Machine Learning in the Clouds

Jeremy Beard

MSDS 696

Capstone

December 2023

> “TRAVEL IS NOT REWARD FOR WORKING, IT’S EDUCATION FOR LIVING”

Anthony Bourdain

AGENDA

- Problem Statement
- Methods and Operations
- Notable EDA Plots
- ML Overview
- ML Results
- Future Work
- Conclusions
- References

Problem Statement

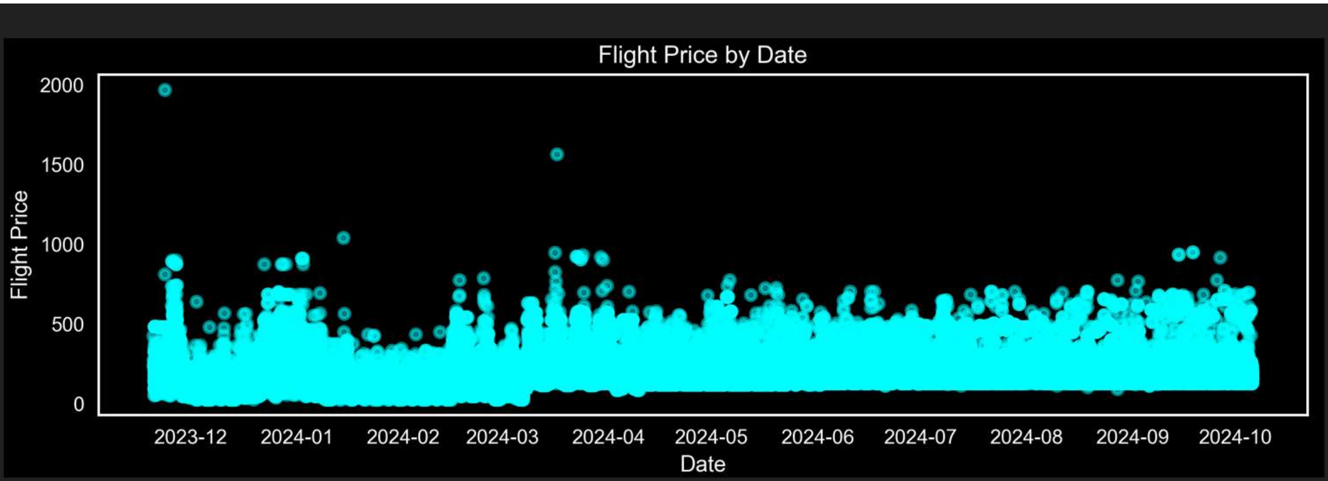
"How can we reliably predict trends and prices for air travel based on common real-life inputs to leverage the same ML tactics corporations are utilizing, but on consumer-side?"

- The challenge is to implement a web-scraping solution that extracts relevant data, including departure details, flight duration, layovers, class types, and current/historical pricing information, from a relevant online source (or sources).
- Solution will enable the creation of a precise predictive model and empower stakeholders with valuable insights for informed decision-making in the airline industry.

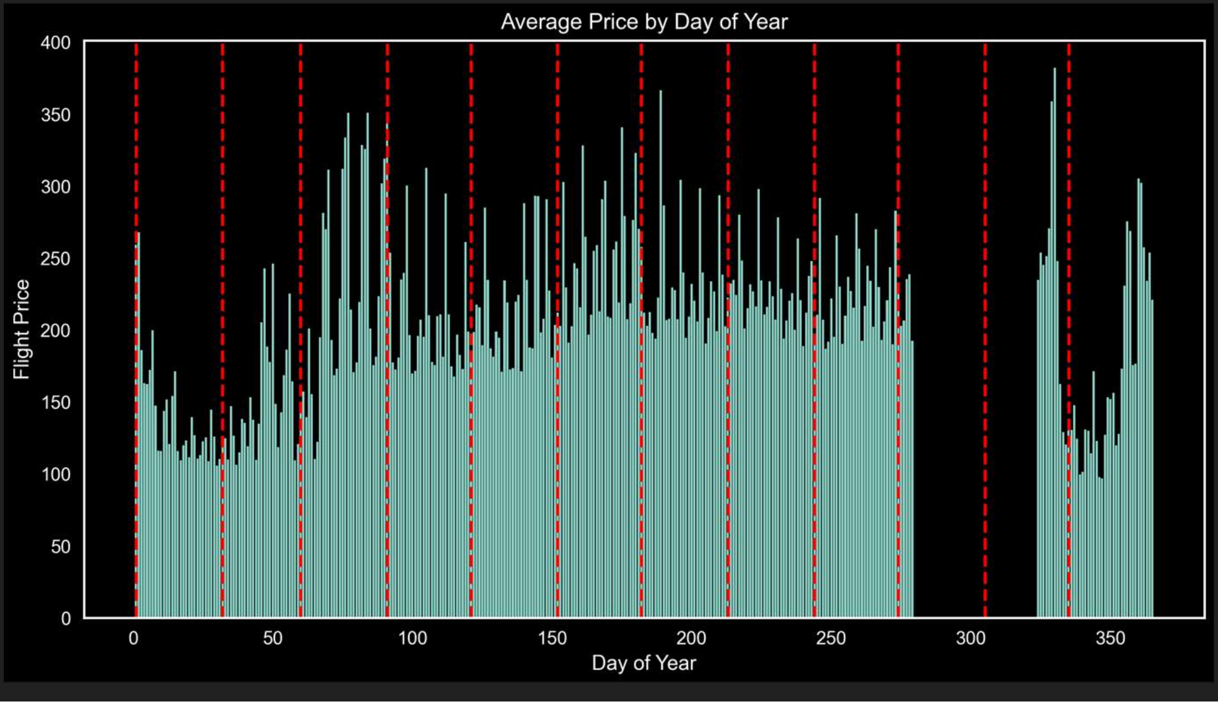
Methods, Operations & Processes

- Web-scraping (via API)
- EDA
- Tableau Dashboarding
- Model Creation
- Model Comparison
- Final Model Selection
- Summary

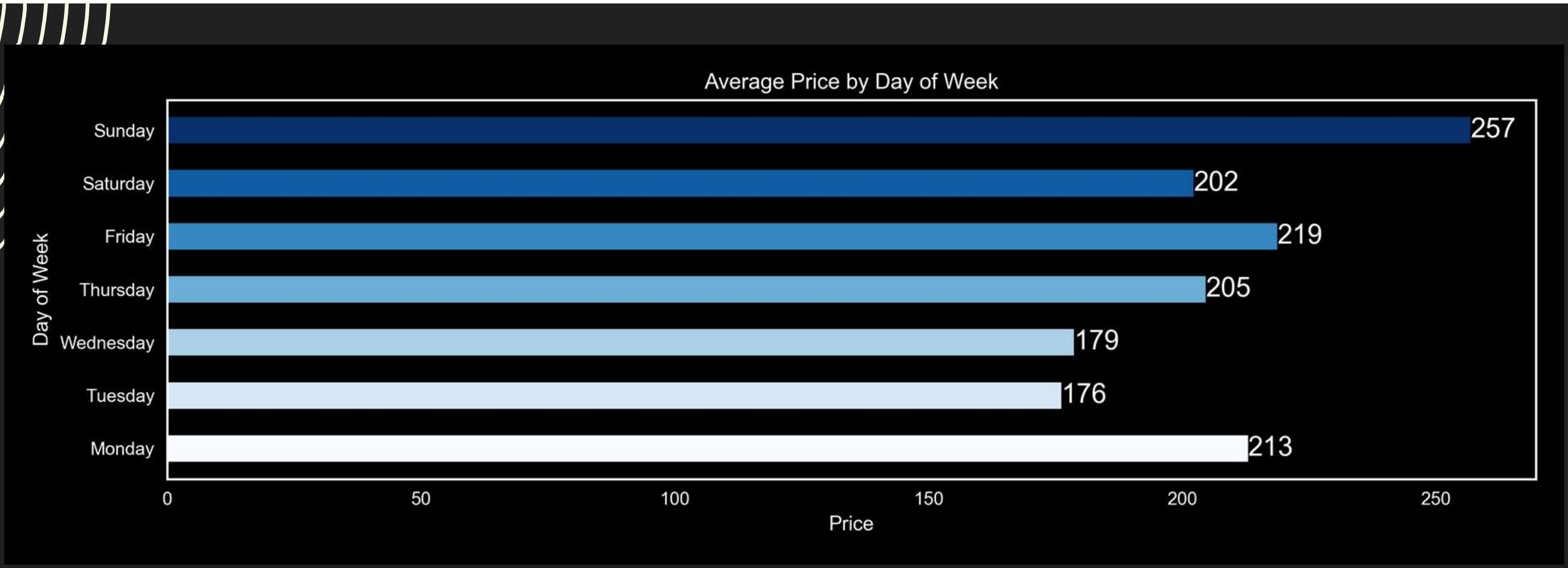
Notable EDA

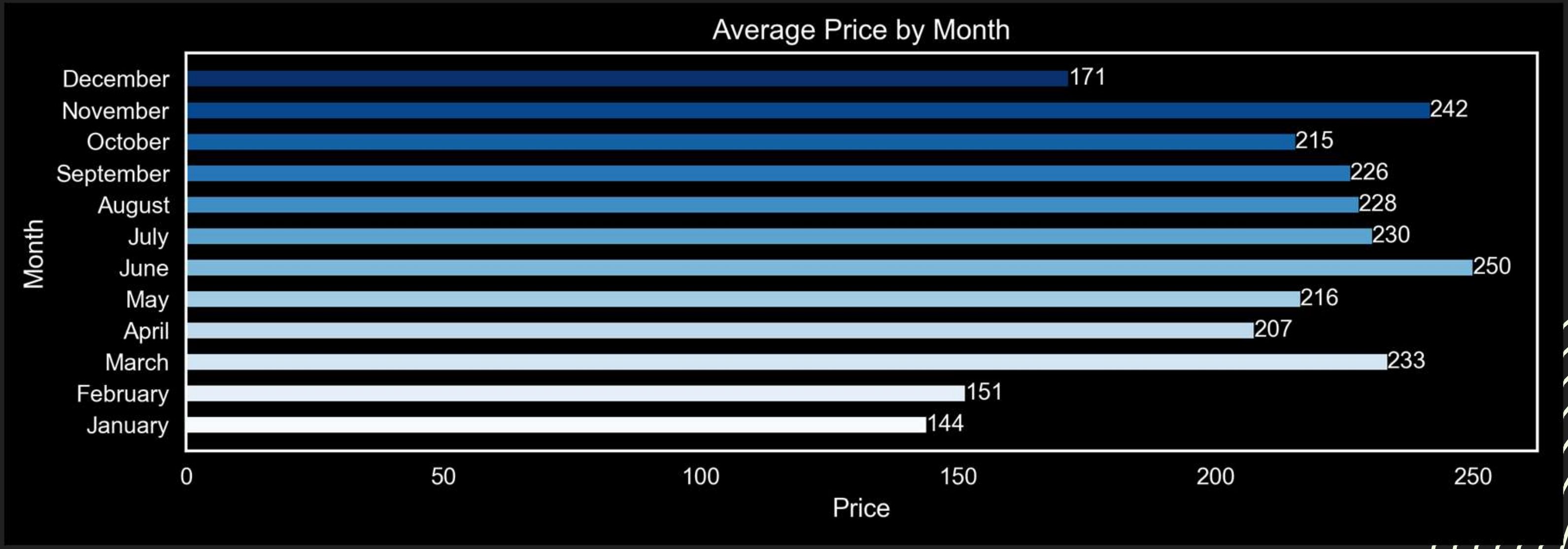


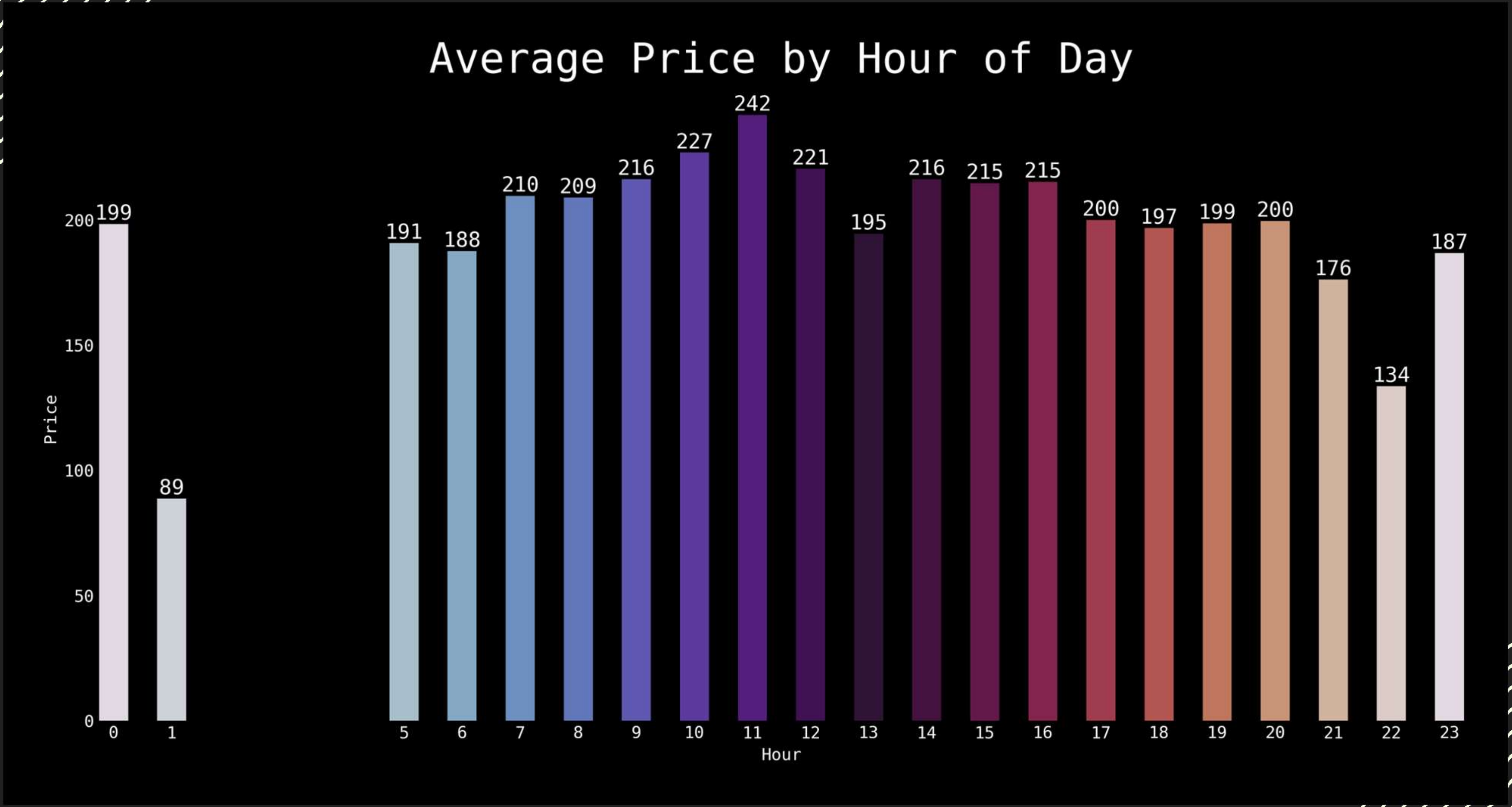
Flight Price by Date

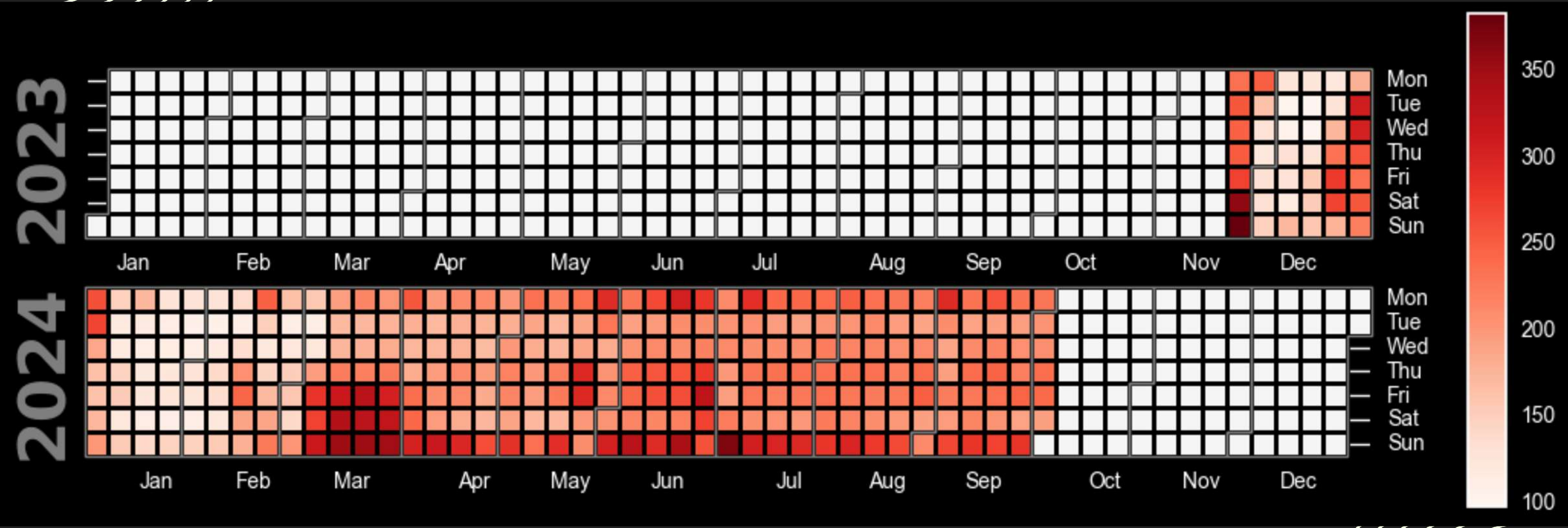


Average Price by Day of Year

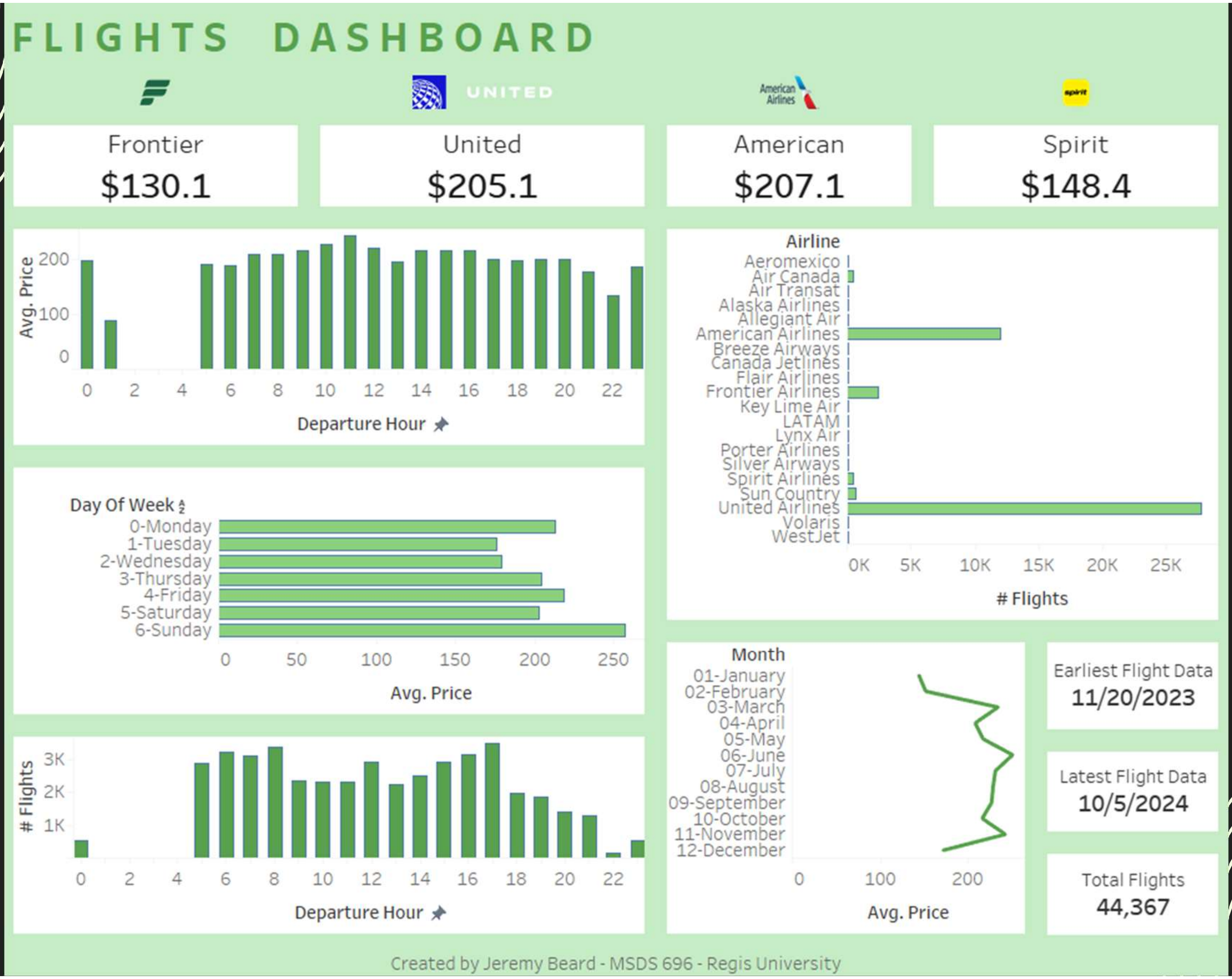

Average Price by Day of Week
Day of Week
Sunday
Saturday
Friday
Thursday
Wednesday
Tuesday
Monday
257
202
219
205
179
176
213
0
50
100
150
200
250
Price


Average Price by Month
Month
December
November
October
September
August
July
June
May
April
March
February
January
171
242
215
226
228
230
250
216
207
233
151
144
0
50
100
150
200
250
Price


Average Price by Hour of Day
Price
Hour
0
50
100
150
200
199
89
191
188
210
209
216
227
242
221
195
216
215
215
200
197
199
200
176
134
187
0
1
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23


2023
Mon
Tue
Wed
Thu
Fri
Sat
Sun
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
2024
Mon
Tue
Wed
Thu
Fri
Sat
Sun
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
350
300
250
200
150
100

Tableau



FLIGHTS DASHBOARD

Frontier	United	American	Spirit
$130.1	$205.1	$207.1	$148.4

Earliest Flight Data 11/20/2023

Latest Flight Data 10/5/2024

Total Flights 44,367

Created by Jeremy Beard - MSDS 696 - Regis University

Machine Learning Overview

ML Overview

- Utilized **pycaret** Python package for rapidly iterating on multiple ML models

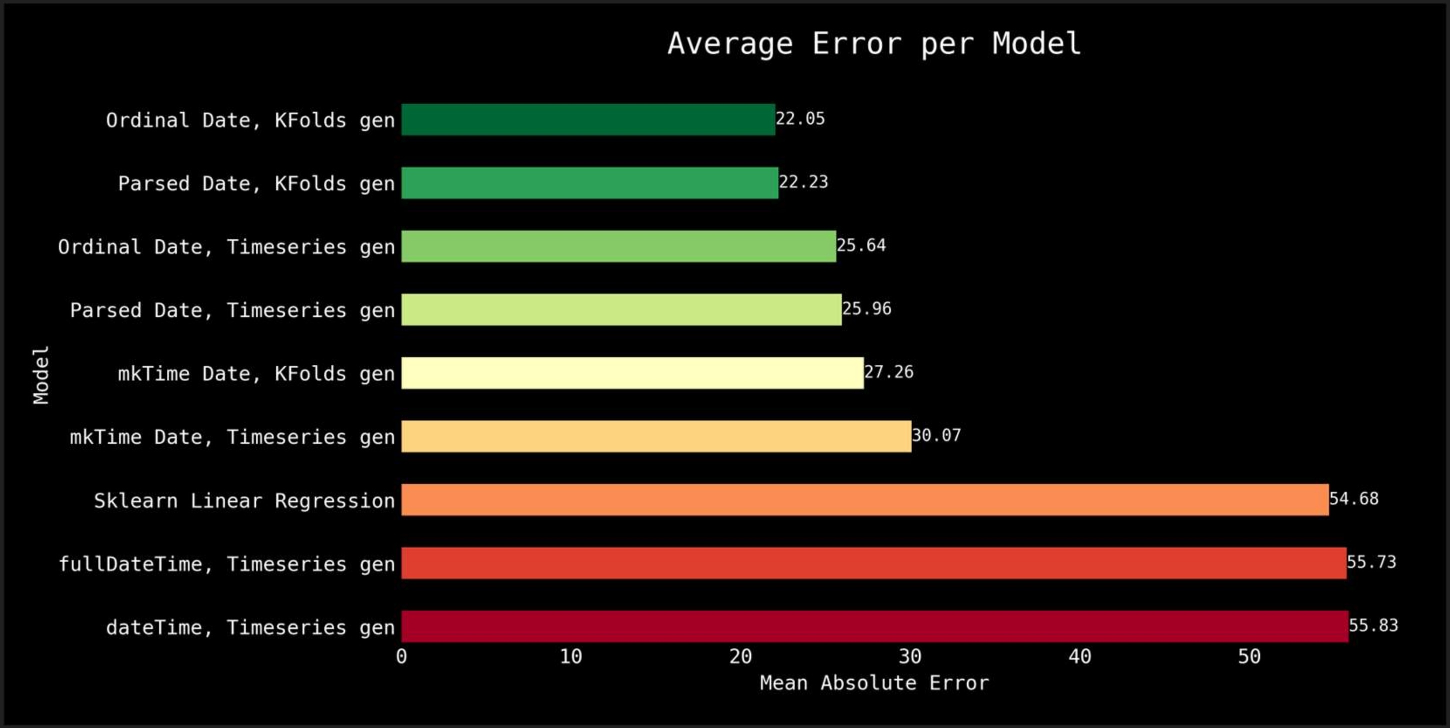
- Compared multiple dimensions of model parameters
    - Focus: timeseries handling
    - Compared 3-5 different methods of handling datetimes

- Created **1** ML model baseline and **8** pycaret iterations through 8-10 ML models each

ML Model Results

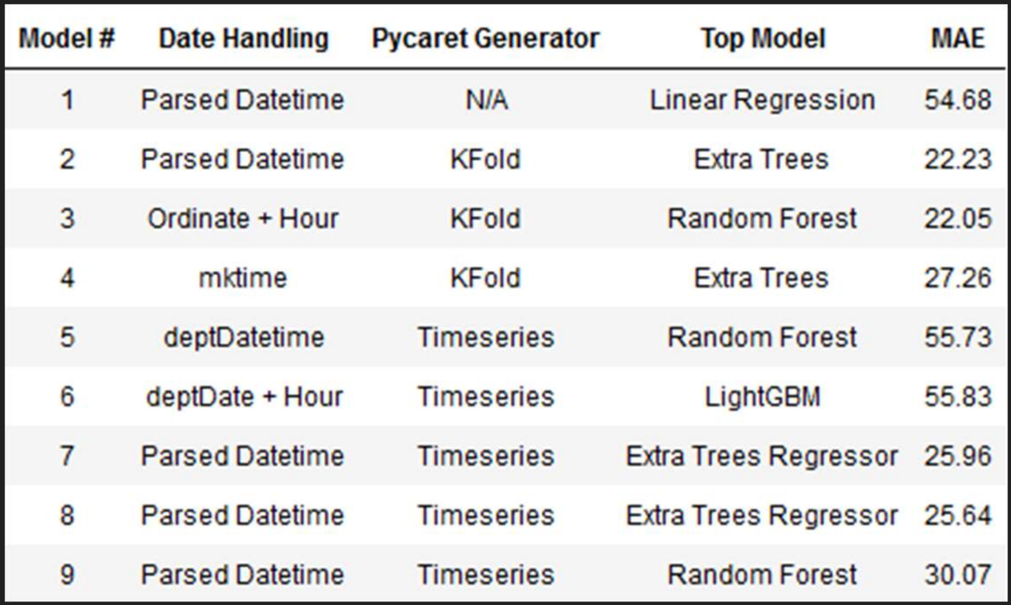
ML Model Results

- **Mean Absolute Error:** Average of the absolute differences between the predicted and actual values

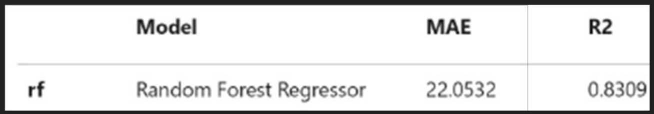


ML Model Results

- Overall good performance
  - MAE: $22.05
  - R^2: 0.83
- Pycaret timeseries generator did not perform as well as expected
- Random Forest and Extra Trees



Model #	Date Handling	Pycaret Generator	Top Model	MAE
1	Parsed Datetime	N/A	Linear Regression	54.68
2	Parsed Datetime	KFold	Extra Trees	22.23
3	Ordinate + Hour	KFold	Random Forest	22.05
4	mktime	KFold	Extra Trees	27.26
5	deptDatetime	Timeseries	Random Forest	55.73
6	deptDate + Hour	Timeseries	LightGBM	55.83
7	Parsed Datetime	Timeseries	Extra Trees Regressor	25.96
8	Parsed Datetime	Timeseries	Extra Trees Regressor	25.64
9	Parsed Datetime	Timeseries	Random Forest	30.07



	Model	MAE	R2
rf	Random Forest Regressor	22.0532	0.8309

Future Work

- More data!
  - More airports, more airlines, continually update the sqlite3 database
  - 5-sec latency API queries made 45,000 datapoints take all weekend
  - Research other API's
  - Implement selenium?
- Implement more feature engineering
- Experiment with more Tableau dashboarding

Conclusions

- Using the Booking.com API, we were able to create a model with $22 Mean Absolute Error prediction
- Data is biased according to when it was scraped
- Best Models: Random Forest, Extra Trees
  - Using Ordinal datetime
- Timeseries still a question in my mind, it should perform better

References

- https://github.com/jeremyabeard5/MSDS696
- https://booking-com13.p.rapidapi.com/flights/one-way
- https://stackoverflow.com/questions/40217369/python-linear-regression-predict-by-date
- https://matplotlib.org/stable/gallery/style_sheets/style_sheets_reference.html
- https://pandas.pydata.org/docs/reference/api/pandas.DatetimeIndex.dayofweek.html
- https://pandas.pydata.org/docs/reference/api/pandas.Series.dt.month.html
- https://stackoverflow.com/questions/71419004/how-to-plot-vertical-lines-at-specific-dates-in-matplotlib
- https://medium.com/analytics-vidhya/calendar-heatmaps-a-perfect-way-to-display-your-time-series-quantitative-data-ad36bf81a3ed
- https://calplot.readthedocs.io/en/latest/index.html
- https://matplotlib.org/stable/users/explain/colors/colormaps.html
- https://medium.com/analytics-vidhya/calendar-heatmaps-a-perfect-way-to-display-your-time-series-quantitative-data-ad36bf81a3ed
- https://pypi.org/project/july/
- https://datascience.stackexchange.com/questions/2368/machine-learning-features-engineering-from-date-time-data
- https://datascience.stackexchange.com/questions/112357/feature-engineering-for-datetime-column
- https://www.reddit.com/r/learnpython/comments/chunas/correlation_with_day_of_week/
- https://mikulskibartosz.name/time-in-machine-learning
- https://www.pycaret.org/tutorials/html/REG101.html
- https://pycaret.readthedocs.io/en/latest/api/regression.html

THANK YOU

Jeremy Beard

jeremyab5@gmail.com

https://www.linkedin.com/in/jeremyab5