Can we predict Which will be the most Popular destinations to Fly to and reduce CO2 emissions??

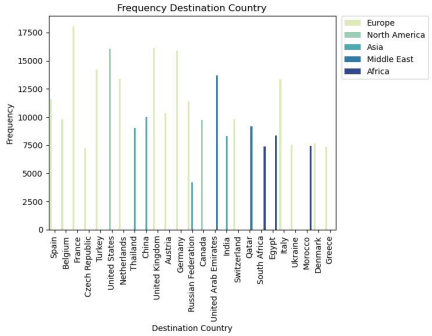
- In this presentation explains
  - The type of data we were provided
  - The thoughts behind our actions
  - And, the results obtained
- The Data Set:
  - A train set with **9 columns** and 2796982 rows
  - A test set with **8 columns** and 69246 rows
  - Our Target to Predict is the **DESTINATION COUNTRY!**

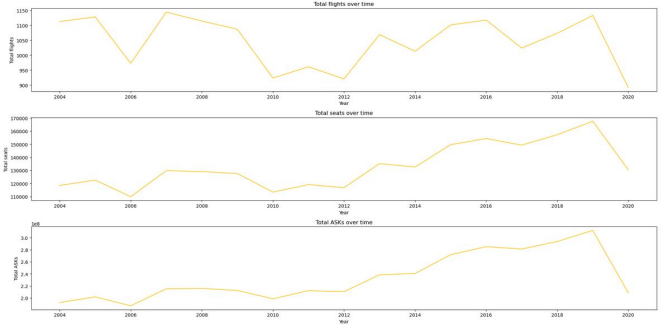
	Year	Month	Origin Country	Origin Continent	Destination Country	Destination Continent	Total flights	Total seats	Total ASKs
0	2009	7	United Kingdom	Europe	13	Europe	9,032	1,531,683	2,447,559,137
1	2008	4	Lebanon	Middle East	9	Europe	5	760	2,389,940
2	2005	4	Switzerland	Europe	11	Europe	1,471	158,661	66,533,450
3	2016	8	Israel	Middle East	19	Europe	117	23,366	61,557,637
4	2019	2	Albania	Europe	8	Europe	80	12,854	9,837,347
...	...	...	...	...	...	...	...	...	...
276977	2012	7	Iraq	Middle East	21	Europe	4	744	2,635,372
276978	2007	6	Cyprus	Europe	4	Europe	64	9,519	19,409,611
276979	2007	9	Panama	Central America	17	North America	442	63,061	161,243,990
276980	2015	11	Russian Federation	Europe	9	Europe	90	13,422	30,147,391
276981	2016	2	Iceland	Europe	13	Europe	10	2,000	8,046,076

276982 rows × 9 columns

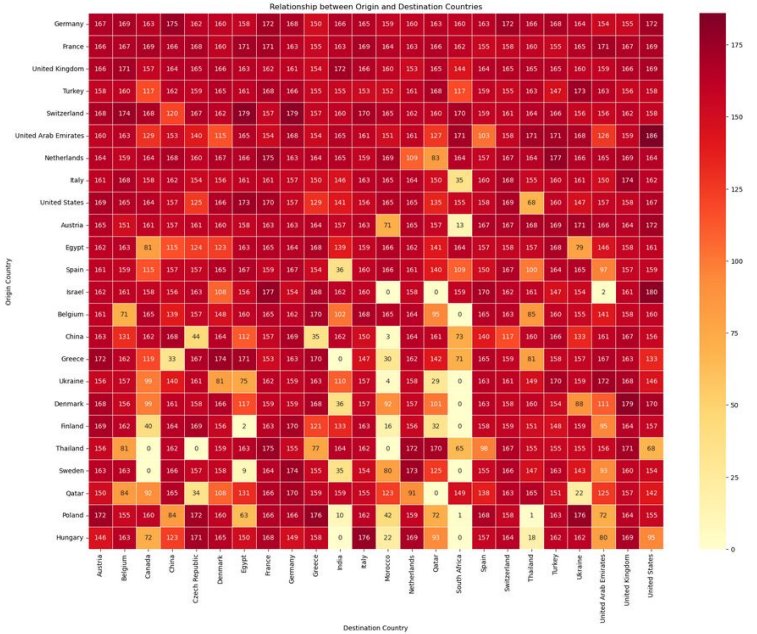
Exploratory Data Analysis (EDA)

- During the EDA we have observed:
  - Total flights over time don’t have a general tendency
  - Total seats and ASKs in general increase
  - We see that total flights, seat and ASKs decrease abruptly in 2019, because of the Covid
  - Per total number of flight, the most frequent country is USA in North America
  - Per frequency of appearance, the most frequent origin country is France, and most frequent Continent, both of origin and destination is Europe.





- Correlation between the top 25 destination countries flights for each combination of origin country and destination country
- The image has a lot of information, the cells with reddish colors, indicate that between these two countries, the one in the x and the one in the y, there are many flights, while the cells with pale yellow color has almost non flight



Model Evaluation

- Our thoughts behind our actions to look for the best pre-processing of the data and the best model:
  - The Origin Continent destination and continents have been treated as Labels with no implicit order
  - The Origin country, since is similar to our target (destination Country), has been treated as numerical, as has been given by the organizers
  - The total flights, seats and ASKS, has been treat as they are numerical with no outliers.
  - We do not treat Year and Month as cyclical data although we have seen it is and it does have some seasonal importance because after checking different preprocessing we have come out that it is not improving the model. This, with more time, could be rethinked and do some feature engineering to see if it is really like this or it could help to predict.

- Models Evaluation:
  - We have checked several models and the one that gave better results was the **Random Forest Classifier**

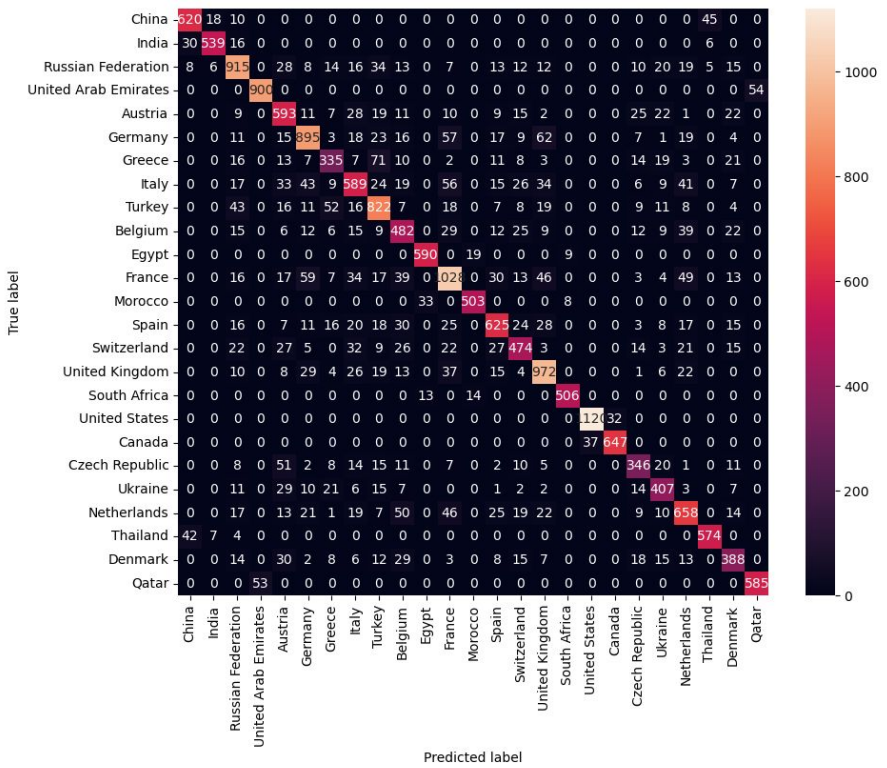
	model	f1__macro_cv	f1_macro_std_cv
0	knc	0.431902	0.006814
1	gnb	0.177379	0.003731
2	dtc	0.658262	0.004981
3	svc	0.023378	0.002447
4	rfc	0.693670	0.003601

- Hyper-Parameter Tuning:
  - We lood for the best parameters of the **Random Forest Classifier** with our dataset.

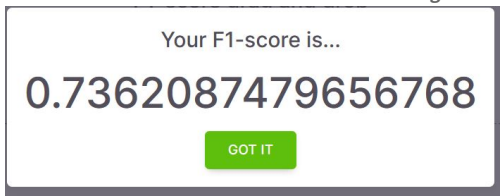
Confusion Matrix with Training data:

- The highest numbers, in reddish, are in the diagonal which corresponds to the **True positive**, so, the ones that the model got right.
- The **bluish colors** corresponds to the **False negative and false Positive**. which are represented by the recall and the precision. Looking at this numbers below, well they are quite similar and equilibrated, so no more, false positive than false negatives.
    - Precision: 0.8060774864234902
    - Recall: 0.80565

- While 0.80 is not a super-high score is not bad, it could be said in general (since all scores are very similar) that **around the 80% of times, these model gets the target right,** we could aim for more, but we didn't had much time.



Final result with test data!unt of flights.



So, as **future work** trying to implement the year and month implicit cyclicity, in the EDA, we clearly have seen a year seasonality of the amount of flights.