CMSE 381 Project Final

Eric Morris

4/16/2021

Introduction and Project Goal

When I was growing up and had to go through anti-drug education, I was always told “Marijuana is a gateway drug.” Meaning that if I started with marijuana early, it would cause me to use harder drugs later down the line like coke, heroin, etc. With the “Drug Use By Age” dataset I aim to test that claim and see if my model can predict drug use in age groups by their respective marijuana use/frequency.

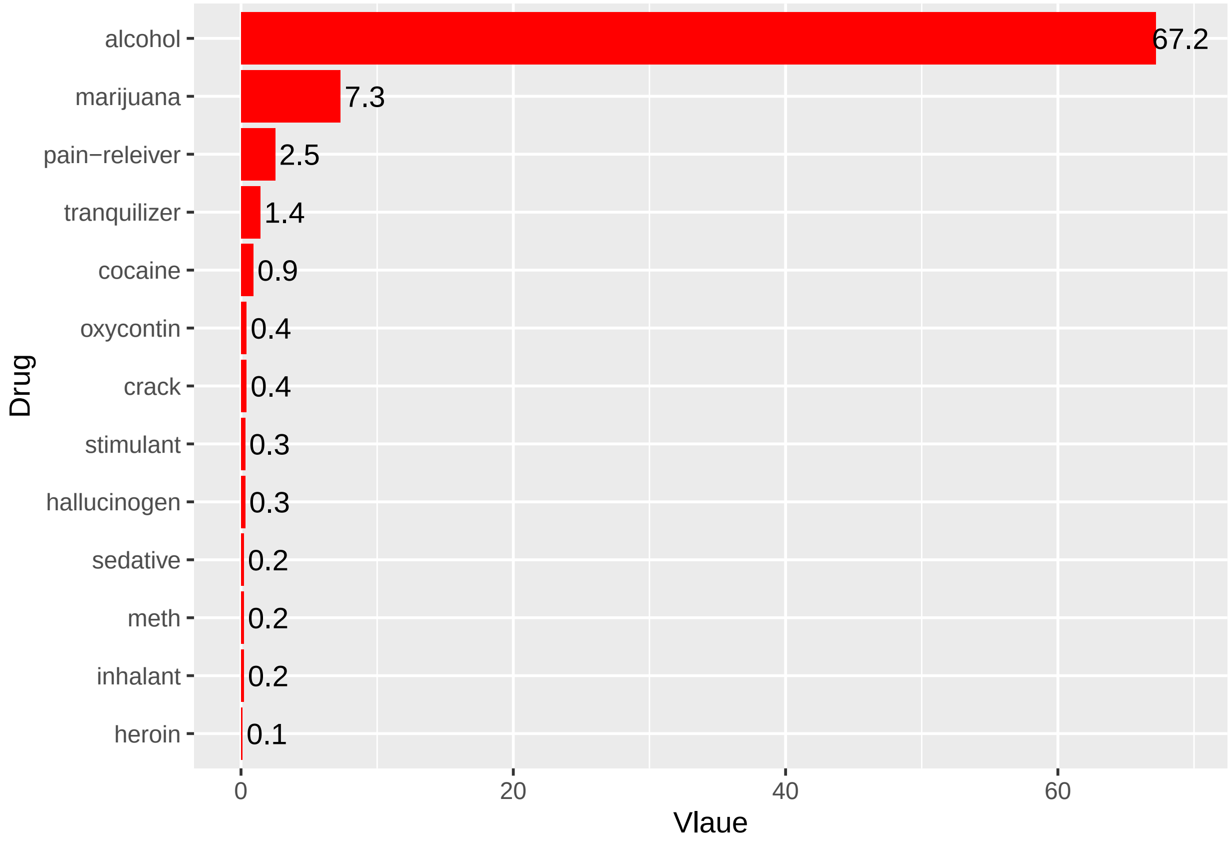
importing data

```
## Warning: package 'caret' was built under R version 4.0.5
```

```
## Warning: package 'readr' was built under R version 4.0.5
```

Recreating Outside Model

The model associated with this data used an horizontal bar chart to show what drug baby boomers (ages 50-64 years) use the most frequently to get high.



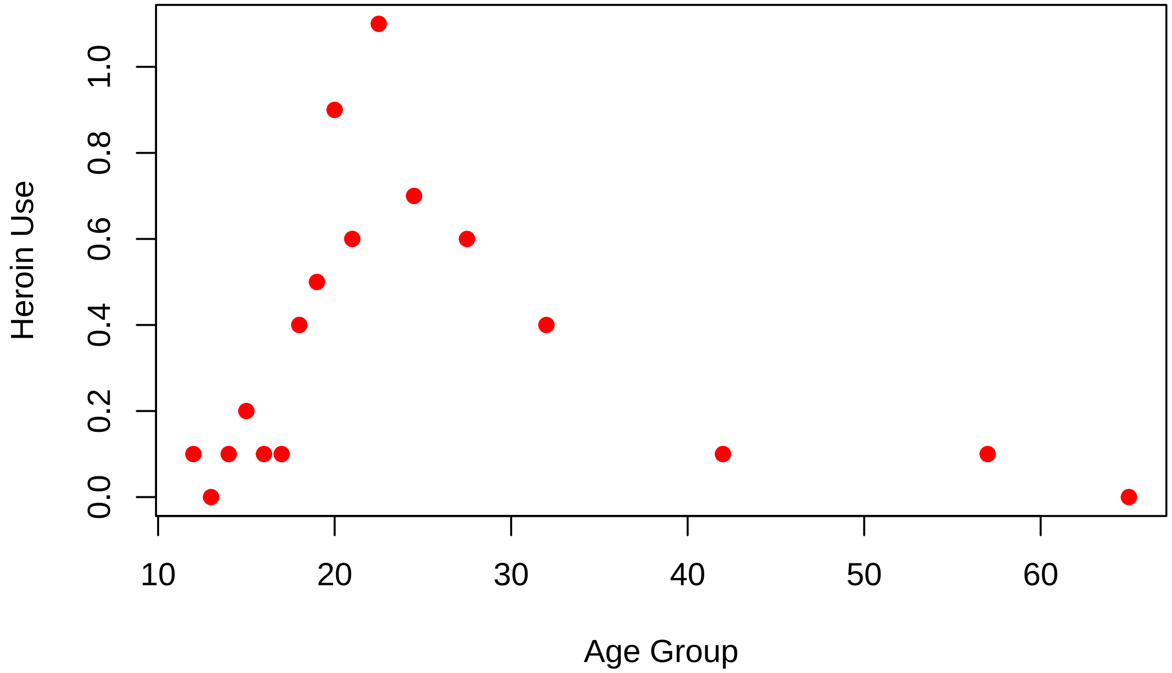
Opinion on recreated model

While this model makes it very clear what boomers are using most often to get drunk/high, that's really it. From this graphic there is not much more we can ask about the presented data. I believe that the original data can answer more complex questions while using more complex models. With my model we will be answering the question of can the age group of someone and their use of marijuana help significantly in predicting their use of heroin. I am choosing because I am from Ohio and the heroin epidemic has ravaged that state for some years now

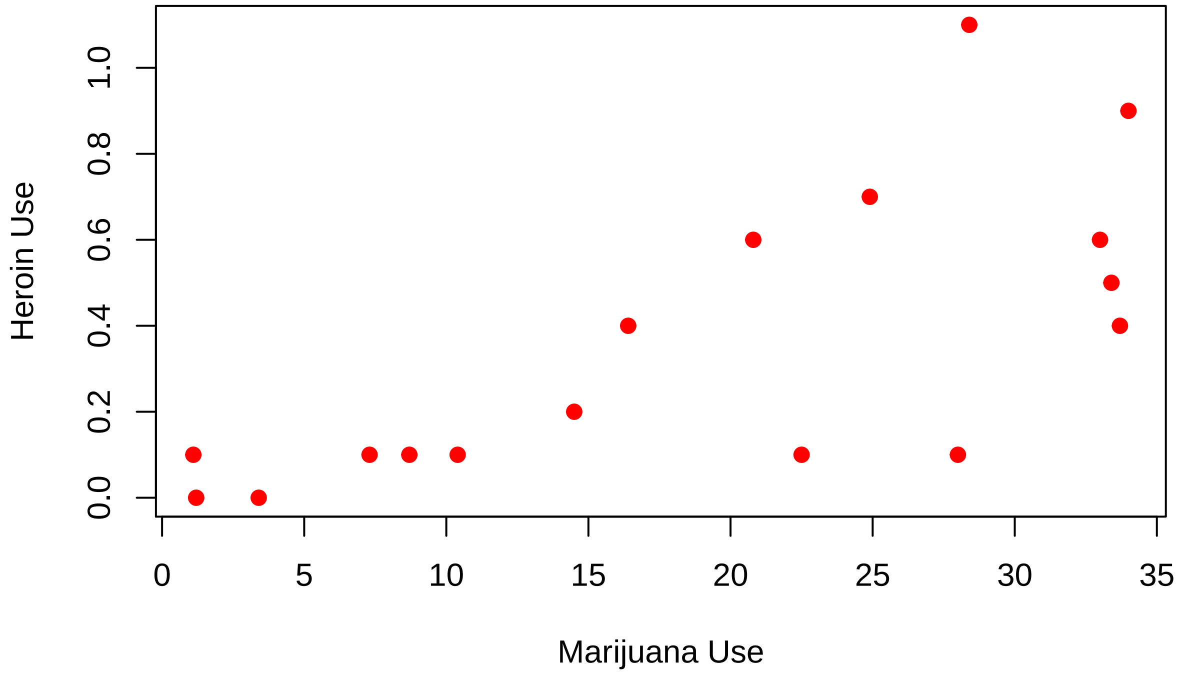
Method

I will be using linear regression as I want to see if an increase in marijuana use shows an increase in heroin use, and also if age groups have a significant affect on heroin use. Before fitting that model I will make simple scatter plots of age vs heroin and marijuana vs age to see if a clear relationship can be seen between these variables or if I need to reevaluate my plan of action. This is mainly to help me predict what I think my results will be when I got to see how well my model preformed. For simplicity I have taken the mean age in the ranges and will be using those numerical values.

Results



It seems like a linear regression might not fit this data so well. After looking at the trend I will use a polynomial regression instead. It will use a degree of 3 for Age.

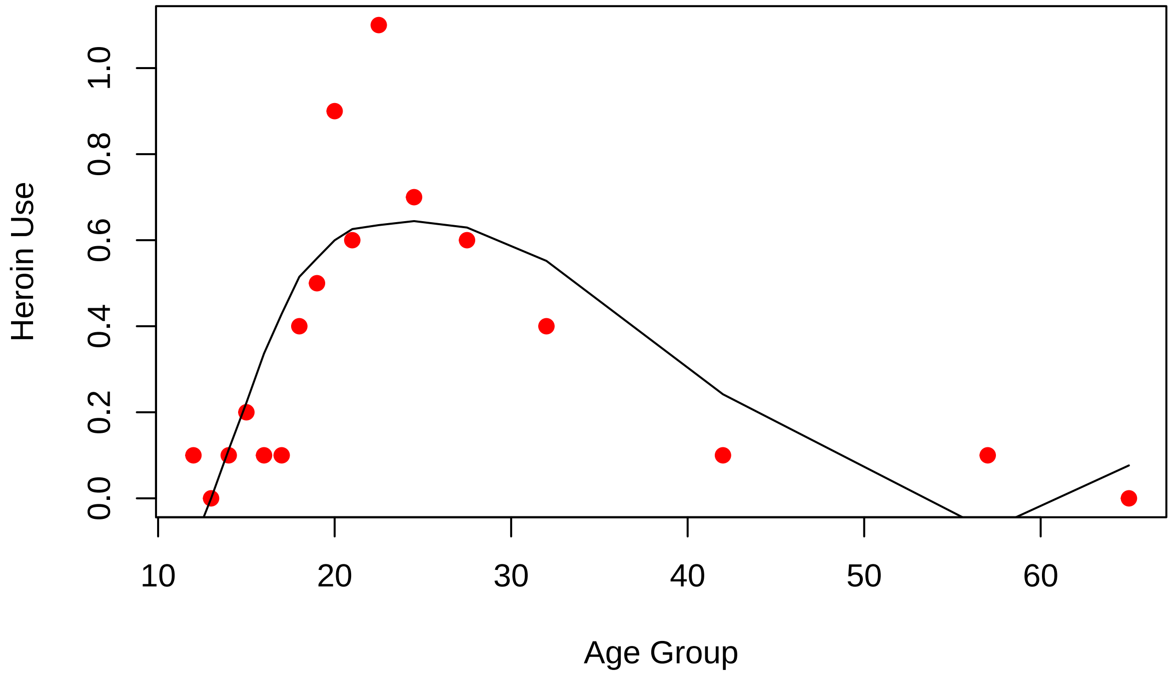


The relationship in this plot is a lot more messy than the previous than before. I will incorporate it into the fit with a degree of 1.

```
##
## Call:
## lm(formula = drug.use$‘heroin-use‘ ~ poly(ages, 3) + drug.use$‘marijuana-use‘)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32872 -0.11491 -0.02582  0.05562  0.46496
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                0.237876   0.167844   1.417   0.1818
## poly(ages, 3)1            -0.156456   0.272681  -0.574   0.5767
## poly(ages, 3)2            -0.610396   0.298937  -2.042   0.0638 .
## poly(ages, 3)3             0.551429   0.315124   1.750   0.1056
## drug.use$‘marijuana-use‘  0.006081   0.008383   0.725   0.4822
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.226 on 12 degrees of freedom
## Multiple R-squared:  0.656,  Adjusted R-squared:  0.5413
## F-statistic: 5.721 on 4 and 12 DF,  p-value: 0.008179
```

From this summary it seems that none of the variables are significant, the most significant is ages^2 with a p-value of .0638. With all variables fit to it, it has and adjusted R^2 value of .5413 suggesting that 54.13%

of the variation in our data can be explained by our model. While this summary shows a very weak model, I will add it to our Age Vs Heroin plot to see how it looks and calculate the training MSE.



The model doesn't look terrible but let's looks at the training MSE before making any further discussion.

```
## [1] 0.03606628
```

This fit on the plot and training MSE give us hopeful results. With an MSE of .03 some However since this is such a small dataset and we don't have a test set we can't make big claims about this model. So I will be preforming LOOCV on this data to get a better idea on how this model actually performed.

```
## Linear Regression
##
## 17 samples
##  2 predictor
##
## No pre-processing
## Resampling: Leave-One-Out Cross-Validation
## Summary of sample sizes: 16, 16, 16, 16, 16, 16, ...
## Resampling results:
##
##   RMSE       Rsquared   MAE
##   0.4091381  0.1199828  0.2706149
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

Conclusion

At first seeing the summary of the model along with the p-values of the coefficients I thought this model would be pretty useless. However, fitting it on the plot showed that it actually fit the data a bit better than I anticipated. Thanks to in part my decision to switch from a strictly linear model to a polynomial. I had figured the training MSE was going to be pretty small as it is measuring against the data it trained with. However I did not anticipate that the RMSE from the LOOCV would be so small. At the same time while RMSE is small, so is the Rˆ2 value from the LOOCV which seems to be the opposite of what I would predict from a small RMSE. The last thing I would like to discuss is potential improvements to my model and the data. I think my data would've been a lot more helpful if the data was formatted differently. I would've liked to look at individuals rather than a collective for each age group. Even with those limitations I believe my model answered the question I set out to answer quite well.

References

https://fivethirtyeight.com/features/how-baby-boomers-get-high/ https://www.statology.org/leave-one-out-cross-validation-in-r/#:~:text=The%20easiest%20way%20to%20perform,the%20caret%20library%20in%20R.