Private Higher School
of Engineering and
Technology

Academic Project Report

« Tourism »

Realized by :

Khalil Baraketi

Mariem Mannai

Molka Kawech

Naziha Ksouri

Youssef Tebourbi

Supervised by :

Sarra Sebai

Ines Chouchane

Ines Mhaya

Ghassen Fodha

Jihen Rezgui

4ERP BI 3

15/05/2023

Thanks

I would like to extend my heartfelt gratitude to the individuals who have contributed to the completion of this work. I would like to express my deep appreciation to Mrs. Ines Chouchen, Ines Mhaya, Jihen rezgui, and Mr. Ghassen Fodha for their assistance and guidance throughout this project. I hope to meet their expectations and convey my profound gratitude through this work.

I would also like to thank Mrs. Sarra Sebai for her exceptional supervision, constructive criticism, and invaluable advice, which have enabled me to achieve my initial goals. I am grateful for her kindness and support.

Furthermore, I would like to express my thanks to the members of the jury for accepting the responsibility of assessing, evaluating, and enhancing this work. Your presence is an honor to me, and I hold you in high regard and deep appreciation.

I would also like to acknowledge and appreciate all my teachers who have contributed to my academic journey. I hold them in the highest respect and extend my infinite gratitude to them.

Table of Contents

General Introduction 1

Chapter 1: Project Context 2

Introduction 3

I. Company presentation 3

II. Context of project 3

1. Study of the existing 3

2. Problematic 5

3. Adopted Solution 5

4. Methodology of work 6

4.1. Classic methodology 6

4.2. advanced methodology 7

Conclusion 8

Chapter 2: Data Modeling 9

Introduction 10

I. Identification of functional and non-functional requirements 10

1. Functional requirements 10

2. Non- Functional requirements 10

II. Modeling techniques 11

1. Star Schema 11

1.1. Advantages 11

1.2. Disadvantages 11

2. Snowflake Schema 12

2.1. Advantages 12

2.2. Disadvantages 12

3. Constellation Schema 12

3.1. Advantages 12

3.2. Disadvantages 12

III. KPI Identification 14

IV. Dimensions Identification 14

Conclusion 14

Chapter 3: Data Integration 15

Introduction 16

I. Working tool: 16

1. Database management tool 16
2. Data integration: Talend 16
3. Data Implementation 18
3.1. Internal Data 18
3.1.1. Trips dimension 18
3.1.2. Date dimension 22
3.1.3. Accommodation dimension (integration of internal and external data) 24
3.1.4. Country dimension 26
3.1.5. Type Arrivals dimension 27
3.1.6. Main purpose dimension 28
3.1.7. Expenditure dimension (internal and external data) 29
3.2. External data 30
3.2.1. Season dimension 30
3.2.2. Landscape dimension 30
3.2.3. Type transport dimension 31
3.2.4. Profil dimension 31
3.2.5. Tourism dimension 32
4. Fact table 32
Conclusion 36
Chapter 4: Data Mining 37
Introduction 38
I. Business Understanding 38
II. Data Understanding 38
1. Descriptive 38
1.1. Profiling 39
1.1.1. Algorithm used 39
1.2. Attractivity 40
1.2.1. Algorithm used 40
1.3. Performance 41
1.3.1. Algorithm used 42
2. Predictive 44
2.1. Profiling 44
2.1.1. Algorithm used 44
2.2. Total Arrivals 45
2.2.1. Algorithm used 45
Conclusion 48

Chapter 5: Data Viz and Realization 49
Introduction 50
I. DataViz 50
1. Working tool: Power BI 50
2. Dashboards 50
2.1. Overview Interface 50
2.2. Comparison interface 51
2.3. Inbound interface 52
2.4. Outbound interface 52
2.5. Domestic interface 53
2.6. Accommodation interface 53
2.7. Profil interface 54
II. Realization 54
1. Working tools 54
1.1. Back-end tool 54
1.2. Font-end tool 55
1.3. DataBase tool 55
2. WebSite 56
2.1. Register interface 56
2.2. Login interface 56
2.3. Home page 57
2.4. Dashboard 57
2.5. Prediction interface 58
2.6. About Us 58
Conclusion 58
General conclusion 59

List of Figures

Figure 1: Logo 3
Figure 2: 1st study of existing Australian international agency 4
Figure 3: 2nd study of existing 4
Figure 4: 3rd study of existing 5
Figure 5: Classic methodology 7
Figure 6:Phases of CRISP DM 8
Figure 7.Datawarehouse 13
Figure 8:Postgresql 16
Figure 9: Talend 17
Figure 10: tFileInputExcel 17
Figure 11: tFilterRow 17
Figure 12: tPostgresqlInput 17
Figure 13: tPostgresqlOutput 17
Figure 14: tRowGenerator 18
Figure 15: tSortRow 18
Figure 16: tMap 18
Figure 17:tUniqRow 18
Figure 18: Trips dimension 19
Figure 19: First tMap 19
Figure 20: tFilterRow 20
Figure 21: tUniqRow 20
Figure 22: Second tMap 20
Figure 23: Output : Add 21
Figure 24: Output: Update 21
Figure 25: TpostgresqlInput 22
Figure 26: Trips postgresql 22
Figure 27: Date dimension 22
Figure 28: tGeneratorRow 23
Figure 29: tSortRow 23
Figure 30: tMap 24
Figure 31:Date postgresql 24
Figure 32:Accommodation dimension 24
Figure 33:1st tMap 25
Figure 34: 2nd tMap 25
Figure 35: 3rd tMap 25
Figure 36:Accommodation postgresql 26
Figure 37:Country dimension 26
Figure 38:tMap 26
Figure 39: Country postgresql 27
Figure 40: Type Arrivals dimension 27
Figure 41: tMap 28
Figure 42: Type Arrivals postgresql 28
Figure 43: Main purpose dimension 28
Figure 44: Main purpose postgresql 29
Figure 45: Expenditure dimension 29

Figure 46: Expenditure Postgresql 29
Figure 47: Season Postgresql 30
Figure 48:Landscape postgresql 30
Figure 49: Type transport PostgreSQL 31
Figure 50: Profil postgresql 31
Figure 51: Tourism postgresql 32
Figure 52: Job1 for accommodation 32
Figure 53: Job2 : Accommodation type 33
Figure 54: Job 3 : Expenditure 33
Figure 55: Job 4: Tourism Domestic 34
Figure 56: Job 5: Type main purpose 34
Figure 57: Job 6: Total arrivals and departures: tourism domestic 35
Figure 58: Job 7: External data 35
Figure 59:Fact table: postgresql 36
Figure 60: Fact table Postgresql 36
Figure 61.Profiling: Result of k_means algorithm 40
Figure 62.Result of algorithm 41
Figure 63.1st result 42
Figure 64.2nd result 43
Figure 65.train a logistic regression model 44
Figure 66. plot 45
Figure 67.Result of logistic regression 45
Figure 68.Result of prediction 45
Figure 69.Steps he model 46
Figure 70.Steps of the model 46
Figure 71.Result of prédiction 47
Figure 72.Logo Power BI 50
Figure 73.OverView 51
Figure 74.Compare 51
Figure 75.Inbound tourism 52
Figure 76.Outbound Tourism 52
Figure 77.Domestic Tourism 53
Figure 78.Accommodation 53
Figure 79.Profil details 54
Figure 80.Logo Flask 54
Figure 81.Logo Bootstrap 55
Figure 82.Logo phpMyAdmin 55
Figure 83.Register interface 56
Figure 84.Login interface 56
Figure 85.home page 57
Figure 86.Dashboard page 57
Figure 87.Prediction interface 58
Figure 88.About us interface 58

List of Tables

Tableau 1: Profiling: variable type 39
Tableau 2: Attractivity: Type of variable 40
Tableau 3: the performance : type variable 41

General Introduction

Business Intelligence (BI) is a technology-driven process that involves analyzing and transforming data into actionable insights and reports to support data-driven decision-making. In the tourism industry, BI can be used to gain a deeper understanding of customers, destinations, and travel trends, which can help businesses improve their competitiveness, optimize their operations, and increase profitability.

Tourism is a critical sector in many countries, and is a major industry worldwide, with millions of people traveling every year for business, leisure, or other reasons, it contributes greatly to their economies. In today's data-driven world, the tourism industry faces many challenges that require innovative solutions to enhance its sustainability and competitiveness.

With the increasing availability of data and technological advancements, it has become essential to leverage these resources to improve the tourism industry. By using these resources, they can gain valuable insights into consumer behavior and preferences, optimize operational efficiency, improve the overall tourism experience, and ultimately drive industry growth.

In this project, we aim to provide a dashboard for the tourism industry that offers insights into key performance indicators (KPIs) and helps stakeholders make data-driven decisions.

Chapter 1: Project Context

Introduction

To commence our project, proper research and exploration are required. First of all, our data needs to be thoroughly analyzed and understood. Then, we will be looking to dive into other solutions, similar to what we are trying to achieve. Furthermore, we need to study and pick from methodologies that enable such projects to be tackled all based on what we're aiming to do and what information is available to us.

I. Company presentation

Figure 1: Logo

As our name suggests, **"Trip Trend",** refers to the prevailing patterns or tendencies in the travel behavior of people. we are a startup that specializes in collecting and analyzing data to link and extract information that helps our customers with decision-making, and predicting changes and fluctuations.

« Trip trend » analysis involves studying various factors that influence travel behavior, such as demographics, economic conditions, social and cultural factors, and technological developments. By identifying and analyzing these trends, travel businesses can gain insights into customer behavior, preferences, and needs, allowing them to tailor their services and offerings accordingly.

II. Context of the project

1. Study of the existing

Nowadays the internet is the first source of information for people, but the problem here is that information on the internet is not always trustworthy. We are going to make a study of existing.it is an important step in implementing a BI Solution. This study supports the project management's understanding of the actual information system.

We'll start with an Australian international agency its ergonomics are perfectly styled. However, the displayed numbers get reset each day so there is no clear record or available history of said numbers

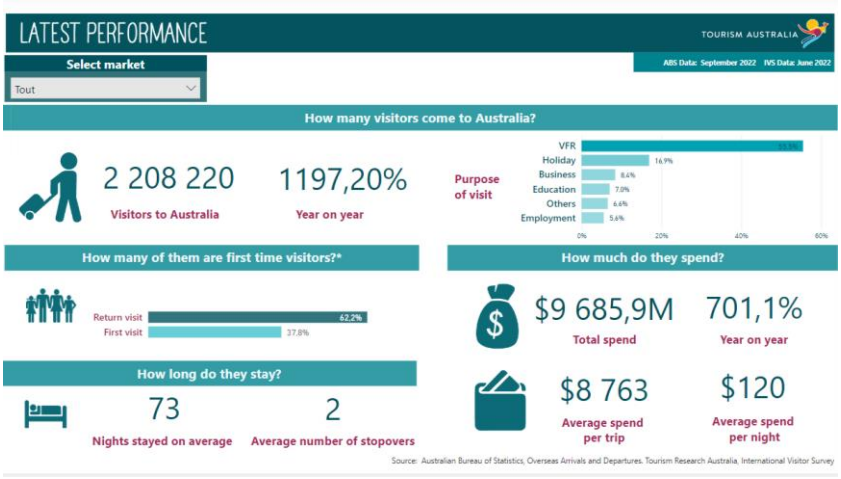


Figure 2: 1st study of existing Australian international agency

The Organization for Economic Co-operation and Development is considered detailed without a doubt but what it gains in details lacks in ergonomics and styling. They plot too many lines on the same graph, it becomes cluttered and hard to read .

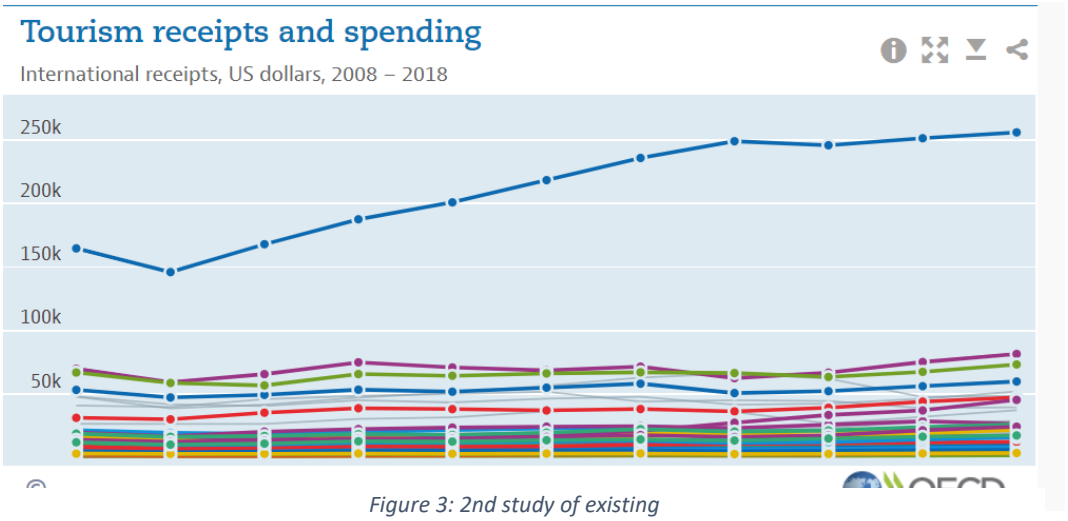


Figure 3: 2nd study of existing

This particular dashboard is not completely detailed when it comes to names of country data Also we can't choose multiple choices of characteristics so we can't compare changes over the same period. In addition, we can't visualize the percentages of trips purpose in the same pie chart so it's difficult to analyze.

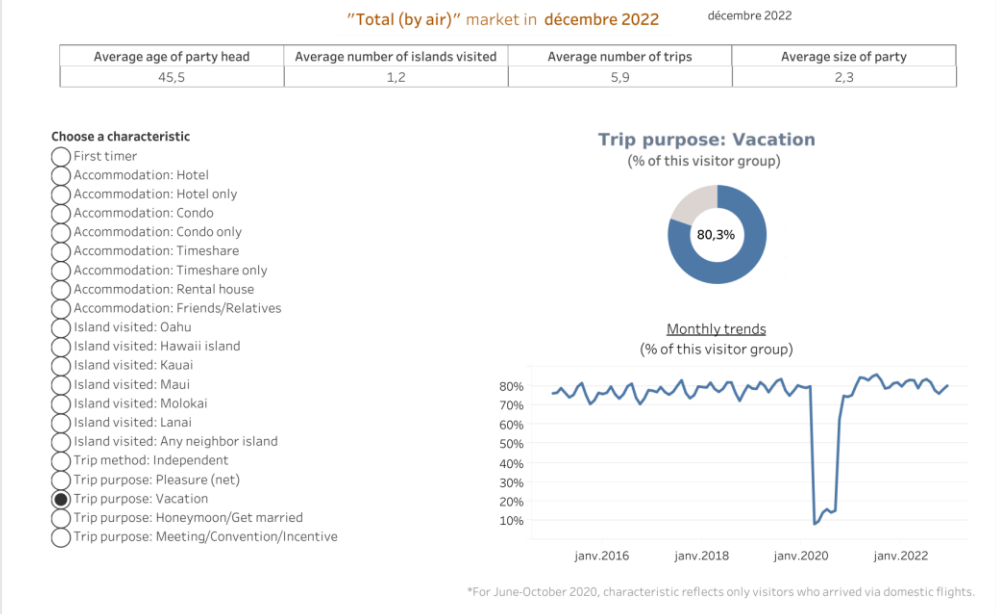


Figure 4: 3rd study of existing

2. Problematic

The COVID-19 pandemic has had a significant impact on the tourism industry, with many countries imposing travel restrictions and lockdowns to control the spread of the virus. Some of the consequences of the pandemic for the tourism industry include :

Decreased travel demand : Due to travel restrictions and concerns about the virus, many people have canceled or postponed their travel plans, resulting in a significant decrease in travel demand.

Shifts in travel behavior : The pandemic has led to shifts in travel behavior, with many people opting for domestic travel or short-haul trips instead of long-haul flights.

all this creates a big problem for travel agencies especially, those who are trying to figure out how to increase the number of reservations.

3. Adopted Solution

One-click deployment of analytical solutions: The deployment of analytical solutions can often be a complex process that requires specialized knowledge and resources. To make it easier for users to access these solutions, we propose a one-click deployment system that automates the installation and configuration process, allowing users to quickly and easily deploy analytical solutions without requiring extensive technical expertise.

Improved accessibility to information: In today's data-driven world, access to information is critical for decision-making. We propose to improve accessibility to information by creating an intuitive and user-friendly interface that enables users to quickly locate the information they need. This could involve features such as advanced search capabilities, personalized dashboards, and interactive data visualization tools.

Predictive analytics: Predictive analytics is a powerful tool that enables organizations to make informed decisions by leveraging data to forecast future outcomes. We propose to implement predictive analytics tools that can help businesses identify patterns and trends in their data, anticipate potential issues, and make proactive decisions.

Customer profiling: Understanding customer behavior is essential for creating effective marketing campaigns and delivering personalized experiences. We propose to develop customer profiling tools that can analyze customer data to identify trends, preferences, and behavior patterns. This information can then be used to tailor marketing campaigns, improve customer engagement, and increase customer satisfaction.

Overall, these proposed solutions are aimed at improving decision-making and empowering businesses to leverage data-driven insights for growth and success.

4. Methodology of work

The importance of project management in business cannot be overstated, because to achieve our objectives we must choose a methodology to be able to organize our work efficiently.

4.1.Classic methodology

The traditional approach to project management is linear where all phases of a process take place in sequence. The concept of the latter is based on tools and a predictable experience. Each project follows the same life cycle which always includes the same steps as shown in the figure

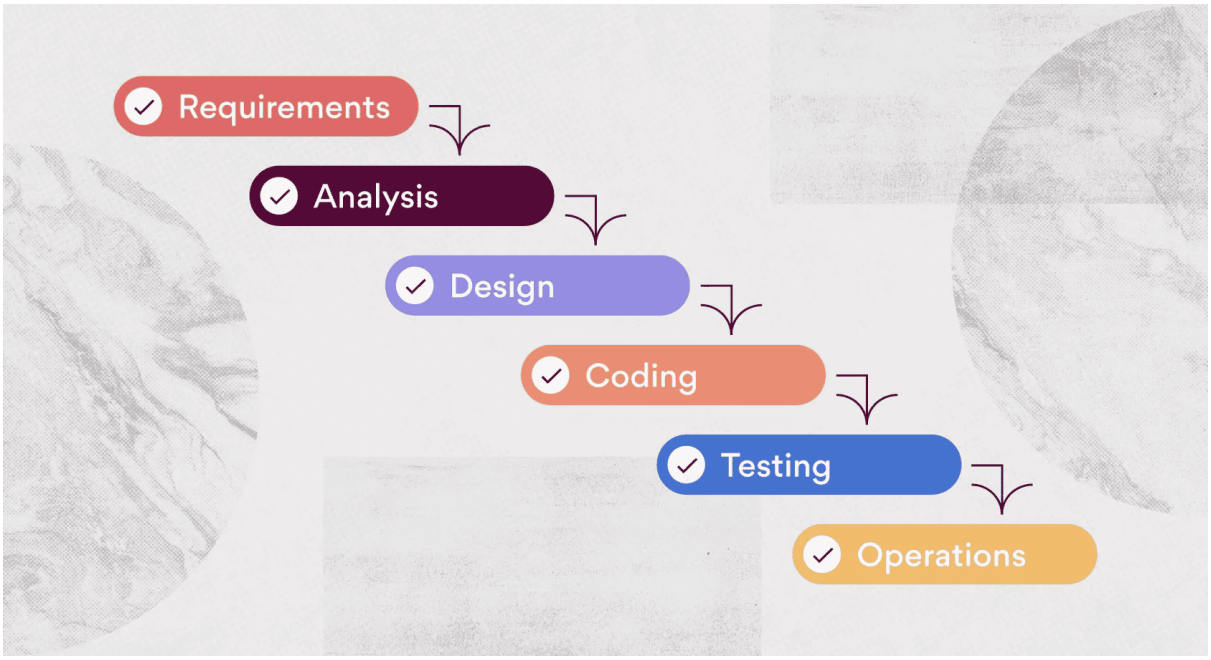


Figure 5: Classic methodology

4.2.advanced methodology

CRISP-DM (Cross-Industry Standard Process for Data Mining) is a widely-used methodology for developing data mining projects. It consists of six main phases:

1- Business Understanding
2- Data Understanding
3- Data Preparation
4- Modeling
5- Evaluation
6- Deployment

The CRISP-DM methodology provides a structured and systematic approach to data mining projects, helping to ensure that projects are well-planned, well-executed, and deliver the desired results. It is widely used in the industry as a standard approach for data mining and analytics projects.

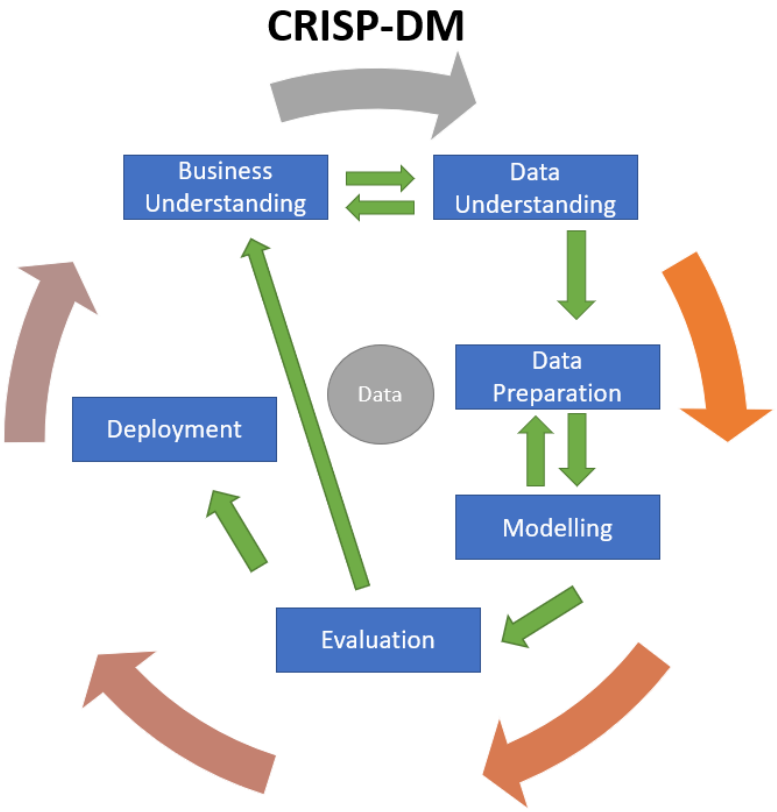


Figure 6:Phases of CRISP DM

After doing research between the two methodologies, we find that the CRISP-DM method is the most adapted to our project.

Conclusion

After exploring multiple solutions similar to ours and researching different methodologies, we have got a general idea of how to approach our project. We were able to get an overview of the projectand , we established our work environment, its structure, and workflow, as indicated by the first phase of the CRISP-DM methodology.

Chapter 2: Data Modeling

Introduction

In chapter 2 of our work, we dive into the technical aspect of our Datawarehouse, but first, we need to specify Functional and Non-functional requirements which are an essential phase in this project, also we need to get acquainted with data modeling techniques and what works best for us and create a model for our warehouse.

I. Identification of functional and non-functional requirements

1. Functional requirements

- **Profiling:**

Personalization: Profiling enables a travel agency to understand the preferences, interests, and travel habits of its customers. With this information, they can personalize their offerings and recommendations, creating a more tailored experience for each customer.

- **real-time access to data:**

Faster decision-making: With real-time access to data, a travel agency can make informed decisions quickly, allowing them to respond to changes in the market and customer demands more efficiently.

Competitive advantage: Real-time data can provide a travel agency with a competitive advantage, allowing them to make more informed decisions and more quickly to changes in the market. This can help them to stay ahead of the competition and attract more customers.

- **The prediction of a country's attractiveness:**

Improved customer experience: Predicting a country's attractiveness can help travel agencies provide a more personalized and satisfying customer experience. They can use this information to recommend destinations and activities that align with their customers' interests and preferences.

2. Non- Functional requirements

Performance: The dashboard should be able to handle large amounts of data and queries without slowing down or crashing. It should also load quickly and be responsive to user interactions.

Reliability: The dashboard should be reliable and available to users at all times. It should be able to recover from errors and failures gracefully.

Security: The dashboard should be secure and protect sensitive data from unauthorized access or disclosure. It should also comply with relevant data protection regulations.

Usability: The dashboard should be easy to use and intuitive, with clear and concise visualizations and navigation. It should also be customizable to meet the specific needs of different users.

Accessibility: The dashboard should be accessible to users with disabilities, such as those who are visually impaired or have limited mobility. It should comply with accessibility standards and provide alternative ways to access information.

II. Modeling techniques

1. Star Schema

Star schema is a popular data modeling technique used in data warehousing that is optimized for querying large datasets. It consists of a central fact table surrounded by a set of dimension tables, with each dimension table representing a specific attribute or characteristic of the data being analyzed.

1.1. Advantages

- *Simplified Querying:* Star schema simplifies querying of large datasets by enabling queries to be executed in a fast and efficient manner.
- *Flexibility:* Star schema provides flexibility in adding new dimensions to the schema.
- Reduced Data Redundancy: Star schema helps to reduce data redundancy because each dimension table only contains a single copy of a specific attribute, rather than storing the attribute in multiple tables.

1.2. Disadvantages

- *Redundancy in dimensions:* Each dimension is stored in a separate dimension table which results in denormalization.

- *Complex diet:* All dimensions do not relate to measurement.

2. Snowflake Schema

Snowflake Schema is the star Schema with dimension normalization. There may be dimension hierarchies to divide dimension tables when they are too large.

2.1.Advantages

- Normalization of dimensions
- *Reduction of redundancy:* A hierarchy between dimensions exists.

2.2.Disadvantages

- More complex model
- *Performance degradation:* It is complex with many often-expensive joins

3. Constellation Schema

Constellation Schema represents many fact relationships that share common dimensions. These deferential fact relationships make up a family that shares dimensions but where each fact relation has its dimensions.

3.1.Advantages

- The dimension tables common to the different fact tables are not subject to redundancies.
- Here, dimensional tables are shared by the number of fact tables.

3.2.Disadvantages

- It is difficult to understand as it is a very complex schema to implement.
- It uses more space in the database compared to the star schema.
- It has many joins between dimensional and fact tables and thus it is difficult to understand.

⇨ After research and comparison, we chose to work with « Star schema » for the following reasons:

- We have one fact table: Fact_Travel_Agency.
- The dimensional tables will be shared by one fact table.

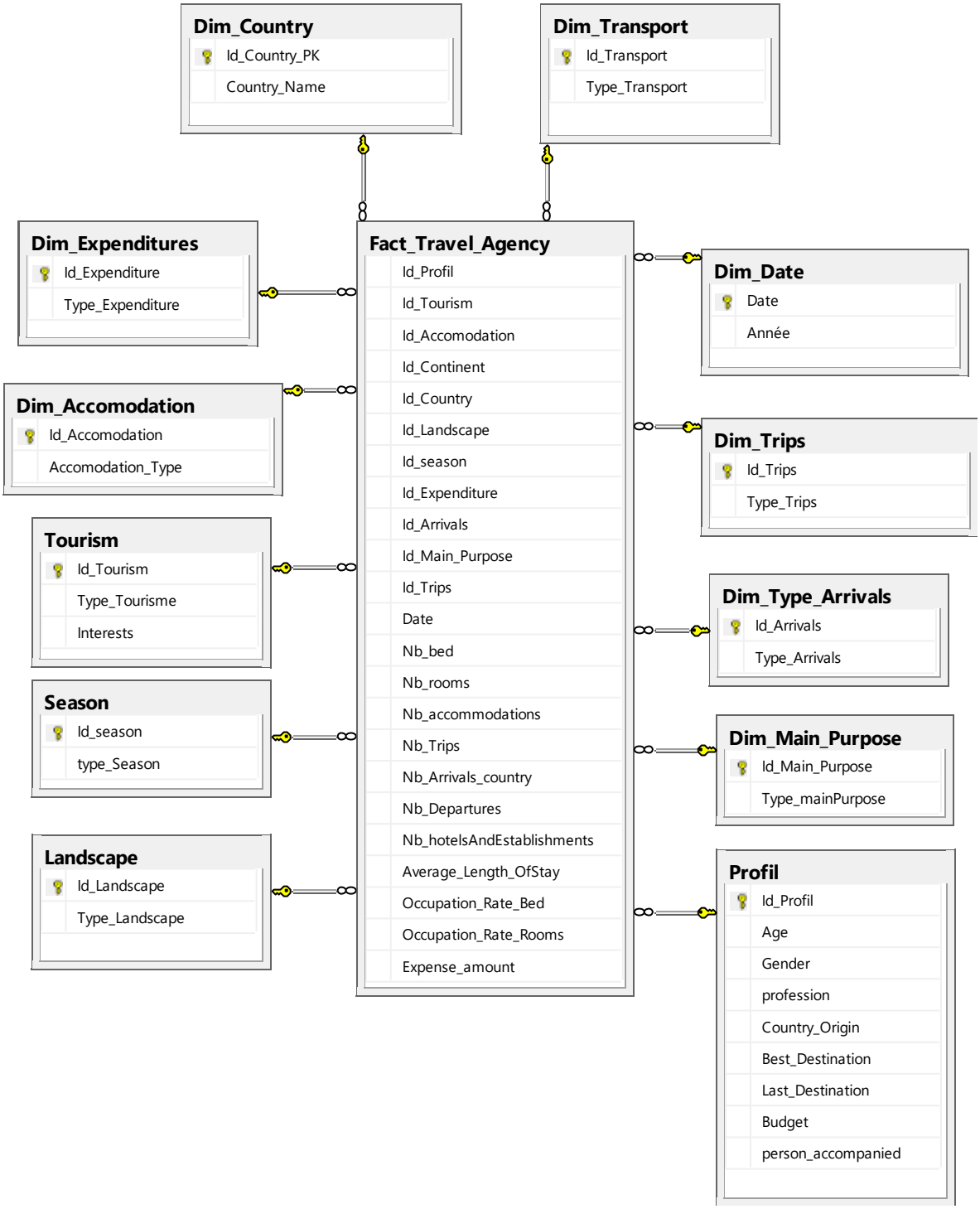


Figure 7.Datawarehouse

III. KPI Identification

The DW must contain useful information used to extract measurements called KPIs or key indicators. To identify all of this information we based ourselves on the information given to us by our tutors which is internal data plus the external data we have scrapped and added to our data.

This is the 5th step of the second phase of the Crisp-dm Methodology: "Choice of indicators".

Number of beds; Number of rooms; Number of accommodations

Number of Trips; Number of Arrivals; Number of Departures

Number of hotels And Establishments; Average Length of Stay

Occupation Rate of Bed; Occupation Rate of Rooms; Expense amount

IV. Dimensions Identification

Dimension identification is the process of identifying the key attributes of data that are important for analysis in a data warehouse. Dimensions are used to categorize, filter, and analyze data, and are a key component of a dimensional data model.

After the needs that we worked out, it was necessary to provide for the creation of the following dimensions which relate to travel agency services:

Dimension Country ; Dimension Expenditures; Dimension Accommodation

Dimension Tourism ; Dimension Season; Dimension Landscape

Dimension Profil ; Dimension Main purpose; Dimension Type Arrivals

Dimension Trips ; Dimension Transport ; Dimension Date

Conclusion

In this phase of our work, we managed to settle on the data modeling technique Star schema after comparison with snowflake schemas and Constellation schemas and modeled our data warehouse accordingly.

Chapter 3: Data Integration

Introduction

In this part, we're looking to integrate our data and impute any missing or unusable values to finally do a full implementation of our dimensions using Talend's data tools.

I. Working tool:

1. Database management tool

PostgreSQL is a powerful, open-source relational database management system (RDBMS) that uses and extends the SQL language. It is known for its reliability, robustness,and performance, and is widely used for large-scale applications and data warehouses.

Figure 8:Postgresql

2. Data integration: Talend

A data integration tool is a software application or platform that enables the seamless and automated integration of data from various sources, formats, and systems into a unified and consistent format. Data integration tools help organizations extract, transform, and load (ETL) data from different sources, such as databases, spreadsheets, flat files, and cloud applications, and integrate them into a single data warehouse or data lake. They also enable the synchronization and replication of data between different systems and applications.

During our project, we used Talend which allows you to create a complete pipeline for ETL and Data integration.

Figure 9: Talend

One of Talend's great strengths is that it can connect to almost any existing data source, business application, and file type. And, thanks to more than 250 components. Among its components, we used:

Figure 10: tFileInputExcel

tFileInputExcel: It's a component that can be used to read data from Excel files and convert them into a format that can be processed by other Talend components.

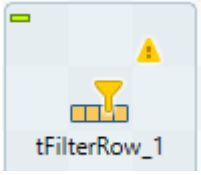


Figure 11: tFilterRow

TFilterRow: It is used to filter rows of data based on specific criteria, allowing users to focus on the data that is most relevant to their needs.

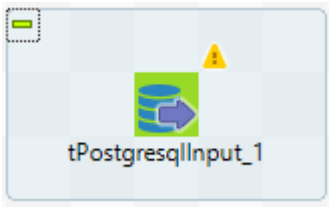


Figure 12: tPostgresqlInput

tPostgresqlInput: It is used to read data from a PostgreSQL database and convert it into a format that can be processed by other Talend components.

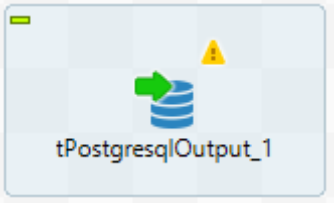


Figure 13: tPostgresqlOutput

tPostgresqlOutput: It is used to write data to a PostgreSQL database, allowing users to insert, update, or delete data in a database table.

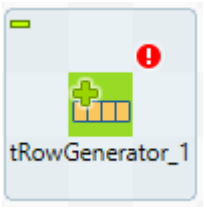


Figure 14: tRowGenerator

tRowGenerator: It is used to generate a specified number of rows of dummy data, which can be used for testing, development, or training purposes.

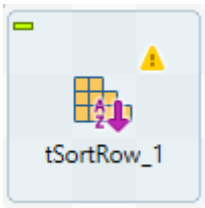


Figure 15: tSortRow

tSortRow: It is used to sort data based on specific criteria, allowing users to order(ascending or descending) data in a way that is most relevant to their needs.

Figure 16: tMap

tMap: It is used to transform and manipulate data from multiple sources, allowing users to combine, filter, and modify data in a way that is most relevant to their needs.

Figure 17:tUniqRow

tUniqRow: It is used to remove duplicates from a dataset, ensuring that each record is unique based on one or more key fields.

3. Data Implementation

The implementation of our data model is a physical realization on a real machine of the components of the abstract machine that together constitute that model alongside the necessary data.

3.1.Internal Data

3.1.1. Trips dimension

We have the trips dimension table which contains the type of overnight trips or same days trips.

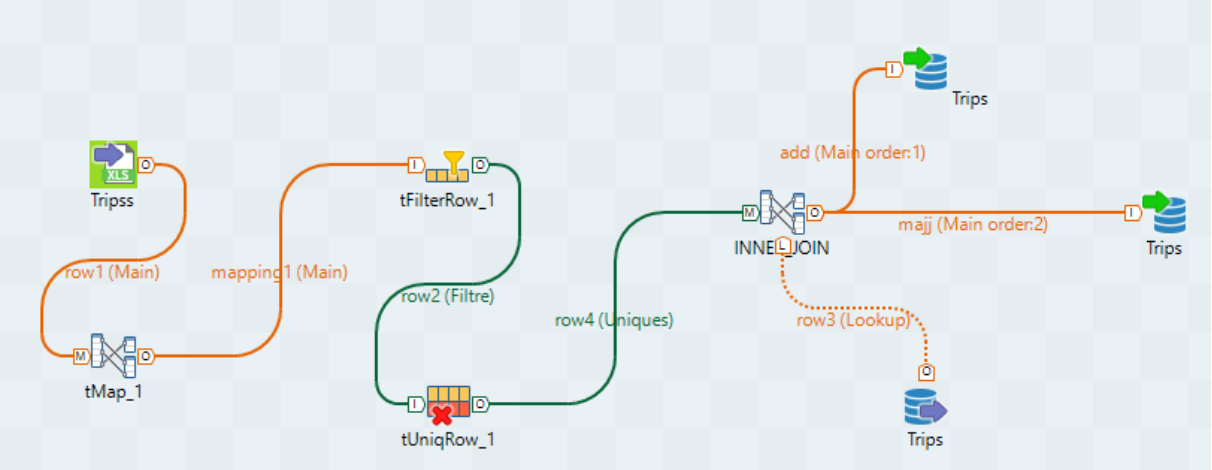


Figure 18: Trips dimension

And here is the tMap configuration. Here we select just the column where we have information about the type of trips.

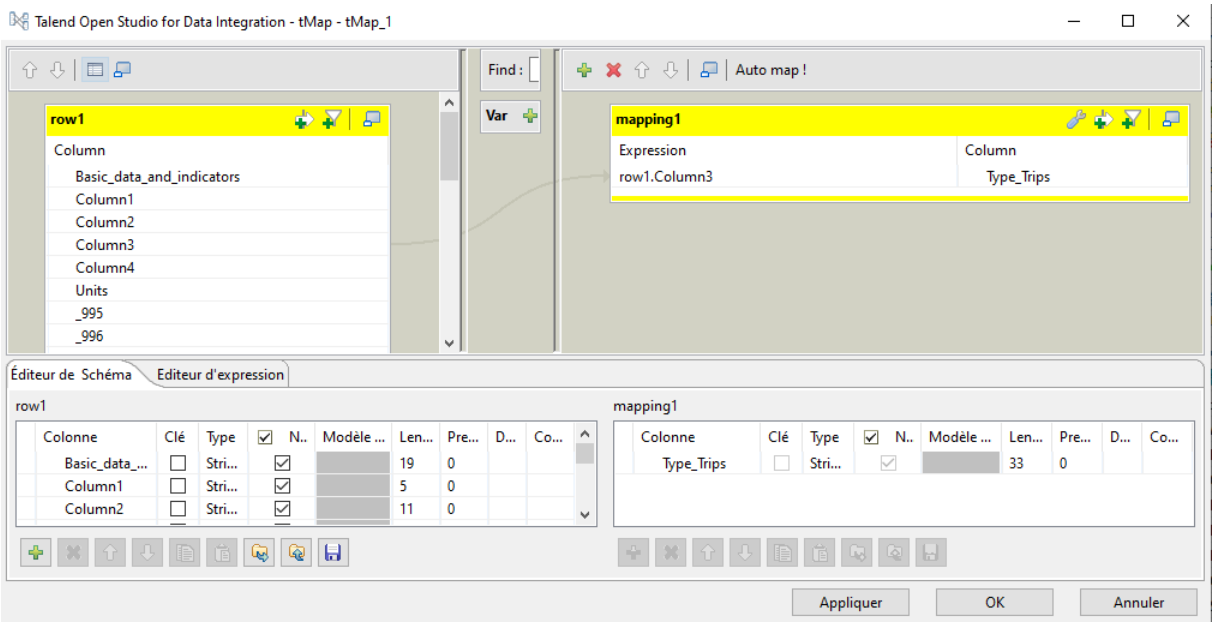


Figure 19: First tMap

here we use the tFilterRow to delete the null rows.

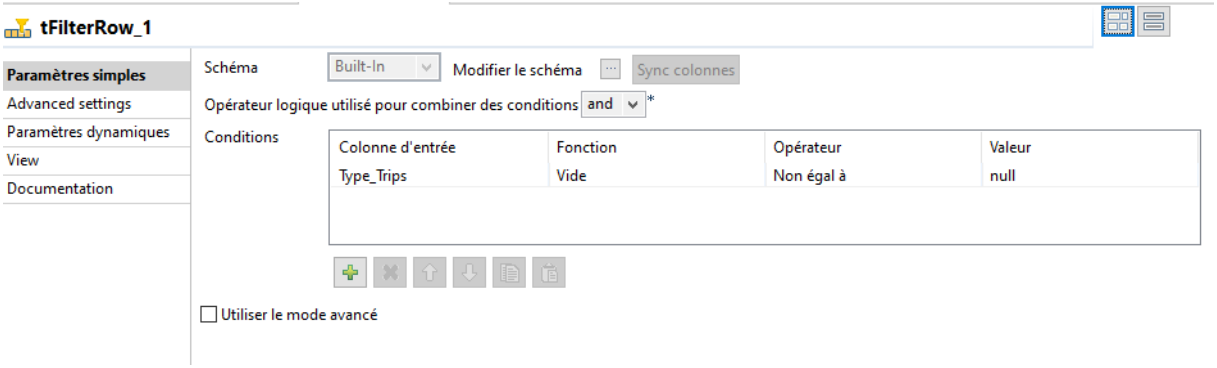


Figure 20: tFilterRow

Then we use tUniqRow to eliminate the redundancy of rows

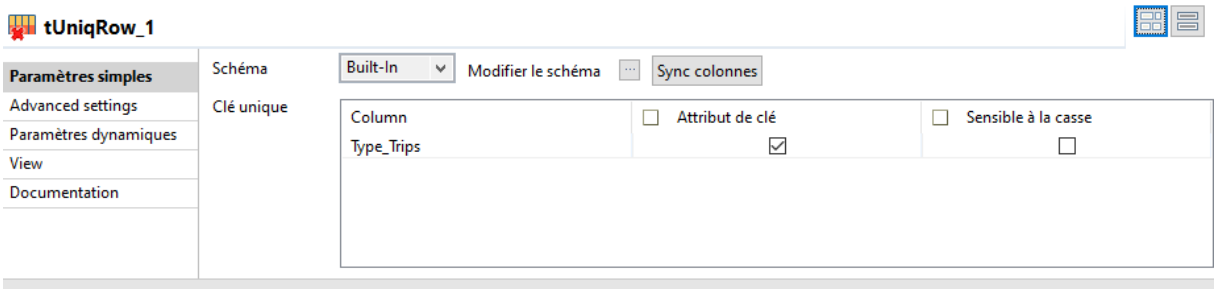


Figure 21: tUniqRow

After that, we added the second tmap where We joined the data to process the added and modified data, and we specify « Id_Trips » as the primary key of this dimension.

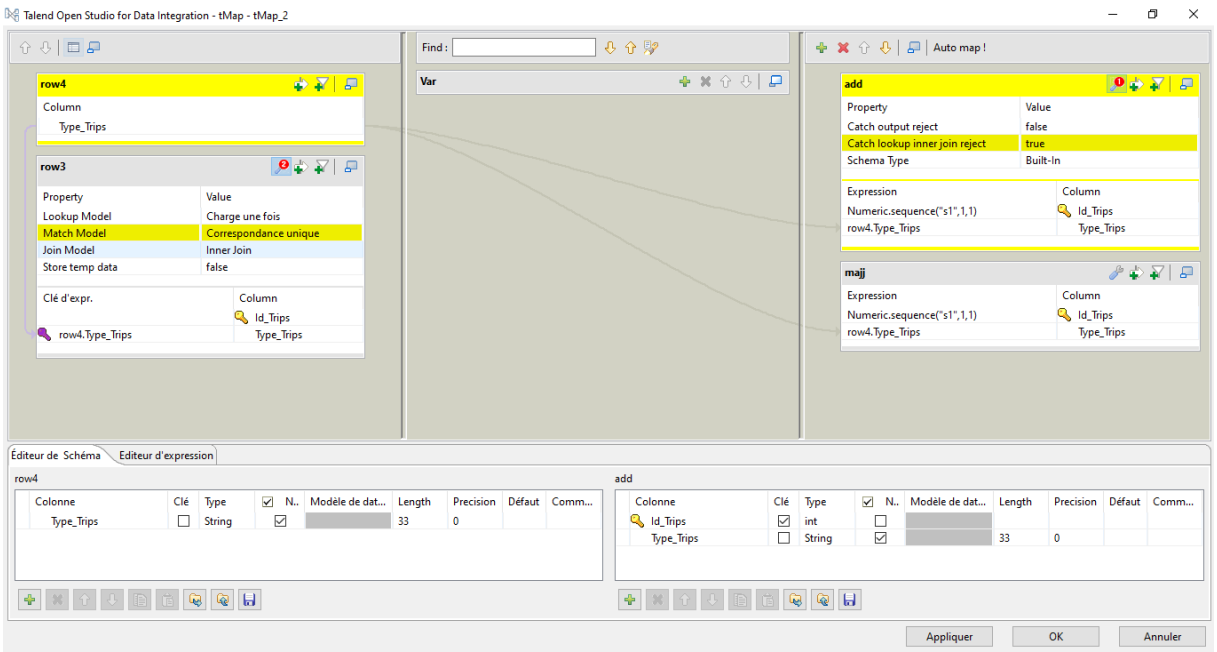


Figure 22: Second tMap

After doing the inner join, we added the first tPostgresOutput for the addition of new data.

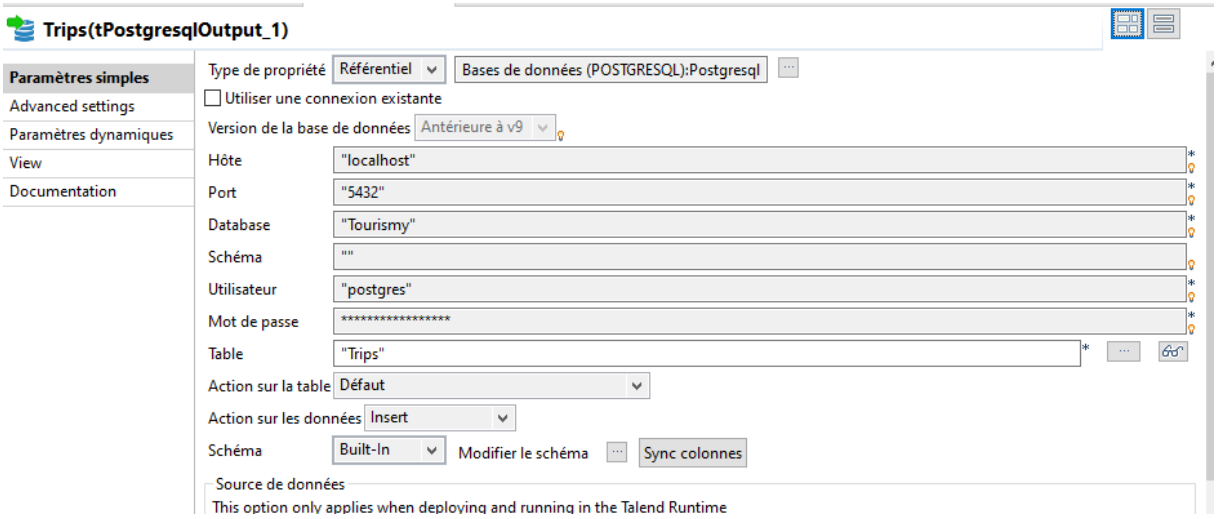


Figure 23: Output : Add

we added also the second tPostgresOutput for the update of data.

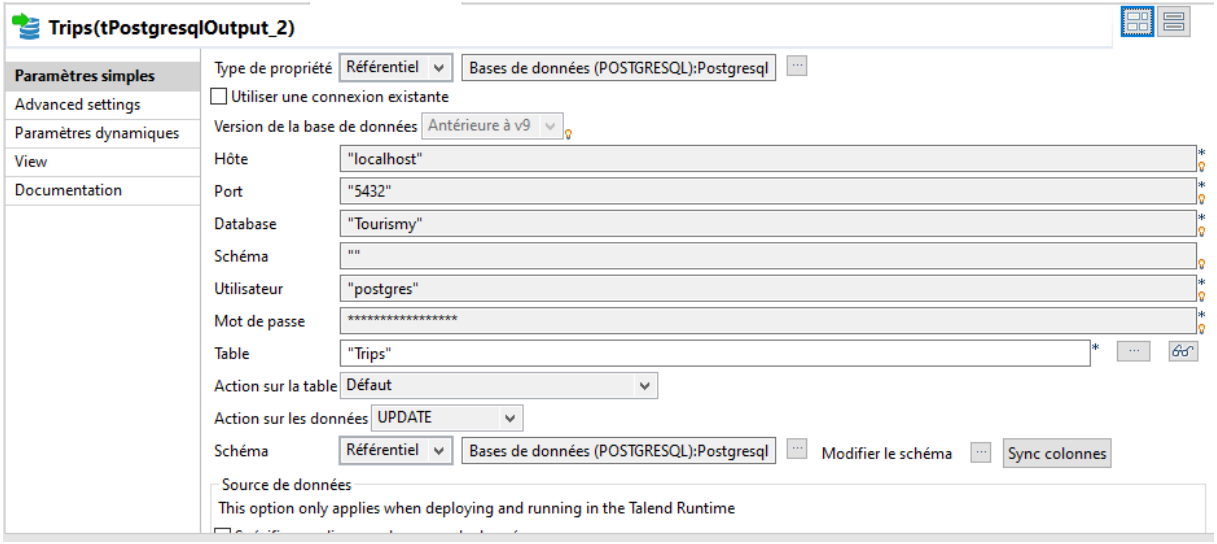


Figure 24: Output: Update

And we use tPostgresqlInput to extract data from a PostgreSQL database.

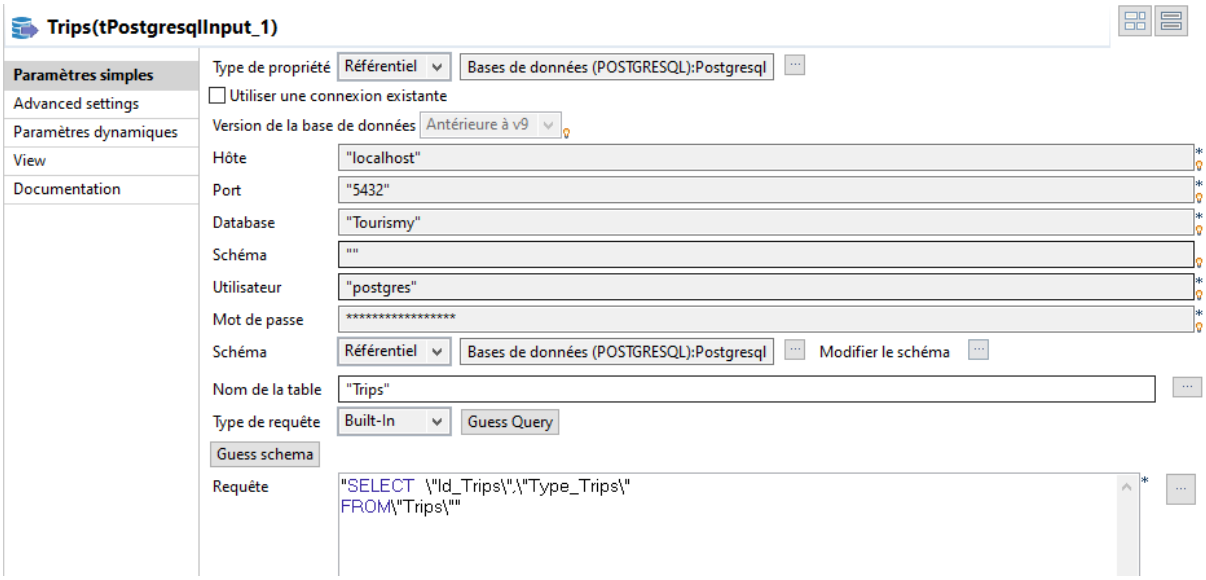


Figure 25: TpostgresqlInput

➪ The result in Postgresql

	Id_Trips [PK] integer	Type_Trips character varying (33)
1	1	Overnights visitors (tourist...
2	2	Same-day visitors (excursi...

Figure 26: Trips postgresql

3.1.2. Date dimension

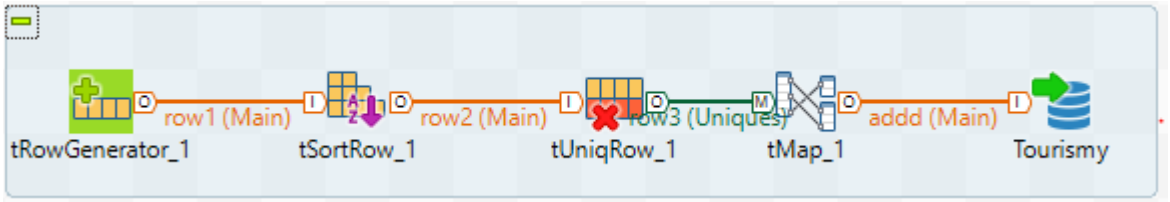


Figure 27: Date dimension

We use tRowGenerator to fixe the min and max « year »

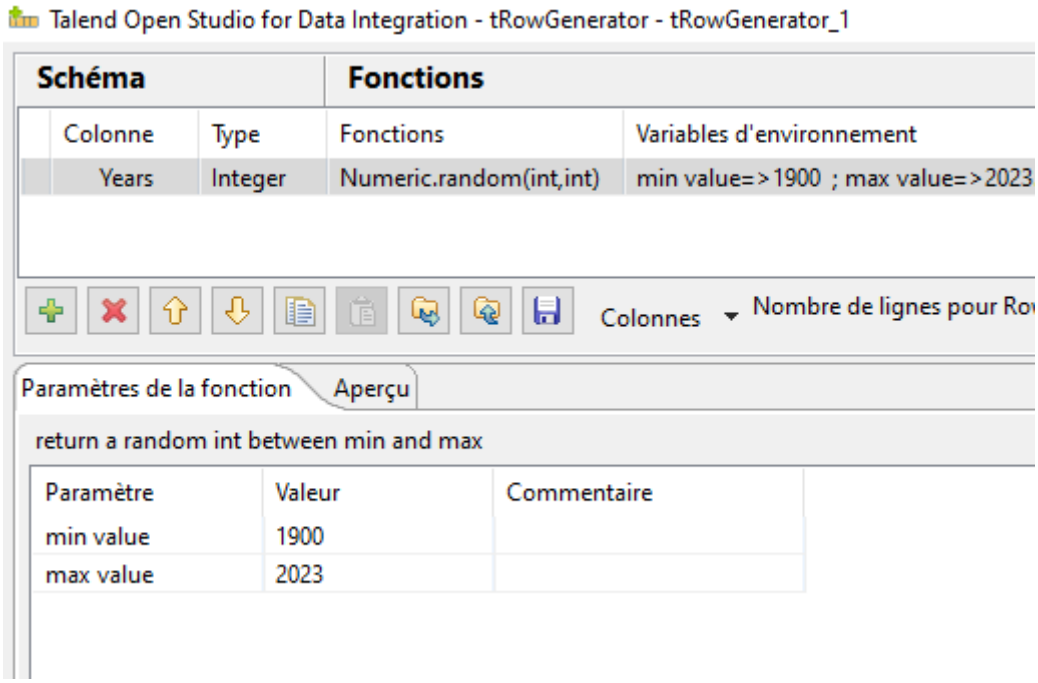


Figure 28: tGeneratorRow

TSort to specify the order of data (ascending)

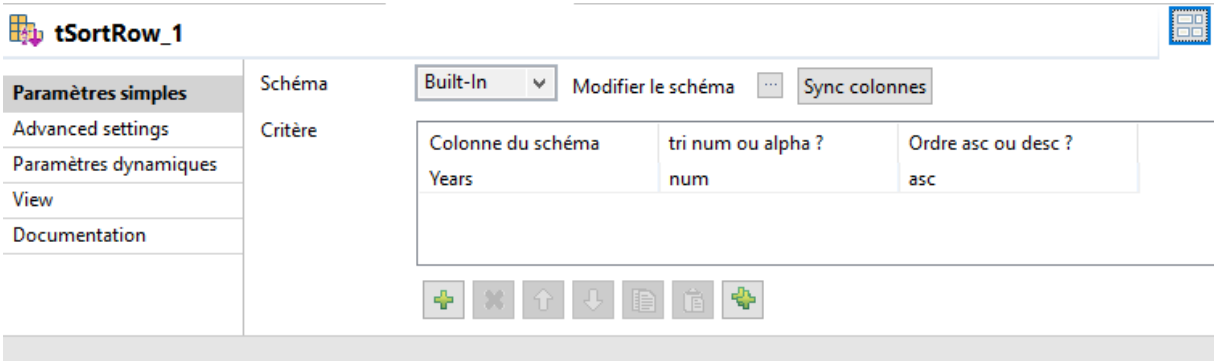


Figure 29: tSortRow

The tMap configuration

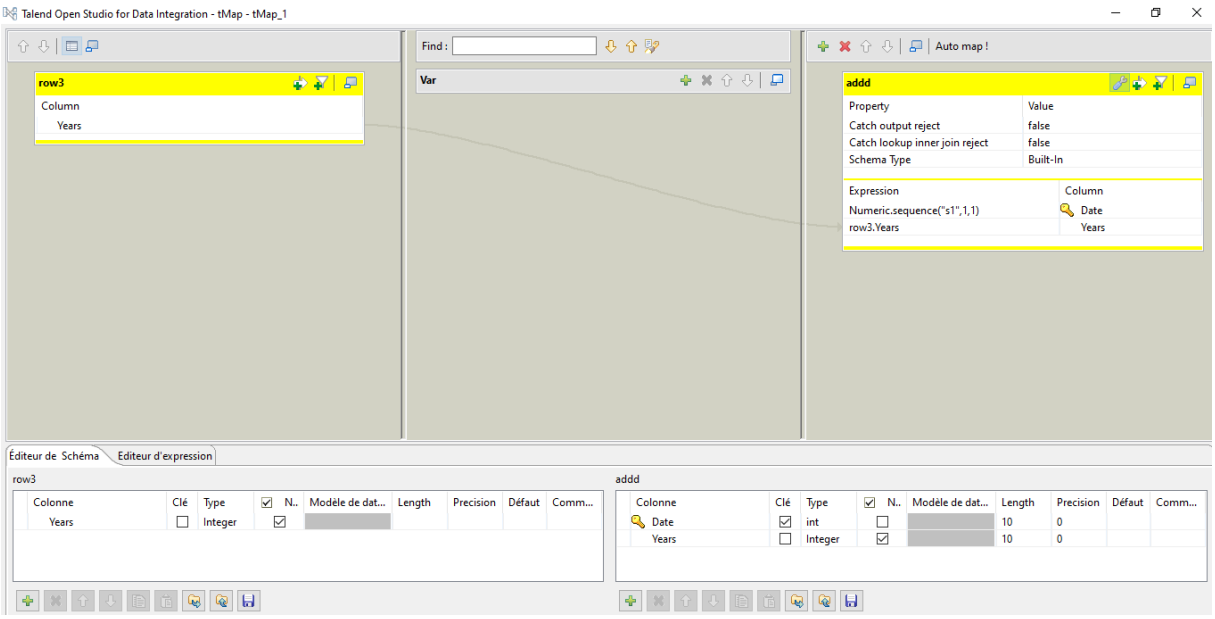


Figure 30: tMap

⇨ The result in Postgresql

	Date [PK] integer	Years integer
1	1	1900
2	2	1901
3	3	1902
4	4	1903
5	5	1904
6	6	1905
7	7	1906
8	8	1907
9	9	1908
10	10	1909
11	11	1910

Figure 31:Date postgresql

3.1.3. Accommodation dimension (integration of internal and external data)

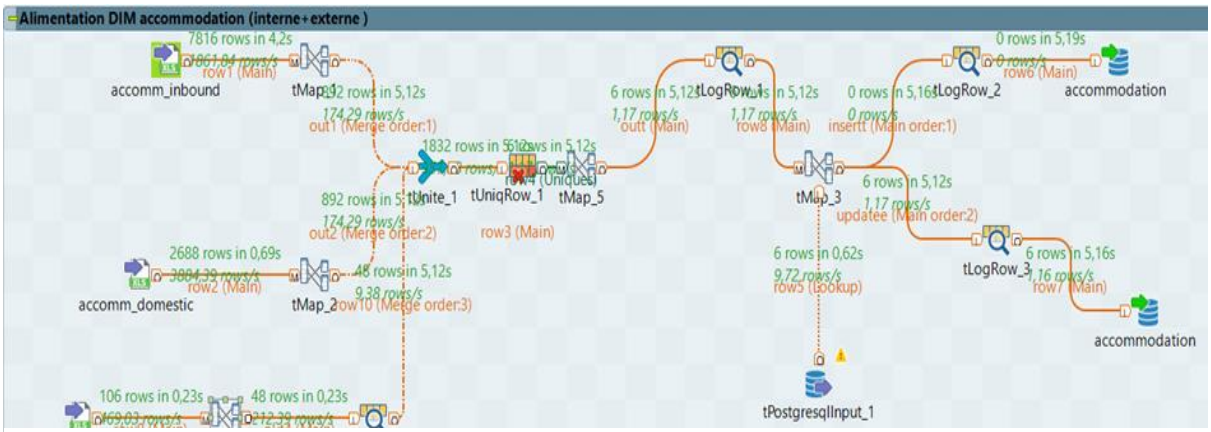


Figure 32:Accommodation dimension

I have configured each TMap to process data according to the t following transformation and filtering rules

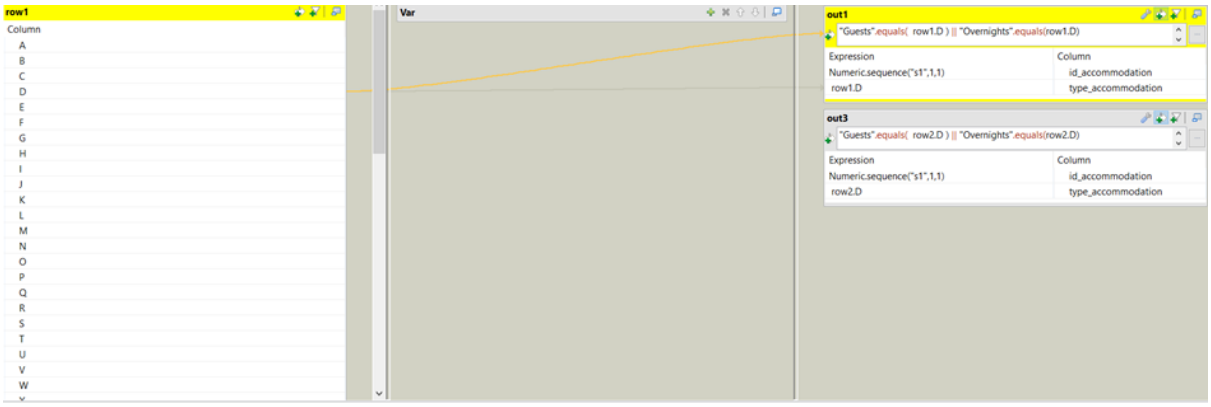


Figure 33:1st tMap

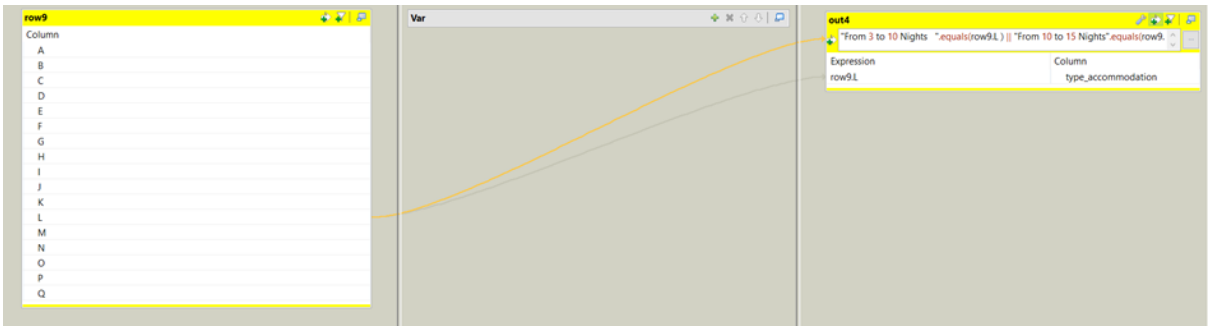


Figure 34: 2nd tMap

To combine the results of threeTMaps in Talend, I used the 'tUnite' component. I added the 'tUniqRow' component to the workspace to remove duplicates and then connected the 'tUniqRow' component to TMAP5. To feed data into the 'accommodation' table in PostgreSQL using Talend, I used the 'tPostgreSQLOutput1' component named 'accommodation to perform the join of the two data sources, I configured TMap. In the 'Join Model' section, I specified the join fields for an 'inner join' between the two data sources.

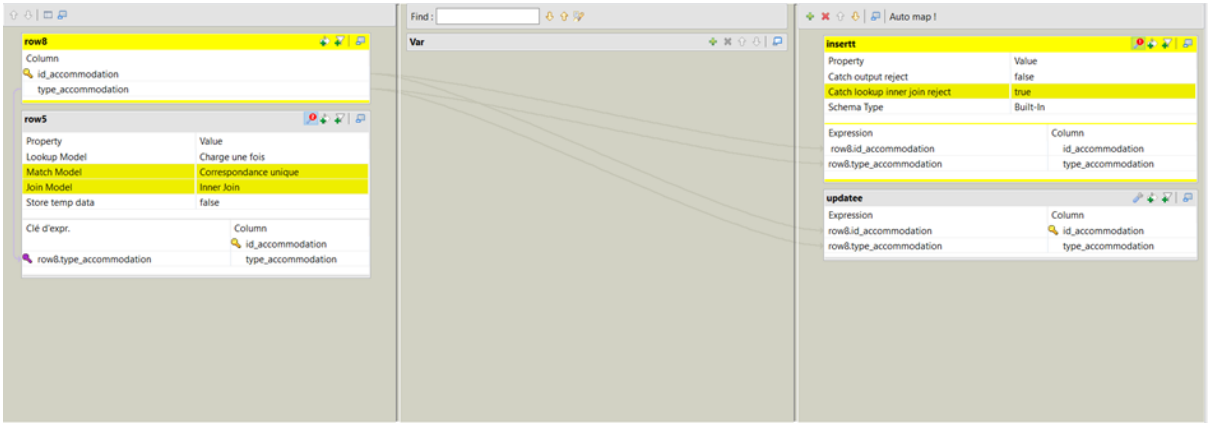


Figure 35: 3rd tMap

⇨ The result in Postgresql

	id_accommodation [PK] integer	type_accommodation character varying
1	1	Guests
2	2	Overnights
3	3	From 10 to 15 Nights
4	4	Less than 3 Nights
5	5	15 Nights or more
6	6	less than 1 Night

Figure 36:Accommodation postgresql

3.1.4. Country dimension

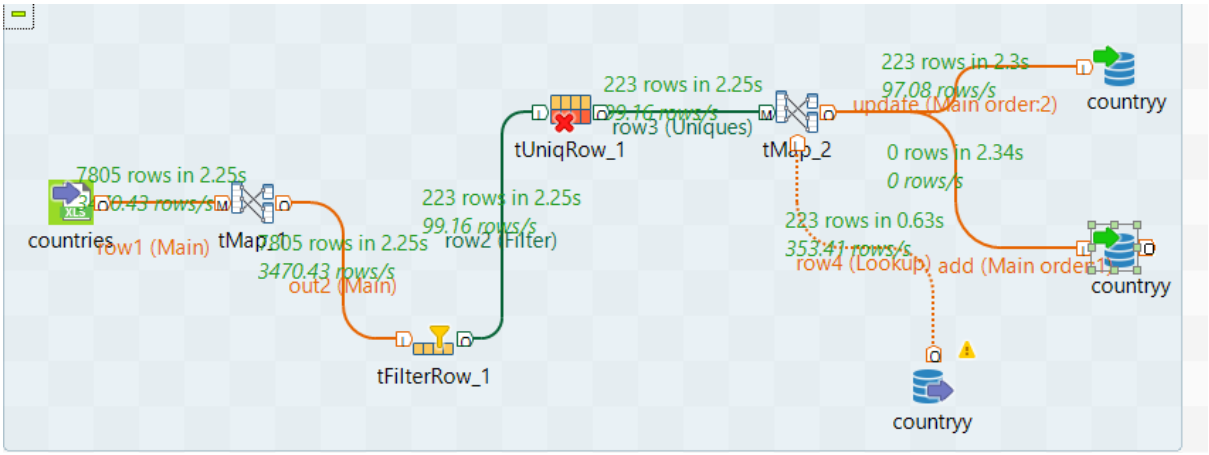


Figure 37:Country dimension

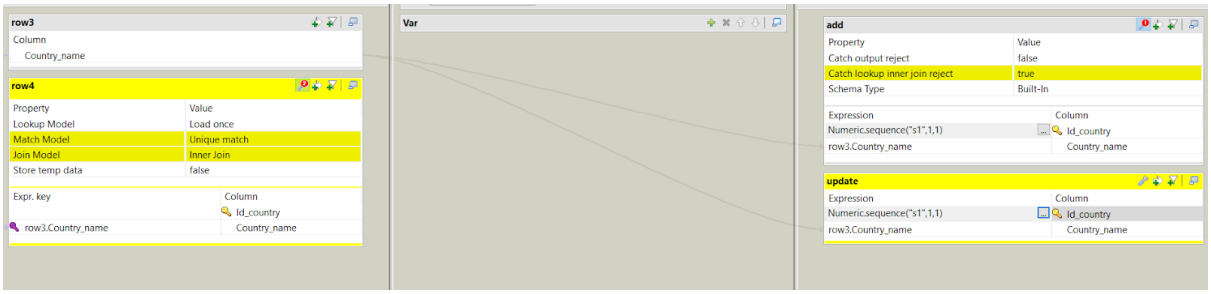


Figure 38:tMap

⇨ The result in Postgresql

	Id_country [PK] integer	Country_name character varying (1000)
1	1	AFGHANISTAN
2	2	ALBANIA
3	3	ALGERIA
4	4	AMERICAN SAMOA
5	5	ANDORRA
6	6	ANGOLA
7	7	ANGUILLA
8	8	ANTIGUA AND BARBUDA
9	9	ARGENTINA

Figure 39: Country postgresql

3.1.5. Type Arrivals dimension

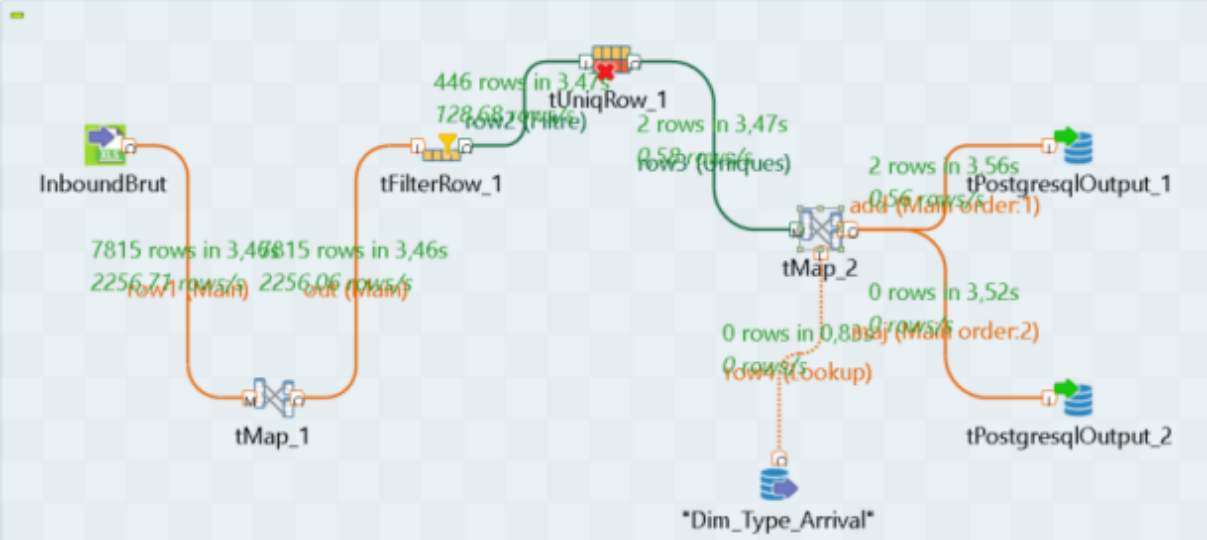


Figure 40: Type Arrivals dimension

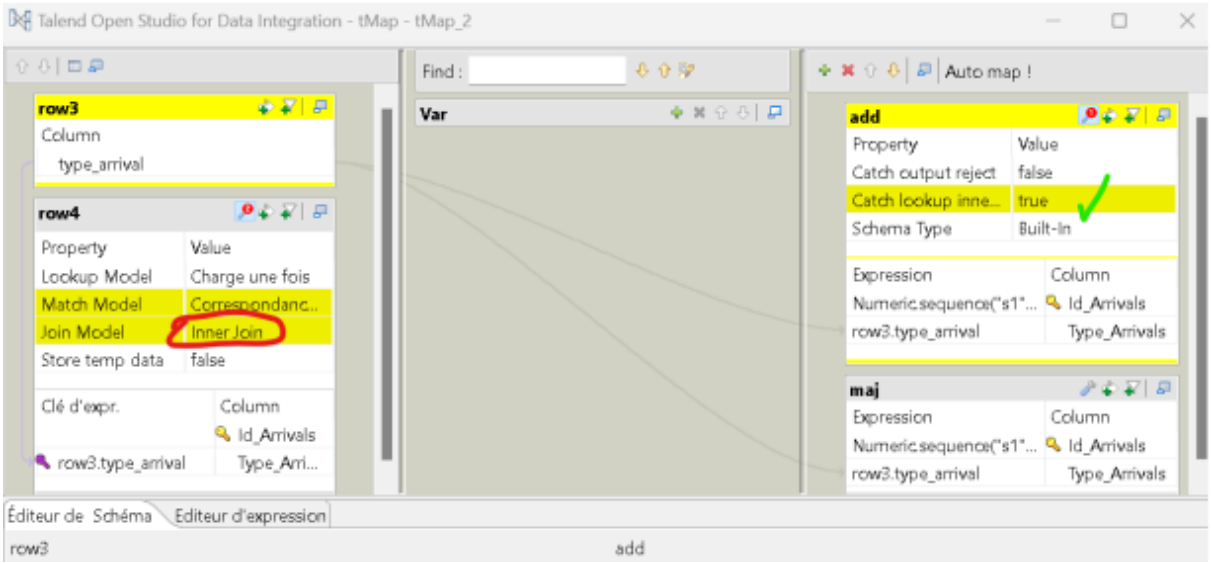


Figure 41: tMap

⇨ The result in Postgresql

Id_Arrivals integer	Type_Arrivals character varying
1	Overnights visitors
2	Same-day visitors

Figure 42: Type Arrivals postgresql

3.1.6. Main purpose dimension

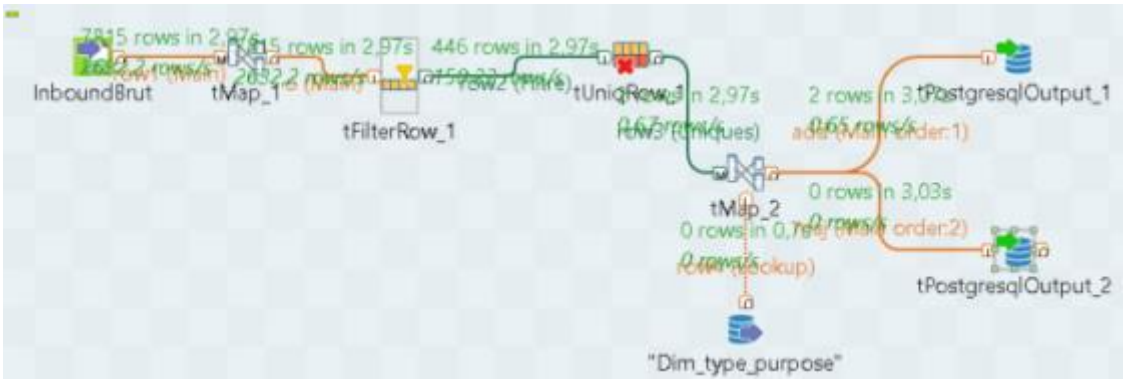


Figure 43: Main purpose dimension

⇨ The result in Postgresql

	Id_Main_Pur [PK] integer	type_Main_P text
1	1	Personal
2	2	Business an
*		

Figure 44: Main purpose postgresql

3.1.7. Expenditure dimension (internal and external data)

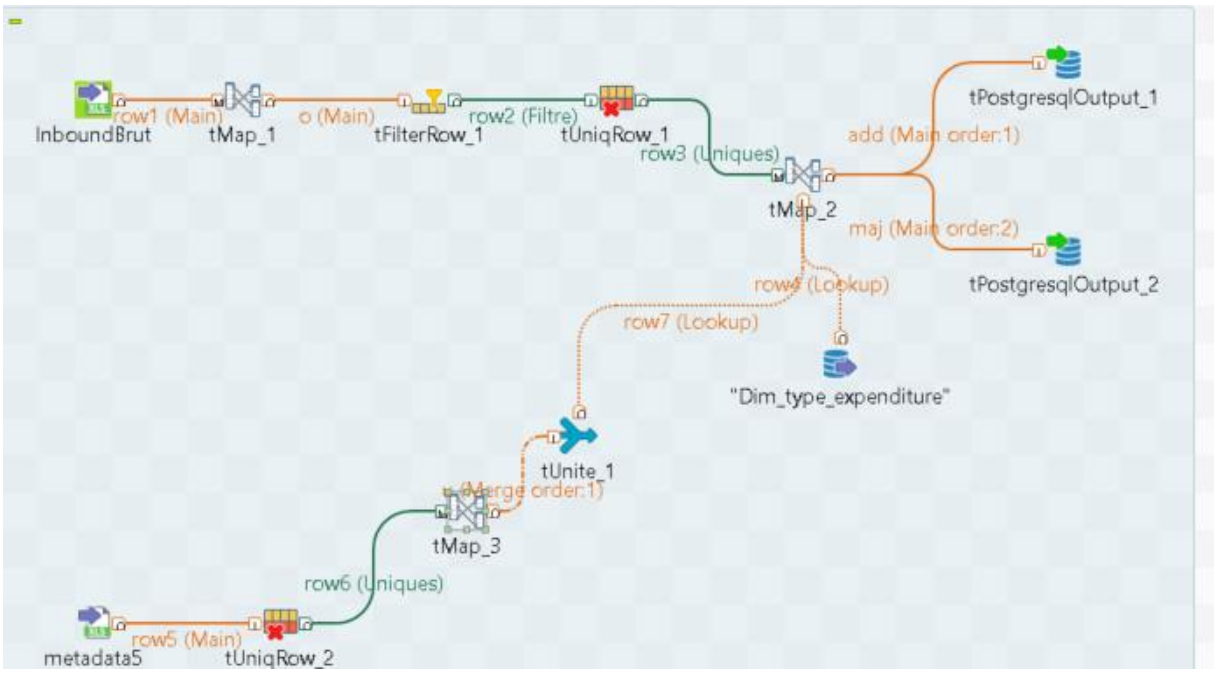


Figure 45: Expenditure dimension

⇨ The result in Postgresql

	Id_type_exp [PK] integer	type_expend character var	budget integer
1	1	Travel	2500
2	2	Travel	1500
3	3	Passenger t	2500
4	4	Passenger t	1500
*			

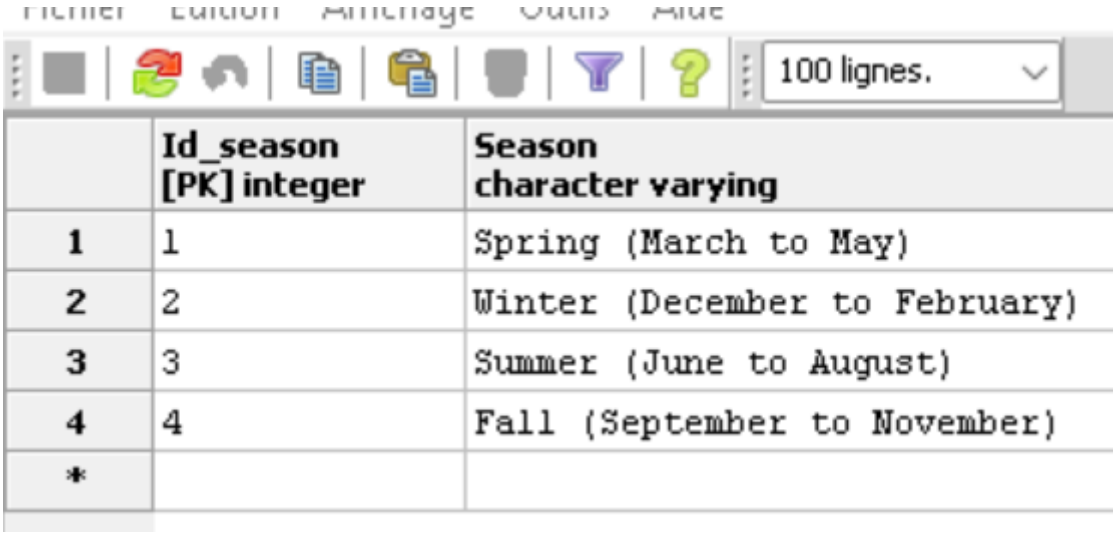
Figure 46: Expenditure Postgresql

3.2.External data

External data is data that is not generated by one of the components of the company's internal information system, or that is not entered internally. There is no major difference between internal data and external data. Simply, the company has less control over the external data when it is entered or generated by an external application.

So, for our project, after the integration of internal data with Talend, we were asked to reintegrate with external data complementary to the already existing data. And these are the data we decided to add so that we would be able to use them for visualizations and the prediction phase.

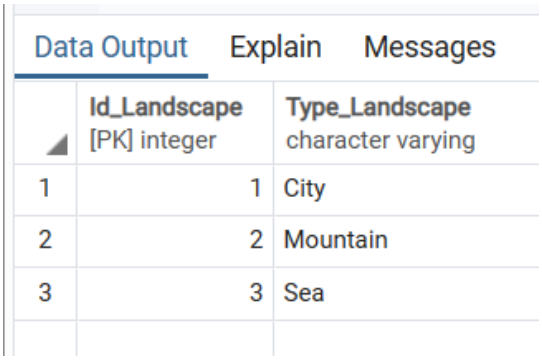
3.2.1. Season dimension



	Id_season [PK] integer	Season character varying
1	1	Spring (March to May)
2	2	Winter (December to February)
3	3	Summer (June to August)
4	4	Fall (September to November)
*		

Figure 47: Season Postgresql

3.2.2. Landscape dimension



	Id_Landscape [PK] integer	Type_Landscape character varying
1	1	City
2	2	Mountain
3	3	Sea

Figure 48:Landscape postgresql

3.2.3. Type transport dimension

Data Output | Explain | Messages

	id_transport [PK] integer	type_transport character varying
1	1	Plane
2	2	Boat
3	3	Car

Figure 49: Type transport PostgreSQL

3.2.4. Profil dimension

Data Output | Explain | Messages | Notifications

Id_Profil [PK] integer	Age character varying (100)	Gender character varying (100)	Profession character varying (100)	Country_origin character varying (100)	Budget integer	Last_destination character varying (100)	Best_destination character varying (100)	travelin charact
71	18 - 25 years old	Female	Student	Tunisia	3000	Europe	Portugal	with
72	25 - 35 years old	Women	Professional	Allemagne	3000	Europe	Italy	alon
73	18 - 25 years old	Male	Professional	Pakistan	2000	Asia	Saudi Arabia	alon
74	25 - 35 years old	Male	Professional	Algeria	5000	Europe	France	alon
75	35 - 65 years old	Male	Professional	France	2600	Asia	Tunisia	alon
76	35 - 65 years old	Female	Independent	Tunisia	2000	Africa	Algeria	alon
77	25 - 35 years old	Female	Professional	Tunisia	3000	Oceania	Denmark	alon
78	35 - 65 years old	Male	Professional	Iraq	1200	Europe	Bahrain	alon
79	25 - 35 years old	Male	Independent	Tunisia	3000	Africa	Libya	alon
80	25 - 35 years old	Male	Student	Tunisia	3000	Europe	Vietnam	alon

Figure 50: Profil postgresql

3.2.5. Tourism dimension

	Id_Tourism [PK] integer	Tourism character varying (100)	interests character varying (100)
1	1	Adventure tourism	adventure
2	2	Adventure tourism	sports
3	3	Business tourism	business meetings
4	4	Business tourism	networking
5	5	Cultural tourism	history
6	6	Cultural tourism	art
7	7	Cultural tourism	cultural
8	8	Education tourism	academic programs
9	9	Education tourism	language courses
10	10	Gastronomic tourism	Gastronomy
11	11	Gastronomic tourism	Food
12	12	Medical tourism	Esthetic
13	13	Nature tourism	Nature

Figure 51: Tourism postgresql

4. Fact table

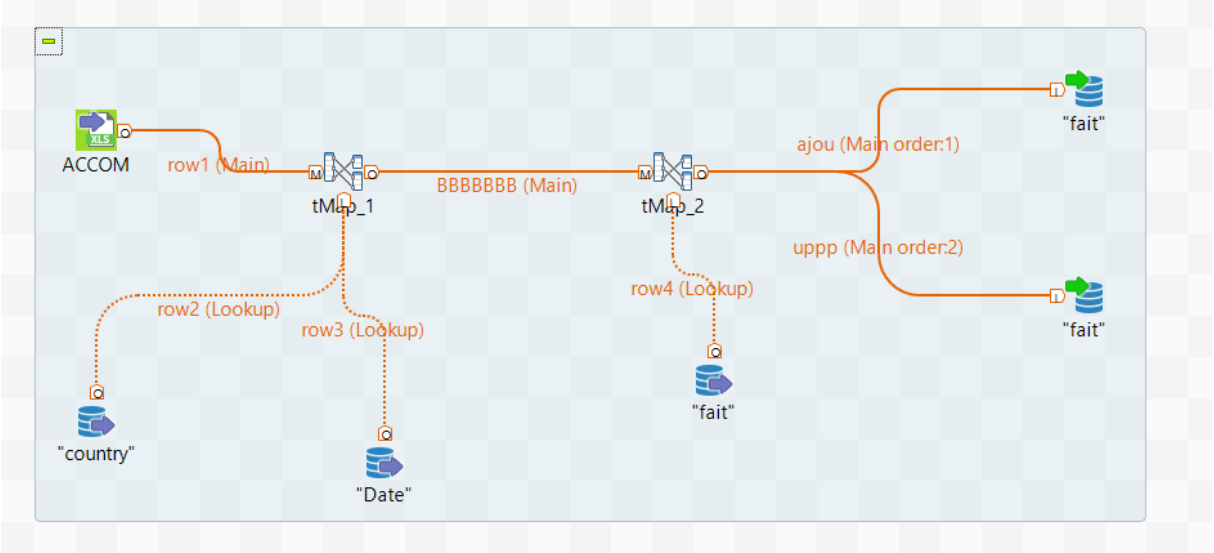


Figure 52: Job1 for accommodation

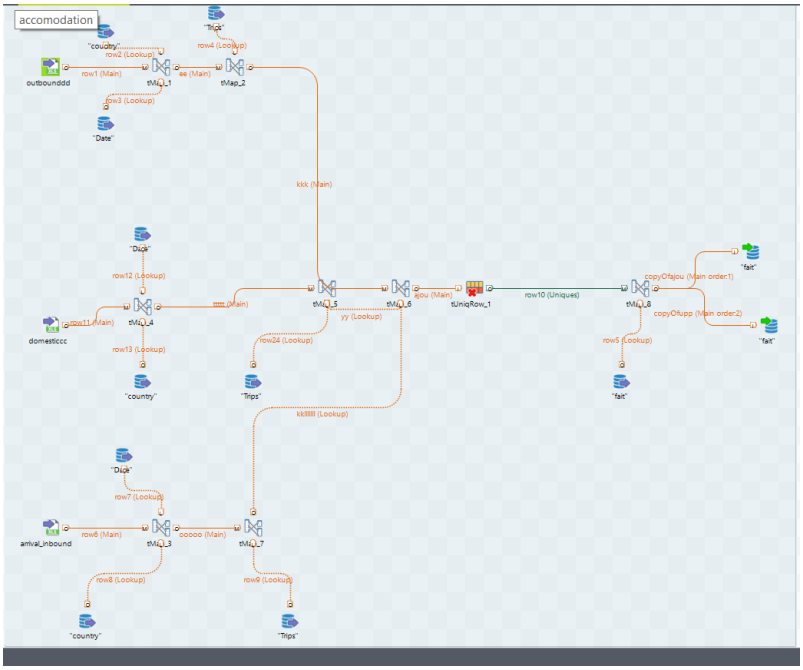


Figure 53: Job2 : Accommodation type

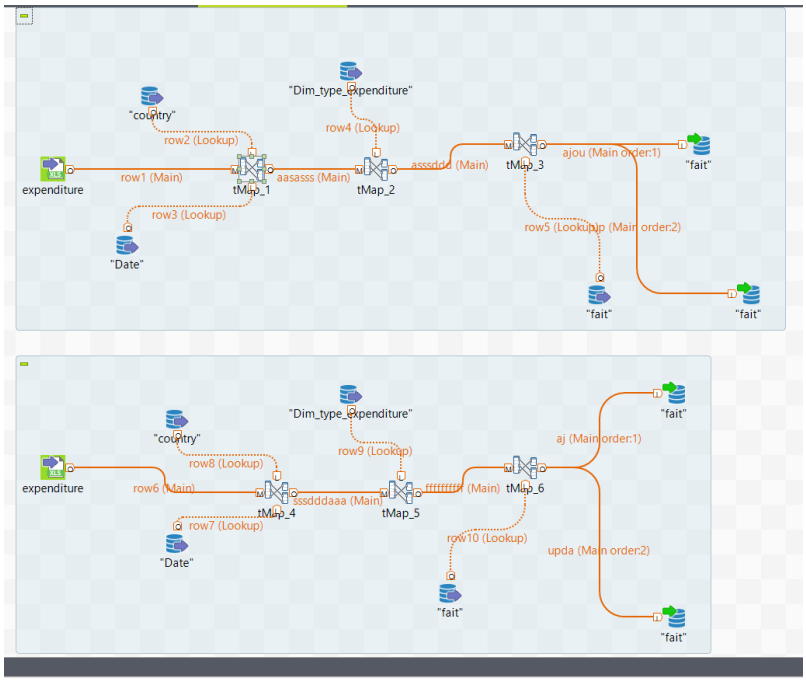


Figure 54: Job 3 : Expenditure

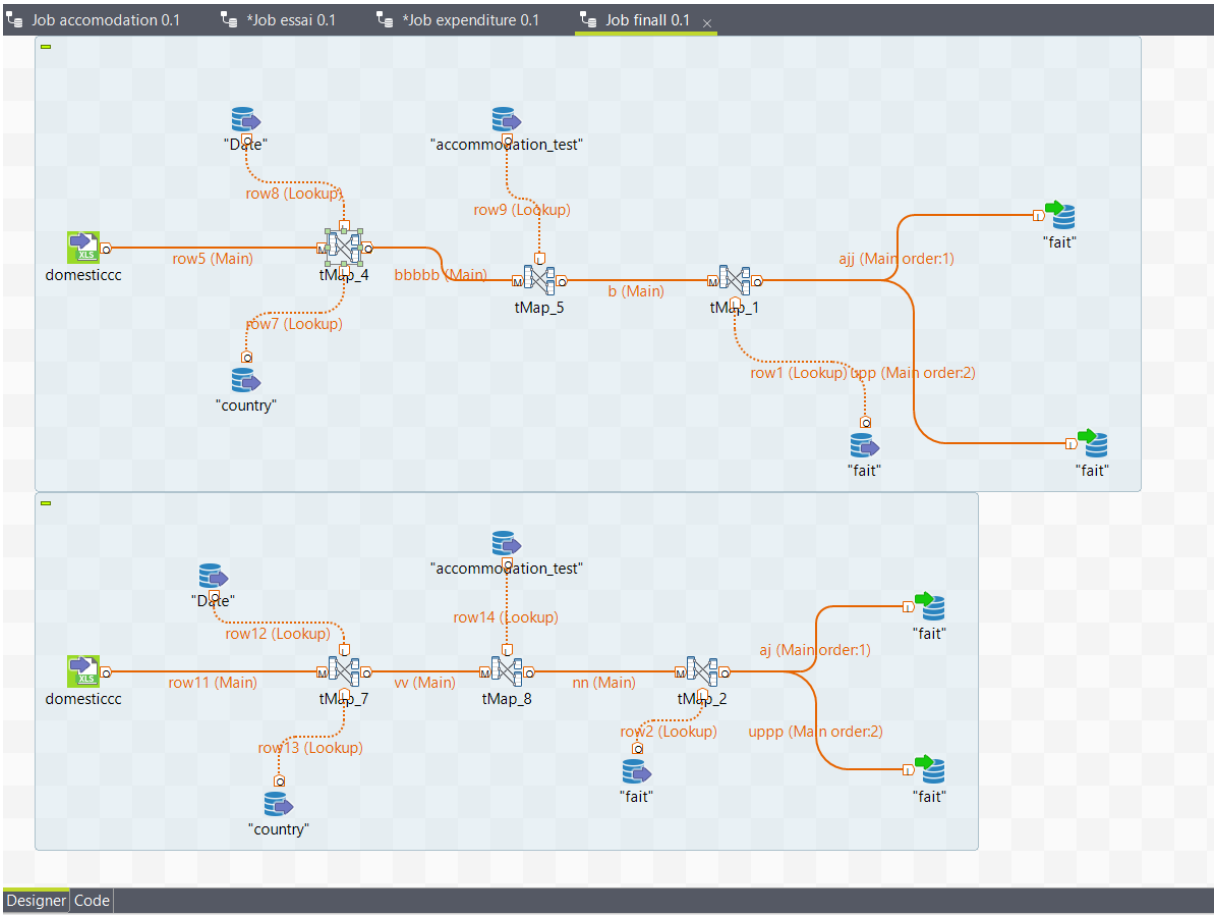


Figure 55: Job 4: Tourism Domestic

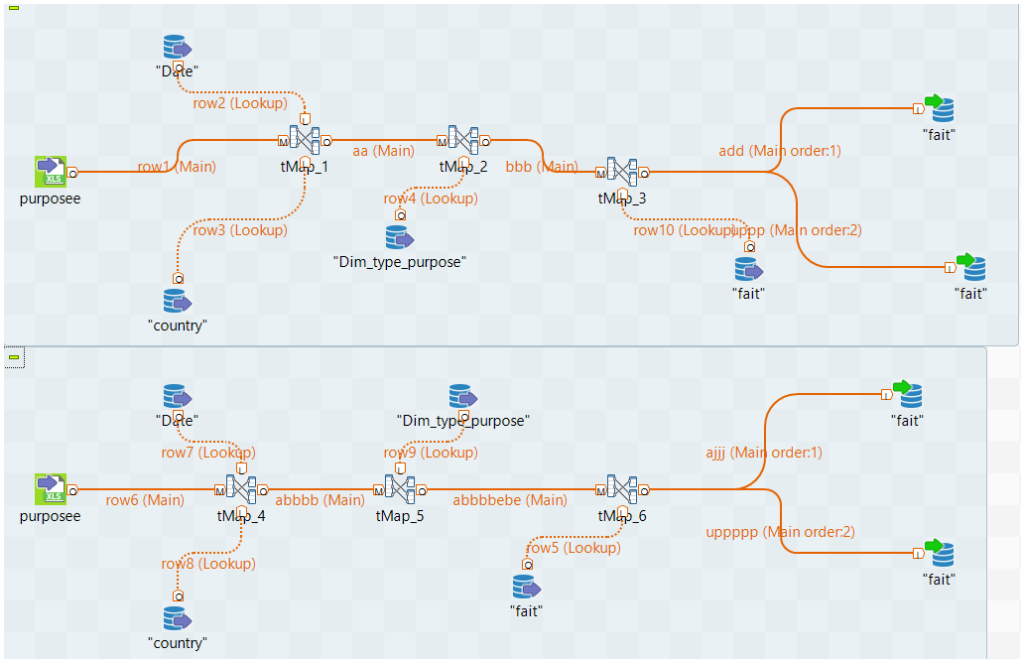


Figure 56: Job 5: Type main purpose

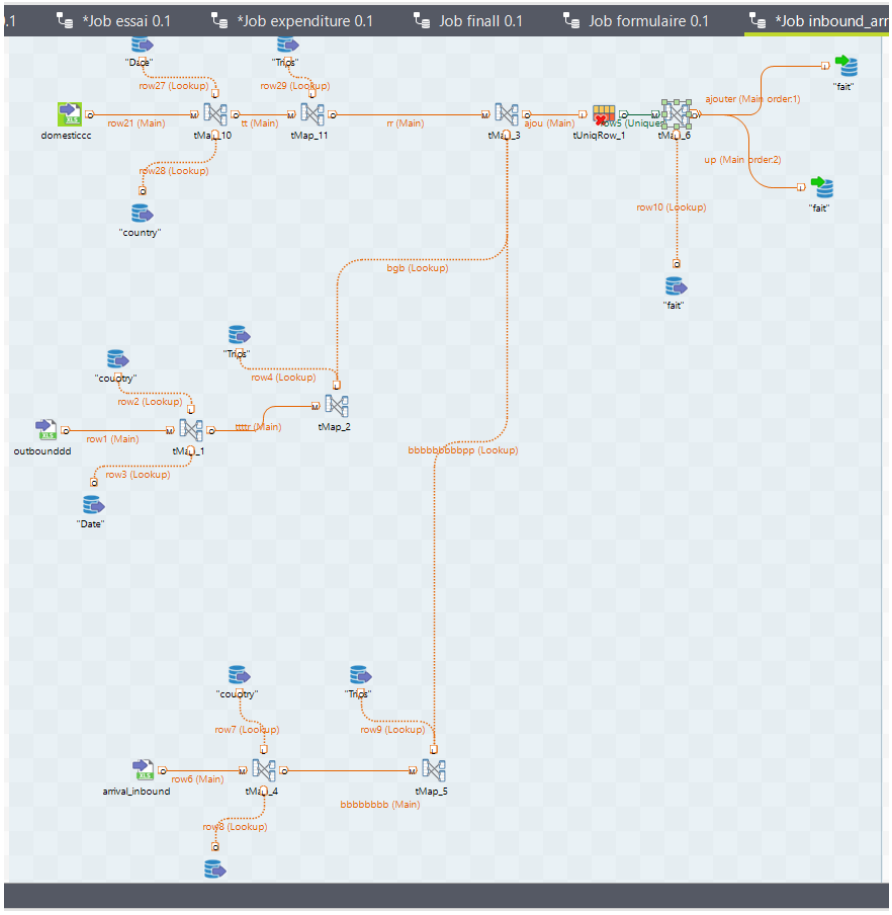


Figure 57: Job 6: Total arrivals and departures: tourism domestic

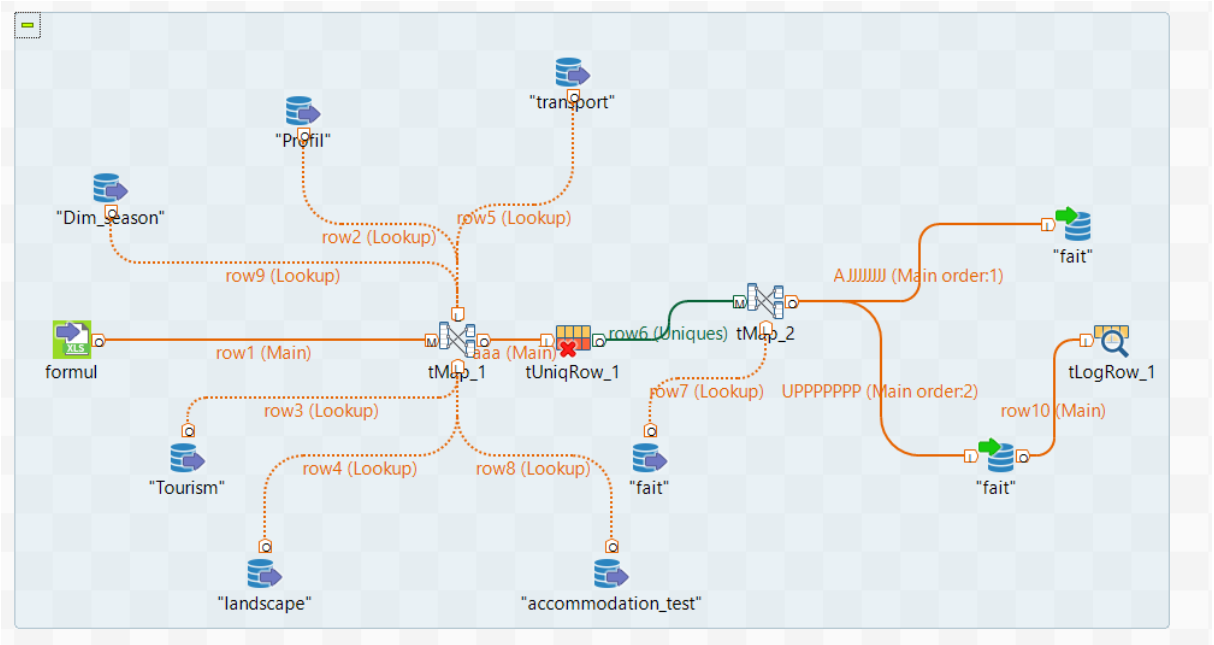


Figure 58: Job 7: External data

⇨ The result in Postgresql

	fkcountry [PK] integer	fk_trips [PK] integer	total_trips double precisio	fk_date [PK] integer	Number_of_est double precisio	Number_of_roc double precisio	Number_of_be double precisio	Occupancy_rat double precisio	Occupancy_rat double precisio	Average_lengt double precisio	Nb_accomodati double precisio	Nb_hotelsande double precisio	fk_accomodati [PK] integer	Nb_departures double precisio	fk_purpose [PK] integer	Nb_purpose double precisio	fk_expenditure [PK] integer	Nb_expenditur double precisio	Nb_a doub
116	2	0		102	486	2283	4532	1.559999942	9.600000381	1.870000004			0		0		0		
117	2	0		102									0		0		1	1693	
118	2	0		102									0		0		2	128	
119	2	0		102									0		1	370	0		
120	2	0		102									0		2	100	0		
121	2	0		102								135	1		0		0		
122	2	0		102								389	2		0		0		
123	2	0		103	487	2337	4956	1.600000023	9.600000381	1.870000004			0		0		0		
124	2	0		103									0		0		1	1943	
125	2	0		103									0		0		2	107	
126	2	0		103									0		1	470	0		
127	2	0		103									0		2	88	0		
128	2	0		103								131	1		0		0		
129	2	0		103								400	2		0		0		
130	2	0		104	488	2376	5539	1.740000005	9.600000381	1.870000004			0		0		0		
131	2	0		104									0		0		1	2186	
132	2	0		104									0		0		2	120	
133	2	0		104									0		1	582	0		
134	2	0		104									0		2	63	0		
135	2	0		104								42	1		0		0		
136	2	0		104								153	2		0		0		
137	2	0		105	489	4341	7583	2	9.600000381	1.870000004			0		0		0		
138	2	0		105									0		0		1	2329	
139	2	0		105									0		0		2	129	
140	2	0		105									0		1	680	0		
141	2	0		105									0		2	68	0		
142	2	0		105								56	1		0		0		
143	2	0		105								214	2		0		0		

Figure 60: Fact table Postgresql

fk_typetransport integer	fk_profil integer	fk_type_tourism integer	fk_saison integer	fk_landscape integer
1	143	2	3	2
1	144	2	2	1
3	145	1	3	2
1	5	1	3	1
1	147	1	4	1
1	152	1	1	2
1	23	3	3	3
2	25	6	2	3
1	26	7	3	1
1	27	6	4	1
1	28	8	3	1
1	29	8	2	1
3	30	10	3	1

Figure 59:Fact table: postgresql

Conclusion

We were able to complete the missing data, we implemented our different dimensions and fact tables. Finally, we added some external data that we needed and that will help us with the next phases of our project.

Chapter 4: Data Mining

Introduction

Data mining is the process of discovering hidden patterns and relationships in large datasets using statistical and computational methods. It involves identifying and extracting valuable insights from vast amounts of data, which can then be used to make more informed business decisions.

I. Business Understanding

Business understanding is a critical step in the data mining process, as it involves defining the problem or opportunity that the organization is seeking to address through data mining. It involves understanding the business context, identifying the goals and objectives, and defining the scope of the project.

Some of the key activities involved in the business understanding phase include:

Identifying the goals and objectives: This involves defining the key performance indicators (KPIs) and metrics that will be used to measure the success of the project.

Our Data Mining objectives are :

- Group the clients according to their travel experiences.
- Analyse the relationships between the number of hotels and establishments, the number of rooms, the room occupancy rate, and the total number of arrivals.
- Analysis of the performance of tourist destinations.
- Predict the total number of arrivals for a country
- Predict the profile of arrivals.

II. Data Understanding

Next is the Data Understanding phase. Adding to the foundation of Business understanding, it drives the focus to identify, collect, and analyze the data sets that can help you accomplish the project goals.

1. Descriptive

Descriptive analysis is the term given to data analysis that describes and summarizes historical data in a meaningful way so that, for example, insights emerge.

1.1.Profiling

	Description	***Variable Type***	
		Quantitative	***Qualitative***
Best destination	Best destination		O
Type landscape	type of landscape		O

Tableau 1: Profiling: variable type

1.1.1. Algorithm used

The KMeans[1] clustering method is used to group similar profiles based on their characteristics. The elbow method is used to determine the optimal number of clusters, which is then used to cluster the data and add cluster labels to the original dataset.

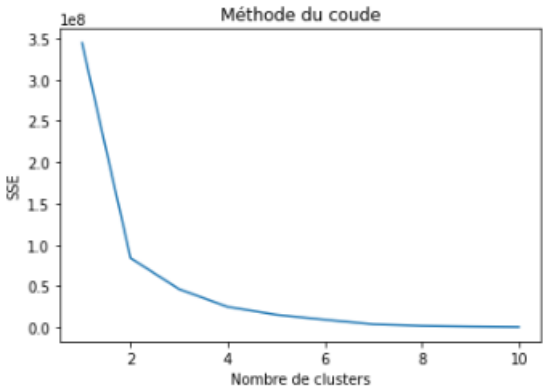


Figure 60: Elbow method

Based on the visualization, it appears that the KMeans algorithm has generated four clusters of tourist profiles based on the variables "Best destination" and "Type of landscape".

Each cluster is represented by a different color, and you can observe how the points are distributed and separated from one another within each cluster. This can provide insight into the different types of tourist profiles that exist and the factors that may influence their travel preferences.

Figure 61.Profiling: Result of k_means algorithm

1.2.Attractivity

	Description	***Variable Type***	
		Quantitative	***Qualitative***
Total_arrivals	***The total of arrivals by country***	✖	
Nb_establishments	***the number of hotels and establishments***	✖	
Nb_rooms	***The number of rooms***	✖	
Occupancy_rate_rooms	***The occupancy rate of rooms***	✖	

Tableau 2: Attractivity: Type of variable

1.2.1. Algorithm used

To manage our Data mining part, we used different analysis algorithms. For the Attractivity part, we barely used the Association rule[2] method to describe our data, which consists used to discover relationships between different variables in large datasets. It involves identifying frequently occurring patterns, correlations, and associations among variables.

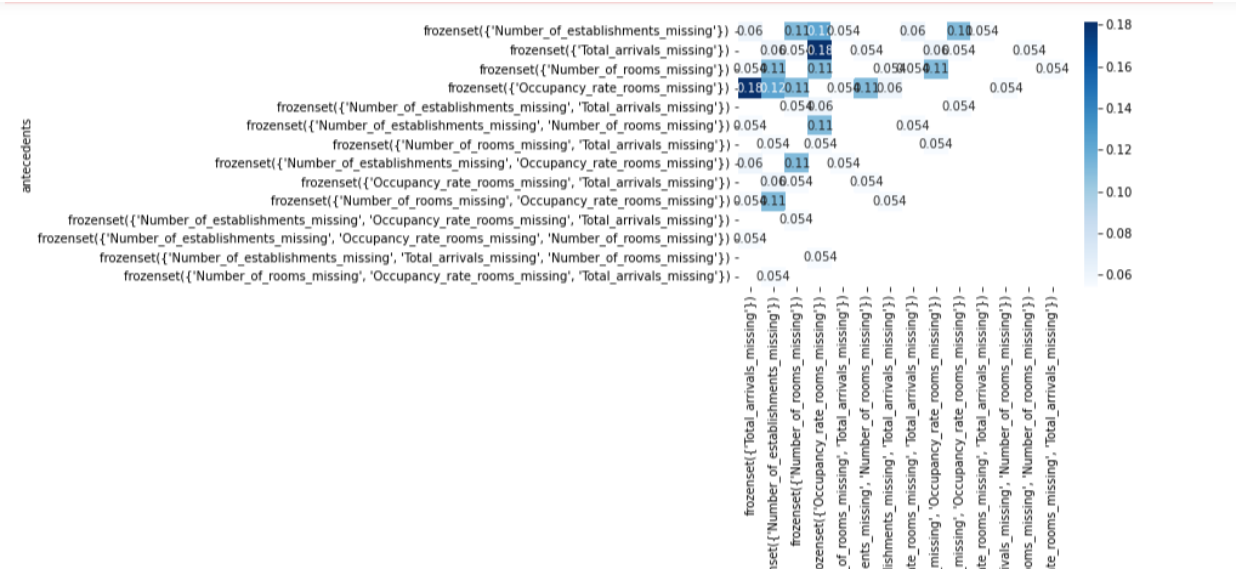


Figure 62.Result of algorithme

As we can see we have a fort correlation between occupancy rate rooms and total arrivals, we have also a relation between the number of rooms and number of establishments. We observe that if the number of establishments increases, then the number of rooms increases, and consequently the number of arrivals increases, and vice versa

1.3.Performance

	Description	***Variable Type***	
		Quantitative	***Qualitative***
Type tourism	***The type of tourism***		O
Interests traveling	***The interest in travel***		O
Type landscape	***The type of landscape***		O

Tableau 3: the performance : type variable

1.3.1. Algorithm used

For the satisfaction part, we barely used the ACP[3] method to describe our data, It is a statistical technique used in data analysis to reduce the dimensionality of a dataset by identifying and removing correlated variables and creating a new set of uncorrelated variables called principal components.

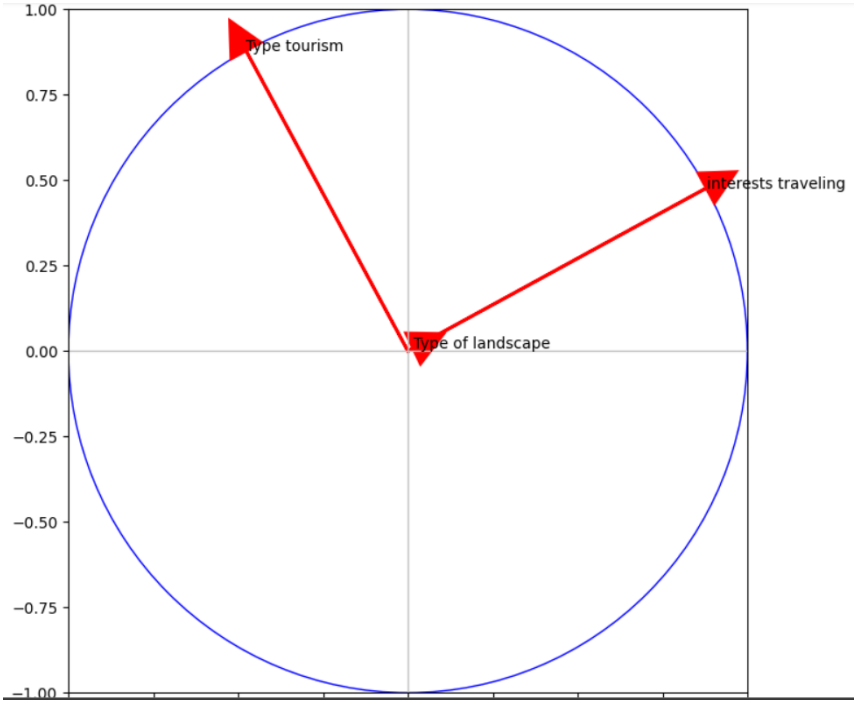


Figure 63.1st result

In the correlation circle plot, when the arrows representing the variables "interests traveling" and "Type of landscape" have a similar trajectory or point in the same direction, it suggests a relationship or association between these two variables. Specifically, it indicates that individuals who have a preference for a certain type of landscape also tend to have similar interests in traveling.

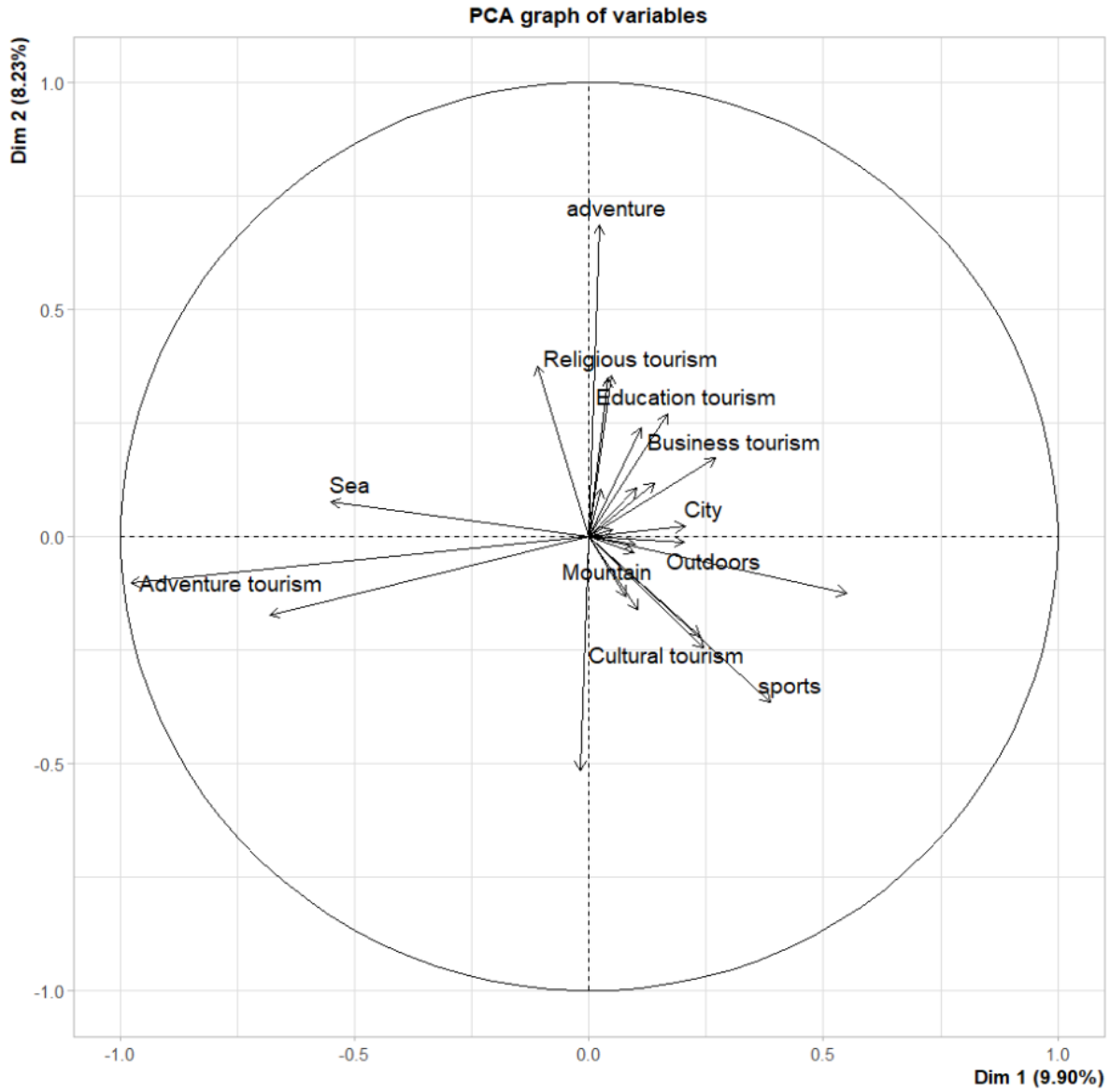


Figure 64.2nd result

For example, if the arrows for "interests traveling" and "Type of landscape" both point towards a specific direction, such as "adventure" or "sports," it suggests that individuals who enjoy adventurous or sports activities are also more likely to prefer landscapes associated with those activities (Mountain). Other individuals who are interested in religious tourism, business tourism, or education tourism are more likely to prefer city landscapes.

This correlation or association between "Type tourism", "interests traveling" and "Type landscape" can provide valuable insights for various purposes. It can help design targeted tourism or travel packages that align with individuals' specific preferences and interests. It can also be useful for segmenting or profiling individuals based on their preferences and tailoring marketing strategies accordingly.

2. Predictive

Predictive analytics encompasses a variety of techniques from statistics, knowledge extraction from data, and game theory that analyze present and past facts to make predictive assumptions about future events.

2.1.Profiling

2.1.1. Algorithm used

To predict the satisfaction of our future customers, we used the Logistic Regression algorithm[4], and the result seems pretty logical according to our descriptive part.

Entrée [8]:
```
# split the data into training and testing sets
X = df_encoded.drop("interests traveling", axis=1)
y = df_encoded["interests traveling"]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# train a logistic regression model
clf = LogisticRegression()
clf.fit(X_train, y_train)

# make predictions on the test set
y_pred = clf.predict(X_test)
```

Entrée [9]:
```
# evaluate the model's accuracy
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)
```

```
Accuracy: 0.47619047619047616
```

Figure 65.train a logistic regression model

The code creates a confusion matrix to evaluate the performance of a classification model. The categories variable contains an array of all unique categories from the actual and predicted values.

Entrée [11]:
```
# get all unique categories from the actual and predicted values
categories = np.unique(np.concatenate((y_test, y_pred)))
# create an empty confusion matrix with rows and columns labeled with the categories
cm = pd.DataFrame(data=np.zeros((len(categories), len(categories))), index=categories, columns=categori

# populate the confusion matrix with the actual and predicted values
for i in range(len(y_test)):
    cm.loc[y_test.iloc[i], y_pred[i]] += 1

# plot the confusion matrix
sns.heatmap(cm, annot=True, cmap="Blues", fmt="g")
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.show()
```

Figure 66. plot

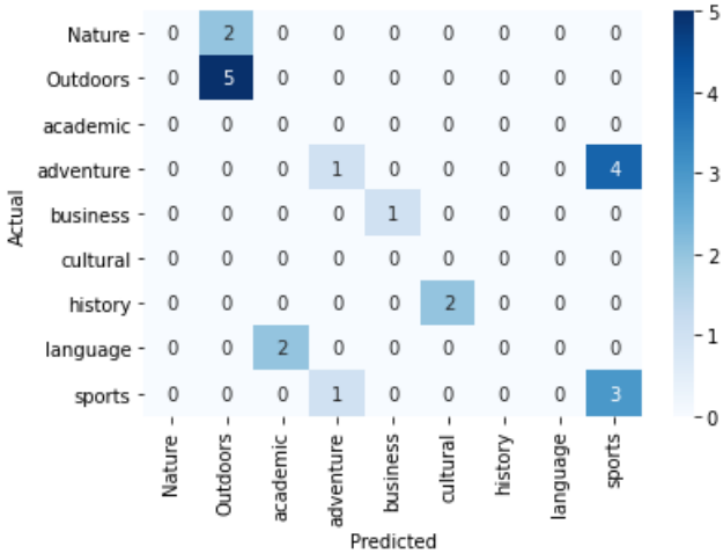


Figure 67.Result of logistic regression

Prediction on a sample Data Point a using Logistic regression Model

```
# create a sample data point
sample = pd.DataFrame({
    "Type tourism": [1],
    "Season": [0]
})

# make a prediction on the sample data point
prediction = model.predict(sample)
prediction_label = encoder.inverse_transform(prediction)

print("Prediction:", prediction_label)
```

```
Prediction: ['business']
```

Figure 68.Result of prediction

2.2.Total Arrivals

2.2.1. Algorithm used

Entrée [9]:

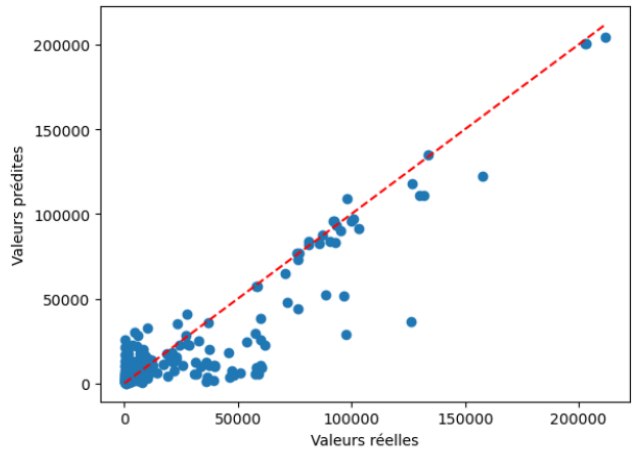
```python
# Importation des librairies nécessaires
import pandas as pd
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.decomposition import PCA
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.metrics import mean_squared_error, r2_score
from math import sqrt
from xgboost import XGBRegressor
import matplotlib.pyplot as plt

# Charger les données
data = pd.read_excel("C:/Users/user/Desktop/fichier_final/2020.xlsx", header=0)
df = data.dropna()

# Encodage de la variable catégorielle "country"
le = LabelEncoder()
df.loc[:, 'country'] = le.fit_transform(df.loc[:, 'country'])

# Séparation des variables explicatives et de la variable cible
X = df.drop(['Total_arrivals', 'Average_length_of_stay'], axis=1)
y = df['Total_arrivals']

# Normalisation des données
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Réduction de dimension avec ACP
pca = PCA(n_components=3)
X_pca = pca.fit_transform(X_scaled)

# Séparation des données en données d'entraînement et de test
X_train, X_test, y_train, y_test = train_test_split(X_pca, y,
test_size=0.2, random_state=42)

# Entraînement du modèle XGBoost avec validation croisée
param_grid = {'learning_rate': [0.1, 0.01], 'max_depth': [3, 5, 7],
'n_estimators': [50, 100, 150]}
model = GridSearchCV(XGBRegressor(), param_grid, cv=5)
model.fit(X_train, y_train)

#affiche les best param
print("best paramters found", model.best_params_)

```

Figure 70.Steps of the model

```python
# Évaluation du modèle sur les données de test
y_pred = model.predict(X_test)
mse = mean_squared_error(y_test, y_pred)
rmse = sqrt(mse)
r2 = r2_score(y_test, y_pred)
print('MSE:', mse)
print('RMSE:', rmse)
print('R^2 score:', r2)

# Courbe de prédiction pour le modèle XGBoost
plt.scatter(y_test, y_pred)
plt.plot([min(y_test), max(y_test)], [min(y_test), max(y_test)], '--',
color='red')
plt.xlabel('Valeurs réelles')
plt.ylabel('Valeurs prédites')
plt.show()

# Prédiction pour un pays et une année donnée
country = input("Enter the country: ")
year = int(input("Enter the year: "))
num_establishments = int(input("Enter the number of establishments: "))
num_rooms = int(input("Enter the number of rooms: "))
occupancy_rate = float(input("Enter the occupancy rate: "))

new_data = pd.DataFrame({'country': [country], 'year': [year],
'Number_of_establishments': [num_establishments], 'Number_of_rooms':
[num_rooms], 'Occupancy_rate_rooms': [occupancy_rate]})

# Vérification si le nouveau pays est dans les données originales
if country in le.classes_:
    new_data['country'] = le.transform(new_data['country'])
    new_data_scaled = scaler.transform(new_data)
    new_data_pca = pca.transform(new_data_scaled)
    prediction = model.predict(new_data_pca)
    print('Predicted total arrivals:', prediction[0])
else:
    print("The country entered is not in the original data.")
```

Figure 69.Steps of the model

```
best paramters found {'learning_rate': 0.1, 'max_depth': 7, 'n_estimators': 50}
MSE: 162133602.27064177
RMSE: 12733.16937257342
R^2 score: 0.8388773731177932
```




```
Enter the country: FRANCE
Enter the year: 2022
Enter the number of establishments: 35000
Enter the number of rooms: 1100000
Enter the occupancy rate: 46
Predicted total arrivals: 90185.766
```

Figure 71.Result of prédiction

This Python script is dedicated to predicting the total number of arrivals in a given country for a specific year. It uses a machine learning approach with the XGBoost Regressor model. To start, the script imports the necessary libraries and loads the dataset from a specified location. It then preprocesses the data, which involves handling missing values, encoding the categorical 'country' variable, scaling the numerical features, and performing a Principal Component Analysis (PCA) for dimensionality reduction.

The dataset is then split into a training set and a testing set, with the 'Total_arrivals' as the target variable. The script uses the XGBoost model and GridSearchCV for hyperparameter tuning, after which it trains the model on the training data. Once the model is trained, it makes predictions on the test data and evaluates the model's performance by calculating metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the coefficient of determination, R^2.

The script then provides a visualization of the predicted versus actual values for 'Total_arrivals', and finally, it allows the user to make predictions for a specific country and year by inputting the required details.

The R^2 value of 0.87 indicates that the model explains 87% of the variability in the response data around its mean. In other words, it suggests that the model fits the data fairly

well. It's a relatively high score, meaning the model has a good predictive ability for the number of arrivals. However, it's worth noting that there's still 13% of variability that the model does not capture, so there could be room for further improvement or additional factors to consider.

Conclusion

In this chapter, we have defined the descriptive and predictive objectives of data mining, identified the type of each variable used to apply algorithms to our data, and finally, we have explored the various algorithms that we used to accomplish our objectives.

Chapter 5: Data Viz and Realization

Introduction

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps. We will be creating interactive dashboards to visualize our data to make the analysis and decision-making clearer and easier to comprehend. Here we're in the 4th phase of the CRISP-DM methodology: The Implementation.

I. DataViz

1. Working tool: Power BI

Power BI[4] is a business analytics service by Microsoft. It aims to provide interactive visualizations and business intelligence capabilities with an interface simple enough for end users to create their reports and dashboards. It offers data warehouse capabilities including data preparation, data discovery, and interactive dashboards.

Figure 72.Logo Power BI

2. Dashboards

These are the full Dashboards we've created using MS Power BI Services, we will be explaining each element in the following section. Our visualizations will be divided into seven interfaces you can access while navigating the website. First, we started with these seven interfaces with Power BI.

2.1.Overview interface

This graph allows Tourism to have a global view and summarize their information.

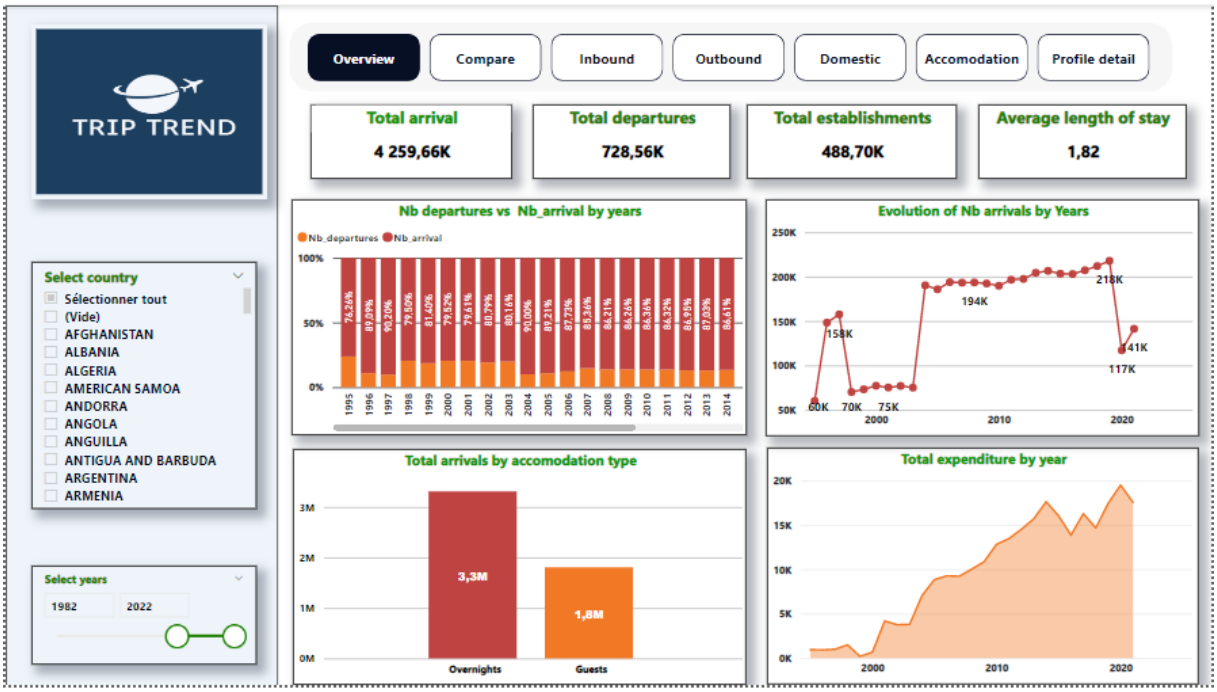


Figure 73.OverView

2.2.Comparison interface

On this page, we will find a comparison between countries regarding the number of departures, number of arrivals, total expenditures, and total domestic trips.

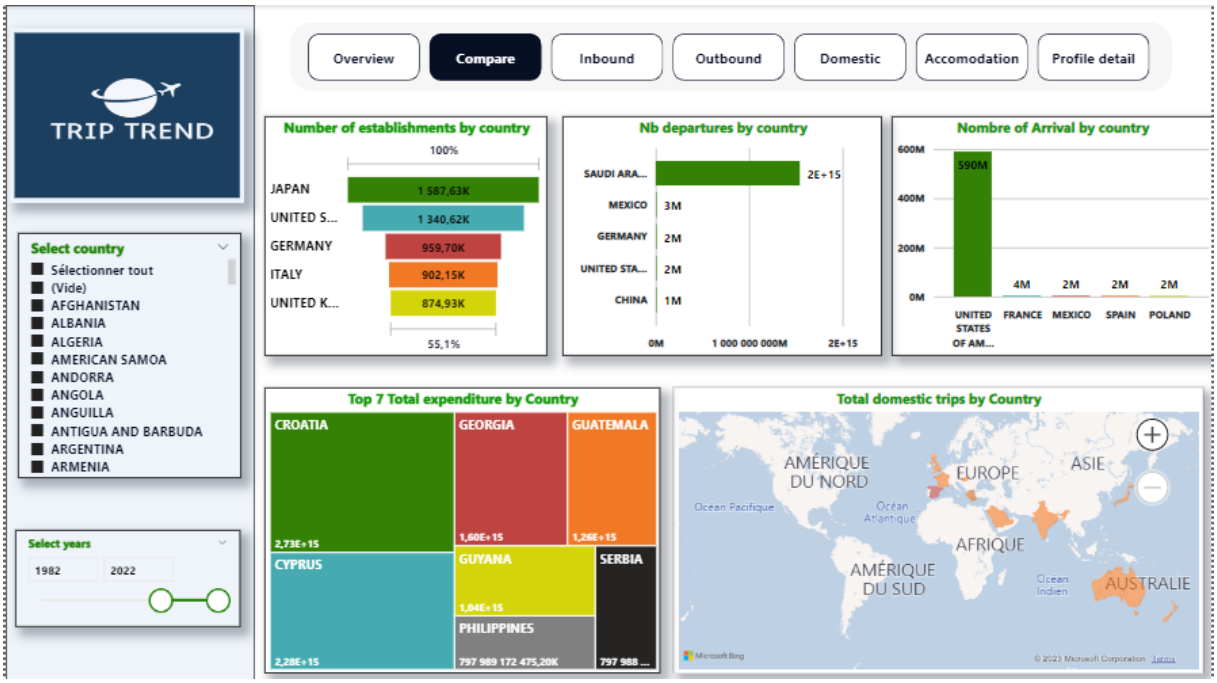


Figure 74.Compare

2.3. Inbound interface

Here we will find all information about Inbound tourism, that which the person who comes to a country.

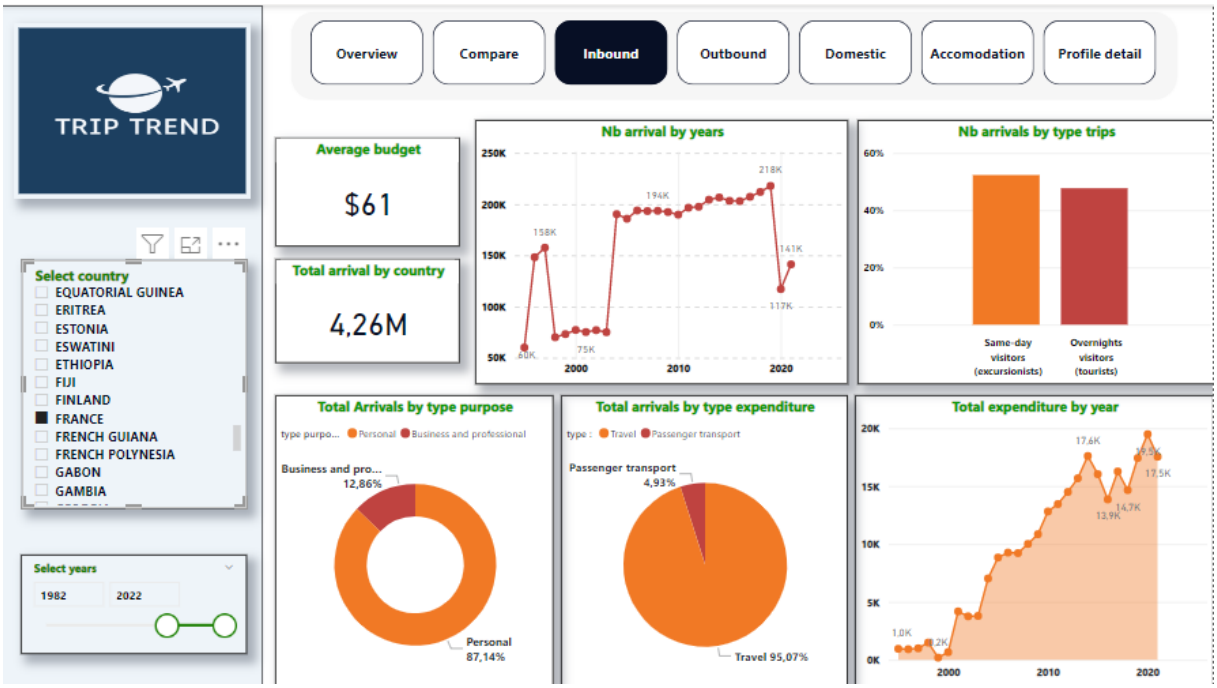


Figure 75.Inbound tourism

2.4.Outbound interface

Here we will find all information about Outbound tourism, which means the person who travels to another country.

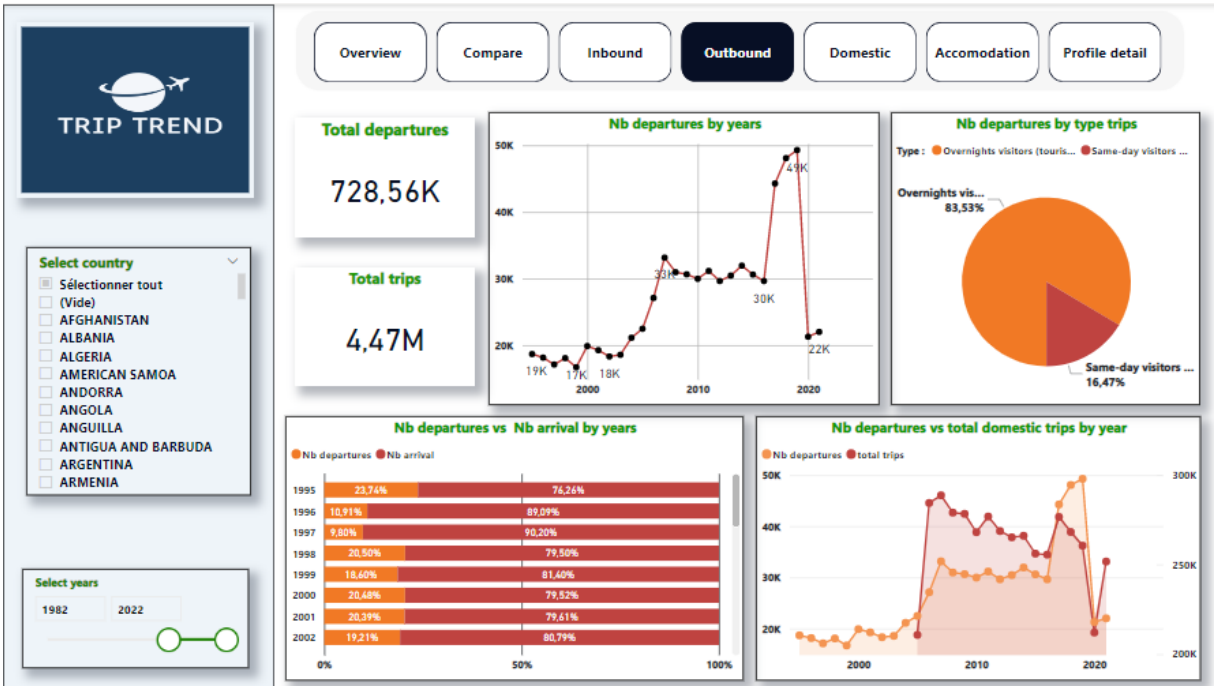


Figure 76.Outbound Tourism

2.5. Domestic interface

On this page, we will find all information about the person who travels inside her country.

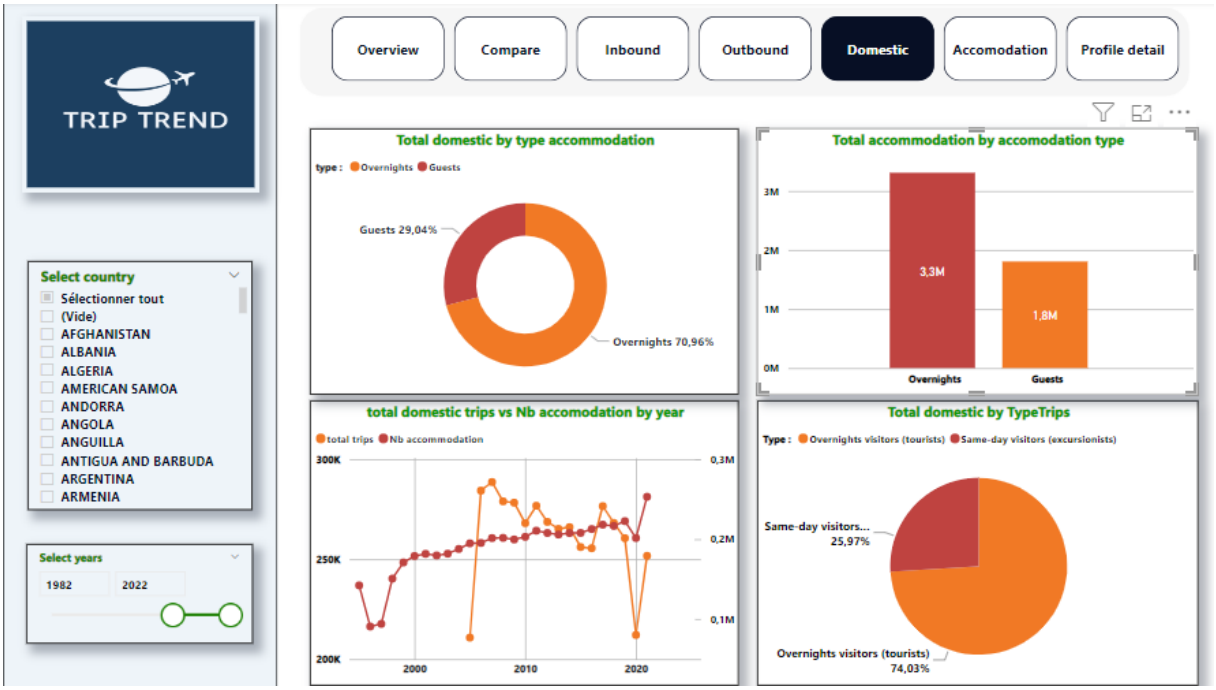


Figure 77.Domestic Tourism

2.6.Accommodation interface

On this page, we will find all information about accommodation by country and by years

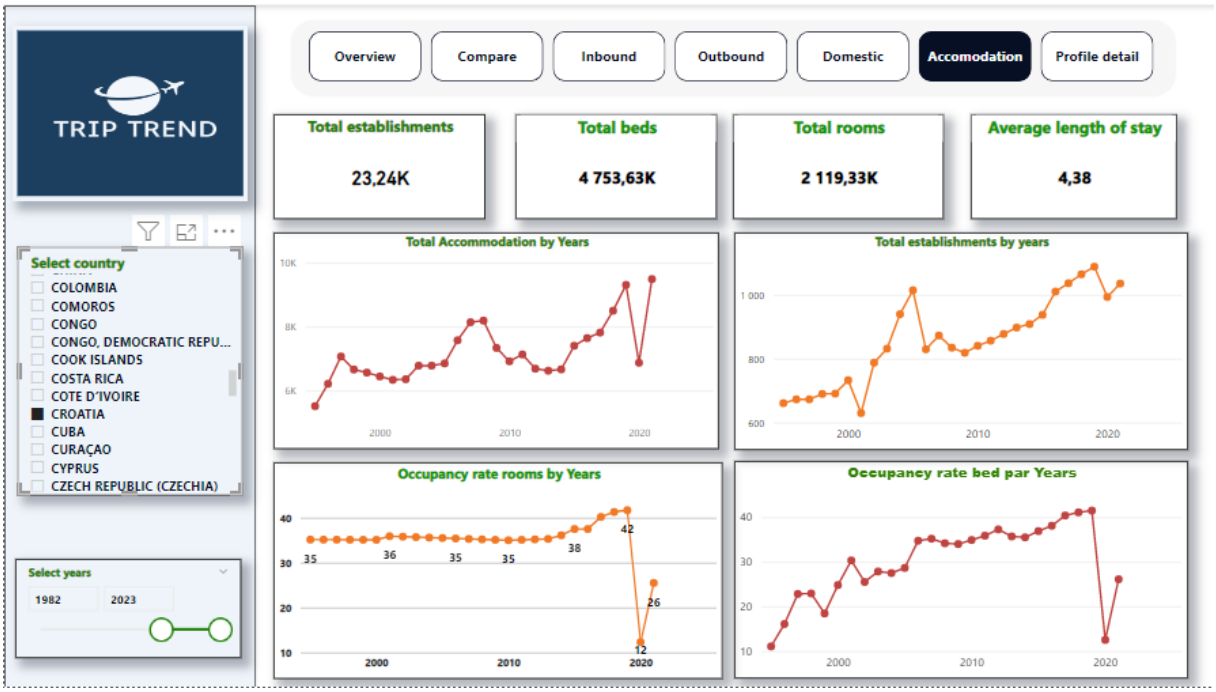


Figure 78.Accommodation

2.7.Profil interface

Here we will find all information about the profile and préférence of travel.

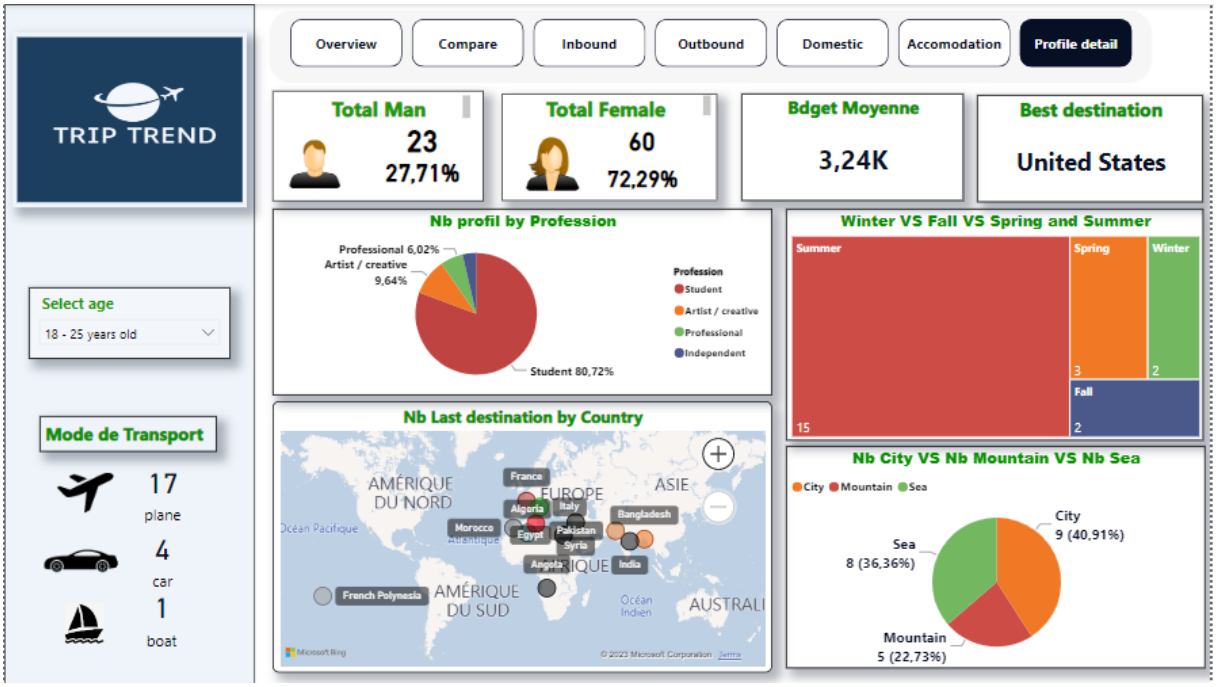


Figure 79.Profil details

II. Realization

1. Working tools

1.1. Back-end tool

Figure 80.Logo Flask

Flask[5] is a lightweight and flexible web framework for Python. It is designed to make web development simple and scalable, with a focus on providing the core functionality needed to build web applications.

1.2. Font-end tool

Figure 81.Logo Bootstrap

Bootstrap is a popular front-end framework for building responsive and mobile-first websites and web applications. It provides a collection of CSS and JavaScript components, such as grids, forms, buttons, navigation menus, and much more, that can be easily incorporated into web projects.

1.3. DataBase tool

Figure 82.Logo phpMyAdmin

phpMyAdmin is a free and open-source web-based application that provides a graphical user interface (GUI) for managing and administering MySQL or MariaDB databases. It allows users to easily perform tasks such as creating databases, managing database tables, executing SQL queries, importing and exporting data, and more, without having to use the command-line interface.

2. WebSite

2.1. Register interface

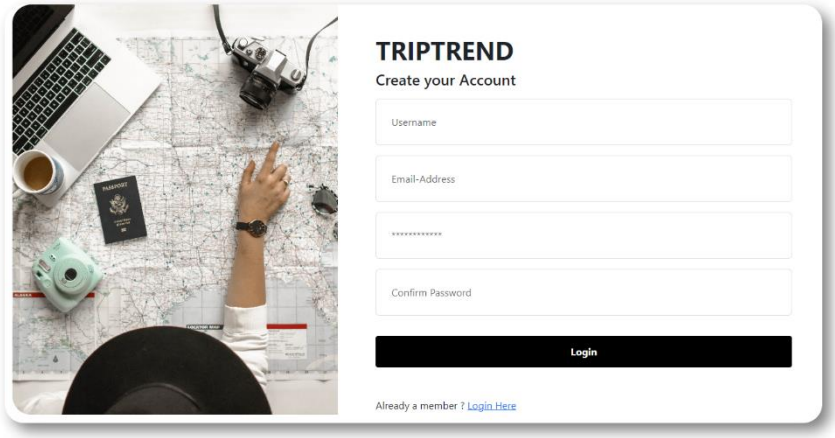


Figure 83.Register interface

2.2. Login interface

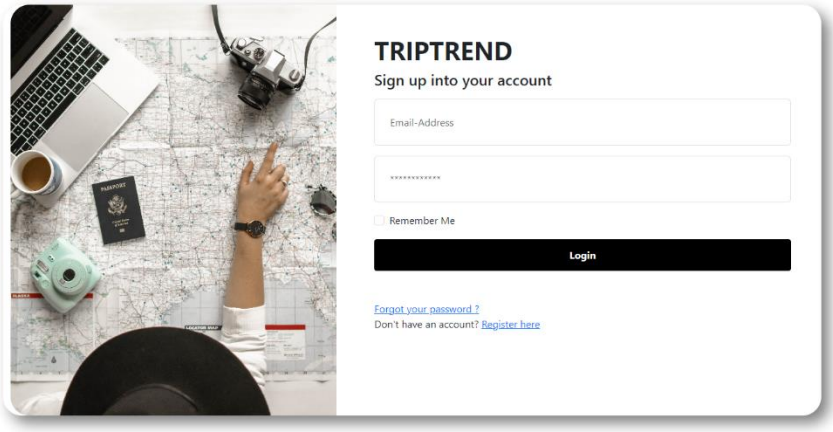


Figure 84.Login interface

2.3. Home page

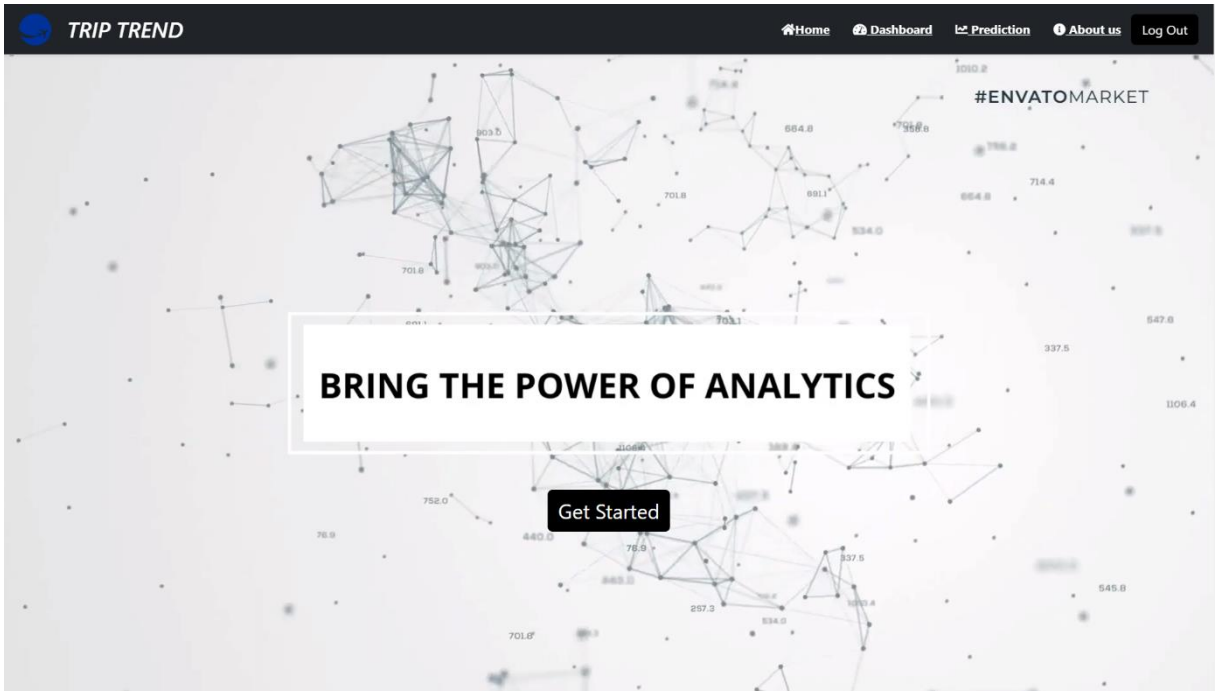


Figure 85.home page

2.4. Dashboard

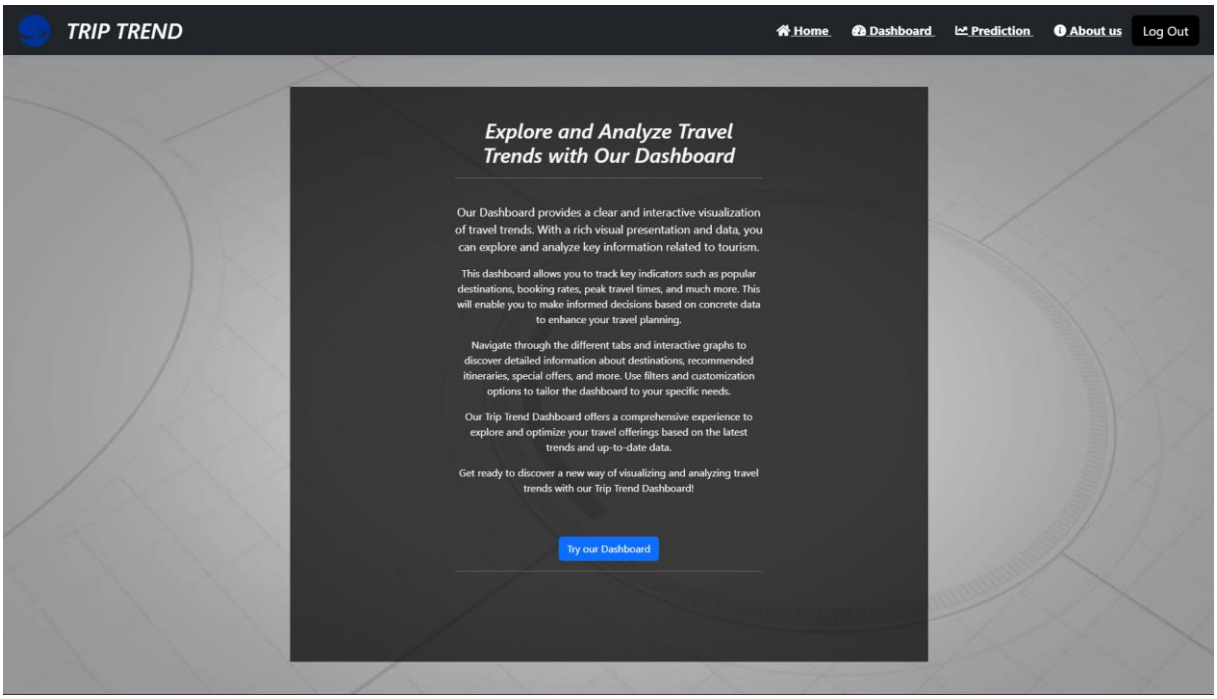


Figure 86.Dashboard page

2.5. Prediction interface

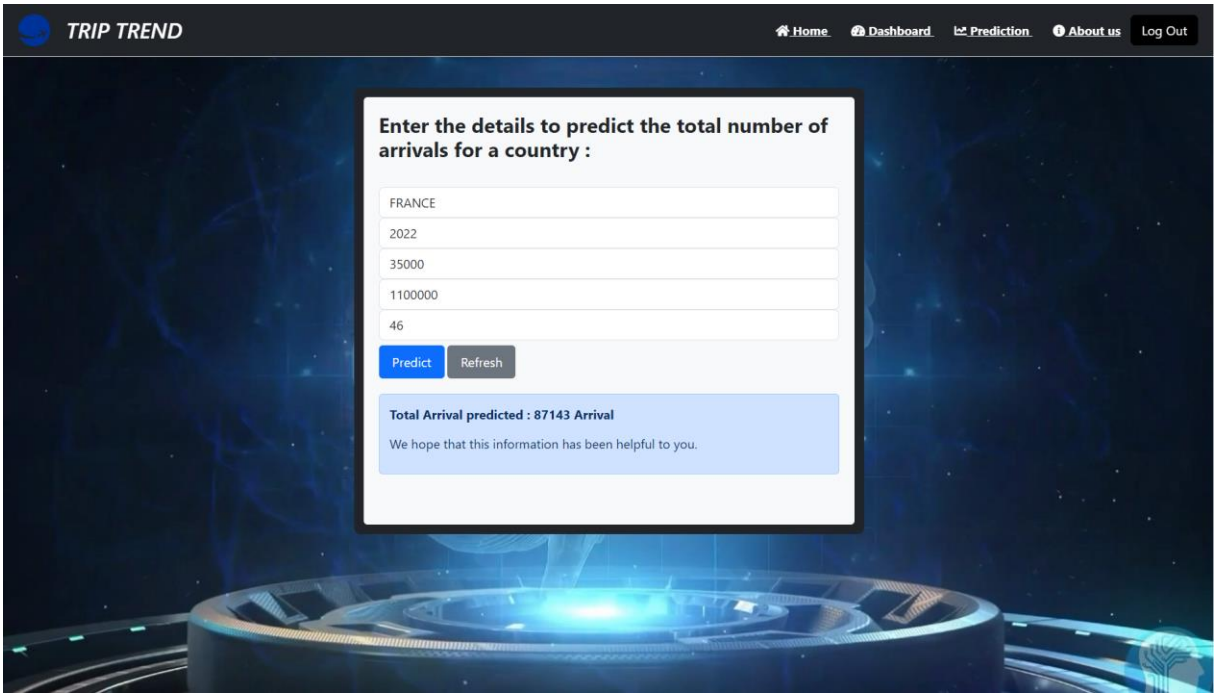


Figure 87.Prediction interface

2.6. About us

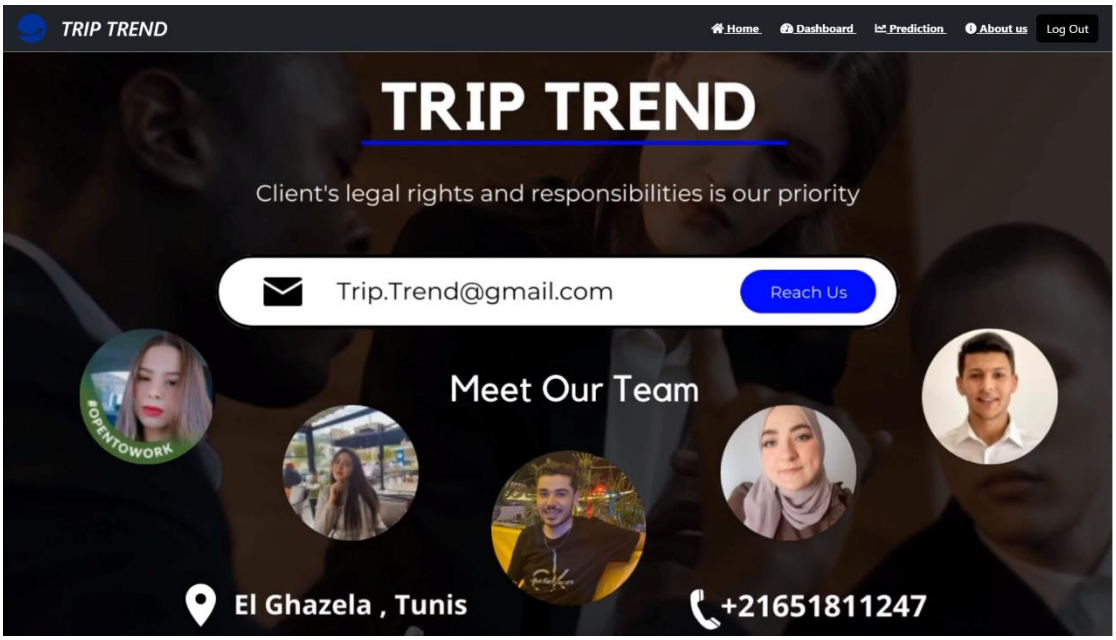


Figure 88.About us interface

Conclusion

In this last chapter, we presented our dashboard and the tool we used to create the different graphs. In the second part, we discussed everything related to our website, including the tools we used to create it and the deployment of the predictive section on our site.

General conclusion

At the end of this project, we were able to use the database provided to implement a full conceptualization of a data warehouse, we integrated our data, imputed missing information, and implemented it.

With that, we were able to present graphical representations that make the analysis and decision-making process easier for such a large sum of information. Finally, through Data mining techniques we were able to further manipulate data through modeling and clustering techniques and algorithms to get accurate predictions according to a multitude of factors.

The result of all this was implemented on a web platform with a simple UI and visual graphs that make understanding the data easy for even non-tech-savvy customers. Throughout this project, we learned to use multiple software solutions like Talend, MS Power BI...

We learned about and applied, Methodologies, used for projects like these. We were able to understand how to mine information from data sources. We got an insight into how a Business Intelligence project is operated and the different challenges we could face. Last but not least, we have got further confirmation of the importance of collaborative work and teamwork, and how to manage time, resources, and stress in a proper work environment.

[1] (https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html, s.d.)

[2] (https://perso.univ-rennes1.fr/valerie.monbet/doc/cours/IntroDM/Chapitre5.pdf, s.d.)

[3] (https://www.xlstat.com/fr/solutions/fonctionnalites/analyse-en-composantes-principales-acp, s.d.)

[4] (https://aws.amazon.com/fr/what-is/logistic-regression/, s.d.)

[5] (https://powerbi.microsoft.com/fr-fr/, s.d.)

[6] (https://flask.palletsprojects.com/en/2.3.x/, s.d.)