IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ed
2024-02-27

Outline

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

Executive Summary

Methodologies

- Data Collection: Using the SpaceX REST API, launch data were collected.
- Data Wrangling: Data was filtered and cleaned to for the analysis and modelling steps.
- Exploratory Analysis: Using SQL, Pandas and Matplotlib, data was explored and analyzed.
- Data Visualization: Using Plotly Dash, Folium, dashboards were built in order to analyze launch records interactively.
- Predictive Analysis (Classification): SVM, Decision Trees, Logistic Regression and KNN was used and compared.

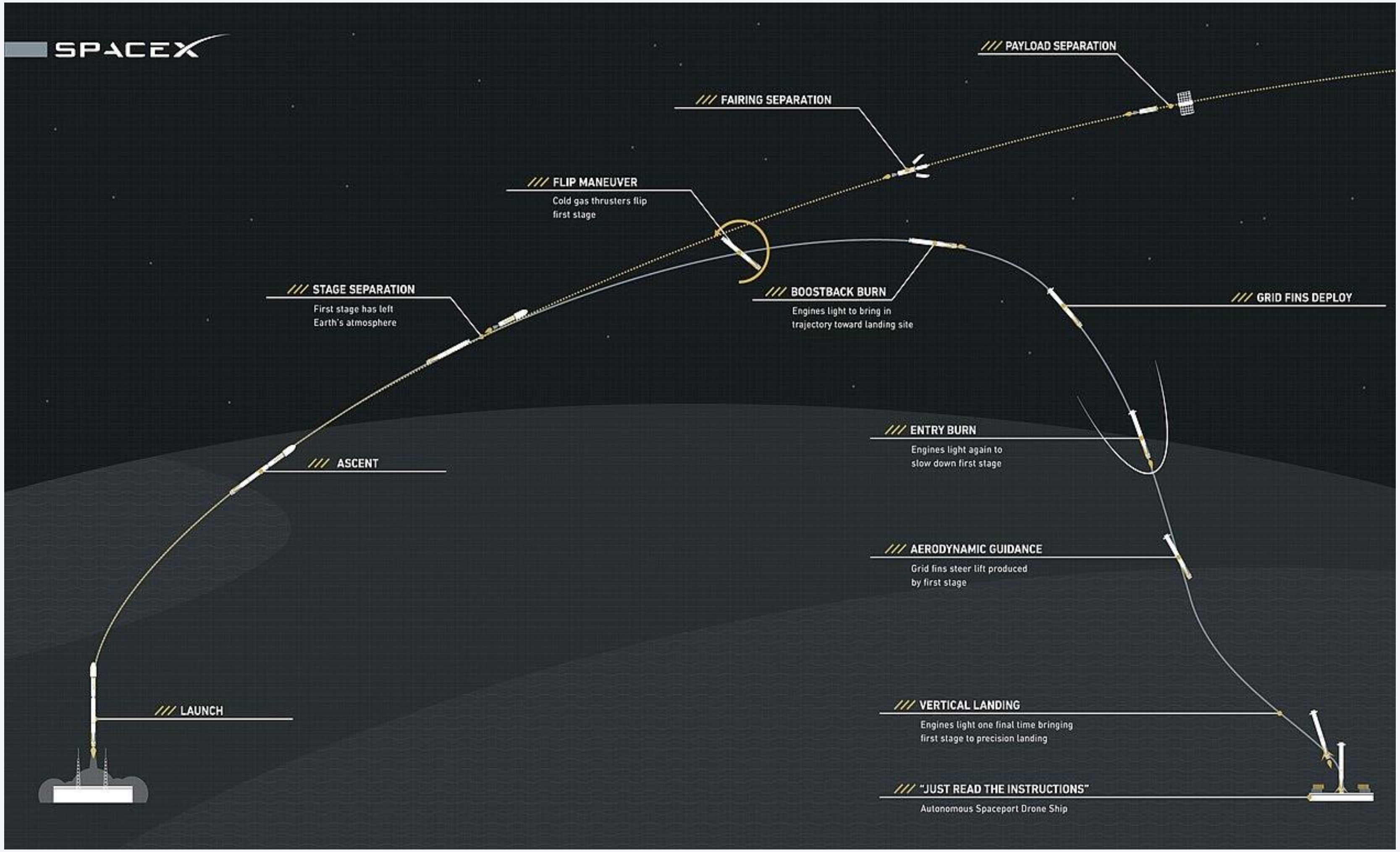
Results

- Space X Falcon 9 launch success has significantly increased from 2013 to 2023; KSC LC-39A having the highest success rate.
- Based on the classification results, each model performed adequately, however, the Decision Tree slightly outperformed the others.

Introduction

- SpaceX was founded in 2002. SpaceX is specifically known for its ability to reuse the first stage of its rockets which has a significant impact on the overall cost of a launch.
- Space X Falcon 9 rockets have a successful track record, launching payloads into space with high reliability and efficiency.
- This Applied data science Capstone project for Space X Falcon 9 rockets is aimed at predicting the successful outcomes of the rocket launches for the first stage using classification models. The objective is to predict whether or not the first stage would be reused

SpaceX reusable first stage technology



Section 1

Methodology

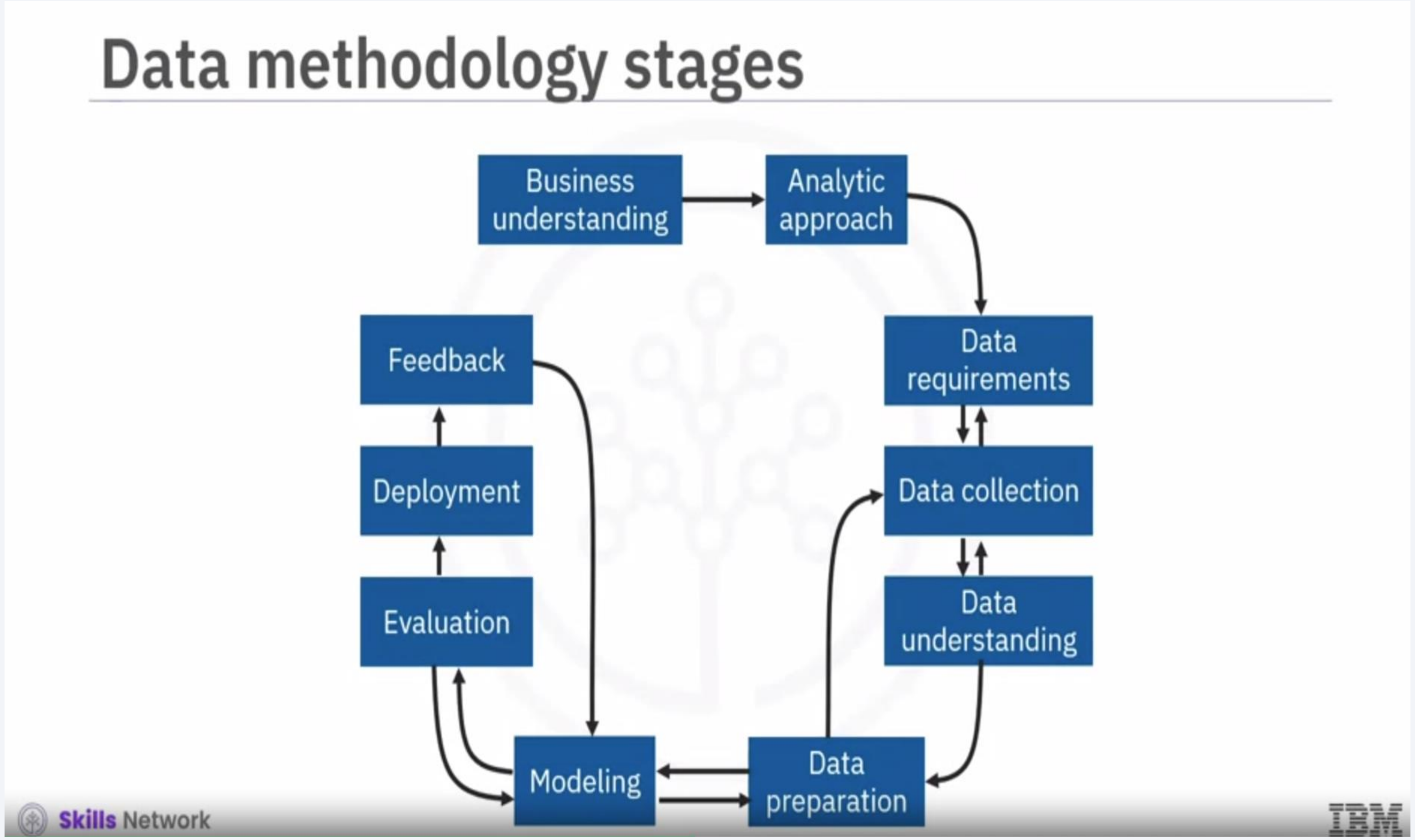
Methodology

Executive Summary

- Data collection methodology:
  - Launch data was collected using the SpaceX REST API.
- Perform data wrangling
  - Data was processed by filtering and cleaning the data.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Determined the best performing method, using training and test data.

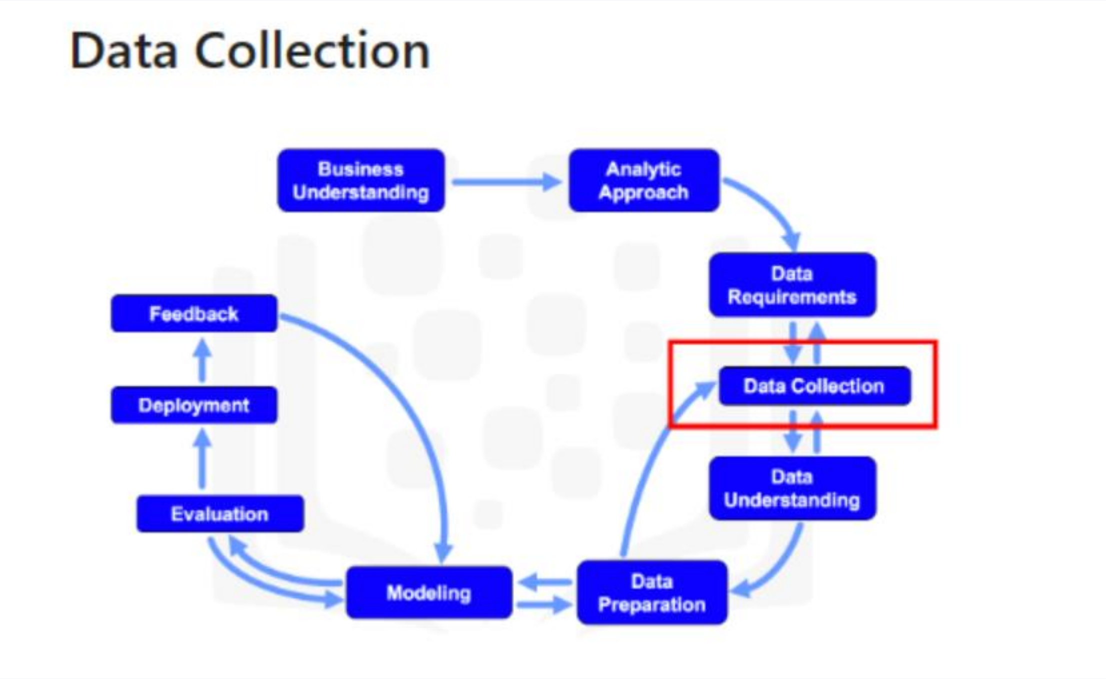
Methodology

- Based on John Rollins Methodology



Data Collection

- Using the SpaceX REST API, launch and landing data and specifications including landing outcomes were collected. The steps describing how the data was collected for both SpaceX API and web scraping are in the next two slides.



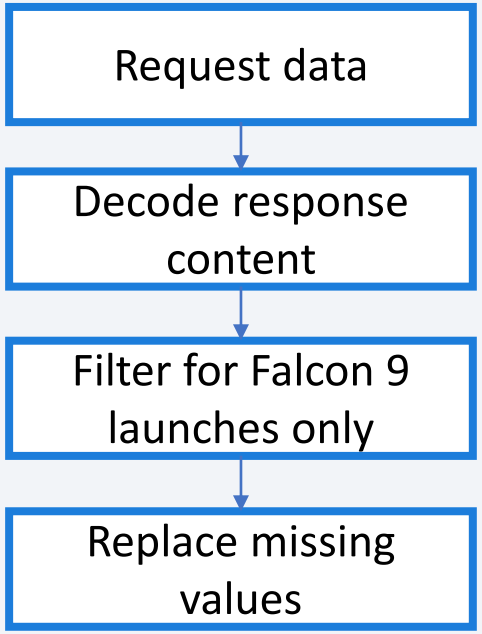
Data Collection – SpaceX API

Objectives:

- **Request to the SpaceX API**
- **Clean the requested data**

Steps followed to collect the data:

1. Request and parse the SpaceX launch data.
2. Decode response content.
3. Filter the data frame to only include Falcon 9 launches.
4. Replace all missing values.

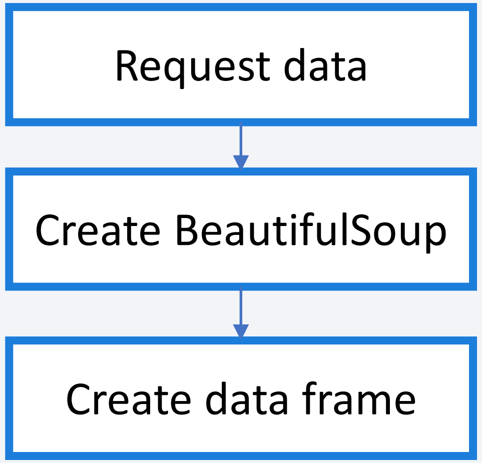


- Link to GitHub URL: https://github.com/Elsbell/Capstone-Project-Falcon-9/blob/main/Spacex%20Data%20Collection.ipynb

Data Collection - Scraping

The following steps provide an overview of the web scraping process:

1. Request data from Wikipedia.
2. Create a BeautifulSoup object from the HTML response.
3. Extract all column names/variables from HTML table header.
4. Parse the launch HTML tables to create a data frame.



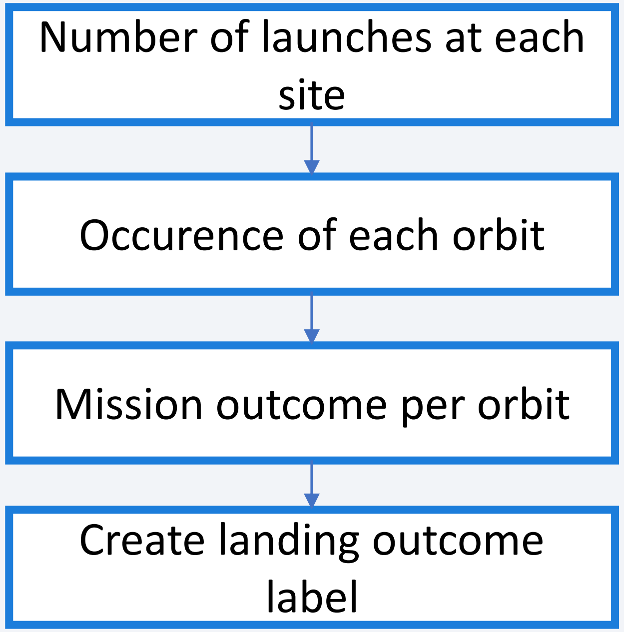
- Link to GitHub URL: https://github.com/Elsbell/Capstone-Project-Falcon-9/blob/main/Spacex%20Webscraping.ipynb

Data Wrangling

The following steps describe how the data was processed:

1. Calculate the number of launches at each site.
2. Calculate the number of each orbit.
3. Calculate the number of mission outcome per orbit.
4. Create a landing outcome label.

- Link to GitHub URL: https://github.com/Elsbell/Capstone-Project-Falcon-9/blob/main/Spacex%20Data%20Wrangling.ipynb



EDA with Data Visualization

The following charts were plotted:

- Scatter plots (3) to show the relationships between flight numbers, payload mass, launch site variables and launch outcomes.
- Bar chat was used to show the relationship between the success rate and orbit type (Plot: Orbit Type vs Success Rate)
- Line chart was used to obtain the average launch success trend. (Plot: Years vs Success Rate)

- Link to GitHub URL: https://github.com/Elsbell/Capstone-Project-Falcon-9/blob/main/Spacex%20EDA%20Data%20Visualization.ipynb

EDA with SQL

Summary of the SQL queries performed:

1. Display unique launch site names.
2. Display launch sites that start with CCA.
3. Display total payload mass carried by boosters launched by NASA (CRS).
4. Display average payload mass carried by booster version F9 v1.1.
5. List first successful landing outcome on ground pad date.

- Link to GitHub URL: https://github.com/Elsbell/Capstone-Project-Falcon-9/blob/main/Spacex%20EDA%20SQL%20Notebook.ipynb

EDA with SQL (continuation)

6. List names of boosters that have success in drone ship and have payload mass between 4000 and 6000 lbs.
7. List total number of successful and failure mission outcomes.
8. List names of boosters which have carried the maximum payload mass.
9. List records such as month names, landing outcomes in drone ship, launch sites for 2015 months.
10. Rank count of landing outcomes between 2010-06-04 and 2017-03-20, in descending order.

- Link to GitHub URL: https://github.com/Elsbell/Capstone-Project-Falcon-9/blob/main/Spacex%20EDA%20SQL%20Notebook.ipynb

Build an Interactive Map with Folium

The main tasks were:

•TASK 1: Mark all launch sites on a map

•TASK 2: Mark the success/failed launches for each site on the map

•TASK 3: Calculate the distances between a launch site to its proximities

The following list summarizes the map objects that were added to the folium map:

- Highlight the area of a specific coordinate with a text label.
- Mark specific launch locations.
- Simplify the map by clustering the same coordinates.
- Find coordinates of any point by hovering with a mouse.
- Draw Lines to show the distances from the launch sites to particular infrastructure.

- Link to GitHub URL: https://github.com/Elsbell/Capstone-Project-Falcon-9/blob/main/Spacex%20Launch%20Sites%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

The following list summaries the plots, graphs and interactions that were added to the dashboard:

1. Dropdown list to select the launch sites.
2. Pie chart showing the total successful launches count for selected launch sites.
3. Slider for the payload range.
4. Scatter chart showing the relationship between payload and launch success.

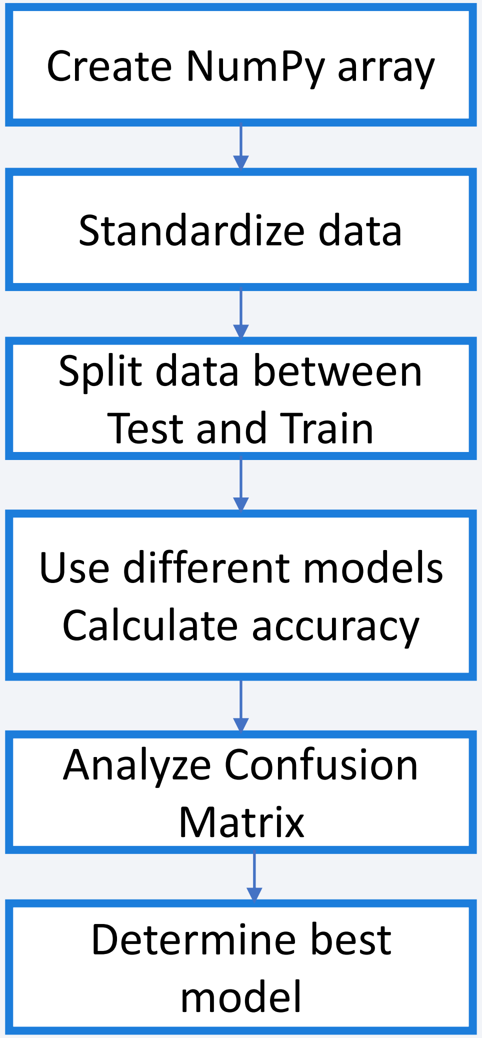
- Link to GitHub URL: https://github.com/Elsbell/Capstone-Project-Falcon-9/blob/main/spacex_dash_app.py.py

Predictive Analysis (Classification)

The following steps summarizes how the best performing classification model was built, improved, and found:

1. Create an NumPy array.
2. Standardize the data.
3. Split the data into training and test data.
4. Calculate the test data accuracy.
5. Analyze the confusion matrix.
6. Repeat for steps four to six for each model (i.e., logistic regression, support vector machine (SVM), decision tree, k nearest neighbors (KNN)) to determine the best performing method.

- Link to GitHub URL: https://github.com/Elsbell/Capstone-Project-Falcon-9/blob/main/Spacex%20Machine%20Learning%20Prediction.ipynb



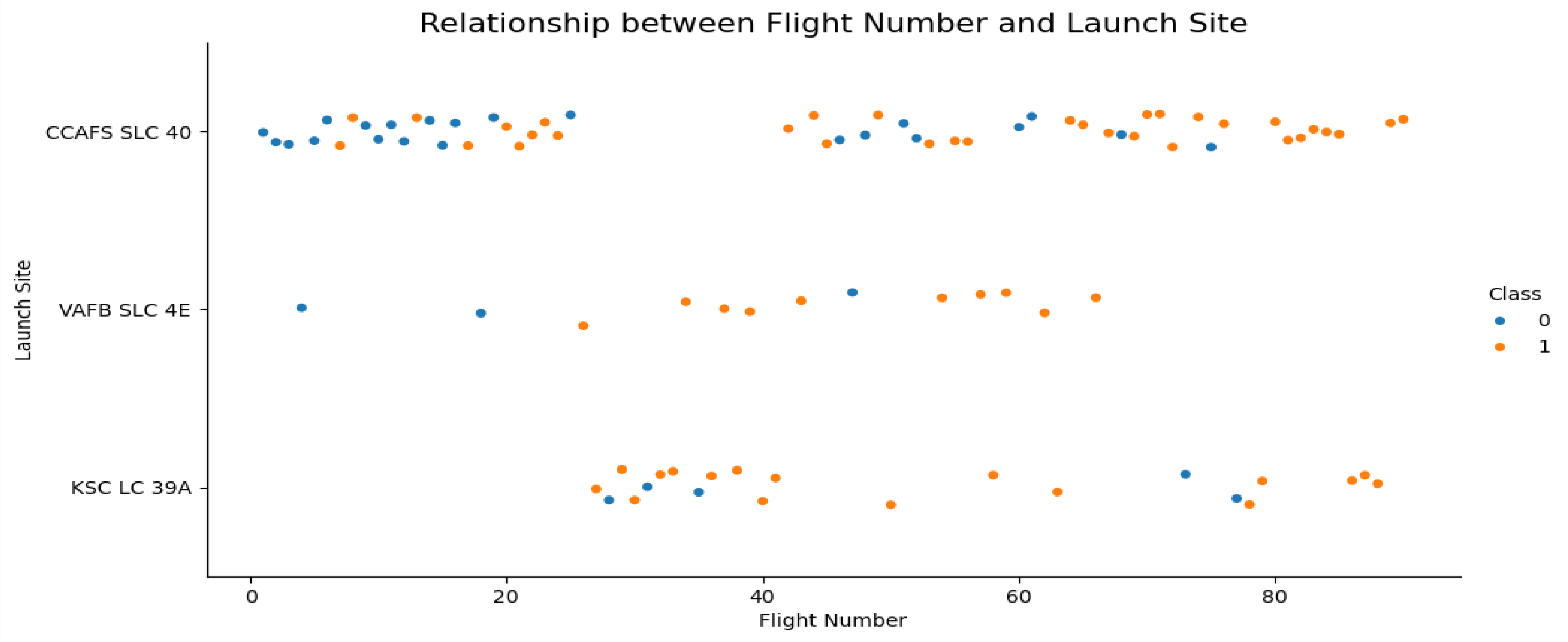
Results

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

Section 2

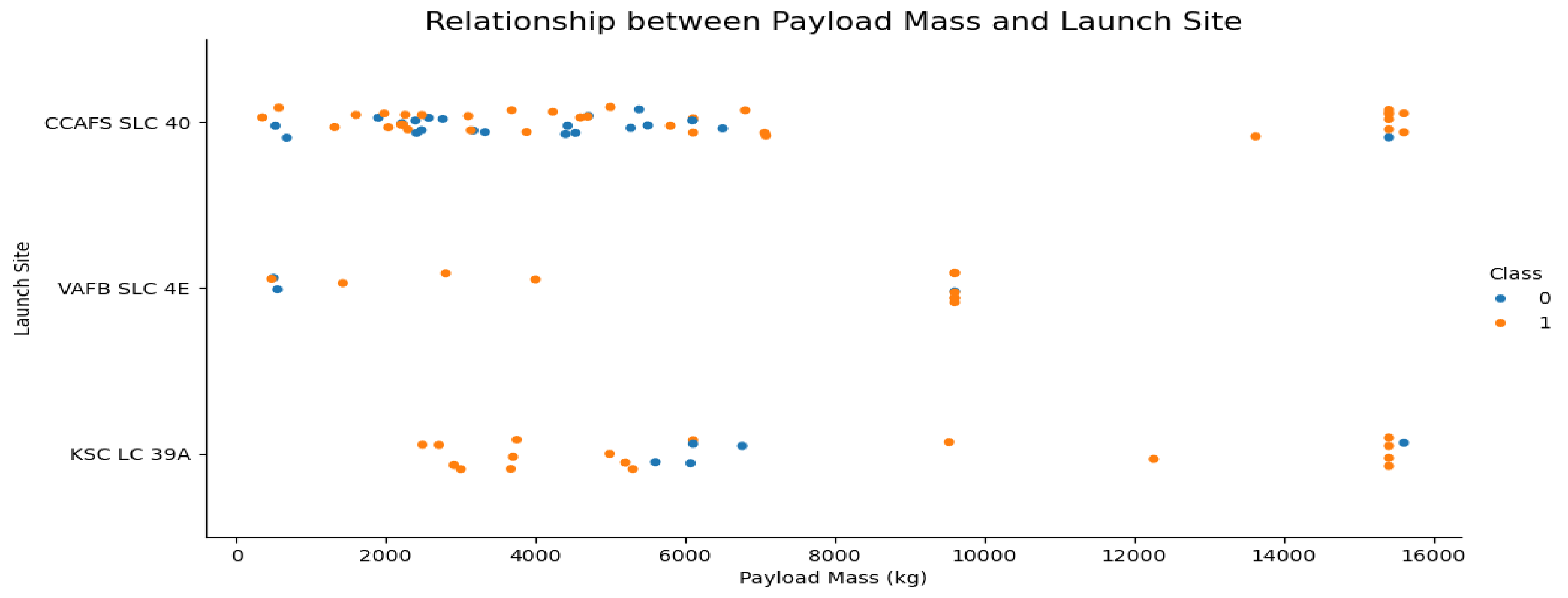
Insights drawn from EDA

Flight Number vs. Launch Site



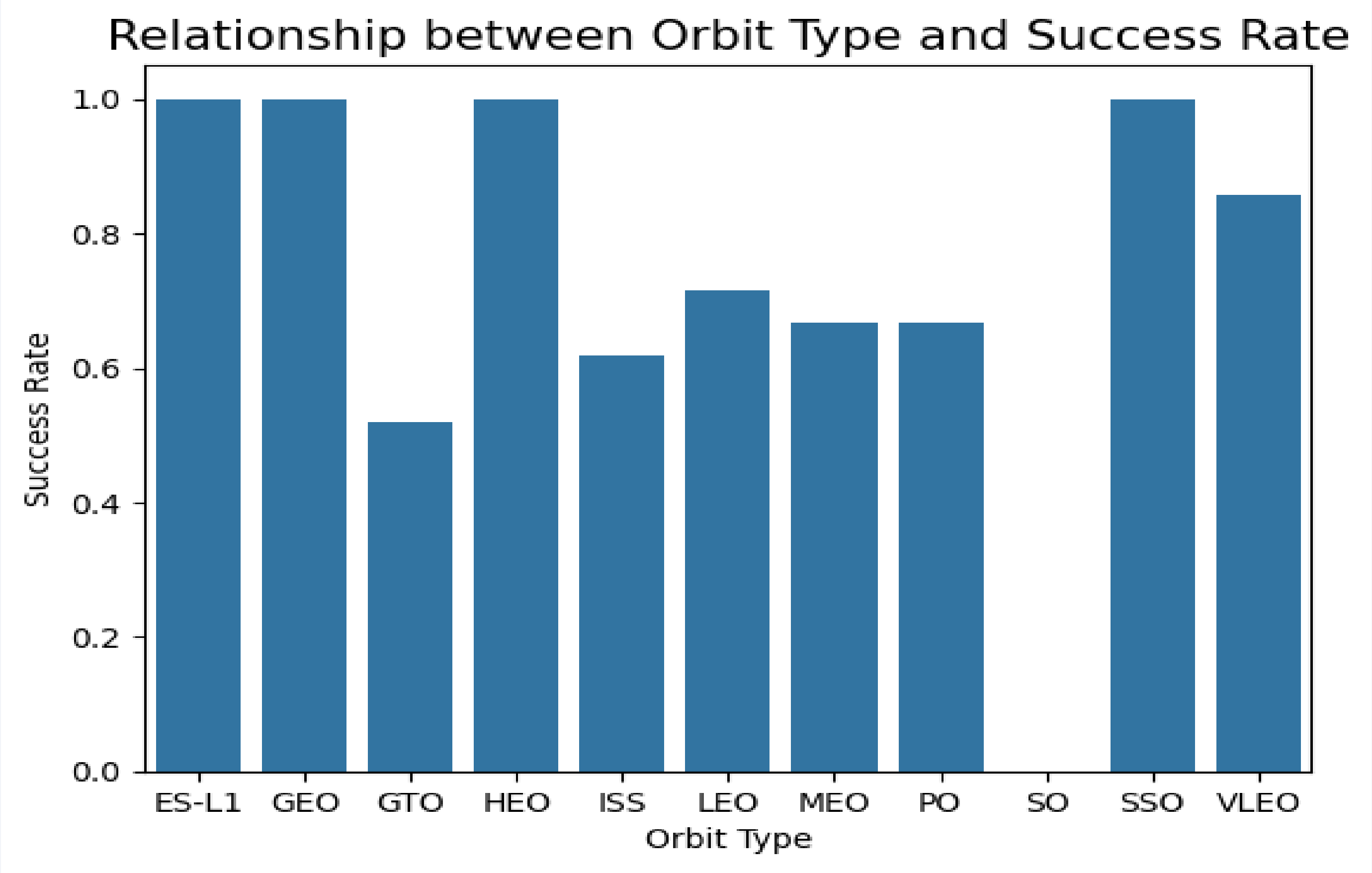
- CCAFS SLC 40 has launched the highest number of rockets compared to the other sites while KSC LC 39A showed a higher success rate.
- KSC LC 39A seems to be the most recent launching site, which could explain its success rate.

Payload vs. Launch Site



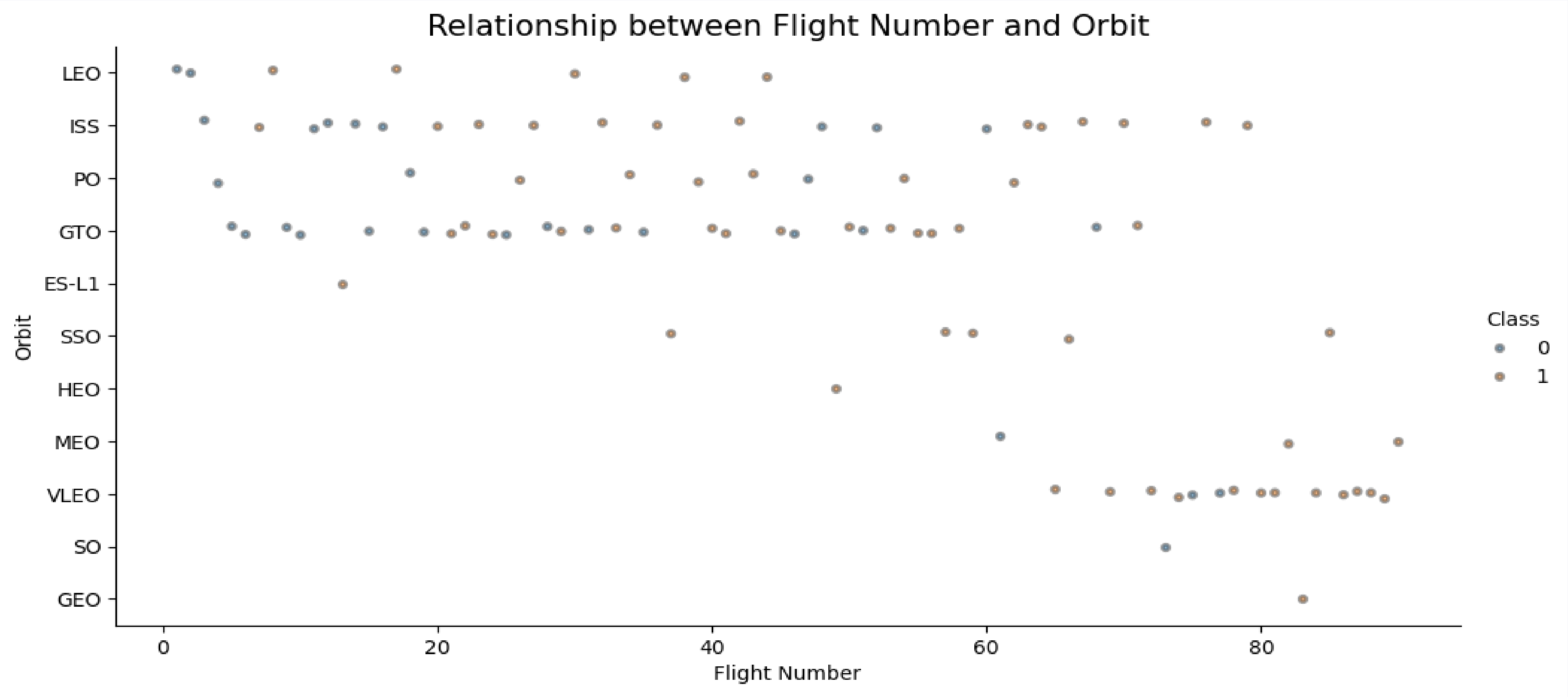
- CCAFS SLC 40 is the busiest launch site.
- Heavy payload mass greater than 10000kg are launched at CCAFS SLC 40 or at KSC LC 39A.
- Large majority of the rockets launched have a payload mass of less than 9000kg.

Success Rate vs. Orbit Type



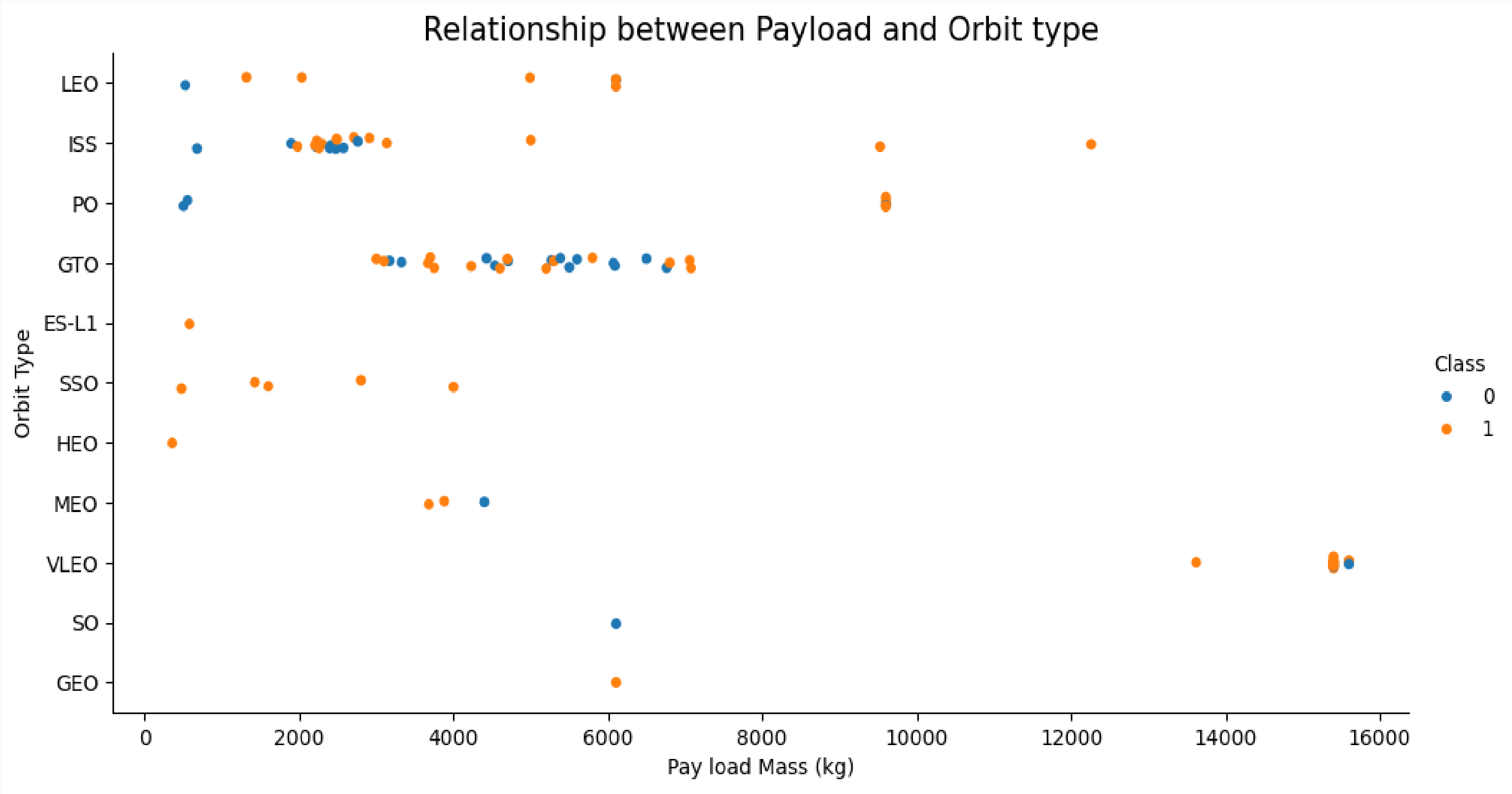
- GTO which is located at 35,786 km from the earth surface, has the lowest success rate of success. But other very high orbits such as GEO and HEO performed better; however there were only one launch of each.

Flight Number vs. Orbit Type



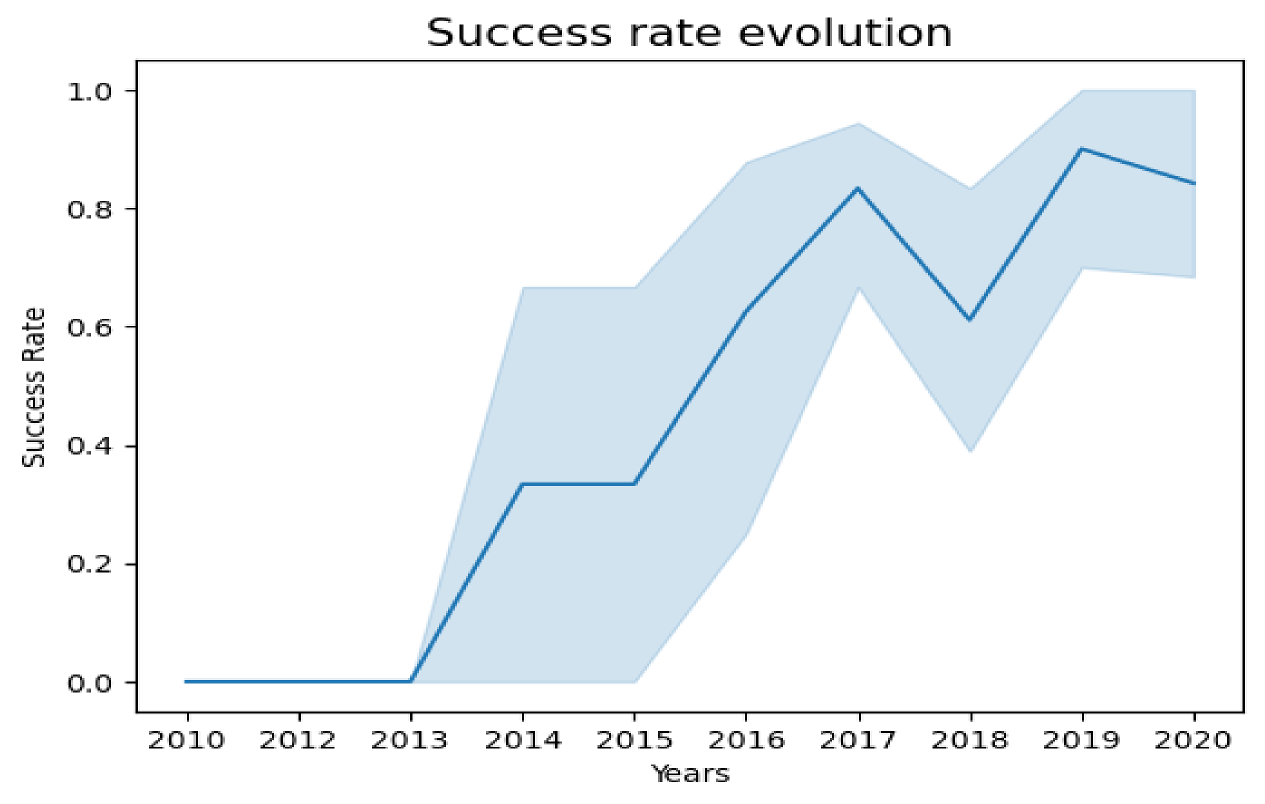
- This slide explains the low success rate of the GTO when compared specifically to the other high orbit launches (GEO and HEO). There were 27 GTO flights vs only one flight for GEO and HEO.

Payload vs. Orbit Type



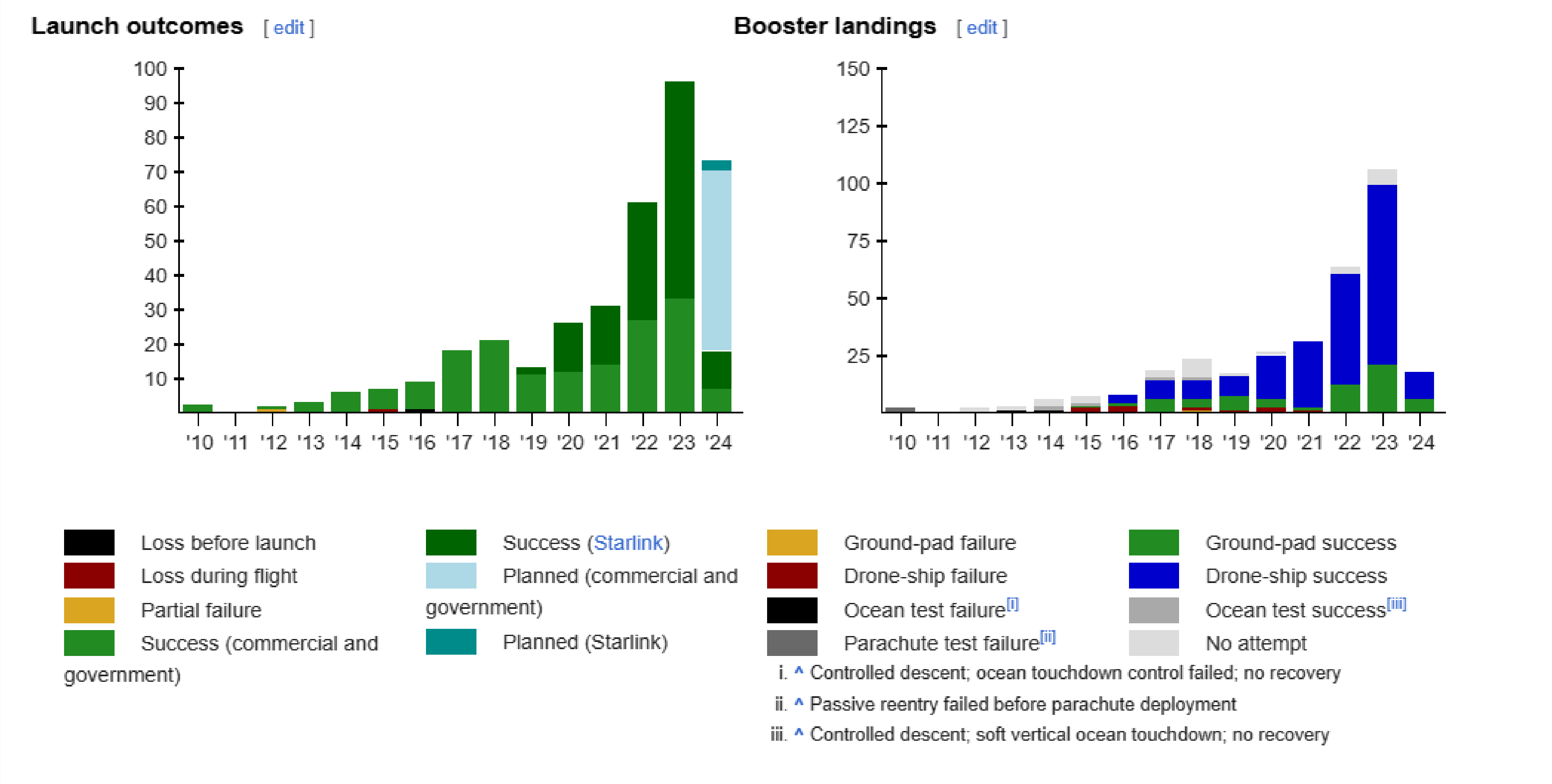
- Heavy payloads are better with LEO, ISS and PO orbits.
- Very high orbits (GEO, GTO and HEO) involve light payloads (below 7000 kg).

Launch Success Yearly Trend



- Steep learning curve between 2013 and 2017 and then a stabilization around a success rate around 0,8.

Launch Success Yearly Trend



- Up to date graphs showing accelerated improvement beyond 2017.

All Launch Site Names

Display the names of the unique launch sites in the space mission

```
%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

- Use Magic SQL command % sql.
- To find the Unique Launch Sites, the command SELECT DISTINCT was used.

Launch Site Names Begin with 'CCA'

Display 5 records where launch sites begin with the string 'CCA'

```
%sql SELECT * FROM spacextbl WHERE launch_site LIKE 'CCA%' LIMIT 50
```

* 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

- Use Magic SQL command % sql.
- Use the command SELECT* with the parameters WHERE and LIKE `CCA%`.
- LIMIT 5 keyword to display only 5 records.

Total Payload Mass

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql SELECT SUM(payload_mass__kg_) FROM spacextbl WHERE Customer LIKE '%CRS%'
```

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

SUM(payload_mass__kg_)
48213

- Use Magic SQL command % sql.
- Use the command SELECT SUM with the parameters WHERE and LIKE `%CRS%`.

Average Payload Mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1

```
%sql SELECT AVG(payload_mass__kg_) FROM spacextbl WHERE Booster_Version LIKE '%F9 v1.1%'
```

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

AVG(payload_mass__kg_)
2534.6666666666665

- Use Magic SQL command % sql.
- Use the command SELECT AVG with the parameters WHERE and LIKE `%F9 v1.1%`.

First Successful Ground Landing Date

List the date when the first succesful landing outcome in ground pad was achieved.

Hint:Use min function

```
%sql SELECT MIN(date) FROM spacextbl WHERE mission_outcome LIKE 'Success'
```

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

MIN(date)
2010-06-04

- Use Magic SQL command % sql.
- Use the command SELECT MIN(date) with the parameters WHERE and LIKE `Success`.

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%sql SELECT Booster_Version FROM spacextbl WHERE Landing_Outcome LIKE '%Success (drone ship)%' AND payload_mass__kg_ BETWEEN 4000 and 6000
```

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

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

- Use Magic SQL command % sql.
- Use the command SELECT FROM with the parameters WHERE and BETWEEN

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
%sql SELECT COUNT(Customer), Mission_Outcome FROM spacextbl GROUP BY Mission_Outcome;
```

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

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

- Use Magic SQL command % sql.
- Use the command SELECT COUNT with the parameters FROM and GROUP BY.

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

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

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

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

- Use Magic SQL command % sql.
- Use a double Query with the command SELECT and WHERE and a second Query with SELECT MAX and FROM.

2015 Launch Records

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%sql SELECT date, landing_outcome, booster_version, launch_site FROM spacextbl WHERE landing_outcome LIKE 'Failure (drone
```

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

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

- Use Magic SQL command % sql.
- Use the command SELECT with the parameters FROM, WHERE and LIKE.

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 counts FROM spacextbl WHERE date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY
%sql SELECT landing_outcome, COUNT(*) AS counts FROM spacextbl WHERE date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY 
```

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

Landing_Outcome	counts
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

- Use Magic SQL command % sql.
- Use the commands SELECT, COUNT(*) with the parameters WHERE, BETWEEN and GROUP BY.

Section 3

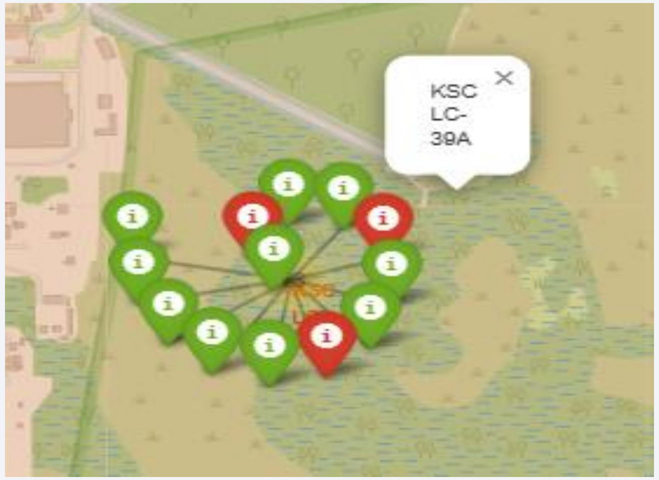
Launch Sites Proximities Analysis

Launch Sites

Four launch sites all located near the coast, 3 in Florida and one in California:

- **VAFB SLC-4E (California)**
- **KSC LC-39A (Florida)**
- **CCAFS LC-40 (Florida)**
- **CCAFS SLC-40 (Florida)**

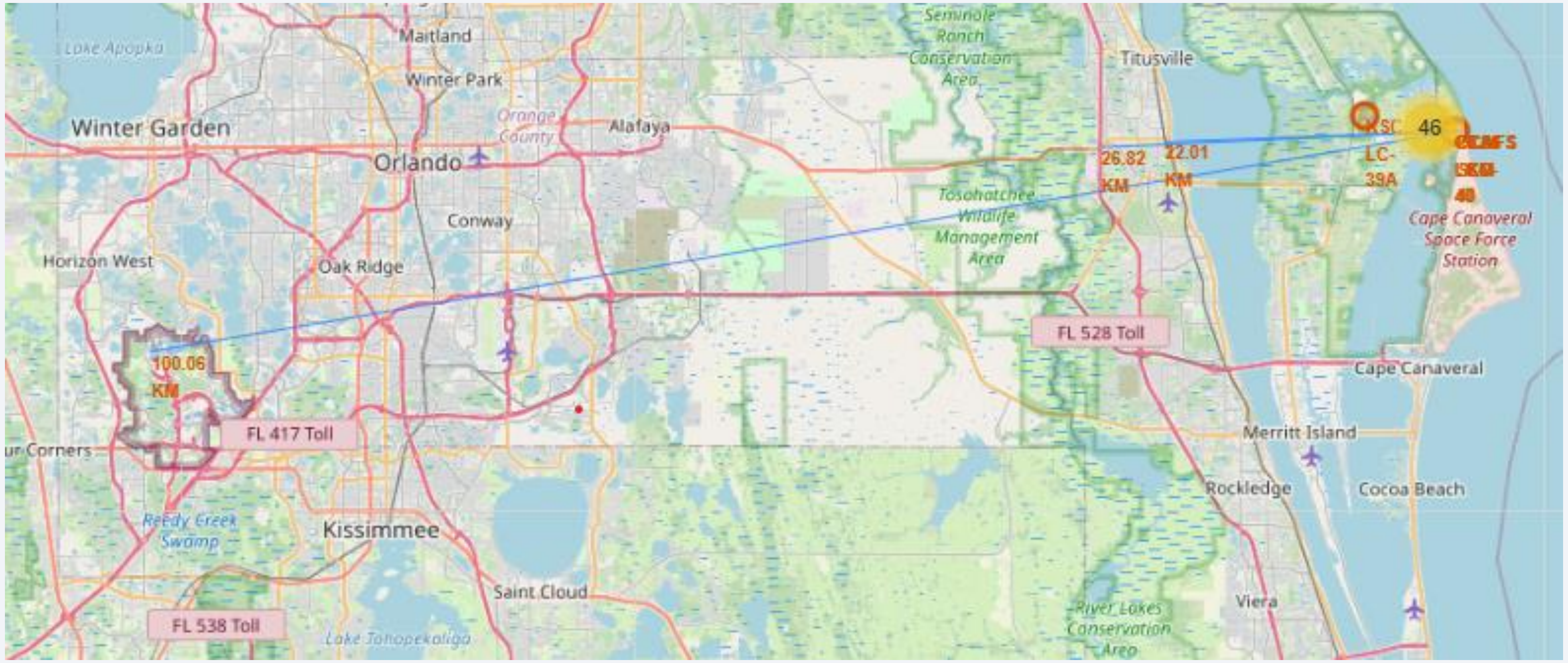
Launch Outcomes



Outcomes:

- Green markers for successful launches
- Red markers for unsuccessful launches

Launch site distance from important infrastructure



Distance CCAFS_SLC40 to Florida East Coast Railway: 22.0 km

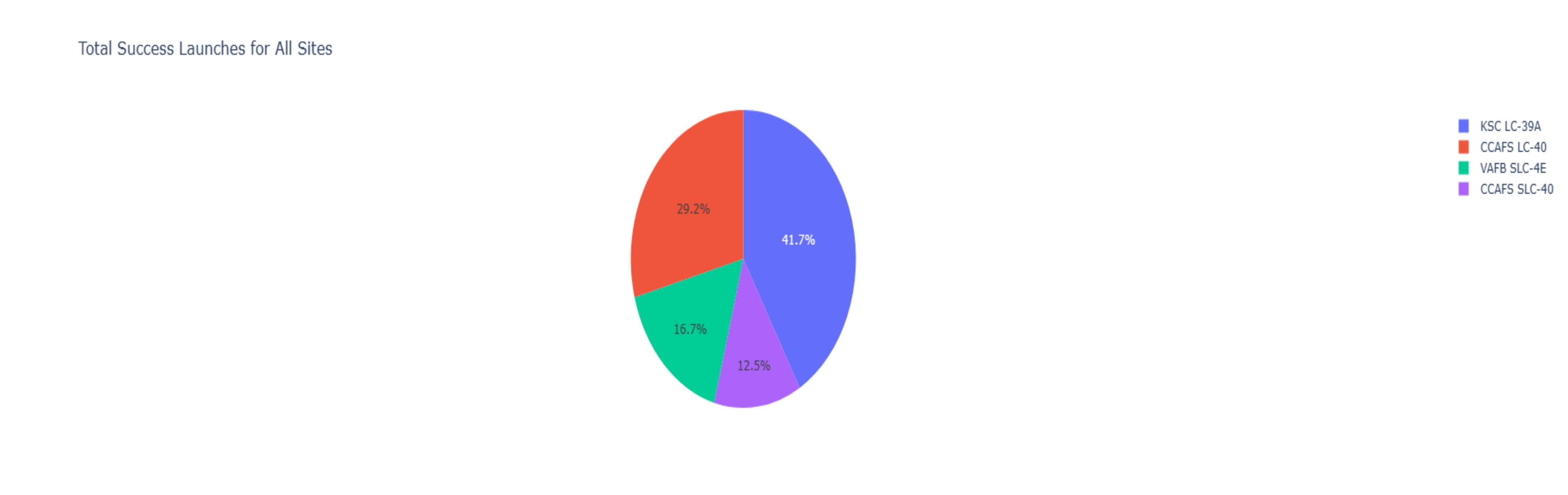
Distance CCAFS_SLC40 to Highway I95: 26.8 km

Distance CCAFS_SLC40 to Walt Disney Orlando: 100.1 km

Section 4

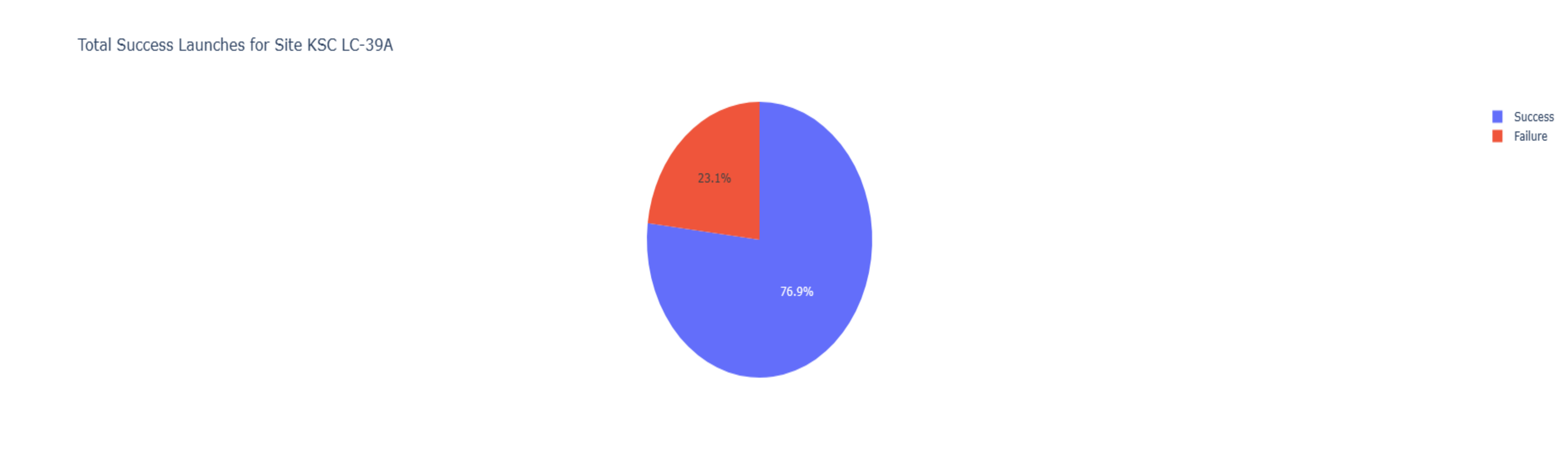
Build a Dashboard with Plotly Dash

Launch success count for all sites



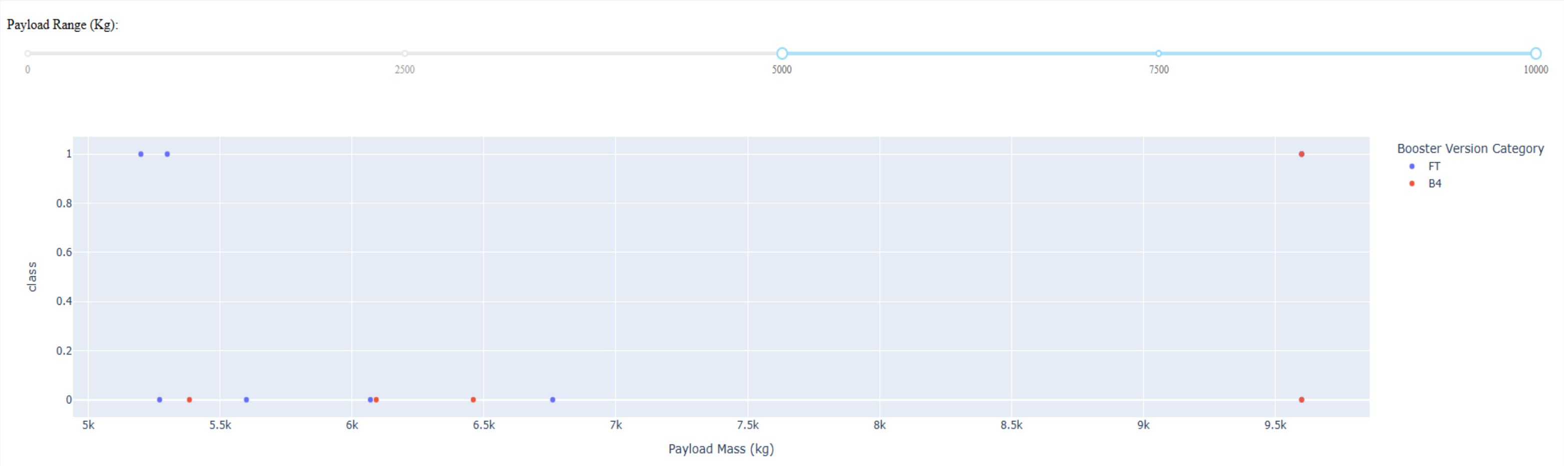
- Most of the successful launches were obtained at KSC LC-39A (41.7%). It is the most recent launching site, and has most likely benefited from the experience gained on other sites.

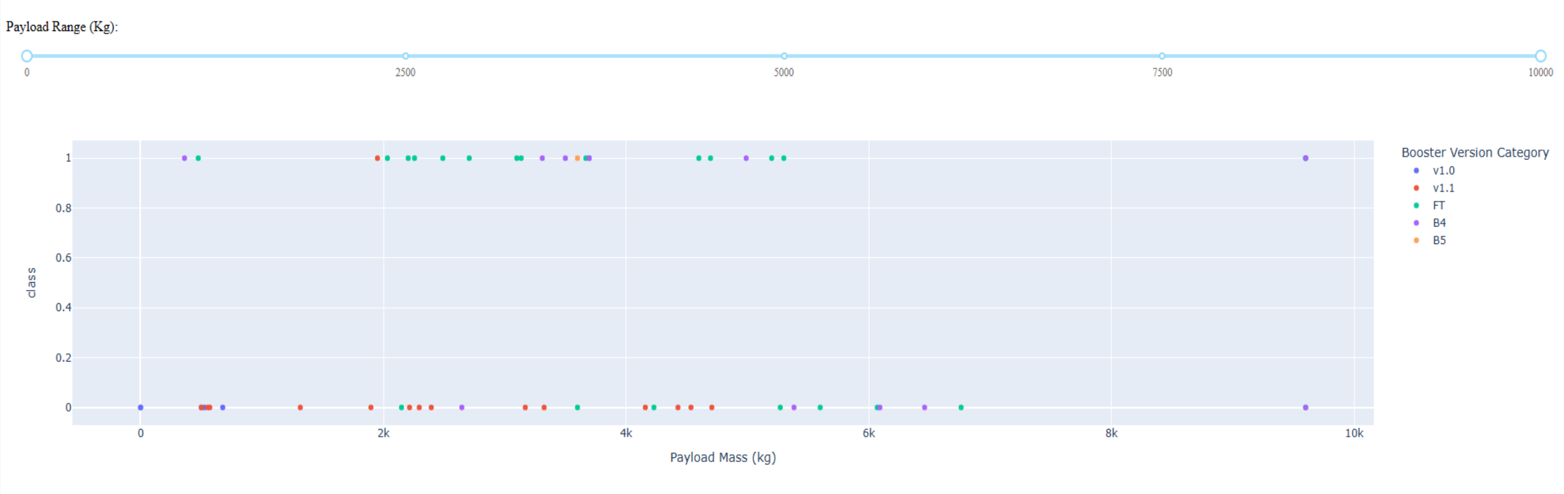
Launch site with the highest success rate



- KSC LC-39A is the site with the highest success rate (76.9%). It is the most recent launching site, and has most likely benefited from the experience gained on other sites.

Launch Outcome for Different Payload



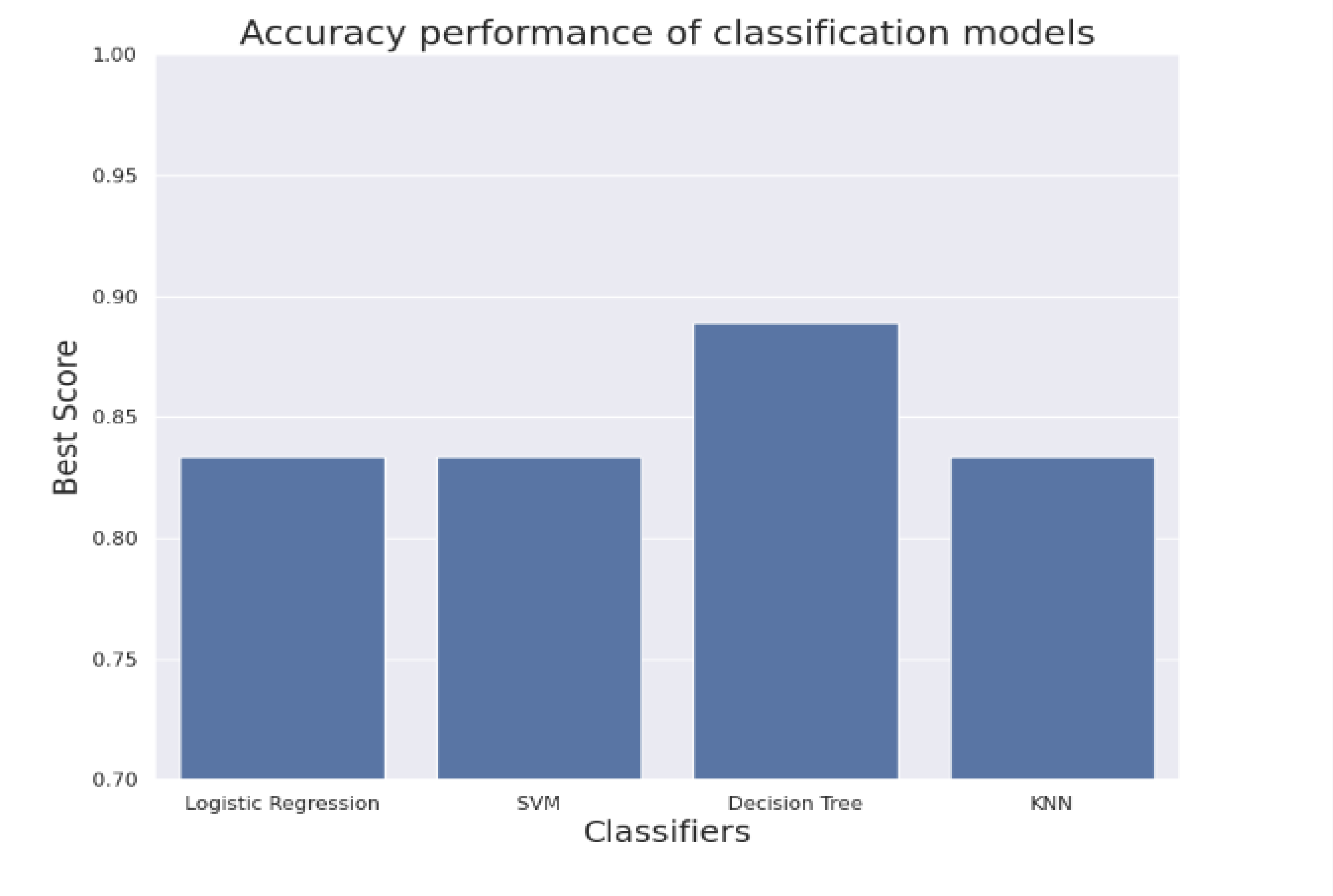


- Payloads lower than 5,000 kg have the highest success rate

Section 5

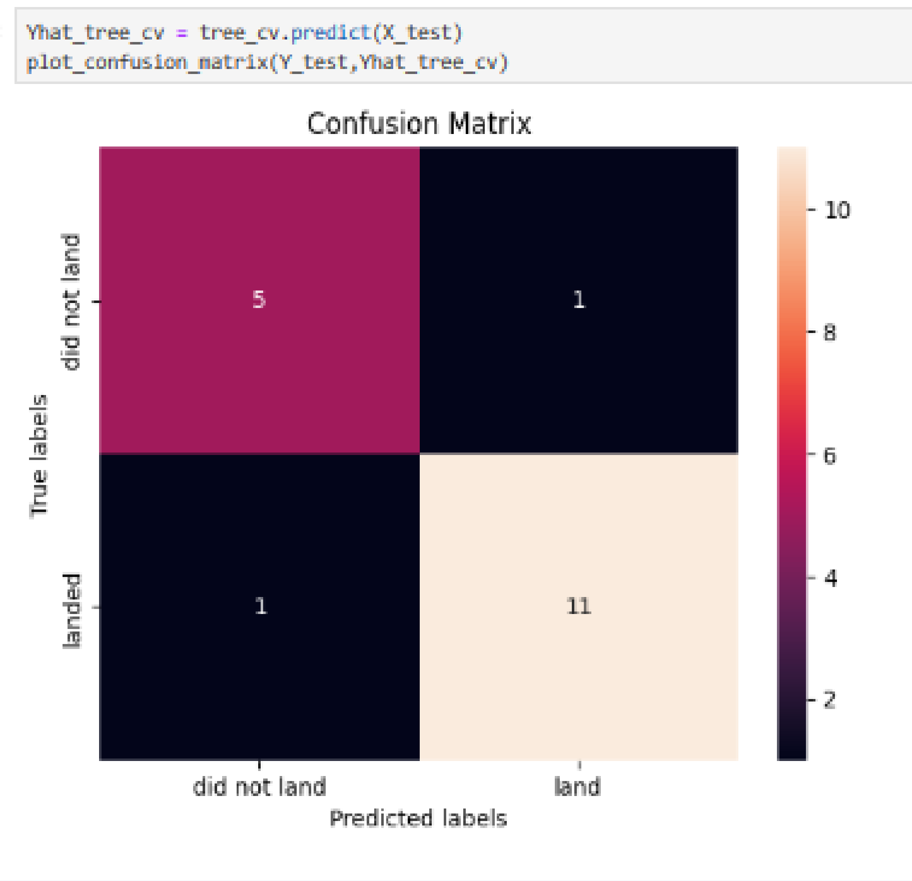
Predictive Analysis (Classification)

Classification Accuracy



- All models have similar Accuracy (over 80%) with a slight advantage to the Decision Tree.

Confusion Matrix



- The Decision Tree Confusion Matrix show the smallest FP (false positive) and FN (false negative)

Conclusions

- Payloads lower than 5,000 kg have the highest Success Rate
- Remarkable improvement since 2013, in terms of Success Rate and number of launches. Improvements required for heavy payloads and higher orbits.
- All the classifiers had adequate accuracy, although Decision Tree had a slightly better performance of approximately 87% making it the most appropriate model for landing outcome prediction.

Thank you!