**Lesson Overview**

This is the first part of a three lesson series on Python for Data Analysis. In this lesson, you will learn about the data analysis process and practice investigating different datasets with Python and Pandas. In the next two lessons, you will perform case studies where you will thoroughly analyze two datasets and learn more about NumPy, Pandas, and Matplotlib.

This lesson is aimed at helping you quickly understand what happens at each step in the data analysis process. **Though you will do a bit of coding just to build your intuition, a deeper dive of each individual topic is provided throughout the rest of the data analysis program.**

**Programming and Python**

These lessons require some familiarity with Python. Here are two awesome free courses you can take to learn or refresh your programming and Python skills!

* [**Intro to Computer Science**](https://www.udacity.com/course/intro-to-computer-science--cs101) to learn about if statements, loops, functions, lists, sets, and dictionaries.
* [**Programming Foundations with Python**](https://www.udacity.com/course/programming-foundations-with-python--ud036) to learn about classes, objects, and modules in Python.

**Other Helpful Resources**

* [**NumPy and Pandas by Udacity**](https://classroom.udacity.com/courses/ud170) to learn all about using NumPy and pandas.
* [**Python for Data Analysis**](http://www.ruxizhang.com/uploads/4/4/0/2/44023465/python_for_data_analysis.pdf) is one of the **best** resources for learning **NumPy**, **pandas**, and **Matplotlib**, as it was written by the creator of the libraries you will be learning about, Wes McKinney. This is an exhaustive guide to using these libraries. I, personally, own two copies: one for work and one for home.

You will be putting these skills to use throughout the rest of this nanodegree program.

**Applications of Data Analysis**

* [**Facebook's blog post**](https://research.fb.com/exposure-to-diverse-information-on-facebook-2/) and [**paper**](https://research.fb.com/publications/exposure-to-ideologically-diverse-information-on-facebook/) on exposure to ideologically diverse information
* [**OkCupid's blog post**](http://blog.okcupid.com/index.php/the-best-questions-for-first-dates/) on the best questions to ask on a first date
* [**Article**](http://www.dezyre.com/article/how-big-data-analysis-helped-increase-walmart-s-sales-turnover/109) on how Walmart used big data analysis to increase sales
* [**Wikipedia page**](https://en.wikipedia.org/wiki/Bill_James) on how Bill James applied data analysis to baseball
* [**Numerate's post**](http://www.numerate.com/numerates-ranking-technology-pharmaceutical-rd-gains-u-s-patent/) on using data analysis to design pharmaceutical drugs

**Anaconda and Jupyter Notebooks**

If you haven't already done so, go through our short course on [**Anaconda and Jupyter notebooks**](https://classroom.udacity.com/courses/ud1111) to set up your computer. You will learn how to set up an environment for this course, which requires Python 3.6, Jupyter Notebook, NumPy, pandas, Matplotlib, and Seaborn.

We organized the data analysis process into five steps: Question, Wrangle, Explore, Draw Conclusions, and Communicate. Below is a review of the key points, but feel free to move ahead. We will practice each step as we go through the next sections, and you will get the whole process down in no time.

**Step 1: Ask questions**

Either you're given data and ask questions based on it, or you ask questions first and gather data based on that later. In both cases, great questions help you focus on relevant parts of your data and direct your analysis towards meaningful insights.

**Step 2: Wrangle data**

You get the data you need in a form you can work with in three steps: gather, assess, clean. You gather the data you need to answer your questions, assess your data to identify any problems in your data’s quality or structure, and clean your data by modifying, replacing, or removing data to ensure that your dataset is of the highest quality and as well-structured as possible.

**Step 3: Perform EDA (Exploratory Data Analysis)**

You explore and then augment your data to maximize the potential of your analyses, visualizations, and models. Exploring involves finding patterns in your data, visualizing relationships in your data, and building intuition about what you’re working with. After exploring, you can do things like remove outliers and create better features from your data, also known as feature engineering.

**Step 4: Draw conclusions (or even make predictions)**

This step is typically approached with machine learning or inferential statistics that are beyond the scope of this course, which will focus on drawing conclusions with descriptive statistics.

More on machine learning: [**Machine Learning Engineer Nanodegree**](https://www.udacity.com/course/machine-learning-engineer-nanodegree--nd009)

**Step 5: Communicate your results**

You often need to justify and convey meaning in the insights you’ve found. Or, if your end goal is to build a system, you usually need to share what you’ve built, explain how you reached design decisions, and report how well it performs. There are many ways to communicate your results: reports, slide decks, blog posts, emails, presentations, or even conversations. Data visualization will always be very valuable.

Before walking through each of these steps with real datasets using Python, let's build a bit of intuition for each step!

One of the biggest reasons that Python is considered the standard language for data science is its powerful packages. [**NumPy**](http://www.numpy.org/), [**pandas**](https://pandas.pydata.org/), and [**Matplotlib**](https://matplotlib.org/) are three core packages for data analysis that we'll be learning to use in this course.

Note: This course aims to introduce these packages with a learning by doing approach as we go through the data analysis process. If you'd like full lessons on NumPy and pandas, you can find them in lessons 2 and 3 of this [**free course**](https://classroom.udacity.com/courses/ud170). You can skip the first lesson, which analyzes data with pure Python, similar to what you did in the previous project.

Let's ask questions about [**tumor data**](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29) from the University of Wisconsin's general surgery lab.

Gathering Data

These are links to useful text editors referenced here.

* [**Atom**](https://atom.io/)
* [**Sublime**](https://www.sublimetext.com/).
* **Selecting Multiple Ranges in Pandas**
* Selecting the columns for the mean dataframe was pretty straightforward - the columns we needed to select were all together (id, diagnosis, and the mean columns). Now we run into a little issue when we try to do the same for the standard errors or maximum values. id and diagnosis are separated from the rest of the columns we need! We can't specify all of these in one range.
* First, try creating the standard error dataframe on your own to see why doing this with just locand iloc is an issue. Then, use this [stackoverflow link](https://stackoverflow.com/questions/41256648/select-multiple-ranges-of-columns-in-pandas-dataframe" \t "_blank) to learn how to select multiple ranges in Pandas and try it below. By the way, to figure this out myself, I just found this link by googling "how to select multiple ranges df.iloc"
* *Hint: You may have to import a new package!*

Using the functions you learned so far, explore census\_income\_data.csv in the Jupyter notebook below to answer quiz questions below the notebook about these characteristics of the dataset:

* number of rows
* number of columns
* data types of several columns
* columns with missing values
* number of unique values of education
* mean age
* 75th percentile of hours-per-week

This data was originally taken from [**here**](https://archive.ics.uci.edu/ml/datasets/Census+Income).