IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Judy Chen>
<2023-01-01>

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- *Summary of methodologies*

- Data Collection
- Data Wrangling
- EDA (Exploratory Data Analysis) with Data Visualization
- EDA with SQL
- Interactive Map with Folium
- Dashboard with Plotly Dash
- Predictive analysis (Classification)

- *Summary of all results*

- EDA results
- Interactive maps and dashboard
- Predictive results

Introduction: Background

- *Working as a data scientist of a new rocket company called SpaceX*
- *Falcon 9 rocket launches with cost greatly lower than that of other providers*
- *Reuse of first stage*

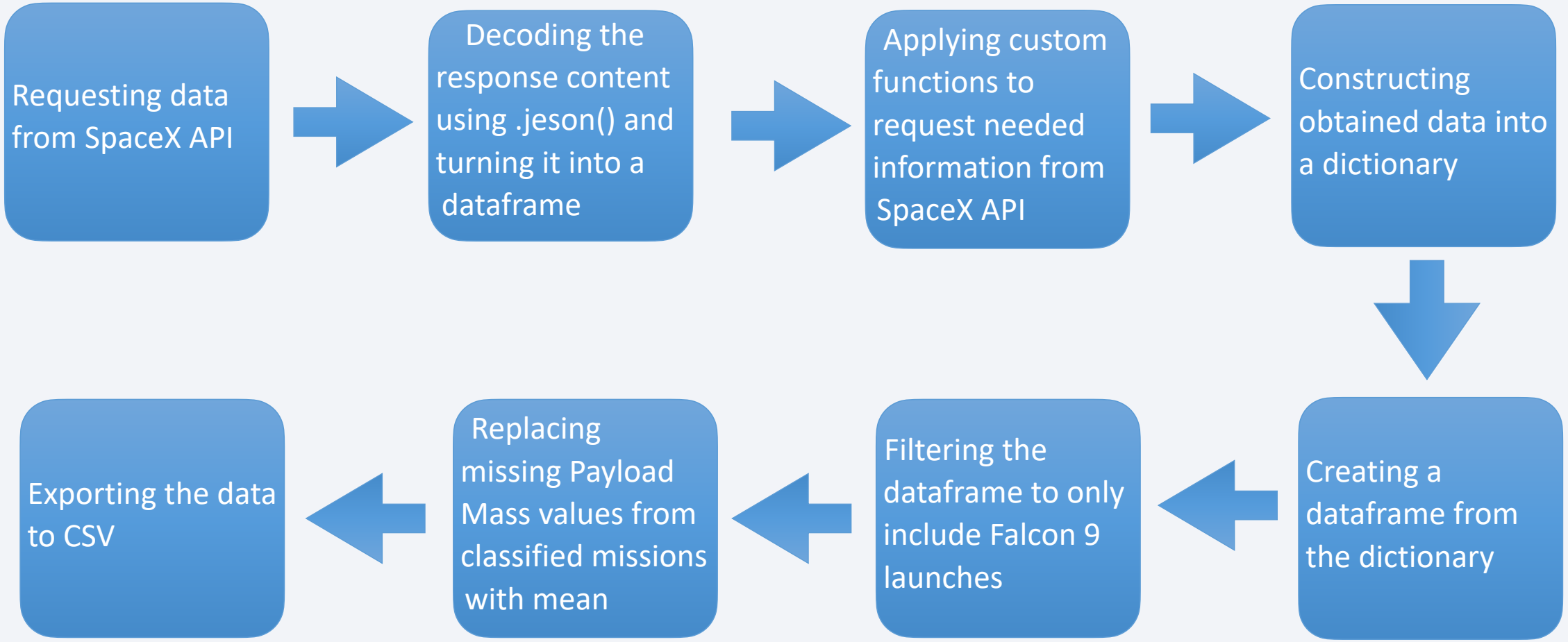
Introduction: Problems to be answered

- *How do variables such as payload mass, launch site, number of flights and orbits affect the success of the first stage landing?*
- *What are the conditions enabling SpaceX to achieve the best landing success rate?*
- *What is the best algorithm used for predicting if the first stage will land?*

Section 1

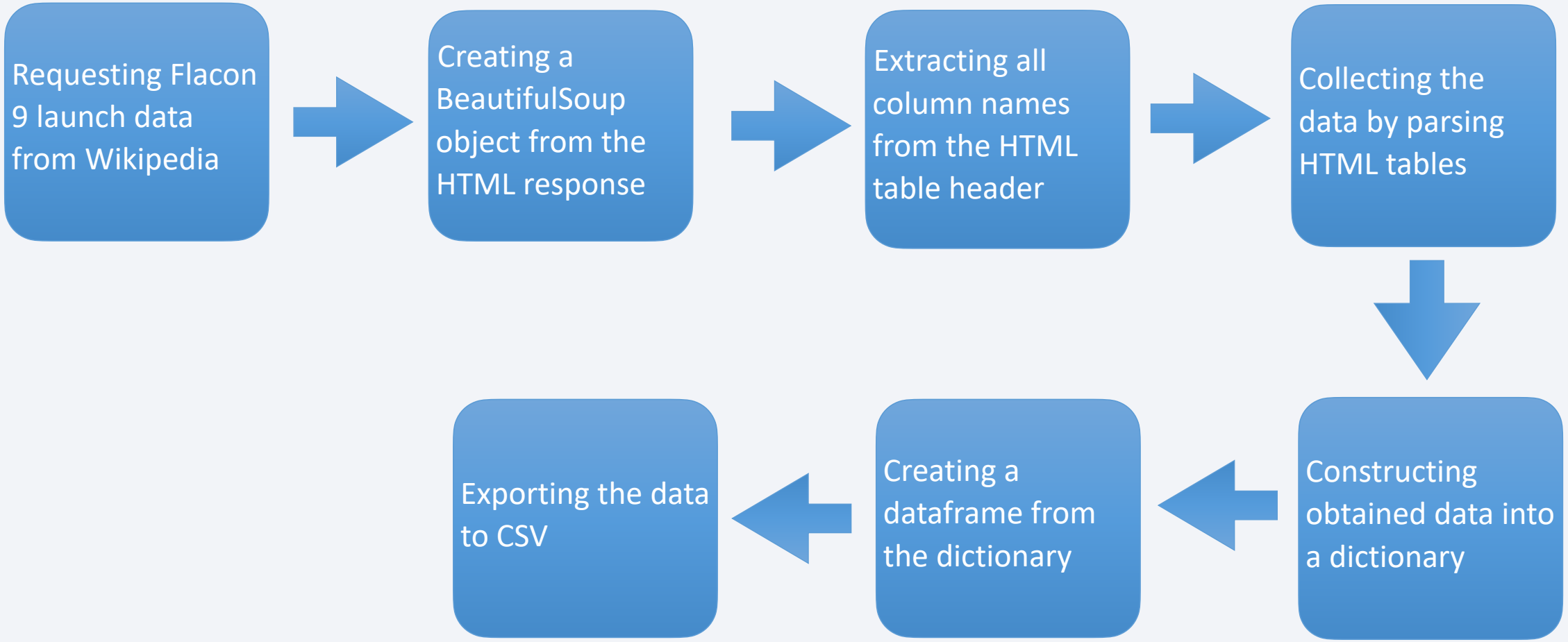
Methodology

Data Collection: SpaceX API



GitHub URL: Data Collection API

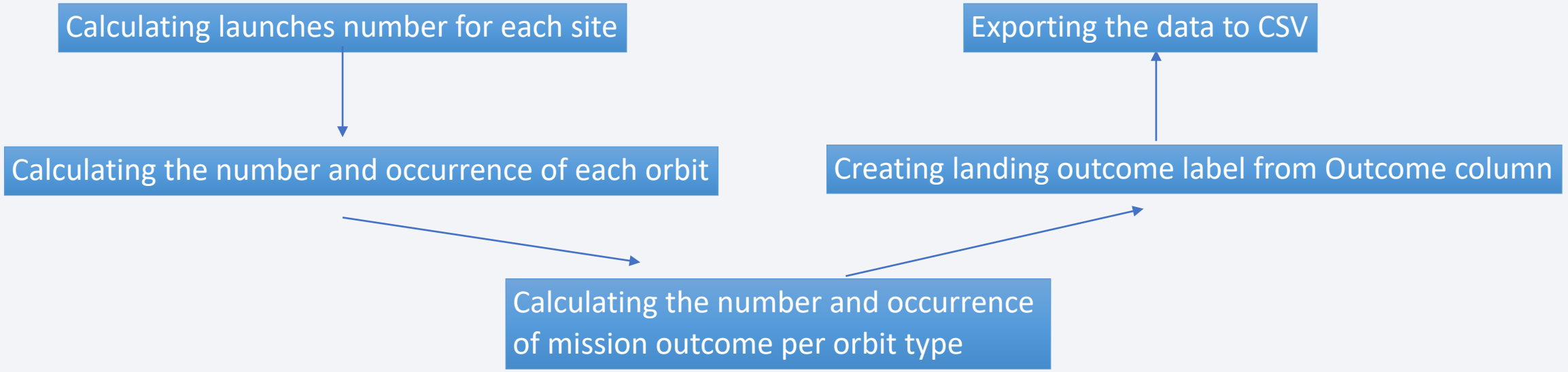
Data Collection: Web Scraping



GitHub URL: Data Collection with Web Scraping

Data Wrangling

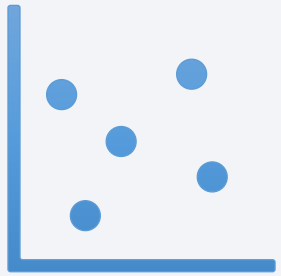
- In the dataset, **True Ocean**, **True RTLS** and **True ASDS** mean the mission was successful, while **False Ocean**, **False RTLS** and **False ASDS** mean the mission was unsuccessful.
- We will mainly convert the outcomes into training labels with **1** means the booster successfully landed **0** means it was unsuccessful.



GitHub URL: Data Wrangling

EDA with Data Visualization

- Scatter Plot
  - *Flight Number vs. Payload Mass*
  - *Flight Number vs. Launch Site*
  - *Payload Mass vs. Launch Site*
  - *Flight Number vs. Orbit Type*
  - *Payload Mass vs. Orbit Type*



Scatter plots show relationship between variables.

- Bar chart
  - *Success Rate vs. Orbit Type*

Bar charts show the relationship between numeric and categoric variables.

- Line chart
  - *Success Rate vs. Year*

Line charts show trends in data over time.

GitHub URL: EDA with Data Visualization

EDA with SQL

- **SQL queries performed**:

GitHub URL: EDA with SQL

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

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

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

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

5. List the date when the first successful landing outcome in ground pad was acheived.

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

7. List the total number of successful and failure mission outcomes.

8. List the names of the booster_versions carrying the maximum payload mass. Use a subquery.

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

10. 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.

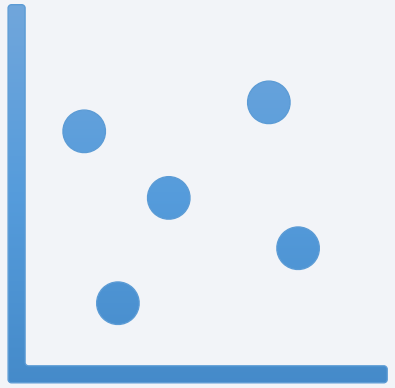
Build an Interactive Map with Folium

- Markers of all Launch Sites:
- *Adding Marker with Circle, Popup Label and Text Label of **NASA Johnson Space Center** using its latitude and longitude coordinates as a start location.*
- *Adding Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.*
- Colouring Markers of the launch outcomes for each Launch Site:
- *Adding coloured Markers of **success** (Green) and **failed** (Red) launches using Marker Cluster to identify which launch site has a relatively high success rate.*
- Distances between a Launch Site and its proximities:
- *Adding coloured lines to show distances between the Launch Site **CCAFS SLC-40** and its proximities like railway, highway, coastline and closest city.*

GitHub URL: Interactive Visual Analytics with Folium

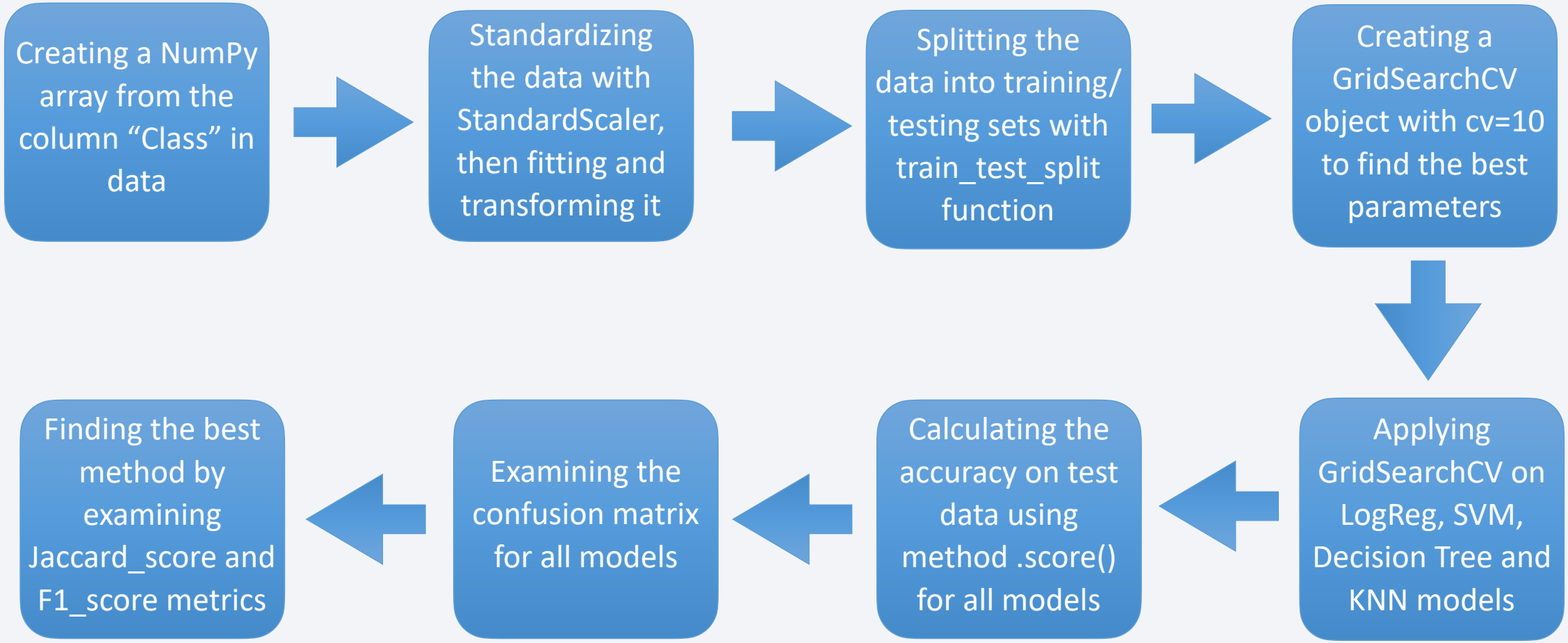
Build a Dashboard with Plotly Dash

- Adding a **dropdown list** to enable Launch Site selection
- Adding a **pie chart** to show the total success and the total failure for the Launch Site chosen from the dropdown list
- Adding a **slider** to enable selection of a payload mass within a fixed range
- Adding a **scatter chart** to show the correlation between payload mass and launch success rate



GitHub URL: SpaceX Dash App

Predictive Analysis (Classification)



GitHub URL: Machine Learning Prediction

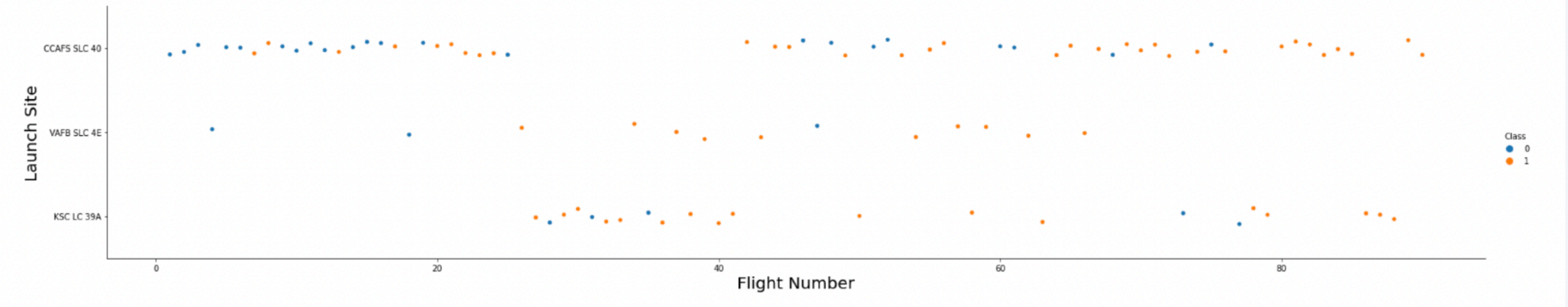
Results

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

Section 2

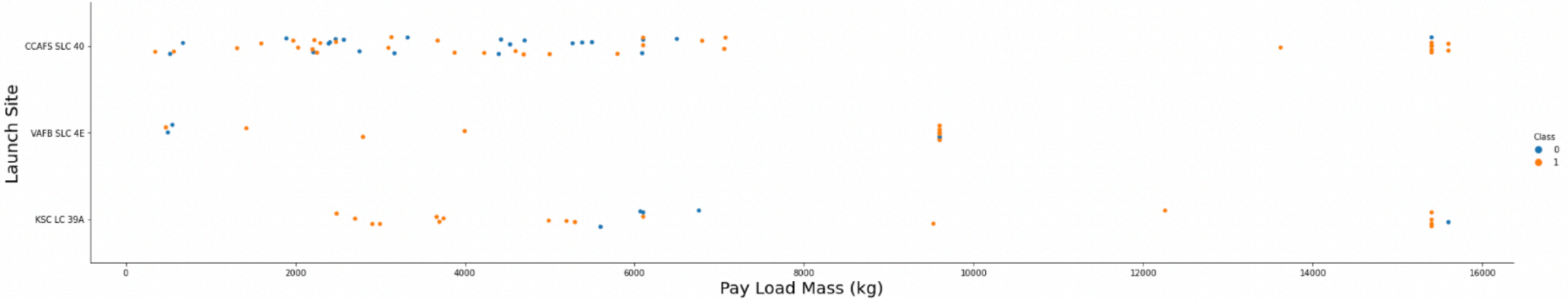
Insights drawn from EDA

Flight Number vs. Launch Site



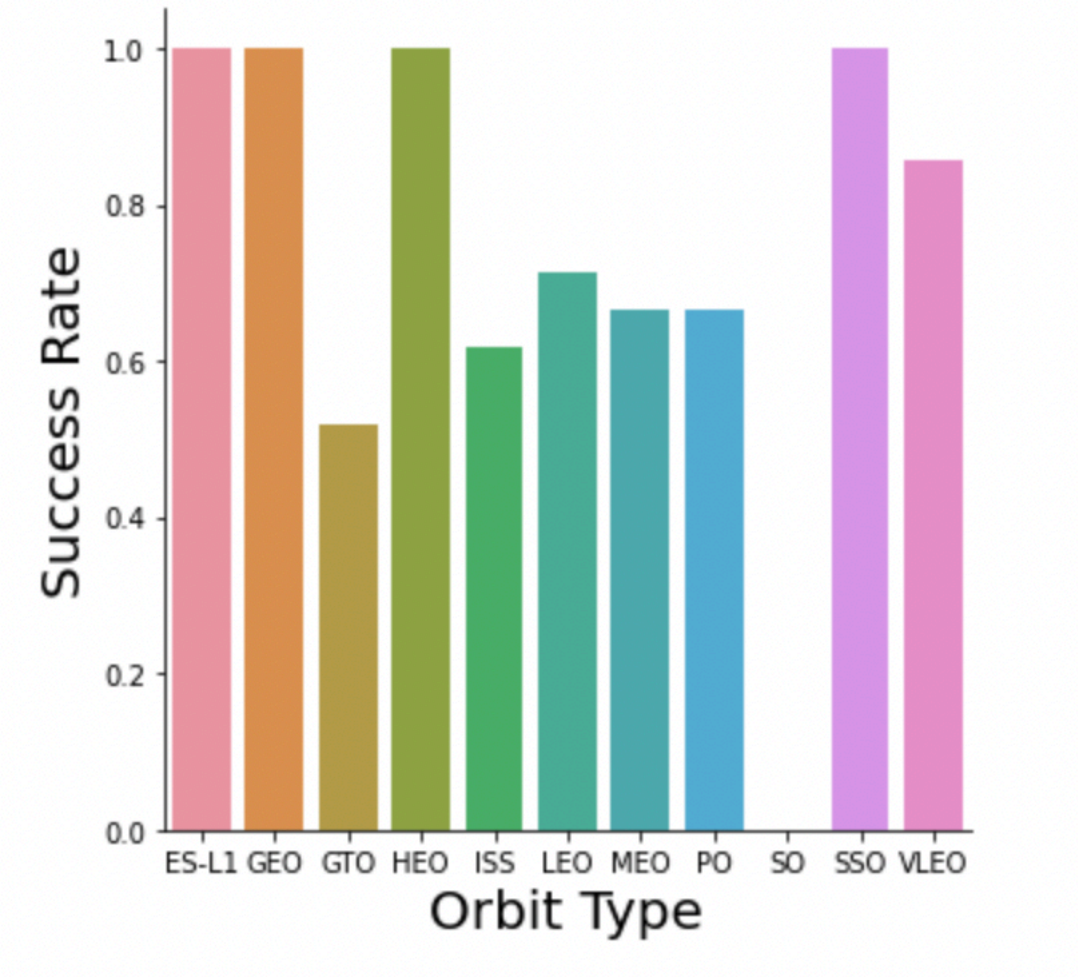
- *The CCAFS SLC 40 launch site has about a half of the total launches.*
- *The earliest launches all failed while the latest launches all succeeded.*
- *For each site, the success rate increases with the flight number.*

Payload vs. Launch Site



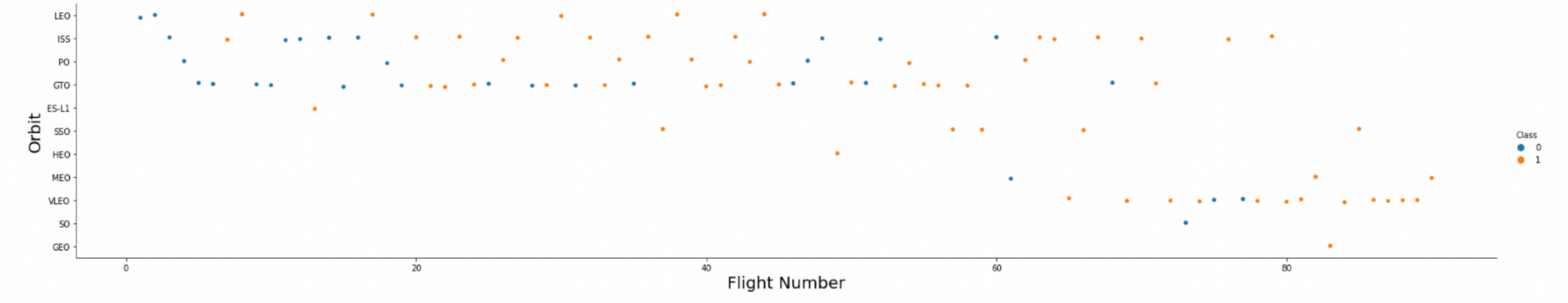
- *KSC LC 39A has a 100% success rate with payload mass under 5500kg.*
- *Most of the launches with payload mass over 7000kg are successful.*

Success Rate vs. Orbit Type



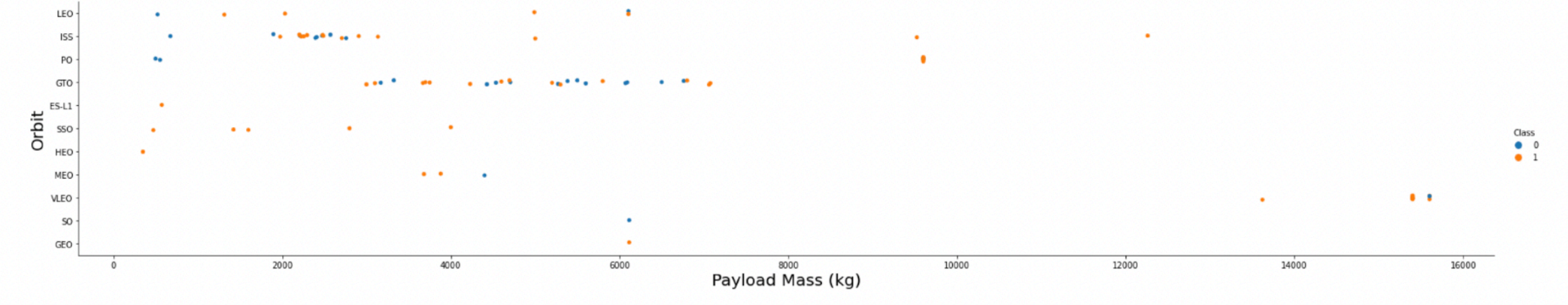
- *Orbits with 100% success rate: ES-L1, GEO, HEO, SSO*
- *Orbits with 0% success rate: SO*

Flight Number vs. Orbit Type



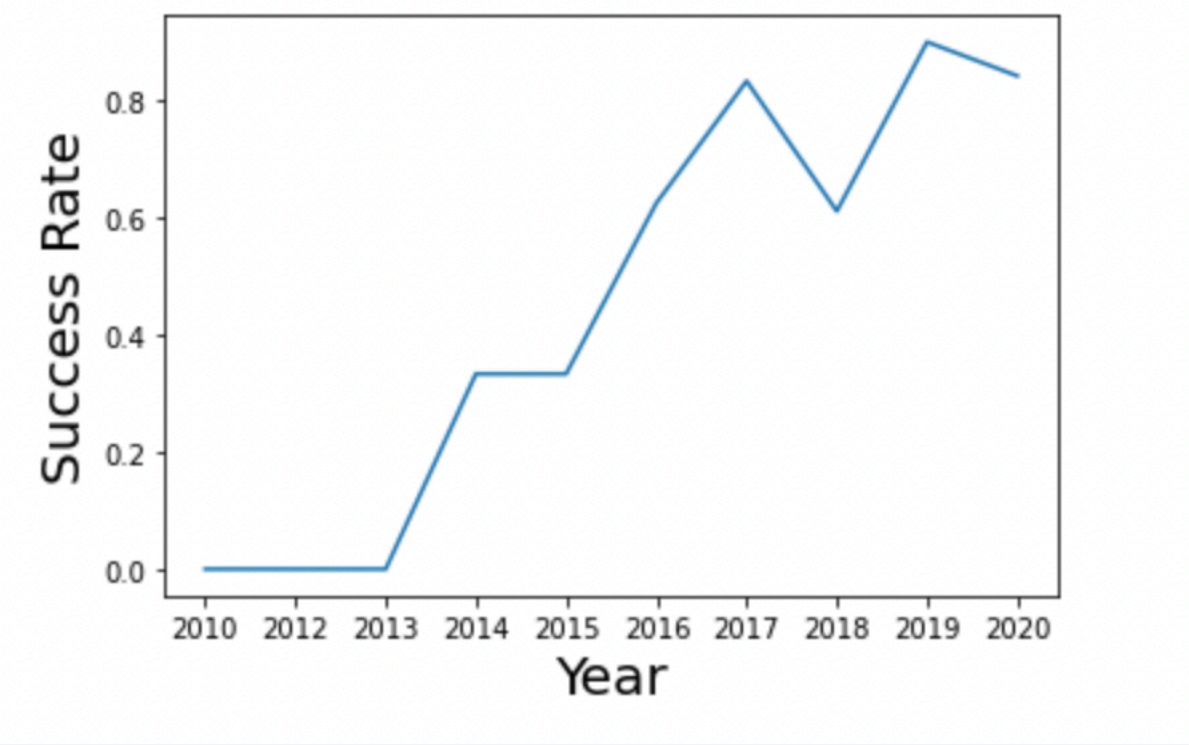
- *For the LEO orbit, the success rate increases with the flight number.*
- *For most orbits, there seems to be no relationship between the flight number and the success rate.*

Payload vs. Orbit Type



- *For the orbits with 100% success rate (ES-L1, GEO, HEO, SSO), the payload mass is under 6500kg.*
- *Heavy payload has a negative influence on GTO orbit and positive influence on LEO and ISS orbit.*

Launch Success Yearly Trend



- *Since 2013, we can see an obvious increase in the success rate.*

All Launch Site Names

```
%sql select distinct launch_site from CLK40286.SPACEX;
```

```
 * ibm_db_sa://clk40286:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

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

*Explanation: Using **distinct** in the query to remove duplicate launch_site*

Launch Site Names Begin with 'CCA'

```
%sql select * from CLK40286.SPACEX where launch_site like 'CCA%' limit 5;
```

```
 * ibm_db_sa://clk40286:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/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

*Explanation: The where clause following by **like** clause filters launch sites containing the substring 'CCA'.*

Total Payload Mass

```
%sql select sum(payload_mass__kg_) as total_payload_mass from CLK40286.SPACEX where customer like '%NASA (CRS)%';
```

```
 * ibm_db_sa://clk40286:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

total_payload_mass
48213

Explanation: Displaying the total payload mass carried by boosters launched by NASA (CRS).

Average Payload Mass by F9 v1.1

```
%sql select avg(payload_mass__kg_) as average_payload_mass from CLK40286.SPACEX where booster_version like '%F9 v1.1%';
```

```
 * ibm_db_sa://clk40286:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

average_payload_mass
2534

Explanation: Displaying the average payload mass carried by booster version F9 v1.1.

First Successful Ground Landing Date

```
%sql select min(date) as first_successful_landing from CLK40286.SPACEX where landing__outcome = 'Success (ground pad)';
```

```
 * ibm_db_sa://clk40286:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

first_successful_landing
2015-12-22

*Explanation: Displaying the data when the first successful landing outcome in ground pad was achieved, wherein **min()** function is used to select the record with the oldest date.*

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select booster_version from CLK40286.SPACEX where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6(
```

```
 * ibm_db_sa://clk40286:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

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

Explanation: Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.

Total Number of Successful and Failure Mission Outcomes

```
%sql select mission_outcome, count(*) as total_number from CLK40286.SPACEX group by mission_outcome;
```

```
 * ibm_db_sa://clk40286:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

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

*Explanation: Listing the total number of successful and failure mission outcomes by using **group by**.*

Boosters Carried Maximum Payload

```
%sql select booster_version from CLK40286.SPACEX where payload_mass__kg_ = (select max(payload_mass__kg_) from CLK40286.SPACEX);
```

```
 * ibm_db_sa://clk40286:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
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

*Explanation: Listing the names of the booster versions which have carried the maximum payload mass by using **max()** function and **subquery**.*

2015 Launch Records

```
%%sql select date, booster_version, launch_site, landing__outcome from CLK40286.SPACEX
      where landing__outcome = 'Failure (drone ship)' and year(date)=2015;
```

```
 * ibm_db_sa://clk40286:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

DATE	booster_version	launch_site	landing__outcome
2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Explanation: Listing the failed landing outcomes in drone ship, their booster versions and launch site names in year 2015.

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

```
%%sql select landing__outcome, count(*) as count_outcomes from CLK40286.SPACEX
      where date between '2010-06-04' and '2017-03-20'
      group by landing__outcome
      order by count_outcomes desc;
```

```
 * ibm_db_sa://clk40286:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

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

*Explanation: Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the data 2010-06-04 and 2017-03-20 in descending order by using **group by** and **order by**.*

Section 3

Launch Sites Proximities Analysis

Folium Map-location markers

Most of Launch sites are in proximity to the Equator line.

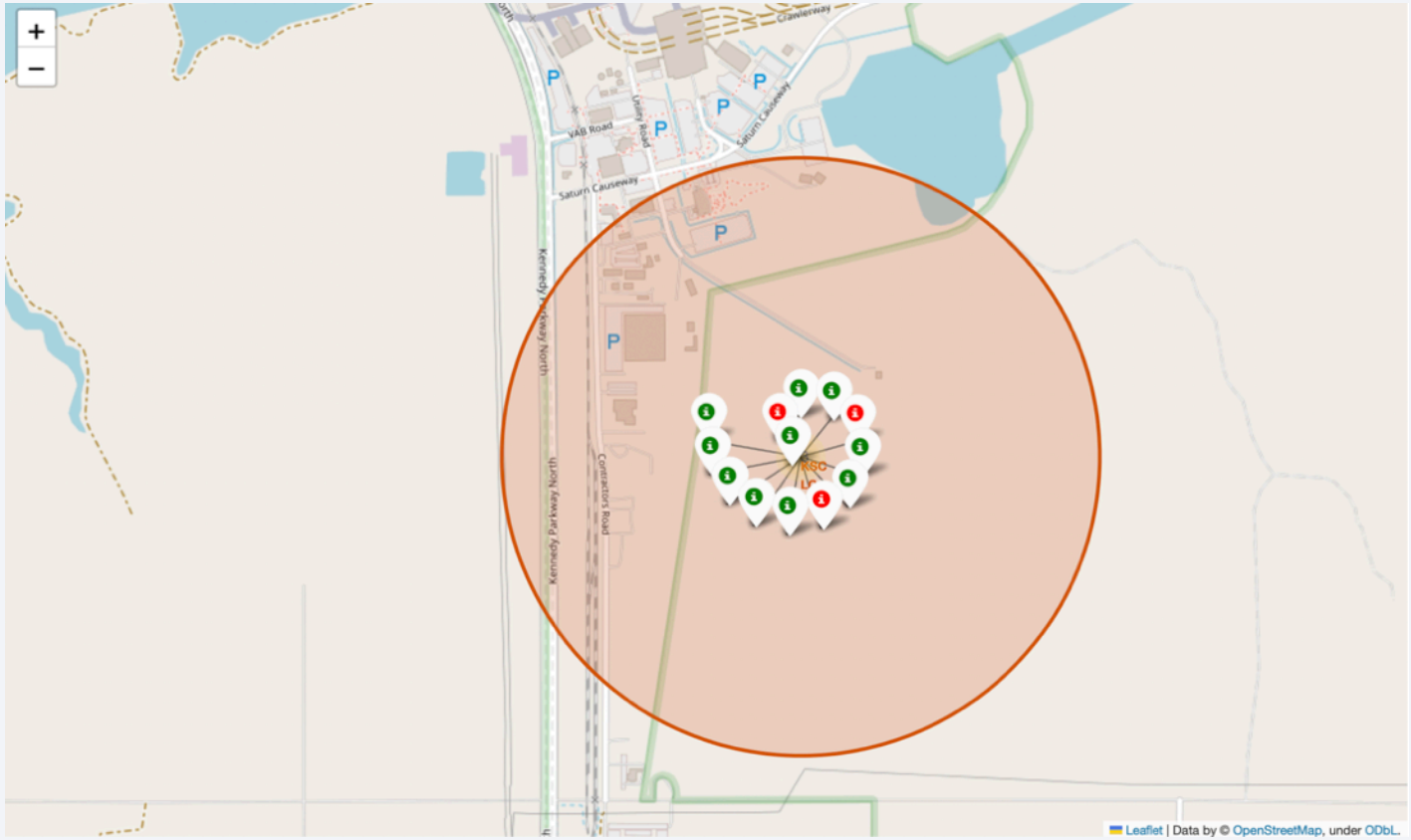
All launch sites are in very close proximity to the coast.

Folium Map-color-labeled markers

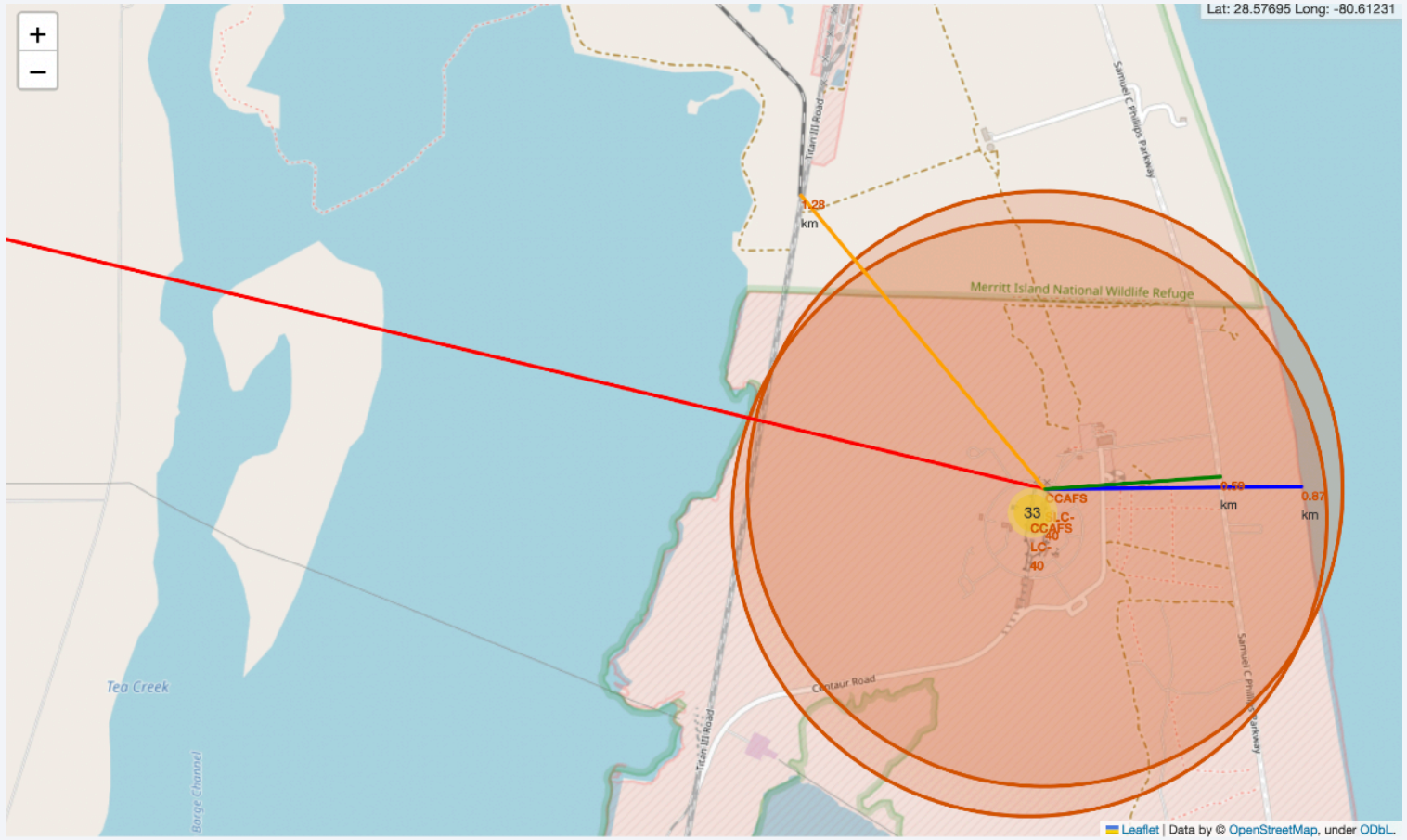
Green Marker - Successful Launch
Red Marker - Failed Launch

From the color-labeled markers in marker clusters, we can easily identify which launch sites have relatively high success rates.

*Launch Site **KSC LC-39A** has a high success rate.*



Folium Map-Distance between CCAFS SLC-40 and its proximities



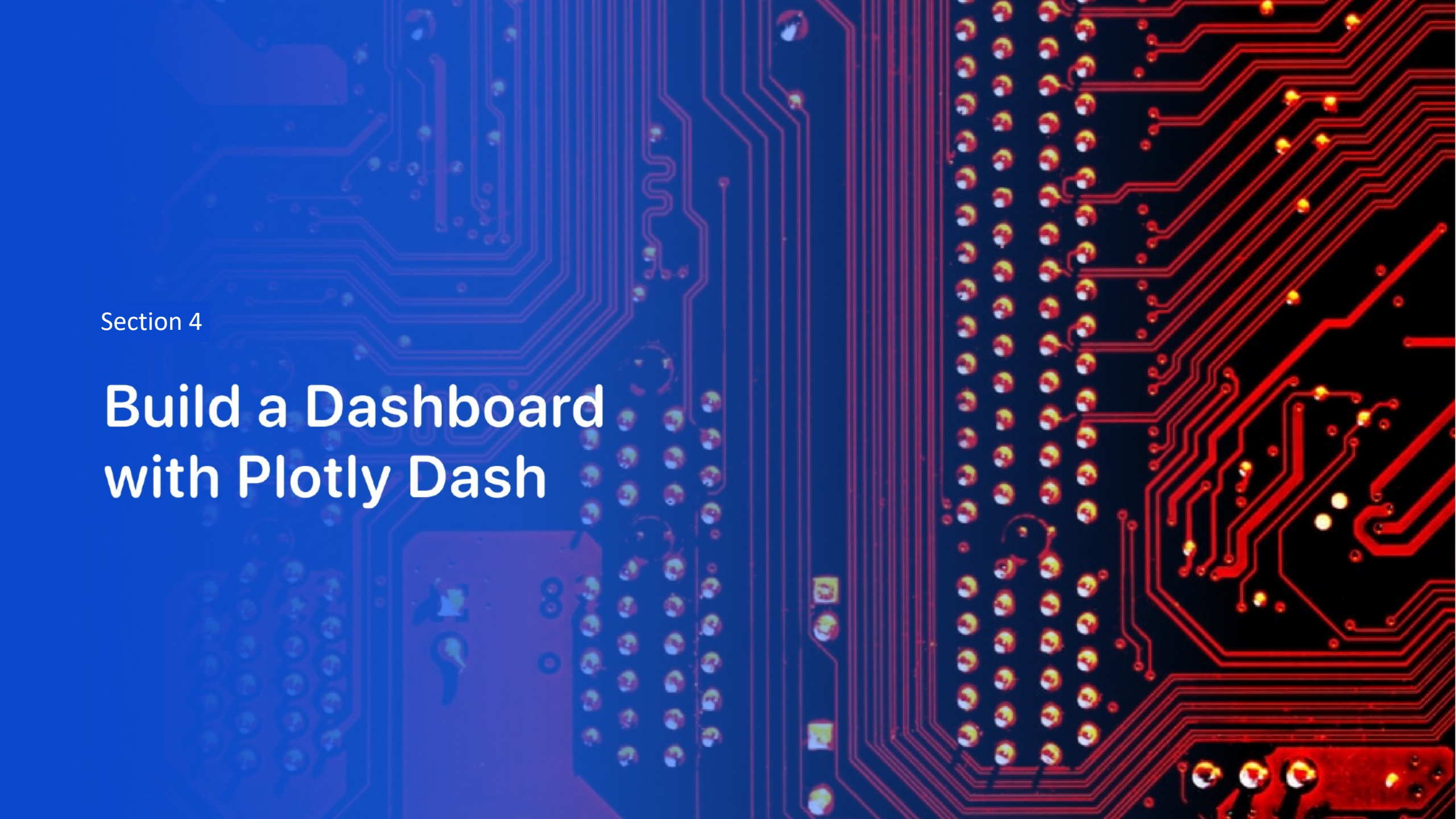
*Launch site **CCAFS SLC-40**:*

In very close proximity to railway: 1.28km

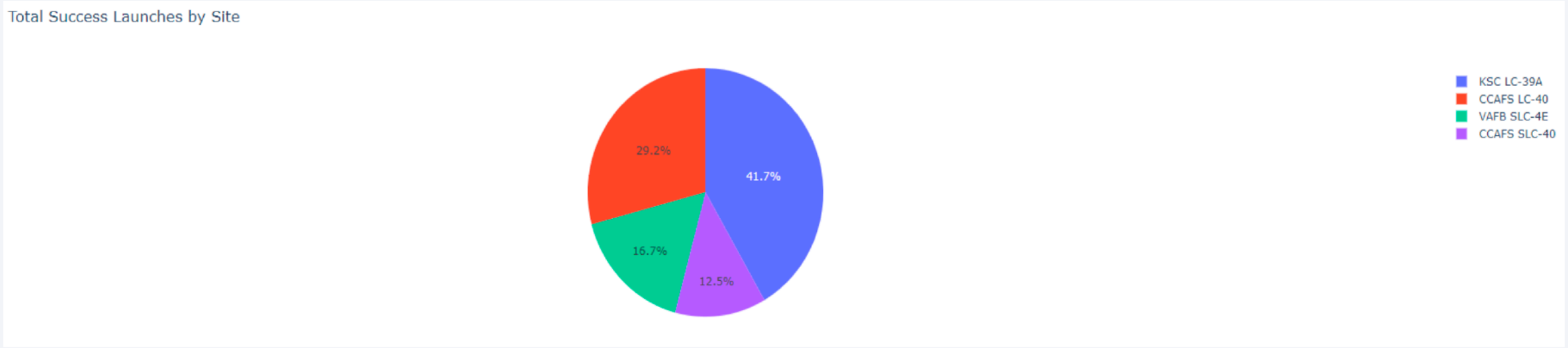
In very close proximity to highway: 0.59km

In very close proximity to coastline: 0.87km

In relative close proximity to the closest city Titusville: 23.21km

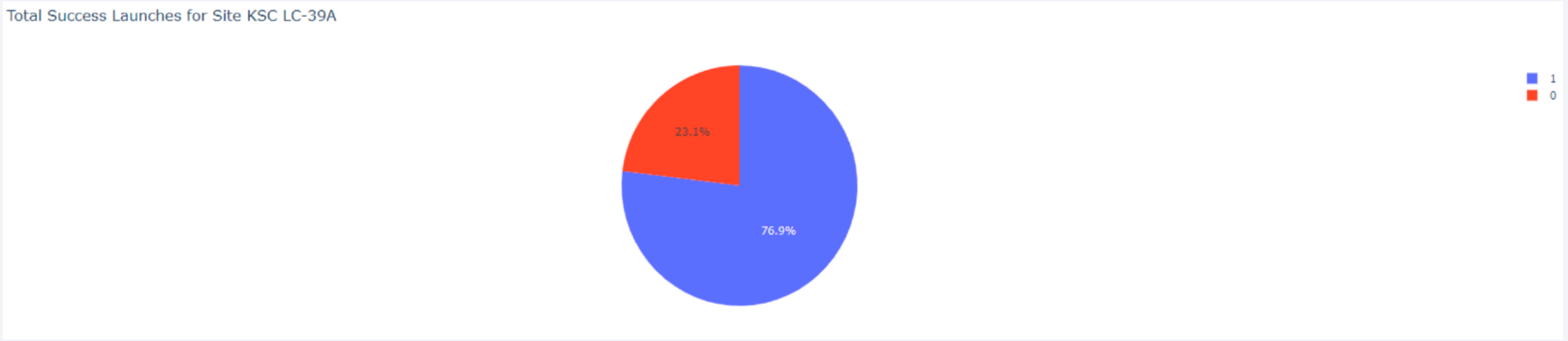

Section 4
Build a Dashboard with Plotly Dash

Dashboard-Total Success Launches by Site



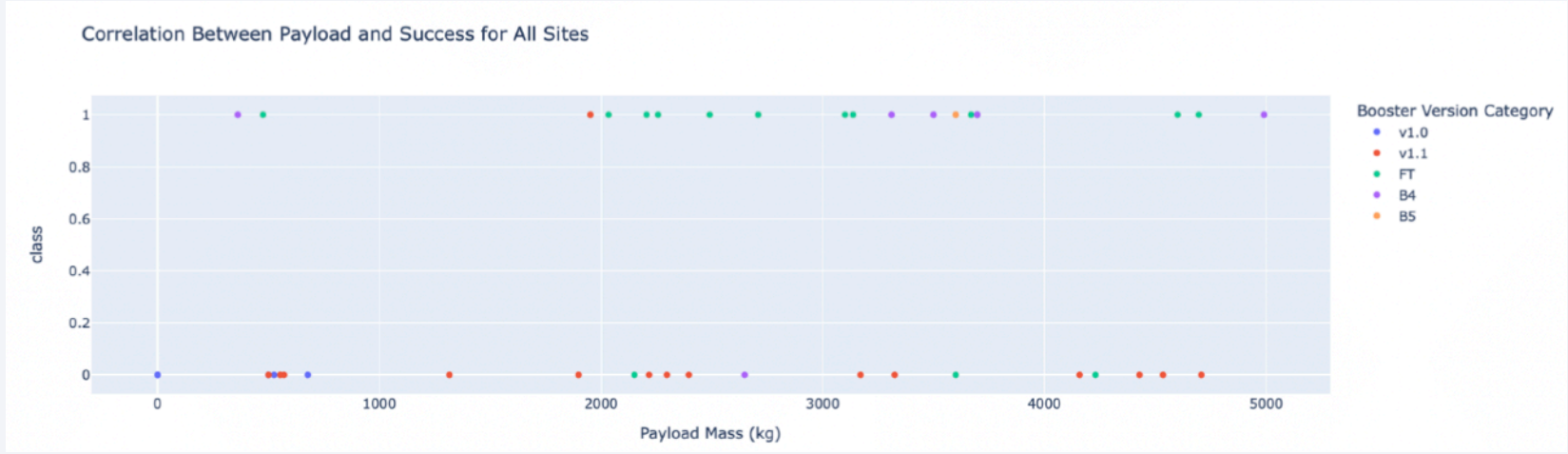
*Explanation: The pie chart clearly indicates that site **KSC LC-39A** has the most successful launches.*

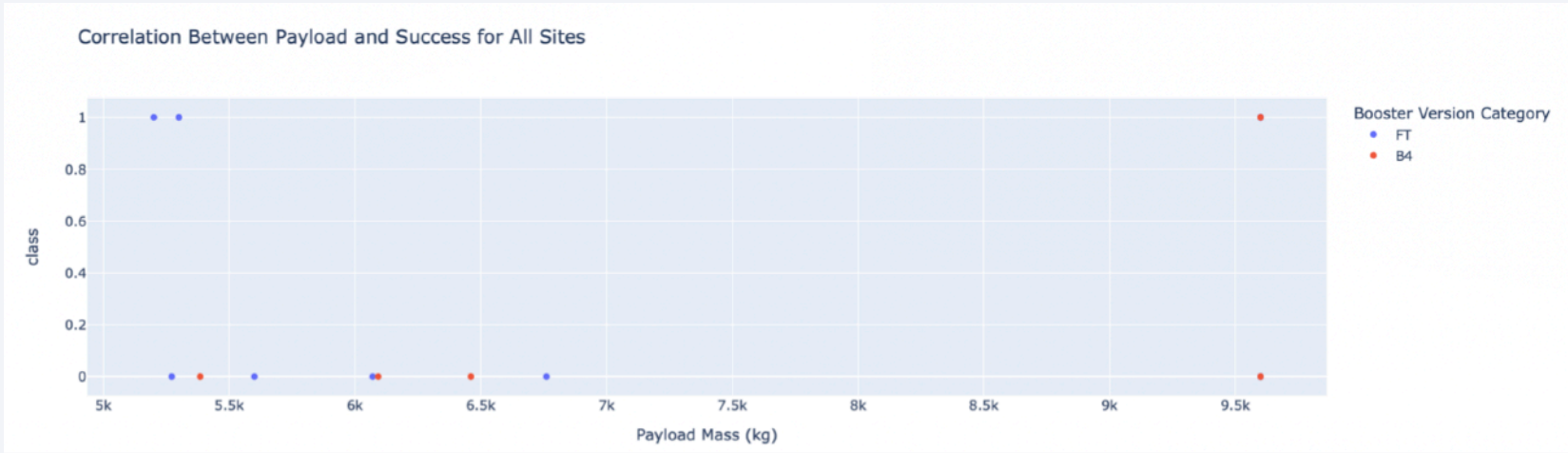
Dashboard-Total Success Launches for Site KSC LC-39A



*Explanation: The pie chart clearly indicates that site **KSC LC-39A** has achieved a **76.9%** success rate.*

Dashboard-Payload Mass vs. Launch Outcome Scatter Plot





Explanation: The chart shows that payload mass between 2000kg and 5500kg has a relatively high success rate, while payload mass between 5500kg and 7000kg has a ***0%*** *success rate.*

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Method Performance on Test Set

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

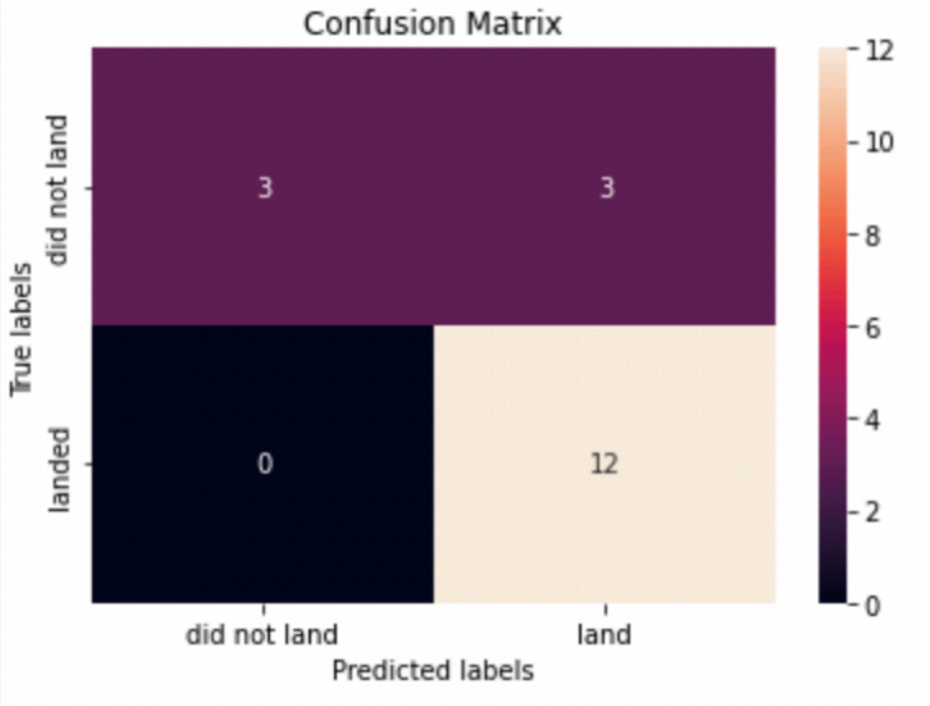
Method Performance on Entire Data Set

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.803030	0.819444
F1_Score	0.909091	0.916031	0.890756	0.900763
Accuracy	0.866667	0.877778	0.855556	0.855556

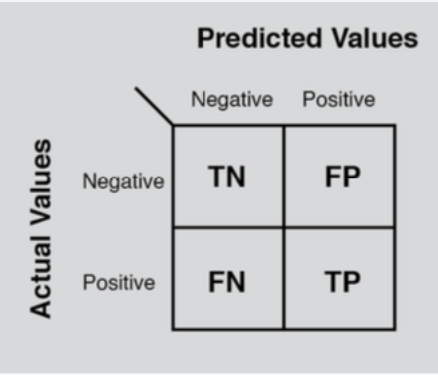
Explanation:
Due to the small size (only 18 samples) of the test set, we could not confirm which method performs best based on the same score.

We try to test all methods based on the entire dataset, and can confirm that the best model is the SVM which performs relatively better on both score and accuracy.

Confusion Matrix



Explanation:
As the methods have the same test accuracy, the confusion matrixes are also the same.
The main problem of these models is false positive.



Conclusions

- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- The success rate of launches increases over the year.
- The orbits with the best success rate are GEO, HEO, SSO and ES-L1.
- KSC LC-39A has the highest success rate of the launches in all sites.
- Launches with a low payload mass generally perform better than launches with a high payload mass.
- Support Vector Machines may be the best algorithm for the dataset.

Thank you!