KU LEUVEN

Statistical Consulting

Project: Travel Agency

Katia Aerts, Thomas Umberto Grava, Huu Duc Luu,
Alexander Saines Fajardo, Gabby Vinco, Ziyue Zhu

KU Leuven

April 28th, 2022

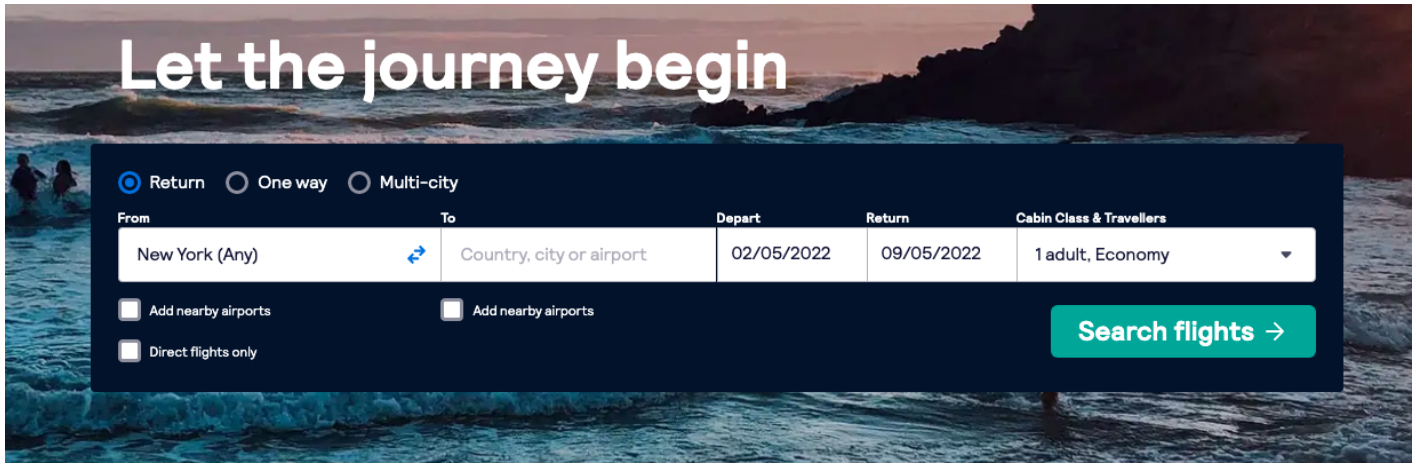
0 Predicting flight delays and cancellations

1. StatTravels
2. Datasets
3. Data preparation
4. Modeling
5. Evaluation
6. Deployment

1 Outline

1 Travel agency: 'StatTravels'

- Web tool where customers can choose the best flight
- Decrease number of complaints about delayed and cancelled flights
- Add information about predicted delays and cancellations to the existing interface.



Existing web tool from www.Skyscanner.net

2 Outline

2 Datasets

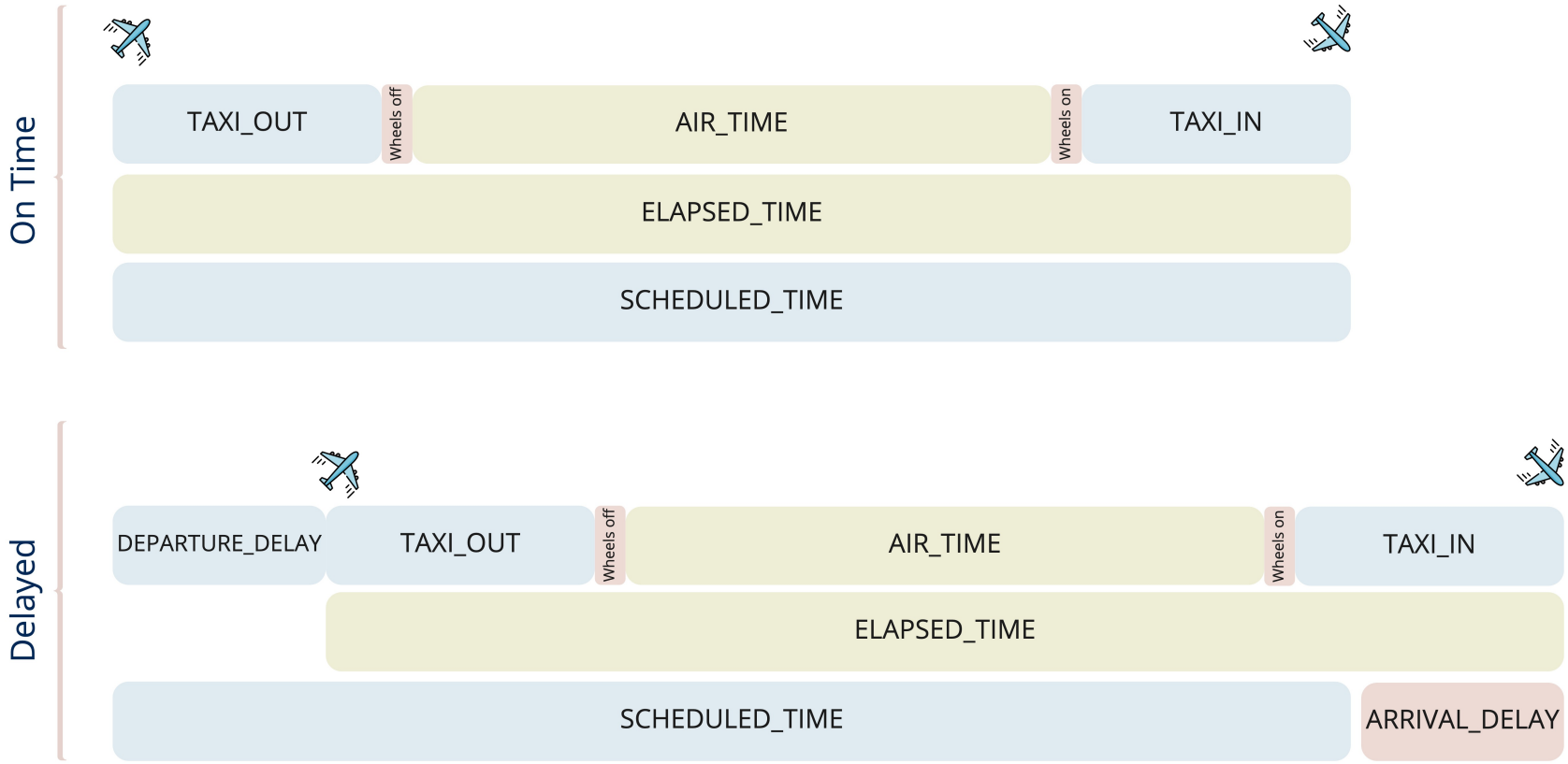
Airports
(n=322)

Airlines
(n=14)

Flight data 2019
March, April, May, June, July
(n=2 458 628)

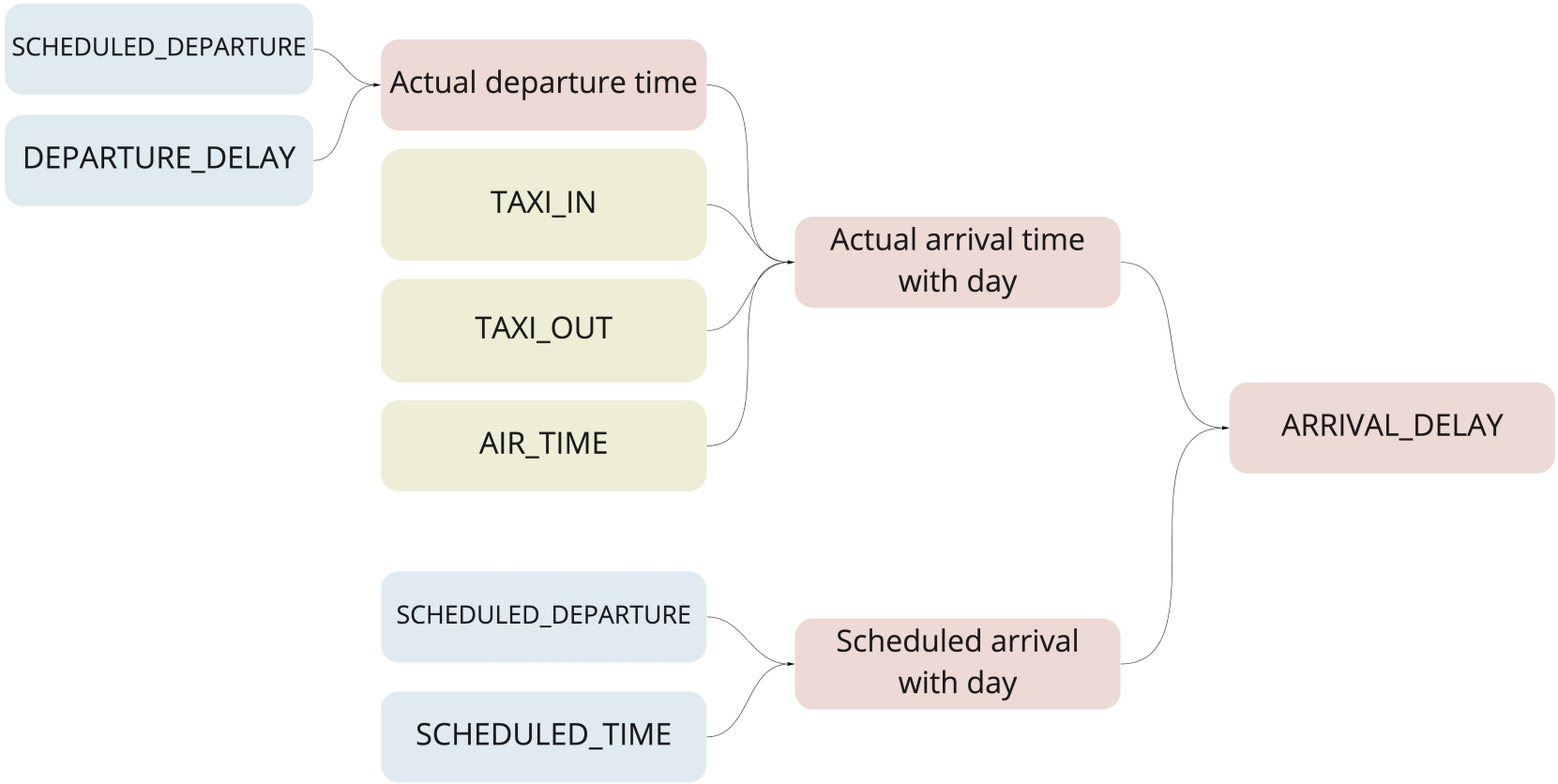
Scheduled flight data 2022
August
(n=469 968)

2 Flight time

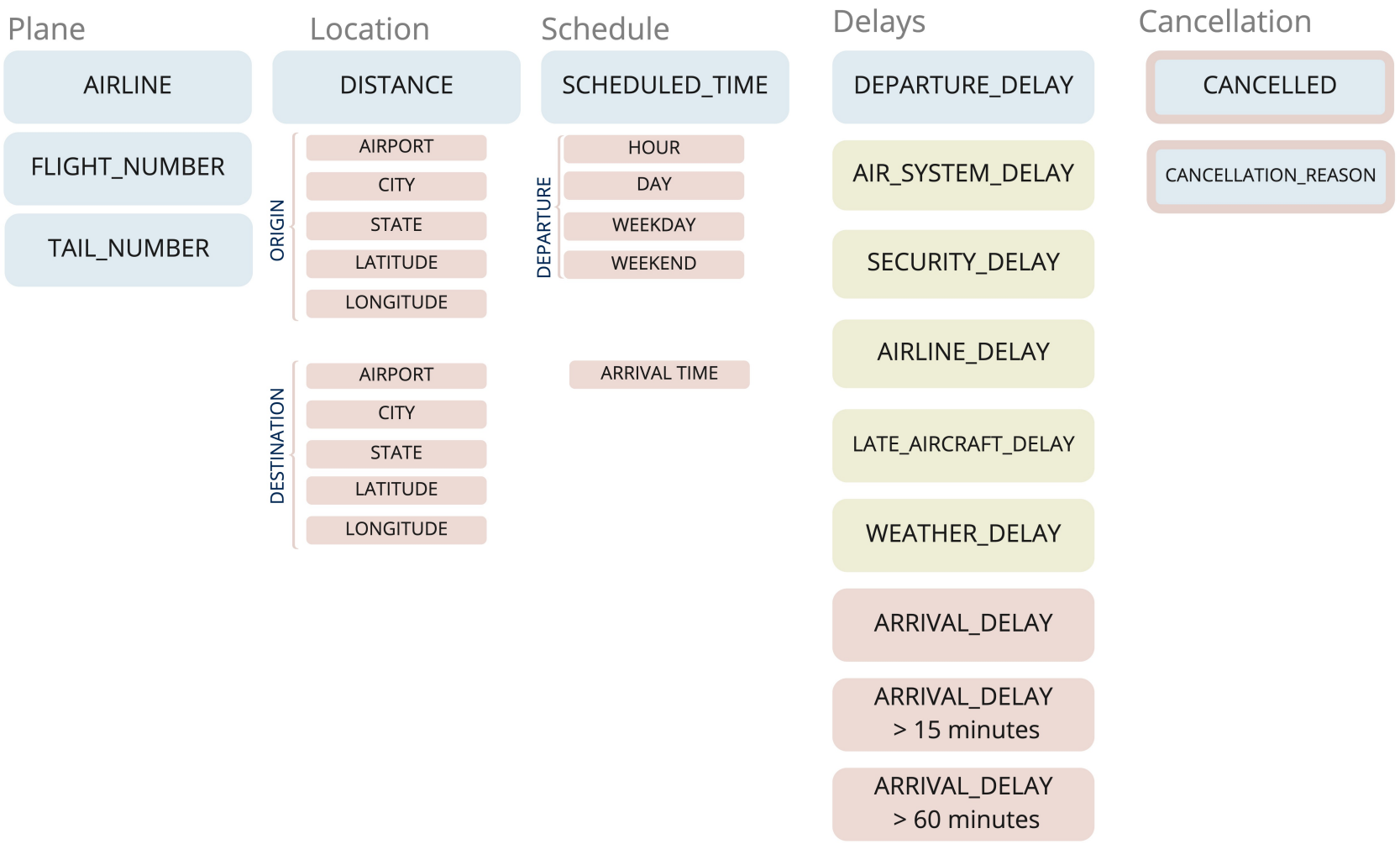


3 Outline

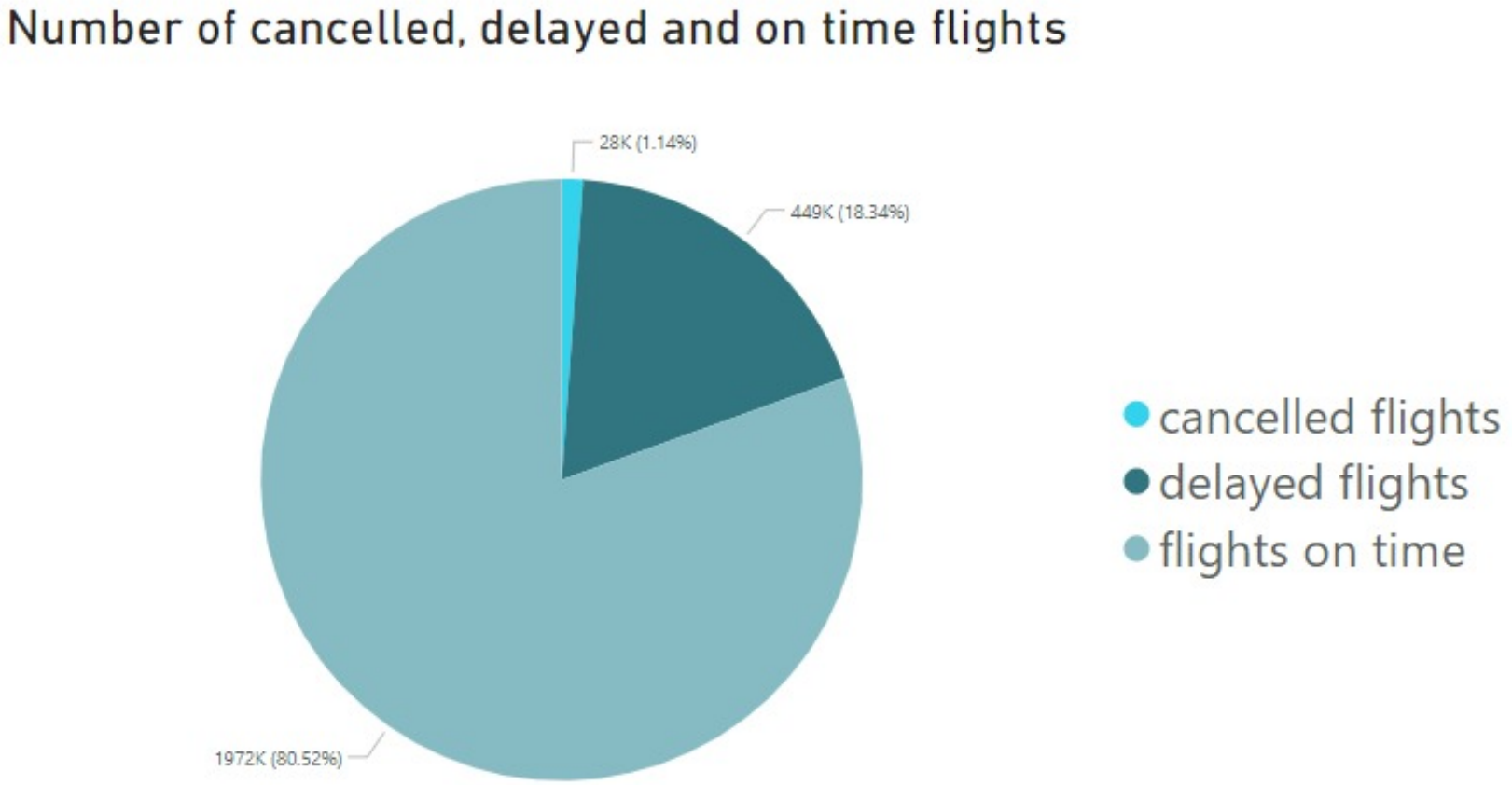
3 Arrival delay



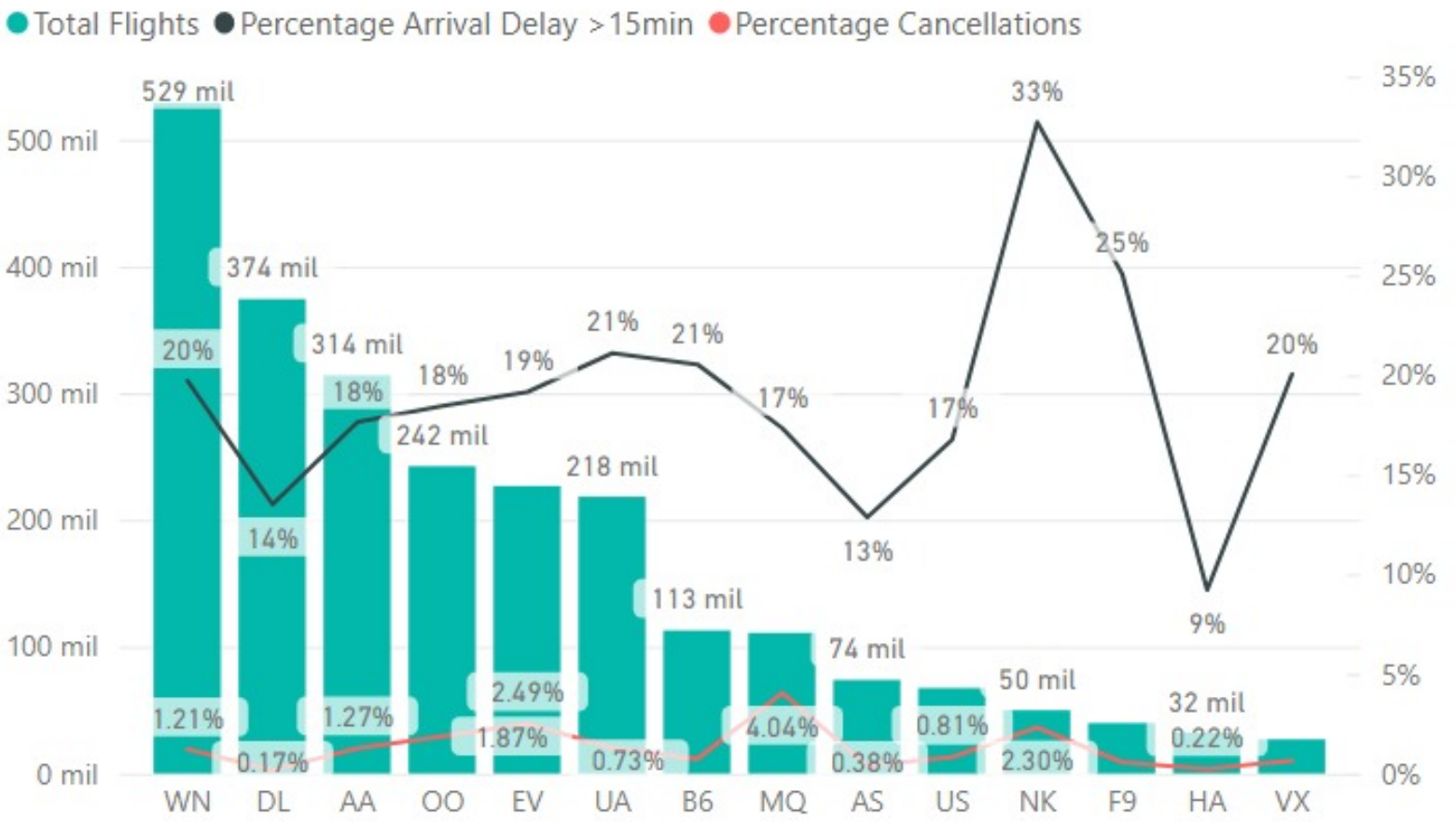
3 Combined dataset



3 Number of Delays

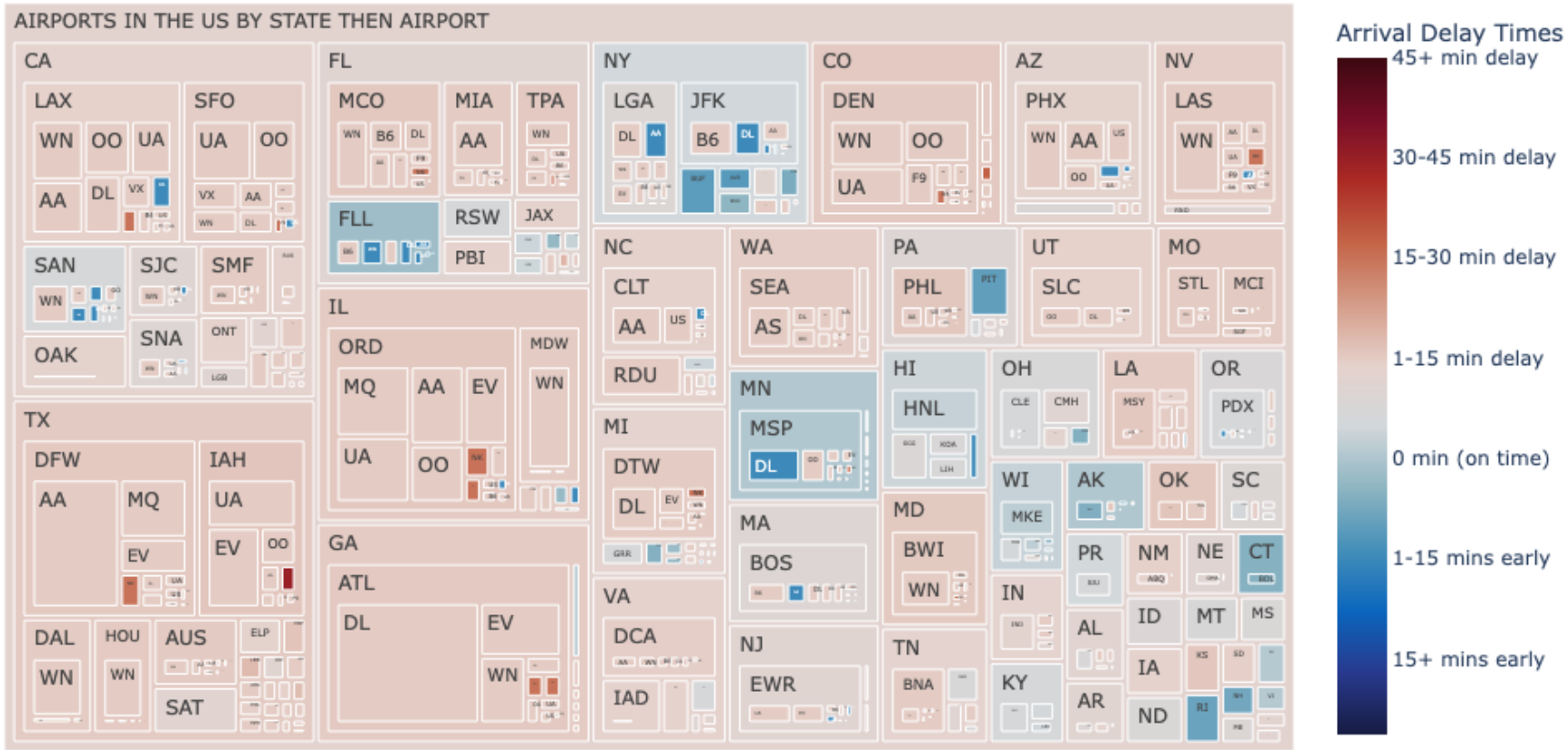


3 Number of Flights and delays by airline



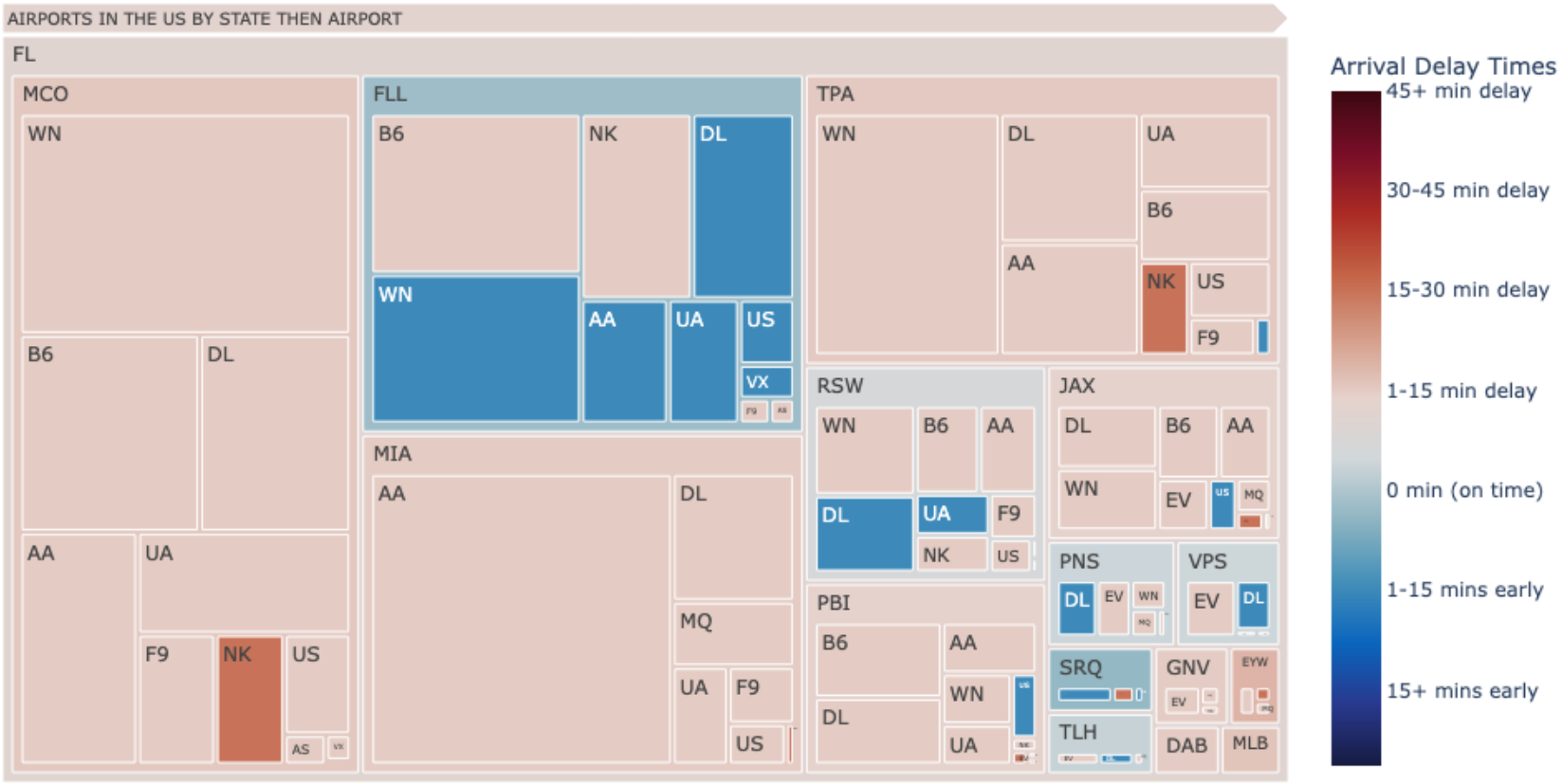
KU LEUVEN

3 Average arrival delay



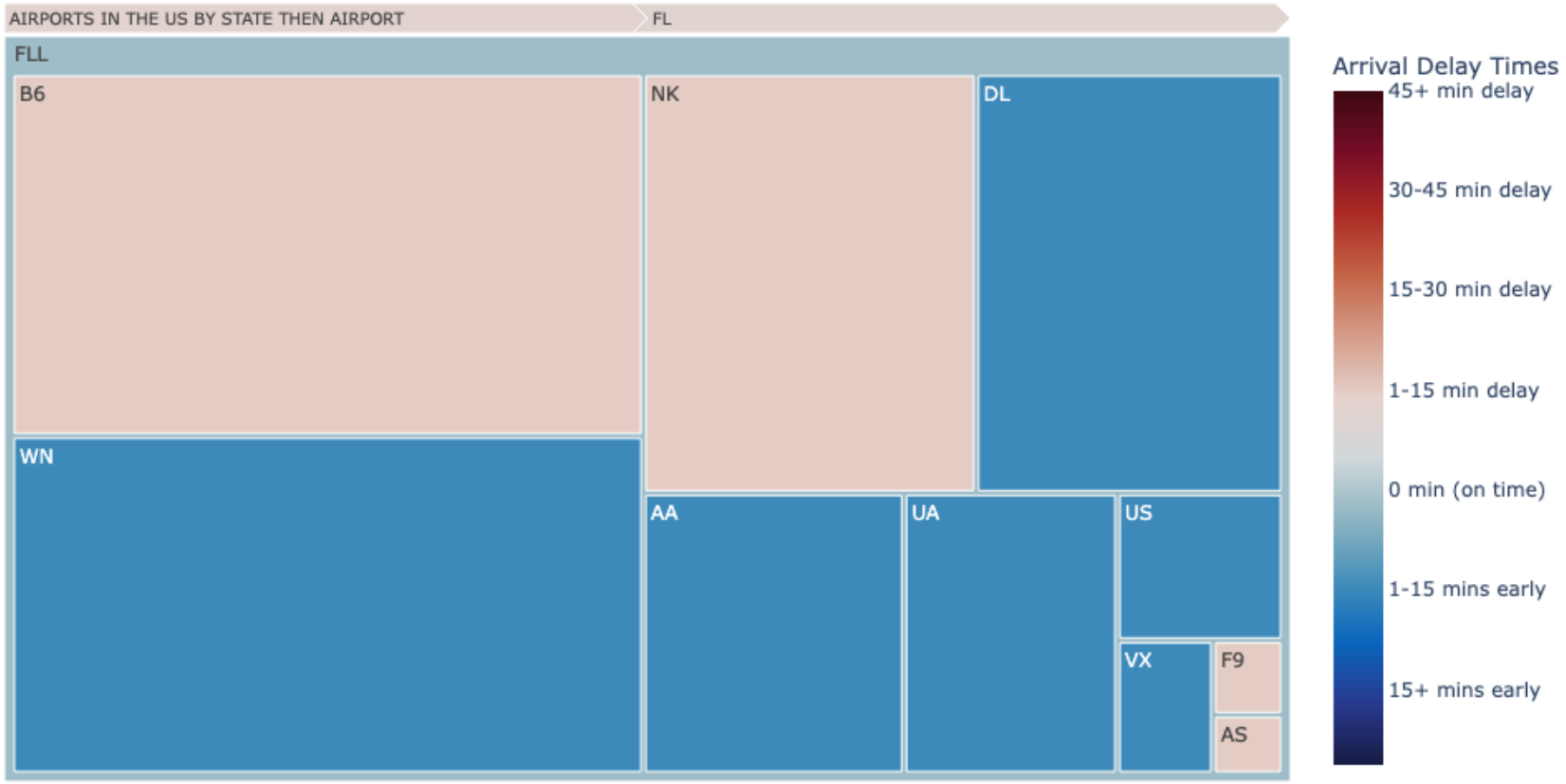
KU LEUVEN

3 Average arrival delay: Florida

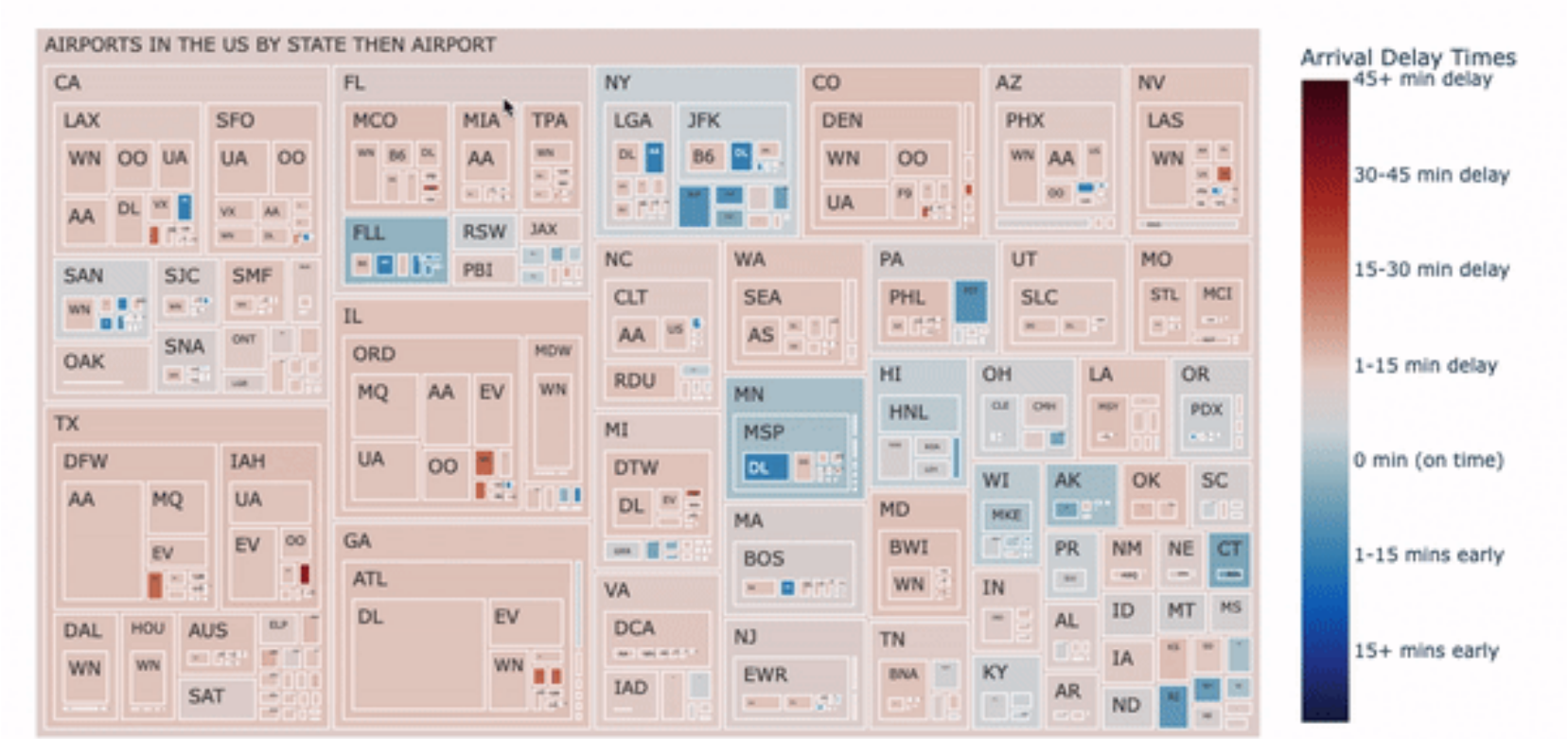


KU LEUVEN

3 Average arrival delay: Florida - Fort Lauderdale - Hollywood International Airport

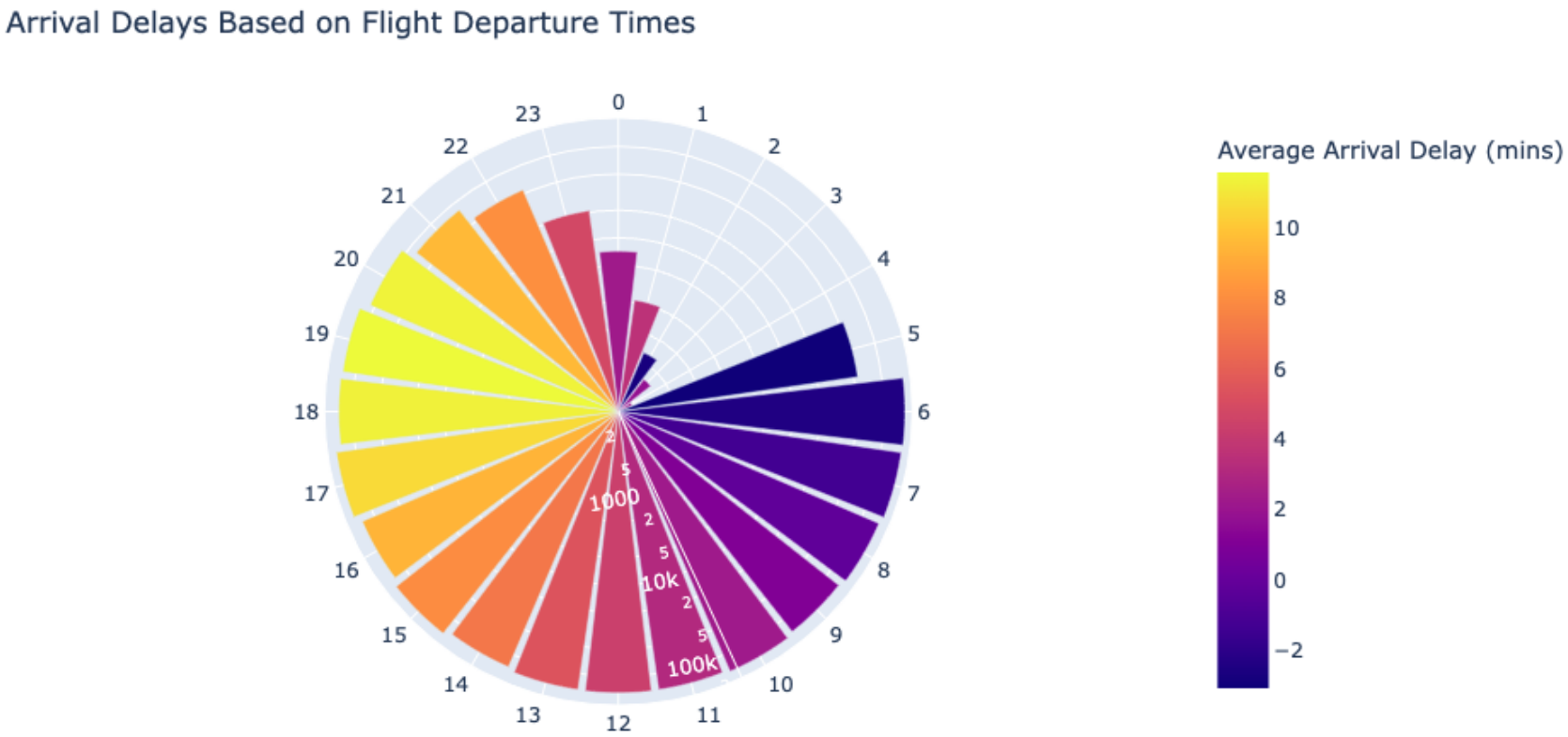


3 Average arrival delay by airport and airline: Florida



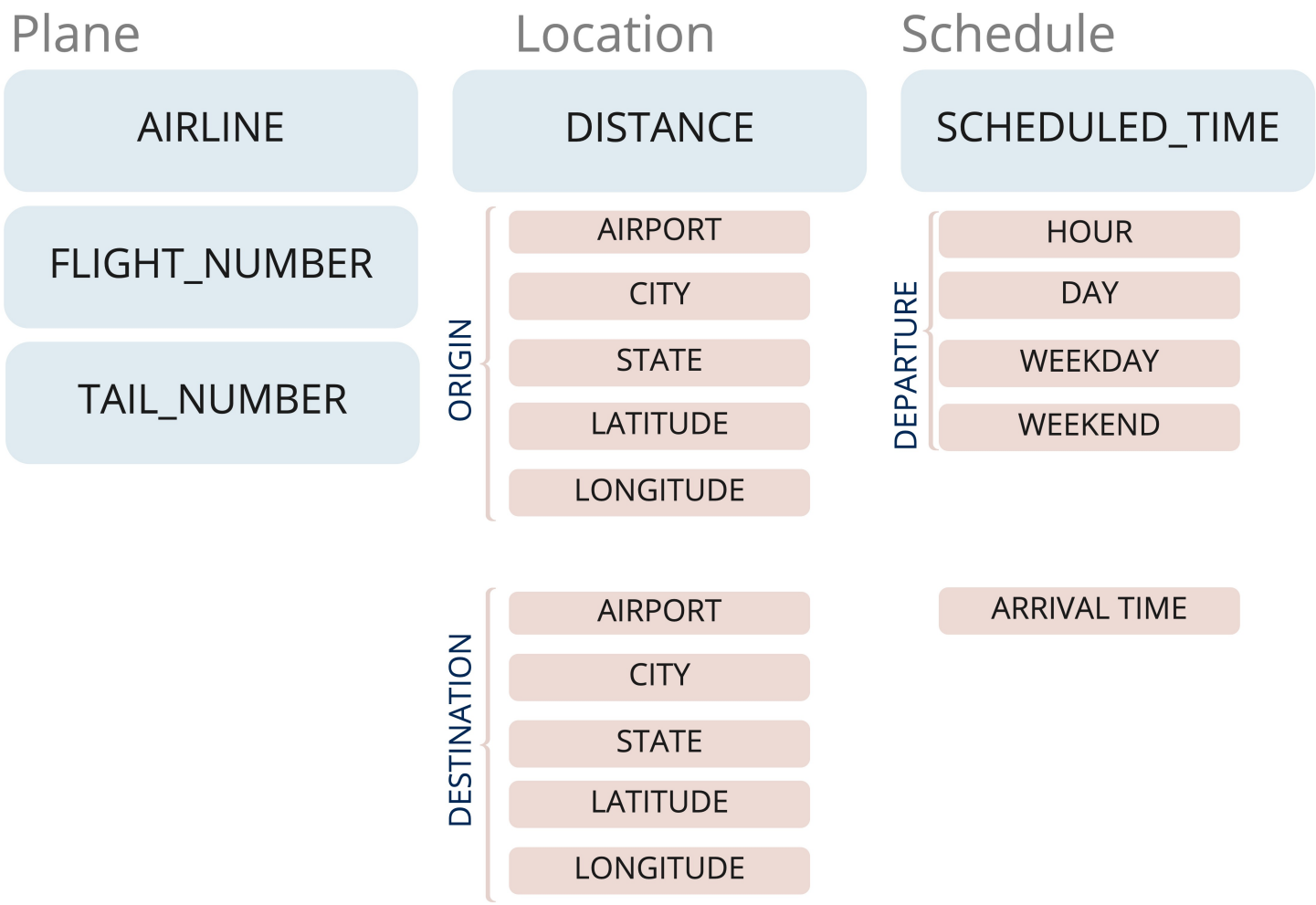
KU LEUVEN

3 Hourly overview

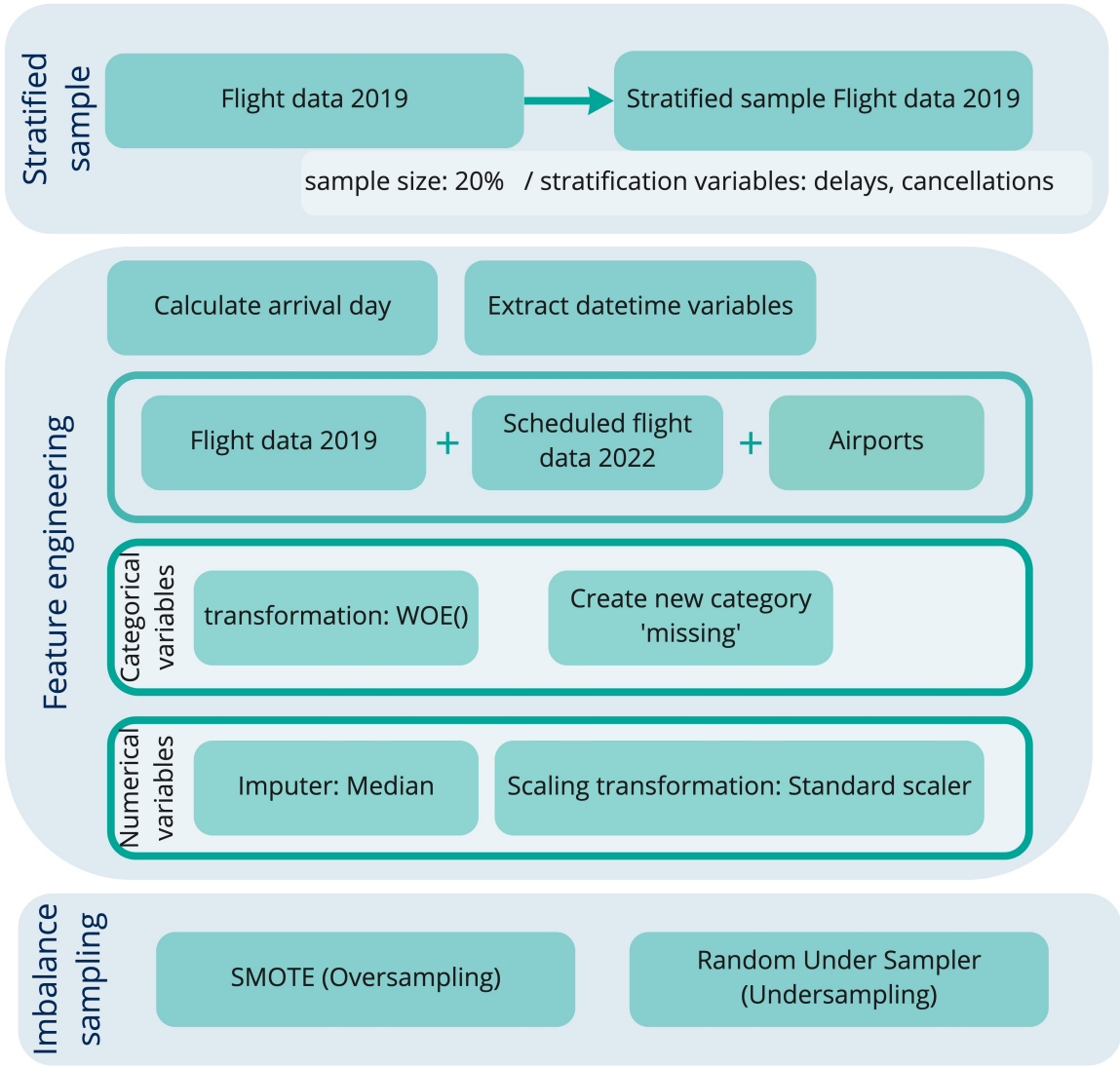


4 Outline

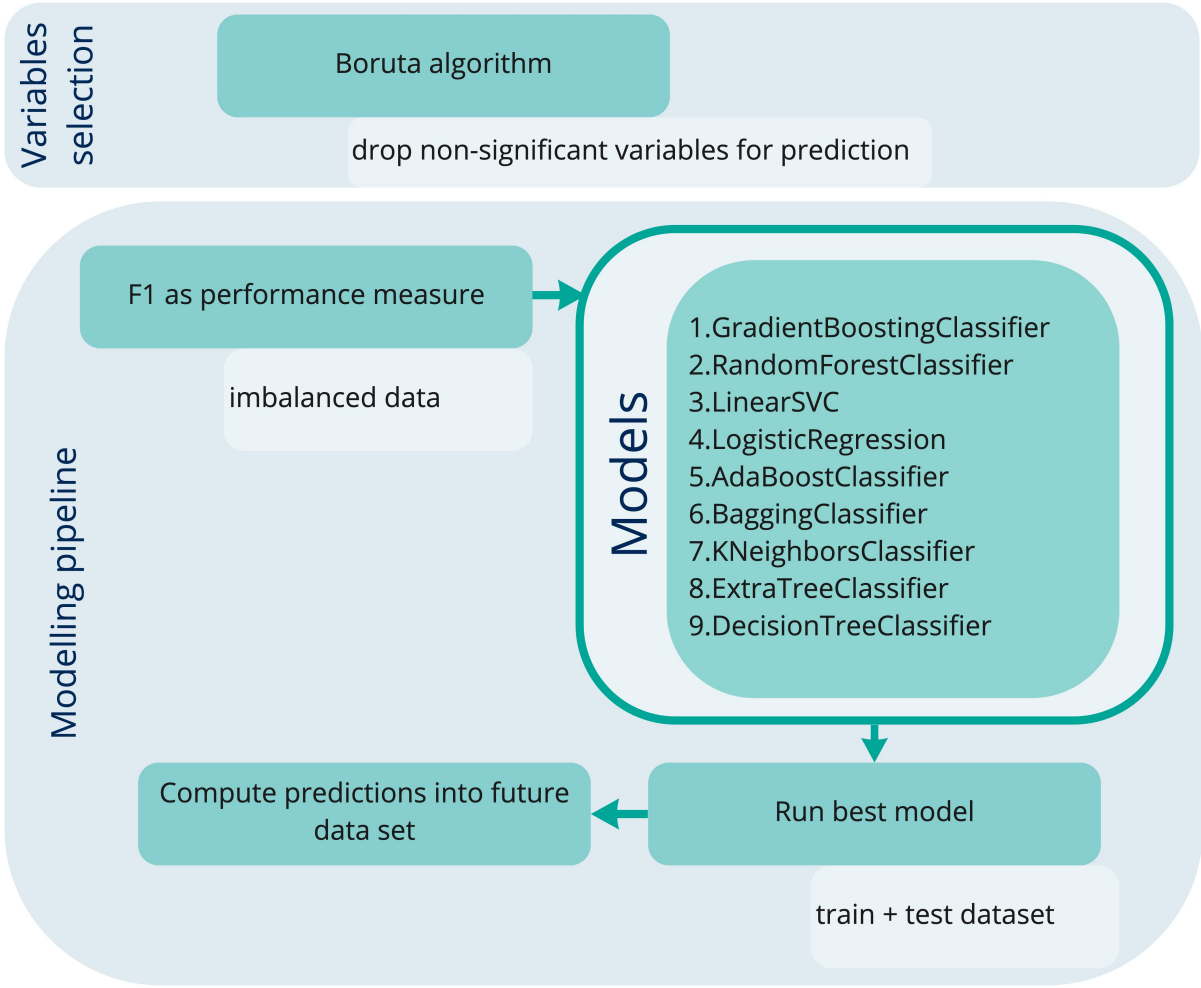
4 Prediction features



4 ML workflow



4 ML workflow (continued)



5 Outline

5 Model comparisons

Delays		
Name	Train F1 Score	Test F1 score
GradientBoostingClassifier	0.7449	0.7906
RandomForestClassifier	1.0000	0.7782
LinearSVC	0.6931	0.7738
LogisticRegression	0.6940	0.7679
AdaBoostClassifier	0.7168	0.7674
BaggingClassifier	0.9823	0.7663
KNeighborsClassifier	0.7953	0.6996
ExtraTreeClassifier	1.0000	0.6670
DecisionTreeClassifier	1.0000	0.6626

Cancellations		
Name	Train F1 Score	Test F1 score
RandomForestClassifier	1.0000	0.9847
BaggingClassifier	0.9988	0.9834
DecisionTreeClassifier	1.0000	0.9665
GradientBoostingClassifier	0.9438	0.9656
ExtraTreeClassifier	1.0000	0.9534
AdaBoostClassifier	0.9138	0.9289
KNeighborsClassifier	0.9696	0.9270
LogisticRegression	0.8303	0.8880
LinearSVC	0.8304	0.8871

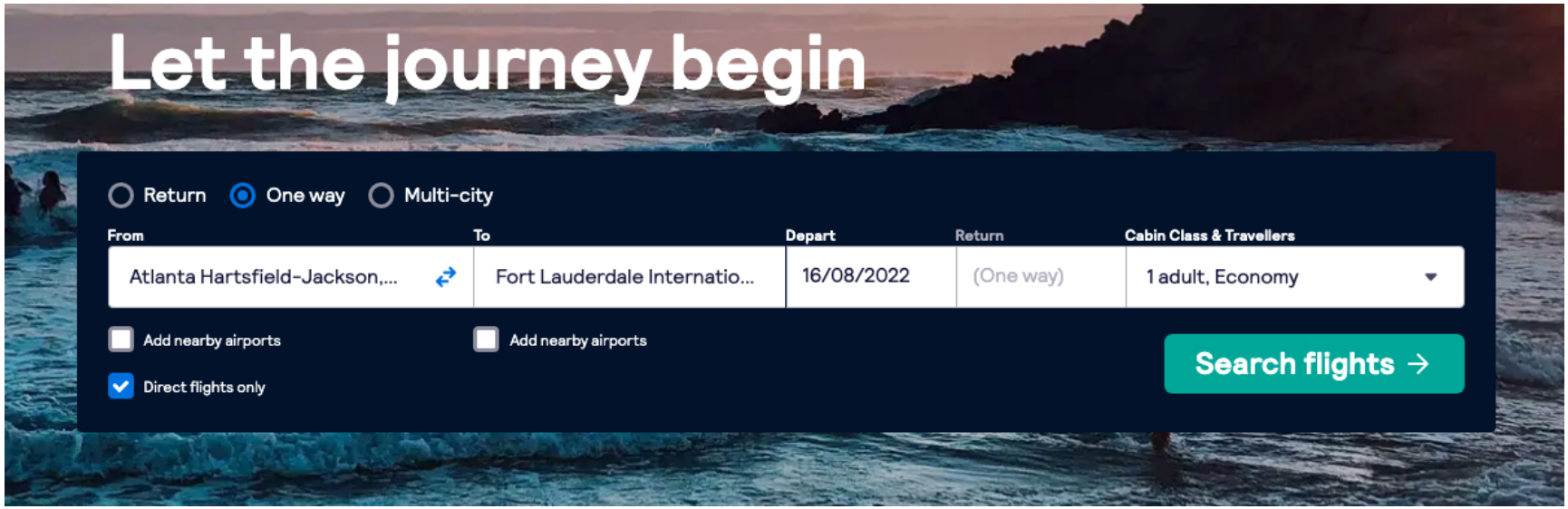
To predict delays, we used the GradientBoostingClassifier.
To predict cancellations, we used the RandomForestClassifier.

5 Feature importance

Features	Importance		Features	Importance	
	Cancellation	Delayed		Cancellation	Delayed
Tail Number	**24%**	15.02%	Destination Airport	3%	0.36%
Weekday	**14%**	**22.64%**	Distance	3%	0.03%
Flight Number	**12%**	11.72%	Destination City	2%	0.34%
Departure Day	9%	**16.93%**	Origin Longitude	2%	\
Airline	5%	1.97%	Destination Longitude	2%	\
Weekend	3%	8.33%	Origin Latitude	2%	\
Origin Airport	3%	1.15%	Destination Latitude	2%	\
Departure Hour	3%	**18.84%**	Origin State	2%	\
Origin City	3%	0.17%	Destination State	2%	\
Arrival Hour	3%	2.51%			

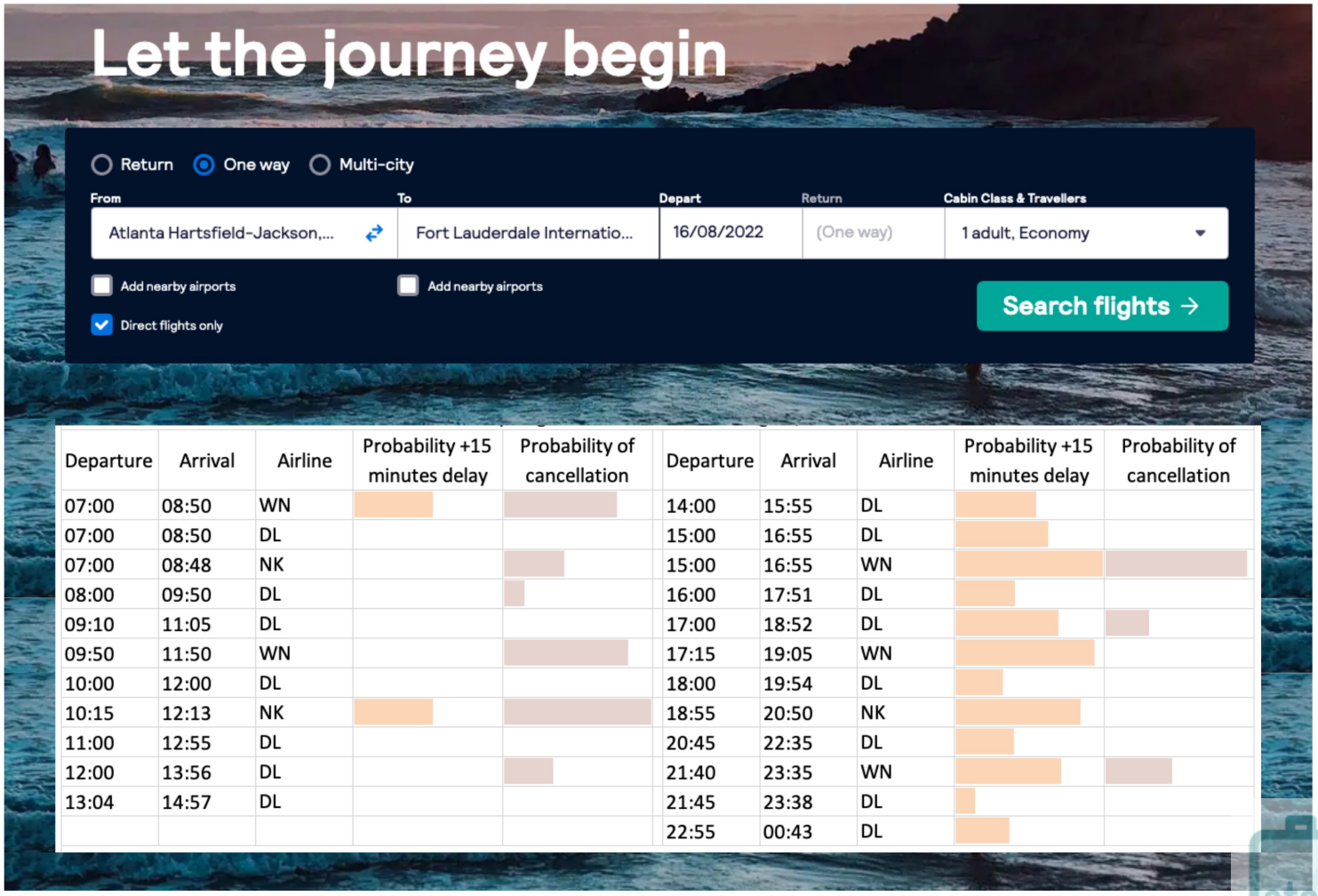
6 Outline

6 Booking a flight at StatTravels



Existing web tool from www.Skyscanner.net

6 Mock-up results



Departure	Arrival	Airline	Probability +15 minutes delay	Probability of cancellation	Departure	Arrival	Airline	Probability +15 minutes delay	Probability of cancellation
07:00	08:50	WN			14:00	15:55	DL		
07:00	08:50	DL			15:00	16:55	DL		
07:00	08:48	NK			15:00	16:55	WN		
08:00	09:50	DL			16:00	17:51	DL		
09:10	11:05	DL			17:00	18:52	DL		
09:50	11:50	WN			17:15	19:05	WN		
10:00	12:00	DL			18:00	19:54	DL		
10:15	12:13	NK			18:55	20:50	NK		
11:00	12:55	DL			20:45	22:35	DL		
12:00	13:56	DL			21:40	23:35	WN		
13:04	14:57	DL			21:45	23:38	DL		
					22:55	00:43	DL		

6 Future improvements

- Which delay time will lead to dissatisfaction?
  - Very personal and context related.
  - Mock-up worked with $>$15 minutes delay
  - Adding a slider to give the client the opportunity to choose the cap.
- Recommend nearby airports

Questions & Answers