# I'VE GOT THE DATA, NOW WHAT?

## EXPLORATORY DATA ANALYSIS WITH PYTHON AND PANDAS

Presented by Ashley Steele

Women Who Code Python – 12/11/2019

https://tinyurl.com/wwceda

# OVERVIEW

I. What is a Data Scientist and what do we do?
II. What is EDA and Why Do We Use it?
III. Introduction to Our Dataset and the Tools We Will Use
IV. Welcome to the Checklist!
V. Checklist Breakdown and Playtime
VI. Question and Answer Time/Further Resources

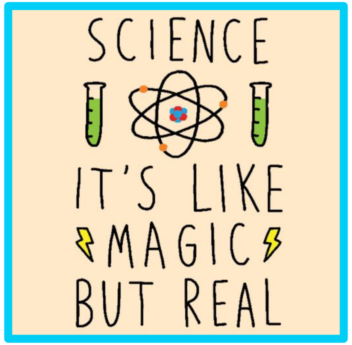
# HELLO!

I’m Ashley Steele!

I am here because I love data and your should too.

You can find me on Twitter at @AshleyCSteele !

# What is a Data Scientist and what do we do?



# WHAT IS A DATA SCIENTIST?

Deep Learning

Artificial Intelligence

Natural Language Processing

Machine Learning

Data Storytelling

Big Data

Time Series Analysis

Analysis and Statistics

# WHAT DO WE REALLY DO?

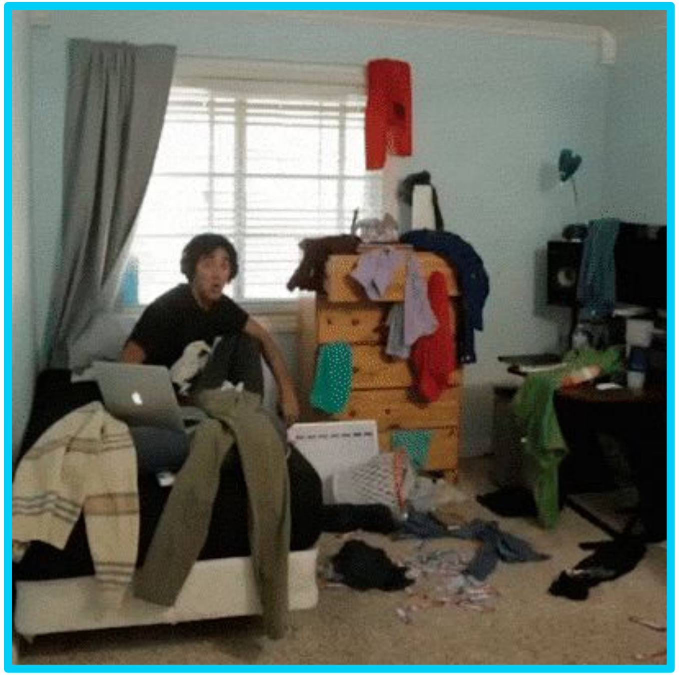
Ugh! I have so much data to clean !

- ★ Data scientists spend 80% of their time cleaning data and 20% applying/modeling their data (source)

# What is EDA and Why Do We Use it?

# WHAT IS EDA?

Exploratory Data Analysis (EDA) is the process of loading, cleaning, and analyzing data.



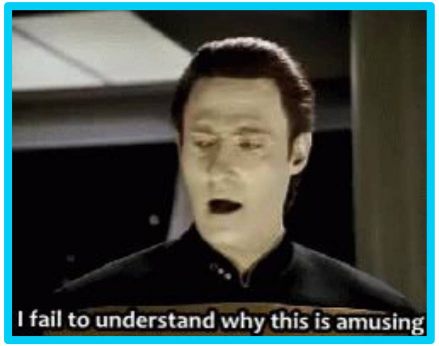
# WHY DO WE USE IT?

- Data without cleaning and organizing is useless.

- We need to know what story our data are telling us.

- Crappy data = crappy models

# Intro to Our Dataset and Tools!



# The Hero of Our Story: The Dataset!

We're going to use the Titanic training data set from here!

- **What is this dataset about?**

Original 1912 data about passengers on the Titanic. (We are looking at a portion only!)

- **Why is it used?**

One of the most common uses is to create a predictive model to test for survival.

Warning!!

Raw data doesn't typically look like this!

# WHAT TOOLS DO WE NEED TO GET THE JOB DONE?

## Our Stars of the Show:

- **Python**- our programing language
- **Pandas-** a Python library made for data science and analysis
- **Jupyter Notebook**- where we will write and run our code

## Supporting Actors:

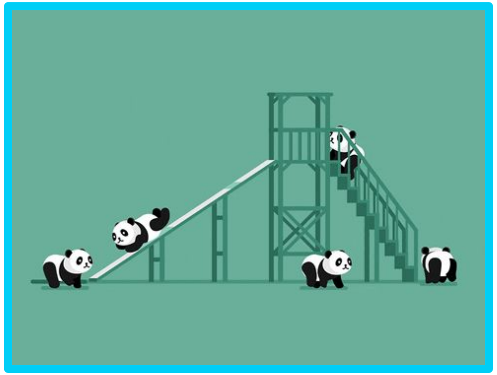
- **Numpy-** a library that lets us do high-level math on multi-dimensional arrays
- **Matplotlib.pyplot-**basis for all plotting in Python
- **Seaborn-** creates BEAUTIFUL plots and visualizations
- **OS**- let's us control our working directory

# Welcome to the Checklist!

# THE CHECKLIST

1. Import libraries (pandas, numpy, matplotlib.pyplot, seaborn, os, etc.)
2. Set up your working directory using os (if you already know you want to work in another folder)
3. Load your data in!
4. Glance at your data and take a quick introductory peek (Things to look at: head(), describe(), info(), columns, dtypes, missing/null values, etc.)
5. Fix any missing values (my favorite way is df.isnull().sum().sort_values(ascending = False)
6. Convert data types (this is especially important if you are using "weird" dtypes, like datetime
7. Feature distribution (A.K.A. What do my features look like, individually?) (Things to do: QQ plots, histograms, look for bias, look for skew in data)
8. Normalize data/outlier analysis
9. Feature engineering and selection (Things to pay attention to: collinearity, multicollinearity, Omitted Variable Bias (OVB))
10. Bivariate Analysis!
11. Relationships (scatterplots, correlations, matrices, etc.)

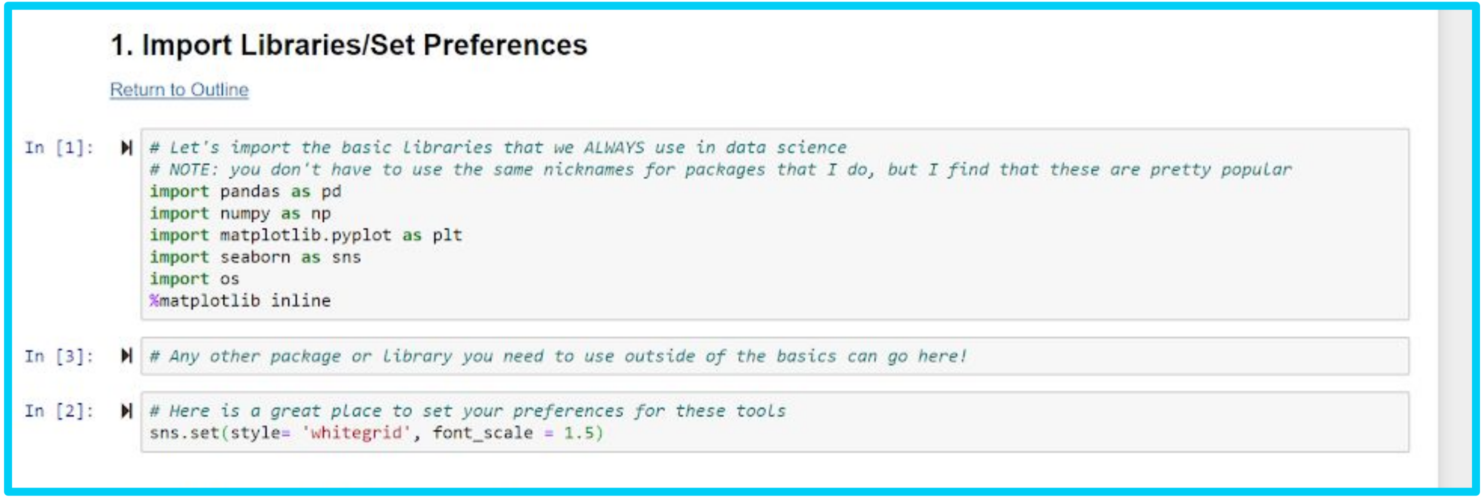
# Checklist Breakdown and Playtime!



# WAIT... SHE SAID PLAY BUT THIS IS A PRESENTATION!!

- Our time is limited so I'm going to do a LOT of showing
- Want to have hands-on experience or follow along later?
- All of the resources for this presentation are here, including my Jupyter Notebook: https://github.com/SteeleAlloy/edaworkshop

# STEP 1: IMPORT YOUR LIBRARIES/SET PREFERENCES



## 1. Import Libraries/Set Preferences

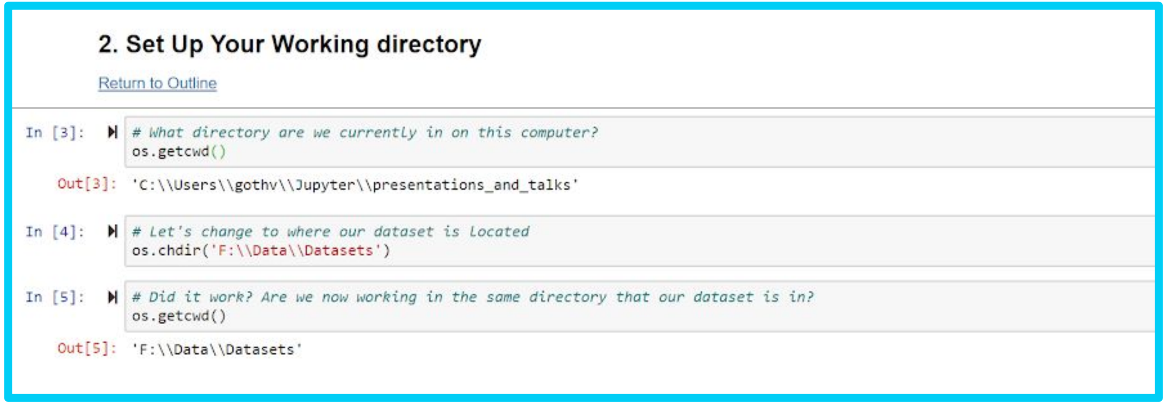
Return to Outline

In [1]:
```
# Let's import the basic libraries that we ALWAYS use in data science
# NOTE: you don't have to use the same nicknames for packages that I do, but I find that these are pretty popular
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import os
%matplotlib inline
```

In [3]:
```
# Any other package or library you need to use outside of the basics can go here!
```

In [2]:
```
# Here is a great place to set your preferences for these tools
sns.set(style= 'whitegrid', font_scale = 1.5)
```

# STEP 2: SET WORKING DIRECTORY

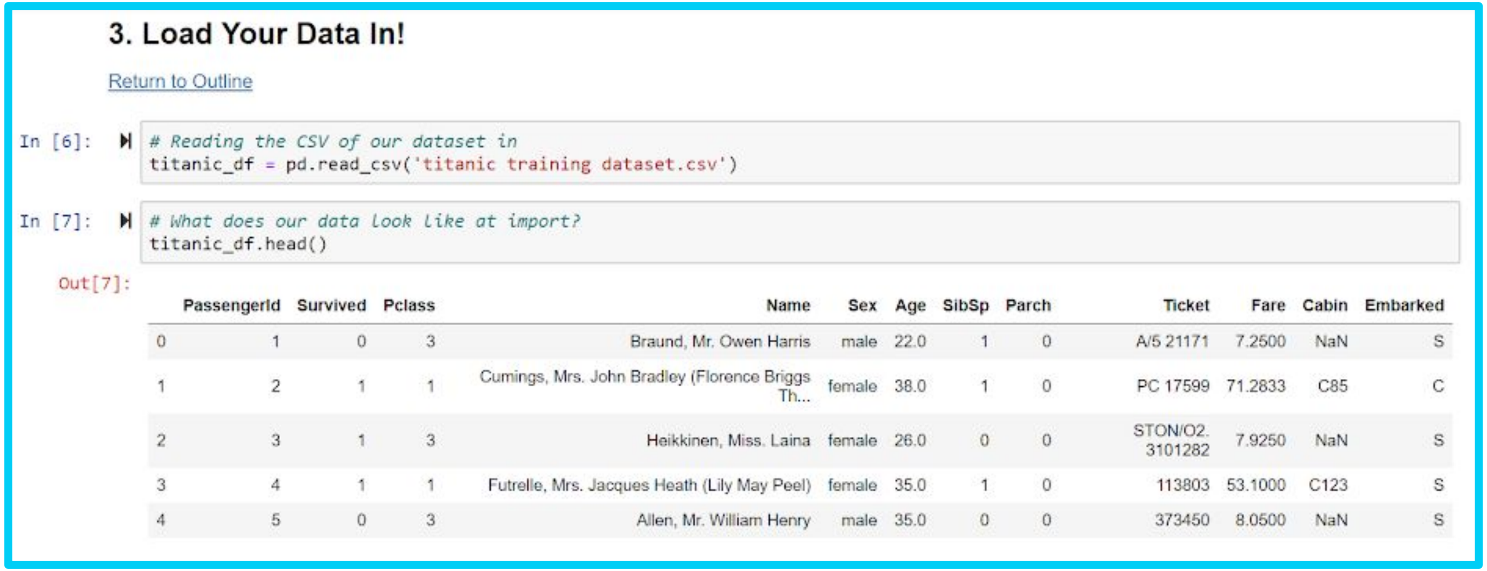


## 2. Set Up Your Working directory

Return to Outline

```
In [3]:  # What directory are we currently in on this computer?
         os.getcwd()

Out[3]:  'C:\\Users\\gothv\\Jupyter\\presentations_and_talks'

In [4]:  # Let's change to where our dataset is located
         os.chdir('F:\\Data\\Datasets')

In [5]:  # Did it work? Are we now working in the same directory that our dataset is in?
         os.getcwd()

Out[5]:  'F:\\Data\\Datasets'
```

# STEP 3: LOAD YOUR DATA



### 3. Load Your Data In!

Return to Outline

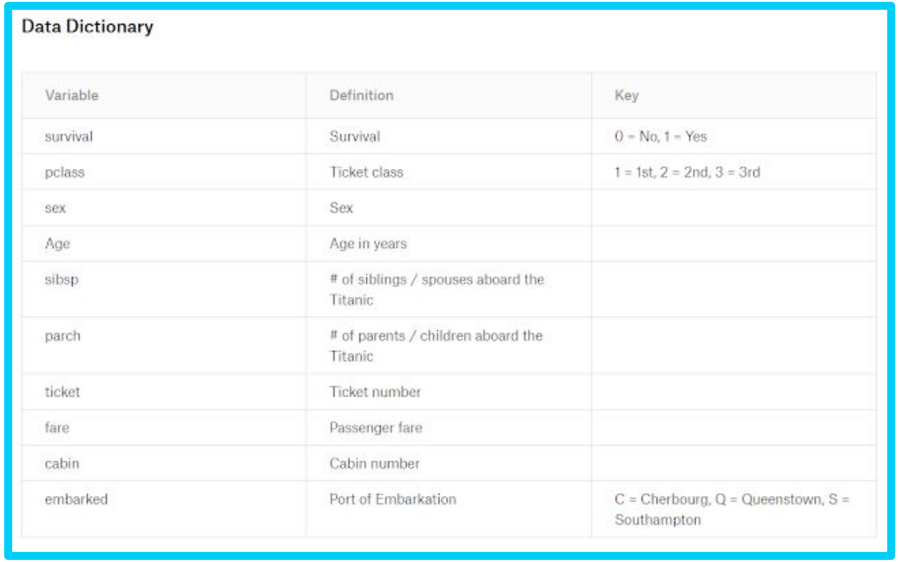
```
In [6]:  # Reading the CSV of our dataset in
         titanic_df = pd.read_csv('titanic training dataset.csv')
```

```
In [7]:  # What does our data look like at import?
         titanic_df.head()
```

Out[7]:

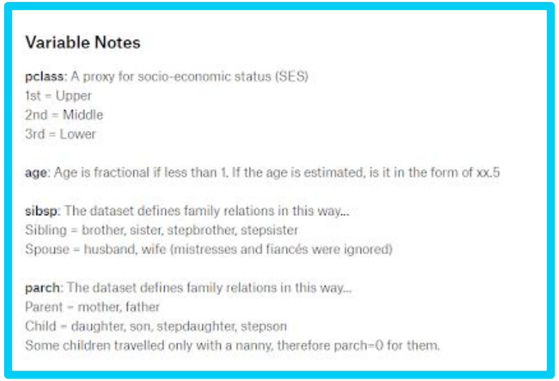
| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 0 | 3 | Braund, Mr. Owen Harris | male | 22.0 | 1 | 0 | A/5 21171 | 7.2500 | NaN | S |
| 1 | 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Th... | female | 38.0 | 1 | 0 | PC 17599 | 71.2833 | C85 | C |
| 2 | 3 | 1 | 3 | Heikkinen, Miss. Laina | female | 26.0 | 0 | 0 | STON/O2. 3101282 | 7.9250 | NaN | S |
| 3 | 4 | 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | 35.0 | 1 | 0 | 113803 | 53.1000 | C123 | S |
| 4 | 5 | 0 | 3 | Allen, Mr. William Henry | male | 35.0 | 0 | 0 | 373450 | 8.0500 | NaN | S |

# STEP 3B: LOOK AT YOUR DATA DICTIONARY



### Data Dictionary

| Variable | Definition | Key |
|---|---|---|
| survival | Survival | 0 = No, 1 = Yes |
| pclass | Ticket class | 1 = 1st, 2 = 2nd, 3 = 3rd |
| sex | Sex | |
| Age | Age in years | |
| sibsp | # of siblings / spouses aboard the Titanic | |
| parch | # of parents / children aboard the Titanic | |
| ticket | Ticket number | |
| fare | Passenger fare | |
| cabin | Cabin number | |
| embarked | Port of Embarkation | C = Cherbourg, Q = Queenstown, S = Southampton |



### Variable Notes

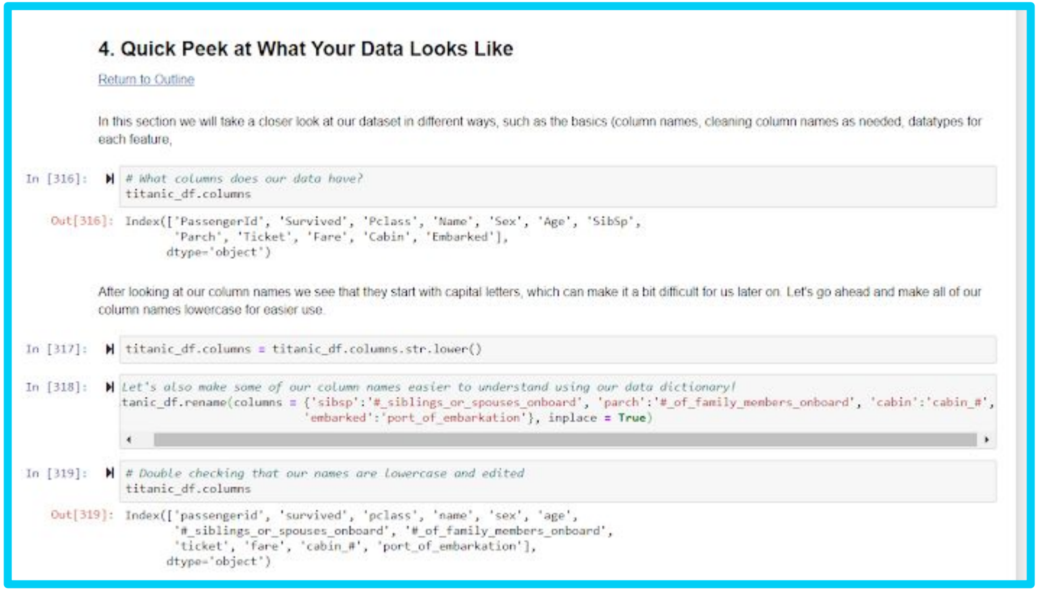
**pclass**: A proxy for socio-economic status (SES)
1st = Upper
2nd = Middle
3rd = Lower

**age**: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5

**sibsp**: The dataset defines family relations in this way...
Sibling = brother, sister, stepbrother, stepsister
Spouse = husband, wife (mistresses and fiancés were ignored)

**parch**: The dataset defines family relations in this way...
Parent = mother, father
Child = daughter, son, stepdaughter, stepson
Some children travelled only with a nanny, therefore parch=0 for them.

# STEP 4: INTRO PEEK AT YOUR DATA



## 4. Quick Peek at What Your Data Looks Like

Return to Outline

In this section we will take a closer look at our dataset in different ways, such as the basics (column names, cleaning column names as needed, datatypes for each feature,

```
In [316]:  # What columns does our data have?
           titanic_df.columns

Out[316]:  Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
                  'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],
                 dtype='object')
```

After looking at our column names we see that they start with capital letters, which can make it a bit difficult for us later on. Let's go ahead and make all of our column names lowercase for easier use.

```
In [317]:  titanic_df.columns = titanic_df.columns.str.lower()

In [318]:  Let's also make some of our column names easier to understand using our data dictionary!
           tanic_df.rename(columns = {'sibsp':'#_siblings_or_spouses_onboard', 'parch':'#_of_family_members_onboard', 'cabin':'cabin_#',
                                      'embarked':'port_of_embarkation'}, inplace = True)

In [319]:  # Double checking that our names are lowercase and edited
           titanic_df.columns

Out[319]:  Index(['passengerid', 'survived', 'pclass', 'name', 'sex', 'age',
                  '#_siblings_or_spouses_onboard', '#_of_family_members_onboard',
                  'ticket', 'fare', 'cabin_#', 'port_of_embarkation'],
                 dtype='object')
```

# Step 5 - 11: Let's just look at our code!

# ANNNNDDDDD.......?

# Question and Answer Time/Additional Resources!

# LET'S CONNECT!

**Twitter**
@AshleyCSteele

**Email**
Scientistofsteele
@gmail.com