CSE 5544
Homework 1
Page Zhu.1786

WriteUp:

D3:

D3, Data Driven Documents, is a JavaScript library that allows for manipulation of Document Object Models using data. D3 is based on JavaScript which is mainly used on web pages using CSS, HTML and SVG. D3 supports a lot of data types as long as it is supported by JavaScript and web, such as Integer, Boolean, Float and String. Also, D3 is capable of manipulating objects as array and customized data structure.

D3 can creates multiple different kinds of visualizations such as charts, plots and diagrams. Maps and even animations are doable as well. API of D3 is considerable that supports many types of data files. D3 is very customizable because it is a JavaScript library.

ParaView:

ParaView is an application for data analysis and visualization which does not need coding to manipulate data within its own interface. ParaView provides multiple different options to achieve and manipulate data visualizations, such as scaling and even animation control.

After reading the tutorial that is packed within ParaView, I found out that this application is mainly focusing on visualizing the data with a 3D view which means it is mainly for data types like coordinates and meshes. Though the tool of ParaView is very strong and comprehensive, it is not customizable when facing some manipulations which are originally not supported.

Python:

Python is a programming language with numeric libraries, tools and API. With the tools developed by different companies, groups or people, python can be used as a data visualization tool to manipulate data. Data type supported by python tools are mainly defined by the libraries.

Supported visualization types of Python can be basically every kind of 2D graphs. Besides that, python can also be capable of customization to fulfill some specific use since it is a programming language.

Tableau:

Tableau is an application dedicated to help big company view and understand data. Tableau helps anyone quickly analyze, visualize, and share information. This is a strong application that needs no coding as well as ParaView. Also, it supports almost every kind of business data type such as xlsx, csv and similar data types.

The Visualization tools are also powerful, it can achieve and manipulate almost every kind of 2D data visualization, such as bar charts, glyphs and volume rendering. However, its tools are also built in like ParaView, so it can sometimes be not capable of customization.

Visualization 1 Report:

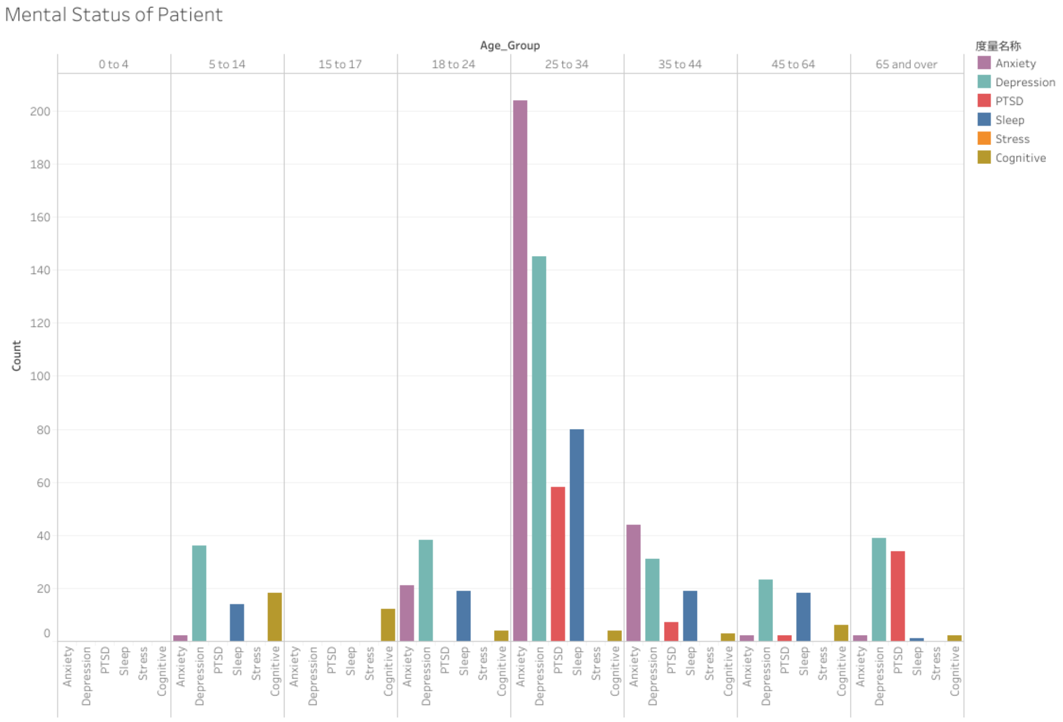
Data Using: EHRdataSample.xlsx

Linked Question: What is patient count for each age group and what are their mental status portion?

Tool using: Tableau Public

For this question, I first tried to use the D3 to visualize the graph. I spent nearly 2 days on finding out how to load the CSV file, dealing with problems like CORS. However, at the end, the CSV file saved from the original sample file EHRdataSample.xlsx was still not read successfully. Then I decided to try ParaView, due to some unsolved reason, loading CSV file was failed in ParaView.

At the end, I used Tableau to fulfill this visualization by making a barchart with different bars and colors. In order to make it interesting and useful, I changed the question to be "What is the Count of patients with Mental Issue in different age group?"



This is the preview of the visualization. As we can see, I used different bar to represent the count of different mental issue reported in certain age group. As the tableau supported, we can emphasize certain issue by hiding the other bars by click on certain bars. In this way, we can clearly see the differences in different age group and make assumption on correlated issue on specific age groups (birth year).

The original graph url:

https://public.tableau.com/profile/.53223230#!/vizhome/HomeWork1_15799933904040/1?publish=yes

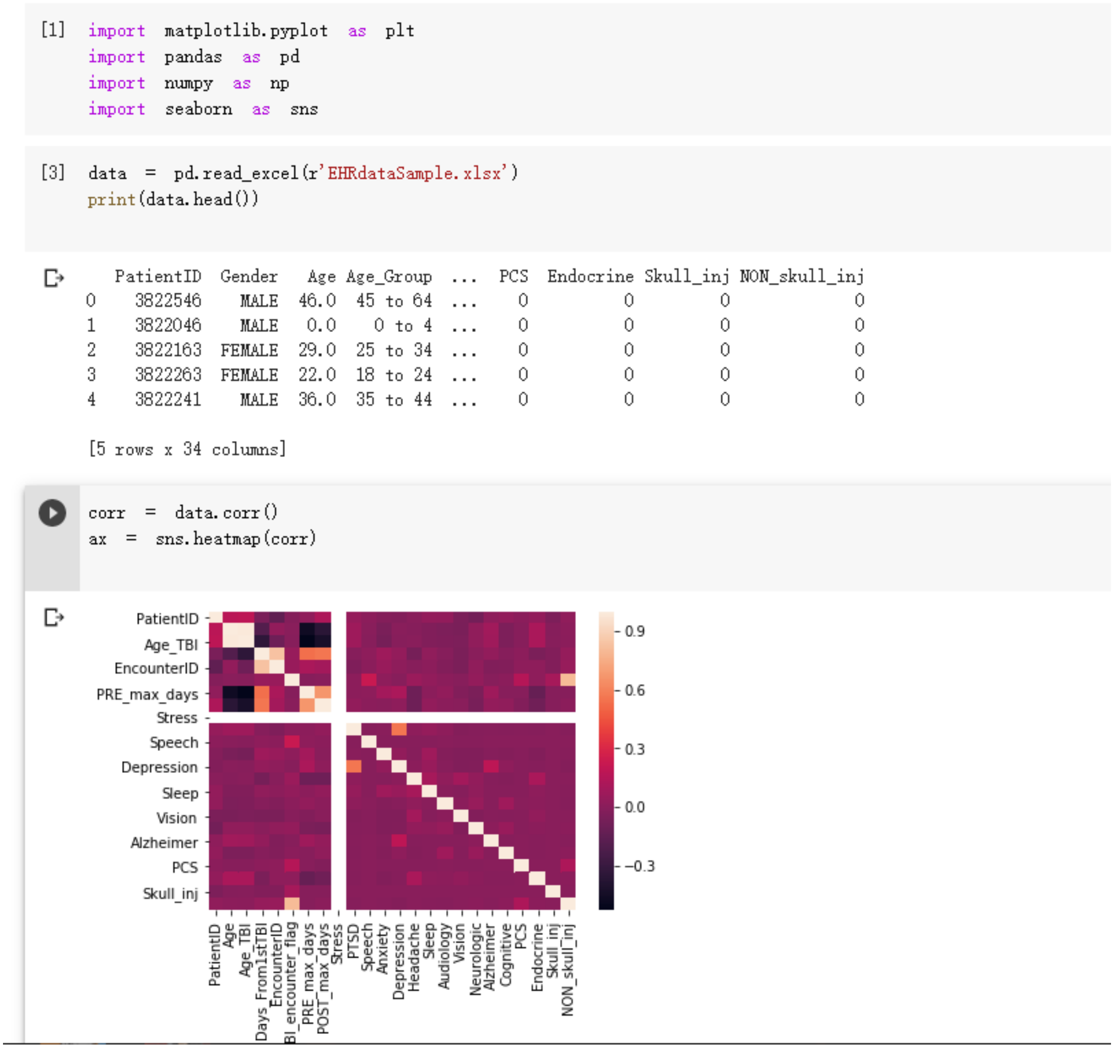
Visualization 2 Report:

Data using: EHRdataSample.xlsx

Linked Question: Correlation of every columns.

Tool using: Python

This is a first testing visualization of python. I used google colab to fulfill this idea. I tried to make a visualization of a general view on the data files. At the first the loading CSV file process trapped me again for a while. Then I found out the problem is in the CSV file itself not the original data nor the applications. So, I decided to directly load the XLSX file instead of the CSV file transferred by myself.



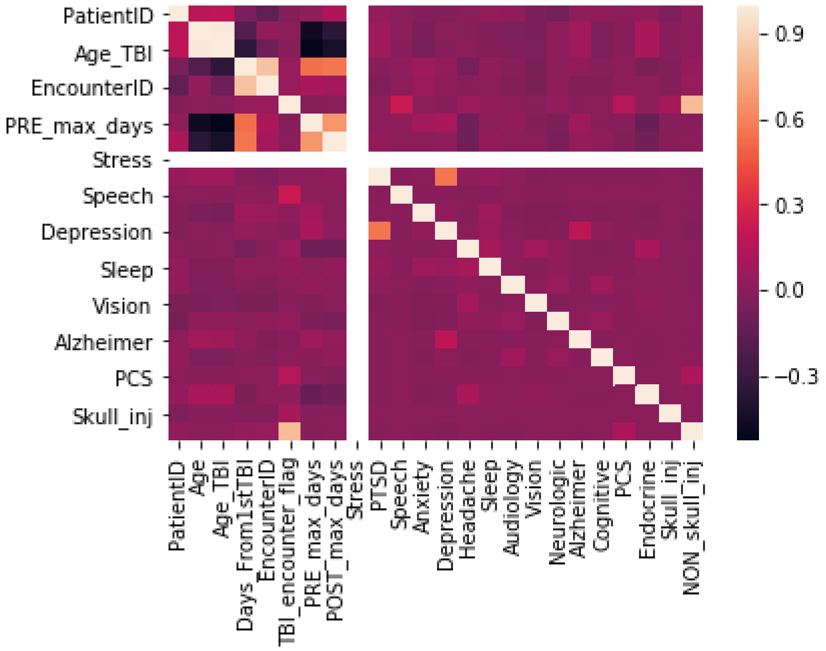
```
[1] import matplotlib.pyplot as plt
    import pandas as pd
    import numpy as np
    import seaborn as sns
```

```
[3] data = pd.read_excel(r'EHRdataSample.xlsx')
    print(data.head())
```

```
     PatientID  Gender   Age Age_Group  ...  PCS  Endocrine  Skull_inj  NON_skull_inj
  0    3822546    MALE  46.0  45 to 64  ...    0          0          0              0
  1    3822046    MALE   0.0    0 to 4  ...    0          0          0              0
  2    3822163  FEMALE  29.0  25 to 34  ...    0          0          0              0
  3    3822263  FEMALE  22.0  18 to 24  ...    0          0          0              0
  4    3822241    MALE  36.0  35 to 44  ...    0          0          0              0

  [5 rows x 34 columns]
```

```
corr = data.corr()
ax = sns.heatmap(corr)
```



The graph shows the correlation of every columns. The darker the color is, the stronger the correlation is. This graph looks interesting to me since its somehow strange but funny. After that, this graph might not be meaningful, since it can not be an answer to any actual question.

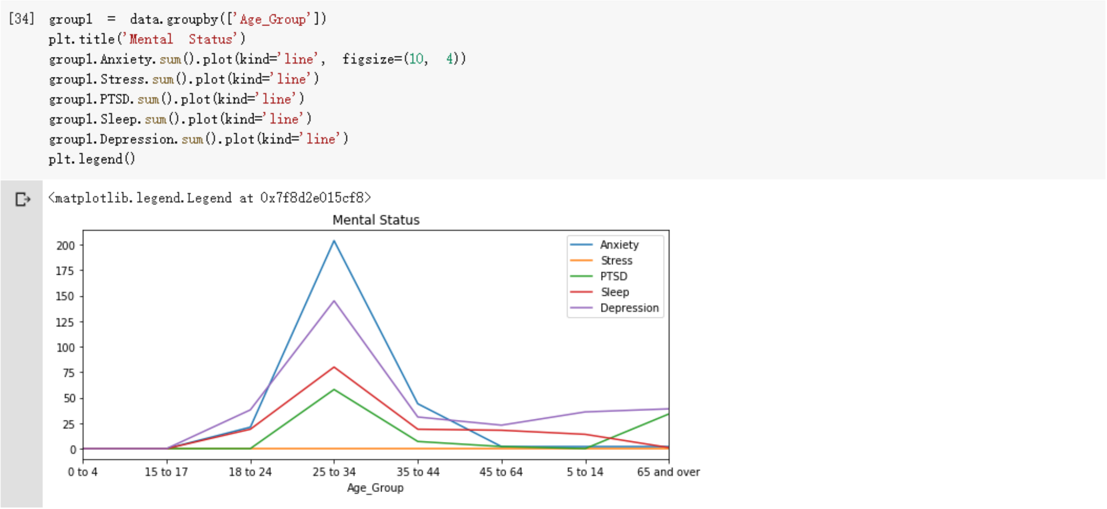
Visualization 3 Report:

Data Using: EHRdataSample.xlsx

Linked Question: What is patient count for each age group and what are their mental status portion?

Tool using: Python (Goodle Co Lab)

The Visualization 3 I decided to go back to the original Question. I then found out the first question can be easily fulfilled in just 2 sentences. However, the graph could be ugly and the further manipulation is confusing and difficult. After tons of hours of trying, I decided to make them as separate line graphs.



```
[34] group1 = data.groupby(['Age_Group'])
     plt.title('Mental Status')
     group1.Anxiety.sum().plot(kind='line', figsize=(10, 4))
     group1.Stress.sum().plot(kind='line')
     group1.PTSD.sum().plot(kind='line')
     group1.Sleep.sum().plot(kind='line')
     group1.Depression.sum().plot(kind='line')
     plt.legend()
```

```
<matplotlib.legend.Legend at 0x7f8d2e015cf8>
```

Conclusion:

My favorite visualization is the one made by Tableau, since its pretty and interactable with more clear information.

The Pro of Tableau is that it is powerful with various tools to fulfill visualization, while it is still simple for user to learn it.

The Con of Tableau is that it can be not very useful when use it visualize some specific data since it is not customizable.

The Pro of Python is that it has multiple different tools that can be added and it is programming language that is highly customizable.

The Con of Python is that it takes a while to first learn it as a new language or new visualization tool.