

# **IDENTIFY THE KEY FACTORS AND PREDICT FLIGHT PRICES**

Fuyi Pao, Jinisha Kande, Kevin Yu, Shu Wang, Yuhan Wang



# AGENDA

- INTRO
- DATA OVERVIEW
- EDA
- REGRESSION
- CONCLUSION



# INTRO

## MOTIVATION

As international students who travel a lot between our home country and US, it is crucial for us to find the best value for flight tickets during the summer vacation period

## GOAL

Identify key factors of flight prices for 04/16/2022 - 10/05/2022 to generate price predictions for customers.

## METHOD

Use PySpark for EDA and Regression Model on a large-scale dataset of flight prices and related variables



# DATA OVERVIEW

The dataset contains purchasable flight tickets available on **Expedia** website during the period **between April 16, 2022 and October 5, 2022**. It contains **82.1 M rows and 28 columns** of information that includes details on flight dates, starting and destination airports, fares, travel duration, and other relevant information.

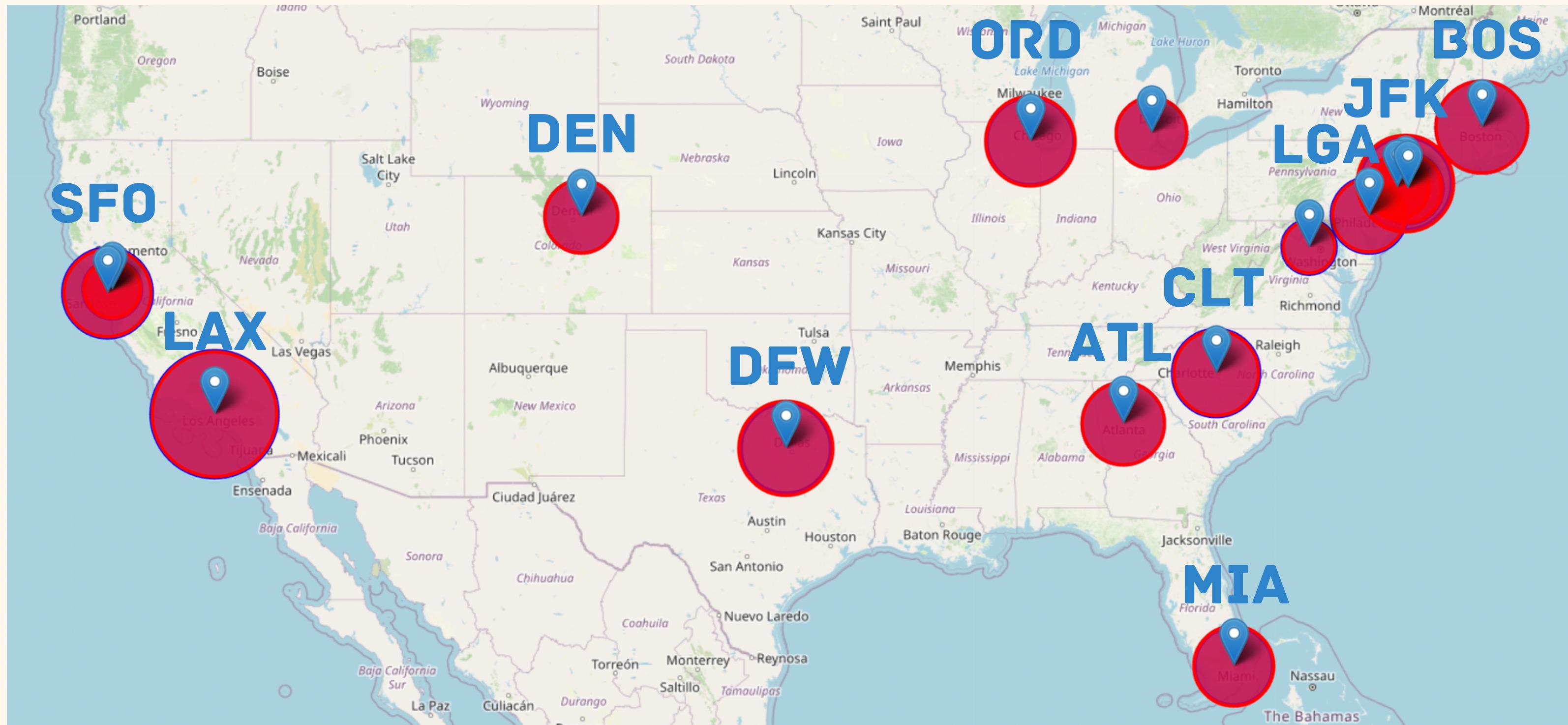
Column type	Attribution
<b>BASIC INFO</b>	<code>legId, searchDate, flightDate, startingAirport, destinationAirport</code>
<b>FARE INFO</b>	<code>fareBasisCode, baseFare, totalFare, and seatsRemaining fields</code>
<b>FLIGHT DETAILS</b>	<code>travelDuration, elapsedDays, isBasicEconomy, isRefundable, isNonStop, totalTravelDistance, segmentsDepartureTimeEpochSeconds, segements</code>

# PART 1 : EDA



# A. TRAVEL ROUTE

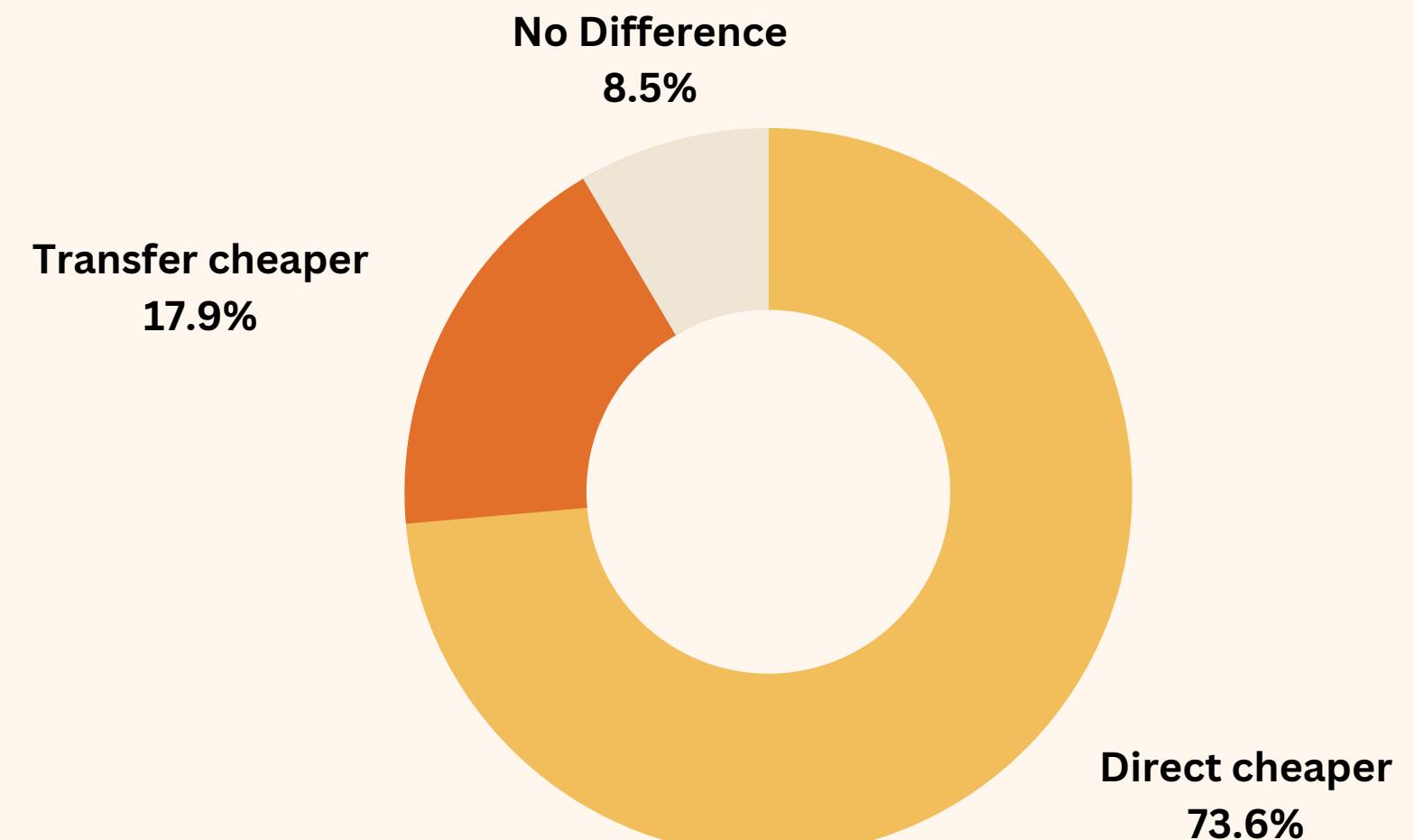
## WHAT ARE SOME POPULAR AIRPORTS



# POPULAR OR DIRECT = EXPENSIVE?

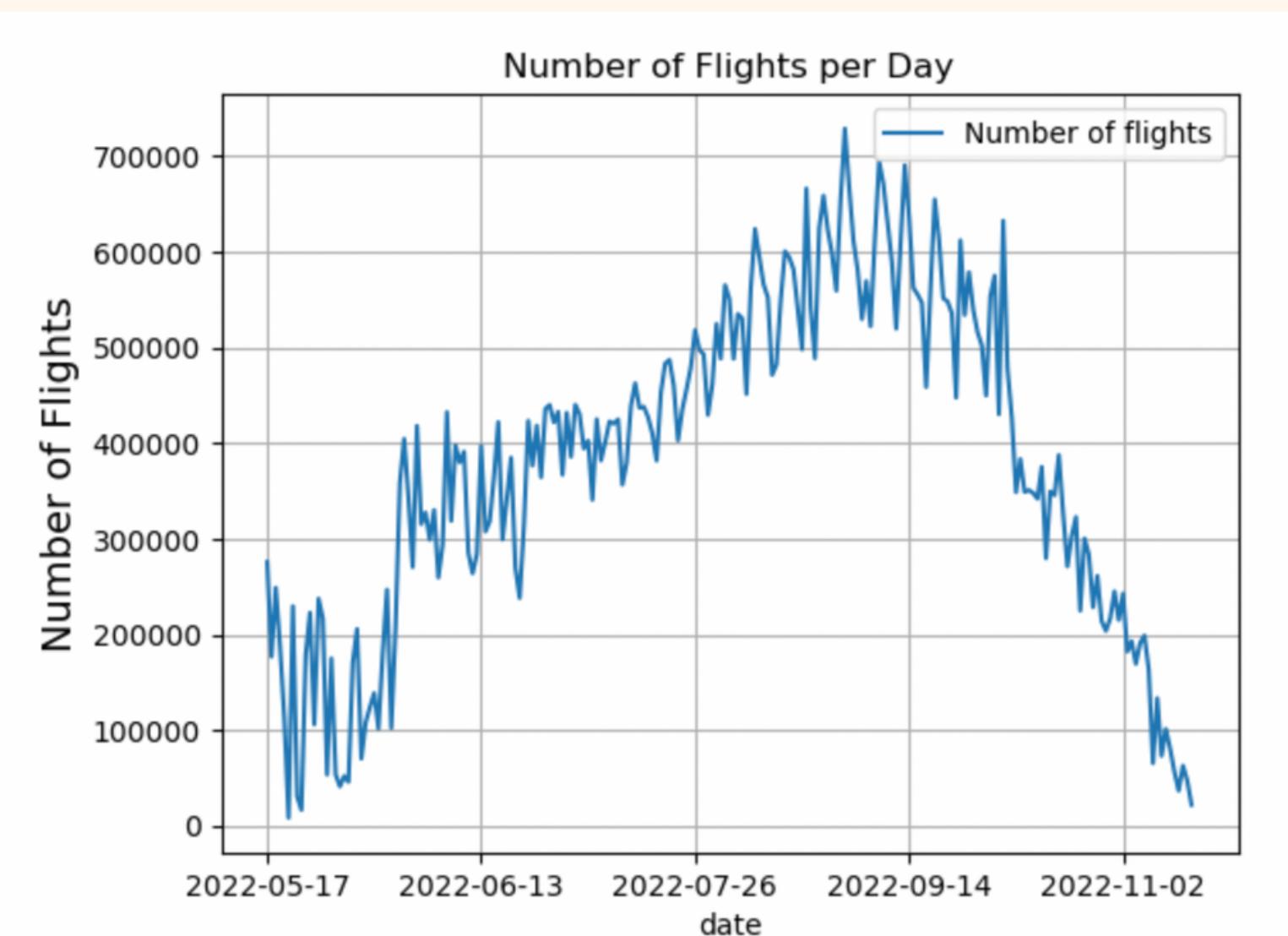
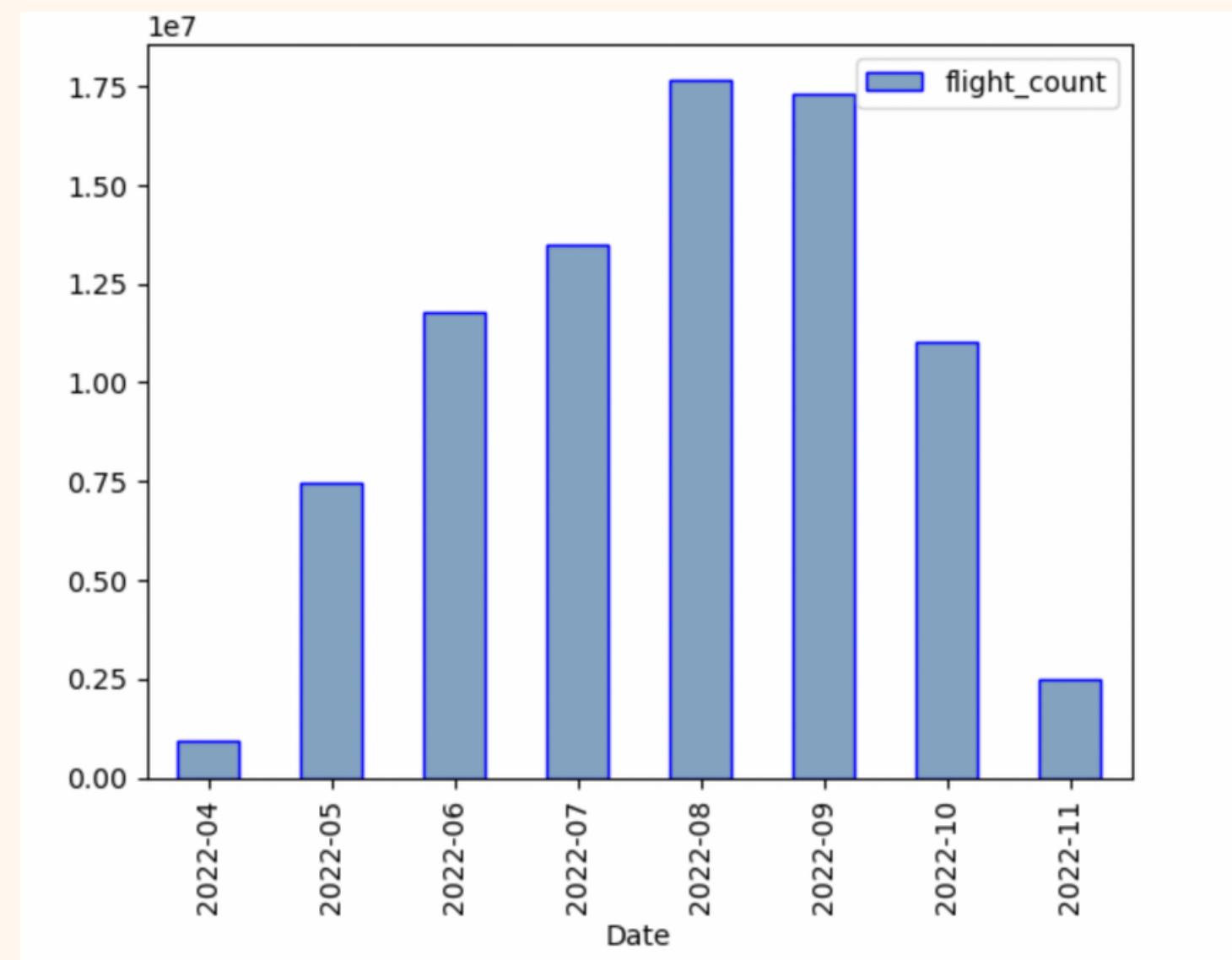
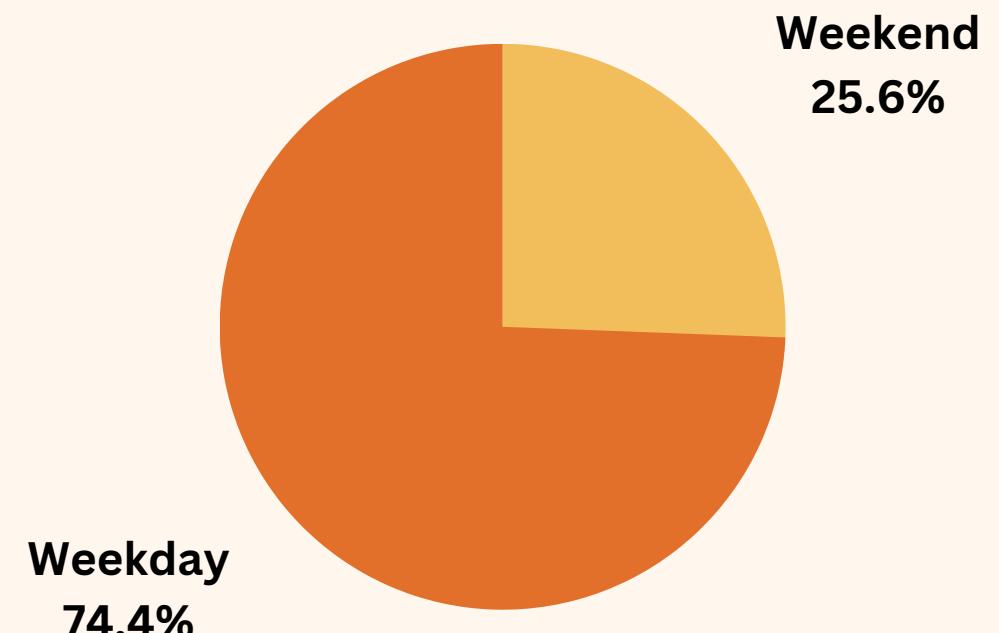


## TRANSFER VS DIRECT ON SAME ROUTE

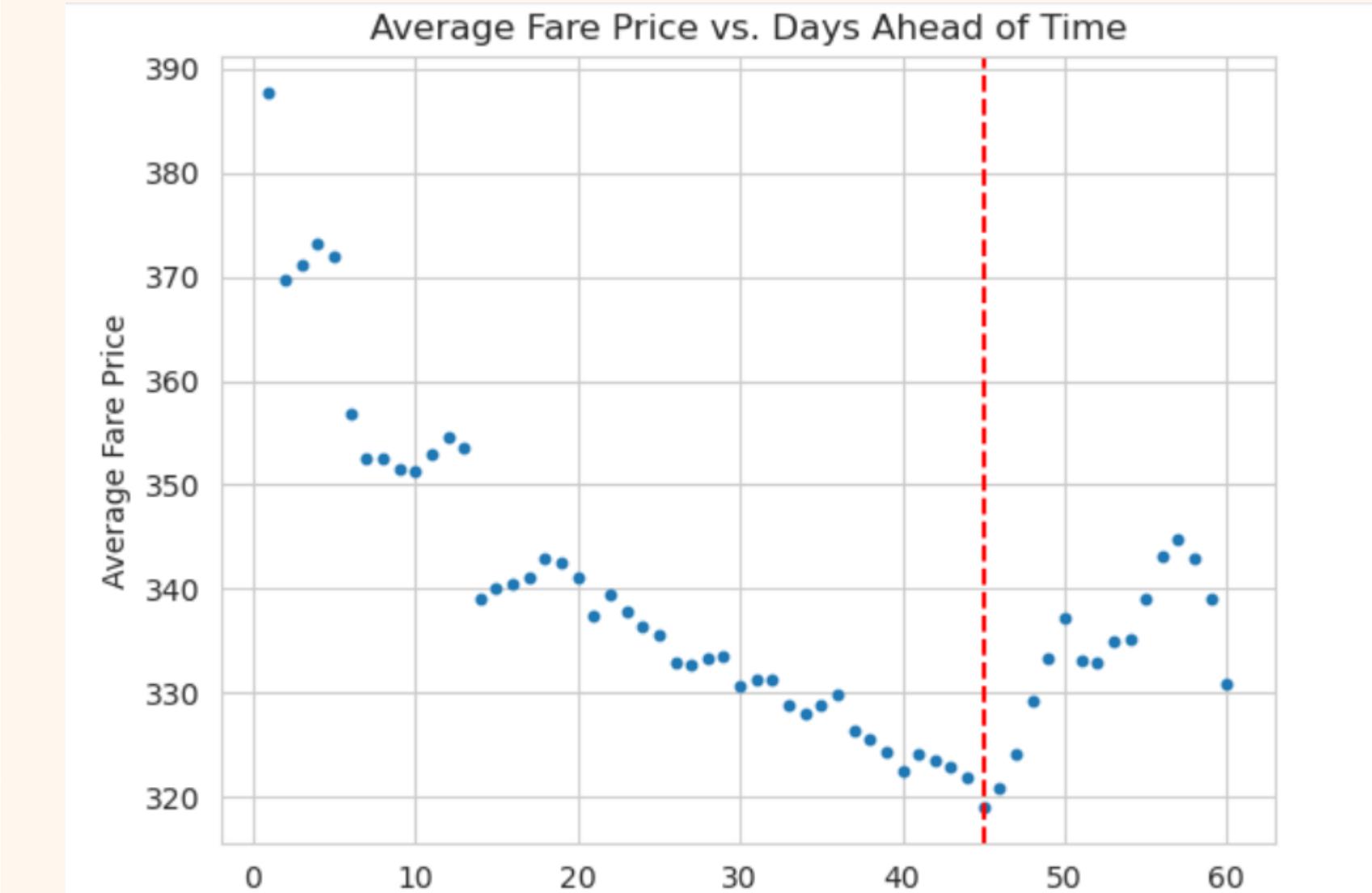
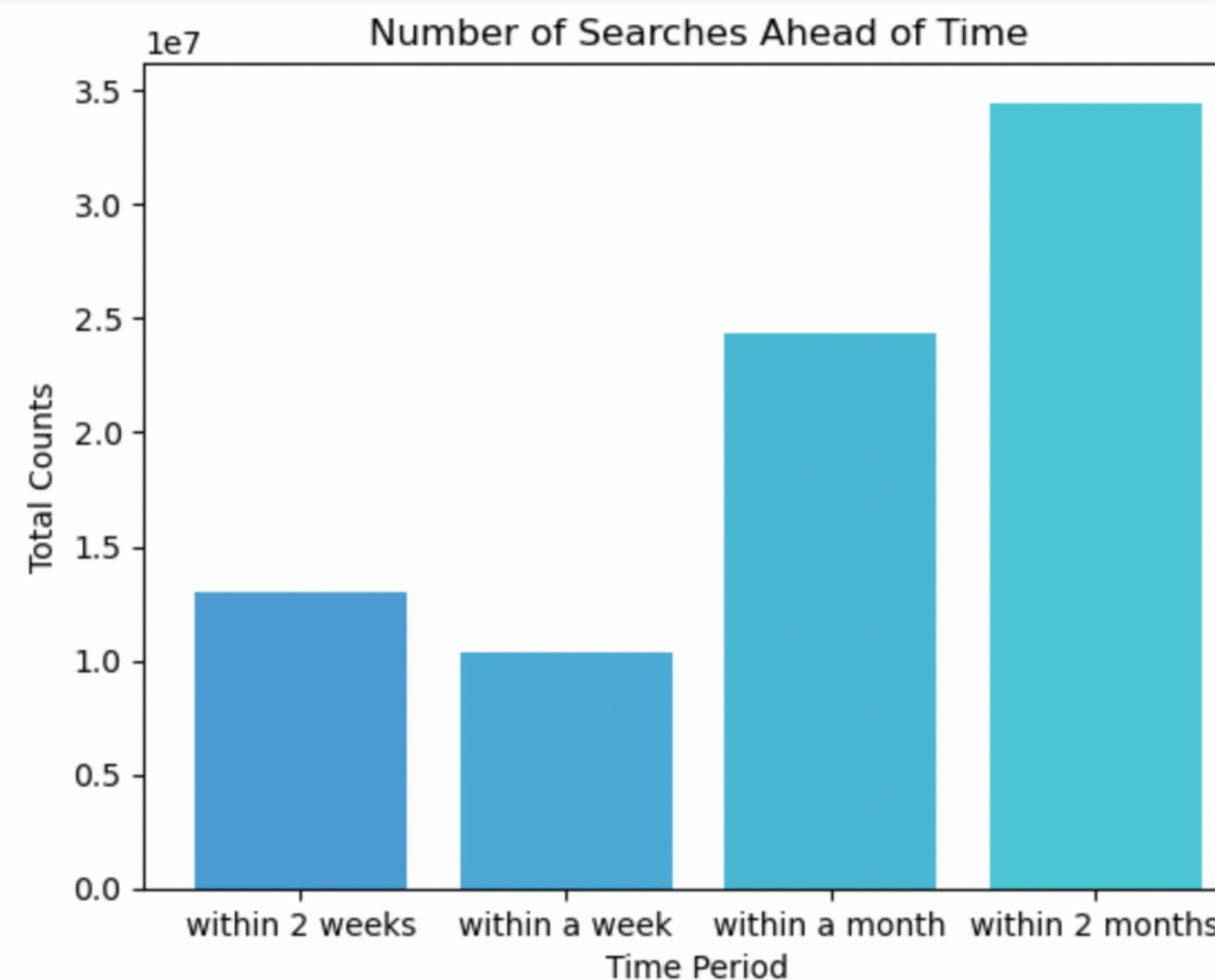


## B. TIME

### DAILY PURCHASABLE TICKETS

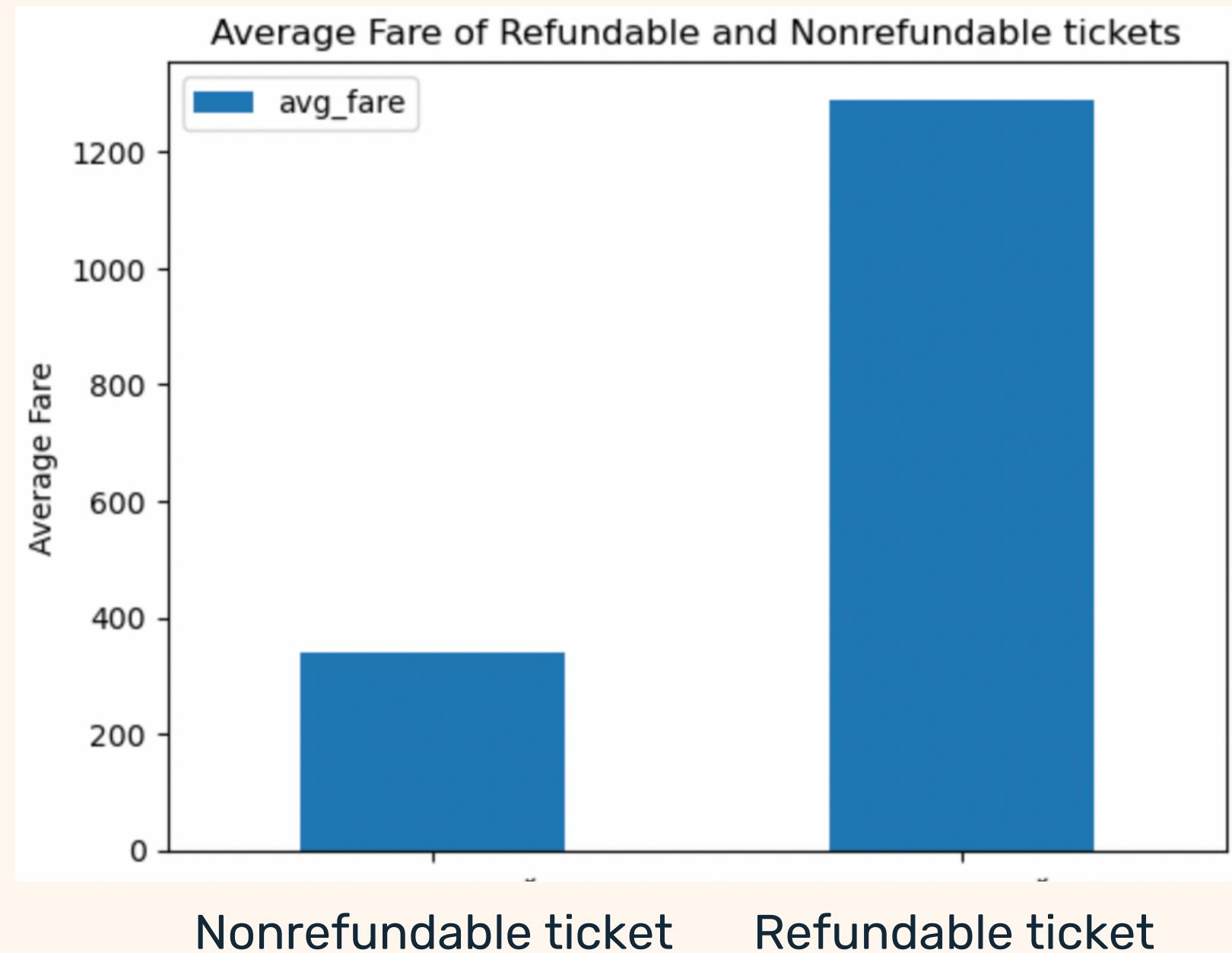


# SEARCH EARLY = CHEAP?



# C. SALES SITUATION

## NON REFUNDABLE TICKET = CHEAP?

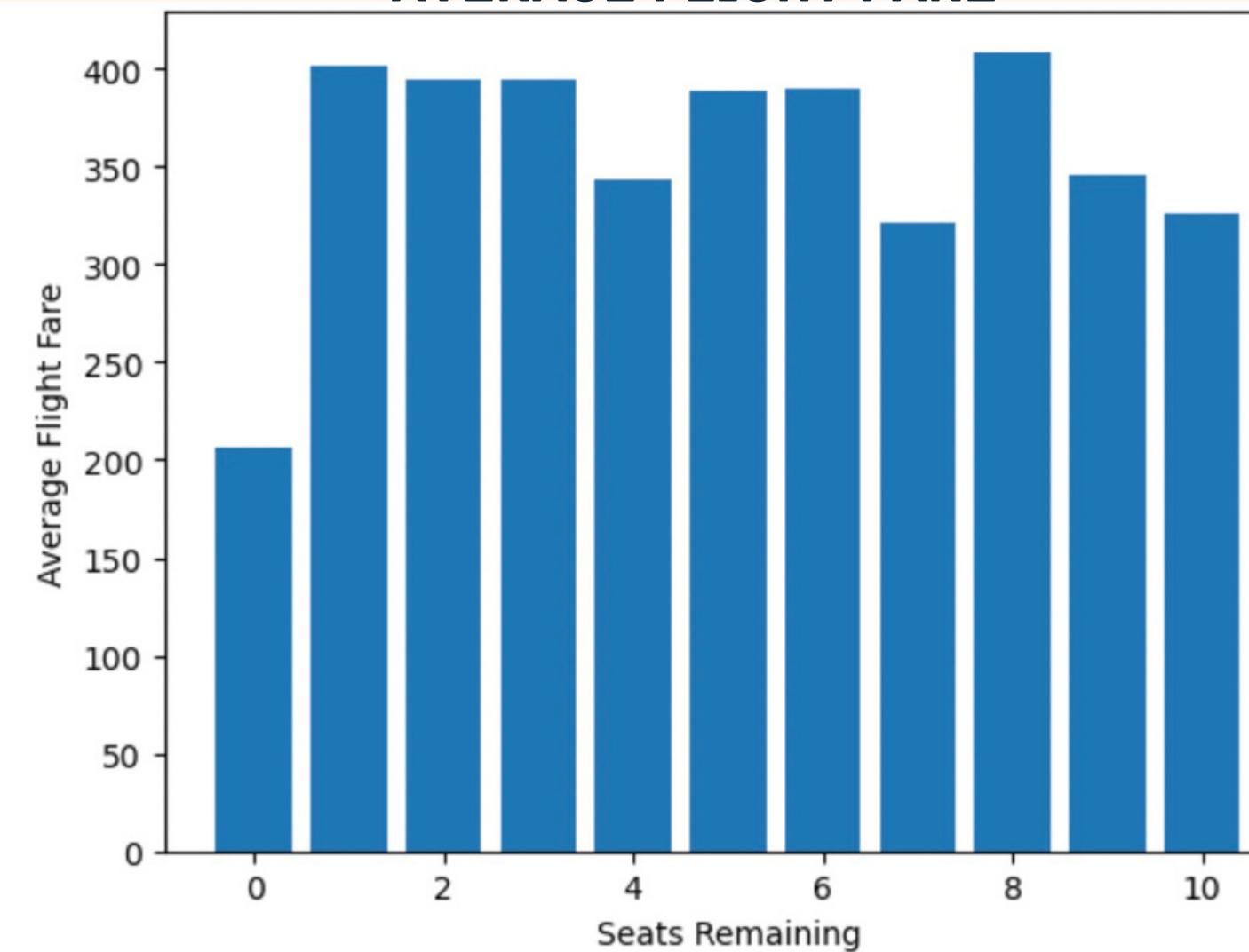


"A significant difference in average flight fare between refundable ticket and non-refundable ticket."



# MORE EMPTY SEATS = CHEAP?

REMAINING SEATS AND CORRESPONDING AVERAGE FLIGHT FARE



OLS REGRESSION ANALYSIS

OLS Regression Results						
Dep. Variable:	avg_fare	R-squared:	0.027			
Model:	OLS	Adj. R-squared:	-0.081			
Method:	Least Squares	F-statistic:	0.2466			
Date:	Sun, 30 Apr 2023	Prob (F-statistic):	0.631			
Time:	11:14:29	Log-Likelihood:	-59.762			
No. Observations:	11	AIC:	123.5			
Df Residuals:	9	BIC:	124.3			
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Intercept	341.6940	34.528	9.896	0.000	263.586	419.802
seatsRemaining	2.8980	5.836	0.497	0.631	-10.305	16.101
Omnibus:	5.707	Durbin-Watson:	1.741			
Prob(Omnibus):	0.058	Jarque-Bera (JB):	2.514			
Skew:	-1.141	Prob(JB):	0.285			
Kurtosis:	3.527	Cond. No.	11.3			

# PART 2: REGRESSION



## FEATURE ENGINEER

- Date difference : Search Date -Flight Date
- Flight duration : Deal with transfer flight
- Drop null value

## TRANSF ORMER

```
RFormula ( formula="totalFare ~ . ",  
featuresCol="features", labelCol="totalFare")
```

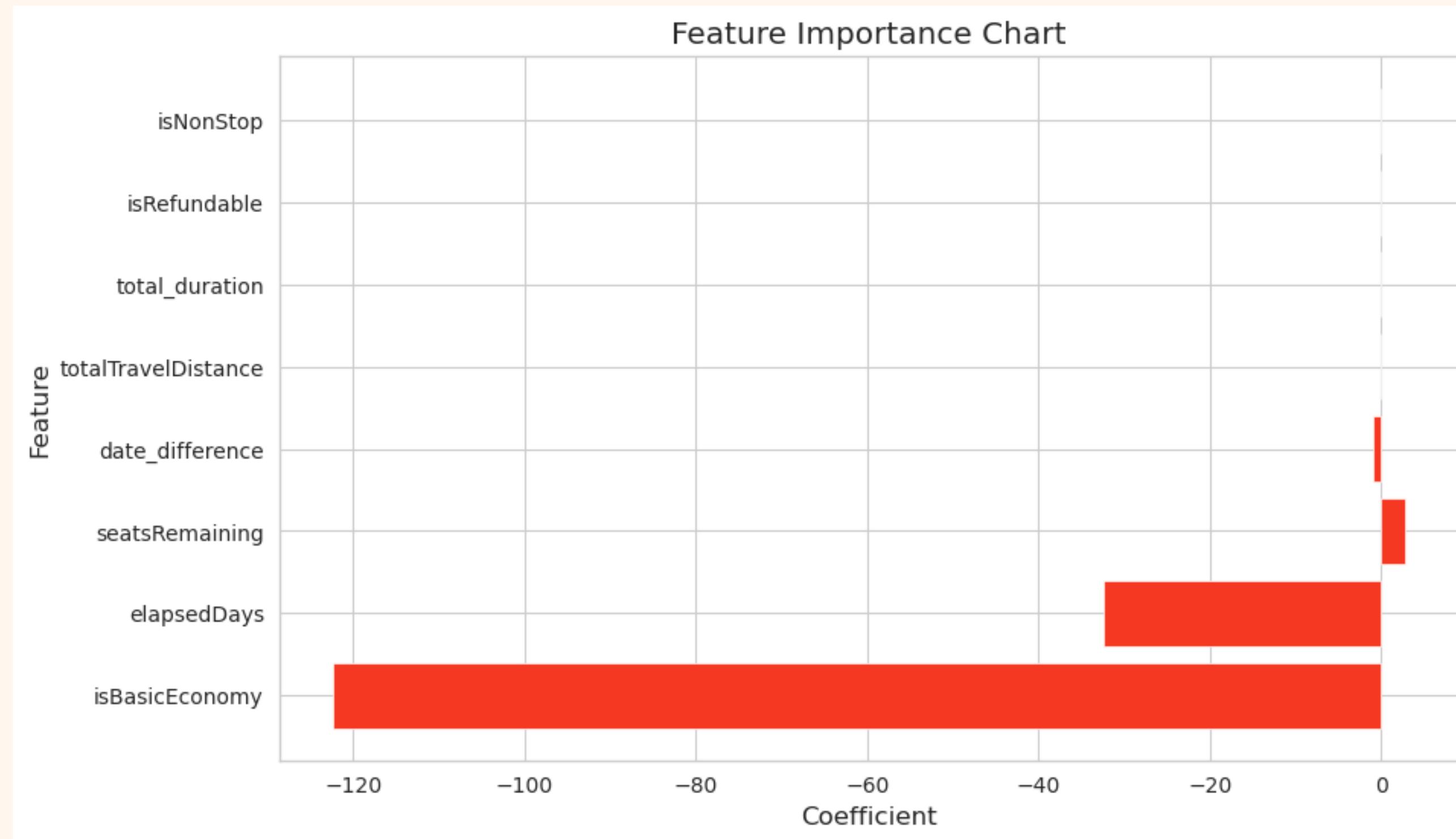
## MODEL FITTING

Linear Regression: R-squared value: 0.5358  
MAE: 72.1836

Random Forest : R-squared value: 0.5055  
MAE: 77.5210

RF has higher MSE, MAE, and RMSE than LR

# FEATURE IMPORTANCE



# CONCLUSION & SUGGESTIONS

- Impactful pricing factors:
  - Class of the ticket listing
  - Elapsed days
  - Seats remaining
  - Search & flight date difference
- Direct flight ticket is cheaper than transfer flight
- 45 days earlier
- Non-refundable
- Avoid peak time periods

## Future Works:

- Expand the dataset to:
  - A whole year
  - Including international flights



**THANK YOU!**

