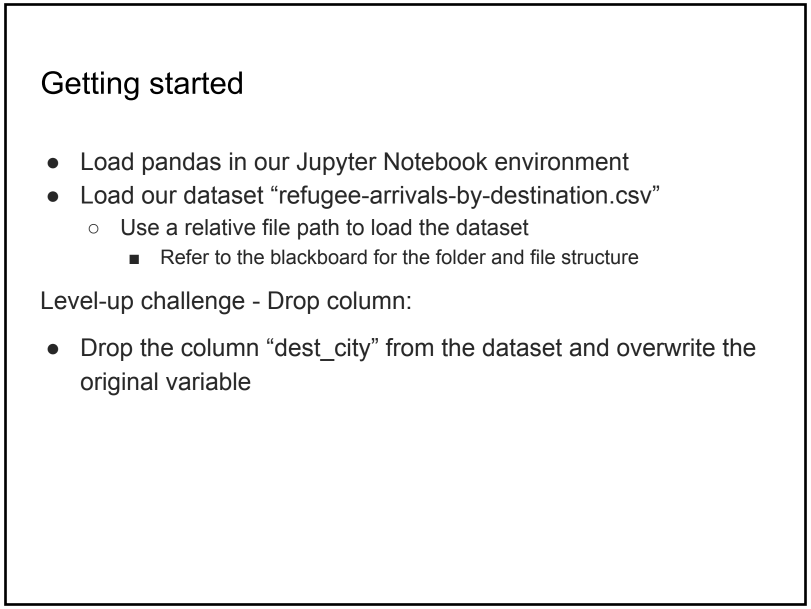
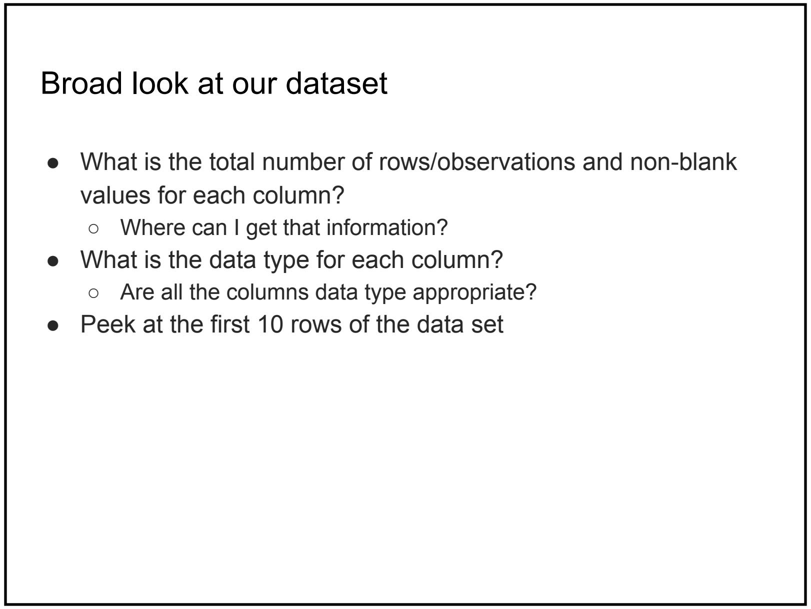


Getting started

- Load pandas in our Jupyter Notebook environment
- Load our dataset “refugee-arrivals-by-destination.csv”
  - Use a relative file path to load the dataset
    - Refer to the blackboard for the folder and file structure

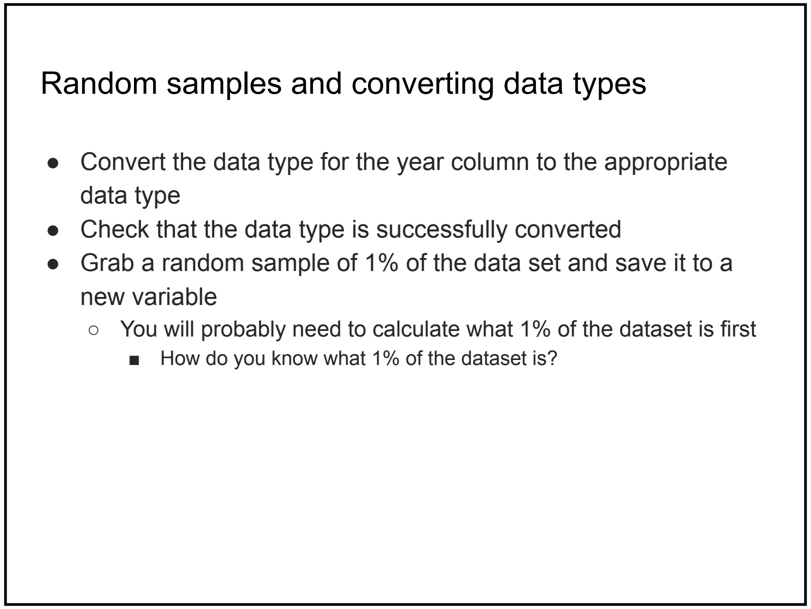
Level-up challenge - Drop column:

- Drop the column “dest_city” from the dataset and overwrite the original variable



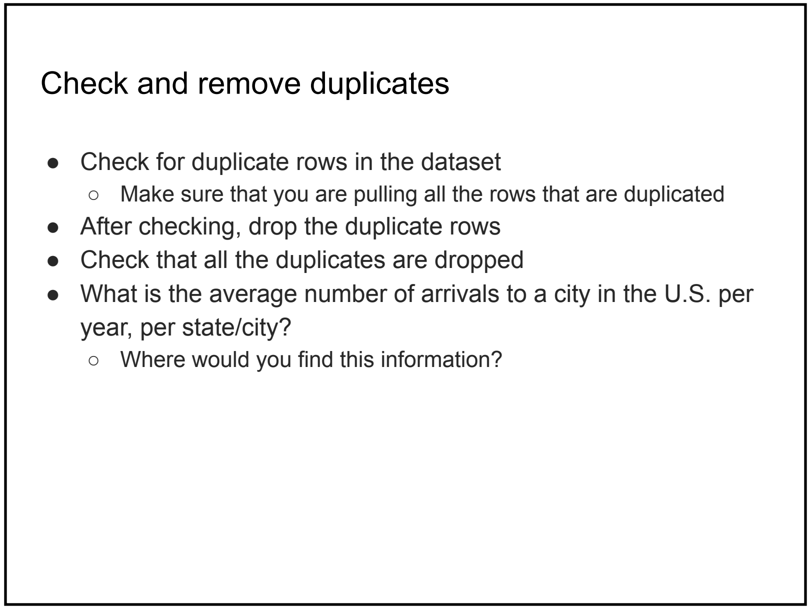
Broad look at our dataset

- What is the total number of rows/observations and non-blank values for each column?
  - Where can I get that information?
- What is the data type for each column?
  - Are all the columns data type appropriate?
- Peek at the first 10 rows of the data set



Random samples and converting data types

- Convert the data type for the year column to the appropriate data type
- Check that the data type is successfully converted
- Grab a random sample of 1% of the data set and save it to a new variable
  - You will probably need to calculate what 1% of the dataset is first
    - How do you know what 1% of the dataset is?



Check and remove duplicates

- Check for duplicate rows in the dataset
  - Make sure that you are pulling all the rows that are duplicated
- After checking, drop the duplicate rows
- Check that all the duplicates are dropped
- What is the average number of arrivals to a city in the U.S. per year, per state/city?
  - Where would you find this information?

Check for blank/NA values

- Identify all column(s) that have blank/NA values
- Pick one column to do further exploration
  - Create a new variable for the rows with NA values
  - How many NA rows are there for your chosen column?
- Replace blank/NA values with “no ______ information recorded”
  - Replace “______” with the appropriate word that represents the column

Top/bottom 10 in a column

Pick a column to explore further

- Identify the top 10 **or** bottom 10 value of your chosen column
- Save the result into a new variable

Level-up challenge - Data visualization:

- Plot the top 10 **or** bottom 10 as a bar graph with an appropriate title

After exploring the dataset…

- What questions can we ask?
- What questions cannot be answered by our dataset?
  - What are the limitations of this dataset?
  - How might who collected the dataset impact our analysis?

After exploring the dataset…

- What questions can we ask?
- What questions cannot be answered by our dataset?
  - What are the limitations of this dataset?
  - How might who collected the dataset impact our analysis?

After exploring the dataset…

- What questions can we ask?
- What questions cannot be answered by our dataset?
  - What are the limitations of this dataset?
  - How might who collected the dataset impact our analysis?

After exploring the dataset…

- What questions can we ask?
- What questions cannot be answered by our dataset?
  - What are the limitations of this dataset?
  - How might who collected the dataset impact our analysis?

After exploring the dataset…

- What questions can we ask?
- What questions cannot be answered by our dataset?
  - What are the limitations of this dataset?
  - How might who collected the dataset impact our analysis?

After exploring the dataset…

- What questions can we ask?
- What questions cannot be answered by our dataset?
  - What are the limitations of this dataset?
  - How might who collected the dataset impact our analysis?