DATA 101

Lecture 2: January 25, 2021



Homework Questions...Not Much To Review But...

Policy and Nits

Jupyter Notebook

Conda Prompt

conda install ...

git clone ...

Let's have a discussion about your datasets...

- What was in them? In data terms
- What might that mean? In data terms
- What problems might you be able to or NOT be able to address? Why?

What are some questions a Data Scientist answers?

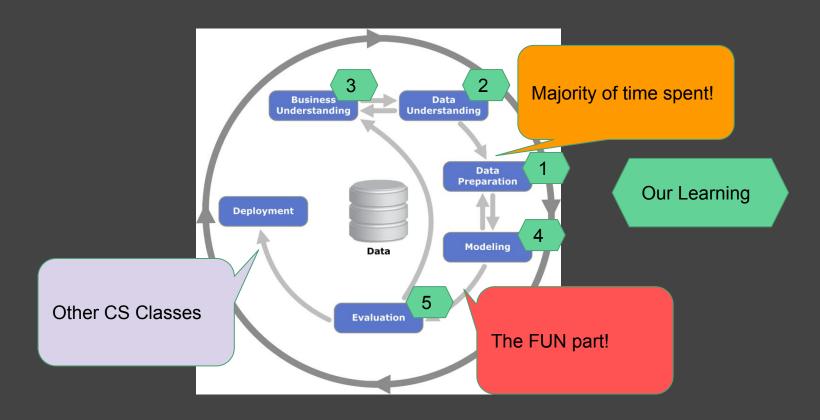
- Trend analysis
- How well do certain columns/rows compare to each other
- Use data to find a solution
- Risk
- Efficiency
- Which 'features' are important (application features)
- Which 'features' are important (data features)

Review of resources (available .ipynb)

https://github.com/wesm/pydata-book

https://github.com/jakevdp/PythonDataScienceHandbook/tree/master/notebooks

How do we do this in practice? (Also this course!)



Exploratory Data Analysis (EDA)

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	quality	1	0.44	0.099	0.054	0.0082	-0.0092	-0.098	-0.11	-0.17	-0.19	-0.21	-0.31	0.
count	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	alcohol	0.44	1	0.12	-0.017	-0.25	-0.076	-0.45	-0.12	-0.45	0.068	-0.36	-0.78	
mean	6.854788	0.278241	0.334192	6.391415	0.045772	35.308085	рН	0.099	0.12	1	0.16	-0.00062	-0.16	-0.19	-0.43	0.0023	-0.032	-0.09	-0.094	0.
std	0.843868	0.100795	0.121020	5.072058	0.021848	17.007137	sulphates	0.054	-0.017	0.16	1	0.059	0.062	-0.027	-0.017	0.13	-0.036	0.017	0.074	
min	3.800000	0.080000	0.000000	0.600000	0.009000	2.000000	free sulfur dioxide	0.0082	-0.25	-0.00062	0.059	1	0.094	0.3	-0.049	0.62	-0.097	0.1	0.29	0.
25%	6.300000	0.210000	0.270000	1.700000	0.036000	23.000000	citric acid	-0.0092	-0.076	-0.16	0.062	0.094	1	0.094	0.29	0.12	-0.15	0.11	0.15	
50%	6.800000	0.260000	0.320000	5.200000	0.043000	34.000000	residual sugar	-0.098	-0.45	-0.19	-0.027	0.3	0.094	1	0.089		0.064	0.089	0.84	
75%	7.300000	0.320000	0.390000	9.900000	0.050000	46.000000	fixed acidity	-0.11	-0.12	-0.43	-0.017	-0.049	0.29	0.089	-1	0.091	-0.023	0.023	0.27	0.
max	14.200000	1.100000	1.660000	65.800000	0.346000	289.000000		20000000				0.62		0.4	0.091	0.001		0.2	0.53	
							total sulfur dioxide	0000000		0.0023		Caralletter-	0.12			1	0.089		100000000	-
							volatile acidity	-0.19	0.068	-0.032	-0.036	-0.097	-0.15	0.064	-0.023	0.089	1	0.071	0.027	
							chlorides	-0.21	-0.36	-0.09	0.017	0.1	0.11	0.089	0.023	0.2	0.071	1	0.26	-(
							density	-0.31	-0.78	-0.094	0.074	0.29	0.15	0.84	0.27	0.53	0.027	0.26	1	
		N/Le	arch Pr	roioot				quality	alcohol	Hd	sulphates	sulfur dioxide	citric acid	esidual sugar	fixed acidity	total sulfur dioxide	volatile acidity	chlorides	density	_

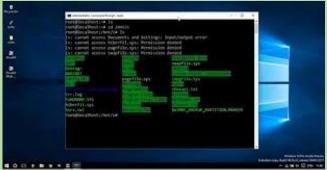
5 Minute Helpers

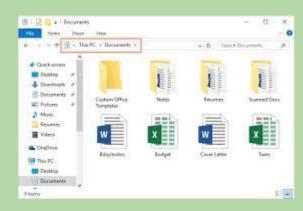
Brief Programming History

Terminals and "Spaces"

File Structure - what is behind that expensive GUI







5 More Minutes

What is data?

How is it stored typically (locally)?

Do we care how its specifically stored?

Some common formats - .csv, .xls, .json, .sql...others...?

Quick Agenda Check

Today:

Learn to code...basics

February 1st:

- 1/3 Statistics
- 1/3 Principles of Logic Statements
- ½ Exploring data in code

February 8th:

- Exploratory Data Analysis Really Begins!
- Wrangling Data
- Organizing data with our thoughts (and our code)
- Begin March Project

February 15th:

• If all goes well...we will start talking about our first Algorithm (confusion sets in)

Jupyter Notebook

PYTHON INTRODUCTION DAY