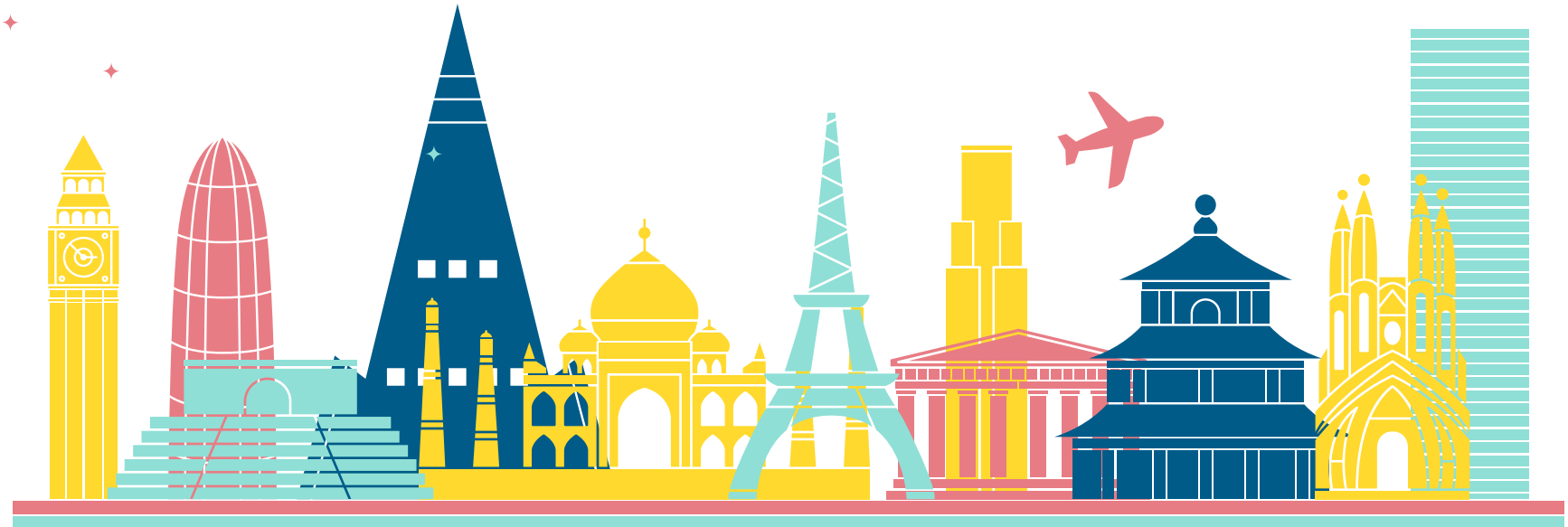
Attempting to predict flight prices



The Goal

I hypothesized that based on key flight features I would be able to build a model that can help predict the price of those flights

Work Flow Chart

01
Scraping Data

Kayak's website code is dynamic, this made scraping difficult and time consuming

02
Cleaning Data

Removing unnecessary features. Cleaning the data collected so it could be interpreted.

03
Feature Engineering

Creating a 'Days to Trip' column. Evaluating pair plots. Manipulating features to understand their relationship to the target.

04
Modeling

Building and evaluating models to find the one with the best score across train and validation sets

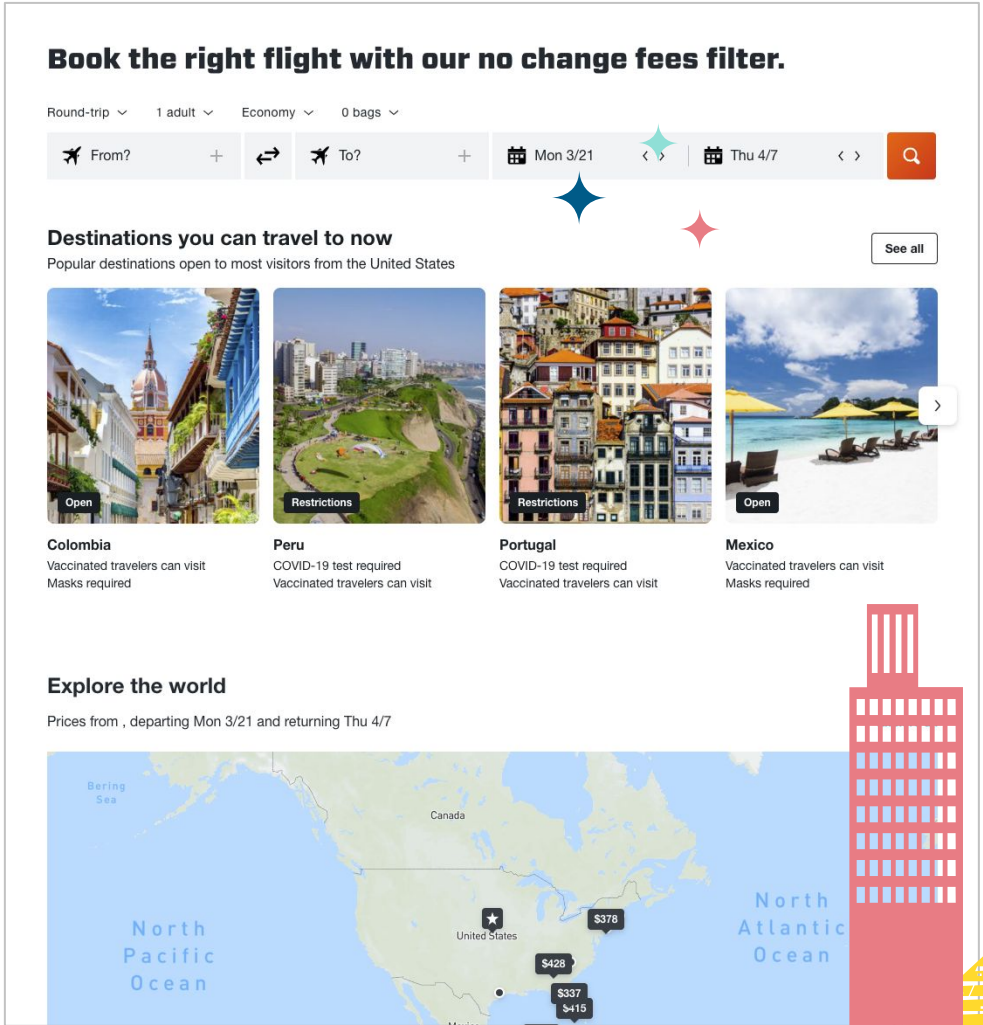
Modeling Framework

Goal:

Predict flight prices base on flight features.

Features:

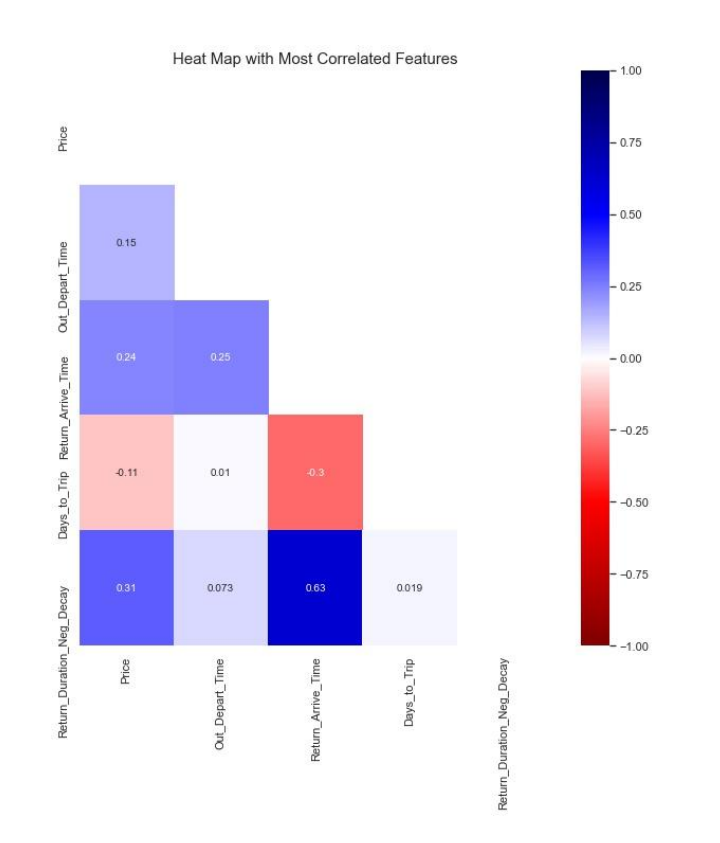
- Days to Trip
- Length of flight
- Departure time
- Arrival time
- Day of week the flight falls on

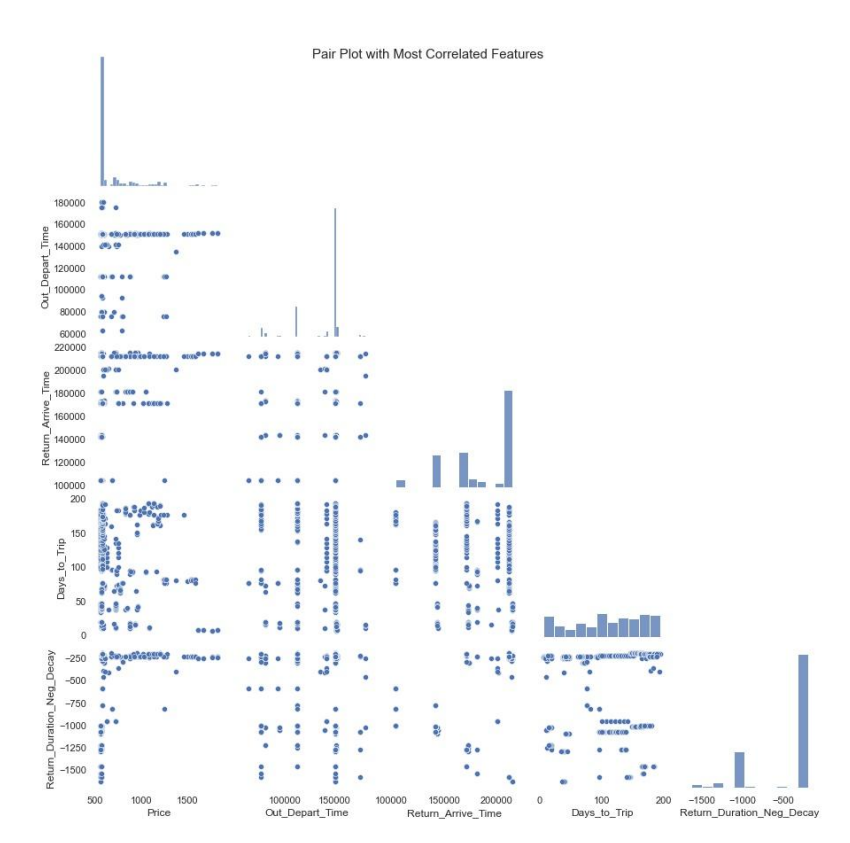


The Data

- Web-Scraped from Kayak.com
- Scraped flights from October 24th of this year to April 24th
- Scraped data for 1027 flights, after removing duplicates, I had 738 flights

Looking for Correlation





Key Features

Days to Trip

Number of days between today and the day the first flight leaves

01

Return Duration

The negative decay of the duration of the return trip

02

03

Departure time

The outgoing time of the flight from the origin

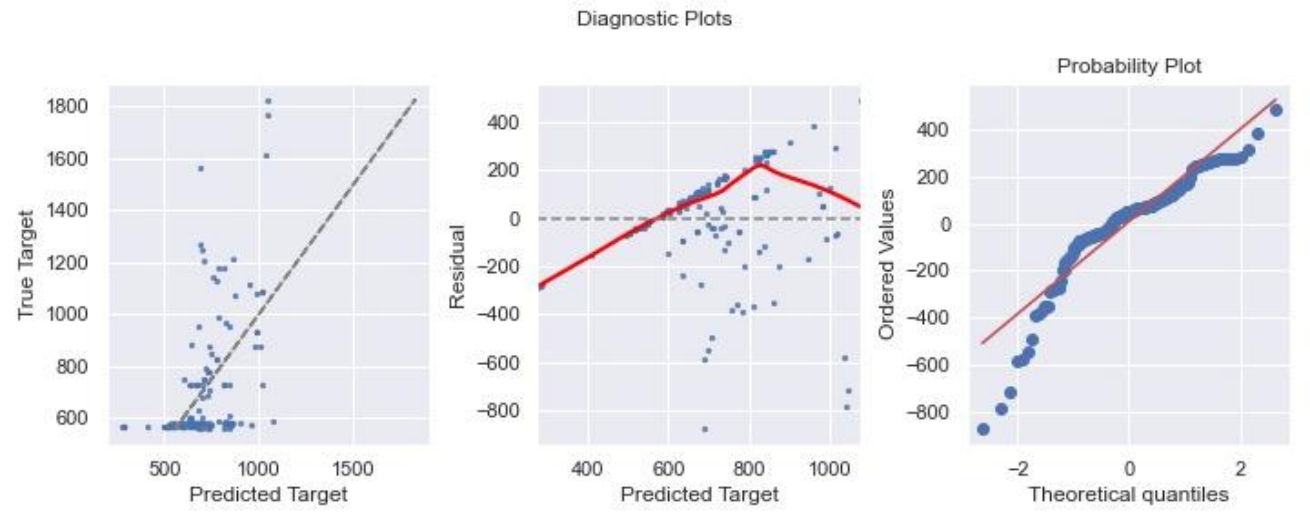
Results

Diagnostic plots for Training Data when using a third degree polynomial function with the negative decay of the return trip duration, the days to the trip and the outgoing departure time

R^2 Score with 5 fold cross validation

Training Data
.1633

Validation Data
.1537



Future Work

- Using a time series model
- Gathering data from different Origin and Destination cities to see how that affects the models accuracy
- Gather data from other travel sites to determine how trends change across travel platforms

Questions?