CSC461

INTRODUCTION TO DATA SCIENCE

Dr. Muhammad Sharjeel

https://muhmmadsharjeel.github.io/

04

DATA VISUALIZATION

DATA VISUALIZATION

There's a story behind numbers, visualizing data brings them to life

- Data visualization is the method to present the data in a pictorial or graphical format
  - To effectively and accurately represent information about the data
- Graphical format allows to identify new trends and patterns in the data easily
- Gives you answers to questions you didn't know you had

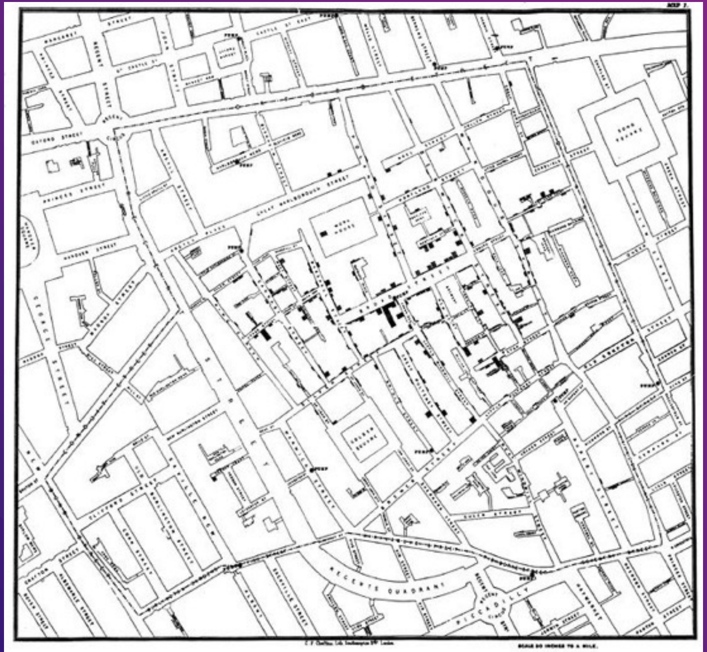
DATA VISUALIZATION

- Some of the main benefits of data visualization
  - Simplifies the complex quantitative information
  - Helps analyze and explore big data easily
  - Identifies the areas that need attention or improvement
  - Identifies the relationship between data points and variables
  - Explores new patterns and reveals hidden patterns in the data

- Goals of data visualization
  - Record
  - Analyze
  - Communicate

WHY VISUALIZE DATA

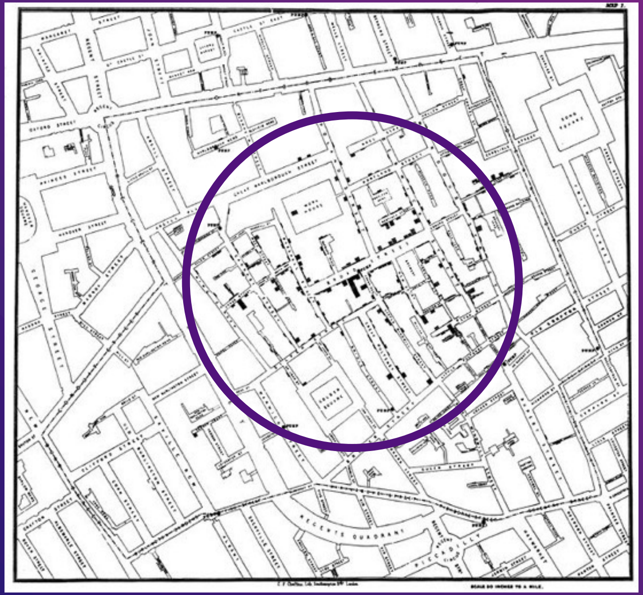
- 1854 London Cholera Epidemic



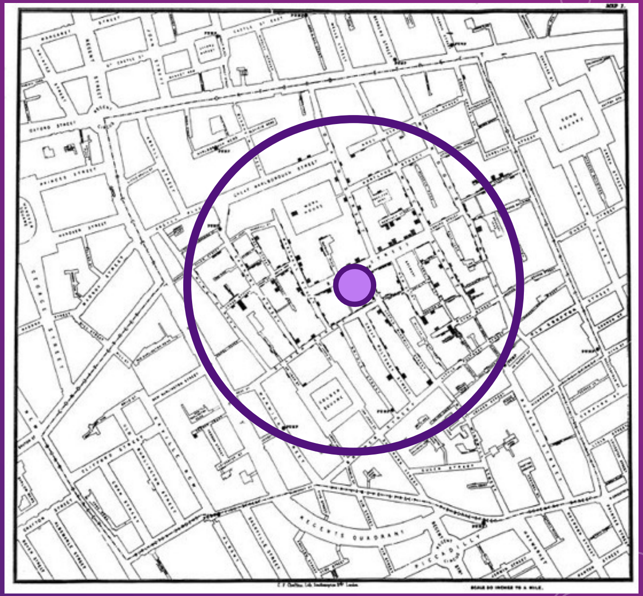
Locations of deaths in the 1854 London Cholera Epidemic

WHY VISUALIZE DATA

- 1854 London Cholera Epidemic



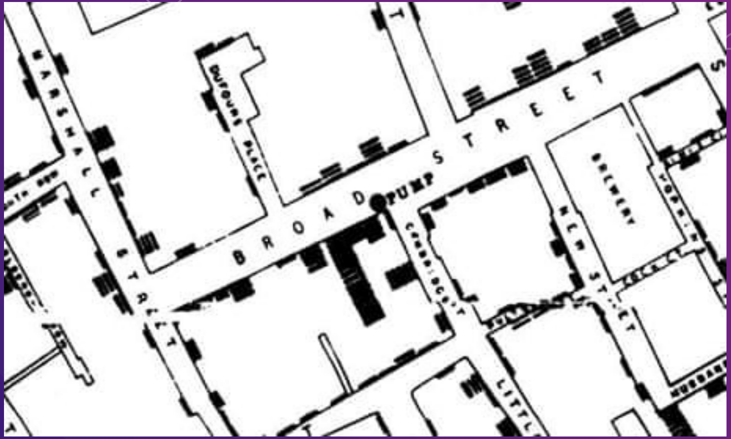
cluster region



cluster center

DATA VISUALIZATION

- 1854 London Cholera Epidemic
- Cholera occurred almost entirely among those who lived near (and drank from) the Broad Street water pump
- By removing the handle of the contaminated pump, the epidemic was controlled, which had taken more than 500 lives



https://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak
https://www.theguardian.com/news/datablog/2013/mar/15/john-snow-cholera-map
https://www.wired.com/2009/09/0908london-cholera-pump/

WHY VISUALIZE DATA

- Anscombe’s Quartet

D-I		D-II		D-III		D-IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Are these four datasets the same?

https://en.wikipedia.org/wiki/Anscombe%27s_quartet

WHY VISUALIZE DATA

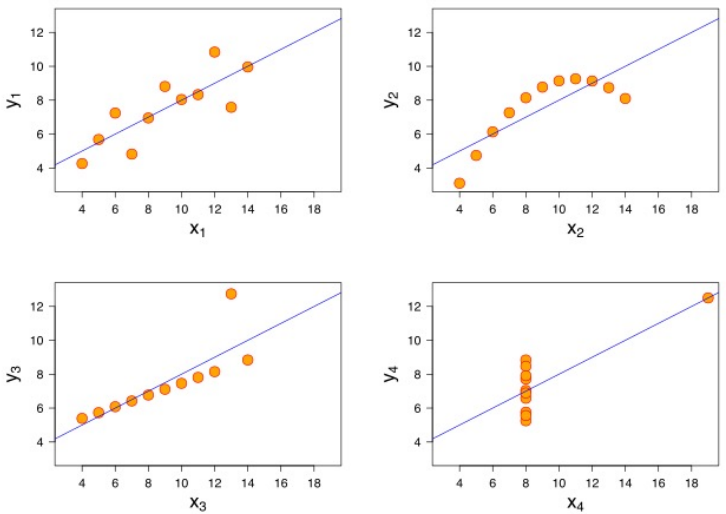
- Anscombe's Quartet

	D-I		D-II		D-III		D-IV	
	x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
	12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
mean	**9.0**	**7.5**	**9.0**	**7.5**	**9.0**	**7.5**	**9.0**	**7.5**
var.	**10.0**	**3.75**	**10.0**	**3.75**	**10.0**	**3.75**	**10.0**	**3.75**
corr.	**0.816**		**0.816**		**0.816**		**0.816**	

Interestingly, they all have the same mean, variance, and correlation

WHY VISUALIZE DATA

- Anscombe's Quartet



However, they appear very different when graphed

WHY VISUALIZE DATA

- Anscombe’s Quartet constructed in 1973 by the statistician Francis Anscombe
- Four datasets with nearly identical simple descriptive statistics
- Demonstrate the importance of graphing data

DATA VISUALIZATION

- Data visualization is the cornerstone of data science
- It is important to first visualize the data before applying more sophisticated data science methods
- There are two types of data visualization:
    - Data exploration visualization
        - Figuring out what is true
    - Data presentation visualization
        - Convincing other people it is true

- Data exploration is to put together the pieces of the puzzle
- Data presentation is to share the solved puzzle with people who can act on the insights

- Before running any analysis, always visualize the data
- If we can’t identify a trend or make a prediction from our dataset, neither will an automated algorithm

DATA VISUALIZATION

- Four important types of data to understand before visualizing data
- Nominal: categorical data with no ordering
  - Example – Pet: {dog, cat, rabbit}
  - Operations: =, ≠
- Ordinal: categorical data with ordering
  - Example – Rating: {1,2,3,4,5}
  - Operations: =, ≠, ≥, ≤, >, <
- Interval: numerical data in which zero has no fixed meaning
  - Example – In surveys, completely agree
  - Operations: =, ≠, ≥, ≤, >, <, +, −
- Ratio: numerical data in which zero has special meaning
  - Example – Temperature
  - Operations: = , ≠ , ≥ , ≤ , > , < , + , − , ÷

DATA VISUALIZATION

- Mostly data visualization revolves around charts and graphs
  - For visualizing data using charts, type and dimensionality of the underlying data is important

- Visualization types
  - 1D: bar chart, pie chart, histogram
  - 2D: scatter plot, line plot, box plot, whisker plot, heatmap
  - 3D+: scatter matrix, bubble chart

DATA VISUALIZATION

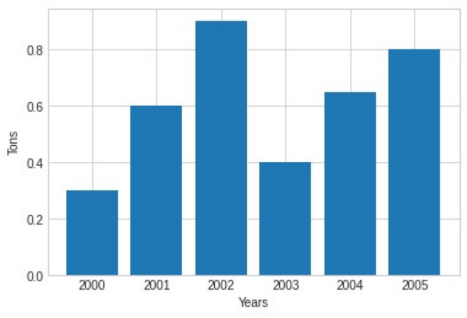
- Data visualization using Python
  - Python offers several plotting libraries with different features for creating informative, customized, and appealing plots to present data in the most simple and effective way
- Standard charts
  - matplotlib, seaborn, ggplot, altair
- Thematic maps
  - folium, basemap, cartopy, iris
- Advanced visualizations
  - bokeh, plotly

DATA VISUALIZATION

- Bar plot

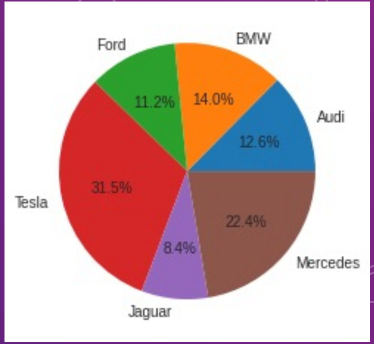
```
import matplotlib.pyplot as plt

years = range(2000, 2006)
apples = [0.3, 0.6, 0.9, 0.4, 0.65, 0.8]
plt.bar(years, apples)
plt.xlabel("Years")
plt.ylabel("Tons")
plt.show()
```



- Pie chart

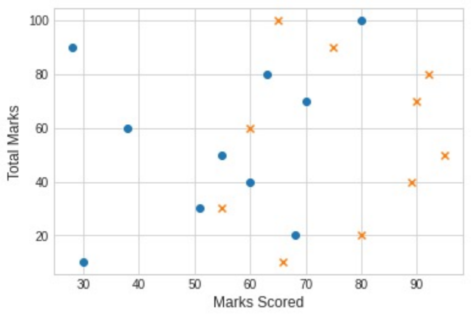
```
cars = ['Audi', 'BMW', 'Ford', 'Tesla', 'Jaguar', 'Mercedes']
data = [18, 20, 16, 45, 12, 32]
plt.pie(data, labels = cars, autopct='%1.1f%%')
plt.show()
```



DATA VISUALIZATION

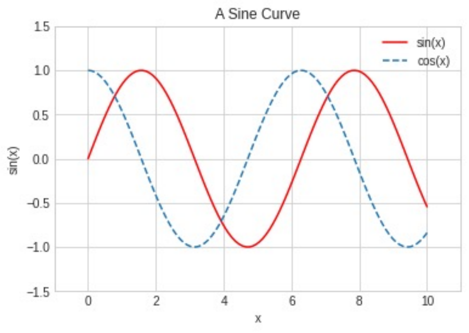
- Scatter plot

```
boys_grades = [30, 68, 51, 60, 55, 38, 70, 63, 28, 80]
girls_grades = [66, 80, 55, 89, 95, 60, 90, 92, 75, 65]
grades_range = [10, 20, 30, 40, 50, 60, 70, 80, 90, 100]
plt.scatter(boys_grades, grades_range, marker='o')
plt.scatter(girls_grades, grades_range, marker='x')
plt.xlabel('Marks Scored', fontsize=12)
plt.ylabel('Total Marks', fontsize=12)
plt.show()
```



- Line plot

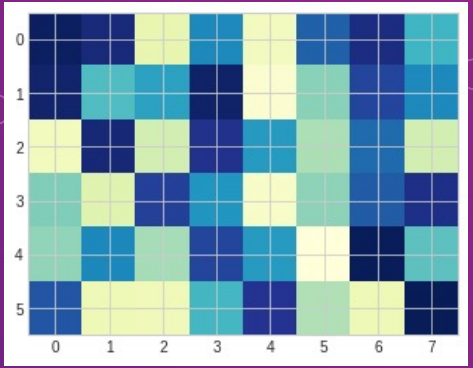
```
x = np.linspace(0, 10, 100)
plt.plot(x, np.sin(x), '-', color='red', label='sin(x)')
plt.plot(x, np.cos(x), '--', label='cos(x)')
plt.axis([-1, 11, -1.5, 1.5])
plt.title("A Sine Curve")
plt.xlabel("x")
plt.ylabel("sin(x)")
plt.legend()
plt.show()
```



DATA VISUALIZATION

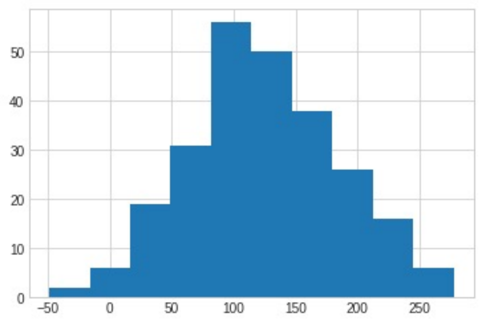
- Heatmap

```
data = np.random.random((6, 8))
plt.imshow(data, cmap='YlGnBu', interpolation='nearest')
plt.show()
```



- Histogram

```
x = np.random.normal(120, 60, 250)
plt.hist(x)
plt.show()
```

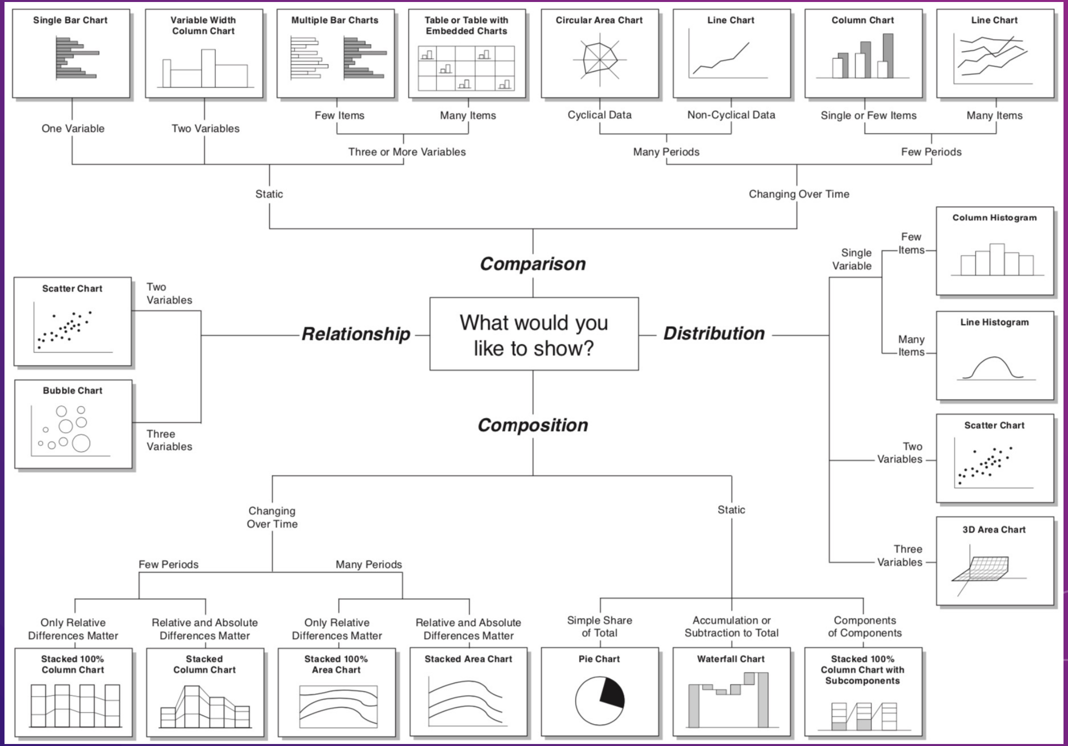


IMPORTANT GUIDELINES FOR CHARTS

- Label everything appropriately
- Work with the numbers
- Choose colors carefully
- Know your audience
- Use the correct chart

CHOOSE THE MOST APPROPRIATE CHART

- Chart Chooser



CHOOSE THE MOST APPROPRIATE CHART

- Comparison
  - Bar chart
    - horizontal bar
    - column chart
- Composition
  - 1d
    - donut
    - pie chart
  - 2d
    - stacked percent
    - stacked column
- Time series
  - Line chart
- Correlation
  - Scatter plot
  - heatmap
  - bubble chart
- Distribution
  - box plot
  - histogram

CHOOSE THE MOST APPROPRIATE CHART

- Question: How many new users are coming every day?
- Goal: Compare values (number of users) over time (days)
- Outcome: Line chart

- Question: From where these new users are coming from?
- Goal: Display composition of data (where new users came from) over time (new users across days)
- Outcome: Area chart

- Question: What time of day sees the highest number of users?
- Goal: Comparing values (number of visits) over time (hours) across multiple dimensions (days)
- Outcome: Overlay line chart

CHOOSE THE MOST APPROPRIATE CHART

- Question: Which referrers are driving the most traffic?
- Goal: Compare values (number of visits) across categories (referrers)
- Outcome: Bar chart

- Question: Which referrers tend to drive more traffic from desktops, and which ones from mobile devices?
- Goal: Comparing values (number of visits) across categories (referrers) and looking at composition within each (mobile vs. web traffic)
- Outcome: Stacked bar chart

- Question: How does the traffic from mobile and desktop stack up across referrers?
- Goal: Comparing values (number of visits) across categories (referrers) in multiple dimensions (mobile and desktop)
- Outcome: Grouped bar chart

CHOOSE THE MOST APPROPRIATE CHART

- Question: Which pages are driving the most engagement based on where users are coming from to those pages?
- Goal: See at the relationship between where the users are coming from and landing pages to see how the different combinations influence average visit duration
- Outcome: Heat map

- Question: How to find out ways to divert more traffic to high-performing pages?
- Goal: See the relationship between high-performing pages and number of visits to those pages to better promote those pages
- Outcome: Scatterplot

THANKS