ROBIN WHITING

DATA ESSENTIAL SKILLS

CONTENTS

INTRODUCTION

I am an aspiring Data Analyst with over 14 years' experience working in the field of analytical chemistry in the pharmaceutical and environmental sectors. I also have experience working in other areas, including business administration and marketing. I am adept in the use of Microsoft Office, with a particular interest in using Microsoft Excel for a wide range of applications, including data handling and analysis.

Robin Whiting BSc (Hons)

Data Analyst I Web Creator I Trainer

https://robin-whiting.co.uk/

In each of my previous roles, I have always been on the lookout for ways to improve the accuracy and efficiency of processes that handle data in my own and other areas of the business, suggesting and implementing more streamlined ways of working. I have always found data analysis the most interesting aspect of my work and aim to transfer existing skills, whilst developing new skills in this field, including Python and SQL; the Cambridge Spark Data Essential Skills Bootcamp was the perfect fit for myself to create a platform on which to further my career development.

This portfolio evidences some of the main areas of my learning during the data bootcamp.

THE VALUE OF DATA

In this section, we take a brief look at some of what I have learnt, regarding the ways the role of data can impact and bring value to organisations in an ever-more digitised economy and society.

WHY IS DATA IMPORTANT TO ORGANISATIONS?

Data is important to organisations because it helps them make informed decisions, improve their performance, and achieve their goals. Data can provide insights into customer needs, market trends, operational efficiency, and competitive advantage. Data can also help organisations identify and solve problems, reduce costs, and increase revenue. Data enhances decision-making capabilities and is the foundation of any successful organisation in the modern world and

HOW DOES DATA ADD VALUE TO ORGANISATIONS?

Data is a vital asset for any organisation that wants to thrive in the modern world. Data can help organisations improve their decision-making, optimise their processes (and minimising risk), enhance their customer experience, gain greater insights into target markets, and create new products or services. Data can also help organisations reduce costs, increase efficiency, and gain a competitive edge in the market, creating targeted strategies and marketing campaigns, as well as using it to identify new product and service opportunities. As just one example, the use of a Customer Relationship Management (CRM) system can help with providing a more personalised service to customers by sharing knowledge across different departments and harmonising the customer experience between customer-facing staff (e.g. sales team and customer support). Data can add value to organisations in many ways, but only if it is collected, stored, analysed, and used effectively and ethically.

WHY HAS THE VALUE OF DATA INCREASED?

The value of data has increased because it has become increasingly essential for many aspects of modern life, such as business, education, health, and entertainment. Technology which was once unheard of is now a part of everyday life for the majority of the population. Data can provide insights, solutions, and opportunities that were not possible before. Data can also help improve efficiency, productivity, and innovation across different departments and functions within an organization, and in various sectors and industries; it has become an increasingly valuable asset that can create competitive advantages and social benefits using modern tools including smartphone technology, artificial intelligence and the Internet of Things (IoT).

SOURCING AND COMBINING DATA

In this exercise, I had to select a question that required myself to combine at least 2 different data sets in Excel. The question I set myself and how I went about solving this follows:

Question:

Is there a higher level of 'happiness' in countries that have democratically elected governments?

Data Tools Used:

- Microsoft Excel, Including:
  - basic functions (e.g. COUNTIF, VLOOKUP)
  - charts
- Seeking Data (Internet)

SOURCING AND COMBINING DATA - EVIDENCE

In order to collect relevant and up-to-date information on world 'happiness' and measures of democracy for each country, I obtained data from the following trustworthy sources:

1. Democracy Index values from WorldPopulationReview.com, an independent for-profit organization committed to delivering up-to-date global population data and demographics (includes data sources and methodology on each webpage),

2. World Happiness Report 2023, a publication of the Sustainable Development Solutions Network, powered by the Gallup World Poll data.

3. Full alphabetical list of countries as exactly recognised for use in Excel Map Charts from www.map-in-excel.com.

The datasets from sources 1. And 2. were combined into an Excel spreadsheet, with both being downloaded first as Excel files. The data was from the very latest reports and *not* subject to change (i.e. not 'live' data, which requires direct linking from a webpage).

SOURCING AND COMBINING DATA - EVIDENCE

	A	B	C	D	E	F	G	H
1	Country	Category	Total Score	Electoral Process / Pluralism	Functioning of Government	Political Participation	Political Culture	Civil Liberties
2	Norway	Full Democracy	9.81	10	9.64	10	10	9.41
3	Iceland	Full Democracy	9.37	10	8.57	8	10	9.41
4	Sweden	Full Democracy	9.26	9	9.29	8	10	9.12
5	New Zealand	Full Democracy	9.25	10	8.93	8	8	9.71
6	Canada	Full Democracy	9.24	9	8.93	8	9	9.41
7	Finland	Full Democracy	9.2	10	8.93	8	8	9.41
8	Denmark	Full Democracy	9.15	10	8.93	8	9	9.12
9	Ireland	Full Democracy	9.05	10	7.86	8	9	9.71
10	Australia	Full Democracy	8.96	10	8.57	7	8	9.71
11	Netherlands	Full Democracy	8.96	9	9.29	8	8	8.82
12	Taiwan Province of Chin	Full Democracy	8.94	10	9.64	7	8	9.71
13	Switzerland	Full Democracy	8.83	9	8.57	7	9	8.82
14	Luxembourg	Full Democracy	8.68	10	8.57	6	8	9.41

Dataset 1: Democracy Index for each Country.

	A	B	C	D	E	F	G	H
1	Country name	Ladder score	Standard error of ladder score	upperwhisker	lowerwhisker	Logged GDP per capita	Social support	Healthy life expectancy
2	Finland	7.804	0.036	7.875	7.733	10.792	0.969	71.150
3	Denmark	7.586	0.041	7.667	7.506	10.962	0.954	71.250
4	Iceland	7.530	0.049	7.625	7.434	10.896	0.983	72.050
5	Israel	7.473	0.032	7.535	7.411	10.639	0.943	72.697
6	Netherlands	7.403	0.029	7.460	7.346	10.942	0.930	71.550
7	Sweden	7.395	0.037	7.468	7.322	10.883	0.939	72.150
8	Norway	7.315	0.044	7.402	7.229	11.088	0.943	71.500
9	Switzerland	7.240	0.043	7.324	7.156	11.164	0.920	72.900
10	Luxembourg	7.228	0.069	7.363	7.093	11.660	0.879	71.675
11	New Zealand	7.123	0.038	7.198	7.048	10.662	0.952	70.350
12	Austria	7.097	0.040	7.176	7.018	10.899	0.888	71.150
13	Australia	7.095	0.044	7.180	7.009	10.821	0.934	71.050
14	Canada	6.961	0.042	7.042	6.879	10.773	0.929	71.400
15	Ireland	6.911	0.044	6.996	6.825	11.527	0.905	71.300
16	United States	6.894	0.047	6.986	6.802	11.048	0.919	65.850
17	Germany	6.892	0.049	6.989	6.795	10.879	0.896	71.300

Dataset 2: World Happiness Report Data.

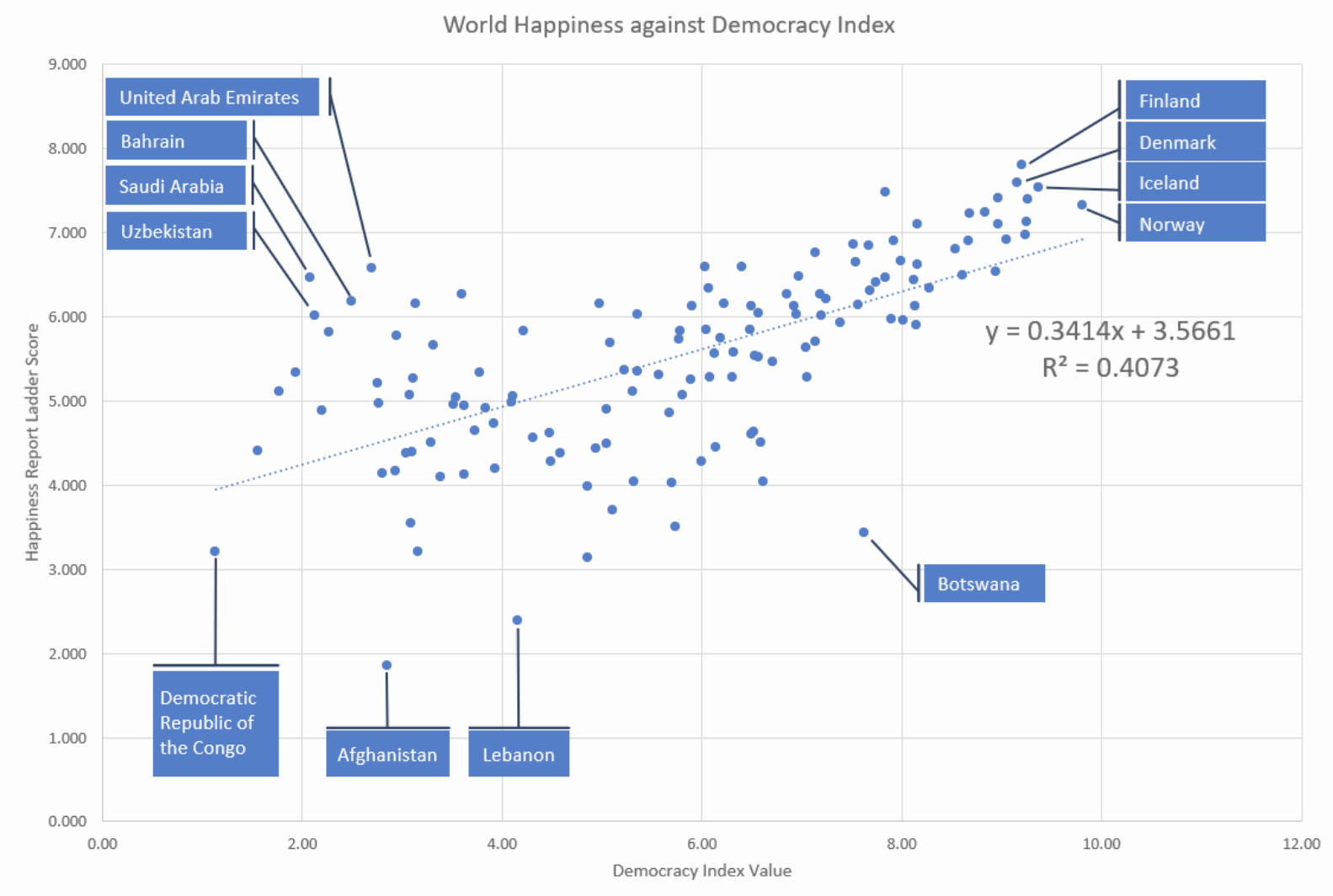
SOURCING AND COMBINING DATA - EVIDENCE

In order to compile a comprehensive list of country names (some of which may not have been listed in 1. and 2. due to availability of data), data source 3. was used. 3. was used to check for consistent naming of countries (e.g. UK, United Kingdom, United Kingdom of Great Britain and Northern Ireland), by using the COUNTIF function to check if an exact name was present in the list and if it needed adjusting. Once the names of countries for each data source had been harmonised, a list of countries that appear in both datasets 1. And 2. was compiled (using Excel formulae), and data for democratic index and world happiness recorded next to each country (using VLOOKUP function and source data files); the combined report is shown on the next page.

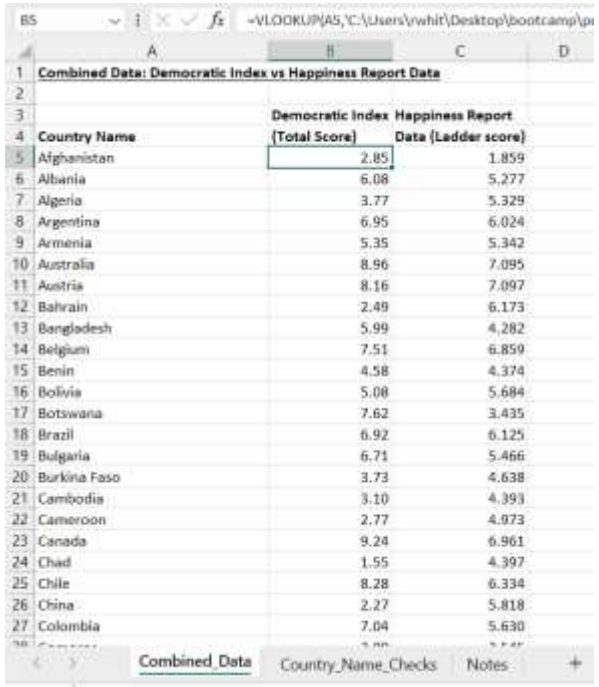
	A	B	C	D	E
1	ISO3 country code table for use in Excel E-Maps				
2					
3	ISO3	Country name			
4	ABW	Aruba			
5	AFG	Afghanistan			
6	AGO	Angola			
7	AIA	Anguilla			
8	ALA	Åland Islands			
9	ALB	Albania			
10	AND	Andorra			
11	ARE	United Arab Emirates			
12	ARG	Argentina			
13	ARM	Armenia			
14	ASM	American Samoa			
15	ATA	Antarctica			
16	ATF	French Southern Territories			

Dataset 3: Full list of countries with exact names as used by Excel (these were needed later in the "Storytelling" section of this portfolio)

SOURCING AND COMBINING DATA - EVIDENCE



Final Combined Data Report



	A	B	C	D
1	Combined Data: Democratic Index vs Happiness Report Data			
2				
3		Democratic Index	Happiness Report	
4	Country Name	(Total Score)	Data (Ladder score)	
5	Afghanistan	2.85	1.859	
6	Albania	6.08	5.277	
7	Algeria	3.77	5.329	
8	Argentina	6.95	6.024	
9	Armenia	5.35	5.342	
10	Australia	8.96	7.095	
11	Austria	8.16	7.097	
12	Bahrain	2.49	6.173	
13	Bangladesh	5.99	4.282	
14	Belgium	7.51	6.859	
15	Benin	4.58	4.374	
16	Bolivia	5.08	5.684	
17	Botswana	7.62	3.435	
18	Brazil	6.92	6.125	
19	Bulgaria	6.71	5.466	
20	Burkina Faso	3.73	4.638	
21	Cambodia	3.10	4.393	
22	Cameroon	2.77	4.973	
23	Canada	9.24	6.961	
24	Chad	1.55	4.397	
25	Chile	8.28	6.334	
26	China	2.27	5.818	
27	Colombia	7.04	5.630	

Combined_Data | Country_Name_Checks | Notes

SOURCING AND COMBINING DATA - EVIDENCE

With reference to the data shown on the previous slide, there does appear to be a greater amount of happiness in more democratic countries, but the correlation between the two is only 41%, meaning 59% of the variability in the outcome data cannot be explained by the model. Some points to note from the data are as follows:

- In terms of most democratic and happiest countries, at the top (unsurprisingly) there is a cluster of Nordic countries.
- Of the less democratic countries, there appears a wider variation of happiness.
    - Of these, the less happy countries face challenges, including Democratic Republic of Congo (poverty & armed conflict), Afghanistan (including women's and girls' rights and economic and humanitarian crises), and Lebanon (including mismanagement, violence, economic crisis and lack of electricity supply.
    - Of these, the atypically happier countries includes a cluster of wealthy oil-rich states on the Arabian Peninsula. A surprising exception to this is Uzbekistan (factors of its relatively high happiness may well be attributable to a number of factors including having a more open and progressive government, open borders and equal rights for women [Reference 4]).
- One other outlier of note is Botswana which has a good score for democracy, but a low score for happiness. This can be attributed to ‘a lack of social support, curtailed healthy life expectancy and, above all, matters of income inequality and of generosity.’ [Reference 5]

SOURCING AND COMBINING DATA - EVIDENCE

A full discussion of all of the countries here is beyond the scope of this brief study and further work could also include looking at the Democracy Index whilst breaking down data for the individual factors used in the calculation of World Happiness Data (e.g. social support, healthy life expectancy, freedom to make life choices, generosity).

Sources of Data and Citations:

1. https://worldpopulationreview.com/country-rankings/democracy-countries, with the following citation: The Democracy Index, an annual report compiled by the Economist Intelligence Unit. The index measures the state of democracy in 167 of the world's countries by tracking 60 indicators in five different categories: electoral process and pluralism, functioning of government, political participation, political culture, and civil liberties.
2. https://worldhappiness.report/ed/2023/#appendices-and-data, with the following citation: Helliwell, J. F., Layard, R., Sachs, J. D., Aknin, L. B., De Neve, J.-E., & Wang, S. (Eds.). (2023). World Happiness Report 2023 (11th ed.). Sustainable Development Solutions Network.
3. Country Codes - List of ISO3 country codes| Excel E-Maps (map-in-excel.com)
4. *Background: Women and Uzbek nationhood* (no date) *Uzbekistan: Sacrificing Women To Save The Family? - Background: Women and Uzbek Nationhood*. Available at: https://www.hrw.org/reports/2001/uzbekistan/
5. *Miraculous unhappiness; does Botswana need a Ministry of Laughter?* (no date) *Miraculous Unhappiness; does Botswana need a Ministry of Laughter? | African Studies Centre Leiden*. Available at: https://www.ascleiden.nl/content/ascl-blogs/rijk-van-dijk/miraculous-unhappiness-does-botswana-need-ministry-laughter (Accessed: 17 November 2023).

NUMBERS FOR DATA ANALYSIS

For this section of the bootcamp, I was required to analyse simple and complex structured and unstructured data using basic statistical methods and algorithms and identify any trends and patterns in the data.
The learning also gave me an opportunity to brush-up on some basic statistics (e.g. quartiles and percentiles, standard deviation).
The dataset used for the following exercise included information on staff names, department, actual sales, and target sales. From analysing this information, I was able to answer questions relating to performance based on targets. Screenshots from this work are shown on the following pages:

Data Tools Used:

- Microsoft Excel, Including:
  - basic maths operations
  - working with tables
  - formulae (e.g. SUM, SUMIFS, AVERAGE, AVERAGEIFS, MAX, MIN)
  - charts
  - basic statistics

NUMBERS FOR DATA ANALYSIS - EVIDENCE

	A	B	C	D
1	NAME	SALES TARGE	ACTUAL	DEPARTMENT
2	Caleb Woodard	$30,000.00	$32,700.00	A
3	Elliot Vaughn	$30,000.00	$33,521.00	A
4	Abbie Forbes	$50,000.00	$45,535.00	A
5	Alisha Branch	$100,000.00	$96,003.00	A
6	Malcolm Marquez	$55,000.00	$58,197.00	A
7	Noel Roach	$100,000.00	$95,640.00	A
8	Jamiya Carney	$65,000.00	$65,349.00	A
9	Darren Kerr	$50,000.00	$56,768.00	A
10	Clinton Hahn	$55,000.00	$51,725.00	A
11	Yusuf Cowan	$55,000.00	$49,705.00	A
12	Elisabeth Hensley	$55,000.00	$50,349.00	A
13	Jax Fuentes	$65,000.00	$68,607.00	A
14	Kylan Mcmillan	$55,000.00	$56,634.00	A
15	Kailey Cruz	$65,000.00	$70,634.00	A
16	Cody Galloway	$30,000.00	$10,000.00	A
17	Lizbeth Cameron	$55,000.00	$57,998.00	A
18	Kaila White	$65,000.00	$67,970.00	A
19	Ryker Mayo	$30,000.00	$32,565.00	A
20	Raquel Hopkins	$65,000.00	$60,292.00	A
21	Monique Butler	$55,000.00	$58,643.00	A

Original Dataset

	A	B	C	D	E	F	G
1	NAME	SALES TARGET	ACTUAL	DEPARTMEN	Target_Achievement_Lev	Bonus Percentage	Bonus Pay Increa
2	Johan Mcdaniel	$100,000.00	$153,470.00	D	0.535	10	$15,347.00
3	Jasper Torres	$30,000.00	$41,259.00	D	0.375	2	$825.18
4	Lola Humphrey	$100,000.00	$114,881.00	B	0.149	10	$11,488.10
5	Marvin Jacobs	$50,000.00	$57,370.00	C	0.147	5	$2,868.50
6	Addison Lloyd	$65,000.00	$74,529.00	B	0.147	5	$3,726.45
7	Darren Kerr	$50,000.00	$56,768.00	A	0.135	5	$2,838.40
8	Axel Moreno	$50,000.00	$56,070.00	C	0.121	5	$2,803.50
9	Yahir Casey	$55,000.00	$61,461.00	D	0.117	5	$3,073.05
10	Elliot Vaughn	$30,000.00	$33,521.00	A	0.117	2	$670.42
11	Braylon Gonzales	$65,000.00	$71,731.00	A	0.104	5	$3,586.55
12	Mason Gay	$30,000.00	$32,990.00	D	0.100	2	$659.80
13	Rocco Hooper	$30,000.00	$32,817.00	D	0.094	2	$656.34
14	Juliette Oneal	$50,000.00	$54,636.00	D	0.093	5	$2,731.80
15	Caleb Woodard	$30,000.00	$32,700.00	A	0.090	2	$654.00
16	Frances Andersen	$30,000.00	$32,665.00	B	0.089	2	$653.30
17	Arianna Mayer	$55,000.00	$59,830.00	D	0.088	5	$2,991.50
18	Helena Avery	$55,000.00	$59,814.00	D	0.088	5	$2,990.70

Dataset in Table with my Additional Calculation Columns

NUMBERS FOR DATA ANALYSIS - EVIDENCE

Sales person who has exceeded the sales target the most:

Johan Mcdaniel 0.535

Sales person who has underperformed the most w.r.t. the sales target:

Cody Galloway -0.667

Average performance of the overall sales team:

0.004

Q: Based on the calculated target achievement level for each agent, and based on average performance calculated above, reflect upon the fairness of sales targets that are set.
A: On average, the all of the sales staff are performing only slightly higher than the target sales amount (0.4%). I would therefore conclude that the sales targets that are set are fair.

Sales information by each departm

Average sales of an agent in each department:

Department	Total Sales	Average Sales of an Agent [A]	Average Target Achievement Level [B]	% Target Met
A	$1,285,338.00	$58,424.45	$58,863.64	-0.75
B	$774,088.00	$59,545.23	$58,461.54	1.85
C	$946,971.00	$52,609.50	$54,166.67	-2.87
D	$2,792,786.00	$59,420.98	$58,085.11	2.30

Q: Based on the above calculations, do you see a significant difference of performance between different departments?
A: In terms of total sales, there are clear differences between departments, for example department D's total sales were twice that of department A. This may be because of the volume to type of product being sold between departments. In terms of meeting targets, on average there were significant, but not big differences between departments, with % Target being met ranging from -2.87% (Dept. C) to +2.30 (Dept. D).

Bonus Percentages

Actual Sales	Bonus Percentage
Above 30000	2
Above 50000	5
Above 75000	8
Above 100000	10

Total budget for the organization to pay for bonuses:

$316,567.81

Total bonus payment for each department:

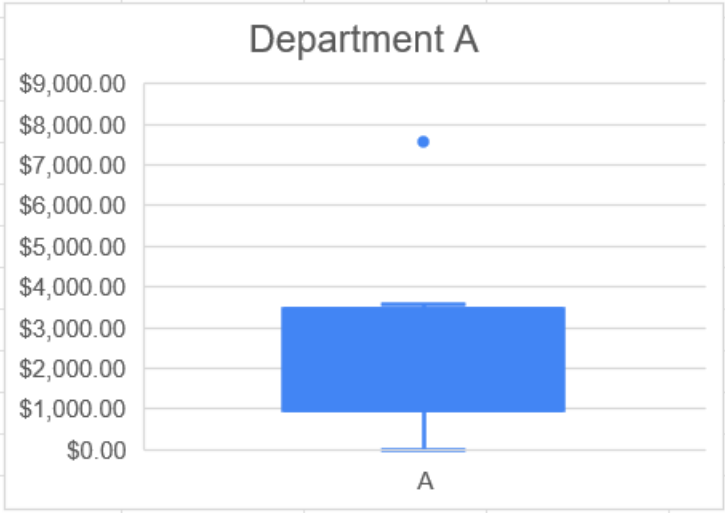
Department	Total Bonus Payment
A	$66,538.57
B	$46,029.10
C	$46,925.76
D	$157,074.38

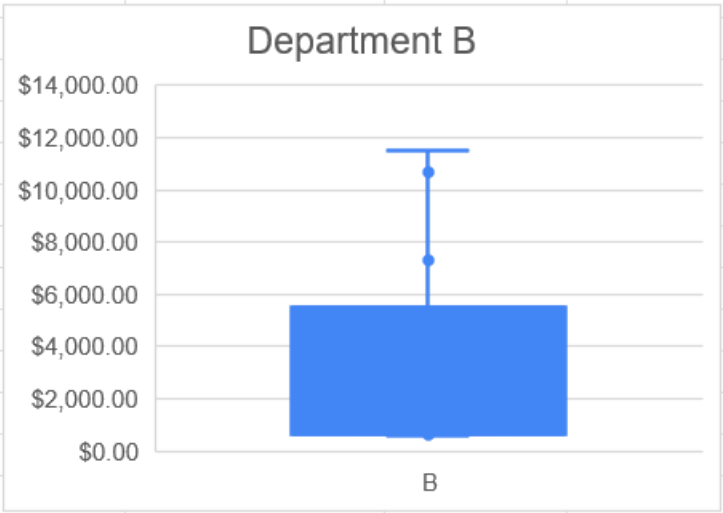
NUMBERS FOR DATA ANALYSIS - EVIDENCE

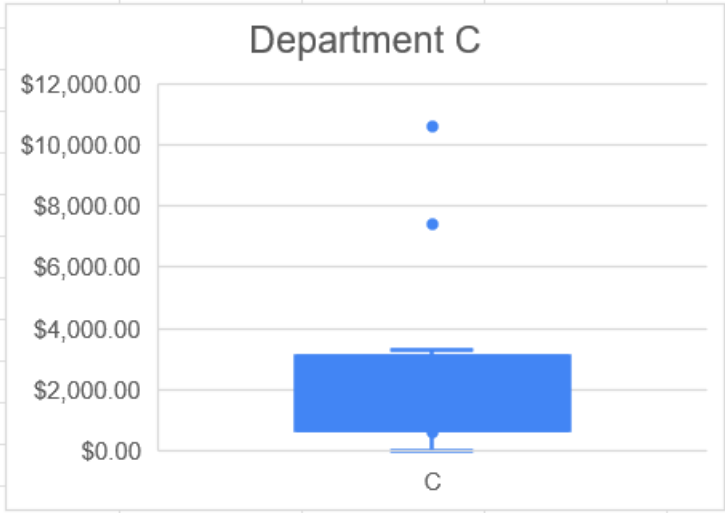
Q: Comment on the variation of performance for different departments based on bonus pay.
A:
Within Departments: With reference to the 'Box and Whisker' diagrams, within department A, the variation of performance based on bonus pay is skewed, with a a higher than normal proportion of staff underachieving. For department B, the opposite is true, with a higher than normal proportion of staff over-achieving. For departments C and D, the distributions are closer to normal distribution, with a slight bias towards underachievement of staff performance.

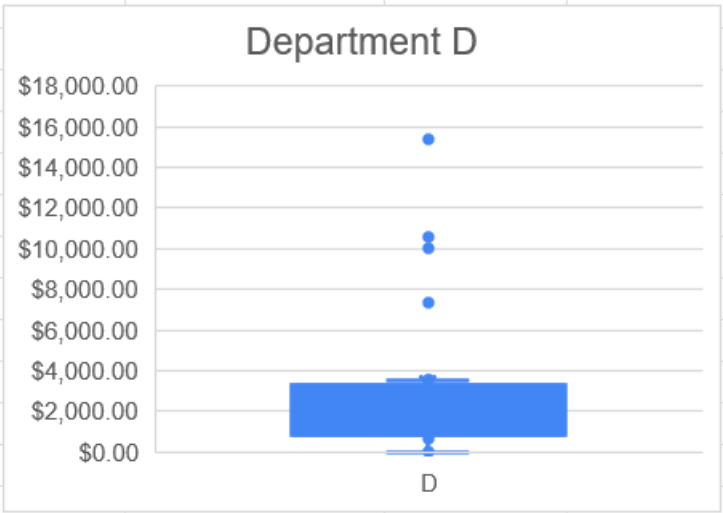
Between Departments: Based on bonus pay, departments B and C are performing at similar levels. Compared with departments B and C, department A is performing about 45% better, whilst department D three times better (data shown above).

Recommendations: New questions may look into why some departments are performing better than others. Is it because the items being sold in one department are higher in value per unit? Is the system of pay increase fair on employees when, as one example, Ciara Leonard in department C and Jerimiah Hart in department D both do not meet their target achievement levels (-0.074 and -0.076 respectively) but Ciara has a pay rise of $7,406.56 whilst Jerimiah does not get any pay rise at all?









DATA PREPARATION, QUALITY AND VALIDATION

This part of the bootcamp involved taking two datasets and demonstrating the following:

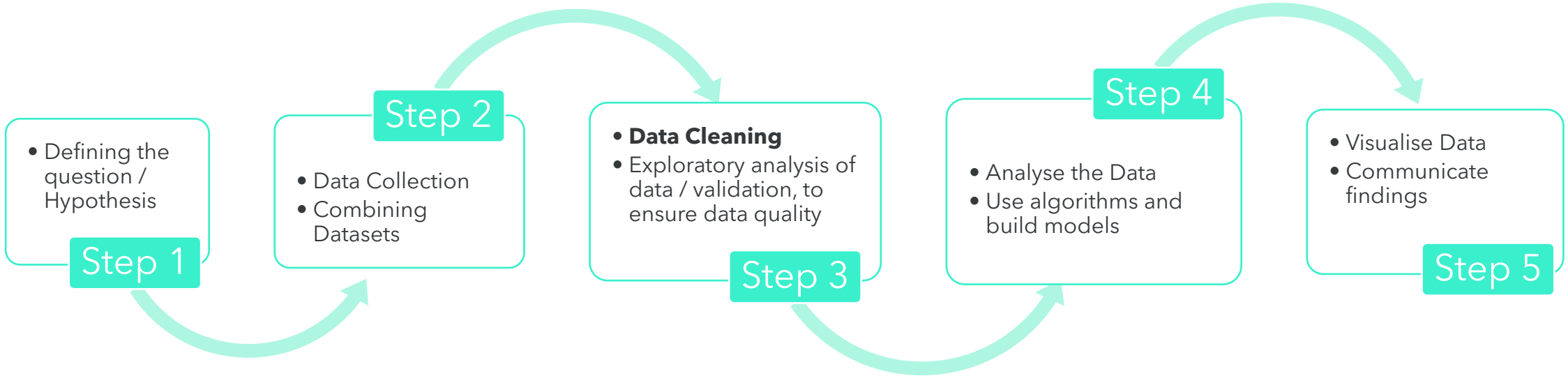
- Cleaning the data
- Analysing the data
- Combining the data
- Making any conclusions that could be drawn from the data

The first set contained heart disease data obtained from 4 different laboratory sites, Hungary (I sometimes refer to this as HU for convenience), Switzerland (CH), Virginia (VA), and Cleveland (CLE). The second set contained data from heart failure clinical records (location not specified).

Data Tools Used:

- Microsoft Excel, Including:
  - working with tables
  - data tools (e.g. sorting, filtering)
  - formulae (e.g. IF, AND, OR, MAXIFS, MINIFS, SUMIFS, AVERAGEIFS, ISNUMBER, ISTEXT, ISBLANK, COUNT, COUNTA, COUNTBLANK, COUNTIF)
  - Charts and basic statistics

DATA PREPARATION, QUALITY AND VALIDATION - EVIDENCE



Data cleaning is an essential process in data analysis. It is the process of identifying and correcting errors, inconsistencies, and outliers in a dataset. It is important because it can improve the quality, reliability, and usability of the data for analysis and decision making. Data cleaning can also reduce the risk of errors, biases, and misleading results that can affect the validity and credibility of the data.

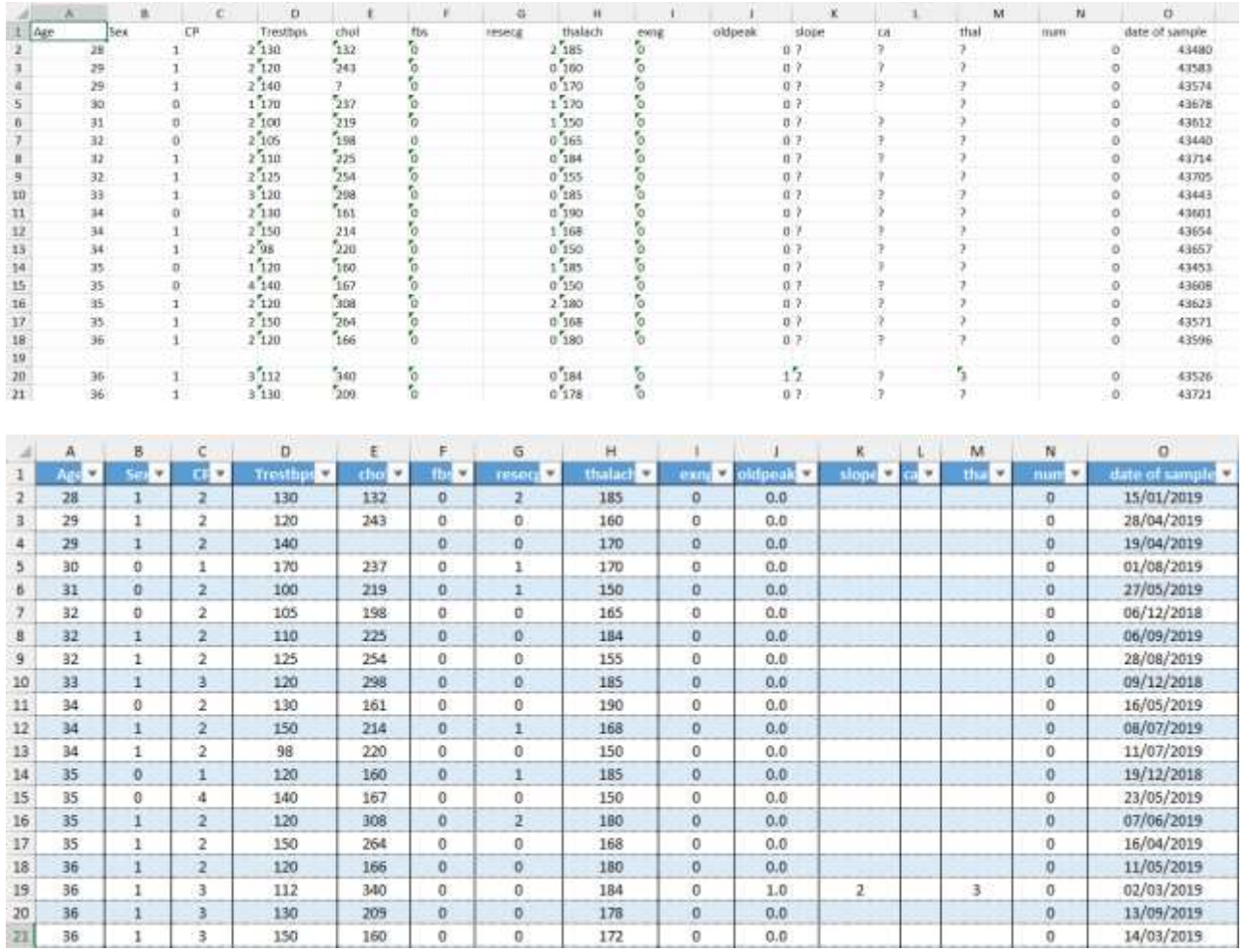
Before analysing the datasets in Excel, I needed to make sure that the source-data was free from any errors or “dirty data”.

DATA PREPARATION, QUALITY AND VALIDATION - EVIDENCE

For this exercise, the datasets were relatively clean and well ordered. However, some data cleaning was required, and this included the following:

- Checking that no columns need to be split and that each column from each data set has the correct column heading and that the columns are in corresponding orders between datasets.
- Converting numbers stored as text into numbers (using copy and paste multiplier method, or using the green (formula error) option in Excel to convert these.
- Getting rid of blank data rows (e.g. data rows without dates on)
- Removing obvious mistakes (by scanning the datasets and also converting them into tables and checking the values in the dropdown boxes at the top of each column) and duplicates (via the Remove Duplicates" dialog box.
- Formatting data (e.g. dates).
- Removing symbols. For example, many cells contained a single question mark. These were removed using the replace function (in Excel, this is done by replacing "~?" with "".

DATA PREPARATION, QUALITY AND VALIDATION - EVIDENCE



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Age	Sex	CP	Trestbps	chol	fbs	resecg	thalach	exng	oldpeak	slope	ca	thal	num	date of sample
2	28	1	2	130	132	0	2	185	0	0.0				0	15/01/2019
3	29	1	2	120	243	0	0	160	0	0.0				0	28/04/2019
4	29	1	2	140		0	0	170	0	0.0				0	19/04/2019
5	30	0	1	170	237	0	1	170	0	0.0				0	01/08/2019
6	31	0	2	100	219	0	1	150	0	0.0				0	27/05/2019
7	32	0	2	105	198	0	0	165	0	0.0				0	06/12/2018
8	32	1	2	110	225	0	0	184	0	0.0				0	06/09/2019
9	32	1	2	125	254	0	0	155	0	0.0				0	28/08/2019
10	33	1	3	120	298	0	0	185	0	0.0				0	09/12/2018
11	34	0	2	130	161	0	0	190	0	0.0				0	16/05/2019
12	34	1	2	150	214	0	1	168	0	0.0				0	08/07/2019
13	34	1	2	98	220	0	0	150	0	0.0				0	11/07/2019
14	35	0	1	120	160	0	1	185	0	0.0				0	19/12/2018
15	35	0	4	140	167	0	0	150	0	0.0				0	23/05/2019
16	35	1	2	120	308	0	2	180	0	0.0				0	07/06/2019
17	35	1	2	150	264	0	0	168	0	0.0				0	16/04/2019
18	36	1	2	120	166	0	0	180	0	0.0				0	11/05/2019
19	36	1	3	112	340	0	0	184	0	1.0	2		3	0	02/03/2019
20	36	1	3	130	209	0	0	178	0	0.0				0	13/09/2019
21	36	1	3	150	160	0	0	172	0	0.0				0	14/03/2019

Original Data (Top) vs Cleaned Data (Bottom)

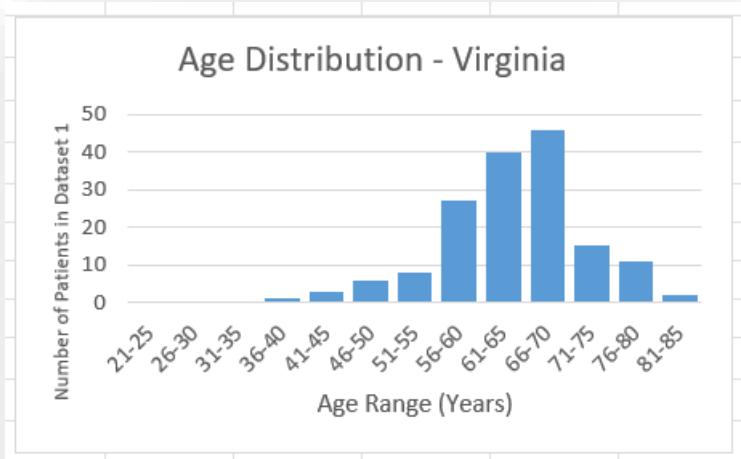
After cleaning the data, I carried out some analysis on the first dataset, according to the following:

- Working out the average age of the patients in each location
- Plotting a histogram of age variables on the dataset (VA only)
- Working out in which month the most samples were received (VA only)

My analysis and any findings are shown on the next page:

DATA PREPARATION, QUALITY AND VALIDATION - EVIDENCE

Average age of the patients:		
Virginia	59.4	Years
Hungary	47.8	Years
Switzerland	55.3	Years
Cleveland	54.3	Years



The above distribution is skewed left.
Note: A distribution is called **skewed left** if the left tail (smaller values) is much longer than the right tail (larger values).

AE18 | =VLOOKUP(AE16,AE4:AF15,2,FALSE)

	AC	AD	AE	AF	AG	AH
2			**Samples Received Each Month**			
3			Virginia			
4			10	January		
5			4	February		
6			7	March		
7			12	April		
8			14	May		
9			11	June		
10			11	July		
11			8	August		
12			13	September		
13			13	October		
14			44	November		
15			12	December		
16		Max	44			
17		No. of Max*	1			
18	Month Max Sam Rec.		***November***			
19						
20						
21						
22						
23						

For Virginia, the most samples were received in November

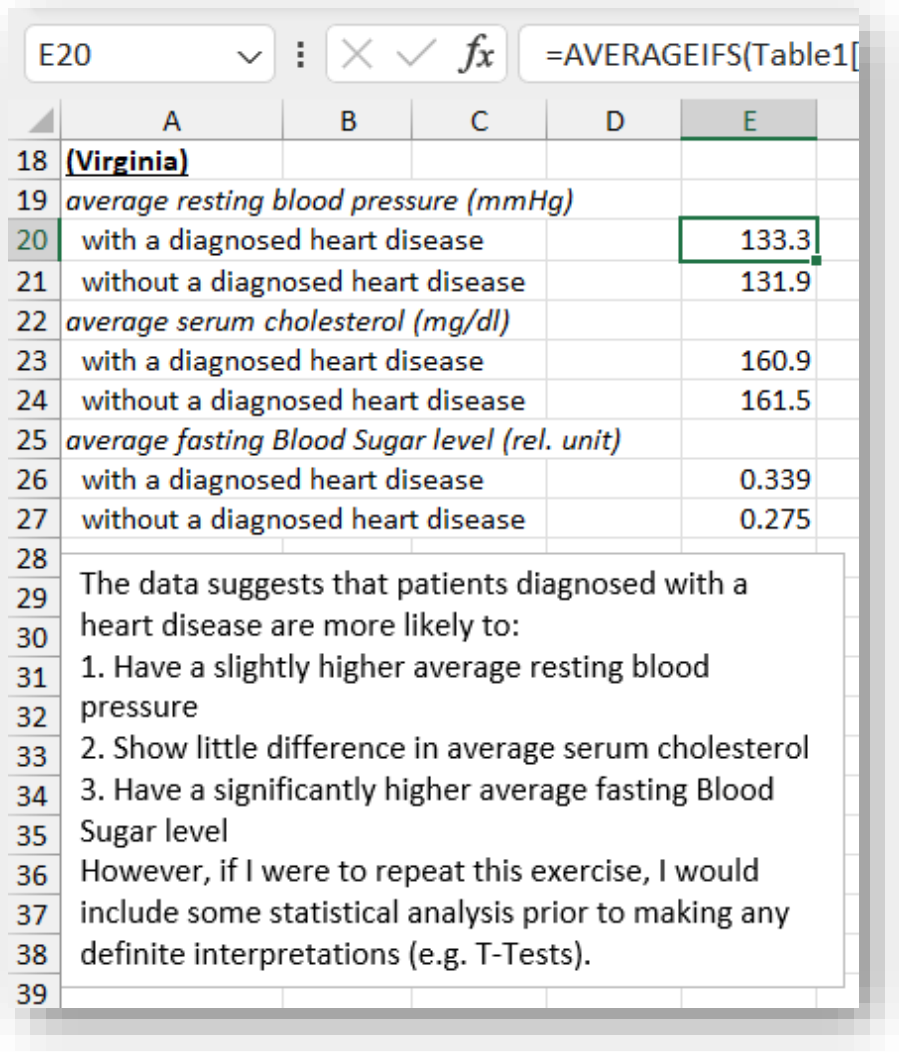
* If value = 1, this verifies that only one month has the maximum number of samples received

DATA PREPARATION, QUALITY AND VALIDATION - EVIDENCE

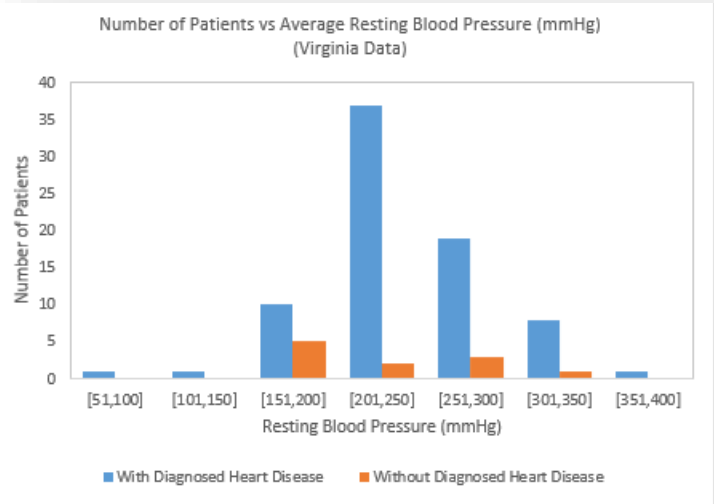
Next, the data was analysed to see if there was any connection between the presence or absence of a heart disease and the following factors:

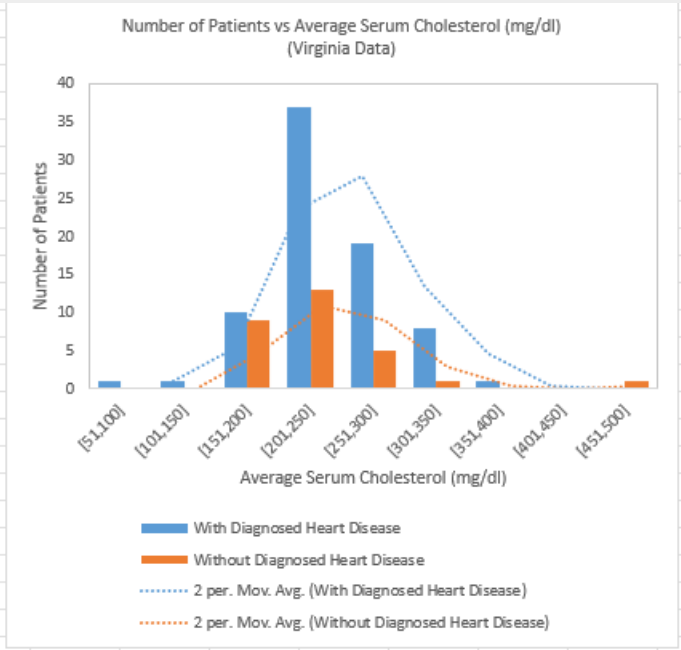
- Resting blood pressure
- Serum cholesterol
- Fasting blood sugar level (f.b.s.)

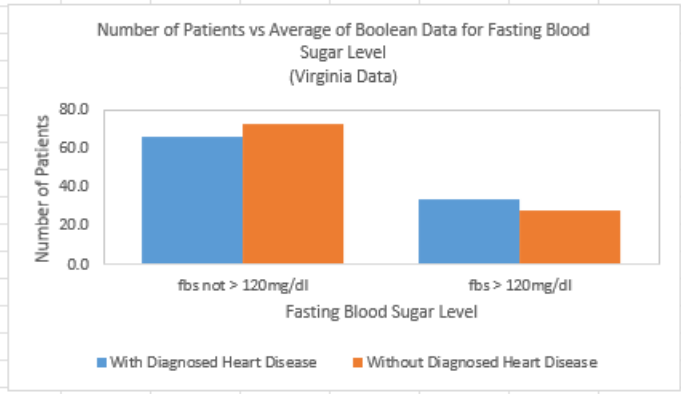
My analysis and any findings are shown here, and on the next page:



DATA PREPARATION, QUALITY AND VALIDATION - EVIDENCE







Looking at the *relative ratios* of each chart, it appears that the higher the blood pressure, the higher is the probability of having a heart disease, as is with the case of cholestero levels and to a lesser degree, fasting Blood Sugar level.

Q: What other analysis measure can you do to investigate the relationship between the key indicators and existence of heart disease?

A: Other data available in the tables that that could be used to investigate any possible relationships with the existence of heart disease include Chest-pain type, Electrocardiogram (ECG) data, Maximum heart rate achieved, exercise induced angina data, ST depression induced by exercise relative to rest, Peak exercise ST segment, the number of major vessels (0–3) colored by flourosopy, and the presence of the inherited disorder thalassemia.

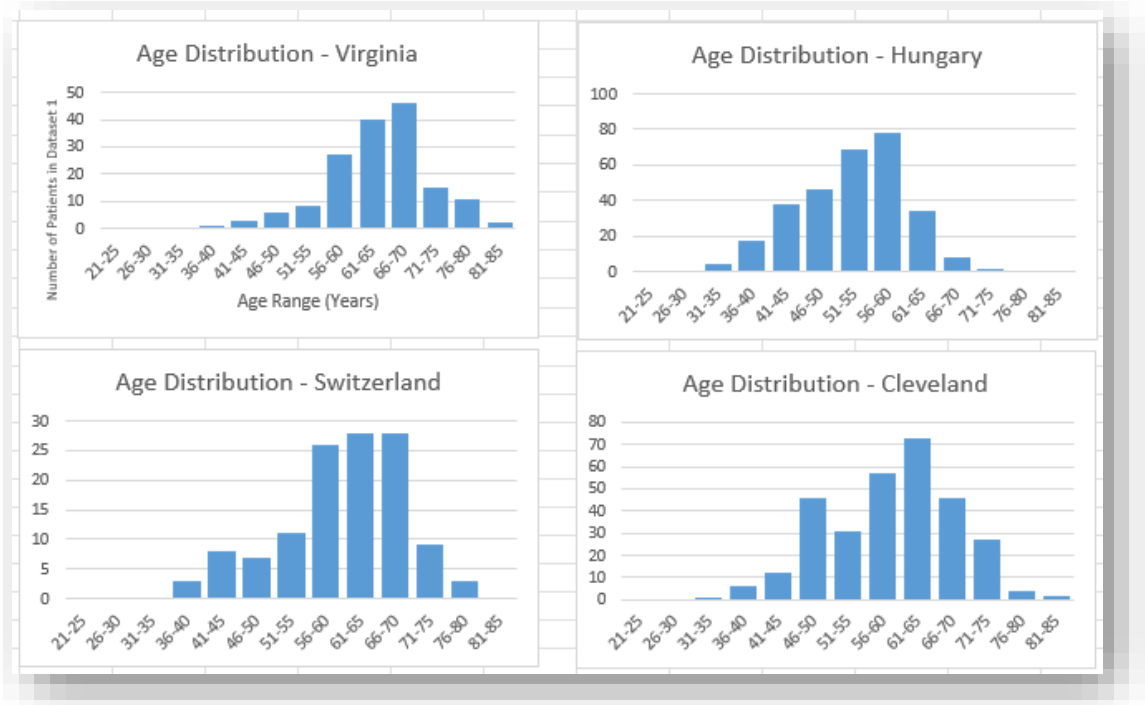
Data could also be shown, for example, in a box and whisker plot for comparisons between the key indicators and existence of heart disease; the shape of the plot would show how the data is distributed.

DATA PREPARATION, QUALITY AND VALIDATION - EVIDENCE

A comparison study was then conducted to see if the same conclusions could be made using data from the different laboratory sites, regarding the presence or absence of a heart disease and 1) Resting blood pressure, 2) Serum cholesterol, and 3) Fasting blood sugar level (fbs).

Again, My analysis and any findings are shown here, and on the next page:

Average age of the patients:		
Virginia	59.4	Years
Hungary	47.8	Years
Switzerland	55.3	Years
Cleveland	54.3	Years



DATA PREPARATION, QUALITY AND VALIDATION - EVIDENCE

	Virginia	Hungary	Switzerland	Cleveland	Combined
average resting blood pressure (mmHg)					
with a diagnosed heart disease	133.3	135.8	130.6	134.6	133.5
without a diagnosed heart disease	131.9	130.7	124.4	129.3	130.0
average serum cholesterol (mg/dl)					
with a diagnosed heart disease	160.9	269.2	No Data	251.5	253.9
without a diagnosed heart disease	161.5	239.6	No Data	242.7	240.1
average fasting Blood Sugar level (rel. unit)					
with a diagnosed heart disease	0.339	0.125	0.106	0.158	0.194
without a diagnosed heart disease	0.275	0.044	No Sig. Data	0.139	0.108

By comparing the different laboratory data, further conclusions were made as follows:

1. Whereas it was noted that in Virginia patients diagnosed with a heart disease were more likely to have a slightly higher average resting blood pressure, the other 3 laboratories provided better support to the idea that those with a heart disease do indeed have a higher average resting blood pressure.
2. Whilst the Virginia laboratory data gave no indication that patients diagnosed with a heart disease have higher average serum cholesterol levels, data from Cleveland and Hungary in particular, suggests that they do (no data from Switzerland).
3. Data from Hungary and Cleveland matches the data from Virginia with respect to patients diagnosed with a heart disease having a significantly higher average fasting Blood Sugar level (no useful data from Switzerland).

By analysing data from all the laboratories combined, is there a relationship between the sex of patient and the existence of heart disease?

Patents by Gender

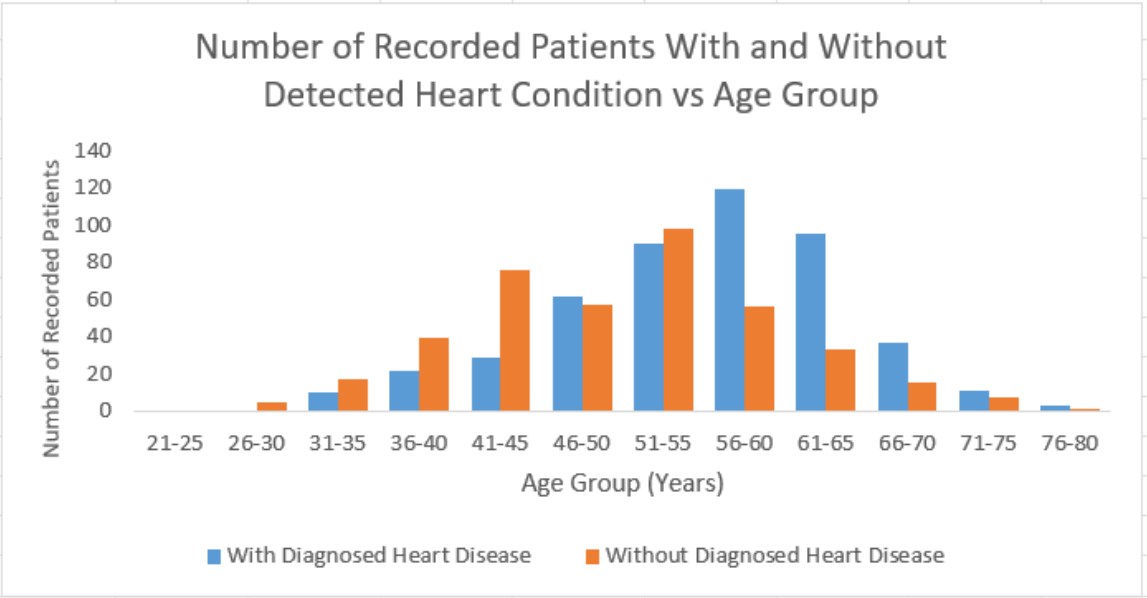
	Male (1)	Female (0)
With a diagnosed heart disease	429	49
Without a diagnosed heart disease	260	144
Total	689	193
% with a diagnosed heart disease	62.3	25.4
% without a diagnosed heart disease	37.7	74.6

From the data is is possible to see that of the patients tested, a much higher percentage of males (62.3%) had a detected heart disease, than females (25.4%).

DATA PREPARATION, QUALITY AND VALIDATION - EVIDENCE

Similarly, investigate if you can see any relationship between the age of a patient and the existence of heart disease. Hint: try to sort the dataset based on age, and then work on several bins (i.e. range of age).

	H.D.	No H.D.
Bin	*Frequency*	*Frequency*
21-25	0	0
26-30	0	5
31-35	10	17
36-40	22	39
41-45	29	76
46-50	62	57
51-55	90	98
56-60	119	56
61-65	95	33
66-70	37	15
71-75	11	7
76-80	3	1
More	0	0



The distribution of the histogram suggests that the the most common age range for patients with heart disease is 55 to 60 years of age, whereas in patients without a diagnosed heart disease, the most common age range is 50 to 55 years of age.

DATA PREPARATION, QUALITY AND VALIDATION - EVIDENCE

The next part of this study involved looking at the fields of data in the Heart failure clinical records dataset, the objective being to expand my investigations based on combining the two datasets. Whilst combining the two sets of data, age was an obvious common-variable between the two data sets, whilst not being a medical practitioner, I was only able to speculate which data variables may or may-not have any relationship to one another. The fields of each dataset are shown below. For dataset 1, I removed data from people with no detected heart condition and age bins where they only appeared in one dataset. The findings from my investigations are shown on the following pages:

1. Age (age)
2. Sex (sex)
3. Chest Pain (cp)
4. Resting blood pressure (trestbps): in mm Hg on admission to the hospital)
5. Serum cholestoral in mg/dl (chol)
6. Fasting blood sugar > 120 mg/dl (fbs): 1 = true; 0 = false
7. Resting electrocardiographic results (restecg)
   - Value 0: normal
   - Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
   - Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8. Maximum heart rate achieved (thalach)
9. Exercise induced angina (exang): 1 = true; 0 = false
10. ST depression induced by exercise relative to rest (oldpeak)
11. The slope of the peak exercise ST segment (slope)
    - Value 1: upsloping
    - Value 2: flat
    - Value 3: downsloping
12. Number of major vessels (0-3) colored by flourosopy (ca)
13. Thal (thal) 3 = normal; 6 = fixed defect; 7 = reversable defect
14. Diagnosis of heart disease (num)
    - Value 0: < 50% diameter narrowing
    - Value 1: > 50% diameter narrowing
15. Date of sample (date of sample)*

1. age: age of the patient (years)
2. anaemia: decrease of red blood cells or hemoglobin (boolean)
3. high blood pressure: if the patient has hypertension (boolean)
4. creatinine phosphokinase (CPK): level of the CPK enzyme in the blood (mcg/L)
5. diabetes: if the patient has diabetes (boolean)
6. ejection fraction: percentage of blood leaving the heart at each contraction (percentage)
7. platelets: platelets in the blood (kiloplatelets/mL)
8. sex: woman or man (binary)
9. serum creatinine: level of serum creatinine in the blood (mg/dL)
10. serum sodium: level of serum sodium in the blood (mEq/L)
11. smoking: if the patient smokes or not (boolean)
12. time: follow-up period (days)
13. [target] death event: if the patient deceased during the follow-up period (boolean)
14. Date of Event (date)

Attributes from dataset 1, Heart Disease Data Set on UCI Repository (left) and attributes from dataset 2, Heart failure clinical records (above)

DATA PREPARATION, QUALITY AND VALIDATION - EVIDENCE

Which fields of the datasets that are related in their meaning? Read the data descriptions carefully.elds of data in the Heart failure clinical records Data Set.

Fields of two datasets possibly related in their meaning:

Equivalent	Age
Related?	High Blood Pressure (Boolean) vs Resting Blood Pressure (mmHg)
Related?	Diabetes (boolean) vs Fasting Blood Sugar (although diabetics may be on insulin)
Related?	Ejection Fraction vs Resting Blood Pressure
Related?	Cholesterol vs Serum Sodium
Related?	Heart Disease vs Death Event
Related?	Smoking vs Diagnosis of Heart Disease
Equivalent	Sex

Q: How can you pre-process some features of the “Heart failure clinical records Data Set” to utilize it in synergy with the “Heart Disease dataset”?
A: Data can be separated into age and/or gender bins, as appropriate to the data anaysis. Numbers checked that they are values and not text. Also, some ages have decimal places and should be rounded down.
The overall age range (lowest to highest) for the second data set is different from the first dataset and so if utilising the two data sets in synergy with each other, data will need to be truncated with respect to age. Some data may also have to be removed if an age range bin has a low sample size (determining exact value of this ithrough the relevant statistical method is beyond the scope of this demonstrative exercise, but as an example value, sample sizes of less than 10 are disregarded in the overall data sets).

Combined Data, where:
1 = Heart Disease Dataset (Combined from 4 Laboratories*)
2 = Heart Failure Clinical Records
* (Patients with Heart Condition Only)

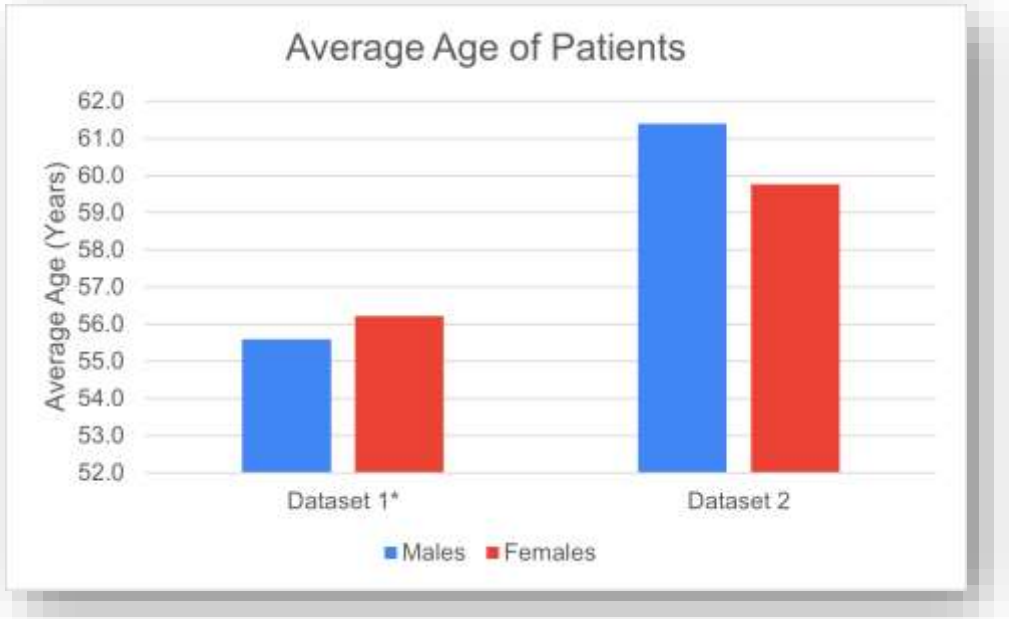
Average Age of Patients

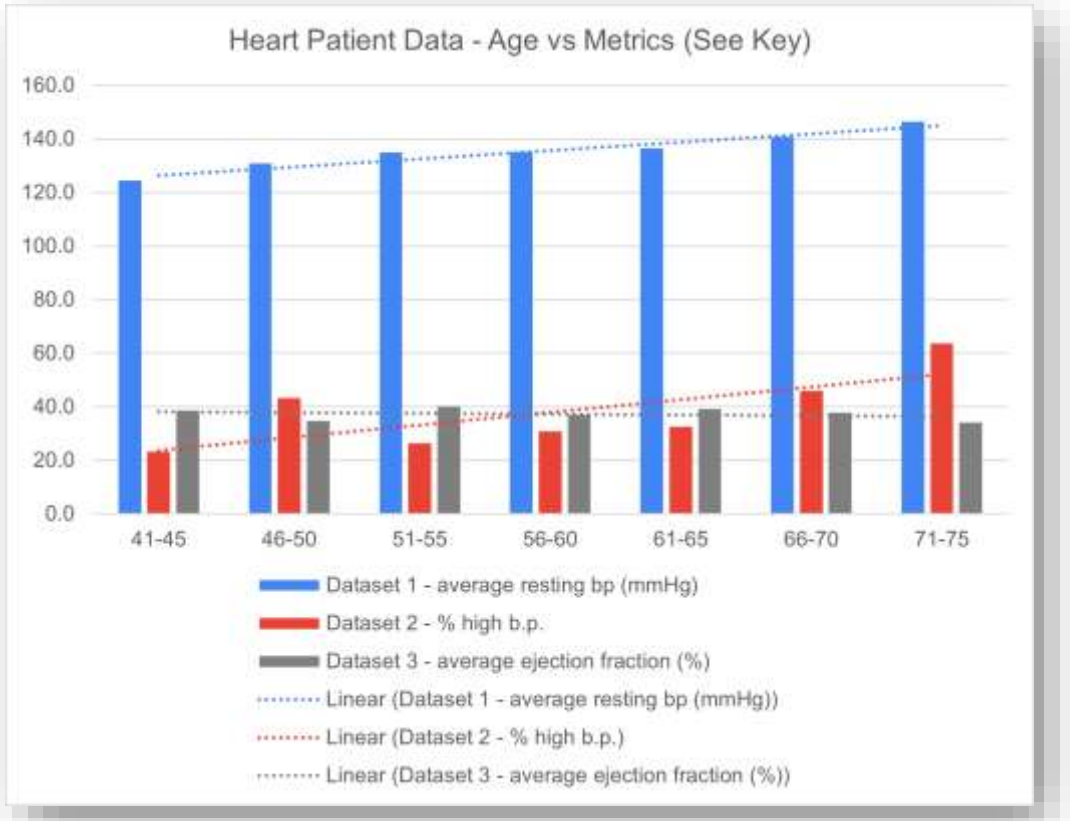
	Males	Females
Dataset 1*	55.6	56.2
Dataset 2	61.4	59.8

Total Number of Patients

	Males	Females
Dataset 1*	429	49
Dataset 2	194	105

DATA PREPARATION, QUALITY AND VALIDATION - EVIDENCE

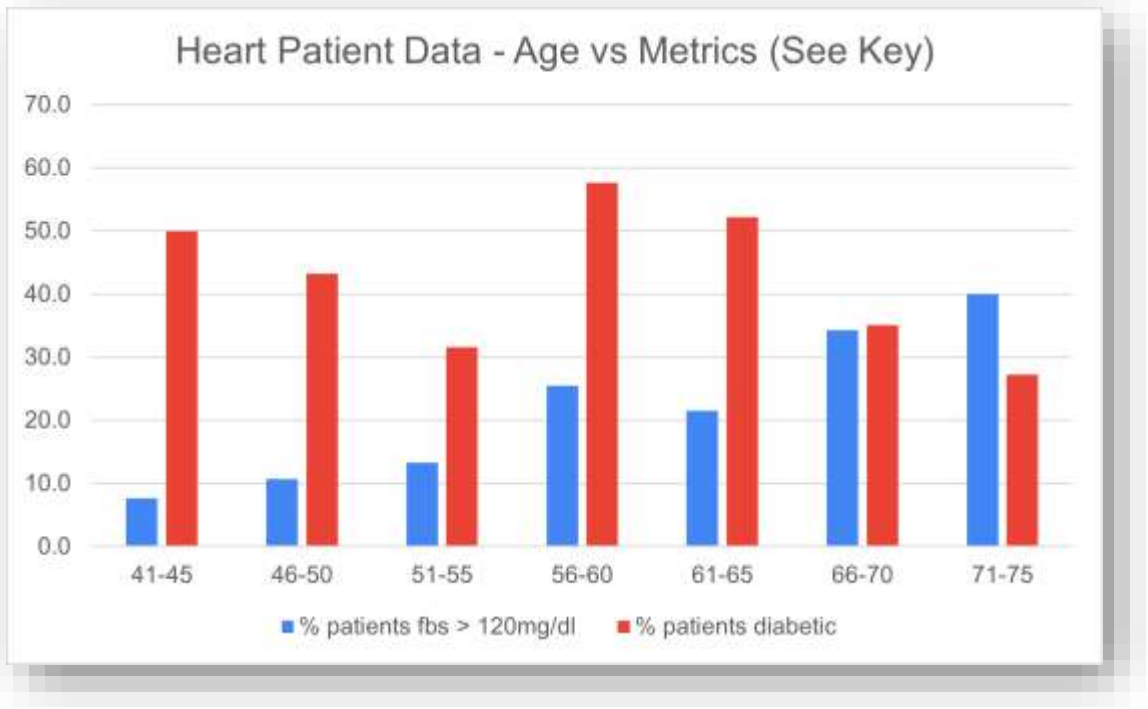




Some observations from the combined data:

1. As one might expect, the average age of patients with reported heart failure is higher than those from dataset 1, Heart Disease Data Set (excluding patients with no detected heart condition)
2. As age increases in dataset 1, so does average resting blood pressure. We see a similar trend in dataset 2 with % high blood pressure, but not with average ejection fraction.
3. Further work would apply statistical analysis of the data before any definite conclusions could be made. Also note, as mentioned earlier in this portfolio, that correlation does not mean causation.

DATA PREPARATION, QUALITY AND VALIDATION - EVIDENCE



Some observations from the combined data:
1. As age increases, whilst the percentage of patients with a high resting blood sugar increases, there is no apparent similar pattern occurring from dataset 2 with the percentage of patients who are diabetic.
2. Without knowing much about the significance of each variable in the data, no further significant trends were observed with the combined data in the time available for this exercise.
3. Again, further work would apply statistical analysis of the data before any definite conclusions could be made. Also note, as mentioned earlier in this portfolio, that correlation does not mean causation.

EXCEL FOR DATA ANALYSIS

This section of the bootcamp involved furthering my understanding of data analysis in Excel, continuing on the theme of collating, analysing, and presenting data according to best practices.

In the following exercise, I used dataset 1 from the previous section again, but this time importing the data using Power Query and manipulating the data in pivot tables.

Data Tools Used:

- Microsoft Excel, Including:
  - power query
  - data tools and formulae mentioned in last section
  - pivot tables

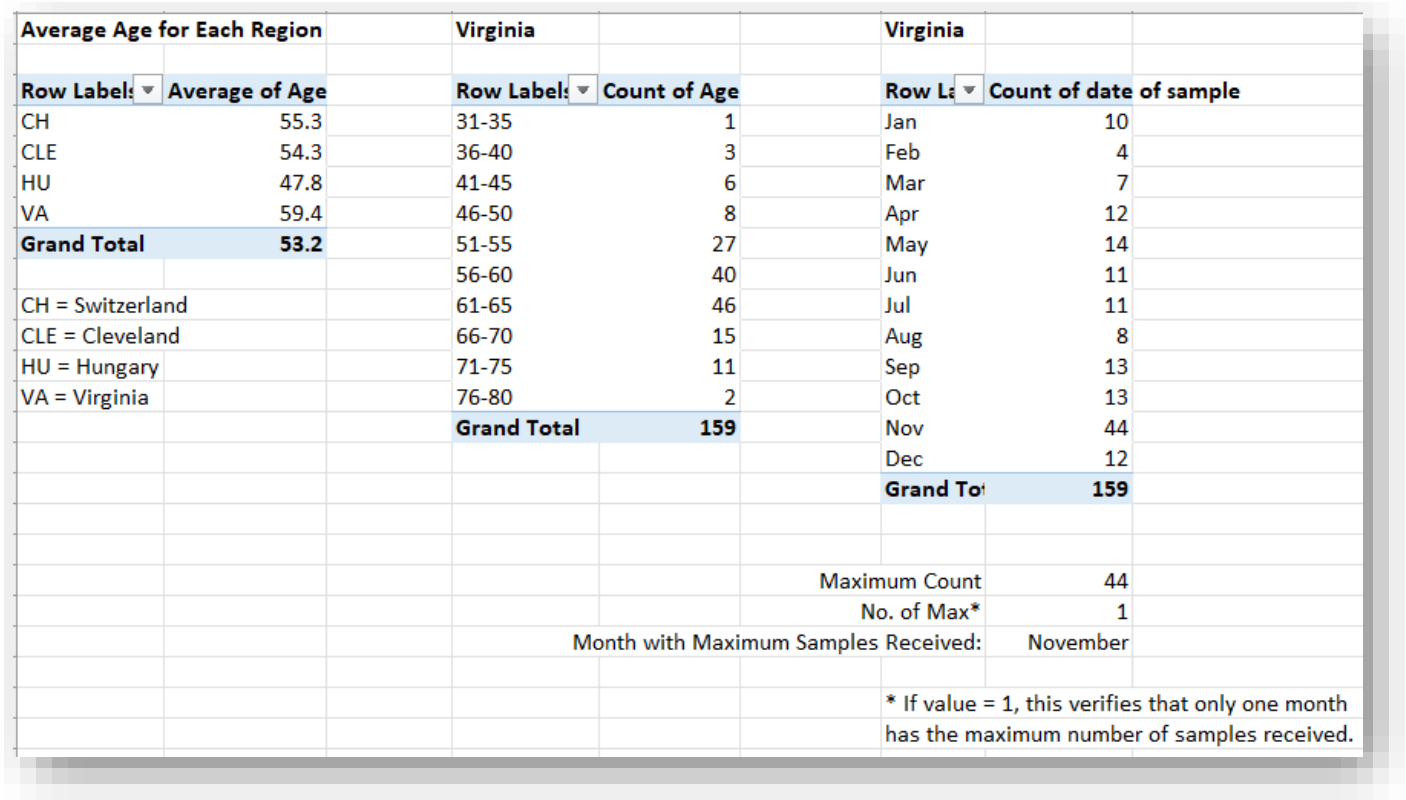
EXCEL FOR DATA ANALYSIS

In this part of the bootcamp, I learnt that Power Query is a useful tool for data analysis and transformation as it allows the data analyst to connect to various data sources, apply filters, merge tables, and create custom columns, amongst other useful features. Power Query helps with automating tasks and simplifying the data preparation process. It can help with creating reports and dashboards that are updated automatically.

I also discovered how useful pivot tables are, as they allow the data analyst to summarise and analyse large amounts of data quickly and easily. Pivot tables can be used to create reports, charts, dashboards, and other visualizations that help with understanding the data better. They also allow the user to filter, sort, group, and rearrange data in different ways, without changing the original source data.

EXCEL FOR DATA ANALYSIS - EVIDENCE

Using Power query to import my pre-cleaned data from the last section of the bootcamp, I created a series of pivot tables to complete some further analysis and produced some charts to help visualise the data. In summary, I found the whole process much more efficient using the pivot tables and they helped me with being able to view different comparisons of data much quicker than previously. The data from this pivot table work also helped confirm my previous findings from the dataset.
Some screenshots of my work from this section of the bootcamp follow.



Average Age for Each Region	
Row Labels	Average of Age
CH	55.3
CLE	54.3
HU	47.8
VA	59.4
Grand Total	**53.2**

CH = Switzerland
CLE = Cleveland
HU = Hungary
VA = Virginia

Virginia	
Row Labels	Count of Age
31-35	1
36-40	3
41-45	6
46-50	8
51-55	27
56-60	40
61-65	46
66-70	15
71-75	11
76-80	2
Grand Total	**159**

Virginia	
Row La	Count of date of sample
Jan	10
Feb	4
Mar	7
Apr	12
May	14
Jun	11
Jul	11
Aug	8
Sep	13
Oct	13
Nov	44
Dec	12
Grand To	**159**

Maximum Count	44
No. of Max*	1
Month with Maximum Samples Received:	November

* If value = 1, this verifies that only one month has the maximum number of samples received.

Pivot table data concurring with previous findings.

EXCEL FOR DATA ANALYSIS - EVIDENCE

All 4 Laboratories: Blood Pressure, Serum Cholesterol, and Fasting Blood Sugar Data (and appended data)

Row Labels	Average of Trestbps	Average of chol	Average of oldpeak	Average of slope
CH	**130.2**	**0.0**	**0.65**	**1.80**
<1	124.4	0.0	0.54	1.75
1-4	130.6	0.0	0.66	1.81
CLE	**131.7**	**246.7**	**1.04**	**1.61**
<1	129.3	242.7	0.59	1.42
1-4	134.6	251.5	1.57	1.83
HU	**132.5**	**250.6**	**0.59**	**1.90**
<1	130.7	239.6	0.22	1.66
1-4	135.8	269.2	1.25	2.01
VA	**133.0**	**161.1**	**1.42**	**2.14**
<1	131.9	161.5	0.77	2.08
1-4	133.3	160.9	1.61	2.15
Grand Total	**132.0**	**196.9**	**0.88**	**1.77**

<1 Value denotes where there was no detected heart disease
1-4 denotes heart disease detected
CH = Switzerland
CLE = Cleveland
HU = Hungary
VA = Virginia
Trestbps (Resting Blood Pressure units = mmHg
Serum Cholesterol Units = mg/dl
Appended Data for investigative purposes:
ST depression induced by exercise relative to rest
Peak exercise ST segment Peak exercise ST segment: 1 = upsloping, 2 = flat, 3 = downsloping)

Row Labels	Count of fbs
CH	**48**
<1	**1**
0	1
1-4	**47**
0	42
1	5
CLE	**305**
<1	**166**
0	143
1	23
1-4	**139**
0	117
1	22
HU	**287**
<1	**183**
0	175
1	8
1-4	**104**

% Patients: Setected Heart Disease and Fasting Blood Sugar

Location	Det. Hear	fbs not > 120mg/d	fbs > 120mg/dl
CH	No	No Sig. Data	No Sig. Data
	Yes	89.4%	10.6%
CLE	No	86.1%	13.9%
	Yes	84.2%	15.8%
HU	No	95.6%	4.4%
	Yes	87.5%	12.5%
VA	No	72.5%	27.5%
	Yes	66.1%	33.9%

EXCEL FOR DATA ANALYSIS - EVIDENCE

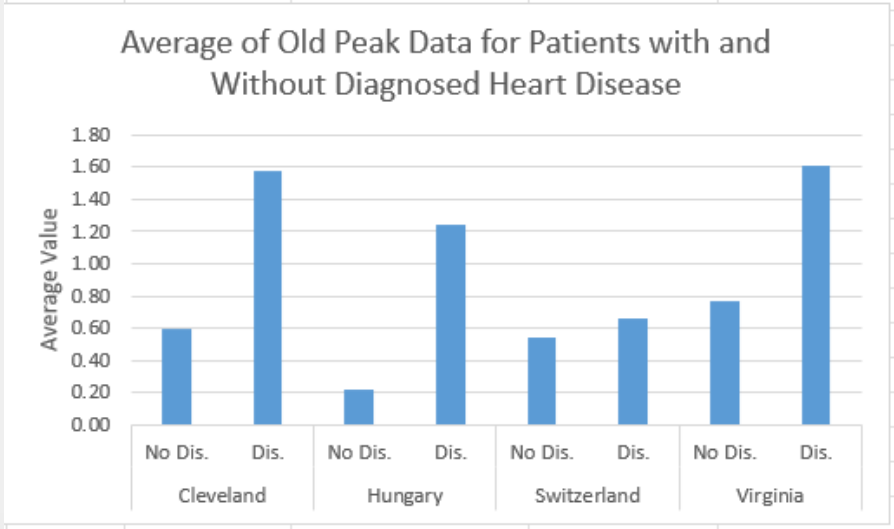
Data Extracted from Pivot Table:

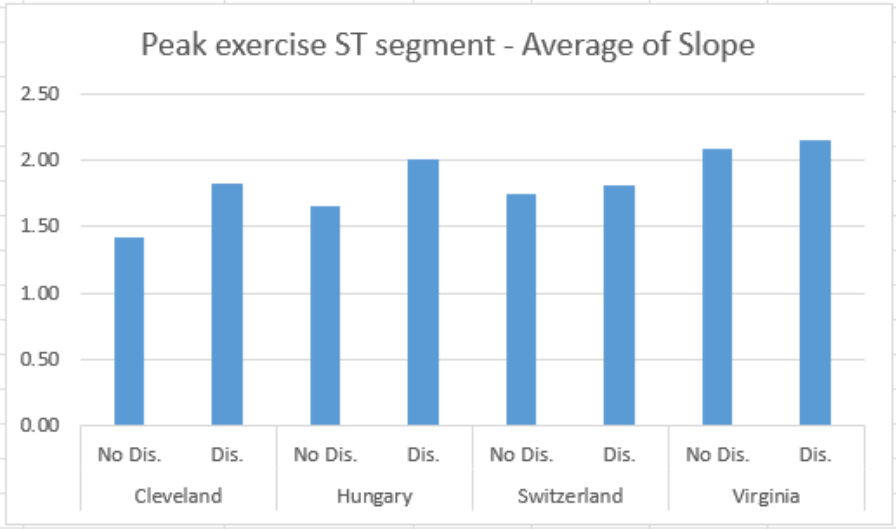
		Average of oldpeak	Average of slope
Cleveland	No Dis.	0.59	1.42
	Dis.	1.57	1.83
Hungary	No Dis.	0.22	1.66
	Dis.	1.25	2.01
Switzerlan	No Dis.	0.54	1.75
	Dis.	0.66	1.81
Virginia	No Dis.	0.77	2.08
	Dis.	1.61	2.15

ST depression induced by exercise relative to rest

Peak exercise ST segment Peak exercise ST segment: 1 = upsloping, 2 = flat, 3 = downsloping)

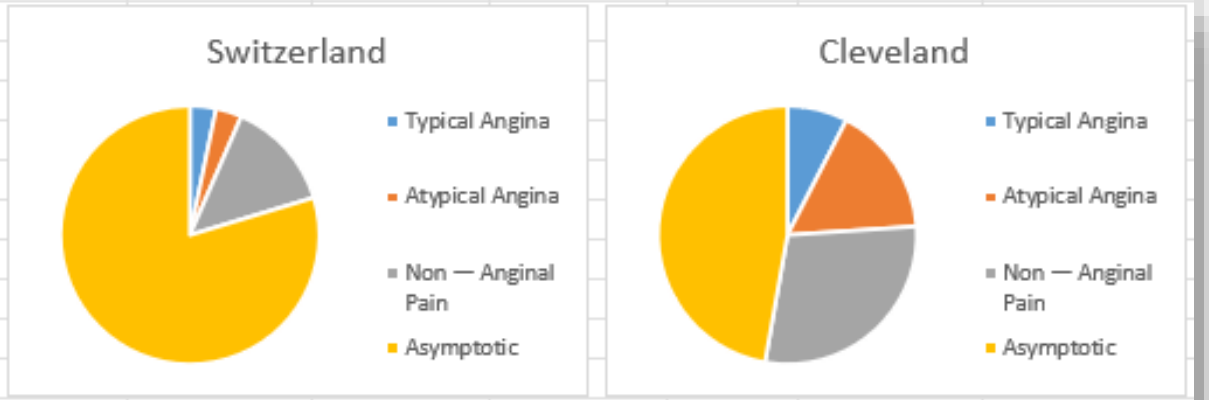
Conclusion from this Data:
All laboratories show identical trends, whereby both Average of Old Peak Data and Peak exercise ST segment (Average of Slope) are increased in people who have a diagnosed heart disease.





EXCEL FOR DATA ANALYSIS - EVIDENCE

Row Labels	Count of CP
CH	123
1	4
2	4
3	17
4	98
CLE	305
1	23
2	50
3	88
4	144
HU	295
1	11
2	107
3	54
4	123
VA	159
1	7
2	11
3	39
4	102
Grand Total	882



Here, some conclusions can be made from the data extracted via a pivot table, including, for example, a much higher proportion of patients with a diagnosed heart disease in Switzerland being asymptomatic, with this value being lowest in Hungary. Futher investigation might reveal Swiss patients undergo better, or routine heart-monitoring, for example.

CHARTS AND VISUALISATIONS

This section of the bootcamp looked at the different ways data can be visualised in Excel via charts and included how to create an Excel dashboard. Creating dashboards in Excel is a way to summarise and visualise data from different sources. Dashboards can help businesses monitor key performance indicators, track trends, and identify patterns.

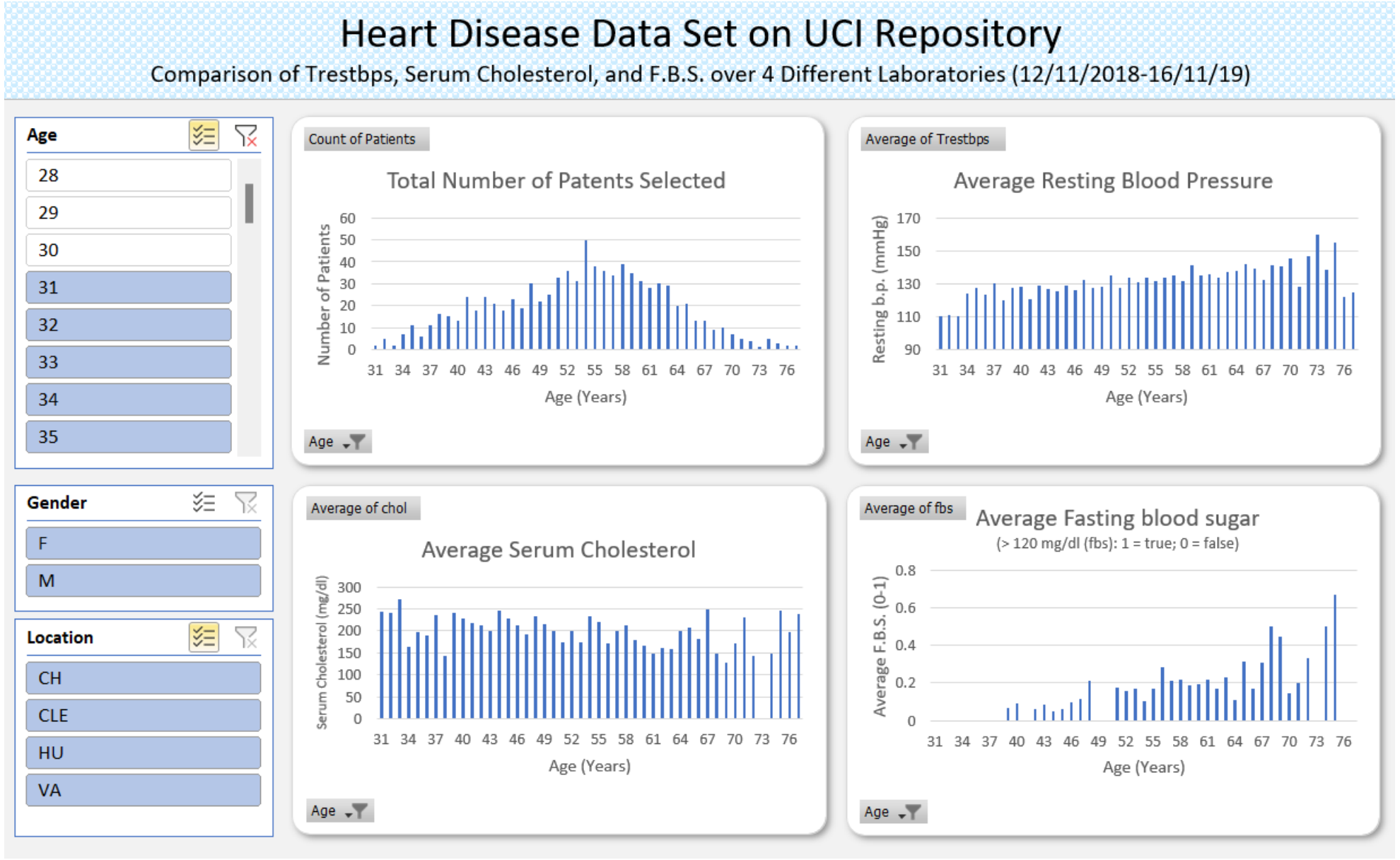
An example of my work is shown on the following page. For this, I decided to use the trestbps, serum cholesterol, and f.b.s. data from the previous exercises and include some sliders to enable the user to compare date by age, gender, and laboratory location. The pre-cleaned data was imported, placed in pivot tables and then pivot charts were created with sliders (linked to each chart using the 'report connections' feature). The page was then formatted for aesthetics.

Note that the choice of charts used in this example was not varied, although they reflected the type of data being presented here, just as an example. If I were presenting sales data, as another example, I could include pie charts showing the popularity of different product categories, histograms or bubble charts to represent demographics, and/or include a timeline slider so that period of sales could be adjusted by the user/viewer of the dashboard.

Data Tools Used:

- Microsoft Excel, Including:
  - power query
  - pivot charts
  - slicers
  - dashboarding

CHARTS AND VISUALISATIONS - EVIDENCE



DASHBOARDING WITH MS POWER BI

Microsoft Power BI is a tool for business intelligence that enables users to connect to and visualise any data. With Power BI, it is possible to create reports and dashboards with various types of charts, maps, and graphs, and share them with others within or external to an organisation.

In this section, I demonstrate an example from my learning whereby I have loaded, transformed, and created relationships between raw data sets, conducted basic analysis, and visually presented the data for review and analysis by others using different tools with Power BI.

Data Tools Used:

- Power Query
- Microsoft Power BI
- DAX (Introductory Level)

DASHBOARDING WITH MS POWER BI - EVIDENCE

In this example, a company (Adventure Works) wishes to use the data contained within its Product Report to understand individual product performance over time and on a monthly basis. They also want a forecast of profit for the next 2 months.

- They would also like to understand more about their customers.
- Which types of customers have the highest number of orders and revenue?

AdventureWorks_Territories	09/11/2023 23:59	Microsoft Excel Comma Separated Values File	1 KB
AdventureWorks_Sales_2017	09/11/2023 23:59	Microsoft Excel Comma Separated Values File	1,357 KB
AdventureWorks_Sales_2016	09/11/2023 23:59	Microsoft Excel Comma Separated Values File	1,102 KB
AdventureWorks_Sales_2015	09/11/2023 23:59	Microsoft Excel Comma Separated Values File	122 KB
AdventureWorks_Returns	09/11/2023 23:59	Microsoft Excel Comma Separated Values File	36 KB
AdventureWorks_Products	09/11/2023 23:59	Microsoft Excel Comma Separated Values File	58 KB
AdventureWorks_Product_Subcategories	09/11/2023 23:59	Microsoft Excel Comma Separated Values File	1 KB
AdventureWorks_Product_Categories	09/11/2023 23:59	Microsoft Excel Comma Separated Values File	1 KB
AdventureWorks_Customers	09/11/2023 23:59	Microsoft Excel Comma Separated Values File	1,981 KB
AdventureWorks_Calendar	09/11/2023 23:59	Microsoft Excel Comma Separated Values File	11 KB

Left: Product Report Data (CSV files)

DASHBOARDING WITH MS POWER BI - EVIDENCE

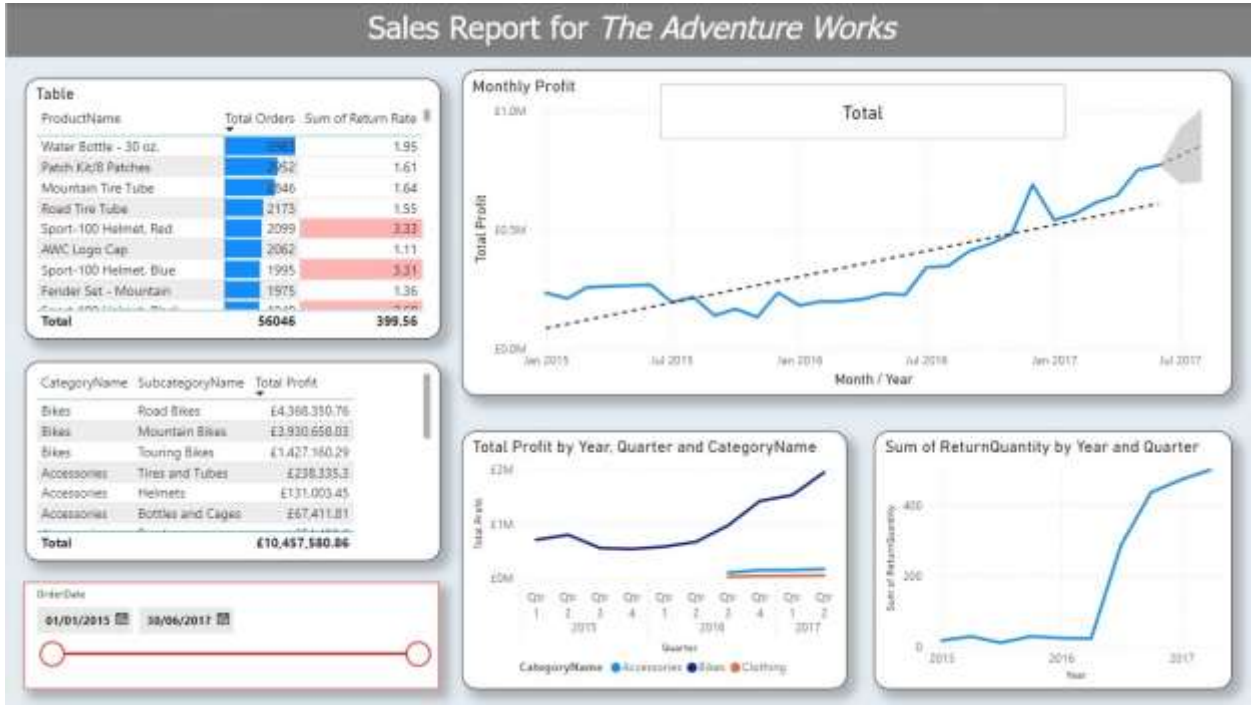
In order to create a dashboard in Power BI, with reference to the points on the previous page, the data source files were initially reviewed, and then imported into Power BI (including appending the different year's sales files into a single table). The data was transformed as necessary (e.g. changing date of birth of customers to ages at time of purchase), relationship links were created between the different datasets (shown right) and then 2 dashboard pages were created in order to display the relevant information; one for sales and one for customer information. These are shown on the following two pages.

DASHBOARDING WITH MS POWER BI - EVIDENCE

In this first dashboard, sales data is reported as monthly profit (top right), and includes a forecast for the two months ahead, based on 95% confidence (the actual forecast values are shown for each of the two months when actively viewing the chart with a mouse cursor). The table on the top left shows individual products by order and sum of return rate. Clicking on an individual product in the table changes the monthly profit chart to show the performance over time of the individual product in question.

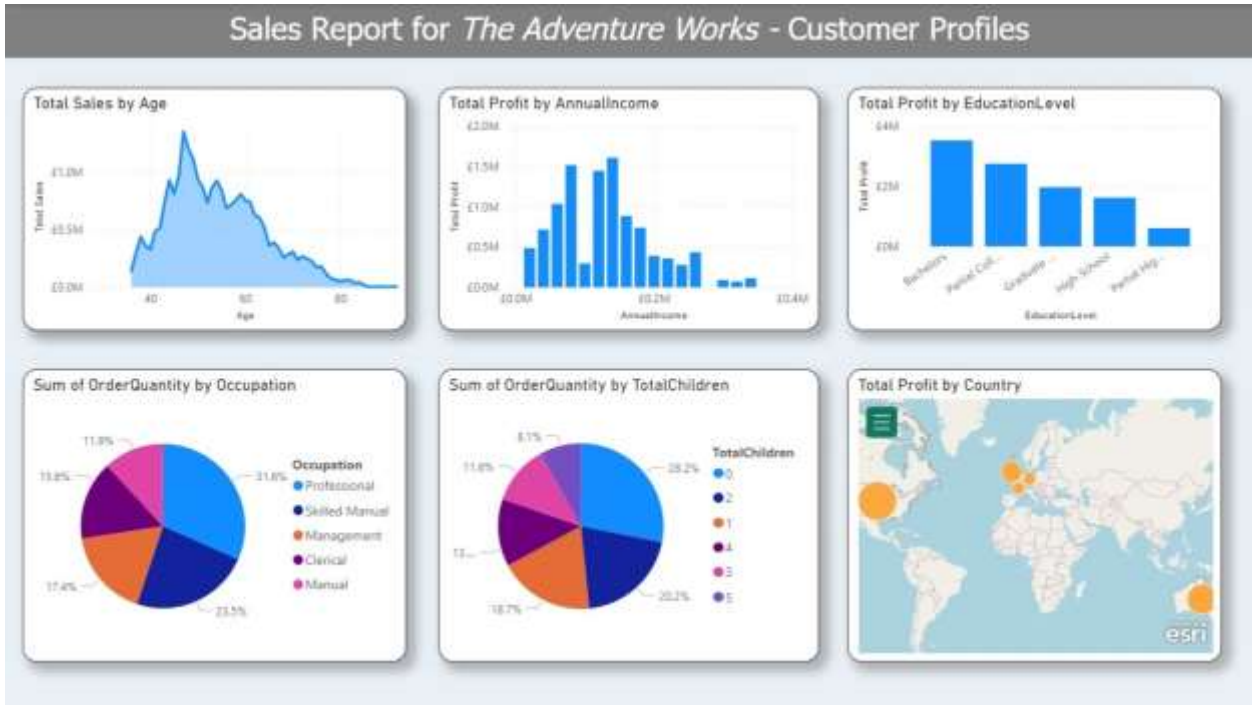
The timeline can be adjusted to focus on specified date ranges. Also shown in chart form are quarterly profits for each of three main product categories and return quantities over time. From the data, it is possible to see that from Quarters 2-3, 2016, and with the expansion of overall sales and product lines, there was a significant increase in the number of items people were returning.

Further investigation could look into whether certain products were being returned more often than others.



DASHBOARDING WITH MS POWER BI - EVIDENCE

In this second dashboard, sales information is broken down into different customer attributes.
The actual number of different ways of presenting all of the data in this exercise with respect to looking for relationships and trends between each variable can be seemingly infinite, and so the charts here are shown as examples.
Whilst viewing the data on the Power BI dashboard, it is possible to see actual values on the top three charts by interacting with the data points, although the main conclusions from the charts presented here follow on the next page.



DASHBOARDING WITH MS POWER BI - EVIDENCE

Conclusions drawn from the charts presented on the previous page:

- Total customer sales by age peaks at 47 (perhaps the age where people are still relatively active, but with higher incomes).
- People earning between £80k and 100k pa generate higher profits for the business, with an exception for people earning 100k (worth investigating why so few customers earn this amount, e.g. is there an error in the data collecting? What tax-brackets are there in other countries?
- It is clear that the more educated customers are, the more profit they generate
- Orders are most likely to come from professionals (31.6%), followed by skilled manual workers (23.5%).
- Orders are more likely to come from people with no children (28.2%), followed by those with two children (20.2%), closely followed by those with one child (18.7%).
- The areas generating the most profit (in order) are the USA, followed by Australia, and then the UK.
- Further investigations could focus on the preferred product lines with respect to demographics.

STORYTELLING WITH DATA

In this part of the *Data Essential Skills Bootcamp*, I learnt about the importance of storytelling with data, and created a data-driven story, communicating results through a basic narrative.

Data Tools Used:

- Excel (Including features and functions described in earlier sections)
- PowerPoint
- Storytelling

THE IMPORTANCE OF STORYTELLING WITH DATA

Storytelling has been an integral part of human behaviour and development ever since the times of our early ancestors. Its main purpose has always been as a means of transferring knowledge to the listener, enforcing beliefs and helping that knowledge be retained in the listener's mind. Storytelling existed long before even the written word, let alone PowerPoint and is this inherent in our nature. For decades, the traditional method for presenting data was to present charts and point out facts based on these, leaving the audience often bored and coming away having retained

THE IMPORTANCE OF STORYTELLING WITH DATA (CONTINUED)

… relatively little information. In today's world, where people are faced with an ever-increasing amount of information, there needs to be a more appealing and captivating way of presenting data to people, which is where storytelling comes in.

Storytelling with data is important because it helps to communicate insights and persuade audiences effectively. Data alone can be dry and hard to understand, but when combined with a narrative, it can become engaging and memorable; multiple cognitive psychology studies have confirmed that facts are remembered more when connected in a narrative than when they are separate. Storytelling with data can also help to establish credibility, build trust, drive a connection with the audience, encourage acceptance, and inspire action. By using techniques such as context, structure, visuals, and emotion, storytellers can craft compelling stories that showcase the value and impact of data.

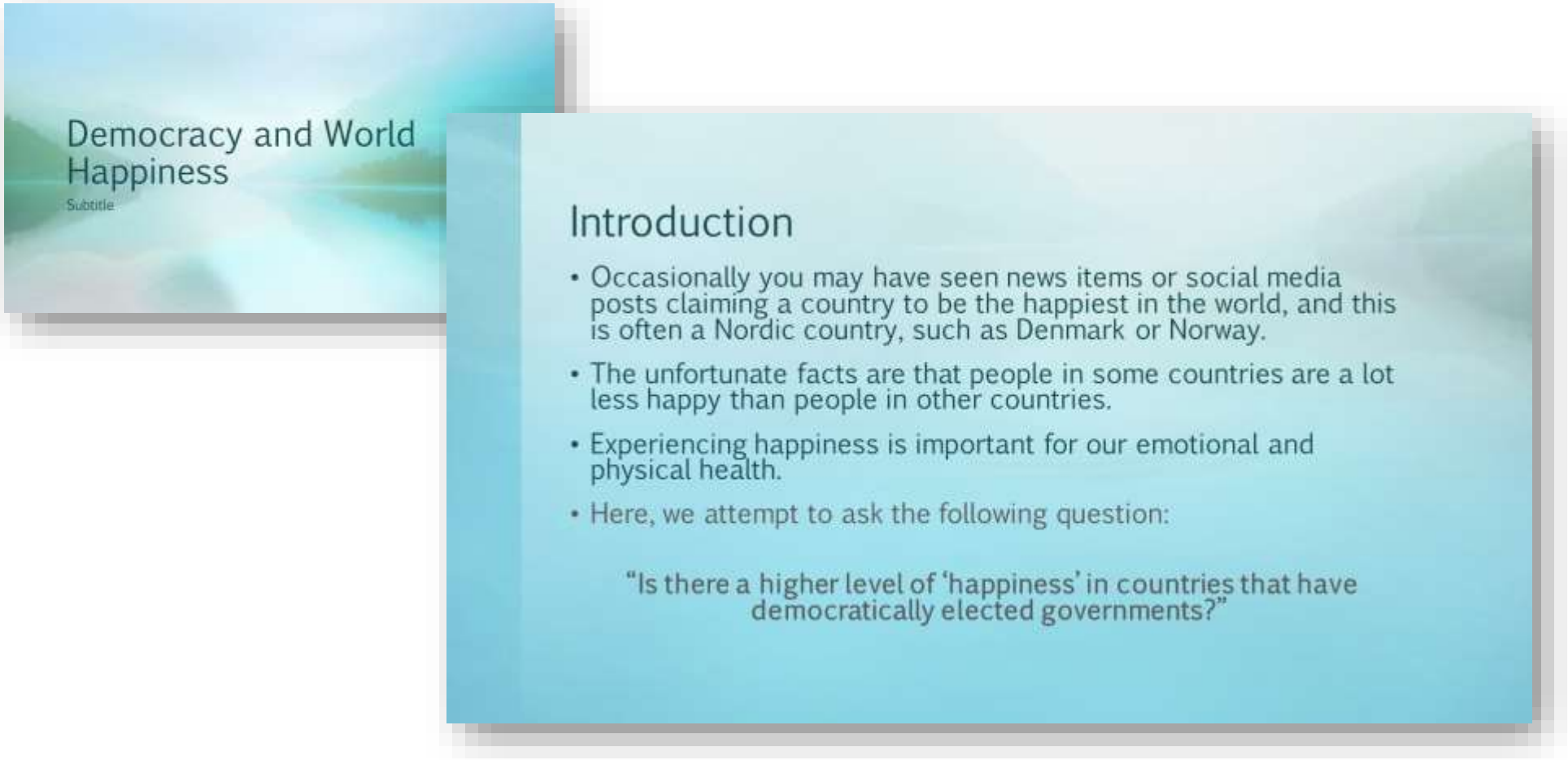
THE IMPORTANCE OF STORYTELLING WITH DATA - EVIDENCE

Building on the data I sourced and combined earlier (pages 7-13), I then went on to present this data as a narrative. The first step was to use a narrative arc content builder, shown below:

Beginning	Middle	End
Problem statement People in some countries are a lot less happy than people in other countries.	**Key plot movements** 1. Countries with a reputation for being happy (e.g. Nordic countries) are democracies 1. If democracy makes people happy, then this is a powerful argument for democracy 2. Will there be any exceptions to any findings?	**Most important message to land** In general, more democratic countries are happier. Outliers can often be explained (e.g. oil-rich Gulf states) Democracy is not linked to every factor that makes up happiness.
Context / why this is important Experiencing happiness is important for our emotional and physical health.	**Pivotal discovery / story climax resulting in a new challenge** There does seem to be a relationship between happiness of countries and democracy Index value. However, there are exceptions. Also note correlation is not necessarily causation.	**Resolution, where outcome is progress** Regime change and war aside, there are certain areas where world happiness could be improved, such as social support, health services, and anti-corruption measures.
Hypothesis Countries which are more democratic have greater levels of happiness.	**Re-framing of problem statement** Democracy is not necessarily linked to every factor that is taken into account when ranking world happiness.	**Tangible action** Increasing support to poorer nations with respect to healthcare. Addressing people's freedoms to make life choices. Increased fight against corruption.

THE IMPORTANCE OF STORYTELLING WITH DATA - EVIDENCE

In order to present my story, a brief PowerPoint presentation was prepared, and sample slides follow:

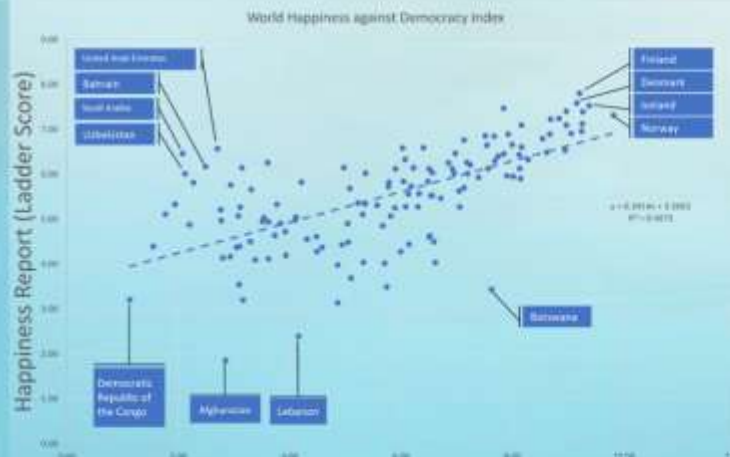


THE IMPORTANCE OF STORYTELLING WITH DATA - EVIDENCE

Different Political Systems Around the World

- In this chart, we can see how the type of political system varies around the world
- You may notice that the United States is categorised as a "Flawed Democracy". It has been lower than normal in recent years, due to the functioning of government and political culture
- But are the more democratic countries happier?

World Happiness and Democracy



- My findings agree that in general, there a higher level of 'happiness' in countries that have more democratic systems of government.
- The correlation between the two is 41%, meaning that 59% of the variability in the outcome data cannot be explained by the model.
- As predicted, countries with a reputation for being happy (e.g. Nordic countries) are more democratic.
- The less democratic countries show a wider variation in happiness.

What could be the cause of the outliers in the data, especially amongst the less democratic countries?

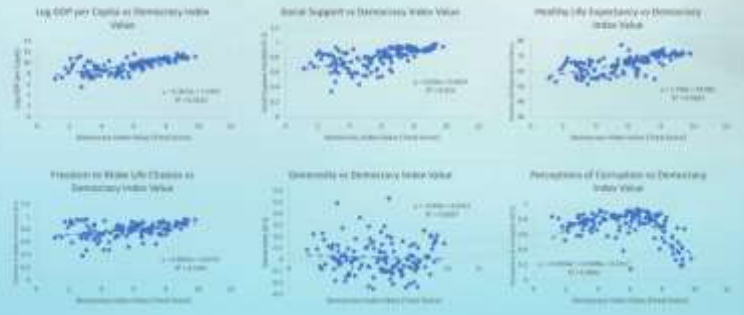
THE IMPORTANCE OF STORYTELLING WITH DATA - EVIDENCE

World Happiness and Democracy

- Possible causes of the outliers in the data, especially amongst the less democratic countries:
  - The countries that are atypically less happy face challenges, including Democratic Republic of Congo (poverty & armed conflict), Afghanistan (including women's and girls' rights and economic and humanitarian crises), and Lebanon (including mismanagement, violence, economic crisis and lack of electricity supply.
  - The countries that are atypically more happy includes a cluster of wealthy oil-rich states on the Arabian Peninsula. A surprising exception to this is Uzbekistan (factors of its relatively high happiness may well be attributable to a number of factors including having a more open and progressive government, open borders and equal rights for women).
- One other outlier of note is Botswana which has a good score for democracy, but a low score for happiness. This can be attributed to a lack of social support, curtailed healthy life expectancy and, above all, matters of income inequality and of generosity.
- Democracy is not necessarily linked to every factor that is taken into account when ranking world happiness.

Let's take just a brief look at how variations in democracy relates to different factors that make people happier:

World Happiness and Democracy – Further Breakdown of Data



- Without going into specifics, from the above data, the democracy index appears to be linked to several factors that make up happiness, including GDP, Social Support, and Life Expectancy.
- There is no obvious link between the democracy index and the generosity of people.
- Perceptions of corruption only become lower with the higher end of the democracy index values. There is also a slight shift lower amongst the much less democratic countries. In some cases, there may be less sense of corruption due to propaganda, or a lack of criticism caused by fear.

World Happiness and Democracy – Conclusions

- Speaking in general terms, the more democratic countries are happier, although this is not without exceptions (which can be attributed to factors, e.g. oil wealth, governmental policy). The data does, however, show a case in favour of democracy being a better form of government in terms of people's happiness.
- Democracy is not necessarily related to every factor that is taken into account when ranking world happiness.
- Regime change and war aside, there are certain areas where world happiness could be improved, such as social support, health services, and anti-corruption measures.
- Tangible actions that could increase happiness in less-happy nations include increasing support to poorer nations with respect to healthcare, addressing people's freedoms to make life choices, and increasing measures to fight corruption.
- It is important to remember in this study, that correlation does not necessarily mean causation.

PORTFOLIO: SUMMARY

During the 10-week bootcamp, I was able to build on my existing knowledge of Microsoft Excel and explore new ways of analysing data which I had not done before (e.g. Pivot Charts and Power Query). I learnt about the importance of cleaning data before analysing it and how the choice of visual is important in order to most effectively present data. The introductions to dashboarding and Microsoft Power BI were extremely interesting, and I felt I learnt a lot during the time available, as I did with the all-important storytelling aspect of presenting data as well – something I had not really considered before.

The course also included intensive introductions to using Python and SQL for data analysis towards the end, leaving no time to evidence them in this first version of my Portfolio.

Looking ahead, the areas I intend on focusing my learning on are Python, Excel VBA, and the fundamentals of Microsoft cloud computing platform Azure.

v. 1.0 26 Nov 2023