CSE 564: Visualization Final Project

Venkata Ravi Teja Takkella
113219890
venkataravite.takkella@stonybrook.edu

Background:

Evaluating an individual's risk of drug consumption and misuse is highly important. One might as well ask how this can be done? We can try to use an individual's personality traits and his/her background information to calculate the probable risk of drug consumption. But the linking of these traits to the risk is an enduring problem. Researchers do return again and again coming up with new data and questions.

But how do we calculate the personality traits of an individual firstly? There are numerous methods and scales to calculate this. They basically contain the person's

preferences, mannerisms and behaviours. The Five Factor Model (FFM) is one of the most popular tests and it provides the scores of the below five traits. The subtraits mentions an individual with a high score.

- Conscientiousness
    1. keep things in order
    2. are goal-driven
- Agreeableness
    1. are always ready to help out
    2. believe the best about others
- Neuroticism
    1. often feel vulnerable or insecure
    2. struggle with difficult situations
- Openness
    1. enjoy trying new things
    2. be more creative
- Extraversion/Extroversion
    1. make friends easily
    2. speak without thinking

A dashboard mentioning the scores of these traits and background information can really help us in obtaining beautiful insights about the risk of drug consumption and misuse. This might be really helpful and can help save lives by learning it early. So our main agenda of this project is to visualise our data onto a dashboard using plots and charts meaningful and derive any insights found.

Problem:

Firstly, each data tells us a different story. It's our job to pull out the information from it and we call it the Information Gain. There are a lot of Data Science models to pull out this information in a plethora of methods. But still some human expertise is required to find the exact information which is interesting from the user's point of view. This requires to visualise the data into plots and graphs for getting a better vantage point of the information. So this is one of the major problems which we need to solve by visualizing our data.

Coming to our specific problem, how do we know that one's personality, age, gender or nationality affect the risk? And is this risk different for different drugs? For example, does the risk of consumption of heroin and the risk of consumption of meth differ for different personality profiles? And do these traits follow a pattern which can be taken as the generalised personality for a specific drug consumer? And out of all these available traits and background values, which category is the most influential in regular consumption? These questions are the focus of this project. And we will try to gain information by visualising the data into the plots which were discussed in class.

Data:

Link: https://bit.ly/3fR2Cip

UCI contains a dataset containing information about 1885 individuals by questioning them in person about their personality traits, background information and rate of consumption of a few sets of drugs. For each person, there are around 12 personality and background attributes have been asked. These contain the scores of the FFM model along with "impulsivity" and "sensation seeking". It also contains the background information like level of education, age, gender, country of residence and ethnicity.

Finally they were also asked about the consumption of around 18 legal and illegal drugs. These contain alcohol, amphetamines, amyl nitrite, benzodiazepine, cannabis, chocolate, cocaine, caffeine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, mushrooms, nicotine and volatile substance abuse and one fictitious drug (Semeron) which was introduced to identify over-claimers.

They answer by choosing one from these values : "Never Used", "Used over a Decade Ago", "Used in Last Decade", "Used in Last Year", "Used in Last Month", "Used in Last Week", and "Used in Last Day".

Even though being categorical variables, most of this data is quantitative with some pre-mentioned values. These can still be categorized as categorical as they follow some specific numbers.

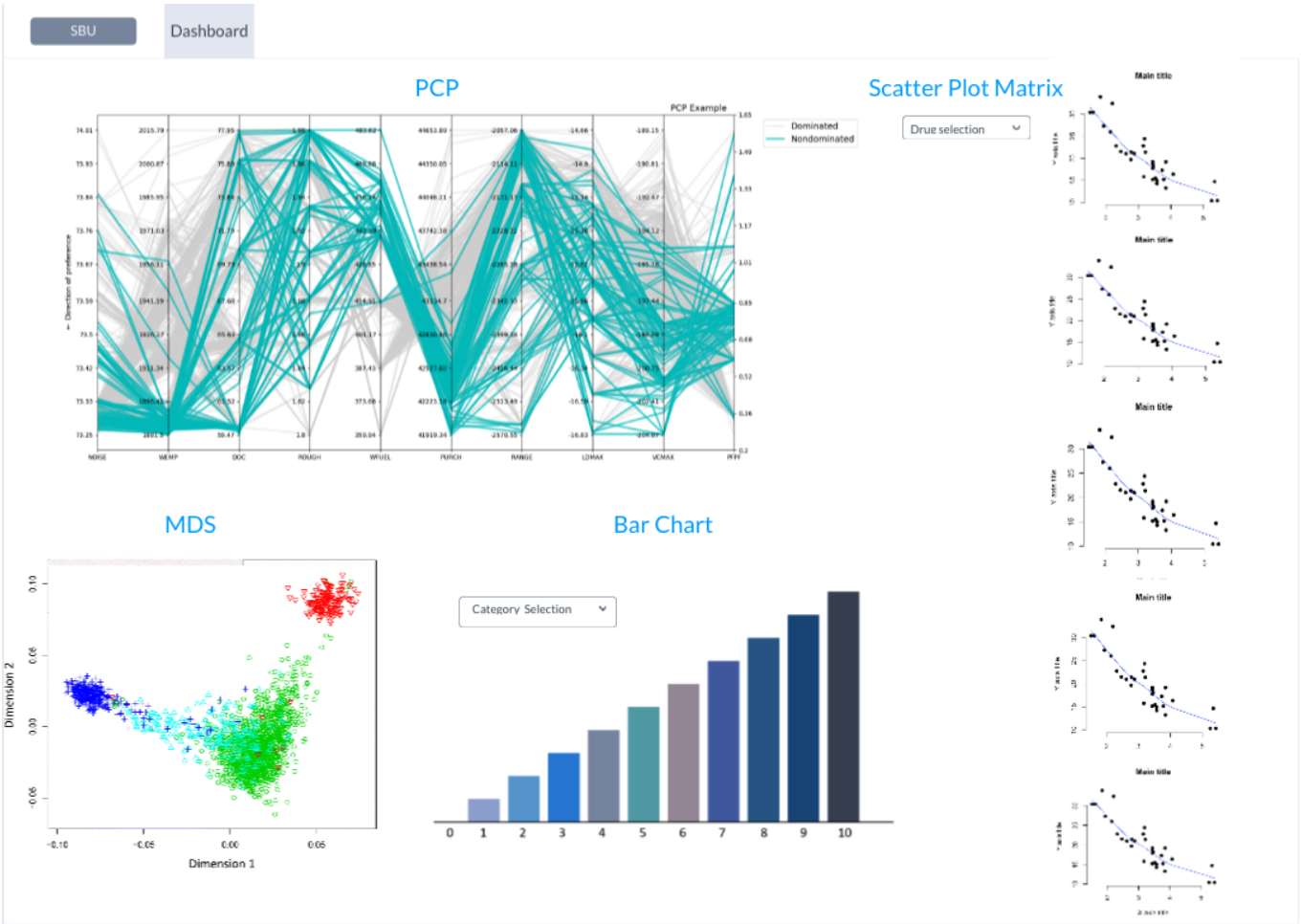
Approach:

Design:

I plan to visualise the data in the below mentioned design.

a) A bar chart showing the count of individuals from each provided category like drug consumer or from a specific location e.t.c.
b) A PCP plot showing all the given data points using K Means can provide interesting observations about the traits of being a drug consumer.
c) A Scatter plot matrix for each drug along with the FFM model traits can provide information which trait is influencing the more to consume the specific drug.
d) A MDS plot with K-means to properly show the entire data clusters on a single plot.
e) Brushing on the PCP or Scatter Matrix plot to provide corresponding information on the other plots too.
f) Interactivity between the graph elements and user input.

So the final dashboard would look like this.



Methodology:

1. Python - for data cleaning, pre-processing.
2. Flask - for data exchange between back-end exposed via various endpoints and front-end to consume to draw various charts on the UI.
3. D3 - for data visualization.
4. HTML, CSS and Javascript
5. Bootstrap