ALTERNATIVE CLASSIFICATIONS OF DRUG USERS

DOMINIC HO KWAN CHAN CID:01190685

THE CONTENTS OF THIS WORK AND THE ASSOCIATED CODE ARE MY OWN UNLESS OTHERWISE STATED

OBJECTIVES

- Show the current binary classification of drug consumption might possibly over simplifies individual drug consumption habits
- Investigate if there are alternative ways to classify drug users.
- Evaluate the practically of classifying individuals according to these new classifications.

NEW CLASSIFICATION NEEDED

- Quite a number of people has low severity in illegal drug consumption yet consume high level of cannabis (the black box).

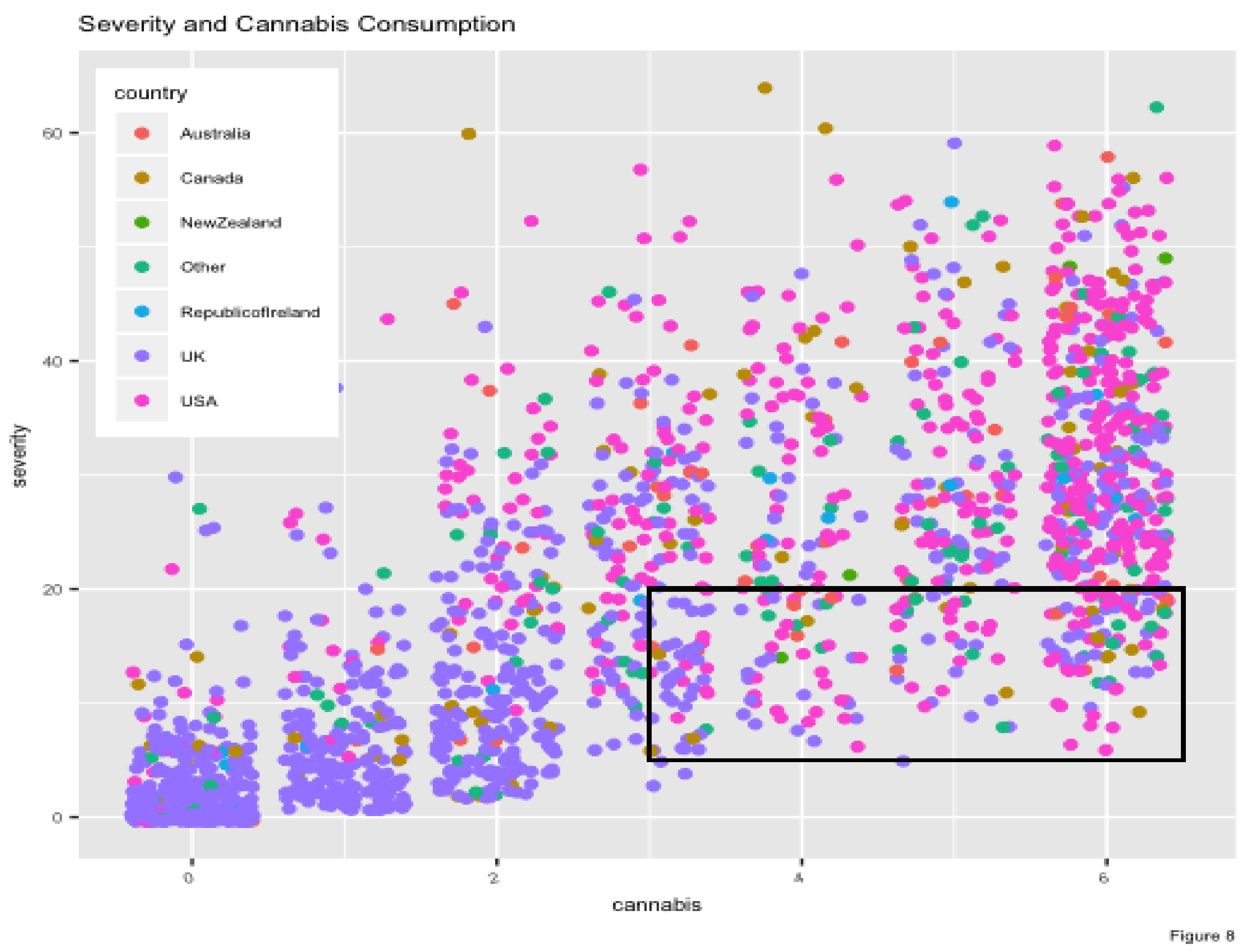


Figure 8

- People in the box are those that *mostly* prefer Cannabis (otherwise their severity shall be higher).
- A Different consumption pattern emerge for other drugs like Crack or Heroin.

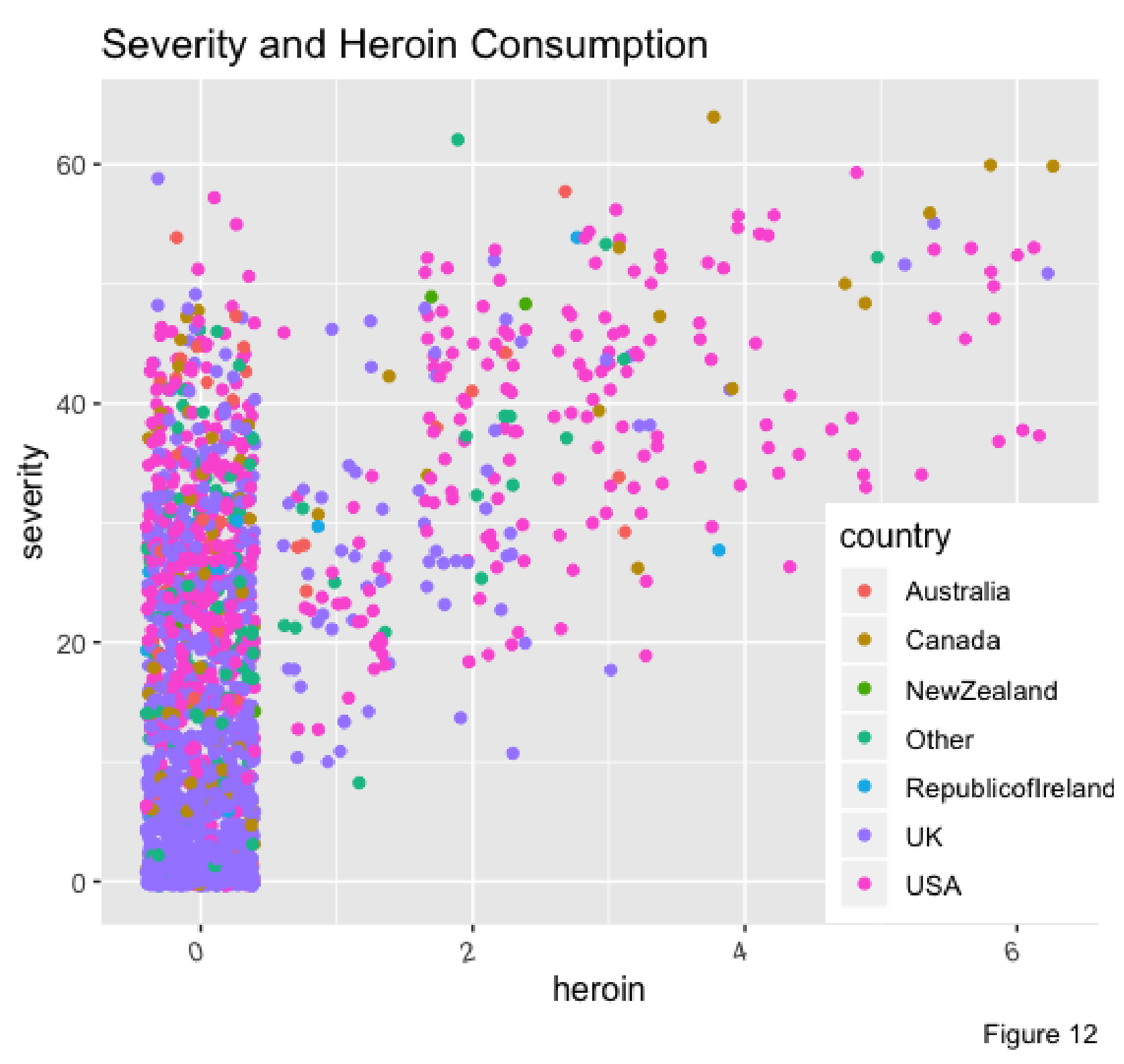


Figure 12

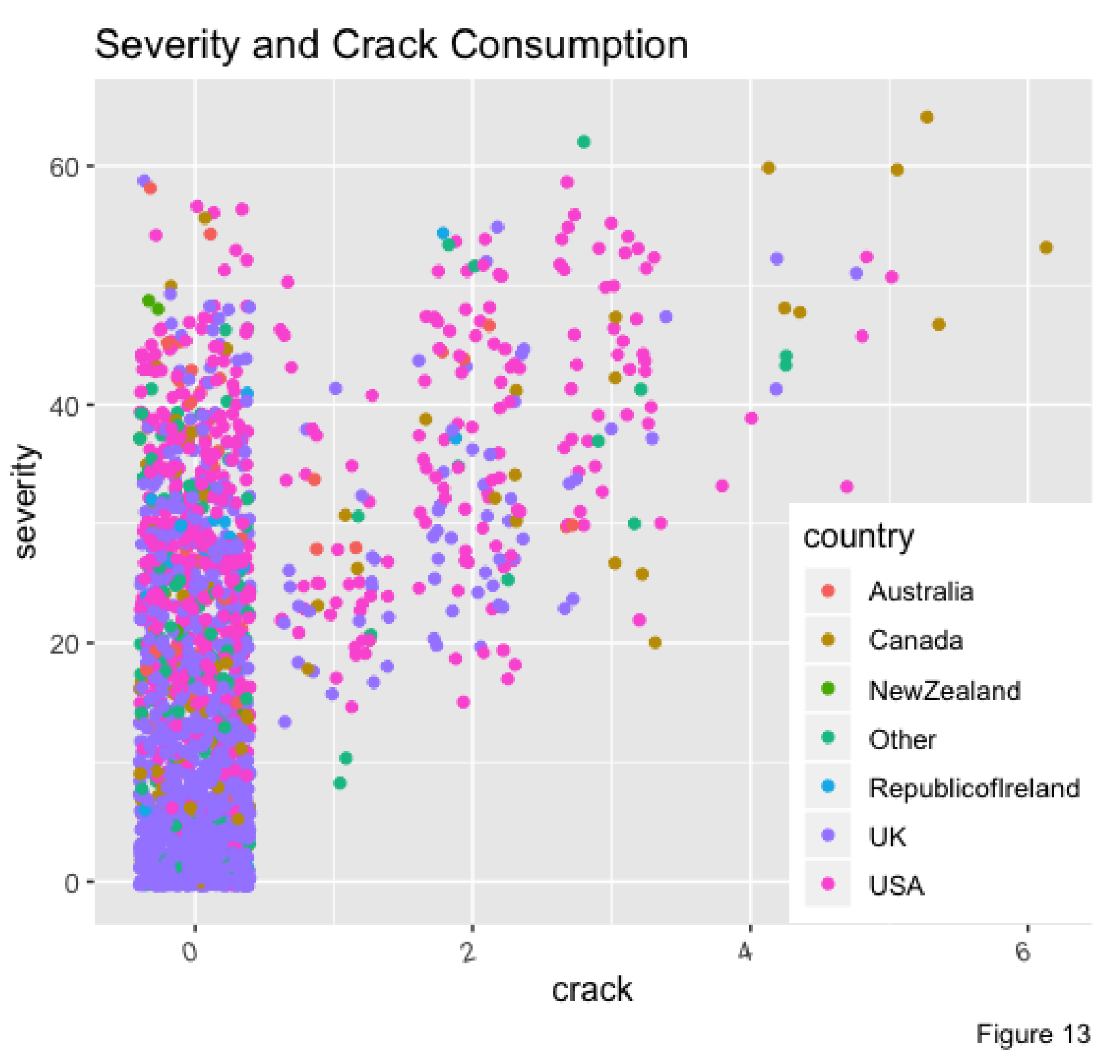
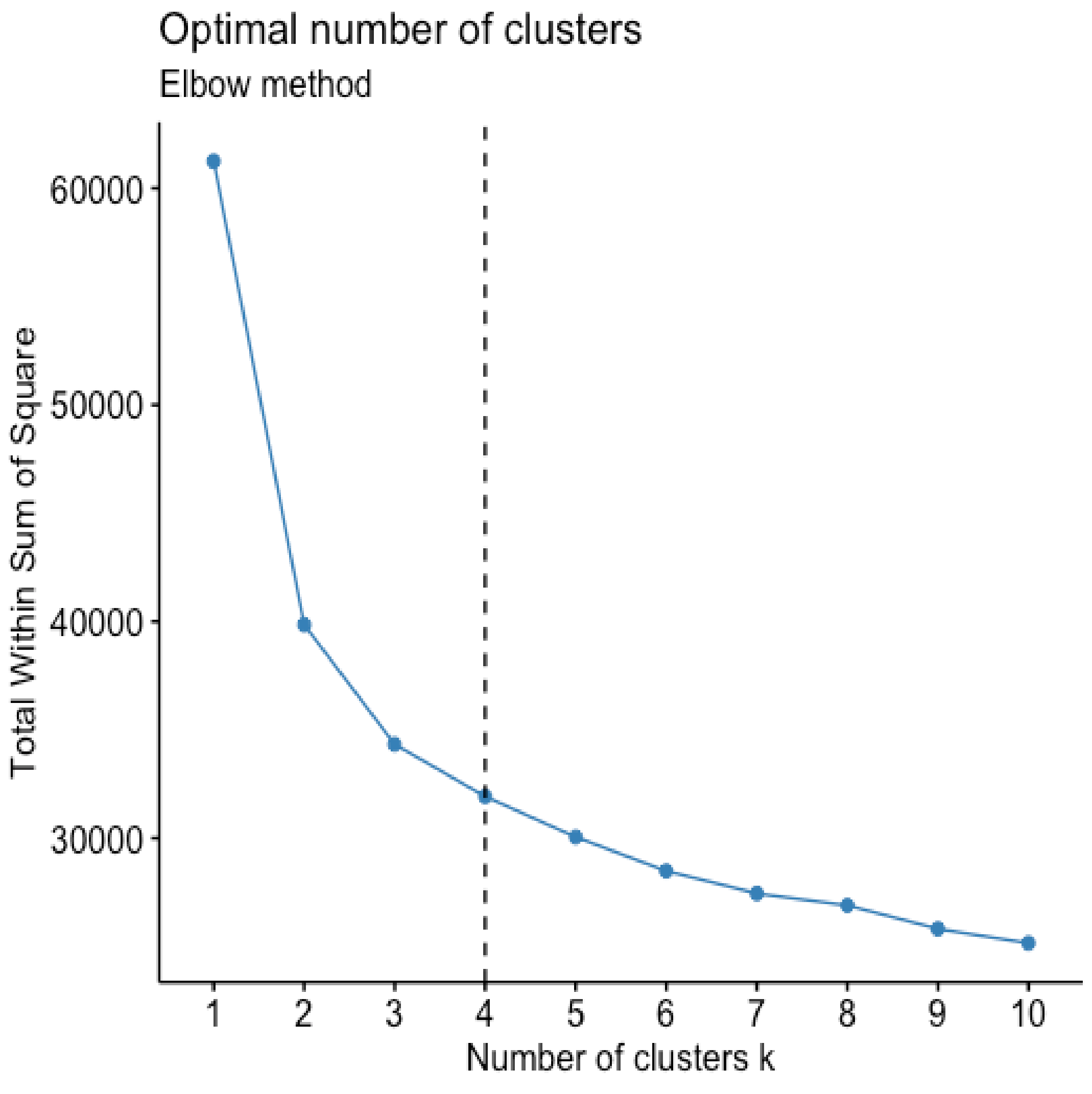


Figure 13

- Individuals are more likely to have high severity if they have high consumption in Crack or Heroin, implying these individuals most likely take other illegal drugs too.
- New classification might be possible to capture these features.

THE OPTIMAL K IN K-MEANS CLUSTERING

- Performing K-means clustering on the illegal drugs consumption level might group drug user of similar preference.
- Elbow Method is used to decide the optimal k values to choose.
- **k = 3 or 4** seem to be the sensible choice, $k \geq 5$ does not give better marginal improvement to within-clusters sum of square.



INTERPRETING THE CLUSTERS

The **3, 4-means** clusters both identify groups of individuals with similar drug consumption habits.

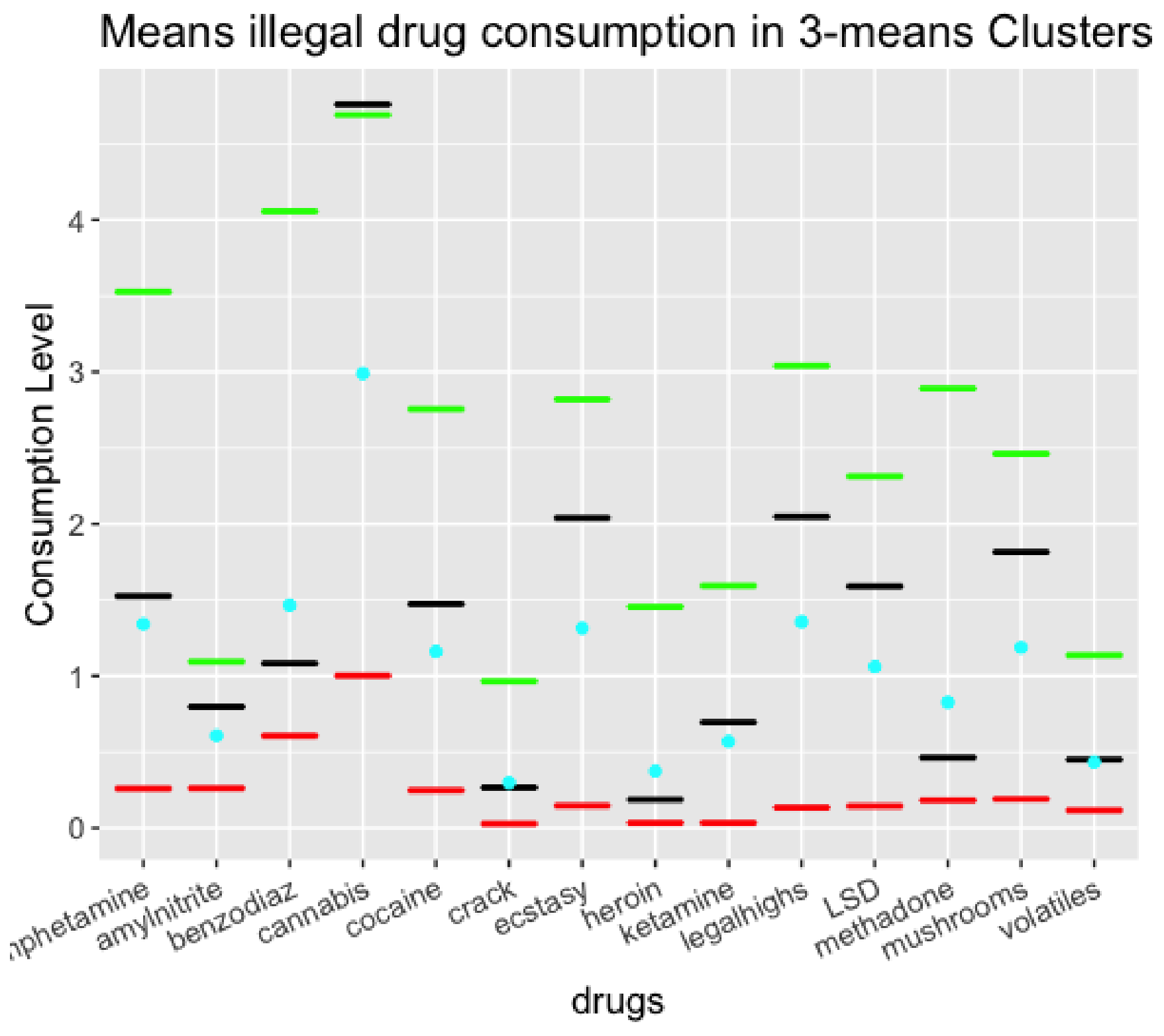


Figure 15

- Red → "Mostly clean/sober" (LOW)
- **Black** → "Mostly Cannabis" (CAN)
- Green → "Drug Addicts" (HIGH)
- Preserves the "Low" and "High" features in binary classification.
- Extra classification identify groups of individual's that mainly preferred cannabis.

- Red → "Mostly clean/sober" (LOW)
- **Black** → "Mostly Cannabis" (CAN)
- Green → "Favour methadone, crack, heroin" (AD 1)
- Blue → "Favour ecstasy, legalhighs, LSD, mushroom" (AD 2)
- 4-means cluster splits the "Drug Addicts" group to and identity more details on their drug preferences.

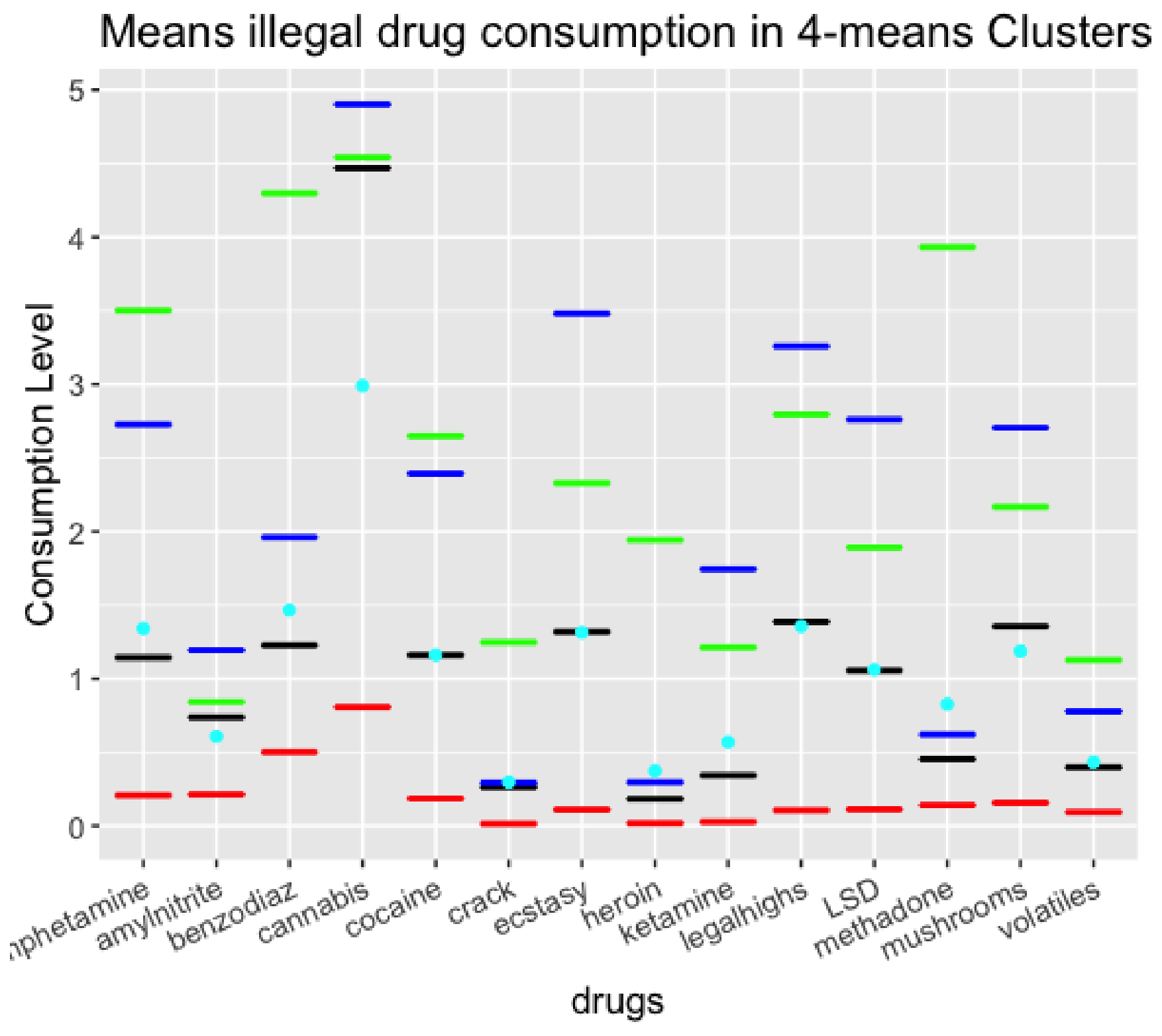


Figure 16

COMPARING THE CLUSTERS

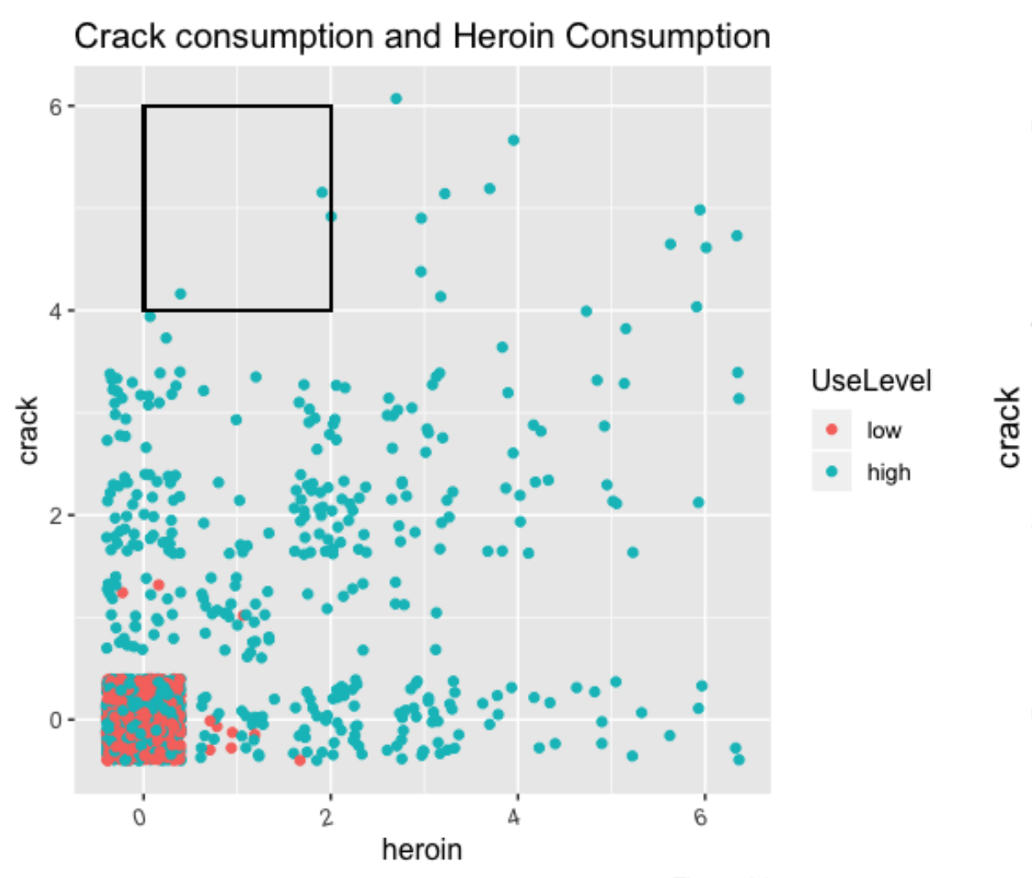


Figure 14

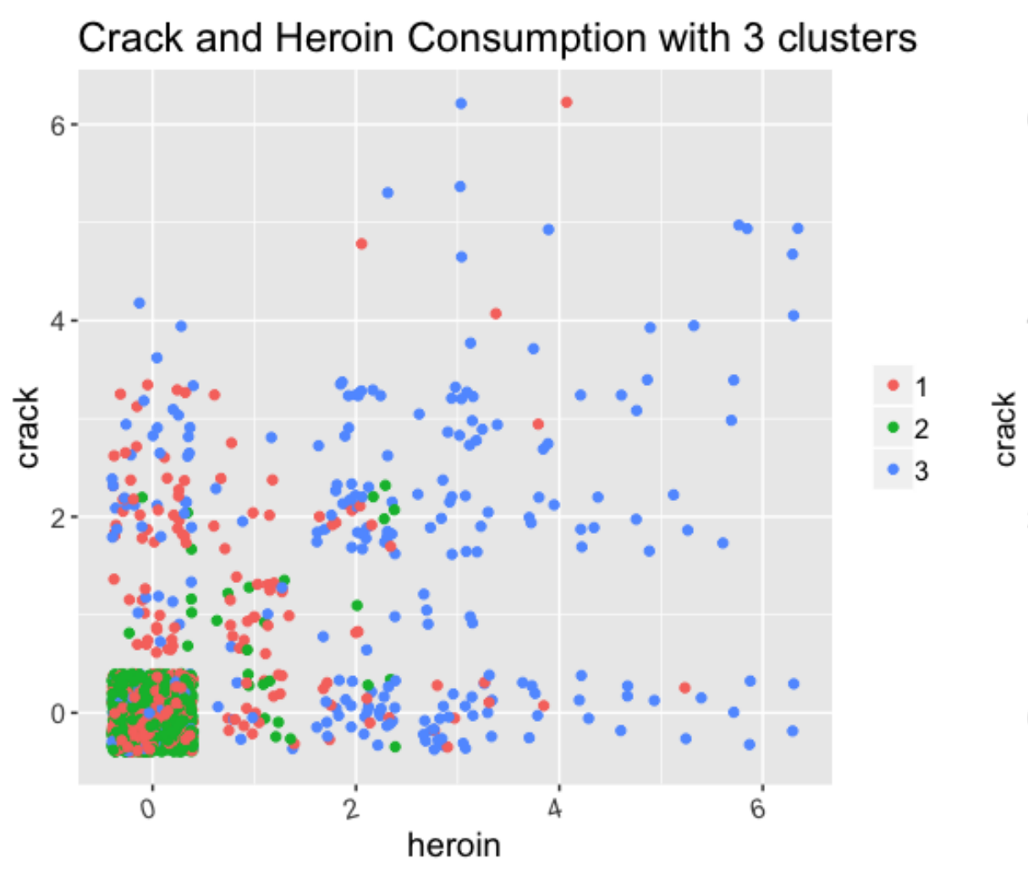


Figure 17

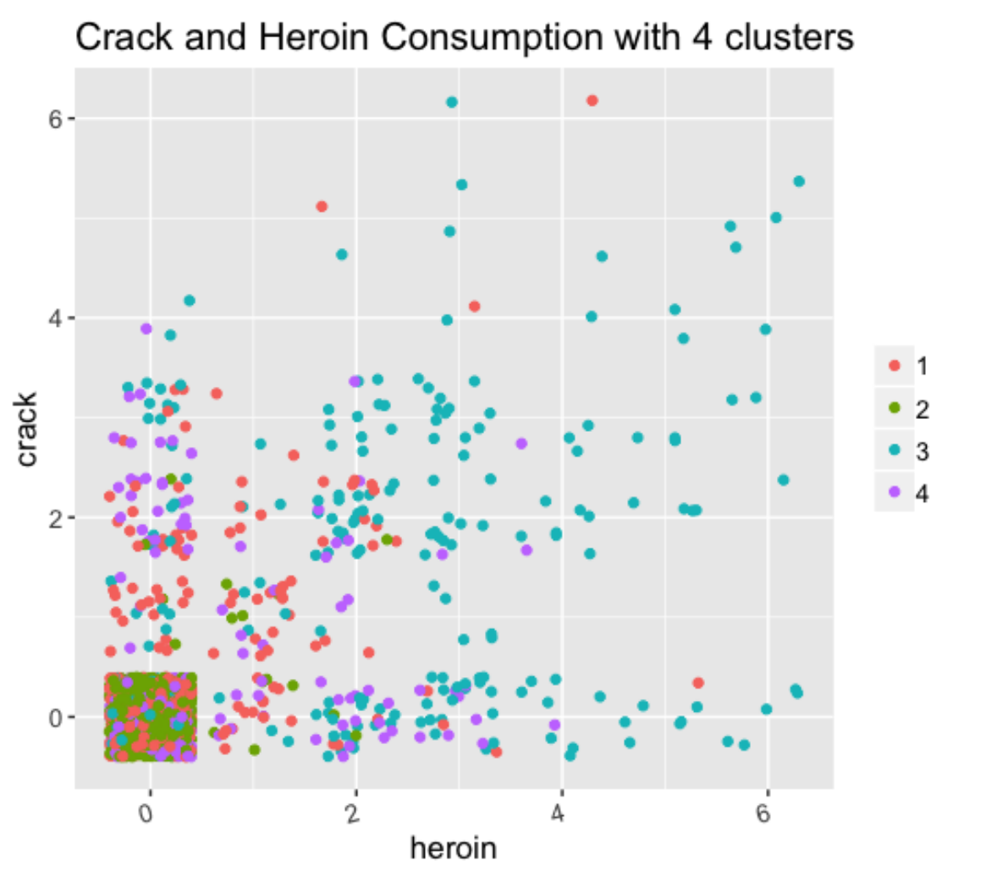


Figure 18

- The three plot shows the degree of information we obtained from the different classification methods.
- The binary class masks features of "High" drug consumption users.
- 3-group classification give distinction of the "High" group in the binary class, breaking them as "Drug Addicts" and "Cannabis" group.
- 4-group classification provides more info on the "Drug Addicts" group from 3-group. Most of these people are split in to one of two drug addict group, blue representing those that like crack and heroin while purple are those drug addicts that prefer other illegal drugs.

PREDICTING THE NEW FOUND CLASSIFICATIONS

RandomForest is used to predict the new classifications with first 1400 sample as training set and the rest as testing. (For reference, the accuracy for the binary classification problem using **KNN** is 80%)

	LOW	CAN	HIGH
LOW	101	22	36
CAN	35	209	12
HIGH	22	3	45

- 73.2% accuracy
- 71.1% 10-fold Cross Validation accuracy.

	LOW	CAN	AD 1	AD 2
LOW	63	16	22	35
CAN	33	200	3	11
AD 1	7	2	24	10
AD 2	19	1	7	2

- 63.7% accuracy
- 61.1% 10-fold Cross Validation accuracy.

- There is a trade-off between more insightful classification and prediction accuracy.
- Accuracy of the prediction drops 10% when considering a 4-group classification problem instead of the 3-group classification problem.
- In light of such trade-off in current prediction methods, using the 3-group classification would be more viable.