

IBM DATA SCIENCE PROFESSIONAL CERTIFICATE

Applied Data Science Capstone:
Final Project

26 December 2021

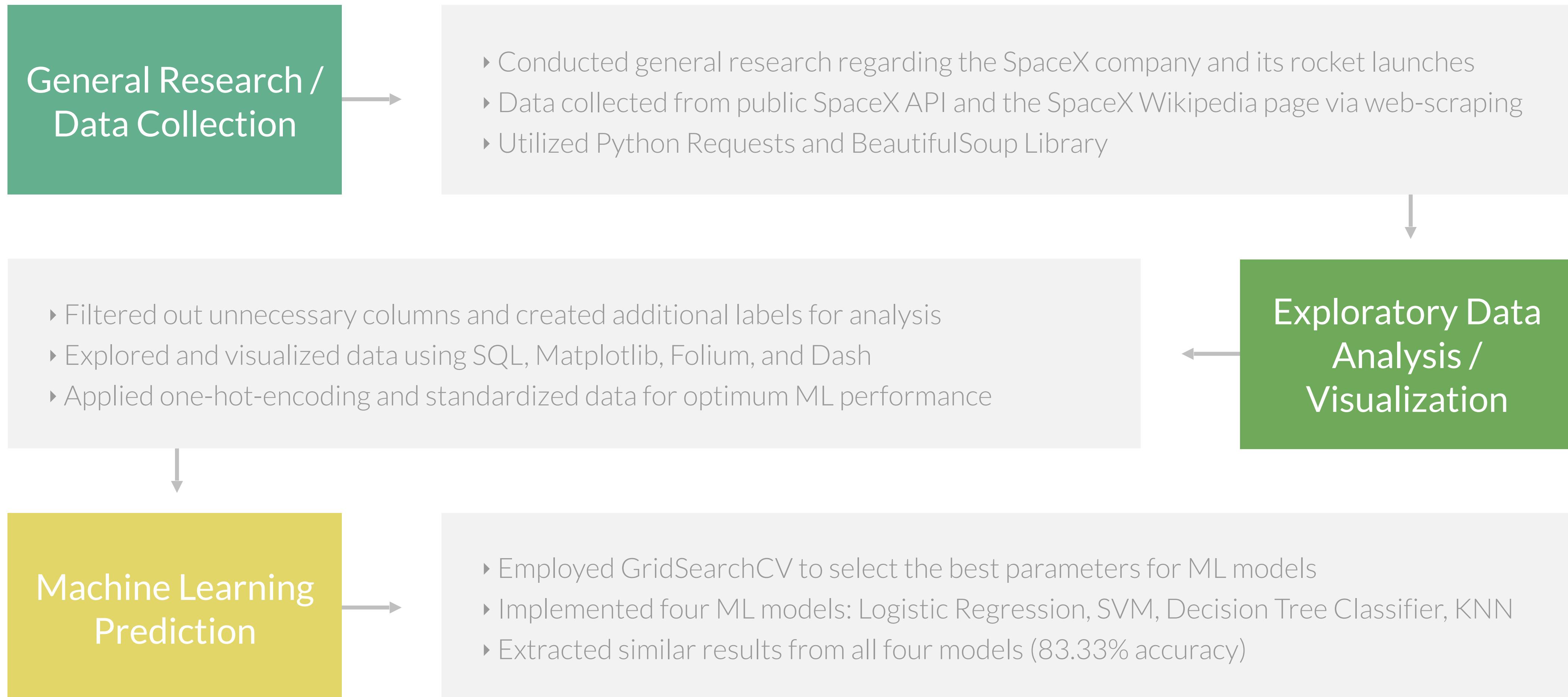
SCOTT KIM

github.com/scottpjkim

Table of Contents

1	Executive Summary	3
2	Introduction	4
3	Methodology	5
4	Predictive Analysis Results	18
5	Conclusion	19
6	Appendix	20

Executive Summary



Introduction



"The Company was founded in 2002 to revolutionize space technology, with the ultimate goal of enabling people to live on other planets."

Improving Lives

- Create alternative habitats for humans
- Prioritizes environmental sustainability

Exceeding Expectations

- Advancements in space exploration
- Ultimate goal is to conquer other planets

Revolutionize Space Technology

- Competitive pricing for space flight
- Stage 1 rocket recovery decreases costs

PROBLEM: Develop and train a machine learning model to predict successful Stage 1 recovery

Methodology Overview

1

Data Collection

Collect data from SpaceX's public API and SpaceX Wikipedia page

2

Data Wrangling

Classify Stage 1 recovery landings as successful VS unsuccessful

3

Exploratory Data Analysis (EDA)

Conduct EDA and Visualization via SQL, Matplotlib, Folium and Dash

4

Machine Learning Models

Perform predictive analysis using ML classification models

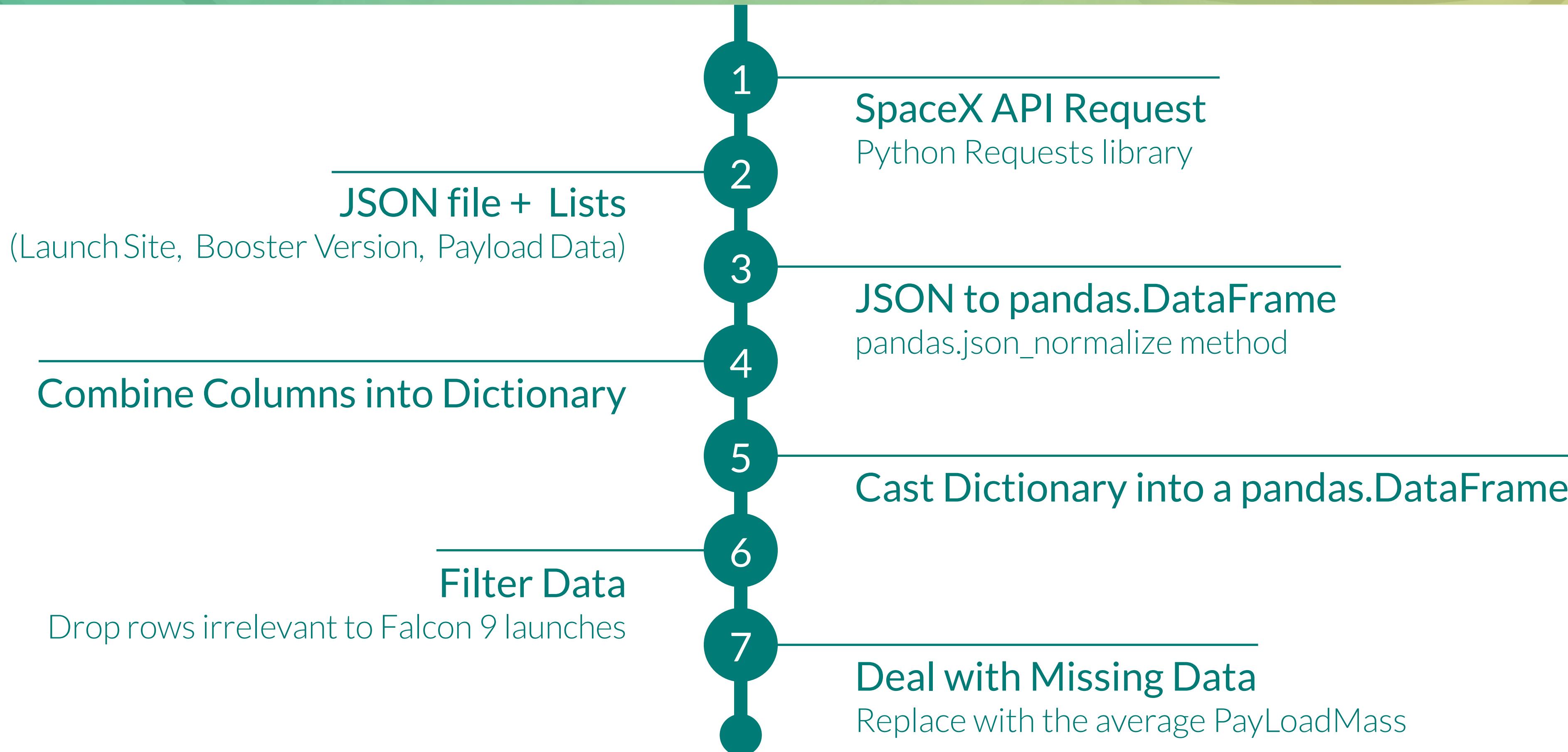
Data Collection

The SpaceX launch dataset was collected from a combination of API requests from SpaceX's public API via the Python Requests library and web-scraping tables of data from the SpaceX Wikipedia site.

SpaceX Public API Data Columns	SpaceX Wikipedia Web-Scrape Columns
<ul style="list-style-type: none">▶ FlightNumber▶ Date▶ BoosterVersion▶ PayloadMass▶ Orbit▶ LaunchSite▶ Outcome▶ Flights <ul style="list-style-type: none">▶ GridFins▶ Reused▶ LegsLandingPad▶ Block▶ ReusedCount▶ Serial▶ Longitude▶ Latitude	<ul style="list-style-type: none">▶ Flight No.▶ Launch site▶ Payload▶ PayloadMass▶ Orbit▶ Customer▶ Launch outcome▶ Version Booster <ul style="list-style-type: none">▶ Booster landing▶ Date▶ Time

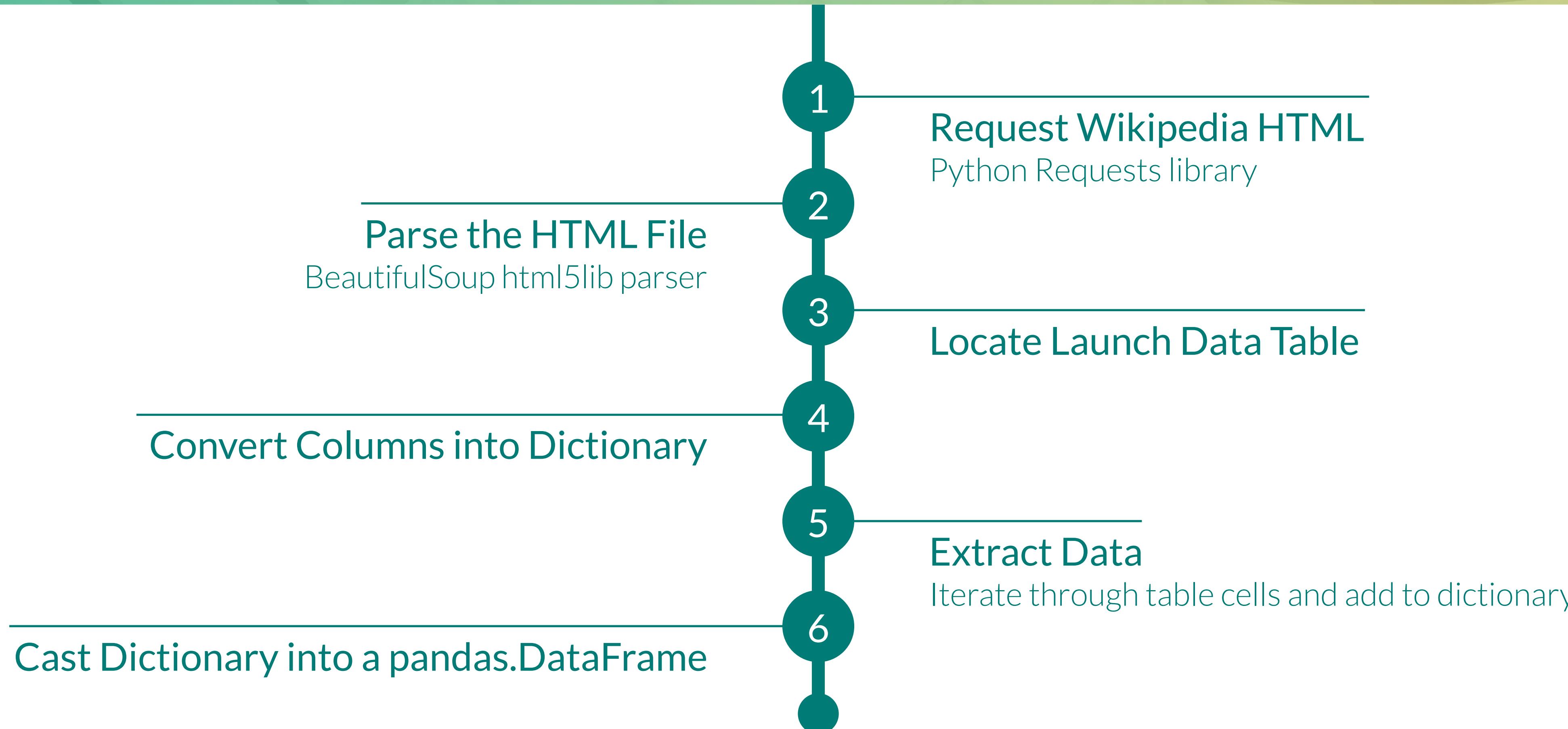
The following slides will reveal a flowchart of the data collection from the SpaceX API and from web-scraping the Wikipedia page.

Data Collection: API



GitHub Link
tinyurl.com/2p8p4m9h

Data Collection: Web-Scraping



Data Wrangling

To optimize the dataset for Machine Learning, a binary landing outcome column was created as part of the dataset where successful and unsuccessful landings were labeled 1 and 0 respectively.

Binary Value Mapping of Falcon 9 Landings	
Set to 0 (Unsuccessful)	True ASDS True RTLS
Set to 1 (Successful)	None None False ASDS None ASDS False Ocean False RTLS

EDA with SQL

Loaded data set into IBM DB2 Database.

Applied SQL queries via iPython Notebook SQL integration.

Queries were made to adjust the data for a more efficient analysis of the data.

Queried information about launch site names, pay load sizes, booster versions, and landing outcomes.

The following slide reveals the SQL queries used to complete these tasks.

EDA with SQL

Task 1

Display the names of the unique launch sites in the space mission

```
%sql select DISTINCT LAUNCH_SITE from SPACEXDATASET
```

Task 2

Display 5 records where launch sites begin with the string 'CCA'

```
%sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5
```

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql select sum(payload_mass_kg_) as sum from SPACEXDATASET where customer like 'NASA (CRS)'
```

Task 4

Display average payload mass carried by booster version F9 v1.1

```
%sql select avg(payload_mass_kg_) as Average from SPACEXDATASET where booster_version like 'F9 v1.1%'
```

Task 5

List the date when the first successful landing outcome in ground pad was achieved.

Hint: Use min function

```
%sql select min(date) as Date from SPACEXDATASET where mission_outcome like 'Success'
```

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%sql select booster_version from SPACEXDATASET where (mission_outcome like 'Success')  
AND (payload_mass_kg_ BETWEEN 4000 AND 6000) AND (landing_outcome like 'Success (drone ship)')
```

Task 7

List the total number of successful and failure mission outcomes

```
%sql SELECT mission_outcome, count(*) as Count FROM SPACEXDATASET GROUP by mission_outcome ORDER BY mission_outcome
```

Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
maxm = %sql select max(payload_mass_kg_) from SPACEXDATASET  
maxv = maxm[0][0]  
%sql select booster_version from SPACEXDATASET where  
payload_mass_kg_=(select max(payload_mass_kg_) from SPACEXDATASET)
```

Task 9

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015

```
%sql select MONTHNAME(DATE) as Month, landing_outcome, booster_version, launch_site  
from SPACEXDATASET where DATE like '2015%' AND landing_outcome like 'Failure (drone ship)'
```

Task 10

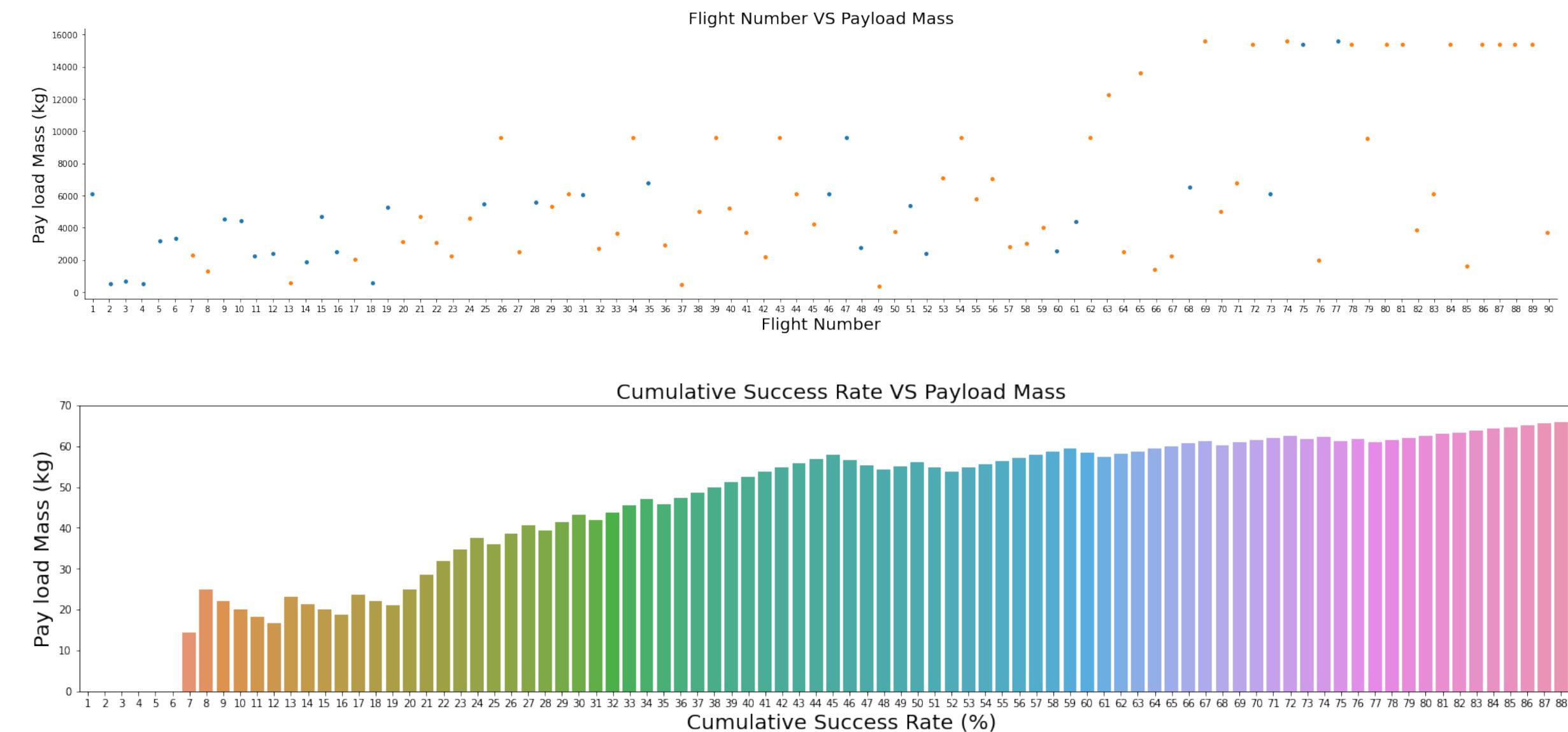
Rank the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

```
%sql select landing_outcome, count(*) as count from SPACEXDATASET  
where Date >= '2010-06-04' AND Date <= '2017-03-20'  
GROUP by landing_outcome ORDER BY count Desc
```

EDA with Visualization

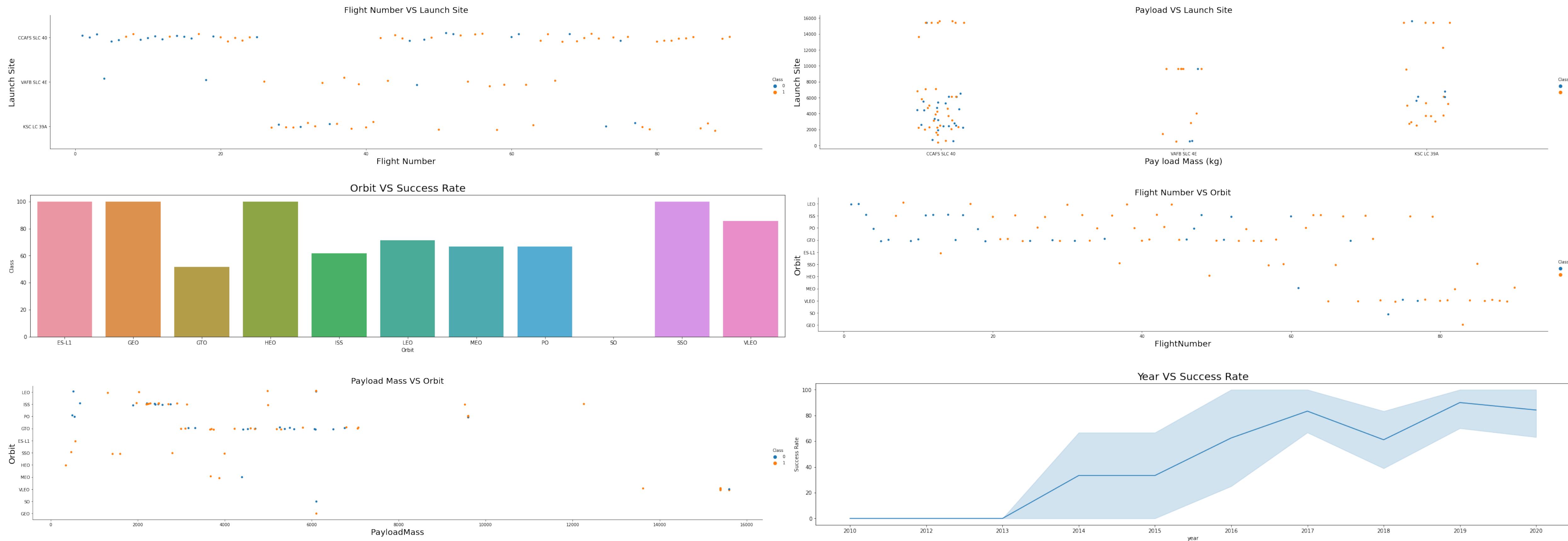
Exploratory Data Analysis performed on Flight Number, Payload Mass, Launch Site, Orbit, Class and Year.

Scatter plots, line charts, and bar plots were used to compare relationships between variables to decide if a relationship exists so that they could be used in training the machine learning model.



GitHub Link
tinyurl.com/2s3vkhj3

EDA with Visualization



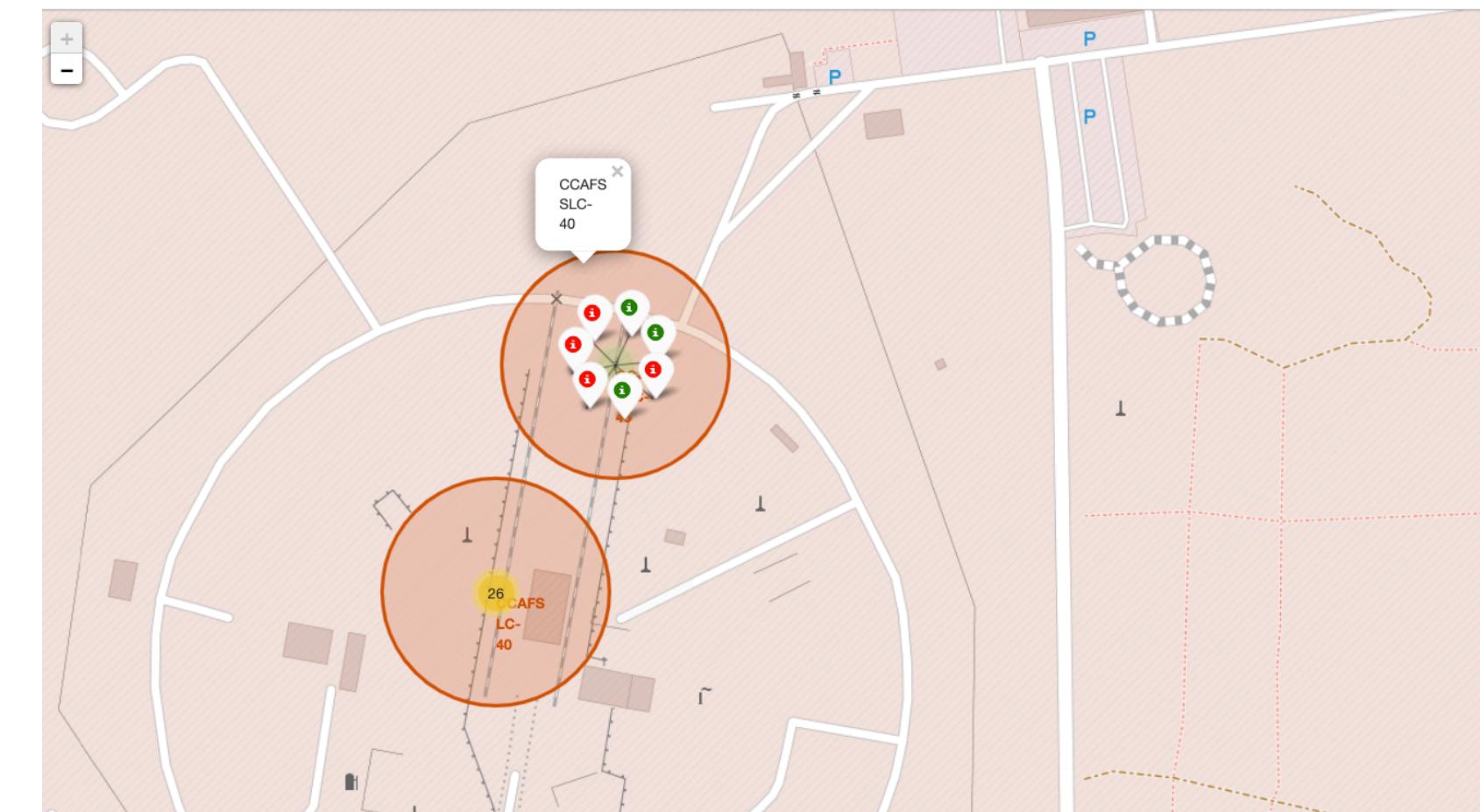
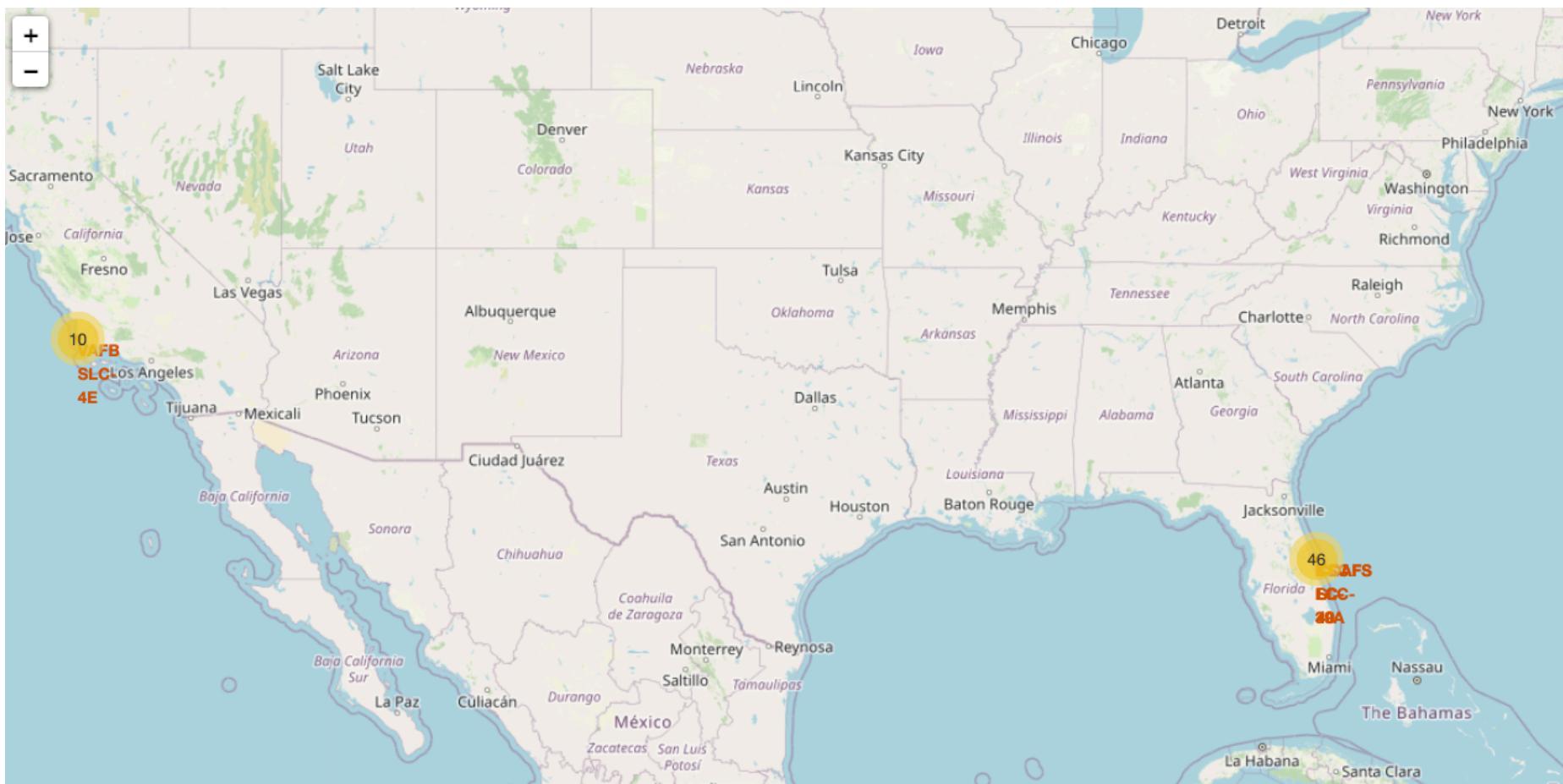
GitHub Link
tinyurl.com/2s3vkhj3

Interactive Map with Folium

Marked all launch sites on a map.

Marked the success/failed launches for each site on the map.

Calculated the distances between a launch site to its proximities.



GitHub Link
tinyurl.com/2p9bxnb3

Dashboard with Plotly Dash

Dashboard includes a pie chart and a scatter plot.

Pie chart can be selected to show distribution of successful landings across all launch sites and can be selected to show individual launch site success rates.

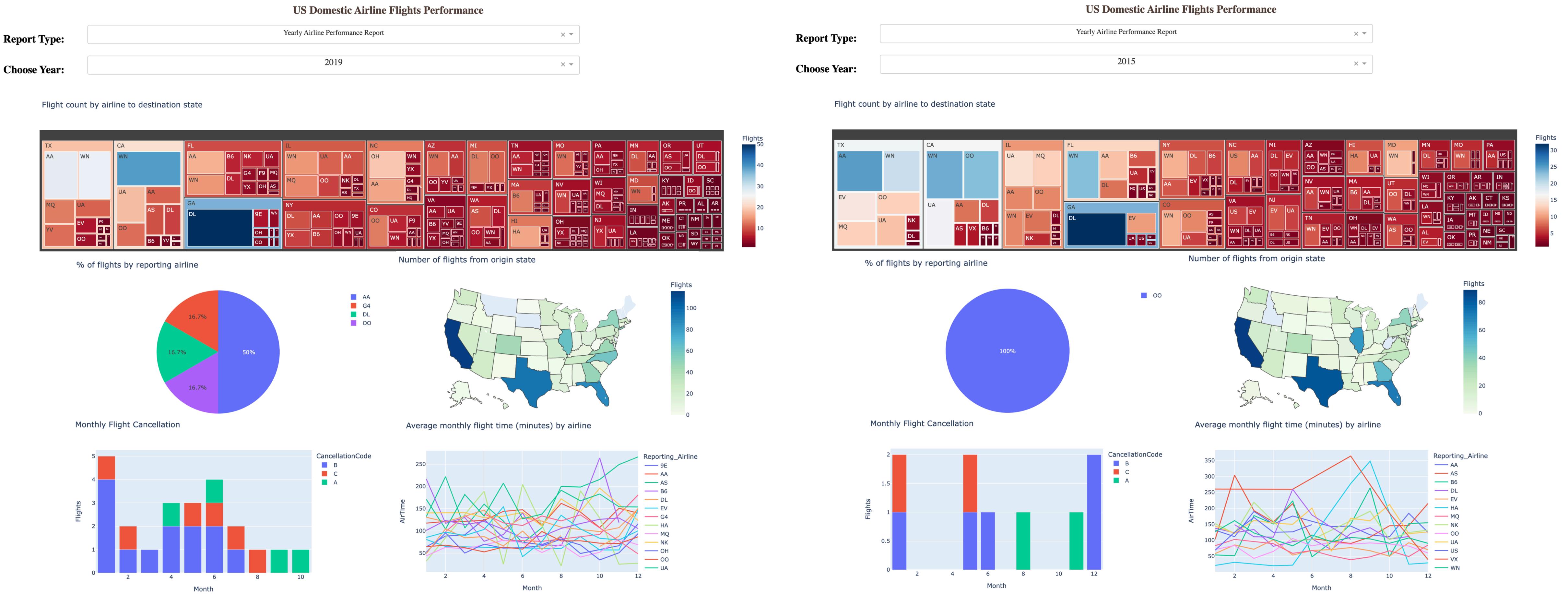
Scatter plot takes two inputs: All sites or individual site and payload mass on a slider between 0 and 10000 kg.

The pie chart is used to visualize launch site success rate.

The scatter plot can help us see how success varies across launch sites, payload mass, and booster version category.

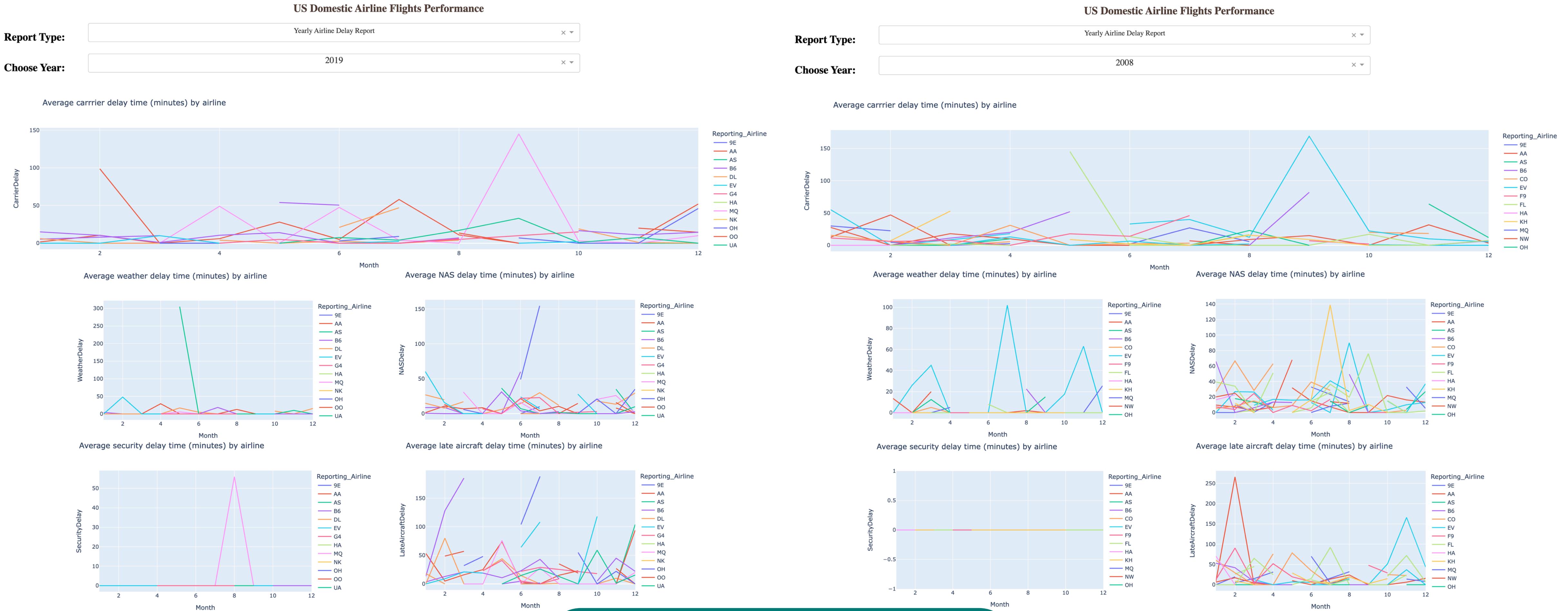
The following slides are screenshots of the dashboard created via Plotly Dash.

Dashboard with Plotly Dash



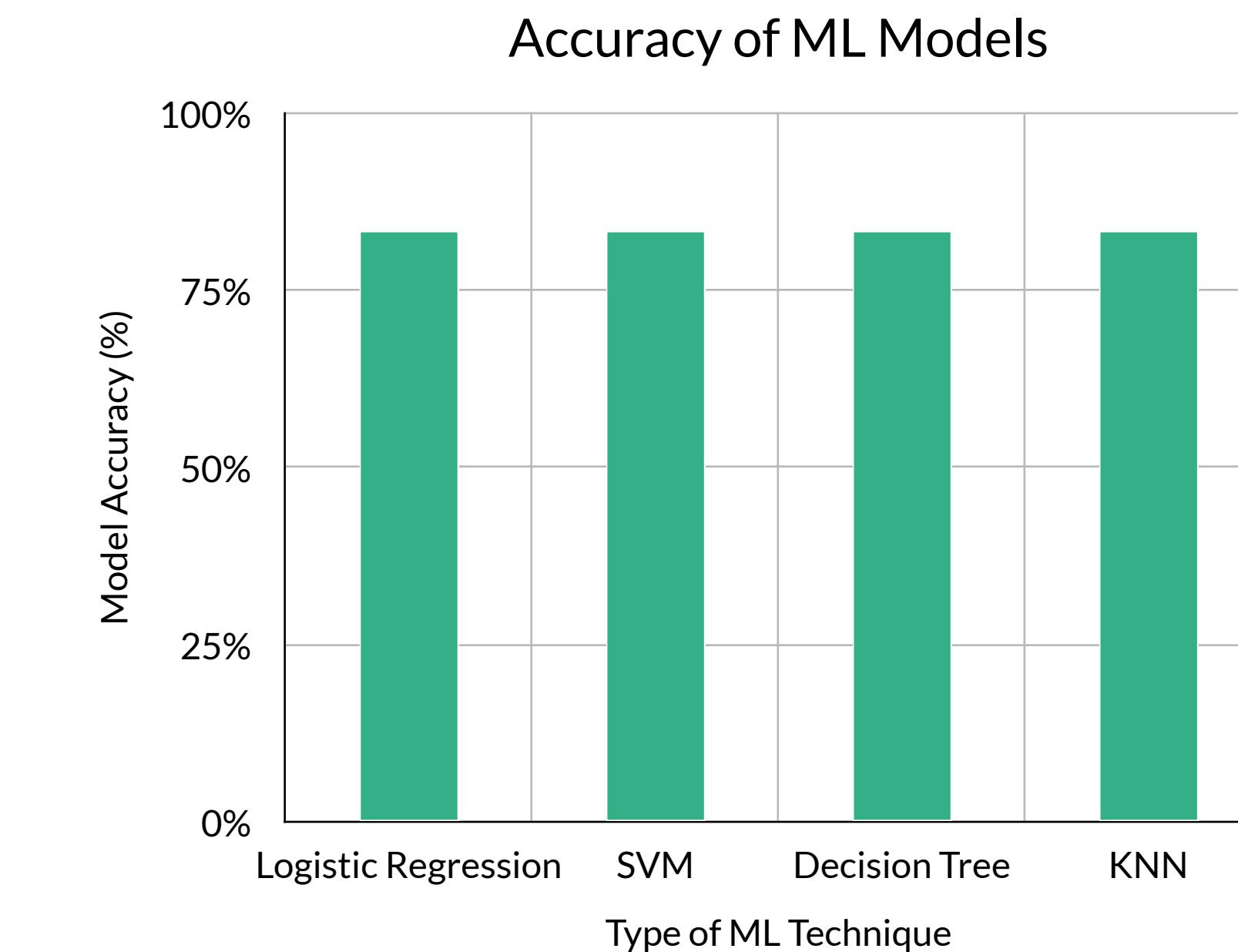
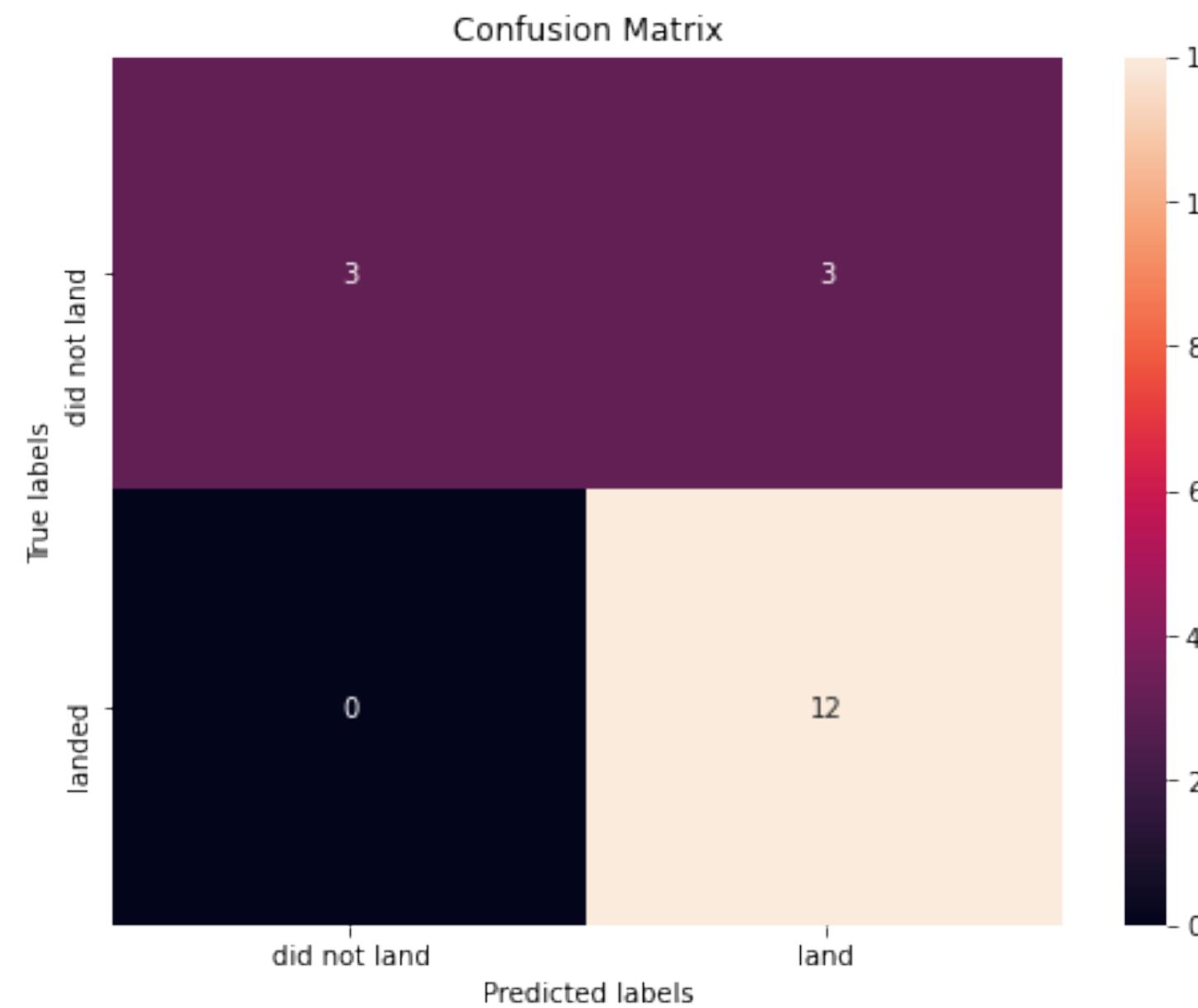
GitHub Link
tinyurl.com/3tbyss4z

Dashboard with Plotly Dash



GitHub Link
tinyurl.com/3tbyss4z

Predictive Analysis Results



Logistic Regression
Accuracy: 83.33%



Support Vector Machine
Accuracy: 83.33%



Decision Tree Classifier
Accuracy: 83.33%



K Nearest Neighbors
Accuracy: 83.33%

Conclusion

Problem: Develop and train a machine learning model to predict successful Stage 1 recovery

Logistic Regression
Accuracy: 83.33%



Support Vector Machine
Accuracy: 83.33%



Decision Tree Classifier
Accuracy: 83.33%



K Nearest Neighbors
Accuracy: 83.33%

Conclusion

Developed and trained Predictive Analysis using Machine Learning with 83.33% accuracy.
Need more launch data to improve accuracy to better determine the optimum ML model.

IBM DATA SCIENCE PROFESSIONAL CERTIFICATE

Appendix



IBM DATA SCIENCE PROFESSIONAL CERTIFICATE

coursera.org/professional-certificates/ibm-data-science

SCOTT KIM

LinkedIn

linkedin.com/in/scottpjkim/

GitHub Link

github.com/scottpjkim