



Data Citizen Skills Bootcamp

Iris Mejuto

Introduction



With a background in Commerce and Marketing, my journey through different industries has been driven by a keen interest in tech and business.

This interest, along with a strong desire to understand how businesses operate and a love for data analysis, guides my career direction. I aim to combine technology, business knowledge, and continuous learning. With skills in Excel, Google Sheets, and a basic understanding of Power BI and Python, I am ready to start a career in data analysis, using my interests and skills.



How does data add value to organisations?

The importance of data in businesses is clear, acting as the main support for making smart decisions and necessary changes.

Based on my experience at Amazon, where decisions are driven by data, it's crucial to focus on key metrics that matter to the business.

With advances in technology and more online interactions, along with the growth of the Internet of Things, the trend of tracking data is increasing. This shows a growing use of data for key business decisions.



INDEX

The portfolio it is divided into 4 parts:

- **Project 1: Salesperson and Targets** 5 - 11
Tool: Excel
Relative to week 3.
- **Project 2: Heath failure data across various locations** 12 - 47
Tool: Excel
Relative to week 4, 5, 6
- **Project 3: Adventure Works** 48 - 54
Tool: Power BI
Relative to week 7.
- **Narrative Arc of Project 3: Adventure Works** 55 - 61
Relative to week 8.

Project 1 – Situation



In my first Project, I apply the skills acquired during the initial weeks of the bootcamp, focusing on analyzing and interpreting sales data in Excel.

This project tasks me with working on a data set that includes detailed information about the sales staff, their departments, and their sales goals. My goal is to deeply analyze this data set and evaluate the performance for each salesperson and the different departments.

Through this project I demonstrate the analytical techniques introduced in the bootcamp.

Project 1 – Task



In the project, I'll perform the following tasks using Excel functions such as VLOOKUP, INDEX, MATCH, IFERROR, AVERAGE, AVERAGEIF, and SUMIF:

- Calculate the "Target_Achievement_Level" for each sales agent, along with their bonus percentages.
- Determine the total bonus budget and calculate the bonus payments for each department.
- Identify the highest and lowest performers based on their sales targets.
- Compute the average performance of the sales team, considering the target achievement level.
- Analyse department-level performance by calculating total sales, average sales per agent, average target achievement level, and the total bonus allocated to each department.

Project 1 – Action - Sourcing data



Data Set:

NAME	SALES TARGET	ACTUAL	DEPARTMENT
Abbie Forbes	\$50.000,00	\$45.535,00	A
Addison Lloyd	\$65.000,00	\$74.529,00	B
Adison Collins	\$30.000,00	\$27.622,00	C
Adriana Escobar	\$30.000,00	\$27.632,00	C
Ainsley Austin	\$50.000,00	\$45.458,00	B
Alexis Chan	\$55.000,00	\$54.811,00	D
Alisha Branch	\$100.000,00	\$96.003,00	A
Alyssa Burns	\$30.000,00	\$31.374,00	D
Amani Parrish	\$65.000,00	\$70.398,00	D
Anne Parsons	\$30.000,00	\$30.472,00	C
Arianna Mayer	\$55.000,00	\$59.830,00	D
Aubree Mitchell	\$100.000,00	\$90.968,00	B

Target Ach. Lvl	Bonus Percentage	Sales Value	Sales Level	Bonus_2
-0,4	0	0 th		0
-0,02	4	30000 t1		2
-0,01	5	50000 r2		5
0,02	7	75000 e3		8
0,05	8	10000 w3		10
0,1	10			

Project 1 – Action - Evidence



Calculate: Target_Achievement_Level, the bonus percentage for each sales agent, the total budget for bonus, the total bonus payment for each department

A	B	C	D	E	F	G
NAME	SALES TARGET	ACTUAL	DEPARTMENT	Target Achievement Level	Bonus Percentage	Budget for bonus
Abbie Forbes	\$50.000,00	\$45.535,00	A	-0,09	0	-
Addison Lloyd	\$65.000,00	\$74.529,00	B	0,15	10	7.452,90
Adison Collins	\$30.000,00	\$27.622,00	C	-0,08	0	-
Adriana Escobar	\$30.000,00	\$27.632,00	C	-0,08	0	-
Ainsley Austin	\$50.000,00	\$45.458,00	B	-0,09	0	-
Alexis Chan	\$55.000,00	\$54.811,00	D	0,00	5	2.740,55
Alisha Branch	\$100.000,00	\$96.003,00	A	-0,04	0	-
Alyssa Burns	\$30.000,00	\$31.374,00	D	0,05	7	2.196,18
Amani Parrish	\$65.000,00	\$70.398,00	D	0,08	8	5.631,84
Anne Parsons	\$30.000,00	\$30.472,00	C	0,02	5	1.523,60
Arianna Mayer	\$55.000,00	\$59.830,00	D	0,09	8	4.786,40
Aubree Mitchell	\$100.000,00	\$90.968,00	B	-0,09	0	-
Axel Moreno	\$50.000,00	\$56.070,00	C	0,12	10	5.607,00
Braylon Gonzales	\$65.000,00	\$71.731,00	A	0,10	10	7.173,10
Britney Blevins	\$55.000,00	\$56.435,00	D	0,03	7	3.950,45
Caleb Woodard	\$30.000,00	\$32.700,00	A	0,09	8	2.616,00
Callie Grant	\$55.000,00	\$56.651,00	D	0,03	7	3.965,57
Callum Sherman	\$55.000,00	\$56.031,00	C	0,02	5	2.801,55
Cherish Mitchell	\$30.000,00	\$32.091,00	C	0,07	8	2.567,28
Ciara Leonard	\$100.000,00	\$92.582,00	C	-0,07	0	-
Clinton Hahn	\$55.000,00	\$51.725,00	A	-0,06	0	-
Clinton Mcknight	\$30.000,00	\$29.874,00	D	0,00	5	1.493,70
Cody Galloway	\$30.000,00	\$10.000,00	A	-0,67	0	-
Cruz Hickman	\$30.000,00	\$30.302,00	B	0,01	5	1.515,10

Target_Achievement_Level

$$=(C2-B2)/B2$$

Bonus % for each Sales Person

$$=IFERROR(VLOOKUP(E2:E101;'Adiccional Task'!A1:E7; 2; TRUE);0)$$

Total budget for bonus

$$=IFERROR(C2 * (VLOOKUP(E2; 'Adiccional Task'!A1:E7; 2; TRUE) / 100);0)$$

Project 1 – Action - Evidence



The best and worst-performing salesperson by sales targets:

Best Sales Person
Johan Mcdaniel
Worse Sales Person
Cody Galloway

```
=INDEX(A2:A100; MATCH(MAX(E2:E100); E2:E100; 0))
```

```
=INDEX(A2:A100; MATCH(MIN(E2:E100); E2:E100; 0))
```

Average performance of the overall sales team based the target achievement level value:

Average Performance
0,01

```
=AVERAGE(E2:E100)
```

- Johan Macdaniel exceeded his target by 53%.
- Cody Galloway's performance is 67% below his sales target.
- The average target overachievement is 10%.

Project 1 – Action - Evidence



Department level performance: Total Sales, Average Sales, Average Target AL and total bonus by department

A	B	C	D	E
Departament	Total Sales	Average Sales	Average Target AL	Bonus Payment
A	\$ 1.285.338	\$ 58.424	-0,01	\$ 2.660,88
B	\$ 774.088	\$ 59.545	0,01	\$ 3.070,36
C	\$ 946.971	\$ 52.610	-0,02	\$ 2.073,52
D	\$ 2.792.786	\$ 59.421	0,02	\$ 2.989,68

=SUMIF('number-democracies-autocracies'!\$D\$2:\$D\$101;"A";'number-democracies-autocracies'!\$C\$2:\$C\$101)

=AVERAGEIF('number-democracies-autocracies'!\$D\$2:\$D\$101;"A";'number-democracies-autocracies'!\$C\$2:\$C\$101)

=AVERAGEIF('number-democracies-autocracies'!\$D\$2:\$D\$101;"A";'number-democracies-autocracies'!\$E\$2:\$E\$101)

=AVERAGEIF('number-democracies-autocracies'!\$D\$2:\$D\$101;"A";'number-democracies-autocracies'!\$G\$2:\$G\$101)

Project 1 - Results



Department A:

It has the lowest bonus payout at \$2,660.88, although they don't have the best sales. It has a negative average target AL (-0.01), suggesting that on average it slightly underperformed its sales targets compared to other departments.

Department B:

It has the second highest bonus payout at \$3,070.36 and a positive average target AL (0.01) but is the worst in terms of total sales at \$774,088.

Department C:

Falls in the middle range for total sales and bonus payments and has the lowest Average Target AL at -0.02, indicating a underperformance relative to its sales targets.

Department D:

Sold \$2,598,057, over double compared to Department B, which sold the least at \$774,088.

Project 2 – Situation



In Project 2, I build on the skills I honed during weeks 4, 5, and 6 of the bootcamp, focusing on the task of combining, cleaning, and analysing data within Excel. This project involves working with two data sets, collected specifically to analyse the impact of the main indicators associated with heart disease.

My goal is to meticulously clean and analyse these data sets to investigate the correlation between various indicators and heart disease and identify the most frequently diagnosed demographics.

This task allows me to apply the analytical techniques learned in the bootcamp.

Project 2- Sourcing data



Dataset: Heart Disease Data set on UCI Repository

Age	Sex	CP	Trestbps	chol	fbs	resecg	thalach	exng	oldpeak	slope	ca	thal	num	date of sample
28	1	2	130	132	0	2	185	0	0 ?	?	?	?	0	43480
29	1	2	120	243	0	0	160	0	0 ?	?	?	?	0	43583
29	1	2	140 ?		0	0	170	0	0 ?	?	?	?	0	43574
30	0	1	170	237	0	1	170	0	0 ?		?	?	0	43678
31	0	2	100	219	0	1	150	0	0 ?	?	?	?	0	43612
32	0	2	105	198	0	0	165	0	0 ?	?	?	?	0	43440
32	1	2	110	225	0	0	184	0	0 ?	?	?	?	0	43714
32	1	2	125	254	0	0	155	0	0 ?	?	?	?	0	43705
33	1	3	120	298	0	0	185	0	0 ?	?	?	?	0	43443
34	0	2	130	161	0	0	190	0	0 ?	?	?	?	0	43601
34	1	2	150	214	0	1	168	0	0 ?	?	?	?	0	43654
34	1	2	98	220	0	0	150	0	0 ?	?	?	?	0	43657
35	0	1	120	160	0	1	185	0	0 ?	?	?	?	0	43453
35	0	4	140	167	0	0	150	0	0 ?	?	?	?	0	43608
35	1	2	120	308	0	2	180	0	0 ?	?	?	?	0	43623
35	1	2	150	264	0	0	168	0	0 ?	?	?	?	0	43571
36	1	2	120	166	0	0	180	0	0 ?	?	?	?	0	43596
36	1	3	112	340	0	0	184	0	1 2	?	3	?	0	43526
36	1	3	130	209	0	0	178	0	0 ?	?	?	?	0	43721
36	1	3	150	160	0	0	172	0	0 ?	?	?	?	0	43538
37	0	2	120	260	0	0	130	0	0 ?	?	?	?	0	43725
37	0	3	130	211	0	0	142	0	0 ?	?	?	?	0	43723
37	0	4	130	173	0	1	184	0	0 ?	?	?	?	0	43708

Project 2– Data Set Presentation



Dataset: Heart Disease Data set on UCI Repository

- Age (age)
- Sex (sex)
- Chest Pain (cp)
- Resting blood pressure (trestbps): in mm Hg on admission to the hospital)
- Serum cholestoral in mg/dl (chol)
- Fasting blood sugar > 120 mg/dl (fbs): 1 = true; 0 = false
- Resting electrocardiographic results (restecg)
- Maximum heart rate achieved (thalach)
- Exercise induced angina (exang): 1 = true; 0 = false

Project 2 - Data Set Presentation



Dataset: Heart Disease Data set on UCI Repository

- ST depression induced by exercise relative to rest (oldpeak)
- The slope of the peak exercise ST segment (slope)
- Number of major vessels (0-3) colored by flourosopy (ca)
- Thal (thal) 3 = normal; 6 = fixed defect; 7 = reversable defect
- Diagnosis of heart disease (num)
 - 0: No
 - 1, 2, 3, 4: Yes
- Date of sample (date of sample)

Project 2 - Task 1



To cleaning and process this Data Set I had to clean the data, to do so I have done the following steps:

- Delete empty columns
- Format date to date column
- Replace the value "?" and replace it with empty field
- Delete rows that contained null data in fields important to the analysis
- Format each field with the appropriate type and format the datasets as a table
- Apply functions such as AVERAGE, MIN, MAX, and COUNTIF to identify trends within the data set.
- Plot histogram of age
- Draw conclusions from the analysis

Project 2 - Evidence - Task 1 - Average age of the patients



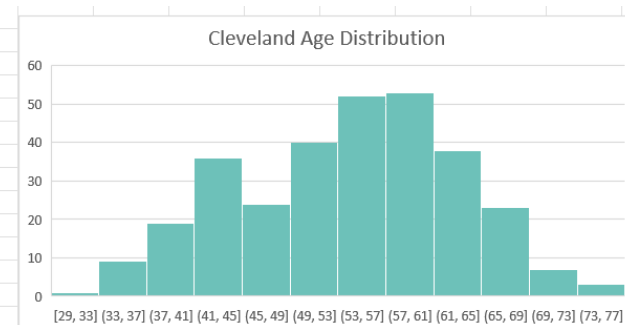
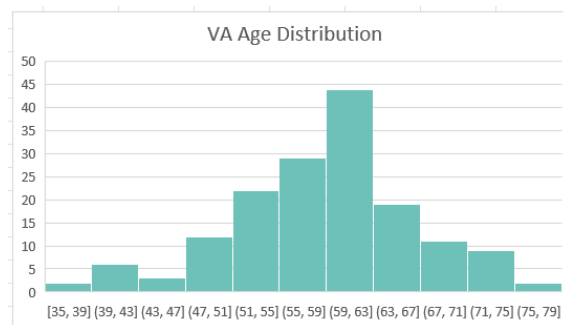
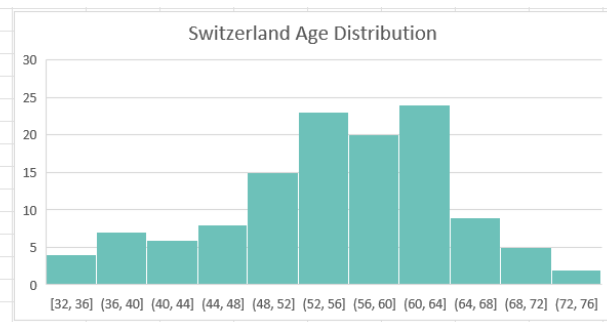
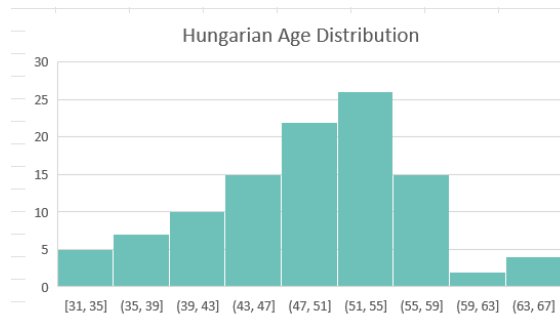
- Average age of the patients:

Worksheets	Age Average	
Hungarian	48	=AVERAGE('processed hungarian'!A2:A296)
Switzerland	55	=AVERAGE('processed switzerland'!A2:A124)
VA	59	=AVERAGE('processed va'!A2:A160)
Cleveland	54	=AVERAGE('processed cleveland'!A2:A306)

Project 2 - Evidence - Task 1 - Plot Histograms



- Histograms of age variable :



Project 2 - Evidence - Task 1 - Draw conclusions - Histograms



- **Hungarian:**

If the histogram shows a bell-shaped curve, it suggests a normal age distribution among Hungarian participants. If there is a right or left bias, it indicates that the study has a higher concentration of younger or older participants, respectively.

- **Switzerland:**

A histogram with multiple peaks could indicate that there are several different age groups that participated in the Swiss study. If the distribution has a long tail on one side, it may suggest a broader range of ages with a gradual decline.

Project 2 - Evidence - Task 1 - Draw conclusions - Histograms



- **VA:**

This histogram has a narrow shape, it would suggest that the majority of VA study participants are clustered within a smaller age range, indicating less variability in participant ages.

- **Cleveland:**

A histogram with a flat distribution would mean that the ages of the Cleveland study participants are fairly evenly distributed across the range, with no distinct concentration in any particular age group.

Project 2 - Evidence - Task 1 – Month the most samples



- Which month the most samples were received:

Month More Samples		August	November	November	April
Month	Nu. Month	Hungarian	Switzerland	VA	Cleveland
January	1	25	9	11	20
February	2	16	13	4	26
March	3	26	8	7	28
April	4	26	13	12	30
May	5	31	8	14	21
June	6	14	8	11	24
July	7	23	13	11	26
August	8	32	7	8	20
September	9	28	10	13	21
October	10	21	12	13	23
November	11	28	15	44	30
December	12	25	7	12	26

Hungarian: **Switzerland:**

August

November

VA:

Cleveland:

November

April

Conditional Formatting Rules Manager				
show formatting rules for: This Table				
<div> <div>New Rule...</div> <div>Edit Rule...</div> <div>Delete Rule</div> <div>Duplicate Rule</div> <div>↑</div> <div>↓</div> </div>				
Rule (applied in order shown)	Format	Applies to	Stop If True	
Formula: =I2=MAX(\$I\$2:\$I\$13)	AaBbCcYyZz	=I\$2:I\$14	↑	<input type="checkbox"/>
Formula: =H2=MAX(\$H\$2:\$H\$13)	AaBbCcYyZz	=H\$2:H\$14	↑	<input type="checkbox"/>
Formula: =G2=MAX(\$G\$2:\$G\$13)	AaBbCcYyZz	=G\$2:G\$14	↑	<input type="checkbox"/>
Formula: =F2=MAX(\$F\$2:\$F\$13)	AaBbCcYyZz	=F\$2:F\$14	↑	<input type="checkbox"/>

=COUNTIF('processed hungarian'!\$Q\$2:\$Q\$296;\$E\$3)

Project 2 - Situation - Task 2



I have conducted an analysis to investigate the relationship between blood pressure, cholesterol levels and blood sugar with the diagnosis of heart disease.

The goal is to determine whether elevated levels of these indicators were frequently observed in patients with heart disease in different regions.

For this task I have used dynamic tables to organize the information and I have calculated the average of the main indicators:

Blood pressure (Trestbps), Blood sugar (fbs) and Cholesterol (CHOL)

Project 2 - Evidence - Task 2 - Pivot Tables



Hungarian				
		Data		
num ▼↑	fbs ▼↑	AV Trestbps	AV Chol	COUNTA
0	0	130	238	157
	1	134	225	7
1	0	135	263	86
	1	141	299	13

Switzerland				
		Data		
num ▼↑	fbs ▼↑	AV Trestbps	AV Chol	COUNTA
0	0	120	0	1
	1	136	0	41
1	0	141	0	5
	1			

VA				
		Data		
num ▼↑	fbs ▼↑	AV Trestbps	AV Chol	COUNTA
0	0	129	130	21
	1	131	240	10
1	0	133	180	66
	1	143	195	35

Cleveland			
		Data	
num ▼↑	fbs ▼↑	AV Trestbps	COUNTA
0	0	129	142
	1	135	23
1	0	133	117
	1	144	22

Project 2 - Results - Task 2 - Results



The data show a trend toward higher blood pressure among patients with heart disease in all settings studied.

- Cholesterol levels also point to a possible correlation, but inconsistencies between regions and an imbalance in sample sizes between patients with and without heart disease prevent a definitive conclusion.
- Blood sugar levels did not demonstrate a clear relationship due to the same data limitations.
- Hypertension is identified as a major risk factor for heart disease.
- High cholesterol and blood sugar levels are potential factors, but the lack of balanced samples requires caution when interpreting their impact.

Project 2 – Situation - Task 2 - Relationship with heart disease

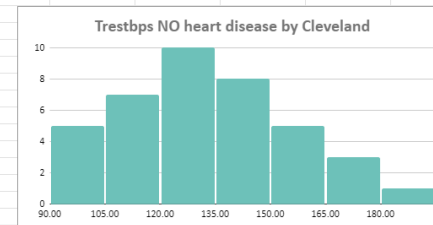
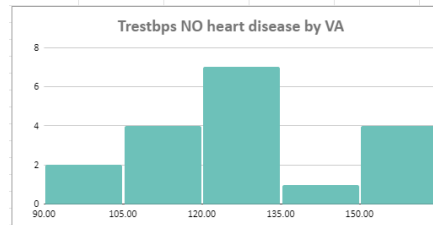
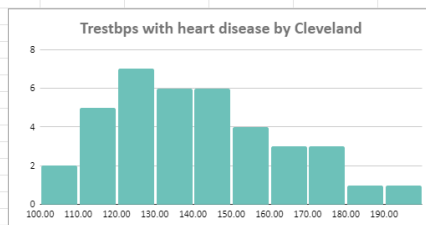
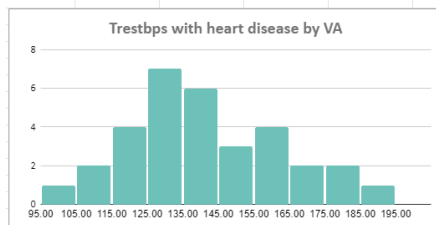
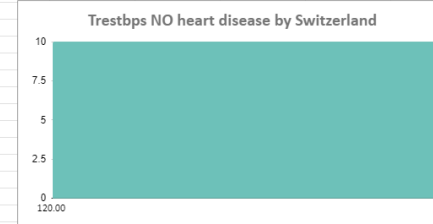
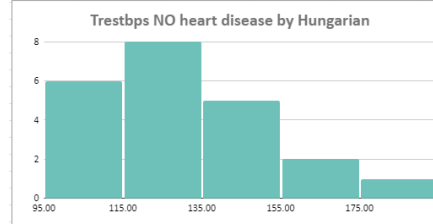
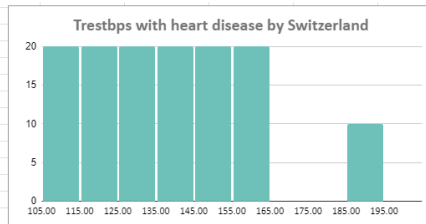
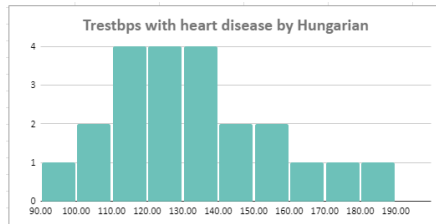


In this analysis, I'm focusing on exploring the relationship between heart disease presence and three key health indicators: blood pressure, cholesterol, and blood sugar.

By using histograms, I'll examine how these indicators are distributed within our study population and how they relate to heart conditions.

Recognized as significant risk factors, these indicators' patterns could provide insights into diagnosing, preventing, and managing heart conditions through a clear and simple visual approach.

Project 2 - Evidence - Task 2 - Relationship with heart disease - Trestbps

**BLOOD PRESSURE (Trestbps)**

Project 2 - Results - Task 2 - Relationship with heart disease - Trestbps

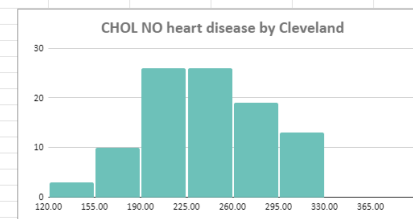
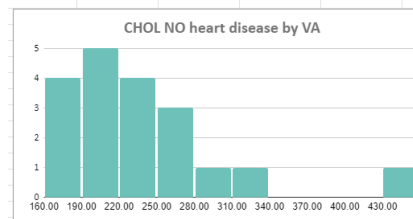
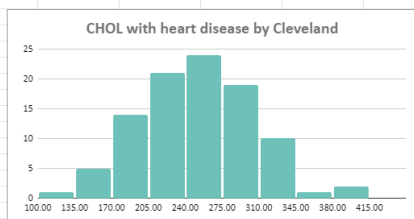
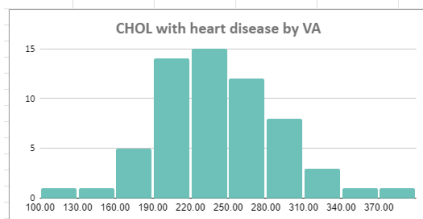
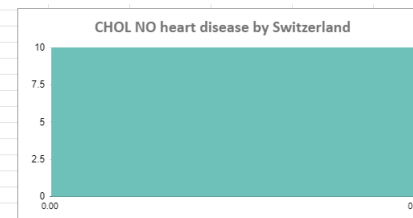
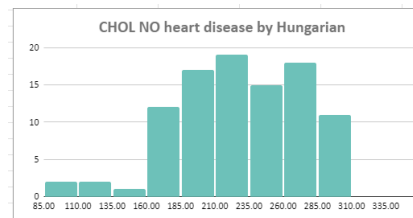
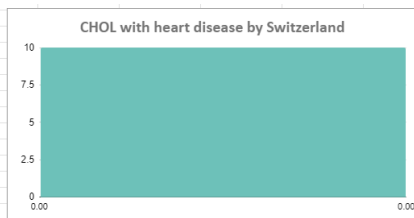
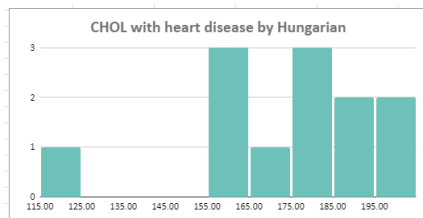


- **Individuals with Heart Disease:**
 - Hungary: Most values are in the 120-140 mmHg range.
 - Switzerland: Data is consistent, with many values in the 115-135 mmHg range.
 - VA: Values spread from 110 mmHg to over 170 mmHg, with peaks in the 125-145 mmHg range.
 - Cleveland: Broad distribution, with values mainly between 110 mmHg and 170 mmHg.
- **Individuals without Heart Disease:**
 - Hungary: Lower values predominate, mainly in the 115-135 mmHg range.
 - Switzerland: Insufficient data to determine the distribution.
 - VA: Values are mostly in the lower range, 90-120 mmHg.
 - Cleveland: Values are spread out, with a significant number in the 150-170 mmHg range.
- This analysis suggests that higher values of "Trestbps" are associated with heart disease. Lower "Trestbps" values are more common in people without heart disease. However, variability and data gaps should be noted, especially in the case of Switzerland.

Project 2 - Evidence - Task 2 - Relationship with heart disease - CHOL



CHOLESTEROL(CHOL)



Project 2 - Results - Task 2 - Relationship with heart disease - CHOL



- **Individuals with Heart Disease:**
 - Hungary: CHOL levels are mainly in the range of 150-175 mg/dL.
 - VA: Broader distribution with levels from 190 mg/dL to over 300 mg/dL.
 - Cleveland: Most common CHOL levels are between 200 and 250 mg/dL.
- **Individuals without Heart Disease:**
 - Hungary: CHOL levels mostly range from 150 to 300 mg/dL, with a concentration around 200-250 mg/dL.
 - Switzerland: No visible data, suggesting a lack of samples or data collection issues.
 - VA: Lower CHOL levels predominate, primarily between 160 mg/dL and 240 mg/dL.
 - Cleveland: CHOL levels vary widely, with a significant number of individuals having levels around 200-260 mg/dL.
- These observations suggest that there is a wide range of cholesterol levels among people with heart disease, which may reflect the complex nature of the disease and the influence of other factors such as diet and genetics. The absence of data for Switzerland on people without heart disease limits the ability to compare between

Project 2 – Situation – Task 3 - Compare the Data Sets - Sex



Relationship between the sex of patient and the existence of heart disease :

To analyse the correlation between patient gender and heart disease, I used pivot tables to categorize and filter the data by gender and heart disease status.

For each gender, the proportion of patients with and without heart disease in various age ranges was calculated in relation to the total number of samples.

Additionally, the total distribution of the sample by gender was determined, which facilitated an analysis that takes into account the sample size of each gender category.

Project 2 – Evidence – Task 3 - Compare the Data Sets - Sex



Relationship between the sex of patient and the existence of heart disease :

- On the right, the table details the distribution by sex correlating to diagnosed heart disease presence for Hungarian and Switzerland.
- On the left, it shows the total sample counts by sex for both Hungarian and Switzerland.

Hungarian				Switzerland		
Num	Count	%		Num	Count	%
0	189			0	8	
Female	69	85%		Male	8	7%
Male	120	56%		1	115	
1	106			Female	10	100%
Female	12	15%		Male	105	93%
Male	94	44%		Grand Total	123	
Grand Total	295					

Hungarian				Switzerland		
Sex	Count	%		Sex	Count	%
Female	81	27%		Female	10	8%
Male	214	73%		Male	113	92%
Grand Total	295			Grand Total	123	

Project 2 – Evidence – Task 3 - Compare the Data Sets - Sex



Relationship between the sex of patient and the existence of heart disease :

- On the right, the table details the distribution by sex correlating to diagnosed heart disease presence for VA and Cleveland.
- On the left, it shows the total sample counts by sex for both VA and Cleveland.

VA			Cleveland		
Num	Count	%	Num	Count	%
0	41		0	166	
Female	2	50%	Female	73	74%
Male	39	25%	Male	93	45%
1	118		1	139	
Female	2	50%	Female	25	26%
Male	116	75%	Male	114	55%
Grand Total	159		Grand Total	305	

VA			Cleveland		
Sex	Count	%	Sex	Count	%
Female	4	3%	Female	98	32%
Male	155	97%	Male	207	68%
Grand Total	159		Grand Total	305	

Project 2 – Results – Task 3 - Compare the Data Sets - Sex



Relationship between the sex of patient and the existence of heart disease :

- In all regions, men have a higher incidence of heart disease than women, which could suggest that males may have a higher risk of certain heart diseases.
- But it should be noted that there are a greater number of samples of men than of women, and the total number of samples is small to draw conclusions.

Project 2 – Situation – Task 3 - Compare the Data Sets - Age



Relationship between the age of patient and the existence of heart disease :

To analyze the relationship between age and the prevalence of heart disease, pivot tables were used to sort and filter the data by age groups and the presence of heart disease.

The percentage of individuals with and without heart disease was calculated for each age range against the total number of samples, revealing prevalence patterns across different demographics.

Furthermore, to assess the representativeness of these findings, the total sample distribution by age group was determined, allowing for a weighted analysis that considers the size of each group.

These methodical steps were crucial to ensure the integrity and accuracy of the final analysis.

Project 2 – Evidence – Task 3 - Compare the Data Sets - Age



Relationship between the age of patient and the existence of heart disease :

- On the right, the table details the distribution by age ranges correlating to diagnosed heart disease presence for Hungarian and Switzerland.
- On the left, it shows the total sample counts by age range for both Hungarian and Switzerland.

Hungarian			Switzerland		
Num	Count	%	Num	Count	%
0	189	64%	0	8	7%
29-39	40	77%	29-39	1	10%
40-49	70	64%	40-49	1	7%
50-59	74	61%	50-59	4	8%
60-69	5	45%	60-69	1	2%
1	106	36%	70-79	1	20%
29-39	12	23%	1	115	93%
40-49	40	36%	29-39	9	90%
50-59	48	39%	40-49	14	93%
60-69	6	55%	50-59	48	92%
Grand Total	295		60-69	40	98%
			70-79	4	80%
			Grand Total	123	

Hungarian			Switzerland		
Age range	Samples	%	Age range	Samples	%
29-39	52	18%	29-39	10	8%
40-49	110	37%	40-49	15	12%
50-59	122	41%	50-59	52	42%
60-69	11	4%	60-69	41	33%
Grand Total	295		70-79	5	4%
			Grand Total	123	

Project 2 – Evidence – Task 3 - Compare the Data Sets - Age



Relationship between the age of patient and the existence of heart disease :

- On the right, the table details the distribution by age ranges correlating to diagnosed heart disease presence for VA and Cleveland.
- On the left, it shows the total sample counts by age range for both VA and Cleveland.

VA			Cleveland		
Num	Count	%	Num	Count	%
0	41	26%	0	166	54%
29-39	1	50%	29-39	12	75%
40-49	7	47%	40-49	51	70%
50-59	14	24%	50-59	65	52%
60-69	17	24%	60-69	32	40%
70-79	2	15%	70-79	6	60%
1	118	74%	1	139	46%
29-39	1	50%	29-39	4	25%
40-49	8	53%	40-49	22	30%
50-59	44	76%	50-59	60	48%
60-69	54	76%	60-69	49	60%
70-79	11	85%	70-79	4	40%
Grand Total	159		Grand Total	305	

VA			Cleveland		
Age range	Samples	%	Age range	Samples	%
29-39	2	1%	29-39	16	5%
40-49	15	9%	40-49	73	24%
50-59	58	36%	50-59	125	41%
60-69	71	45%	60-69	81	27%
70-79	13	8%	70-79	10	3%
Grand Total	159		Grand Total	305	

Project 2 – Results – Task 3 - Compare the Data Sets - Age



Relationship between the sex of patient and the existence of heart disease :

- Analysis of the data suggests an increasing trend of suffering from heart disease with age, particularly beyond the threshold of 50 to 59 years.
- However, it is worth noting that the sample size is not evenly distributed across age groups, with a larger proportion of the data coming from the 50-59 age range.
- This disproportionate sample size requires cautious interpretation of the data.

Project 2 – Task 4 - Sourcing data



Data Set: Heart failure clinical records data set

age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT	Date of event
75	0	582	0	20	1	265000	1,9	130	1	0	4	1	43419
55	0	7861	0	38	0	263358,03	1,1	136	1	0	6	1	43417
65	0	146	0	20	0	162000	1,3	129	1	1	7	1	43420
50	1	111	0	20	0	210000	1,9	137	1	0	7	1	43416
65	1	160	1	20	0	327000	2,7	116	0	0	8	1	43419
90	1	47	0	40	1	204000	2,1	132	1	1	8	1	43417
75	1	246	0	15	0	127000	1,2	137	1	0	10	1	43419
60	1	315	1	60	0	454000	1,1	131	1	1	10	1	43419
65	0	157	0	65	0	263358,03	1,5	138	0	0	10	1	43418
80	1	123	0	35	1	388000	9,4	133	1	1	10	1	43416
75	1	81	0	38	1	368000	4	131	1	1	10	1	43416
62	0	231	0	25	1	253000	0,9	140	1	1	10	1	43416
45	1	981	0	30	0	136000	1,1	137	1	0	11	1	43419
50	1	168	0	38	1	276000	1,1	137	1	0	11	1	43420
49	1	80	0	30	1	427000	1	138	0	0	12	0	43418
82	1	379	0	50	0	47000	1,3	136	1	0	13	1	43417
87	1	149	0	38	0	262000	0,9	140	1	0	14	1	43417
45	0	582	0	14	0	166000	0,8	127	1	0	14	1	43416
70	1	125	0	25	1	237000	1	140	0	0	15	1	43416
48	1	582	1	55	0	87000	1,9	121	0	0	15	1	43417
65	1	52	0	25	1	276000	1,3	137	0	0	16	0	43416
65	1	128	1	30	1	297000	1,6	136	0	0	20	1	43418
68	1	220	0	35	1	289000	0,9	140	1	1	20	1	43419
53	0	63	1	60	0	368000	0,8	135	1	0	22	0	43416

Project 2 – Task 4 - Sourcing data



The data Set contains 14 attributes:

- age: age of the patient (years)
- anaemia: Presence of anaemia (0 = No, 1 = Yes)
- creatinine_phosphokinase: Level of creatine phosphokinase in the blood (mcg/L).
- diabetes: Whether the patient has diabetes (0 = No, 1 = Yes).
- ejection_fraction: Percentage of blood leaving the heart at each contraction.
- high_blood_pressure: Whether the patient has high blood pressure (0 = No, 1 = Yes).
- platelets: Number of platelets in the blood (kilo-platelets/mL).

Project 2 – Task 4 - Sourcing data



- serum_creatinine: Level of serum creatinine in the blood (mg/dL).
- serum_sodium: Level of serum sodium in the blood (mEq/L).
- sex: Sex of the patient (1 = Male, 0 = Female).
- smoking: Whether the patient smokes (0 = No, 1 = Yes).
- time: Follow-up time (days).
- DEATH_EVENT: Whether a death event occurred during the follow-up period (0 = No, 1 = Yes).
- Date of event: Date of the event (likely related to the follow-up or death event), in Excel format.

Project 2 – Evidence - Task 4 - Related Fields in Meaning Across Datasets



Related Fields in Meaning Across Datasets: The fields that coincide in their meaning between the two datasets are:

- Age/age: Age of the patient.
- Sex/sex: Sex of the patient.
- fbs/diabetes: Fasting blood sugar and presence of diabetes, respectively. Though not identical, both relate to diabetes.
- Trestbps/high_blood_pressure: Resting blood pressure and presence of high blood pressure, respectively. While one is a specific value and the other a binary indicator, both are related to high blood pressure.

Project 2 – Evidence - Task 4 - Pre-processing for Synergy



Pre-processing for Synergy:

To utilize the "Heart failure clinical records Data Set" in synergy with the "Heart Disease dataset," you can perform the following pre-processing steps:

- Handling missing values: identify, impute and exclude missing values. For example, replace '?' for an empty cell
- Type all columns with the appropriate type: give date format to the date column, number format when necessary
- Replace binary values with "Yes" or "No" for readability
- Create intervals for the age field like I created in the other data set

Project 2 – Evidence - Task 4



- Average Age and Age Distribution:

Age Average	61
Min Age	40
Max Age	95

=AVERAGE('Cleaning Data'!A2:A300)

=MIN('Cleaning Data'!A2:A300)

=MAX('Cleaning Data'!A2:A300)

- Percentage of Patients with Anemia:

% of Patients with Anemia	43%
---------------------------	-----

=COUNTIF('Cleaning Data'!C2:C300;"Yes")/COUNTA('Cleaning Data'!C2:C300)

- Average Ejection Fraction and Distribution:

Average Ejection Fraction	38
Min	14
Max	80

=AVERAGE('Cleaning Data'!F2:F300)

=MIN('Cleaning Data'!F2:F300)

=MAX('Cleaning Data'!F2:F301)

Project 2 – Evidence - Task 4



- Percentage of Patients with High Blood Pressure:

% of Patients with High Blood Pressure	35%	→	=COUNTIF('Cleaning Data'!G2:G300;"Yes")/COUNTA('Cleaning Data'!G2:G300)
--	-----	---	---

- Average Serum Creatinine Level:

Serum Creatinine Level	1,39	→	=AVERAGE('Cleaning Data'!I2:I300)
------------------------	------	---	-----------------------------------

- Percentage of Death Events (DEATH_EVENT):

Percentage of Death Events	31%	→	=COUNTIF('Cleaning Data'!N6:N304;"Yes")/COUNTA('Cleaning Data'!N6:N304)
----------------------------	-----	---	---

Project 2 – Results - Task 4

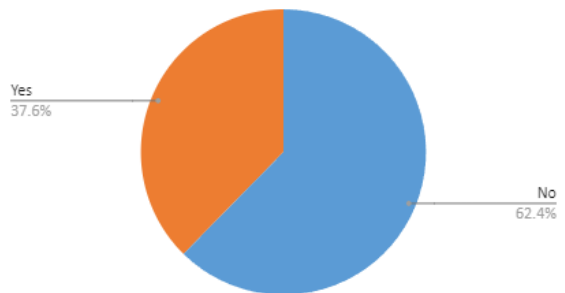


- The average age of the patients is approximately 61 years, with a distribution from 40 to 95 years.
- About 43% of the patients have anemia.
- The average ejection fraction, a key indicator of heart function, is 38%, with values ranging from 14% to 80%.
- Around 35% of the patients have high blood pressure.
- The average serum creatinine level in the blood, important for assessing kidney function, is approximately 1.39 mg/dL.
- The death event (DEATH_EVENT) occurred in 31% of the patients during the follow-up period.

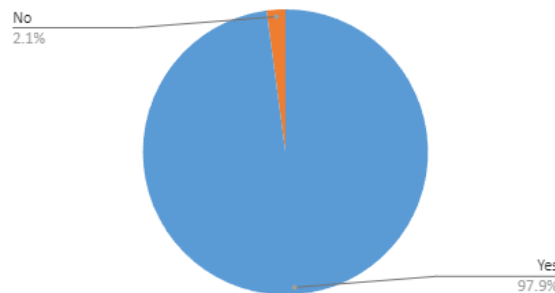
Project 2 – Pie Charts



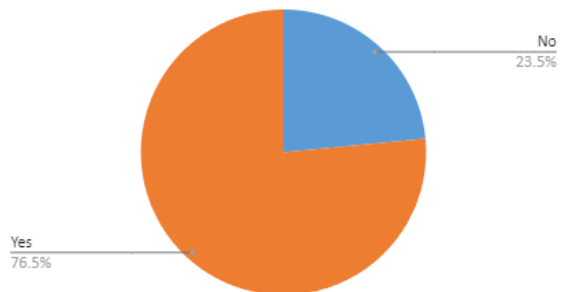
Hungarian Diagnosed heart disease



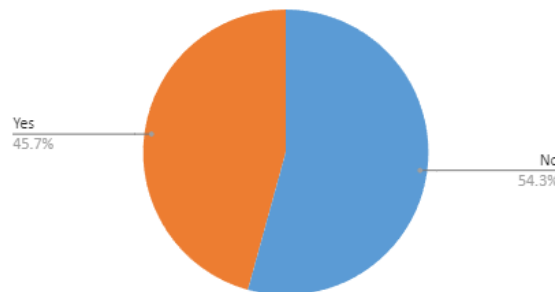
Switzerland Diagnosed heart disease



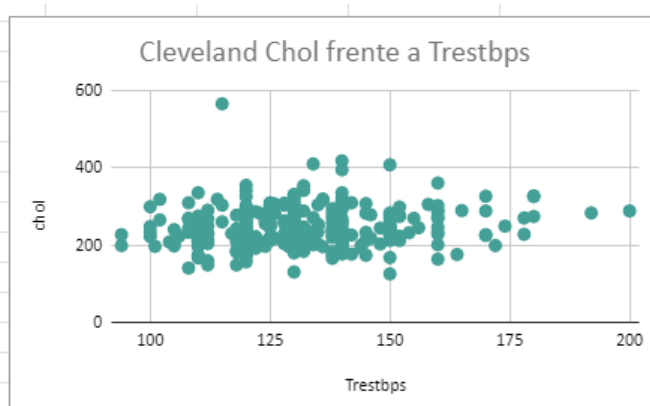
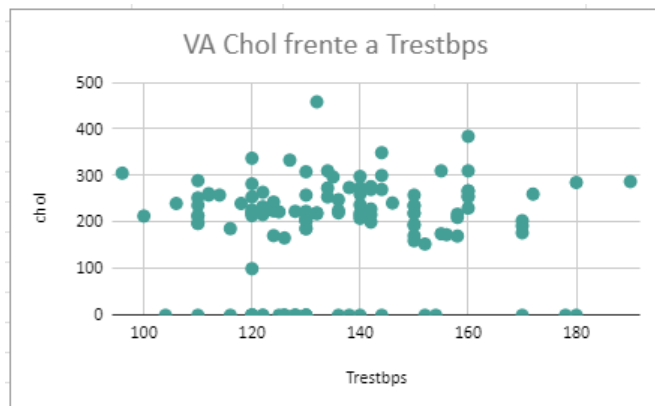
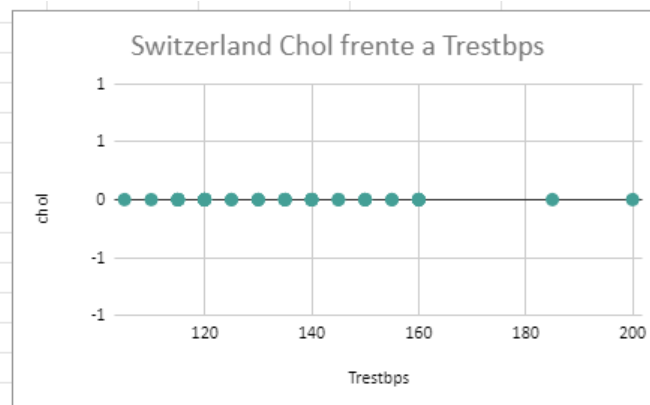
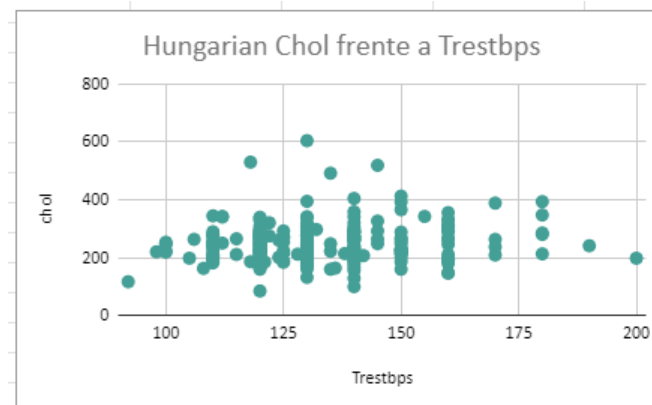
VA Diagnosed heart disease



Cleveland Diagnosed heart disease



Project 2 – Scatter Charts



Project 3 - Adventure Work - Situation



This project was an exercise in applying the skills I learned during my bootcamp, focused on Power BI for data cleaning, modeling, exploration and visualization. I examined the sales development of Adventure Works from 2015 to 2017, shedding light on market dynamics in the United States, Canada, France, Germany, Australia and the United Kingdom. The goal of the project was to look at sales trends across different continents, understand customer preferences and behaviours, and determine which products had the highest rates of profitability and returns.

Project 3 - Adventure Work - Action













Visualizations and Functions Used:

- I created advanced DAX measures to calculate key performance indicators (KPIs) such as total sales, profitability, and comparisons against targets.
- I implemented drill-down functionality to reveal details by product category, allowing for deeper understanding of each segment.
- I applied hierarchies to territory fields to facilitate analysis by geographic location.
- I used conditional formatting to enhance the visual interpretation of data and highlight areas of interest.
- I designed KPI cards that reflect the company's real-time achievement of objectives.

Project 3 - Adventure Work - Sourcing Data

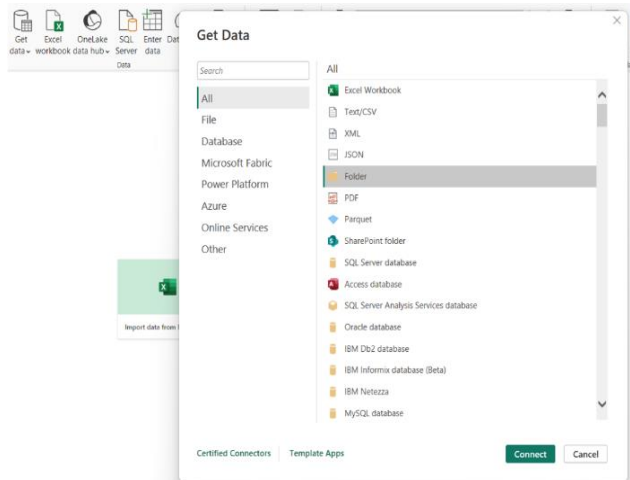


Files:

-  AdventureWorks_Calendar
-  AdventureWorks_Customers
-  AdventureWorks_Product_Categories
-  AdventureWorks_Product_Subcategories
-  AdventureWorks_Products
-  AdventureWorks>Returns
-  AdventureWorks_Sales_2015
-  AdventureWorks_Sales_2016
-  AdventureWorks_Sales_2017
-  AdventureWorks_Territories

Load data in Power BI:

- Folder connector to upload csv files



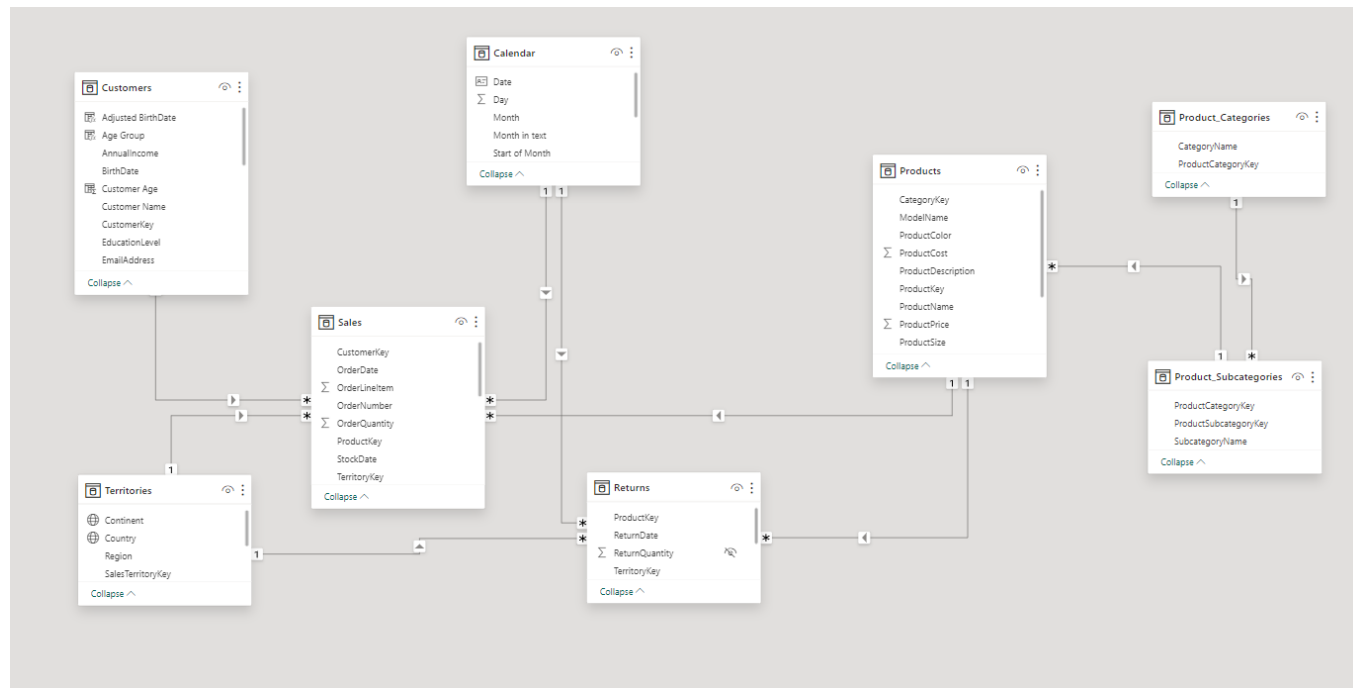
Content	Name	Extension	Date accessed	Date modified	Date created	Attributes	
Binary	AdventureWorks_Calendar.csv	.csv	24/02/2024 10:23:31	24/02/2024 10:23:31	26/01/2022 22:55:40	Record	X\jr
Binary	AdventureWorks_Customers.csv	.csv	24/02/2024 10:23:31	24/02/2024 10:23:31	26/01/2022 22:44:48	Record	X\jr
Binary	AdventureWorks_Products.csv	.csv	24/02/2024 10:23:31	24/02/2024 10:23:31	26/01/2022 22:49:12	Record	X\jr
Binary	AdventureWorks_Product_Categories.csv	.csv	24/02/2024 10:23:31	24/02/2024 10:23:31	22/03/2018 17:53:40	Record	X\jr
Binary	AdventureWorks_Product_Subcategories.csv	.csv	24/02/2024 10:23:31	22/03/2018 17:53:06		Record	X\jr
Binary	AdventureWorks>Returns.csv	.csv	24/02/2024 10:23:31	24/02/2024 10:23:31	26/01/2022 12:04:22	Record	X\jr
Binary	AdventureWorks_Sales_2015.csv	.csv	24/02/2024 10:23:31	24/02/2024 10:23:31	26/01/2022 22:57:46	Record	X\jr
Binary	AdventureWorks_Sales_2016.csv	.csv	24/02/2024 10:23:31	24/02/2024 10:23:31	26/01/2022 23:00:02	Record	X\jr
Binary	AdventureWorks_Sales_2017.csv	.csv	24/02/2024 10:23:31	24/02/2024 10:23:31	26/01/2022 23:04:16	Record	X\jr
Binary	AdventureWorks_Territories.csv	.csv	24/02/2024 10:23:31	22/03/2018 17:51:12		Record	X\jr
Binary	AdventureWorks_Sales_2015.csv	.csv	24/02/2024 10:23:31	24/02/2024 10:23:31	26/01/2022 22:57:46	Record	X\jr
Binary	AdventureWorks_Sales_2016.csv	.csv	24/02/2024 10:23:31	24/02/2024 10:23:31	26/01/2022 23:00:02	Record	X\jr
Binary	AdventureWorks_Sales_2017.csv	.csv	24/02/2024 10:23:31	24/02/2024 10:23:31	26/01/2022 23:04:16	Record	X\jr

Combine Load Transform Data Cancel

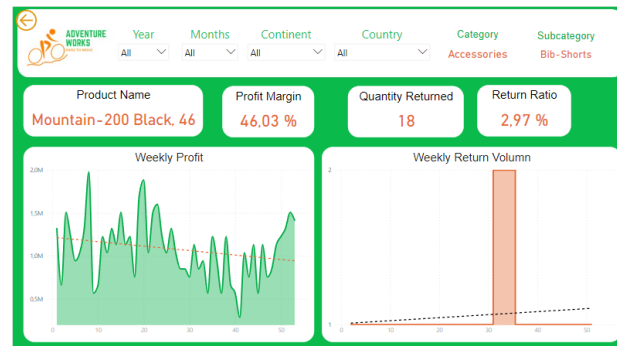
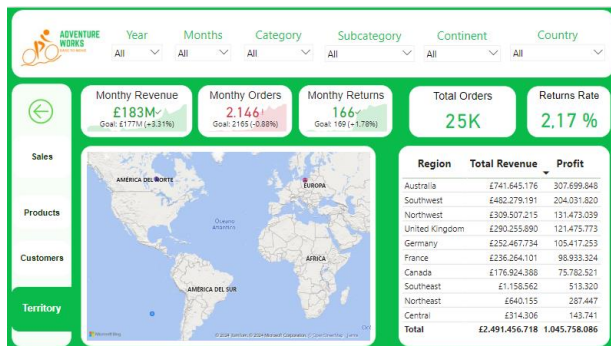
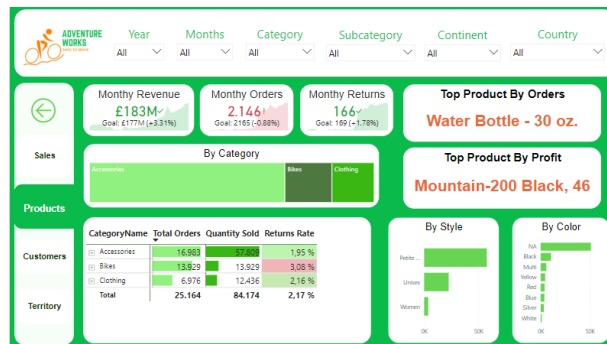
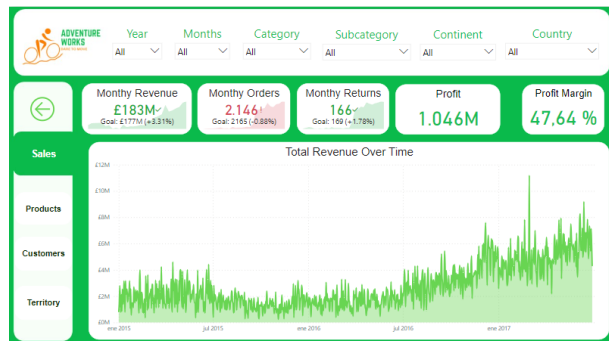
Project 3 - Adventure Work - Evidence



Modeling: Create relationships between tables



Project 3 - Adventure Work - Evidence



Project 3 - Adventure Work - Results



In this project, I applied the skills learned from the bootcamp in cleaning, transforming, modelling, and exploring data with Power BI.

I also worked on creating charts and calculating key business metrics to make sense of the data.

To see the interactive report



[Report Adventure Work](#)

Narrative Arc of Project 3 - Situation



To practically apply the knowledge acquired during week 8 of the bootcamp, I have decided to design a narrative act based on one of the projects I have done previously. The chosen project is Adventure Works. I will use this framework to explore and tell the story within the data, employing all the narrative and storytelling techniques learned.

My goal is not only to demonstrate my understanding of these methods, but also how they can transform a raw data set into an understandable and engaging story.

Through this exercise, I hope to illustrate the power of data through storytelling, using Power BI to bring to life this visual narrative spanning the period from 2015 to 2017. The story will focus on uncovering sales trends, customer preferences, customers and product return patterns, all with the purpose of providing key information that facilitates strategic and tactical decision making at Adventure Works.



Narrative Arc of Project 3 - Evidence

Beginning

Problem statement

The company needs to identify effective strategies to increase revenue and reduce returns by understanding who buys and which are the star products.

Context / why this is important

Addressing this challenge is crucial for market leadership, improving product and customer satisfaction, and ensuring repeat business.

Hypothesis

By analyzing the data on product sales and returns, the company can identify which products are top-sellers with the least returns to optimize the product portfolio and focus on the most profitable items that align with customer satisfaction.

Middle

Key plot movements

1. "Water Bottle - 30 oz" dominates in orders, "Mountain 200 Black" in profit.
2. Bikes show highest returns with "Touring - 2000 Blue, 46" at 8.33%.
3. Australia, UK, France exceed sales goals.

Pivotal discovery

High return rate of 8.33% for "Touring - 2000 Blue, 46" reveals key quality or satisfaction issues, emphasizing the need for strategy reassessment.

Re-framing of problem statement

The focus expands from simply increasing sales and reducing returns to ensuring quality and aligning customer expectations, after identifying high return rates on key products.

End

Most important message to land

High return rates spotlight the need for enhanced product quality and better alignment with customer expectations.

Resolution, where outcome is progress

Initiatives to address quality and expectations significantly reduce returns and strengthen customer trust.

Tangible action

- Establish quality KPIs.
- Adjust product offerings based on thorough analysis.
- Launch feedback mechanisms for continuous improvement.

Narrative Arc of Project 3 - Evidence



Why this is important?

- Market leadership
- Improving product
- Increase customer satisfaction

Problem

Identify strategies to:

- Increase revenue
- Reduce returns

How

- What are the star products?
- Detecting which products with higher return rates
- Understanding who buys

Narrative Arc of Project 3 - Evidence



Key plot movements

1. Best sellers

Top Product By Orders

Water Bottle - 30 oz.

Top Product By Profit

Mountain-200 Black, 46

2. Category with highest returns:

Bikes: *Touring - 2000 Blue*, 46

CategoryName	Total Orders	Quantity Sold	Returns Rate	% Revenue per Product
⊕ Bikes	13.929	13.929	3,08 %	94,89 %
⊕ Clothing	6.976	12.436	2,16 %	1,47 %
⊕ Accessories	16.983	57.809	1,95 %	3,64 %
Total	25.164	84.174	2,17 %	100,00 %

CategoryName	Total Orders	Quantity Sold	Returns Rate	% Revenue per Product
Bikes				
Touring-2000 Blue, 46	96	96	8,33 %	0,47 %

3. Regions exceed sales goals last month

- Australia:
10.56% with 45 million
- United Kingdom:
25.07% with 24 million
- France:
22.41% with 19 million

Narrative Arc of Project 3 - Evidence



Pivotal discovery

High return rate of 8.33% for "Touring - 2000 Blue, 46" reveals:

- Key quality or satisfaction issues.
- Emphasizing the need for strategy reassessment.

Re-framing of problem statement

The focus expands from simply increasing sales and reducing returns to:

- Ensuring quality .
- Aligning customer expectations.
- Identifying high return rates on key products.

Narrative Arc of Project 3 - Evidence



Most important message to land

High return rates spotlight the need for enhanced product quality and better alignment with customer expectations.

Resolution, where outcome is progress

Initiatives to address quality and expectations significantly reduce returns and strengthen customer trust.

Tangible action

- Establish quality KPIs.
- Adjust product offerings based on thorough analysis.
- Launch feedback mechanisms for continuous improvement.

Narrative Arc of Project 3 - Results



- Facing the challenge of effectively communicating the results of a complex data analysis, a structured approach was required to tell the story of the findings.
- I decided to create a narrative arc for the Adventure Works project, which would clearly narrate the process and the results of the analysis.
- Using the narrative arc structure, I identified the key points of the project and applied storytelling techniques to craft a coherent story. This included establishing the context, highlighting the pivotal discovery, and framing the resolution and actions taken.
- The narrative arc significantly improved the presentation and understanding of the analysis results. I learned to appreciate the importance of storytelling in data analytics and to use narrative to make data accessible and compelling to a broader audience."