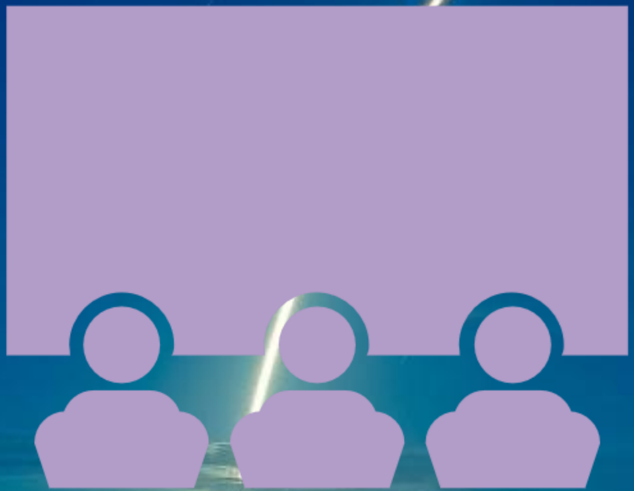
Data Science Capstone

Lingxin (Alan) Ma

August 29, 2021

Outline



- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- Summary of methodologies:
  - Data collection
  - Data wrangling
  - Exploratory data analysis
  - Interactive visual analytics
  - Predictive analysis
- Summary of all results:
  - Tree classifier has the most accuracy on predicting the success rate of a launch
  - First Stage landing is most likely to success if launched in KSC LC-39A Launch site with smaller payload mass (<5500kg), FT, BF, or B5 booster versions, and on ES-L1, GEO, HEO, or SSO orbits.

Introduction

The cost of SpaceX's Falcon 9 rocket launches is advertised with a cost of 62 million dollars, which is less than half of the cost from other providers. It is all because of the reuse of the first stage. In this capstone project, we will predict if the Falcon 9 first stage will land successfully.

By collecting and analyzing SpaceX data, we could discover:

1. The geolocation of each launch sites, and what is special about these locations
2. What factors contribute to Falcon 9's landing outcomes
3. Predict whether Falcon 9 first stage will land successfully given certain conditions

Methodology

- Data collection methodology
- Perform data wrangling
- Exploratory data analysis (EDA) using visualization and SQL
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classification models

Methodology

Data collection

- How data was collected:

Spacexdata API → Filter dataframe to only include Falcon 9 Launches

Web scraping Launches Records from Wikipedia → Falcon 9 and Falcon Heavy Launches Records, with detailed booster version

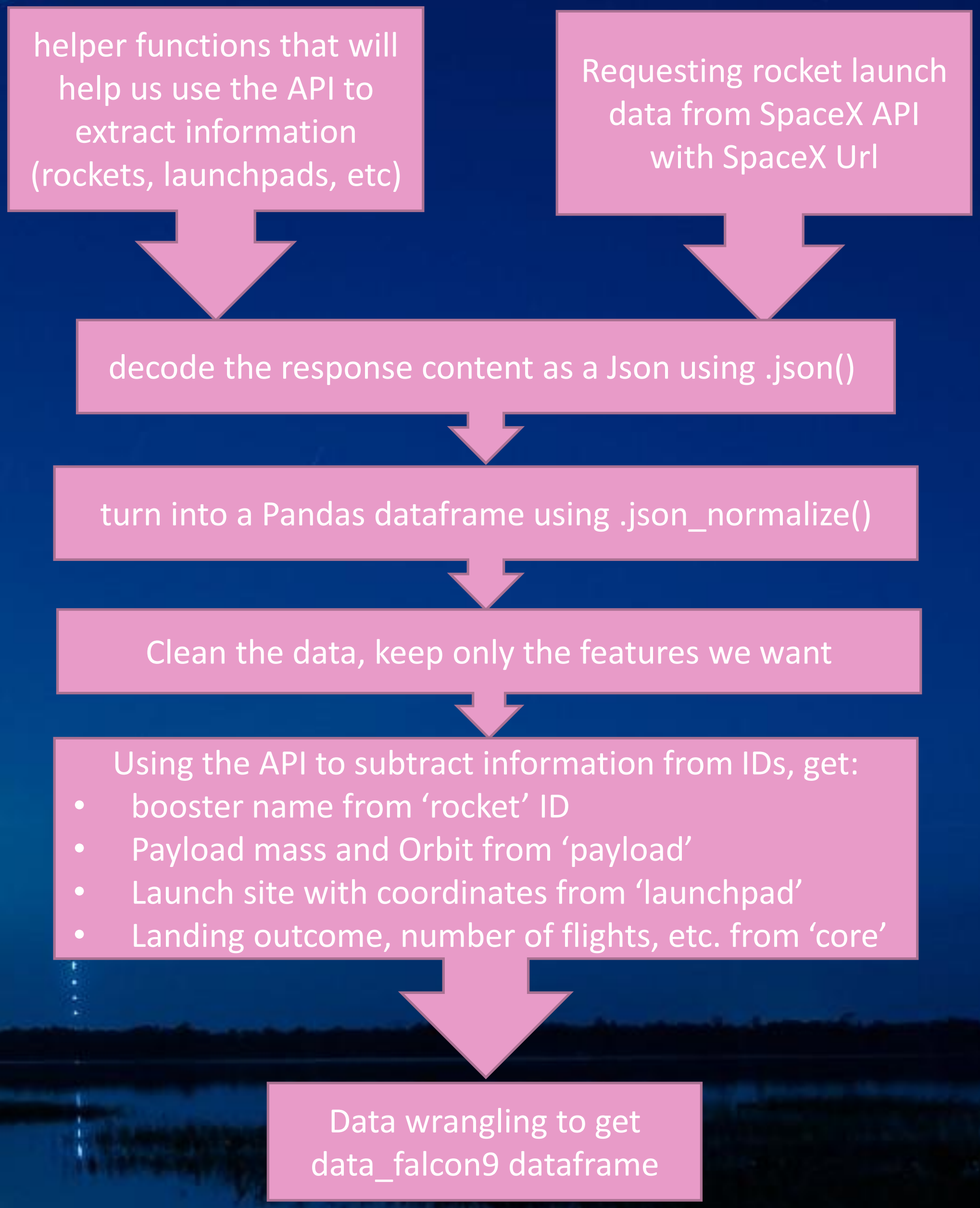
Data collection – SpaceX API

Data collection with SpaceX REST calls flowchart:

3 Launch Sites were recorded in the dataframe:

CSFS SLC 40, VAFB SLC 4E , and KSC LC 39A

Refer to Appendix for GitHub URL of the completed SpaceX API calls notebook



Data collection – Web scraping

Web scraping process flowchart:

Refer to Appendix for GitHub URL of the completed Webscraping notebook

Data wrangling

- Data wrangling process flowchart:

- Refer to Appendix for GitHub URL of the completed Data wrangling notebook

Import dataset

↓

Identify the missing values, numerical and categorical values

↓

Replace missing value of PayloadMass with mean value

↓

Create landing outcome labels from ‘Outome’ column

↓

Create data frame

EDA with data visualization

Charts Plotted:

1. Scatter plot of Flight Number vs. Launch Site
   - Indicate how the FlightNumber (indicating the continuous launch attempts.) and Payload variables would affect the launch outcome.
2. Scatter plot of Payload vs. Launch Site
   - Visualize the relationship between Payload and Launch Site
3. Bar chart of Success rate vs. Orbit type
   - Visualize the relationship between success rate of each orbit type
4. Scatter plot of Flight Number vs. Orbit type
   - Visualize the relationship between FlightNumber and Orbit type
5. Scatter plot of Payload vs. Orbit type
   - Visualize the relationship between Payload and Orbit type
6. Line chart of Launch success yearly trend
   - Visualize the launch success yearly trend

Refer to Appendix for GitHub URL of the completed EDA with data visualization notebook

EDA with SQL

Performed SQL queries:

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first succesful landing outcome in ground pad was achieved.
- List the name of boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
- 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
- Rank the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

Refer to Appendix for GitHub URL of the completed EDA with data visualization notebook

Build an interactive map with Folium

Added features to map:

- folium.Circle to add a highlighted circle area with a text label on each launch site’s coordinate
- MarkerCluster() object to simplify the map containing many markers having the same coordinate
- Color based marker to mark the success/failed launches for each site
- Add MousePosition() to get coordinate for a mouse over a point on map
- Plot distance lines from each site to the proximities (Railway, Highway, Coastline, and Nearest City)

Refer to Appendix for GitHub URL of the completed interactive map with Folium map notebook

Build a Dashboard with Plotly Dash

Features and plots added to the Plotly Dash:

- Dropdown to choose between All sites
- Pie Charts of success counts for all launch sites
- If a specific launch site was selected, show the Success vs. Failed counts for the site
- Scatter chart to show the correlation between payload and launch success
- Rangeslider to better visualize how success outcomes distributes among different payload mass for each launch site

Refer to Appendix for GitHub URL of the completed Plotly Dash lab python script

Predictive analysis (Classification)

Standardize the data X, and let Y be the 'Class', with 1 being 'Success', and 0 being 'Failure'

After train test split, with test_size=0.2, there are only 18 test samples

Choose 4 models for classifications

- Logistic Regression (LogReg)
- Support Vector Machine (SVM)
- Tree classifier
- K Nearest Neighbors (knn)

Use GridSearchCV to find the best parameters for each model

Calculate the accuracy for each model with best parameters on training sets

Calculate the accuracy on the test data using the method score

Plot confusion matrix to visualize how each model can distinguish between the different classes

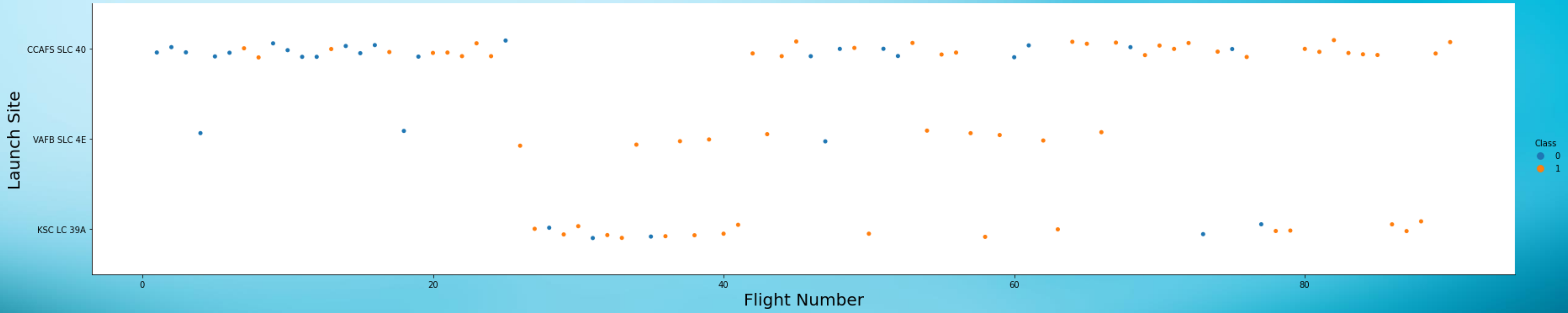
Refer to Appendix for GitHub URL of the completed predictive analysis lab notebook

Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

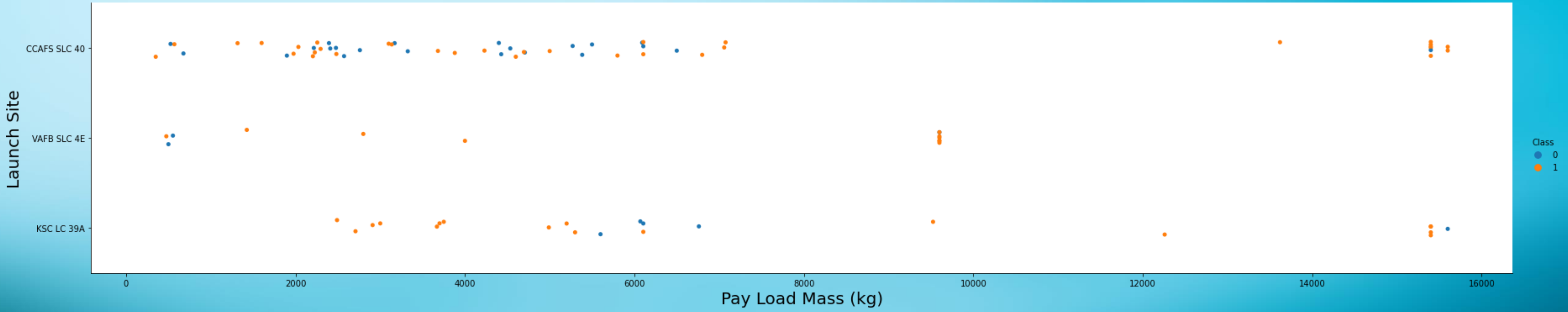
EDA with Visualization

Flight Number vs. Launch Site



x axis to be Flight Number and y axis to be the launch site, and hue to be the class value

Payload vs. Launch Site



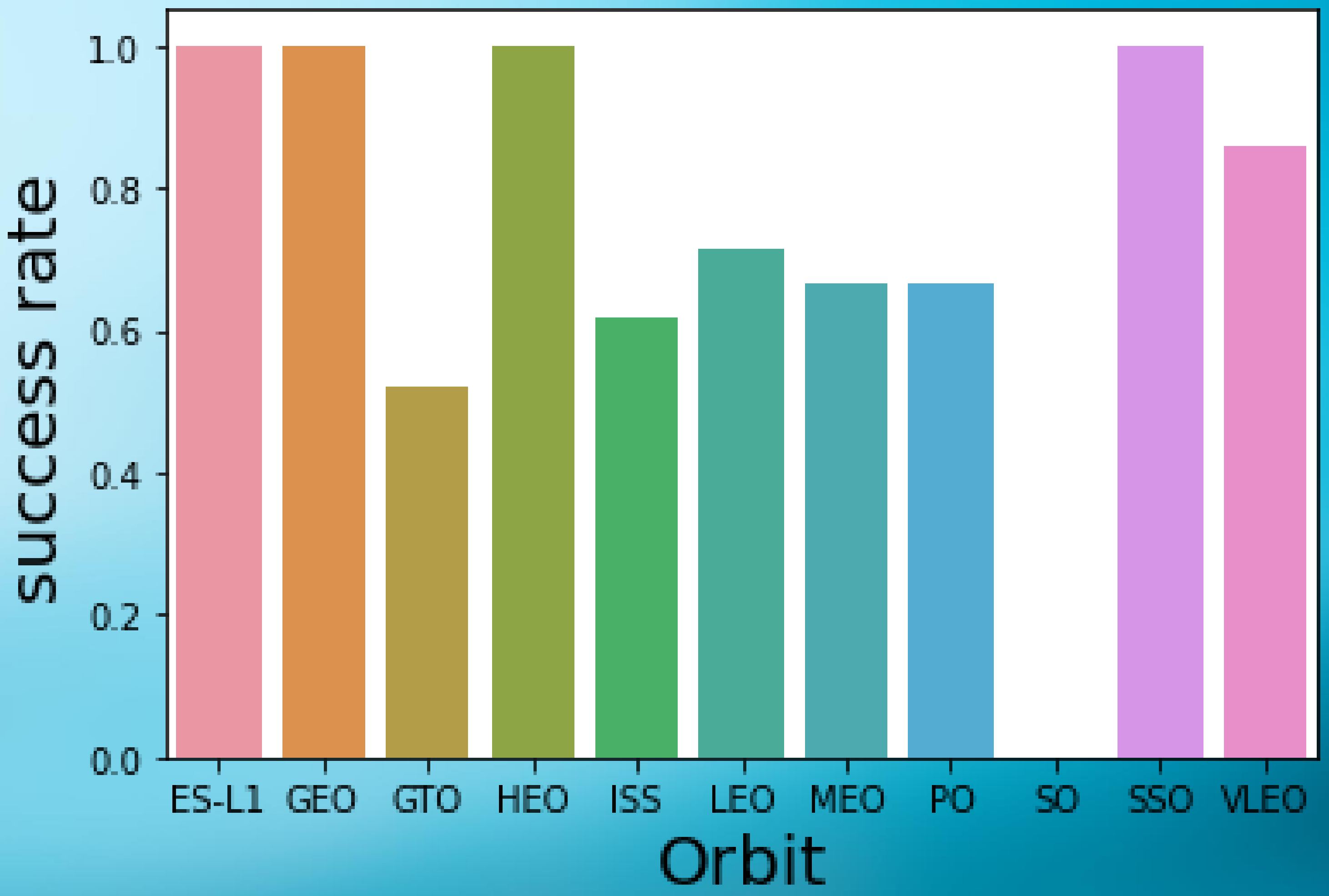
x axis to be Flight Number and y axis to be the launch site, and hue to be the class value

Success rate vs. Orbit type

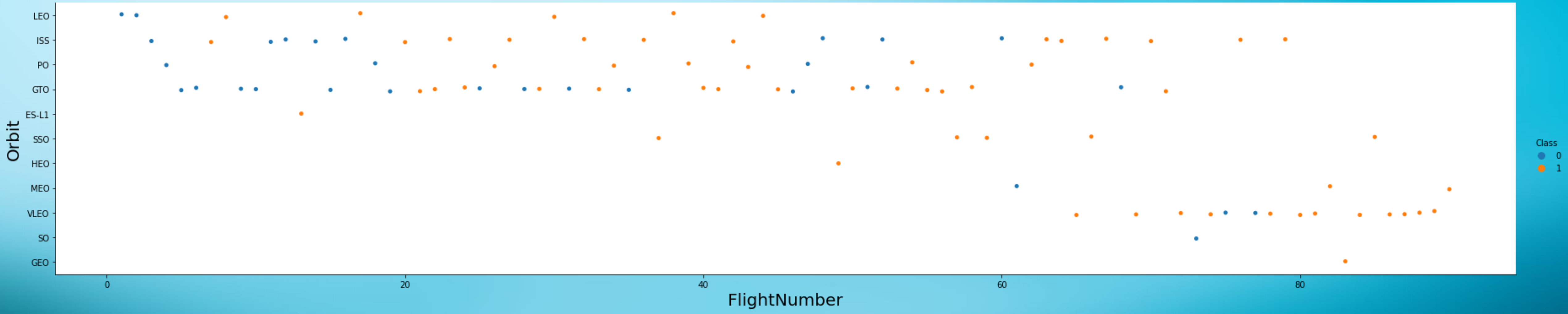
Barchart for the success rate of each orbit type

X axis to be the Orbit Type, y axis to be success rate

We notice that Falcon9 has never succussed on SO orbit

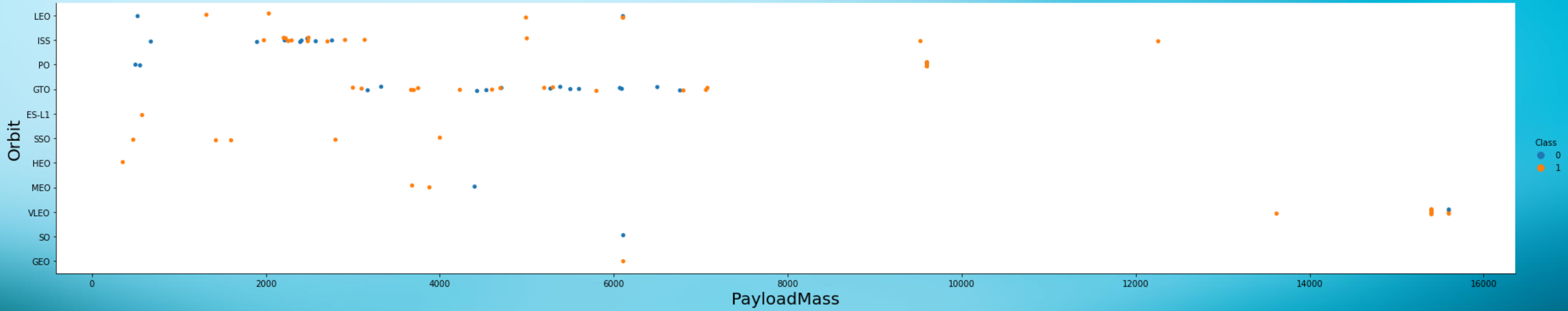


Flight Number vs. Orbit type



x axis to be Flight Number and y axis to be the Orbit, and hue to be the class value

Payload vs. Orbit type

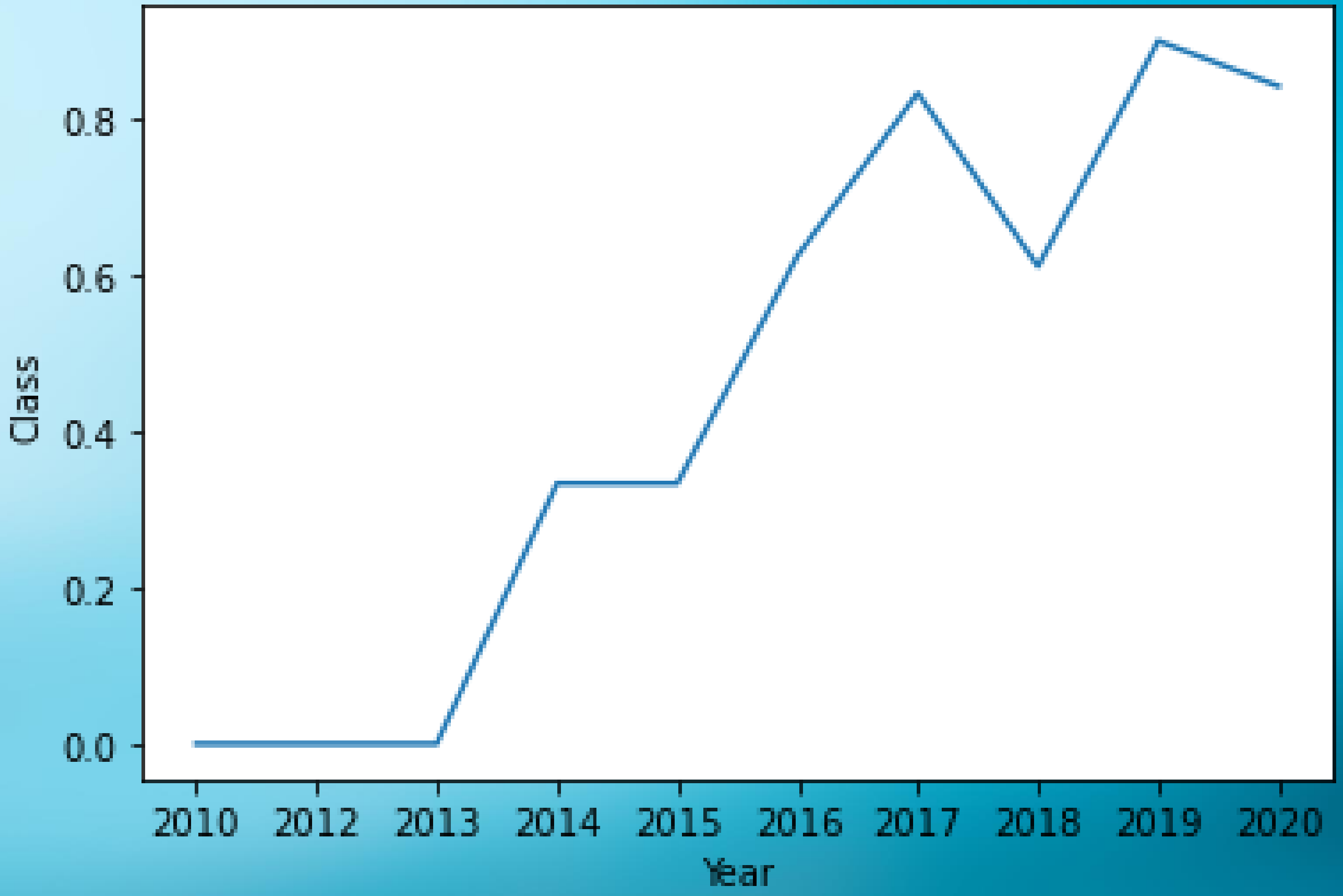


x axis to be Payload and y axis to be the Orbit, and hue to be the class value

Launch success yearly trend

Line chart of yearly average success rate

success rate has been increasing since 2013



EDA with SQL

All launch site names

Find the names of the unique launch sites:

```
SELECT Distinct LAUNCH_SITE FROM SPACEXTBL
```

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

Launch site names begin with `CCA`

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

SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	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)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total payload mass

Calculate the total payload carried by boosters from NASA (CRS)

```
SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER='NASA (CRS)'
```

45596

Average payload mass by F9 v1.1

Calculate the average payload mass carried by booster version F9 v1.1

```
SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE BOOSTER_VERSION='F9 v1.1’
```

2928

First successful ground landing date

The date when the first successful landing outcome in ground pad

```
SELECT min(DATE) FROM SPACEXTBL WHERE LANDING__OUTCOME='Success (ground pad)'
```

2015-12-22

Successful drone ship landing with payload between 4000 and 6000

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

```
SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ between 4000 and 600
0 AND LANDING__OUTCOME='Success (drone ship)’
```

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

Total number of successful and failure mission outcomes

Calculate the total number of successful and failure mission outcomes

```
SELECT COUNT(*) FROM SPACEXTBL WHERE MISSION_OUTCOME LIKE '%Success%' OR MISSION_OUTCOME LIKE '%Failure%'
```

101

Boosters carried maximum payload

List of booster which have carried the maximum payload mass

SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)

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

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

```
SELECT TO_CHAR(TO_DATE(MONTH("DATE"), 'MM'), 'MONTH') AS MONTH_NAME, LANDING__OUTCOME AS LANDING__OUTCOME, BOOSTER_VERSION AS BOOSTER_VERSION, LAUNCH_SITE AS LAUNCH_SITE FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND "DATE" LIKE '%2015%’
```

month_name	landing__outcome	booster_version	launch_site
JANUARY	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
APRIL	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Rank success count between 2010-06-04 and 2017-03-20

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

We can see that each date has at most 1 success

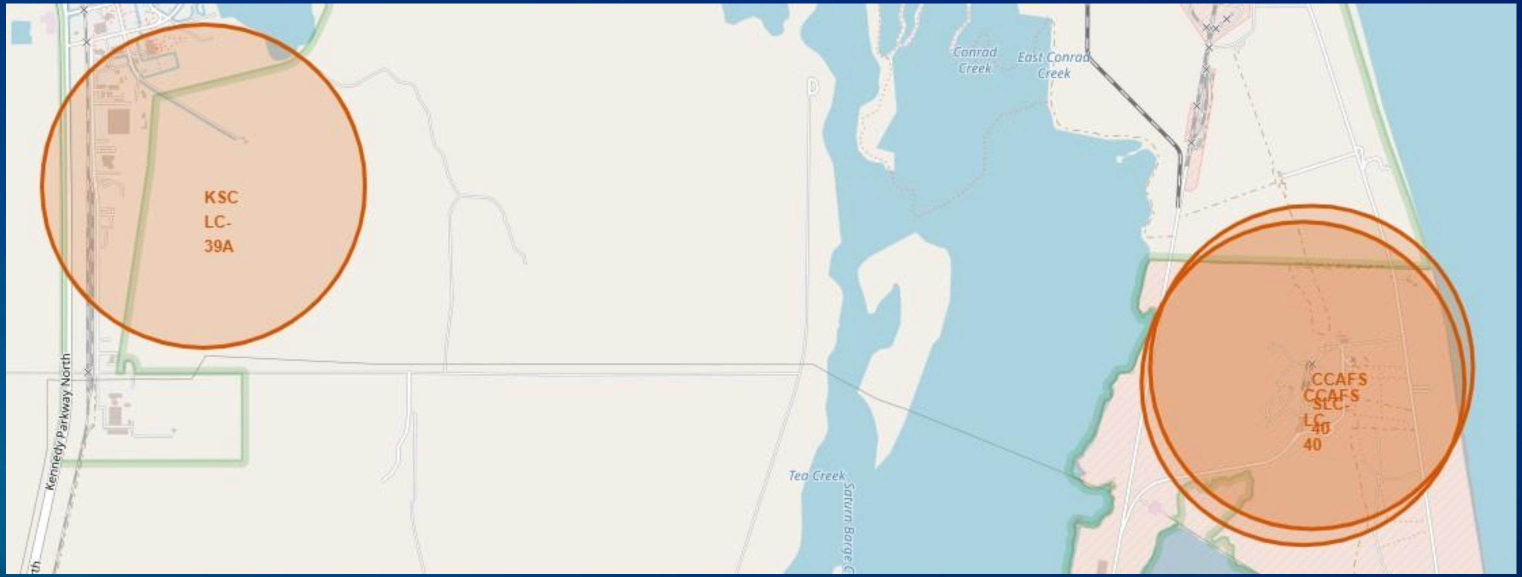
```
SELECT "DATE", COUNT(LANDING__OUTCOME) as COUNT FROM SPA
CEXTBL
    WHERE "DATE" BETWEEN '2010-06-04' and '2017-03-
20' AND LANDING__OUTCOME LIKE '%Success%'
    GROUP BY "DATE“
    ORDER BY COUNT(LANDING__OUTCOME) DESC
```

DATE	COUNT
2015-12-22	1
2016-04-08	1
2016-05-06	1
2016-05-27	1
2016-07-18	1
2016-08-14	1
2017-01-14	1
2017-02-19	1

Interactive map with Folium

All Launch Sites’ Location

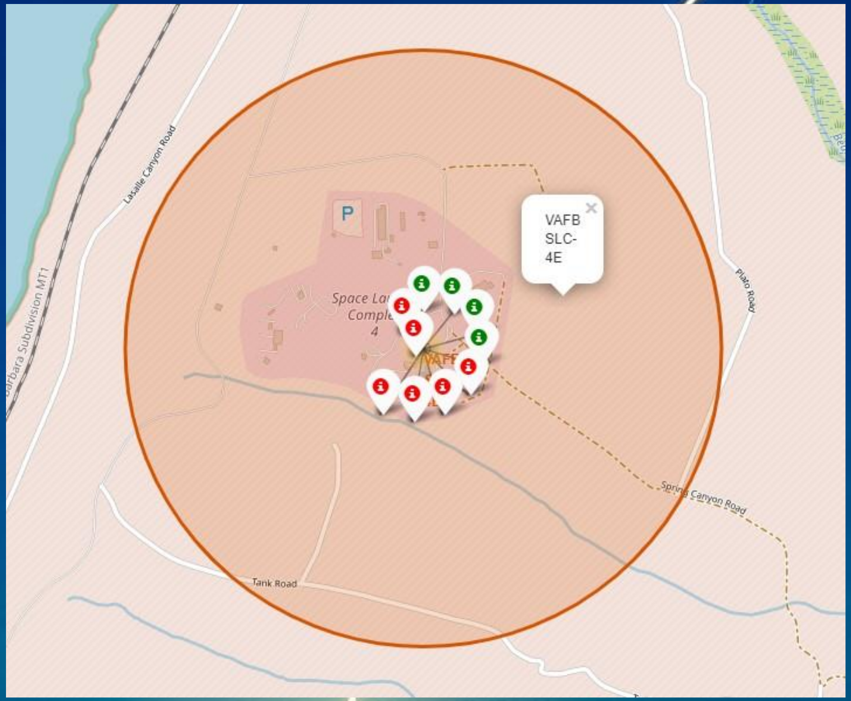
All Launch Sites’ Location (Cont.)

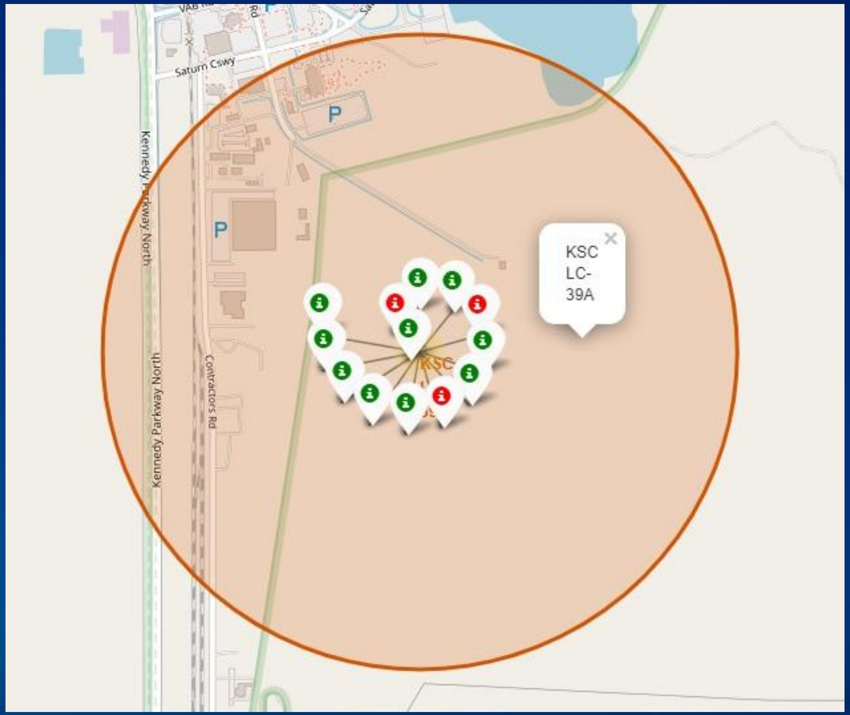


1 Launch Site in California

3 Launch Sites in Florida

Color-Labeled Launch Records

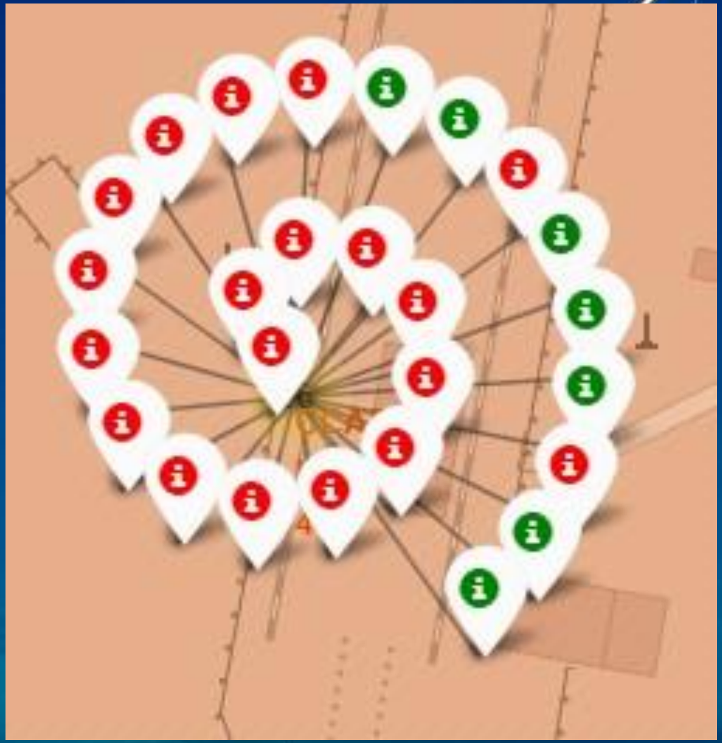




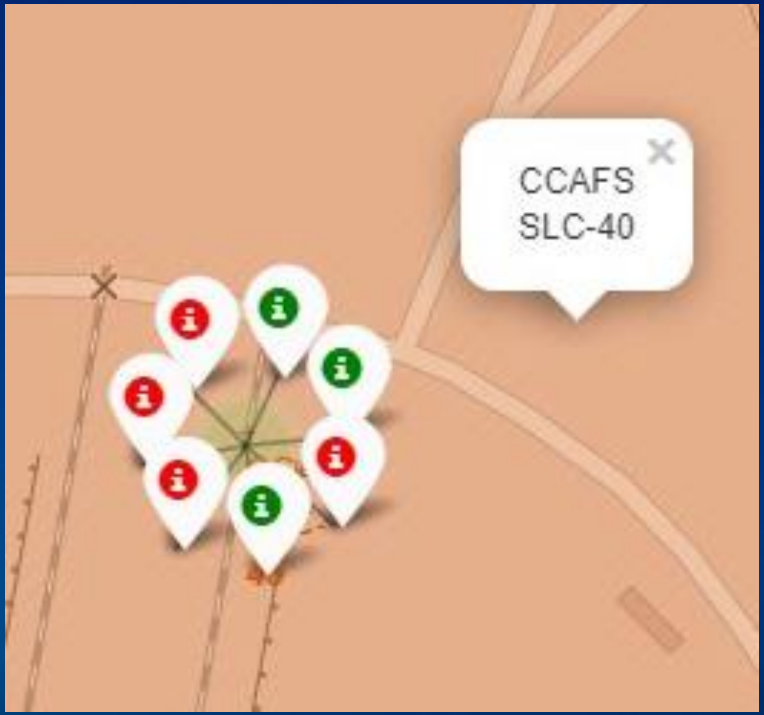
Total of 10 Launches in VAFB SLC-4E
With 4 Success landings

Total of 13 Launches in KSC LC-39A
With only 3 Failure landings

Color-Labeled Launch Records (Cont.)

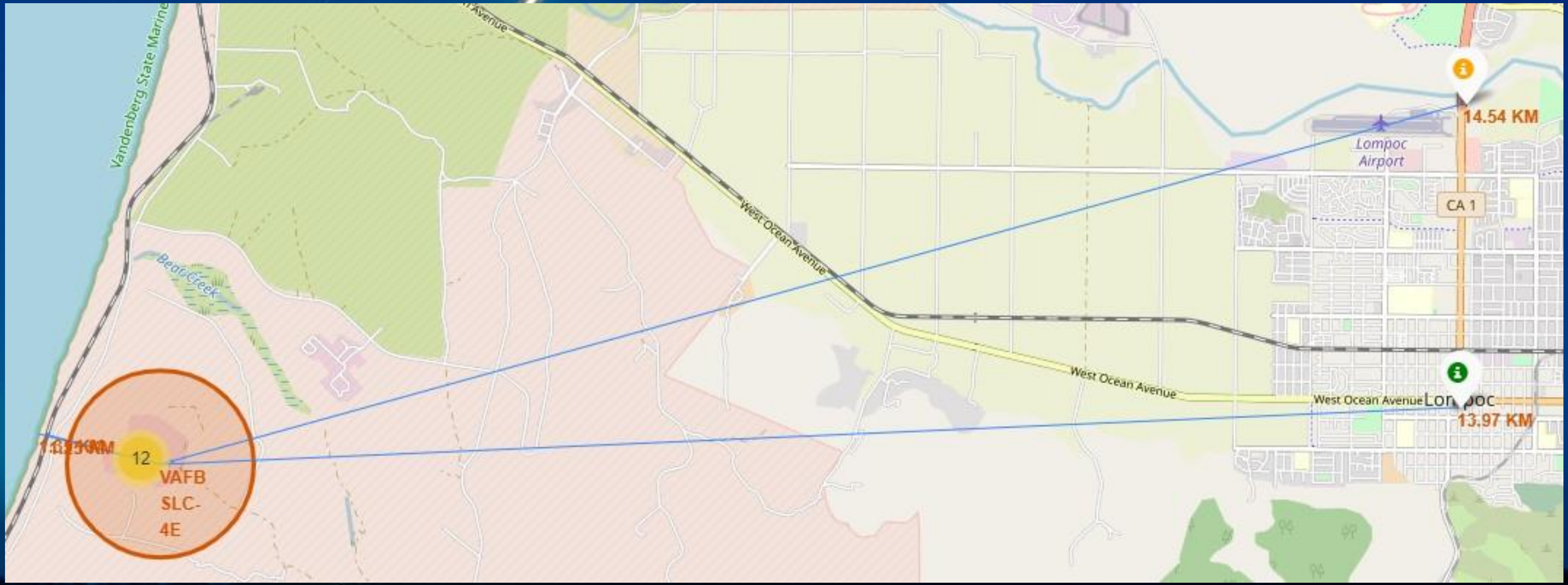


Total of 26 Launches in CCAFS LC-40
With only 7 Success landings



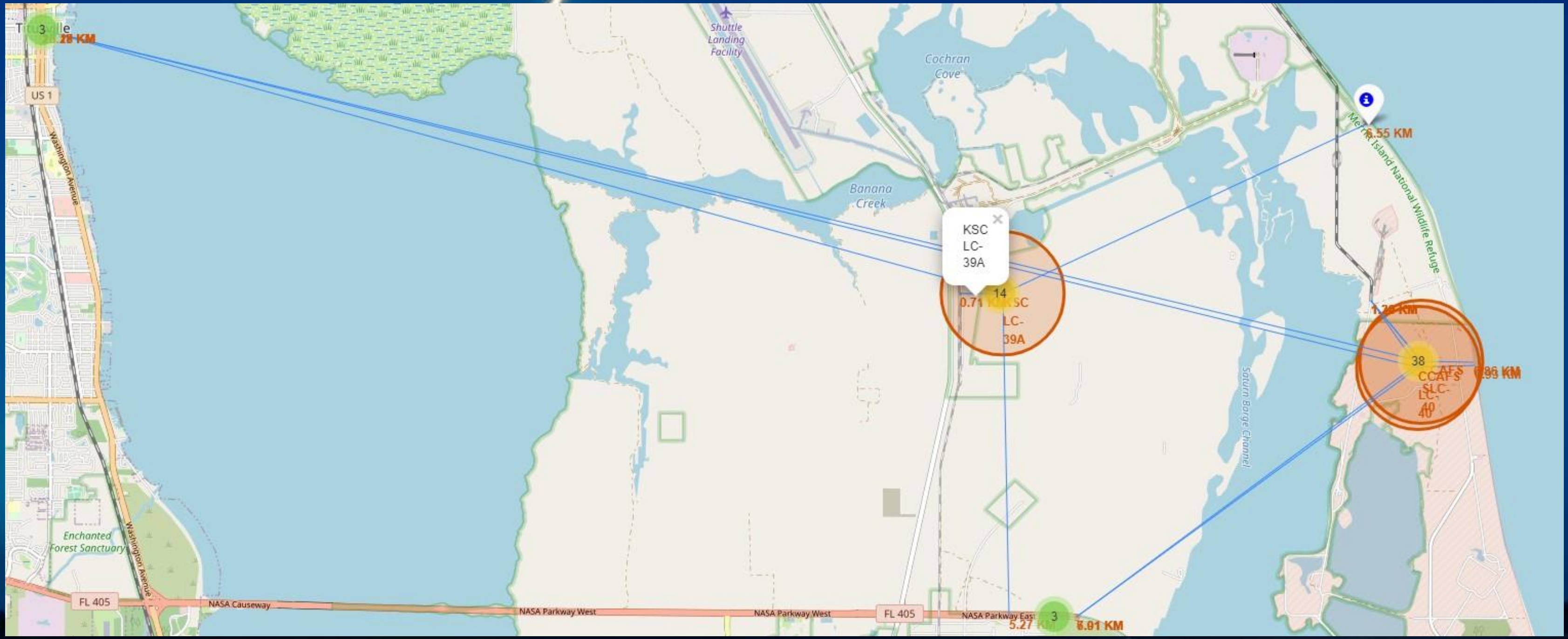
Total of 7 Launches in CCAFS SLC-40
With 3 Success landings

VAFB SLC-4E to its proximities Railway, Coastline, Highway, and City



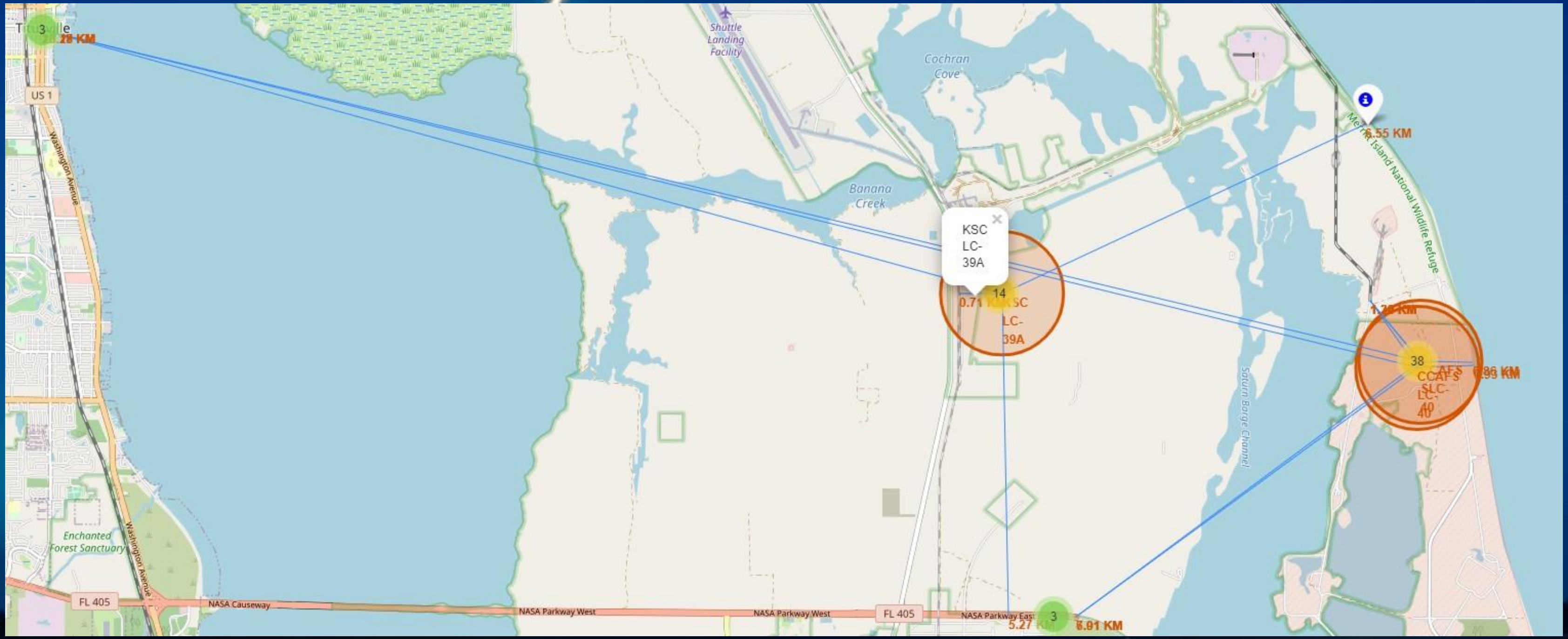
- 1.25 km to Railway
- 1.35 km to Coastline
- 14.54 km to highway
- 13.97 km to nearest city

KSC LC-39A to its proximities Railway, Coastline, Highway, and City



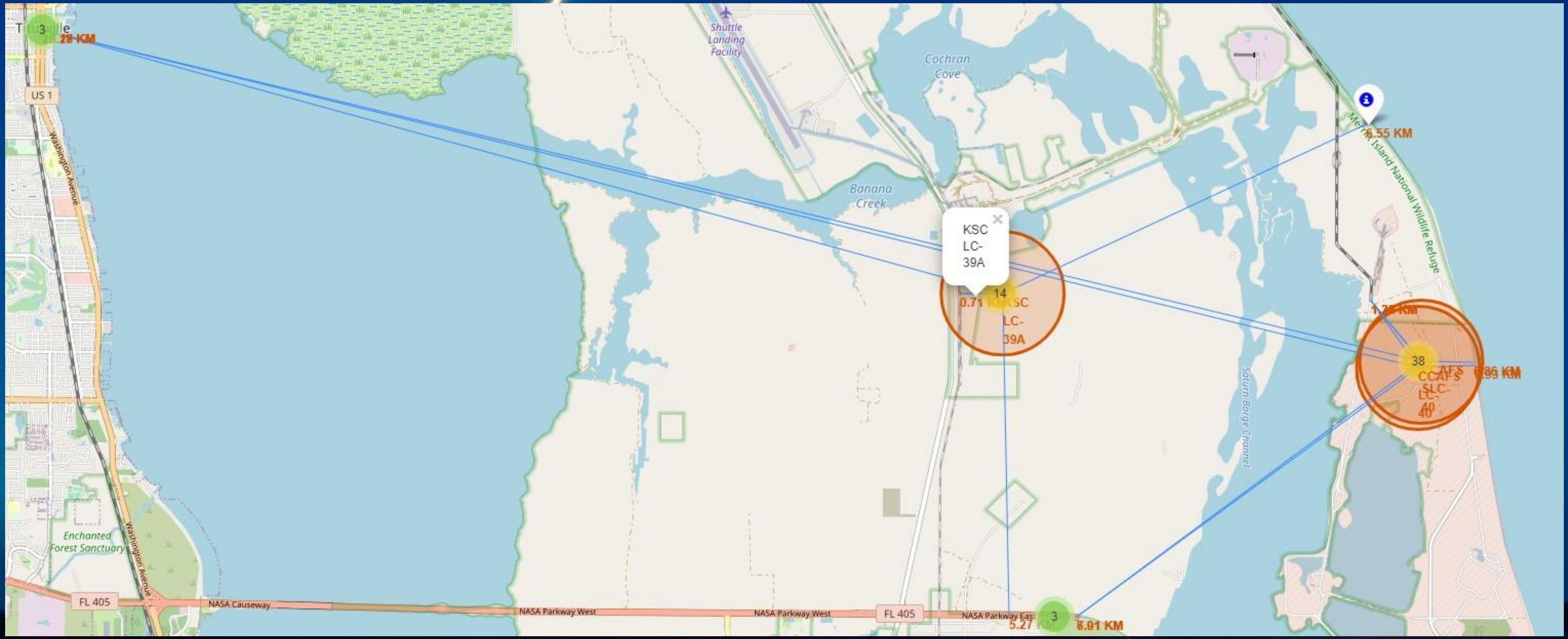
- 0.71 km to Railway
- 6.55 km to Coastline
- 5.27 km to highway
- 16.28 km to nearest city

CCAFS LC-40 to its proximities Railway, Coastline, Highway, and City



- 1.34 km to Railway
- 0.93 km to Coastline
- 6.91 km to highway
- 23.17 km to nearest city

CCAFS SLC-40 to its proximities Railway, Coastline, Highway, and City

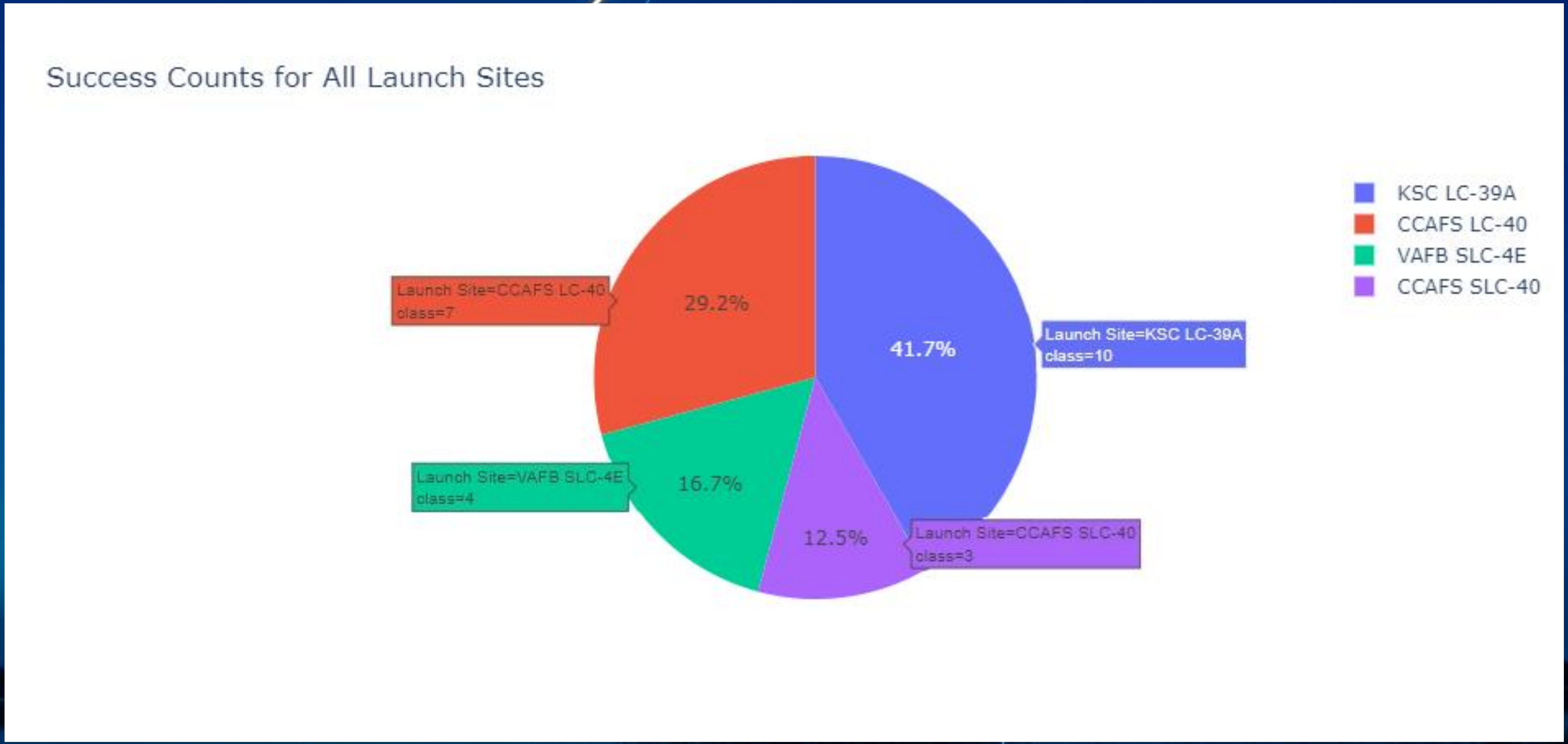


- 1.29 km to Railway
- 0.86 km to Coastline
- 7.01 km to highway
- 23.20 km to nearest city

Dashboard with Plotly Dash

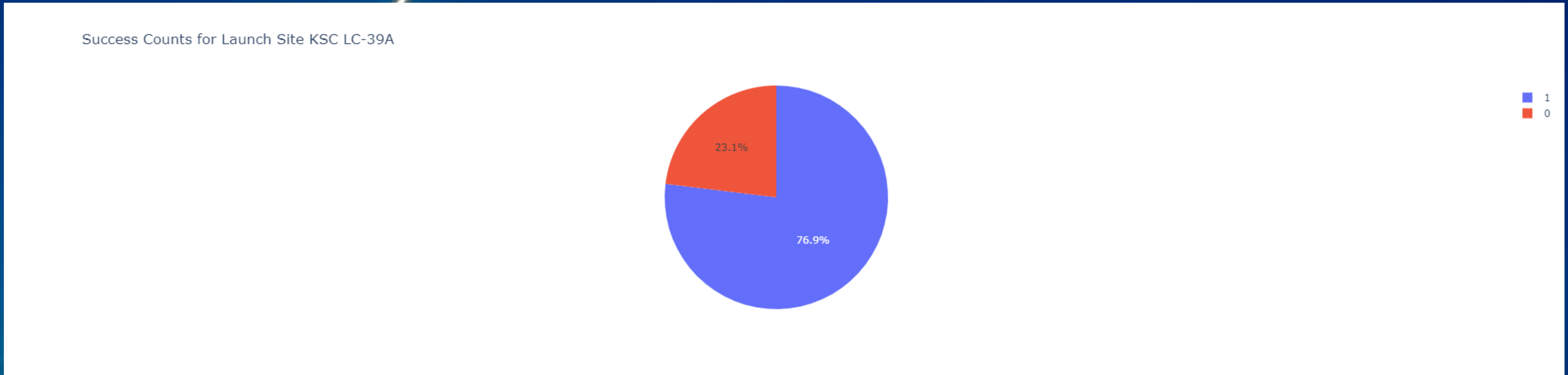
Using datasets that only consider payload mass from 0 – 10k [kg]

Launch Success for All Sites



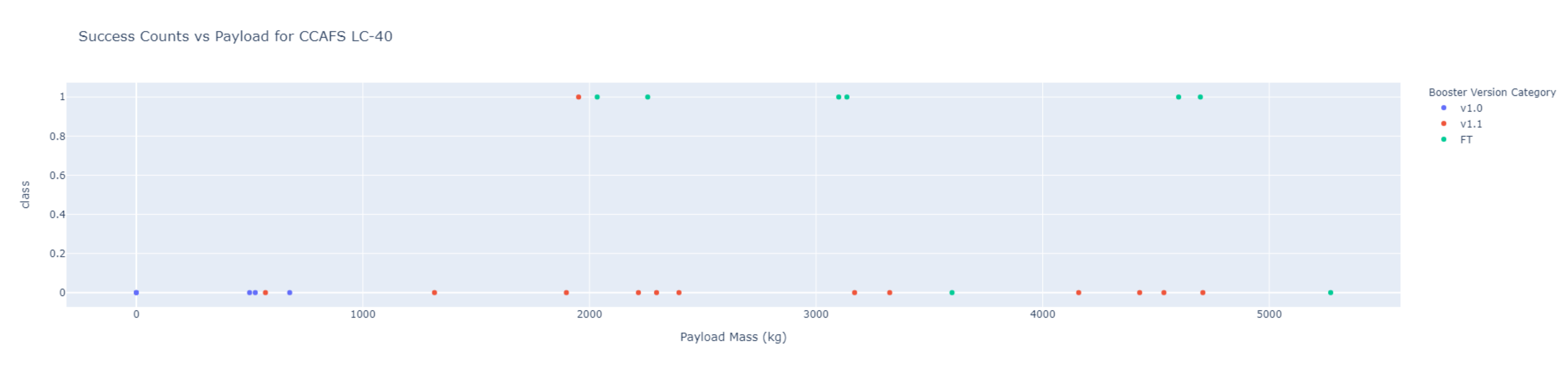
With KSC LC-39A having the most success

Launch Site with Highest Success Ratio



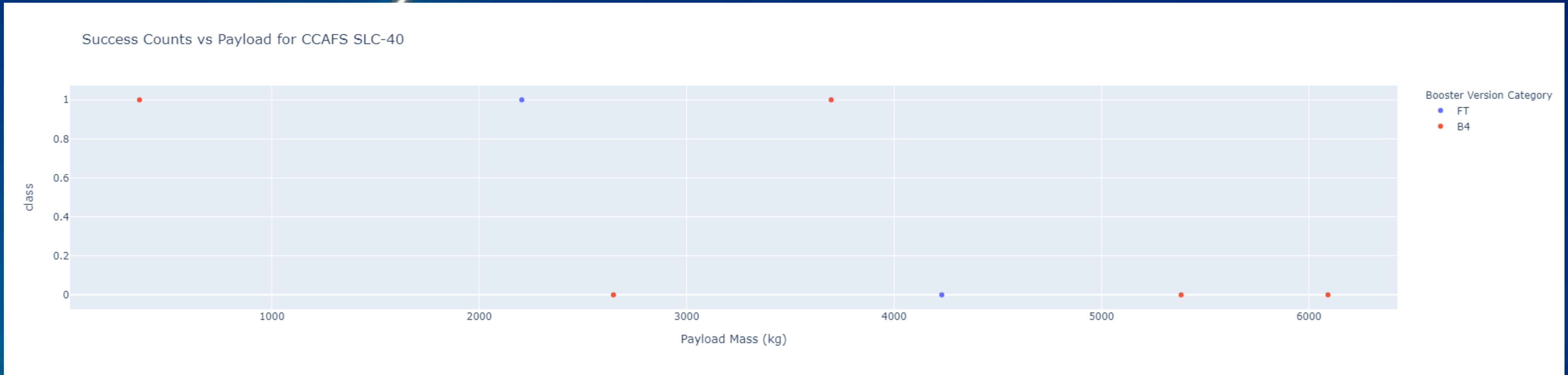
With KSC LC-39A having the most success ratio of 76.9%

CCAFS LC-40
Payload vs. Launch Outcome scatter plot



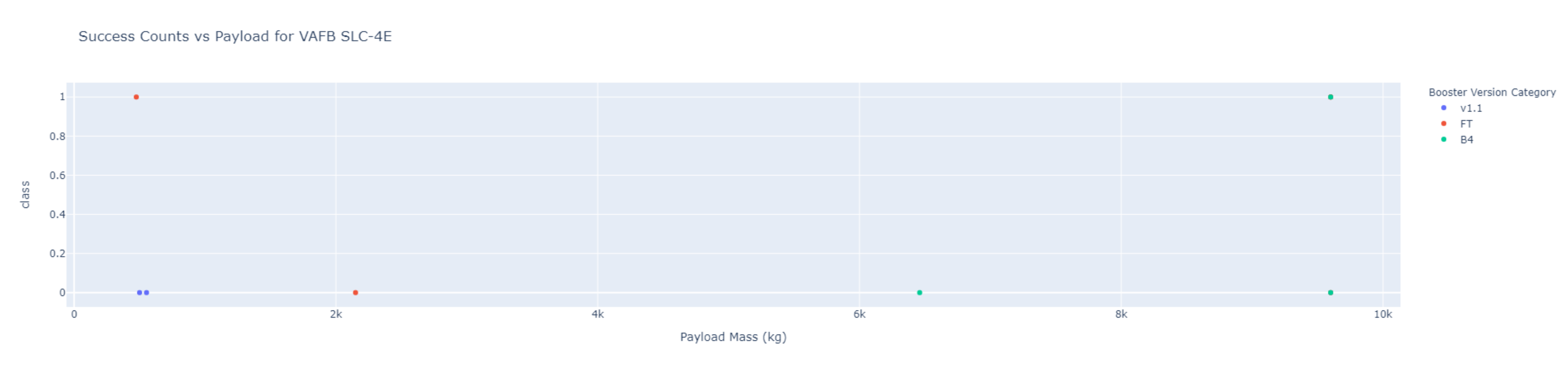
Success Rate is higher in the 2000 – 5000 kg range

CCAFS SLC-40
Payload vs. Launch Outcome scatter plot



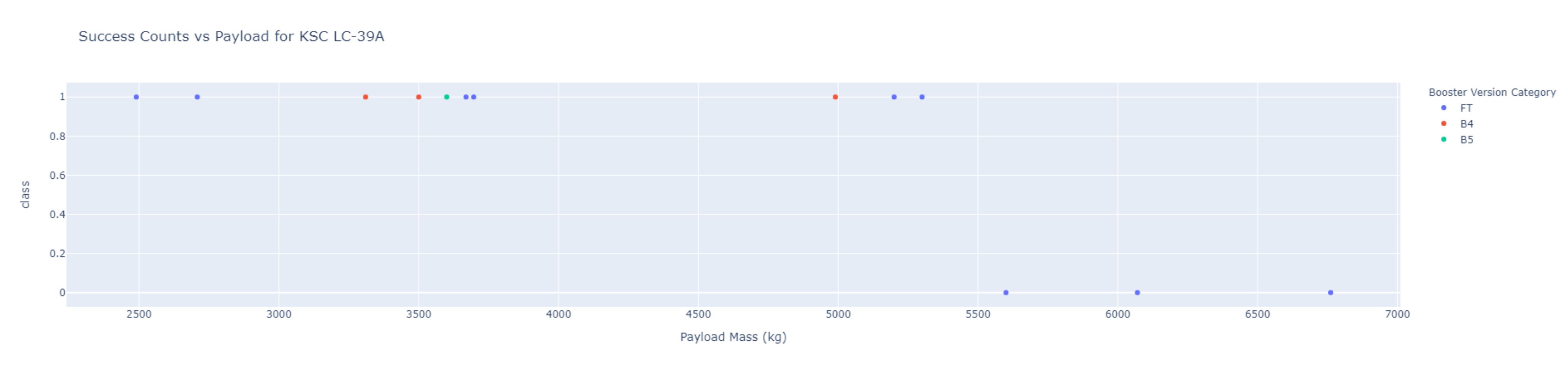
Success Rate is higher in the smaller payload mass range

VAFB SLC-4E
Payload vs. Launch Outcome scatter plot



Success Rate is higher in the small and large payload mass range

KSC LC-39A
Payload vs. Launch Outcome scatter plot



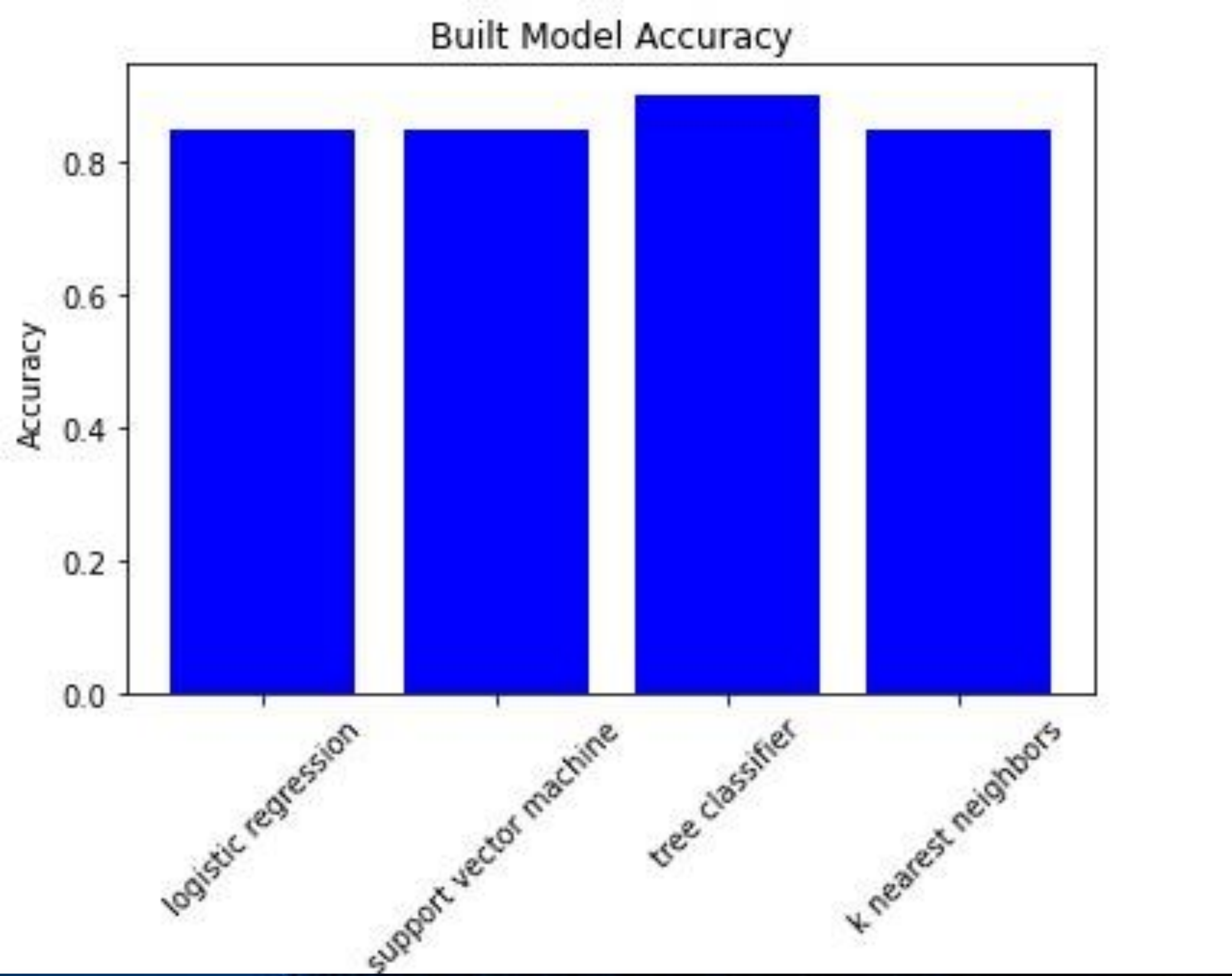
Success Rate is extremely high in the smaller payload mass range

Predictive analysis (Classification)

Classification Accuracy

Barchart showcasing all the built model accuracy for all built models

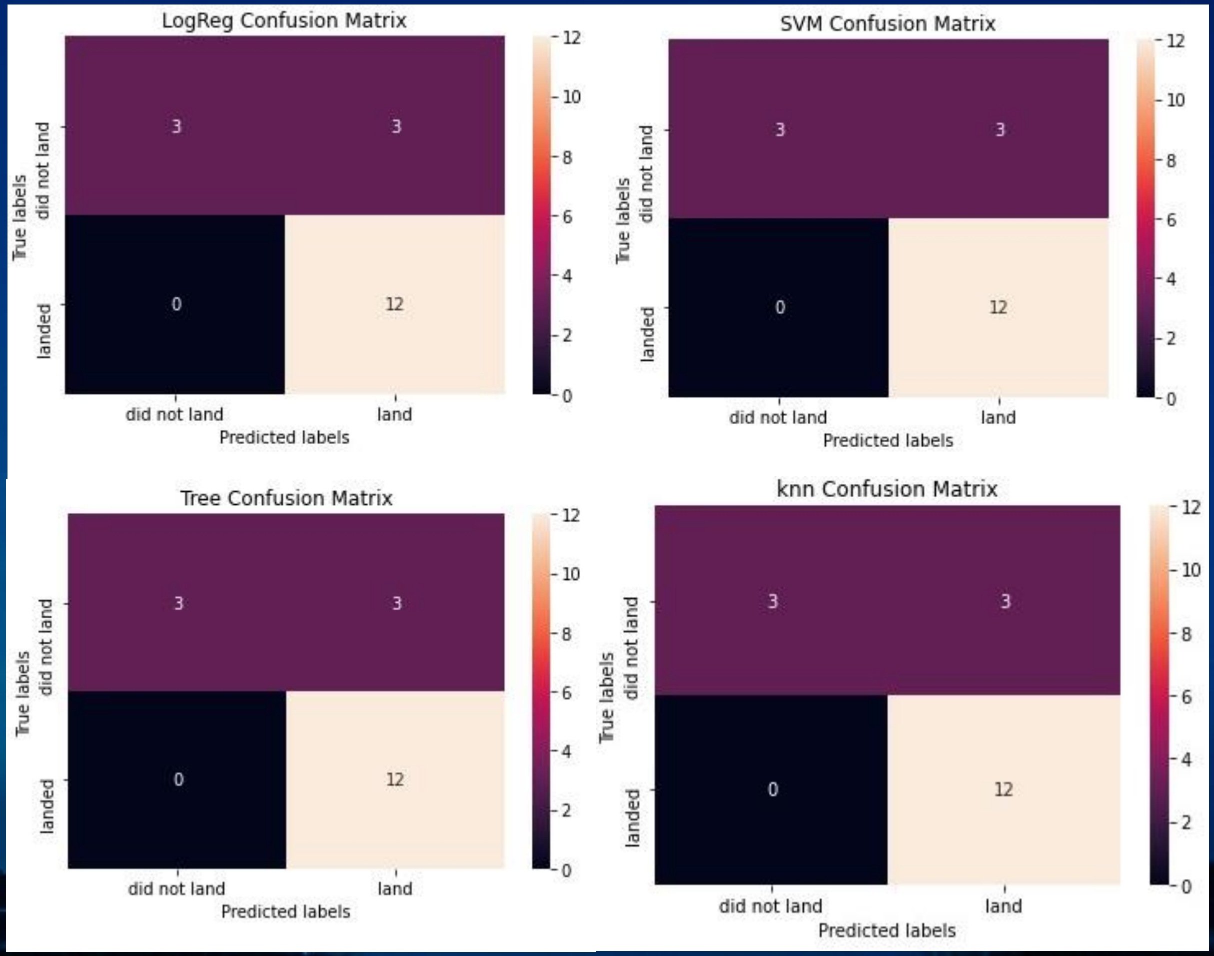
Tree Classifier has the highest accuracy



Confusion Matrix

All 4 models perform the same on the Test Data.

As all 4 models do not have a drastic performance difference on the training set, it is possible that due to the small test sample size (18 samples), all 4 models have the same confusion matrix.



CONCLUSION

- As the flight number increases, the first stage is more likely to land successfully
- More massive the payload, the less likely the first stage will return, except when payload mass over 10000kg, recent flights have a promising success rate
- CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
- large payload mass at "VAFB SLC 4E" launch site often brings success
- Small payload mass (<6000kg) and Large payload mass (>9000kg) at "KSC LC 39A" Launch site have high success rate
- Most flights are focusing on LEO, ISS, PO, GTO, and VLEO orbits
- High success rate is found on ES-L1, GEO, HEO, and SSO orbits
- Success rate is improving since 2013
- Launch sites typically are near railways, highways, and coastlines for ease of transportation, and are all far from nearest cities to minimize the effect on local residence and infrastructure
- Tree classifier has the most accuracy on predicting the success rate of a launch

Appendix and Reference

- Data sets and API
  1. https://github.com/r-spacex/SpaceX-API
  2. https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetwork-Channel-SkillsNetworkCoursesIBMDS0321ENSkillsNetwork26802033-2021-01-01
  3. https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/dataset_part_1.csv
  4. https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/api/dataset_part_2.csv
  5. https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/spacex_launch_geo.csv
  6. https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/api/dataset_part_3.csv

- Background
  1. https://www.wallpaperbetter.com/en/hd-wallpaper-zkfqt

Appendix (Cont.)

Github Urls

1. https://github.com/AtticusAlan/Coursera_Applied_Data_Science_Capstone/blob/master/Week1_jupyter-labs-spacex-data-collection-api.ipynb
2. https://github.com/AtticusAlan/Coursera_Applied_Data_Science_Capstone/blob/master/Week1_jupyter-labs-webscraping.ipynb
3. https://github.com/AtticusAlan/Coursera_Applied_Data_Science_Capstone/blob/master/Week1_labs-jupyter-spacex-Data%20wrangling.ipynb
4. https://github.com/AtticusAlan/Coursera_Applied_Data_Science_Capstone/blob/master/Week2_jupyter-labs-eda-dataviz.ipynb
5. https://github.com/AtticusAlan/Coursera_Applied_Data_Science_Capstone/blob/master/Week2_jupyter-labs-eda-sql-coursera.ipynb
6. https://github.com/AtticusAlan/Coursera_Applied_Data_Science_Capstone/blob/master/Week3_lab_jupyter_launch_site_location.ipynb
7. https://github.com/AtticusAlan/Coursera_Applied_Data_Science_Capstone/blob/master/spacex_dash_app.py
8. https://github.com/AtticusAlan/Coursera_Applied_Data_Science_Capstone/blob/master/Week4_SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb