DATA SCIENCE: CAREER OF THE FUTURE

INTRODUCTION TO DATA SCIENCE

SANJAY RAJVANSHI

SCHEDULE

Session	Date	Time	Topic
1	Sep 25	7:00 pm – 8:00 pm	Introduction to data science and associated tools.
2	Oct 2	7:00 pm – 8:00 pm	Introduction to Python. Learn how to use Python for data analysis. Python is simple, yet powerful language that is often used in data science.
3	Oct 9	7:00 pm – 8:00 pm	Data wrangling with Python. Learn how to gather data and make it useful for analysis.
4	Oct 16	7:00 pm – 8:00 pm	Data visualization and analysis with Python. Learn how to create useful visualizations to aid in the analysis of the data.
5	Oct 23	7:00 pm – 8:00 pm	Brief introduction to artificial intelligence and machine learning. Get a peek into how to make data based predictions.

Note: All classes are on Wednesdays.

SESSION 3 – RECAP

- Identification of relevant dataset(s)
- Data wrangling or munging or management
- pandas library
- DataFrame data structure
  - Viewing, <TAB> completion
  - Column labels, counts, data types, size,
  - Selection, Addition, Deletion
  - Arithmetic and logical operations at cell, row, column, DataFrame levels
  - Statistical, Transpose, Sorting, Boolean Indexing, Setting, Missing values
  - Append, Grouping, Selective assignment

SESSION 4: DATA VISUALIZATION AND ANALYSIS WITH PYTHON

SESSION 4 – AGENDA

- Data visualization and analysis with Python.
- Create useful visualizations to aid in the analysis of the data.
- Create and customize various types of graphs
- Learn some statistical techniques

DATA SCIENCE SOLUTION LIFECYCLE

- Data Science solution lifecycle (iterative):
  - Problem identification
  - Identify data
  - Clean, transform data
  - **Analyze, visualize**
  - Identify algorithm(s)
  - Implement
  - Maintain and support

PROBLEM IDENTIFICATION

- For this introductory class, we will work on a simple problem.
- Of course, a problem becomes even simpler if the data is readily available.
- Problem: Find out the number of students attending Montgomery College by campuses.

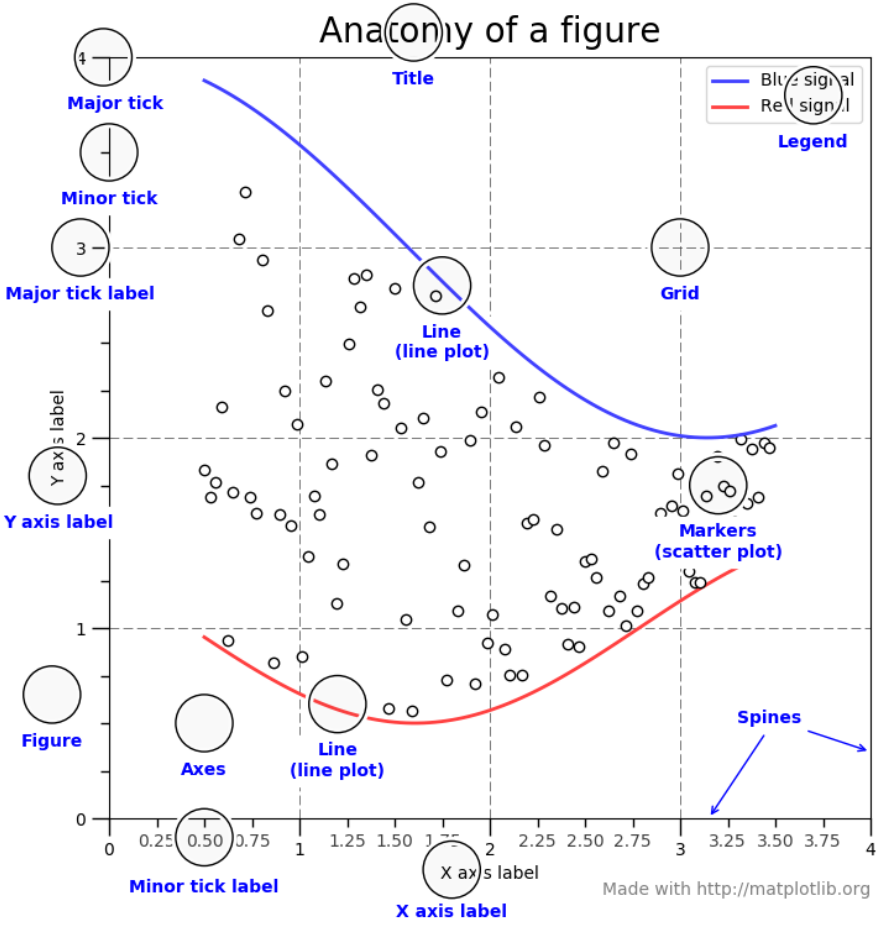
IDENTIFY DATA

- Montgomery College enrollment data is published by Montgomery County, MD and made available via its Open Data Portal website.
- Download *Montgomery College Enrollment Data* from https://data.montgomerycountymd.gov/Education/Montgomery-College-Enrollment-Data/wmr2-6hn6

DATA WRANGLING

- Making data suitable for analysis
  - Cleaning data
    - Missing values
  - Transforming data
    - String to numbers or vice versa
    - Conversion of coded values
  - Handling outliers
    - Values that are exceptionally out of place
  - Normalize data
    - Technique to adjust the spread of data
- Do pretty much any type of data management that increases the data suitability for the analysis

ANATOMY OF A FIGURE



- https://matplotlib.org/tutorials/introductory/usage.html#sphx-glr-tutorials-introductory-usage-py

PLOT LIBRARIES

- matplotlib, pyplot
  - matplotlib – Python 2D plotting library for publication quality figures
  - pyplot – Module that makes matplotlib MATLAB like
  - matplotlib is the whole package and pyplot is a module in it
  - Types of plots – generic, bar, pie, scatter, histogram, boxplot, contours, polygons, polar, stackplot, stem, step, violin plot
  - Reference
    - https://matplotlib.org/index.html
    - https://matplotlib.org/contents.html
    - https://matplotlib.org/api/_as_gen/matplotlib.pyplot.html#module-matplotlib.pyplot
    - https://matplotlib.org/tutorials/introductory/pyplot.html
    - https://matplotlib.org/resources/index.html#tutorials

PLOT LIBRARIES

- seaborn
  - A library for making statistical graphics in Python. It is built on top of matplotlib and closely integrated with pandas data structures.
  - Types of plots – relational, categorical, distribution, regression, matrix,
  - References
    - https://seaborn.pydata.org/introduction.html#introduction
    - https://seaborn.pydata.org/index.html
    - https://seaborn.pydata.org/tutorial.html
    - https://seaborn.pydata.org/tutorial/categorical.html#categorical-tutorial
    - https://seaborn.pydata.org/api.html

ANALYZE, VISUALIZE

- Continue with our Session 3 dataset (*Montgomery College Enrollment Data* from https://data.montgomerycountymd.gov/Education/Montgomery-College-Enrollment-Data/wmr2-6hn6
- We will also use couple of built-in datasets

CLASS EXERCISE-1

- Create a Python file with name "S4-Exx1"
- We will cover all the topics in previous slides in this exercise working directly in the Jupyter notebook

CLASS EXERCISE-2

- Create a Python file with name "S4-Exx2"
- Work with the Montgomery college enrollment dataset
  - Compute the number of students in one campus in each of the "HS Category" as well as "MC Program Description"
  - Suppose you are in the college administration and you need to decide about what staff you need to have for evening classes. What kind of analysis would you do? What data elements you would look at? Identify the applicable data elements and come up with a solution for this problem.
  - Suppose you are in the college administration and you are debating whether the three campus locations are adequate or there is a need for another campus(es). What kind of analysis would you do? What data elements you would look at? Identify the applicable data elements and come up with a solution for this problem. Discuss with everyone on your table.

CLASS EXERCISE-3

- Create a Python file with name "S4-Exx3"
- Work with the Boston house prices dataset
  - Explore relationship (scatter plots) between the price and crime rate (CRIM), price and nitric oxide (NOX) concentration (with appropriate labels)
  - Draw the heatmap (larger sized)
  - Based on the correlation between the variables, draw 2 good graphs and share with others on your table.

SESSION 4 – HOME WORK

- Pyplot tutorial (https://matplotlib.org/3.1.1/tutorials/introductory/pyplot.html)
  - Try all plot types and various configuration settings
- An Introduction to seaborn (https://seaborn.pydata.org/introduction.html)
  - Try all plot types and various configuration settings
- Identify some datasets of interest to you and do analysis and visualization you have learned so far
  - You could download the following dataset but you have to comply and agree with the age requirements, terms and conditions of Kaggle.com and the dataset owners.
    - House Sales in King County, USA - https://www.kaggle.com/harlfoxem/housesalesprediction
- Review linear regression code in the exercise and research about it on the internet
- Review all the references mentioned in this session (and try at least a few)

SESSION 5 – AGENDA

- More data visualization and analysis with Python.
- Learn more statistical techniques as they come up.
- Brief introduction to artificial intelligence and machine learning.
- Get a peek into how to make data based predictions.

REFERENCES

Note: you are not required to sign-up for an account on any of the sites to read these articles.

1. *pyplot –*
   a. https://matplotlib.org/index.html
   b. https://matplotlib.org/contents.html
   c. https://matplotlib.org/api/pyplot_summary.html
   d. https://matplotlib.org/api/_as_gen/matplotlib.pyplot.html#module-matplotlib.pyplot
   e. https://matplotlib.org/tutorials/introductory/pyplot.html
   f. https://matplotlib.org/resources/index.html#tutorials

REFERENCES

Note: you are not required to sign-up for an account on any of the sites to read these articles.

2. *seaborn –*
    a. https://seaborn.pydata.org/introduction.html
    b. https://seaborn.pydata.org/index.html
    c. https://seaborn.pydata.org/tutorial.html
    d. https://seaborn.pydata.org/tutorial/categorical.html
    e. https://seaborn.pydata.org/api.html