CSE 564: Visualization Final Project

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

1) Proposal Report:

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.

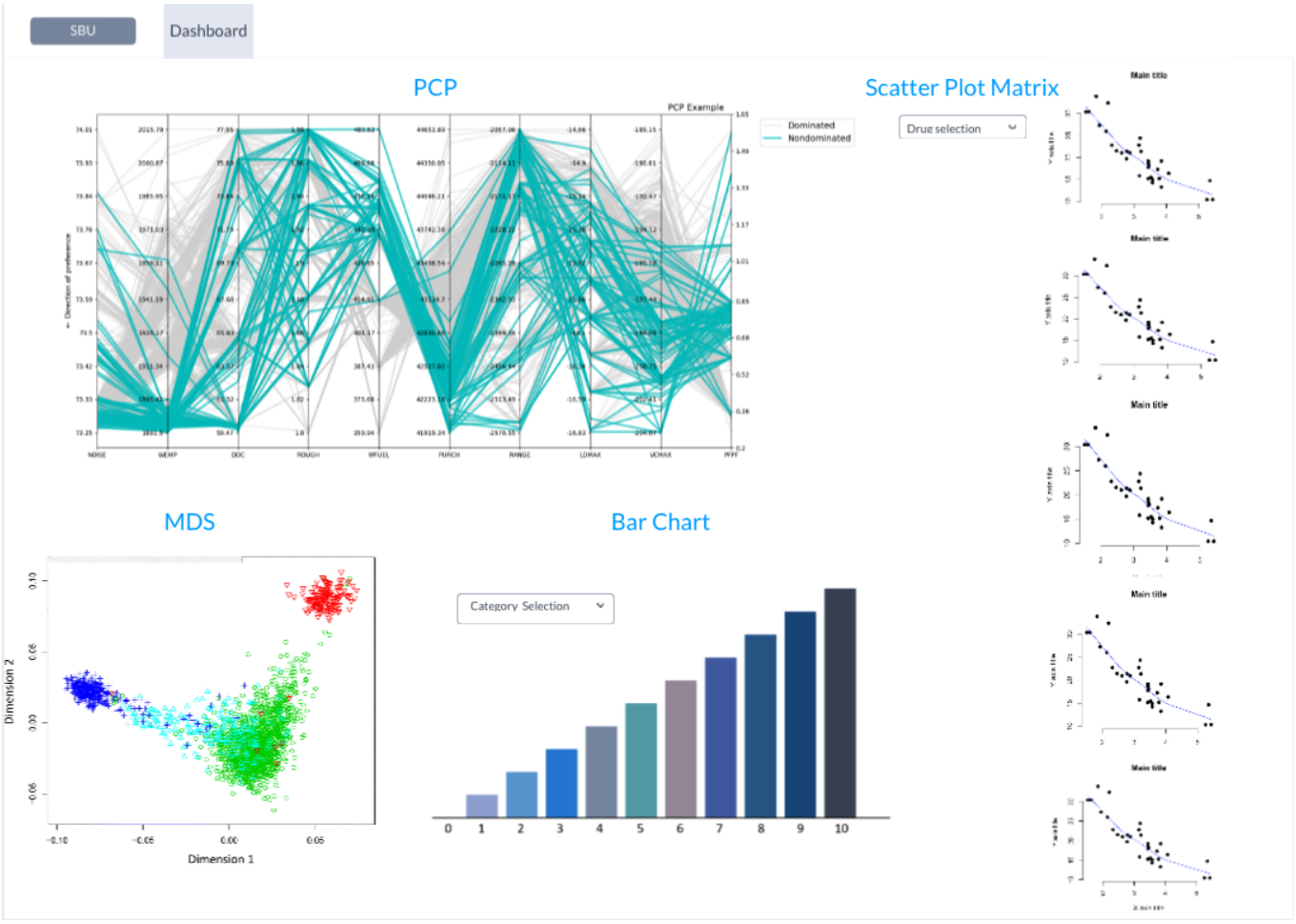
Approach:

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

2) Prelim Report:

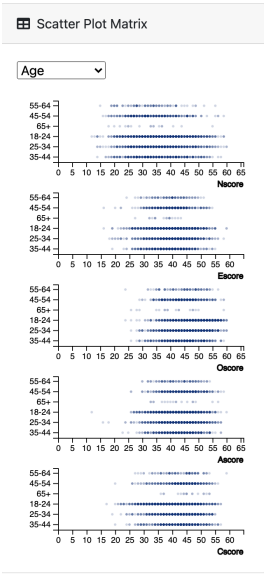
As mentioned per the requirements, the below report consists the information about the progress of the project till now.

Modifications to the Proposal :

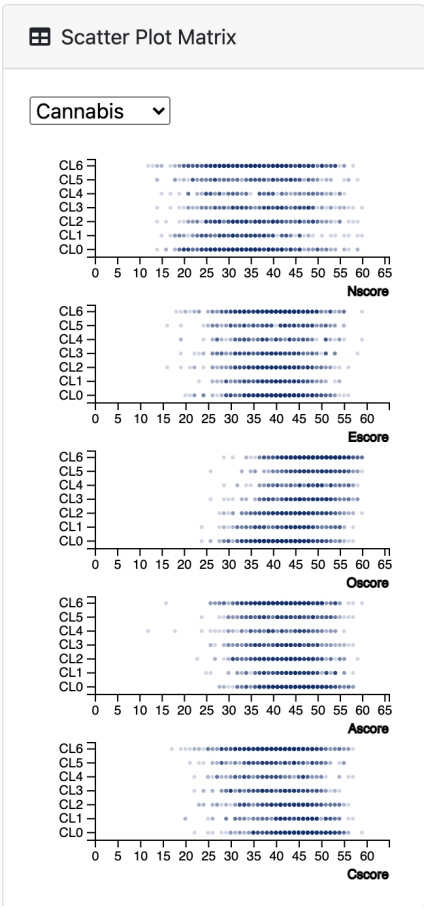
a) Changed the Euclidean MDS plot of the data points to a precomputed correlation plot of the features. This plot helps to better explain the correlation between features which is our primary goal of this project.
b) Used Ordinal Encoding for Categorical variables to be shown on the MDS feature plot.
c) Added Histogram plots too for numerical features.
d) Made the traits of each person limited to the five scores :Nscore, Escore, Oscore, Ascore and Cscore along with their background information.
e) Added tooltips for axis labels to provide more information to the user

Progress:

a) Scatter Plot Matrix:



Added a scatter plot matrix with the X-axis values fixed to the five primary traits and given an option to change the Y-axis values to other features apart from the primary traits.



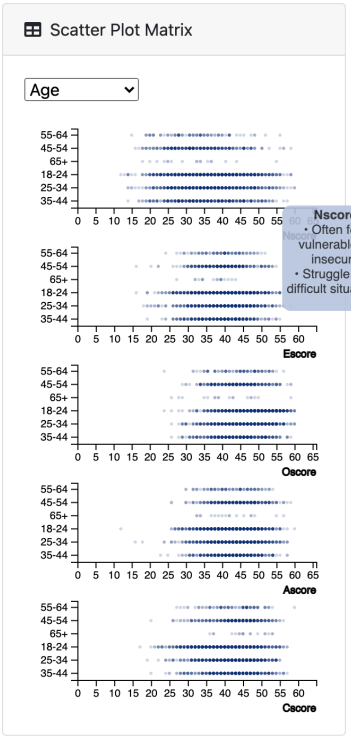
For example, if Cannabis is selected as the drug. The scatterplot matrix shows the plots of consumption against the primary traits. We can get information about the consumption of each drug and the traits which might incline towards the usage.

Insights:

We can see a few insights just from this plot too. If you observe, higher the Oscore(Enjoy to try new things) the individual is more likely to consume the drug.

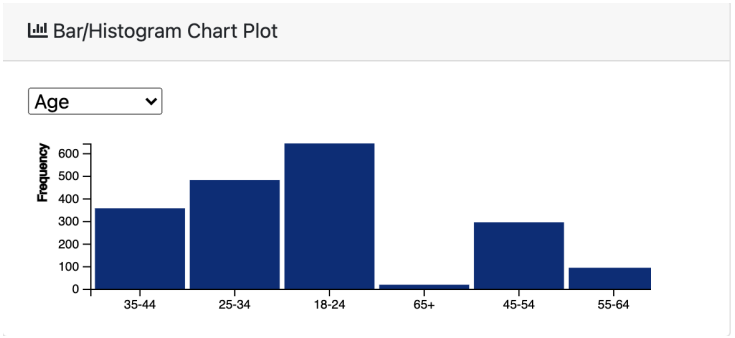
Similarly for Ascore(Helping others) and Cscore(Goal driven) are negatively correlated with Cannabis. So we can say people who have these scores high are not more inclined to consume Cannabis.

We can also observe that the plot is pretty dense even at the top, which conveys us that Cannabis is consumed highly compared to other drugs.

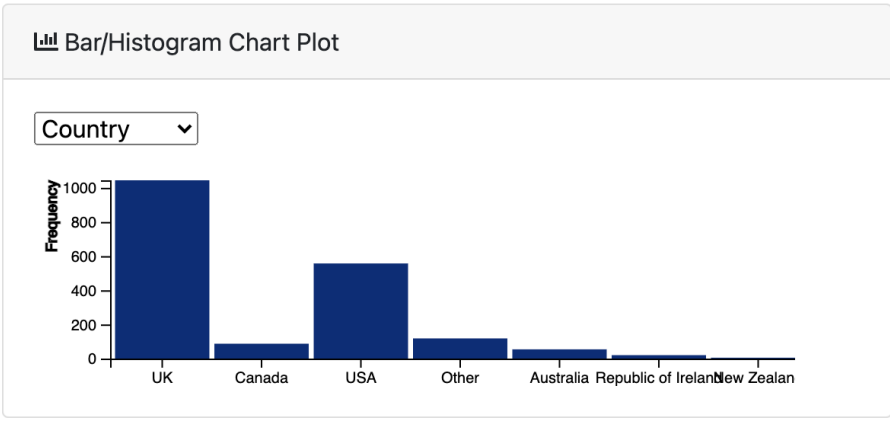


Added the tooltips for each axis to get better information about each trait if the user doesn't have much knowledge about it.

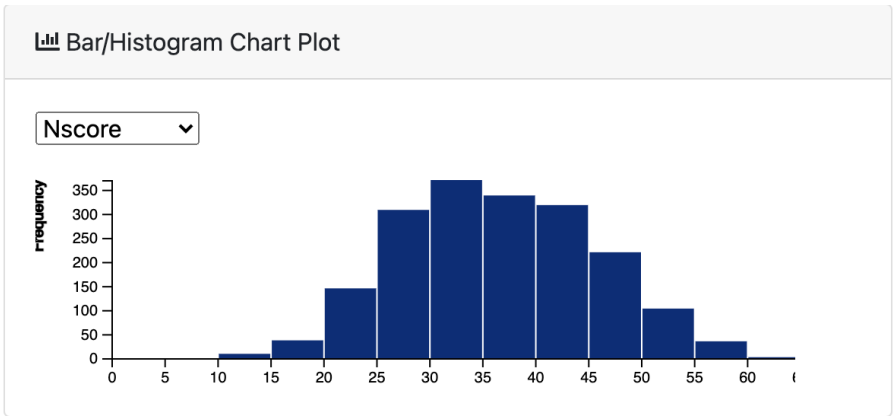
b) Bar/Histogram Plot:



Implemented a bar and histogram plots for each feature depending on whether it's a categorical or numerical feature. These plots help to provide us information on how diverse the initial dataset is. It helps us to provide a numerical value of different values of each feature.



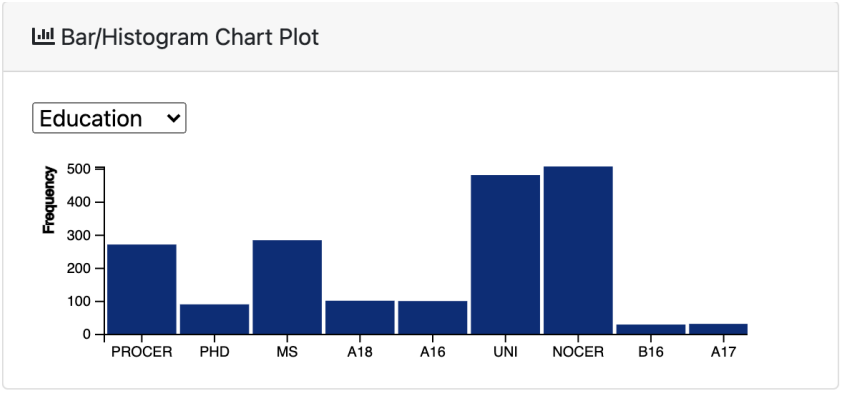
For example, if we select a country as the feature. We get a Bar plot and can see how diverse the country of each individual from the dataset is.



And selecting a numerical feature like Nscore will provide us a Histogram plot showing the ranges of scores.

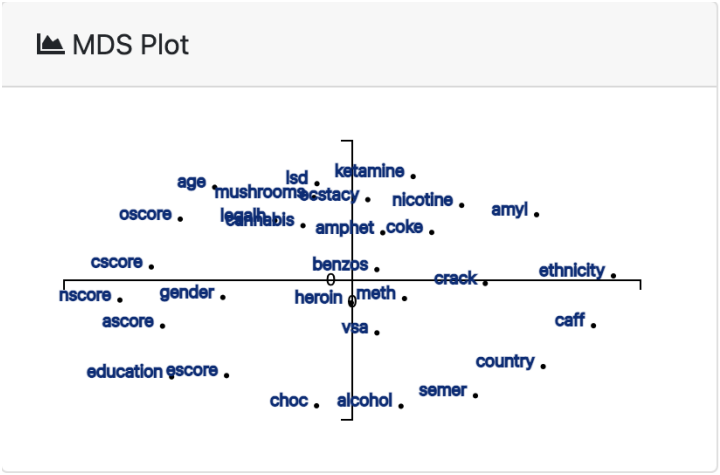
Insights:

As explained, these plots will provide us data on how diverse our dataset is actually. This will help us to take into consideration whether few insights are actually reasonable or not.



For example, if you see over here. There are more individuals who are either from University and dropouts among our dataset. This might help us out when the brushing is implemented by the end

c) MDS Correlation Feature Plot:

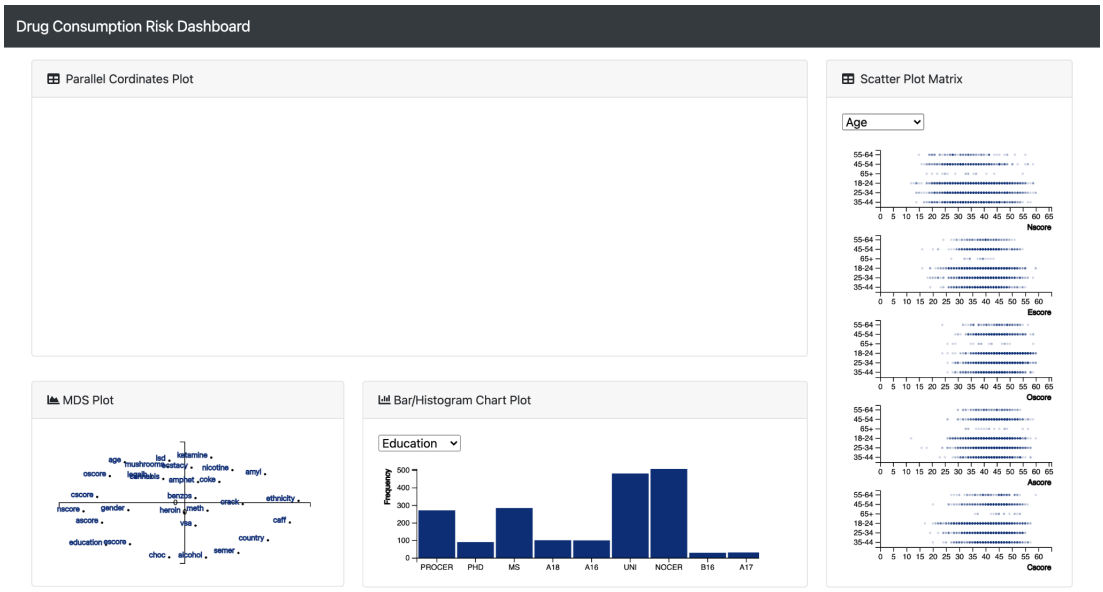


This plot helps us to provide the correlation between all the features. Features which appear close to each other are more correlated compared to distant ones. I used Ordinal Encoding to convert all the categorical features individually to numerical values to calculate the Correlation matrix and finally plot onto the screen.

Insights:

Knowing correlation between features is really important as this our primary goal of the project. Knowing which traits or features inclines the individual to consume a specific drug is important. If you see from the plot above, education and cannabis or alcohol aren't that positively correlated. This shows the general stereotype why a person who is too into higher education is not more likely to consume these drugs that frequently.

d) Dashboard:



The dashboard likes this by the end of the preliminary report and will include the PCP plot by the final submission.

Future Work:

a) Will be implementing the PCP plot to complete the provided plan.
b) Enable brushing on PCP and Scatter plot matrix.
c) Provide interaction between the plots to get more specific information.
d) Add more tooltips wherever required.
e) Implement K-mean once PCP is completed and integrate that to the Scatter Plot matrix.

How would these help ?

1) PCP plot will provide us information at a higher level showing the entire data onto a single plot and help us to gather any paths which are common for few features.
2) Brushing would help us to get an intrinsic subset of the data.
3) Interaction between plots would finally provide a meaning to the dashboard to integrate all the information and provide insights which are not available just form a single plot