IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jinrong Yan
Oct-26-2021

Outline

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

Executive Summary

- In the research, we use data collection, wrangling to clean up the data, and visualize these data in interactive maps and charts to explore. Further, we use classification methods to make predictions.
- The Falcon 9 series has increase landing success rate as time goes. However, it does not strongly improve in some orbits. Our machine learning is great predicting the landing success rate, but the sample size is not large enough to make accurate predictions.

Introduction

- To beat the Musk's Space X empire, we have to take some action to avoid the monopoly.
- To take the shortcut, we need to find out, how long time did Space X need to improve its performance, which orbits has greater success rate, and which launch site has the best performance.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data was collected through SpaceX API and web scrapping on SpaceX wikipedia
- Perform data wrangling
  - Simplified multiple criterion of landing outcome as binary elements.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Split the dataset into train set and test set randomly. Then used multiple types of classification method to train the data, and used grid search to tune the algorithm, to find the best fitted model. Finally validated the model with test set.

Data Collection

- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts

Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

- https://github.com/yanjinrong1/SpaceXestimator/blob/main/Data%20Collection%20API.ipynb

- Use the GET request, request and get a response file from SpaceX API
- Use response.json() to decode into json
- Use pd.json_normalized() to transform into pandas dataframe
- Transform and formatt data
- Substitute variables
- Filtered the dataframe to only include Falcon 9 launches
- Replaced missing payload mass with mean value

Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts

- Add the GitHub URL of the completed web scraping notebook, as an external reference and peer-review purpose

- Request the Falcon9 launch from Wikipedia url
- Use beautifulsoup to organize html response
- Extract data on the html tags and made into lists
- Create a panda data frame based on the parsed lists.
- https://github.com/yanjinrong1/SpaceXestimator/blob/main/Data%20Collection%20with%20Web%20Scraping_done.ipynb

Data Wrangling

- In the data set, there are multiple conditions of landing, some are recognized as successful, while others are unsuccessful. We labelled the successful category as 1, and unsuccessful as 0 for convenience.
- Process:
  - Calculate the number of launches on each site
  - Calculate the number and occurrence of each orbit
  - Calculate the number of mission outcome per orbit type
  - Create a landing outcome label from outcome column
- https://github.com/yanjinrong1/SpaceXestimator/blob/main/Data%20wrangling.ipynb

EDA with Data Visualization

- We plotted
  - The Flight number vs. Payload vs. launch site with landing outcome overlay
  - The successful rate vs. orbit and year
  - The flight number vs. orbit vs. payload mass with landing outcome overlay
- We learnt the successful rate increase with flight number, payload mass, and there is no strong relation with launch site.

https://github.com/yanjinrong1/SpaceXestimator/blob/main/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- I queryed
  - The total payload mass by NASA
  - The average payload mass carried by booster version F9 v1.1
  - The first successful landing outcome on ground pad.
  - The boosters at specific payload mass range
  - Total number of successful and failure mission outcome
  - ….More detail please searched my github
- https://github.com/yanjinrong1/SpaceXestimator/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- In my folium map, these widgets are added:
  - Successful/failed launched icons were added
  - A line from the launch site to its nearest coast
- By adding these objects, it allowed us to explore：
  - The location of each launch site
  - The distribution of flight of each launch site
  - The distance from launch site to coastline, transportation, populated area, etc
- https://github.com/yanjinrong1/SpaceXestimator/blob/master/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

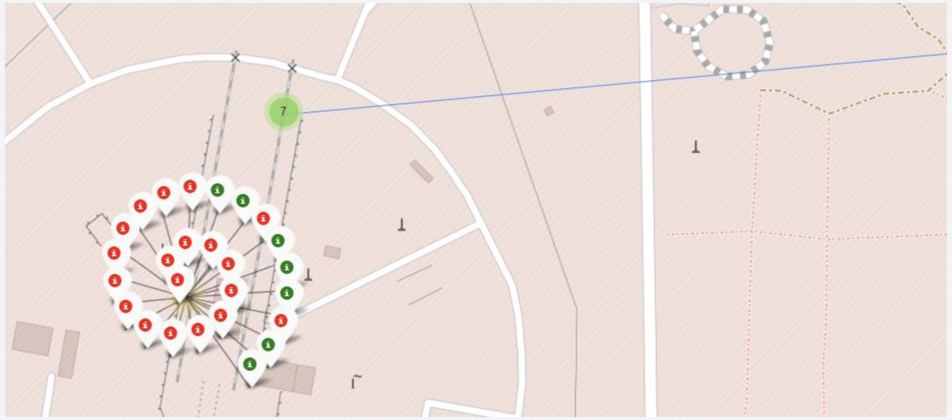
- Two plots were generate in the dashboard
  - The pie chart for all launch site and each launch sites
  - The landing outcome vs. payload mass interactive chart
- Adding those chart can help us better understand the correlation between payload and successful rate for the falcon 9
- https://github.com/yanjinrong1/SpaceXestimator/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Four common types of classification model were applied in this project. We split the data randomly into train set and test set. Used the train set and Grid search CV to find the best fitting model for each type of classification model, then predicting the test data, to see what extent we match the given data.
- Process:
  - Preprocessing the data into scaler form
  - Use GridSearchCV method model to find the best fitting parameters
  - Test the model with test set and plot the confusion matrix
- https://github.com/yanjinrong1/SpaceXestimator/blob/main/Machine%20Learning%20Prediction.ipynb

Results

- Exploratory data analysis results
  - Successful rate increase with year, flight number, and payload mass
  - Some orbits have great successful rate while other didn't
  - Most mission outcomes are success
- Interactive analytics demo in screenshots
- Predictive analysis results
  - The logistic regression model best fit the set of data



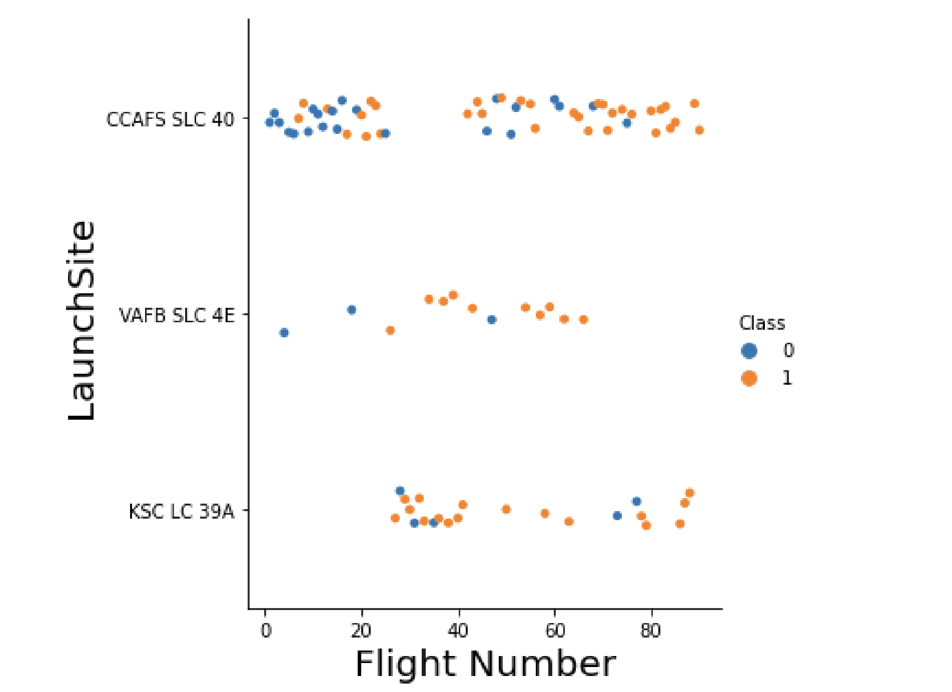
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- Show a scatter plot of Flight Number vs. Launch Site

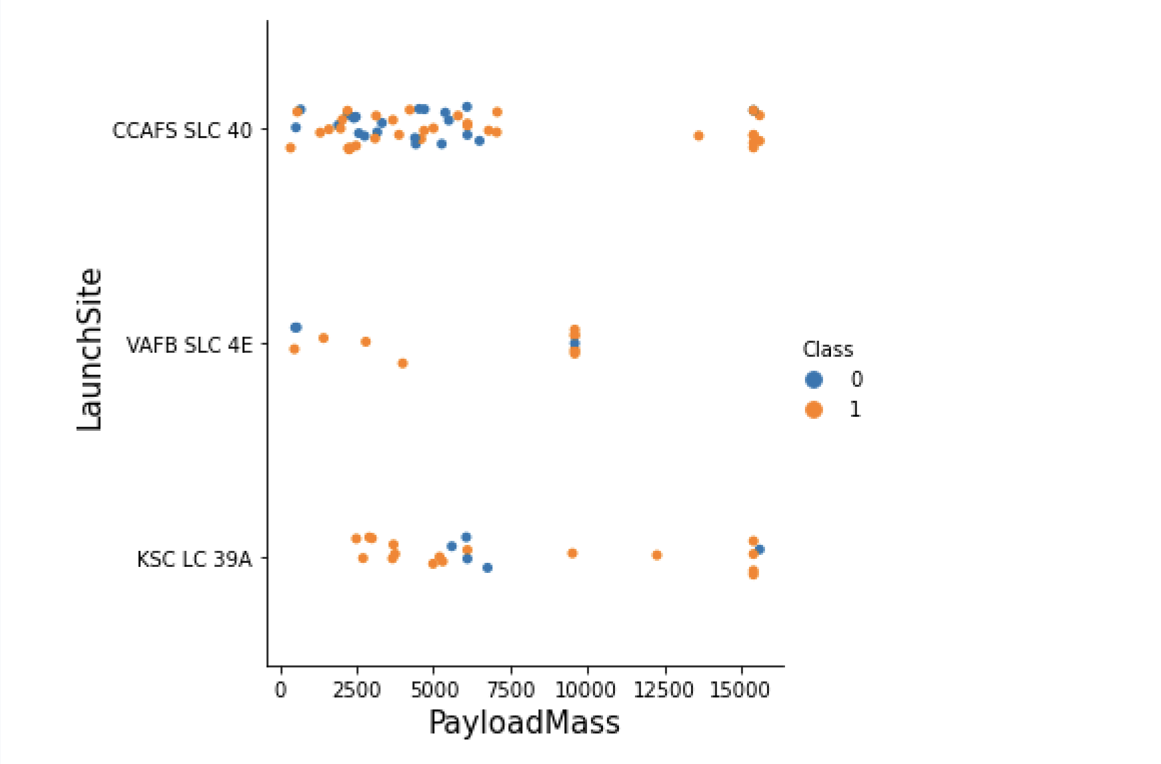
- As flight number increases, all flights from all launch sites increases the successful rate.



Payload vs. Launch Site

- Show a scatter plot of Payload vs. Launch Site

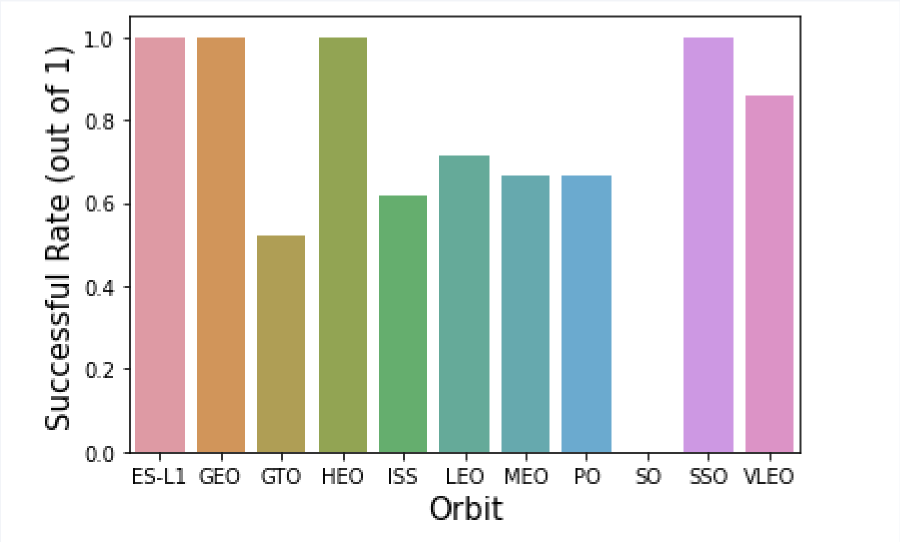
- The gaining of payload mass increase the successfully rate as well, especially for CCAFS SLC 40, and KSC LC39A departs flights



Success Rate vs. Orbit Type

- Show a bar chart for the success rate of each orbit type

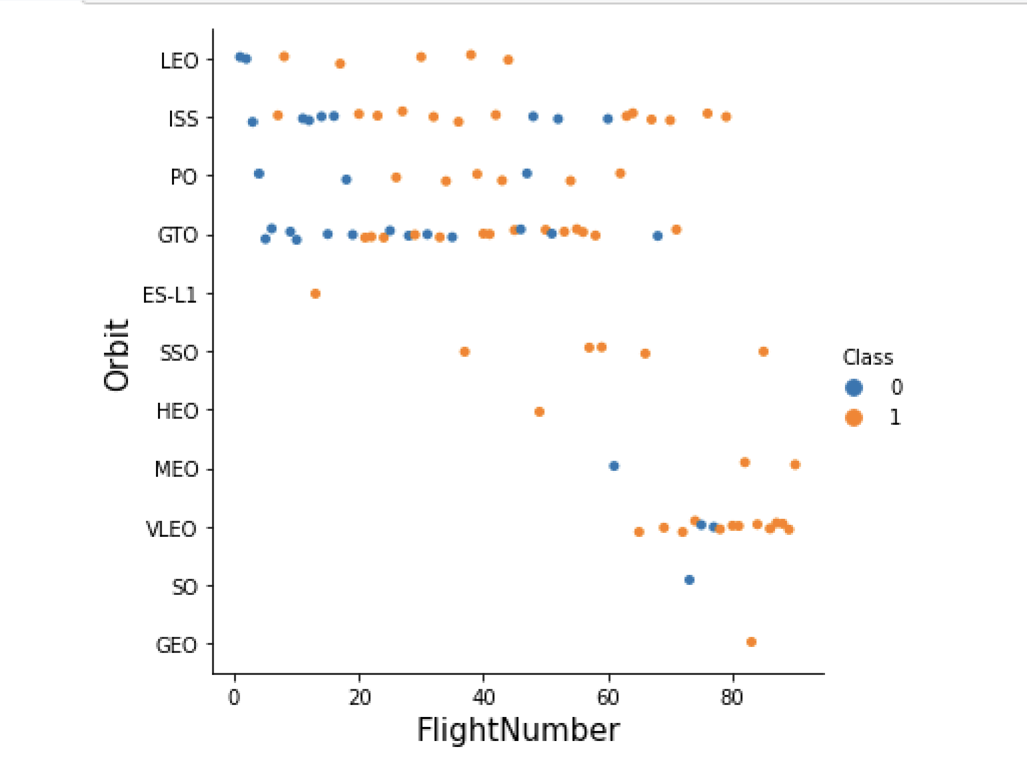
- In ES-L1, GEO, HEO, SSO oribit flights, it has a successful rate close to 1, while the SO orbit has an rate close to 0.



Flight Number vs. Orbit Type

- Show a scatter point of Flight number vs. Orbit type

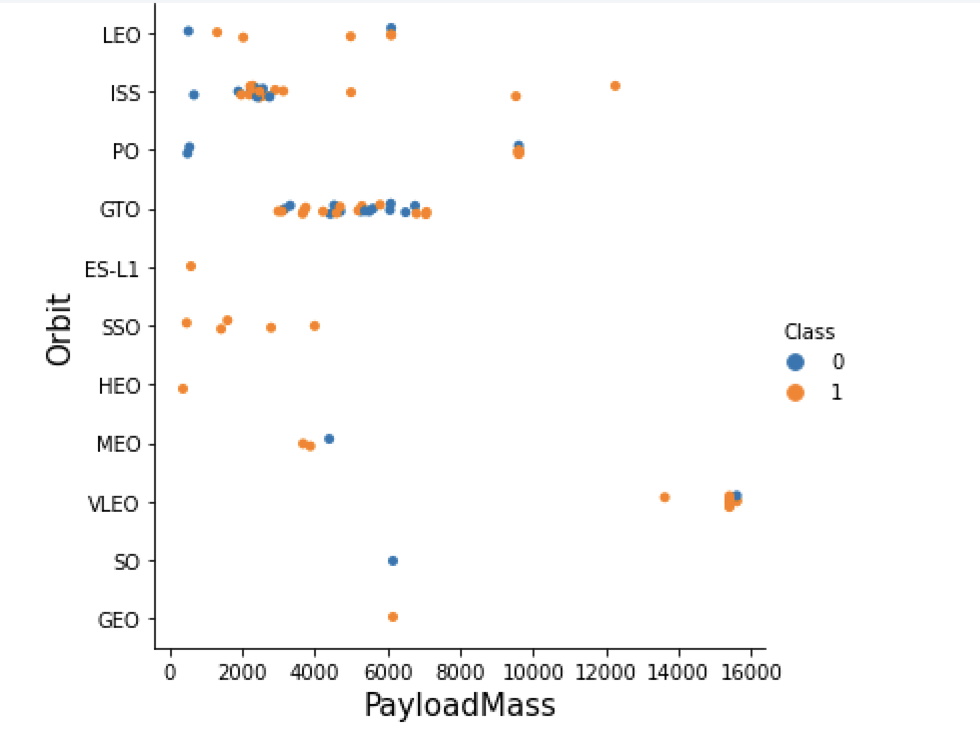
- This plot shows the successful rate does not increase as flight number increases for each orbit.



Payload vs. Orbit Type

- Show a scatter point of payload vs. orbit type

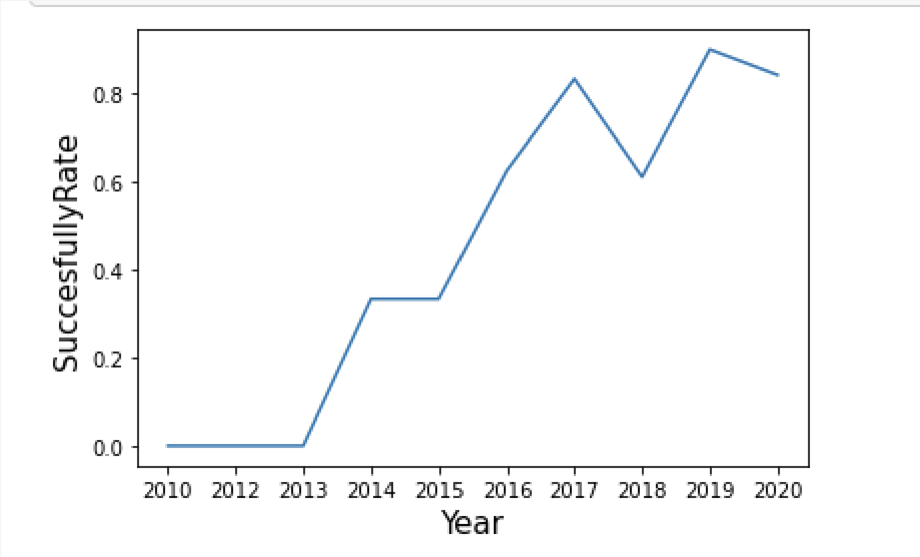
- This plot shows the successful rate does not change as payload mass increases for each orbit.



Launch Success Yearly Trend

- Show a line chart of yearly average success rate

- Since 2013, the success rate rocket up till 2019, to about 0.8



All Launch Site Names

- Four launch sites were query.
- The first and second item should be the same, it was caused by not well clean up before.

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Results is here as:

Out[9]:

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
- The total payload (kg) from NASA is 45596 kg

Out[12]:

total_payload_mass_kg_by_nasa
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Query result is 2928 kg on average

Out[16]:

1
2928

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- The first successful landing is as follow.

```
In [39]: %sql SELECT MIN(Date) FROM SPACEX WHERE "Landing _Outcome" LIKE 'Success (ground pad)'

 * ibm_db_sa://fvd84318:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdom
Done.
Out[39]:
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- There are four booster versions have payload mass between 4000 to 6000

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

```
In [40]: %sql SELECT Booster_Version FROM SPACEX WHERE PAYLOAD_MASS__KG_ >= 4000 AND PAYLOAD_MASS__KG_ <= 6000 AND "Landing _Outcome" LIKE 'Succes
s (drone ship)'

 * ibm_db_sa://fvd84318:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[40]:

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
- In terms of mission outcome, most mission were success, only 1 failure at flight

List the total number of successful and failure mission outcomes

In [43]:
```
%sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) FROM SPACEX GROUP BY MISSION_OUTCOME
```

```
 * ibm_db_sa://fvd84318:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[43]:

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- The name of booster carried max payload is as follows.

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

In [46]:
```
%sql SELECT BOOSTER_VERSION, PAYLOAD_MASS__KG_ FROM SPACEX WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEX)
```

```
 * ibm_db_sa://fvd84318:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[46]:

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

In [49]:
```
%%sql

SELECT "Landing _Outcome", BOOSTER_VERSION, LAUNCH_SITE
FROM SPACEX
WHERE YEAR(DATE) = 2015 AND "Landing _Outcome" = 'Failure (drone ship)'
```

```
 * ibm_db_sa://fvd84318:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08k
Done.
```

Out[49]:

Landing _Outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

- The failure attempts are shown in the table.

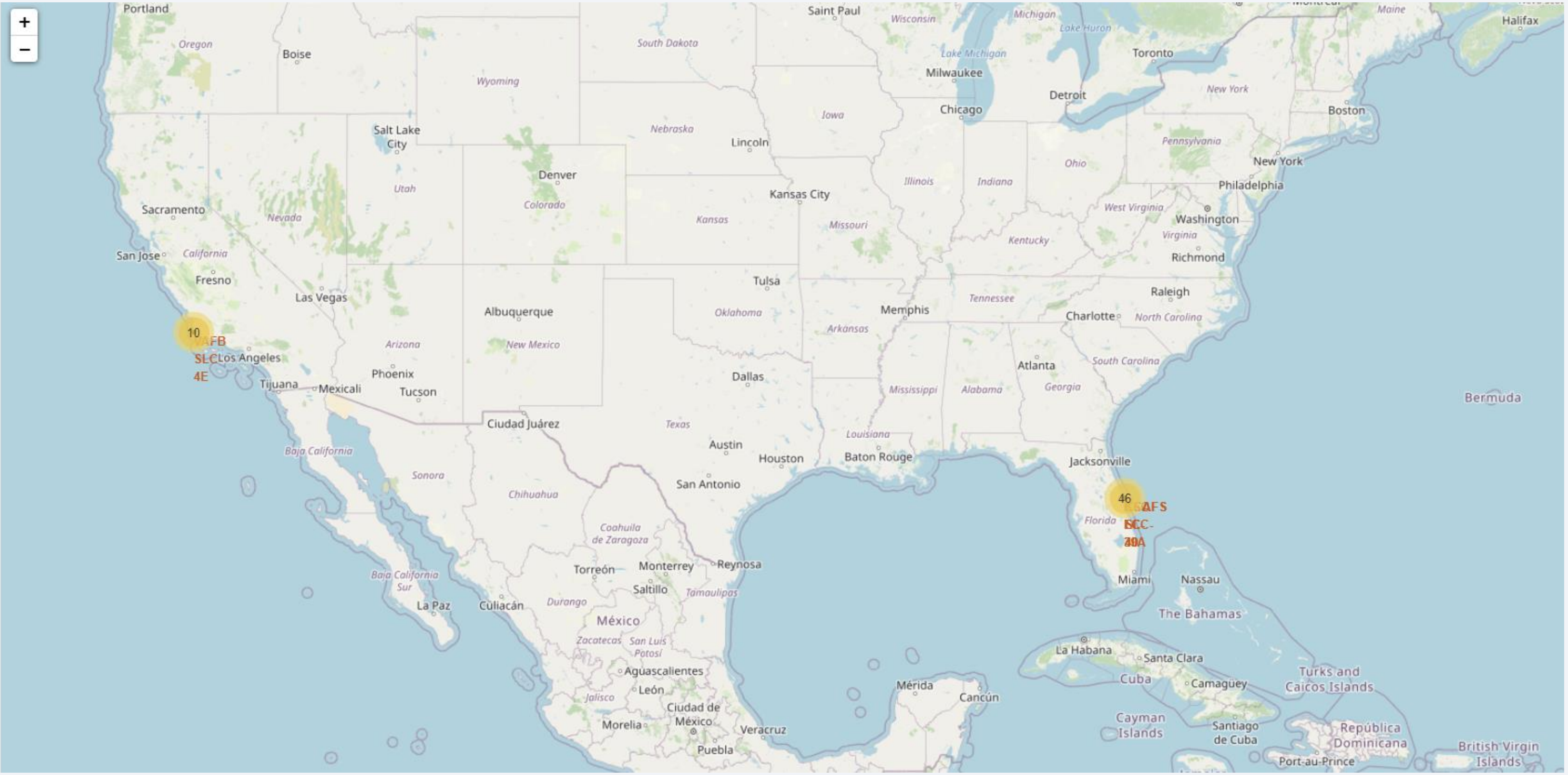
Out[53]:

Landing _Outcome	COUNT
Controlled (ocean)	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	10
Precluded (drone ship)	1
Success (drone ship)	5
Success (ground pad)	3
Uncontrolled (ocean)	2

Section 4

Launch Sites Proximities Analysis

Map of launch site

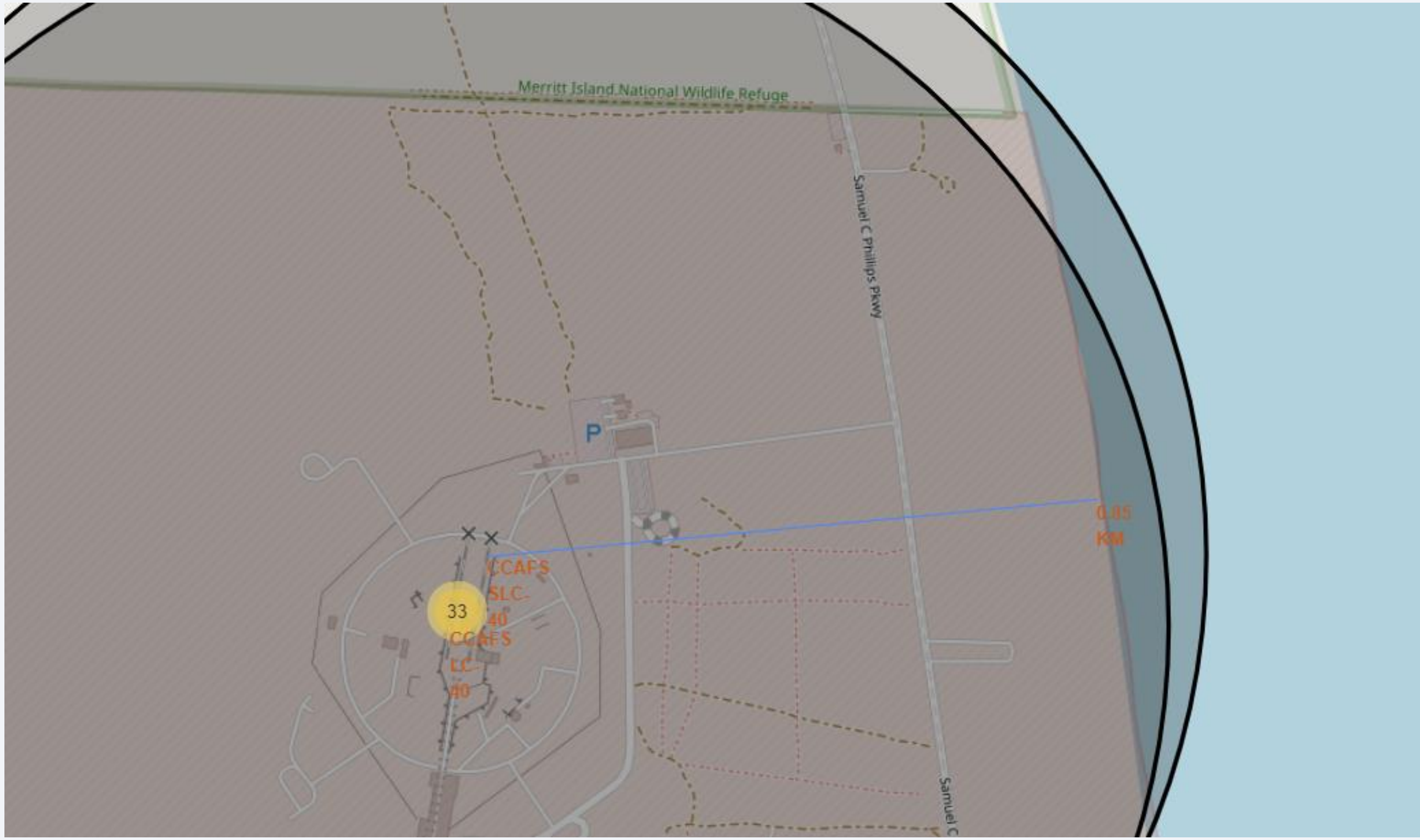


- From this map, we learnt all launch sites are near the coast, and at the south end of US

Landing outcome of flights from CCAFS LC-40

- Many flights were failed at this site

Nearby object of the launch site

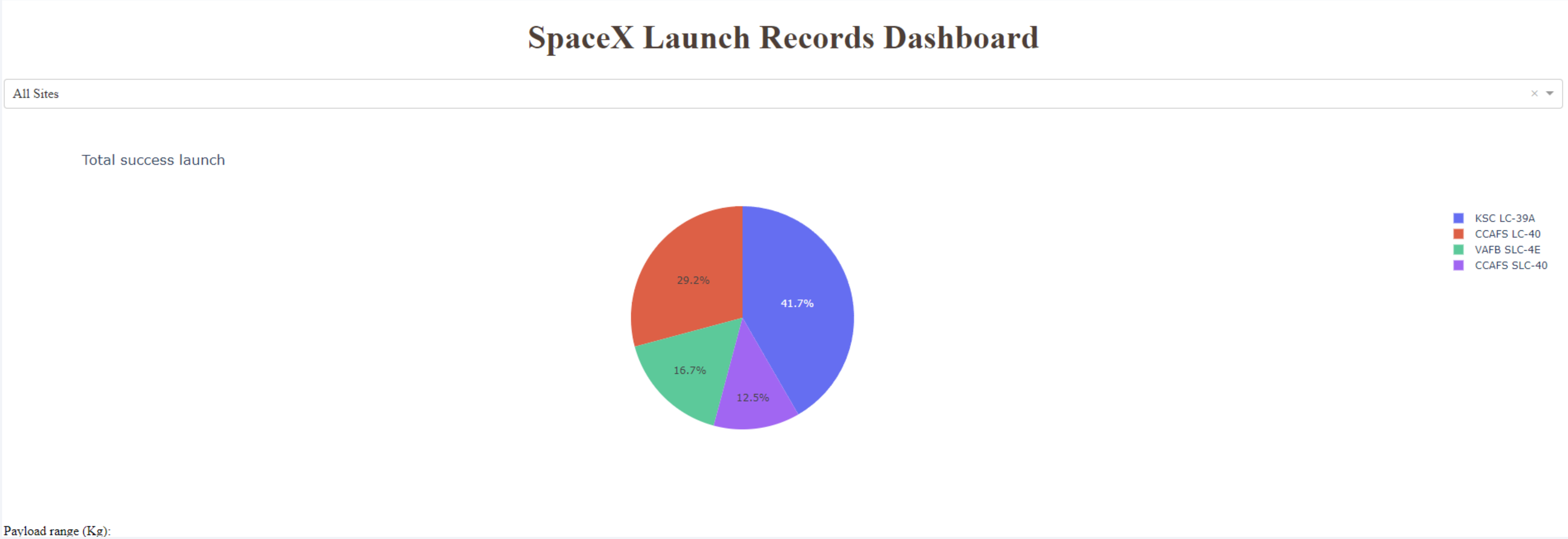


- The launch site is so close to the coastline

Section 5

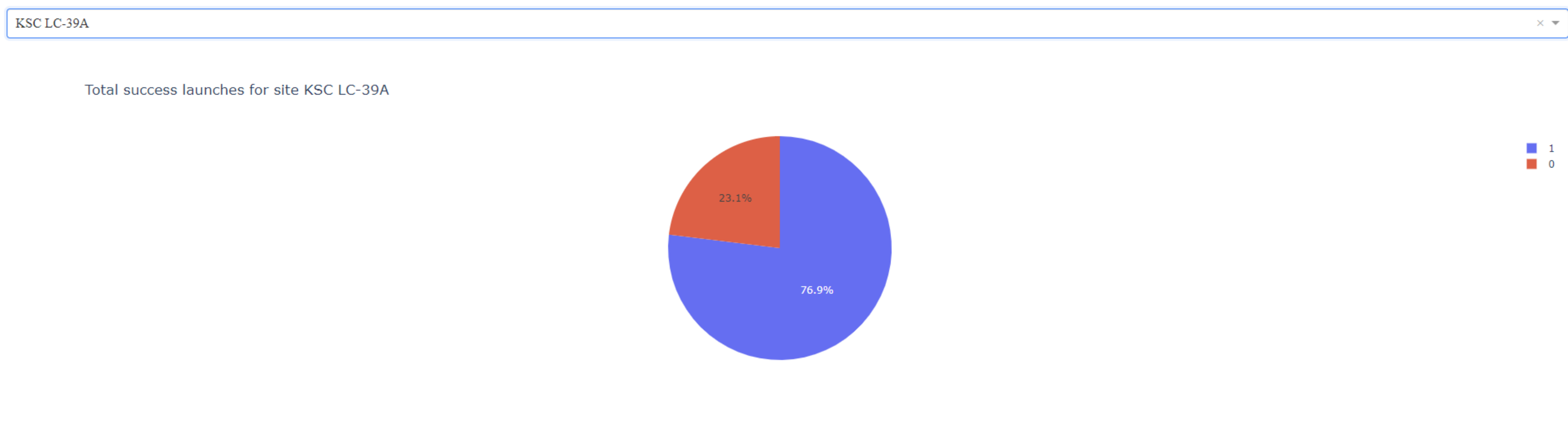
Build a Dashboard with Plotly Dash

Total success launch of each launch site



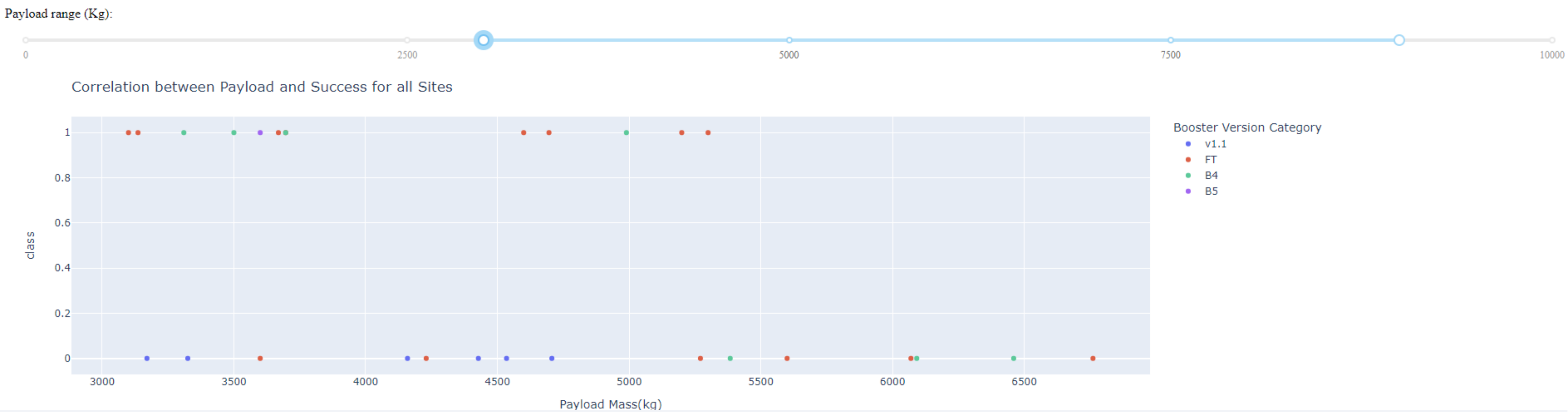
- The VAFB and CCAFS SLC-40 launch site have a relatively low success rate

Launch success rate of KSC LC-39A



- The success rate is 76.9% at KSC LC-39A

Correlation between payload and success for all sites

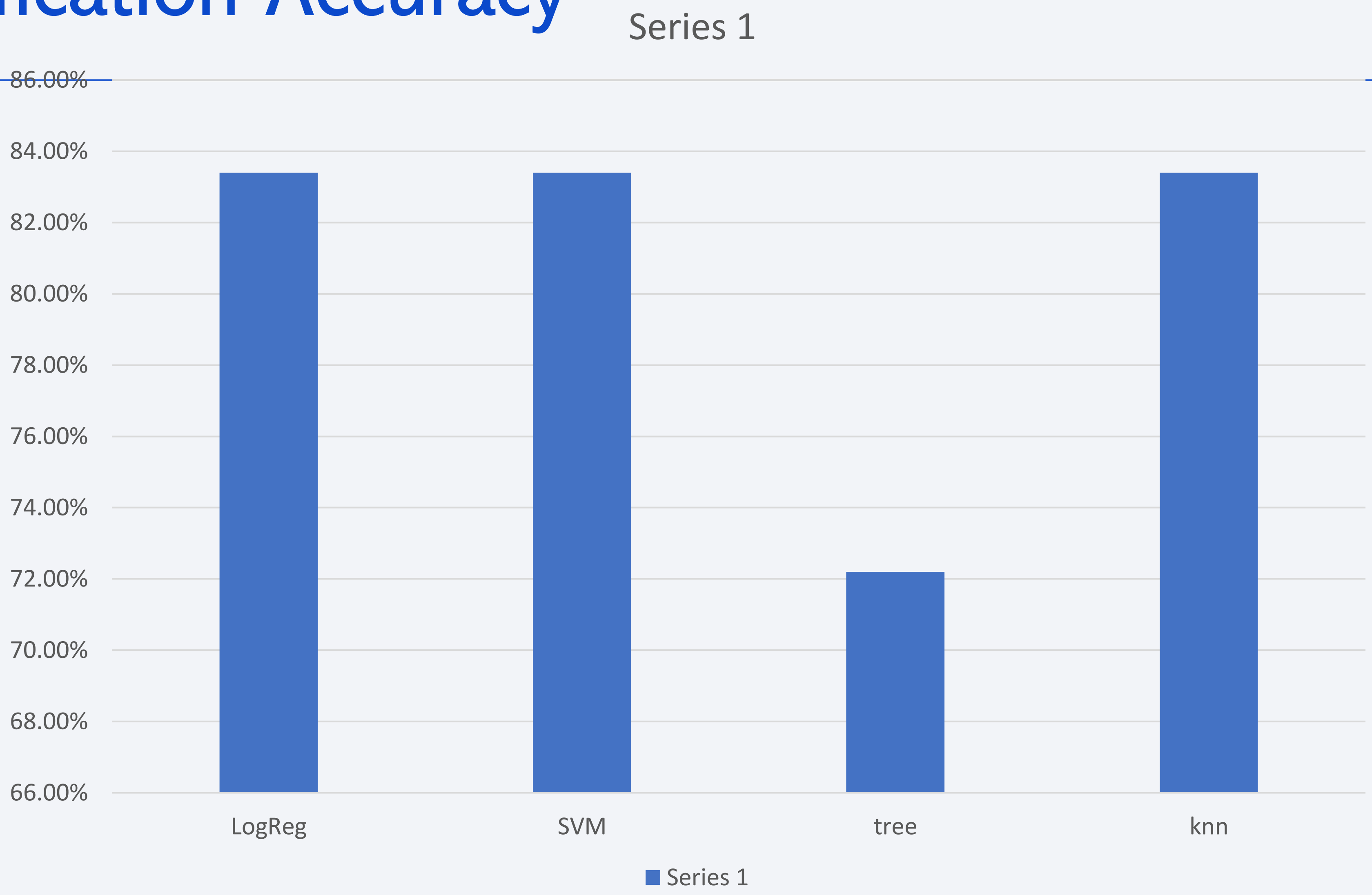


- The v1.1 booster version has terrible successful rate.

Section 6

Predictive Analysis (Classification)

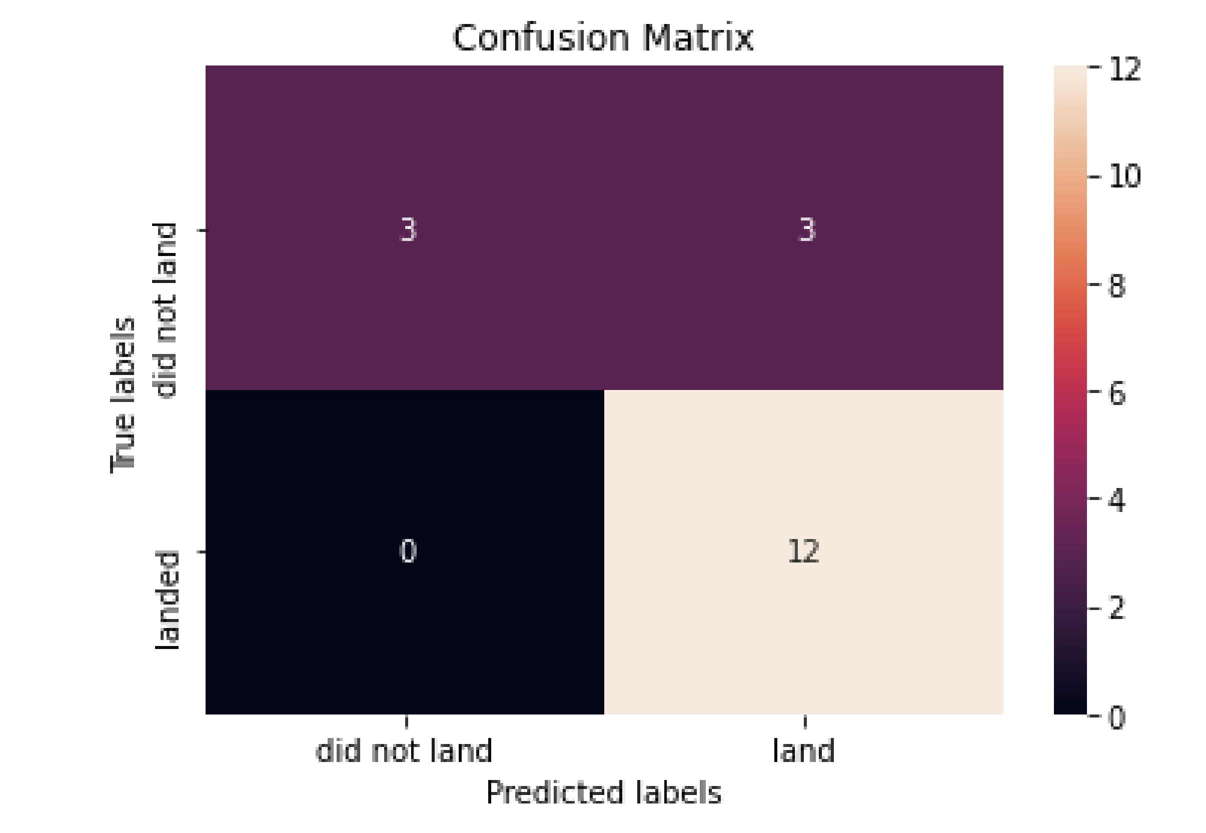
Classification Accuracy



- Log Reg, SVM, and k nearest has similar accuracy

Confusion Matrix

- The Logistic Regression model has one of the best accuracy of prediction, which has 80% accuracy when it will land, while 100% accuracy when it is not landed.



Conclusions

- The successful rate of SpaceX booster landing increase with year
- But it is still unstable at some orbits, no improvement was obtained in some orbits
- The sample for machine learning is too small, which making all methods have similar results.

Appendix

- Sorry, no extra material

Thank you!