IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Steven Collareno
6/14/2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection through API and web scraping
  - Data Wrangling
  - EDA with SQL and Data Visualization
  - Visual Analytics featuring Folium
  - Machine Learning Model and prediction
- Summary of all results
  - EDA result
  - Interactive analytics
  - Predictive analytics

Introduction

- Project background and context
  - Space X says that the Falcon 9 rocket launches with a cost of 62 million dollars. However, other companies say that the cost is about 165 million dollars. Why the discrepancy? Space X can reuse the first stage of their rockets. This, if we can determine if the first stage is going to land it will determine how much the launch will cost. If you were an alternate company-you could use this information to bid against Space X. This is the ultimate purpose of this project-use machine learning to find if the first stage will land successfully.
- Problems you want to find answers
  - What are the factors that will determine a successful launch?
  - What are the interactions between different features that have the highest correlation in determining success?
  - What are the optimal operating conditions to ensure a successful landing?

Section 1

Methodology

Methodology

Executive Summary

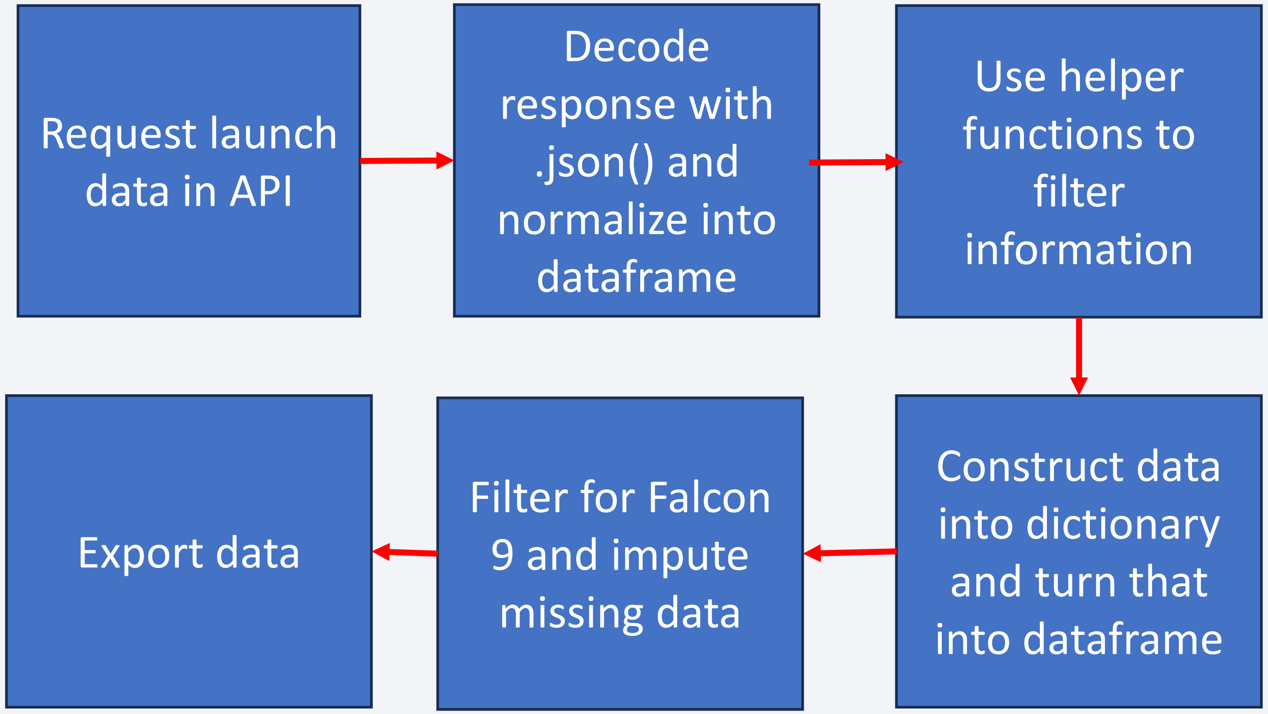
- Data collection methodology:
  - Collected data through SpaceX API and web scraping from Wikipedia.
- Perform data wrangling
  - Categorical features were one-hot encoded
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Classification models were built and tuned according to the requirements need to finish the project.

Data Collection

- The initial data was collected through a the SpaceX API
  - This initial data was left in Json format-so it was turned into a pandas dataframe using the normalize function
  - As per usual-any missing data was filled in using imputation or even just removing the data entirely.
- Notably-there was not enough data to go around with just the API.
  - BeautifulSoup, a python library for pulling data out of HTML files, was used on Wikipedia to get more Falcon 9 launch records
  - This data met the same fate as the initial API data-it was converted into a pandas dataframe

Data Collection – SpaceX API

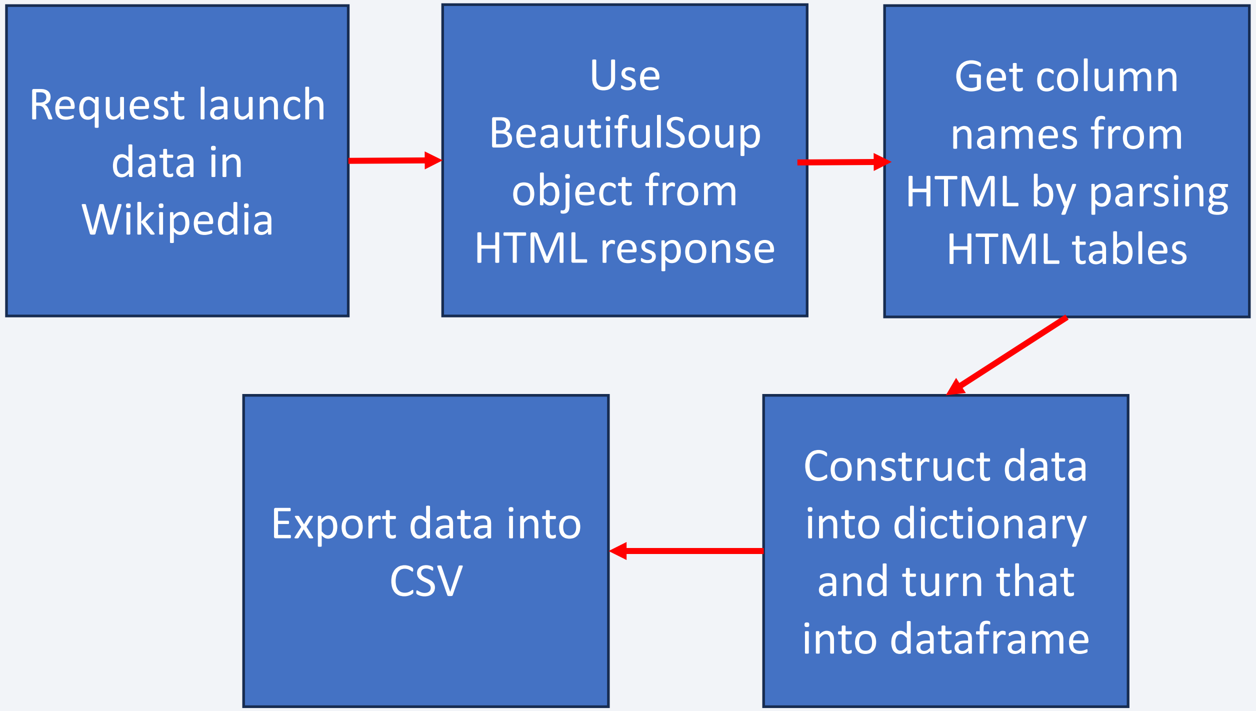
- To the right is a flowchart for how I constructed the initial data collection
- Github URL: https://github.com/scollareno28/IBM-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api%20(2).ipynb



Data Collection - Scraping

- Flowchart to the right about process

- Github for web scraping https://github.com/scollareno28/IBM-Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb

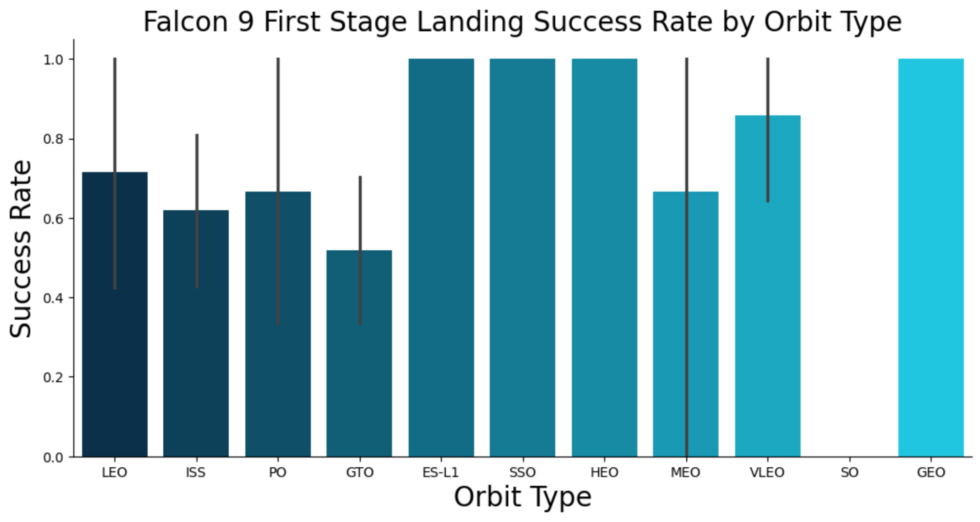


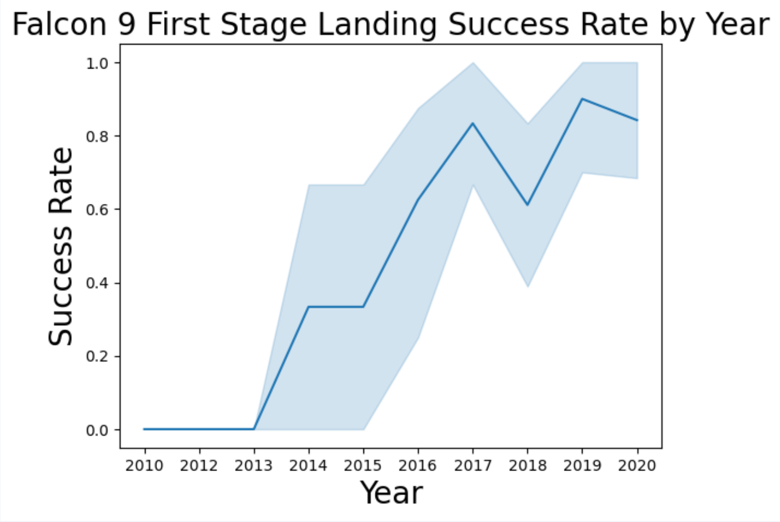
Data Wrangling

- Describe how data were processed
- You need to present your data wrangling process using key phrases and flowcharts
- Add the GitHub URL of your completed data wrangling related notebooks, as an external reference and peer-review purpose

EDA with Data Visualization

- The chart that were used to find the relationship between launch site and success was a simple bar chart – which was helpful since there's only a handful of launch sites. The success rate as a change over year was a line graph to see each trend. Those two plots are listed below, and here is the link to the github: https://github.com/scollareno28/IBM-Data-Science-Capstone/blob/main/edadataviz.ipynb





EDA with SQL

- Using Pandas and SQLlite, I created a database in order to get a better view of the data and different ways to look at things:
  - First successful landing site ever for a Falcon 9
  - Payload mass carried (on average)
  - Payload mass within a range
  - The amount of successful launches and failure launches
  - Where each drone ship was launched from and whether it was a success
- Notebook link: https://github.com/scollareno28/IBM-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- All launch sites were marked with circles, lines, markers in order to show if a launch was a success or failure on the folium maps. Each site was marked wit ha 0 for failure and a 1 for success.

- This was done because we need to know if a launch site was close to a highway, railway or maybe even the coast. We also needed to understand how far away a launch is from a city.
- Github link: https://github.com/scollareno28/IBM-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- The dashboard was made utilizing Plotly dash, python, and cognitiveclass.ai. The dashboard was created in order to look at 5 different sites and how they fared (success vs failure). You could see if a flight was successful out of each site as well as play around with different payloads to see how that affected success/failure
- Github page: https://github.com/scollareno28/IBM-Data-Science-Capstone/blob/main/dashboard.py

Predictive Analysis (Classification)

- Using Numpy and Pandas, data was split into training and testing. Then, the data was hyperparameter tuned using GridSearch
- Accuracy was the metric for our models
- Github: https://github.com/scollareno28/IBM-Data-Science-Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

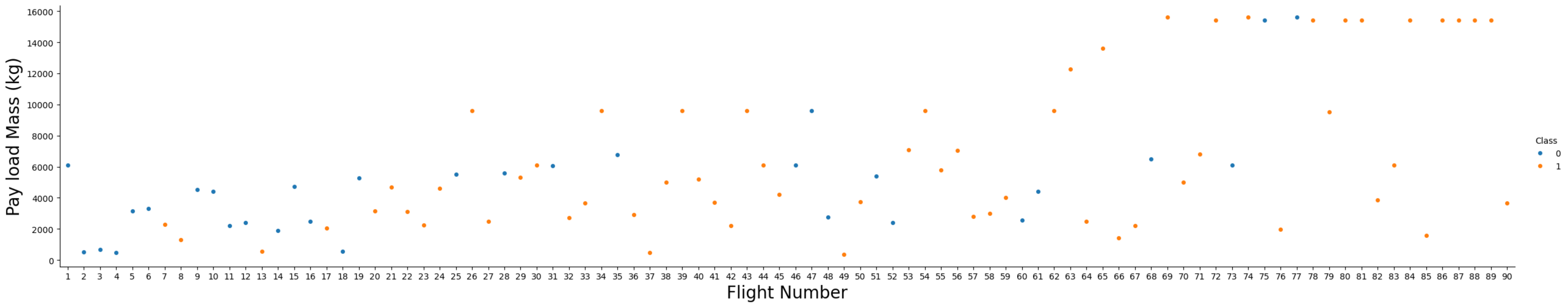
Results

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

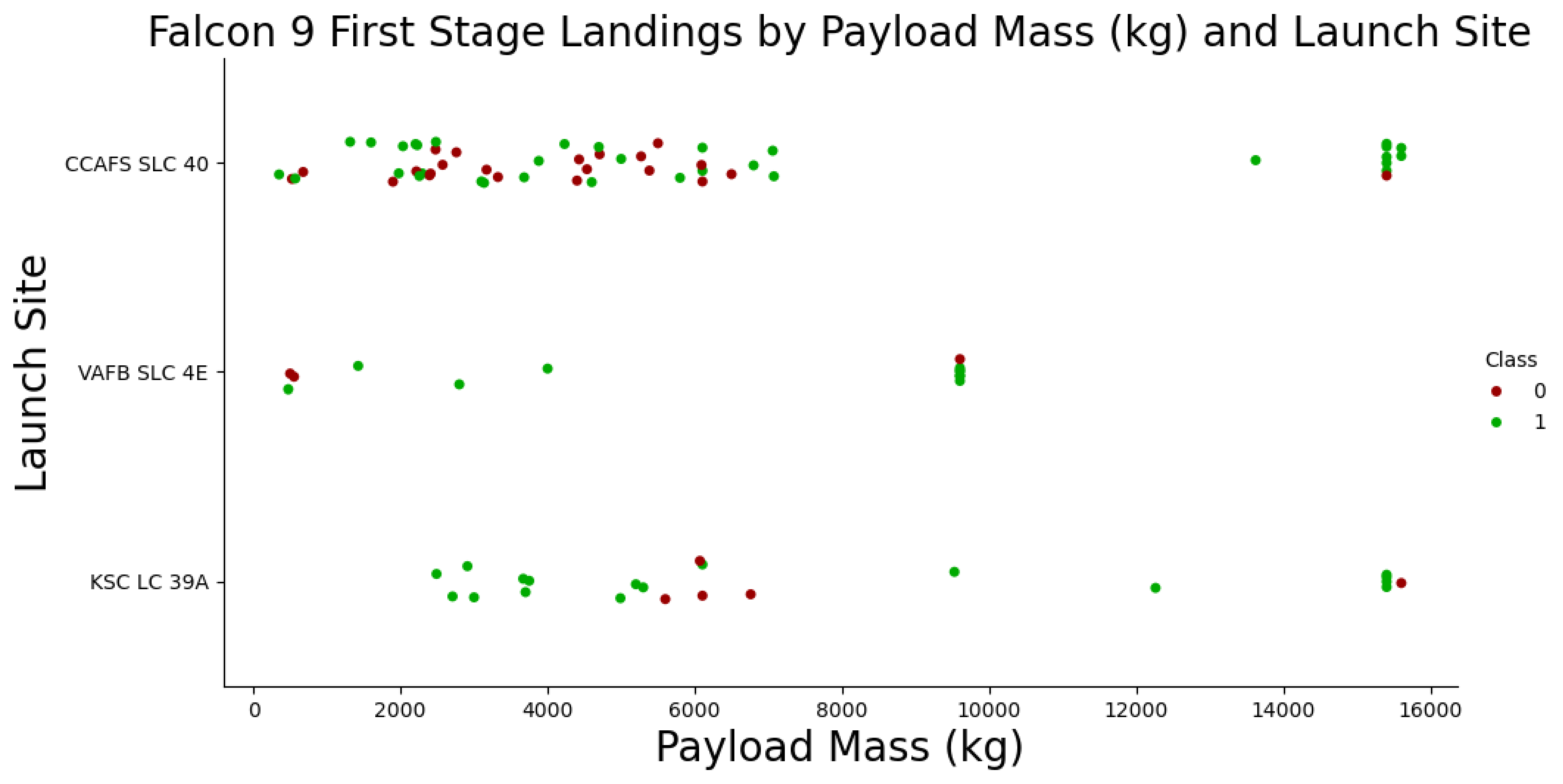
Section 2

Insights drawn from EDA

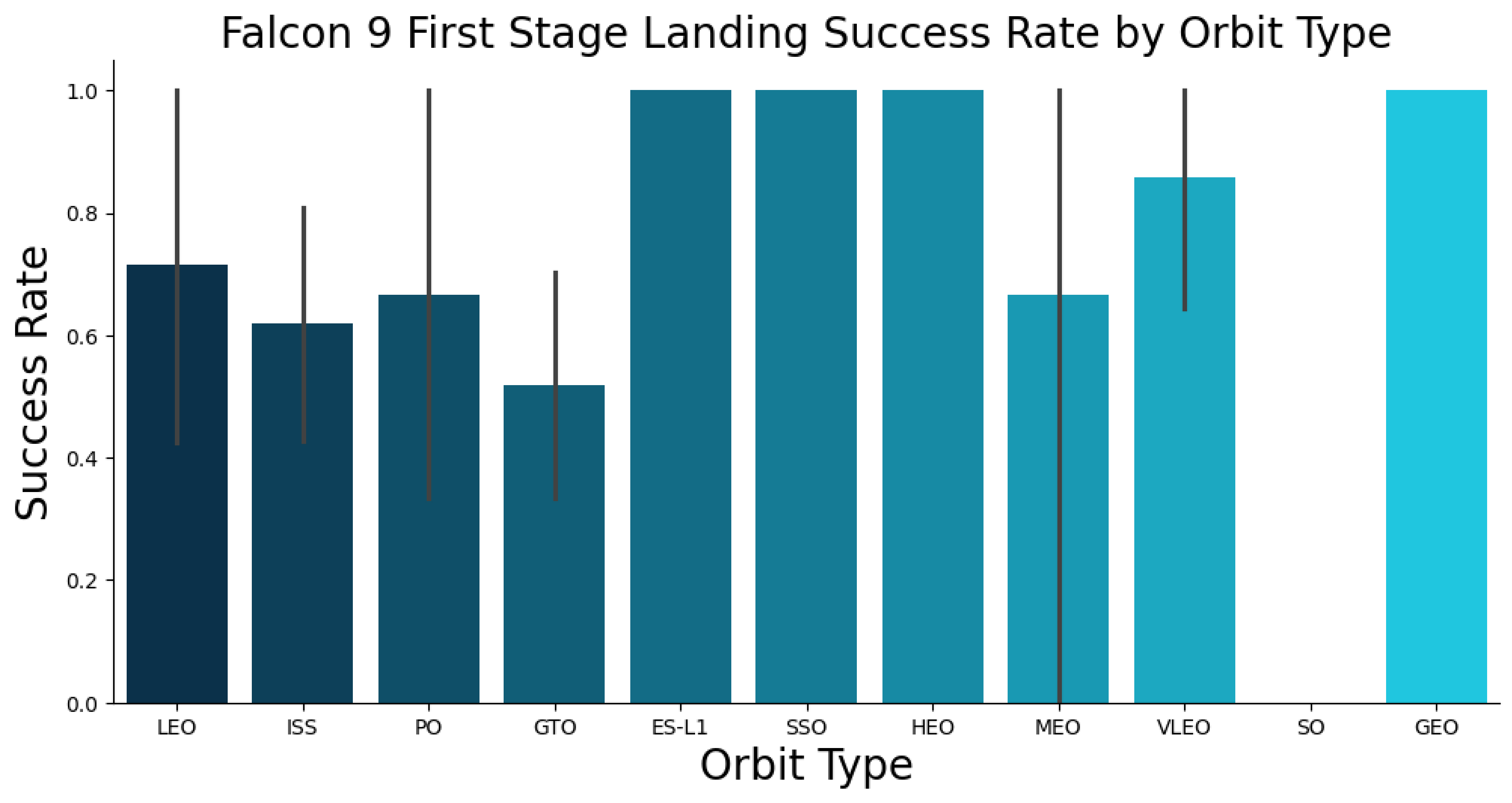
Flight Number vs. Launch Site



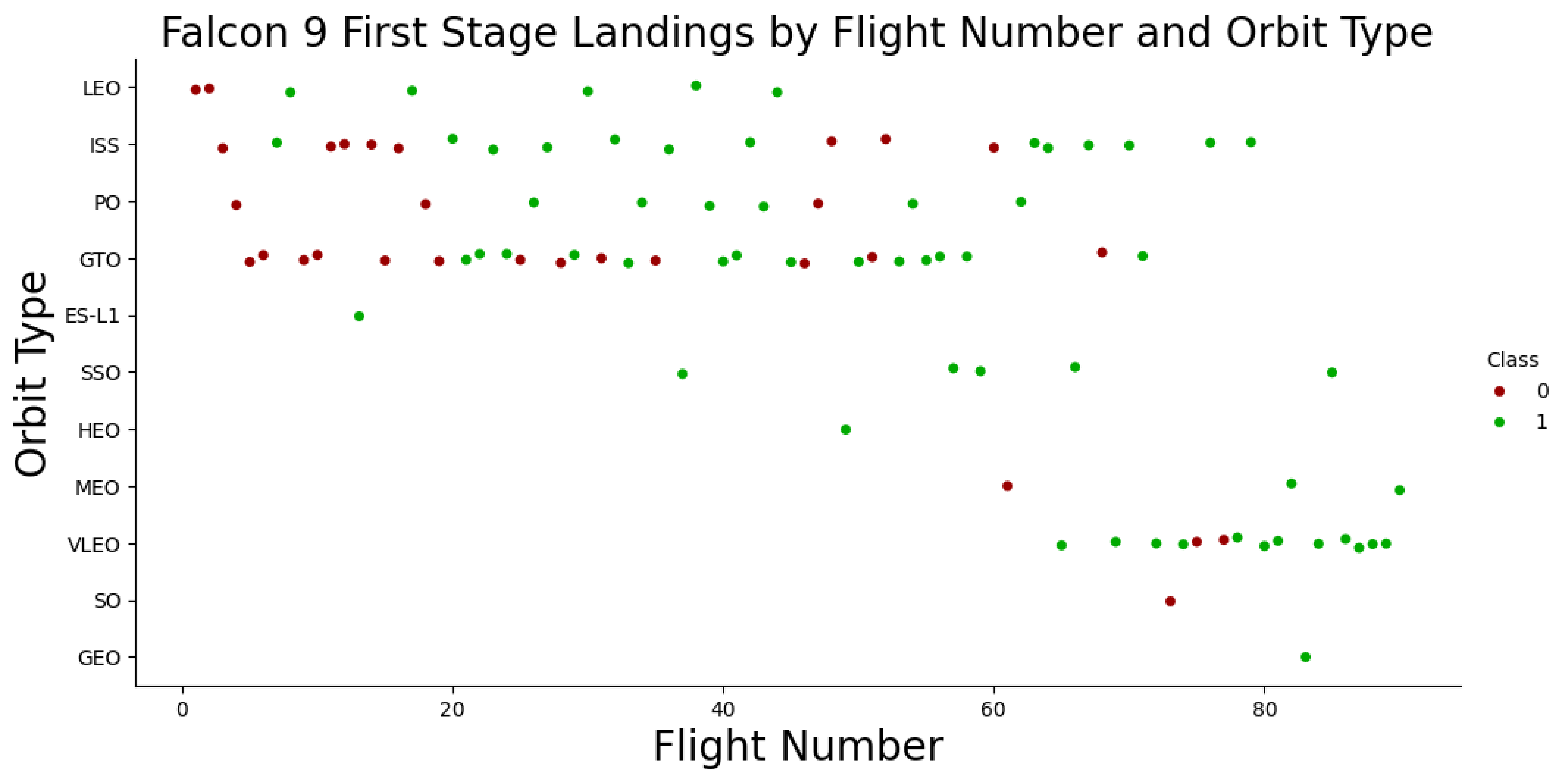
Payload vs. Launch Site



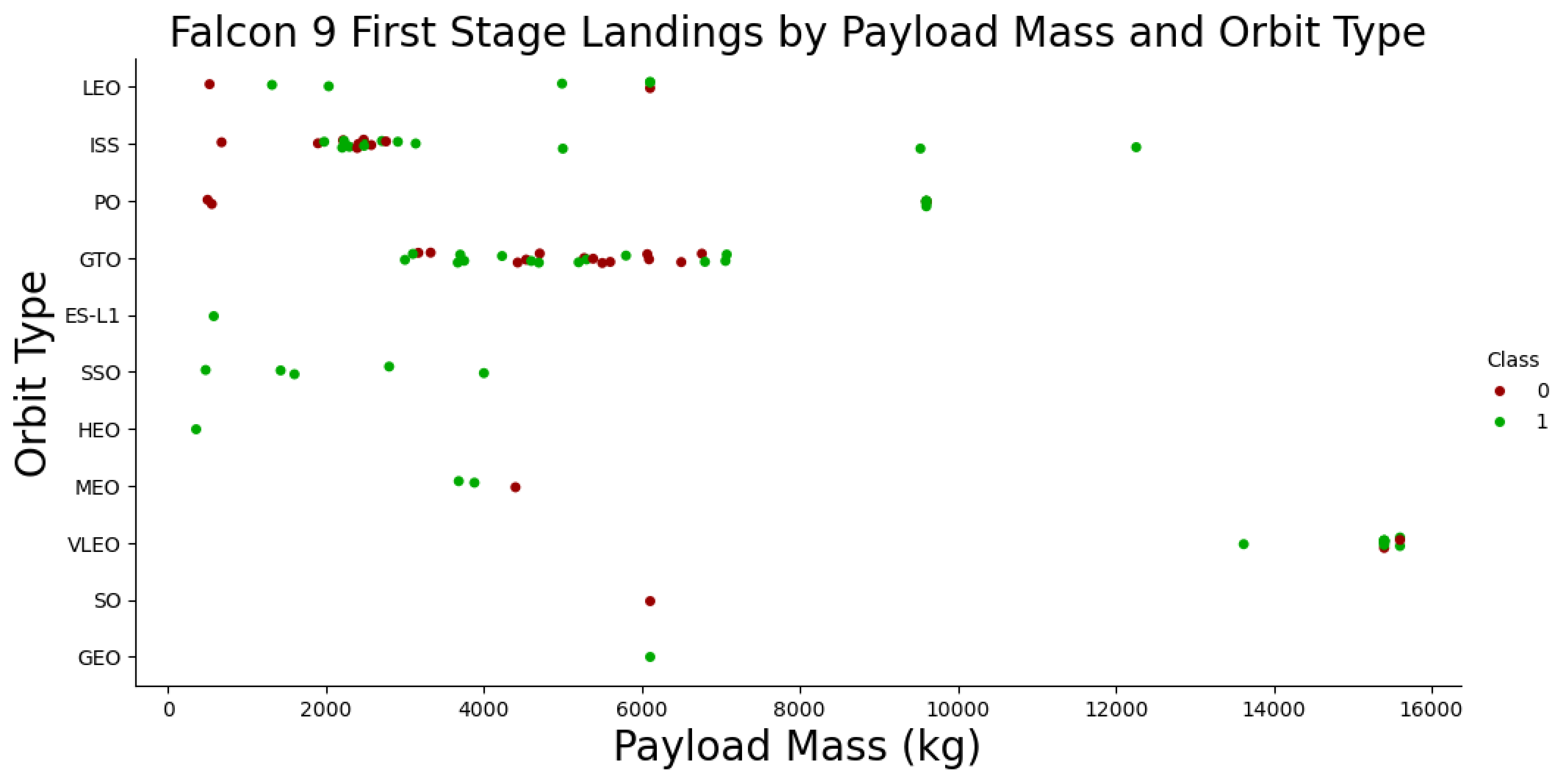
Success Rate vs. Orbit Type



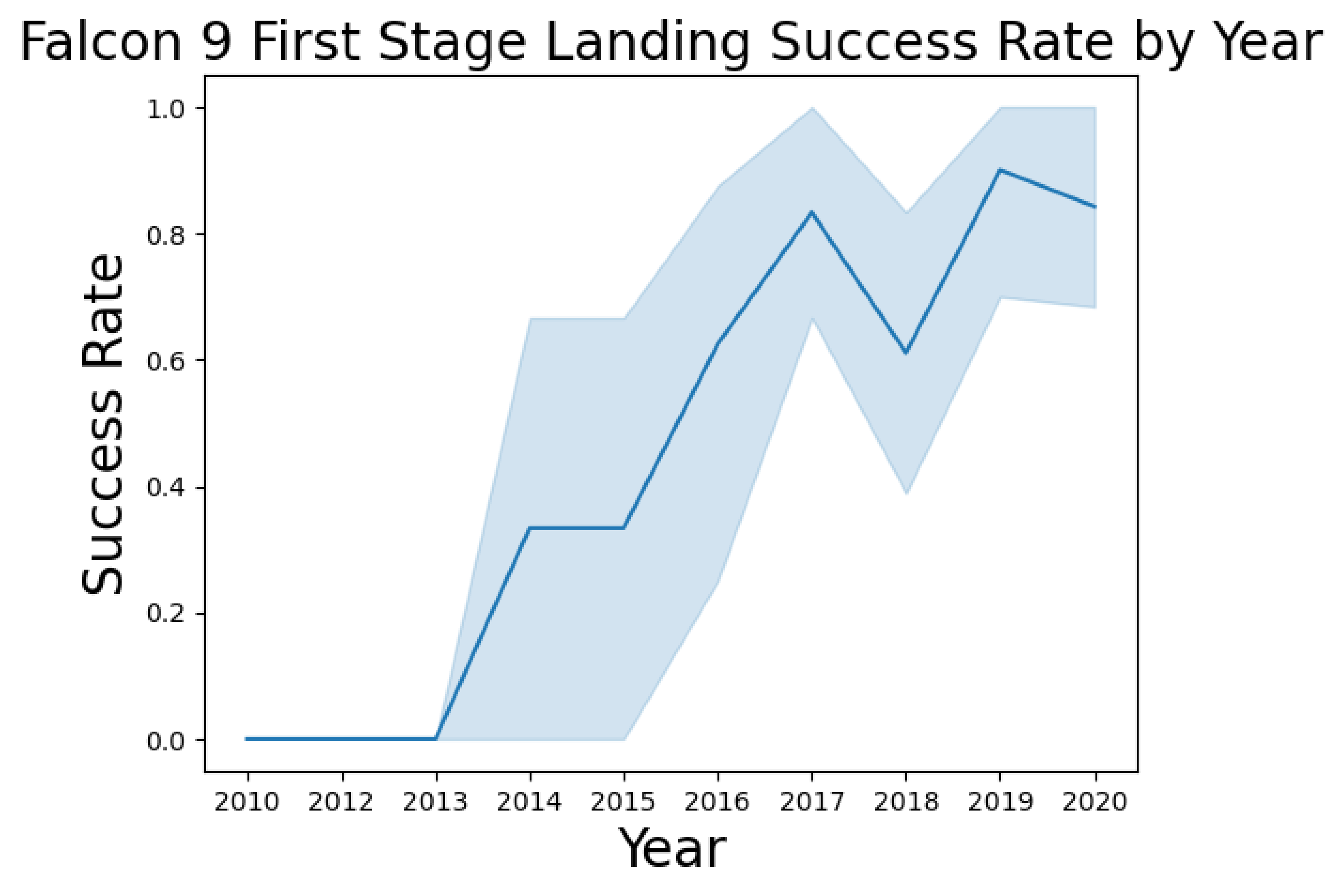
Flight Number vs. Orbit Type



Payload vs. Orbit Type



Launch Success Yearly Trend



All Launch Site Names

```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL;
```

```
 * sqlite:///my_data1.db
Done.
```

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

Launch Site Names Begin with 'CCA'

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

```
 * sqlite:///my_data1.db
Done.
```

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

```
%sql SELECT sum(PAYLOAD_MASS__KG_) AS "Total Payload Mass (kg)" FROM SPACEXTBL WHERE Customer LIKE '%NASA (CRS)%';
```

```
 * sqlite:///my_data1.db
Done.
```

Total Payload Mass (kg)
48213

Average Payload Mass by F9 v1.1

```
%sql SELECT sum(PAYLOAD_MASS__KG_) / COUNT(PAYLOAD_MASS__KG_) AS "Average Payload Mass (kg)" FROM SPACEXTBL WHERE booster_ve
```

```
 * sqlite:///my_data1.db
Done.
```

Average Payload Mass (kg)
2928

First Successful Ground Landing Date

```
%sql SELECT min(DATE) AS "First Successful Landing Outcome Date" FROM SPACEXTBL WHERE landing_outcome LIKE 'Success (ground
```

```
* sqlite:///my_data1.db
one.
```

First Successful Landing Outcome Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT DISTINCT booster_version FROM SPACEXTBL WHERE landing_outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ BE
```

```
 * sqlite:///my_data1.db
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT count(*) AS "Success" FROM SPACEXTBL WHERE landing_outcome LIKE 'Success%';
```

```
 * sqlite:///my_data1.db
Done.
```

Success
61

```
%sql SELECT count(*) AS "Failure" FROM SPACEXTBL WHERE landing_outcome not LIKE 'Success%';
```

```
 * sqlite:///my_data1.db
Done.
```

Failure
40

Boosters Carried Maximum Payload

```
%sql SELECT booster_version, payload_mass__kg_ FROM SPACEXTBL WHERE payload_mass__kg_ = (SELECT max(payload_mass__kg_) FROM
```

```
 * sqlite:///my_data1.db
Done.
```

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

```
%sql SELECT CASE substr(Date, 6, 2) WHEN '01' THEN 'January' WHEN '02' THEN 'February' WHEN '03' THEN 'March' WHEN '04' THEN
```

```
 * sqlite:///my_data1.db
Done.
```

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

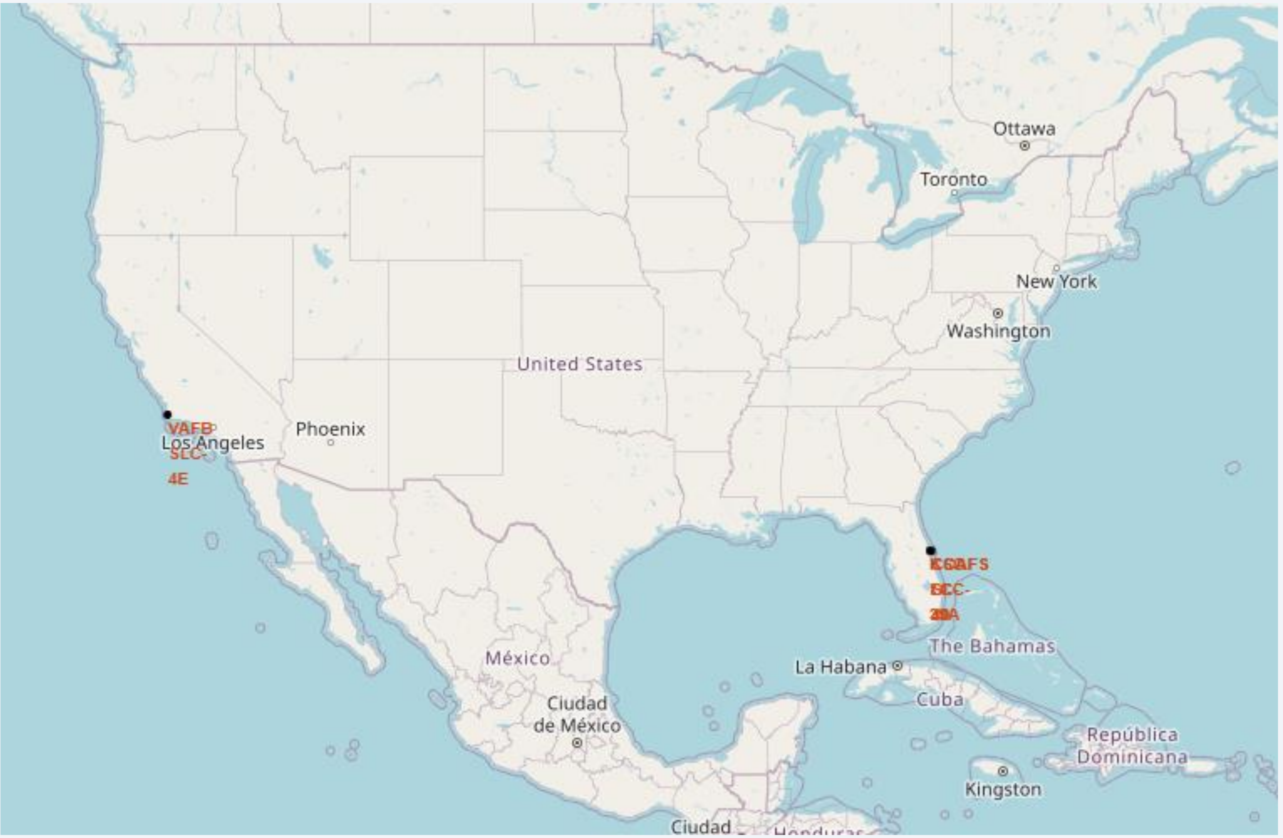
```
%sql SELECT landing_outcome, COUNT(*) AS count FROM SPACEXTBL WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY land
```

	landingoutcome	count
0	No attempt	10
1	Success (drone ship)	6
2	Failure (drone ship)	5
3	Success (ground pad)	5
4	Controlled (ocean)	3
5	Uncontrolled (ocean)	2
6	Precluded (drone ship)	1
7	Failure (parachute)	1

Section 3

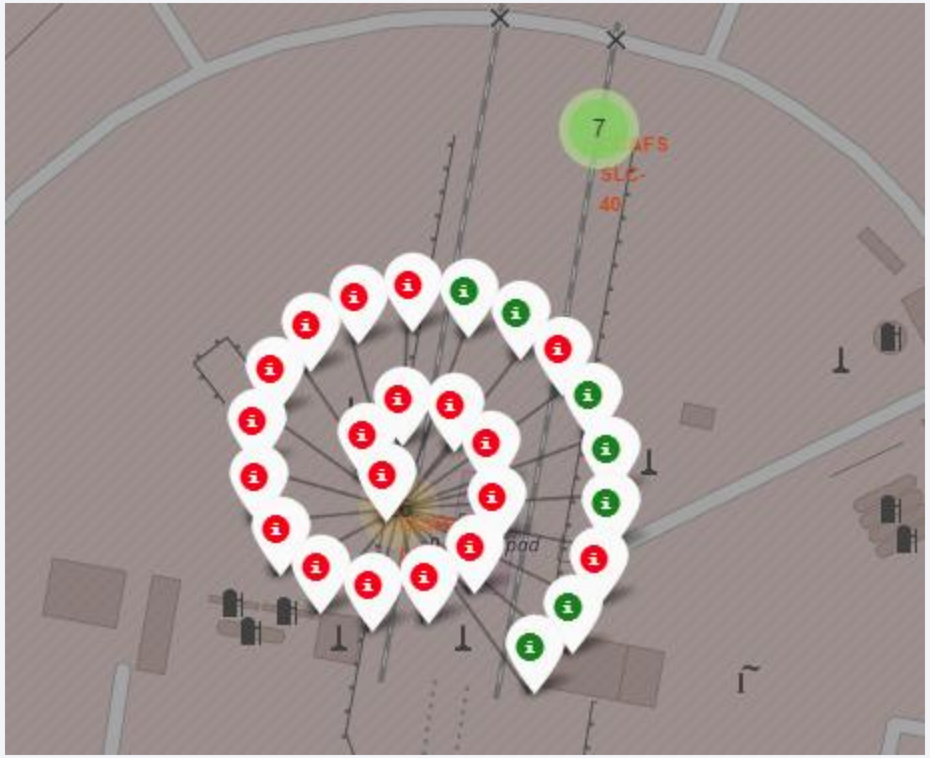
Launch Sites Proximities Analysis

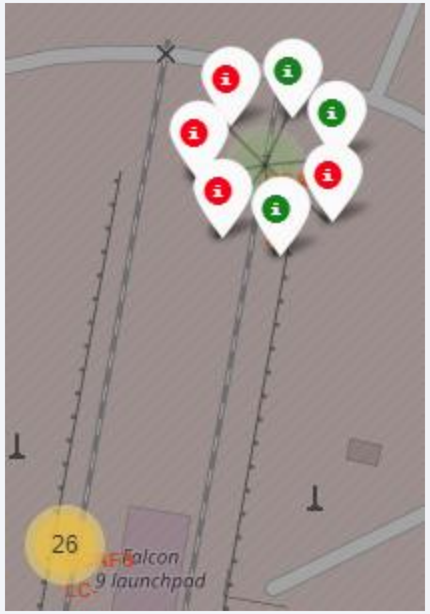
Global Launch Sites

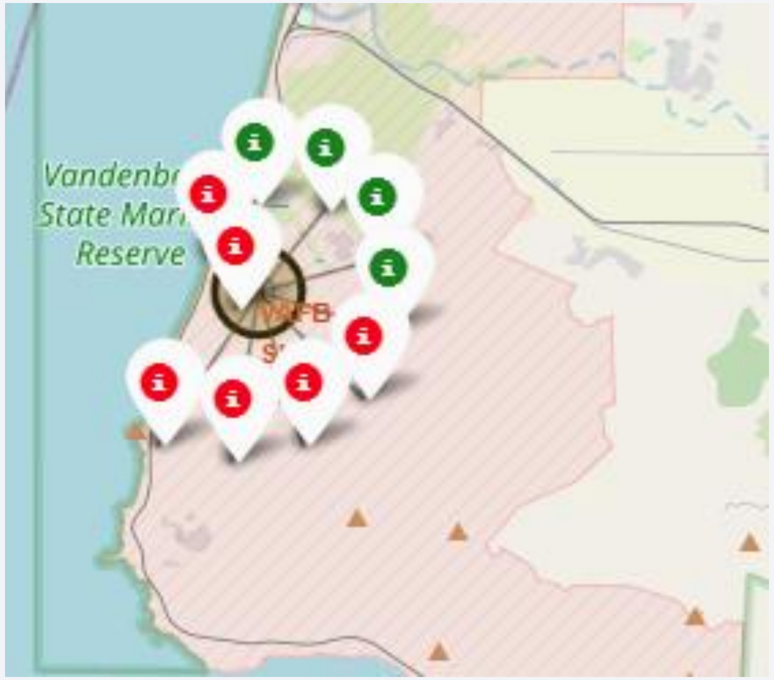


- Launch sites in Florida and California

Successful and Unsuccessful Launches



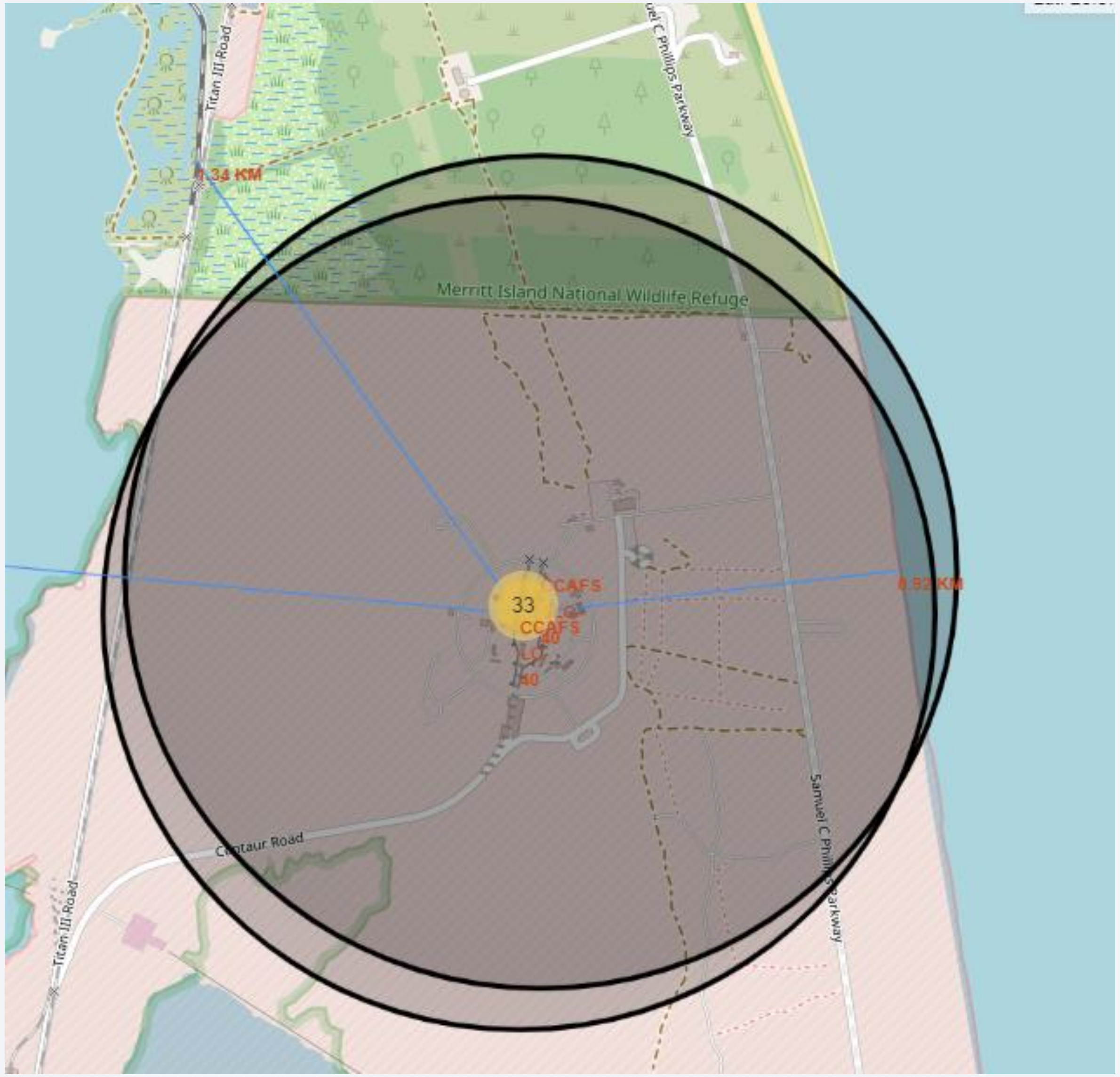




- Green denotes success, red denotes failure

Railway and coastline distances

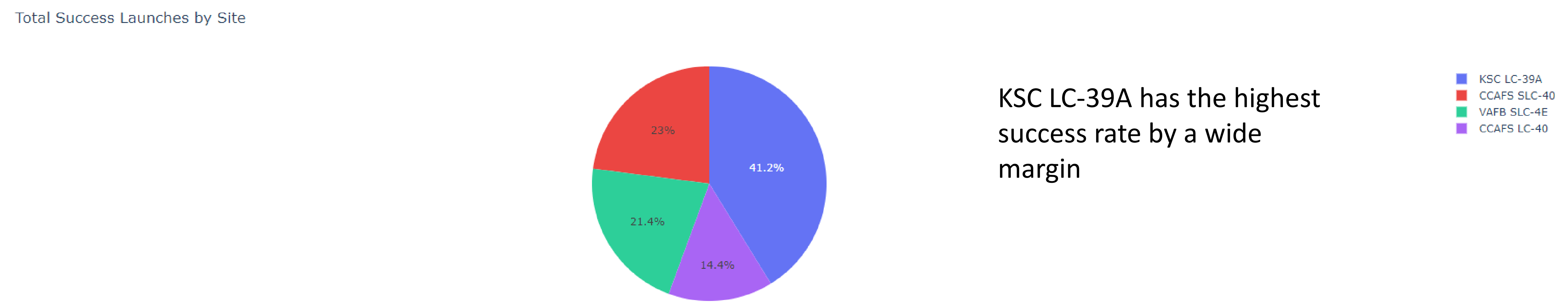
- Distance of 0.82KM designates distance to coastline
- 1.34KM designates distance to Railway



Section 4

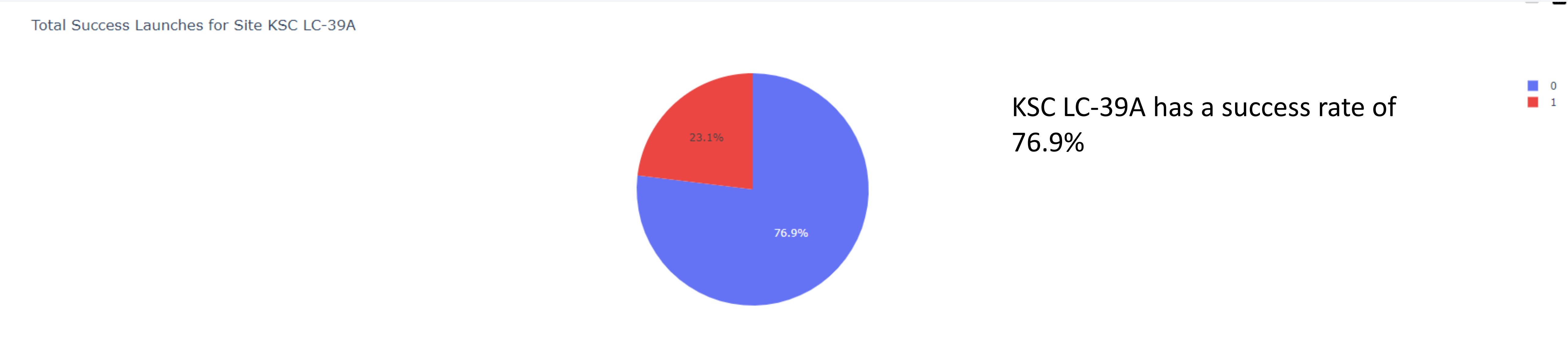
Build a Dashboard with Plotly Dash

Dashboard: Success Rate



KSC LC-39A has the highest success rate by a wide margin

DSC LC-39A Success/Failure Rate



KSC LC-39A has a success rate of 76.9%

Payload Rnage

Section 5

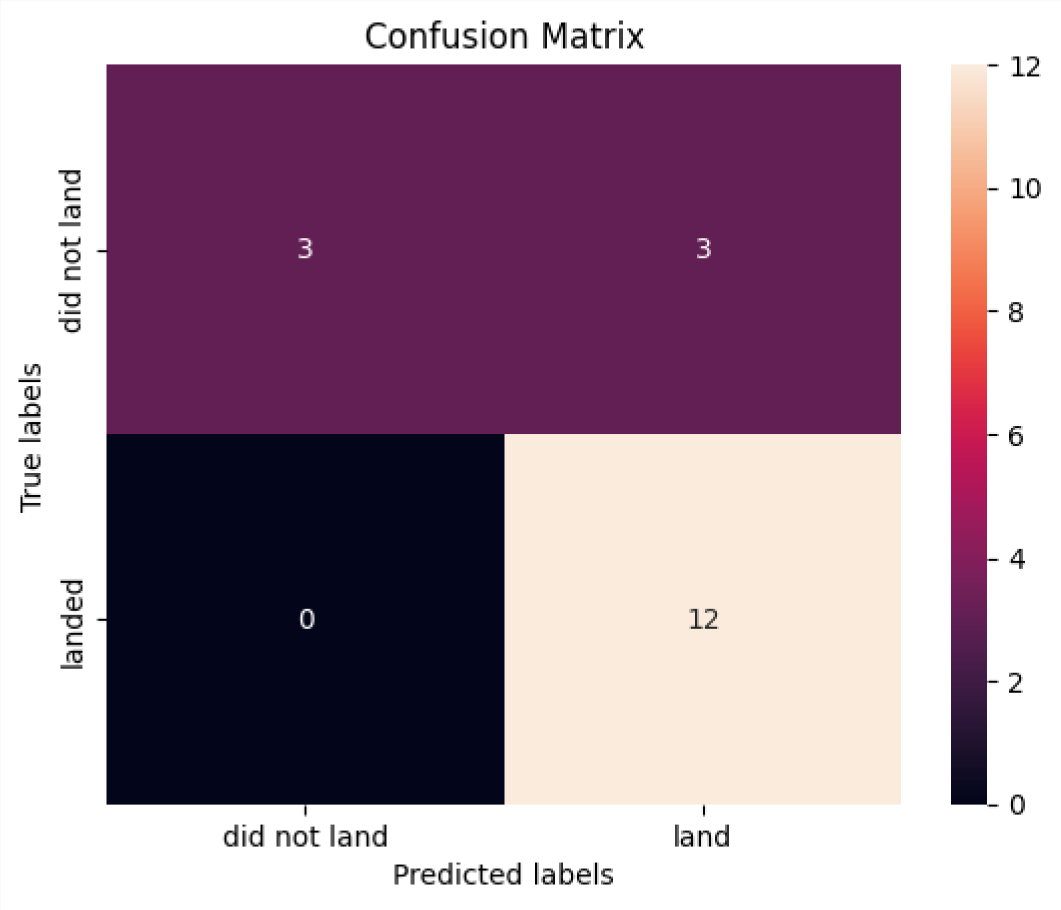
Predictive Analysis (Classification)

Classification Accuracy

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.805556	0.819444
F1_Score	0.909091	0.916031	0.892308	0.900763
Precision	0.833333	0.845070	0.828571	0.830986
Recall	1.000000	1.000000	0.966667	0.983333
Accuracy	0.866667	0.877778	0.844444	0.855556

- Accuracy at bottom

Confusion Matrix



- SVM Confusion Matrix, best performing model

Conclusions

- There is a direct correlation between the amount of launches at a site and how well a launch will go.
- Every year until 2020 there was an increase in success reate
- VLEO, ES-L1, HEO had high success rates
- KSC LC-39A had the most successful launches at any site period
- SVM had the best accuracy score compared to other models although there are more details and other tests that were taken

Thank you!