programming for data science

KAYAK FLIGHT TICKETS ANALYSIS

GROUP 8

FEDERICO PASCHETTA
KARLA GONZALEZ ROMERO
YACINE MERIOUA
ARTH JAIN

0 24563 84926 54 2

PART 1 - WEB SCRAPING

WEBSITE SELECTED:

KAYAK

www.kayak.com

RESOURCES CHOSEN:

- Top 10 Airports in Europe per passengers traffic
https://aeroaffaires.com/europes-20-biggest-airports/
- Flight tickets between those airports from Kayak
- One day per month (19th of each month), for a year

PART 1 - WEB SCRAPING

LHR: LONDON
CDG: PARIS
AMS: AMSTERDAM
FRA: FRANKFURT
MAD: MADRID
BCN: BARCELONA
IST: ISTANBUL
MUC: MUNICH
FCO: ROME
DUB: DUBLIN

airports.txt

2024-04-19
2024-05-19
2024-06-19
2024-07-19
2024-08-19
2024-09-19
2024-10-19
2024-11-19
2024-12-19
2025-01-19
2025-02-19
2025-03-19

dates.txt

https://www.kayak.com/flights/MAD-DUB/2024-04-19?sort=bestflight_a

PART 1 - WEB SCRAPING

GENERATED FILES

- 1 .csv file for each airport (10 files .csv in total)
- More than 3000 data about flights departing from each airport

STRUCTURE FOR EVERY ROW

e.g. from Madrid.csv file

Carrier	DepTime	ArrTime	Price	Duration	Day	DepAirport	ArrAirport	DayAfter*
Iberia	20:50	22:10	$249	2h 20m	2024-04-19	Madrid	London	False
easyJet	20:45	22:50	$45	2h 05m	2024-06-19	Madrid	Paris	False

*DayAfter attribute is True when the flight arrives to the arrival airport the day after the departure, False otherwise

PART 2 - PREPROCESSING

DATA CLEANING AND TRANSFORMATION

- Convert price text into numerical values
- Combine repeated names into a single entity (e.g., 'Wizz Air', 'Wizz Air UK', 'Wizz Air Brussels' -> 'Wizz Air')
- Eliminate non-airline services from the dataset (e.g., 'ALSA', 'RENFE')

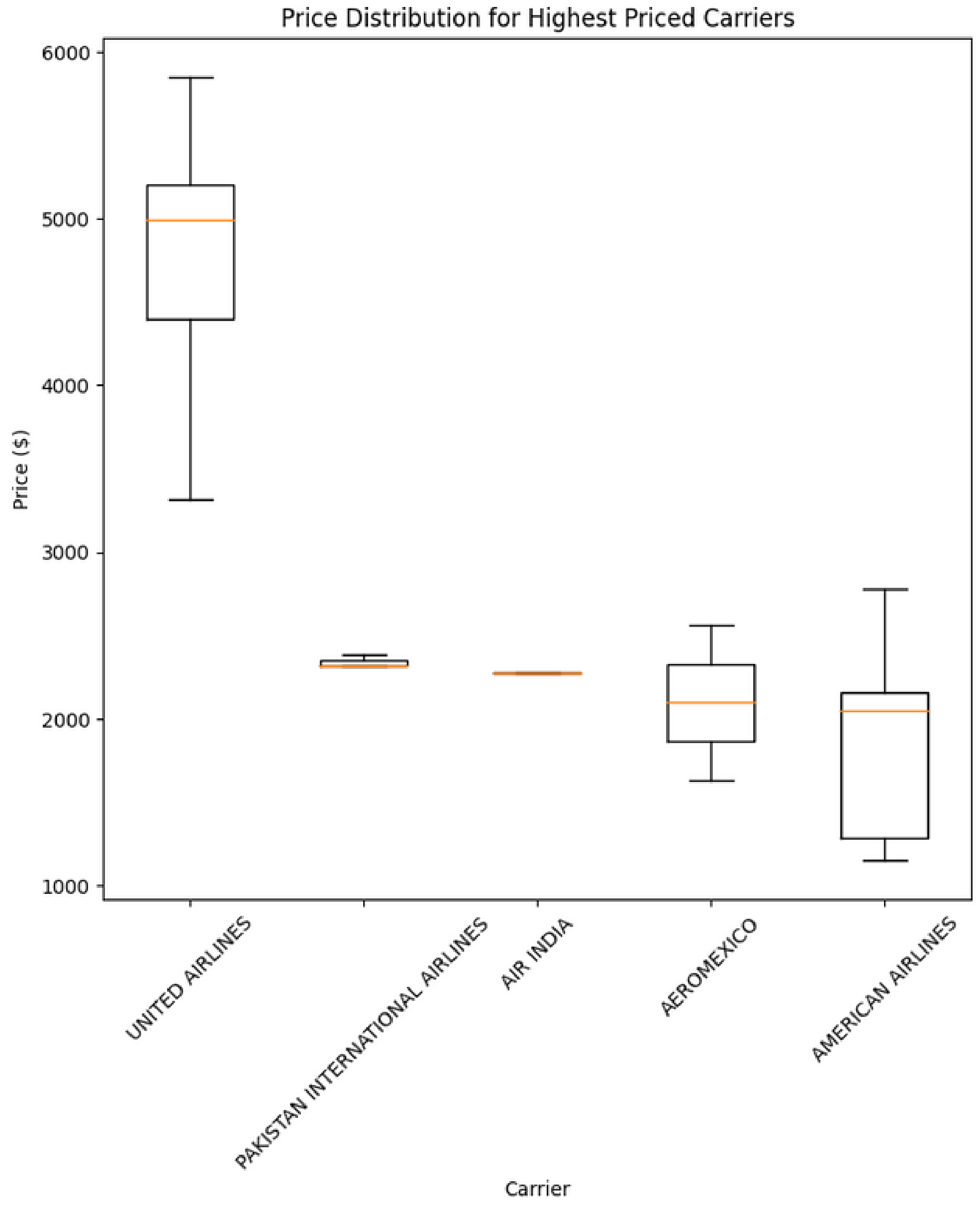
Code Snippet:

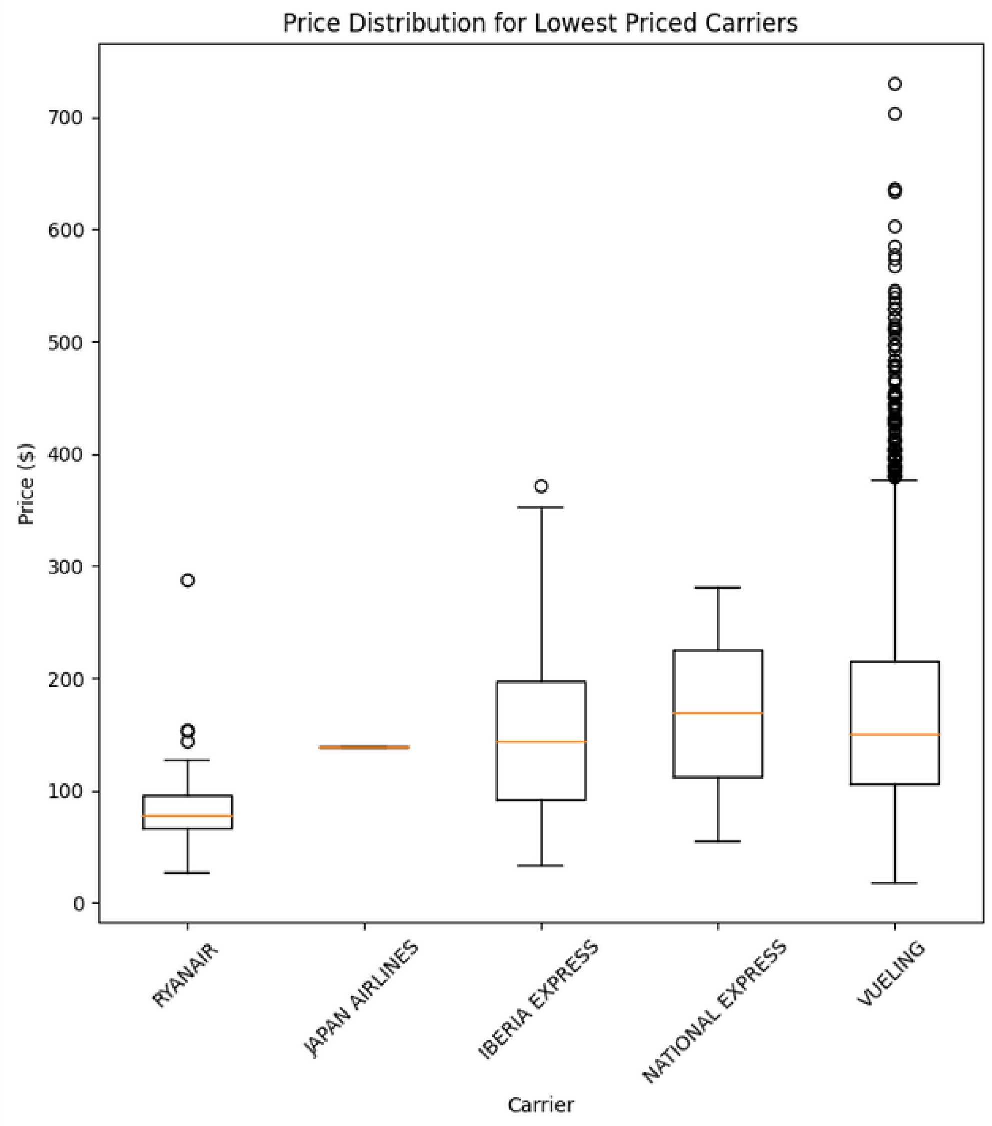
```
import re

def unify_carrier_name(carrier_name):
# Use regular expressions to match common patterns in carrier names
if re.match(r'^WIZZ AIR', carrier_name):
return 'WIZZ AIR'
elif re.match(r'^EASYJET', carrier_name):
return 'EASYJET'
elif re.match(r'^TUI', carrier_name):
return 'TUI AIRWAYS'
# Add more patterns as needed
else:
return carrier_name
```

PART 3 - ANALYSIS

ANALYZING TICKET PRICES AND TRENDS

- Price Distribution by Carrier





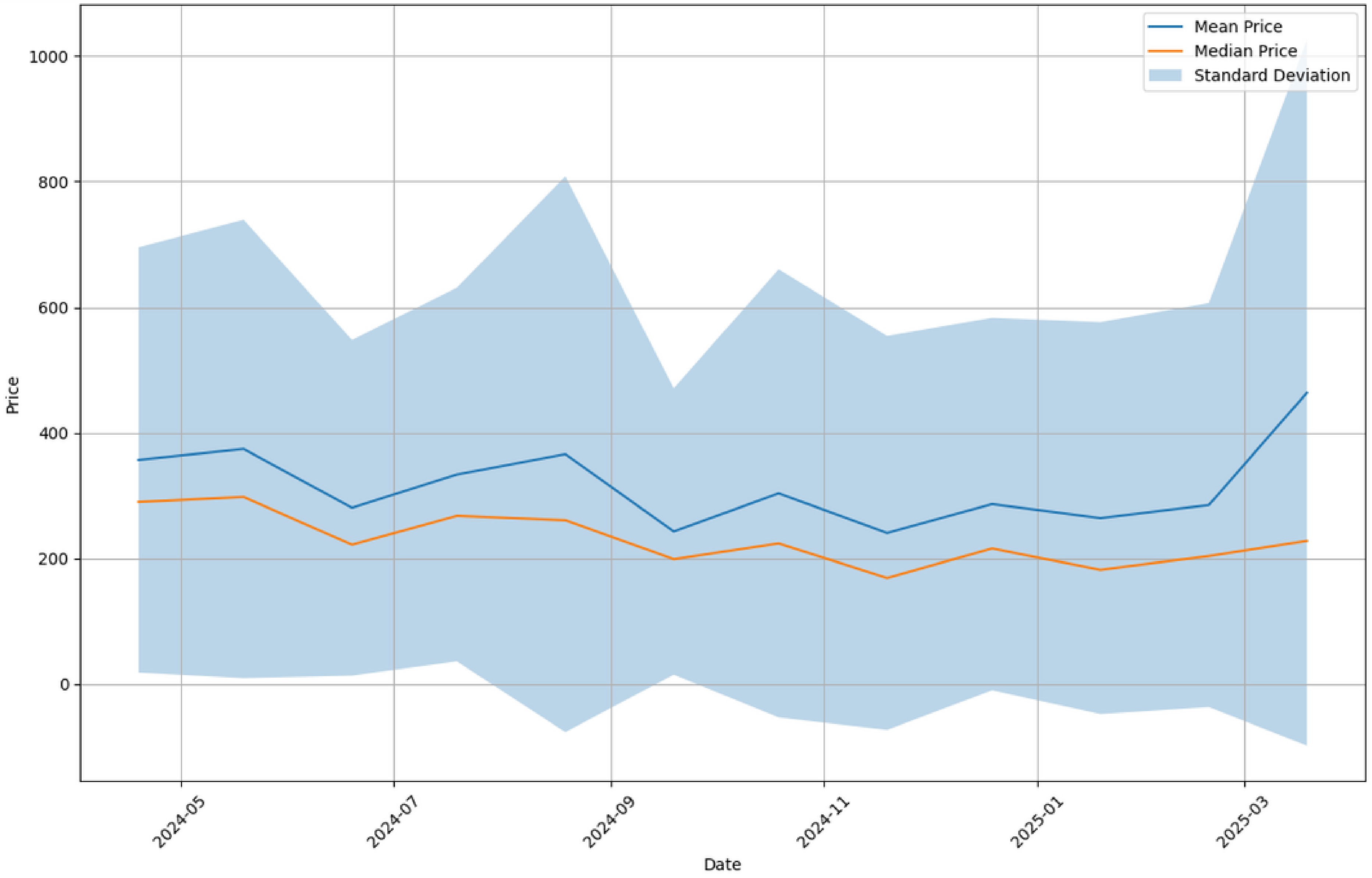
Most Expensive: United Airlines (4749)

Least Expensive: RyanAir (87)

PART 3 - ANALYSIS

ANALYZING TICKET PRICES AND TRENDS

- Price Trends Over Time*



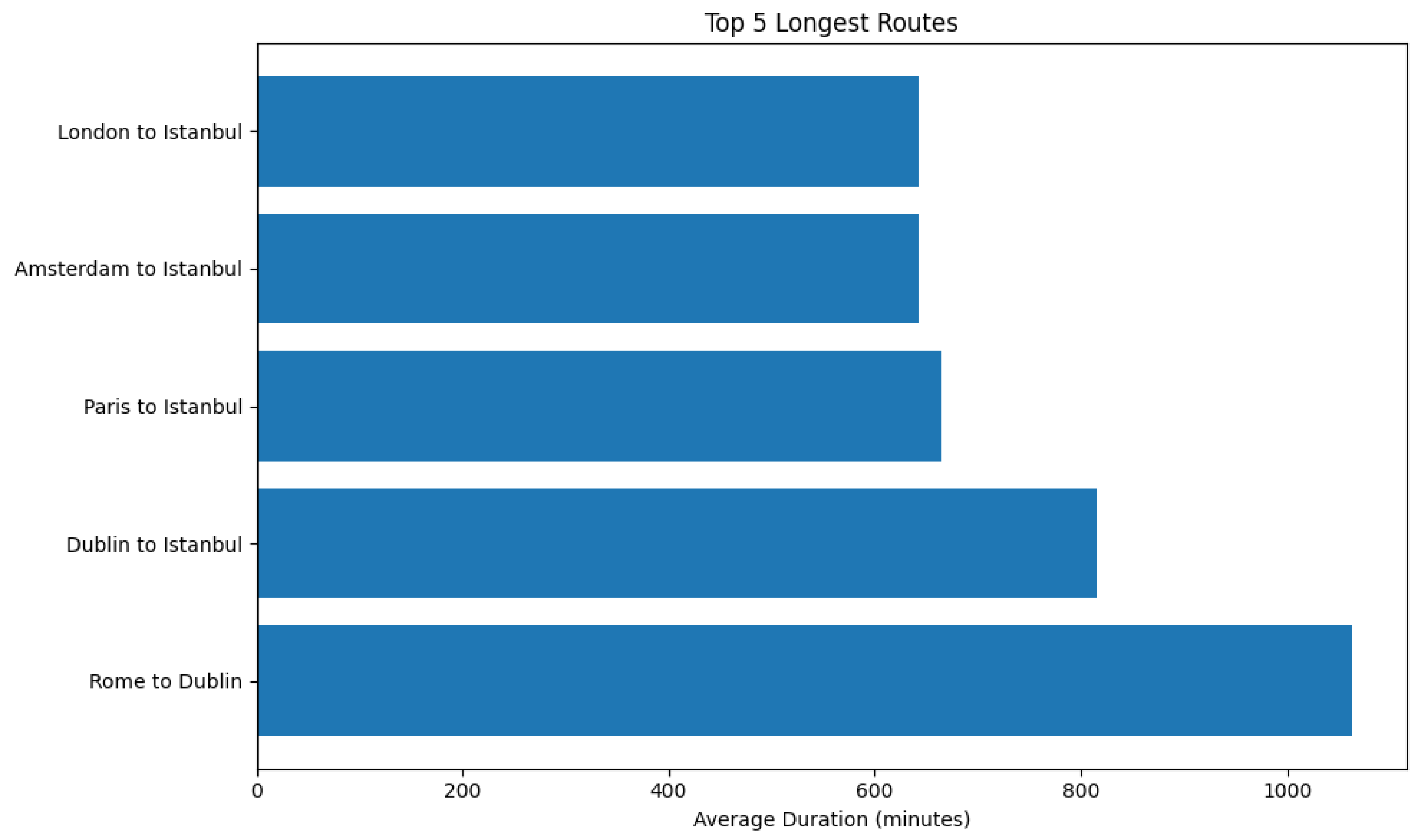
*Projected for the next year

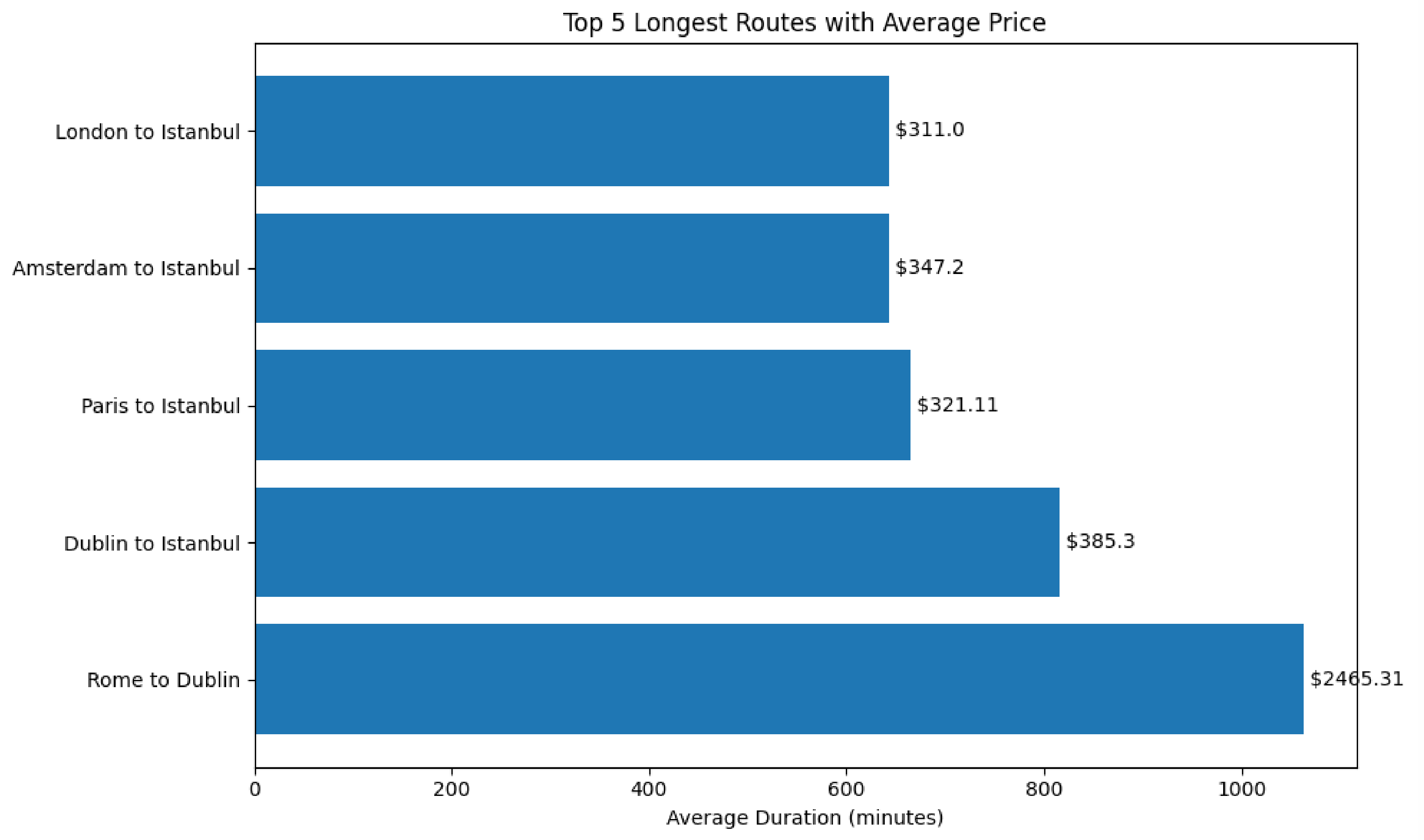
Most Expensive:
Spring (03-05)

Least Expensive:
Winter (10-02)

PART 3 - ANALYSIS

LONGEST ROUTES AND AVERAGE PRICES





- The average prices for flights are directly proportional to the average flight time

PART 3 - ANALYSIS

KEY FINDINGS AND INSIGHTS SUMMARY

- The price of a flight depends on various factors such as the distance, time of booking and the airline you book through
- United Airlines and RyanAir proved to be the most expensive and cheapest airlines respectively
- Spring was the most expensive season to fly with Winter being the cheapest on avarage
- The flight price is likely to be higher for longer distances

RECOMMENDED STEPS

- Check websites in advance to avoid surcharge pricing programming for data science

THANK YOU

WISH YOU CHEAP FLIGHTS

0 24563 84926 54 2