IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Fernández Acuña Jorge Manuel
10/25/2023

Outline

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

Executive Summary

- Summary of methodologies
  - To collect the data, the SpaceX web API was used and through web scraping with BeautifulSoup library.
  - The data wrangling was mainly made with the Pandas library and SQL.
  - To visualize the data, the Folim geolocation tool was used, and for the representation, Matplotlib and Dash were used.
  - To generate models, the tools from the Scikit-learn library were used.
- Summary of all results
  - The data was extracted correctly
  - Significant graphic representations were obtained
  - An adequate prediction model was obtained

Introduction

- Project background and context
  - SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch. This goal of the project is to create a machine learning pipeline to predict if the first stage will land successfully.
- Problems you want to find answers
  - What are the factors for the rocket to land successfully? Considering its location, payload and the type of orbit

Section 1

Methodology

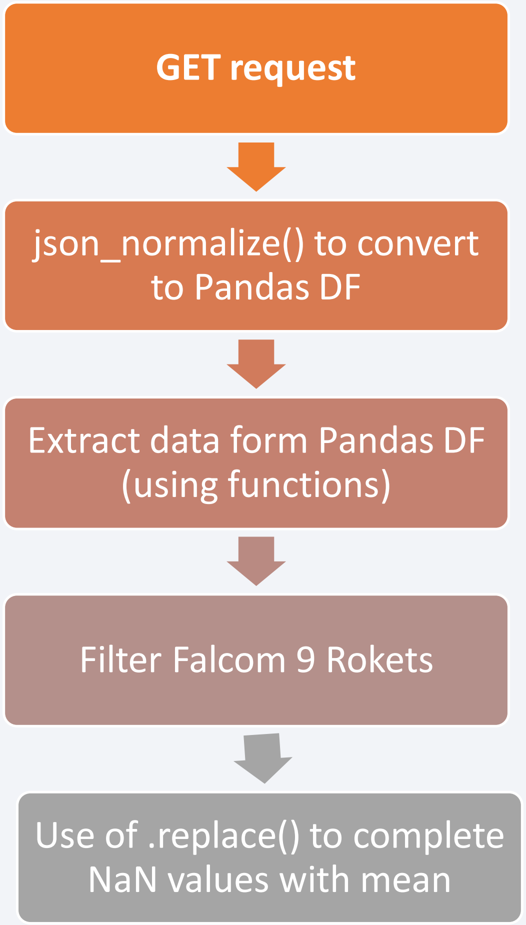
Methodology

Executive Summary

- Data collection methodology:
  - To collect the data via SpaceX Api, it was accessed through the following url: https://api.spacexdata.com/v4/launches/past.
  - The Web Scraping was from here: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches
- Perform data wrangling
  - The data was processed in a dataframe obtained with Pandas, the rows with zero values and unnecessary columns were eliminated
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

From the API the dataset was collected whit this steps:



GET request

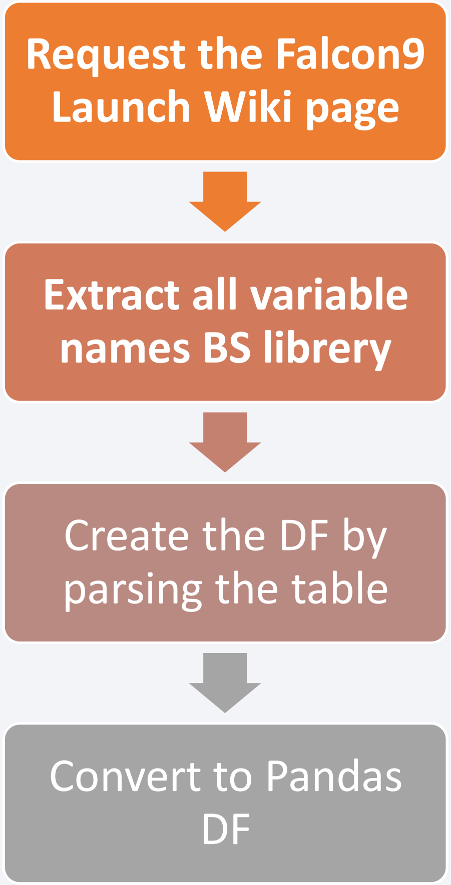
json_normalize() to convert to Pandas DF

Extract data form Pandas DF (using functions)

Filter Falcom 9 Rokets

Use of .replace() to complete NaN values with mean

From the WebScraping the dataset was collected whit this steps:



Request the Falcon9 Launch Wiki page

Extract all variable names BS librery

Create the DF by parsing the table

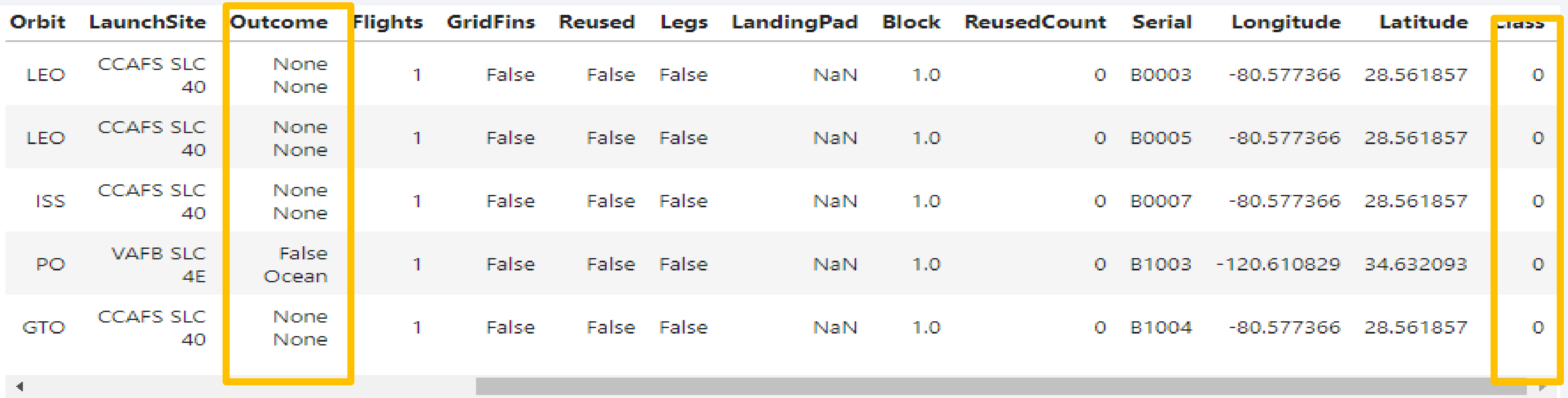
Convert to Pandas DF

URL: https://github.com/JorgeMFernandezAcuna/Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

URL: https://github.com/JorgeMFernandezAcuna/Applied-Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

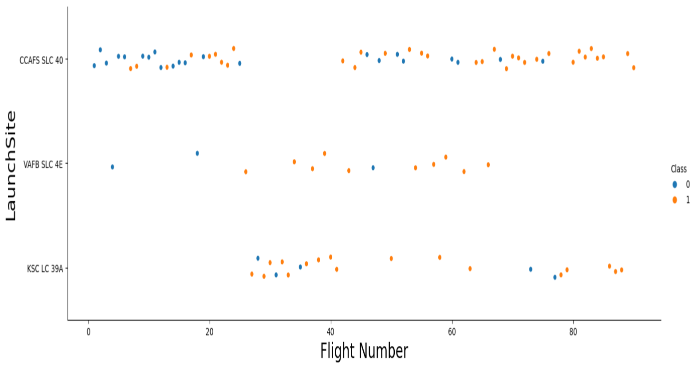
- The objective of this stage is to perform **Exploratory Data Analysis** on the **Dataset to Determine Training Labels**
- For that **the number of launches on each site** and **the number occurrence of each orbit** was determined.
- Then a column is added to the dataset that determines the **landing outcome.**

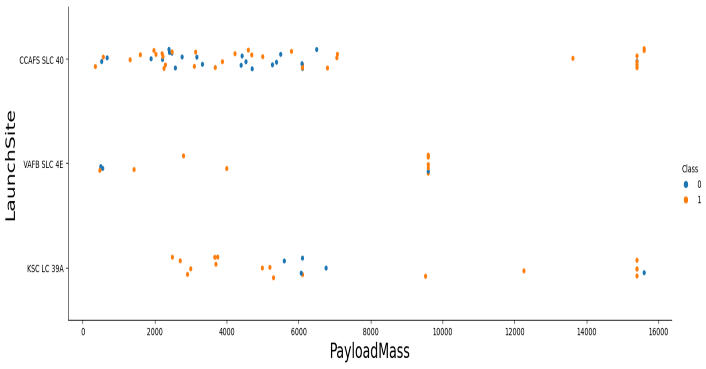


Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude	Class
LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.561857	0
LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.561857	0
ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.561857	0
PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610829	34.632093	0
GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28.561857	0

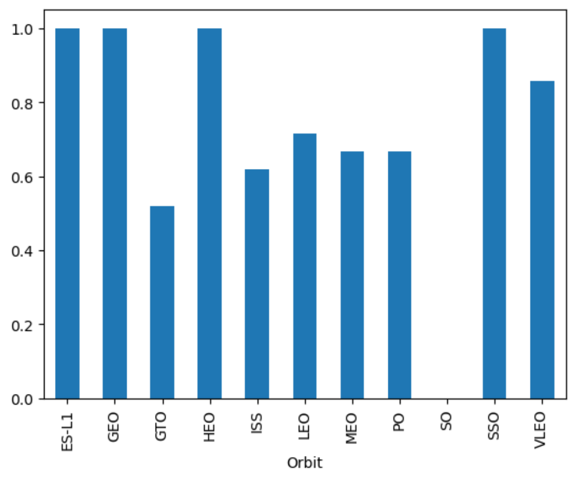
https://github.com/JorgeMFernandezAcuna/Applied-Data-Science-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

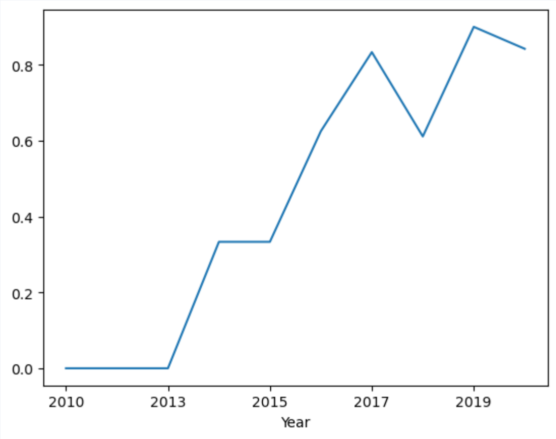




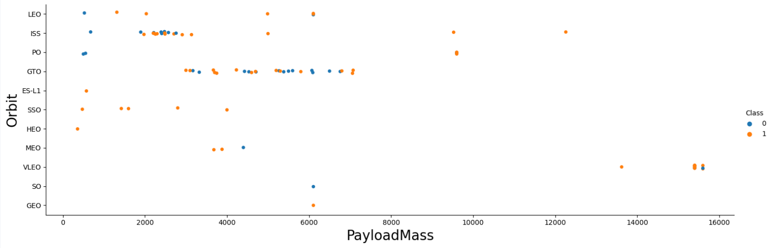
Scatterplots were used to see the relationship between **Flight Number and Launch Site** and **Payload and Launch Site. It** seems the more massive the payload, the less likely the first stage will return.



The relationship was compared between **success rate and orbit type. And payload and orbit**. It is observed in which orbits are not accessed with a certain payload.



Finally, the **year** vs. **the launch success** was graphed.
To demonstrate improvement over the years



https://github.com/JorgeMFernandezAcuna/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-dataviz%20(1).ipynb

EDA with SQL

The following SQL queries were performed:

- 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 average payload mass carried by booster version F9 v1.1
- List the date when the first succesful 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 mas
- 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 landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.

https://github.com/JorgeMFernandezAcuna/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- With markers, mark all launch sites on a map and mark the success/failed launches for each site on the map.
- The success/failed launches for each site was mark with color-labeled clusters
- With lines we show sites near railways, highways and coastlines.
- https://github.com/JorgeMFernandezAcuna/Applied-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location%20(1).ipynb

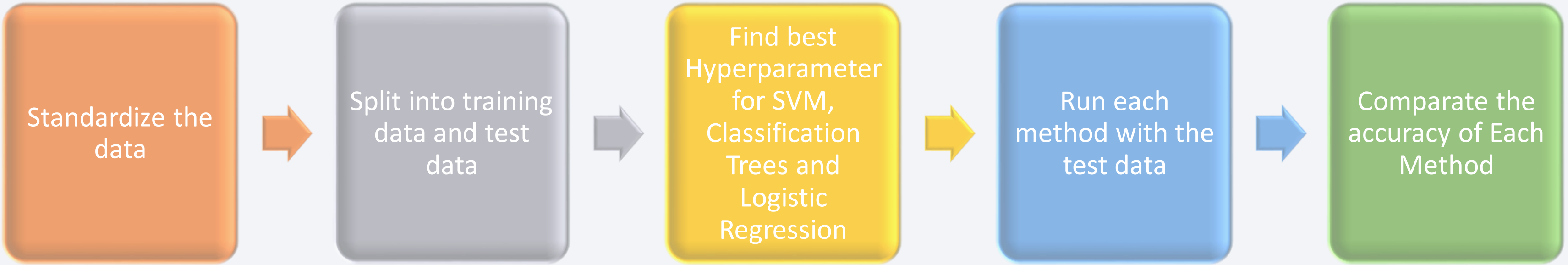
Build a Dashboard with Plotly Dash

- Two interactive graphs were generated with Dash.
- The first is a pie chart that allows you to see the percentage of successful launches according to the site.
- The second is points and allows you to select the payload to see the booster used.
- With the agreement of both you can see which is the best payload for each launch site.

https://github.com/JorgeMFernandezAcuna/Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- The goal was to find the method performs best using test data.
- For that we follow this sequence:


Standardize the data
Split into training data and test data
Find best Hyperparameter for SVM, Classification Trees and Logistic Regression
Run each method with the test data
Comparate the accuracy of Each Method

http://localhost:8888/lab/tree/SpaceX_Machine_Learning_Prediction.ipynb

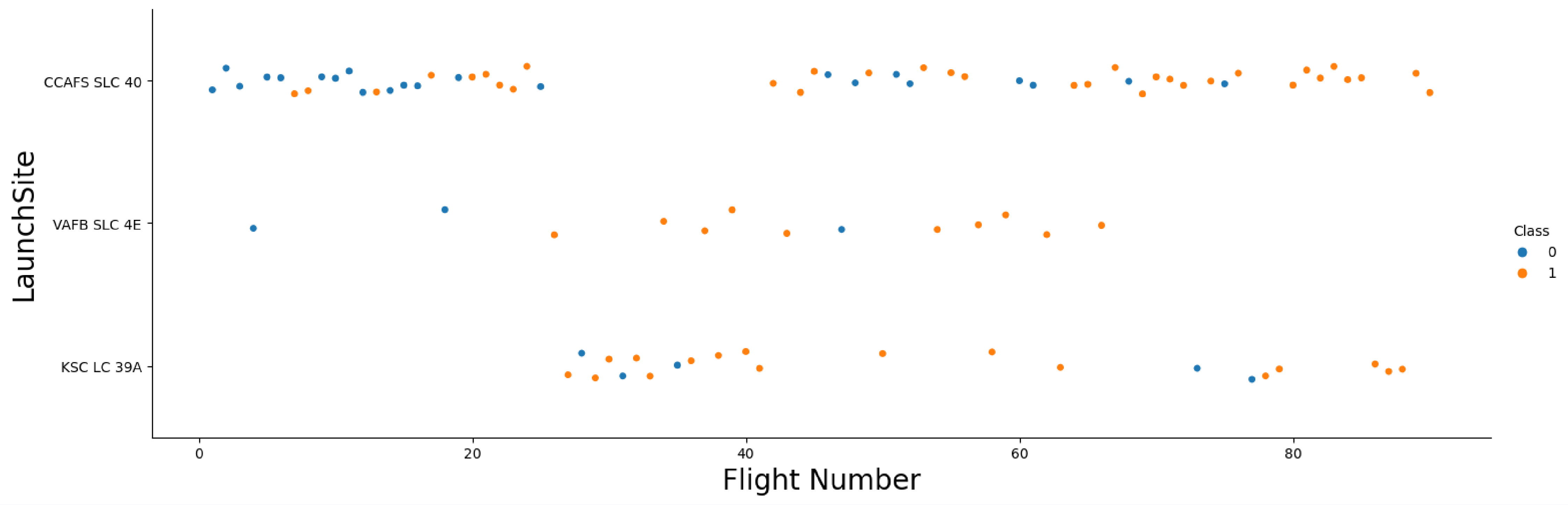
Results

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

Section 2

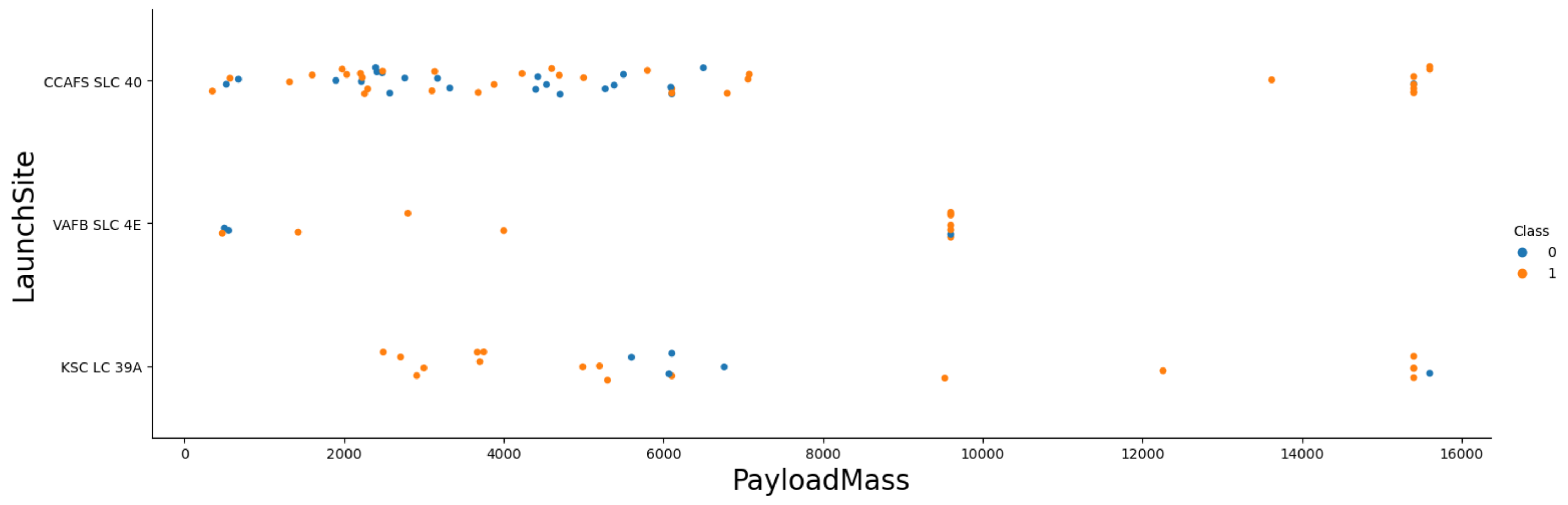
Insights drawn from EDA

Flight Number vs. Launch Site



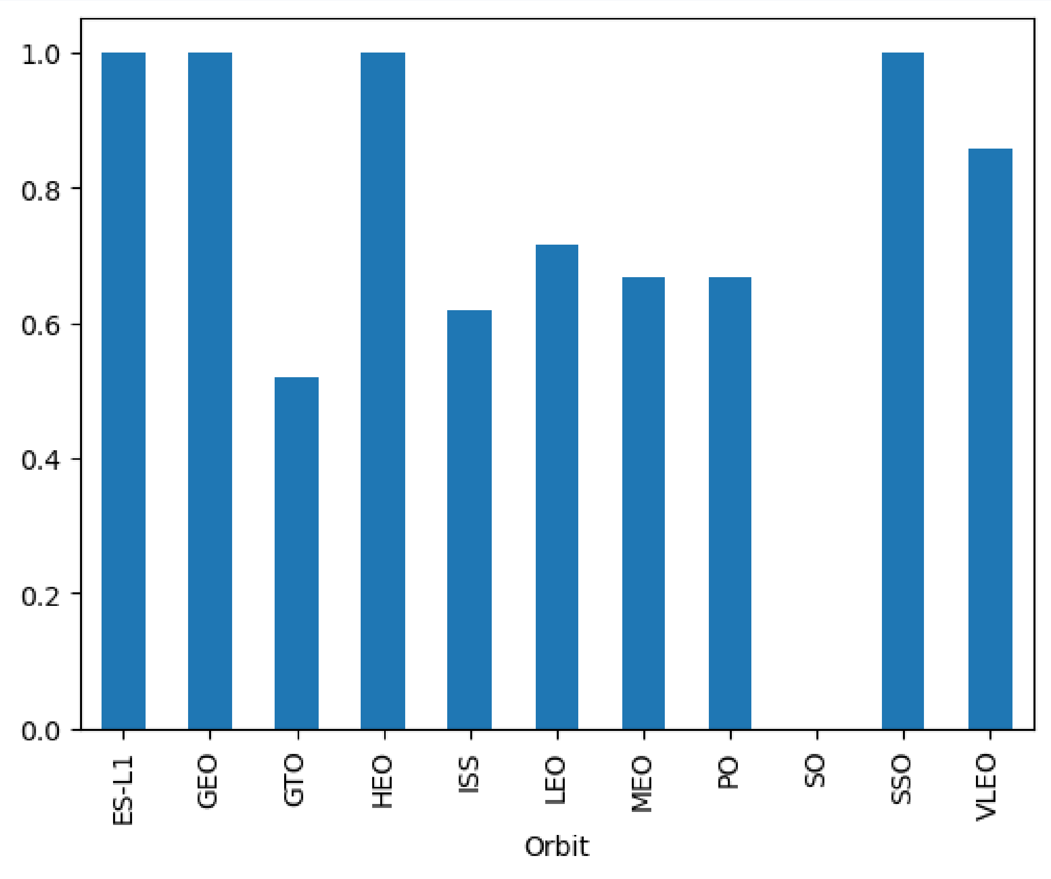
- The best launch site is CCAFS SLC 40 at the same time it is the one with the highest number of launches currently.
- Also KSC LC 39A has high success ratios. In conclusion, the sites with launches closest in time are the most successful.

Payload vs. Launch Site



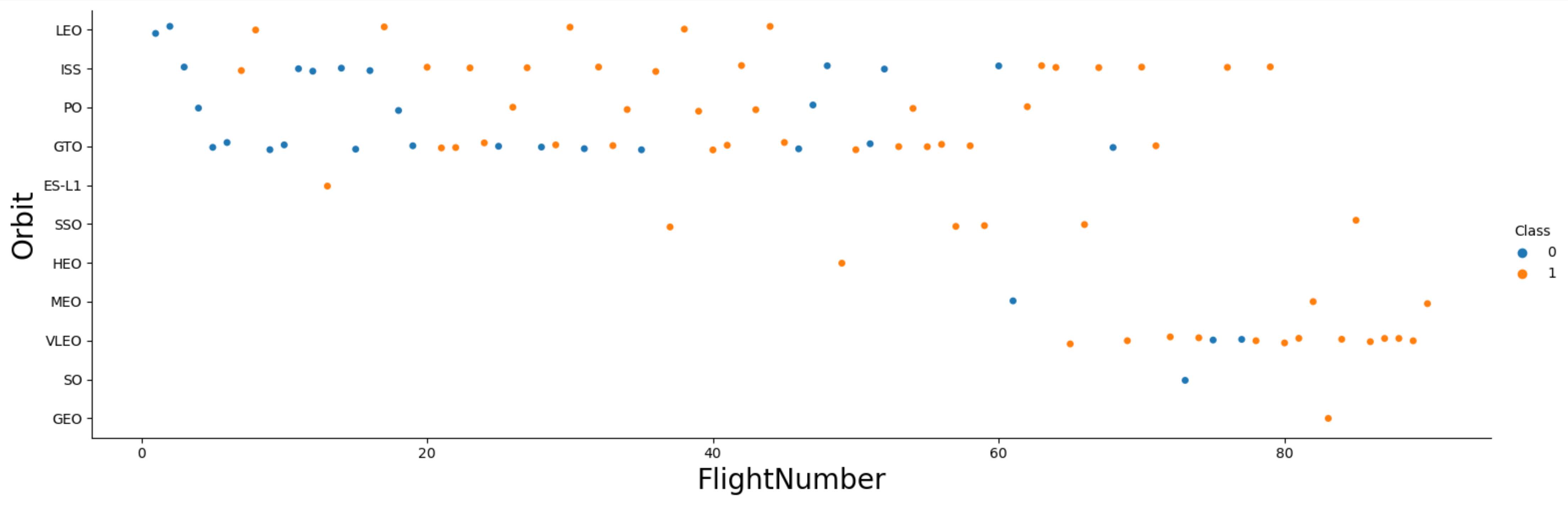
- The trend is that the greater the payload, the greater the success at launch.
- The payload mass cut is approximately 9000 kg.
- These occur on all the launch sites but in CCAF SLC 40 and KSC LC 39 A the trend is observed to use larger payload masses

Success Rate vs. Orbit Type



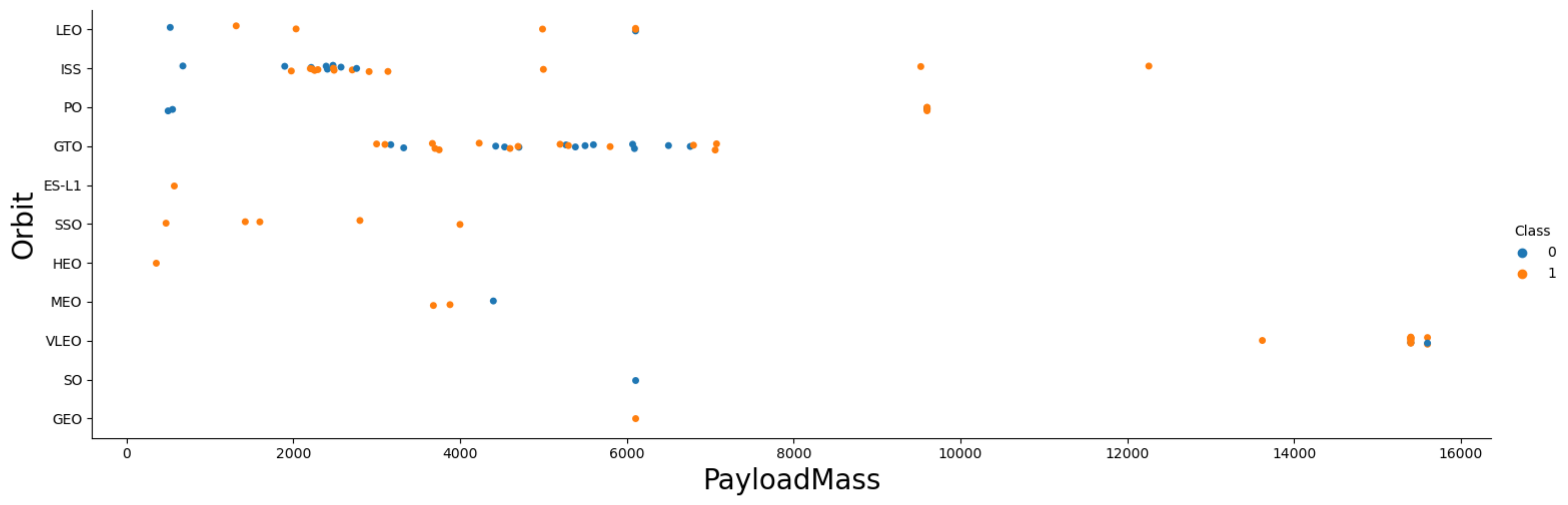
- ES-L1, GEO, HEO, SSO, VLEO had the most success rate.

Flight Number vs. Orbit Type



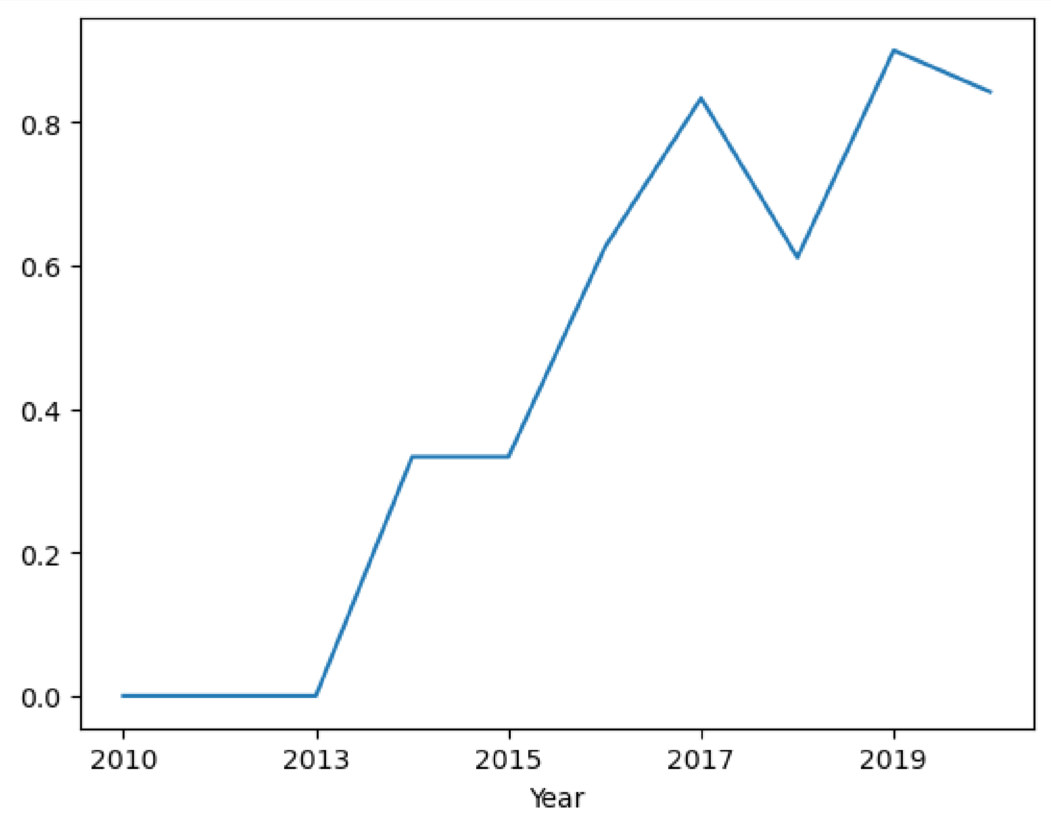
- Launch success depending on the chosen orbit increases with time, except apparently for GTO.
- The VLEO orbit is new and has a great success rate.

Payload vs. Orbit Type



- The heavy load favorably affects the PO, LEO, ISS and VLEO orbits.
- Not so for GTO.

Launch Success Yearly Trend



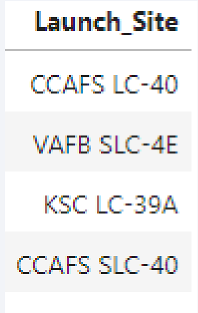
- Clearly the early years were all about development and improvement.
- The success trend increases until it reaches almost 100% by the end of 2020.

All Launch Site Names

The Distinct statement allows you to select unique values

```
%sql SELECT Distinct LAUNCH_SITE FROM SPACEXTABLE
```

%sql wildcard allows you to execute SQL commands in Python

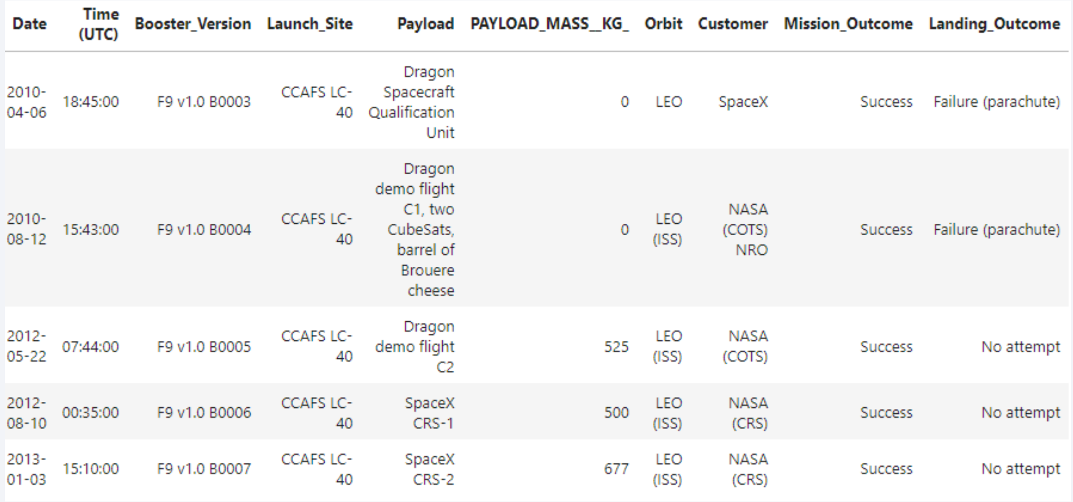


Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEXTABLE WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
```

The LIKE statement allows you to search for everything that begins with what is indicated by marking the cut with the wildcard %

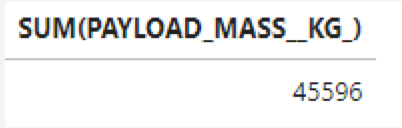


Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight 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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	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 SUM statement performs the arithmetic sum of the entire selected column

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM
SPACEXTABLE WHERE CUSTOMER='NASA (CRS)'
```

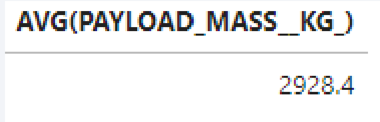


SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

The AVG statement performs the arithmetic average of the entire selected column

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE
WHERE BOOSTER_VERSION='F9 v1.1'
```

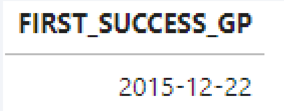


AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

The MIN statement
filters out the data
with the lowest value,
in this case, the
oldest date

```
%sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP
FROM SPACEXTABLE WHERE LANDING_OUTCOME =
'Success (ground pad)'
```

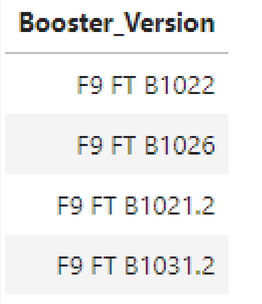


FIRST_SUCCESS_GP
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

The BETWEEN statement selects data from a range of values

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ between
4000 and 6000 AND LANDING_OUTCOME='Success (drone ship)'
```

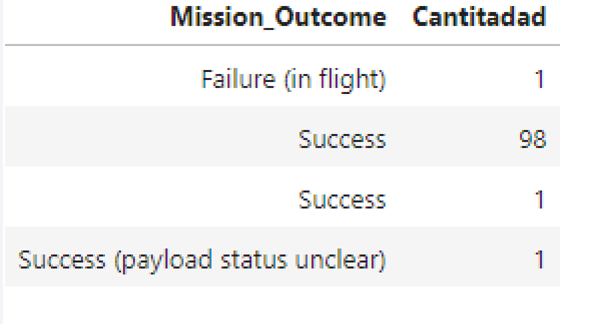


Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

The COUNT statement allows you to count the number of heats of a selected criterion

```
%sql SELECT MISSION_OUTCOME, COUNT(*) AS Cantitadad FROM SPACEXTABLE GROUP BY MISSION_OUTCOME
```

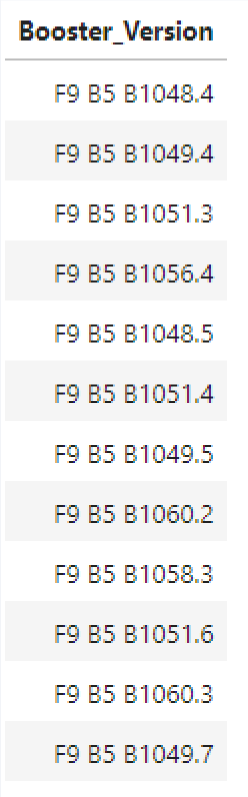


Mission_Outcome	Cantitadad
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)

In this case, a subquery is used to select the cases where the payload mass is maximum

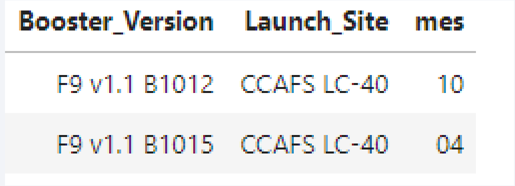


Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

```
%sql SELECT BOOSTER_VERSION, LAUNCH_SITE, substr(Date, 6,2) as mes
FROM SPACEXTABLE WHERE LANDING_OUTCOME = 'Failure (drone ship)'
AND substr(Date,0,5) = '2015'
```

We use the AND statement to restrict to two conditions

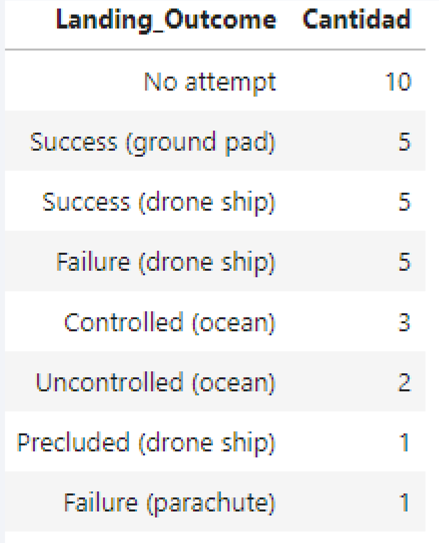


Booster_Version	Launch_Site	mes
F9 v1.1 B1012	CCAFS LC-40	10
F9 v1.1 B1015	CCAFS LC-40	04

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

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

The ORDER BY statement allows us to display the results in ascending or descending order.

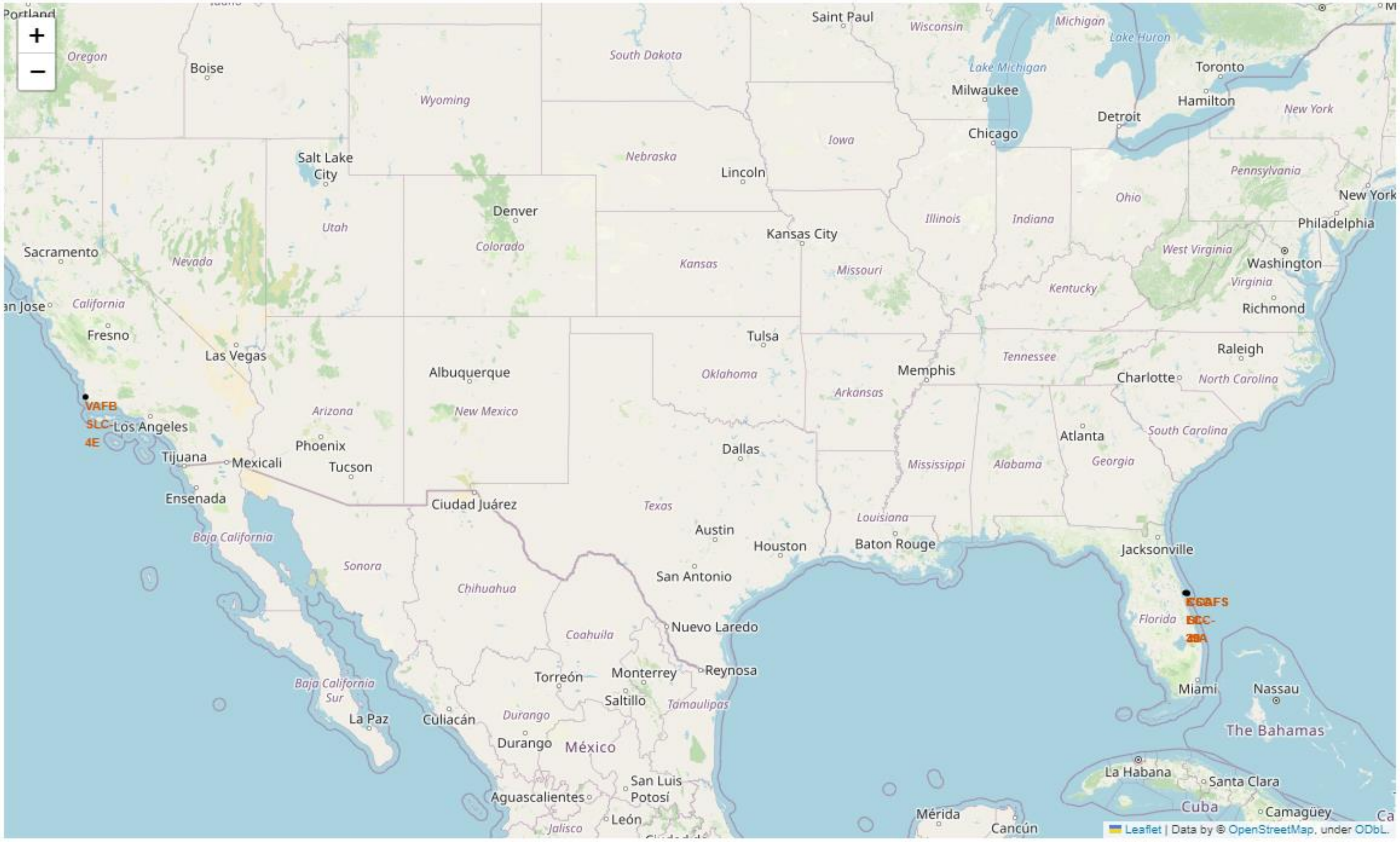


Landing_Outcome	Cantidad
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

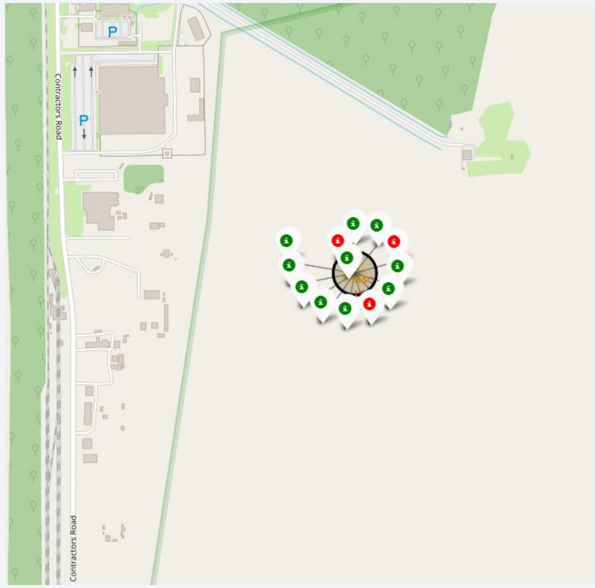
Launch Sites Proximities Analysis

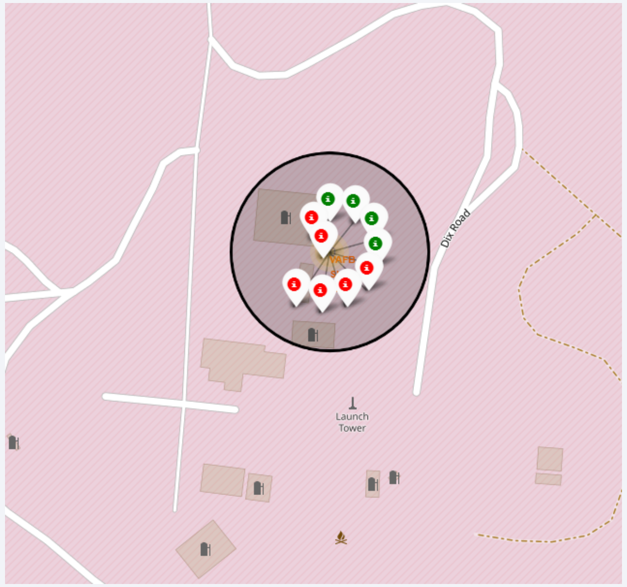
Global map of Lunch Sites

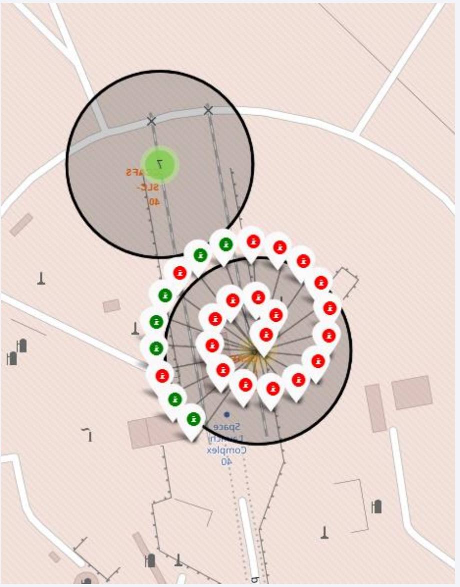


The chosen launch sites are on the Atlantic and Pacific coasts of the United States. Probably for security reasons.

Launch Outcome

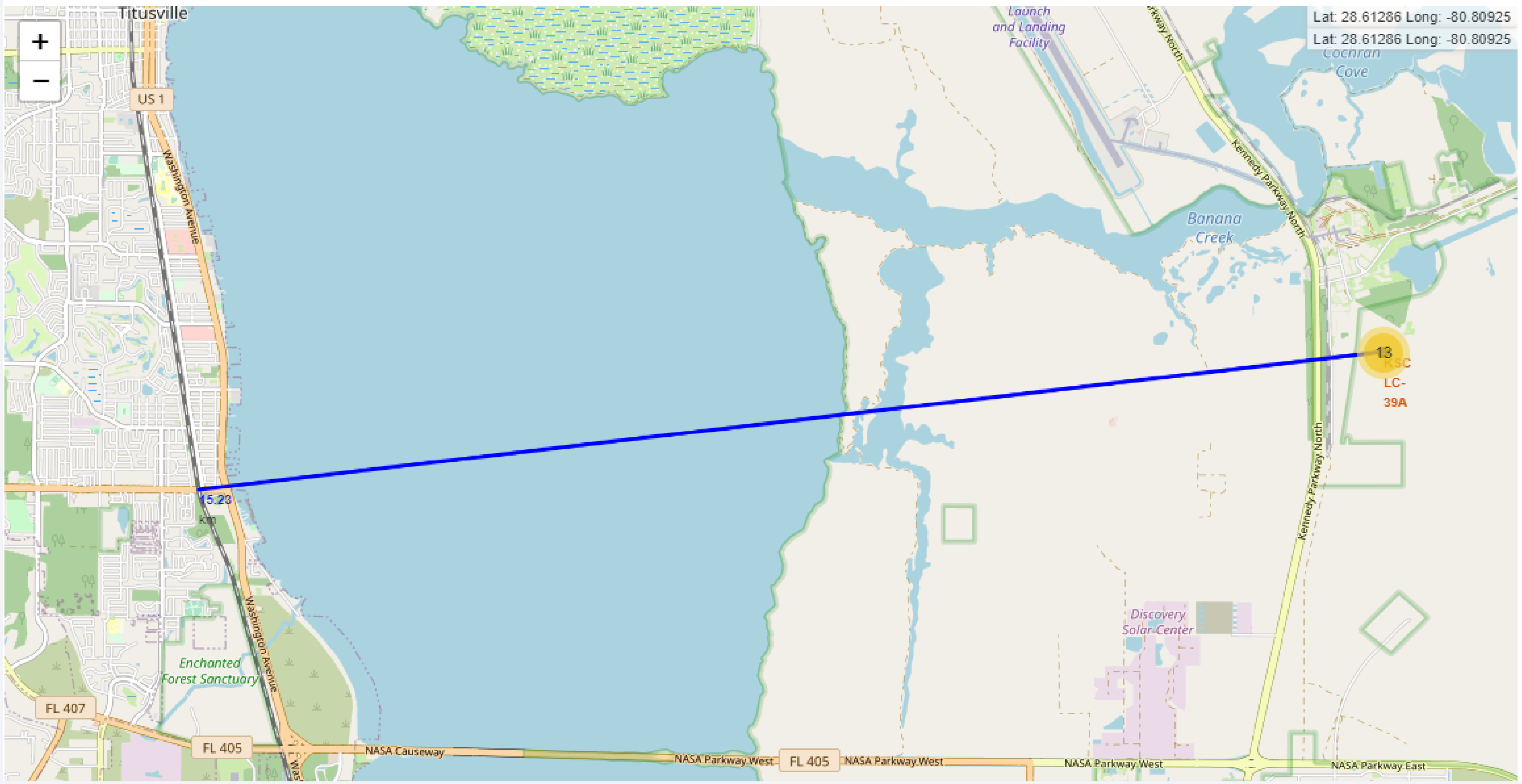






RED are the fail launches
GREEN are the successful

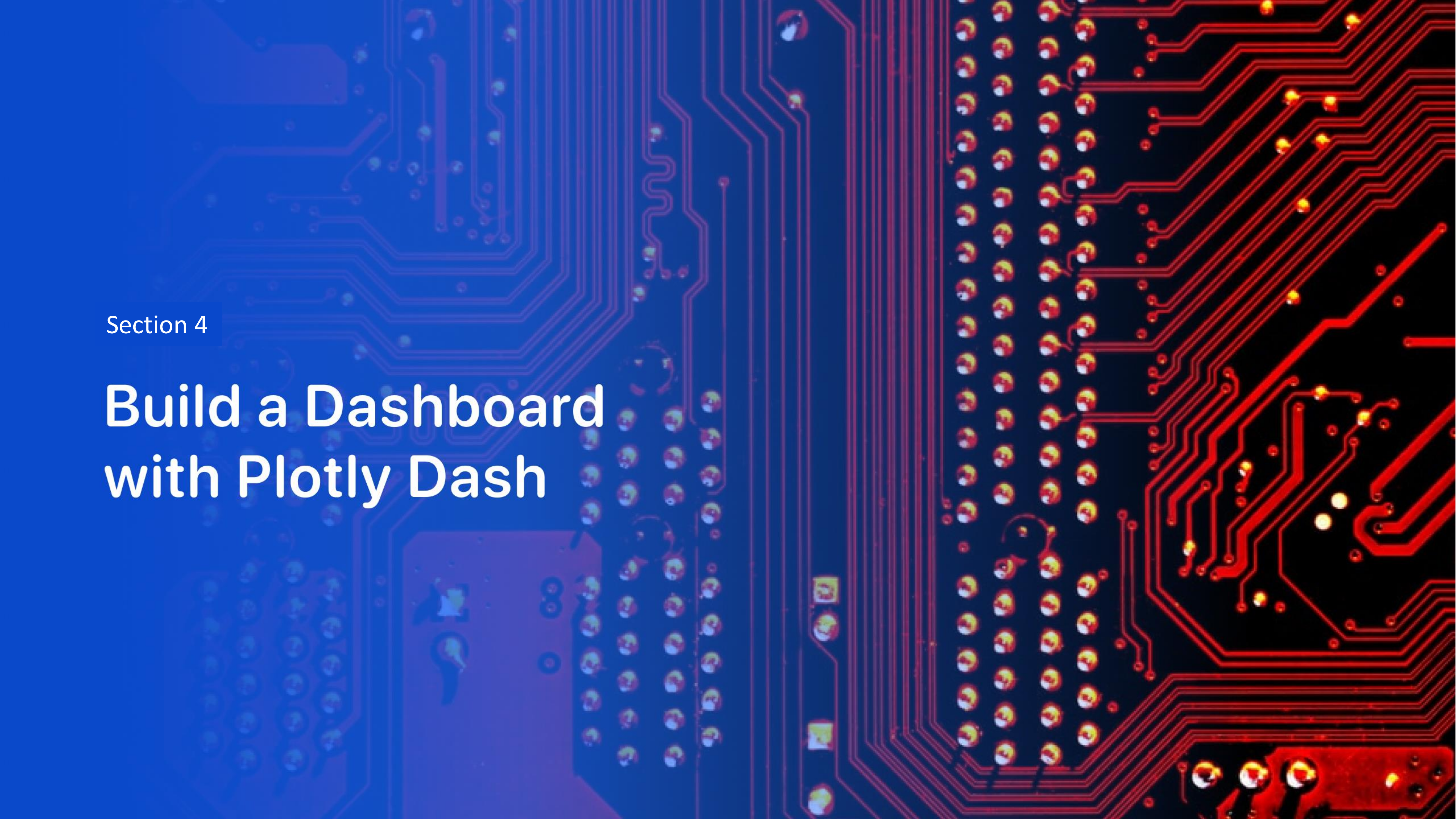
Security Decisions



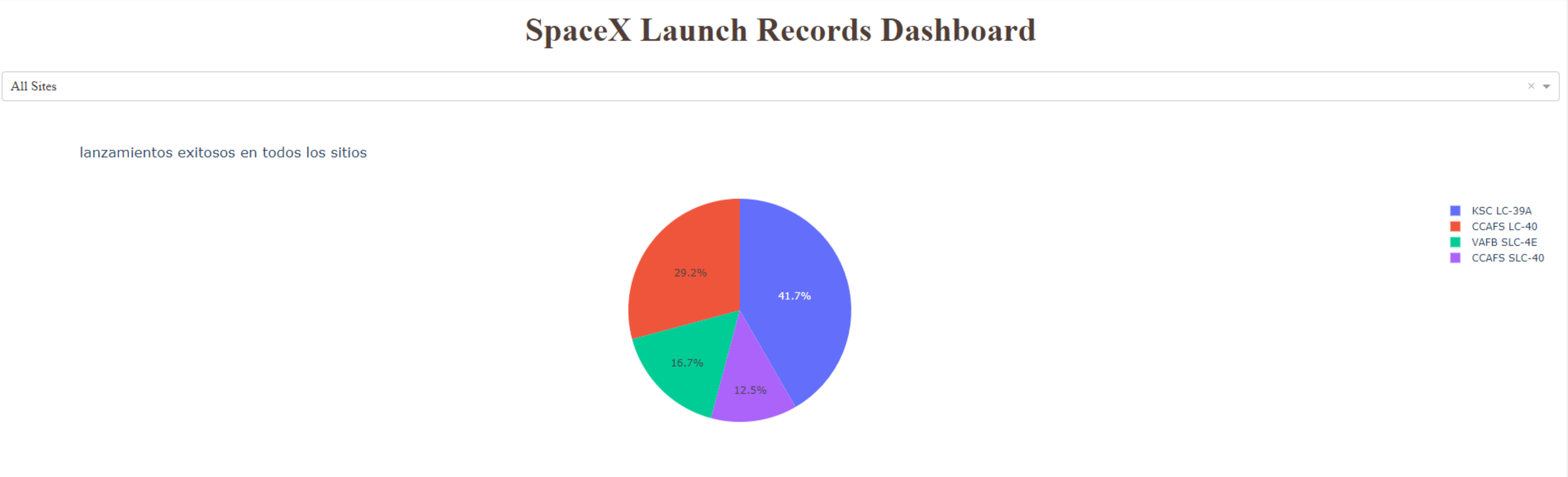
The distance from the launch site RSC LC-39 A to the railway is 15.32 km

Section 4

Build a Dashboard with Plotly Dash



The success percentage in each launch site



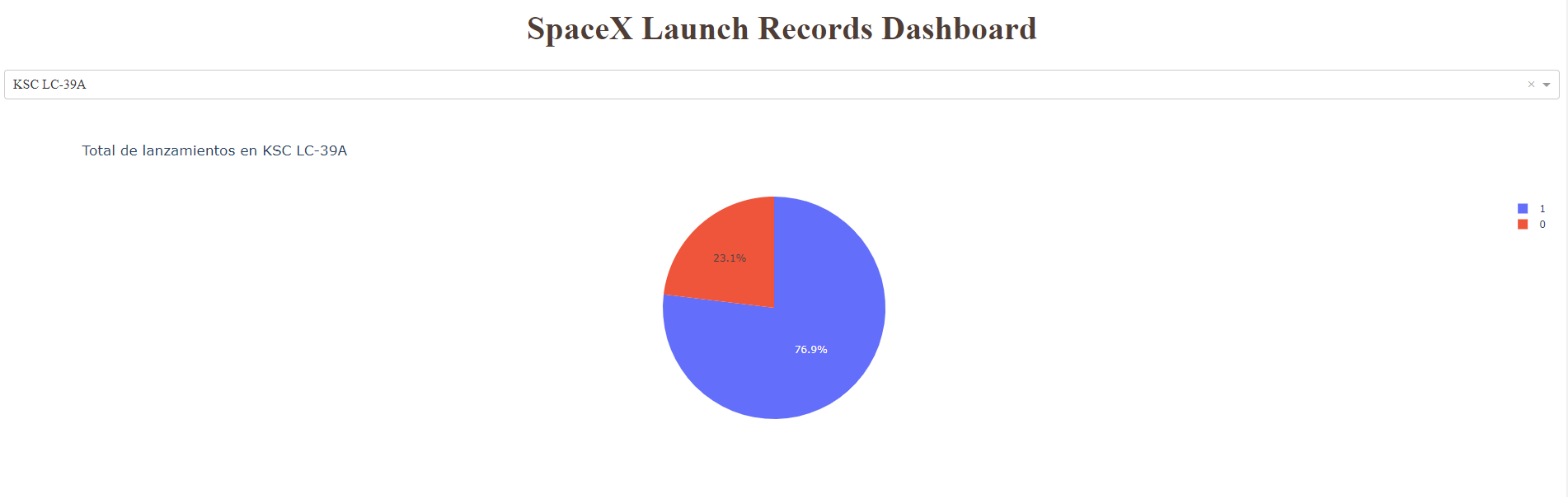
SpaceX Launch Records Dashboard

All Sites

lanzamientos exitosos en todos los sitios

- The KSC LC-39a launch site is the most successful

Ratio of the best launch site



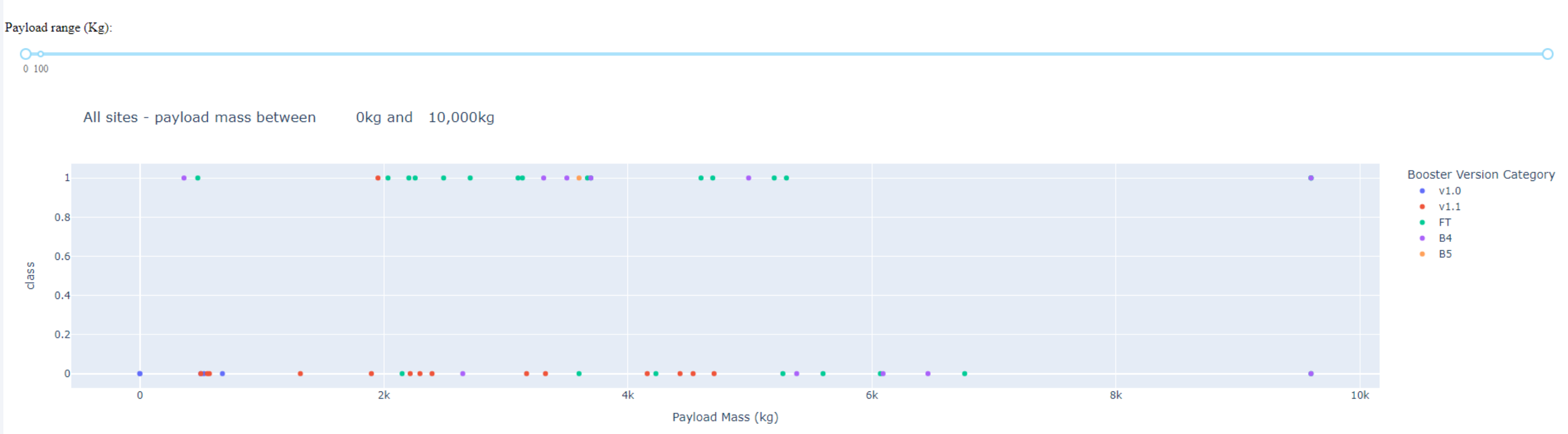
SpaceX Launch Records Dashboard

KSC LC-39A

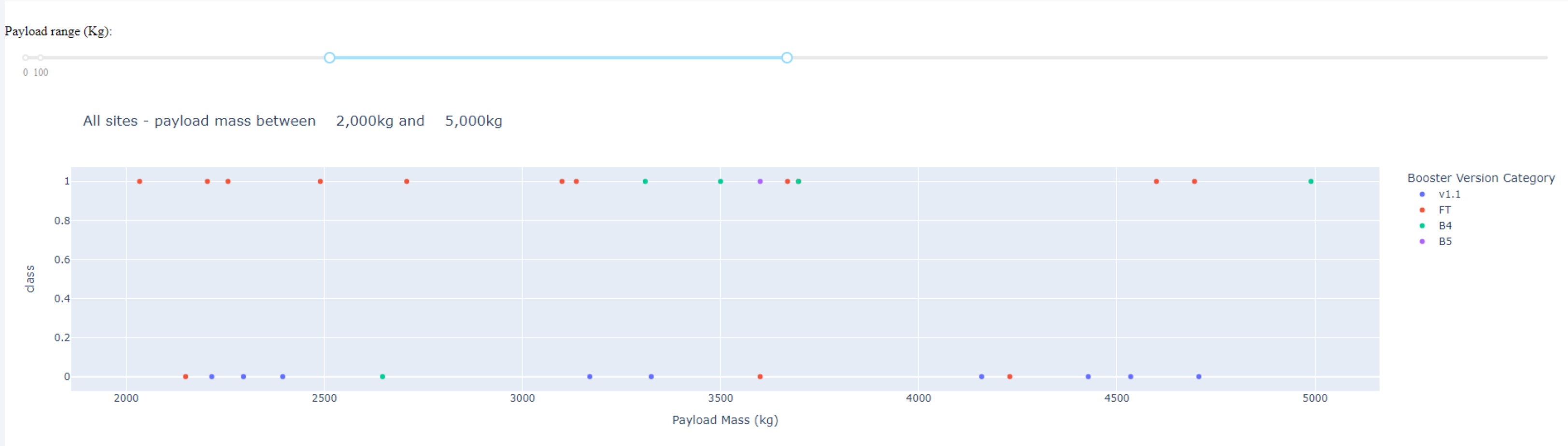
Total de lanzamientos en KSC LC-39A

KSC LC-39ª has a success rate of almost 80% which makes it very efficient.

Payload mass vs. launch outcome


Payload range (Kg):

All sites - payload mass between 0kg and 10,000kg


Payload range (Kg):

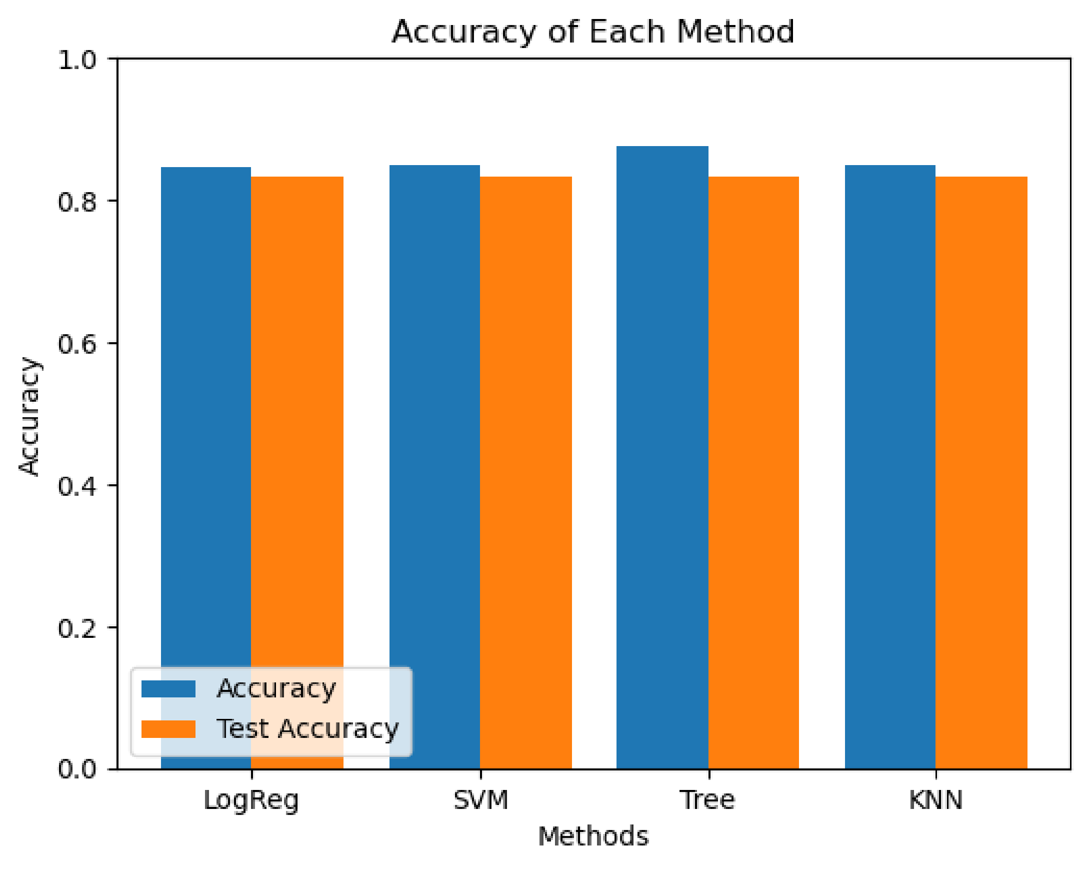
All sites - payload mass between 2,000kg and 5,000kg

- Launchs with a load between 2000 kg and 3300kg are the most successful.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

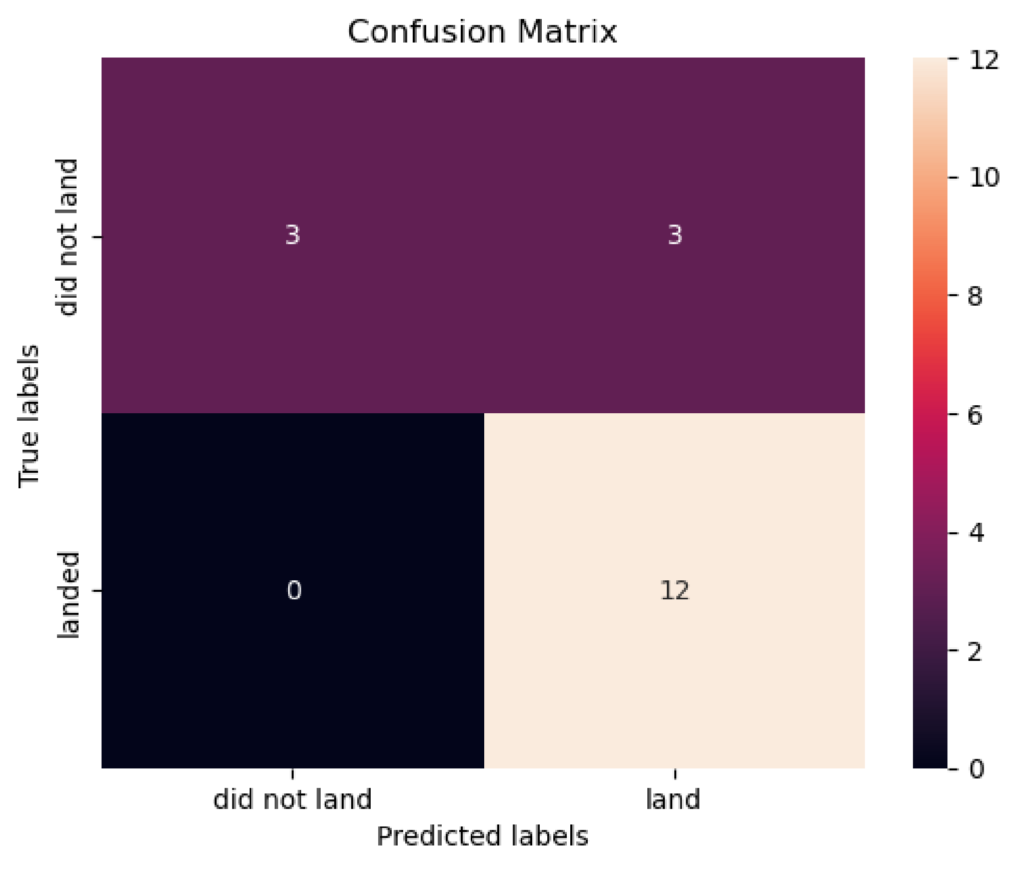


In the comparison of the models it is deduced that the best is the decision tree.

The ratios are:

Model	Accuracy
LogReg	0.84643
SVM	0.84821
Tree	0.875
KNN	0.84821

Confusion Matrix



The model can very effectively predict which launches landed. In the case of those that did not land it is not as efficient.

Conclusions

- The best prediction model is the decision tree, with an Accuracy of 83%.
- KSC LC 39A has high success launch ratios. Launches with payload mass above 7000kg are more successful.
- As time goes by, launch success increases significantly.
- The most successful orbits are: ES-L1, GEO, HEO, SSO, VLEO.

Thank you!