DATA ANALYSIS report

Pokémon

Prepared by

Muhammad owais

Muneeba Siddiqui

Zain ul abdeen

Dated:januray2021

AbstracT

Institute :UNIVERSITY OF KARACHI

DATED : JANUARY20 21

DEGREE PROGRAM: BS FINANCIAL MATHAMETICS

NAME OF PROJECT : DATA ANALYSIS USING PANDAS ON CSV FILE

Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data.Data science is related to data mining, machine learning and big data.

Data science is a "concept to unify statistics, data analysis and their related methods" in order to "understand and analyze actual phenomena" with data. It uses techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, domain knowledge and information science.

INTRODUCTION

Data analytics is the investigation of separating unrefined data to settle on choices about the information. A noteworthy number of the strategies and methods of information investigation have been automated into mechanical systems and figurines that work

over unrefined data for human use. Data analytics methods can uncover patterns and measurements that would some way or another be lost in the mass of data. This data would then be able to be utilized to improve procedures to expand the general productivity of a business or framework. Data analytics is a wide term that incorporates numerous various kinds of information examination. Any kind of data can be exposed to information examination methods to get the knowledge that can be utilized to improve things. For instance, fabricating organizations regularly record the runtime, personal time, and work line for different machines and after that investigate the information to even more likely arrangement the remaining tasks at hand, so the machines work nearer to crest limit.

DATA ANALYSIS

Data Analysis is a way of thinking about information from social events and then set it up for major conferences. Information analysts discuss the use of notable methods related to the description and control of information. Every one of these bits of knowledge permits the organizations to define better procedures and to settle on remotely enhanced choices. Data Analysis is characterized as a procedure of cleaning, changing, and displaying data to find valuable data for business basic leadership. The motivation behind Data Analysis is to extricate valuable data from information and getting the preference reliant on the data analysis. Similarly, Data Analysis is a procedure of examining, purging, changing and displaying data with the objective of finding helpful data and establishing basic

leadership. Data analysis has many aspects and methods. They combine different methods under many different names and used in different fields of business, science, and sociology. In today's business world, data analysis is responsible for making progressive logical choices and assistance to organizations work more successfully. (Thumar 2019.)

Data analysis is a procedure of gathering and demonstrating information with the objective of finding the necessary data. The outcomes are conveyed, proposing ends, and supporting basic leadership. Occasionally, data perception depicts the simplicity of finding valuable examples in the data for the data. Data analysis along with Data modelling statements indicate the consequent

Data Science

When the world entered to the time of large information, the requirement for data science stockpiling additionally developed. Data science is an interdisciplinary field that utilizes logical techniques, procedures, calculations and frameworks to remove information and insights from organized and unstructured information. Data science is identified with information mining and big data.

Pandas

In computer programming, for data control and examination in Python programming language a product library composes which is known as Pandas. Specifically, it extends data structure and activities for regulating mathematical tables and time arrangement. It is free programming language released under the three-provision BSD permit. The name is grown from the expression "panel data", an econometrics term for informational collections that incorporate perceptions over various timeframes for similar people. Pandas (Python data analysis) is an indisputable requirement in the Data Science life cycle. It is the most famous and broadly utilized Python library for Data Science, alongside NumPy in matplotlib. With around 17,00 remarks on GitHub and a functioning network of 1,200 benefactors, it is actively utilized for data analysis and cleaning. Pandas give quick, adaptable information structures, for example, information outline CDs, which are intended to work with organized data rapidly and instinctively.

Matplotlib

Matplotlib is a plotting library in which mathematical science reinforcement is NumPy for Python programming language. Matplotlib gives an article arranged API implanting plots through useful toolbox such as Tkinter, wxPython, Qt, or GTK+ are utilizing broadly applications. There is additionally a procedural "pylab" interface dependent on a state machine (like OpenGL), intended to attentively resemble of MATLAB,

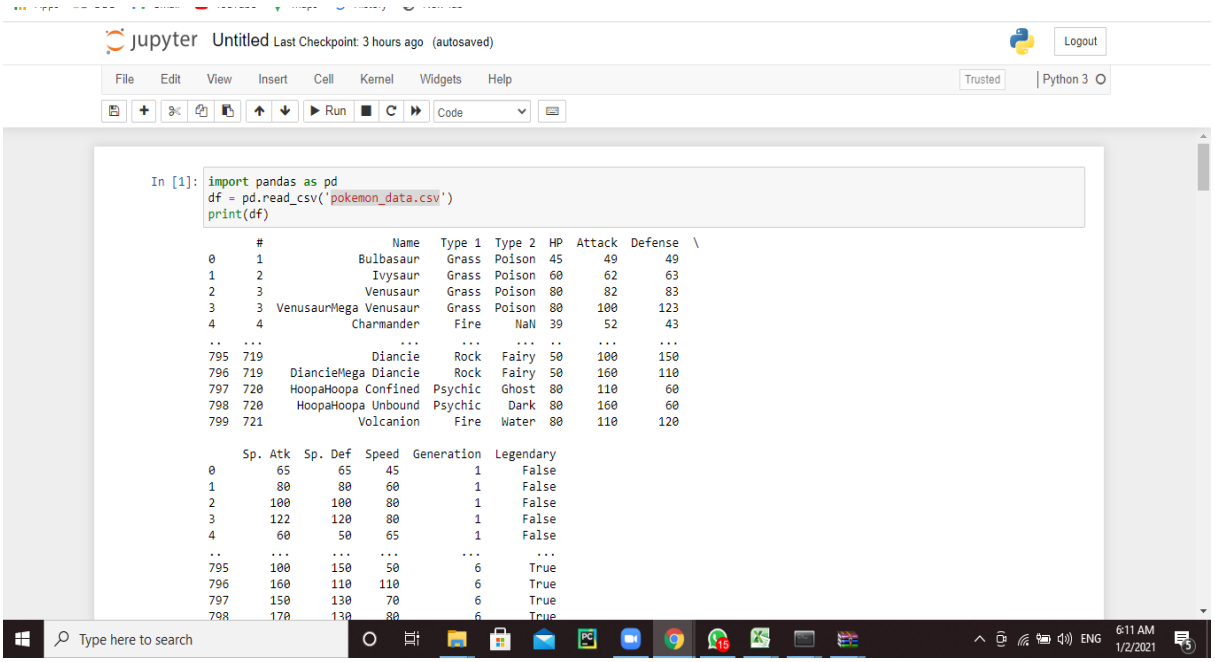
however, its utilization is debilitated. SciPy utilizes Matplotlib. Matplotlib has incredible pleasant perceptions. It is a plotting library for Python with around 26,000 remarks on GitHub and and

exceptionally energetic network of around 700 benefactors. On account of the charts and plots that it delivers, it is widely utilized for data representation. It likewise gives an item situated API, which can be utilized to implant those plots into applications.

Seaborn

Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas data structures. Seaborn helps you explore and understand your data.

Step no 1:-



Importing all libraries

First we import import libraries for data analysis and then read our selected csv file

Step no 2:-

jupyter Untitled Last Checkpoint: 3 hours ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run Code

```
[800 rows x 12 columns]
```

In [2]:
```
#first five row of dataset
df.head()
```

Out[2]:

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

In [3]:
```
#Last five row fo the dataset
df.tail()
```

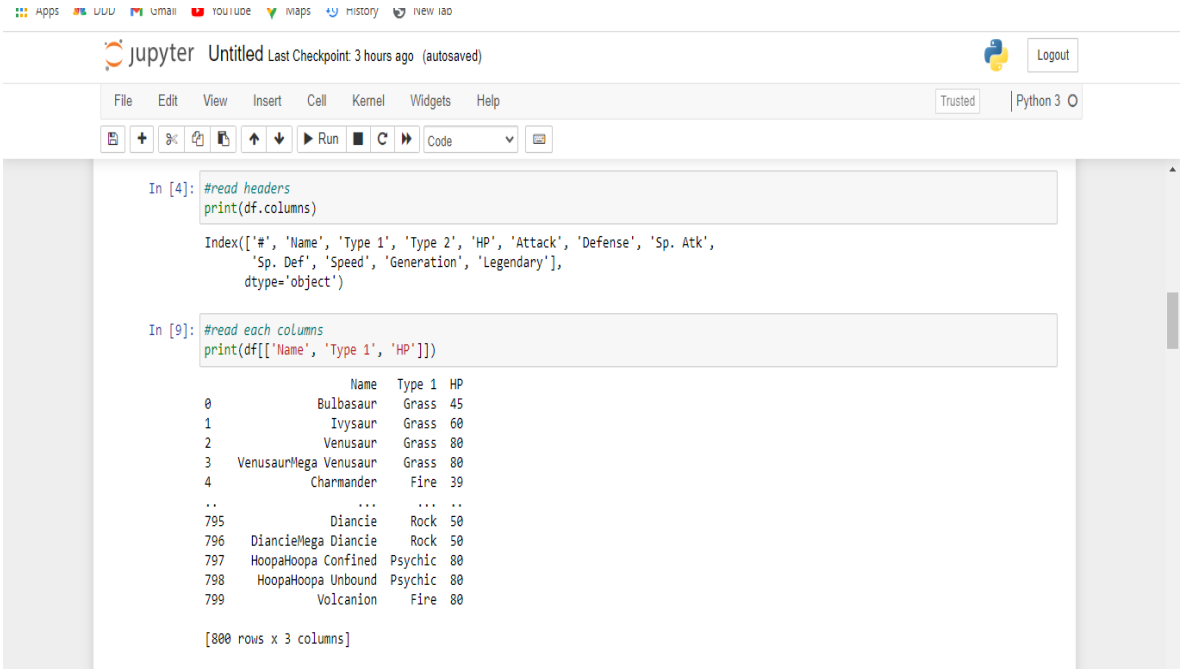
Out[3]:

	#	Name	Type 1	Type 2	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
795	719	Diancie	Rock	Fairy	50	100	150	100	150	50	6	True
796	719	DiancieMega Diancie	Rock	Fairy	50	160	110	160	110	110	6	True
797	720	HoopaHoopa Confined	Psychic	Ghost	80	110	60	150	130	70	6	True
798	720	HoopaHoopa Unbound	Psychic	Dark	80	160	60	170	130	80	6	True
799	721	Volcanion	Fire	Water	80	110	120	130	90	70	6	True

First five and last five rows of data set

* **We use** head for 1 five rows and tail for last five rows from the data set

Step no 3:-



Read header and read each columns

In this step we read all header and those columns which we want to read

Step no 4:-

jupyter Untitled Last Checkpoint: 3 hours ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run Code

In [11]:
```
#read the specific location
print(df.iloc[2,2])
```

```
Grass
```

In [13]:
```
df.loc[df['Type 1'] == "Grass"]
```

Out[13]:

	#	Name	Type 1	Type 2	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
0	1	Bulbasaur	Grass	Poison	45	49	49	65	65	45	1	False
1	2	Ivysaur	Grass	Poison	60	62	63	80	80	60	1	False
2	3	Venusaur	Grass	Poison	80	82	83	100	100	80	1	False
3	3	VenusaurMega Venusaur	Grass	Poison	80	100	123	122	120	80	1	False
48	43	Oddish	Grass	Poison	45	50	55	75	65	30	1	False
...	...	...	...	...	...	...	...	...	...	...	...	...
718	650	Chespin	Grass	NaN	56	61	65	48	45	38	6	False
719	651	Quilladin	Grass	NaN	61	78	95	56	58	57	6	False
720	652	Chesnaught	Grass	Fighting	88	107	122	74	75	64	6	False
740	672	Skiddo	Grass	NaN	66	65	48	62	57	52	6	False
741	673	Gogoat	Grass	NaN	123	100	62	97	81	68	6	False

70 rows × 12 columns

Read specific location

In this step we give a specific location to our csv file .We use command iloc to perform this step.

Step no 5:-

Desktop/pandas-master/ | Untitled - Jupyter Notebook | (538) Complete Python Pandas

localhost:8888/notebooks/Desktop/pandas-master/Untitled.ipynb?kernel_name=python3#

Apps DDD Gmail YouTube Maps History New Tab

jupyter Untitled Last Checkpoint: 3 hours ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help

In [44]:
```
df.to_csv('modified.csv', index=False)
```

In [65]:
```
import re

df.loc[df['Name'].str.contains('^pi[a-z]*', flags=re.I, regex=True)]
```

Out[65]:

	#	Name	Type 1	Type 2	Legendary	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation
20	16	Pidgey	Normal	Flying	False	446	40	45	40	35	35	56	1
21	17	Pidgeotto	Normal	Flying	False	627	63	60	55	50	50	71	1
22	18	Pidgeot	Normal	Flying	False	857	83	80	75	70	70	101	1
23	18	PidgeotMega Pidgeot	Normal	Flying	False	1037	83	80	80	135	80	121	1
30	25	Pikachu	Electric	NaN	False	550	35	55	40	50	50	90	1
136	127	Pinsir	Bug	NaN	False	915	65	125	100	55	70	85	1
137	127	PinsirMega Pinsir	Bug	Flying	False	1095	65	155	120	65	90	105	1
186	172	Pichu	Electric	NaN	False	350	20	40	15	35	35	60	2
219	204	Pineco	Bug	NaN	False	565	50	65	90	35	35	15	2
239	221	Piloswine	Ice	Ground	False	850	100	100	80	60	60	50	2
438	393	Piplup	Water	NaN	False	588	53	51	53	61	56	40	4
558	499	Pignite	Fire	Fighting	False	781	90	93	55	70	55	55	5
578	519	Pidove	Normal	Flying	False	485	50	55	50	36	30	43	5

Type here to search ENG 6:14 AM 1/2/2021

Modified the csv file

Desktop/pandas-master/ | Untitled - Jupyter Notebook | (538) Complete Python Pandas

localhost:8888/notebooks/Desktop/pandas-master/Untitled.ipynb?kernel_name=python3#

Apps DDD Gmail YouTube Maps History New Tab

jupyter Untitled Last Checkpoint: 3 hours ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help

Flying	677.750000	0.500000	867.500000	70.750000	78.750000	66.250000	94.250000	72.500000	102.500000	5.500000
Fighting	363.851852	0.000000	766.814815	69.851852	96.777778	65.925926	53.111111	64.703704	66.074074	3.370370
Fairy	449.529412	0.058824	777.764706	74.117647	61.529412	65.705882	78.529412	84.705882	48.588235	4.117647
Normal	319.173469	0.020408	731.816327	77.275510	73.469388	59.846939	55.816327	63.724490	71.551020	3.051020

In [74]:
```
df = pd.read_csv('modified.csv')
df['count'] = 1
df.groupby(['Type 1', 'Type 2']).count()['count']
```

Out[74]:
```
Type 1  Type 2
Bug     Electric     2
        Fighting     2
        Fire         2
        Flying      14
        Ghost        1
                    ..
Water   Ice          3
        Poison       3
        Psychic      5
        Rock         4
        Steel        1
Name: count, Length: 136, dtype: int64
```

In [ ]:

In [ ]:

Type here to search ENG 6:15 AM 1/2/2021

Step NO: 06

Modified CSV file group by

In [23]:
```
df =pd.read_csv('modified.csv')
df.groupby(['Type 1']).mean().sort_values('Defense', ascending=False)
```

Out[23]:

Type 1	#	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
Steel	442.851852	65.222222	92.703704	126.370370	67.518519	80.629630	55.259259	3.851852	0.148148
Rock	392.727273	65.363636	92.863636	100.795455	63.340909	75.477273	55.909091	3.454545	0.090909
Dragon	474.375000	83.312500	112.125000	86.375000	96.843750	88.843750	83.031250	3.875000	0.375000
Ground	356.281250	73.781250	95.750000	84.843750	56.468750	62.750000	63.906250	3.156250	0.125000
Ghost	486.500000	64.437500	73.781250	81.187500	79.343750	76.468750	64.343750	4.187500	0.062500
Water	303.089286	72.062500	74.151786	72.946429	74.812500	70.517857	65.964286	2.857143	0.035714
Ice	423.541667	72.000000	72.750000	71.416667	77.541667	76.291667	63.458333	3.541667	0.083333
Grass	344.871429	67.271429	73.214286	70.800000	77.500000	70.428571	61.928571	3.357143	0.042857
Bug	334.492754	56.884058	70.971014	70.724638	53.869565	64.797101	61.681159	3.217391	0.000000
Dark	461.354839	66.806452	88.387097	70.225806	74.645161	69.516129	76.161290	4.032258	0.064516
Poison	251.785714	67.250000	74.678571	68.821429	60.428571	64.392857	63.571429	2.535714	0.000000
Fire	327.403846	69.903846	84.769231	67.769231	88.980769	72.211538	74.442308	3.211538	0.096154
Psychic	380.807018	70.631579	71.456140	67.684211	98.403509	86.280702	81.491228	3.385965	0.245614
Electric	363.500000	59.795455	69.090909	66.295455	90.022727	73.704545	84.500000	3.272727	0.090909
Flying	677.750000	70.750000	78.750000	66.250000	94.250000	72.500000	102.500000	5.500000	0.500000
Fighting	363.851852	69.851852	96.777778	65.925926	53.111111	64.703704	66.074074	3.370370	0.000000
Fairy	449.529412	74.117647	61.529412	65.705882	78.529412	84.705882	48.588235	4.117647	0.058824
Normal	319.173469	77.275510	73.469388	59.846939	55.816327	63.724490	71.551020	3.051020	0.020408

Step NO:06

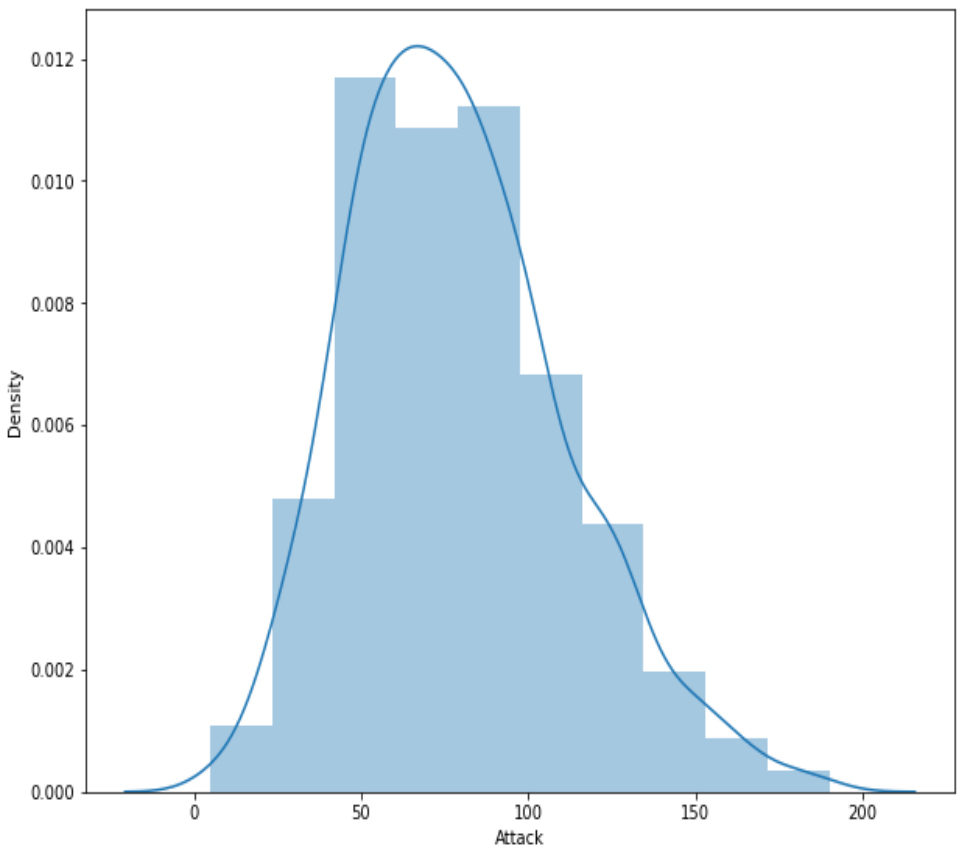
New Filtered CSV file

In [24]:
```
new_df = df.loc[(df['Type 1'] == 'Grass') & (df['Type 2'] =='poison') & (df['HP'] > 70)]

new_df.reset_index(drop=True, inplace=True)

new_df

new_df.to_csv('filtered.csv')
```

Step No:07

In [29]:
```
df = pd.read_csv('modified.csv')
f, ax = plt.subplots(figsize=(10,8))
x = df['Attack']
ax = sns.distplot(x, bins=10)
plt.show()
```

```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\distributions.py:2551: FutureWarning: `distplot` i
s 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 fo
r histograms).
  warnings.warn(msg, FutureWarning)
```



Step No: 08

In [40]:
```
fig, axs = plt.subplots(2, 2, figsize = (12,12))

ax1 = plt.subplot2grid((8,8), (0,0), rowspan=3, colspan=3)
ax2 = plt.subplot2grid((8,8), (4,0), rowspan=3, colspan=3)
ax3 = plt.subplot2grid((8,8), (0, 4), rowspan=3, colspan=3)
ax4 = plt.subplot2grid((8,8), (4, 4), rowspan=3, colspan=3)

fig.tight_layout()

ax1.set_title("Plot1: HP and Attack", fontsize =18)
ax2.set_title("Plot2: HP and Attack", fontsize =18)
ax3.set_title("Plot3: HP and Attack", fontsize =18)
ax4.set_title("Plot4: HP and Attack", fontsize =18)

# Plot 1
sns.regplot(x='HP', y='Attack',
            data=df, ax=ax1)

# Plot 2
sns.regplot(x='HP', y='Attack',
            data=df, fit_reg = False, color = 'green', marker ="^", ax=ax2)

# Plot 3
sns.regplot(x='HP', y='Attack',
            data=df, fit_reg = True, x_bins = 6, color = 'orange', ax=ax3)

# Plot 4
sns.regplot(x='HP', y='Attack',
            data=df, fit_reg = False, x_bins = 12, ci = 99, color = 'red', ax=ax4)

plt.show()
```