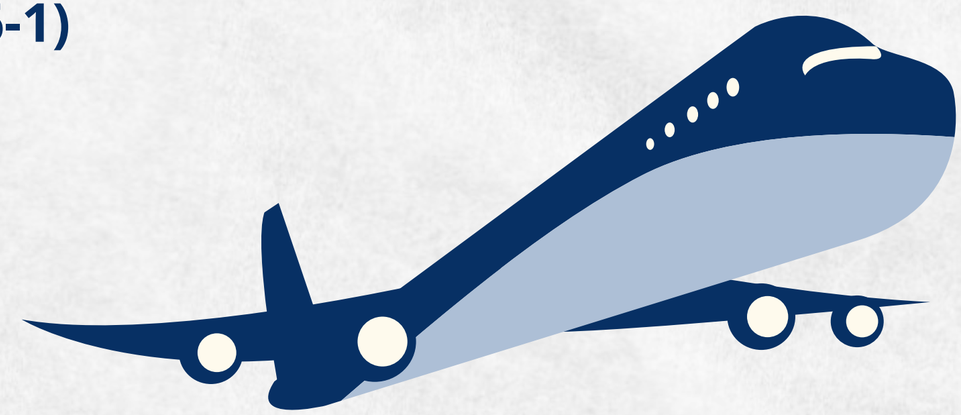
PREDICTING FLIGHT PRICES

Project Presentation (ECON2306-1)

Céline Lam & Adelia Rahmawati



June 7, 2023

AGENDA

01

Webscraping

- Selenium
- Booking.com
- 11,619 data

02

Data Preparation

- Data cleaning
- Preprocessing

03

Data Visualization

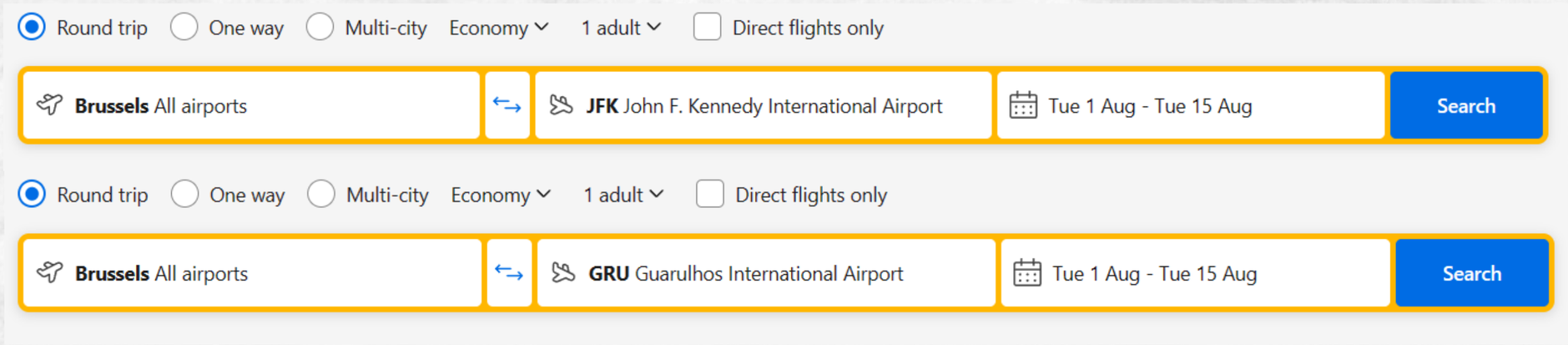
- Seaborn
- Matplotlib
- Plotnine

04

Data Modelling

- Logistic Regression
- Linear Regression
- Price prediction

MOTIVATIONS & OBJECTIVE



For travelers seeking to minimize costs and maximize their travel budgets

- Model that can accurately **predict** the prices of round-trip airline tickets
- **Optimal time** to purchase a ticket for each destination
- Cost **savings** in air travel during peak travel times

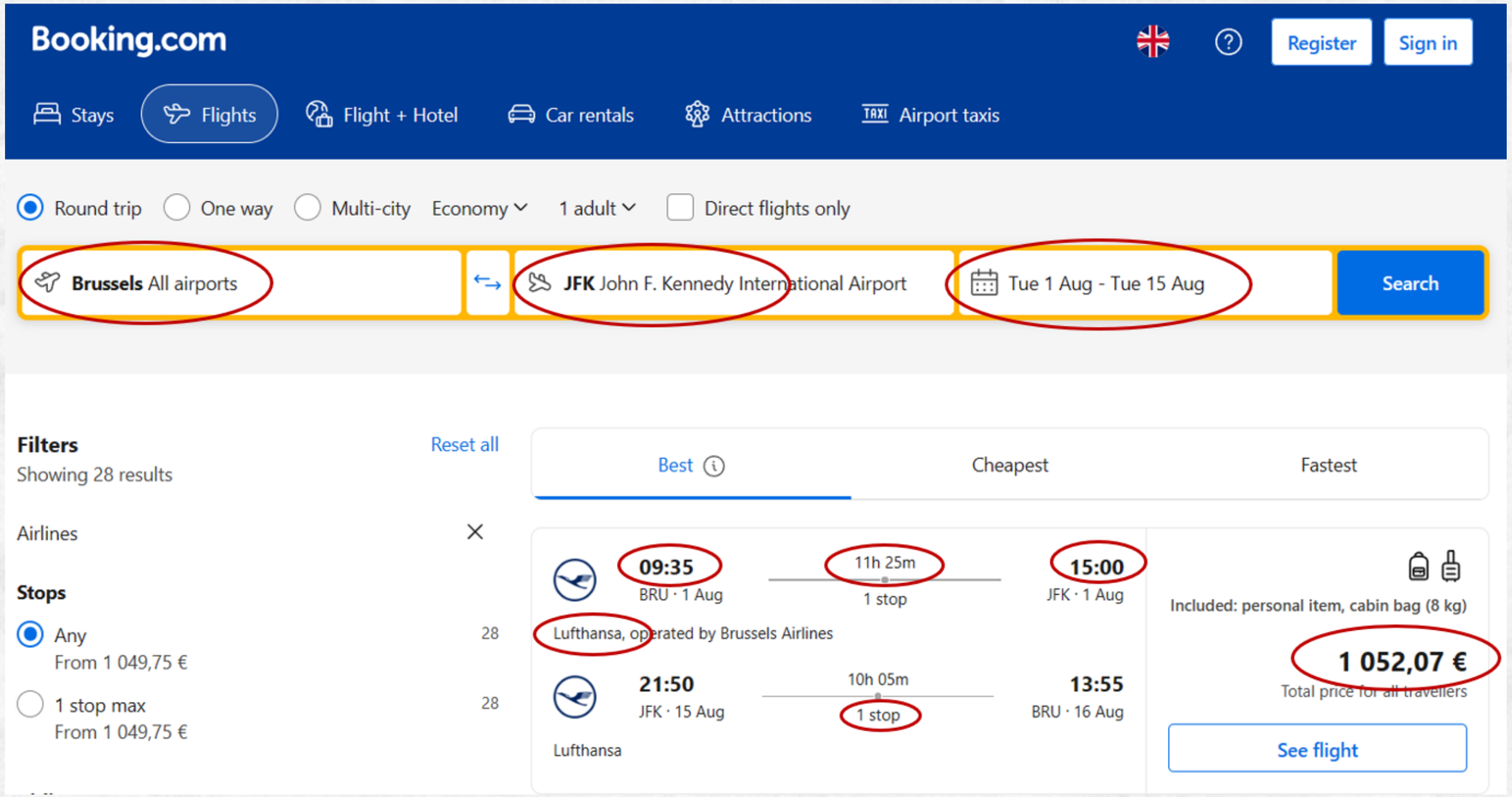
Lufthansa

SWISS

RESEARCH QUESTIONS

- What are the primary **factors** that influence ticket prices for Lufthansa and Swiss flights to Sao Paulo and NYC?
- How does the time at which flight ticket data is collected **impact ticket prices**?
- When is the **optimal time** to collect flight ticket data to minimize cost?
- How accurately can future flight prices be predicted using a **model trained** on historical data?

WEBSCRAPING



From May 20 to May 31

- Departure city
- Arrival city
- Departure/arrival date
- Departure/arrival time
- Airline company
- Number of stops
- Duration of flight
- Flight ticket price

+ current date and time

DATA PREPARATION

	out_airline_company	in_airline_company	dep_city	arr_city	out_dep_date	out_dep_time	out_duration	out_stop_num	out_arr_date	out_arr_time	in_dep_date	in_dep_time	in_duration	in_stop_num	in_arr_date	in_arr_time	hour_scrap	day_scrap	ticket_price
0	Lufthansa	Lufthansa	BRU	JFK	2023-08-01	09:35	0 days 11:25:00	1	2023-08-01	15:00	2023-08-15	15:55	0 days 09:40:00	1	2023-08-16	07:35	16	20	1176.01
1	Lufthansa	Lufthansa	BRU	JFK	2023-08-01	13:15					2023-08-15	17:30	0 days 16:20:00	1	2023-08-16	15:50	16	20	923.39
2	Lufthansa	Lufthansa	BRU	JFK	2023-08-01	08:50					2023-08-15	15:55	0 days 09:40:00	1	2023-08-16	07:35	16	20	1176.01
3	Lufthansa	Lufthansa	BRU	JFK	2023-08-01	09:35					2023-08-15	15:55	0 days 10:40:00	1	2023-08-16	08:35	16	20	1176.01
4	Lufthansa	Lufthansa	BRU	JFK	2023-08-01	09:35					2023-08-15	17:30	0 days 10:50:00	1	2023-08-16	10:20	16	20	1176.01
...	...	...	...	...	...	...					...	...	...	...	...	...	...	...	...
11614	Swiss	Swiss	BRU	GRU	2023-08-01	15:00					2023-08-15	18:45	1 days 08:35:00	2	2023-08-17	08:20	12	31	2126.92
11615	Swiss	Swiss	BRU	GRU	2023-08-01	09:15					2023-08-15	18:45	1 days 08:35:00	2	2023-08-17	08:20	12	31	1976.92
11616	Swiss	Swiss	BRU	GRU	2023-08-01	09:15					2023-08-15	18:45	1 days 08:35:00	2	2023-08-17	08:20	12	31	1976.92
11617	Swiss	Swiss	BRU	GRU	2023-08-01	09:45					2023-08-15	18:45	1 days 08:35:00	2	2023-08-17	08:20	12	31	2100.92
11618	Swiss	Swiss	BRU	GRU	2023-08-01	09:15					2023-08-15	18:45	1 days 08:35:00	2	2023-08-17	08:20	12	31	2126.92

Missing values? Incorrect values?
Convert to correct data types, etc.

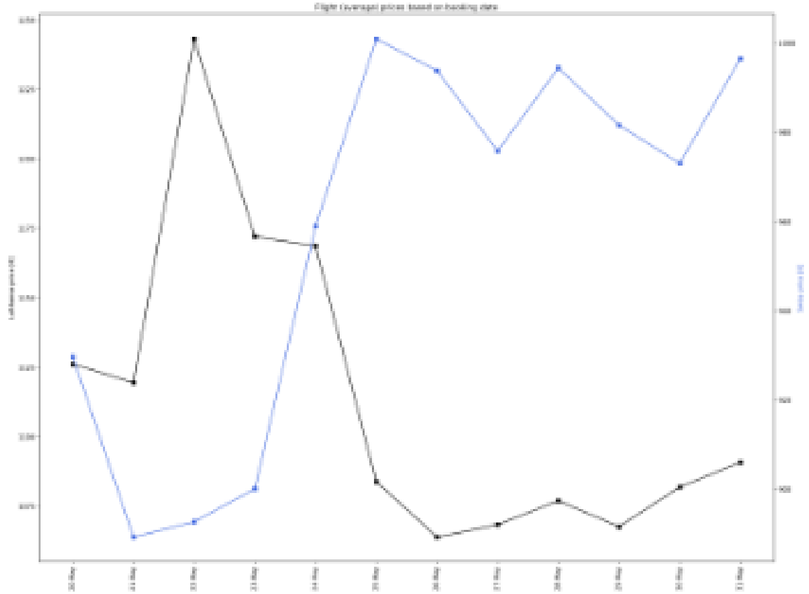
DATA VISUALIZATION

Bar charts (1)

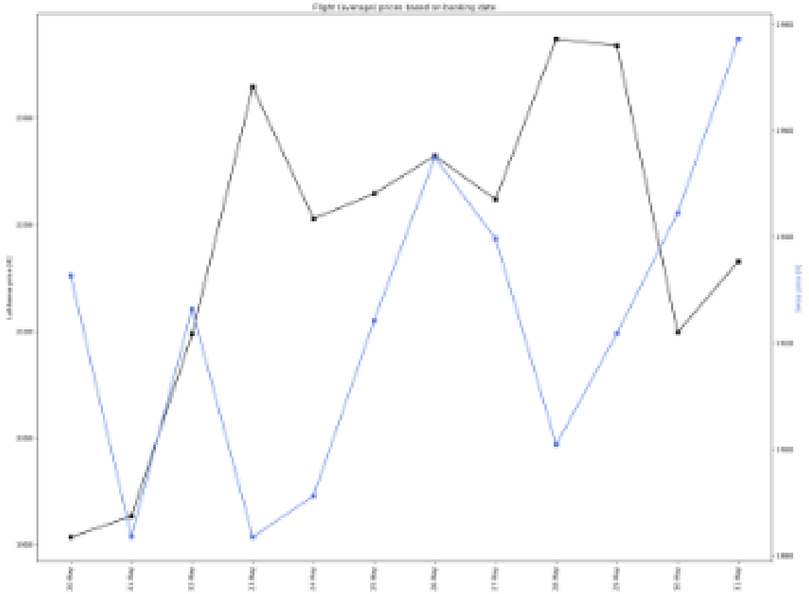
(a) Stacked bar chart for flights to New York (b) Stacked bar chart for flights to Sao Paulo

DATA VISUALIZATION

Line plots (2)



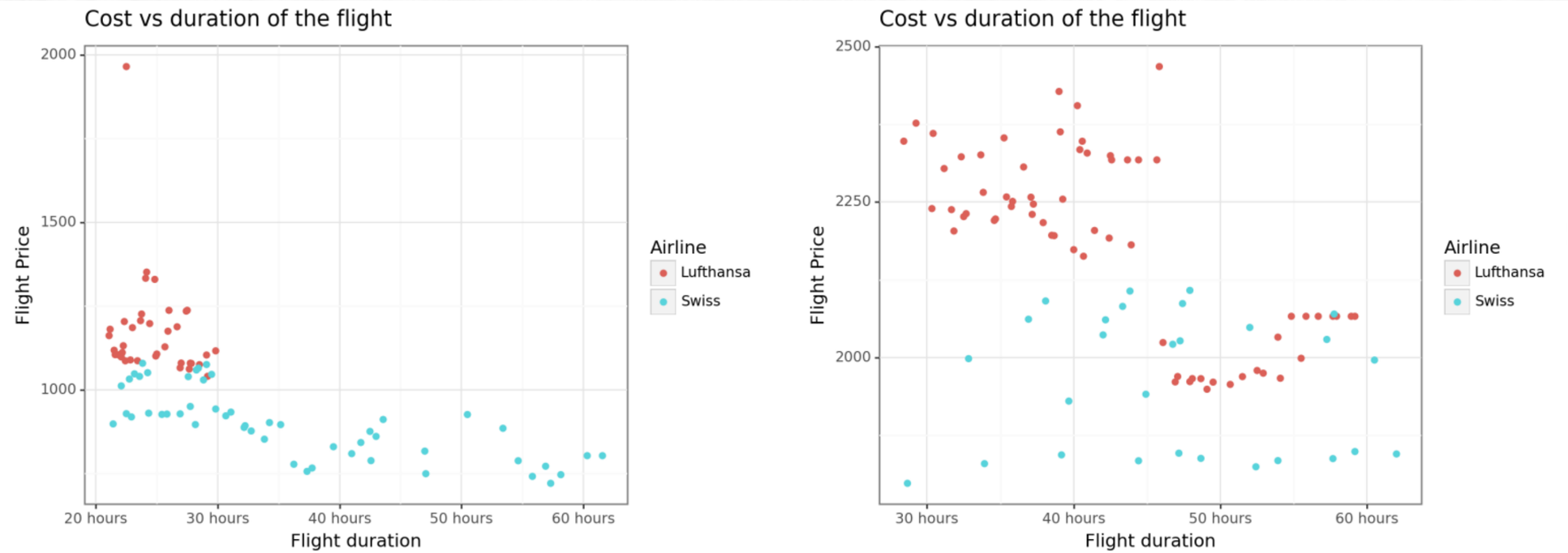
(a) Flight price trend to New York



(b) Flight price trend to São Paulo

DATA VISUALIZATION

Scatter plots (3)

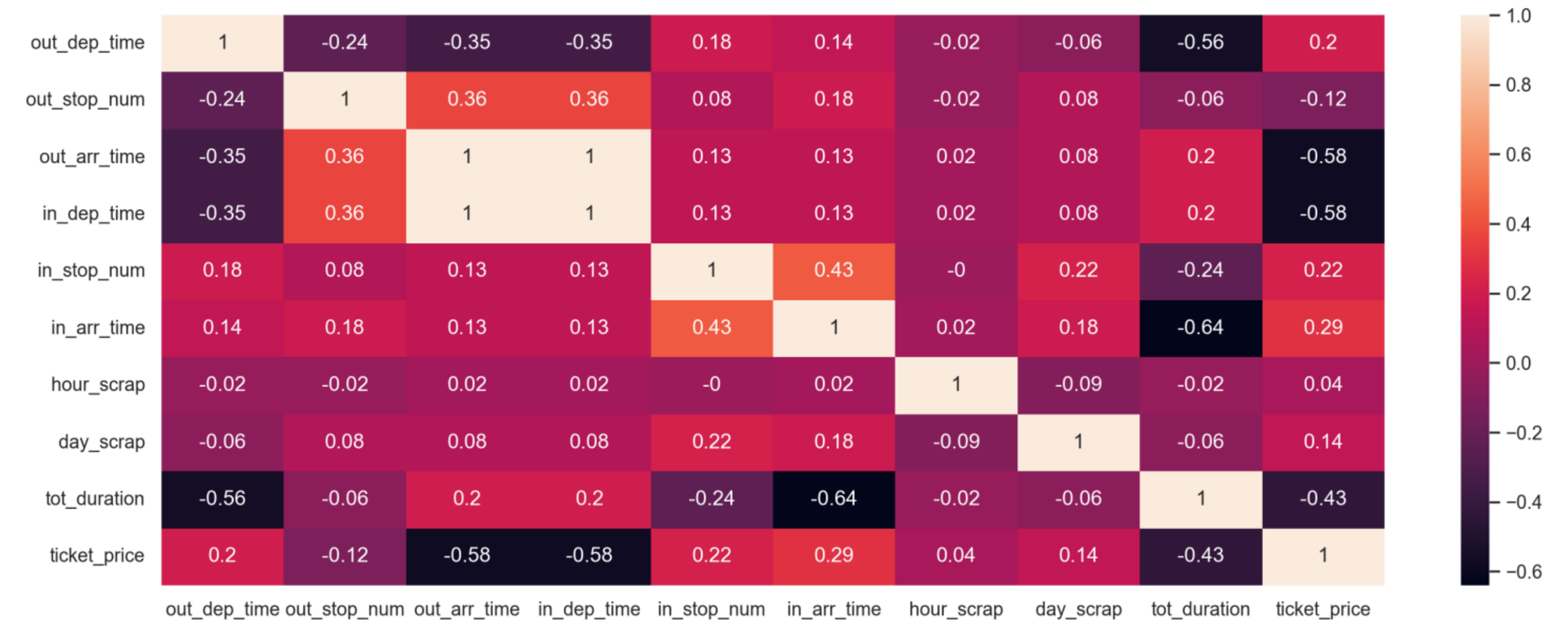


(a) Left figure

(b) Right figure

DATA VISUALIZATION

Heatmaps (4)



DATA MODELLING

Ticket price (y) = β0 + β1 × tot_duration_seconds + β2 × hour_scrap + β3 × airline_company_dummy+ β4 × destination_dummy + β5 × out_dep_time_dummy + β6 × in_arr_time_dummy + β7 × day_of_week_

⟶ Linear regression

	Values
Intercept	2097.4909712064773
Training RMSE	522.364425
Test RMSE	524.069444
Cross-validation RMSE	516.31323
Mean Absolute Error	171.879278
Mean Squared Error	274648.782087
Root Mean Squared Error	524.069444
R-squared	0.552468

Table 1: Different metrics to assess the model and their results

% of variance explained by the independent variables

	coef	p-values
const	2103.6217	1.613956e-219
tot_duration_seconds	-0.0011	5.714598e-08
hour_scrap	-2.8844	5.003298e-02
day_scrap	16.5914	2e-24
airline_company_dummy	-291	2e-128
destination_dummy	-1207	0
out_dep_time_dummy	-159	1e-33
in_arr_time_dummy	89	2e-09

Table 2: Coefficients of the model and their p-values

Statistically significant

DATA MODELLING

feature	VIF
const	142.799773
tot_duration_seconds	2.393894
hour_scrap	1.008921
day_scrap	1.054136
airline_company_dummy	1.175422
destination_dummy	2.411166
out_dep_time_dummy	1.460508
in_arr_time_dummy	1.249996

Table 3: VIF values for each feature

LINEAR REGRESSION

+ **relevant** independent variables (p-value)
+ reasonable level of predictive **accuracy** (R-squared)
+ **multicollinearity** is not a major concern (VIF)

Which will be the predicted **lowest ticket price** for each combination of destination, airline and day of the week ?

RESULTS

1) Primary factors that influence ticket prices:

- tot_duration_seconds, hour_scrap, airline_company_dummy, destination_dummy, out_dep_time_dummy, in_arr_time_dummy

2) The **time** at which flight ticket data is collected **has a significant impact** on ticket prices

3) The **optimal time** to book **depends on the specific combination** of destination, airline company, and day of week.

4) Our predictive model trained on historical data proved **reasonably accurate** in predicting future ticket prices.

LIMITATIONS & CONCLUSION

Scope of data collection could have been extended

DATA COLLECTION

The reliance on Booking.com, a comparable website

DATA SOURCE

Results from other models may provide better accuracy

MODELLING

Extraneous factors, airline pricing strategies, etc.

OTHER FACTORS

THANK YOU