IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Evgeny Pinigin
February 13, 2024

Table of contents

Executive Summary

- The goals of this project were Data analysis of SpaceX Falcon 9 launches and Machine learning prediction of first stage successful landing. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.
- The project contains some stages:
  - Data collection. Data was collected from SpaceX REST API and via Web scraping of Wikipedia Falcon 9 page.
  - Data wrangling. The different true and false outcomes were converted to the binary outcome label "Class". Furthermore, all categorical variables were converted to binary variables.
  - Exploratory data analysis (EDA) using visualization and SQL.
  - Interactive visual analytics using Folium map and Plotly Dash.
  - Predictive Machine Learning analysis. Four classifiers were used: The Logistic Regression, SVM, Decision Trees, and K-nearest neighbours.
- Successful mission outcome rate is almost 100%, but there are still problems with successful landings, which rate is only 66.67%.
- From 2010 to 2012 all landings were failed. But the success landing rate since 2013 kept increasing till 2020 and is about 80%.

Executive Summary

- For the VAFB-SLC launch site there are no rockets launched for payload mass greater than 10000 kg.
- ES-L1, GEO, HEO and SSO orbits have 100% success landing rate, whereas SO orbit's landing rate is zero.
- In the LEO orbit the success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in other orbits.
- With heavy payloads the successful landing rate are more for PO, LEO and ISS. However, for GTO we cannot distinguish this well as both positive landing rate and negative landing are both there here.
- The Falcon 9 missions were launched from four sites: three launch sites are on the Cape Canaveral, and one is in California.
- All launch sites are away from cities but have railways and highways. Furthermore, all launch sites are close to ocean coastline.

Executive Summary

- The created interactive Dashboard is useful for stakeholders and tells them about percentage and count of successful and failed landings for every launch site along with a scatter plot of relationship between Payload mass and Landing outcomes.
- KSC LC-39A launch site has the largest successful launches and the highest landing success rate.
- The landing outcome depends on payload mass: the payload mass range 2700-3700 kg has the highest landing success rate, whereas a mass range 5380-6800 kg has the lowest landing success rate.
- The Logistic Regression, SVM, Decision Trees, and K-nearest neighbours models give the same accuracy of 0.83 with current dataset of 101 launches. The false negatives are zero, but the false positives are 20%.
- The testing accuracy of Decision Tree classifier usually is 0.83. But it changes randomly upon repeated fitting launches from 0.5 to 0.94. The reason is random model parameters changes.
- Over time the launches dataset will be increased, as a result the model accuracy will be increased also after repeated training.

Introduction

- SpaceX promotes Falcon 9 rocket launches on its website at a price of $62 million, significantly undercutting other providers whose costs soar above $165 million per launch. A significant portion of this price advantage stems from SpaceX's ability to reuse the first stage of the rocket.
- Thus, by assessing the likelihood of successful first stage recovery, one can estimate the cost of a launch. This insight could prove invaluable for competing companies seeking to bid against SpaceX for rocket launch contracts.

- The goal of this project is a developing of a prediction model that could be used for estimating the probability of successful first stage recovery.

Section 1

Methodology

Methodology

Executive Summary

- Data was collected from SpaceX REST API and via Web scraping of Wikipedia Falcon 9 page
- At a data wrangling stage, the different true and false outcomes were converted to the binary outcome label "Class"
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis. Normalized the independent variables, split the data into training and test sets, fitted and tuned some classification models, found the best classification method.

Data Collection

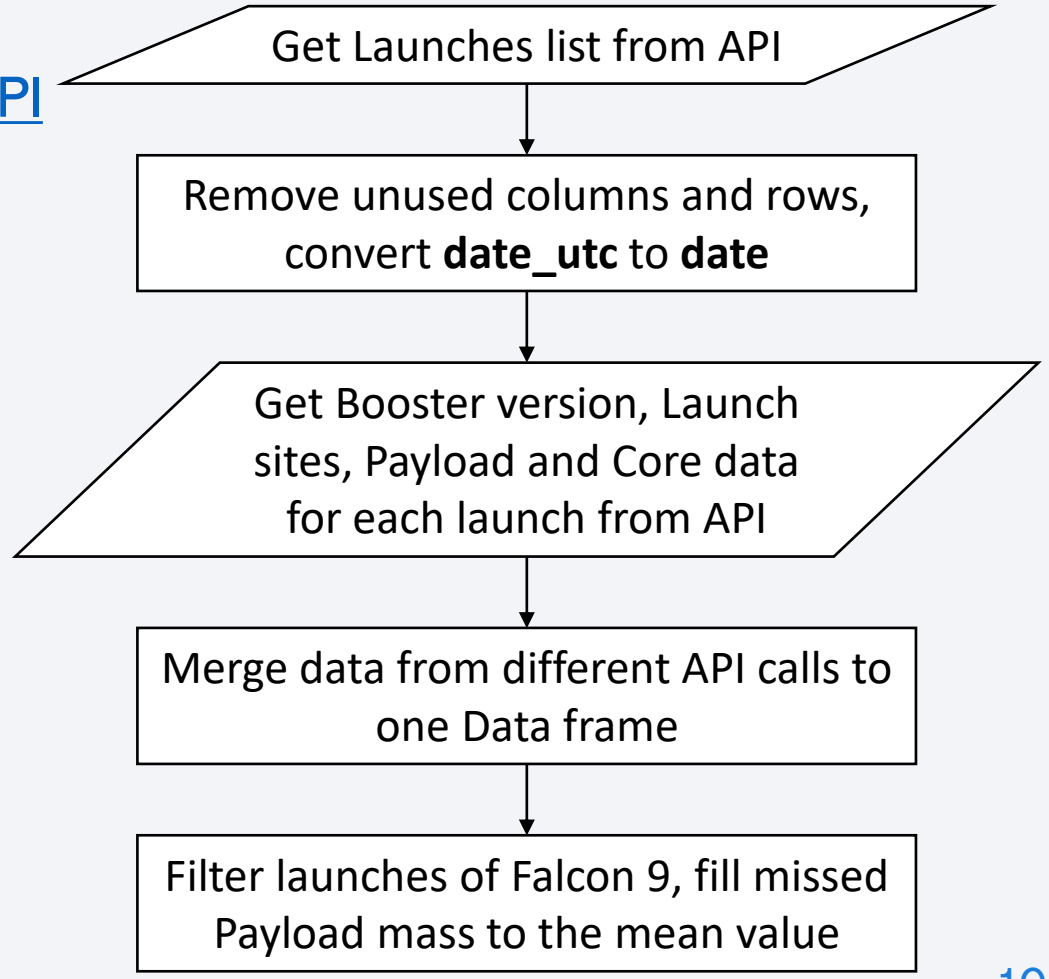
Data sets were collected using two different methods:

- Collecting from SpaceX REST API
- Collecting using web scraping from Wikipedia Falcon 9 page

Collecting from REST API is preferable because the received data is uniform and does not need additional parsing and cleansing

Data Collection – SpaceX API

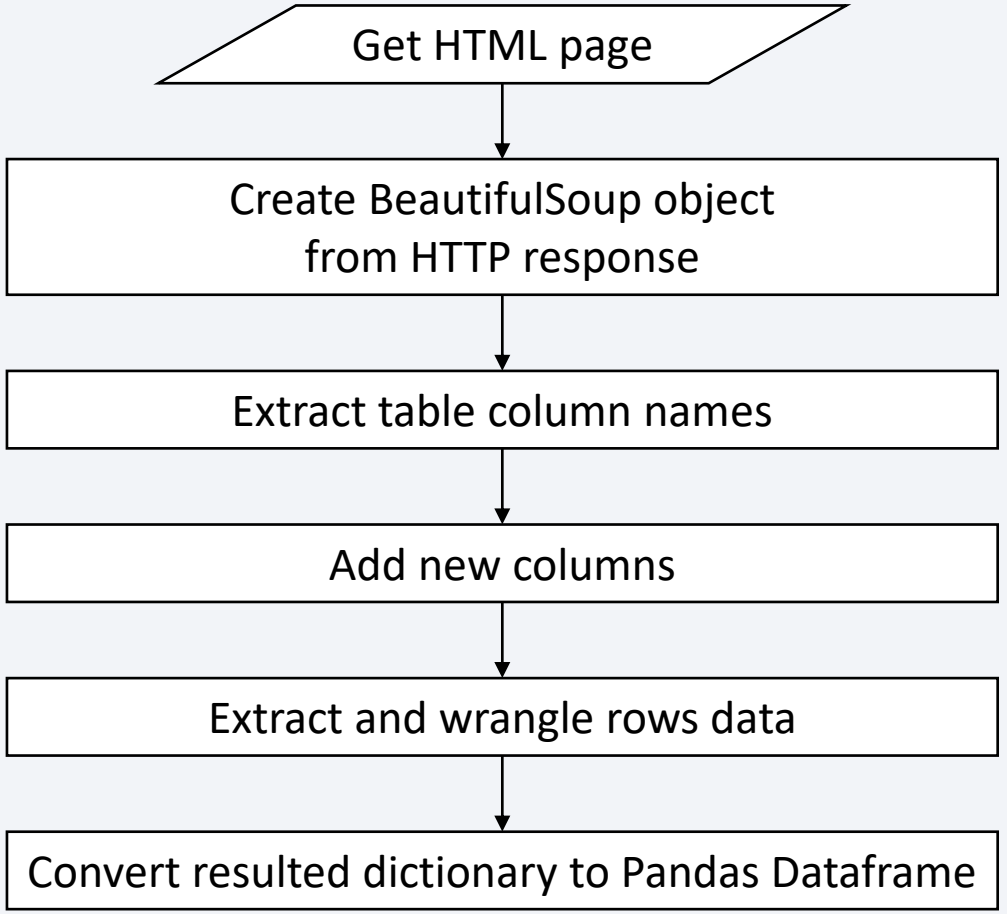
- The data was collected from SpaceX REST API
- Request, Pandas and Numpy libraries were used
- Notebook in Github



Data Collection - Scraping

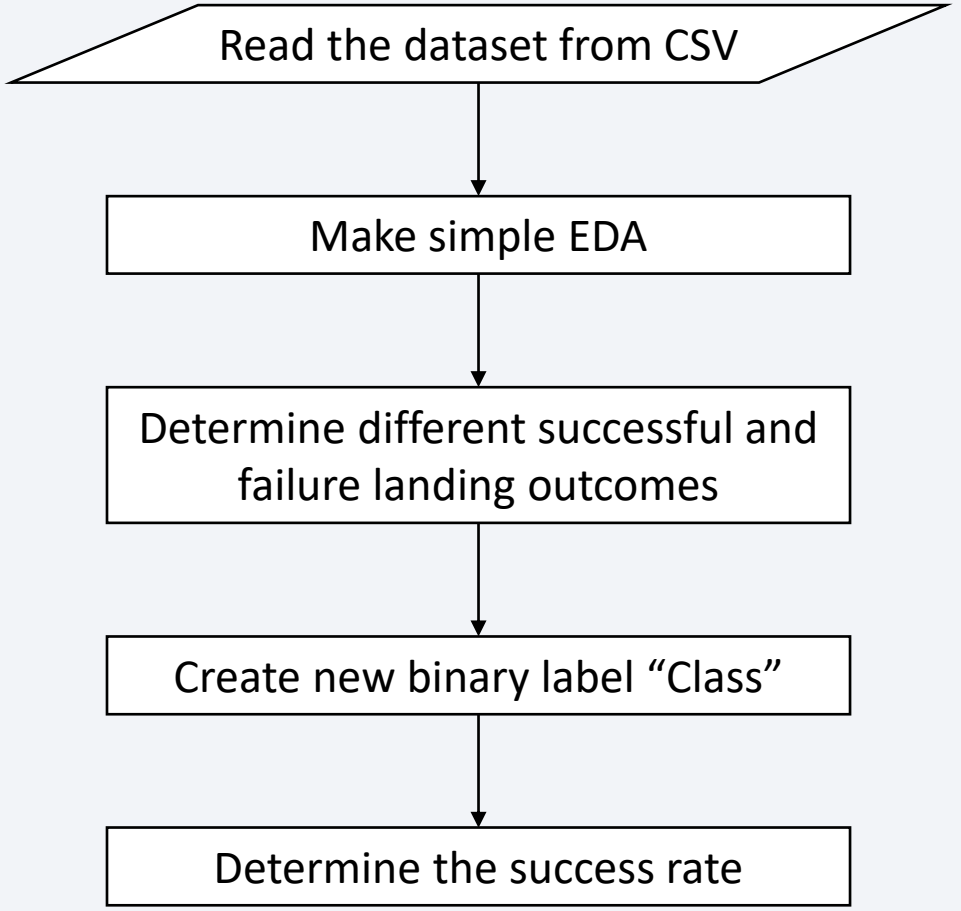
- The data was collected from Wikipedia
- Request, BeautifulSoup and Pandas libraries were used
- The data contains annotations, missed values and noises and need to be cleaned

- Notebook in Github



Data Wrangling

- Some simple exploratory data analysis were made:
  - Percentage of the missing values in each column
  - Column types
  - Number of launches on each site
  - Number of each used orbit
- The various spelling of success and failure in “Outcome” column were converted to the new binary column “Class”
- The success rate was determined



- Notebook in Github

EDA with Data Visualization

- Scatter plots, a bar chart and a line chart were plotted for EDA with Data Visualization using Seaborn library.
- These charts were used for determining relationships between different variables and the launch outcome.
- Besides, the feature engineering was made: all categorical variables were converted to binary via one hot encoding.

The plotted charts

1. Scatter plot “Flight Number vs Payload Mass”
2. Scatter plot “Flight Number vs Launch site”
3. Scatter plot “Payload Mass vs Launch Site”
4. Bar plot “Orbit type vs Success rate”
5. Scatter plot “Flight Number vs Orbit”
6. Scatter plot “Payload mass vs Orbit”
7. Line chart “Year vs Success rate”

- Completed EDA with data visualization notebook in Github

EDA with SQL

Different SQL queries were performed for EDA with SQL:

- The names of the unique launch sites in the space mission.
- The 5 records where launch sites begin with the string 'CCA'.
- The total payload mass carried by boosters launched by NASA (CRS).
- The average payload mass carried by booster version F9 v1.1.
- The date when the first successful landing outcome in ground pad was achieved.
- The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- The total number of successful and failure mission outcomes.
- The names of the booster versions which have carried the maximum payload mass.
- The month names, failure outcomes in drone ship, booster versions, launch sites for the months in year 2015.
- Ranking of the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.

Completed EDA with SQL notebook in Github

Build an Interactive Map with Folium

The Folium map with different objects was created:

- Every launch site was marked with Circle and Marker objects. Circle object is useful for displaying of a color circle near some coordinates. Marker object is used for adding text or graphic labels on a map.
- Every launch outcome was marked with graphical color Marker. MarkerCluster object was used to simplify a map containing many markers having the same coordinate.
- MousePosition object was used to get coordinate for a mouse over a point on the map. It is useful to find the coordinates of any points of interests.
- PolyLine and Marker objects were used to display the distances between a launch site and the closest objects such as city, railway, highway etc.

Completed interactive map with Folium map

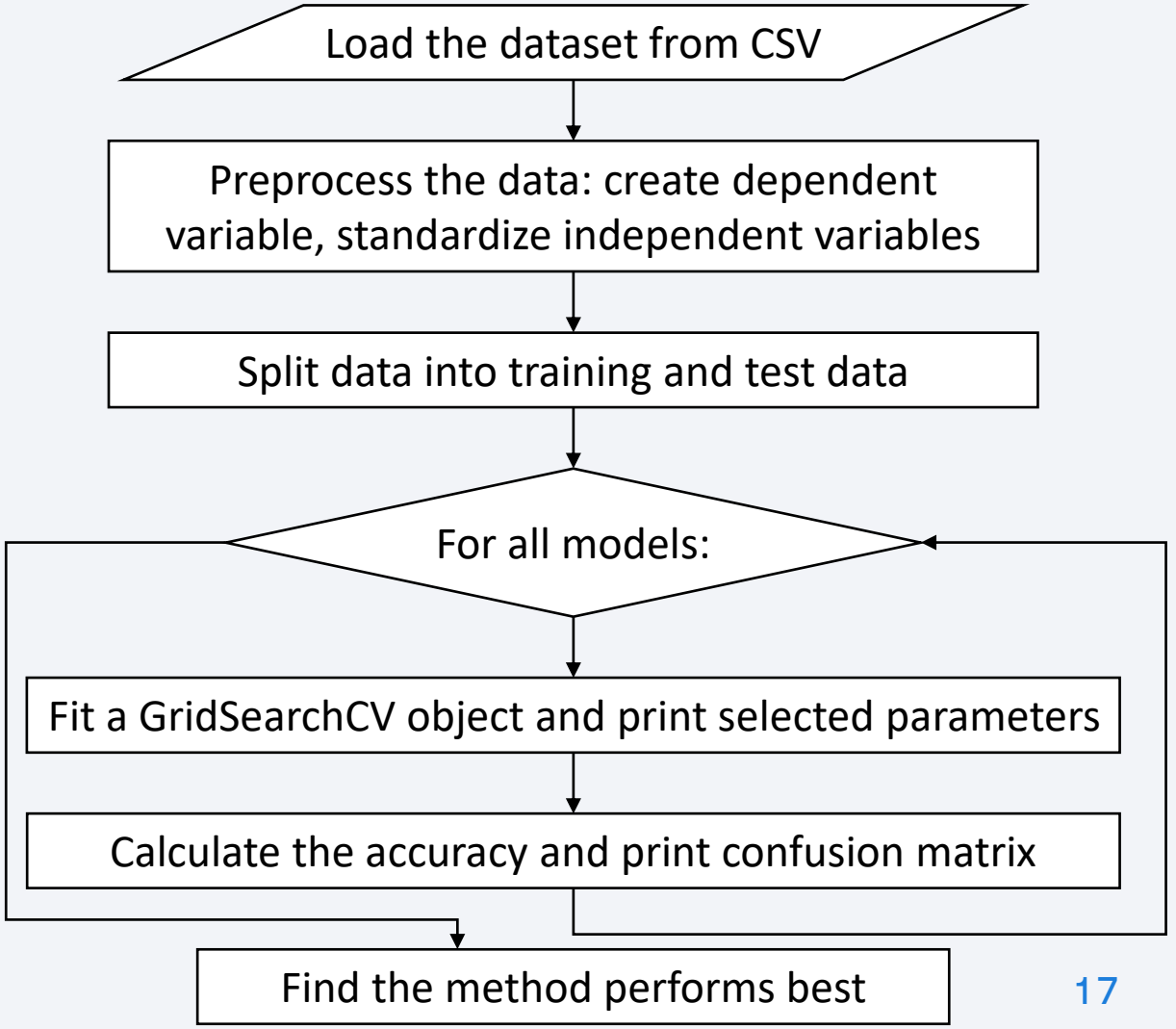
Build a Dashboard with Plotly Dash

- The pie chart and scatter plot were added to the dashboard.
- The launch sites dropdown list and payload mass range slider were added to the dashboard for interactivity.
- The scatter plot displays:
  - The total success launches for every launch site when ALL sites are selected.
  - The percentages of successful and failed launches when a specific launch site is selected.
- The scatter plot displays a relationship between Payload Mass and Launch Outcome(class) for all launch sites or selected launch site. The scatter plot data are filtered for selected Payload Mass range.

- Completed Plotly Dash lab in Github

Predictive Analysis (Classification)

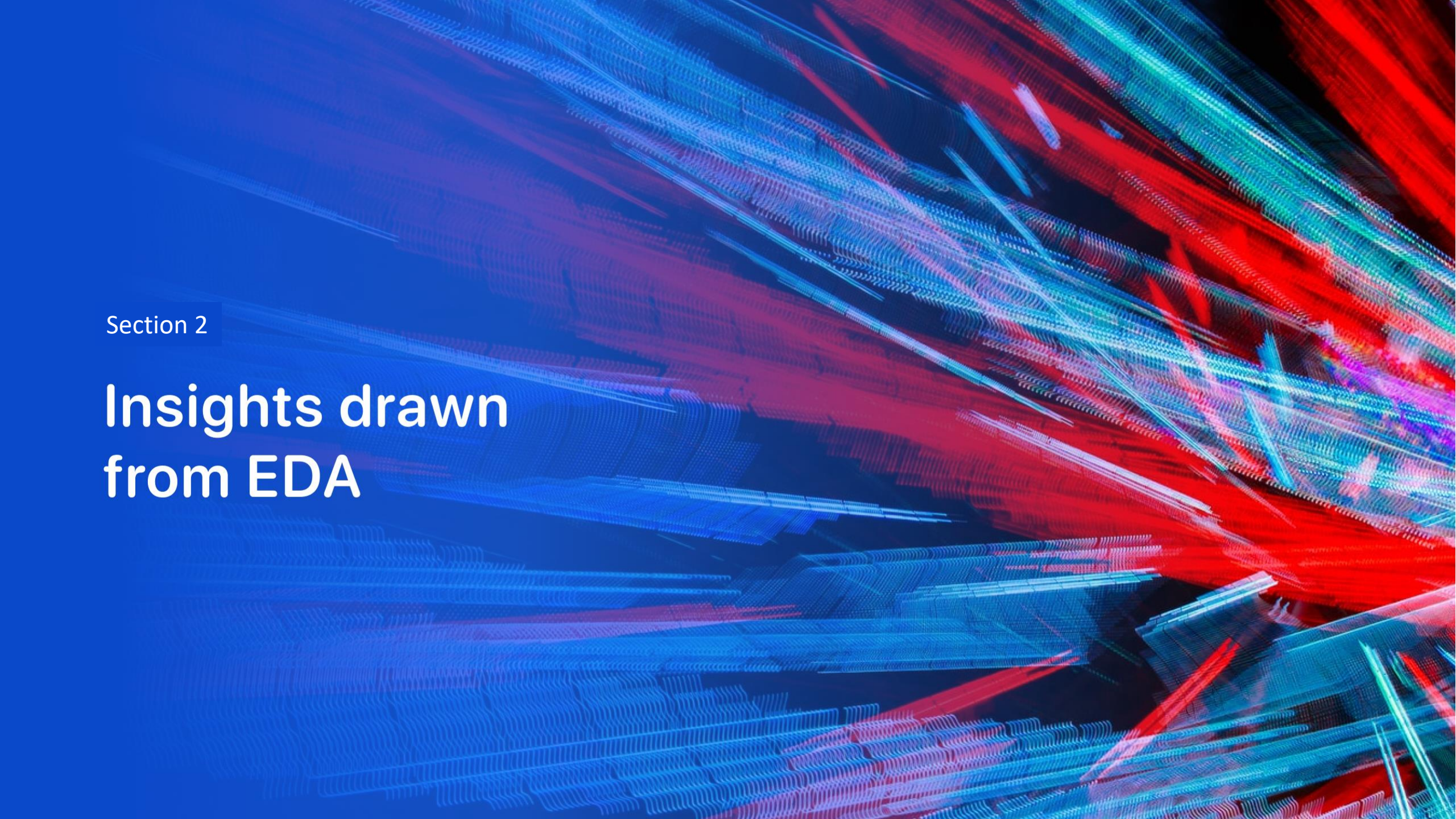
- Scikit-learn, Pandas and Numpy libraries are used for a model building.
- Matplotlib and Seaborn are used for visualizing a confusion matrix.
- Before fitting estimators, the independent variables are standardized via Scikit-learn StandardScaler.
- Four classification models are checked for best performing: Logistic Regression, SVM, Decision Trees and K-nearest Neighbours.
- GridSearchCV is used for model parameters optimization.
- Completed predictive analysis notebook in Github



Results

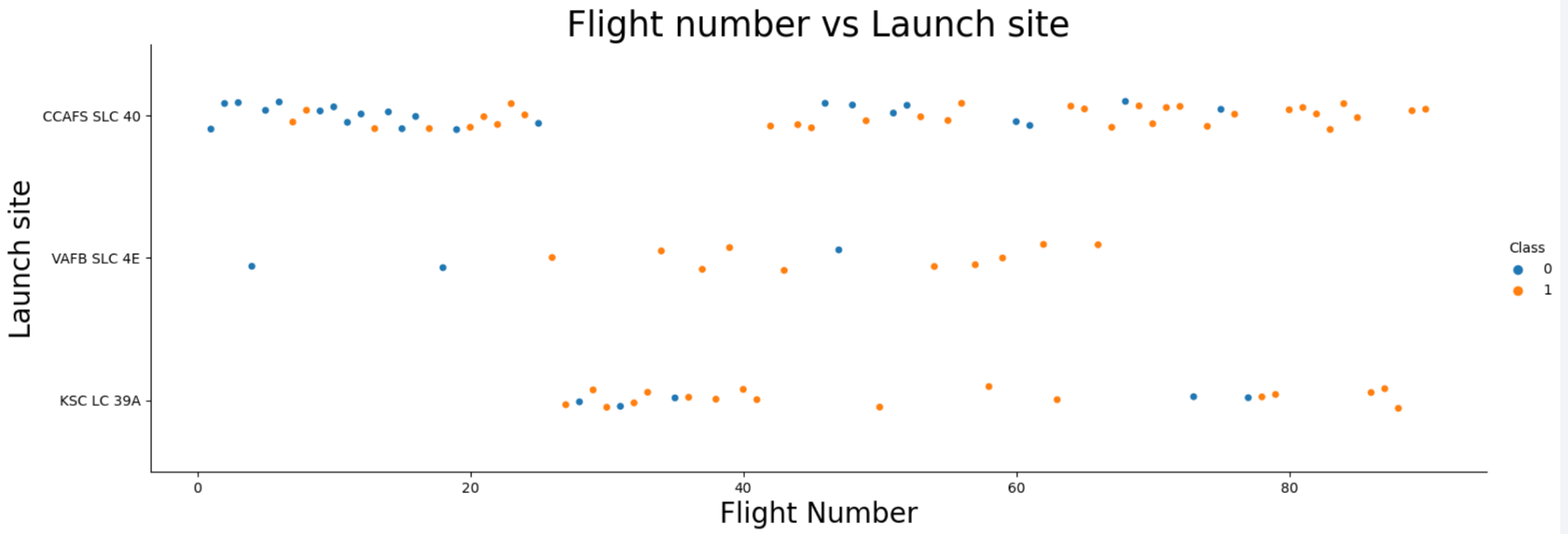
On the next slides we will see the project results:

- Exploratory data analysis results
- Launch site proximities analysis
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn from EDA

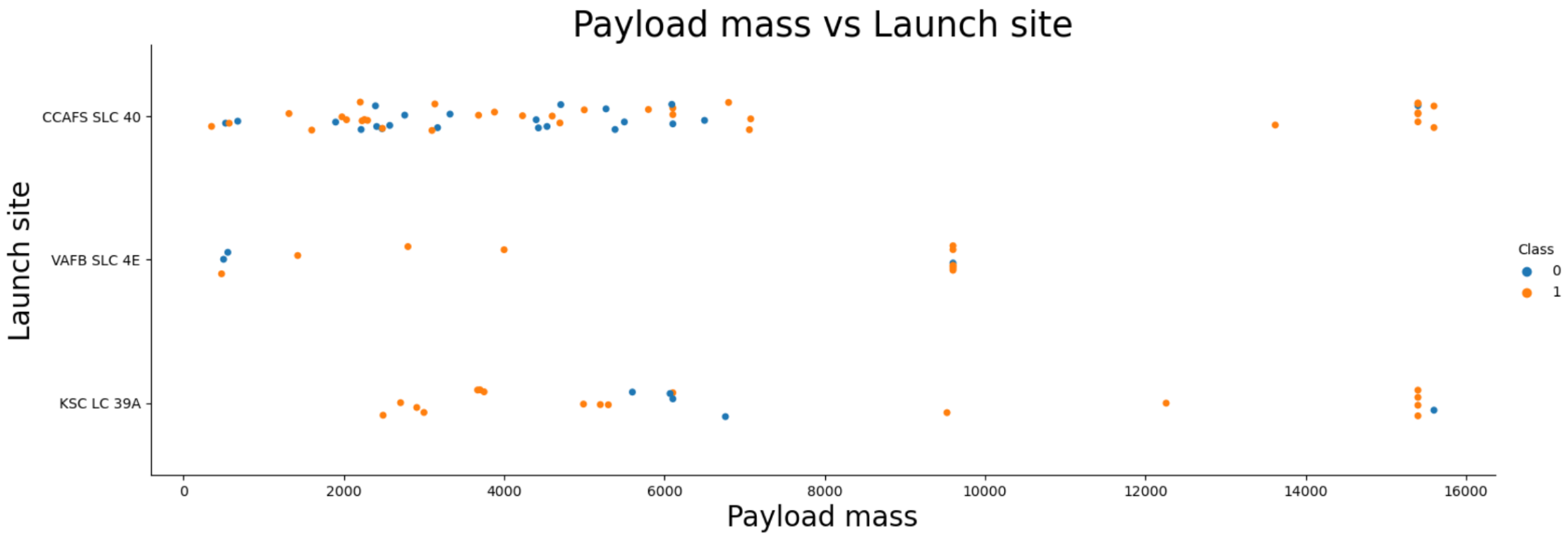
Flight Number vs. Launch Site

For KSC LC 39A Launch site there is no strong relationship between a flight number and a success rate. But for the other launch sites the success rate is increased with a flight number increasing.



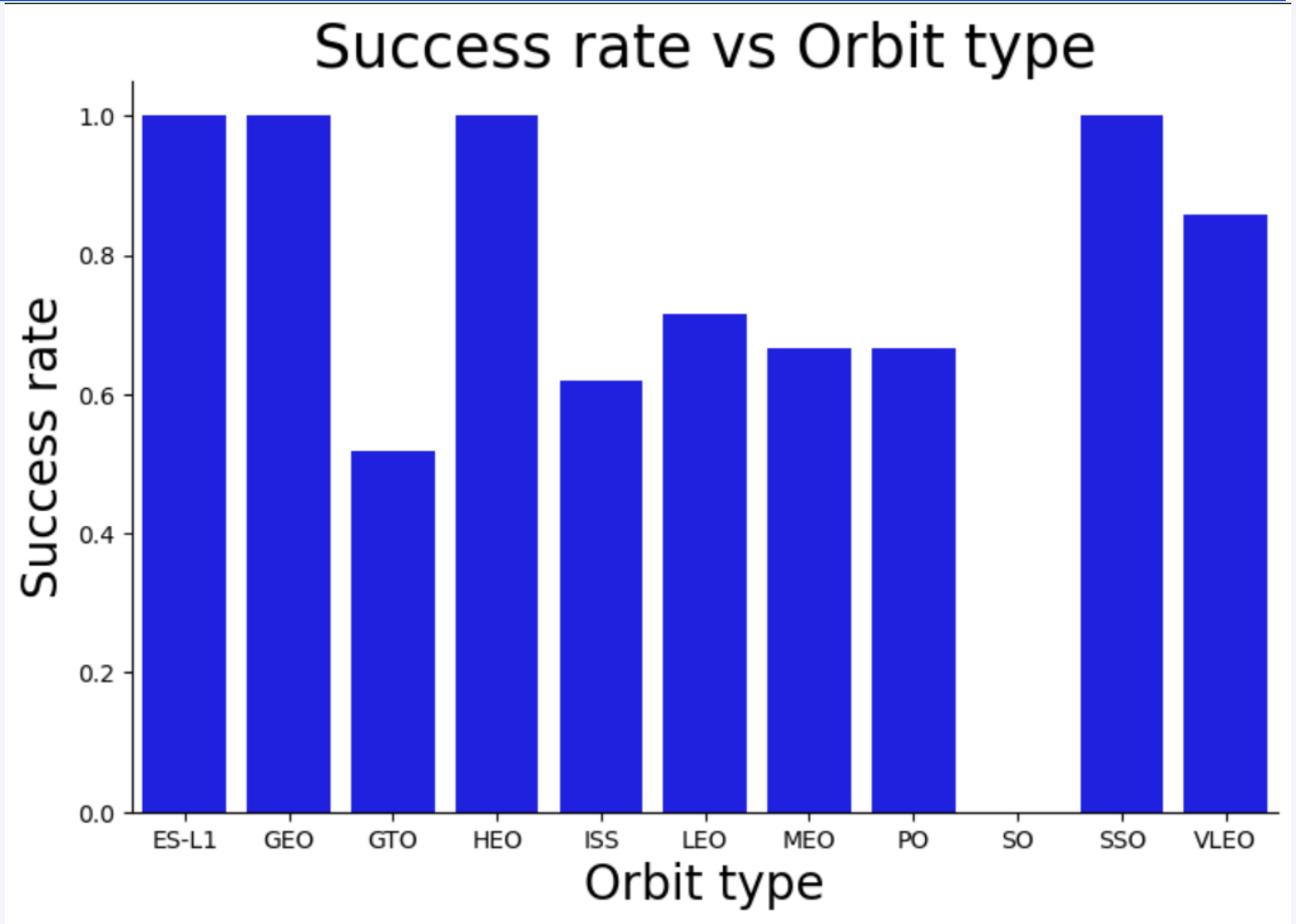
Payload vs. Launch Site

For the VAFB-SLC launch site there are no rockets launched for payload mass greater than 10000 kg.



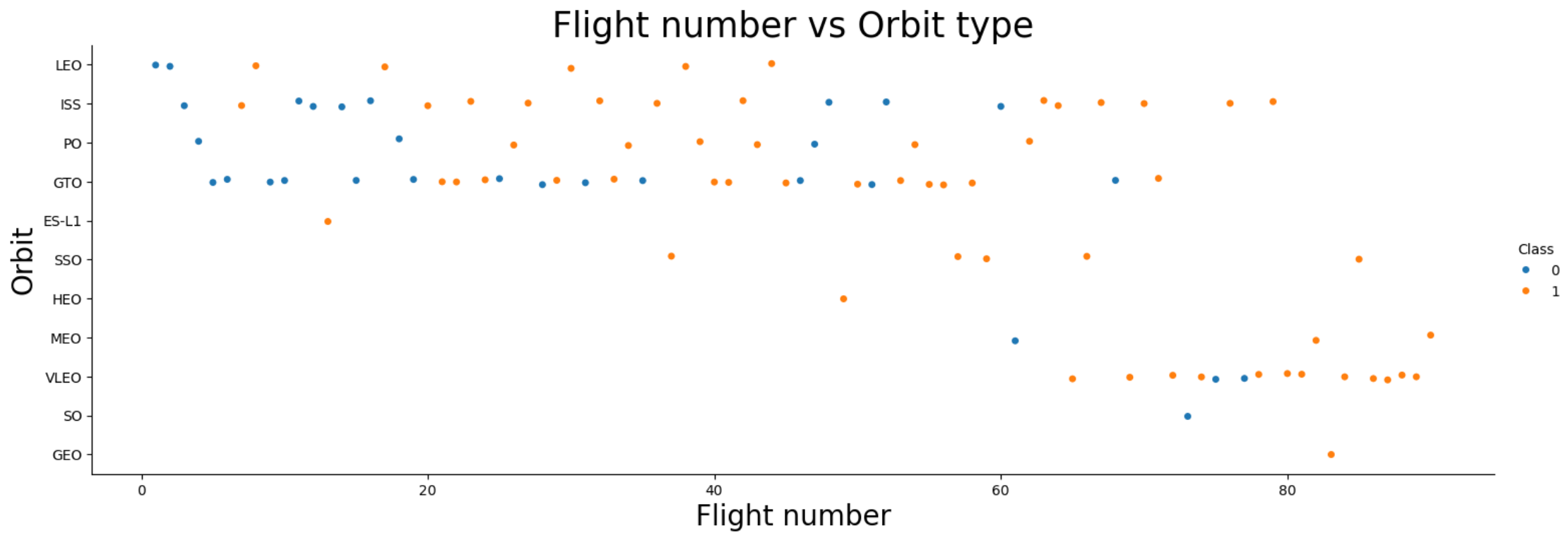
Success Rate vs. Orbit Type

ES-L1, GEO, HEO and SSO orbits have 100% success landing rate, whereas SO orbit's landing rate is zero.



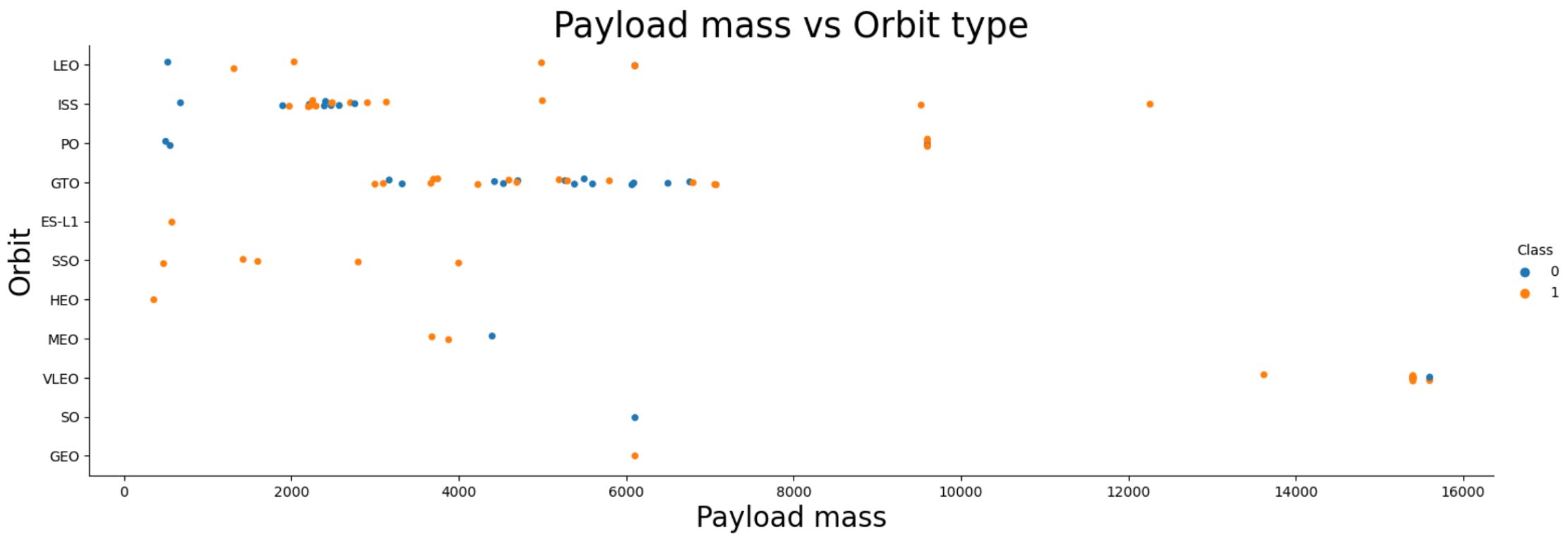
Flight Number vs. Orbit Type

In the LEO orbit the success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in other orbits.



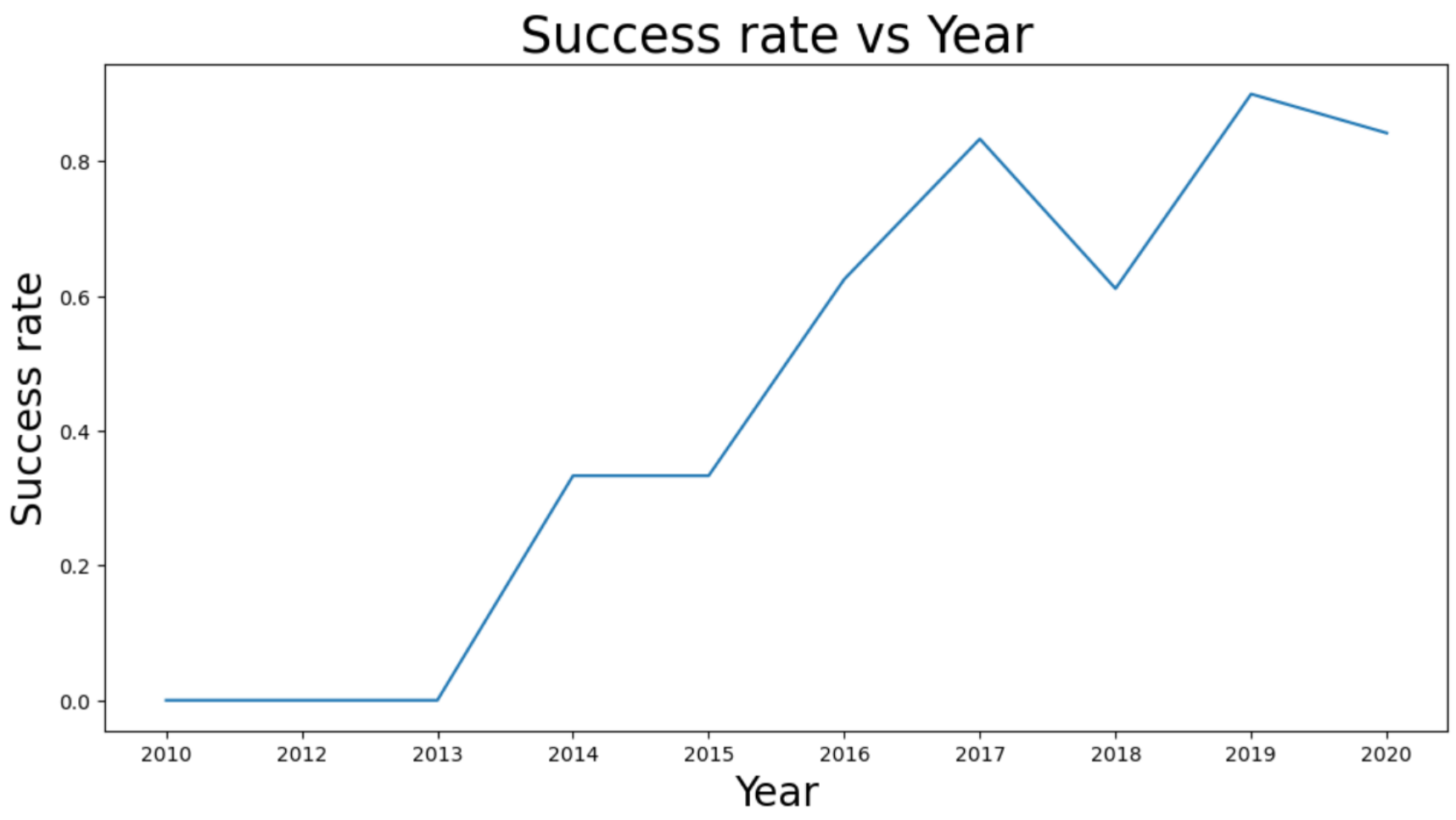
Payload vs. Orbit Type

With heavy payloads the successful landing rate are more for PO, LEO and ISS. However, for GTO we cannot distinguish this well as both positive landing rate and negative landing are both there here.



Landing Success Yearly Trend

The success landing rate since 2013 kept increasing till 2020.



All Launch Site Names

The Falcon 9 missions were launched from these four sites:

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

Launch Site Names Begin with 'CCA'

The 5 records where launch sites begin with `CCA` are:

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

Total Payload Mass

The total payload mass carried by boosters launched by NASA (CRS) is 45596 kg.

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1 is 2928.4 kg.

First Successful Ground Landing Date

- The first successful landing outcome in ground pad was achieved on December 22, 2015.
- In the original SpaceX dataset, the such type of successful outcomes are named “Success (ground pad)”.

Successful Drone Ship Landing with Payload between 4000 and 6000

The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 kg, but less than 6000 kg are:

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

In the original SpaceX dataset, the such type of successful outcomes are named "Success (drone ship)".

Total Number of Successful and Failure Mission Outcomes

The total number of successful and failure mission outcomes are:

- Successful – 100
- Failure – 1

Successful mission outcome rate is almost 100%, but there are still problems with successful landings, which rate is only 66.67%.

Boosters Carried Maximum Payload

The names of the boosters which have carried the maximum payload mass are:

- 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

The failed landing outcomes in drone ship in year 2015:

Month	Landing Outcome	Booster Version	Launch Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Ranking of Landing Outcomes from 2010-06-04 to 2017-03-20

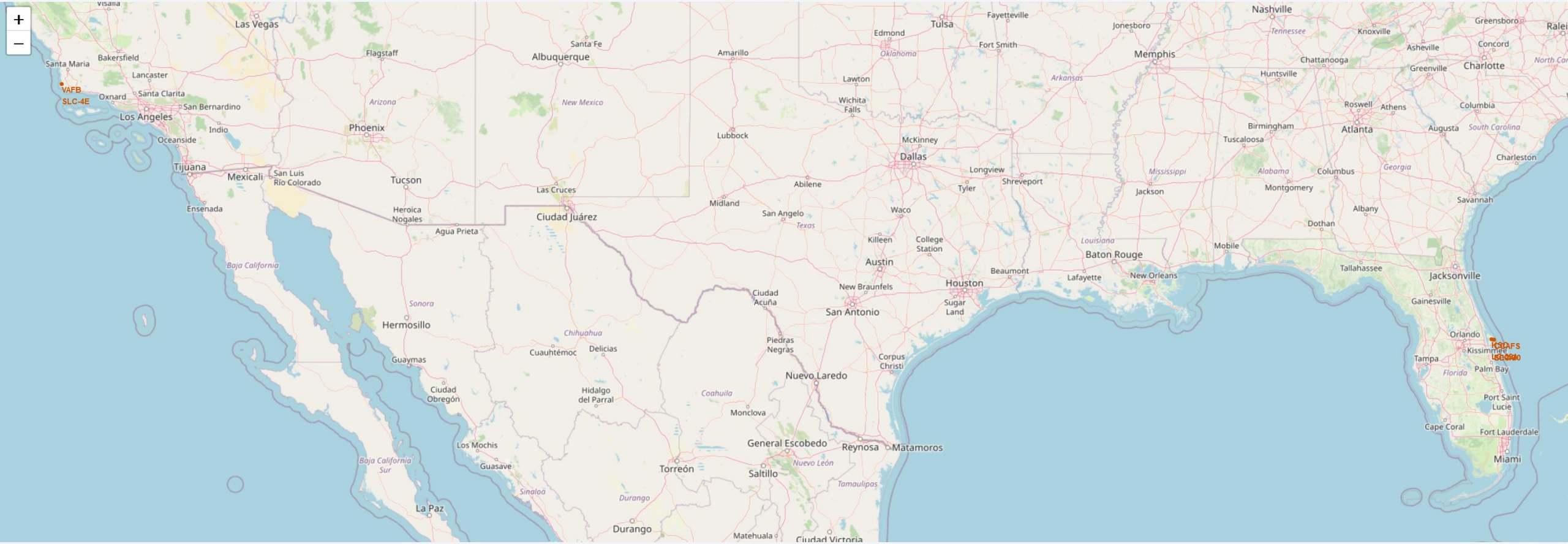
- The landing outcomes between 2010-06-04 and 2017-03-20, in descending order.
- The most common failed landing outcomes in this period were “No attempt”.

Landing Outcome	Landing count
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

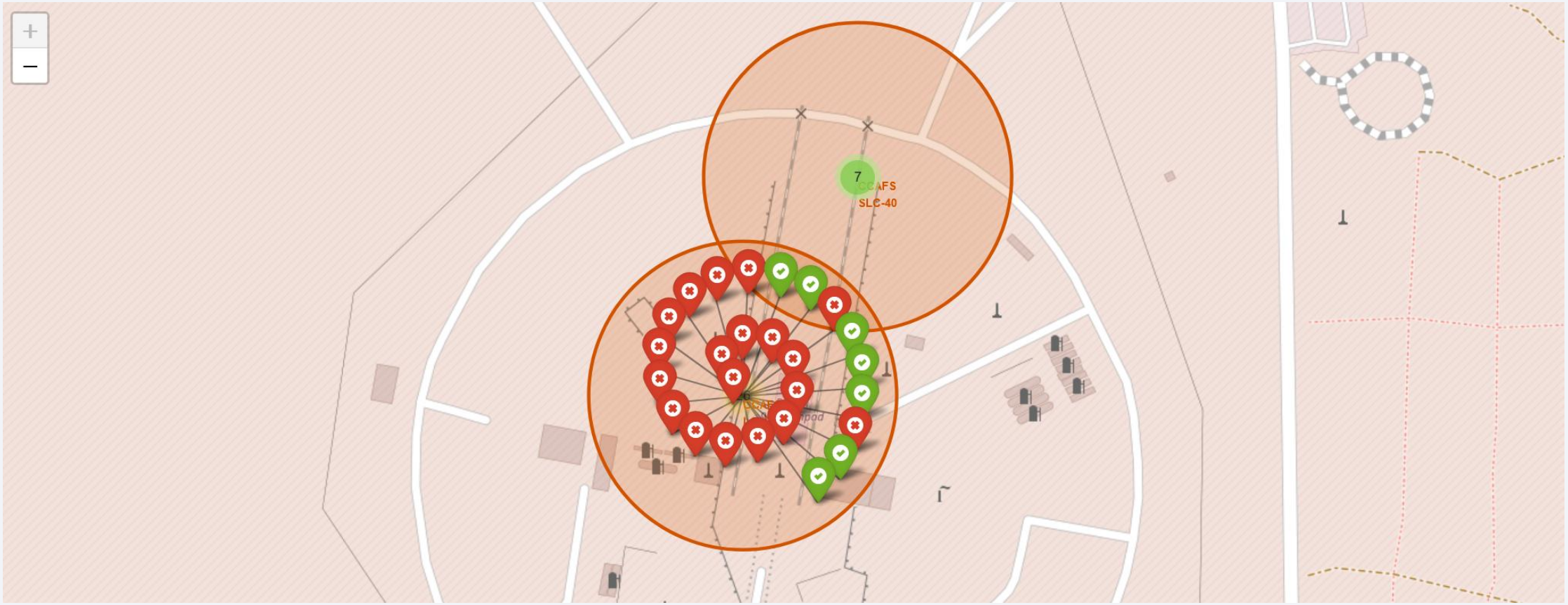
Falcon 9 Launch Sites on the map

Three launch sites are on the Cape Canaveral, and one is in California.



The success/failed launches markers on the map

For every launch site the successful launches are marked with green, the failed launches are marked with red.

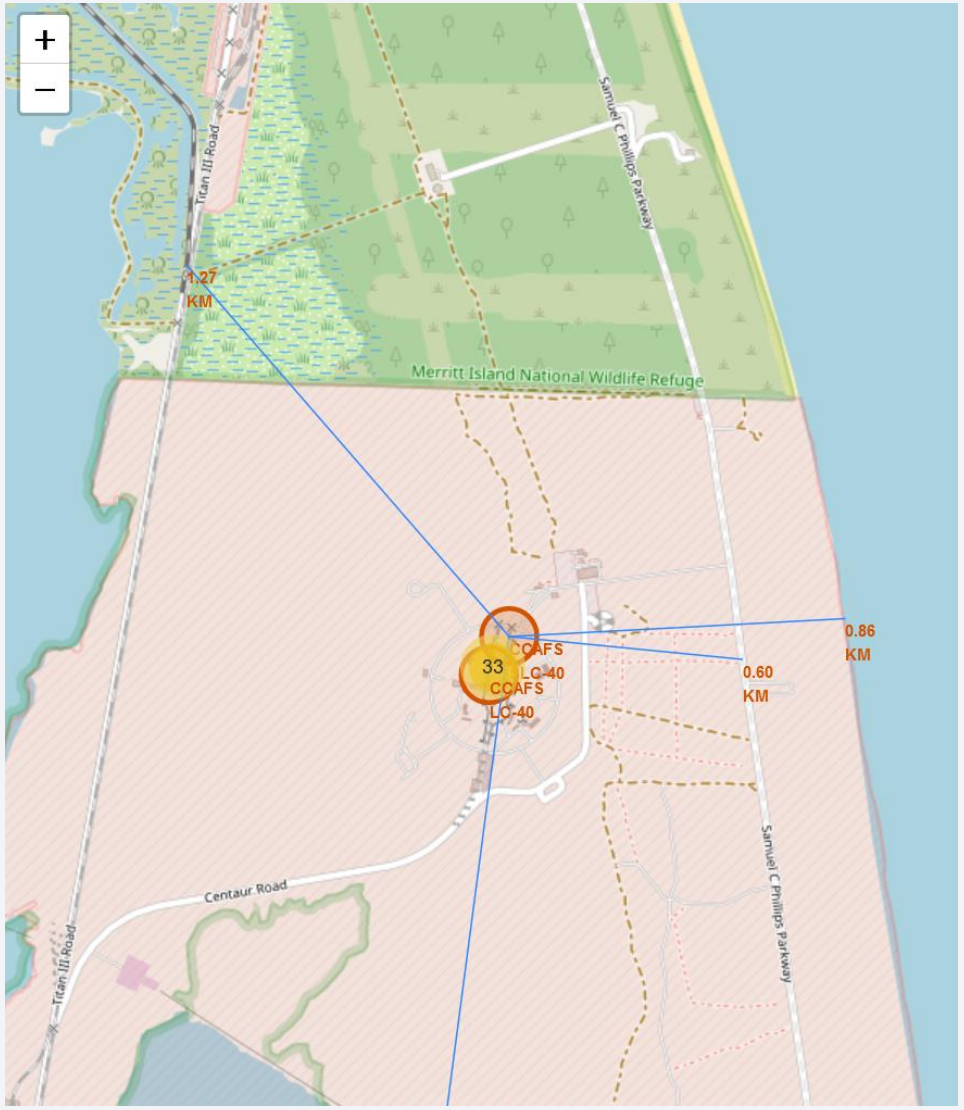


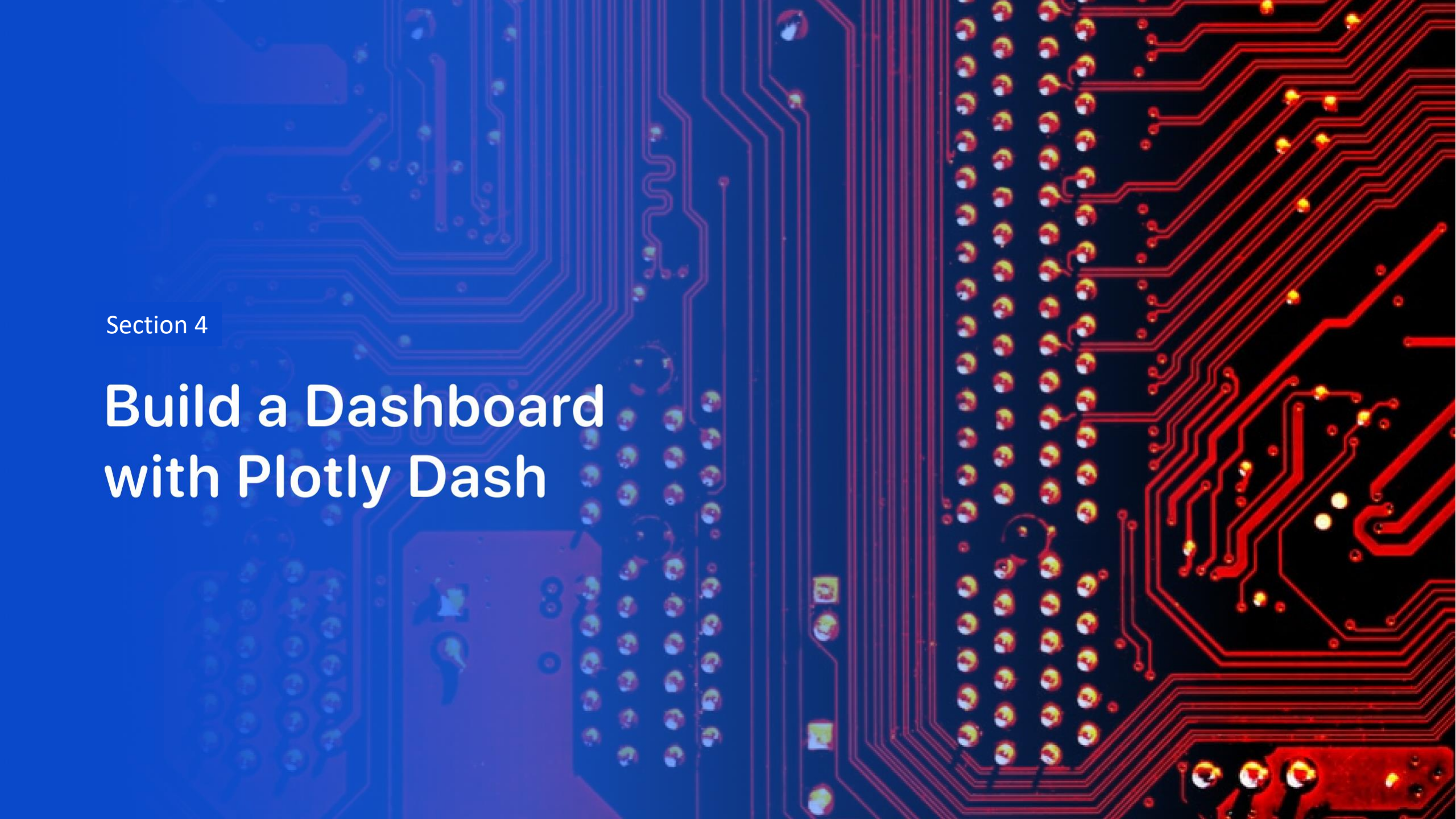
The distances from a launch site to the proximities

The distances from CCAFS SLC-40 launch site to the civil infrastructure:

- The Titan III Road railway is 1.27 km from the launch site.
- The eastern coastline is 0.86 km from the launch site.
- The Samuel C Philips Parkway is 0.60 km from the launch site.
- The city of Melbourne is 47.75 km from the launch site.

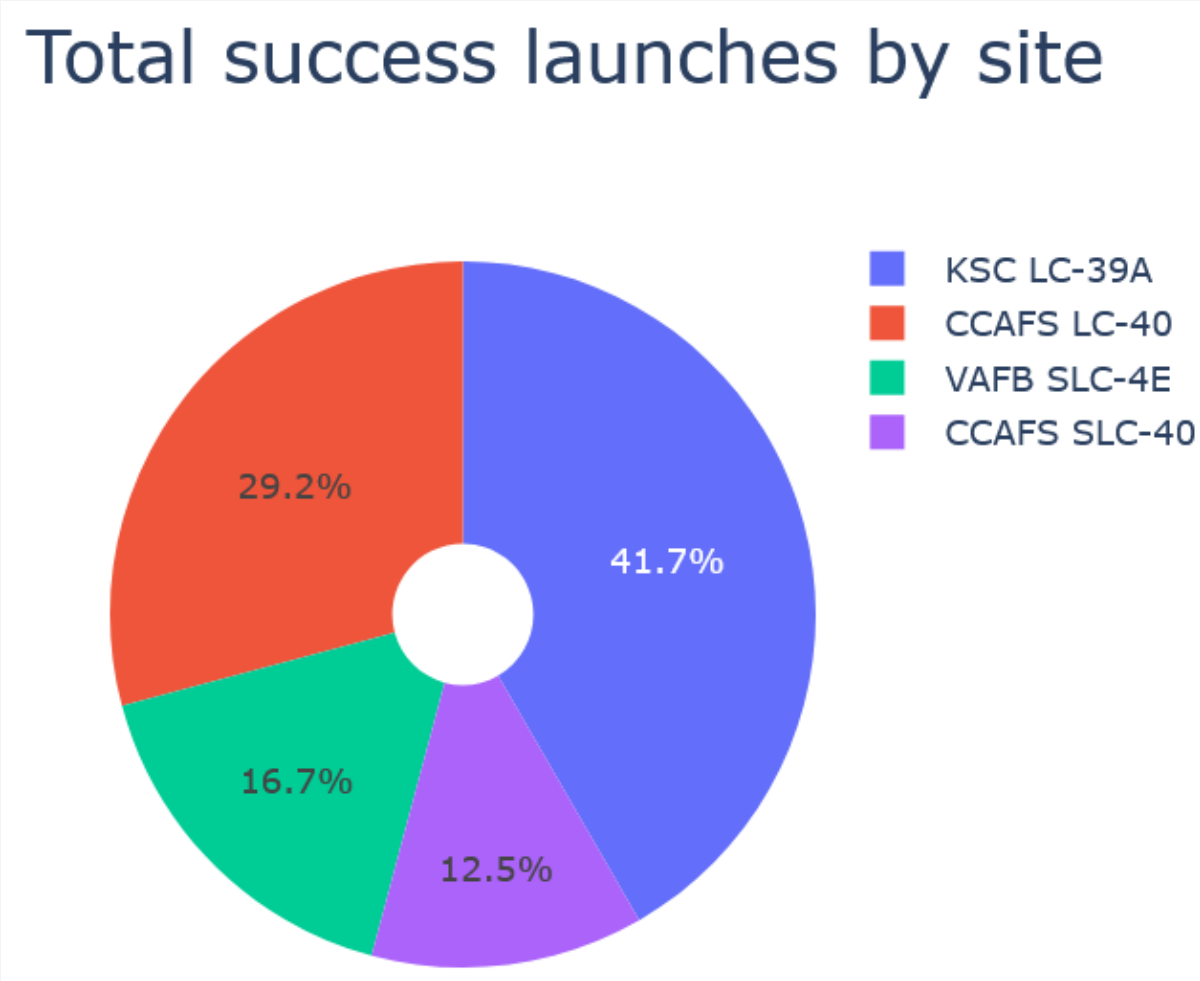
For other launch sites there is no cities close to them. All launch sites are close to the ocean coastline.




Section 4
Build a Dashboard with Plotly Dash

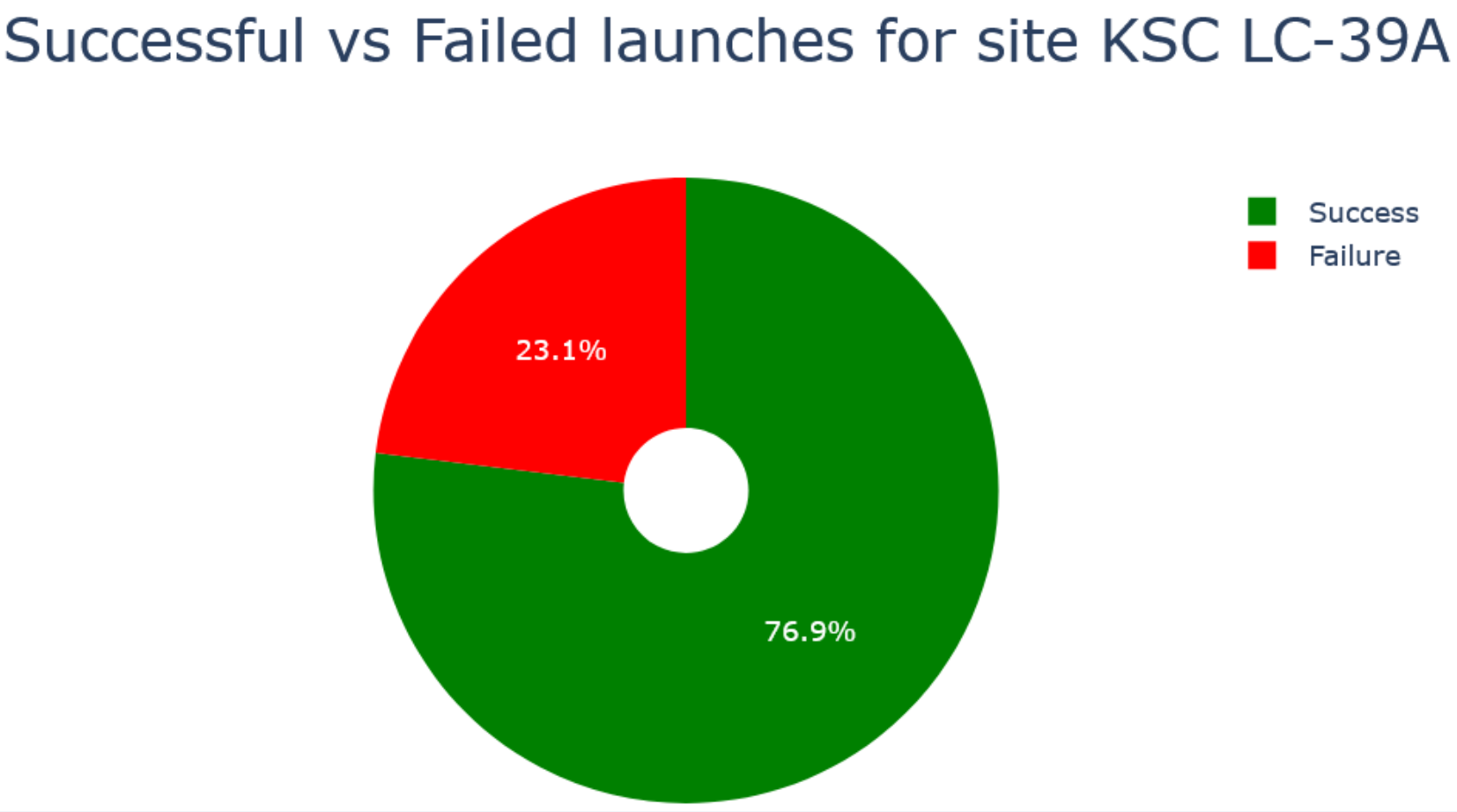
Launch success count for all sites

- KSC LC-39A launch site had 10 successful launches,
- Whereas CCAFS SLC-40 had only 3 successful launches.



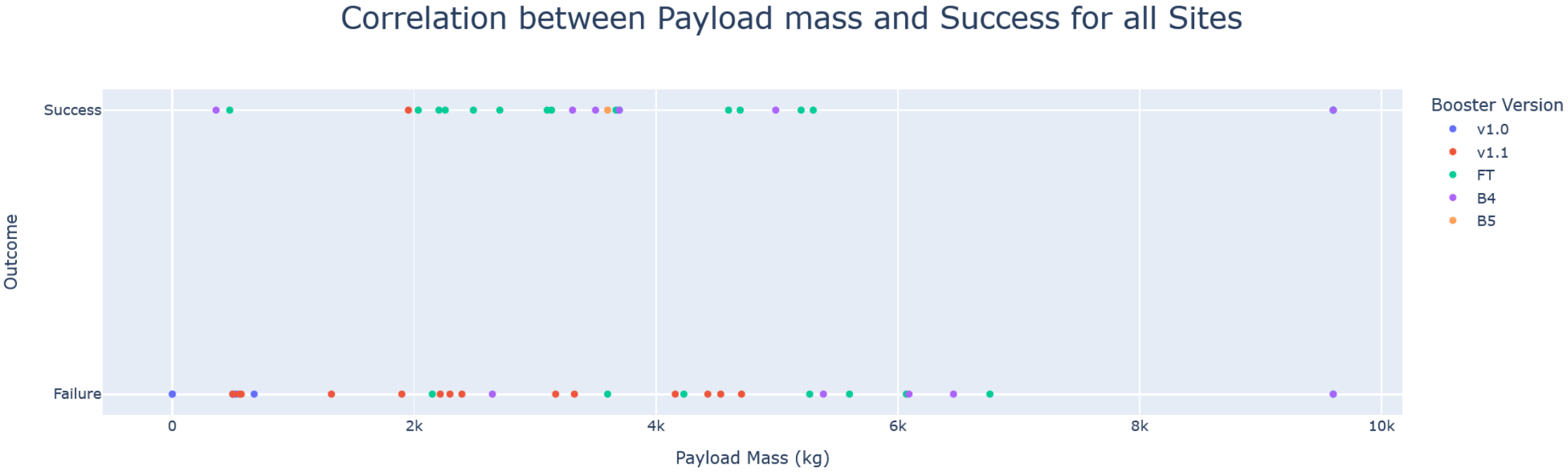
The launch site with highest launch success ratio

KSC LC-39A is the launch site with highest success ratio – 76.9%.

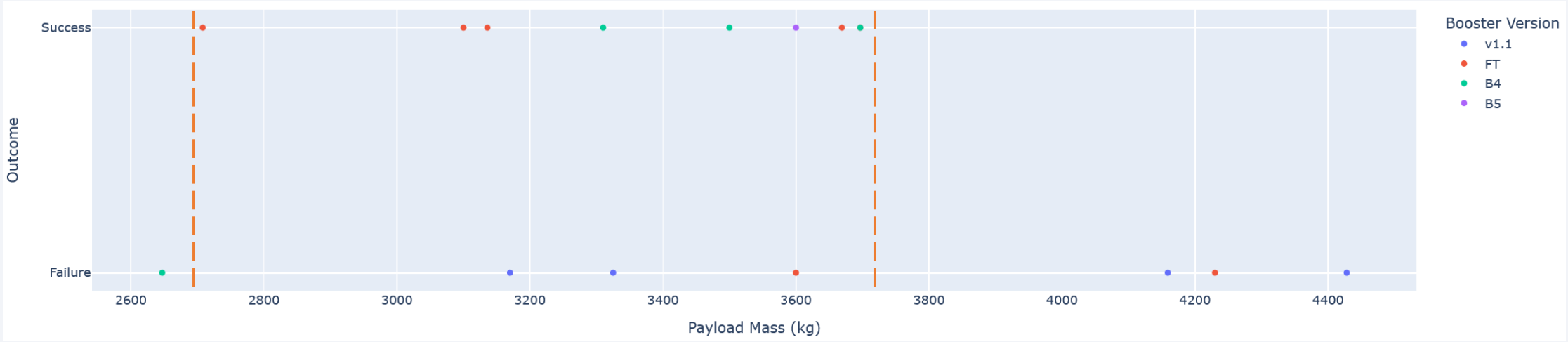


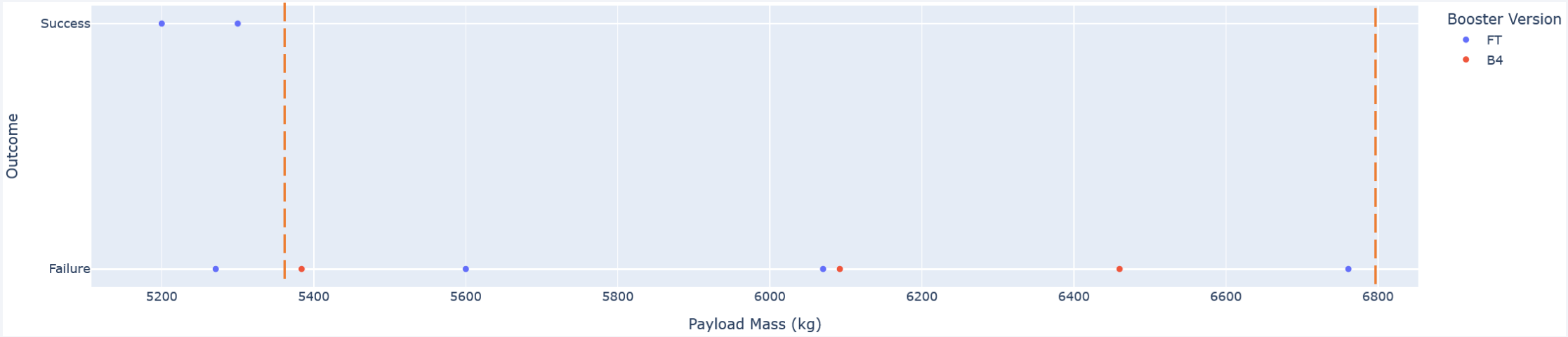
Payload vs. Landing Outcome for all sites

- FT booster version has the highest success rate, whereas V1.1 has the lowest one.
- The payload mass range 2700-3700 kg has the highest launch success rate.
- The payload mass range 5380-6800 kg has the lowest launch success rate.



Best and worst Payload Mass ranges



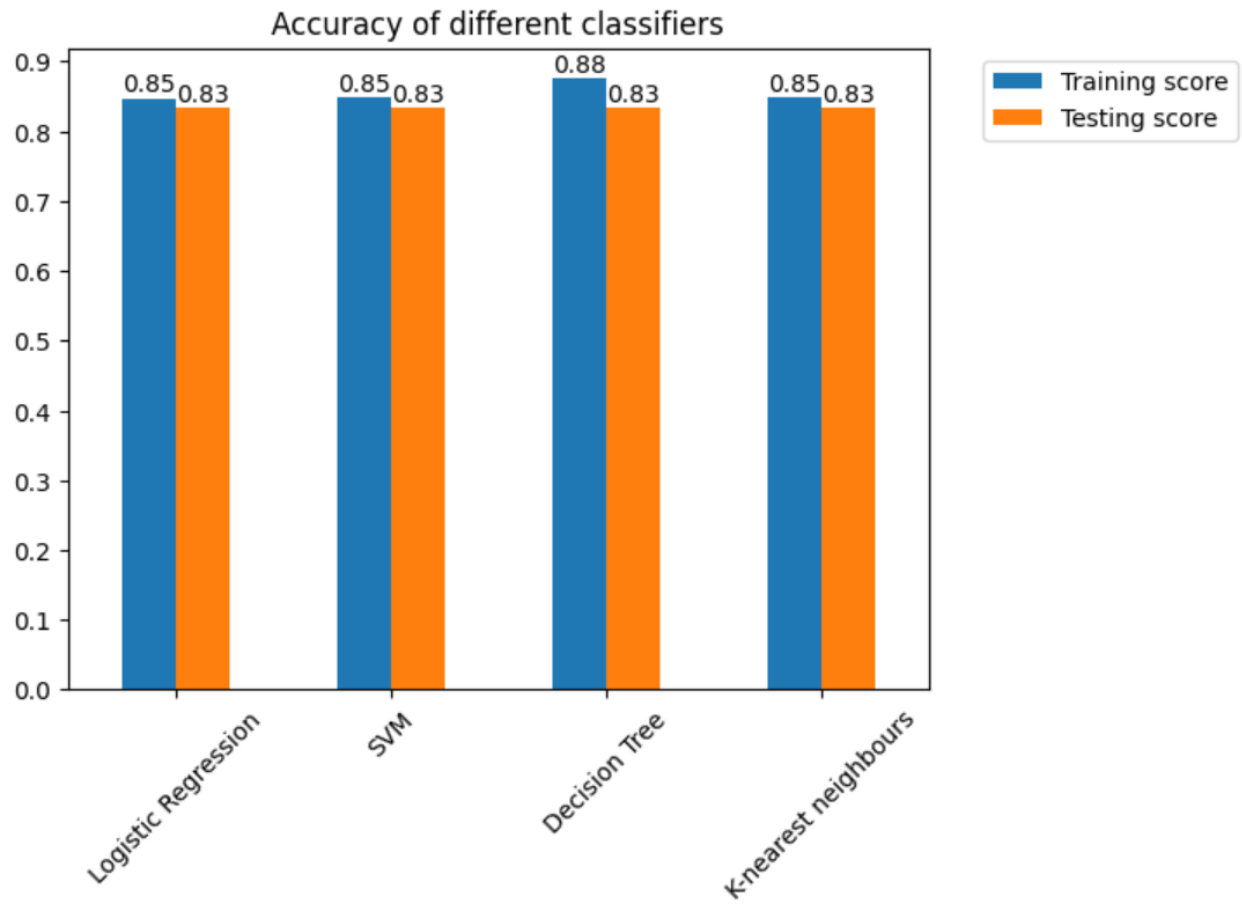


Section 5

Predictive Analysis (Classification)

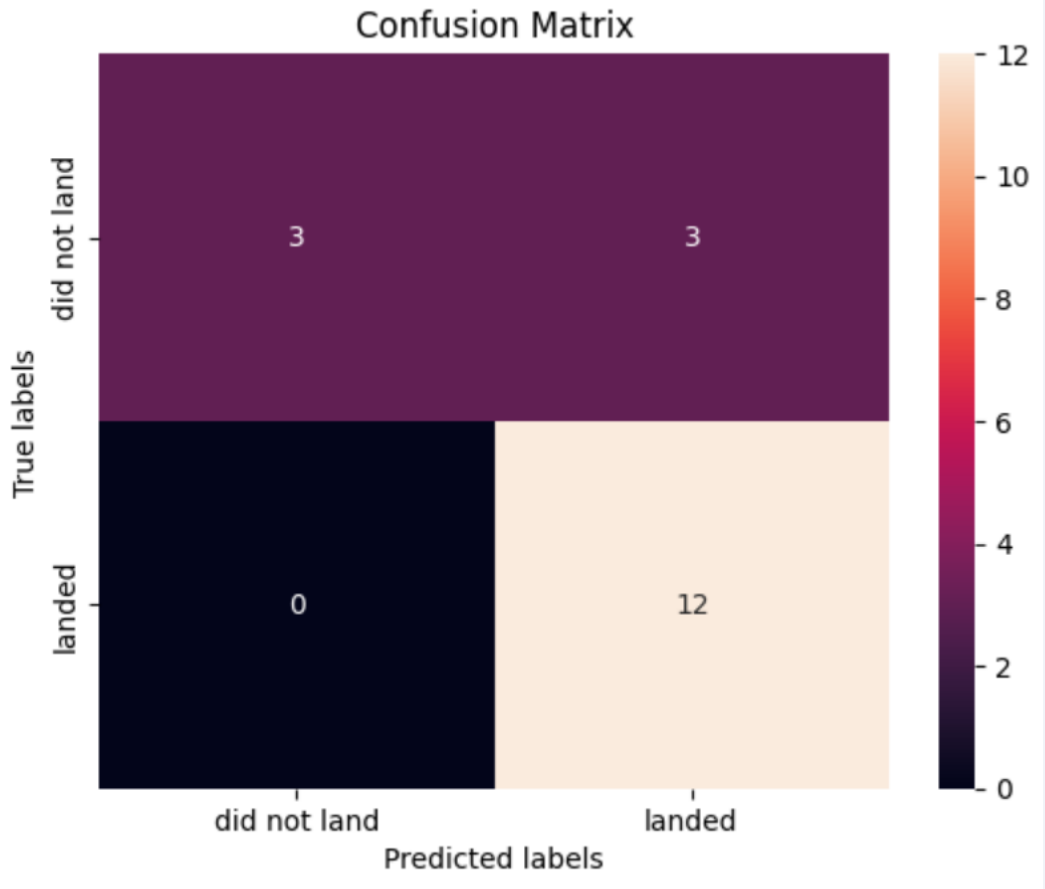
Classification Accuracy

- The Logistic regression, SVM, and KNN models are with the same testing accuracy of 0.83 for this data set.
- The testing accuracy of Decision Tree classifier usually is 0.83. But it changes randomly upon repeated fitting launches from 0.5 to 0.94. The reason is random model parameters changes.



Confusion Matrix

- The confusion matrix of all four models is the same.
- The exception is fluctuations of Decision Tree model because of random model parameters changes.
- The major problem for all four models is false positives. The false negatives are 0.



Conclusions

- Successful mission outcome rate is almost 100%, but there are still problems with successful landings, which rate is only 66.67%.
- The successful landing rate is increased over time for almost all launch sites, besides KSC LC 39.
- The Logistic Regression, SVM, Decision Trees, and K-nearest neighbours models give the same accuracy of 0.83 with current dataset of 101 launches. The false negatives are zero, but the false positives are 20%(3/15).
- Over time the launches dataset will be increased, as a result the model accuracy will be increased also after repeated training.

Appendix

- Coursera IBM Data Science Professional Certificate course
- SpaceX REST API
- Falcon 9 in Wikipedia (latest release)
- The project notebooks in Github:
  - Data Collection from REST API
  - Data Collection via web scraping
  - Data Wrangling
  - EDA with data visualization
  - EDA with SQL
  - Interactive map with Folium
  - Dashboard with Plotly Dash
  - Predictive analysis with Machine Learning
- Github does not fully support image rendering of Jupyter notebooks, so it is recommended to use Jupyter Nbviewer instead.

Thank you!