IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Orion Miller
2022/03/21

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - Exploratory Data Analysis with Data Visualization
  - Exploratory Data Analysis with SQL
  - Building an Interactive Map with Folium
  - Building a Dashboard with Plotly Dash
  - Predictive Analysis (Classification)
- Summary of all results
  - Exploratory Data Analysis Results
  - Interactive Analytics Demo
  - Predictive Analysis Comparison

Introduction

- Project background and context
  - SpaceX is the most successful private company of the commercial space age, dramatically lowering the cost of launches. The company advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. Based on public information and various machine learning models, we seek to predict whether SpaceX can reuse the first stage for a given launch.
- Problems you want to find answers
  - How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
  - Does the rate of successful landings increase across time?
  - What is the best algorithm that can be used for binary classification in this situation?

Section 1

Methodology

Methodology

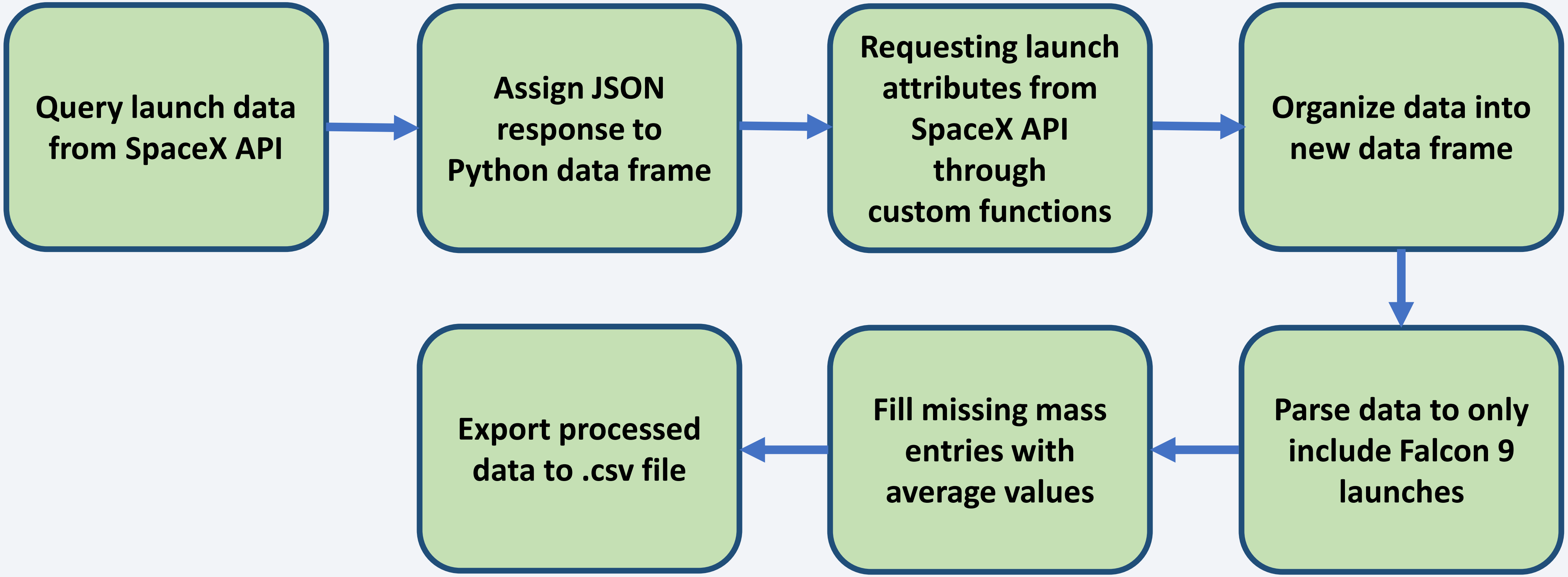
Executive Summary

- Data collection methodology:
  - Collected data directly from the SpaceX Rest API, and also web-scraped from Wikipedia
- Perform data wrangling
  - Filtering, handling of empty values, and one hot encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Built models for logistic regression, SVM, decision tree, and K-nearest neighbors and evaluated their output using multiple metrics

Data Collection

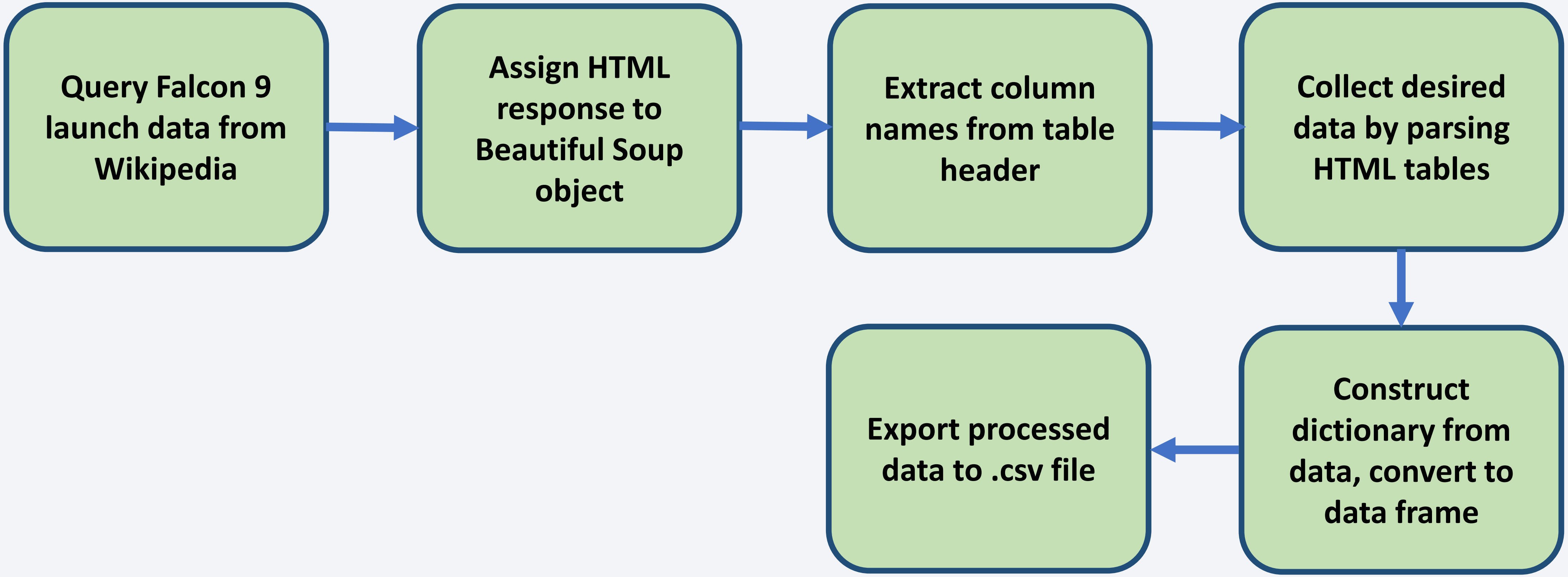
- Data collection process involved a combination of both API requests from the SpaceX REST API and Web Scraping data from SpaceX's Wikipedia page.
- Both of these data collection methods were necessary to get complete enough information about the launches for a detailed analysis.
  - Data columns obtained by using SpaceX REST API:
    - FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
  - Data Columns obtained by using Wikipedia Web Scraping:
    - Flight Number, Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



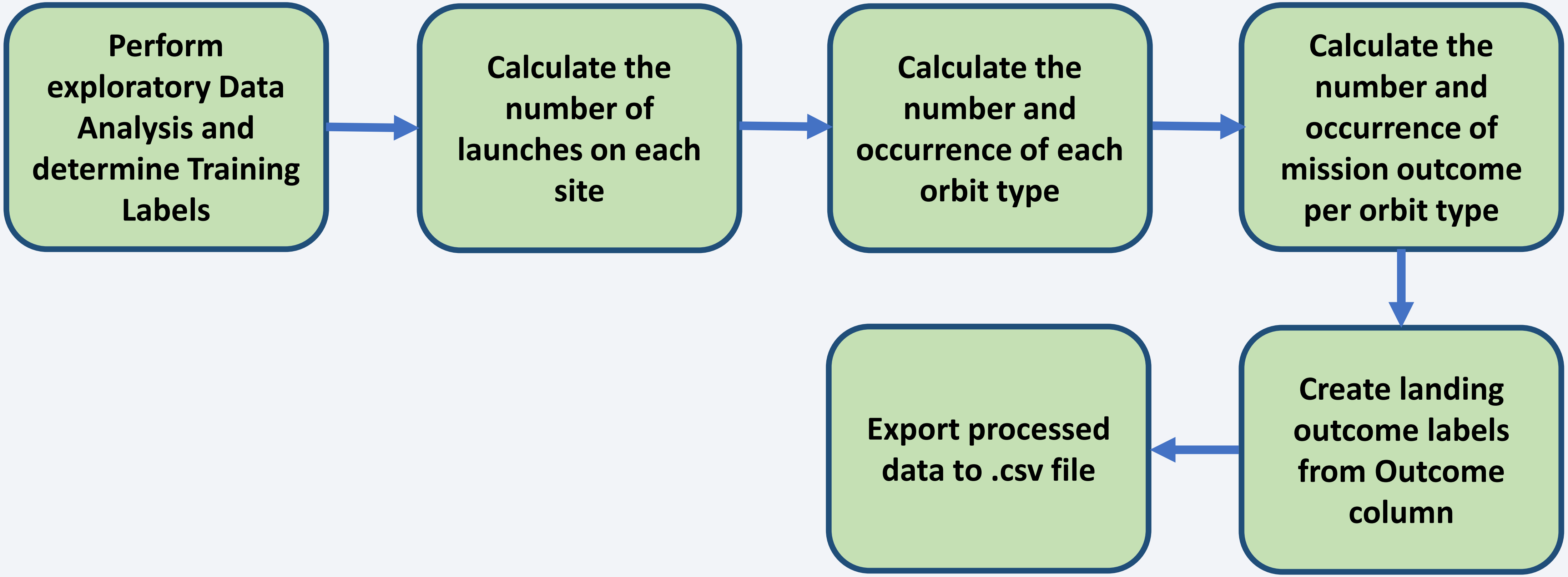
[SpaceX API Notebook]

Data Collection - Scraping



Web Scraping Notebook

Data Wrangling



[Data Wrangling Notebook]

EDA with Data Visualization

- The following relationships were plotted:
  - Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type and Success Rate Yearly Trend
- Scatter plots show the relationship between variables. If a relationship exists, they could be used in a machine learning model.
- Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value.
- Line charts show trends in data across time.

[EDA & Visualization Notebook]

EDA with SQL

- Performed SQL queries:
  - Display the names of the unique launch sites across launch events
  - 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
  - Listing the date when the first successful landing outcome in ground pad was achieved
  - Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - Listing the total number of successful and failure mission outcomes
  - Listing the names of the booster versions which have carried the maximum payload mass
  - Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
  - Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) across time in descending order

EDA with SQL Notebook

Build an Interactive Map with Folium

- Markers for all Launch Sites:
  - Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.
  - Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.
- Colored Markers of the launch outcomes for each Launch Site:
  - Added colored Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.
- Distances between a Launch Site to its proximities:
  - Added colored Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City.

[Interactive Folium Map Notebook]

Build a Dashboard with Plotly Dash

- Launch Sites Dropdown List:
  - Added a dropdown component to allow Launch Site selection.
- Pie Chart showing Success Launches (All Sites/Certain Site):
  - Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.
- Slider of Payload Mass Range:
  - Added a slider to select Payload range.
- Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:
  - Added a scatter chart to visualize the correlation between Payload and Launch Success.

Plotly Dash App

Predictive Analysis (Classification)

1. Create NumPy array from the column "Class" in data
2. Standardize the data, fit it, and then transform it
3. Split the data into training and testing sets with train_test_split function
4. Create a GridSearchCV object to find the best parameters
5. Applying GridSearchCV on LogReg, SVM, Decision Tree, and KNN models
6. Calculating the accuracy on the test data using the method .score() for all models
7. Examine the confusion matrix for all models
8. Evaluate the best method with Jaccard_score and F1_score metrics

[Predictive Analysis Notebook]()

Results

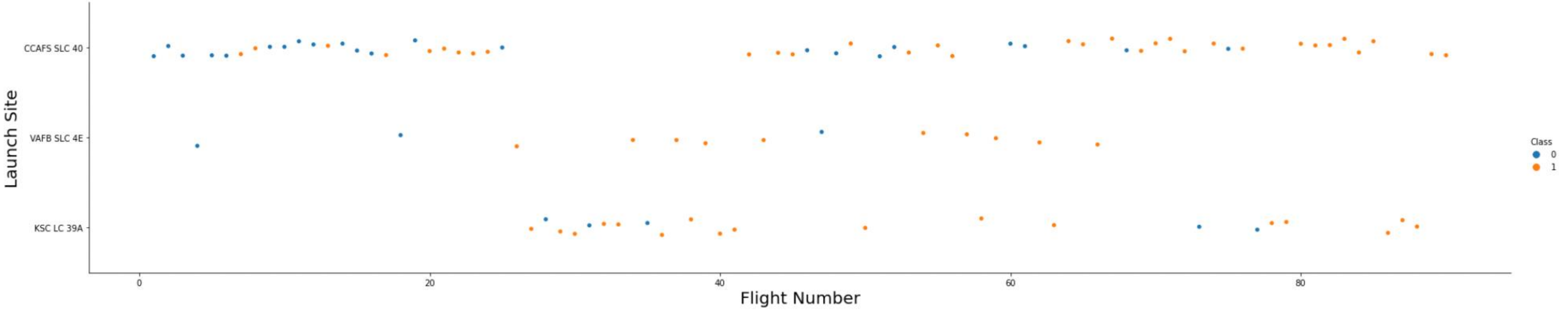
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

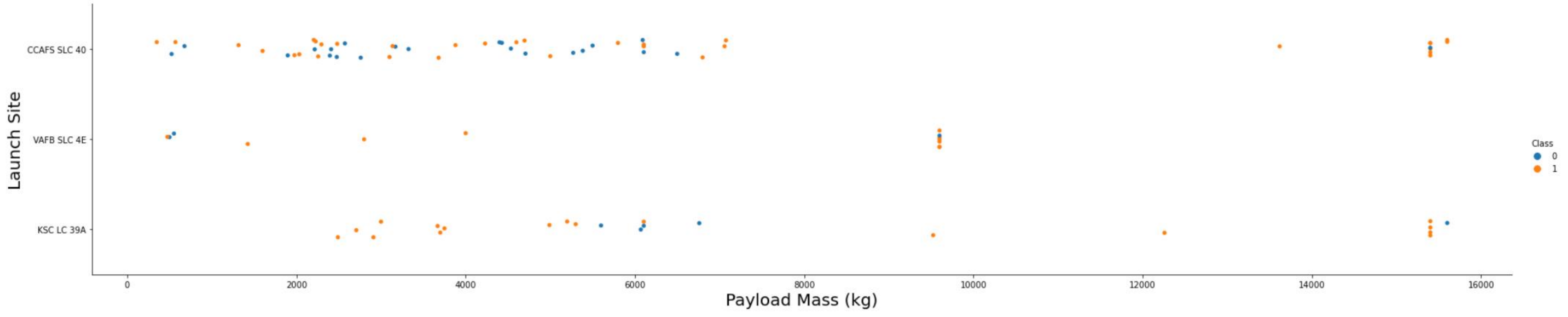
Flight Number vs. Launch Site

- Early flights had a high rate of failure, while later flights had a very low rate of failure. There is a notable change that happens around flight 20 where subsequent flights become more successful. This may be related to a particular improvement to the system made around that time.
- The CCAFS SLC 40 launch site has about a half of all launches. VAFB SLC 4E and KSC LC 39A seem to have higher success rates than CCAFS SLC 40, but its unclear whether this is attributable to any particular attributes of those sites, rather than the simple fact that they went into use at a later time in Falcon 9 development.



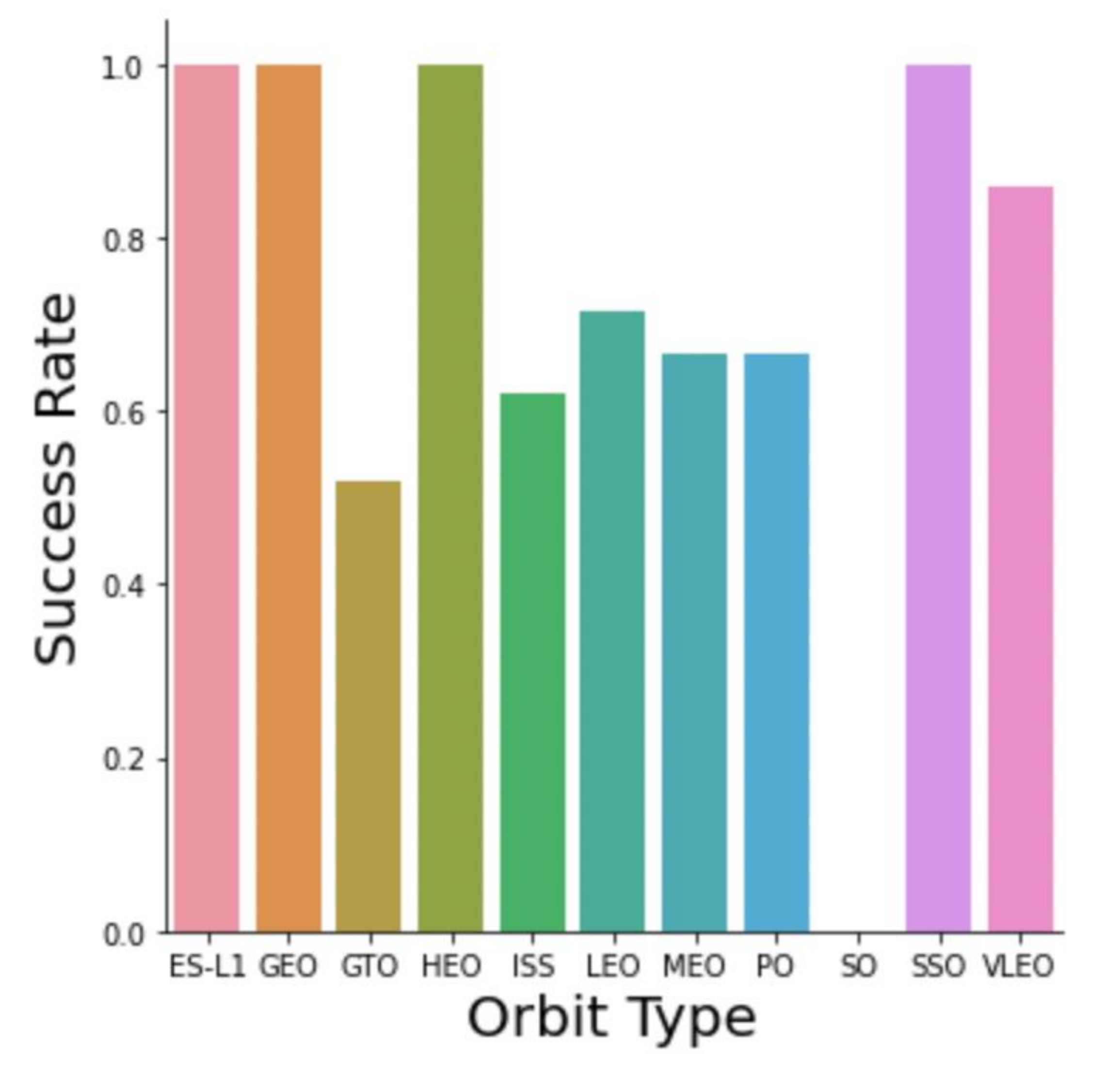
Payload vs. Launch Site

- The vast majority of launches carried payloads of under 7000 kg.
- The relationship between payload mass and launch success is not very clear from this plot. Particularly with CCAFS SLC 40, the failures (blue) seem to be mostly evenly distributed.
  - Most of the launches with payload mass over 7000 kg were successful. This could be because SpaceX was more willing to use a large payload in higher confidence situations.
  - KSC LC 39A has a 100% success rate for payload mass under 5500 kg.



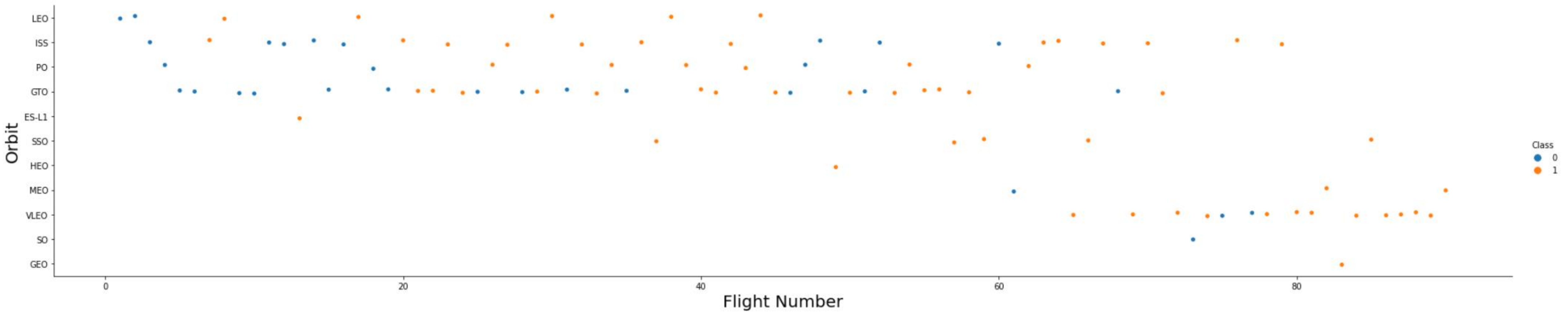
Success Rate vs. Orbit Type

- Orbits with 100% success rate:
  - ES-L1, GEO, HEO, SSO
- Orbits with success rate between 50% and 90%:
  - GTO, ISS, LEO, MEO, PO, VLEO
- Orbits with 0% success rate:
  - SO



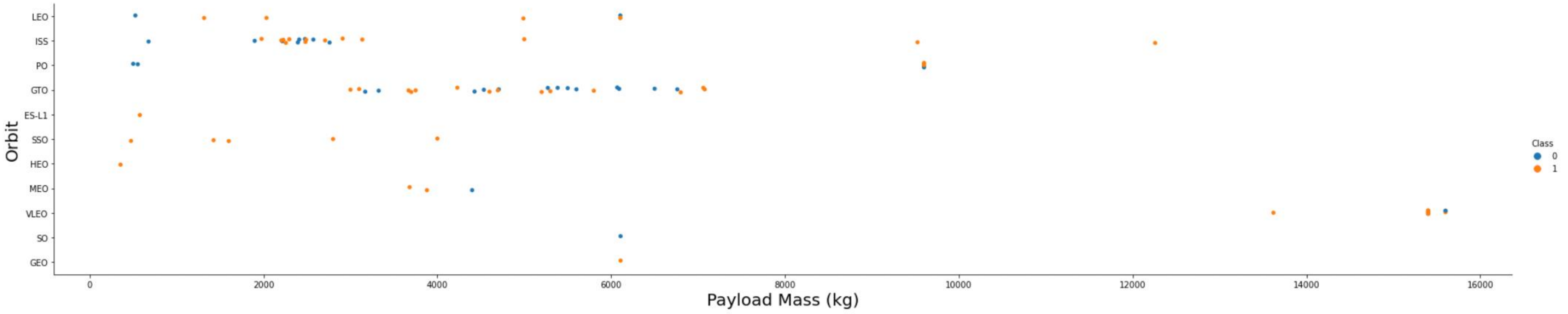
Flight Number vs. Orbit Type

- GTO orbit seems to show the clearest relationship of the success rate improving as flight number increases. This also happens with LEO and PO, but those have comparatively lower sample sizes.
- It seems that ISS, GTO, and VLEO are the more common orbit types.



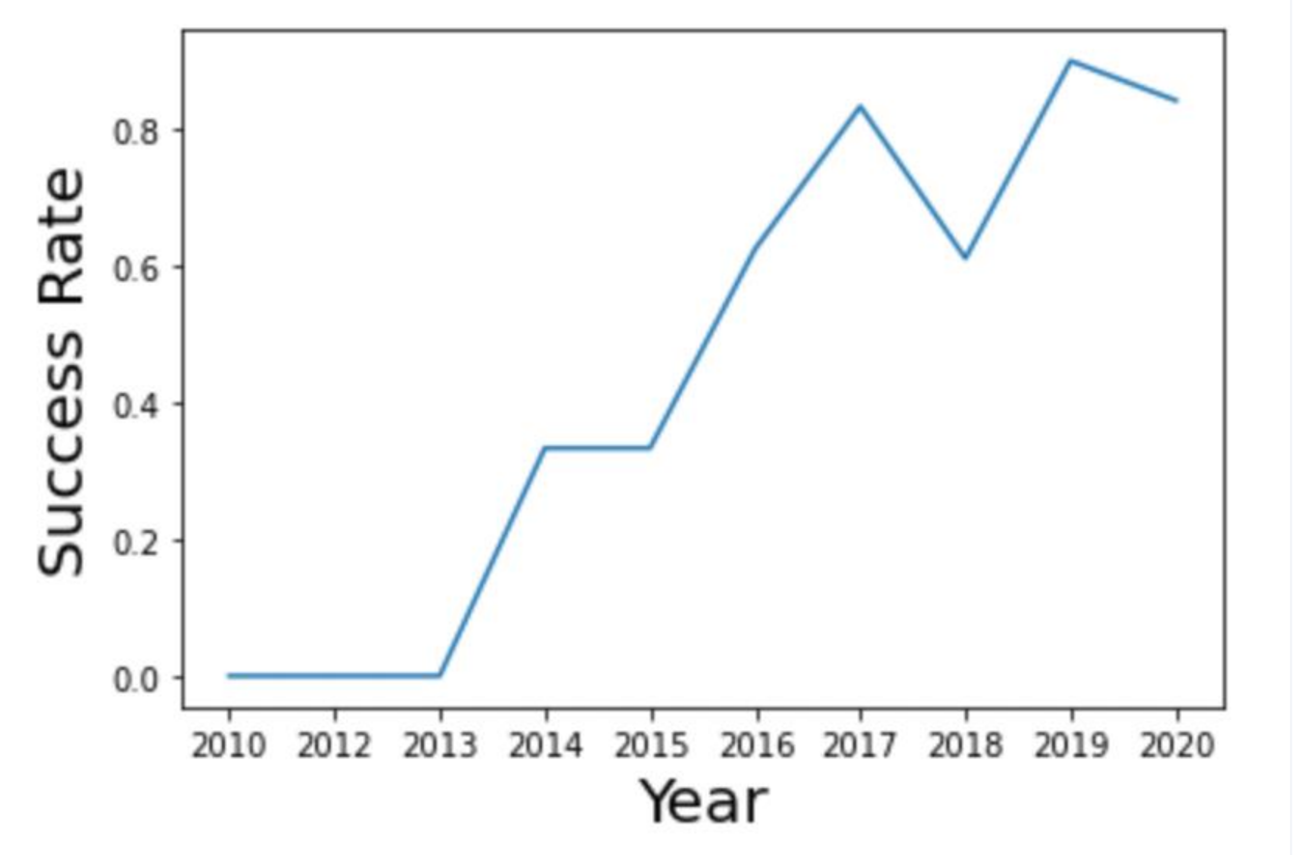
Payload vs. Orbit Type

- Heavy payloads have a negative influence on GTO orbits and positive on ISS orbits. Its unclear if this is a causal relationship.
- Trends are not clear for other orbit types.



Launch Success Yearly Trend

- The increase success rate across time mostly occurred between 2013 and 2017.
- There was a reduction in success in 2018.
- Success rate has yet to surpass around 90%.



All Launch Site Names

- There are four unique launch site names:
  - CCAFS LC-40
  - CCASF LSC-40
  - KSC LC-39A
  - VAFB SLC-4E

In [4]:
```
%sql select distinct launch_site from SPACEXDATASET;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[4]:

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

Launch Site Names Begin with 'CCA'

- Showing first 5 entries where **launch_site** began with string 'CCA'

```
In [5]: %sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[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

- Displaying sum total of payload masses delivered for NASA (CRS)

```
In [6]: %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[6]:
```

total_payload_mass
45596

Average Payload Mass by F9 v1.1

- Displaying average payload mass carried by booster version F9 v1.1.

```
In [7]: %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%';

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[7]:

average_payload_mass
2534

First Successful Ground Landing Date

- Listing the date when the first successful landing outcome in ground pad was achieved. The first successful launches began occurring in 2014 so it took about another year to successfully land on ground.

In [8]:
```
%sql select min(date) as first_successful_landing from SPACEXDATASET where landing__outcome = 'Success (ground pad)';
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[8]:

first_successful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Listing the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.

In [9]:
```
%sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4
000 and 6000;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[9]:

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

Total Number of Successful and Unsuccessful Mission Outcomes

- Showing the total number of successful and unsuccessful mission outcomes.

```
In [10]: %sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[10]:

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

Boosters Carried Maximum Payload

```
In [11]: %sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET);

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[11]:

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

- Listing the names of the booster versions which carried the maximum payload mass

2015 Launch Records

- Listing the failed landing outcomes in drone ship, their booster versions, and launch site names for 2015.

```
In [12]: %%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET
               where landing__outcome = 'Failure (drone ship)' and year(date)=2015;

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[12]:
```

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- Ranking the number of landing outcomes between 2010-06-04 and 2017-03-20 in descending order. Drone ship landings were the most common scenario.

```
In [13]: %%sql select landing__outcome, count(*) as count_outcomes from SPACEXDATASET
               where date between '2010-06-04' and '2017-03-20'
               group by landing__outcome
               order by count_outcomes desc;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[13]:

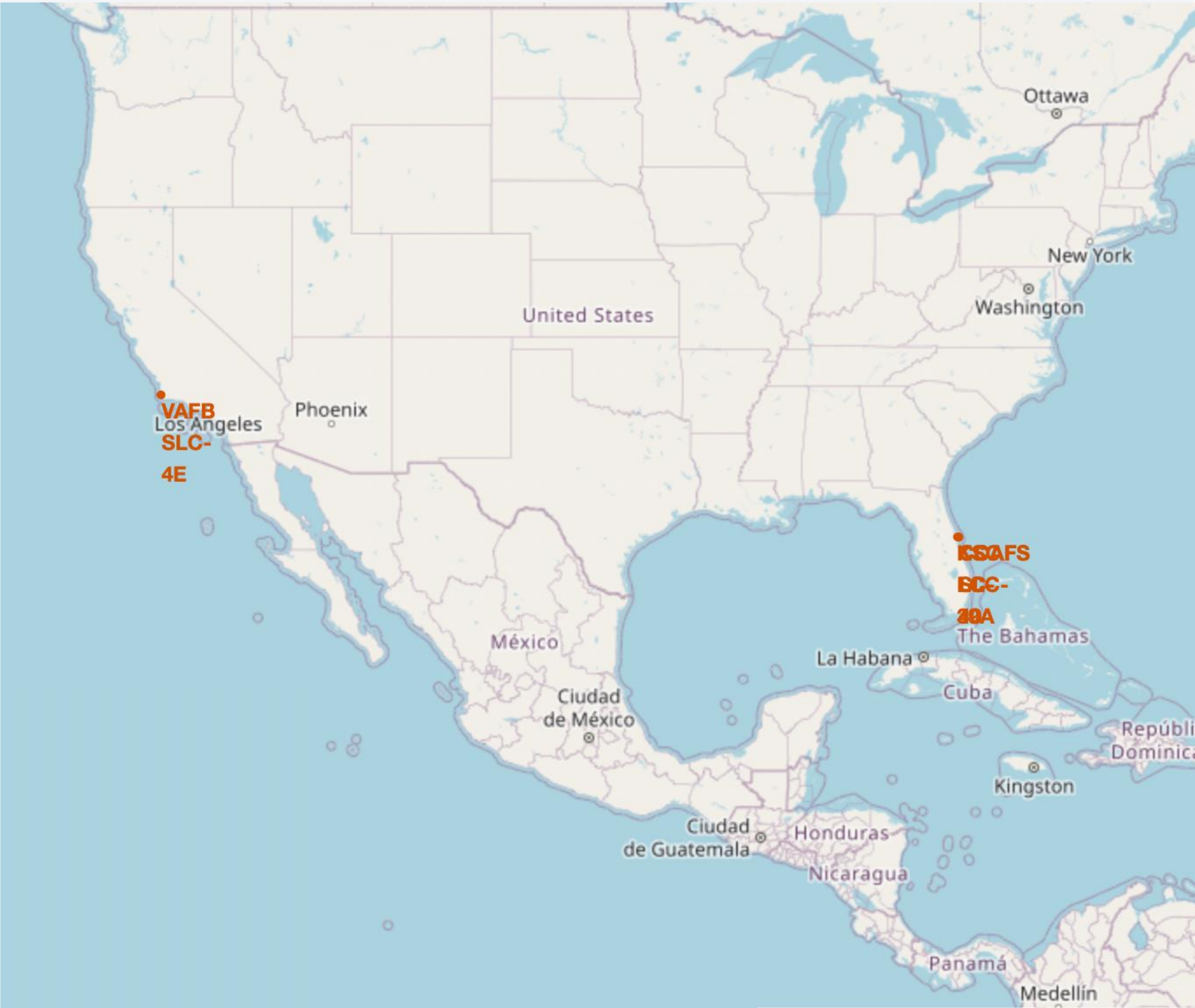
landing__outcome	count_outcomes
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

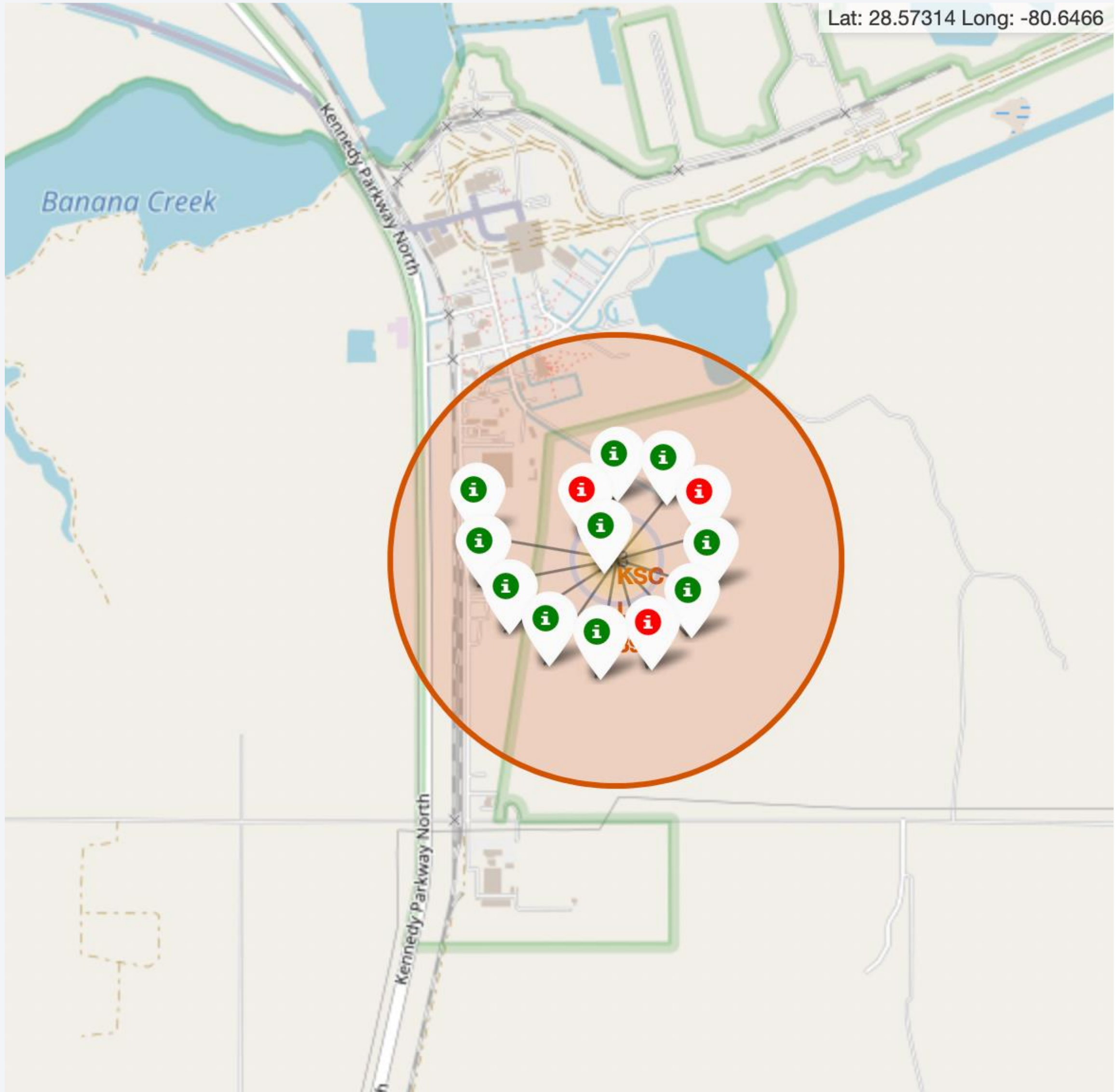
Launch Site Map Locations

- Most of Launch sites are in proximity to the Equator line. The land is moving faster at the equator than any other place on the surface of the Earth. Anything on the surface of the Earth at the equator is already moving at 1670 km/hour. If a ship is launched from the equator it goes up into space, and it is also moving around the Earth at the same speed it was moving before launching. This is because of inertia. This speed will help the spacecraft keep up a good enough speed to stay in orbit.
- All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimizes the risk of having any debris dropping or exploding near people.



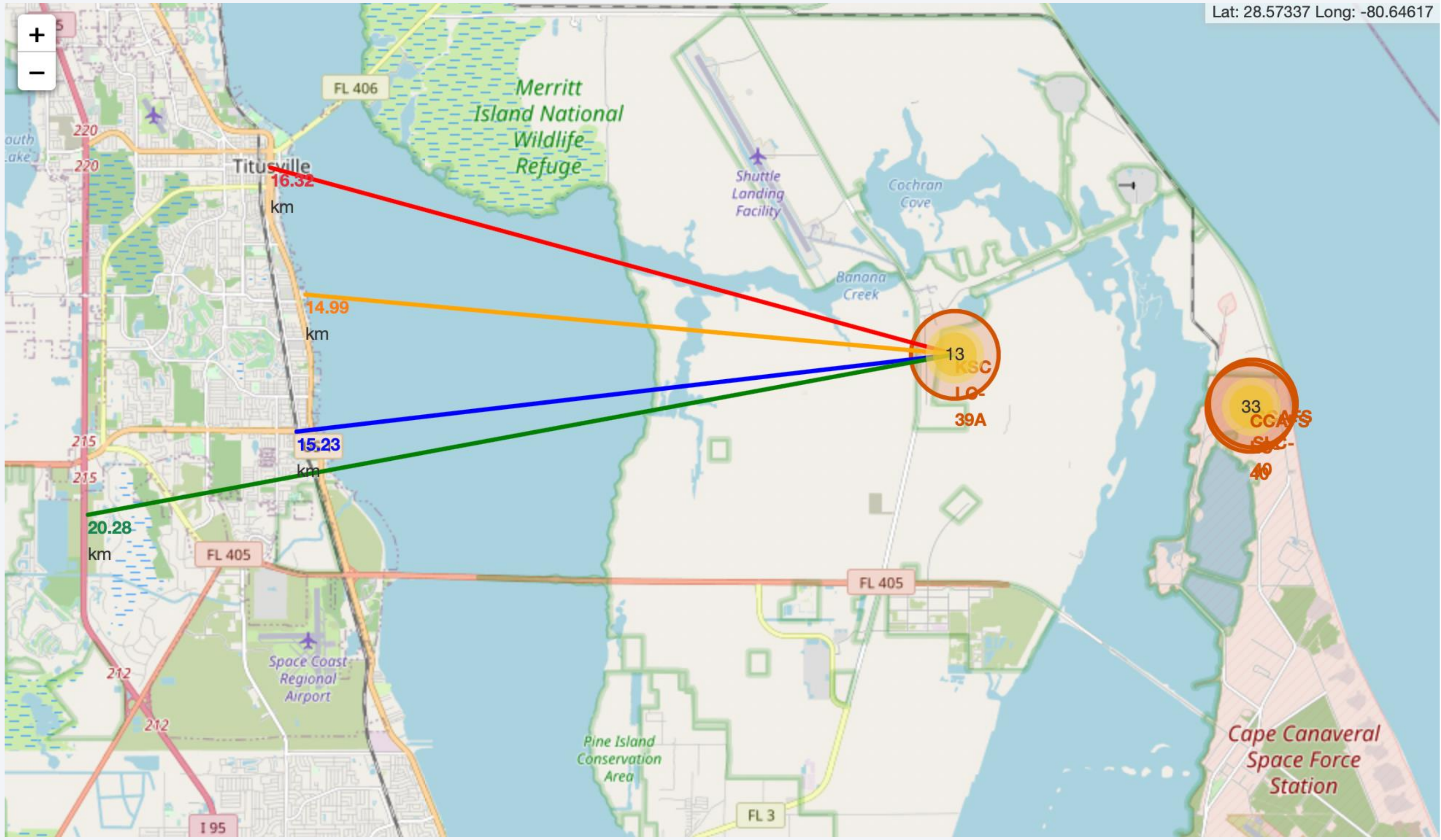
Color-Labeled Launch Outcomes

- Using color-labeled markers we can easily identify which launch sites have high success rates.
  - Green Marker = Successful Launch
  - Red Marker = Failed Launch
- Launch Site KSC LC-39A (shown here) has a very high Success Rate.



Proximity of Launch Site to Specific Amenities/Locations

- From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:
  - relatively close to a railway (15.23 km)
  - relatively close to a highway (20.28 km)
  - relatively close to the coastline (14.99 km)
- Also the launch site KSC LC-39A is relative close to its closest city Titusville (16.32 km).

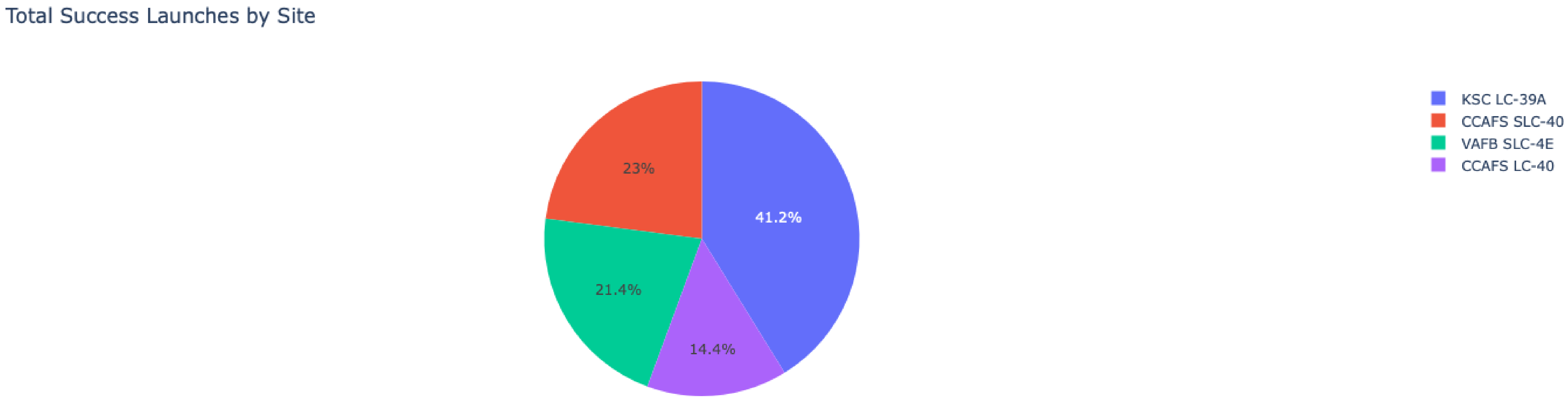


Section 4

Build a Dashboard with Plotly Dash

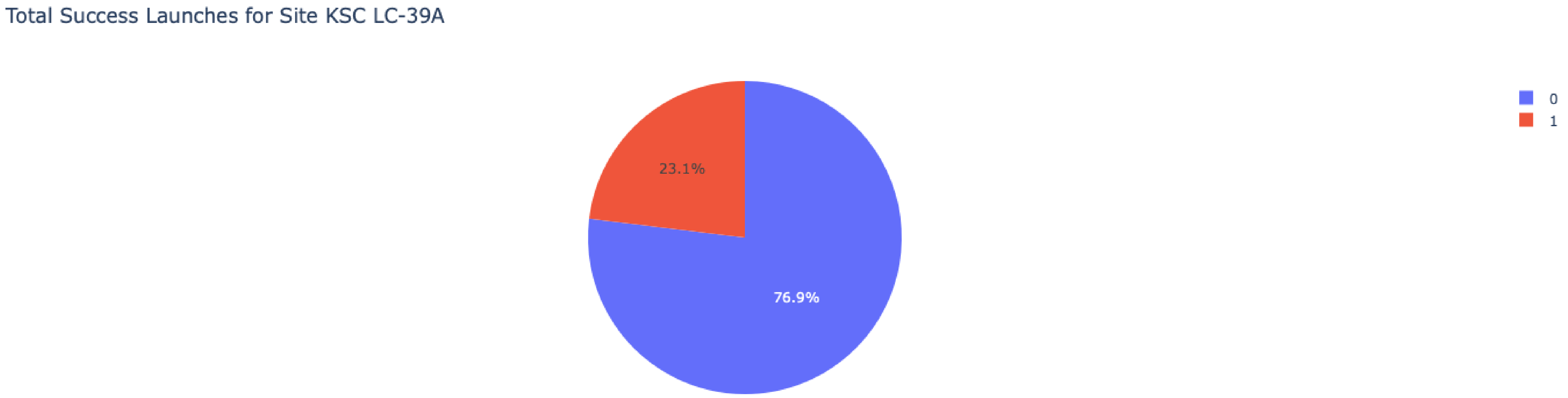
Launch Success Count Between Sites

- KSC LC-39A has the most successful events, where CCAFS LC-40 has the least.



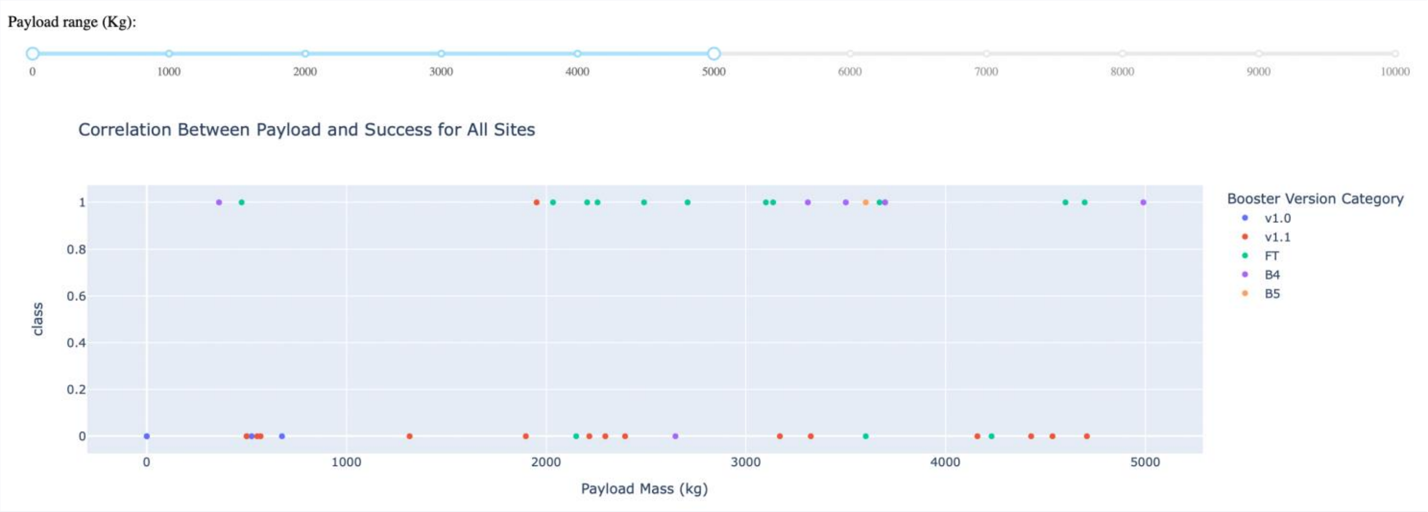
Launch Site with Highest Success Ratio

- KSC LC-39A has the highest launch success rate, with 10 successful and
- only 3 failed landings.



Payload Mass vs. Launch Outcome Between Sites

- The charts show that payloads between 2000 and 5500 kg have a high success rate.

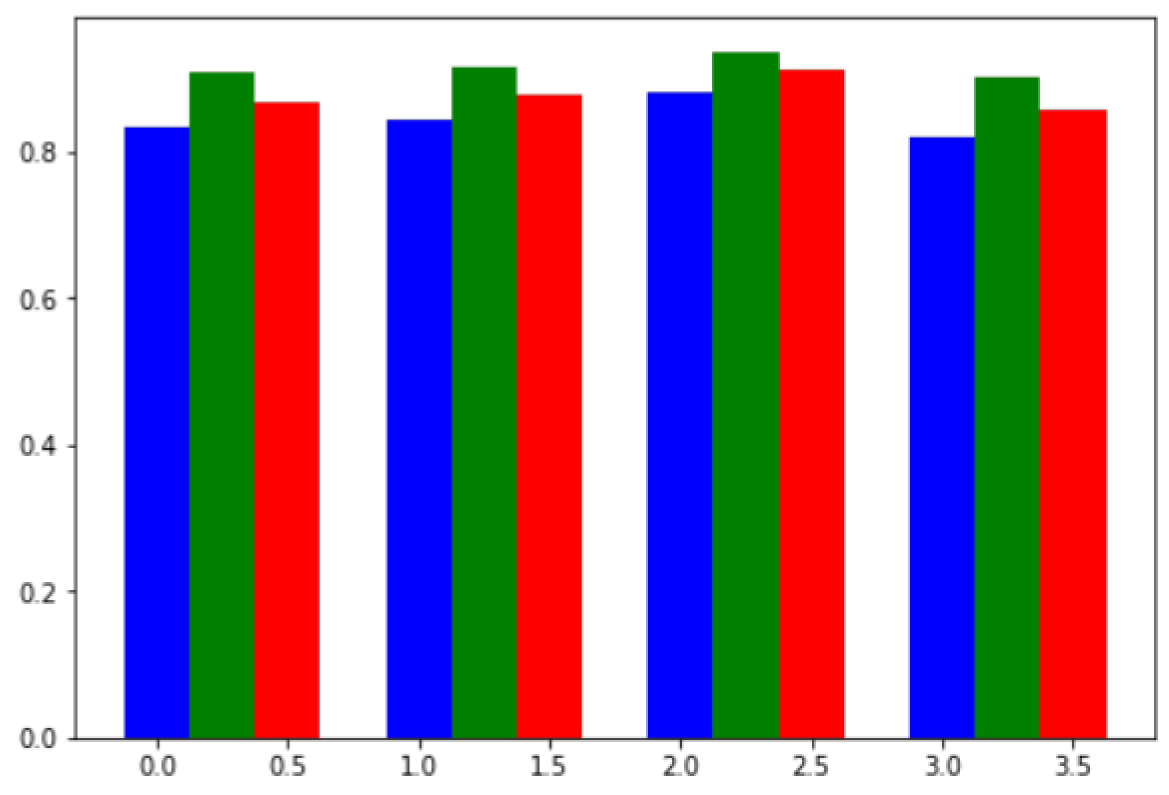


Section 5

Predictive Analysis (Classification)

Classification Accuracy

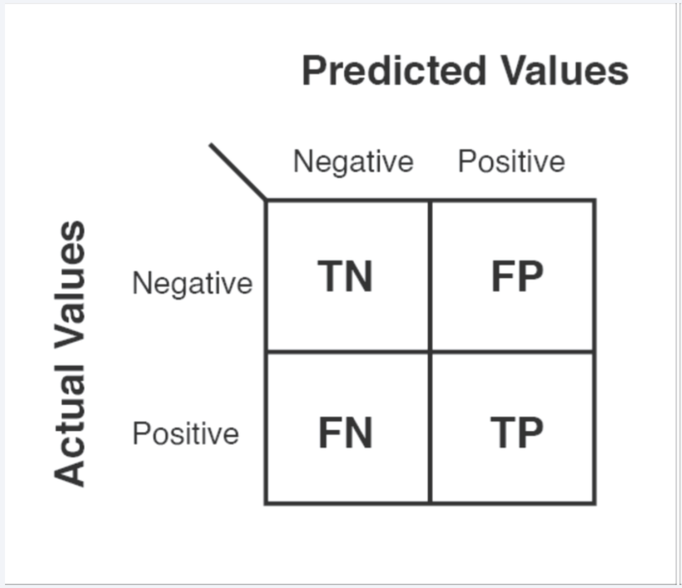
- The Decision Tree has the best performance against the full dataset of all evaluated models, by 2-4% depending on the metric.
- Blue = Jaccard Score
- Green = F1 Score
- Red = Accuracy

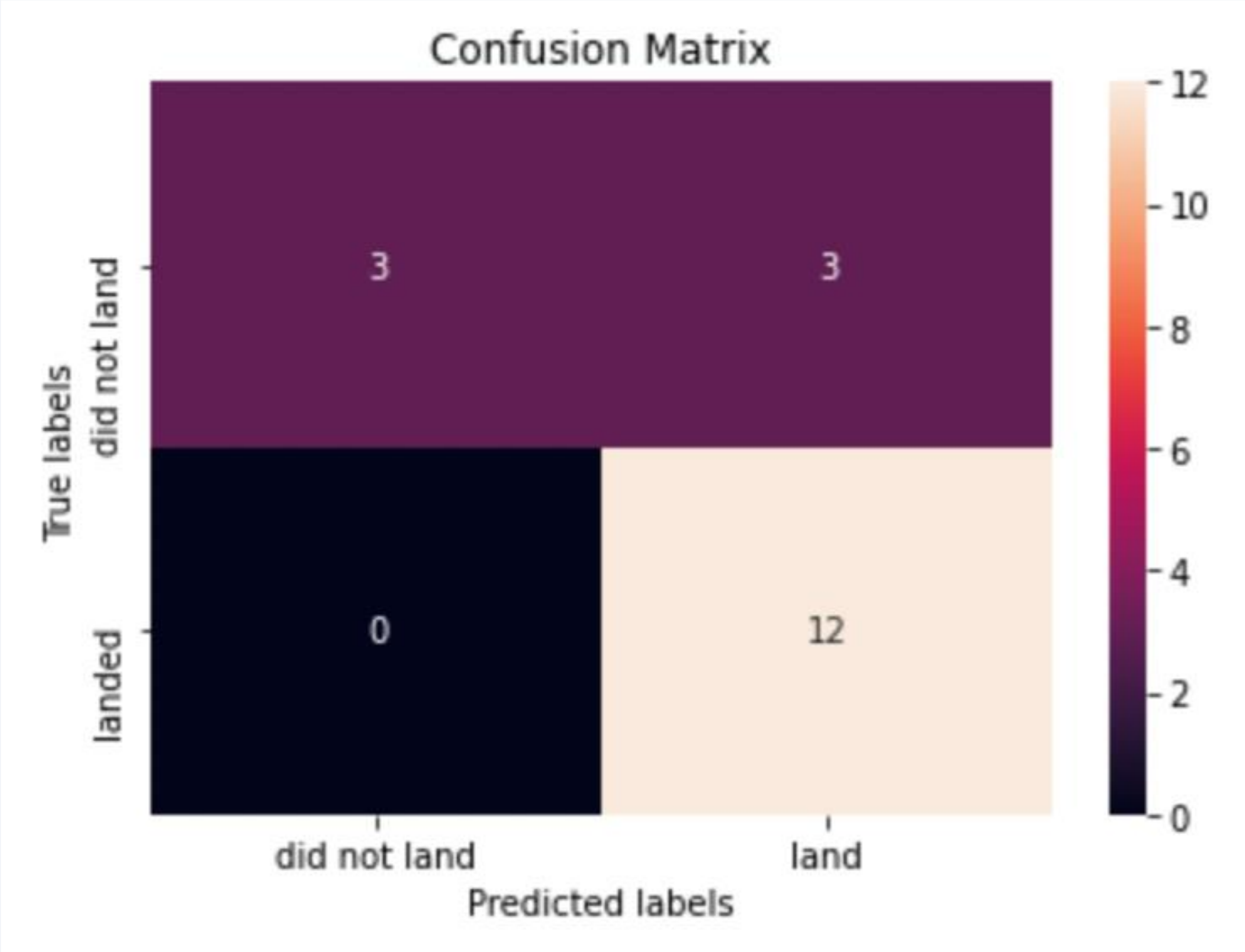


	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

Confusion Matrix

- The decision tree shows that only one type of error occurs, false positives. For a SpaceX's perspective, it would be much better to overpredict false negatives as opposed to false positives, because of the need to have a high factor of safety for these launches.





Conclusions

- Decision Tree Model is the best algorithm for this dataset, based on this process.
- The success rate of launches increases over the years, with most of the improvement occurring between 2013 and 2017.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- It is unclear if there is a strong relationship between payload mass and launch success.

Thank you!