DataDive Week 2:

Data Collection & Understanding

SheffDataSoc

Data Science Society

DataSoc Promo!

Welcome! If you aren't already a member of DataSoc, consider joining!

https://su.sheffield.ac.uk/activities/view/data-science-society

We have some very interesting socials coming up soon, including ice-skating (25th Feb)!

Find out more about our socials/events from our social media, and by signing up to the newsletter!

In Week 4, there will be a **career panel** session before the Data Dive (**4th March, 3-4pm, Diamond LT9**), with visiting speakers from industry. This will be a really insightful event and we encourage everyone to attend!

Session Outline

- Where do we get data from?
- Credibility Check
- Web-Scraping
- Exploring the data
- Creating Visualisations

If you want to follow along with the session, or refer back to these resources, head to the Data Dive GitHub repository:

https://github.com/sheffdatasoc/datadive-resources

Where can we get data from?

Kaggle

Online platform providing different kinds of datasets and resource to aid data science learning

https://www.kaggle.com/datasets

Hugging Face

Open platform for AI models, datasets, and ML tools.

https://huggingface.co/datasets

ONS

Open datasets from the Office for National Statistics about the UK (equivalents for other countries may be available)

https://www.ons.gov.uk/

DISCLAIMER

With many of these online platforms open to all collaborators, we have to look at **HOW** the data is collected and **WHO** collected it for **WHAT** purpose.

Credibility Check

- **Source**: Who produced/collected the data?

  Reputation of the source
- **Purpose**: Why was the data collected?

  Ex: Research, marketing, etc.
- **Collection Method**: How was the data collected?

  Ex: Survey, Web-scraping

Some sources (such as the ONS) are likely to be more reliable and accurate than sources like Kaggle, however they may not have the data you are looking for.

Judges may want to ask about your data sources in your project!

Example: Kaggle Dataset

https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data

- **Who?**

  'Inside Airbnb' collected data from Airbnb's website

- **Why?**

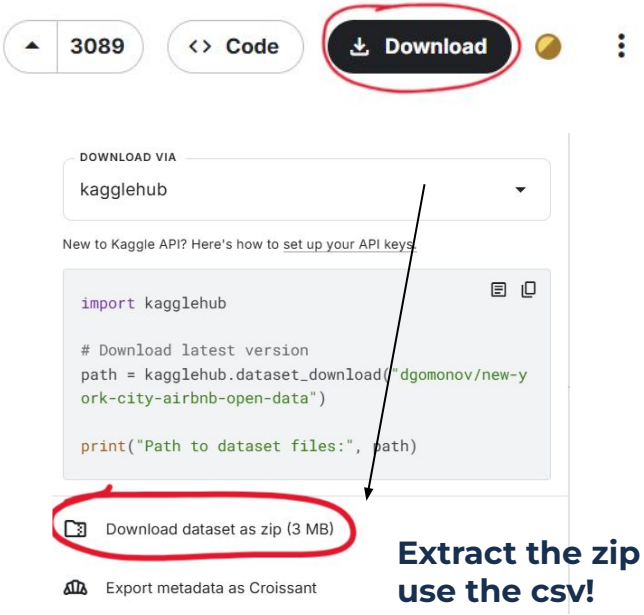
  Advocacy and education about how Airbnbs are impacting residential communities

- **How?**

  Web scraping publicly available information directly from the Airbnb website

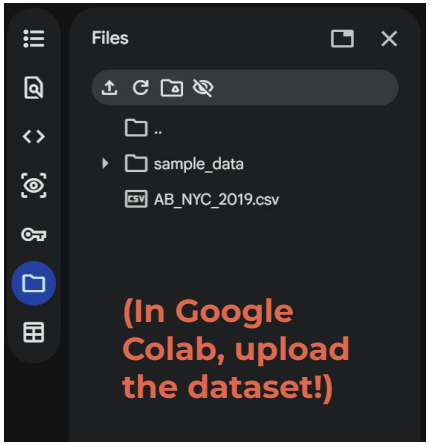
Example: Loading The Data

1. Download the data into csv format from Kaggle



Extract the zip file to use the csv!

2. Load the data in Python (likely a jupyter notebook or Google Colab, using your own blank notebook or the template provided):




```
import pandas as pd
df = pd.read_csv('AB_NYC_2019.csv')
```

Example: Exploring The Data

- How many rows and columns does the dataset have?

```
df.shape
```

```
(48895, 16)
```

- What are the columns in the dataset?

```
df.columns
```

```
Index(['id', 'name', 'host_id', 'host_name', 'neighbourhood_group',
       'neighbourhood', 'latitude', 'longitude', 'room_type', 'price',
       'minimum_nights', 'number_of_reviews', 'last_review',
       'reviews_per_month', 'calculated_host_listings_count',
       'availability_365'],
      dtype='object')
```

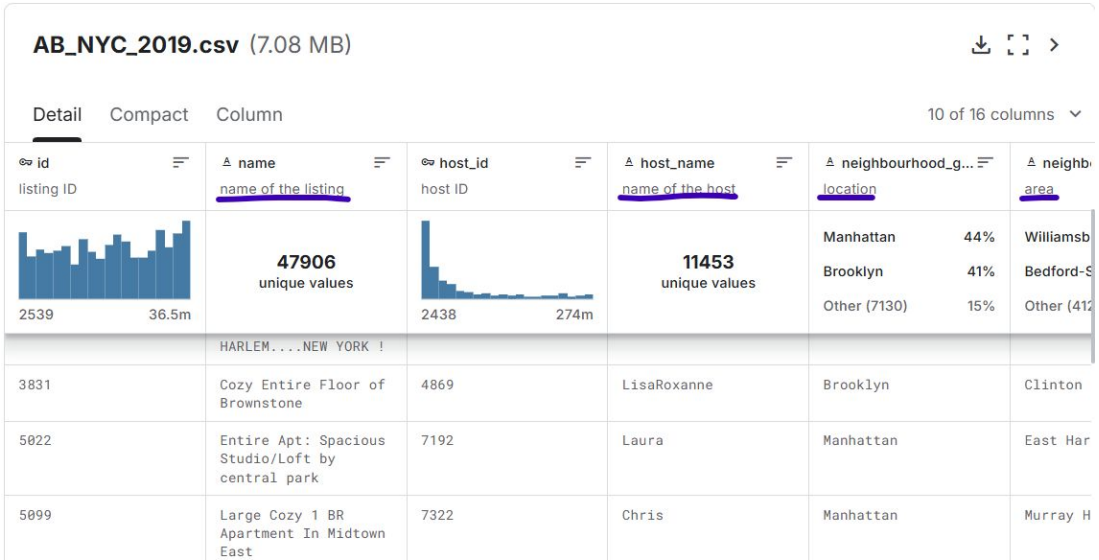
- What does the dataset look like?

```
df.head()
```

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	n
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	

Example: Exploring The Data

- What are the variables and their types? (You can use pandas, or you may be able to refer to Kaggle/the original source for information on the dataset and its features)



AB_NYC_2019.csv (7.08 MB)

Detail | Compact | Column | 10 of 16 columns

id listing ID	name name of the listing	host_id host ID	host_name name of the host	neighbourhood_g... location	neighb area
2539 — 36.5m	47906 unique values	2438 — 274m	11453 unique values	Manhattan 44% Brooklyn 41% Other (7130) 15%	Williamsb Bedford-S Other (412
	HARLEM....NEW YORK !				
3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton
5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Har
5099	Large Cozy 1 BR Apartment In Midtown East	7322	Chris	Manhattan	Murray H

```
df.dtypes
✓ 0.0s
```

```
id                                  int64
name                               object
host_id                             int64
host_name                          object
neighbourhood_group                object
neighbourhood                      object
latitude                          float64
longitude                         float64
room_type                          object
price                               int64
minimum_nights                      int64
number_of_reviews                   int64
last_review                        object
reviews_per_month                 float64
calculated_host_listings_count      int64
availability_365                    int64
dtype: object
```

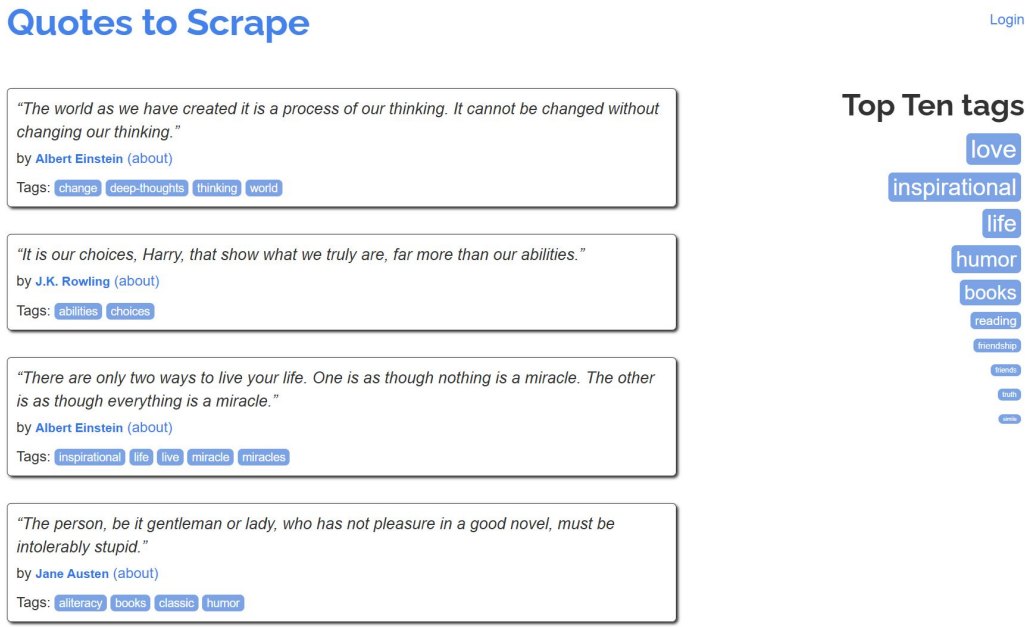
Spend some more time exploring the dataset! What other insights can you find?

Web-Scraping

Another method of gathering data, by ‘scraping’ it from available webpages. This can be useful if a page displays publicly available data in a table (such as Wikipedia).

However, there are risks and ethical considerations, just like with the previous data sources mentioned!

Perform a credibility check.



Example: Web-Scraping

Bs4_tutorial.py > ...

```python
from bs4 import BeautifulSoup
import requests
import csv

url = "https://quotes.toscrape.com"
response = requests.get(url)
response.raise_for_status()

soup = BeautifulSoup(response.text, "html.parser")
quotes = soup.find_all("span", class_="text")
authors = soup.find_all("small", class_="author")

with open("bs4_scraped_quotes.csv", "w", encoding="utf-8-sig", newline="") as file:
    writer = csv.writer(file)
    writer.writerow(["Quote", "Author"])

    for quote, author in zip(quotes, authors):
        print(f"{quote.text} - {author.text}")
        writer.writerow([quote.text, author.text])
```

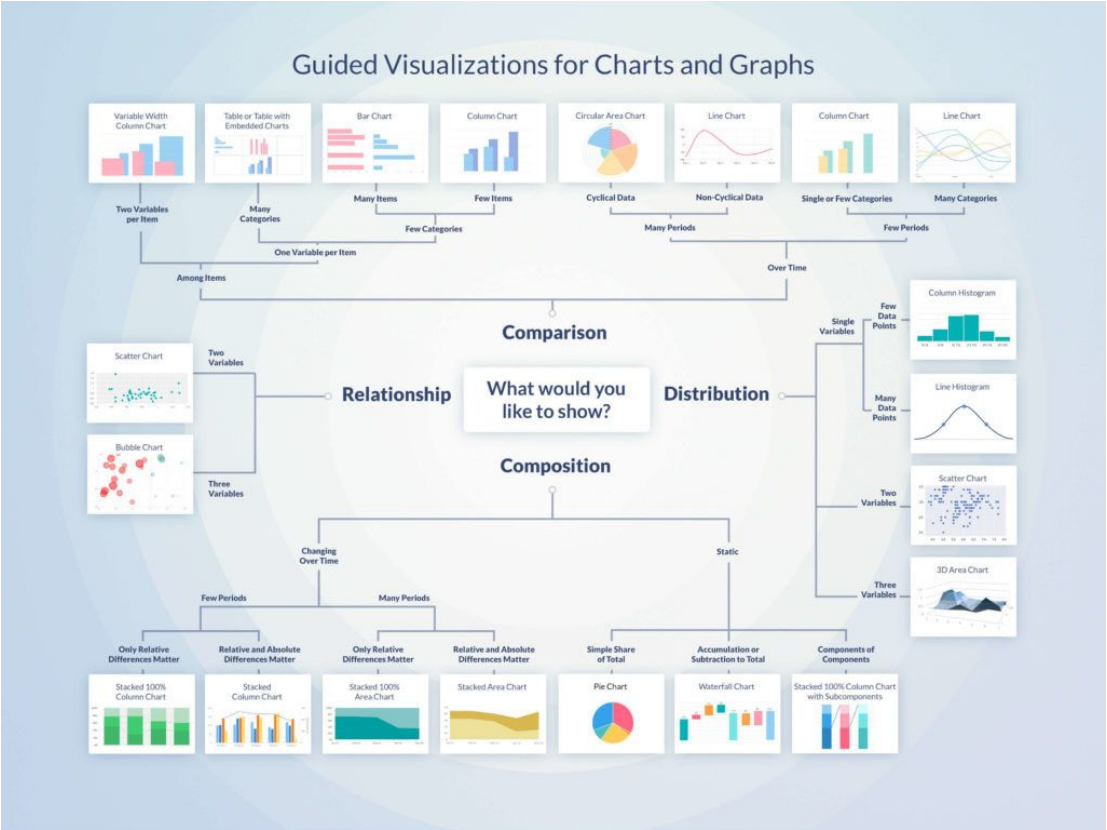
If you have generated a csv from web-scraping, you can load it like any other dataset!

	Quote	Author
0	“The world as we have created it is a process ...	Albert Einstein
1	“It is our choices, Harry, that show what we t...	J.K. Rowling
2	“There are only two ways to live your life. On...	Albert Einstein
3	“The person, be it gentleman or lady, who has ...	Jane Austen
4	“Imperfection is beauty, madness is genius and...	Marilyn Monroe
5	“Try not to become a man of success. Rather be...	Albert Einstein
6	“It is better to be hated for what you are tha...	André Gide
7	“I have not failed. I've just found 10,000 way...	Thomas A. Edison
8	“A woman is like a tea bag; you never know how...	Eleanor Roosevelt
9	“A day without sunshine is like, you know, nig...	Steve Martin

Visualisations

The types of visualisation you create will depend on the types of variables you want to show. Examples include:

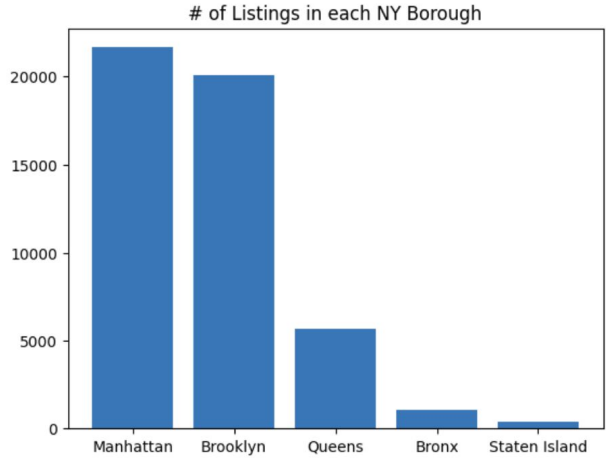
- **Categorical variables**: bar/column charts, pie charts
- **Single numerical variables**: histograms, boxplots
- **Relationships between numerical variables**: line charts, scatter plots



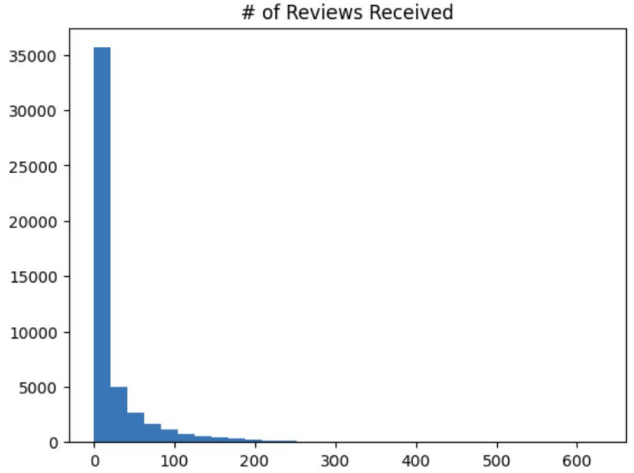
source: https://www.tapclicks.com/blog/data-visualization-types

Example: Visualisations in Python (using matplotlib)

```
plt.bar(x='neighbourhood_group', height='count', data=borough_frequencies)
plt.title('# of Listings in each NY Borough')
plt.show()
```




```
plt.hist(x='number_of_reviews', data=df, bins=30)
plt.title('# of Reviews Received')
plt.show()
```



Can you create some more exciting visualisations?

note: in the real project, it may be best to create visualisations *after* data cleaning (covered next week!)

What comes next?

The remaining DataDive ‘education’ sessions will be →

Week	Session	Where/When
3	Data Cleaning & Pre-processing	**Wednesday, 4-6pm Diamond LT9**
4	Creating ML Models	
5	Evaluating & Interpreting ML Results (& Team Formation)	

Datasoc will also be hosting a **career panel** in Week 4 (before the Data Dive session) with speakers from industry giving insights into their careers. Please come along!

If you think you might want to be a *Project Lead* for a Data Dive project, applications for this will open in Week 3!

Interested in participating the Data Dive? Let us know below!

https://tinyurl.com/sheffdatadive

Find the resources from today:

https://github.com/sheffdatasoc/datadive-resources

Not a member of the society yet? Join us!

https://su.sheffield.ac.uk/activities/view/data-science-society

Data Science Society