Project Summary:

I analyzed drug usage data obtained from UC Irvine Machine Learning Repository. The aim of the work was to show if there were any statistically significant differences in a certain physical, emotional, and psychological scores of users of different drugs and chemicals.

The data was obtained from an online survey of 1885 respondents, but 12 attributes measured were of interest. In the work, I try to address hypothesis that includes:

1. Compare between the user classes of a specific drug Cannabis based on the attributes/variables using Hotelling's T2 and/or MANOVA.

2. Compare between the psychological variables for respondents between different gender, country or age using Hotelling's T2 and/or MANOVA.

In all these hypotheses tests, the null hypothesis was:
Null: All means of the groups are equal for the different drugs and chemicals used.
Alternative: At least one of the means is different.

Findings:

1. After conducting the Hotelling's Test of equality of means NEO-FFI-R scores between Users and Non-Users of Cannabis, I found that there is convincing evidence that the mean personality (NEO-FFI-R) scores of cannabis Users is different from the mean personality (NEO-FFI-R) scores of non-users of cannabis. (p-value (<0.001), significance level=0.05).
2. There is convincing evidence that the mean personality (NEO-FFI-R) scores of heroin users is different from the mean personality (NEO-FFI-R) scores of non-users of heroin. (p-value (<0.001), significance level=0.05)
3. There is convincing evidence that the mean personality (NEO-FFI-R) scores of cannabis Users is different among users of different Education backgrounds or levels (p-value<0.001, significance level=0.05).
4. There is no evidence that the mean personality (NEO-FFI-R) scores of Heroin Users is different among users of different Education backgrounds or levels (p-value=0.8775, significance level=0.05).

Other Work:

1. I did Principal Component Analysis to reduce the dimension of the data and to determine the optimal linear combination of the quantitative features to use.
2. Some cluster analysis and classification techniques were also employed to identify if users and non-users of cannabis could be separated effectively using these distance-based algorithms like the k-medoids clustering algorithm. Random forest algorithm was employed in the classification. Overall, the methods performed fairly on the data.