Survey data notebook to do list

Step 0:

* Data csv exploration
* How many questions? (Q1-52)
* Format of question number and name

1. Q26\_1, Q26\_2 2 sub-questions under Q26, single questions but are similar

2. Q13 single question

3. Q12\_0\_1 rank of answer

4. Q11\_4\_TEXT indicate the “Other” choice is numbered/coded as 4, the 4th choice is a blank for you to fill in

* what types of questions are there? (likert, degree, text open ended, rank, yes/no, choices, importance degree, )

1. survey respondents’ information ()

2. IVs (college, gender, age, transfer, ethnicity, sexuality, )

3. questions

* which columns are IVs? Which are questions? [Column A to Q, col 1-17] [Column col 18- ]
* What each question asks about? How to analyze each question (question analysis readme file, discuss with Monica)
* Which question is a sub-question of another?
* What columns to drop? Are id, duration, finished, and progress important for the data analysis?
* Choose questions that have fewer missing values
* Responses all from Berkeley students in our data
* Probably want to not use columns that are only answered if another question is answered in a particular way

But many questions that have this pattern, some questions only for students who voted and registered for voting ???

* Cleaning and make notebook question, go through data cleaning process, at the same time
* Cleaning data and find the “need to improve” part of the survey design
* Too many options – people don’t read all of the options
* Qualtrics has the option to randomize the order of options (original student probably didn’t choose this, so look if there is a tendency to pick the first few options)
* Double barrelled question – responders may have conflicting views but not able to distinguish between these
* Leaving a text box makes more NaNs than yes/no Many NaNs in our data mean “no”, Signal is “yes”, “Decline to answer”
* People may have tried to breeze through and not seriously answer in order to be able to enter raffle (i.e. all first option, skip all questions)

Step 1: Data cleaning

1. missing value percentage for each question, non-real answers, non-serious answers

2. how to deal with missing values (delete/replace with median?/ or?)

3. replace question title string with numerical value (done)

4. select useful columns, drop columns (which questions to drop)

5. drop rows (decide which response to kept depends on percentage completion)

6. replace likert question answer with numerical scale (one hot encoding 001 010 …)

7. datascience package teaching (borrow from data8 lab) while data cleaning walk through

8. write some internal functions that used to solve some complicated issues and just tell them how to use the function to reach their goal

**Data excel explore 101**

1. First row is good for question and variable label , second and third row can be deleted

2. columns that can be deleted:

First name, last name, email, external data reference,

3. no Q2, Q38 in the original data

4. skip questions

No answer for Q10 – Q19 (if Age = 18, 19)

5. multiple answer questions: Q8, Q20

6. Q 39 - Q 50 political opinions, campus political, voting system evaluation

7. question answer that need to convert to numerical values: Q26, Q39, Q40, Q42, Q43, Q47, Q48

Step 2: EDA

* Comparisons each question between multiple variables (IV for main questions: gender, major, age, ~~voter/nonvoter~~)
* Summary statistic and visualization on multiple choice responses (questions with numerical answers)
* Histogram (continuous variable x-axis, how to read it, height of bar, area)
* Bar chart (discrete variable on x-axis)
* See if there are examples of more discrete or more continuous data
* Text/open ended/type in question analysis? most of them are the “Other” choice within each question. sentiment analysis like DS100?
* Ordinal vs nominal data
* Pivot table
* Explore correlation between responses of questions (responses are discrete)
* Fairly representative sample of university-wide ethnic makeup, less so for gender & STEM/humanities
* Question about whether or not students are comfortable discussing their political views was significant
* Survey design addressing and recommendation while exploring data

ASK:

1. which columns should we choose? ()

2. CANNOT omit all the questions that can be answered if some other question answered because the first and second main part of the survey is based on the answer if you vote or not

Check missing percent for each question, see if the skipped questions have more missing percent than not skipped questions, then drop.

3. how deep do you want to go for causation and correlation ?

4. how many questions answered/ how many percent answer count as complete?

5. what issues/ topics the second notebook should include and address

6. easy/ hardness of the notebook, for example, XXXXX = surveydata.select(....) or XXXXX = ….

Fill out each code for them for now, just like an solution, then can delete later

7. does notebook has a limited memory? Run out of memory?

8. If we borrow other notebook and data8 lab, textbook, should we write citation?

Try to use our own words and write a “adapted from lab 2” if use their code and their language a lot

9. do we combine introduction to tables and data cleaning together? I mean teach how to use Table while doing the cleaning, or intro first and then apply? Do we need another cvs table as example to teach them how to use datascience Table or just teach them while doing cleaning?

Intro to python

Teach Tables while walk them through the cleaning part

11. how to link to an exactline in a html?

Need the html has the section link already

12. do you think it would be good if we wrap up cleaning steps into one function?

Too many cleaning stuff

13. why cannot replace ‘nan’ with np.nan? is there a better way to handle missing value ‘nan’ in the Table? Using pandas?

14. convert to numerical value, function

Write a function for the converting:

Within the utility function, check unique ?= type in list [“never”, ‘always’, ….]

Bc there might be no one answer “never”, throw an error if there

Arg 1 table, arg 2 column, arg 3 a list of level factors,