WhiteHat Session 3 - Tools of the Trade

6 Hours

Overview

In this module we will be covering various tools apprentices will need to enhance their analytics performance in their roles. In the first half of the session we will give an overview of Big Data, considering how it works and what the benefits and drawbacks are of using it. This is followed by a discussion around data analytics platforms and comparing their usage to coding it yourself.

In the second part of the module we will be looking into the statistical programming language R, briefly looking at how many of the processes we learned in Python can be performed in R.

Prerequisites

- Apprentices will need to have R installed on their computer (link)
- They should also install RStudio (link)

Learning objectives

- Understand fundamental concepts of **Big Data**
- Justify the use (or lack of) Big Data technologies **in your analysis**
- Critically evaluate the differences in **using a data platform** and **coding it yourself**
- Understand how **R** can be utilised in your analysis

Assignment

Task

Justify the use (or lack of) Big Data technologies in your role

What we are looking for

A written report (max 1500 words) detailing how Big Data technologies are used in your role and the business impact this decision has. Consider also how Big Data technology could be used in your role and what the potential business implications would be. Additionally, you should consider how use of these technologies comply with GDPR and other policies your workplace has around data protection.

Technical Knowledge

- Identify, collect and migrate data to/from a range of internal and external systems (TC1)
- Interpret and apply the organisations data and information security standards, policies and procedures to data management activities (TC3)
- Perform routine statistical analyses and ad-hoc queries (TC6)
- Use a range of analytical techniques such as data mining, time series forecasting and modelling techniques to identify and predict trends and patterns in data (TC7)
- Apply the tools and techniques for data analysis, data visualisation and presentation (TC8)
- Works with organisation's data architecture (TC11)
- Assist with data quality checking and cleansing (TC12)

Skills, Attributes and Behaviours

- Logical and creative thinking skills (SAB1)
- A thorough and organised approach (SAB5)
- Maintain productive, professional and secure working environment (SAB8)

KM1 Data Analysis Tools Syllabus

- Explain the nature and challenges of data volumes being process through integration activities and how a programming approach can improve this:
    - Big Data (1.5)
- Capabilities and functions of statistical programming language: R (2.1)

Recordings (for coach reference)

https://drive.google.com/drive/folders/1uXw2RfTfzVR1QEMRoAiX6hvWiZn4xzR8?usp=sharing

Session Overview	Timing (approximate)
Class Introduction	10 Minutes
Understanding Big Data	20 Minutes
The Five V's	25 Minutes
Developing the Future	20 Minutes
How Does it Work?	10 Minutes

Break	10 Minutes
Using Big Data in Your Role	10 Minutes
Advantages and Disadvantages	20 Minutes
Big Data Products	5 Minutes
Data Platforms	20 Minutes
Platform vs Coding Yourself	20 Minutes
Setting up R	5 Minutes
Session 1 Recap	5 Minutes
Session 2 Intro	10 Minutes
Features of R	30 Minutes
Control Flow in R	30 Minutes
EDA in R	35 Minutes
Break	15 Minutes
Visualisation in R	35 Minutes
Linear Modelling in R	15 Minutes
RStudio	5 Minutes
Session 2 Recap	5 Minutes

Additional Resources

- Basic R Tutorial

Topic	Class Introduction	Duration	10 minutes

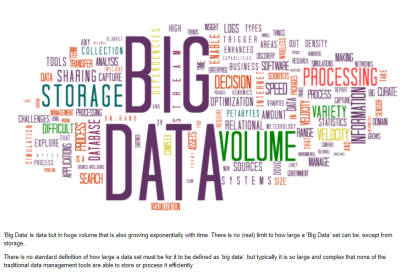
Objectives
• To provide an overview of the class agenda and the expected learning objectives

Section No	Section	Notes	Timing
1	**Session 1: Tools of the Trade** Data Analytics is taking over the world and many companies are taking the opportunity to develop products and softwares to help companies leverage their data more efficiently without having to upskill their staff. In this session we will be learning about the fundamentals of Big Data and how it can impact your analysis. We will also consider the various data tools available to help you write about them in your portfolio. **Session Outline** • Understanding Big Data • The Five V's • Developing The Future • How It Works • Advantages and Disadvantages • Big Data Products • Data Platforms • Platforms vs Coding Yourself • Setting Up R • Recap **Learning Objectives** • Understand fundamental concepts of Big Data • Justify the use (or lack of) Big Data technologies in your analysis • Critically evaluate the differences in using a data platform and coding it yourself	Coach welcome to session, run an ice breaker from here. Run through session plan and learning objectives	

Topic	Understanding Big Data	Duration	20 Minutes

Objectives
• Define what we mean by 'Big Data'

Section No	Section	Notes	Timing
1	 Top **Aim:** Define what we mean by 'Big Data'	The first thing we will be covering today is the fundamentals of Big Data and why it is important to you and your organisation.	
2	**What is Big Data?**	Ask apprentices what their understanding of Big Data is (through annotate, chat, etc). Some follow up questions: • How would you use it in a working environment? • What type of data do we mean by 'Big Data'? (i.e. structured/unstructured/ quantitative/qualitative etc)	

		• Does your organisation have a Big Data strategy?	
3	**Big Data is data whose scale, distribution, diversity and or timeliness require the use of new technical architectures and analytics to enable insights that unlock new sources of business value.** *McKinsey & Co.; Big Data: The Next Frontier for Innovation, Competition, and Productivity*	McKinsey's definition of Big Data implies that organisations will need new data architectures, new tools, methods and an integration of multiple skills to be competitive when it comes to recruiting talent and generating business insights. Data is created constantly and at an ever-increasing rate. Mobile phones, social media, imaging technologies- all these and more create new data and that must be stored somewhere. Merely keeping up with this huge influx of data is difficult, but what is even more challenging is analysing the vast amounts of it, especially when it does not conform to traditional notions of data structure. This makes identifying meaningful patterns and extracting useful information harder.	
4	 'Big Data' is data but in huge volume that is also growing exponentially with time. There is no (real) limit to how large a 'Big Data' set can be, except from storage. There is no standard definition of how large a data set must be for it to be defined as 'big data', but typically it is so large and complex that none of the traditional data management tools are able to store or process it efficiently.	'Big Data' is data but in huge volume, but also growing exponentially with time. Like the way the universe is continually expanding there is no (real) limit to how large a 'big data' set can be. Except from storage… There is no standard definition of how large a data set must be for it to be defined as 'big data', but typically it is so voluminous that traditional data processing software (tableau, excel, your laptop) can't manage it. Such data is so large and complex	

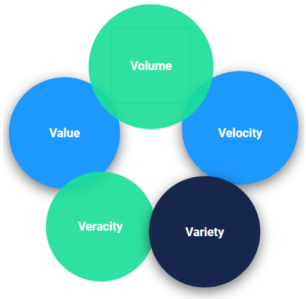
		that none of the traditional data management tools are able to store or process it efficiently.	
5	 Big Data is not just about size or scope but also the technology that supports it. Until recently analysing Big Data sets was impossible as the tools had not been invented!	Therefore, Big Data as a topic is not only about the size and scope of the data, but the technology that supports it. Until recently, analysing big data sets was impossible- the tools literally hadn't been invented! Let's look at some examples of Big Data. How many tweets do you think are sent every day?	
6	Generates 500 million Tweets **every day** Source: Internet Live Stats	Just think about the amount of data that is! Every character, link, image is data which Twitter is processing. Click on the source link to get live statistics	
7	NYSE Generates 1 TeraByte of trading data **every day** Source: New York Stock Exchange	Consider the NY stock exchange and the sheer volume of stocks and transactions that are getting processed. All this data needs to be stored somewhere!	
8	Generates 4 petabytes of user data **every day** Source: Facebook	Each of these companies generate an enormous amount of data every day. These examples show not just the size of the big data but the continuous growth of data that it represents. Question: What other kinds of data sources can you think of that generates data at this scale? [Could say email (est. 168 million per second), sms messages (est 11 million per second) or google search results (est 698445 per second) for example).	

		Current estimates are that 1820 TB of new data are created every second over the planet.	
9	**79% of enterprise executives agree that companies that do not embrace Big Data will lose their competitive position and face extinction** Source: Accenture	As Data Analysts it's important we are able to properly understand what Big Data means. Particularly because there are big opportunities for those companies that embrace these newtools, analytical methods and data architectures. Why should we care? As data analysts we have a lot more data sets to use to inform decisions in our companies. This shouldn't be surprising, think about all the devices in our homes today that gather data (Alexa, FitBit, CCTV, etc that produce 5 quintillion bytes of data daily worldwide).	

Topic	The Five V's	Duration	25 Minutes

Objectives
• Understand the characteristics that make up big data

Section No	Section	Notes	Timing
1	 Top Aim: Understand the characteristics that make up 'Bio Data'	Now we know what we mean by 'big data' we are going to look at five characteristics we need to be aware of. For data analysts it is important to understand the nature of a data set being considered for	

		inclusion as part of their analysis.	
2		The 5 v's are volume, velocity, variety, veracity and value. Don't worry about memorising these right now, we'll go through each one in detail. The 5 V's provide a useful toolbox of characteristics that can be deployed to assess each data set, determine any risks and create parameters for how to use the data. You should use the 5 V's as a mental checklist to employ whenever you are given a big dataset to work with.	
3	Volume  • Relates to how much data there is • Whether a dataset is considered 'big' or not is dependent on the volume • Data is considered Big when it is too voluminous for traditional data storage methods • Big Data platforms provide a way to store incredibly large amounts of data as well the ability to process it efficiently	The first characteristic to consider is volume, or how much data is in the dataset. Size of data plays a very crucial role in determining value you can get out of the data. Whether a particular data set can actually be considered as Big Data or not is dependent on the volume of the data. Typically petabytes (1000 terabytes) or exabytes (1000 petabytes) of data consisting of billions to trillions of records of millions of people- all from different sources (e.g. Web, social media, mobile data and so on). What do we think some issues that may arise due to the volume of our data?	

4	**Velocity**  • Refers to the speed of data generation • For example, how fast data flows in from sources such as application logs, networks and social media sites • Typically the flow will be massive and continuous • Increasingly companies are looking to stream analytics and data so it available at the right time to make appropriate business decisions • Big Data platforms offer solutions to dealing with fast flowing data in terms of storage and processing	Refers to the speed of generation of data Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks and social media sites, for example Typically for Big Data the flow of data is massive and continuous Increasingly companies are looking to stream analytics and data The data has to be available at the right time to make appropriate business decisions. The more agile a business the more the need for streaming. What other benefits or drawbacks can we think of to do with the velocity of data?	
5	**Variety**  • Refers to the heterogeneity of the data sources • In other words, how different the data sources are (unstructured vs structured etc) • With large flows of data from several of sources it is possible that there will be a variety of data types and structures • Big Data platforms can handle a variety of sources, structures and data types	Variety refers to data types. What are some of the different data types we saw back in module 1? Heterogeneity (how different it is) of data sources is important to consider with big data. For example, we need to consider whether it is structured or unstructured or what the file types are. Different types of data require different strategies which you will need to consider as part of your analysis.	

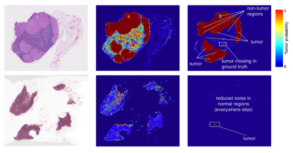
6	**Veracity**  • Refers to the quality of the data • Checking the veracity of data sources is always challenging- **never trust data as given** • Enquire about the data preparation- has it been normalised already? Has it been subject to statistical manipulation (i.e. outliers removed)? • Where possible you should always ask to look at raw data • Determining veracity can be challenging to an organisations policies as well as externally determining authorship/ownership	Veracity is the quality of being true or the habit of telling the truth Veracity of sources is always challenging Rule of thumb: **never trust data as given** Enquire and ask about the data preparation. Has it been normalised already? Been subject to statistical manipulation (i.e. outliers removed)? Where possible, you should always ask to look at the raw data Determining veracity can be challenging to an organisations (politics and policies) as well as externally (determining authorship/ownership) This is especially true of big data where data potentially could have drawn from a variety of disparate sources.	
7	**Value**  • Refers to the business value of the data • Having access to vast amounts of fast flowing data is useless unless it can leveraged into something of value • How does this data add value to your business? What is the ROI?	It's all well and good to have access to big data, but unless we can turn it into something valuable it is useless Does it provide added value to your business? Is the organisation working on big data achieving a high return on investment? If working with Big Data is not adding to a businesses profits, it is useless. Where do you think utilising Big Data will add value to your business?	

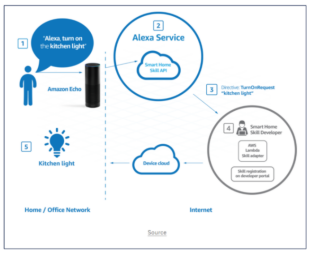
8	**Activity** • Think about some of the data you use regularly, how does it stack up against the 5 V's? • Share your thoughts with your group and be prepared to feedback to the rest of the class	In your groups look for some datasets you are familiar with (could be from work if closed cohort, or something they used for bootcamp, hackathon, presentation) and discuss the five v's. I.e. What is the value in analysis? How can we check the veracity of the data? WHat is the volume or velocity?	10 Minutes

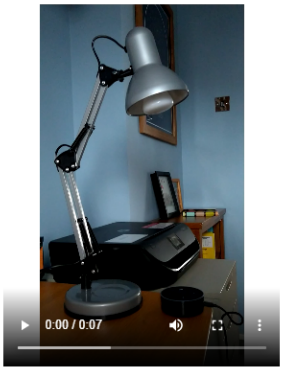
Topic	Developing the Future	Duration	20 Minutes

Objectives
• Understand the potential Big Data Analysis can have on our industries

Section No	Section	Notes	Timing
1	 Tips Aim: Understand the potential Big Data Analysis can have on our industries	Big Data as a subject area is still being invested in heavily and is at the forefront of some of the most recent developments in computing. Social media and genetic sequencing are among the fastest growing sources of big data and examples of non traditional sources of data being used for analyses. How do you think machine learning helps diagnoses in medicine?	

2	**Medicine**  Genetic sequencing and human genome mapping provide a detailed understanding of genetic makeup and lineage. The healthcare industry is looking toward these advances to help predict which illnesses a person is likely to get in their lifetime and take steps to avoid these maladies or reduce their impact through the use of personalised medicine and treatment. While data has grown, the cost to perform this work has fallen dramatically. The cost to sequence one human genome fell from $100m in 2001 to $10k in 2011. Today it costs around $1k. Companies are able to leverage the vast amount of new data available to build machine learning tools to predict results based on past experience. A 2017 report showed it was possible to train a model to detect tumour in breast cancer patients that either matched or exceeded the performance of a pathologist who had unlimited time to examine the slides.	Genetic sequencing and human genome mapping provide a detailed understanding of genetic makeup and lineage. The healthcare industry is looking toward these advances to help predict which illnesses a person is likely to get in their lifetime and take steps to avoid these maladies or reduce their impact through the use of personalised medicine and treatment. While data has grown, the cost to perform this work has fallen dramatically. The cost to sequence one human genome fell from $100m in 2001 to $10k in 2011. Today it costs around $1k. Companies are able to leverage the vast amount of new data available to build machine learning tools to predict results based on past experience. For example, a 2017 report showed it was possible to train a model to detect tumours in breast cancer patients that either matched or exceeded the performance of a pathologist who had unlimited time to examine the slides.	
3	**NLP**  Natural Language Processing is a powerful tool that allows teh analysis of text. Companies like Google and Amazon both make use of NLP and other technologies to give us a virtual assistant experience. **Activity:** In groups discuss and sketch how you think software such as Amazon Alexa produces personalised recommendations Read more about this story	Natural Language Processing is the ability of a computer to understand human language as spoken. The most obvious examples that people can relate to these days are Google Home and Amazon Alexa. Both use NLP and other technologies to give us a virtual assistant experience.	10 Minutes

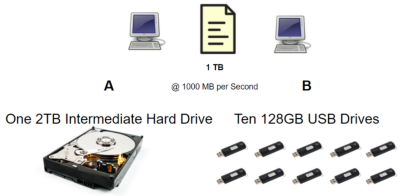
		Activity- In groups discuss and sketch how you think Alexa produces personalised recommendations for you (e.g. music, restaurants, movies, etc)	
4	Alexa  • Your words are recorded • The recording is sent to Amazon's server to be analysed (due to interpreting sounds taking up a l ot of computational power • Important words are identified and corresponding functions are carried out (e.g. 'basketball' would open a sports app) • The server then sends the information back to your device. If Alexa needs to say something it will follow a reverse process (identify the words it wants to give and convert them to the correct sounds)	Alexa is built on natural language processing, a process of converting speech into words, sounds and ideas Alexa will record your words. Interpreting sounds takes up a lot of computational power however, so the recording is sent to Amazon's servers to be analysed more efficiently (use of big data) There, the recording will be broken down into individual sounds. It then consults a database containing various word's pronunciations to find which words most closely correspond to the combinations of individual sounds. It then identifies important words to make sense of the tasks and carry out corresponding functions. For example, if Alexa 'hears' basketball, it would open a sports app Amazon's servers then send the information back to your device. Alexa may speak at this point. If Alexa needs to say something it will go through the same process, but in reverse (i.e. identify the words it wants to give, convert them into audio files and send them back).	

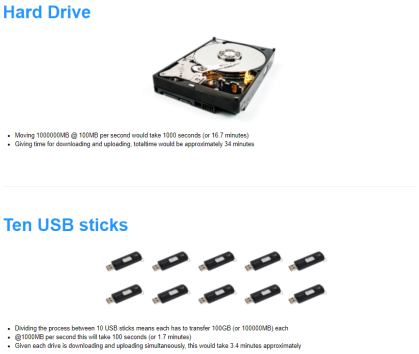
5		For the fun of it, I have included a video of Alexa in action… What other uses of big data are you aware of?	

Topic	How Does it Work?	Duration	10 Minutes

Objectives
• Understand the principles behind how Big Data technologies work

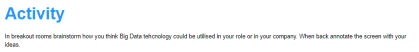
Section No	Section	Notes	Timing
1		Let's now consider how Big Data technologies work. Take for example a SQL query or a python script which you can run on your local computer. How efficiently do you think it will process data which is in terabytes? This is where Big Data comes in. Instead of downloading all the data to your machine, you instead take your code to the data.	

2	**Divide and Conquer**  Most Big Data softwares use an approach called 'Divide and Conquer' where the data will be split into smaller 'chunks' and processed simultaneously across different nodes and the results combined. This process happens over several servers so if one fails the analysis can continue and the remaining can pick up the slack.	Big Data uses an approach called 'divide and conquer' where your data is split into smaller chunks and processed simultaneously, after which the results are combined. To do this efficiently it will place these chunks over several servers. This has a couple of benefits- by spreading out the processing across multiple servers you are requiring each one individually to do less processing which overall is faster. Also, by spreading the chunks about it means if one server fails for any reason the analysis can continue uninterrupted on the others. Once complete each server has completed their part and the results combined, the final output is then sent back to you. Let's do an example to show why it is quicker.	
3	**Which is Faster?**  	Let's do some maths! [This can be done in breakout rooms] Imagine you want to transfer 1TB (1000GB) of data between two computers. Assuming the transfer speed is 1000 MB per second, how long will it take to transfer the data using: a) One 2TB intermediate hard-drive or b) Ten 128GB USB Flash Drives? Assume both computers have 10 USB ports and transferring into multiple USB drives	5 minutes

		simultaneously does not affect speed.	
4		You can see then that the divide and conquer approach is much faster than doing it all in one go. This is the basic principle behind Big Data. Divide your analyses into different chunks and process each simultaneously. It means processing terabytes of data can happen in minutes instead of hours.	

Topic	Using Big Data in Your Role	Duration	10 Minutes

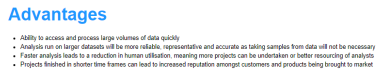
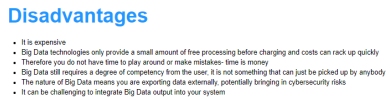
Objectives
• Consider how Big Data technologies can be utilised in apprentices own roles

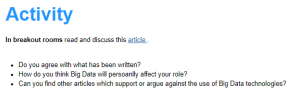
Section No	Section	Notes	Timing
1		Over to you, let's now consider how Big Data could be used to enhance your role.	
2		We are going to go into breakout rooms where I would like you to discuss how Big Data could be used to enhance your roles Once back, ask apprentices to annotate the slide to show their	7 Minutes (5 Minutes in room and 2 to annotate)

		ideas. When writing up a project for your portfolio, a good thing to include is your justification (or lack of) big data technologies in your analysis. We will look into this in the following section.	

Topic	Advantages (and Disadvantages)	Duration	20 Minutes

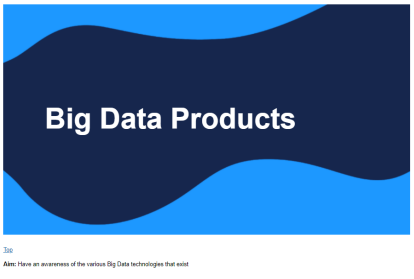
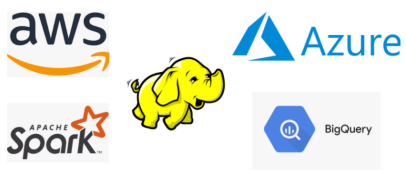
Objectives
• Justify the use of Big Data in analysis by considering the benefits and drawbacks

Section No	Section	Notes	Timing
1		So what are the benefits of utilising big data in your analysis? When justifying your use or non use you will want to refer to some of the information provided in this section.	
2		Run through the advantages, can the apprentices think of any others?	
3		Run through the disadvantages, can the apprentices think of any others? In most cases the advantages of using Big Data will outweigh the disadvantages, but it is important to be aware of the downsides when writing up	

		your portfolio.	
4		In breakout rooms read and discuss the following article- https://www.datamation.com/big-data/big-data-pros-and-cons.html Do you agree with what the writer has put? Can you find other articles which support or argue against the use of big data? What do you think and how can it personally affect your role?	10 minutes

Topic	Big Data Products	Duration	5 Minutes

Objectives
• Be aware of the various Big Data technologies that exist

Section No	Section	Notes	Timing
1		Finally, let's look at what Big Data products are available. We will not look into any of them individually, but it is important you know what exists if you want to try them out.	
2		There are many platforms or services that allow for the efficient processing of Big Data. Hadoop was one of the first built and many companies still use it today to process their data. Apache Spark provides a similar service and both can be easily integrated with python.	

		BigQuery is run by Google while AWS is run by Amazon, both have a wide suite of functions allowing you to run queries and scripts efficiently. AWS even has functionality to bring python scripts live. Azure is owned by Microsoft and offers a cloud computing platform with a wide degree of functionality Many more exist, what does your company use?	

Topic	Data Platforms	Duration	20 Minutes

Objectives
• Consider the various data platforms that exist • Reflect on what products you use and justify their use

Section No	Section	Notes	Timing
1	 Top Aim: Reflect on the products you use in your role and justify their use	An aspect of data analytics you need to consider and justify is the use of data platforms. Companies invest in platforms to make analysis faster and easier to perform. For example, instead of building a dashboard in python you could use Tableau or PowerBI. In this section we will consider what platforms there are and discuss their potential uses and how you can apply them in your role.	

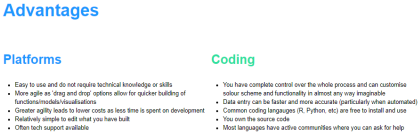
2	Visualisation Platforms 	Already in this course we have come across Tableau and PowerBI. These platforms allow us to build visualisations in a simple way and have a varied suite of options as well as other analysis functions. Ask the apprentices to give any examples of how they have used these softwares in their role.	
	Data Management 	Many of you are familiar with software like Microsoft Excel for spreadsheet and database management. It is easy to perform EDA, calculating aggregates and making visualisations as well as producing tables for showcasing your results. The GUI is fairly intuitive and macros allow for a great range of functionality including adding regression and other ML techniques)	
3	Statistics Platforms 	Has anyone used SPSS before? SPSS stands for statistical package for social sciences and allows you to perform a wide variety of statistical operations on a dataset, including linear and logistic regression, ANOVAs and much more. While it is expensive to license and has a steep learning curve, you can perform statistical analysis incredibly quickly through a few clicks of a button once you know what to do.	

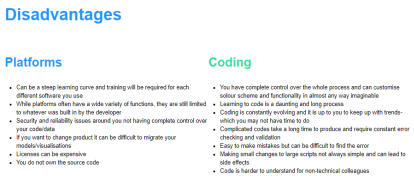
4	Cloud Based Platforms 	Then you have software like Microsoft Azure. Azure is a cloud based computing platform which allows you to perform data analytics and is used to supplement companies servers (or even replace them). Does anyone have any experience of using Azure?	
5	Common Platforms 	There are a wide range of data platforms on the market which allow you to do anything from data mining and data retrieval to exploratory data analysis and predictive analytics. Ask the apprentices what softwares they use and give examples of how they have used them.	
6	Activity In Breakout Rooms: • Discuss what software platforms you use in your role and how you use them • What are the advantages and disadvantages of using them?	In breakout rooms, discuss what softwares you use in your role and what advantages (and disadvantages are of using them). Ask for examples once everyone has returned.	8 Minutes

Topic	Platforms vs Coding it Yourself	Duration	20 Minutes

Objectives
• Justify the use of platforms by comparing them to coding it yourself

Section No	Section	Notes	Timing

1		We have seen various platforms and softwares you can use in your role if your company has licensed them, but are they the optimal way for your analytics? All platforms depend on pre-written code, but could it be better to just code it yourself? When writing your portfolio you need to be able to justify why you used (or didn't use) certain programmes and part of this is considering the advantages and disadvantages of coding it yourself. The following slides will list some of the advantages and disadvantages of each you can use in your write ups.	
2		First the advantages. When using software you are using external code with a graphical user interface meaning you can build what you want without having to type all the code yourself. Projects can be completed quicker, reducing costs and freeing up your time. Coding it yourself gives you complete control over the whole process and costs nothing in terms of licensing. You also own the code, meaning it is more secure and giving you intellectual property rights.	

3		While platforms can make building models quick and easy, you still need to learn how to use the programme, and if your company licenses a new software then you will need to receive more training, as most programmes have unique methods of useage. On top of this, while platforms often contain a wide suite of processes they are still limited compared to what you can do yourself. A bigger issue also relates to data security- you are not in complete control of your code and data, if the platform provider is hacked your data potentially could be compromised. Also consider what happens if you want to switch providers or they go out of business- migrating your properties across platforms can be difficult. Coding yourself has its challenges too. Not everyone can code and with it being an ever changing process can be a daunting process to begin learning. Even seasoned coders can be unfamiliar with new concepts. Complicated models can take a long time to build and can also be frustrating- it can be difficult to identify where errors will occur as often they won't appear until you're ready to run the entire script. If it is a large script finding errors can be challenging, and minor changes could lead to serious unintended consequences later down the line. Using a platform or coding it	

		yourself is a decision you must make when completing a project and a decision you must justify in your portfolio. When writing up the task/action section write down what method you used to achieve your goal and give reasons for why you did and not any other method. This could be to prioritise stakeholder needs as they require output in a particular way.	
4	**Activity** Mr Jones works in a data analytics department and has been given a project to complete. He must design a dashboard that displays daily KPIs for his stakeholders. In teams you will be each be assigned a different product (Tableau, Python, etc) and will come up with arguments to convince Mr Jones to use your product. Each team will then be given 1 minute to pitch their product.	7 minutes in breakout rooms	10 Minutes

Topic	Setting Up R	Duration	5 Minutes

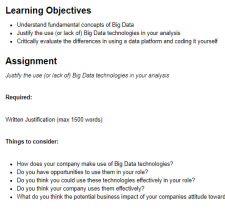
Objectives
• Ensure apprentices have correct software for the following R section

Section No	Section	Notes	Timing
1	 Top **Aim:** Have the correct software installed to use R	Before we finish this session let's take some time to set up R which we will be looking at tomorrow.	
2	1. Open Anaconda 2. Click Environments 3. Click Create 4. Name it R and check the R box 5. Wait a few minutes for the environments to be created 6. Once ready, click on the triangle next to your new environment and select 'Open with Jupyter Notebook' 7. Whenever you want to use R make sure you use this new environment. If you want to use Python make sure you use the environment called base(root) 8. To create a R notebook, select new in the Jupyter navigator and select R	1. Open Anaconda 2. Click Environments 3. Click Create	

		4. Name it R and check the R box 5. Wait a few minutes for the environment to be created 6. Once ready, click on the triangle next to R and select ‘Open with Jupyter Notebook’	

Topic	Session 1 Recap	Duration	5 Minutes

Objectives
• Review the days learning and introduce the assignment

Section No	Section	Notes	Timing
1		Review the day	
2			

Topic	Session 2 Intro	Duration	10 Minutes

Objectives
• To provide an overview of the class agenda and the expected learning objectives

Section No	Section	Notes	Timing
55	**Session 2: R** In this session we will be exploring R, a computing language which has been optimised for statistical analysis and machine learning. While it does not have the range of functionality Python has it more than makes up for it with the suite of packages it has for data manipulation, analysis and visualisation This notebook will cover the basics of R, exploratory data analysis and visualisations. **Session Outline** • Features of R • Control Flow in R • Exploratory Data Analysis in R • Visualisation in R • Linear Modelling in R • RStudio • Recap **Learning Objectives** • Understand the Basics of R • Compare and Contrast the functions of R with Python • Utilise R to perform Data Analytics	Coach welcome to session, run an ice breaker from here. Run through session plan and learning objectives	

Topic	Features of R	Duration	30 Minutes

Objectives
• Understand the Basic Features of R

Section No	Section	Notes	Timing
1	**Features of R** Top **Aim:** Understand the Basic Features of R R is a statistical language realeased in 1995 by Ross Ihaka and Robert Gentleman at the University of Auckland. It was named after the first letter of both author's names, and as a play on S, the computing language R was built upon. R is mostly used for statistical computation and analysis but has also gained attention for developing machine learning products. Like Python, R's statistical computing capabilities can be extended with packages (called packages). R is also an object orientated language (OOP) but is not as intuitive as python. However, the wide range of statistical tests and the ease of building visualisations makes R a favourite of statisticians across the globe.	Explain that R is a language built for statistical analysis built in 1995. It carries much of the same functionality as python but has different syntax. In this session you will be showing the R equivalents to different concepts learned previously. There will be chances to practice as well.	

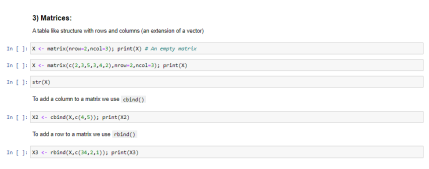
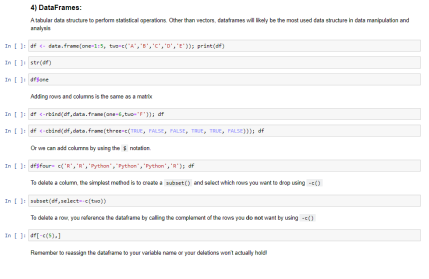
2	**Packages** R packages are collections of functions that allow for reusability, modularity and building cleaner code through extending R capabilities. Packages are stored in repositories called Comprehensive R Archive Networks or CRANs. As with Python, R is open source meaning anyone can write packages and you can install any package available in a CRAN. A script in R is a text file with extension `.R` where you can write your code and save anywhere.	In R repositories are called CRAN.	
3	**Data Types in R** The table below shows the different data types in R and what their python equivalents are. Data Type (R) | Python Equivalent | Example Character | String | 'a', 'MultiVerse', 'a123abc' Integer | Integer | 3, 7, 902 Numeric | Float | 10.5, 3.14, 2.3 Logical | Boolean | TRUE, FALSE NA | NaN Note that for logical data types they are all capitalised as opposed to python where only the first letter is uppercase. If you are ever stuck in R, remebar there is the `help()` function to bring up the documentation. Otherwise, stackoverflow has plenty of solutions to common problems.	Introduce the different data types and how they relate to those from python.	

4	**Declaring Variables in R**
	For a single variable
	`x <- 'my variable'`
	Traditionally we use `<-` to assign a variable, but `=` works just as well.
	`x = 'my variable'`
	Like python, R has a print function to see what is stored in a variable.
	`print(x)`
	Or we could just call the variable name itself to see what is stored.
	`x`
	We can also calla collection of integers
	`y <- 1:5`
	`y`
	You can check the datatype of a variable by using the command `typeof()`
	`typeof(x)`
	`# We can add comments in R the same way we do in Python`
	`"But there are other methods of doing it too"`
		Demonstrate how to use basic R functions like calling a variable.
		
		Note that you can use = or <- when assigning a variable.	

5	**Mathematical Operators**
	Like python, R has a range of in built operators
	`1 + 1`
	`10-15`
	`5 * 6`
	`2 ** 5`
	`10 %% 3 #Modulo`
	`10/5`
	`sqrt(25)`
	`mean(3,5,6)`
	Data Structures in R
		Go through the different mathematical operators. Note how square root and mean are already built in, whereas in python you had to import a library.	

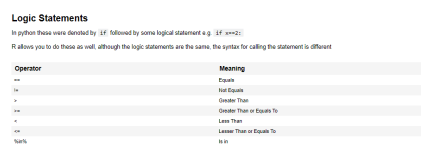
6	**Data Structures in R**
	Like in python there exist data structures we use to store our data for analyis.
	1) List:
	A collection of elements that can hold a combination of data types. Note, a list is also similar to a dictionary in python
	`l <- list('a'=1, 'b'=2,'c'=3)`
	`l`
	If I want to call an element from a list I need to use the `$` notation and reference the tag.
	`l$a`
	Lists do not require tags to be called however.
	`l2 <- list(1,'some text',3, TRUE)`
	`l2`
	We can reference elements from a list by using their key. Lists in R are more similar to dictionaries in python.
	`l2[4]`
	`l2['a']`
	If I want to retrieve this value for further use I need to use [[]]
	`l2[[4]]`
	`l[['a']]`
	To add items to a list you need to call the tag of where you want it to be placed and assign it the value you want.
	`l[['d']]='Multiverse'`
	`l`
	Editing a value in a list follows the same format.
		In the following section you will be explaining much of the different data structures in R.
		
		Start with lists, note how these are more similar to dictionaries in python. They can hold a combination of data types and have tags.
		
		Demonstrate how to reference lists and manipulate them.	

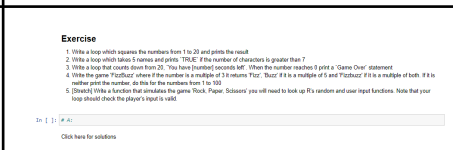
7	**2) Vectors:**
	A collection of elements of the same data type, similar to numpy arrays
	We use `c()` to define a vector
	`numbers <- c(1,2,3); numbers`
	`str(numbers)`
	To edit, append or view values from a vector you just use the same notation as a list. Note that vectors are not zero indexed.
	`numbers[1]`
	`numbers[c(1,3)] #access several elements`
	`numbers[3]= 9; numbers`
	`numbers[4]=10; numbers`
	We can make it more like a dictionary and apply tags to the values.
	`more_numbers <- c('first'=1, 'second'=2, 'third'=3); more_numbers`
	`more_numbers['first']`
	`more_numbers[['first']]`
	What is the difference between a list and a vector?
	`# A:`
	Click here for solution
		Show vectors and highlight they are different to lists as they must contain the same data type.
		
		Demo basic functions.	

8		Introduce matrices as an extension of a vector.	
9		Introduce dataframes and link them to pandas. State how you will be using them a lot with data manipulation. Demonstrate how you can columns or rows and how to reference a column	
10		Give them 5 minutes to complete this exercise	5 Minutes

Topic	Control Flow in R	Duration	30 Minutes

Objectives
• Understand how to set up loops and functions in R

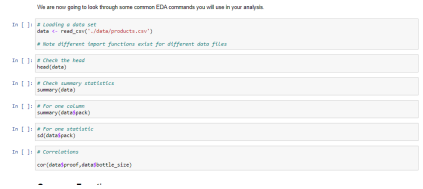
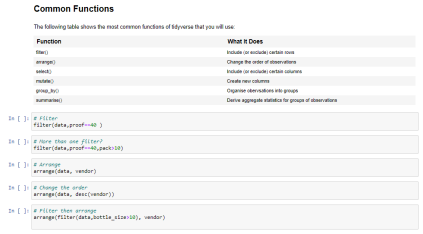
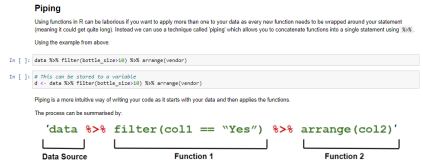
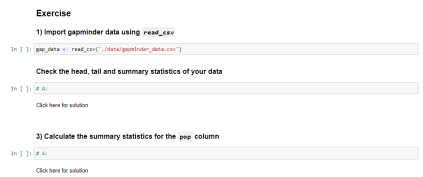
Section No	Section	Notes	Timing
1		Depending on the skill level of the group, this can be skipped. In python they learned how to write loops to parse data and automate processes. We will now show them how to do this in R	
2		Show the different logic statements, note how similar they are to python. Demonstrate if statements and build in else then else if. Note	

		the difference from python (not elif) and the way the statement is called ({} instead of :)	
3		Show how to build a for loop and how to add in a logic statement	
4		Do the same with while loops.	
5		Show how to build a function, note the use of the return statement.	
6		Solutions embedded to notebook	10 Minutes

Topic	Exploratory Data Analysis in R	Duration	35 Minutes

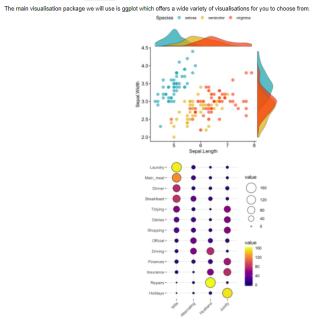
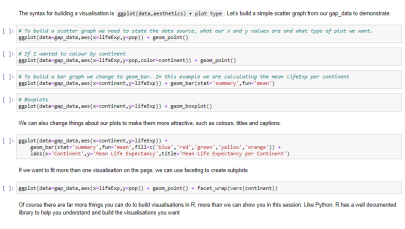
Objectives
• Consider how to perform EDA in R

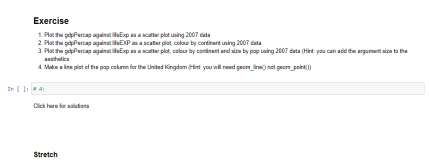
Section No	Section	Notes	Timing
1		Link back to previous lessons on EDA (python, SQL, tableau/PowerBI, Excel) and reiterate its importance. R has many functions readily available to help with this which are easy to use.	

2		Packages are the same as libraries in python (many packages in R can be used in python as well). Show how to install (only need to do once) and then how to call it up.	
3		Show how to read in data and how to perform the basic checks we saw in python.	
4		Introduce the common functions and state that you will be demonstrating them shortly. Note the odd syntax of trying to apply more than one function.	
5		As wrapping functions can be confusing, piping is a better way of adding in several layers to the same command. Demonstrate how much simpler piping is compared to wrapping.	
6		Show how to slice a dataframe, select and rename columns.	
7		Demonstrate how to add new columns or amend an existing one, how to group by and aggregate data frames.	
8		Solutions embedded	10 Minutes

Topic	Visualisation in R	Duration	35 Minutes

Objectives
• Build Visualisations in R

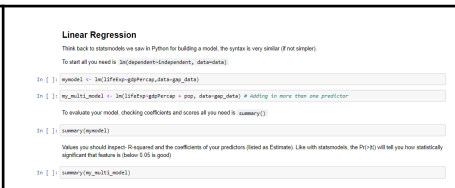
Section No	Section	Notes	Timing
1		Visualisation is an important part of EDA and presenting your findings. It is good for your portfolio to have a variety of methods, so consider building some in R.	
2		The library we will be using is ggplot2 (comes with tidyverse). It is simple to use and has a wide array of functionality. While we will be keeping it simple, there are many ways you can build outstanding visualisations. Note: plotly dash can also be run in R	
3		Show how to build a scatter plot, bar chart and box plot. Show how to add colour and other features to the plots. Show facet_wrap as as a way of putting multiple plots on the page. Note the structure of geom_graph and state that this is the general structure of	

		ggplots.	
4		Solutions embedded, encourage those who are more confident to attempt the stretch questions.	15 Minutes

Topic	Linear Modelling in R	Duration	15 Minutes

Objectives
• Learn about different modelling techniques in R

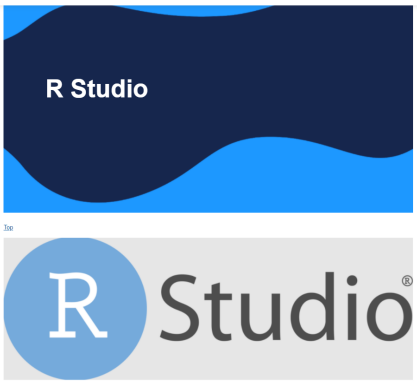
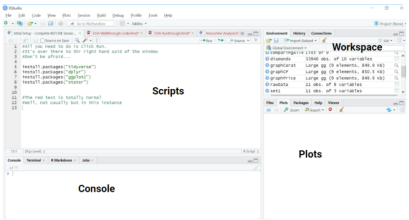
Section No	Section	Notes	Timing
1	 Top Aim: Learn about different modelling techniques in R The final part of our whistle-stop tour of R is to see how easy it is do build a linear model. R has a wide range of models you can use as well as allowing you to perform statistical tests such as ANOVA, t-tests and generalised linear modelling. The examples below show how to start building these models/tests but if you want to learn more will need to do your own research.	This section can be skipped if short on time or the group not confident enough. Amongst R's many applications is statistical analysis and machine learning. In fact, R is possibly better suited for it due to the wide range of built in statistical tests and models already built in. In python we saw different types of hypothesis tests, R has more and the syntax is easier! It can also easily do ANOVA, chi-square test, f-test and more. We will look at a few now.	

2		Demonstrate how to build a linear regression. Show how to call a model and note the simplicity. Train test split is possible, but we won't be showing it today. Walk through the summary and draw particular attention to p-values and r squared.	
3		Do the same for logistic regression. Walk through the data engineering steps (not seen yet) and then how to build and interpret the model. Note the similarity to the python statsmodels package. Also note the different name from python.	
4		Finally show a t-test. Stress the importance that normally you would do checks such as looking at distributions, summary statistics, etc but for today we are only going to look at how to run and interpret one.	

Topic	RStudio	Duration	5 Minutes

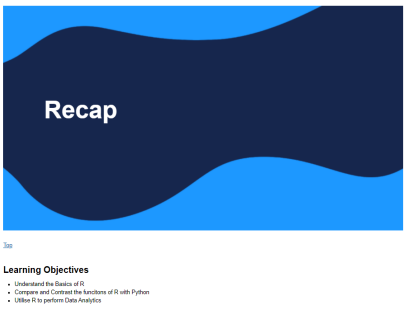
Objectives
• Be aware that RStudio exists

Section No	Section	Notes	Timing

1		We used Jupyter because apprentices are already familiar with it In most industry cases they would use RStudio.	
2		If you have time, download and bring up RStudio to show how it works and although the GUI is different, the syntax is the same.	

Topic	Recap	Duration	5 Minutes

Objectives
• Recap the day

Section No	Section	Notes	Timing
1		Recap the day, remind apprentices about OTJ, assignments and SALs. Assignment is in other notebook titled ‘assignment.’ Solutions are in repo.	