Group number: 19
Student1_id: 24280045
Student2_id: 24280063

Contribution:
We divided our work for fetching data as in Reddit was done by me, kaggle was done by Ahsan and third one we both came up with options and finalized the script together.

Overview of topic:
We chose **Healthcare**, given the amount of data we could get was of critical nature and in this progressing period of technology. There is still some lags we are experiencing using the field of AI when we know machine only learns based on the data provided by us and how impactful this can be, provided the severity of this field as we're dealing with patients, quality of care and medications.
The expectation from the data was to analyze Patient data, what medications they're taking, prognostics of their disease, from stock data what medications is top seller based on their stocks, insurance claims, and from last one reddit posts how people are reacting to movements related to healthcare, what are their perspectives, and all.

Data Collection Process:

- For Kaggle it was fetched using API setup, which started with downloading their library using "pip" statement, and then provided the dataset name we wanted to download. The challenge faced was API was not authenticating from our local machine which further created problems as we did not have any URL/API key to provide in the setup. Then, after troubleshooting we figured we just needed to download the API.json setup in our local machine and VIOLA!!! It worked.
- For second, We used the smae approach as python provided a library to download stock trends named "yfinance" and then we started looking for stocks, which were owned by healthcare companies like "Pfizer" and others. Rest of the process was fairly simple. We didn't run into any challenge per se for this one.
- Last but not least Reddit was fairly challenging as we had to create an "app" in teh admin console of Reddit and after that for fetching posts we had to use script option as it was mentioned explicitly to use this option we then provided the redirect URL as in "port:8080" and then the name we signed into reddit with. The main challenge was to send request to reddit for fetching posts but it was not working as we were using password body type to fetch data, when we signed into to reddit using google account which was done without creating password and just OAuth Authentication was done, to address this problem we then changed our approach to not use password field and we were able to fetch post based on some healthcare provided keywords.

Initial Observations:

```
Public Dataset

                Age  Billing Amount   Room Number
count  55500.000000    55500.000000  55500.000000
mean      51.539459    25539.316097    301.134829
std       19.602454    14211.454431    115.243069
min       13.000000    -2008.492140    101.000000
25%       35.000000    13241.224652    202.000000
50%       52.000000    25538.069376    302.000000
75%       68.000000    37820.508436    401.000000
max       89.000000    52764.276736    500.000000
Healthcare Stock Data

             Open        High         Low       Close        Volume   Dividends  Stock Splits
count  630.000000  630.000000  630.000000  630.000000  6.300000e+02  630.000000         630.0
mean   170.624371  172.473825  168.701388  170.477668  1.381975e+07    0.014063           0.0
std    201.746130  203.719030  199.675522  201.632037  1.491583e+07    0.143853           0.0
min     24.298299   24.691793   24.081876   24.396671  1.581900e+06    0.000000           0.0
25%     40.948750   42.435001   39.490999   40.829999  4.724700e+06    0.000000           0.0
50%     58.370001   59.340837   57.669056   58.364288  8.044500e+06    0.000000           0.0
75%    160.844403  161.936391  159.693642  160.603851  1.630472e+07    0.000000           0.0
max    617.212685  628.320065  611.574296  622.861023  1.310744e+08    2.100000           0.0
Reddit Healthcare posts
               Score
count     200.000000
mean     5714.490000
std     14538.920929
min         0.000000
25%         3.000000
50%        27.000000
75%      5161.750000
max    136537.000000
```

AI-powered system:
Based on this data I fetched I can create AI-healthcare based Analytics platform, providing predictive insights, cost analysis, Medical Fraud Detection system, healthcare sentiment analysis using NLP, Disease prediction giving better prognosis, and many more.

Legal Obligation while Fetching data:
We can face many legal obligations as well as technical ones as well, given how they've configured their system against DDoS attacks, using rate-limiting to not exhaust their systems when large query is ran, which can lead to IP ban or API ket revocation. For instance Reddit users own their content, meaning you cannot redistribute, or sell their info, and collecting sensitive posts could breach privacy policies. Then in case of Finance Yahoo, the data is available to public but misuse may breach copyright, and automated scrapping from many sites now a days can detect bot behvior and blacklist that IP.

Data Quality Issues:

The main issue would be the acquisition of data from multiple sources can result in diverse datasets, conflicting the commonality which won't exist as data storage is not a standard protocol and data can be in structured as well as unstructured format. Then comes the step to process the data for analysis purpose which is crucial as this step will define the course of data mdoling in Macine Learning, based on which predictions can be made and any inconsistency in data can result in false positives or faulty predictions.

Ways to store this type of data:
Like discussed in class we can store data on multiple platforms, based on data format we have, For instance for unstructured data we have an option provided by AWS S3 bucket concept, for structured and if it's relatable data like hospital managment info then we can store in PostgreSQL, MySQL. Last but not least in case we have not relational data like Reddit texts, we can store them in MongoDB, Cassandra.

Visualizations for each dataset:
We've added a file named Visualizations.py which can be viewed in repository.

Repository link: https://github.com/saadmuzammil098/Project-Mosaic