Udacity Provided Project

If you decide not to create your own capstone project, you can use the one we've provided. For this project, we have some datasets available to you and some ideas for the project. However, it's still open-ended in nature and you'll have to define the data model and the corresponding use cases for your final deliverable.

Datasets

The following datasets are included in the project workspace. We purposely did not include a lot of detail about the data and instead point you to the sources. This is to help you get experience doing a self-guided project and researching the data yourself. If something about the data is unclear, make an assumption, document it, and move on. Feel free to enrich your project by gathering and including additional data sources.

- **I94 Immigration Data:** This data comes from the US National Tourism and Trade Office. A data dictionary is included in the workspace. This is where the data comes from. There's a sample file so you can take a look at the data in csv format before reading it all in. You do not have to use the entire dataset, just use what you need to accomplish the goal you set at the beginning of the project.
- **World Temperature Data:** This dataset came from Kaggle. You can read more about it here.
- **U.S. City Demographic Data:** This data comes from OpenSoft. You can read more about it here.
- **Airport Code Table:** This is a simple table of airport codes and corresponding cities. It comes from here.

Accessing the Data

Some of the data is already uploaded to the workspace, which you'll see in the navigation pane within Jupyter Lab. The immigration data and the global temperate data is in an attached disk.

Immigration Data

You can access the immigration data in a folder with the following path: `../../data/18-83510-I94-Data-2016/`. There's a file for each month of the year. An example file name is `i94_apr16_sub.sas7bdat`. Each file has a three-letter abbreviation for the month name. So a full file path for June would look like this: `../../data/18-83510-I94-Data-2016/i94_jun16_sub.sas7bdat`. Below is what it would look like to import this file into pandas. Note: these files are large, so you'll have to think about how to process and aggregate them efficiently.

```
fname = '../../data/18-83510-I94-Data-2016/i94_apr16_sub.sas7bdat'
df = pd.read_sas(fname, 'sas7bdat', encoding="ISO-8859-1")
```

The most important decision for modeling with this data is thinking about the level of aggregation. Do you want to aggregate by airport by month? Or by city by year? This level of aggregation will influence how you join the data with other datasets. There isn't a right answer, it all depends on what you want your final dataset to look like.

Temperature Data

You can access the temperature data in a folder with the following path: `../../data2/`. There's just one file in that folder, called `GlobalLandTemperaturesByCity.csv`. Below is how you would read the file into a pandas dataframe.

```
fname = '../../data2/GlobalLandTemperaturesByCity.csv'
df = pd.read_csv(fname)
```