GENERAL ASSEMBLY

DATA SCIENCE

SYD DAT 6

Week 2 - Data Visualisation
Monday 17th October

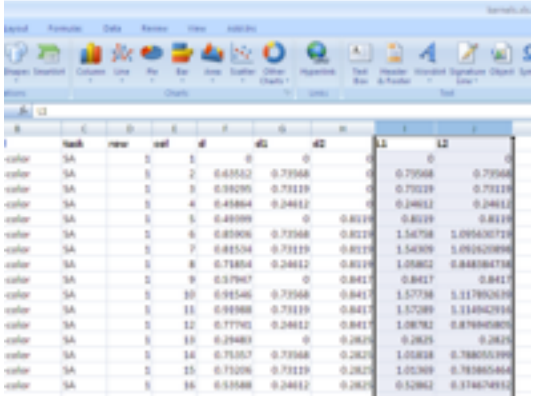
AGENDA

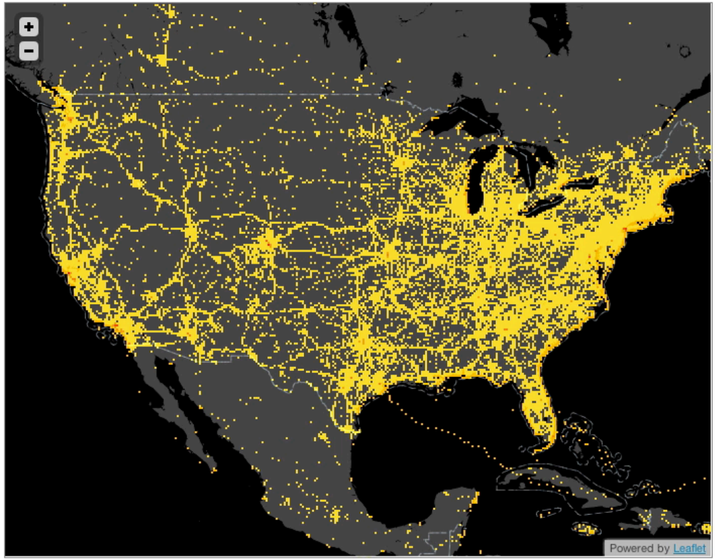
1. What is Data Visualisation?
2. Why do we visualise data?
3. How do we visualise data?
4. Tools for visualising data
5. Git - retrieving new materials
6. Git - make changes, push to origin and make a pull request
7. Lab - visualisation
8. Discussion

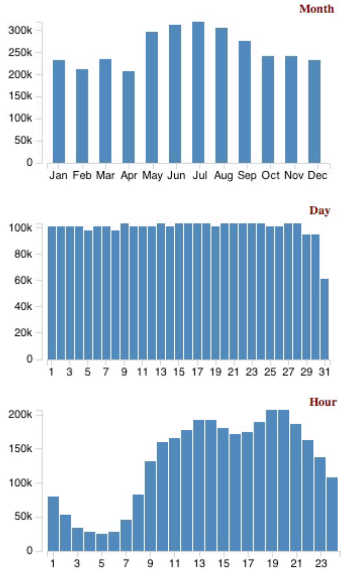
DATA SCIENCE PART TIME COURSE

WHAT IS DATA VISUALISATION?

WHAT IS DATA VISUALISATION?





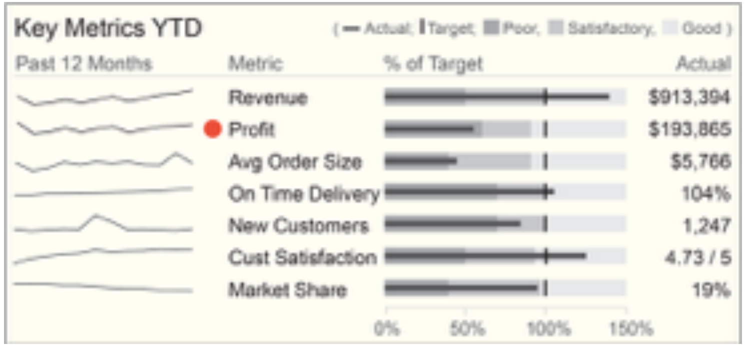


- Present information that is intuitive and clear for the viewer
- Turn numbers in a spreadsheet into something people can interpret and extract insights

WHY VISUALISE DATA?

Reporting

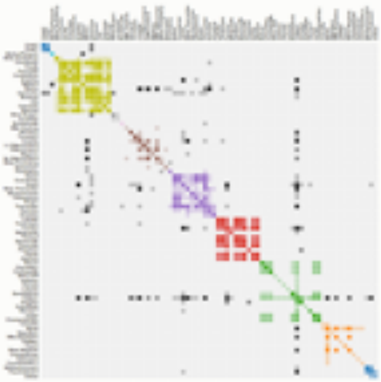
- Dashboards and Business Intelligence
- Know the questions you want answers to
- Can detect changes from the norm
- Good for taking a 30,000 foot view of the problem

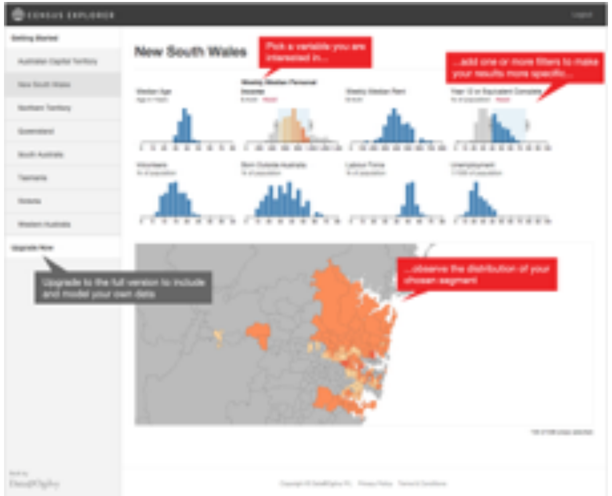


WHY VISUALISE DATA?

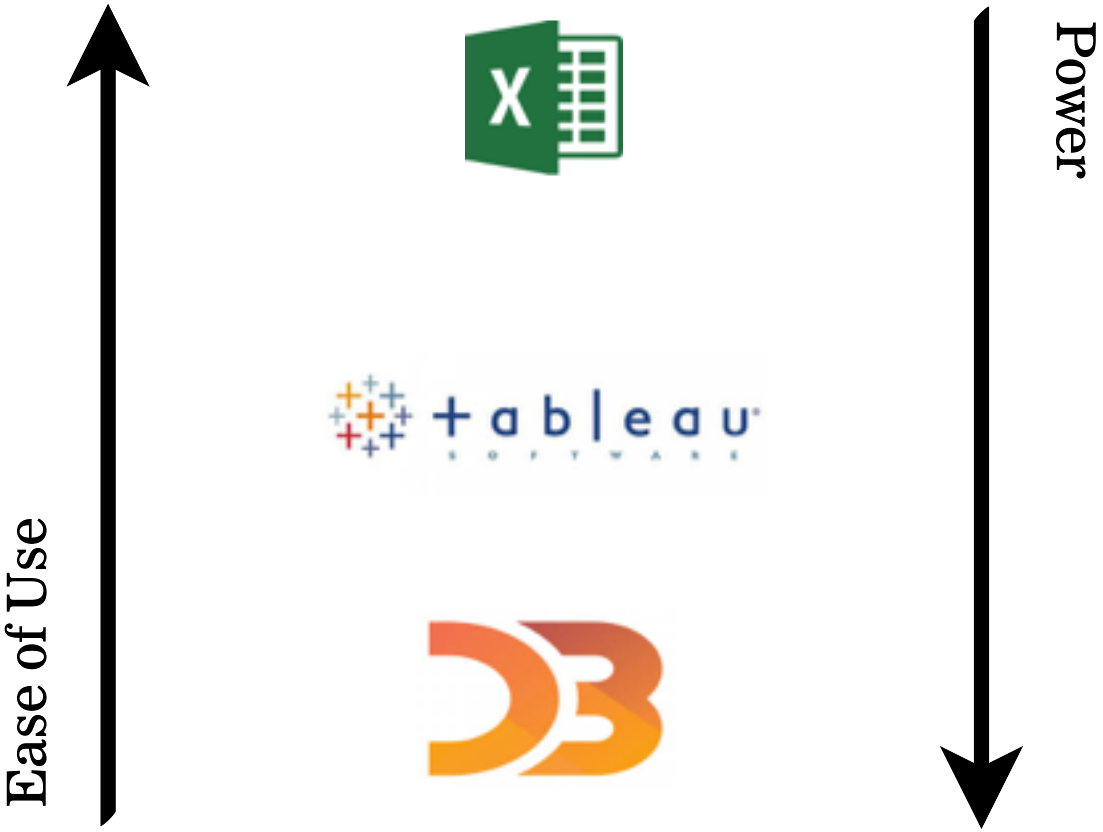
Exploring

- 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




CENSUS EXPLORER
Logout
Getting Started
Australian Capital Territory
New South Wales
Northern Territory
Queensland
South Australia
Tasmania
Victoria
Western Australia
Upgrade Now
New South Wales
Pick a variable you are interested in...
...add one or more filters to make your results more specific...
Median Age
Weekly Median Personal Income
Weekly Median Rent
Year 12 or Equivalent Completion
Volunteers
Born Outside Australia
Labour Force
Unemployment
Upgrade to the full version to include and model your own data
...observe the distribution of your chosen segment

HOW DO WE VISUALISE DATA?



Ease of Use

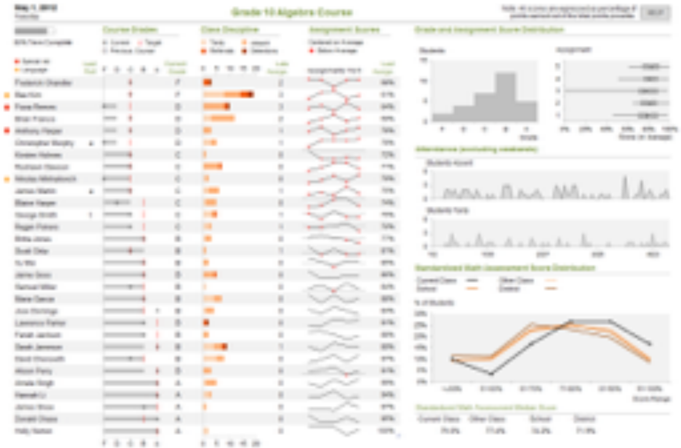
Power

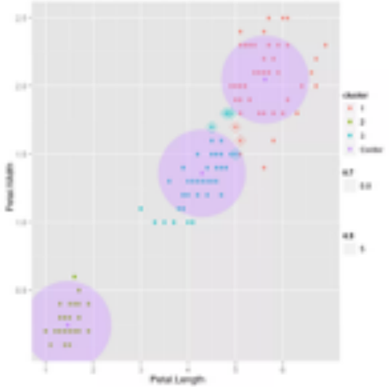
TOOL TRADEOFFS

	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



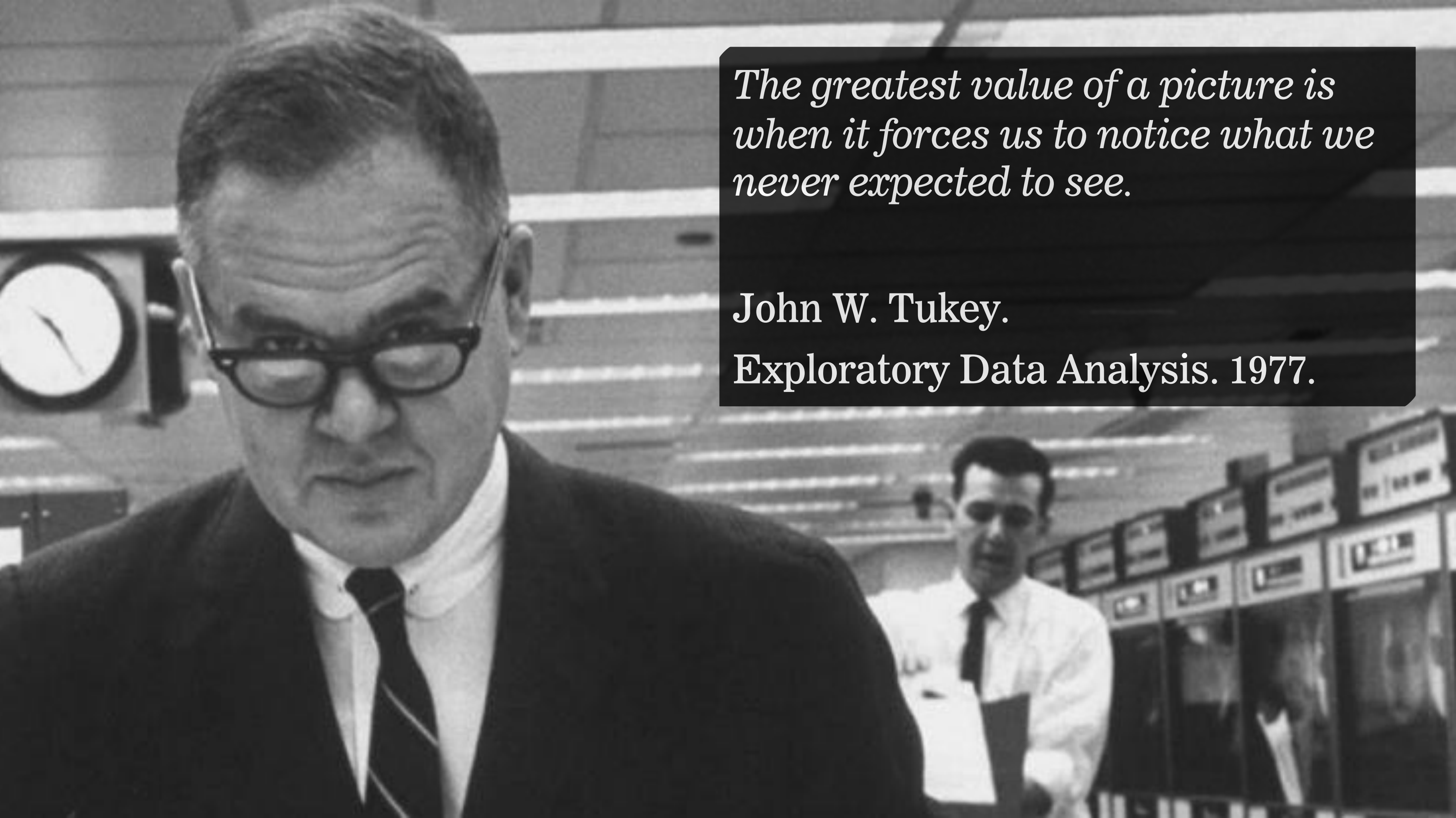


INTERESTING SITES

- http://idl.cs.washington.edu/
- https://www.windytv.com/?-33.459,151.260,6
- http://www.nytimes.com/interactive/2016/upshot/presidential-polls-forecast.html
- http://junkcharts.typepad.com/junk_charts/2014/11/a-rule-breaking-cliche-defying-punch-carrying-chart-worthy-of-the-election.html
- http://flowingdata.com
- tools available: http://selection.datavisualization.ch/

DATA SCIENCE PART TIME COURSE

Philosophy in Data Analysis



The greatest value of a picture is when it forces us to notice what we never expected to see.

John W. Tukey.

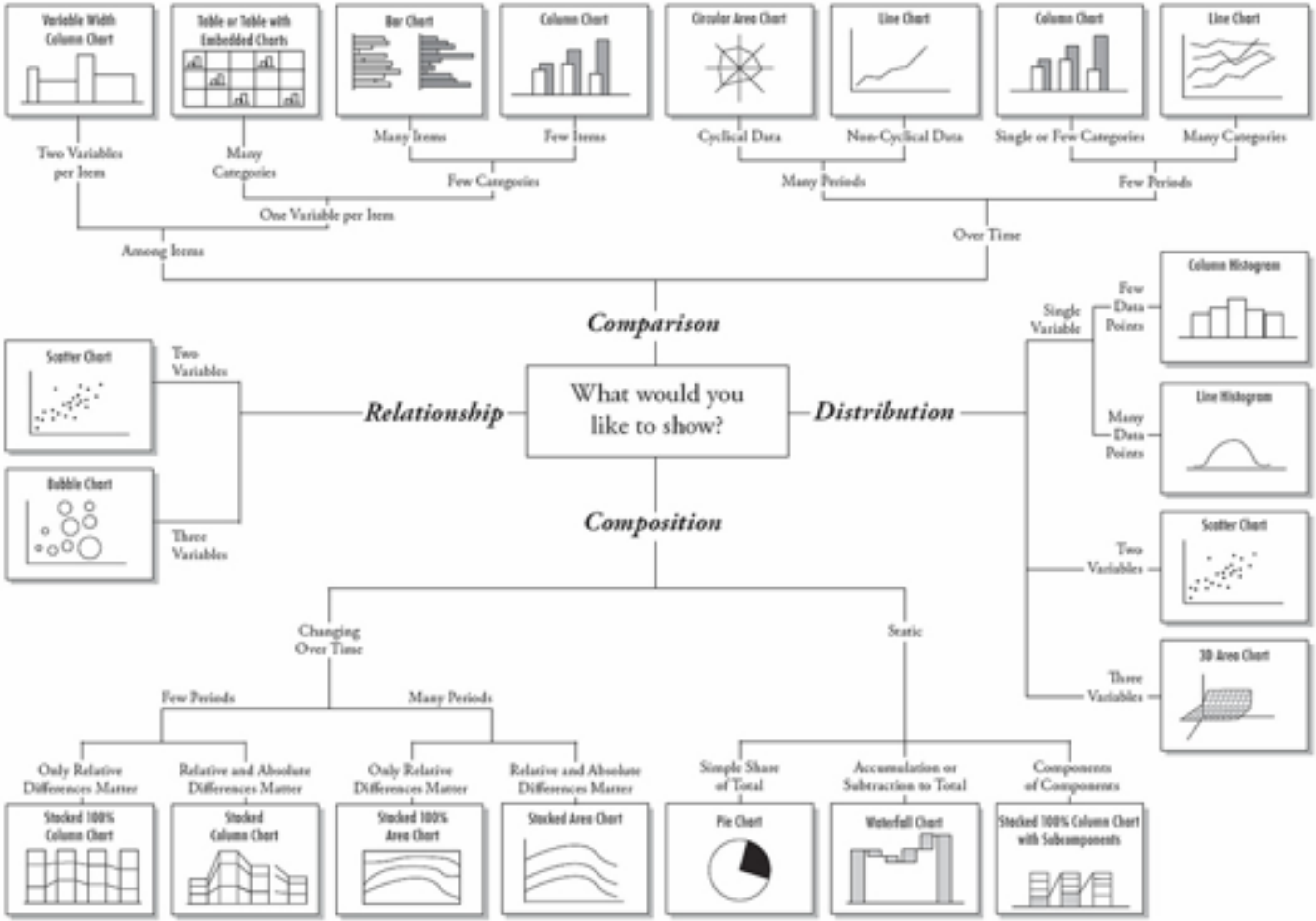
Exploratory Data Analysis. 1977.

Philosophy on Data Analysis

> “More data analysis efforts seem to go bad because of an excess of sophistication rather than a lack of it.” Phillip K Janert

There is nothing wrong with speaking of the “range over which points spread”. Once we start talking about “standard deviations,” this clarity is gone.

Chart Suggestions—A Thought-Starter



What would you like to show?

Comparison

Among Items

Two Variables per Item

One Variable per Item

Many Categories

Few Categories

Many Items

Few Items

Over Time

Many Periods

Cyclical Data

Non-Cyclical Data

Few Periods

Single or Few Categories

Many Categories

Relationship

Two Variables

Three Variables

Distribution

Single Variable

Few Data Points

Many Data Points

Two Variables

Three Variables

Composition

Changing Over Time

Few Periods

Only Relative Differences Matter

Relative and Absolute Differences Matter

Many Periods

Only Relative Differences Matter

Relative and Absolute Differences Matter

Static

Simple Share of Total

Accumulation or Subtraction to Total

Components of Components

Index	Cuisine	Price	Rating
0	Mexican	$68	1
1	Italian	$58	1
2	Thai	$86	3
3	Mexican	$63	4
4	Thai	$89	3
5	Thai	$14	3
6	Thai	$25	3
7	Mexican	$37	1
8	Mexican	$15	1
9	Italian	$33	2
10	Italian	$72	4

DATA SCIENCE PART TIME COURSE

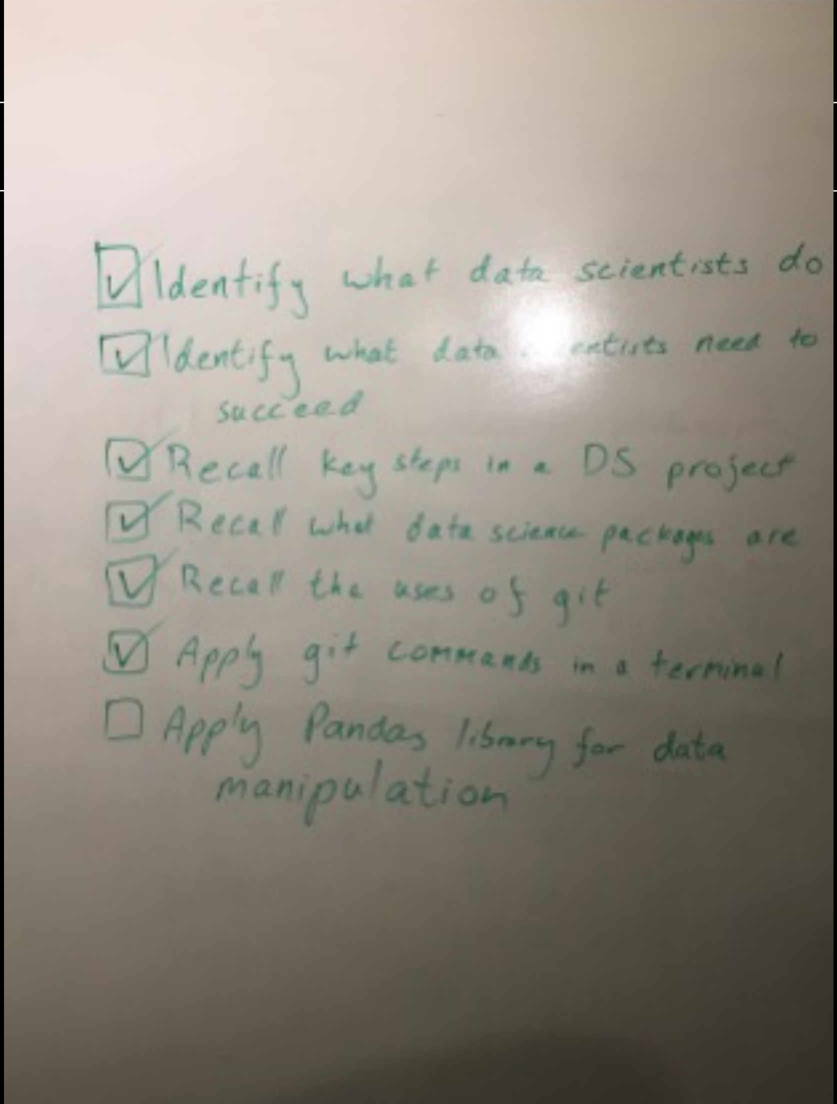
Git
handling changes

DATA VISUALISATION LAB

DISCUSSION TIME

- Review of last week
- Further Reading for Data Visualisation
- Check in with homework/course project

WEEK 1



☑ Identify what data scientists do

☑ Identify what data scientists need to succeed

☑ Recall key steps in a DS project

☑ Recall what data science packages are

☑ Recall the uses of git

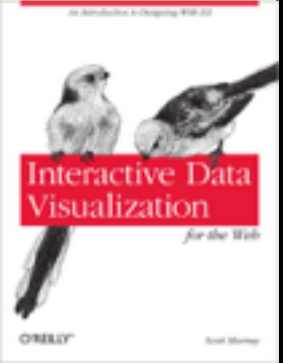
☑ Apply git commands in a terminal

□ Apply Pandas library for data manipulation

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
- New York Times (Upshot)


An Introduction to Designing With D3
Interactive Data Visualization
for the Web
O'REILLY
Scott Murray

DATA SCIENCE – Week 2 Day 1

DISCUSSION TIME

Homework/Course Project

- Homework1.ipynd - due Friday
- Read Chapter 3 of Introduction to Statistical Learning - Linear Regression
- Course Project: Prepare 3 concepts for a project