IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Roshan Sara Cherian
14-01-2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Project objective was to predict whether the first stage of Falcon 9 rocket will land successfully based on analysis of data of past launches.
- Methodology and Results:
  - Data collection through Spacex API and Webscraping of Wikipedia page on Falcon 9 launches
  - Data Wrangling done to deal with missing values and preparing data for analysis
  - Parameters impacting landing outcome identified through exploratory data analysis – launch site, payload mass, flight number, orbit type, booster version, type of landing pad etc. identified
  - Use of Folium Map and interactive dashboard further confirmed that landing outcome is impacted by choice of launch site and payload mass
  - 4 different machine learning classification models built and trained on prepared data
  - Threee models accurately predicted all unsuccessful landings and predicted successful landings with 80% accuracy

Introduction

Background

- SpaceX advertises Falcon 9 rocket launches with a cost of 62 million dollars.
- Other rocket launch providers cost upward of 165 million dollars each.
- SpaceX can quote so low because SpaceX can reuse the first stage.
- A competitor wanting to bid against SpaceX can determine the cost of a launch, if it can be determined if the first stage will land

Objective

- To predict if the Falcon 9 first stage will land successfully

Section 1

Methodology

Methodology

Executive Summary

- Data collection from two sources – from Spacex API and from the Wikipedia page on Falcon 9 launches
- Data wrangling carried out to deal with missing values, ensure data format consistency and to retain only such data as is relevant for analysis
- Exploratory data analysis (EDA) carried out using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models such as Logistic Regression, Support Vector Machine, Decision Tree and KNN -
    - Classification models were evaluated on accuracy score and confusion matrix.
    - Model with the highest accuracy score used to draw conclusions

Data Collection

From Spacex API

- All past rocket launches data collected from the SpaceX API
- Data received in JSON format converted into DataFrame
- Data of Only Falcon 9 launches, and only the corresponding relevant columns - 'rocket', 'payloads', 'launchpad', 'cores', 'flight_number', 'date_utc' filtered out.
- SpaceX APIs used to get additional data values such as booster name, launchsite name and location, payload mass etc. using data available in ID form.

From Wikipedia page on Falcon 9 launches

- Wikipedia HTML page scraped to get data on Falcon 9 rocket launches
- Webscraping output (i.e. the HTML response) stored as a `BeautifulSoup` object
- HTML tables parsed to extract,
  - Column names from HTML Table header rows
  - Data from other rows (other than header rows)
- The extracted data is then converted into a DataFrame

Data Collection – SpaceX API

- Data on SpaceX launches available from SpaceX API.
- Use requests.get() to obtain all past launches data from the SpaceX API in JSON format
- Data in JSON format converted to DataFrame using json_normalize()
- Helper functions used to extract information using ID numbers in the launch data using SpaceX API.
  - From the rocket column - booster name
  - From launchpad – name of launchsite, latitude and longitude
  - From payload – mass of payload and orbit its going to
  - From cores - outcome of the landing, the type of the landing, number of flights with that core, landing pad used, etc.
- Culled out data converted into a DataFrame

```
spacex_url=https://api.spacexdata.com/v4/launches/past

response = requests.get(spacex_url)

data = pd.json_normalize(response.json())
```

requests.get() used again to obtain more specific information such as name of rocket, name of launch-site etc. E.g:

requests.get (“https://api.spacexdata.com/v4/rockets/” + str(x)).json(), where x is the name of rocket

Github Link: https://github.com/rscherian/rscherian/blob/main/Module%201_Lab1_spacex-data-collection-api.ipynb

Data Collection - Scraping

- Data of all Falcon 9 Launches is at "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"
- HTTP GET method used to request the Falcon9 Launch HTML page.
- Create a `BeautifulSoup` object from the HTML `response`
- Find all tables on the wiki page
- Collect all relevant column names from the HTML table header of the table which has the launch records
- Iterate through the <th> elements and extract column name one by one
- Create a dictionary with column names as keys and fill in launch record values by parsing the launch HTML table
- Create a DataFrame from the dictionary with launch data

```
response = requests.get(static_url)

r = response.text

soup = BeautifulSoup(r)

html_tables = soup.find_all("table")

headers = first_launch_table.find_all("th")

for x in headers:

  if(extract_column_from_header(x) != None and len(extract_column_from_header(x)) > 0):


column_names.append(extract_column_from_header(x))

df= pd.DataFrame({ key:pd.Series(value) for key, value in launch_dict.items() })
```

where launch_dict is a dictionary with launch records extracted from table rows

GitHub Link: https://github.com/rscherian/rscherian/blob/main/Module1_Lab1-webscraping.ipynb

Data Wrangling

- DataFrame had 5 missing values for “PayloadMass” and 27 missing values for “Landingpad”.
- Missing values for PayloadMass replaced with mean Payload Mass.
- **value_counts()** method used to determine the following:
  - No. of launches on each launch site – maximum for CCAFS SLC 40 (55 launches)
  - No. of launches on each orbit type – maximum for GTO (27 launches)
  - No. and occurence of mission outcome for each orbit type – maximum 13 success ASDS for GTO orbit
- Landing Outcome Label column “class” was created from “Outcome” column.
- Success rate was determined using **.mean()** method on the “class” column

- GitHub URL:

https://github.com/rscherian/rscherian/blob/main/Module%201_Lab1_spacex-data-collection-api.ipynb

https://github.com/rscherian/rscherian/blob/main/Module1_Lab2-spacex-data_wrangling.ipynb

EDA with Data Visualization

- One of the objectives of the EDA was to identify those variables which are relevant for predicting whether a launch will be successful.
- 4 types of charts were used for EDA
  - ❑ Scatter Plots and Categorical Plots: to analyse the relationship between two variables
  - ❑ Bar Charts: To analyse categorical data
  - ❑ Line Charts: to analyse trend of data points over a period
- The following graphs were plotted:
  - ❑ Categorical plot of FlightNumber vs. PayloadMass, overlaid with the outcome of the launch.
  - ❑ Categorical plot and Scatterplot of FlightNumber vs LaunchSite, overlaid with the outcome of the launch.
  - ❑ Scatter point chart of Payload Vs. Launch Site
  - ❑ Categorical plot of Launch Site Vs PayloadMass, overlaid with the outcome of the launch
  - ❑ Bar Chart of success rate of each orbit – to visualize the relationship between success rate and orbit type
  - ❑ Categorical plot and Scatterplot of FlightNumber Vs Orbit type.
  - ❑ Categorical plot and Scatterplot of PayloadMass Vs Orbit type
  - ❑ Line chart of success rates in each year to visualize the trend

- GitHub URL: https://github.com/rscherian/rscherian/blob/main/Module2_lab2-eda-dataviz.ipynb

EDA with SQL

The following information was extracted using SQL Queries:

1. Names of the unique launch sites in the space mission
2. Five launch records belonging to launch sites whose name begin with the string 'KSC’
3. Total payload mass carried by boosters launched by NASA (CRS)
4. Average payload mass carried by booster version F9 v1.1
5. Date of successful landing outcome in drone ship
6. List of boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
7. Count of successful mission outcomes and failed mission outcomes
8. List of the booster_versions which have carried the maximum payload mass.
9. List of launches with succesful landing_outcomes in ground pad in year 2017 along with the month of launch, booster version, and launch_site
10. Count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, ranked in descending order

GitHub URL: https://github.com/rscherian/rscherian/blob/main/Module2_Lab1_sql-edx_sqllite.ipynb

Build an Interactive Map with Folium

- Launch success rate may depend on many factors. One such may be the location and proximities of a launch site, i.e., the initial position of rocket trajectories.

- Finding an optimal location for building a launch site involves many factors. Analyzing the existing launch site locations may assist in identifying those factors. Accordingly,

  - ❑ All the launch sites were marked on a Folium Map using their latitude and longitude data
  - ❑ All the launches, both successful and failed, were also marked onto the same map
  - ❑ A highlighted circle area with a text label was added to each launch site for easy identification
  - ❑ Markers were used to identify each launch outcome (Green, if successful & Red, if failure)
  - ❑ Since each launch site had multiple launches, and therefore multiple markers, marker clusters were used to simplify visualization of multiple markers having the same coordinate.
  - ❑ MousePosition was added on the map to get coordinate for a mouse over a point on the map
  - ❑ Using PolyLine, a line was drawn from one launch site to the closest city, coastline, highway and railway line.
  - ❑ A marker with distance from the launch site was placed at each identified proximate point

- GitHub URL: https://github.com/rscherian/rscherian/blob/main/Module3_lab1_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- The Dashboard has
  - ❑a dropdown list to enable Launch Site selection, with the default select value ‘ALL sites’
  - ❑a slider is to select payload range
- Based on the selection in the dropdown list and the range selected on the slider,
  - ❑A pie chart is generated which either shows the total successful launches count for all sites or, if a specific launch site is selected, will show the Success vs. Failed counts for that site
  - ❑A scatter chart is generated which shows the correlation between payload and launch success
- GitHub URL: https://github.com/rscherian/rscherian/blob/main/RSC_spacex_dash_app.py

Predictive Analysis (Classification)

- The objective is to create a machine learning pipeline to predict if the first stage will land given the data of previous launches
- Data of past launches, after data wrangling, is standardized using preprocessing.StandardScaler()
- The standardized data is then split into training data and test data using train_test_split()
- 4 Classification Models were chosen viz. Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Decision Trees and Logistic Regression
- The best Hyperparameter for SVM, Decision Trees, KNN and Logistic Regression were determined using GridSearchCV()
- Accuracy of each classification model on the test data is calculated using the method score
- The method which performs best using test data is preferable for predicting if the first stage will land in proposed launches.

GitHub URL:
https://github.com/rscherian/rscherian/blob/main/Module4_SpaceX_Machine_Learning_Prediction_Part_5.ipynb

Results

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

Section 2

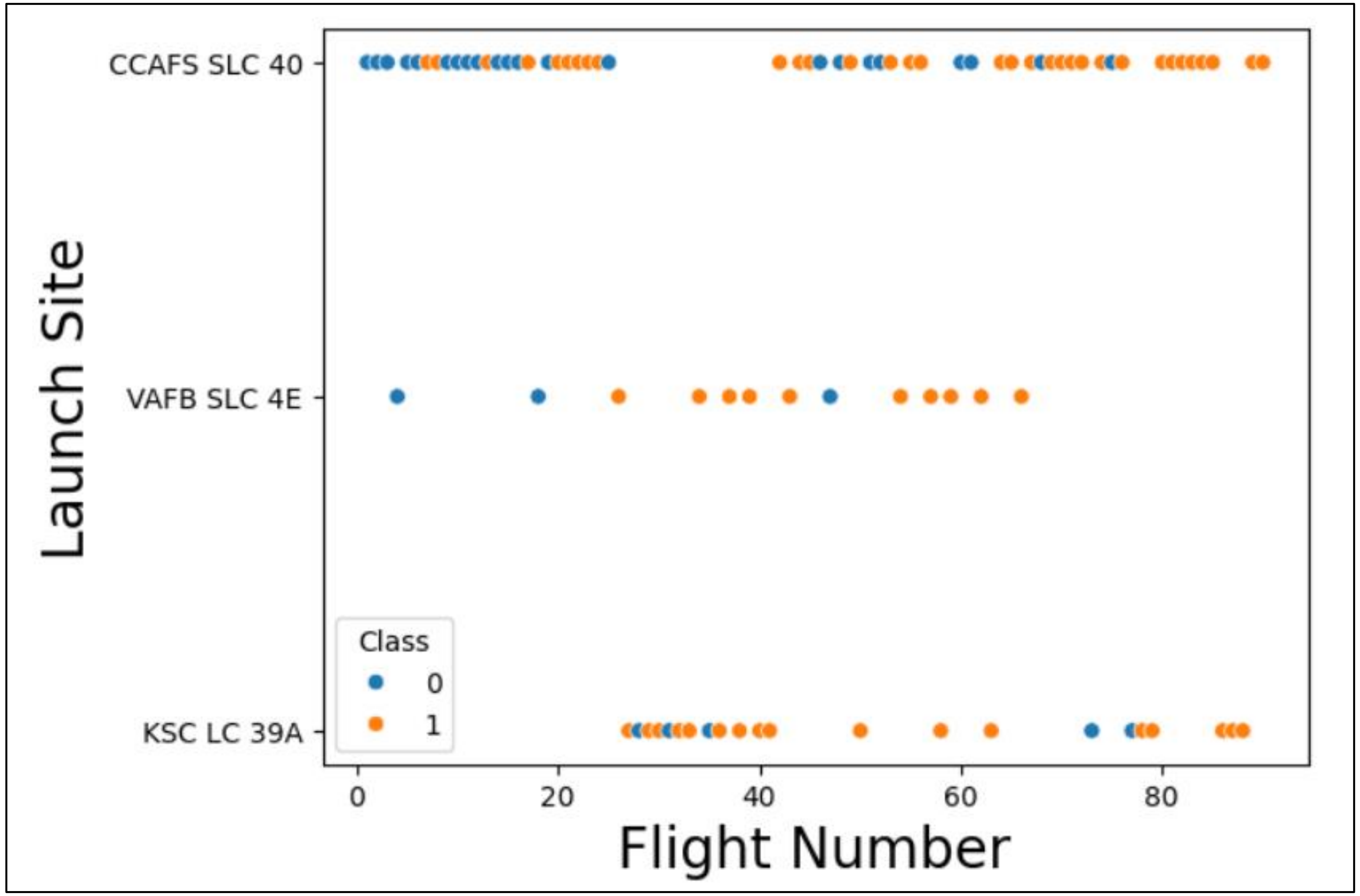
Insights drawn from EDA

Flight Number vs. Launch Site

In case of CCAFS SLC 40 and VAFB SLC 4E sites, the initial flight numbers were unsuccessful, while the latest flight numbers are successful.

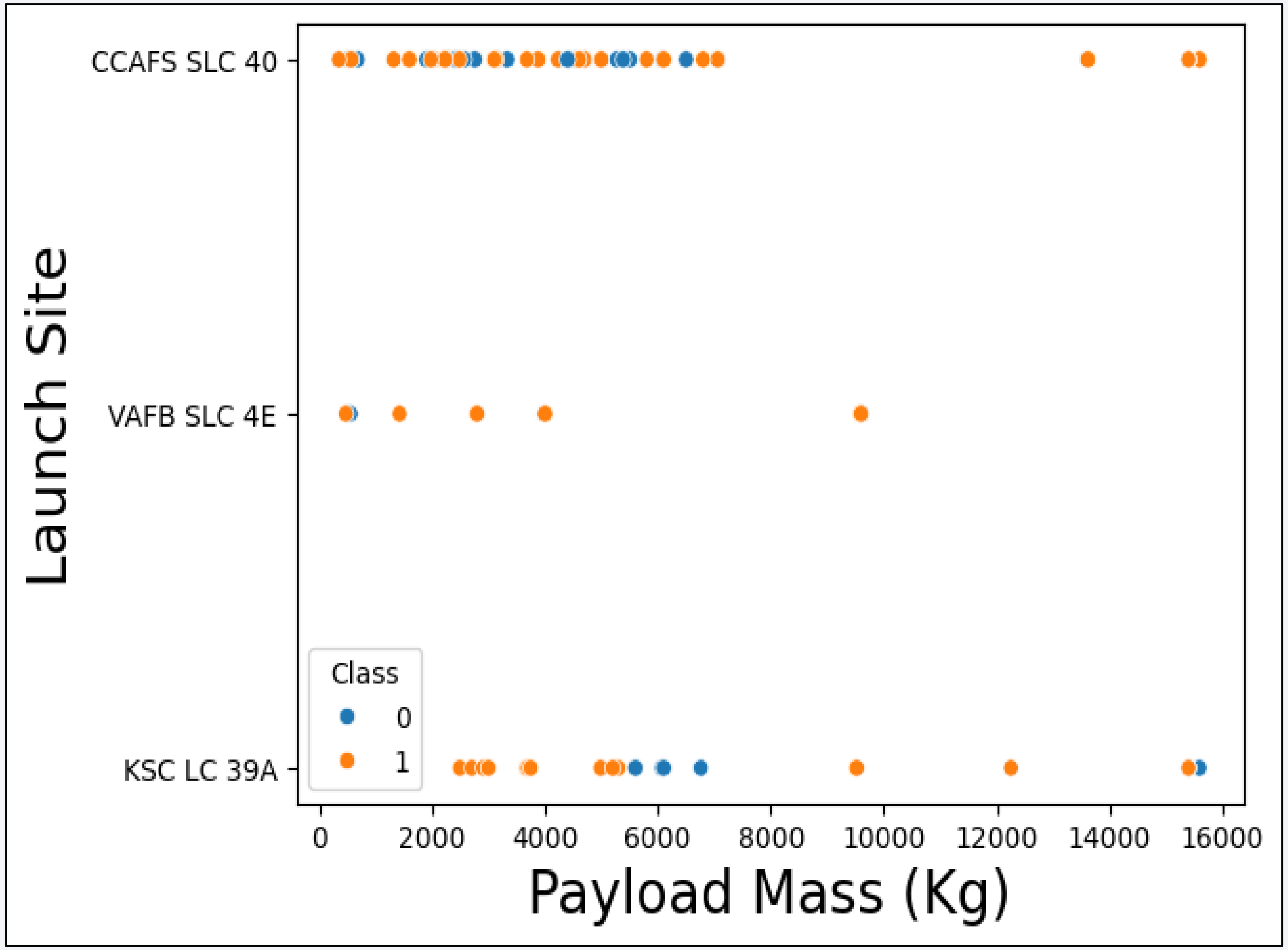
However, in case of KSC LC 39A, where launches started quite late, there were successful launches even in the beginning

Scatter plot of Flight Number vs. Launch Site



Payload vs. Launch Site

Scatter plot of Payload vs. Launch Site



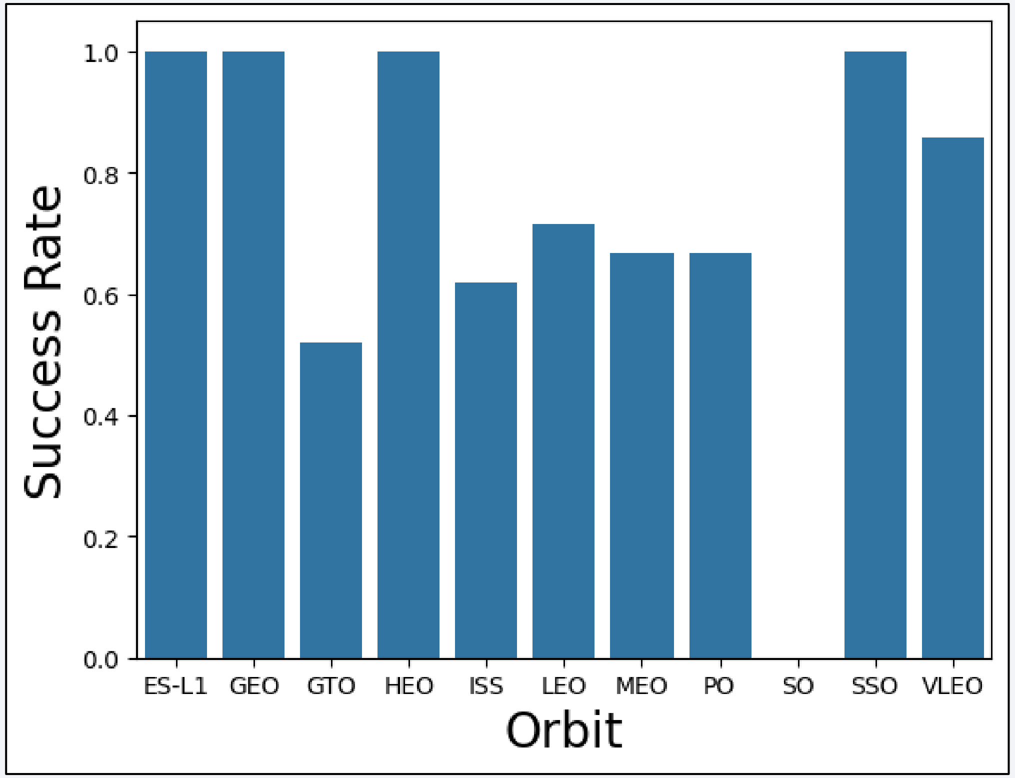
For the VAFB-SLC launch site, there are no rockets launched for heavy payload mass (greater than 10000)

For the KSC LC 39A launch site, light payload mass launches have had successful landing outcome

For the CCAFS SLC 40 launch site, there have been both successes and failures for light payload mass (less than 7000 kg), but heavy payload mass launches have had successful landing outcome

Success Rate vs. Orbit Type

Bar chart of the success rate of each orbit type



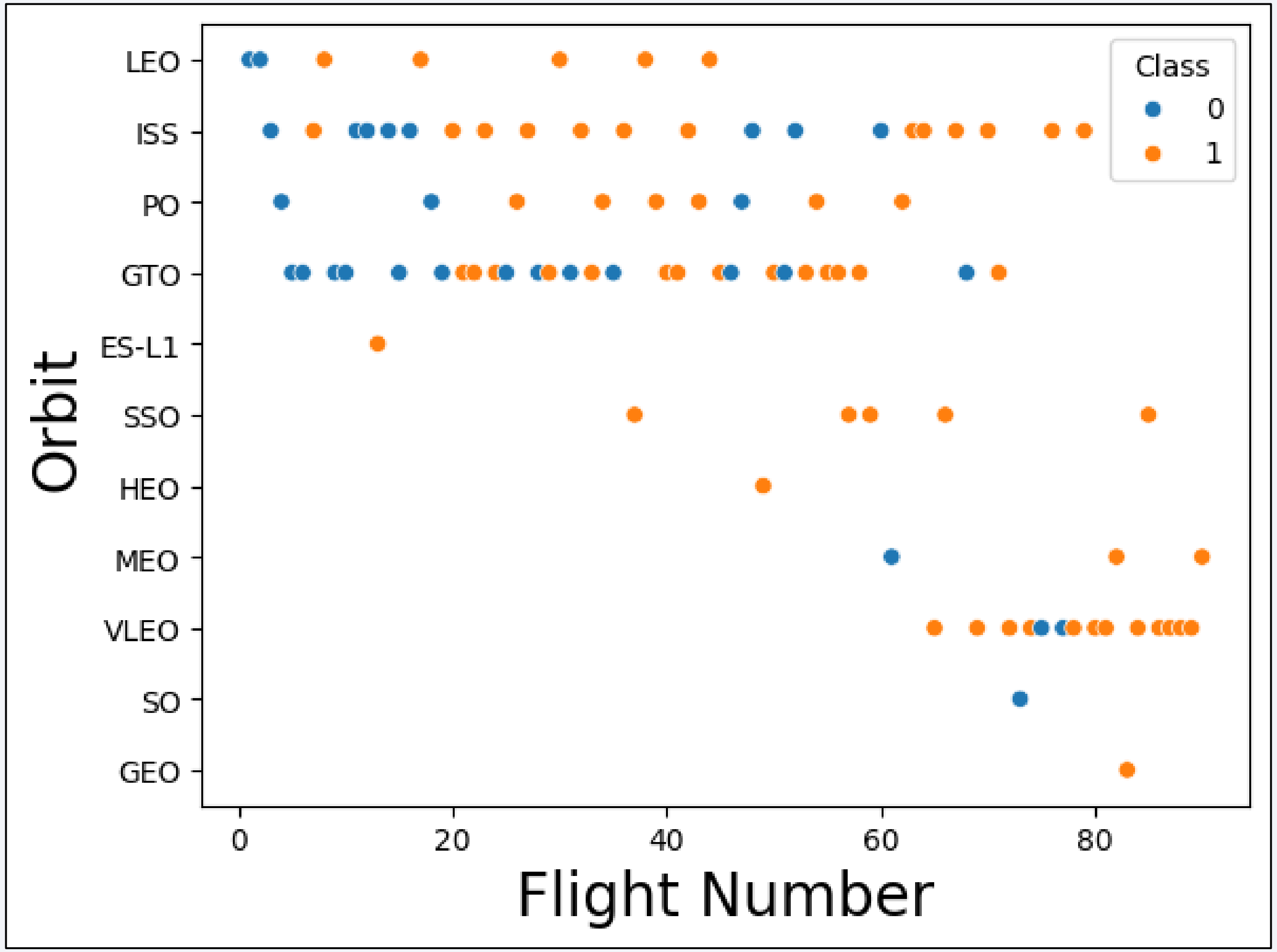
ES-L1, GEO, HEO, SSO orbits have high success rate

GTO orbit type had the lowest success rate of 55%

There were no launches of Falcon 9 for SO orbit type

Flight Number vs. Orbit Type

Scatter plot of Flight number vs. Orbit type

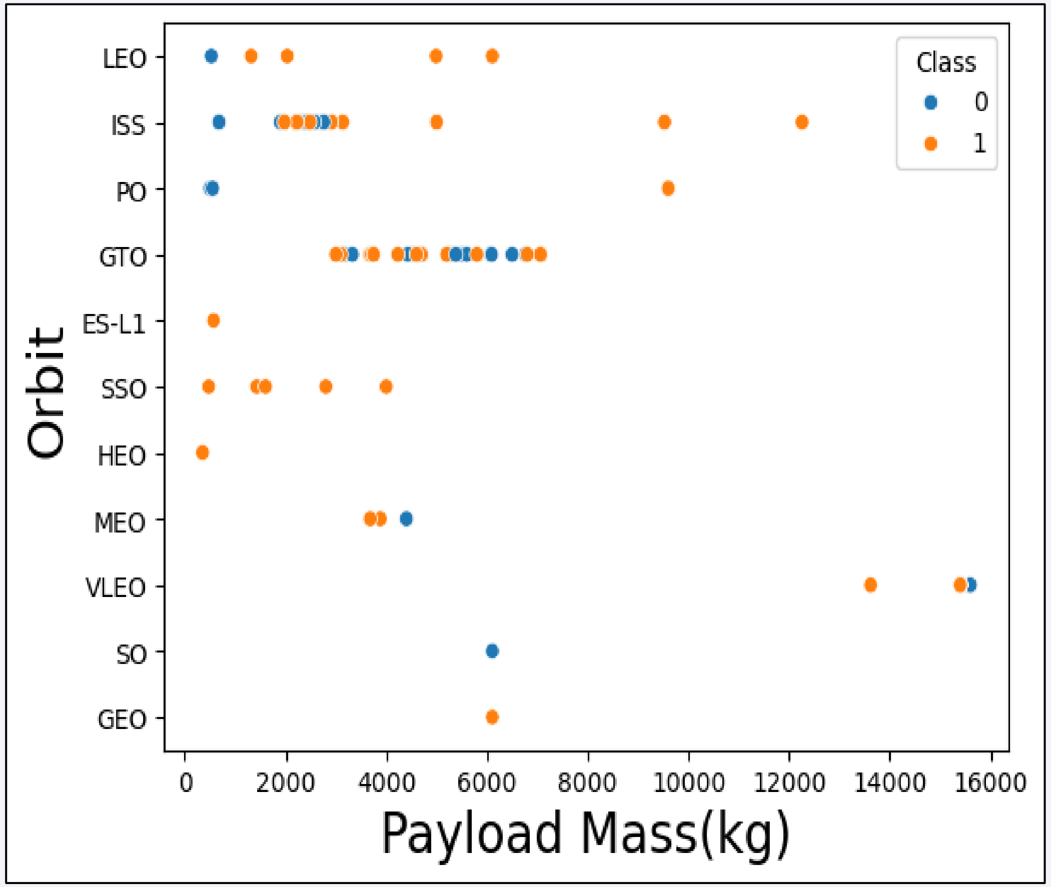


In the LEO orbit - Success appears related to the number of flights

In the GTO orbit - there seems to be no relationship between flight number and orbit

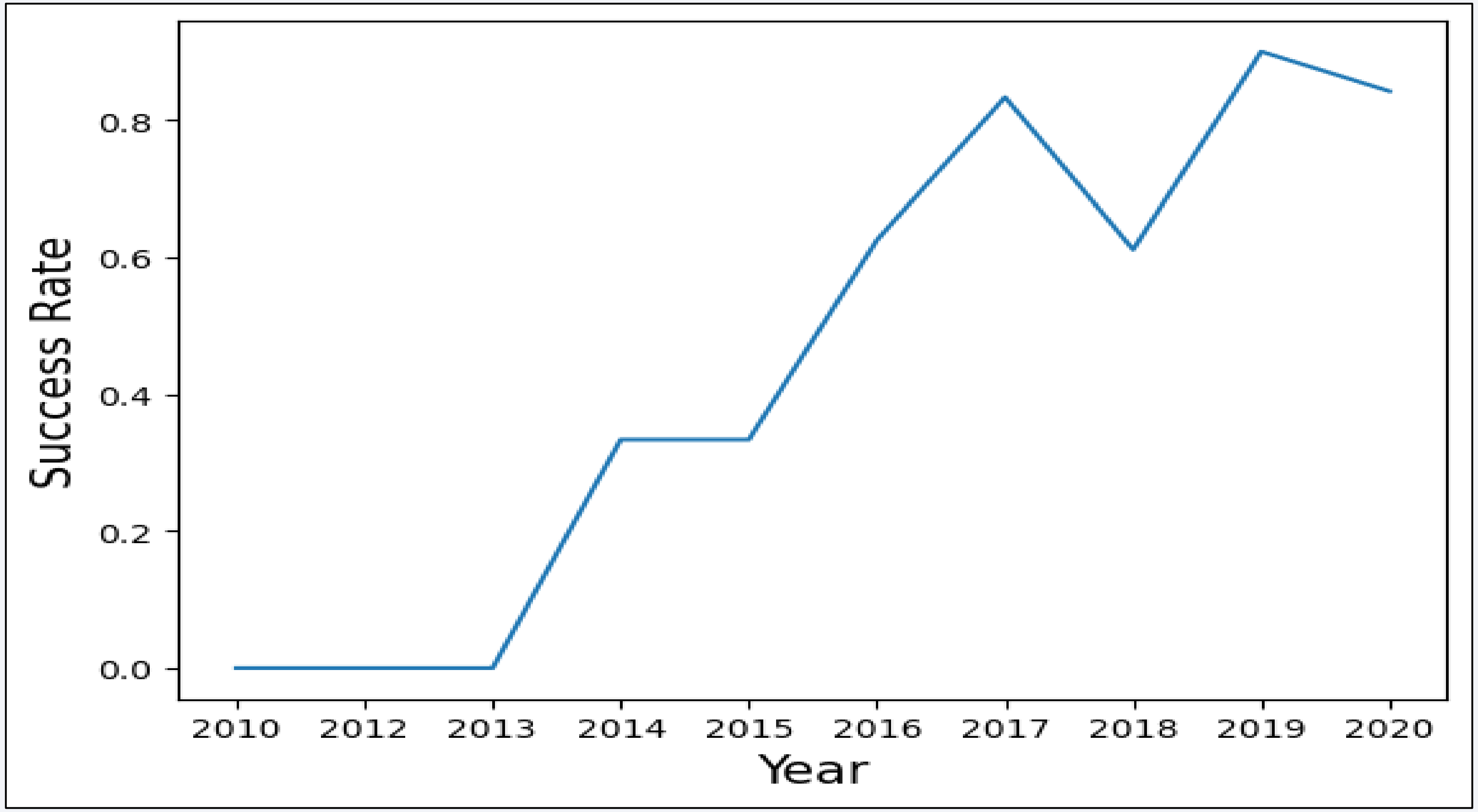
Payload vs. Orbit Type

Scatter point of payload vs. orbit type



- With heavy payloads the successful landing rates are more for Polar, LEO and ISS Orbits.
- All launches from SSO, HEO, GEO and ES-L1 orbit types had successful landing outcome. All had light payloads.
- For GTO orbit – there are no conclusive findings

Launch Success Yearly Trend



The success rate since 2013 kept increasing till 2020

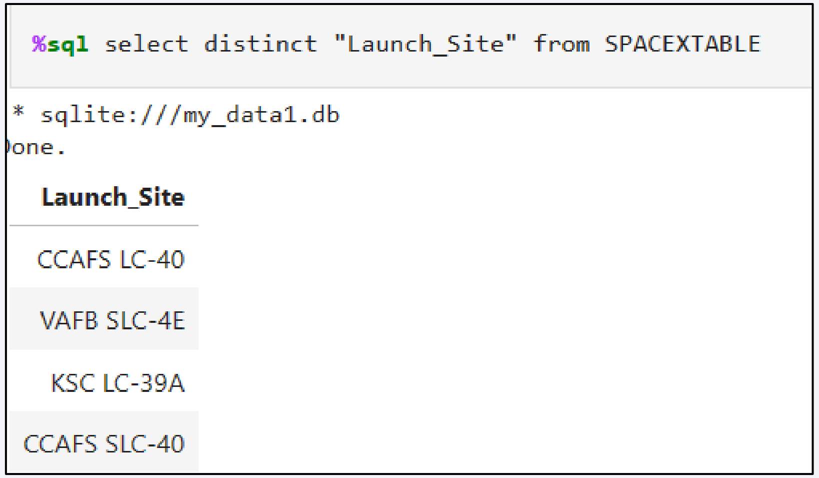
All Launch Site Names

- Code to find the names of the unique launch sites

```
%sql select distinct "Launch_Site" from SPACEXTABLE;
```

- Output: There are 4 launch sites from where Falcon 9 rockets are launched

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



```
%sql select distinct "Launch_Site" from SPACEXTABLE
```

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

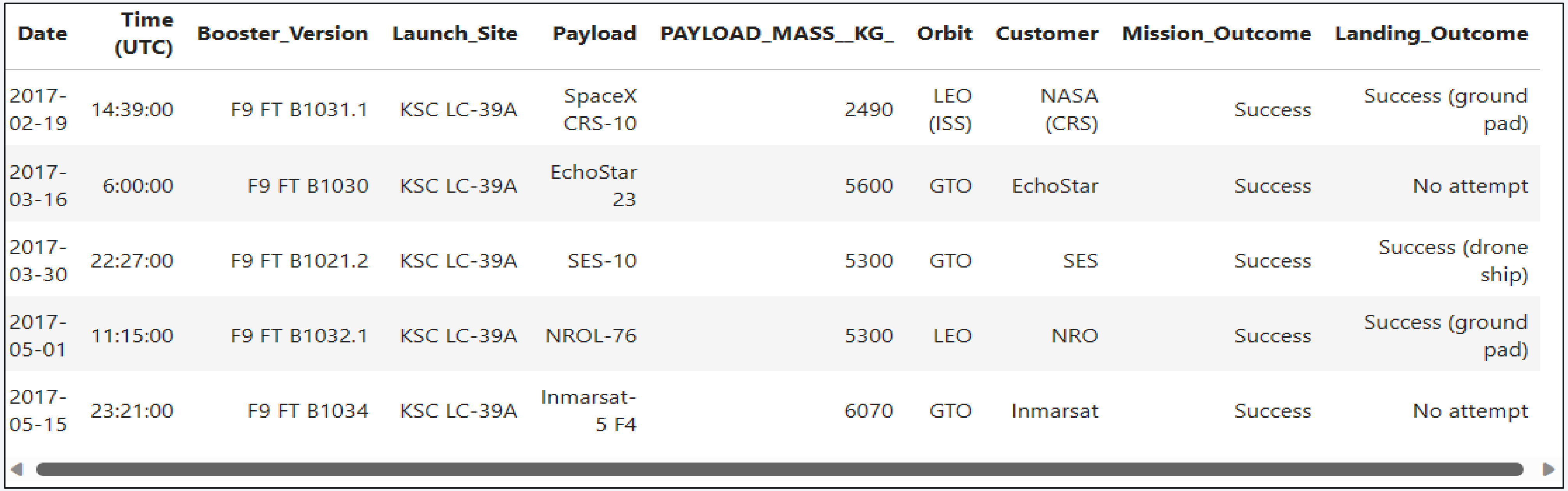
Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'KSC'

- Code to find 5 records where launch sites’ names start with `KSC`

```
%sql select * from SPACEXTABLE where "Launch_Site" like ("KSC%") limit 5;
```

- Output:



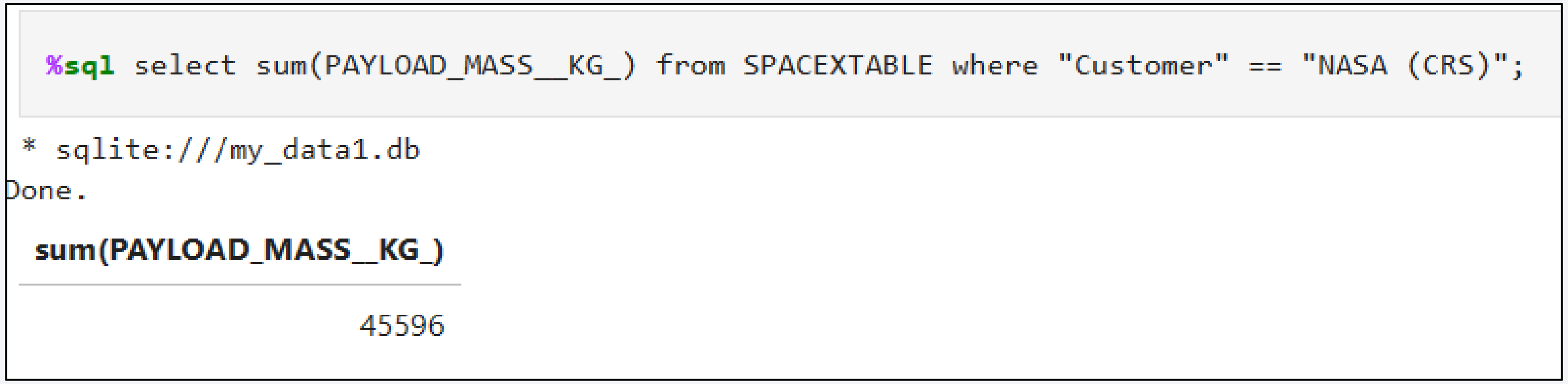
Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-03-16	6:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2017-05-15	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

Total Payload Mass

- Code to find out the total payload carried by boosters from NASA:

  %sql select sum(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE "Customer" == "NASA (CRS)";

- Output: 45596 Kgs



```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where "Customer" == "NASA (CRS)";
```

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

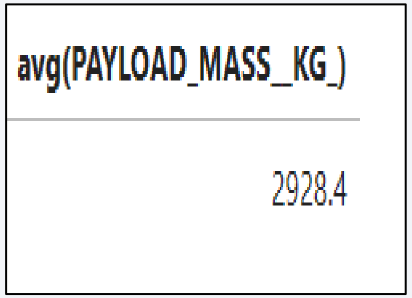
sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Code to calculate the average payload mass carried by booster version F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_) FROM
SPACEXTABLE WHERE "Booster_Version" == "F9 v1.1";
```

- Output: 2928.4 Kgs



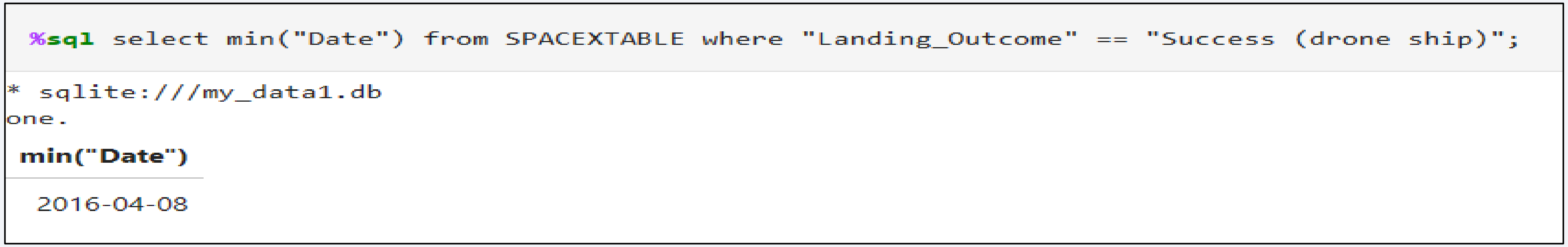
avg(PAYLOAD_MASS_KG_)
2928.4

First Successful Ground Landing Date

- Code to find the dates of the first successful landing outcome on drone ship.

  %sql select min("Date") FROM SPACEXTABLE WHERE "Landing_Outcome" == "Success (drone ship)";

- Output: 08-August-2016



```
%sql select min("Date") from SPACEXTABLE where "Landing_Outcome" == "Success (drone ship)";
```

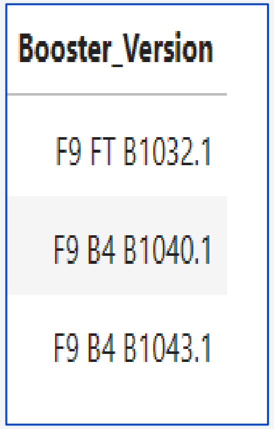
```
* sqlite:///my_data1.db
one.
 min("Date")
  2016-04-08
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- Code to find out the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
%sql select "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome"
== "Success (ground pad)" and PAYLOAD_MASS__KG_ between 4000 and 6000;
```

- Output:



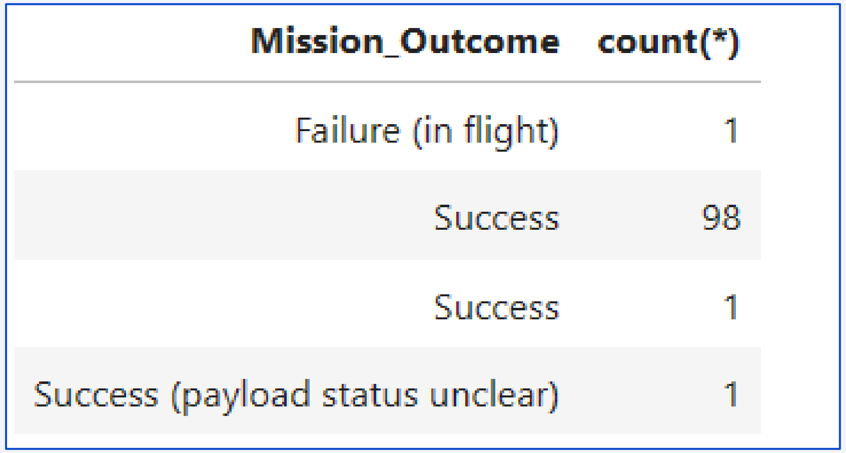
Booster_Version
F9 FT B1032.1
F9 B4 B1040.1
F9 B4 B1043.1

Total Number of Successful and Failure Mission Outcomes

- Code to calculate the total number of successful and failure mission outcomes

```
%sql select "Mission_Outcome", count(*) from SPACEXTABLE group by
"Mission_Outcome";
```

- Output:



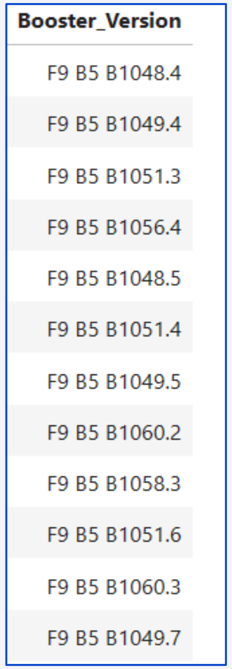
Mission_Outcome	count(*)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Code to find the names of the boosters which have carried the maximum payload mass

```
%sql select "Booster_Version" from SPACEXTABLE where PAYLOAD_MASS__KG_ == (select
max(PAYLOAD_MASS__KG_) from SPACEXTABLE);
```

- Output:



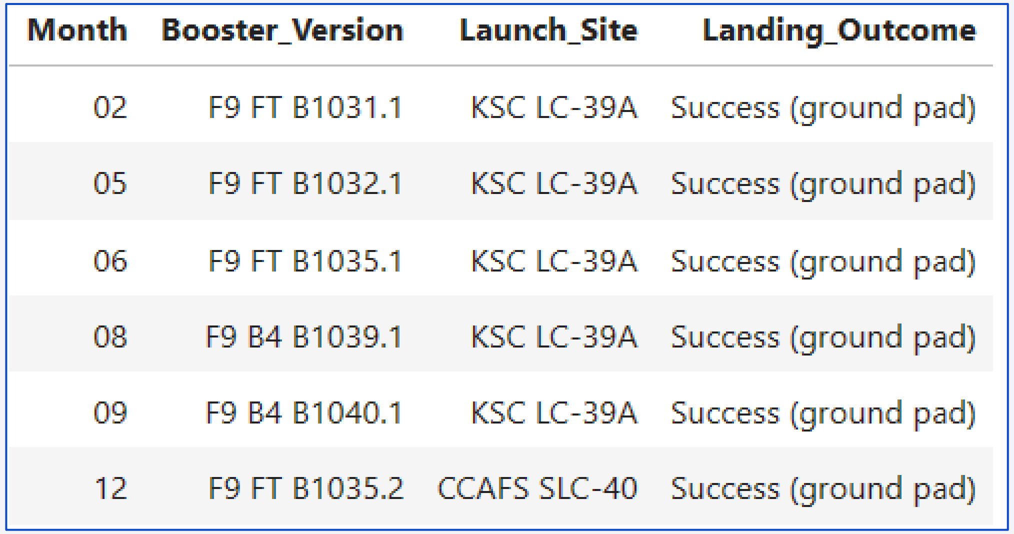
Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2017 Launch Records

- Code to obtain the records which will display the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017

```
sql select substr("Date", 6,2) as "Month", "Booster_Version", "Launch_Site", "Landing_Outcome"
from SPACEXTABLE where "Landing_Outcome" == "Success (ground pad)" and substr("Date", 1,4)
== "2017";
```

- Output:



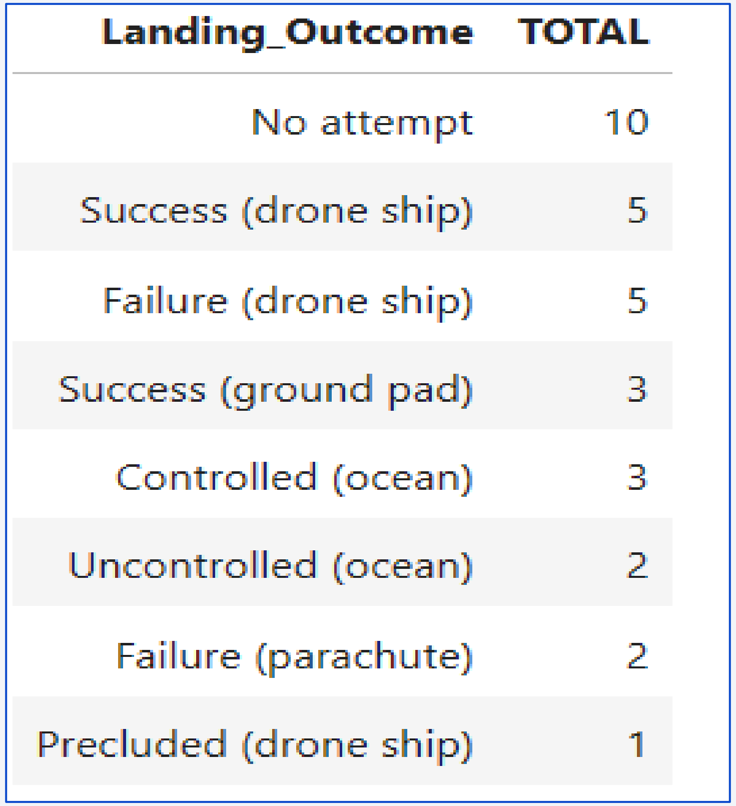
Month	Booster_Version	Launch_Site	Landing_Outcome
02	F9 FT B1031.1	KSC LC-39A	Success (ground pad)
05	F9 FT B1032.1	KSC LC-39A	Success (ground pad)
06	F9 FT B1035.1	KSC LC-39A	Success (ground pad)
08	F9 B4 B1039.1	KSC LC-39A	Success (ground pad)
09	F9 B4 B1040.1	KSC LC-39A	Success (ground pad)
12	F9 FT B1035.2	CCAFS SLC-40	Success (ground pad)

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

- Code to obtain the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, ranked 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 TOTAL DESC;
```

- Output:

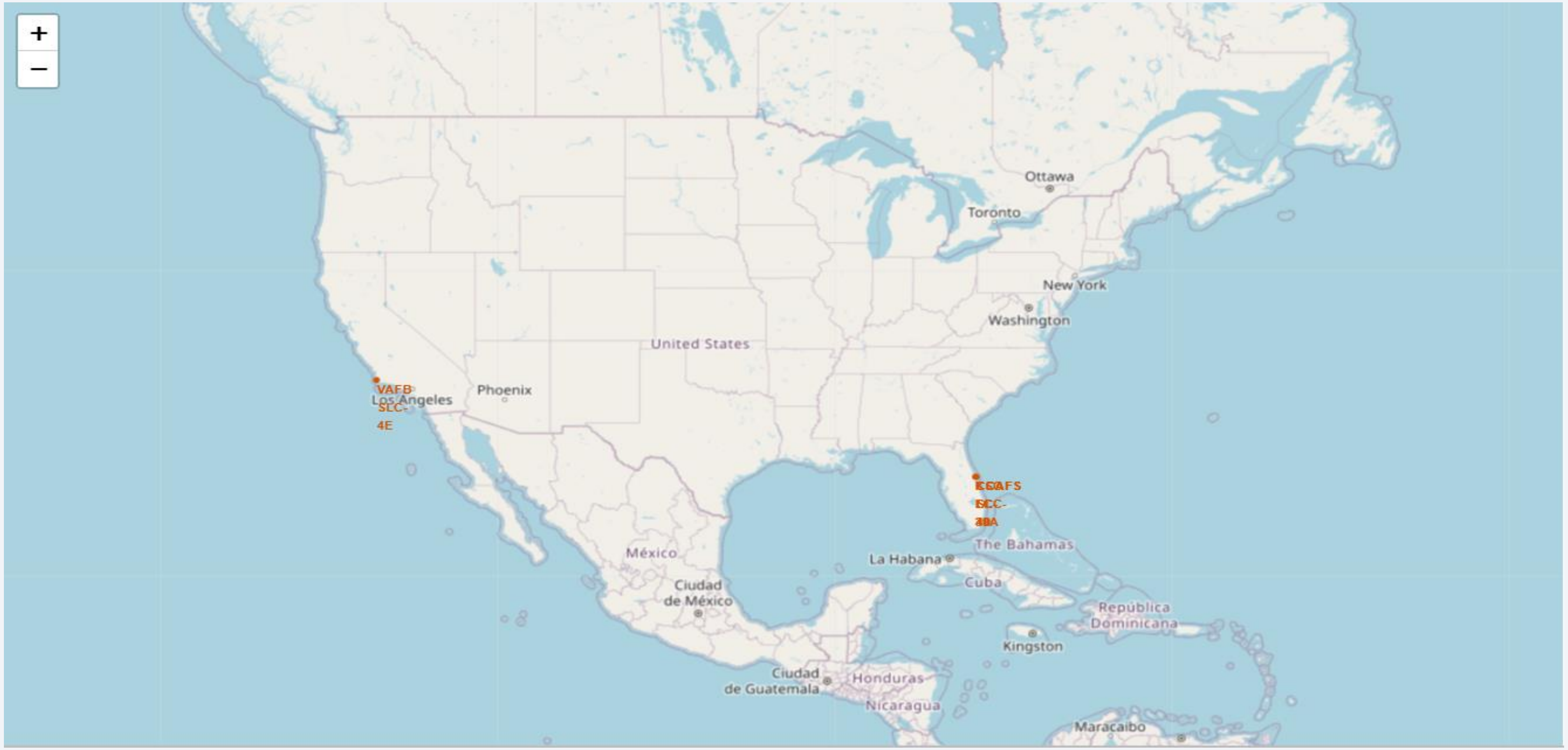


Landing_Outcome	TOTAL
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

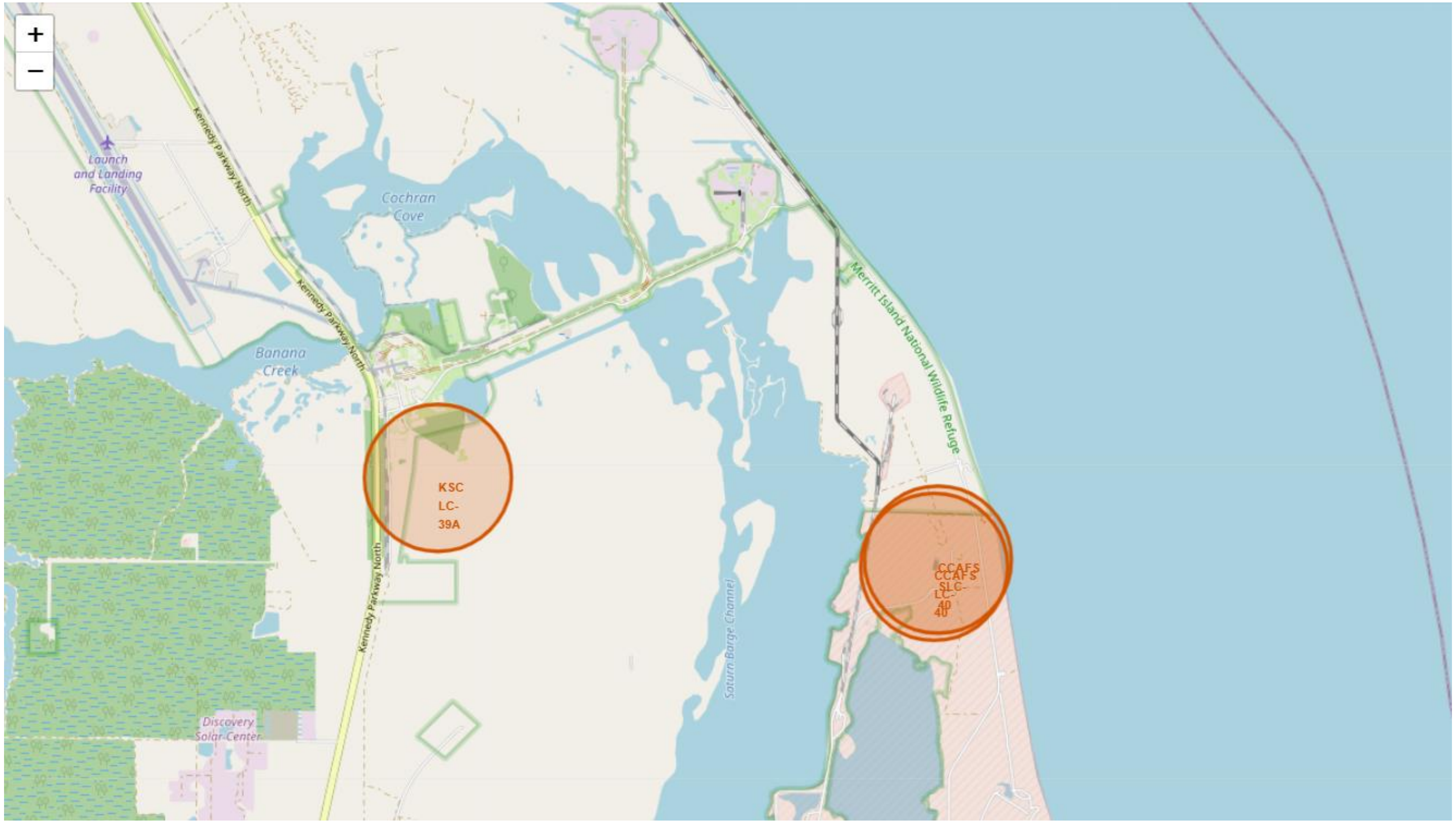
Launch Sites Proximities Analysis

Launch Sites Location Analysis



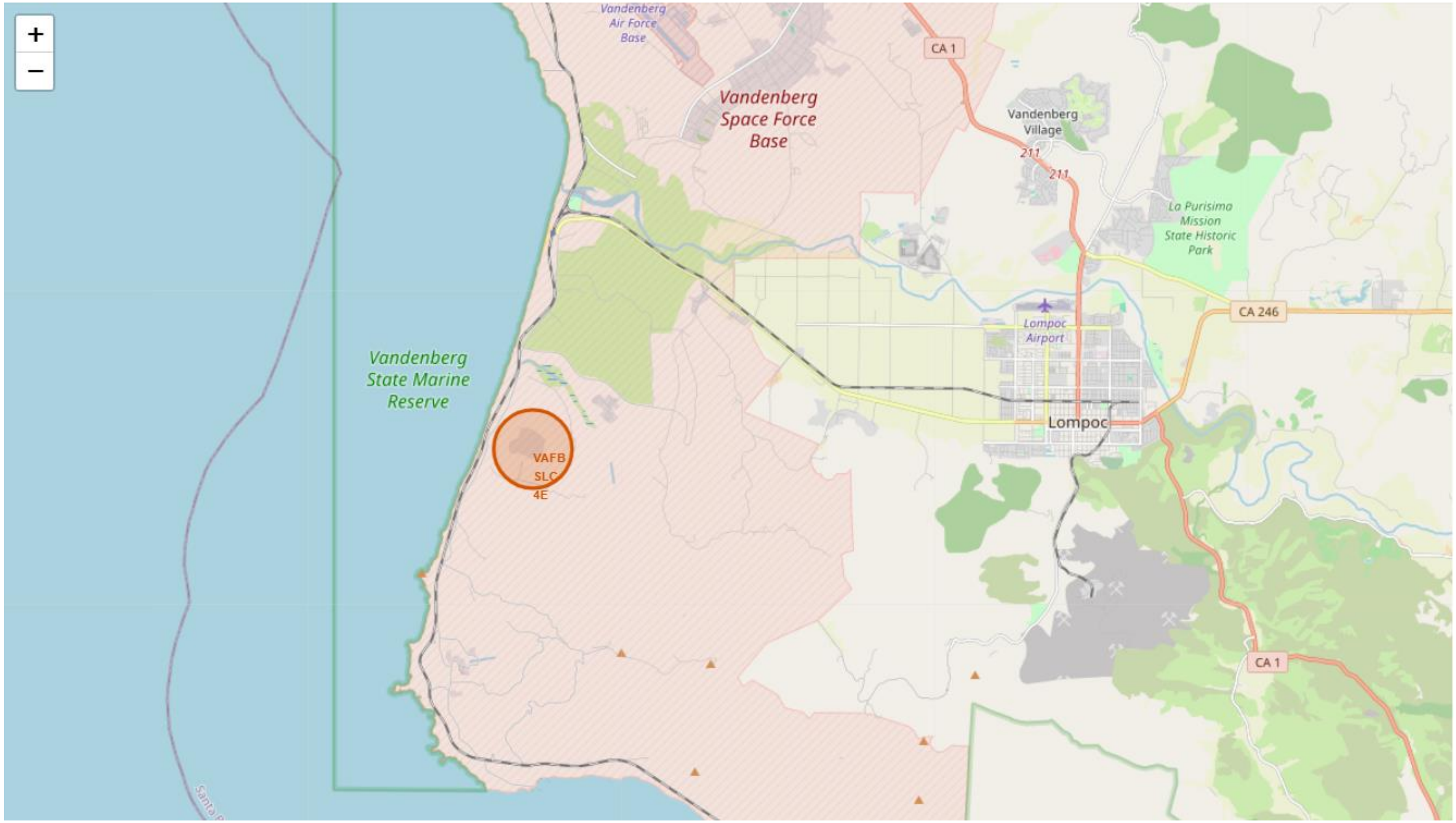
All launch sites in proximity to the Equator line

Launch Sites Location Analysis



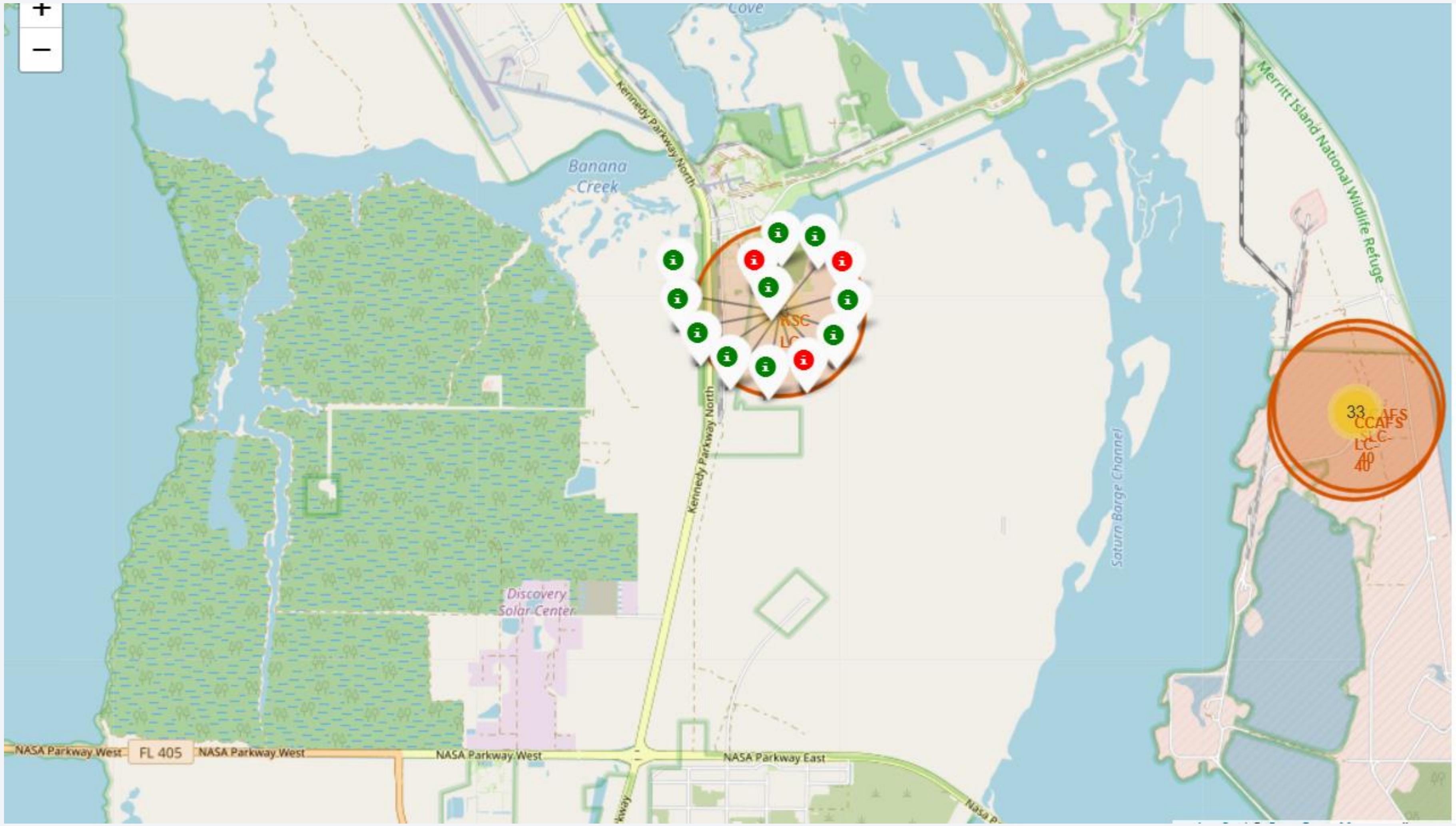
All launch sites in close proximity to the Coast

Launch Sites Location Analysis


+
−
Vandenberg Air Force Base
Vandenberg Space Force Base
CA 1
Vandenberg Village
211
La Purisima Mission State Historic Park
CA 246
Lompoc Airport
Lompoc
Vandenberg State Marine Reserve
VAFB SLC 4E
CA 1

All launch sites in close proximity to the coast

Launch outcomes for KSC LC-39A

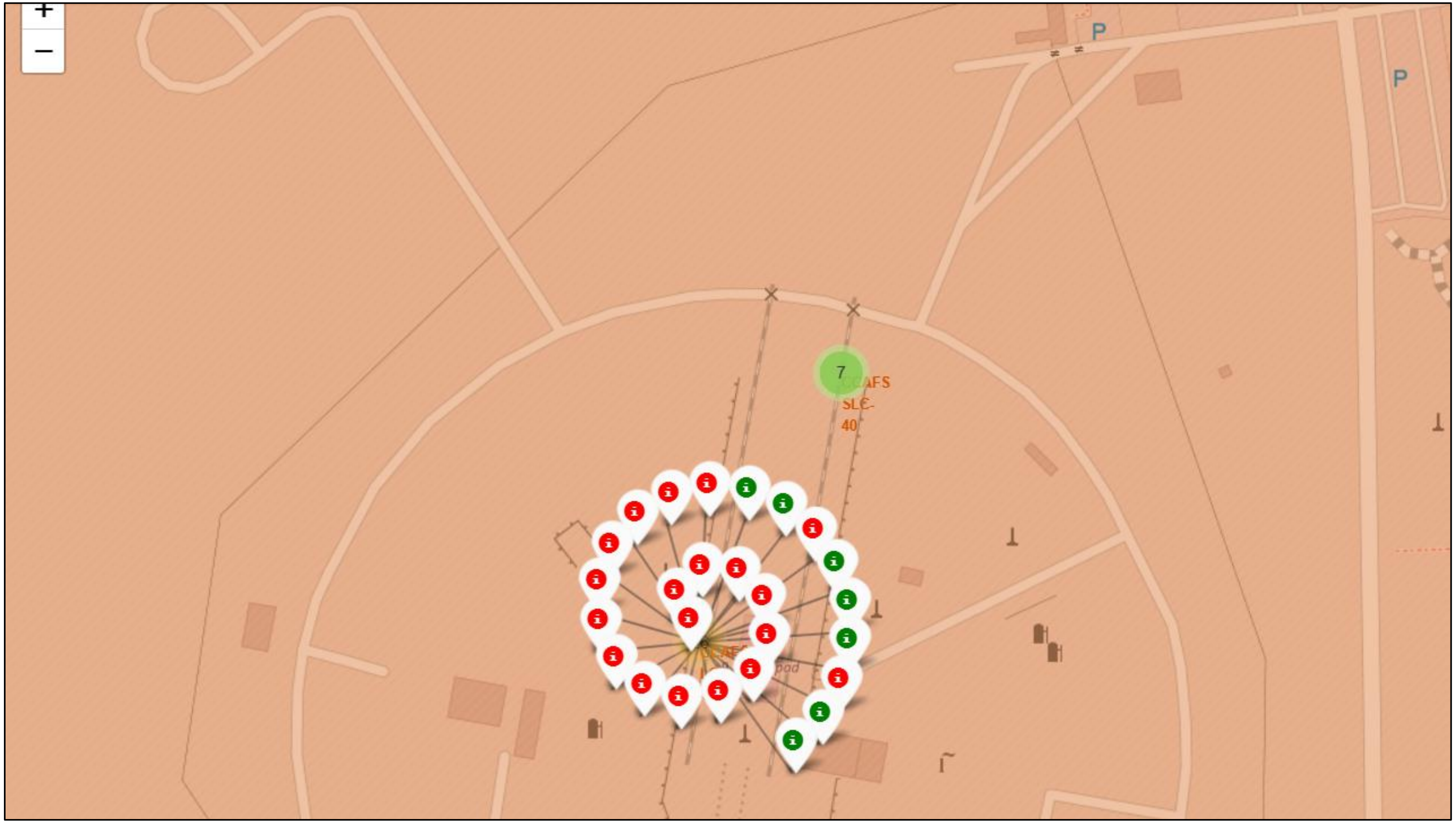


Green marker indicates success

Red marker indicates failure

Success rate is high

Launch outcomes for CCAFS LC-40

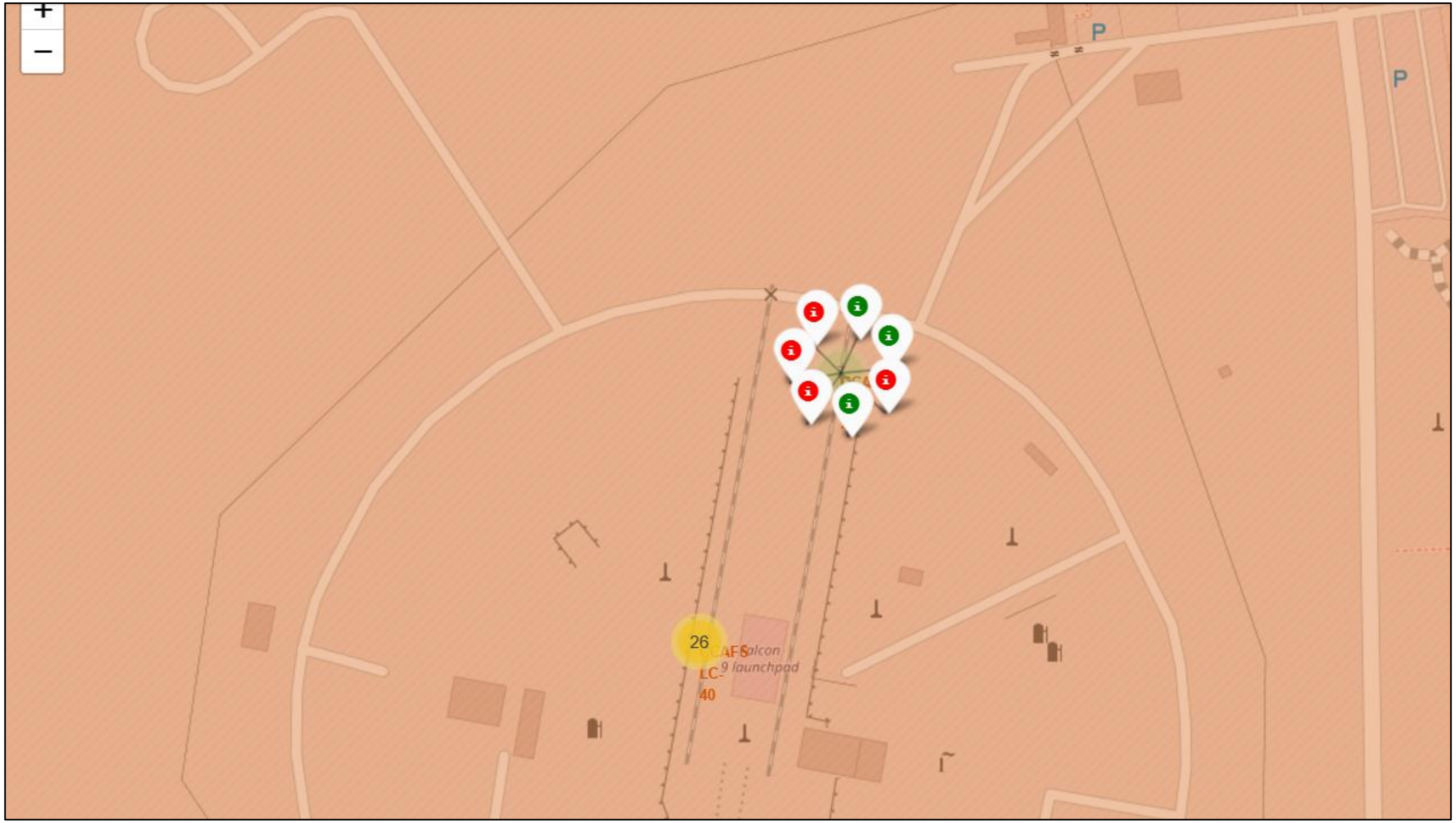


Green marker indicates success

Red marker indicates failure

Success rate is low

Launch outcomes for CCAFS SLC-40

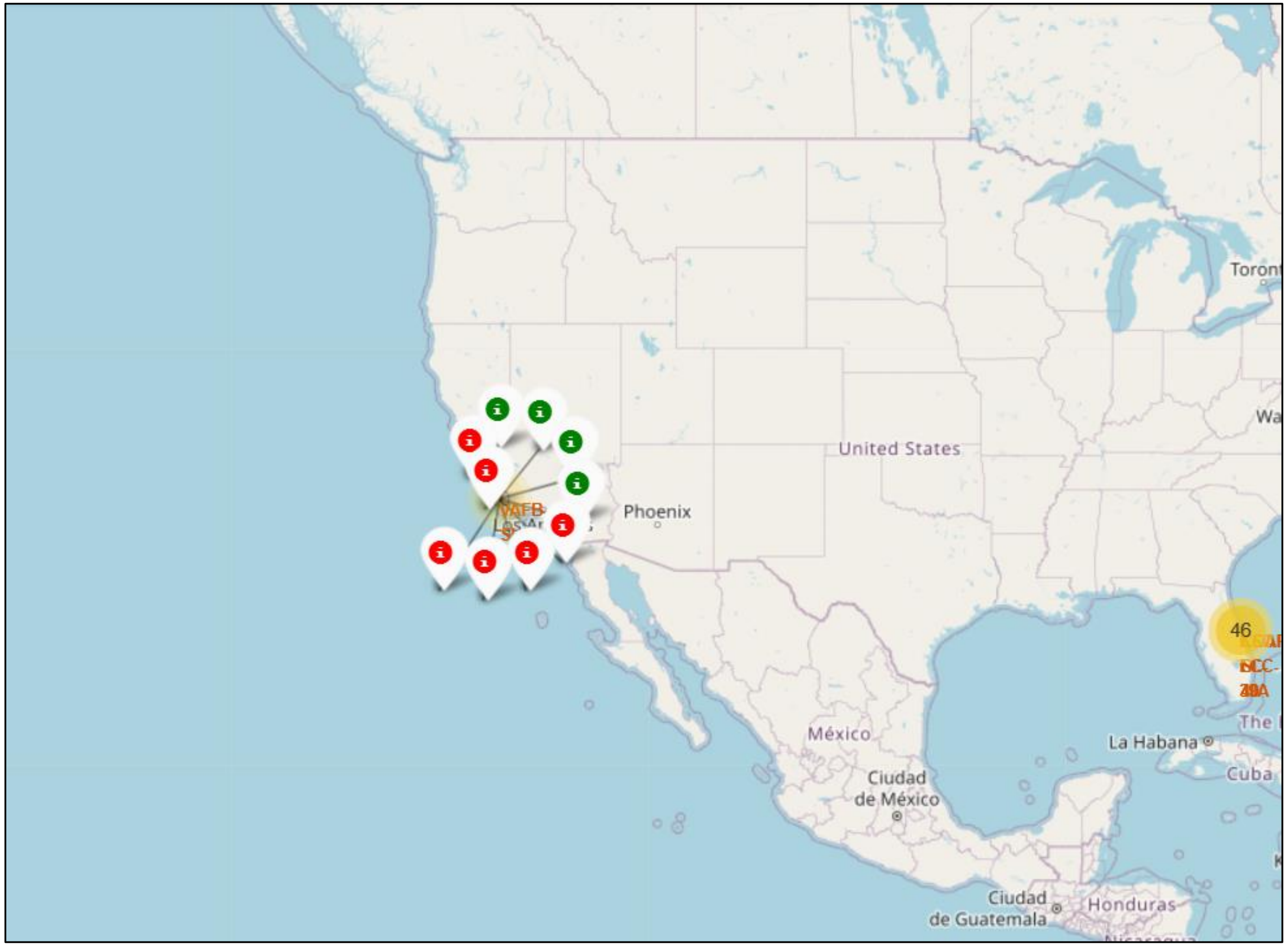


Green marker indicates success

Red marker indicates failure

Success rate is moderate

Launch outcomes for VAFB SLC 4E



Success rate is low

Green marker indicates success

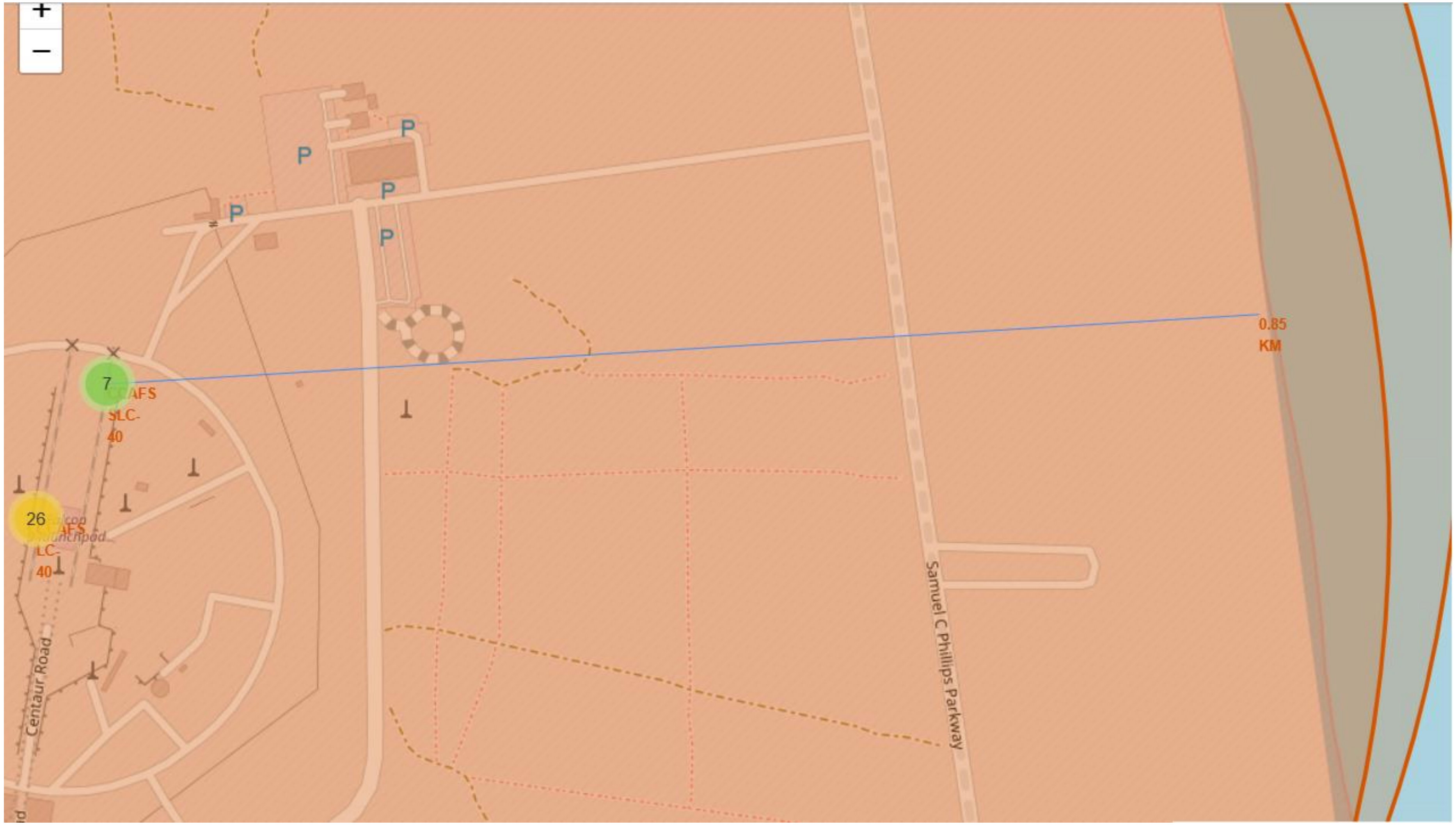
Red marker indicates failure

Analysis of launch site proximities

Proximity of launch sites to the following was analysed:

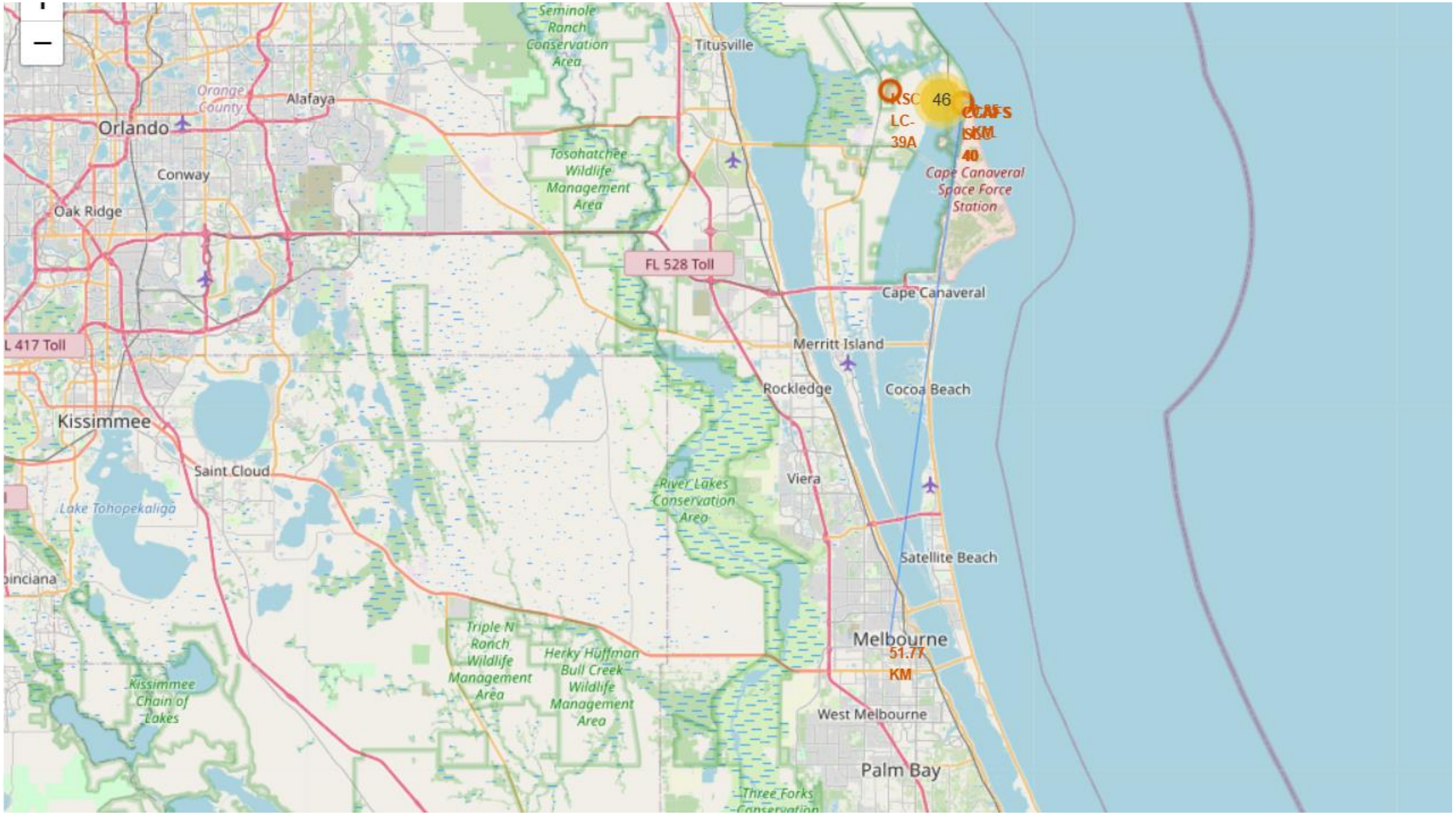
- Railways
- Highways
- Coastline
- Cities

Launch Sites Proximity to Coastline



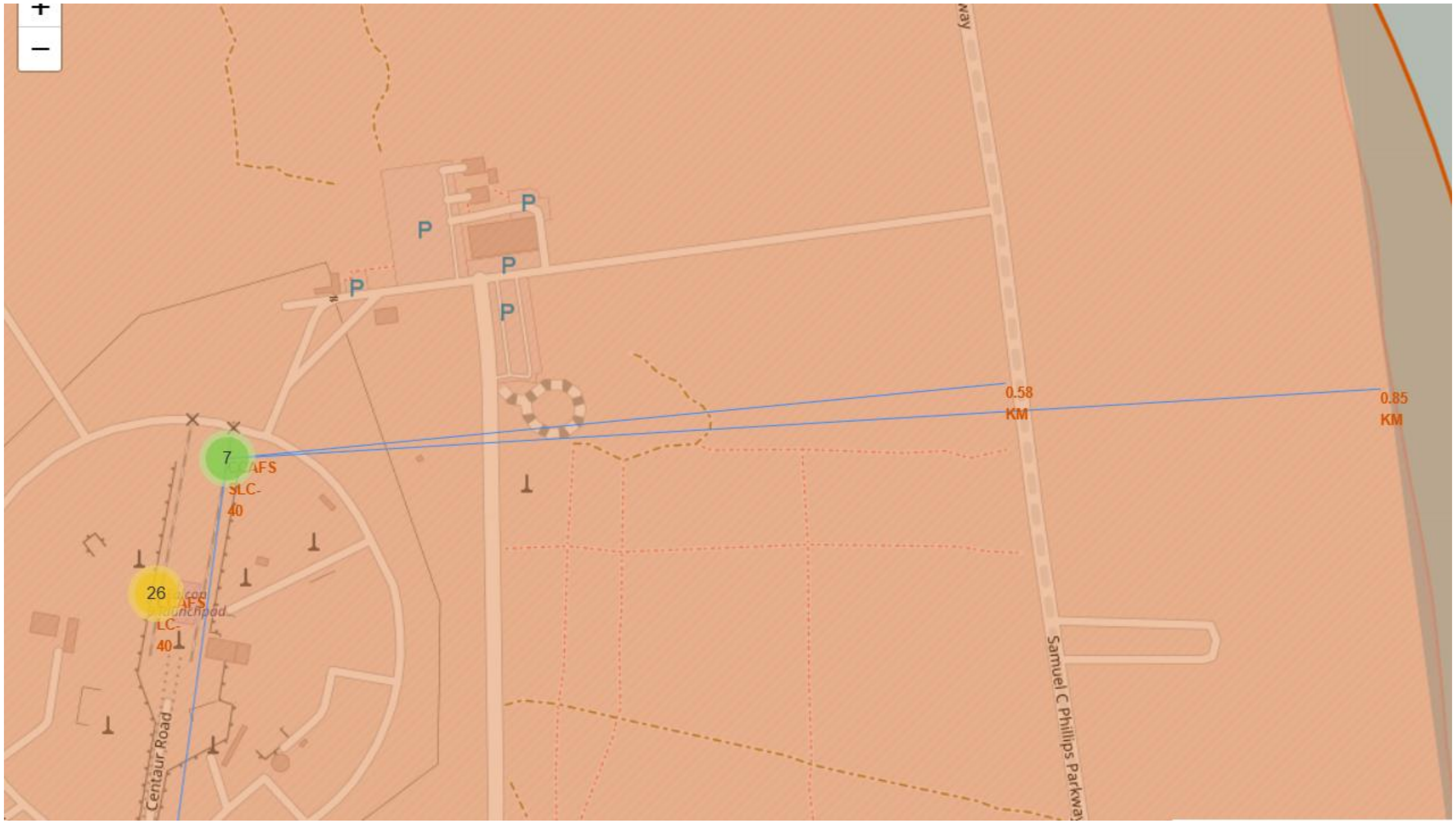
Launch sites are close to coastline

Launch Sites Proximity to Cities



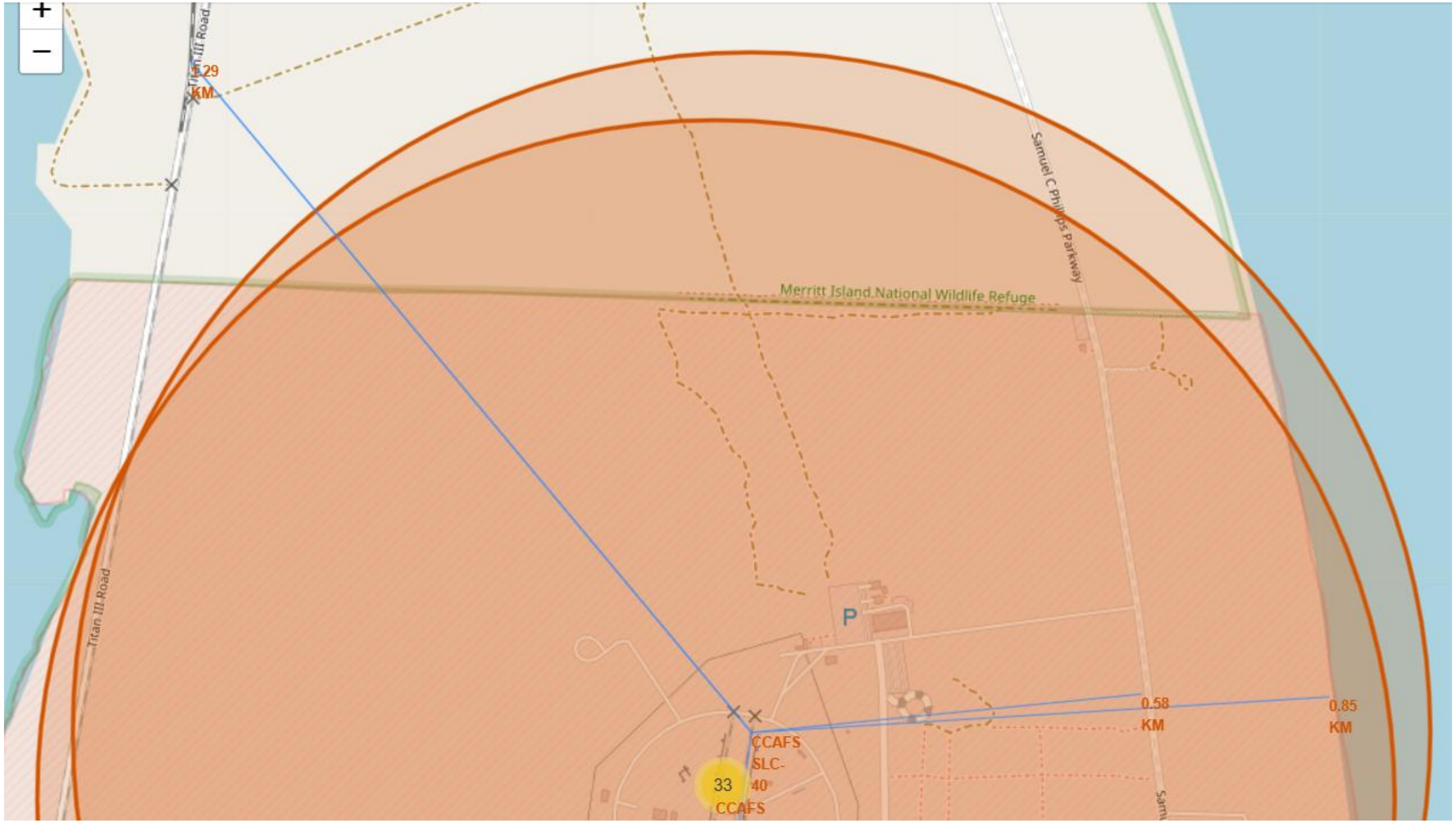
Launch sites are away from cities

Launch Sites Proximity to Highways



Launch sites are close to highways

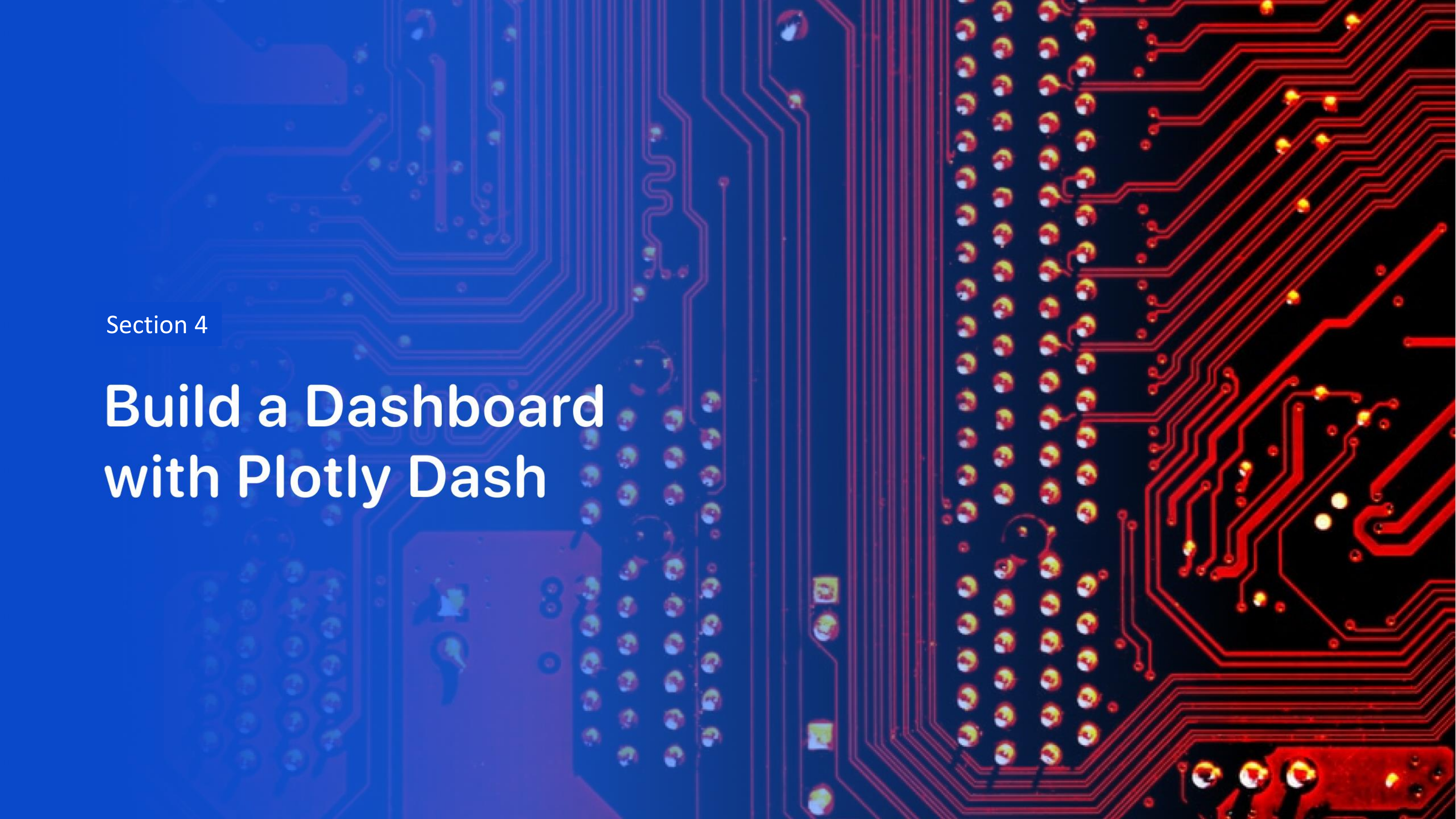
Launch Sites Proximity to Railways



Launch sites are close to Railways

Section 4

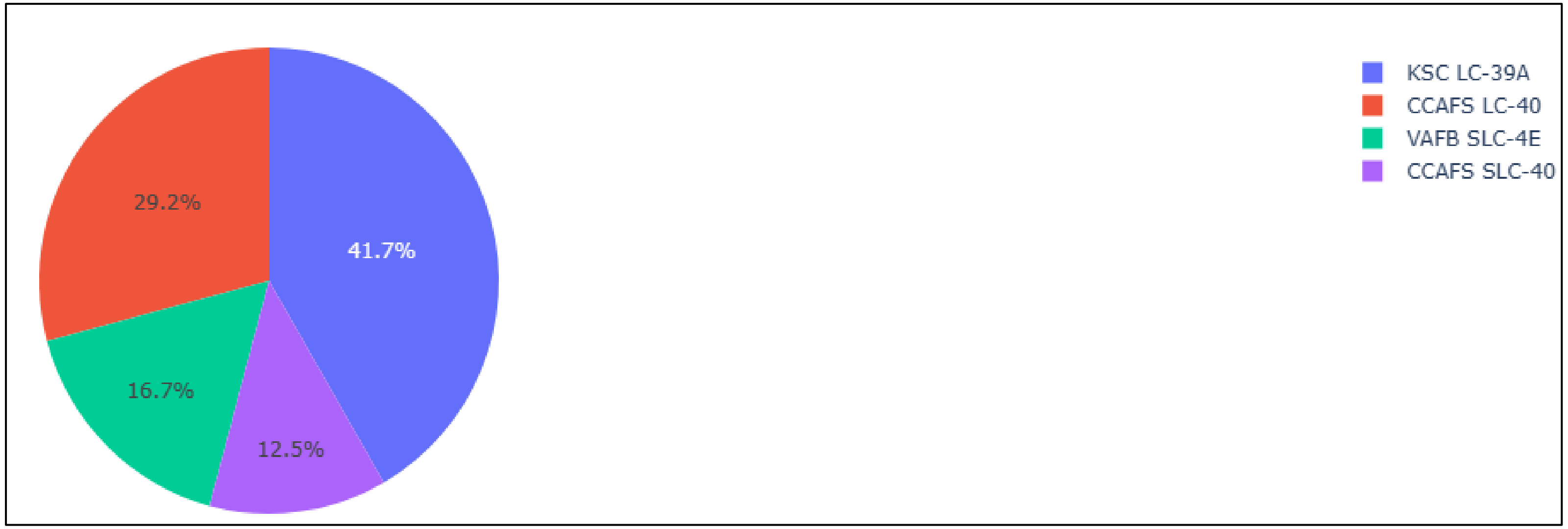
Build a Dashboard with Plotly Dash



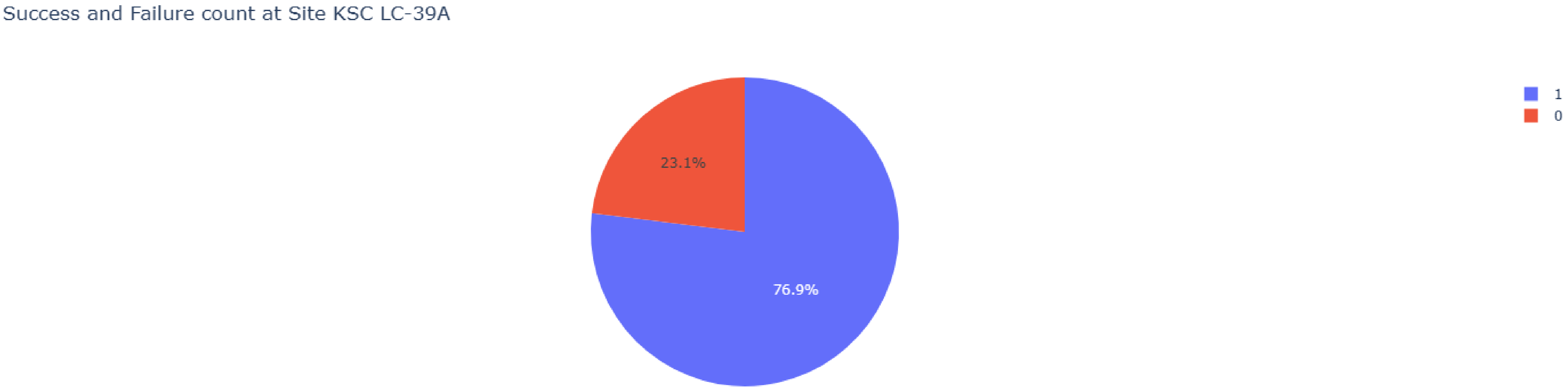
Launch Success Count for all sites

Most number of successful launches is from KSC LC-39A

The lowest number of successful launches is from CCAFS SLC-40



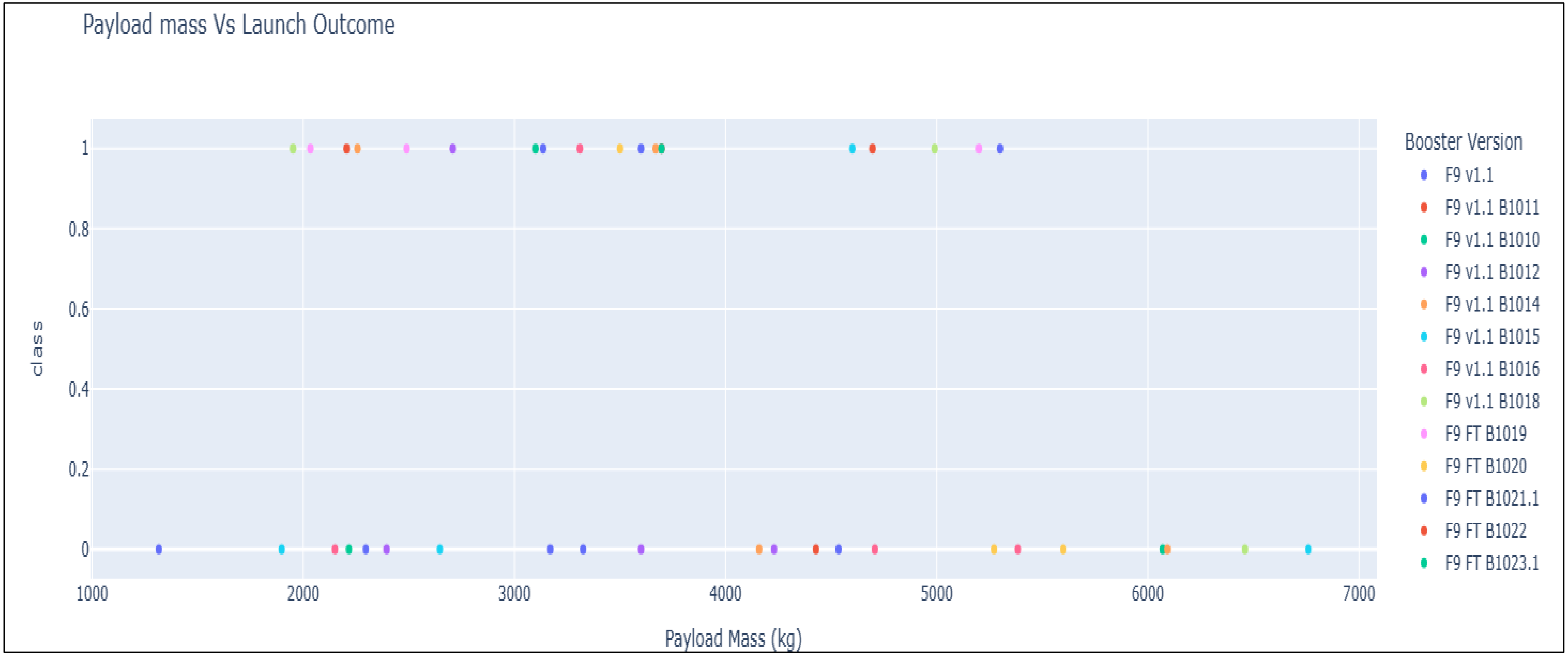
Highest Launch Success Ratio



Success and Failure count at Site KSC LC-39A

KSC LC-39A has the highest launch success ratio at 76.9%

Payload vs. Launch Outcome scatter plot

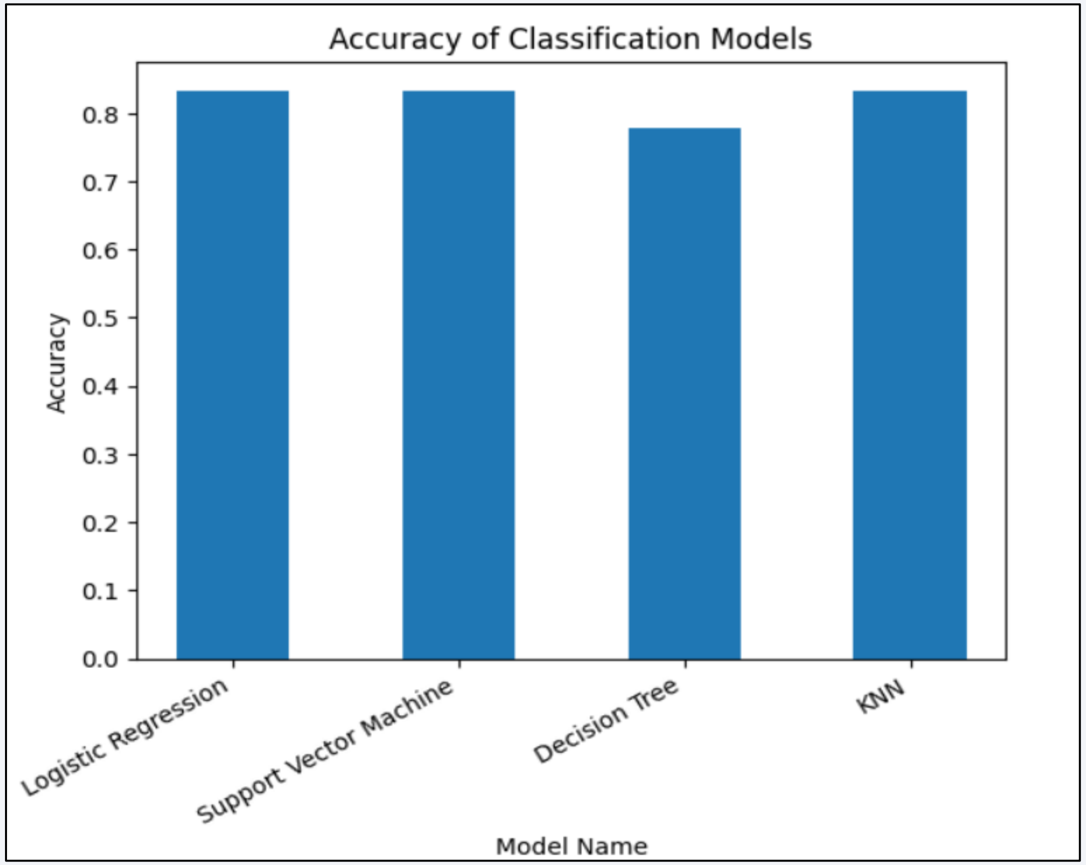


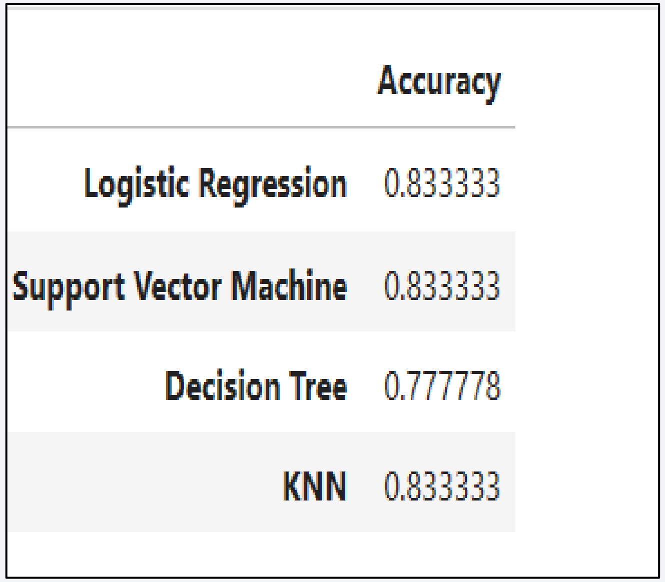
Launches having medium payload range (2000 kg to 4000 kg) have the highest success rate while those lower than 2000 kg or higher than 5500 kg have low success rate

Section 5

Predictive Analysis (Classification)

Classification Accuracy



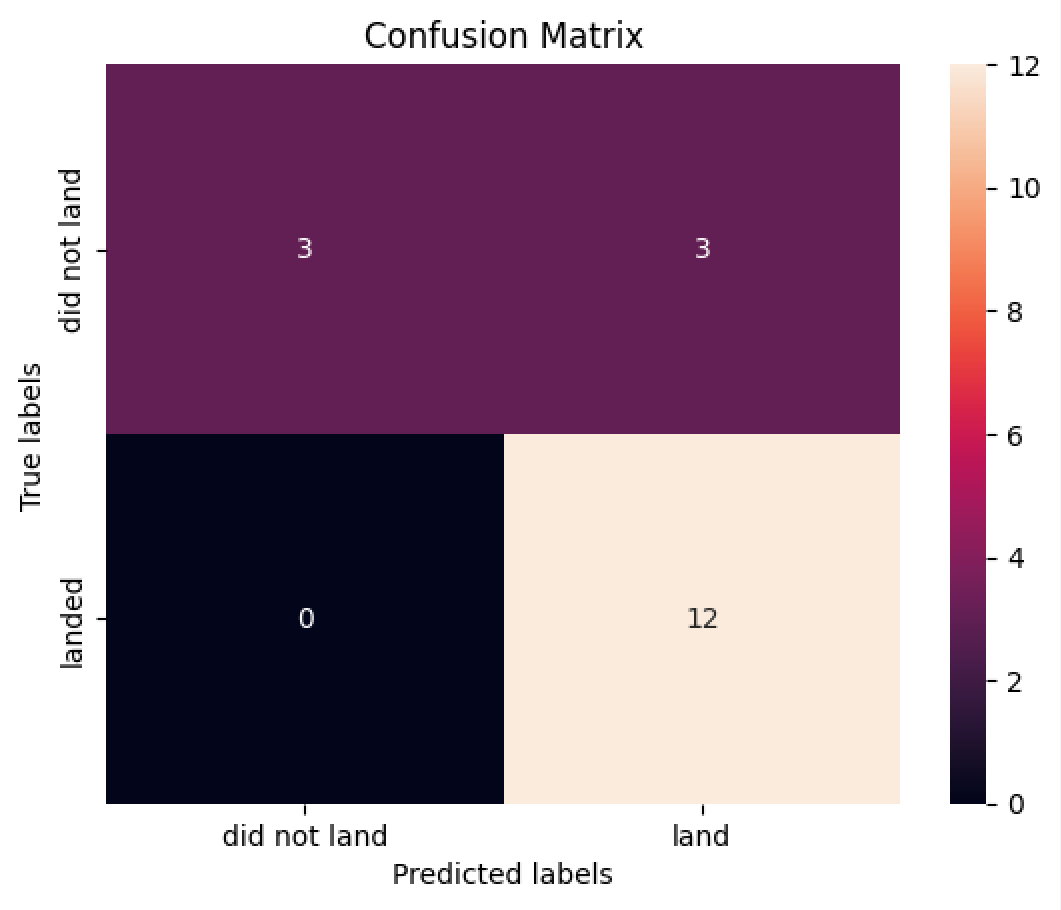


	Accuracy
Logistic Regression	0.833333
Support Vector Machine	0.833333
Decision Tree	0.777778
KNN	0.833333

Logistic regression, SVM and KNN have the same accuracy score of 83.3%, on test data.

Accuracy of decision tree model was slightly lower at 77.8%.

Confusion Matrix



Confusion matrix of the 3 models with same accuracy score- viz. Logistic regression, SVM and KNN , is given on the slide.

All 3 models are able to predict landing failure with 100% accuracy. The accuracy of prediction of successful landing is slightly less at 80%.

Conclusions

A. Objective of the analysis was to determine if in a launch, the Falcon 9 first stage will land successfully.

B. Factors which influence success of landing of the rocket were identified based on Exploratory data Analysis. Some of the factors are Booster Version, Payload mass, launchpad, No. of cores, Orbit Type, Flight Number, etc. The following conclusions are drawn from analysis of past Falcon 9 rocket launches:

1. Landing success increases with increase in flight number
2. Landing success decreases with increase in payload mass. Medium payload range (2000 kg to 4000 kg) have the highest launch success rate
3. All launches for orbit types ES-L1, GEO, HEO and SSO were successful
4. For heavy payloads, Polar, LEO and ISS orbits may be preferable for launch, as these have more successful landings in such payload range
5. Success rate of launches is highest for KSC LC-39A launch site.
6. SLC-40 has the lowest share of successful launches

Thank you!