Data Science CSCI 2022

E+ Exchange

Erasmus mobility statistics 2014 To 2019

Abdelrahman Alarqan
120181168

What is Erasmus+?

Erasmus+ is the EU's programme to support education, training, youth and sport in Europe.

It supports priorities and activities set out in the European Education Area, Digital Education Action Plan and the European Skills Agenda. The programme also

- Supports the European Pillar of Social Rights
- Implements the EU Youth Strategy 2019-2027
- Develops the European dimension in sport

Erasmus+ offers mobility and cooperation opportunities in:

- higher education
- vocational education and training
- school education (including early childhood education and care)
- adult education
- youth
- sport

About The Dataset

Erasmus mobility statistics 2014_2019

This dataset contains 10,000 of data rows and 24 columns (feature) for Erasmus+ mobility for students and staff from 2014 to 2019.

Metadata

- Data Source: **Kaggle**
- Data Owner: **Lucas Arbabyazd**

Row Example 01

Columns	Values
Project Reference	2017-1-PT01-KA103-035561
Academic Year	2018-2019
Mobility Start Month	2018-02
Mobility End Month	2018-06
Mobility Duration	152
Activity (mob)	Student mobility for studies between Programme Countries
Field of Education	Business and administration, not further defined
Participant Nationality	PT

Row Example 02

Education Level	ISCED-6 - First cycle / Bachelorâ€™s or equivalent level (EQF-6)
Participant Gender	Male
Participant Profile	Learner
Special Needs	No
Fewer Opportunities	No
Group Leader	No
Participant Age	21
Sending Country Code	UK

Row Example 03

Sending City	FARO
Sending Organization	UNIVERSIDADE DO ALGARVE
Sending Organisation Erasmus Code	P FARO02
Receiving City	POZNAN
Receiving Country Code	PL
Receiving Organization	UNIWERSYTET EKONOMICZNY W POZNANIU
Receiving Organisation Erasmus Code	PL POZNAN03
Participants	4

Tools And Techniques

Tool	Used For
Power BI	Data visualization and analysis
Python	Data preparation and transformation
Knime	Predictive and Descriptive Analytics
Google Slides	Presenting and reporting

Data Exploration

- Number of rows: **10,000**
- Number of Columns: **24** (Features)
- Number of missing values: **0** (There is no missing)
- Duplicated Values: **0** (There is no duplicated)

10.00K

Count of Project Reference

Data Exploration (Continuous Data)

Start with continuous data

1- Participant Age

13	73	25
Min of Participant Age	Max of Participant Age	Average of Participant Age

Data Exploration (Continuous Data)

2- Mobility Duration

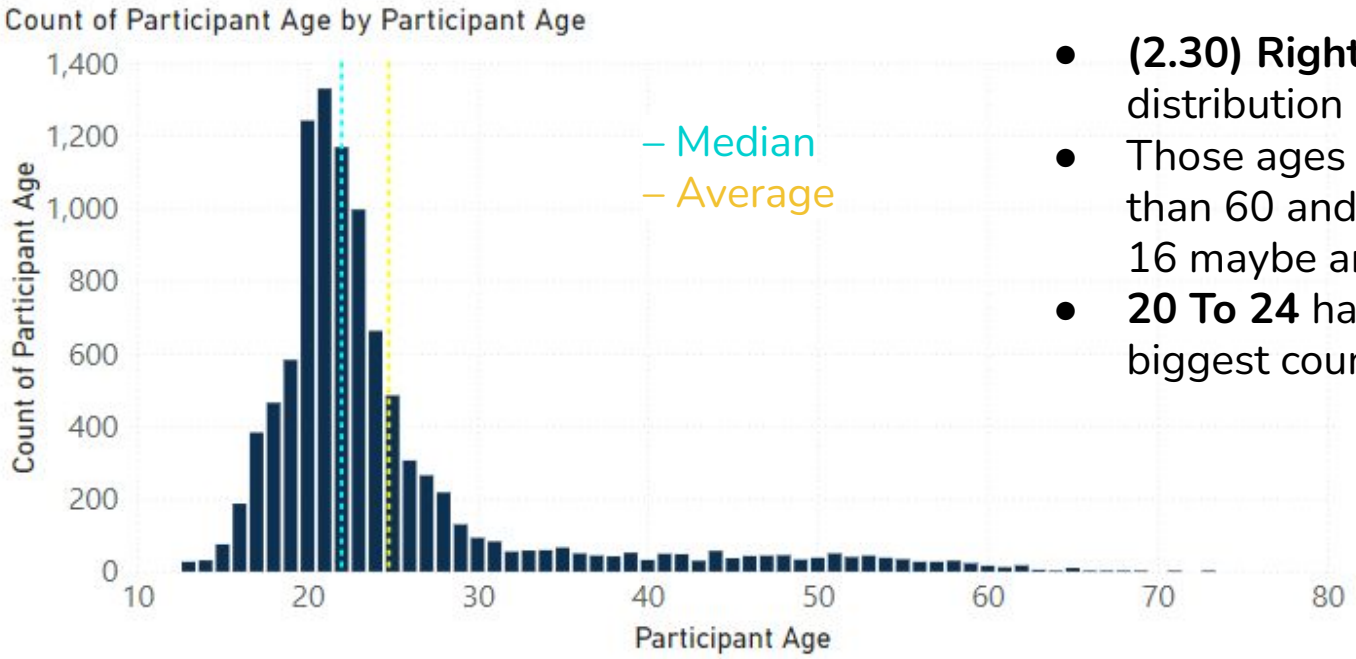
1	541	89
Min of Mobility Duration	Max of Mobility Duration	Average of Mobility Duration

3- Participants

1	13
Min of Participants	Max of Participants

Data Exploration (Continuous Data)

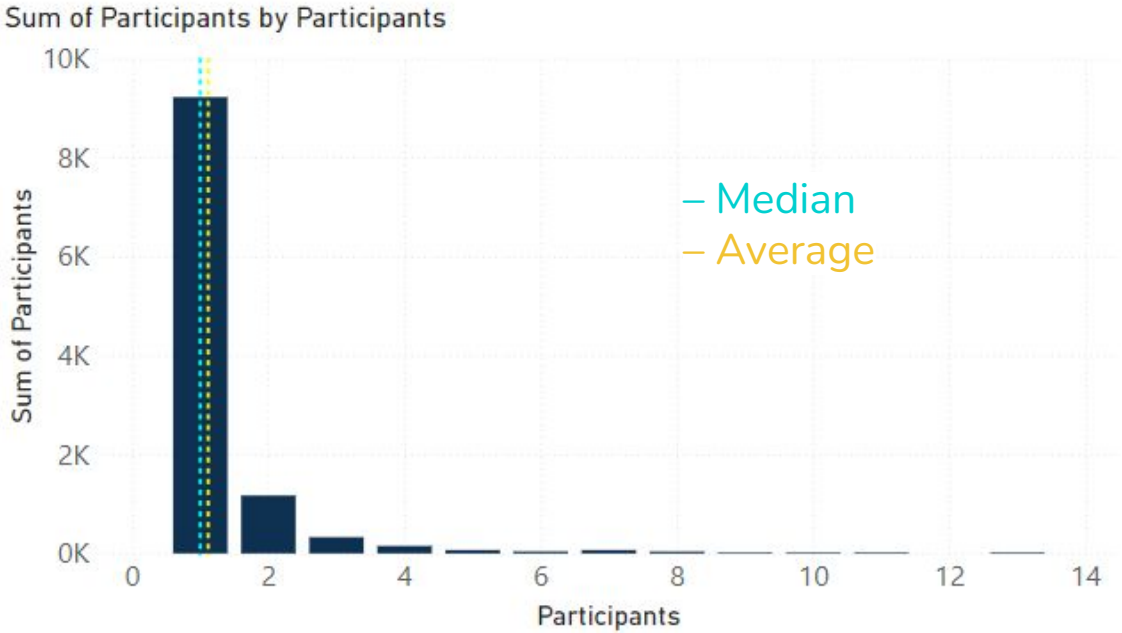
Data Distribution of Participant Age



- **(2.30) Right-skewed** distribution
- Those ages are more than 60 and less than 16 maybe an **outliers**
- **20 To 24** have the biggest count

Data Exploration (Continuous Data)

Data Distribution of Participants



9225 of 10K Projects is one participant, more than two participant are **outliers**

Data Exploration (Correlations)

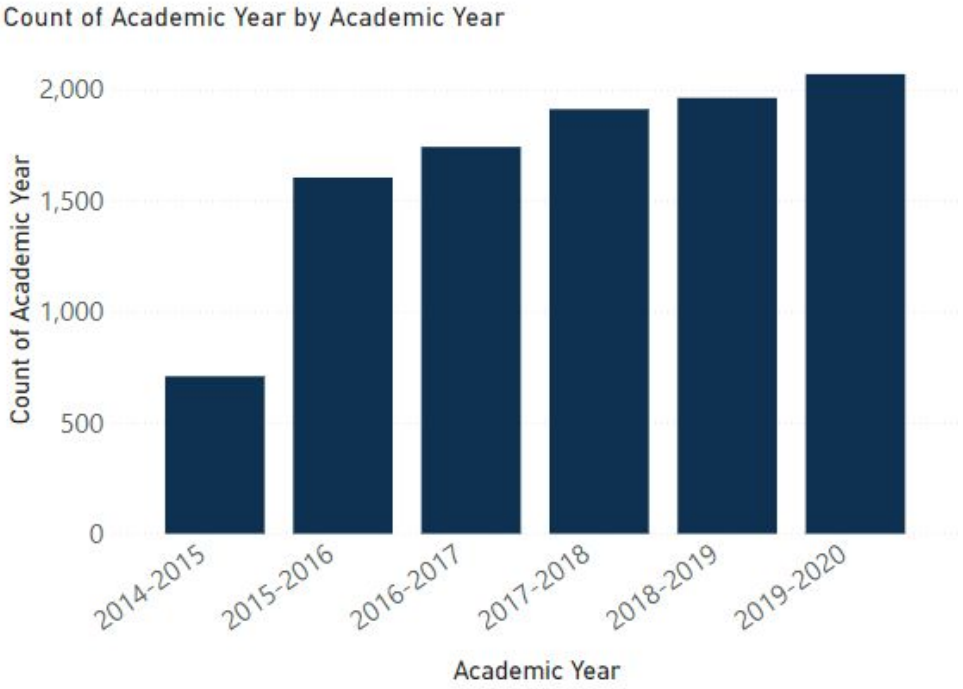
Relationship between continuous data

	Mobility Duration	Participant Age	Participants
Mobility Duration	1.000000	-0.251680	-0.122827
Participant Age	-0.251680	1.000000	-0.109811
Participants	-0.122827	-0.109811	1.000000

There is **no relationship** between the continuous data

Data Exploration (Categorical Data)

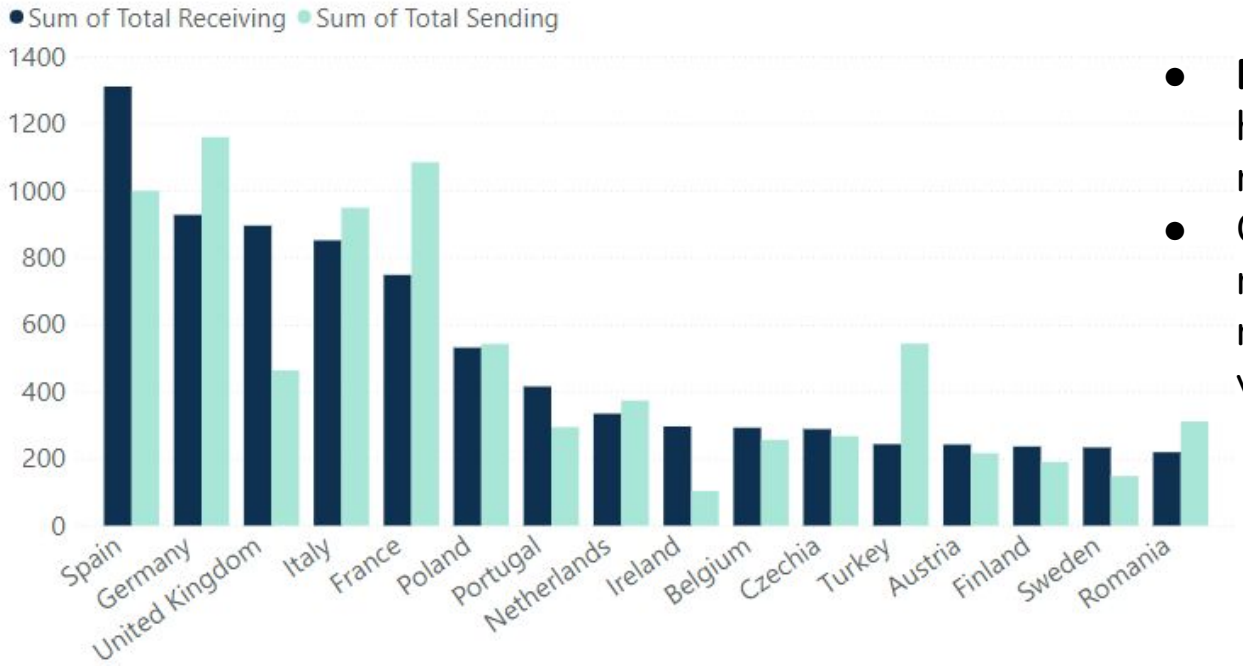
Data Distribution of Academic Year



- The count of participant is **increased** year by year from 2014 to 2020
- From 2014 to 2016 the count has been increased more than the half of participant

Data Exploration (Categorical Data)

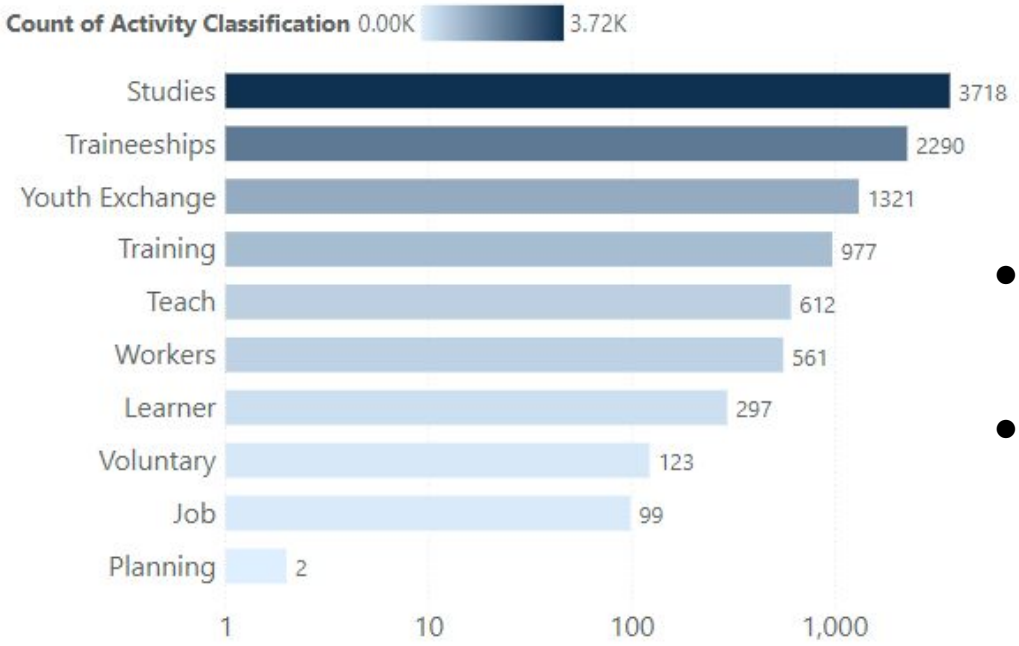
Difference between Sending and Receiving Countries



- **Poland and Italy** have a close numbers
- Germany sends more than it receives, and vice versa with Spain

Data Exploration (Categorical Data)

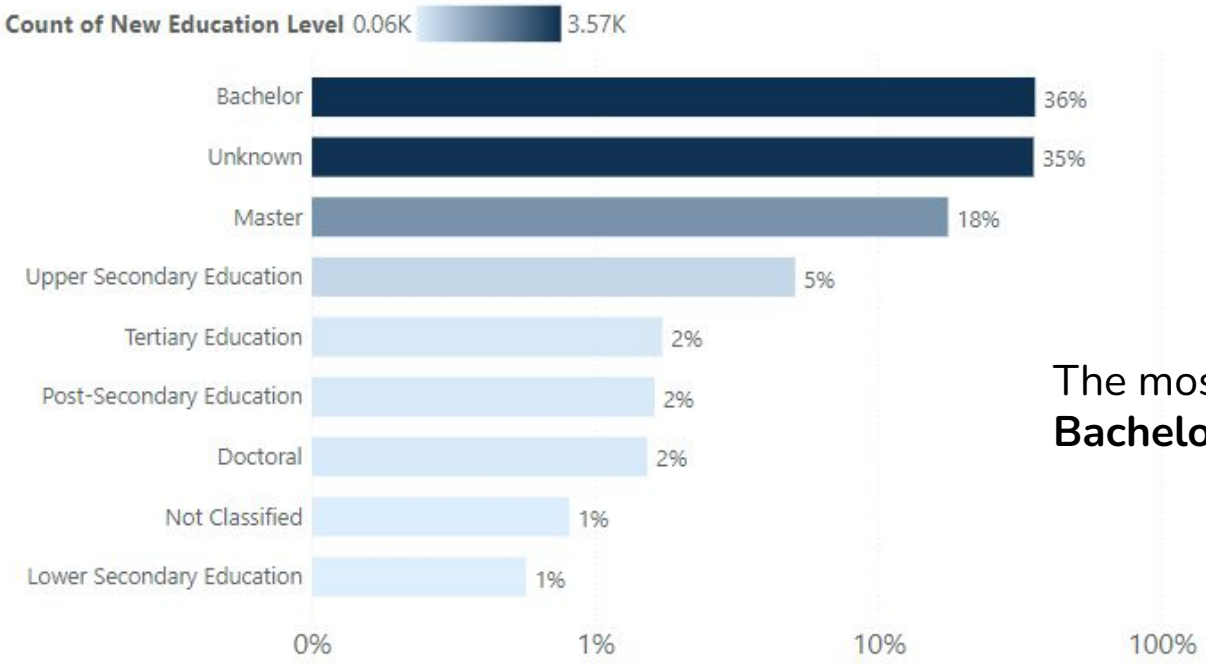
Data Distribution of Activity Classification



- As we see on this visual, most of activities are **Studies and Traineeships**.
- Planning has two activities only, that could be an **outliers**

Data Exploration (Categorical Data)

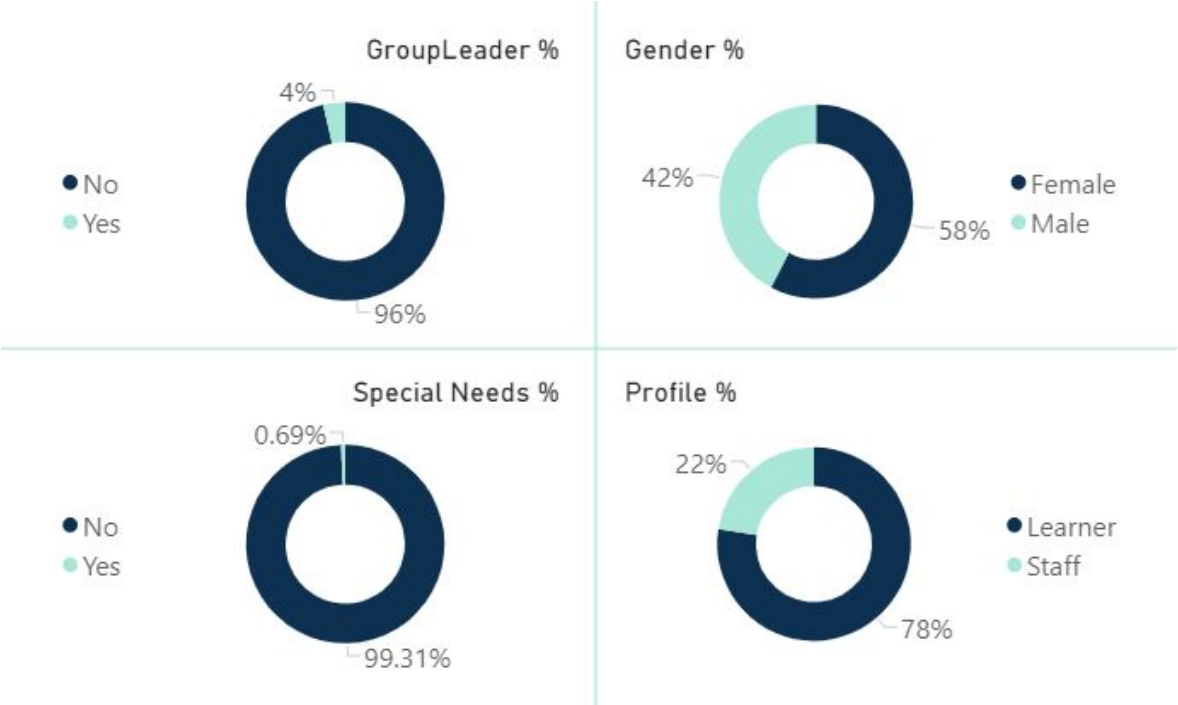
Data Distribution of Education Level



The most of exchanges is **Bachelor Level (36%)**

Data Exploration (Categorical Data)

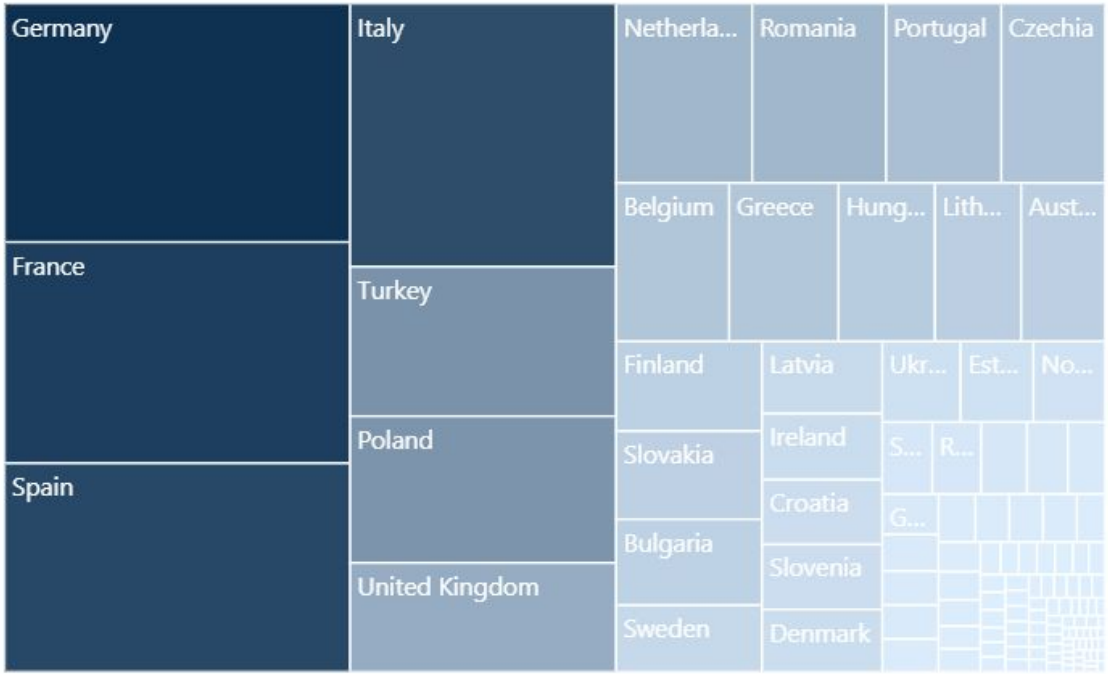
Group Leader | Gender | Special Needs | Profile



- Female is **more than** male
- 4% are **group leaders**
- Less than 1% **Special Needs**
- More students, more **Learners**

Data Exploration (Categorical Data)

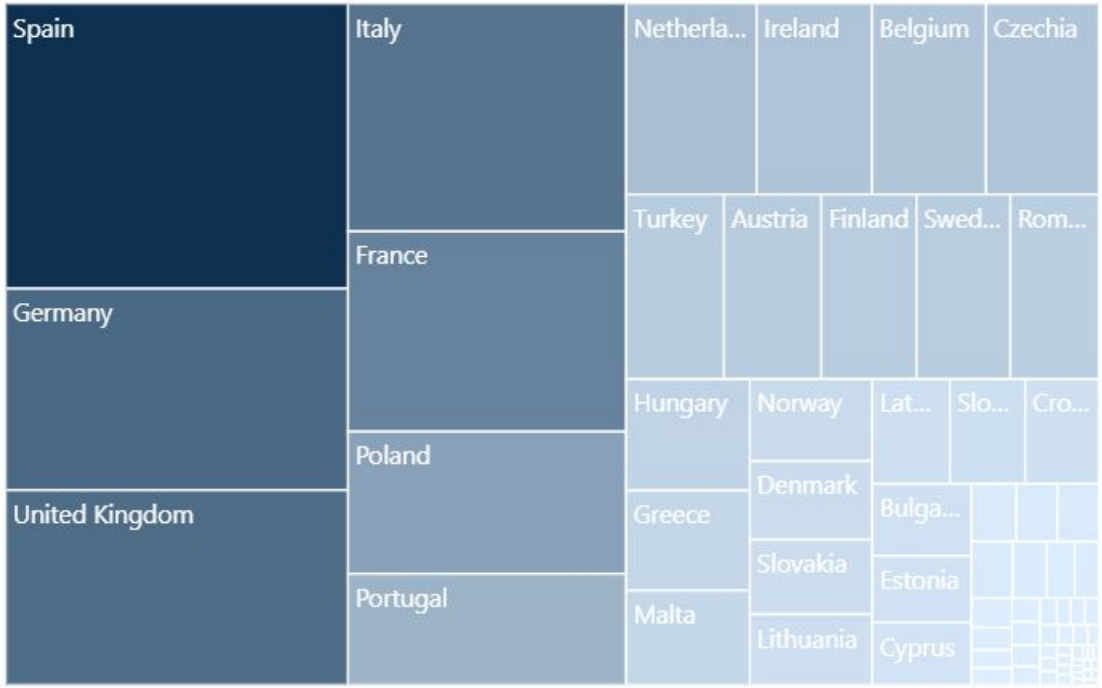
Data Distribution of Sending Countries



- Germany has the **big count** of participants
- France, Spain and Italy are **too close**
- **Seven** countries are the half of 121 Countries

Data Exploration (Categorical Data)

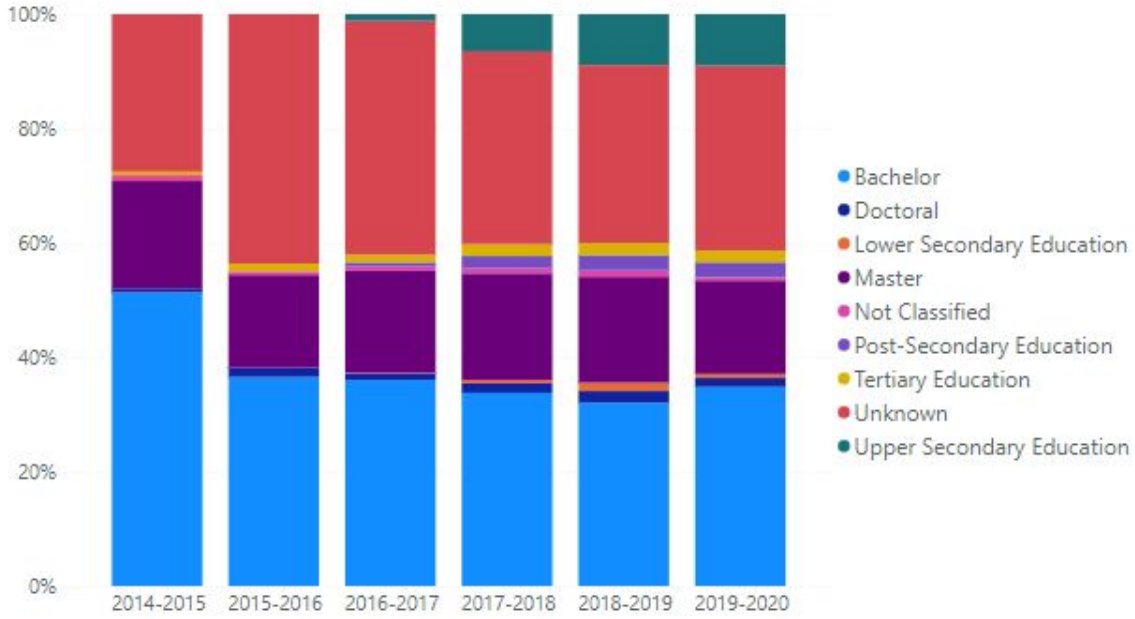
Data Distribution of Receiving Countries



- Spain has the **big count** of Receiving participants
- Germany and UK are **too close**
- **Seven** countries are the half of 121 Countries

Data Exploration (Categorical Data)

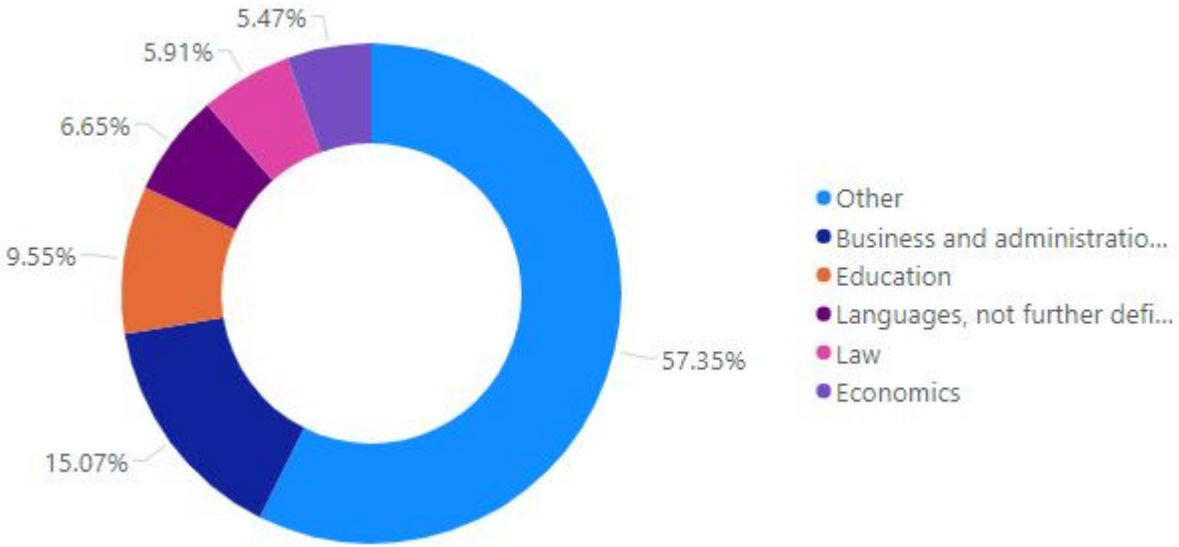
Data Distribution of Levels with Academic Year



As we see here the distribution of levels is **increased** from 2016 To 2020 and starts in 2017

Data Exploration (Categorical Data)

The most popular field of studies



Business and administration is the most **popular** field

Data Exploration (Categorical Data)

Sending only	Receiving only	
Haiti	Algeria	Kazakhstan
Indonesia	Australia	Madagascar
Kenya	Bhutan	Nepal
Korea, Republic of	Burkina Faso	Pakistan
Mozambique	Cambodia	Singapore
South Africa	Cuba	Taiwan
Sri Lanka	Iraq	Uganda
	Uruguay	Uzbekistan

Number of Receiving is **more than** Sending

Data Preparation

The dataset needs a lot of processing and transformation processes, and I will explain it in the following slides through steps and points

- Cleaning
- Transformation
- Extraction
- ETL
- Processing
- Type conversion
- ETC...

Data Preparation

- Number of rows: **10,000**
- Number of Columns: **24** (Feature)
- Number of missing values: **0** (There is no missing)
- Duplicated Values: **0** (There is no duplicated)
- There are a lot of Incorrect values, need to cleaning

Data Preparation

Columns need correct data type

- String → Integer (Mobility Duration, Participant Age, Participants)

Handle Incorrect values such as

- “-” → Nan
- “??? Unknown????” → Unknown or Other
- Handling missing value using pandas ffill()

Drop some of columns, those give the same value

- Project reference
- Country code
- ETC

Data Transformation 01

New Columns added to make clear visualization

↓→	Main Column	New Column Added
Column Name	**Activity (mob)**	**Activity Classification**
Example	Staff training abroad	Training
Column Name	**Participant Nationality**	**New Participant Nationality**
Example	UK	United Kingdom
Column Name	**Education Level**	**New Education Level**
Example	ISCED-6 - First cycle / Bachelor’s or equivalent level (EQF-6)	Bachelor

Data Transformation 02

New Columns added to make clear visualization

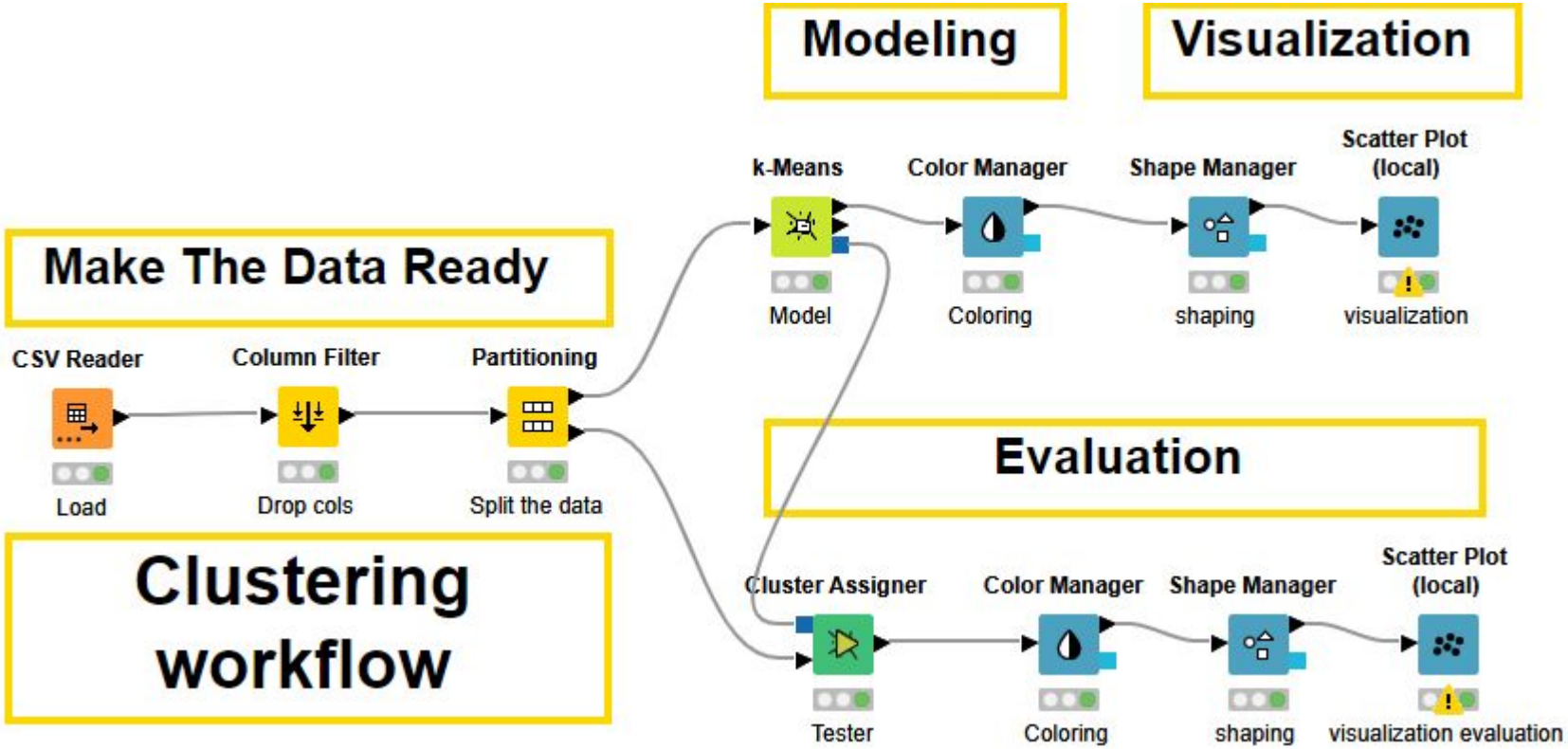
↓→	Main Column	New Column Added
Column Name	**Sending Country Code**	**Sending Country Name**
Example	PS	Palestine
Column Name	**Receiving Country Code**	**Receiving Country Name**
Example	US	United States

Data Preparation

Splitting The data

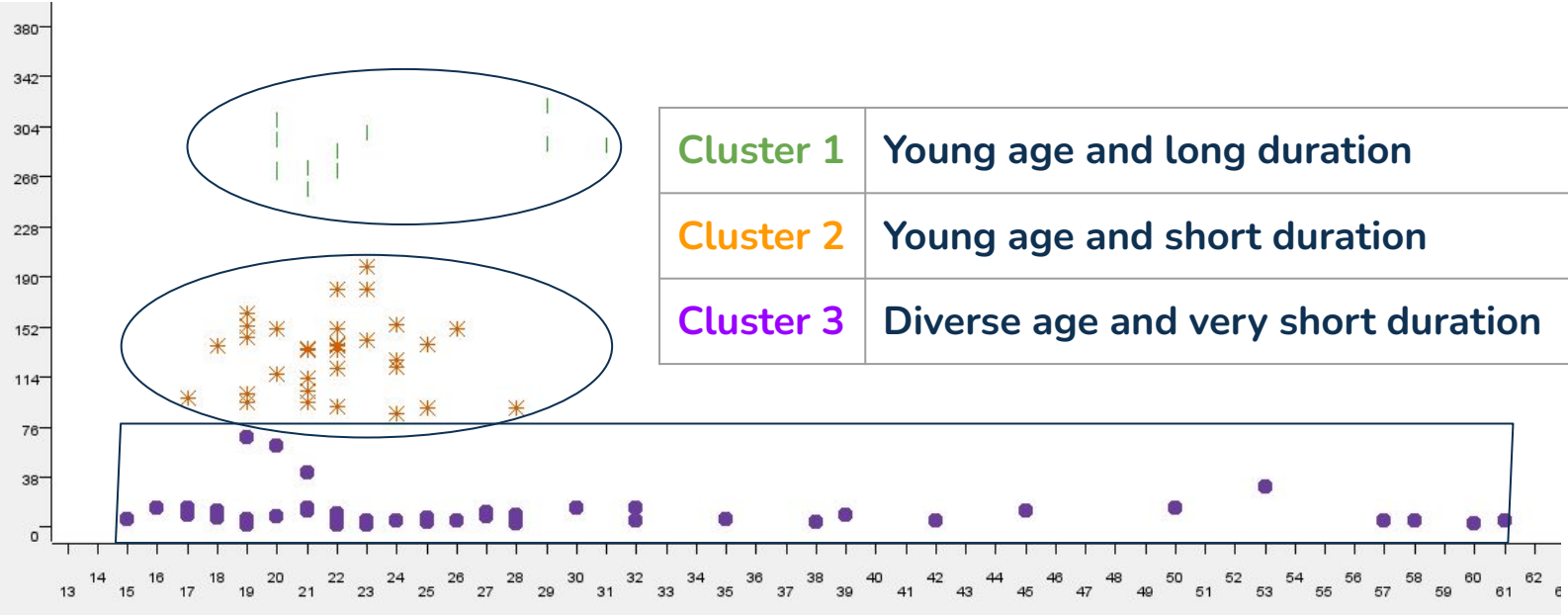
Parts	Number Of Records	Percentage %
Train	8000	80%
Test	2000	20%

Descriptive Analytics (Clustering)



Descriptive Analytics (Clustering)

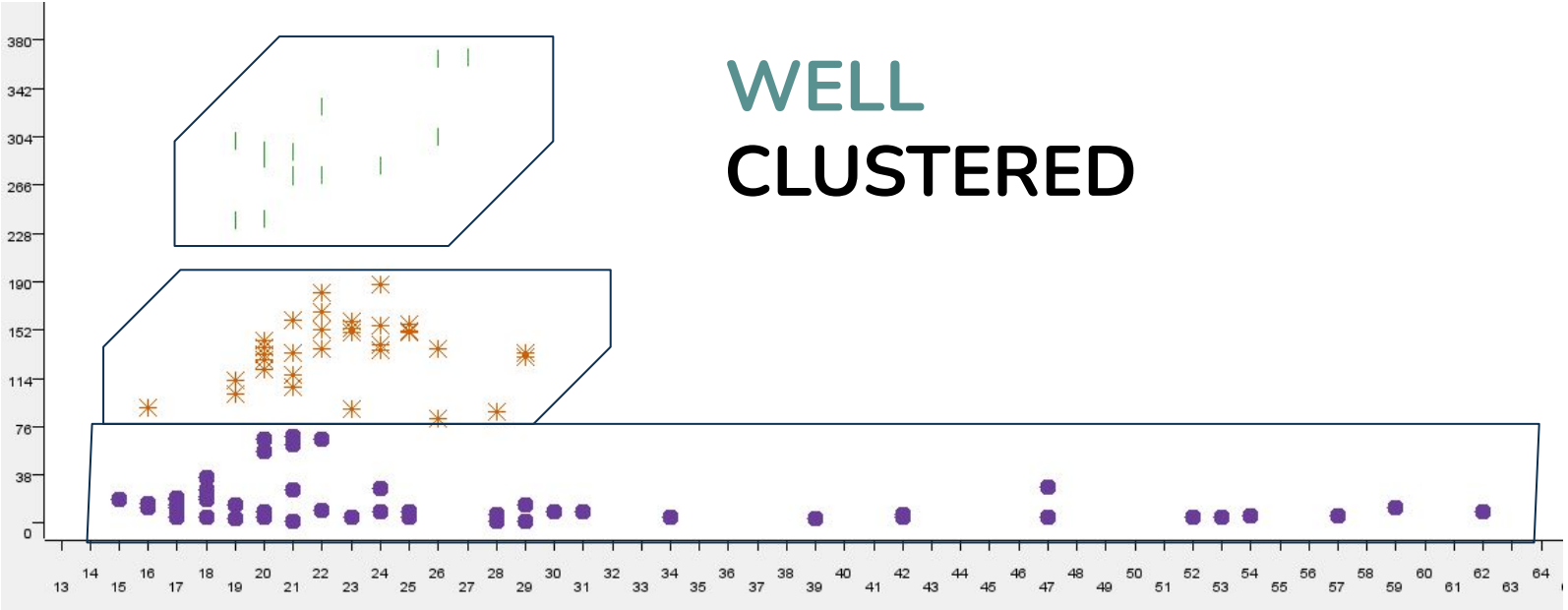
Clustering The Dataset (Clusters Training Result)



Cluster 1	Young age and long duration
Cluster 2	Young age and short duration
Cluster 3	Diverse age and very short duration

Descriptive Analytics (Clustering)

Clustering The Dataset (Clusters Evaluation Test Result)



WELL
CLUSTERED

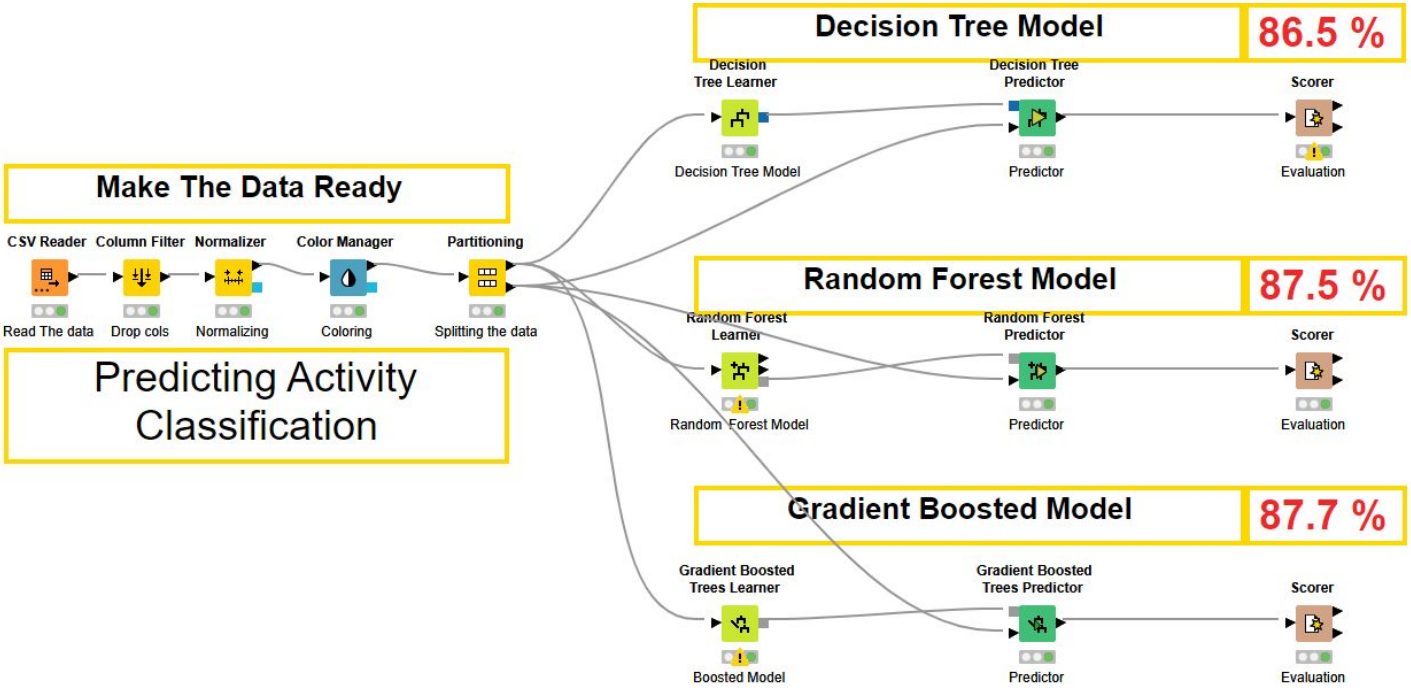
Predictive Analytics

Predict The Future(Activity Classification Prediction)

- Studies
- Traineeship
- Training
- Teaching
- Youth Exchange
- Voluntary
- Learner
- Planing
- Job

Predictive Analytics

Predict The Future(Activity Classification Prediction)



Predictive Analytics

Predict The Future(Activity Classification Prediction)

Model	Accuracy %
Decision Tree Model	86.6 %
Random Forest Model	87.5 %
Gradient Boosted Model	87.7 %

Conclusions

- Most of the exchanges target people between the ages of **20 to 25.**
- The educational level is **increasing distribution** every year.
- The exchanges focus on educational activity the most.
- Business and administration is the most **popular** field.
- **(Germany, United kingdoms, Italy, Spain, France)** are the most countries who send and receive exchanges.


THANK YOU