IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Leo Pacheco
9/24/2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - EDA with data visualization and SQL
  - Building an Folium interactive map
  - Building a Plotly Dash Dashboard
  - Performing predictive analysis
- Summary of all results
  - Interactive analytics
  - EDA Results
  - Results of predictive analysis

Introduction

- Project background and context

  Space X advertises on its website that's its Falcon 9 rocket launches cost 62 million dollars. Other providers cost upward of 165 million each. The differences in price is said to be because Space X reuses its first stage. Therefore if the success of the first stage landing can be predicted we can predict the cost of a launch.

- Problems you want to find answers
  - What factors contribute to a rockets successful landing?
  - What operating conditions are needed to be in place to ensure a successful landing?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Space X API
  - Web scraping from Wikipedia
- Perform data wrangling
  - Replaced problematic null values and created new target variable Landing outcomes in binary for successful and unsuccessful outcomes.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Collected data was divided into test and training sets and fitted to different classification. Then the accuracy of these models was calculated to determine the best fit.

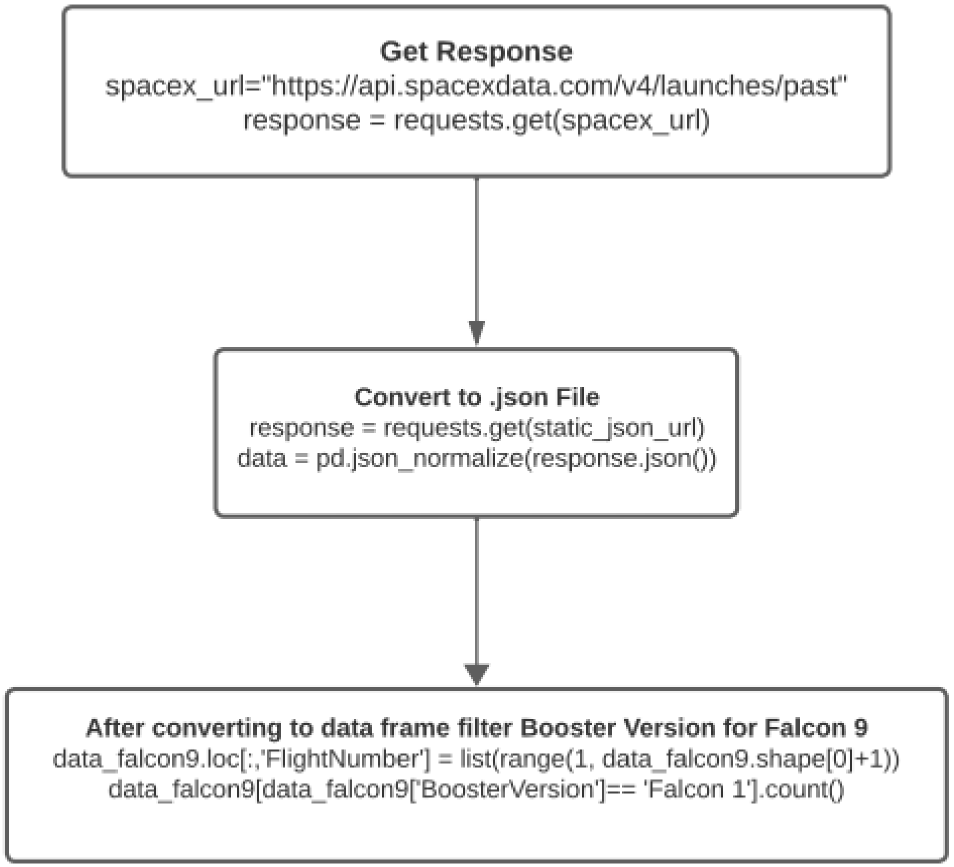
Data Collection

- Describe how data sets were collected.
- Data sets were collected directly from the Space X API and through the Space X launch Wikipedia page utilizing web scrapping techniques.

Data Collection – SpaceX API

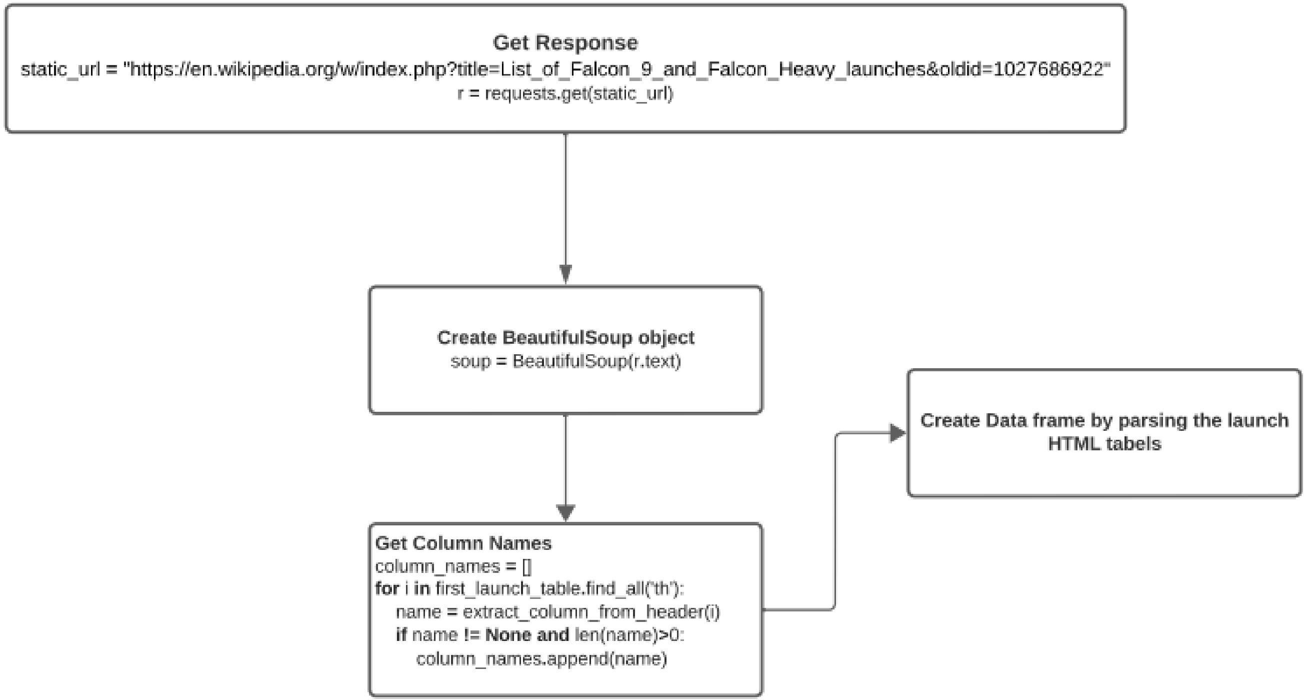
- Using public API offered by Space X data can be obtained and used using get methods.

https://github.com/leotrisol/Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



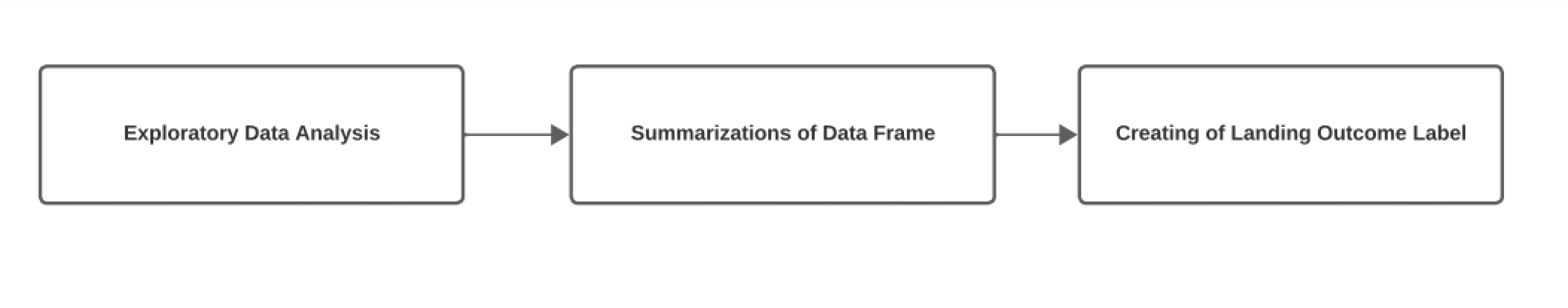
Data Collection - Scraping

- Further data can be obtained using web scraping of data tables on launch statistics from Wikipedia pages.
- https://github.com/leotrisol/Capstone/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Initial exploratory data analysis is performed calculating the occurrences of types of orbits and launches from specific sites. Null values are checked and handled. A label variable for Landing outcome is created.



- https://github.com/leotrisol/Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

- For exploratory purposes the following charts were created.
  - Payload mass X Flight Number Scatterplot
    - To examine the relation of successful returns with mass and number.
  - Launch site X Flight Number Scatterplot
    - To look for correlations with success based on flight number and site.
  - Launch Site X Payload Mass Scatterplot
    - To see if certain launch sites perfume better with more or less mass.
  - Orbit X Class Bar Chart
    - Plotted to see which sites had higher success rates.
  - Orbit and Flight Number Scatter Plot
  - Success Rate % and Year line plot
    - Plotted to see trend in success based off of year

https://github.com/leotrisol/Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- Found distinct names of landing sites
- Display 5 records where launch sites begin with the string 'CCA'
- Found total mass of payloads launched by NASA (CRS)
- Calculated average mass of payload carried by Booster F9 v1.1
- Found date where first successful landing on ground pad
- Found the boosters that had a successful drone ship landing with payload between 4000 and 6000
- Found total amount of successful and failure mission outcomes
- Found all booster names that carried the maximum payload
- Listed the month and booster, outcome and sites of launches in 2015
- Ranked the count of outcomes between 2010-06-03 and 2017-03-20

https://github.com/leotrisol/Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Markers used to indicate points of interest like launch sites
- Circles were used to indicate specific areas on the map like the NASA Johnson Space center that take up more area than just a launch site
- Marker Clusters are used to make groupings of events like launches from launch site and colored to indicate success of failure
- Lines were implemented to show distances between coastline and points of interest.

https://github.com/leotrisol/Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- Plotted in the dashboard is a pie chart to show the distribution of total launces at specific sites
- A Scatter graph is plotted to demonstrate the relationship between outcome and payload differentiated by booster version.
- The dash board had two dropdowns for selecting the data viewed to either be all or sorted into a specific launch sites data

https://github.com/leotrisol/Capstone/blob/main/spacex_dash_app.py

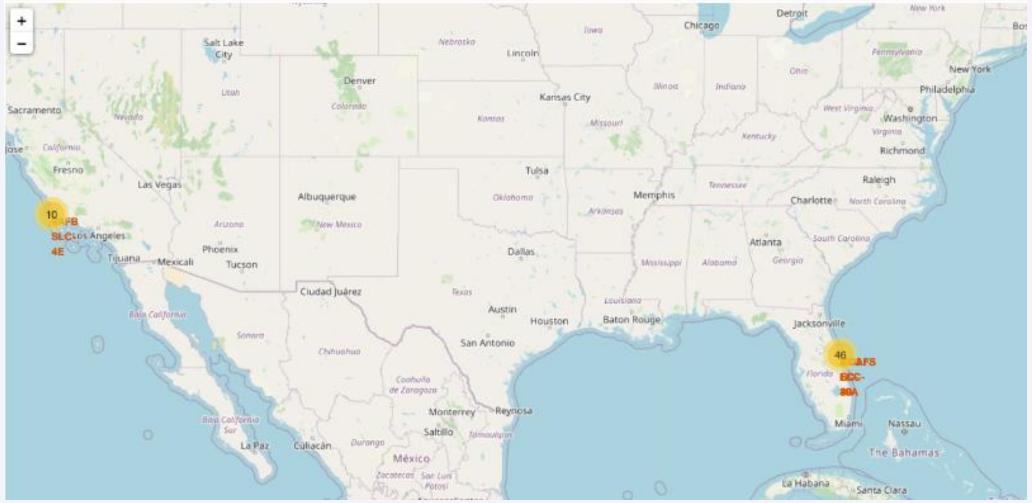
Predictive Analysis (Classification)

- Utilizing NumPy and pandas the data is transformed and split into our train and testing sets.
- These sets are then past to several different learning models and fine tuned to specific hyperparameters.
- Then these models are tested for the accuracy and metric for each model fine tuning and applying feature engineering to improve accuracy.
- Once finished the best performing model is selected based on its test metrics.

https://github.com/leotrisol/Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite%20(1).ipynb

Results

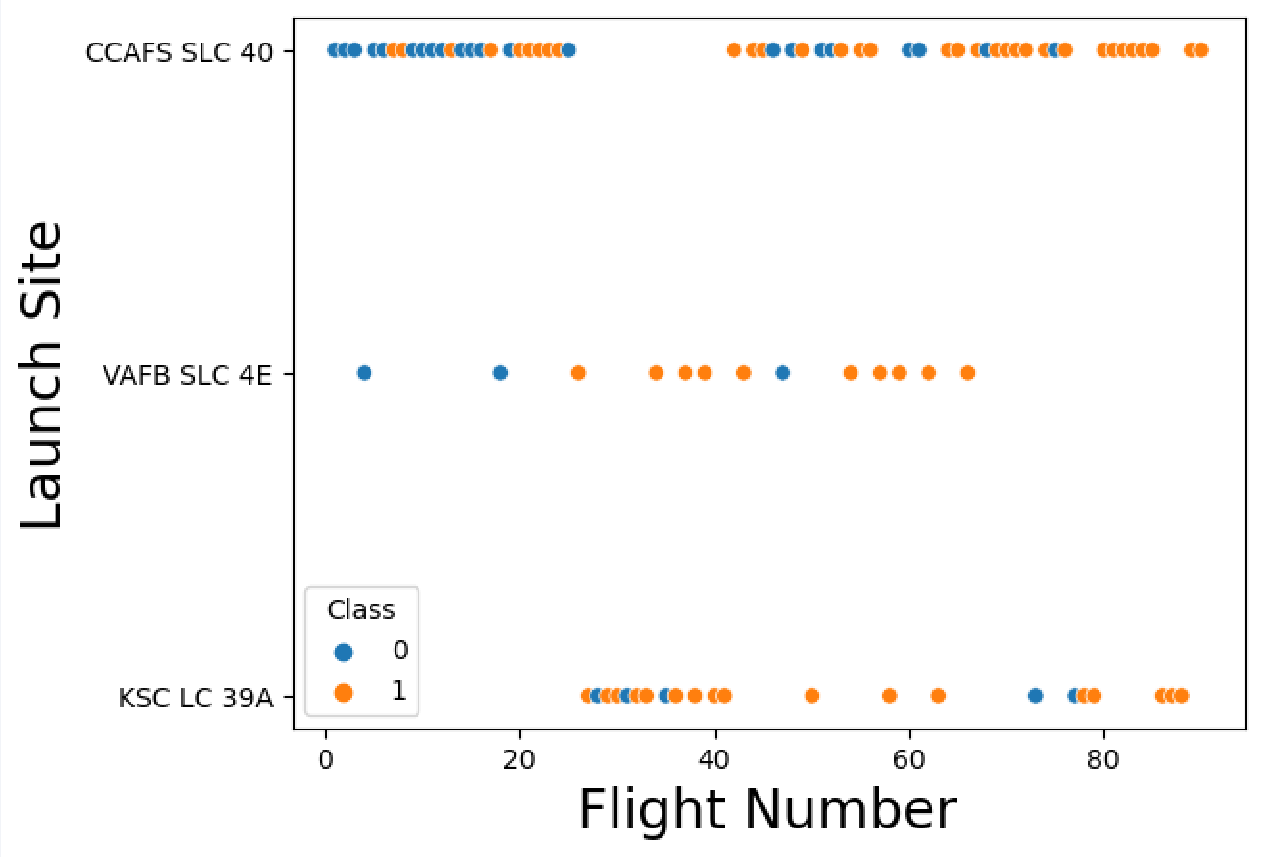
- From EDA lower weighted payloads performed better and as time passes the rates of success are increasing for all launces.
- Average payload per launch of F9 v1.1 booster is 2928 kg.
- Orbits of GEO, HEO, SSO, ES L1 had the highest success rates.



Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

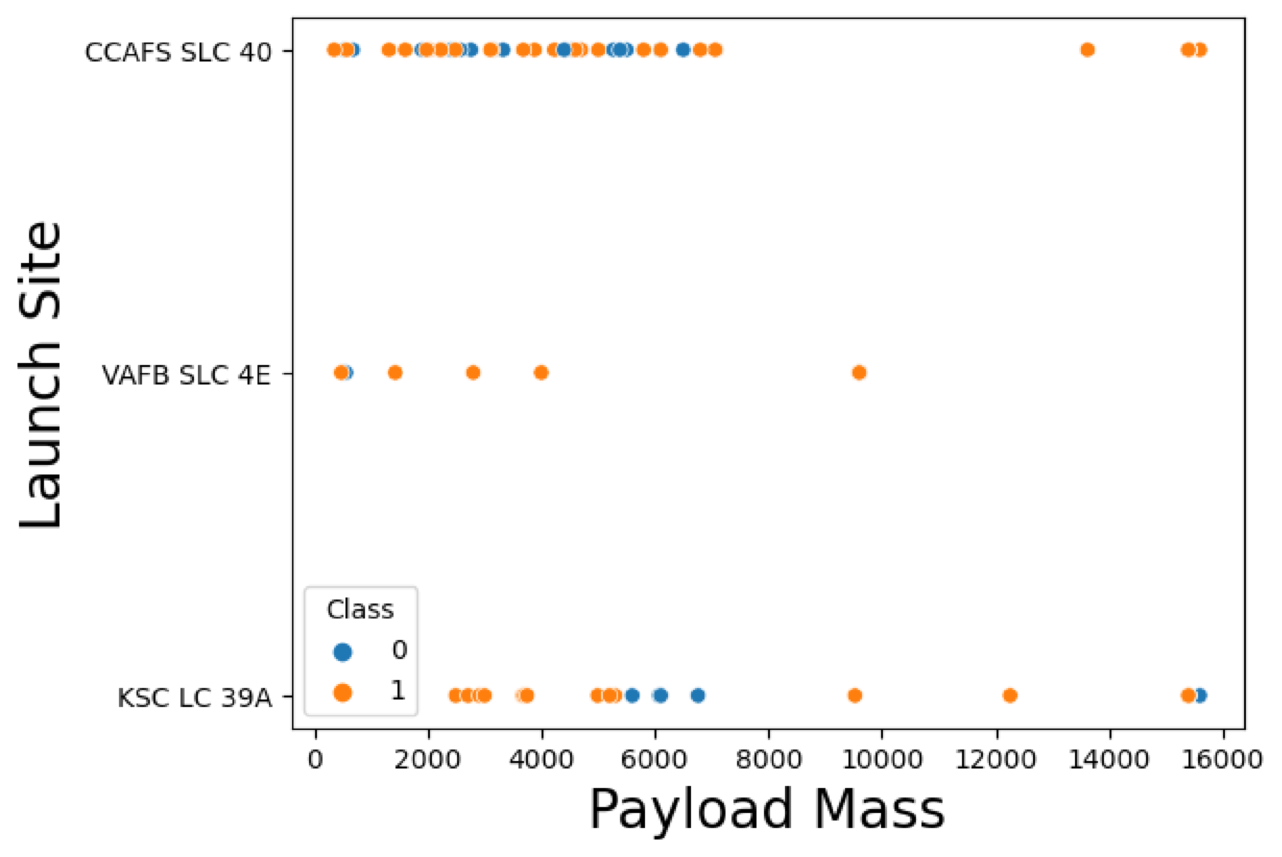


Launches from CCAFS SLC 40 are higher than other launch sites.

Further successful flights increase with flight number.

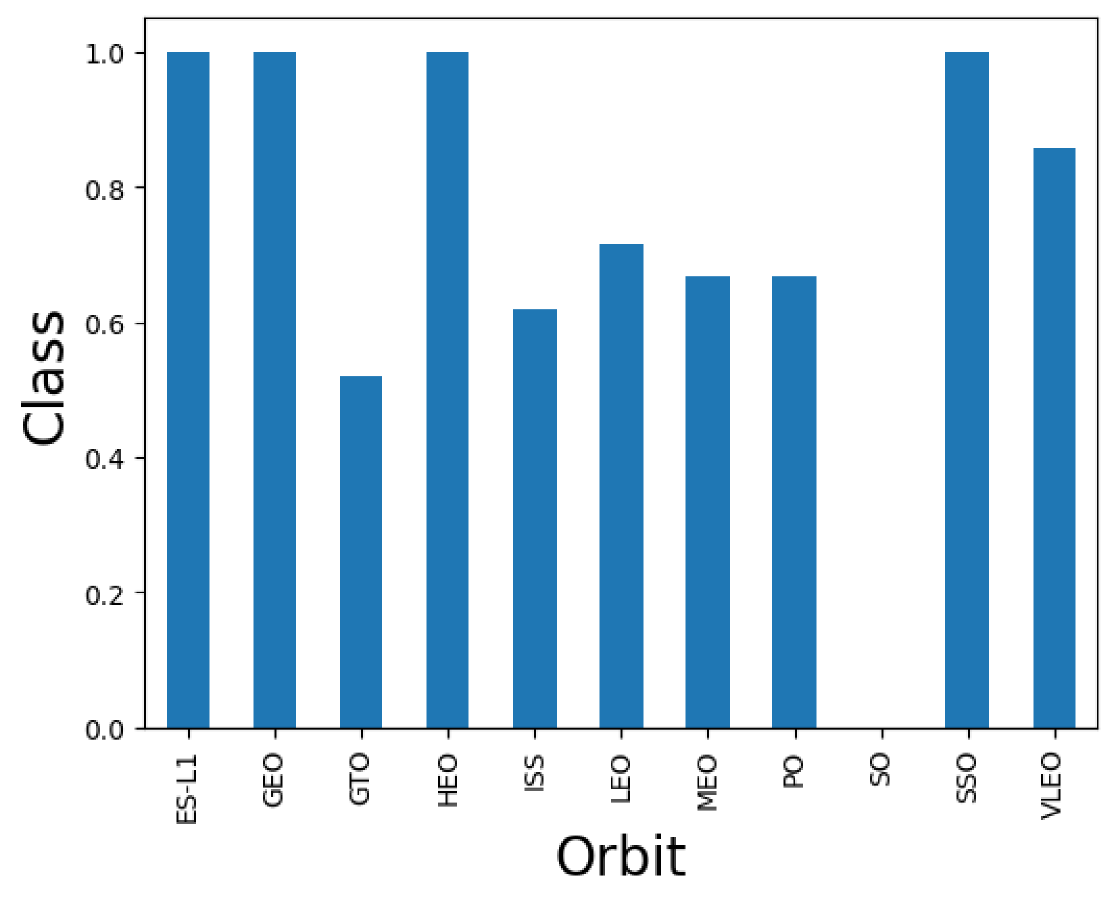
Payload vs. Launch Site

- From the graph the majority of launches have been if mass lower than 10000 kg.
- Site VAFB SLC 4E hasn't had a launch with a payload more than 10,000 kg.



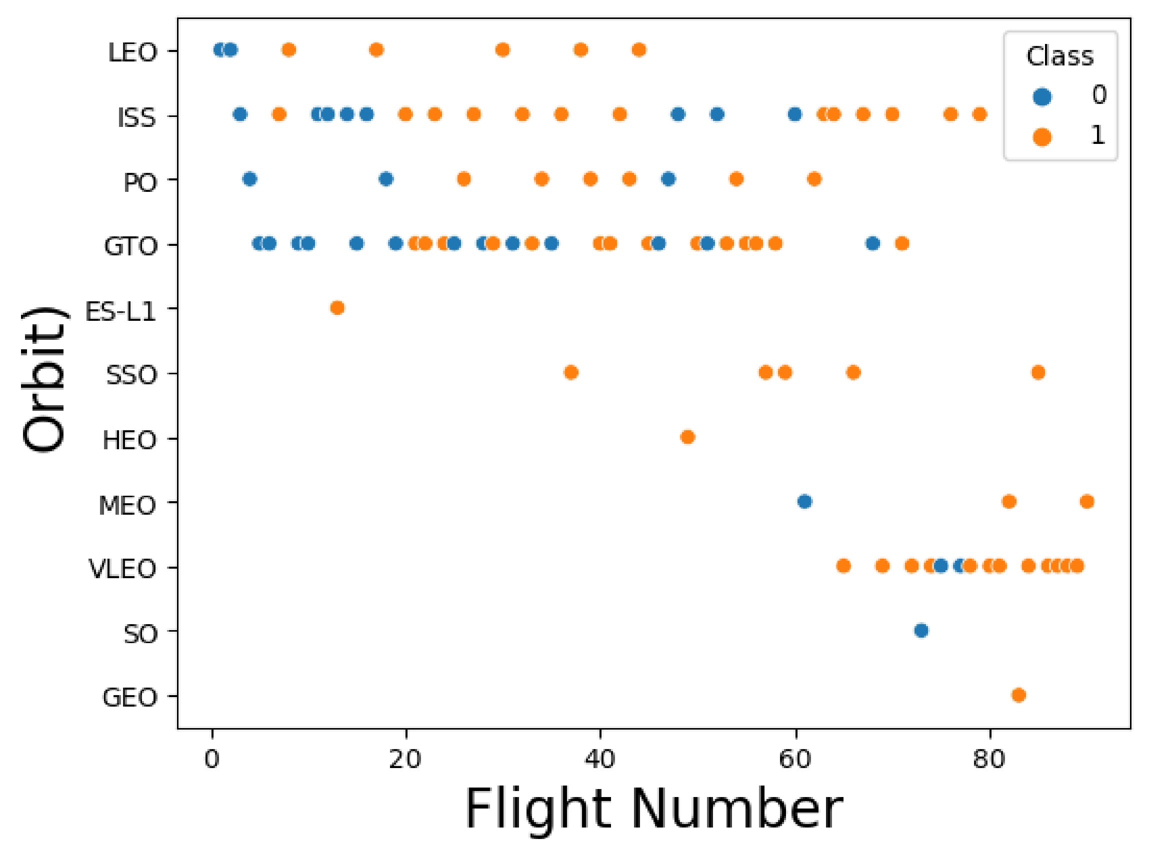
Success Rate vs. Orbit Type

- The plot shows than SSO, ES-L1, GEO, and HEO orbits have the highest rates of success.
- With SO having a success rate of 0.



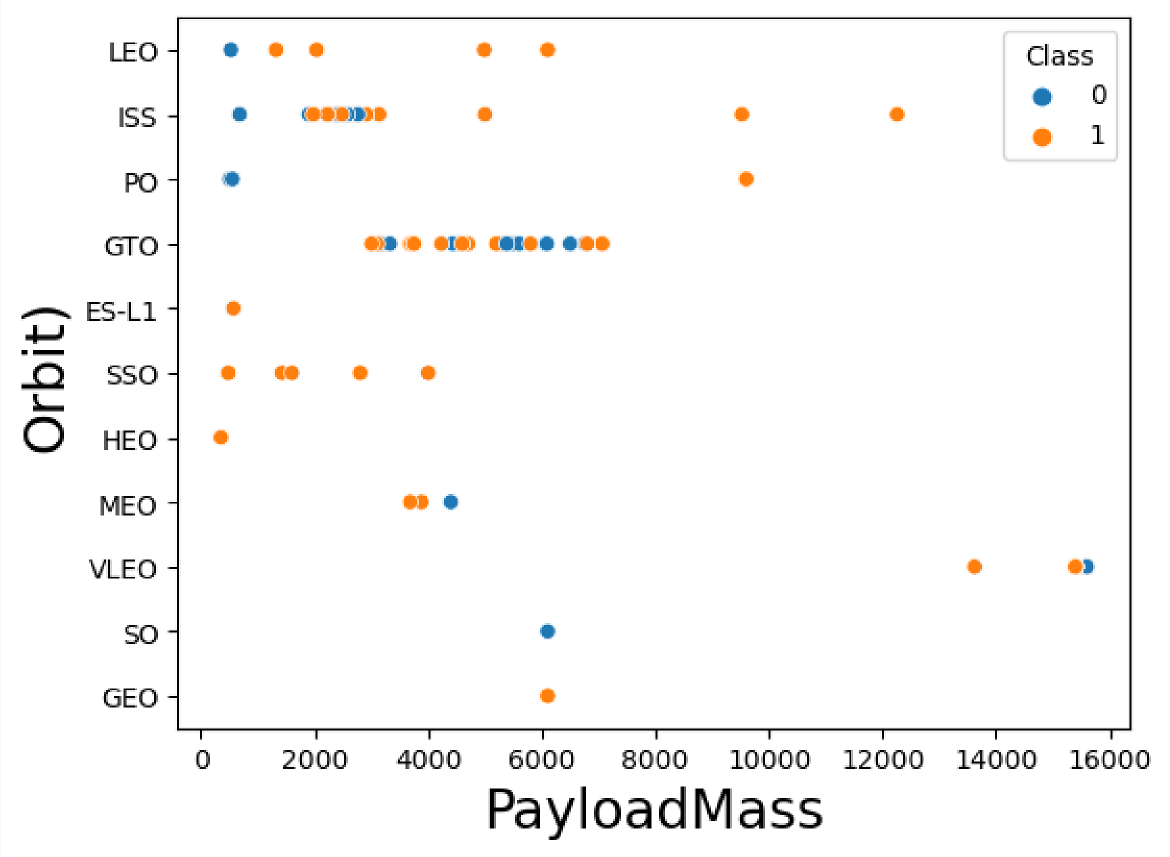
Flight Number vs. Orbit Type

- This scatter shows that VLEO orbit launces did not occurs until flight numbers surpassed 60.
- There is also a trend in LEO orbit success between increasing flight numbers.



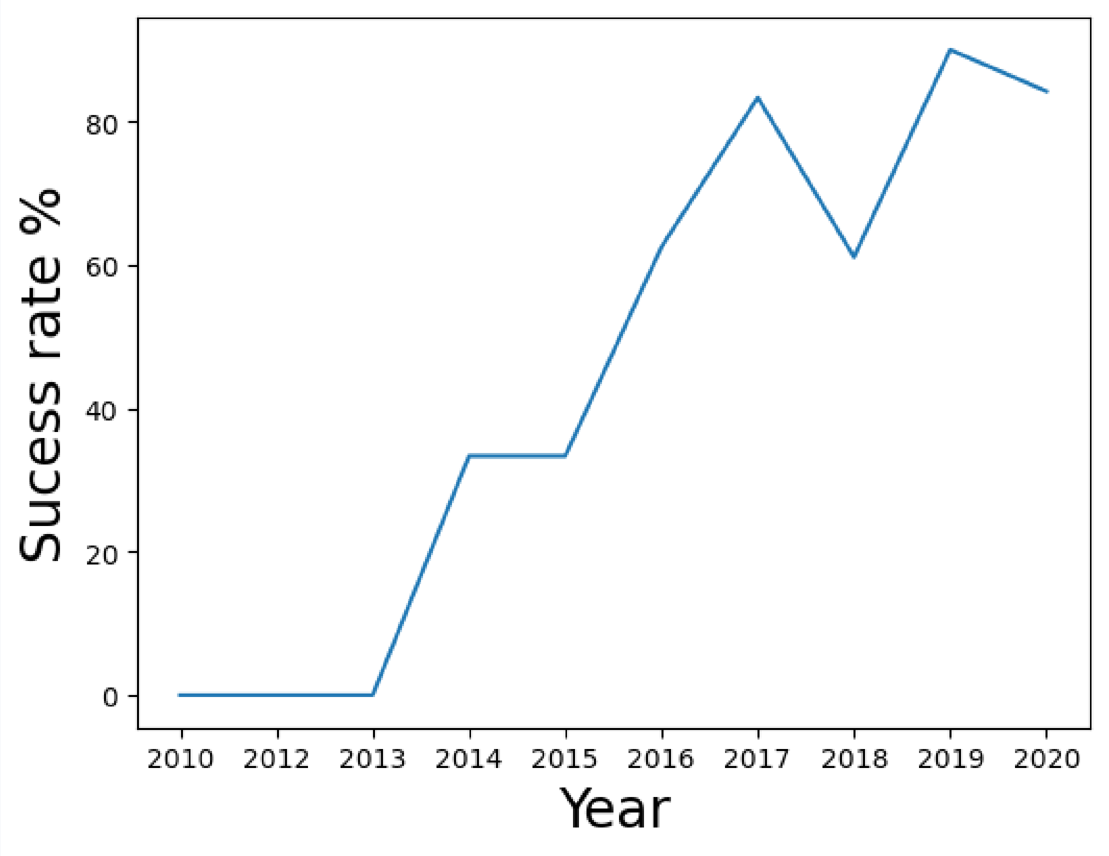
Payload vs. Orbit Type

- For PO ISS and LEO orbits heavier payloads are corelated to success launches.
- Any other launches have no apparent patter to payload mass and orbit.



Launch Success Yearly Trend

- This line graph demonstrates the trend of increasing success rate with increasing year.



All Launch Site Names

- This query was used to find the unique names of Launch sites utilizing Distinct key word.

```
In [57]: %sql SELECT Distinct Launch_Site From SPACEXTABLE

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

Out[57]:

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA` using Limit and the Like key words to filter the Launch site

```
%sql Select * From SPACEXTABLE 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
2010-04-06	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)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	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

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- Using the where function to sort the customer and the Sum key word to find the total

```
%sql Select Sum(PAYLOAD_MASS__KG_) as 'Total Payload Mass By NASA (CRS)' From SPACEXTABLE Where Customer= 'NASA (CRS)'
```

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

Total Payload Mass By NASA (CRS)
45596

Average Payload Mass by F9 v1.1

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

%sql Select AVG(PAYLOAD_MASS__KG_) as 'average payload mass by F9 v1.1 booster' From SPACEXTABLE Where Booster_Version = 'F9 v1.1’

Uses the where key work to sort the booster version and the avg key word to kind the mean.

average payload mass by F9 v1.1 booster
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- The column Date is sorted using the where key word with landing outcome equal to success (ground pad) then the MIN key work is used to find the min date. In other words the first successful ground pad landing.

```
%sql Select Min("Date") From SPACEXTABLE Where Landing_Outcome = 'Success (ground pad)'
```

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

Min("Date")
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

```
%sql Select Distinct(Booster_Version), Landing_Outcome, PAYLOAD_MASS__KG_ from SPACEXTABLE Where
Landing_Outcome = 'Success (drone ship)' and (PAYLOAD_MASS__KG_ between 4000 and 6000)
```

Booster_Version	Landing_Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

%sql Select (Select Count(*) From SPACEXTABLE where Mission_Outcome in ('Success', 'Success (payload status unclear')), (Select Count(*) From SPACEXTABLE where Mission_Outcome = 'Failure (in flight)') From SPACEXTABLE Limit 1

- Using nested statements to pull out the counts of successful filter missions and failures

(Select Count(*) From SPACEXTABLE where Mission_Outcome in ('Success', 'Success (payload status unclear'))	(Select Count(*) From SPACEXTABLE where Mission_Outcome = 'Failure (in flight)')
98	1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

%sql Select Distinct Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTABLE where PAYLOAD_MASS__KG_ = (Select Max(PAYLOAD_MASS__KG_) from SPACEXTABLE)

- Using Max to find the maximum payload and passing it to a where key word to sort the rows into only maximum payloads. Then displaying the unique booster versions that carried that amount.

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

%sql Select SUBSTRING("Date", 6, 2) as Month, Booster_Version, Landing_Outcome, Launch_Site from SPACEXTABLE where SUBSTRING("Date",1,4)= '2015' and Landing_Outcome Like 'Failure (drone ship)’

- Using substring to pull out month and year from the date string to sort into the year 2015 and displaying the information of the months launchs happened.

Month	Booster_Version	Landing_Outcome	Launch_Site
10	F9 v1.1 B1012	Failure (drone ship)	CCAFS LC-40
04	F9 v1.1 B1015	Failure (drone ship)	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 Landing_Outcome, Count(*) as total from SPACEXTABLE where "Date" between '2010-06-04' and '2017-03-20' group by Landing_Outcome order by Count(*) desc

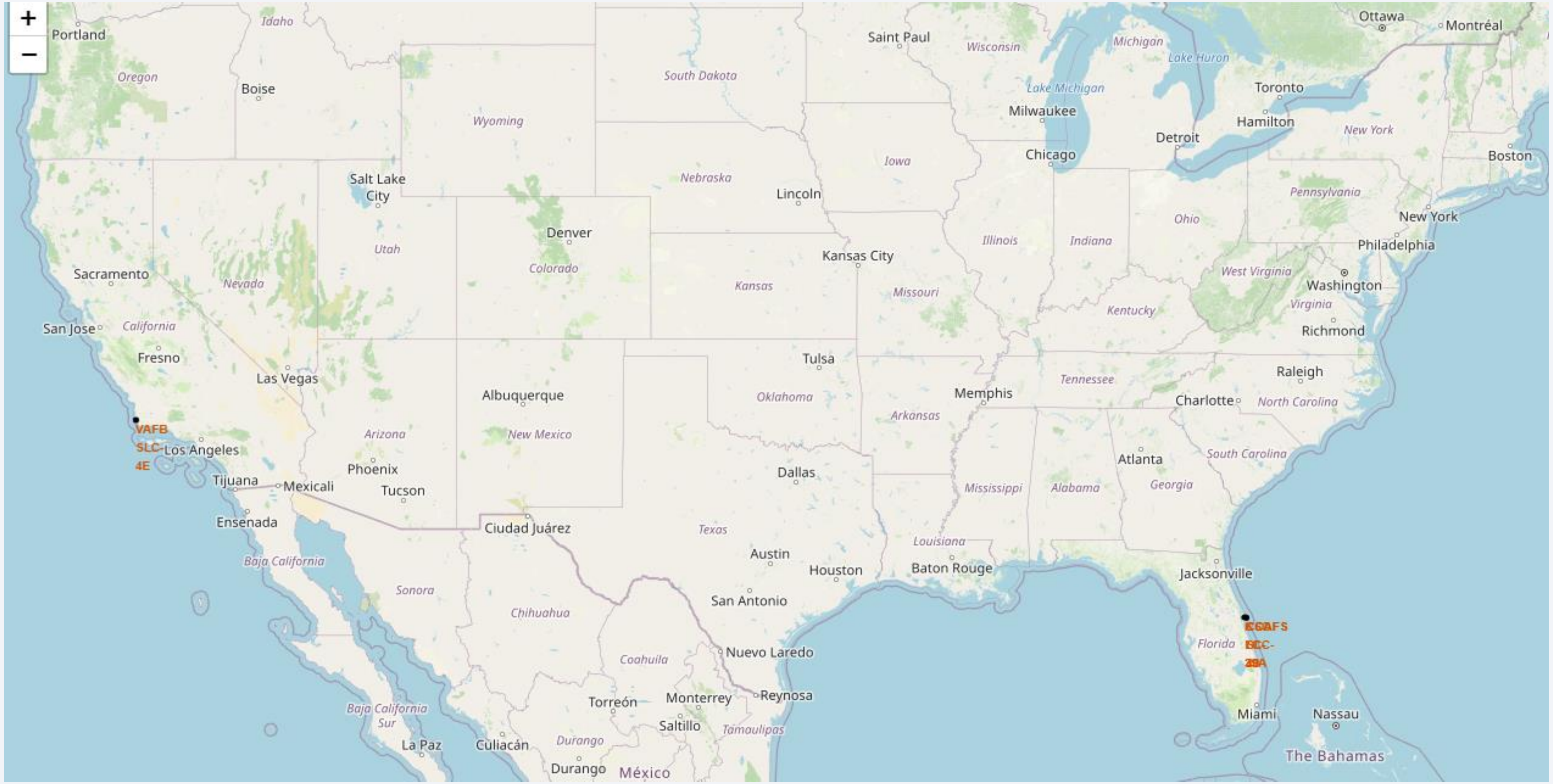
- Group by landing outcomes and filtering by the date range using the between key word. Then Ordering by desc count to get the ranking and count.

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

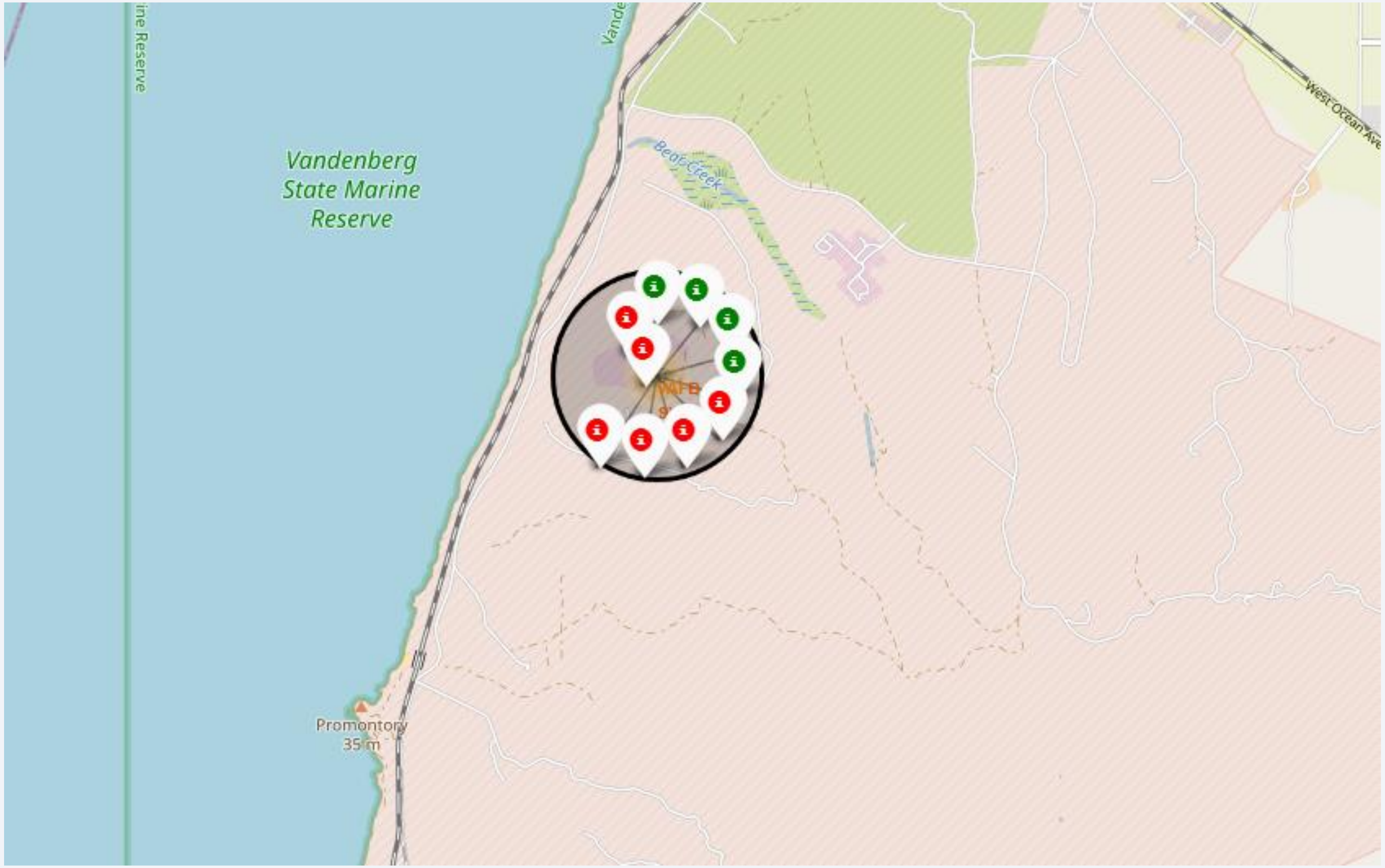
Section 3

Launch Sites Proximities Analysis

Full map of Launch Locations



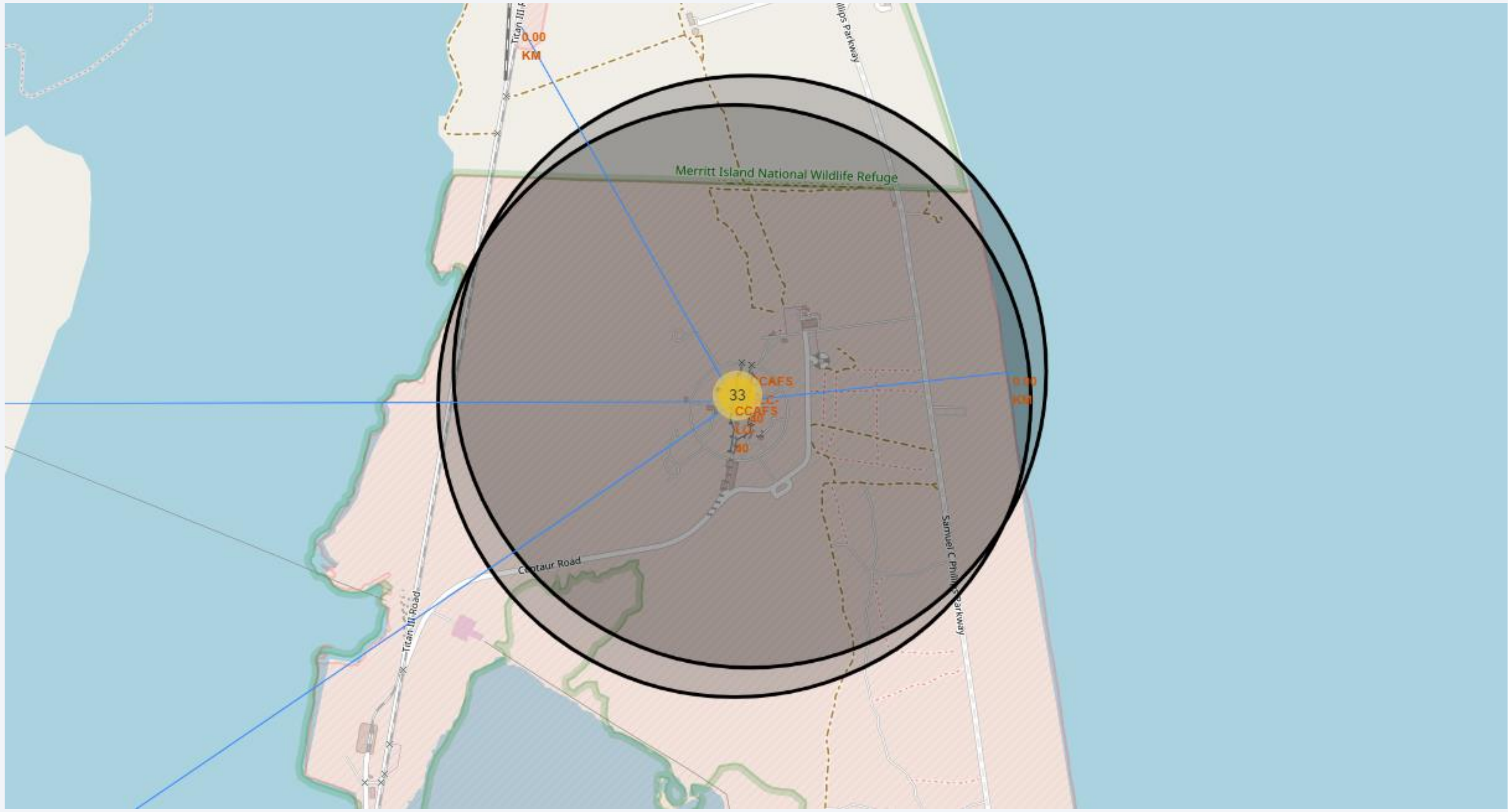
Launch site with labeled and classed launches



- Pop ups represent launches from this site.
- Red labels symbolize flailed mission outcome and green for successful.

Folium Map With Distance Lines

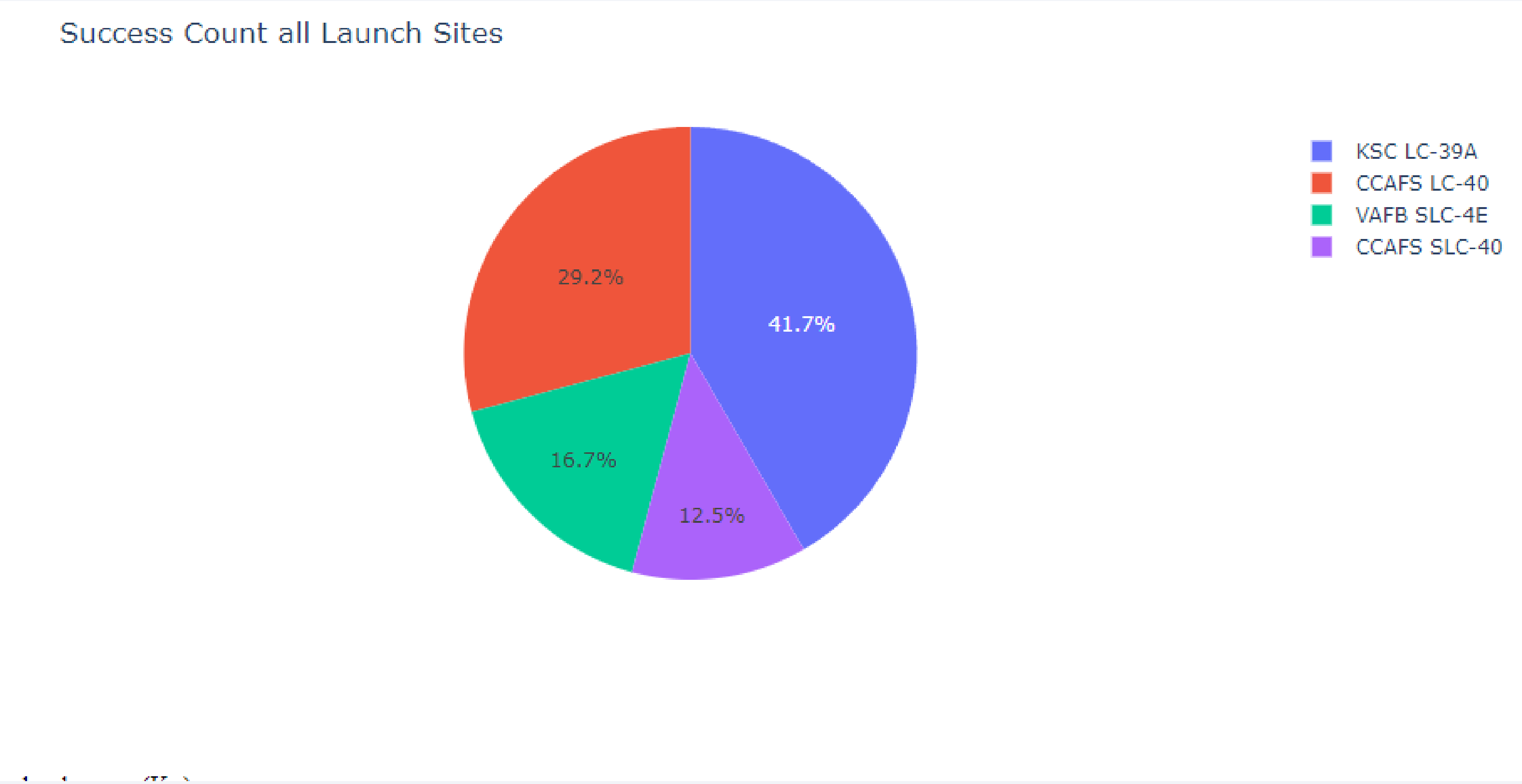
- The lines generated from the elements represent and calculate the distances from circles generated. Including km away from coast line roads and city center.



Section 4

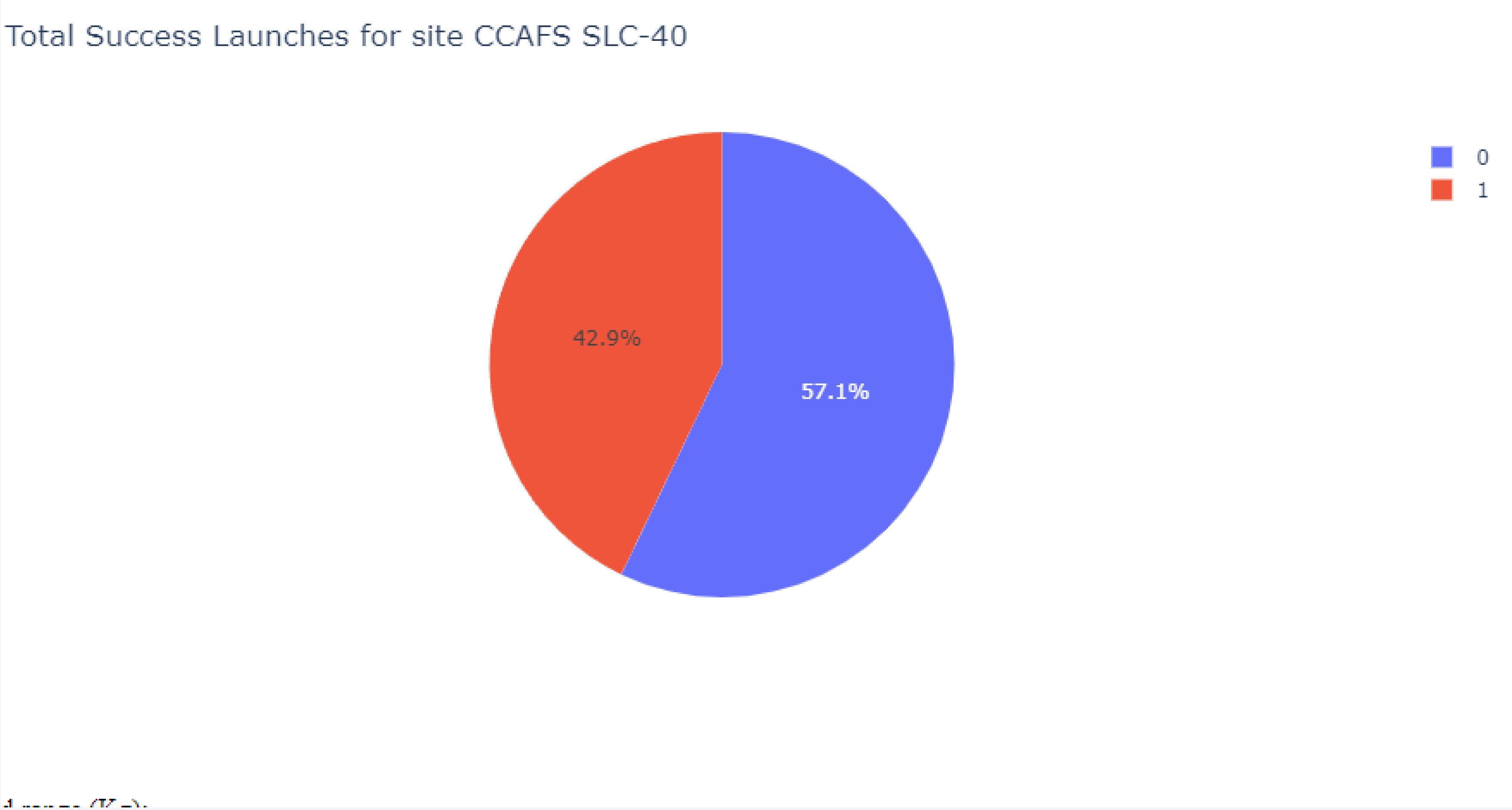
Build a Dashboard with Plotly Dash

Chart of All launches grouped by order



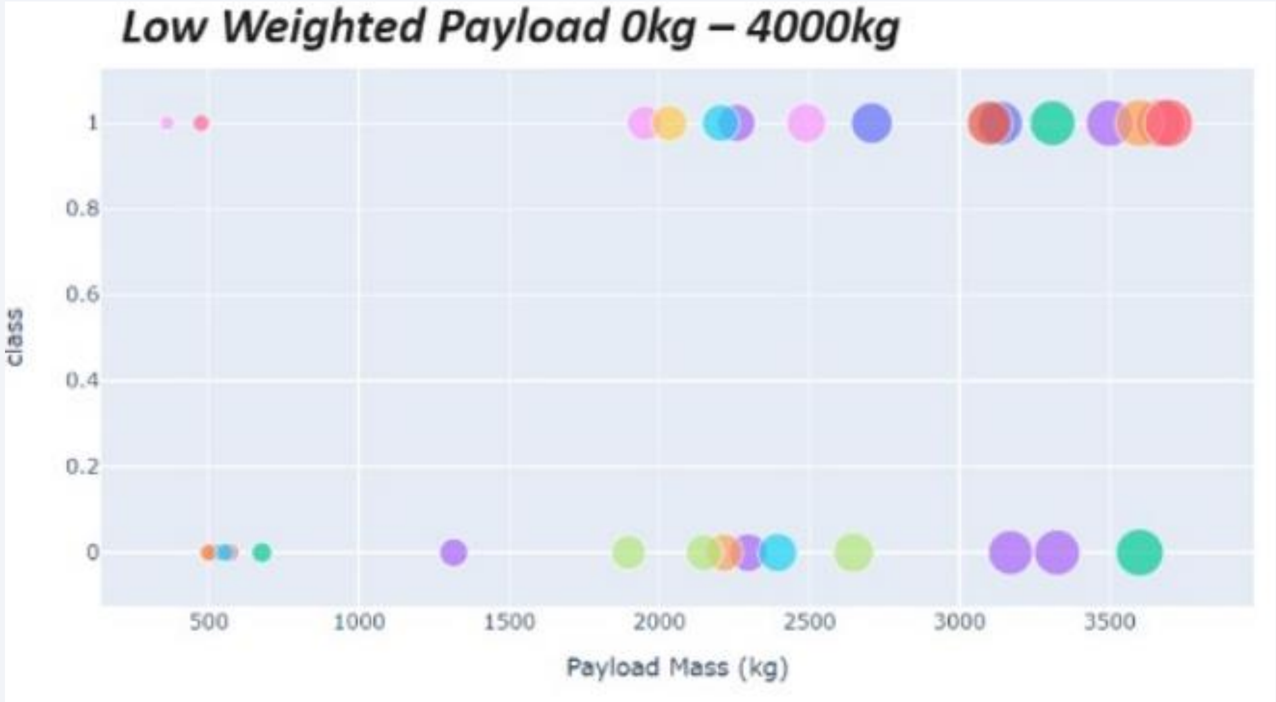
- The pie chart represents the total launches split up into the launch sites. Showing the most launches occurs at KSC LC-39A.

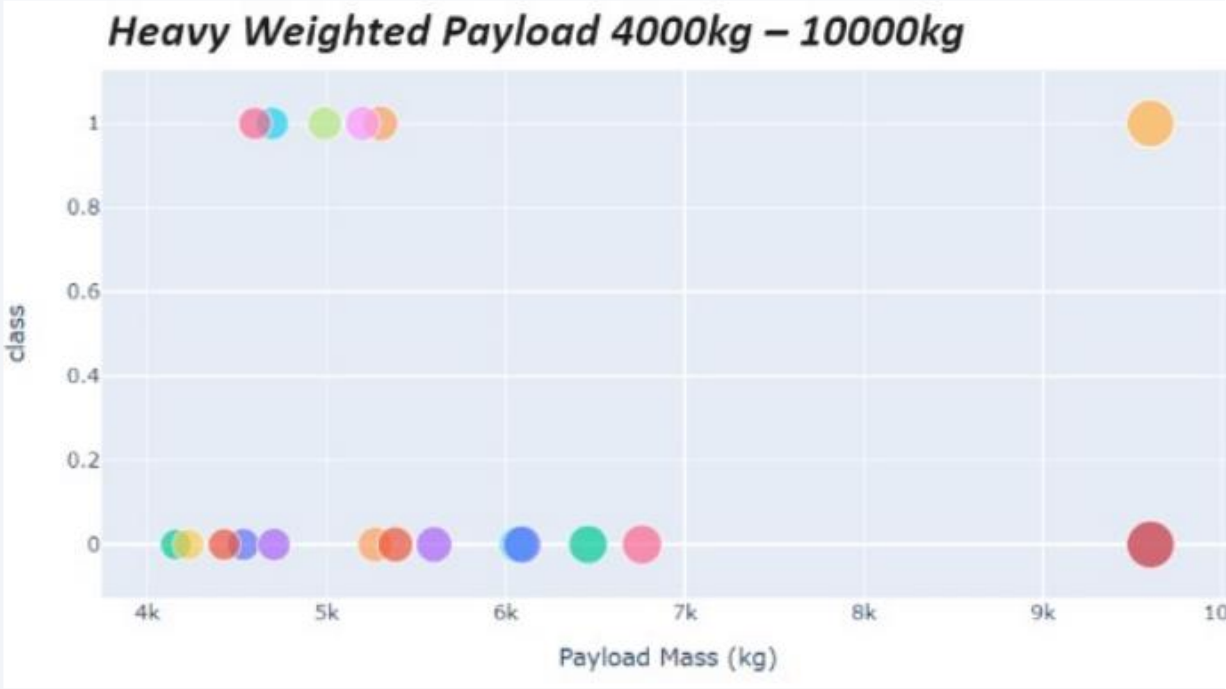
Total Launches from CCAFS SLC-40



- The chart displays the total amount of launches from CCAFS SLC-40. Split between successful launches 1 and unsuccessful 0.

Scatter plot of Weights vs Class





- Its seen that lower weighted payloads have a higher success rate then heavy weighted payloads.

Section 5

Predictive Analysis (Classification)

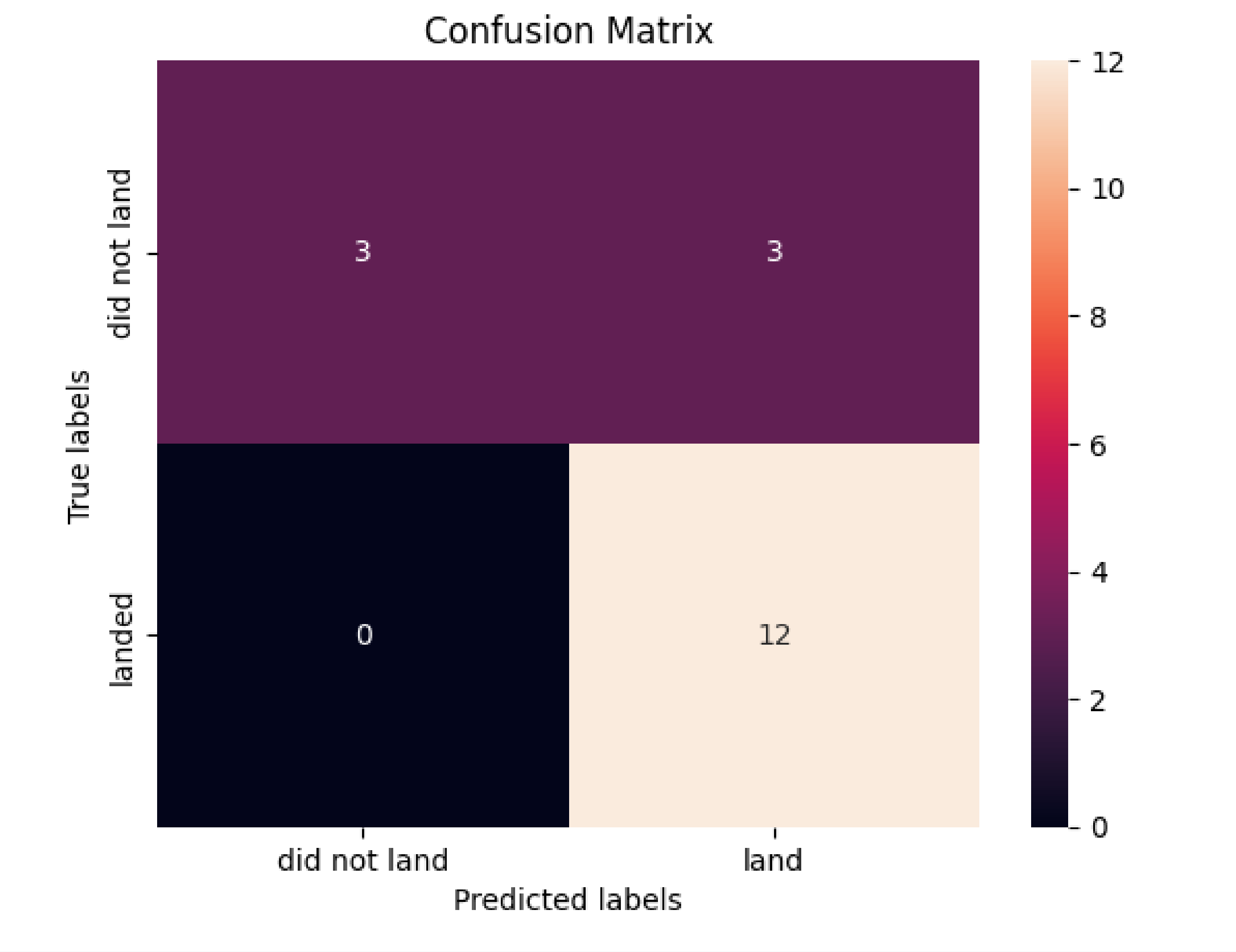
Classification Accuracy

- The accuracy of each classifier was determined to be nearly the same but for the sake of the rest of the presentation we will pick Decision tree as the best.

In [33]:
```
print('Accuracy for Logistics Regression method:', logreg_cv.score(X_test, Y_test))
print( 'Accuracy for Support Vector Machine method:', svm_cv.score(X_test, Y_test))
print('Accuracy for Decision tree method:', tree_cv.score(X_test, Y_test))
print('Accuracy for K nearsdt neighbors method:', knn_cv.score(X_test, Y_test))
```

```
Accuracy for Logistics Regression method: 0.8333333333333334
Accuracy for Support Vector Machine method: 0.8333333333333334
Accuracy for Decision tree method: 0.8333333333333334
Accuracy for K nearsdt neighbors method: 0.8333333333333334
```

Confusion Matrix



Conclusions

- Lower weighted payloads have a higher rate of success
- The launch site that had the highest rates of success was KSC LC-39A, should further examine infrastructure of that launch site to determine if it had a relationship to success rates.
- All the classifiers did a decent job of prediction outcomes so either more data or fine tuning is needed to find the best fit.
- There has been a positive correlation between success rates and they year/number if launch.
- The Orbits with the most successful mission outcome is ES-L1, HEO, VLEO and SSO.


Thank you!