In this exercise you'll work with data collected over three years from persons at the San Francisco International Airport (SFO).

Like many organizations, SFO regularly assesses who its customers are, and what they like and don't like.

You can find out more about the SFO data at <http://www.flysfo.com/media/customer-survey-data>

The data you'll be working with is stored in different file formats and coded in somewhat different ways year to year. Some variables have the same names year to year, while others have names that vary over the years. It will become apparent to you as you review the "data dictionaries" provided. You'll be producing some summary results from the data and creating and storing new data sets for subsequent analyses and research.

This exercise, like much of 420, involves the use of the pandas Python package for cleaning and transforming data. You'll be using Python and pandas to input data and to do common data manipulation tasks. pandas is a very extensive package that provides many methods for working with and summarizing data. Online documentation is at [http://pandas.pydata.org](http://pandas.pydata.org/) is current, but can be a little overwhelming at times. The book by McKinney (the original author of pandas) is accessible but is dated. If you want a gentler introduction to pandas, Try Harrison's "Learning Pandas" book. It's an easy read. For starters pay particular attention to the DataFrames sections.

You'll find a lot about pandas and Python in general online at sites like [http://www.stackoverflow.com](http://www.stackoverflow.com/), [http://www.github.com](http://www.github.com/), and others.

Post your questions, issues, and other things about this assignment to the GrEx1 Huddle on Canvas. You're encouaged to help each other. But be sure to turn in your own work.

**The Data**

The SFO survey data for each of the years 2014, 2015, and 2016 are in zip files. Each zip file contains a data file that's either an xls file or a .csv, and a .pdf document that is the *data dictionary* for the file. This document indicates what the codes mean in each data field of the data file it is for. To understand what the data mean for any particular year, you need to use the information that's in the data dictionary for that year.

By inspecting the files and their data dictionaries you'll note that the file contents are similar from year to year, but are not identical. For example, the numeric or alphabetic codes used to represent specific things that survey respondents said may differ. In one year, a 1 may mean a "dirty fountain" cleanliness comment, while in a different year the same comment may be coded as a 9. So when working with the data from more than one year, attention to what the data codes are is necessary.

In addition to the survey data there is a csv file select\_resps.csv that contains the respondent IDs for respondents that are to be invited to participate in follow-up focus groups.

**What to Do**

In a Python session or a Jupyter Notebook, read each data set into a pandas DataFrame. Use the pandas input method appropriate for each file's type. You'll probably find the pandas input methods to be the most convenient way to read the data files. (If you are adventurous, read the xls files using the openpyxl package.)

Then, do the next five parts. Note that most of the following parts have more than one objective and/or deliverable.

**---Part 1---**

**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, safety-ness 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 responses 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" different 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 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.**

**A Little More on What Should Be in Your new Data Set for Part 1.**

The ratings scale variables that you should include are in the following categories:

* General assessments of aspects of the airport. (In the 2016 data, these are Q7ART through Q7ALL)
* Assessments of cleanliness (Q9BOARDING through Q9ALL in 2016)
* Impression of safety (Q10SAFE in 2016)
* Impression of PreCheck (Q12PRECHECKRATE in 2016)
* Impression of ease of getting to the airport (Q13GETRATE in 2016)
* Ease of finding way around the airport (Q14FIND in 2016)
* Ease of getting through security (Q14PASSTHRU in 2016)

Your data set should include ratings variables for the above categories that were measured *in any of the three years* of data collection. For example, if in 2015 a rating of loadspeaker announcement understandability was collected, this rating variable should be included in the data set you create, even if the measure was not used in 2016 and 2014. In other words, the ratings variables you should include should be the *union* of the sets of variables measured as ratings in any of the years, and not the intersection of the sets of rating variables measured.

You'll also need to include for each respondent an unique respondent identifier, the year that the respondent was surveyed, and where the respondent lives relative to the airport (Q16LIVE in 2016). You'll need the latter for #**3**, below.

**---Part 2---**

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

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

**---Part 3---**

**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.

**---Part 4---**

**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.csv identifies 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.**

**(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 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
* Year surveyed
* Destination geographic area
* Size of destination market
* Purpose(s) of travel (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?
* Checked baggage?
* Purchased from a store?
* Purchased in a restaurant?
* Used free WiFi?
* Times Flown in last 12 mo.
* First time flying out of SFO?
* How Long Using SFO?
* Residence Location? Bay Area, or ...? (Q16LIVE)
* Age
* Gender
* Income
* Fly 100K miles or more per year?
* Language version of questionnaire
* Have used the San Jose airport
* Have used the Oakland airport

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

**(c) Tabulate the frequencies 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.

**How to Do It**

Do this entire assignment using Python. (It's important to practice your Python skills for working with data. That's what this course is about.)

For each of the above objectives, include in your assignment submission syntactically correct, commented Python code, along with any results that each part requests. Do not first report all of your commented code, and then all requested results. Report any results for each part, part by part.

Note that your Python code must be *syntactically correct* in order to get credit for it. "Syntactically correct" doesn't mean that it necessarily works to produce the required result. What it means is that it is consistent with Python style guidelines, that indenting where necessary is correct, that lines of code are not truncated because of their length, that all parentheses and brackets are "balanced," and so on. "Correct" means that a person could copy and paste it and it would work without producing syntax errors. It also means that the comments included as sufficient to explain to another person what the code does. (Or is supposed to do.)

**Tips and Tricks**

Begin by reading all of the above, carefully.

Next, *plan out* what you want to do for each of the above parts of this assignment before jumping headlong into wailing away at the data. Outlining a plan of attack will help you organize your thinking about what needs to be done, which is useful even if you end up "disrespecting" your plan as you dig into the data.

Many things are easier to do if they are broken down into small steps. When you work with the survey data for the three years, think about first working with each year's data separately, and reviewing the dictionary for each one to make sure you understand what the variables are, and how they are coded. If you're not clear on the variables' names and coding, you're unlikely to use the data correctly for the above assignment parts.

There are some pandas methods that are handy for doing what it is you need to do. One often useful pandas "trick" is that you can stack (concatenate, append) DataFrames and columns in them that have the same name will (usually) end up being combined appropriately.

Another trick is that you can rename multiple columns of a DataFrame by providing a dictionary of old name, new name pairs to the DataFrame's .rename() method. A third trick is that you can use methods like .melt() to change a "short and wide" DataFrame to a "tall and narrow" one. We'll cover some of these in a Sync Session, but don't delay on looking into them yourself.

A good thing to do when starting on Part 1 is to "inventory" the ratings variables in the data sets for the three years to see which ones are in common across the three years of data, which ones aren't in all years of data, and whether the names of any need to be changed before combining the data from the three years. This is probably unavoidable variable "bookkeeping" if the assignment's objectives are going to be correctly met.