Welcome to General Assembly

- **WiFi** GA Guest
- **Password** yellowpencil

GENERAL ASSEMBLY

DATA SCIENCE

10 WEEK PART TIME COURSE

Week 2 Lesson 1 - Data Visualisation

AGENDA & Learning Plan

Key: Interactive Paul Presenting

1. Course Plan revisited
2. Recap last time & Homework Presentations
3. Review – Git & GitHub setup
4. What is Data Visualisation?
5. Lab 1
6. Why do we visualise data?
7. How do we visualise data?
8. Different types of Charts
9. Lab 2
10. Basic rules for creating a graph
11. Discussion

Course Plan

UNITS

UNIT 1: FOUNDATIONS OF DATA MODELING	Introduction to Data Science	Lesson 1
	Elements of Data Science	Lesson 2
	Data Visualisation	Lesson 3
	Linear Regression	Lesson 4
	Logistic Regression	Lesson 5
	Model Evaluation	Lesson 6
	Regularisation	Lesson 7
	Clustering	Lesson 8
UNIT 2: DATA SCIENCE IN THE REAL WORLD	Recommendations	Lesson 9
	SQL + Productivity	Lesson 10
	Decision Trees	Lesson 11
	Ensembles	Lesson 12
	Natural Language Programming	Lesson 13
	Cloud Computing	Lesson 14
	Time Series	Lesson 15
	Soft Skills	Lesson 16
	Network Analysis	Lesson 17
	Neural Networks	Lesson 18
	Final Projects Presentations	Lesson 19
	Final Projects Presentations	Lesson 20

GA

Recap – last lesson

- Lab1 done?
- Lab2 done?
- Lab3 done?

Homework Presentations ~ 2mins each

Data Science Homework 1

DAT11 | Lesson 2 | Homework 1

Investigating Data

Instructions for the homework - time required: ~2 hours

1. Browse through the list of Data Science problems on Kaggle: https://www.kaggle.com/datasets
2. Choose 3 data sets that interest you,
3. Download the data and read it into Python Notebook on Jupyter Hub
4. Have a look at the datasets you have downloaded, and consider:

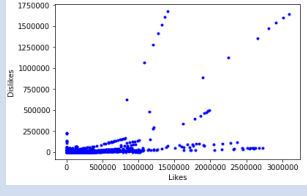
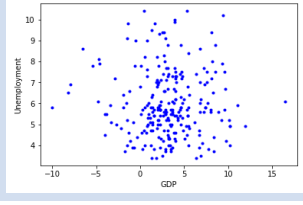
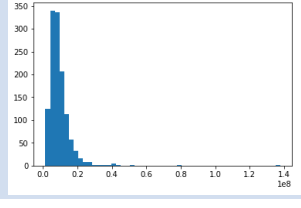
a. What is the data for?,
b. What are the data-types? (Numerics, Characteristics, etc?)
c. What alternative uses could you think of for the dataset?

5. Derive some simple, interesting facts from each dataset to report back to the class (plot a trend, or determine a relationship)

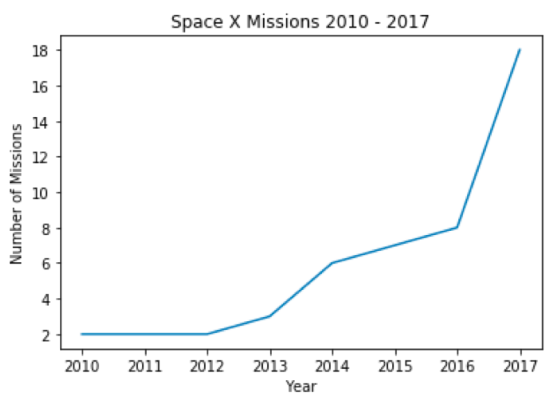
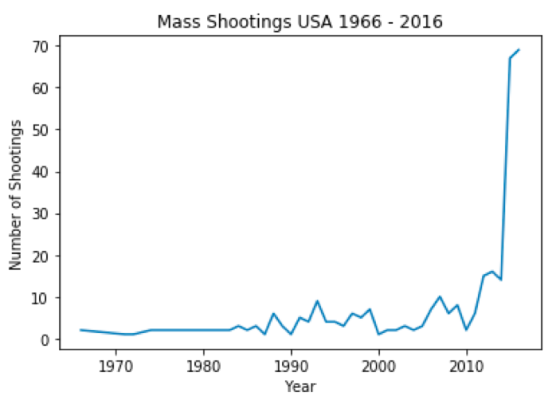
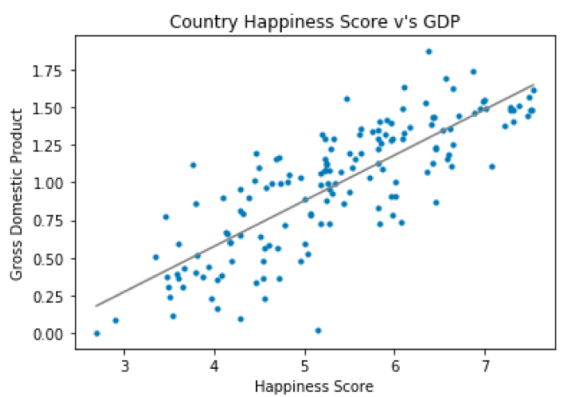
For the next class (Monday 26th Feb), please prepare:

```
* 1 single slide about your investigations - plan to keep your presentation to <= 5mins *
```

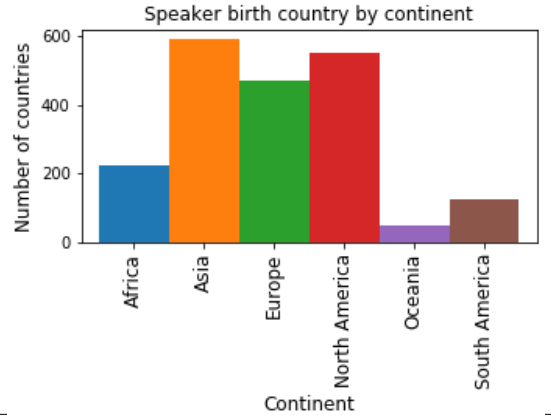
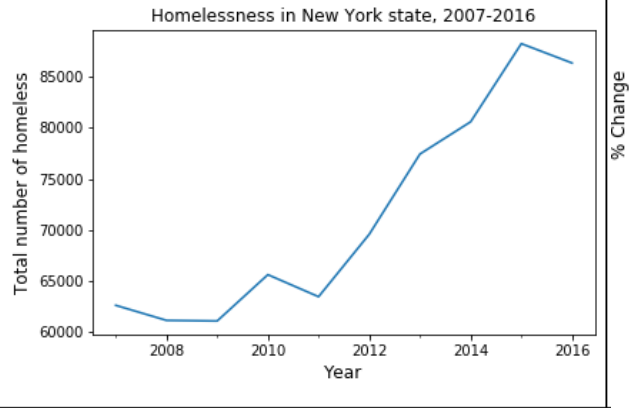
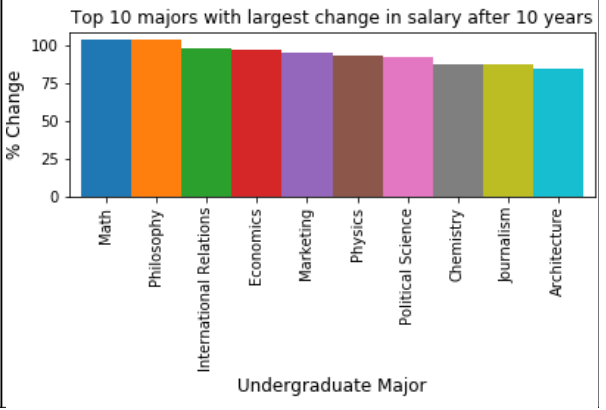
Tostee

	YouTube Trending Video	Federal Reserve Interest Rate	S&P Stock Data
Data Purpose	This dataset is a daily record of the top trending YouTube videos. • Sentiment analysis in a variety of forms • Categorising YouTube videos based on their comments and statistics. • Training ML algorithms like RNNs to generate their own YouTube comments. • Analysing what factors affect how popular a YouTube video will be. • Statistical analysis over time.	This dataset includes data on the economic conditions in the United States on a monthly basis since 1954. The federal funds rate is the interest rate at which depository institutions trade federal funds (balances held at Federal Reserve Banks) with each other overnight. The rate that the borrowing institution pays to the lending institution is determined between the two banks; the weighted average rate for all of these types of negotiations is called the effective federal funds rate. How does economic growth, unemployment, and inflation impact the Federal Reserve's interest rates decisions? How has the interest rate policy changed over time? Can you predict the Federal Reserve's next decision? Will the target range set in March 2017 be increased, decreased, or remain the same?	The data is presented in a couple of formats to suit different individual's needs or computational limitations. I have included files containing 5 years of stock data This dataset lends itself to a some very interesting visualizations. One can look at simple things like how prices change over time, graph an compare multiple stocks at once, or generate and graph new metrics from the data provided. From these data informative stock stats such as volatility and moving averages can be easily calculated. The million dollar question is: can you develop a model that can beat the market and allow you to make statistically informed trades!
Data Types	Data includes the video title, channel title, publish time, tags, views, likes and dislikes, description, and comment count. Data is Strings, Numeric, DateTime	Federal Funds Rate, Real GDP, Unemployment Rate, Inflation Rate Numeric, String	All the files have the following columns: Date - in format: yy-mm-dd Open - price of the stock at market open (this is NYSE data so all in USD). High, Low, Close, Volume
Alternative Uses	Measurin	Identifying how changes in interest rates will impact GDP over the subsequent periods	Predict volumes based on historical patterns. Identify seasonality flows.
Interesting Fact	Max Likes 3,093,544 Max Dislikes 1,674,420 Dislikes Avg 3,042, Median 332 Likes Avg 9011, Median 332	Max unemployment 10.8% Highest Growth 16.5%	AAL Median Volume 8,111,323 AAL Max Volume 137,767,165 AAL American Airlines
Plot / Relationship			

Jon

Space X Missions	Mass Shootings	World Happiness
• Dataset relates to Space X missions (rocket launches) between 2010 – 2017 • Average days between missions - 61.3 • Average days between missions 2017 – 20.27 • 55% of missions have been low earth orbit • Increased trend in missions	• Dataset relates to mass shootings in the USA between 1966 – 2016 • Average mass shootings p/a 1966 – 2014 (4.5) • Average mass shootings p/a 2015 - 2016 (68.0) • Significant increase in mass shootings 2015	• Dataset relates to a 2017 survey of peoples happiness in 155 countries • Top 5 countries are European (Norway, Denmark, Iceland, Switzerland, Finland) • Australia ranked 10th • Strongest linear relationship of happiness score was a countries GDP (0.82)
		

Mai

Speech Accent Archive	Homelessness in the US	Where It Pays to Attend College
2140 recordings by people from 77 countries and 214 native languages reading the same English text aloud	Point-in-time estimates of homeless people reported nationally and by state from 2007 to 2016	US starting and mid-career (after 10 years) salaries of people with a bachelor's degree by college, region and undergraduate major
Data types: strings, integers, boolean	Data types: strings, integers, dates	Data types: strings (!!!)
• 27% are native speakers of English • Top 3 non-English native language are Spanish, Arabic and Mandarin • 18% of speakers in the archive are from the US • 45% of countries are represented by one speaker	• Categories of homelessness included sheltered/unsheltered individuals/families, chronic homelessness and homeless veterans • California had the highest count of homeless people (1.2 million) • Top 3 states with the most homeless youth under 25 in 2016 were California, New York and Florida	• People with a Physician Assistant major had the highest median starting salary ($74 300) • Majors that had a median mid-career salary > $100 000 are all in engineering • Majors surveyed did not include Data Science
• Speech recognition • Teaching resource for ESL	• Inform government policy on housing • Help charities allocate funds	• Influence student choice of institution/major • Inform students/colleges of earning potential
		

Jesse

S&P 500 Index Price Data

Data Purpose?

- Daily stock price data, 5 years of daily trading history across all 500 stocks

Data Types?

- Numerical, dynamic (monthly refresh)

What's in Data?

- Two files, all stocks and individual stocks
- Five years of historical figures
- Labelled by stock
- Ticker, opening price, low/high price each day, volume of shares traded

Alternative Uses?

- Micro vs Macro factors
- Turnover of particular sectors (IT vs mining vs banking etc)
- Correlation between closing price and opening price

Facts?

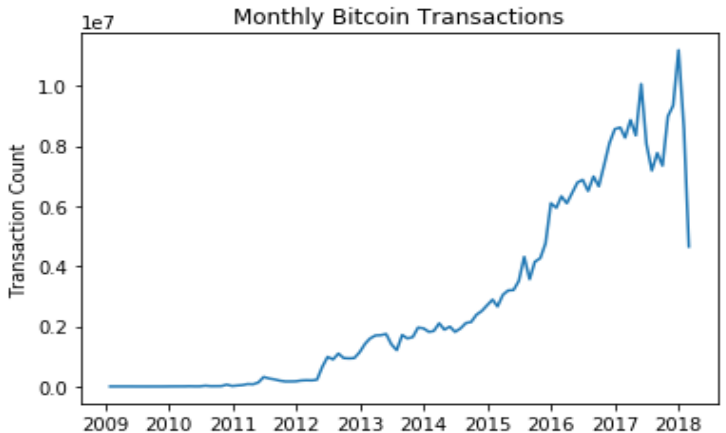
- $24 trillion total value as at 8-Feb-2018
- Top traded share was AAPL

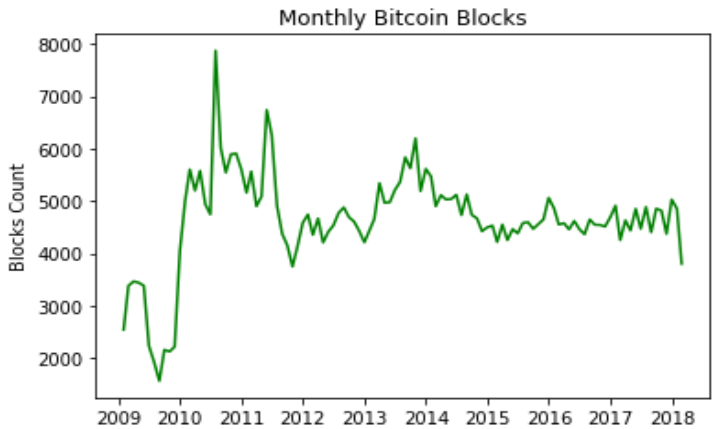
Jesse Imer- Feb 26, 2018

John

The Relationship between Bitcoin Transactions and Blocks:

From emergence to present day



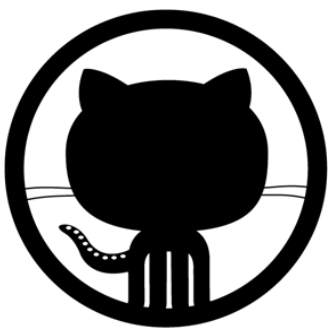


What is the data for?	Daily bitcoin blockchain information on the blocks and transaction data dating from Jan 2009
What are the data types?	Strings and Numeric. Time data had to be converted to a readable format
Alternative use?	- Can be joined with price data - which addresses receive bitcoin each day

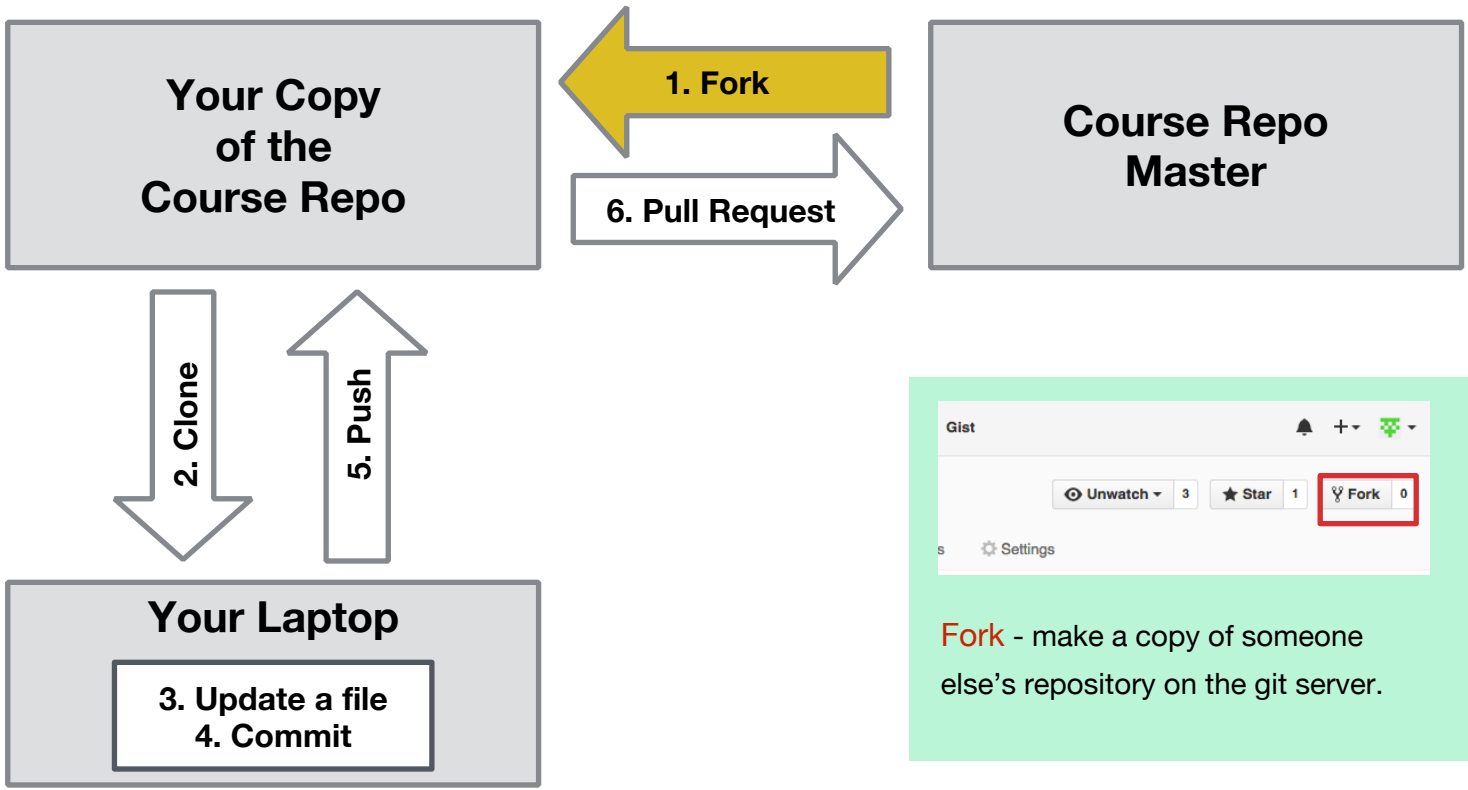
John.N

Student

Review:
Git and GitHub

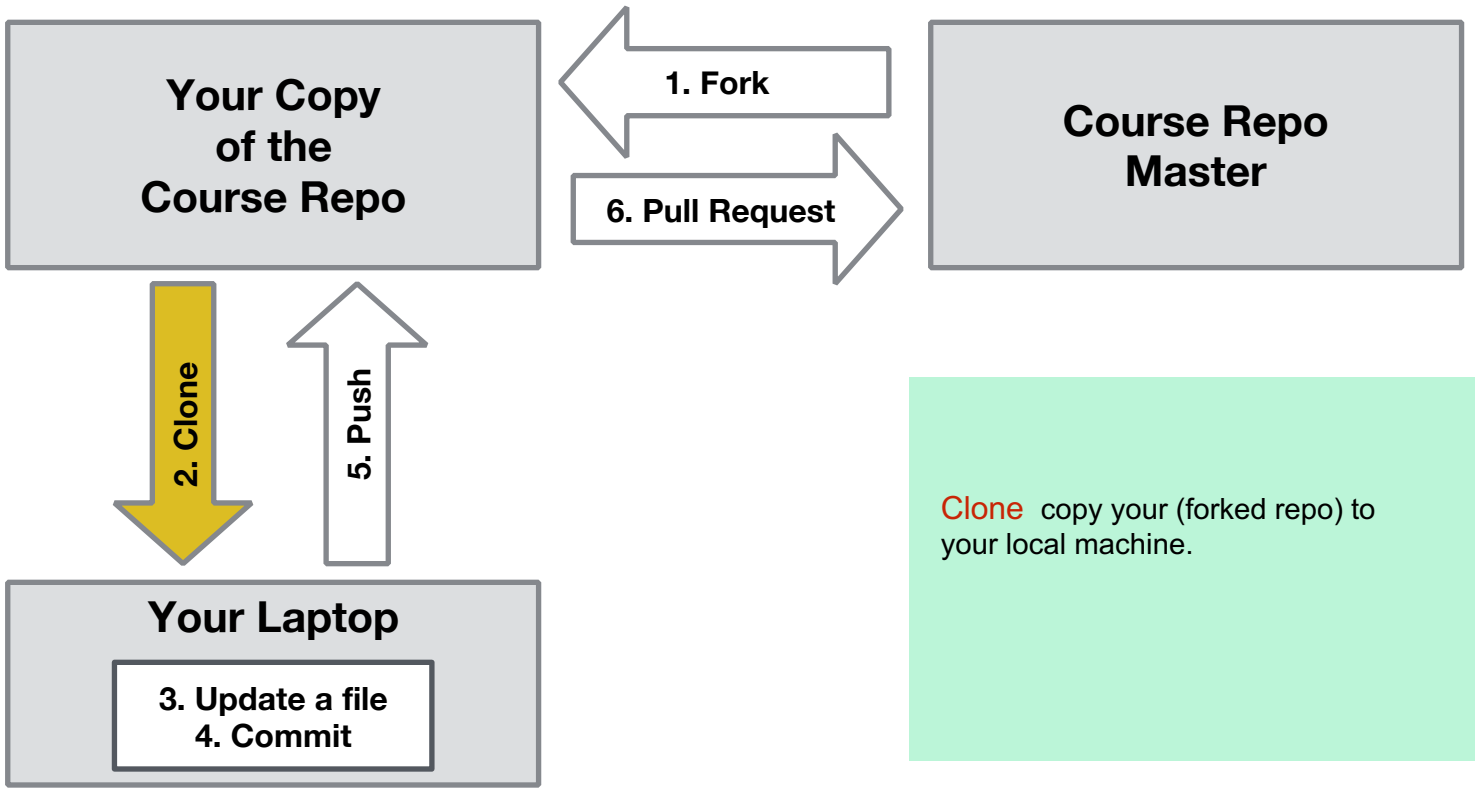


Using Git



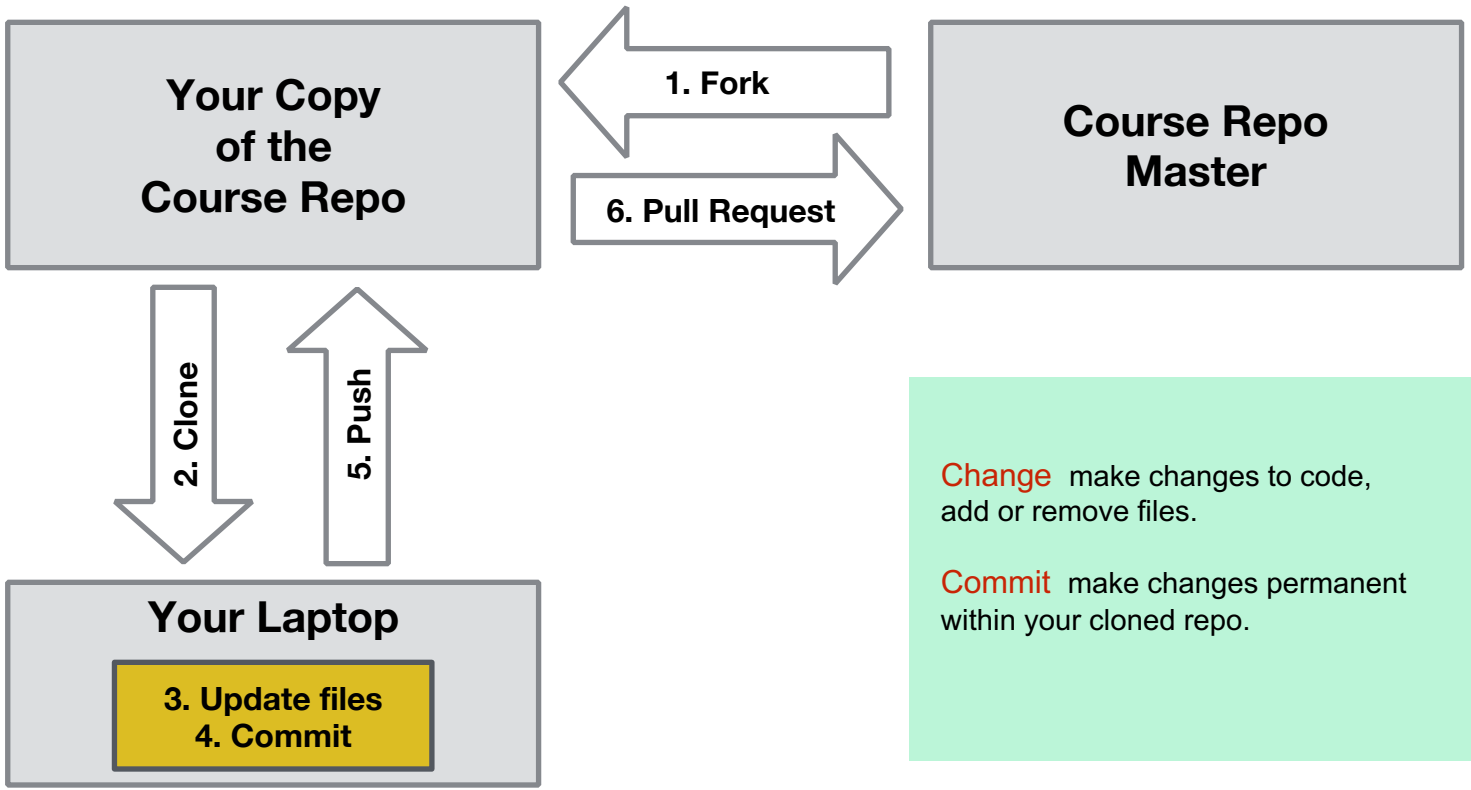
Fork - make a copy of someone else's repository on the git server.

Using Git



Clone copy your (forked repo) to your local machine.

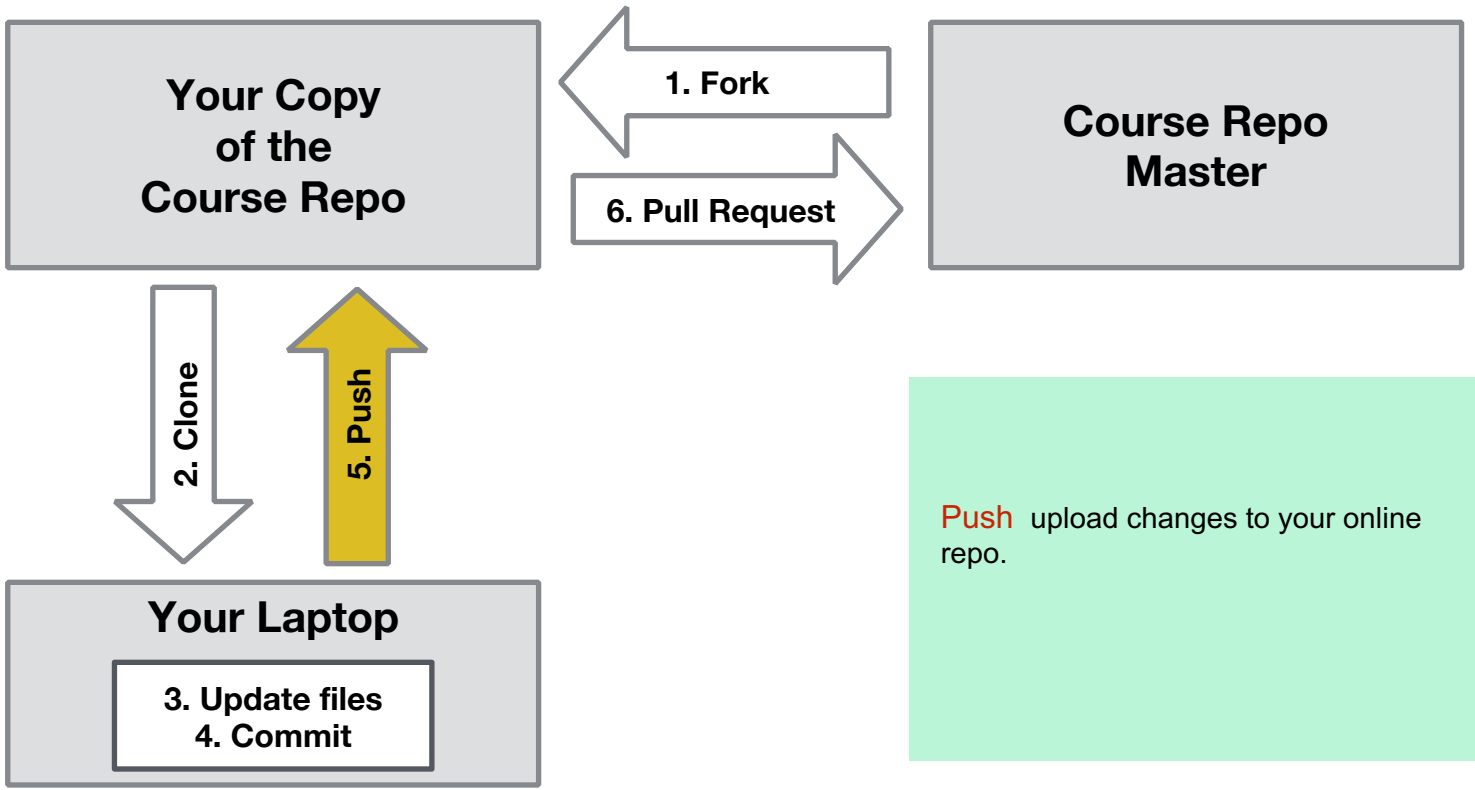
Using Git



Change make changes to code, add or remove files.

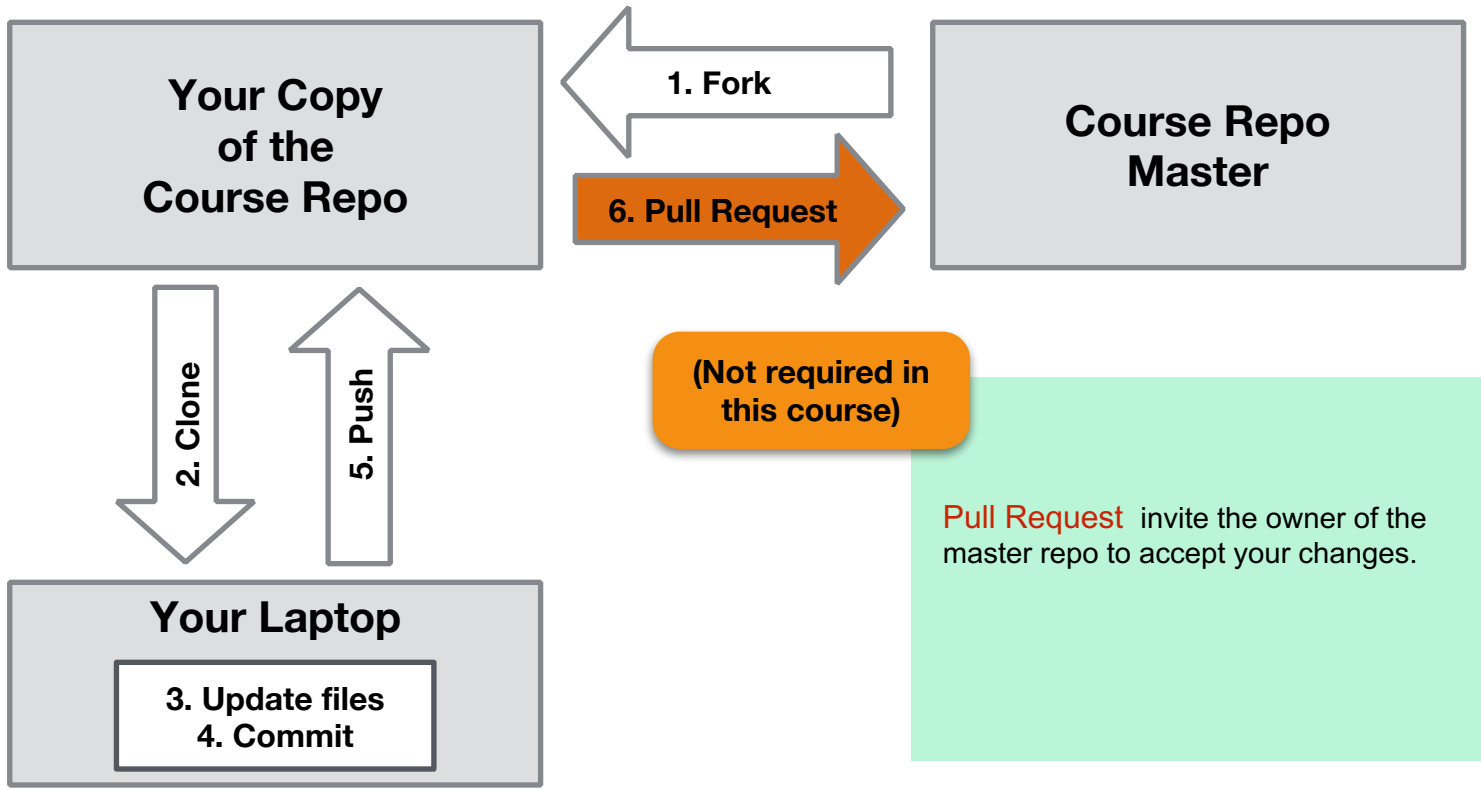
Commit make changes permanent within your cloned repo.

Using Git



Push upload changes to your online repo.

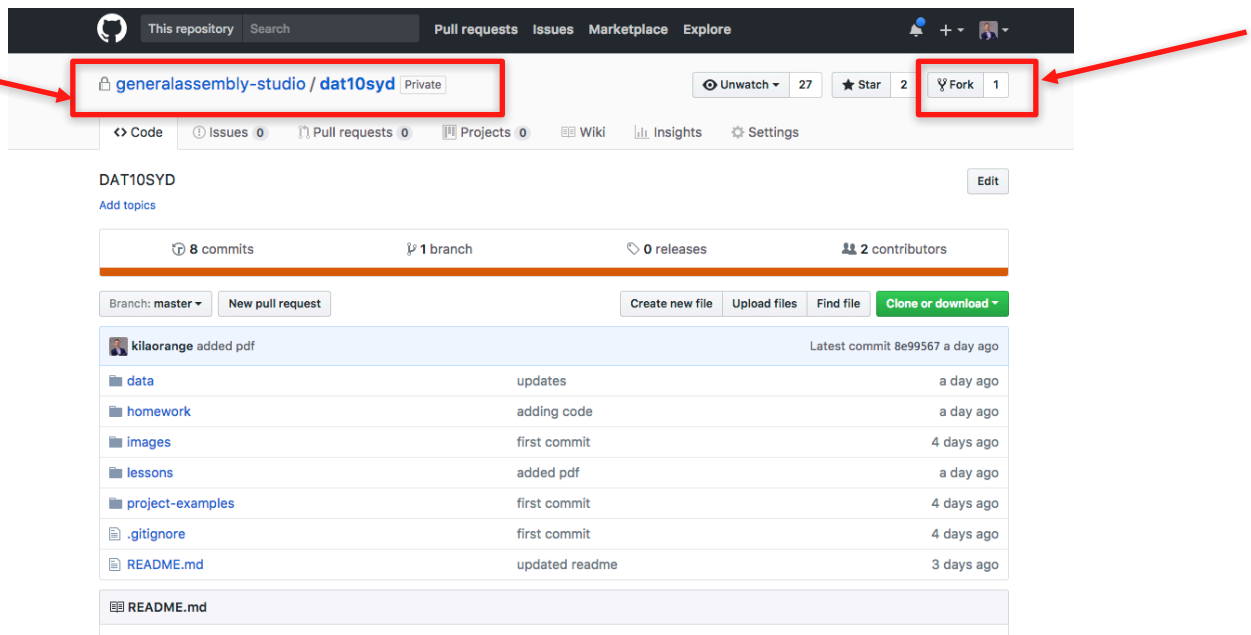
Using Git


Your Copy of the Course Repo
1. Fork
Course Repo Master
6. Pull Request
2. Clone
5. Push
(Not required in this course)
Pull Request invite the owner of the master repo to accept your changes.
Your Laptop
3. Update files
4. Commit

Preparing and Using Your Local Git

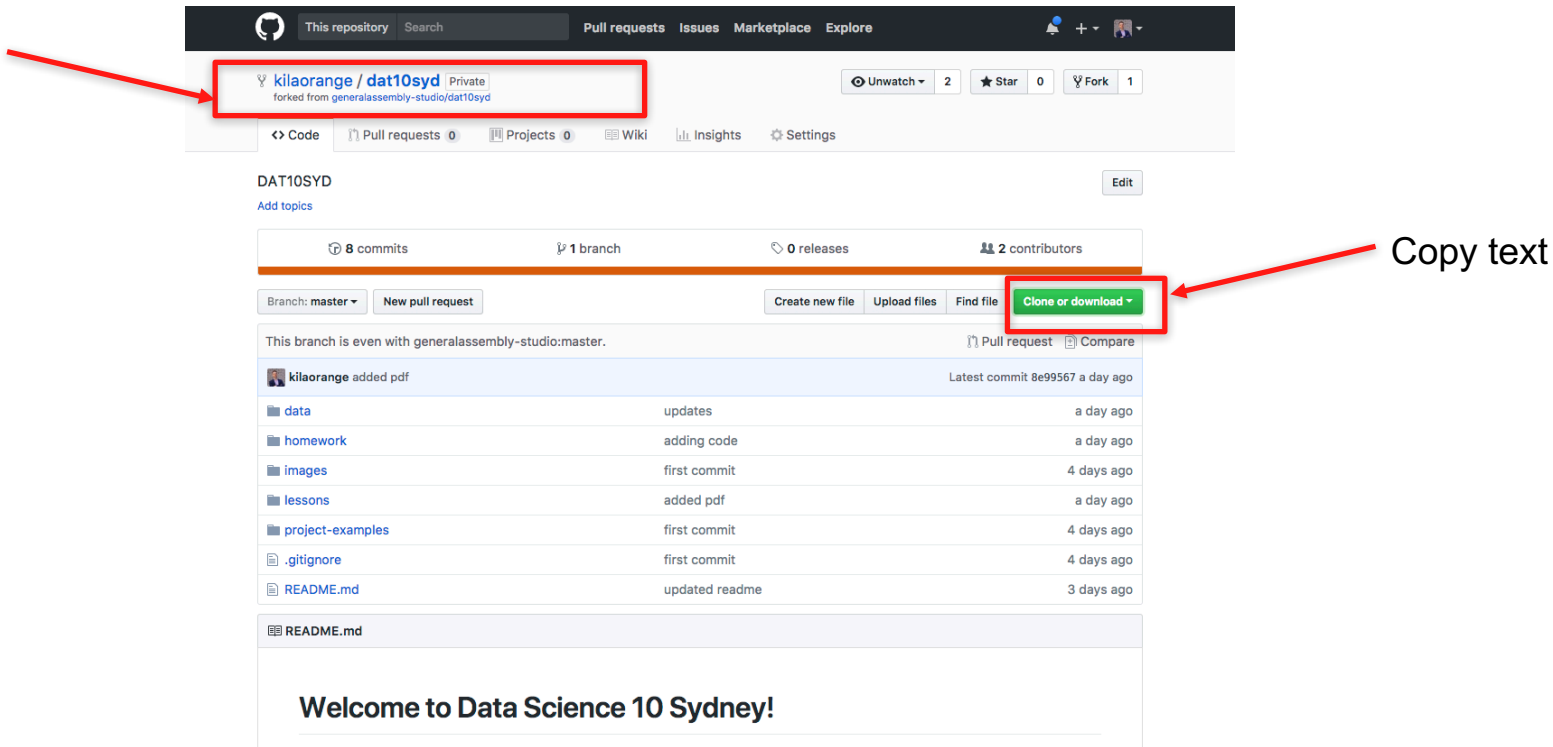
1. Improvement: delete the local Repo we cloned last week
2. Fork the GA Repo to your own GitHub account https://github.com/generalassembly-studio/dat11syd.git

dat11syd



Preparing and Using Your Local Git

1. It now looks like this…



Preparing and Using Your Local Git

1. Fork the master repo to your GitHub account
2. Open a terminal (Mac, Linux) or git bash (Windows)
3. Change working directory to where you want the local repo

    $ cd yourpath

1. Clone

    $ git clone https://github.com/***yourgithubname***/yourgithubrepo

Preparing and Using Your Local Git:
Configure Automatic Updates (Optional)

5. Configure your local clone to point to the official course repository

```
$ git remote –v
        origin https://github.com/pgoodall1984/dat11syd.git (fetch)
        origin https://github.com/pgoodall1984/dat11syd.git (push)

$ git remote add upstream https://github.com/generalassembly-studio/dat11syd


$ git remote -v

  origin      https://github.com/pgoodall1984/dat11syd.git (fetch)
  origin      https://github.com/pgoodall1984/dat11syd.git (push)
  upstream    https://github.com/generalassembly-studio/dat11syd (fetch)
  upstream    https://github.com/generalassembly-studio/dat11syd (push)
```

Download new material from official course repo (upstream) and merge it

1. Ensure you're in the master branch

   $ git checkout master

1. Grab the latest changes from the master

   $ git fetch upstream

1. Merge the master changes with your repo

   $ git merge upstream/master

WARNING: Be careful not to overwrite files you have already changed in your repo unless you want to replace them with the master versions!
(Consider renaming yours or doing a PULL REQUEST.)

Commit Changes to Your Local Git and Push to Your Master Repo

1. Commit

```
$ git status                          # show changes
$ git add filename                    # stages one file
$ git add .                           # stages all changed files
$ git commit -m your comments         # commits file(s), with comments
$ git status                          # check
```

1. Push

```
$ git push origin master              # push request auto accepted
```

origin = your GitHub repo (forked from course repo)
master = course repo

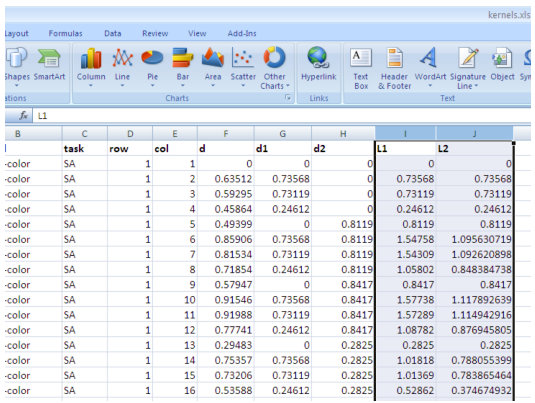
SYNCHING YOUR FORK WITH THE COURSE REPO

update

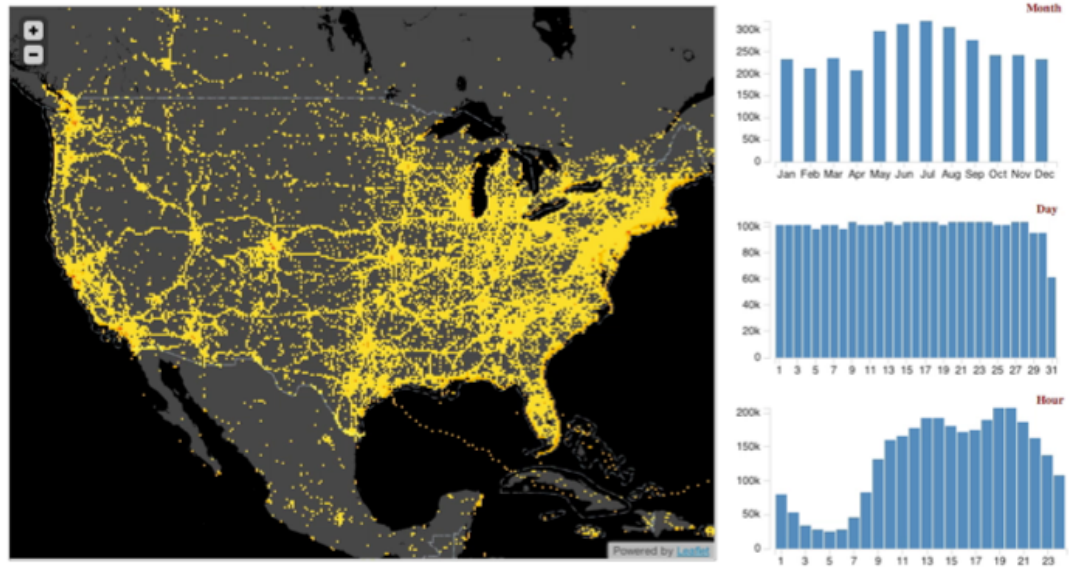
1. Copy your labs from the dat11syd/lessons dir to your dat11syd/students dir
2. cd <path to the root of your dat11syd local repo>
3. commit your changes ahead of sync
   - git status
   - git add .
   - git commit -m “descriptive label for the commit”
   - git status
4. download new material from official course repo (upstream) and merge it
   - git checkout master (ensures you are in the master branch)
   - git fetch upstream
   - git merge upstream/master

DATA SCIENCE PART TIME COURSE

WHAT IS DATA VISUALISATION?



	task	row	col	d	d1	d2	L1	L2
-color	SA	1	1	0	0	0	0	0
-color	SA	1	2	0.63512	0.73568	0	0.73568	0.73568
-color	SA	1	3	0.59295	0.73119	0	0.73119	0.73119
-color	SA	1	4	0.45864	0.24612	0	0.24612	0.24612
-color	SA	1	5	0.49399	0	0.8119	0.8119	0.8119
-color	SA	1	6	0.85906	0.73568	0.8119	1.54758	1.095630719
-color	SA	1	7	0.81534	0.73119	0.8119	1.54309	1.092620898
-color	SA	1	8	0.71854	0.24612	0.8119	1.05802	0.848384738
-color	SA	1	9	0.57947	0	0.8417	0.8417	0.8417
-color	SA	1	10	0.91546	0.73568	0.8417	1.57738	1.117892639
-color	SA	1	11	0.91988	0.73119	0.8417	1.57289	1.114942916
-color	SA	1	12	0.77741	0.24612	0.8417	1.08782	0.876945805
-color	SA	1	13	0.29483	0	0.2825	0.2825	0.2825
-color	SA	1	14	0.75357	0.73568	0.2825	1.01818	0.788055399
-color	SA	1	15	0.73206	0.73119	0.2825	1.01369	0.783865464
-color	SA	1	16	0.53588	0.24612	0.2825	0.52862	0.374674932

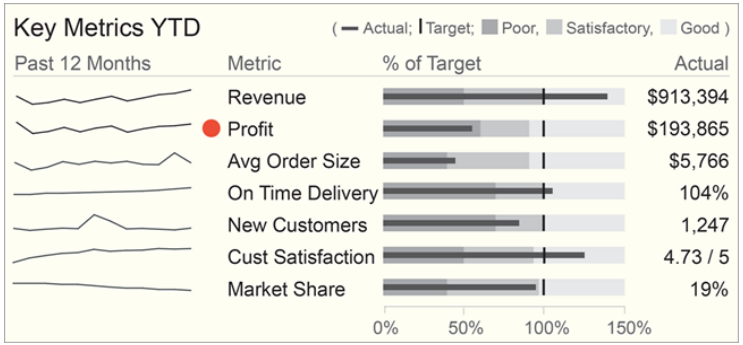


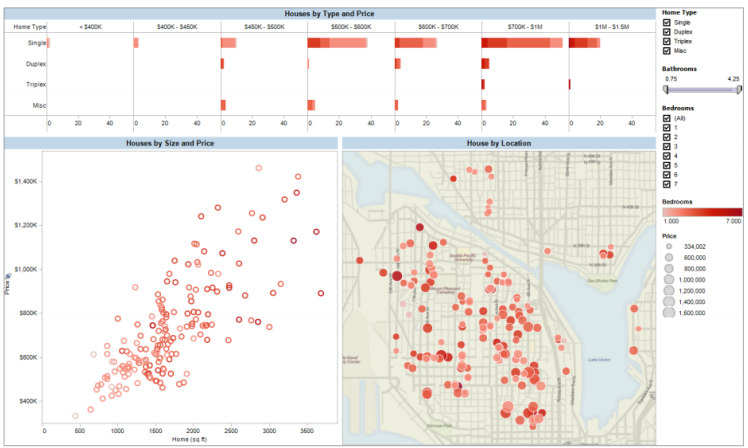
- Present information that is intuitive and clear for the viewer
- Turn numbers in a spreadsheet into something people can interpret and extract insights

WHY VISUALISE DATA?

Reporting

- Dashboards and Business Intelligence
- Know the questions you want answers to
- Can detect changes from the norm
- Good for taking a 30,000 foot view of the problem

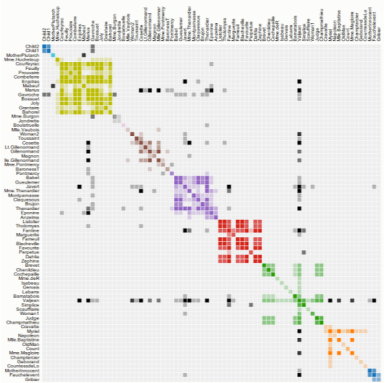


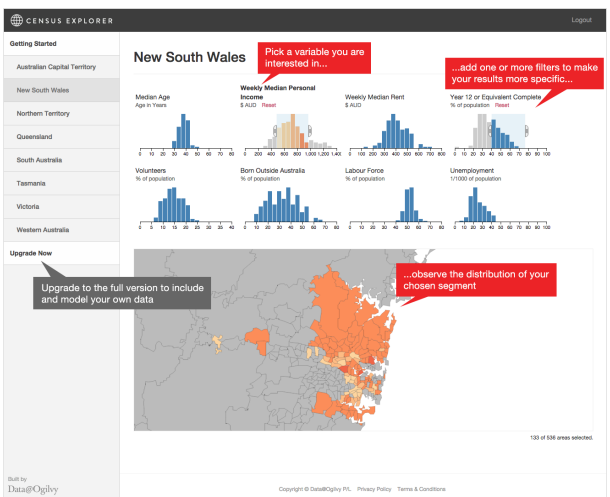


WHY VISUALISE DATA?

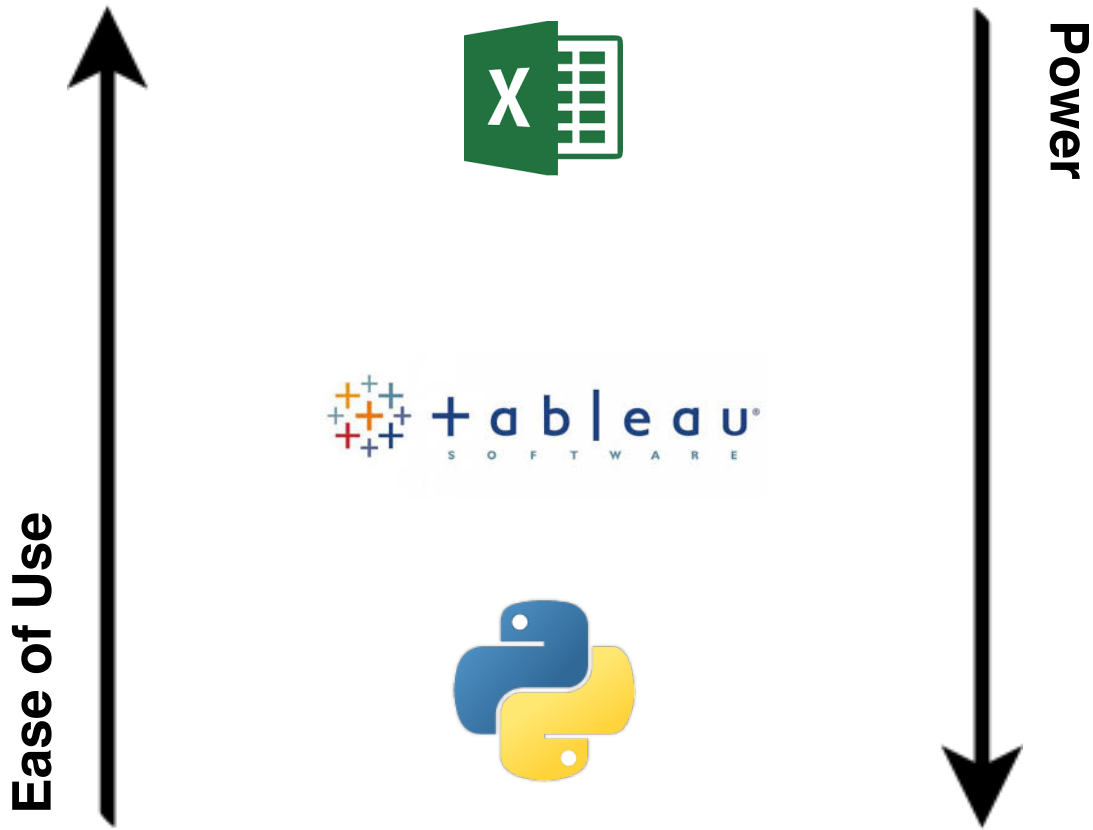
Exploring

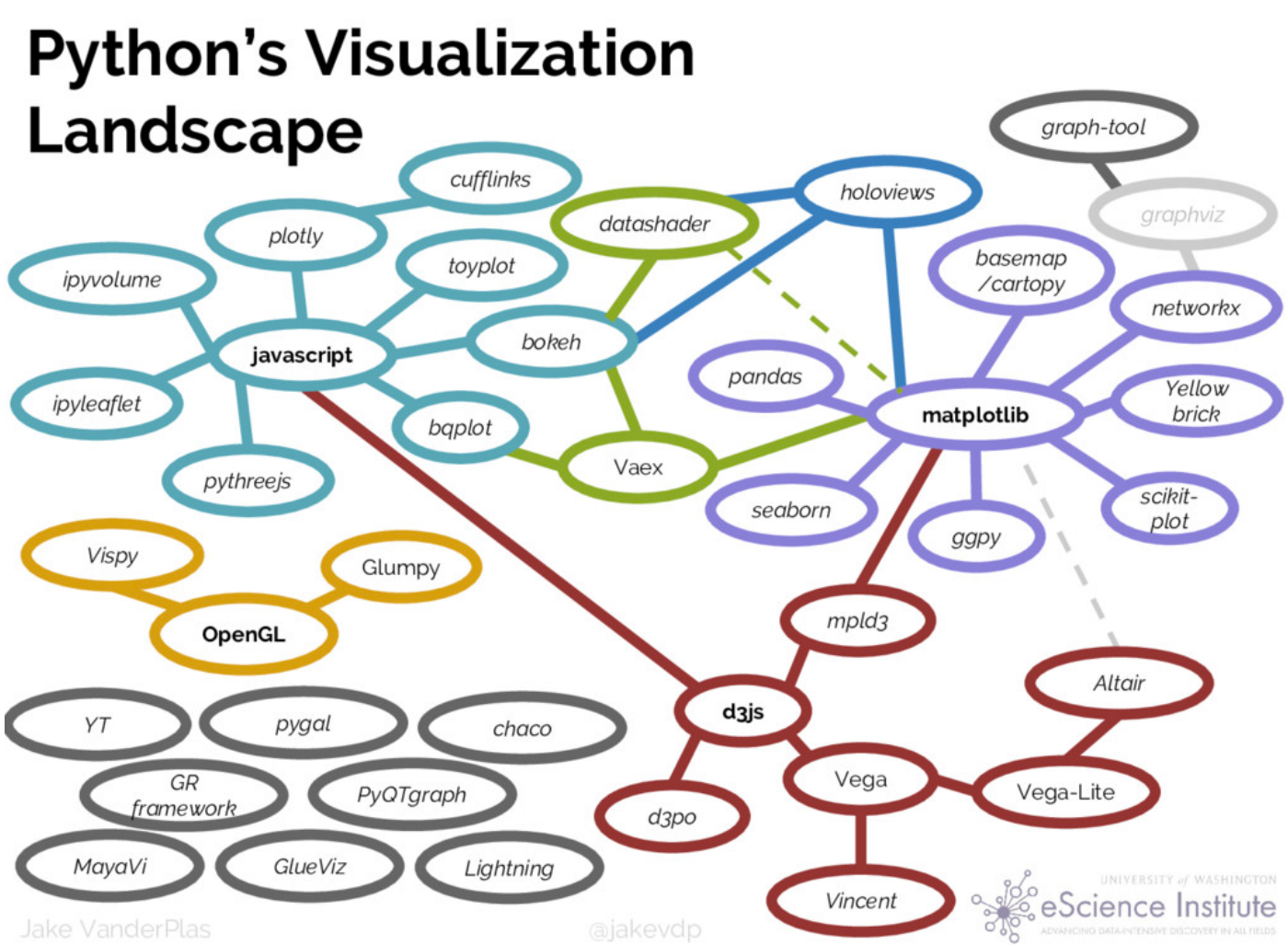
- Exploratory Data Analysis
- Combines multiple data sources for single view of a problem
- Technical analysis of data
- Combined with modelling allows for the discovery of new problems and solutions





HOW DO WE VISUALISE DATA?





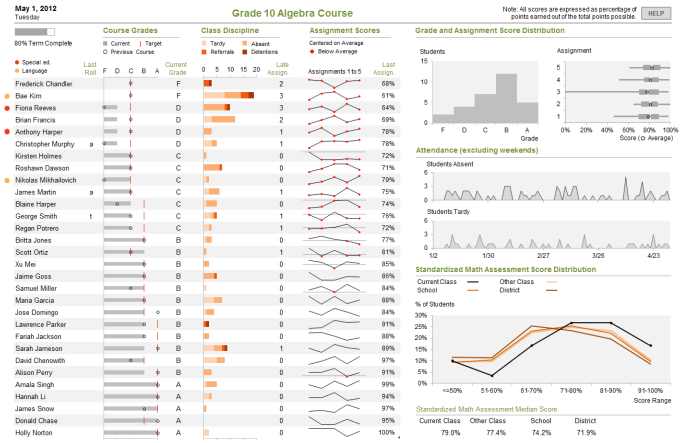
Jake VanderPlas, https://speakerdeck.com/jakevdp/pythons-visualization-landscape-pycon-2017

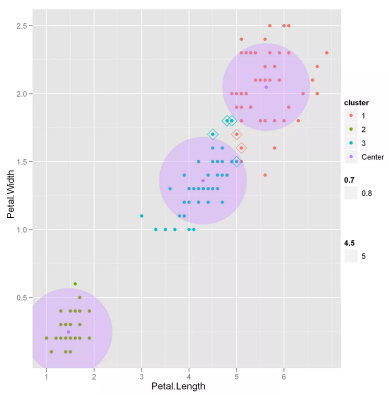
TOOL TRADEOFFS

	Easy to Use	Powerful
Advantages	• Provides a useful starting point • Familiar to a large audience • Prototyping and design time is reduced • Default settings reduce the options and thinking that goes into producing a graph	• Scales to larger datasets • Customised visualisations can create engaging visualisations • Open-source (so free to run and extend) • Non-obvious insights can be discovered with modelling tools • Re-use code to produce similar charts for different data
Disadvantages	• Reproducing analysis requires lots of manual effort • Limited to relative small data sets • Solves known problems and cannot answer complex questions • Licensing can be expensive	• Requires specialist skills to produce a graph • Training and education for some of the output might be necessary

THE VALUE OF DATA VISUALISATION

- Communicate what's happening within the business
- Support decisions with information
- Measure and report the impact of decisions
- Discover ways to improve the business





Enough talking…

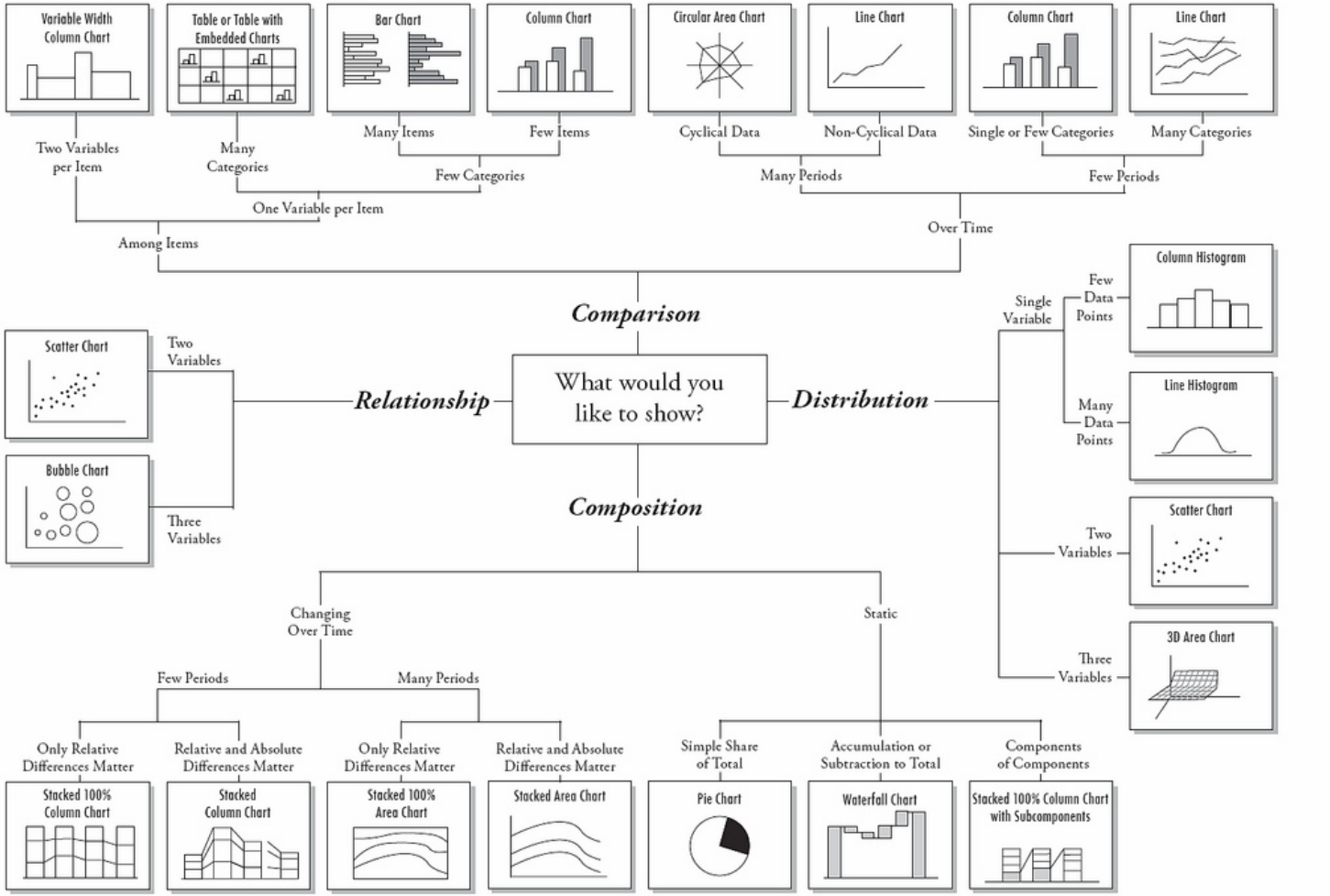
Let's do some lab work:

dat11syd/lessons/lesson-03/lab/lesson-03-lab-01_ReadingData

DATA SCIENCE PART TIME COURSE

DIFFERENT TYPES OF CHARTS

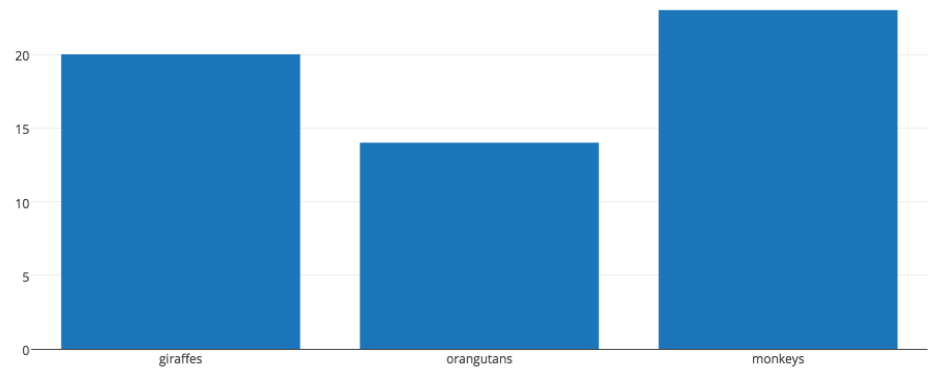
Chart Suggestions—A Thought-Starter



BAR CHART

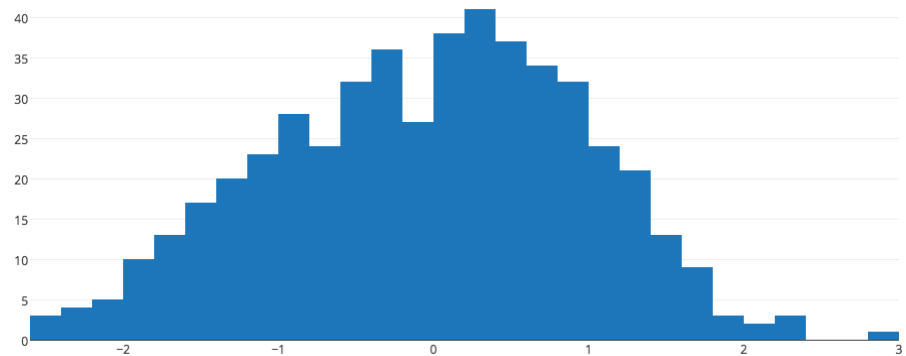
- Shows numeric summaries across different categories (either horizontally or vertically)
- Each bar represents a different category in the data

```
import plotly.plotly as py
import plotly.graph_objs as go

data = [go.Bar(
            x=['giraffes', 'orangutans', 'monkeys'],
            y=[20, 14, 23]
    )]
py.iplot(data, filename='basic-bar')
```



HISTOGRAM

- Shows the distribution of data over a continuous interval
- Allows us to see the shape of our data

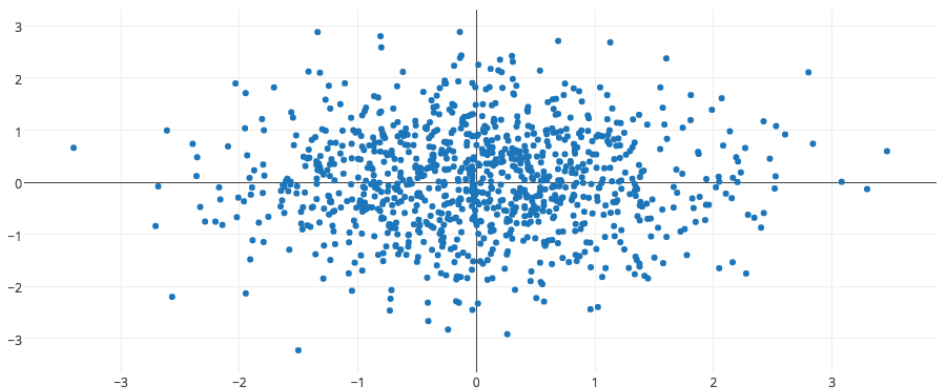



```
import plotly.plotly as py
import plotly.graph_objs as go

import numpy as np

x = np.random.randn(500)
data = [go.Histogram(x=x)]

py.iplot(data, filename='basic histogram')
```

SCATTER PLOT

- Shows values between two variables, one on each axis.
- Used to see a relationship between variables

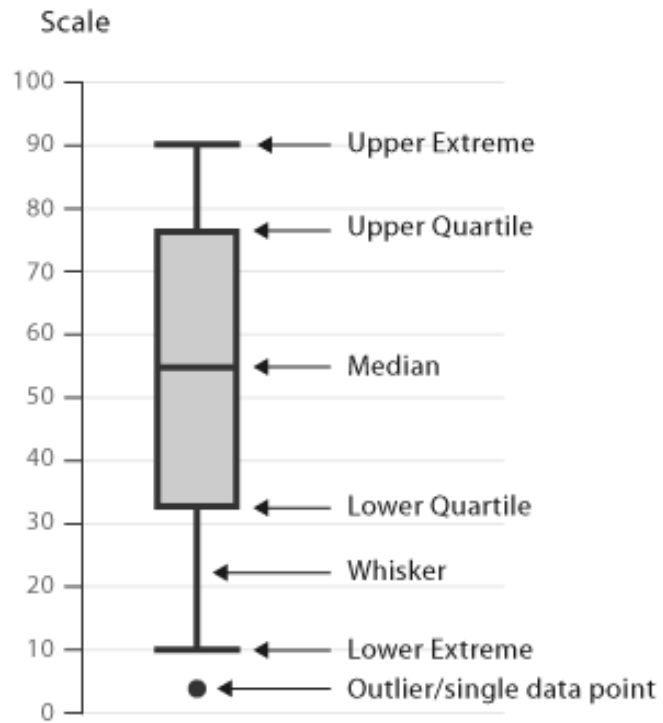
```
import numpy as np
N = 1000
random_x = np.random.randn(N)
random_y = np.random.randn(N)
trace = go.Scatter(
    x = random_x,
    y = random_y,
    mode = 'markers'
)
data = [trace]
py.iplot(data, filename='basic-scatter')
```



BOX AND WHISKER PLOT

- Displays numerical distribution summaries by groups through quartiles
- Can compare different distributions

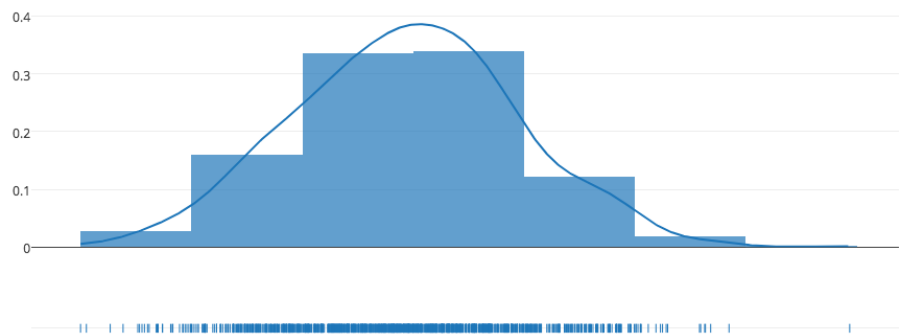
```
import numpy as np
y0 = np.random.randn(50)-1
y1 = np.random.randn(50)+1
trace0 = go.Box(
    y=y0
)
trace1 = go.Box(
    y=y1
)
data = [trace0, trace1]
py.iplot(data)
```



http://www.datavizcatalogue.com/methods/box_plot.html

DENSITY PLOT

- Similar to a histogram but smooths out the distribution with a continuous line
- Not affected by bin choices

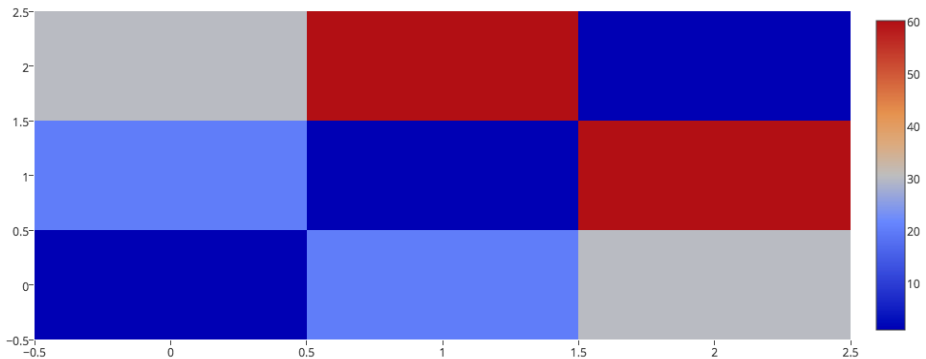



```
import numpy as np

x = np.random.randn(1000)
hist_data = [x]
group_labels = ['distplot']

fig = ff.create_distplot(hist_data, group_labels)
py.iplot(fig, filename='Basic Distplot')
```

HEATMAP

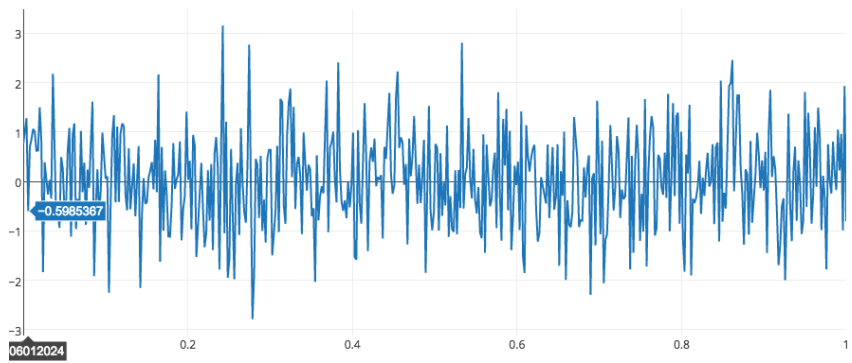
- Colour coding applied to tabular data.
- Provides a generalised view of the data by each cell

```
trace = go.Heatmap(z=[[1, 20, 30],
                      [20, 1, 60],
                      [30, 60, 1]])
data=[trace]
py.iplot(data, filename='basic-heatmap')
```



LINE GRAPH

- Used to display a numeric value over a continuous value or time
- Used to observe trends and changes over time




```
import numpy as np
N = 500
random_x = np.linspace(0, 1, N)
random_y = np.random.randn(N)
trace = go.Scatter(
    x = random_x,
    y = random_y
)
data = [trace]
py.iplot(data, filename='basic-line')
```

- Allows us to plot points geographically
- We can overlay information on a map, usually loaded as a collection of ‘tiles’.

```
import folium


map_object = folium.Map(location=[-33.8, 151.2], zoom_start=6,
tiles="Stamen toner")

marker = folium.features.Marker([-33.869824, 151.206423],
popup=“General Assembly!”)

map_object.add_children(marker)
```

Enough talking…

Let's do some lab work:

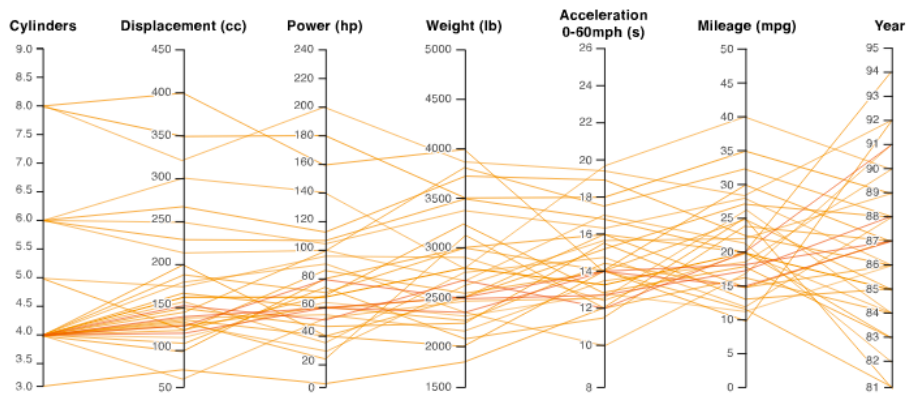
dat11syd/lessons/lesson-03/lab/lesson-03-lab-02_Visualisation

DATA SCIENCE PART TIME COURSE

DIFFERENT TYPES OF CHARTS

PARALLEL COORDINATES

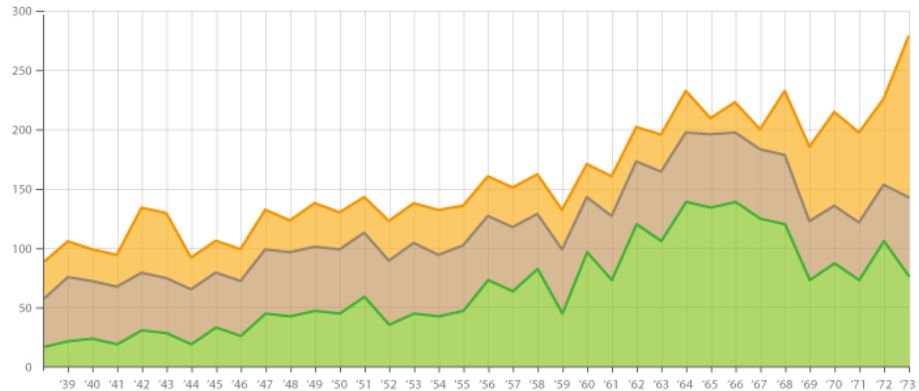
- Plot multiple numeric variables across each observation
- Each axis is scaled and each line through the graph is an observation



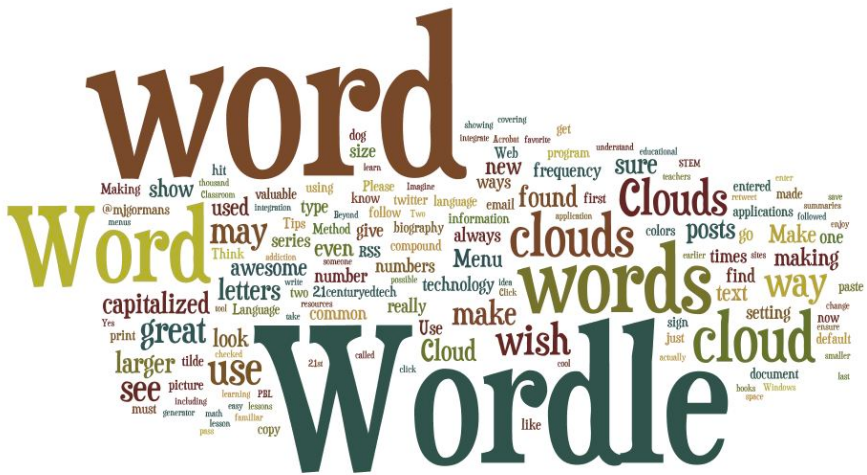

```
import plotly.plotly as py
import plotly.graph_objs as go

data = [go.Bar(
            x=['giraffes', 'orangutans', 'monkeys'],
            y=[20, 14, 23]
    )]
py.iplot(data, filename='basic-bar')
```

STACKED AREA CHARTS

- Try to track changes in a numeric variable across multiple categories
- The area under each category is represented by different colours
- Better to plot multiple lines rather than stacking the area



- Collection of words in a text or summary of a corpus
- Better to apply some form of categorical detection otherwise similar words will be lots of collection



PIE CHARTS

- Represent a proportion that a category makes up of the whole
- Better to use column charts in most cases

DATA SCIENCE - Week 3 Day 1

DISCUSSION TIME

- Review of last week
- Further Reading for Data Visualisation
- Homework due Friday 24th Nov
- Check in with course project
- Pre-reading for next lesson

WEEK 1 Review

DISCUSSION TIME

- Course Overview
- Data Science Overview
- Pre-work
- Python basics
- Git Basics

DATA SCIENCE - Further Reading

DISCUSSION TIME

Further Reading

Edward Tufte, The Visual Display of Quantitative Information

Leland Wilkinson, The Grammar of Graphics

Scott Murray, Interactive Data Visualisation for the Web (free online)

flowingdata.com

DATA SCIENCE - Week 2 Day 1

DISCUSSION TIME

Homework/Course/Project

How’s it going ?

DATA SCIENCE - Week 2 Day 1

PRE-READING

An Introduction to Statistical Learning

Chapter 3 - Linear Regression