Data Exploration & Preparation

Train passengers count data analysis

Pitch

Your company is a startup who wants to help other companies adjusting the number of employees based on business volume prediction.

In the specific case of this practical work, you're in charge to explore possibilities of predicting what would be the number of train passengers for a given day, for the account of SNCF (French train transportation company)

Data

SNCF does not have (at least they don't want to open it to you) accurate statistics about what is the volume of passengers per day and per train station, so they can't provide those history data, they although have the number of daily lost & found items per train station. In this exercise, we assume that number is correlated to daily the number of passengers.

Datasets

Lost and found data (2019, 2020, 2021, 2022)

These are the main datasets you will have for this exercise; it represents the characteristics of each lost and found event.

Train stations are identified by their **UIC code** (« **U**nion **I**nternationnale des **C**hemins de fer »). This code is unique and can be found in other datasets related to train stations.

Column description:

Label	Type	description
Date	Datetime (Text)	Date of the item loss Follows format yyyy-mm-dd HH:mm:ss
Date et Heure de restitution	Datetime (Text)	Date of the item restitution Follows format yyyy-mm-dd HH:mm:ss
Gare	Text	Train station name
Nature d'objets	Category	Represents the detailed kind of lost belonging
Code UIC	Number	The uic code representing the train station
Type d'objets	Category	The category representing the item lost
Type enregistrement	Category	Records the lost and found event source

List of train stations

The list of train stations. It will contain among other data, geographical position of each train station.

Train stations are identified by their **UIC code**

Label	**Type**	**description**
Libelle	Text	name of the train station
PK	Text	"Point Kilométrique" – represents the location using the train network identification
commune	Text	City
idgaia	Text	Reference in the internal SNCF directory
geo_point2d	Text	Coordinates using geopoint notation
geo_shape	Text	Draws the area represented by the train station
code_uic	Integer	Unique code representing this train station
fret	Category	"O": means this train station handles freight "N": means this train station does not handle freight
voyageurs	Category	"O": means this train station handles travelers "N": means this train station does not handle travelers
code_ligne	Category	The code representing the trip line
rg_troncons	Number	The number of line sections impacted
departement	Category	The French department where the train station is located
Idreseau	Number	The id identifying the train station on the French train network
x_l93	Number	The longitude following l93 map standard
y_l93	Number	The latitude following l93 map standard
x_wgs84	Number	The longitude following wgs84 map standard
y_wgs84	Number	The latitude following wgs84 map standard
c_geo.1	Number	The latitude
c_geo.2	Number	The longitude

Train station frequentations

This dataset gathers the total frequentation per train station and per year from year 2015 to year 2020.

Label	**Type**	**description**
Nom de la gare	Text	name of the train station
Code UIC	Number	Unique code representing this train station
Code postal	Number	The postal code where the train station is located
Segmentation	Category	?
Total Voyageurs "N"	Number	Number of passengers for year N
Total Voyageurs + Non Voyageurs « N »	Number	Coordinates using geopoint notation

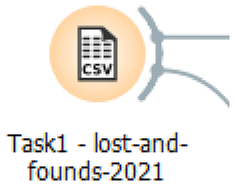
Goals

1. Load the data and explore them to determine what would be the ideal preparation for presenting your added value, for the prediction of train stations passengers loads.
2. Thanks to a refined analysis, prepare the data to make them ready for training a statistical system based on the history data of lost and found events.

Guided section

Important Notice

- Prefix all the widgets you put in your workflow by the name of the task you're covering, as in the following example (loading data for Task1)



  If it is not clear for what task you're using this widget, it will not be taken in account for the grading.
- Give meaningful name to your widgets (avoid Datatable (1), Datatable (2) etc.)
- For all the following tasks, the tasks submitted during the exam will be graded with a bonus factor of **1.35**, the tasks submitted after with factor of **0.65**

Task 1. Load the lost and found csv data (1pt)

a. Load the 4 files provided for lost and found data in a single datatable (use concatenate widget)
b. Observe the events distribution across months, split the series upon year (you should have 4 series) – Use distribution Widget

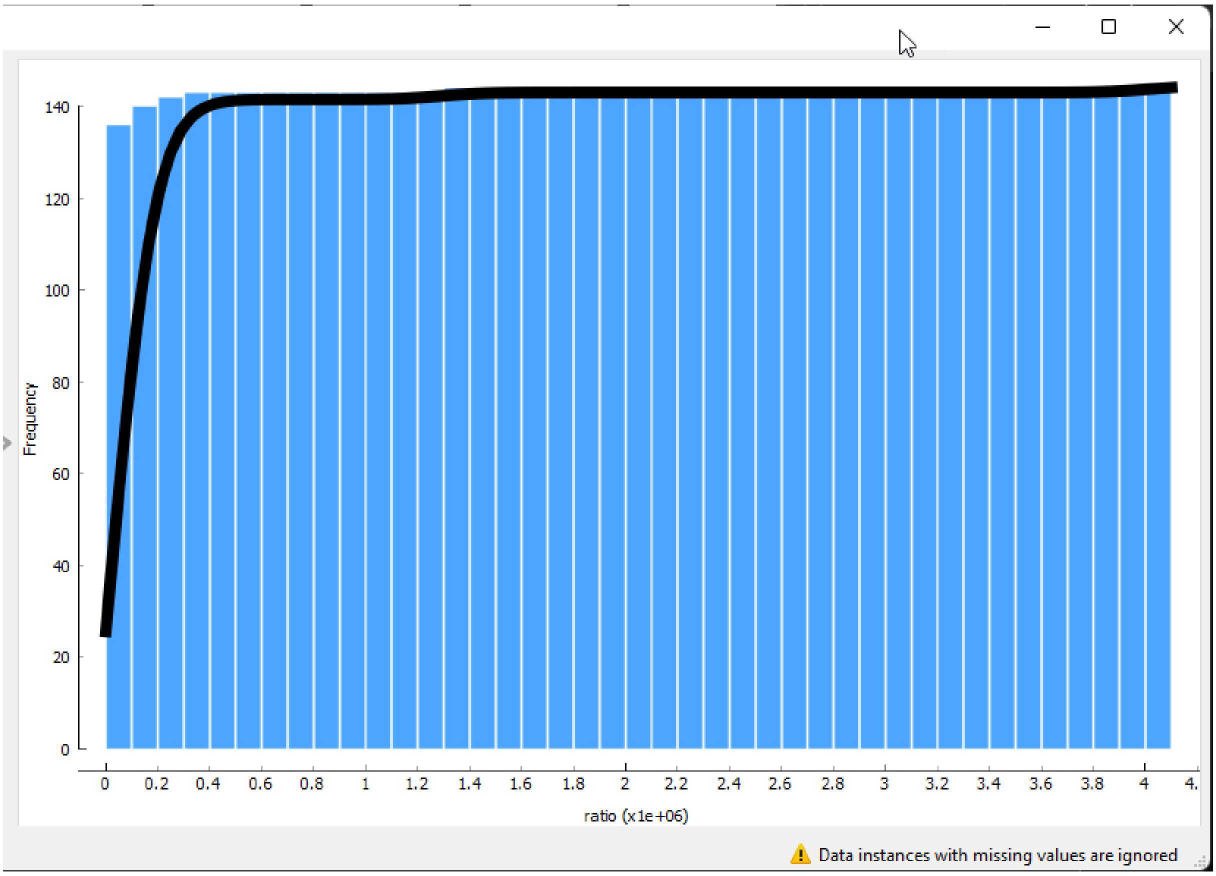
Task 2. Load the total passenger frequentation data (1pt)

a. Load the file provided in a datatable.
b. Isolate the "Total Voyageurs" for each year column (remove unnecessary columns)

Task 3. Find the correlation between lost and found events and frequentation (3pts)

The goal of this task is to validate that there is a reliable correlation between lost and found events and passenger traffic.

a. Compute the total number of lost and found events per train station and per year (use group by widget)
b. Isolate the data from 2019 on both side (“frequentation” and “lost and found”)
c. Compute the ratio for each train station between the frequentation and the lost and found items (use Feature constructor widget) for year 2019
d. Project the ration dispersion across data (use distribution widget), show cumulative distribution as in the following extract:



Data instances with missing values are ignored

e. Is it relevant to take all the data (i.e. what is the range that this ratio can take)? Or a subset can be better (removing outliers)

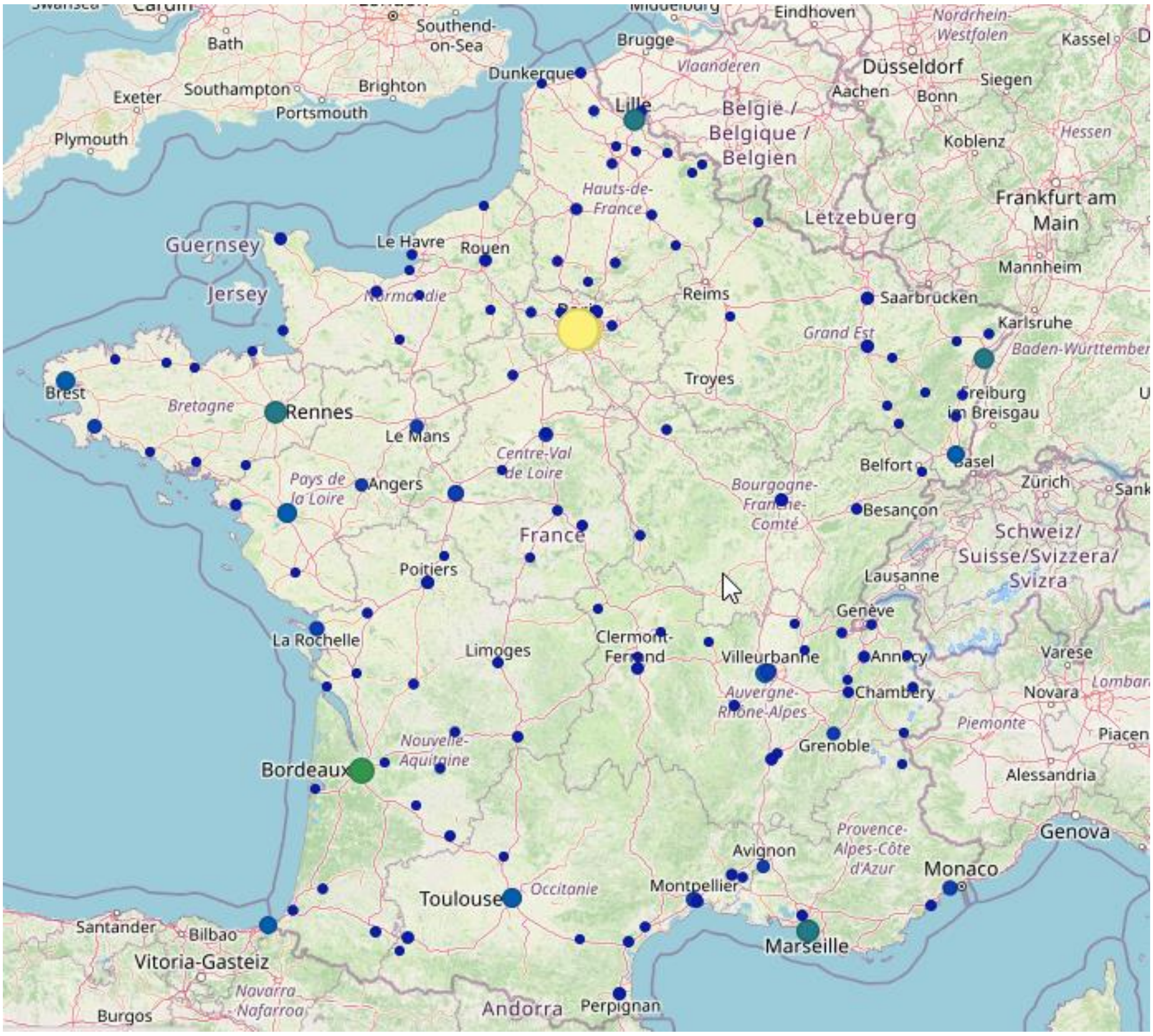
f. Is an average ratio relevant to represent this correlation? If not use what you think is the most relevant. You can summarize the “ratio” feature data thanks to the widget “feature statistics”

Task 4. Load the train station list (3pts)

The goal of this task is to validate that geographical data can have an impact on the frequentation prediction.

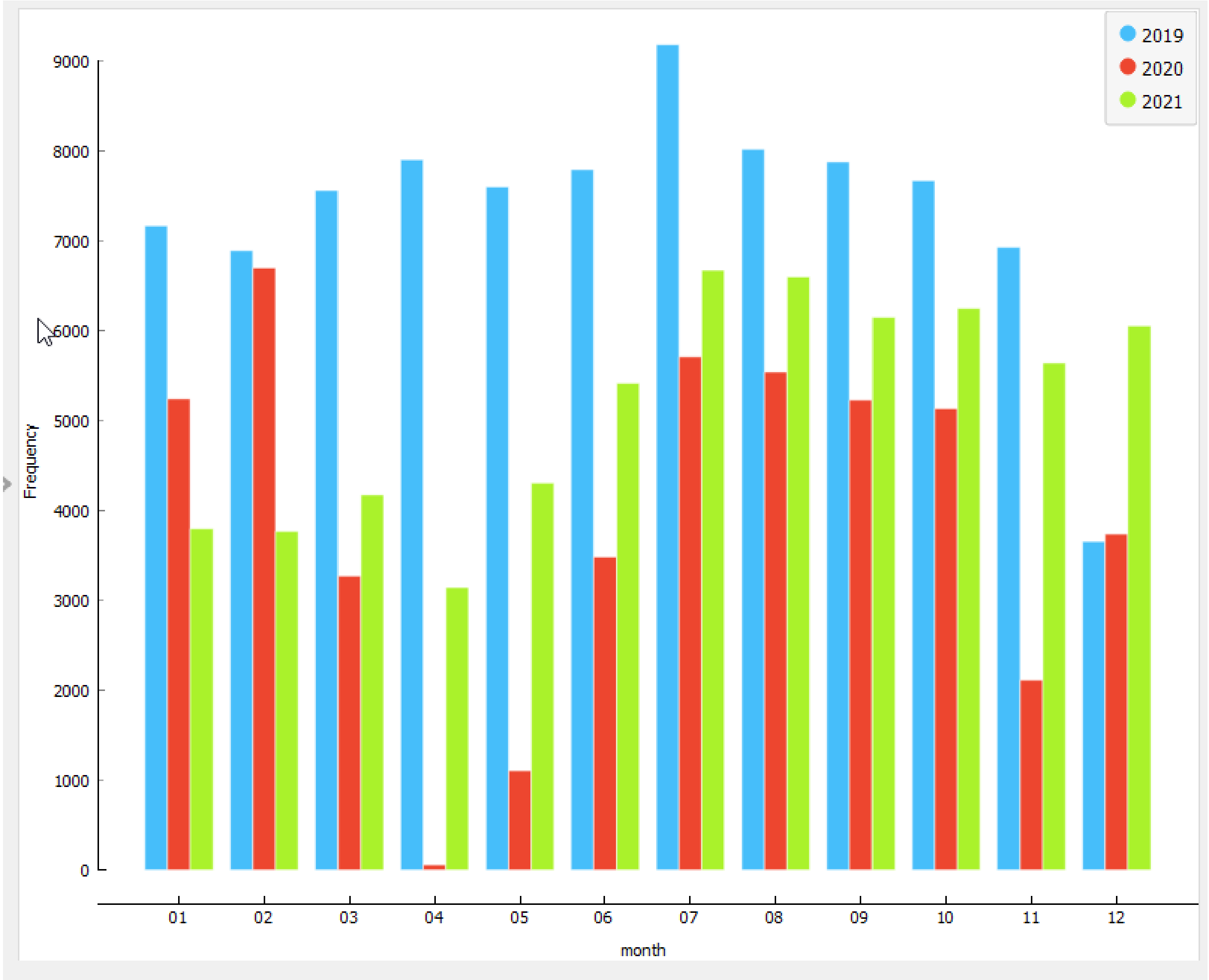
a. Load the dataset “liste-des-gares.csv”. Keep only the ones where “voyageurs” column value is “O”.
b. Merge the dataset on Code UIC
c. Summarize the lost and found count for year 2020 & 2021 per Code UIC
d. Project the volume of lost and found items count on a geomap widget

Expected result



Task 5. Reasoning on time and cycles (3pts)

a. Summarize the items loss events per month as in the following example:



b. What can you observe? Is there an impact of the month on the lost and found events?

c. Do you think we should keep year 2020 data?

End of mandatory section for the practical exam

Task 6 – Working with weekdays (3pts)

We try to find what impacts could be positively correlated with the volume of passengers.

For each day, determine if it is a holiday and then add it to a new column “isWeekday” as a categorical column that takes 2 values (“T” or “F”).

You can refine the mechanism by putting “T” to days preceding the holiday (as people could take the train in the evening the day before the holiday). You can consider week-ends as holidays.

You can inspire from this script to encode the day as the “day in week” (0 to 6, 0 = Monday… 6= Sunday)

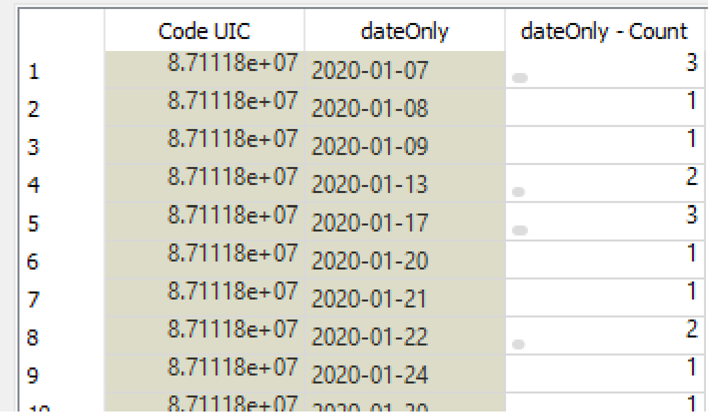
```python
import datetime
import pytz
from pytz import timezone

timezone = timezone('Europe/Paris')
for data in in_data:
    timestamp: str = str(data['timestamp'])
    date = timezone.localize(datetime.datetime.strptime(timestamp, '%Y-%m-%d %H:%M:%S'))
    weekday: int = date.weekday()
    data['weekday'] = weekday

out_data = in_data
```

Task 8 – compute the number of lost and found event per day (6pts)

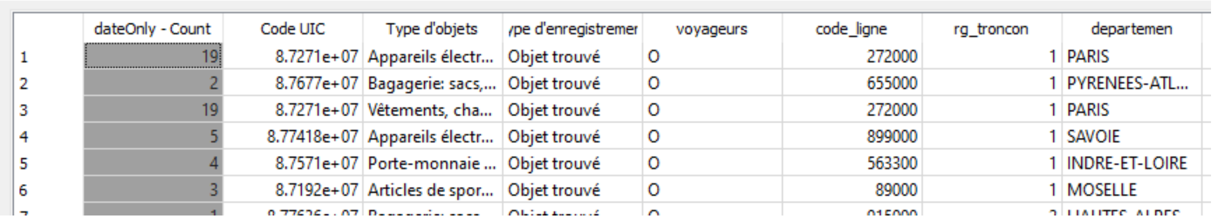
a. Regroup (using a group by widget) the occurrences per day and count the associated number of occurrences. An example of the expected result is found below:



	Code UIC	dateOnly	dateOnly - Count
1	8.71118e+07	2020-01-07	3
2	8.71118e+07	2020-01-08	1
3	8.71118e+07	2020-01-09	1
4	8.71118e+07	2020-01-13	2
5	8.71118e+07	2020-01-17	3
6	8.71118e+07	2020-01-20	1
7	8.71118e+07	2020-01-21	1
8	8.71118e+07	2020-01-22	2
9	8.71118e+07	2020-01-24	1
10	8.71118e+07	[illegible]	1

b. Merge again this computed value with matching codeuic and date parameters, you should have the count per date and per train station (with all the attributes of the train station)

c. You should define occurrences by day count as your target variable



	dateOnly - Count	Code UIC	Type d'objets	ype d'enregistremer	voyageurs	code_ligne	rg_troncon	departemen
1	19	8.7271e+07	Appareils électr...	Objet trouvé	O	272000	1	PARIS
2	2	8.7677e+07	Bagagerie: sacs,...	Objet trouvé	O	655000	1	PYRENEES-ATL...
3	19	8.7271e+07	Vêtements, cha...	Objet trouvé	O	272000	1	PARIS
4	5	8.77418e+07	Appareils électr...	Objet trouvé	O	899000	1	SAVOIE
5	4	8.7571e+07	Porte-monnaie ...	Objet trouvé	O	563300	1	INDRE-ET-LOIRE
6	3	8.7192e+07	Articles de spor...	Objet trouvé	O	89000	1	MOSELLE
7	1	[illegible]	Bagagerie: sacs,...	Objet trouvé	O	[illegible]	2	HAUTES-ALPES

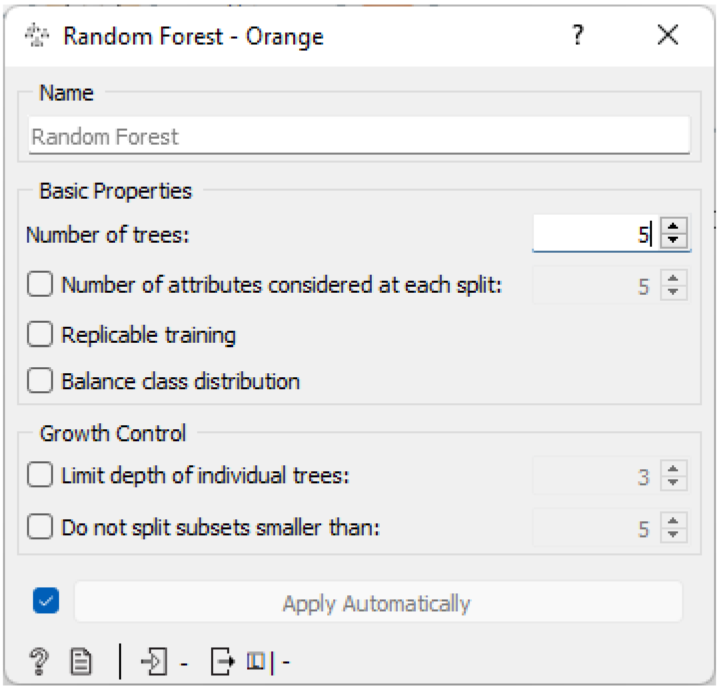
Task 7 – Build a predictive system using the input data (2pts)

As the goal is to predict the number of transiting passengers (so, a continuous variable), the exercise of predicting this value is called “regression”. We are using this predictive system as a black box for now.

For that purpose, put a random forest (initialized with 5 trees) and a test and score widget.

The goal is to observe the improvement of the RMSE (Root Mean Squared Error) metric, which symbolizes the errors the predictive systems has made while under test, the lower is the better.

The fact that we observe an error metric is because we try here to predict a number of passengers, and not a class.



Random Forest - Orange

Name

Random Forest

Basic Properties

Number of trees: 5

☐ Number of attributes considered at each split: 5

☐ Replicable training

☐ Balance class distribution

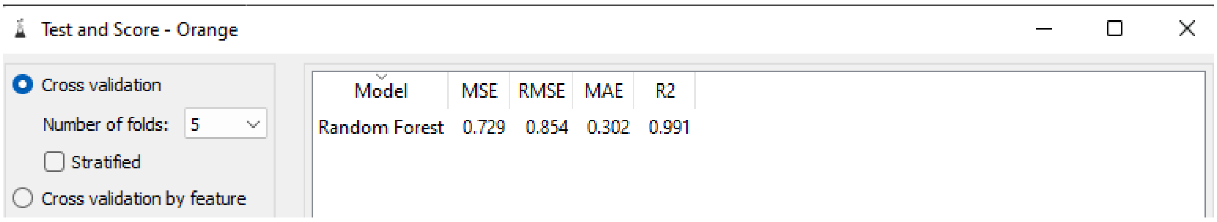
Growth Control

☐ Limit depth of individual trees: 3

☐ Do not split subsets smaller than: 5

☑ Apply Automatically

As a reference: a proper baseline RMSE (when only date is selected as an input variable) value is around **1.0** for this exercise.



Test and Score - Orange

● Cross validation

Number of folds: 5

☐ Stratified

○ Cross validation by feature

Model	MSE	RMSE	MAE	R2
Random Forest	0.729	0.854	0.302	0.991

Non guided section

Task 8- Identify input variables that could be relevant to predict the volume of passengers (1pt)

Identify what data are available among all your datasets, select the features available before the predictions.

Task 9 – find other hypothesis or datasets to improve the accuracy of your system. (3pts)

Find other datasets and /or transformation ideas on the existing data and combine them together with the existing ones to figure out what can improve the RMSE metric of the predictive system.

Try at least 3 hypotheses.

Possible ideas:

- Holidays / Public holidays
- Train station connectivity (how much the train station is connected to other)
- Train station activity (how many trips are departing/arriving to a train station)
- Special events: football games, concerts etc.
- Weather (for last minute travelers)