IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Chiara Brambilla
9 March 2023

Outline

Executive Summary

In order to determine the cost of a launch, we predict whether the first stage of SpaceX Falcon 9 will land successfully.

The methodologies used to achieve this goal include Data Collection, Data Wrangling and Data Pre-processing, Exploratory Data Analysis (with SQL and visual tools), and Machine Learning algorithms.

Our analysis emphasises that some features are strictly related to a higher rate of success of rocket launches. Moreover, our research points out which is the best machine learning algorithm for this problem.

Introduction

SpaceX declares that Falcon 9 rocket launches cost of 62 million, while other providers cost upward of 165 million dollars each. Much of the savings is because SpaceX can reuse the first stage. If we can determine if the Falcon 9 first stage will land, we can determine the cost of a launch. This information can be used to bid against SpaceX for a rocket launch.

The main goal of this capstone is to collect and analyze data to answer the following question:

Will the Falcon 9 first stage land successfully?

Section 1

Methodology

Methodology

Executive Summary

Data was collected both from SpaceX API and from Wikipedia by Web Scraping. In the first case we perform data wrangling by cleaning and ensuring data structure, while in the other case we parse HTML table and convert it data frame using Python's Pandas library.

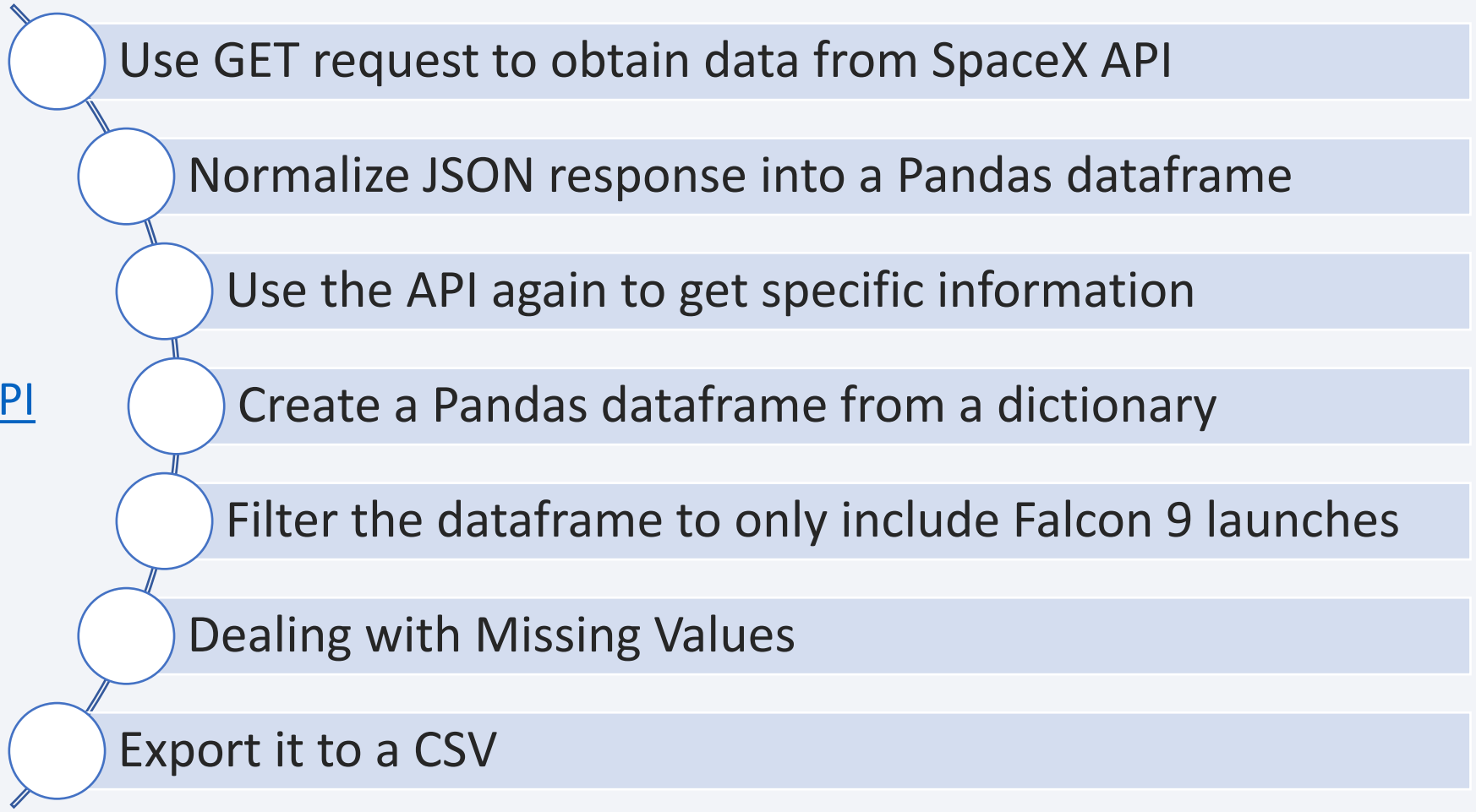
We perform some Exploratory Data Analysis (EDA) to find some patterns in the data. We visualize information using Python's Pandas and Matplotlib libraries and use SQL queries to understand the data set. We perform interactive visual analytics using Folium and Plotly Dash.

Finally, we perform predictive analysis using different classification models. All models were trained and their best hyperparameter's were found. Then, we evaluate them to find which is the best one.

Data Collection – SpaceX API

GitHub URL:

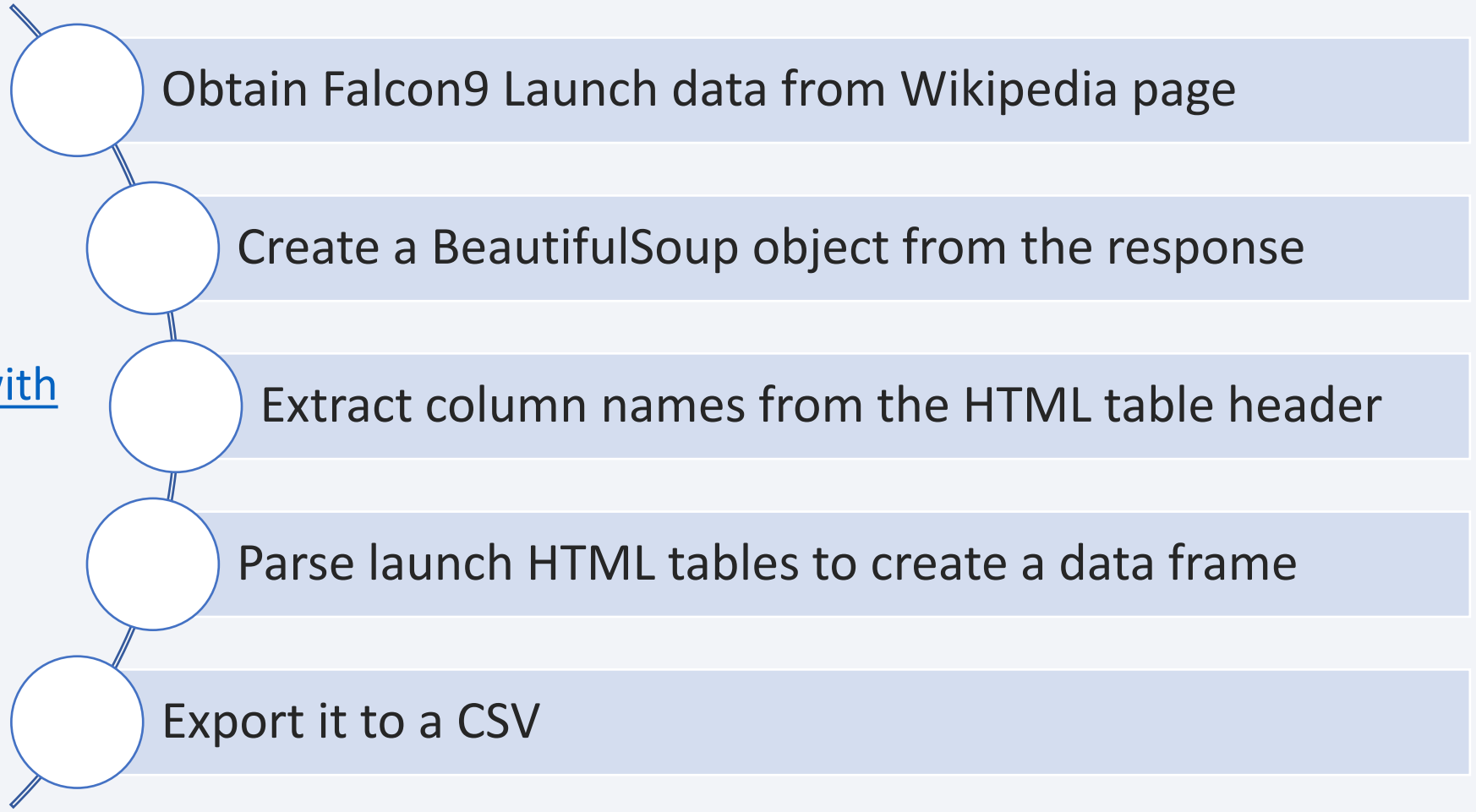
Data Collection API



Data Collection - Scraping

GitHub URL:

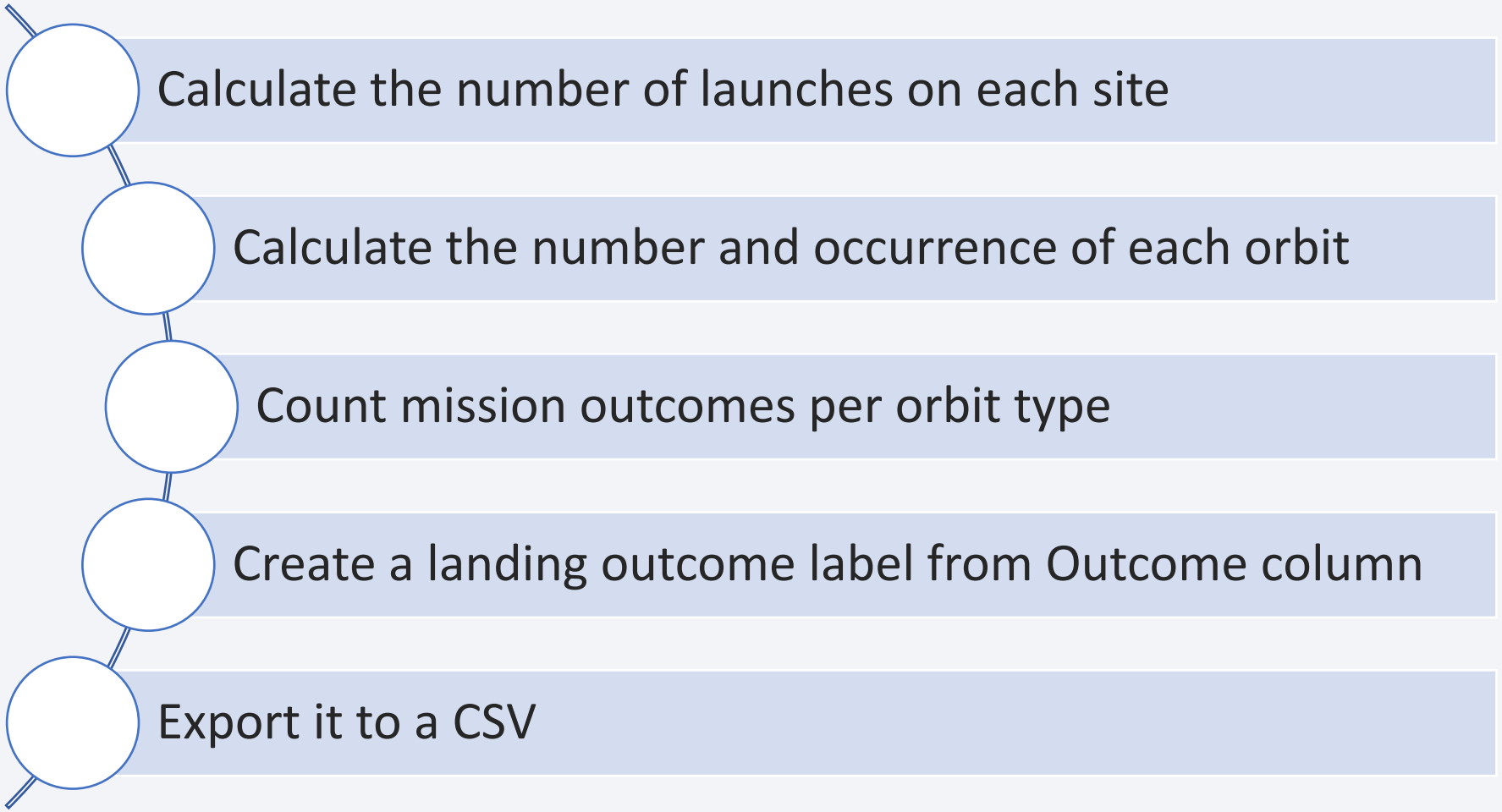
Data Collection with Web Scraping



Data Wrangling

GitHub URL:

EDA



EDA with Data Visualization

GitHub URL:

EDA with Data Visualization

- *Scatter plots*: to display how different variables would affect the launch outcome. Different couples were considered:

  Flight Number vs Payload Mass, Flight Number vs Launch Site, Payload vs Launch Site, Flight Number vs Orbit type, Payload vs Orbit type.

- *Bar chart*: to visualize the relationship between different variables, such as *Success Rate and Orbit type*

- *Line chart*: to visualize trends over time, such as *Launch Success yearly trend*

EDA with SQL

GitHub URL:

EDA with SQL

- Display:
  - names of the unique launch sites in the space mission
  - 5 records where launch sites begin with the string '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 landing outcome in ground pad
  - boosters with success in drone ship and payload mass between 4000 and 6000
  - total number of successful and failure mission outcomes
  - booster versions' names which have carried the maximum payload mass
  - failed landing outcomes in drone ship, with booster versions and launch site names for 2015
- Rank:
  - count of landing outcomes between 2010-06-04 and 2017-03-20 , in descending order

Build an Interactive Map with Folium

GitHub URL:
Interactive Visual Analytics with Folium lab

To find geographical patterns about launch sites, we complete the tasks:

1. Add *markers* to identify launch sites and *circles* to highlight the area
2. Add *markers* to display success/failed launches for each site
3. Add *lines* to compute distances of a launch site and answer the questions:
   - Are launch sites close to railways? YES
   - Are launch sites close to highways? YES
   - Are launch sites close to coastline? YES
   - Do launch sites keep certain distance away from cities? YES

Build a Dashboard with Plotly Dash

The Dashboard application display:

- a *pie chart* to show the total successful launches count for all sites or the success rate for specific selected site
- a *scatter chart* to show the correlation between payload and launch success for a selected site and range of payload mass

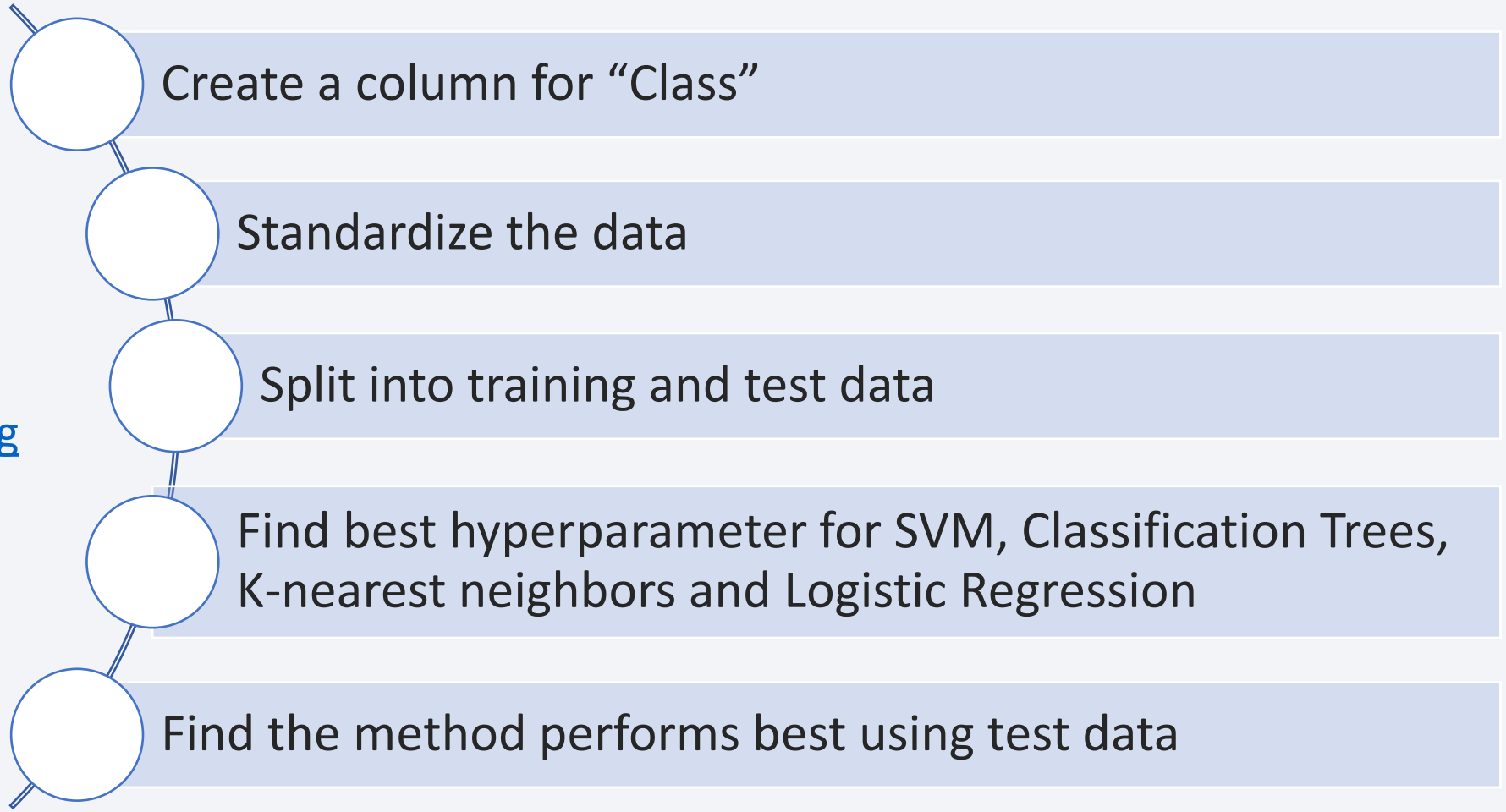
GitHub URL:

SpaceX Dash App

Predictive Analysis (Classification)

GitHub URL:

Machine Learning Prediction

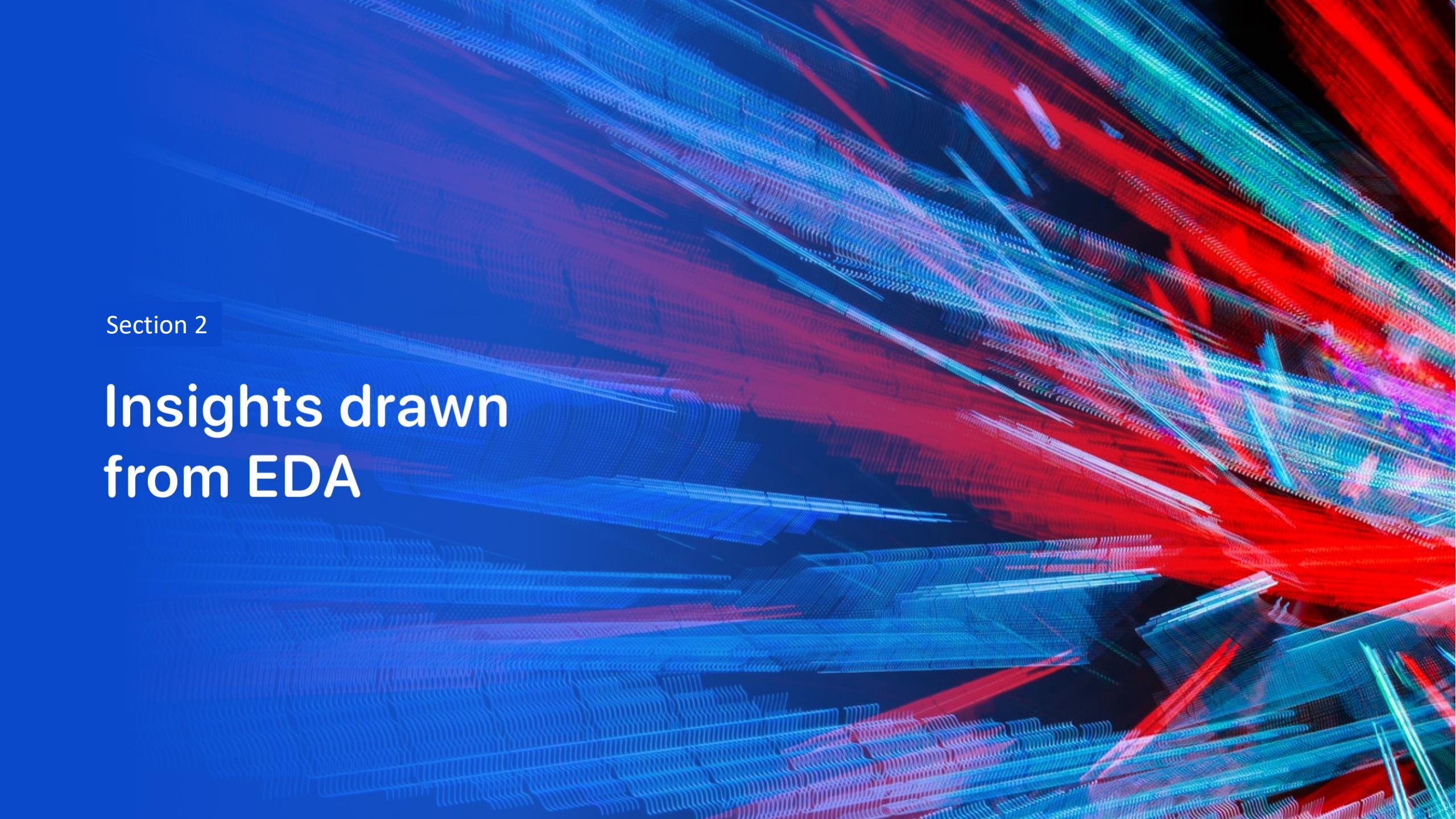


Results

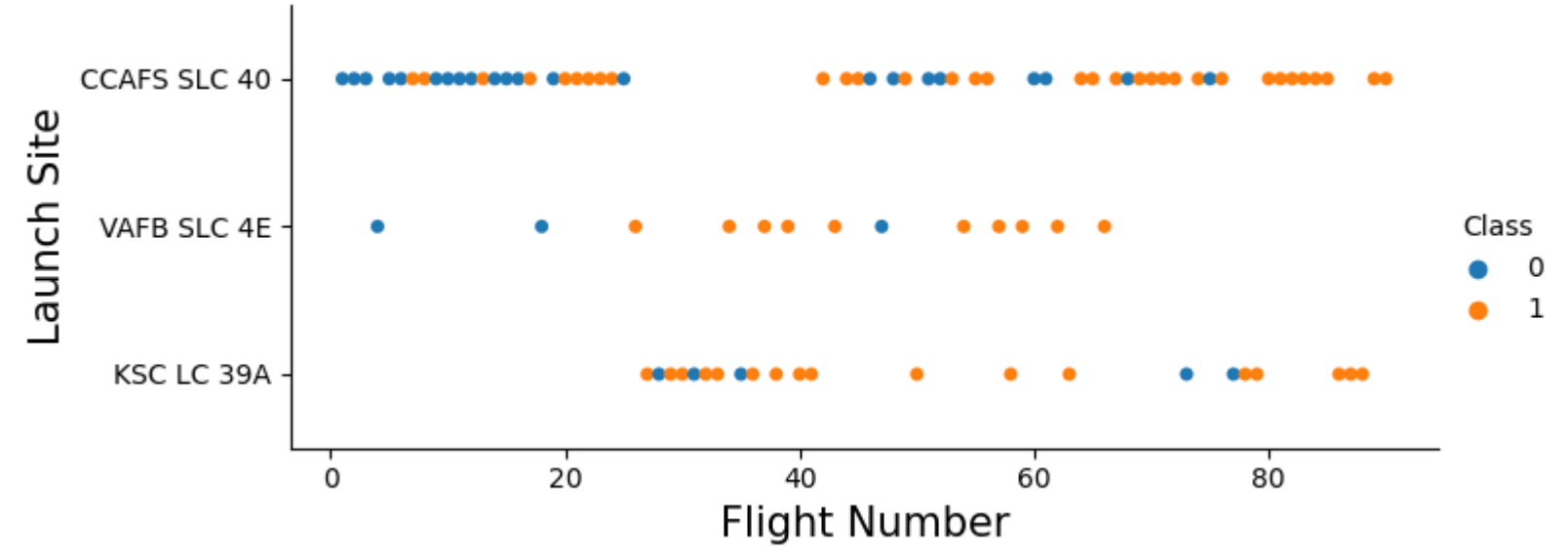
Exploratory data analysis shows that the success rate since 2013 kept increasing till 2020.

Interactive analytics is very useful to understand that launch sites are very close to highways and railways making transport easier, but they are quite distant from cities for security reasons.

Predictive analysis compares different methodologies and points out that decision tree classifier gets the best test accuracy score and fit accuracy.


Section 2
Insights drawn from EDA

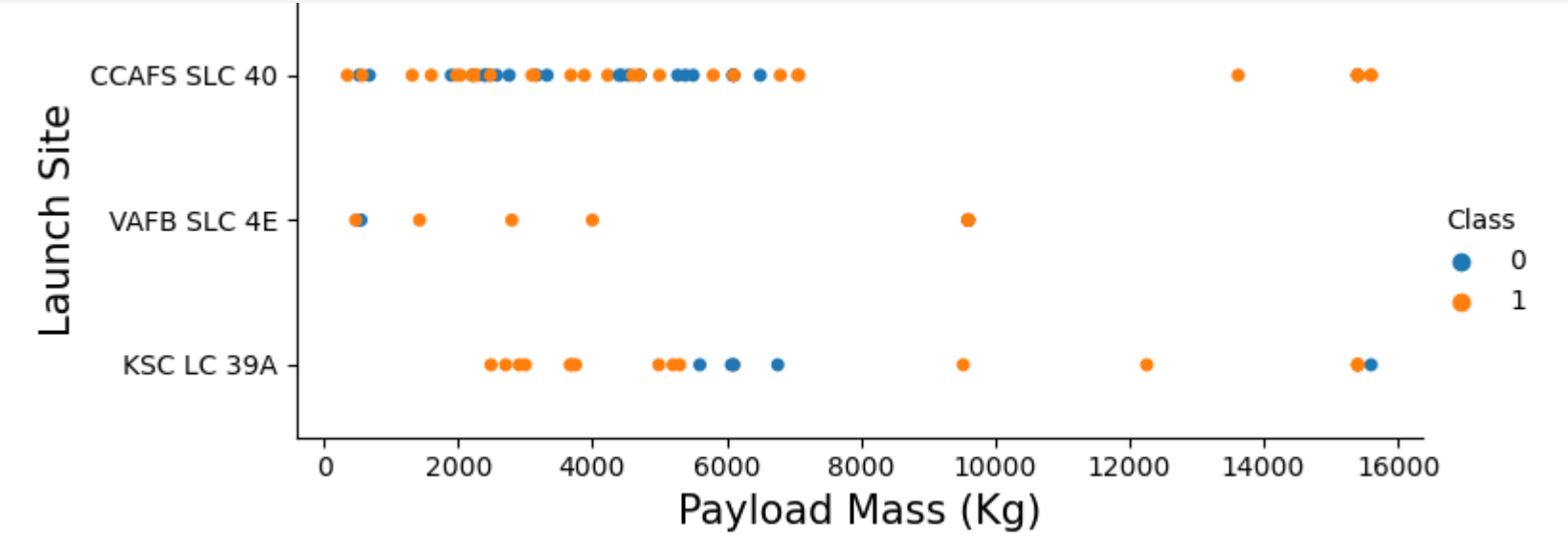
Flight Number vs. Launch Site



- Class 0 represents the unsuccessful launches
- Class 1 represents successful launches

Success rate increased with the number of flight.
Different launch sites have different success rates: CCAFS LC-40 has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E have a success rate of 77%.

Payload vs. Launch Site

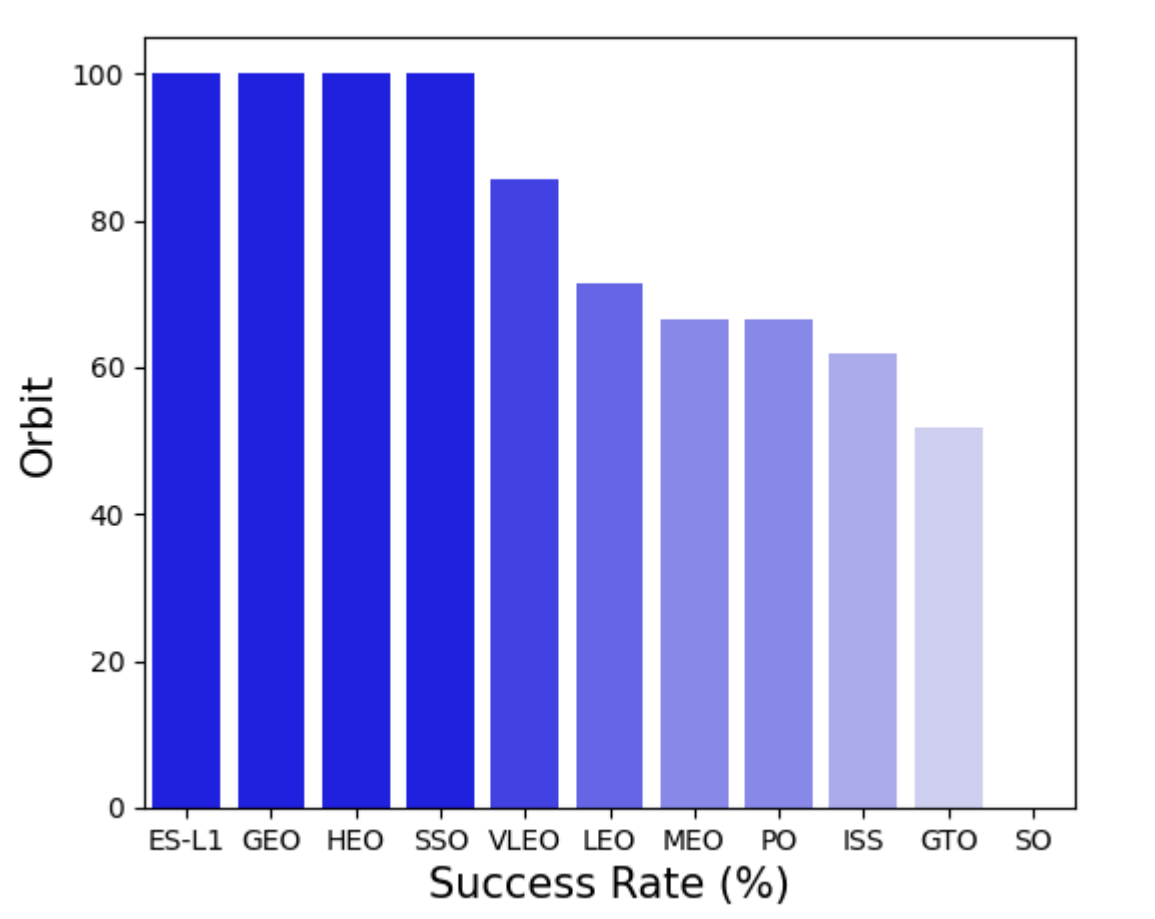


- Class 0 represents the unsuccessful launches
- Class 1 represents successful launches

No relationship is pointed out.

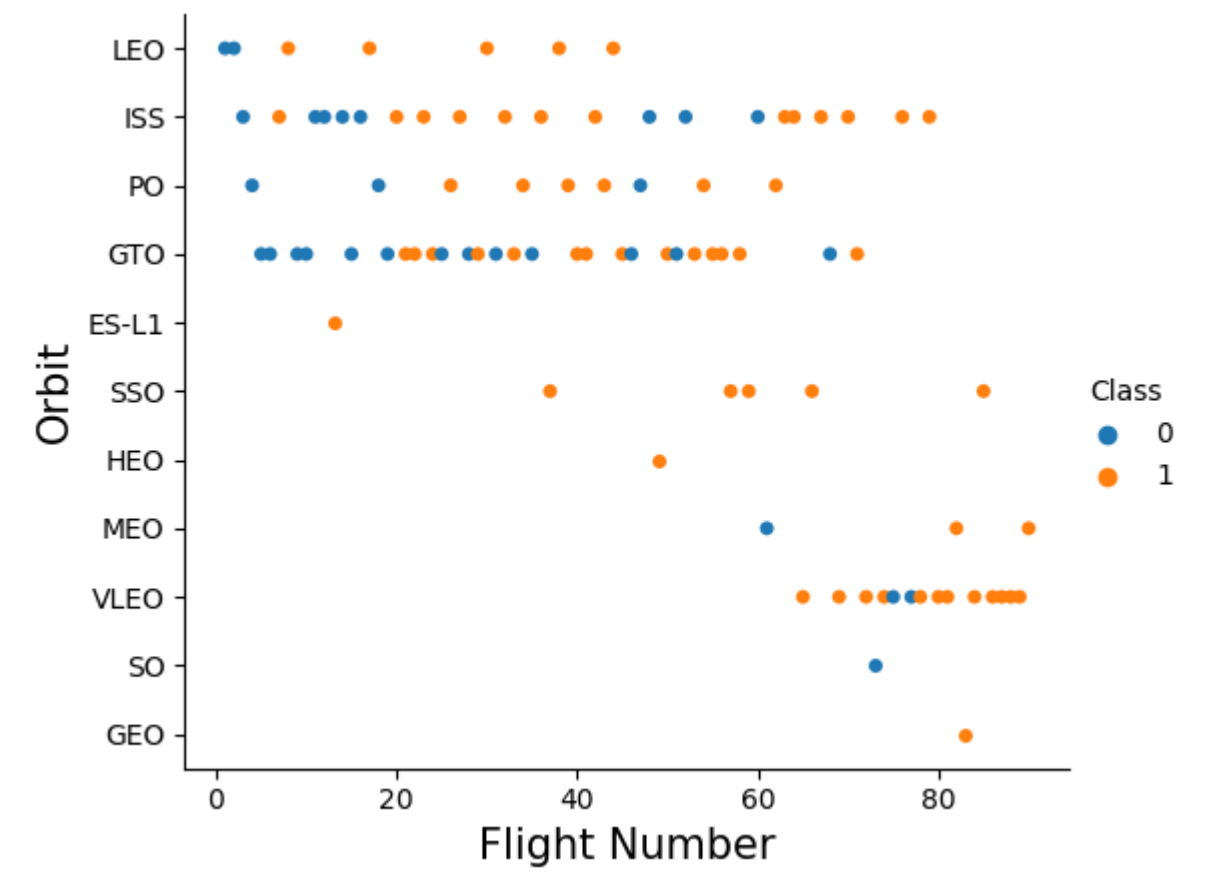
However, we observe that VAFB-SLC launch site has no rockets launched for heavy payload mass, i.e., greater than 10000.

Success Rate vs. Orbit Type



- Four orbits have success rates equal to 100%. They are ES-L1, GEO, HEO and SSO.
- SO orbit has a success rate of 0%.

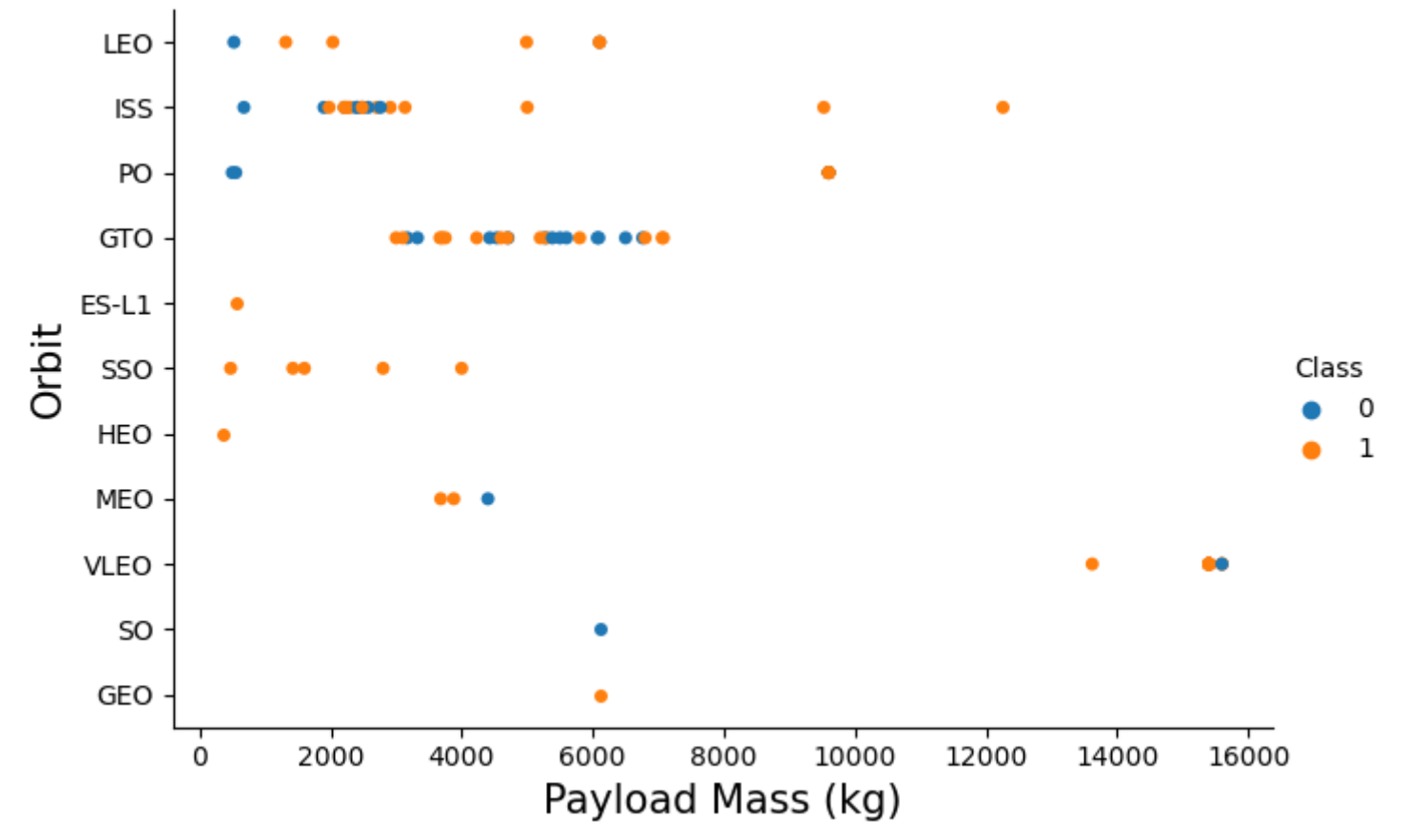
Flight Number vs. Orbit Type



- Class 0 represents the unsuccessful launches
- Class 1 represents successful launches

- LEO orbit increase its success with the number of flights.
- SSO has a 100% rate of success.
- No relationship can be found considering GTO orbit. In general, success rate increase with number of flights.

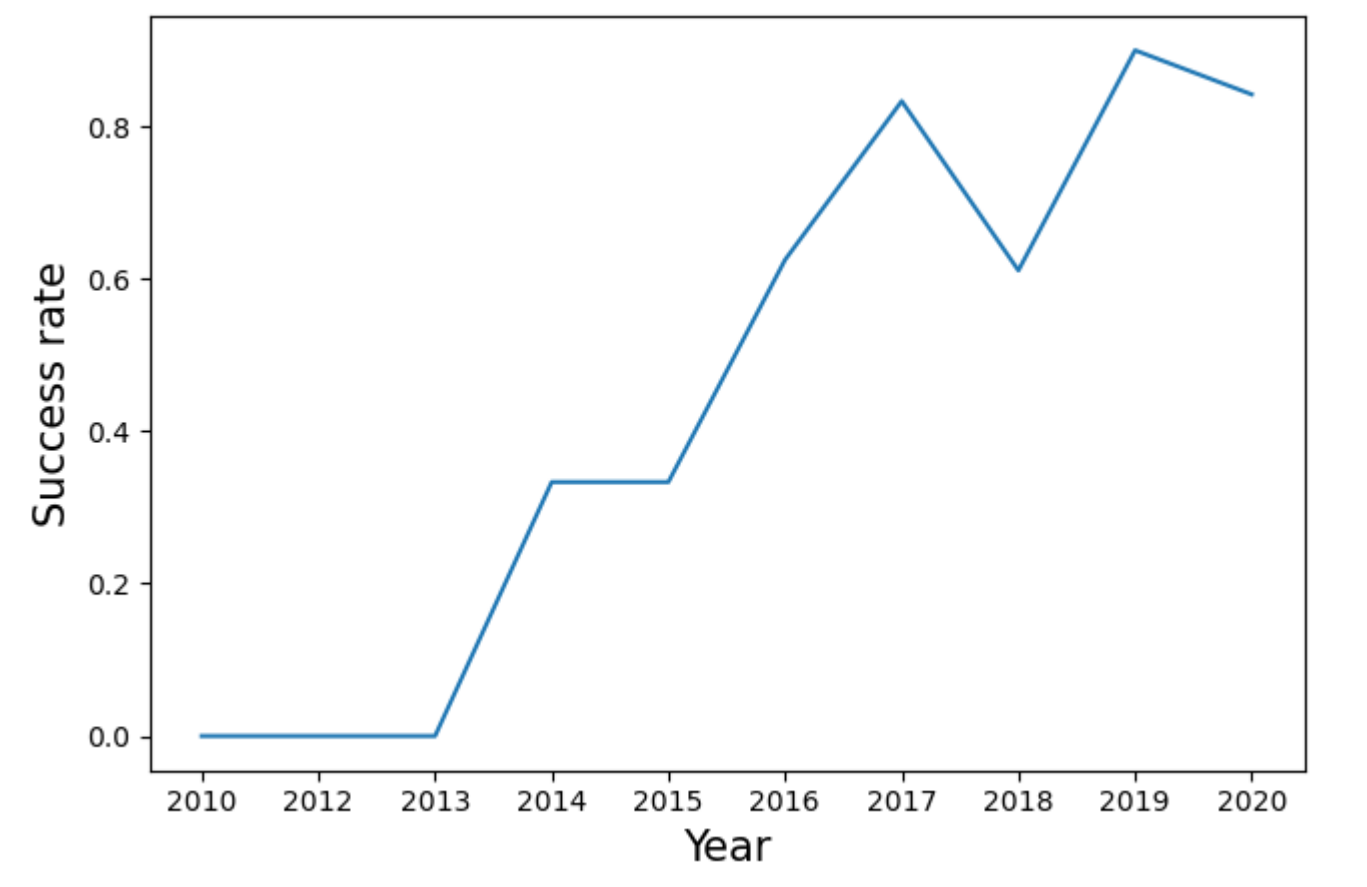
Payload vs. Orbit Type



- Class 0 represents the unsuccessful launches
- Class 1 represents successful launches

- Success landing rate increases with heavy payloads for LEO, ISS and Polar.
- We cannot conclude for GTO.

Launch Success Yearly Trend



The success rate has a general increasing shape from 2013 to 2020.

All Launch Site Names

Using DISTINCT in the SQL query, we select all launch sites.

They are:

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

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

Launch Site Names Begin with 'CCA'

The data displays 5 records where launch sites begin with `CCA`.

The SQL query uses LIKE in the WHERE clause and LIMIT.

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

The total payload mass (in Kg) carried by boosters from NASA is displayed below.

The SQL query uses SUM and a WHERE clause to restrict customer to ‘NASA (CRS)’.

total_payload_mass_by_nasa_crs
45596

Average Payload Mass by F9 v1.1

The average payload mass (in Kg) carried by booster version F9 v1.1 is displayed below.

The SQL query uses AVG and a WHERE clause to restrict the booster version to ‘F9 v1.1’.

avg_payload_mass_by_f9_v11
2928

First Successful Ground Landing Date

The date (in YYYY-MM-DD) of the first successful landing outcome on ground pad is displayed below.

The SQL query uses MIN and a WHERE clause to restrict landing outcome to ‘Success (ground pad)’.

first_successful_landing_date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

List of the names of booster versions which have successfully landed on drone ship and with a payload mass greater than 4000 but less than 6000.

The SQL query uses a WHERE clause to restrict landing outcome to ‘Success (drone ship)’ and BETWEEN for payload mass interval.

booster_version	payload_mass__kg_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

Total Number of Successful and Failure Mission Outcomes

Total number of successful and failure mission outcomes. The successes are 99 plus 1 with unclear payload status; failure is only 1.

The SQL query uses COUNT to compute the amount of mission outcomes and GROUO BY to group with respect to mission outcomes.

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

Boosters Carried Maximum Payload

List of the names of the booster versions which have carried the maximum payload mass (equal to 15.600 Kg).

The SQL query uses subquery with the MAX function in the WHERE clause to require that the payload carried is equal to the maximum.

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

List of the failed landing outcomes in drone ship, their booster versions, and launch sites names for the year 2015.

The list contains two different booster version, but all launches took place at CCAFS LC-40.

The SQL query uses LIKE to select failure and the YEAR function to filter for 2015 in the WHERE clause.

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

landing_outcome	COUNT
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

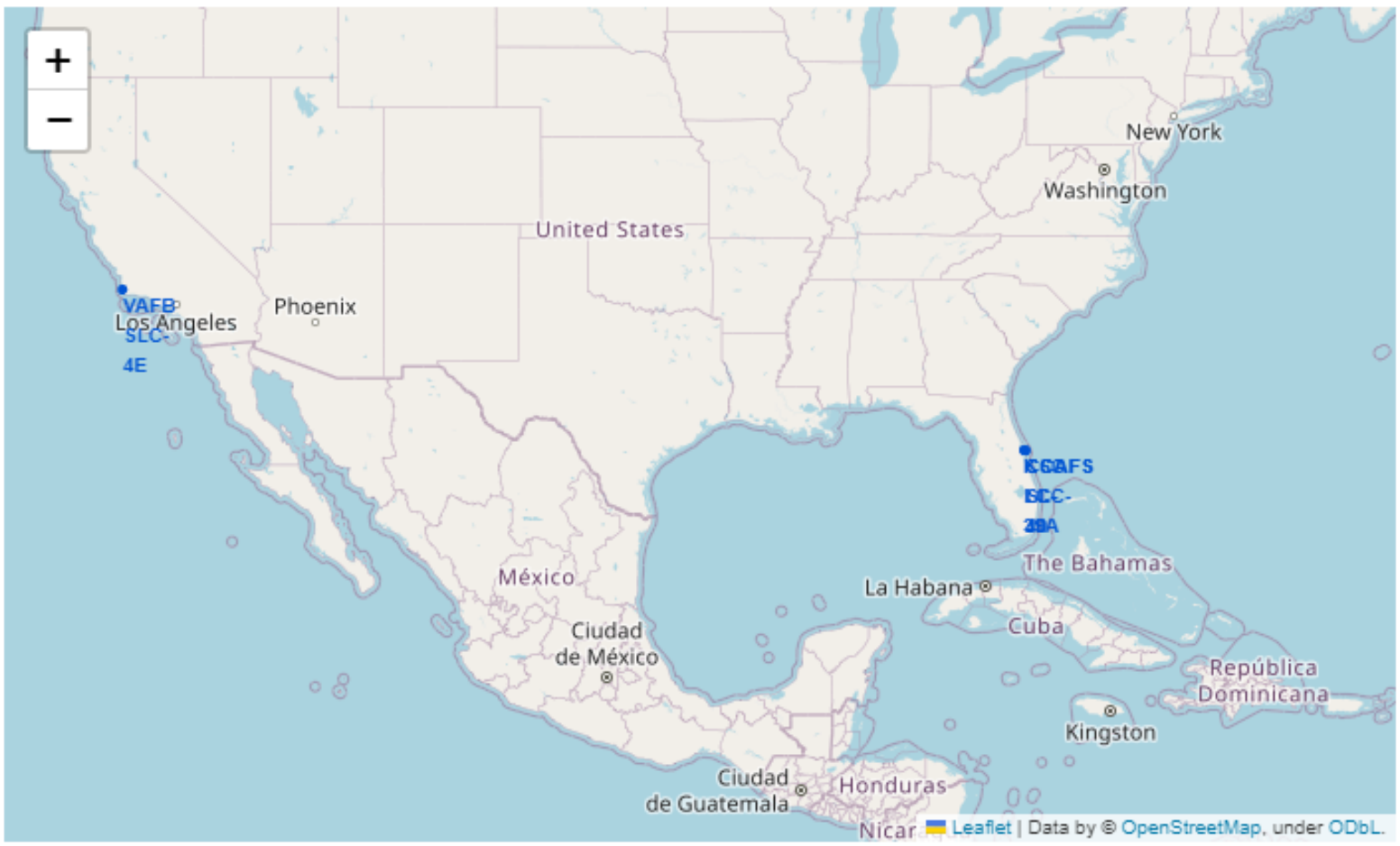
Descendent rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.

The SQL query COUNT to compute the amount of landing outcomes. Moreover, it uses BETWEEN to select the time interval in the WHERE clause, GROUP BY to group by landing outcomes and the ORDER request in DESC mode.

Section 3

Launch Sites Proximities Analysis

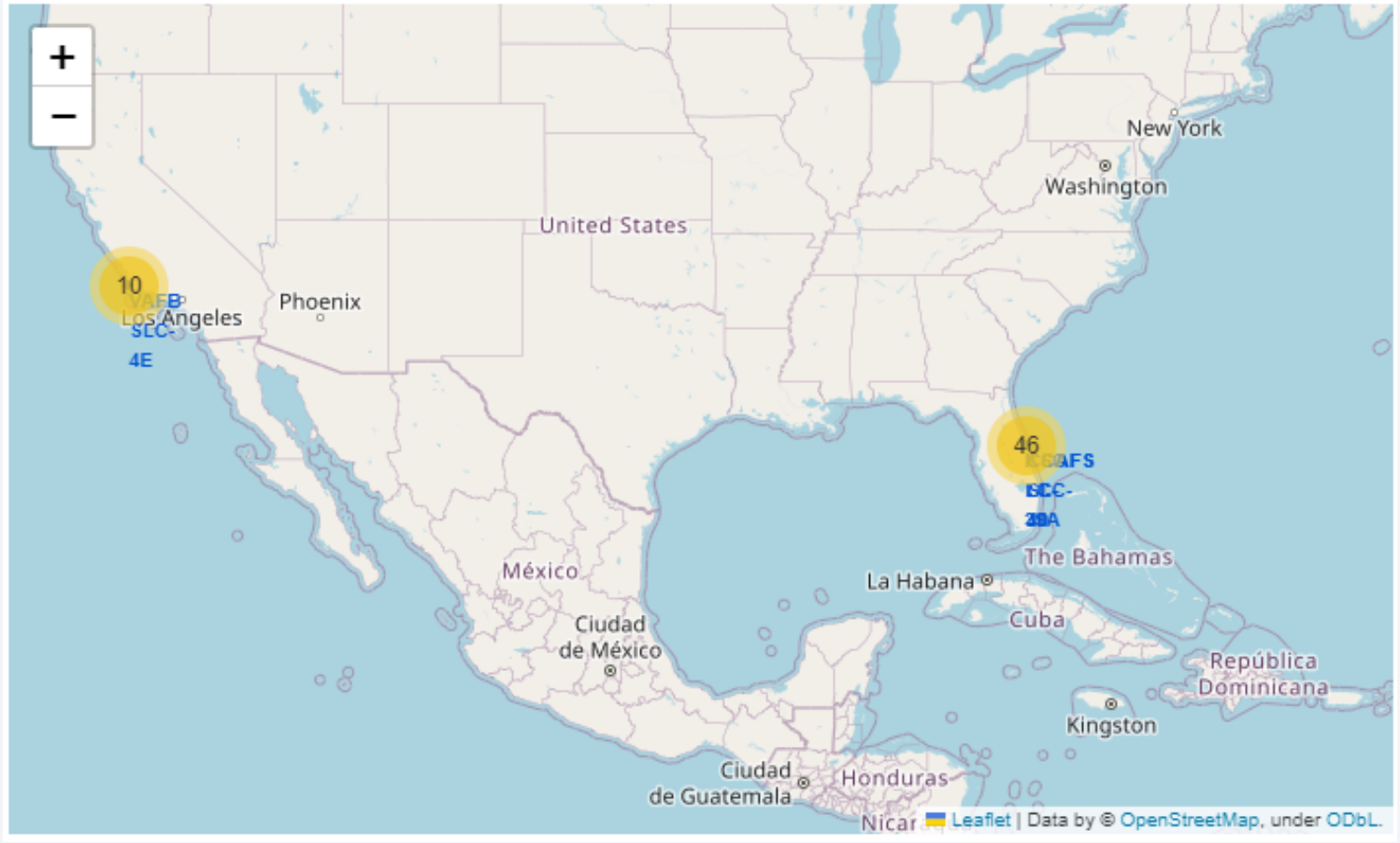
SpaceX Launch Sites



Blue markers indicate where launch sites are placed.

The map clearly show that all launch sites are positioned to be near the coast.

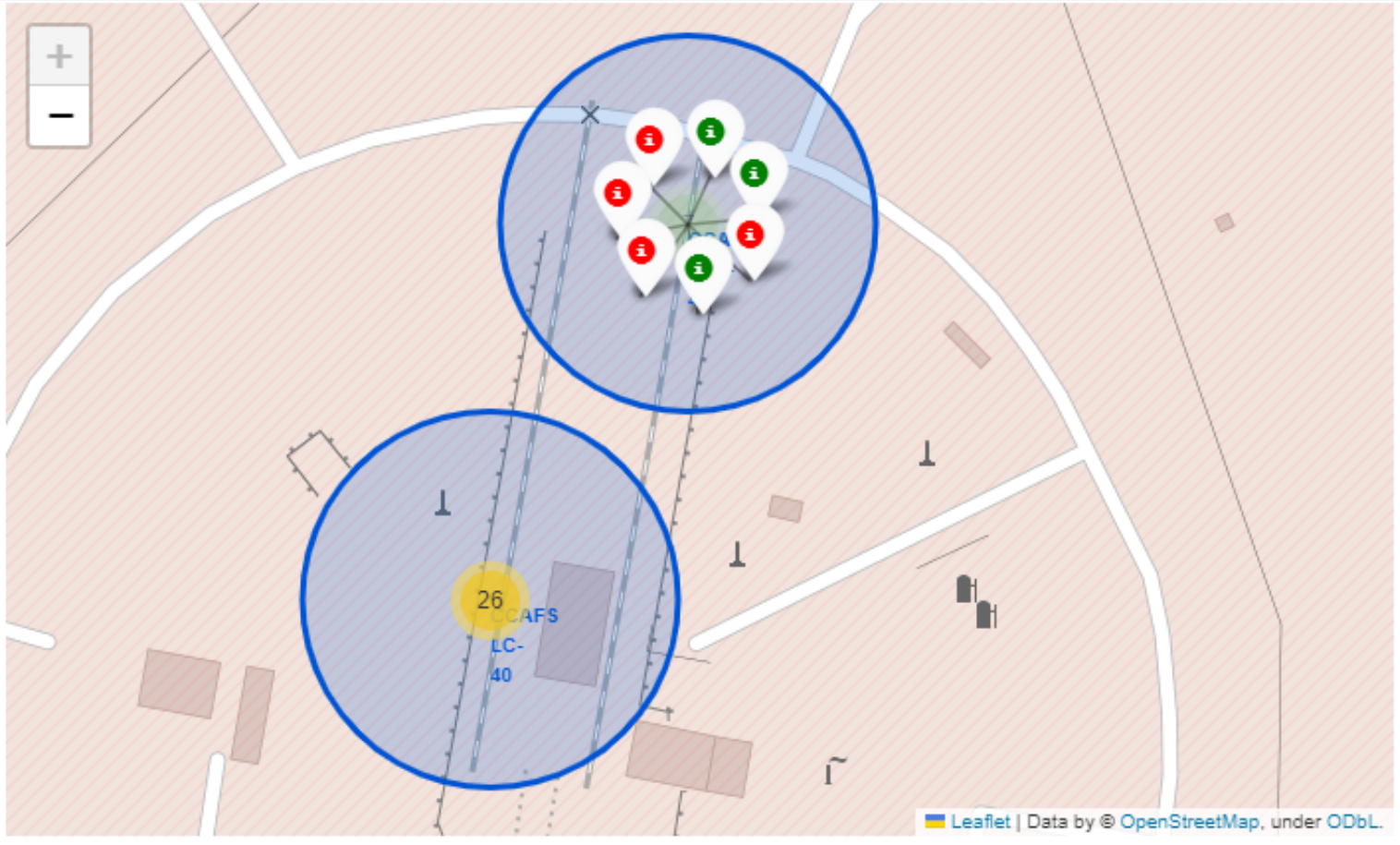
Success and Failure for Launch Sites



Yellow circles are added to launch sites. The circles display the number of launch for each region.

Zooming in the iterative map the number of launch become more precise by separating on specific different launch sites.

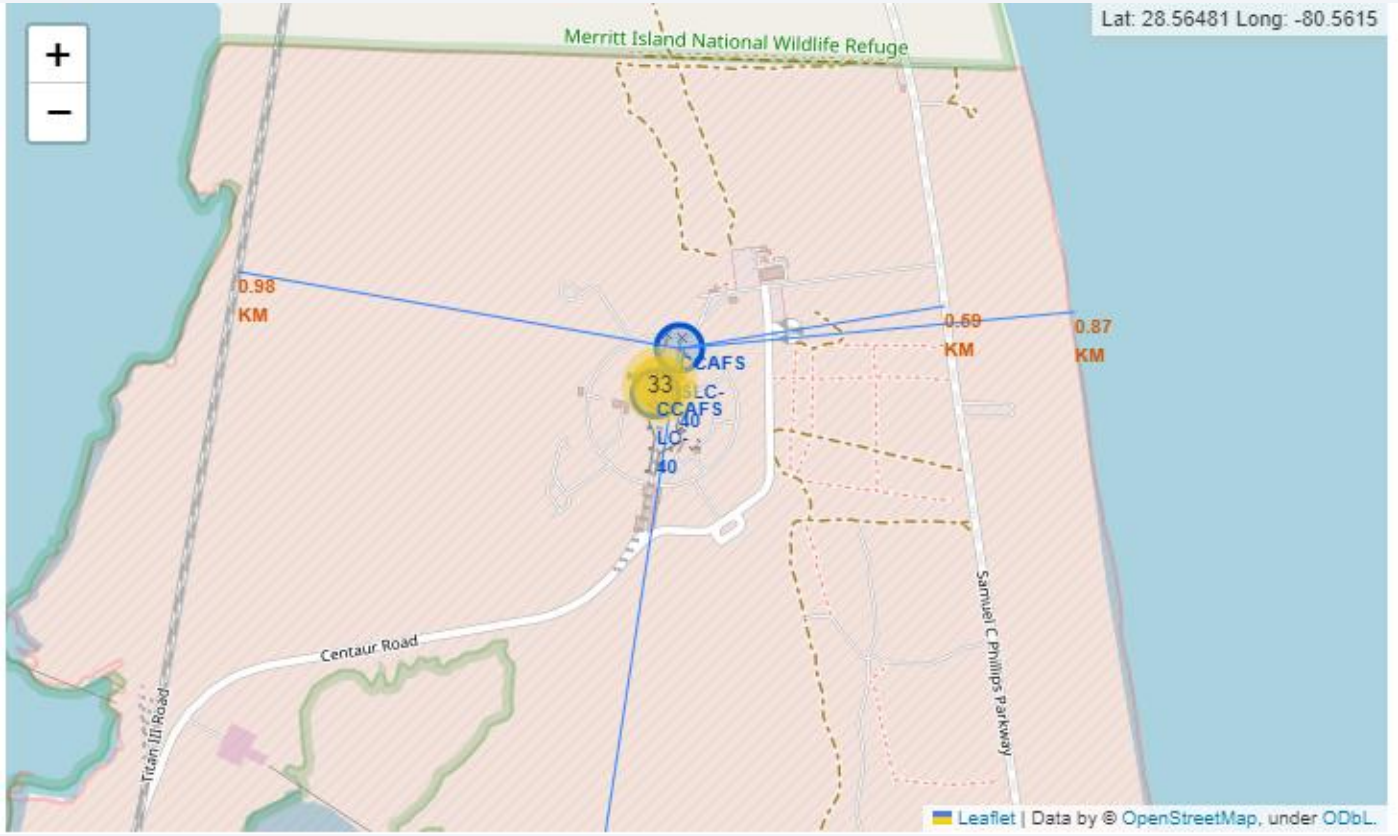
Success and Failure for Launch Sites



Keeping zoom in, the map display for each launch site a popup markers denoting whether launches were successful or not.

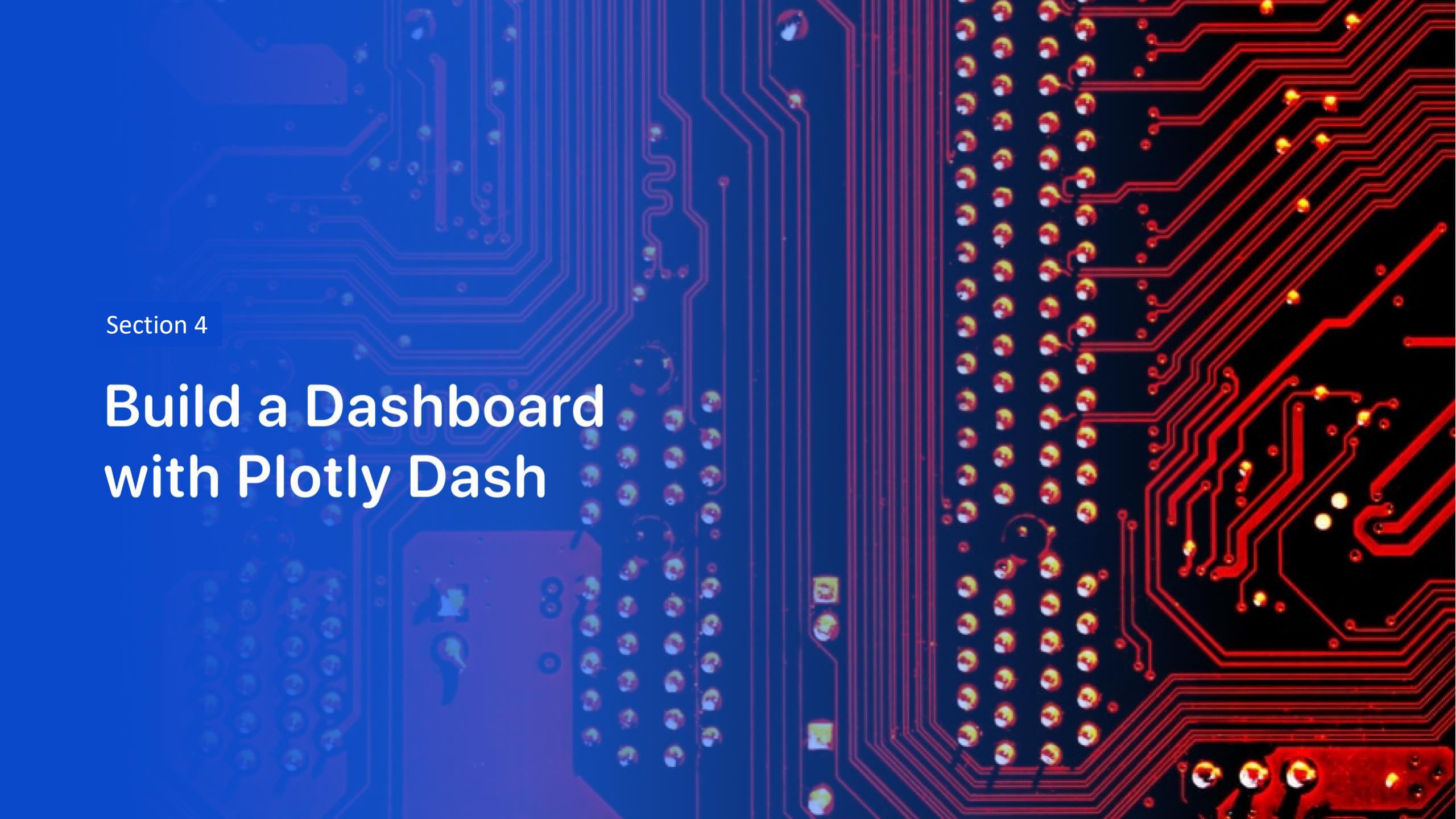
Red markers denote a failure while green denote a success.

Launch Sites Distances with Proximities



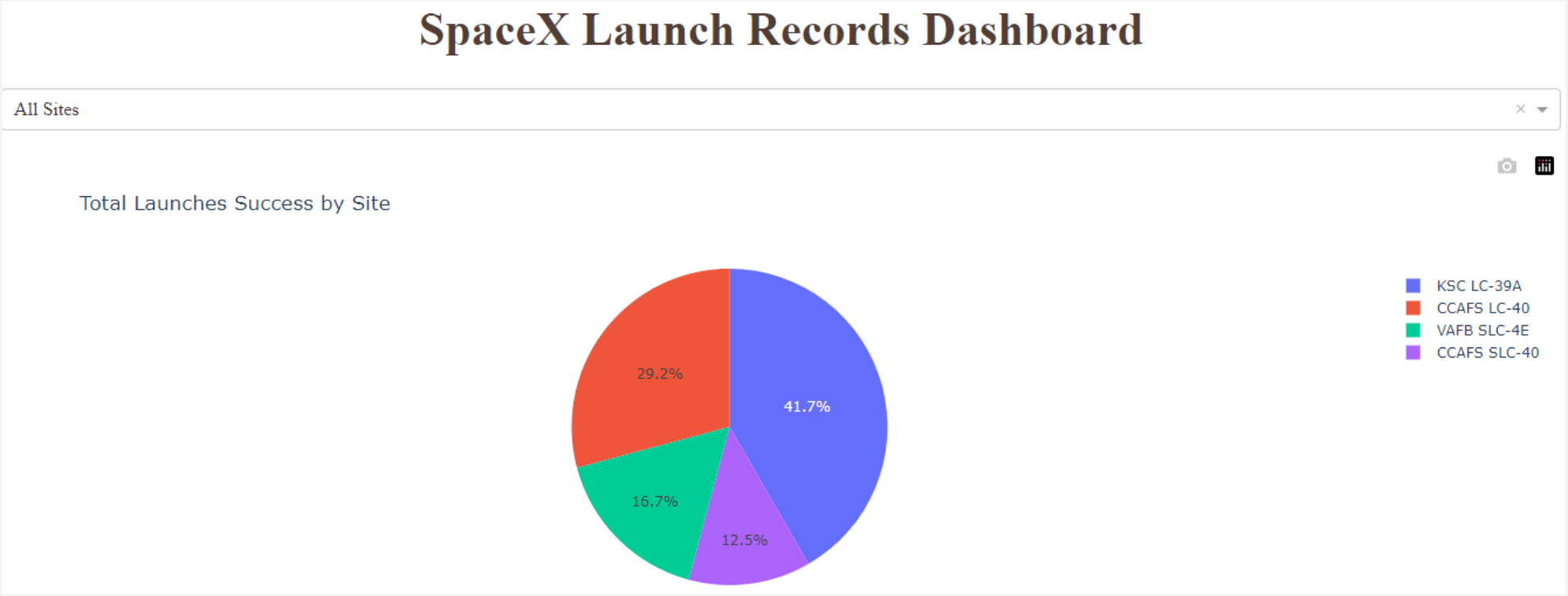
The map shows for a selected launch site the distances to its proximities: railway (0.98 Km), highway (0.59 Km), coastline (0.87 Km) and city (18.16 Km).

Launch sites are near highways and railways for transport reasons and distant from cities for security reasons.


Section 4
Build a Dashboard with Plotly Dash

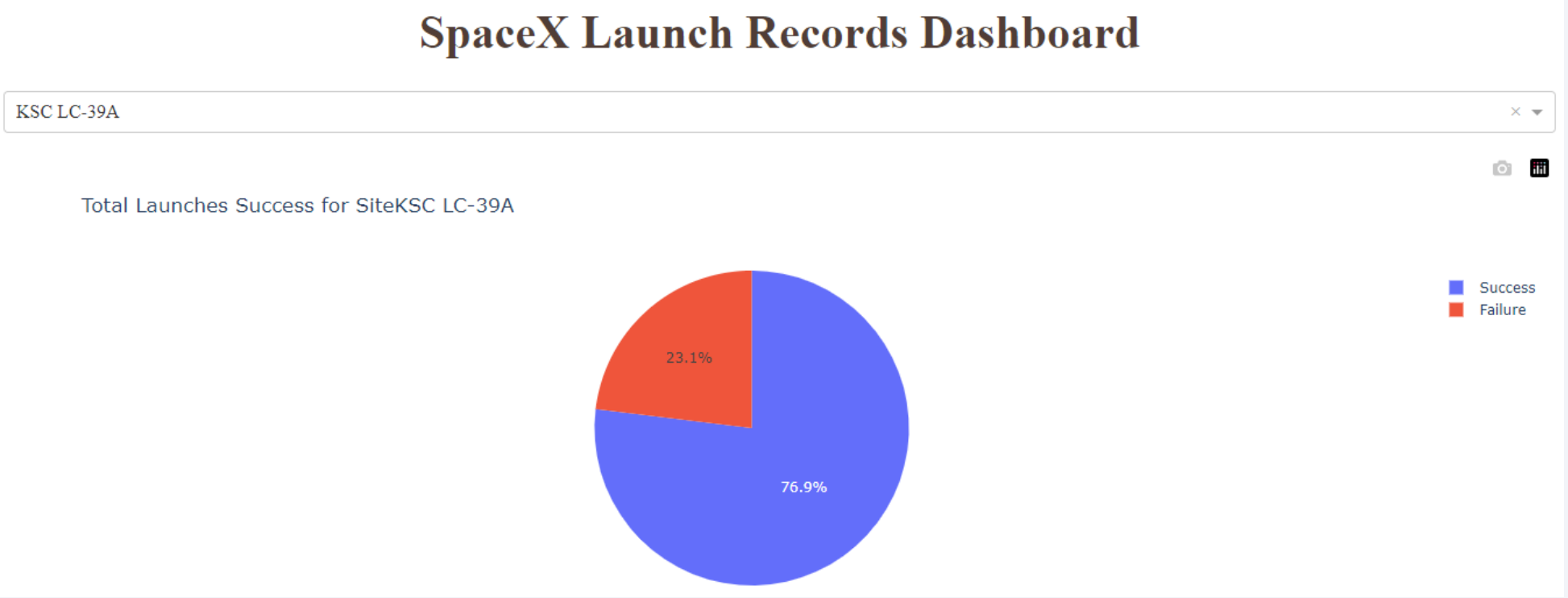
Total Launches Success by Site

The pie-chart show that KSC LC-39A has the higher success amount (41.7% of all successes); the worst is CCAFS SCL-40 with only the 12.5%.



Success vs Failure for KSC LC-39A

The pie-chart display for KSC LC-39A (the launch site with highest launch success ratio) its percentage of success (76.9%) and of failure (23.1%).



Payload vs. Launch Outcome

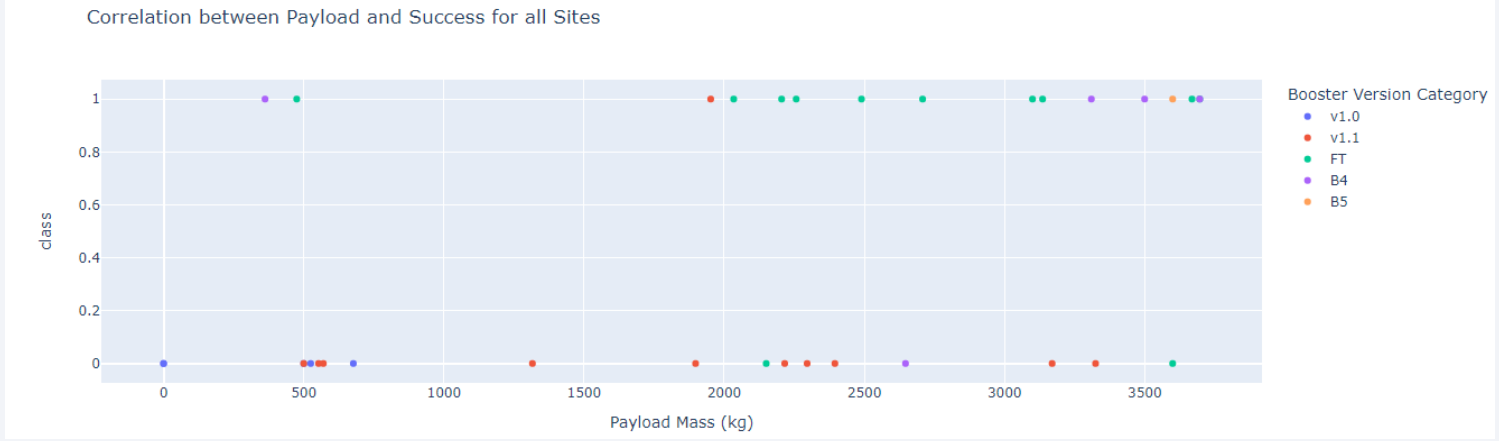
We display a payload mass vs. launch outcome scatter plot for all launch sites.

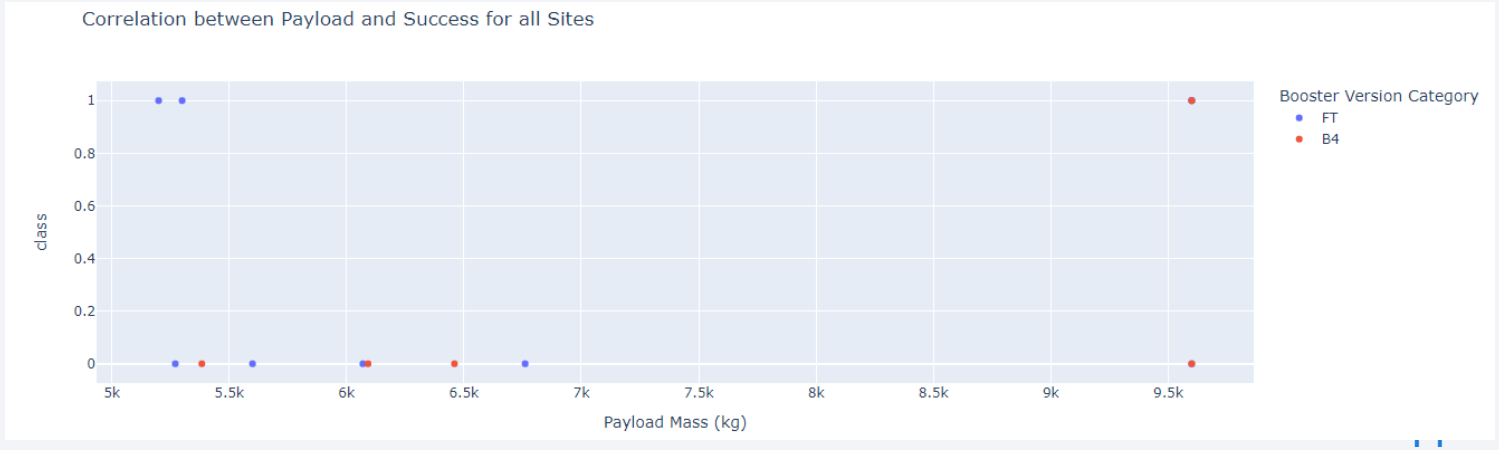
Payload mass: 0 - 4000 Kg.

A good success rate for FT and B4 booster version is shown.

Payload mass: ≥ 5000 Kg.

Only FT and B4 booster versions are considered. The general success is lower than above.





Section 5

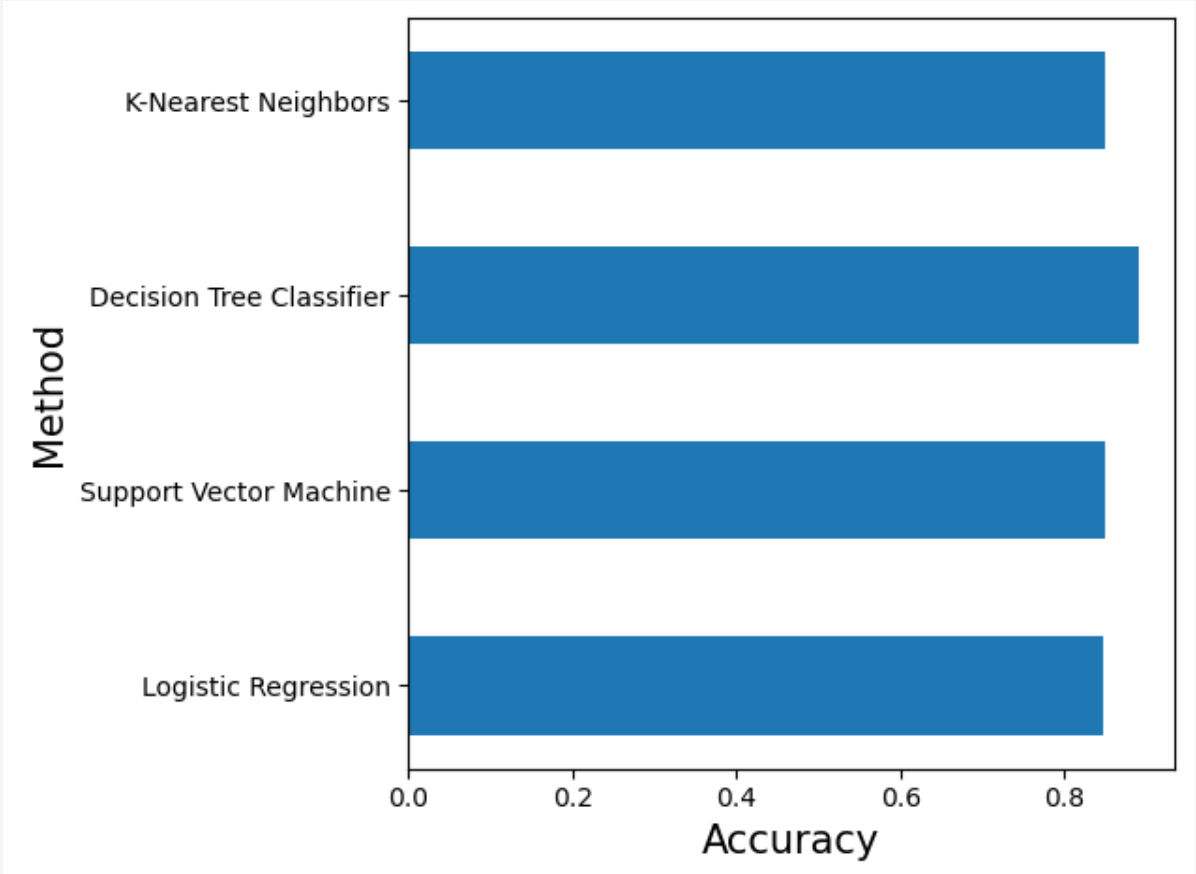
Predictive Analysis (Classification)

Classification Accuracy

We compare the accuracy of the following methods:

- Logistic Regression
- Support Vector Machine
- Decision Tree Classifier
- K-Nearest Neighbor.

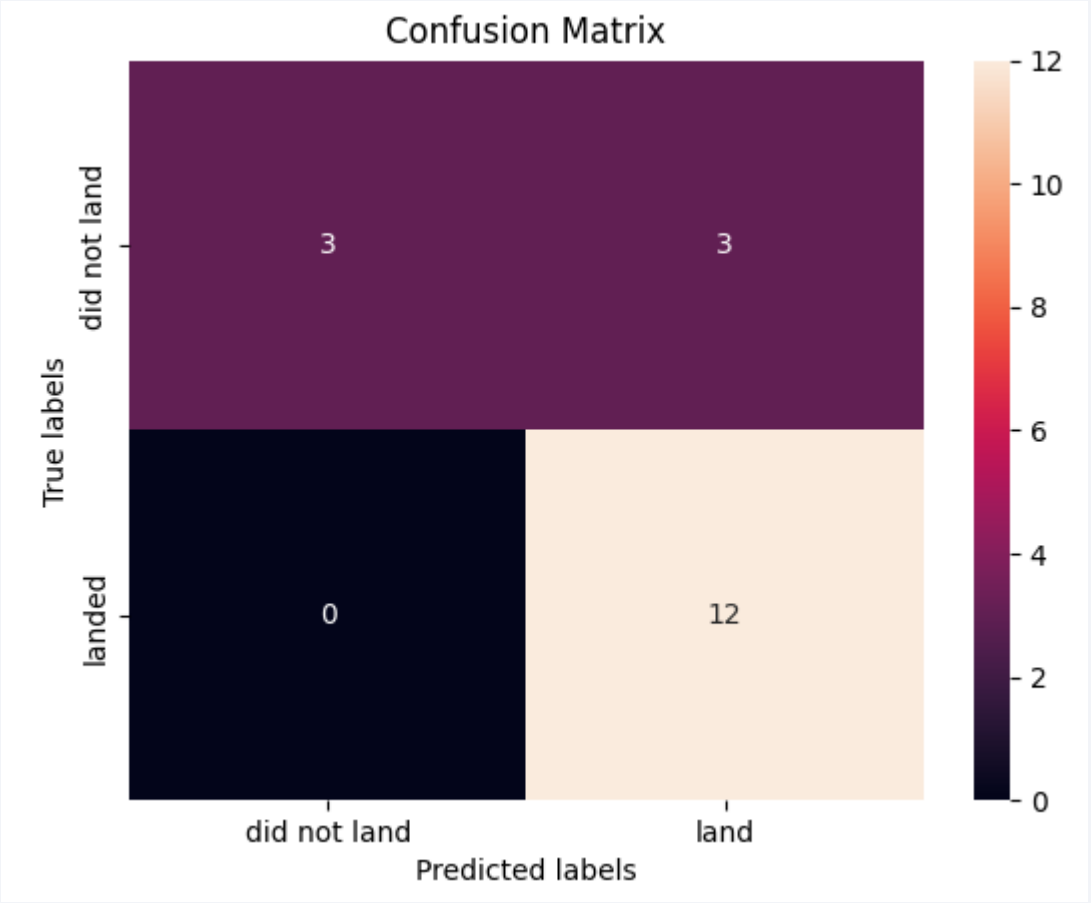
The method with the highest classification accuracy is Decision Tree with an accuracy of 0.89.



Confusion Matrix

Here displayed the confusion matrix for the best performing model.

The model predict correctly all the 12 successful landing (True Positive), while it predicts correctly only 3 over 6 unsuccessful landings (True Negative) and fails by predicting successful 3 unsuccessful landings (False Positive).



Conclusions

- The analysis verifies an increasing success rate trend for rocket launches over years. The success is also positively correlated to the number of flight.
- The orbits where launches were the most successful ES-L1, GEO, HEO and SSO.
- The launch site with highest launch success ratio is KSC LC-39A.
- The success rate seems to be correlated to payload mass: lower payload masses correspond to a higher success rate.
- The launch sites are placed near transportation infrastructure and far away from cities for safety reasons.
- The best predictive method is the Decision Tree Classifier with an accuracy of 89%.

Appendix

- GitHub Repository: https://github.com/ChiaraBrambilla96/Final-Assignment-Data-Science

Thank you!