IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ayesheh
4 September 2023

Outline

Executive Summary

Summary of Methodologies

The goal of the research is to pinpoint what makes a rocket landing effective. The following approaches were employed to arrive at this conclusion:

- **Collect** data obtained using the SpaceX REST API and web scraping methods
- **Wrangle** Using information, establish a success/fail result variable.
- **Explore** using data visualization approaches while considering the following factors: launch site, flight number, payload, and annual trend
- **Analyze** using SQL, the following statistics were determined: payload, total payload
- range for successful launches, as well as the overall number of successful and unsuccessful outcomes
- **Explore** success rates at launch sites and proximity to geographic landmarks
- **Visualize** the launch locations and payload ranges that have had the most success
- **Create** models employing support vector machines and logistic regression to forecast landing outcomes decision tree, K closest neighbor (KNN), and SVM

Results

Exploratory Data Analysis:

- Over time, launch success has increased
- Among landing locations, KSC LC-39A has the best success rate.
- 100% of attempts on orbits ES-L1, GEO, HEO, and SSO succeed.

Visualization/Analytics:

- All launch points are close to the shore, and most are near the equator.

Predictive Analytics:

- On the test set, all models performed equally. The decision tree mode fared somewhat better.

Introduction

Leading space company SpaceX works to lower the cost of space travel for all people. Its successes include building a satellite network that offers internet access, launching manned space flights, and delivering spacecraft to the international space station. SpaceX can accomplish this because to their inventive reuse of the first stage of its Falcon 9 rocket, which makes rocket launches relatively affordable ($62 million each launch). Other suppliers cost more than $165 million apiece and cannot utilize the first stage. The cost of the launch may be calculated by calculating if the first stage will land. To achieve this, we can forecast whether SpaceX or a rival business will be able to reuse the first stage using open data and machine learning algorithms.

Problem Statements:

What elements determine whether the rocket will successfully land?

What are the different elements interacting to affect the likelihood of a successful landing?

What operational requirements must be met to guarantee a successful landing program?

Section 1

Methodology

Methodology

Executive Summary

- Utilize web scraping and the SpaceX REST API to get data.
- Prepare data for analysis and modeling by filtering the data, addressing missing values, and using one hot encoding.
- SQL and data visualization tools are used to explore data using EDA.
- Use Plotly Dash and Folium to visualize the data.
- Create models employing categorization techniques to forecast landing results. To identify the optimal model and parameters, tune and assess your models.

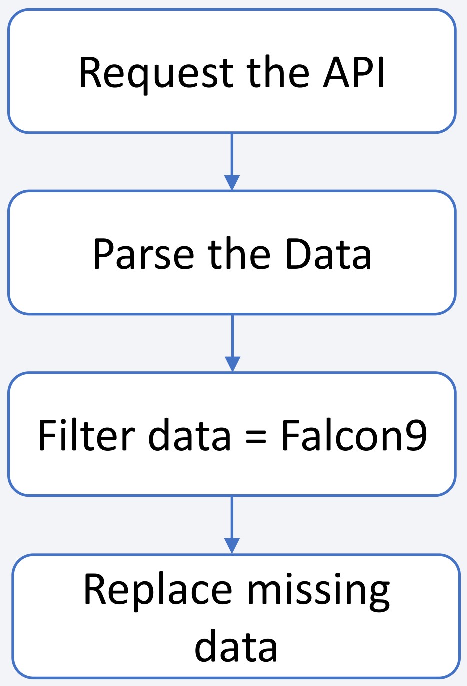
Data Collection

Datasets were collected from SpaceX API (https://api.spacexdata.com/v4/rockets/) and from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches), using web scraping techniques.

- Requested file was in .json format
- Turned into Panda Dataframe

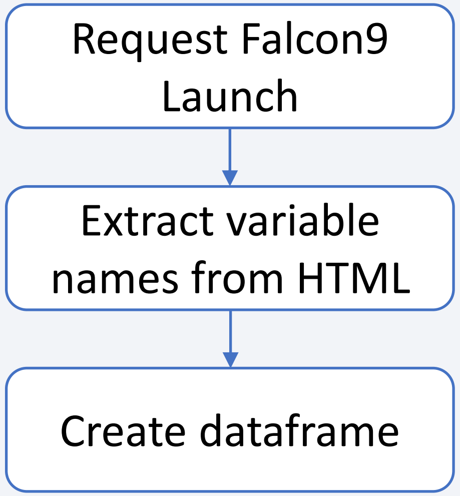
Data Collection – SpaceX API

- Request information (rocket launch data) from the SpaceX API.
- Utilizing. json () to decode the response and json_normalize to transform it into a dataframe.
- Utilize custom functions to ask the SpaceX API for information on the launches.
- Make a dictionary using the information.
- From the dictionary, create a dataframe.
- Set a filter to only include Falcon 9 launches in the dataframe.
- Calculated.mean() should be used to replace any missing values for Payload Mass.
- Data export to a CSV file.



Data Collection – Scraping

- Request information from Wikipedia on the Falcon 9 launch
- From an HTML answer, create a BeautifulSoup object.
- Take the column names out of the HTML table header.
- Gather information by analyzing HTML tables
- create a dictionary using the information
- from the dictionary, create a dataframe
- Data export to a CSV file



Data Wrangling

EDA is performed, and data labels are chosen.

Calculate: the number of launches for each site, the frequency of orbits, and the likelihood that missions will succeed for each type of orbit.

Create a column for the binary landing outcome (dependent variable).

Data export to a CSV file.

Landing result: It wasn't always successful to land.

- False Ocean: a failed attempt to land in a certain oceanic area.
- True RTLS indicated that the mission's landing attempt on a ground pad was successful.
- False RTLS: depicted a failed landing attempt on a ground pad.
- True ASDS: referred to the mission's successful drone ship landing.
- False ASDS: portrayed a failed landing on a drone ship translated results into 1 for a successful landing and 0 for a failed landing.
- True Ocean: mission outcome had a successful landing to a specific region of the ocean.

EDA with Data Visualization

Charts

- Flight Number vs. Payload
- Flight Number vs. Launch Site
- Payload Mass (kg) vs. Launch Site
- Payload Mass (kg) vs. Orbit type

EDA with Visualization Analysis

- Scatter plots may be used to visualize relationships. If there is a link between the variables, machine learning might use them.
- Use bar charts to compare differences across distinct groups. Bar graphs display the connections between the categories and a quantified value.

EDA with SQL

Display:

- Names of unique launch sites
- 5 records where launch site begins with ‘CCA’
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1.

List:

- Date of the first successful ground landing names of boosters whose landings on drone ships were successful larger than 4,000 but less than 6,000 pound payload mass.
- total number of missions both successful and unsuccessful.
- names of the booster iterations that have transported the most weight.
- Failure to land results in drone ship, launch, and booster version website for the months of 2015.
- Number of landings between 2010 and2017 March 20 and 06 04 (desc).

Build an Interactive Map with Folium

Markers Indicating Launch Sites

- Added blue circle at **NASA Johnson Space Center's coordinate** with a **popup label** showing its name using its latitude and longitude coordinates
- Added red **circles** at **all launch sites coordinates with** a **popup label** showing its name using its name using its latitude and longitude coordinates

Map with Folium

- **Colored Markers of Launch Outcomes**
- Added **colored markers** of **successful**(green) and **unsuccessful**(red) **launches at** each launch site to show which launch sites have high success rates

Distances Between a Launch Site to Proximities

- Added **colored lines** to **show distance between launch** site **CCAFS SLC-40 and its** proximity to the **nearest coastline, railway, highway, and city**

Build a Dashboard with Plotly Dash

Dropdown List with Launch Sites

- Allow user to select all launch sites or a certain launch site

Slider of Payload Mass Range

- Allow user to select payload mass range

Pie Chart Showing Successful Launches

- Allow user to see successful and unsuccessful launches as a percent of the total

Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version

- Allow user to see the correlation between Payload and Launch Success

Predictive Analysis (Classification)

Charts

- **Create** NumPy array from the Class column
- **Standardize** the data with StandardScaler. Fit and transform the data.
- **Split** the data using train_test_split
- **Create** a GridSearchCV object with cv=10 for parameter optimization
- **Apply** GridSearchCV on different algorithms: logistic regression (LogisticRegression()), support vector machine (SVC()), decision tree (DecisionTreeClassifier()), K-Nearest Neighbor (KNeighborsClassifier())
- **Calculate** accuracy on the test data using .score() for all models
- **Assess** the confusion matrix for all models
- **Identify** the best model using Jaccard_Score, F1_Score and Accuracy

Results

Exploratory Data Analysis

- Over time, launch success has increased
- Among landing locations, KSC LC-39A has the best success rate.
- 100% of attempts on orbits ES-L1, GEO, HEO, and SSO succeed.

Visual Analytics

- All launch points are close to the shore, and most are near the equator.
- Launch locations are far enough from potential damage sources (cities, highways, and railroads) without preventing access to people and supplies needed for launch activities.

Predictive Analytics

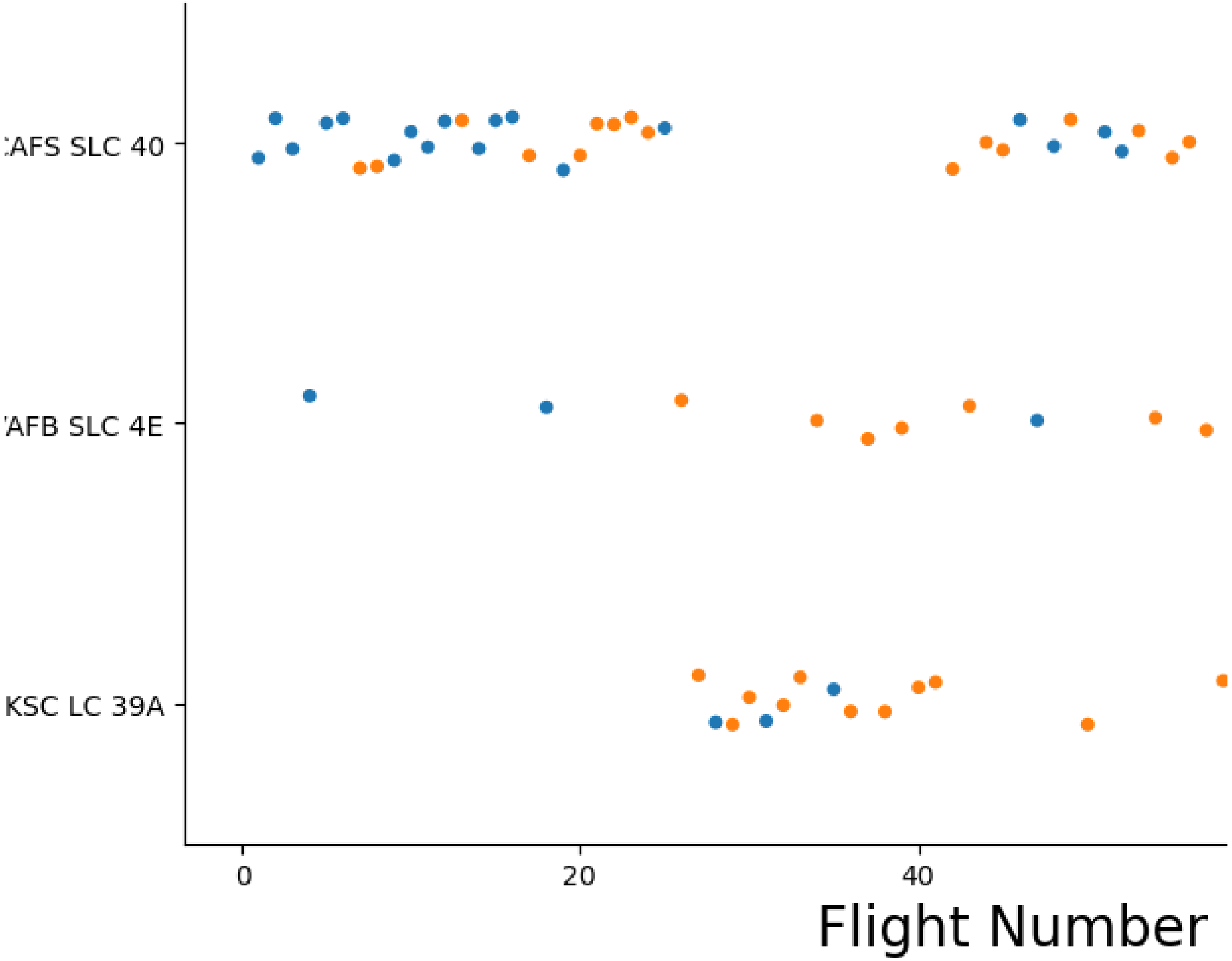
- The most accurate prediction model for the dataset is the decision tree model.

Section 2

Insights drawn from EDA

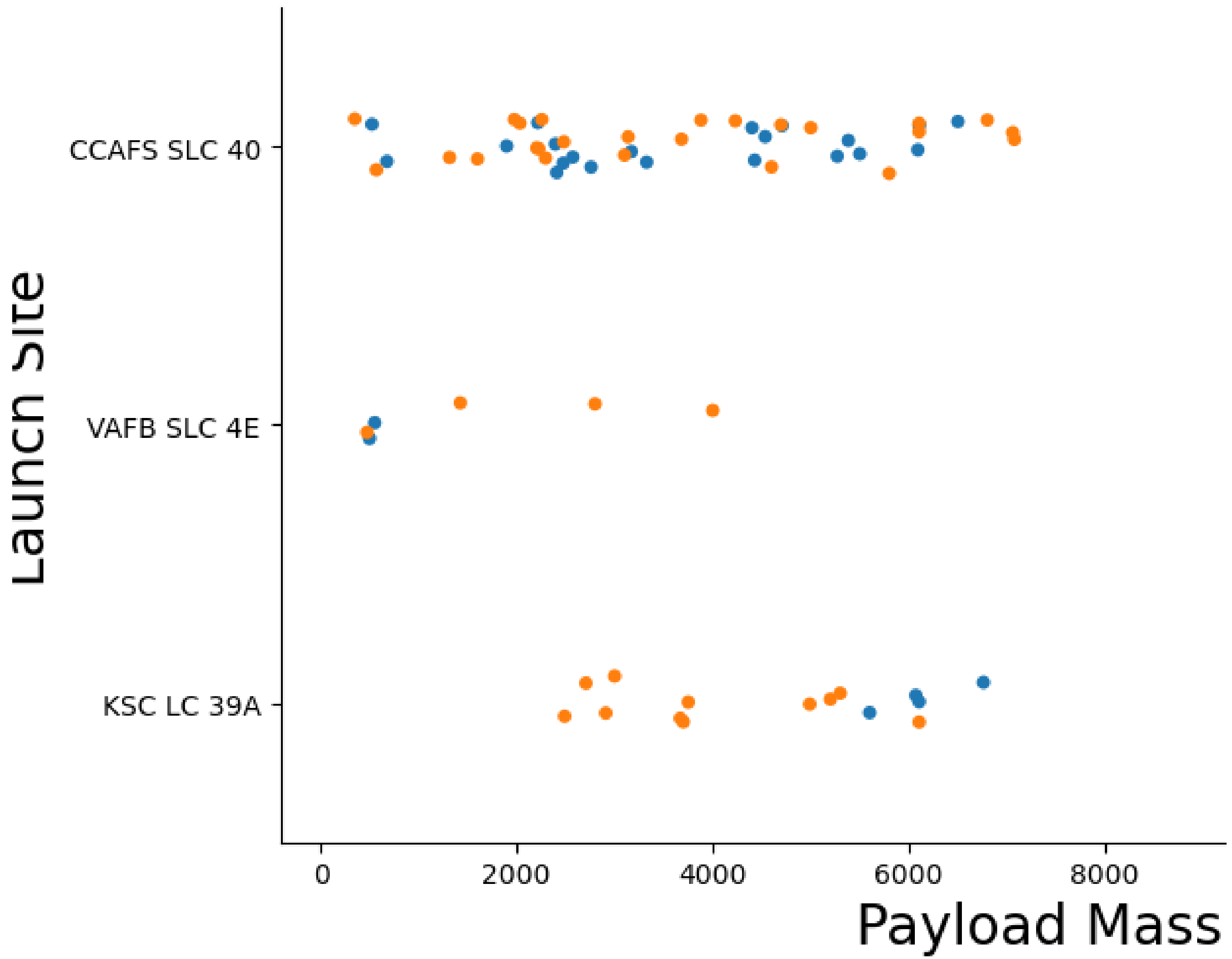
Flight Number vs. Launch Site

- Early flights had a lower percentage of success. failed in blue
- Later flights saw greater success rates. success is shown by orange.
- VAFB SLC 4E and KSC LC 39A had greater success rates than the CCAFS SLC 40 launch location, where approximately fifty percent of launches were made.
- New launches have a greater success rate, as we may deduce.



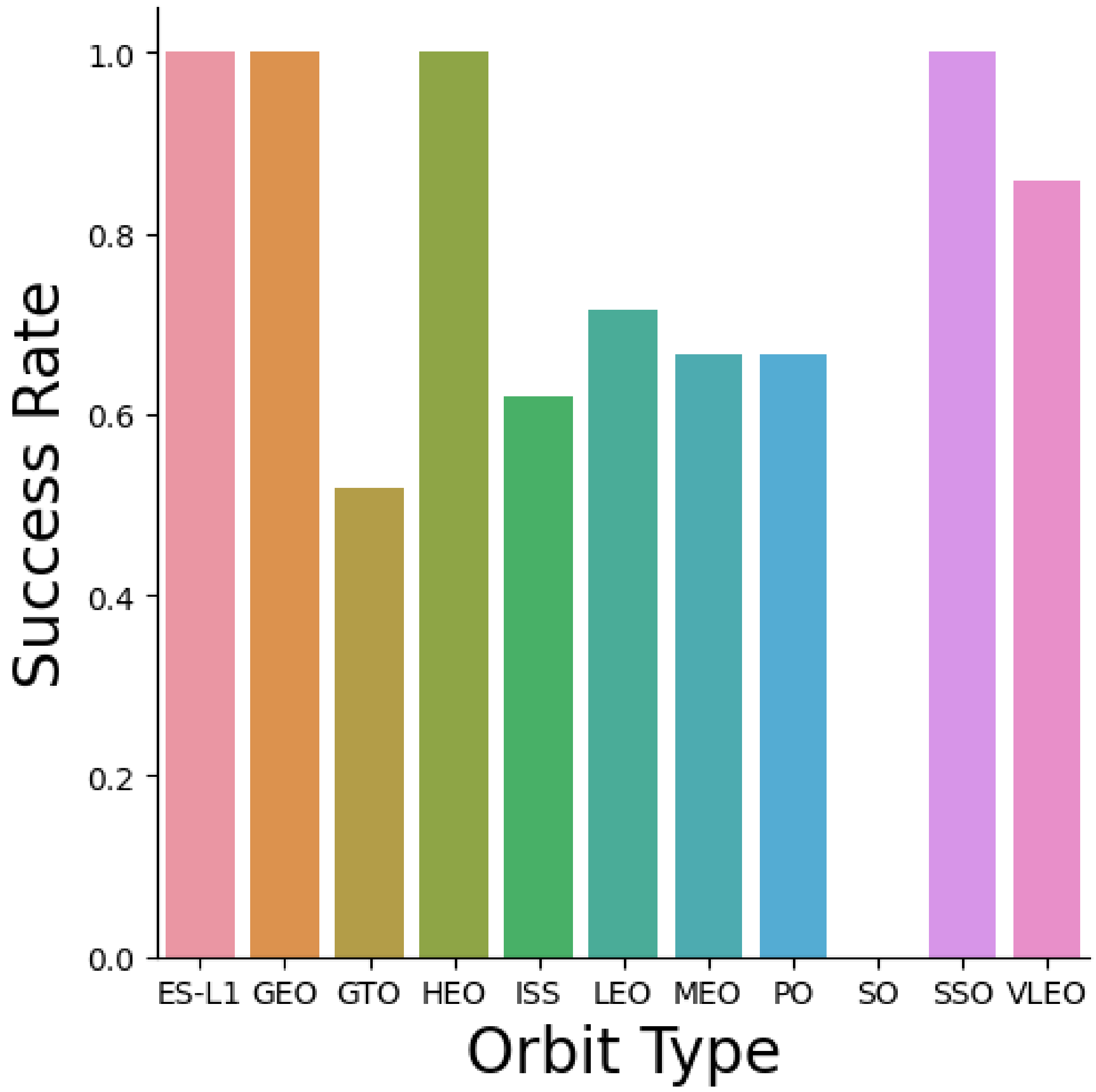
Payload vs. Launch Site

- Typically, the success rate increases with increasing payload mass (kg).
- Most launches with payloads of more than 7,000 kg were successful.
- For launches of less than 5,500 kg, KSC LC 39A has a success record of 100%.
- There hasn't been a launch at VAFB SKC 4E that weighs more than 10,000 kg.



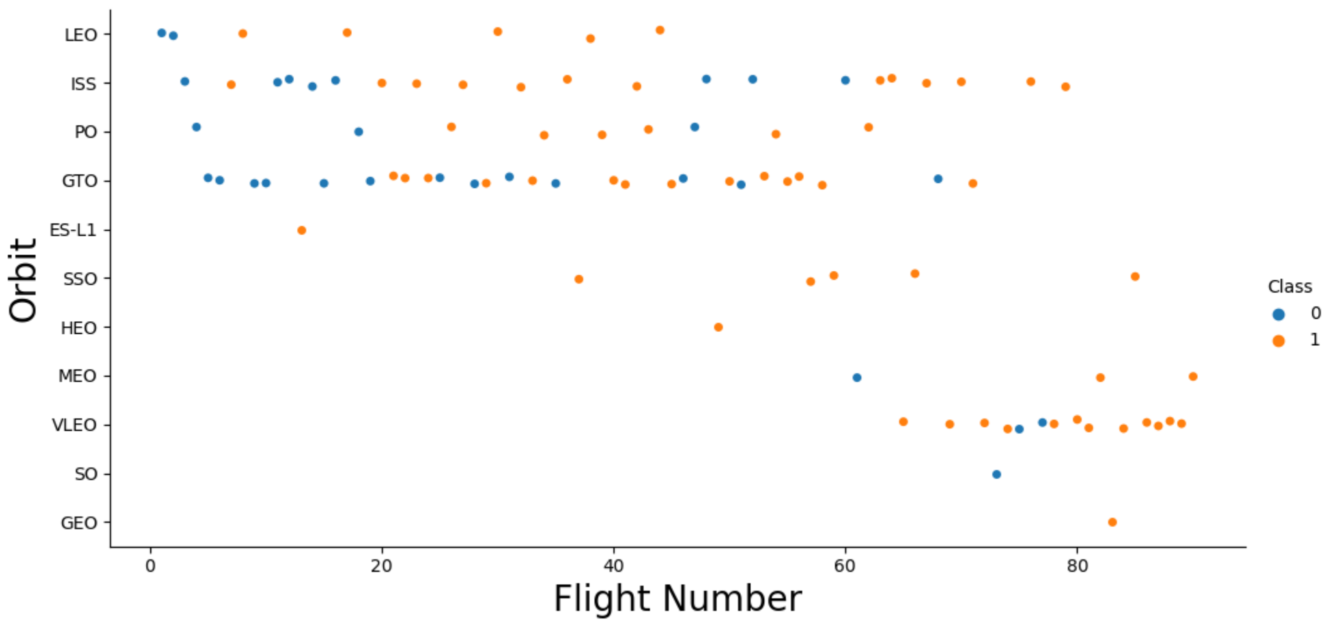
Success Rate vs. Orbit Type

- **100% Success Rate**: ES-L1, GEO, HEO and SSO
- **50%-80% Success Rate**: GTO, ISS, LEO, MEO, PO
- **0% Success Rate**: SO



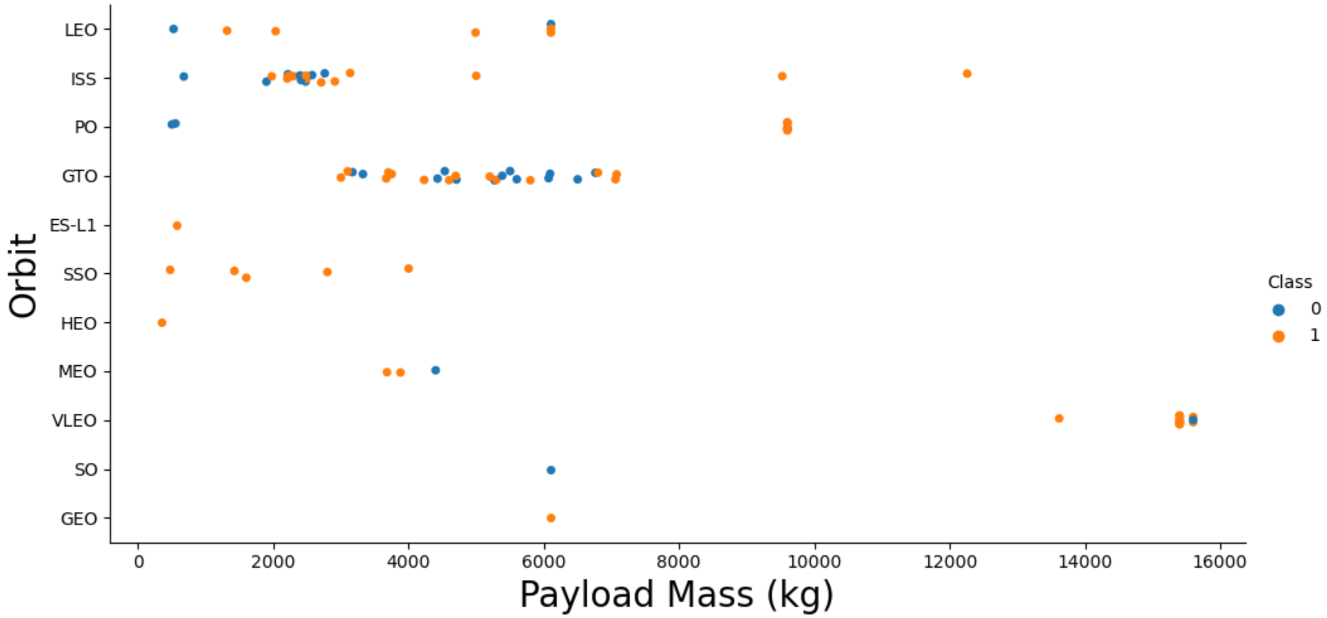
Flight Number vs. Orbit Type

- For each orbit, the success rate normally rises with the number of flights.
- this relationship is particularly clear for the LEO orbit.
- the GTO orbit, however, deviates from this pattern.



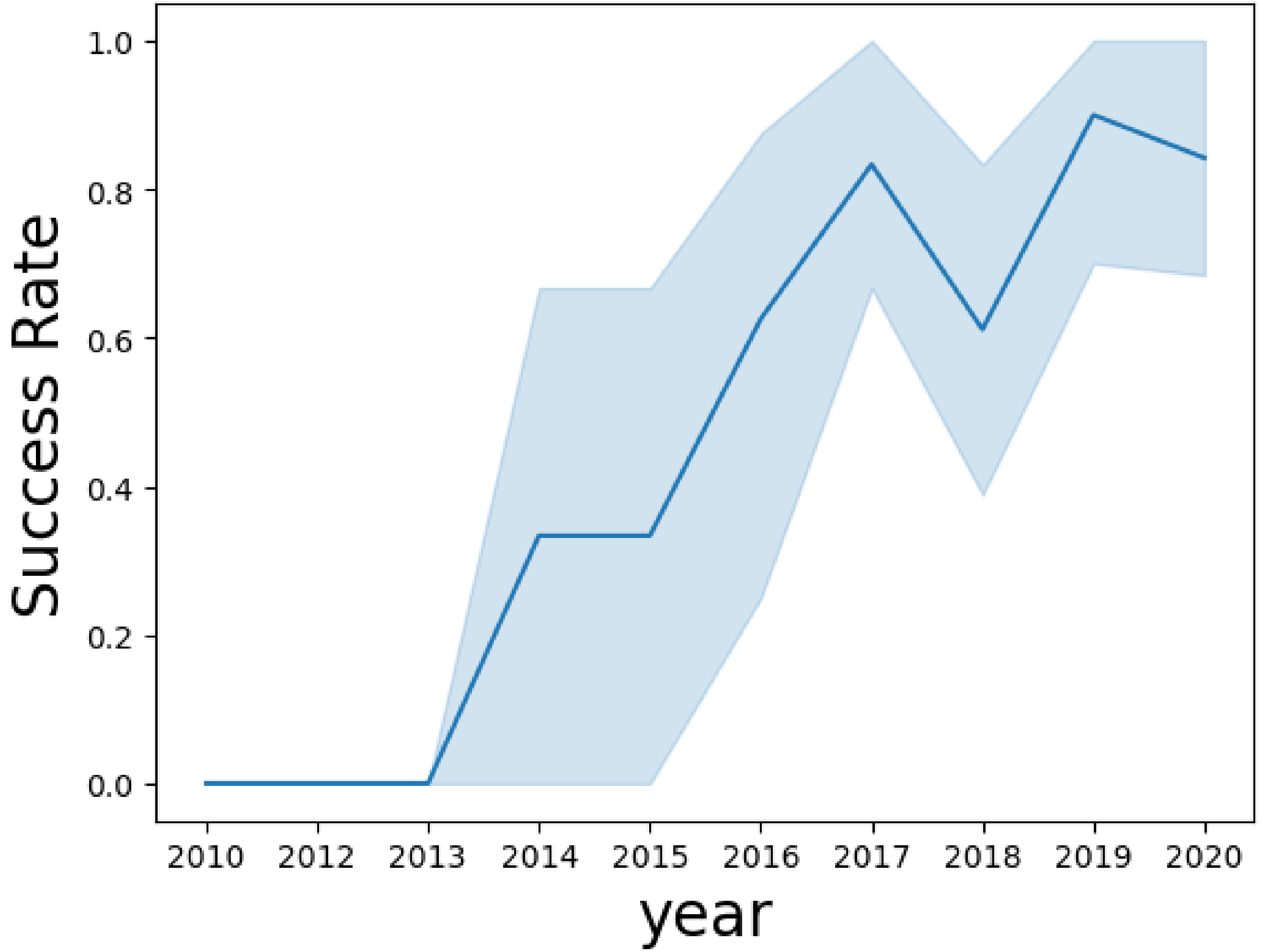
Payload vs. Orbit Type

- LEO, ISS, and PO orbits are preferable for heavy payloads.
- Mixed results may be achieved with bigger payloads in the GTO orbit.



Launch Success Yearly Trend

- The success rate:
- Increased from 2013 to 2017 and from 2018 to 2019
- Fell from 2017 to 2018 and from 2019 to 2020
- Overall increased since 2013.



All Launch Site Names

- Launch sites name, latitude and longitude are presented using the key word **DISTINCT** to show only unique launch sites from the SpaceX data.

	Launch Site	Lat	Long
0	CCAFS LC-40	28.562302	-80.577356
1	CCAFS SLC-40	28.563197	-80.576820
2	KSC LC-39A	28.573255	-80.646895
3	VAFB SLC-4E	34.632834	-120.610745

Launch Site Names Begin with 'CCA'

- Five of the records with launch site starting with CCA are displayed here. The records is gathered using a SQL query.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outc
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parach
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 (parach
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No atte
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No atte
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No atte

Total Payload Mass & Average Payload Mass by F9 v1.1

- NASA's boosters carried a total of 45,596 kg.
The calculation was done using SQL SUM function.

SUM(PAYLOAD_MASS__KG_)
45596

- The average weight of the booster version F9 v1.1 is 2,928.4 kg. The calculation was done using SQL AVG function.

AVG(PAYLOAD_MASS__KG_)
2928.4

Landing & Mission Information

- **First Successful Landing in Ground Pad was at**12/22/2015.

- **Successful Drone Ship Landing with Payload between 4000 and 6000 are** SCAT-14, JSCAT-16, SES-10, and SES-11 / EchoStar 105.

- **Total Number of Successful and Failure Mission Outcomes are** one Failure in Flight and hundred Success with one of the success as (payload status unclear).

MIN(DATE)
2015-12-22

Payload
JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105

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

Boosters Carried Maximum Payload

Using a subquery in the WHERE clause and the MAX() method, we were able to identify the booster that had carried the most payload. Twelve boosters were displayed.

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

2015 Launch Records

- For the year 2015, we filtered for failure landing results in drone ship, their booster versions, and launch site names using sets of the WHERE clause, LIKE, AND, and BETWEEN conditions.

month	Date	Booster_Version	Launch_Site	Landing_Outcome
10	2015-10-01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- To filter for landing events BETWEEN 2010-06-04 and 2010-03-20, we choose Landing events and the COUNT of landing events from the data.
- The landing results were categorized using the GROUP BY clause, and they were then put in decreasing order using the ORDER BY clause.

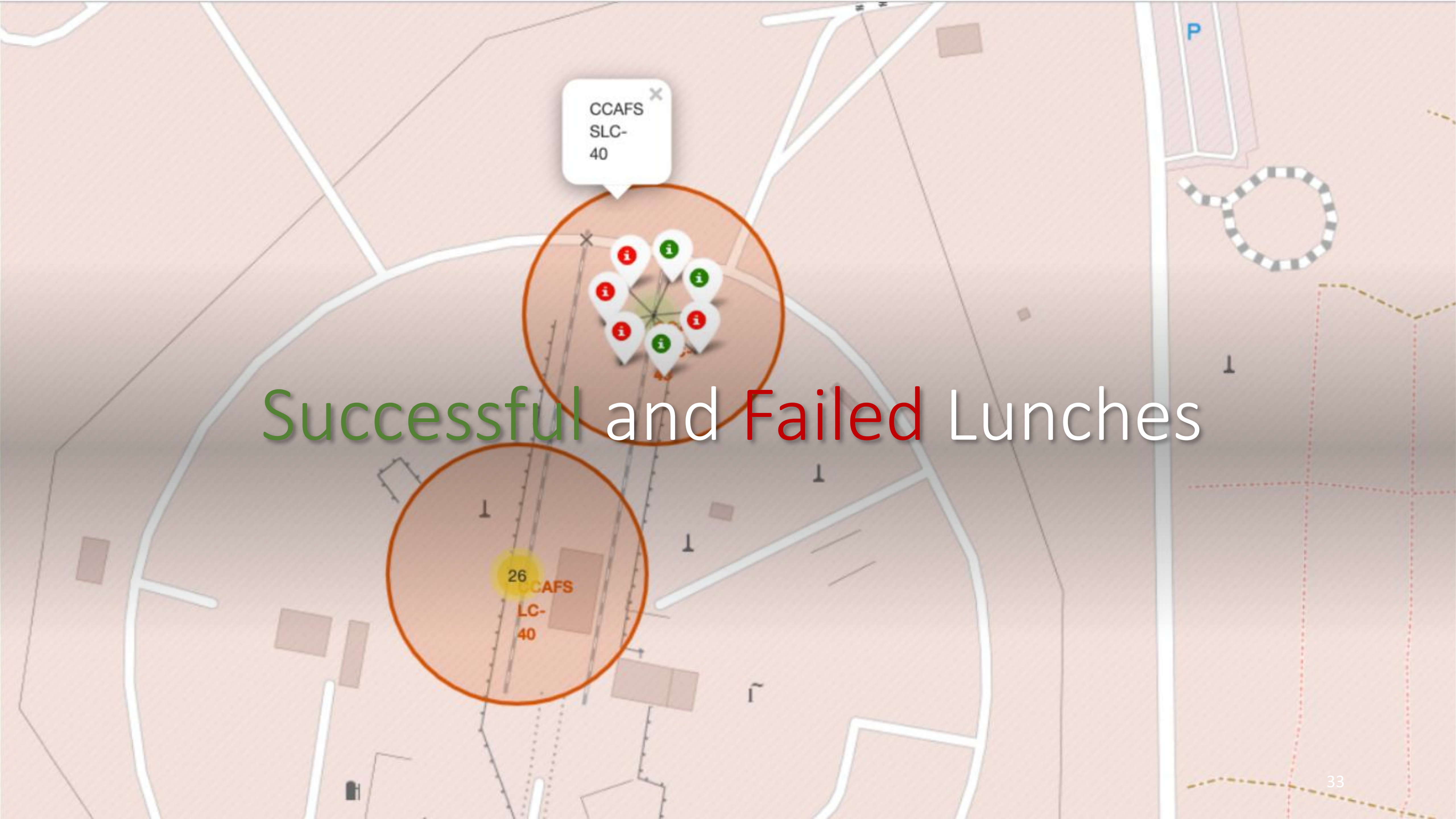
Landing_Outcome	count_outcomes
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

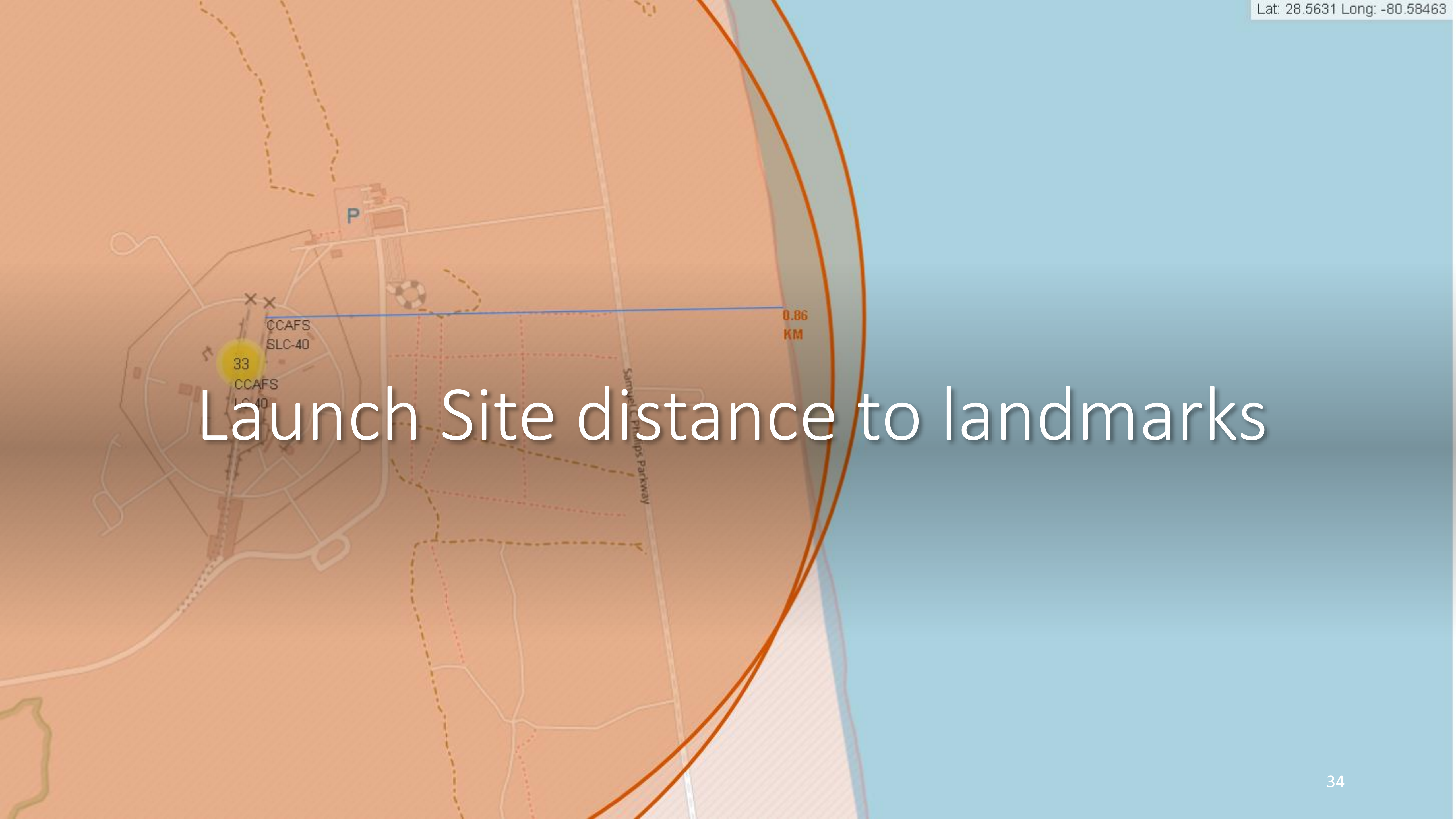
Launch Sites Proximities Analysis

Launch Sites Markers Globally

Successful and Failed Lunches



Launch Site distance to landmarks

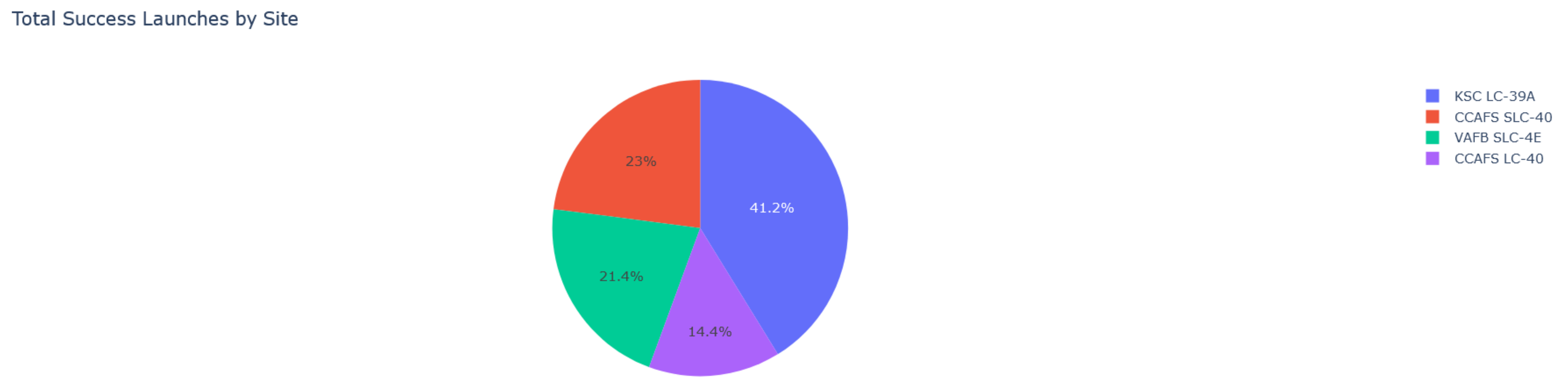


Section 4

Build a Dashboard with Plotly Dash

SpaceX Launch Records Dashboard

All Sites

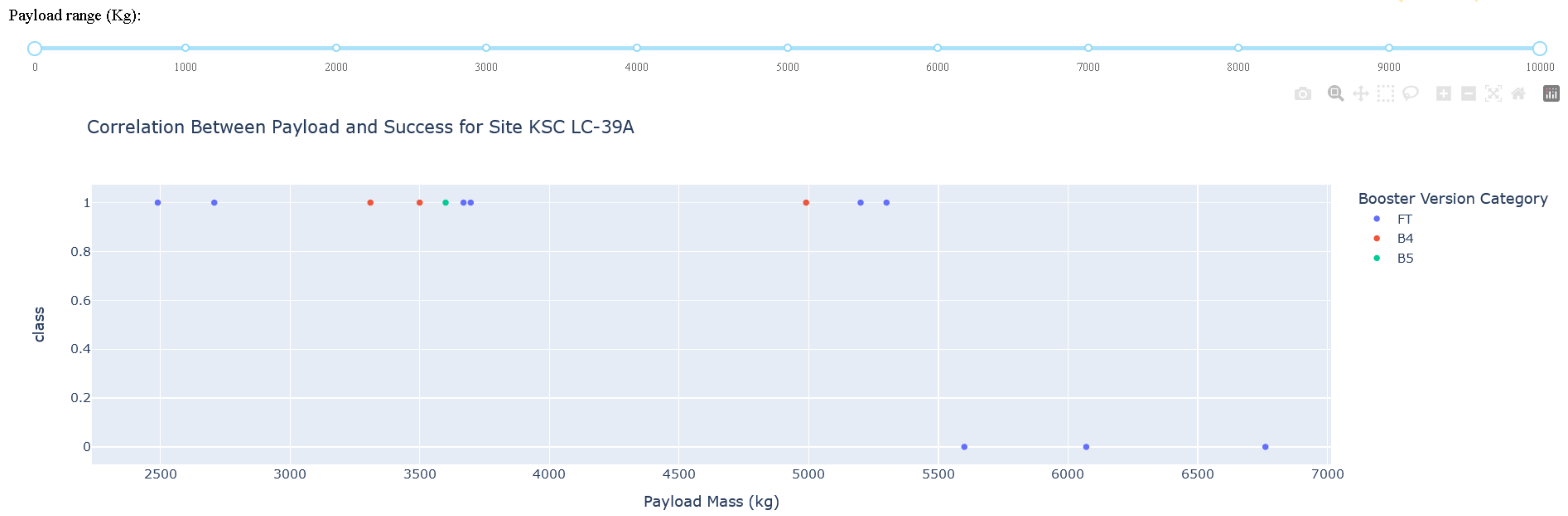


SpaceX Launch Records Dashboard

KSC LC-39A

Total Success Launches for Site KSC LC-39A

KSC LC-39A has the **highest success rate**



Payloads between2,000 kg and **5,000 kg** have the **highest success rate**

Section 5

Predictive Analysis (Classification)

Classification Accuracy

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.705882	0.800000
F1_Score	0.888889	0.888889	0.827586	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

- All the models performed about at the same level, had comparable scores, and were equally accurate. Most likely, this is caused by the limited dataset. When looking at best_score_, which is the average of all cv folds for a particular combination of the parameters.
- The Decision Tree model marginally beat the others.

```
models = {'KNeighbors':knn_cv.best_score_,
              'DecisionTree':tree_cv.best_score_,
              'LogisticRegression':logreg_cv.best_score_,
              'SupportVector': svm_cv.best_score_}

bestalgorithm = max(models, key=models.get)
print('Best model is', bestalgorithm,'with a score of', models[bestalgorithm])
if bestalgorithm == 'DecisionTree':
    print('Best params is :', tree_cv.best_params_)
if bestalgorithm == 'KNeighbors':
    print('Best params is :', knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best params is :', logreg_cv.best_params_)
if bestalgorithm == 'SupportVector':
    print('Best params is :', svm_cv.best_params_)
```

```
Best model is DecisionTree with a score of 0.8714285714285716
Best params is : {'criterion': 'gini', 'max_depth': 2, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 1
0, 'splitter': 'random'}
```

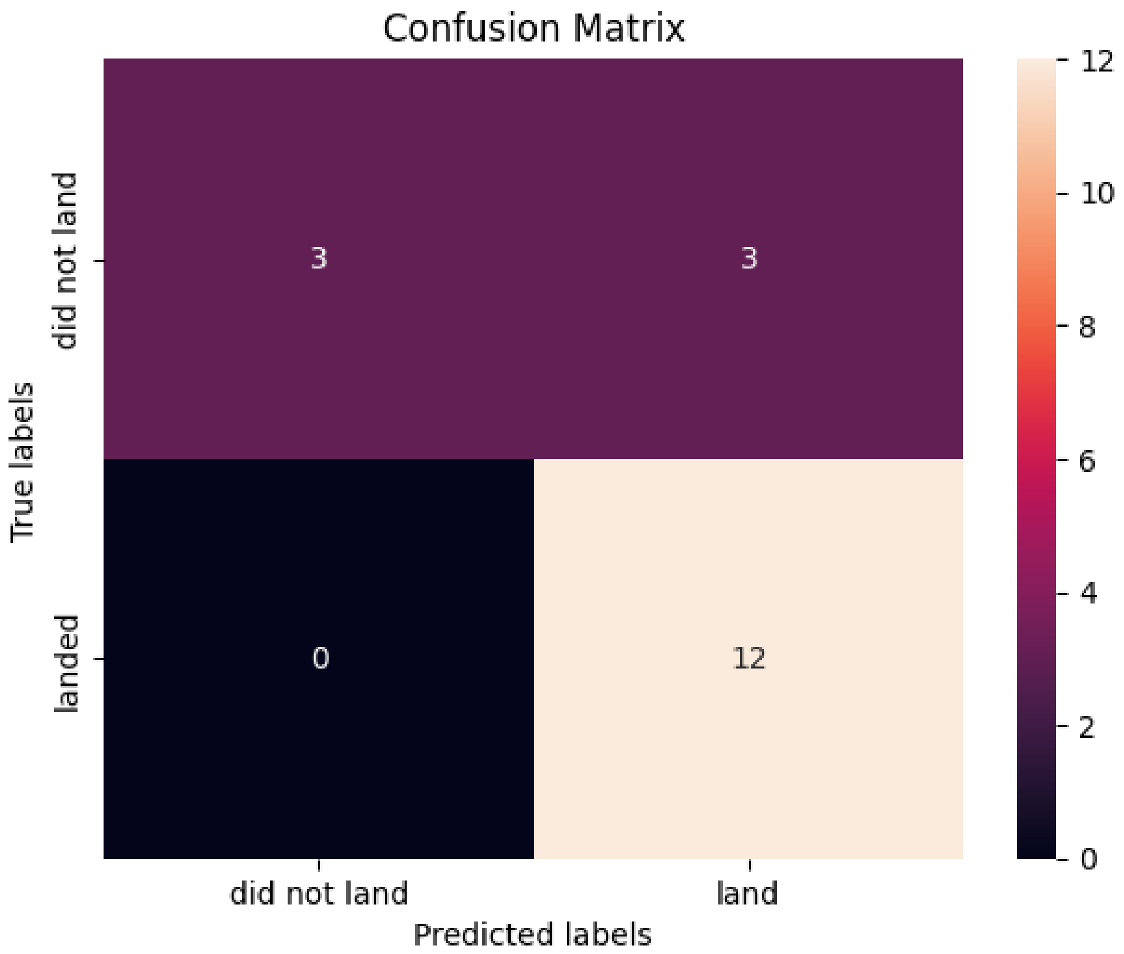
Confusion Matrix

The fact that there are false positives (Type 1 error) is not good

Confusion Matrix Outputs:

- 12 Ture positive
- 3 True negative
- **3 False positive**
- 0 False Negative

- **Precision**= TP / (TP + FP)12 / 15 = .80
- **Recall**= TP / (TP + FN)12 / 12 = 1
- **F1 Score**= 2 * (Precision * Recall) / (Precision + Recall)2 * (.8 * 1) / (.8 + 1) = .89
- **Accuracy**= (TP + TN) / (TP + TN + FP + FN) = .833



Conclusions

- Model Performance: The decision tree model marginally outperformed the other models on the test set.
- Equator: Due to the earth's spinning speed and the additional natural boost, it provides, most launch sites are located close to the equator, which reduces the need for additional fuel and boosters.
- Among launch locations, KSC LC 39A has the best success rate.
- Shore: All launch sites are near the shore.
- Launch Success: Increases over time. has a 100% success rate for launches weighing less than 5,500 kg; ES L1, GEO, HEO, and SSO orbits have 100% success rates; and the bigger the payload mass (kg), the higher the success rate across all launch locations.

Appendix

- Like to the GitHub codes and datasets:

https://github.com/ayesha2938/SpaceY.git

Thank you!