ASSIGNMENT - 2
DATA VISUALIZATION

- Rohith Reddy Vangala (016762109)

Question :

- On the web, find examples of Three different datasets from web.
- For each dataset:

a. familiarize yourself with the data
b. Describe the dataset:
    i. Format
    ii. Number of Records
    iii. Columns
    iv. Anything else of interest
c. Describe the data wrangling.
d. Describe how you would display the data
    i. Type of chart
    ii. Data transformations that would be required

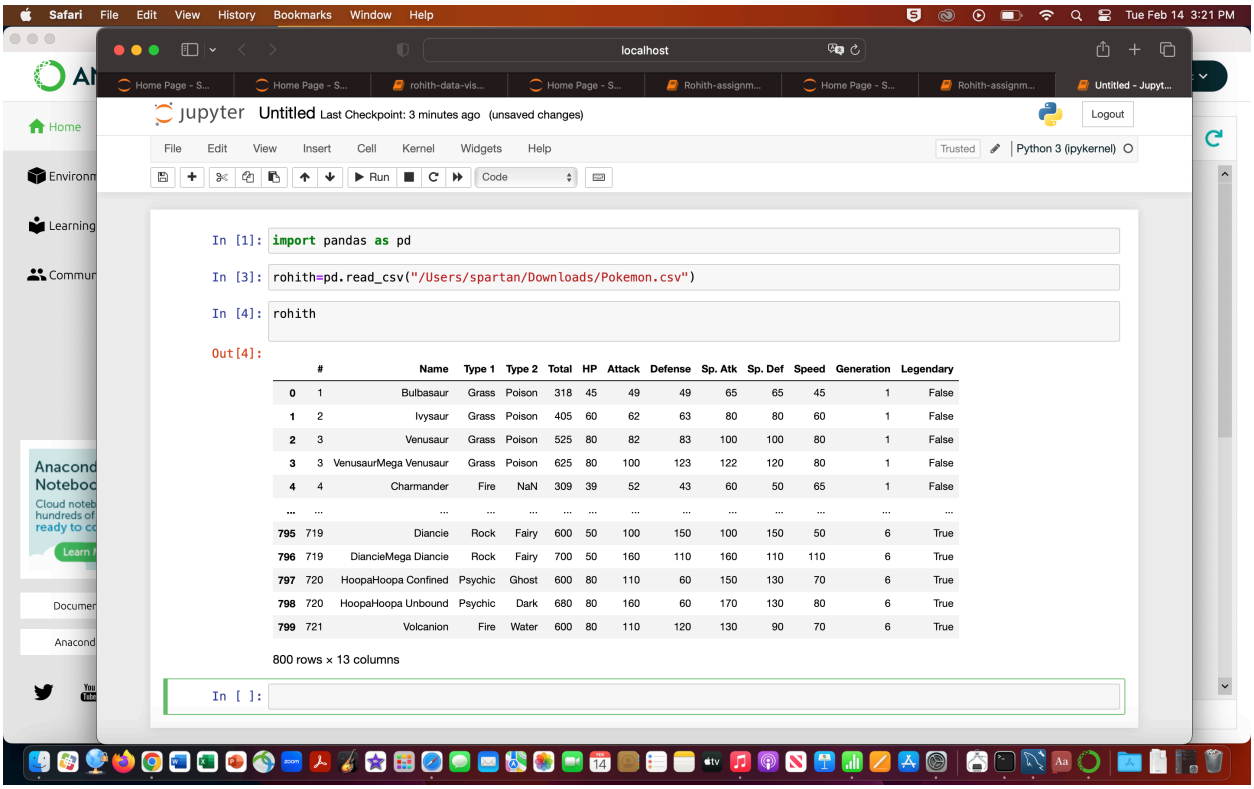
Answer :

Dataset -1 : Pokemon with stats
Link : https://www.kaggle.com/datasets/abcsds/pokemon

- The dataset contains information about Pokemon, including their type, abilities, and base stats. Here I have used pandas and Jupyter notebook to display the data and check the efficiency of the data.

- **Describing the dataset :**

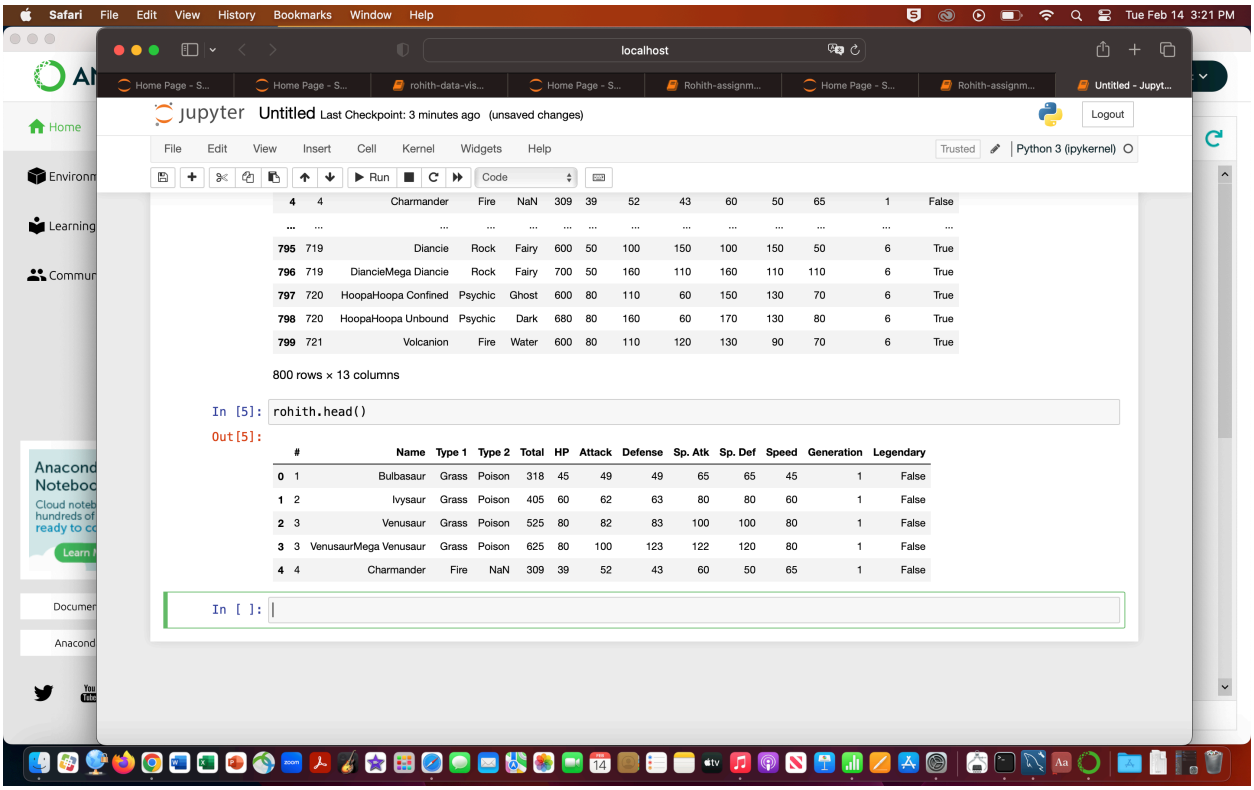
    i. CSV format
    ii. 800 records
    iii. Columns include: #, Name, Type 1, Type 2, Total, HP, Attack, Defense, Sp. Atk, Sp. Def, Speed, Generation, Legendary
    iv. The dataset includes information on different types of Pokemon with various abilities and base stats.



	#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
0	1	Bulbasaur	Grass	Poison	318	45	49	49	65	65	45	1	False
1	2	Ivysaur	Grass	Poison	405	60	62	63	80	80	60	1	False
2	3	Venusaur	Grass	Poison	525	80	82	83	100	100	80	1	False
3	3	VenusaurMega Venusaur	Grass	Poison	625	80	100	123	122	120	80	1	False
4	4	Charmander	Fire	NaN	309	39	52	43	60	50	65	1	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...
795	719	Diancie	Rock	Fairy	600	50	100	150	100	150	50	6	True
796	719	DiancieMega Diancie	Rock	Fairy	700	50	160	110	160	110	110	6	True
797	720	HoopaHoopa Confined	Psychic	Ghost	600	80	110	60	150	130	70	6	True
798	720	HoopaHoopa Unbound	Psychic	Dark	680	80	160	60	170	130	80	6	True
799	721	Volcanion	Fire	Water	600	80	110	120	130	90	70	6	True

800 rows × 13 columns

In [ ]:



4	4	Charmander	Fire	NaN	309	39	52	43	60	50	65	1	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...
795	719	Diancie	Rock	Fairy	600	50	100	150	100	150	50	6	True
796	719	DiancieMega Diancie	Rock	Fairy	700	50	160	110	160	110	110	6	True
797	720	HoopaHoopa Confined	Psychic	Ghost	600	80	110	60	150	130	70	6	True
798	720	HoopaHoopa Unbound	Psychic	Dark	680	80	160	60	170	130	80	6	True
799	721	Volcanion	Fire	Water	600	80	110	120	130	90	70	6	True

800 rows × 13 columns

	#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
0	1	Bulbasaur	Grass	Poison	318	45	49	49	65	65	45	1	False
1	2	Ivysaur	Grass	Poison	405	60	62	63	80	80	60	1	False
2	3	Venusaur	Grass	Poison	525	80	82	83	100	100	80	1	False
3	3	VenusaurMega Venusaur	Grass	Poison	625	80	100	123	122	120	80	1	False
4	4	Charmander	Fire	NaN	309	39	52	43	60	50	65	1	False

In [ ]:

- **Data Wrangling :**

The dataset required cleaning and preparation, which included dealing with missing data. For example, the Type 2 column had many missing values that were filled in with "None." Not just that, but also here there are four fields: attack, defense, special attack, and special defense, which can be made into two fields using average.

- **Display Suggestion :**

i. A scatter plot could be used to display the relationship between Attack and Defense. A bar chart could be used to display the distribution of Pokemon by Type 1.
Ii. Data transformations such as Data Filtering, Data Cleaning are required to ensure that variables are on the same scale and to get the perfect data.

Dataset -2 : StackOverflow Annual Developer Survey 2020
Link : https://www.kaggle.com/datasets/aitzaz/stack-overflow-developer-survey-2020

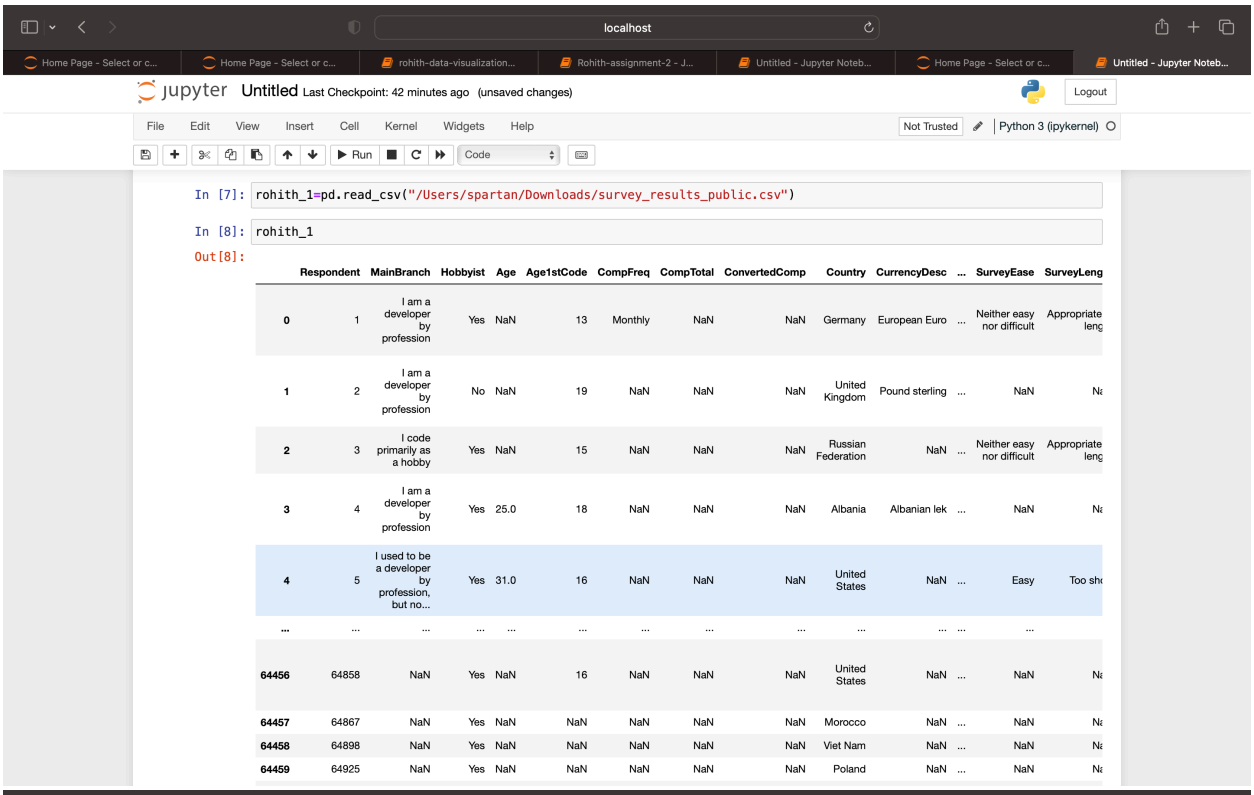
- The dataset contains information about developers, including their demographics, job roles, and technologies they use.. Here I have used pandas and Jupyter notebook to display the data and check the efficiency of the data.

- **Describing the dataset :**

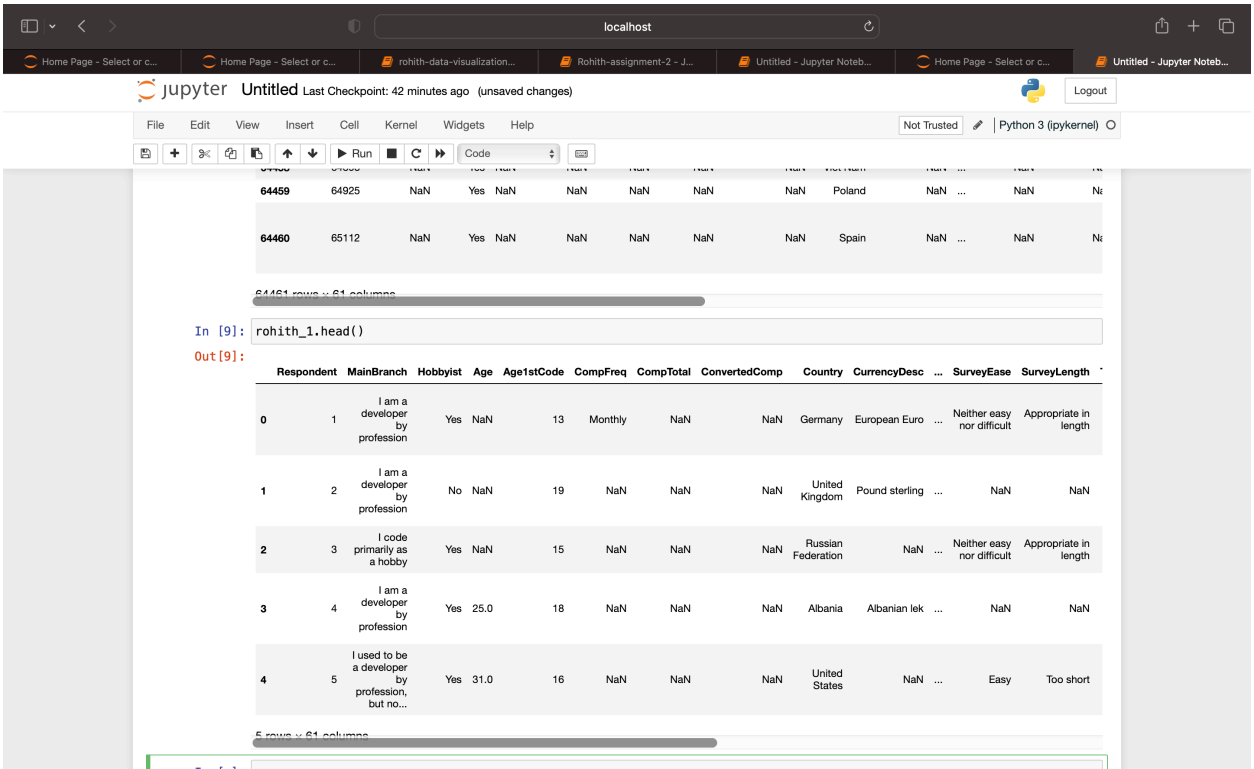
  i. CSV format
  ii. 64,461 records
  iii. Columns include: Respondent, MainBranch, Hobbyist, Age, Age1stCode, CompFreq, CompTotal, ConvertedComp, Country, CurrencyDesc, CurrencySymbol, DatabaseDesireNextYear, DatabaseWorkedWith, DevType, EdLevel, Employment, Ethnicity, Gender, JobFactors, JobSat, LanguageDesireNextYear, LanguageWorkedWith, MiscTechDesireNextYear, MiscTechWorkedWith, NEWCollabToolsDesireNextYear, NEWCollabToolsWorkedWith, NEWDevOps, NEWJobHunt, NEWLearn, NEWOffTopic, NEWOnboardGood, NEWOtherComms, NEWOvertime,

NEWPurchaseResearch, NEWPurpleLink, NEWSOSites, NEWStuck, OpSys, OrgSize, PlatformDesireNextYear, PlatformWorkedWith, PurchaseWhat, Sexuality, SOAccount, SOComm, SOPartFreq, SOVisitFreq, SurveyEase, SurveyLength, Trans, UndergradMajor, WebframeDesireNextYear, WebframeWorkedWith, WelcomeChange, WorkChallenge, WorkLoc, WorkPlan, WorkRemote, WorkWeekHrs

iv. The dataset includes information on developers from various countries, with different job roles and technologies used.



	Respondent	MainBranch	Hobbyist	Age	Age1stCode	CompFreq	CompTotal	ConvertedComp	Country	CurrencyDesc	...	SurveyEase	SurveyLeng
0	1	I am a developer by profession	Yes	NaN	13	Monthly	NaN	NaN	Germany	European Euro	...	Neither easy nor difficult	Appropriate leng
1	2	I am a developer by profession	No	NaN	19	NaN	NaN	NaN	United Kingdom	Pound sterling	...	NaN	Na
2	3	I code primarily as a hobby	Yes	NaN	15	NaN	NaN	NaN	Russian Federation	NaN	...	Neither easy nor difficult	Appropriate leng
3	4	I am a developer by profession	Yes	25.0	18	NaN	NaN	NaN	Albania	Albanian lek	...	NaN	Na
4	5	I used to be a developer by profession, but no...	Yes	31.0	16	NaN	NaN	NaN	United States	NaN	...	Easy	Too sh
...	...	...	...	...	...	...	...	...	...	...	...	...	
64456	64858	NaN	Yes	NaN	16	NaN	NaN	NaN	United States	NaN	...	NaN	Na
64457	64867	NaN	Yes	NaN	NaN	NaN	NaN	NaN	Morocco	NaN	...	NaN	Na
64458	64898	NaN	Yes	NaN	NaN	NaN	NaN	NaN	Viet Nam	NaN	...	NaN	Na
64459	64925	NaN	Yes	NaN	NaN	NaN	NaN	NaN	Poland	NaN	...	NaN	Na



64459	64925	NaN	Yes	NaN	NaN	NaN	NaN	NaN	Poland	NaN	...	NaN	Na
64460	65112	NaN	Yes	NaN	NaN	NaN	NaN	NaN	Spain	NaN	...	NaN	Na

64461 rows × 61 columns

	Respondent	MainBranch	Hobbyist	Age	Age1stCode	CompFreq	CompTotal	ConvertedComp	Country	CurrencyDesc	...	SurveyEase	SurveyLength
0	1	I am a developer by profession	Yes	NaN	13	Monthly	NaN	NaN	Germany	European Euro	...	Neither easy nor difficult	Appropriate in length
1	2	I am a developer by profession	No	NaN	19	NaN	NaN	NaN	United Kingdom	Pound sterling	...	NaN	NaN
2	3	I code primarily as a hobby	Yes	NaN	15	NaN	NaN	NaN	Russian Federation	NaN	...	Neither easy nor difficult	Appropriate in length
3	4	I am a developer by profession	Yes	25.0	18	NaN	NaN	NaN	Albania	Albanian lek	...	NaN	NaN
4	5	I used to be a developer by profession, but no...	Yes	31.0	16	NaN	NaN	NaN	United States	NaN	...	Easy	Too short

5 rows × 61 columns

- **Data Wrangling :**

The dataset required cleaning and preparation, which included dealing with missing data and also dropping the data which is having Nan, because there are 64,461 records . After dropping the useless data we can still have maximum data for any purpose to use, also we can remove the years code pro because we can know it by the years code field. Additionally, the Age1stCode column had some values that were not numeric and requires conversion.

- **Display Suggestion :**

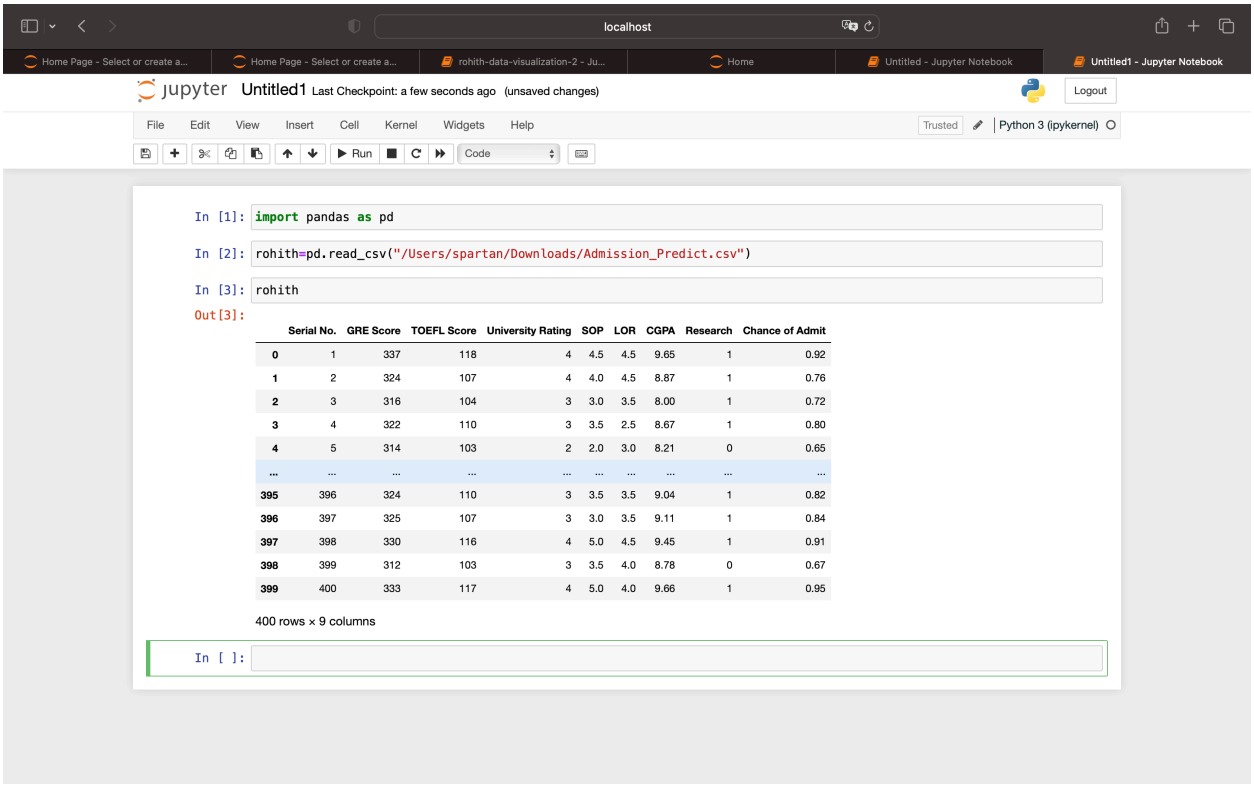
i. A bar chart could be used to display the distribution of developers by country. A scatter plot could be used to display the relationship between Years of Experience and Salary.
ii. Data transformations such as Data Filtering, Data Cleaning are required to ensure that variables are on the same scale and to get the perfect data.

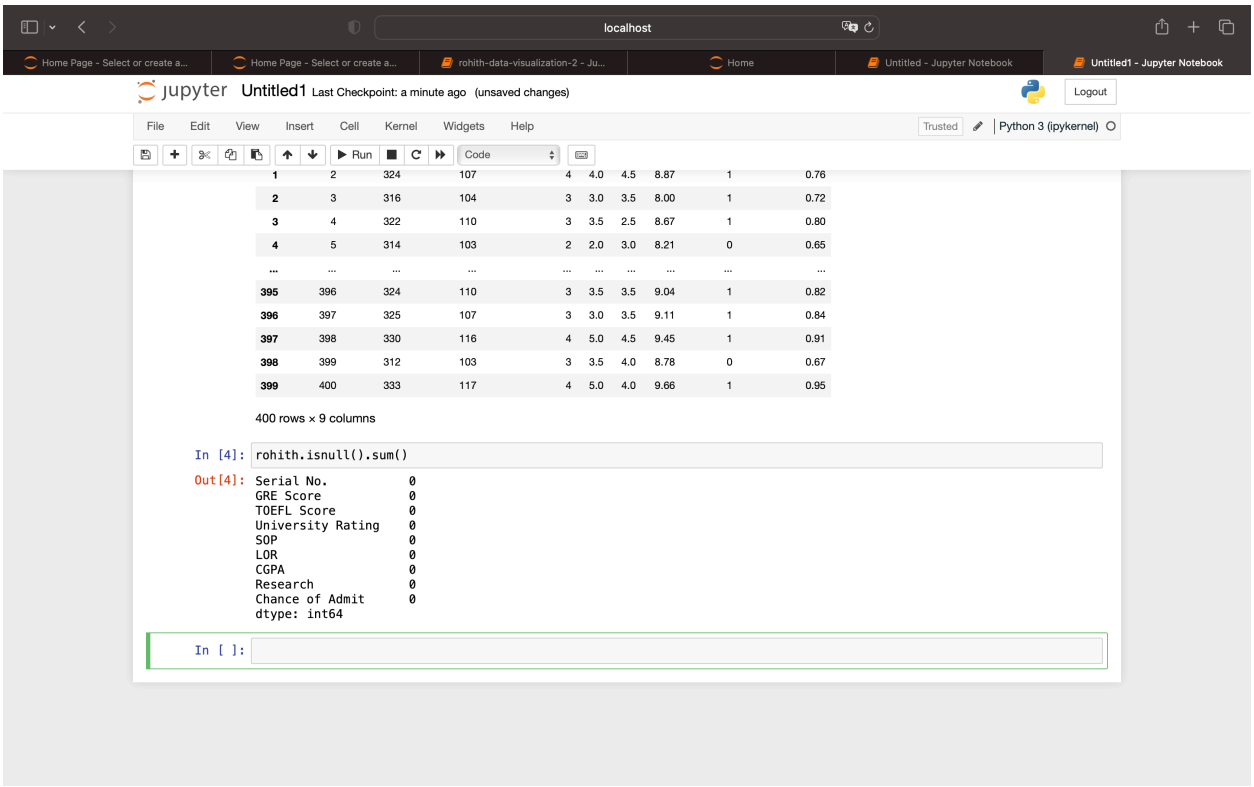
Dataset -3 : Graduate Admissions
Link : https://www.kaggle.com/datasets/mohansacharya/graduate-admissions

- The dataset contains data on graduate school admissions, including the applicant's GRE score, TOEFL score, university rating, statement of purpose, letter of recommendation strength, undergraduate GPA, and admission decision.
- Here I have used pandas and Jupyter notebook to display the data and check the efficiency of the data.

- **Describing the dataset :**

i. CSV format
ii. 400 records
iii. Columns include: GRE Score, TOEFL Score, University Rating, SOP, LOR, CGPA, Research, and Chance of Admit.
iv. The dataset provides insight into the factors that influence graduate school admissions decisions.


localhost
Home Page - Select or create a...
Home Page - Select or create a...
rohith-data-visualization-2 - Ju...
Home
Untitled - Jupyter Notebook
Untitled1 - Jupyter Notebook
jupyter Untitled1 Last Checkpoint: a few seconds ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
In [1]: import pandas as pd
In [2]: rohith=pd.read_csv("/Users/spartan/Downloads/Admission_Predict.csv")
In [3]: rohith
Out[3]:
Serial No. GRE Score TOEFL Score University Rating SOP LOR CGPA Research Chance of Admit
0 1 337 118 4 4.5 4.5 9.65 1 0.92
1 2 324 107 4 4.0 4.5 8.87 1 0.76
2 3 316 104 3 3.0 3.5 8.00 1 0.72
3 4 322 110 3 3.5 2.5 8.67 1 0.80
4 5 314 103 2 2.0 3.0 8.21 0 0.65
... ...
395 396 324 110 3 3.5 3.5 9.04 1 0.82
396 397 325 107 3 3.0 3.5 9.11 1 0.84
397 398 330 116 4 5.0 4.5 9.45 1 0.91
398 399 312 103 3 3.5 4.0 8.78 0 0.67
399 400 333 117 4 5.0 4.0 9.66 1 0.95
400 rows × 9 columns
In [ ]:


localhost
Home Page - Select or create a...
Home Page - Select or create a...
rohith-data-visualization-2 - Ju...
Home
Untitled - Jupyter Notebook
Untitled1 - Jupyter Notebook
jupyter Untitled1 Last Checkpoint: a minute ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
1 2 324 107 4 4.0 4.5 8.87 1 0.76
2 3 316 104 3 3.0 3.5 8.00 1 0.72
3 4 322 110 3 3.5 2.5 8.67 1 0.80
4 5 314 103 2 2.0 3.0 8.21 0 0.65
... ...
395 396 324 110 3 3.5 3.5 9.04 1 0.82
396 397 325 107 3 3.0 3.5 9.11 1 0.84
397 398 330 116 4 5.0 4.5 9.45 1 0.91
398 399 312 103 3 3.5 4.0 8.78 0 0.67
399 400 333 117 4 5.0 4.0 9.66 1 0.95
400 rows × 9 columns
In [4]: rohith.isnull().sum()
Out[4]: Serial No. 0
GRE Score 0
TOEFL Score 0
University Rating 0
SOP 0
LOR 0
CGPA 0
Research 0
Chance of Admit 0
dtype: int64
In [ ]:

- **Data Wrangling :**

The dataset does not require any cleaning or preparation; I have checked the nulls and identifiers. Everything is on point without any duplicates or inefficient data. Not only that, but the fields here are ideal for the requirement and analysis.

- **Display Suggestion :**

i. A scatter plot could be used to display the relationship between GRE scores and admission decisions. A box plot could be used to compare the distribution of CGPA scores between admitted and not admitted applicants.
ii. Data transformation is not required.