Quantitative Data Reasoning

OUTLINE

1. Data Driven Decision Making
2. Statistical Data Analysis
3. Data Analysis using Power Pivot
4.
5.
6.

Data-Driven Decision Making

1

Ask Questions to your Data !

Data and Analytics Framework

Putting the Framework into Action

Discovery

Define the problem

- What is the key opportunity?
- Engage stakeholders for perspective and concerns

Develop Hypotheses

- Answer 'what is likely to happen'?
- Use information from stakeholders and other knowledge to refine hypotheses
- Choose the hypothesis for which the best data exists

Collect Data

- Collect relevant internal and external data sets
- Validate the accuracy of the data

Insights

Explore Data

- Explore data sets to understand how they would help in accepting or refuting the hypotheses

Analyze Data

- Use Qualitative and Quantitative analysis techniques to use data to validate the hypotheses
- Convert outputs into user-friendly formats and visualizations that will help different stakeholders understand the analysis

Actions

Link Insights

- Use actionable data insights to explain past outcomes and predict the future landscape
- Link insights to financial and operational metrics to specify impact and aid decision making

Provide Recommendations

- Prioritize insights to build actionable plans
- Provide solutions that help business to address future challenges

Outcomes

Execute Plan

- Develop clear pathways of how insights will be delivered to the right stakeholders at the right time
- Ensure the plans meet long term business objectives and help refine solutions in the future

Data Driven Decision Making - Simulation Exercise

https://bit.ly/3dm-blink

Discovery

Define the problem

- What is the key opportunity?
- Engage stakeholders for perspective and concerns

Develop Hypotheses

- Answer 'what is likely to happen'?
- Use information from stakeholders and other knowledge to refine hypotheses
- Choose the hypothesis for which the best data exists

Collect Data

- Collect relevant internal and external data sets
- Validate the accuracy of the data

Insights

Explore Data

- Explore data sets to understand how they would help in accepting or refuting the hypotheses

Analyze Data

- Use Qualitative and Quantitative analysis techniques to use data to validate the hypotheses
- Convert outputs into user-friendly formats and visualizations that will help different stakeholders understand the analysis

Actions

Link Insights

- Use actionable data insights to explain past outcomes and predict the future landscape
- Link insights to financial and operational metrics to specify impact and aid decision making

Provide Recommendations

- Prioritize insights to build actionable plans
- Provide solutions that help business to address future challenges

Outcomes

Execute Plan

- Develop clear pathways of how insights will be delivered to the right stakeholders at the right time
- Ensure the plans meet long term business objectives and help refine solutions in the future

2

Understand your Numbers !

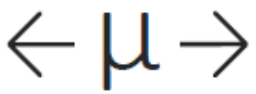
Statistics

Descriptive Statistics

Understand what your sample data looks like

Probability Distributions

If the sample data fits a probability distribution, use it as a **model** for the entire population



Confidence Intervals

If the sample doesn't fit a distribution, use the central limit theorem to make **estimates** about population parameters

Hypothesis Tests

Continue to leverage the central limit theorem to draw **conclusions** about what a population looks like based on a sample

Regression Analysis

Use additional variables to increase the accuracy of your estimates and make **predictions** based on their relationships

Statistics

MAVEN PIZZA PARLOR | PROJECT BRIEF

You are a BI Consultant that has just been approached by **Maven Pizza Parlor**, a new pizza place in New Jersey that needs help with their demand planning

From: **Mary Margherita** *(Owner)*

Subject: **Daily Pizza Sales**

Hi!

We we're extract our daily pizza sales from our POS system, and we want to use this for planning, but none in the team is data savvy.

Is that something you could help us with?

We want to know how many pizza sales to expect every day, how much they typically vary, and if they fluctuate by day of the week.

Thank you!

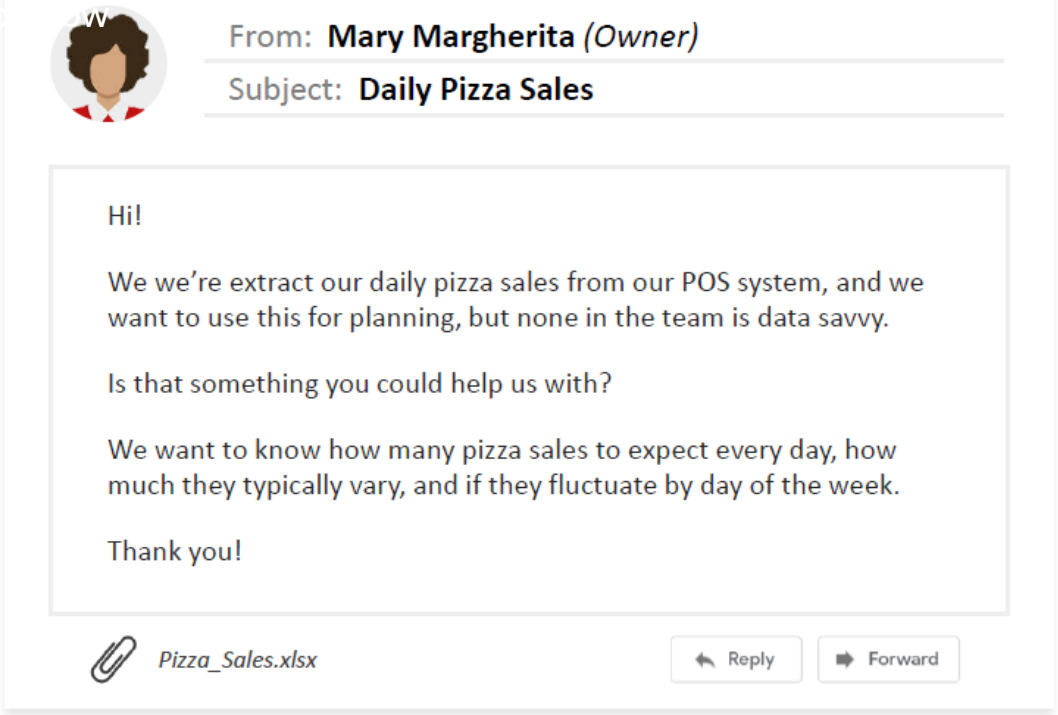
Pizza_Sales.xlsx

Reply Forward

Statistics

MAVEN PIZZA PARLOR | PROJECT BRIEF

You are a BI Consultant that has just been approached by **Maven Pizza Parlor**, a new pizza place in New Jersey that needs help with their demand planning



From: **Mary Margherita** *(Owner)*

Subject: **Daily Pizza Sales**

Hi!

We we're extract our daily pizza sales from our POS system, and we want to use this for planning, but none in the team is data savvy.

Is that something you could help us with?

We want to know how many pizza sales to expect every day, how much they typically vary, and if they fluctuate by day of the week.

Thank you!

Pizza_Sales.xlsx

Reply Forward

Key Objectives

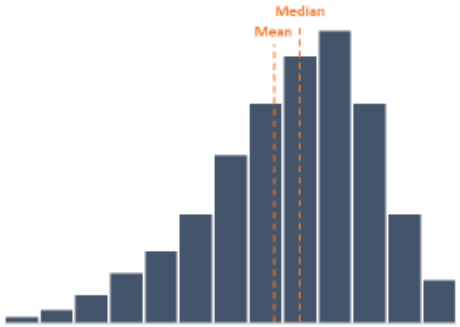
1. Summarize the daily pizza sales by using descriptive statistics

SKEW

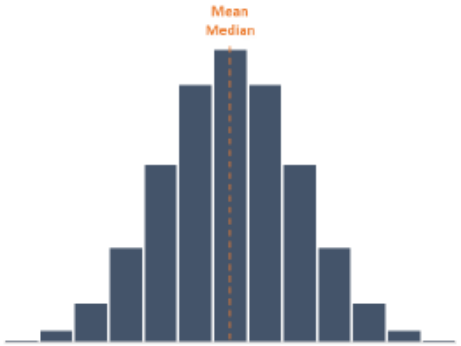
The **skew** represents the asymmetry of a distribution around its mean

- In a **zero-skewed** distribution, the mean and median are equal
- In a **right-skewed** *(or positive)* distribution, the mean is typically greater than the median
- In a **left-skewed** *(or negative)* distribution, the mean is typically smaller than the median

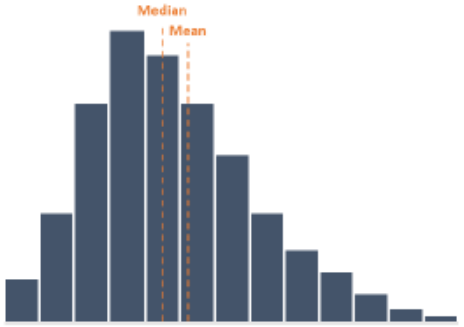
Left skew



Zero skew



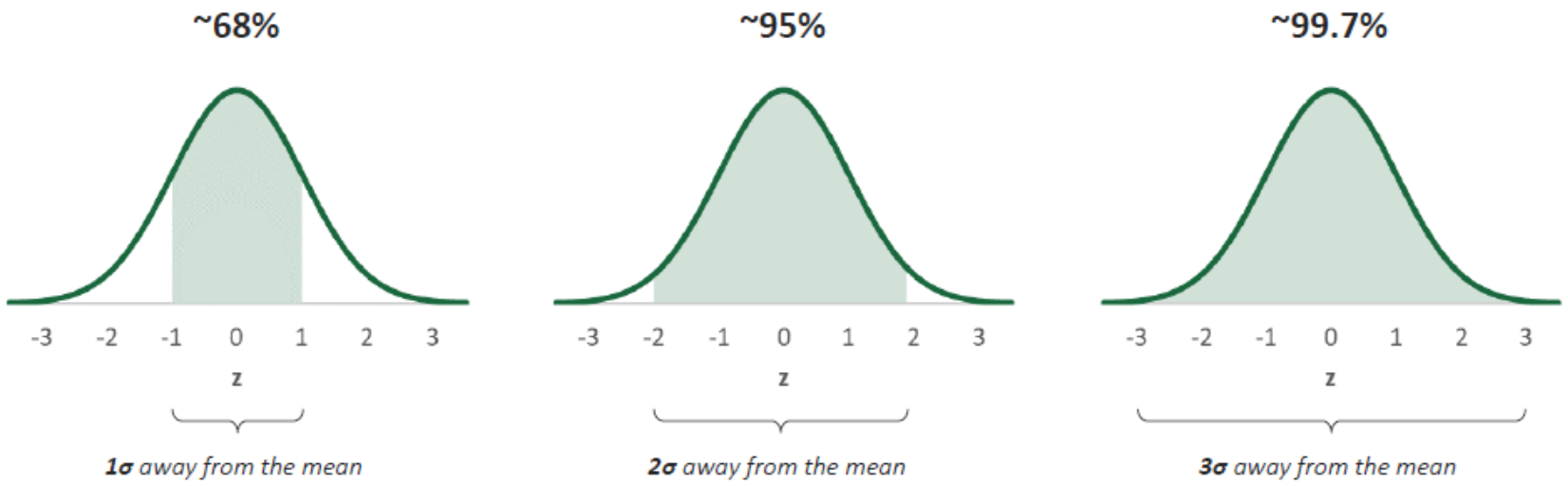
Right skew



*This is one of the properties of a **normal distribution** (more on that later!)*

THE EMPIRICAL RULE

The **empirical rule** outlines where most values fall in a normal distribution



PRO TIP: Beyond using a histogram to determine whether your data is distributed normally, check if it follows the empirical rule

INTERQUARTILE RANGE

The **interquartile range** is the spread of the *middle half* of the values in a variable

- In other words, it's the spread from the **first quartile** to the **third quartile**

Undergrad Grade	
	63.3
25%	66
	67.5
	68.1
	68.7
25%	70.1
	74
	74.6
	75.3
25%	75.6
	76
	78.9
	79.3
25%	82.9
	88.8
	96.3

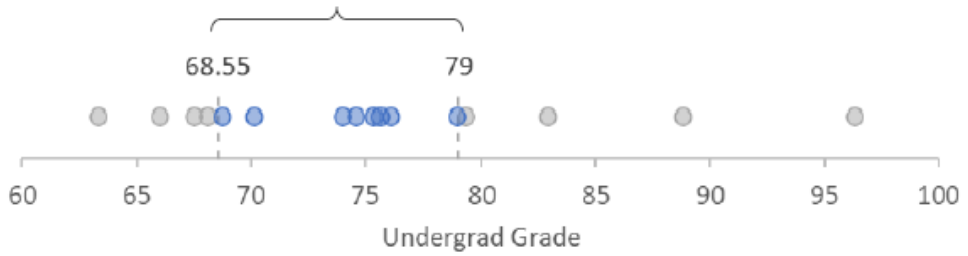
Q1 = 68.55

Median

Q3 = 79

n=16

$$IQR = 79 - 68.55 = \mathbf{10.45}$$



HEY THIS IS IMPORTANT!

The quartiles in this example are calculated by *including* the median, but Excel also lets you use the *exclusive* method

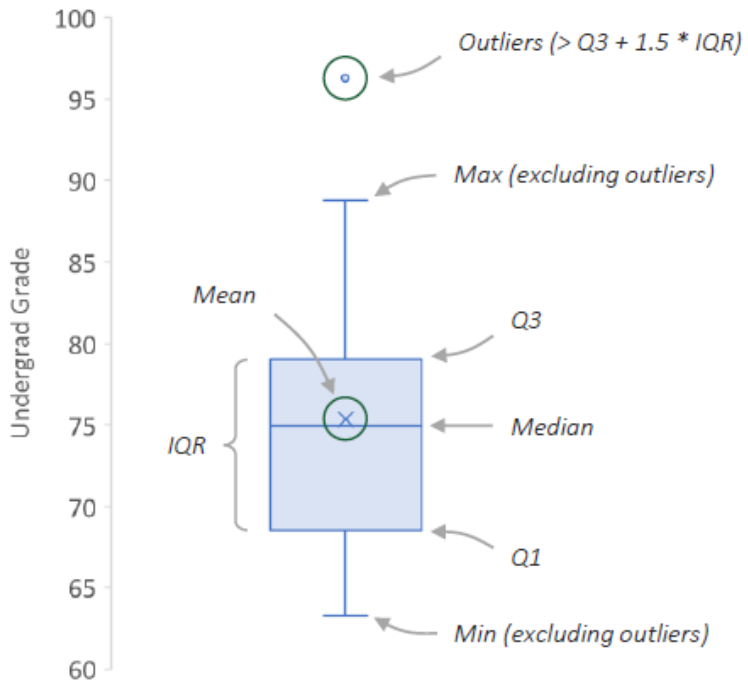
There is no right or wrong method to use, but many prefer the inclusive method, as it leads to a narrower range

BOX & WHISKER PLOTS

Box & whisker plots are used to visualize key descriptive statistics

Undergrad Degree	Undergrad Grade
Business	78.9
Business	74
Business	74.6
Engineering	79.3
Engineering	70.1
Business	88.8
Business	66
Art	82.9
Business	96.3
Business	75.6
Finance	67.5
Computer Science	68.7
Business	76
Engineering	75.3
Engineering	68.1
Finance	63.3

n=16



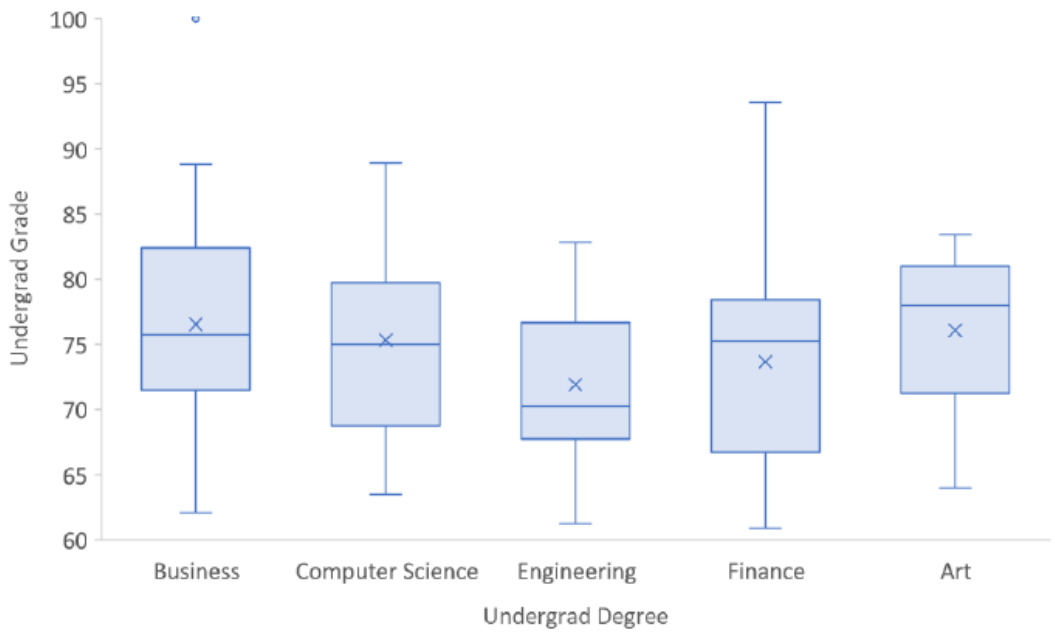
BOX & WHISKER PLOTS

Box & whisker plots are used to visualize key descriptive statistics

- They can be used to quickly compare statistical characteristics between categories

Undergrad Degree	Undergrad Grade
Business	78.9
Business	74
Business	74.6
Engineering	79.3
Engineering	70.1
Business	88.8
Business	66
Art	82.9
Business	96.3
Business	75.6
Finance	67.5
Computer Science	68.7
Business	76
Engineering	75.3
Engineering	68.1
Finance	63.3

n=95



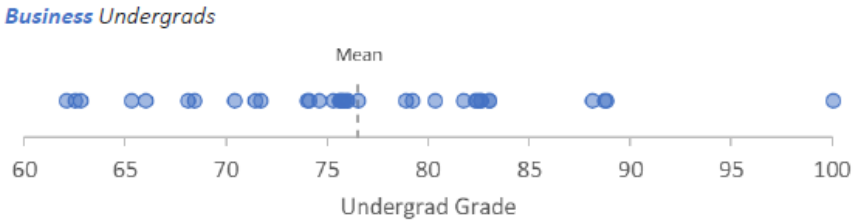
STANDARD DEVIATION

The **standard deviation** measures, on average, how far each value lies from the mean

- The *higher* the standard deviation, the *wider* a distribution is *(and vice versa)*

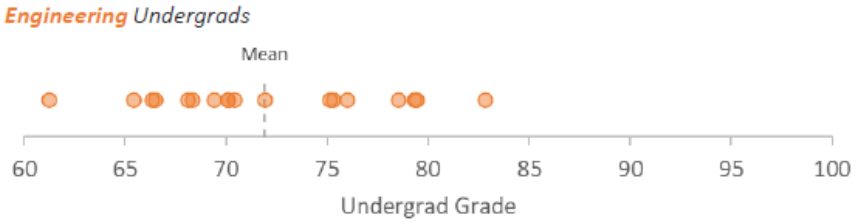
Undergrad Degree	Undergrad Grade
Business	78.9
Business	74
Business	74.6
Engineering	79.3
Engineering	70.1
Business	88.8
Business	66
Art	82.9
Business	96.3
Business	75.6
Finance	67.5
Computer Science	68.7
Business	76
Engineering	75.3
Engineering	68.1
Finance	63.3

n=95



Std Dev

= **8.17**



= **5.79**

Statistics

MAVEN MEDICAL CENTER | PROJECT BRIEF

You are a Data Analyst at the **Maven Medical Center** in Springfield, MA and just received a project request from the chief gynecologist

From: **Betty Birth** *(Chief Gynecologist)*

Subject: **Need some probability figures**

Good morning!

We've had over 30% of the babies born this year weigh under 2.5kg, which is considered low. The percentage itself seems a little high to me though. Is there any way you could check what the probability of a baby weighing under 2.5kg is with the data we have?

I could also use the number of births we've had so far in the top & bottom 1% if possible.

Thank you!

Birth_Weights.xlsx

Reply Forward

Statistics

MAVEN MEDICAL CENTER | PROJECT BRIEF

You are a Data Analyst at the **Maven Medical Center** in Springfield, MA and just received a project request from the chief gynecologist

From: **Betty Birth** *(Chief Gynecologist)*

Subject: **Need some probability figures**

Good morning!

We've had over 30% of the babies born this year weigh under 2.5kg, which is considered low. The percentage itself seems a little high to me though. Is there any way you could check what the probability of a baby weighing under 2.5kg is with the data we have?

I could also use the number of births we've had so far in the top & bottom 1% if possible.

Thank you!

Birth_Weights.xlsx

Reply Forward

Key Objectives

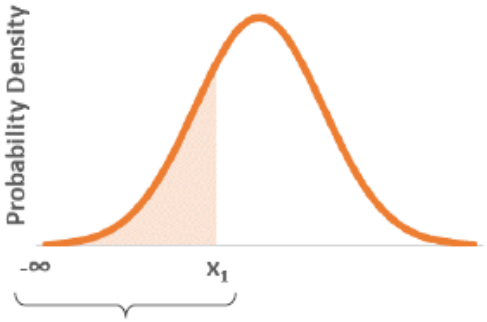
1. Check if the weights can be assumed to follow a normal distribution
2. If so, calculate the probability of a baby weighing 2.5kg or less
3. Estimate the values at the 1% and 99% cumulative probabilities
4. Count the number of births under and over those thresholds

CALCULATING PROBABILITIES

If a variable follows a normal distribution, you can **calculate the probability** of randomly obtaining a value within a specified range

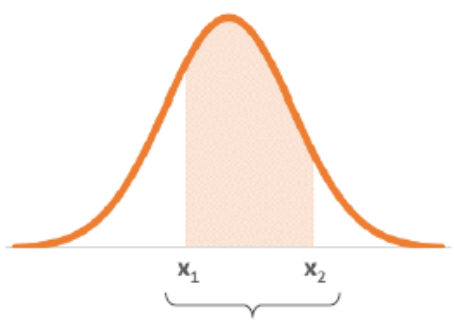
- This is determined by the area under the curve in that range

Less than or equal to "x_1"



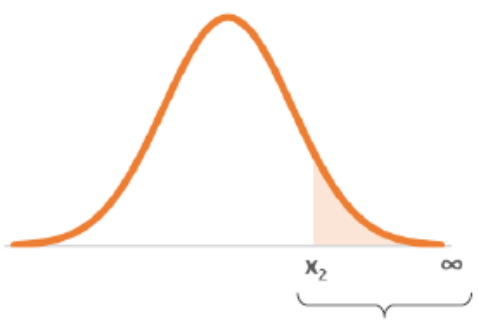
*The area from negative infinity to "x_1" is the **cumulative probability***

Between "x_1" and "x_2"



This is the cumulative probability of "x_2" minus the cumulative probability of "x_1"

Greater than or equal to "x_2"



This is 1 (the entire area under the curve) minus the cumulative probability of "x_2"

HEY THIS IS IMPORTANT!

You CANNOT calculate the probability of obtaining an x value *exactly* – there's no area under a single point!

THE NORM.DIST FUNCTION

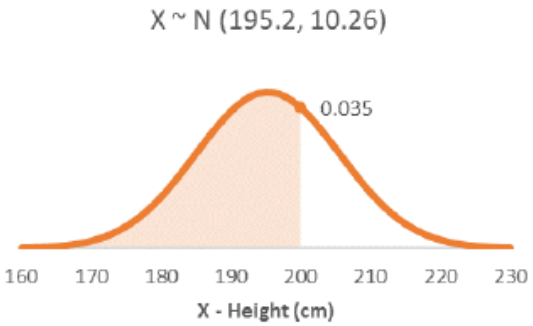
NORM.DIST() — *Returns the cumulative probability or the probability density at "x" from a normal distribution*

=**NORM.DIST**(x, mean, standard_dev, cumulative)

- *The **value** to calculate the probability for*
- *The **mean** & **standard deviation** for the normal distribution of the population*
- ***TRUE**: The area under the curve*
  ***FALSE:** The height of the curve*

Possible question:

*"What's the probability of an Olympic Basketball Player being **2 meters tall or shorter**?"*



=**NORM.DIST**(200, 195.2, 10.26, TRUE) **= 0.68** ← *This is the probability!*

=**NORM.DIST**(200, 195.2, 10.26, FALSE) **= 0.035** ← *This is just the height of the curve*

THE NORM.DIST FUNCTION

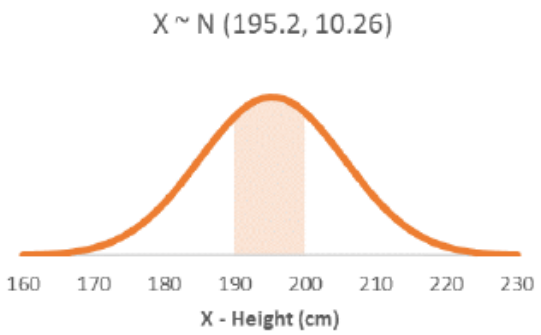
NORM.DIST() — *Returns the cumulative probability or the probability density at "x" from a normal distribution*

=**NORM.DIST**(x, mean, standard_dev, cumulative)

- *The **value** to calculate the probability for*
- *The **mean** & **standard deviation** for the normal distribution of the population*
- ***TRUE**: The area under the curve*
  ***FALSE:** The height of the curve*

Possible question:

*"What's the probability of an Olympic Basketball Player being **between 1.9 and 2 meters tall**?"*



=**NORM.DIST**(200, 195.2, 10.26, TRUE) = **0.68**

=**NORM.DIST**(190, 195.2, 10.26, TRUE) = **0.3061**

=0.68-0.306 = **0.3739**

This is the probability!

THE NORM.DIST FUNCTION

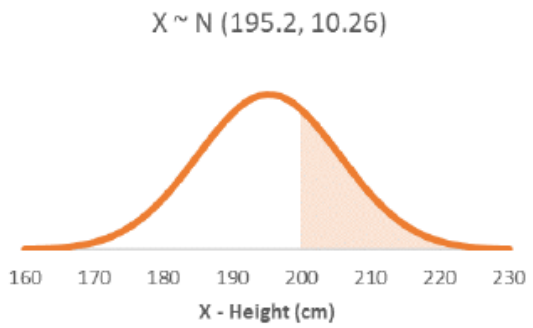
NORM.DIST() — *Returns the cumulative probability or the probability density at "x" from a normal distribution*

=**NORM.DIST**(x, mean, standard_dev, cumulative)

- *The **value** to calculate the probability for*
- *The **mean** & **standard deviation** for the normal distribution of the population*
- ***TRUE**: The area under the curve*
  ***FALSE:** The height of the curve*

Possible question:

*"What's the probability of an Olympic Basketball Player being **at least 2 meters tall**?"*



=**NORM.DIST**(200, 195.2, 10.26, TRUE) = **0.68**

=1-**NORM.DIST**(190, 195.2, 10.26, TRUE) = **0.32**

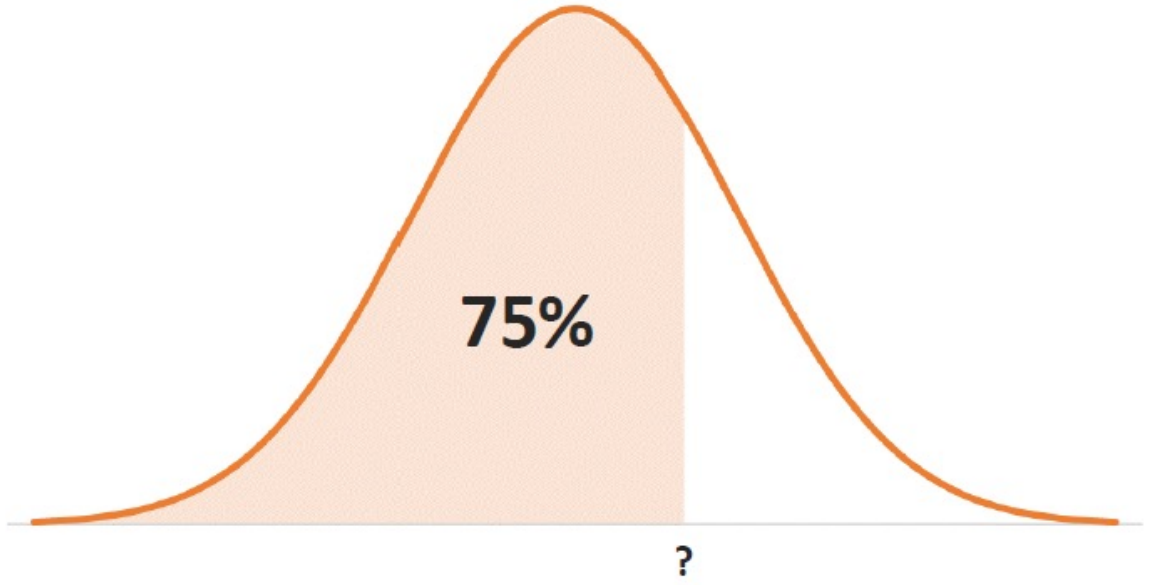
This is the probability!

The cumulative probability under the entire curve is equal to 1

(it's every value possible!)

ESTIMATING VALUES

If a variable follows a normal distribution, you can **estimate the value of "x" or "z"** at a specified cumulative probability



THE NORM.INV FUNCTION

NORM.INV() — *Returns the x value in a normal distribution at a specified cumulative probability*

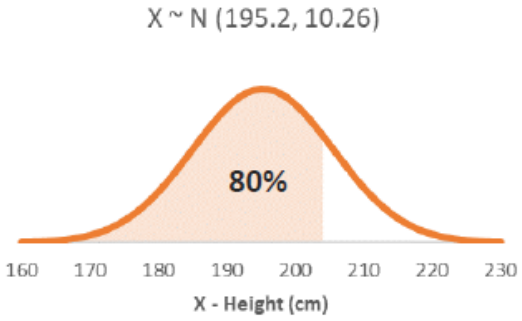
=**NORM.INV**(probability, mean, standard_dev)

*The **cumulative probability** for the value you want*

*The **mean** & **standard deviation** for the normal distribution of the population*

Possible question:

*"How tall do you need to be to be **taller than 80%** of Olympic Basketball Players?"*



=**NORM.INV**(0.8, 195.2, 10.26) = **203.8** cm

3

Visualize your Data!


HR Analysis:
Employee Retention

Employee Turnover Analysis

Problem Statement: Management wants to understand how to reduce employee turnover.

Goal: HR wants to create an employee retention program.

Task: Analysis, hypothesis and data story on reasons for churn.

Data: ~15,000 employee records.

Questions from Management:

- What is the main cause of turnover?
- Is there something surprising in the data?
- What segment should we focus on?
- Which department has the highest turnover?
- Do we need to increase X or decrease X?
- Where should we put our pilot program

Insight Development

How to develop insights?(W.H.W)

1. **What's the goals of the business?**

   Make money/reduce employee churn/limit recruitment cost

2. **What is the metric of success or failure?**

   Employee retention/churn

3. **What are the trends?(positive or negative)**

   Departments with high and low churn

4. **What influences our metrics and trends?**

   Other metrics' affect on churn

5. **How can we fix the trends?**

   Lowering/increasing X may lower or increase

Tools & Techniques

Tools: Excel and PowerPoint

Techniques: Pivot Table, Power Query, DAX

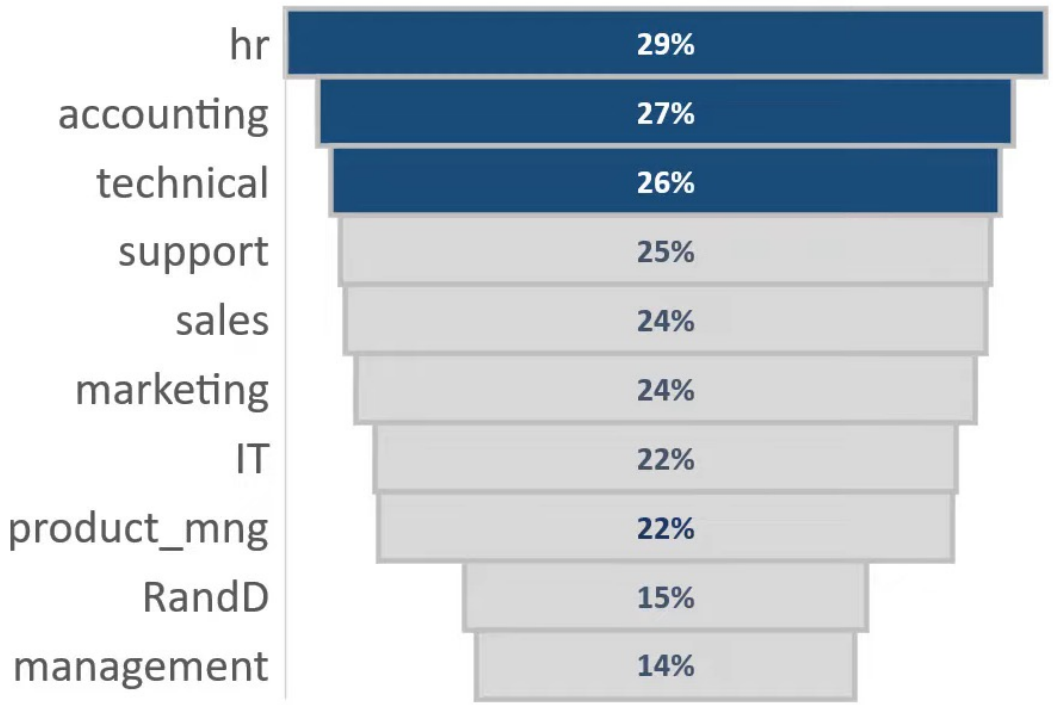
Statistics: Mean, Median, Sum, Count, Percentile

Visuals: Stacked Bar, Boxplot, Funnels, Pie Charts

Where Do We Have the Most Churn?

24%

Company Turnover



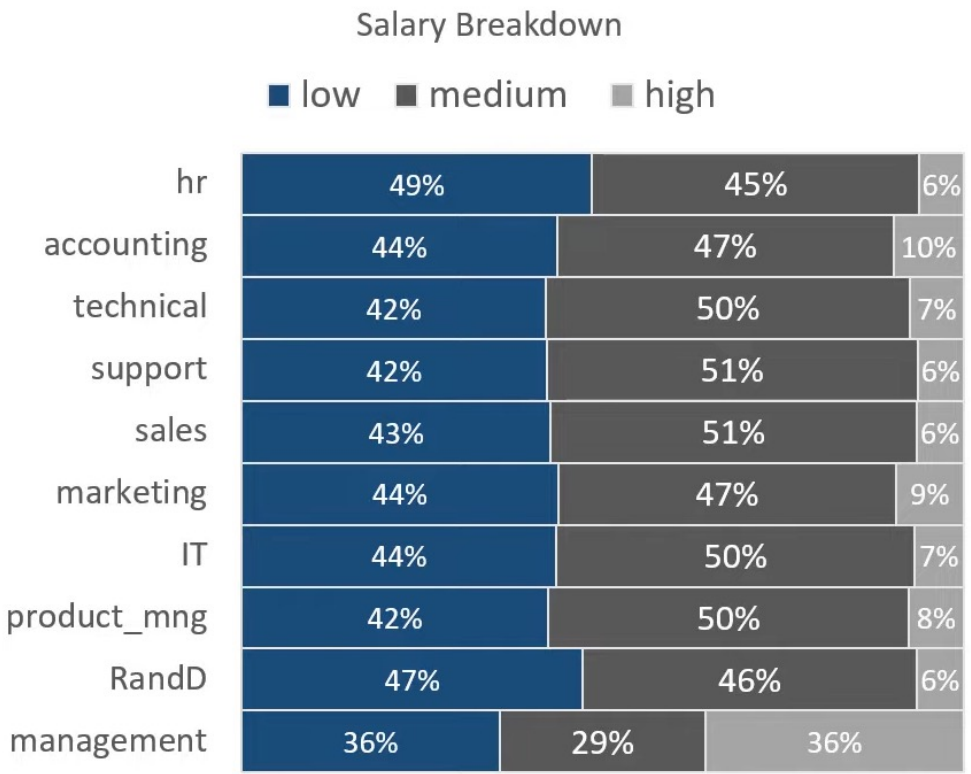
DEPARTMENT TURNOVER

These departments have the most churn. However, we need to ask what is the representation of these departments in the company and what is driving this churn?

Does Salary Affect Employee Retention?

High Churn & Low Salary

The departments with the most churn also have the most employee in the low salary range.

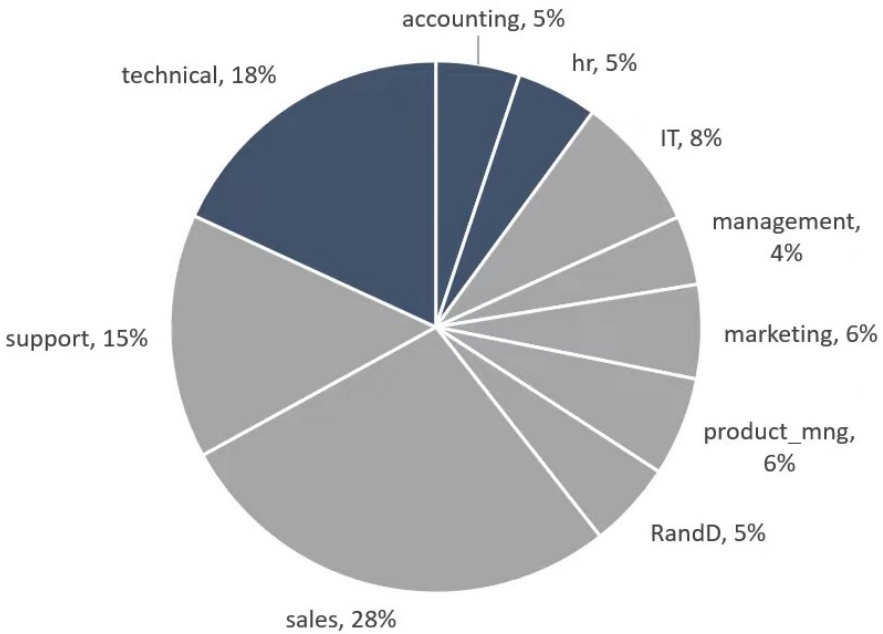


Salary

Although salary are lower for the top 3 departments with the lowest retention. Not all the categories have the lowest salaries. However, high medium and high salaries do show greater retention.

Does Salary Affect Employee Retention?

Where are most employees concentrated?



Top 3 Departments by Churn and Employees

Although these departments have the most churn there don necessarily equal large volume of employee. However, we should evaluate these departments difficult in recruitment.

Is Working Long Hours Affected Churn ?

Top 3 Departments by Churn and Long Hours

When evaluating the long hours outliers which would be at the 90th percentile. It's easy to determine that the technical department has the highest amount of employee in this segment.

Summary & Recommendations

Summary:

The **overall churn of the companies sits at 24%.** This indicates that there may be an issue since the **industry average is between 12 and 15%.**

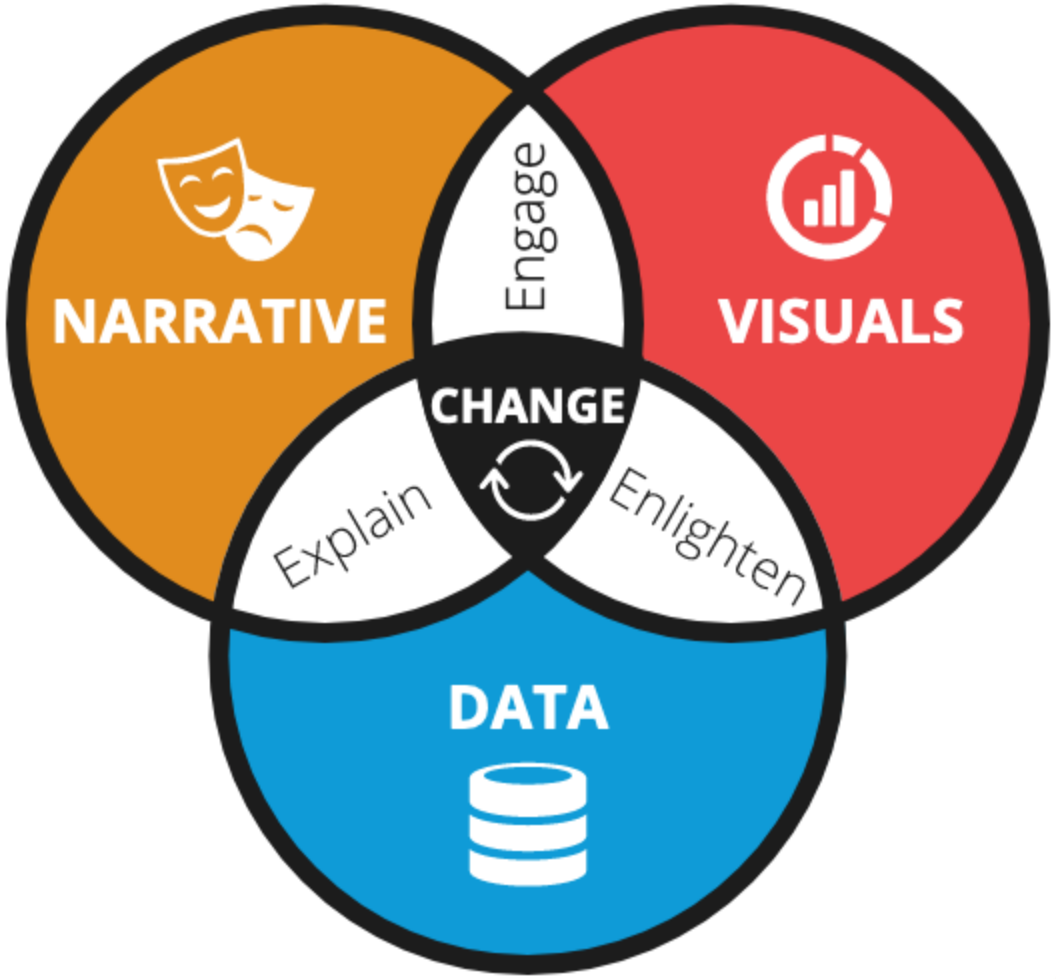
We have identified 3 candidates for a pilot program who have the highest churn. Out three segments, the **technical has the greatest number of employees at 18% at churn of 26% while HR(29% churn) and accounting(26 % churn) make up 5% of employee each, respectively**.

Salary and work hours may factor into the department churn with these segments having the majority of employees in the low and mid salary ranges. **Technical employees have 11% of employee working more than 267 hours month or more per month. This would be the best candidate for the pilot program.**

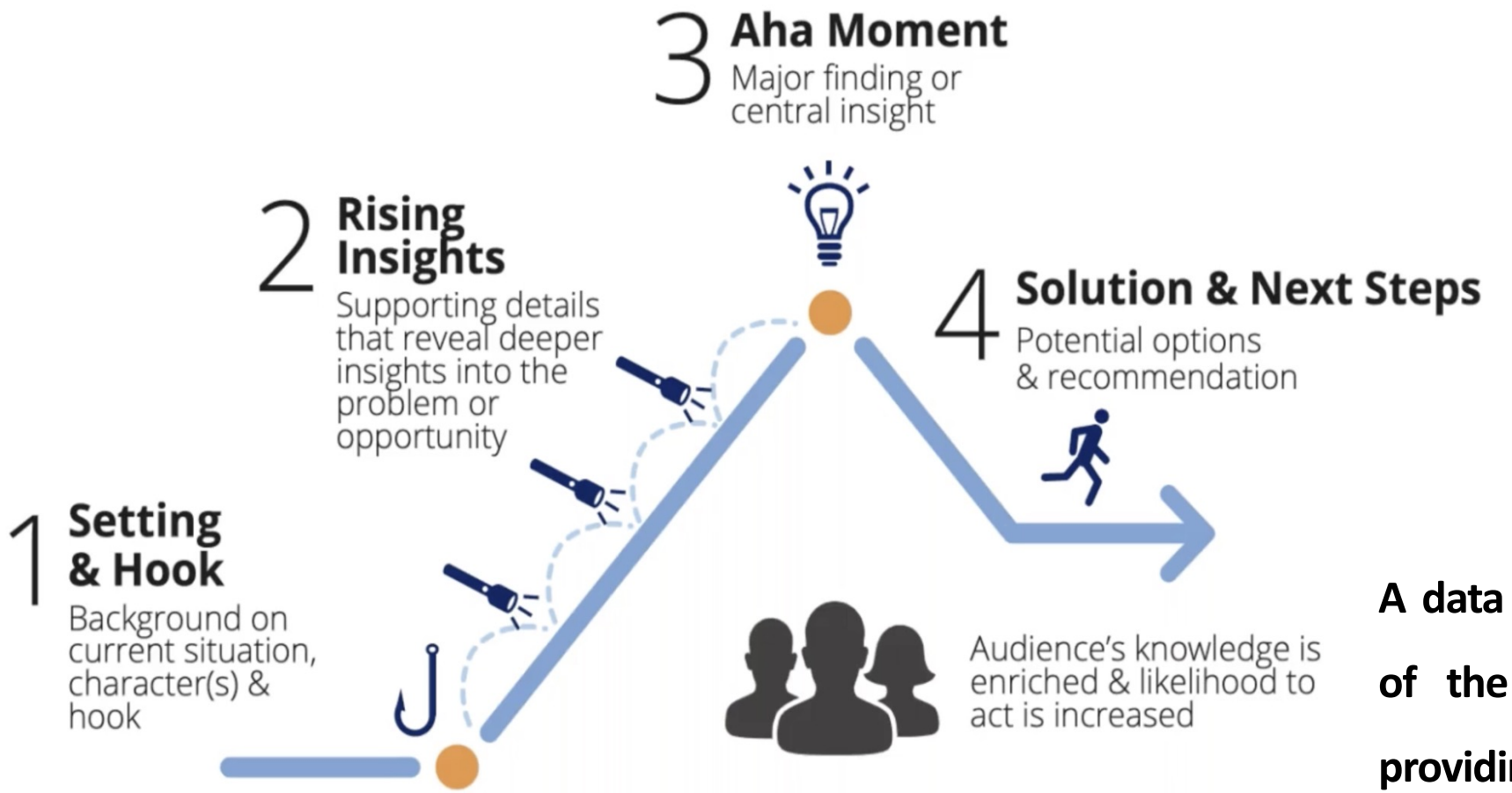
4

Data Storytelling Connect Dots!

Telling Effective Data Stories with Narrative, Data, and Visuals



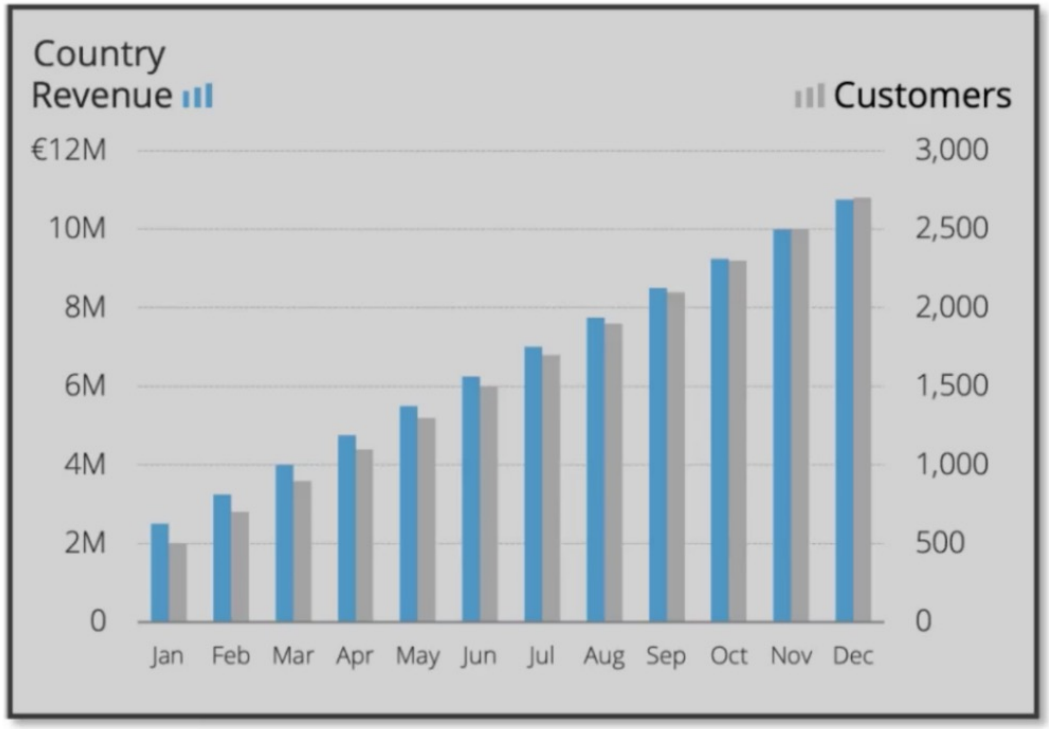
Narrative Structure

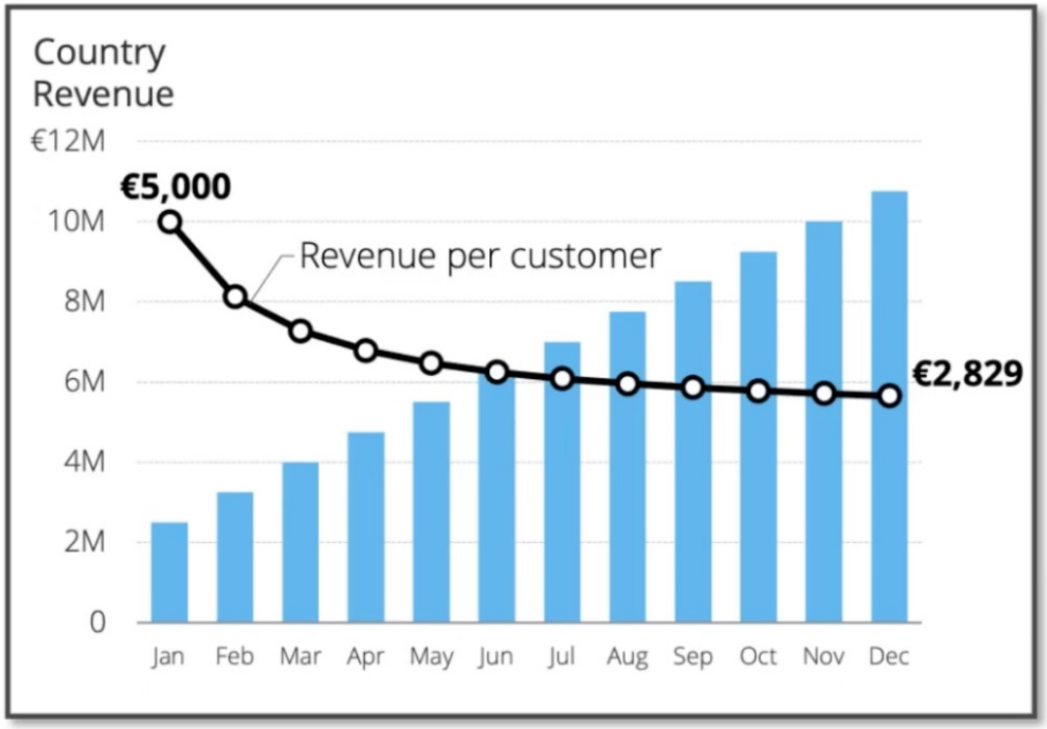


A data story begins by setting the scene of the current situation, proceeds by providing insights that lead up to the central insight, and ends with relevant recommendations.

Identify Right Data for Your Data Story

Calculated metrics may be more insightful than total values.

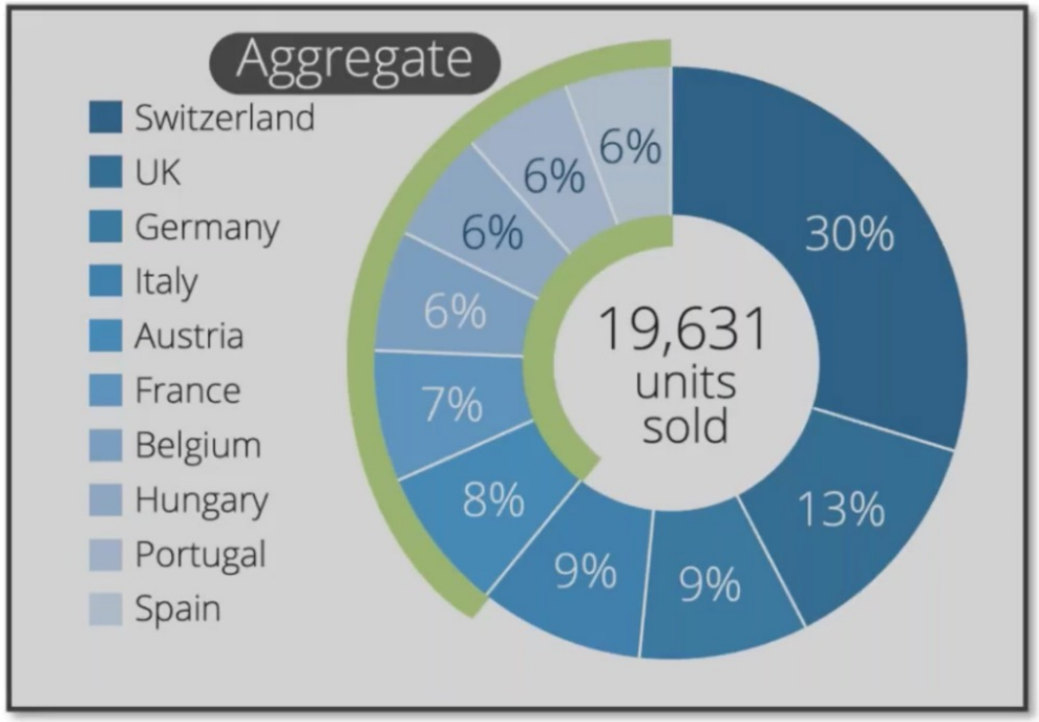


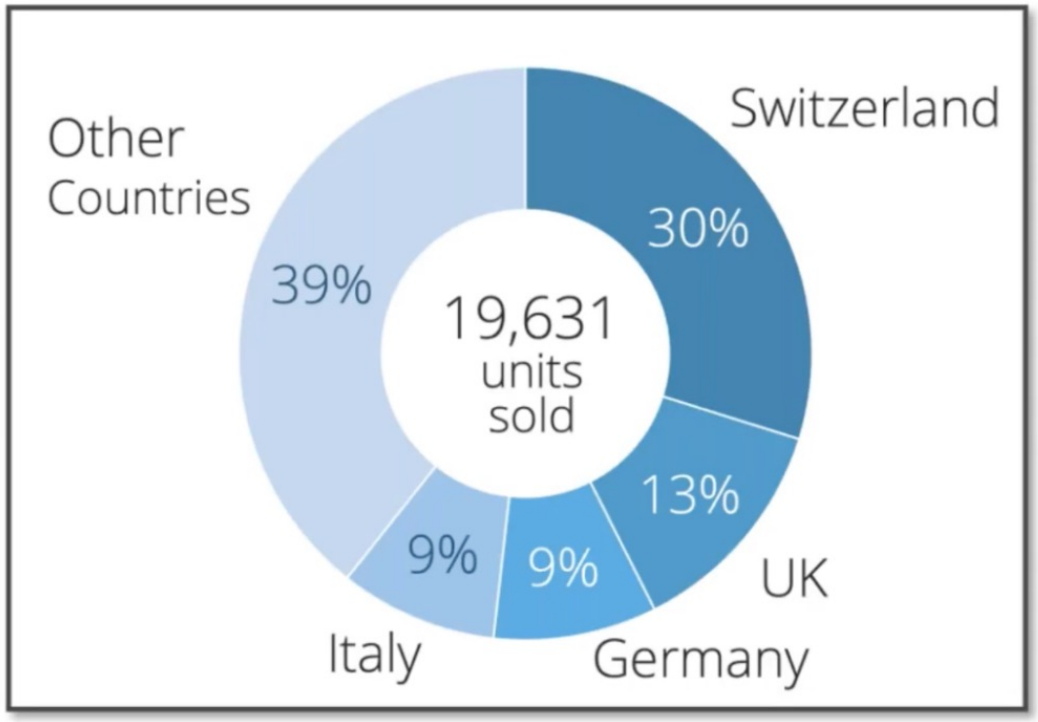


Explicitly demonstrating that the revenue per customer is falling (right) is a better choice than plotting the total revenue and customer side-by-side (left)

Aggregate Less Important Information

To simplify charts, you can **aggregate less critical data** to reduce the cognitive load.





The market shares of the largest markets become apparent when the smallest markets are aggregated.

1

Ask Questions to your Data !

2

Understand your Numbers !

3

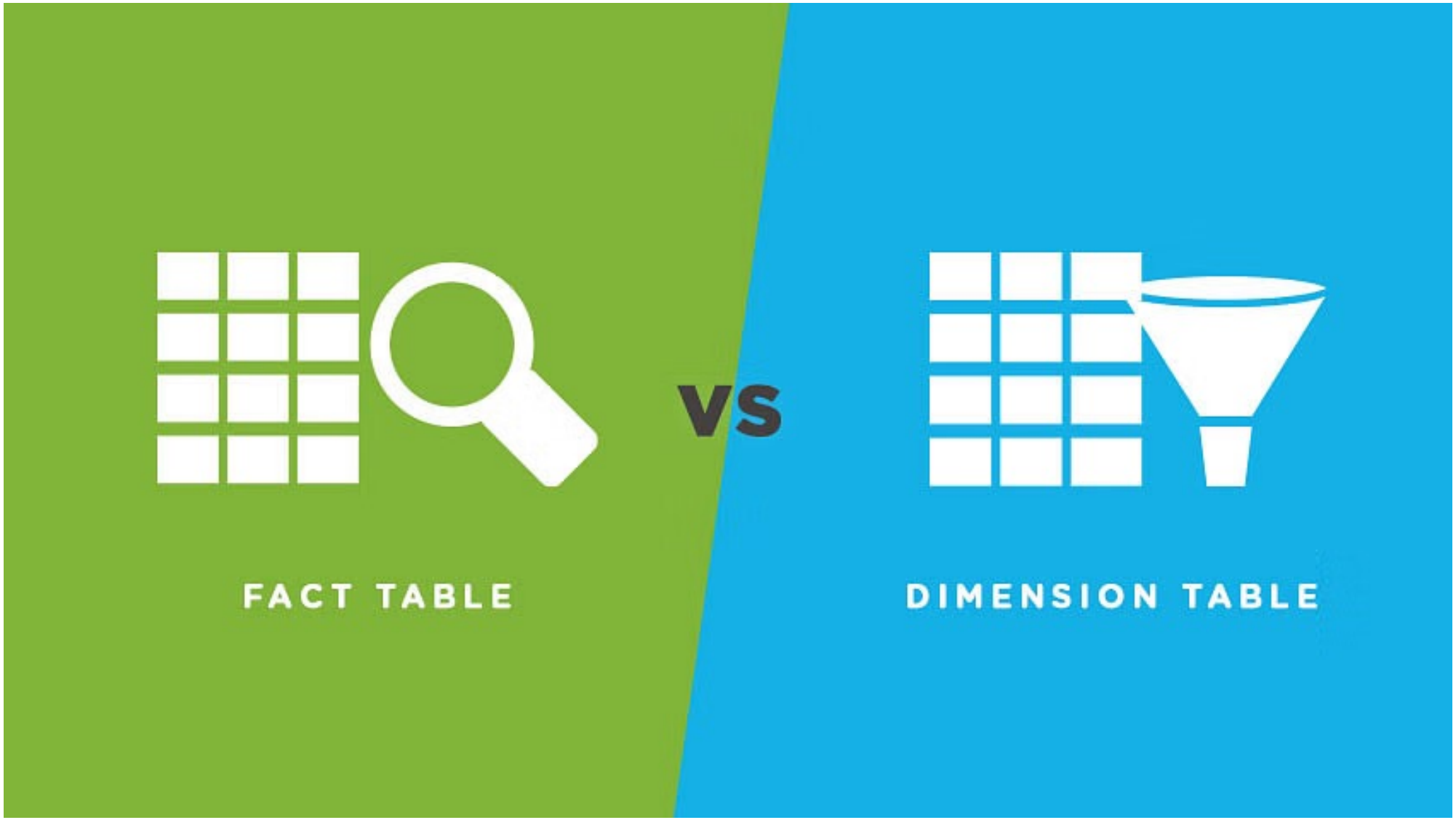
Visualize your Data!

4

Data Storytelling Connect Dots!

5

Data Modelling
Mingle your Data!


VS
FACT TABLE
DIMENSION TABLE

Fact Table

Order ID	Date	Product ID	Customer ID	Quantity	Price	Total Order Amount
123456	12-04-2000	1555	4564	3	1000	3000

Dimension Table

Product ID	Product Name	Category	Sub-Category	Brand	Price
1555	Chair	Furniture	Household	ABC	1000

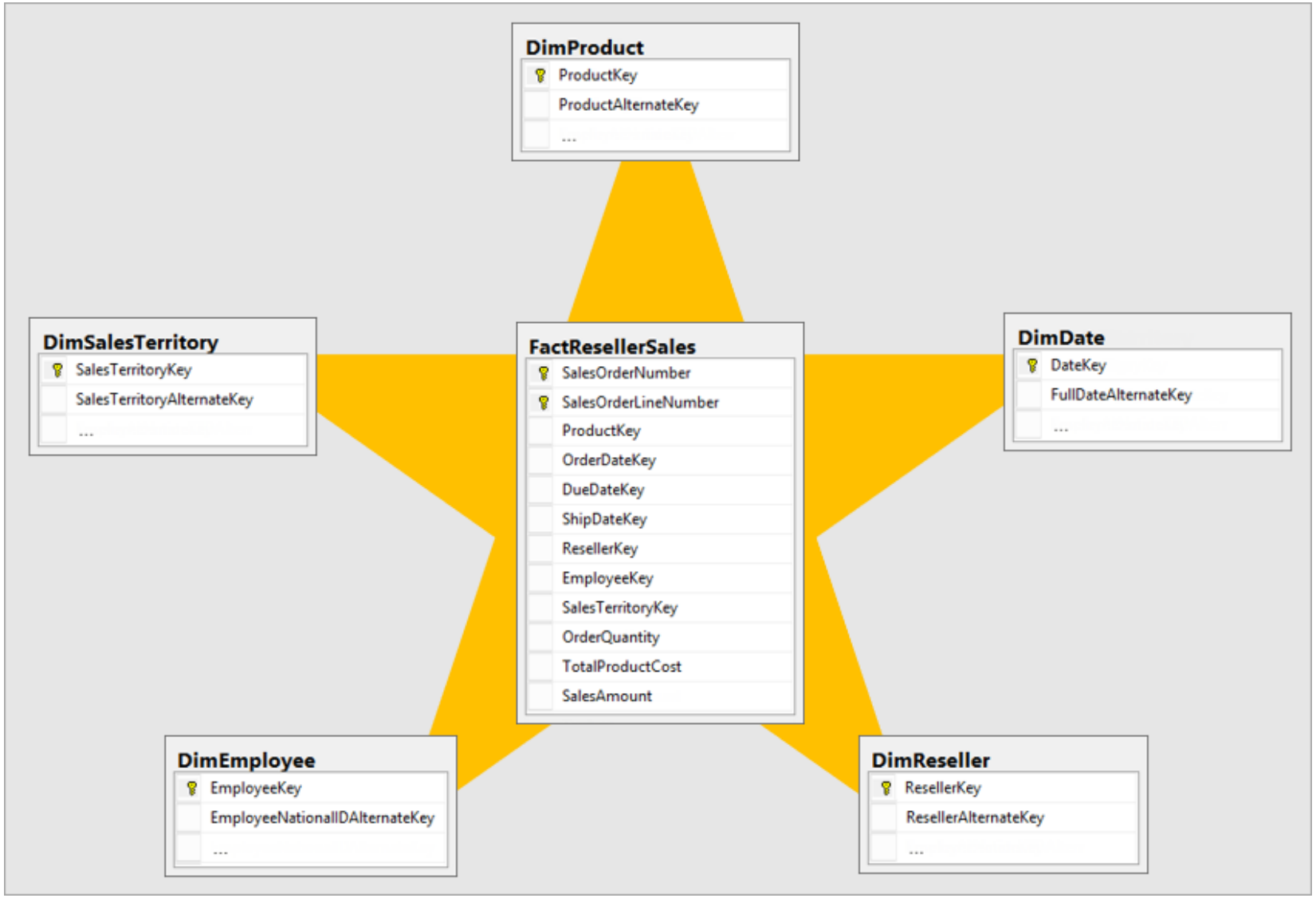
Fact vs Dimension Table

Basis	Fact Table	Dimension Table
Contents	Numeric values and transactional data.	Categorical data and descriptive attributes.
Purpose	Stores quantitative measures and metrics.	Provides descriptive attributes and context.
Size	Larger in terms of data volume.	Smaller in terms of data volume.
Aggregation	Aggregates data for analysis and reporting.	Provides context for data aggregation.
Querying	Provides data for analysis and calculations.	Used for filtering and categorization
Examples	Sales transactions and inventory levels.	Date, product, store, and customer dimensions
Rows	Many rows.	Fewer rows.

Star Schema


BEST RUN SHOES
TIME
DAY
MONTH
YEAR
QUARTER
HOUR
SHOE SALES
FACTS
ITEM_KEY
TIMESTAMP
CUSTOMER_ID
PAYMENT
QTY_SOLD
AMT_SOLD
AVG_SALES
ITEM
ITEM_NAME
BRAND
COLOR
SIZE
CUSTOMER
CUSTOMER ID
ADDRESS
CITY
COUNTRY
PAYMENT
PAYMENT METHOD
INSTALLMENTS

Star Schema


DimProduct
ProductKey
ProductAlternateKey
...
DimSalesTerritory
SalesTerritoryKey
SalesTerritoryAlternateKey
...
FactResellerSales
SalesOrderNumber
SalesOrderLineNumber
ProductKey
OrderDateKey
DueDateKey
ShipDateKey
ResellerKey
EmployeeKey
SalesTerritoryKey
OrderQuantity
TotalProductCost
SalesAmount
DimDate
DateKey
FullDateAlternateKey
...
DimEmployee
EmployeeKey
EmployeeNationalIDAlternateKey
...
DimReseller
ResellerKey
ResellerAlternateKey
...

Let’s do that in using Power Pivot