IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jorge Ivan Fuentes
Rosado
Agosto 2022

Outline

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

Executive Summary

- Summary of methodologies:
  - The project started with collecting and cleaning public data from SpaceX and Wikipedia by API and Web scrapping respectively. After replacing the missing data with the average of each feature and exploratory analysis was performed. A Dashboard is provided to analyze the data in real time. Four Models were training to find if the mission will succeed or not. In this project Success means if the mission landed. To select the best model Cross Validation was performed in all the models.
- Summary of all results:
  - The best model was the KNN with 94 % of accuracy to determine if the mission will succeed or not. The Confusion Matrix shows that the model did not have False Positive. It has only one False Negative.

Introduction

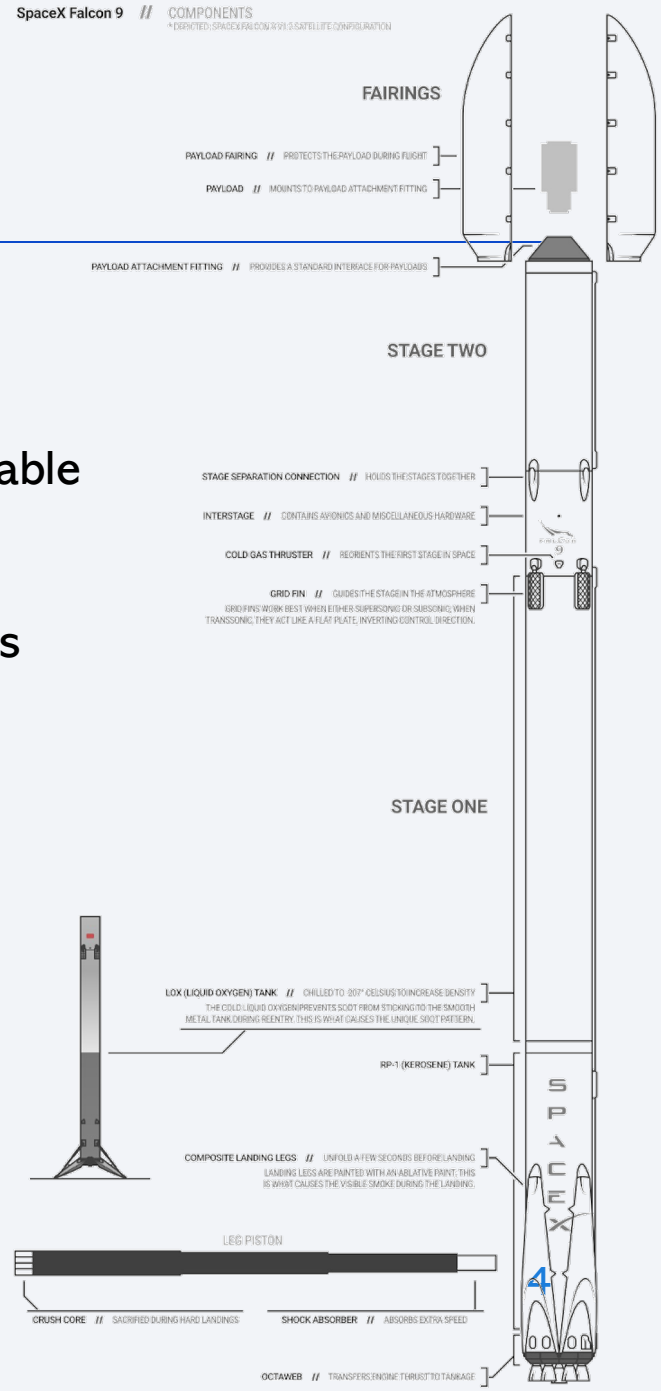
Background and Context

The commercial space age is here, companies are making space travel affordable for everyone.

The shuttle is composed of two stages. Stage two, or the second stage, helps bring the payload to orbit, but most of the work is done by the first stage.

Problems to Solve

- Determining the price of each launch
- Determining if SpaceX will reuse the first stage.



Section 1

Methodology

Methodology

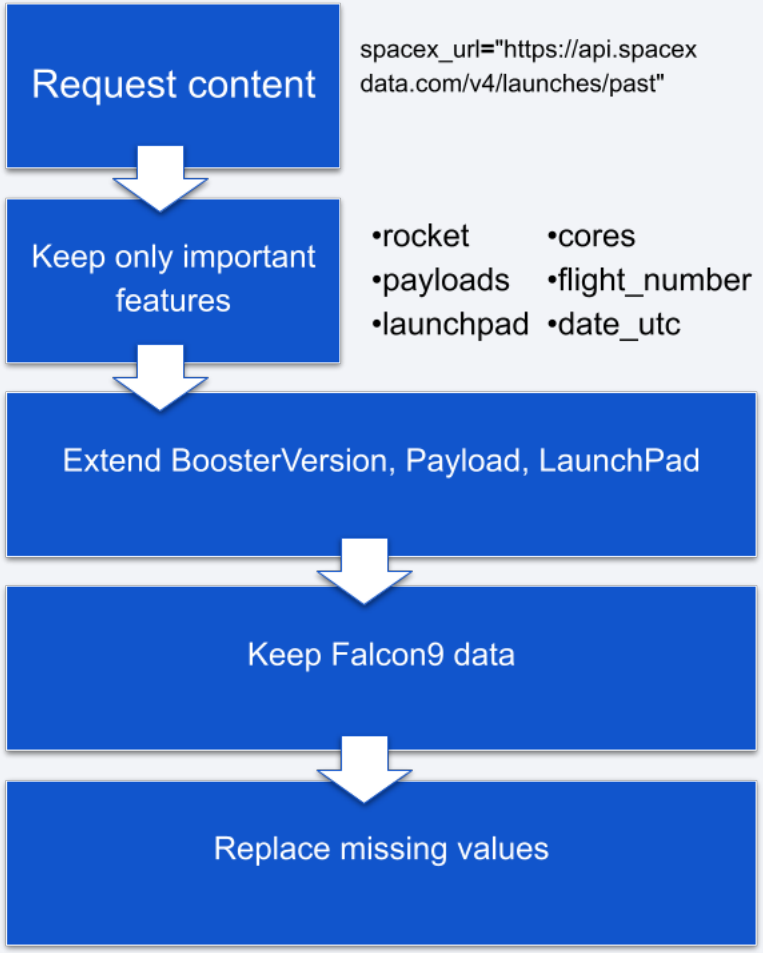
Executive Summary

- Data collection methodology:
  - Using SpaceX API to get the data. Falcon 9 data was used in this project
- Perform data wrangling
  - After collecting the data, some features were extended to express more information about the mission. The missing values were replaced by the average
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Four models were used. All of them were trained with Cross Validation and GridSearch to tune para hyperparameters.

Data Collection – SpaceX API

1. Request Data
2. Get Important Features for prediction model
3. Extend BoosterVersion, Payload, LaunchPad
4. Keep Falcon9 data
5. Replace missing values

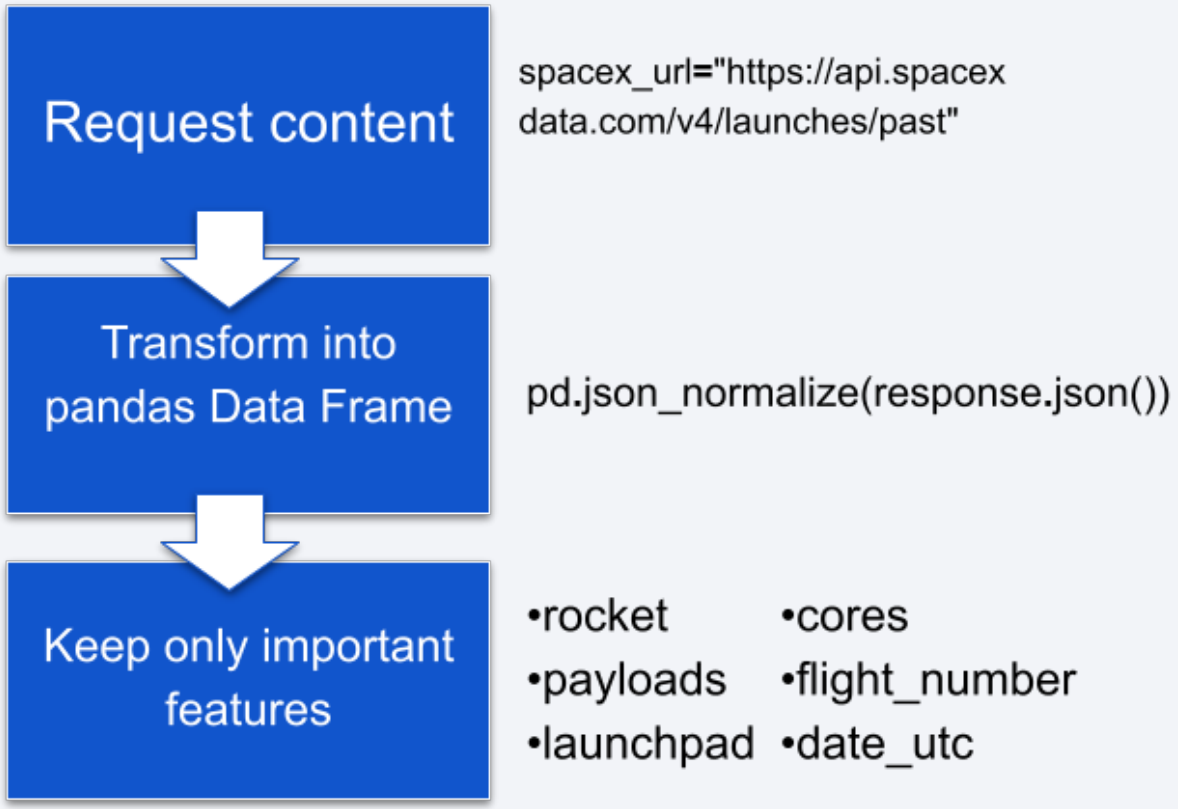
[Github Code]



Data Collection - Scraping

1. Request Data
2. Save it on a pandas Data Frame
3. Get Important Features for prediction model

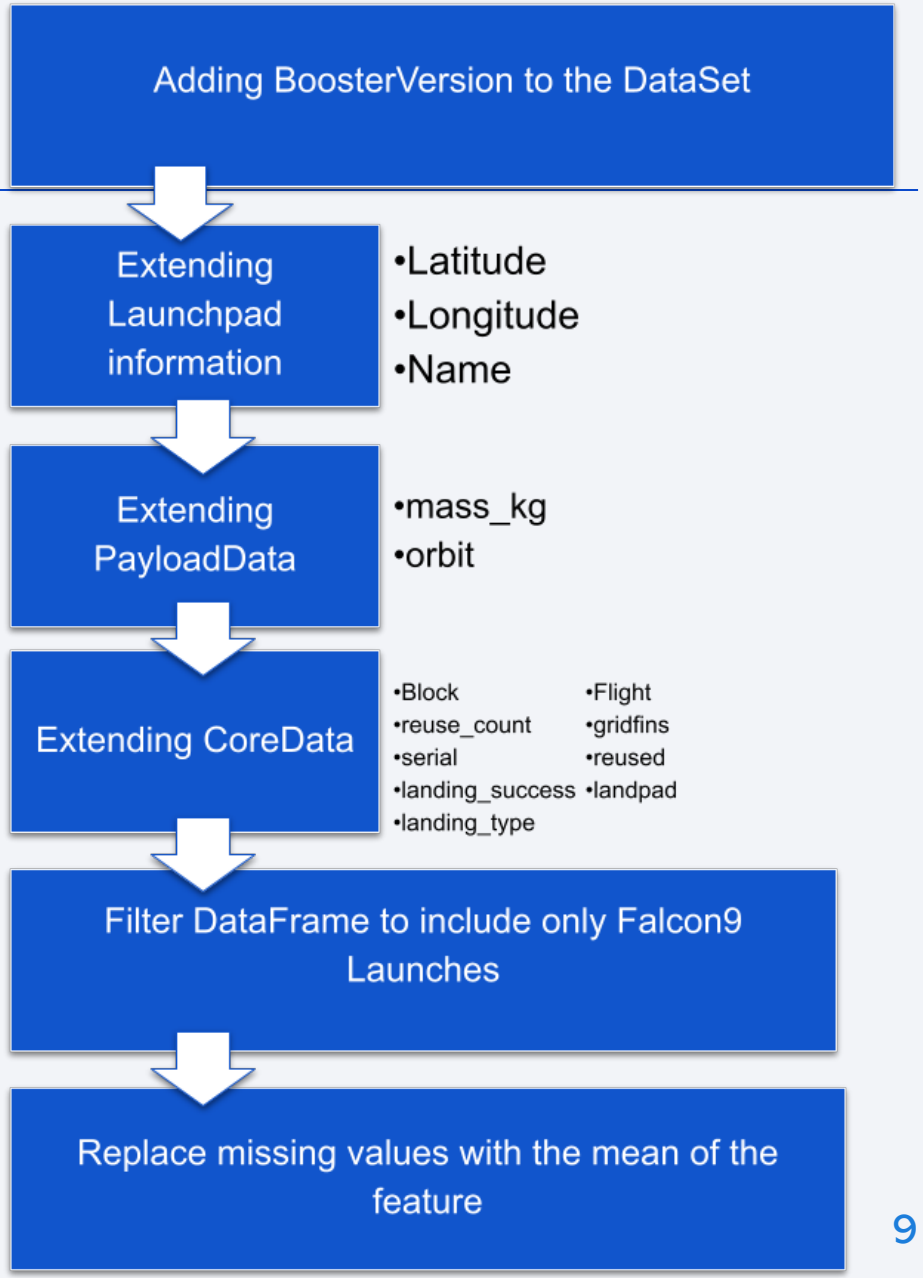
[Github Code]



Data Wrangling

1. Adding BoosterVersion to the data set
2. Extend Launchpad Information
3. Extending Payload
4. Filter data to keep Falcon 9 Launches
5. Replace missing values with the mean of the feature

[Github Code]



EDA with Data Visualization

- Among the plots that were used are:
  - Scatter plot to identify some patterns in the data
  - Bar plot provides a visual representation of the distribution of a dataset
  - Line graphs are used to track changes over short and long periods of time

[Github Code]

EDA with SQL

- Among the SQL queries are:
  - Display the names of the unique launch sites in the space mission
  - Display 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 successful landing outcome in ground pad was achieved.
  - 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. Use a subquery
  - 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.

- [Github Code]

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
- Explain why you added those objects
- Add the GitHub URL of your completed interactive map with Folium map, as an external reference and peer-review purpose

[Github Code]

Build a Dashboard with Plotly Dash

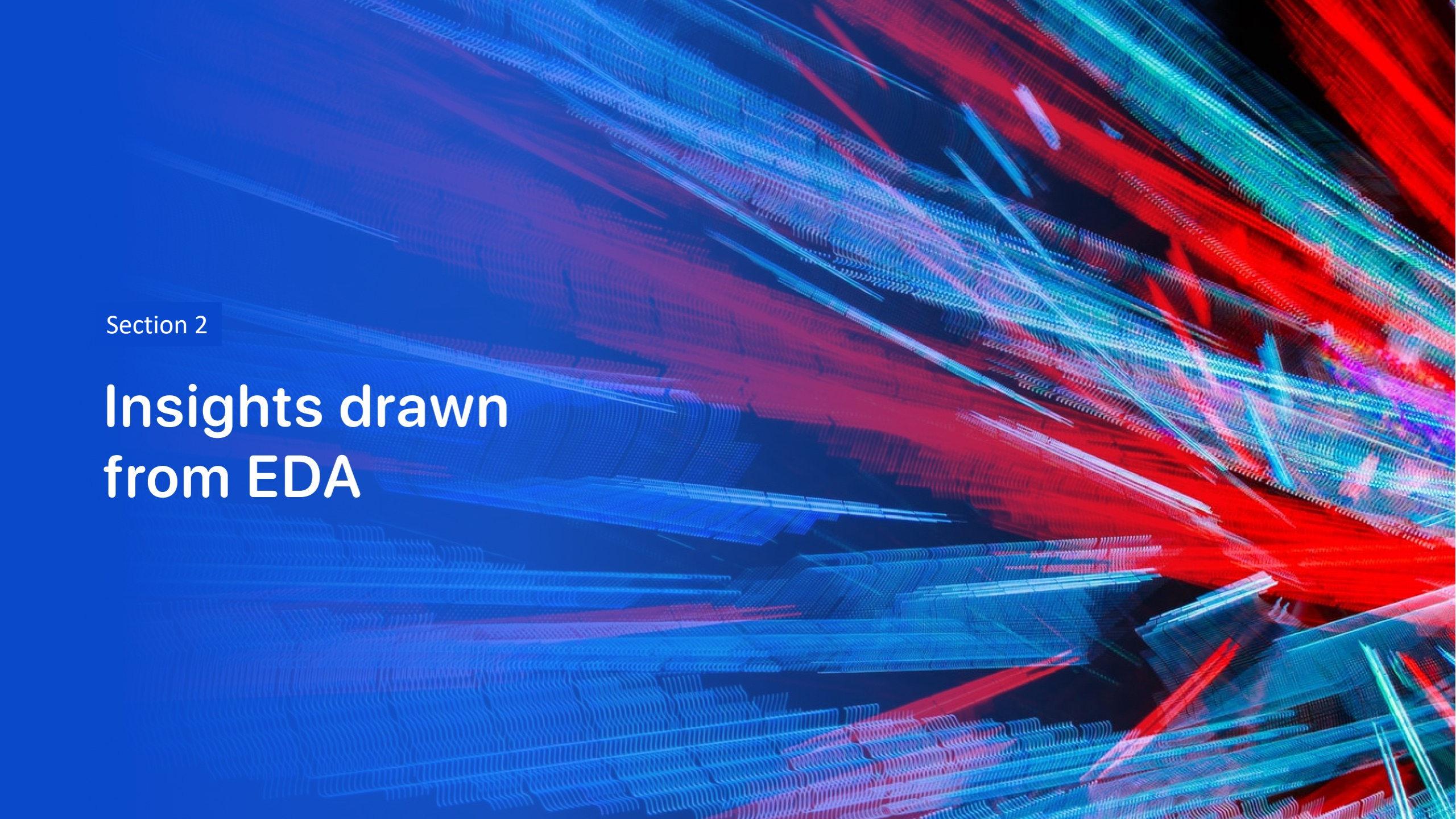
- The Dashboard includes two plots:
  - Pie Chart to show the success/failure ratio of all the Launch Sites. This plot is useful to show percentage of a whole.
  - Scatter Plot to show Payload and success for each launch Site. Scatter plots shows the relationship between two variables.

- [Github code]

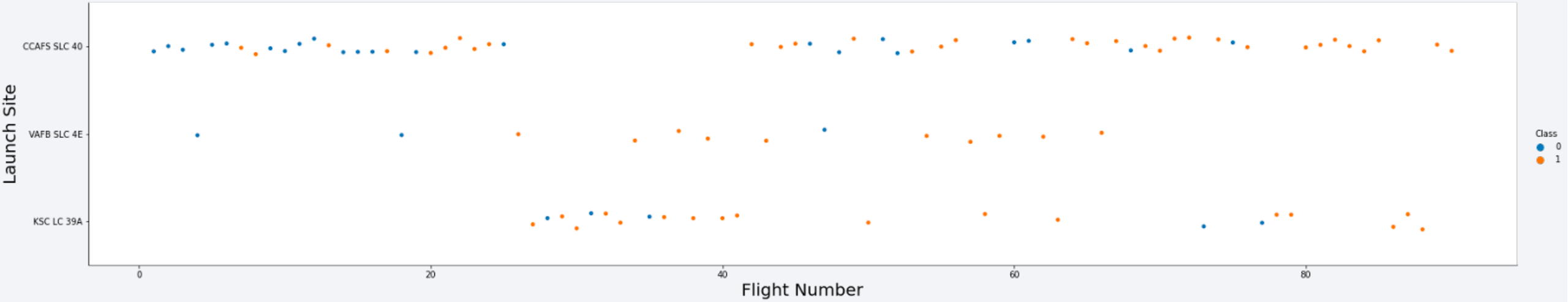
Predictive Analysis (Classification)

- Separate the dataset in Training and Testing, 80% and 20% respectively
- Create a Model
- Set the hyperparameters to tune. Each model has a set of hyperparameters to tune.
- Set the GridSearch over all those hyperparameters and Cross Validation with K=10.
    - Crossvalidation divide the training set in K subsets and it trains the model with k-1 and test with the one that was not used to train. All the k sets are used as testing once. After finishing the training the average of each metrics will represents the metric of the whole model
- Get the accuracy, f1, precision and recall to decide which model is the best.

- [Github Code]

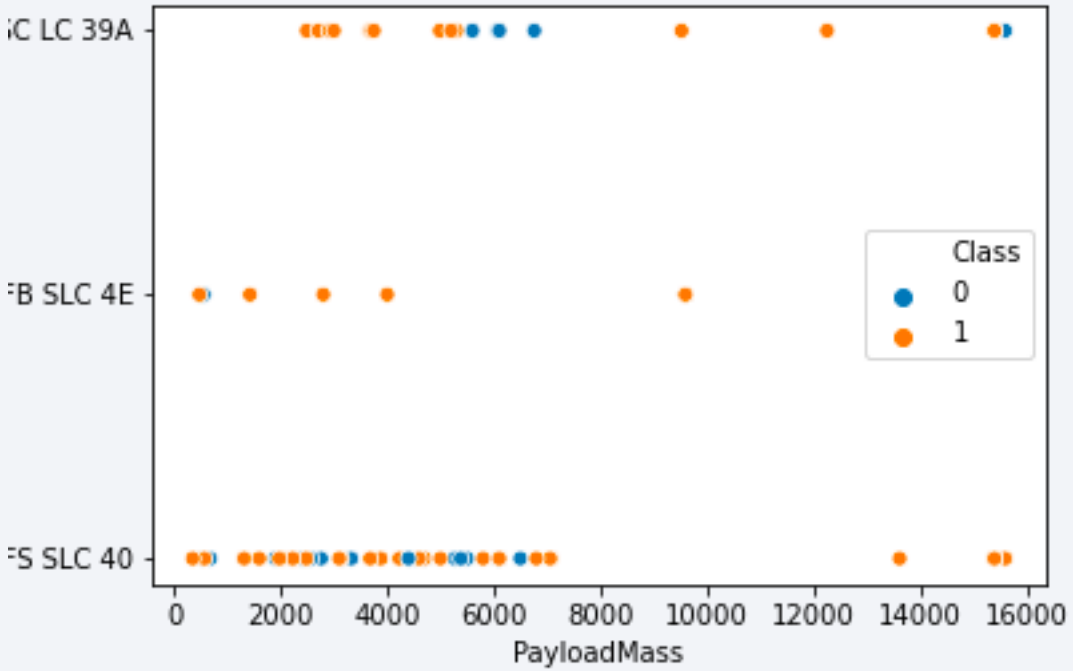

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



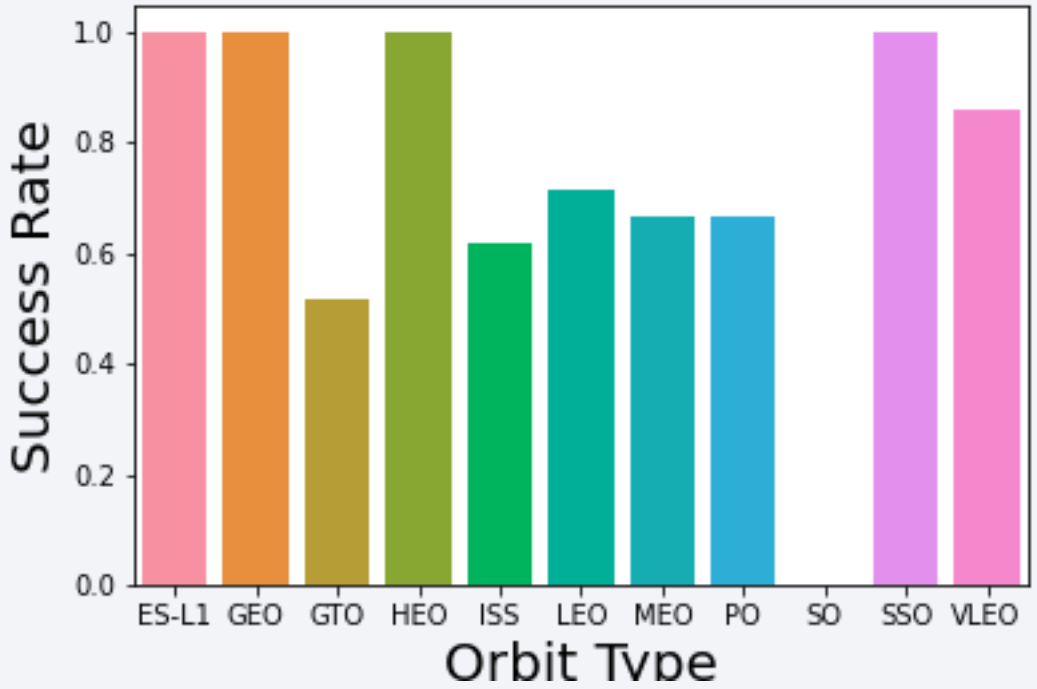
The first Flights that were launched from CCAFS SLC 40 seems to be more successful, meaning, the first stage will return

Payload vs. Launch Site



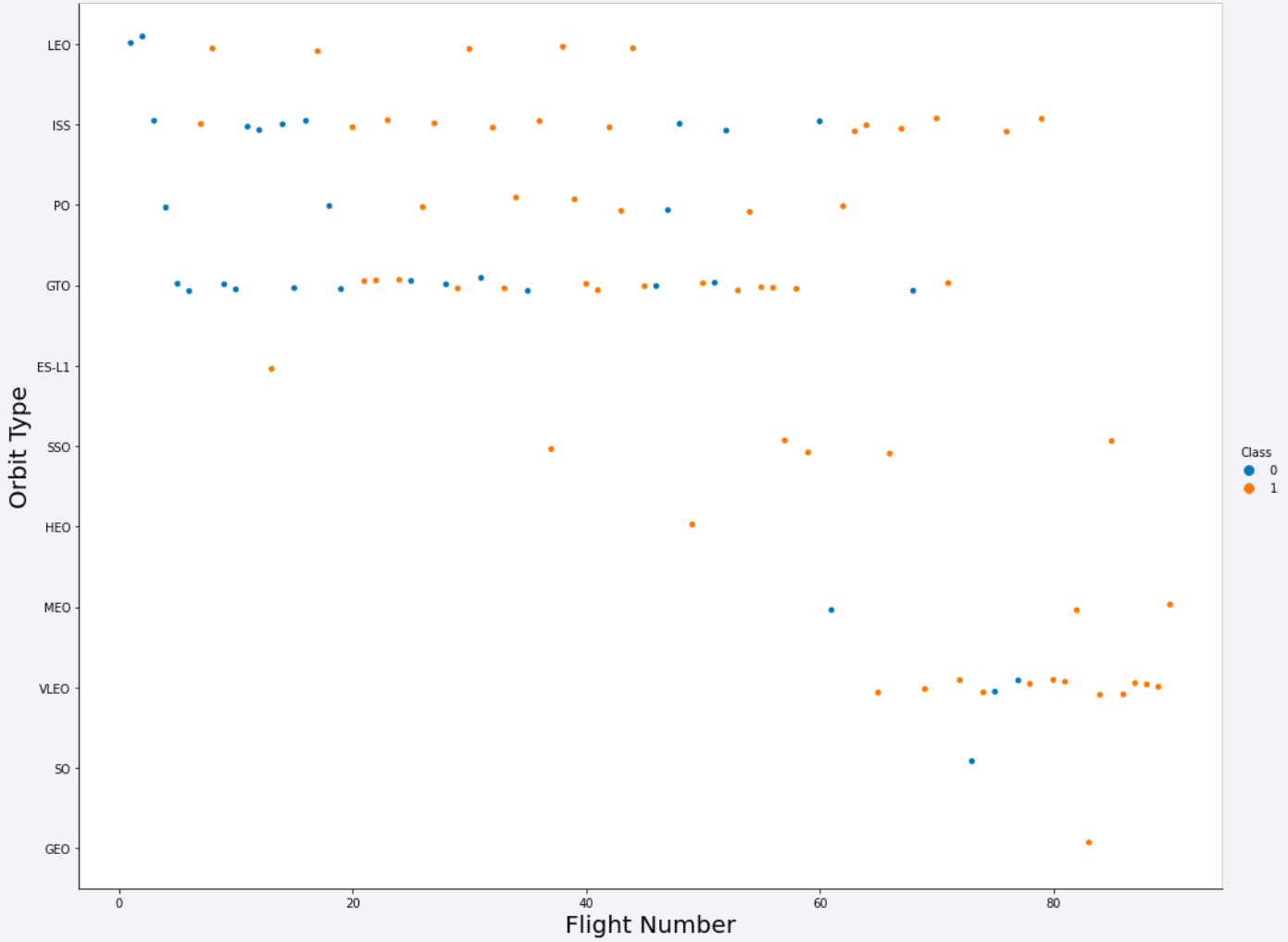
The payload and launch site might not be correlated.
We can see that SLC4E had the least launches but all of them successful. SLC 40 had the most launches, the most of the launches were between 0 and 8000. However the heaviest between 13000 and 16000 were all successful. LC 39A failed when the payload was between 6000 and 8000. The heaviest also failed.

Success Rate vs. Orbit Type



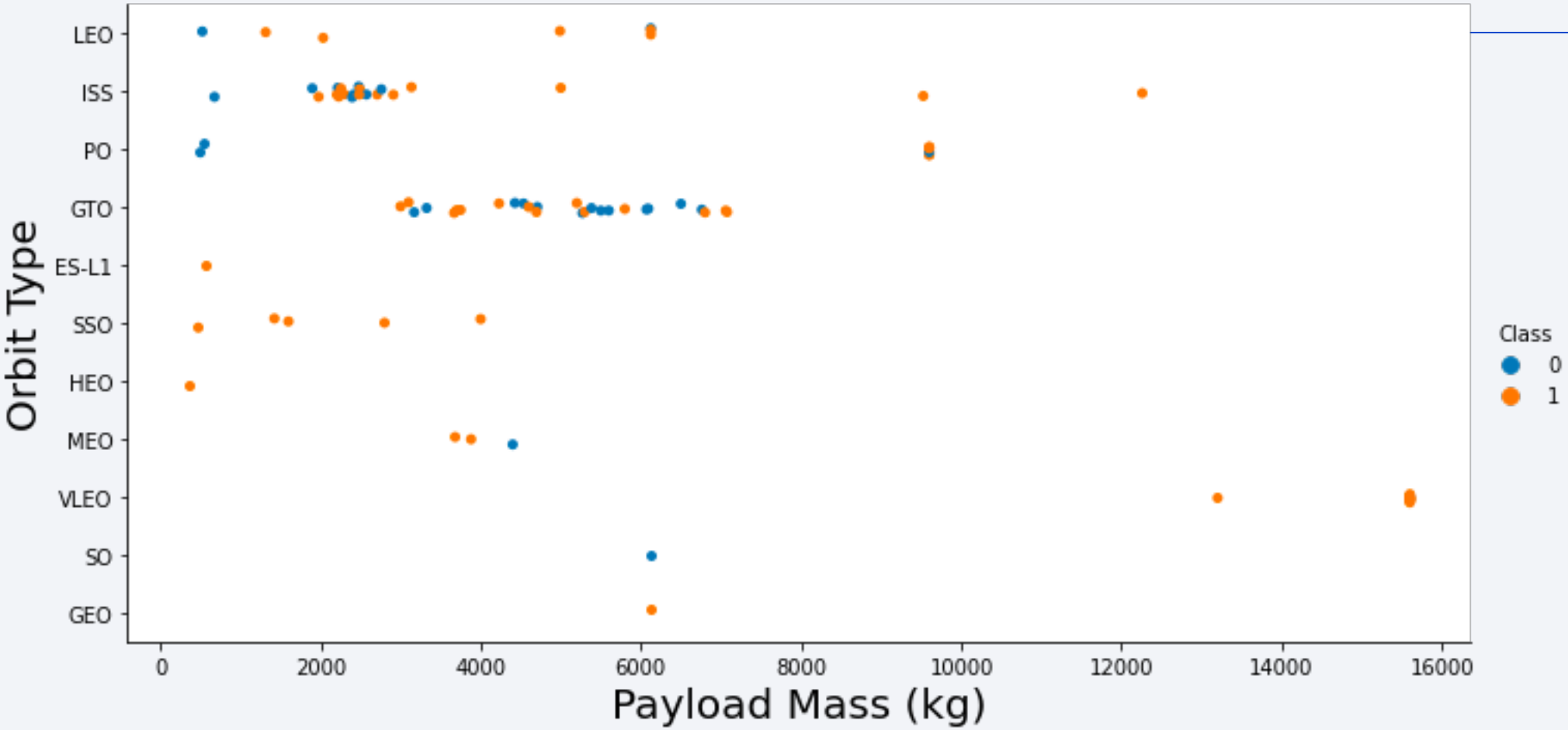
ES-L1, GEO, HEO, S50 are the orbits with more successful rate, while SO is the orbit with the less successful rate.
On the other hand, GTO has the least successful rate.

Flight Number vs. Orbit Type



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 GTO orbit.

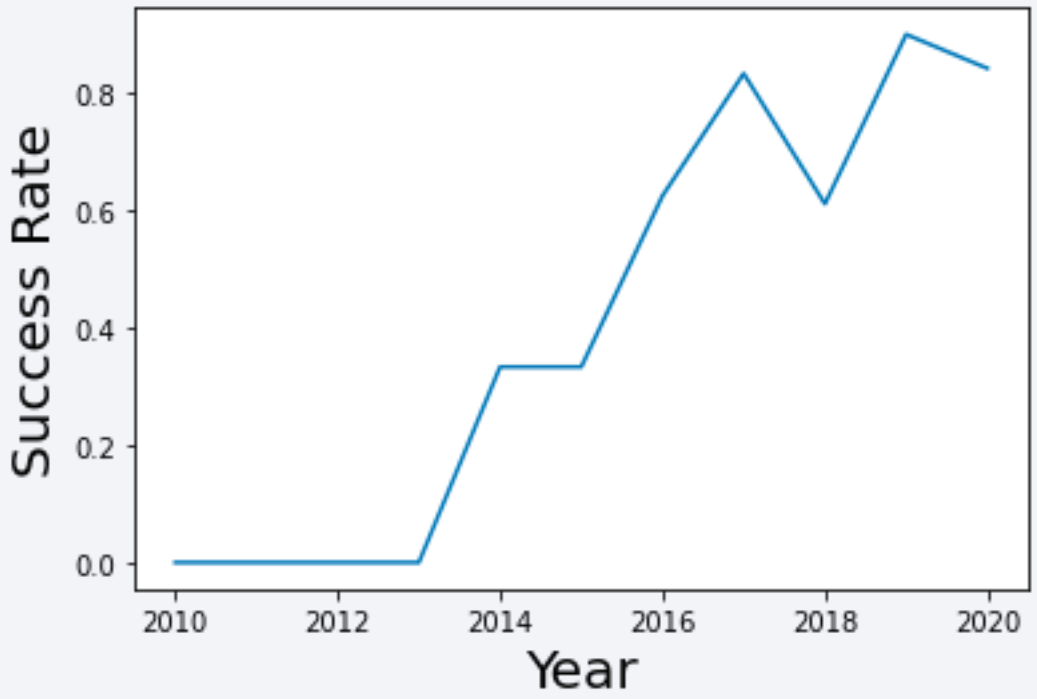
Payload vs. Orbit Type



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



A hypothesis about this trend is related to the normal improvement of technology over the years.

All Launch Site Names

Task 1

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

```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTBL;
```

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

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

I used SQL to get the Site Names. Using the instruction DISTINCT I was able to filter only the different names.
As see in the figure, there are 4 different Launch Sites in the Data Set

Launch Site Names Begin with 'CCA'

Task 2

Display 5 records where launch sites begin with the string 'CCA'

In [19]:
```
%sql SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE "CCA%" LIMIT 5
```

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

Out[19]:

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

- To find the Launch Sites that start with with CCA I used the operator LIKE and I limited the search to 5 with LIMIT keyword.

Total Payload Mass

Task 3

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

In [21]:
```
%sql SELECT sum(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer = "NASA (CRS)";
```

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

Out[21]:

sum(PAYLOAD_MASS__KG_)
45596

- To calculate the total payload mass it was necessary to use the SUM operator.
- It was filtered to include only the customer NASA (CRS)

Average Payload Mass by F9 v1.1

Task 4

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

In [24]:
```
%sql SELECT avg(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version = "F9 v1.1";
```

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

Out[24]:

avg(PAYLOAD_MASS__KG_)
2928.4

- To calculate the average payload mass it was necessary to use the AVG operator.

First Successful Ground Landing Date

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

In [27]:
```
%sql SELECT min(Date) FROM SPACEXTBL WHERE "Landing _Outcome" = "Success (ground pad)"
```

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

Out[27]:

min(Date)
01-05-2017

- To find the first successful landing. I defined as Success ground pad and select the minimum date.

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [28]:
```
%sql SELECT Booster_Version FROM SPACEXTBL WHERE "Landing _Outcome" = "Success (drone ship)" and PAYLOAD_MASS__KG_ between 4000 and 6000;
```

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

Out[28]:

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

- The dataset was filtered to include only Successful (drone ship) and using the operator BETWEEN I set the values to be 4000 and 6000

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

In [32]:
```
%sql SELECT Mission_Outcome, count(Mission_Outcome) FROM SPACEXTBL GROUP BY Mission_Outcome
```

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

Out[32]:

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

- To get the number of successful and failed missions I grouped the data set using the feature mission outcome, then I used the operator count

Boosters Carried Maximum Payload

Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [47]:
```
%sql CREATE VIEW SUM_PAYLOAD_MASS AS SELECT Booster_Version, sum(PAYLOAD_MASS__KG_) as SUM_PAYLOAD_MASS_KG FROM SPACEXTBL GROUP BY Booster
%sql SELECT Booster_Version  FROM SUM_PAYLOAD_MASS WHERE SUM_PAYLOAD_MASS_KG == (SELECT MAX(SUM_PAYLOAD_MASS_KG) FROM SUM_PAYLOAD_MASS)
```

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

Out[47]:

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

- First it was necessary to get the payload by carrier
- Then I filtered selecting the maximum payload by carrier.

2015 Launch Records

Task 9

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

In [54]:
```
%sql SELECT CASE substr(Date, 4, 2) when '01' then 'January' when '02' then 'Febuary' when '03' then 'March' when '04' then 'April' when
```

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

Out[54]:

MONTH	Landing _Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- The two months with failure outcomes were January and April

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

Task 10

Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

In [60]:
```
%sql SELECT "Landing _Outcome", count("Landing _Outcome") FROM SPACEXTBL WHERE "Landing _Outcome" LIKE "Success%" and Date BETWEEN "04-06-
```

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

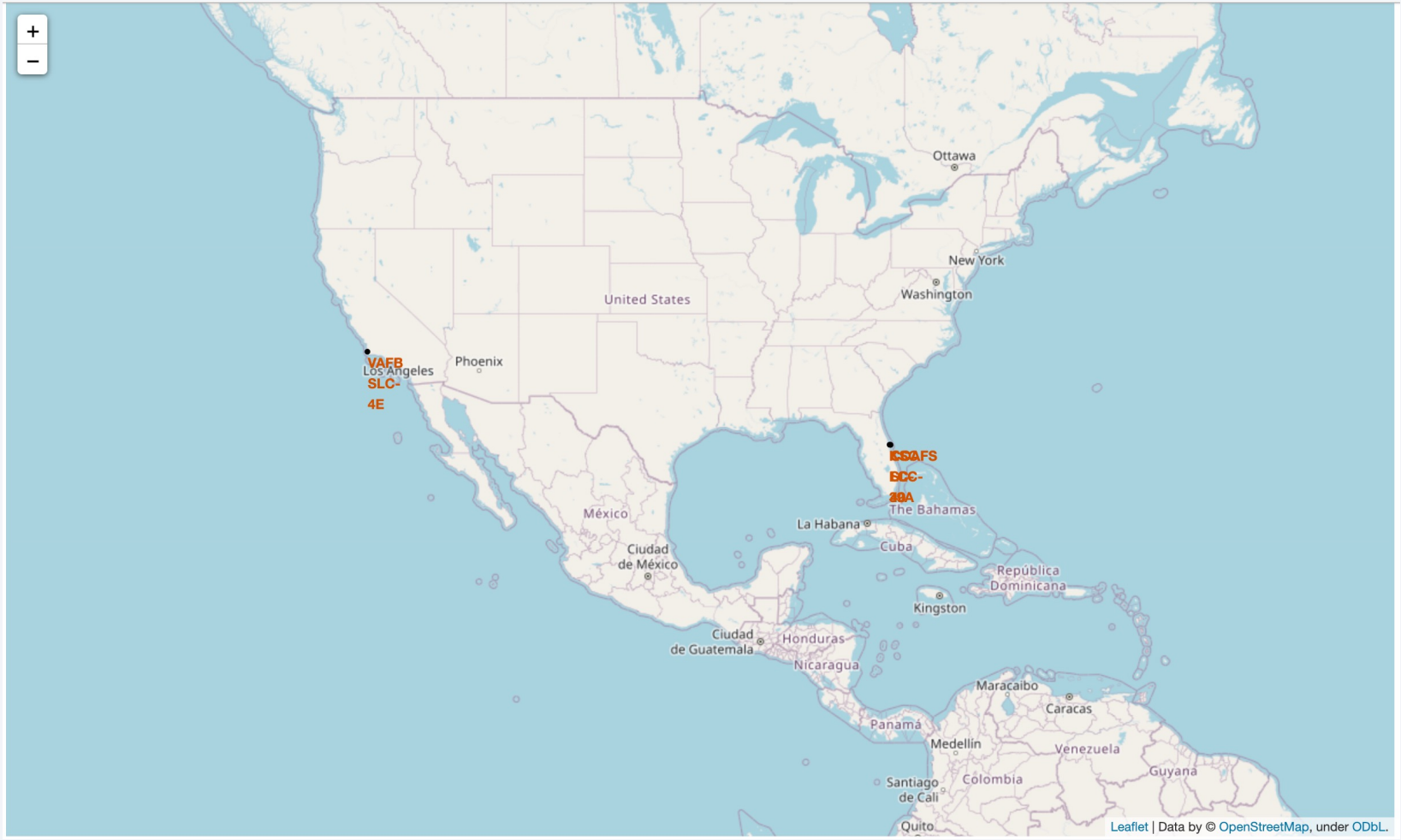
Out[60]:

Landing _Outcome	count("Landing _Outcome")
Success	20
Success (drone ship)	8
Success (ground pad)	6

Because Success was defined in different ways, I had to use LIKE operator to get them, after that I applied the count and order DESC

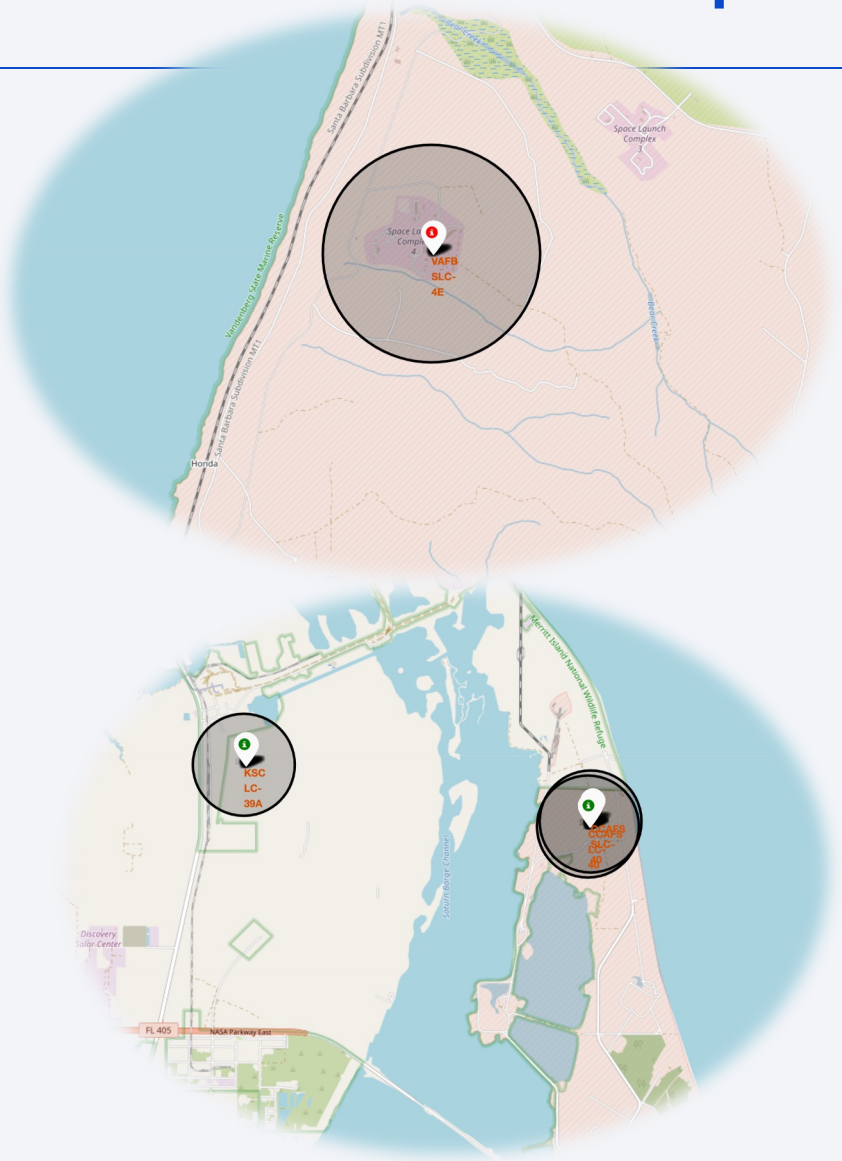

Section 3
Launch Sites
Proximities Analysis

Launch sites on a map



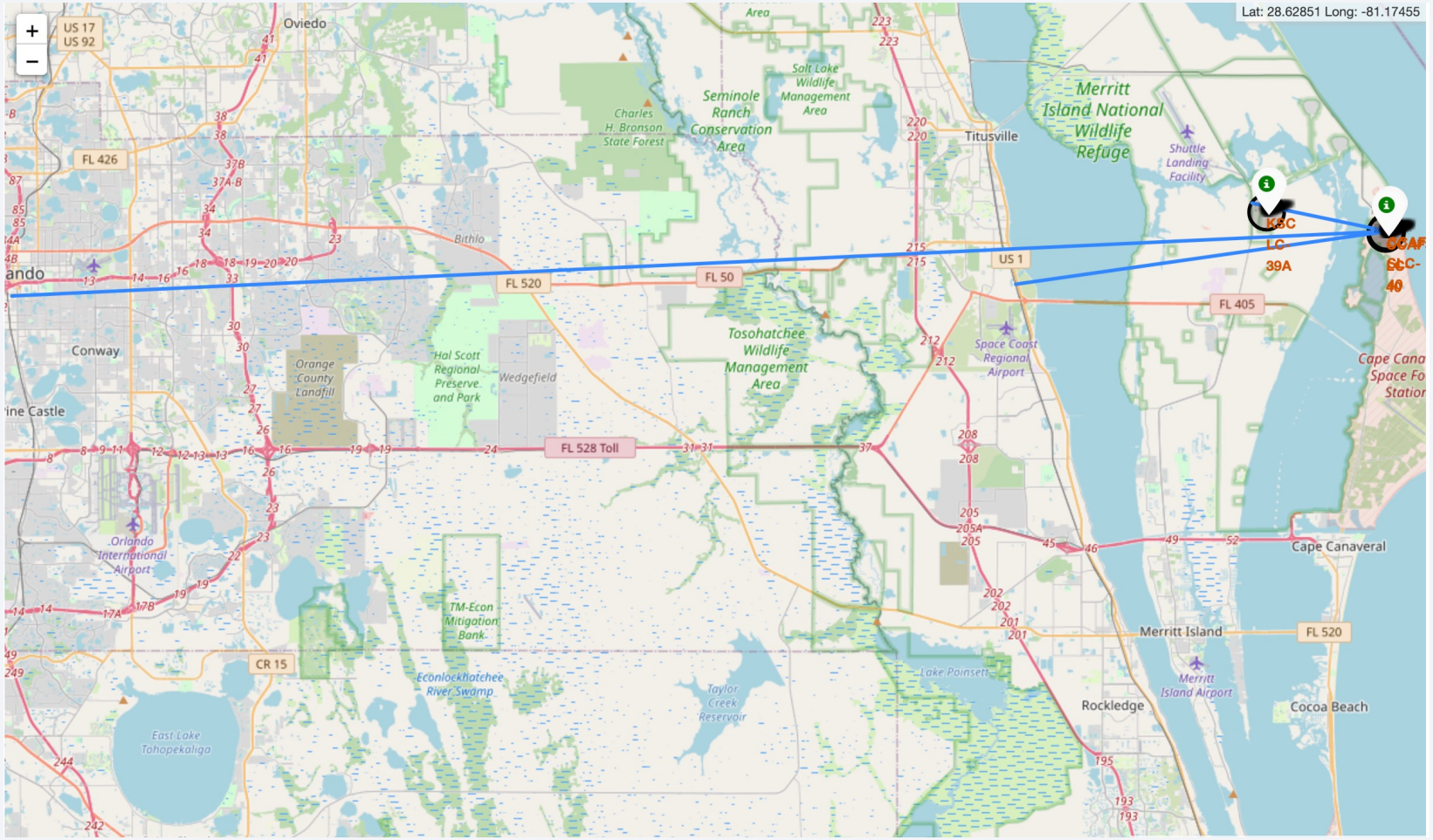
- There are basically three launch sites. One is located in California and the other two in Florida

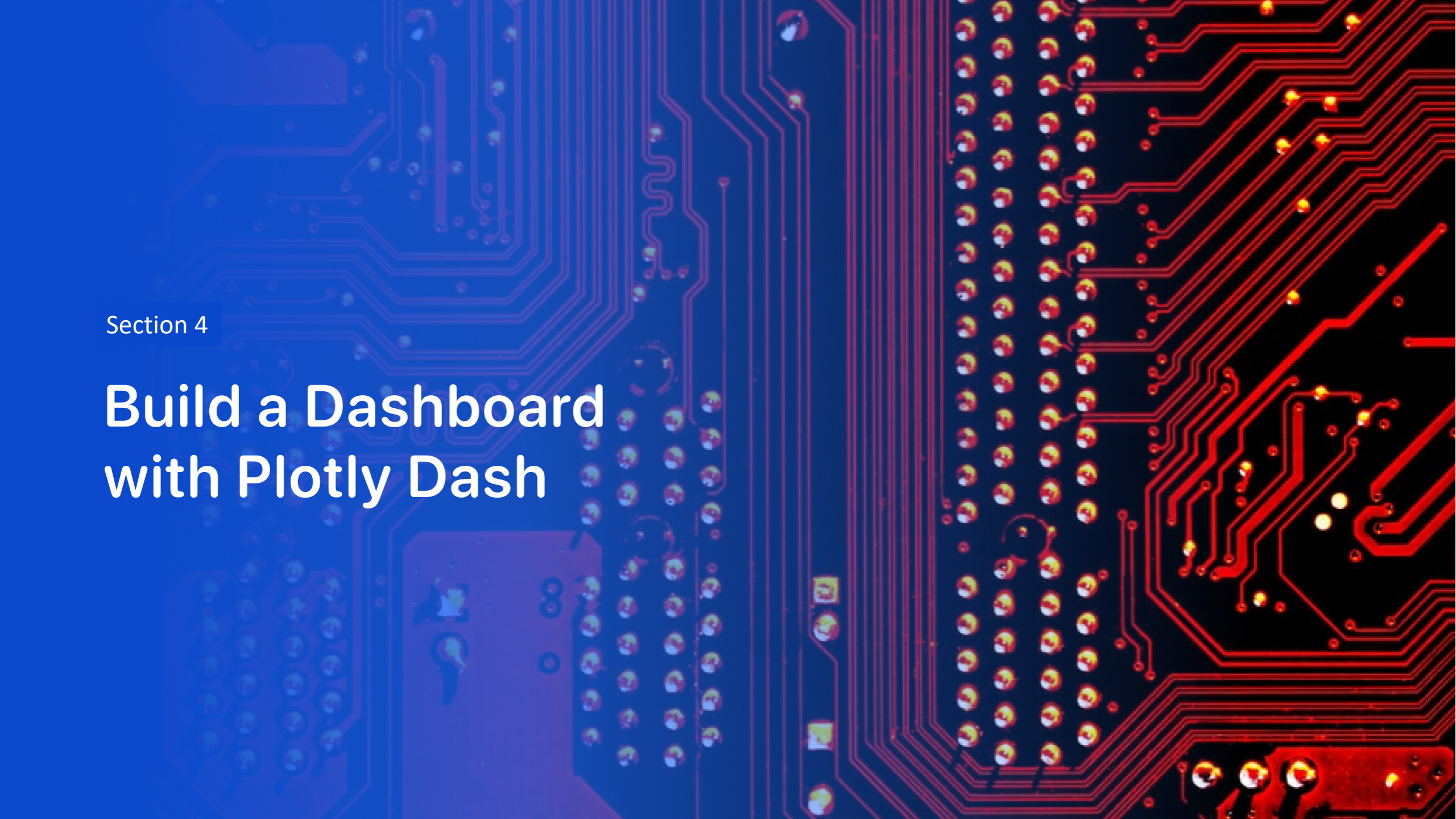
Success/failed launches for each site on the map



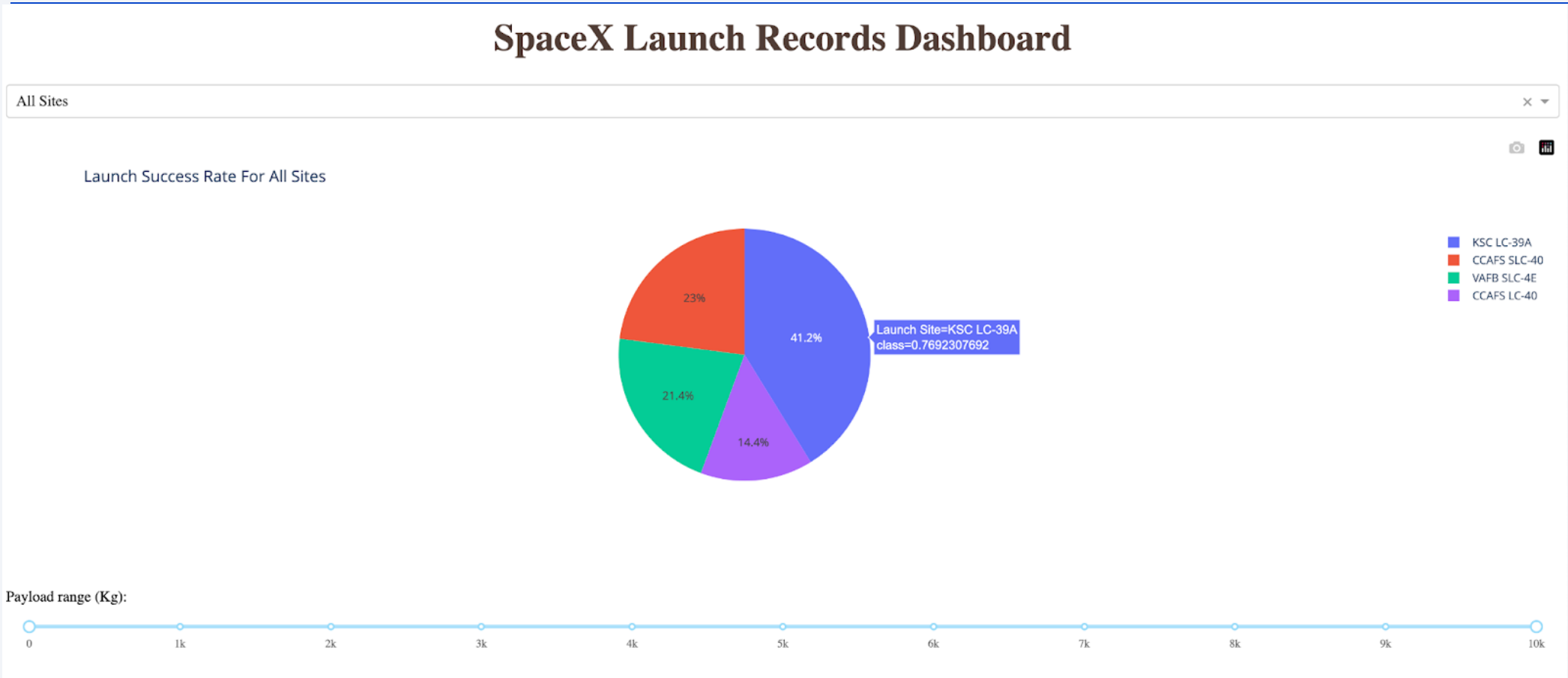
The distances between a launch site to its proximities

- Distances From CCAFS SLC 40 to:
  - Orlando
  - Kennedy Parkway North
  - Florida East Coast Rallway



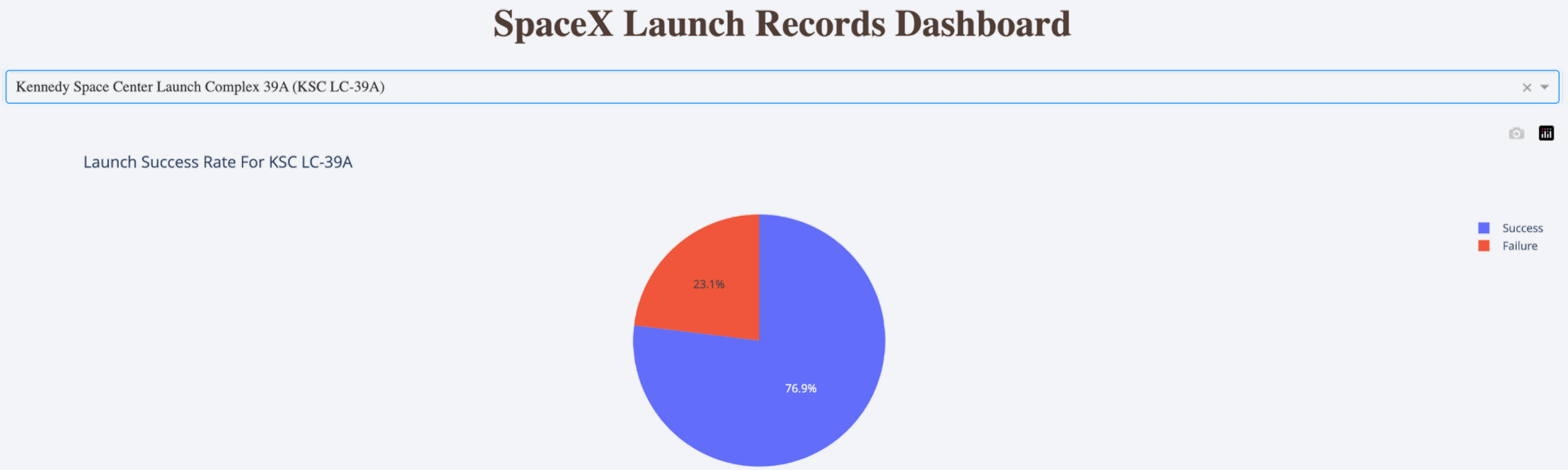

Section 4
Build a Dashboard with Plotly Dash

Launch Success Rate for all Sites



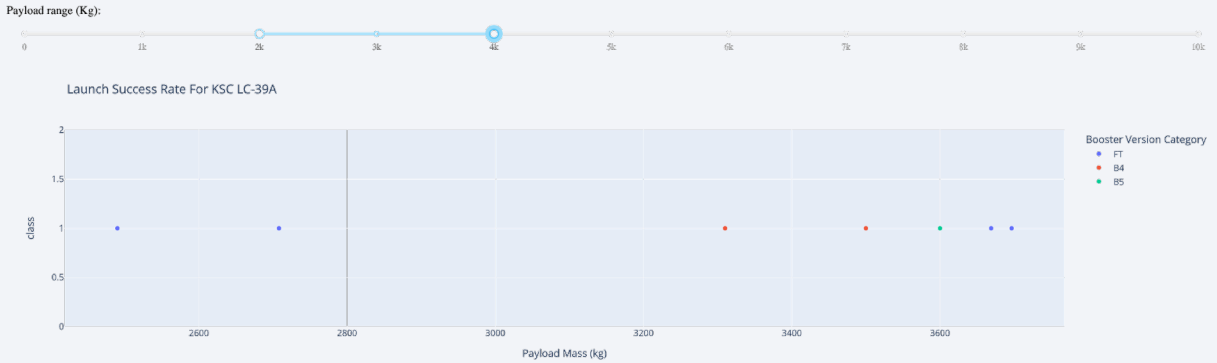
The most successful site was KSC LC 39A with 41.2% while the least successful one was CCAFS LC-40

The launch site with highest launch success ratio

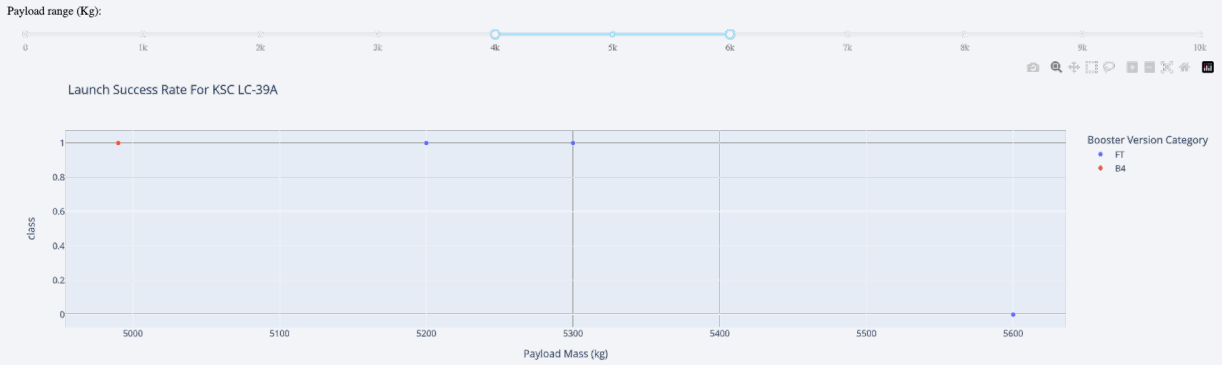


- Kennedy Space Center has the most success ratio from all the launch sites. The success ratio is 76.9%.

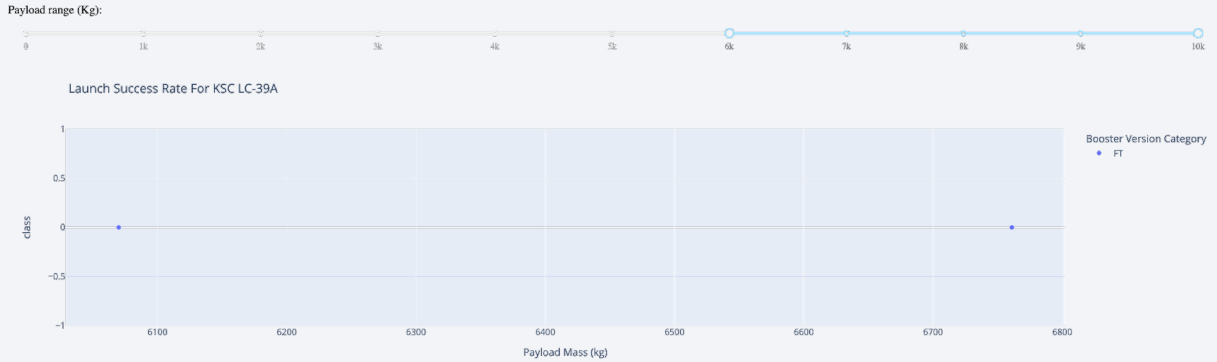
Payload vs Launch Outcome



Payload in (2k and 4k)



Payload in (4k and 6k)

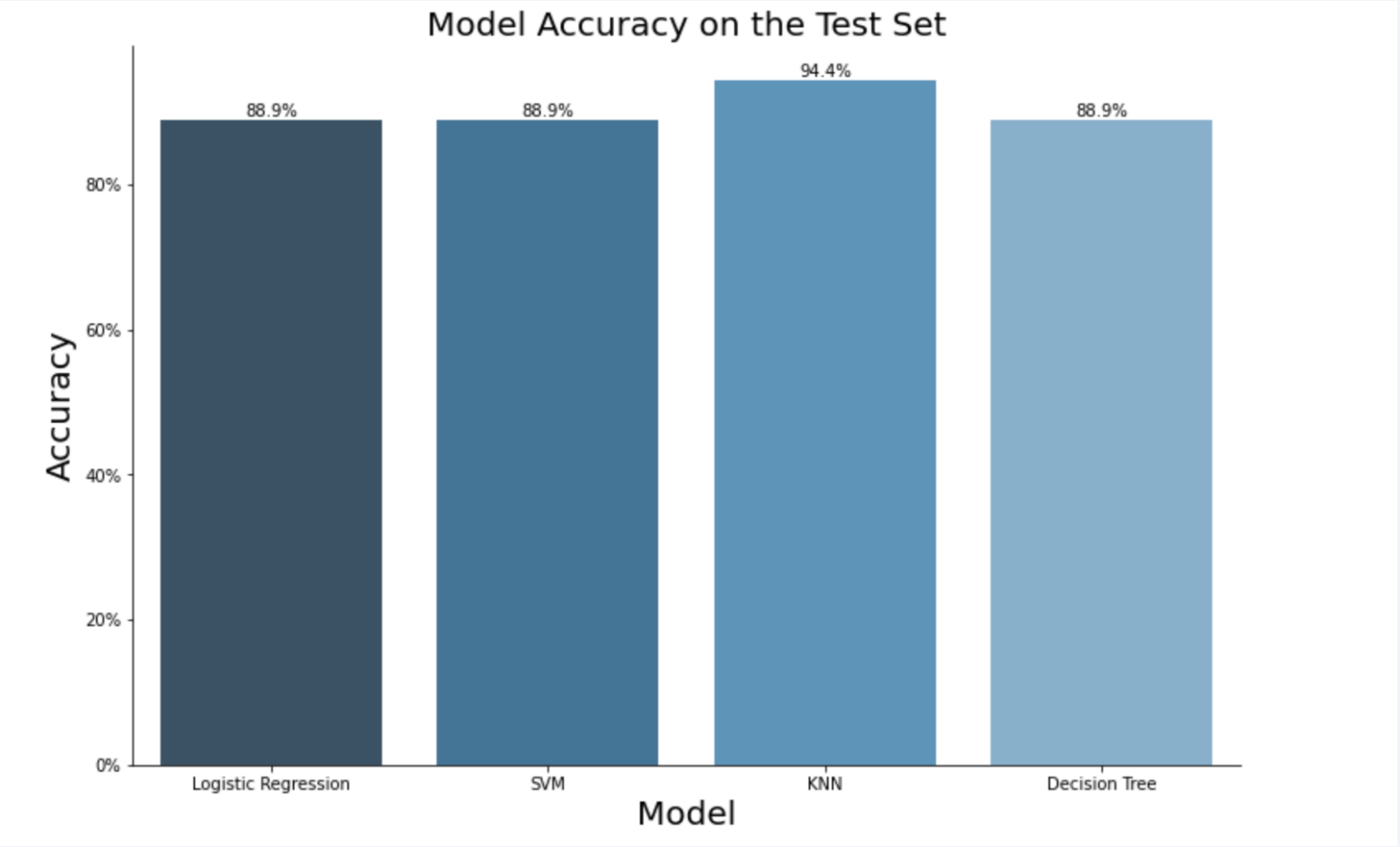


Payload greater than 6K

Section 5

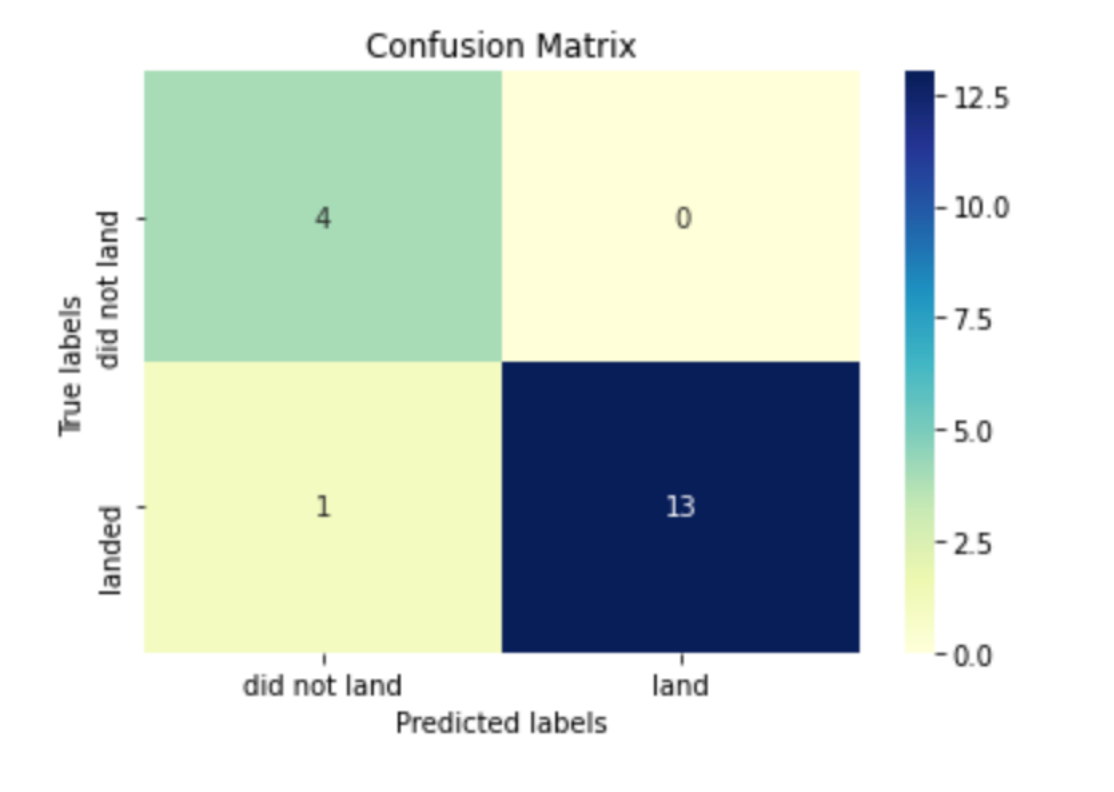
Predictive Analysis (Classification)

Classification Accuracy



In this experiment, the highest accuracy in the test set was KNN

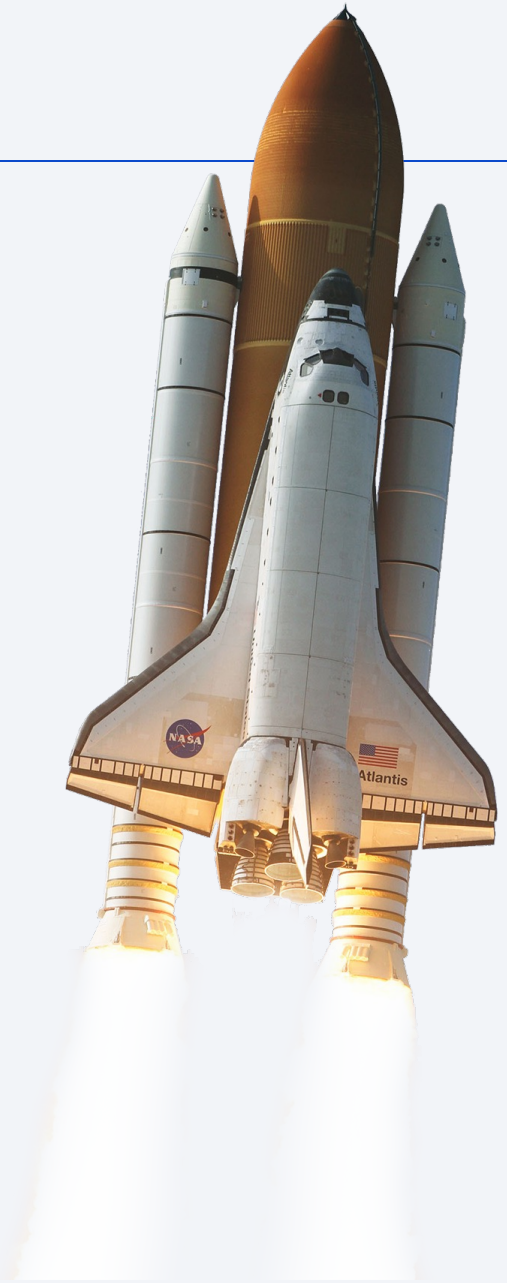
Confusion Matrix



- The Confusion Matrix shows that the model did not have False Positive.
- It has only one False Negative.

Conclusions

- This project shows that is possible to determine is a launch could land or not successfully.
- There is a correlation between payload and success rate.
- More deep research is need it.



Thank you!