IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Alena Baklanova
22/06/2023

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

- SpaceX – currently the most successful commercial space agency.
- It provides the much cheaper rocket launches relative to competitors due to the reusable first stage.

New company Space Y would like to compete with SpaceX, so we need to:

- Determine the price of each SpaceX launch
- Determine if SpaceX will reuse the first stage

Section 1

Methodology

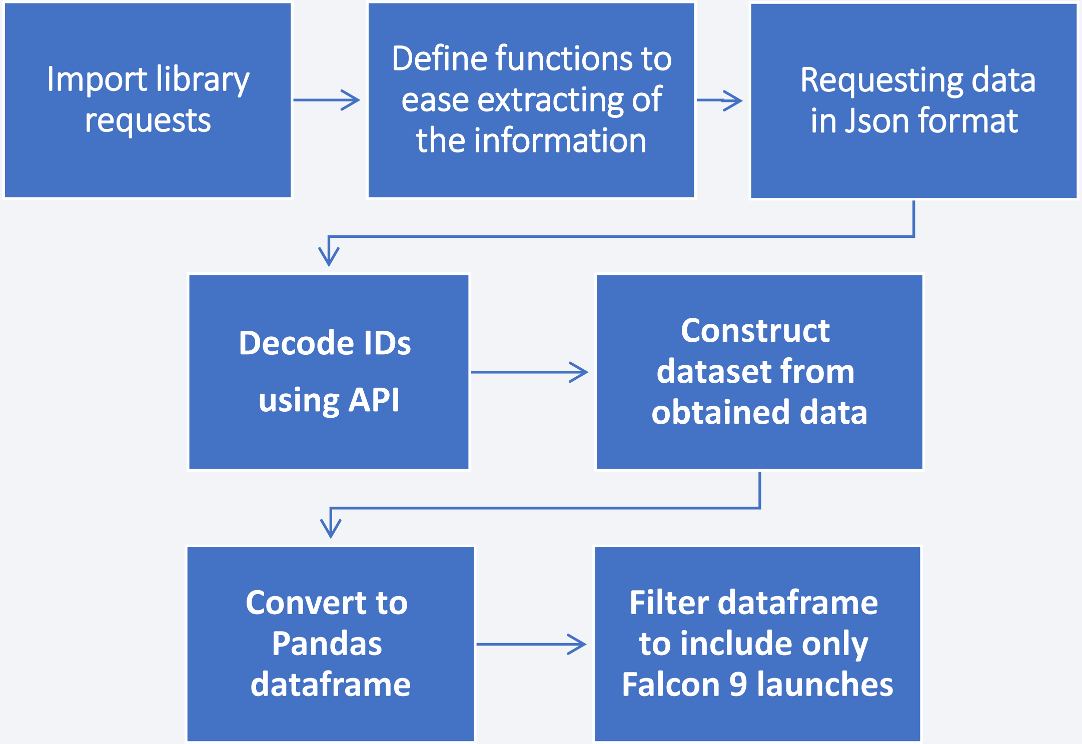
Methodology

Executive Summary

- Data collection methodology:
  - SpaceX REST API
  - Wikipedia
- Perform data wrangling
  - Wrangling Data using an API. Decode data using other tables from SpaceX REST API
  - Sampling Data. Filter the data to only include Falcon 9 launches
  - Dealing with Nulls and missing values
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection – SpaceX API

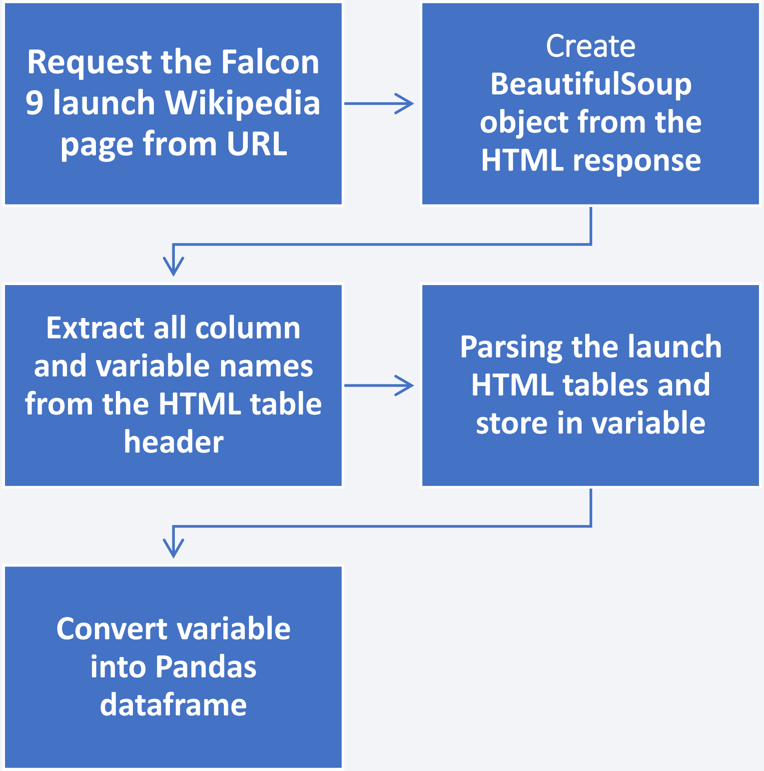
- HTTP requests were used to connect to API, so library requests was imported before commection.
- We obtained data about rocket type, launchpad, payload and information related to core.
  - Ex.: *response = requests.get("https://api.spacexdata.com/v4/payloads/" +load).json()*



GitHub - [Capstone_Coursera/jupyter-labs-spacex-data-collection-api.ipynb at Week-1 · Aleninok-pixel/Capstone_Coursera (github.com)]

Data Collection - Scraping

- Falcon 9 historical launch records could be obtained from Wikipedia from HTML table "*List of Falcon 9 and Falcon Heavy launches*"
- Relevant table was requested using HTTP GET method and was transformed into BeautifulSoup object to then parse the table
- Table was converted into a Pandas dataframe for further analysis



GitHub - *Capstone_Coursera/jupyter-labs-webscraping.ipynb at Week-1 · Aleninok-pixel/Capstone_Coursera · GitHub*

Data Wrangling

1. Exploratory Data Analysis
   - Identified and calculated the percentage of the missing values in each attribute
   - Calculated the number and occurrence of mission outcome per orbit type

2. Determination Training Labels
   - Created a column with binary landing outcome (bad outcome = 0, good outcome = 1)

GitHub - *Capstone_Coursera/labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb at Week-1 · Aleninok-pixel/Capstone_Coursera · GitHub*

EDA with Data Visualization

Following charts were plotted with overlay of launch outcomes:

- **Flight number vs. Payload mass**
  - To identify relationship between launch success and payload of rocket
  - To show that as the flight number increases, the first stage is more likely to land successfully
- **Flight number vs. Launch site**
  - To identify dependence on success from launch site
- **Launch site vs. Payload mass**
  - To understand what payload mass rockets were launched from each site
- **Orbit vs. Flight number**
  - To show relationship between orbit type and launch success
- **Orbit vs. Payload mass**
  - To identify what payload mass of rocket is likely to be launched on each orbit type
- **The average launch success rate of each orbit type**
- **The average launch success trend through years**

GitHub - *https://github.com/Aleninok-pixel/Capstone_Coursera/blob/Week-2/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb*

EDA with SQL

In purpose of SpaceX dataset understanding we used SQL queries:

- Find unique launch sites
  - *%sql SELECT DISTINCT(Launch_Site) FROM SPACEXTBL*
- Display 5 records where launch sites begin with the string 'CCA'
  - *%sql SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5*
- Display the total payload mass of NASA (CRS) rockets
  - *%sql SELECT SUM(PAYLOAD_MASS__KG_) as Total_payload_mass FROM SPACEXTBL WHERE Customer = "NASA (CRS)"*
- Show the average payload mass of Falcon9 rockets
  - *%sql SELECT AVG(PAYLOAD_MASS__KG_) as Average_payload_mass FROM SPACEXTBL WHERE Booster_Version LIKE '%F9 v1.1%'*
- List the date of the first successful landing outcome in ground pad
  - *%sql SELECT MIN(DATE) FROM SPACEXTBL WHERE Landing_Outcome LIKE '%ground%'*
- Find the boosters that have successful drone ship landing with payload mass between 4000 and 6000
- List the total number of successful and failure mission outcomes
- List the boosters with the maximum payload mass
- List the records with the month names, failure outcomes in drone ship ,boosters and launch sites in year 2015

GitHub - *https://github.com/Aleninok-pixel/Capstone_Coursera/blob/Week-2/jupyter-labs-eda-sql-coursera_sqllite.ipynb*

Build an Interactive Map with Folium

Following objects were created and added to folium map:

- Markers of all launch sites with circle object and corresponding name as a popup label
- Markers of all the successful and failed launches for each site on the map
- Lines to closest coastlines and cities from each launch sites with distance labels

All these visualized data help in further analysis to understand possible dependences.

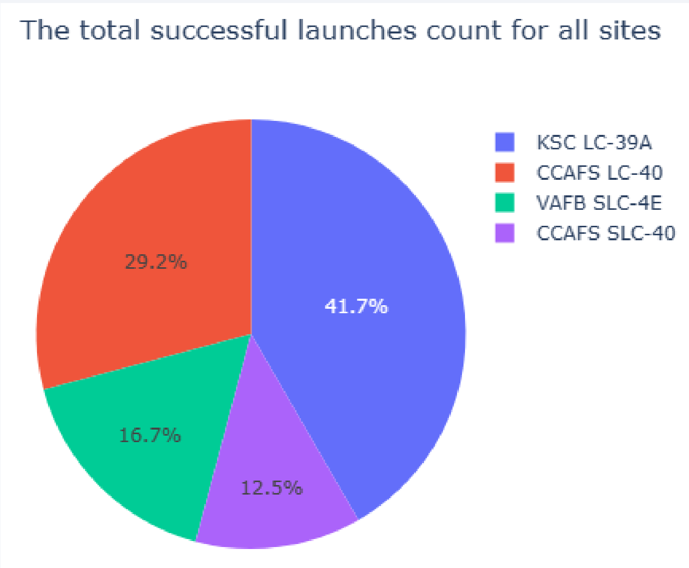
GitHub - *Capstone_Coursera/lab_jupyter_launch_site_location.jupyterlite (1).ipynb at Week-3 · Aleninok-pixel/Capstone_Coursera · GitHub*

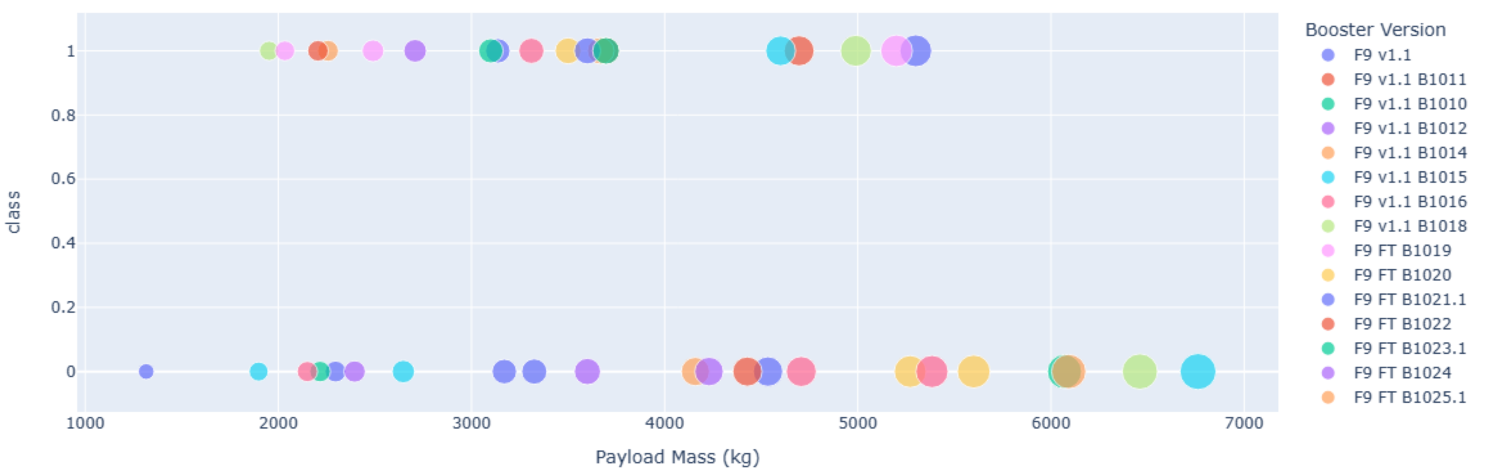
Build a Dashboard with Plotly Dash

Dashboard was built to visualize correlations between launch sites, success rate of launches and payload of rockets.

Following charts were added:

- Interactive pie chart of successful launches for each launch site and for all sites
- Interactive scatter plot of success rate in relation to payload mass depending on selected payload range for each site and all sites





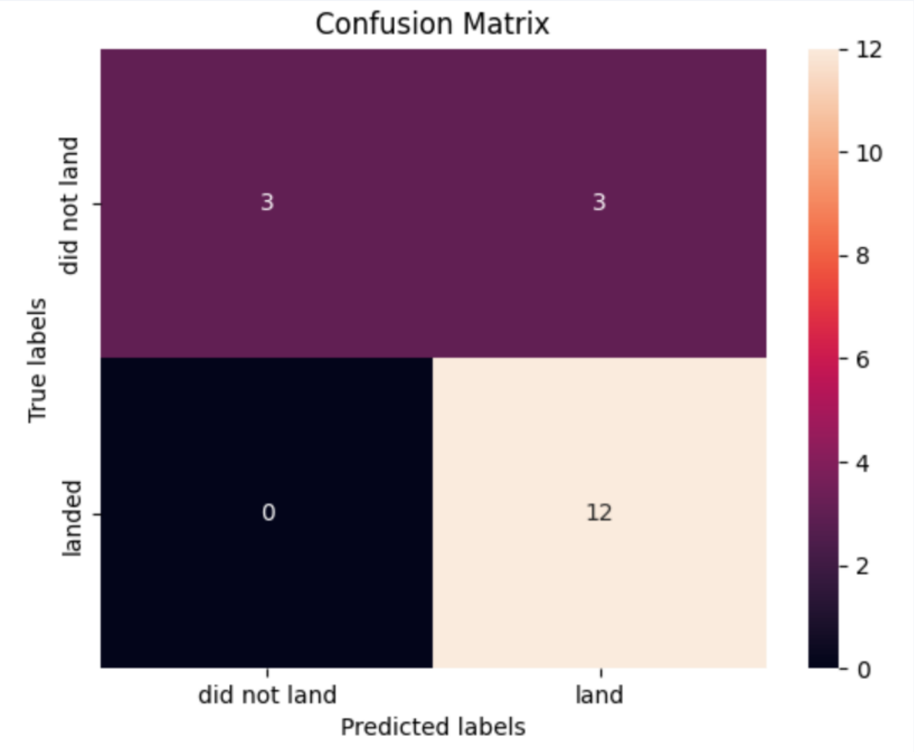
GitHub - [Capstone_Coursera/spacex_dash_app (1).py at Week-3 · Aleninok-pixel/Capstone_Coursera (github.com)]

Predictive Analysis (Classification)

Model development:

1. Prepared the data (standardize and split into training and test data)
2. Created and fitted logistic regression
3. Created and fitted SVM
4. Created and fitted Decision tree
5. Created and fitted K-nearest neighbors
6. For each model calculated accuracy and plotted confusion matrix

Most of the models showed the same accuracy around 83.3% and confusion matrix



As accuracy of all models are the same, we chose the simplest model, which is Logistic regression.

GitHub - [Capstone_Coursera/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb at Week-4 · Aleninok-pixel/Capstone_Coursera (github.com)]

Results

Exploratory data analysis results are:

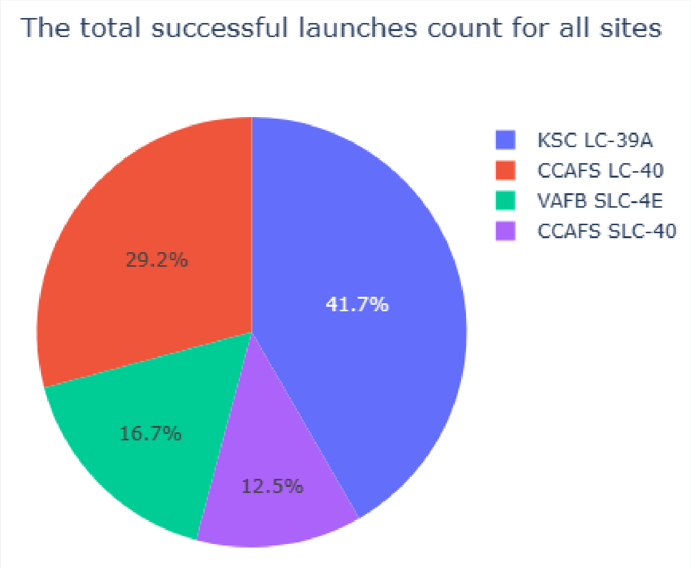
- There are no rockets launched for heavy payload mass(greater than 10000) from VAFB-SLC launch site
- The highest success rates of landing has several orbits: ES-L1, GEO, HEO, SSO.
- The LEO orbit's success appears to be related to the number of flights
- No relationship between flight number when in GTO orbit
- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

Interactive analytics demo in screenshots:

- The most successful launch site is KSC LC-39A

Predictive analysis results are:

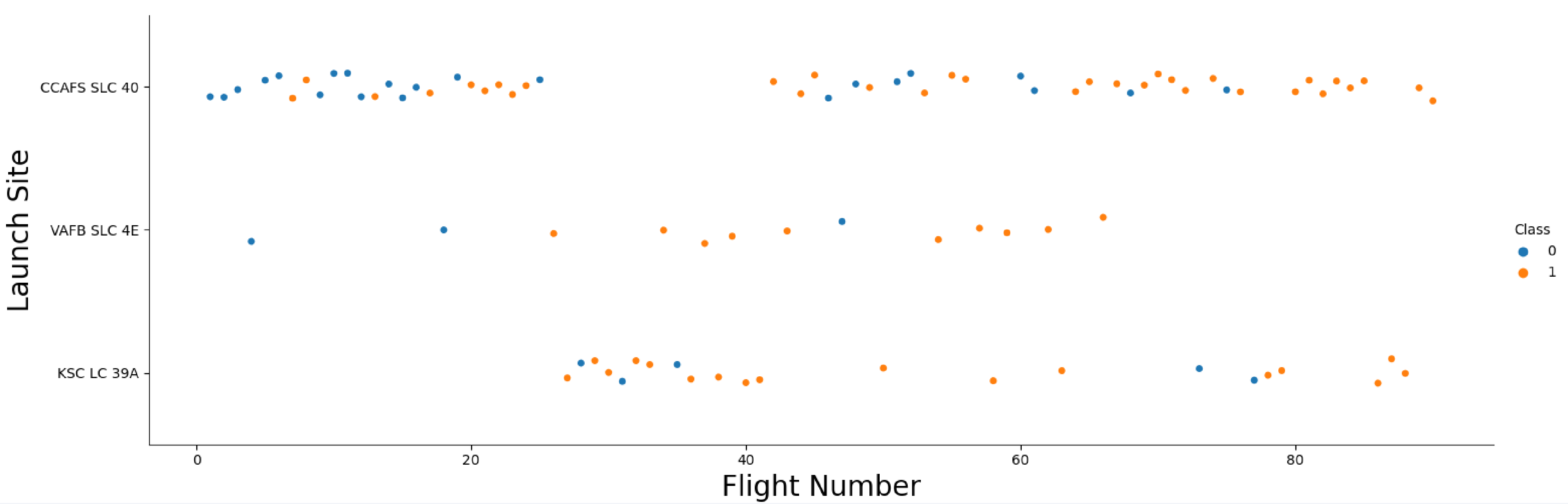
- The best model for prediction – Logistic regression



Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

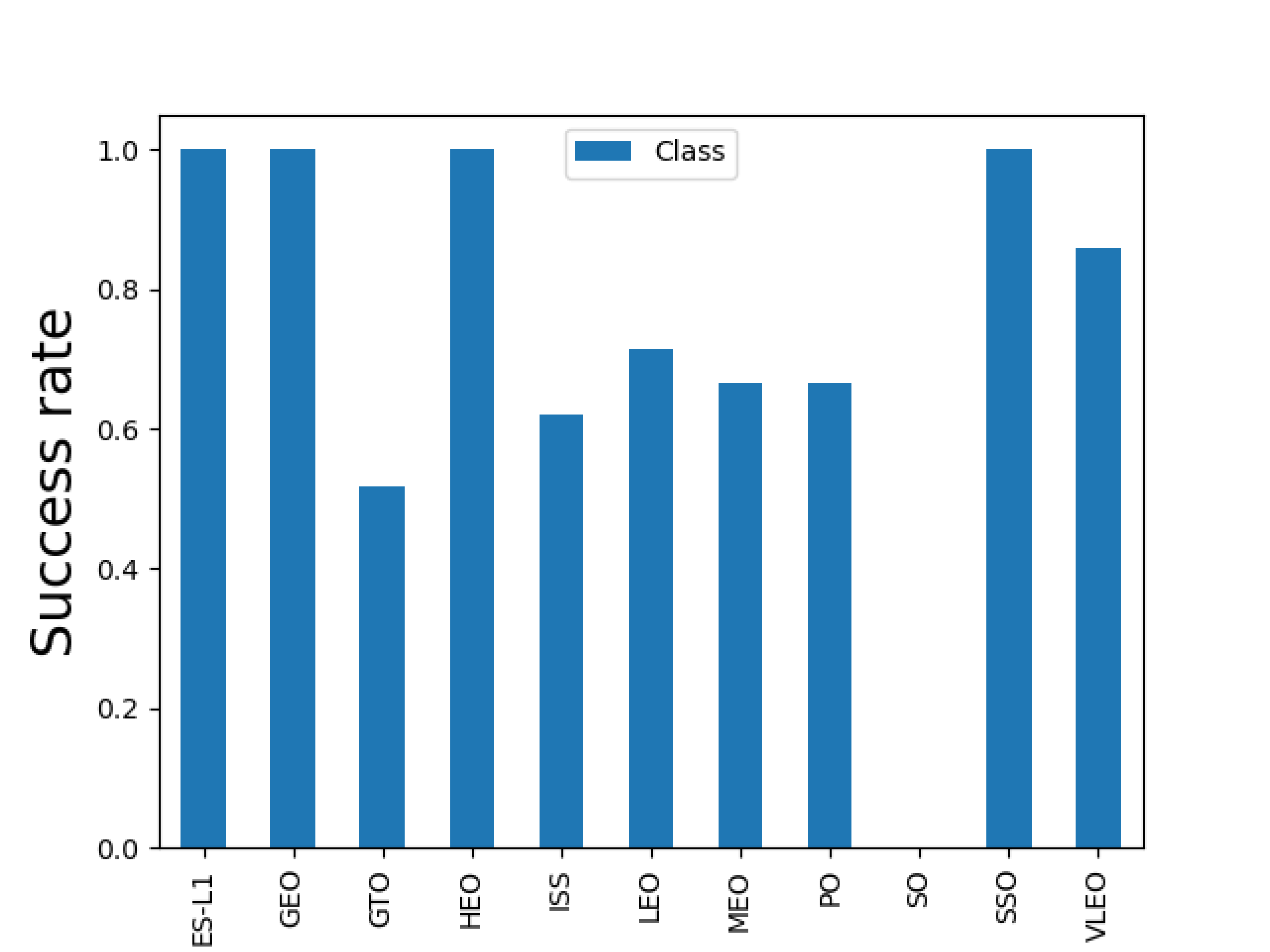


- Higher flight number – higher success of landing

Payload vs. Launch Site

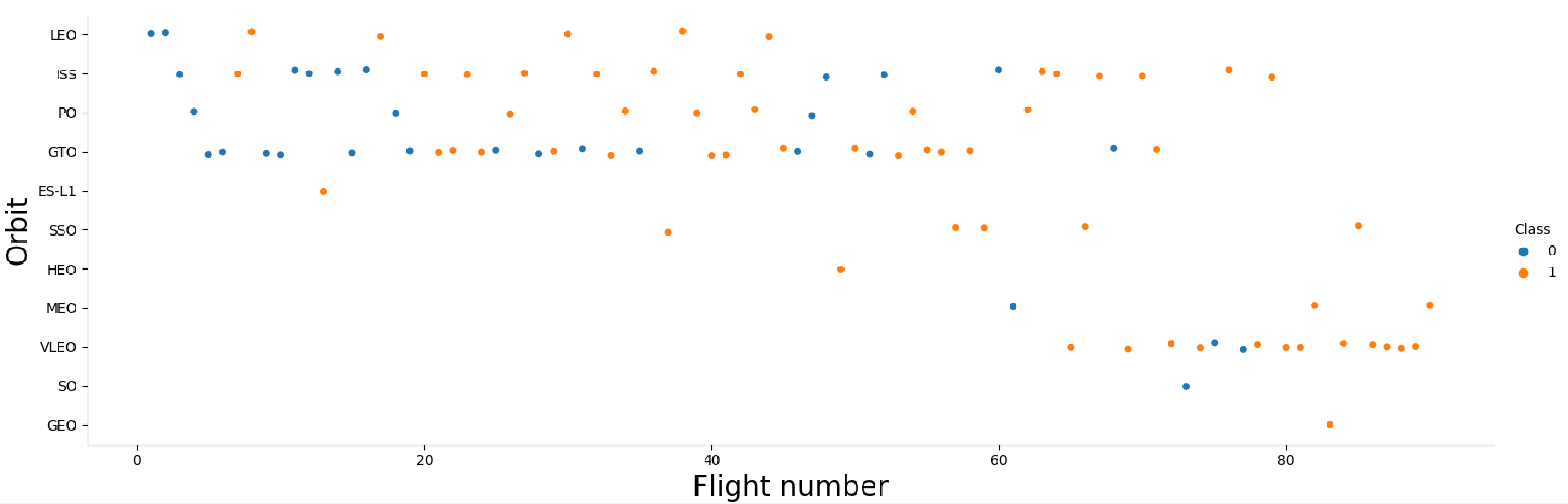
- For the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000)
- For payload more than 8000 kg there is a high chance of successful landing

Success Rate vs. Orbit Type



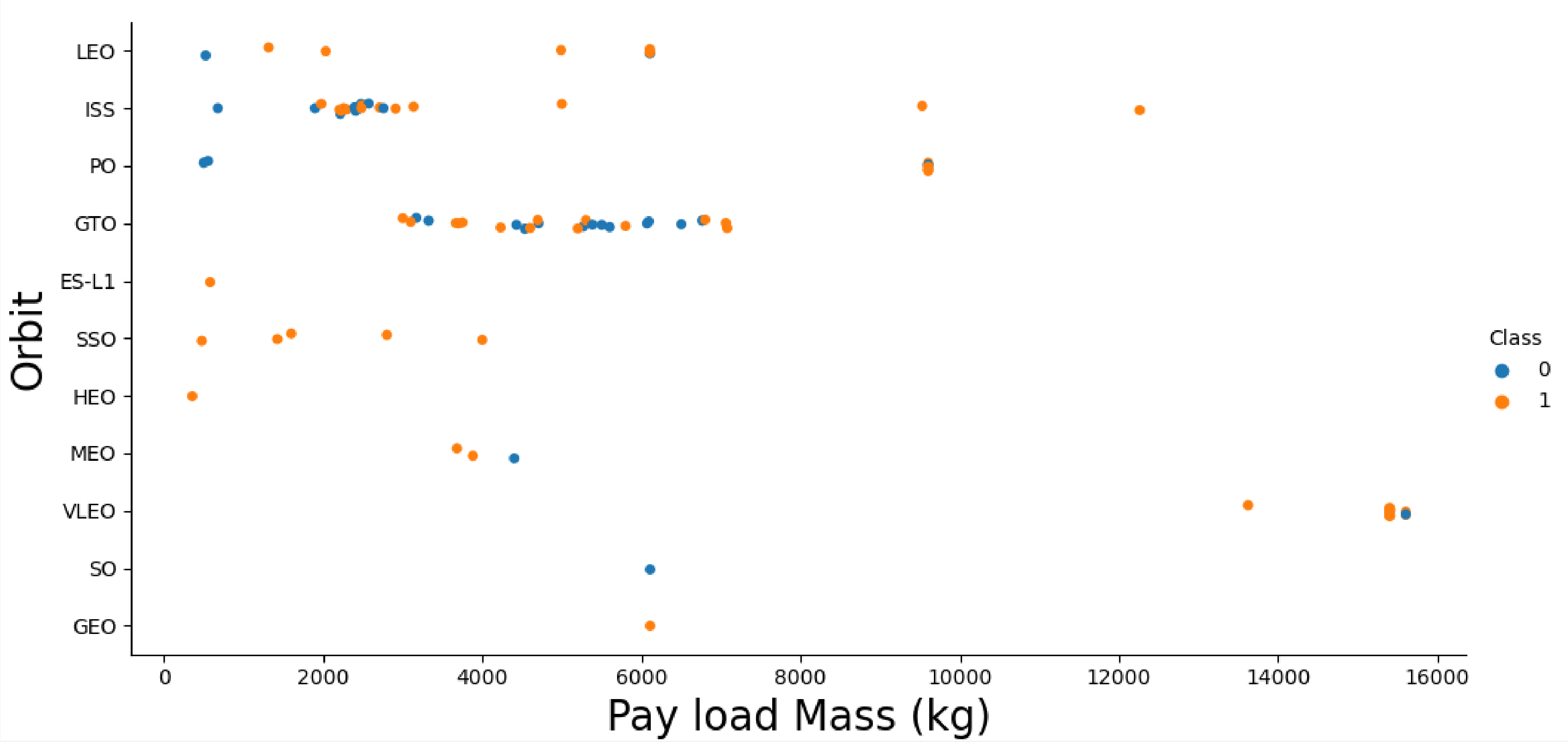
- The highest success rates of landing has several orbits: ES-L1, GEO, HEO, SSO.
- The most unsuccessful orbit is SO

Flight Number vs. Orbit Type



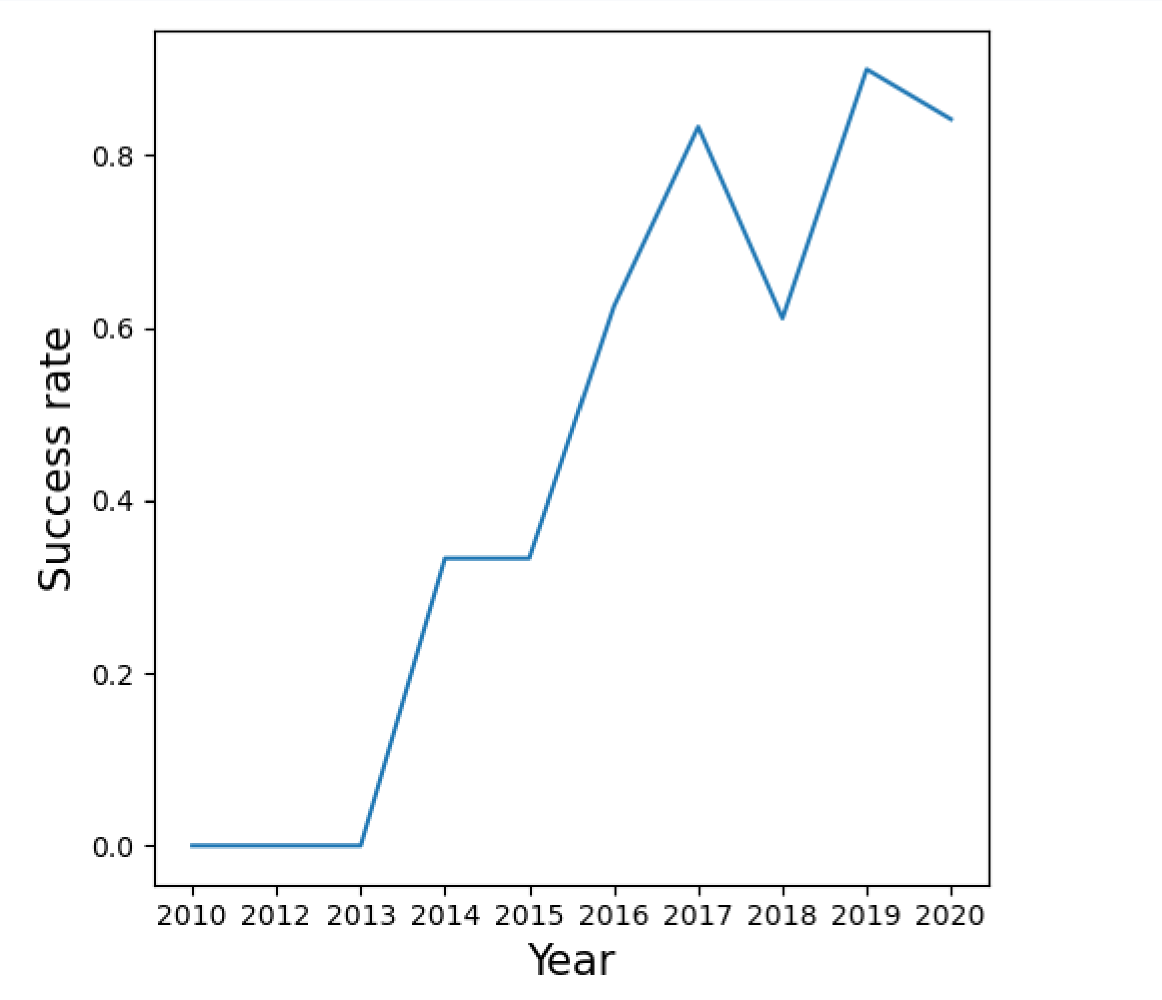
- In the LEO orbit the Success appears related to the number of flights
- There is no relationship between flight number when in GTO orbit

Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS
- For GTO we cannot distinguish this well as both positive landing rate and negative landing are both there here

Launch Success Yearly Trend



- The success rate since 2013 kept increasing till 2020

All Launch Site Names

The names of the unique launch sites:

%sql SELECT DISTINCT(Launch_Site) FROM SPACEXTBL

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

There are four launch sites, that are presented on the left.

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outc
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (paracl
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (paracl
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No atte
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No atte
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No atte

Five records where launch sites begin with `CCA`:

%sql SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5

Launch site CCAFS LC-40 was likely to be used in the early 2010

Total Payload Mass

The total payload carried by boosters from NASA:

%sql SELECT SUM(PAYLOAD_MASS__KG_) as Total_payload_mass FROM SPACEXTBL WHERE Customer = "NASA (CRS)"

Total_payload_mass
45596.0

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1:

%sql SELECT AVG(PAYLOAD_MASS__KG_) as Average_payload_mass
FROM SPACEXTBL WHERE Booster_Version LIKE '%F9 v1.1%’

Average_payload_mass
2534.6666666666665

First Successful Ground Landing Date

The date of the first successful landing outcome on ground pad:

%sql SELECT MIN(DATE) FROM SPACEXTBL WHERE Landing_Outcome LIKE '%ground%'

MIN(DATE)
01/08/2018

Successful Drone Ship Landing with Payload between 4000 and 6000

The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000:

%sql SELECT Booster_Version FROM SPACEXTBL WHERE Landing_Outcome LIKE '%drone%' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000

Booster_Version
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

The total number of successful and failure mission outcomes:

```
%sql SELECT Mission_Outcome, COUNT(Mission_Outcome) AS Total_number_of_outcome \
    FROM SPACEXTBL \
    GROUP BY Mission_Outcome
```

Mission_Outcome	Total_number_of_outcome
None	0
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

The name of the booster which have carried the maximum payload mass:

```
%sql SELECT Booster_Version, Total_payload \
FROM \
    (SELECT Booster_Version, SUM(PAYLOAD_MASS__KG_) as Total_payload \
     FROM SPACEXTBL \
     GROUP BY Booster_Version) \
ORDER BY Total_payload DESC \
LIMIT 1
```

Booster_Version	Total_payload
F9 B5 B1048.4	15600.0

2015 Launch Records

List of the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015:

```
%sql SELECT substr(Date, 4, 2) AS Month, Landing_Outcome, Booster_Version, Launch_Site \
FROM SPACEXTBL \
WHERE Landing_Outcome = 'Failure (drone ship)' AND substr(Date,7,4)='2015'
```

Month	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	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:

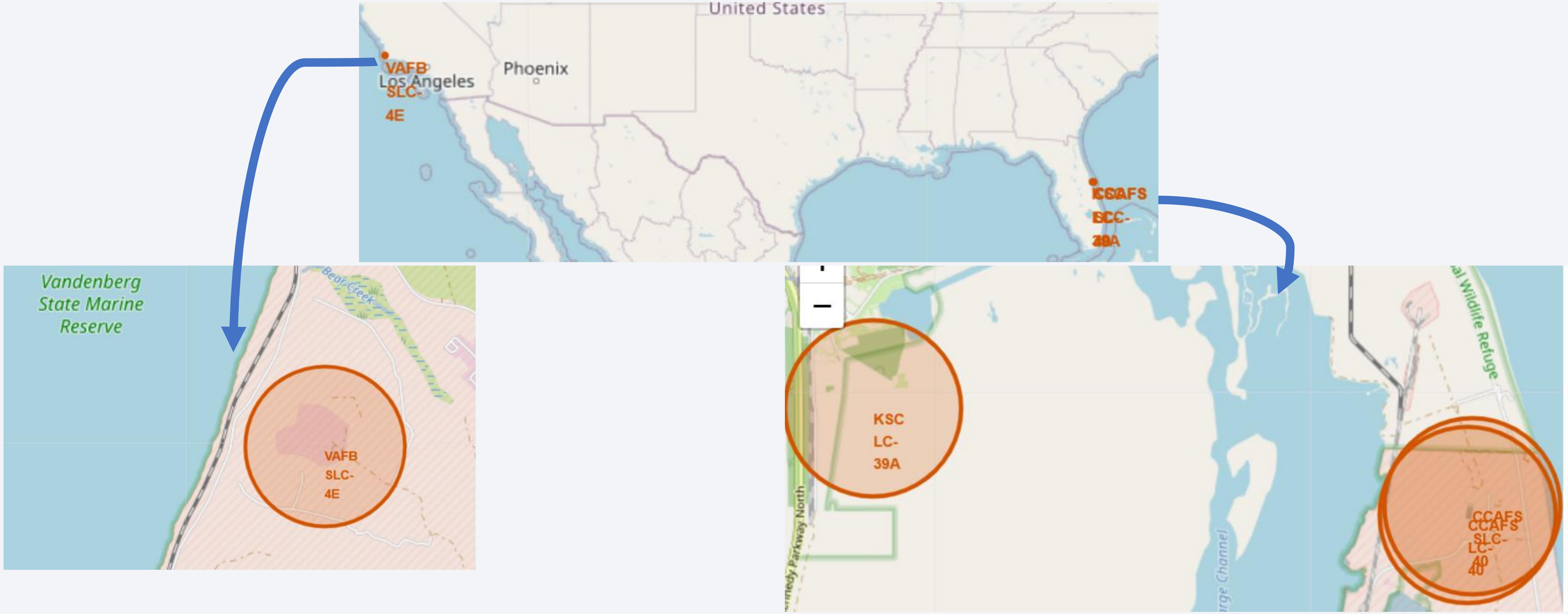
```
%sql SELECT COUNT(Landing_Outcome) AS Count, Landing_Outcome \
FROM SPACEXTBL \
GROUP BY Landing_Outcome \
HAVING Landing_Outcome LIKE '%Success%' \
ORDER BY Count DESC
```

Count	Landing_Outcome
38	Success
14	Success (drone ship)
9	Success (ground pad)

Section 3

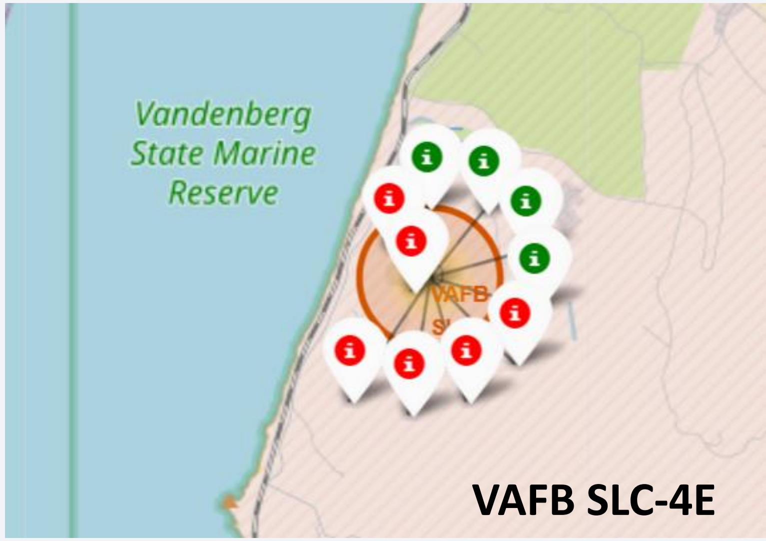
Launch Sites Proximities Analysis

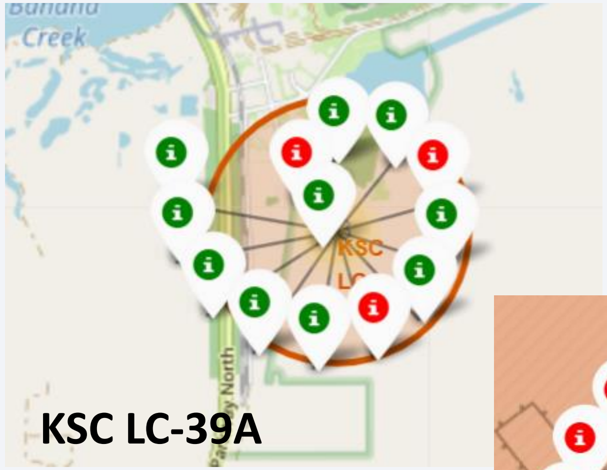
Launch sites' locations on world map

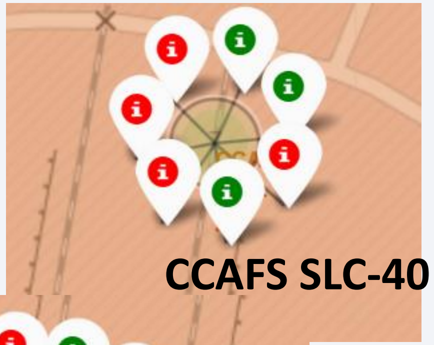


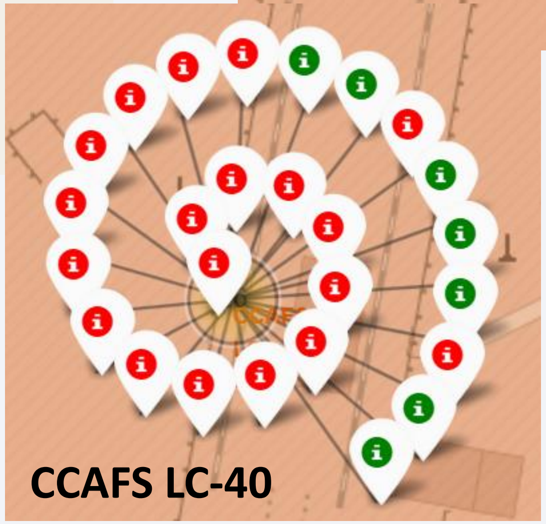
- Launch sites are located not in the proximity to the Equator line
- All sites are close to coastline

Launch sites' success rate of launches



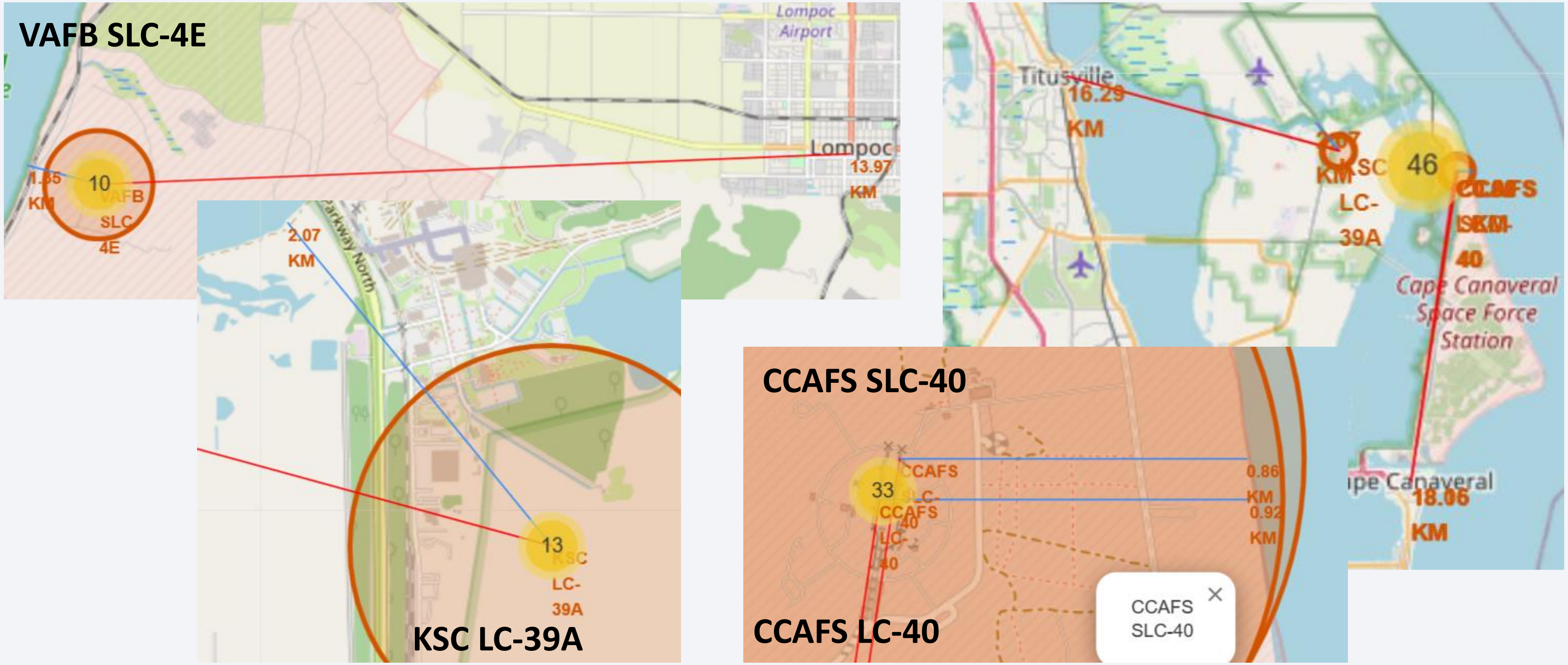






It can be seen that KSC LC-39A has the highest success rate

Launch sites' proximities to coastline and cities

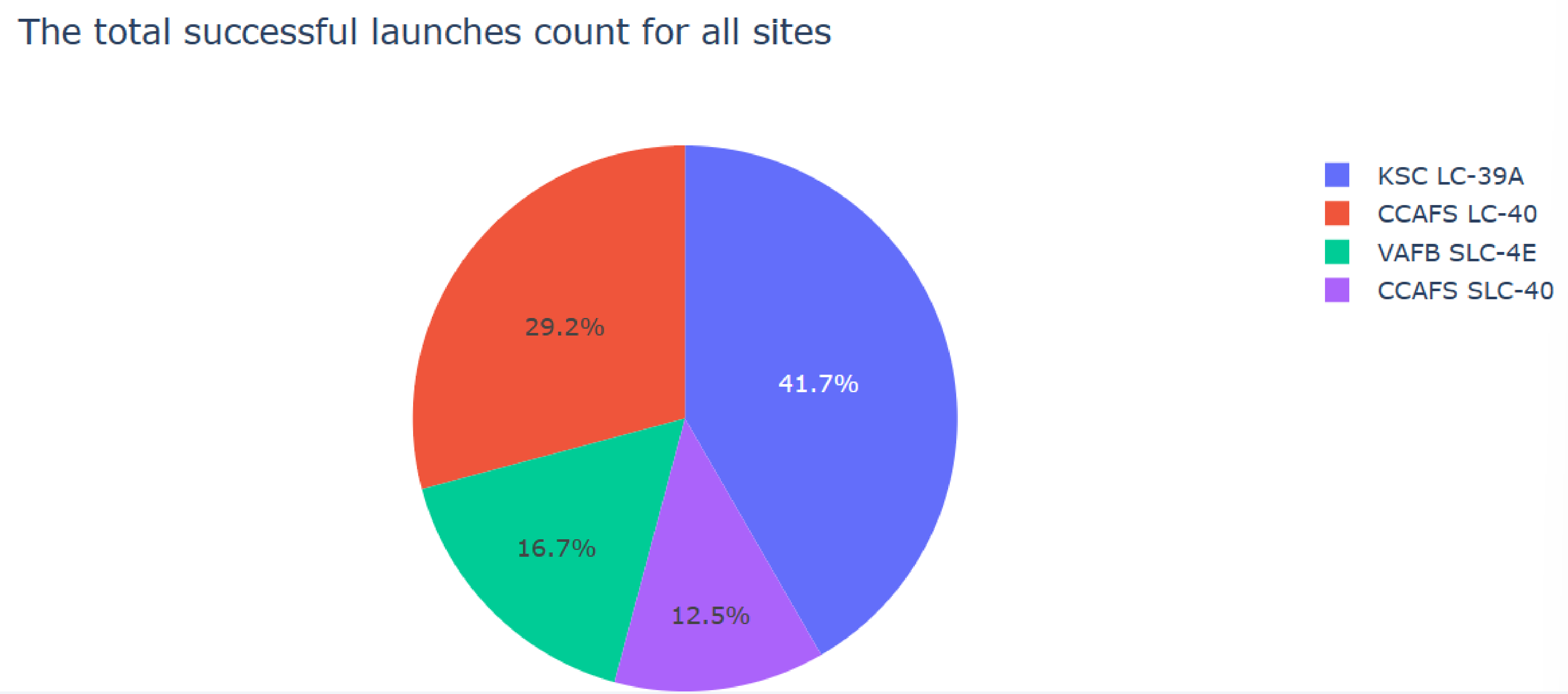


- All launch sites in close proximity to coastline and railways
- Launch sites keep certain distance away from cities

Section 4

Build a Dashboard with Plotly Dash

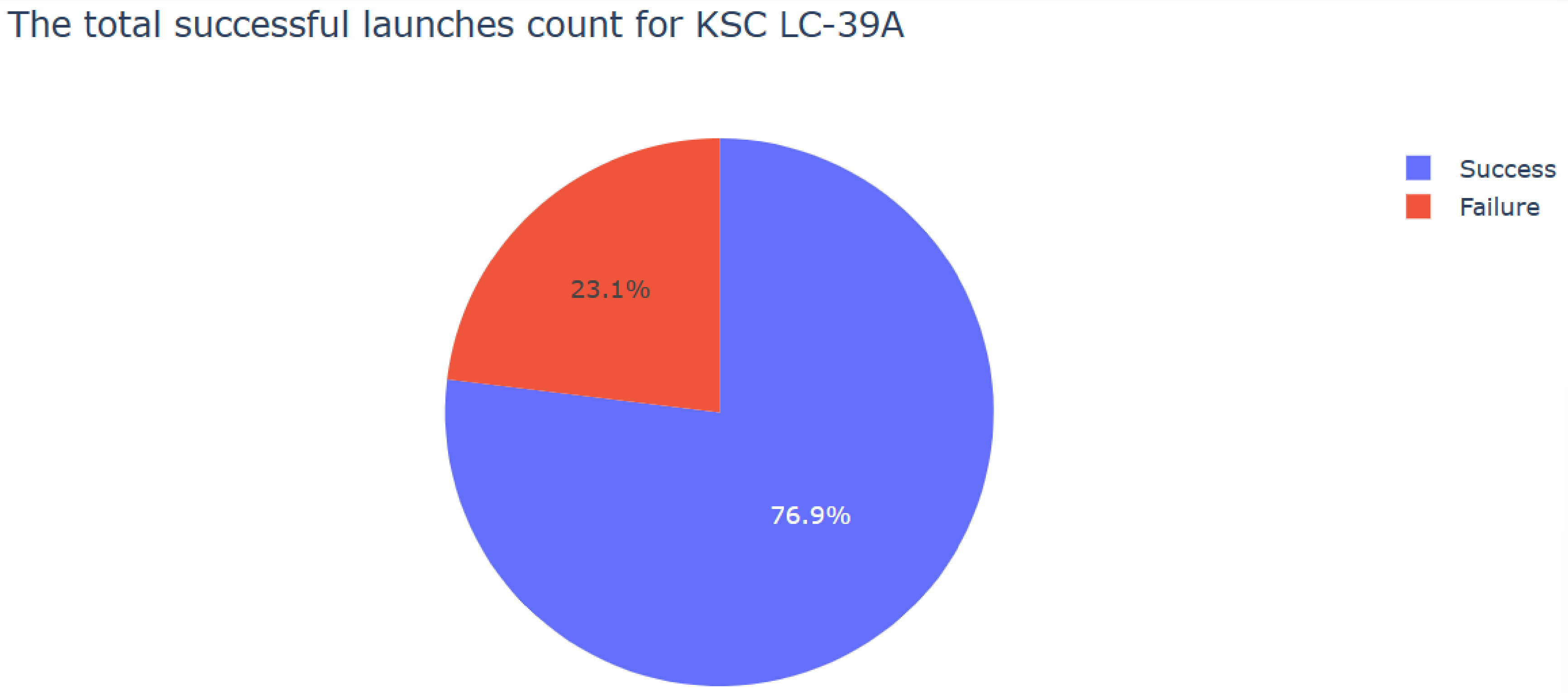
The total successful launches count for all sites



The most successful launch site – KSC LC-39A

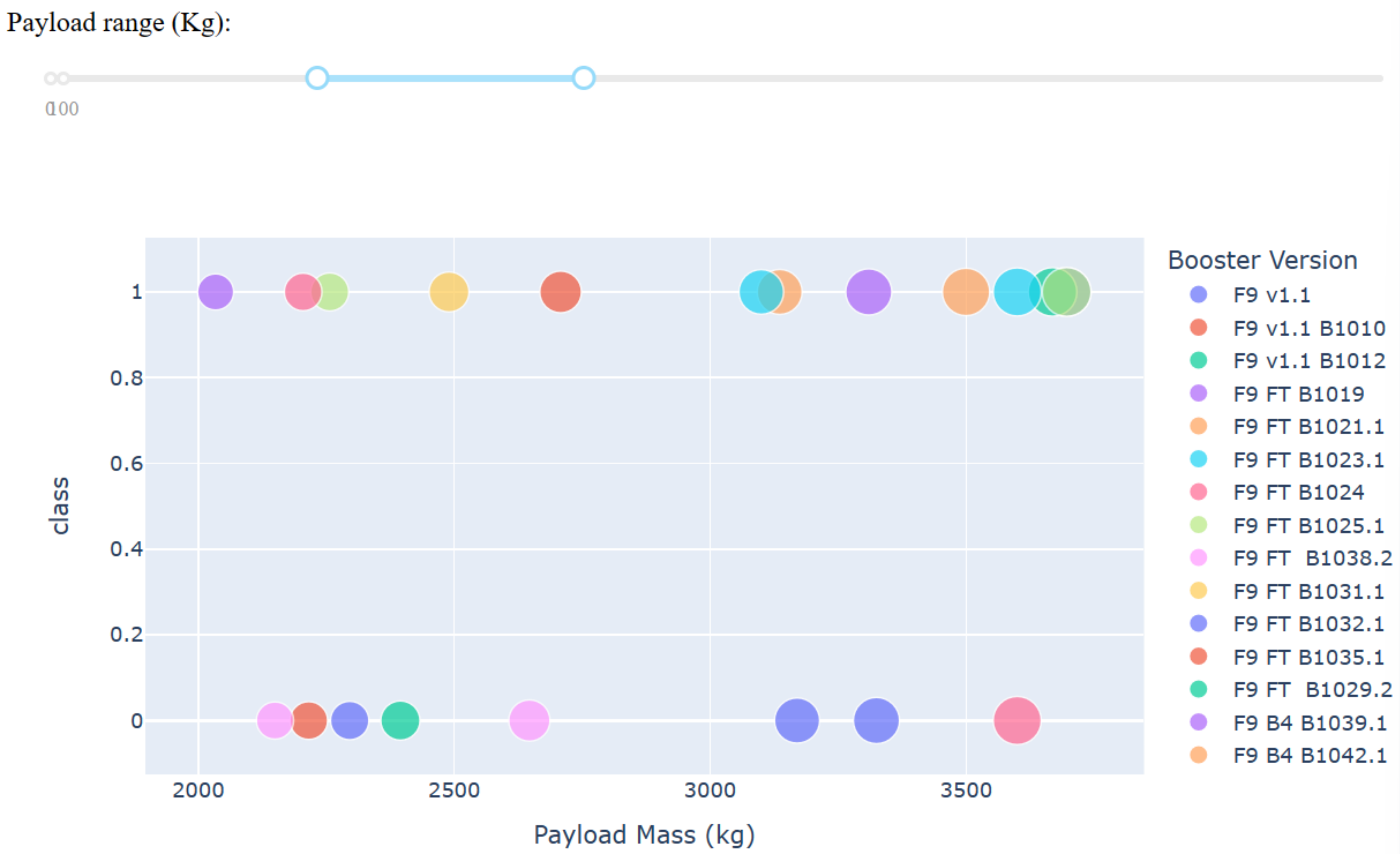
The least – CCAFS SLC-40

The KSC LC-39A's total successful launches count



Almost ¾ of all launches had positive outcome

Payload vs. Launch Outcome scatter plot for all sites

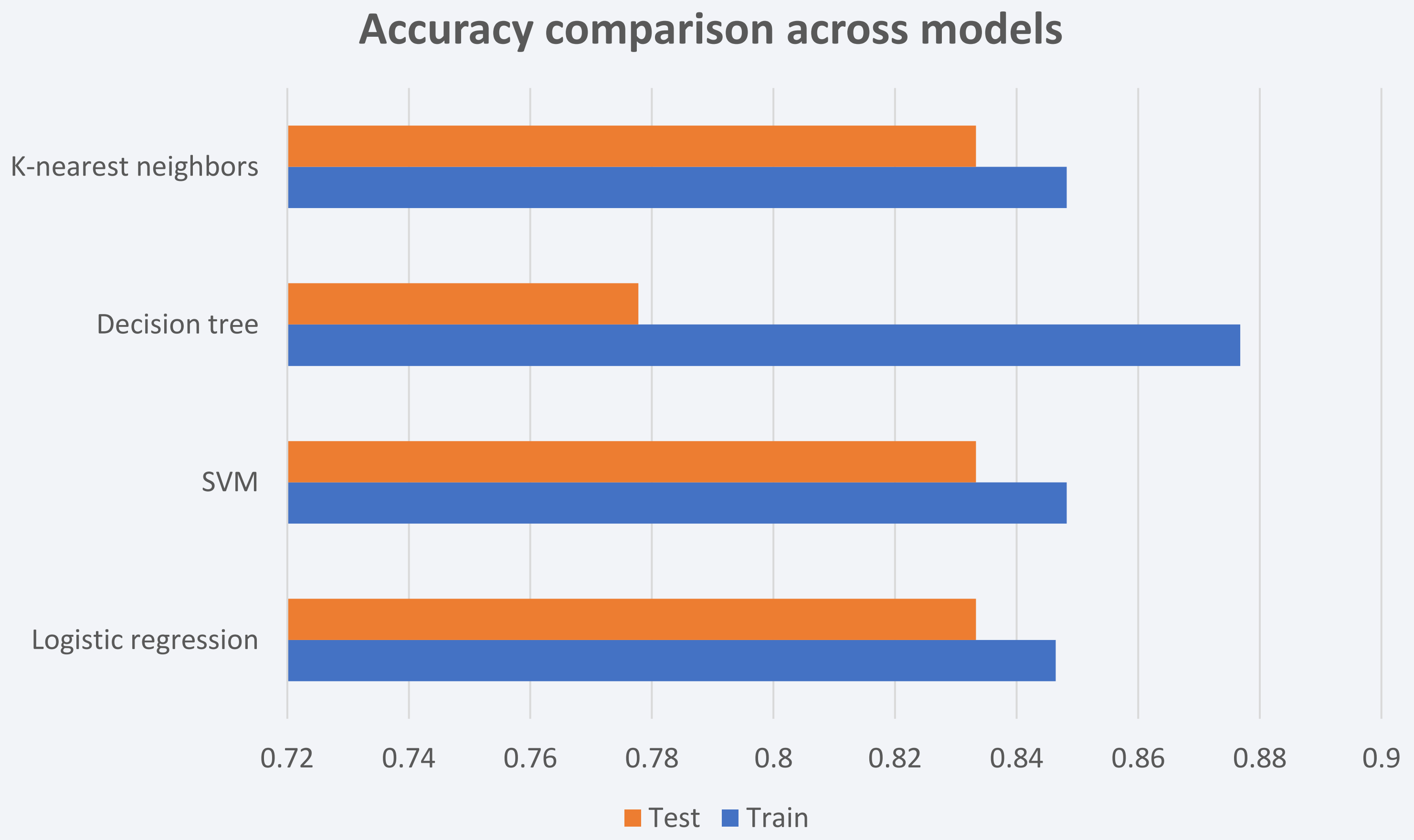


- Payload range from 2000 to 4000 kg has the highest success rate

Section 5

Predictive Analysis (Classification)

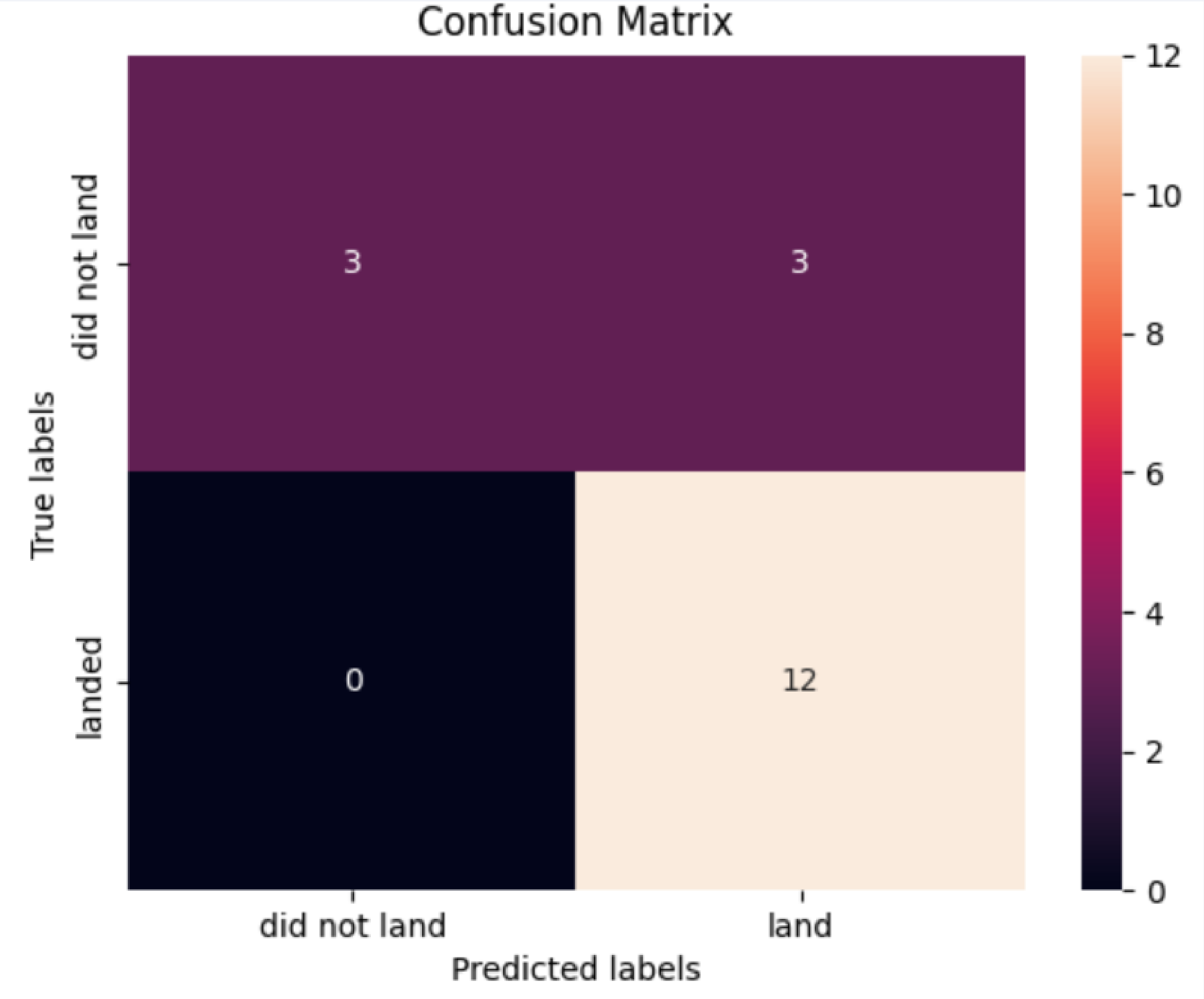
Classification Accuracy



According to accuracy based on train data, the best model is Decision tree. But that model has the lowest test data accuracy.

Other models are quite similar in terms of rest and train data accuracy, so we chose the simplest model – Logistic regression.

Confusion Matrix



Confusion matrix for logistic regression model

It has high probability to predict positive outcome but has problem with False positive outcome.

Conclusions

- The best model to predict landing outcome – logistic regression
- If launch is from KSC LC-39A launch site, then there is quite high chance to have successful landing
- If payload mass is in range between 2000 and 4000 kilograms, there is a possibility to successfully retrieve the first stage
- The most suitable for positive outcome launches are orbits ES-L1, GEO, HEO and SSO, as they have the highest successful landing

Appendix

See GitHub - Aleninok-pixel/Capstone_Coursera (github.com)

Thank you!