Introduction to Data Analysis I

Agenda - Schedule

1. Intro to Data Analytics
2. Basics & Charts
3. Functions & Pivot Tables
4. Break
5. Sheets Lab

When it comes to simple data analysis, spreadsheet software cannot be beat

Agenda - Goals

- Modify column names, sort columns, and format columns in Google Sheets
- Create bar-plots, line-plots, and scatter plots in Google sheets
- Calculate aggregate metrics using the functions
- Create and read pivot tables in Google Spreadsheets

Agenda - Announcements

- **Week 6 Pre-Class Quiz** due 4/15 (*2 attempts*)
- **Career Class on 4/17** (*this time for real*)
- **TLAB #2** due 4/21

Warm Up

Hypothesis Testing Warmup

We are data analysts running a hypothesis test on our website. We would like to figure out if adding a **ChatGPT Agent** leads to more sales.

We provide a version of our website **with** the ChatGPT agent to our **experimental group**. Our **null hypothesis** states that this group will spend an average of **$10 per customer.**

The experimental group spends on average **$2 per customer.** This gives us a **z-score of -4. Is this z-score in the critical region? Is this a desirable z-score?**

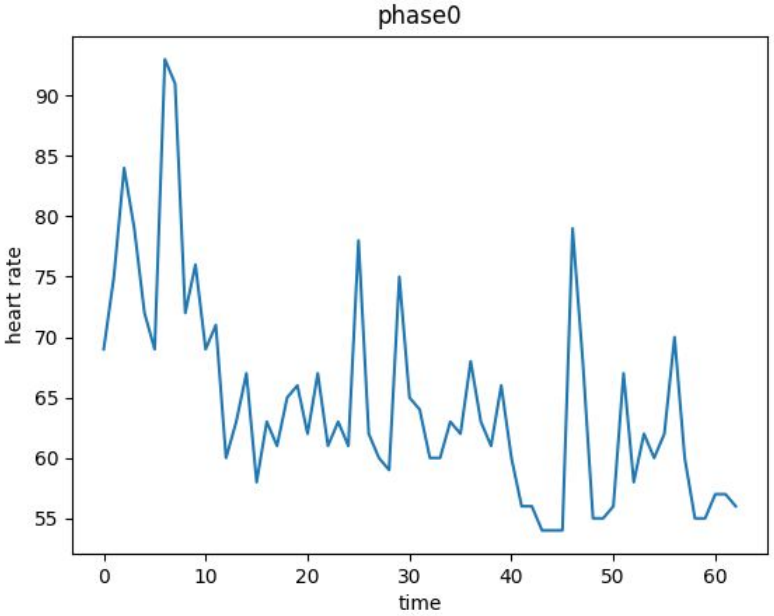
Data Analytics

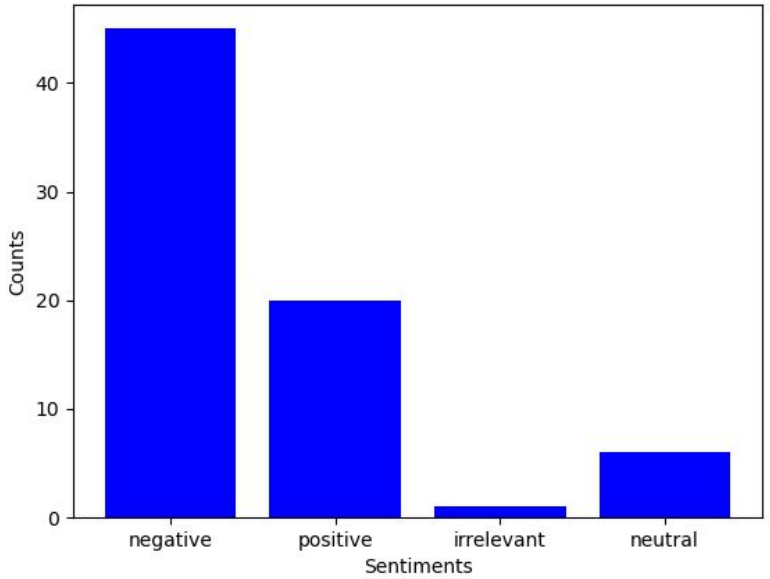
Introduction to Data Analytics

Now that we have the foundational computational knowledge needed to work through a series of data pipeline problems, let's begin applying our **statistical knowledge** to datasets.

Remember, it's vital to be able to code the pipelines needed to extract the data.

Now let's go about **analyzing and reasoning** about the data.





In both **TLAB 1 and TLAB 2,** we went about the process of generating visualizations, which we then wrote reports on. But, what are the **patterned tools** to completing such workflows?

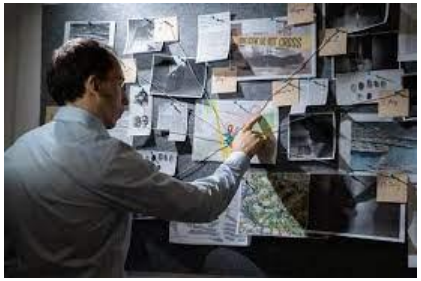
Data Analysis

Well, it's not actually **not** base Python.

Instead it entails a series of powerful data manipulation tools which will allow you to **transform and eventually analyze** entire data files.

We will dive into one of these tools today, but before that, let's go over what **data analysis** actually is.

Data Analysis



Data analysis could be defined as the following:

- Exploratory or intentional research of **insights, patterns, and trends…**
- …on a **consistent** set of data objects…
- …where you look for **reliable explanations** of how things occurred in the **past**, or how things will **occur in the future** assuming certain conditions.

This is a loaded set of statements that we must peel back for the remainder of phase 1.

Data Analysis

Before we begin to f**ormally define** which visualizations & steps you might take to accomplish the goals listed on the previous slide, let's run through an example.

You are a travel agency company. Like every company, you are trying to ensure your business survives and remains profitable. Which questions might you have about your business?

- ...

Data Analysis

You are a travel agency company. Like every company, you are trying to ensure your business survives and remains profitable. Which questions might you have about your business?

- **How much money did we spend and make last year?**
- **Which countries did people visit the most?**
- **How have TV-shows shaped our clients travel desires?**
- **Will tariffs increase or decrease travel?**

The next question is, **how do we answer these questions?**

Data Analysis

Well the answer first entails **getting the data**.

The next step entails **hiring a data analyst (you)** who is knowledge of the statistics and technology needed to effectively and reliably assess and make predictions on your data!

This entails doing a few things (some of which we've gone over already).

- **Transforming** your dataset
- Calculating **descriptive statistics**
- Making **pivot tables**
- Making **visualizations**

Data Analysis Spreadsheets

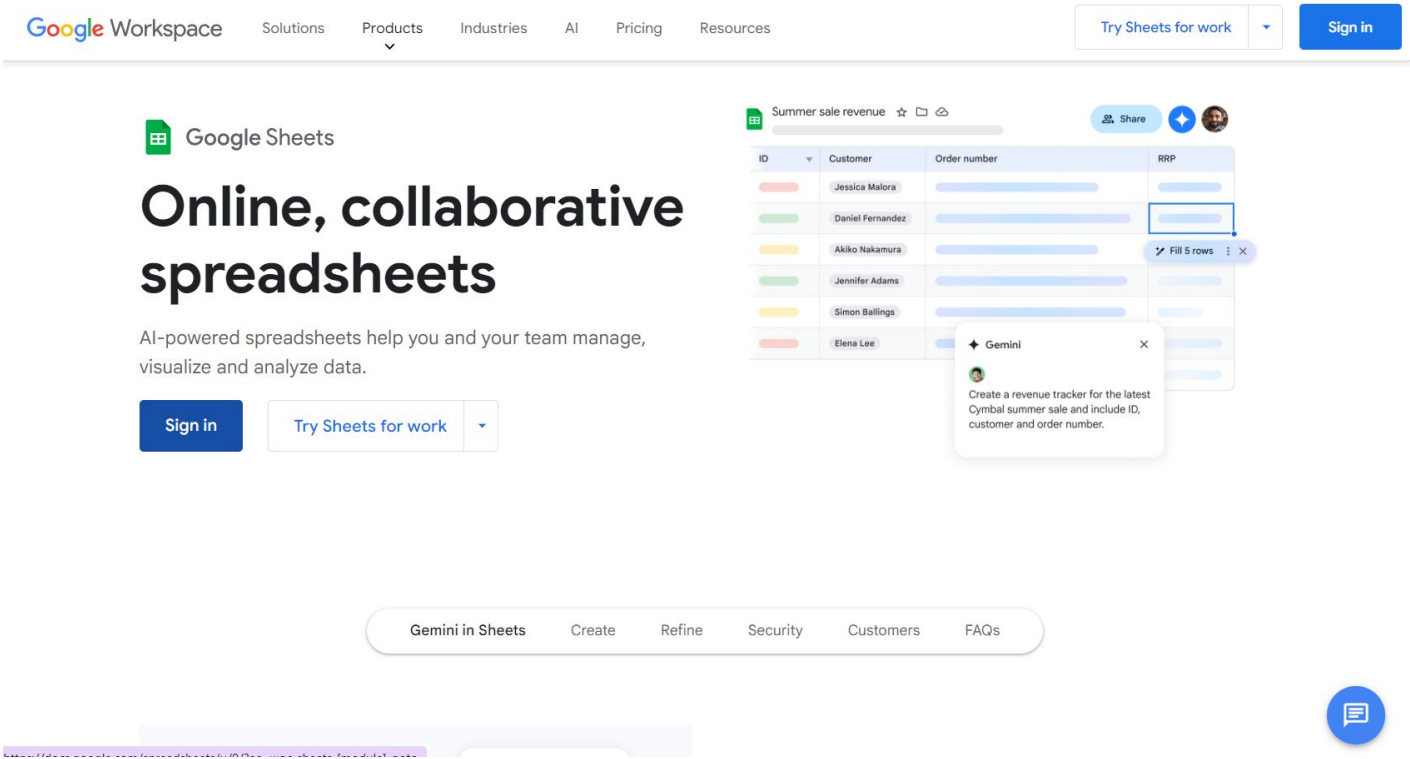
To introduce you all to this process, we will begin working with **Google Sheets**.

While not as “high-tech” as other data analysis solutions out there, it is still a reliable tool and effectively does what other Python-based tools do! In fact, some data analysis teams **solely** use spreadsheet technology to get their work done.

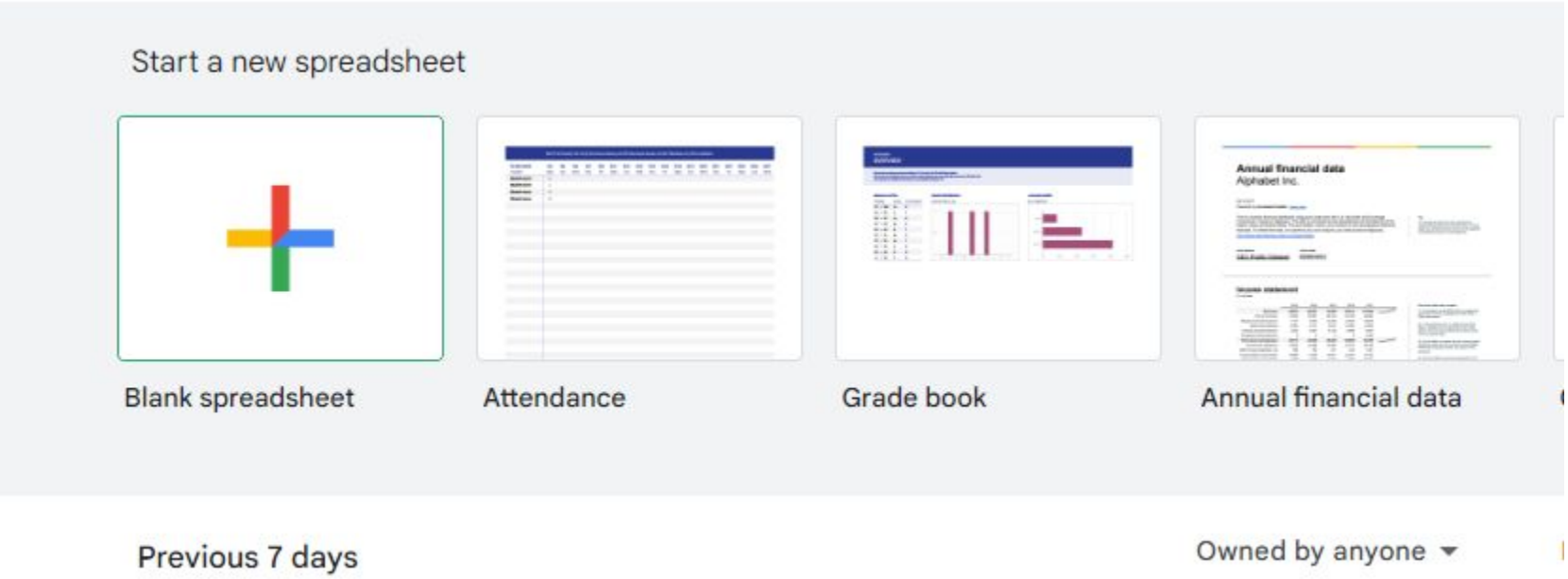
However, we generally don’t recommend limiting yourself to spreadsheets, as there are some downsides to this.

code	iso	country	PMD.raw.2	PMD.raw.2	PMD.raw.2	PMD.raw.2	PMD.raw.2019
826	GBR	United King	294.898	289.6565	286.4196	285.8509	280.5128
840	USA	United Stat	283.5238	267.2755	252.9543	250.8517	249.4297
32	ARG	Argentina	624.5758	620.0179	596.8604	585.5217	580.4381
156	CHN	China	1921.886	1854.628	1782.889	1749.996	1743.355
566	NGA	Nigeria	1929.854	1904.94	1849.285	1838.035	1869.209
818	EGY	Egypt	4388.482	4248.828	4089.566	4034.413	3993.224
76	BRA	Brazil	568.7507	568.0072	552.8456	546.0555	542.7444
392	JPN	Japan	262.7909	261.7103	257.2666	256.6441	256.0192
643	RUS	Russia	975.8511	871.8861	763.8981	754.5885	764.4997
276	DEU	Germany	375.4855	353.7205	338.2791	335.3877	334.5227

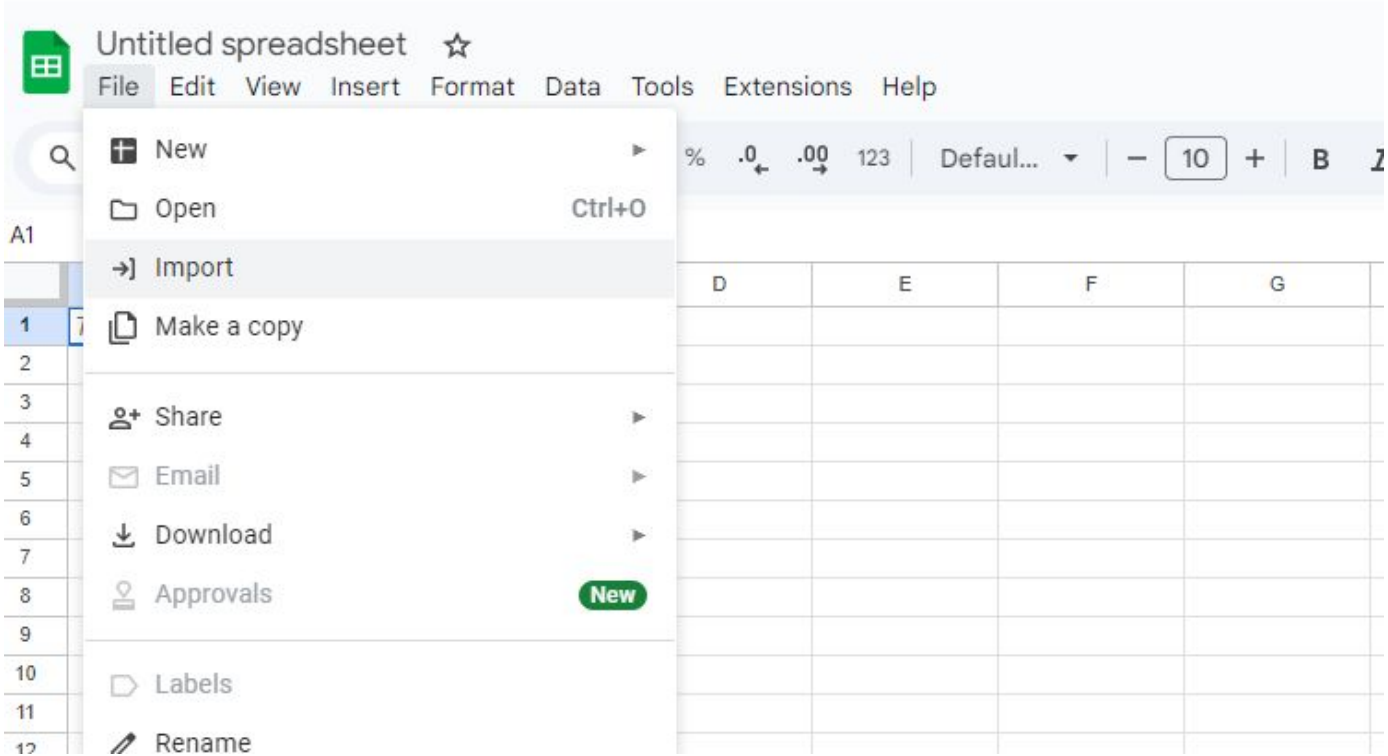
For today's spreadsheet demonstration we will work through the data file titled **sample.csv**. **Download and extract** the folder titled **spreadsheets_lab** to get this file.



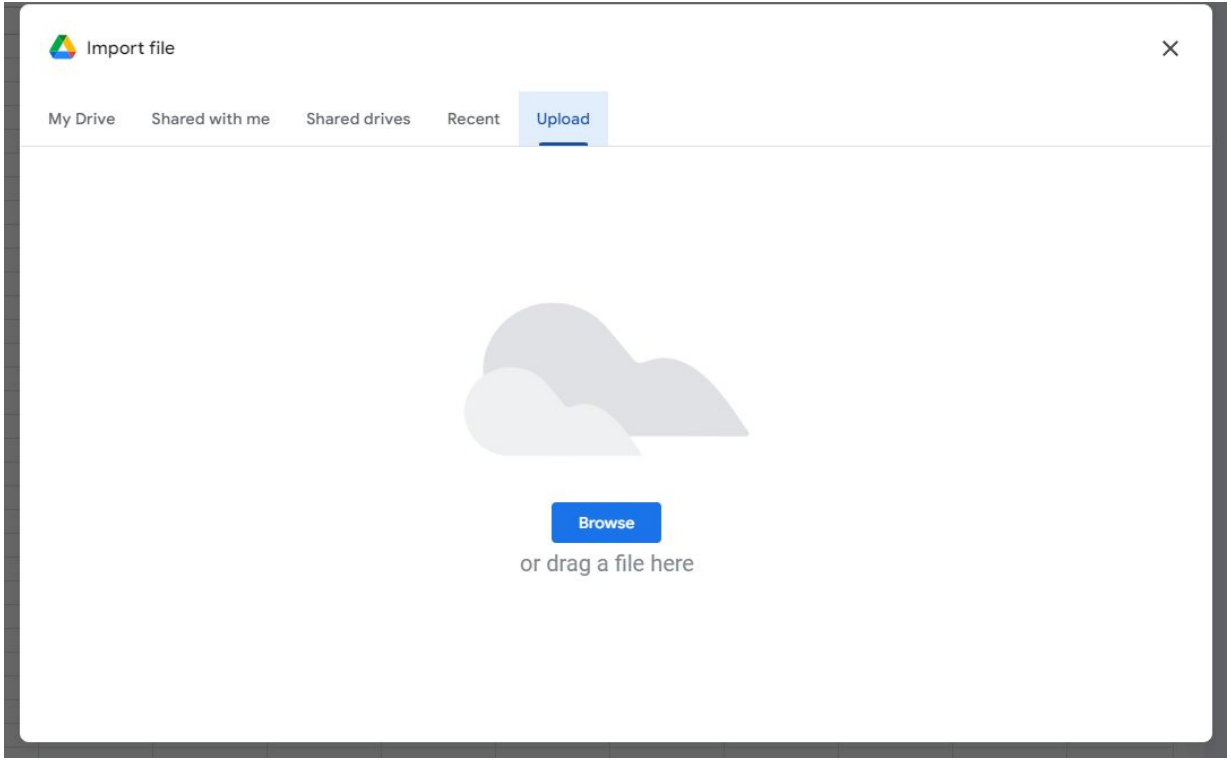
You can open this file by going to https://workspace.google.com/products/sheets/ and clicking on “Sign In”



From there, select “**Blank Spreadsheet**”



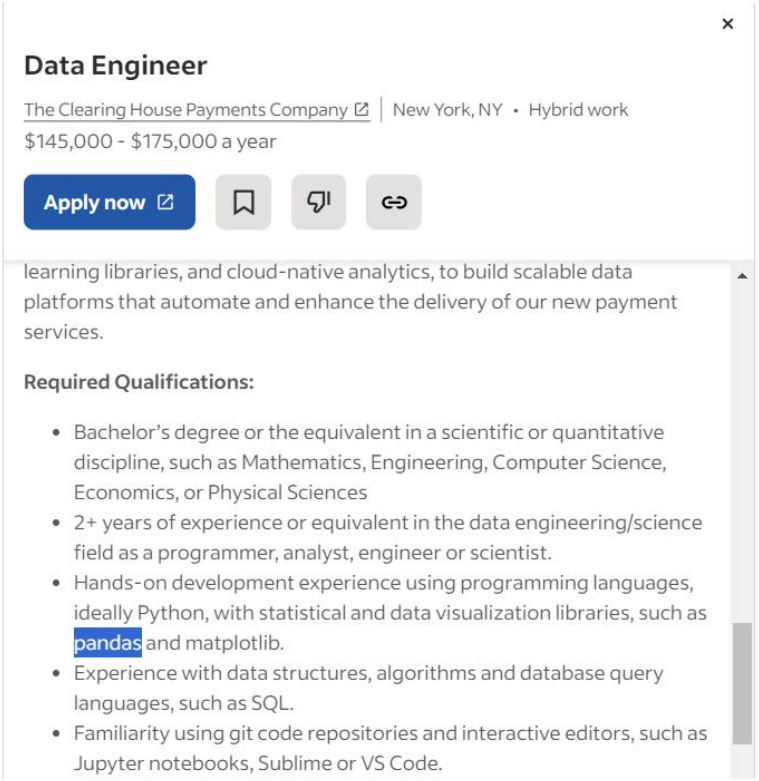
Next click **File** and then **Import.**



Then you can select the **Upload** tab, where you can click on **Browse** to search for the **sample.csv** file in your **spreadsheet_lab** folder.

A	B	C	D	E	F	G	H
code	iso	country	PMD.raw.2015	PMD.raw.2016	PMD.raw.2017	PMD.raw.2018	PMD.raw.2019
826	GBR	United Kingdom	294.8980096	289.6564515	286.4196291	285.8509262	280.5127785
840	USA	United States of	283.5238203	267.2755296	252.9543121	250.8516978	249.4296641
32	ARG	Argentina	624.5758182	620.0178524	596.8603743	585.5216743	580.4380542
156	CHN	China	1921.885946	1854.62846	1782.888749	1749.996404	1743.3549
566	NGA	Nigeria	1929.854375	1904.940118	1849.284647	1838.035469	1869.209248
818	EGY	Egypt	4388.481759	4248.828354	4089.565822	4034.412844	3993.224209
76	BRA	Brazil	568.7506785	568.0071932	552.8455853	546.0554702	542.7443866
392	JPN	Japan	262.7909262	261.7102781	257.2665629	256.6441396	256.0192158
643	RUS	Russia	975.8511373	871.8861424	763.8980703	754.5885114	764.4997137
276	DEU	Germany	375.4855158	353.7205444	338.2790574	335.3876598	334.522656

Upon successful import, you will be able to view the following file.



Data Engineer

The Clearing House Payments Company | New York, NY • Hybrid work

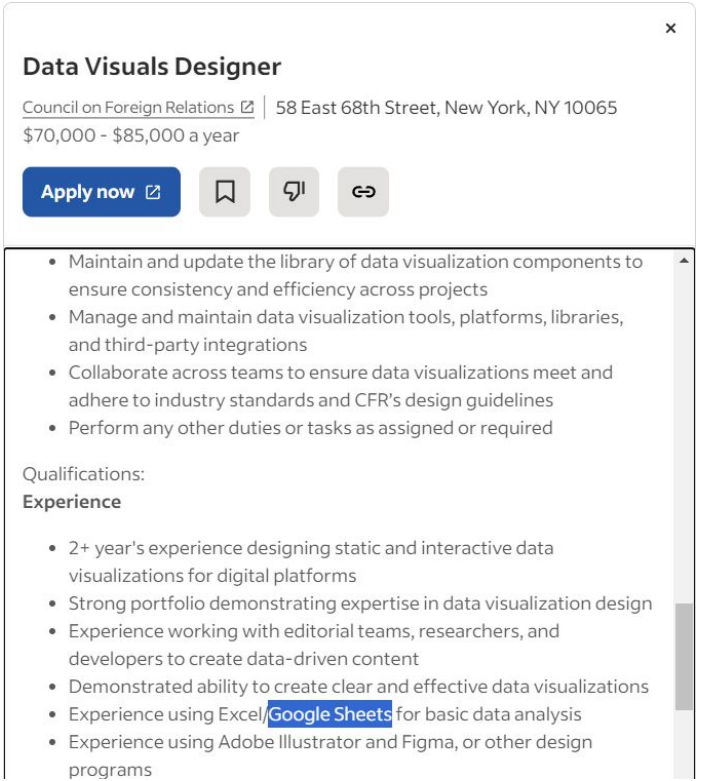
$145,000 - $175,000 a year

Apply now

learning libraries, and cloud-native analytics, to build scalable data platforms that automate and enhance the delivery of our new payment services.

Required Qualifications:

- Bachelor's degree or the equivalent in a scientific or quantitative discipline, such as Mathematics, Engineering, Computer Science, Economics, or Physical Sciences
- 2+ years of experience or equivalent in the data engineering/science field as a programmer, analyst, engineer or scientist.
- Hands-on development experience using programming languages, ideally Python, with statistical and data visualization libraries, such as pandas and matplotlib.
- Experience with data structures, algorithms and database query languages, such as SQL.
- Familiarity using git code repositories and interactive editors, such as Jupyter notebooks, Sublime or VS Code.



Data Visuals Designer

Council on Foreign Relations | 58 East 68th Street, New York, NY 10065

$70,000 - $85,000 a year

Apply now

- Maintain and update the library of data visualization components to ensure consistency and efficiency across projects
- Manage and maintain data visualization tools, platforms, libraries, and third-party integrations
- Collaborate across teams to ensure data visualizations meet and adhere to industry standards and CFR's design guidelines
- Perform any other duties or tasks as assigned or required

Qualifications:

Experience

- 2+ year's experience designing static and interactive data visualizations for digital platforms
- Strong portfolio demonstrating expertise in data visualization design
- Experience working with editorial teams, researchers, and developers to create data-driven content
- Demonstrated ability to create clear and effective data visualizations
- Experience using Excel/Google Sheets for basic data analysis
- Experience using Adobe Illustrator and Figma, or other design programs

A **good chunk** of data-related position will have you working with the tools we'll be discussing.

Spreadsheets - Basics

Spreadsheets

When beginning a data analysis project, the best thing we can do is to simply **view the data in Excel** (or Google Sheets in our context).

No technologist is too good for spreadsheets!Viewing in spreadsheets allows you to:

- *Quickly understand data*
- *Notice missing values*
- *Find dataset features*

Spreadsheets

Furthermore, when working in tech-adjacent roles where your **coworkers are not programmers**, sheets are a nice reprieve from reading code.

Example: *Your growth analyst coworker would like you to modify a dataset of travel data.*

Spreadsheets

Giving them a spreadsheet is usually an easier “sell” than **making them learn Python**.

Think of spreadsheets as the bridge between your computational knowledge and your stakeholders domain knowledge.

```python
import csv

import time
from datetime import datetime


You, 2 weeks ago | 1 author (You)
class StockData:
    def __init__(self, path):
        self.path = path
        self.data = None

    def load(self):
        """A function to assing the csv dataset to a list of lists.
        Does not include headers.

        All data is expected to be a string
        """
        # to keep us in compliance with EU standards, we must log the datetime
        # of all data loads
        epoch = time.time()
        self._date = datetime.utcfromtimestamp(epoch).\
            strftime('%Y-%m-%d %H:%M:%S')
        print("DATA LOADED AT", self._date)

        # open data using csv reader
        with open(self.path, newline='') as file:
            reader = csv.reader(file)
            # skip header row
```

Date	Monday	Tuesday	Wednesd	Thursday	Friday
3/2/2020	97.6975	95.4495	98.7915	96.2015	95.0545
3/9/2020	90.0305	94.591	91.043	83.8305	89.25
3/16/2020	84.4575	90.392	91.5	94.0465	92.3045
3/23/2020	95.1415	97.005	94.292	97.7745	95.005
3/30/2020	98.1975	97.486	95.385	95.9415	95.3295
4/6/2020	99.8795	100.58	102.15	102.138	
4/13/2020	108.4435	114.166	115.384	120.4095	118.75
4/20/2020	119.6805	116.406	118.1745	119.9725	120.511
4/27/2020	118.8	115.704	118.6355		

code	iso	country	PMD.raw.2015	PMD.raw.2016	PMD.raw.2017	PMD.raw.2018	PMD.raw.2019
826	GBR	United Kingdom	294.8980096	289.6564515	286.4196291	285.8509262	280.5127785
840	USA	United States of	283.5238203	267.2755296	252.9543121	250.8516978	249.4296641
32	ARG	Argentina	624.5758182	620.0178524	596.8603743	585.5216743	580.4380542
156	CHN	China	1921.885946	1854.62846	1782.888749	1749.996404	1743.3549
566	NGA	Nigeria	1929.854375	1904.940118	1849.284647	1838.035469	1869.209248
818	EGY	Egypt	4388.481759	4248.828354	4089.565822	4034.412844	3993.224209
76	BRA	Brazil	568.7506785	568.0071932	552.8455853	546.0554702	542.7443866
392	JPN	Japan	262.7909262	261.7102781	257.2665629	256.6441396	256.0192158
643	RUS	Russia	975.8511373	871.8861424	763.8980703	754.5885114	764.4997137
276	DEU	Germany	375.4855158	353.7205444	338.2790574	335.3876598	334.522656

Lets analyze a dataset of global air health metrics solely in google sheets. **No Python, no pandas, and no numpy**.

Spreadsheets - Notations

To get started with sheets, lets iron out a few key phrases:

Column

A **vertical** arrangement of data indicated by letters.

Row

A **horizontal** arrangement of data indicated by numbers.

Cell

An **individual square** in our sheet that aligns with some **row/column**

code	iso	country	MD.raw.2015	PMD.raw.2016	PMD.raw.2017	PMD.raw.2018	PMD.raw.2019
826	GBR	United Kingdom	294.8980096	289.6564515	286.4196291	285.8509262	280.5127785
840	USA	United States of	283.5238203	267.2755296	252.9543121	250.8516978	249.4296641
32	ARG	Argentina	624.5758182	620.0178524	596.8603743	585.5216743	580.4380542
156	CHN	China	1921.885946	1854.62846	1782.888749	1749.996404	1743.3549
566	NGA	Nigeria	1929.854375	1904.940118	1849.284647	1838.035469	1869.209248
818	EGY	Egypt	4388.481759	4248.828354	4089.565822	4034.412844	3993.224209
76	BRA	Brazil	568.7506785	568.0071932	552.8455853	546.0554702	542.7443866
392	JPN	Japan	262.7909262	261.7102781	257.2665629	256.6441396	256.0192158
643	RUS	Russia	975.8511373	871.8861424	763.8980703	754.5885114	764.4997137
276	DEU	Germany	375.4855158	353.7205444	338.2790574	335.3876598	334.522656

Column: vertical arrangement of data. We often label our own columns as they correspond to a **specific "attribute" of our dataset.** We can call this column "country" or "C"

code	iso	country	PMD.raw.2015	PMD.raw.2016	PMD.raw.2017	PMD.raw.2018	PMD.raw.2019
826	GBR	United Kingdom	284.8880896	280.6564515	286.4196201	285.8590262	280.5127795
840	USA	United States of	283.5238203	267.2755296	252.9543121	250.8516978	249.4296641
32	ARG	Argentina	624.5758182	620.0178524	596.8603743	585.5216743	580.4380542
156	CHN	China	1921.885946	1854.62846	1782.888749	1749.996404	1743.3549
566	NGA	Nigeria	1929.854375	1904.940118	1849.284647	1838.035469	1869.209248
818	EGY	Egypt	4388.481759	4248.828354	4089.565822	4034.412844	3993.224209
76	BRA	Brazil	568.7506785	568.0071932	552.8455853	546.0554702	542.7443866
392	JPN	Japan	262.7909262	261.7102781	257.2665629	256.6441396	256.0192158
643	RUS	Russia	975.8511373	871.8861424	763.8980703	754.5885114	764.4997137
276	DEU	Germany	375.4855158	353.7205444	338.2790574	335.3876598	334.522656

Row: horizontal arrangement of data. We sometimes label this as well, but not too often. Usually a row indicates a **new sample**. We can call this row 3.

code	iso	country	PMD.raw.2015	PMD.raw.2016	PMD.raw.2017	PMD.raw.2018	PMD.raw.2019
826	GBR	[illegible]	294.8980096	289.6564515	286.4196291	285.8509262	280.5127785
840	USA	United States of	283.5238203	267.2755296	252.9543121	250.8516978	249.4296641
32	ARG	Argentina	624.5758182	620.0178524	596.8603743	585.5216743	580.4380542
156	CHN	China	1921.885946	1854.62846	1782.888749	1749.996404	1743.3549
566	NGA	Nigeria	1929.854375	1904.940118	1849.284647	1838.035469	1869.209248
818	EGY	Egypt	4388.481759	4248.828354	4089.565822	4034.412844	3993.224209
76	BRA	Brazil	568.7506785	568.0071932	552.8455853	546.0554702	542.7443866
392	JPN	Japan	262.7909262	261.7102781	257.2665629	256.6441396	256.0192158
643	RUS	Russia	975.8511373	871.8861424	763.8980703	754.5885114	764.4997137
276	DEU	Germany	375.4855158	353.7205444	338.2790574	335.3876598	334.522656

Cell: an individual square of our data. When referring to a cell we first mention row then column. In this case it's **C3**
(*aka row 3, column C, or row 3, column "country", or row 3 column 3*)

Spreadsheets - Basic Controls

The key features of Sheets include

Navigating Cells

Using our mouse or keyboard to move across the dataset.

Modifying Cells

Updating the value in a cell by simply clicking in a cell and typing.

Automating Updates

Letting Sheets “intelligently” auto-complete our cells by dragging.

code	iso	country	PMD.raw.2015	PMD.raw.2016	PMD.raw.2017	PMD.raw.2018	PMD.raw.2019
826	GBR	United Kingdom	294.8980096	289.6564515	286.4196291	285.8509262	280.5127785
840	USA	United States of	283.5238203	267.2755296	252.9543121	250.8516978	249.4296641
32	ARG	Argentina	624.5758182	620.0178524	596.8603743	585.5216743	580.4380542
156	CHN	China	1921.885946	1854.62846	1782.888749	1749.996404	1743.3549
566	NGA	Nigeria	1929.854375	1904.940118	1849.284647	1838.035469	1869.209248
818	EGY	Egypt	4388.481759	4248.828354	4089.565822	4034.412844	3993.224209
76	BRA	Brazil	568.7506785	568.0071932	552.8455853	546.0554702	542.7443866
392	JPN	Japan	262.7909262	261.7102781	257.2665629	256.6441396	256.0192158
643	RUS	Russia	975.8511373	871.8861424	763.8980703	754.5885114	764.4997137
276	DEU	Germany	375.4855158	353.7205444	338.2790574	335.3876598	334.522656

Modification: let's say I dislike the names of these year columns, let's update them

code	iso	country	2015	PMD.raw.2016	PMD.raw.2017	PMD.raw.2018	PMD.raw.2019
826	GBR	United Kingdom	294.8980096	289.6564515	286.4196291	285.8509262	280.5127785
840	USA	United States of	283.5238203	267.2755296	252.9543121	250.8516978	249.4296641
32	ARG	Argentina	624.5758182	620.0178524	596.8603743	585.5216743	580.4380542
156	CHN	China	1921.885946	1854.62846	1782.888749	1749.996404	1743.3549
566	NGA	Nigeria	1929.854375	1904.940118	1849.284647	1838.035469	1869.209248
818	EGY	Egypt	4388.481759	4248.828354	4089.565822	4034.412844	3993.224209
76	BRA	Brazil	568.7506785	568.0071932	552.8455853	546.0554702	542.7443866
392	JPN	Japan	262.7909262	261.7102781	257.2665629	256.6441396	256.0192158
643	RUS	Russia	975.8511373	871.8861424	763.8980703	754.5885114	764.4997137
276	DEU	Germany	375.4855158	353.7205444	338.2790574	335.3876598	334.522656

Modification: let's say I dislike the names of these year columns, let's update them

code	iso	country	201	2016	PMD.raw.2017	PMD.raw.2018	PMD.raw.2019
826	GBR	United Kingdom	294.8980096	[illegible]	286.4196291	285.8509262	280.5127785
840	USA	United States of	283.5238203	267.2755296	252.9543121	250.8516978	249.4296641
32	ARG	Argentina	624.5758182	620.0178524	596.8603743	585.5216743	580.4380542
156	CHN	China	1921.885946	1854.62846	1782.888749	1749.996404	1743.3549
566	NGA	Nigeria	1929.854375	1904.940118	1849.284647	1838.035469	1869.209248
818	EGY	Egypt	4388.481759	4248.828354	4089.565822	4034.412844	3993.224209
76	BRA	Brazil	568.7506785	568.0071932	552.8455853	546.0554702	542.7443866
392	JPN	Japan	262.7909262	261.7102781	257.2665629	256.6441396	256.0192158
643	RUS	Russia	975.8511373	871.8861424	763.8980703	754.5885114	764.4997137
276	DEU	Germany	375.4855158	353.7205444	338.2790574	335.3876598	334.522656

Modification: let's say I dislike the names of these year columns, let's update them

code	iso	country	2015	2016	MD.raw.2017	PMD.raw.2018	PMD.raw.2019
826	GBR	United Kingdom	294.8980096	289.6564515	286.4196291	285.8509262	280.5127785
840	USA	United States of	283.5238203	267.2755296	252.9543121	250.8516978	249.4296641
32	ARG	Argentina	624.5758182	620.0178524	596.8603743	585.5216743	580.4380542
156	CHN	China	1921.885946	1854.62846	1782.888749	1749.996404	1743.3549
566	NGA	Nigeria	1929.854375	1904.940118	1849.284647	1838.035469	1869.209248
818	EGY	Egypt	4388.481759	4248.828354	4089.565822	4034.412844	3993.224209
76	BRA	Brazil	568.7506785	568.0071932	552.8455853	546.0554702	542.7443866
392	JPN	Japan	262.7909262	261.7102781	257.2665629	256.6441396	256.0192158
643	RUS	Russia	975.8511373	871.8861424	763.8980703	754.5885114	764.4997137
276	DEU	Germany	375.4855158	353.7205444	338.2790574	335.3876598	334.522656

Automatic Updates: Let's autocomplete by dragging across my two updated cells over **my previously created columns**

code	iso	country	2015	2016	2017	2018	2019
826	GBR	United Kingdom	294.8980096	289.6564515	286.4196291	285.8509262	280.5127785
840	USA	United States of	283.5238203	267.2755296	252.9543121	250.8516978	249.4296641
32	ARG	Argentina	624.5758182	620.0178524	596.8603743	585.5216743	580.4380542
156	CHN	China	1921.885946	1854.62846	1782.888749	1749.996404	1743.3549
566	NGA	Nigeria	1929.854375	1904.940118	1849.284647	1838.035469	1869.209248
818	EGY	Egypt	4388.481759	4248.828354	4089.565822	4034.412844	3993.224209
76	BRA	Brazil	568.7506785	568.0071932	552.8455853	546.0554702	542.7443866
392	JPN	Japan	262.7909262	261.7102781	257.2665629	256.6441396	256.0192158
643	RUS	Russia	975.8511373	871.8861424	763.8980703	754.5885114	764.4997137
276	DEU	Germany	375.4855158	353.7205444	338.2790574	335.3876598	334.522656

This automatically updates all subsequent columns.

Spreadsheets - Basic Controls

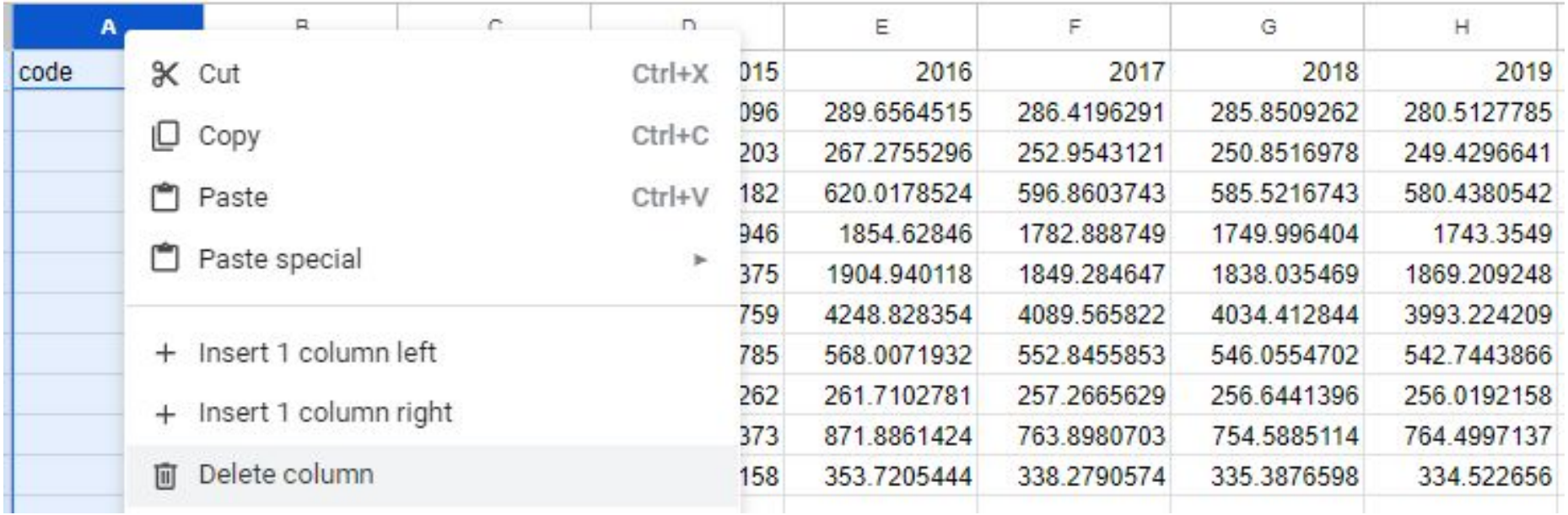
It also pays to know that there are more advanced features by "right-clicking" on a column, row, or cell. We can:

Delete

Remove data

Sort

Arrange your data



Delete: I don't care about the "code" column so I will just delete it.

A	B	C	D	E	F	G
iso	country	PMD.raw.2015	PMD.raw.2016	PMD.raw.2017	PMD.raw.2018	PMD.raw.2019
GBR	United Kingdom	294.8980096	289.6564515	286.4196291	285.8509262	280.5127785
USA	United States of	283.5238203	267.2755296	252.9543121	250.8516978	249.4296641
ARG	Argentina	624.5758182	620.0178524	596.8603743	585.5216743	580.4380542
CHN	China	1921.885946	1854.62846	1782.888749	1749.996404	1743.3549
NGA	Nigeria	1929.854375	1904.940118	1849.284647	1838.035469	1869.209248
EGY	Egypt	4388.481759	4248.828354	4089.565822	4034.412844	3993.224209
BRA	Brazil	568.7506785	568.0071932	552.8455853	546.0554702	542.7443866
JPN	Japan	262.7909262	261.7102781	257.2665629	256.6441396	256.0192158
RUS	Russia	975.8511373	871.8861424	763.8980703	754.5885114	764.4997137
DEU	Germany	375.4855158	353.7205444	338.2790574	335.3876598	334.522656

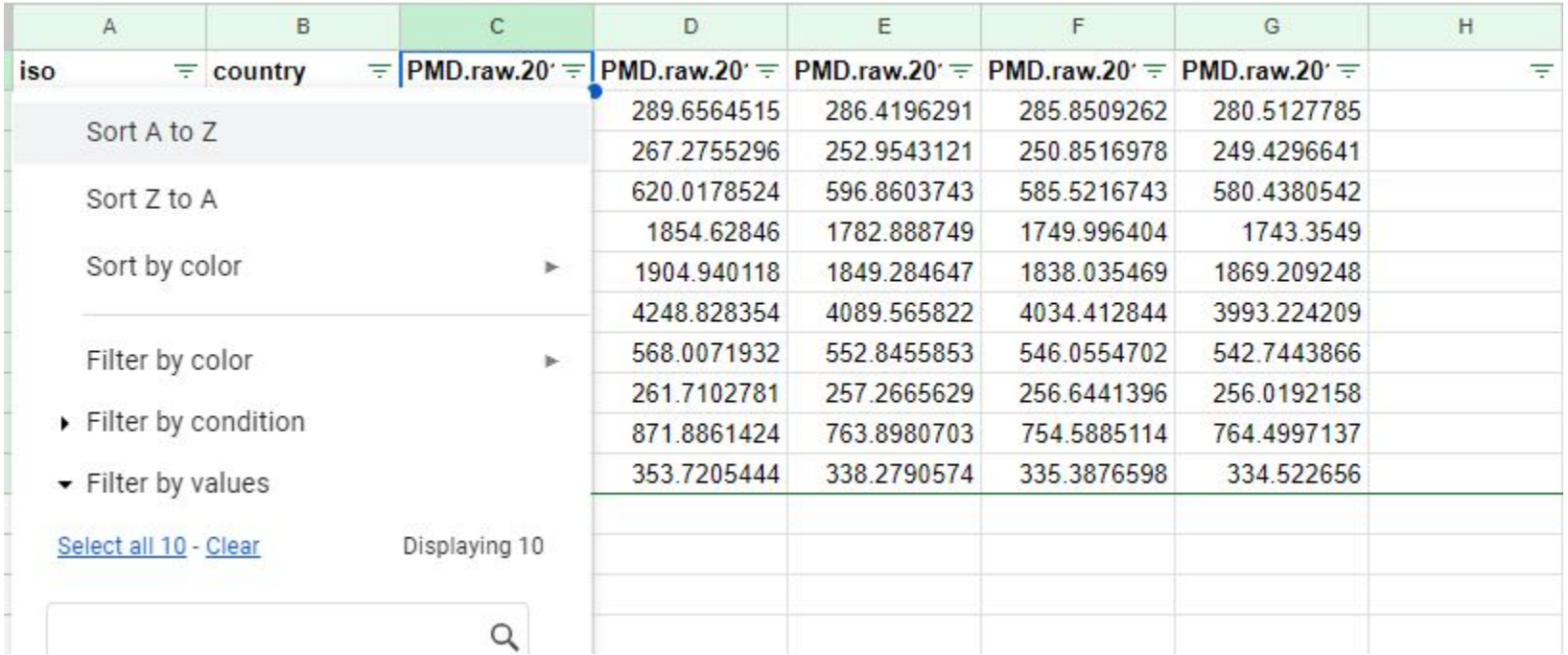
Sort: Lastly, I want to arrange my data in such a way that I can see who had the lowest years of life lost per 100,000 people for poor air quality (for 2015)

	A	B	C	D	E	F	G	H
1	iso	country	PMD.raw.2015	PMD.raw.2016	PMD.raw.2017	PMD.raw.2018	PMD.raw.2019	
2	GBR	United Kingdom	294.8980096	289.6564515	286.4196291	285.8509262	280.5127785	
3	USA	United States of	283.5238203	267.2755296	252.9543121	250.8516978	249.4296641	
4	ARG	Argentina	624.5758182	620.0178524	596.8603743	585.5216743	580.4380542	
5	CHN	China	1921.885946	1854.62846	1782.888749	1749.996404	1743.3549	
6	NGA	Nigeria	1929.854375	1904.940118	1849.284647	1838.035469	1869.209248	
7	EGY	Egypt	4388.481759	4248.828354	4089.565822	4034.412844	3993.224209	
8	BRA	Brazil	568.7506785	568.0071932	552.8455853	546.0554702	542.7443866	
9	JPN	Japan	262.7909262	261.7102781	257.2665629	256.6441396	256.0192158	
10	RUS	Russia	975.8511373	871.8861424	763.8980703	754.5885114	764.4997137	
11	DEU	Germany	375.4855158	353.7205444	338.2790574	335.3876598	334.522656	
12								

Sort: I select the first row which represents my column names. From there, I select the “funnel” icon which will apply a sort to all these columns

iso	country	PMD.raw.20'	PMD.raw.20'	PMD.raw.20'	PMD.raw.20'	PMD.raw.20'
GBR	United Kingdom	294.8980096	289.6564515	286.4196291	285.8509262	280.5127785
USA	United States of	283.5238203	267.2755296	252.9543121	250.8516978	249.4296641
ARG	Argentina	624.5758182	620.0178524	596.8603743	585.5216743	580.4380542
CHN	China	1921.885946	1854.62846	1782.888749	1749.996404	1743.3549
NGA	Nigeria	1929.854375	1904.940118	1849.284647	1838.035469	1869.209248
EGY	Egypt	4388.481759	4248.828354	4089.565822	4034.412844	3993.224209
BRA	Brazil	568.7506785	568.0071932	552.8455853	546.0554702	542.7443866
JPN	Japan	262.7909262	261.7102781	257.2665629	256.6441396	256.0192158
RUS	Russia	975.8511373	871.8861424	763.8980703	754.5885114	764.4997137
DEU	Germany	375.4855158	353.7205444	338.2790574	335.3876598	334.522656

We've created a filter, now let's apply a sort on the column that represents 2015.



We click on the filter icon for our column of interest and **specify the arraignment** (*least to greatest A to Z*) or (*greatest to least Z to A*)

A	B	C
iso	country	PMD.raw.2015
JPN	Japan	262.7909262
USA	United States of	283.5238203
GBR	United Kingdom	294.8980096
DEU	Germany	375.4855158
BRA	Brazil	568.7506785
ARG	Argentina	624.5758182
RUS	Russia	975.8511373
CHN	China	1921.885946
NGA	Nigeria	1929.854375
EGY	Egypt	4388.481759

Now we see which country has the highest and lowest air-quality EPI metric for 2015.

Spreadsheets - Format Controls

Just like Python, we have different data-types. In sheets, we have the ability to **modify data-types according to human-readable types**. A couple of important types to be aware of include:

Currency: Dollars & cents. Will cut off decimal values.

Data + Time: Will force numbers into dates

Number: Safe bet. Will keep a number as is.

A	B	C
iso	country	PMD.raw.2015
JPN	Japan	262.79
USA	United States of	283.52
GBR	United Kingdom	294.90
DEU	Germany	375.49
BRA	Brazil	568.75
ARG	Argentina	624.58
RUS	Russia	975.85
CHN	China	1921.89
NGA	Nigeria	1929.85
EGY	Egypt	4388.48

By using the data format controls above, I can manipulate my column to represent an easier-to-read real number.

Spreadsheets - Graphs

Spreadsheets - Graphs

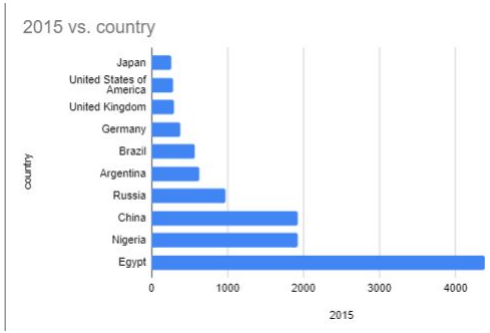
Recall that during exploratory data analysis, we generate visualizations to reveal trends and dataset patterns.

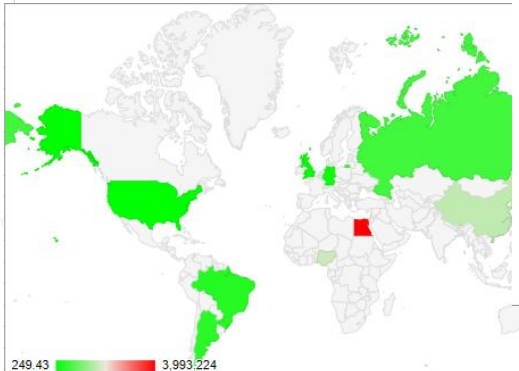
Some common visualizations include:

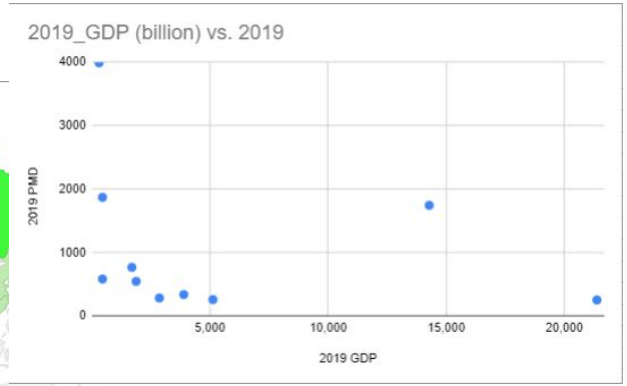
- **Bar Charts**
- **Line Plots**
- **Scatter Charts**

Spreadsheets - Graphs

Using our table as our primary reference, w**e can make a variety of visualizations**.

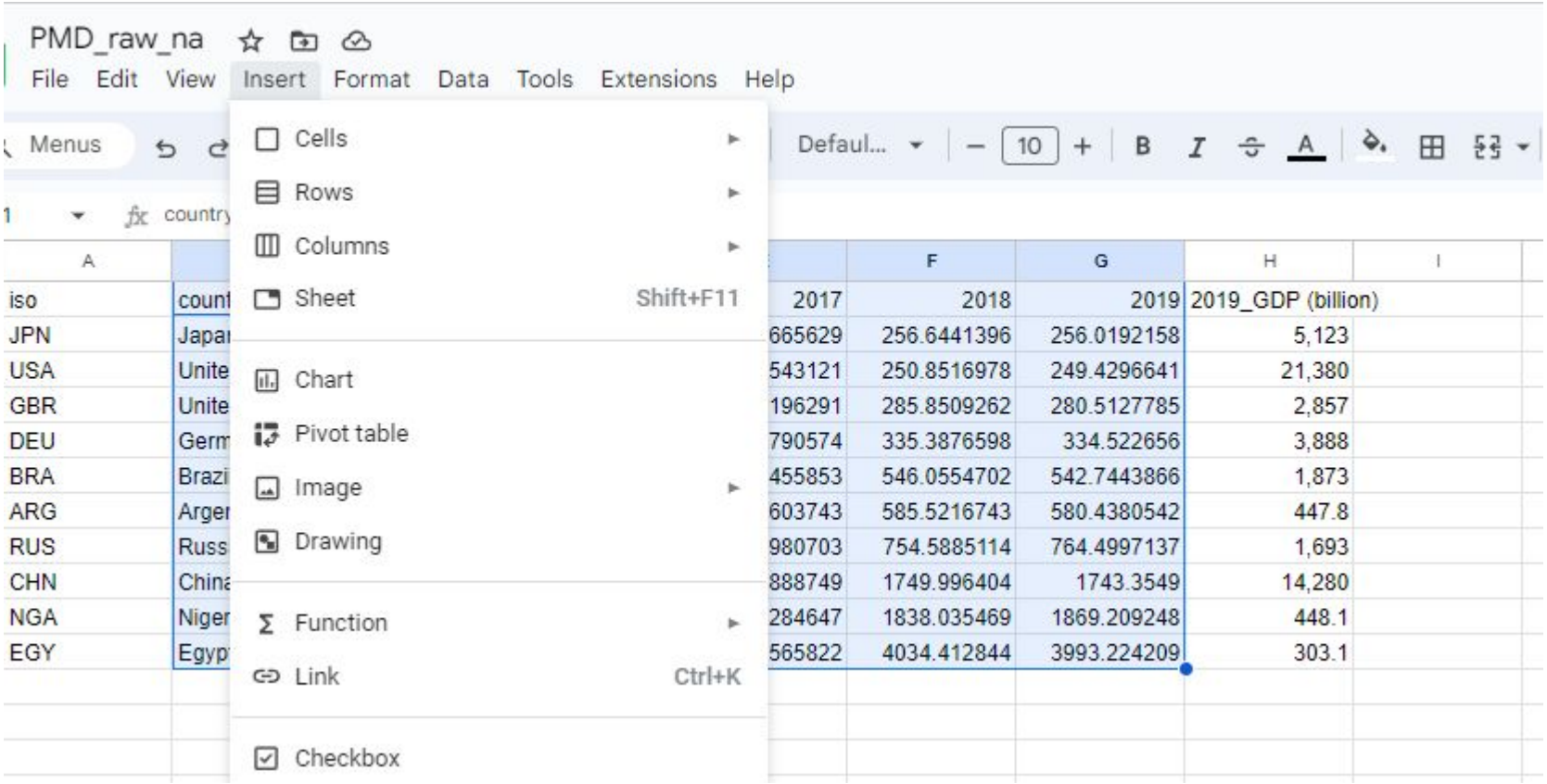




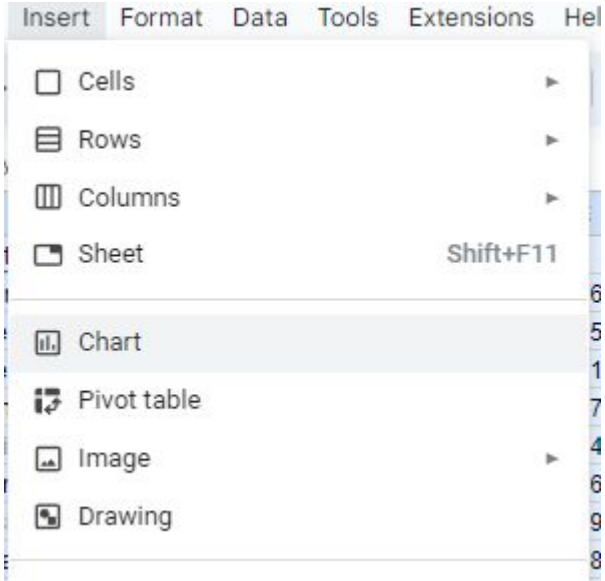


A	B	C	D	E	F	G	H
iso	country	2015	2016	2017	2018	2019	2019_GDP (billio
JPN	Japan	262.7909262	261.7102781	257.2665629	256.6441396	256.0192158	5,123
USA	United States of	283.5238203	267.2755296	252.9543121	250.8516978	249.4296641	21,380
GBR	United Kingdom	294.8980096	289.6564515	286.4196291	285.8509262	280.5127785	2,857
DEU	Germany	375.4855158	353.7205444	338.2790574	335.3876598	334.522656	3,888
BRA	Brazil	568.7506785	568.0071932	552.8455853	546.0554702	542.7443866	1,873
ARG	Argentina	624.5758182	620.0178524	596.8603743	585.5216743	580.4380542	447.8
RUS	Russia	975.8511373	871.8861424	763.8980703	754.5885114	764.4997137	1,693
CHN	China	1921.885946	1854.62846	1782.888749	1749.996404	1743.3549	14,280
NGA	Nigeria	1929.854375	1904.940118	1849.284647	1838.035469	1869.209248	448.1
EGY	Egypt	4388.481759	4248.828354	4089.565822	4034.412844	3993.224209	303.1

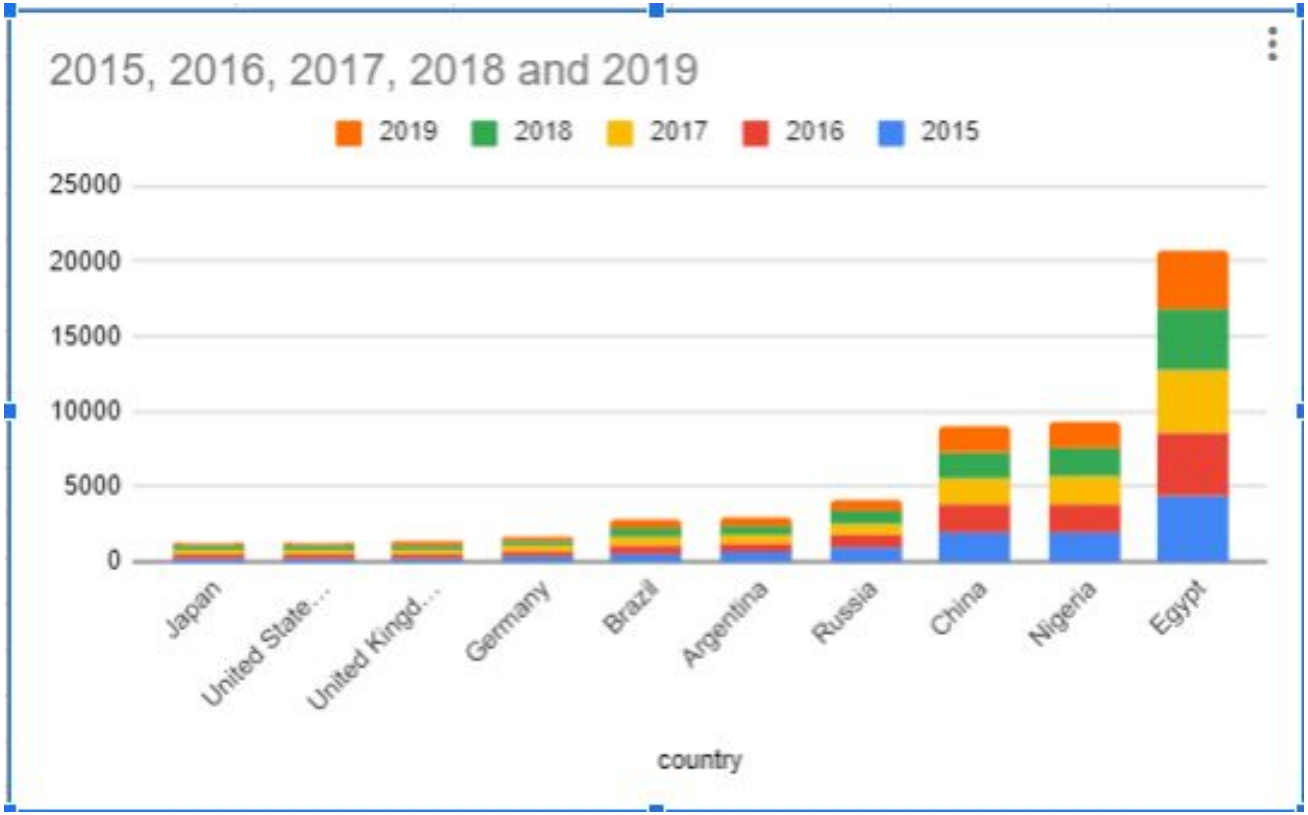
I first highlight my columns of interest by dragging my mouse across my table.



And then select insert on the menu-bar.



From there, notice how I have the ability to select a chart.



Sheets will make a best-guess as to which chart I need.

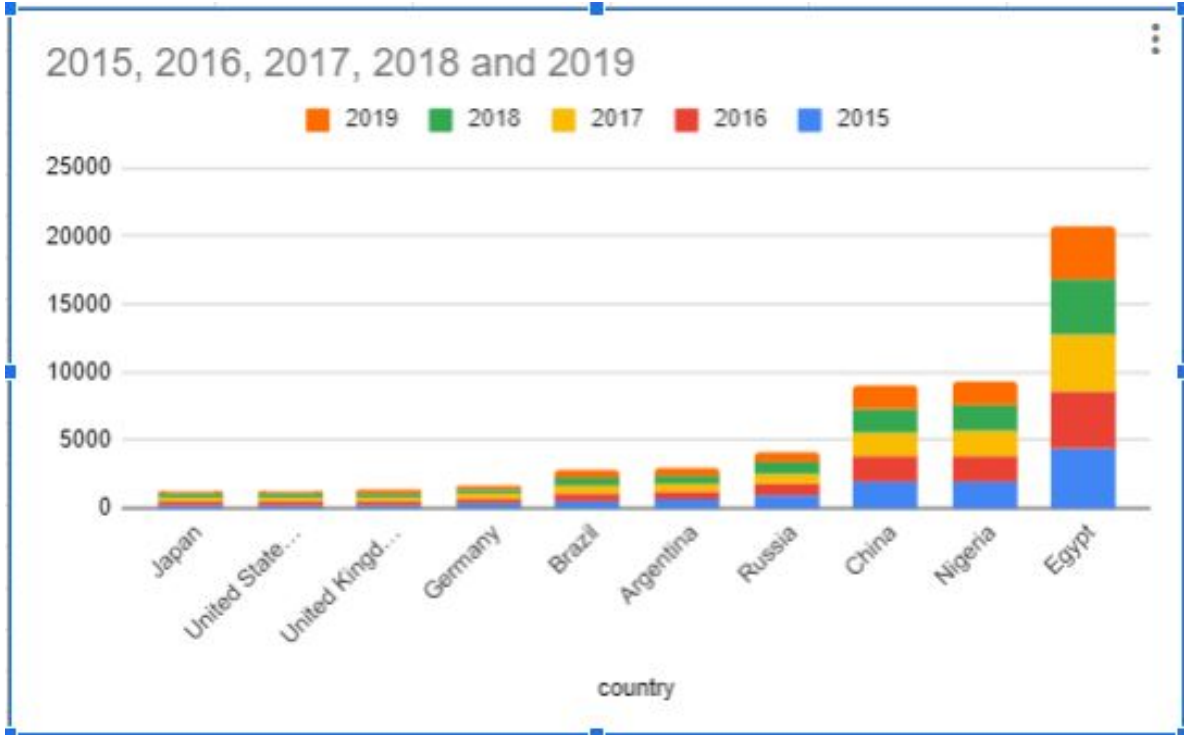


Chart editor

Setup | Customize

Chart type

Stacked column chart

Chart type

Stacking

Standard

Data range

B1:G11

X-axis

country

Aggregate

Series

2015

2016

2017

2018

2019

Add Series

Switch rows / columns

Use row 1 as headers

Use column B as labels

On the right-hand side however, I can modify the presentation and axes.

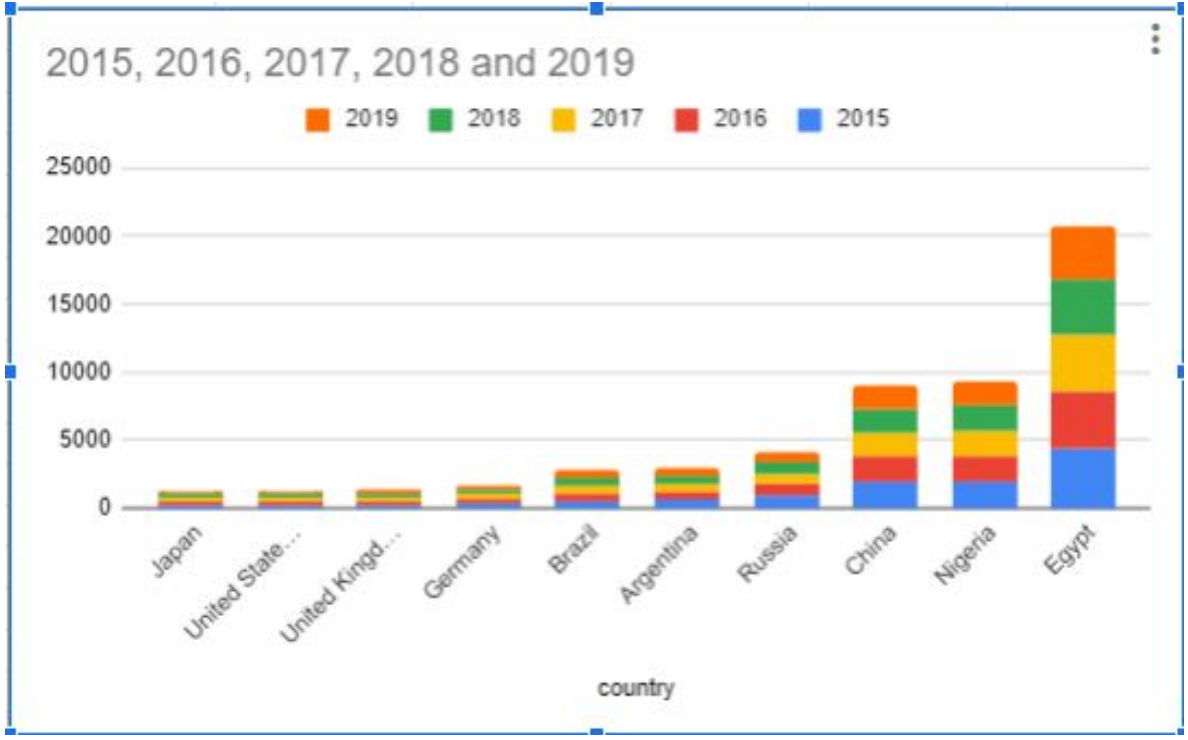


Chart editor

Setup | Customize

Chart type

Stacked column chart

Chart type

Stacking

Standard

Data range

B1:G11

X-axis

country

Aggregate

Series

2015

2016

2017

2018

Add Series

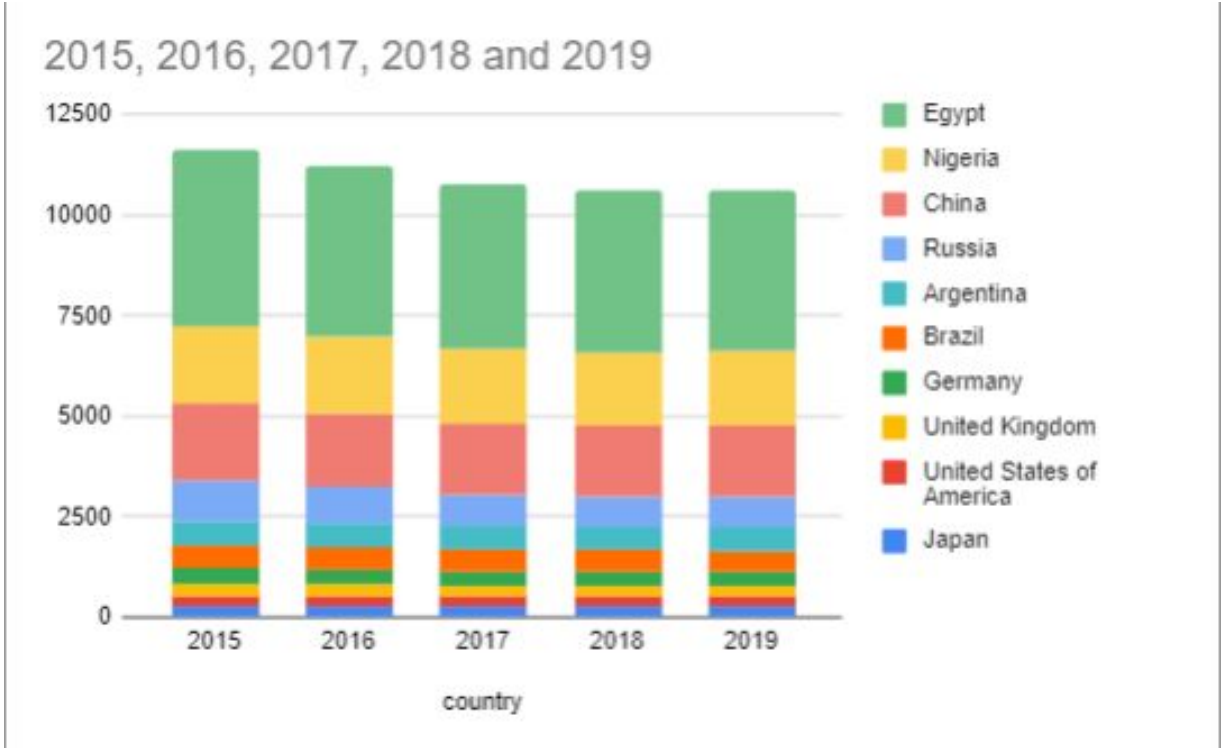
Switch rows / columns

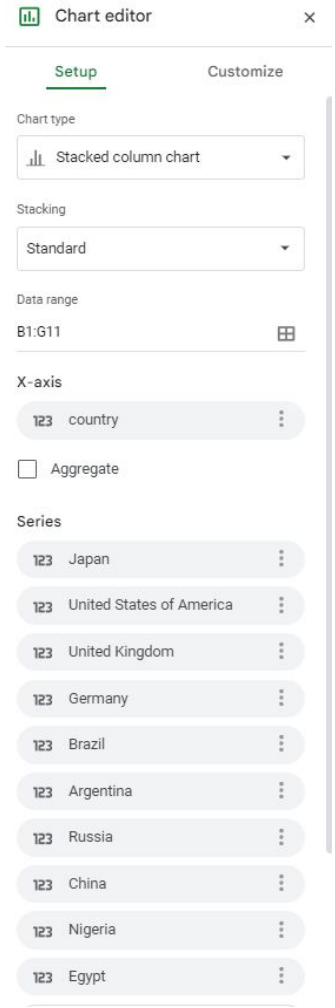
Use row 1 as headers

Use row 1 as headers

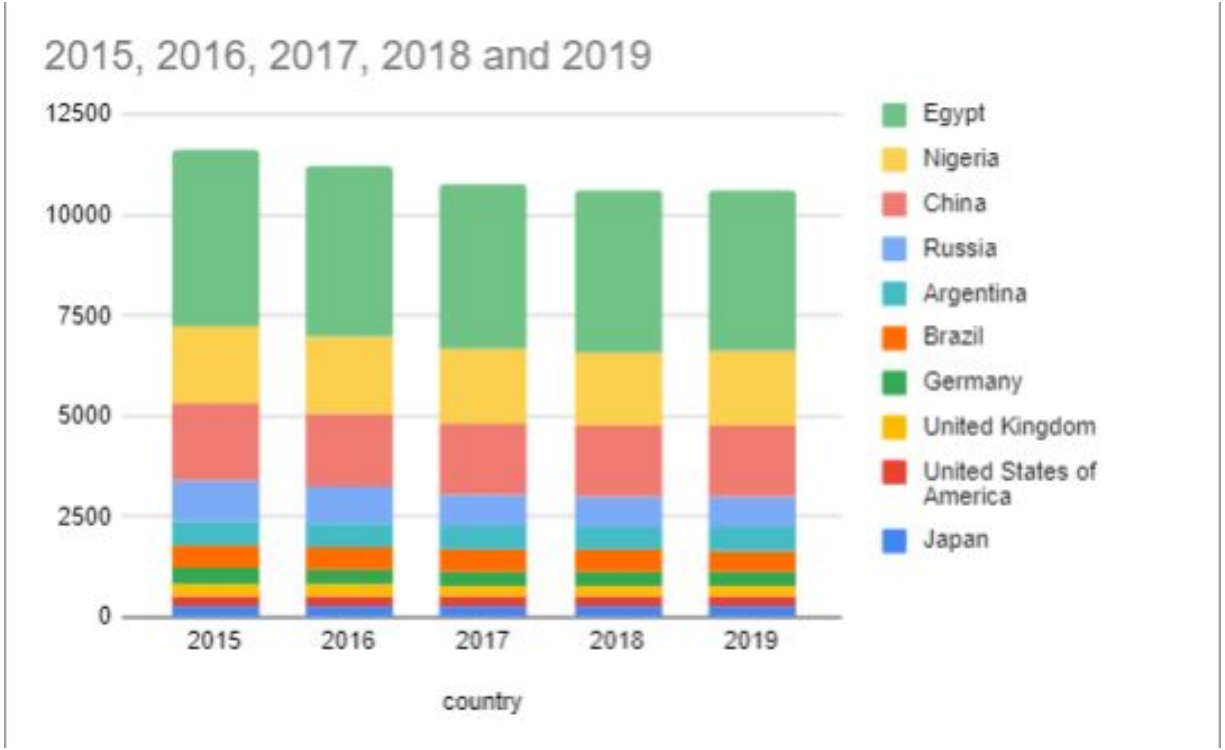
Use column B as labels

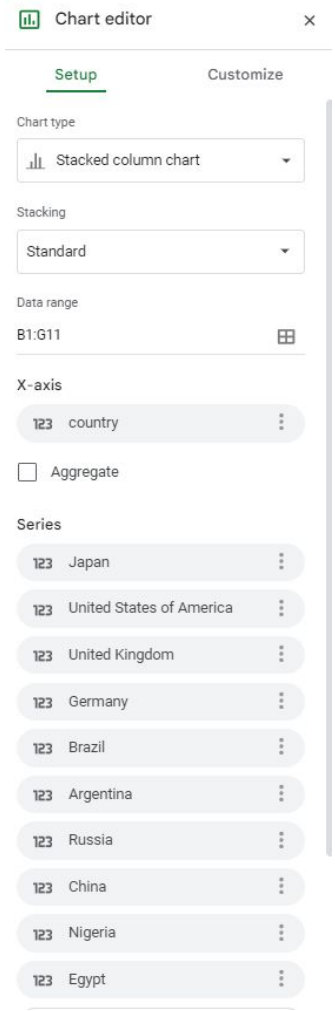
Let's say I want to limit my choice to the US, and plot a line-chart across the years.



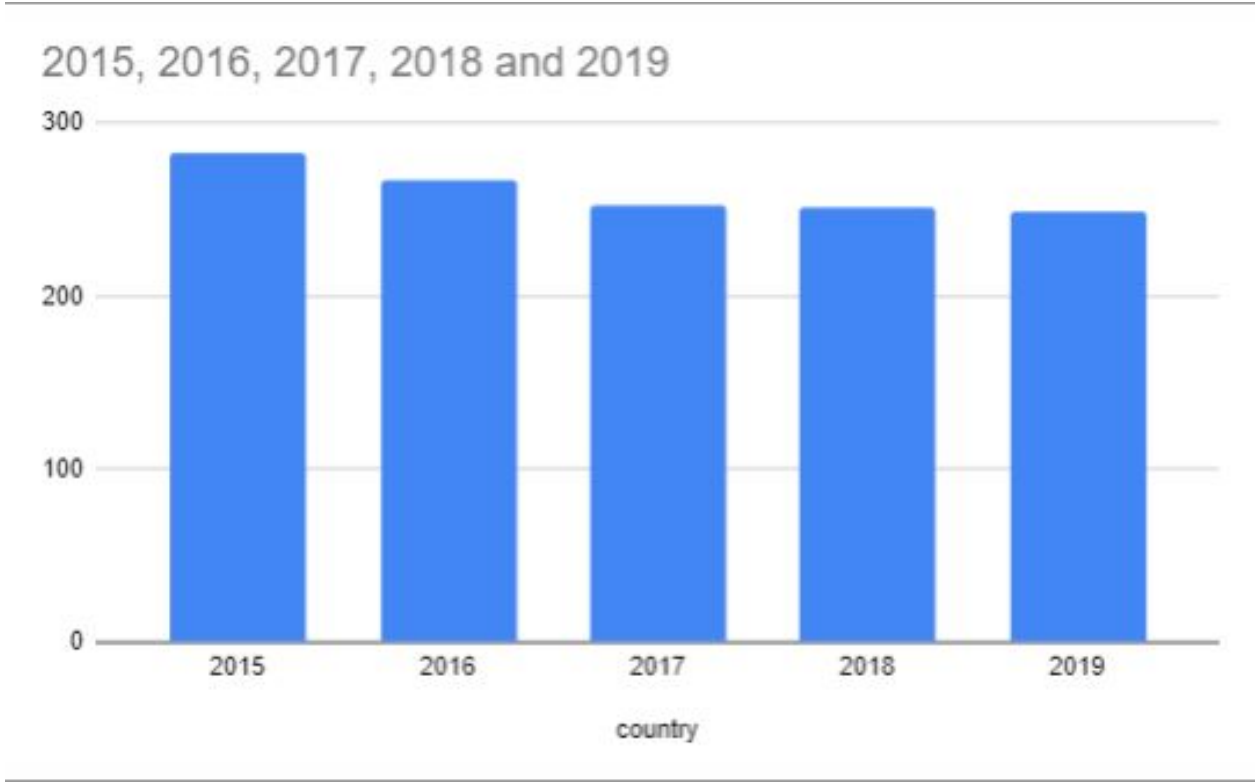


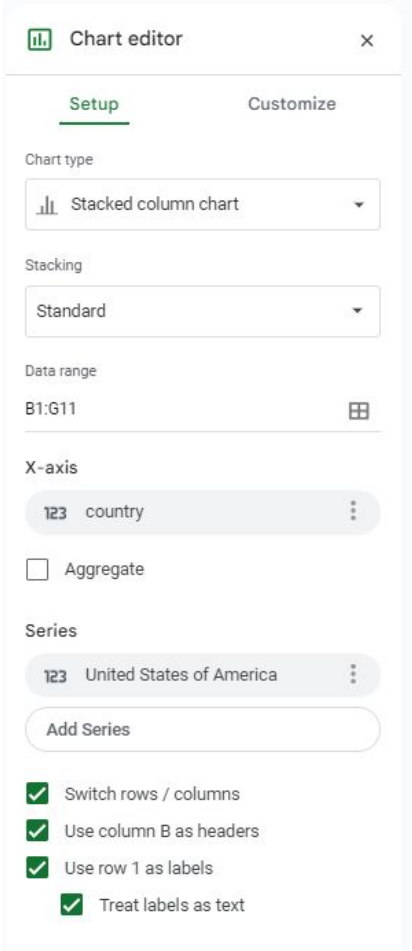
I switch my rows & columns.



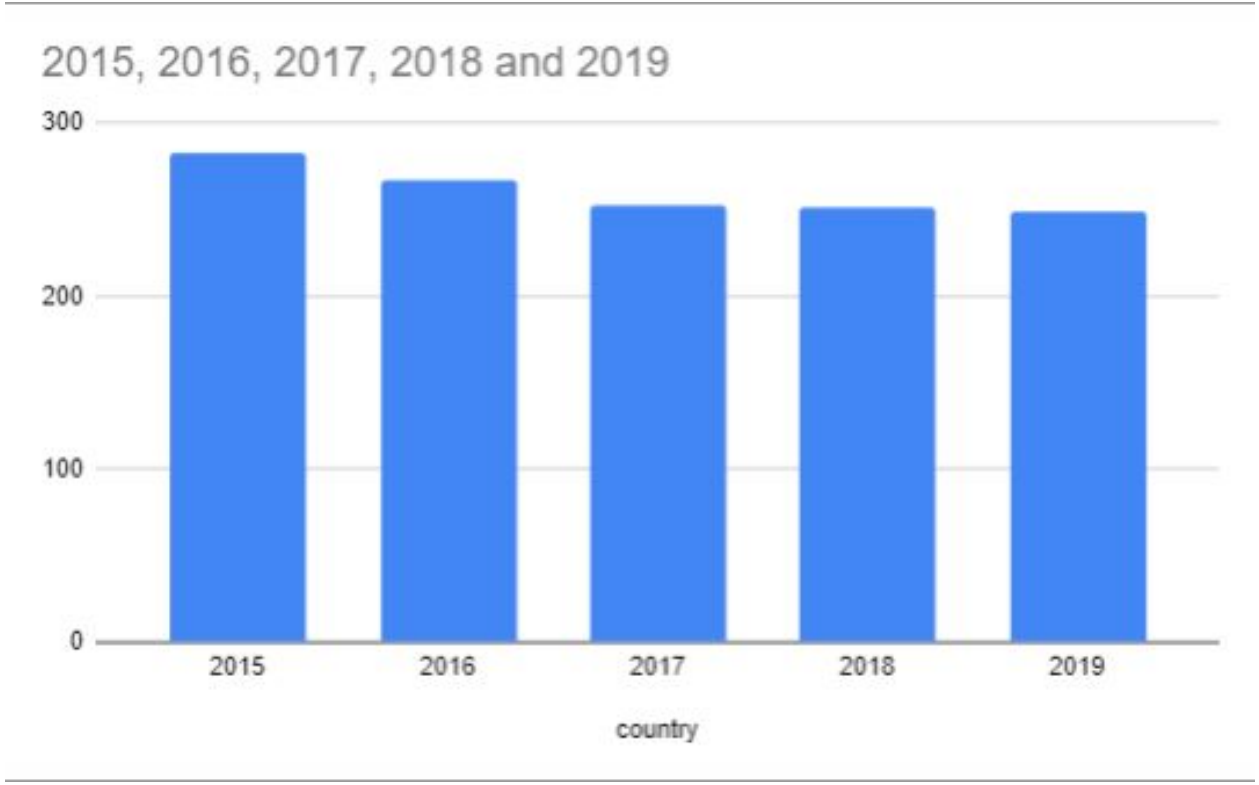


Let’s remove all series except for the US

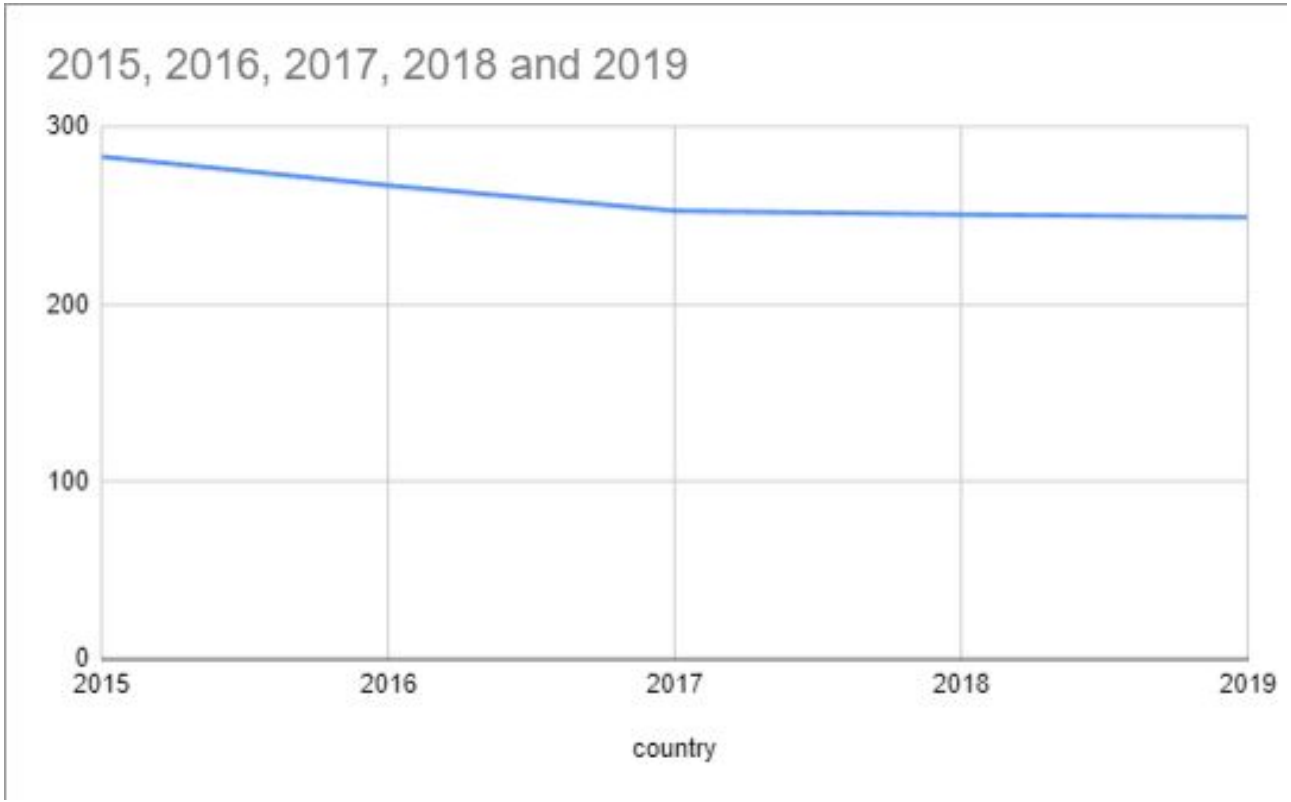


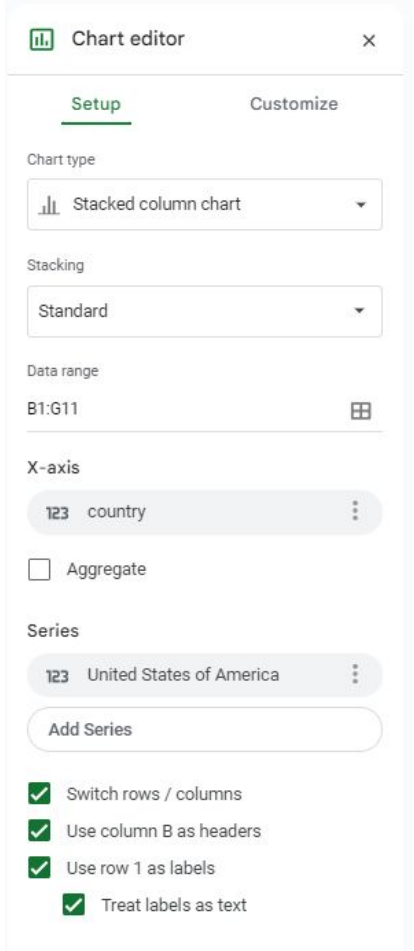


And finally, let's change this to a line plot using the "chart type" menu.

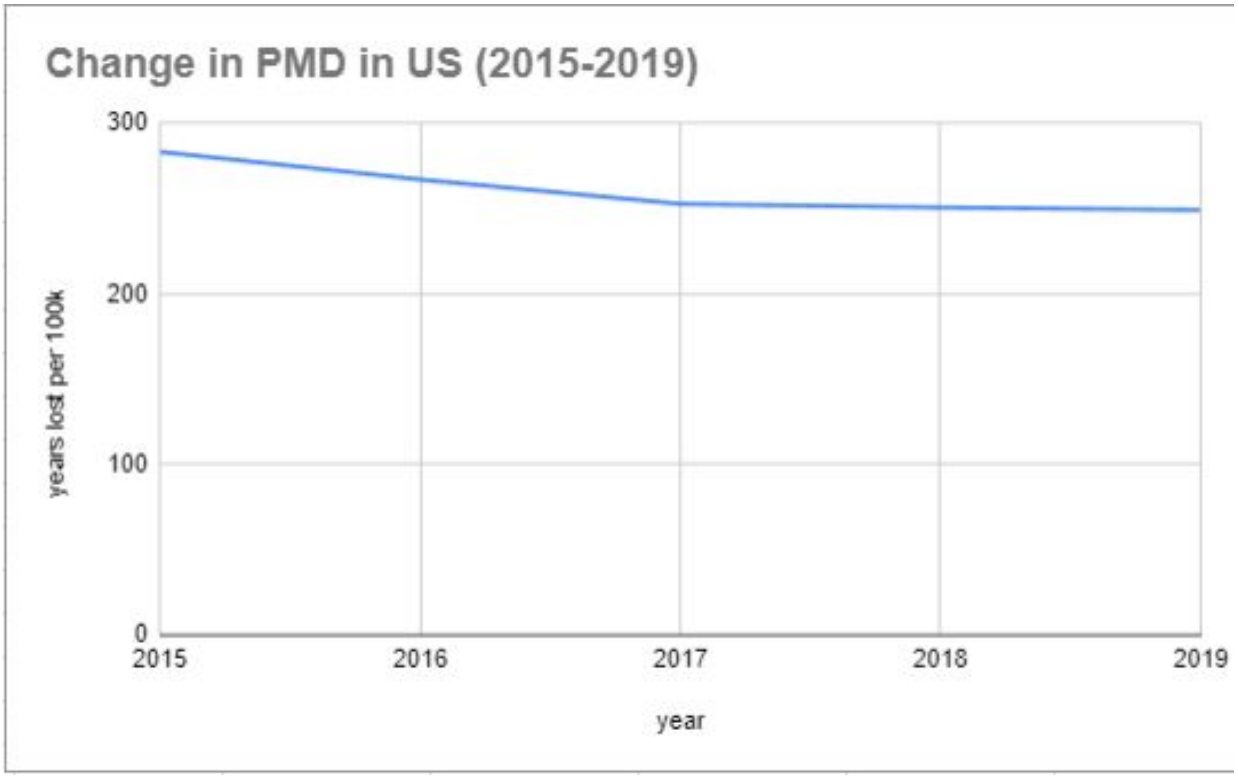


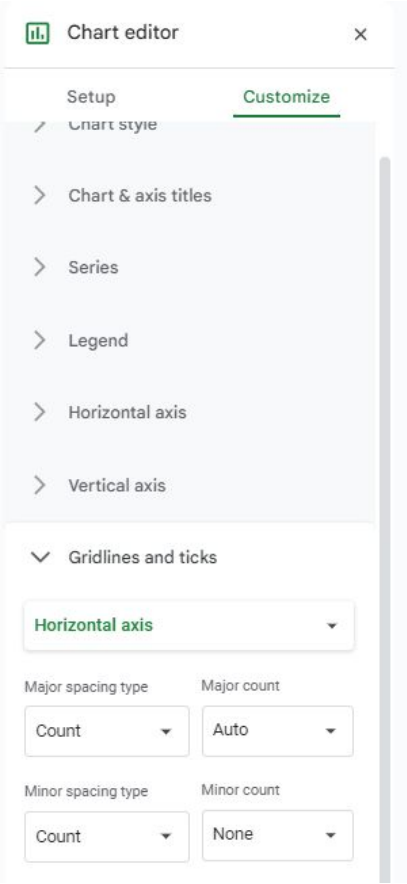
Notice the variety of charts available to us.





And now we have created a (somewhat uninteresting) line chart.
What do you notice about the labels however. Did those change?





We encourage you to experiment and explore features to improve your charts.

Spreadsheets - General Tips

When creating a data visualization, we have a lot of options when it comes to customizing look. A couple of tips for making charts interesting & impactful include:

- *Always **label** axes & titles*
- *Don't hesitate to use **signifiers to point to patterns/events***
  - ***...but still be economical with space***
- ***Large swaths of empty space** usually indicates underutilized space*
- *Don't hesitate to include **text***

Spreadsheets - Functions

Spreadsheets - Formulas

Much like in Python, we use ***functions*** to calculate metrics & implement behavior.

Filter	SORTN	`SORTN(range, [n], [display_ties_mode], [sort_column1, is_ascending1], ...)`	Returns the first n items in a data set after performing a sort. Learn more
Filter	UNIQUE	`UNIQUE(range)`	Returns unique rows in the provided source range, discarding duplicates. Rows are returned in the order in which they first appear in the source range. Learn more
Financial	ACCRINT	`ACCRINT(issue, first_payment, settlement, rate, redemption, frequency, [day_count_convention])`	Calculates the accrued interest of a security that has periodic payments. Learn more

Spreadsheets - Functions/Formulas

We implement functions in Sheets by clicking on a cell and by starting the the “equals” symbol:

=

Spreadsheets - Functions/Formulas

Afterwards, we type in the name of the function we are interested in. Just like in VSCode, there are autocomplete features to assist us here

= AVERAGE()

Spreadsheets - Functions/Formulas

Sometimes however, we will have to specify the range of data ourselves. We type in the **label of the starting cell,** a **colon**, and then **the name of the ending cell**.

= AVERAGE(C2:G2)

Spreadsheets - Functions/Formulas

If something is wrong with our input data, sheets will provide a message that indicating that an error occurred.

#NAME?

#ERROR

country	2015	2016	2017	2018	2019	2019_GDP (billic	stdev
Japan	262.7909262	261.7102781	257.2665629	256.6441396	256.0192158	5,123	
United States of	283.5238203	267.2755296	252.9543121	250.8516978	249.4296641	21,380	
United Kingdom	294.8980096	289.6564515	286.4196291	285.8509262	280.5127785	2,857	
Germany	375.4855158	353.7205444	338.2790574	335.3876598	334.522656	3,888	
Brazil	568.7506785	568.0071932	552.8455853	546.0554702	542.7443866	1,873	
Argentina	624.5758182	620.0178524	596.8603743	585.5216743	580.4380542	447.8	
Russia	975.8511373	871.8861424	763.8980703	754.5885114	764.4997137	1,693	
China	1921.885946	1854.62846	1782.888749	1749.996404	1743.3549	14,280	
Nigeria	1929.854375	1904.940118	1849.284647	1838.035469	1869.209248	448.1	
Egypt	4388.481759	4248.828354	4089.565822	4034.412844	3993.224209	303.1	

I've created a new column to discover the standard deviation of each country's PMD

B	C	D	E	F	G	H	I
country	2015	2016	2017	2018	2019	2019_GDP (billic	stdev
Japan	262.7909262	261.7102781	257.2665629	256.6441396	256.0192158	5,123	=STD
United States of	283.5238203	267.2755296	252.9543121	250.8516978	249.4296641	21,380	
United Kingdom	294.8980096	289.6564515	286.4196291	285.8509262	280.5127785	2,857	
Germany	375.4855158	353.7205444	338.2790574	335.3876598	334.522656	3,888	
Brazil	568.7506785	568.0071932	552.8455853	546.0554702	542.7443866	1,873	
Argentina	624.5758182	620.0178524	596.8603743	585.5216743	580.4380542	447.8	
Russia	975.8511373	871.8861424	763.8980703	754.5885114	764.4997137	1,693	
China	1921.885946	1854.62846	1782.888749	1749.996404	1743.3549	14,280	
Nigeria	1929.854375	1904.940118	1849.284647	1838.035469	1869.209248	448.1	
Egypt	4388.481759	4248.828354	4089.565822	4034.412844	3993.224209	303.1	

STDEV
Standard deviatio
STDEVA
STDEVP
STDEV.P
STDEV.S
STDEVPA
Tab to accept

Notice that by typing in the first few characters of my function, I get an autocomplete menu. “My sheets can’t do that!!!’

2015	2016	2017	2018	2019	2019_GDP (billic	stdev
262.7909262	261.7102781	257.2665629	256.6441396	256.0192158	5,123	=STDEV(
283.5238203	267.2755296	252.9543121	250.8516978	249.4296641	21,380	
294.8980096	289.6564515	286.4196291	285.8509262	280.5127785	2,857	
375.4855158	353.7205444	338.2790574	335.3876598	334.522656	3,888	
568.7506785	568.0071932	552.8455853	546.0554702	542.7443866	1,873	
624.5758182	620.0178524	596.8603743	585.5216743	580.4380542	447.8	
975.8511373	871.8861424	763.8980703	754.5885114	764.4997137	1,693	
1921.885946	1854.62846	1782.888749	1749.996404	1743.3549	14,280	
1929.854375	1904.940118	1849.284647	1838.035469	1869.209248	448.1	
4388.481759	4248.828354	4089.565822	4034.412844	3993.224209	303.1	

Just like in Python, I follow up my function name with parentheses,

2015	2016	2017	2018	2019	2019_GDP (billi	3.126178233
262.7909262	261.7102781	257.2665629	256.6441396	256.0192158	5,123	=STDEV(C2:G2)
283.5238203	267.2755296	252.9543121	250.8516978	249.4296641	21,380	
294.8980096	289.6564515	286.4196291	285.8509262	280.5127785	2,857	
375.4855158	353.7205444	338.2790574	335.3876598	334.522656	3,888	
568.7506785	568.0071932	552.8455853	546.0554702	542.7443866	1,873	
624.5758182	620.0178524	596.8603743	585.5216743	580.4380542	447.8	
975.8511373	871.8861424	763.8980703	754.5885114	764.4997137	1,693	
1921.885946	1854.62846	1782.888749	1749.996404	1743.3549	14,280	
1929.854375	1904.940118	1849.284647	1838.035469	1869.209248	448.1	
4388.481759	4248.828354	4089.565822	4034.412844	3993.224209	303.1	

Just like in Python, I follow up my function name with parentheses, and then I feed in the arguments.

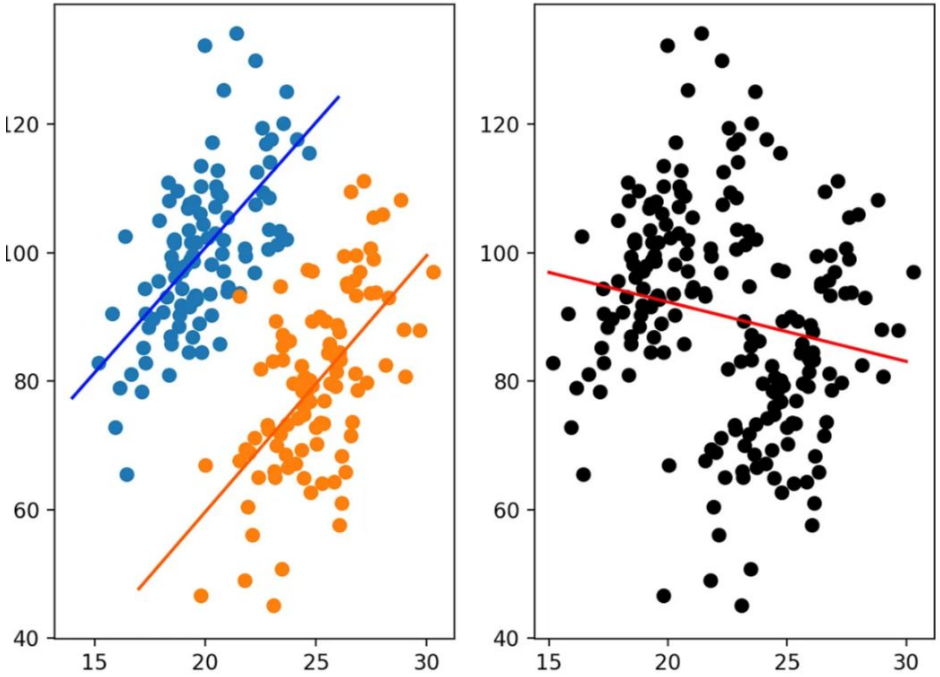
2015	2016	2017	2018	2019	2019_GDP (billic	stdev
262.7909262	261.7102781	257.2665629	256.6441396	256.0192158	5,123	3.126178233
283.5238203	267.2755296	252.9543121	250.8516978	249.4296641	21,380	
294.8980096	289.6564515	286.4196291	285.8509262	280.5127785	2,857	
375.4855158	353.7205444	338.2790574	335.3876598	334.522656	3,888	
568.7506785	568.0071932	552.8455853	546.0554702	542.7443866	1,873	
624.5758182	620.0178524	596.8603743	585.5216743	580.4380542	447.8	
975.8511373	871.8861424	763.8980703	754.5885114	764.4997137	1,693	
1921.885946	1854.62846	1782.888749	1749.996404	1743.3549	14,280	
1929.854375	1904.940118	1849.284647	1838.035469	1869.209248	448.1	
4388.481759	4248.828354	4089.565822	4034.412844	3993.224209	303.1	

I created the calculation, but now my hair-sprayed coworker wants me to fill in the rest of rows!

2015	2016	2017	2018	2019	2019_GDP (billic	stdev
262.7909262	261.7102781	257.2665629	256.6441396	256.0192158	5,123	3.126178233
283.5238203	267.2755296	252.9543121	250.8516978	249.4296641	21,380	14.56118403
294.8980096	289.6564515	286.4196291	285.8509262	280.5127785	2,857	5.293474213
375.4855158	353.7205444	338.2790574	335.3876598	334.522656	3,888	17.47875929
568.7506785	568.0071932	552.8455853	546.0554702	542.7443866	1,873	12.15316729
624.5758182	620.0178524	596.8603743	585.5216743	580.4380542	447.8	19.97399144
975.8511373	871.8861424	763.8980703	754.5885114	764.4997137	1,693	96.56517526
1921.885946	1854.62846	1782.888749	1749.996404	1743.3549	14,280	76.2982106
1929.854375	1904.940118	1849.284647	1838.035469	1869.209248	448.1	38.44901885
4388.481759	4248.828354	4089.565822	4034.412844	3993.224209	303.1	164.4900063

I simply double-click on the cross icon in the corner of my cell, and it auto completes the rest of the rows. Wow! My career is saved.

Spreadsheets - Pivot Tables



Sometimes, calculating metrics in the **full aggregate** hides information that is only revealed when **considering groups**.

Batter \ Year	1995		1996		Combined	
Derek Jeter	12/48	.250	183/582	.314	195/630	**.310**
David Justice	104/411	**.253**	45/140	**.321**	149/551	.270

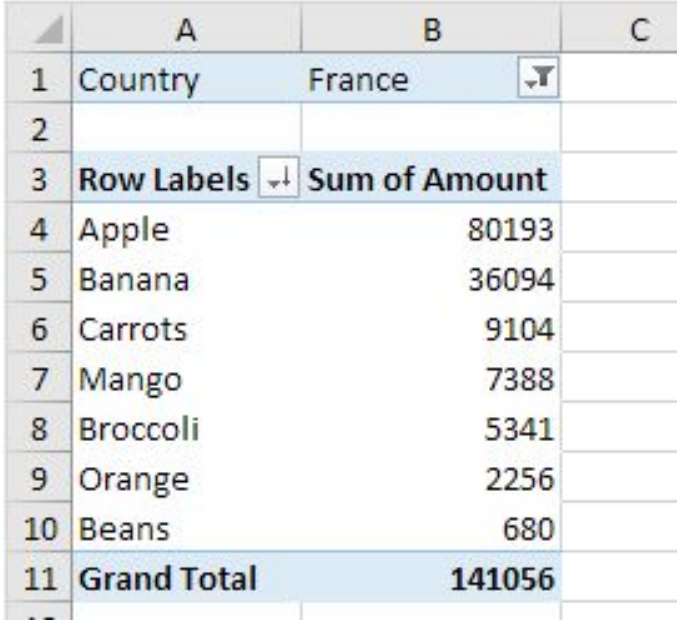
For example, it's **entirely possible** for one baseball player to have a higher batting average than another player, yet have a lower batting average across each year. This occurs due to differences in sample size.

Spreadsheets - Pivot Tables

This is why it's important to be able to calculate metrics **across groups**.

Instead of painstakingly creating the correct formulas to do this, we instead use **pivot tables**.

This allows us to quickly generate aggregate metrics across multiple groups.



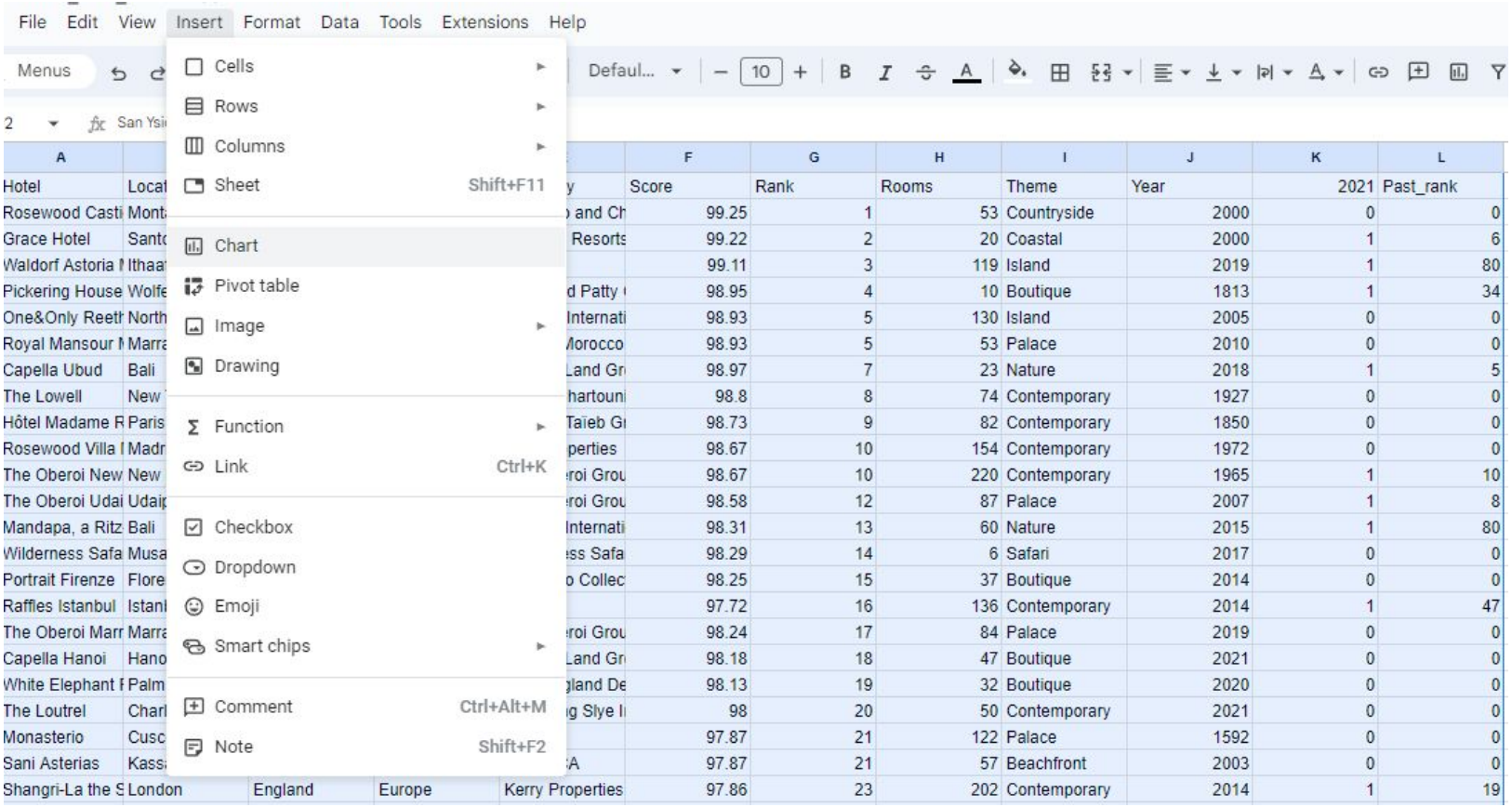
	A	B	C
1	Country	France	
2			
3	**Row Labels**	**Sum of Amount**	
4	Apple	80193	
5	Banana	36094	
6	Carrots	9104	
7	Mango	7388	
8	Broccoli	5341	
9	Orange	2256	
10	Beans	680	
11	**Grand Total**	**141056**	

Hotel	Location	Country	Region	Company	Score	Rank	Rooms	Theme	Year	2021	Past_rank
Rosewood Casti	Montalcino	Italy	Europe	Massimo and Ch	99.25	1	53	Countryside	2000	0	0
Grace Hotel	Santorini	Greece	Europe	Auberge Resorts	99.22	2	20	Coastal	2000	1	6
Waldorf Astoria I	Ithaafushi Island	Maldives	Southeast Asia	Hilton	99.11	3	119	Island	2019	1	80
Pickering House	Wolfeboro	United States	North America	Peter and Patty (	98.95	4	10	Boutique	1813	1	34
One&Only Reeth	North Malé Atoll	Maldives	Southeast Asia	Kerzner Internati	98.93	5	130	Island	2005	0	0
Royal Mansour M	Marrakech	Morocco	Africa	King of Morocco	98.93	5	53	Palace	2010	0	0
Capella Ubud	Bali	Indonesia	Southeast Asia	Pontiac Land Gr	98.97	7	23	Nature	2018	1	5
The Lowell	New York	United States	North America	Fouad Chartouni	98.8	8	74	Contemporary	1927	0	0
Hôtel Madame R	Paris	France	Europe	Laurent Taïeb Gr	98.73	9	82	Contemporary	1850	0	0
Rosewood Villa I	Madrid	Spain	Europe	RLH Properties	98.67	10	154	Contemporary	1972	0	0
The Oberoi New	New Delhi	India	Southeast Asia	The Oberoi Grou	98.67	10	220	Contemporary	1965	1	10
The Oberoi Udai	Udaipur	India	Southeast Asia	The Oberoi Grou	98.58	12	87	Palace	2007	1	8
Mandapa, a Ritz	Bali	Indonesia	Southeast Asia	Marriott Internati	98.31	13	60	Nature	2015	1	80
Wilderness Safa	Musanze	Rwanda	Africa	Wilderness Safa	98.29	14	6	Safari	2017	0	0
Portrait Firenze	Florence	Italy	Europe	Lungarno Collec	98.25	15	37	Boutique	2014	0	0
Raffles Istanbul	Istanbul	Turkey	Middle East	Accor	97.72	16	136	Contemporary	2014	1	47
The Oberoi Marr	Marrakech	Morocco	Africa	The Oberoi Grou	98.24	17	84	Palace	2019	0	0
Capella Hanoi	Hanoi	Vietnam	Southeast Asia	Pontiac Land Gr	98.18	18	47	Boutique	2021	0	0
White Elephant I	Palm Beach	United States	North America	New England De	98.13	19	32	Boutique	2020	0	0
The Loutrel	Charleston	United States	North America	Spaulding Slye I	98	20	50	Contemporary	2021	0	0
Monasterio	Cusco	Peru	Latin America	LVMH	97.87	21	122	Palace	1592	0	0
Sani Asterias	Kassandra	Greece	Europe	LBRI SCA	97.87	21	57	Beachfront	2003	0	0

For this example, we will take a look at the top 100 hotels in the world.

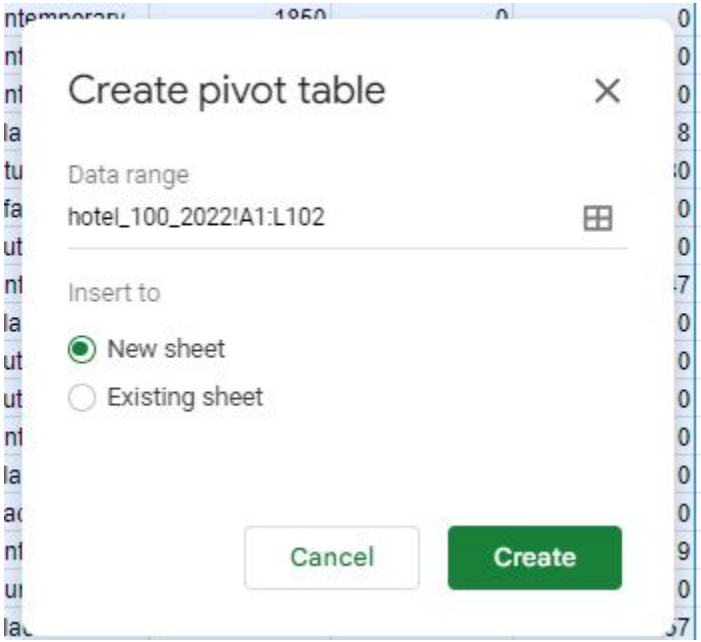
Hotel	Location	Country	Region	Company	Score	Rank	Rooms	Theme	Year	2021	Past_rank
Rosewood Casti	Montalcino	Italy	Europe	Massimo and Ch	99.25	1	53	Countryside	2000	0	0
Grace Hotel	Santorini	Greece	Europe	Auberge Resorts	99.22	2	20	Coastal	2000	1	6
Waldorf Astoria M	Ithaafushi Island	Maldives	Southeast Asia	Hilton	99.11	3	119	Island	2019	1	80
Pickering House	Wolfeboro	United States	North America	Peter and Patty (	98.95	4	10	Boutique	1813	1	34
One&Only Reeth	North Malé Atoll	Maldives	Southeast Asia	Kerzner Internati	98.93	5	130	Island	2005	0	0
Royal Mansour M	Marrakech	Morocco	Africa	King of Morocco	98.93	5	53	Palace	2010	0	0
Capella Ubud	Bali	Indonesia	Southeast Asia	Pontiac Land Gr	98.97	7	23	Nature	2018	1	5
The Lowell	New York	United States	North America	Fouad Chartouni	98.8	8	74	Contemporary	1927	0	0
Hôtel Madame R	Paris	France	Europe	Laurent Taïeb Gr	98.73	9	82	Contemporary	1850	0	0
Rosewood Villa I	Madrid	Spain	Europe	RLH Properties	98.67	10	154	Contemporary	1972	0	0
The Oberoi New	New Delhi	India	Southeast Asia	The Oberoi Grou	98.67	10	220	Contemporary	1965	1	10
The Oberoi Udai	Udaipur	India	Southeast Asia	The Oberoi Grou	98.58	12	87	Palace	2007	1	8
Mandapa, a Ritz	Bali	Indonesia	Southeast Asia	Marriott Internati	98.31	13	60	Nature	2015	1	80
Wilderness Safa	Musanze	Rwanda	Africa	Wilderness Safa	98.29	14	6	Safari	2017	0	0
Portrait Firenze	Florence	Italy	Europe	Lungarno Collec	98.25	15	37	Boutique	2014	0	0
Raffles Istanbul	Istanbul	Turkey	Middle East	Accor	97.72	16	136	Contemporary	2014	1	47
The Oberoi Marr	Marrakech	Morocco	Africa	The Oberoi Grou	98.24	17	84	Palace	2019	0	0
Capella Hanoi	Hanoi	Vietnam	Southeast Asia	Pontiac Land Gr	98.18	18	47	Boutique	2021	0	0
White Elephant I	Palm Beach	United States	North America	New England De	98.13	19	32	Boutique	2020	0	0
The Loutrel	Charleston	United States	North America	Spaulding Slye I	98	20	50	Contemporary	2021	0	0
Monasterio	Cusco	Peru	Latin America	LVMH	97.87	21	122	Palace	1592	0	0
Sani Asterias	Kassandra	Greece	Europe	LBRI SCA	97.87	21	57	Beachfront	2003	0	0
Shangri-La the S	London	England	Europe	Kerry Properties	97.86	23	202	Contemporary	2014	1	19
Coquillade Prove	Gargas	France	Europe	Coquillade	97.82	24	63	Countryside	2008	0	0
Taj Palace	New Delhi	India	Southeast Asia	India Hotels Con	97.6	25	403	Palace	1903	1	57
Pendry Chicago	Chicago	United States	North America	Montage Hotels	97.6	25	364	Contemporary	1927	0	0
Nayara Tented C	La Fortuna	Costa Rica	Latin America	Nayara Resorts	97.55	27	29	Nature	2019	1	2

By hitting CTRL + A (or CMD + A), we can select the entire table.

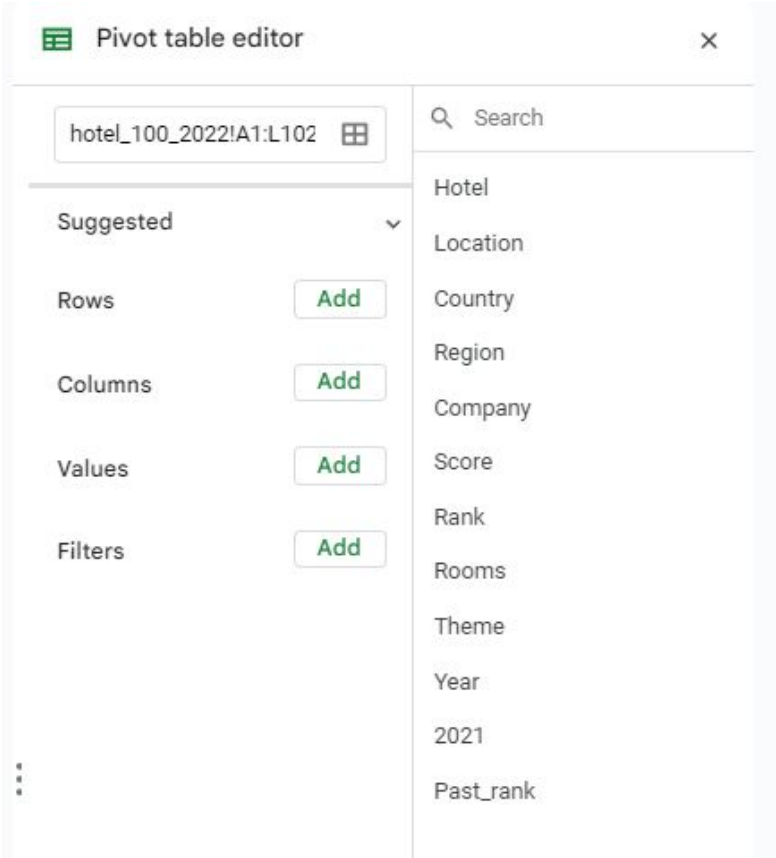


A					F	G	H	I	J	K	L
Hotel	Locat			y	Score	Rank	Rooms	Theme	Year	2021	Past_rank
Rosewood Casti	Monta			and Ch	99.25	1	53	Countryside	2000	0	0
Grace Hotel	Santo			Resorts	99.22	2	20	Coastal	2000	1	6
Waldorf Astoria I	Ithaa				99.11	3	119	Island	2019	1	80
Pickering House	Wolfe			d Patty	98.95	4	10	Boutique	1813	1	34
One&Only Reeth	North			Internati	98.93	5	130	Island	2005	0	0
Royal Mansour M	Marra			Morocco	98.93	5	53	Palace	2010	0	0
Capella Ubud	Bali			Land Gr	98.97	7	23	Nature	2018	1	5
The Lowell	New			hartouni	98.8	8	74	Contemporary	1927	0	0
Hôtel Madame R	Paris			Taïeb Gr	98.73	9	82	Contemporary	1850	0	0
Rosewood Villa I	Madr			perties	98.67	10	154	Contemporary	1972	0	0
The Oberoi New	New			roi Grou	98.67	10	220	Contemporary	1965	1	10
The Oberoi Udai	Udaip			roi Grou	98.58	12	87	Palace	2007	1	8
Mandapa, a Ritz	Bali			Internati	98.31	13	60	Nature	2015	1	80
Wilderness Safa	Musa			ss Safa	98.29	14	6	Safari	2017	0	0
Portrait Firenze	Flore			o Collec	98.25	15	37	Boutique	2014	0	0
Raffles Istanbul	Istanl				97.72	16	136	Contemporary	2014	1	47
The Oberoi Marr	Marra			roi Grou	98.24	17	84	Palace	2019	0	0
Capella Hanoi	Hano			Land Gr	98.18	18	47	Boutique	2021	0	0
White Elephant I	Palm			gland De	98.13	19	32	Boutique	2020	0	0
The Loutrel	Charl			g Slye I	98	20	50	Contemporary	2021	0	0
Monasterio	Cusc				97.87	21	122	Palace	1592	0	0
Sani Asterias	Kass			A	97.87	21	57	Beachfront	2003	0	0
Shangri-La the S	London	England	Europe	Kerry Properties	97.86	23	202	Contemporary	2014	1	19

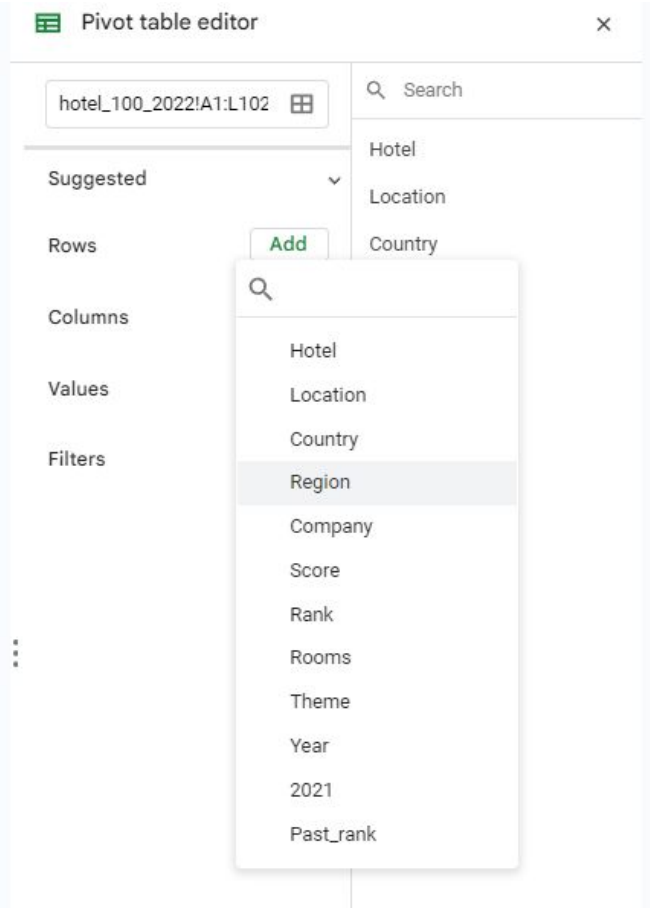
By hitting CTRL + A (or CMD + A), we can select the entire table.



A modal asking us if we want to create a new sheet will appear. You most likely want to say yes.

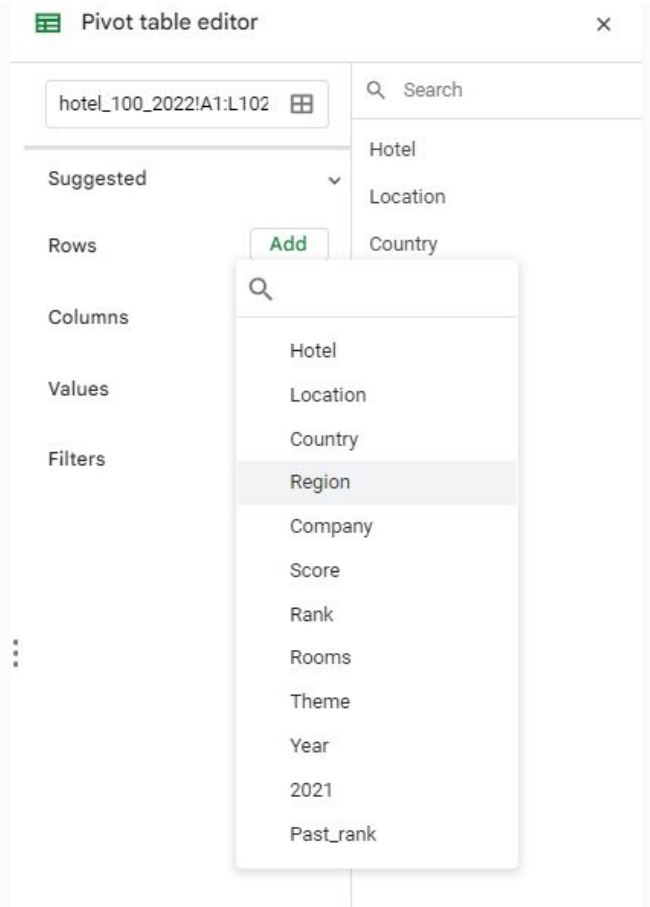


We are given an empty pivot table and an editor.



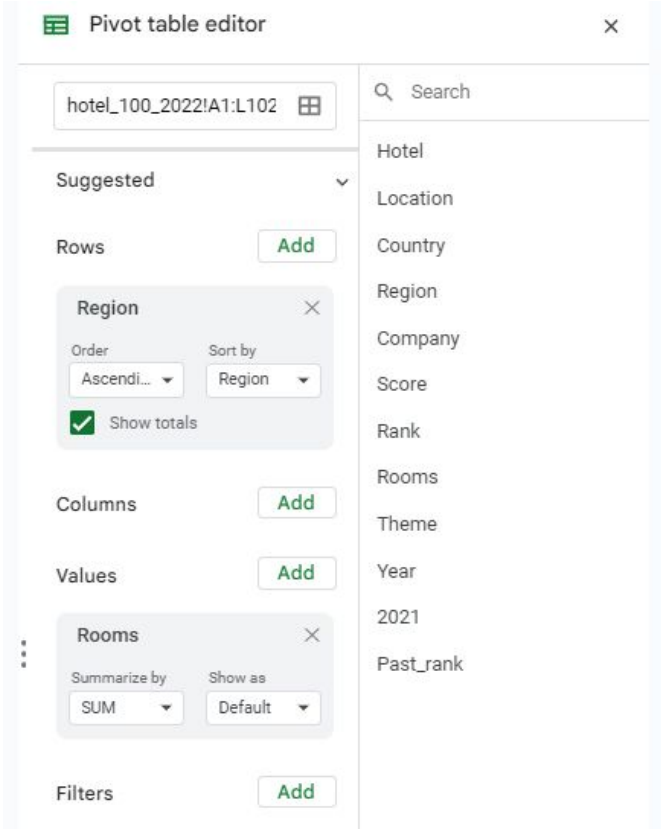
Let’s say we want to get the median number of rooms for all hotels grouped by region.

Region
Africa
Asia
Caribbean
Europe
Latin America
Middle East
North America
Oceania
Southeast Asia
Grand Total



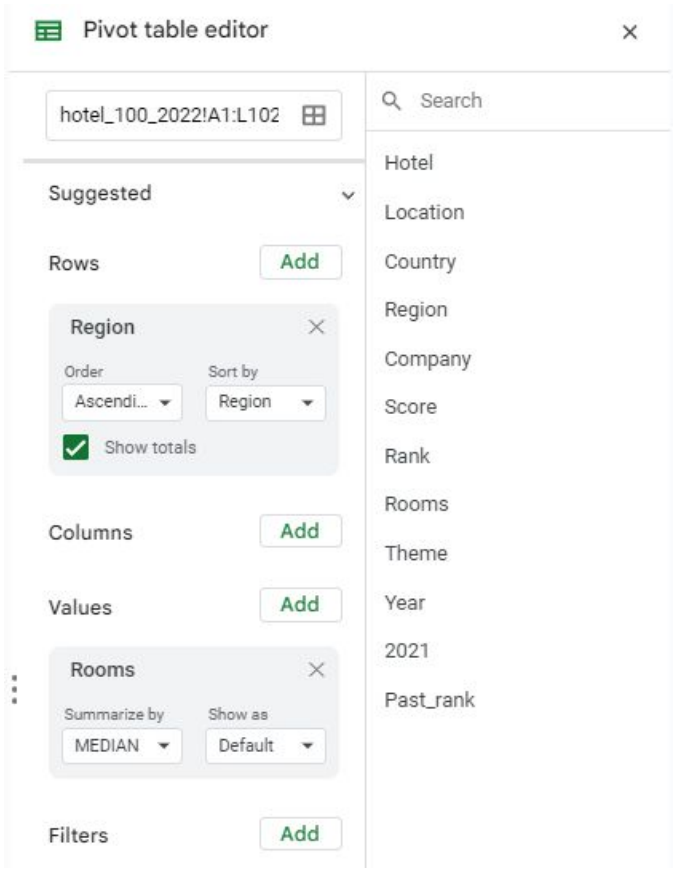
We'll first specify our row to be region.

Region	SUM of Rooms
Africa	438
Asia	123
Caribbean	386
Europe	2052
Latin America	596
Middle East	322
North America	1779
Oceania	19
Southeast Asia	3835
Grand Total	**9550**



And next we will specify our values to be rooms. Notice how we get “sum” by default.

Region	MEDIAN of Roo
Africa	29
Asia	123
Caribbean	72
Europe	66
Latin America	27.5
Middle East	161
North America	50
Oceania	19
Southeast Asia	83
Grand Total	**63**



By clicking on the dropdown menu, we can select “median” instead.

Spreadsheets - Limitations

Limitations to Spreadsheets

As we can see, Google Sheets is a powerful tool. However you might notice some **downsides** to using this in a technical workspace.

Can you think of any ways that Sheets is *not* as good as Python?

- ...

Limitations to Spreadsheets

As we can see, Google Sheets is a powerful tool. However you might notice some **downsides** to using this in a technical workspace.

Can you think of any ways that Sheets is *not* as good as Python?

- **Manual** data transformations
- Running data transformations on a **schedule** is clunky
- Will **not** prevent disastrous errors
- ~~Telling people you use Google Sheets is not as cool as telling people you use Python~~

iso	country	$2,015.00	7/8/1905	2017	2018	2019
JPN	Japan	$262.79	9/17/1900	257.2665629	256.6441396	256.0192158
USA	United States of America	$283.52	9/23/1900	252.9543121	250.8516978	249.4296641
GBR	United Kingdom	$294.90	10/15/1900	286.4196291	285.8509262	280.5127785
DEU	Germany	$375.49	12/18/1900	338.2790574	335.3876598	334.522656
BRA	Brazil	$568.75	7/21/1901	552.8455853	546.0554702	542.7443866
ARG	Argentina	$624.58	9/11/1901	596.8603743	585.5216743	580.4380542
RUS	Russia	$975.85	5/20/1902	763.8980703	754.5885114	764.4997137
CHN	China	$1,921.89	1/27/1905	1782.888749	1749.996404	1743.3549
NGA	Nigeria	$1,929.85	3/18/1905	1849.284647	1838.035469	1869.209248
EGY	Egypt	$4,388.48	8/18/1911	4089.565822	4034.412844	3993.224209

Notice how Sheets does not attempt to stop me from doing something wrong. It assumes I know exactly what I want.

SCIENCE / TECH / MICROSOFT

Scientists rename human genes to stop Microsoft Excel from misreading them as dates

Illustration by Alex Castro / The Verge

/ Sometimes it's easier to rewrite genetics than update Excel

By James Vincent, a senior reporter who has covered AI, robotics, and more for eight years at The Verge.

Aug 6, 2020, 8:44 AM EDT | 0 Comments / 0 New

If you buy something from a Verge link, Vox Media may earn a commission. See our ethics statement.

Sometimes automated solutions bring about more problems than they solve: https://www.theverge.com/2020/8/6/21355674/human-genes-rename-microsoft-excel-misreading-dates

Excel: Why using Microsoft's tool caused Covid-19 results to be lost

5 October 2020

Choosing the right tool is important!
[https://www.bbc.com/news/technology-54423988]

Limitations to Spreadsheets

For all these reasons, we will explore **pandas** in tomorrow's lecture.

However, remember that Sheets still has its place in the data analysts toolkit.

Spreadsheets Lab

Google Sheets Lab

Open the **README.md** file inside of the “spreadsheets_lab” folder and begin by reading the instructions!

Complete this lab in your groups!

Take the last 10 minutes of lecture to revisit this lab together.

This could be you if you keep up the practice!

Tuesday

Tuesday will entail:

- Data analysis in pandas!
- Data manipulation in pandas!
- Data selection in pandas!

Pandas: SettingWithCopy Warning

If you understand what you're doing, you're not learning anything. - Anonymous