ANALYZING AND VISUALIZING RIDERSHIP PATTERNS IN ÎLE-DE-FRANCE RAIL NETWORK

Master 2 SSIO

AUTHORS:

- KALBOUSSI EYA
- MILOUDI MOHAMED ZAKARIA

SUPERVIZED BY:
ETIENNE CÔME

2024/2025

Table des matières

Table des figures

1 Introduction

In this project, we commence with the data collection phase, where the primary focus is on gathering Île-de-France's railway station ridership from 2017 to the frist semester of 2023. The initial steps encompass the crucial tasks of cleaning the data, detecting and rectifying missing values, and eliminating outliers to ensure the dataset's integrity. Subsequently, we delve into the Exploratory Data Analysis (EDA), unraveling the intricate tapestry of ridership patterns. This phase involves identifying overarching trends, scrutinizing seasonality and monthly variations, and meticulously assessing potential outliers that may influence the data. Following the EDA, the focus shifts to comparing ridership against established norms. This involves defining a baseline for a "normal" week and delving into deviations observed during holiday and non-holiday periods, (Christmas, Easter, National Day), vacations, the start of the school year, the impact of the coronavirus pandemic, and the Olympic Games including an assessment of the influence of vacations and school breaks on ridership patterns. Our goal is to understand the impact of these events on the number of validations, the types of transport tickets purchased, and the locations visited, comparing them with the number of validations on regular days. Finally, the journey culminates in the development of an interactive dashboard using the Shiny framework in R. This dashboard will serve as a dynamic interface, incorporating key visualizations that showcase overall ridership trends, weekly variations, and comprehensive comparisons with the established norms, providing stakeholders with an intuitive tool to monitor and comprehend ridership dynamics

2 Presentation and collection data from Open Data Île-de-France

Collecting data from Open Data Île-de-France . The first dataset [1], is the historical validation data on the rail network spanning from 2017 to 2022.
For each year, there are four files :

— **Rail network validations : Number of validations per day for the frist semester (NbValidS1)**

Variables	Description
JOUR	Date
COD_STIF_TRNS	Carrier's STIF Code
COD_STIF_ARRET	STIF code for the stop/station
LIBELLE_ARRET	Stop/Station label
ID_REFA_LDA	STIF référentiel Identifier of stop/station
CATEGORIE_TITRE	The type of Transport ticket
NB_VALD	Number of validations (input on the network)

- **Rail network validations : Hourly profiles per day type for the frist semester(ProfilFerS1)**

Variables	Description
JOUR	Date
COD_STIF_TRNS	Carrier's STIF Code
COD_STIF_ARRET	STIF code for the stop/station
LIBELLE_ARRET	Stop/Station label
ID_REFA_LDA	STIF référentiel Identifier of stop/station
CAT_JOUR	The type of Transport ticket
TRNC_HORR_60	One hour time slot
POURC_VALIDATIONS	validation pourcentage

- **Surface network validations : Number of validations per day (NbValidS2)**

Variables	Description
JOUR	Date
COD_STIF_TRNS	Carrier's STIF Code
COD_STIF_ARRET	STIF code for the stop/station
LIBELLE_ARRET	Stop/Station label
ID_REFA_LDA	STIF référentiel Identifier of stop/station
CATEGORIE_TITRE	The type of Transport ticket
NB_VALD	Number of validations (input on the network)

- **Surface network validations : Hourly profiles per day type (ProfilFerS2)**

Variables	Description
JOUR	Date
COD_STIF_TRNS	Carrier's STIF Code
COD_STIF_ARRET	STIF code for the stop/station
LIBELLE_ARRET	Stop/Station label
ID_REFA_LDA	STIF référentiel Identifier of stop/station
CAT_JOUR	The type of Transport ticket
TRNC_HORR_60	One hour time slot
POURC_VALIDATIONS	validation pourcentage

* Additionally, there is another dataset [2] which provides information on validations on the rail network, specifically the number of validations per day during the first semester of 2023. Another dataset[3], zone dataset serves as a reference for stops, detailing the stop zones.

2.1 Some data explanations :

The type of Transport ticket are :

1. JOHV : Weekday outside of School Holidays.
2. SAHV : Saturday outside of School Holidays.
3. JOVS : Weekday during School Holidays.
4. SAVS : Saturday during School Holidays.
5. DIJFP : Sunday and Public Holidays, including bridge days.

The validation data is classified by the category of transport passes :

- "IMAGINE R" : includes the annual Imagine R Scolaire and Imagine R Etudiant passes, reserved for students and apprentices, allowing unlimited travel throughout the Île-de-France region for the entire year.
- "NAVIGO" : includes the Navigo Annuel, Navigo Mois, and Navigo Semaine passes.
- "AMETHYSTE" : tallies the Améthyste passes, reserved for elderly or disabled individuals meeting certain income or status conditions and residing in Île-de-France. This annual pass allows unlimited travel on all modes of transport within the valid zones.
- "TST" : groups weekly and monthly reduced-rate passes granted to beneficiaries of the Réduction Solidarité Transport, enabling travel within selected zones across all modes of transport in Île-de-France.
- "FGT" : tallies the Forfaits Navigo Gratuité Transport, a pass that allows certain social aid beneficiaries to travel for free throughout Île-de-France.
- "AUTRE TITRE" : counts special passes.
- "NON DEFINI" : counts validations with undefined pass types (anomalies).

3 Preliminary Data Analysis

In this step, for each data from 2017 to the frist semester of 2023 we will focus on understanding the data, exploring its distribution, and identifying potential issue by using **summary() :** Provides key statistical measures for numeric variables, aiding in a quick overview of data. And using **str()** provides detailed structural information, while **head()** gives a snapshot of the initial rows for a quick overview. **dim()** uncovers the dataset's dimensions (rows and columns), while **names()** lists the column names, crucial for grasping the data structure. Moreover, **nrow()** facilitates determination of the number of rows in the dataset, and **length()** aids in assessing the object's size or complexity. These functions collectively serve as indispensable tools during the initial stages of exploring and comprehending a dataset. Both are valuable during the early stages of data exploration.

4 Cleaning data

The data cleaning process involves, first and foremost, handling missing values, outliers, and any inconsistencies. Exploring, handling missing values, and removing outliers are essential steps to prepare and visualize data effectively for analysis.

4.1 Handling missing values

In this stage, the primary focus was on addressing data-related challenges, including handling various data types and ensuring proper formatting. We delved into strategies for detecting and managing specific symbols like "NA," "ND," empty values, "null," "Inconnu," , " ?" and "NON DEFINI" within all the dataset from 2017 to the frist semstrer of 2023 . The provided code in R Markdown snippets offered a structured approach to identify, visualize, and potentially remove or replace rows and columns containing these symbols.
To assess the impact of data cleaning, we emphasized the importance of checking dataset dimensions before and after the cleaning process. Additionally, inspecting a subset of the data was recommended. This holistic approach is designed to elevate data quality, paving the way for more meaningful analyses in R.

4.2 Detecting and moving Outliers

Outliers can significantly impact data visualizations by distorting the scale and representation of patterns. In graphs, outliers may cause misleading trends or exaggerate the spread of data, affecting the overall interpretation of visualizations.

5 Exploratory Data Analysis (EDA) and comparison with Norms

For data visualization, first, we will attempt to understand the distribution of data using different plots. Afterward, we focused on comparing ticket validation during various periods : school holidays, Christmas holidays, the coronavirus pandemic period, the World Cup period, and PSG matches in the Île-de-France region. We will analyze ticket validations based on zones and validation stations to gain a thorough understanding. Our goal is to compare events to normalcy to observe and comprehend buyer behavior, as well as the types of purchased tickets (such as Navigo, Imagin'R, etc.) associated with each event or destination.

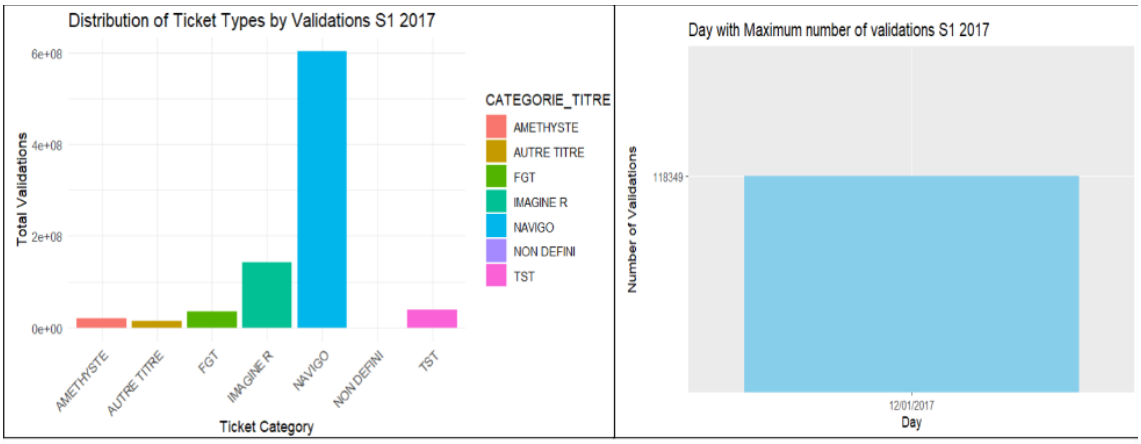
5.1 Temporal and spatial analyses, trends, and data visualization

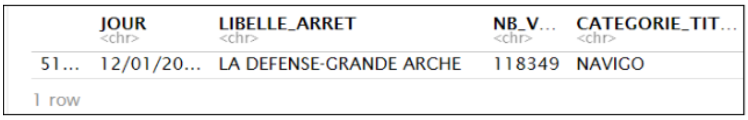
The analysis of temporal trends involves graphically representing the evolution of the number of validations over a specific period, whether it's a day, a month, or a year. This graph allows visualizing the fluctuations over time of the conducted validations. By examining this data, it becomes possible to identify the days of the week or the months displaying the highest levels of validations, providing an overview of the periods with the most activity. Moreover, the distribution of validations by stop helps locate specific points where these validations are concentrated, thus offering a detailed insight into the most significant activity zones.

5.1.1 Analysis of the 2017 dataset involves examining the number of validations for S1 and S2, as well as the profiles for S1 and S2.

— Identify the stops, days, and tickets with the highest number of validations for semester 1

During the first semester of 2017 figure 1, starting in January and ending in June, our graphical analyses revealed that "Navigo" had the highest number of validations among all transportation passes 118394. We can conclude that during this period, "Navigo" was the best-selling ticket. Additionally, the stations recording the highest number of validations were "La Défense - Grande Arche". It can be inferred that users likely have a profile of professionals, given that "La Défense"is the heart of Paris's economic activity.Following Navigo, the transportation passes "Autre Titre" and "TST"were also notable.Also We found that the stations"ABBABES" and "Assemblee nationel" have the fewest validations, specifically fewer than 5 With transport ticket title is "Autre Titre". The"12/01/2017 " is the day that had the highest number of validations, totaling 118,394





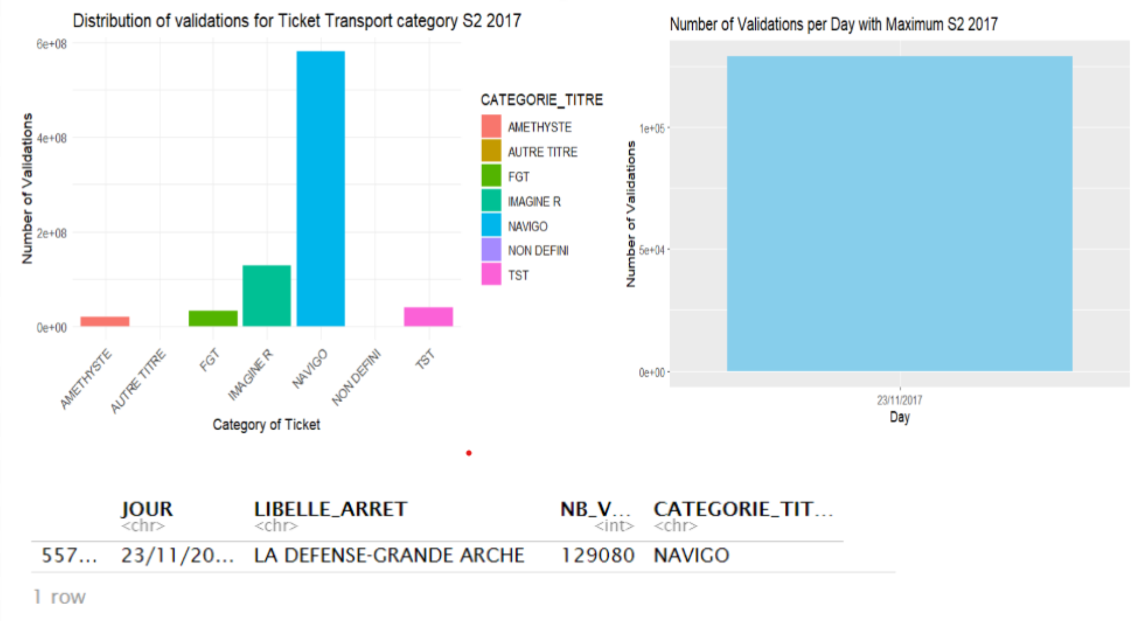
	JOUR <chr>	LIBELLE_ARRET <chr>	NB_V... <chr>	CATEGORIE_TIT... <chr>
51...	12/01/20...	LA DEFENSE-GRANDE ARCHE	118349	NAVIGO

1 row

FIGURE 1 – Identify the stops, days, and tickets categories with the highest number of validations for semester 1 2017

— Identify the stops, days, and tickets with the highest number of validations for semester 2

During the second semester of 2017 figure 2 , starting in July and ending in december, our graphical analyses revealed that "Navigo" had the highest number of validations among all transportation passes 129080. We can conclude that during this period, "Navigo" was the best-selling ticket. Additionally, the stations recording the highest number of validations were "La Défense - Grande Arche". It can be inferred that users likely have a profile of professionals, given that La Défense is the heart of Paris's economic activity.Following NAVIGO, the transportation passes Imagine R and TST were also notable.Also We found that the stations "BUTTES-CHAUMONT" and "EGLISE D'AUTEUIL" have the fewest validations, specifically fewer than 5 With transport ticket title is "Autre Titre. The"23/11/2017 " is the day that had the highest number of validations, totaling 129080.



	JOUR	LIBELLE_ARRET	NB_V...	CATEGORIE_TIT...
	<chr>	<chr>	<int>	<chr>
557...	23/11/20...	LA DEFENSE-GRANDE ARCHE	129080	NAVIGO

1 row

FIGURE 2 – Identify the stops, days, and tickets categories with the highest number of validations for semester 2 of 2017

1. **Effect of Summer Vacation on IMAGINE R Transport Ticket Validations in Semester 2 of 2017**

   To assess the impact of summer vacation on the validation count of IMAGINE R transport tickets, we compared two periods within the second semester of 2017. figure 3

   The first period encompassed the summer school break, starting from ("2017-07-08") to ("2017-09-03"). During this time, we found that the validation count for NAVIGO transport tickets was 144403148, while for IMAGINE R tickets, it was 24600169.

   The second period represents the regular timeframe from ("2017-07-04") to ("2017-10-10"). Here, the validation count for NAVIGO tickets was 130261826, and for IMAGINE R tickets, it reached 31434154.

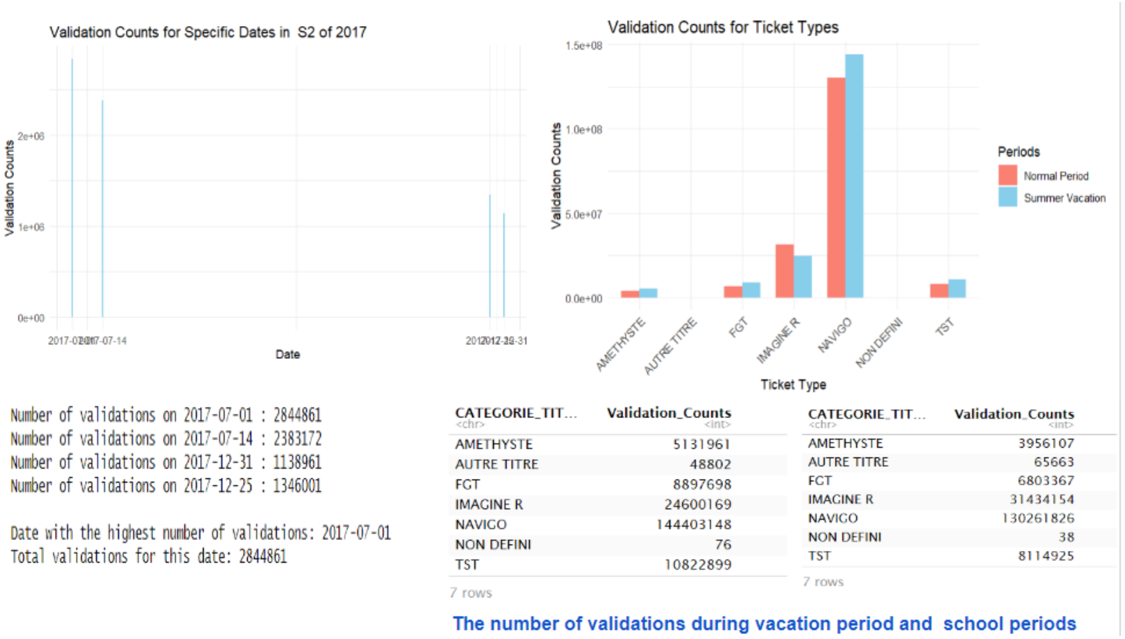
   => The rate of increase in IMAGINE R ticket validations between Period 1 and Period 2 is approximately 27.78%, we observe a significant impact from the summer school break on the validation count.

2. **Comparison between the number of validations on a normal day, Christmas Day, the last day of the year, and National Day**

   We found that on the date 2017-12-31, there were the fewest validations, equal to 1138961, which confirms our hypothesis as the French government offered free access without ticket validations from the transport terminals starting at 5 PM on 2017-12-31.

   The date 2017-07-01 had the highest number of validations : 2844861.

Additionally, we observed a significant number of validations on 14/07/2017, indicating higher public movement due to the celebration of the French National Day."



```
Number of validations on 2017-07-01 : 2844861
Number of validations on 2017-07-14 : 2383172
Number of validations on 2017-12-31 : 1138961
Number of validations on 2017-12-25 : 1346001

Date with the highest number of validations: 2017-07-01
Total validations for this date: 2844861
```

CATEGORIE_TIT... <chr>	Validation_Counts <int>
AMETHYSTE	5131961
AUTRE TITRE	48802
FGT	8897698
IMAGINE R	24600169
NAVIGO	144403148
NON DEFINI	76
TST	10822899

7 rows

CATEGORIE_TIT... <chr>	Validation_Counts <int>
AMETHYSTE	3956107
AUTRE TITRE	65663
FCT	6803367
IMAGINE R	31434154
NAVIGO	130261826
NON DEFINI	38
TST	8114925

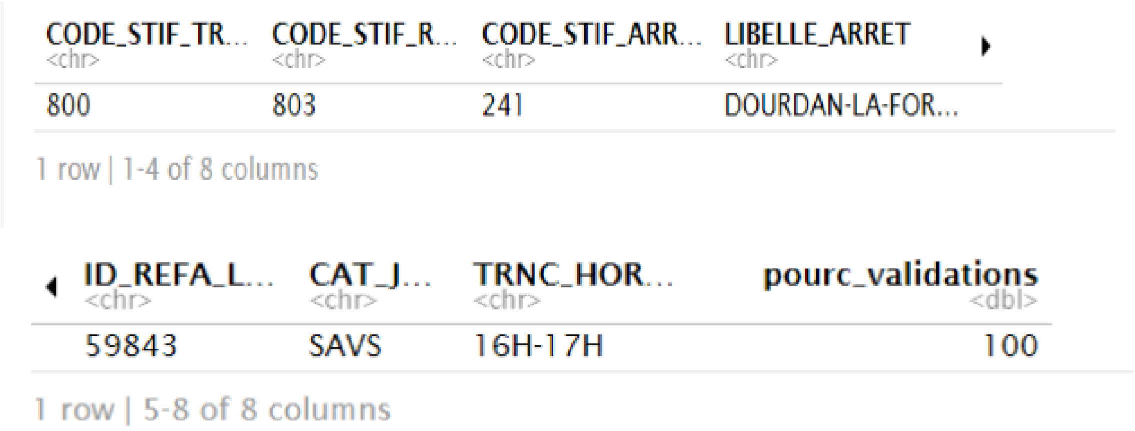
7 rows

The number of validations during vacation period and school periods

Figure 3 – Analysis of IMAGINE R Transport Ticket Validations : Impact of Summer Vacation and Distinctive Patterns Across Specific Dates in 2017

Determining the High Validation Percentage with Day Category, Locations, and Time Slot :

While analyzing the ProfilFerS1 dataset to determine the type of day with the most validations, we discovered that 'SAVS' corresponds to a Saturday during school holidays, with a validation percentage of 100%. The location label is 'Dourdan la foret,' and the time slot is from 4 PM to 5 PM.



CODE_STIF_TR... <chr>	CODE_STIF_R... <chr>	CODE_STIF_ARR... <chr>	LIBELLE_ARRET <chr>
800	803	241	DOURDAN-LA-FOR...

1 row | 1-4 of 8 columns

ID_REFA_L... <chr>	CAT_J... <chr>	TRNC_HOR... <chr>	pourc_validations <dbl>
59843	SAVS	16H-17H	100

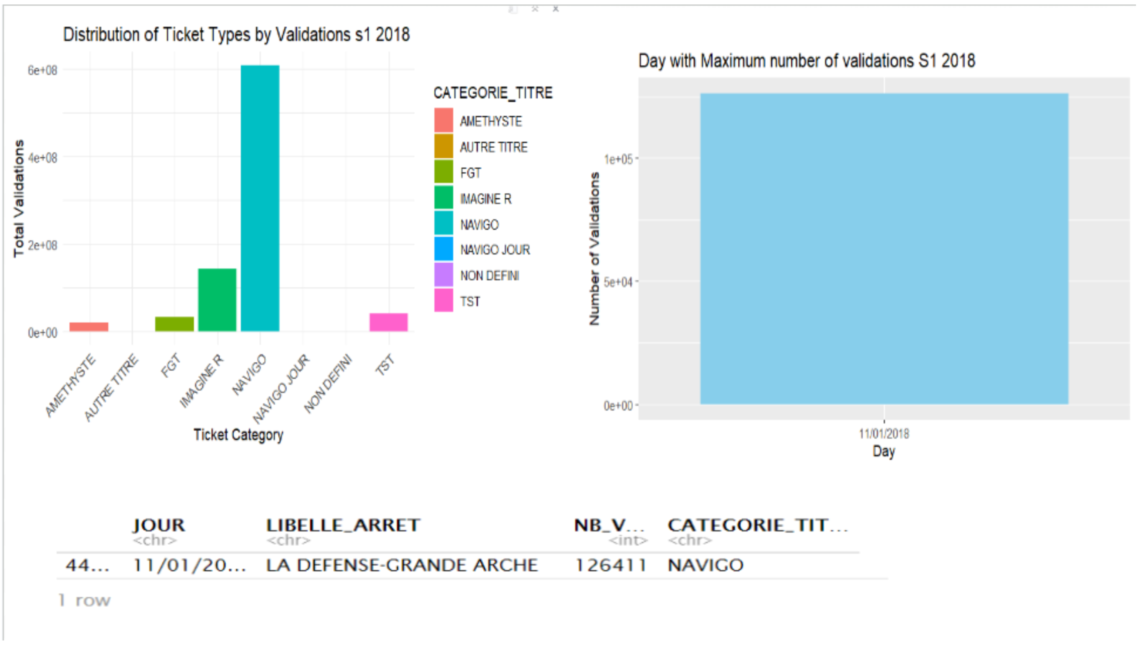
1 row | 5-8 of 8 columns

FIGURE 4 – the High Validation Percentage with Day Category, stop name, and Time Slot

5.1.2 Analysis of the 2018 dataset involves examining the number of validations for S1 and S2, as well as the profiles for S1 and S2.

— Identify the stops, days, and tickets with the highest number of validations for semester 1

During the first semester of 2018 figure 5, starting in January and ending in June, our graphical analyses revealed that "NAVIGO" had the highest number of validations among all transportation passes 126411.
We can conclude that during this period, "NAVIGO" was the best-selling ticket. Additionally, the stations recording the highest number of validations were "La Défense - Grande Arche". It can be inferred that users likely have a profile of professionals, given that "La Défense"is the heart of Paris's economic activity.
Following Navigo, the transportation passes "IMAGINE R" and "TST"were also notable.Also We found that the stations "PORTE MAILLOT" "ALESIA" have the fewest validations, specifically fewer than 5 With transport ticket title is "Navigo Jour" in 01/01/2018.
The"11/01/2018 " is the day that had the highest number of validations, totaling 126411 ;
Comparing 2017 and 2018, we see that the number of validations for transport tickets of type NAVIGO increased from 118394 to 126411.
=> So, the progress rate or percentage increase between 2017 and 2018 for the number of validations of NAVIGO transport tickets is approximately 6.77

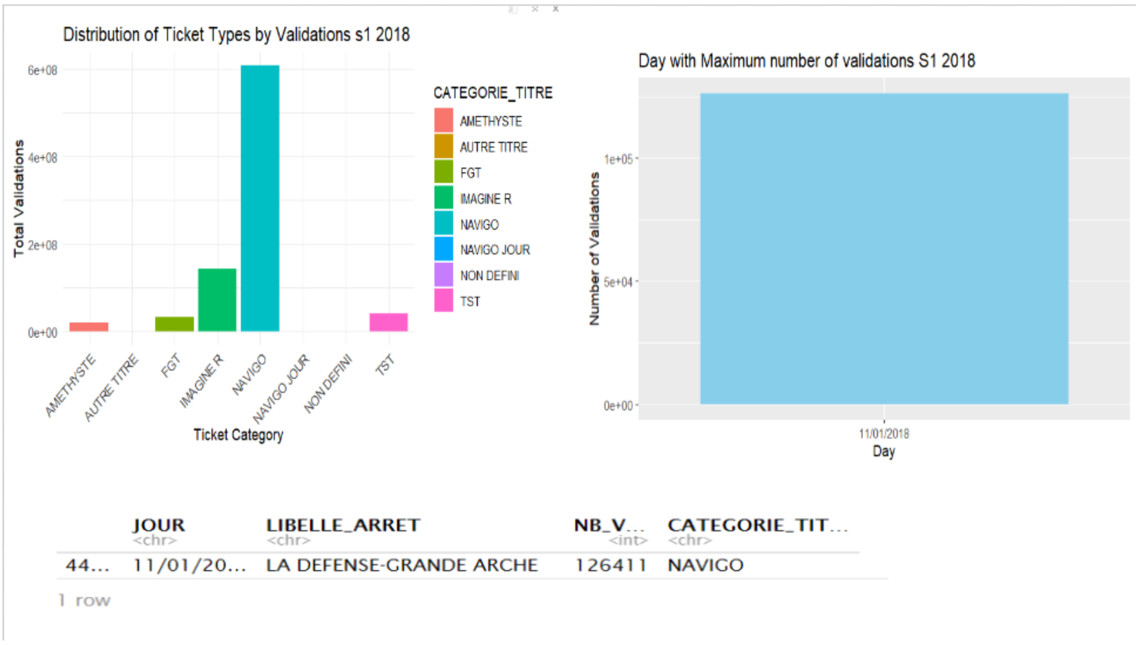


```
        JOUR          LIBELLE_ARRET              NB_V...   CATEGORIE_TIT...
        <chr>         <chr>                        <int>   <chr>
44...   11/01/20...   LA DEFENSE-GRANDE ARCHE     126411   NAVIGO
1 row
```

FIGURE 5 – Identify the stops, days, and tickets categories with the highest number of validations for semester 1 2018

— Identify the stops, days, and tickets with the highest number of validations for semester 2

During the second semester of 2018, figure 6 starting in July and ending in december, our graphical analyses revealed that "NAVIGO" had the highest number of validations among all transportation passes 125508.
We can conclude that during this period, "NAVIGO" was the best-selling ticket. Additionally, the stations recording the highest number of validations were "La Défense - Grande Arche". It can be inferred that users likely have a profile of professionals, given that La Défense is the heart of Paris's economic activity.
Following Navigo, the transportation passes Imagine R and TST were also notable.Also We found that the stations "BOULOGNE-PONT DE SAINT CLOUD"and "LES AGNETTES-ASNIERES-GENNEVILLIERS" have the fewest validations, specifically fewer than 5 With transport ticket title is "NAVIGO JOUR" in 01/07/2018. The"16/10/2018 " is the day that had the highest number of validations, totaling 125508.



```
        JOUR          LIBELLE_ARRET               NB_V...  CATEGORIE_TIT...
        <chr>         <chr>                        <int>   <chr>
44...   11/01/20...   LA DEFENSE-GRANDE ARCHE     126411   NAVIGO
1 row
```

Figure 6 – Identify the stops, days, and tickets categories with the highest number of validations for semester 2 2018

1. **Effect of Summer Vacation on IMAGINE R Transport Ticket Validations in Two Semesters of 2018**

   To assess the impact of summer vacation on the validation count of IMAGINE R transport tickets, we compared two periods within the second semester of 2018. figure 7

   The first period encompassed the summer school break, starting from ("2018-07-08") to ("2018-09-03"). During this time, we found that the validation count for NAVIGO transport tickets was 150304398, while for IMAGINE R tickets, it was 26000185.

   The second period represents the regular timeframe from ("2018-07-04") to ("2018-10-10"). Here, the validation count for NAVIGO tickets was 130201172 and for IMAGINE R tickets, it reached 32815673.

   We can also observe that the number of Navigo ticket validations during the summer is higher than during the normal period. This suggests that during the vacation period, there might be a higher number of tourists compared to the normal period.

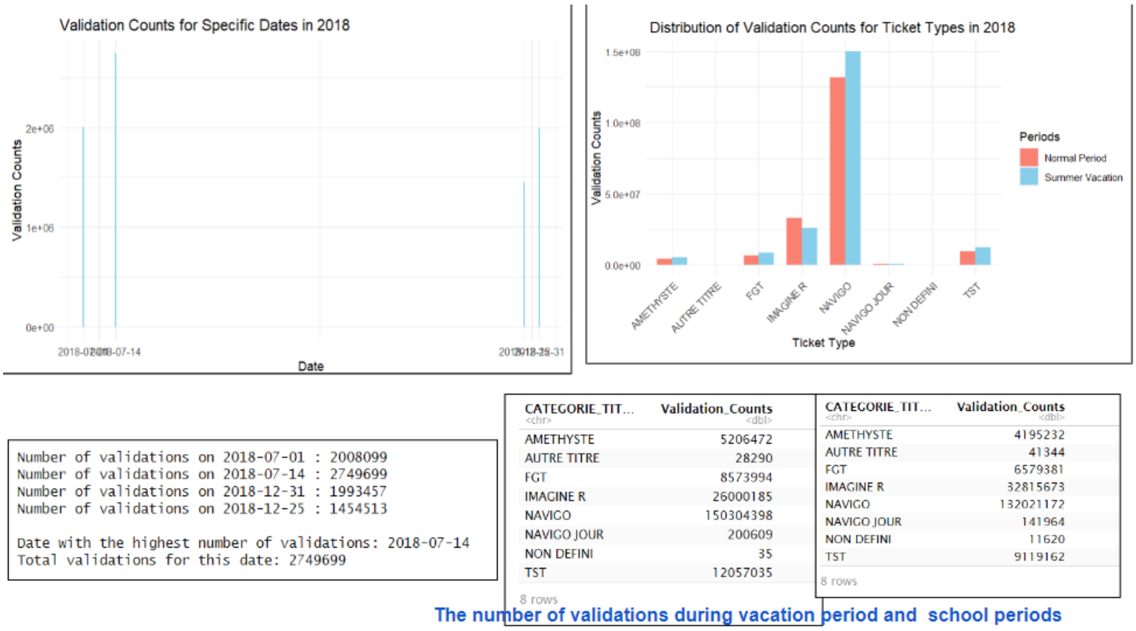
   => The rate of increase in IMAGINE R ticket validations between Period 1 and Period 2 is approximately 26.19 %., we observe a significant impact from the summer school break on the validation count.

2. **Comparison between the number of validations on a normal day, Christmas Day, the last day of the year, and National Day**

We found on 25/12/2018, which is Christmas Day, a slightly lower validation number compared to other days, 1454513, which confirms a lower validation rate. This suggests that people celebrate Christmas in their homes with their families and friends

The date 2018-07-14 had the highest number of validations : 2749699. Additionally, we observed a significant number of validations on 14/07/2018, indicating higher public movement due to the celebration of the French National Day."

Our goal is to analyse and discuss the number of validations on different specific dates, outlining the observations and the reasoning behind the variations in validation counts.



```
Number of validations on 2018-07-01 : 2008099
Number of validations on 2018-07-14 : 2749699
Number of validations on 2018-12-31 : 1993457
Number of validations on 2018-12-25 : 1454513

Date with the highest number of validations: 2018-07-14
Total validations for this date: 2749699
```

CATEGORIE_TIT... <chr>	Validation_Counts <dbl>
AMETHYSTE	5206472
AUTRE TITRE	28290
FGT	8573994
IMAGINE R	26000185
NAVIGO	150304398
NAVIGO JOUR	200609
NON DEFINI	35
TST	12057035

8 rows

CATEGORIE_TIT... <chr>	Validation_Counts <dbl>
AMETHYSTE	4195232
AUTRE TITRE	41344
FGT	6579381
IMAGINE R	32815673
NAVIGO	132021172
NAVIGO JOUR	141964
NON DEFINI	11620
TST	9119162

8 rows

The number of validations during vacation period and school periods

FIGURE 7 – Analysis of IMAGINE R Transport Ticket Validations : Impact of Summer Vacation and Distinctive Patterns Across Specific Dates in 2018

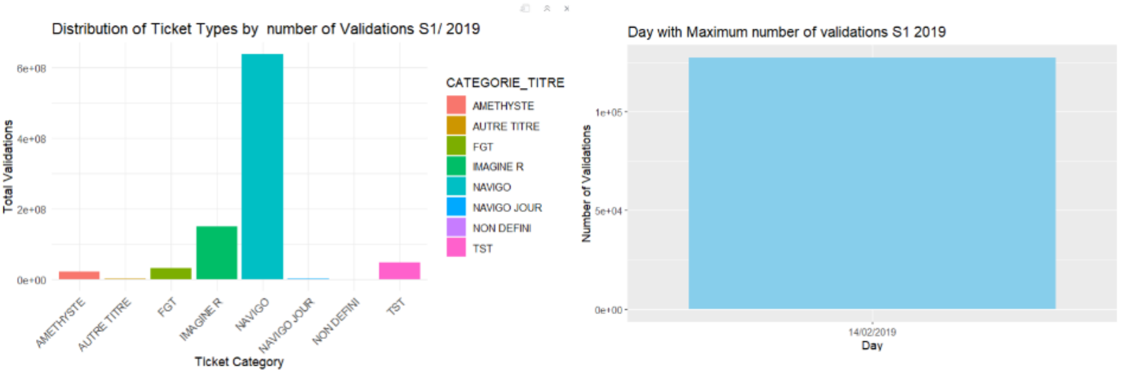
5.1.3 Analysis of the 2019 dataset involves examining the number of validations for S1 and S2, as well as the profiles for S1 and S2.

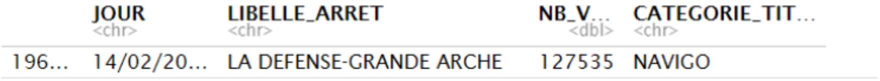
- Identify the stops, days, and tickets with the highest number of validations for semester 1

  During the first semester of 2019,figure 8 starting in January and ending in June, our graphical analyses revealed that "NAVIGO" had the highest number of validations among all transportation passes 127535.

  We can conclude that during this period, "NAVIGO" was the best-selling ticket. Additionally, the stations recording the highest number of validations

were "La Défense - Grande Arche". It can be inferred that users likely have a profile of professionals, given that "La Défense"is the heart of Paris's economic activity.
Following Navigo, the transportation passes "IMAGINE R" and "TST"were also notable.Also We found that the stations "PORTE MAILLOT" "ALESIA" have the fewest validations, specifically fewer than 5 With transport ticket title is "Navigo Jour" in 01/01/2019.
The"14/02/2018 " is the day that had the highest number of validations, totaling 127535.
Comparing 2018 and 2019, we see that the number of validations for transport tickets of type NAVIGO increased from 126411 to 127535.
=> So, the progress rate or percentage increase between 2018 and 2019 for the number of validations of NAVIGO transport tickets is approximately 0.88%.



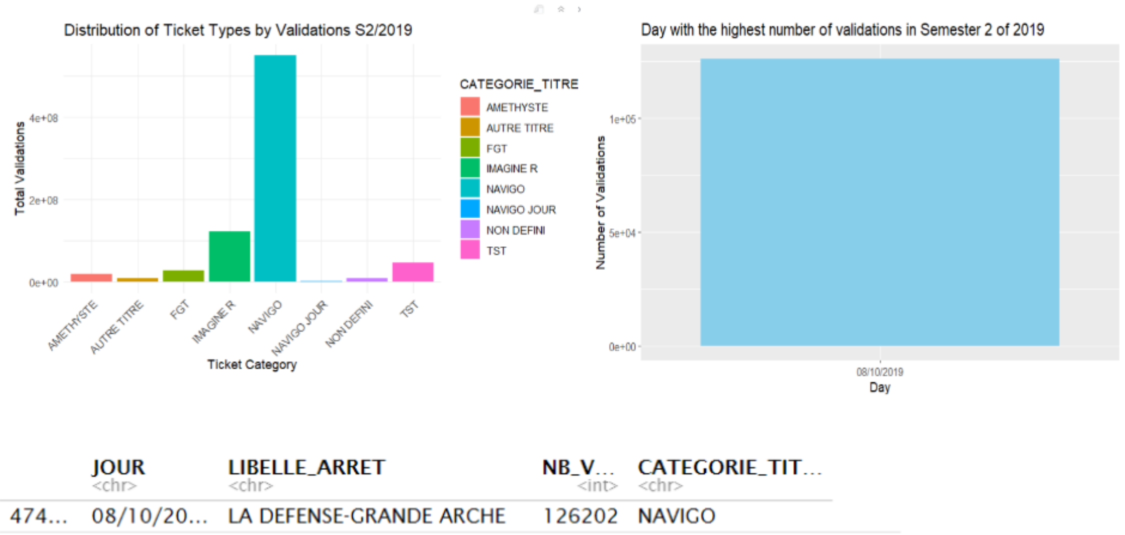


	JOUR <chr>	LIBELLE_ARRET	NB_V… <dbl>	CATEGORIE_TIT… <chr>
196…	14/02/20…	LA DEFENSE-GRANDE ARCHE	127535	NAVIGO

FIGURE 8 – Identify the stops, days, and tickets categories with the highest number of validations for semester 1 2019

— Identify the stops, days, and tickets with the highest number of validations for semester 2
During the second semester of 2019, figure 9 starting in July and ending in december, our graphical analyses revealed that "NAVIGO" had the highest number of validations among all transportation passes 126202.
We can conclude that during this period, "NAVIGO" was the best-selling ticket. Additionally, the stations recording the highest number of validations were "La Défense - Grande Arche". It can be inferred that users likely have a profile of professionals, given that La Défense is the heart of Paris's economic

activity.
Following Navigo, the transportation passes Imagine R and TST were also notable.Also We found that the stations "BOULOGNE-PONT DE SAINT CLOUD"and "PORTE MAILLOT" and "ALESIA" have the fewest validations, specifically fewer than 5 With transport ticket title is "NON DEFINI" in 01/07/2019.
The"08/10/2019 " is the day that had the highest number of validations, totaling 126202.



	JOUR <chr>	LIBELLE_ARRET <chr>	NB_V… <int>	CATEGORIE_TIT… <chr>
474…	08/10/20…	LA DEFENSE-GRANDE ARCHE	126202	NAVIGO

Figure 9 – Identify the stops, days, and tickets categories with the highest number of validations for semester 2 2019

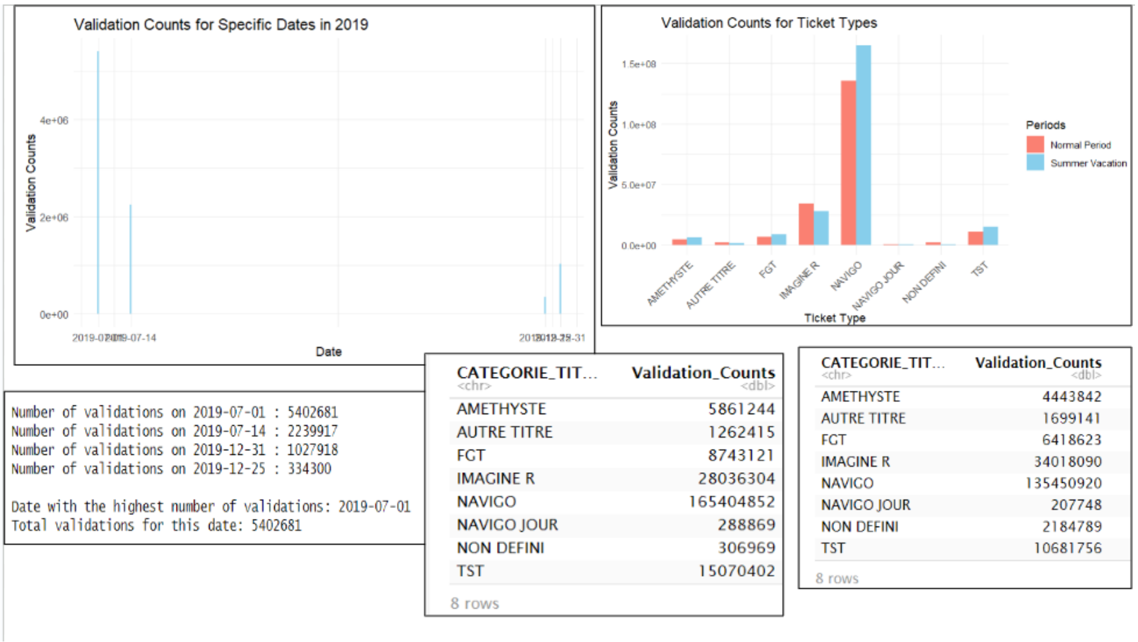
1. **Effect of Summer Vacation on IMAGINE R Transport Ticket Validations in Two Semesters of 2019**
   To assess the impact of summer vacation on the validation count of IMAGINE R transport tickets, we compared two periods within the second semester of 2019. figure 10

   To assess the impact of summer vacation on the validation count of IMAGINE R transport tickets, we compared two periods within the second semester of 2019.
   The first period encompassed the summer school break, starting from ("2019-07-08") to ("2019-09-03"). During this time, we found that the validation count for NAVIGO transport tickets was 165404852, while for IMAGINE R tickets, it was 28036304.
   The second period represents the regular timeframe from ("2019-07-04") to

("2019-10-10"). Here, the validation count for NAVIGO tickets was 135450920 and for IMAGINE R tickets, it reached 34018090.
We can also observe that the number of Navigo ticket validations during the summer is higher than during the normal period. This suggests that during the vacation period, there might be a higher number of tourists compared to the normal period.

=> The rate of increase in IMAGINE R ticket validations between Period 1 and Period 2 is approximately 21.35%., we observe a significant impact from the summer school break on the validation.

2. **Comparison between the number of validations on a normal day, Christmas Day, the last day of the year, and National Day**
The date 2019-07-14 had a important number of validations compraing with others event in ile de france : 2239917. Additionally, we observed a significant number of validations on 14/07/2019, indicating higher public movement due to the celebration of the French National Day."
Our goal is to analyse and discuss the number of validations on different specific dates, outlining the observations and the reasoning behind the variations in validation counts.



```
Number of validations on 2019-07-01 : 5402681
Number of validations on 2019-07-14 : 2239917
Number of validations on 2019-12-31 : 1027918
Number of validations on 2019-12-25 : 334300

Date with the highest number of validations: 2019-07-01
Total validations for this date: 5402681
```

CATEGORIE_TIT... <chr>	Validation_Counts <dbl>
AMETHYSTE	5861244
AUTRE TITRE	1262415
FGT	8743121
IMAGINE R	28036304
NAVIGO	165404852
NAVIGO JOUR	288869
NON DEFINI	306969
TST	15070402

8 rows

CATEGORIE_TIT... <chr>	Validation_Counts <dbl>
AMETHYSTE	4443842
AUTRE TITRE	1699141
FGT	6418623
IMAGINE R	34018090
NAVIGO	135450920
NAVIGO JOUR	207748
NON DEFINI	2184789
TST	10681756

8 rows

FIGURE 10 – Analysis of IMAGINE R Transport Ticket Validations : Impact of Summer Vacation and Distinctive Patterns Across Specific Dates in 2019

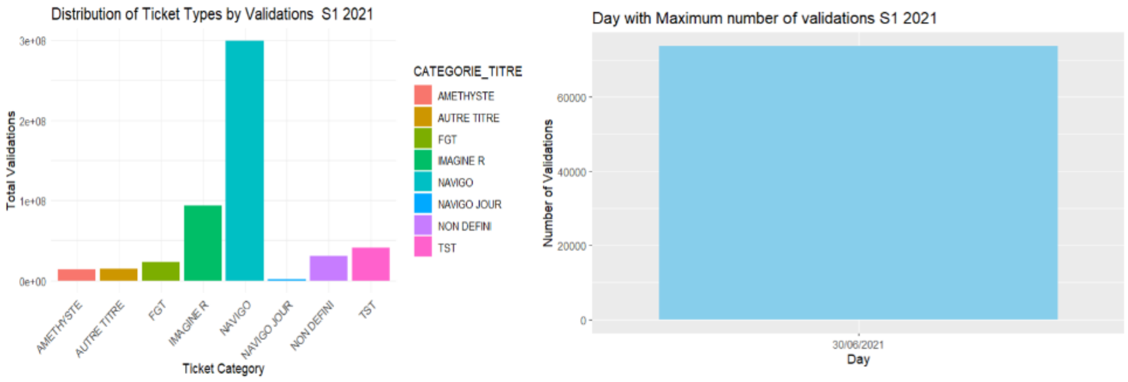
5.1.4 Analysis of the 2021 dataset involves examining the number of validations for S1 and S2, as well as the profiles for S1 and S2.

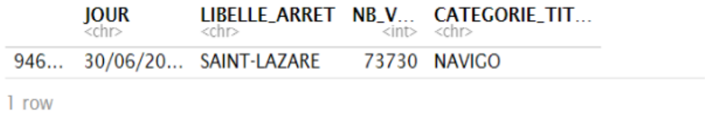
— Identify the stops, days, and tickets with the highest number of validations for semester 1

During the first semester of 2021,figure 11 starting in January and ending in June, our graphical analyses revealed that "Navigo" had the highest number of validations among all transportation passes 73730 .
We can conclude that during this period, "Navigo" was the best-selling ticket.
Additionally, the stations recording the highest number of validations were "SAINT-LAZARE".
Following Navigo, the transportation passes "IMAGINE R" and "TST"were also notable.Also We found that the stations" PORTE MAILLOT" and" BOULOGNE-PONT DE SAINT" have the fewest validations, specifically fewer than 5 With transport ticket title is "NAVIGO JOUR" in 01/01/2021.
The"30/06/2021 " is the day that had the highest number of validations, totaling 73730.
Comparing 2021 and 2019, we see that the number of validations for transport tickets of type NAVIGO increased from 127535 to 73730.
So, the regression rate os rate or percentage increase between 2021 and 2019 for the number of validations of NAVIGO transport tickets is approximately 24.17%

5.1.5 The impact of the coronavirus on the number of validations

The impact of the coronavirus on the number of validations explains the decrease in the number of validations due to the lockdown the period of confinement from March 20th to May 3rd, 2021 and the emergence of telecommuting during the years 2017 to 2019. During that time, La Défense - Grande Arche was the station with the most validations in Île-de-France. However, in 2021, Saint Lazare has the highest number of validations.





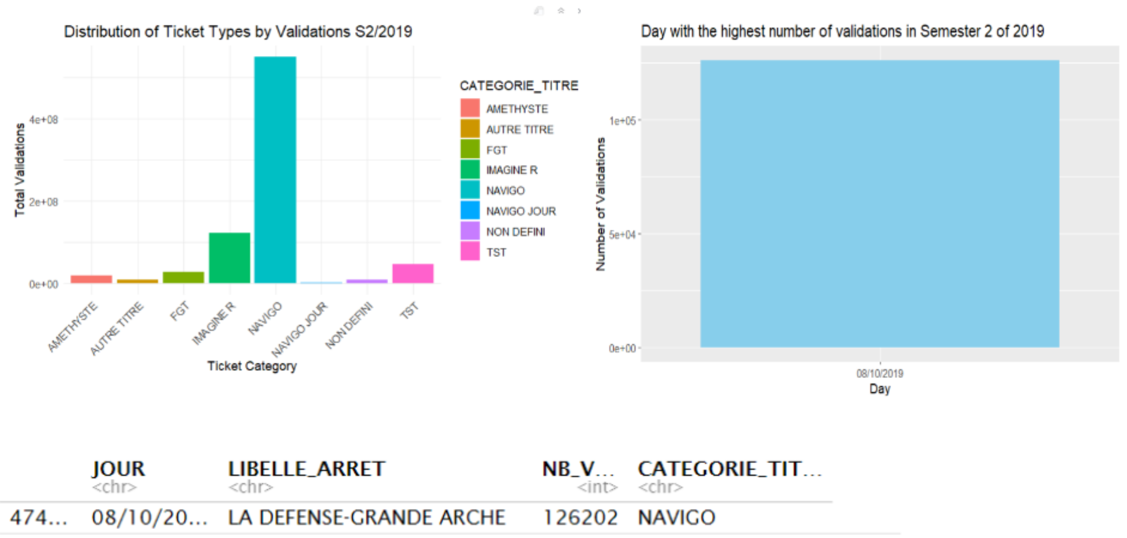
	JOUR <chr>	LIBELLE_ARRET <chr>	NB_V... <int>	CATEGORIE_TIT... <chr>
946...	30/06/20...	SAINT-LAZARE	73730	NAVIGO

1 row

FIGURE 11 – Identify the stops, days, and tickets categories with the highest number of validations for semester 1 2021

— Identify the stops, days, and tickets with the highest number of validations for semester 2

During the second semester of 2021, figure 12 starting in July and ending in december, our graphical analyses revealed that "NAVIGO" had the highest number of validations among all transportation passes 90901.
We can conclude that during this period, "NAVIGO" was the best-selling ticket. Additionally, the stations recording the highest number of validations were "SAINT-LAZARE".
Following Navigo, the transportation passes Imagine R and TST were also notable.Also We found that the stations "BREGUET-SABIN" and"CAMPO-FORMIO" have the fewest validations, specifically fewer than 5 With transport ticket title is "NAVIGO JOUR" in 01/07/2021.
The"18/11/2021 " is the day that had the highest number of validations, totaling 90901.



```
      JOUR         LIBELLE_ARRET             NB_V…   CATEGORIE_TIT…
      <chr>        <chr>                      <int>  <chr>
474…  08/10/20…    LA DEFENSE-GRANDE ARCHE   126202  NAVIGO
```

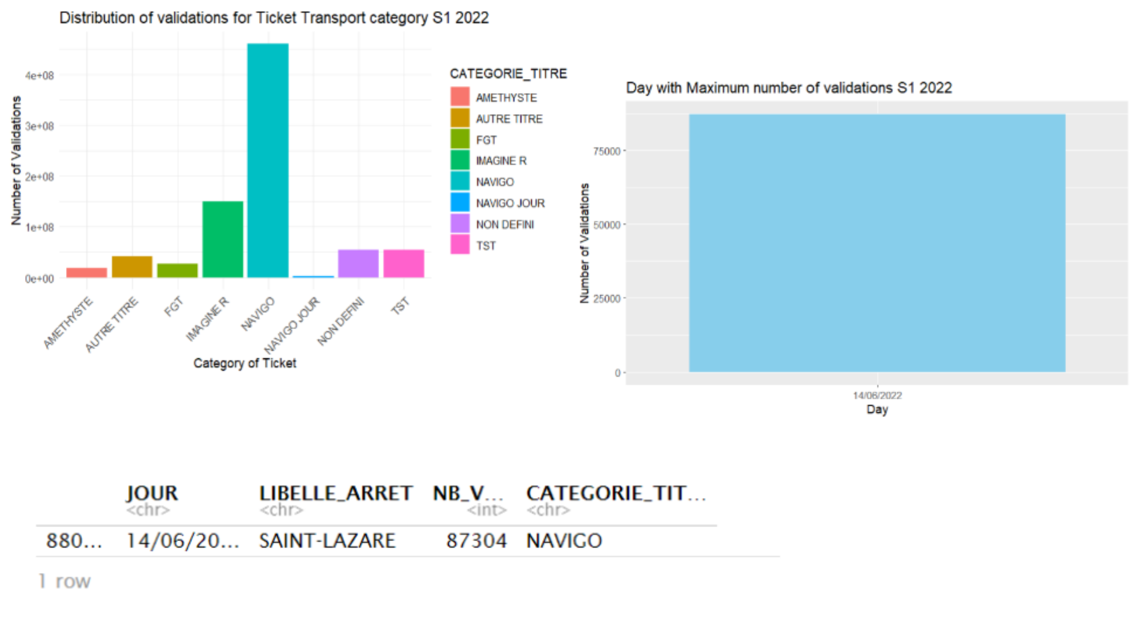
FIGURE 12 – Identify the stops, days, and tickets categories with the highest number of validations for semester 2 2021

1. **Comparison of Ticket Validation Counts between Summer Vacation and Normal School Periods after the strict lockdown in 2021**
   To assess the impact of summer vacation on the validation count of IMAGINE R transport tickets, we compared two periods within the second semester of 2021. **??** To assess the impact of summer vacation on the validation count of IMAGINE R transport tickets, we compared two periods within the second semester of 2021.
   The first period encompassed the summer school break, starting from ("2021-07-08") to ("2021-09-03"). During this time, we found that the validation count for NAVIGO transport tickets was 101089683 , while for IMAGINE R tickets, it was 24381950.
   The second period represents the regular timeframe from ("2021-07-04") to ("2021-10-10"). Here, the validation count for NAVIGO tickets was 92485334, and for IMAGINE R tickets, it reached 31424058
   => The rate of progress in IMAGINE R ticket validations between Period 1 and Period 2 is approximately 28.87 %., we observe a significant impact from the summer school break on the validation count.

5.1.6 Analysis of the 2022 dataset involves examining the number of validations for S1 and S2, as well as the profiles for S1 and S2.

— Identify the stops, days, and tickets with the highest number of validations for semester 1

During the first semester of 2022,figure 13 starting in January and ending in June, our graphical analyses revealed that "Navigo" had the highest number of validations among all transportation passes 87304.
We can conclude that during this period, "Navigo" was the best-selling ticket. Additionally, the stations recording the highest number of validations is Saint-Lazare station.
The Saint-Lazare station is served by regional trains, long-distance trains, and suburban lines, facilitating travel for both local and national passengers.
Following Navigo, the transportation passes "Imagine r" and "TST"were also notable.Also We found that the stations "ALESIA" and "BOULOGNE-PONT DE SAINT" have the fewest validations, specifically fewer than 5 With transport ticket title is "Navigo Jour" at 01/01/2022.
"14/06/2022" is the day that had the highest number of validations, totaling 87304.



	JOUR <chr>	LIBELLE_ARRET <chr>	NB_V... <int>	CATEGORIE_TIT... <chr>
880...	14/06/20...	SAINT-LAZARE	87304	NAVIGO

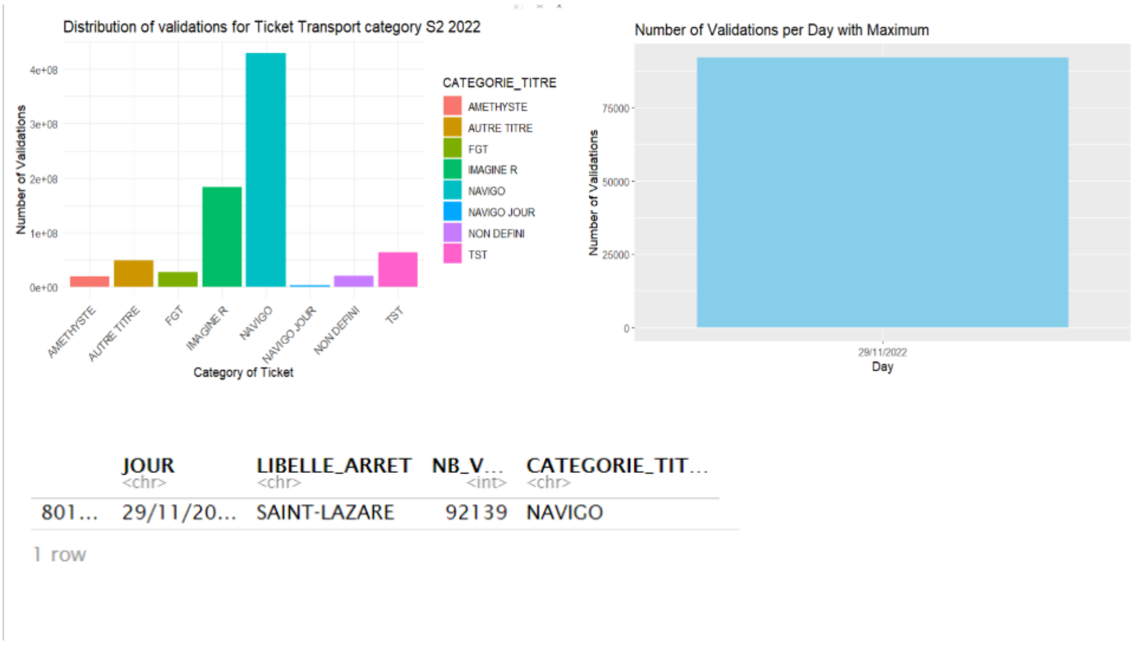
1 row

FIGURE 13 – Identify the stops, days, and tickets categories with the highest number of validations for semester 1 2022

— Identify the stops, days, and tickets with the highest number of validations for semester 2

During the second semester of 2022, figure 14 starting in july and en-

ding in decembre, our graphical analyses revealed that "Navigo" had the highest number of validations among all transportation passes 92139.
We can conclude that during this period, "Navigo" was the best-selling ticket. Additionally, the stations recording the highest number of validations is Saint-Lazare station.
The Saint-Lazare station is served by regional trains, long-distance trains, and suburban lines, facilitating travel for both local and national passengers.
Following Navigo, the transportation passes "Imagine r" and "TST"were also notable.Also We found that the stations "BOULOGNE-ST.CL" and "CHARDON-LAGACH " have the fewest validations, specifically fewer than 5 With transport ticket title is "Navigo Jour" at 01/07/2022.
"29/11/2022" is the day that had the highest number of validations, totaling 87304.



	JOUR <chr>	LIBELLE_ARRET <chr>	NB_V... <int>	CATEGORIE_TIT... <chr>
801...	29/11/20...	SAINT-LAZARE	92139	NAVIGO

1 row

Figure 14 – Identify the stops, days, and tickets categories with the highest number of validations for semester 2 2022

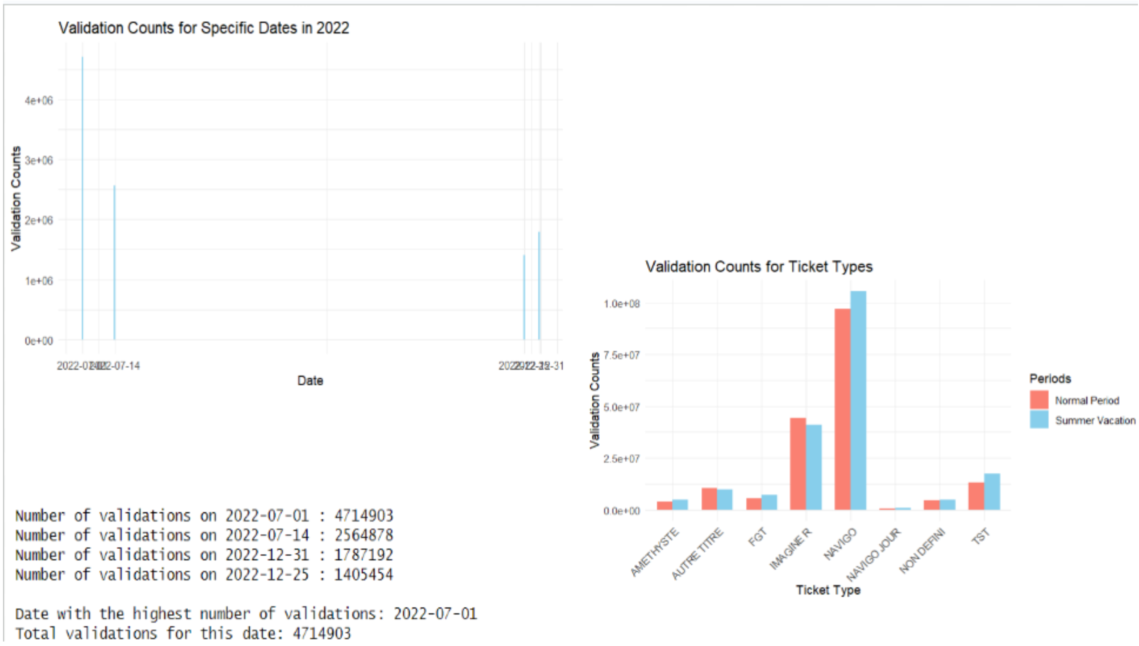
(a) Comparison of Ticket Validation Counts between Summer Vacation and Normal School Periods after the strict lockdown in 2021

To assess the impact of summer vacation on the validation count of IMAGINE R transport tickets, we compared two periods within the second semester of 2022. figure 15 The first period encompassed the summer school break, starting from ("2022-07-08") to ("2022-09-03"). During this

time, we found that the validation count for NAVIGO transport tickets was 105640379 , while for IMAGINE R tickets, it was 41019728.
The second period represents the regular timeframe from ("2022-09-04") to ("2022-10-10"). Here, the validation count for NAVIGO tickets was 97082966, and for IMAGINE R tickets, it reached 44207729.

=> The rate of progess in IMAGINE R ticket validations between Period 1 and Period 2 is approximately 7.77%., we observe a significant impact from the summer school break on the validation count.

(b) **Comparison between the number of validations on a normal day, Christmas Day, the last day of the year, and National Day**
To assess the impact of summer vacation on the validation count of IMAGINE R transport tickets, we compared two periods within the second semester of 2022.
The first period encompassed the summer school break, starting from ("2022-07-08") to ("2022-09-03"). During this time, we found that the validation count for NAVIGO transport tickets was 105640379 , while for IMAGINE R tickets, it was 41019728.
The second period represents the regular timeframe from ("2022-09-04") to ("2022-10-10"). Here, the validation count for NAVIGO tickets was 97082966, and for IMAGINE R tickets, it reached 44207729.
=> The rate of progess in IMAGINE R ticket validations between Period 1 and Period 2 is approximately 7.77%., we observe a significant impact from the summer school break on the validation count.



```
Number of validations on 2022-07-01 : 4714903
Number of validations on 2022-07-14 : 2564878
Number of validations on 2022-12-31 : 1787192
Number of validations on 2022-12-25 : 1405454

Date with the highest number of validations: 2022-07-01
Total validations for this date: 4714903
```

FIGURE 15 – Identify the stops, days, and tickets categories with the highest number of validations for semester 2 2022

— Identify the stops, days, and tickets with the highest number of validations for semester 1

During the first semester of 2023,figure 16 starting in January and ending in June, our graphical analyses revealed that "NAVIGO" had the highest number of validations among all transportation passes 94766.

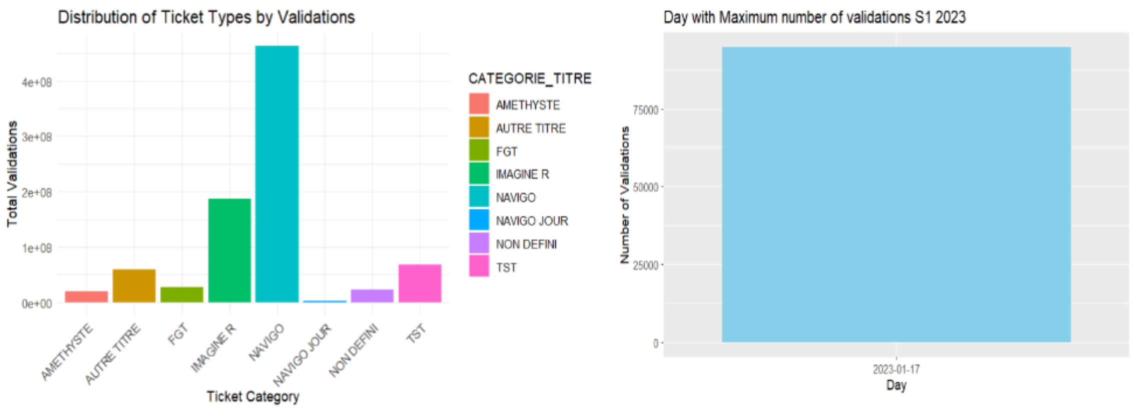
We can conclude that during this period, "NAVIGO" was the best-selling ticket. Additionally, the stations recording the highest number of validations were "La Défense". It can be inferred that users likely have a profile of professionals, given that "La Défense"is the heart of Paris's economic activity.

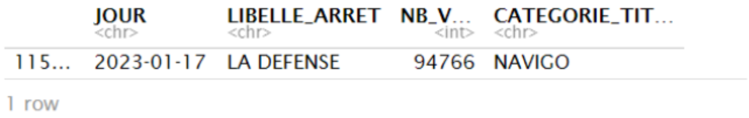
Following Navigo, the transportation passes "IMAGINE R" and "TST"were also notable.Also We found that the stations "NOINTEL MOURS" and " VALMONDOIS" have the fewest validations, specifically fewer than 5 With transport ticket title is "Navigo Jour" in 26/06/2023.

The"17/01/2023 " is the day that had the highest number of validations, totaling 94766.

Comparing 2022 and 2023, we see that the number of validations for transport tickets of type NAVIGO increased from 87304 to 94766 .

=> So, the progress rate or percentage increase between 2022 and 2023 for the number of validations of NAVIGO transport tickets is approximately 8.55





	JOUR <chr>	LIBELLE_ARRET <chr>	NB_V... <int>	CATEGORIE_TIT... <chr>
115...	2023-01-17	LA DEFENSE	94766	NAVIGO

1 row

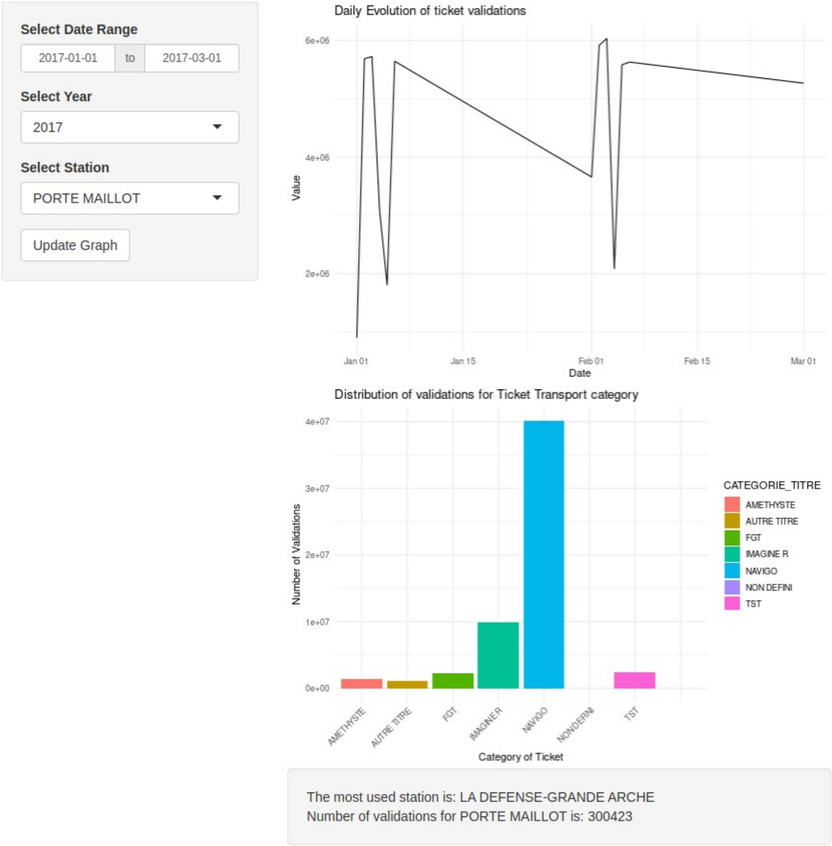
FIGURE 16 – Identify the stops, days, and tickets categories with the highest number of validations for semester 1 2023

6 Dashboard Development using Shiny

In this section we present the Dashboard and what the user can do to view the ridership data.

The user can choose the year and the date interval of the ridership statistics , for each interval the evolution of the number of daily validations and the usage statistics for each category of ticket in the chosen interval.

The use can also select a station from the list of all stations to show the number of total validations in the chosen interval To access the :Shiny App



Select Date Range

2017-01-01 to 2017-03-01

Select Year

2017

Select Station

PORTE MAILLOT

The most used station is: LA DEFENSE-GRANDE ARCHE
Number of validations for PORTE MAILLOT is: 300423

Figure 17 – Enter Caption

7 Conclusion

In this project, our initial phase involves collecting data on Île-de-France's railway station ridership from 2017 up to the first half of 2023. The primary focus is on data cleaning, rectifying missing values, and removing outliers to ensure data integrity. Following this, we move into Exploratory Data Analysis (EDA) to uncover ridership patterns. This phase includes identifying overarching trends, examining seasonality and monthly variations, and meticulously assessing potential outliers that could impact the data.

After the EDA, our focus shifts to comparing ridership against established norms. This includes establishing a baseline for a 'normal' week and examining deviations during holidays (such as Christmas, Easter, National Day), vacations, the start of the school year, the impact of the coronavirus pandemic on the number of validations. We aim to understand how these events influence the number of validations, types of transport tickets purchased, and locations visited compared to regular days.

The final stage involves developing an interactive dashboard using the Shiny framework in R. This dashboard will serve as a dynamic interface, featuring key visualizations showcasing overall ridership trends, weekly variations, and comprehensive comparisons with established norms. It will provide stakeholders with an intuitive tool to monitor and understand ridership dynamics

Références

[1] Open Data ile de france, *Historique des données de validation sur le réseau ferré (2017-2022)*, `https://data.iledefrance-mobilites.fr/explore/dataset/histo-validations-reseau-ferre/information/`.

[2] Open Data ile de france, *Validations sur le réseau ferré : Nombre de validations par jour (1er semestre 2023)*, `https://data.iledefrance-mobilites.fr/explore/dataset/validations-reseau-ferre-nombre-validations-par-jour-1er-semestre/information/`.

[3] Open Data ile de france, *Référentiel des arrêts : Zones d'arrêts*, `https://data.iledefrance-mobilites.fr/explore/dataset/zones-d-arrets/information/?disjunctive.zdatype`.