COGS9-Intro to Data Science

Spring24 - Prof. Kyle Shannon

Discussion Section A01
Week 3

Teaching Assistant (TA): Abdullah
Instructional Assistant (IA): Kyra

Discussion Sections Outline: Mostly Hands-on

- Week 2: Introductions, Making teams, Reading 1 (Part 1)
- Week 3: Reading 1 (Part 2), Python Basics with Jupyter Notebook
- Week 4: Reading 2, Getting data and wrangling it using Pandas
- Week 5: Reading 3, Assignment 1, Basics of programming for data science
- Week 6: Reading 4, Final Project Part 1 reviews/discussions
- Week 7: Assignment 2, Data Visualization and EDA demo
- Week 8: Assignment 3, Machine Learning demo
- Week 9: Reading 5, Closing thoughts
- Week 10: Final Project Part 2 reviews/discussions

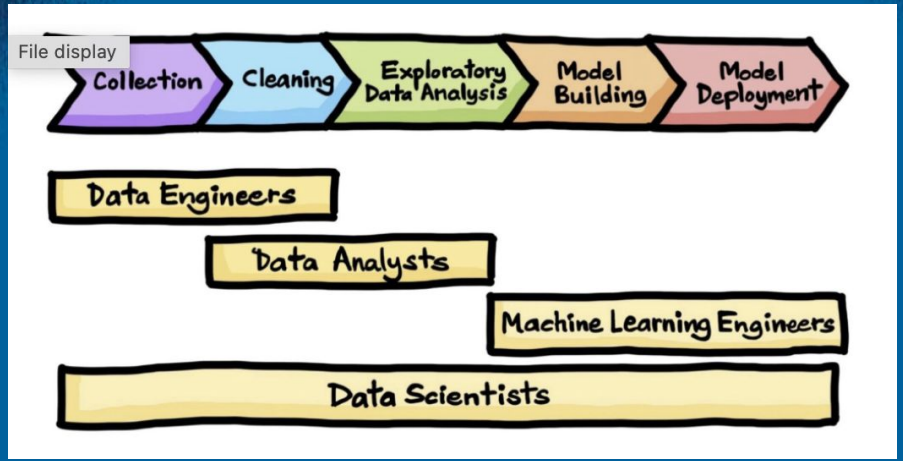
Today’s Outline

- Reading 1(Part 2)
- Basic Analysis (Excel to Python)
- Python Basics
  - Data Types (int, float, string)
  - Basic Data Structures (dictionary, list, tuple)
  - Conditional and Loops
  - Functions
  - Classes and Objects

Reading 1 (Part 2)
50 Years of Data Science

Donoho’s Six Divisions

- Data Gathering, Preparation, and Exploration
- Data Representation and Transformation
- Computing with Data
- Data Modeling
- Data Visualization and Presentation
- Science about Data Science



Data Gathering, Preparation, and Exploration

For example, a data team can gather data
- About patient demographics, medical history and drug efficacy, from clinical trials, electronic health record, and public datasets,

prepare the data,
- By data cleaning, such as removing any missing or inconsistent values,

And explore the data,
- By creating visualizations, such as histograms and scatter plots,
- To understand the distribution and identify patterns from the data.

There are missing values in the dataframe that need to be handled properly.

File display

	Employee Name	Job Title	Base Pay	Overtime Pay	Other Pay	Benefits	Total Pay
0	David xxxxx	Fire Battalion Chief	81917.0	172590.0	68870.00	21784.0	323377.0
1	Scott xxxxx	Chief Operating Officer	255000.0	NaN	31164.00	49921.0	NaN
2	Glen xxxxx	NaN	85904.0	120682.0	99408.00	26470.0	305994.0
3	David xxxxx	Fire Battalion Chief	100110.0	118798.0	62895.00	28142.0	281803.0
4	Daniel xxxxx	NaN	41389.0	196284.0	42027.00	20125.0	279700.0
5	Mark xxxxx	Retirement Administrator	240000.0	NaN	6190.00	52051.0	NaN
6	Edward xxxxx	NaN	46020.0	171896.0	59944.00	19669.0	277860.0
7	Andrea xxxxx	Independent Budget Anlyst	224099.0	NaN	13413.00	47651.0	NaN
8	Stacey xxxxx	Asst Chief Oper Ofcr	215000.0	NaN	20352.00	49139.0	NaN
9	Eric xxxxx	Fire Engineer	31869.0	149615.0	61107.00	32243.0	242591.0

Data Representation and Transformation

- after exploring the data, the team would represent and transform the data in a way that is suitable for analysis and modeling
- This could include feature engineering, normalization, and dimensionality reduction

Computing with Data

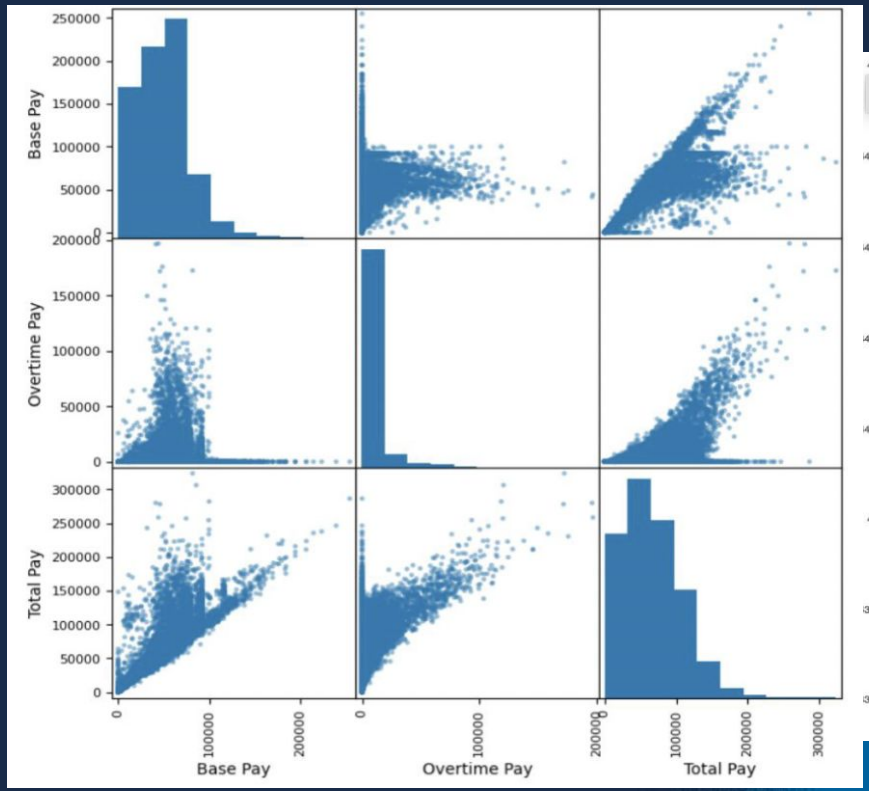
- Involves using computational techniques to analyze the data, such as statistical inference, machine learning, and data mining

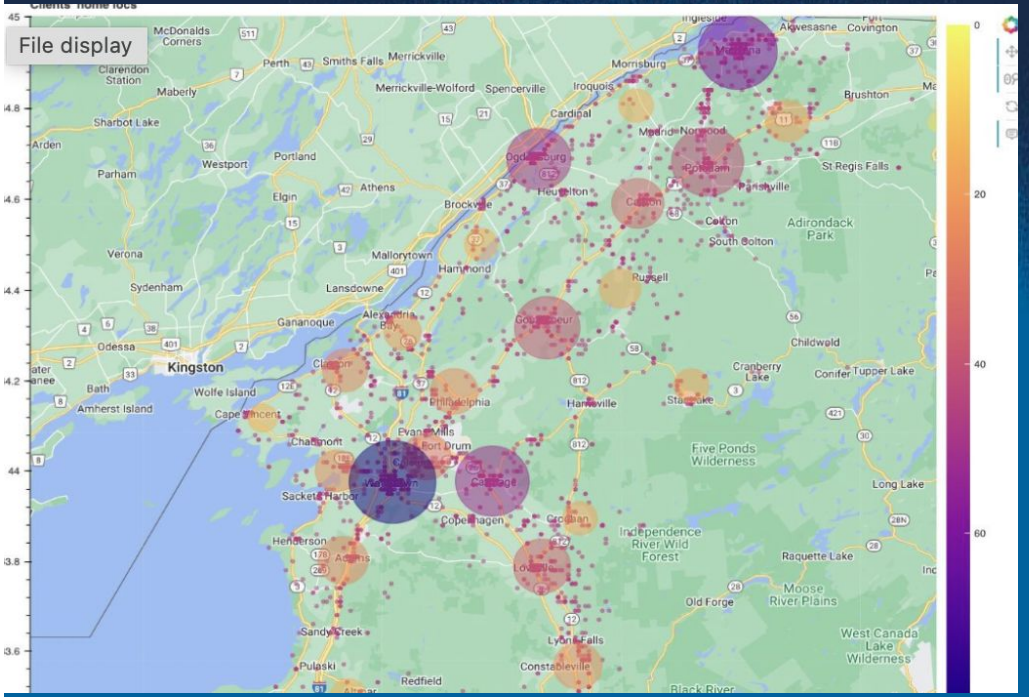
- These can include popular languages such as R and Python, and many more

Data Visualization and Presentation

- The data team would create visual representations of the data, such as heatmaps and bar charts, to make it easier to understand and interpret the data
- For example, they could create interactive dashboards that allow the medical team to explore the data and gain insights, and also prepare the results of the project in a way that is easy to understand and present to stakeholders

Examples of Data Visualization



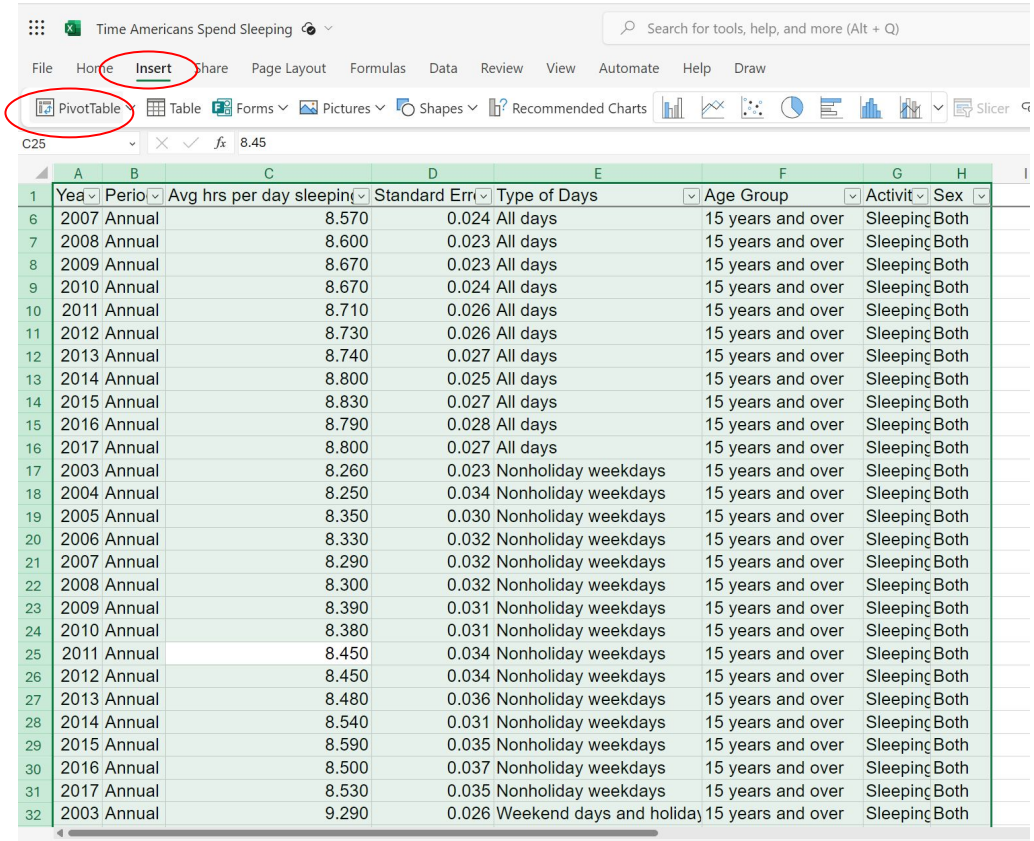


Science about Data Science

- Tukey proposed that “a science of data analysis” exists and should be recognized as among the most complicated of all sciences”
- It involves monitoring the performance of the model, validating the findings, and understanding the ethical and legal implications of the results.
- Additionally, it involves staying current with the latest developments and trends in data science and being able to reflect on the processes and methods used throughout the project

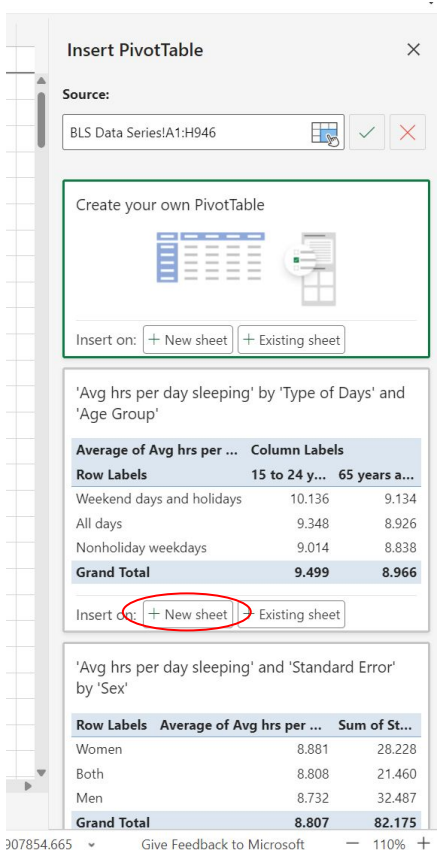
Excel

Step 1:
Select All
Go to Insert -> Pivot Table



	A	B	C	D	E	F	G	H
1	Yea	Perio	Avg hrs per day sleepin	Standard Err	Type of Days	Age Group	Activit	Sex
6	2007	Annual	8.570	0.024	All days	15 years and over	Sleepinc	Both
7	2008	Annual	8.600	0.023	All days	15 years and over	Sleepinc	Both
8	2009	Annual	8.670	0.023	All days	15 years and over	Sleepinc	Both
9	2010	Annual	8.670	0.024	All days	15 years and over	Sleepinc	Both
10	2011	Annual	8.710	0.026	All days	15 years and over	Sleepinc	Both
11	2012	Annual	8.730	0.026	All days	15 years and over	Sleepinc	Both
12	2013	Annual	8.740	0.027	All days	15 years and over	Sleepinc	Both
13	2014	Annual	8.800	0.025	All days	15 years and over	Sleepinc	Both
14	2015	Annual	8.830	0.027	All days	15 years and over	Sleepinc	Both
15	2016	Annual	8.790	0.028	All days	15 years and over	Sleepinc	Both
16	2017	Annual	8.800	0.027	All days	15 years and over	Sleepinc	Both
17	2003	Annual	8.260	0.023	Nonholiday weekdays	15 years and over	Sleepinc	Both
18	2004	Annual	8.250	0.034	Nonholiday weekdays	15 years and over	Sleepinc	Both
19	2005	Annual	8.350	0.030	Nonholiday weekdays	15 years and over	Sleepinc	Both
20	2006	Annual	8.330	0.032	Nonholiday weekdays	15 years and over	Sleepinc	Both
21	2007	Annual	8.290	0.032	Nonholiday weekdays	15 years and over	Sleepinc	Both
22	2008	Annual	8.300	0.032	Nonholiday weekdays	15 years and over	Sleepinc	Both
23	2009	Annual	8.390	0.031	Nonholiday weekdays	15 years and over	Sleepinc	Both
24	2010	Annual	8.380	0.031	Nonholiday weekdays	15 years and over	Sleepinc	Both
25	2011	Annual	8.450	0.034	Nonholiday weekdays	15 years and over	Sleepinc	Both
26	2012	Annual	8.450	0.034	Nonholiday weekdays	15 years and over	Sleepinc	Both
27	2013	Annual	8.480	0.036	Nonholiday weekdays	15 years and over	Sleepinc	Both
28	2014	Annual	8.540	0.031	Nonholiday weekdays	15 years and over	Sleepinc	Both
29	2015	Annual	8.590	0.035	Nonholiday weekdays	15 years and over	Sleepinc	Both
30	2016	Annual	8.500	0.037	Nonholiday weekdays	15 years and over	Sleepinc	Both
31	2017	Annual	8.530	0.035	Nonholiday weekdays	15 years and over	Sleepinc	Both
32	2003	Annual	9.290	0.026	Weekend days and holiday	15 years and over	Sleepinc	Both

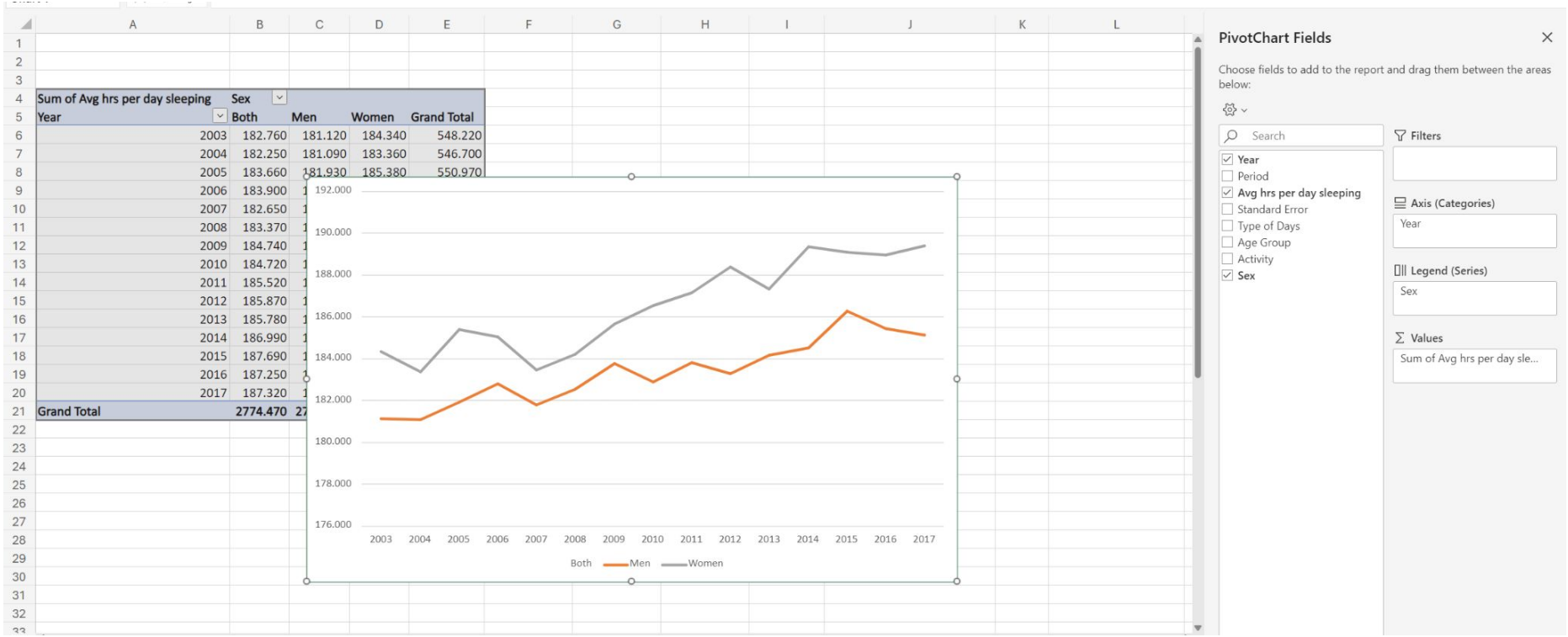
Step 2:
Insert on new sheet



Step 3:

Drag and drop columns into Axis (x-axis), Values (y-axis), Legend (groups)

Then click on the table and add any chart



jxf@mastodon.social
@jxxf · **Follow**

Optimist: The glass is ½ full.
Pessimist: The glass is ½ empty.
Excel: The glass is January 2nd.

5:33 PM · May 7, 2022

Read the full conversation on Twitter

259.2K