Weekly Report - 2

Since our discussion, significant advancements have been made towards enhancing the quality and scope of the project.

Upon receiving your valuable feedback, a pivotal decision was made to procure a more comprehensive dataset, leading to the adoption of the MIMIC-III dataset. This dataset, renowned for its extensive collection of clinical notes and patient records, presented a wealth of opportunities for deeper analysis and insights into healthcare trends.

Following that, intensive work was put into finishing two essential training modules that are necessary for successfully navigating and utilising the intricacies of the MIMIC-III dataset. Meanwhile, administrative work was done with diligence, and the necessary paperwork for obtaining the dataset were successfully submitted and approved.

With all of the data contained in the MIMIC-III dataset, acquiring the data itself was no easy task. It took almost ten hours to download the complete dataset, which was close to ten gigabytes. To streamline initial analysis and mitigate computational constraints, a pragmatic approach was adopted, focusing on a subset of data points for preliminary exploration.

Initial exploratory data analysis (EDA) centered on deciphering the dataset's structure and characteristics. Notably, attention was directed towards understanding patient outcomes, with a particular emphasis on the incidence of expiration during hospital stays. Moreover, demographic profiling encompassing patient ethnicities provided valuable insights into the dataset's diversity.

An essential aspect of the EDA involved elucidating the relationship between patient age, length of stay, and potential correlations therein. This was achieved through the visualization of scatter plots, facilitating the identification of underlying trends and patterns.

In conclusion, significant strides have been made in advancing the capstone project, with the successful acquisition and initial analysis of the MIMIC-III dataset laying a robust groundwork for upcoming work.

Below are some plots from initial analysis of MIMIC Dataset:

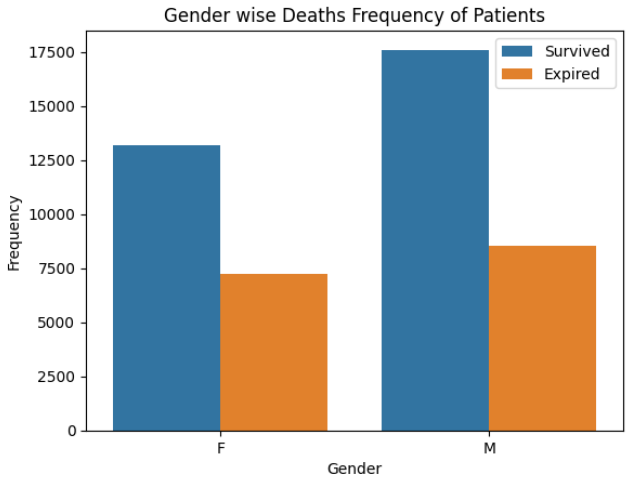
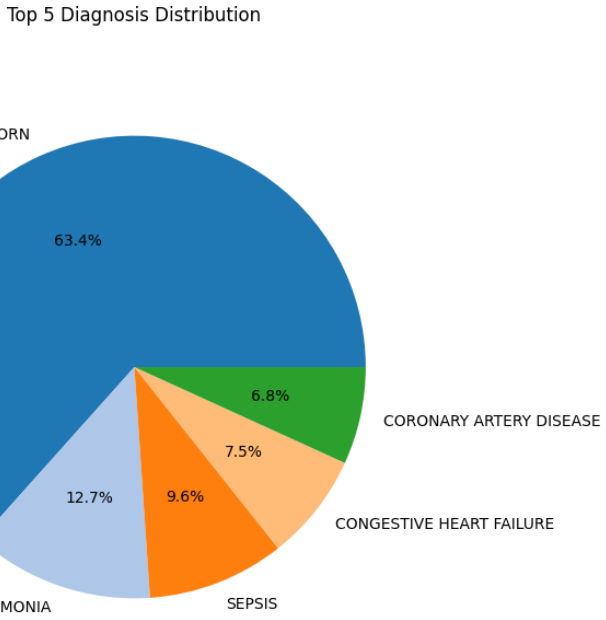


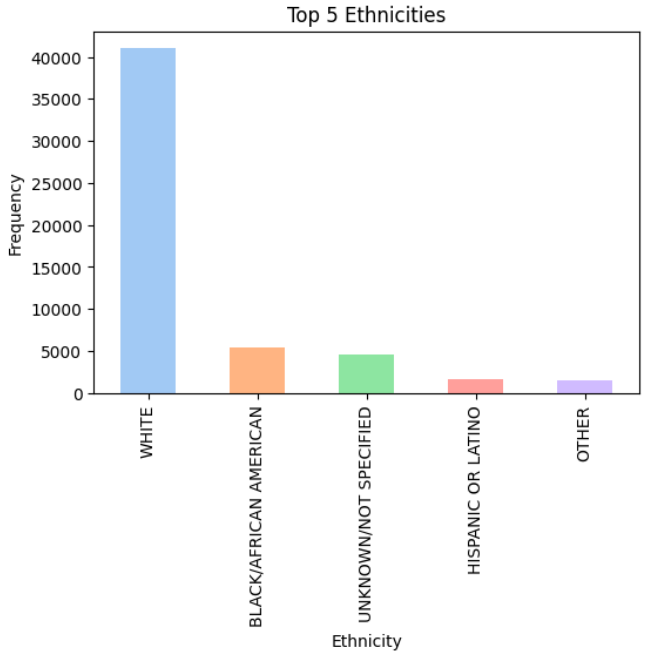
Fig. 1: Gender wise death Ratio in hospital




```
NEWBORN                                  7823
PNEUMONIA                                1566
SEPSIS                                   1184
CONGESTIVE HEART FAILURE                  928
CORONARY ARTERY DISEASE                   840
                                          ...
DIAPHRAGM RUPTURE                           1
RIGHT ANTERIOR CEREBRAL ARTERY STROKE       1
HYPOXIA, ACUTE RENAL FAILURE                1
S/P MOTOR VECHICLE ACCIDENT                 1
JOINT EFFUSION                              1
Name: DIAGNOSIS, Length: 15691, dtype: int64
```

Fig. 2: Top 5 Diagnosis Distribution

The above graphs summarize the top 5 diagnosis in the dataset. As we can observe, the most entries in the dataset are for Newborn, other meaningful diagnosis are Pneumonia, Sepsis, Congestive Heart Failure and Coronary Artery Disease. We have adequate data for analysis.




```
Top 5 ethnicity:
 WHITE                          40996
BLACK/AFRICAN AMERICAN          5440
UNKNOWN/NOT SPECIFIED           4523
HISPANIC OR LATINO              1696
OTHER                           1512
Name: ETHNICITY, dtype: int64
```

Fig. 3: Ethnicity Distribution of Patients

The above charts give us an idea about the patient's ethnicity. We can see that the dataset dominated by the "White" ethnicity, and other popular ethnicities include "Black/ African American" and "Hispanic or Latino".

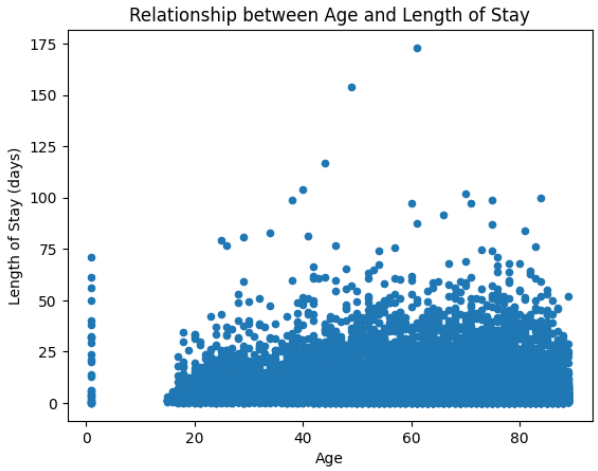


Fig. 4: Length of Stay based on Age

The above scatter plot shows how long did different aged people stay at the hospital. It can be noticed that the highest length of hospital stay is somewhere in the age of 60s. There are very less or no hospital stays around the age of 20, so it can be inferred that in older age the duration of hospital stay is more.

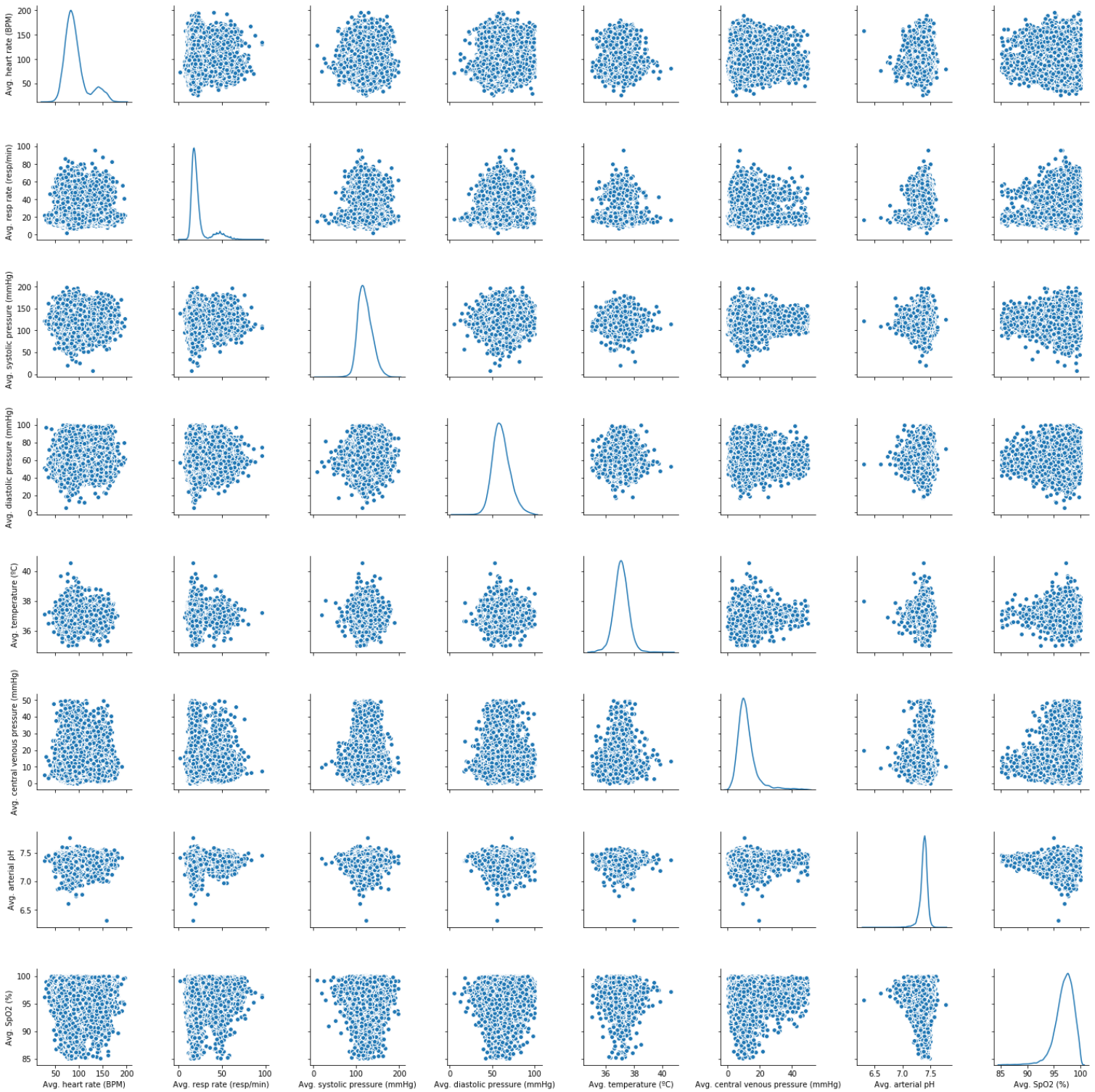


Fig. 5: Physio Data Pairplot

The above chart gives pairplot for the physio data.