# Welcome New Analysts!

First, Decide what you want to investigate or analyze. What problem do we want to solve?



* How can we translate data into dollars?
* What impact do I want to make with this data?
* What business value does our model bring to the table?
* What will save us lots of money?
* What can be done to make our business run more efficiently?

# Set up your environment

* Install Anaconda
  + Set up Jupyter Notebook
  + Get familiar with how it works, such as the import statements are placed in only one cell near the top
    - import pandas as pd
    - import numpy as np
  + Get familiar with how it works such as stop using print() everywhere. The output happens when you run the cell
  + Get familiar with markdown cells so you have notes, comments, conclusions
    - Always comment your code!
* You will be using Python and the Pandas Library

# Collect your Data

* Databases
  + SQL, PostgreSQL
* Log files
* API’s (Application Programming Interface) for web services
  + Collect with the *requests* package
* Web scraping from a websites HTML
  + Beautiful Soup
* Internet resources that provide data for download
  + Kaggle’s [Titanic: Machine Learning from Disaster](https://www.kaggle.com/c/titanic) challenge



# Data Wrangling

Be prepared for some issues you may encounter: human error, computer error, unexpected values, incomplete information, relevance of the fields, format of the data.

## Examine the data

* Look at a random sample
  + .sample()
* Number of columns/rows
  + .shape
* Column names
  + Need to remove spaces
  + Make them short and descriptive
* Data types of each column
  + .info()

## Summary Statistics with .describe()

* Get summary statistics for the numerical columns
  + example\_data.describe(include=[np.number]).T
* Get summary statistics for the object (string) columns
  + example\_data.describe(include=[np.object]).T
* Transpose the dataframe for more readable output
  + .T will improve readability for dataframes with many columns as it allows more data to fit on the screen without scrolling

## Change Data Types

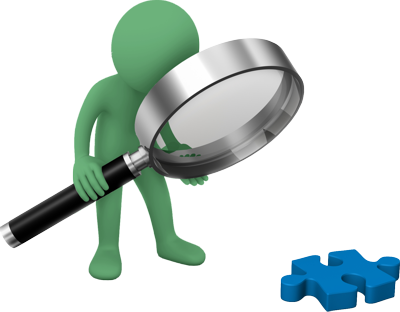
* Reduce memory by changing data types
  + ‘object’ data type takes the most memory - possibly change to category
  + If 1 or 0, convert to int8
* Change DateTime as needed
  + parse\_dates parameter from read\_csv
  + pd.to\_datetime after loading data

# Exploratory Data Analysis (EDA)

* Check for missing
* Check for duplicates
* My personal EDA list:
* data.head()
* data.tail() #bottom is more likely corrupt.
* data.sample(10) # 10 random rows
* data.shape #-or- len(data) #length of columns, same as number of rows
* data.info() # easy to see missing values
* data.describe() #shows 0.00000 - 25% is 0.0000
* #this specifically looks for zero - many of the games had zero sales in a region
  + for i in data.columns:
    - print(i, len(data[data[i] == 0]))
* data.describe(include=['object']) #shows categorical
* data.duplicated().sum()
* data.drop\_duplicates()
* data.[‘Platform’].value\_counts()
* data.[‘Rating’].unique()
* data.[‘Rating’].isna().sum()

# Draw Conclusions at each step

* Does the data help us answer the questions we have or give insight into the problem we are investigating?
* Do we need to collect new or additional data?
* Did we notice any patterns or relationships within the data?

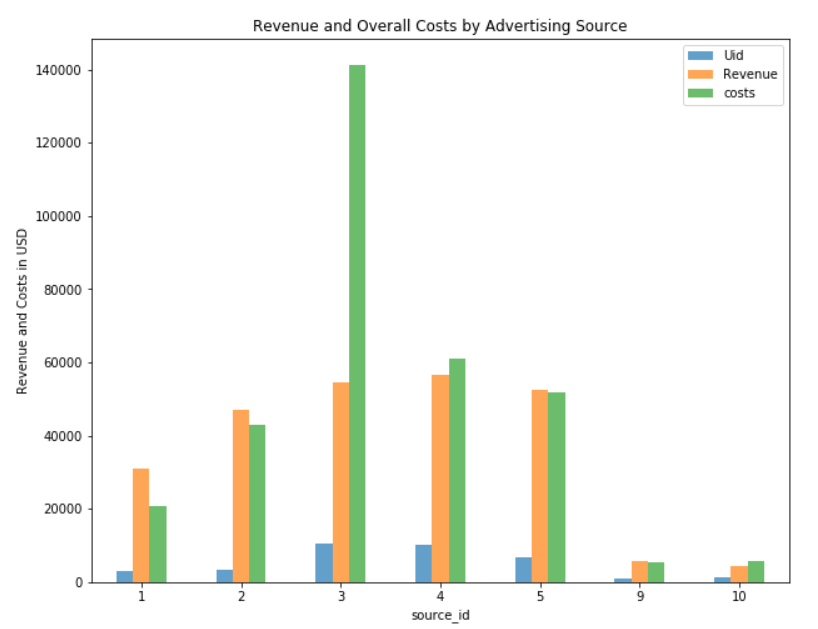


# Analyze the data

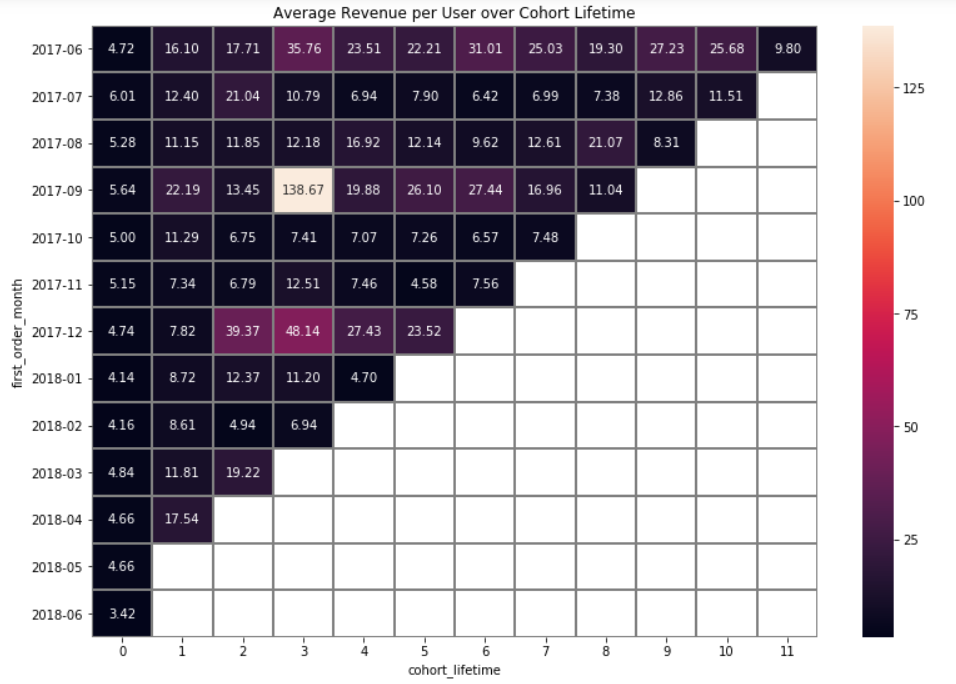
* Plot histograms
* Describe the distributions
* Group and reshape data
  + Merge
  + Pivot
* Hypothesis Testing
* A/B Testing
* Get familiar with statistical methods
  + z-score
  + t-test
  + p-value
  + mann-whitney u

# Visualize

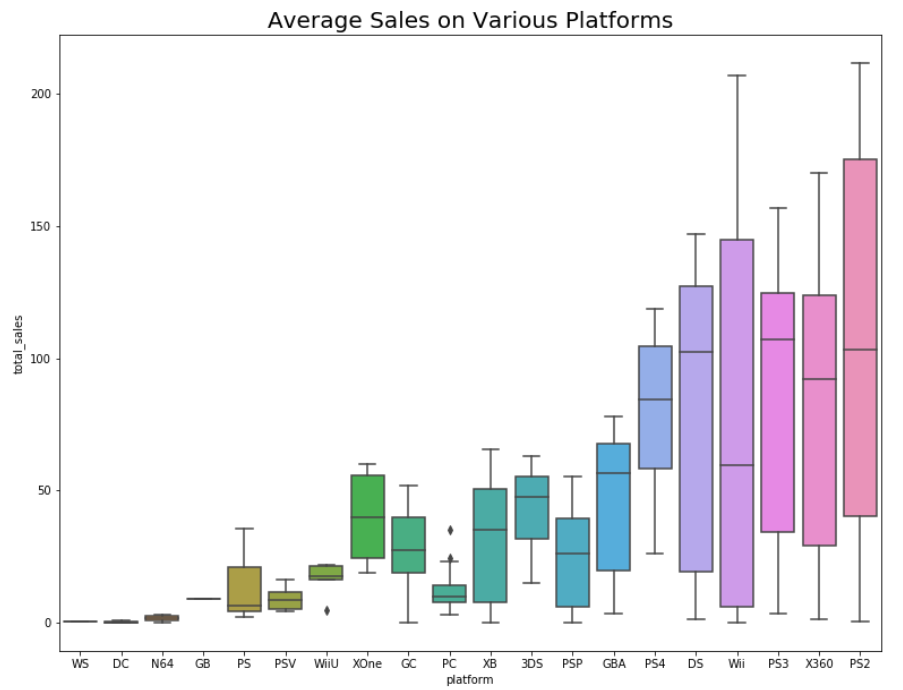
* Always label your graphs!
  + X axis
  + Y axis
  + Title
  + Legend if it helps to clarify
* Choose an appropriate type of graph or chart
  + Check out the gallery for ideas
  + Check the documentation for help with labeling
* Matplotlib
  + https://matplotlib.org/gallery/index.html
  + Bar chart



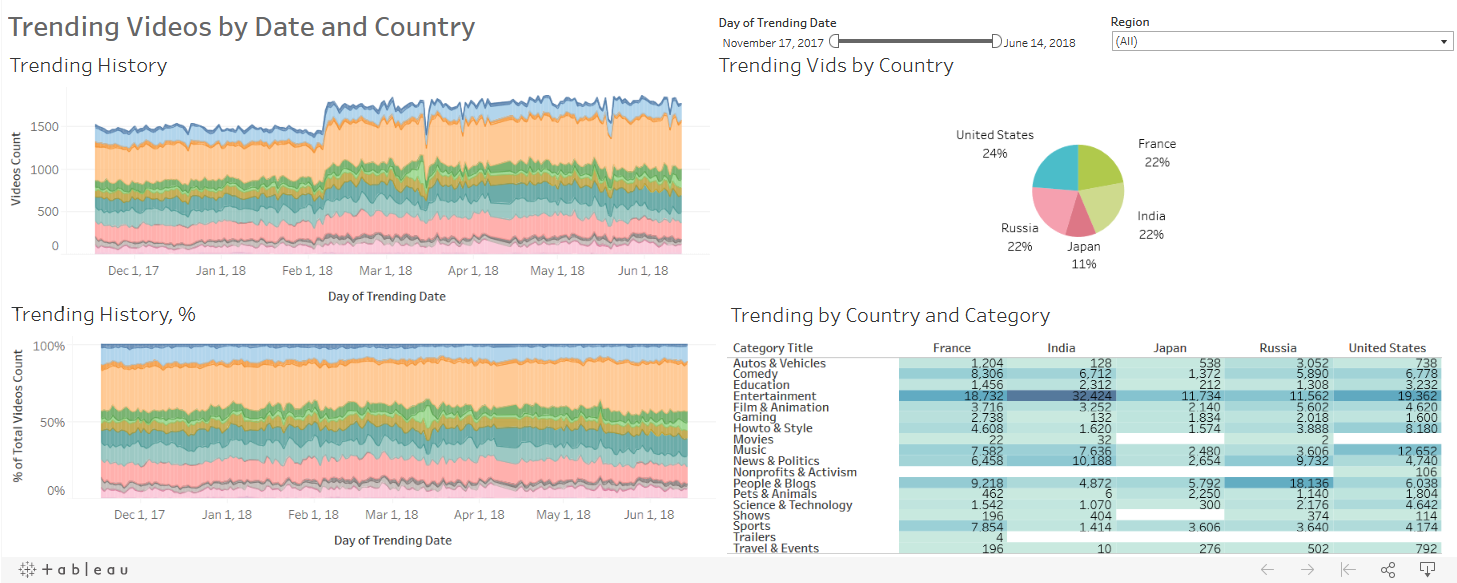
* Seaborn
  + https://seaborn.pydata.org/examples/index.html
  + Heatmap

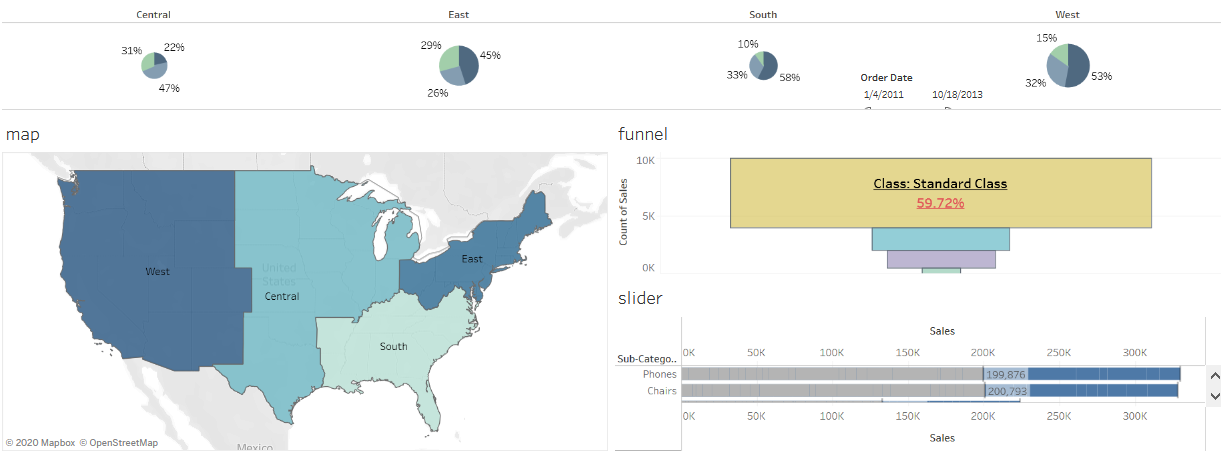


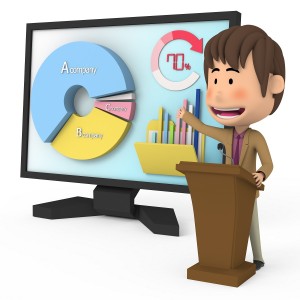
* + Boxplot



* Tableau
  + Interactive Dashboards







# Presentation

“If you can’t explain it to a six year old, you don’t understand it yourself.” — Albert Einstein

## Create a slide deck in Powerpoint, save as a pdf to distribute.

* **Identify business insights:** return back to business problem
* **Visualize your findings accordingly:** keep it simple and priority driven
* **Tell a clear and actionable story:** effectively communicate to non-technical audience