IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Case Study

by Mihai Doroftei

23.04.2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection using different methods
  - Data wrangling
  - Exploratory data analysis and visualizations
  - Predictive analysis using Machine Learning models
- Summary of all results
  - Success/failure rates for different parameter pairs
  - Best combination of booster – payload
  - Best launch site
  - Predicting future launches outcomes

Introduction

Project background and context

- Space conquer is a priority for humankind, not necessarily for find new natural resources (that are not renewable on earth, in a short period of time), but for fulfilling our curiosity, exploring the part of the Universe we live in, expand our horizons and increasing our survival as a species also.
- The project studies Space X launch outcomes between 2010 and 2020

Investors are massively attracted by this innovator company, as an alternative to the existing state backed companies, hence they are highly interested in its overall capacity of achieving success in this field.

The project tries to predict the future outcomes of launches, based on historical data and actual trend

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - From SpaceX API using the Requests module
  - Web scrapping using the Beautiful Soup module
- Perform data wrangling
  - Slicing and filtering the dataset in order to keep only information of interest
  - Dealing with missing values (find and replace them based on the particularity of the column)

Methodology

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Create a column for the class (outcome)
  - Standardize the data
  - Split into training and test data
  - Find the best hyperparameter for Support Vector Machine, Classification Tree, Logistic Regression and K Neighbors Classifier using Grid Search CV
  - Find the method which performs the best on test data

Data Collection

- Data Collection Procedures (using Requests)

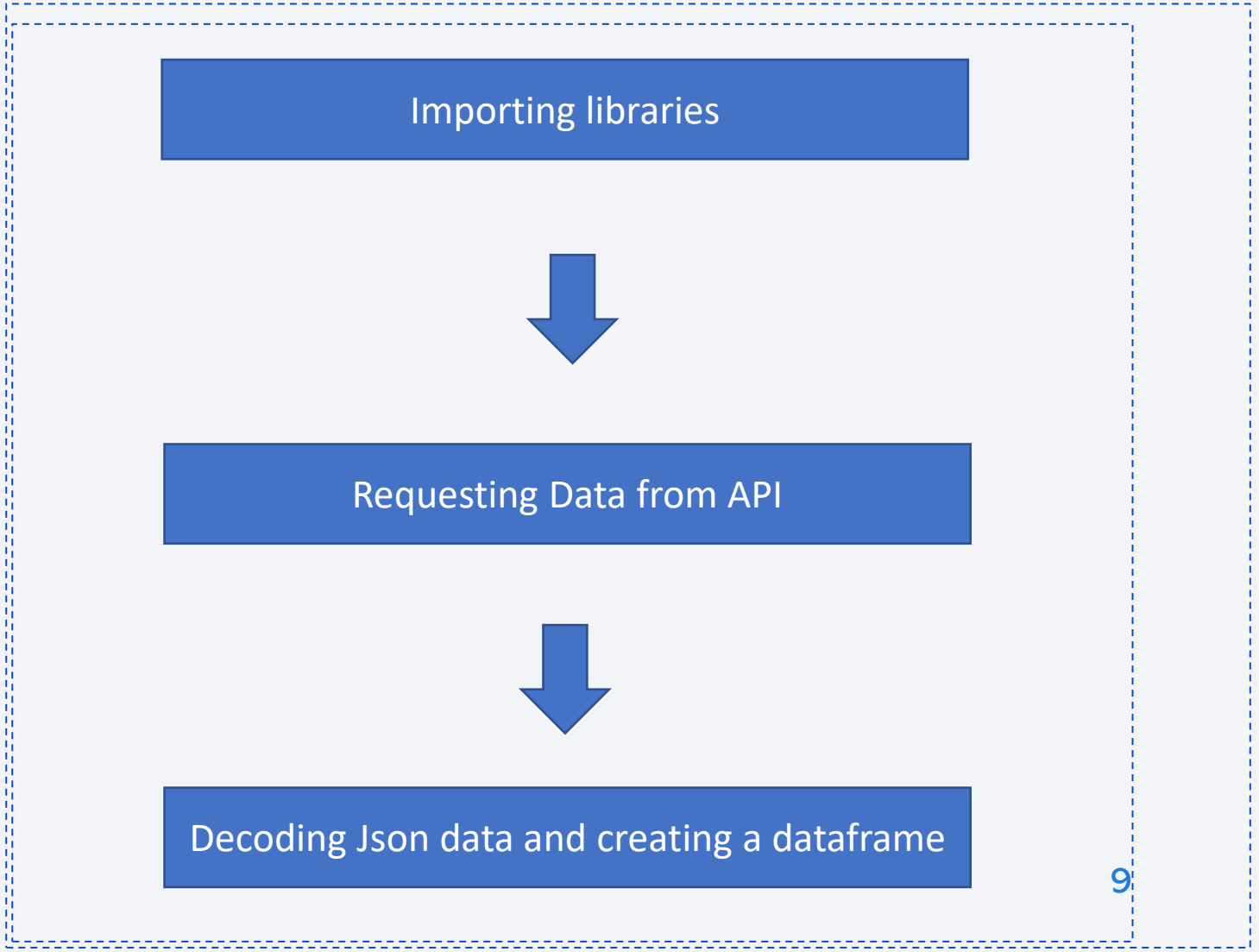
  - Using the Requests module (making HTTP requests which were used to get data from the SpaceX API)
  - Web scraping and parsing the results using the Beautiful Soup module

Data Collection – SpaceX API

- Flowchart

- GitHub URL of the completed SpaceX API calls notebook:

https://github.com/cityzenmike/Space-Race/blob/main/1.1%20jupyter-labs-spacex-data-collection-api.ipynb

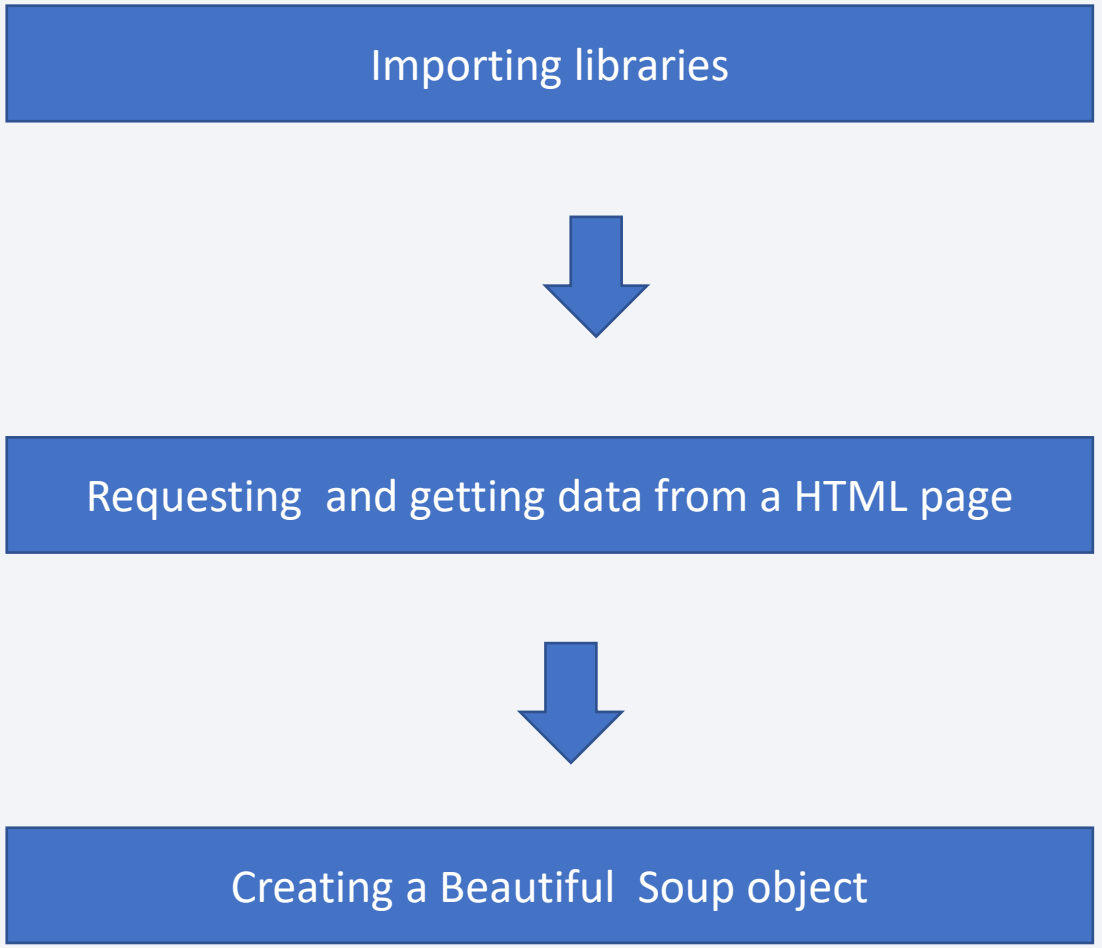


Data Collection - Scraping

- Flowchart

- GitHub URL of the completed web scraping notebook:

https://github.com/cityzenmike/Space-Race/blob/main/1.2%20jupyter-labs-webscraping.ipynb



Data Wrangling

https://github.com/cityzenmike/Space-Race/blob/main/1.3%20labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

- Identifying the columns of interest
- Searching for missing data
- Finding the number of launches for each site
- Finding the number and occurences of each orbit
- Finding the number and occurencies of mission outcome per orbit type
- Creating a landing outcome numerical column

EDA with Data Visualization

- Charts plotted and reasons of plotting them
  - Payload Mass vs Flight Number
  - Launch Site vs Payload Mass
  - Orbit vs Payload Mass
  - Yearly Success Rate
  - Success Rate vs Orbit
  - Launch Site vs Flight Number
  - Orbit vs Flight Number

The charts help to identify the outcome of successful landing of the first stage of the rocket based on the orbit, payload mass and launch site and the yearly trend.

The GitHub URL of your completed EDA with data visualization notebook

- https://github.com/cityzenmike/Space-Race/blob/main/1.4%20IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

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

EDA with SQL

- GitHub URL of the completed EDA with SQL notebook

https://github.com/cityzenmike/Space-Race/blob/main/2.%20jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- List of map objects created and added to the folium map
  - Circle
  - Markers
  - Marker Cluster
  - Mouse Position
  - Lines
- Those objects have been added in order to map the positions of the launch sites and find the distances between them and the nearest points of interests (coastlines, roads, railways).

Each marker offers information about the number successful and failed launches also.

Build an Interactive Map with Folium

- GitHub URL of the completed interactive Folium map

https://github.com/cityzenmike/Space-Race/blob/main/3.1%20IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite%20(2)%20-%20FINAL.ipynb

Build a Dashboard with Plotly Dash

- Plots/graphs added to the dashboard
  - Pie chart: Success rate for each launch site and for all of them as a whole
  - Scatter plot: success rate vs payload for each site or for all of them
- Plots/graphs added to the dashboard
  - Drop-down menu for site selection
  - Slide bar for payload selection within a range
- Reasons for adding those plots and interactions
  - Interactive selection of every possible combination of launch site and payload, in order to identify the best combination of the two factors, the one with the best success rate.

Build a Dashboard with Plotly Dash

The GitHub URL of the completed Plotly Dash lab:

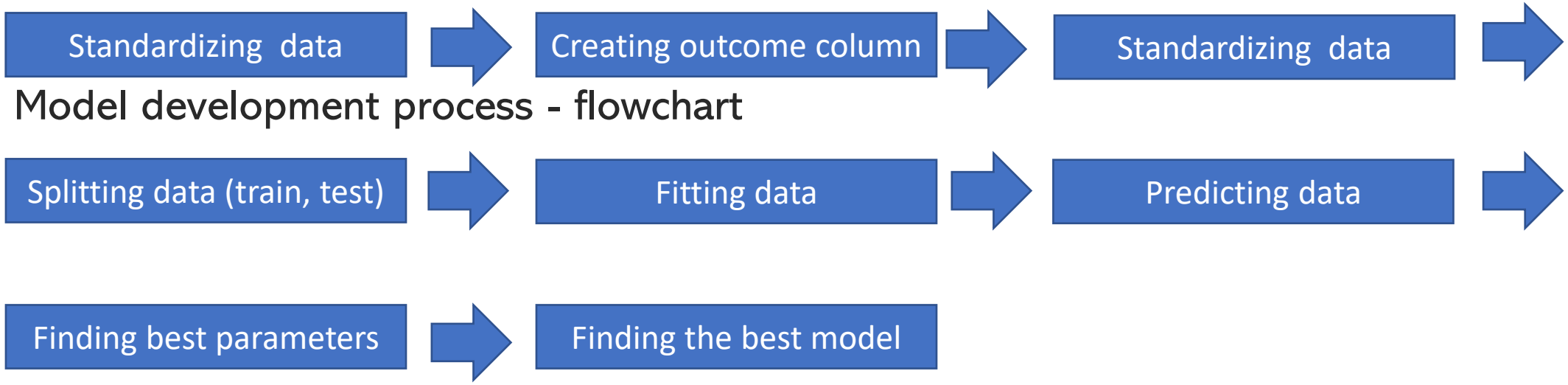
https://labs.cognitiveclass.ai/v2/tools/cloud-ide-kubernetes?ulid=ulid-86c283a8ce5e0f718bd45fee78022802a4ed2b95

Predictive Analysis (Classification)

- Summary of building, evaluating, improving and founding the best performing classification model

  - Subsetting the dataset
  - Creating a column for the outcome(class)
  - Standardizing the data
  - Splitting into training data and test data
  - Finding the best hyperparameters for SVM, Classification Trees, Logistic Regression and K Neighbours using Grid Seach CV
  - Finding the method that performs best using test data, based on the accuracy od different models

Predictive Analysis (Classification)

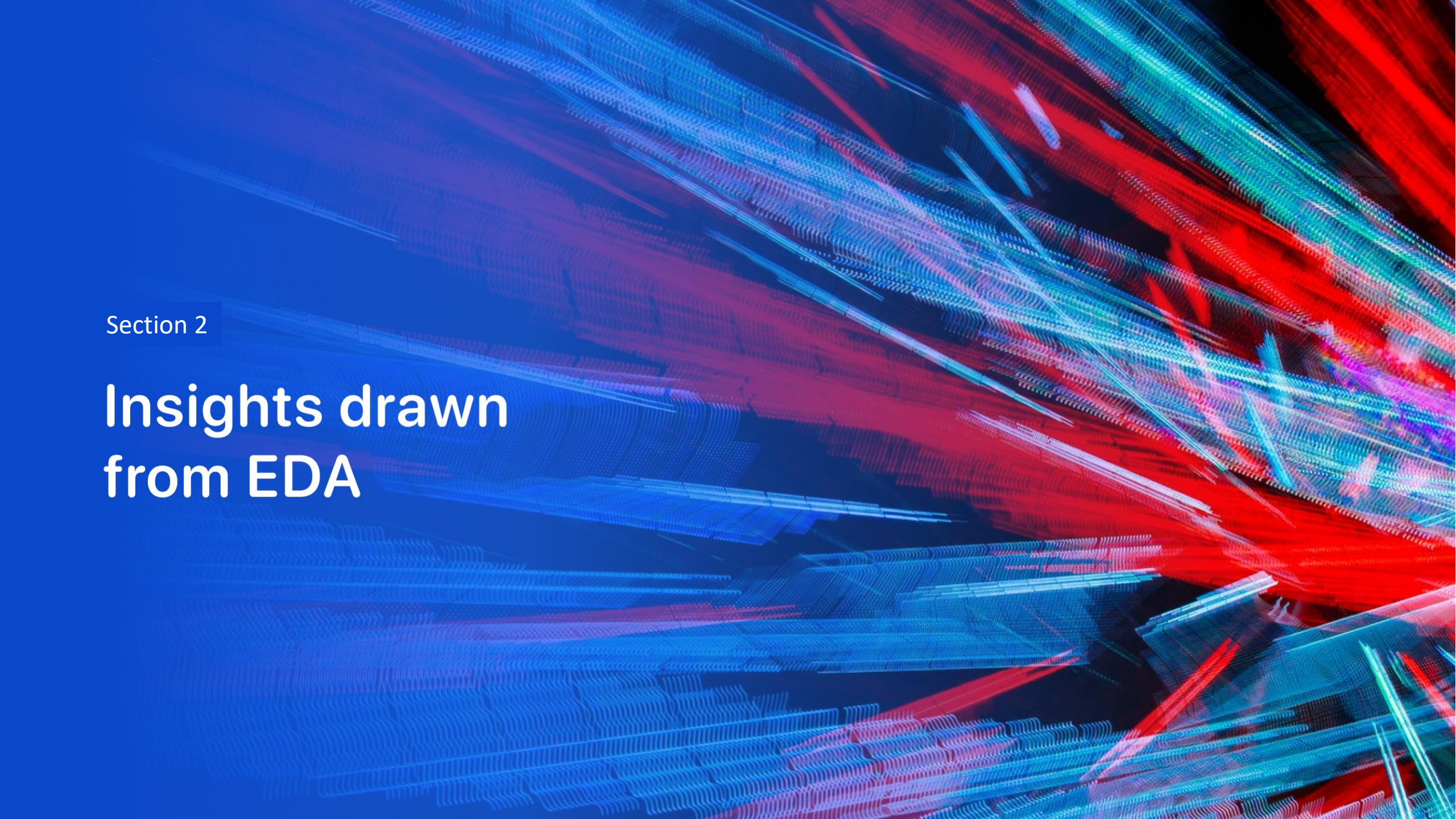


- Model development process - flowchart

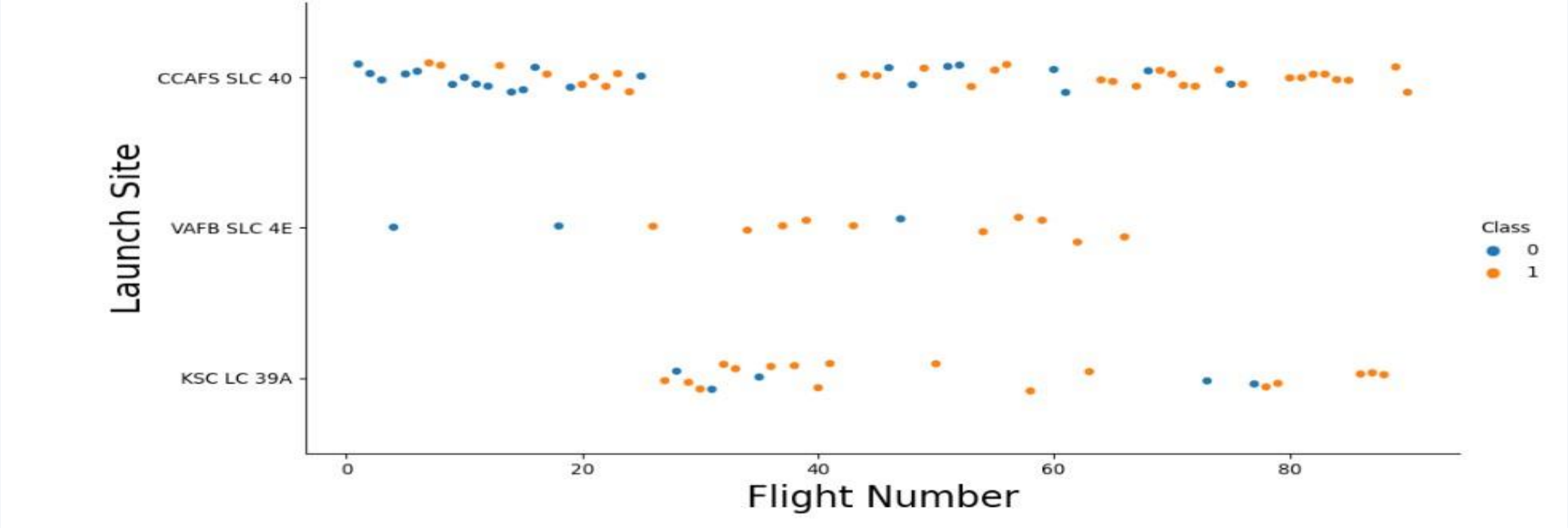
https://github.com/cityzenmike/Space-Race/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite%20(3).ipynb

Results

- Exploratory data analysis results (insights drawn from EDA)
- Interactive analytics demo in screenshots (launch sites proximity analysis)
- Predictive analysis results

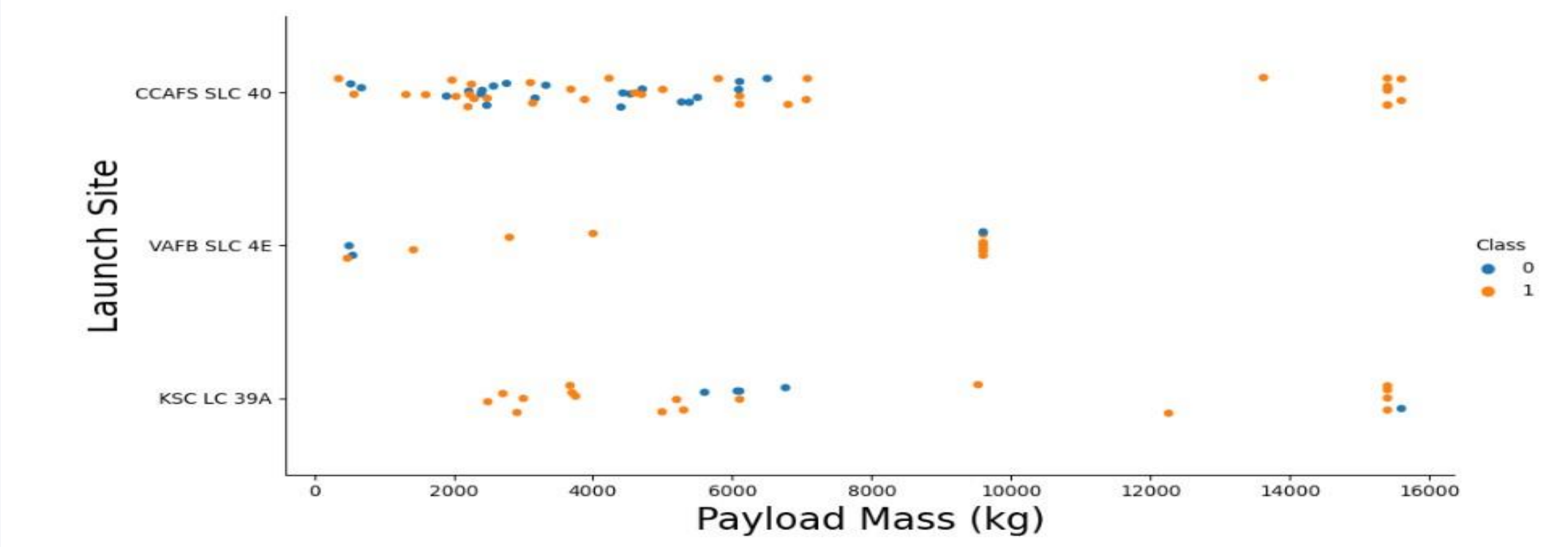

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



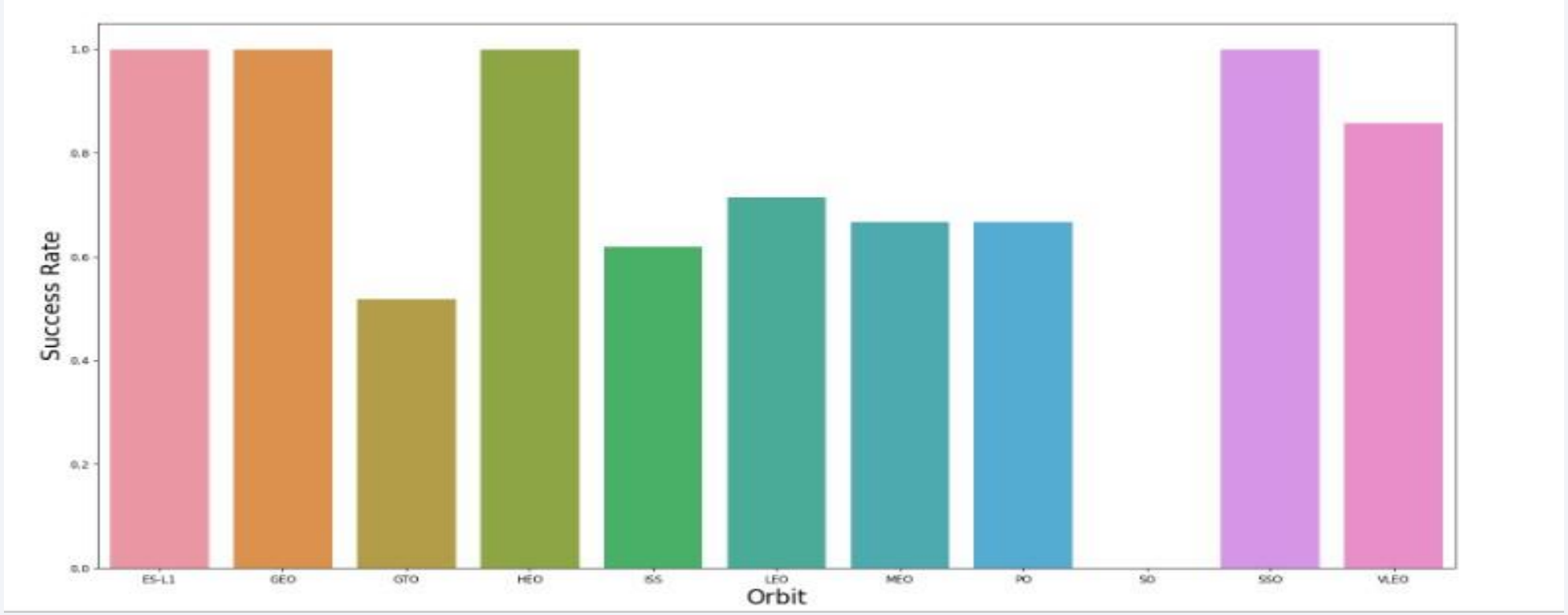
- The chart shows that, despite the initial failures, the rate of success has increased during the analyzed period of time for each launching site. Actually, all recent launches were successful.

Launch Site vs. Payload



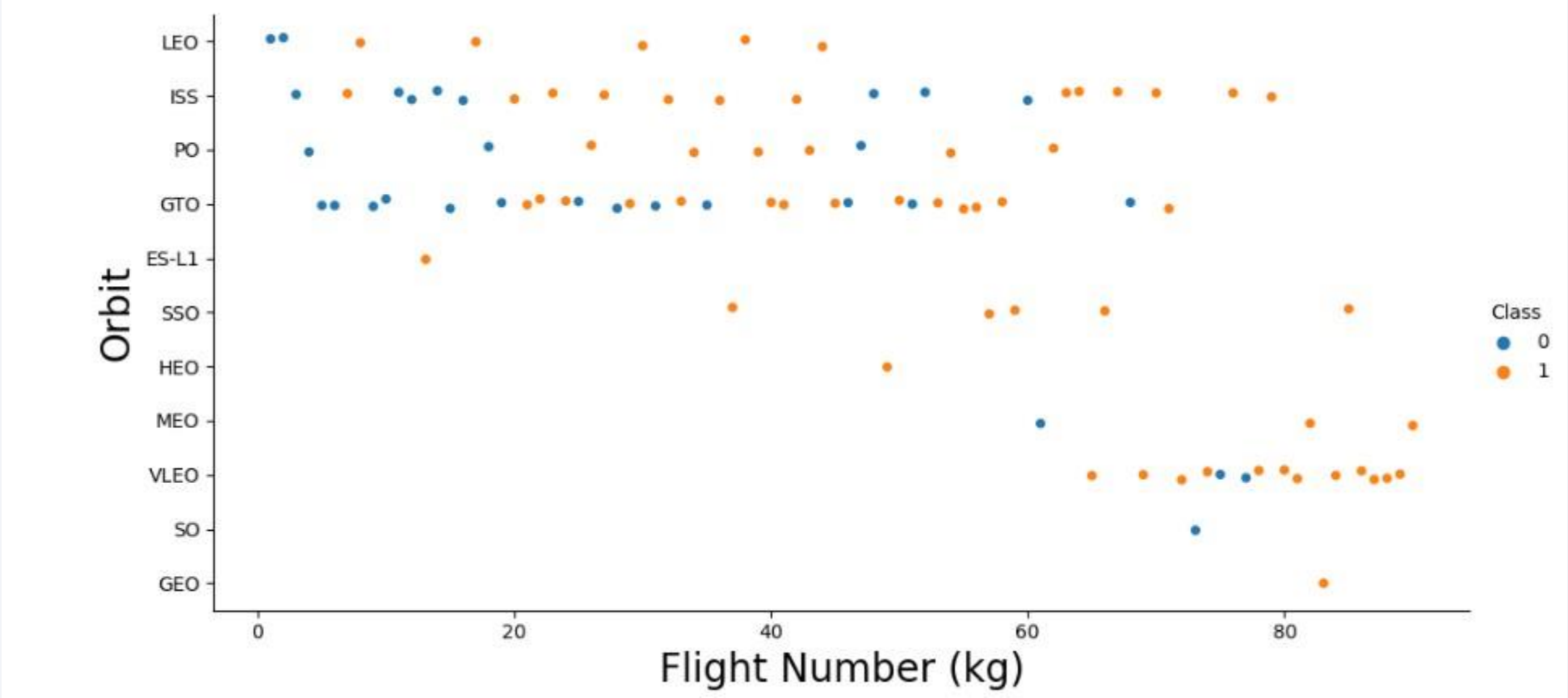
Except for two, all of the launches having a payload superior to 7,000kg were successful for each launch site. In the case of VAFB SLC-4E there were no payloads above 10,000 kg.

Success Rate vs. Orbit Type



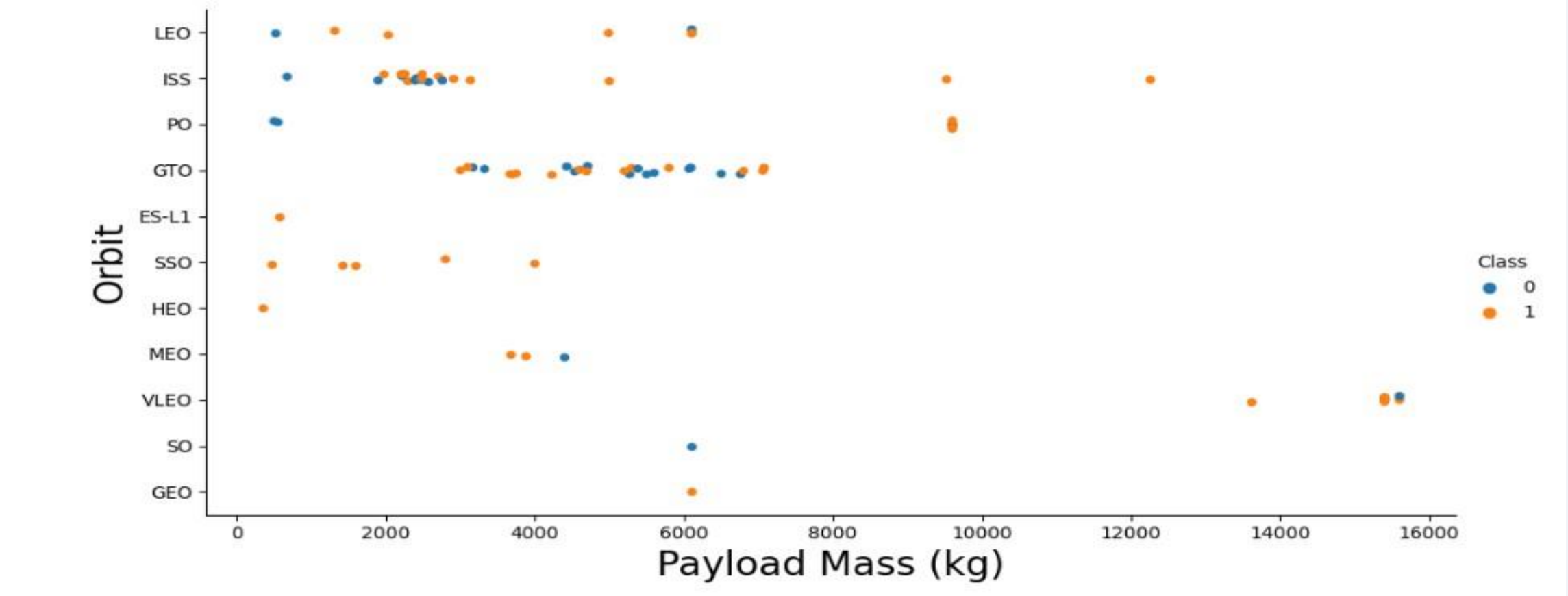
The success rate was 100% for 4 orbits (ES-L1, GEO, HEO, SSO), 85% for VLEO, between 50 and 75% for other 5 orbits (GTO, ISS, LEO, MEO, PO) and 0% for SO (single unsuccessful launch).

Flight Number vs. Orbit Type



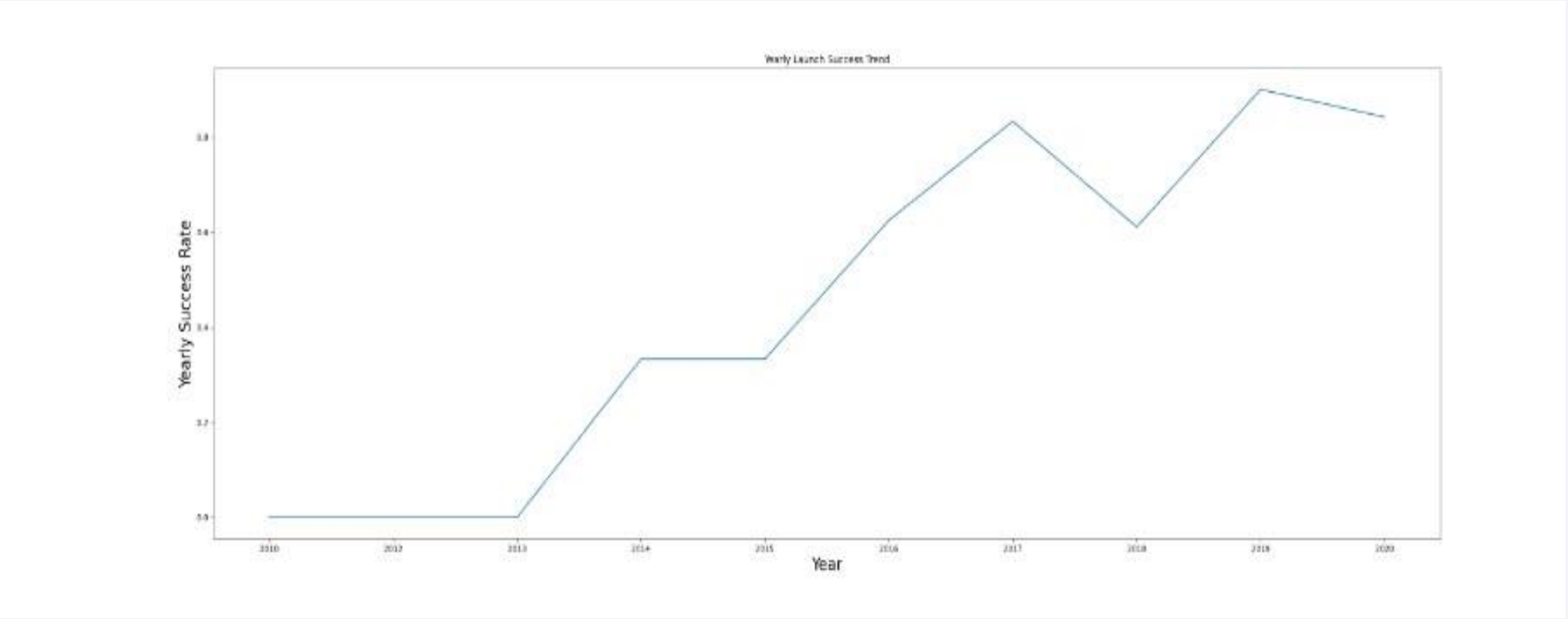
The last 30 launches focused on low (ISS) and very low (VLEO) orbits and most of them were successful.

Payload vs. Orbit Type



The heaviest payloads (above 8,000 kg) were sent on very low (VLEO) or low (ISS, PO) orbits. Except one, the others were successful.

Launch Success Yearly Trend



The average success rate has increased and remained over 80% in the past 4 years (except 2017). The overall trend is positive. It's very likely that most of the future launches will be successful.

All Launch Site Names

- The names of the unique launch sites

```
[7]: %sql select distinct Launch_Site from spacextbl;
 * sqlite:///my_data1.db
Done.
```

[7]:

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

- The “DISTINCT” SQL command was used in order to filtering the dataset and find the names of all 4 unique launch sites.

5 Launch Site Names Begin with 'CCA'

```
[8]: %sql select * from spacextbl where Launch_Site like "CCA%" limit 5;
 * sqlite:///my_data1.db
Done.
```

[8]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

In order to find 5 launch sites whose names begin with “CCA”, a “LIKE” statement and the magic card character “%” were used. In fact, the pattern searches for “CCA” followed by any other characters in the column “Launch_Site”.

Total Payload Mass Carried by Booster from NASA

```
#%sql select distinct Customer from spacextbl;
```

```
%sql select sum(PAYLOAD_MASS__KG_) from spacextbl where Customer="NASA (CRS)"
```

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

sum(PAYLOAD_MASS__KG_)
45596

Total payload was calculated using the aggregate function “SUM”, adding up all the records in the “Payload Mass” column, where the name of the customer was “NASA (CRS)”.

Average Payload Mass by 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

Average payload mass was calculated by using the aggregate function “AVG”, by finding the mean of all records within the column “Payload Mass” where the value in the “Booster Version” column starts with “F9 v1.1” (using “LIKE” inside a “WHERE” clause and the magic card character “%” which stands for “Any other characters”.

First Successful Ground Landing Date

```
%sql select date from spacextbl WHERE "Landing _Outcome" = "Success (ground pad)" limit 1;
```

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

Date
22-12-2015

The specific date was calculated as the date corresponding the first successful ground landing outcome.

(It should have been : %sql SELECT MIN(date) FROM spacextbl WHERE “Landing_Outcome”=“Success (ground pad)”, but the command doesn’t work at all, for unknown reasons.)

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select distinct booster_version from spacextbl \
            where "landing _outcome" = "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

The query uses the “and” operator in the “where” clause in order to select the unique booster versions that successfully landed on a drone ship, having a payload between 4,000 and 6,000 kg.

Total Number of Successful and Failure Mission Outcomes

```
%sql select mission_outcome, count(mission_outcome) from spacextbl group by mission_outcome;
```

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

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

The query groups the data based on the mission outcome, then aggregates (counts) their occurrences. The results show that 100 out of 101 landings on a drone ship were successful.

Boosters Carried Maximum Payload

```
%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

This shows the case of using subqueries: the inner query is executed first, calculating the maximum payload, which became the criteria for the outer query, in order to select only those booster versions whose payload is equal to the calculated value.

2015 Launch Records

```
%sql select substr(Date,7,4),substr(Date,4,2) as "Month", "Landing _Outcome", Booster_Version, Launch_Site from spacextbl \
              where ("Landing _Outcome" like "%Failure%drone%") and substr(Date,7,4)='2015';
```

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

substr(Date,7,4)	Month	Landing _Outcome	Booster_Version	Launch_Site
2015	01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015	04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

This query uses the logical operator “and” within a “where” clause to find the failed drone ship landings, order by month, in 2015.

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

```
%sql select "Landing _Outcome", count(*) as Outcome_Count from spacextbl \
   where (substr(Date,7,4) || substr(Date,4,2) || substr(Date,1,2) between '20100604'and '20170320') and ("landing _Outcome" like "Success%") \
   group by "Landing _Outcome" order by outcome_count desc;
```

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

Landing _Outcome	Outcome_Count
Success (drone ship)	5
Success (ground pad)	3

This query looks up for dates between two dates, groups the records and aggregates the data by counting the values.

Section 3

Launch Sites Proximities Analysis

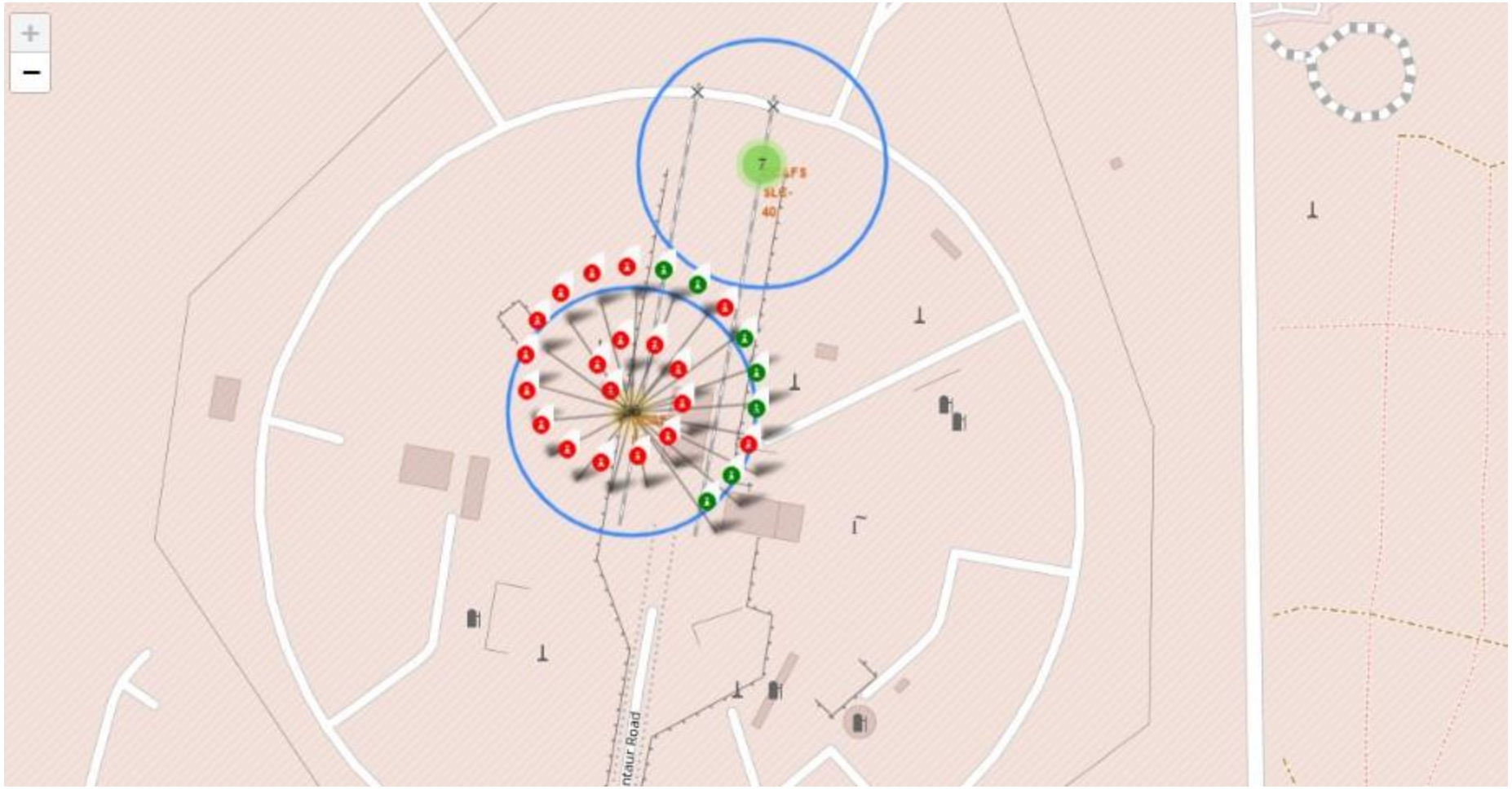
Launch Sites positions on Global Map

The launches sites are in the South of the country, as close as possible to the equator. Nasa explains the reason: “If a spacecraft is launched from a site near Earth's equator, it can take optimum advantage of the Earth's substantial rotational speed. Sitting on the launch pad near the equator, it is already moving at a speed of over 1650 km per hour relative to Earth's center”.

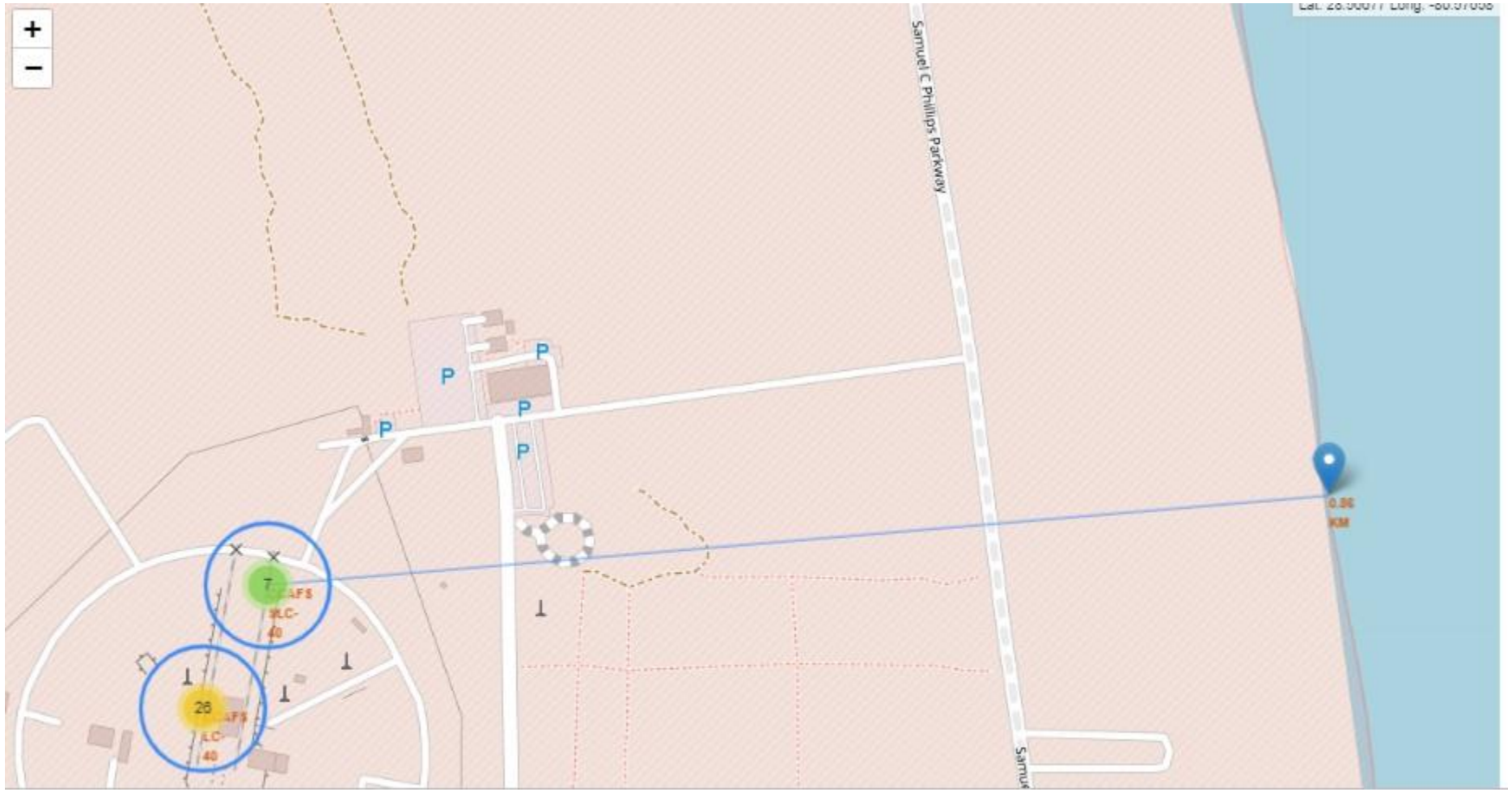
The launch sites need to be near railways, highways and cities (for logistical reasons), but far enough, at the same time, from cities. The proximity to the coastline is an advantage, in order to lower the probability of casualties in case of failed launches or explosion.

Launch Sites positions on Global Map

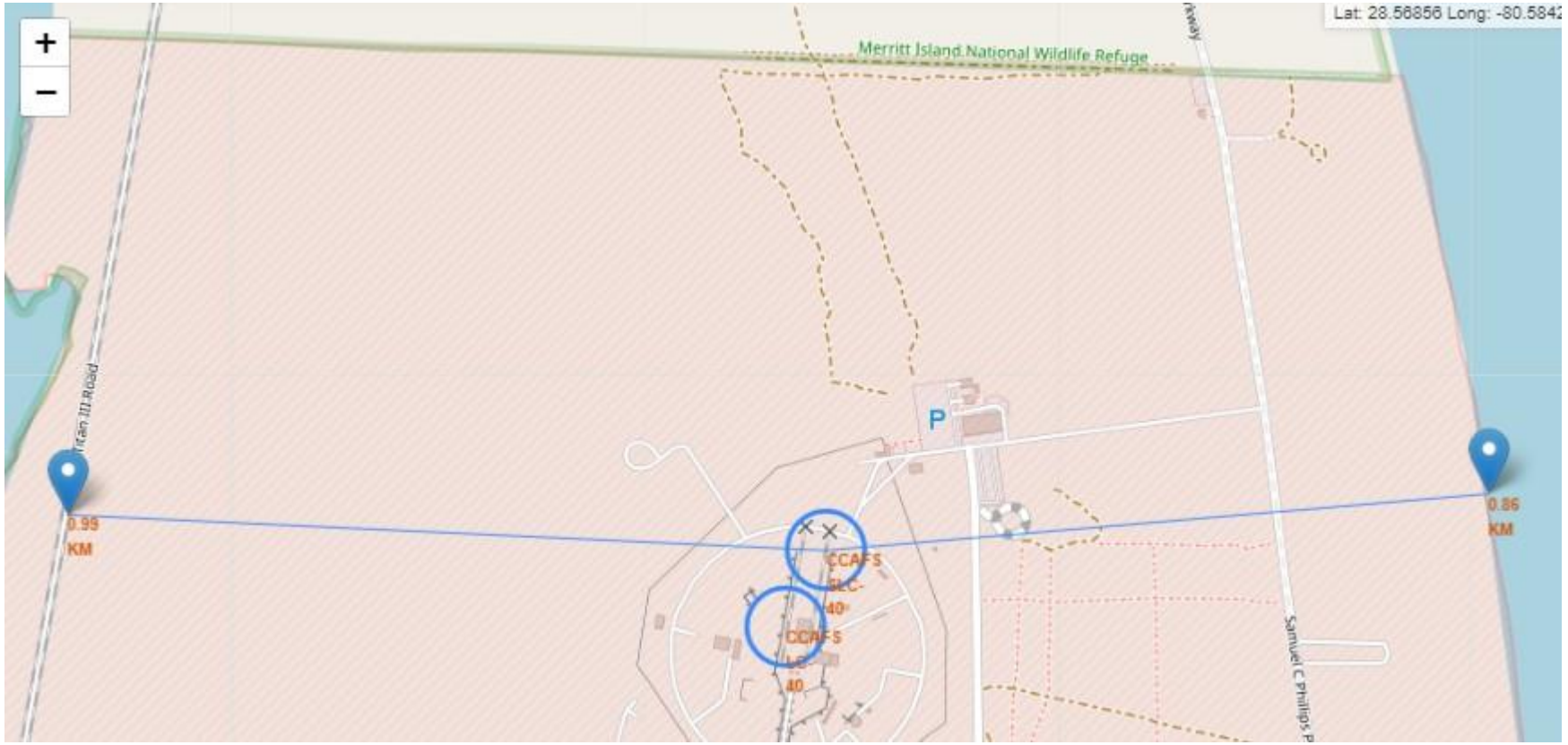
Rate of success/failure for a launch site



Launch Site Nearest Places - Coastline

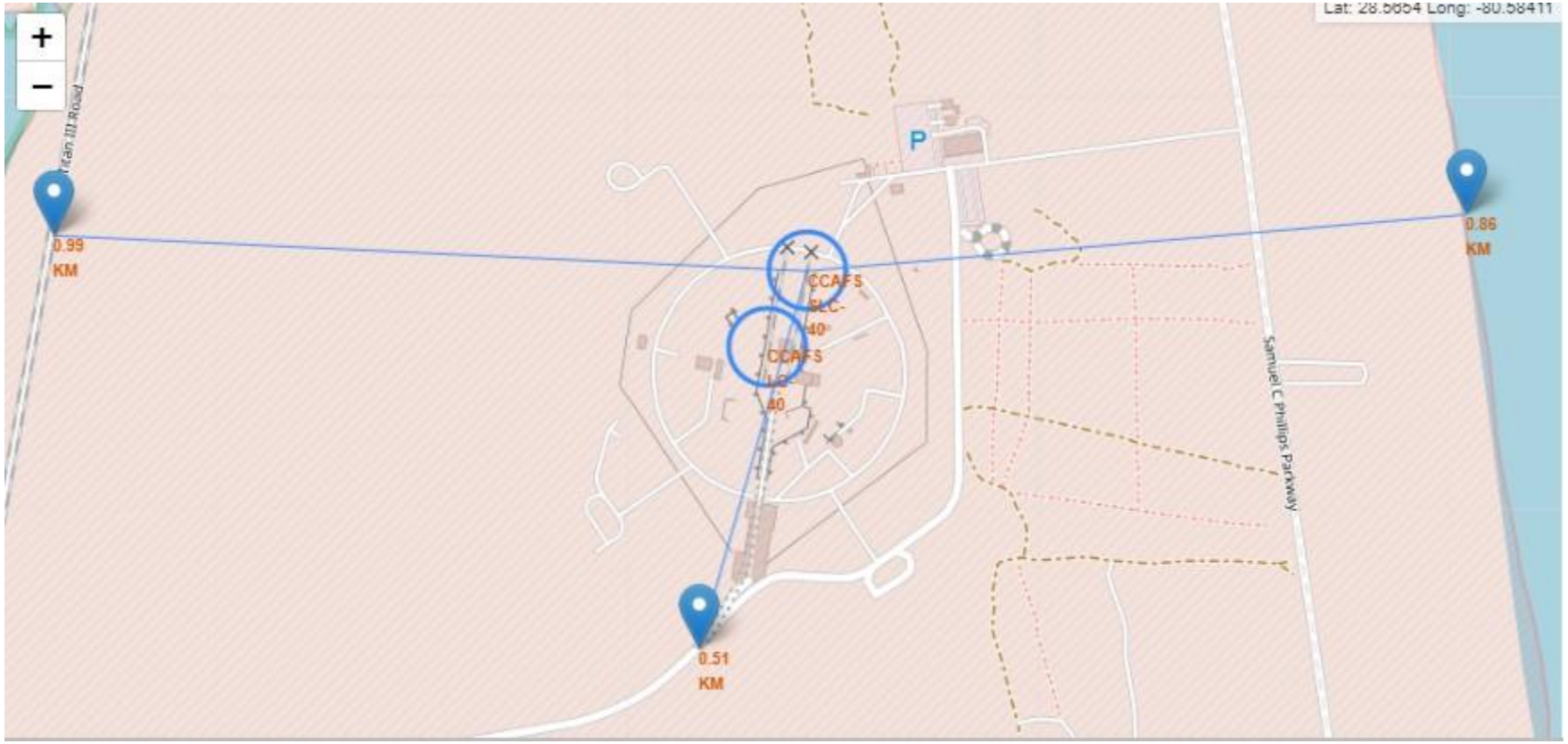


Launch Site Nearest Places - Railway



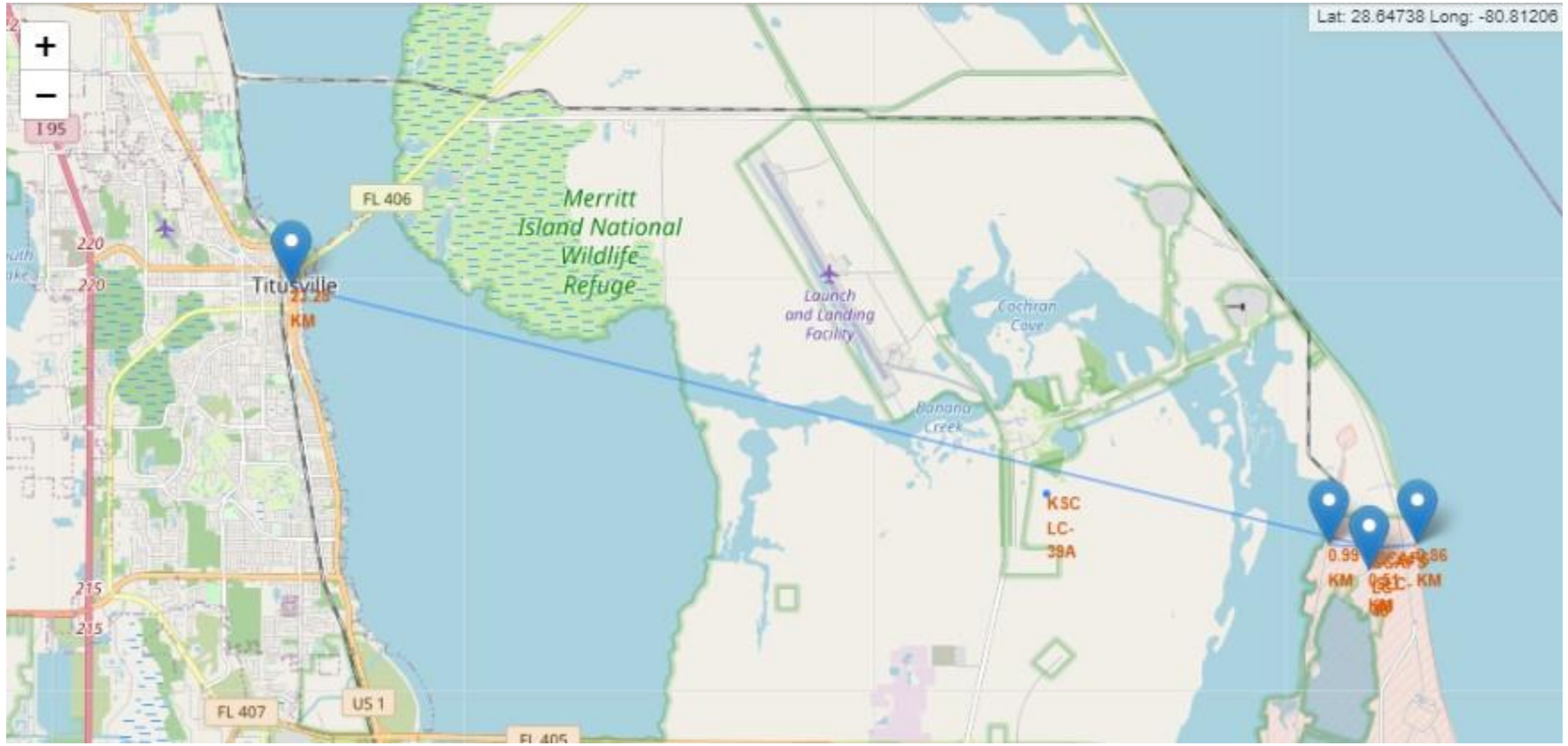
Railway

Launch Site Nearest Places – Access Road



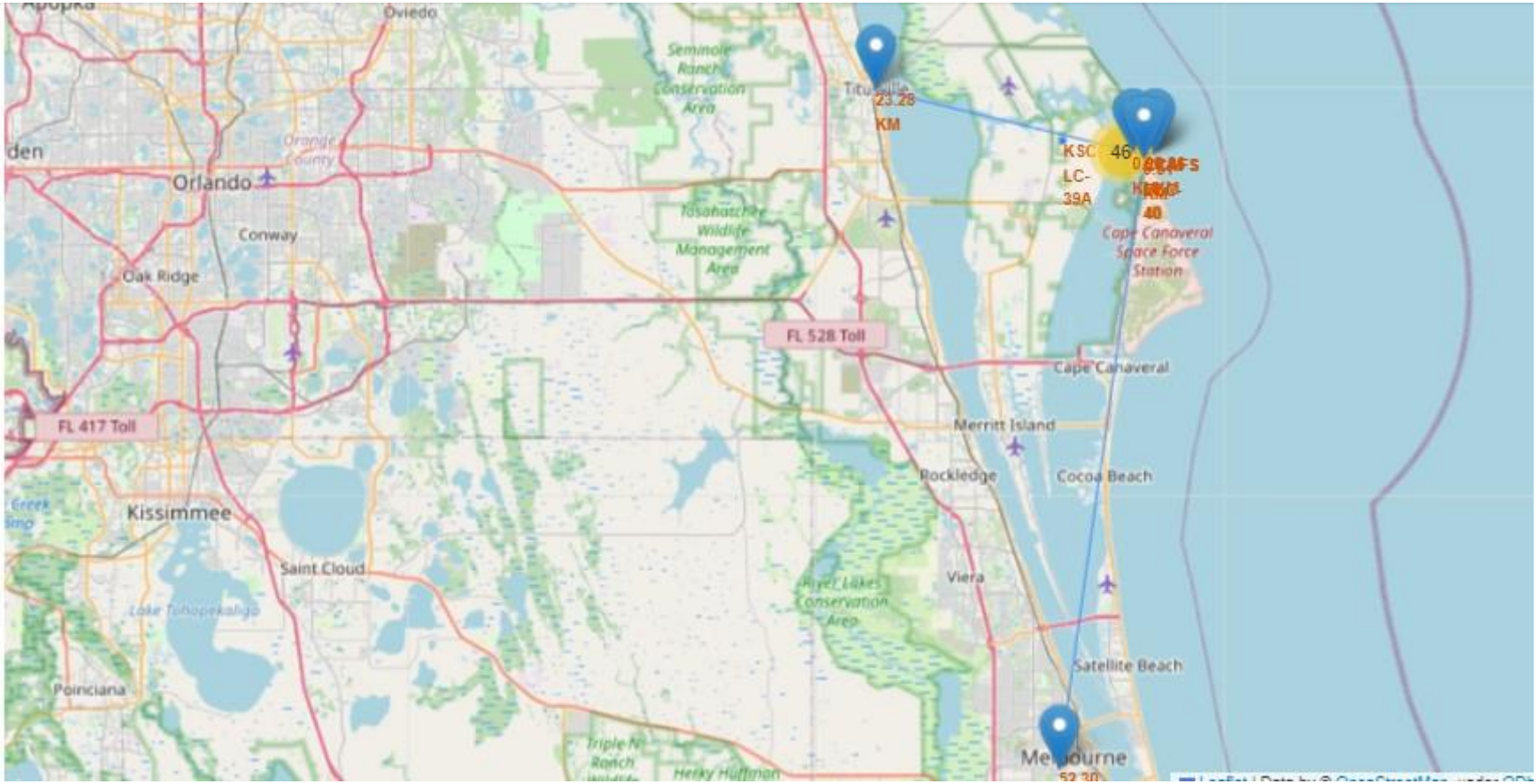
Highway

Launch Site Nearest Places - Highway

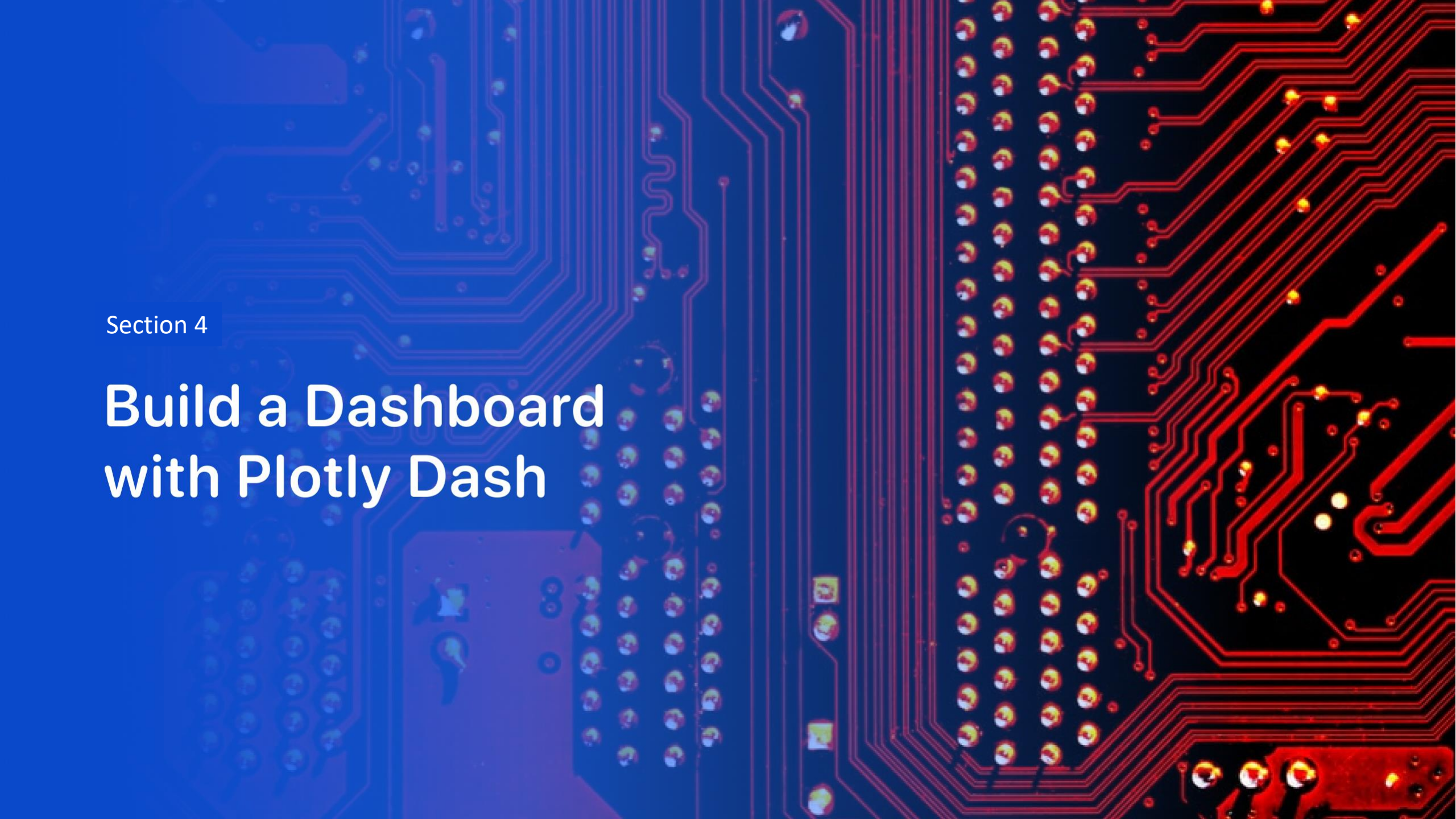


Highway

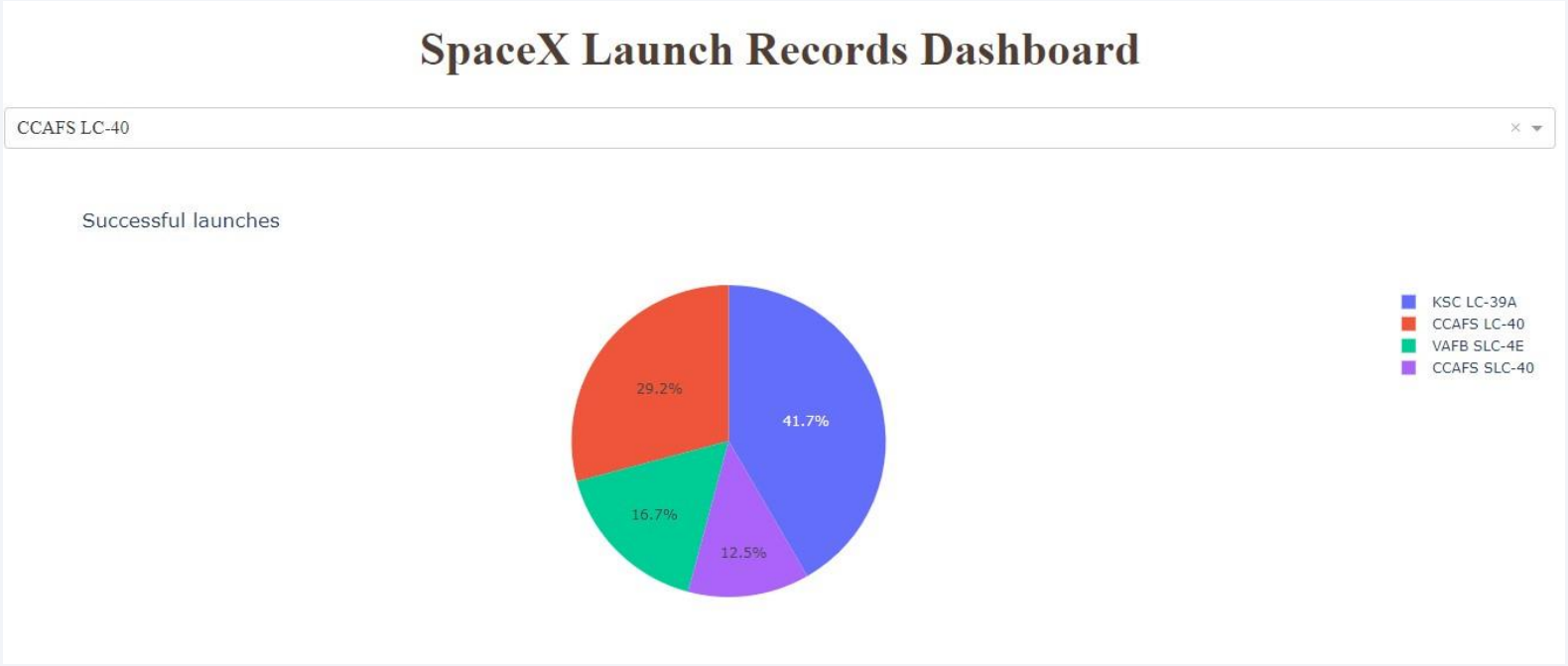
Launch Site Nearest Places - Cities



Highway

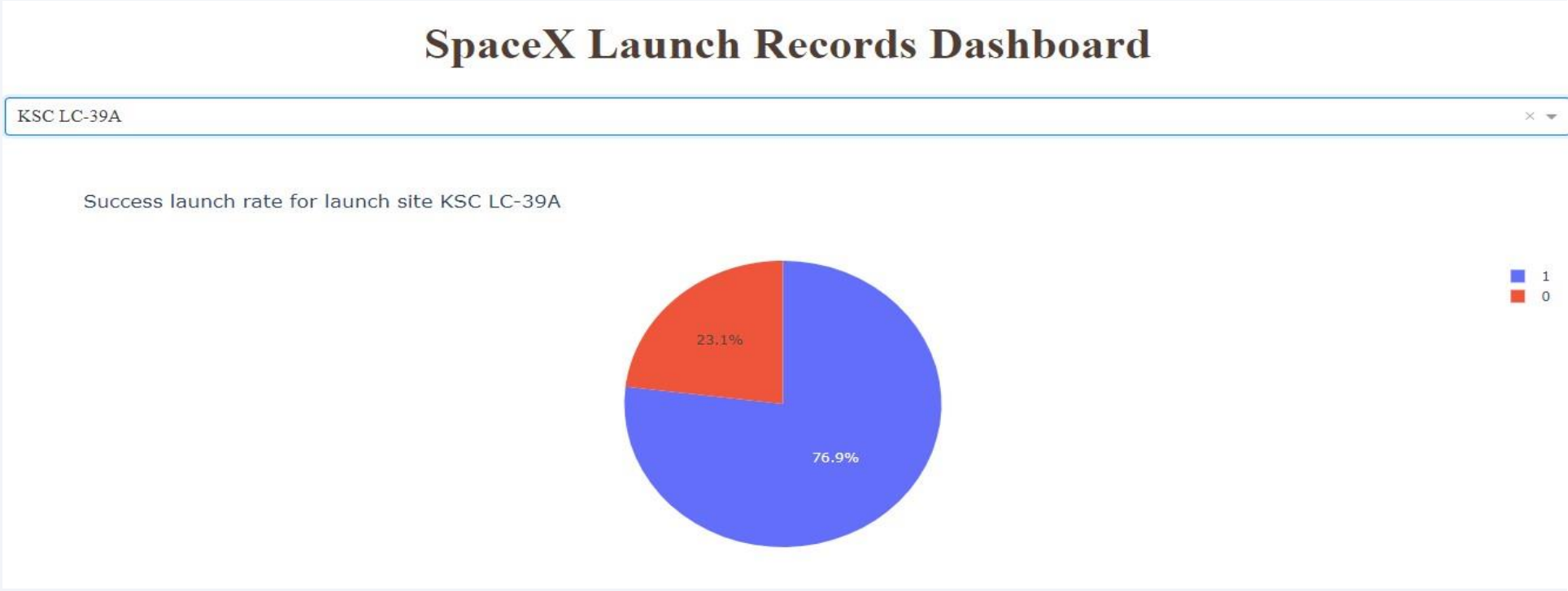

Section 4
Build a Dashboard with Plotly Dash

Success rate for each launch site



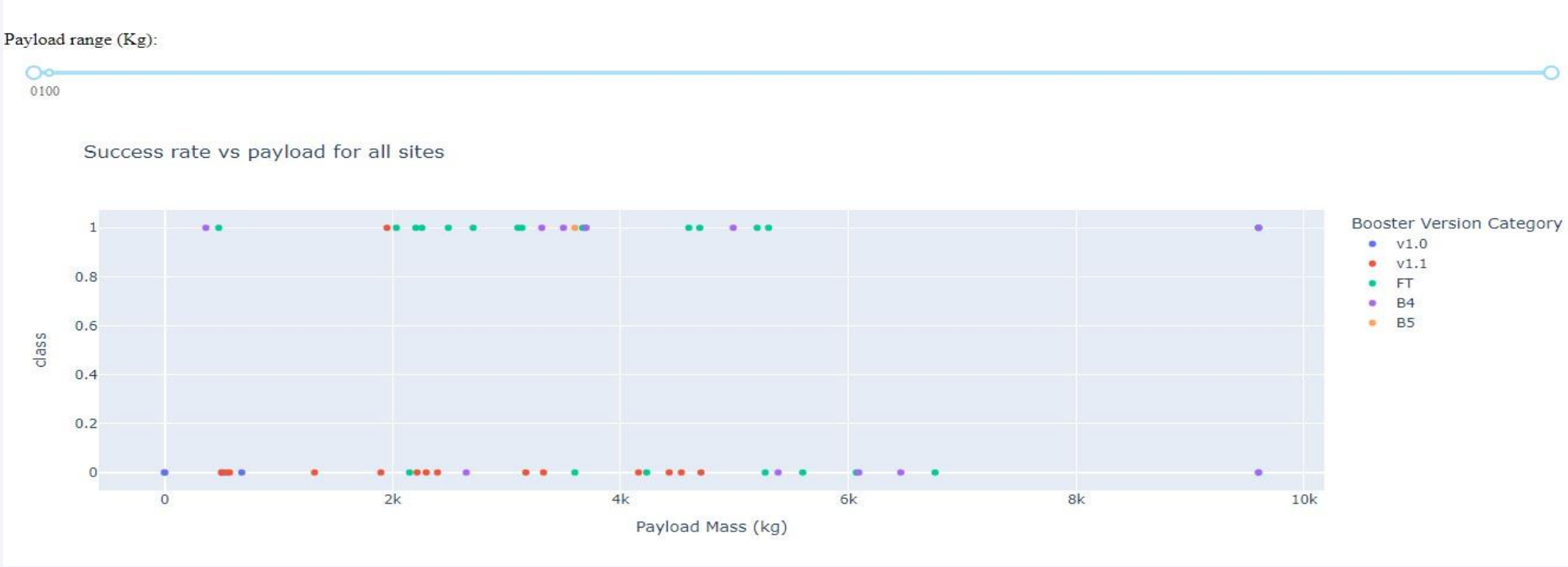
The highest percentage of successful launches were at the site KSC LC39-A.

Highest Success Ratio Site



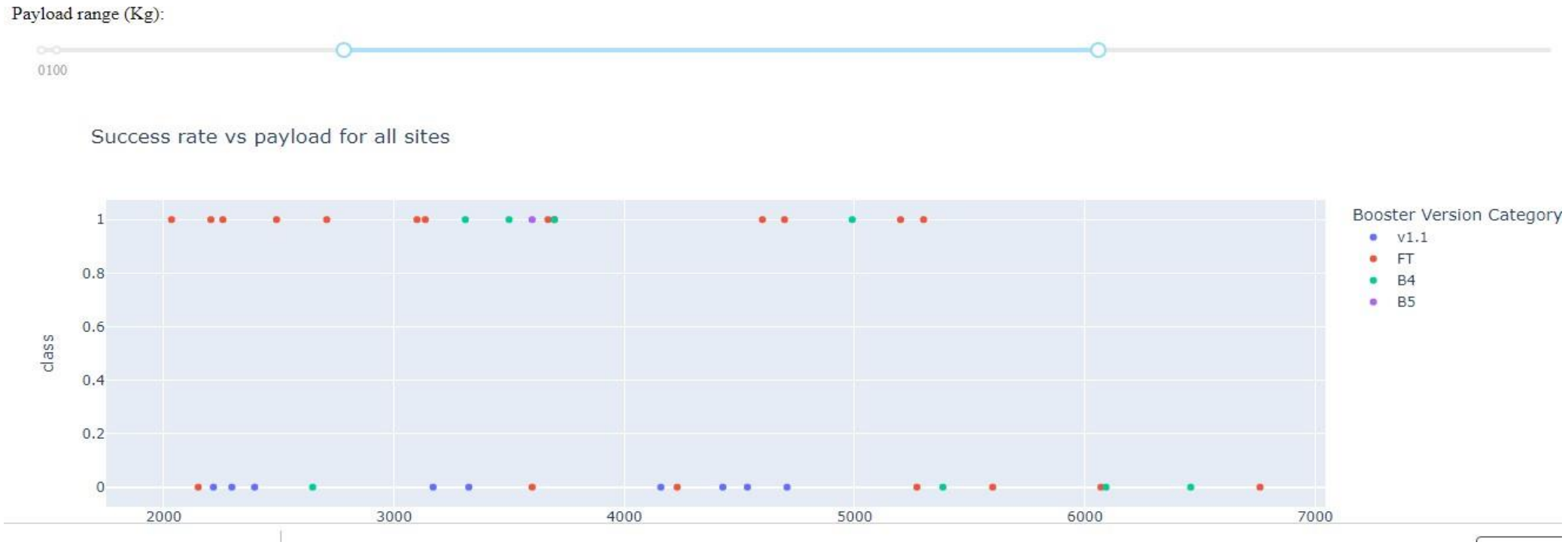
According to the chart, the highest success ratio site's percentage of successful missions is of 76.9%.

Payload vs Launch Outcome - All Sites



Most of the successful launches carried a payload below 6,000 kg. Only a mission carrying more than 6,000 kg was successful.

Payload vs Launch Outcome –All Sites Different Payloads



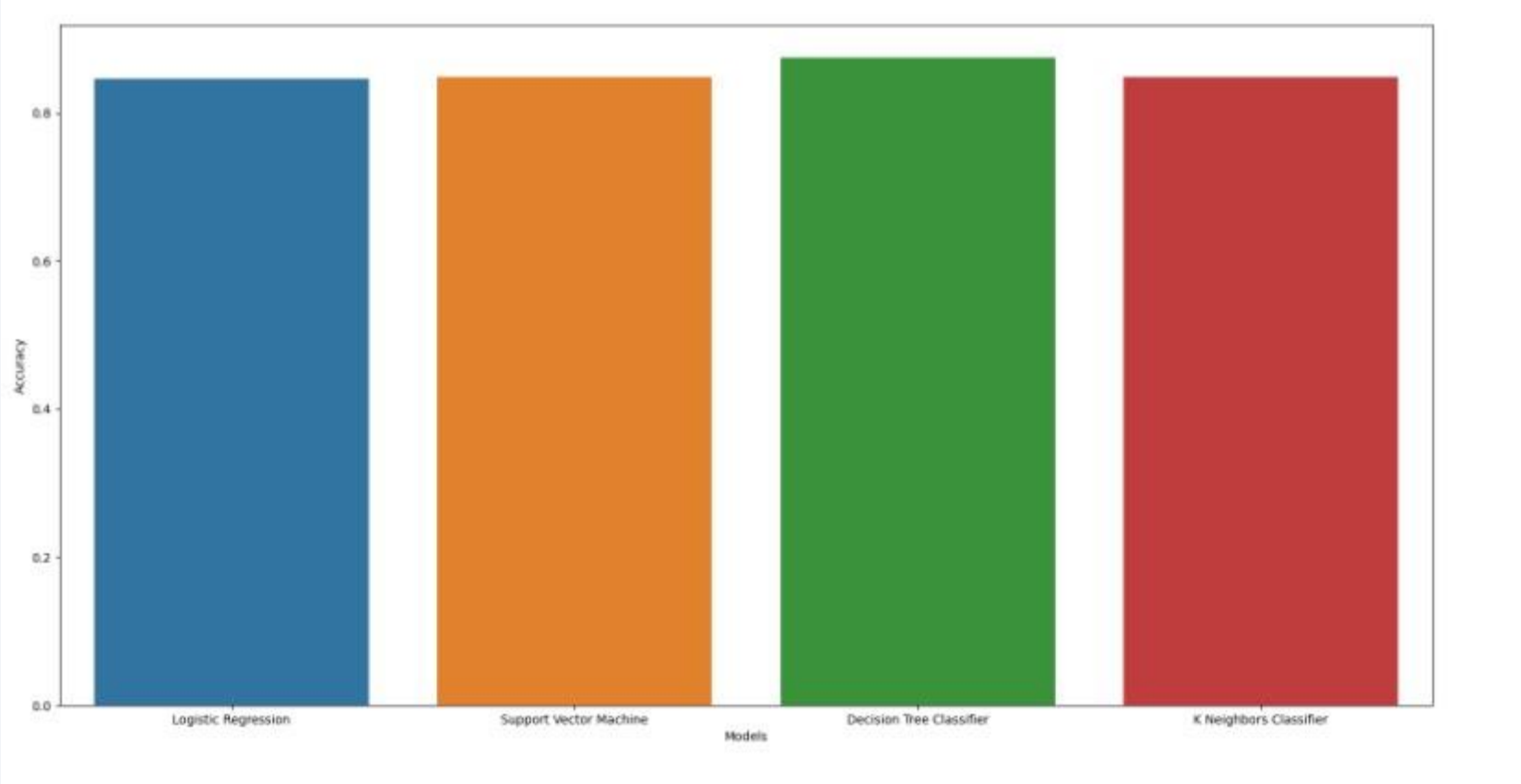
Most of the successful launches carried a payload below 6,000 kg. Only a mission carrying more than 6,000 kg was successful.

Section 5

Predictive Analysis (Classification)

Predictive Analysis Notebook

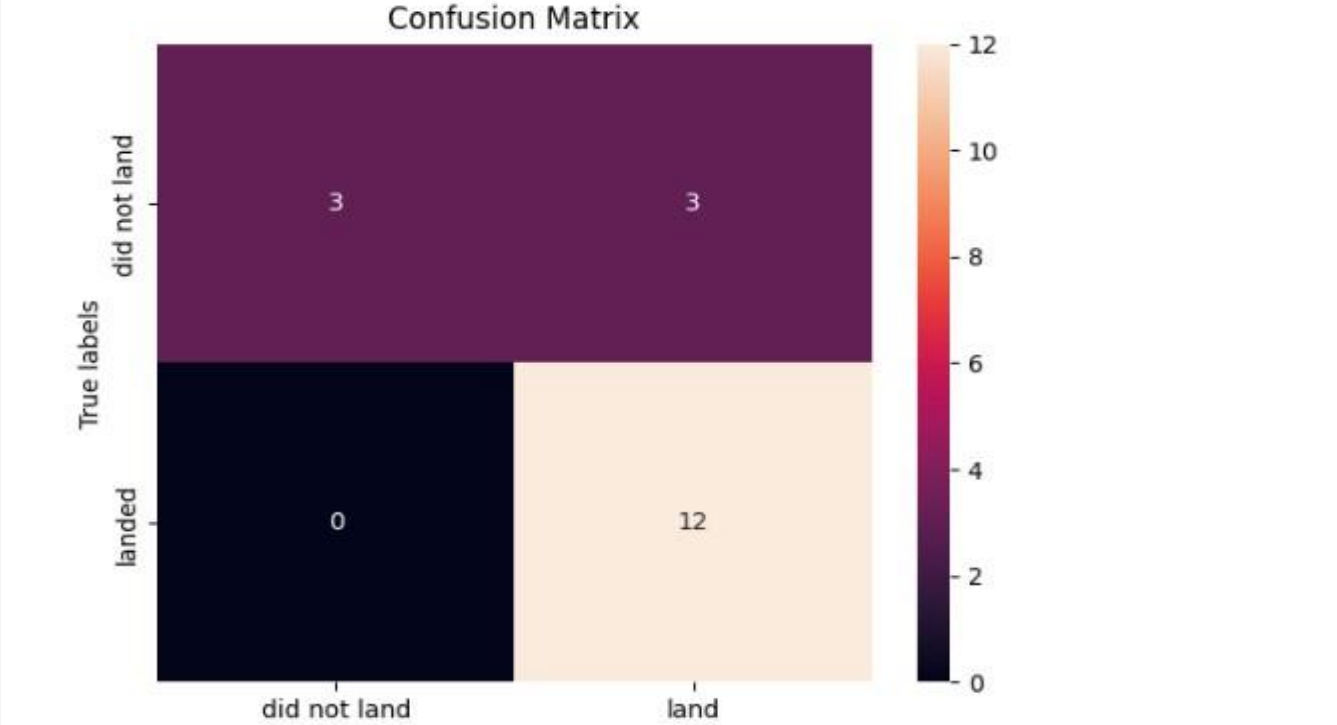
Classification Accuracy



The model that performed best is the Decision Tree Classifier (0.875).

Confusion Matrix – Best Model (Decision Tree)

```
yhat = tree_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



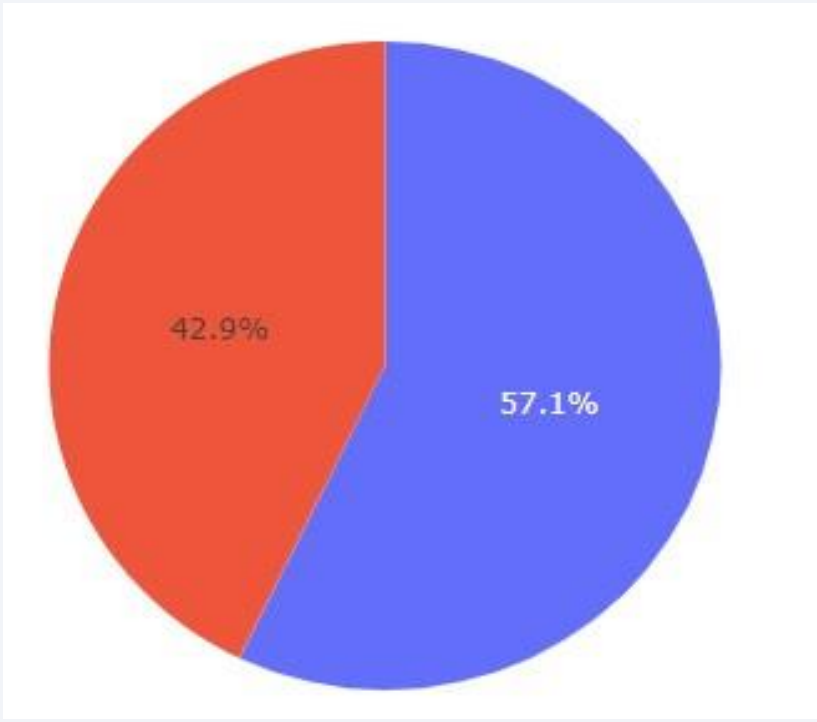
The model correctly identified all the true cases (no false negatives), but yielded 3 false positives (50% of total cases).

Conclusions

- Despite the frequent failures in the beginning of its space program, SpaceX has managed to improve their rockets which resulted in a high rate of success lately.
- The outcome of a launch depends on the chosen orbit and the carried payload (the success rate is higher for relatively low orbits and lower payloads – 6,000 kg or less)
- The company mostly conducts tests on lower orbits in order to improve the performance of its rockets, launching, from time to time, missions on higher orbits and assess the results.
- The 4 classification models used in the predictive analysis, yielded similar results (around 85% accuracy) in predicting the outcome of a mission, therefore, we could rely on them to predicting the outcome of future launches.
- The overall trend throughout the whole period of activity is definitely positive, given the average rate of success for the past 4 years was above 85%, we could conclude that SpaceX will reach its main objective to land on Mars in the near future.

Appendix

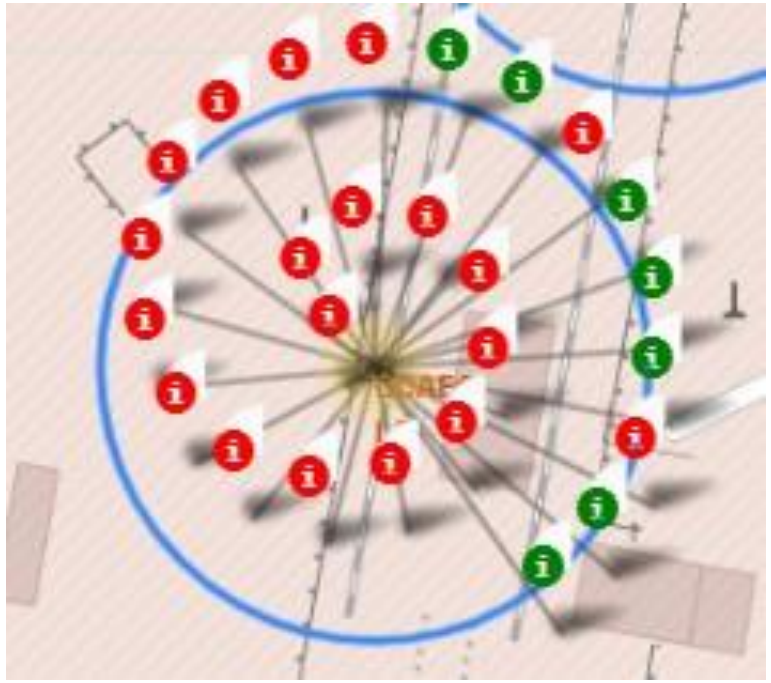
Success rate for site CCAFS SLC-40

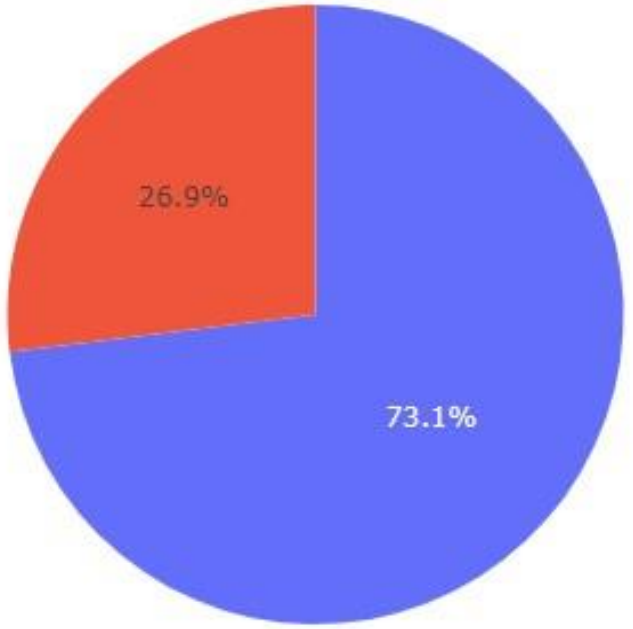


Success rate 42.9% (3 out of 7 launches). Pie chart: blue – failure, orange - success

Appendix

Success rate for site CCAFS LC-40

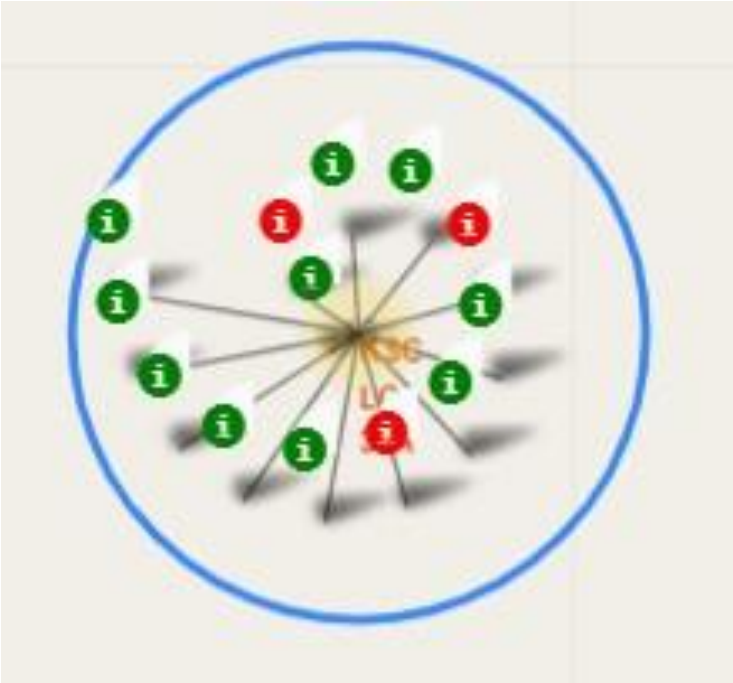


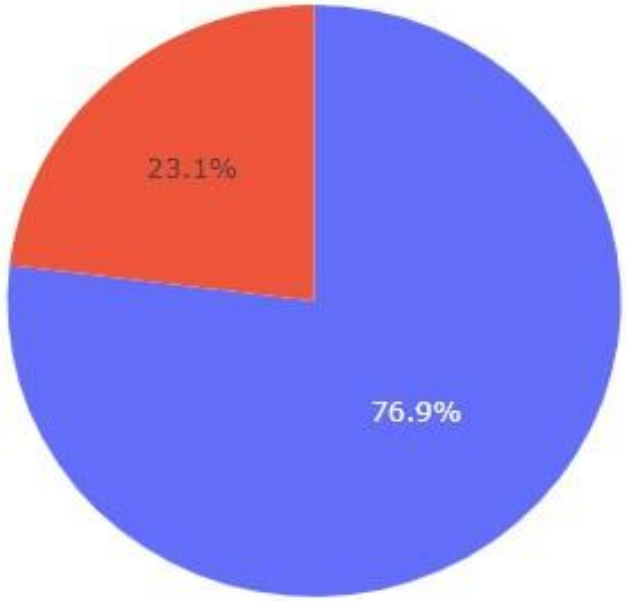


Success rate 26.9% (7 out of 26 launches). Pie chart: blue – failure, orange - success

Appendix

Success rate for site KSC LC-39A

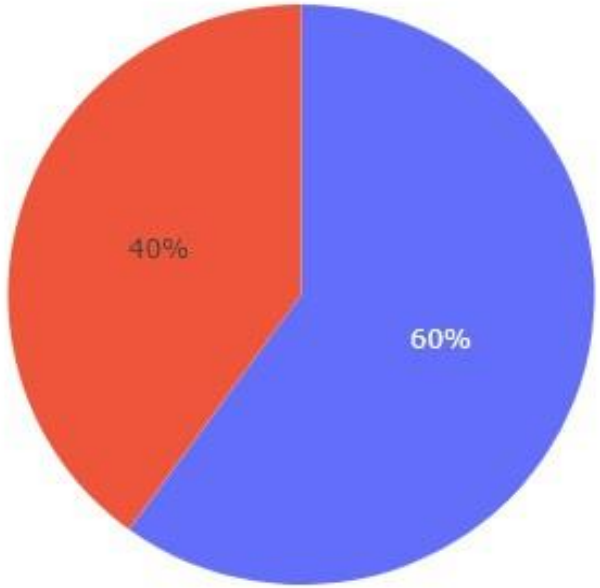




Success rate 76.9% (10 out of 13 launches). Pie chart: blue – success, orange - failure

Appendix

Success rate for site VAFB SLC-4E



Success rate 40.0% (4 out of 10 launches). Pie chart: blue – failure, orange - success

Thank you!