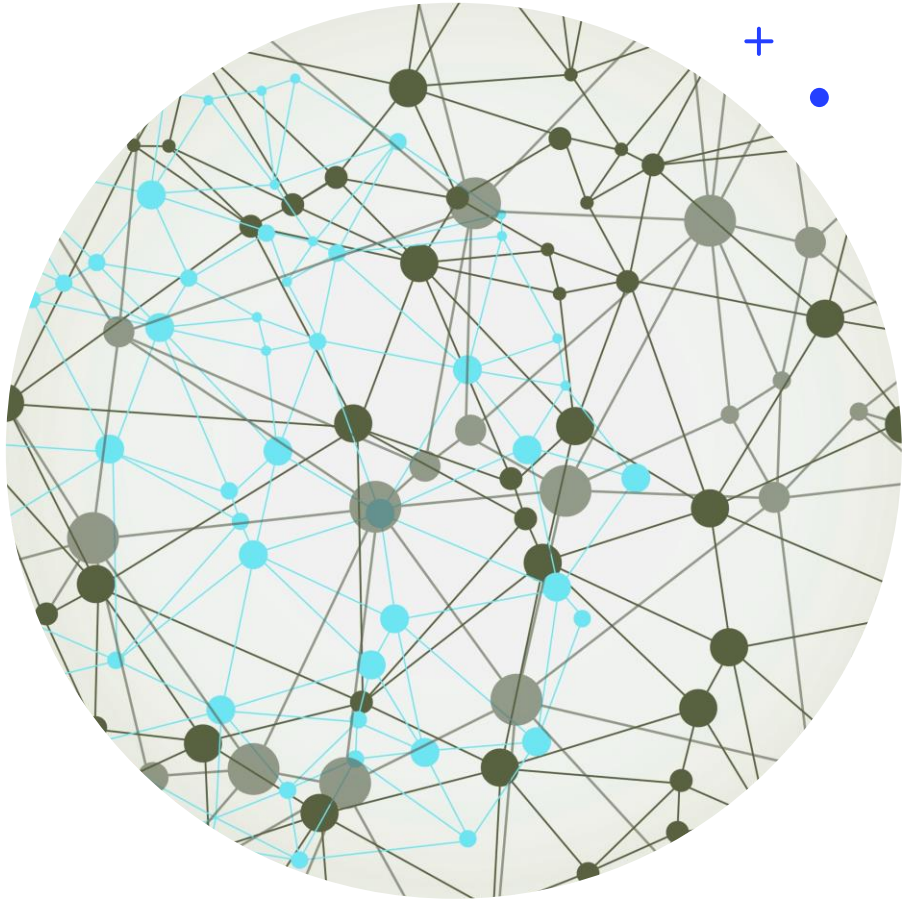
PROJECT: DATASET VISUALIZATION

Prepared By:

Dipkumar Patel

Noopa Jagadeesh

Prasanth Varma



Datasets to visualize

Bar Crawl: Detecting Heavy Drinking Data Set

Breast Cancer Wisconsin (Diagnostic) Data Set

Human Activity Recognition Using Smartphones Data Set

A study of Asian Religious and Biblical Texts Data Set

Student Performance Data Set

1. Bar Crawl: Detecting Heavy Drinking Data Set

Dataset Link

Dataset

Participant accelerometer data

TAC Reading: ankle bracelets - 30 minute intervals

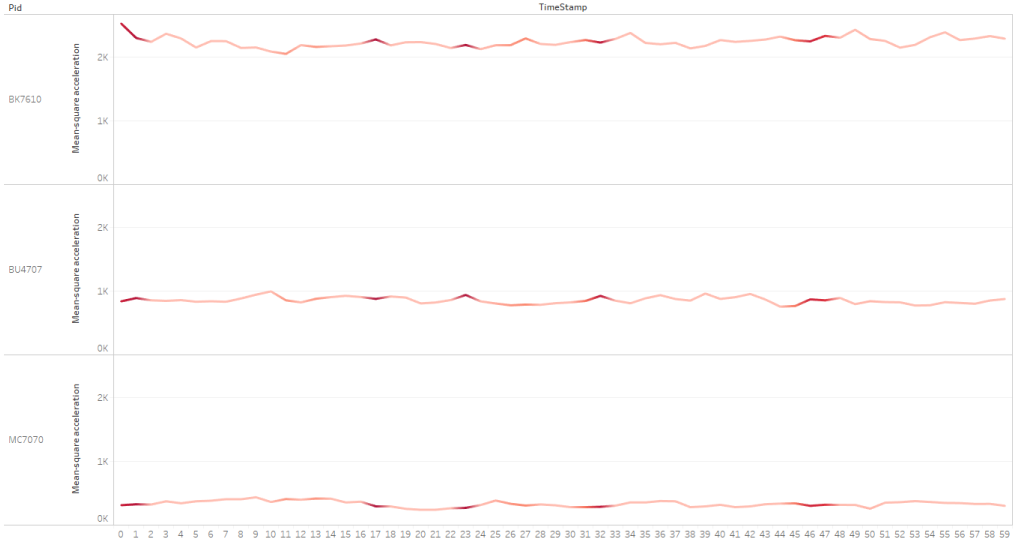
In total

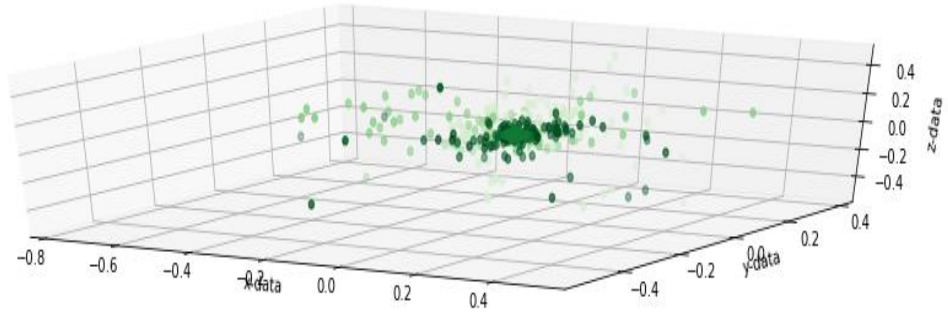
Accelerometer readings: 14,057,567

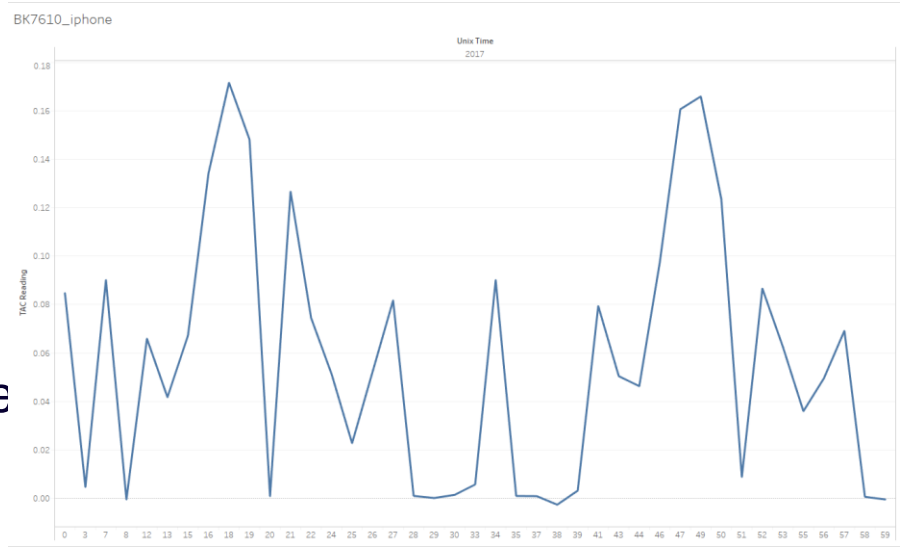
TAC readings: 715

Participants: 13

Insight







Questions:

1. Why did you apply this/these visualization
2. What kind of pattern(s) have you discovere
3. What is your final conclusion?

2. Breast Cancer Wisconsin (Diagnostic) Data Set

Dataset Link

Dataset

32 attributes

569 observations

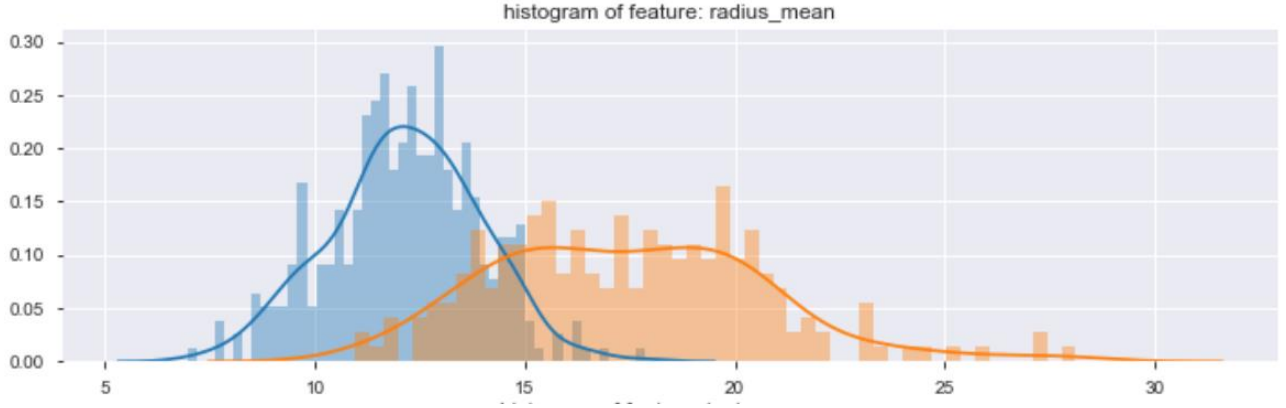
Target

M=malignant(cancerous),
B = benign(non-cancerous)

Problem Statement

- To understand the statistics of breast cancer dataset to fit the most suitable ML algorithm on the dataset.

Histogram Insights

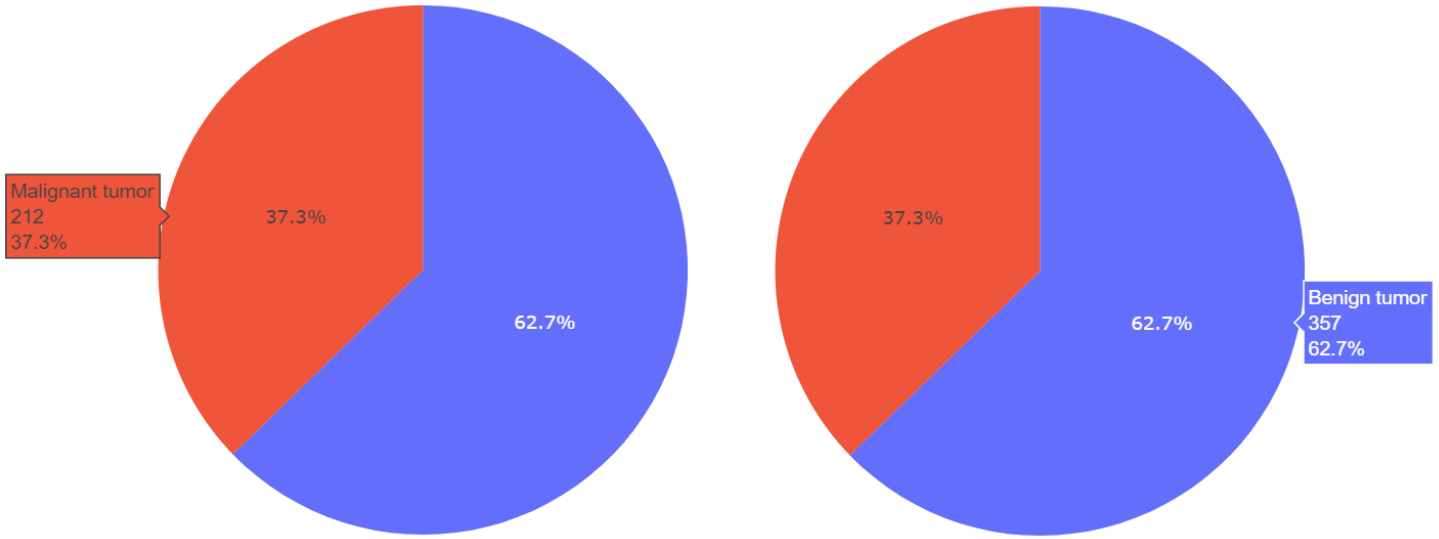


Insights:

1. Dataset is **not normally distributed** as some attributes have normal distribution while others are either right or left skewed.

Tool Used:
Matplotlib, python

Pie Chart Insights

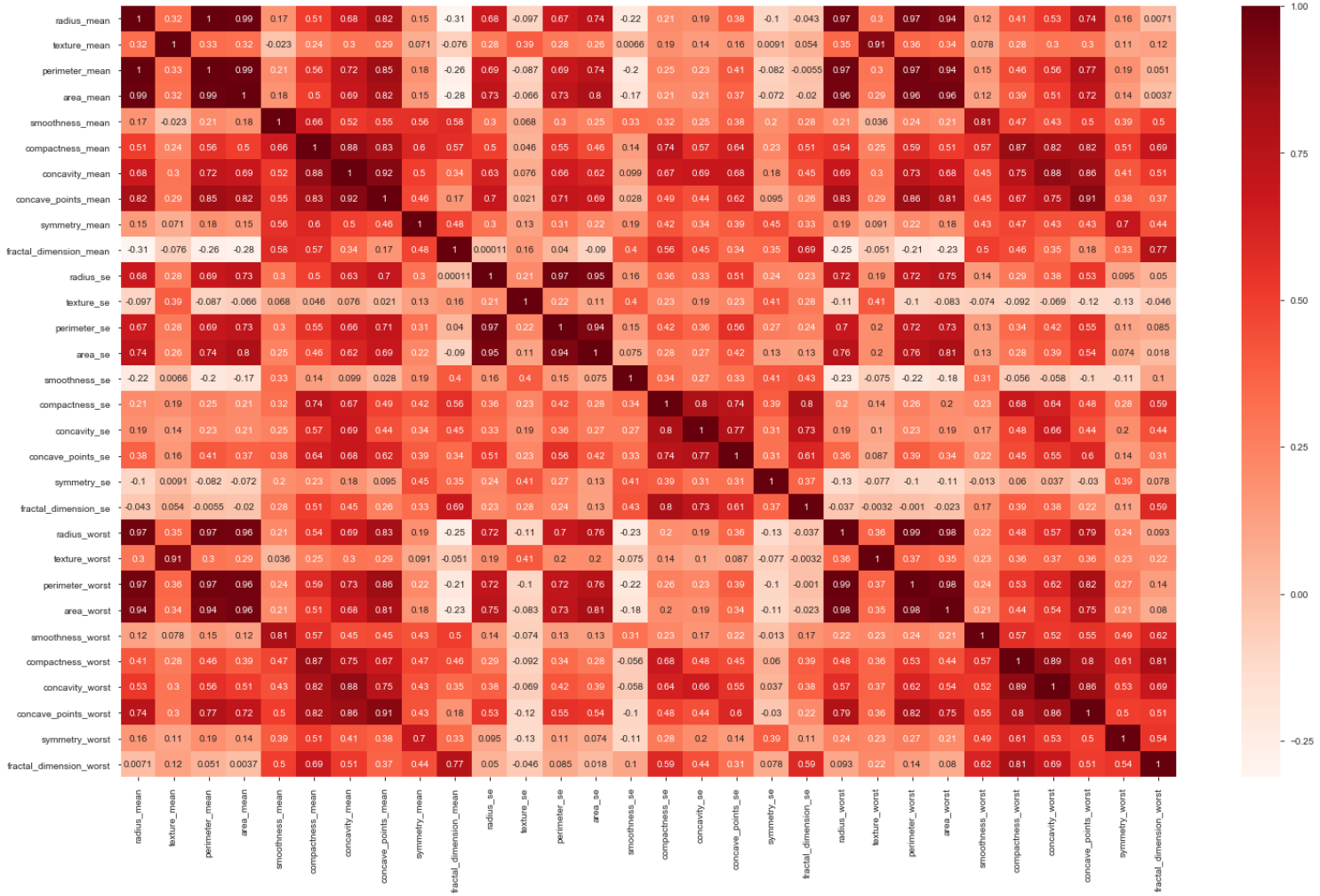


Tool Used:
Plotly, python

Insights:

1. Dataset is **not balanced.** No of instances of benign tumor class is way more than Malignant class

Heatmap Insights

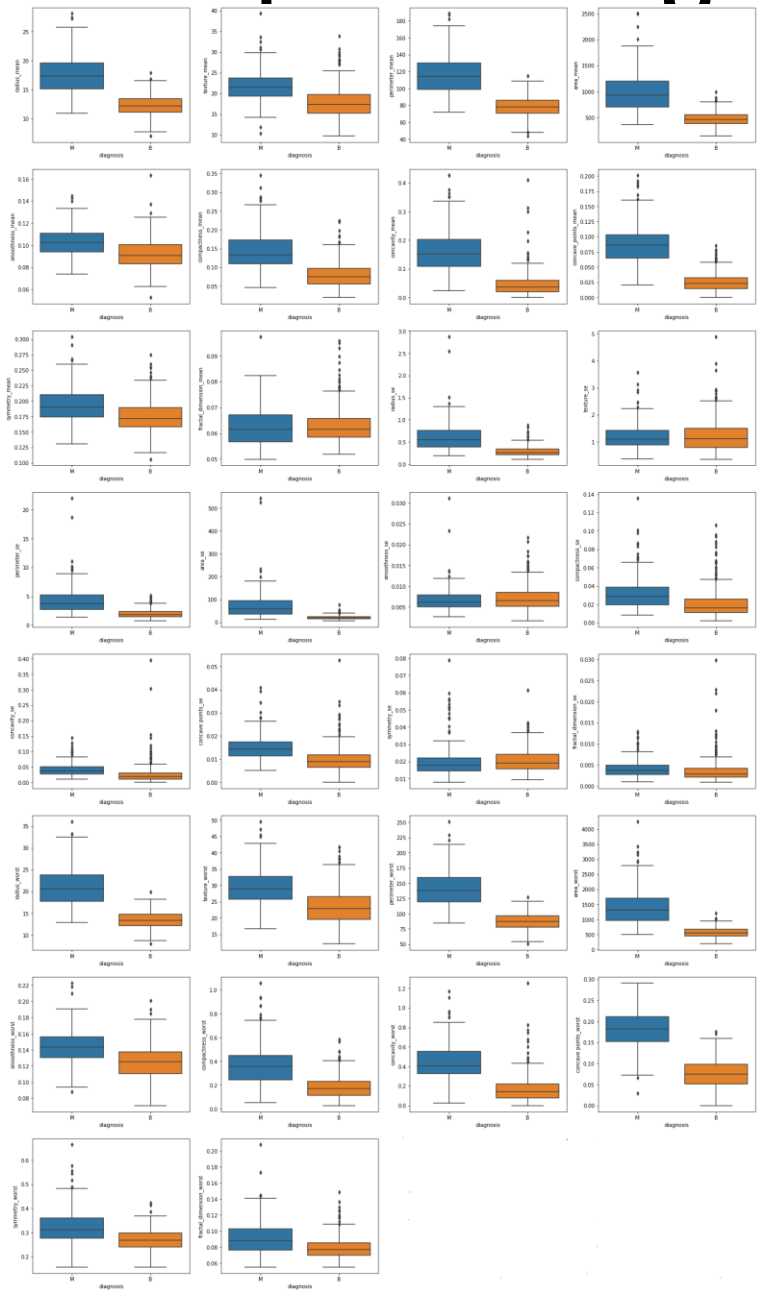


Insights:

1. [('perimeter_mean', 'radius_mean') has a very high correlation of 1
2. ('perimeter_worst', 'radius_worst')] has a correlation of .99

Tool Used:
Matplotlib, python

Boxplot Insights

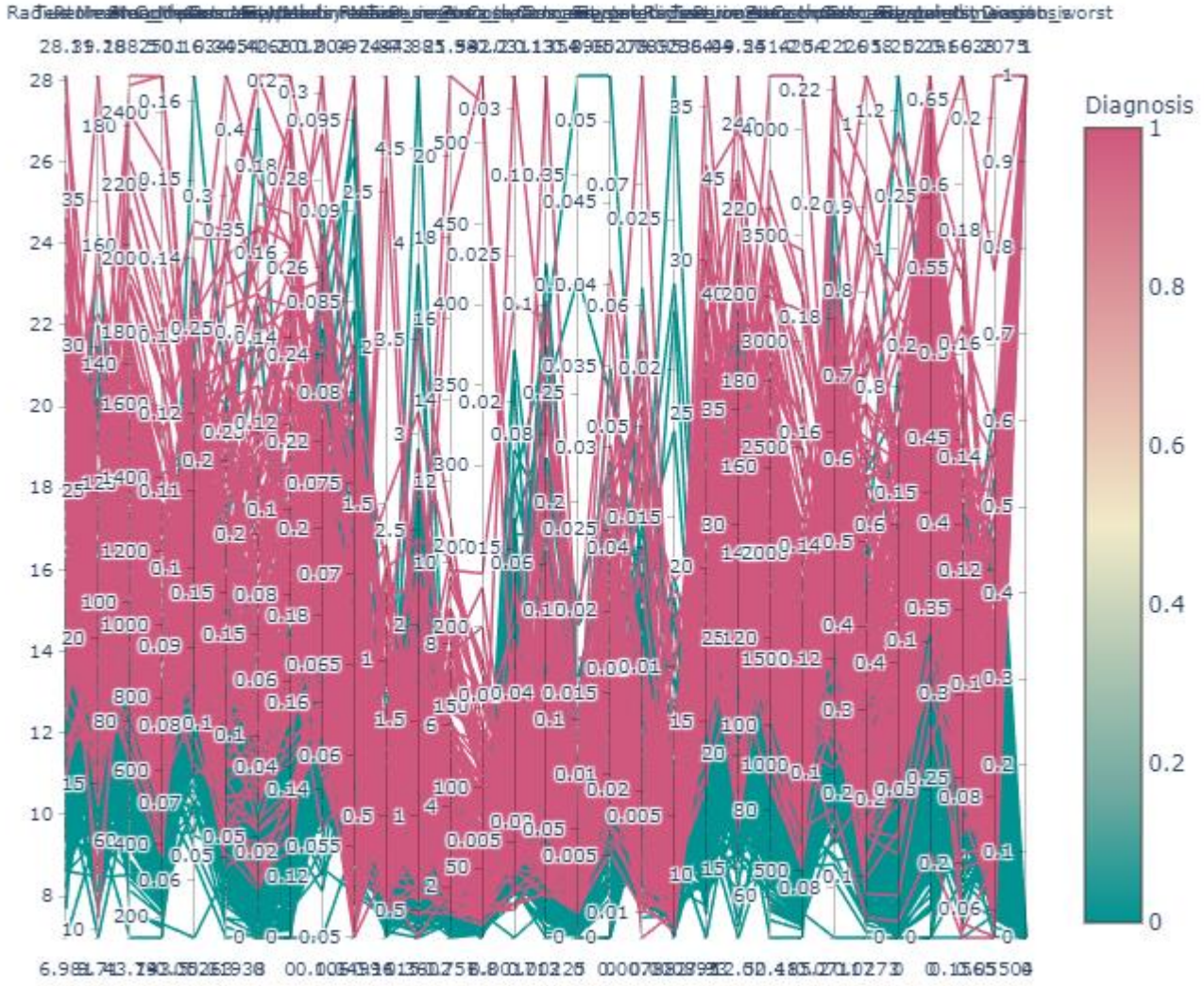


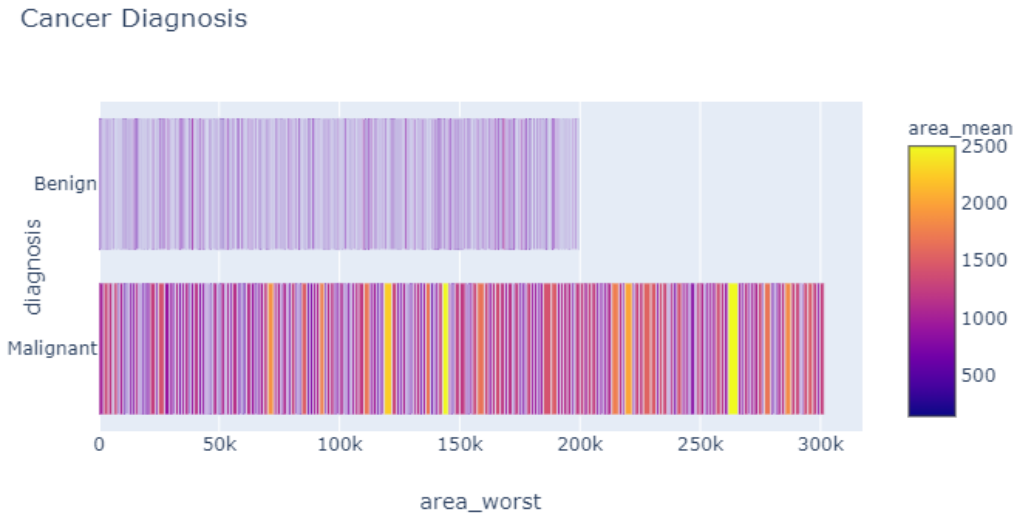
Insights:

1. All the attributes have outliers.

Tool Used:
matplotlib, python

Parallel coordinates and Horizontal bar chart

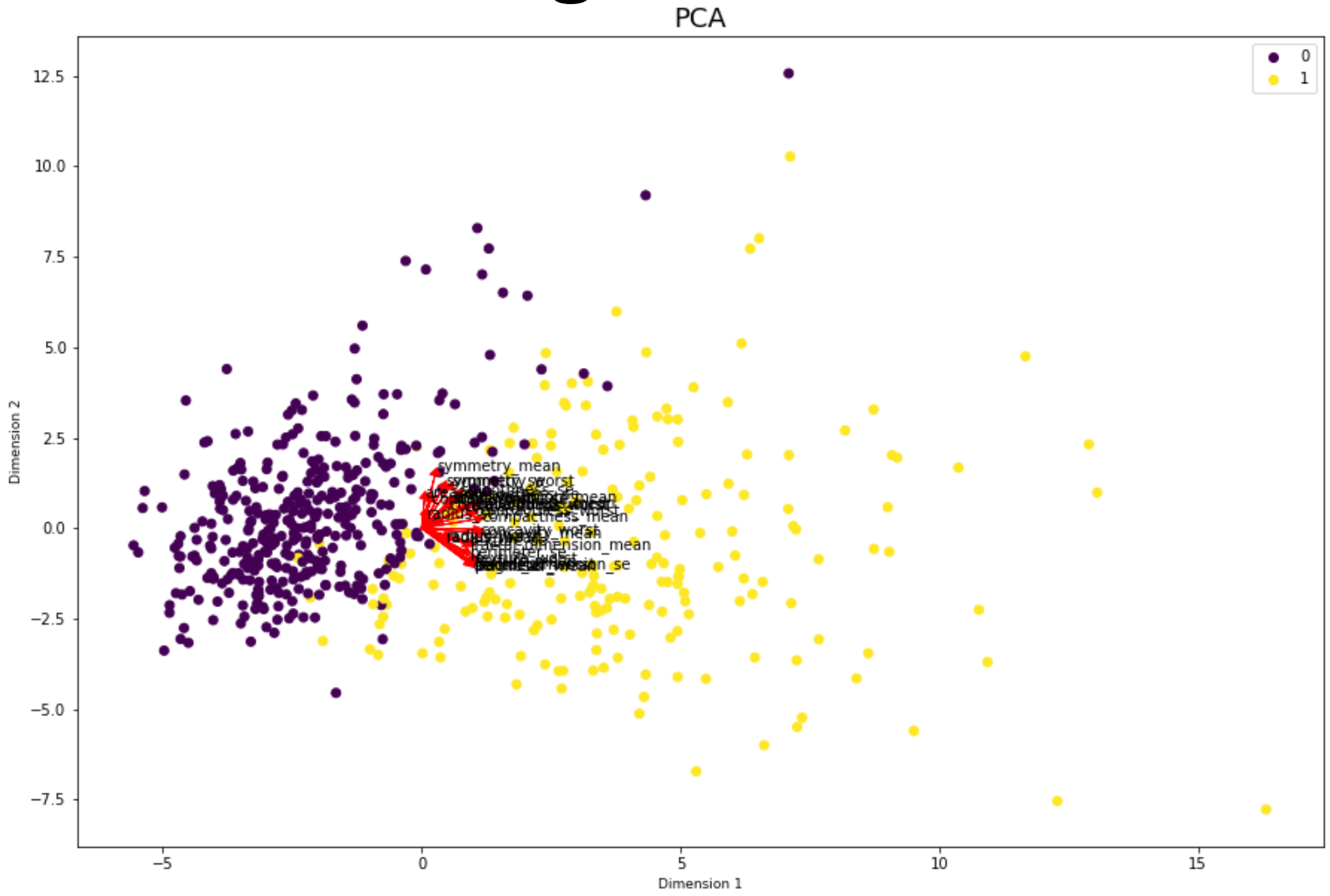




Visualizes the entire dataset

Tool Used:
Plotly, python

PCA Insights



Insights:

1. Shows the clustering of the dataset.
2. Benign class is well clustered than malignant class.

Tool Used:
matplotlib, seaborn, python

3. Human Activity Recognition Using Smartphones Data Set

[Dataset link]

No. of Instances: 10299 (50Hz)

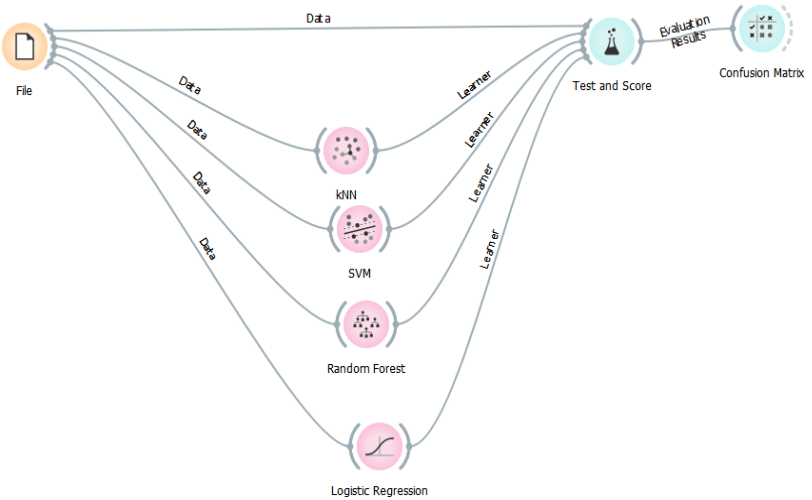
No. of Attributes: 561

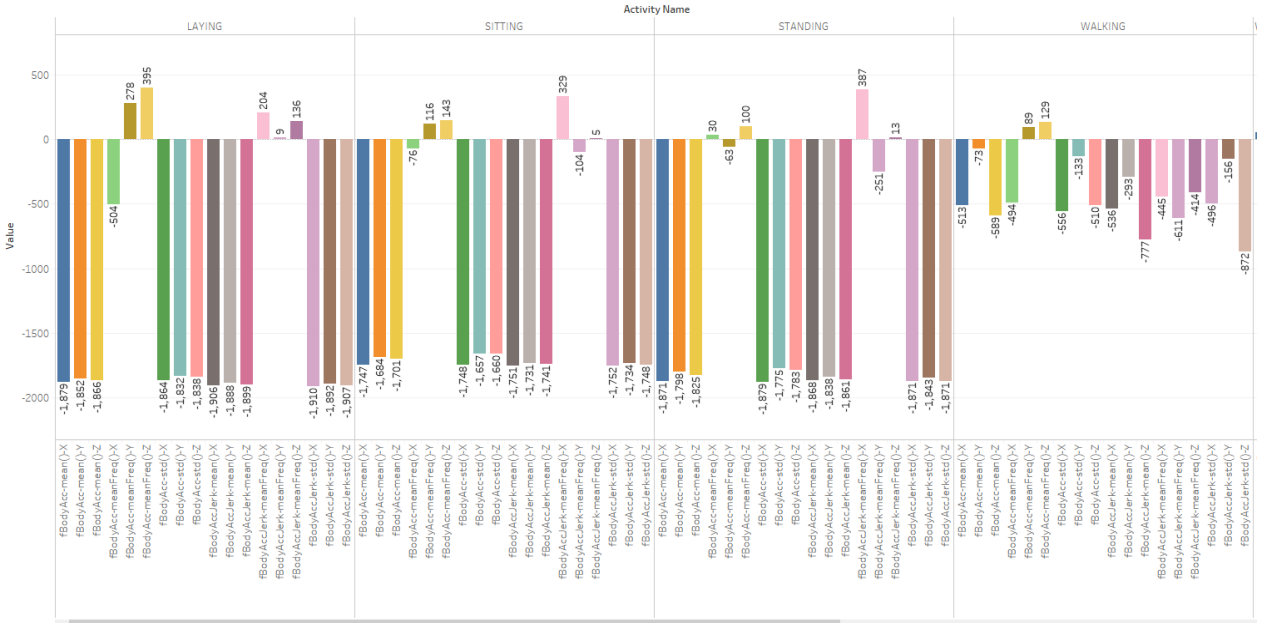
30 volunteers

Activities Classified:

WALKING
WALKING_UPSTAIRS
WALKING_DOWNSTAIRS
SITTING
STANDING
LAYING

Insight





Questions:

1. Why did you apply this/these visualization method(s)?
2. What kind of pattern(s) have you discovered?
3. What is your final conclusion?

4. A study of Asian Religious and Biblical Texts Data Set

Dataset Link

Dataset

8265 attributes

590 observations

Religion

Hinduism
Buddhism
Taoism
Christianity

Books

Yogasutras
Upanishads
Four Noble Truth of Buddhism
Tao Te Ching
Book of Proverb
Book of Ecclesiastes
Book of Ecclesiasticus
Book of Wisdom

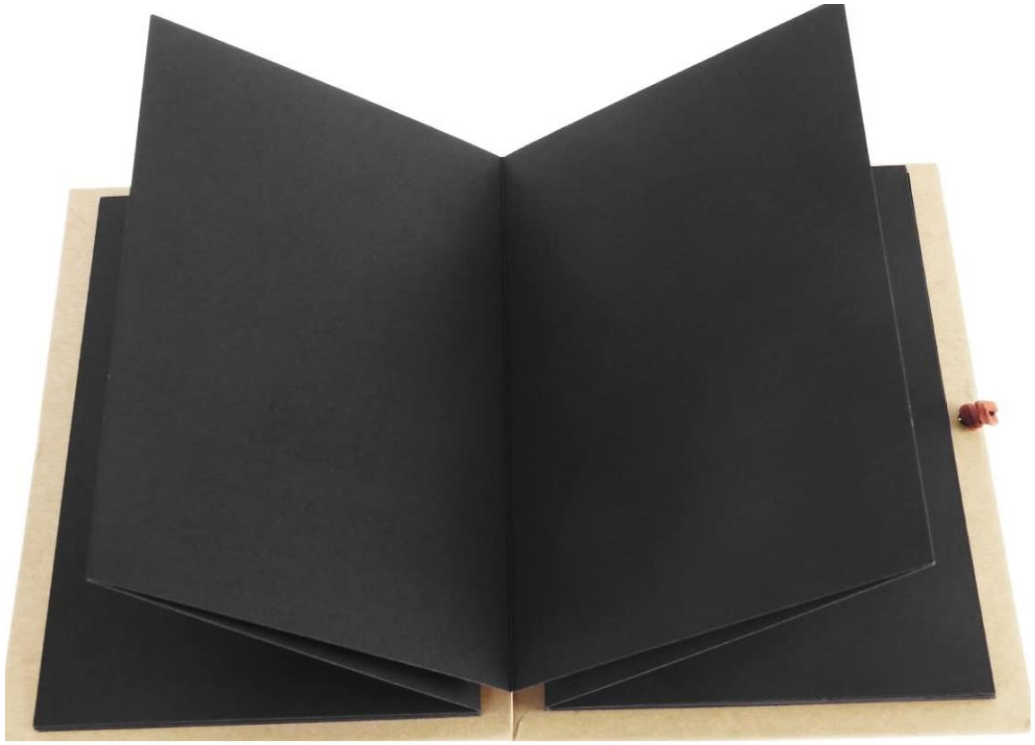
Word Cloud: Chapter-1Buddhism

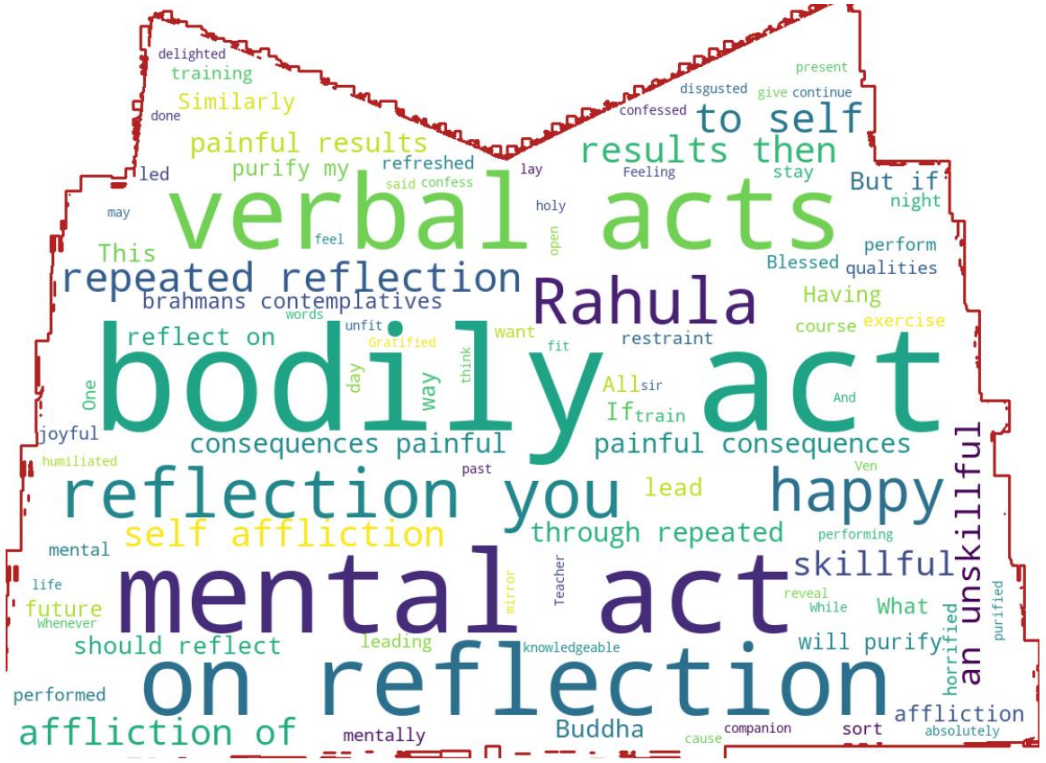
Insights:

1. Chapter 1 of Buddhism text have the most frequent word **acts(39),reflection(19)**

Tool Used: matplotlib, python

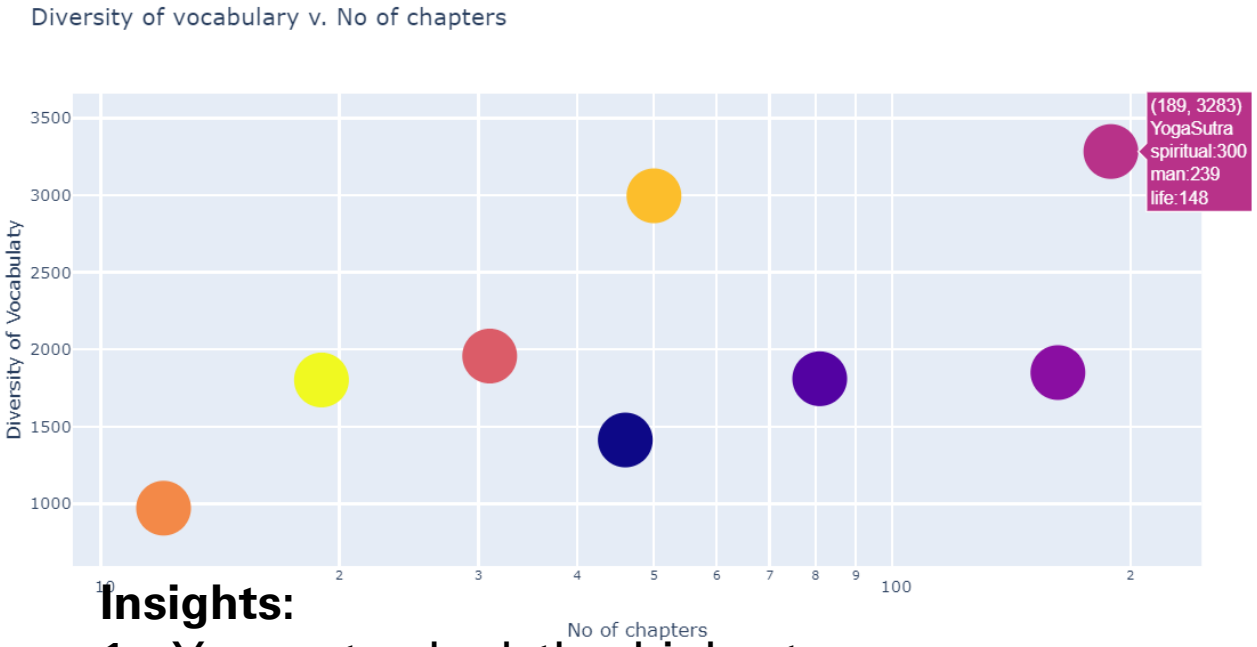
Word Cloud masked on a Image





Tool Used:
matplotlib, python

Relation between no of chapters and the diversity of vocabulary of a book

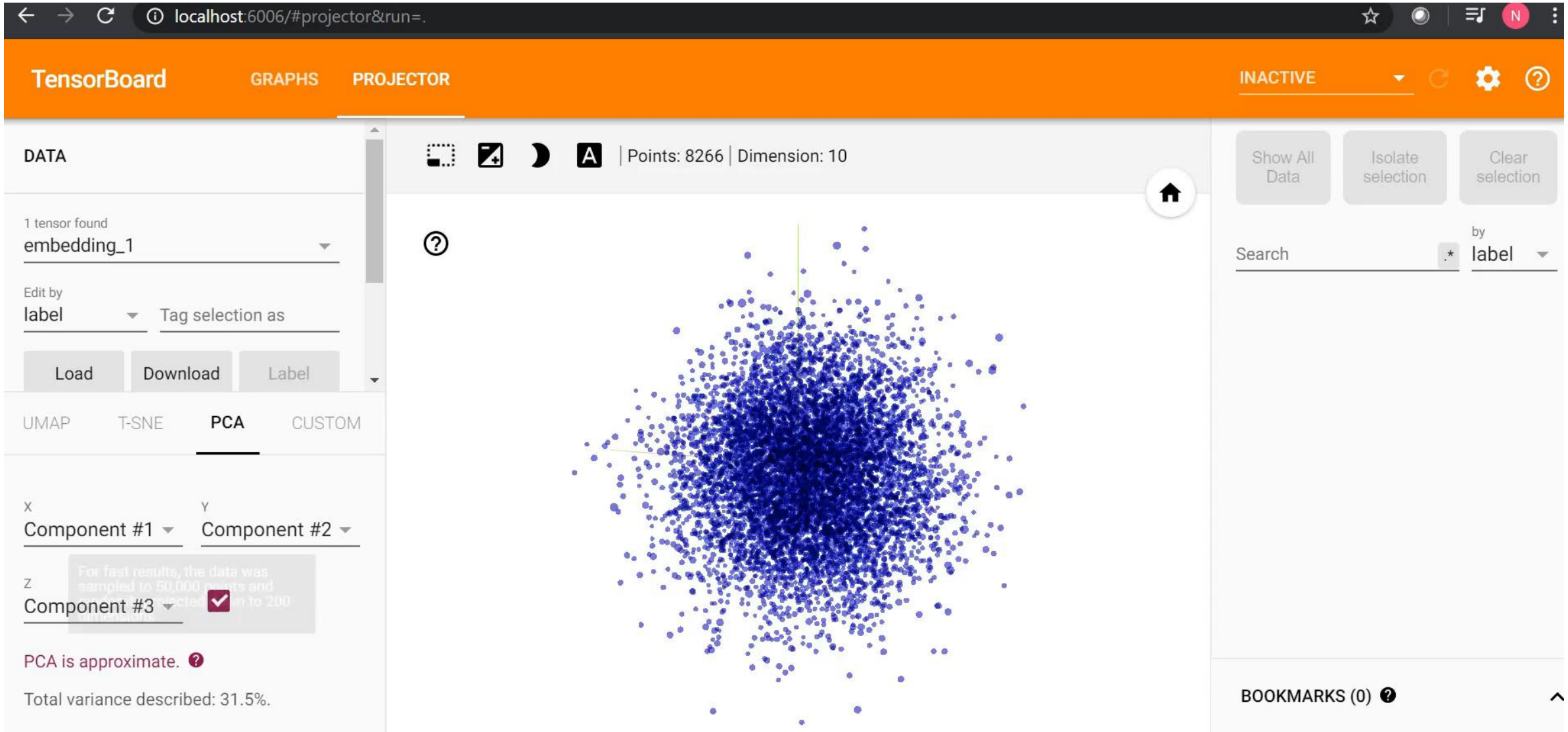


Insights:

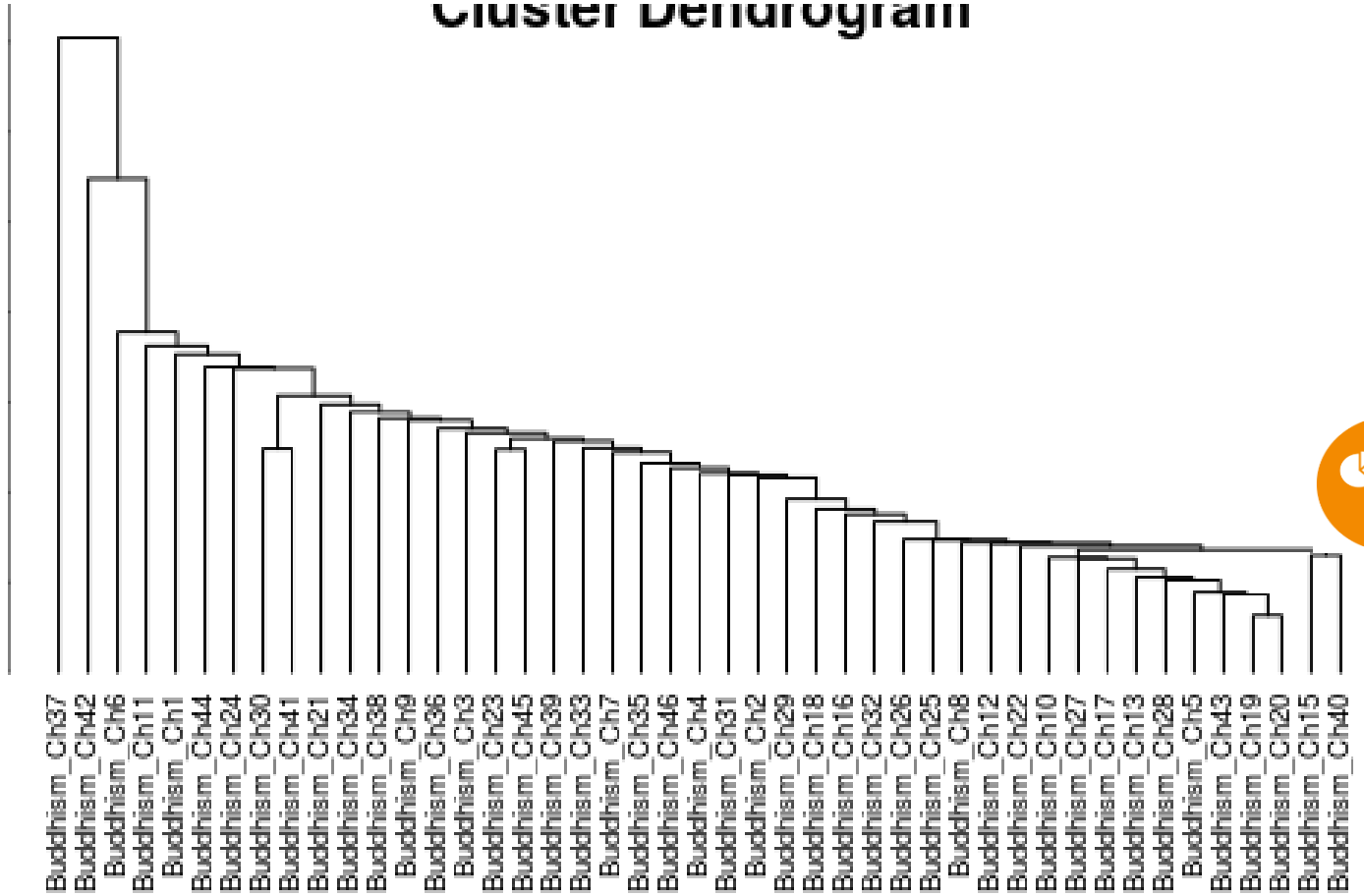
1. Yogasutra had the highest no of chapters and richest vocabulary.
2. Ecclesiasticus have only 50 chapters but very diverse corpus of 2996
3. Word **shall** is the most frequent word In the books for Christianity

Tool Used:
Plotly, python

Tensorboard visualization of vocabulary

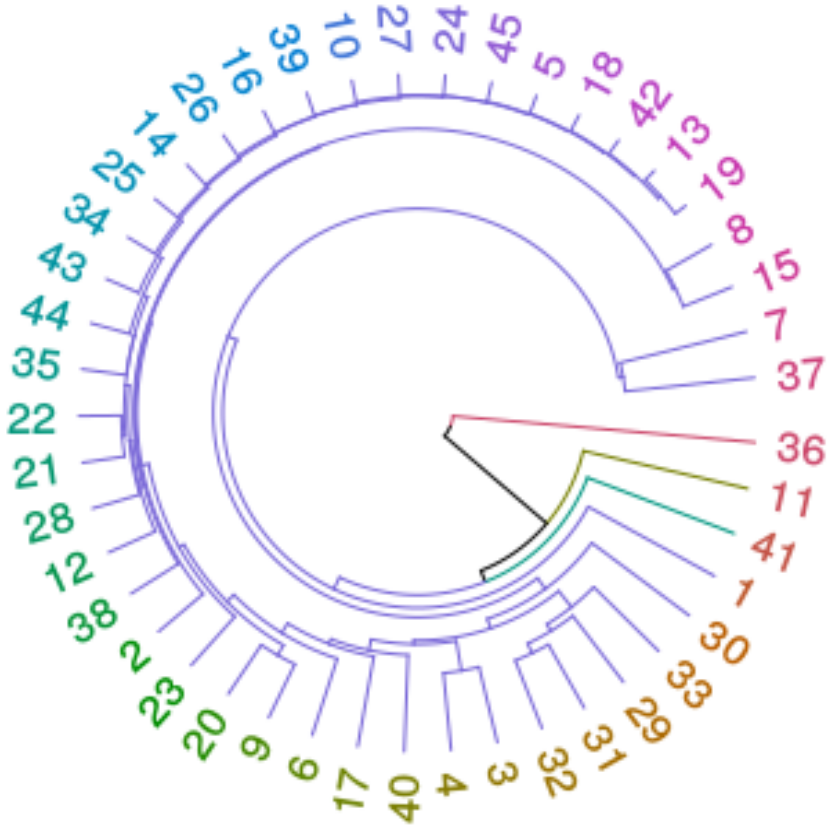


Hierarchical clustering Of Buddhism book chapters using dendrogram



Tool Used:
ggplot, R

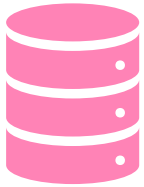
Hierarchical clustering Of Buddhism book chapters using radial plot



Tool Used:
ggplot, R

5. Student Performance Data Set

Dataset Link



Datasets- 2

Student Performance Results

In Math and Portuguese

In total

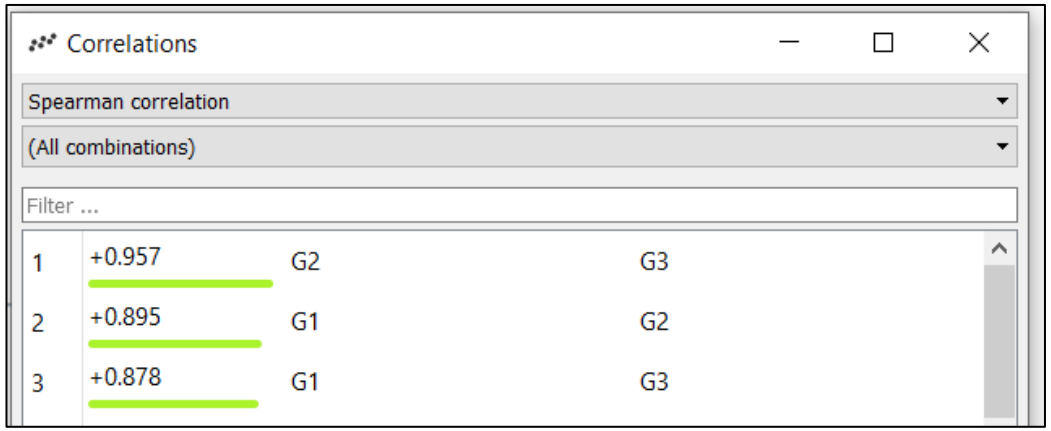
No of Attributes :- 33

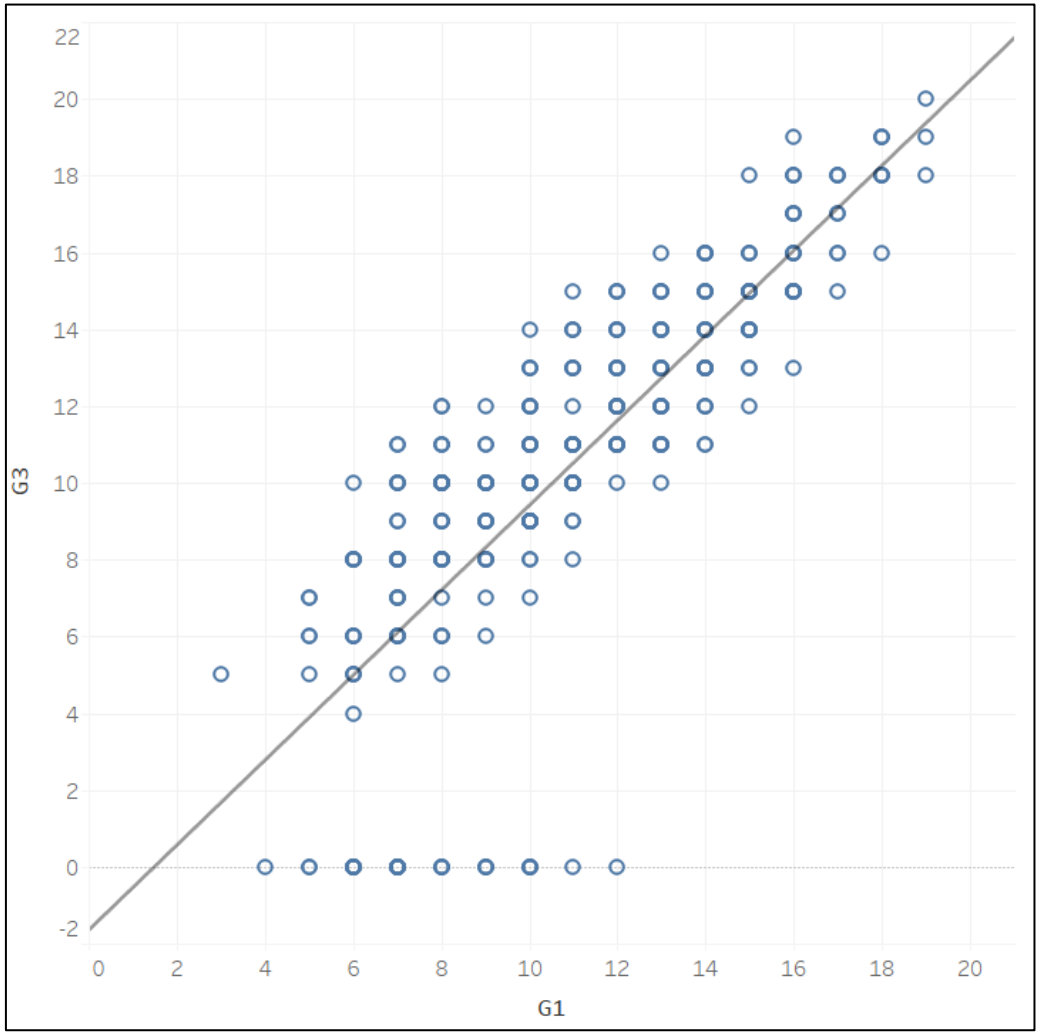
No of Instances- Math :- 395

No of Instances- Portuguese :- 649

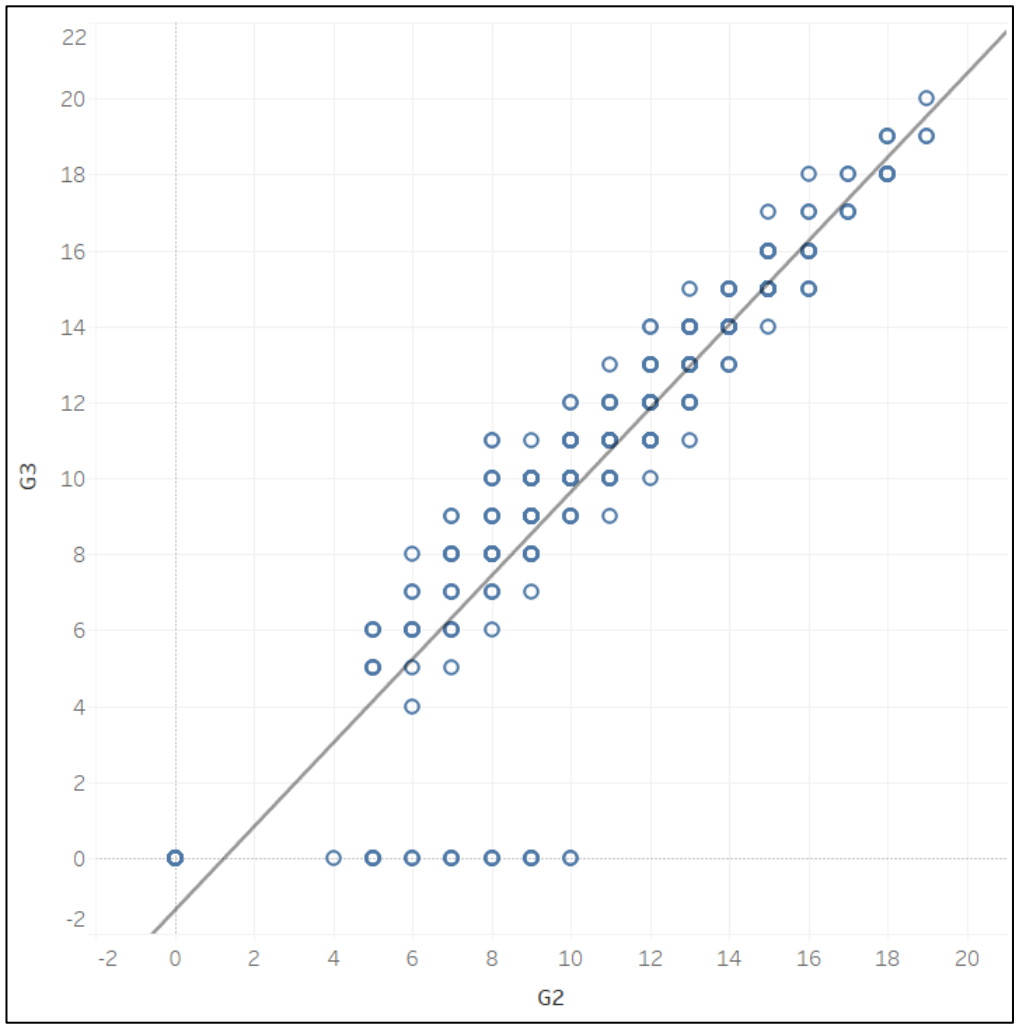
Insight for Math Dataset

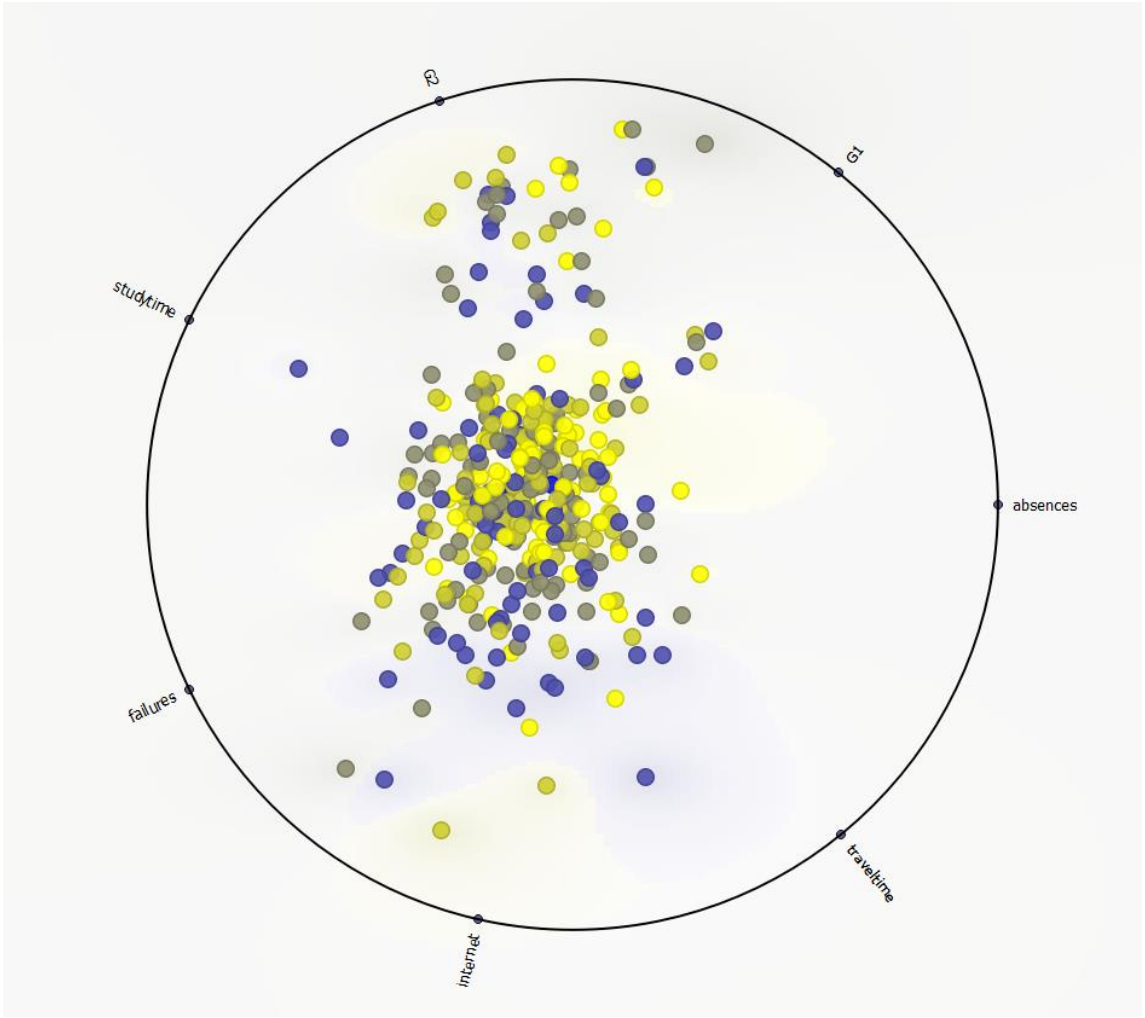
Tools Used:- Orange 3 (3.23.1) and Tableau



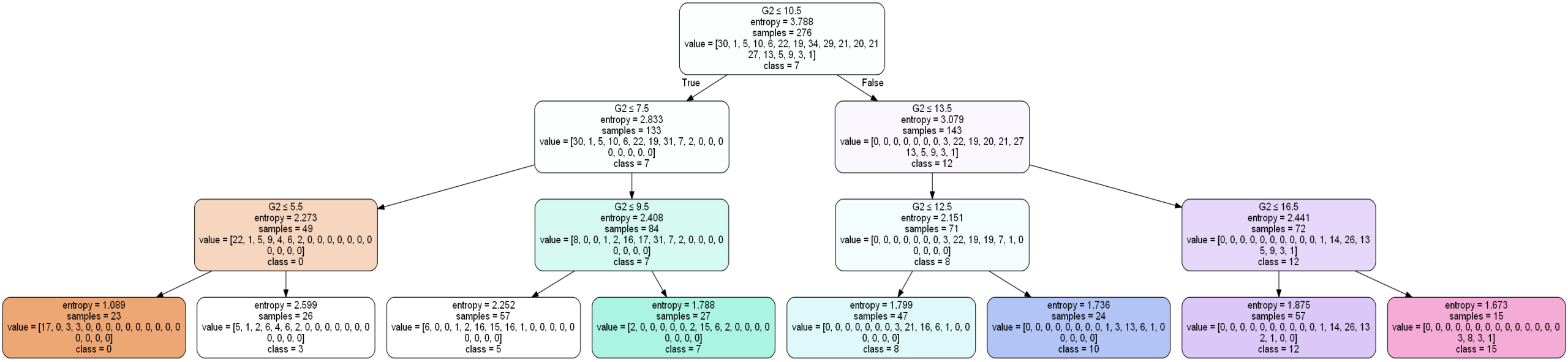


Insight for Math Dataset (cont..)

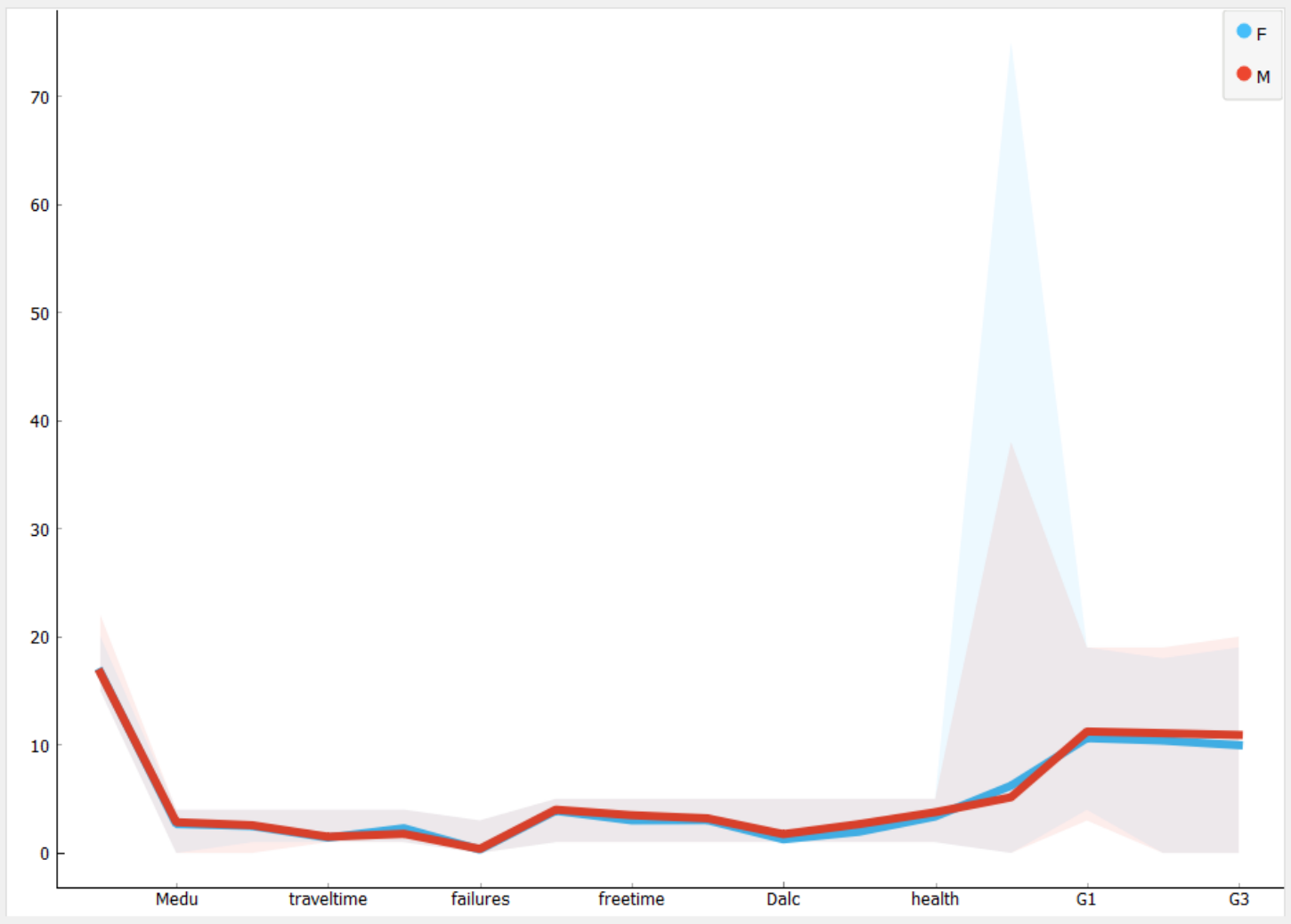




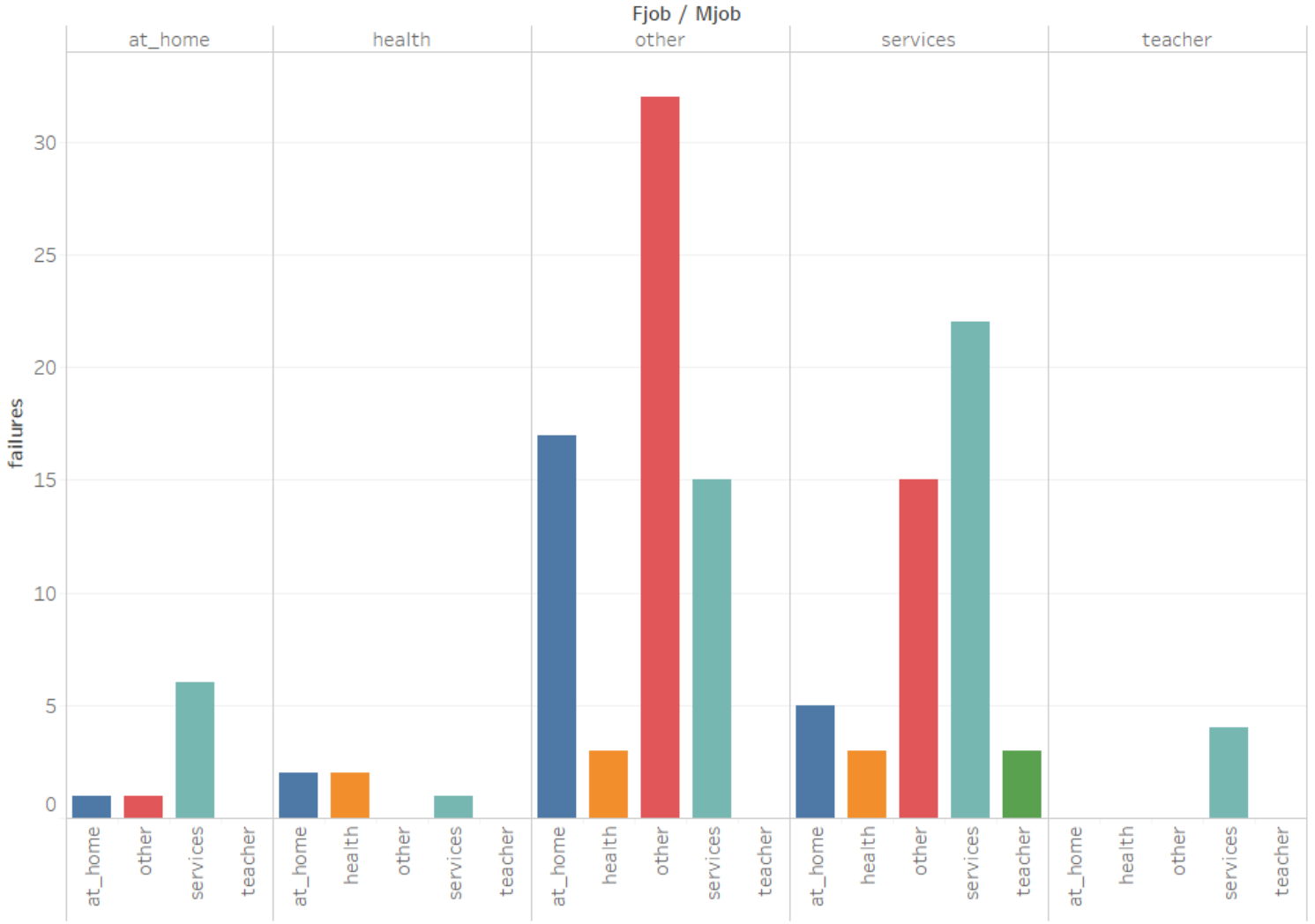
Insight for Math Dataset (cont..)



Insight for Por Dataset



Insight for Por Dataset



THANK YOU!