My list in order of preference

1. Student alcohol consumption (interesting topic, and the data set I found has a ton of stuff to work with)
   1. ~~Data set we found can be used to answer this question: what behaviors correlate with student success (edit: someone already did a question like this on Kaggle, and it goes very in depth)?~~
      1. [~~https://www.kaggle.com/uciml/student-alcohol-consumption~~](https://www.kaggle.com/uciml/student-alcohol-consumption)
   2. ~~How does alcohol consumption affect student performance?~~
      1. [~~https://www.kaggle.com/uciml/student-alcohol-consumption~~](https://www.kaggle.com/uciml/student-alcohol-consumption)
   3. **What factors in students are correlated with alcohol consumption?**
      1. <https://www.kaggle.com/uciml/student-alcohol-consumption>
2. ~~NBA-Related~~
   1. ~~2015-2016 Warriors had the best season in NBA history (73-9)~~
      1. ~~What factors contribute to Warriors' (or teams in general) success in the 2015-2016 NBA seasons?~~
         1. [~~https://www.kaggle.com/daverosenman/nba-finals-team-stats~~](https://www.kaggle.com/daverosenman/nba-finals-team-stats)
         2. [~~https://www.basketball-reference.com/teams/GSW/2016.html~~](https://www.basketball-reference.com/teams/GSW/2016.html)
   2. ~~2015-2016 Warriors won the championship~~ 
      1. ~~How impactful is the high 3-point percentage for Warriors’ success?~~
         1. [~~https://www.kaggle.com/daverosenman/nba-finals-team-stats~~](https://www.kaggle.com/daverosenman/nba-finals-team-stats)
   3. ~~2017-2018, 2018-2019 Warriors won back-to-back championships~~
      1. ~~How critical was XYZ for winning two consecutive championships?~~
         1. ~~https://www.kaggle.com/daverosenman/nba-finals-team-stats~~
3. ~~Detecting diseases, possibly COVID-19? (another interesting topic with a ton of topics and data sets online)~~
   1. [~~https://data-flair.training/blogs/python-machine-learning-project-detecting-parkinson-disease/~~](https://data-flair.training/blogs/python-machine-learning-project-detecting-parkinson-disease/)
   2. [~~https://github.com/owid/covid-19-data/tree/master/public/data/~~](https://github.com/owid/covid-19-data/tree/master/public/data/)
   3. [~~https://www.kaggle.com/andrewmvd/heart-failure-clinical-data~~](https://www.kaggle.com/andrewmvd/heart-failure-clinical-data)
4. ~~Sentiment Analysis~~
   1. [~~https://data-flair.training/blogs/data-science-r-sentiment-analysis-project/~~](https://data-flair.training/blogs/data-science-r-sentiment-analysis-project/)
   2. [~~https://analyticsindiamag.com/10-popular-datasets-for-sentiment-analysis/~~](https://analyticsindiamag.com/10-popular-datasets-for-sentiment-analysis/)
5. ~~Student performance on exams~~
   1. [~~https://www.kaggle.com/spscientist/students-performance-in-exams?select=StudentsPerformance.csv~~](https://www.kaggle.com/spscientist/students-performance-in-exams?select=StudentsPerformance.csv)
6. ~~Fake news detection (a bit difficult to find good data sets for this)~~
   1. [~~https://github.com/BuzzFeedNews/2016-10-facebook-fact-check~~](https://github.com/BuzzFeedNews/2016-10-facebook-fact-check)
   2. [~~https://figshare.com/articles/dataset/PHEME\_dataset\_of\_rumours\_and\_non-rumours/4010619~~](https://figshare.com/articles/dataset/PHEME_dataset_of_rumours_and_non-rumours/4010619)
7. ~~College Ranking and Major: Does it actually matter?~~
   1. ~~https://www.kaggle.com/chris95cam/forbes-americas-top-colleges-2019~~
8. ~~The Effects of COVID-19 Lockdown on Air Pollution in LA Count~~
   1. [~~https://www.kaggle.com/thaddeussegura/california-air-quality-2020-through-sept10th~~](https://www.kaggle.com/thaddeussegura/california-air-quality-2020-through-sept10th)
   2. ~~https://www.kaggle.com/adityavipradas/covid19-dataset-for-california-counties~~
9. ~~Disparity between socioeconomic levels and internet usage~~
10. ~~How does GDP per capita correlate to the popularity of certain movie genres in USA?~~
11. ~~What is the relationship between marijuana usage, age, and socioeconomic status throughout the United States?~~
12. ~~What is the relationship between an adult population’s awareness of climate change and the amount of carbon dioxide (CO2) released by that population’s region?~~
13. ~~What is twitter’s overall sentiment about Kevin Durant and Giannis Antetokounmpo on days where their team wins vs days where their team loses?~~
14. ~~How do measures of climate change and destruction – specifically temperature and mixing layer depth – impact the presence of harmful algal blooms along the coast of California?~~
15. ~~Can we create a classification model to predict the 2020 presidential bias of politically-focused TikTok videos?~~
16. https://www.kaggle.com/daverosenman/nba-finals-team-stats

Research Questions

Professor Ellis’s email:

Allen,

Thanks for reaching out.

Going to answer your second set of questions first.

**- We found out that some of these questions have already been answered. Do we need to drop these ideas or are we allowed to use them as long as we cover previous analyses in the Background and Prior Work section?**

If others have done this before, your work has to build on their work in some way or another. For example, for Q#1(alcohol and student performance) - your analysis could use multiple datasets or different geographic locations or a new time period/updated time period to answer the question...but it cannot simply replicate (repeat) the work of others.

**- Are we on the right track for these questions?**

You're on the right track, but some could be more specific. For example (Q2), "student success" can be measured in many different ways - I want your proposal to be very clear exactly how success is being defined...and what behaviors are being measured.

**- Can we narrow these questions down as the quarter progresses?**

Yes - you can change/modify/pivot at any point in time (we don't need to approve this but are happy to discuss). Most students will change their question at least slightly. But, the question (or set of questions) you propose must be specific, clear, and answerable with data for the proposal

As for the specific questions, you're definitely on the right track. I'm going to add a few bits of feedback for them:

1) how is performance going to be measured? are these undergraduate students? graduate? in the states? elsewhere? over what time period?

2) behaviors needs to be better defined as well as what metrics of success you intend to use (GPA? Job? Salary? etc.)

3) similar to #1 - factors needs to be better defined. And is this a yes/no on alcohol consumption (binary) or how much is consumed? - it could be either; be sure wording is clear

4) Your background section here would need to be clear as to why you're focusing on that season in particular. A related question could be to see similarities in the team metrics for the two teams that went to the finals over time (you could decide the time period).

5) If you go this route, I'm going to encourage you to put this in context to other factors (i.e home court advantage, number of all stars, free throw shooting %, overall shooting %, number of 3 pt shots taken, etc.) so that you're able to answer/quantify this relative to other aspects of the game, 3pt shooting has more/less of an impact)

6) I like this general idea....specifics to be fleshed out.

Any of these would be fine avenues. Be careful with #1. There are some datasets out there that are better than others (For example, students sometimes fall to using a dataset that's popular on kaggle...that is alcohol consumption in a different country with grades they don't understand. If you want to specifically answer this question in that country, that's fine. But if this is to be in the states/anywhere else, be sure your data match. There are survey data in the states with this information...so encouraging when the time comes, if you go this route, to spend the time finding the data you need)

Hope this helps!

Prof Ellis

Background and Prior Work (previous research on the alcohol topic)

1. Correlation between alcohol consumption and academic performance: <https://rstudio-pubs-static.s3.amazonaws.com/365904_90b5ff49711448b5aa6ba2f09ffb4fc3.html>
2. Do students drink more on weekends? <https://www.kaggle.com/jhuno137/machine-learning-on-student-alcohol-consumption>
3. How alcohol consumption affects student performance: <https://www.kaggle.com/mukultaneja/analysis-student-alcohol-consumption>
   1. This one is interesting because they used the same data set as us, except that they looked at the data as alcohol impacting certain columns. Our question is more about what columns impact alcohol, so I think our question is still viable.
4. Outside of our data set, many of the projects are seen are done on college students
   1. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3125885/>
   2. <https://www.collegedrinkingprevention.gov/media/journal/040_090.pdf>
5. Finding correlation between personal factors and alcohol consumption and potential impacts on school performance: <https://www.rpubs.com/heke0495/419722>
   1. This one is concerning because our question is based on this, and Professor Ellis doesn’t want us to replicate work. I’ll try to think of something else.
6. Similar to 6
   1. <https://www.slideshare.net/DeminDamianWang/students-alcohol-consumption-data-analysis>

# **COGS 108 - Final Project Proposal**

# **Names**

* Allen Zou
* Brian Han
* Dillen Padhiar
* Lian Su

# **Research Question**

Are there certain factors that correlate with frequent alcohol use by high school students in the United States, particularly with individual factors such as sex and age, parental and family factors such as their education, size, income, etc., school performance factors such as workload/free time, educational support, grades, etc.?

## **Background and Prior Work**

Alcohol consumption is a topic of concern that has raised many interests and questions because of its known negative consequences. Much research has already been done to examine the reasons for large alcohol intake and the effects of such intake, but we noticed that they tend to be conducted on the general population, usually in the scope of an entire country like the United States or the world itself. We want to examine a smaller category of people.

The group that we’re interested in are students, and we want to see if there are perhaps any new correlations or factors specific to students regarding alcohol consumption. From what we found, most research on students has been done on college students, and the first 3 references we included below elucidates some findings. For example, college students tend to drink only slightly more than people who don’t attend college when trait-based risk is low, but this is exacerbated when trait-based risk is high [1]. Additionally, social norms tend to play a large role in college drinking, college students tend to have more alcohol-related problems than those who don’t attend college, having an early Friday class correlates with more drinking before that day (strangely does not correlate with no Friday classes or late Friday classes), etc. [2] The third reference describes some roadblocks encountered with research about college alcohol consumption; given that a college campus environment has a constantly changing social scene and culture, this confounds the factors that this group tried to consider, making it difficult to identify and establish fixed variables in research. Their situation is further convoluted by the fact that individual factors can play a role in alcohol consumption, which often aren’t related to the environmental factors in the first place [3]. This is just a small snippet of the total amount of research conducted, but for the sake of length, we’ll cut it short here.

Interestingly, there is a large difference between the amount of research on college students and on high school students, which may be due to the limited number of datasets that exist on high school students. We found that the conclusions for high school student alcohol consumption tend to focus on their academic performance. For example, reference 4 found that the level of alcohol consumption (scale of 1 to 5) negatively correlates with school performance (measured on a scale of 1 - 20), but there seems to be no correlation between level of alcohol consumption and study hours [4]. We found another research attempt on the relation between alcohol consumption and academic performance (reference 5), but the conclusions aren’t clear. Nevertheless, the main takeaway from reference 5 is that similarly to reference 4, the level of alcohol consumption does negatively correlate with school performance, particularly in terms of grades in that study (same scale of 1 - 20) [5]. Finally, there are a few attempts at determining what factors correlate with high school student alcohol consumption (reference 6 for example); however, the majority of these projects approach this from a pure statistical and machine learning perspective. The study in reference 6 found that students tend to drink more on weekends, parental cohabitation status don’t have any noticeable correlation with alcohol consumption, etc. [6] Thus, we want to add onto this part of research by using an approach that focuses on data analysis and visualization instead. We hope that a different methodology of research would help reveal any new correlations and confirm (or correct) existing discoveries.

References (include links):

* 1) <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3125885/>
* 2) <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1530-0277.2007.00402.x>
* 3) <https://www.collegedrinkingprevention.gov/media/journal/040_090.pdf>
* 4) <https://www.rpubs.com/heke0495/419722>
* 5) <https://www.kaggle.com/kanncaa1/does-alcohol-affect-success>
* 6) <https://www.kaggle.com/jhuno137/machine-learning-on-student-alcohol-consumption>

# **Hypothesis**

High school alcohol consumption is a persistent issue in the United States. Considering that the legal drinking age is 21, this issue poses a major concern for future young generations if it continues to be prominent. We hypothesize that parental and family factors, in contrast to school performance or individual factors, have stronger correlations with frequent alcohol use by high school students in the United States.

# **Data**

The ideal data set for us would consist of the following attributes:

* The most important attribute for an ideal data set is that it’s tabular.
* It should contain all the specified information in our research question. To reiterate, it should at least have individual factors such as sex and age, parental and family factors such as their education, size, income, etc., and school performance factors such as workload/free time, educational support, grades, etc.
* The ideal data set should also store the aforementioned data categories in individual columns.
* The data in the columns should either be quantified or formatted in a way that allows us to be able to quantify them.
* The data set should cover a relatively wide range of high school students. The ideal data set in terms of high school student coverage should be at country-level, with the amount of data gathered from each state proportional to the number of high school students in that state.
* Because some factors such as grades could could be measured differently, the ideal data set would have one type of measurement for all types of factors (for example, in our Background and Prior Work section, we noted that there is a grading scale that only goes up to 20, which is different from the grading scale of 100 most teachers and professors use in the United States). Preferably, the data are gathered from the same survey around the same time period to reduce variation in responses.
* We don’t have a specific number of observations in mind, but for reasonable inferences to be made, the data set should have a large number of data, perhaps at least in the ten thousands to provide a solid basis for our analysis.
* Ideally, to reduce bias, data shouldn’t be based on surveys responses from students but instead by people who observe students such as teachers. This does raise some ethical and privacy concerns, however, especially when trying to figure out how much a student is drinking in the first place.
* To further reduce bias, it should have no missing information or gaps.
* Because we’re not trying to see how high school alcohol consumption changes over time, the data set will probably be most useful if the data collected is recent and over a relatively short period of time, such as a year.

# **Ethics & Privacy**

There are quite a few ethical and privacy concerns when conducting research on high school students. As noted above in the notes for our ideal data set for this project, some students may not be comfortable with giving as much information as we desire, especially information regarding their individual factors, familial factors, and alcohol consumption frequency. In order to mitigate these privacy concerns, the data may need to be from surveys, which are most likely to contain data that have been consented to. Additionally, we can potentially further mitigate these concerns by using public data, especially one from a well respected institution that follows ethical guidelines, and data that excludes any student identification components. Furthermore, the usage of public data will help remove the need to encrypt and consider secure storage on our part.

Regarding any biases in our data, because we’re most likely going to use data from surveys, there is a good chance that students may skew their responses. For our group, there isn’t much of a way to correct for this, but what we can do to try to address this issue is to make sure we follow what is discussed in the paragraph above: find data that have been collected without student identification components. That way, they are less inclined to misrepresent themselves in their responses. There is also a concern with bias due to data entries themselves. There could be entries with missing information, which could skew our analysis. We plan to mitigate these biases by simply pruning all the missing information.

We also need to consider the stakeholders and how they may be affected, which include parents, students, friends, schools, and possibly even the states involved. The most important one to consider are the students themselves. If it turns out that we find strong correlations with alcohol consumptions for specific factors, especially with individual factors such as sex, age, and etc., our findings may not be appealing to them and could contribute to increased negative effects for these individuals. Students that are associated with alcohol consumption-correlated factors may also be treated with bias from peers, educational institutions, or law enforcements. Like other projects, our project will only investigate correlation (in our case, between alcohol consumption and individual factors), which generally isn’t enough to prove causation from a statistical standpoint. However, there is a risk that stakeholders such as the public, institutions, etc. may misinterpret our findings as causes of illegal drinking for minors. For example, if we find this to be true, the public may misinterpret “low income correlates with alcohol consumption” as low income causes alcohol consumption. Moreover, there may be the possibility of some misunderstanding of our data which could end up causing misconceptions that influence institutions/policy makers.

**Team Expectations**

* The team will have weekly work meetings where we will dedicate an hour to collaborating on the project.
* In cases where the team needs to meet more frequently or dedicate more hours, everyone should be actively communicating in the group chat to arrange extra time.
* Should any emergency arise, one should make sure to notify the group chat for the rest to be aware of the situation.
* If anyone is struggling with designated work, he or she should communicate with the group in order for the rest to help or accommodate (don’t be afraid to ask for help!).

TEAM EXPECTATIONS: Read over the [COGS108 Team Policies](https://github.com/COGS108/Projects/blob/master/COGS108_Team_Policies.pdf) individually. Then, include your group’s expectations of one another for successful completion of your COGS108 project below. Discuss and agree on what all of your expectations are. Discuss how your team will communicate throughout the quarter and consider how you will communicate respectfully should conflicts arise. By including each member’s name above and by adding their name to the submission, you are indicating that you have read the COGS108 Team Policies, accept your team’s expectations below, and have every intention to fulfill them. These expectations are for your team’s use and benefit — they won’t be graded for their details.

Excel sheet for all the previous projects: <https://docs.google.com/spreadsheets/d/1Y6okRFiiSlcIUTRxFOJx0m-1Fl3zdsnseotNYSwBhvY/edit#gid=0>

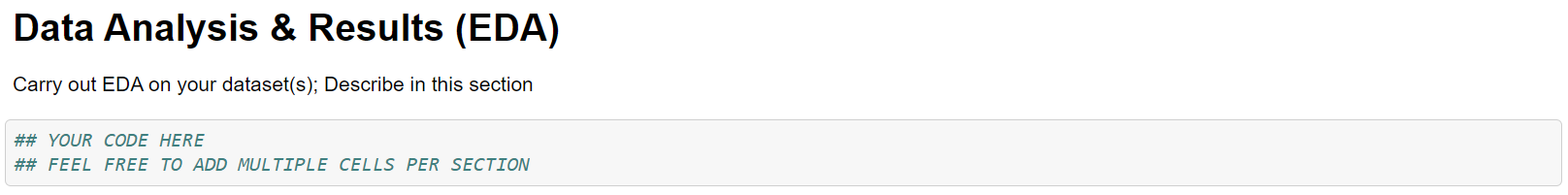
Ethics checklist:

<https://deon.drivendata.org/#data-science-ethics-checklist>

Project guidelines:

<https://github.com/COGS108/Projects/blob/master/FinalProject_Guidelines.md>

**Project Timeline Proposal**



|  |  |  |  |
| --- | --- | --- | --- |
| Date | Time | Before Meeting | What to Discuss |
| 1/31 | 2PM | Search for datasets that relate to our research question | Go over the datasets and decide which would be best, what analysis we would like to do on them |
| 2/7 | 2PM | Data cleaning assigned to each member | Assign roles for the project report (mainly coding, writing, analysis etc) |
| 2/14 | 2PM | Brainstorm ideas for EDA, specifically what parts of the data we want to compare and how we will visualize them | Review progress on EDA, discuss which analysis methods will work best |
| 2/21 | 2PM | Same as the previous row: continue brainstorming ideas for EDA | Review progress on project for the check-in |
| 2/28 | 2PM | Start writing code based on our ideas from the previous two meetings | Share each other’s code and start figuring out what to put on Github |
| 3/7 | 2PM | Have code and analysis finished, begin writing conclusion and finalizing report | Discuss the video portion of the project in terms of editing and content, review conclusion |
| 3/14 | 2PM | Finish up video and finalizing report (conclusion, formatting, cleaning up code) | Turn in the project and group survey! |

Datasets for high school alcohol consumption:

<https://data.world/nces/elem-and-secondary-stimulants>

Never mind, it’s really hard to find data with a level of quality that even comes close to the current one we have

<https://chronicdata.cdc.gov/Youth-Risk-Behaviors/DASH-Youth-Risk-Behavior-Surveillance-System-YRBSS/q6p7-56au> (not sure if this is helpful, very complicated dataset)

<https://data.world/data-society/student-alcohol-consumption>

(check this one out)

Relevant previous projects:

[~~https://github.com/COGS108/FinalProjects-Wi20/blob/master/FinalProject\_group077.ipynb~~](https://github.com/COGS108/FinalProjects-Wi20/blob/master/FinalProject_group077.ipynb) ~~(missing EDA)~~

[~~https://github.com/COGS108/FinalProjects-Wi20/blob/master/FinalProject\_group049.ipynb~~](https://github.com/COGS108/FinalProjects-Wi20/blob/master/FinalProject_group049.ipynb) ~~(missing EDA)~~

<https://github.com/COGS108/FinalProjects-Sp19/blob/master/FinalProject_Group102.ipynb>

[~~https://github.com/COGS108/FinalProjects-Sp19/blob/master/FinalProject\_Group138.ipynb~~](https://github.com/COGS108/FinalProjects-Sp19/blob/master/FinalProject_Group138.ipynb) ~~(EDA is there, but it’s all visuals with no explanations)~~

[~~https://github.com/COGS108/FinalProjects-Sp19/blob/master/FinalProject\_group174.ipynb~~](https://github.com/COGS108/FinalProjects-Sp19/blob/master/FinalProject_group174.ipynb)

~~(missing EDA)~~

<https://github.com/COGS108/FinalProjects-Fa20/blob/main/FinalProject_group008.ipynb>

<https://github.com/COGS108/FinalProjects-Sp20/blob/master/FinalProject_group31.ipynb>

COGS 108 Lectures:

<https://github.com/COGS108/Lectures-Wi21>

Link to our repository:

<https://github.com/COGS108/group030_wi21/blob/main/DataCheckpoint_group030.ipynb>

<https://www.kaggle.com/uciml/student-alcohol-consumption>

~~Data cleaning ideas:~~

* ~~Quantify variables:~~
  + ~~Sex (binary)~~
  + ~~Famsize (binary)~~
  + ~~Pstatus (binary)~~
  + ~~Mjob~~
    - ~~Teacher: 0~~
    - ~~Health: 1~~
    - ~~Services: 2~~
    - ~~At\_home: 3~~
    - ~~Other: 4~~
  + ~~Fjob~~
    - ~~Teacher: 0~~
    - ~~Health: 1~~
    - ~~Services: 2~~
    - ~~At\_home: 3~~
    - ~~Other: 4~~
  + ~~Guardian (binary)~~
  + ~~Schoolsup (binary)~~
  + ~~Famsup (binary)~~
  + ~~Paid (binary)~~
  + ~~Activities (binary)~~
  + ~~Higher (binary)~~
  + ~~Internet (binary)~~
  + ~~Romantic (binary)~~
* ~~Merge Dalc and Walc~~
  + ~~Average them together~~
  + ~~Add them together for a scale of 2-10~~
* ~~Deal with G1, G2, and G3~~
  + ~~Average them~~
  + ~~Add them together~~
  + ~~Only use G3 since that’s the final grade~~
* ~~Failures: check for non-numeric entries (edit: do this for all columns)~~
* ~~Check for out of range values~~
* ~~“Transpose” the table that describes the column names and the codings~~

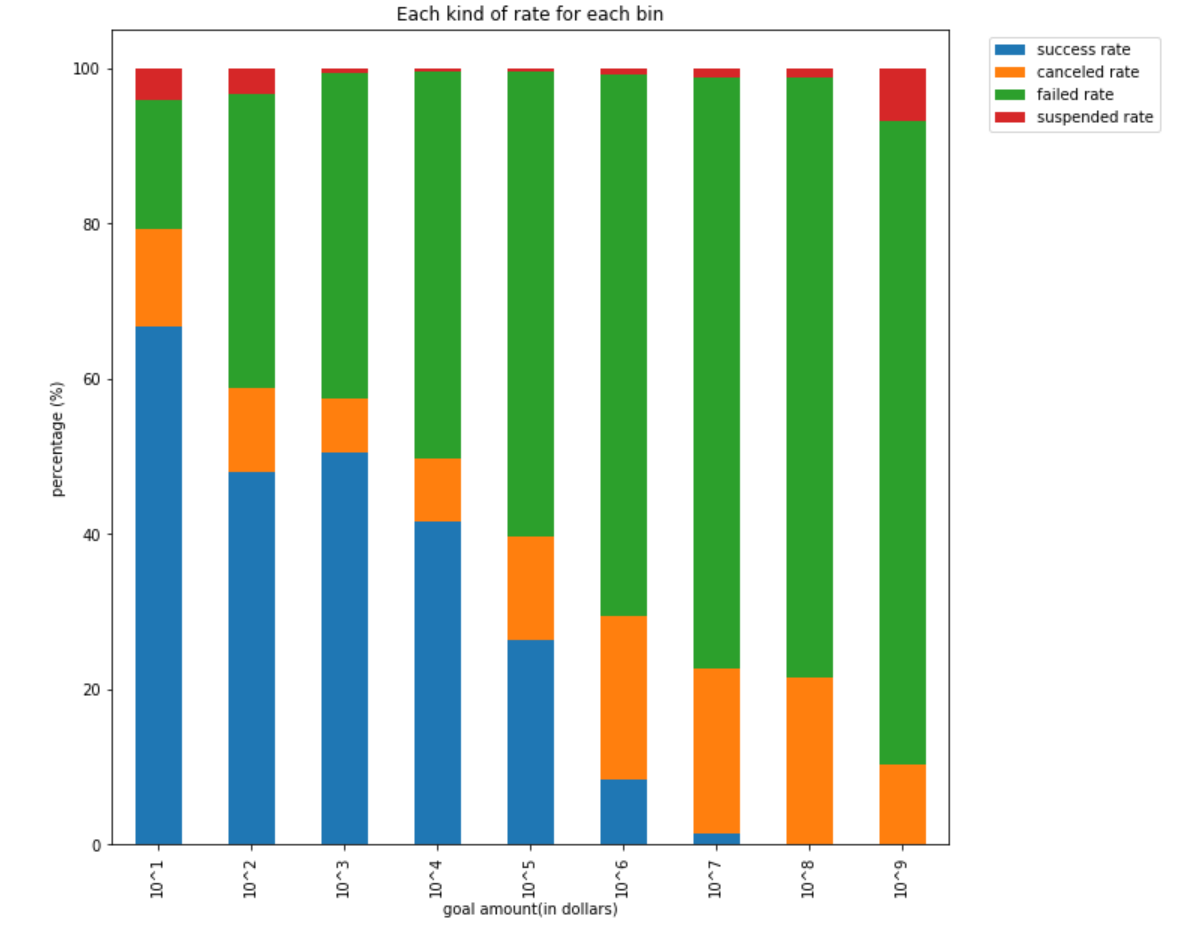
Some things to consider for descriptive analysis:

* Transform skewed data by using log transformation and maybe a shift
* Consider outliers and if they should be pruned (don’t prune them if they are legitimate)
  + Prune if you can confirm they are due to poor sampling errors, data entry errors, machine errors, etc.
* Talk about central tendency:
  + Mode: Use when talking about categorical variables
  + Median: Use when there are some outliers or severe skews
  + Mean: Use this otherwise
* Talk about variability:
  + Range: Not really used much other than to just describe your data
  + IQR:
  + Variance and standard deviation:
* Don’t use describe() to show information. You should only highlight some things, not all.

Things to consider for EDA:

* Structure:
  + What format is your data in?
    - Tabular (this is how our data is formatted)
    - Nested
  + Is each observation in a separate row?
    - Our data is already formatted like this, so we don’t need to do anything
  + Join multiple datasets?
    - Doesn’t apply to us
* Granularity:
  + What does each row represent?
    - A single student in our dataset
* Scope:
  + We need to change our question to Portuguese students
* Temporality: there are no reported times in our dataset. It’s all from 2008
* Faithfulness:
  + Are these values realistic? Are they reasonable?
  + Determines how trustworthy the dataset is

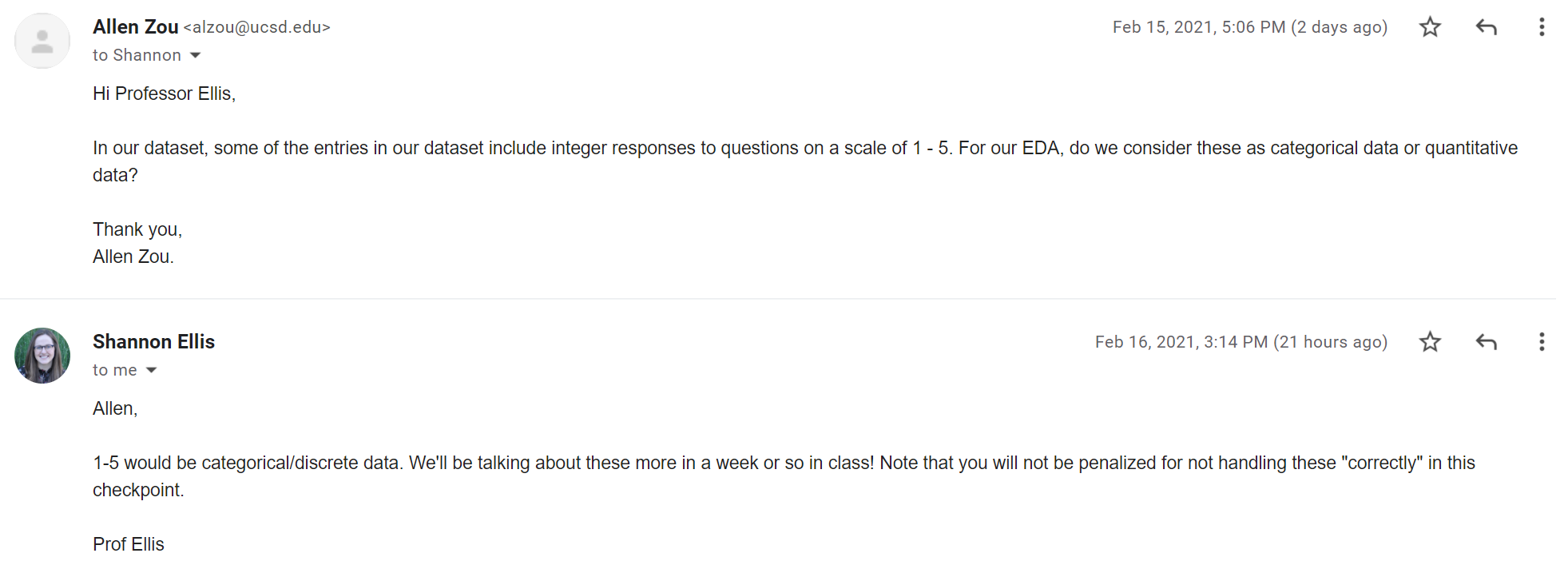
Ideas for our dataset:

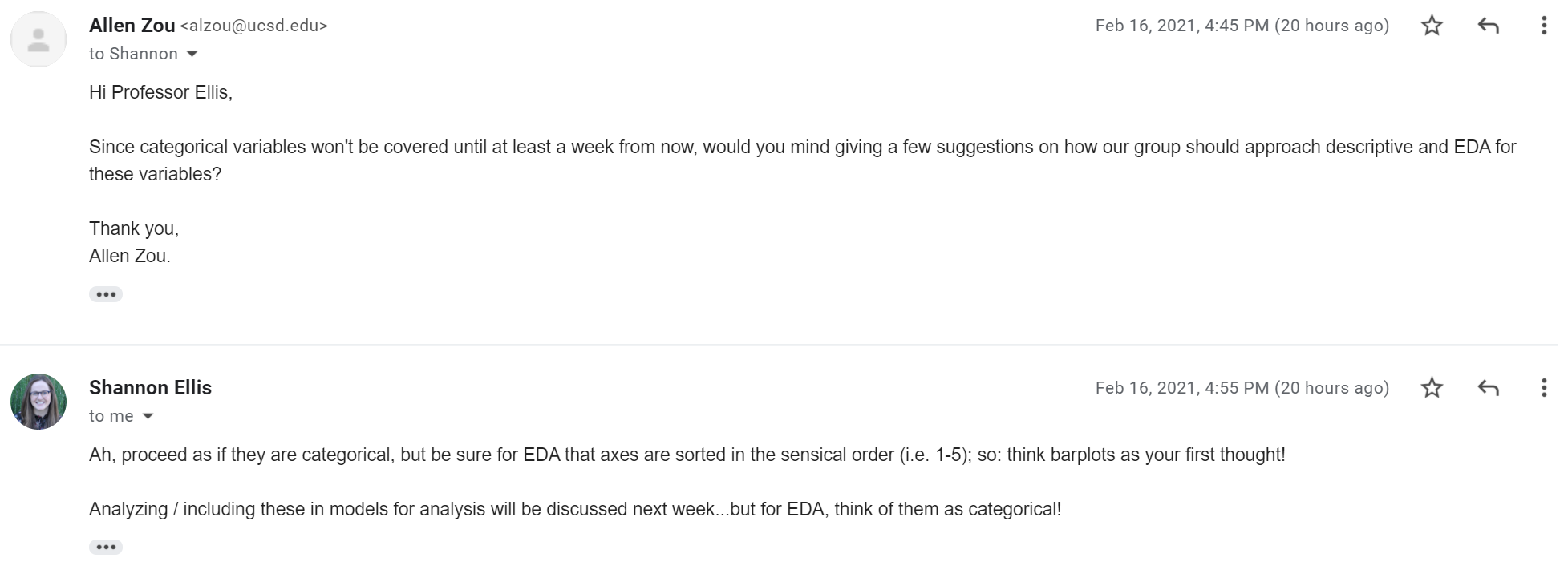
* After rewatching the lectures and looking at the previous projects, it seems like a ton of the group projects don’t have a section labeled specifically for EDA. Rather, their EDA is actually merged with another part of the project. We should try to keep EDA separate from other parts of the project.
  + I think some of the things we did in data cleaning should actually be in EDA.
* Talk about structure, granularity, scope, temporality, and faithfulness throughout the EDA. Faithfulness should probably be saved for the end of EDA.
* Should first identify which variables are categorical or continuous.
  + Do the following with continuous variables:
    - Central tendency (mean and median only)
    - Check the legitimacy of outliers
    - Compare variability among different data
    - Check the relationship between variables
      * Perhaps do a scatterplot matrix?
    - Check the distribution. Do tests to check if they are normal.
  + Do the following with categorical variables:
    - Central tendency (mode only)
    - One thing to note is that most of our responses, while they are initially given as numbers, might be categorical because responses are discrete values
    - Probably should use bar charts or pie charts for these
    - Maybe look at frequencies and proportions?
      * Something like this may work
    - Can still use IQRs and medians for ordinal categorical data (which our data is most likely to be)
    - Needs confirmation, but I think you can still see relationships between variables if the data is categorical

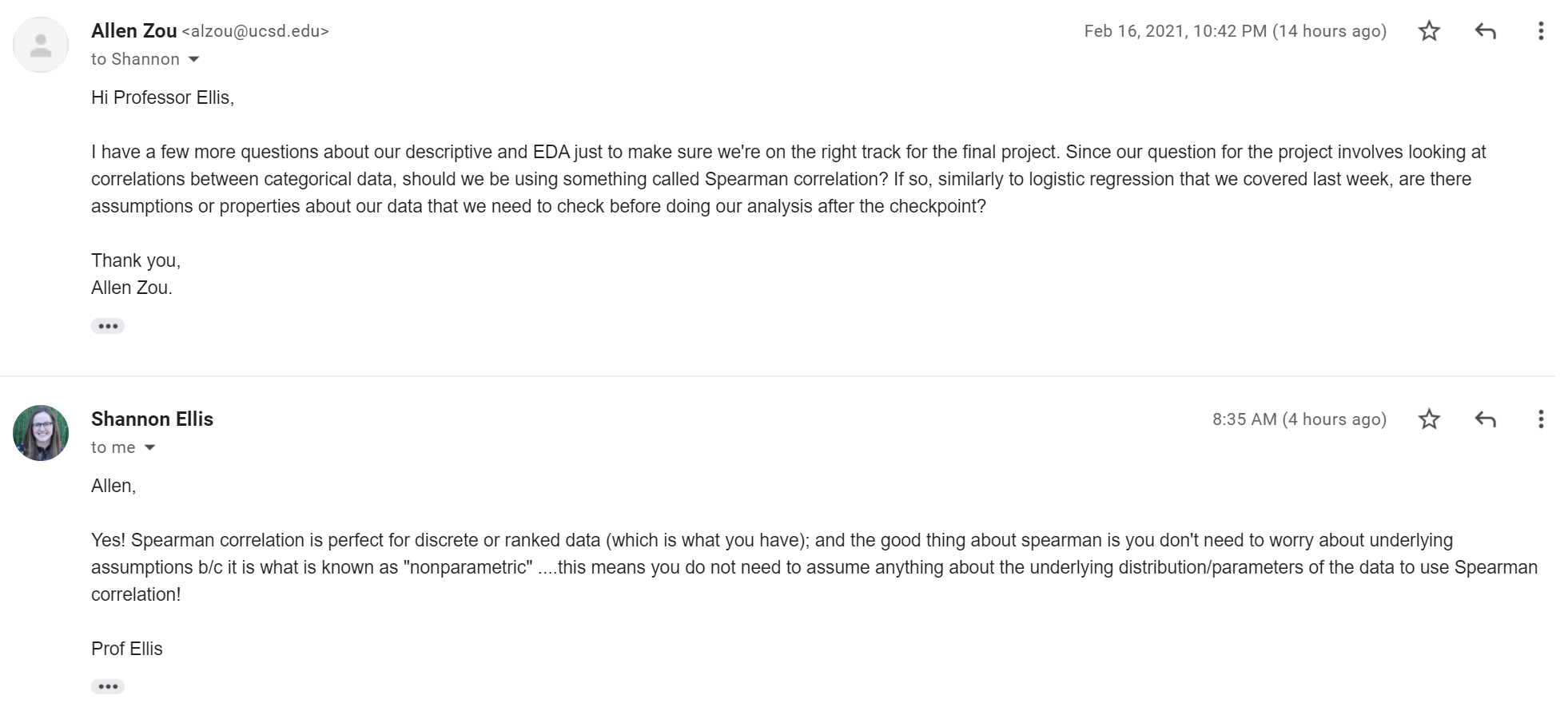
Questions (in red) for Professor Ellis during office hours:

* Should some stuff in the data cleaning section of our project be moved to descriptive and EDA?
  + We have a part that takes a look at period 1 grades, period 2 grades, and final grades to see how much of a difference there is between the period grades and final grades.
    - Additional question: we pruned the ones that have a large difference between the period grades and final grades, although there weren’t many that we removed, specifically 43 students out of the 649 we had. Should we revert this or is this fine?
  + A part that merges the weekday and weekend alcohol consumption frequencies (data cleaning)
  + A part that performs a ton of checks to see if the response values are valid
* ~~Because our question is about correlation, can we still do linear regression even with categorical data?~~
  + Yes, we should do something called Spearman correlation, which doesn’t require any assumption checks; however, we do have another question:
    - Our final grade column consists of integer values. Are they considered categorical or quantitative? If it’s quantitative, then how do we find the correlation between a quantitative and categorical variable? Unless we also use Spearman correlation for this, are there any assumptions or checks we need to make for this case?
    - Change question to what is the relationship between the variables
    - How do we want to model the predictors with a categorical outcome - read more into this
    - Find additional US dataset
  + Based on what I’m seeing, we should be using something called Spearman correlation instead of Pearson. In lecture, the checks and assumptions for regression/correlation typically involve checking for normal distributions, doing log transformations, or doing checks for linear regression, etc. but this may not be the same for ordinal data. I’m not sure what we should be checking in our EDA for this before we do analyses in the final project.
    - There are no checks or assumptions for Spearman correlation
* Do we need as many columns as we currently have? Update: should we focus only on columns/variables that we asked in our question?
  + Question and hypothesis involves the following categories (variables/columns are listed under those variables):
    - Personal factors
      * Sex
      * Age
      * In a romantic relationship
      * How often the student hangs out with friends
      * Health rating of student
    - Parental and familial factors
      * Family size
      * Parental cohabitation status
      * Mother’s education level
      * Father’s education level
      * Mother’s job
      * Father’s job
      * Student’s guardian
      * Quality of family relationship
    - School performance factors
      * Study time
      * Past failures
      * Extra educational support
      * Family educational support
      * Paid extra classes within the Portuguse subject
      * Extra curricular activities
      * Has access to internet
      * Amount of free time
      * Absences
      * Final grade
    - Alcohol consumption
      * Alcohol frequency
  + ~~If all of these are fine, then is there anything we should be wary of when answering our question? How should we go about answering our question given the large amount of independent variables?~~
  + Will we need these new datasets by the time we turn in the second checkpoint?
  + Possibly a general question, but would you mind explaining what our new approach should be given that we’re now measuring the relationship between certain factors?
    - Question: What is the relationship between certain factors and alcohol use by high school students in the United States, particularly with individual factors such as sex and age, parental and family factors such as their education, size, income, etc., school performance factors such as workload/free time, educational support, grades, etc.?
    - Outcome: alcohol use
    - Predictors: the variables in the question
    - Model: Similarly to what we’ve done in lecture (poverty percentage, violent crime, etc.), we should be building a model.
* Questions about the assignment rather than the project
  + In lecture, we were doing logistic regression on Teen birth rate, poverty percentage, and violent crime, which are quantitative variables, but on assignment 3, we’re doing logistic regression on height, major, and gender, the latter 2 being nominal categorical variables. This creates rows with gender[T.male] and major[T.COMPSCI], so I’m not sure if I’m doing regression correctly.

Possibly a helpful email conversation:







Meeting for 2/17 (Notes for Professor):

* Having too many variables to study; confounding variables
* Clean up questions to ask for OH
* Spearman correlation studying after

YRBS Dataset:

* There are a ton of questions in these surveys, some of which go up to 100 questions
* There are missing entries, so we should select the columns we want first before pruning them
* Some years have missing information

Link to the new dataset:

<https://www.cdc.gov/healthyyouth/data/yrbs/data.htm>

-this survey contains a lot of questions on relevant risk behaviors (drink and drive, marajuana, cigarettes etc) but little information about family and personal factors. We can potentially analyze the correlation between various behaviors? (i.e. Are high school students who smoke more likely to drink? Are high school students who have suicidal thoughts more likely to drink? Are high school students who eat healthy necessarily drink less? etc)

Columns that might be relevant:

Q1. How old are you?

Q2. What is your sex?

Q40. How old were you when you had your first drink of alcohol other than a few sips?

Q41. During the past 30 days, on how many days did you have at least one drink of alcohol?

Q42. During the past 30 days, on how many days did you have **4** or more drinks of alcohol in a row, that is, within a couple of hours (if you are **female**) or **5** or more drinks of alcohol in a row, that is, within a couple of hours (if you are **male**)?

Q43. During the past 30 days, what is the largest number of alcoholic drinks you had in a row, that is, within a couple of hours?

Q44. During the past 30 days, how did you **usually** get the alcohol you drank?

Q89. During the past 12 months, how would you describe your grades in school?

1999 Data Manuals

1. How old are you?

2. What is your sex?

3. In what grade are you? (Can be excluded)

4. How do you describe yourself? (Can be exluded)

5. During the past 30 days, on how many days did you carry a weapon such as a gun, knife, or club?

6. During the past 30 days, how many days did you not go to school because you felt you would be unsafe at school or on your way to or from school?

7. During the past 12 months, how many times were you in a physical fight?

8. During the past 12 months, did you ever feel so sad or hopeless almost every day for two weeks or more in a row that you stopped doing some usual activities?

9. During the past 12 months, did you ever seriously consider attempting suicide?

10. During the past 30 days, on how many days did you smoke cigarettes?

11. During the past 30 days, on how many days did you use chewing tobacco or snuff, such as Redman, Levi Garrett, Beechnut, Skoal, Skoal Bandits, or Copenhagen?

12. During your life, on how many days have you had at least one drink of alcohol?

13. During the past 30 days, on how many days did you have at least one drink of alcohol?

14. During the past 30 days, on how many days did you have five or more drinks of alcohol in a row, that is, within a couple of hours?

15.

During your life, how many times have you used marijuana?

During your life, how many times have you used any form of cocaine, including powder, crack, or freebase?

During your life, how many times have you sniffed glue, or breathed the contents of aerosol spray cans, or inhaled any paints or sprays to get high?

During your life, how many times have you used heroin (also called smack, junk, or China White)?

During your life, how many times have you used methamphetamines (also called speed, crystal, crank, or ice)?

During your life, how many times have you taken steroid pills or shots without a doctor's prescription?

16.

During the past seven days, how many times did you drink 100% fruit juices such as orange juice, apple juice, or grape juice?

During the past seven days, how many times did you eat fruit?

During the past seven days, how many times did you eat green salad?

During the past seven days, how many times did you eat potatoes?

During the past seven days, how many times did you eat carrots?

During the past seven days, how many times did you eat other vegetables?

During the past seven days, how many glasses of milk did you drink?

17.

On how many of the past seven days did you exercise or participate in physical

activities for at least 20 minutes that made you sweat and breathe hard?

On how many of the past seven days did you do exercises to strengthen or tone your muscles, such as push-ups, sit-ups, or weight lifting?

Questions after checking for issues in all manuals (red means unusable, orange means usable starting from 1999, and green means no issues):

How old are you?

What is your sex?

In what grade are you? (Can be excluded)

During the past 30 days, on how many days did you carry a weapon such as a gun, knife, or club?

During the past 30 days, how many days did you not go to school because you felt you would be unsafe at school or on your way to or from school?

During the past 12 months, how many times were you in a physical fight?

During the past 12 months, did you ever seriously consider attempting suicide?

During the past 30 days, on how many days did you smoke cigarettes?

During the past 30 days, on how many days did you have five or more drinks of alcohol in a row, that is, within a couple of hours?  
During the past 30 days, on how many days did you have at least one drink of alcohol?

During your life, how many times have you used marijuana?  
During your life, how many times have you used any form of cocaine, including powder, crack, or freebase?  
During your life, how many times have you taken steroid pills or shots without a doctor's prescription?

During the past 12 months, did you ever feel so sad or hopeless almost every day for two weeks or more in a row that you stopped doing some usual activities?

During the past 30 days, on how many days did you use chewing tobacco or snuff, such as Redman, Levi Garrett, Beechnut, Skoal, Skoal Bandits, or Copenhagen?  
During your life, how many times have you sniffed glue, or breathed the contents of aerosol spray cans, or inhaled any paints or sprays to get high?  
During your life, how many times have you used heroin (also called smack, junk, or China White)?  
During your life, how many times have you used methamphetamines (also called speed, crystal, crank, or ice)?

During the past seven days, how many times did you drink 100% fruit juices such as orange juice, apple juice, or grape juice?  
During the past seven days, how many times did you eat fruit?  
During the past seven days, how many times did you eat green salad?  
During the past seven days, how many times did you eat potatoes?  
During the past seven days, how many times did you eat carrots?  
During the past seven days, how many times did you eat other vegetables?  
During the past seven days, how many glasses of milk did you drink?

How do you describe yourself? (Can be excluded)

During your life, on how many days have you had at least one drink of alcohol?

During the past seven days, how many times did you eat potatoes?

On how many of the past seven days did you exercise or participate in physical activities for at least 20 minutes that made you sweat and breathe hard?

On how many of the past seven days did you do exercises to strengthen or tone your muscles, such as push-ups, sit-ups, or weight lifting?

Some things to note about these questions:

* Only 5 questions can’t be used
* About half of the remaining questions are only usable starting from 1997, a large portion of which are questions about healthy eating habits
* I groups questions together that we’re most likely combining (removed the spacing between questions)
* Number of grouped questions in each color section:
  + 10 green groups
  + 3 orange groups
  + 5 red groups
  + If we’re considering using the orange questions, then one of the orange groups can be combined with a green group, which makes it 12 groups instead of 13 groups.

Overall, I think we should prune the 1991 - 1997 datasets and use the orange questions. We don’t need to decide how to group these columns for the EDA checkpoint, but we can certainly start brainstorming now if you guys want.