IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Joseph Wraga
01/04/23

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Goal: Will the Falcon 9 first stage land successfully?
- Methodology
  - Data collection
  - EDA with visualization and SQL
  - Interactive visual analytics using Folium and Plotly Dash
  - Predictive analysis using classification models
- Summary of results
  - Launch successes have been trending upward and recently plateaued
  - Site KSC LC-39A has the best success rate
  - Decision tree classifier is ideal, with 94% accuracy on test set

Introduction

- Commercial space exploration is becoming more affordable
- Falcon 9 costs $62 million because can reuse the first stage
  - Competition: Over $165 million
- If first stage will land, we can determine the cost
- Can we predict launch success based on launch location, orbit type, payload mass, etc?

Section 1

Methodology

Methodology

Executive Summary

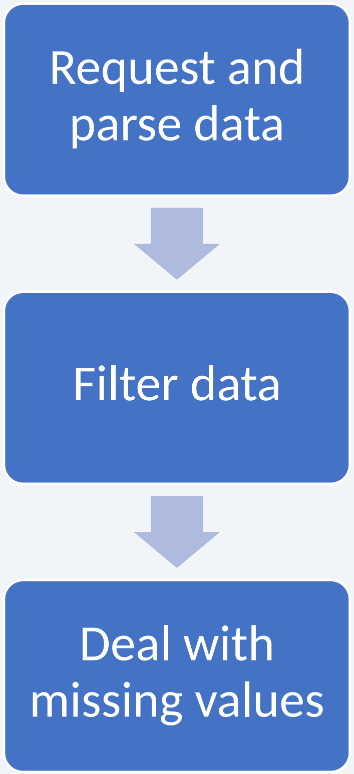
- Data collection methodology:
  - SpaceX API
- Perform data wrangling
  - Filter to include only desired rocket, fill in missing data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Define classification models, determine best hyperparameters, and determine accuracy on test set

Data Collection

- Made get request to SpaceX API, saved as pandas DataFrame
- Webscraped wikipedia with BeautifulSoup
- Filtered to include only Falcon 9 and filled in missing values

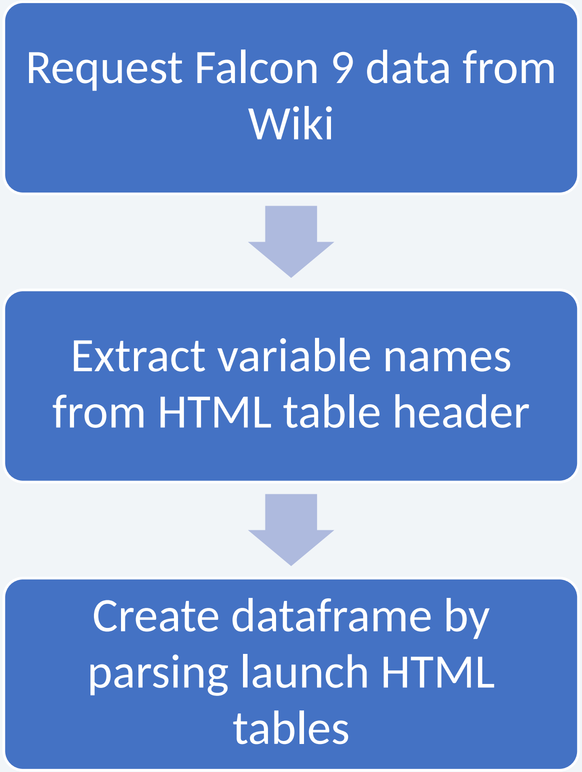
Data Collection – SpaceX API

- Request and parse data from SpaceX REST with GET request
- Remove Falcon 1 data
- Missing values in PayloadMass filled in with mean value
- https://github.com/joeastro5/IBM_Capstone/blob/main/data-collection-api.ipynb



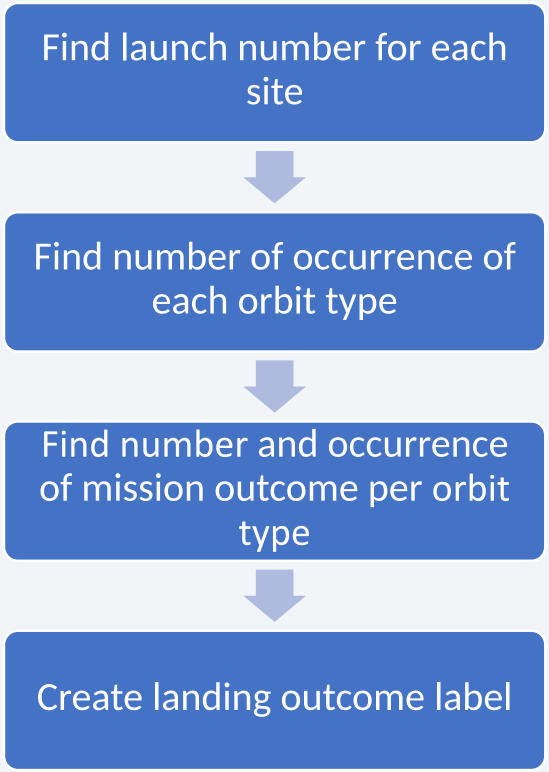
Data Collection - Scraping

- Webscraped to collect historical data from Wikipedia, with BeautifulSoup
- Parse table and convert to pandas
- https://github.com/joeastro5/IBM_Capstone/blob/main/data-collection-scraping.ipynb



Data Wrangling

- There are several launch sites and orbit types
- Converted outcomes to training labels from various strings to 0 and 1
- https://github.com/joeastro5/IBM_Capstone/blob/main/data-collection-wrangling.ipynb



EDA with Data Visualization

- Used Seaborn catplots of Flightnumber, Payload, Orbit type, and Flight number to determine relative importance of variables
- CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E have a success rate of 77%
- Also ranked orbits in terms of success rate with a bar chart
- https://github.com/joeastro5/IBM_Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- Various SQL queries, including:
  - Find total payload mass carried by NASA boosters
  - AVG mass carried by F9
  - Date when first successful landing in ground pad
  - Total number of successes
  - Names of booster versions which carried max payload mass
  - Number of successes between 04/06/2010-20/03/2017
- https://github.com/joeastro5/IBM_Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- How does success depend on location and proximity of launch sites?
- Folium Map: Circles are launch sites and other significant sites, red/green marker clusters are groups of events (success/failures), lines show distances between coordinates
- https://github.com/joeastro5/IBM_Capstone/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb (May need https://nbviewer.org/ to view maps)

Build a Dashboard with Plotly Dash

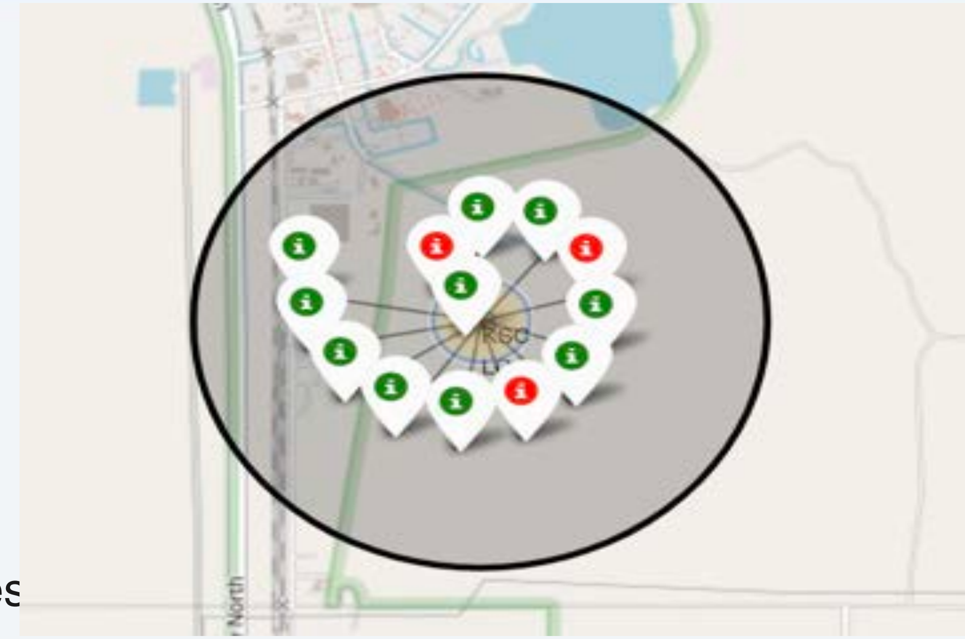
- Included in dashboard:
  - Pie charts for success count for all sites, and for success/failure of individual sites.
  - Success/Fail status for each booster, where the user can select range of payload mass
- Interactivity allows user to look more closely at significant payload mass ranges
- https://github.com/joeastro5/IBM_Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- First, data processed and split into test and training set

- Model selection:
  - SVM, Classification Tree, KNN, and Logistic Regression (sklearn)
  - Optimal hyperparameters with GridSearchCV
  - Evaluated with the score method and confusion matrices
- https://github.com/joeastro5/IBM_Capstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

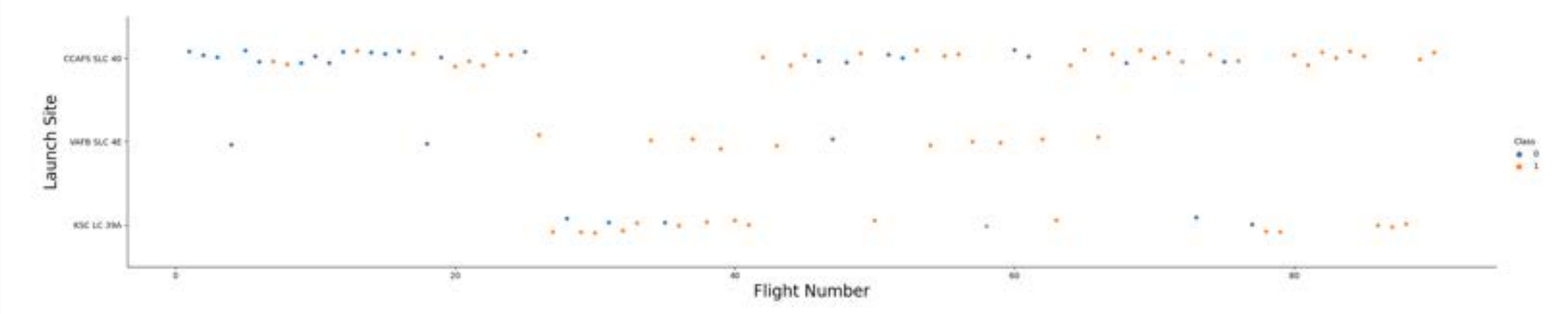


- EDA results from plotly dashboard:
  - KSC LC-39A has highest percentage of successes
  - Most successes in range 2000-6000 kg; very low at higher masses.
  - The booster with highest success rate was the FT.
- Predictive analysis results: Decision tree performs best on test data.

Section 2

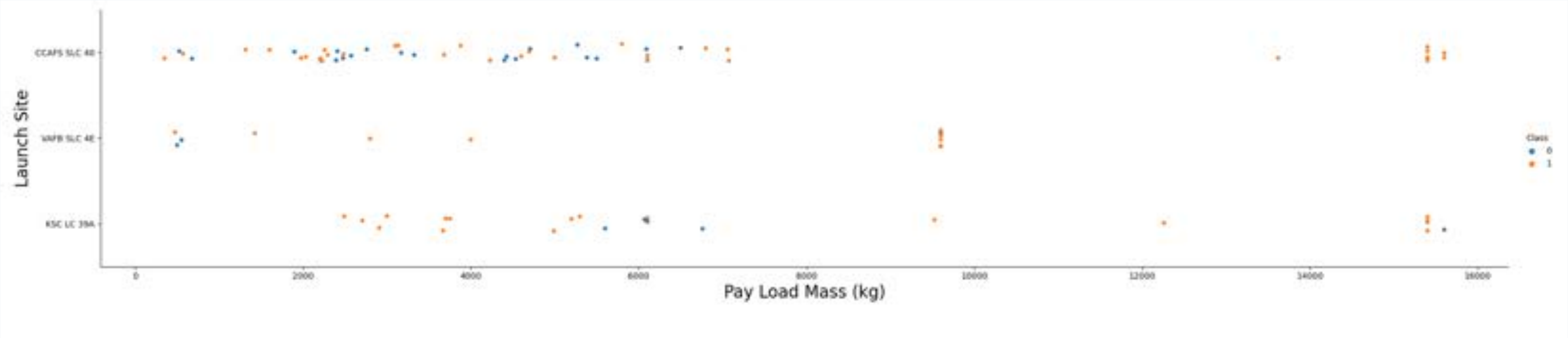
Insights drawn from EDA

Flight Number vs. Launch Site



- There are relatively few at VAFB SLC 4E
- Type KSC LC 39A has a higher success rate

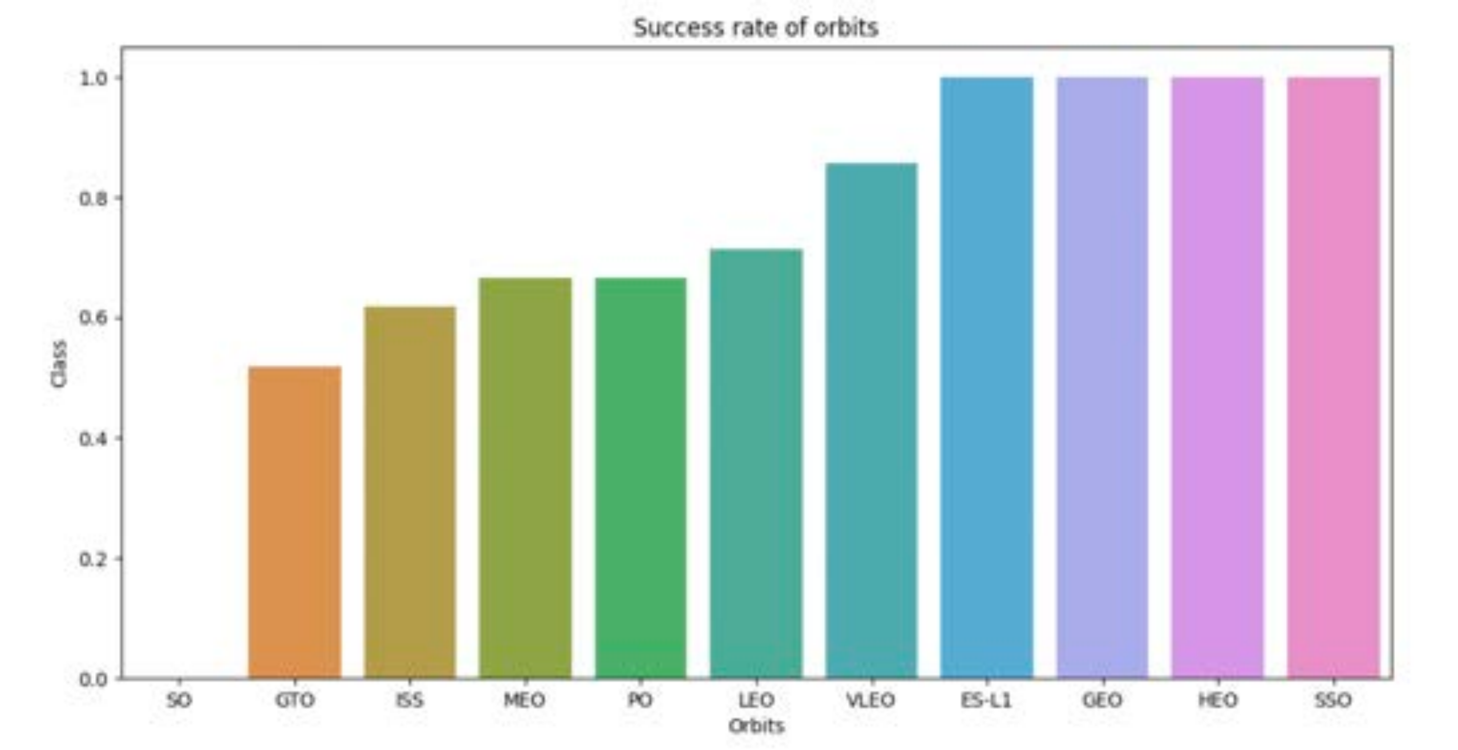
Payload vs. Launch Site



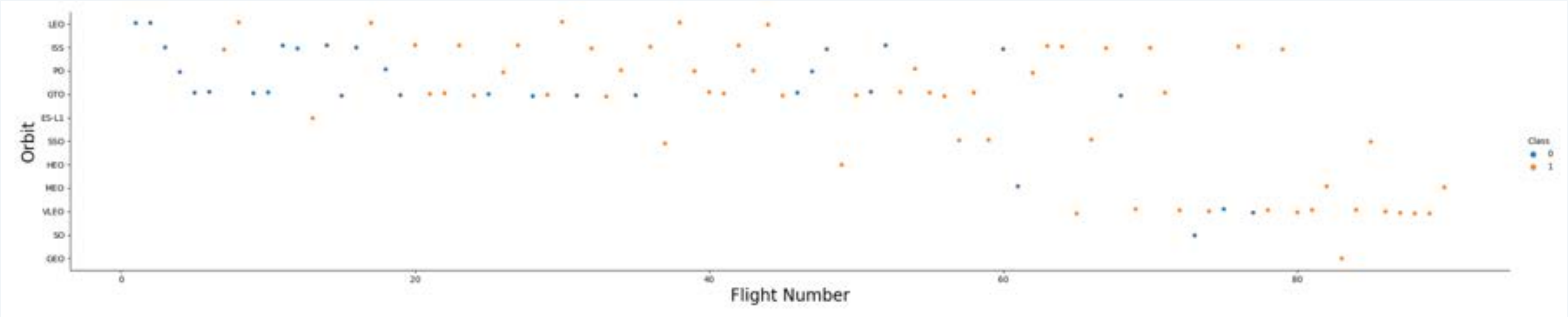
- Relatively few with very high mass, but still some successes
- Site VAFB SLC 4E does not have high mass payloads

Success Rate vs. Orbit Type

- All have at least a 50% success rate
- The four best orbit types have near 100%

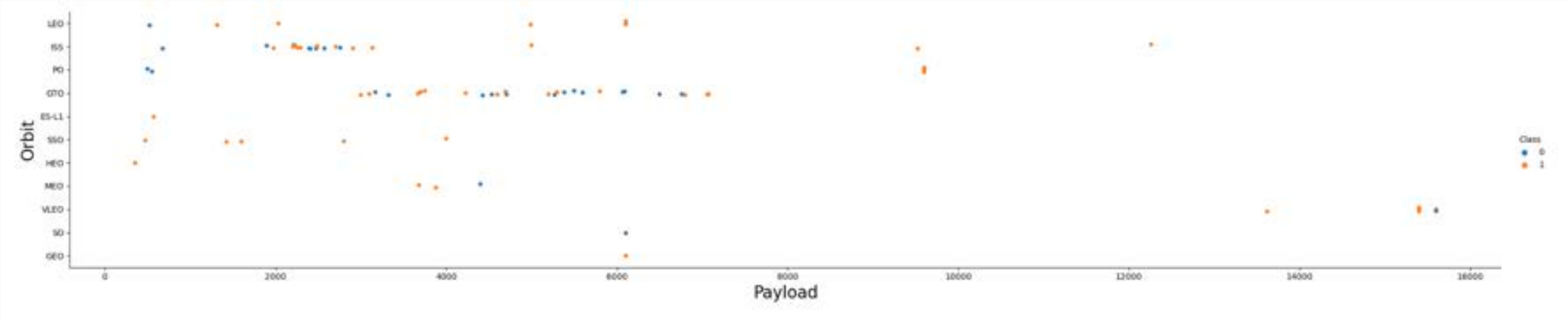


Flight Number vs. Orbit Type



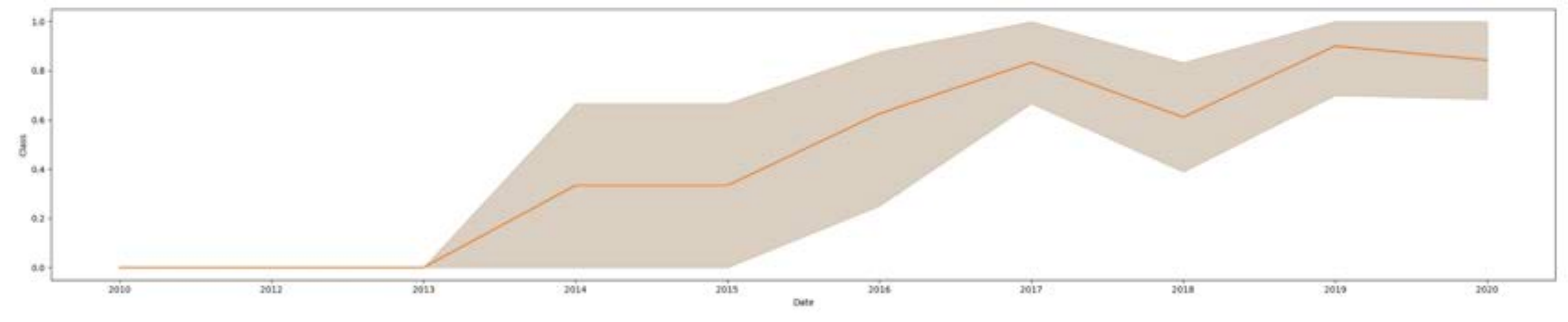
- Flights more successful over time
- Recent emphasis on orbits GEO, SO, VLEO, MEO, HEO,SSO

Payload vs. Orbit Type



- Payloads above 8000 kg not attempted often
- GTO and ISS orbits have most attempts

Launch Success Yearly Trend



- Improvement over time, which appears to have plateaued
- Narrower spread since 2015

All Launch Site Names

```
[12]: %sql SELECT DISTINCT Launch_Site FROM SPACEXTBL

 * sqlite:///my_data1.db
Done.
```

[12]:

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

- There are only four unique launch sites

Launch Site Names Begin with 'CCA'

```
[95]: %sql SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5

 * sqlite:///my_data1.db
Done.
```

[95]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- The top five, with very different payload masses
- Same orbit, LEO

Total Payload Mass

```
[26]: %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer LIKE 'NASA (CRS)'

 * sqlite:///my_data1.db
Done.

[26]: SUM(PAYLOAD_MASS__KG_)
                     45596
```

- Total payload mass from NASA

Average Payload Mass by F9 v1.1

```
[15]: %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version LIKE 'F9 v1.1'
 * sqlite:///my_data1.db
Done.
```

[15]:

AVG(PAYLOAD_MASS__KG_)
2928.4

- Avg payload mass for F9 v1.1

First Successful Ground Landing Date

```
[93]: %sql SELECT Date,MIN(substr(Date,7,4)) FROM SPACEXTBL WHERE "Landing _Outcome" LIKE '%Success%'

 * sqlite:///my_data1.db
Done.
```

[93]:

Date	MIN(substr(Date,7,4))
22-12-2015	2015

- The first success was on Dec 22, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

- The boosters which have successfully landed on drone ship and had payload mass between 4000 and 6000

Total Number of Successful and Failure Mission Outcomes

COUNT(Mission_Outcome)
100

100 Successes

COUNT(Mission_Outcome)
1

Only one failure

Boosters Carried Maximum Payload

- These boosters all carried the maximum payload that was attempted
- A subquery was used to select this subset

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

- The failed landing outcomes, and their booster versions, and launch site names in 2015
- All but one are successes
- All from the same Launch site

Month	Mission_Outcome	Booster_Version	Launch_Site
01	Success	F9 v1.1 B1012	CCAFS LC-40
02	Success	F9 v1.1 B1013	CCAFS LC-40
03	Success	F9 v1.1 B1014	CCAFS LC-40
04	Success	F9 v1.1 B1015	CCAFS LC-40
04	Success	F9 v1.1 B1016	CCAFS LC-40
06	Failure (in flight)	F9 v1.1 B1018	CCAFS LC-40
12	Success	F9 FT B1019	CCAFS LC-40

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- The count of landing outcomes between date 2010-06-04 and 2017-03-20, in descending order

Landing _Outcome	count
Success	20
Success (drone ship)	8
Success (ground pad)	6

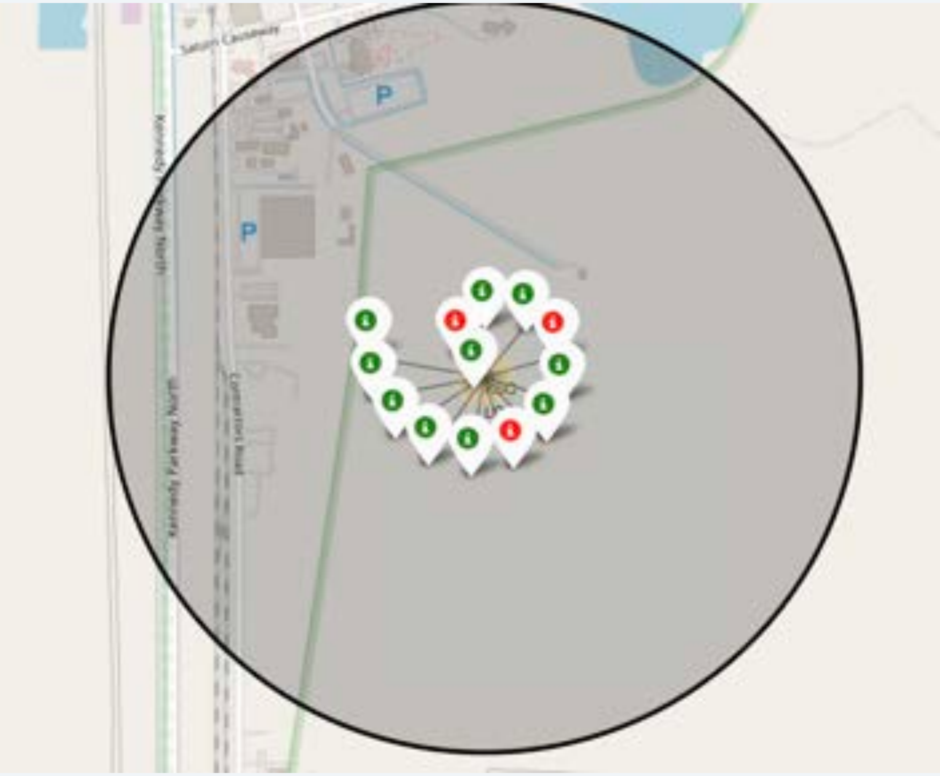
Section 3

Launch Sites Proximities Analysis

Launch Site Locations

- The generated folium map with all launch sites' location markers on a global map
- All in the US near the coasts

Example of Outcomes Near Launch Site



- Folium map with the color-labeled launch outcomes on the map
- Mostly successful launches at site KSC LC-39A

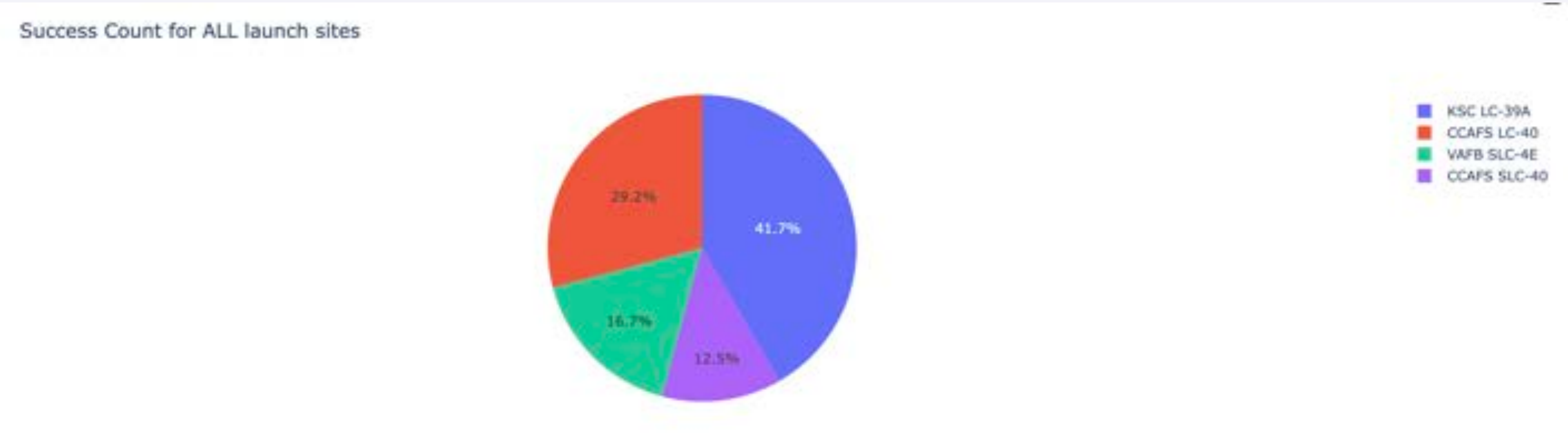
Distances from geographic features

- Map with selected launch site to any nearby locations, with distance calculated and displayed
- Here we find distance to coast (0.92 km)

Section 4

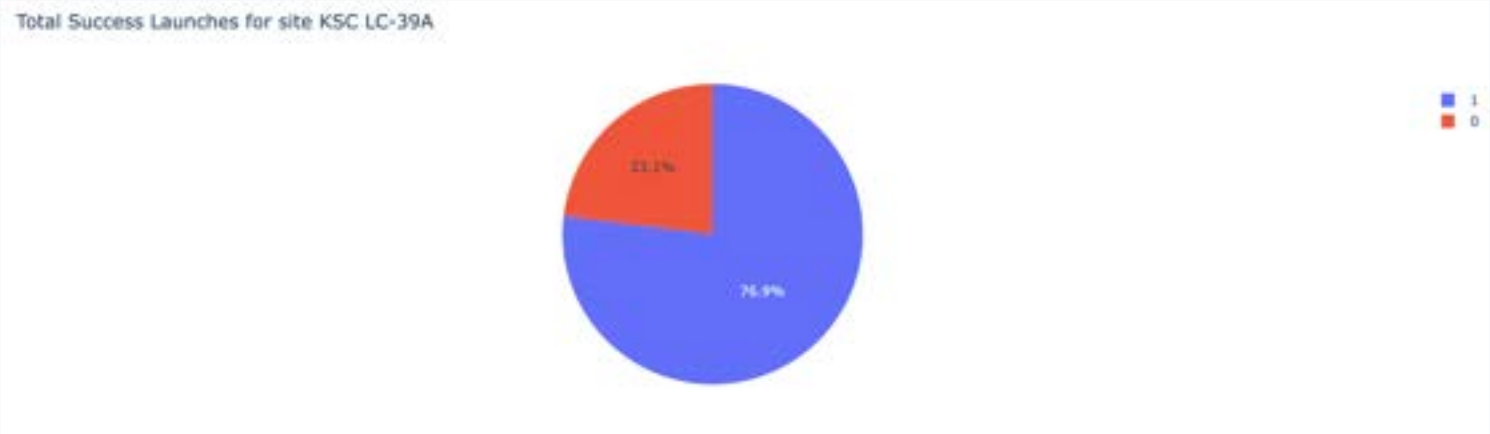
Build a Dashboard with Plotly Dash

Launch Success Rate for All Sites



- Site KSC LC-39A has the highest success rate
- Site CCAFS SLC-40 has the worst

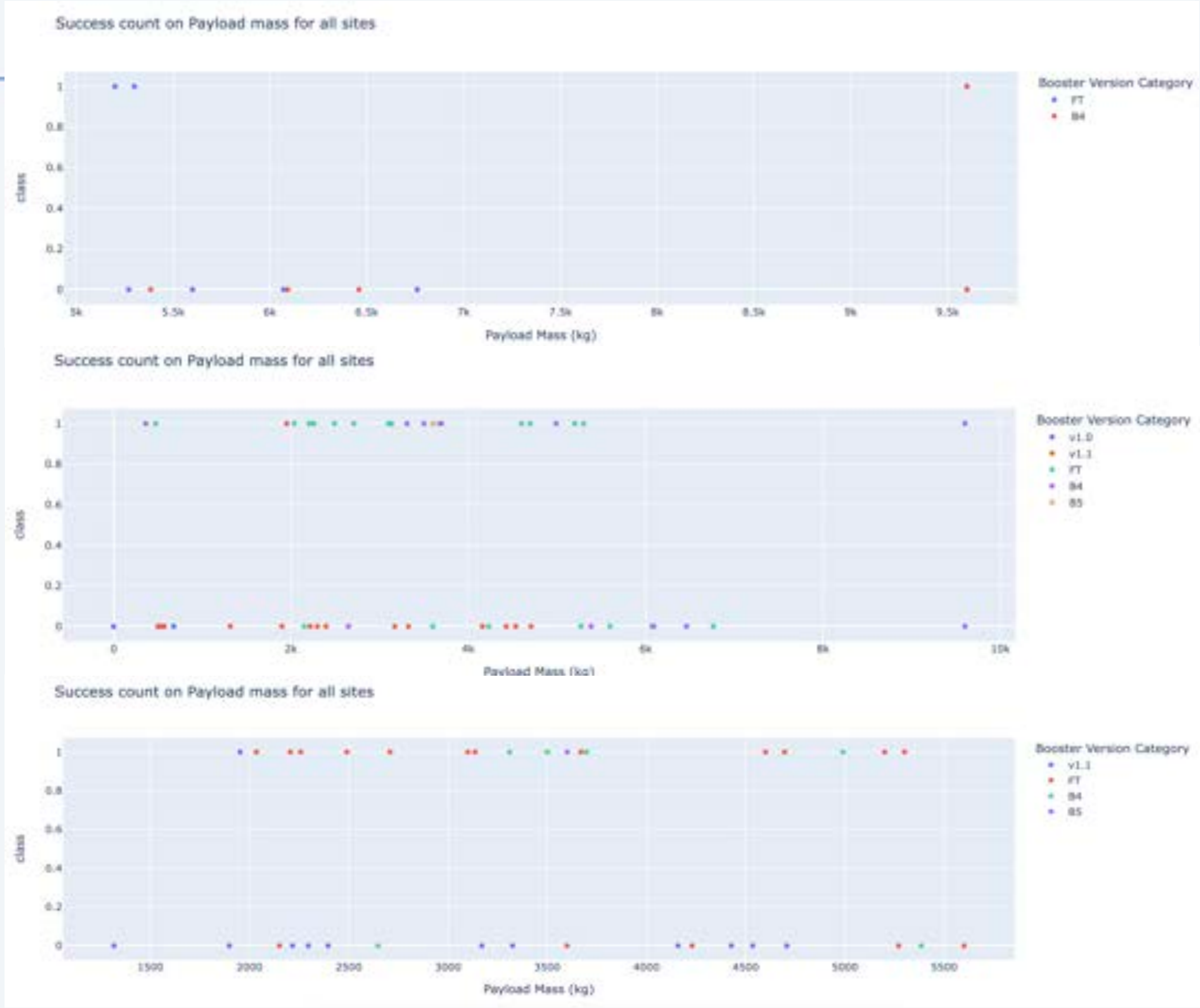
Success and failure of most successful site



- Launch site with highest launch success ratio, KSC LC-39A
- More than 75% success rate

Payload vs Launch Outcome (All sites)

- User can select range of payload masses
- Most successful launches in range 2000-6000kg
- FT boosters in this range perform best



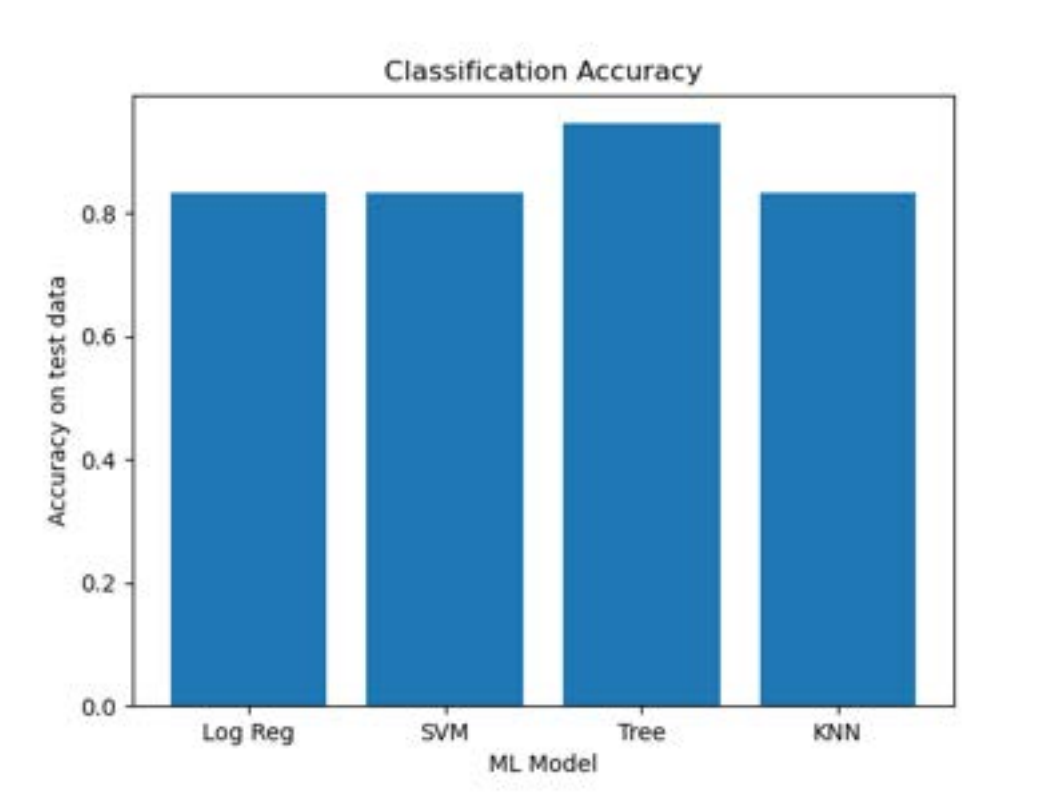
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- All attempted classification models fairly successful on test data

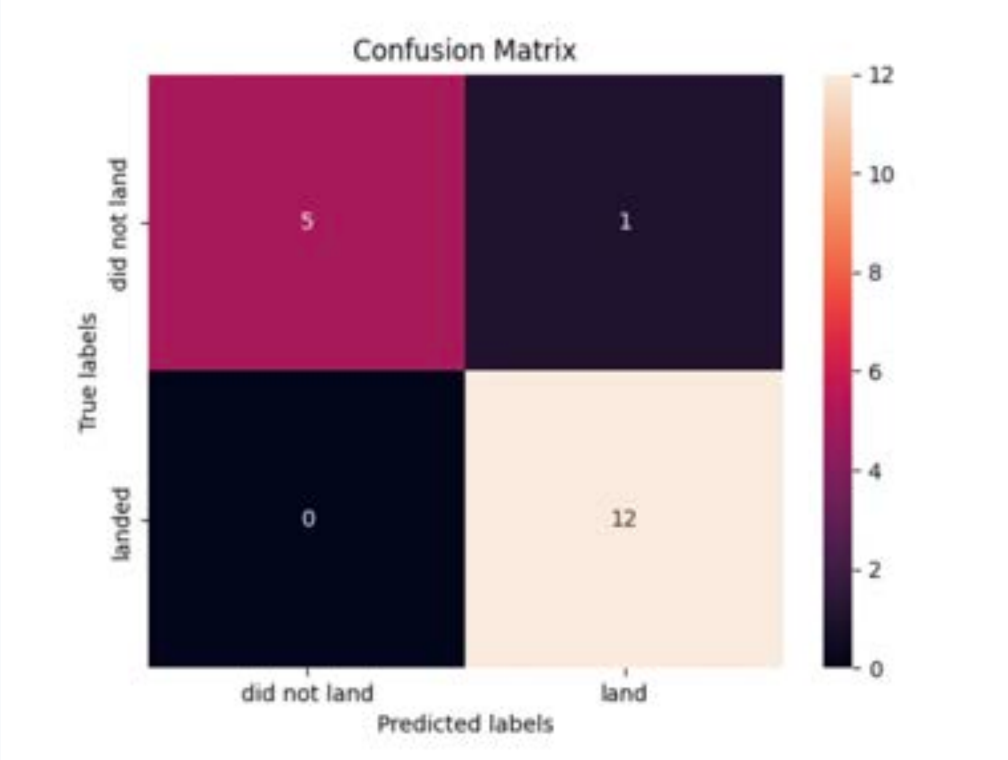
- Decision tree has the highest classification accuracy



Confusion Matrix

- The confusion matrix of the best performing model (decision tree)

- Only one misclassified



Conclusions

- Best launch site: KSC LC-39A
- Best payload range: 2000-6000 kg
- Launches have been improving with time
- Decision tree classifier ideal for prediction

Thank you!