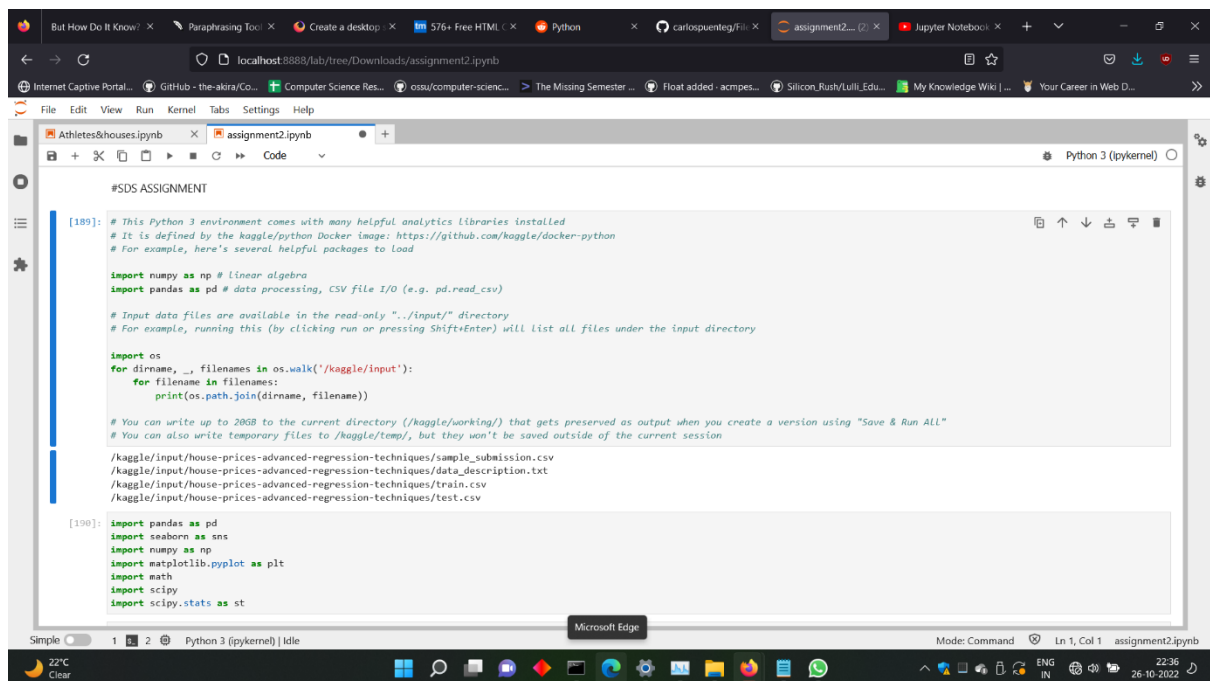


Nandan N

PES1UG21CS361

Roll no. 63, F Section

SDS Assignment



The screenshot shows a Jupyter Notebook titled 'assignment2.ipynb' in a web browser. The code in the notebook is as follows:

```
#SDS ASSIGNMENT

[189]: # This Python 3 environment comes with many helpful analytics libraries installed
# It is defined by the kaggle/python Docker image: https://github.com/kaggle/docker-python
# For example, here's several helpful packages to load

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

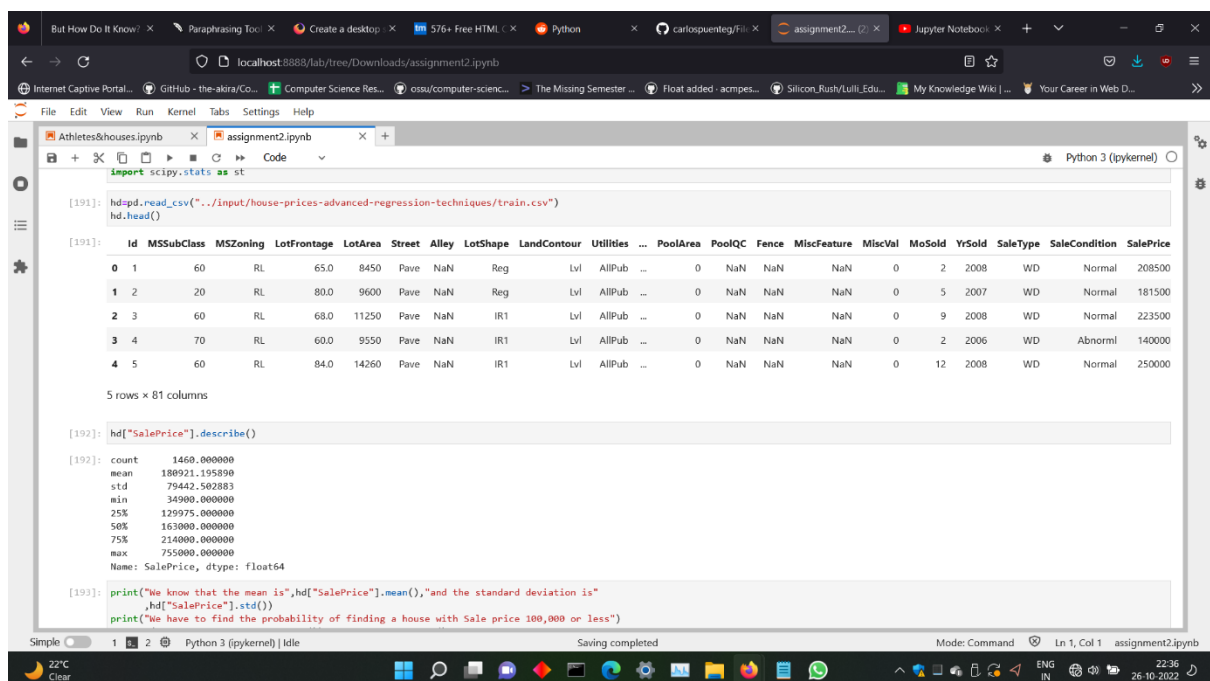
# Input data files are available in the read-only "../input/" directory
# For example, running this (by clicking run or pressing Shift+Enter) will list all files under the input directory

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

# You can write up to 20GB to the current directory (/kaggle/working/) that gets preserved as output when you create a version using "Save & Run All"
# You can also write temporary files to /kaggle/temp/, but they won't be saved outside of the current session

/kaggle/input/house-prices-advanced-regression-techniques/sample_submission.csv
/kaggle/input/house-prices-advanced-regression-techniques/data_description.txt
/kaggle/input/house-prices-advanced-regression-techniques/train.csv
/kaggle/input/house-prices-advanced-regression-techniques/test.csv

[190]: import pandas as pd
import seaborn as sns
import numpy as np
import matplotlib.pyplot as plt
import math
import scipy
import scipy.stats as st
```



The screenshot shows the same Jupyter Notebook with the following code and output:

```
[191]: hd=pd.read_csv("../input/house-prices-advanced-regression-techniques/train.csv")
hd.head()
```

[191]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	PoolArea	PoolQC	Fence	MiscFeature	MiscVal	MoSold	YrSold	SaleType	SaleCondition	SalePrice	
0	1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	2	2008	WD	Normal	208500
1	2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	5	2007	WD	Normal	181500
2	3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	9	2008	WD	Normal	223500
3	4	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	2	2006	WD	Abnorml	140000
4	5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	12	2008	WD	Normal	250000

5 rows x 81 columns

```
[192]: hd["SalePrice"].describe()
```

[192]:

```
count    1460.000000
mean    180921.195890
std      79442.502883
min      34900.000000
25%     129975.000000
50%     163000.000000
75%     214000.000000
max      755000.000000
Name: SalePrice, dtype: float64
```

```
[193]: print("We know that the mean is",hd["SalePrice"].mean(),"and the standard deviation is",
hd["SalePrice"].std())
print("We have to find the probability of finding a house with Sale price 100,000 or less")
```

But How Do It Know? x Paraphrasing Tool x Create a desktop x 576+ Free HTML x Python x carlospueg/Fil x assignment2... (2) x Jupyter Notebook x + -

localhost:8888/lab/tree/Downloads/assignment2.ipynb

File Edit View Run Kernel Tabs Settings Help

Athletes&houses.ipynb x assignment2.ipynb Python 3 (ipykernel)

We know that the mean is 189921.1959941895 and the standard deviation is 79442.50288288662
We have to find the probability of finding a house with Sale price 100,000 or less
The probability of finding a house with Sale price 100,000 or less is 0.1541932762727009

```
[194]: import pandas as pd
import seaborn as sns
import numpy as np
import matplotlib.pyplot as plt
import math
import scipy
import scipy.stats as st
hdfpd.read_csv("../input/house-prices-advanced-regression-techniques/train.csv")
hd["BedroomAbvGr"].describe()
```

```
[194]: count    1460.000000
      mean     2.866438
      std     0.815778
      min      0.000000
      25%      2.000000
      50%      3.000000
      75%      3.000000
      max      8.000000
      Name: BedroomAbvGr, dtype: float64
```

```
[195]: three_bd=hd[hd['BedroomAbvGr']==3]
bd=len(hd)
three_bd_len=len(three_bd)
prob_threebd = three_bd_len/bd
print("the probability of finding a three bedroom house is",prob_threebd)

the probability of finding a three bedroom house is 0.5506849315068493
```

Simple 1 2 Python 3 (ipykernel) | Idle Mode: Command Ln 1, Col 1 assignment2.ipynb 22°C Clear 22:36 26-10-2022

But How Do It Know? x Paraphrasing Tool x Create a desktop x 576+ Free HTML x Python x carlospueg/Fil x Athletes%26h... x Jupyter Notebook x + -

localhost:8888/lab/tree/Downloads/Athletes%26houses.ipynb

File Edit View Run Kernel Tabs Settings Help

Athletes&houses.ipynb x assignment2.ipynb Python 3 (ipykernel)

```
[ ]:
```

```
[62]: print("Data Sheet used for the study of Olympic Athletes")
athlete
```

Data Sheet used for the study of Olympic Athletes

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NaN
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	NaN
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	NaN
3	4	Edgar Lindénau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 500 metres	NaN
...
271111	135569	Andrzej ya	M	29.0	179.0	89.0	Poland-1	POL	1976 Winter	1976	Winter	Innsbruck	Luge	Luge Mixed (Men)'s Doubles	NaN
271112	135570	Piotr ya	M	27.0	176.0	59.0	Poland	POL	2014 Winter	2014	Winter	Sochi	Ski Jumping	Ski Jumping Men's Large Hill, Individual	NaN
271113	135570	Piotr ya	M	27.0	176.0	59.0	Poland	POL	2014 Winter	2014	Winter	Sochi	Ski Jumping	Ski Jumping Men's Large Hill, Team	NaN
271114	135571	Tomasz Ireneusz ya	M	30.0	185.0	96.0	Poland	POL	1998 Winter	1998	Winter	Nagano	Bobsleigh	Bobsleigh Men's Four	NaN
271115	135571	Tomasz Ireneusz ya	M	34.0	185.0	96.0	Poland	POL	2002 Winter	2002	Winter	Salt Lake City	Bobsleigh	Bobsleigh Men's Four	NaN

271116 rows x 15 columns

```
[63]: print("Deriving the Statistical Parameters")
```

Simple 1 2 Python 3 (ipykernel) | Idle Mode: Command Ln 1, Col 1 Athletes&houses.ipynb 22°C Clear 22:37 26-10-2022

But How Do It Know? x Paraphrasing Tool x Create a desktop x 576+ Free HTML x Python x carlospueg/Fil x Athletes%26h... x Jupyter Notebook x + - x

localhost:8888/lab/tree/Downloads/Athletes%26houses.ipynb

Internet Captive Portal... GitHub - the-akira/Co... Computer Science Res... ossu/computer-scienc... The Missing Semester ... Float added -acmpes... Silicon_Rush/Lullu_Edu... My Knowledge Wiki |... Your Career in Web D...

File Edit View Run Kernel Tabs Settings Help

Athletes&houses.ipynb x assignment2.ipynb x +

Python 3 (ipykernel)

```
[63]: print("Deriving the Statistical Parameters")
athlete.describe()

Deriving the Statistical Parameters

[63]:
```

	ID	Age	Height	Weight	Year
count	271116.000000	261642.000000	210945.000000	208241.000000	271116.000000
mean	68248.954396	25.556898	175.338970	70.702393	1978.378480
std	39022.286345	6.393561	10.518462	14.348020	29.877632
min	1.000000	10.000000	127.000000	25.000000	1896.000000
25%	34643.000000	21.000000	168.000000	60.000000	1960.000000
50%	68205.000000	24.000000	175.000000	70.000000	1988.000000
75%	102097.250000	28.000000	183.000000	79.000000	2002.000000
max	135571.000000	97.000000	226.000000	214.000000	2016.000000

```
[64]: zvalue = (190 - athlete["Height"].mean())/athlete["Height"].std()
zvalue
zprob = scipy.stats.norm.cdf(zvalue)
print("z value for height above 190cm:",zvalue)
print("probability that the height of the player will be above than 190cm: ",(1-zprob))

z value for height above 190cm: 1.3938377888285878
probability that the height of the player will be above than 190cm: 0.08158329699239463

[65]: print("Athlete mean: ",athlete["Height"].mean())
print("Athlete SD:",athlete["Height"].std())

Athlete mean: 175.33896987366376
```

Simple 1 2 Python 3 (ipykernel) | Idle Mode: Command Ln 1, Col 1 Athletes&houses.ipynb 22°C Clear 26-10-2022

But How Do It Know? x Paraphrasing Tool x Create a desktop x 576+ Free HTML x Python x carlospueg/Fil x Athletes%26h... x Jupyter Notebook x + - x

localhost:8888/lab/tree/Downloads/Athletes%26houses.ipynb

Internet Captive Portal... GitHub - the-akira/Co... Computer Science Res... ossu/computer-scienc... The Missing Semester ... Float added -acmpes... Silicon_Rush/Lullu_Edu... My Knowledge Wiki |... Your Career in Web D...

File Edit View Run Kernel Tabs Settings Help

Athletes&houses.ipynb x assignment2.ipynb x +

Python 3 (ipykernel)

```
Athlete mean: 175.33896987366376
Athlete SD: 10.51846222679224

[66]: zvalue2 = (170 - athlete["Height"].mean())/athlete["Height"].std()
zprob2 = scipy.stats.norm.cdf(zvalue2)
print("z value for height less than 170cm:",zvalue2)
print("probability that the height of the player will be less than 170cm: ",zprob2)

z value for height less than 170cm: -0.5075808383997632
probability that the height of the player will be less than 170cm: 0.3058736657039933

[67]: zvalue3 = (195 - athlete["Height"].mean())/athlete["Height"].std()
zprob3 = scipy.stats.norm.cdf(zvalue3)
print("z value for height less than 195cm:",zvalue3)
print("probability that the height of the player will be less than 195cm: ",zprob3)

z value for height less than 195cm: 1.8691924456356757
probability that the height of the player will be less than 195cm: 0.9692081977544262

[68]: print("The probability that the height of the player will be between 170 and 195 centimeter: ",zprob3-zprob2)

The probability that the height of the player will be between 170 and 195 centimeter: 0.6633283118402686
```

Simple 1 2 Python 3 (ipykernel) | Idle Mode: Command Ln 1, Col 1 Athletes&houses.ipynb 22°C Clear 26-10-2022