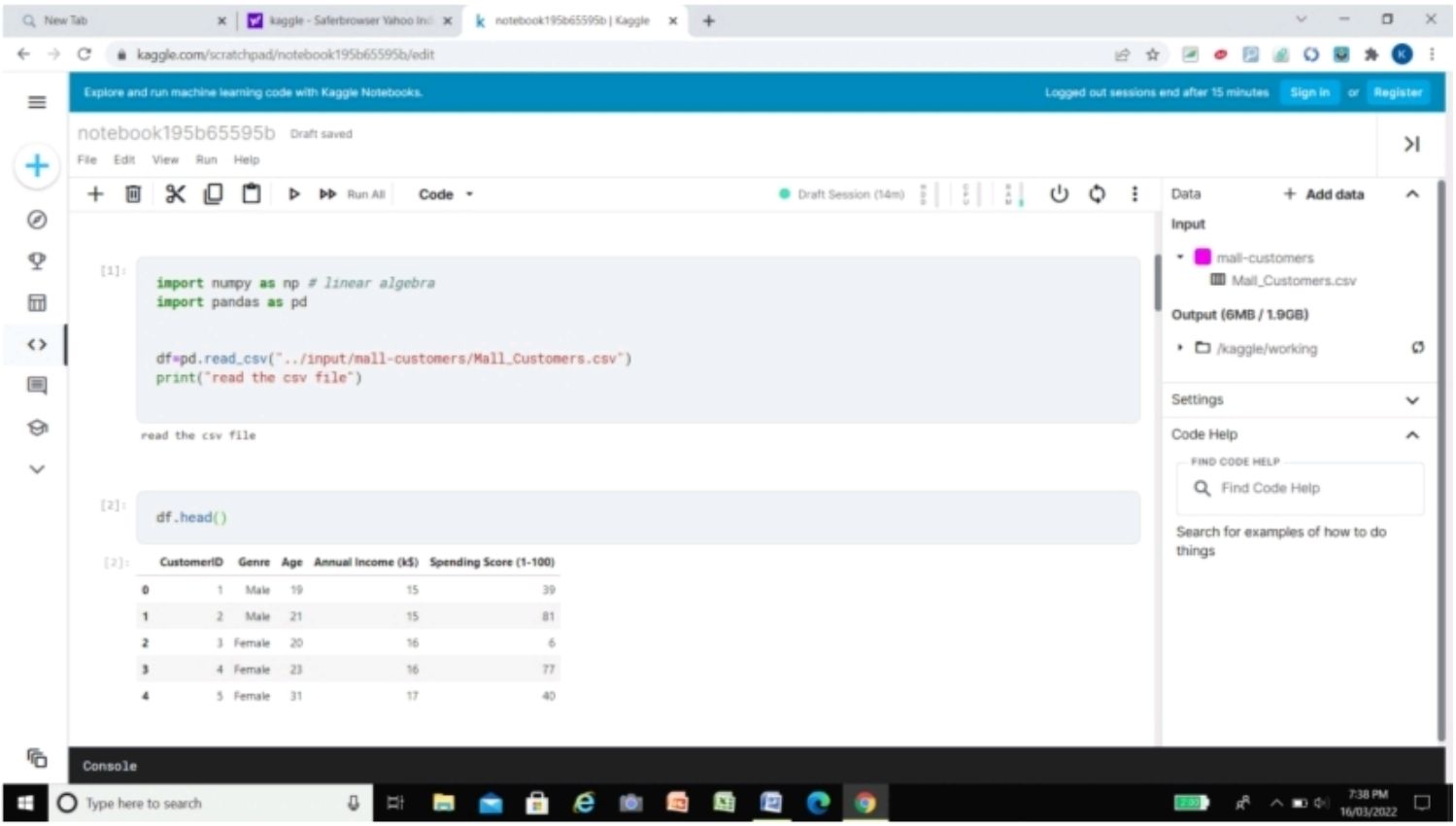
Perform the following operations on any open source dataset (e.g., data.csv)

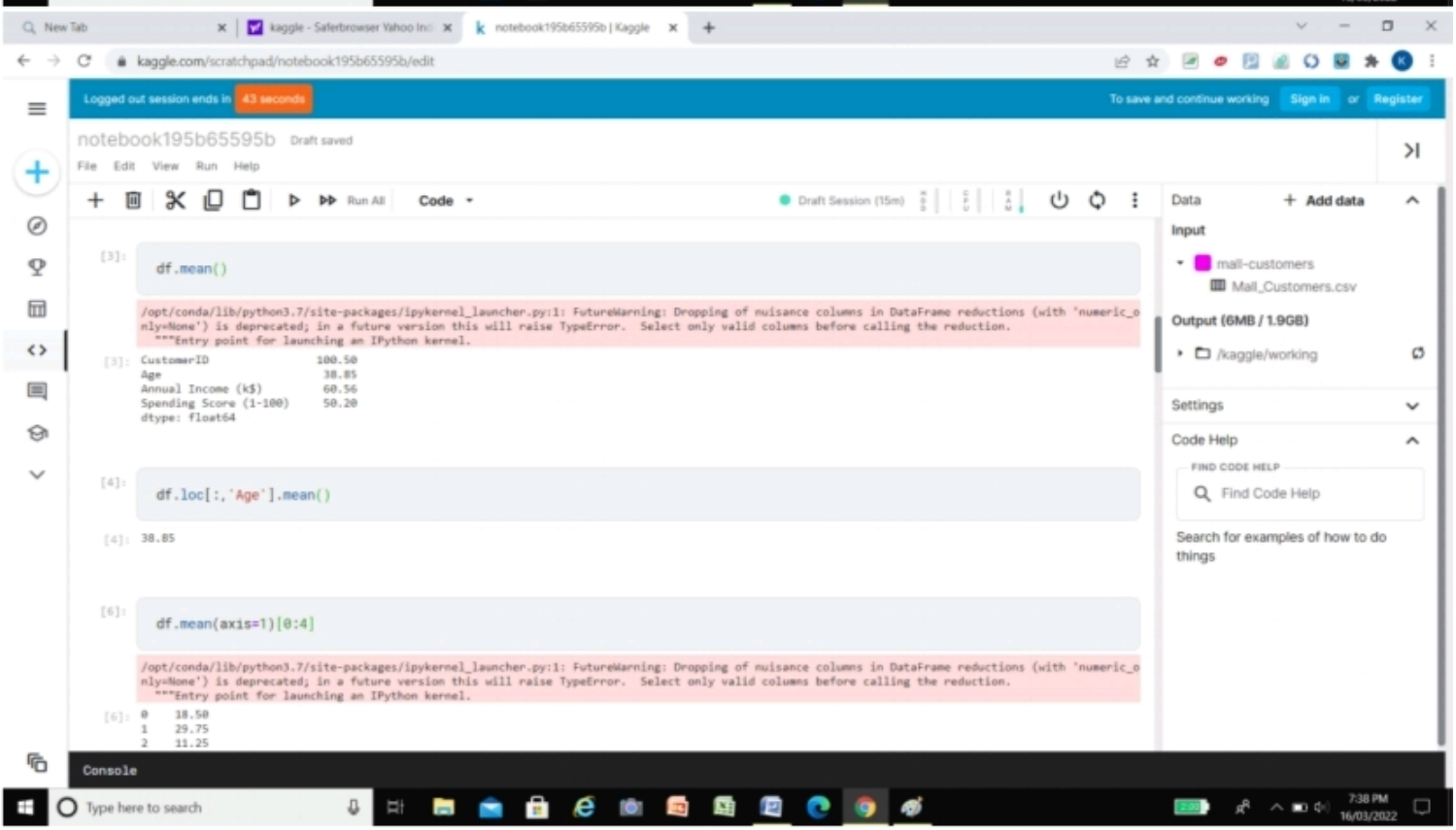
1. Provide summary statistics (mean, median, minimum, maximum, standard deviation) for a dataset (age, income etc.) with numeric variables grouped by one of the qualitative (categorical) variables. For example, if your categorical variable is age groups and quantitative variable is income, then provide summary statistics of income grouped by the age groups. Create a list that contains a numeric value for each response to the categorical variable.

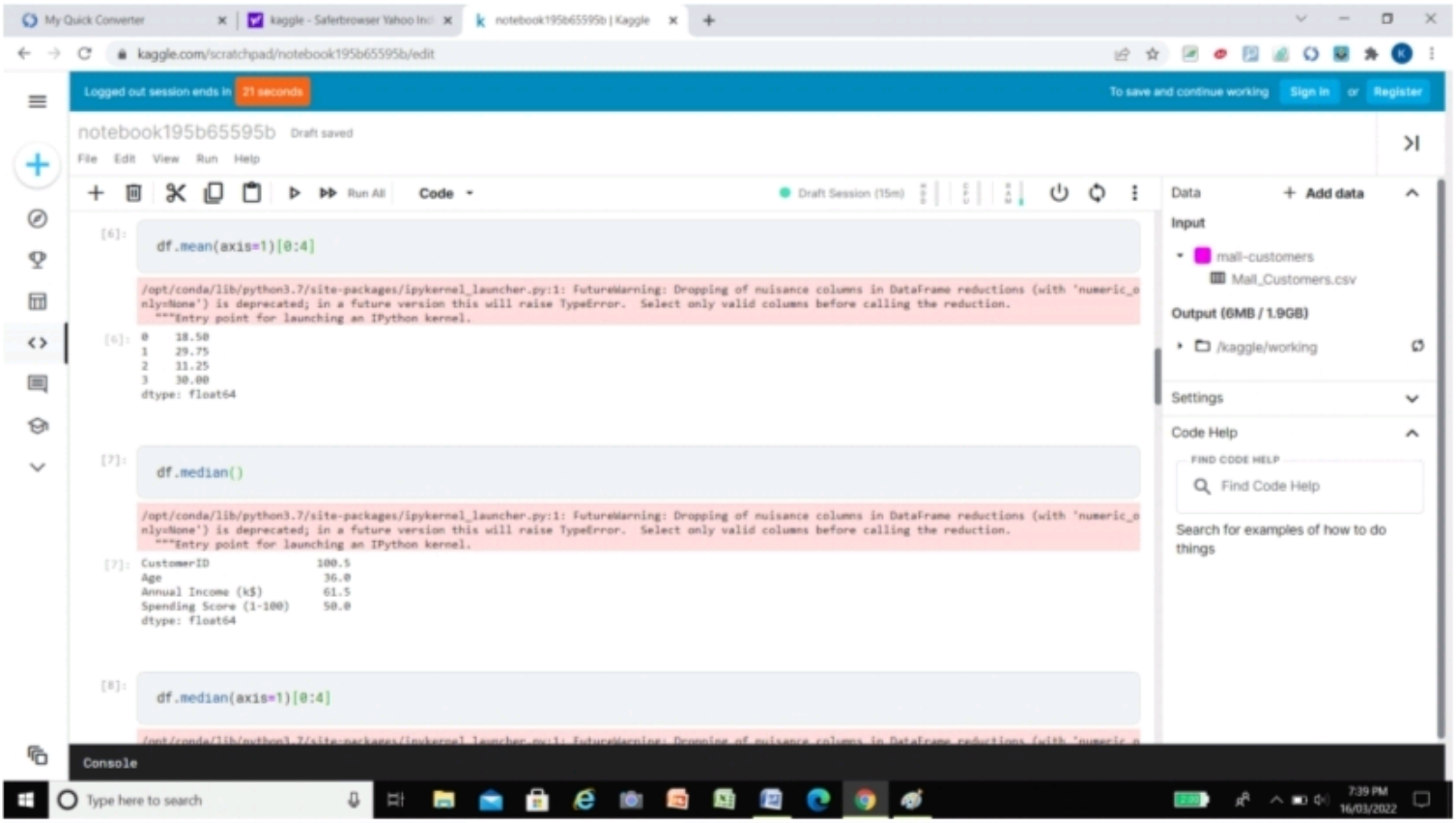
2. Write a Python program to display some basic statistical details like percentile, mean, standard deviation etc. of the species of 'Iris-setosa', 'Iris-versicolor' and 'Iris-versicolor' of iris.csv dataset.

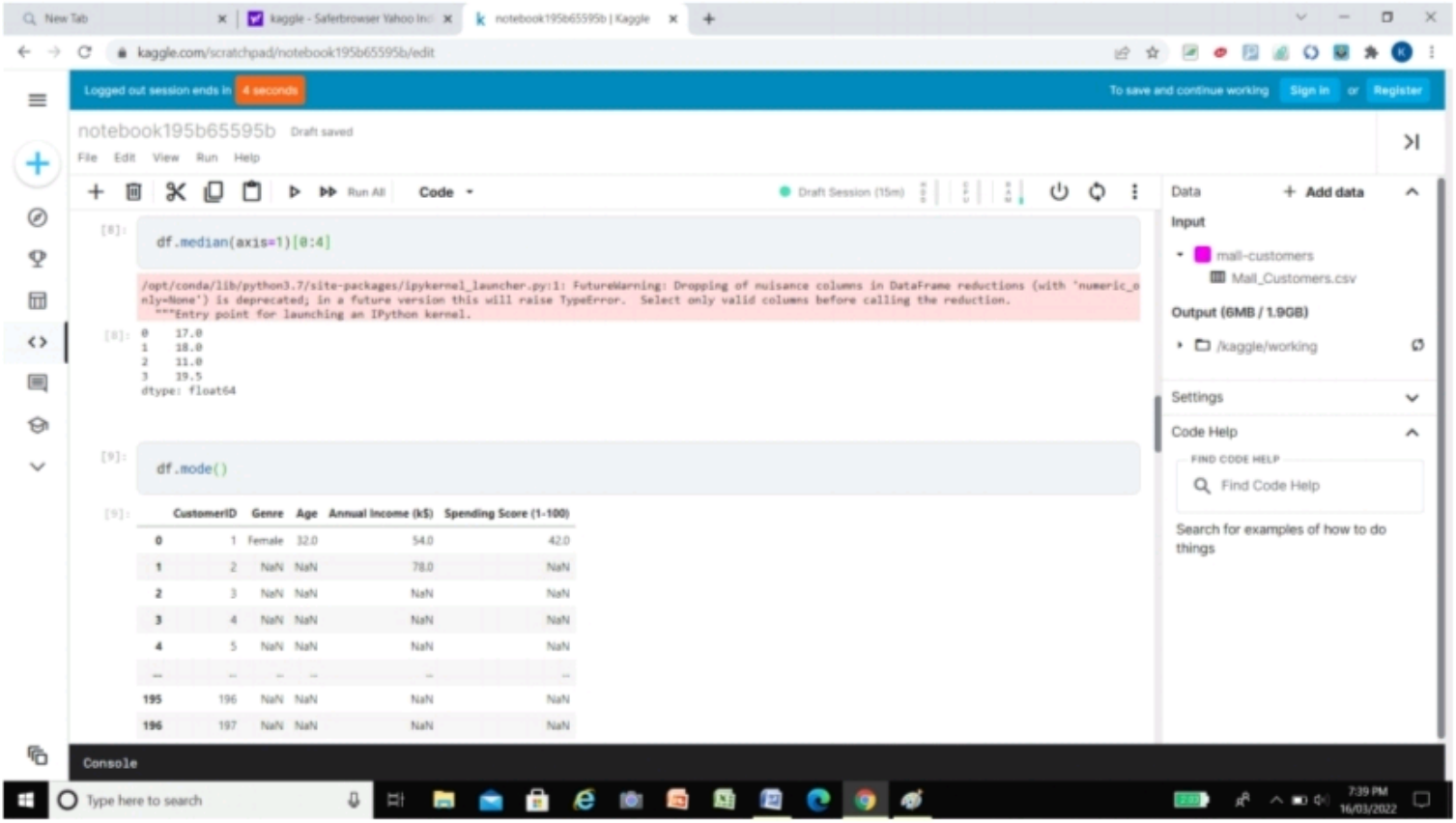
Provide the codes with outputs and explain everything that you do in this step.

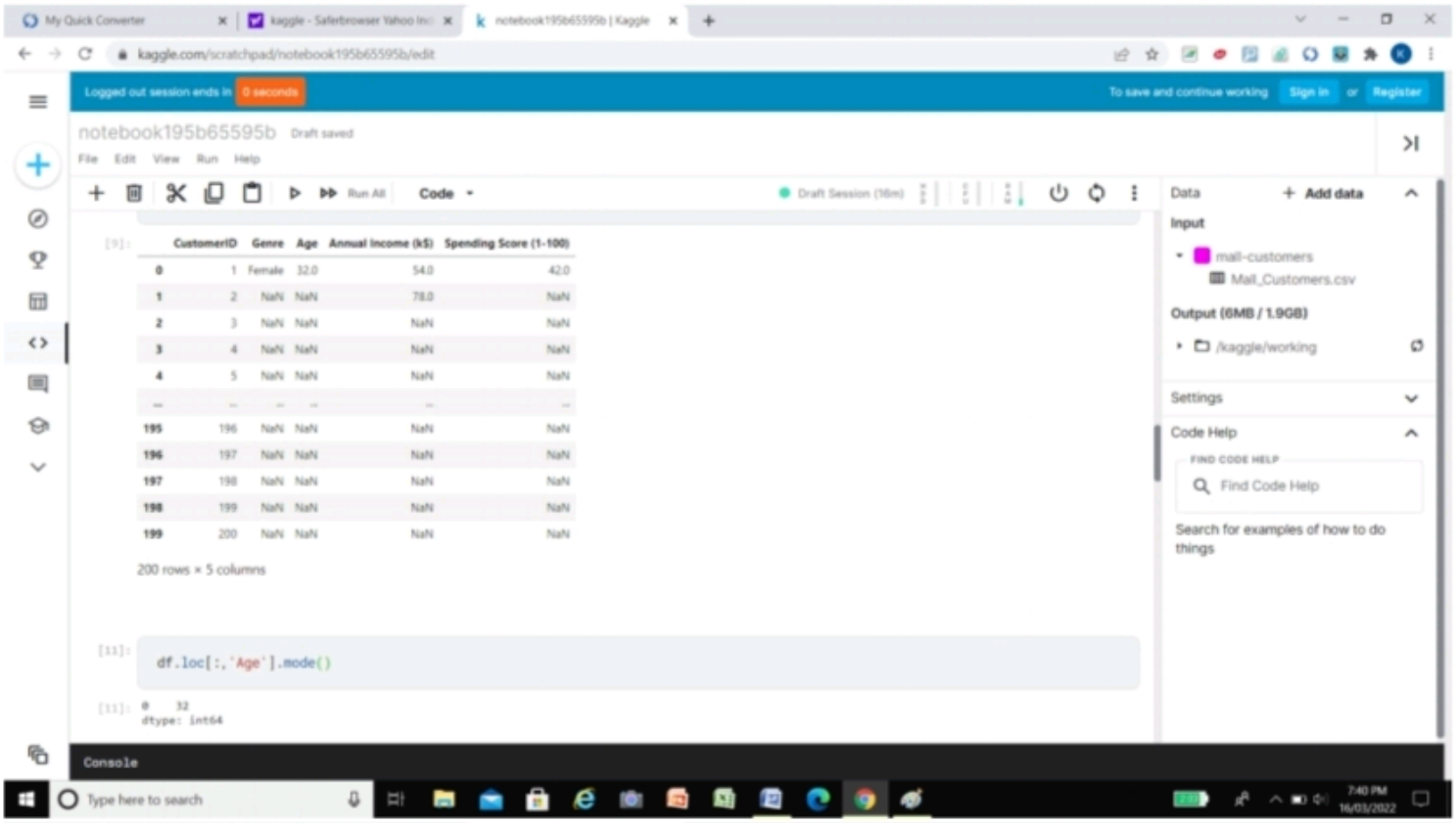
Output :

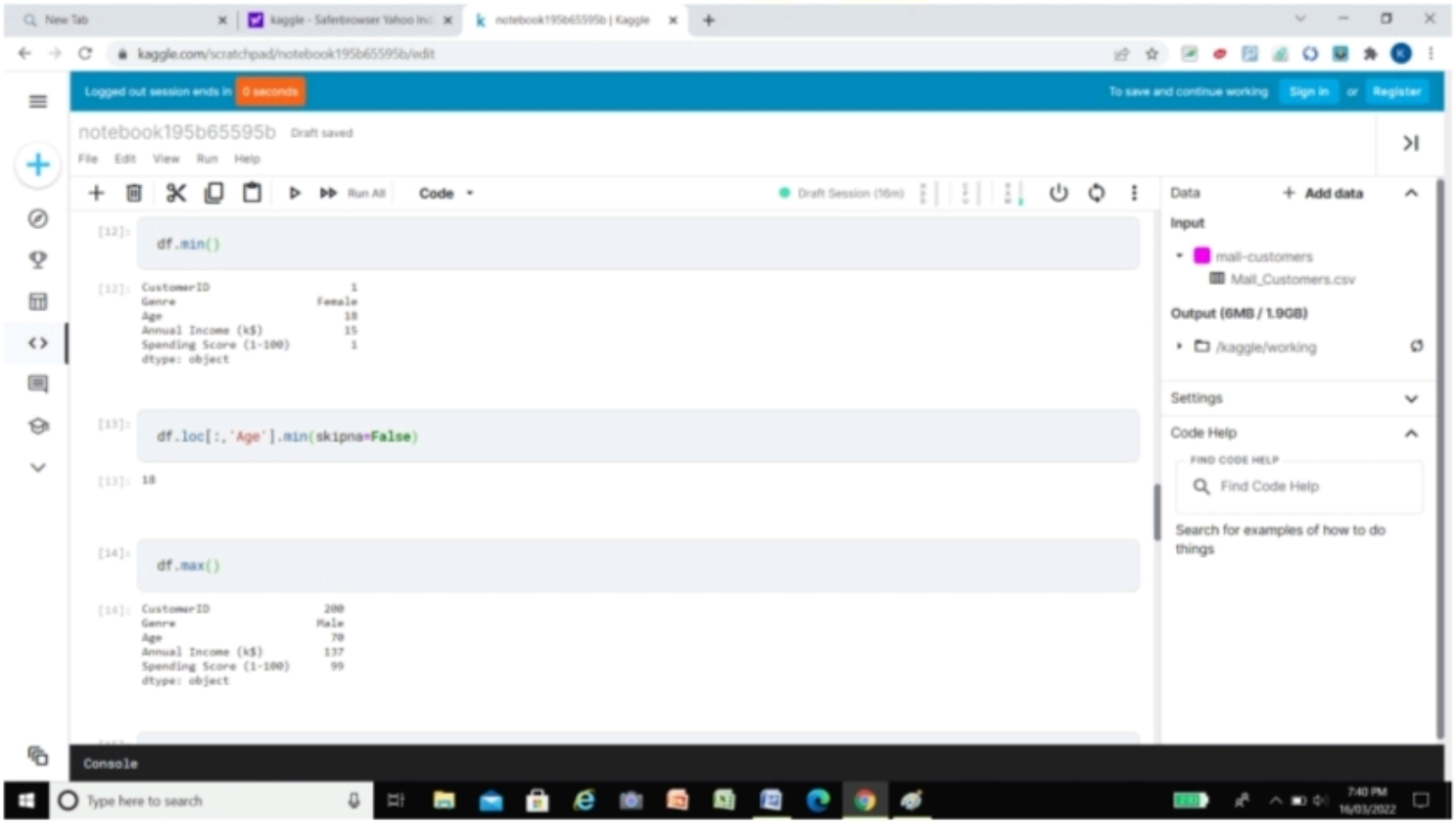

notebook195b65595b | Kaggle
kaggle.com/scratchpad/notebook195b65595b/edit
Explore and run machine learning code with Kaggle Notebooks.
Logged out sessions end after 15 minutes
Sign in
Register
notebook195b65595b
Draft saved
File Edit View Run Help
Run All
Code
Draft Session (14m)
Data
Add data
Input
mall-customers
Mall_Customers.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
[1]:
import numpy as np # linear algebra
import pandas as pd
df=pd.read_csv("../input/mall-customers/Mall_Customers.csv")
print("read the csv file")
read the csv file
[2]:
df.head()
[2]:
CustomerID Genre Age Annual Income (k$) Spending Score (1-100)
0 1 Male 19 15 39
1 2 Male 21 15 81
2 3 Female 20 16 6
3 4 Female 23 16 77
4 5 Female 31 17 40
Console
Type here to search
7:38 PM
16/03/2022

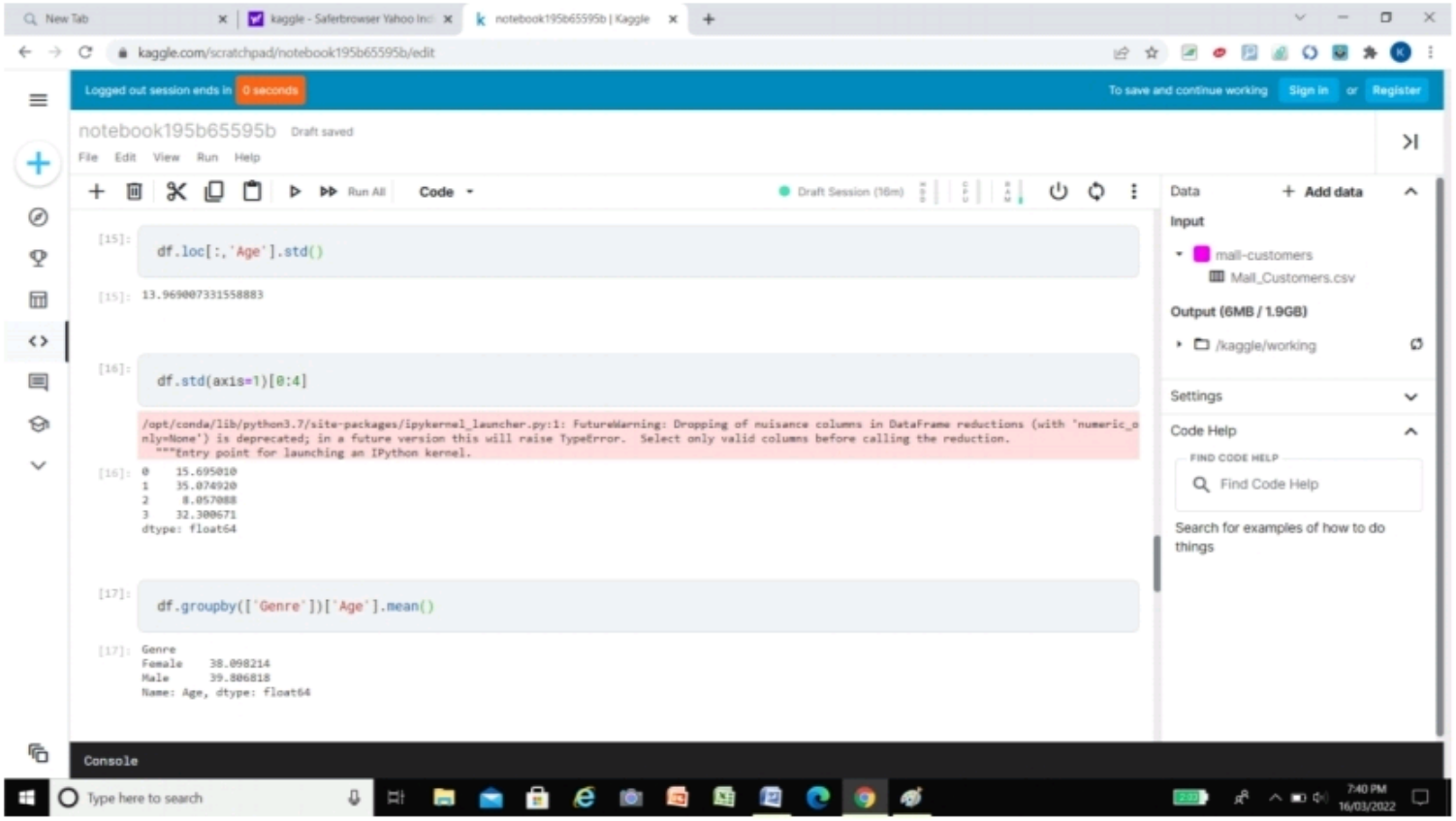

notebook195b65595b | Kaggle
kaggle.com/scratchpad/notebook195b65595b/edit
Logged out session ends in 43 seconds
To save and continue working
Sign in
Register
notebook195b65595b
Draft saved
File Edit View Run Help
Run All
Code
Draft Session (15m)
Data
Add data
Input
mall-customers
Mall_Customers.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
[3]:
df.mean()
/opt/conda/lib/python3.7/site-packages/ipykernel_launcher.py:1: FutureWarning: Dropping of nuisance columns in DataFrame reductions (with 'numeric_only=None') is deprecated; in a future version this will raise TypeError. Select only valid columns before calling the reduction.
"""Entry point for launching an IPython kernel.
[3]:
CustomerID 100.50
Age 38.85
Annual Income (k$) 60.56
Spending Score (1-100) 50.20
dtype: float64
[4]:
df.loc[:,'Age'].mean()
[4]:
38.85
[6]:
df.mean(axis=1)[0:4]
/opt/conda/lib/python3.7/site-packages/ipykernel_launcher.py:1: FutureWarning: Dropping of nuisance columns in DataFrame reductions (with 'numeric_only=None') is deprecated; in a future version this will raise TypeError. Select only valid columns before calling the reduction.
"""Entry point for launching an IPython kernel.
[6]:
0 38.50
1 29.75
2 11.25
Console
Type here to search
7:38 PM
16/03/2022

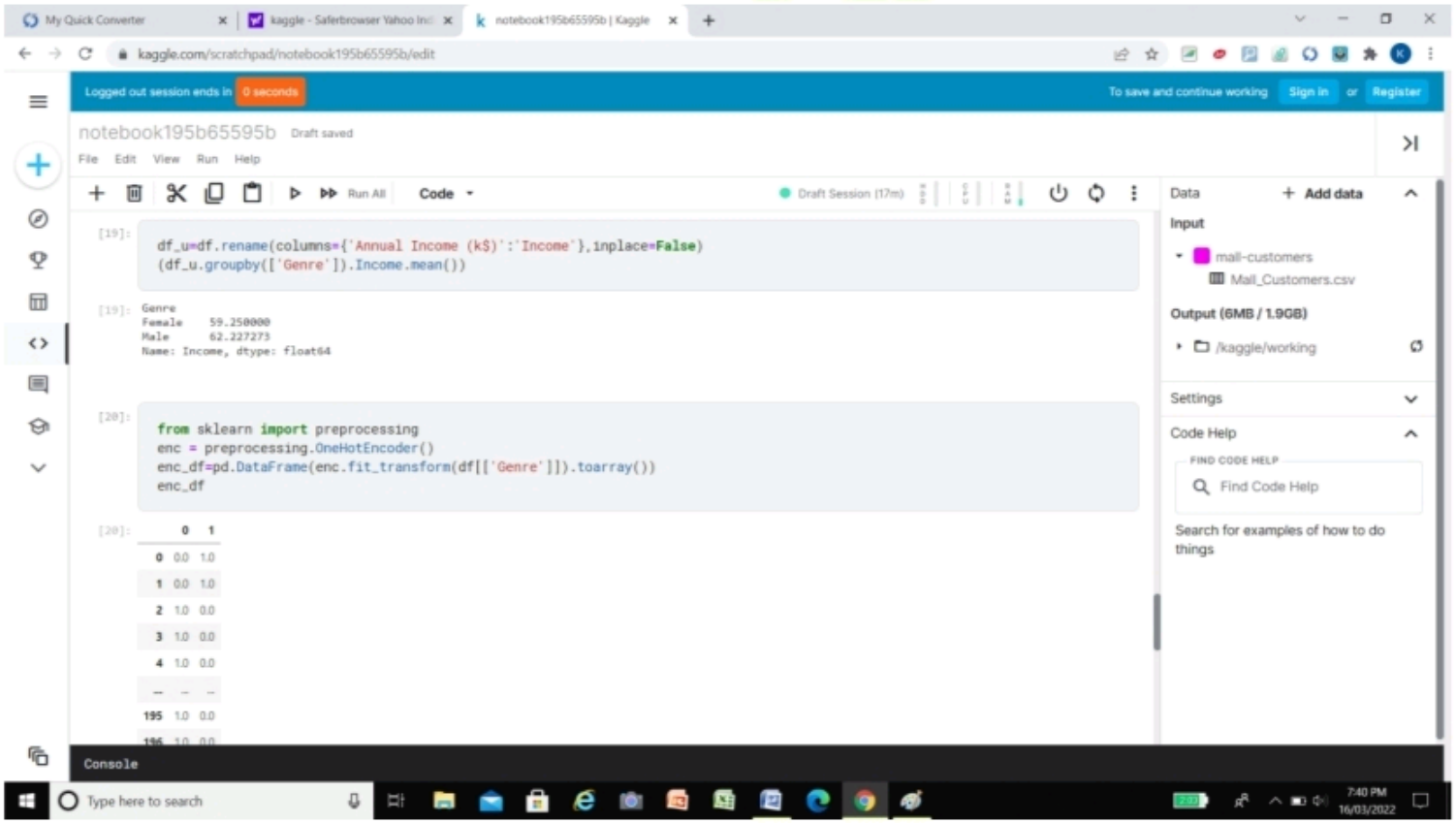

My Quick Converter
kaggle - Saferbrowser Yahoo Inc
notebook195b65595b | Kaggle
kaggle.com/scratchpad/notebook195b65595b/edit
Logged out session ends in 21 seconds
To save and continue working Sign in or Register
notebook195b65595b Draft saved
File Edit View Run Help
Run All Code Draft Session (15m)
Data Add data
Input
mall-customers
Mall_Customers.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
[6]: df.mean(axis=1)[0:4]
/opt/conda/lib/python3.7/site-packages/ipykernel_launcher.py:1: FutureWarning: Dropping of nuisance columns in DataFrame reductions (with 'numeric_only=None') is deprecated; in a future version this will raise TypeError. Select only valid columns before calling the reduction.
"""Entry point for launching an IPython kernel.
[6]: 0 18.50
1 29.75
2 11.25
3 30.00
dtype: float64
[7]: df.median()
/opt/conda/lib/python3.7/site-packages/ipykernel_launcher.py:1: FutureWarning: Dropping of nuisance columns in DataFrame reductions (with 'numeric_only=None') is deprecated; in a future version this will raise TypeError. Select only valid columns before calling the reduction.
"""Entry point for launching an IPython kernel.
[7]: CustomerID 100.5
Age 36.0
Annual Income (k$) 61.5
Spending Score (1-100) 50.0
dtype: float64
[8]: df.median(axis=1)[0:4]
Console
Type here to search
7:39 PM
16/03/2022

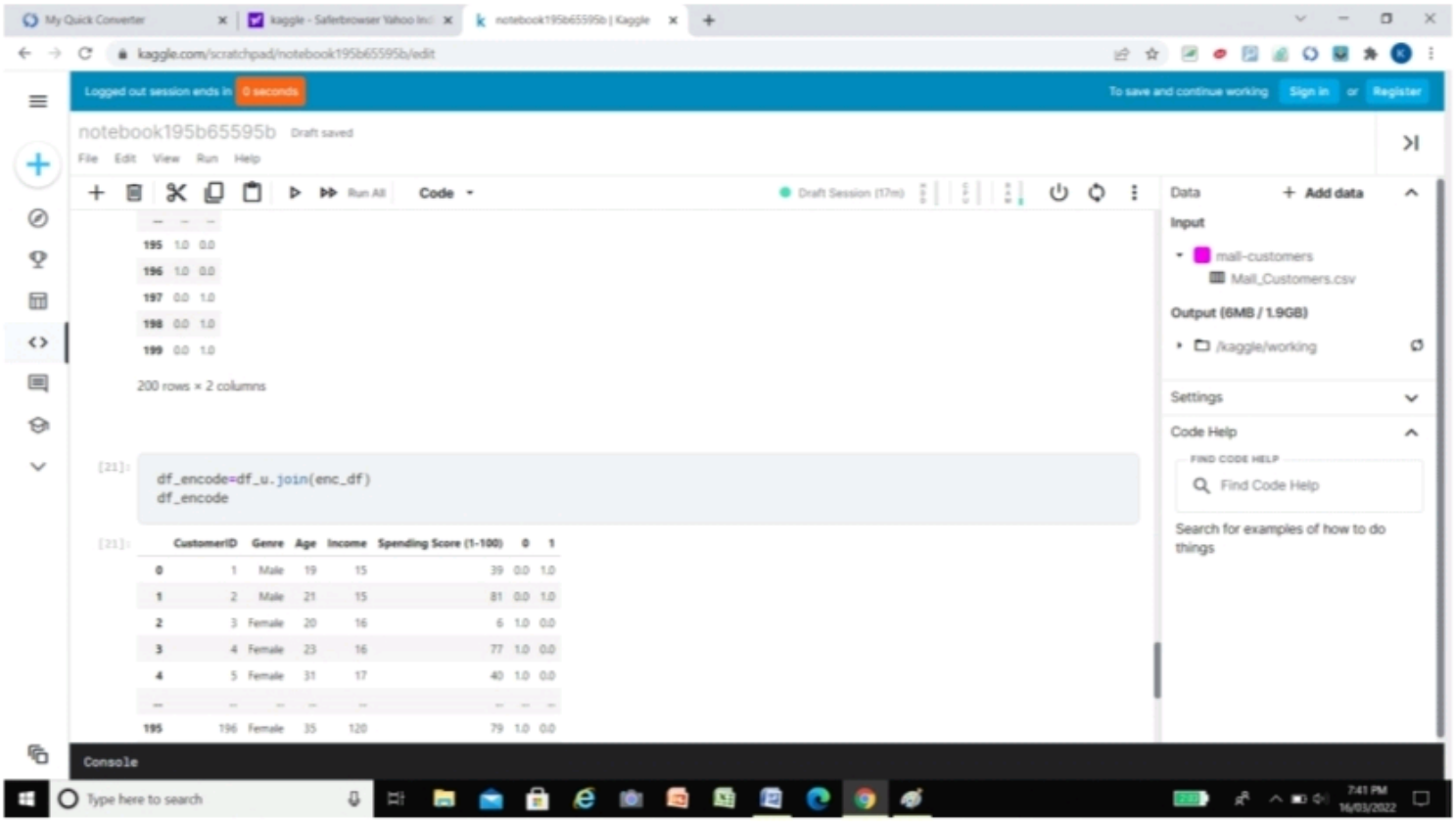

New Tab
kaggle - Saferbrowser Yahoo Inc
notebook195b65595b | Kaggle
kaggle.com/scratchpad/notebook195b65595b/edit
Logged out session ends in 4 seconds
To save and continue working Sign in or Register
notebook195b65595b Draft saved
File Edit View Run Help
Run All Code Draft Session (15m)
Data Add data
Input
mall-customers
Mall_Customers.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
[8]: df.median(axis=1)[0:4]
/opt/conda/lib/python3.7/site-packages/ipykernel_launcher.py:1: FutureWarning: Dropping of nuisance columns in DataFrame reductions (with 'numeric_only=None') is deprecated; in a future version this will raise TypeError. Select only valid columns before calling the reduction.
"""Entry point for launching an IPython kernel.
[8]: 0 17.0
1 18.0
2 11.0
3 19.5
dtype: float64
[9]: df.mode()
[9]:
CustomerID Genre Age Annual Income (k$) Spending Score (1-100)
0 1 Female 32.0 54.0 42.0
1 2 NaN NaN 78.0 NaN
2 3 NaN NaN NaN NaN
3 4 NaN NaN NaN NaN
4 5 NaN NaN NaN NaN
... ...
195 196 NaN NaN NaN NaN
196 197 NaN NaN NaN NaN
Console
Type here to search
7:39 PM
16/03/2022


My Quick Converter
kaggle - Saferbrowser Yahoo Inc
notebook195b65595b | Kaggle
kaggle.com/scratchpad/notebook195b65595b/edit
Logged out session ends in 0 seconds
To save and continue working Sign in or Register
notebook195b65595b Draft saved
File Edit View Run Help
Run All Code Draft Session (16m)
Data Add data
Input
mall-customers
Mall_Customers.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
[9]:
CustomerID Genre Age Annual Income (k$) Spending Score (1-100)
0 1 Female 32.0 54.0 42.0
1 2 NaN NaN 78.0 NaN
2 3 NaN NaN NaN NaN
3 4 NaN NaN NaN NaN
4 5 NaN NaN NaN NaN
195 196 NaN NaN NaN NaN
196 197 NaN NaN NaN NaN
197 198 NaN NaN NaN NaN
198 199 NaN NaN NaN NaN
199 200 NaN NaN NaN NaN
200 rows × 5 columns
[11]:
df.loc[:,'Age'].mode()
[11]:
0 32
dtype: int64
Console
Type here to search
7:40 PM
16/03/2022


New Tab
kaggle - Saferbrowser Yahoo Inc
notebook195b65595b | Kaggle
kaggle.com/scratchpad/notebook195b65595b/edit
Logged out session ends in 0 seconds
To save and continue working Sign in or Register
notebook195b65595b Draft saved
File Edit View Run Help
Run All Code Draft Session (16m)
Data Add data
Input
mall-customers
Mall_Customers.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
[12]:
df.min()
[12]:
CustomerID 1
Genre Female
Age 18
Annual Income (k$) 15
Spending Score (1-100) 1
dtype: object
[13]:
df.loc[:,'Age'].min(skipna=False)
[13]:
18
[14]:
df.max()
[14]:
CustomerID 200
Genre Male
Age 70
Annual Income (k$) 137
Spending Score (1-100) 99
dtype: object
Console
Type here to search
7:40 PM
16/03/2022


kaggle.com/scratchpad/notebook195b65595b/edit
Logged out session ends in 0 seconds
To save and continue working Sign in or Register
notebook195b65595b Draft saved
File Edit View Run Help
Run All Code Draft Session (16m)
Data + Add data
Input
mall-customers
Mall_Customers.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
[15]: df.loc[:,'Age'].std()
[15]: 13.969007331558883
[16]: df.std(axis=1)[0:4]
/opt/conda/lib/python3.7/site-packages/ipykernel_launcher.py:1: FutureWarning: Dropping of nuisance columns in DataFrame reductions (with 'numeric_only=None') is deprecated; in a future version this will raise TypeError. Select only valid columns before calling the reduction.
"""Entry point for launching an IPython kernel.
[16]: 0 15.695010
1 35.074920
2 8.057088
3 32.300671
dtype: float64
[17]: df.groupby(['Genre'])['Age'].mean()
[17]: Genre
Female 38.098214
Male 39.806818
Name: Age, dtype: float64
Console
Type here to search
7:40 PM
16/03/2022


kaggle.com/scratchpad/notebook195b65595b/edit
Logged out session ends in 0 seconds
To save and continue working Sign in or Register
notebook195b65595b Draft saved
File Edit View Run Help
Run All Code Draft Session (17m)
Data + Add data
Input
mall-customers
Mall_Customers.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
[19]: df_u=df.rename(columns={'Annual Income (k$)':'Income'},inplace=False)
(df_u.groupby(['Genre']).Income.mean())
[19]: Genre
Female 59.250000
Male 62.227273
Name: Income, dtype: float64
[20]: from sklearn import preprocessing
enc = preprocessing.OneHotEncoder()
enc_df=pd.DataFrame(enc.fit_transform(df[['Genre']]).toarray())
enc_df
[20]:
0 1
0 0.0 1.0
1 0.0 1.0
2 1.0 0.0
3 1.0 0.0
4 1.0 0.0
195 1.0 0.0
Console
Type here to search
7:40 PM
16/03/2022



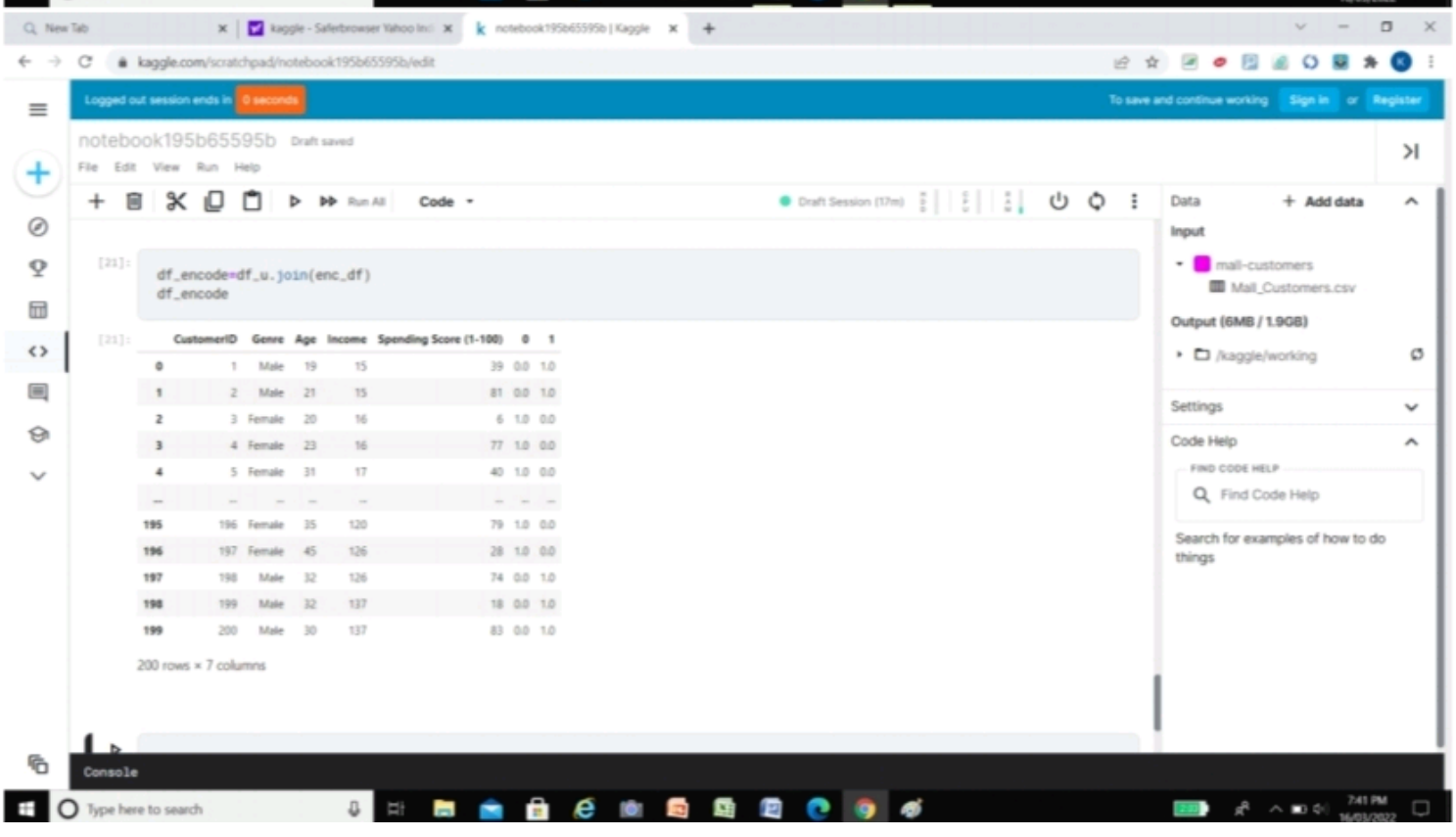

```
195  1.0  0.0
196  1.0  0.0
197  0.0  1.0
198  0.0  1.0
199  0.0  1.0
```

200 rows × 2 columns

[21]:
```
df_encode=df_u.join(enc_df)
df_encode
```

[21]:

	CustomerID	Genre	Age	Income	Spending Score (1-100)	0	1
0	1	Male	19	15	39	0.0	1.0
1	2	Male	21	15	81	0.0	1.0
2	3	Female	20	16	6	1.0	0.0
3	4	Female	23	16	77	1.0	0.0
4	5	Female	31	17	40	1.0	0.0
...	...	...	...	...	...	...	...
195	196	Female	35	120	79	1.0	0.0

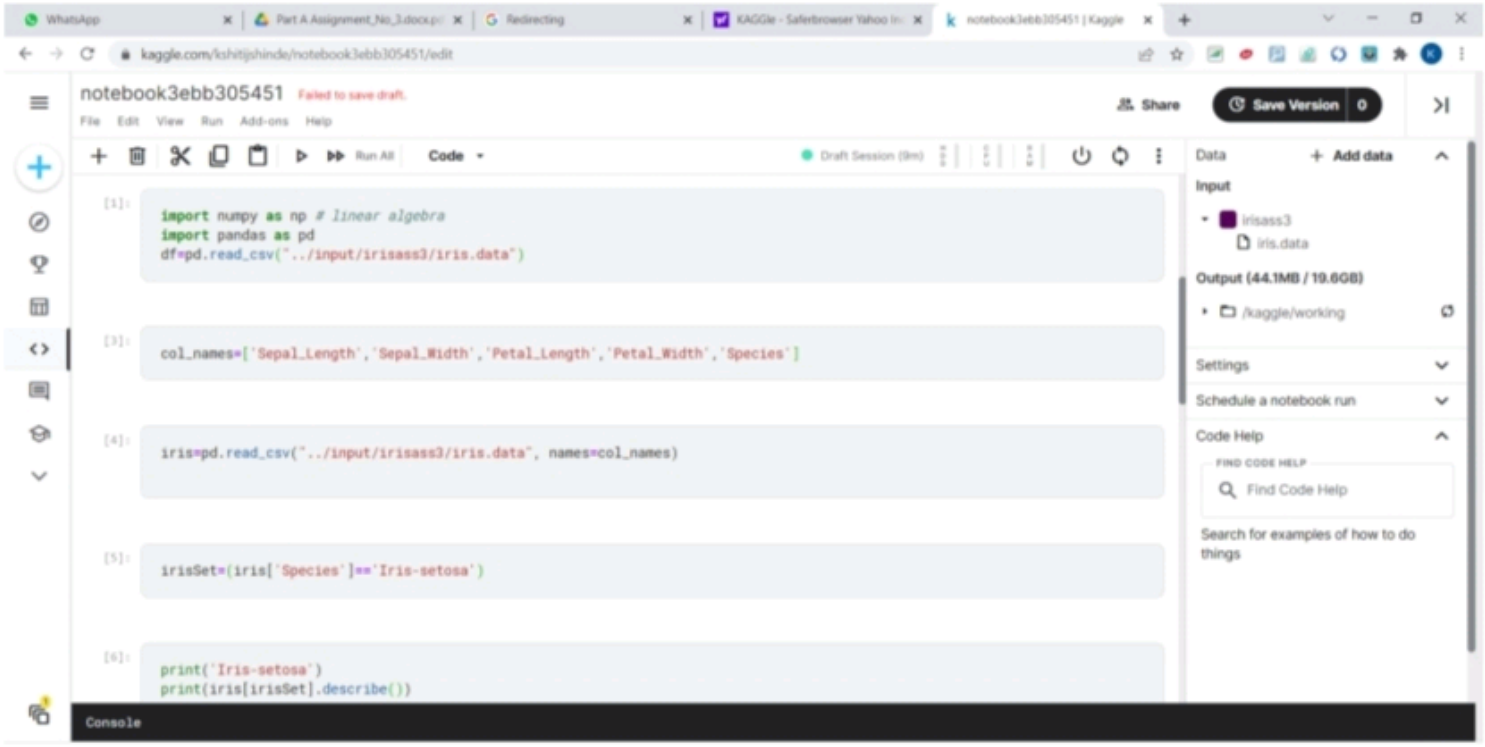


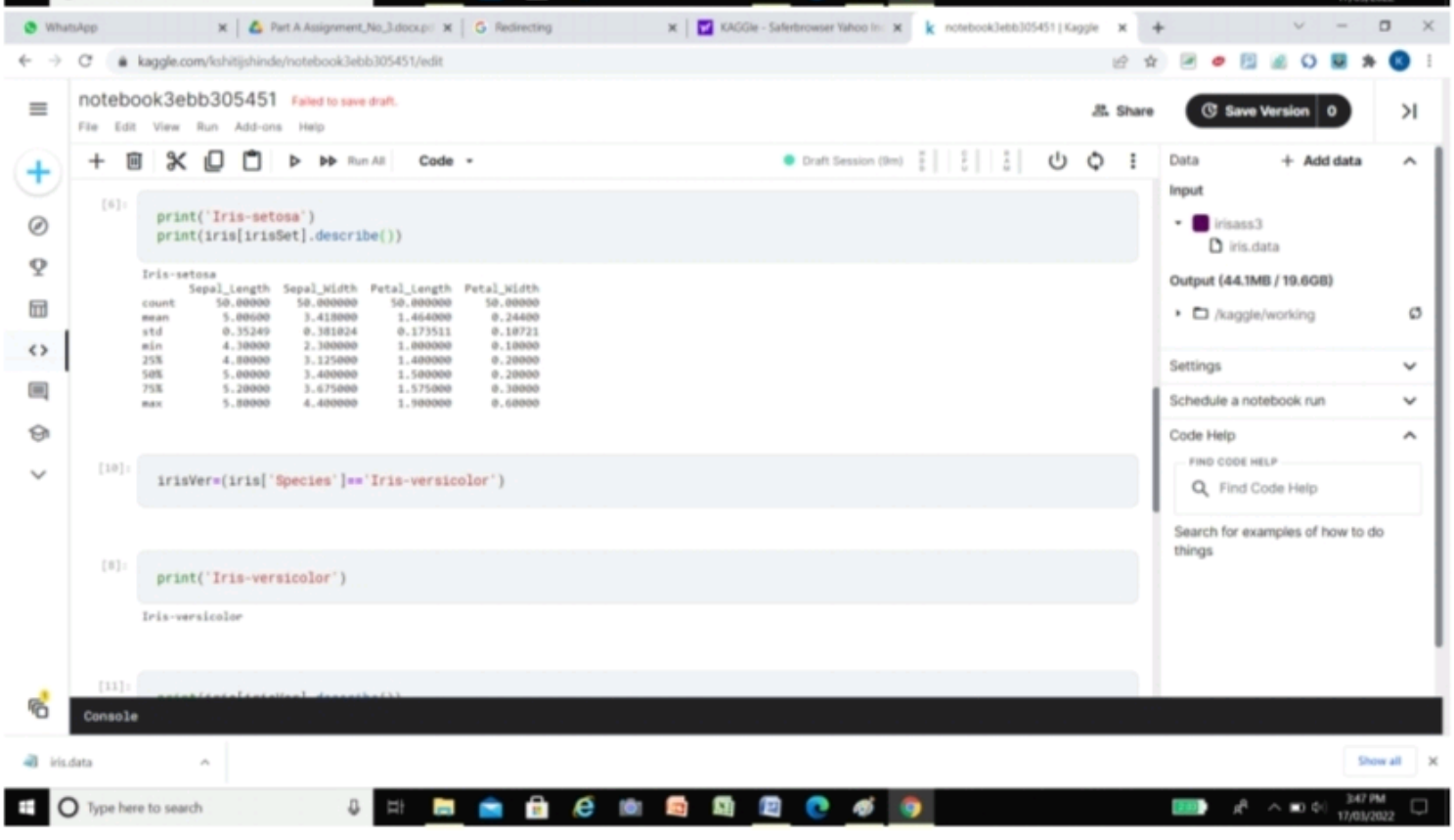
[21]:
```
df_encode=df_u.join(enc_df)
df_encode
```

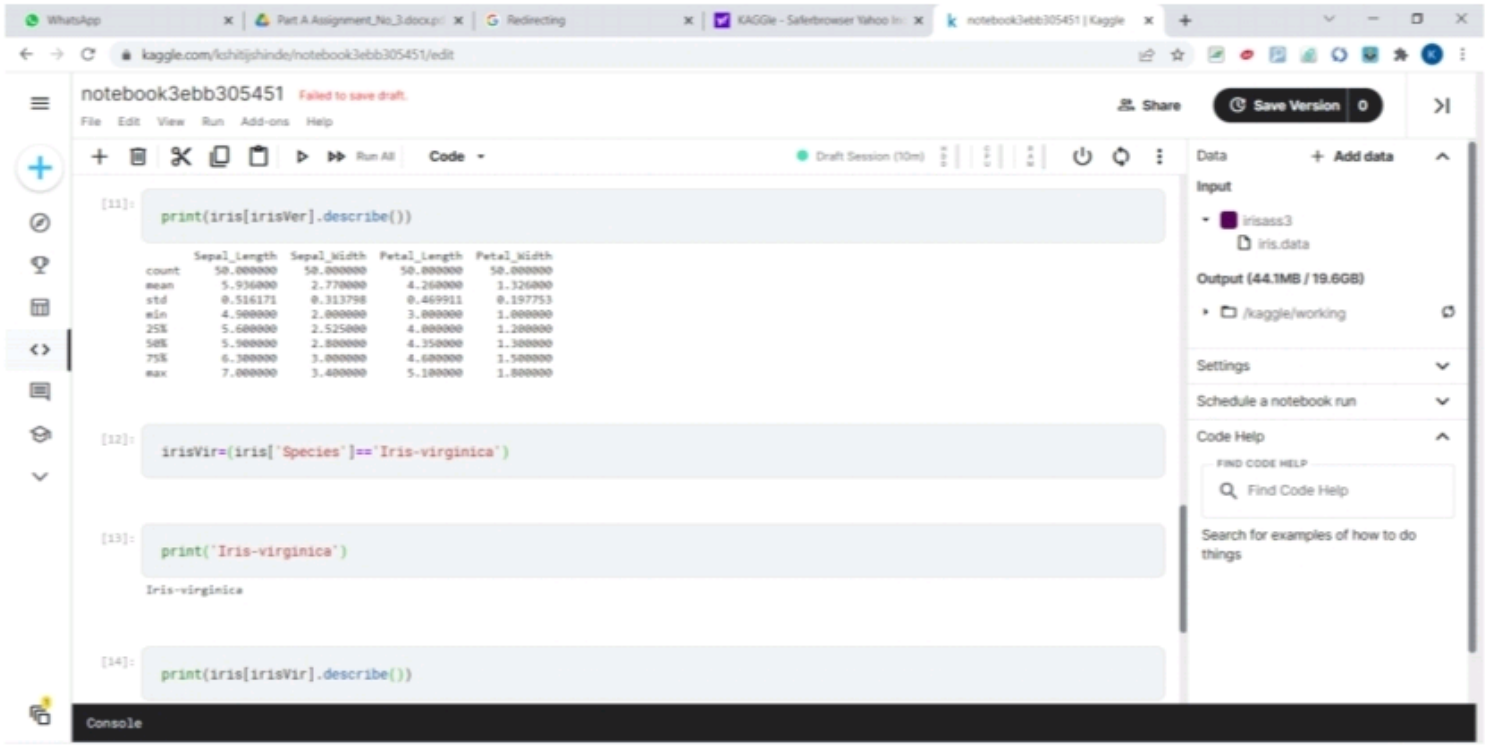
[21]:

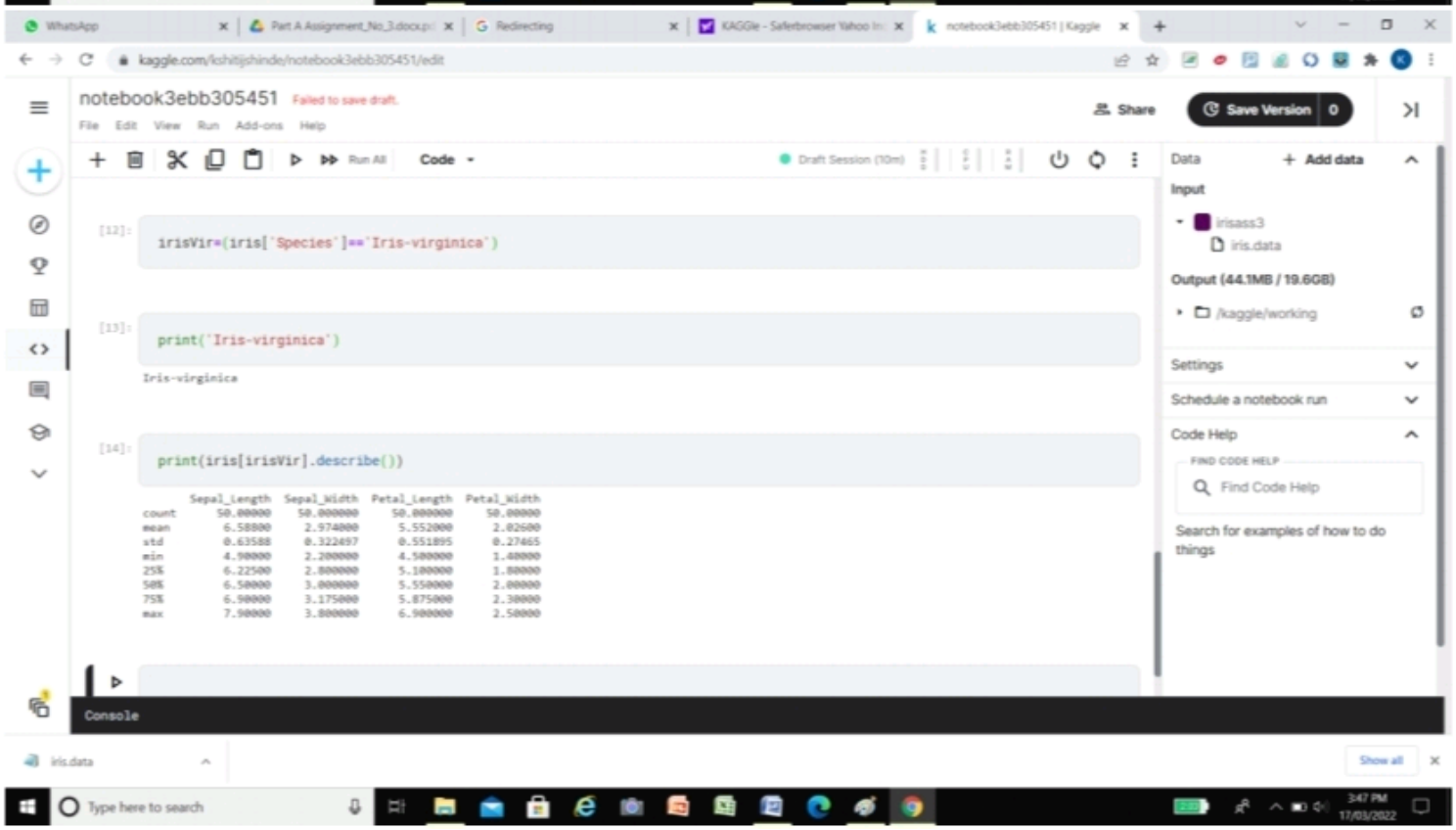
	CustomerID	Genre	Age	Income	Spending Score (1-100)	0	1
0	1	Male	19	15	39	0.0	1.0
1	2	Male	21	15	81	0.0	1.0
2	3	Female	20	16	6	1.0	0.0
3	4	Female	23	16	77	1.0	0.0
4	5	Female	31	17	40	1.0	0.0
...	...	...	...	...	...	...	...
195	196	Female	35	120	79	1.0	0.0
196	197	Female	45	126	28	1.0	0.0
197	198	Male	32	126	74	0.0	1.0
198	199	Male	32	137	18	0.0	1.0
199	200	Male	30	137	83	0.0	1.0

200 rows × 7 columns


notebook3ebb305451
Failed to save draft.
File Edit View Run Add-ons Help
Share
Save Version 0
Run All
Code
Draft Session (9m)
[1]:
import numpy as np # linear algebra
import pandas as pd
df=pd.read_csv("../input/irisass3/iris.data")
[3]:
col_names=['Sepal_Length','Sepal_Width','Petal_Length','Petal_Width','Species']
[4]:
iris=pd.read_csv("../input/irisass3/iris.data", names=col_names)
[5]:
irisSet=(iris['Species']=='Iris-setosa')
[6]:
print('Iris-setosa')
print(iris[irisSet].describe())
Console
Data
Add data
Input
irisass3
iris.data
Output (44.1MB / 19.6GB)
/kaggle/working
Settings
Schedule a notebook run
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things


notebook3ebb305451
Failed to save draft.
File Edit View Run Add-ons Help
Share
Save Version 0
Run All
Code
Draft Session (9m)
[6]:
print('Iris-setosa')
print(iris[irisSet].describe())
Iris-setosa
Sepal_Length Sepal_Width Petal_Length Petal_Width
count 50.00000 50.000000 50.000000 50.00000
mean 5.00600 3.418000 1.464000 0.24400
std 0.35249 0.381024 0.173511 0.10721
min 4.30000 2.300000 1.000000 0.10000
25% 4.80000 3.125000 1.400000 0.20000
50% 5.00000 3.400000 1.500000 0.20000
75% 5.20000 3.675000 1.575000 0.30000
max 5.80000 4.400000 1.900000 0.60000
[10]:
irisVer=(iris['Species']=='Iris-versicolor')
[8]:
print('Iris-versicolor')
Iris-versicolor
Console
Data
Add data
Input
irisass3
iris.data
Output (44.1MB / 19.6GB)
/kaggle/working
Settings
Schedule a notebook run
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things


notebook3ebb305451
Failed to save draft.
File Edit View Run Add-ons Help
Share
Save Version 0
Run All
Code
Draft Session (10m)
Data
Add data
Input
irisass3
iris.data
Output (44.1MB / 19.6GB)
/kaggle/working
Settings
Schedule a notebook run
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
[11]:
print(iris[irisVer].describe())
Sepal_Length Sepal_Width Petal_Length Petal_Width
count 50.000000 50.000000 50.000000 50.000000
mean 5.936000 2.770000 4.260000 1.326000
std 0.516171 0.313798 0.469911 0.197753
min 4.900000 2.000000 3.000000 1.000000
25% 5.600000 2.525000 4.000000 1.200000
50% 5.900000 2.800000 4.350000 1.300000
75% 6.300000 3.000000 4.600000 1.500000
max 7.000000 3.400000 5.100000 1.800000
[12]:
irisVir=(iris['Species']=='Iris-virginica')
[13]:
print('Iris-virginica')
Iris-virginica
[14]:
print(iris[irisVir].describe())
Console
iris.data
Show all
Type here to search
3:47 PM
17/03/2022


notebook3ebb305451
Failed to save draft.
File Edit View Run Add-ons Help
Share
Save Version 0
Run All
Code
Draft Session (10m)
Data
Add data
Input
irisass3
iris.data
Output (44.1MB / 19.6GB)
/kaggle/working
Settings
Schedule a notebook run
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
[12]:
irisVir=(iris['Species']=='Iris-virginica')
[13]:
print('Iris-virginica')
Iris-virginica
[14]:
print(iris[irisVir].describe())
Sepal_Length Sepal_Width Petal_Length Petal_Width
count 50.00000 50.000000 50.000000 50.00000
mean 6.58800 2.974000 5.552000 2.02600
std 0.63588 0.322497 0.551895 0.27465
min 4.90000 2.200000 4.500000 1.40000
25% 6.22500 2.800000 5.100000 1.80000
50% 6.50000 3.000000 5.550000 2.00000
75% 6.90000 3.175000 5.875000 2.30000
max 7.90000 3.800000 6.900000 2.50000
Console
iris.data
Show all
Type here to search
3:47 PM
17/03/2022

Write-up	Correctness of Program	Documentation of Program	Viva	Timely Completion	Total	Dated Sign of Subject Teacher
4	4	4	4	4	20	

Group A

Assignment No: 4

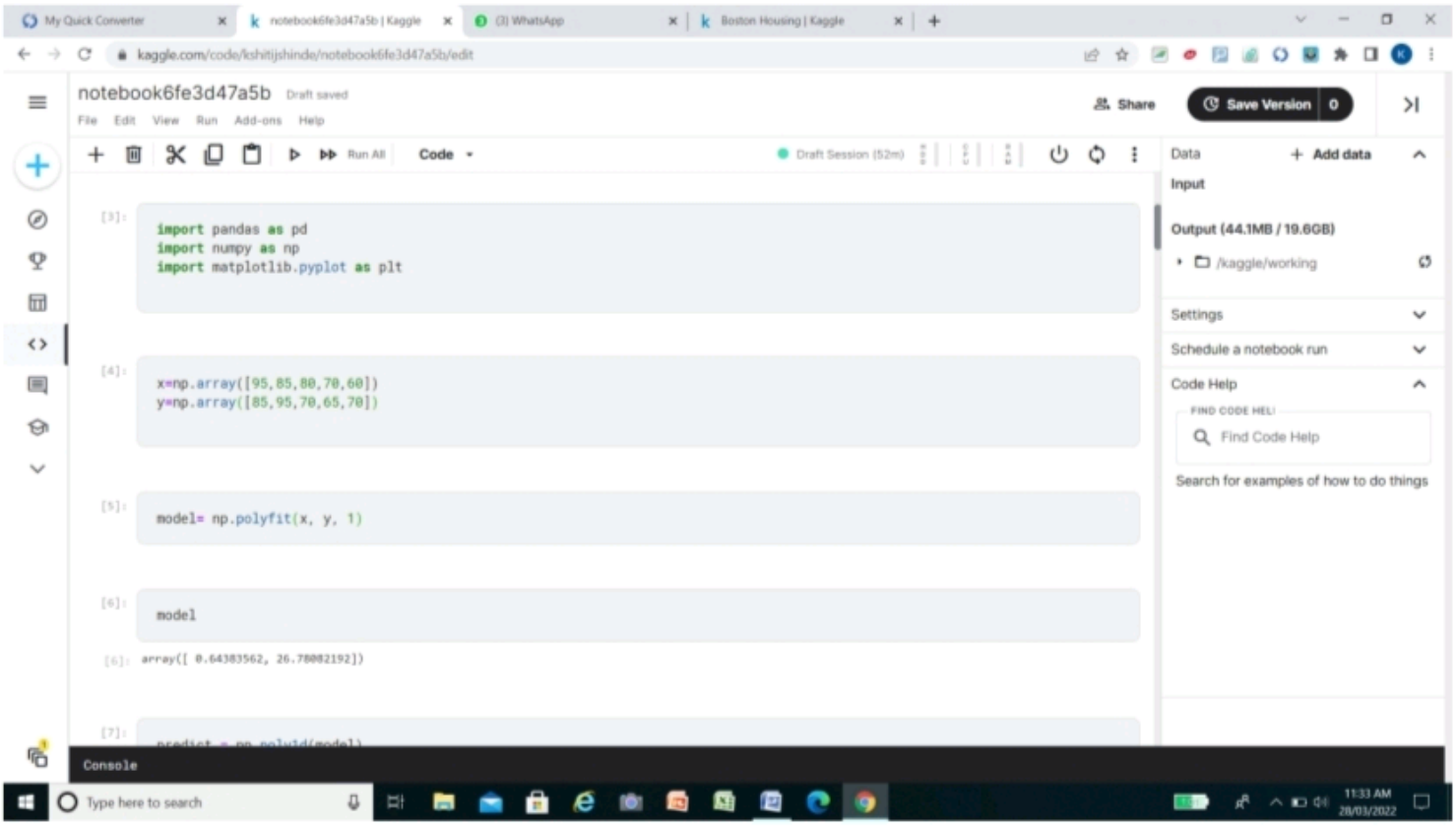
Title of the Assignment:

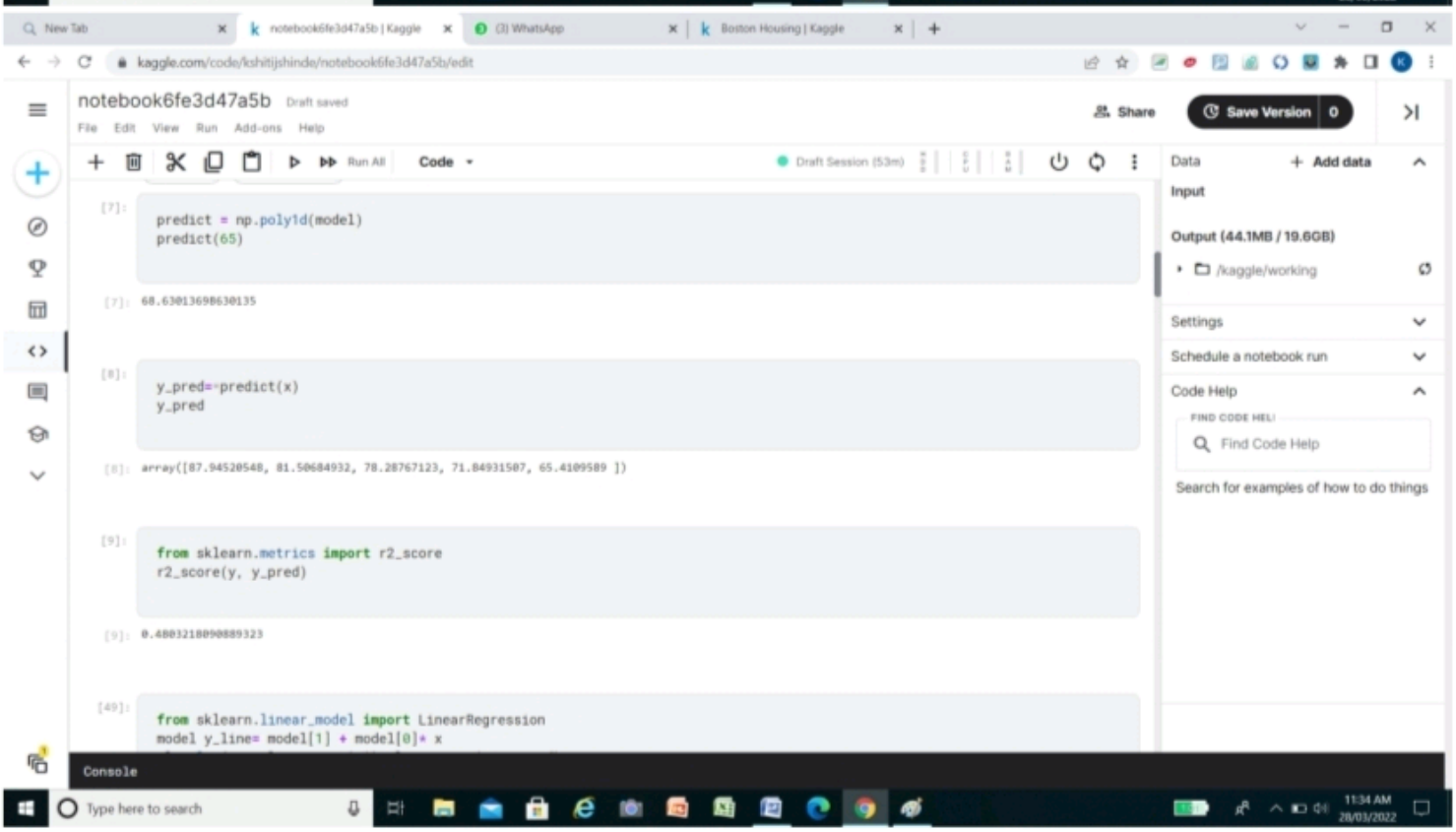
Create a Linear Regression Model using Python/R to predict home prices using Boston Housing Dataset (https://www.kaggle.com/c/boston-housing).

The Boston Housing dataset contains information about various houses in Boston through different parameters. There are 506 samples and 14 feature variables in this dataset.

The objective is to predict the value of prices of the house using the given features.

OUTPUT :


notebook6fe3d47a5b
Draft saved
File Edit View Run Add-ons Help
Share
Save Version 0
Run All
Code
Draft Session (52m)
Data
Add data
Input
Output (44.1MB / 19.6GB)
/kaggle/working
Settings
Schedule a notebook run
Code Help
Find Code Help
Search for examples of how to do things
[3]:
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
[4]:
x=np.array([95,85,80,70,60])
y=np.array([85,95,70,65,70])
[5]:
model= np.polyfit(x, y, 1)
[6]:
model
[6]:
array([0.64383562, 26.78082192])
[7]:
Console


notebook6fe3d47a5b
Draft saved
File Edit View Run Add-ons Help
Share
Save Version 0
Run All
Code
Draft Session (53m)
Data
Add data
Input
Output (44.1MB / 19.6GB)
/kaggle/working
Settings
Schedule a notebook run
Code Help
Find Code Help
Search for examples of how to do things
[7]:
predict = np.poly1d(model)
predict(65)
[7]:
68.63013698630135
[8]:
y_pred=predict(x)
y_pred
[8]:
array([87.94520548, 81.50684932, 78.28767123, 71.84931507, 65.4109589])
[9]:
from sklearn.metrics import r2_score
r2_score(y, y_pred)
[9]:
0.480321809889323
[49]:
from sklearn.linear_model import LinearRegression
model y_line= model[1] + model[0]* x
Console

My Quick Converter | notebook6fe3d47a5b | Kaggle | (3) WhatsApp | Boston Housing | Kaggle

kaggle.com/code/kshitijshinde/notebook6fe3d47a5b/edit

notebook6fe3d47a5b

Draft saved

File Edit View Run Add-ons Help

Share | Save Version 0

Run All | Code | Draft Session (53m)

Data | + Add data

Input

Output (44.1MB / 19.6GB)

/kaggle/working

Settings

Schedule a notebook run

Code Help

FIND CODE HELP

Find Code Help

Search for examples of how to do things

[21]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

[22]:
```
from sklearn.datasets import load_boston
boston_dataset = load_boston()
boston_dataset.keys()
```

```
/opt/conda/lib/python3.7/site-packages/sklearn/utils/deprecation.py:87: FutureWarning: Function load_boston is deprecated; `load_boston` is deprecat
ed in 1.0 and will be removed in 1.2.

    The Boston housing prices dataset has an ethical problem. You can refer to
    the documentation of this function for further details.

    The scikit-learn maintainers therefore strongly discourage the use of this
    dataset unless the purpose of the code is to study and educate about
    ethical issues in data science and machine learning.

    In this special case, you can fetch the dataset from the original
    source::

        import pandas as pd
        import numpy as np
```

Console

Type here to search

11:35 AM 28/03/2022

New Tab | notebook6fe3d47a5b | Kaggle | (3) WhatsApp | Boston Housing | Kaggle

kaggle.com/code/kshitijshinde/notebook6fe3d47a5b/edit

notebook6fe3d47a5b

Draft saved

File Edit View Run Add-ons Help

Share | Save Version 0

Run All | Code | Draft Session (54m)

Data | + Add data

Input

Output (44.1MB / 19.6GB)

/kaggle/working

Settings

Schedule a notebook run

Code Help

FIND CODE HELP

Find Code Help

Search for examples of how to do things

```
        data_url = "http://lib.stat.cmu.edu/datasets/boston"
        raw_df = pd.read_csv(data_url, sep="\s+", skiprows=22, header=None)
        data = np.hstack([raw_df.values[::2, :], raw_df.values[1::2, :2]])
        target = raw_df.values[1::2, 2]

    Alternative datasets include the California housing dataset (i.e.
    :func:`~sklearn.datasets.fetch_california_housing`) and the Ames housing
    dataset. You can load the datasets as follows::

        from sklearn.datasets import fetch_california_housing
        housing = fetch_california_housing()

    for the California housing dataset and::

        from sklearn.datasets import fetch_openml
        housing = fetch_openml(name="house_prices", as_frame=True)

    for the Ames housing dataset.

  warnings.warn(msg, category=FutureWarning)
```

[22]:
```
dict_keys(['data', 'target', 'feature_names', 'DESCR', 'filename', 'data_module'])
```

[24]:
```
boston = pd.DataFrame(boston_dataset.data, columns=boston_dataset.feature_names)
boston.head()
```

[24]:

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94

Console

Type here to search

11:35 AM 28/03/2022

New Tab | notebook6fe3d47a5b | Kaggle | (3) WhatsApp | Boston Housing | Kaggle

kaggle.com/code/kshitijshinde/notebook6fe3d47a5b/edit

notebook6fe3d47a5b Draft saved

File Edit View Run Add-ons Help

Share | Save Version 0

Run All | Code | Draft Session (54m)

Data | + Add data

Input

Output (44.1MB / 19.6GB)

/kaggle/working

Settings

Schedule a notebook run

Code Help

FIND CODE HELP

Find Code Help

Search for examples of how to do things

[24]:
```
boston = pd.DataFrame(boston_dataset.data, columns=boston_dataset.feature_names)
boston.head()
```

[24]:

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33

[25]:
```
boston['MEDV'] = boston_dataset.target
```

[26]:
```
boston.isnull().sum()
```

[26]:
```
CRIM       0
ZN         0
INDUS      0
CHAS       0
NOX        0
RM         0
```

Console

Type here to search

11:35 AM 28/03/2022

New Tab | notebook6fe3d47a5b | Kaggle | (3) WhatsApp | Boston Housing | Kaggle

kaggle.com/code/kshitijshinde/notebook6fe3d47a5b/edit

notebook6fe3d47a5b Draft saved

File Edit View Run Add-ons Help

Share | Save Version 0

Run All | Code | Draft Session (54m)

Data | + Add data

Input

Output (44.1MB / 19.6GB)

/kaggle/working

Settings

Schedule a notebook run

Code Help

FIND CODE HELP

Find Code Help

Search for examples of how to do things

[26]:
```
boston.isnull().sum()
```


[26]:
```
CRIM       0
ZN         0
INDUS      0
CHAS       0
NOX        0
RM         0
AGE        0
DIS        0
RAD        0
TAX        0
PTRATIO    0
B          0
LSTAT      0
MEDV       0
dtype: int64
```

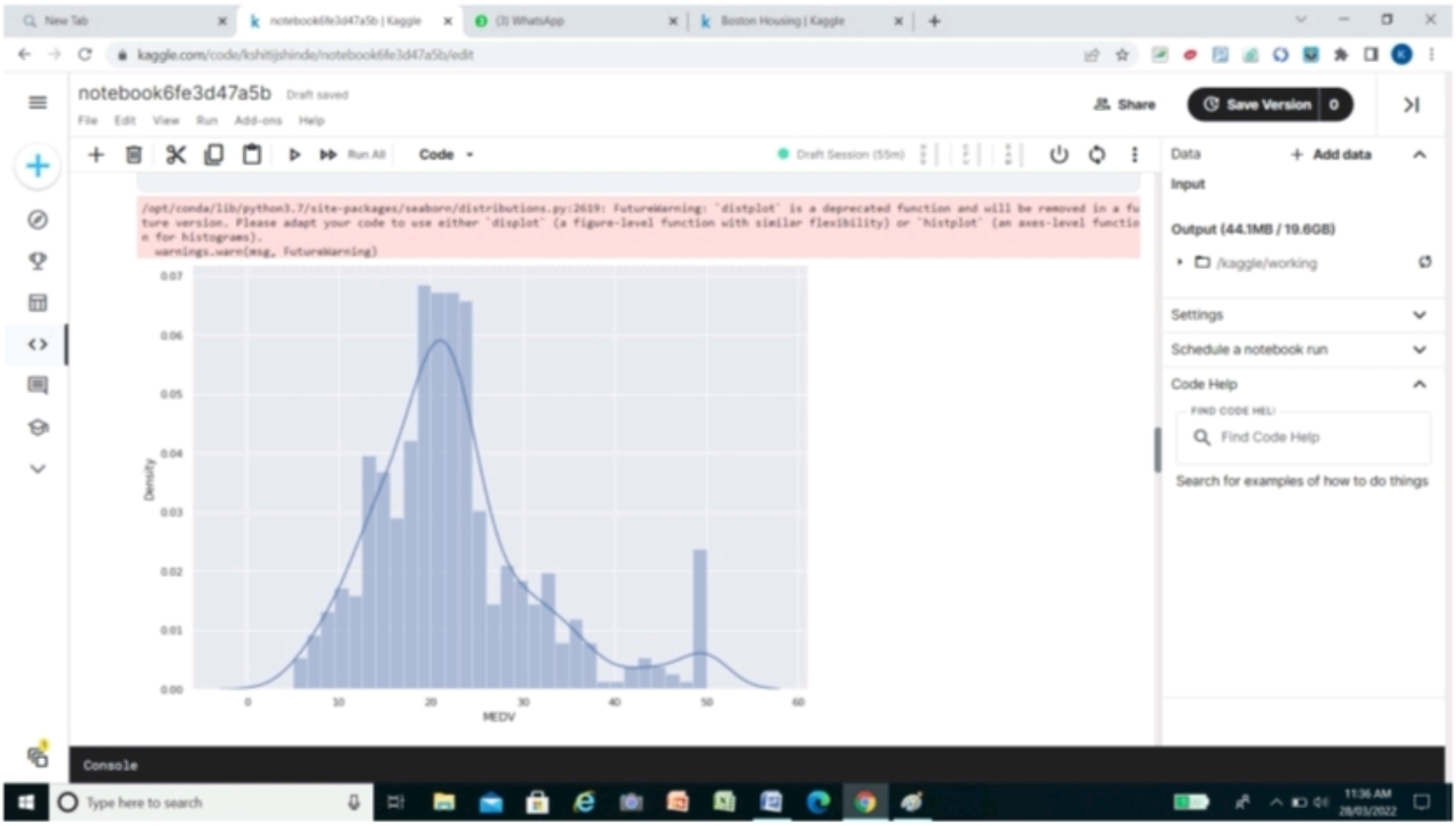
[27]:
```
sns.set(rc={'figure.figsize':(11.7,8.27)})
sns.distplot(boston['MEDV'], bins=30)
plt.show()
```

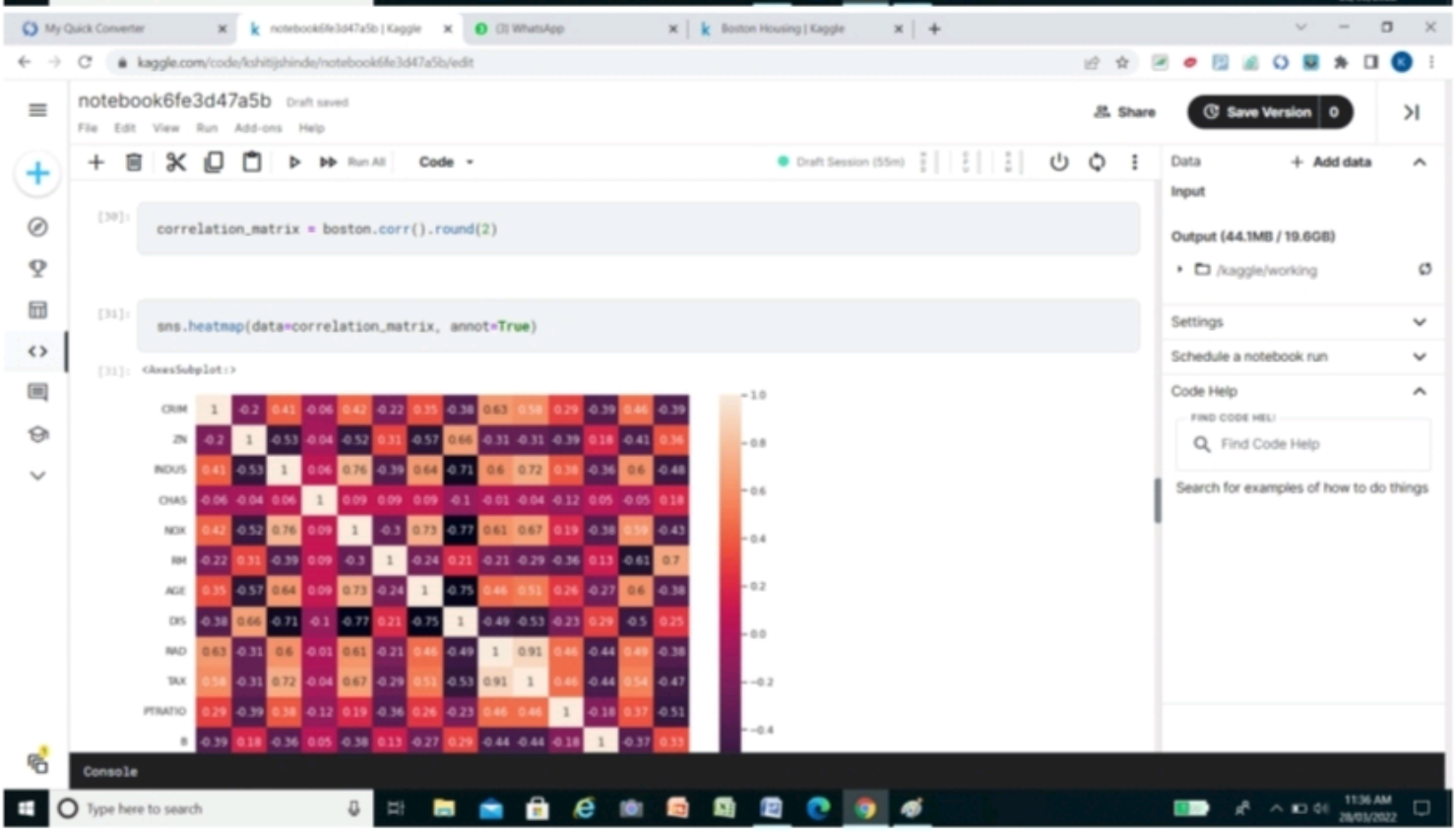
```
/opt/conda/lib/python3.7/site-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated function and will be removed in a fu
ture version. Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level functio
n for histograms).
  warnings.warn(msg, FutureWarning)
```

Console

Type here to search

11:35 AM 28/03/2022


notebook6fe3d47a5b | Kaggle
(3) WhatsApp
Boston Housing | Kaggle
kaggle.com/code/kshitijshinde/notebook6fe3d47a5b/edit
notebook6fe3d47a5b
Draft saved
File Edit View Run Add-ons Help
Share
Save Version 0
Run All
Code
Draft Session (55m)
/opt/conda/lib/python3.7/site-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).
warnings.warn(msg, FutureWarning)
Density
MEDV
Data
Add data
Input
Output (44.1MB / 19.6GB)
/kaggle/working
Settings
Schedule a notebook run
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Console
Type here to search
11:36 AM
28/01/2022


My Quick Converter
notebook6fe3d47a5b | Kaggle
(3) WhatsApp
Boston Housing | Kaggle
kaggle.com/code/kshitijshinde/notebook6fe3d47a5b/edit
notebook6fe3d47a5b
Draft saved
File Edit View Run Add-ons Help
Share
Save Version 0
Run All
Code
Draft Session (55m)
[30]:
correlation_matrix = boston.corr().round(2)
[31]:
sns.heatmap(data=correlation_matrix, annot=True)
[31]:
<AxesSubplot:>
CRIM ZN INDUS CHAS NOX RM AGE DIS RAD TAX PTRATIO B
Data
Add data
Input
Output (44.1MB / 19.6GB)
/kaggle/working
Settings
Schedule a notebook run
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Console
Type here to search
11:36 AM
28/01/2022

My Quick Converter | notebook6fe3d47a5b | Kaggle | (3) WhatsApp | Boston Housing | Kaggle

kaggle.com/code/kshitijshinde/notebook6fe3d47a5b/edit

notebook6fe3d47a5b Draft saved

File Edit View Run Add-ons Help

Share | Save Version 0

Run All | Code | Draft Session (55m)

Data | + Add data

Input

Output (44.1MB / 19.6GB)

/kaggle/working

Settings

Schedule a notebook run

Code Help

FIND CODE HELP

Find Code Help

Search for examples of how to do things

[31]:
```
sns.heatmap(data=correlation_matrix, annot=True)
```

[31]: <AxesSubplot:>

Console

Type here to search

11:37 AM 28/03/2022

New Tab | notebook6fe3d47a5b | Kaggle | (3) WhatsApp | Boston Housing | Kaggle

kaggle.com/code/kshitijshinde/notebook6fe3d47a5b/edit

notebook6fe3d47a5b Draft saved

File Edit View Run Add-ons Help

Share | Save Version 0

Run All | Code | Draft Session (56m)

Data | + Add data

Input

Output (44.1MB / 19.6GB)

/kaggle/working

Settings

Schedule a notebook run

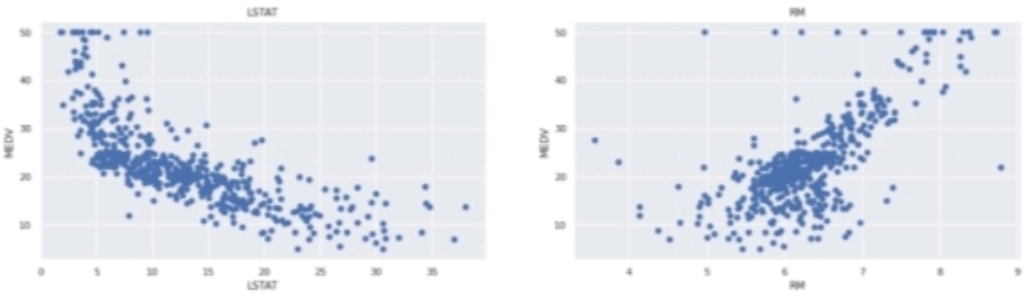
Code Help

FIND CODE HELP

Find Code Help

Search for examples of how to do things

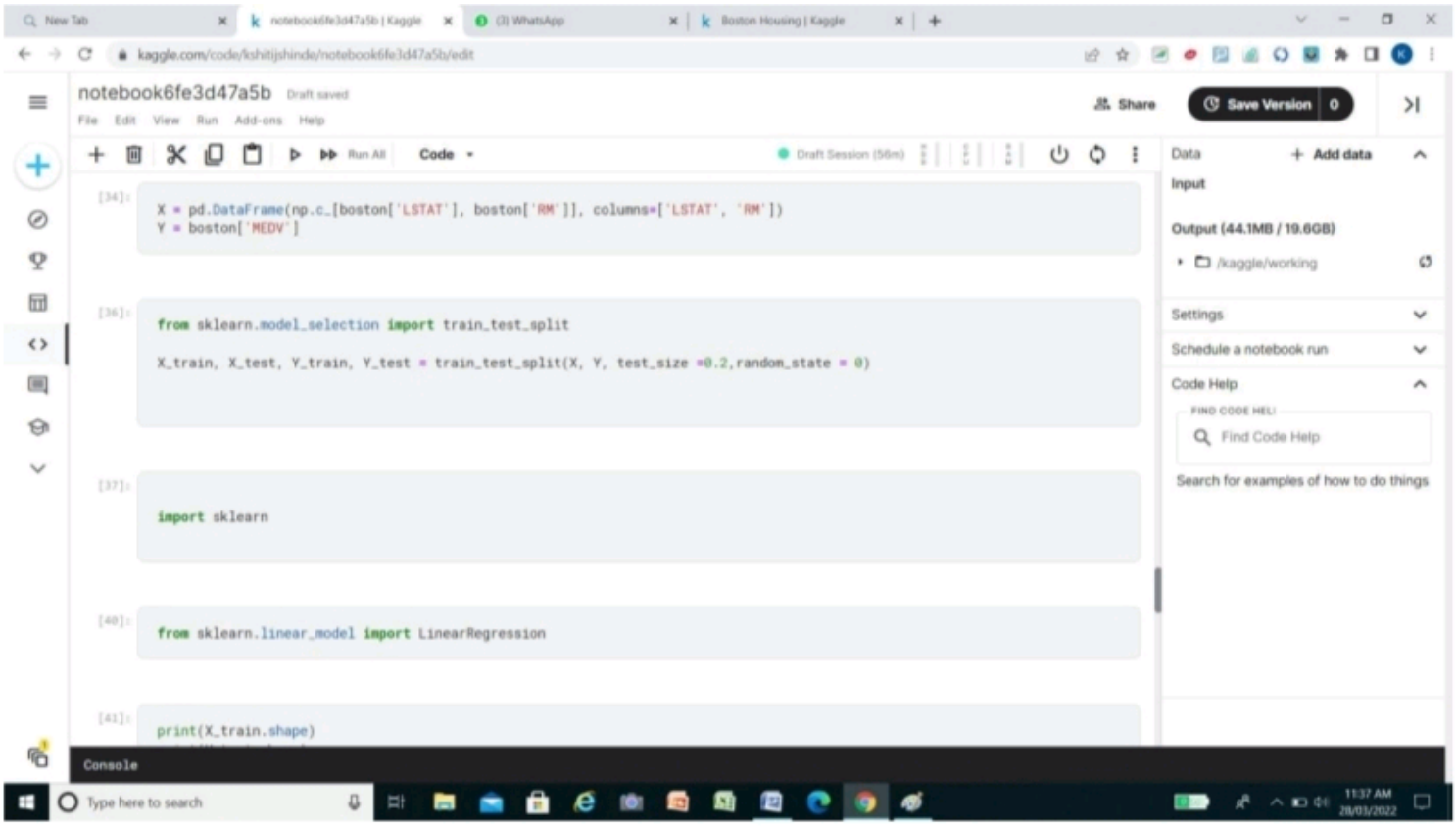
[33]:
```
plt.figure(figsize=(20, 5))
features = ['LSTAT' , 'RM']
target = boston['MEDV']
for i, col in enumerate(features):
    plt.subplot(1, len(features), i+1)
    x = boston[col]
    y = target
    plt.scatter(x, y, marker='o')
    plt.title(col)
    plt.xlabel(col)
    plt.ylabel('MEDV')
```

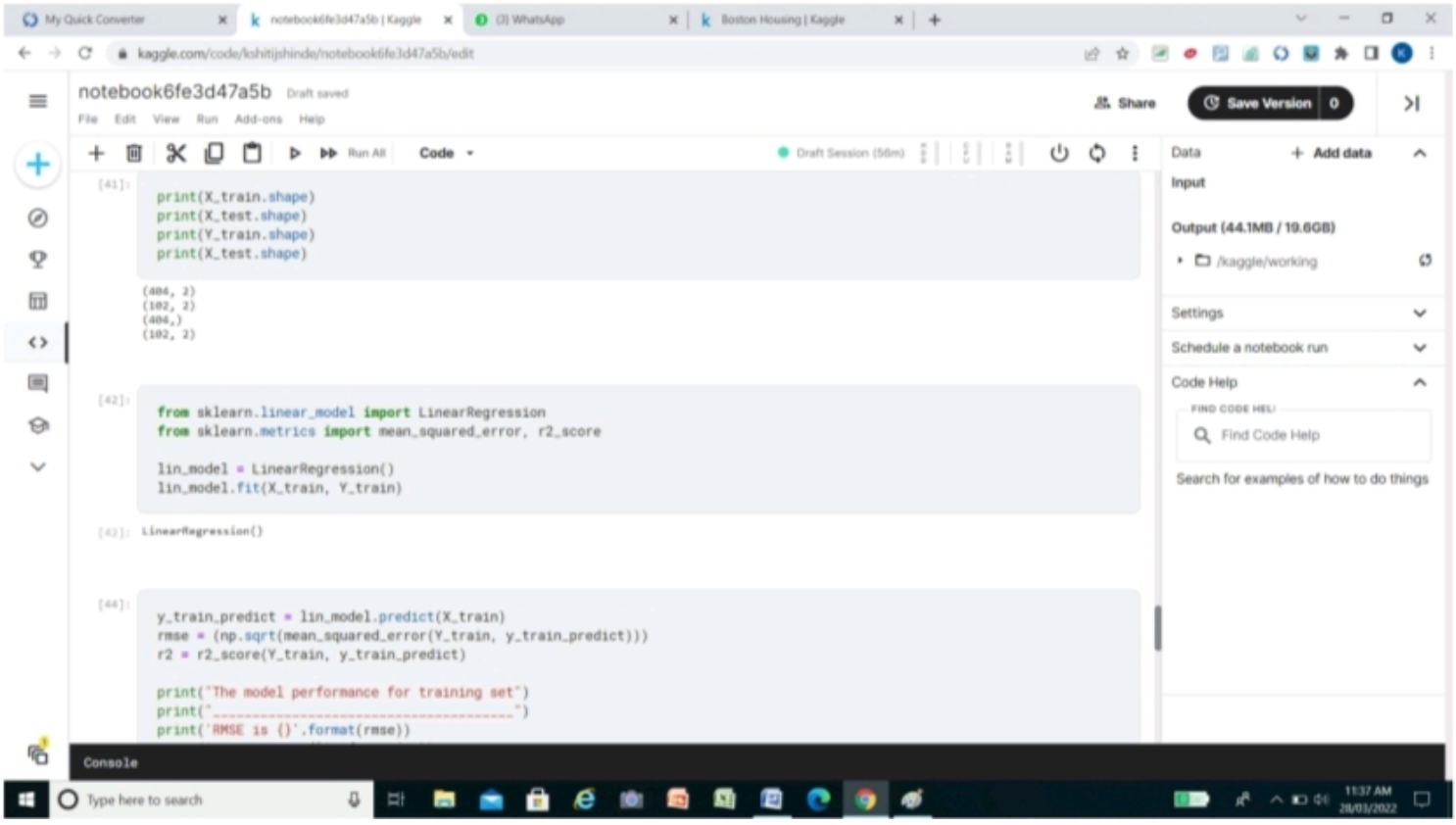


Console

Type here to search

11:37 AM 28/03/2022


notebook6fe3d47a5b
Draft saved
File Edit View Run Add-ons Help
Share
Save Version 0
Run All
Code
Draft Session (56m)
Data
Add data
Input
Output (44.1MB / 19.6GB)
/kaggle/working
Settings
Schedule a notebook run
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
[34]:
X = pd.DataFrame(np.c_[boston['LSTAT'], boston['RM']], columns=['LSTAT', 'RM'])
Y = boston['MEDV']
[36]:
from sklearn.model_selection import train_test_split
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size =0.2,random_state = 0)
[37]:
import sklearn
[40]:
from sklearn.linear_model import LinearRegression
[41]:
print(X_train.shape)
Console


notebook6fe3d47a5b
Draft saved
File Edit View Run Add-ons Help
Share
Save Version 0
Run All
Code
Draft Session (56m)
Data
Add data
Input
Output (44.1MB / 19.6GB)
/kaggle/working
Settings
Schedule a notebook run
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
[41]:
print(X_train.shape)
print(X_test.shape)
print(Y_train.shape)
print(X_test.shape)
(404, 2)
(102, 2)
(404,)
(102, 2)
[42]:
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
lin_model = LinearRegression()
lin_model.fit(X_train, Y_train)
[42]: LinearRegression()
[44]:
y_train_predict = lin_model.predict(X_train)
rmse = (np.sqrt(mean_squared_error(Y_train, y_train_predict)))
r2 = r2_score(Y_train, y_train_predict)
print("The model performance for training set")
print("--------------------------------------")
print('RMSE is {}'.format(rmse))
Console

New Tab | notebook6fe3d47a5b | Kaggle | (3) WhatsApp | Boston Housing | Kaggle

kaggle.com/code/kshitijshinde/notebook6fe3d47a5b/edit

notebook6fe3d47a5b Draft saved

File Edit View Run Add-ons Help

Share | Save Version 0

Run All | Code | Draft Session (57m)

Data | + Add data

Input

Output (44.1MB / 19.6GB)

/kaggle/working

Settings

Schedule a notebook run

Code Help

FIND CODE HELP

Find Code Help

Search for examples of how to do things

[44]:

```
y_train_predict = lin_model.predict(X_train)
rmse = (np.sqrt(mean_squared_error(Y_train, y_train_predict)))
r2 = r2_score(Y_train, y_train_predict)

print("The model performance for training set")
print("--------------------------------------")
print('RMSE is {}'.format(rmse))
print('R2 score is {}' .format(r2))
print("/ln")

y_test_predict = lin_model.predict(X_test)
rmse = (np.sqrt(mean_squared_error(Y_test, y_test_predict)))

r2 = r2_score(Y_test, y_test_predict)

print("The model performance for training set")
print("--------------------------------------")
print('RMSE is {}'.format(rmse))
print('R2 score is {}' .format(r2))
```

```
The model performance for training set
--------------------------------------
RMSE is 5.365657134224422
R2 score is 0.6618625964841893
/ln
The model performance for training set
--------------------------------------
RMSE is 6.114172522817781
R2 score is 0.5489084827186417
```

Console

Type here to search

11:38 AM 28/03/2022

My Quick Converter | notebook6fe3d47a5b | Kaggle | (3) WhatsApp | Boston Housing | Kaggle

kaggle.com/code/kshitijshinde/notebook6fe3d47a5b/edit

notebook6fe3d47a5b Draft saved

File Edit View Run Add-ons Help

Share | Save Version 0

Run All | Code | Draft Session (57m)

Data | + Add data

Input

Output (44.1MB / 19.6GB)

/kaggle/working

Settings

Schedule a notebook run

Code Help

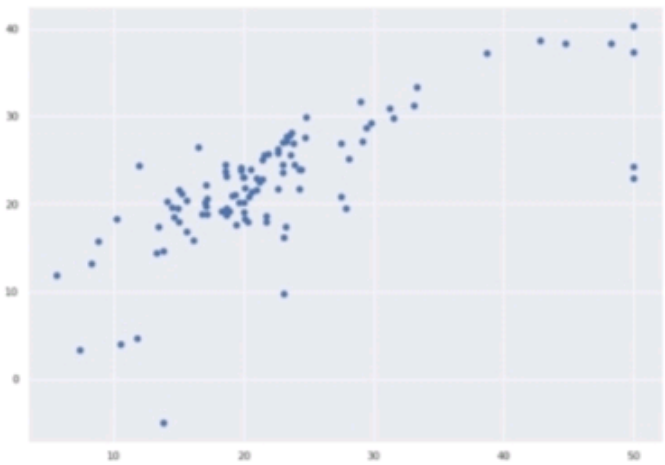
FIND CODE HELP

Find Code Help

Search for examples of how to do things

[45]:

```
plt.scatter(Y_test, y_test_predict)
plt.show()
```



Console

Type here to search

11:38 AM 28/03/2022

Assignment No. 5

Aim: Implement logistic regression using python/r to perform classification on Social Network Ads.csv dataset.

OUTPUT :

In [1]:

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd

dataset = pd.read_csv('../input/Social_Network_Ads.csv')
dataset.head()
```

Out[1]:

	User ID	Gender	Age	EstimatedSalary	Purchased
0	15624510	Male	19	19000	0
1	15810944	Male	35	20000	0
2	15668575	Female	26	43000	0
3	15603246	Female	27	57000	0
4	15804002	Male	19	76000	0

In [2]:

```
X = dataset.iloc[:, [2, 3]].values
y = dataset.iloc[:, 4].values

print(X[:3, :])
print('-'*15)
print(y[:3])
```

```
[[   19 19000]
 [   35 20000]
 [   26 43000]]
---------------
[0 0 0]
```

In [3]:

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.25, random_state = 0)

print(X_train[:3])
print('-'*15)
print(y_train[:3])
print('-'*15)
print(X_test[:3])
print('-'*15)
print(y_test[:3])
```

```
[[    44  39000]
 [    32 120000]
 [    38  50000]]
---------------
[0 1 0]
---------------
[[   30 87000]
 [   38 50000]
 [   35 75000]]
---------------
[0 0 0]
```

In [4]:

```
from sklearn.preprocessing import StandardScaler
sc_X = StandardScaler()
X_train = sc_X.fit_transform(X_train)
X_test = sc_X.transform(X_test)
```

```
/opt/conda/lib/python3.6/site-packages/sklearn/utils/validation.py:595: DataConversionWarning: Data with input dtype int64 was converted to float64 by StandardScaler.
  warnings.warn(msg, DataConversionWarning)
/opt/conda/lib/python3.6/site-packages/sklearn/utils/validation.py:595: DataConversionWarning: Data with input dtype int64 was converted to float64 by StandardScaler.
  warnings.warn(msg, DataConversionWarning)
/opt/conda/lib/python3.6/site-packages/sklearn/utils/validation.py:595: DataConversionWarning: Data with input dtype int64 was converted to float64 by StandardScaler.
  warnings.warn(msg, DataConversionWarning)
```

In [5]:

```
print(X_train[:3])
print('-'*15)
print(X_test[:3])
```

```
[[ 0.58164944 -0.88670699]
 [-0.60673761  1.46173768]
 [-0.01254409 -0.5677824 ]]
---------------
[[-0.80480212  0.50496393]
 [-0.01254409 -0.5677824 ]
 [-0.30964085  0.1570462 ]]
```

In [6]:

```
from sklearn.linear_model import LogisticRegression
classifier = LogisticRegression(random_state = 0, solver='lbfgs' )
classifier.fit(X_train, y_train)
y_pred = classifier.predict(X_test)

print(X_test[:10])
print('-'*15)
print(y_pred[:10])
```

```
[[-0.80480212  0.50496393]
 [-0.01254409 -0.5677824 ]
 [-0.30964085  0.1570462 ]
 [-0.80480212  0.27301877]
 [-0.30964085 -0.5677824 ]
 [-1.10189888 -1.43757673]
 [-0.70576986 -1.58254245]
 [-0.21060859  2.15757314]
 [-1.99318916 -0.04590581]
 [ 0.8787462  -0.77073441]]
---------------
[0 0 0 0 0 0 0 1 0 1]
```

In [7]:

```
print(y_pred[:20])
print(y_test[:20])
```

```
[0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 1 0]
[0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0]
```

In [8]:

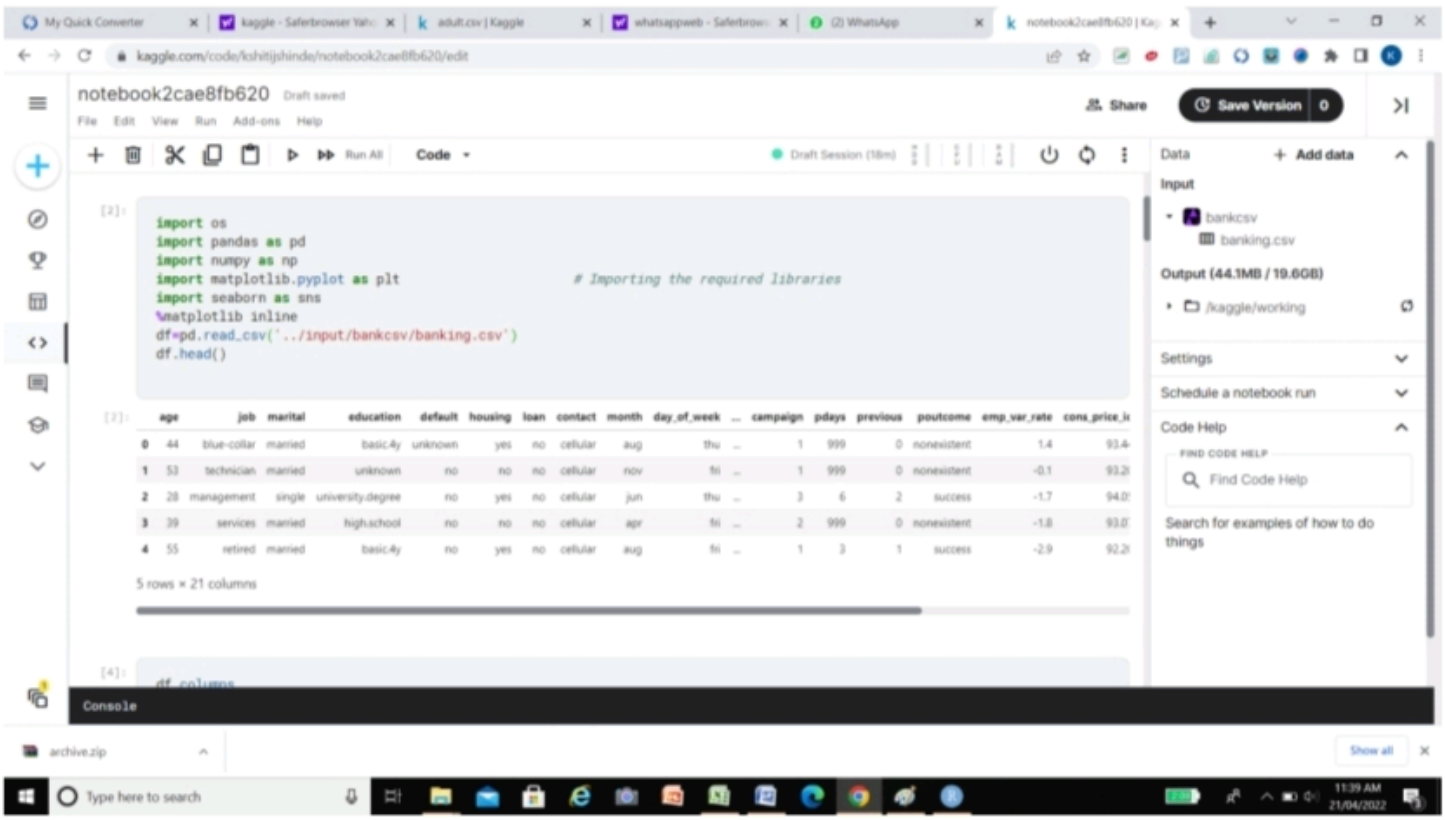
```
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, y_pred)
print(cm)
```

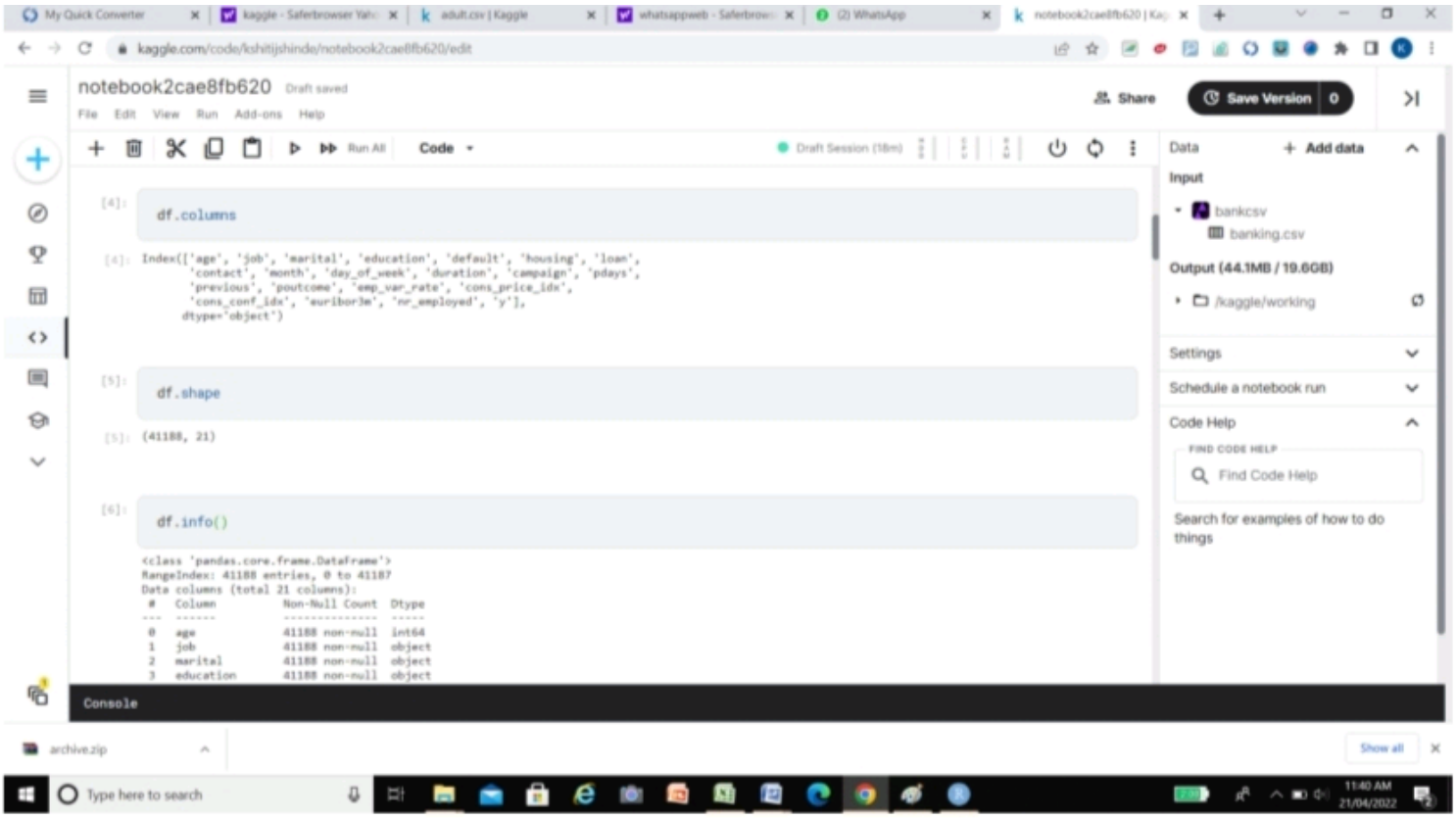
```
[[65  3]
 [ 8 24]]
```

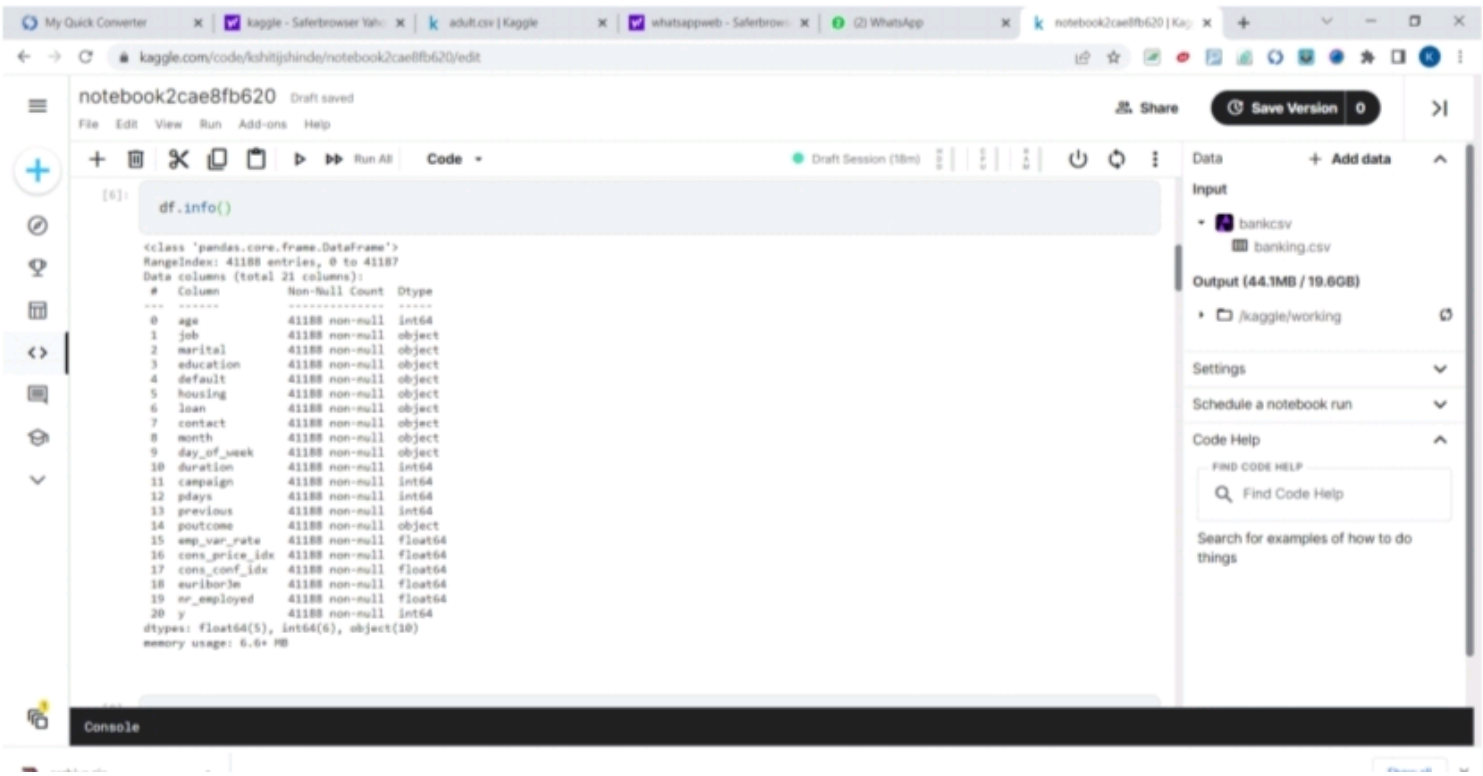
Assignment No. 5.2

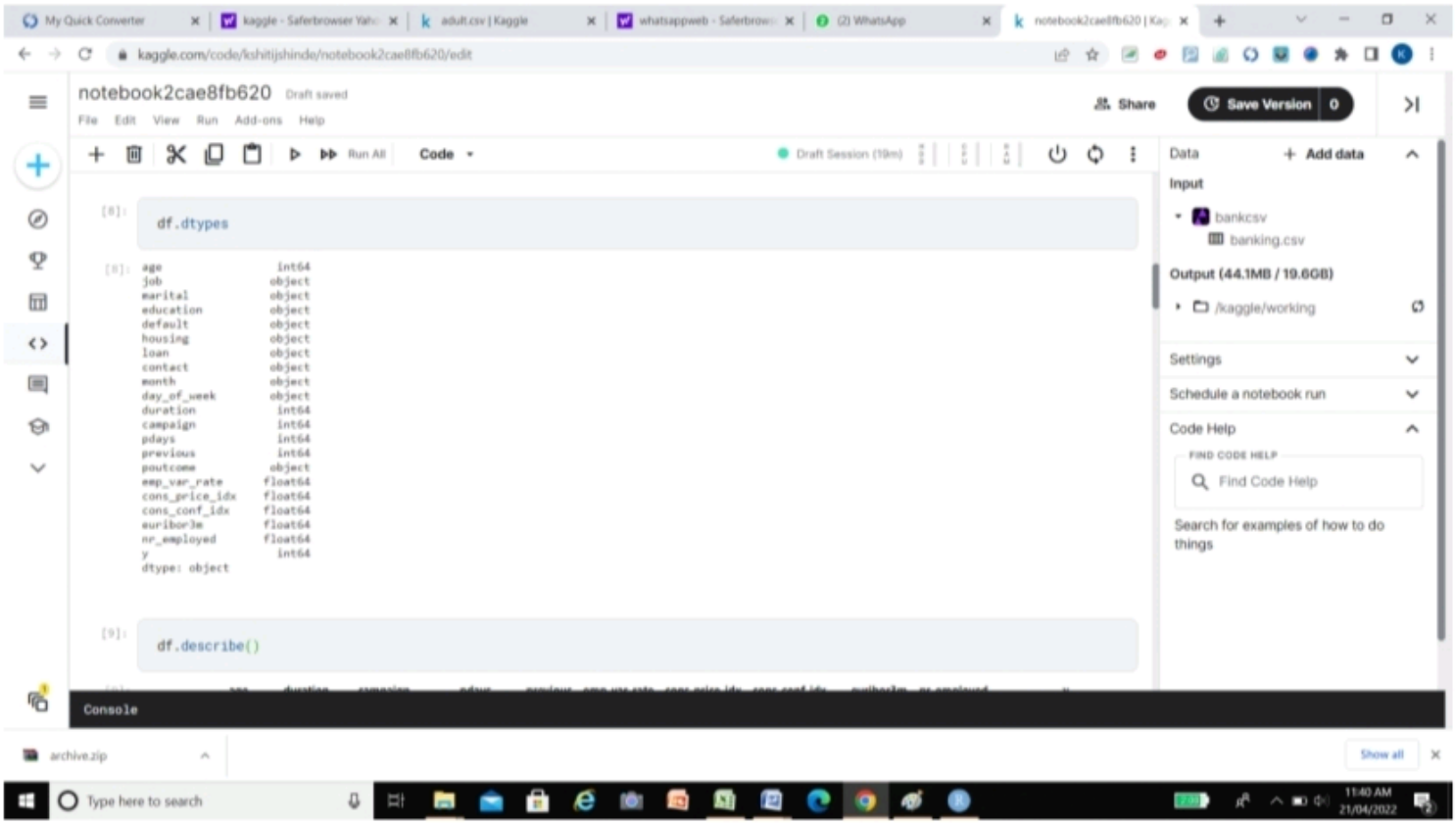
Aim: Data Analytics 2 : Compute Confusion matrix to find TP, FP, TN, FN, Accuracy, Error rate, Precision, Recall on the given set.

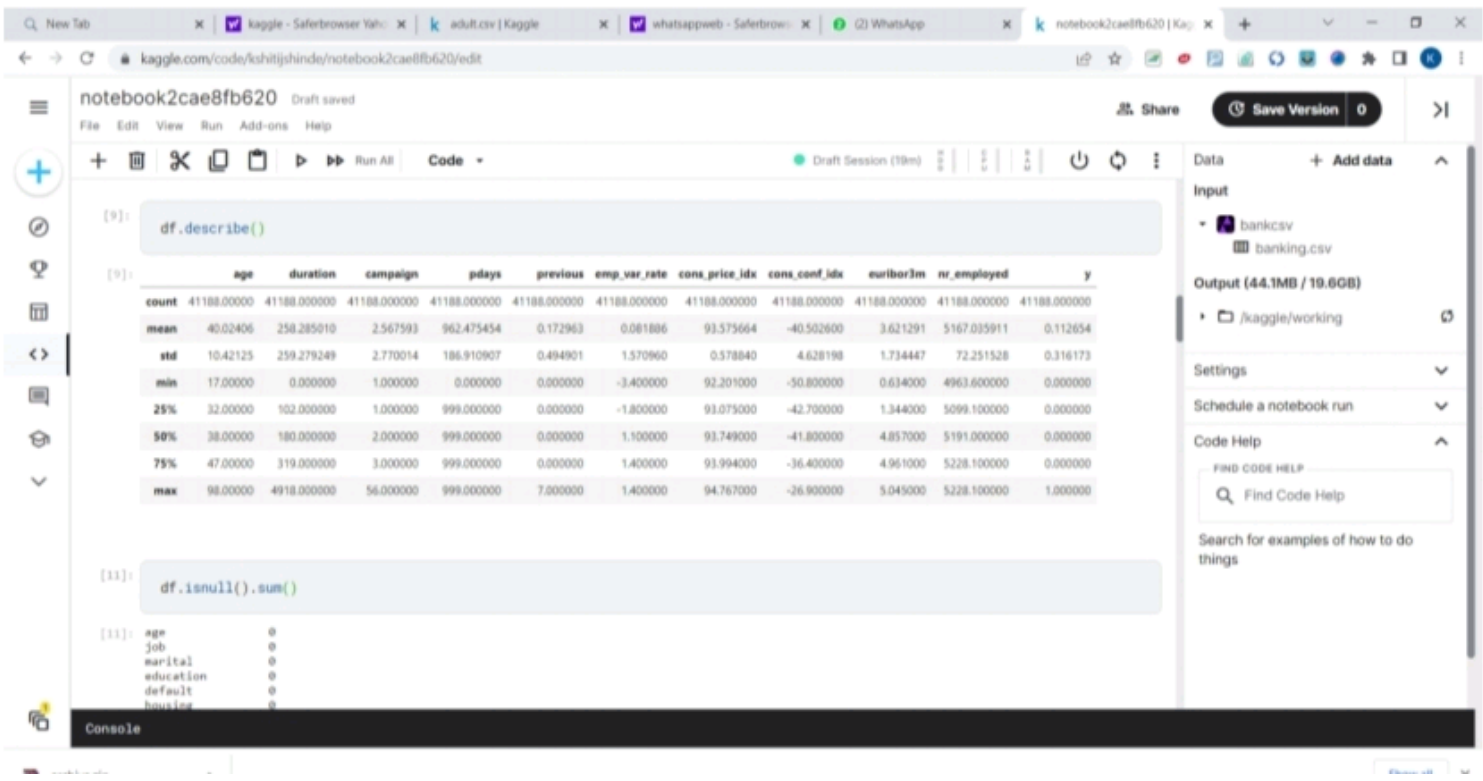
OUTPUT :

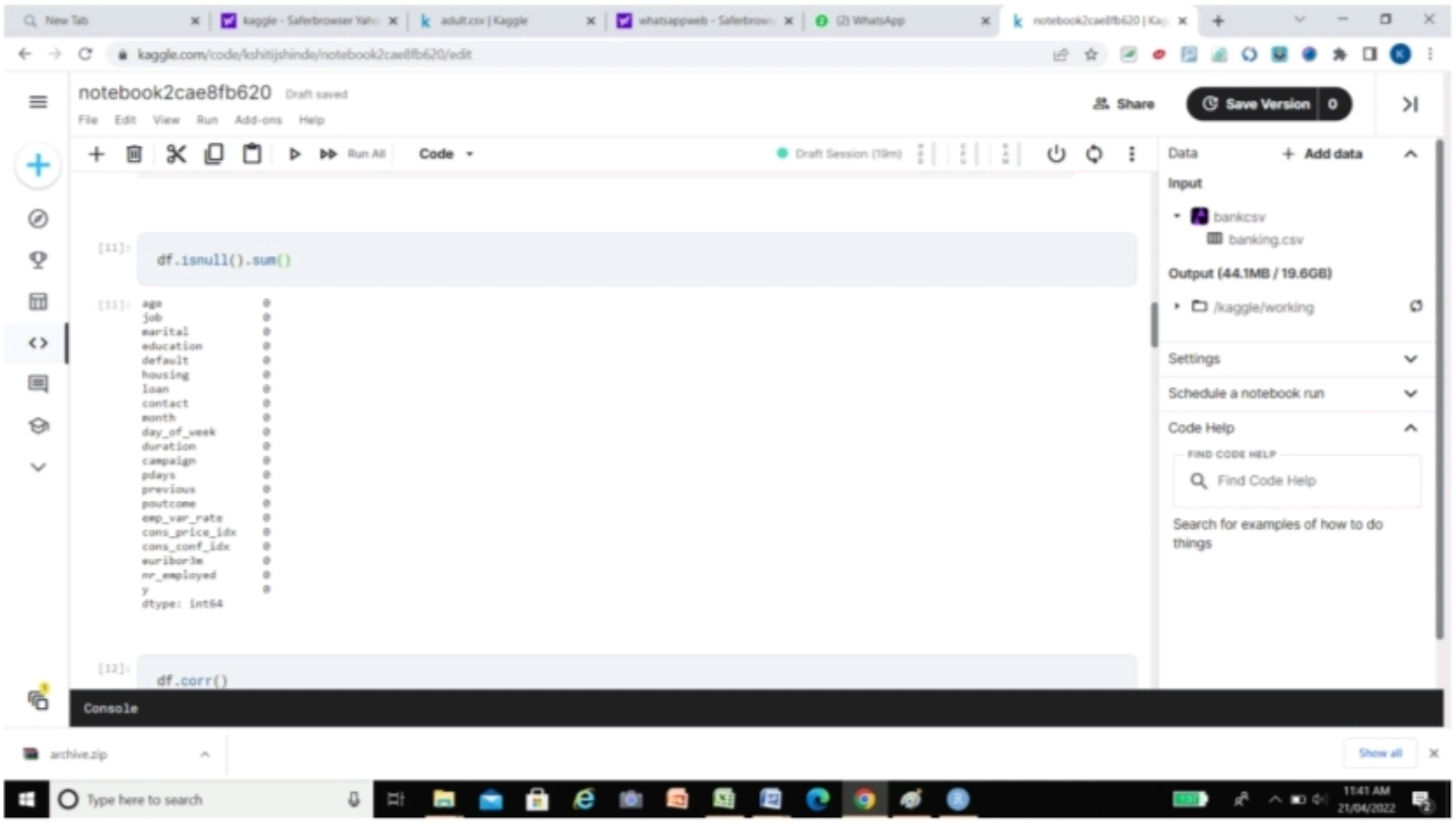


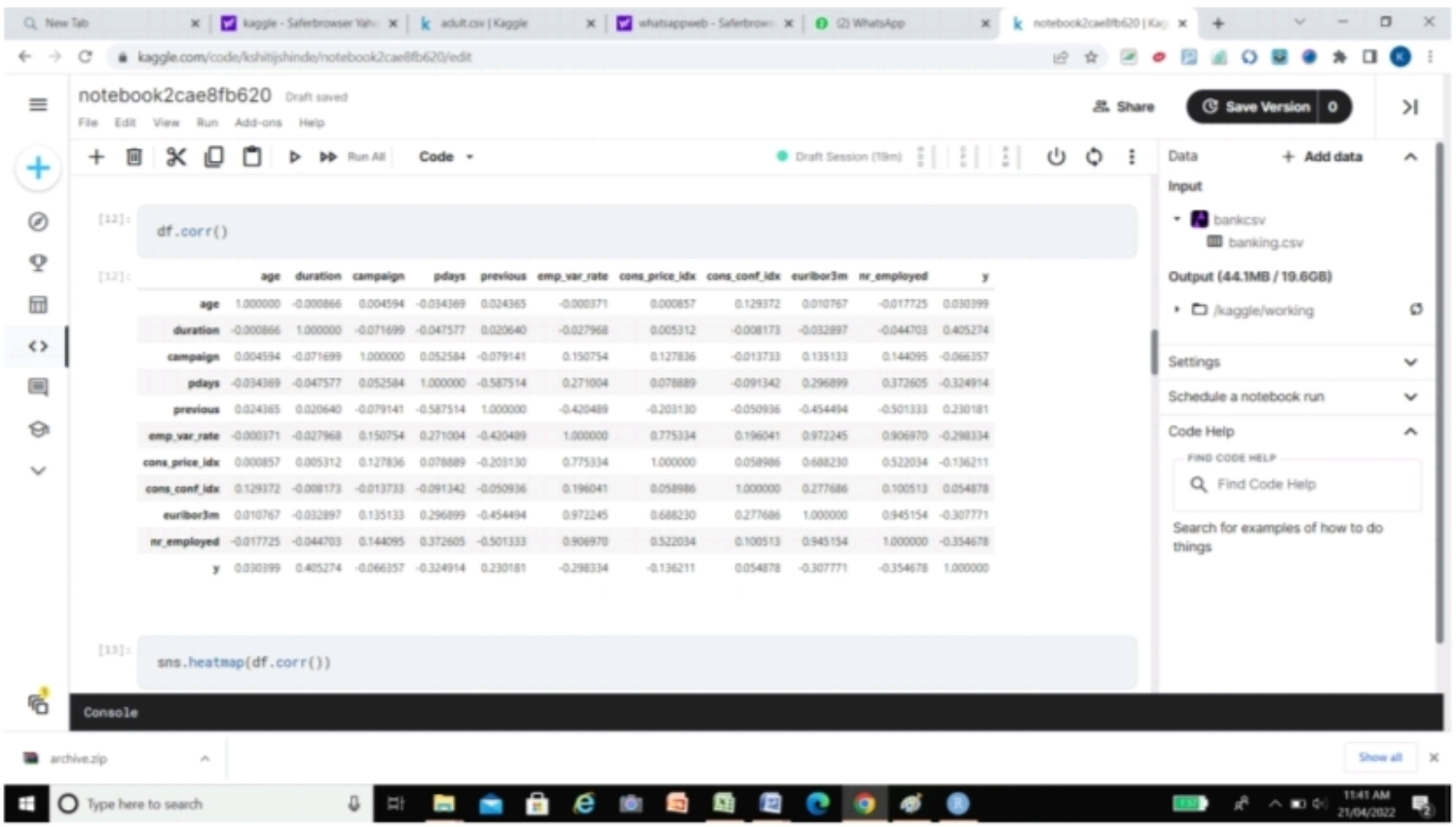

My Quick Converter
kaggle - Saferbrowser Yah
adult.csv | Kaggle
whatsappweb - Saferbrow
(2) WhatsApp
notebook2cae8fb620 | Kag
kaggle.com/code/kshitijshinde/notebook2cae8fb620/edit
notebook2cae8fb620
Draft saved
File Edit View Run Add-ons Help
Share
Save Version 0
Run All
Code
Draft Session (18m)
[4]: df.columns
[4]: Index(['age', 'job', 'marital', 'education', 'default', 'housing', 'loan',
'contact', 'month', 'day_of_week', 'duration', 'campaign', 'pdays',
'previous', 'poutcome', 'emp_var_rate', 'cons_price_idx',
'cons_conf_idx', 'euribor3m', 'nr_employed', 'y'],
dtype='object')
[5]: df.shape
[5]: (41188, 21)
[6]: df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41188 entries, 0 to 41187
Data columns (total 21 columns):
Column Non-Null Count Dtype
--- ------ -------------- -----
0 age 41188 non-null int64
1 job 41188 non-null object
2 marital 41188 non-null object
3 education 41188 non-null object
Data
Add data
Input
bankcsv
banking.csv
Output (44.1MB / 19.6GB)
/kaggle/working
Settings
Schedule a notebook run
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Console
archive.zip
Show all
Type here to search
11:40 AM
21/04/2022

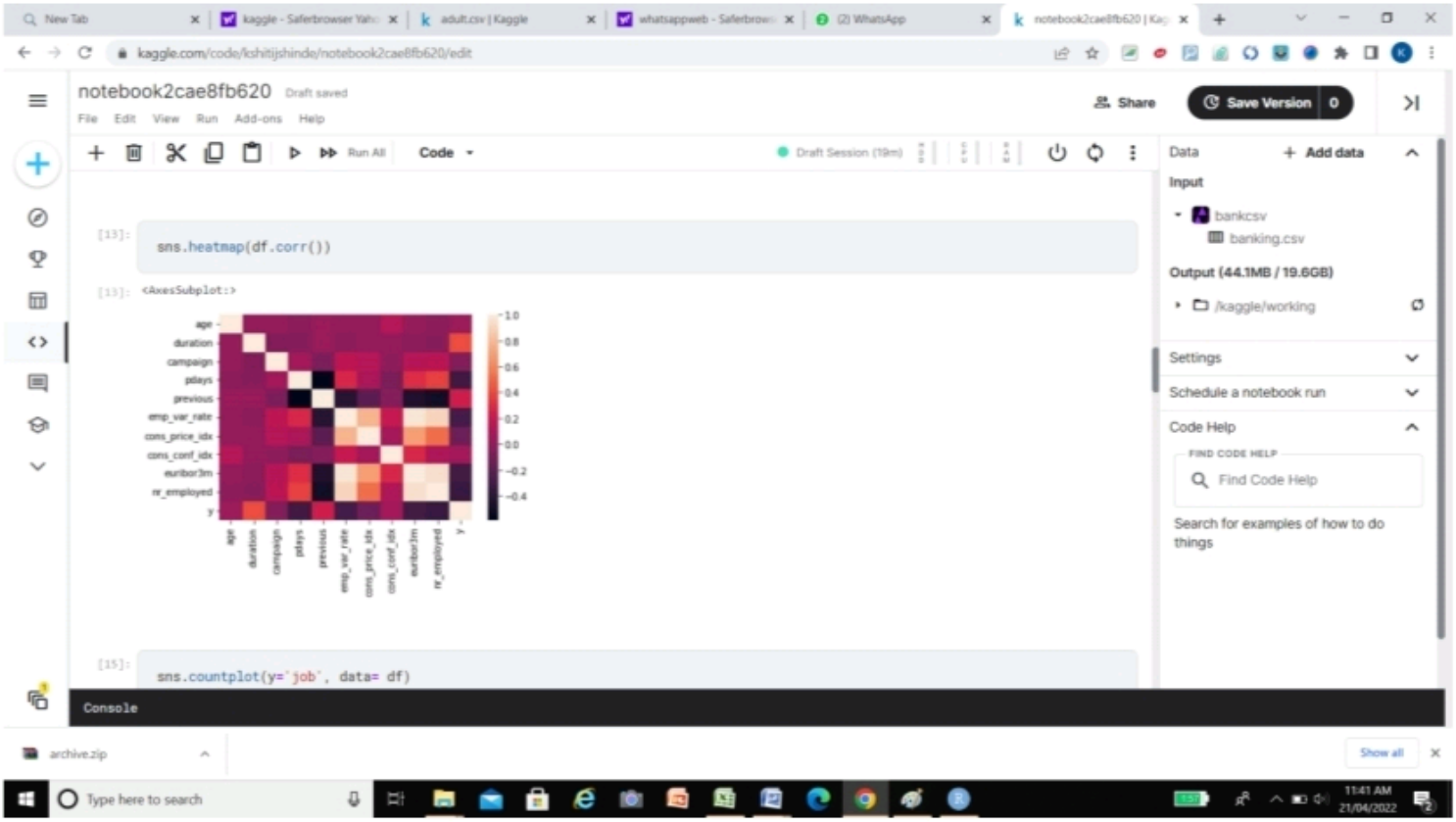

My Quick Converter
kaggle - Saferbrowser Yah
adult.csv | Kaggle
whatsappweb - Saferbrow
(2) WhatsApp
notebook2cae8fb620 | Kag
kaggle.com/code/kshitijshinde/notebook2cae8fb620/edit
notebook2cae8fb620
Draft saved
File Edit View Run Add-ons Help
Share
Save Version 0
Run All
Code
Draft Session (18m)
[6]: df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41188 entries, 0 to 41187
Data columns (total 21 columns):
Column Non-Null Count Dtype
--- ------ -------------- -----
0 age 41188 non-null int64
1 job 41188 non-null object
2 marital 41188 non-null object
3 education 41188 non-null object
4 default 41188 non-null object
5 housing 41188 non-null object
6 loan 41188 non-null object
7 contact 41188 non-null object
8 month 41188 non-null object
9 day_of_week 41188 non-null object
10 duration 41188 non-null int64
11 campaign 41188 non-null int64
12 pdays 41188 non-null int64
13 previous 41188 non-null int64
14 poutcome 41188 non-null object
15 emp_var_rate 41188 non-null float64
16 cons_price_idx 41188 non-null float64
17 cons_conf_idx 41188 non-null float64
18 euribor3m 41188 non-null float64
19 nr_employed 41188 non-null float64
20 y 41188 non-null int64
dtypes: float64(5), int64(6), object(10)
memory usage: 6.6+ MB
Data
Add data
Input
bankcsv
banking.csv
Output (44.1MB / 19.6GB)
/kaggle/working
Settings
Schedule a notebook run
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Console

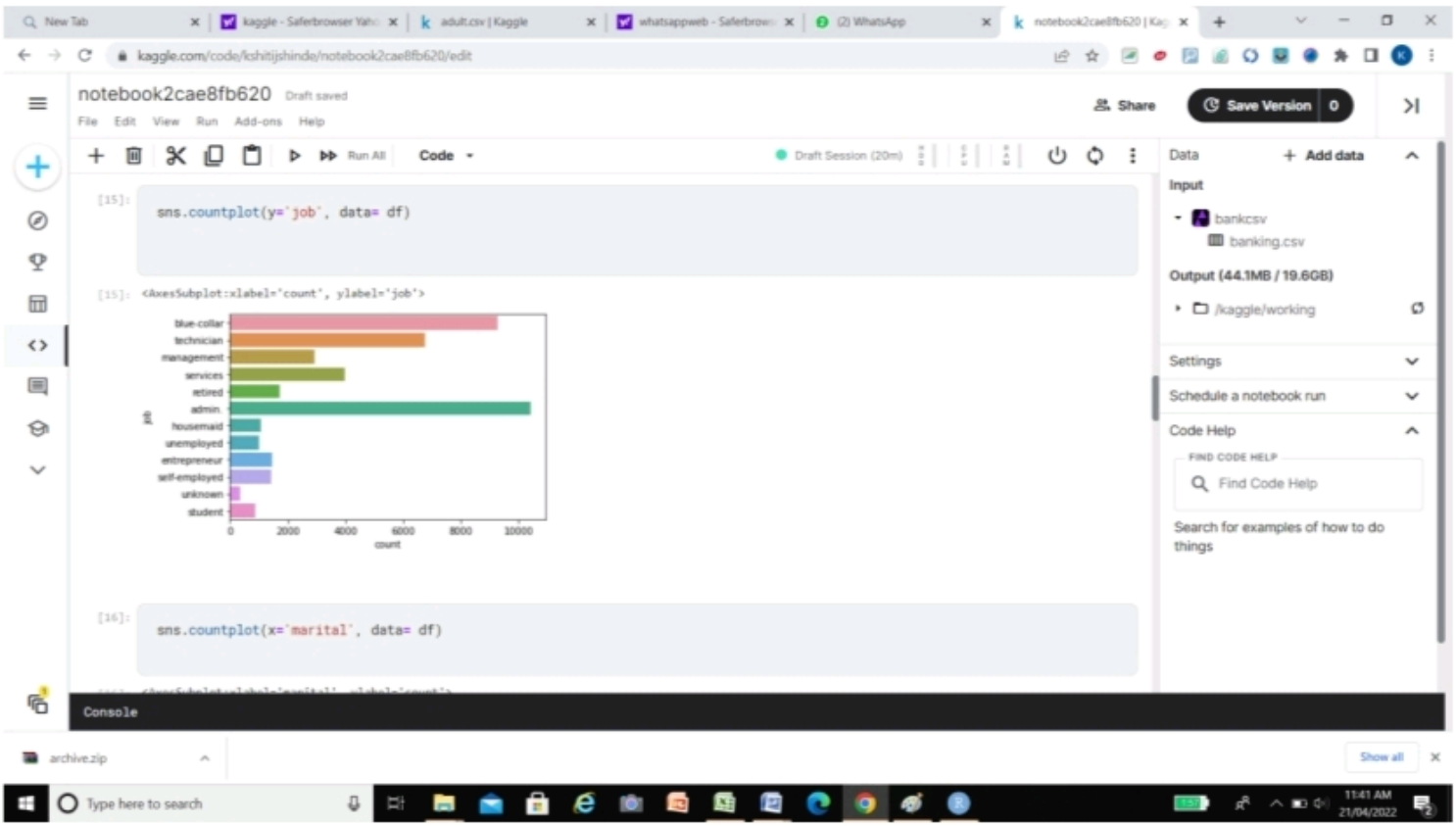

notebook2cae8fb620
Draft saved
File Edit View Run Add-ons Help
Share
Save Version 0
Run All
Code
Draft Session (19m)
[8]:
df.dtypes
age int64
job object
marital object
education object
default object
housing object
loan object
contact object
month object
day_of_week object
duration int64
campaign int64
pdays int64
previous int64
poutcome object
emp_var_rate float64
cons_price_idx float64
cons_conf_idx float64
euribor3m float64
nr_employed float64
y int64
dtype: object
[9]:
df.describe()
Data
Add data
Input
bankcsv
banking.csv
Output (44.1MB / 19.6GB)
/kaggle/working
Settings
Schedule a notebook run
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Console
archive.zip
Show all
Type here to search
11:40 AM
21/04/2022


notebook2cae8fb620
Draft saved
File Edit View Run Add-ons Help
Share
Save Version 0
Run All
Code
Draft Session (19m)
[9]:
df.describe()
[9]:
age duration campaign pdays previous emp_var_rate cons_price_idx cons_conf_idx euribor3m nr_employed y
count 41188.00000 41188.000000 41188.000000 41188.000000 41188.000000 41188.000000 41188.000000 41188.000000 41188.000000 41188.000000 41188.000000
mean 40.02406 258.285010 2.567593 962.475454 0.172963 0.081886 93.575664 -40.502600 3.621291 5167.035911 0.112654
std 10.42125 259.279249 2.770014 186.910907 0.494901 1.570960 0.578840 4.628198 1.734447 72.251528 0.316173
min 17.00000 0.000000 1.000000 0.000000 0.000000 -3.400000 92.201000 -50.800000 0.634000 4963.600000 0.000000
25% 32.00000 102.000000 1.000000 999.000000 0.000000 -1.800000 93.075000 -42.700000 1.344000 5099.100000 0.000000
50% 38.00000 180.000000 2.000000 999.000000 0.000000 1.100000 93.749000 -41.800000 4.857000 5191.000000 0.000000
75% 47.00000 319.000000 3.000000 999.000000 0.000000 1.400000 93.994000 -36.400000 4.961000 5228.100000 0.000000
max 98.00000 4918.000000 56.000000 999.000000 7.000000 1.400000 94.767000 -26.900000 5.045000 5228.100000 1.000000
[11]:
df.isnull().sum()
[11]:
age 0
job 0
marital 0
education 0
default 0
housing 0
Data
Add data
Input
bankcsv
banking.csv
Output (44.1MB / 19.6GB)
/kaggle/working
Settings
Schedule a notebook run
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Console


notebook2cae8fb620
Draft saved
File Edit View Run Add-ons Help
Share
Save Version 0
Run All
Code
Draft Session (19m)
[11]:
df.isnull().sum()
[11]:
age 0
job 0
marital 0
education 0
default 0
housing 0
loan 0
contact 0
month 0
day_of_week 0
duration 0
campaign 0
pdays 0
previous 0
poutcome 0
emp_var_rate 0
cons_price_idx 0
cons_conf_idx 0
euribor3m 0
nr_employed 0
y 0
dtype: int64
[12]:
df.corr()
Console
Data
Add data
Input
bankcsv
banking.csv
Output (44.1MB / 19.6GB)
/kaggle/working
Settings
Schedule a notebook run
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things


notebook2cae8fb620
Draft saved
File Edit View Run Add-ons Help
Share
Save Version 0
Run All
Code
Draft Session (19m)
[12]:
df.corr()
[12]:
age duration campaign pdays previous emp_var_rate cons_price_idx cons_conf_idx euribor3m nr_employed y
age 1.000000 -0.000866 0.004594 -0.034369 0.024365 -0.000371 0.000857 0.129372 0.010767 -0.017725 0.030399
duration -0.000866 1.000000 -0.071699 -0.047577 0.020640 -0.027968 0.005312 -0.008173 -0.032897 -0.044703 0.405274
campaign 0.004594 -0.071699 1.000000 0.052584 -0.079141 0.150754 0.127836 -0.013733 0.135133 0.144095 -0.066357
pdays -0.034369 -0.047577 0.052584 1.000000 -0.587514 0.271004 0.078889 -0.091342 0.296899 0.372605 -0.324914
previous 0.024365 0.020640 -0.079141 -0.587514 1.000000 -0.420489 -0.203130 -0.050936 -0.454494 -0.501333 0.230181
emp_var_rate -0.000371 -0.027968 0.150754 0.271004 -0.420489 1.000000 0.775334 0.196041 0.972245 0.906970 -0.298334
cons_price_idx 0.000857 0.005312 0.127836 0.078889 -0.203130 0.775334 1.000000 0.058986 0.688230 0.522034 -0.136211
cons_conf_idx 0.129372 -0.008173 -0.013733 -0.091342 -0.050936 0.196041 0.058986 1.000000 0.277686 0.100513 0.054878
euribor3m 0.010767 -0.032897 0.135133 0.296899 -0.454494 0.972245 0.688230 0.277686 1.000000 0.945154 -0.307771
nr_employed -0.017725 -0.044703 0.144095 0.372605 -0.501333 0.906970 0.522034 0.100513 0.945154 1.000000 -0.354678
y 0.030399 0.405274 -0.066357 -0.324914 0.230181 -0.298334 -0.136211 0.054878 -0.307771 -0.354678 1.000000
[13]:
sns.heatmap(df.corr())
Console
Data
Add data
Input
bankcsv
banking.csv
Output (44.1MB / 19.6GB)
/kaggle/working
Settings
Schedule a notebook run
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things


notebook2cae8fb620
Draft saved
File Edit View Run Add-ons Help
Share
Save Version 0
Run All
Code
Draft Session (19m)
Data
Add data
Input
bankcsv
banking.csv
Output (44.1MB / 19.6GB)
/kaggle/working
Settings
Schedule a notebook run
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
[13]:
sns.heatmap(df.corr())
[13]:
<AxesSubplot:>
age
duration
campaign
pdays
previous
emp_var_rate
cons_price_idx
cons_conf_idx
euribor3m
nr_employed
y
[15]:
sns.countplot(y='job', data= df)
Console
archive.zip
Show all
Type here to search
11:41 AM
21/04/2022


notebook2cae8fb620
Draft saved
File Edit View Run Add-ons Help
Share
Save Version 0
Run All
Code
Draft Session (20m)
Data
Add data
Input
bankcsv
banking.csv
Output (44.1MB / 19.6GB)
/kaggle/working
Settings
Schedule a notebook run
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
[15]:
sns.countplot(y='job', data= df)
[15]:
<AxesSubplot:xlabel='count', ylabel='job'>
blue-collar
technician
management
services
retired
admin.
housemaid
unemployed
entrepreneur
self-employed
unknown
student
job
0
2000
4000
6000
8000
10000
count
[16]:
sns.countplot(x='marital', data= df)
Console
archive.zip
Show all
Type here to search
11:41 AM
21/04/2022

kaggle.com/code/kshitijshinde/notebook2cae8fb620/edit

notebook2cae8fb620 Draft saved

File Edit View Run Add-ons Help

Share | Save Version 0

Run All | Code | Draft Session (20m)

Data | + Add data

Input

bankcsv

banking.csv

Output (44.1MB / 19.6GB)

/kaggle/working

Settings

Schedule a notebook run

Code Help

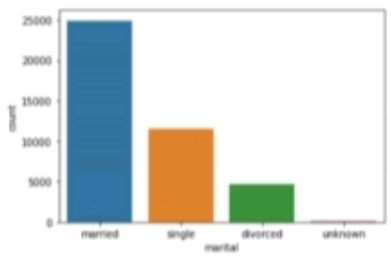
FIND CODE HELP

Find Code Help

Search for examples of how to do things

[16]:
```
sns.countplot(x='marital', data= df)
```

[16]:
```
<Axes:xlabel='marital', ylabel='count'>
```



[17]:
```
sns.countplot(x='y', data= df)
```

Console

archive.zip

Show all

Type here to search

11:41 AM
21/04/2022

kaggle.com/code/kshitijshinde/notebook2cae8fb620/edit

notebook2cae8fb620 Draft saved

File Edit View Run Add-ons Help

Share | Save Version 0

Run All | Code | Draft Session (20m)

Data | + Add data

Input

bankcsv

banking.csv

Output (44.1MB / 19.6GB)

/kaggle/working

Settings

Schedule a notebook run

Code Help

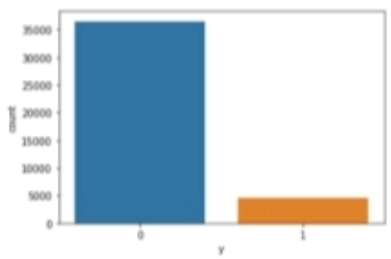
FIND CODE HELP

Find Code Help

Search for examples of how to do things

[17]:
```
sns.countplot(x='y', data= df)
```

[17]:
```
<Axes:xlabel='y', ylabel='count'>
```



[18]:
```
from sklearn import preprocessing
from sklearn.preprocessing import LabelEncoder
from sklearn import model_selection
from sklearn.linear_model import LogisticRegression
```

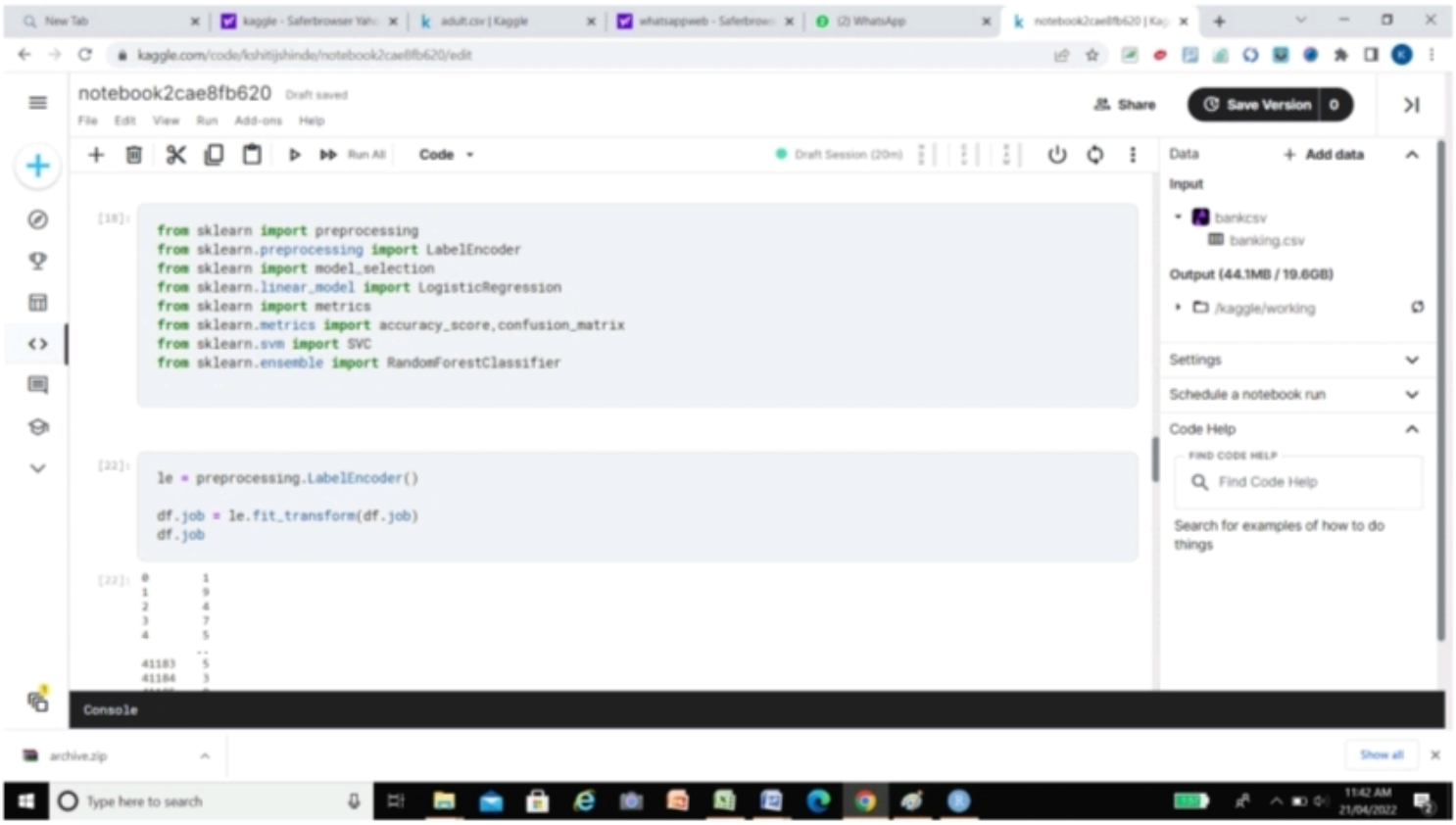
Console

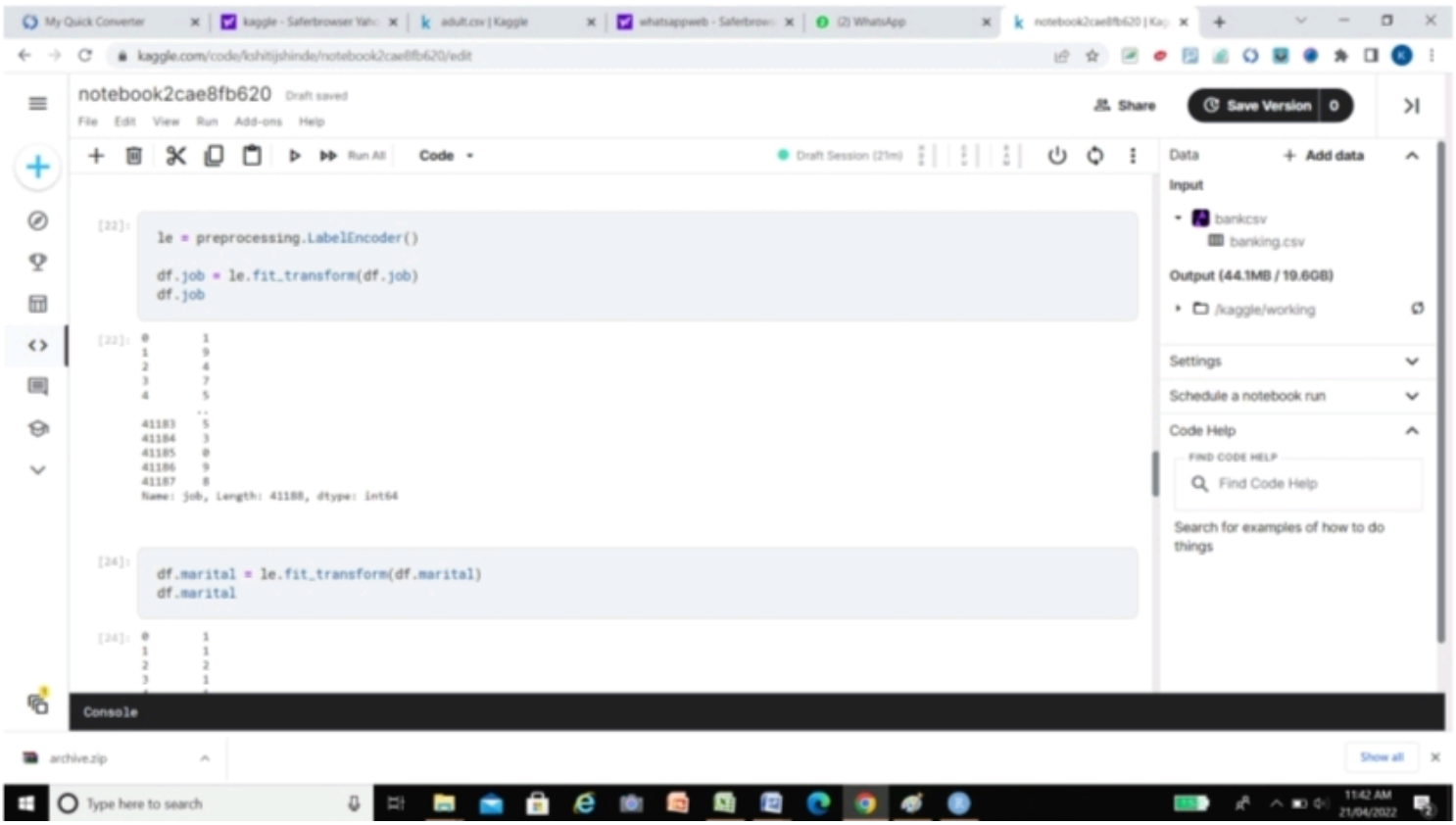
archive.zip

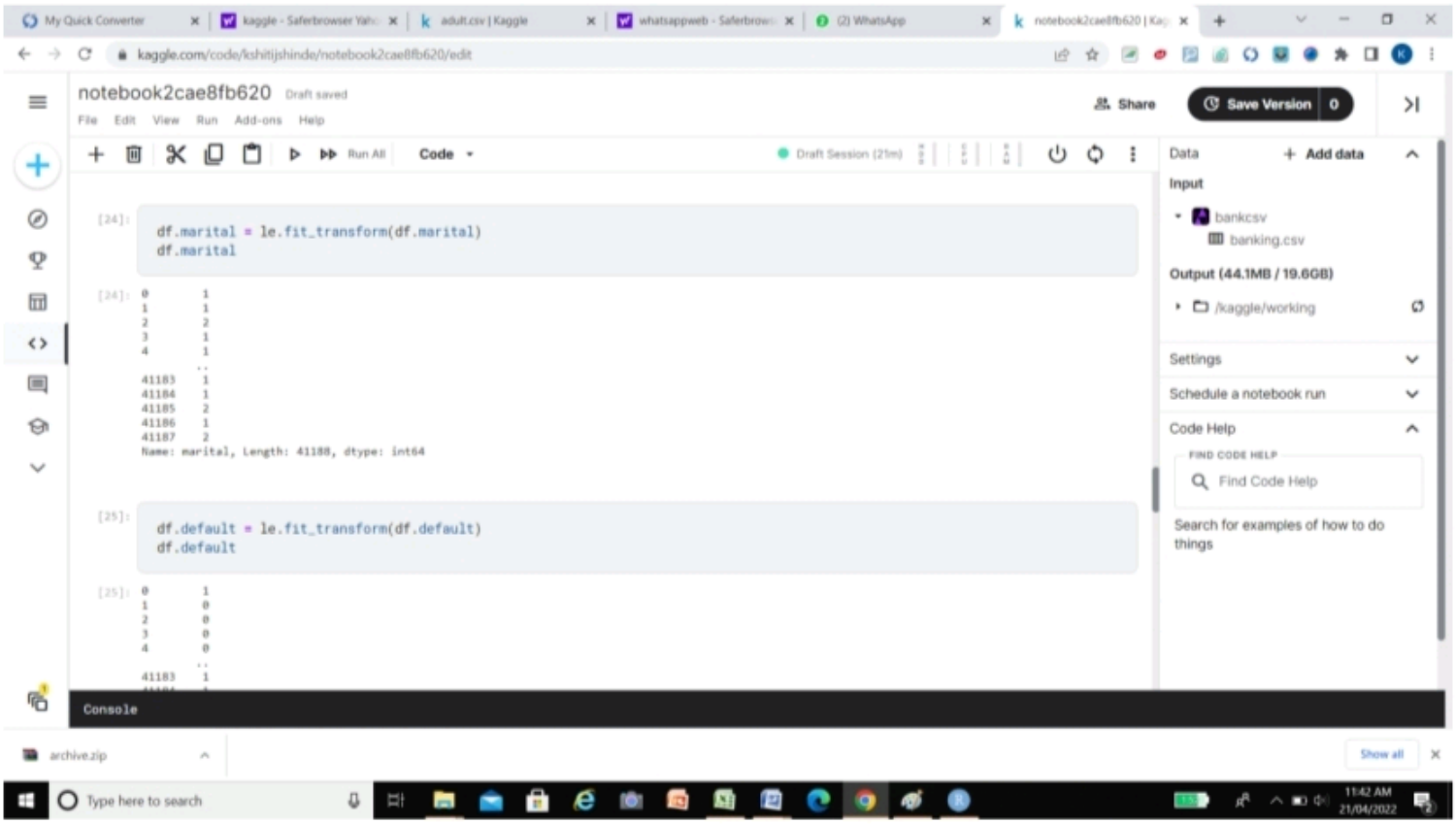
Show all

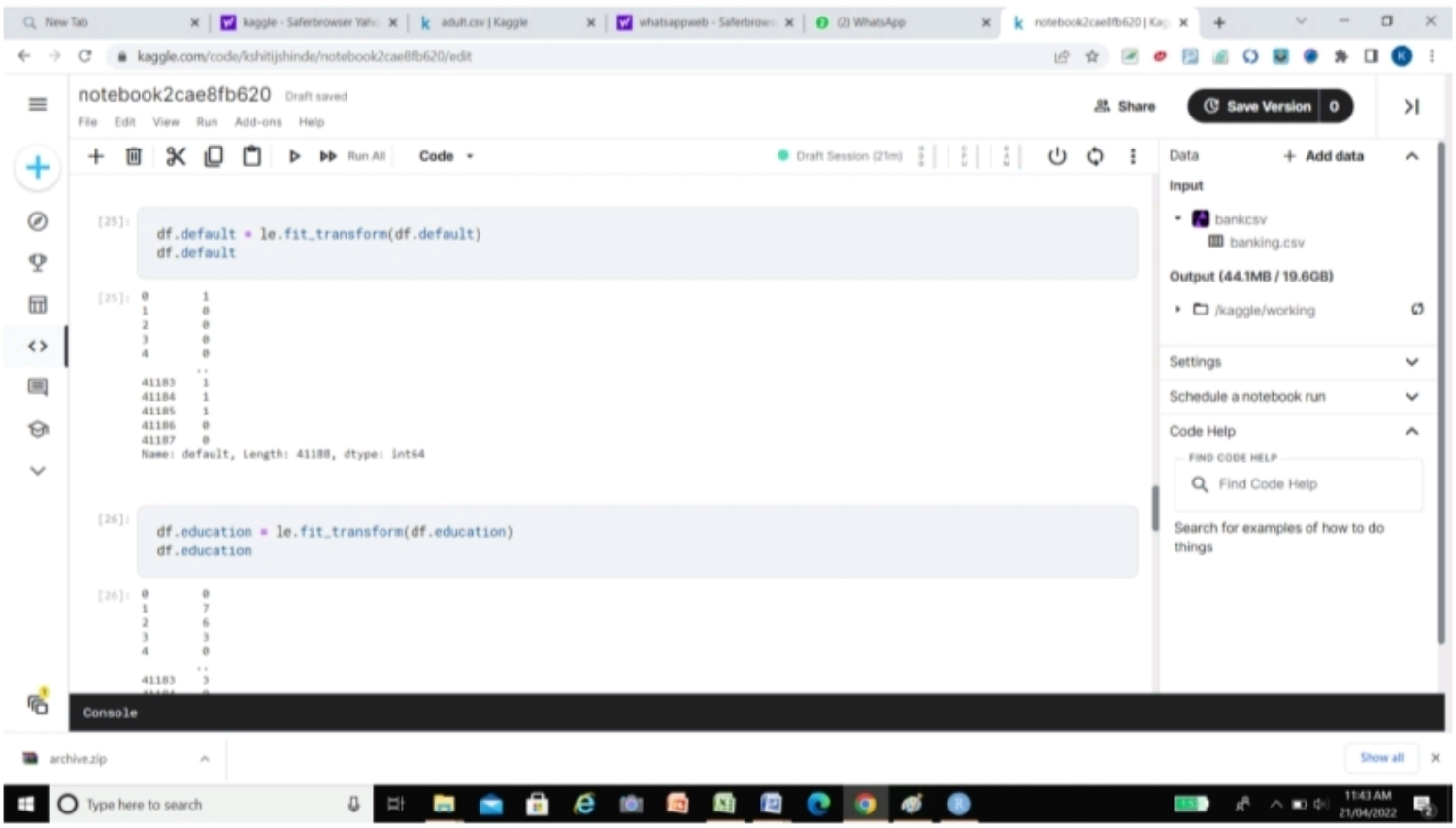
Type here to search

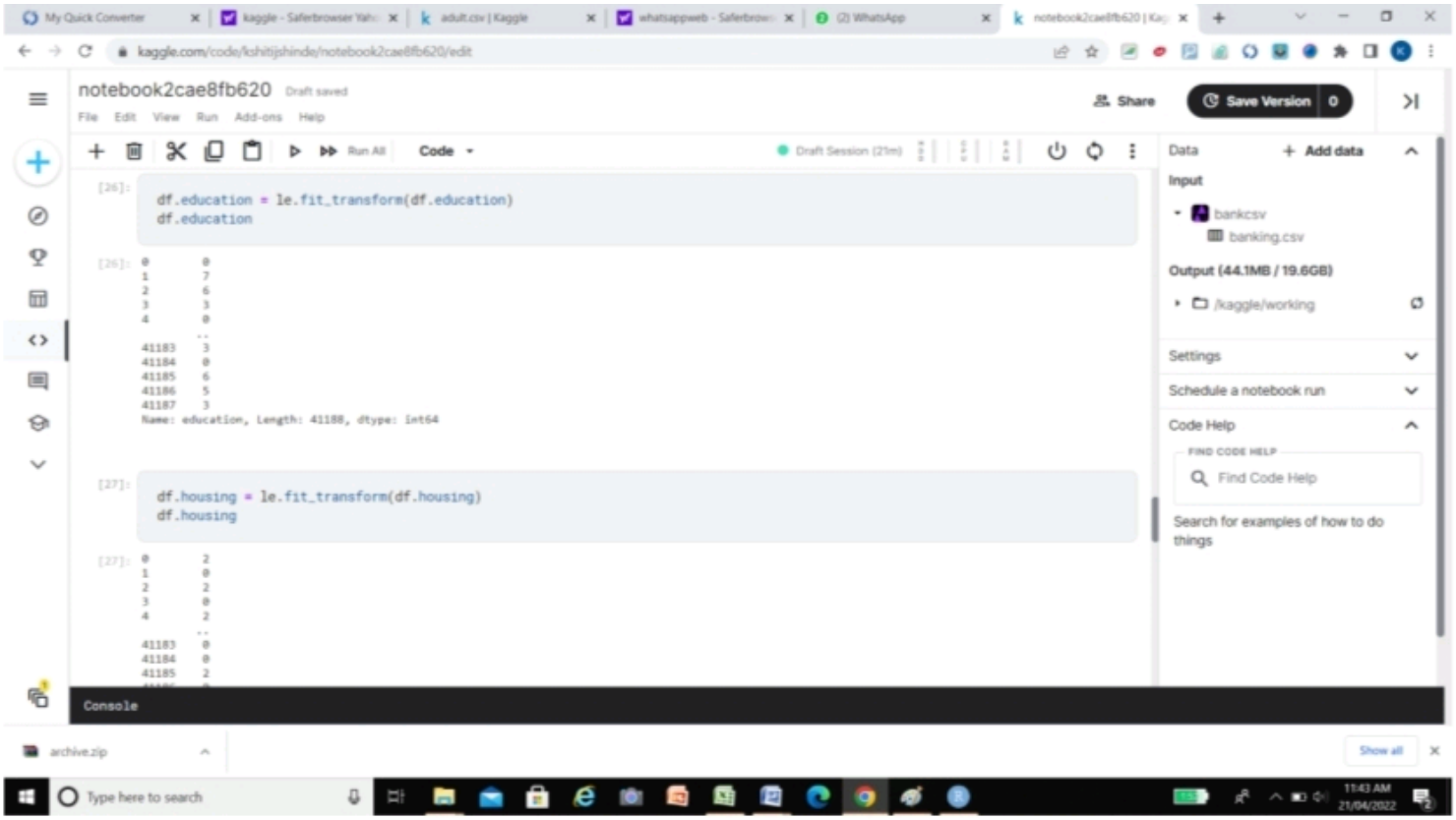
11:42 AM
21/04/2022

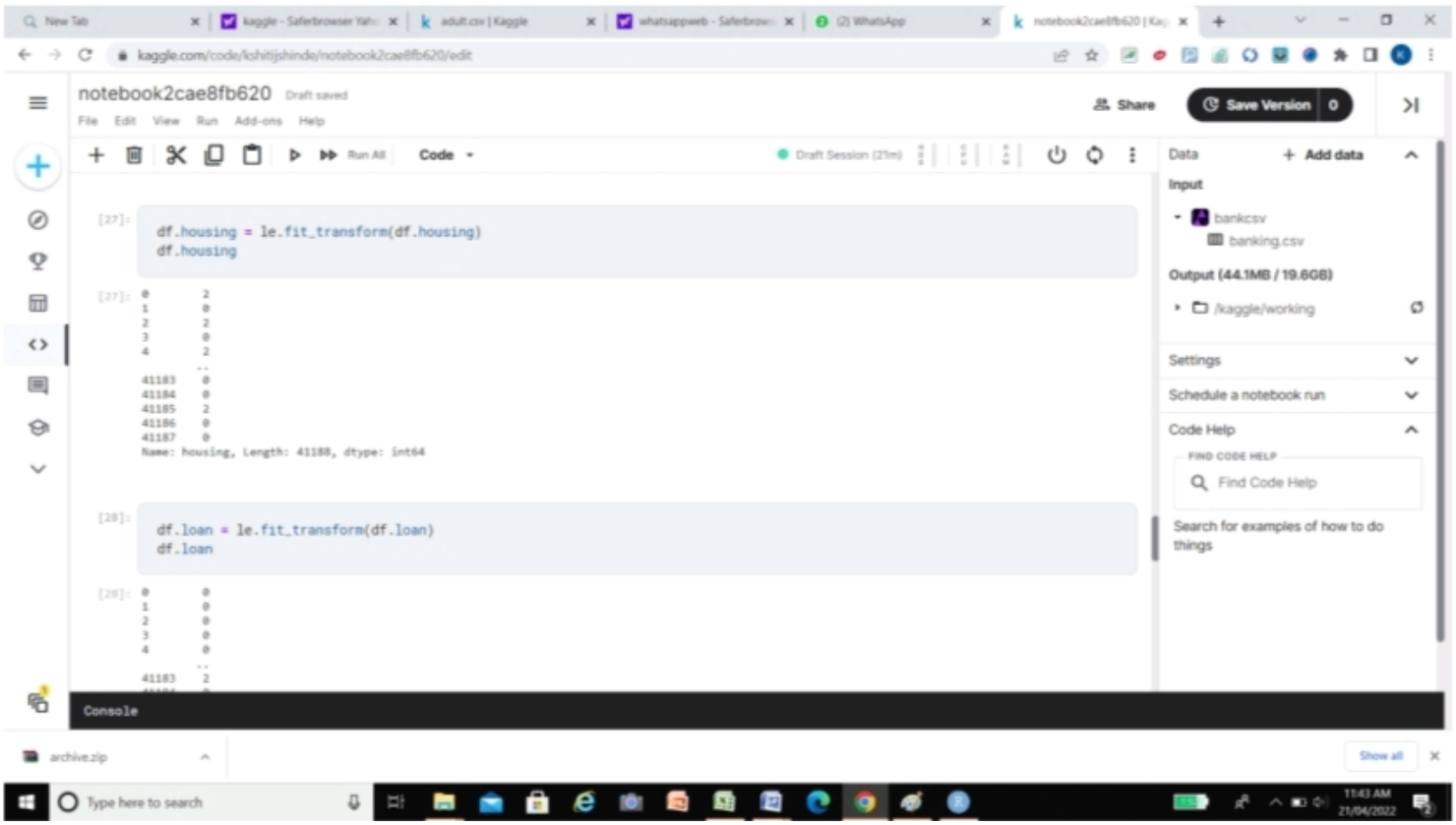

kaggle.com/code/kshitijshinde/notebook2cae8fb620/edit
notebook2cae8fb620
Draft saved
File Edit View Run Add-ons Help
Share
Save Version 0
Run All
Code
Draft Session (20m)
Data
Add data
Input
bankcsv
banking.csv
Output (44.1MB / 19.6GB)
/kaggle/working
Settings
Schedule a notebook run
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
[18]:
from sklearn import preprocessing
from sklearn.preprocessing import LabelEncoder
from sklearn import model_selection
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
from sklearn.metrics import accuracy_score,confusion_matrix
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier
[22]:
le = preprocessing.LabelEncoder()
df.job = le.fit_transform(df.job)
df.job
[22]:
0 1
1 9
2 4
3 7
4 5
..
41183 5
41184 3
Console
archive.zip
Show all
Type here to search
11:42 AM
21/04/2022

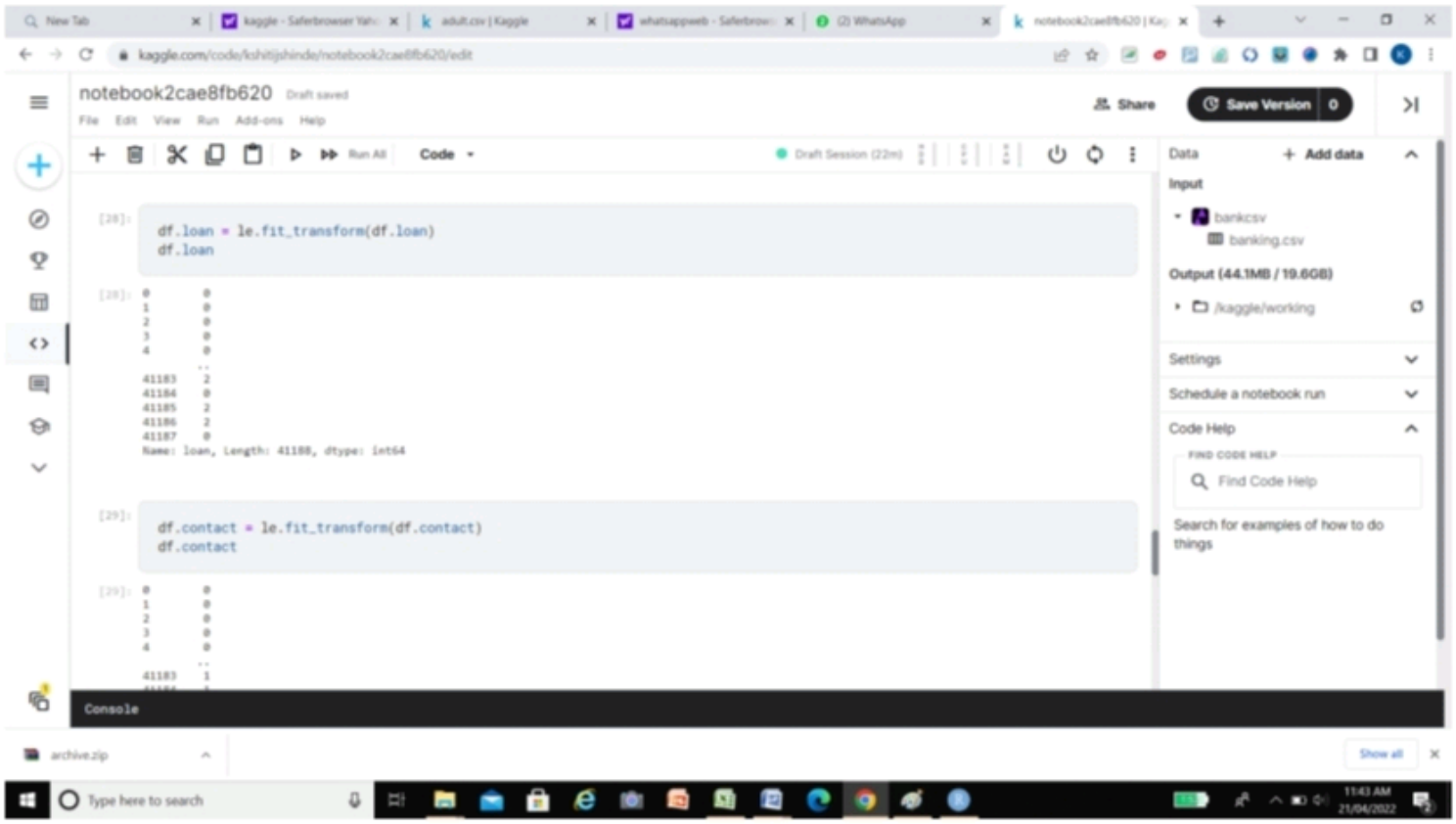

kaggle.com/code/kshitijshinde/notebook2cae8fb620/edit
notebook2cae8fb620
Draft saved
File Edit View Run Add-ons Help
Share
Save Version 0
Run All
Code
Draft Session (21m)
Data
Add data
Input
bankcsv
banking.csv
Output (44.1MB / 19.6GB)
/kaggle/working
Settings
Schedule a notebook run
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
[22]:
le = preprocessing.LabelEncoder()
df.job = le.fit_transform(df.job)
df.job
[22]:
0 1
1 9
2 4
3 7
4 5
..
41183 5
41184 3
41185 0
41186 9
41187 8
Name: job, Length: 41188, dtype: int64
[24]:
df.marital = le.fit_transform(df.marital)
df.marital
[24]:
0 1
1 1
2 2
3 1
Console
archive.zip
Show all
Type here to search
11:42 AM
21/04/2022

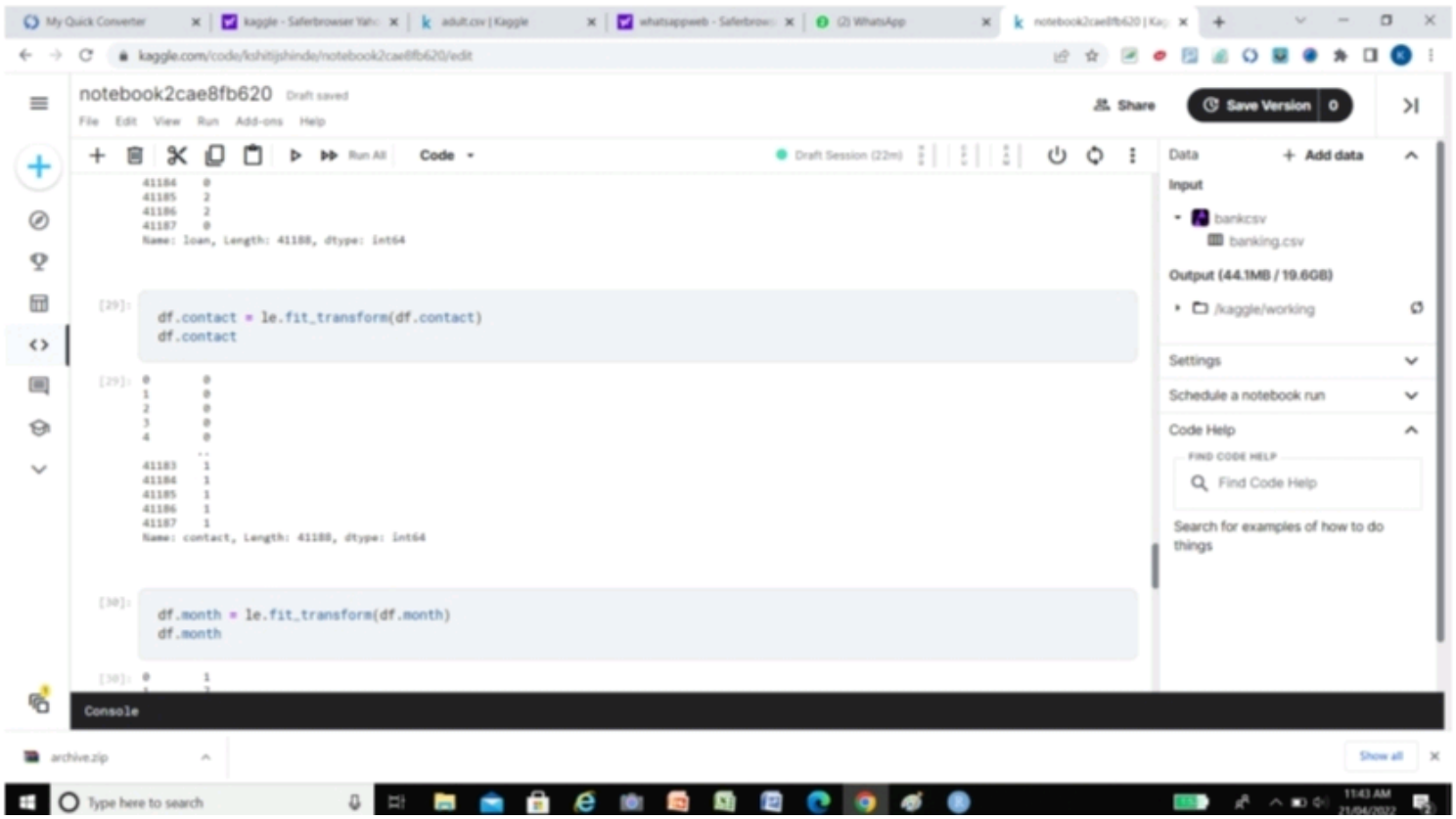

notebook2cae8fb620
Draft saved
File Edit View Run Add-ons Help
Share
Save Version 0
Run All
Code
Draft Session (21m)
[24]:
df.marital = le.fit_transform(df.marital)
df.marital
[24]:
0 1
1 1
2 2
3 1
4 1
..
41183 1
41184 1
41185 2
41186 1
41187 2
Name: marital, Length: 41188, dtype: int64
[25]:
df.default = le.fit_transform(df.default)
df.default
[25]:
0 1
1 0
2 0
3 0
4 0
..
41183 1
Console
Data
Add data
Input
bankcsv
banking.csv
Output (44.1MB / 19.6GB)
/kaggle/working
Settings
Schedule a notebook run
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
archive.zip
Show all
Type here to search
11:42 AM
21/04/2022

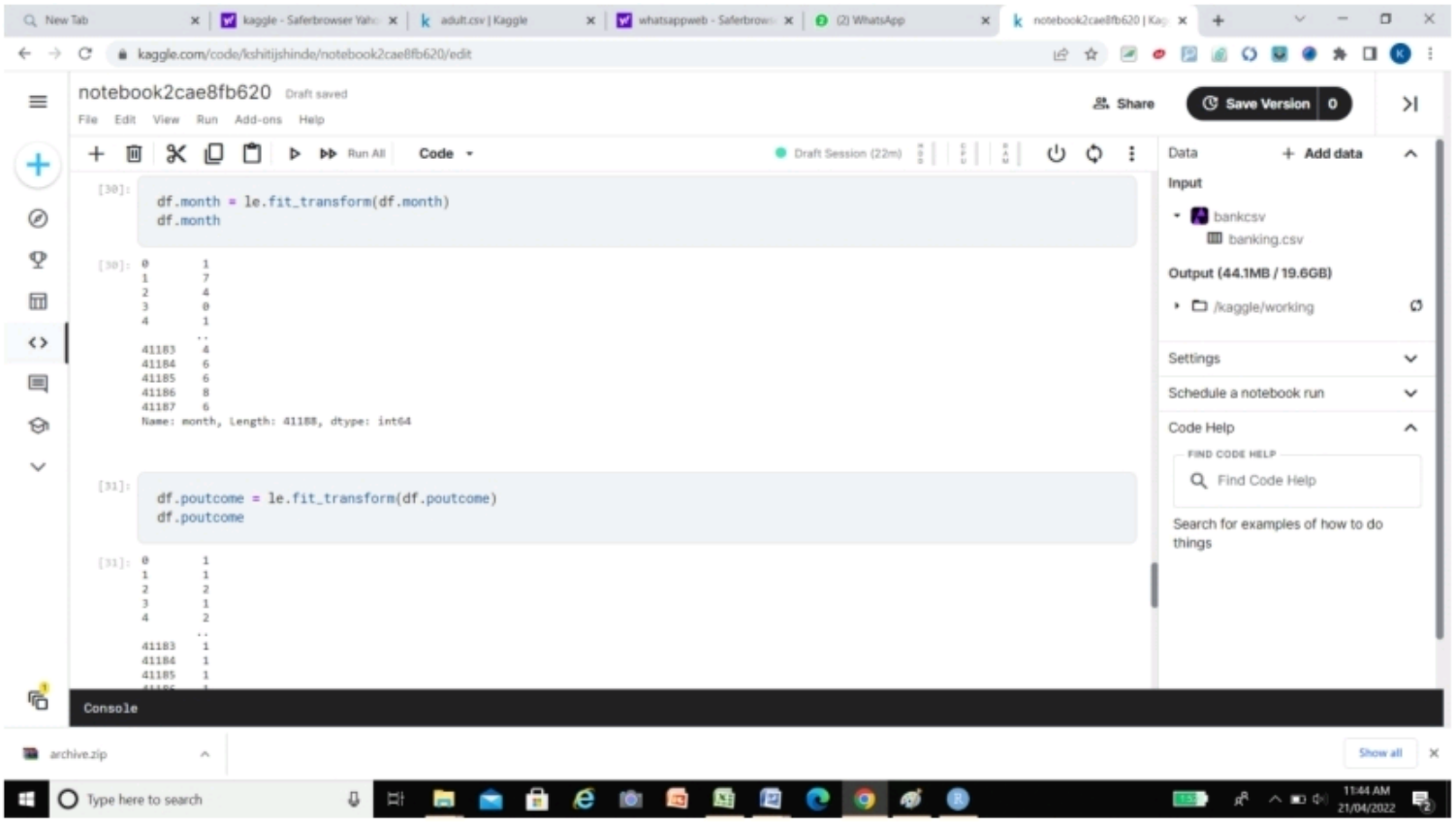

notebook2cae8fb620
Draft saved
File Edit View Run Add-ons Help
Share
Save Version 0
Run All
Code
Draft Session (21m)
[25]:
df.default = le.fit_transform(df.default)
df.default
[25]:
0 1
1 0
2 0
3 0
4 0
..
41183 1
41184 1
41185 1
41186 0
41187 0
Name: default, Length: 41188, dtype: int64
[26]:
df.education = le.fit_transform(df.education)
df.education
[26]:
0 0
1 7
2 6
3 3
4 0
..
41183 3
Console
Data
Add data
Input
bankcsv
banking.csv
Output (44.1MB / 19.6GB)
/kaggle/working
Settings
Schedule a notebook run
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
archive.zip
Show all
Type here to search
11:43 AM
21/04/2022

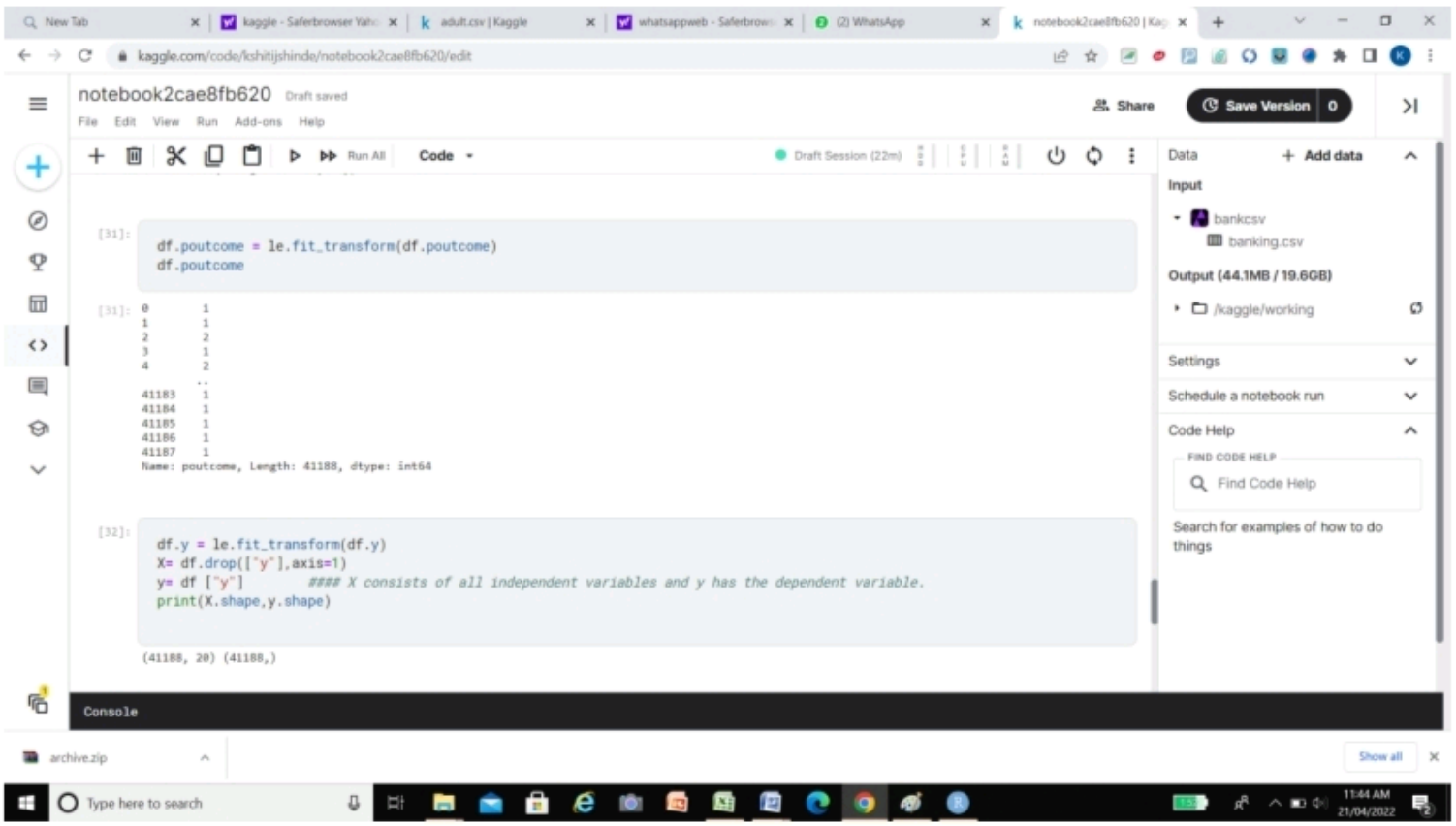

notebook2cae8fb620
Draft saved
File Edit View Run Add-ons Help
Share
Save Version 0
Run All
Code
Draft Session (21m)
[26]:
df.education = le.fit_transform(df.education)
df.education
[26]:
0 0
1 7
2 6
3 3
4 0
..
41183 3
41184 0
41185 6
41186 5
41187 3
Name: education, Length: 41188, dtype: int64
[27]:
df.housing = le.fit_transform(df.housing)
df.housing
[27]:
0 2
1 0
2 2
3 0
4 2
..
41183 0
41184 0
41185 2
Console
Data
Add data
Input
bankcsv
banking.csv
Output (44.1MB / 19.6GB)
/kaggle/working
Settings
Schedule a notebook run
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
archive.zip
Show all
Type here to search
11:43 AM
21/04/2022

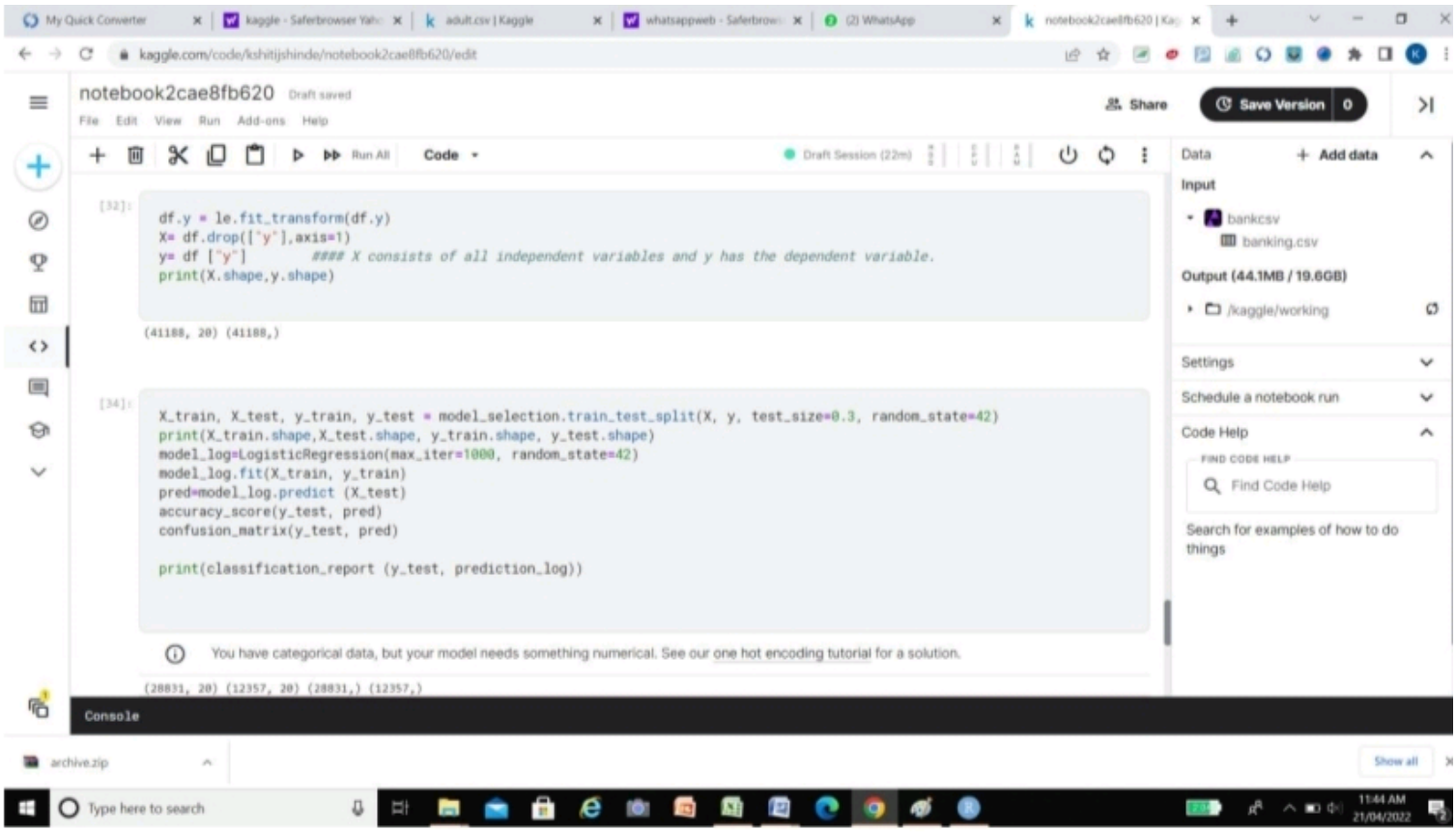

notebook2cae8fb620
Draft saved
File Edit View Run Add-ons Help
Share
Save Version 0
Run All
Code
Draft Session (21m)
[27]:
df.housing = le.fit_transform(df.housing)
df.housing
[27]:
0 2
1 0
2 2
3 0
4 2
..
41183 0
41184 0
41185 2
41186 0
41187 0
Name: housing, Length: 41188, dtype: int64
[28]:
df.loan = le.fit_transform(df.loan)
df.loan
[28]:
0 0
1 0
2 0
3 0
4 0
..
41183 2
Console
Data
Add data
Input
bankcsv
banking.csv
Output (44.1MB / 19.6GB)
/kaggle/working
Settings
Schedule a notebook run
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
archive.zip
Show all
Type here to search
11:43 AM
21/04/2022


notebook2cae8fb620
Draft saved
File Edit View Run Add-ons Help
Share
Save Version 0
Run All
Code
Draft Session (22m)
[28]:
df.loan = le.fit_transform(df.loan)
df.loan
[28]:
0 0
1 0
2 0
3 0
4 0
..
41183 2
41184 0
41185 2
41186 2
41187 0
Name: loan, Length: 41188, dtype: int64
[29]:
df.contact = le.fit_transform(df.contact)
df.contact
[29]:
0 0
1 0
2 0
3 0
4 0
..
41183 1
Console
Data
Add data
Input
bankcsv
banking.csv
Output (44.1MB / 19.6GB)
/kaggle/working
Settings
Schedule a notebook run
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
archive.zip
Show all
Type here to search
11:43 AM
21/04/2022


notebook2cae8fb620
Draft saved
File Edit View Run Add-ons Help
Share
Save Version 0
Run All
Code
Draft Session (22m)
41184 0
41185 2
41186 2
41187 0
Name: loan, Length: 41188, dtype: int64
[29]:
df.contact = le.fit_transform(df.contact)
df.contact
[29]:
0 0
1 0
2 0
3 0
4 0
..
41183 1
41184 1
41185 1
41186 1
41187 1
Name: contact, Length: 41188, dtype: int64
[30]:
df.month = le.fit_transform(df.month)
df.month
[30]: 0 1
Console
Data
Add data
Input
bankcsv
banking.csv
Output (44.1MB / 19.6GB)
/kaggle/working
Settings
Schedule a notebook run
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
archive.zip
Show all
Type here to search
11:43 AM
21/04/2022


notebook2cae8fb620
Draft saved
File Edit View Run Add-ons Help
Share
Save Version 0
Run All
Code
Draft Session (22m)
Data
Add data
Input
bankcsv
banking.csv
Output (44.1MB / 19.6GB)
/kaggle/working
Settings
Schedule a notebook run
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
[30]:
df.month = le.fit_transform(df.month)
df.month
[30]:
0 1
1 7
2 4
3 0
4 1
..
41183 4
41184 6
41185 6
41186 8
41187 6
Name: month, Length: 41188, dtype: int64
[31]:
df.poutcome = le.fit_transform(df.poutcome)
df.poutcome
[31]:
0 1
1 1
2 2
3 1
4 2
..
41183 1
41184 1
41185 1
Console


notebook2cae8fb620
Draft saved
File Edit View Run Add-ons Help
Share
Save Version 0
Run All
Code
Draft Session (22m)
Data
Add data
Input
bankcsv
banking.csv
Output (44.1MB / 19.6GB)
/kaggle/working
Settings
Schedule a notebook run
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
[31]:
df.poutcome = le.fit_transform(df.poutcome)
df.poutcome
[31]:
0 1
1 1
2 2
3 1
4 2
..
41183 1
41184 1
41185 1
41186 1
41187 1
Name: poutcome, Length: 41188, dtype: int64
[32]:
df.y = le.fit_transform(df.y)
X= df.drop(["y"],axis=1)
y= df ["y"] #### X consists of all independent variables and y has the dependent variable.
print(X.shape,y.shape)
(41188, 20) (41188,)
Console



[38]:

```
X_train, X_test, y_train, y_test = model_selection.train_test_split(X, y, test_size=0.3, random_state=42)
print(X_train.shape,X_test.shape, y_train.shape, y_test.shape)
model_log=LogisticRegression(max_iter=1000, random_state=42)
model_log.fit(X_train, y_train)
pred=model_log.predict(X_test)
accuracy_score(y_test, pred)
confusion_matrix(y_test, pred)


print(classification_report (y_test, prediction_log))
```

You have categorical data, but your model needs something numerical. See our one hot encoding tutorial for a solution.

```
(28831, 20) (12357, 20) (28831,) (12357,)
```

Assignment No. 7

Aim: Test Analytics:

1) Extract Sample document and apply following document preprocessing methods: Tokenization, POS Tagging, stop words removal, Stemming and Lemmatization.
2) Create representation of document by calculating Term Frequency and Inverse Document Frequency.

Sol^{n} :

Step 1: Analysing Dataset

The first step in any of the Machine Learning tasks is to analyse the data. So if we look at the dataset, at first glance, we see all the documents with words in English.

Step 2: Extracting Title & Body:

There is no specific way to do this, this totally depends on the problem statement at hand and on the analysis, we do on the dataset.

Filename	Size	Description of the Textfile
sre01.txt	11278	SRE: The Saga Of The Best SRE Game Ever Played! By Josh Renaud
sre02.txt	5862	Solar Realms Elite: The True Story of the Unsung Heroes, by Josh Renaud
sre03.txt	8555	Solar Realms Elite: Ultra's Untold Story by Josh Renaud
sre04.txt	44198	Solar Realms Elite IV: The Confrontation, by Josh Renaud
sre05.txt	20787	Solar Realms Elite V: The Underground, by Josh Renaud
sre06.txt	26731	Solar Realms Elite VI: The Alliance Restored, by Josh Renaud
sre07.txt	23597	Solar Realms Elite 7: Petros, by Josh Renaud
sre08.txt	33170	Solar Realms Elite VIII: Kazik, by Josh Renaud
sre09.txt	26073	Solar Realms Elite IX: Survival of the Fittest, by Josh Renaud
sre10.txt	25725	Solar Realms Elite X: Legacies, by Josh Renaud
sre_feqh.txt	20054	Solar Realms Elite: The Feqh Galaxy, by Josh Renaud
sre_finl.txt	33158	Solar Realms Elite: The Finale by Josh Renaud
sre_sei.txt	20753	Solar Realms Elite: Galaxy Sei, by Josh Renaud
sretrade.txt	9008	The SRE Commerce and Trade Theories, by Josh Renaud
srex.txt	37128	Solar Realms Elite: X1 and X2, by Josh Renaud

There are 15 files for

Now we can find that **folders** give extra / for the root folder, so we are going to remove it.

```
folders[0] = folders[0][:len(folders[0])-1]
```

```
<TR VALIGN=TOP><TD ALIGN=TOP><A HREF="13chil.txt">13chil.txt</A>  <tab t
o=T><TD> 8457<BR><TD> The Story of the Sly Fox
<TR VALIGN=TOP><TD ALIGN=TOP><A HREF="14.lws">14.lws</A>  <tab to=T><TD>
5261<BR><TD> A Smart Bomb with a Language Parser
<TR VALIGN=TOP><TD ALIGN=TOP><A HREF="16.lws">16.lws</A>  <tab to=T><TD>
15294<BR><TD> Two Guys in a Garage, by M. Pshota
<TR VALIGN=TOP><TD ALIGN=TOP><A HREF="17.lws">17.lws</A>  <tab to=T><TD>
10853<BR><TD> The Early Days of a High-Tech Start-up are Magic (November
18, 1991) by M. Peshota
<TR VALIGN=TOP><TD ALIGN=TOP><A HREF="18.lws">18.lws</A>  <tab to=T><TD>
26624<BR><TD> The Couch, the File Cabinet, and the Calendar, by M. Pesho
ta (December 9, 1991)
<TR VALIGN=TOP><TD ALIGN=TOP><A HREF="19.lws">19.lws</A>  <tab to=T><TD>
17902<BR><TD> Engineering the Future of American Technology by M. Peshot
a (January 5, 1992)
<TR VALIGN=TOP><TD ALIGN=TOP><A HREF="20.lws">20.lws</A>  <tab to=T><TD>
13588<BR><TD> What Research and Development Was Always Meant to Be, by
M. Peshota
```

Names and titles variables have the list of all names and titles.

```
names = re.findall('><A HREF="(.*)">', text)
titles = re.findall('<BR><TD> (.*)\n', text)
dataset = []for i in folders:
    file = open(i+"/index.html", 'r')
    text = file.read().strip()
    file.close()    file_name = re.findall('><A HREF="(.*)">', text)
    file_title = re.findall('<BR><TD> (.*)\n', text)

    for j in range(len(file_name)):
        dataset.append((str(i) + str(file_name[j]), file_title[j]))
```

```
1  dataset
```

```
[('/Users/williamscott/Desktop/IR/Assignments/Assignment 2/stories/100west.txt',
  'Going 100 West by 53 North by Jim Prentice (1990)'),
 ('/Users/williamscott/Desktop/IR/Assignments/Assignment 2/stories/13chil.txt',
  'The Story of the Sly Fox'),
 ('/Users/williamscott/Desktop/IR/Assignments/Assignment 2/stories/14.lws',
  'A Smart Bomb with a Language Parser'),
 ('/Users/williamscott/Desktop/IR/Assignments/Assignment 2/stories/16.lws',
  'Two Guys in a Garage, by M. Pshota'),
 ('/Users/williamscott/Desktop/IR/Assignments/Assignment 2/stories/17.lws',
  'The Early Days of a High-Tech Start-up are Magic (November 18, 1991) by M. Peshota'),
 ('/Users/williamscott/Desktop/IR/Assignments/Assignment 2/stories/18.lws',
  'The Couch, the File Cabinet, and the Calendar, by M. Peshota (December 9, 1991)'),
 ('/Users/williamscott/Desktop/IR/Assignments/Assignment 2/stories/19.lws',
  'Engineering the Future of American Technology by M. Peshota (January 5, 1992)'),
 ('/Users/williamscott/Desktop/IR/Assignments/Assignment 2/stories/20.lws',
  'What Research and Development Was Always Meant to Be, by M. Peshota'),
 ('/Users/williamscott/Desktop/IR/Assignments/Assignment 2/stories/3gables.txt',
  'The Adventure of the Three Gables'),
```

simply use a conditional checker to remove it.

```
if c == False:
    file_name = file_name[2:]
    c = True
```

Step 3: Preprocessing

Preprocessing is one of the major steps when we are dealing with any kind of text model.Few mandatory preprocessing are: converting to lowercase, removing punctuation, removing stop words and lemmatization/stemming. In our problem statement, it seems like the basic preprocessing steps will be sufficient.

Lowercase: Numpy has a method that can convert the list of lists to lowercase at once.

```
np.char.lower(data)
```

Stop words

Stop words are the most commonly occurring words that don't give any additional value to the document vector.

```
1 print(stopwords.words('english'))
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you',
'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'sh
t's", 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves
t', "that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'b
ng', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'i
f', 'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into',
e', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'o
e', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'bo
me', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than',
'don', "don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o', 'r
n', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't
"isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "ne
n't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn',
```

Iterate over all the stop words and not append them to the list if it's a stop word.

```
new_text = ""
for word in words:
    if word not in stop_words:
        new_text = new_text + " " + word
```

Punctuation

Punctuation is the set of unnecessary symbols that are in our corpus documents.

```
symbols = "!\"#$%&()*+-./:;<=>?@[\]^_`{|}~\n"
for i in symbols:
    data = np.char.replace(data, i, ' ')
```

Apostrophe

Note that there is no ‘ apostrophe in the punctuation symbols

```
return np.char.replace(data, "'", "")
```

Single Characters

Single characters are not much useful in knowing the importance of the document and few final single characters might be irrelevant symbols, so it is always good to remove the single characters.

```
new_text = ""
for w in words:
    if len(w) > 1:
        new_text = new_text + " " + w
```

Stemming

This is the final and most important part of the preprocessing. stemming converts words to their stem.

```
stemmer = PorterStemmer()
stemmer.stem("swimming")
```

```
'swim'
```

Lemmatisation

Lemmatisation is a way to reduce the word to the root synonym of a word.

Stemming vs Lemmatization

stemming — need not be a dictionary word, removes prefix and affix based on few rules

lemmatization — will be a dictionary word. reduces to a root synonym.

Word	***Lemmatization***	***Stemming***
was	be	wa
studies	study	studi
studying	study	study

Converting Numbers

achieve this we are going to use a library called **num2word**.

```
num2words(100500)
```

```
'one hundred thousand, five hundred'
```

Preprocessing

Finally, we are going to put in all those preprocessing methods above in another method and we will call that preprocess method.

```
def preprocess(data):
    data = convert_lower_case(data)
    data = remove_punctuation(data)
    data = remove_apostrophe(data)
    data = remove_single_characters(data)
    data = convert_numbers(data)
    data = remove_stop_words(data)
    data = stemming(data)
    data = remove_punctuation(data)
    data = convert_numbers(data)
```

Step 4: Calculating TF-IDF

Calculating DF

```
DF = {}
for i in range(len(processed_text)):
    tokens = processed_text[i]
    for w in tokens:
        try:
            DF[w].add(i)
        except:
            DF[w] = {i}
```

```
for i in DF:
    DF[i] = len(DF[i])
DF
```

```
{'sharewar': 1,
 'trial': 1,
 'project': 4,
 'freewar': 1,
 'need': 6,
 'support': 2,
 'continu': 4,
 'one': 10,
 'hundr': 8,
```

To find the total unique words in our vocabulary, we need to take all the keys of DF.

```
total_vocab = [x for x in DF]
print(total_vocab)
```

```
['sharewar', 'trial', 'project', 'freewar', 'ne
'north', 'jim', 'prentic', 'copyright', 'thousa
c', 'phrase', 'spoken', 'mumbl', 'thought', 'in
'map', 'label', 'degr', 'presenc', 'indic', 'h:
ct', 'mind', 'intern', 'border', 'writer', 'poe
le', 'man', 'eat', 'mosquito', 'murder', 'hord
'dog', 'stori', 'record', 'break', 'trout', 'wa
'forest', 'crash', 'moo', 'tear', 'brush', 'tre
uck', 'feed', 'quiet', 'pond', 'placid', 'bay',
r', 'authent', 'northern', 'scene', 'wildlif',
d', 'person', 'live', 'farther', 'becom', 'appa
```

Calculating TF-IDF

```
tf_idf = {}
for i in range(N):
    tokens = processed_text[i]
    counter = Counter(tokens + processed_title[i])
    for token in np.unique(tokens):
        tf = counter[token]/words_count
        df = doc_freq(token)
        idf = np.log(N/(df+1))
        tf_idf[doc, token] = tf*idf
```

Step 4: Ranking using Matching Score

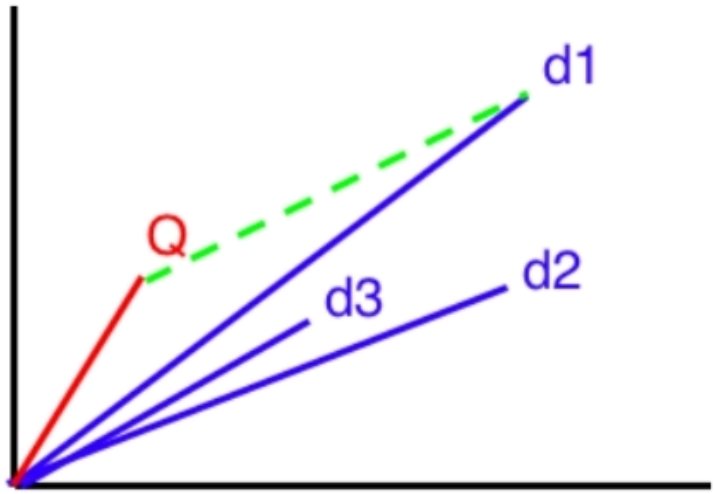
Matching score is the simplest way to calculate the similarity, in this method, we **add tf_idf values of the tokens that are in query for every document.**

```
def matching_score(query):
    query_weights = {}
    for key in tf_idf:
```

```
    if key[1] in tokens:
        query_weights[key[0]] += tf_idf[key]
```

Step 5: Ranking using Cosine Similarity

It will mark all the documents as vectors of tf-idf tokens and measures the similarity in cosine space (the angle between the vectors).



Vectorization

To compute any of the above, the simplest way is to convert everything to a vector and then compute the cosine similarity.

```
# Document Vectorization
D = np.zeros((N, total_vocab_size))
for i in tf_idf:
    ind = total_vocab.index(i[1])
    D[i[0]][ind] = tf_idf[i]
Q = np.zeros((len(total_vocab)))
counter = Counter(tokens)
words_count = len(tokens)
query_weights = {}
for token in np.unique(tokens):
    tf = counter[token]/words_count
```

```
df = doc_freq(token)
idf = math.log((N+1)/(df+1))
```

Analysis

Short Query

```
Matching Score

Query: Without the drive of Rebeccah's insistence, Kate lost her momentum. She stood next a slatted oak bench, canist
ers still clutched, surveying

['without', 'drive', 'rebeccah', 'insist', 'kate', 'lost', 'momentum', 'stood', 'next', 'slat', 'oak', 'bench', 'cani
st', 'still', 'clutch', 'survey']

[166, 200, 352, 433, 211, 350, 175, 187, 188, 294]
```

```
Cosine Similarity

Query: Without the drive of Rebeccah's insistence, Kate lost her momentum. She stood next a slatted oak bench, canist
ers still clutched, surveying

['without', 'drive', 'rebeccah', 'insist', 'kate', 'lost', 'momentum', 'stood', 'next', 'slat', 'oak', 'bench', 'cani
st', 'still', 'clutch', 'survey']

[200 166 433 175 169 402 211  87 151 369]
```

Long Query

```
Matching Score

Query: And then she recalled his stiff body stretched out in the little bed over the garages. Another pearl had come
loose from the strand, seeming to want to search out its old home in a far away oyster bed. She would have those pear
ls laid out n

['recal', 'stiff', 'bodi', 'stretch', 'littl', 'bed', 'garag', 'anoth', 'pearl', 'come', 'loo', 'strand', 'seem', 'wa
nt', 'search', 'old', 'home', 'far', 'away', 'oyster', 'bed', 'would', 'pearl', 'laid']

[352, 351, 43, 269, 17, 272, 9, 67, 416, 208]
```

```
Cosine Similarity

Query: And then she recalled his stiff body stretched out in the little bed over the garages. Another pearl had come
loose from the strand, seeming to want to search out its old home in a far away oyster bed. She would have those pear
ls laid out n

['recal', 'stiff', 'bodi', 'stretch', 'littl', 'bed', 'garag', 'anoth', 'pearl', 'come', 'loo', 'strand', 'seem', 'wa
nt', 'search', 'old', 'home', 'far', 'away', 'oyster', 'bed', 'would', 'pearl', 'laid']

[ 16   3 200 151  65 364 169  87  27  82]
```

ASSIGNMENT NO. 8

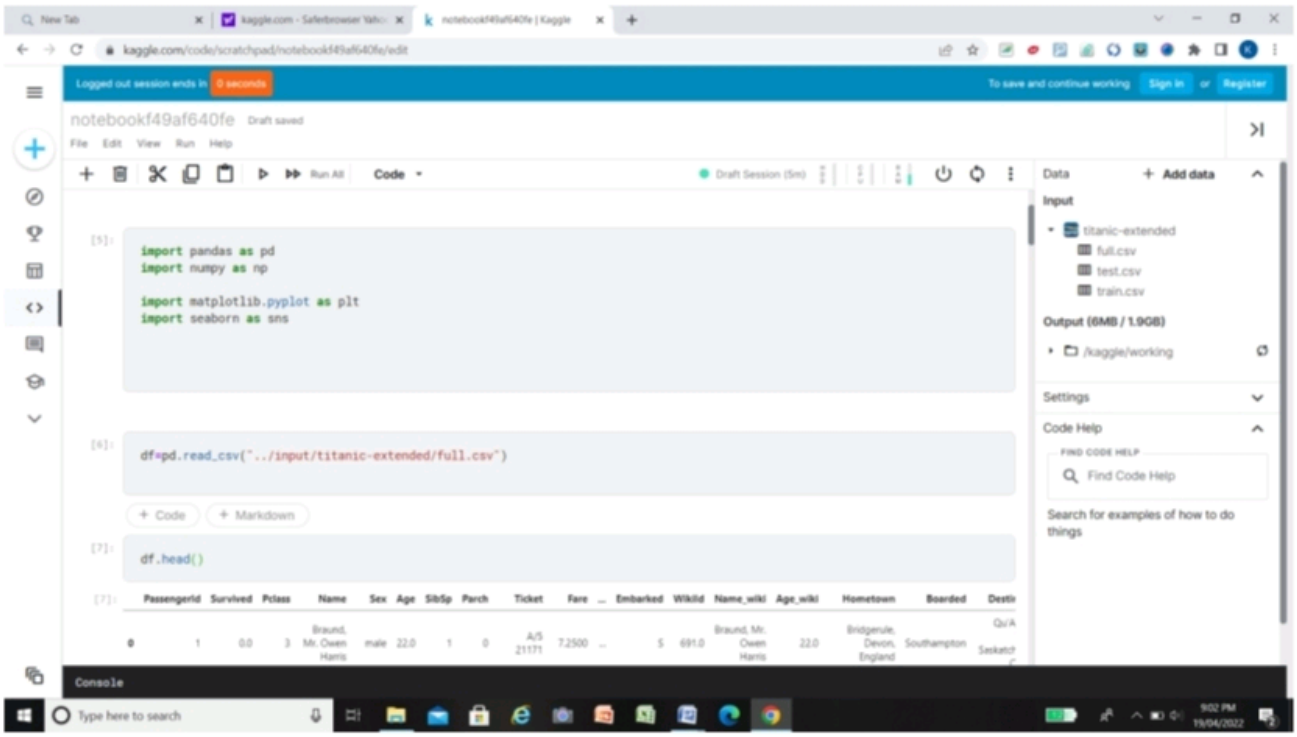
Aim: Data Visualization I :

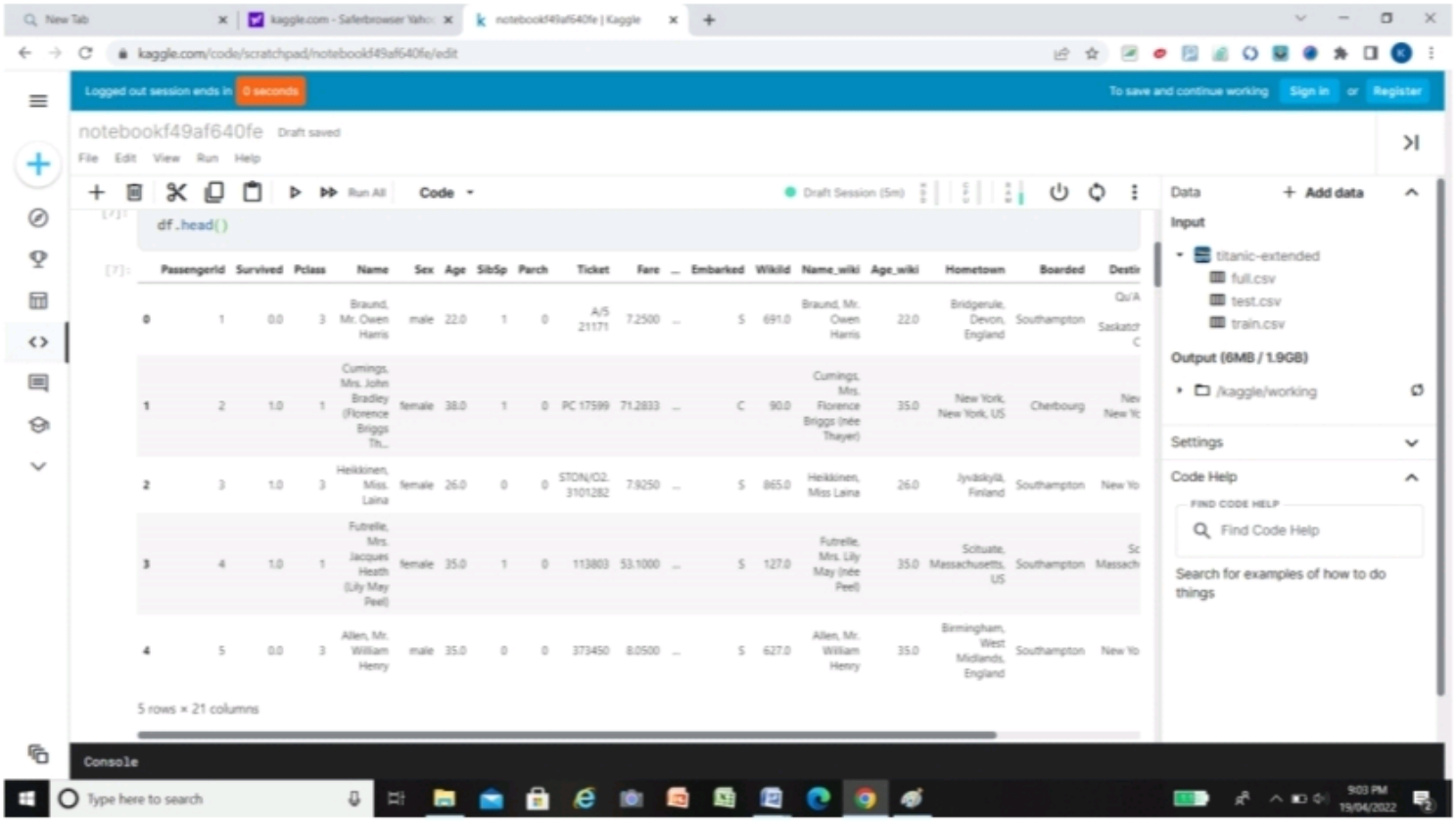
i) Use the inbuilt dataset 'titanic'. The dataset contains 891 rows and contains information about the passengers who boarded the unfortunate Titanic ship. Use seaborn library to see if we can find any patterns in the data.

ii) Write a code to check how the price of the ticket (column name: 'fare') by each passenger is distributed by plotting a histogram.

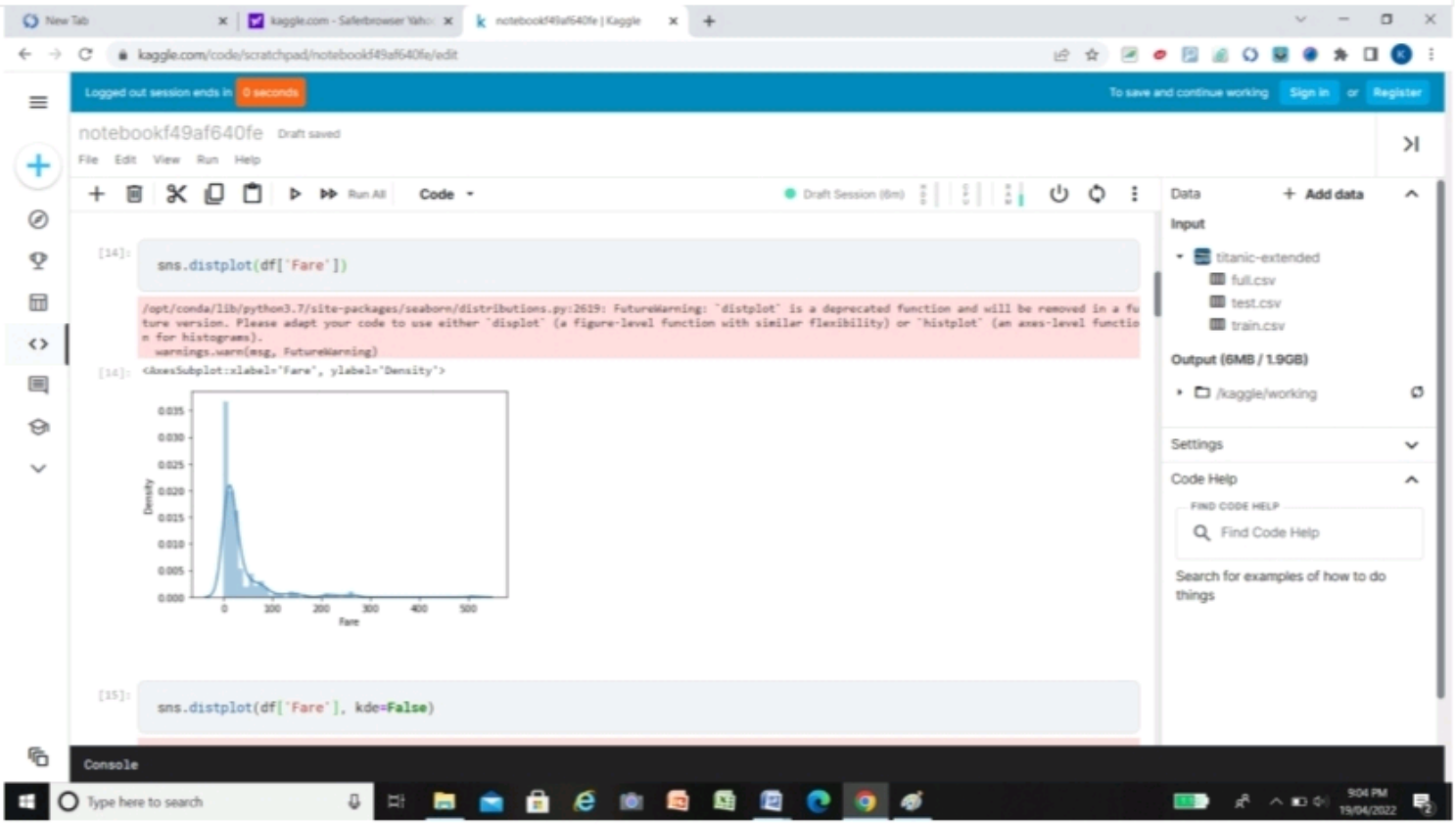
Sol^{n} :

OUTPUT :





The Dist Plot:



My Quick Converter | kaggle.com - Saferbrowser Yahoo | notebookf49af640fe | Kaggle

kaggle.com/code/scratchpad/notebookf49af640fe/edit

Logged out session ends in 0 seconds. To save and continue working Sign in or Register

notebookf49af640fe Draft saved

File Edit View Run Help

Run All | Code | Draft Session (7m)

Data | + Add data

Input

titanic-extended
- full.csv
- test.csv
- train.csv

Output (6MB / 1.9GB)

/kaggle/working

Settings

Code Help

FIND CODE HELP

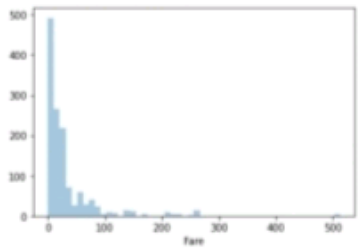
Find Code Help

Search for examples of how to do things

[15]:
```
sns.distplot(df['Fare'], kde=False)
```

```
/opt/conda/lib/python3.7/site-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated function and will be removed in a fu
ture version. Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level functio
n for histograms).
  warnings.warn(msg, FutureWarning)
```

[15]: <AxesSubplot:xlabel='Fare'>



[16]:
```
sns.distplot(df['Fare'], kde=False, bins=10)
```

Console

Type here to search

9:04 PM 19/04/2022

My Quick Converter | kaggle.com - Saferbrowser Yahoo | notebookf49af640fe | Kaggle

kaggle.com/code/scratchpad/notebookf49af640fe/edit

Logged out session ends in 0 seconds. To save and continue working Sign in or Register

notebookf49af640fe Draft saved

File Edit View Run Help

Run All | Code | Draft Session (7m)

Data | + Add data

Input

titanic-extended
- full.csv
- test.csv
- train.csv

Output (6MB / 1.9GB)

/kaggle/working

Settings

Code Help

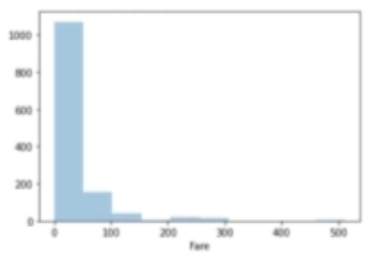
FIND CODE HELP

Find Code Help

Search for examples of how to do things

[16]:
```
sns.distplot(df['Fare'], kde=False, bins=10)
```

[16]: <AxesSubplot:xlabel='Fare'>



[18]:
```
sns.jointplot(x='Age', y='Fare', data=df)
```

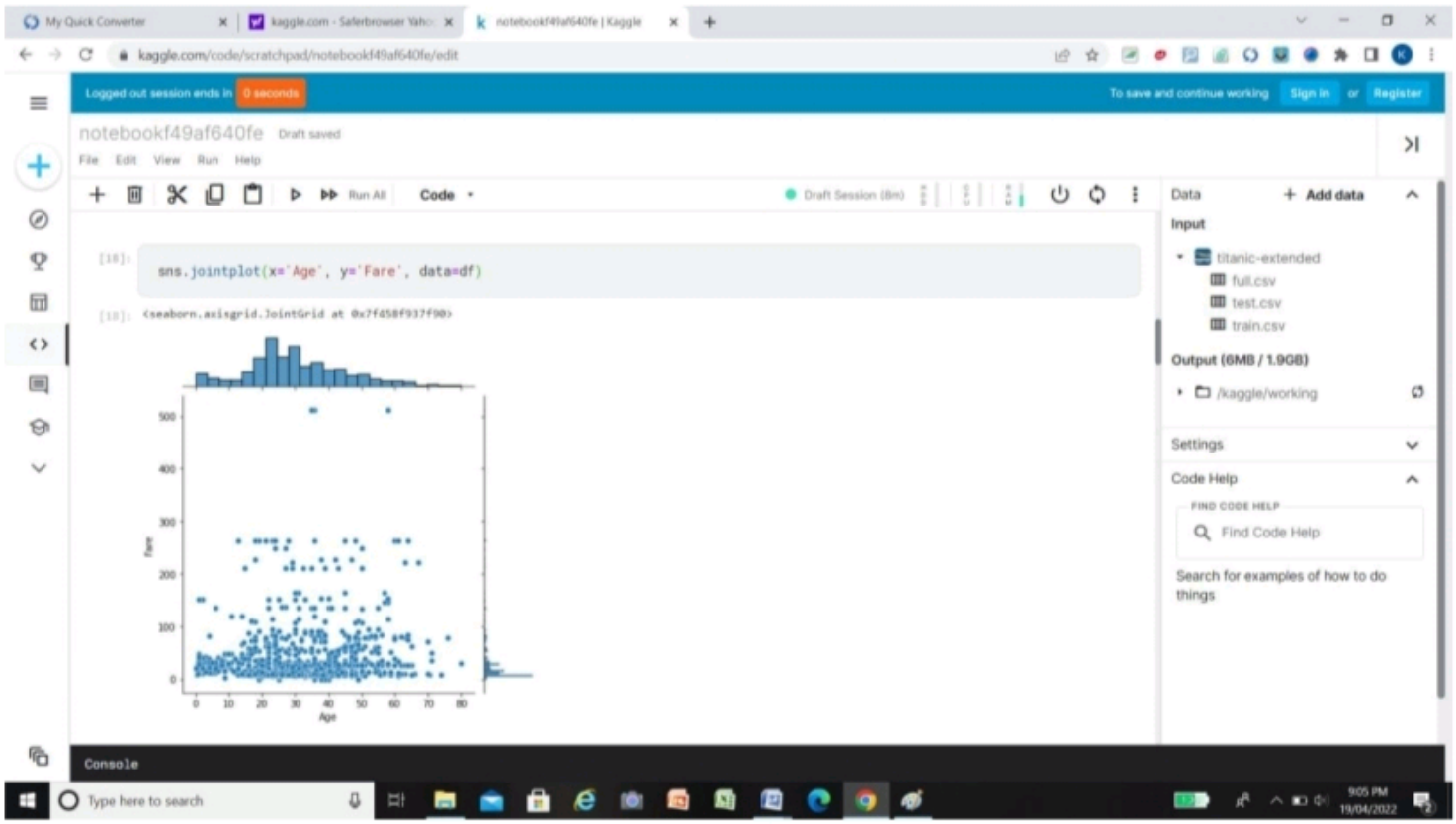
[18]: <seaborn.axisgrid.JointGrid at 0x7f458f937f90>

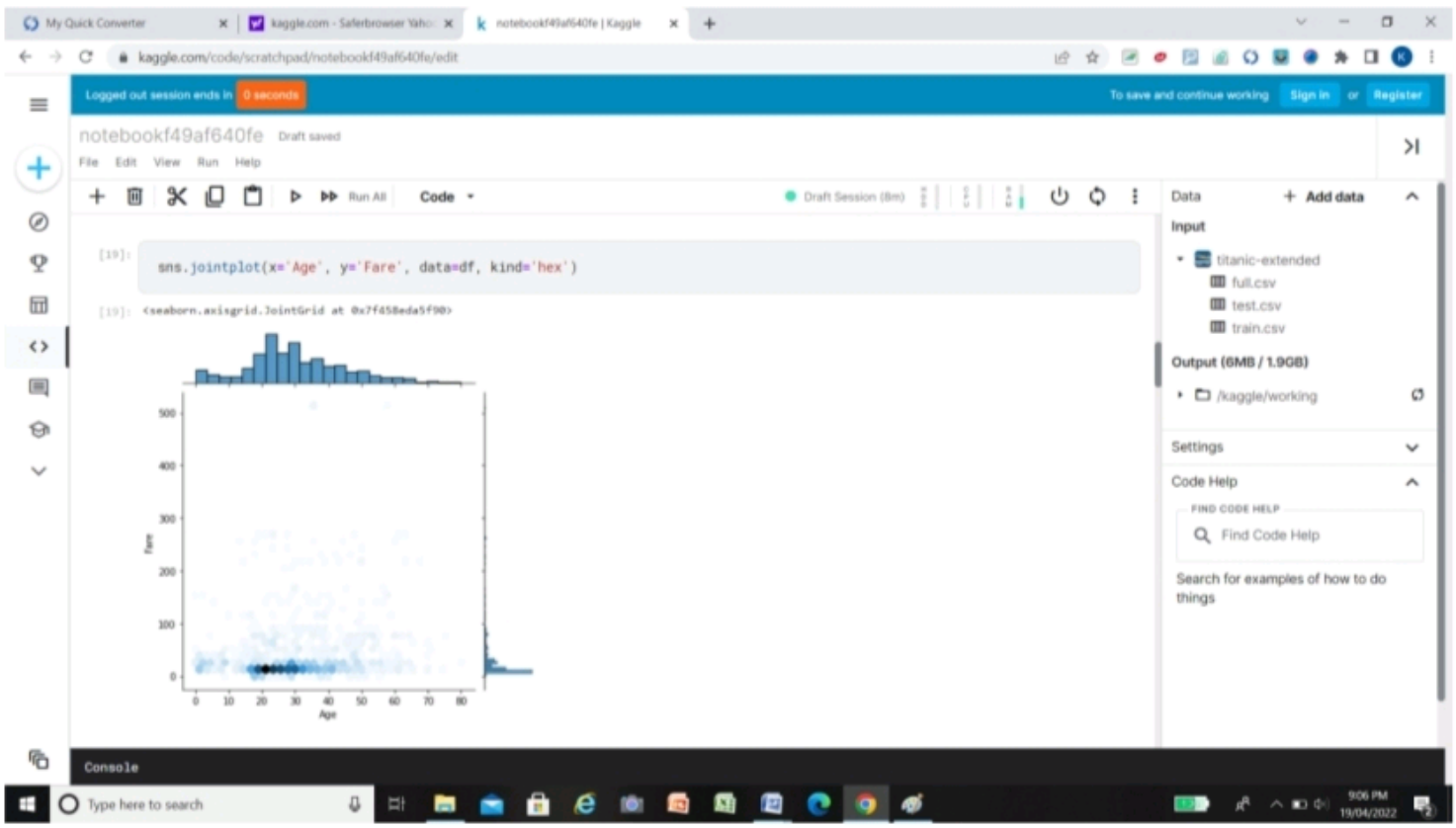
Console

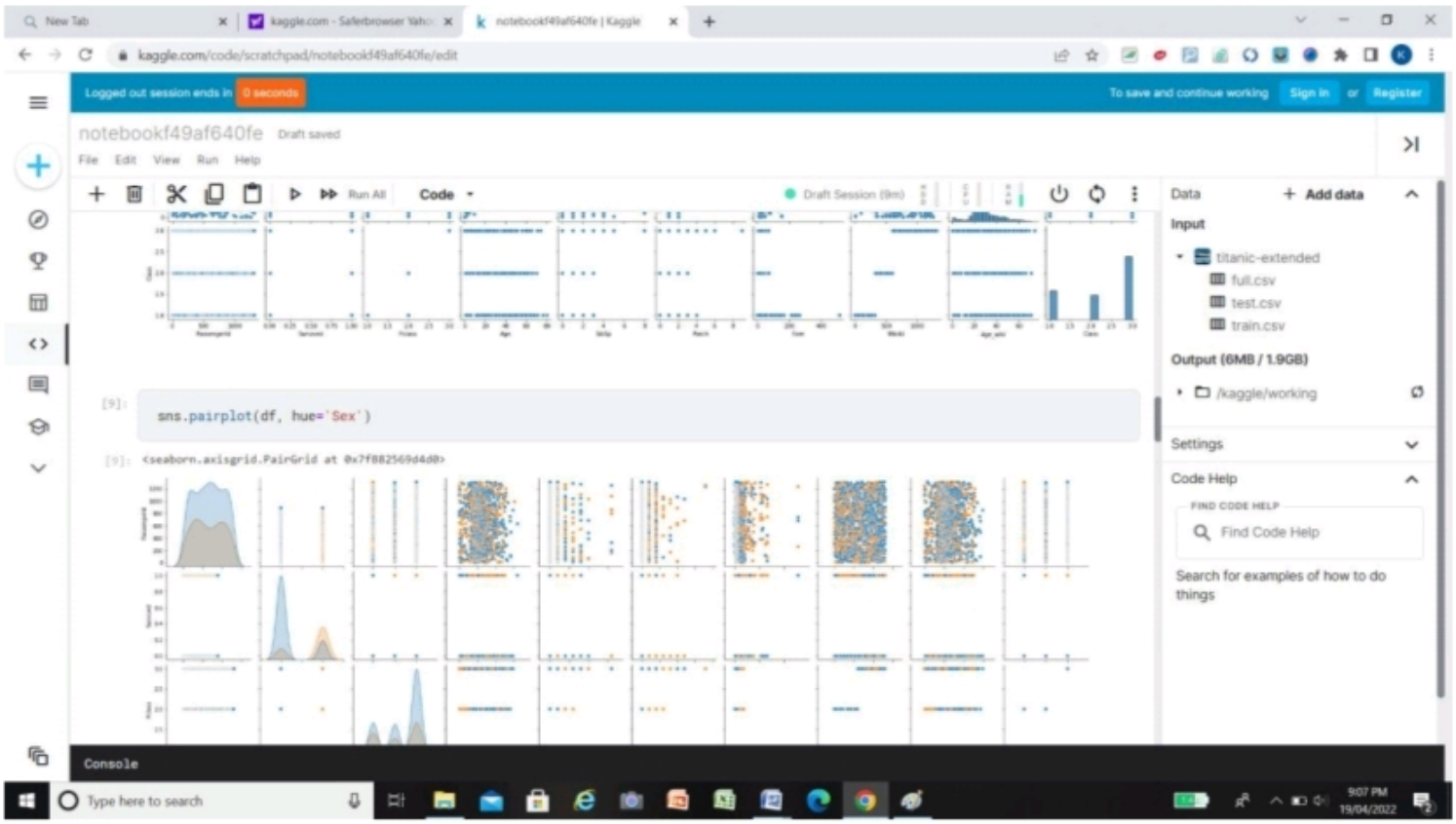
Type here to search

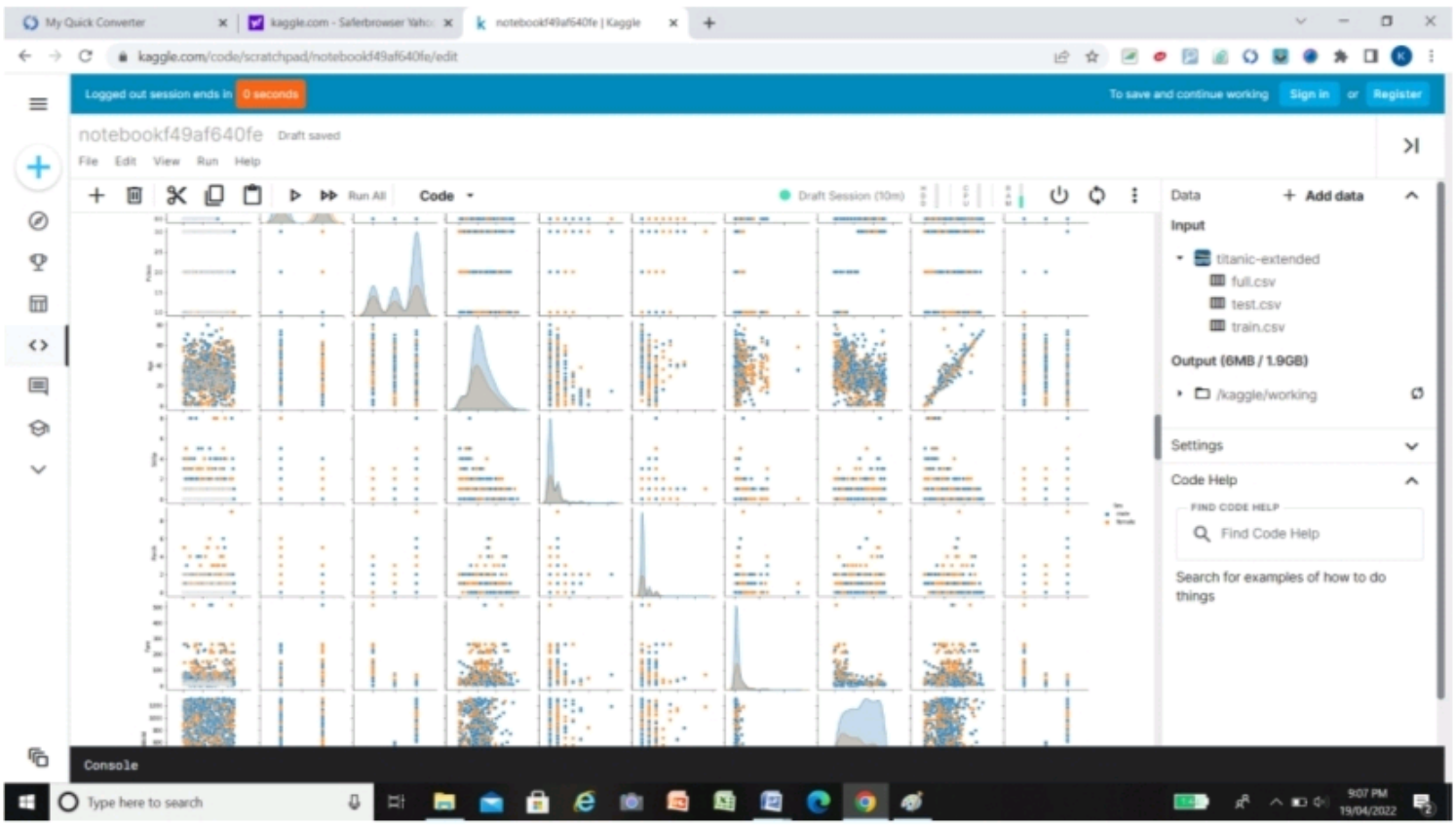
9:04 PM 19/04/2022

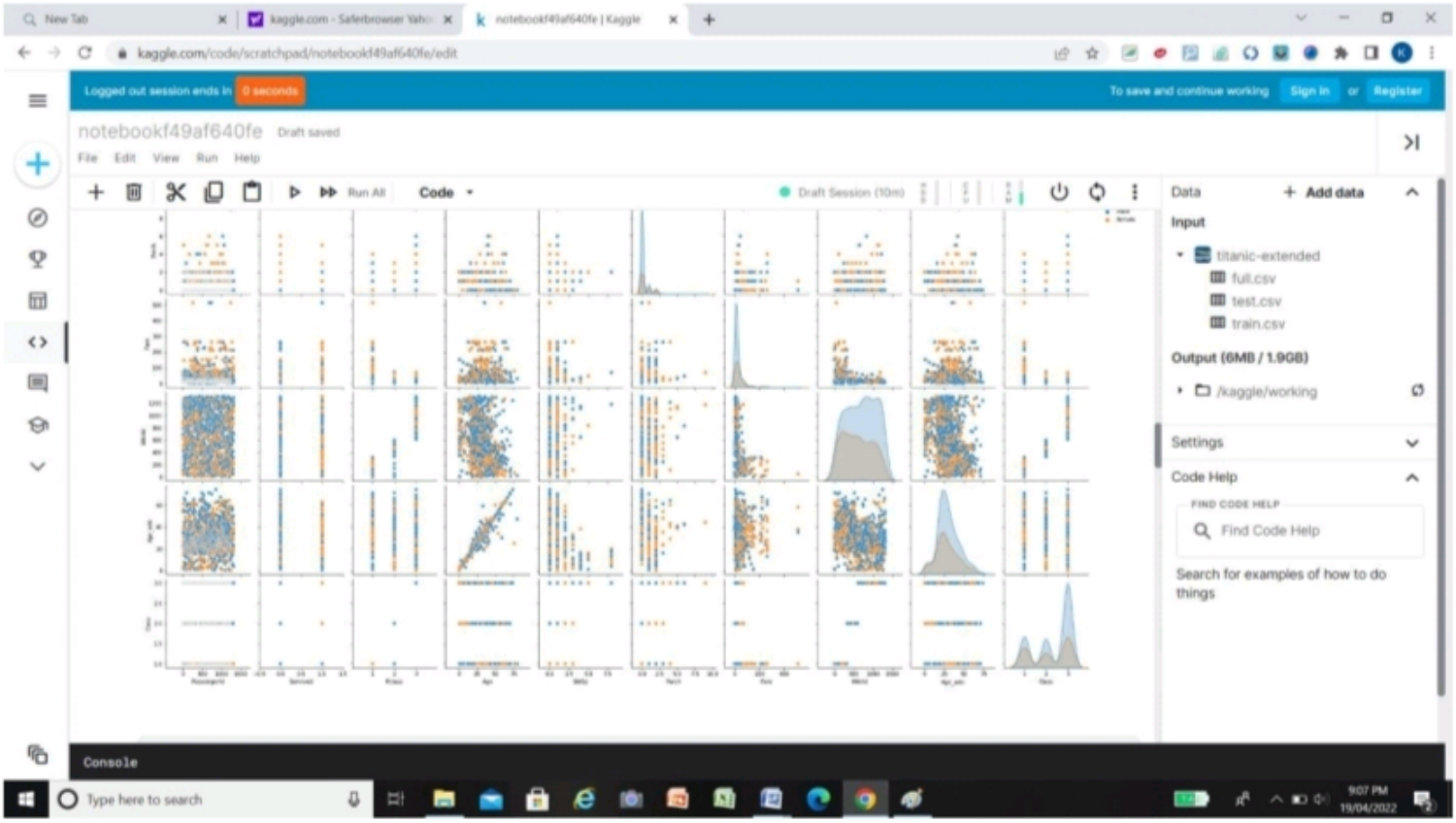
The Joint Plot :



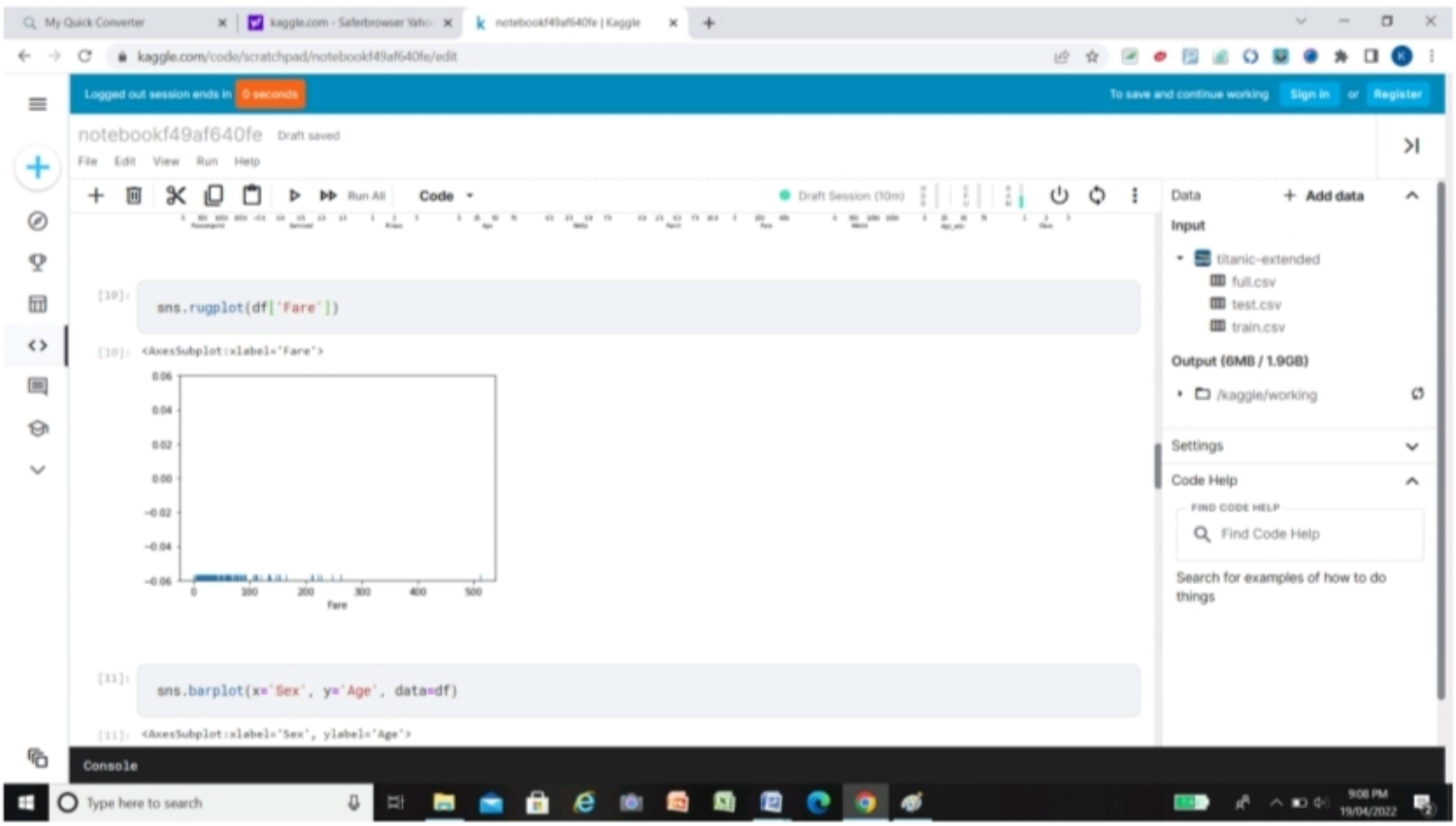



kaggle.com/code/scratchpad/notebookf49af640fe/edit
Logged out session ends in 0 seconds
To save and continue working Sign In or Register
notebookf49af640fe Draft saved
File Edit View Run Help
Run All Code Draft Session (9m)
Data + Add data
Input
titanic-extended
full.csv
test.csv
train.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
[9]: sns.pairplot(df, hue='Sex')
[9]: <seaborn.axisgrid.PairGrid at 0x7f882569d4d0>
Console
Type here to search
9:07 PM
19/04/2022

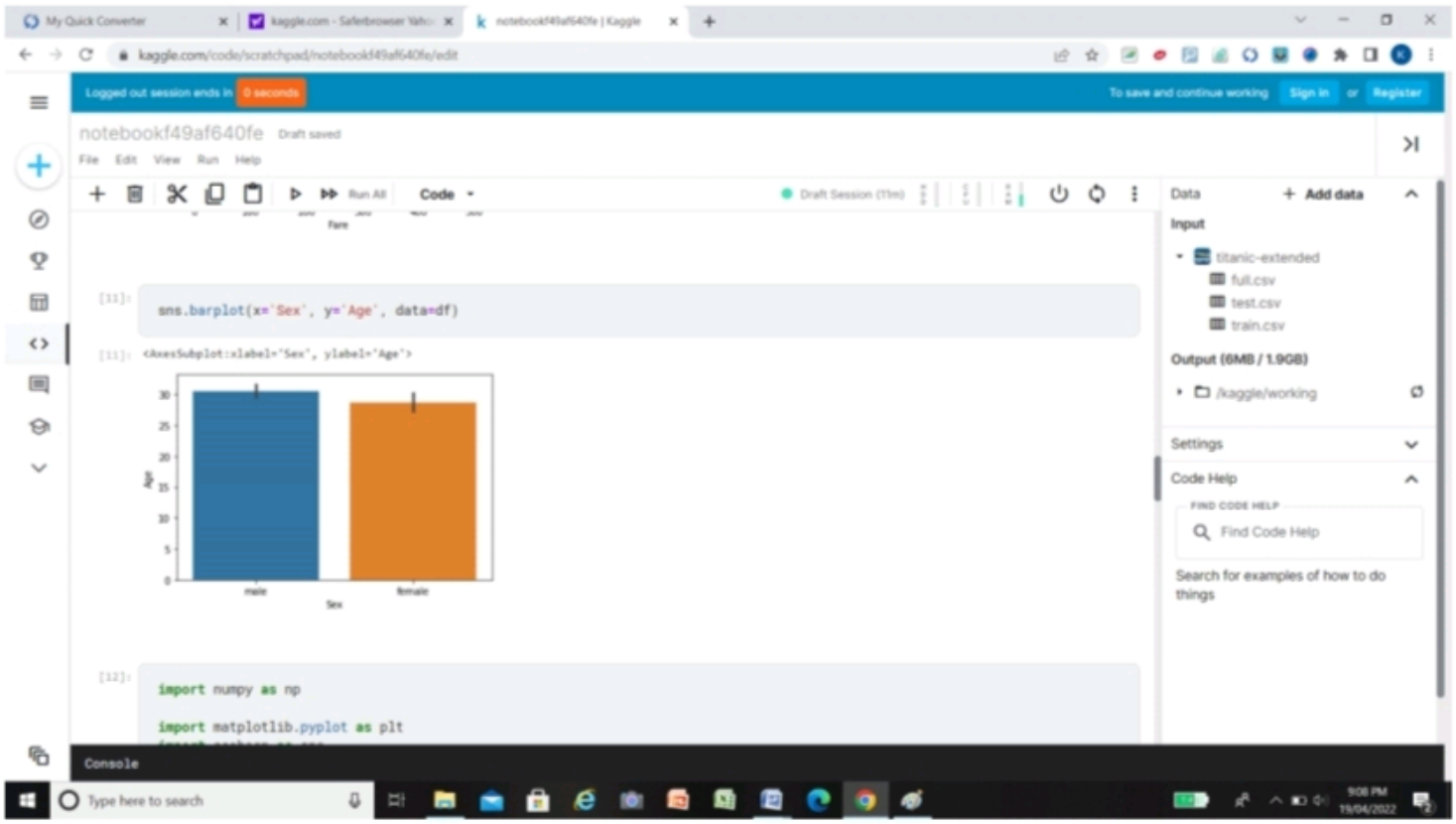

kaggle.com/code/scratchpad/notebookf49af640fe/edit
Logged out session ends in 0 seconds
To save and continue working Sign In or Register
notebookf49af640fe Draft saved
File Edit View Run Help
Run All Code Draft Session (10m)
Data + Add data
Input
titanic-extended
full.csv
test.csv
train.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Console
Type here to search
9:07 PM
19/04/2022

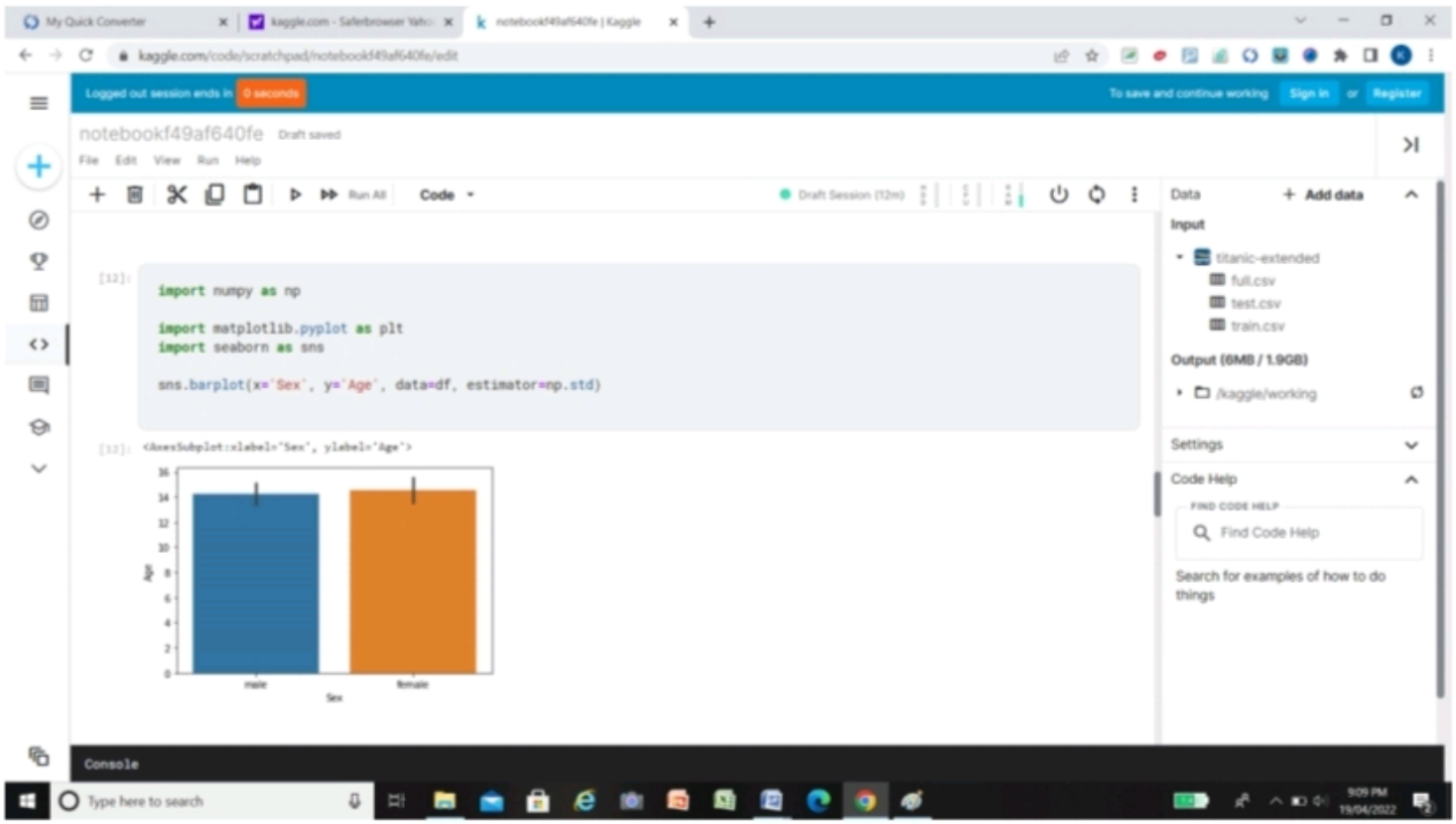

notebookf49af640fe
Draft saved
File Edit View Run Help
Run All
Code
Draft Session (10m)
Data
Add data
Input
titanic-extended
full.csv
test.csv
train.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
Console

The Rug Plot :

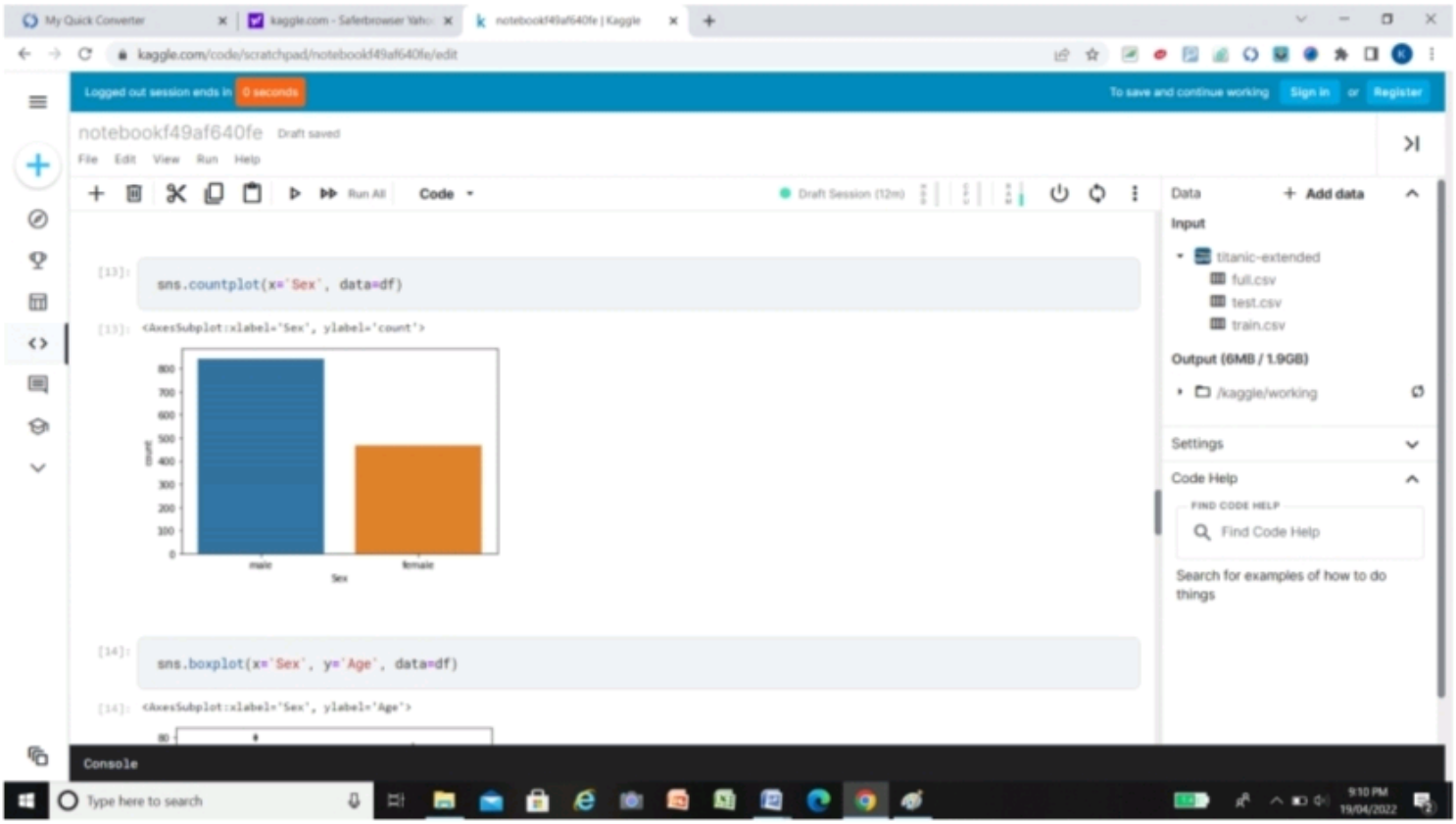

notebookf49af640fe
Draft saved
File Edit View Run Help
Run All
Code
Draft Session (10m)
Data
Add data
Input
titanic-extended
full.csv
test.csv
train.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
[10]:
sns.rugplot(df['Fare'])
[10]:
<AxesSubplot:xlabel='Fare'>
Fare
[11]:
sns.barplot(x='Sex', y='Age', data=df)
[11]:
<AxesSubplot:xlabel='Sex', ylabel='Age'>
Console

The Bar Plot :

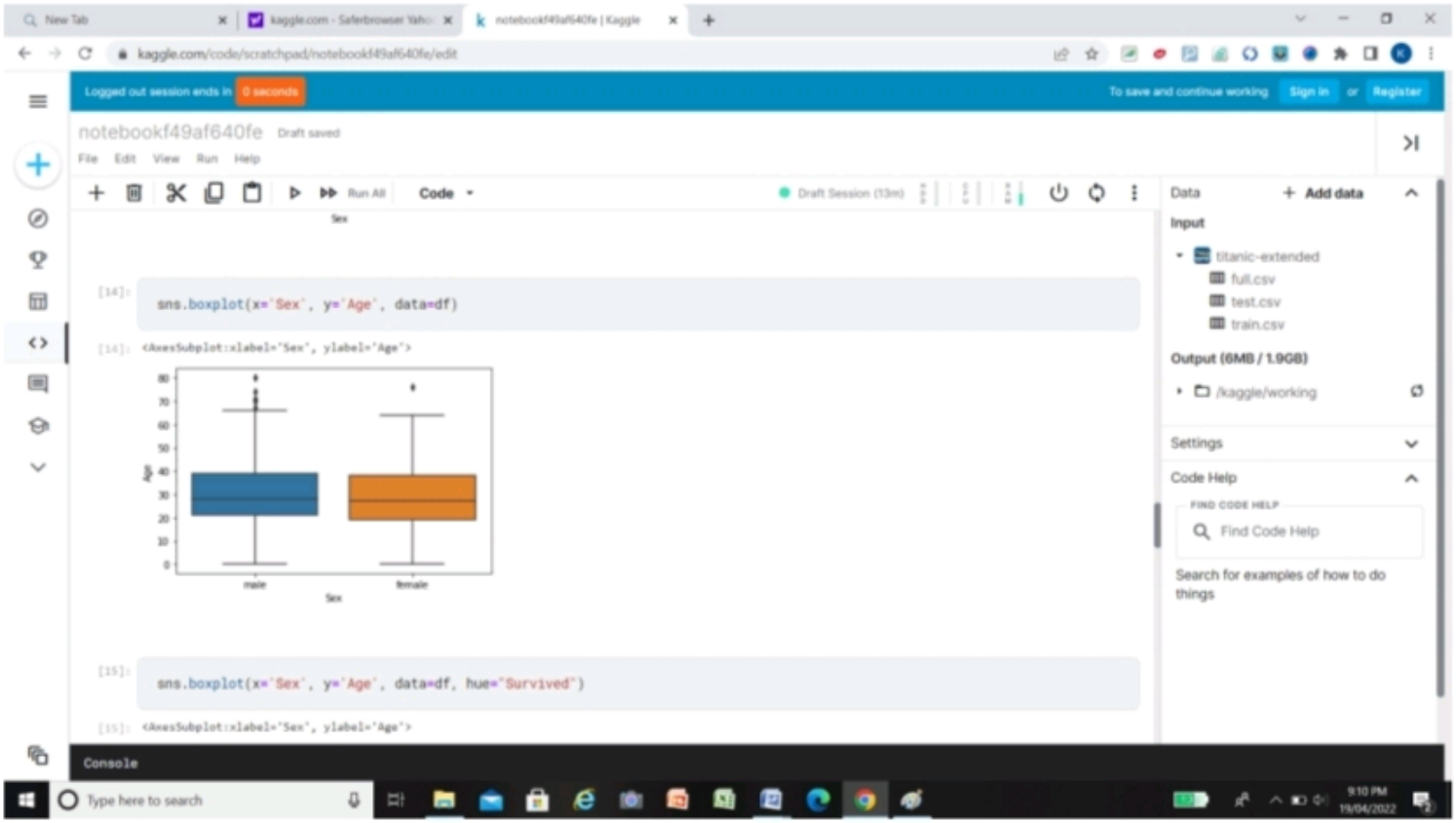

notebookf49af640fe
Draft saved
File Edit View Run Help
Run All
Code
Draft Session (11m)
Data
Add data
Input
titanic-extended
full.csv
test.csv
train.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
[11]:
sns.barplot(x='Sex', y='Age', data=df)
[11]:
<AxesSubplot:xlabel='Sex', ylabel='Age'>
Age
Sex
male
female
[12]:
import numpy as np
import matplotlib.pyplot as plt
Console

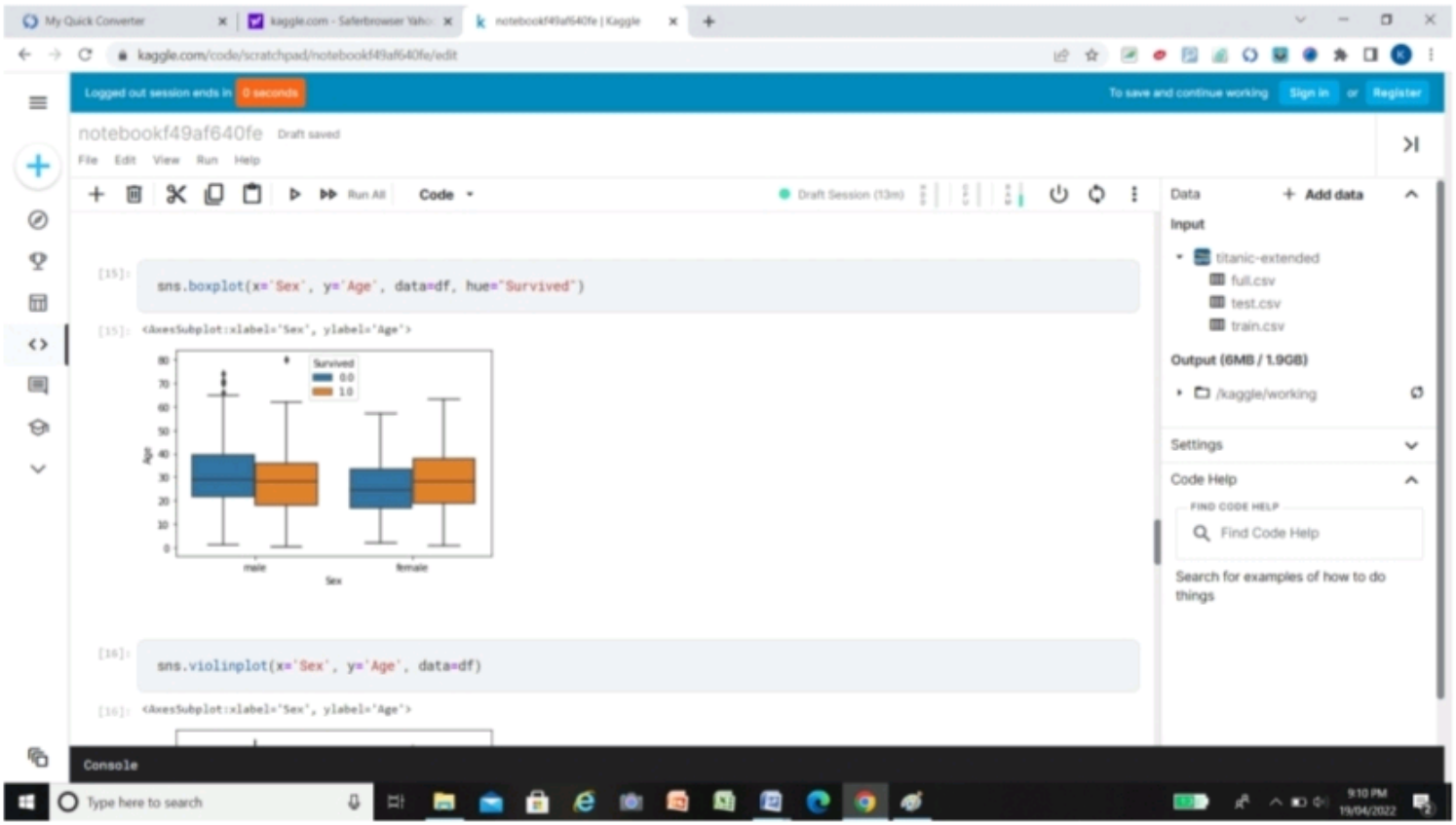

notebookf49af640fe
Draft saved
File Edit View Run Help
Run All
Code
Draft Session (12m)
Data
Add data
Input
titanic-extended
full.csv
test.csv
train.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
[12]:
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
sns.barplot(x='Sex', y='Age', data=df, estimator=np.std)
[12]:
<AxesSubplot:xlabel='Sex', ylabel='Age'>
Age
Sex
male
female
Console

The Count Plot :

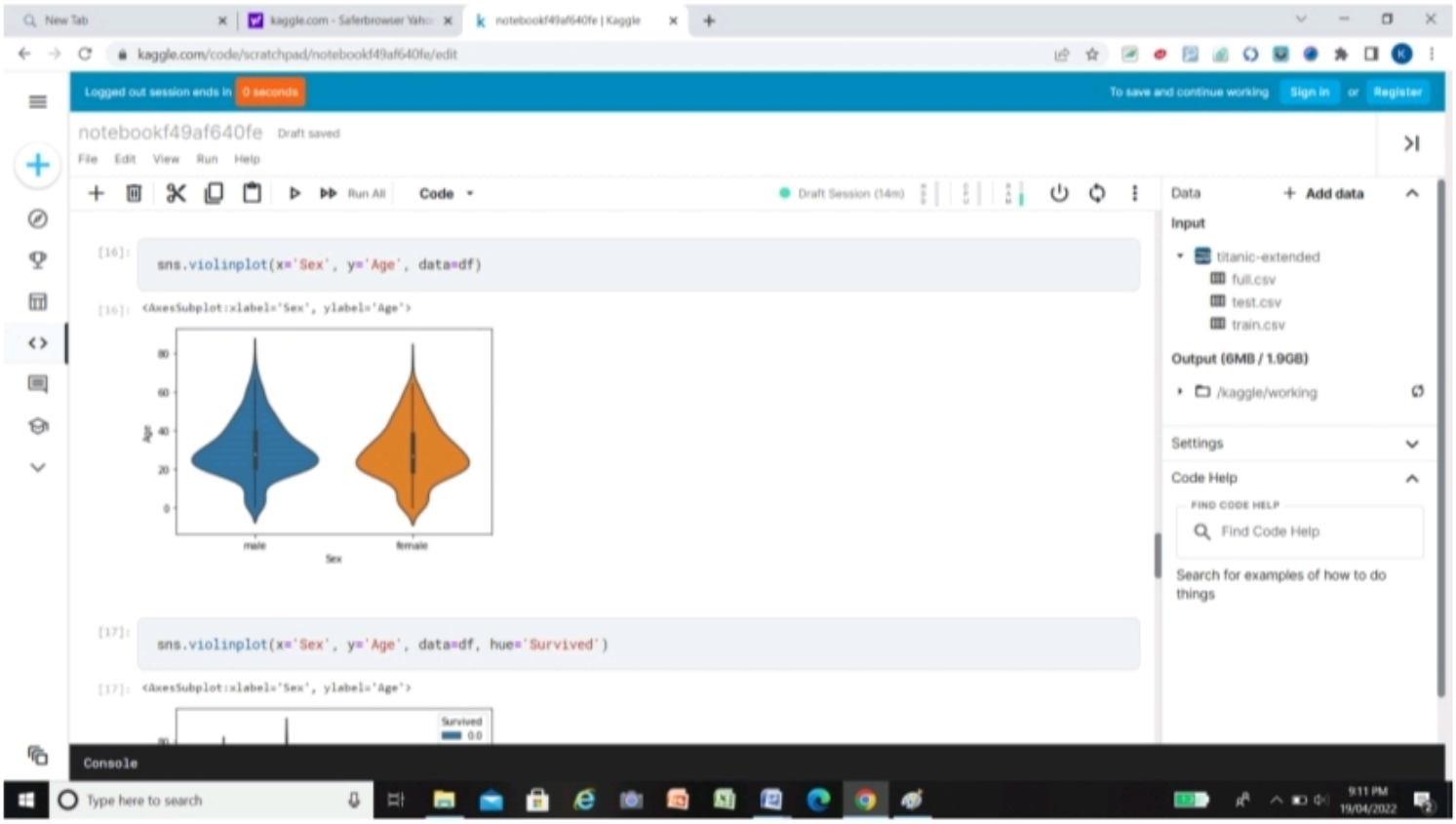


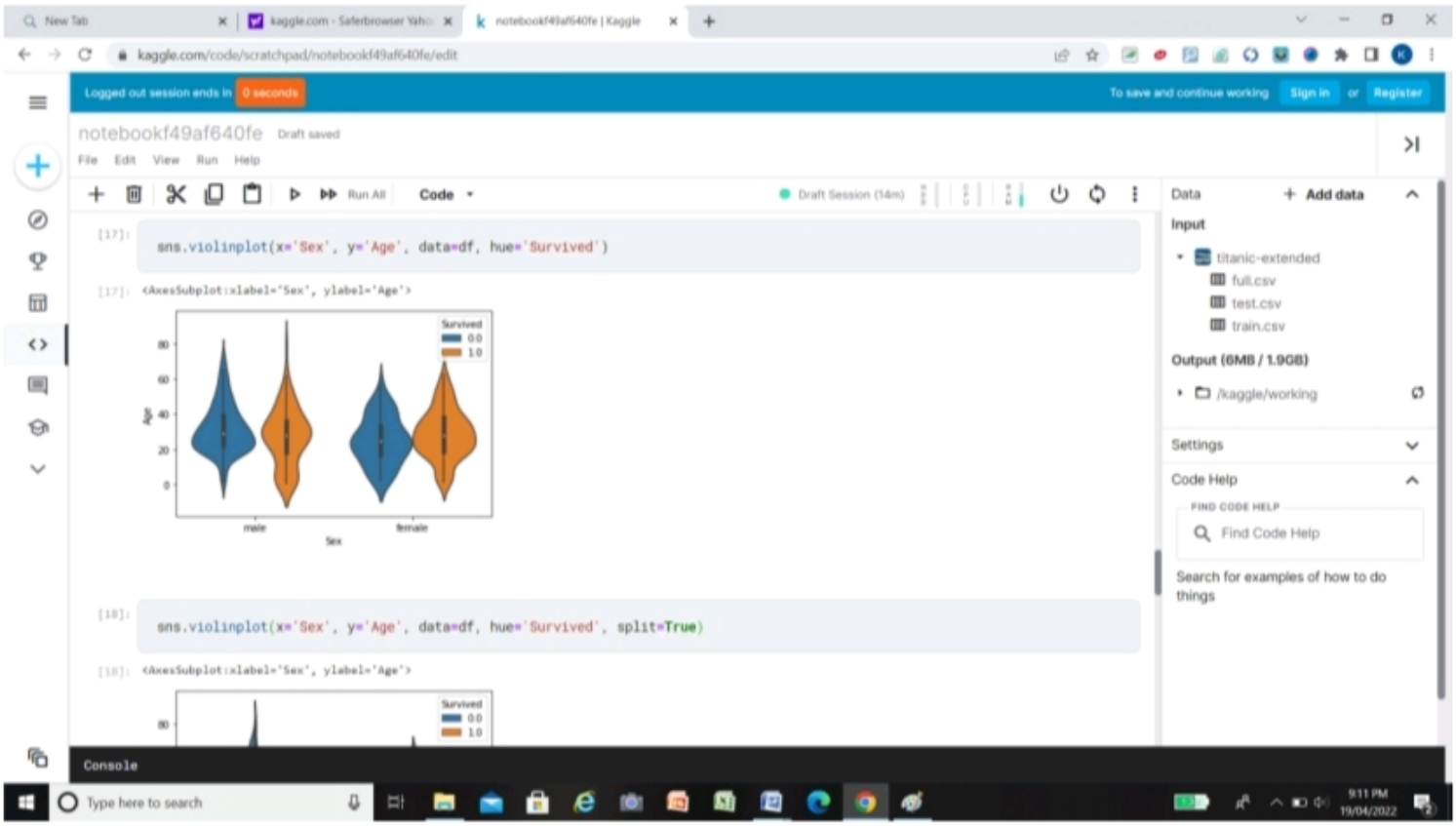
The Box Plot :


kaggle.com/code/scratchpad/notebookf49af640fe/edit
notebookf49af640fe
Draft saved
File Edit View Run Help
Run All
Code
Draft Session (13m)
Logged out session ends in 0 seconds
To save and continue working Sign in or Register
Data
Add data
Input
titanic-extended
full.csv
test.csv
train.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
[14]: sns.boxplot(x='Sex', y='Age', data=df)
[14]: <AxesSubplot:xlabel='Sex', ylabel='Age'>
Age
Sex
male
female
[15]: sns.boxplot(x='Sex', y='Age', data=df, hue="Survived")
[15]: <AxesSubplot:xlabel='Sex', ylabel='Age'>
Console
Type here to search
9:10 PM
19/04/2022


kaggle.com/code/scratchpad/notebookf49af640fe/edit
notebookf49af640fe
Draft saved
File Edit View Run Help
Run All
Code
Draft Session (13m)
Logged out session ends in 0 seconds
To save and continue working Sign in or Register
Data
Add data
Input
titanic-extended
full.csv
test.csv
train.csv
Output (6MB / 1.9GB)
/kaggle/working
Settings
Code Help
FIND CODE HELP
Find Code Help
Search for examples of how to do things
[15]: sns.boxplot(x='Sex', y='Age', data=df, hue="Survived")
[15]: <AxesSubplot:xlabel='Sex', ylabel='Age'>
Survived
0.0
1.0
Age
Sex
male
female
[16]: sns.violinplot(x='Sex', y='Age', data=df)
[16]: <AxesSubplot:xlabel='Sex', ylabel='Age'>
Console
Type here to search
9:10 PM
19/04/2022

The Violin Plot :





My Quick Converter | kaggle.com - Safebrowser Yahoo | notebookf49af640fe | Kaggle

kaggle.com/code/scratchpad/notebookf49af640fe/edit

Logged out session ends in 0 seconds — To save and continue working Sign in or Register

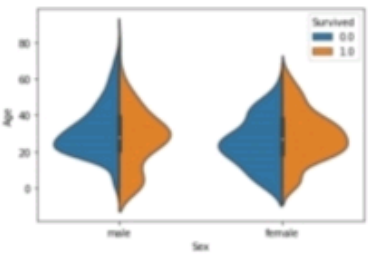
notebookf49af640fe Draft saved

File Edit View Run Help

Run All | Code | Draft Session (14m)

[18]:
```
sns.violinplot(x='Sex', y='Age', data=df, hue='Survived', split=True)
```

[18]: <AxesSubplot:xlabel='Sex', ylabel='Age'>



[19]:

Console

Data | + Add data

Input

titanic-extended
full.csv
test.csv
train.csv

Output (6MB / 1.9GB)

/kaggle/working

Settings

Code Help

FIND CODE HELP

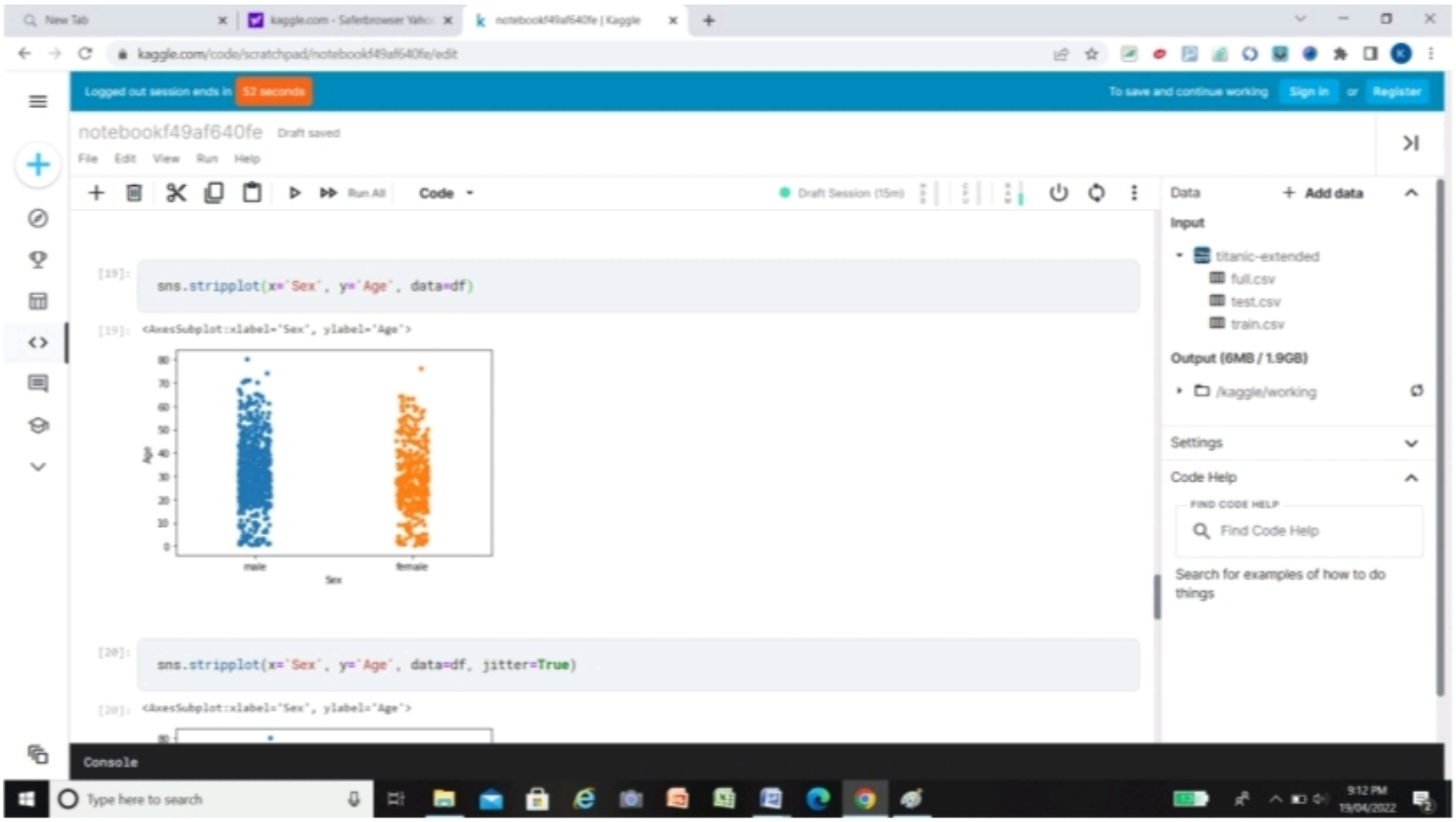
Find Code Help

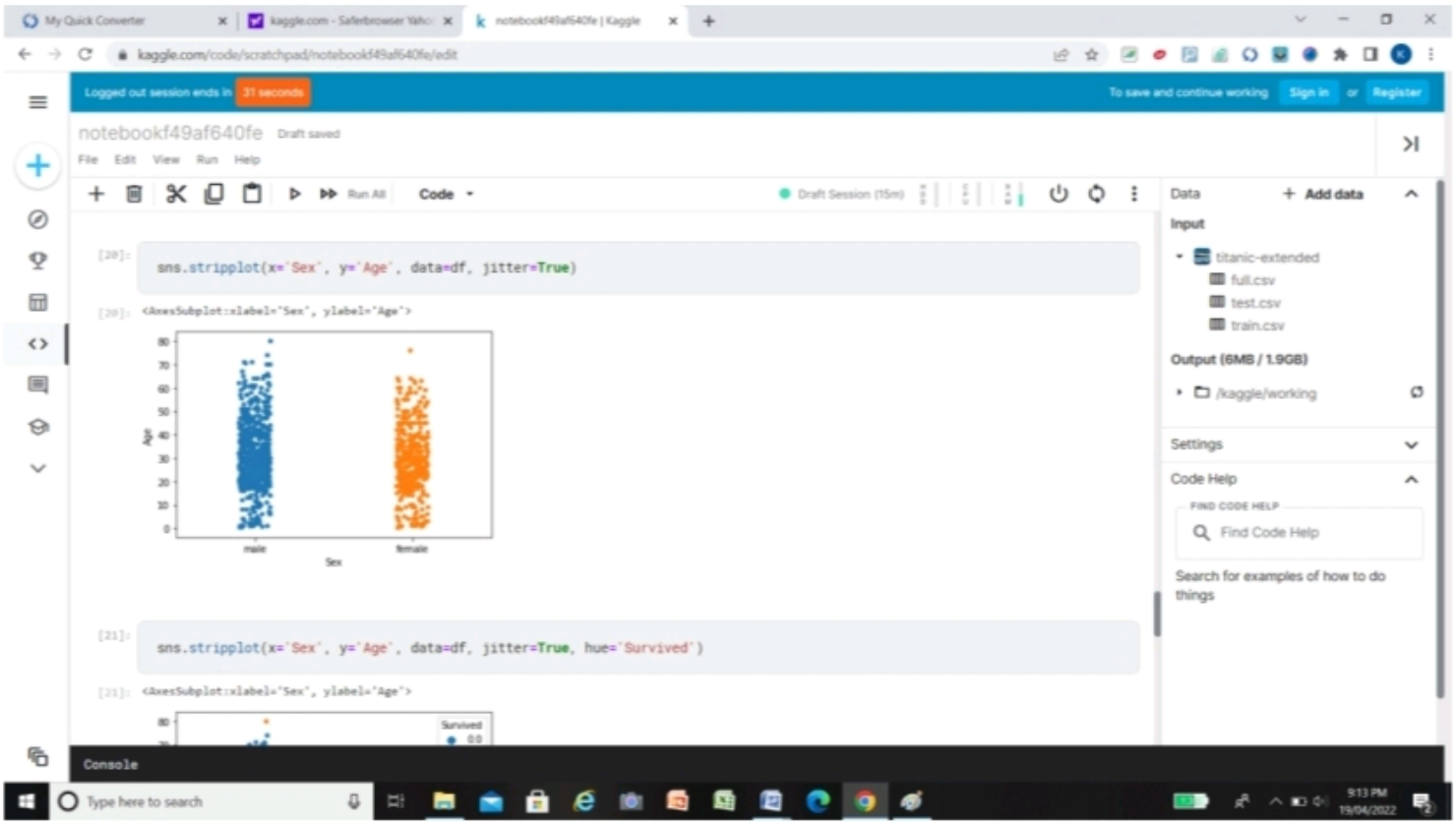
Search for examples of how to do things

Type here to search

9:12 PM
19/04/2022

The Strip Plot :





New Tab | kaggle.com - Saferbrowser Yahoo | notebookf49af640fe | Kaggle

kaggle.com/code/scratchpad/notebookf49af640fe/edit

Logged out session ends in 17 seconds — To save and continue working Sign in or Register

notebookf49af640fe Draft saved

File Edit View Run Help

Run All | Code | Draft Session (15m)

Data | + Add data

Input

titanic-extended
- full.csv
- test.csv
- train.csv

Output (6MB / 1.9GB)

/kaggle/working

Settings

Code Help

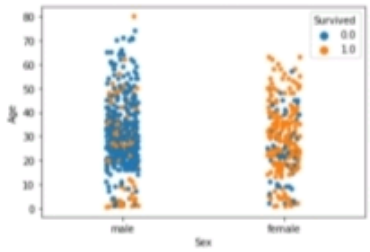
FIND CODE HELP

Find Code Help

Search for examples of how to do things

[21]:
```
sns.stripplot(x='Sex', y='Age', data=df, jitter=True, hue='Survived')
```

[21]:
```
<AxesSubplot:xlabel='Sex', ylabel='Age'>
```



[22]:
```
sns.stripplot(x='Sex', y='Age', data=df, jitter=True, hue='Survived', split=True)
```

```
/opt/conda/lib/python3.7/site-packages/seaborn/categorical.py:2805: UserWarning: The 'split' parameter has been renamed to 'dodge'.
  warnings.warn(msg, UserWarning)
```

Console

Type here to search — 9:13 PM 19/04/2022

My Quick Converter | kaggle.com - Saferbrowser Yahoo | notebookf49af640fe | Kaggle

kaggle.com/code/scratchpad/notebookf49af640fe/edit

Logged out session ends in 3 seconds — To save and continue working Sign in or Register

notebookf49af640fe Draft saved

File Edit View Run Help

Run All | Code | Draft Session (15m)

Data | + Add data

Input

titanic-extended
- full.csv
- test.csv
- train.csv

Output (6MB / 1.9GB)

/kaggle/working

Settings

Code Help

FIND CODE HELP

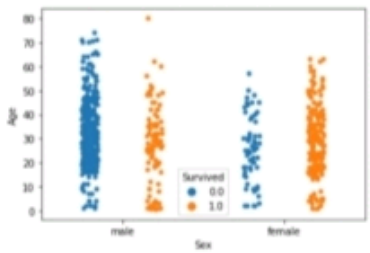
Find Code Help

Search for examples of how to do things

[22]:
```
sns.stripplot(x='Sex', y='Age', data=df, jitter=True, hue='Survived', split=True)
```

```
/opt/conda/lib/python3.7/site-packages/seaborn/categorical.py:2805: UserWarning: The 'split' parameter has been renamed to 'dodge'.
  warnings.warn(msg, UserWarning)
```

[22]:
```
<AxesSubplot:xlabel='Sex', ylabel='Age'>
```

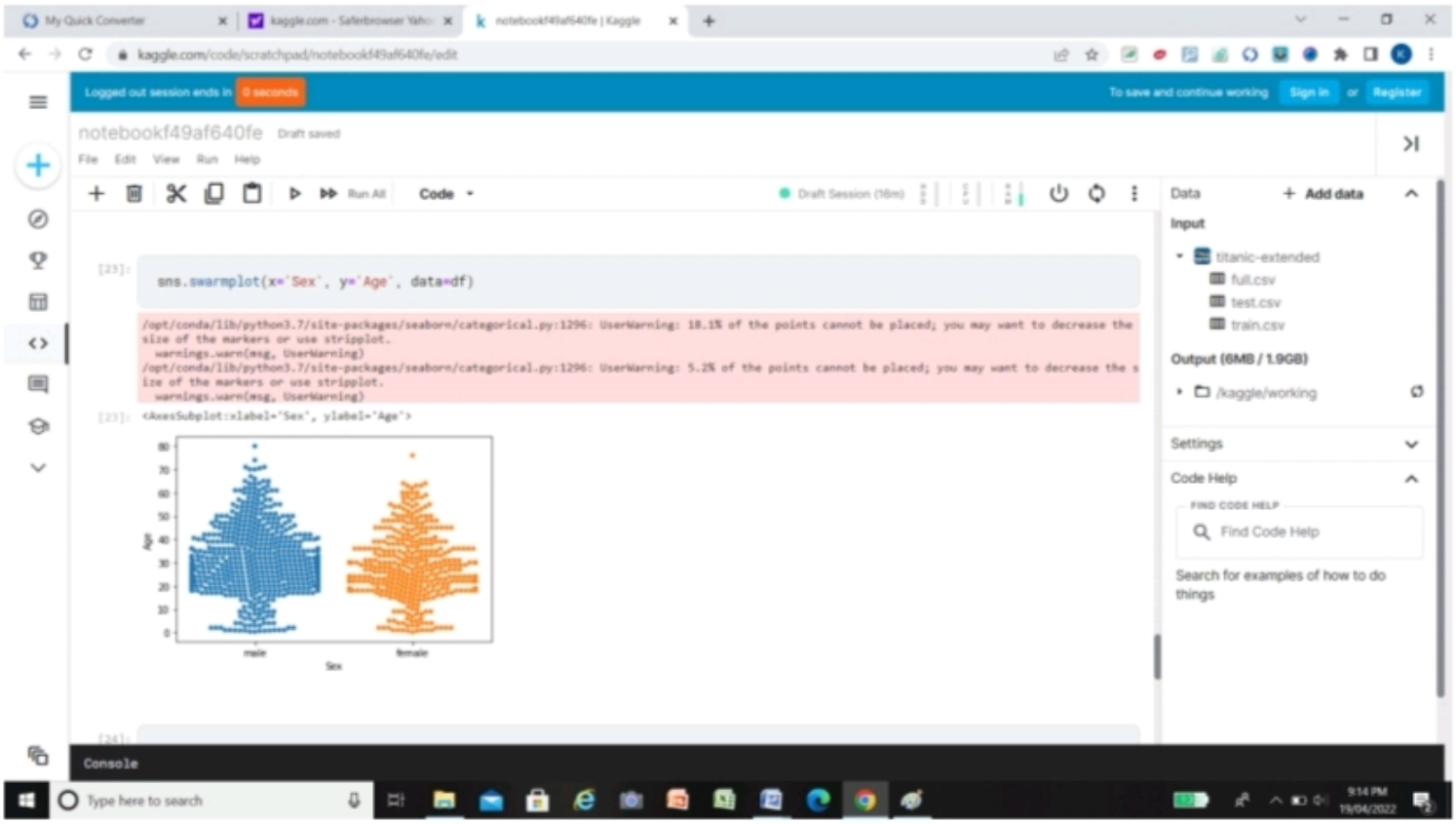


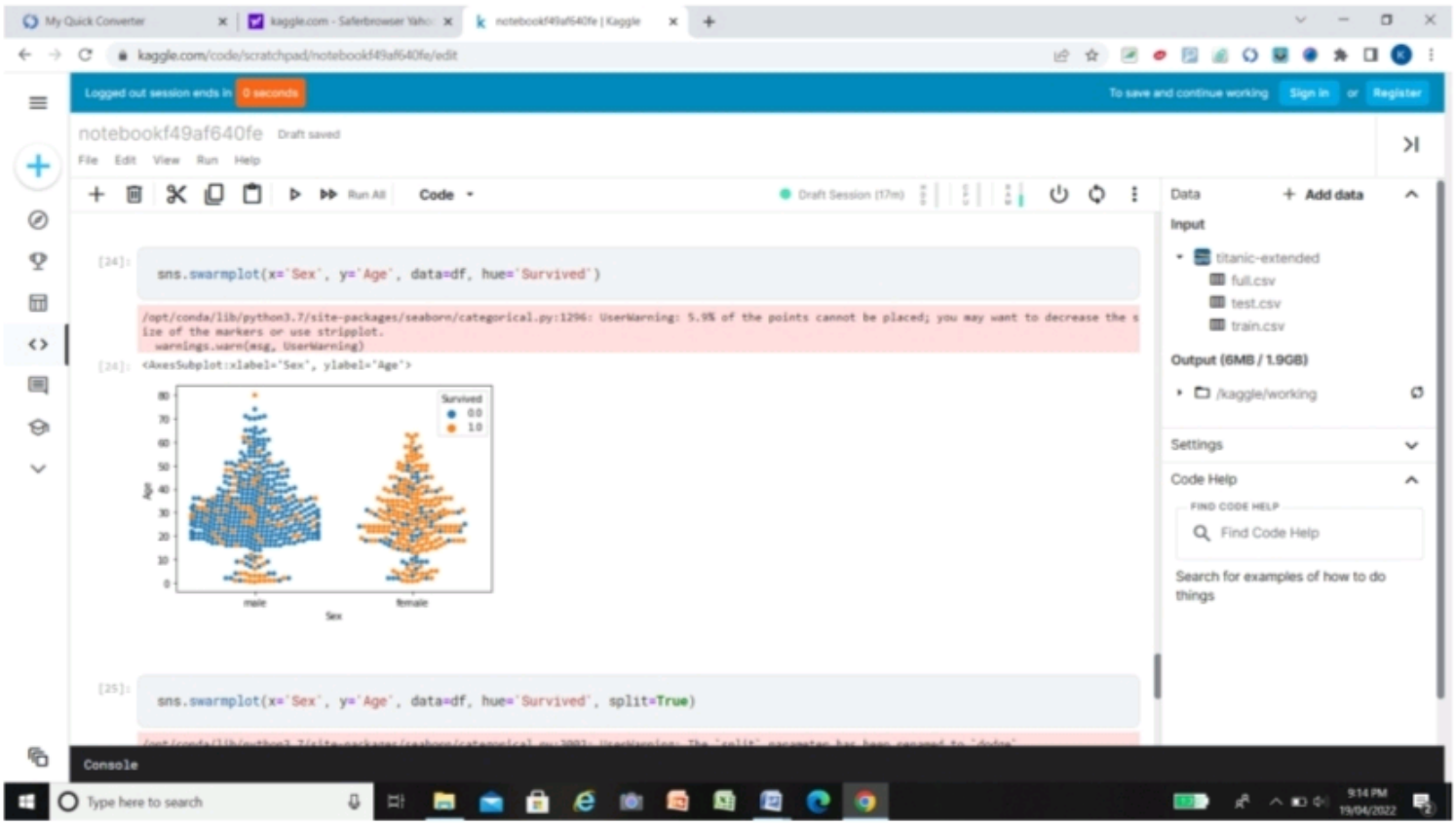
[23]:
```
sns.swarmplot(x='Sex', y='Age', data=df)
```

Console

Type here to search — 9:13 PM 19/04/2022

The Swarm Plot :





My Quick Converter | kaggle.com - Saferbrowser Yahoo | notebookf49af640fe | Kaggle

kaggle.com/code/scratchpad/notebookf49af640fe/edit

Logged out session ends in 0 seconds — To save and continue working Sign in or Register

notebookf49af640fe Draft saved

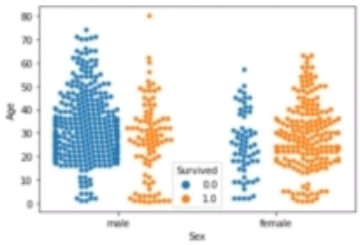
File Edit View Run Help

Run All | Code | Draft Session (17m)

[25]:
```
sns.swarmplot(x='Sex', y='Age', data=df, hue='Survived', split=True)
```

```
/opt/conda/lib/python3.7/site-packages/seaborn/categorical.py:3002: UserWarning: The `split` parameter has been renamed to `dodge`.
  warnings.warn(msg, UserWarning)
/opt/conda/lib/python3.7/site-packages/seaborn/categorical.py:1296: UserWarning: 21.6% of the points cannot be placed; you may want to decrease the
size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
/opt/conda/lib/python3.7/site-packages/seaborn/categorical.py:1296: UserWarning: 7.7% of the points cannot be placed; you may want to decrease the s
ize of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
```

[25]: <AxesSubplot:xlabel='Sex', ylabel='Age'>



Data — + Add data

Input

- titanic-extended
  - full.csv
  - test.csv
  - train.csv

Output (6MB / 1.9GB)

/kaggle/working

Settings

Code Help

FIND CODE HELP

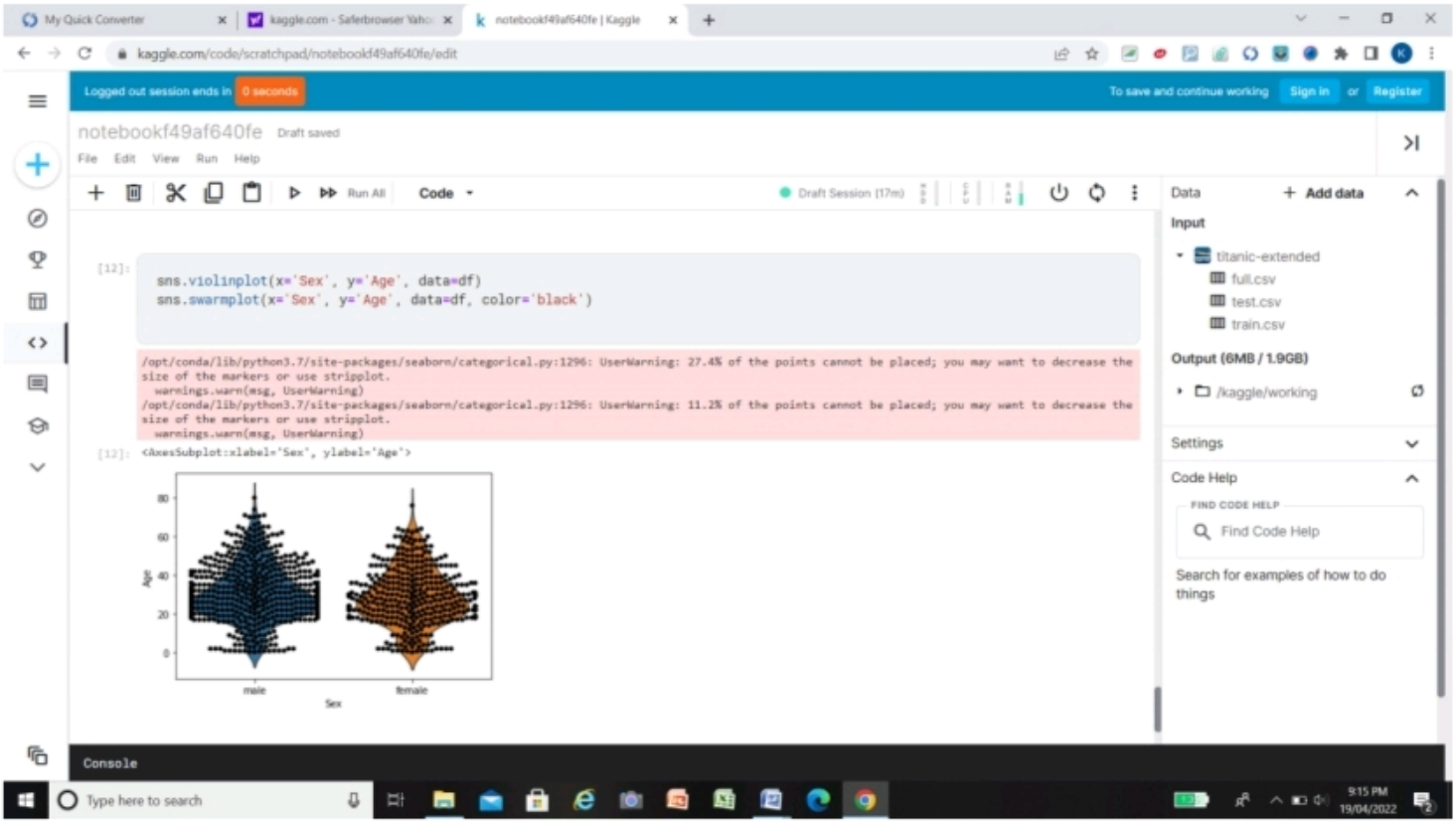
Find Code Help

Search for examples of how to do things

Console

Type here to search — 9:14 PM 19/04/2022

Combining Swarm and Violin Plot :



Assignment No. 10

Aim : Data Visualization III

Download the Iris flower dataset or any other dataset into a DataFrame. (e.g.,

https://archive.ics.uci.edu/ml/datasets/Iris). Scan the dataset and give the inference as:

1. List down the features and their types (e.g., numeric, nominal) available in the dataset.

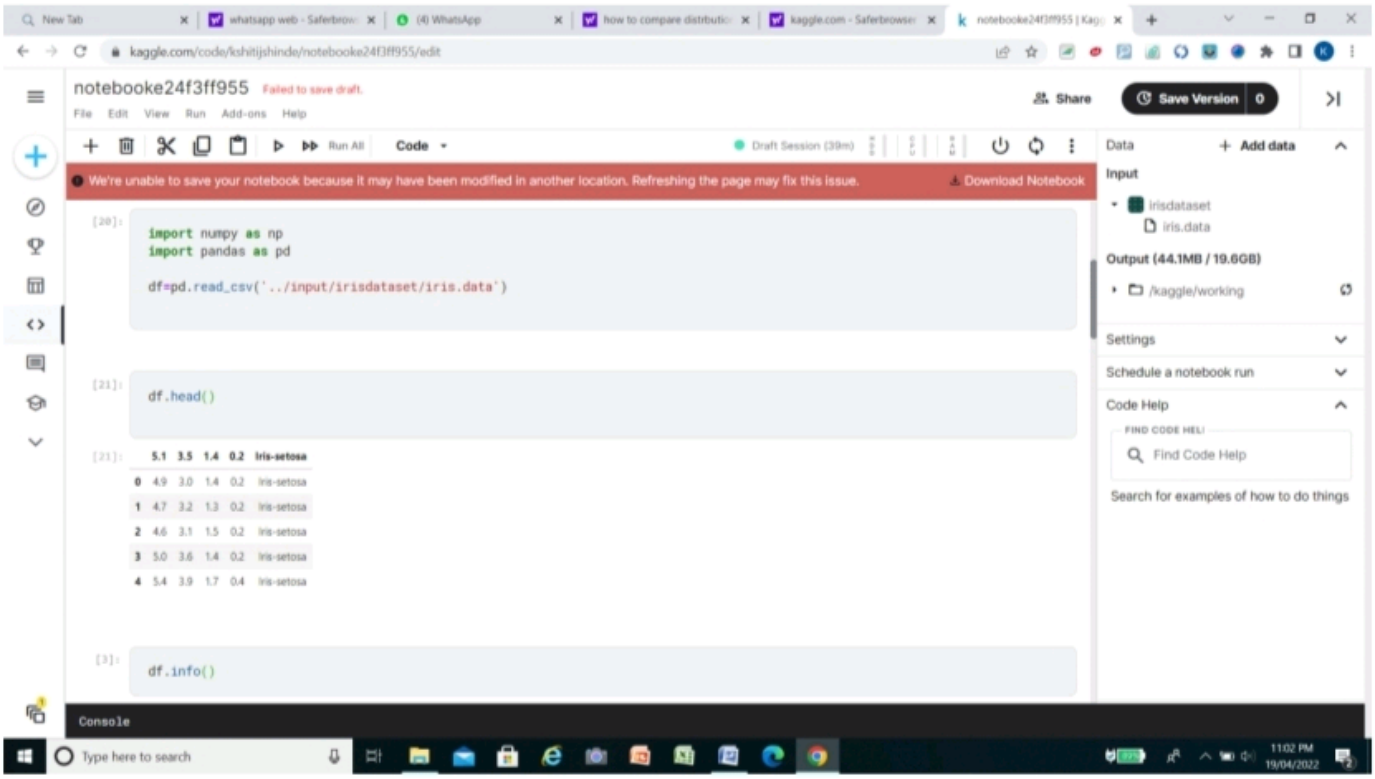
2. Create a histogram for each feature in the dataset to illustrate the feature distributions.

3. Create a box plot for each feature in the dataset.

4. Compare distributions and identify outliers.

Sol^{n} :

OUTPUT :


notebooke24f3ff955
Failed to save draft.
File Edit View Run Add-ons Help
Share
Save Version 0
Run All
Code
Draft Session (39m)
We're unable to save your notebook because it may have been modified in another location. Refreshing the page may fix this issue.
Download Notebook
[20]:
import numpy as np
import pandas as pd
df=pd.read_csv('../input/irisdataset/iris.data')
[21]:
df.head()
[21]:
5.1 3.5 1.4 0.2 Iris-setosa
0 4.9 3.0 1.4 0.2 Iris-setosa
1 4.7 3.2 1.3 0.2 Iris-setosa
2 4.6 3.1 1.5 0.2 Iris-setosa
3 5.0 3.6 1.4 0.2 Iris-setosa
4 5.4 3.9 1.7 0.4 Iris-setosa
[3]:
df.info()
Console
Data
Add data
Input
irisdataset
iris.data
Output (44.1MB / 19.6GB)
/kaggle/working
Settings
Schedule a notebook run
Code Help
Find Code Help
Search for examples of how to do things
Type here to search
11:02 PM
19/04/2022