Data Science Unit 2

Data Visualisation

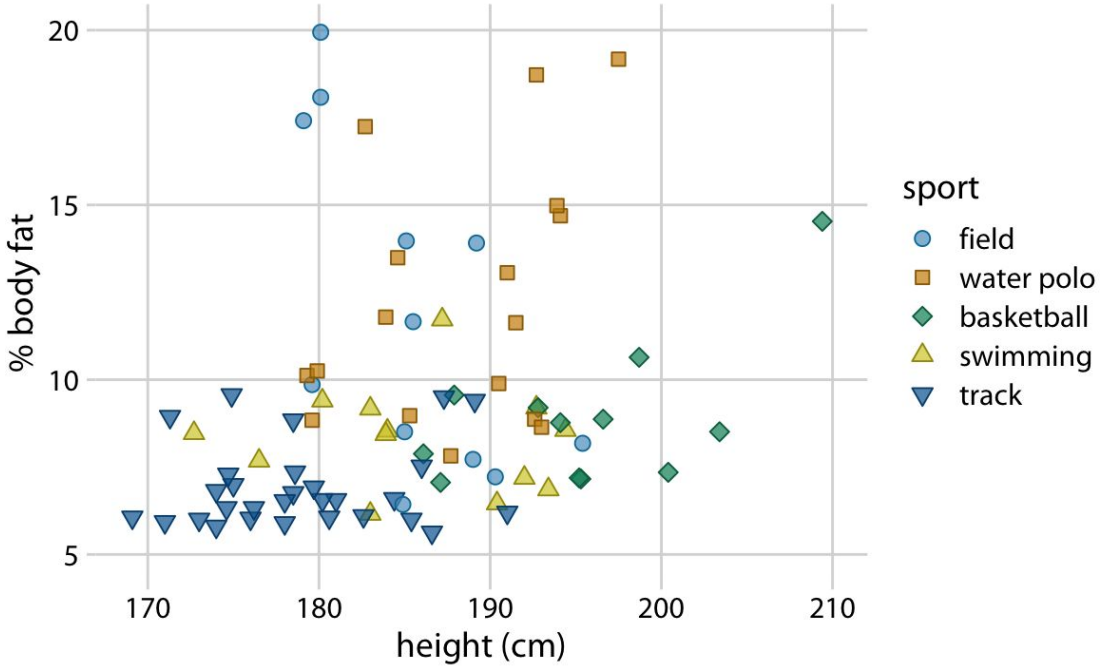
In this session we will...

1. Understand why Data Visualisation is Important
2. Identify what makes a ‘Good’ Data Visualisation
3. Identify the Most Appropriate Visualisation

Why use Data Visualisation

Why do you think Data Visualization is useful?

Humans are exceptional at processing complex information when it is presented in visual form



Without visualization, we rely on statistics to interpret data

But...

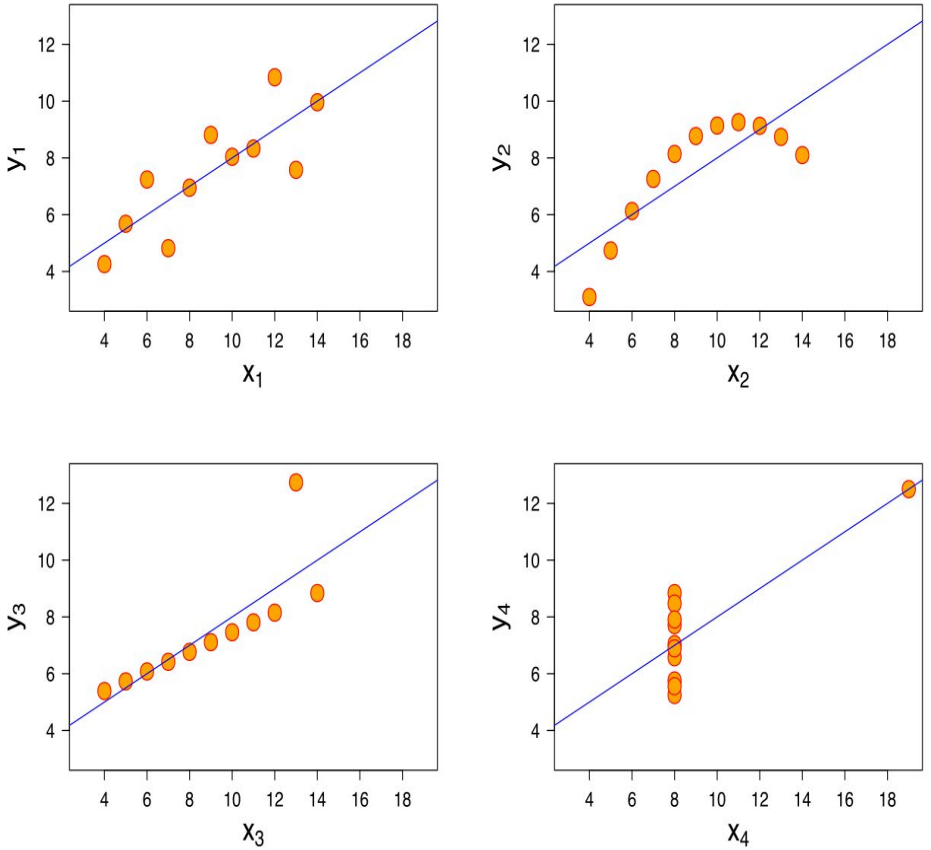
What happens if our basic metrics are the same?

Anscombe's Quartet

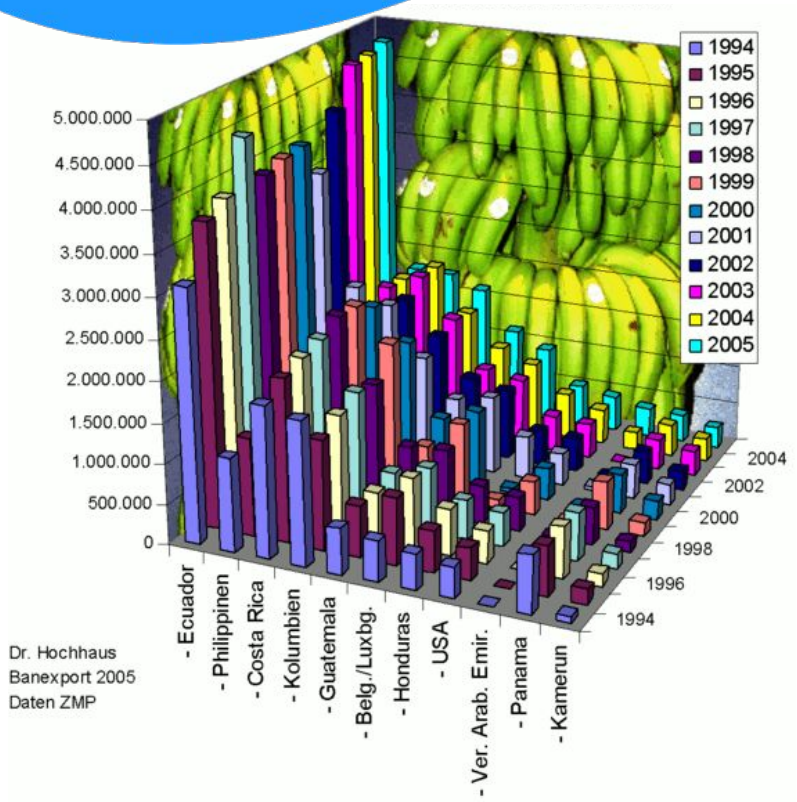
I		II		III		IV	
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Visualization complements statistics

Property	Value
x-mean in each case:	9 (exact)
x-variance in each case:	11 (exact)
y-mean in each case:	7.50
y-variance in each case:	4.122 or 4.127
Correlation between x & y:	0.816
Linear regression line:	y = 3.00 + 0.500x



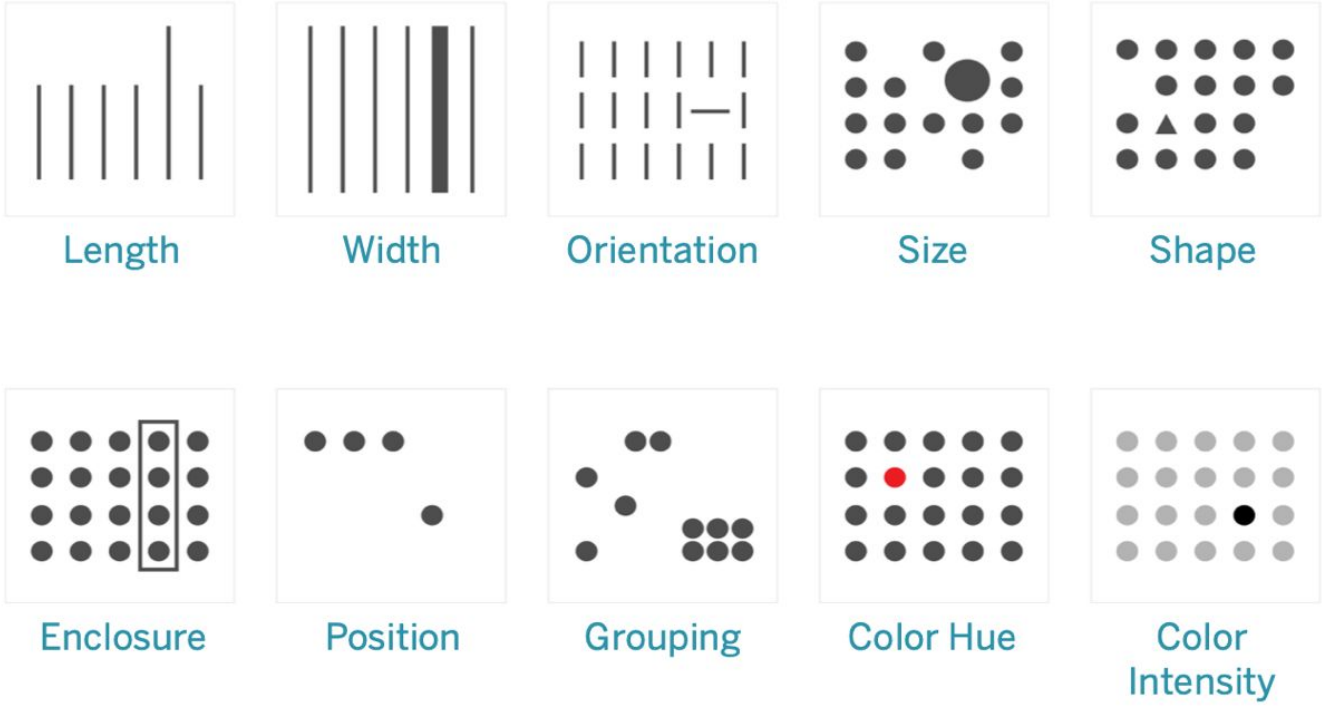
Attributes of a good visualisation



Human eyes good at seeing visual patterns!...

Sometimes.

Preattentive attributes of visual perception



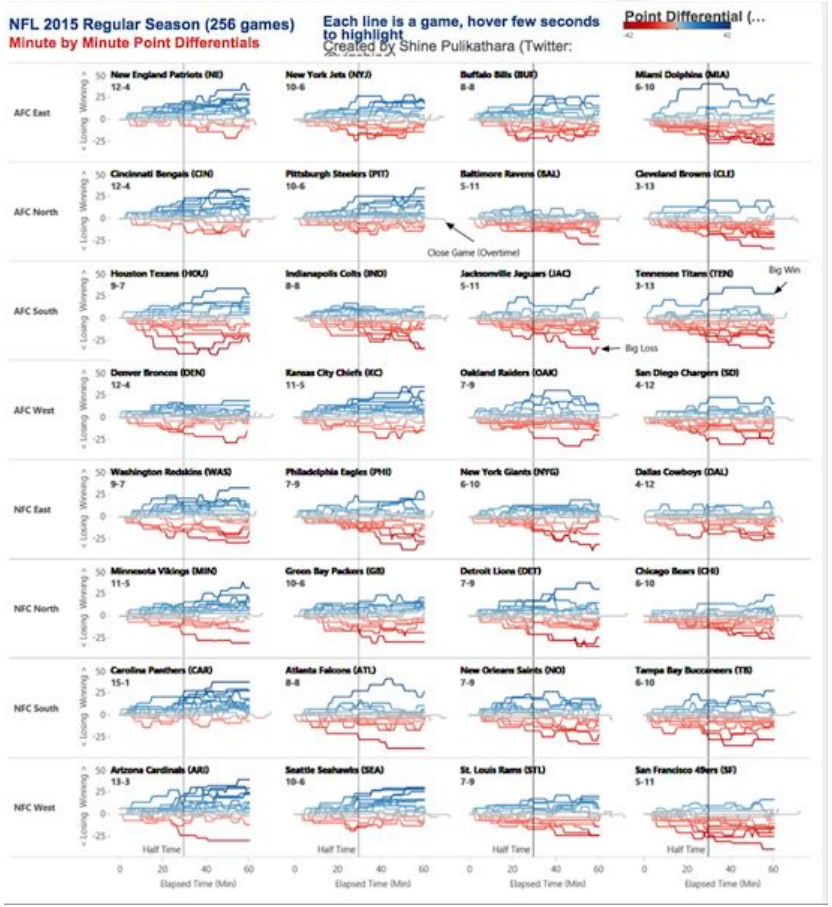
Some attributes have more of an effect on our brains than others

Let's look at three visualizations. Which one catches your attention most? Why?

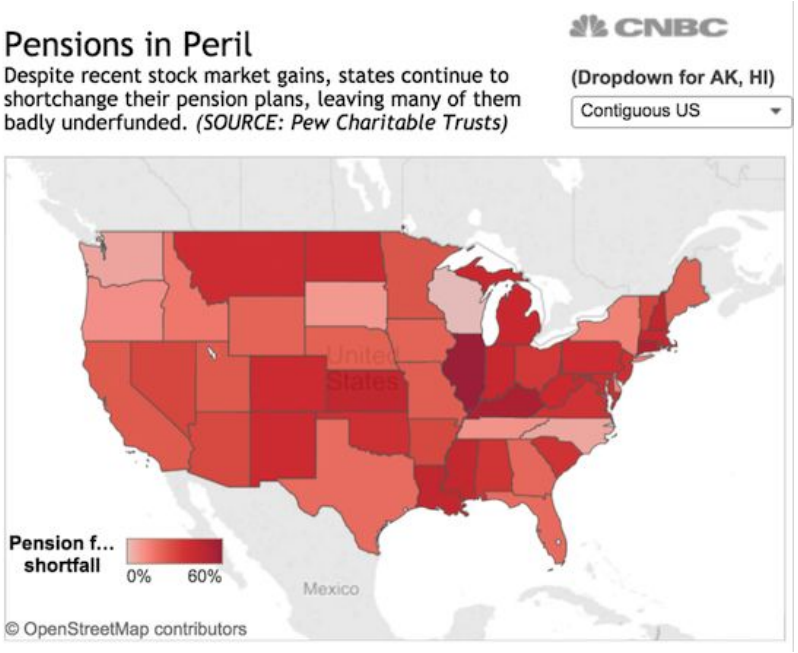
You can use color in one of three ways:

- Divergent
- Sequential
- Categorical

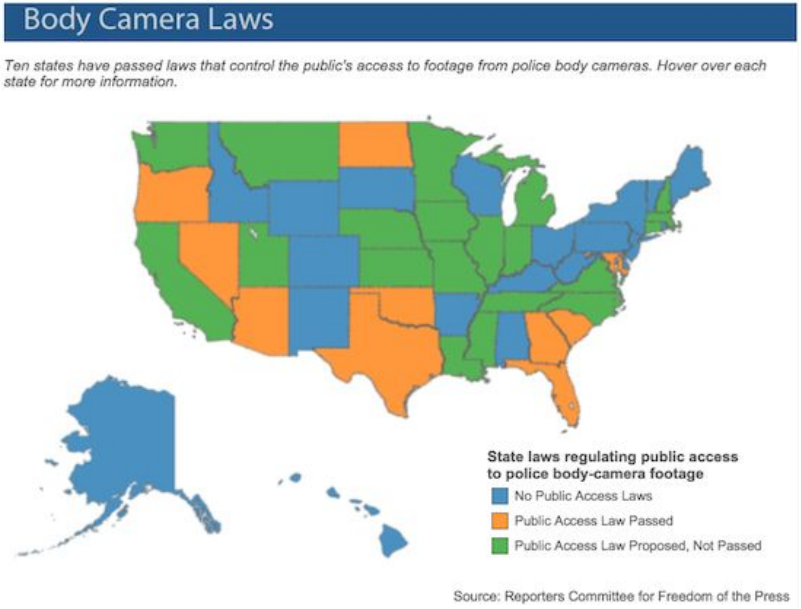
Divergent colors are used to show ordered values that have a critical midpoint



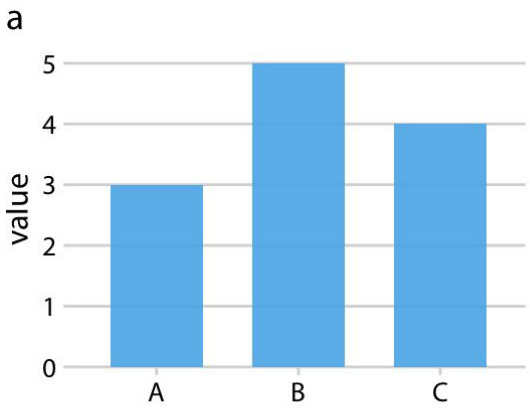
Sequential colors are used to show values ordered from low to high

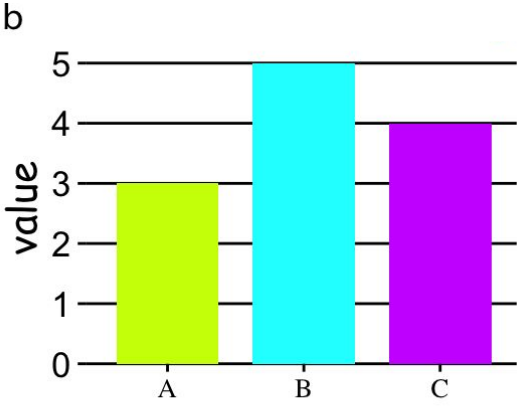


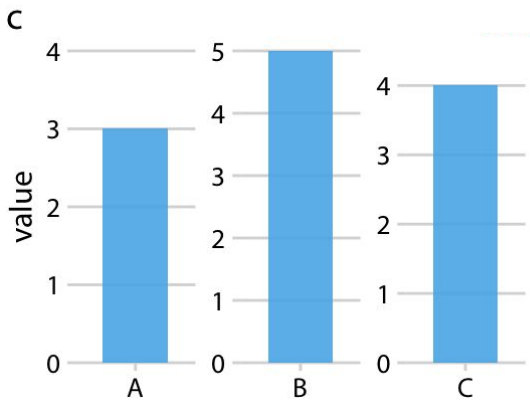
Categorical colors are used to distinguish data that falls into distinct groups

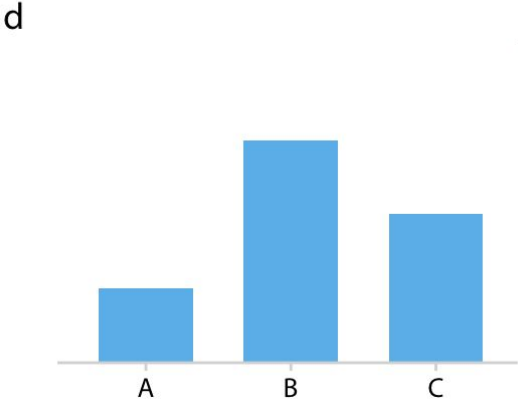


Which chart follows good visualization principles? Why?







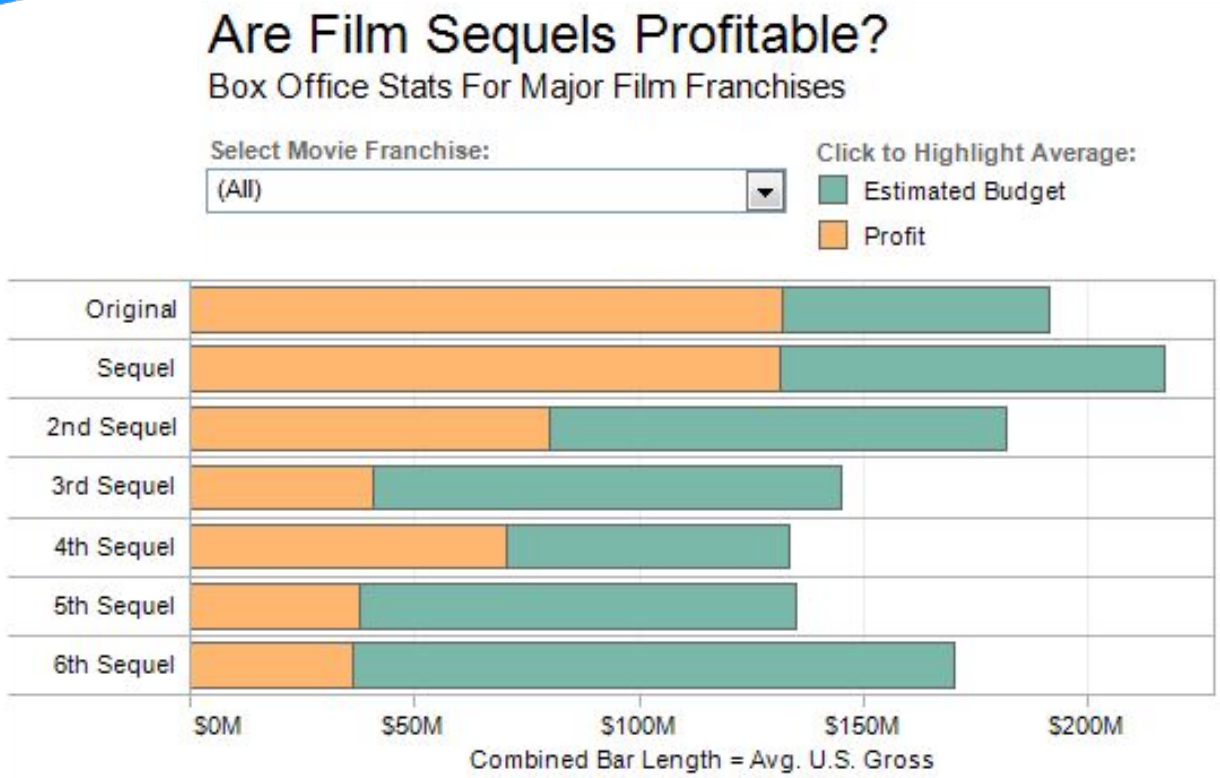


Choosing the right visualization

Visualising Amounts

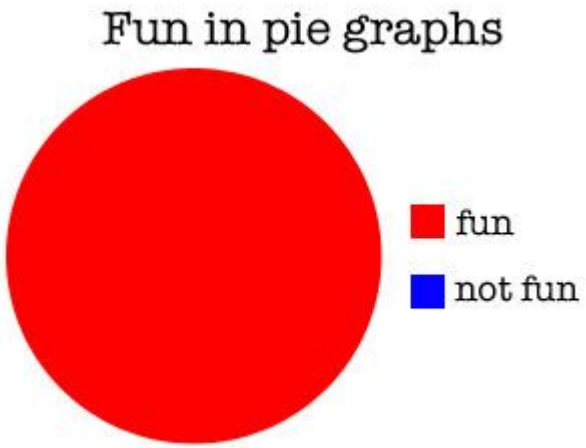
Bar charts are one of the most common ways of visualizing data. Why?

Bar chart- Example



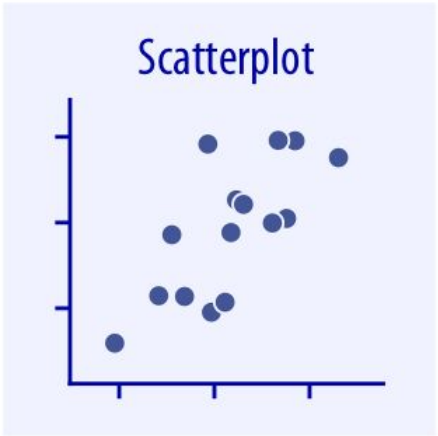
Visualising proportions

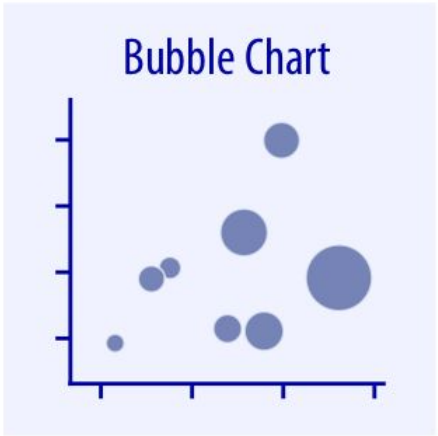
Pie charts can be used to visualize proportion, but they are the most commonly misused chart type.

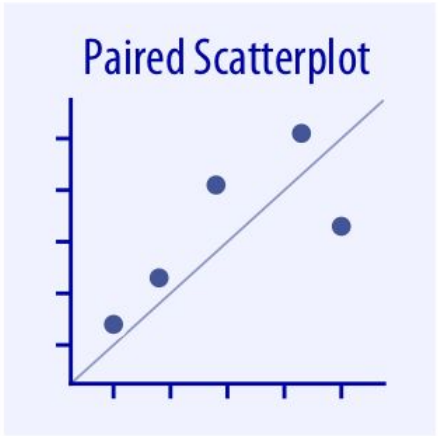


X-Y relationships

Scatter plots are a great way to give you a sense of trends, concentrations, and outliers, which provides a clear idea of what you may want to investigate further..

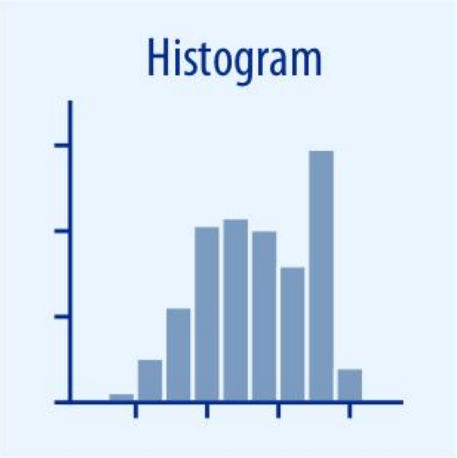


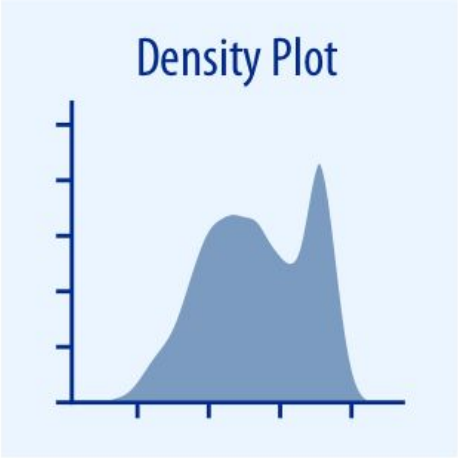




Visualising distributions

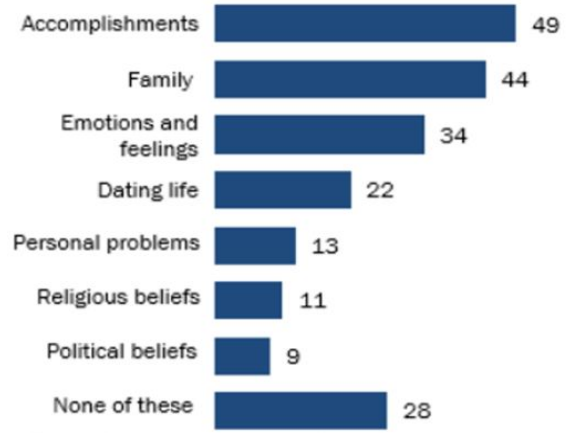
Scatter plots are a great way to give you a sense of trends, concentrations, and outliers. This will provide a clear idea of what you may want to investigate further..





Which chart shows a distribution?

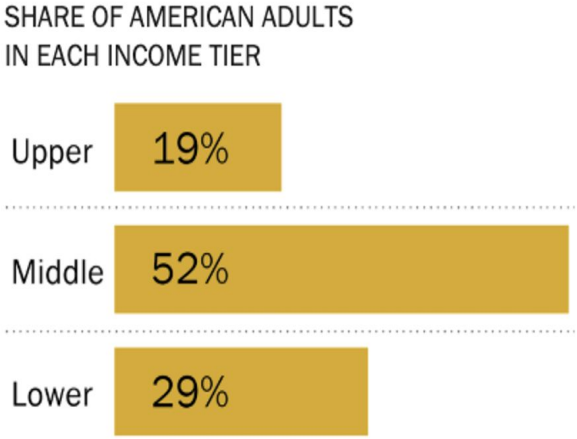
A



While about half of teens post their accomplishments on social media, few discuss their religious or political beliefs

% of U.S. teens who say they ever post about their ___ on social media

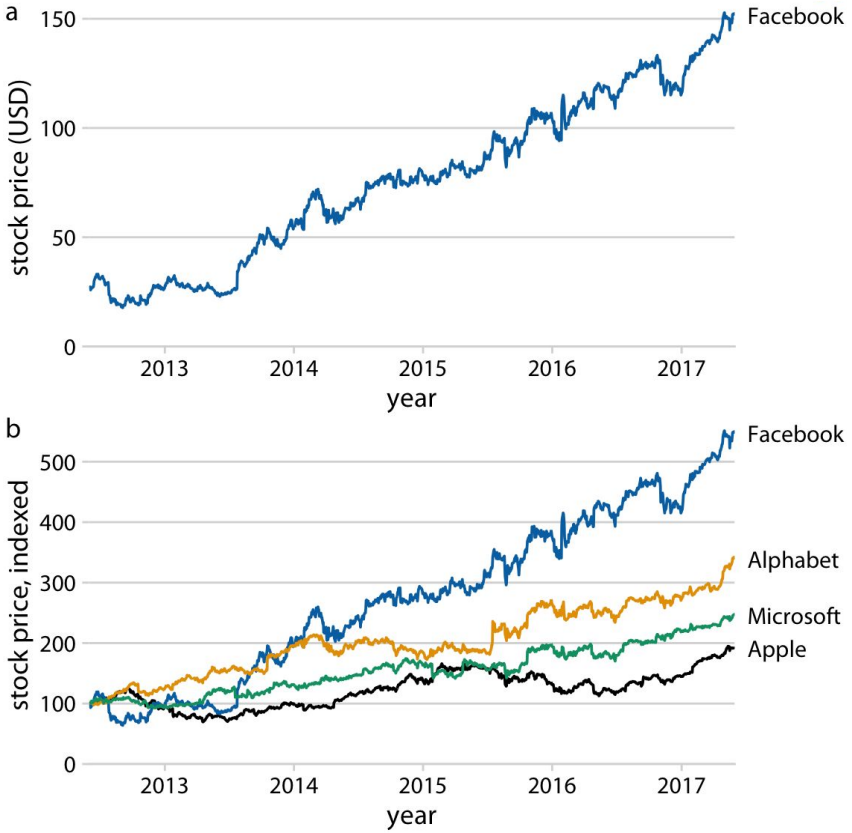
B



The figure shows growth of Facebook stock price over a five-year interval and comparison with other tech stocks?

What is wrong with this figure? Why?

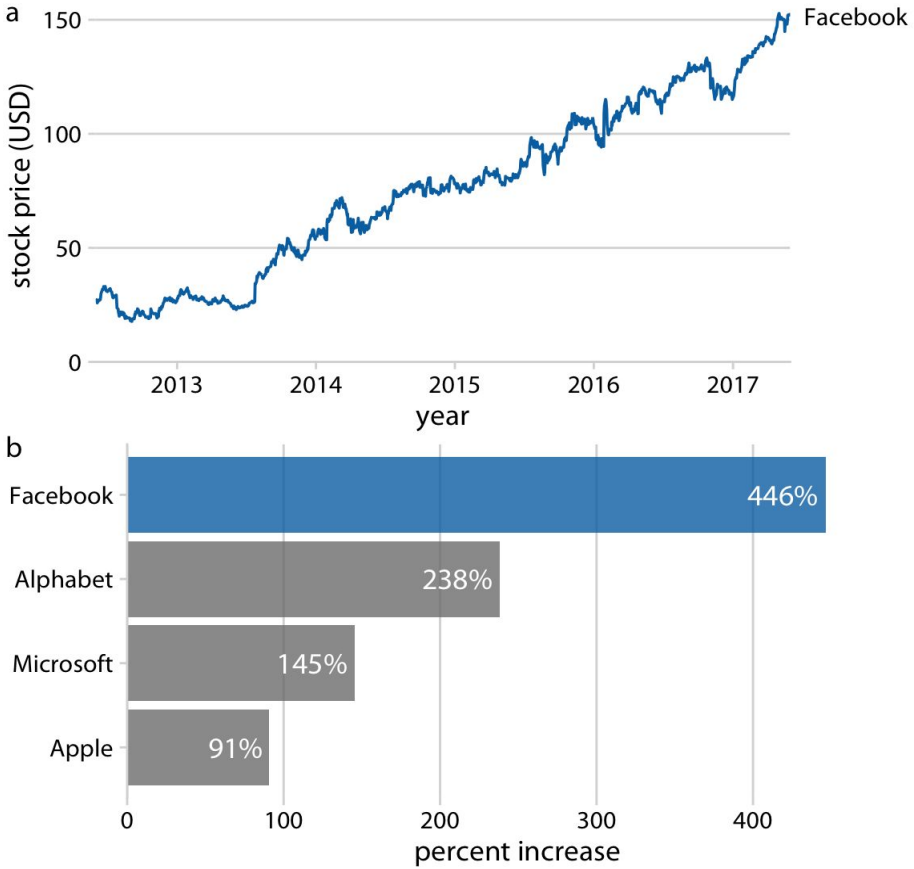
Suggest which chart type we can use to make this a good visualization?



Possible suggestion:

Leave part (a) as is but replace part (b) with a bar plot showing percent increase

Now we have two distinct figures that each make a unique, clear point and that work well in combination.



Python Libraries

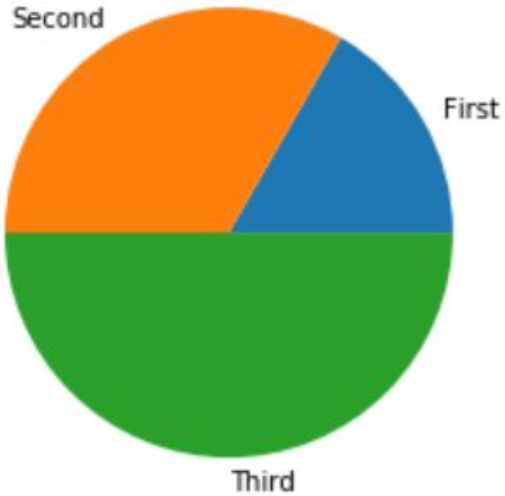
To effectively use data visualizations, you must be proficient with both the principles of visualization and the programming tools to generate plots.

In this lesson, we will use the Python libraries Matplotlib (Python plotting) and Seaborn (statistical data visualization).

matplotlib

```
import matplotlib.pyplot as plt
```

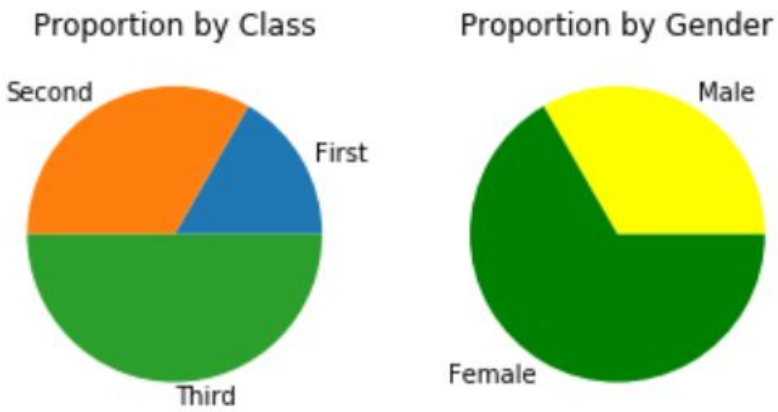
```
plt.pie([1,2,3],data=df.Pclass,labels=['First','Second','Third']);
```



matplotlib

```
import matplotlib.pyplot as plt
```

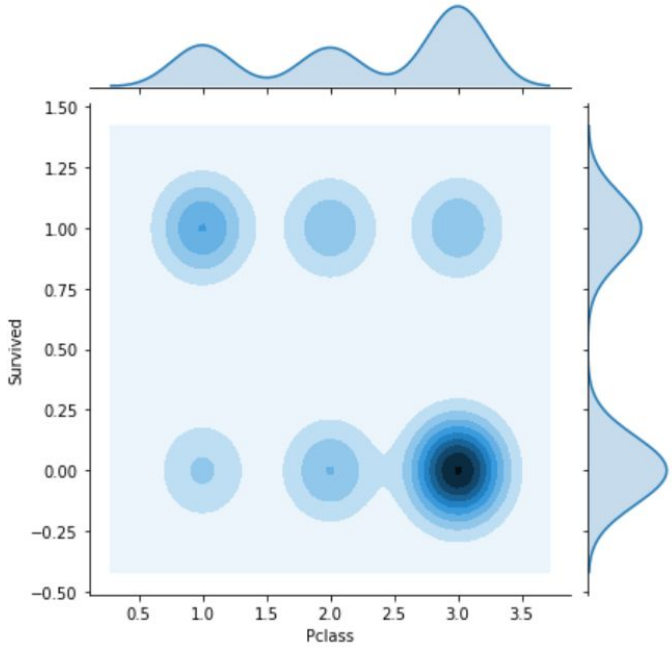
```
fig,ax=plt.subplots(ncols=2)
ax[0].pie([1,2,3],data=df.Pclass,labels=['First','Second','Third'])
ax[0].set_title('Proportion by Class')
ax[1].pie([1,2],data=df.Gender,labels=['Male','Female'],colors=['Yellow','Green'])
ax[1].set_title('Proportion by Gender');
```



Seaborn

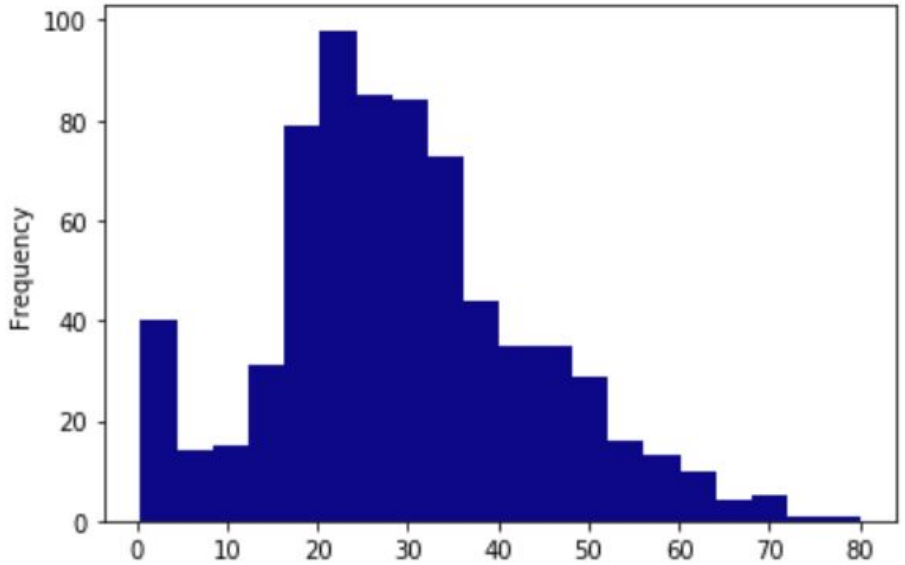
```
import seaborn as sns
```

```
sns.jointplot(x='Pclass',y='Survived',data=df,kind='kde');
```



pandas

```
df.Age.plot(kind='hist',bins=20,colormap='plasma');
```



Other libraries

Many other Python libraries exist for making visualizations. Some of the most popular include:

- Bokeh: Python visualization library that targets the web browser (e.g., in Jupyter). Makes interactive plots, dashboards, data applications, etc.
- Graphviz: Popular visualization library for graph data structures (e.g., edges, vertices, etc). Has Python extensions.
- Basemap: Python Matplotlib extension for drawing static maps. There are many other Python libraries for plotting geographic data, including ones that might be easier to use, but many are not actively developed.
- D3.js: JavaScript library for interactive web visualizations

Other tools

Although this course emphasizes a Python approach to data science, a variety of non-programming tools are also used in industry. Often, these tools can be applied much more quickly than creating a custom Python solution.

For example:

- Excel: For quick data cleaning and simple graphs
- Power BI: A suite of business analytics tools
- Tableau: Business intelligence and analytics software
- Periscope Data: Data analysis platform
- Plotly: Create charts and dashboards

Independent Practice

Python Plotting With Pandas and Seaborn

Open up the independent research notebook to explore plotting the sales data with Python.

Summary

Summary

- Why is data visualization so important?
- What are some considerations to keep in mind when creating a visualization?
- Describe when you would use the following types of charts or graphs:
  - Bar chart
  - Pie chart
  - Scatter plot
  - Histogram

Any Questions

OTJ and SAL