Final Project

Miss Oriade Latifah Simpson

Spring 2022

Group Number 20

Introduction

html_document ioslides_presentation. 10 slides in 10 minutes.

Here I communicate insights to stakeholders and explain why I decided on that particular path in the analysis.

During the project I aimed to reproduce results from the West et. al paper. In this process, processed raw data was processed in preparation for data analysis. There were certain decisions made with regard to data processing. The reasons for these decisions are mentioned below. There were many challenging problems in coding to which I found an elegant solution. For example:

Materials and Methods

The data used was microarray data contained in an .RData file in the following Github repository: https://github.com/ramhiser/data microarray/blob/master/data/west.RData

The R code was separated into six scripts and the HPC Cluster format outlined in the instructions was used. This bio data set required loading, cleaning and tidying, augmenting, modelling and visualisation.
The process of arriving at results in a reproducible manner is important, especially was the scripts were run several times after changes were made. The code in the scripts was refined by use of functions and these functions were bundled up into a package. The functions were made to aid reproducibility.

There is a flow chart of the data journey

Over thirteen packages were used to aid in this analysis pipeline, including packages in the tidyverse, golem, reactable and packages for styling the code such as the stylr package.

Results

As mentioned, t he data was processed several in stages. (**inert plots here**) During the first analysis there was a lot of noise and so the analysis was run again.

The genes that were not significant enough were filtered out and re plotted.

Principal component analysis simplifies the complexity in high-dimensional data by transforming the data into fewer dimensions, which act as summaries of features.

PCA is an unsupervised learning method which finds patterns without reference to prior knowledge about whether the samples come from different groups.

Results for PCA and K-means clustering were produced using ggplot2 and ggthemes packages. In comparison with the research paper, the visualisation was slightly improved in terms of the colour scheme chosen and the size of the points.

In addition to this Github repository and the powerpoint

Discussion

It was computationally intensive to perform the modelling part of the analysis given there were 7130 features and 49 observations of data. Because of this computational expense I found it difficult to find a workaround to deploy the application online.