WESTERN SYDNEY
UNIVERSITY

301111 Discovery Project

Submitted Spring 2019, in partial fulfillment of
the conditions for the award of the degree **Bachelor of Computer Science.**

Sajit Gopal Gurubacharya
18892488

Supervised by Doctor Yi Guo

School of Computing, Engineering and Mathematics
Western Sydney University

I hereby declare that this report is all my own work, except as indicated in the text:

Signature _____

Date 27/10/2019

I hereby declare that I have all necessary rights and consents to publicly distribute this dissertation via Western Sydney University's archive.

Public access to this dissertation is restricted until: 31/12/2019

Abstract

The unit 301111 Discovery Project is a capstone unit in the field of data science that aims to utilize the data science skills learned through-out the degree to produce a final discovery project report. As a final year computer science student with an interest in data science, this is an excellent unit to finish my degree.

This report aims to assist in the current development of a classification model to detect possible diseases with predictive modelling. An initial time series data set of past patients with various characteristics are analyzed and fine tuned. Then multiple dimensional reduction techniques are considered and used to develop consistent transformation functions of the data to a two dimensional scale which aids in its visualization. Finally, this is then used to classify different states of patients based on their profile.

Dimensional reduction techniques used are PCA and t-SNE with various adjustments to suit and better visualize the data set. Aim of this project is to be able to draft a function to map patient data matrix to visualize a simple graph output and check the feasibility of such a transformation.

Contents

Chapter 1

Project Proposal Report

1.1 Motivation

Patients in the emergency department(ED) in a hospital need extensive care due to the time sensitivity of the condition. Unfortunately, hospital departments like any other sector have limited resources in terms of both infrastructure and hospital staff. Thus, patients must be prioritized by the doctors before they are treated.

In Australia, the first step after a patient reaches the emergency department is their assessment through the Australasian Triage Scale. There is a need for categorization as someone with an urgent and life-threatening condition must be treated first regardless the arrived later than someone else with a less urgent condition. The triage nurse will record the patients vital signs including blood pressure, body temperature, pulse and respiratory rate. Addition information including the patients medical history, allergies and medications will also be recorded. Aligning this information along the set guidelines, the patient is placed into one of five categories and prioritized. These vital signs are re-taken once every hour to check the state of the patient.

There are a number of vital signs that are taken regularly which are normalized based on additional factors of the patients such as their age and sex so that each category is scaled relatively which will enable direct comparison between all patients regardless of their background.

To simple view these values in a standard table format is both time consuming and error prone as it is not easy to visualize all factors at the same time, especially when there is data from many hours for a single patient. In such a time sensitive environment, it is the

doctors are also not able to give much time to assess each and every patient's vital signs. A visualization technique would be a Spider or Radar chart as shown below.

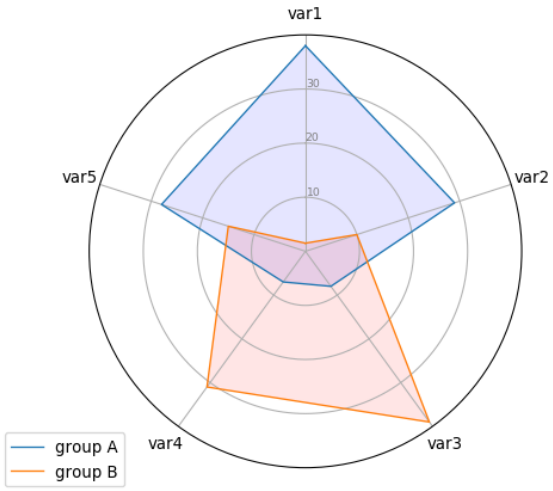

Figure 1.1: A Spider/Radar Chart.

The diagram shows the normalized values of five categories of two separate instances in a single diagram. This enables an easier visualization of all factors, it would still be difficult to make sense of it when more factors and multiple instances are piled up on top of each other.

The motivation of this project is to discover techniques to develop accurate visualizations a patient's vital signs to simplifying the process of classifying a patients state. This in turn will enable us to assess the possibility of using classification and machine learning techniques to aid doctors in speeding up the process of prioritizing patients in the Emergency Department.

1.2 Aims and Objectives

This project is particular aims to assist doctors in emergency departments of Hospitals to predict the possibility of a patients acquiring a condition called Sepsis when in the emergency department. There are seven vital signs that are taken for each patient and it is re-taken every hour. The objective for this project are as follows:

1. Discovering techniques to visualize time series data.

2. Researching on demensionality reduction techniques that accurately and consistently produces expected outputs from higher to lower dimension.

3. Trial and test such demensionality reduction techniques on data from the emergency department in hospitals.

   (a) This includes ensuring output is not merely a representation but a transformation of the input factors which assumes a presence of an underlying lower dimension model.

   (b) Training and validation sets are to be used ensure the implementation of the techniques are applicable.

   (c) Trail and test using a range of parameters to find better consistencies in the transformation.

4. Once reliable transformations have been produced, look up on machine learning and classifying techniques to produce accurate predictions of patients with Sepsis based on previous patient data.

1.3 Description of the work

The study for this project includes both taking inputs from recent advancements in classification techniques and also from the knowledge of sepsis related causes. Factoring both of these together will ensure suitable analytically techniques are used for sepsis related illnesses.

The key items to look for information are as follows:

1. Past journal articles include the current state of predictive analytics of patients in the emergency department.

   (a) Benefits and flaws of the current state of the Australasian Triage System will be taken into consideration when moving forward.

   (b) A sense of the public's perception on use of technology in hospitals for prediction tasks will enable us to gauge trust and demand of such new advancements.

   (c) Predictive techniques already in place will enable us to avoid initial mistakes.

2. Dimensionality reduction techniques such as Principle Component Analysis (PCA) and t-Stochastic Neighbour Embedding (t-SNE) and have different affects in identifying the underlying lower dimensional model or hyper planes in it. Applying

them to the current data set of ED patients will reveal if accurate and consistent transformations can be produced.

3. Classification techniques such as linear regression and SVM will play a key role once dimensionality reduction has taken place.

1.4 Data Sources

The primary source of the data for this project will be provided from the hospitals in NSW. It is expected that over 500 patient data with their vital signs and outcomes will be used to build up training, testing and validation data segments. With this result, various additional parameters such as sensitivity and specificity of the test will be used to indicate the probability of the reliability of the data and our model.

Chapter 2

Literature Review

2.1 Classifying ED Patients

Below are summaries of some of the studies that have gone into similar topics pertaining to this report.

2.1.1 Noninvasive hemodynamic monitoring in the emergency department

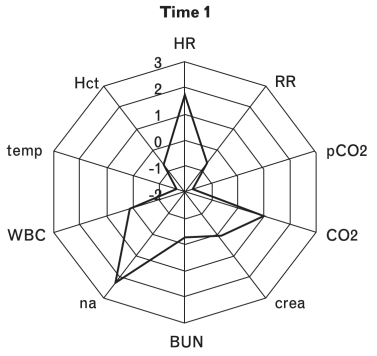

Figure 2.1: Visualization of a patient at one instance

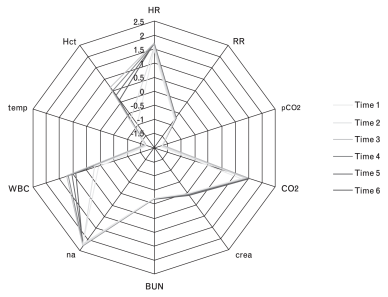

Figure 2.2: Visualization of a patient in multiple instance

This study by P.M. Middleton and S.R. Davies lays the foundation for this project[6]. Emergency departments need accurate risk assessments in a time sensitive scenario and efficient management can benefit patients critically. Hemodynamic monitoring refers to a noninvasive and valid technique to differentiate low and high risk patients. A range of these techniques are already in use and more are being developed. The factors that have been found to be most valuable in diagnosis are Pulse oximetry wave forms, electrocardio-

gram based heart rate variability, Doppler and B-mode ultrasound, echocardiograpy, trans thoracic bio impedance, pressure pulse waveform analysis and near-infrared spectroscopy. These data combined with the advancements in computer visualization has the potential to positively affect the prioritization of patients.

The drawback from these visualizations is when the medical staff need to look at hundreds of such diagrams. Similarly, when generating a predictive model, it would be challenging to classify patients based on multiple factors. In such a case, dimensionality reduction techniques can help reveal potential underlying models to better classify and prioritize patients.

2.1.2 Sepsis/SIRS Physiologic Classification

Sepsis is an illness with varying levels of intensity. Milder levels of sepsis can be treated while higher levels can prove to be fatal. The paper aims to develop a quantitative severity scale within the framework of Physiologic State Classification (PSSC)[8]. The study looked at 338 critically ill patients and captured 17 cardiopulmonary and metabolic variables and then classified them into 7 states of severity. The study concluded L2PDEATH provided a better indicator of the severity of sepsis for post-trauma patients. To visualize it, a similar spider chart was used, but with 17 variables instead of 7.
This approach did prove to be effective by comparing similarity between different patients based on their distance in a multidimensional hyperspace, but then converted the data into a logistic model which turned out to be: log(p/1-p)=-7.62+2.1x Distance from A-1.03x Distance from B, where A and B are different states of normal stress response and metabolic insufficiency respectively. We can see from here how layers of classifications are used to generate different models for different categories and find what variables generate greater value during model creation.

2.1.3 Exploring the Potential of Predictive Analytics and Big Data

Resource intensive casual inference the dominant way to judge a patient, big data sources remains relatively unexplored. Predictive analytics can provide simple heuristics for risk stratification. Factors apart from the patient's details such as time of day, day of the week or season could meaningfully contribute to present probability. Potential strategies

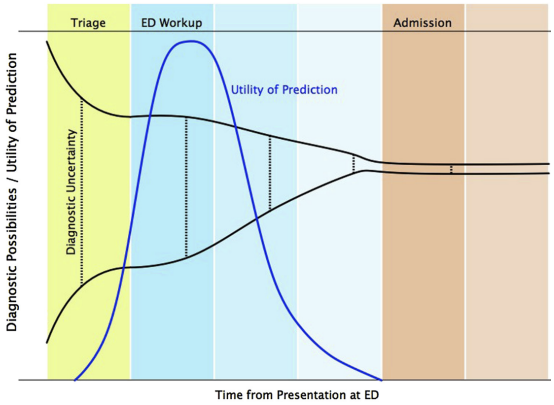

Figure 2.3: The utility of prediction

to predict include Bayesian networks, decision tree learning and markov and monte carlo simulations[3].

The figure above shows the ideal probabilities for diagnoses (black) and the utility of prediction (blue) during the different stages of ED presentation that results in a hospital stay. As time progresses, clinicians are able to narrow down the problem. For predictive utility, initially it is low as there is not enough data, but as there is more data from the patient, it increases, but decreases again later on as the clinicians can independently determine the problem. Thus the largest utility is during the early clinical encounter in the ED before resource committing decisions need to be made. The potential of analytics is no where more salient than in emergency and acute care setting.

2.1.4 Prediction of In-hospital Mortality in ED with Sepsis: Machine Learning Approach

A big data driven, machine learning approach is compared to clinical decision rules (CDR) and traditional analytic methods, in the domain of sepsis mortality rate[9]. ED patients were divided randomly into a 80/20 training and validation split and a random forest model was created with over 500 clinical variables from four hospitals. This random forest machine learning approach had , based on an ROC comparison, an accuracy of 86% compared to 69% for the CART model, 76% for the logistic regression model among others which showed that data driven machine learning techniques out performed CDRs as well as traditional analytic styles.

2.1.5 Improved PCA for Anomaly Detection: Application to ED

Conventional PCA based techniques such as Hotelling's T2 and the Q statistics are ill suited to detect small abnormalities because they use information from recent observations. In such a case, multivariate cumulative sum (MCUSUM) control scheme is better suited. For the first case of anomaly, the abrupt anomaly, T2 statistic generally failed to distinguish anomalies and while the Q statistic did do so, it could not detect smaller changes[2]. The MCUSUM could detect both. In the latter case of gradual anomaly, both Q and T2 statistic failed to detect anomalies and didn't cut the threshold mainly due to the noise during non-anomaly periods. The PCA based MCUSUM techniques could distinctly detect gradual anomalies and took the impact of noise significantly less.

2.2 Dimensionality Reduction Techniques

2.2.1 PCA

PCA is a technique which enables the extracting of a new set of variables form an existing large set of variables which are called Principle Components[10].
A principal component is a linear combination of the original variables. The first principal component explains maximum variance in the dataset. The second principal component attempts to explain the remaining variance in the dataset and is not correlated to the first principal component. The third principal component attempts to explain the variance not explained by the first two principal components and the iterative process continues.
Some of the assumptions and limits of using PCA are as follows:

1. It relies on an underlying model linearity.
2. The mean and the variance fully describe a probability distribution.
3. The principal components with greater associated variances correlate to dynamics, while lower variances represent noise.
4. The principal components are orthogonal.

2.2.2 t-SNE

t-SNE is a relatively new dimensionality reduction technique to output data points in a two or three dimensional map. It is a variation of Stochastic Neighbor Embedding [5]

which is easier to optimize and produces better visualizations by decreasing the likeliness of crowding points together at the center of the output space. The algorithm aims to evaluate the conditional probability of a certain point being a neighbor to another point nearby. It is a pair wise selection method which means the probability of the point being it's own neighbour is set to zero.

For SNE, when comparing the conditional probability of both higher and lower dimension data points, a cost function is applied to minimize the sum of Kullback-Leibler divergences. The cost function used by t-SNE differs because it uses a symmetrized version of the one SNE uses and a Student-t distribution rather than a Gaussian to compute the similarity in lower dimension space. This way the initial crowding/clustering and optimization problem is alleviated. interative process continues.

Some of the weaknesses of using t-SNE are as follows:

1. It is not clear on the performance when reducing to dimensions other than two or three.

2. As it reduces dimensions based on local properties of the data, it is sensitive to high intrinsic dimensionality of the data and with an underlying manifold that is highly varying.

3. The cost function is not convex which then requires several optimization parameters to be chosen.

2.2.3 LDA

Linear Discriminant Analysis (LDA) is defined as a dimensionality reduction technique by authors but some argue it works as a linear classifier. The algorithm tries to find a new feature space to output the data which maximizes class separability. Ronald Fisher in 1998 proposed a solution which maximizes the difference between the mean of each class (between classes) and minimize the spread of the class (within class), meaning this is a supervised algorithm. This relies on the data set following a Normal distribution, hence non-Gaussian data might not be suited for LDA[4].

2.2.4 Auto encoder

Autoencoders are an unsupervised learning technique which uses neural networks that imposes a bottleneck forcing a compressed knowledge representation of the input data. It is capable of learning nonlinear relationships and does not attempt to discover lower dimensional hyperplanes. To ensure the model is not memorizing the input data, the number of nodes in the hidden layer is sufficiently restricted. Because they rely on learning data compression based on the input, they are able to reconstruct data similar to the observations[7]. It is well suited for applications such as anomaly detect ions, information retrieval or image in painting.

2.3 Sepsis

Sepsis is a potentially life-threatening condition which occurs after a body's response to an infection. Normally, the body releases chemicals to the bloodstream when the infection occurs but when Sepsis occurs, these chemicals are out of balance which in-turn can damage multiple organs. Septic shock during Sepsis can cause blood pressure to drops dramatically[1].
Sepsis is most prevalent in the following age groups.

1. Older adults
2. Pregnant women
3. Children younger than 1
4. People with chronic conditions eg. diabetes, cancer
5. Weakened immune system

The early treatment of sepsis is generally with antibiotics and large amounts of intravenous fluids.

2.3.1 Symptoms

To be diagnosed with sepsis, you must have a probable or confirmed infection and all of the following signs: To diagnose sepsis, some of the following symptoms are taken into consideration:

1. Altered mental status
2. Systolic pressure less or equal to 100 millimeters of mercury (mm Hg)
3. Respiratory rate higher greater or equal to 22 breaths/minute

2.3.2 Causes

Some known causes of Sepsis are as follows:

1. Pneumonia
2. Infection in the digestive system (stomach, colon)
3. Infection in the urinary system (kidney, bladder)
4. Infection in the bloodstream

2.3.3 Complications

Sepsis ranges from less to more severe. As the severity of Sepsis increases, the blood flow to the vital organs such as the brain, heart and kidneys becomes impaired. Additionally, blood clots can form in the organs leading to organ failure and tissue death.
Patients do recover from mild sepsis, but on average the mortality rate for septic shock is around 40 percent. There is also a higher risk of future infections after an episode of sever sepsis.

Chapter 3

Report

3.1 Introduction to Data

This section describes the initial data set the project was worked on.

3.1.1 Background

The dataset "Raw Data.csv" was the dataset worked on to begin with. It consisted of data of a number of patients from a hospital in Sydney. These patients came in to the emergency department and were kept there for upto 24 hours where their vitals were recorded at certain time intervals and the end result of their stay were also recorded. Patients could stay for longer than 24 hours in which case they would have been admitted to the ward to taken out of the emergency care department to the ICU, but still technically still in the same hospital. Information collected were on patients arriving in the date interval of 07/08/2018 to 13/08/2018 and departure data interval of 07/08/2018 to 14/08/2018. While this was not a large time period, there were multiple data points recorded for each patient across the 24 hour time period of different vital signs resulting in a good amount of data to start working with.

Each patient had their characteristic information - age and gender to define them. They were each labelled as a character followed by a number, where the character is set by their pre-existing complaint and the number followed set by the nth patient to have a similar complaint. For example, C35 indicated the patient complained about chest pain and was the 35th patient to come in with the complaint during this time period. Each patient also had their result labelled after their time in the emergency department.

As information on patients were confidential, all work on the data set were carried out fully locally and in a secure manner. Data on patients in this report have been confidential by making them unidentifiable using bare minimum definitions such as age and gender, while only patients with over 5 instances have been reported on.

3.1.2 Defining Categories

Patients in this dataset had a number of defining categories relating to their stay in the emergency departments. These variables were used to further classify patients and make sense of them on what they really mean. They are listed below:

1. Patient Definitions
    - Record ID - Primary
    - Age
    - Gender
2. Emergency Department Details
    - Presenting Complaint - Chest Pain or Suspected Sepsis or Trauma
    - Time of Triage - ED Entering time in dd/mm/yy hh:mm format
    - Complete? - All patients are labeled completed
    - Diagnosis - Final Diagnosis by the Doctor
    - ED Discharge Time - ED Departure time in dd/mm/yy hh:mm format
    - ED length of stay - ED Discharge Time minus Time of Triage
    - Disposition - Result after stay in ED
    - Hospital length of stay - Time after disposed from ED
3. Vitals
    - Systolic blood pressure
    - Diastolic blood pressure
    - Mean arterial pressure
    - Heart rate
    - Respiratory rate

- Oxygen saturation
- Body temperature
- Variations of Vitals
  - actual - Recorded vitals
  - normal - Normal vitals based on patient
  - normalised - actual vitals on a normalized scale
  - Hours - initial, 1, 2, 3, 4, 6, 8, 12, 16, 20, 24 - intervals of vitals taken

3.1.3 Facts and Figures

1. Below are some facts about the dataset.
   - Patients of ages range from 7 - 92 averaging 55 years of age.
   - 139 instances of patients
   - 3 different initial complaints
   - ED Arrival time interval from 7/08/2018 to 13/08/2018
   - ED Departure time interval from 07/08/2018 to 14/08/2018
   - 7 Vitals recorded
   - 11 time intervals
   - 121 length of stay
   - 77 diagnosis
   - 2 Genders - Male and Female
   - 4 Dispositions - Discharged Home, Admitted to Ward, Admitted to ICU and Admitted to Coronary Care

Using these details of our data, we can generate meaningful data accordingly.

3.2 Making Sense of Data

Now as the dataset has been defined, we look into what each variable and categories mean and if they relate to each other, in order to make sense of the whole dataset. Some flaws of the dataset are looked into, then the R code that filtered the working matrix are described along with the final database.

The main flaw with the given database was the lack of fully usable patient data. A number of vital signs were missing for each patient time period. This meant that if a set of 7 vitals were not recorded for each time period for a certain patient, then that instance of the recordings needed to be removed. This also meant after removal, the time initially set intervals were not consistent among patients.

3.2.1 Ambiguities

Variables in this dataset are understandable but the meaning is generated from how they relate to each other. When doing this the vagueness of some variables raised concerns. For example, it was not clear why there were only three initial complaints including chest pain, suspected sepsis and trauma. Could this have been due to initial filtering out of patients based on sepsis related diagnosis or simply respiratory related complaints. If other presenting complaints had been included in the dataset, we could have been able to better classify respiratory vs non respiratory related conditions based on the differing vitals to narrow down on sepsis itself. This would have also resulted in overall more instances of patients giving us more data to work with.

The vital signs of the patients were recorded as is initially and in the set time intervals. Reasoning to why time intervals were not kept consistent is not known but it is assumed to be the case because of the stabilization of patients and their vitals the more time they spend getting treated in the ED. Similarly based on the recorded vital signs, they were all normalized. There were standard normal vital signs of each patient to which the recorded vitals were normalized on a scale of 0-2. The reason for which the normal vital values of each patient was different is not known but it is assumed to be based on factors such as their age, gender and initial complaint. For example, older patients might generally have greater blood pressure while teenagers might not, thus both of their vitals being normalized on the same 120/80 scale would have generated incorrect normalization values. There were a number of normalized vitals that had the value of 1.0000 which could entail there were errors or a great amount of simplification of the values during the normalization process.

It is also not fully clear why only the first 24 hours of vitals were recorded for patients who stayed in the emergency department for more than 24 hours. Since there were two

columns based on length of stay - one for ED stay and the other for hospital stay which is accounted only after the patient leaves the ED. It is assumed this was done as standard protocol of taking vital signs for all patients regardless of their length of stay to keep their records consistent. This in turn causes a further ambiguity where it is unclear if patients who were discharged from ED earlier than the 24 hour time period due to either head home or head to the ICU had their vitals not recorded either because they were not present in the ED or if the ED staff were not able to record their vitals for other reasons. Regardless, such patients who did not have their vitals recorded for a certain time period were not accounted for.

Similarly, it is unclear why only one or two vital recordings were not present while five other vitals were recorded. It would have been assumed that the hospital staff would record all vitals at once for each patient, but the data set has numerous patients with only one vital out of seven not being recorded for a time instance. Further more, for the same patient on a different time instance, it is often the case that another vital (different from the one before) is not recorded, but the previous vital sign is. If only one vital sign was consistently not recorded for all time instances, we could have removed that vital sign and still used the data, but since that was not the case, numerous instances for a number of patients needed to be filtered out.

3.2.2 Filtering with R

It was decided that only patients with a full set of seven recorded vital signs for a time period would be taken into account. This meant all patient matrices would have seven columns but variable rows of time. This again means these instances would not evenly be spread out. Even initially the recorded time periods were unevenly spread out, but removing further instances caused patient data time recordings to be inconsistent relative to other patients. Some patients could have 9 recordings taken while others could have 4, but in both case, the first instance correlated to the earliest recording while the last instance correlated to the latest recording. This consistency of forward time recordings was used later on to visualize the trend of these vital signs as time passed.

To filter out the dataset, it was trickier than initially thought as it was my first time working on filtering a dataset rather than simply using one or using a pre-built in one. Concepts of data.frames, matrices and arrays were new in terms of their use in R which

meant the same code could not have been used to process the same data on different data storage types even though the logic would have been the same. Other factors such as numeric data types, header data and different requirements for libraries also came into play when filtering the dataset. Similarly, since one patient data was spread across one single row with all vitals recorded in different time periods, the columns needed to be broken down into more rows for the same patient so it could be later processed more easily.

Below is the main section of code to break down the time intervals vital signs into their own rows. The hours had to be hard coded manually, but the rest was automated.

```
i<-1
list <- array(0, dim=c(77,139))
for (i in 1:verticalLength) {
    result = as.matrix(cbind(hr0[i,], hr1[i,],hr2[i,],hr3[i,],hr4[i,],hr6[i
    ,], hr8[i,],hr12[i,],hr16[i,],hr20[i,],hr24[i,]))
    dimnames(result) <-list(rep("", dim(result)[1]), rep("", dim(result)
    [2]))
    list[1:77,i] <-result
}
```

From 139 patients, we had 1529 rows of vital signs combined - 11 per patient, but after omitting out all blank values, we ended up with 382 rows of usable vital signs instances, averaging 2.74 recordings per patient. At this time in the project, it was realized that we could possibly encounter problems with a lack of usable data set as each instance had seven variables that was planned to accurately get reduced to a 2 or 3 dimensional plane, but even using only 3 sets of 7D points to plot it into a 2D point seemed to not be enough to provide an accurate consistent function. Nevertheless, the project was continued upon using industry standard dimensional reduction techniques. Below is the filtering code that automated the addition of Emergency Department related variables per patient. It was done this way as initially these details were set for one patient in one row, but now needed to be reproduced to be labelled accurately across multiple instances for the same patient.

```
start<-1
for (i in 1:139) { #columns/patients
  for (j in 1:11){ # hours
    patientInstance<-list[start:(start+6),i]
    patientInstance<-c(patientID[i],patientInstance)
```

```
    patientInstance<-c(patientInstance,LengthOfStay[i]) #8th col Length of
    Stay
    patientInstance<-c(patientInstance,DiagnosisAsFactors[i]) #9th col
    Diagnosis
    patientInstance<-c(patientInstance,ComplaintAsFactors[i]) #10th col
    Complaint
    patientInstance<-c(patientInstance,GenderAsFactors[i]) #11th col Gender
    patientInstance<-c(patientInstance,DispositionFactors[i]) #12th col
    Disposition
    data2<-rbind(data2,patientInstance)
    start<-start+7
  }
  start<-1
}
```

3.2.3 Usable Data set

It was my first year statistics professor that without good data to work with, any analysis that we do could be meaningless. A good amount of this project was spent on fixing the "Raw Data.csv" file to be used. Some examples of the matrices of patients ended up using are shown below:

```
> data2[45:54,] #C105
      [,1]   [,2]      [,3]     [,4]       [,5]          [,6]          [,7]          [,8]          [,9]   [,10] [,11] [,12] [,13]
 [1,] "C105" "1.05625" "1.075"  "1.065625" "1.05952381"  "1.333333333" "1.030927835" "0.991803279" "23.4" "11"  "1"   "2"   "3"
 [2,] "C105" "0.99375" "1.25"   "1.121875" "1.023809524" "0.888888889" "1.010309278" "1.008196721" "23.4" "11"  "1"   "2"   "3"
 [3,] "C105" "1.06875" "1.15"   "1.109375" "1"           "1.222222222" "1"           NA            "23.4" "11"  "1"   "2"   "3"
 [4,] "C105" NA        NA       NA         NA            NA            NA            NA            "23.4" "11"  "1"   "2"   "3"
 [5,] "C105" "0.975"   "1.1125" "1.04375"  "0.988095238" "1.166666667" "1"           NA            "23.4" "11"  "1"   "2"   "3"
 [6,] "C105" "1.0625"  "1.175"  "1.11875"  "1.035714286" "1"           "0.979381443" NA            "23.4" "11"  "1"   "2"   "3"
 [7,] "C105" "1.075"   "1.725"  "1.4"      "1.023809524" NA            "1.020618557" "1.008196721" "23.4" "11"  "1"   "2"   "3"
 [8,] "C105" "1.05625" "1.1875" "1.121875" "0.952380952" "1.111111111" "1.020618557" NA            "23.4" "11"  "1"   "2"   "3"
 [9,] "C105" "0.9625"  "1"      "0.98125"  "1"           "0.888888889" "1.010309278" "1.008196721" "23.4" "11"  "1"   "2"   "3"
[10,] "C105" "1.08125" "1.2"    "1.140625" "0.964285714" "1.111111111" "0.979381443" "0.983606557" "23.4" "11"  "1"   "2"   "3"
```

Figure 3.1: Patient C105, Chest Pain, Hour 3 record not taken, Respiratory rate for Hour 8 not recorded and multiple Body Temperatures not taken.

```
> data2[1409:1419,] #T14
      [,1]  [,2]          [,3]          [,4]          [,5]          [,6]          [,7]          [,8]          [,9]          [,10] [,11] [,12] [,13]
 [1,] "T14" "1.141732283" "1.177419355" "1.15936255"  "0.906666667" "1.444444444" "0.989583333" NA            "24.16666667" "41"  "2"   "1"   "2"
 [2,] "T14" "1.007874016" "0.822580645" "0.916334661" "0.92"        "1.111111111" "1.010416667" "1.016393443" "24.16666667" "41"  "2"   "1"   "2"
 [3,] "T14" NA            NA            NA            NA            NA            NA            NA            "24.16666667" "41"  "2"   "1"   "2"
 [4,] "T14" NA            NA            NA            NA            NA            NA            NA            "24.16666667" "41"  "2"   "1"   "2"
 [5,] "T14" NA            NA            NA            NA            NA            NA            NA            "24.16666667" "41"  "2"   "1"   "2"
 [6,] "T14" "0.850393701" "0.838709677" "0.844621514" "0.946666667" "0.833333333" "0.989583333" "1.008196721" "24.16666667" "41"  "2"   "1"   "2"
 [7,] "T14" "0.826771654" "0.983870968" "0.90438247"  "1.066666667" "1.111111111" "1.020833333" "1"           "24.16666667" "41"  "2"   "1"   "2"
 [8,] "T14" "0.763779528" "0.741935484" "0.752988048" "1.093333333" "0.944444444" "1.020833333" NA            "24.16666667" "41"  "2"   "1"   "2"
 [9,] "T14" "0.850393701" "0.822580645" "0.836653386" NA            "1.166666667" "1.010416667" "1.010928962" "24.16666667" "41"  "2"   "1"   "2"
[10,] "T14" "0.968503937" "0.822580645" "0.896414343" "1"           "0.888888889" "1.010416667" NA            "24.16666667" "41"  "2"   "1"   "2"
[11,] "T14" "0.834645669" "0.725806452" "0.780876494" "0.986666667" "1.166666667" "1.020833333" NA            "24.16666667" "41"  "2"   "1"   "2"
```

Figure 3.2: Patient T14, Trauma, Hour 2,3,4 record not taken, Heart rate for Hour 16 not recorded and multiple Body Temperatures not taken.

In figure 3.1 and 3.2 above, columns 9 to 13 entail ED details such as Disposition, Diagnosis and Gender as numerical categories.

```
> noNaData[221:229,]
          [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]
 [1,] 0.9455782 1.1971831 1.0692042 1.1012658 0.8333333 1.0102041 1.0054795
 [2,] 0.9523810 1.1830986 1.0657439 1.1392405 0.7777778 0.9795918 1.0000000
 [3,] 0.9387755 1.1267606 1.0311419 1.1898734 1.0555556 0.9693878 1.0383562
 [4,] 0.8027211 0.8873239 0.8442907 0.9873418 1.2222222 1.0000000 0.9972603
 [5,] 0.8027211 0.9154930 0.8581315 0.9873418 0.7222222 1.0102041 1.0136986
 [6,] 0.8435374 1.1126761 0.9757785 1.0126582 0.7222222 0.9693878 1.0000000
 [7,] 0.9652174 1.0298507 1.0000000 1.0250000 1.2857143 0.9600000 1.0081967
 [8,] 0.8782609 0.9104478 0.8955823 1.0375000 1.0000000 0.9500000 1.0000000
 [9,] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
```

Figure 3.3: Odd patient matrix with numerous normalized values of "1.0000000".

```
> noNaData[337:342,] #S22 with 8 instances
         [,1]      [,2]      [,3]      [,4]     [,5]      [,6]     [,7]
[1,] 1.174419 0.7794118 0.9324324 1.0120482 1.235294 1.0102041 1.038251
[2,] 1.104651 0.7647059 0.8963964 1.0361446 1.352941 1.0000000 1.021858
[3,] 1.348837 0.9117647 1.0810811 0.9879518 1.058824 1.0102041 1.005464
[4,] 0.954955 1.0714286 1.0199203 1.5487805 1.705882 0.8958333 1.054348
[5,] 1.054054 0.9571429 1.0000000 1.3414634 1.176471 0.9895833 1.046196
[6,] 0.963964 0.9714286 0.9681275 1.2560976 1.176471 1.0104167 1.048913
```

Figure 3.4: Example of patient S22, Suspected Sepsis, of its usable vital matrix.

It can be seen that patient data are not consistent amongst each other but vital signs are. Similarly the ED details for each patient were used for labelling and categorization for further analysis. Figure 3.3 and 3.4 above are examples of matrices for two different patients, the former with 9 instances of vitals fully recorded while the latter with only 6.

3.3 Visual Analytics

The aim of this project is to attempt to find suitable Visualizations of patients for doctors and hospital staff to look to judge to state of the patient instead of patient data like in figures 3.1 to 3.4 which is difficult to make sense of. Firstly we look into line graphs and radar charts as discussed in the Literature Review to see their effectiveness and feasibly, then we look into PCA and t-SNE to see the results they produced.

3.3.1 Libraries Used

In R, the libraries used are an important factor to consider as it is their algorithm that provides limitations and restrictions to data types that can be used, parameters that can be defined and variations that can be implemented on the final output. Some libraries and functionality might only be available in other programming languages such as Python

but not in R, but luckily in this project, all major libraries used existed for R. Below are the R libraries used.

1. For Visualising
    - tidyverse - collection of R packages designed for data science
    - tidyr - (under tidyverse) designed specifically for data tidying
    - dplyr - (under tidyverse) for working with data frame like objects, both in memory and out of memory.
    - ggplot2 - (under tidyverse) graphics language for creating elegant and complex plots
    - ggridges - to draw ridge lines
    - RColorBrewer - Provides color schemes for maps
2. For Implementing Algorithms
    - dimRed - initial PCA implementation attempt
    - Rtsne - tSNE implementation
    - e1071 - Short time Fourier transform, fuzzy clustering, support vector machines and others
    - factoextra - Provides some easy-to-use functions to extract and visualize the output of multivariate data analyses, including 'PCA'
    - tensorflow - open-source software library for Machine Intelligence
    - keras - under tensorflow, for high-level interface for neural networks, with a focus on enabling fast experimentation
    - DAAG - Data Analysis and Graphics Data and Functions
    - lattice - High-level data visualization system inspired by Trellis graphics, with an emphasis on multivariate data.
    - caret - Classification And REgression Training
    - MASS - Functions and datasets to support Venables and Ripley, "Modern Applied Statistics with S"
    - rpart - to validate PCA transformations

3.3.2 Simple Plots

To start off with the visualizations, simple plots such as bar/line graphs and radar charts were looked into. Here we can see the variance between patients our first vital sign. There

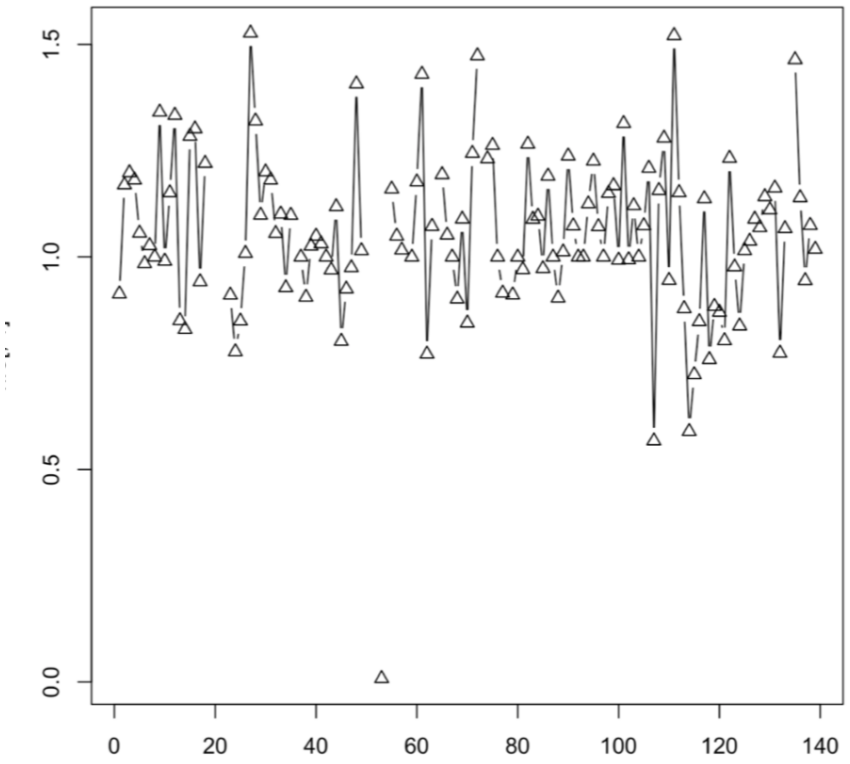

Figure 3.5: Normalized systolic blood pressure of all patients during initial time period

is one anomaly of the 50th range patient, but we cannot discard it as incorrect data as the the value is not NULL. In general, this vital sign does not seem to differ much, but towards the end the patients are of category Trauma or Suspected Sepsis and they seem to have a higher variance than the majority of Chest Pain patients. There were 104 Chest Pain patients, 19 with Suspected Sepsis and 15 with Trauma as their initial complaint. Thus after the 100 mark, the line plot refers to vitals of these patients.

In Figure 3.6 we try to visualise all data for one patient at once - something a doctor would do at first glance. While we can identify individual trends and spikes over the plot, say like for heart rate (4 on the X-Axis) or the patient when they first came in and how it normalizes as time passes (by looking at the dots on 4 on the X-Axis), and say point 6, which is the Oxygen Saturation level, which seems to stay constant throughout, overall it does look chaotic and the time periods are hard to track, even with a legend.

In Figure 3.7 we try to visualize the transpose of Figure 3.6 and see how the same data

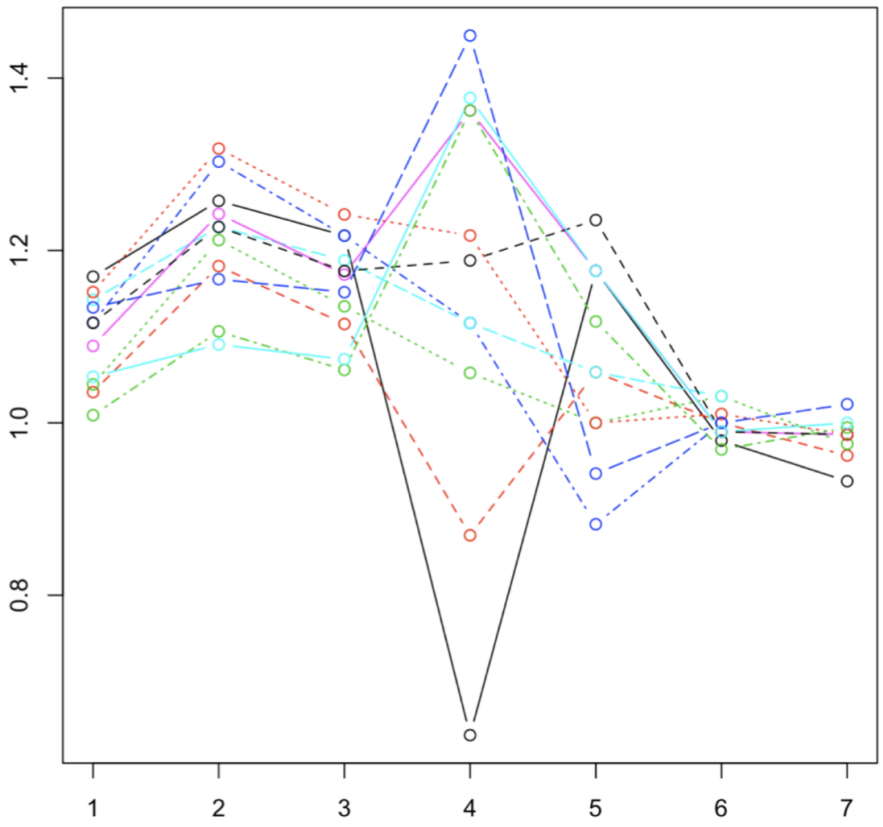

Figure 3.6: Each line is vitals of one patient at one time, eg black line is the values of a patient in hour (initial)

can be look at in two different ways. Instead of the black line referring to hour (initial), it now refers to one vital sign across the 11 time intervals. Personally, I prefer this one as it is more intuitive to understand and can put the focus on vitals instead of time periods. I assume it is because in this context, the time intervals are truly the independent variable (X-Axis) as they are predefined and thus should be represented in the axis accordingly, and while vitals are also predefined, they are categorical variables which do not make logical sense when plotted on an axis that is meant to represent numerical data. Hence in conclusion, the initial line graph should have been a bar graph to state the X-Axis as a categorical variable. In Figure 3.7, we can see that the heart rate of this patient (blue line) is in fact increasing over time instead of stabilizing. Similarly, we even clearly identify the gap in time period 4 and 5 of the dashed black line, something that was not as clear in the plot before. Analysing these two graphs shows how something simple as a line graph can be plotted more accurately to make the same data easily understandable. Nonetheless, this better visualization is still chaotic so we look into our next option - a Spider/Radar chart.

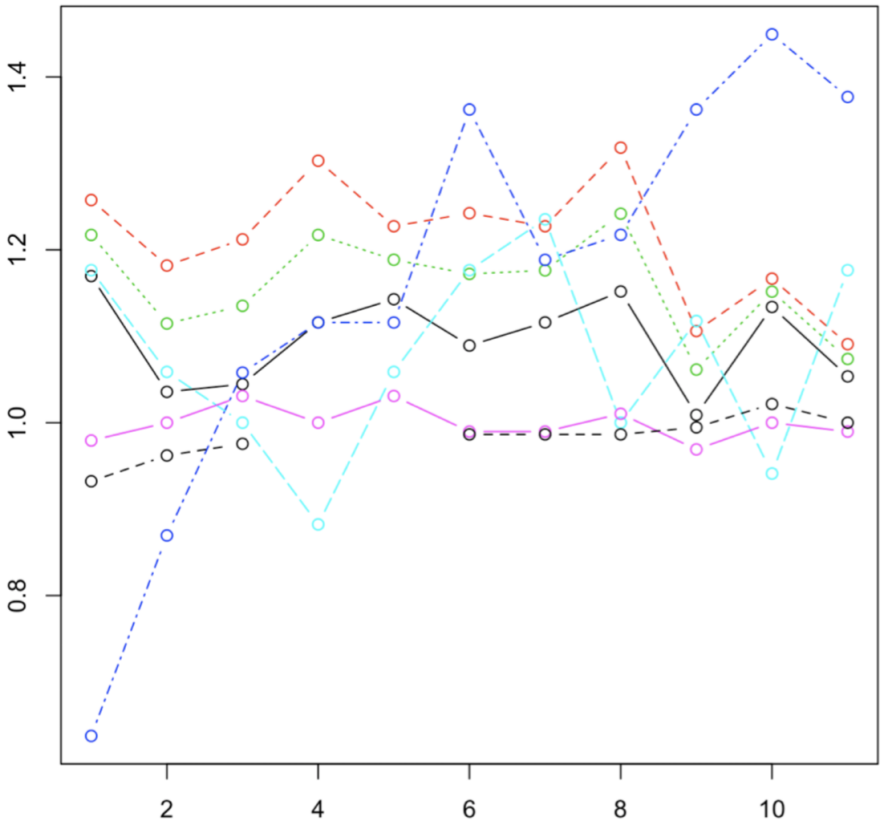

Figure 3.7: Each line is data of one patient's one variable, eg black line is the systolic blood pressure of a patient across the hours

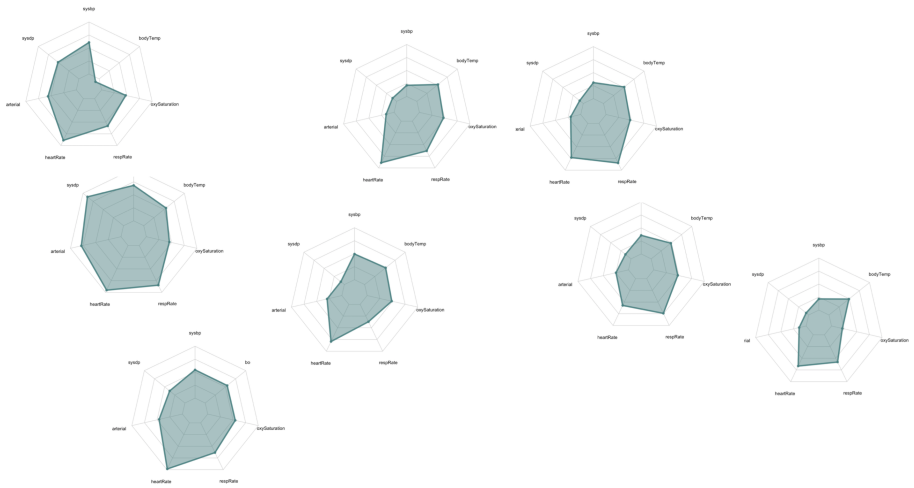

Figure 3.8: Vitals of patient with Tonsillitis over a 24 hour time period

Spider charts are a different way to visualize the same dataset. Each of the seven edges here represent the seven vital signs and their distance away from the center signify their value. When a number of charts are put side by side like so, we can get an idea of the magnitudes and trends of our data set with a quick glance. Here we can see the vitals are generally getting smaller for most vitals except for the heart rate for this patient.

Similarly, vitals tend to be a bit erratic towards the beginning while stabilizing to similar values as more time spent in the ED. If we try to differentiate such charts amongst different patients though, it becomes a problem.

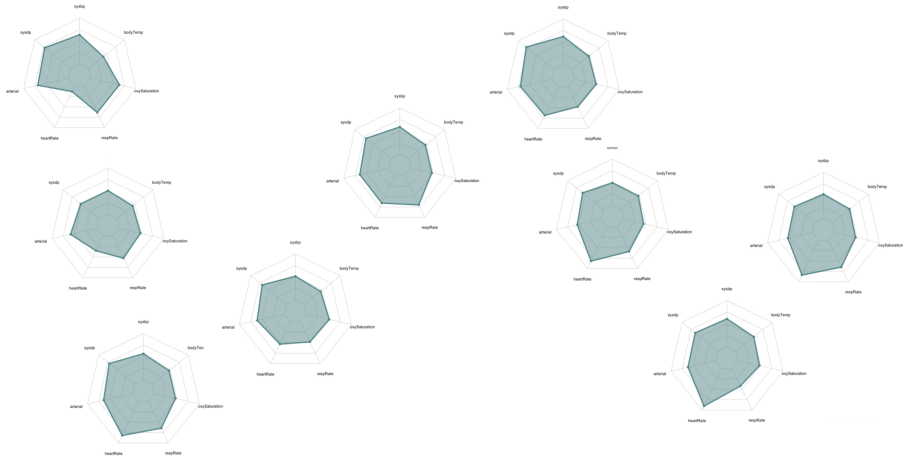

Figure 3.9: Vitals of patient with Chest Pain over a 24 hour time period

Trying to categorize the above charts for a Chest Pain patient versus the one in Figure 3.8 for a Tonsillitis patient is not possible. It could be because there are a number of charts for the same patient and therefore too many points to look into. We can still identify trends such as a stable heart rate and similar shapes of the chart as more time is spent in the ED.

Figure 3.10: Previous charts layered on top of each other

In this attempt, all charts are overlapped on top of each other. While this does give us an idea of the variations of the vitals, it is difficult to visualize the time interval of the chart on which comes first. Radar charts make it difficult to read the values because they are in a circular layout making it hard to differentiate the variations. Placement of categories also have a strong impact as the might induce bias when looking at certain sections of the chart. The scale of the chart also matters as values closer to the edge will have a larger

impact on the overall area of one chart misguiding the viewer of the true extent of its variation overall.

3.3.3 PCA

Principle Component Analysis is a procedure that transforms sets of possibly related observations to linearly uncorrelated variables called principle components. It does so by attempting to draw hyper planes across the dimensions to get the most correlated data together. Overall a seven dimensional data set will have seven principle components with the first few components largely describing the data indicating they are most correlated. As below in Figure 3.11, the 7x9 matrix is plotted on a 2D plane with 9 points indicating their transformation of the vital signs in each time interval. We can see a small circular trend of the points, but the transformation was only made for this matrix, thus cannot be used to compare data from another patient. Figure 3.12 is a Scree plot that shows how much of the data is explained by the principle components. In this instance, the first two components accounted for over 90 percent of the correlation, thus the 2D plot using the first two competes are a fairly accurate representation of the matrix as a whole.

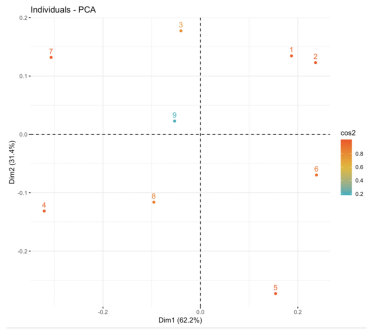

Figure 3.11: Patient's vitals over a time period

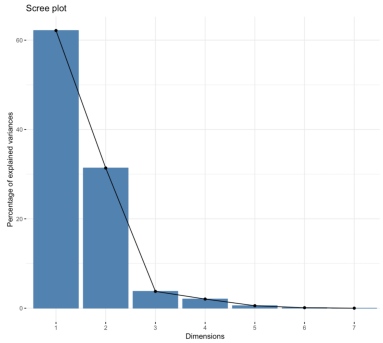

Figure 3.12: Scree plot of the transformation

```
> noNaData[221:229,]
          [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]
[1,] 0.9455782 1.1971831 1.0692042 1.1012658 0.8333333 1.0102041 1.0054795
[2,] 0.9523810 1.1830986 1.0657439 1.1392405 0.7777778 0.9795918 1.0000000
[3,] 0.9387755 1.1267606 1.0311419 1.1898734 1.0555556 0.9693878 1.0383562
[4,] 0.8027211 0.8873239 0.8442907 0.9873418 1.2222222 1.0000000 0.9972603
[5,] 0.8027211 0.9154930 0.8581315 0.9873418 0.7222222 1.0102041 1.0136986
[6,] 0.8435374 1.1126761 0.9757785 1.0126582 0.7222222 0.9693878 1.0000000
[7,] 0.9652174 1.0298507 1.0000000 1.0250000 1.2857143 0.9600000 1.0081967
[8,] 0.8782609 0.9104478 0.8955823 1.0375000 1.0000000 0.9500000 1.0000000
[9,] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
```

Figure 3.13: Patient matrix

Using data from the first 50 instances of a number of patients, we get to compare a number of patients at once relative to each other in the same principle components. Initially this was done for all patients, but visualization of each patient proved to be a mess, thus for the purposes of this project to understand these visualizations, the matrix of 13 Chest Patients have been plotted in a 2D graph in Figure 3.14 below.

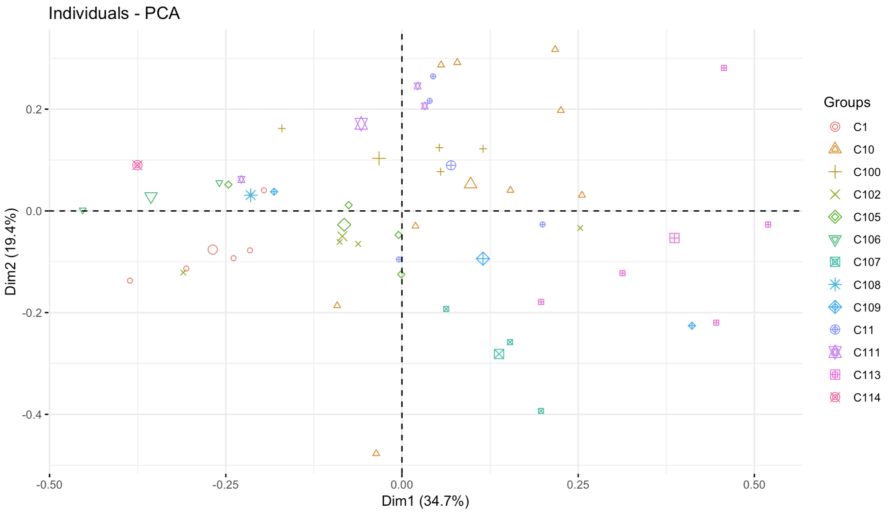

Figure 3.14: Plot of 13 patients

Here we can see each patient with its own symbol plotted in one space. Some patients only tend to lie in a specific region which differentiates them from other patients. These patients could be different in reality due to various factors such as their age, gender, or final diagnosis, giving us an idea of different patients having different vital signs which will help us categorise them more easily.

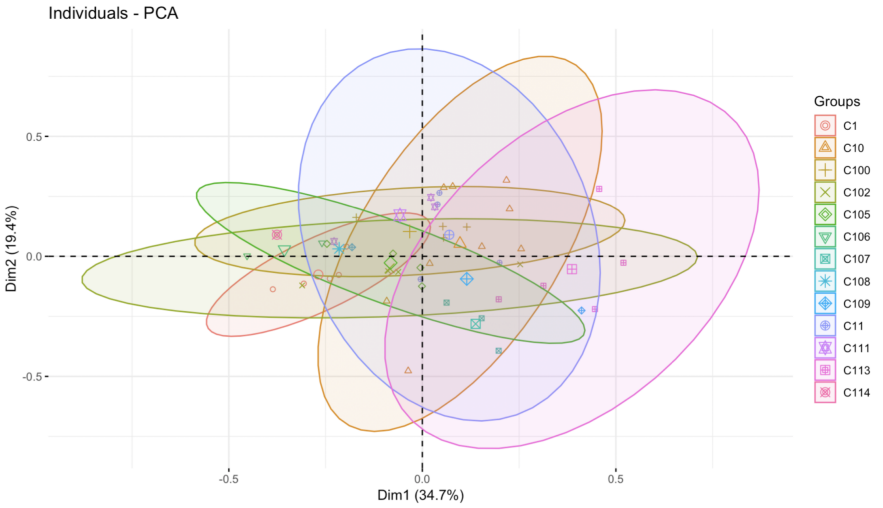

Figure 3.15: Ellipses over region of same patient

Plotting ellipses over each patient's plots, we can now more clearly see regions that the plots lie in. For example, patient C1's plots only line in close proximity to left of the center while patient C106's plots tend to stay flat on the y-Axis but go across the X-Axis indicating it relies more on the first principle component. One drawback of this plot is that the two highest components only account for 50 percent of the variation. This is inevitable as more data is used to generate these plots, the correlation between these points will be more spread out across more components as it is unlikely only a few vital signs are responsible in defining our patients.

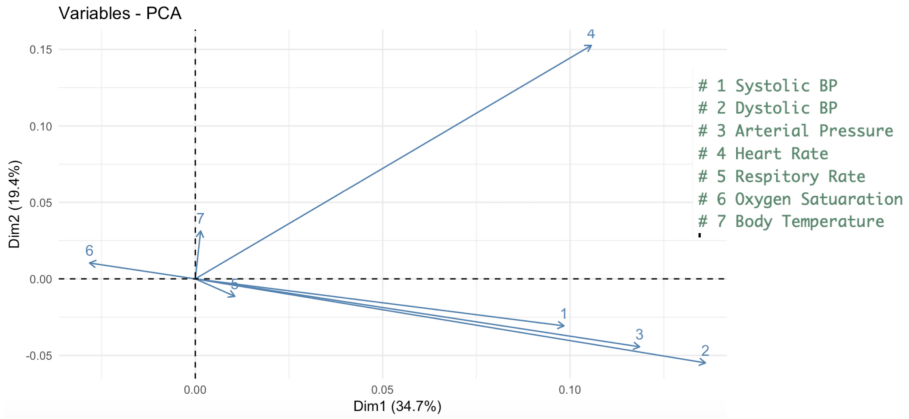

Figure 3.16: Variable correlation

When looking at variables in Figure 3.16, we can see some variables are more correlated than others. Even intuitively in makes sense that Systolic and Diastolic blood pressure along with the Arterial Pressure would show similar trends as one would not differ relatively much from the other. On the other hand, Oxygen Saturation is not showing the same trends as the other variables, thus is in the opposite direction. we can infer that Oxygen Saturation might not be a useful factor to consider when minimizing differences as it would cause disturbances to otherwise already correlated variables.

3.3.4 t-SNE

t-SNE stands for t-Stochastic Neighbourhood Embedding which is a non-linear dimensionality reduction technique which pushes away dissimilar points and brings closer similar points based on the t-distribution. Inherently, t-SNE selects high dimensional points randomly to compare. This random in turn means each time the data is reduced, it is likely to form different transformations.

We see the same shape for both of these plots as the seed was set initially to fix the ran-

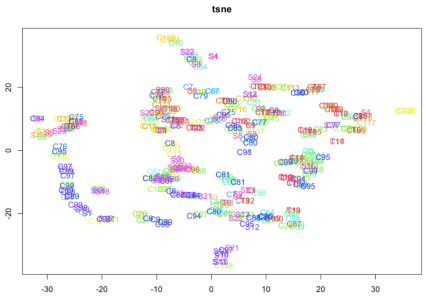

Figure 3.17: 121 Length of Stays

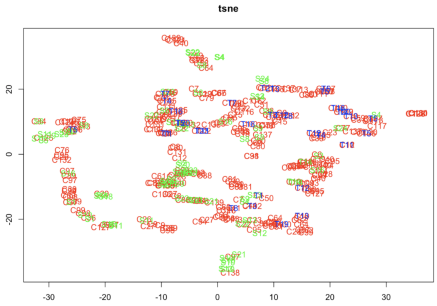

Figure 3.18: 3 Complaints

domness. The different colors represent the different categories the patients fit in. Most variables in our dataset had more than 5 categories making them difficult to visualize, as can be seen from comparing the two plots above. As for the three initial complaints, it can be seen that Trauma patients are only plotted on one section of the plot (blue), while the other two types of patients are plotted across. Some regions do only plot Chest Pain patients (red), but this could simply be due to randomness.

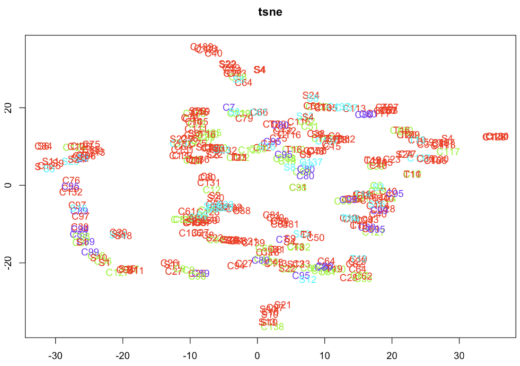

Figure 3.19: 4 Dispositions

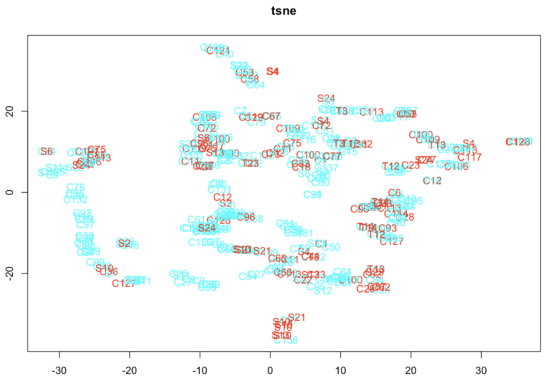

Figure 3.20: 2 Genders

With Disposition and Gender color filtered above, we still cannot see a clear trend in any category. This could either mean the categories are meaningless in predicting the vital values, or that the randomness of t-SNE is not well suited for the relatively low number of instances these plots represents and that the differences are minute enough to not be noticed by this algorithm.

3.4 Predictive Modelling

After visualizing our data, we look into what can be done to predict and classify out patients.

Apart from PCA and tSNE, LDA (Linear Discriminant Analysis), SVM (Support Vector Machines) and simple linear regression were applied on a section of the dataset. There were errors both in coding but also the data that were encountered, thus not being able to complete these sections during this semester. Patient C80 was used to test out the functions in order to test its suitability.

In the LDA attempt, the coefficients of the linear discriminats were achieved, but not further work has been dome in determining it's accuracy.

```
Coefficients of linear discriminants:
            LD1         LD2
c80_1  1.974747 -10.2988770
c80_2  6.228354  -7.4418891
c80_3 16.236562   0.6059572
c80_4 -1.415979   4.0679431
c80_5 -1.147045   4.1015385
c80_6 27.257121  14.0210083
c80_7 23.588660   1.8252247
c80_8  3.311731   6.4901249

Proportion of trace:
   LD1    LD2
0.9022 0.0978
```

In the SVM attempt, I encountered errors with the data type and the correct parameter that need to be put in. With simple linear regression, I was able to produce a few plots but no work has been done in predicting their accuracy nor its suitability as a classifier.

```
all:
lm(formula = patient_time ~ c80_1 + c80_2 + c80_3 + c80_4 + c80_5 +
    c80_6 + c80_7 + c80_8, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-19.0000  -5.0000  -0.2857   6.7143  24.0000

Coefficients: (2 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -795.99    1410.53  -0.564    0.576
c80_1       -1735.70    3696.43  -0.470    0.641
c80_2        2722.89    5704.98   0.477    0.636
```

```
c80_3        -1427.71     3025.30   -0.472     0.639
c80_4          456.30     1000.17    0.456     0.651
c80_5           28.07       94.15    0.298     0.767
c80_6              NA          NA       NA        NA
c80_7          832.73     1607.91    0.518     0.607
c80_8              NA          NA       NA        NA

Residual standard error: 10.51 on 42 degrees of freedom
Multiple R-squared:  0.2833,  Adjusted R-squared:  0.1809
F-statistic: 2.766 on 6 and 42 DF,  p-value: 0.02338
```

After the above attempts, it was decided that PCA was the most suitable function to accurately and consistently be implemented on our patient dataset, as tSNE could not produce consistent results.

After calculating the principal components on our initial 13 patient set, we try to test it with an additional separate patient. One thing to note is to not combine the initial and new set to obtain the components as this would mean we use all of our data to develop the new hyperplanes. To avoid data from the new test set to spread into the development of the principle components, we cannot simply perform PCA on the two sets separately as the resulting vectors from the two sets will have different directions, making them not comparable. To ensure both sets are compared on the same axis, the same transformation are applied to both the initial and new set, including the scaling and centering features.

```
> noNaData[50:56,]
          [,1]     [,2]      [,3]     [,4]      [,5]     [,6]      [,7]
[1,] 0.8296296 1.239437 1.0397112 1.288462 1.2631579 0.979798 0.9752747
[2,] 1.2837838 1.428571 1.3503650 1.650794 1.1428571 1.020833 1.0081522
[3,] 1.0067568 1.301587 1.1423358 1.507937 1.0000000 1.010417 1.0135870
[4,] 1.0000000 1.000000 1.0000000 1.000000 1.0000000 1.000000 1.0000000
[5,] 1.3015873 1.025641 1.1489362 1.178947 0.8888889 0.989899 0.9756757
[6,] 1.1507937 1.051282 1.0957447 1.052632 0.8888889 0.989899 1.0000000
[7,] 0.9416058 1.038462 0.9931741 1.333333 1.3750000 1.020408 0.9945205
> rpart.prediction
        1         2         3         4         5         6         7
1.1386095 1.1386095 1.1386095 0.9875407 1.0636958 1.0483427 1.0483427
```

Figure 3.21: New instances of patient data reduced using the same transformation.

Here, using the package "rpart" we include the principle components previously derived into our inital dataset. Then we develop a model under the ANOVA method using rpart, and use our next set of data to test and predict the values. Currently, it is not clear what

the predicted values mean but it does seem to be in the correct direction in the process of transforming both datasets the same way to hopefully generate new insights.
For the purposes of this project, auto encoders were not used because they focus on learning to re-generate the same data, mainly with images. As it does not take into consideration any time series factor of the dataset, auto encoders were not used but due to the lack of meaningful time series data, they can still be looked into if assumed the time intervals have less meaning.

3.5 Conclusion and Future Work

To conclude, the work in this discovery project mainly revolved around analysing understanding and making meaning out of patient dataset, visualizing techniques and implementing dimensionality reduction techniques. From the initial scope of attempting to create a function to consistently map a 7D data into a 2D space, the focus has diverted considerably into more visualising techniques.
The data initially given was thought to be plenty to train and test the data set, but due to the additional complexities such as complaints, dispositions, diagnosis etc. on top of the seven vital signs, after filtering it to only our usable dataset, it seemed not to be enough to classify patients individually.
PCA turned out to be a suitable technique amongst others including t-SNE, radar charts and out bar graphs as it highly simplifies multiple patient data while still preserving some of the meaning behind it.
For future work, there are a number of things to look into. This project came into the conclusion of the non-feasibility of t-SNE and shined a light on PCA. Even for PCA, there can be additional experiments trailed such as categorizing the reductions based on the ED details, between different Chest Pain/Trauma/Suspected Sepis patients. In regards to Sepsis as it was initially the focused disease, more direct research can be done into known patterns of sepsis patients and compared with the data there was on the sepsis patients directly. Finally, the SVM, LDA, linear regressions need to be looked into in detail to check their feasibility along with the possibility of using auto encoders with additional assumptions.
I would like to thank my professor, Yi Guo and Western Sydney University for giving me

the opportunity to take part in this project as a final year computer science student to implement my learned skills and build a strong portfolio in the ever growing field of data science.

Bibliography

[1] Bone, R. C., Balk, R. A., Cerra, F. B., Dellinger, R. P., Fein, A. M., Knaus, W. A., Schein, R. M., and Sibbald, W. J. Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. *Chest 101*, 6 (1992), 1644–1655.

[2] Harrou, F., Kadri, F., Chaabane, S., Tahon, C., and Sun, Y. Improved principal component analysis for anomaly detection: Application to an emergency department. *Computers & Industrial Engineering 88* (2015), 63–77.

[3] Janke, A. T., Overbeek, D. L., Kocher, K. E., and Levy, P. D. Exploring the potential of predictive analytics and big data in emergency care. *Annals of emergency medicine 67*, 2 (2016), 227–236.

[4] Li, M., and Yuan, B. 2d-lda: A statistical linear discriminant analysis for image matrix. *Pattern Recognition Letters 26*, 5 (2005), 527–532.

[5] Maaten, L. v. d., and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research 9*, Nov (2008), 2579–2605.

[6] Middleton, P., and Davies, S. Noninvasive hemodynamic monitoring in the emergency department. *Current opinion in critical care 17*, 4 (2011), 342–350.

[7] Ng, A., et al. Sparse autoencoder. *CS294A Lecture notes 72*, 2011 (2011), 1–19.

[8] Rixen, D., Siegel, J. H., and Friedman, H. P. " sepsis/sirs," physiologic classification, severity stratification, relation to cytokine elaboration and outcome prediction in posttrauma critical illness. *Journal of Trauma and Acute Care Surgery 41*, 4 (1996), 581–598.

[9] Taylor, R. A., Pare, J. R., Venkatesh, A. K., Mowafi, H., Melnick, E. R., Fleischman, W., and Hall, M. K. Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data–driven, machine learning approach. *Academic emergency medicine 23*, 3 (2016), 269–278.

[10] Wold, S., Esbensen, K., and Geladi, P. Principal component analysis. *Chemometrics and intelligent laboratory systems 2*, 1-3 (1987), 37–52.