IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Patricio Robles
1.15.2023

Outline

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

Executive Summary

- In order to determine the probability of the first stage of a rocket landing back successfully data was obtained from SpaceX REST API. Exploratory Data analysis was done with visualizations in Python as well as SQL queries.
- An interactive Map was made with Folium to gain geographical insights of the Launch Site / Success rate relationship.
- A dashboard was created to visualize the success rate of each launch site (and in relation to each other) as well as the relationship between payload mass and success rate.
- A classification model was created and trained to determine the probability of a launch having a successful landing with an accuracy score of 83%

Introduction

- The space industry is thriving with many companies rising to meet the demand for rocket launches. One of the main players is SpaceX, who is able to launch rockets for $62 Million which is over $100 Million cheaper than the closest competitor. SpaceX is able to achieve this costs because they reuse the first stage of their rockets.

- If we can determine the probability of reusing the first stage in a rocket launch, we can determine the cost of a launch. This information can help a competing Space firm bid for contracts in this industry.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX flight Data was collected using their public API, as well as webscraping their Wikipedia page.
- Perform data wrangling
  - All relevant data from various SpaceX APIs was collected and then processed using Python Pandas. Data was filtered for Falcon 9 rocket and missing values were replaced with their column means.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

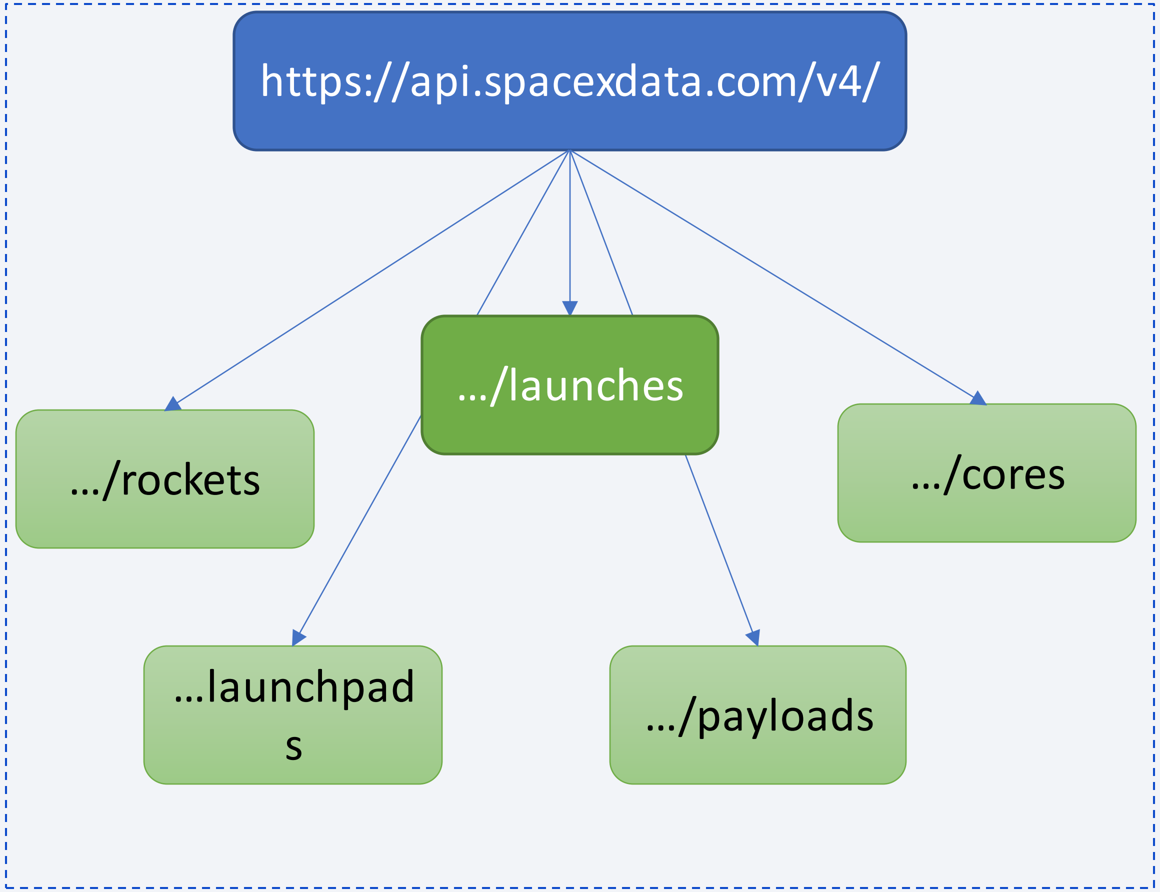
- Data sets were collected from the SpaceX REST API as well as web scraping related wiki pages using Beautiful Soup. Details about the specific flow of each Data Collection pipeline is described in the following slides.

Data Collection – SpaceX API

- Data was collected through various databases from the SpaceX REST API

- Data Collection Code

https://github.com/patorobles/Coursera-Data-Science-Capstone/blob/main/Collecting%20the%20DataAPI.ipynb

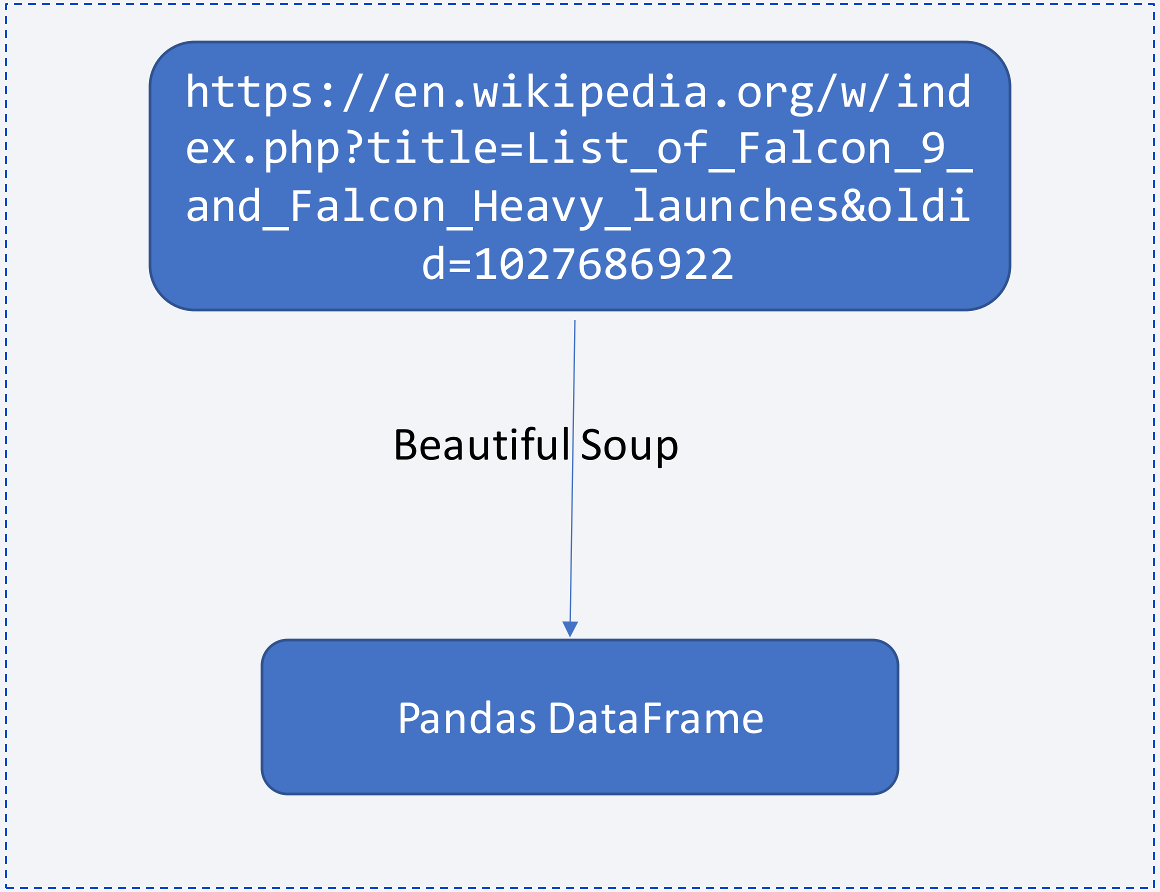


Data Collection - Scraping

- Data was also scraped from Falcon 9 related wiki pages using the beautiful Soup libraries

- Web Scraping code

https://github.com/patorobles/Coursera-Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Data was first retrieved from an API and then processed in the following way.

1) ID references in /launches Data were imported from other API calls to complete de Dataset

2) Data was filtered to only include desired launched (Falcon 9 Rockets)

3) Missing data was replaced for its mean value

4) One hot encoding was used to work with categorical variables in our Dataset

- https://github.com/patorobles/Coursera-Data-Science-Capstone/blob/main/Data%20Wrangling.ipynb

EDA with Data Visualization

- The relationship between **Flight Number and Launch Site** was visualized. Through it it was determined that the **CCAFS LC-40** Launch Site was the only one used for the first ~20 launches, after which other sites with higher success rates were used too.
- The relationship between **Payload Mass and Launch Site** was visualized and we found the VAFB-SLC site has no launches for a mass greater than 10,000 kg
- The relationship between the **success rate of each orbit** was visualized and it was found that ES-L1, GEO, HEO and SSO orbits have the best success rates
- The relationship between **Flight Number and Orbit type** was visualized but no conclusive relationships appeared
- The relationship between **Payload and Orbit Type** was visualized and it was determined that heavier landings were more successful on the Polar, Leo and ISS orbits.
- The **yearly launch success trend** was visualized which showed that over time, the likelihood of landings being successful has increased dramatically

- https://github.com/patorobles/Coursera-Data-Science-Capstone/blob/main/ML%20Prediction%20Models.ipynb

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
  - Names of unique launch sites
  - 5 records of launch site beginning with 'CCA'
  - Total Payload Mass carried by boosters launched by NASA
  - Average payload mass carried by Falcon 9 booster
  - Date of first successful landing in ground pad
  - Names of Boosters with successful drone ship landings and payload mass between 4000 and 6000
  - Total number of successful and failed missions
  - Names of Boosters that have carried the maximum payload mass
  - Failed outcomes in droneship in 2015
  - Landing outcomes by type between 2010 and 2017
- https://github.com/patorobles/Coursera-Data-Science-Capstone/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- An interactive Map using Folium was created to facilitate visualization of all launch sites, their success/failed attempts as well as proximity to nearby objects. For this a Map object was created and Markers, Circles, Marker Clusters and Lines were used to show the desired information.

- https://github.com/patorobles/Coursera-Data-Science-Capstone/blob/main/Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- A Plotly Dash dashboard was created to show the relationship between Successful launches and Launch sites, as well as the correlation between payload mass and success with launch sites.
- The interactive dashboard lets stakeholders explore these relationships by individual launch site (or all of them) as well as by a range of payload mass. This interactivity gives freedom to the stakeholder to support his findings.

- https://github.com/patorobles/Coursera-Data-Science-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Data was standardized and divided into a training and testing set in order to create ML models to predict whether the landing is successful or not.
- A GridSearchCV object was fit and scored on each of the following **model types** to determine its best parameters
  - Logistic Regression 83% accuracy
  - Support Vector Machine 83% accuracy
  - Decision Tree 83% accuracy
  - K Nearest Neighbors 83% accuracy

  It was determined that for the small sample size all models had the same accuracy rates
- https://github.com/patorobles/Coursera-Data-Science-Capstone/blob/main/ML%20Prediction%20Models.ipynb

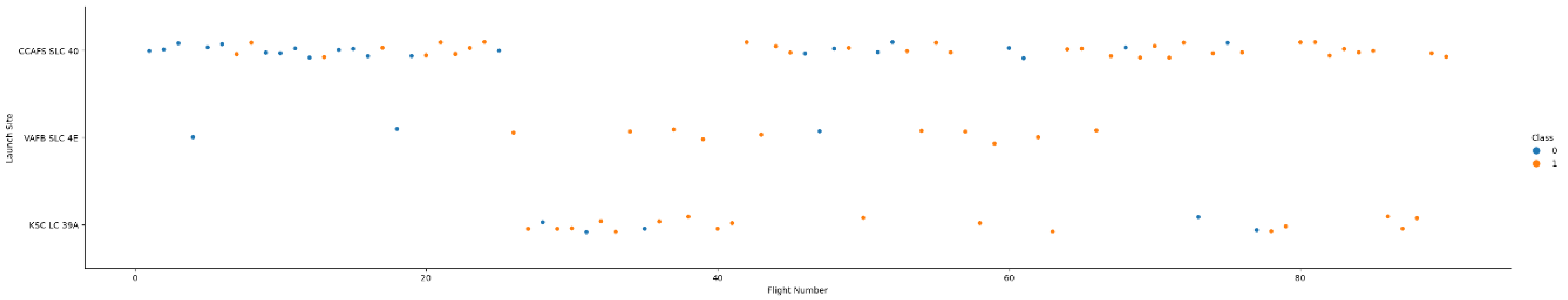
Results

- The following slides will share the results obtained in these phases of the project
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

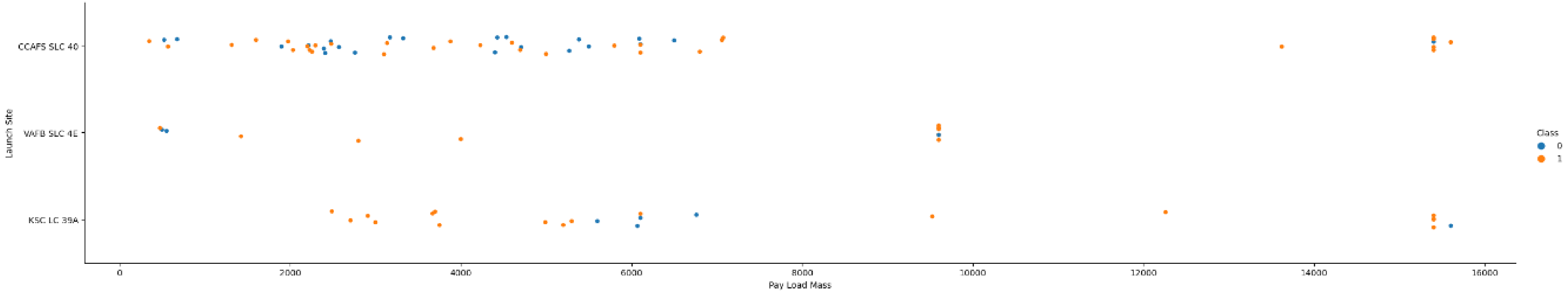
Flight Number vs. Launch Site



Now try to explain the patterns you found in the Flight Number vs. Launch Site scatter point plots.

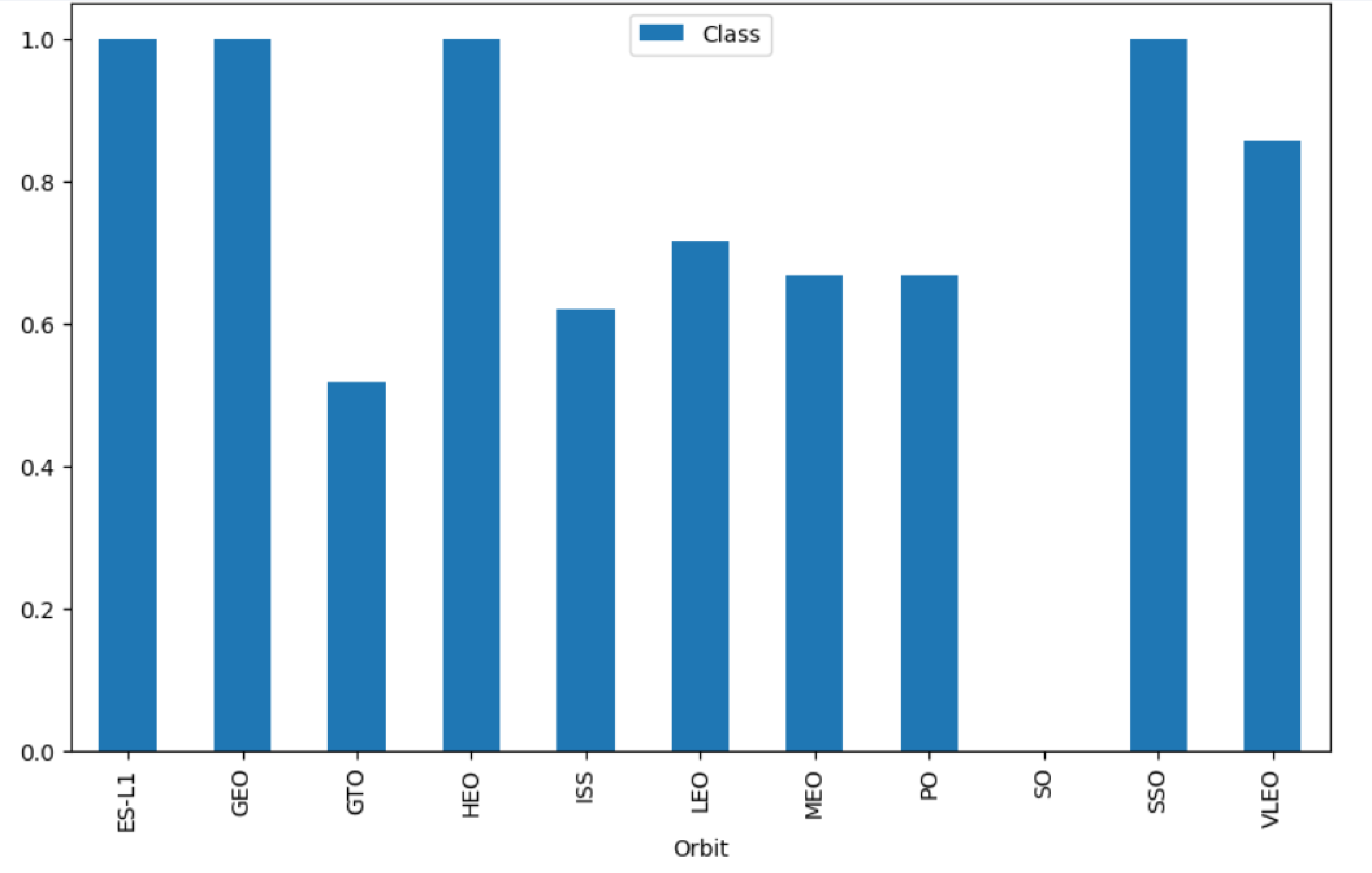
Most of the early launches were from CCAFS and it was more common to have a failed than successful landing. KSC starts being used as a launch site around the 25~ flight and after that launches are sent from all sites with a much more succesful rate.

Payload vs. Launch Site

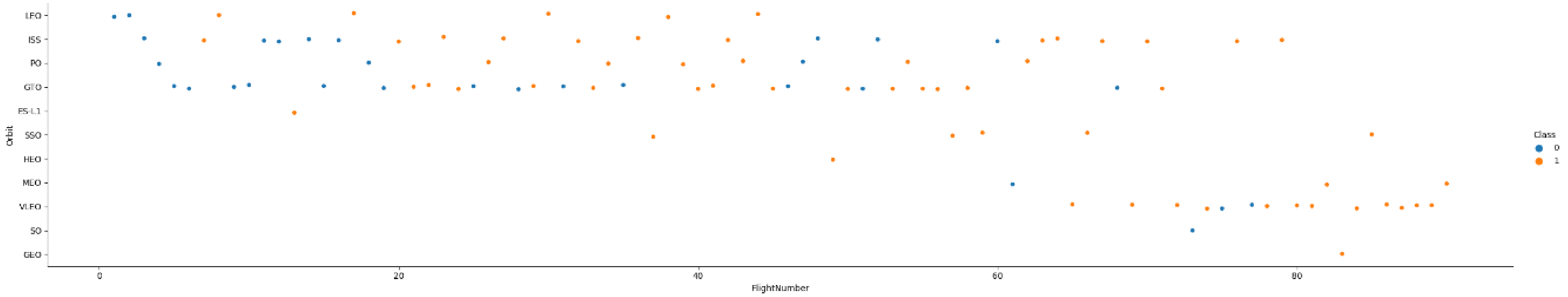


Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type

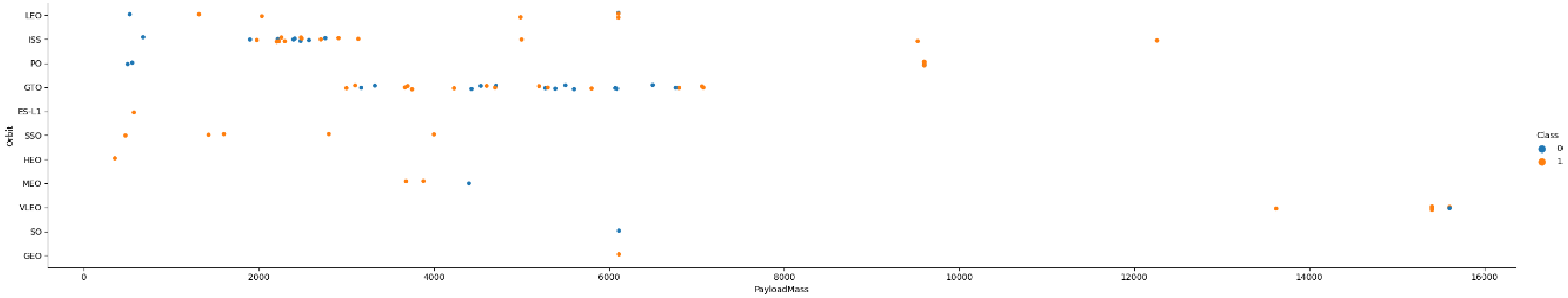


Flight Number vs. Orbit Type



You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

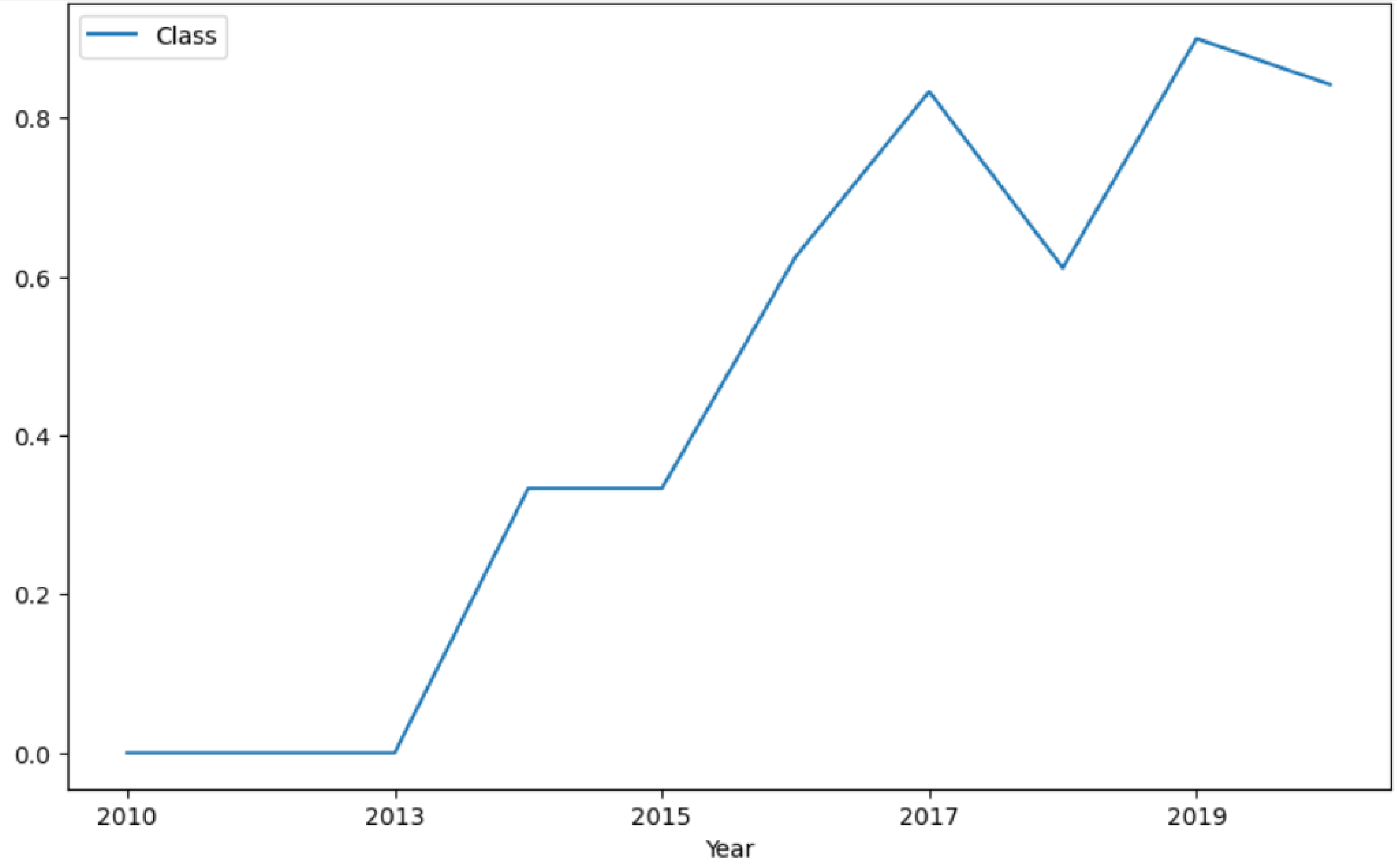
Payload vs. Orbit Type



With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend



you can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

- Find the names of the unique launch sites
- Present your query result with a short explanation here

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

```
%sql select distinct(launch_site) from space
```

```
 * ibm_db_sa://tnt93877:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32304/bludb
Done.
```

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`
- Present your query result with a short explanation here

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

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

```
 * ibm_db_sa://tnt93877:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32304/bludb
Done.
```

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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)
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-12	22:41:00	F9 v1.1	CCAFS LC-40	SES-8	3170	GTO	SES	Success	No attempt

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- Present your query result with a short explanation here

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

```
%sql select sum(PAYLOAD_MASS__KG_) from Space where customer= 'NASA (CRS)'
```

```
 * ibm_db_sa://tnt93877:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32304/bludb
Done.
```

1
22007

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Present your query result with a short explanation here

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

```
%sql select sum(PAYLOAD_MASS__KG_) from Space where booster_version= 'F9 v1.1'
```

```
 * ibm_db_sa://tnt93877:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32304/bludb
Done.
```

1
11030

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Present your query result with a short explanation here

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

Hint:Use min function

```
%sql select min(date) from space where landing__outcome='Success (ground pad)'
```

```
 * ibm_db_sa://tnt93877:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32304/bludb
Done.
```

1
2017-01-05

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

- Present your query result with a short explanation here

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 space where landing__outcome='Success (drone ship)' AND payload_mass__kg_ between 4000 and 6000
```

```
 * ibm_db_sa://tnt93877:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32304/bludb
Done.
```

booster_version
F9 FT B1022
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Present your query result with a short explanation here

List the total number of successful and failure mission outcomes

```
%sql select count(mission_outcome), mission_outcome from space group by mission_outcome
```

```
 * ibm_db_sa://tnt93877:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32304/bludb
Done.
```

1	mission_outcome
44	Success
1	Success (payload status unclear)

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Present your query result with a short explanation here

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

```
%sql select booster_version, payload_mass__kg_ from space \
where payload_mass__kg_ = (select max(payload_mass__kg_) from space)
```

```
 * ibm_db_sa://tnt93877:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32304/bludb
Done.
```

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600

2015 Launch Records

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

- Present your query result with a short explanation here

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

```
%sql select booster_version, launch_site, LANDING__OUTCOME, date from space where LANDING__OUTCOME = 'Failure (drone ship)' and year(date) = 2015
```

```
 * ibm_db_sa://tnt93877:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32304/bludb
Done.
```

booster_version	launch_site	landing__outcome	DATE
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)	2015-10-01

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

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), landing__outcome from space \
where date between '2010-06-04' and '2017-03-20' \
group by landing__outcome \
order by count(landing__outcome) desc
```

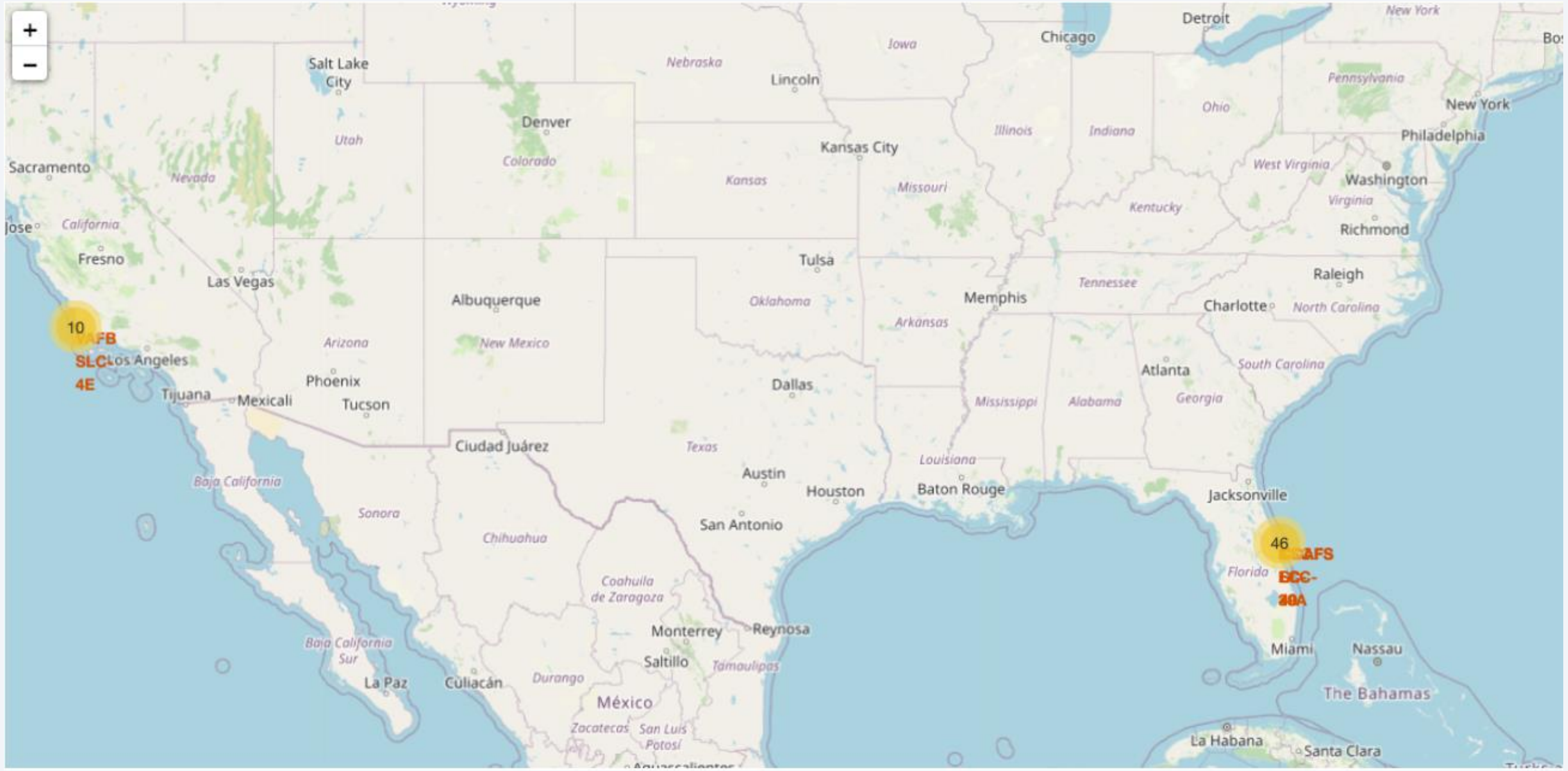
```
 * ibm_db_sa://tnt93877:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32304/bludb
Done.
```

1	landing__outcome
7	No attempt
2	Failure (drone ship)
2	Success (drone ship)
2	Success (ground pad)
1	Controlled (ocean)
1	Failure (parachute)

Section 3

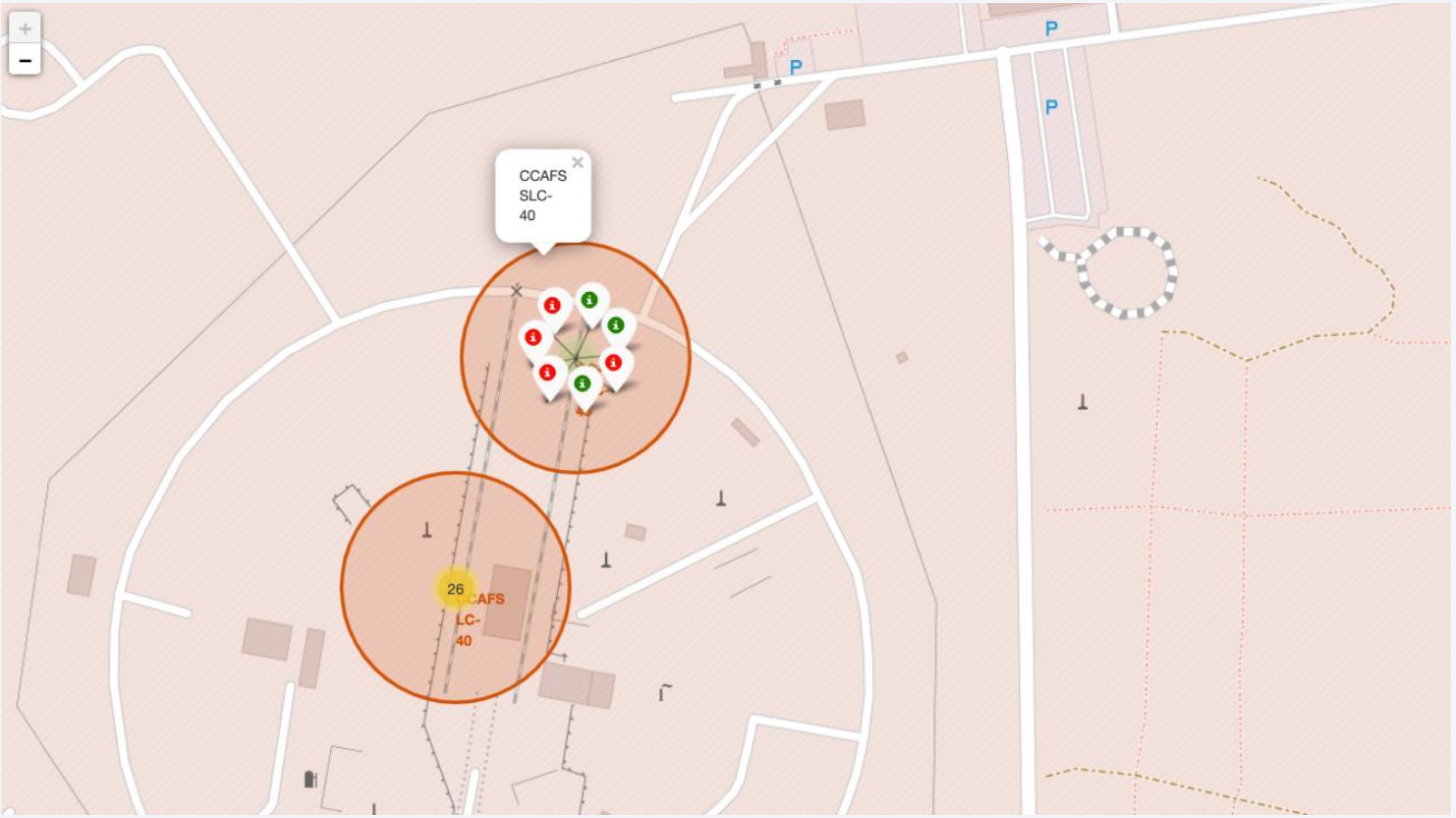
Launch Sites Proximities Analysis

Launch Sites Map



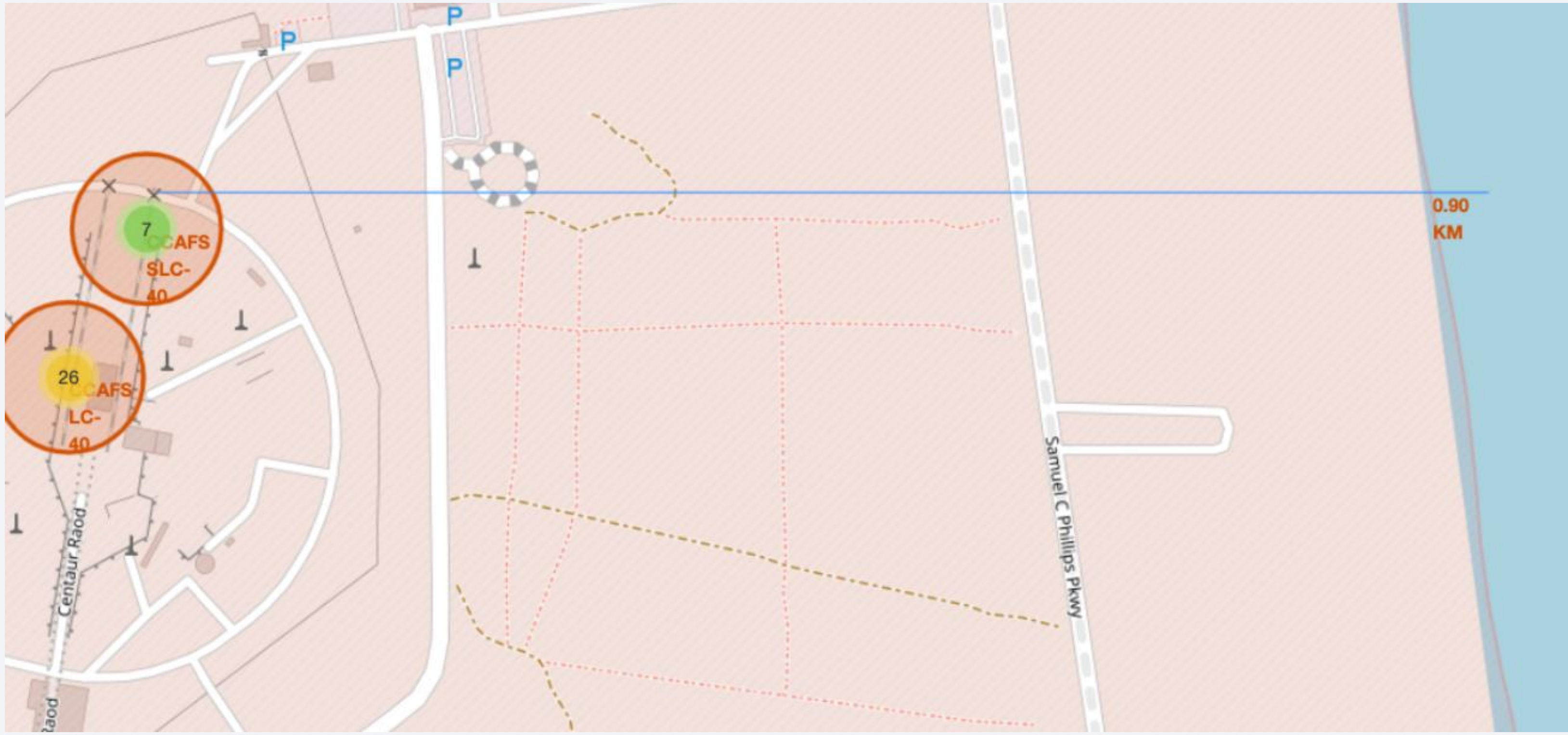
- We can observe the location of the Launch sites as well as the number of launches at each location. It is important to note how the sites are always near the coast and tropics.

Launch Outcomes



- The following screenshot shows us the detailed launch outcomes for the CCAFS SLC Site

Distance to Coastline

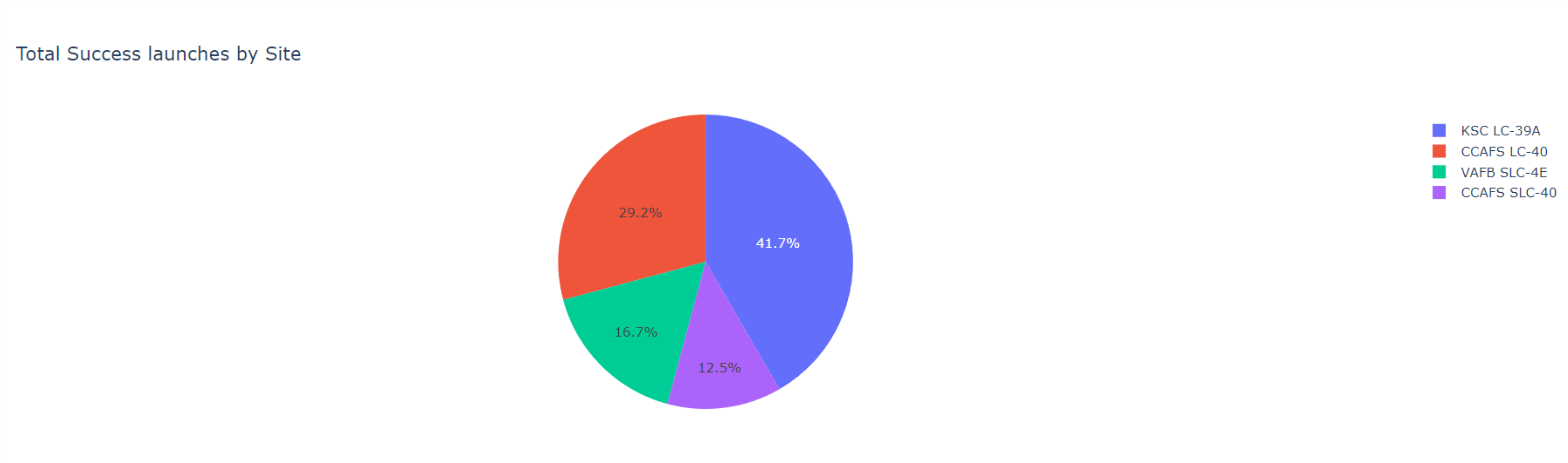


- This map screenshot shows the distance from the CCAFS SLC Site and its nearest coastline

Section 4

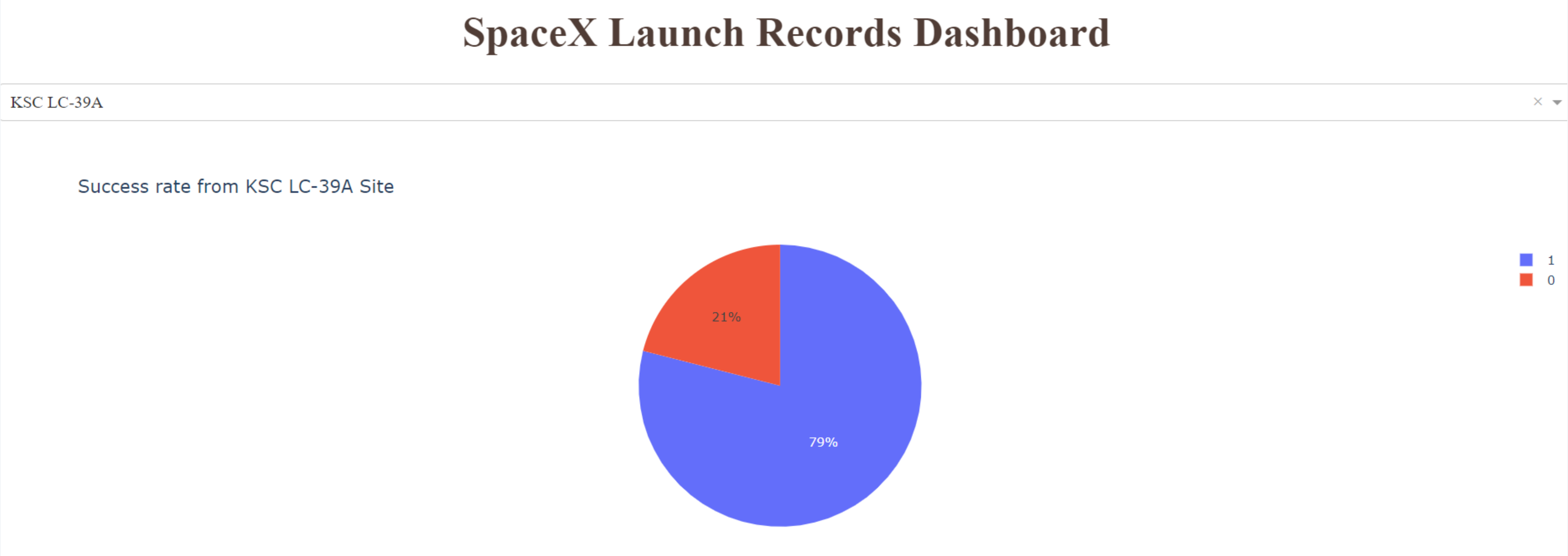
Build a Dashboard with Plotly Dash

Succes Count for all Sites



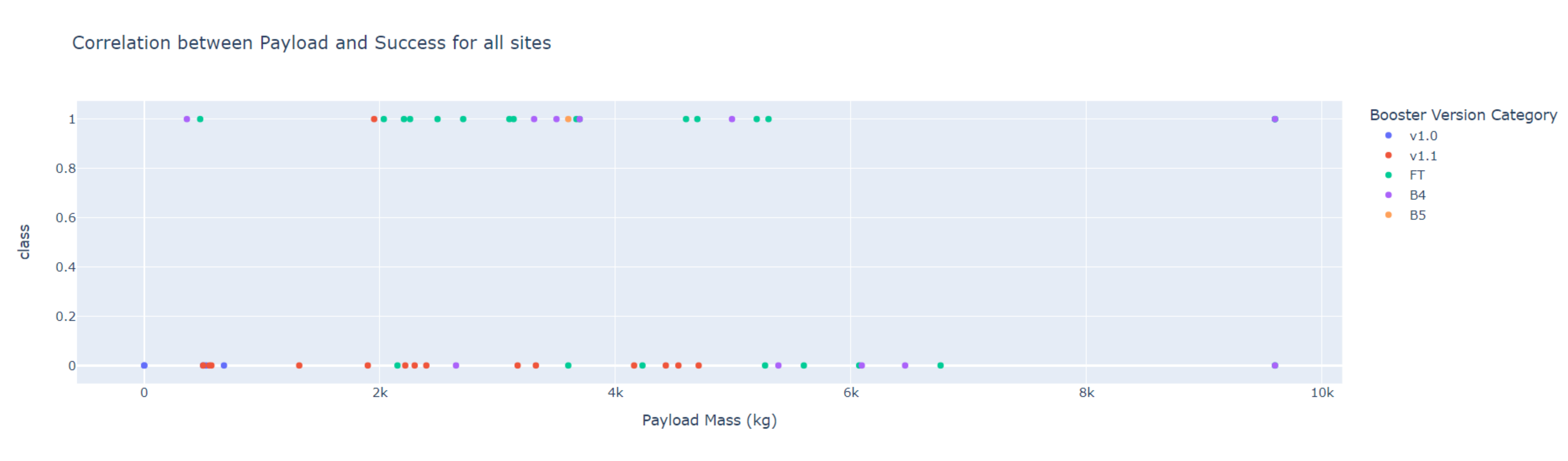
- We can see the success rate of each site as a fraction of all sites. With this we can easily see which site is more likely to have a successful landing

Success Rate from KSC LC-39A Site



- This chart shows the success rate of the most successful launch site

Payload Mass vs launch outcome



- We can observe that launches with a Payload Mass Greater than 6000kg are more likely to be unsuccessful

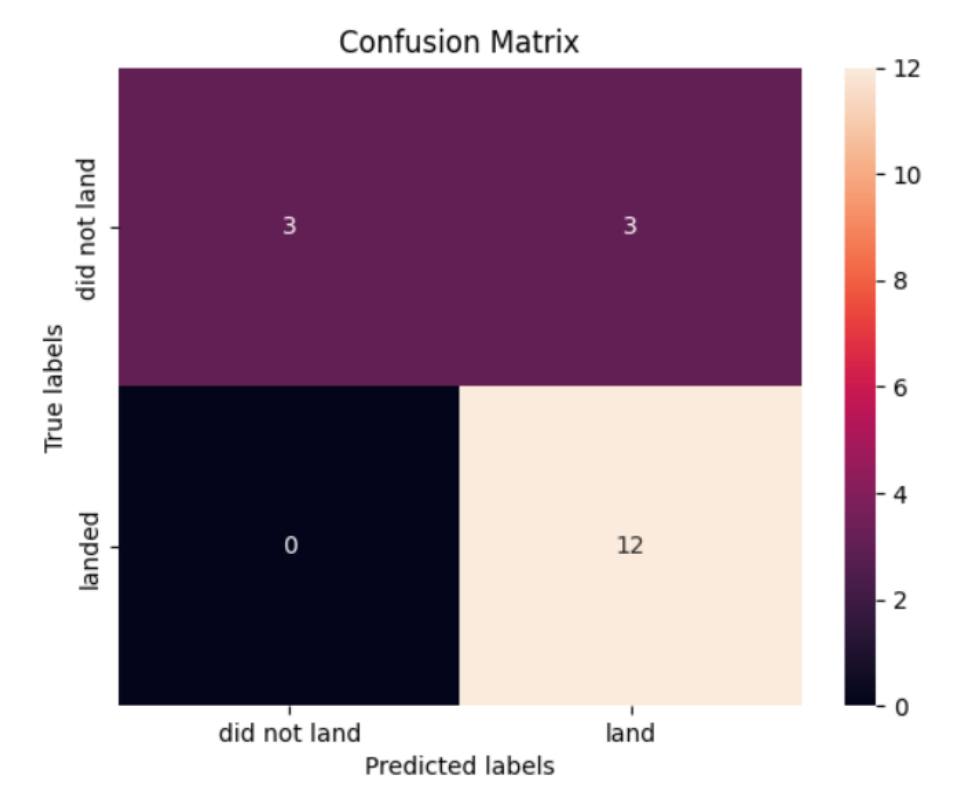
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Visualize the built model accuracy for all built classification models, in a bar chart
- Find which model has the highest classification accuracy

Confusion Matrix



- The same confusion matrix was obtained on all models, predicting 15/18 test values correctly for an accuracy of 83% The only error was a 17% chance of false positives.

Conclusions

- All of the models analyzed (Logistic Regression, Support Vector Machine, Decision Tree and KNN) had the same 83% accuracy score
- It is believed all models have the same score because the test and train samples are quite small.
- An analysis of the full dataset would provide larger differences between the models
- The Logistic Regression model is preferred because of its simplicity and equal accuracy to more complex models.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!