In this analysis our team investigated how people in different age groups misuse (in a way not authorized by their healthcare provider) prescription drugs. Initially starting with exploratory data analysis, we found that the percentage of drug misusers per age group can be over 20% for categories like RX Painkillers. In order to determine exactly how and why people are misusing these types of drugs, our team looked into the text data given by the survey variable 'OTH_RX_DRUG_USE_SPFY' which was a free response question on the survey asking people to specify in words how exactly they were misusing drugs.

The first step in looking at the text data was to group the responses by age category. We identified 6 age groups to use based on traditional grouping procedures as well as evenly distributing the responses in each category. The age groups we identified were [18,24] [25,29] [30,34] [35-39] [40,59] [60,99]. Next, we analyzed how the abuse of prescription drugs varies by age group using question categories B4.1 through B4.4 by generating a rudimentary logistic regression model for each prescription drug category (pain relievers, stimulants, etc), which illustrated that abuse of all categories stayed relatively level as age group increased until the 40-59 group, at which point all categories decreased dramatically.

Then, our team was able to create a bag-of-words model that showed the occurrence of each word in the entire set of vocabulary from respondents for each respondent. Then by sorting this model by frequency, we generated a histogram of the most frequently used words by age group. Interesting points we found were that younger people tended to use drugs like adderall and mention college -- so a possible linkage to those two factors could be seen. And, as for the older respondents, words like 'sleep' and 'pain' became much more prevalent. This shows that older people and younger people are misusing drugs for different reasons

For this reason, our team wanted to find out whether or not there was a trend towards positive or negative sentiment between age groups. We were able to run the VADER sentiment intensity analyzer from the NLTK package. This assigned a score on a scale of -1,1 on how positive or negative each answer given by a respondent was. Calculating this for every age group we only saw slight differences between older and younger respondents. However, age groups older than 30 did have more negative associations for their answers.

Finally, to assign different topics to respondents' answers, we used an unsupervised topic modelling technique. Using Latent Dirichlet Allocation, we were able to come up with 5 unique topics for each one of our age groups. Within these topics, they mostly mirrored the most frequently occurring words, but made it easier to see patterns in between them. For every age group, a common theme persisted which was 'using someone else's prescription.'

In conclusion, our team determined that between age groups, the similarity of drug misuse was the fact that they were using someone else's prescription, and as age increases, the reason for drug misuse changes e.g. pain and sleeping for older respondents.

All code can be found in the following notebooks:
https://colab.research.google.com/drive/19vHWKWtdFOoLGdm-MZyvAZT7ZI2gValp#scrollTo=o5pvVumEMroZ
https://colab.research.google.com/drive/1SjFkkkTCRAlbURdBjgqxv_VmbzI1PoOs#scrollTo=yz3IBwyjkF5i