**1 Create a single summary data set of ratings data** using the data from all three years. Your new data set should include all rating scale response data (e.g. "Rating SFO" responses, "Cleanliness of SFO" responses, safetiness item responses, getting to the airport items, and so on; see below), unique respondent IDs, residence location (Q16LIVE in the 2016 data).

The response data are numeric data (although no necessarily metric) and should be treated as such. Make sure that the records in your data set can be tied back to the original data sets by way of the unique respondent IDs or by other means. Include a variable indicating the year of data collection. Each row in this data set should correspond to one survey respondent. The columns (or fields in the csv file you'll write) should be the ratings and other variables.

Rating scale data are data that are captured by getting reponses on a numeric scale that indicates some kind of magnitude, degree, or ordering. For example, a cleanliness rating scale might have response categories ranging from 1 = poor to 5 = excellent. Other numeric codes may have been used in various places in the data. Note how the coding may differ from year to year. You'll need to "reconcile" diffeent codings in order to put the data from the different years together in a way that makes sense. Be sure you treat missing values in a sensible manner. (When a value is missing, it means that it's not there.)

This data set you are creating is intended for use in multi-year trend analyses. (You or your classmates may end up needing to use it later in the course.)

(a) List the the variables from each year that you used to create this data set using the original names they had in the data set they appeared in.

(b) Document any variable name changes so that it's clear what the original variables are whose names you changed. (Otherwise, a user of your data wouldn't be able to know what these variables are, or how to use them.)

(c) Describe this data set you created in terms of its size, the variables in it, and how the variables are coded. How many missing values do you have on the ratings variables?

(d) Write your new data set to a csv file with an initial header record that includes the variable names.

**2** **Identify the top three (3) comments made by gender of respondent in survey years 2015 and 2016.** Report the top three (3) most common comments based on their relative frequency for each gender category. Include each of the comments' text that's indicated in the data dictionary documents. For each comment, indicate its relative fequency.

The comments to consider for this part are Q8COM1-Q8COM3 in the 2015 data and Q8COM1-Q8COM5 in the 2016 data.

**3** ***Using the data you created in****1****, summarize the distribution of the SFO Airport "as a whole" ratings by respondent residence location category and report your results.*** These are ratings about the airport, overall. For example, in the 2016 data, the variable is called Q7ALL. Make sure how you do this takes into account the nature of the data for this variable.

**4** ***Profile respondents for follow-up participation in qualitative research by creating a data set that describes them.*** Further investigation of some of the issues apparent in the survey data is going to be pursued by inviting selected respondents to participate in follow-up research. The file select\_resps.csvidentifies the respondents to be targeted. These respondents were surveyed in either 2015 or 2016.

(a) Create a data set of these respondents that includes data on their demographics and on the travel behaviors they reported when they were surveyed. (See below.) Include for each respondent the date and time he or she was interviewed. (**INTDATE**)

(b) Save this data set as a headered csv file. Describe it in terms of its size and the variables in it. Be sure that you are clear on what what a demographic variable is, and what 'travel behavior' is. Be sure to preserve original variable names.

Here are the kinds of variables you should include in your profile data set:

* Respondent ID (**RESPNUM**)
* Year surveyed (**YEAR**)
* Destination geographic area (**DESTGEO**)
* Size of destination market (**DESTMARK**)
* Purpose(s) of travel (**Q2PURP1-Q2PURP3**) (Make sure that the meanings of the codes are consistent, e.g. that a code of "5" means the same thing year to year.)
* Used parking? (**Q3PARK**)
* Checked baggage? (**Q4BAGS**)
* Purchased from a store? (**Q4STORE**)
* Purchased in a restaurant? (**Q4FOOD**)
* Used free WiFi? (**Q4WIFI**)
* Times Flown in last 12 mo. (**Q5TIMESFLOWN**)
* First time flying out of SFO? (**Q5FIRSTTIME**)
* How Long Using SFO? (**Q6LONGUSE**)
* Residence Location? Bay Area, or ...? (**Q16LIVE**)
* Age (**2015: Q18Age, 2016: Q19Age**)
* Gender (**2015: Q19Gender, 2016: Q20Gender**)
* Income (**2015: Q20Income, 2016: Q21Income**)
* Fly 100K miles or more per year? (**2015: Q21Fly, 2016: Q22Fly**)
* Language version of questionnaire (**Lang**)
* Have used the San Jose airport (**2015: Q22SJC, 2016: Q23SJC**)
* Have used the Oakland airport (**2015: Q22OAK, 2016: Q23OAK**)

Be sure to "reconcile" any differences in coding across the years.

(c) Tabulate the fequencies of the codes for Parking, Times Flown, How Long Using SFO. (This means for each variable create a frequency table showing how many times each coded value occurs.)

**----Part 5---**

**5** ***Save each of the data sets you have created above by "pickling."*** Pickling is a venable Python technique for "serializing" Python objects. Verify that your pickling worked by unpickling. (If you are feeling adventurous, also try saving your data sets in a shelve database. A shelve database is like a permanent Python dictionary.)

Be sure to save all of your data as you might need some of it later in the course. It's usually a good idea to save your documented code, too, of course. Save early and often.