Kaggle Mavericks

Unleashing your Inner Data Science

Waiz Wafiq

11th November 2023

The Cubes, Block A
Faculty of Computer Science & Information Technology, UM

Workshop Tentative

	Time	Activity
Saturday 11/11/2023	10:10 am – 10:20 am	Introduction
	10:20 am – 12:10 pm	Workshop Session 1
	12:10 pm – 1:10 pm	Lunch Break!
	1:10pm – 2:45pm	Workshop Session 2
	2:45 pm – 3:00 pm	Workshop Ends; Photo Session

QnA – Slido Link

https://qrco.de/beXIuo

What is Kaggle?

The popular platform for

data science competitions

≥ 50,000 **public** datasets

Industry-related;
Also suitable for
computer vision

Our Objectives

free lunch la aiyo

Our Objectives

Equip participants with a solid understanding of **exploratory data analysis** (EDA) techniques and descriptive analysis to unveil patterns, trends, and insights within datasets.

Enable participants to effectively choose and **engineer features** that significantly impact model performance.

Delve into **predictive analytics**, guiding participants through the process of building machine learning models.

Can Kaggle help become a data scientist?

☑ **YES**

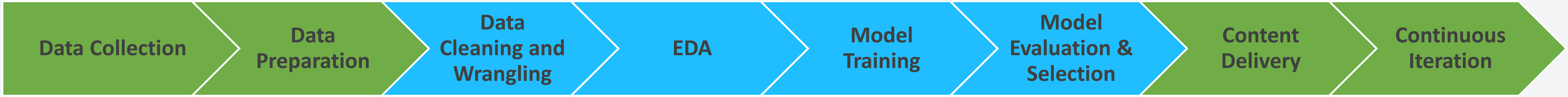
- A good approach to real-life problems
- Learn new libraries in R or Python
- Learn from the Kaggle community sharing their explorations/solutions

✕ **NO**

- Skips data collection
- Overemphasis on the machine learning part of data science, which is a minority part of the job

General Data Science Process

“Turn **data** into **insights**”



Data Collection

Data Preparation

Data Cleaning and Wrangling

EDA

Model Training

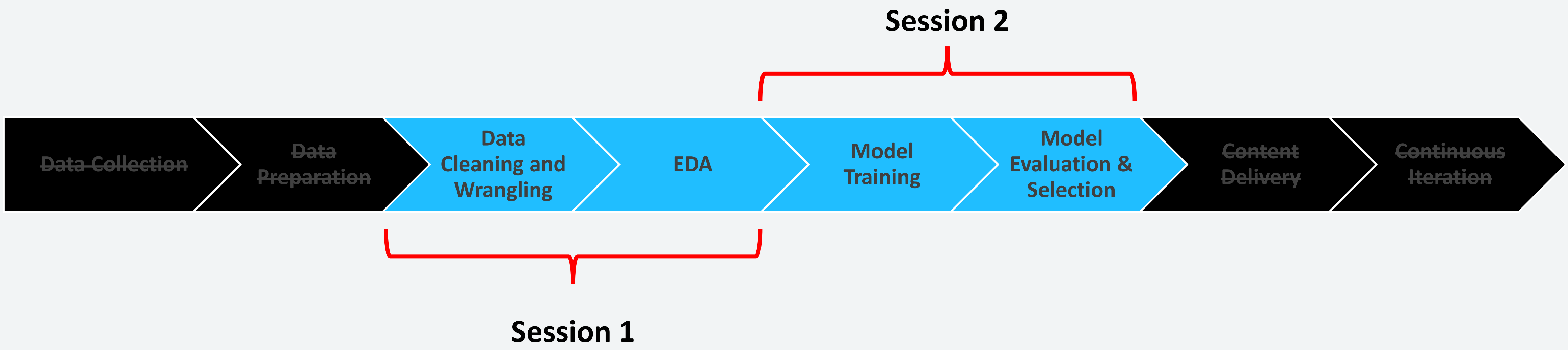
Model Evaluation & Selection

Content Delivery

Continuous Iteration

The Kaggle Process

“Turn **data** into **prediction**”



Session 2

~~Data Collection~~

~~Data Preparation~~

Data Cleaning and Wrangling

EDA

Model Training

Model Evaluation & Selection

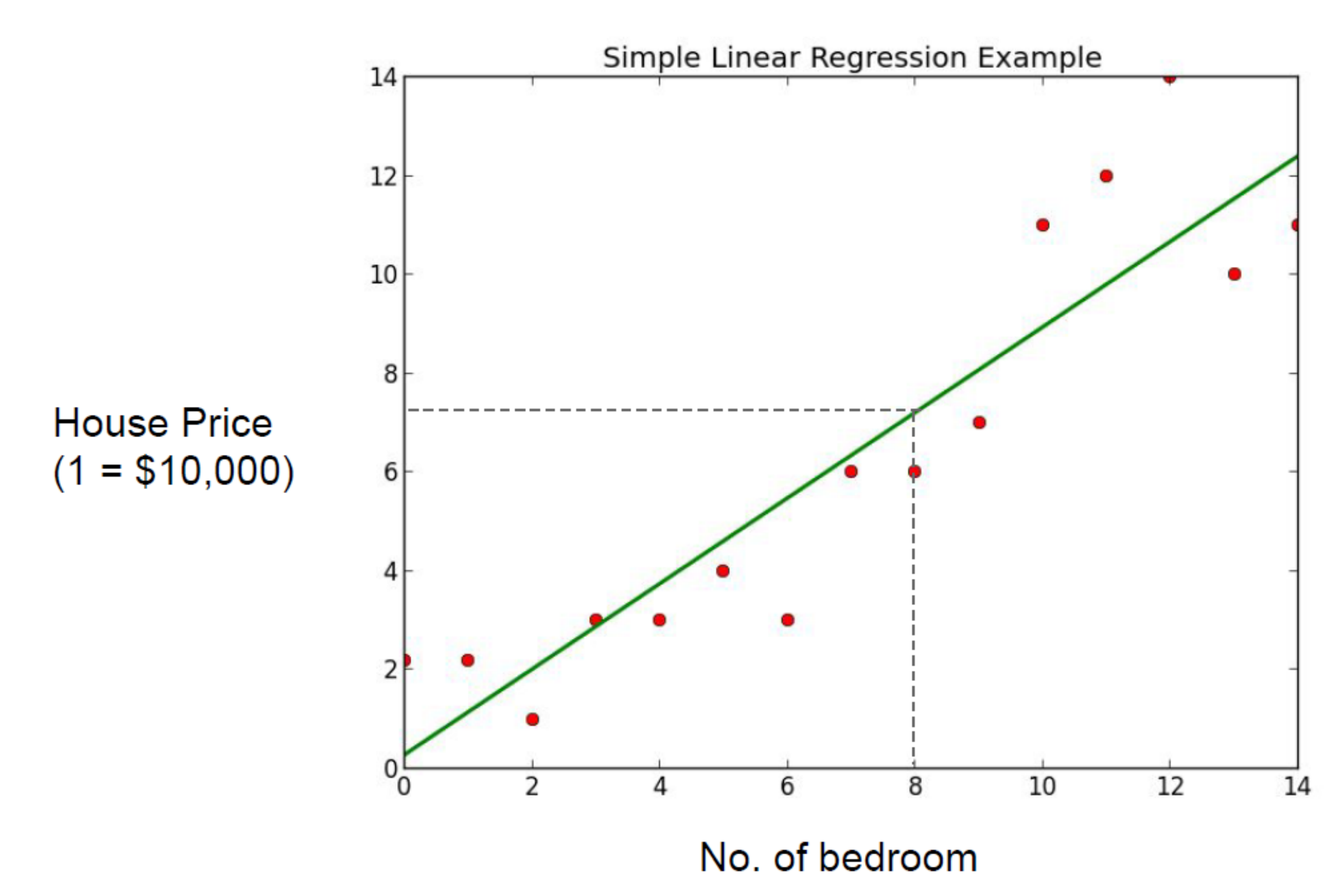
~~Content Delivery~~

~~Continuous Iteration~~

Session 1

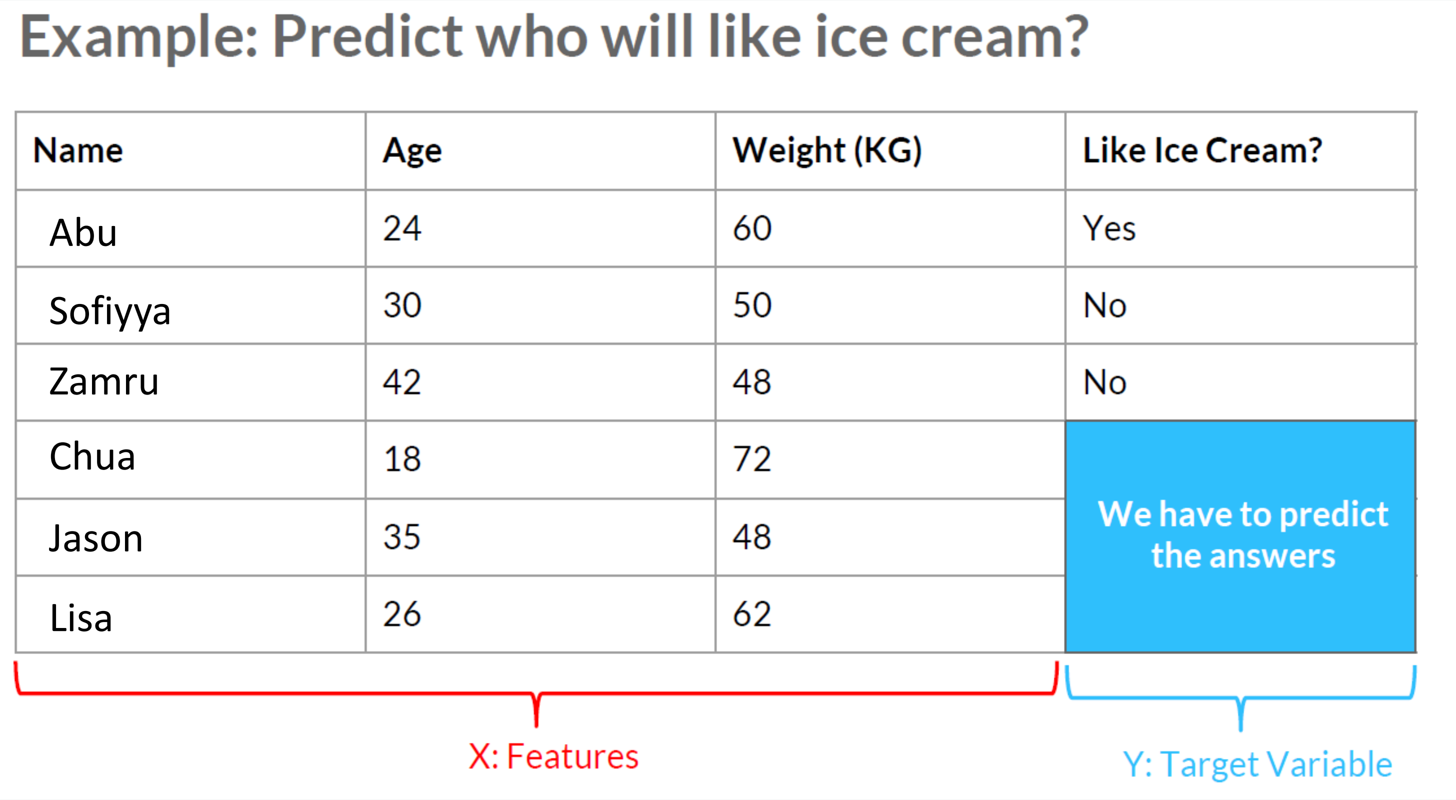
Session 1

Supervised Learning



Regression Problem!

Supervised Learning



Example: Predict who will like ice cream?

Name	Age	Weight (KG)	Like Ice Cream?
Abu	24	60	Yes
Sofiyya	30	50	No
Zamru	42	48	No
Chua	18	72	We have to predict the answers
Jason	35	48	
Lisa	26	62	

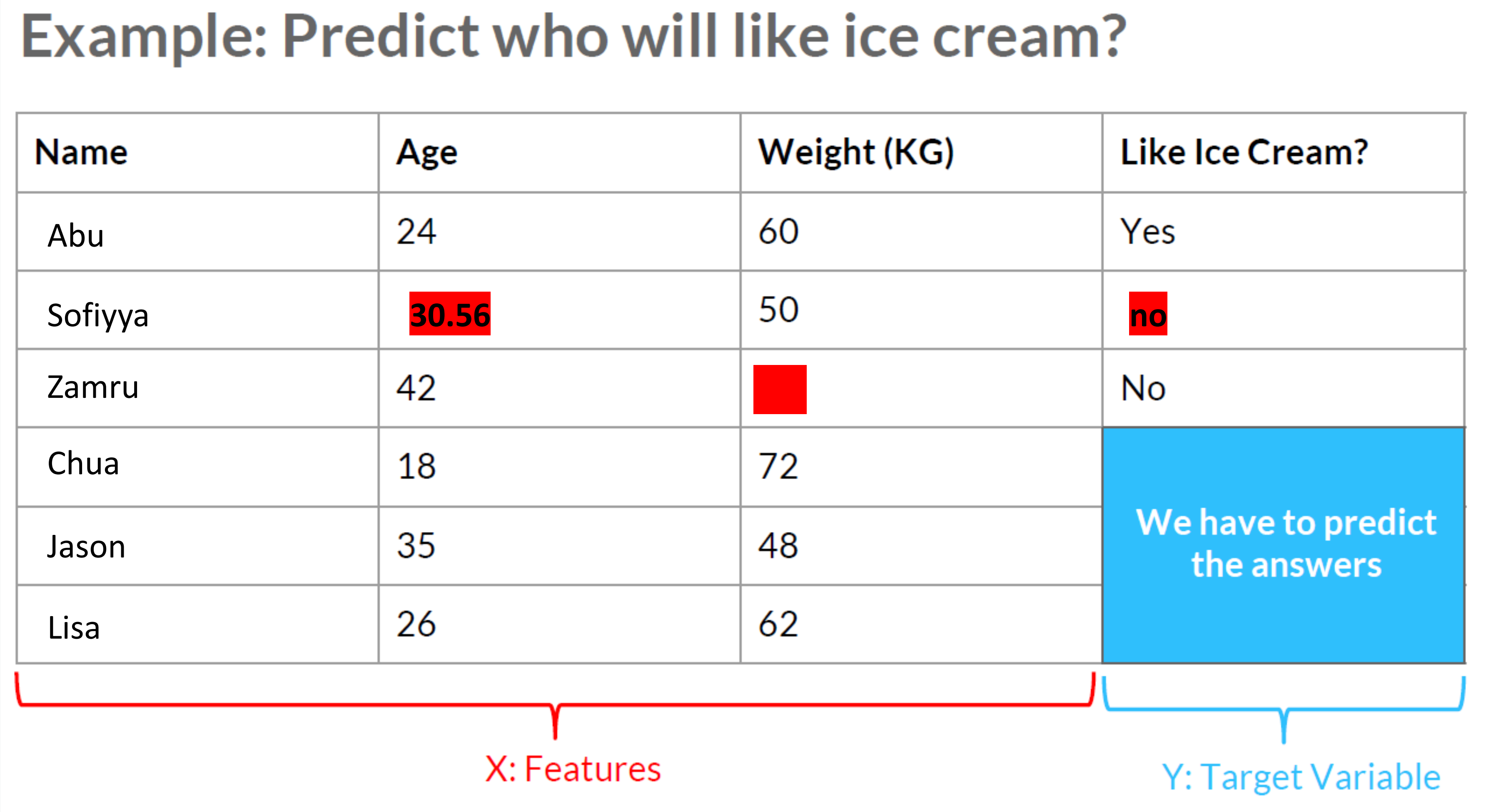
X: Features

Y: Target Variable

Classification Problem!

Introduction to Data Wrangling

The process of converting raw data into a usable form.



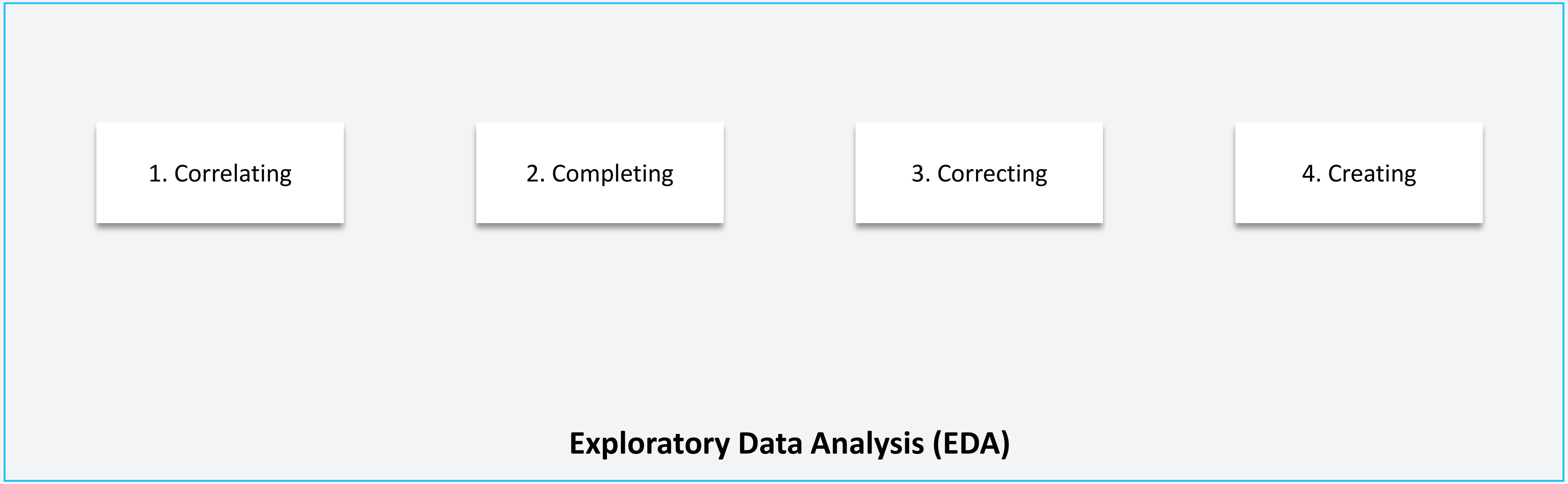
Example: Predict who will like ice cream?

Name	Age	Weight (KG)	Like Ice Cream?
Abu	24	60	Yes
Sofiyya	30.56	50	no
Zamru	42	🟥	No
Chua	18	72	We have to predict the answers
Jason	35	48	
Lisa	26	62	

X: Features

Y: Target Variable

Data Wrangling – Workflow Goals



1. Correlating

2. Completing

3. Correcting

4. Creating

Exploratory Data Analysis (EDA)

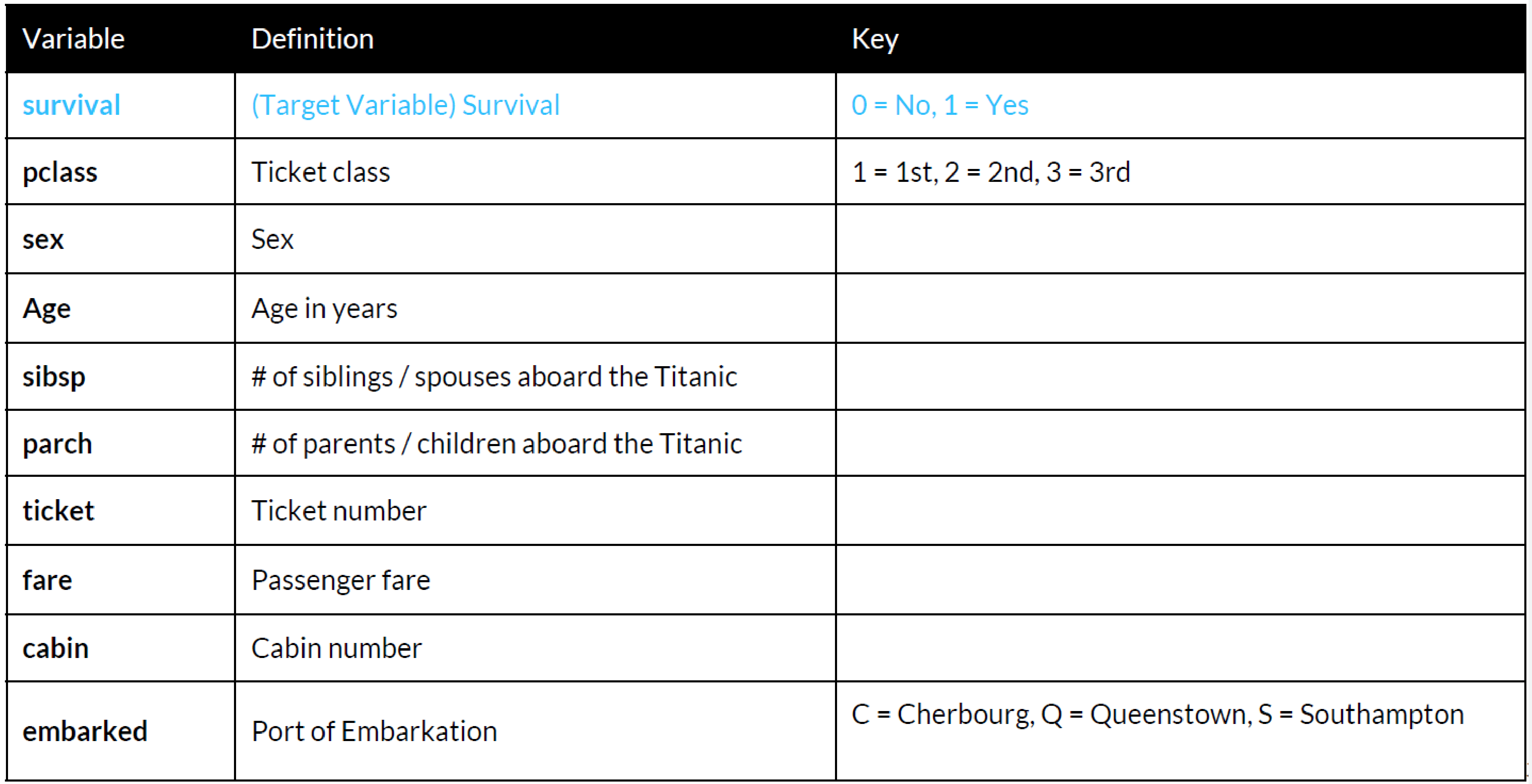
The Titanic Survival Dataset

Question: Who will **survive** the Titanic?

Context:

- On 15th April 1912, the Titanic sank after colliding with an iceberg, killing **1502 out of 2224** passengers and crews. (*32.46% survival rate*)
- There were lack of lifeboats for the passengers and crews.
- Some groups of people are more likely to survive than others, such as women, children, and the upper-class.

Titanic: Data Dictionary

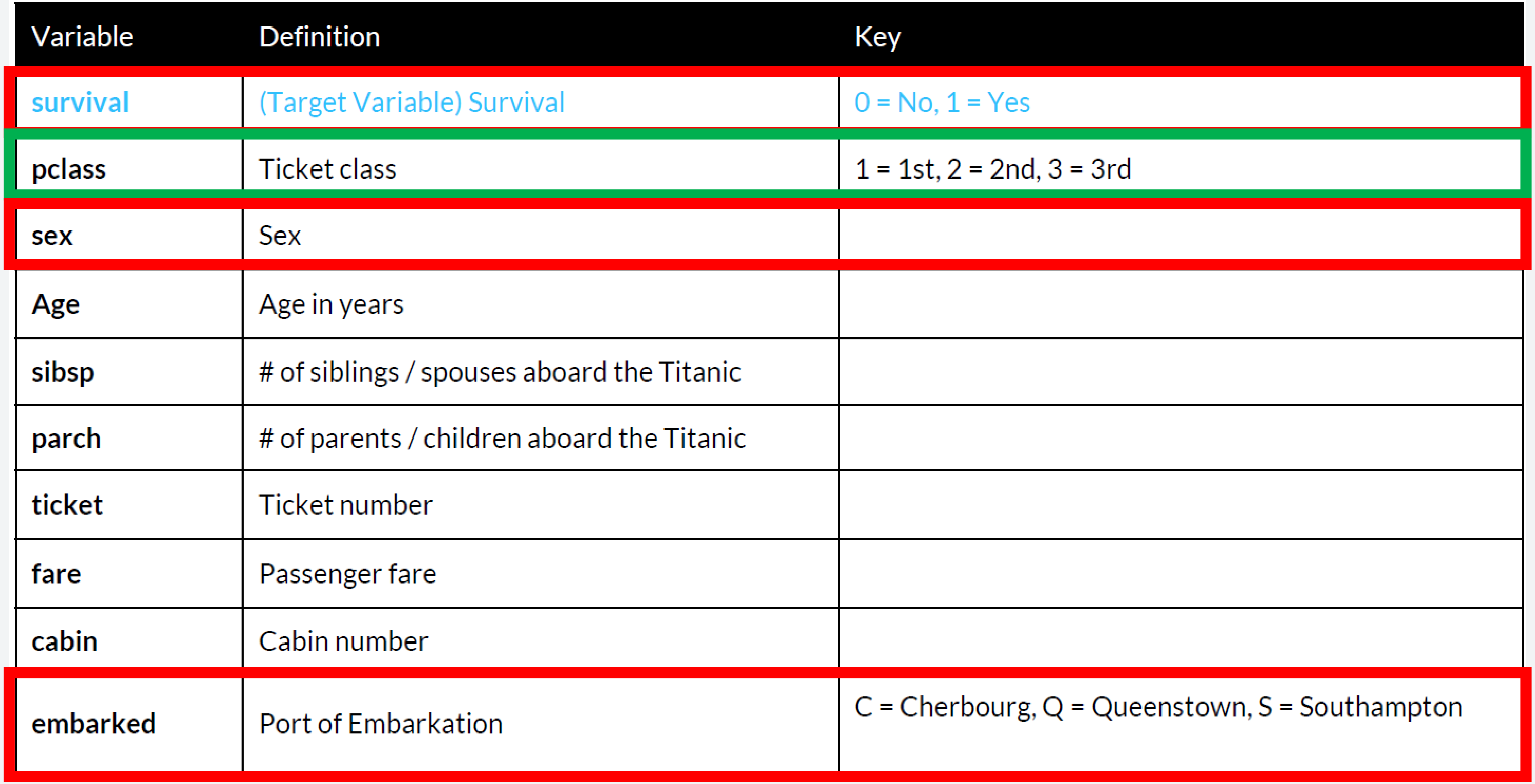


Variable	Definition	Key
survival	(Target Variable) Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Which features are categorical?

Which features are numerical?

Categorical Data

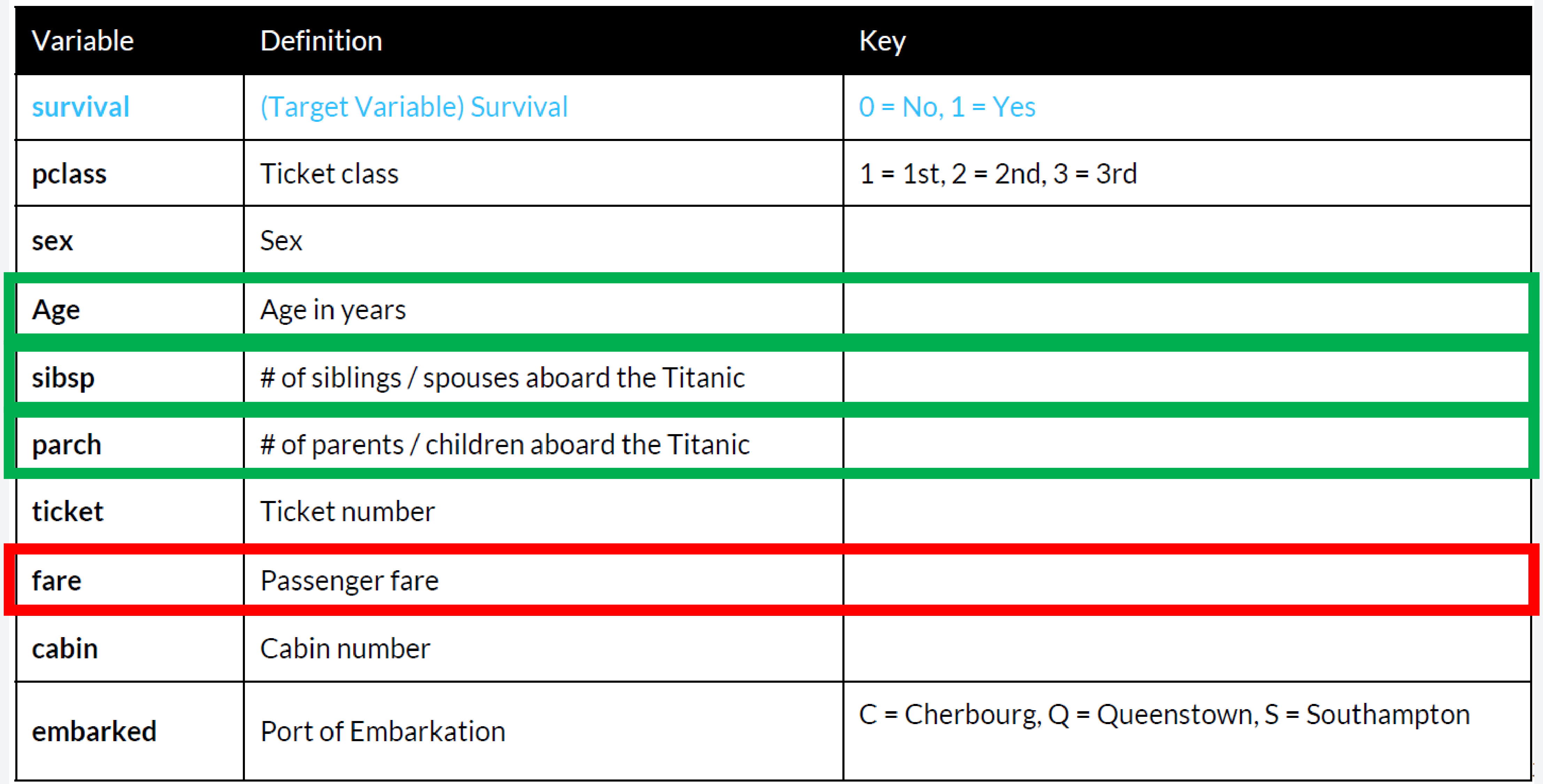


Variable	Definition	Key
survival	(Target Variable) Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

categorical

ordinal

Numerical Data



Variable	Definition	Key
survival	(Target Variable) Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

discrete

continuous

Coding Time!

QnA – Slido Link

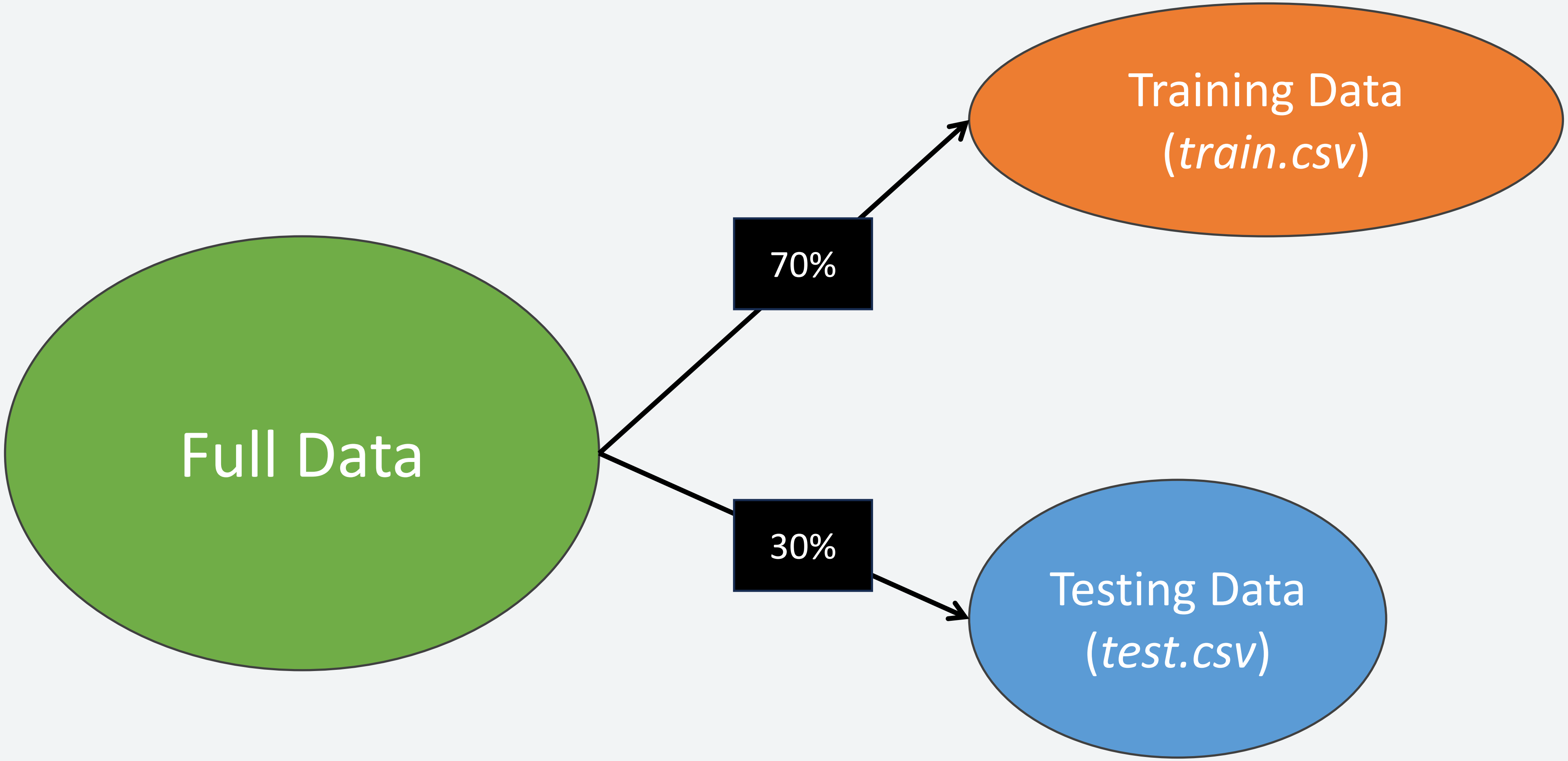
https://qrco.de/beXIuo

Lunch Break

See you guys at 1:10 pm!

Session 2

Data needs to be split



Full Data

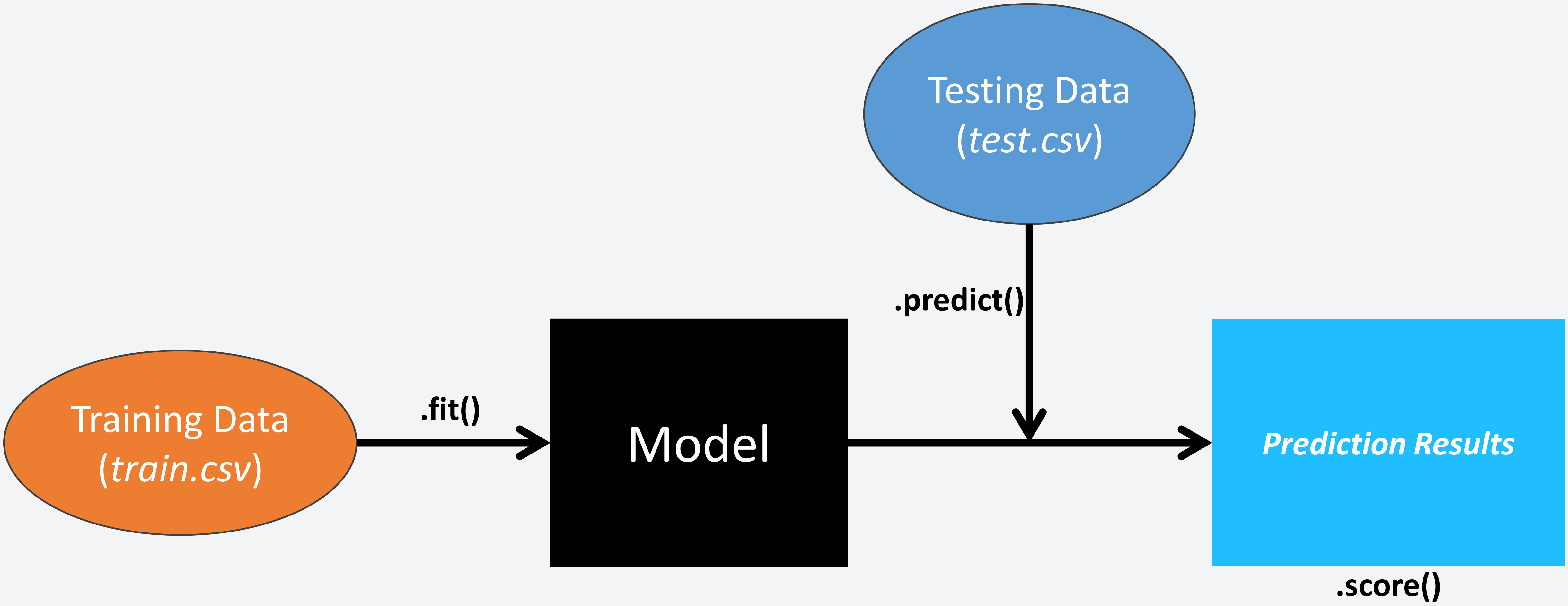
70%

Training Data
(*train.csv*)

30%

Testing Data
(*test.csv*)

Building a Predictive Model



Testing Data (*test.csv*)

.predict()

Training Data (*train.csv*)

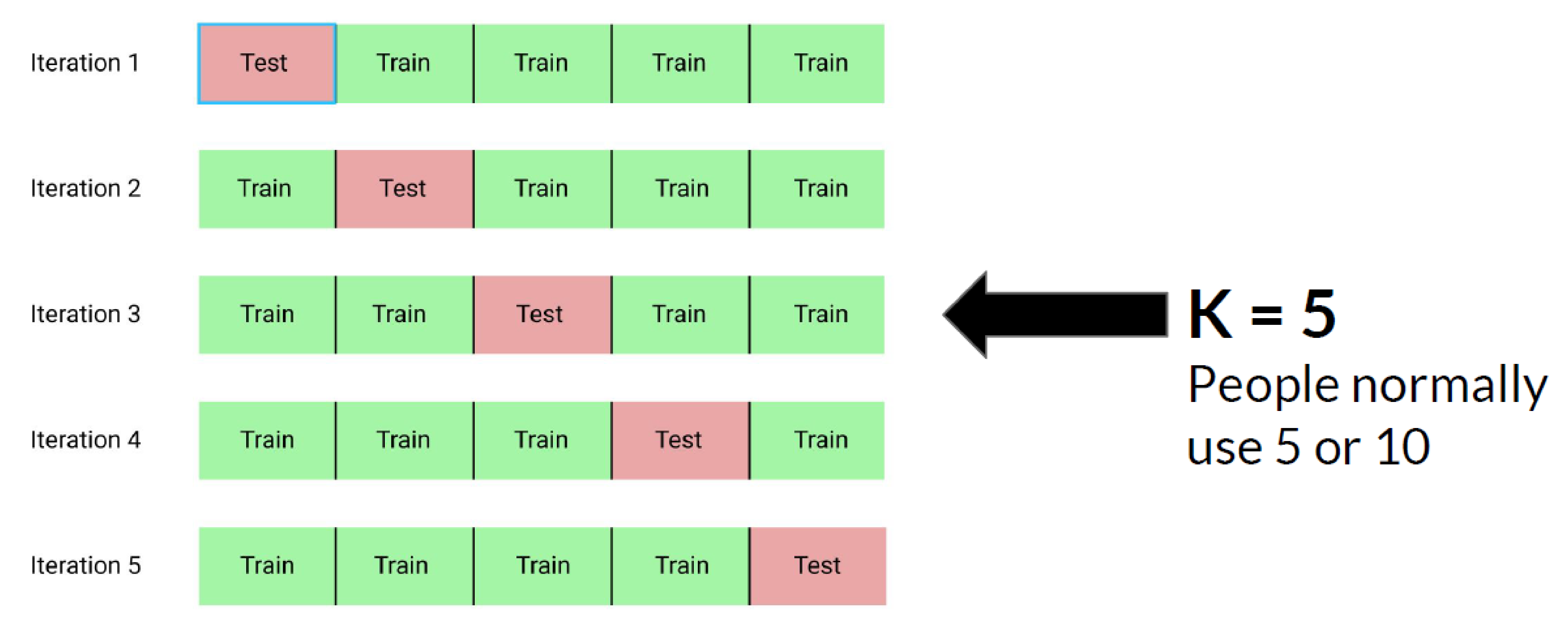
.fit()

Model

Prediction Results

.score()

K-Fold Cross Validation



Iteration 1	Test	Train	Train	Train	Train
Iteration 2	Train	Test	Train	Train	Train
Iteration 3	Train	Train	Test	Train	Train
Iteration 4	Train	Train	Train	Test	Train
Iteration 5	Train	Train	Train	Train	Test

K = 5

People normally use 5 or 10

Coding Time!

QnA – Slido Link

https://qrco.de/beXIuo

Thank you!