

# Predicting Flight Ticket Prices



Web-Scraping / Regression  
Project Presentation

SRIDHAR N

# Goal



# Goal



- ❧ Forecast flight prices
  - ❧ Selecting optimum time for travel
  - ❧ Selecting the cheapest flight to the desired destination





# Scraping



# Scraping



KAYAK

Business Now Sign in SAR

One-way ▾ 1 Adult ▾ Economy ▾ 0 Bags ▾

Riyadh (RUH) × + ⇄ Moscow (SVO) × + Tue 11/30 < > 🔍

**Price rise expected**  
**Buy now** >  
44 SAR rise is expected in the next 28 days  
Track prices ☐

45 of 115 flights

**Stops**  
☐ Nonstop  
☒ 1 stop 1,920 SAR  
☒ 2+ stops 1,248 SAR

**Fare Assistant** ⓘ  
Checked bag - 0 +  
Payment method  
Visa Debit ▾

**Flexible options** ^

Cheapest	Best Ⓢ	Quickest	Other sort
1,248 SAR • 17h 30m	1,920 SAR • 17h 05m	3,237 SAR • 11h 05m	
<b>No change fees</b>			
1 1			
5:10 pm – 7:50 am <sup>+1</sup> RUH King Khale... - SVO Sheremety... Qatar Airways	1 stop	14h 40m	2,224 SAR Economy Classic Select
Sponsored Result			
<b>Best</b>			
0 1			
4:40 am – 9:45 pm RUH King Khale... - SVO Sheremety... Egypt Air, Aeroflot	1 stop ▲ CAI	17h 05m	1,920 SAR Economy Promo Select
<b>Cheapest</b>			
0 1			
4:10 am – 9:40 pm RUH King Khale... - SVO Sheremety... Multiple Airlines	3 stops ▲ DOH, SAW, ...	17h 30m	1,248 SAR Economy Classic Select

Source & Destination

Date (Feb to April, 2022)

Price

Duration

Total Stops

Airline

# Scraping Tool



Please enter -1 when done.

-----

From which city?

RUH

Where to?

NYC

Start Date, Please use YYYY-MM-DD format only 2022-02-01

End Date, Please use YYYY-MM-DD format only

2022-04-30

Please enter -1 when done.

-----

From which city?

RUH

Where to?

NYC

-----

From which city?

SVO

Where to?

PAR

-----

From which city?

-1

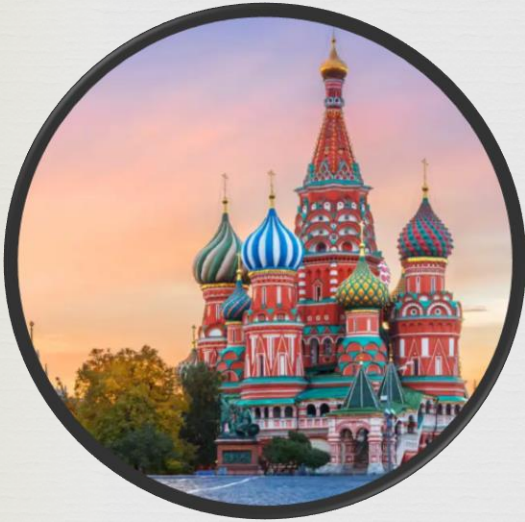
Routes:

RUH => NYC

SVO => PAR



# Scraped Routes



SVO



NYC

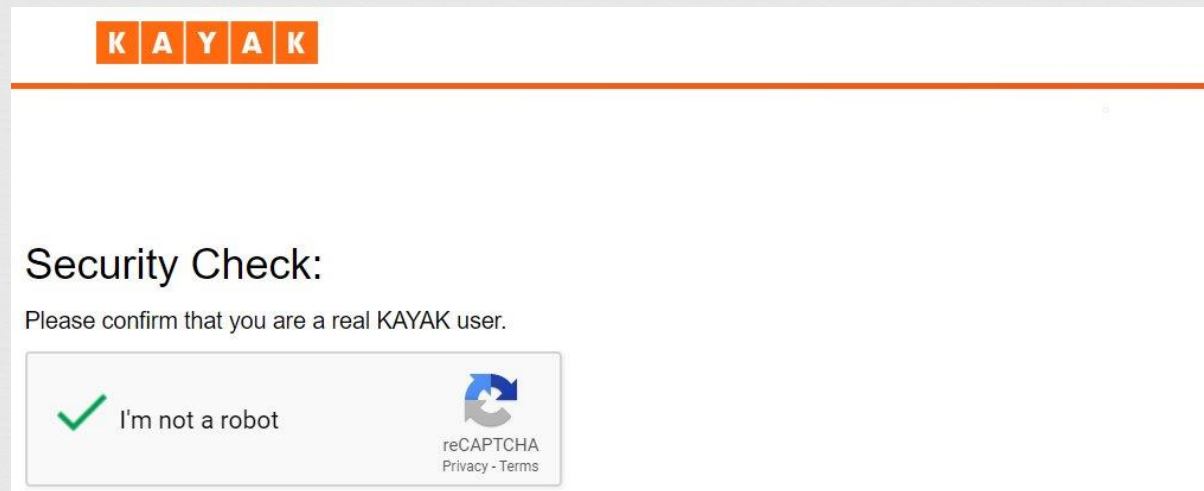


PAR



RUH

# Problems



- ❧ Closing and opening the driver every 10 days (pages).
- ❧ Alert the user to go solve the captcha and resume scraping.



# Scraped Dataset



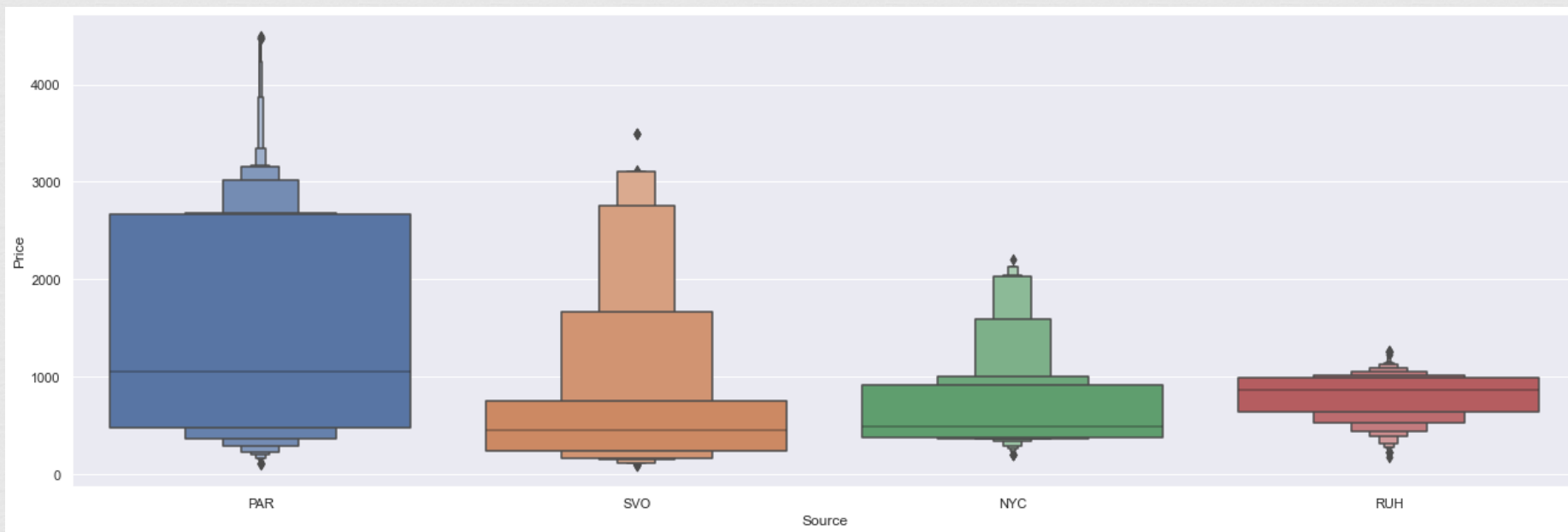
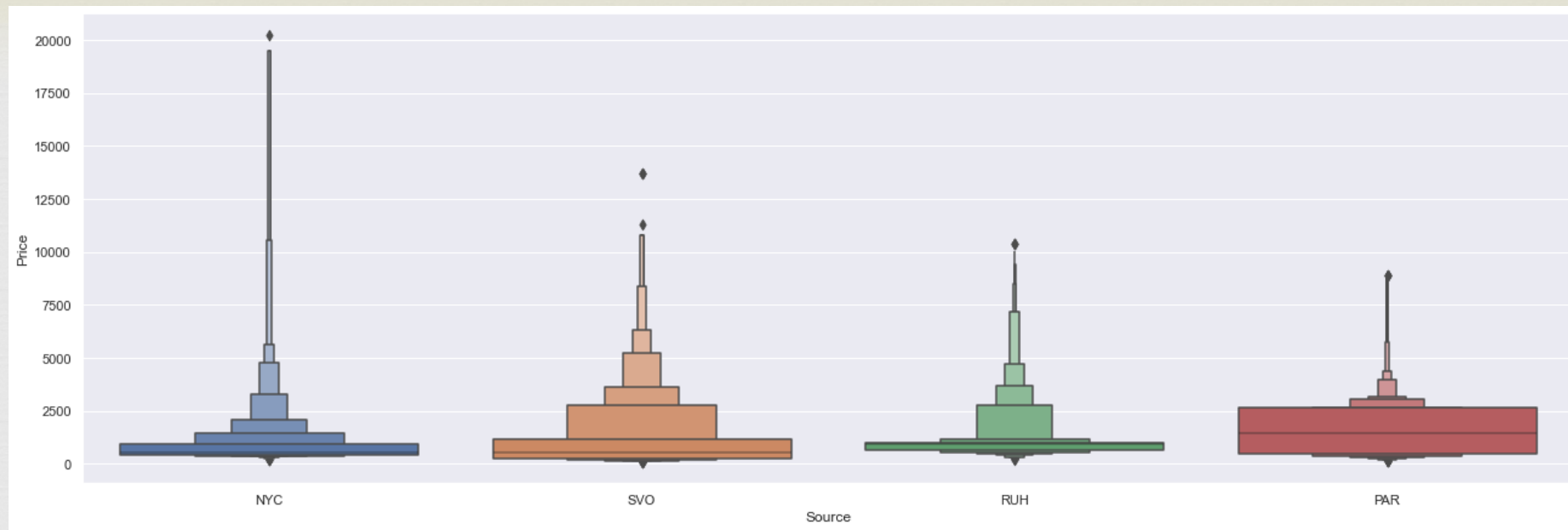
	Airline	Source	Destination	Duration	Total stops	Price	Date
0	TAP AIR PORTUGAL	PAR	NYC	12h 50m	1 stop	1,423 SAR	2022-02-01
1	KLM, Aeroflot	NYC	SVO	15h 45m	1 stop	2,851 SAR	2022-02-01
2	KLM, Aeroflot	NYC	SVO	12h 50m	1 stop	2,851 SAR	2022-02-01
3	Air France	NYC	SVO	19h 05m	1 stop	2,563 SAR	2022-02-01
4	Air France	NYC	SVO	18h 05m	1 stop	2,563 SAR	2022-02-01
...	...	...	...	...	...	...	...
55358	Air France	PAR	RUH	14h 25m	2 stops	15,260 SAR	2022-04-30
55359	Air France	PAR	RUH	15h 25m	2 stops	15,260 SAR	2022-04-30
55360	SAUDIA	PAR	RUH	19h 20m	2 stops	2,364 SAR	2022-04-30
55361	Finnair, American Airlines	SVO	NYC	18h 46m	2 stops	11,636 SAR	2022-04-30
55362	Delta	NYC	PAR	13h 00m	1 stop	1,793 SAR	2022-04-30

55,363 Rows

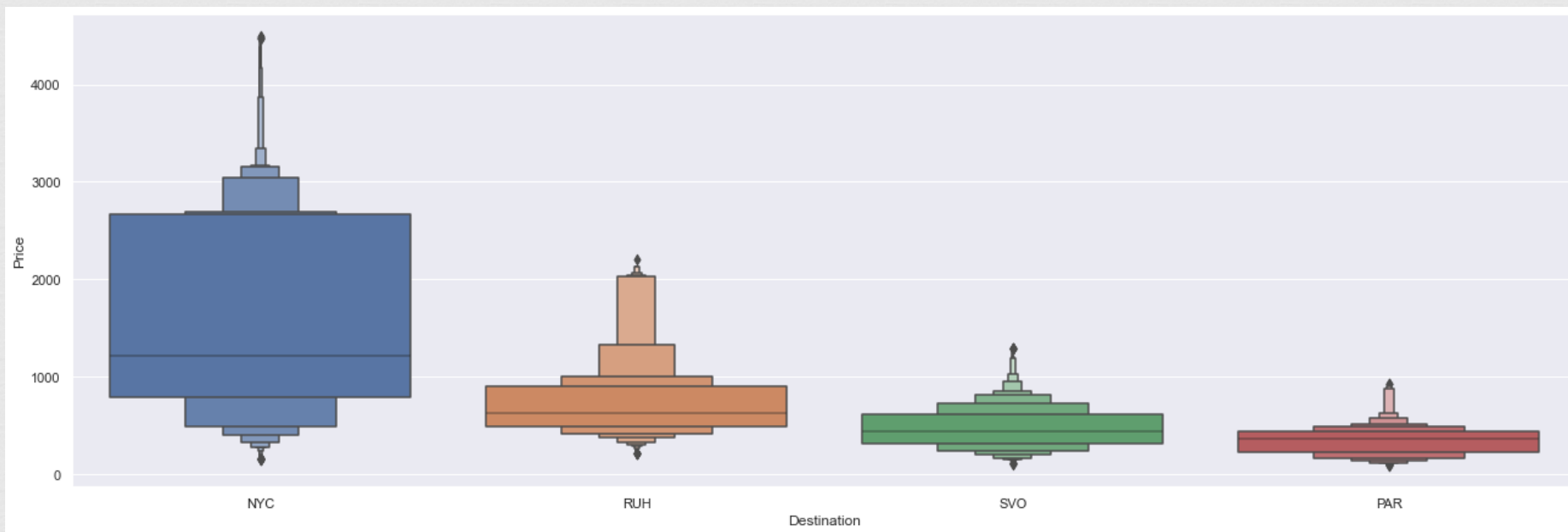
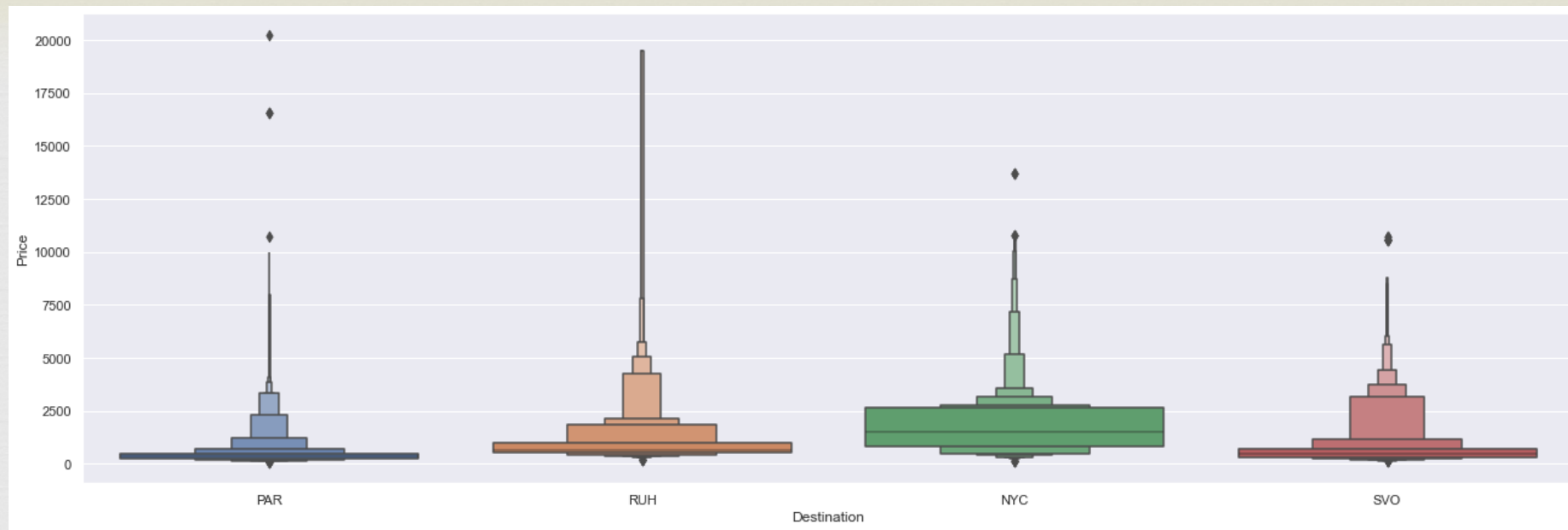
7 Columns

# EDA



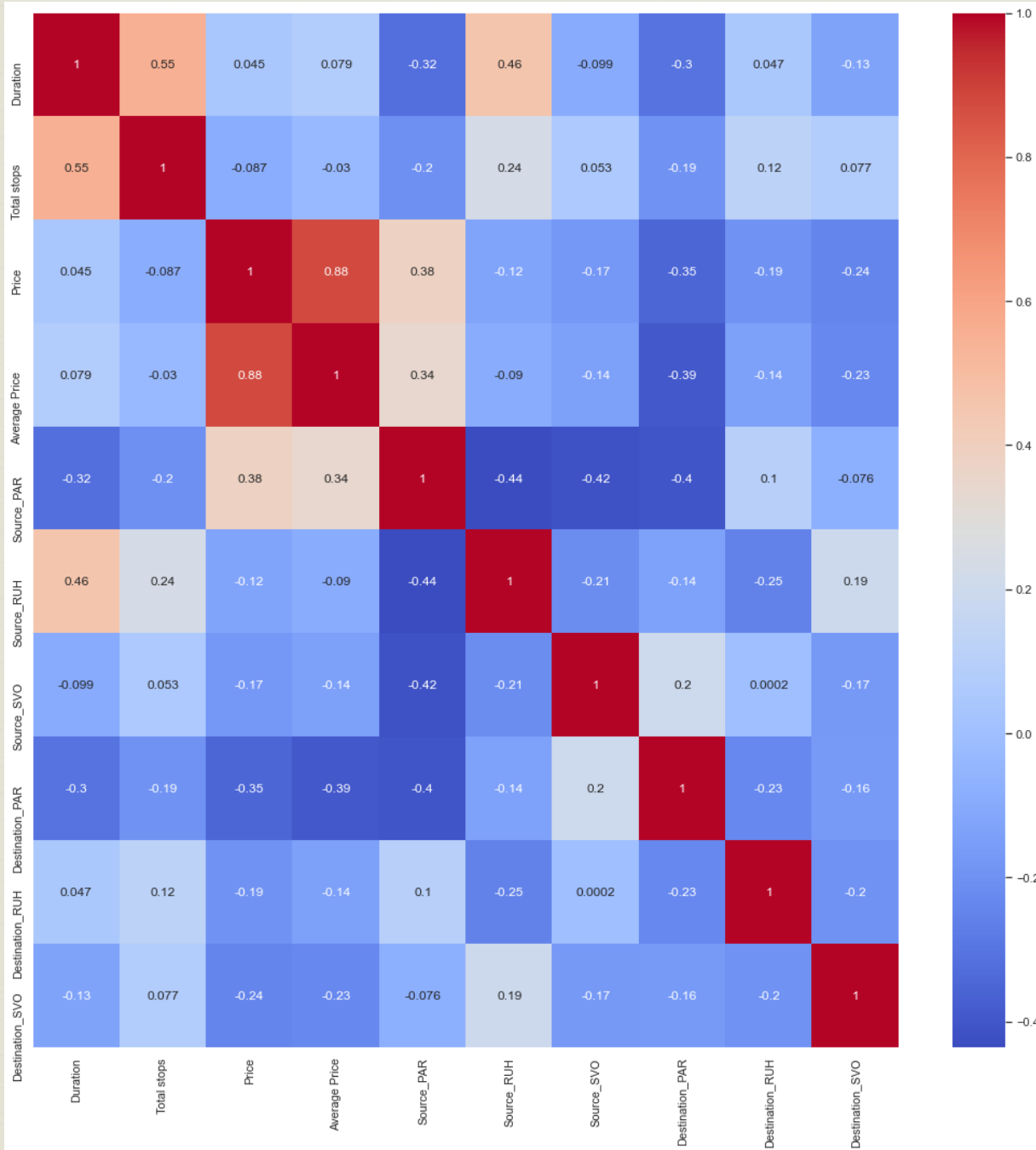






# Features

- ✧ Source (4)
- ✧ Destination (4)
- ✧ Total Stops
- ✧ Average Price per Airline
- ✧ Duration
- ✧ Price



# Results and Experiments

---





# Results



Model / Metric	Train (60%)	Val (20%)	MAE	MSE	RMSE
Linear Regression	0.804	0.789	225.09	152995.68	391.15
Polynomial (5)	0.888	0.871	149.72	93366.04	305.56
Lasso	0.803	0.789	223.49	153097.06	391.28
Ridge	0.804	0.789	225.08	152995.61	391.15
Elastic Net	0.790	0.775	224.41	163034.63	403.78
Random Forest	0.965	0.945	61.72	40035.32	200.09

# Other Experiments



- ❧ Feature Importance
- ❧ Feature Scaling
- ❧ Hyperparameter Tuning
  - ❧ Randomized Search CV

# Result



- Retraining the random forest model on train + val and scoring on test.
- Before and after hyperparameter tuning using random search CV:

Metric	MAE	MSE	RMSE
Before	61.87	40409.87	201.02
After	64.19	41409.93	203.49



# Result



	Predicted Price (USD)	Actual Price (USD)
0	2676.530000	2676.53
1	182.571321	248.53
2	254.362293	265.07
3	447.386056	490.93
4	643.470000	643.47
...	...	...
10015	2682.130000	2682.13
10016	1605.962800	1608.00
10017	233.286587	316.00
10018	1533.600000	1533.60
10019	657.900585	686.93

≈ \$61.87

# Recommendations



# Recommendations



- ❧ Consider the date as a feature
- ❧ Consider autocorrelation
- ❧ More sources / destinations
- ❧ Increase the period (dates)



# App Demo

