IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Flávio Zanette de Angeli
21/06/2022

Outline

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

Executive Summary

- Summary of methodologies:
    1. Data collection using webscraping and API;
    2. Data wrangling;
    3. Exploratory Data Analysis using SQL and Data visualization;
    4. Machine Learning models for prediction of successful launch.
- Summary of all results:
    1. Will be presented a efficient way to collect the data from different sources;
    2. A brief exploratory analysis to find the most important features;
    3. Machine Learning results with the best models and parameters to predict launches.

Introduction

- Project background and context

    1. We have the objective to understand the viability and costs of starting in the rocket launch bussiness.

- Problems you want to find answers

    1. How to predict successful launches to estimate and reduce the cost of the company;
    2. Best location to perform launches and build new lauch sites.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data collected from the API from Space X
  - Data collected by a web scraping in the Wikipedia page of Falcons launchs.
- Perform data wrangling
  - A data wrangling was made, dealing with missing data and other problems.
- 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 split between training and test sets and use o grid search to find best parameters combination for each model.

Data Collection

- Data collected from the API from Space X
  - https://api.spacexdata.com/v4/rockets/
- Data collected by a web scraping in the Wikipedia page of Falcons launchs
  - https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches

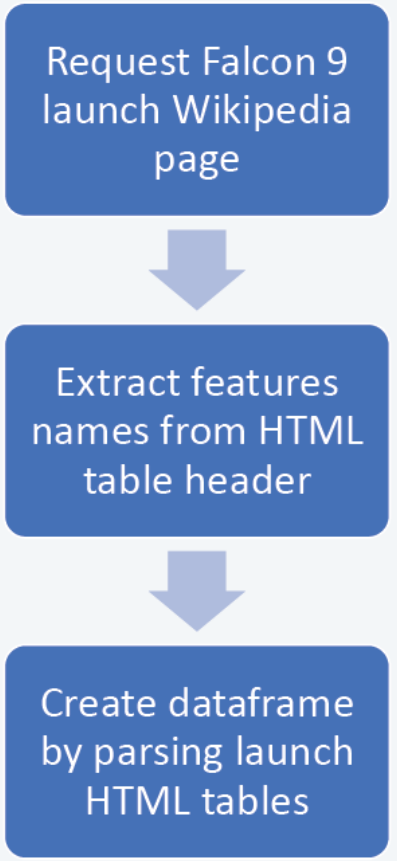
Data Collection – SpaceX API

- Were used the SpaceX API, it is a public APO to access data from launchs.
- GitHub URL: https://github.com/FlavioZanette/final-task-applied-data-science-ibm/blob/9302b3c0e2285ae2b8154fb48296c9c4c96f757c/Data%20collection%20api.ipynb

Data Collection - Scraping

- Data was collected using web scraping methods in the Wikipedia page of the lauchs.

- GitHub URL: https://github.com/FlavioZanette/final-task-applied-data-science-ibm/blob/9302b3c0e2285ae2b8154fb48296c9c4c96f757c/Webscraping.ipynb



Data Wrangling

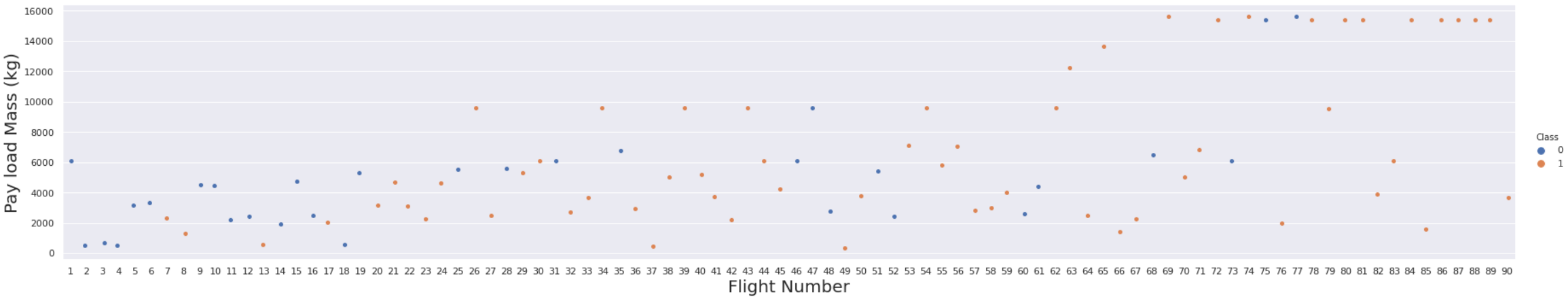
- Initial EDA with SQL was performed to understand the dataset.
- The data war summarise like launches per site and occurrence of each orbit.
- Definition of landing outcome label.

- GitHub URL: https://github.com/FlavioZanette/final-task-applied-data-science-ibm/blob/9302b3c0e2285ae2b8154fb48296c9c4c96f757c/Data%20wrangling.ipynb

EDA with Data Visualization

- To explore data, scatterplots and barplots were used to visualize the relationship between pair of features:
    - Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, Payload and Orbit

GitHub URL: https://github.com/FlavioZanette/final-task-applied-data-science-ibm/blob/9302b3c0e2285ae2b8154fb48296c9c4c96f757c/EDA%20Dataviz.ipynb



EDA with SQL

- The following SQL queries were performed:
  1. Names of the unique launch sites in the space mission
  2. Top 5 launch sites whose name begin with the string 'CCA'
  3. Total payload mass carried by boosters launched by NASA (CRS);
  4. Average payload mass carried by booster version F9 v1.1
  5. Date when the first successful landing outcome in ground pad was achieved
  6. Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg
  7. Total number of successful and failure mission outcomes
  8. Names of the booster versions which have carried the maximum payload mass
  9. Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
  10. Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 04/06/10 and 20/03/17

- GitHub URL: https://github.com/FlavioZanette/final-task-applied-data-science-ibm/blob/9302b3c0e2285ae2b8154fb48296c9c4c96f757c/EDA%20SQL.ipynb

Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters were used with Folium Maps
- Markers indicate points like launch sites
- Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center
- Marker clusters indicates groups of events in each coordinate, like launches in launch site
- Lines are used to indicate distances between two coordinates

- GitHub URL:https://github.com/FlavioZanette/final-task-applied-data-science-ibm/blob/9302b3c0e2285ae2b8154fb48296c9c4c96f757c/Launch%20site%20location.ipynb

Build a Dashboard with Plotly Dash

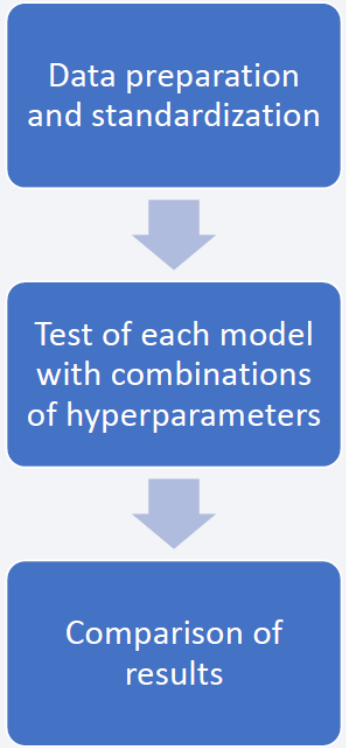
- The following graphs and plots were used to visualize data
    1. Percentage of launches by site
    2. Payload range
- This combination allowed to quickly analyze the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads.

- GitHub URL:https://github.com/FlavioZanette/final-task-applied-data-science-ibm/blob/9302b3c0e2285ae2b8154fb48296c9c4c96f757c/Dash%20code.py

Predictive Analysis (Classification)

Four classification models were tested:

1. Logistic Regression
2. SVM
3. Decision Tree
4. KNN



- GitHub URL:https://github.com/FlavioZanette/final-task-applied-data-science-ibm/blob/9302b3c0e2285ae2b8154fb48296c9c4c96f757c/Machine%20Learning%20Prediction.ipynb

Results

- Exploratory data analysis results:
    1. Space X uses 4 different launch sites
    2. Majority of flights took off from CCAFS SLC 40
    3. VAFB site has the least amount of flight, with almost 77% of success rate, same as KSC
    4. The first success landing outcome happened in 2015 fiver year after the first launch
    5. Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average
    6. With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS
    7. Success rate since 2013 kept increasing till 2020

Results

- Exploratory data analysis results:
    1. Space X uses 4 different launch sites
    2. Majority of flights took off from CCAFS SLC 40
    3. VAFB site has the least amount of flight, with almost 77% of success rate, same as KSC
    4. The first success landing outcome happened in 2015 fiver year after the first launch
    5. Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average
    6. With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS
    7. Success rate since 2013 kept increasing till 2020

Results

- Launch site location analysis results:
    1. Launch sites are close to proximity to railways, this facilitate the transport of heavy and big materials.
    2. Launch sites are close to highways, this facilitate the transport of workers and general supply.
    3. Launch sites are close to coastline so it is possible to fly over the ocean.
    4. Launch sites are not close to cities, which decrease the risk of accident with people.

Results

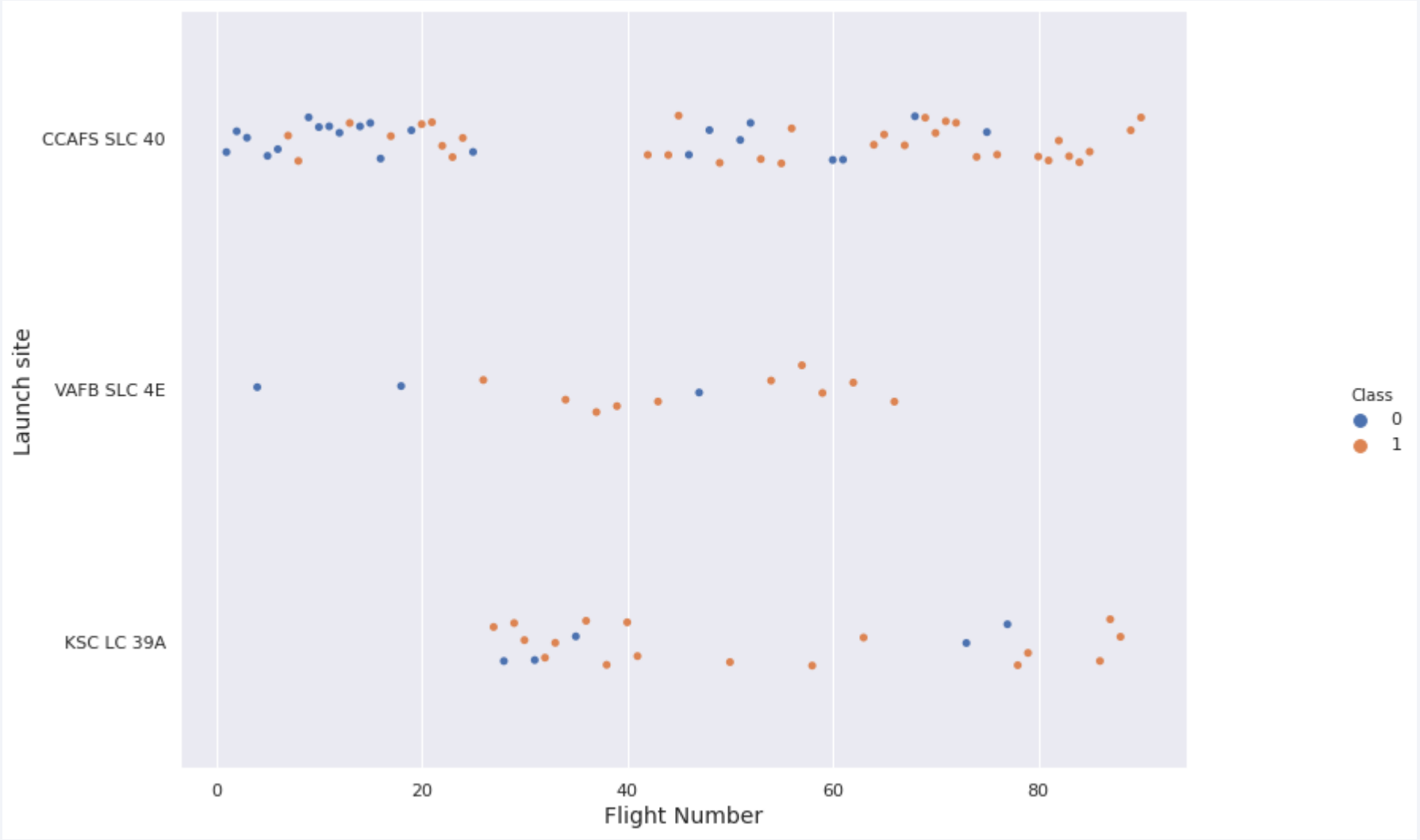
- Predictive Analysis showed that SVM, LogReg and KNN are the best models to predict successful landings, having train accuracy over 84,72% and test accuracy for test data over 83,33%.

	Model	Train	Test
0	LogReg	0.847222	0.833333
1	SVM	0.847222	0.833333
2	DecTree	0.875	0.666667
3	KNN	0.847222	0.833333

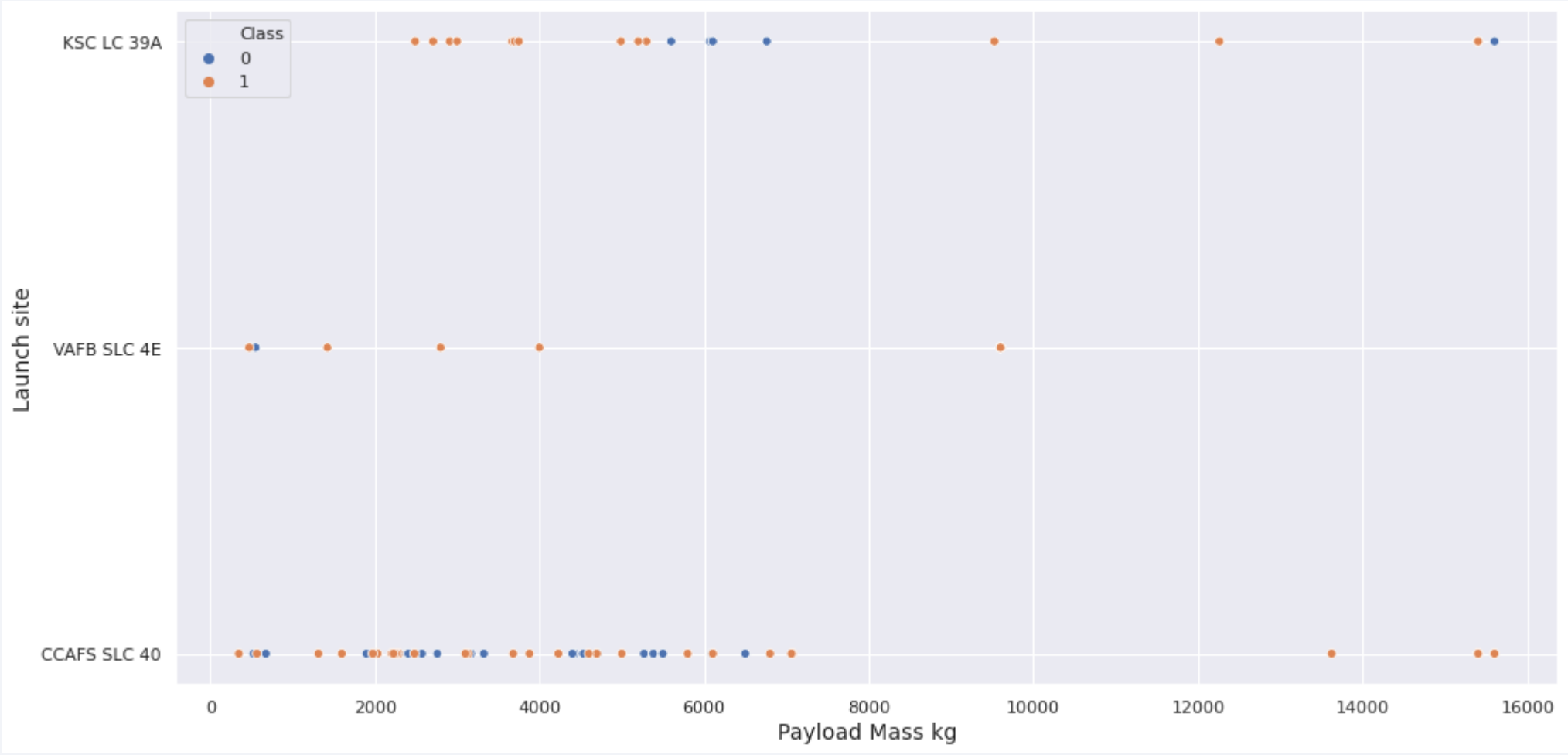
Section 2

Insights drawn from EDA

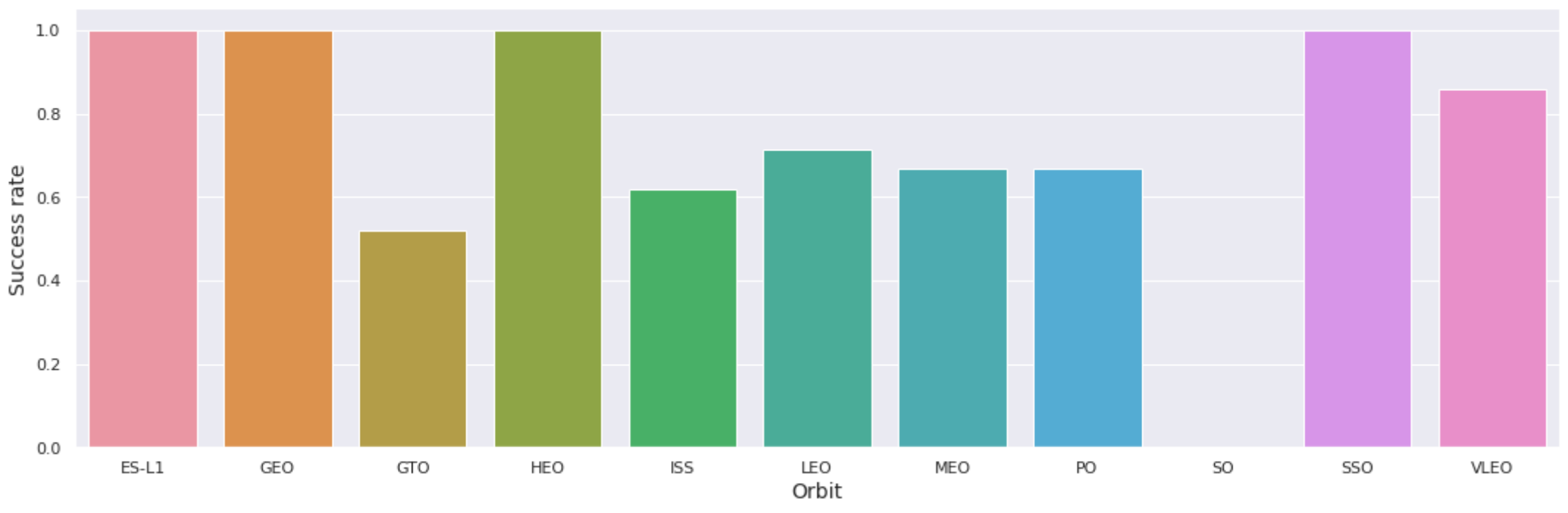
Flight Number vs. Launch Site



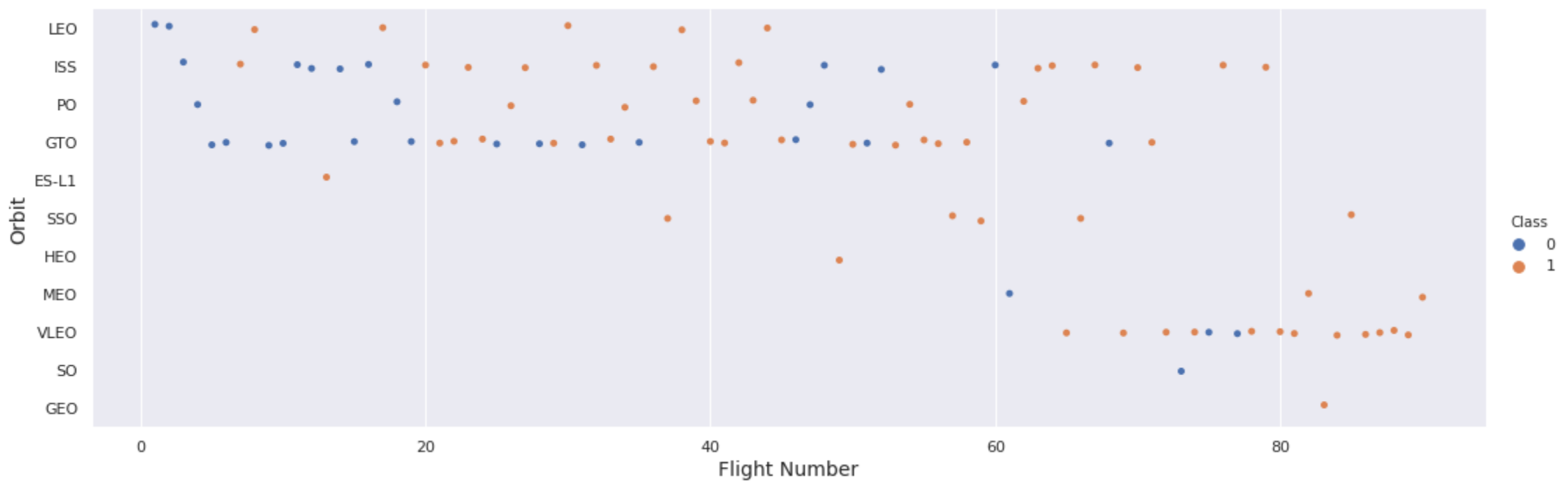
Payload vs. Launch Site



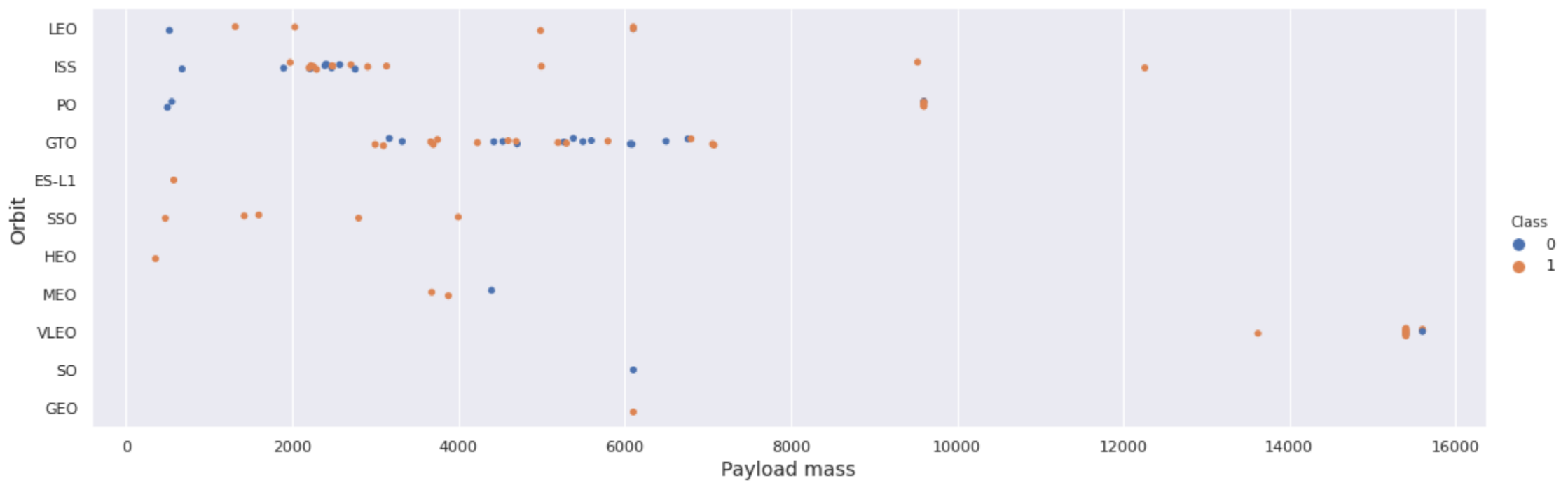
Success Rate vs. Orbit Type



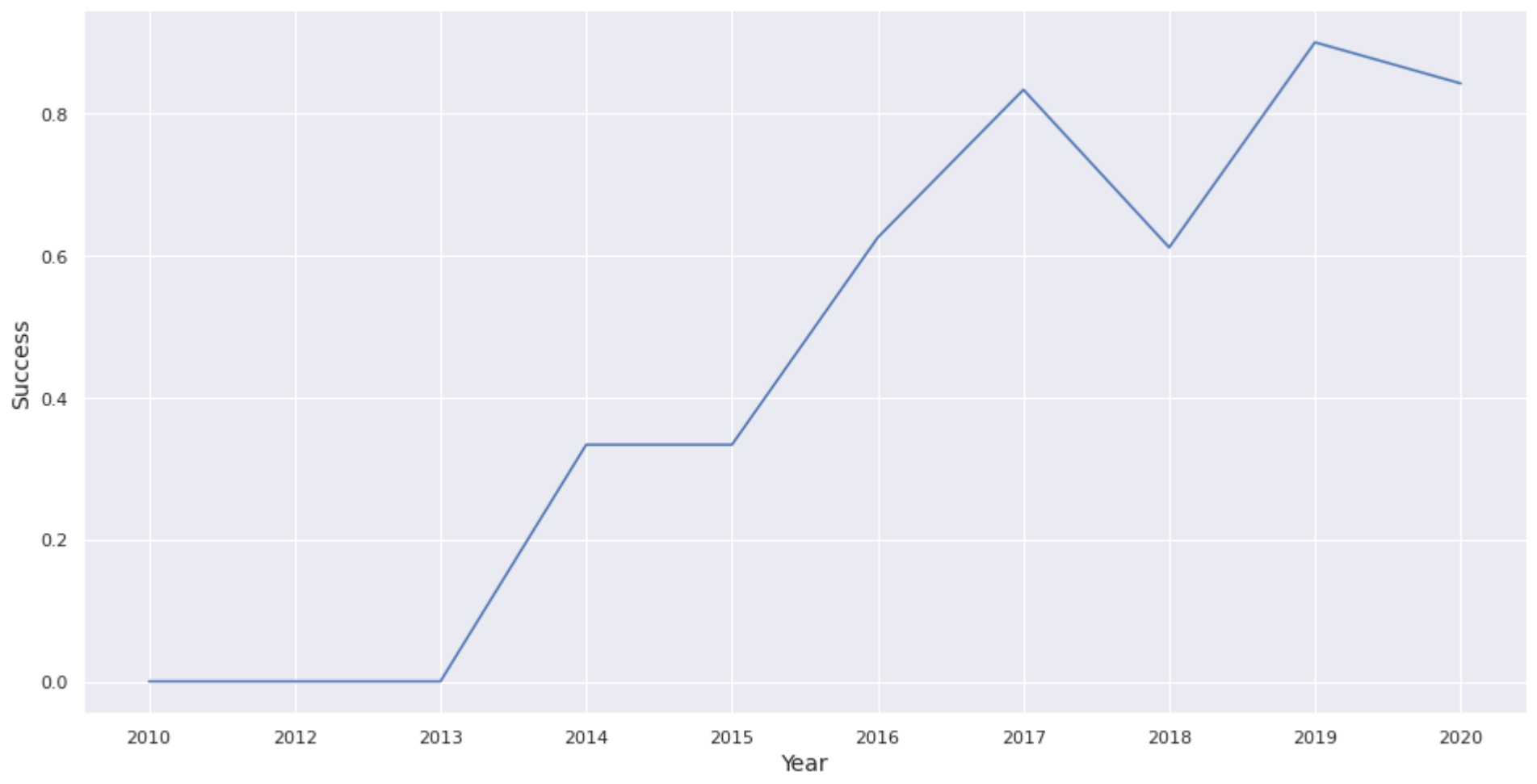
Flight Number vs. Orbit Type



Payload vs. Orbit Type



Launch Success Yearly Trend



All Launch Site Names

In [34]:
```
%sql select DISTINCT(LAUNCH_SITE) from SPACEXTBL
```

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

Out[34]:

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

Launch Site Names Begin with 'CCA'

In [26]:
```
%sql select * from SPACEXTBL where Launch_Site like 'CCA%' LIMIT 5
```

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

Out[26]:

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

Total Payload Mass

```
In [27]: %sql select SUM(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer = 'NASA (CRS)'
 * sqlite:///my_data1.db
Done.
Out[27]: SUM(PAYLOAD_MASS__KG_)
                          45596
```

Average Payload Mass by F9 v1.1

In [28]:
```
%sql select AVG(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version = 'F9 v1.1'
```

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

Out[28]:

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

In [42]:
```
%sql select MIN(Date) from SPACEXTBL where "Landing _Outcome" like '%ground pad%' and Mission_Outcome = 'Success'
```

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

Out[42]:

MIN(Date)
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

In [51]:
```
%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[51]:

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

Total Number of Successful and Failure Mission Outcomes

In [48]:
```
%sql select Mission_Outcome, count(Mission_Outcome) from SPACEXTBL group by Mission_Outcome
```

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

Out[48]:

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

Boosters Carried Maximum Payload

In [53]:
```
%sql select Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

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

Out[53]:

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

In [56]:
```
%sql SELECT substr(Date, 4, 2), Mission_Outcome, Booster_Version, Launch_Site FROM SPACEXTBL where substr(Date,7,4) = '2015'
```

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

Out[56]:

substr(Date, 4, 2)	Mission_Outcome	Booster_Version	Launch_Site
01	Success	F9 v1.1 B1012	CCAFS LC-40
02	Success	F9 v1.1 B1013	CCAFS LC-40
03	Success	F9 v1.1 B1014	CCAFS LC-40
04	Success	F9 v1.1 B1015	CCAFS LC-40
04	Success	F9 v1.1 B1016	CCAFS LC-40
06	Failure (in flight)	F9 v1.1 B1018	CCAFS LC-40
12	Success	F9 FT B1019	CCAFS LC-40

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

In [60]:
```
%sql SELECT "Landing _Outcome", count("Landing _Outcome")  FROM SPACEXTBL WHERE DATE BETWEEN '04-06-2010' AND '20-03-2017' GROUP BY "Landing _Outcome"
```

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

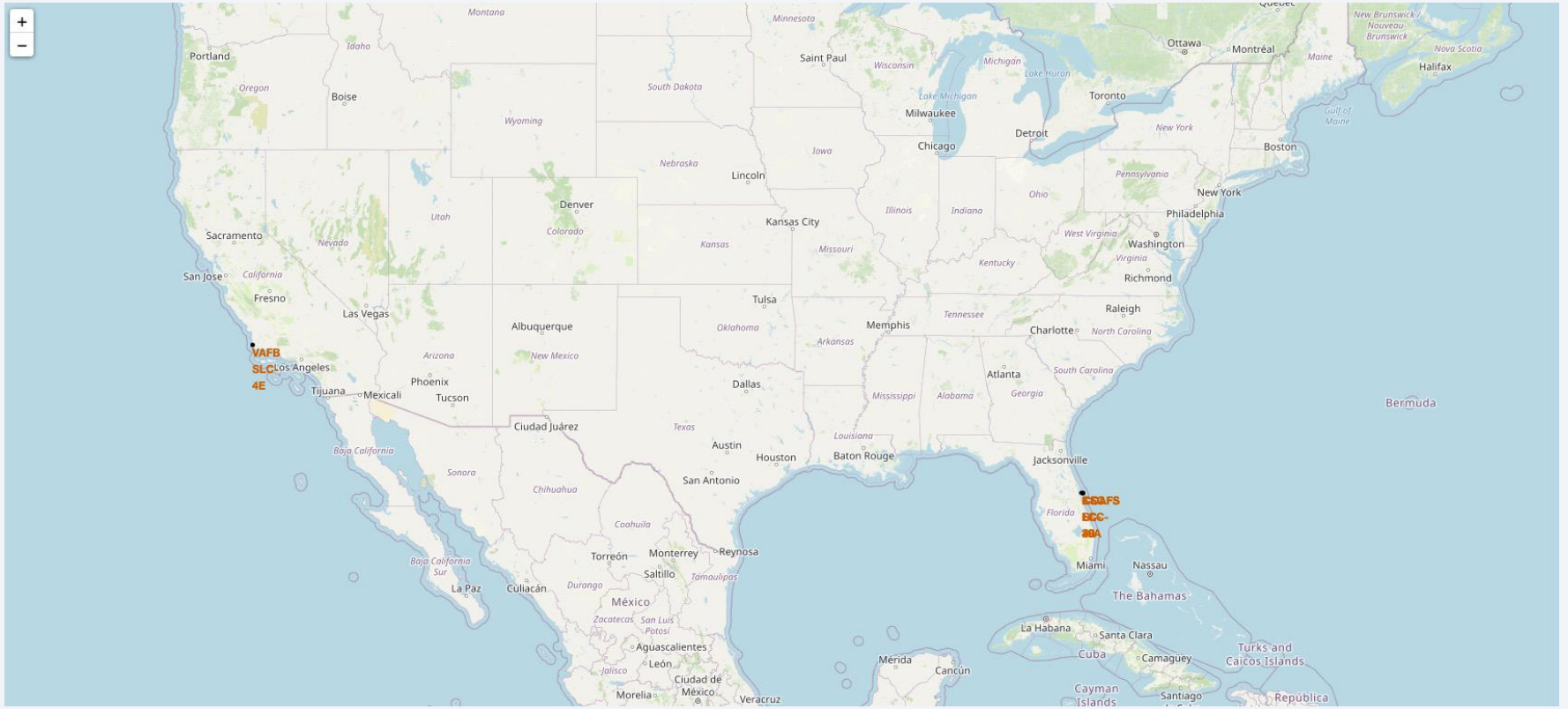
Out[60]:

Landing _Outcome	count("Landing _Outcome")
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

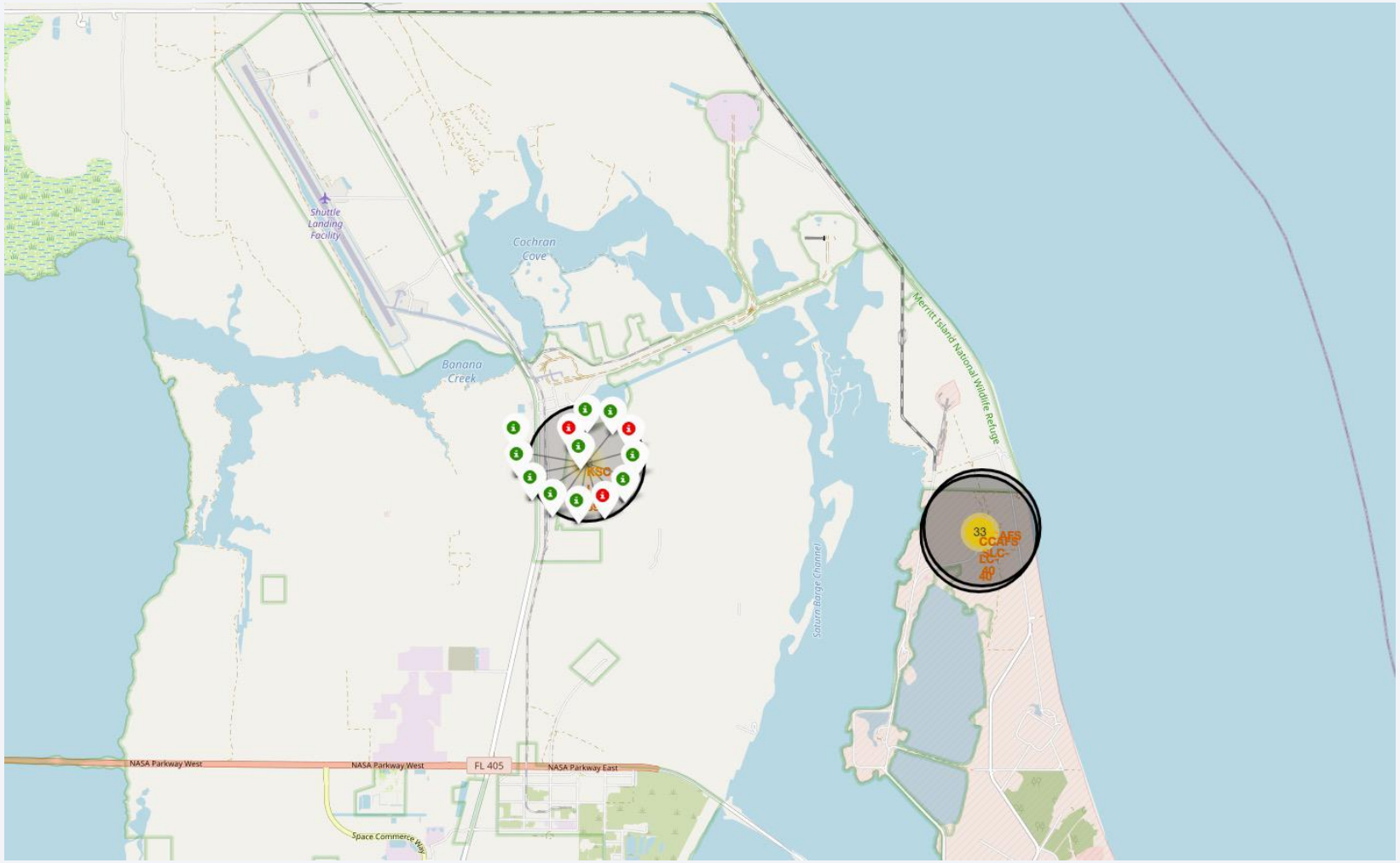
Section 3

Launch Sites Proximities Analysis

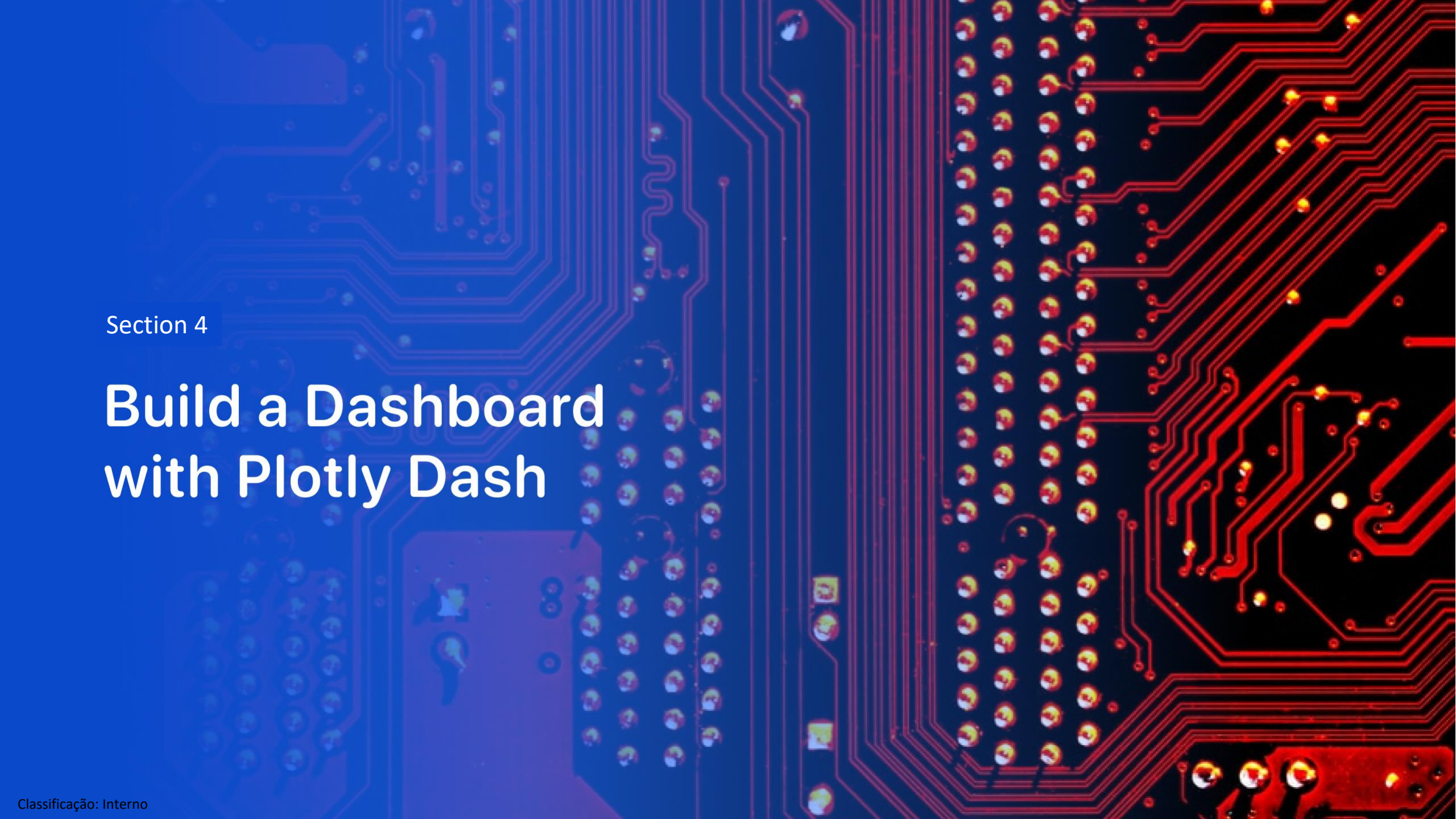
Location Sites



Lauch outcome



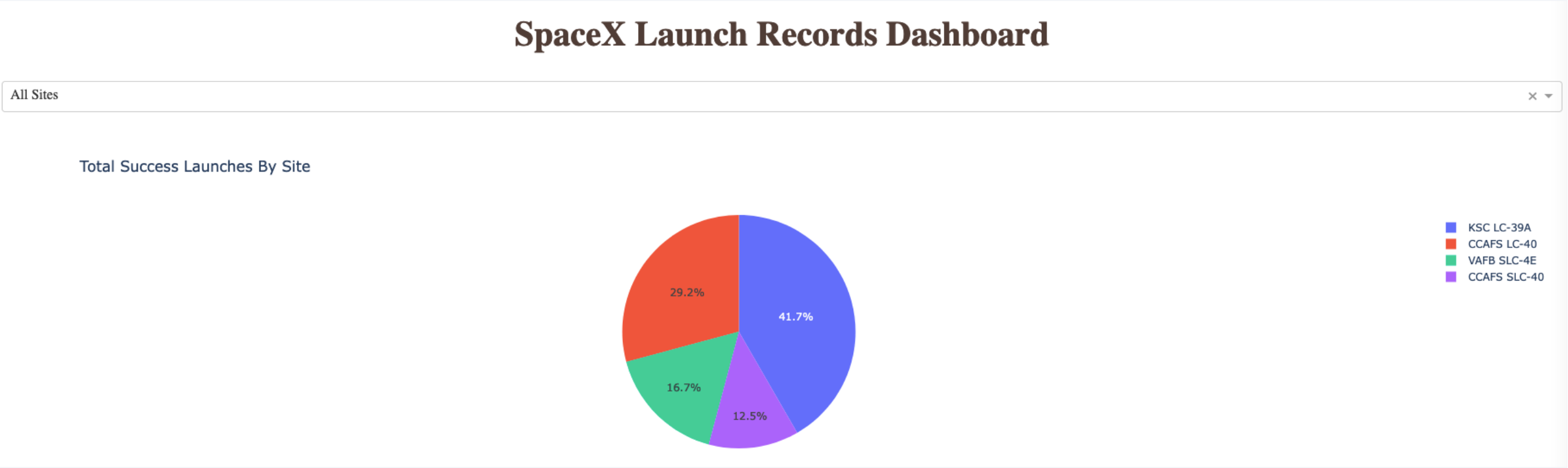
Infrastructure



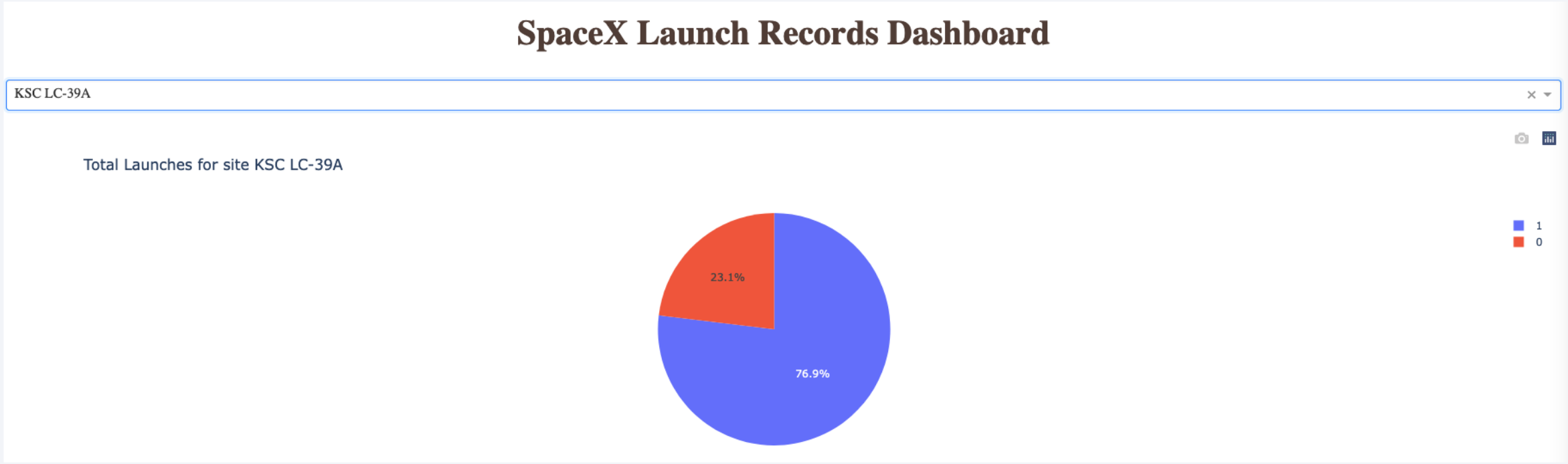
Section 4

Build a Dashboard with Plotly Dash

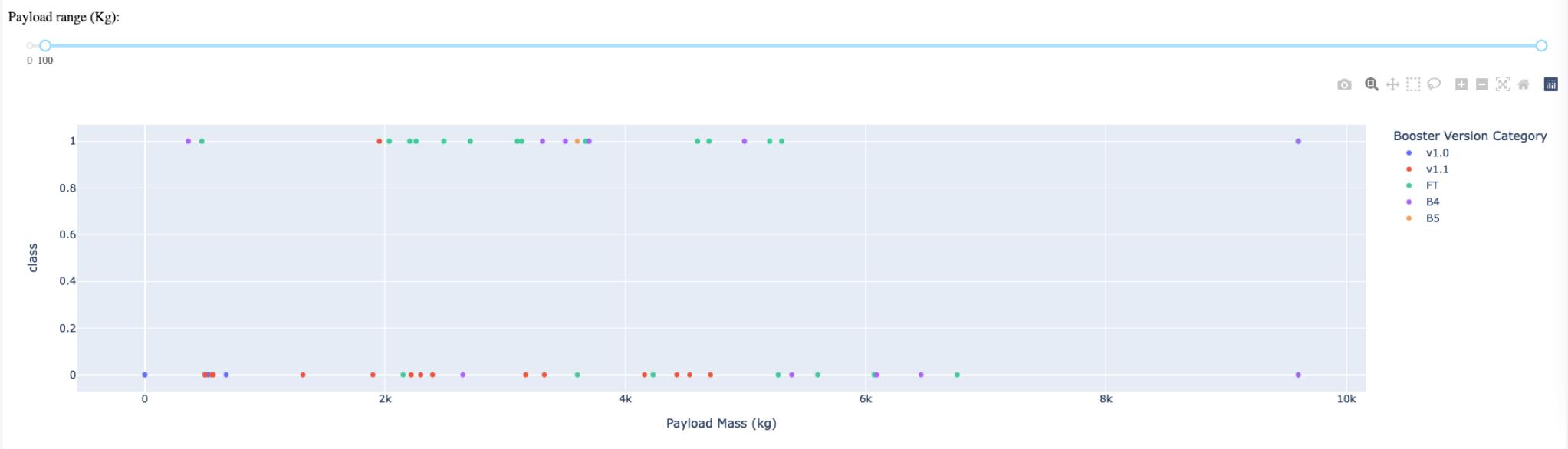
Successful Launches



Launch Success Ratio of KSC LC-39A



Payload vs. Launch Outcome



Section 5

Predictive Analysis (Classification)

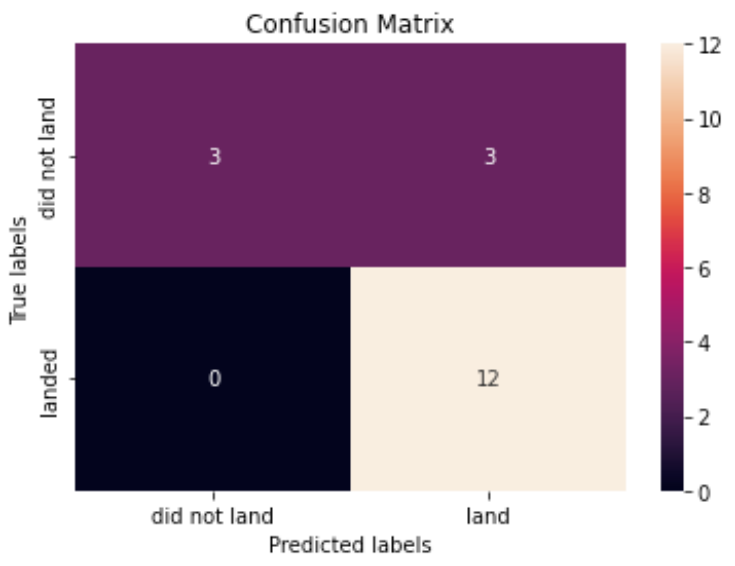
Classification Accuracy

- Train and test accuracy from all four classification models tested
- The model with the highest train accuracy (around 87,5%) is the Decision Tree Classifier

	Model	Train	Test
0	LogReg	0.847222	0.833333
1	SVM	0.847222	0.833333
2	DecTree	0.875	0.666667
3	KNN	0.847222	0.833333

Confusion Matrix

- Confusion matrix shows that the Decision Tree model can correctly predict all successful landed launch, but have a median performance to correctly predict failed launches



Conclusions

- The best launch site is KSC LC-39A with 76.9% of success;
- Launches above 7,000kg are tend to success;
- Successful landing outcomes seem to improve over time;
- Decision Tree Classifier can be used to predict successful landings and increase reduce the costs of the company.

Thank you!