IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Tran Dinh Bao
09/06/2022

Outline

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

Executive Summary

Methodologies were used to collect, analyze and predict SpaceX Falcon 9 first stage Landing

- Collecting data by calling API and scraping
- Wrangling data and Exploratory Data Analysis with visualization and SQL
- Build an Interactive Map with Folium and Dashboard with Plotly Dash
- Build a machine learning pipeline to predict if the first stage of the Falcon 9 lands successfully

Summary of all results

- Data was collected by calling API from SpaceX or scraping from wiki
- Data was standardized and visualized by wrangling and visualization methods
- Find out the rate of the first stage of the Falcon 9 lands successfully by some ML methods

Introduction

Project background and context

SpaceX 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.

Problems you want to find answers

- Determine if the first stage will land, we can determine the cost of a launch
- Predict if the first stage will land given the data from the preceding labs

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Collect data using SpaceX API
  - Collect data using web scraping from Wiki
- Perform data wrangling
  - Analyze data and create label for determining column
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Split data into training data and test data to find the best Hyperparameter for SVM, Classification Trees, and Logistic Regression.
  - Find the method that performs best using test data.

Data Collection – SpaceX API

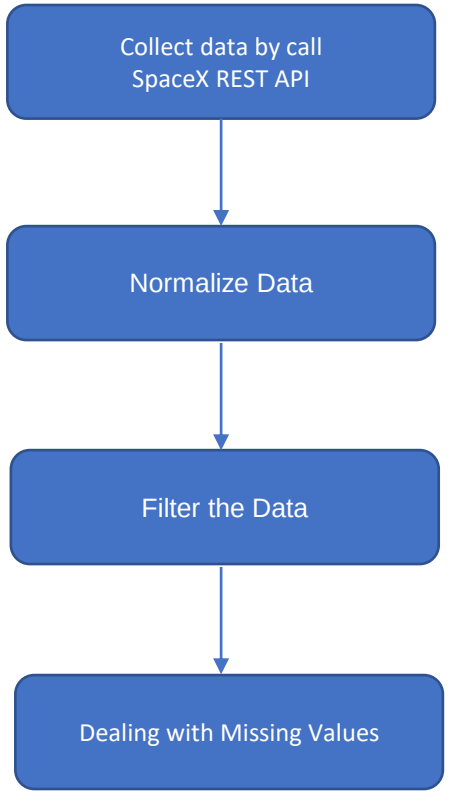
Collecting data by call SpaceX API

- Collect booster version data
- Collect launch site data
- Collect payload data
- Collect core data

Clean the requested data

- Normalize data
- Filter data
- Dealing with missing data

Python for collecting data by call API: [Applied-Data-Science-Capstone/Data Collection API Lab.ipynb]



Data Collection - Scraping

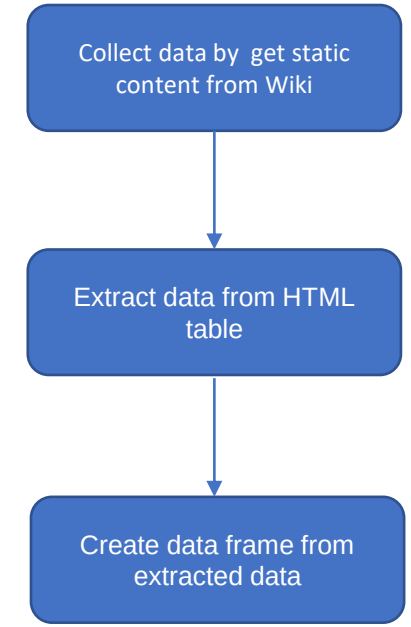
Collecting data by scraping

Scrap data from web static content (Wiki)

Extract the data from data raw (HTML)

Parse the table and convert it into a Pandas data frame

Reference for collecting scraping: [Applied-Data-Science-Capstone/Web Scraping lab.ipynb]

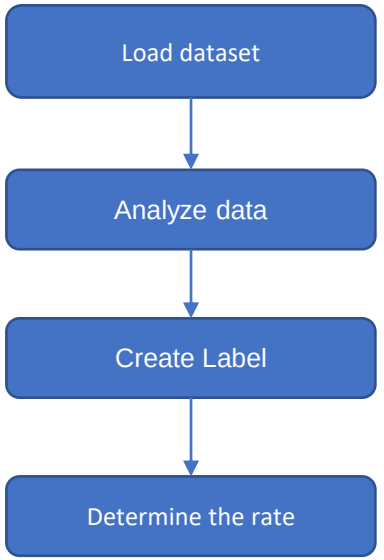


Data Wrangling

How data is processed:

- Load dataset from CSV file
- Exploratory Data Analysis
- Determine Training Labels
- Determine the success rate

Python for data wrangling: [Applied-Data-Science-Capstone/EDA Wrangling lab.ipynb]



EDA with Data Visualization

Visualize analyzed data

- Visualize the relationship between Flight Number and Launch Site by scatter plot. Using scatter plot because it helps to show which Flight Numbers are used for each Launch Site
- Visualize the relationship between Payload and Launch Site by scatter plot because it is easy to show how rockets launched for heavy payload mass are distributed for each Launch Site
- Visualize the relationship between success rate of each orbit type using bar chart because it easy to find which orbits have high success rate
- Visualize the relationship between Flight-Number and Orbit type by scatter plot because it helps to 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
- Visualize the launch success yearly trend by line chart because it helps to see how the success rate since 2013 kept increasing till 2020

Python for Data Visualization: [Applied-Data-Science-Capstone/EDA with Visualization lab.ipynb at master · baotd86/Applied-Data-Science-Capstone (github.com)]

EDA with SQL

Exploratory Data Analysis with SQL

- Display unique launch site from table SpeaceX using SELECT query with UNIQUE
- Display 5 records where launch sites begin with the string 'CCA’ using SELECT query with WHERE and LIMIT
- Display total payload mass carried by boosters launched by NASA (CRS) using SELECT query with SUM and WHERE
- Display average payload mass carried by booster version F9 v1.1 using SELECT with AVG function
- List the date when the first successful landing outcome in ground pad was achieved using SELECT with MIN function
- Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg
- Total number of successful and failure mission outcomes
- Names of the booster versions which have carried the maximum payload mass
- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.

Python for Data Visualization: Exploratory Data Analysis using SQL

Build an Interactive Map with Folium

Building the maps with markers, objects to analyze Launch Site location:

- Add the Circle object to show the area of Launch Site
- Add Marker to highlight and show information for an area
- Add Polyline with a label to visualize the distance between to points

Python for Map: Applied-Data-Science-Capstone/Data Visualization with Folium

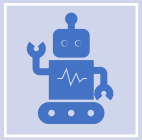
Build a Dashboard with Plotly Dash

Display the total successful launches count for launch sites:

- Using dropdown to select a specific launch site or all launch sites
- Using pie chart to present to show the total successful launches count for all sites or success/fail for the site

Python for Plotly Dash: Applied-Data-Science-Capstone/spacex_dash_app.py

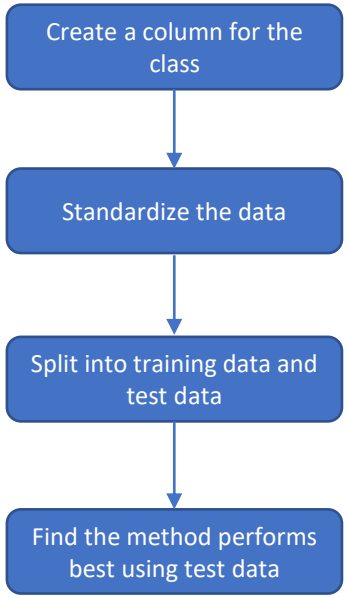
Predictive Analysis (Classification)



Exploratory Data Analysis and determine Training Labels

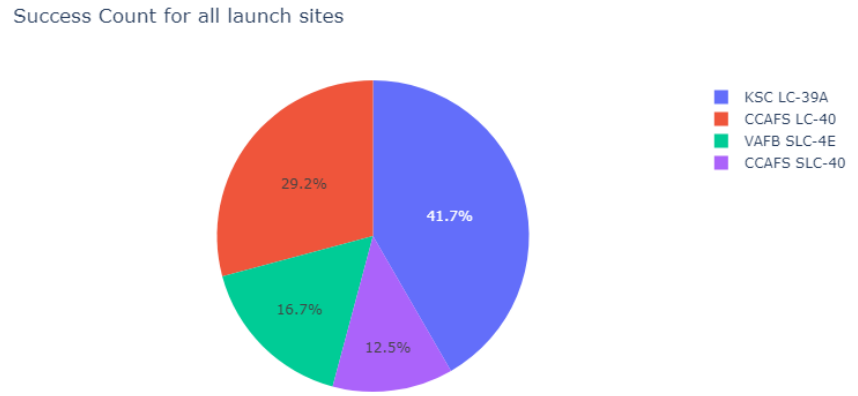
- Load data
- Standardize data
- Select method and train data
- Calculate the accuracy

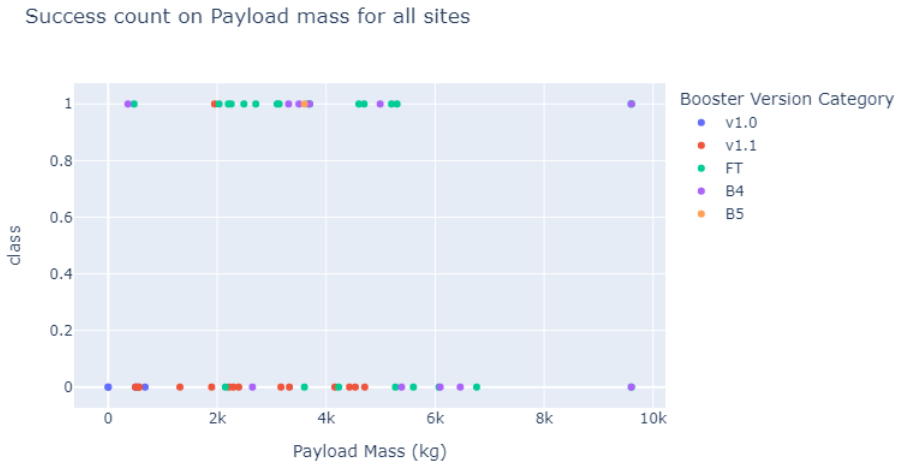
Python for Map: [Applied-Data-Science-Capstone/Machine Learning Prediction]

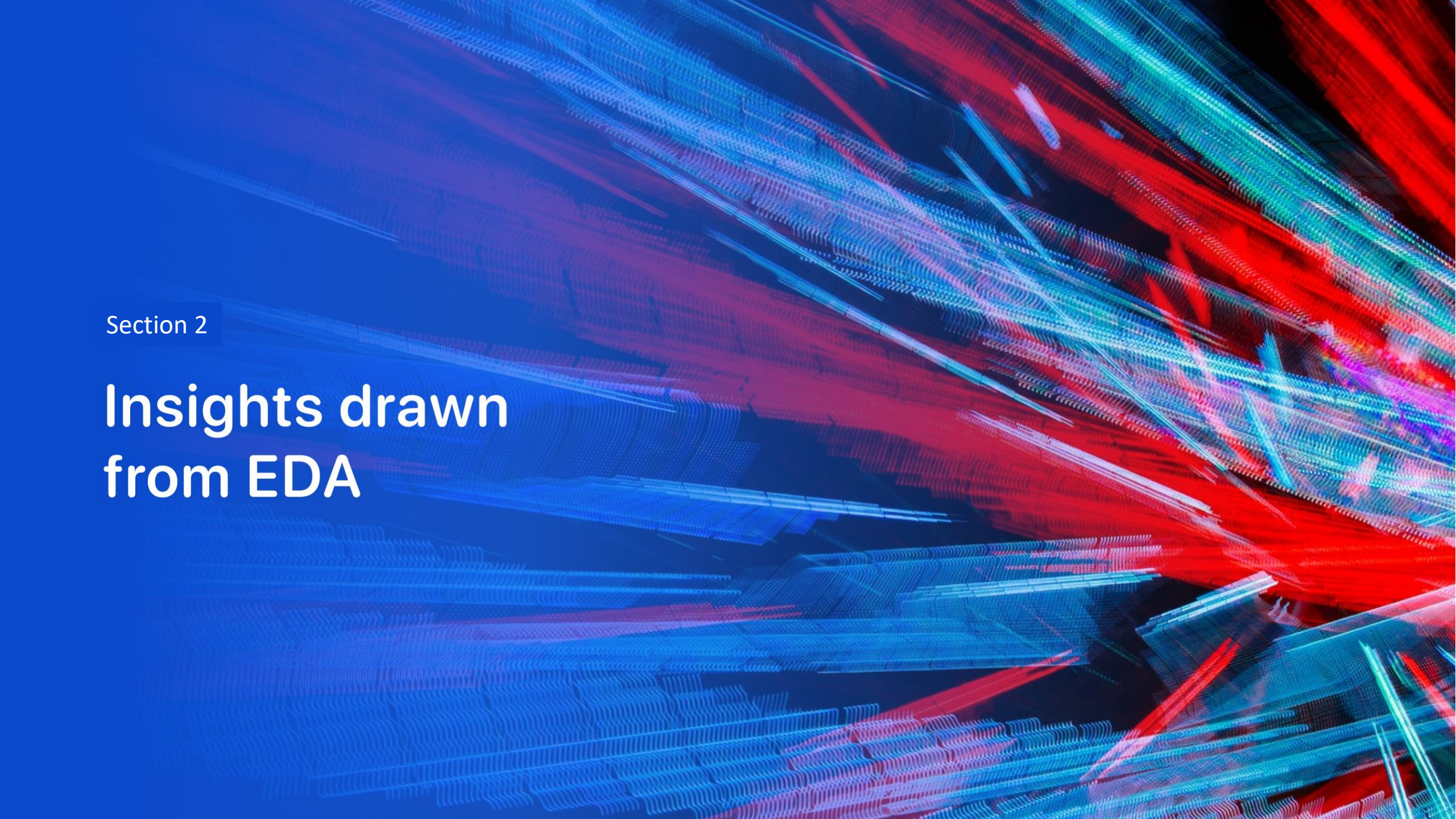


Results

- Exploratory data analysis results
  - KSC LC-39A has highest success rate
  - CCAFS SLC-40 has lowest success rate
- Predictive analysis results
  - Most of launches have payload great than 6k were failed
  - The Booter Version FT has highest success rate

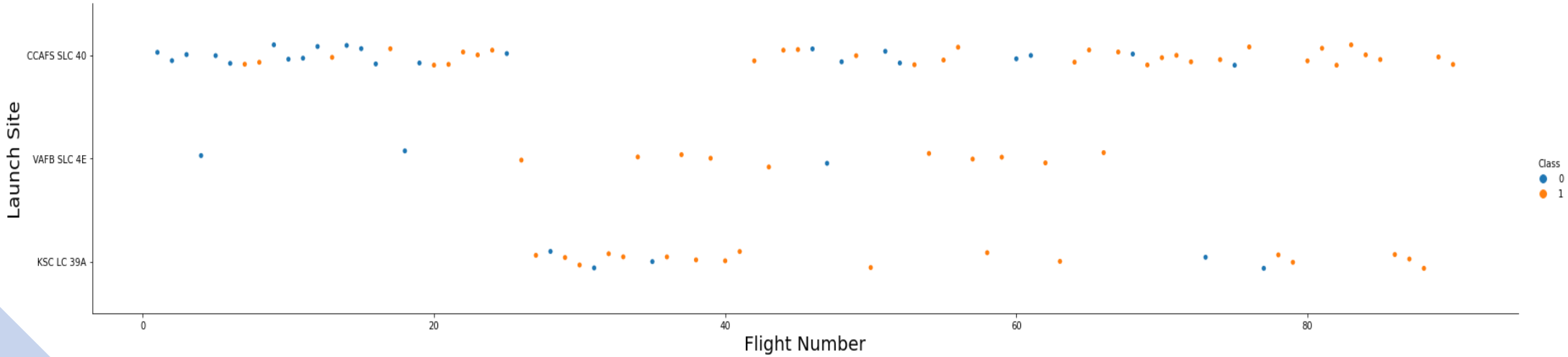





Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

- Most of flight numbers from 1-25, from 41- 80 and great than were distributed for CCAFS LC-40
- There are 13 flight numbers were performed with VAFB SLC 4E
- Most of flight numbers from 25 – 40 were set for KSC LC-39A and there are some flight number from 60- great than 80 also set here

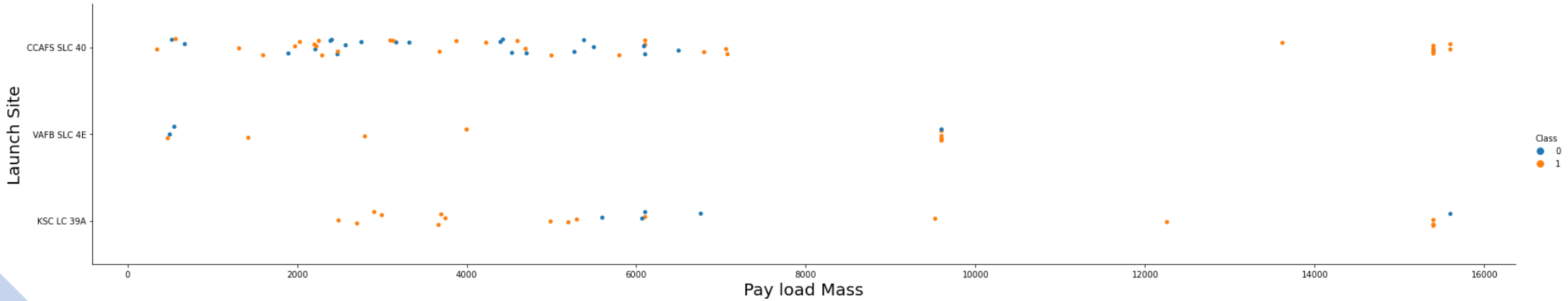


Payload vs. Launch Site

Payloads less than 8000 mainly launched from CCAFS SLC 40, and their failure rate is higher others

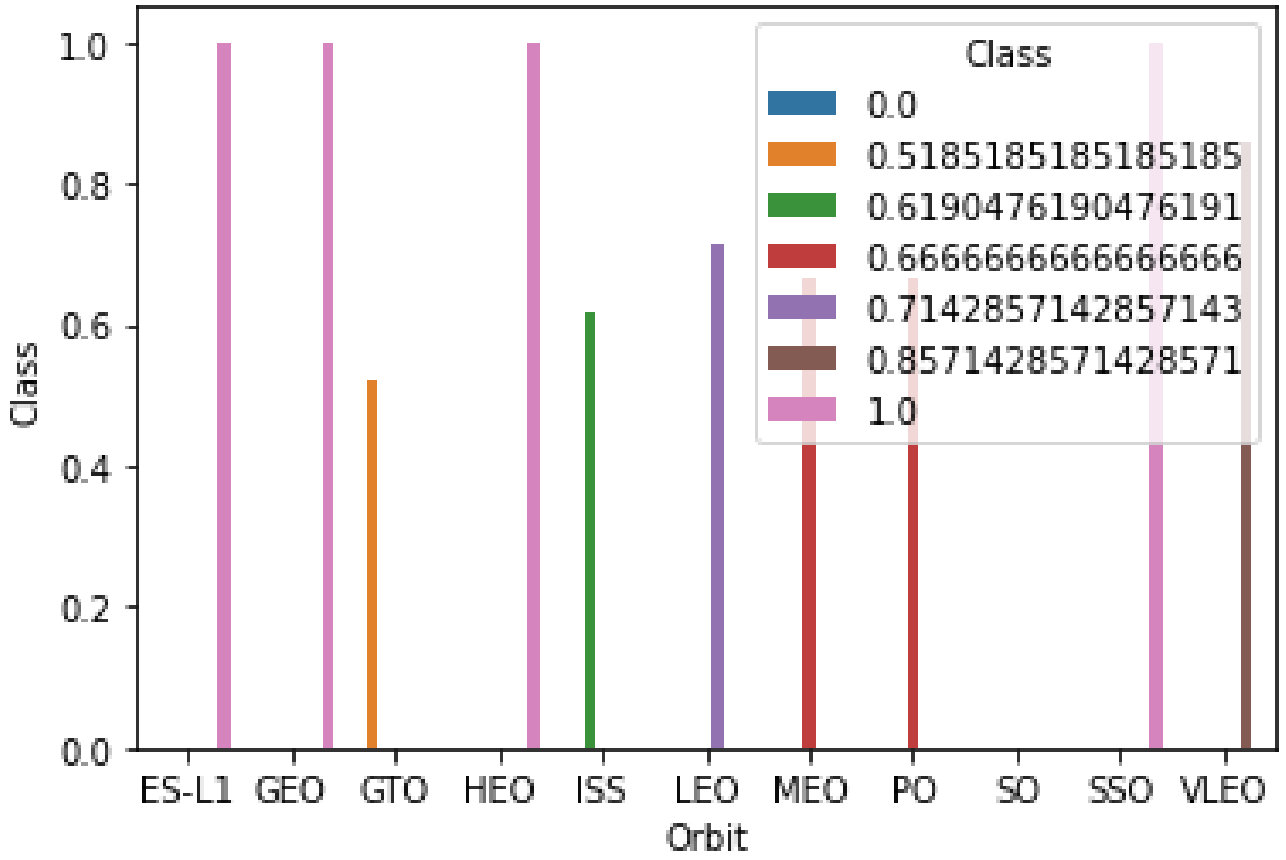
There are several rockets with payloads significant than 1000 launched from KSC LC 39A and CCAFS SLC 40 and almost them were a success

Almost of lunches from VAFB SLC 4E were a success



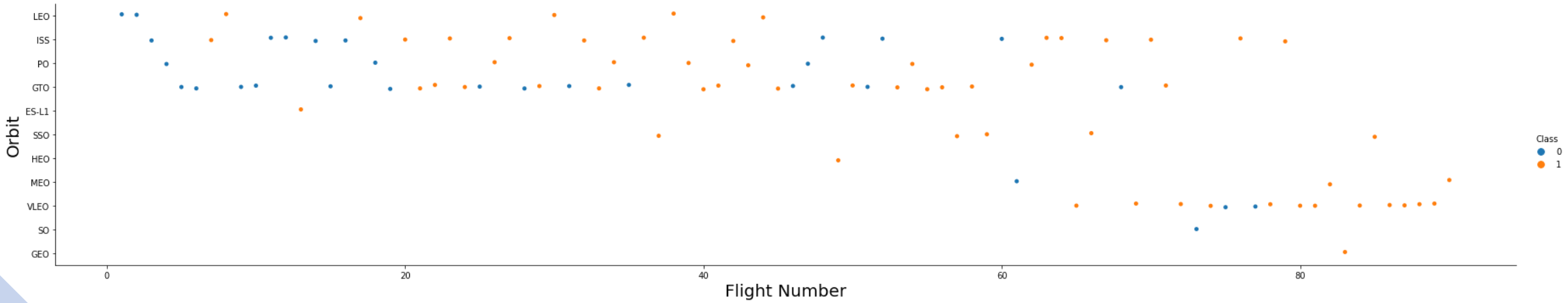
Success Rate vs. Orbit Type

- The orbits have high success rate are ES-L1, GEO, HEO, SSO
- The orbit has lowest success rate is GTO



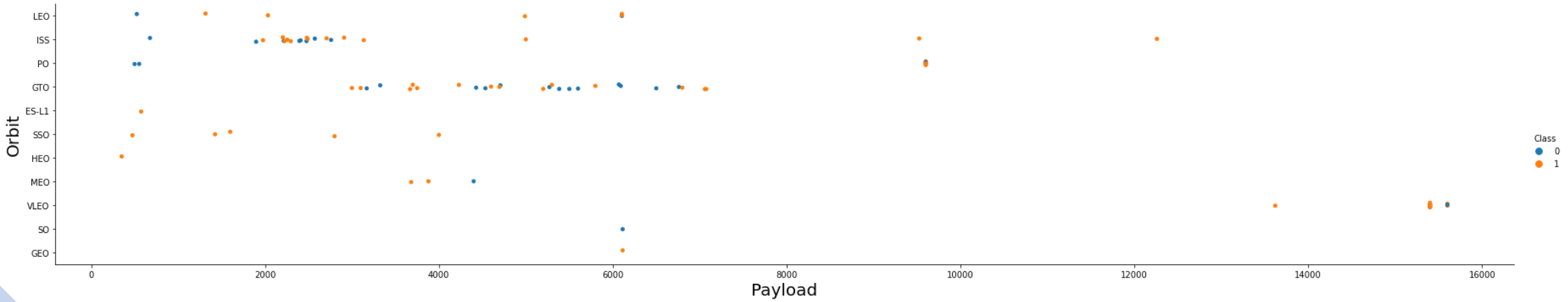
Flight Number vs. Orbit Type

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.



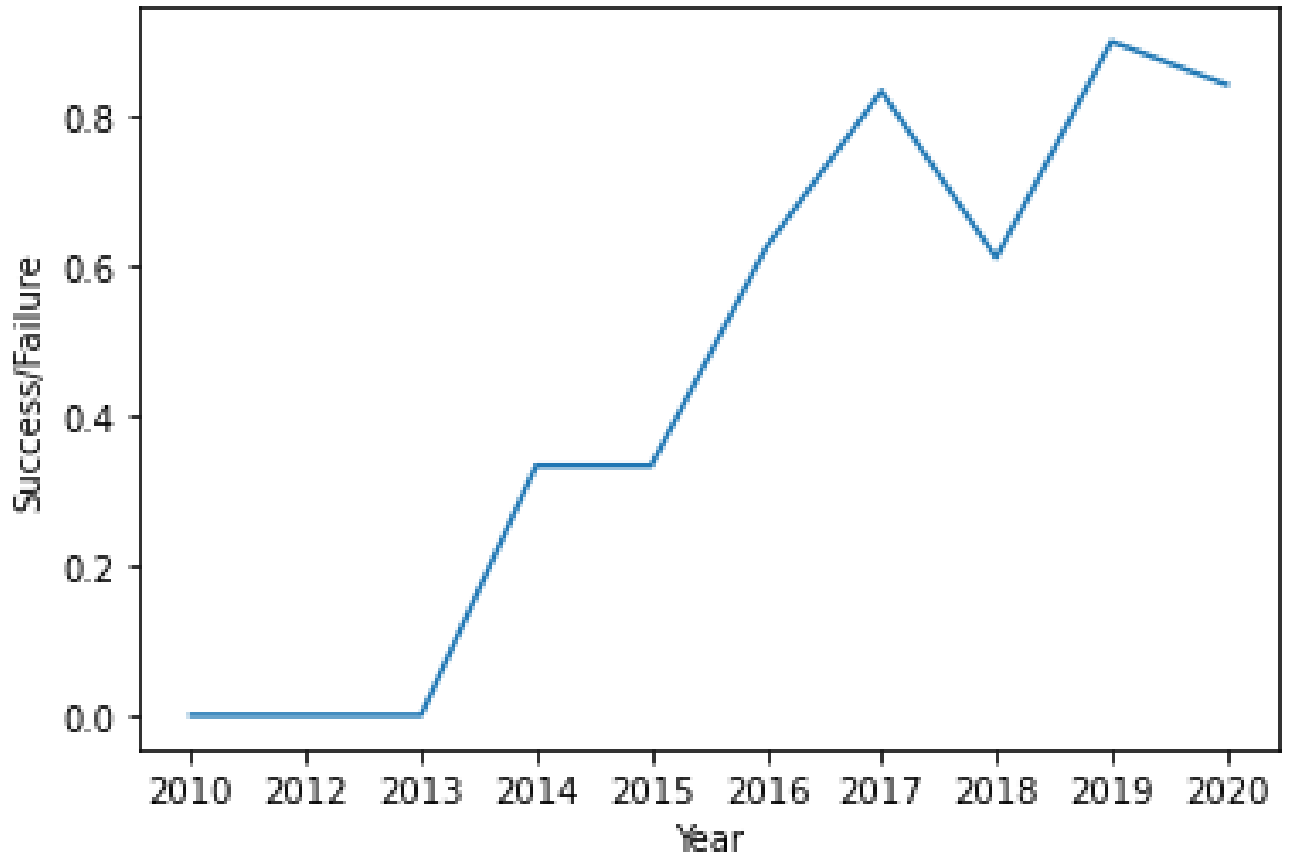
Payload vs. Orbit Type

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

- Almost of launch from 2010-2013 were failed
- C



All Launch Site Names

The unique launch sites: CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E

```
%sql SELECT UNIQUE(launch_site) FROM spacex
```

```
 * ibm_db_sa://kqm41026:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

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

Launch Site Names Begin with 'CCA'

5 records where launch sites begin with `CCA`

```
%sql SELECT * FROM spacex WHERE launch_site LIKE 'CCA%' LIMIT 5
```

```
 * ibm_db_sa://kqm41026:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
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	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

Calculate the total payload carried by boosters from NASA

```
%sql SELECT SUM(payload_mass__kg_) AS payload FROM spacex WHERE customer LIKE 'NASA%'
```

```
 * ibm_db_sa://kqm41026:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

payload
99980

25

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 payload FROM spacex WHERE booster_version LIKE 'F9 v1.1%'
```

```
 * ibm_db_sa://kqm41026:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

payload
2534

26

First Successful Ground Landing Date

Find the dates of the first successful landing outcome on ground pad

```
%sql SELECT MIN(DATE) AS first_date FROM spacex WHERE landing__outcome LIKE '%Success (ground pad)%'
```

```
 * ibm_db_sa://kqm41026:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

first_date
2015-12-22

27

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

```
SELECT booster_version, payload_mass__kg_, landing__outcome FROM spacex WHERE landing__outcome LIKE '%Success (drone ship)%' AND payload_mass__kg_ BETWEEN 4000 AND 6000
```

```
%%sql
SELECT booster_version, payload_mass__kg_, landing__outcome FROM spacex WHERE landing__outcome LIKE '%Success (drone ship)%' AND payload_mass__kg_ BET
```

```
 * ibm_db_sa://kqm41026:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

booster_version	payload_mass__kg_	landing__outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

Calculate the total number of successful and failure mission outcomes

```
%%sql
SELECT mission_outcome, count(*) AS result FROM spacex group by mission_outcome
```

```
 * ibm_db_sa://kqm41026:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

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

Boosters Carried Maximum Payload

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

```
%%sql
SELECT booster_version, payload_mass__kg_ FROM spacex WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM spacex)
```

```
 * ibm_db_sa://kqm41026:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
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

2015 Launch Records

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

```
%%sql
SELECT landing__outcome, booster_version, launch_site FROM spacex WHERE landing__outcome LIKE 'Failure%' AND year(date) = 2015
```

```
 * ibm_db_sa://kqm41026:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
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 landing__outcome FROM spacex WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' ORDER BY DATE DESC;
```

```
 * ibm_db_sa://kqm41026:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

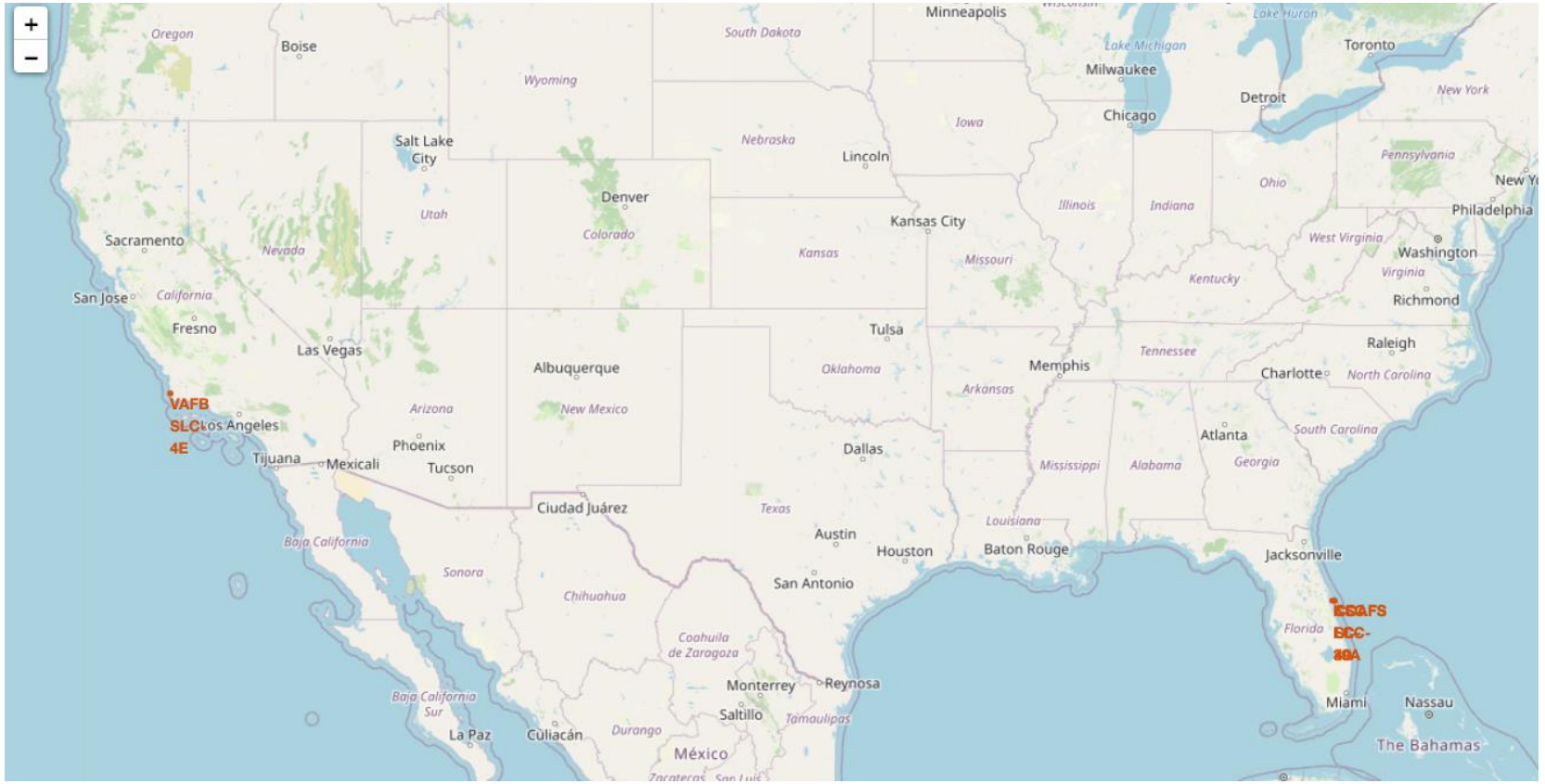
landing__outcome
No attempt
Success (ground pad)
Success (drone ship)
Success (drone ship)
Success (ground pad)

Section 3

Launch Sites Proximities Analysis

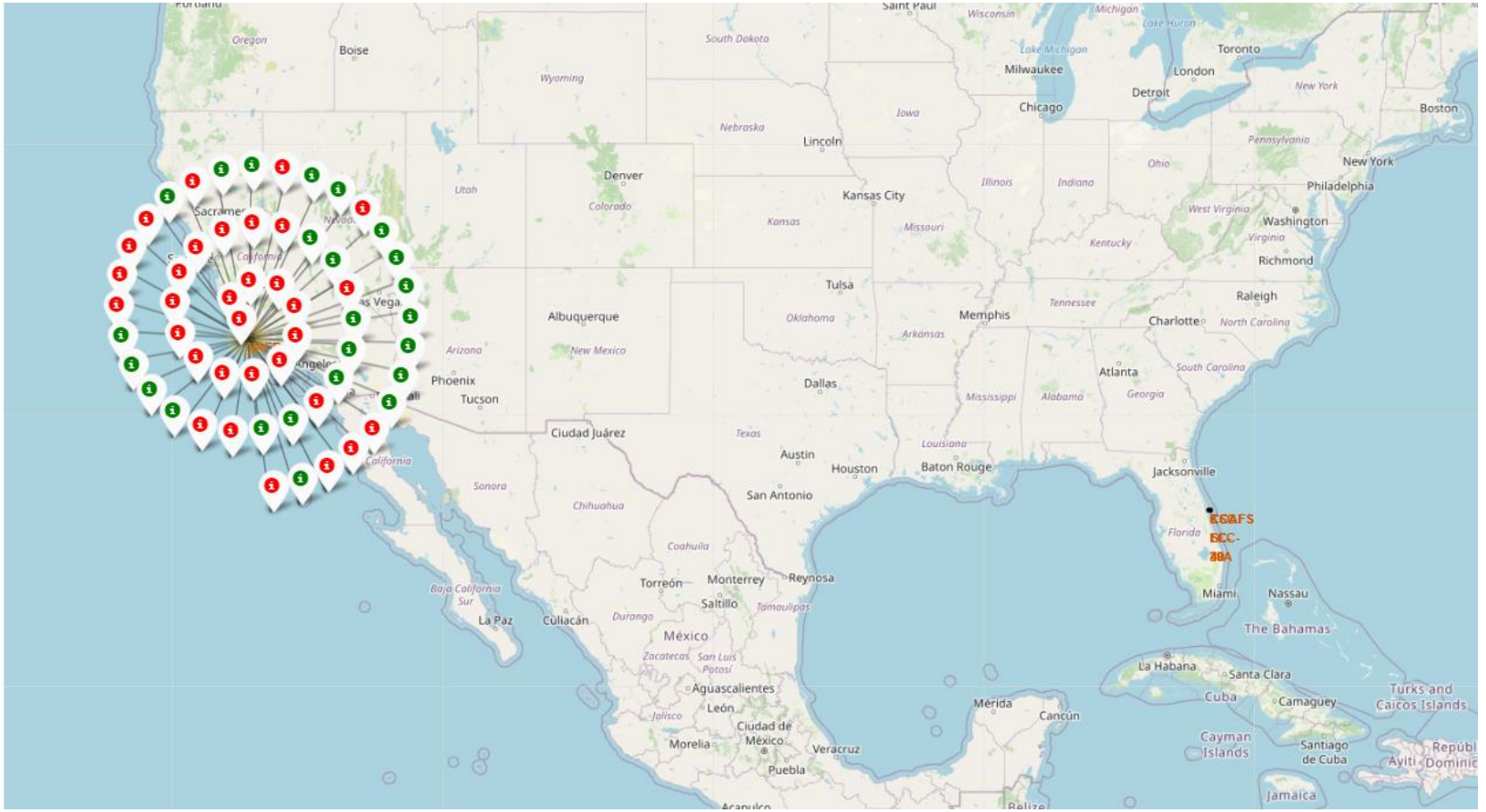
Mark all launch sites on a map

All launch sites in very close proximity to the coast



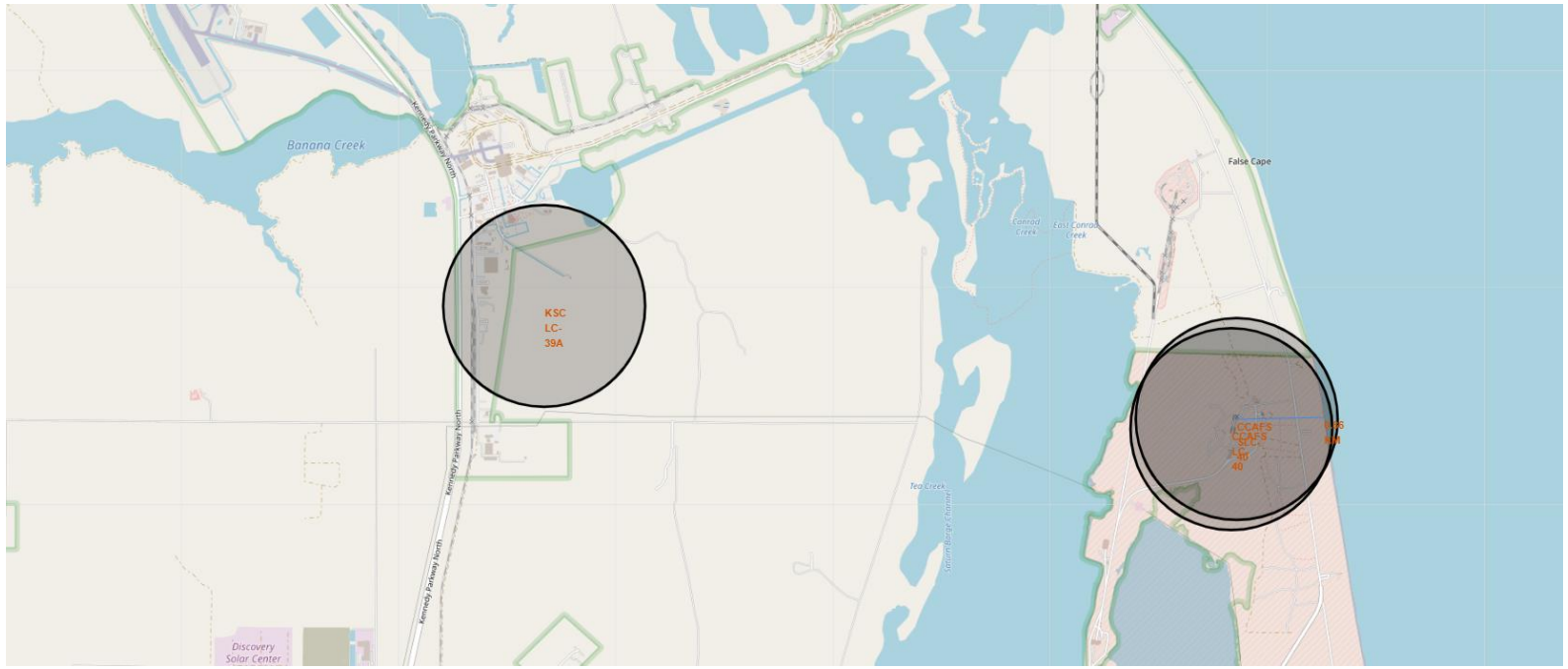
Mark the success/failed launches for each site on the map

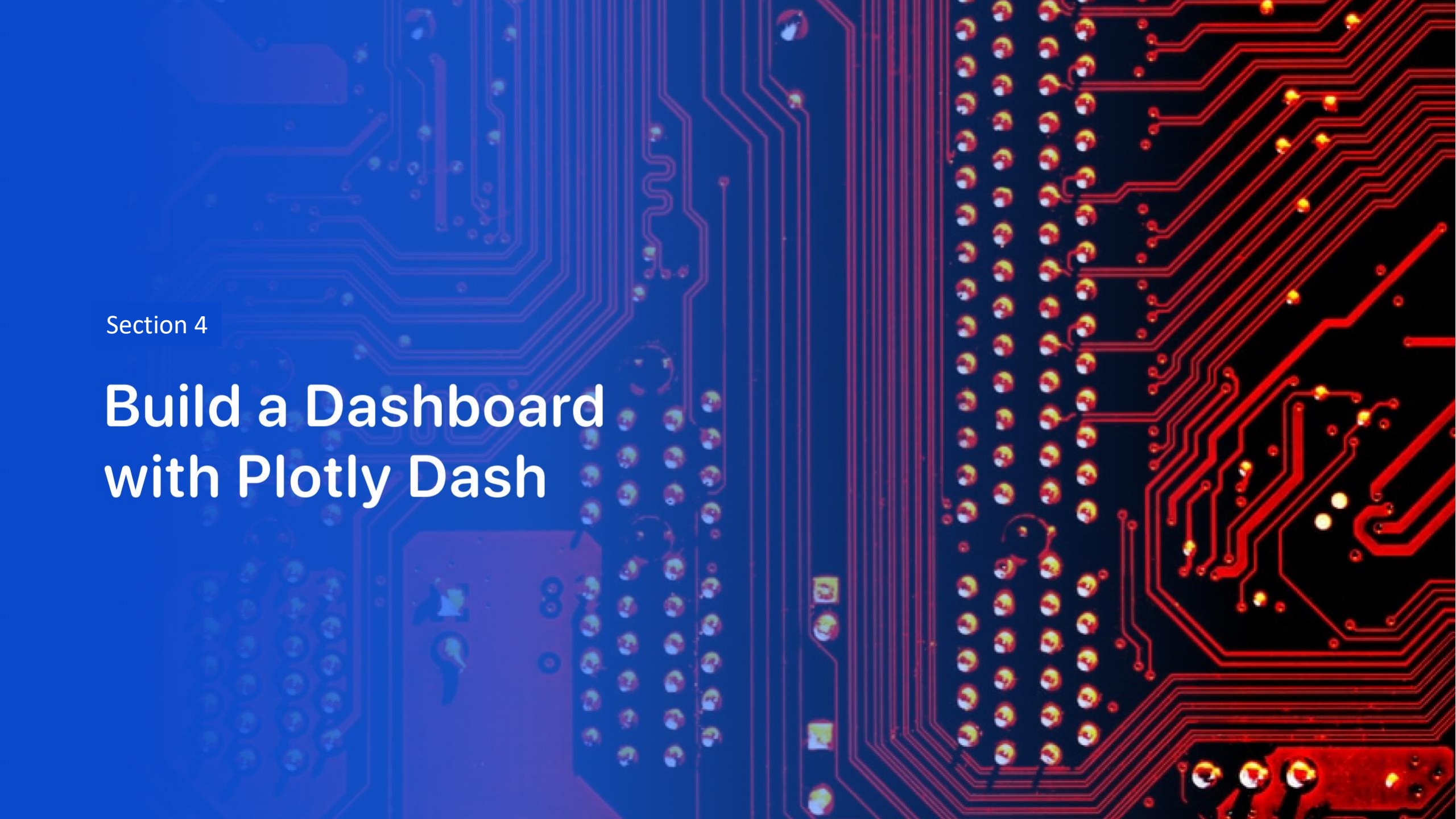
The success rate and fail rate are the same



Calculate the distances between a launch site to its proximities

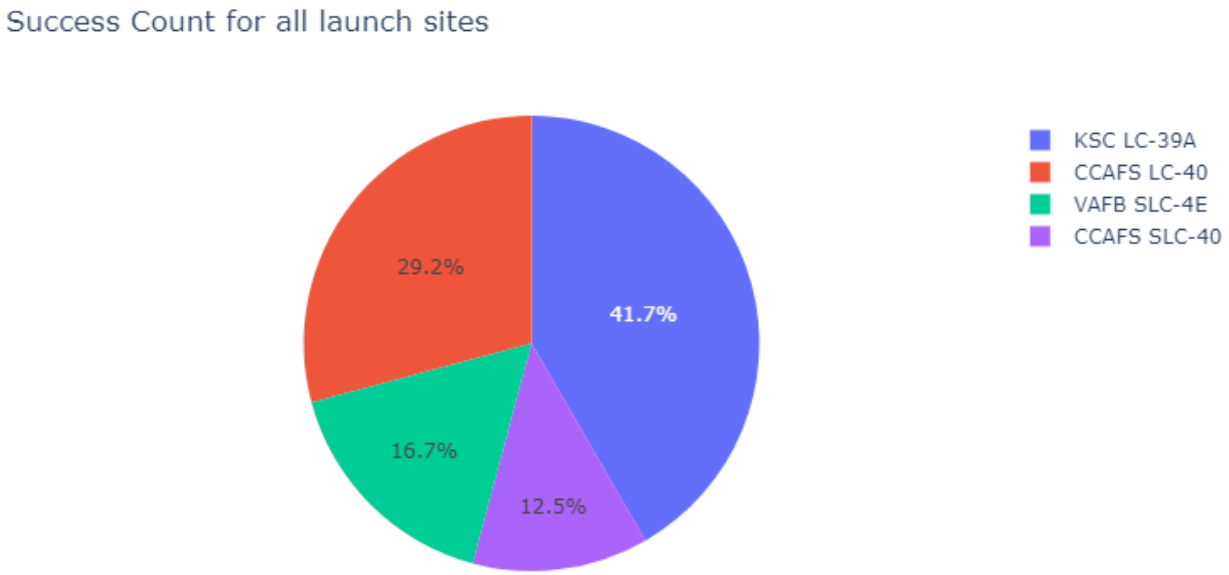
- Launch sites in close proximity to coastline
- Launch sites in close proximity to railways
- Launch sites keep certain distance away from cities




Section 4
Build a Dashboard with Plotly Dash

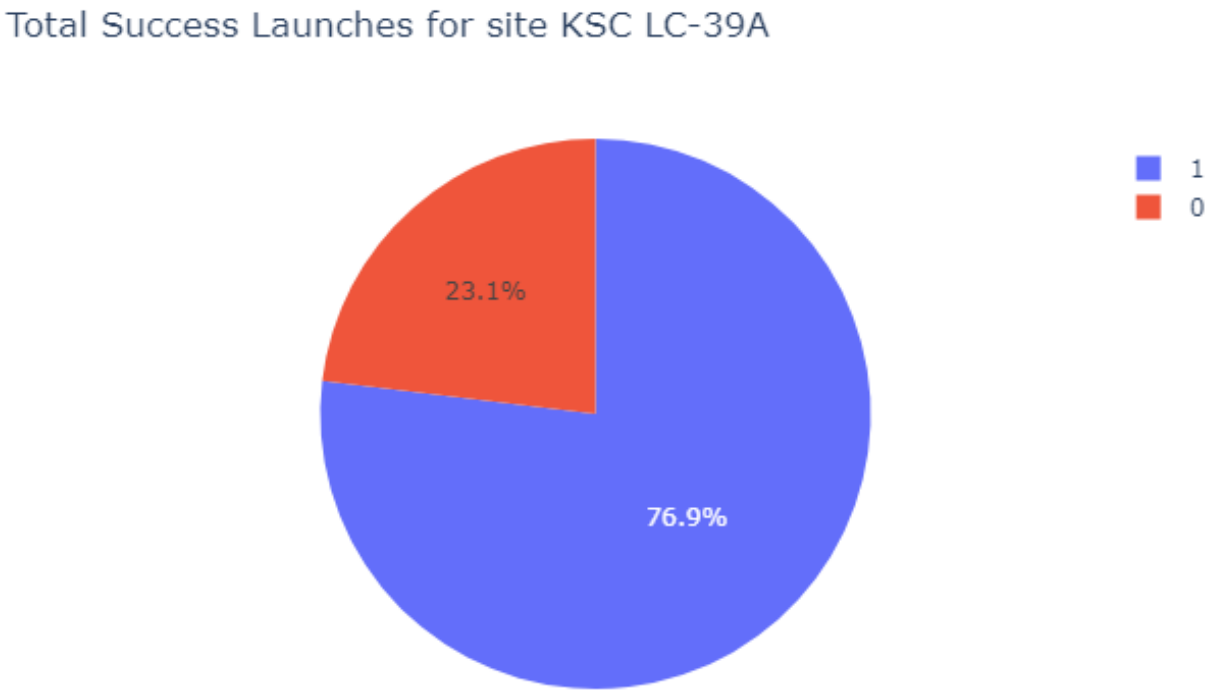
Total Success Launch Site

- Launch site



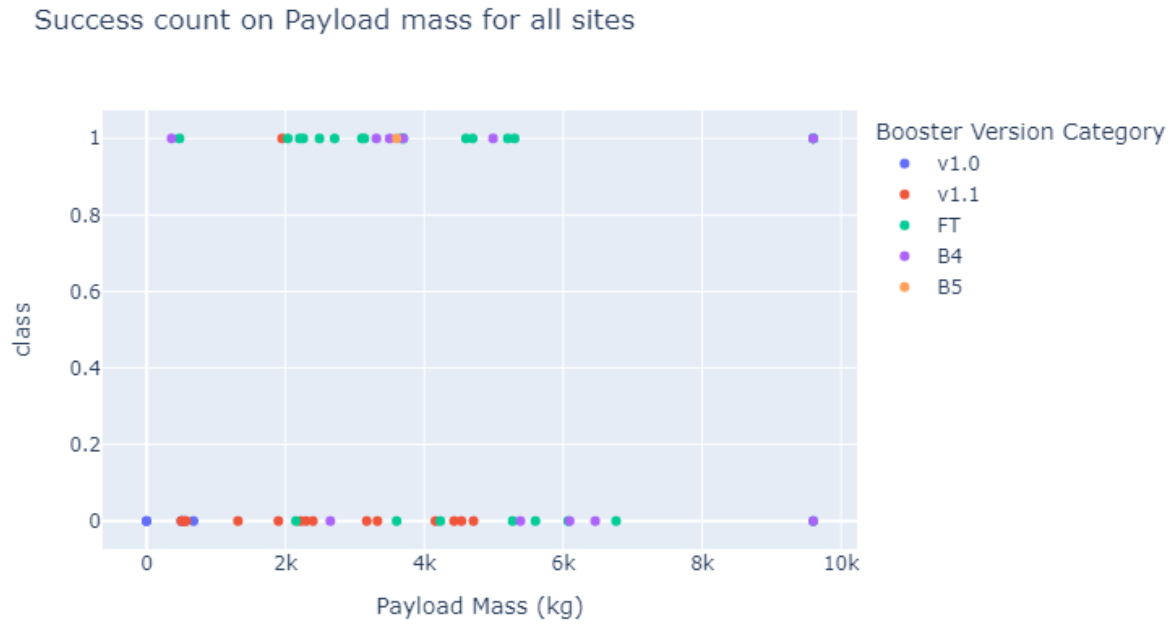
Highest Launch Success

- Explain the important elements and findings on the screenshot



Payload Vs. Launch Outcome For All Sites

- The success rate for low payload is higher than for



Section 5

Predictive Analysis (Classification)

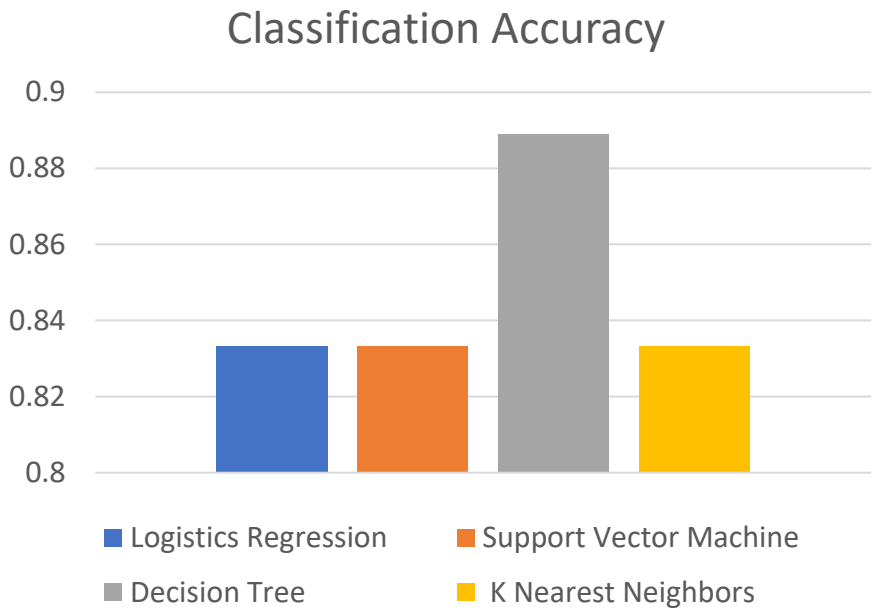
Classification Accuracy

- The Decision Tre has the highest classification accuracy

Find the method performs best:

```
print('Logistics Regression :', logreg_cv.score(X_test, Y_test))
print('Support Vector Machine:', svm_cv.score(X_test, Y_test))
print('Decision Tree:', tree_cv.score(X_test, Y_test))
print('K nearsdt neighbors:', knn_cv.score(X_test, Y_test))
```

```
Logistics Regression : 0.8333333333333334
Support Vector Machine: 0.8333333333333334
Decision Tree: 0.8888888888888888
K nearsdt neighbors: 0.8333333333333334
```



Confusion Matrix

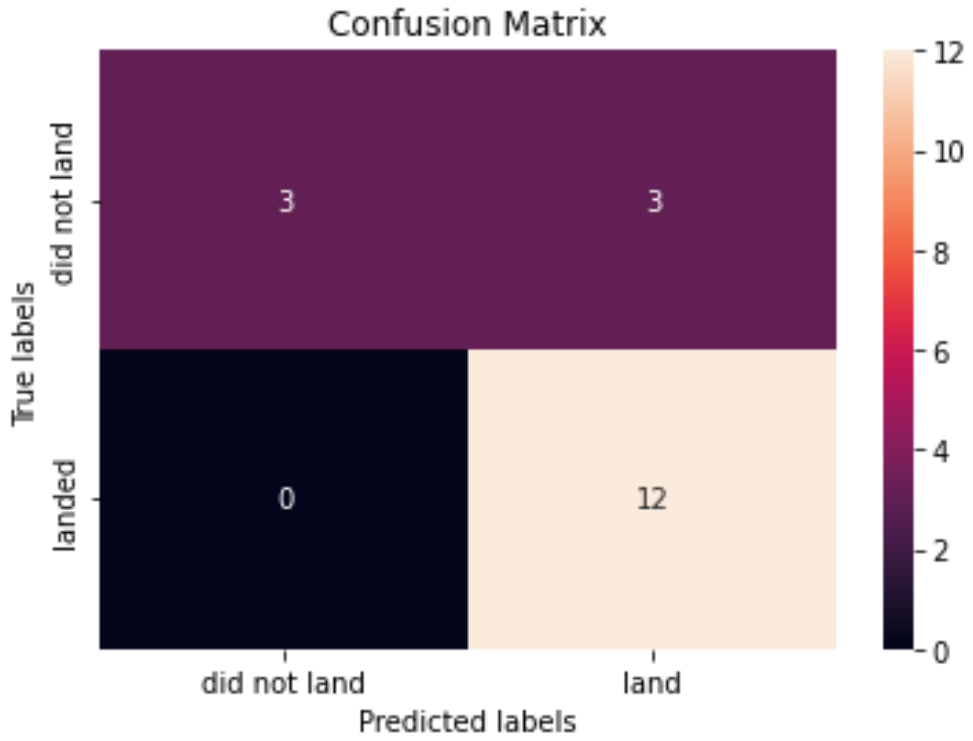
- The actual result for the success landed mostly matched with expectations predicted

```
tree_cv.score(X_test,Y_test)
```

```
0.8888888888888888
```

We can plot the confusion matrix

```
yhat = svm_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



Conclusions

- The orbits have high success rate are ES-L1, GEO, HEO, SSO
- All launch sites in very close proximity to the coast
- KSC LC-39A had the most successful launches of any sites.
- The Decision Tre has the highest classification accuracy

Thank you!