PRINCIPAL COMPONENT ANALYSIS

Group 12:

Jap Purohit : AU1940109

Nihar Patel : AU1940119

Mohit Prajapati : AU1940171

Raj Gariwala : AU1940118

Purvam Sheth : AU1940151

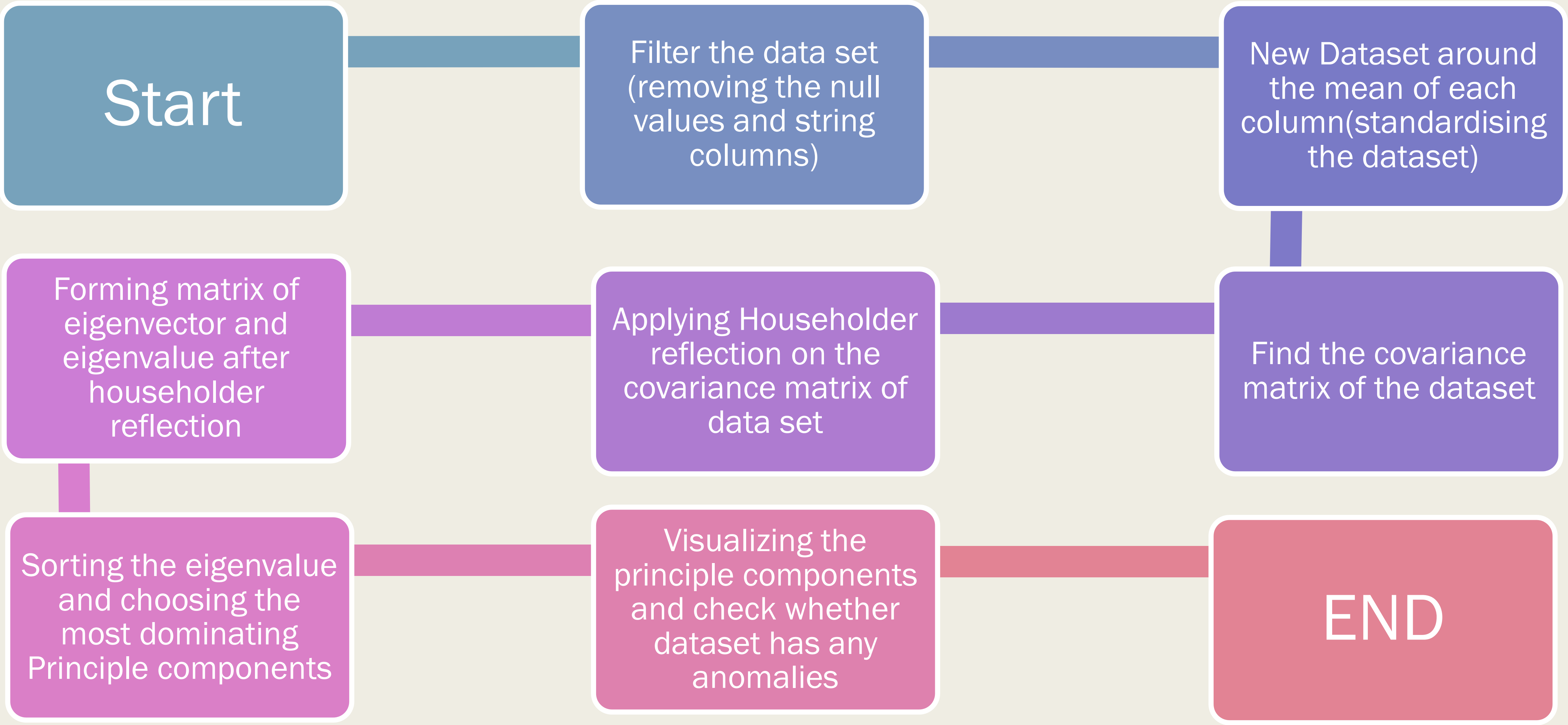
Introduction

- Principle Component Analysis is technique for dimensional reduction.
- It helps in summarizing the dataset which has large number of variables by reducing less affecting variables without loss of significant information
- There are many dimensional reduction method but PCA is considered on the most effective method.
- We have used Householder reflection for QR decomposition.

Steps for PCA

- **Step-1:** Standardize and clean the data-set.
- **Step-2:** Calculate the Covariance matrix of dataset.
- **Step-3:** Calculate eigenvalues and eigen vectors using Householder Reflection QR decomposition or any other method
- **Step-4:** Sorting of eigenvalues and their corresponding eigenvectors in decreasing order.
- **Step-5:** Consider the eigenvalues which contributes the most significant column(i.e. the eigenvalues with maximum value) which can found using proportion by variance.
- **Step-6:** The matrix obtained by the eigenvectors corresponding to these eigenvalues gives the Principle Component Analysis.

FLOW CHART



Start

Filter the data set (removing the null values and string columns)

New Dataset around the mean of each column(standardising the dataset)

Find the covariance matrix of the dataset

Applying Householder reflection on the covariance matrix of data set

Forming matrix of eigenvector and eigenvalue after householder reflection

Sorting the eigenvalue and choosing the most dominating Principle components

Visualizing the principle components and check whether dataset has any anomalies

END

Softwares used:

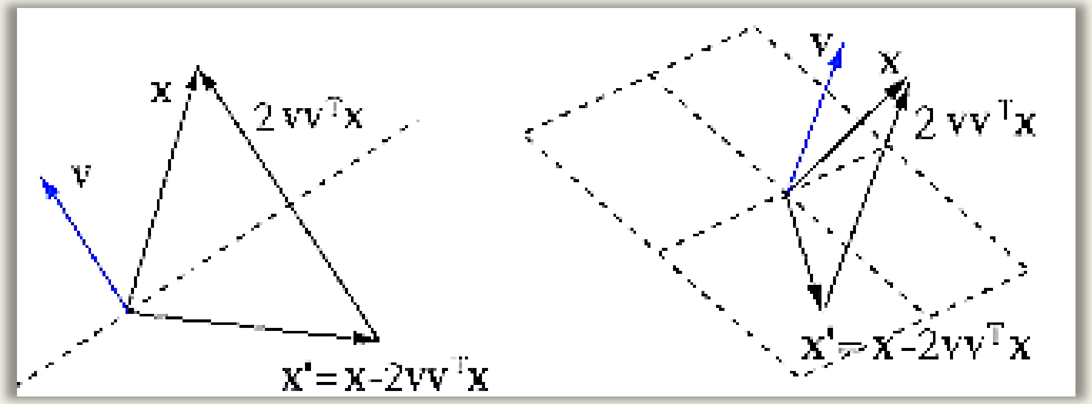
- **Coding the Principal Component Analysis in Python we used Google-Colab environment.**
- **Libraries of Python like: Matplotlib, NumPy, Pandas are used.**
- **The file will be in the form of (Python notebook) ipynb.**
- **Reason of usage of this python libraires**
  - ***Pandas used for reading and extracting data from the provided datasets.***
  - ***NumPy used for building arrays.***
  - ***Matplotlib is used for plotting the graph.***

Approach Used

- QR Decomposition using householder reflection

Householder reflection is one of the important aspects in finding QR decomposition. Householder reflection technique is used in finding reflection of any vector in any dimension which respect to any plane in the given subspace. **H = (*I* – 2(vv^T))**, this equation is known as householder reflection equation, where **v** is the orthonormal vector to the plane and **H** is the Householder reflector matrix.

$$\begin{aligned} x' &= Hx \\ &= x - 2(xv)v \\ &= x - 2v(xv) \\ &= Ix - 2v(v^T x) \\ &= Ix - 2(vv^T)x \\ &= (I - 2(vv^T))x \end{aligned}$$

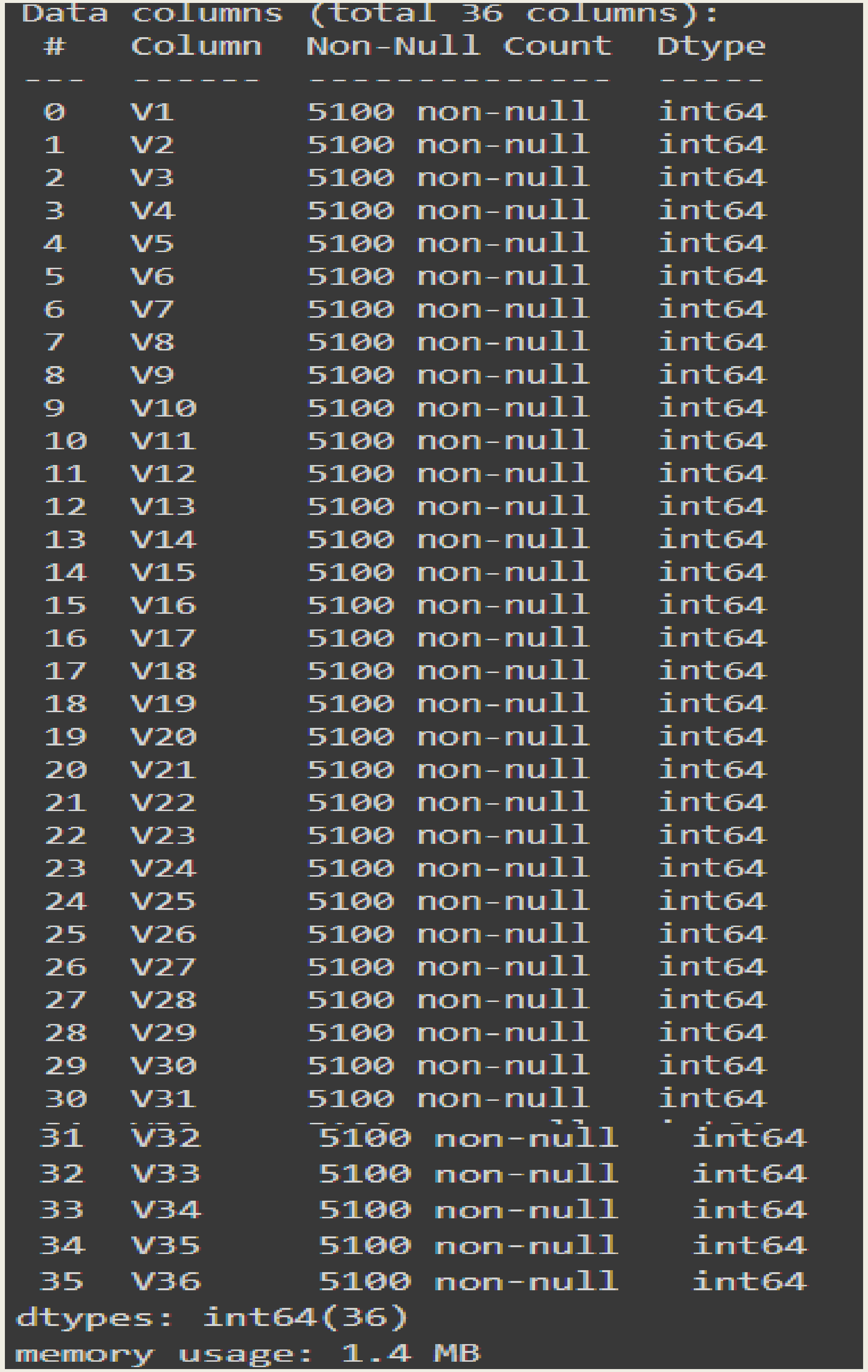


Representation of vector and it’s reflection with respect to a plane.

About the Dataset

- The satellite dataset comprises of features extracted from satellite observations. In particular, each image was taken under four different light wavelengths, two in visible light (green and red) and two infrared images. The task of the original dataset is to classify the image into the soil category of the observed region.
- The dataset has 36 columns and 5100 rows collected from satellite observations. And each column is pixel of the image.

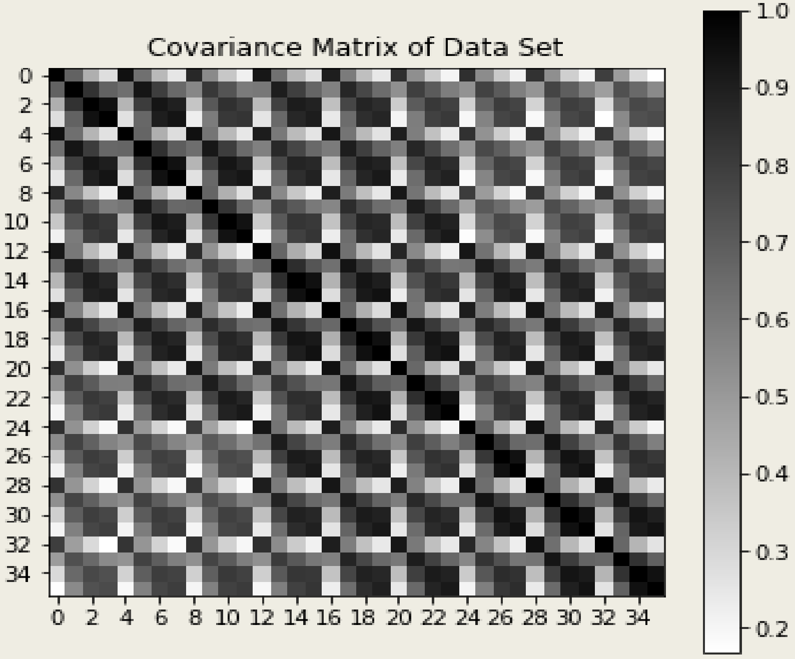
Source: https://www.openml.org/d/40900



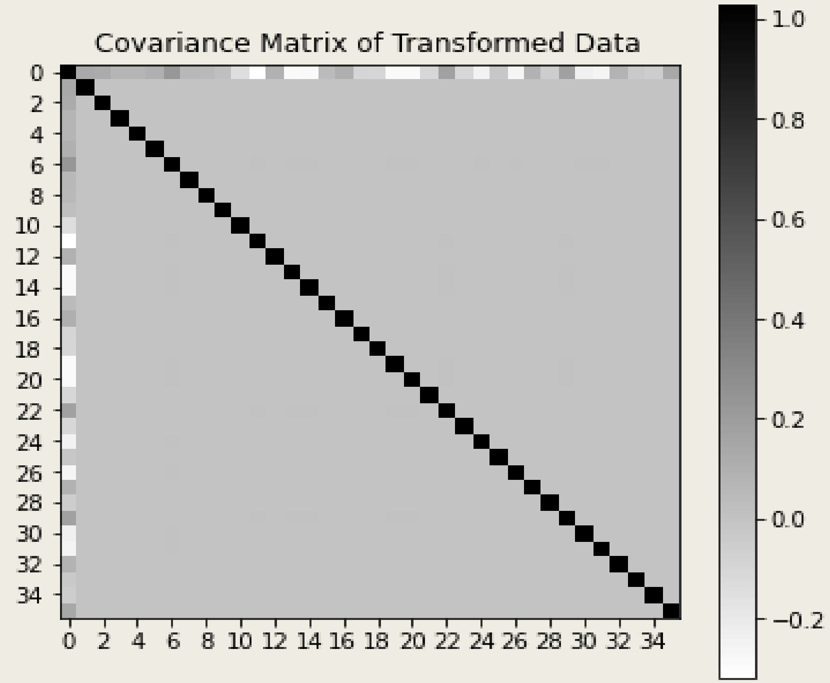
```
Data columns (total 36 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   V1      5100 non-null   int64
 1   V2      5100 non-null   int64
 2   V3      5100 non-null   int64
 3   V4      5100 non-null   int64
 4   V5      5100 non-null   int64
 5   V6      5100 non-null   int64
 6   V7      5100 non-null   int64
 7   V8      5100 non-null   int64
 8   V9      5100 non-null   int64
 9   V10     5100 non-null   int64
 10  V11     5100 non-null   int64
 11  V12     5100 non-null   int64
 12  V13     5100 non-null   int64
 13  V14     5100 non-null   int64
 14  V15     5100 non-null   int64
 15  V16     5100 non-null   int64
 16  V17     5100 non-null   int64
 17  V18     5100 non-null   int64
 18  V19     5100 non-null   int64
 19  V20     5100 non-null   int64
 20  V21     5100 non-null   int64
 21  V22     5100 non-null   int64
 22  V23     5100 non-null   int64
 23  V24     5100 non-null   int64
 24  V25     5100 non-null   int64
 25  V26     5100 non-null   int64
 26  V27     5100 non-null   int64
 27  V28     5100 non-null   int64
 28  V29     5100 non-null   int64
 29  V30     5100 non-null   int64
 30  V31     5100 non-null   int64
 31  V32     5100 non-null   int64
 32  V33     5100 non-null   int64
 33  V34     5100 non-null   int64
 34  V35     5100 non-null   int64
 35  V36     5100 non-null   int64
dtypes: int64(36)
memory usage: 1.4 MB
```

Simulation and Output

- Covariance of Data before PCA

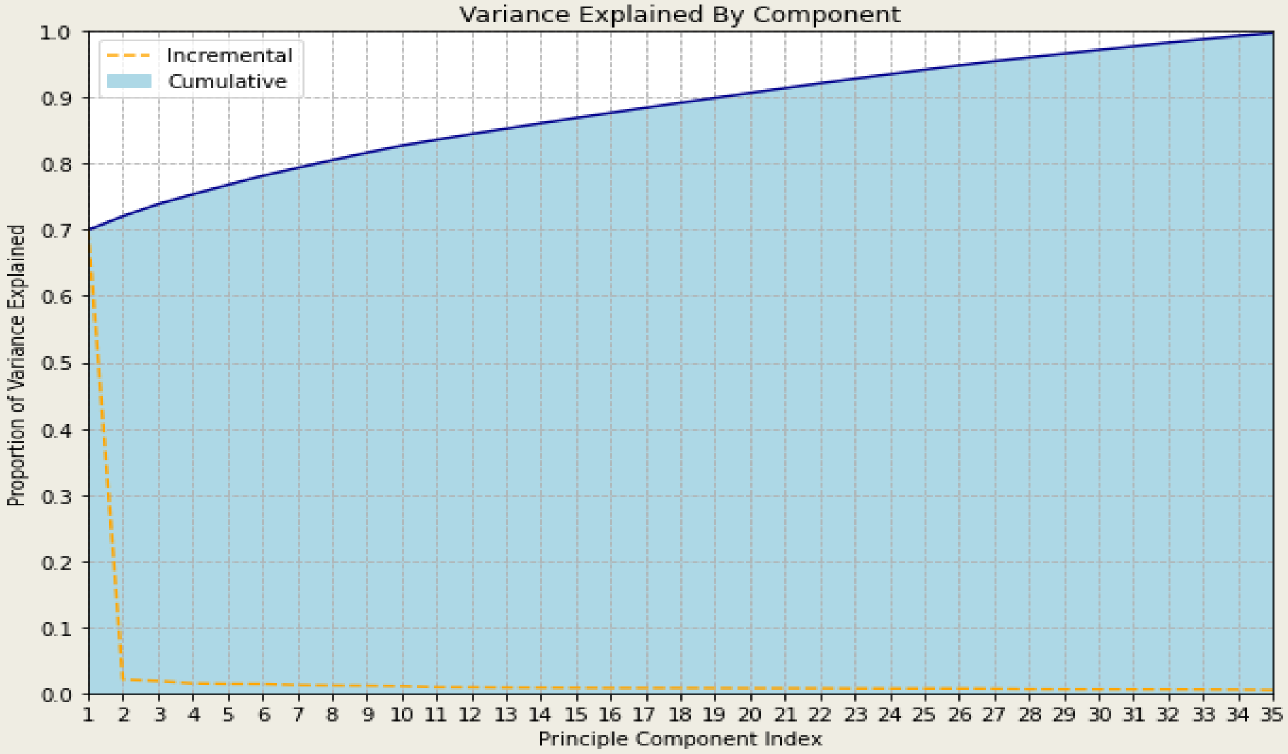


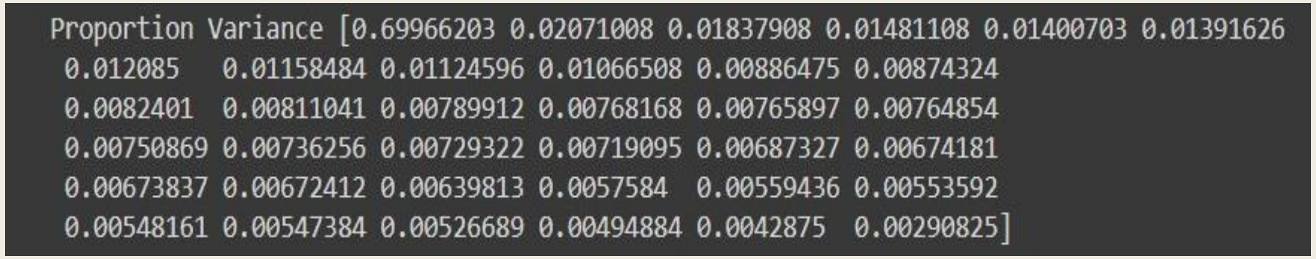
Covariance of Data after PCA



Simulation and Output

- Variance Explained of PCA



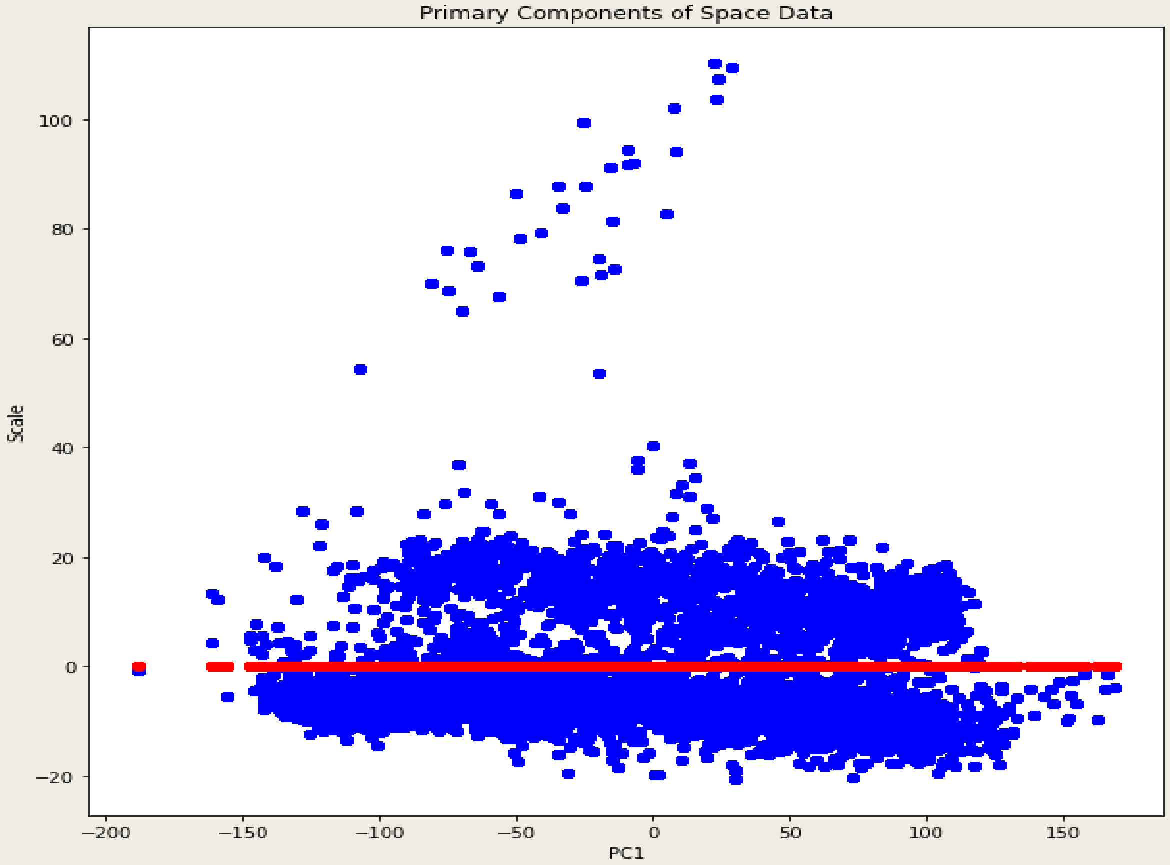


```
Proportion Variance [0.69966203 0.02071008 0.01837908 0.01481108 0.01400703 0.01391626
 0.012085   0.01158484 0.01124596 0.01066508 0.00886475 0.00874324
 0.0082401  0.00811041 0.00789912 0.00768168 0.00765897 0.00764854
 0.00750869 0.00736256 0.00729322 0.00719095 0.00687327 0.00674181
 0.00673837 0.00672412 0.00639813 0.0057584  0.00559436 0.00553592
 0.00548161 0.00547384 0.00526689 0.00494884 0.0042875  0.00290825]
```

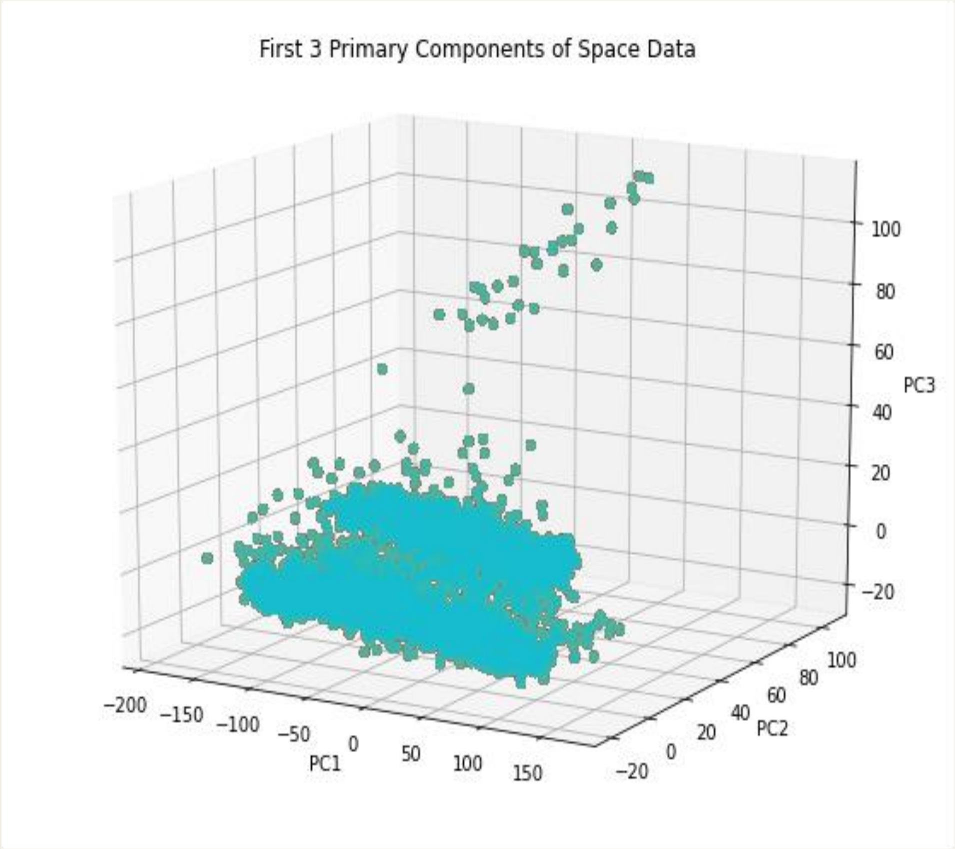
We can see that after performing the PCA we have significance of around 70%. And maximum data variance is across the principle component one.

Simulation and Output

- Graphs along Principle Component 1



Following graph is across 3 principle component



From the graph we can see that the dataset has some outliers

Conclusion

- On a closing note this project was a great chance for us to learn the principles of linear algebra such as QR Decomposition, Householder Reflection.
- In addition to the fact that we learned substantially more about them and applied it. This was an extraordinary hand on movement for us and we saw how the ideas are applied, considering all things.
- In addition to this, there are many other elements of linear algebra that we still have to discover and apply.
- But apart from that, engaging with the team was a remarkable opportunity, and getting their input on how to go forward to tackle a specific challenge and eventually finished the project.