IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jaime Olarte
11/11/2022

Outline

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

Executive Summary

Summary of methodologies

- First, data were obtained, then an Exploratory Analysis (using SQL and Interactive Visual Analytics and Dashboards) was performed, followed by a geographical analysis using maps, and finally the best prediction was sought.

Summary of all results

- In general, after the data was collected and cleaned, EDA showed us relationships between the payload mass, the number of launches, and the success rate that had an impact on the performance of the project, just as geographical analysis showed that geographical access to resources was also important to a good success rate. At the end, four models were created in order to obtain the best accuracy with predictions.

Introduction

This work was carried out as a requirement to obtain the professional Data Scientist certificate provided by IBM on the Coursera platform, the data for the development of this work was taken from the SpaceX API and Wikipedia.

This presentation is meant to show the results of the work that was done over the course of a month. During that time, different explorations were done to produce a model that can automate part of the process of building a SPACE-X rocket for a hypothetical competition.

In these terms, the main available elements were analyzed, and the development of a model that can predict the success or failure of a launch was achieved in 94% of cases.

Section 1

Methodology

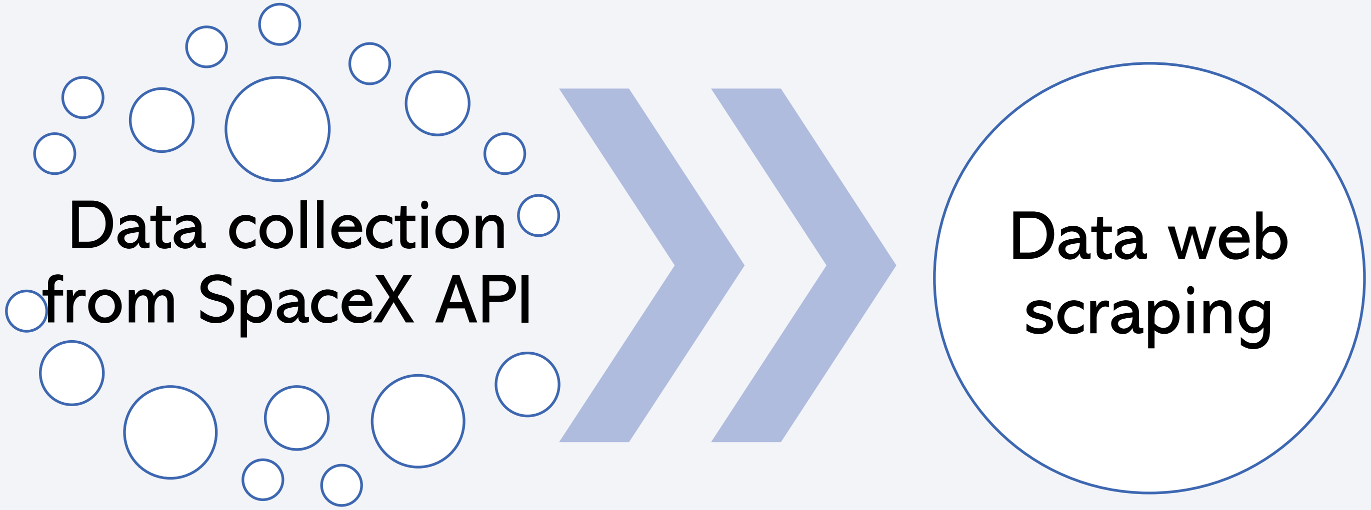
Methodology

Executive Summary

- Data collection methodology:
  - Data web scraping from Wikipedia
  - Data collection from SpaceX API
- Perform data wrangling
  - Data was cleaned, null values were replaced, categorical data was replaced by numeric data.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Data was taken from SpaceX API and Data web scraping, data was cleaned, null values were replaced, categorical data was replaced by numeric data.



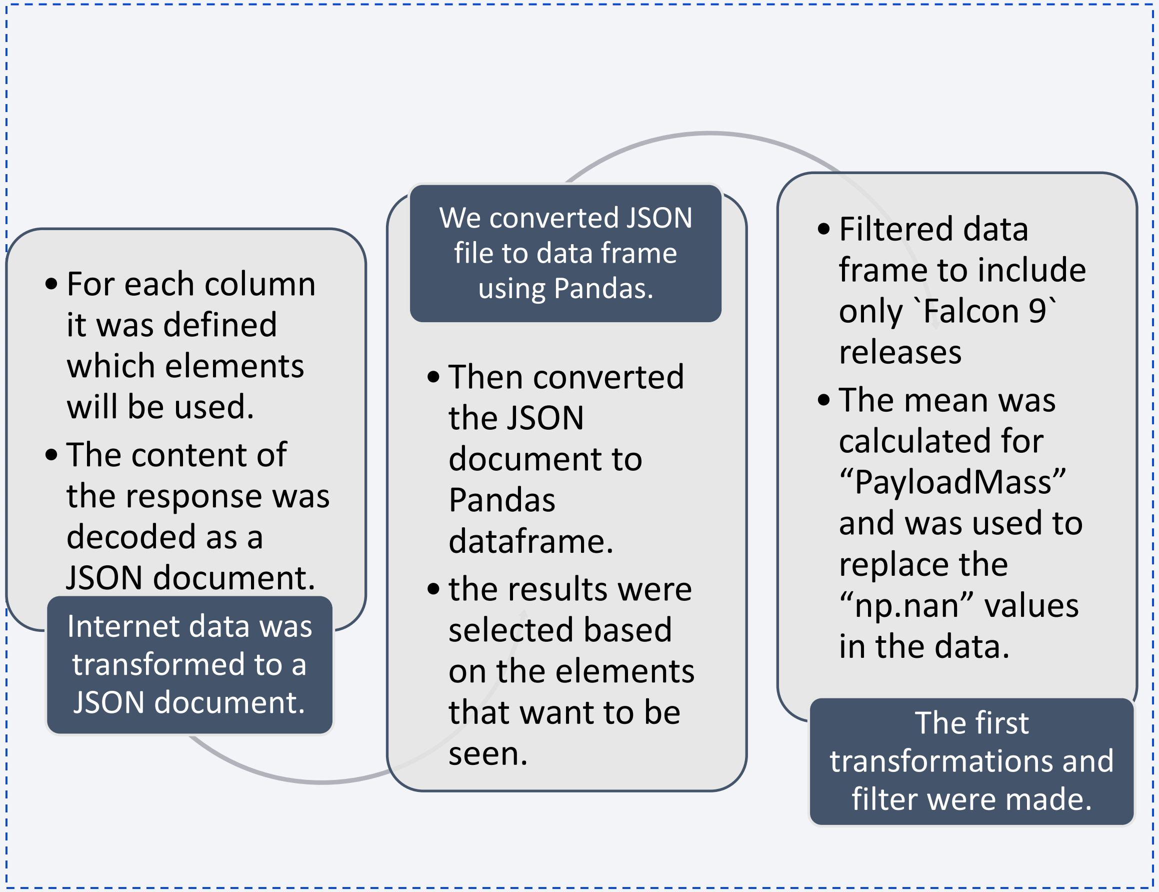
- Internet data was transformed to a JSON document.
- We converted JSON file to data frame using Pandas.
- The first transformations were made.

- A Beautiful Soup were created, and all column/variable names were extracted.
- The “launch_dict” was filled up with launch records extracted from table rows.
- Table were parsed and converted into a Pandas data frame

Data Collection – SpaceX API

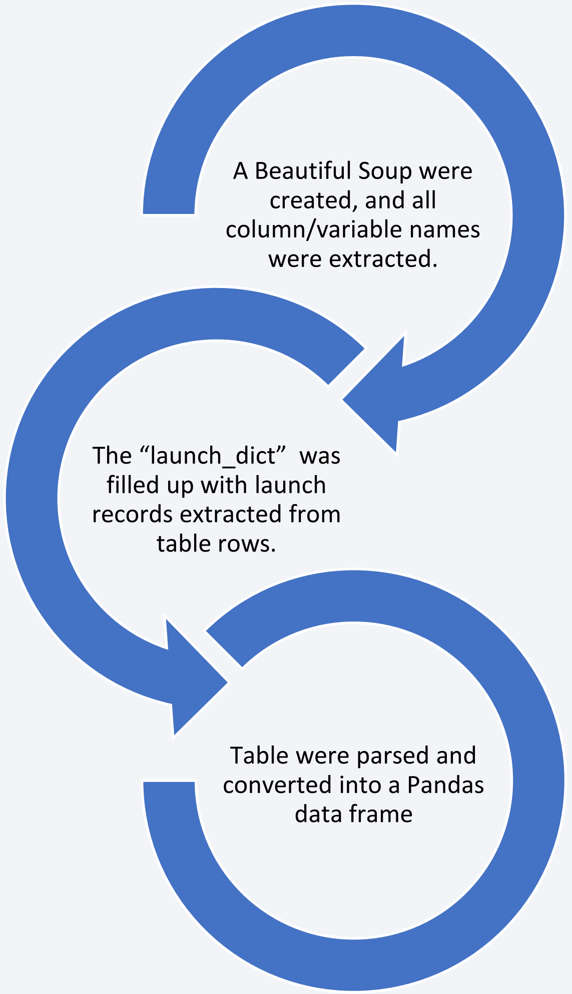
- The Jupyter notebook is hosted on GITHUB and can be peer reviewed at the following link:
  - https://github.com/jaolartem/capstone-data-sciences/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

- The diagram located to the right of this text summarizes the process carried out for this purpose.



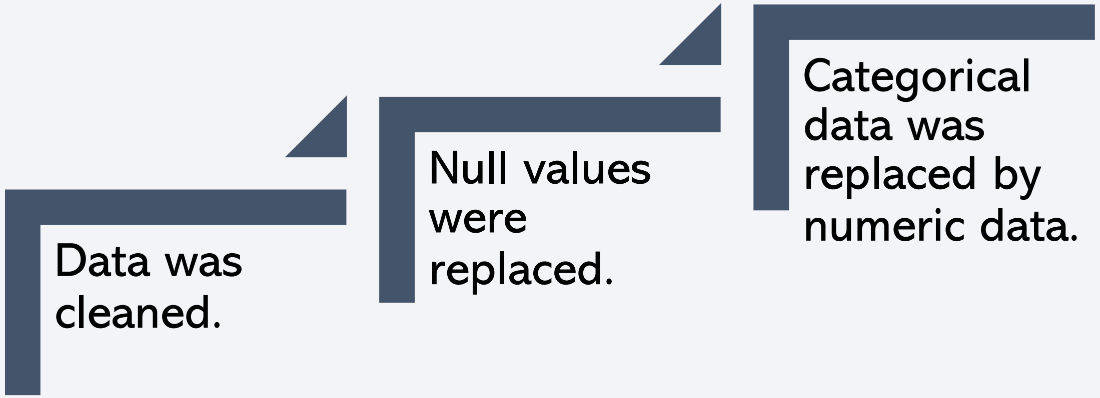
Data Collection - Scraping

- The Jupyter notebook is hosted on GITHUB and can be peer reviewed at the following link:
  - https://github.com/jaolartem/capstone-data-sciences/blob/main/jupyter-labs-webscraping.ipynb
- The diagram located to the right of this text summarizes the process carried out for this purpose.



Data Wrangling

- The Jupyter notebook is hosted on GITHUB and can be peer reviewed at the following link:
  - https://github.com/jaolartem/capstone-data-sciences/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb
- The diagram located to the right of this text summarizes the process carried out for this purpose.



EDA with Data Visualization

- The Jupyter notebook is hosted on GITHUB and can be peer reviewed at the following link:

  https://github.com/jaolartem/capstone-data-sciences/blob/main/jupyter-labs-eda-dataviz.ipynb

The relationship between flight number and launch site was visualized in order to see how different launch sites have different success rates.

The relationship between payload in kg and launch site was visualized to be able to see the influence of payload on success rate.

The relationship between the success rate of each orbit type was visualized in order to see the influence of each one on the success rate.

The relationship between flight number and orbit type was visualized, which allowed us to understand the evolution of the success rate for each kind of orbit.

The relationship between payload and orbit type was visualized because it was necessary to contrast the carrier capacity with respect to each kind of orbit.

The launch success yearly trend was visualized due to the importance of this trend in terms of the evolution of successful.

EDA with SQL

1. Discover the names of the lunch locations analyzed for this project.

2. Five records with launch site names beginning with "CCA" were discovered.

3. The total mass of payload carried by NASA (CRS) boosters was determined.

4. The average mass of the payload carried by booster version F9 v1.1 was determined.

5. The date when the first successful landing on a ground pad was achieved

6. The boosters that are successful in drone ships and have payload masses greater than 4000 but less than 6000 kilograms have been loaded.

7. The total number of successful and unsuccessful mission outcomes was queried.

8. The versions of boosters that have carried the heaviest payloads were presented.

9. The 2015 landing failures of the drone ship were discovered.

10. Between 2010-06-04 and 2017-03-20, the number of landing outcomes (such as Failure (drone ship) and Success (ground pad)) were ranked in descending order.

The Jupyter notebook is hosted on GITHUB and can be peer reviewed at the following link:

https://github.com/jaolartem/capstone-data-sciences/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

All launch sites were marked on a map whit a circle so that all launch locations and their respective advantages could be visualized.

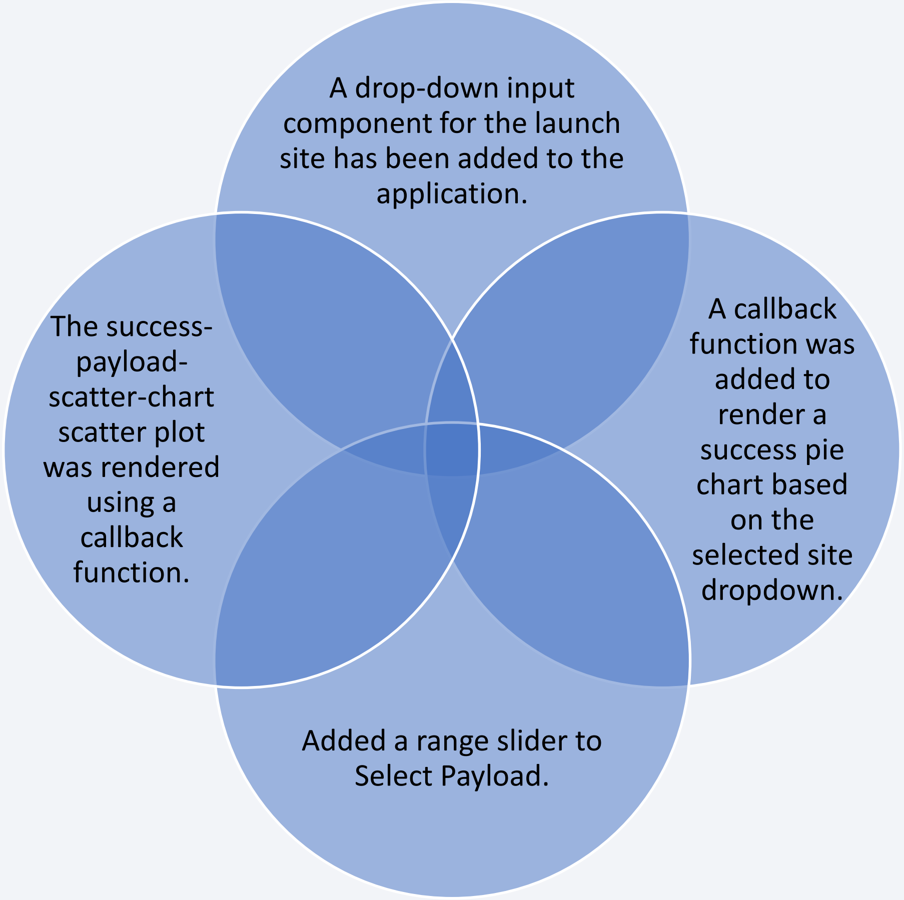
The successful or unsuccessful launches of each site on the map were noted in each circle as a point in order to evaluate the performance of each site.

To determine access to transport routes and the coast, the distances between a launch site and its surroundings were computed, and a line was drawn to represent those distances.

The Jupyter notebook is hosted on GITHUB and can be peer reviewed at the following link:

https://github.com/jaolartem/capstone-data-sciences/blob/main/jupyter-labs-eda-dataviz.ipynb

Build a Dashboard with Plotly Dash



The Jupyter notebook is hosted on GITHUB and can be peer reviewed at the following link:

https://github.com/jaolartem/capstone-data-sciences/blob/main/dash_interactivity.py

Predictive Analysis (Classification)

The column Class in data was converted into a NumPy array using the to numpy() method, which was then assigned to the variable Y. Standardize the data in X, then reassign it to the variable X using the provided transform. Using the train test split function, the data X and Y were separated into training and test data. Random state was set to 2 and test size was set to 0.2.

The GridSearchCV object "logreg cv" with cv = 10 is created following the creation of a logistic regression object. To determine the optimal dictionary parameters, the location of the object was used. The method score was used to calculate the test data's precision.

After creating a support vector machine object, the svm cv with cv - 10 GridSearchCV object was created. The object was calibrated so that the optimal dictionary parameters could be determined. The method score was used to calculate the precision of the test data.

tree cv with cv - 10 GridSearchCV was created after a decision tree classifier object was created. In order to determine the optimal dictionary parameters, the object was calibrated. The method score was utilized to calculate the test data's precision.

The knn_cv with cv - 10 GridSearchCV object was created after the k nearest neighbors object was created. In order to determine the optimal dictionary parameters, the object was calibrated. The method score was utilized to calculate the test data's precision.

The Jupyter notebook is hosted on GITHUB and can be peer reviewed at the following link:

https://github.com/jaolartem/capstone-data-sciences/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

It was found that the greater the weight, the less success

There are platforms with a higher success rate

There is a relationship between the type of orbit and the success rate

There is a technological development that is manifested in the reduction of failures throughout the new releases.

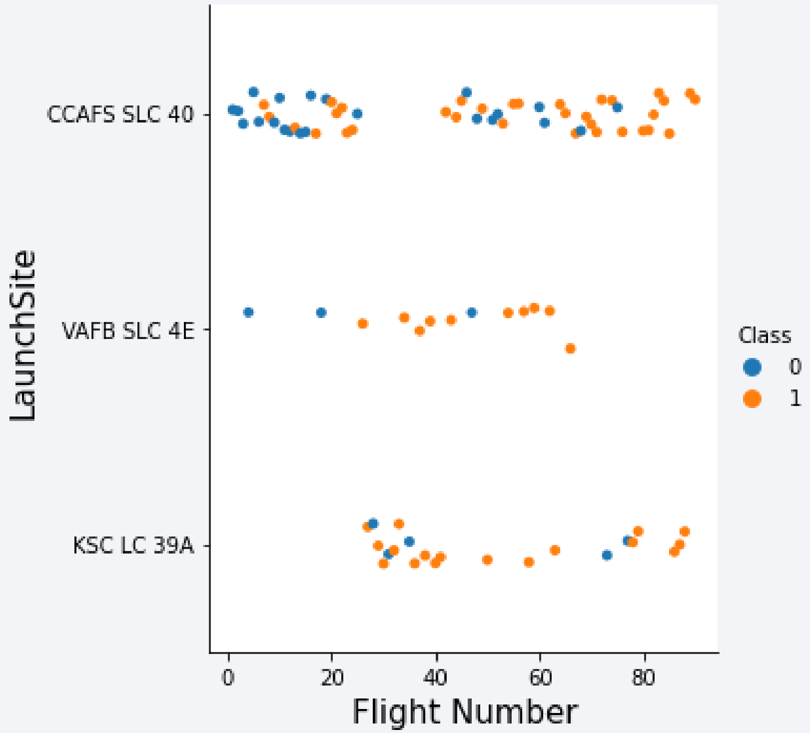
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

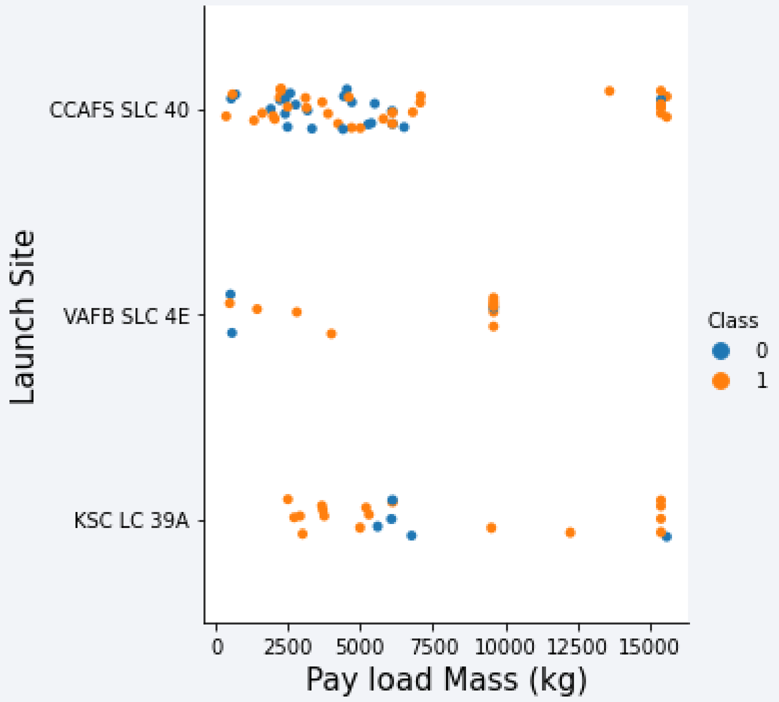
+ Código + Markdown

It is evident that the CCAFS SLC 40 launch site has more failed launches; however, they are concentrated in the beginning and middle, but it has the most launches overall. Now, it is possible that the success rate at the aforementioned location is lower due to the concentration of initial misses.



Payload vs. Launch Site

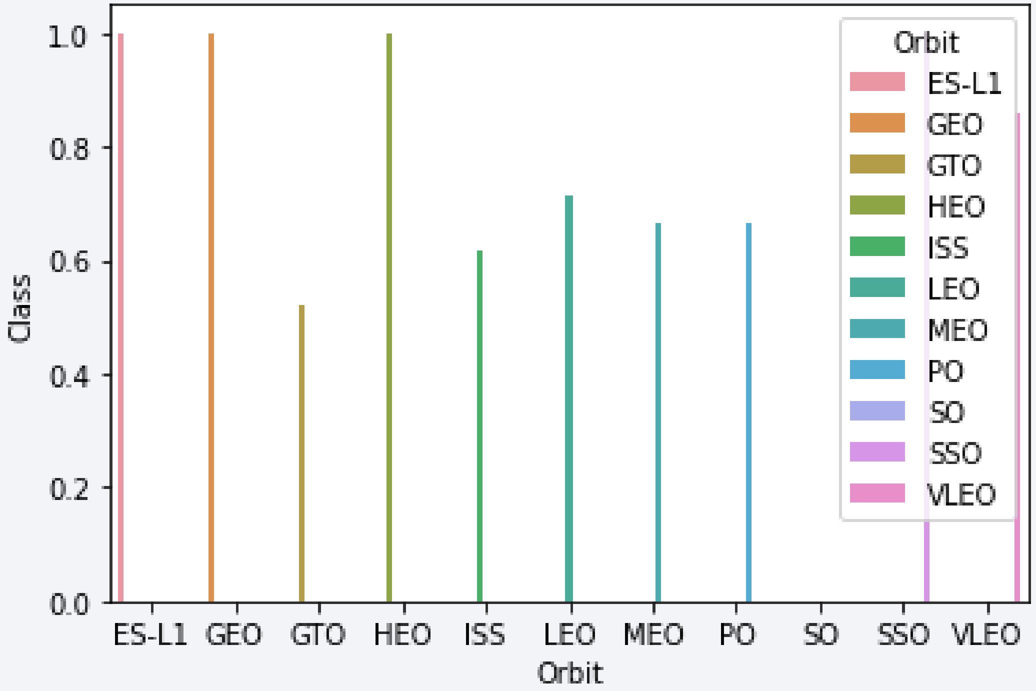
It is evident that the CCAFS SLC 40 launch site has more failed launches; however, they are concentrated in the beginning and middle, but it has the most launches. Now, it is possible that the success rate is lower at the aforementioned location due to the concentration of missed shots near the end.



Success Rate vs. Orbit Type

+ Código + Markdown

The SO orbit has not been successful, whereas the ES-L1, GEO, SSO, and HEO orbits have a 100 percent success rate, and the others range from 61 to 85 percent. The SO Orbit was clearly supplanted by the SSO Orbit, which shares the same characteristics.

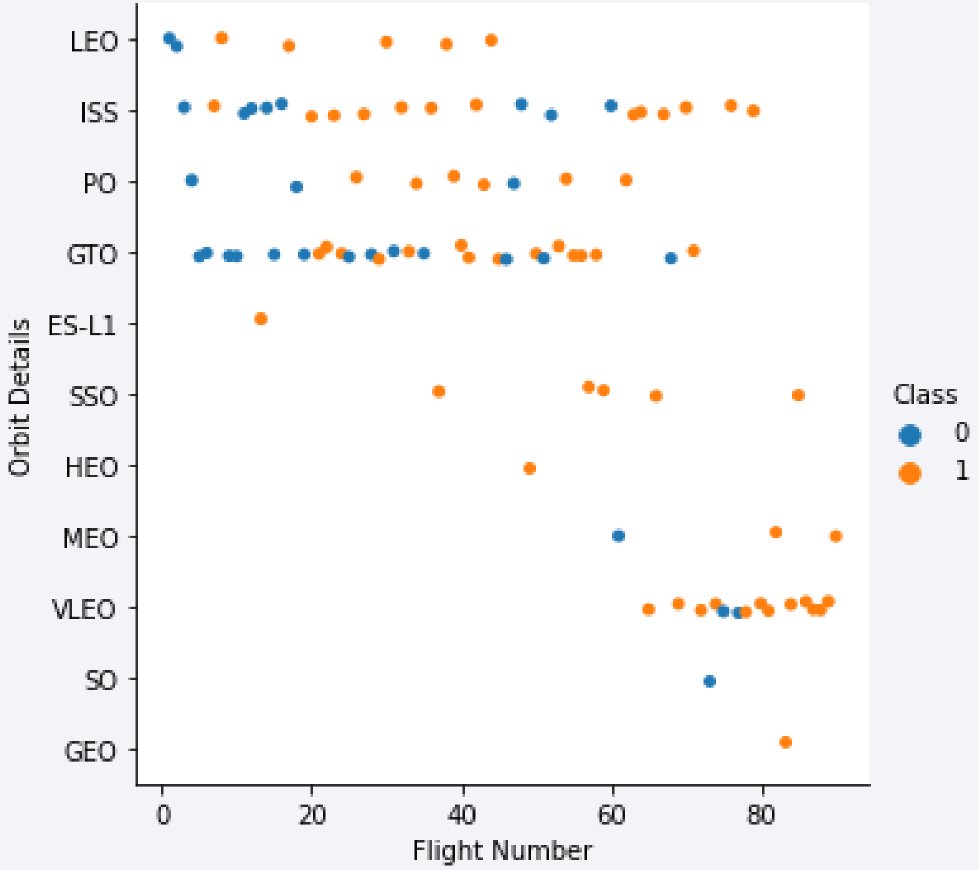


Flight Number vs. Orbit Type

There is a correlation between the success of the LEO orbit and the smaller number of flights associated with it, but the correlation between the success rate and the number of launches appears to be stronger. In other words, it is reliant on technological development.

In contrast, the GTO orbit, which is related to climate management and is located in Ecuador, has more launches. However, there is a significant decline in failures as the number of flights increases, indicating a clear technical advancement.

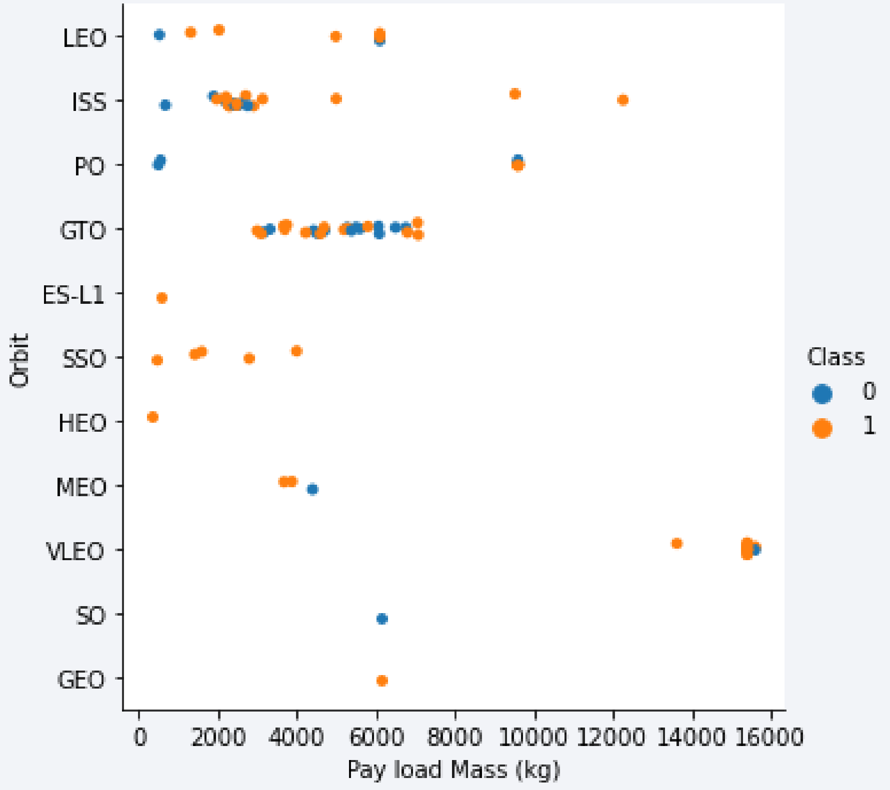
+ Código + Markdown



Payload vs. Orbit Type

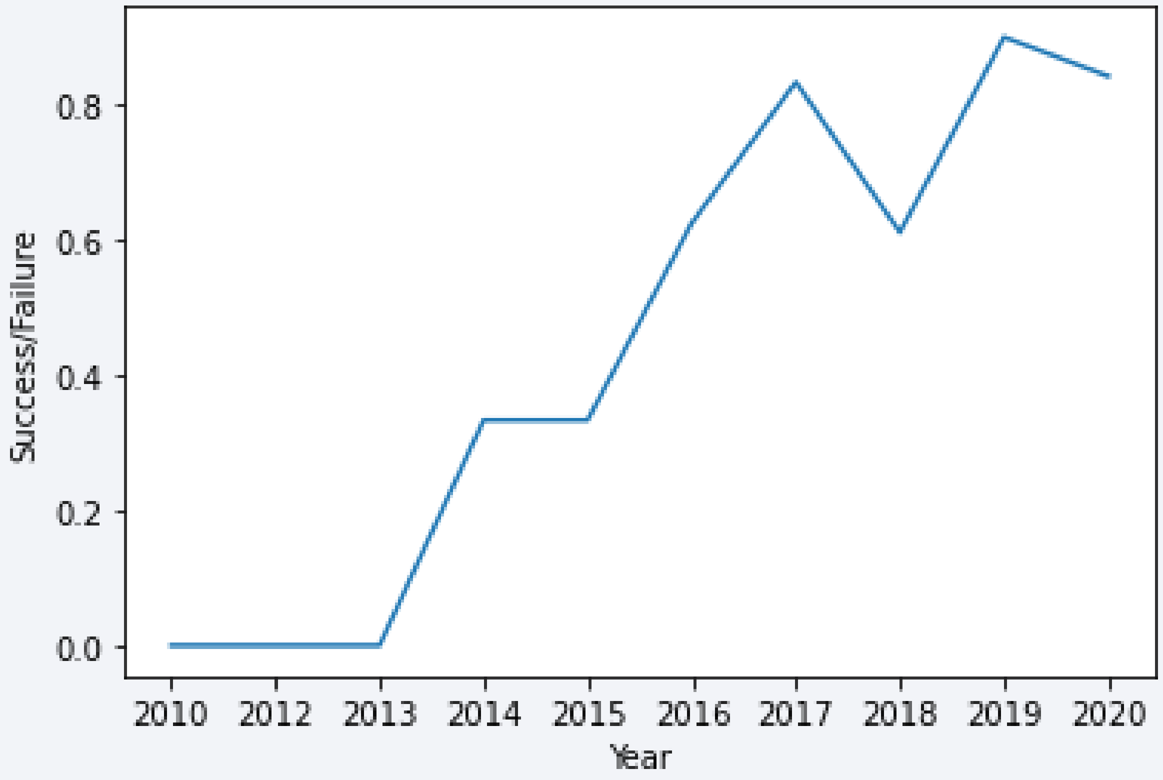
With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.



Launch Success Yearly Trend

The success rate begins to rise in 2013, and continues to rise steadily until 2017, when it begins to decline; however, in 2018 it continued to rise until 2019, when it began to decline.



All Launch Site Names

In this project, five launch sites have been analyzed, and they are listed below.

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

```
%%sql
SELECT UNIQUE (LAUNCH_SITE) AS "NAMES_OF_LAUNCH_SITES"
FROM SPACEXTBL;
```

[4] Python

```
 * ibm_db_sa://bnp18302:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30
Done.
```

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

Launch Site Names Begin with 'CCA'

Below are the details for five platform launch records whose names begin with "CCA."

```
%%sql
SELECT  * FROM SPACEXTBL
WHERE (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5;
```

[19]

Python

```
 * ibm_db_sa://bnp18302:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/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

Total Payload Mass

The total payload mass carried by boosters launched by NASA (CRS) is 45496 kg

Task 3

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

```
%%sql
SELECT SUM (PAYLOAD_MASS__KG_) as NASA_CRS_payloadmass
FROM SPACEXTBL
WHERE (customer) LIKE 'NASA (CRS)';
```

[6] Python

```
 * ibm_db_sa://bnp18302:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30
Done.
```

nasa_crs_payloadmass_avg
45596

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1 is 2534 kg.

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

+ Código + Markdown

```
%%sql
SELECT AVG (PAYLOAD_MASS__KG_) as payloadmass_avg
FROM SPACEXTBL
WHERE (booster_version) LIKE 'F9 v1.1%';
```

[7] Python

```
 * ibm_db_sa://bnp18302:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30
Done.
```

payloadmass_avg
2534

First Successful Ground Landing Date

The date when the first successful landing on a ground pad was achieved is December 22, 2015.

```
%%sql
SELECT MIN(DATE)
FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Success (ground pad)';
```

[8] Python

```
 * ibm_db_sa://bnp18302:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30
Done.
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

There are four boosters that have success in drone ships and have payload masses greater than 4000 but less than 6000.

```
%%sql
SELECT DISTINCT(BOOSTER_VERSION), LANDING__OUTCOME, PAYLOAD_MASS__KG_
FROM SPACEXTBL
WHERE (LANDING__OUTCOME) LIKE 'Success (drone ship)'
AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

[9]

Python

```
 * ibm_db_sa://bnp18302:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

booster_version	landing__outcome	payload_mass__kg_
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600

Total Number of Successful and Failure Mission Outcomes

There are 99 successful and four failed mission outcomes.

```
%%sql
SELECT MISSION_OUTCOME, COUNT (MISSION_OUTCOME)
FROM SPACEXTBL
GROUP BY (MISSION_OUTCOME);
```

[10] Python

```
 * ibm_db_sa://bnp18302:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30
Done.
```

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

There are 12 booster versions that have carried the maximum payload mass.

```
%%sql
select UNIQUE BOOSTER_VERSION as boosterversion, PAYLOAD_MASS__KG_
from SPACEXTBL
where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

[11] Python

```
 * ibm_db_sa://bnp18302:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

boosterversion	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

2015 Launch Records

In 2015, there were two failed landing outcomes on the drone ship; their booster versions are F9 v1.1 B1012 and F9 v1.1 B1015, and both were launched in CCAFS LC-40.

```
%%sql
SELECT DATE, BOOSTER_VERSION,LAUNCH_SITE,
LANDING__OUTCOME
FROM SPACEXTBL
WHERE (LANDING__OUTCOME) LIKE 'Failure (drone ship)'
AND Date BETWEEN '2015-01-01' AND '2015-12-31'
```

```
 * ibm_db_sa://bnp18302:***@125f9f61-9715-46f9-9399-
c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30
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)

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

The count of landing outcomes (such as failure (drone ship) or success (ground pad)) between the dates 2010-06-04 and 2017-03-20, in descending order.

```
%%sql
SELECT  LANDING__OUTCOME AS "LANDING_OUTCOME", COUNT (LANDING__OUTCOME) AS "COUNT" FROM  SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING__OUTCOME
ORDER BY COUNT (LANDING__OUTCOME) DESC
```

[27] ✓ 0.3s Python

```
 * ibm_db_sa://bnp18302:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

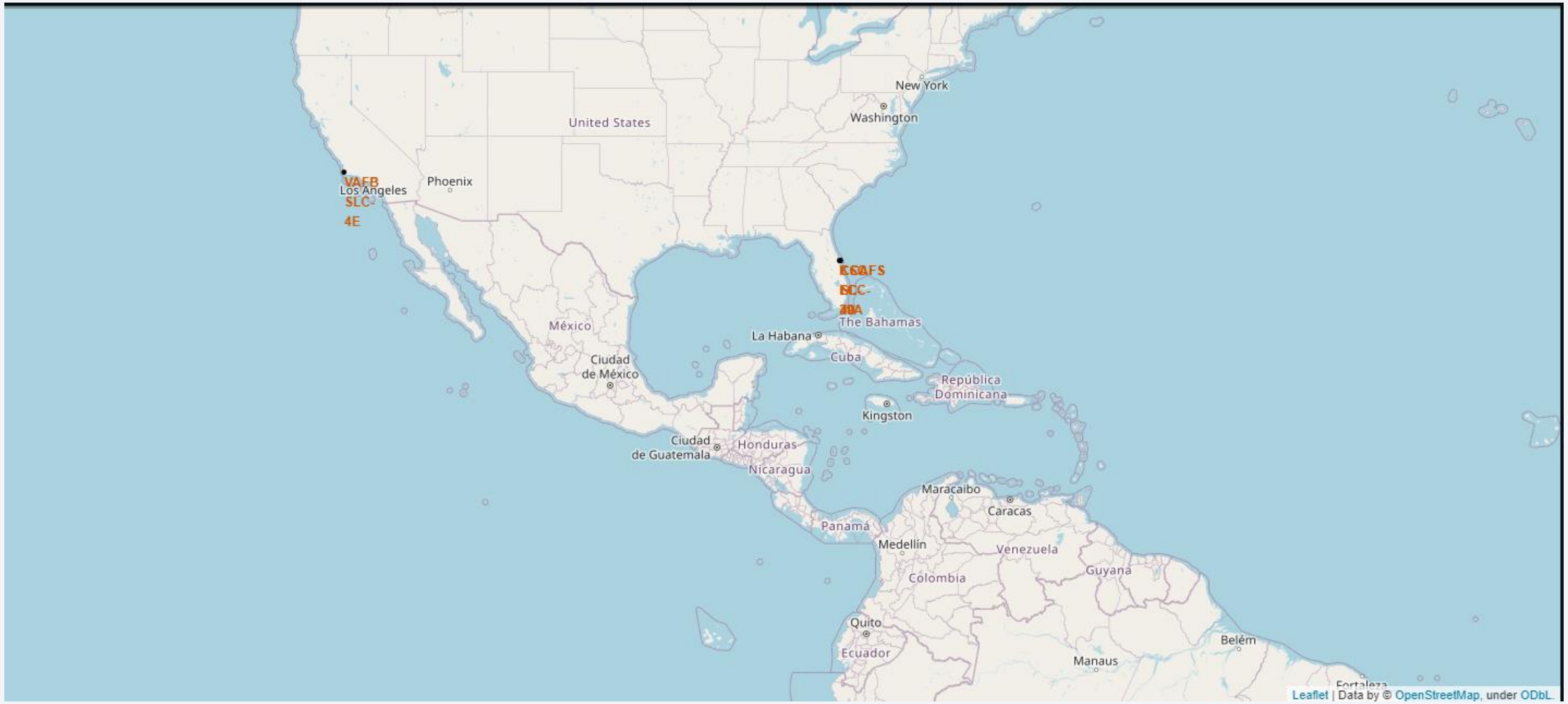
landing_outcome	COUNT
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

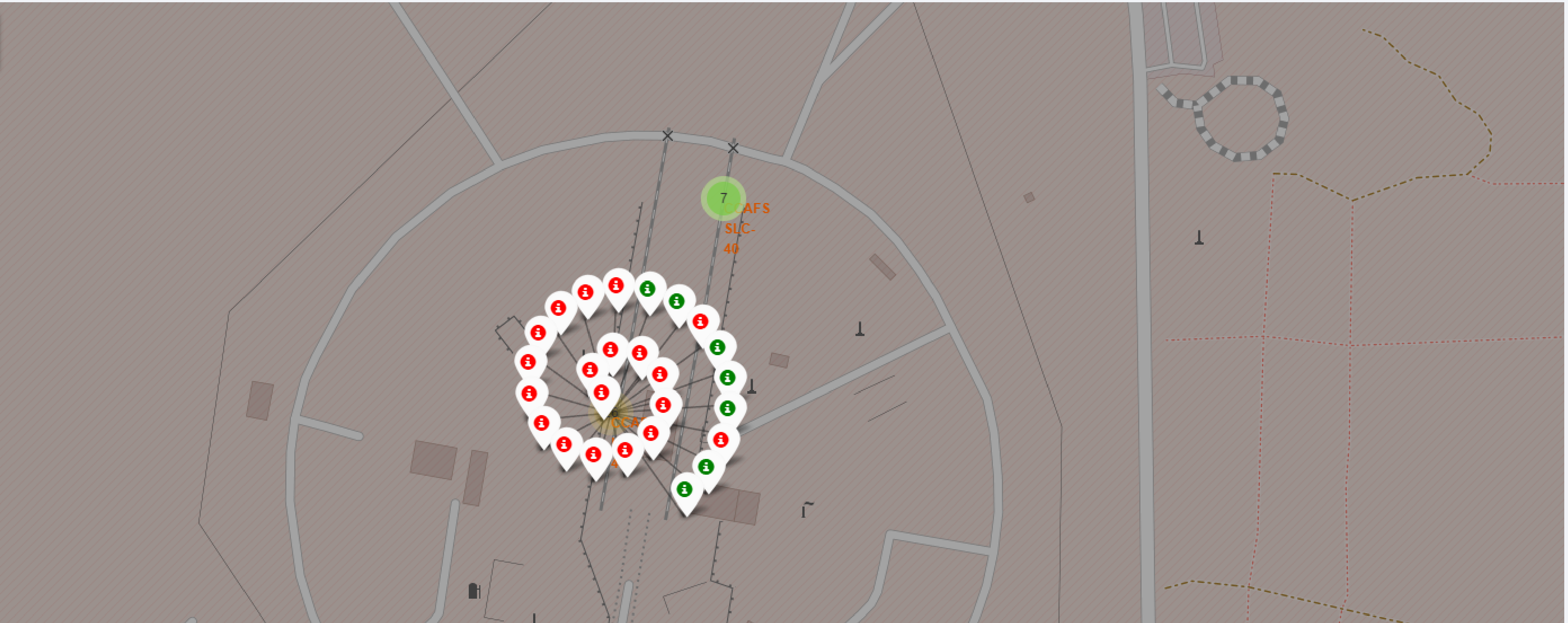
Global Location of Launch Sites

It is essential to note that the three sites are located at the closest point to the Equator on the East and West coasts of the United States, which are also very close to their respective coasts.



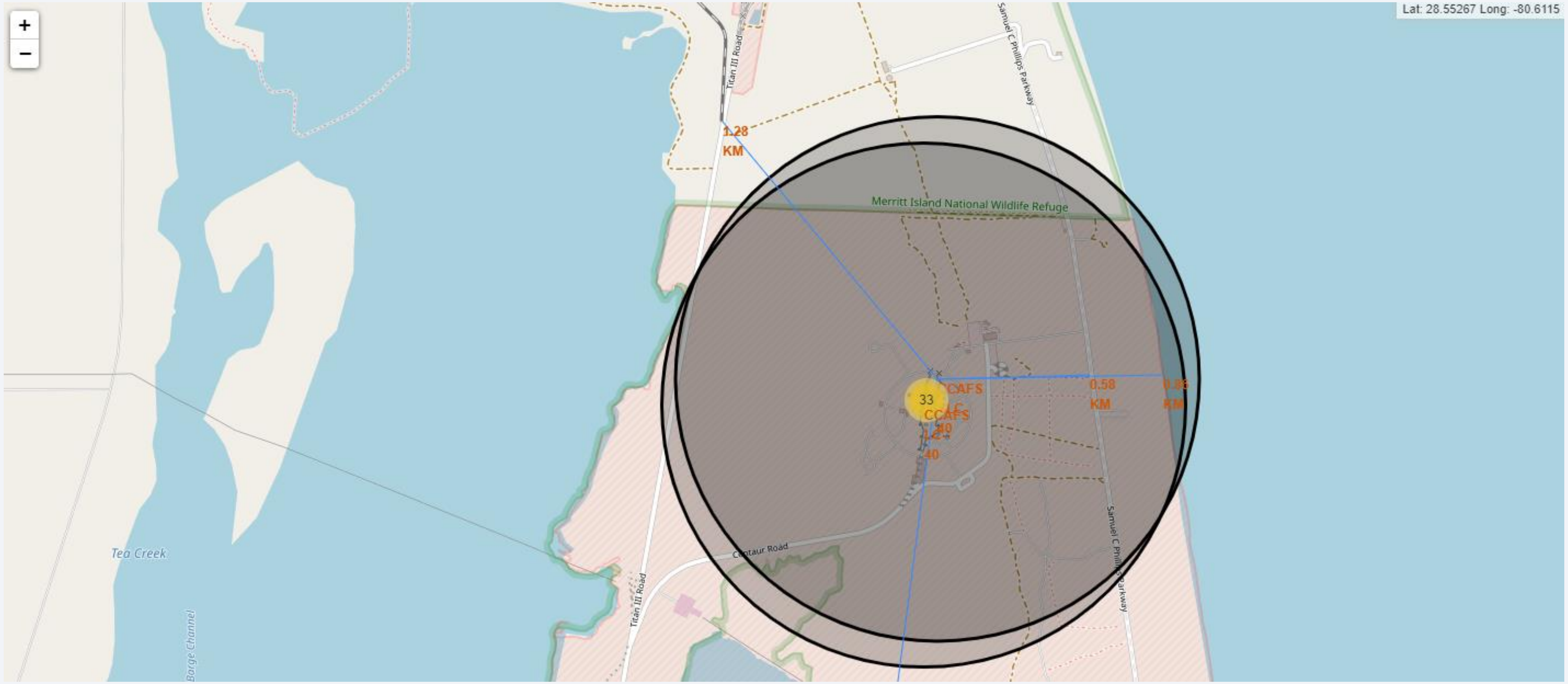
Color Label Markup Example

As an illustration of color labeling, the number of failures and low success rate of platform CCAFSLC-40 are readily apparent. It is possible to review the others in the Jupyter notebook.



CCAFSLC-40 Proximity

The following map illustrates the platform CCAFSLC-40 proximity to various modes of transportation: the nearest train station is 1.28 kilometers away, the nearest beach is 0.5 kilometers away, but the nearest airport is 55 kilometers away and therefore should not be depicted.

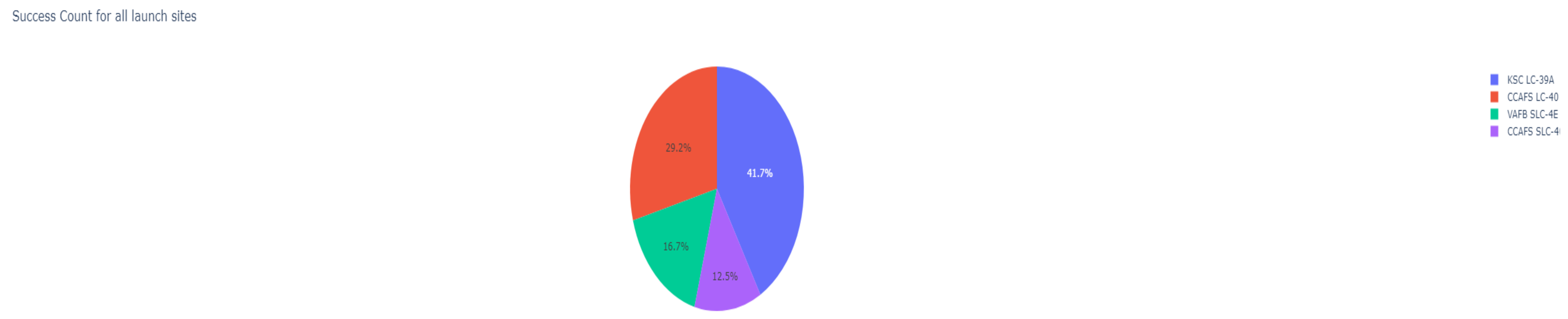


Section 4

Build a Dashboard with Plotly Dash

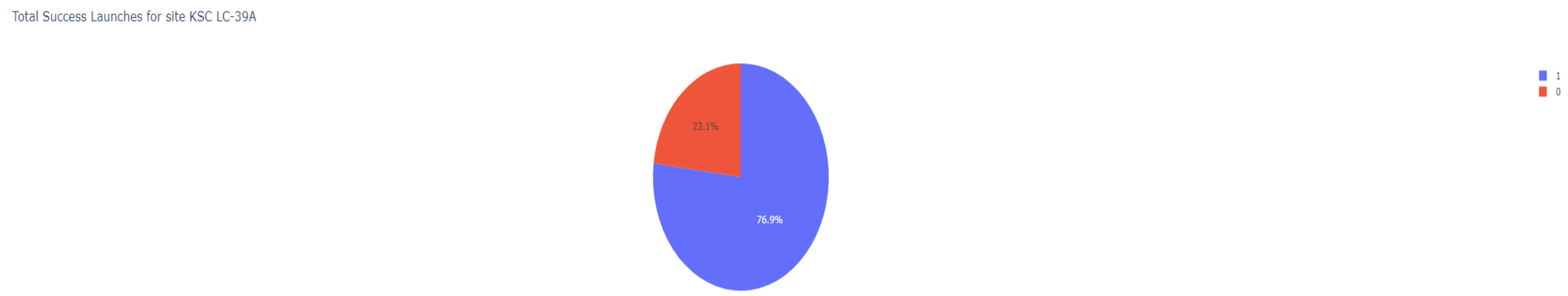
Success Counts for all launch sites

In the graph below it is possible to see the success rate of each of the launch platforms.



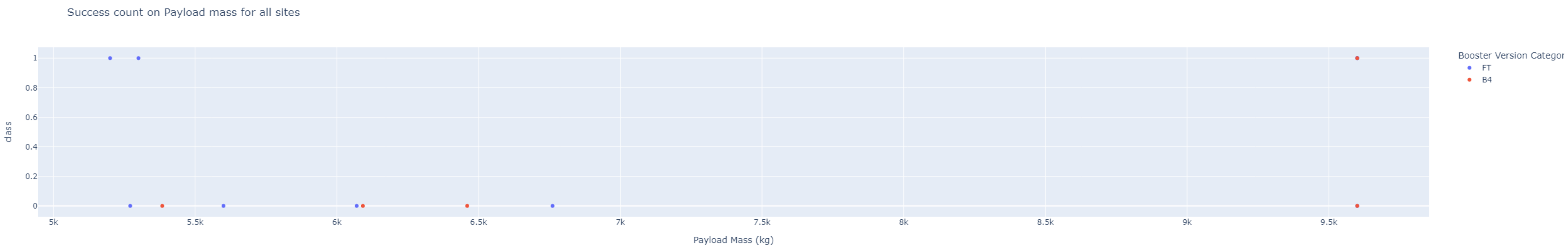
Total Success: KSCLC-89K site

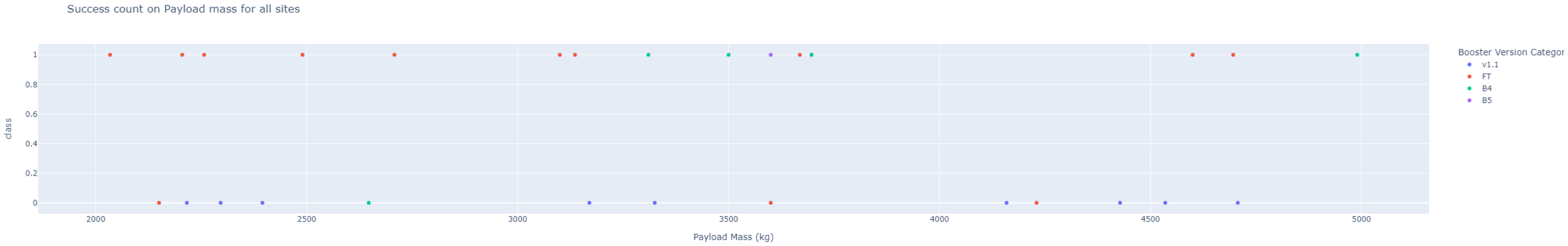
The graph below depicts the success rate of the launch site. KSCLC-89K



Success count on Payload mass compared for all sites

This indicates that the likelihood of failure increases as the load increases. In the lower screenshot, the success rate is considerably more variable than in the higher screenshot, where it is considerably lower.



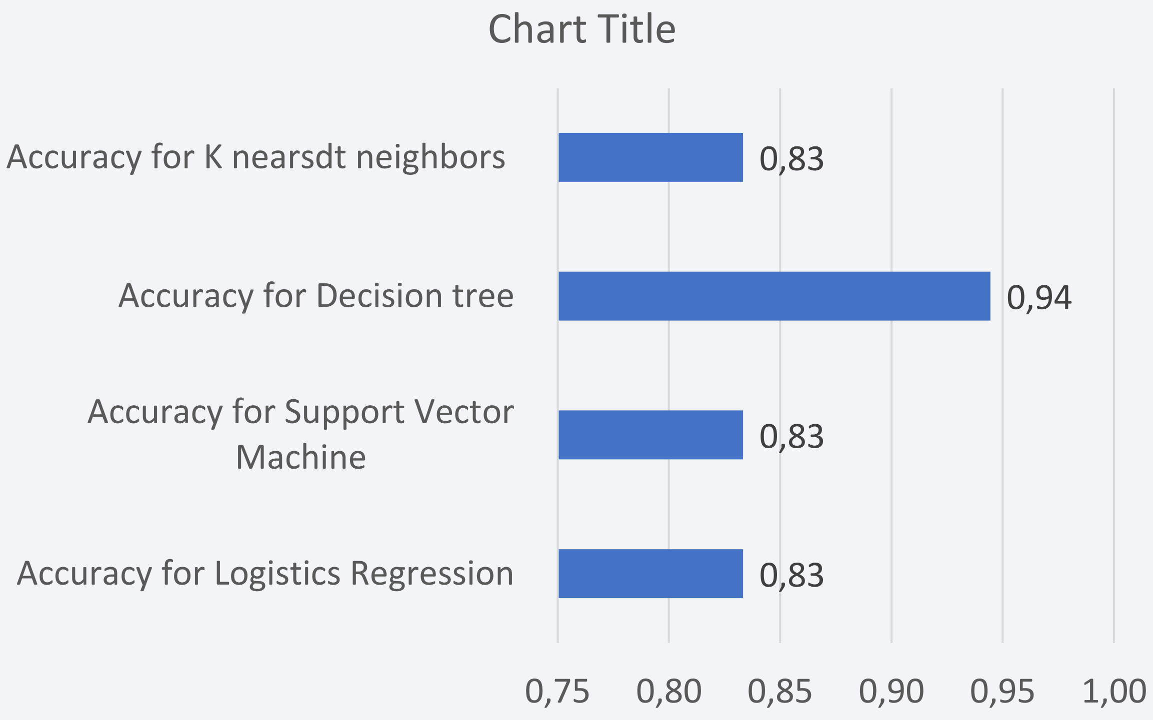


Section 5

Predictive Analysis (Classification)

Classification Accuracy

As can be seen, practically all predictive models have similar results; however, the decision tree is significantly larger, so it is assumed that it is the most suitable model for data analysis.



Confusion Matrix

In the presented model, there is only one false negative; consequently, it was able to predict all positive results and did not make any failures that would compromise the safety of future flights, establishing an exceptional performance within the scope of this project.

Conclusions

- There is a technological advancement that manifests itself in the decrease of failures in new releases.
- Undoubtedly, the decision tree is the most successful model in terms of prediction accuracy, but the performance of the other models is also quite good.
- There are complex relationships within the process between the loaded weight, the type of orbit, the technology used, and the technical progress of the launch processes that are sufficient to generate a reliable predictive model.

Thank you!