IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Reda Gerges Ayad
February 11, 2022

Outline

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

Executive Summary

Nowadays we became interested in space, and even many of our services are relevant to Earth orbit such as satellites, hence many companies working on rockets which used to launch such satellites. A rocket launch costs upward of 165 million dollars however, SpaceX advertises that a rocket launch could cost 62 million dollars that is because SpaceX can reuse the first stage of the rocket. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch.

We will collect and use data related to different rocket launches to be provided to an appropriate machine-learning algorithm to predict if the first stage of the rocket, which is about to launch, will land successfully to be used again in another launch. Therefore, we can predict the cost of a launch.

Introduction

In order to predict if the launch will successfully land then the cost of the launch, we will begin with collecting data using SpaceX API then data wrangling, cleaning, EDA analysis, data visualization, and predictive modeling to predict the land of Falcon 9 rocket.

There will some problems to be solved such as data cleaning, features engineering and selection to get most accurate model.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - We will collect data using SpaceX API
- Perform data wrangling
  - First, we will get intuition about the data we have and specify our label for the model
  - Perform encoding for categorical data to get full numeric dataset
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - We will split data into train and test dataset and apply different classification models and test their accuracy on test dataset and choose the most accurate one.

Data Collection

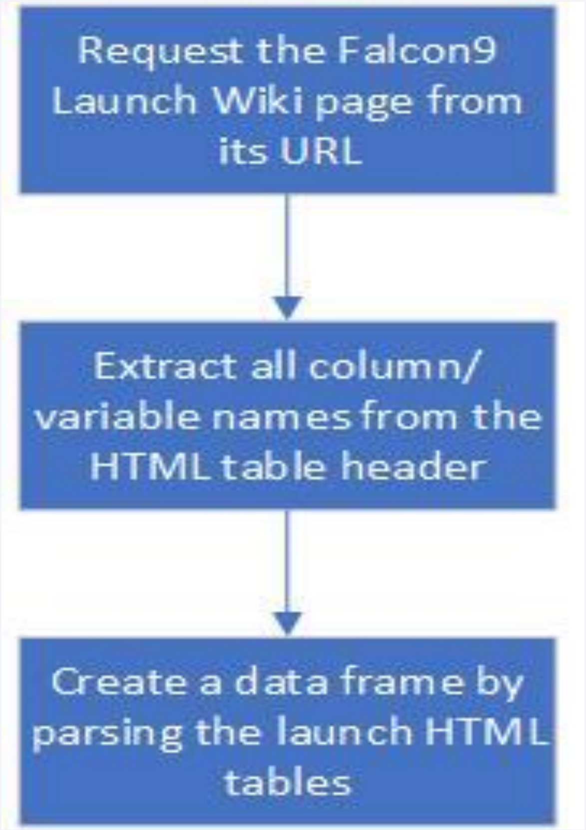
In data collection process, we will collect desired data by tow different ways. The first way by using SpaceX API, and then filtering data for some specific features. The second one by using web scraping for a Wikipedia page and scraping tables we want to work on.

Data Collection – SpaceX API

- Data collection with REST keys using SpaceX API
- This link direct to a GitHub notebook of applying SpaceX API calls

Data Collection - Scraping

- web scraping process steps are shown in the flowchart
- This GitHub URL of the completed web scraping notebook



Data Wrangling

In data wrangling process, we will perform some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.

In the data set, there are several different cases where the booster did not land successfully like 'False Ocean', and 'False RTLS', there are also several different cases for landing successfully such as 'True Ocean', and 'True ASDS' so in this process we will convert those outcomes into training labels with 1 means the booster successfully landed and 0 means it was unsuccessful.

- This GitHub URL of the completed data wrangling process.

EDA with Data Visualization

- In this step we will perform EDA and feature engineering using Pandas, Matplotlib, and Seaborn.
- We will plot different types of charts because we will represent different ideas. We will use scatter plots to explain relationships among columns, bar charts and line plots to show insights in any individual column
- This GitHub URL of the completed EDA with data visualization notebook.

EDA with SQL

In EDA process we will perform queries :

- ➢ select distinct launch_site from spacexdataset ;
- ➢ Select avg(payload_mass_kg_) from spacexdataset where booster_version like ‘F9 v1.1%’ ;
- ➢ select mission_outcome, count(*) from spacexdataset group by mission_outcome ;
- ➢ Select landing_outcome, count(*) as landing_count from spacexdataset where date between ‘20100604’ and ‘20170320’ group by landing_outcome order by landing_count desc;

- This GitHub URL of the completed EDA with SQL notebook.

Build an Interactive Map with Folium

In this section we will do interactive Folium map contains the following:

- Marks of all launch sites
- Marks of the success/failed launches for each site
- Draw PolyLines between launch sites to different selected points

All of that is done to make more interactive visual analytics

This GitHub URL of the completed interactive map with Folium map.

Build a Dashboard with Plotly Dash

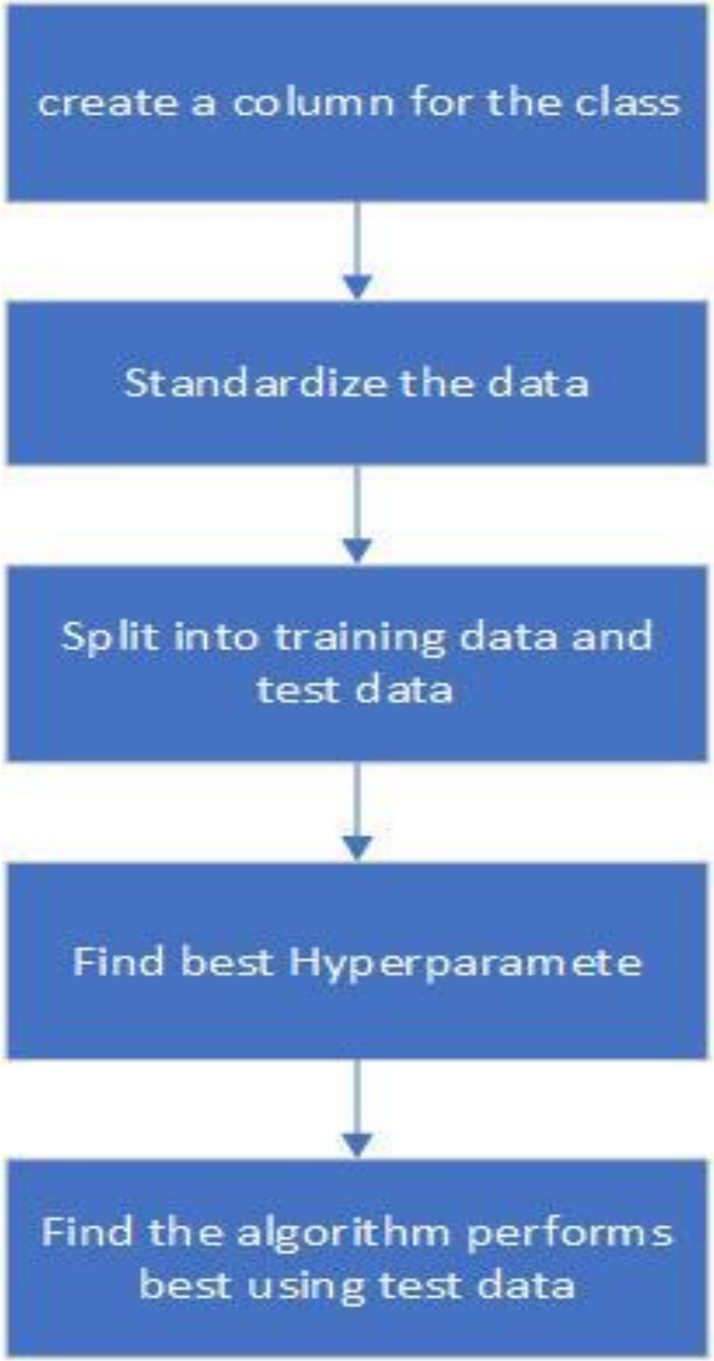
- We will build a dashboard with input components and interactive plots like a pie chart to show the success rate of each launch site, and a scatter plot with a range slider.
- To perform interactive visual analytics on Space launch data in real-time.
- This GitHub URL of the completed Plotly Dash lab.

Predictive Analysis (Classification)

We will convert the categorical data into numeric using get_dummies() method, standardize the dataset and the label column we processed in the data wrangling process, then we will split the dataset, using train_test_split() method, into train part to train different classification algorithms (Logistic Regression, SVM, KNN, and Decision Tree) and tuning hyper parameters of each one using GridSeachCV() and finally we evaluate them using test part to found the best performing classification model.

- This GitHub URL of the completed predictive analysis lab.

Predictive Analysis Flowchart

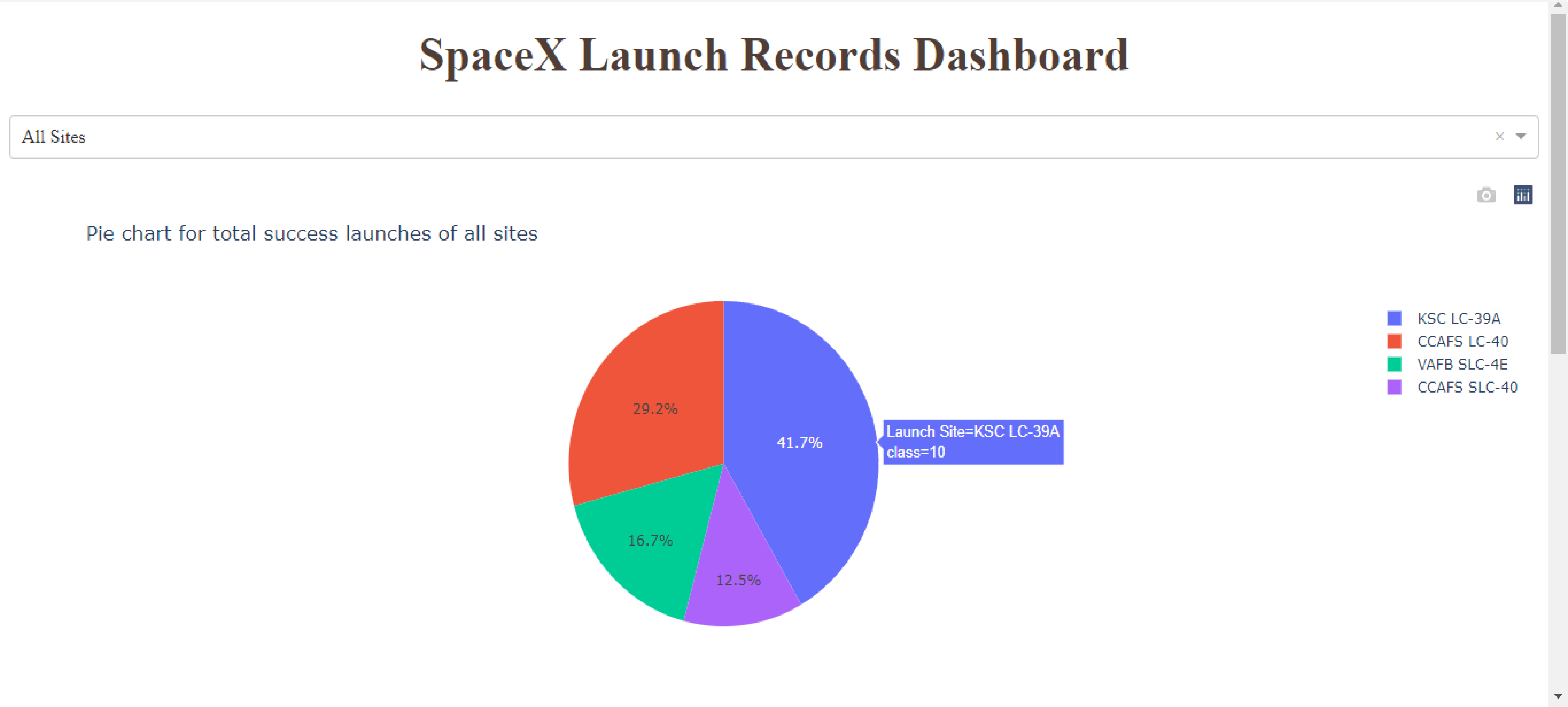


Results

Exploratory data analysis results:

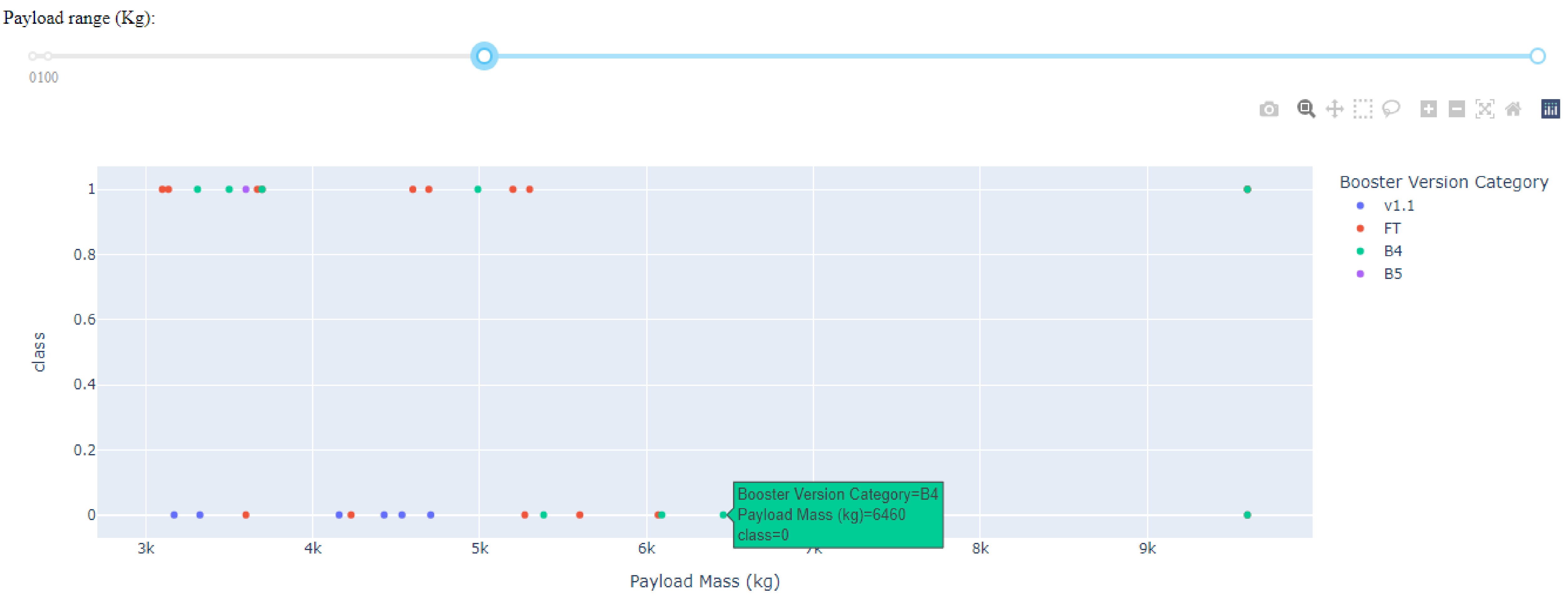
- We got the dataset from SpaceX API or web scraping
- Intuition about dataset
- We got the data cleaned
- Data standardized
- Create the label column

Results



SpaceX Launch Records Dashboard
All Sites
Pie chart for total success launches of all sites
KSC LC-39A
CCAFS LC-40
VAFB SLC-4E
CCAFS SLC-40
29.2%
41.7%
16.7%
12.5%
Launch Site=KSC LC-39A
class=10

Results


Payload range (Kg):
0100
1
0.8
0.6
0.4
0.2
0
class
3k
4k
5k
6k
8k
9k
Payload Mass (kg)
Booster Version Category
v1.1
FT
B4
B5
Booster Version Category=B4
Payload Mass (kg)=6460
class=0

Results

After evaluating the different algorithms on the test data, we got the accuracy for each model as follows:

Find the method performs best:

```
print('Logistic Regression test accuracy is {}'.format(logreg_cv.best_estimator_.score(X_test, Y_test)))
print('SVM test accuracy is {}'.format(svm_cv.best_estimator_.score(X_test, Y_test)))
print('KNN test accuracy is {}'.format(knn_cv.best_estimator_.score(X_test, Y_test)))
print('Decision tree test accuracy is {}'.format(tree_cv.best_estimator_.score(X_test, Y_test)))
```

```
Logistic Regression test accuracy is 0.8333333333333334
SVM test accuracy is 0.8333333333333334
KNN test accuracy is 0.8333333333333334
Decision tree test accuracy is 0.8888888888888888
```

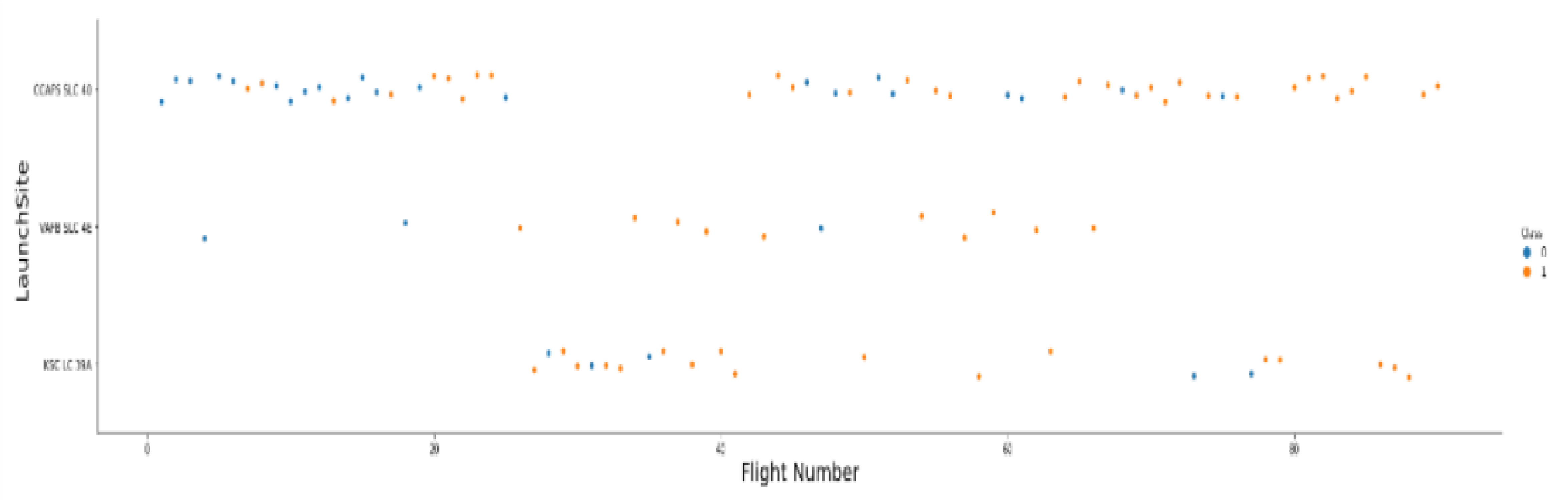
As show above the best performing method is Decision Tree with accuracy of 89%

Section 2

Insights drawn from EDA

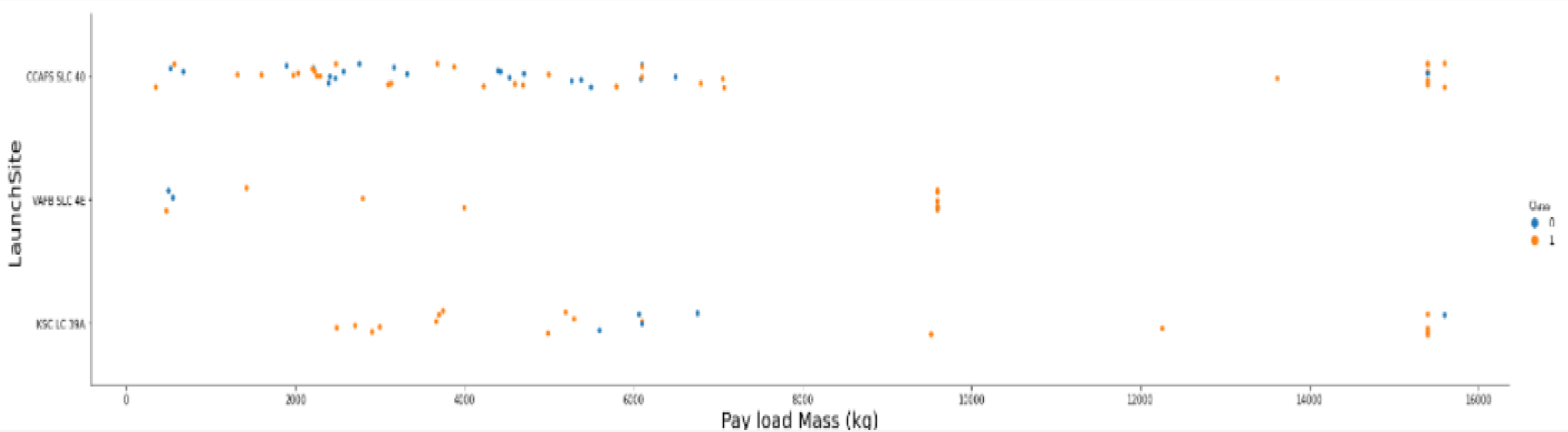
Flight Number vs. Launch Site

- Here is the relationship between flight number and launch site in scatter plot with each point colored to indicate the class of the landing 1 for success and 0 for fail.



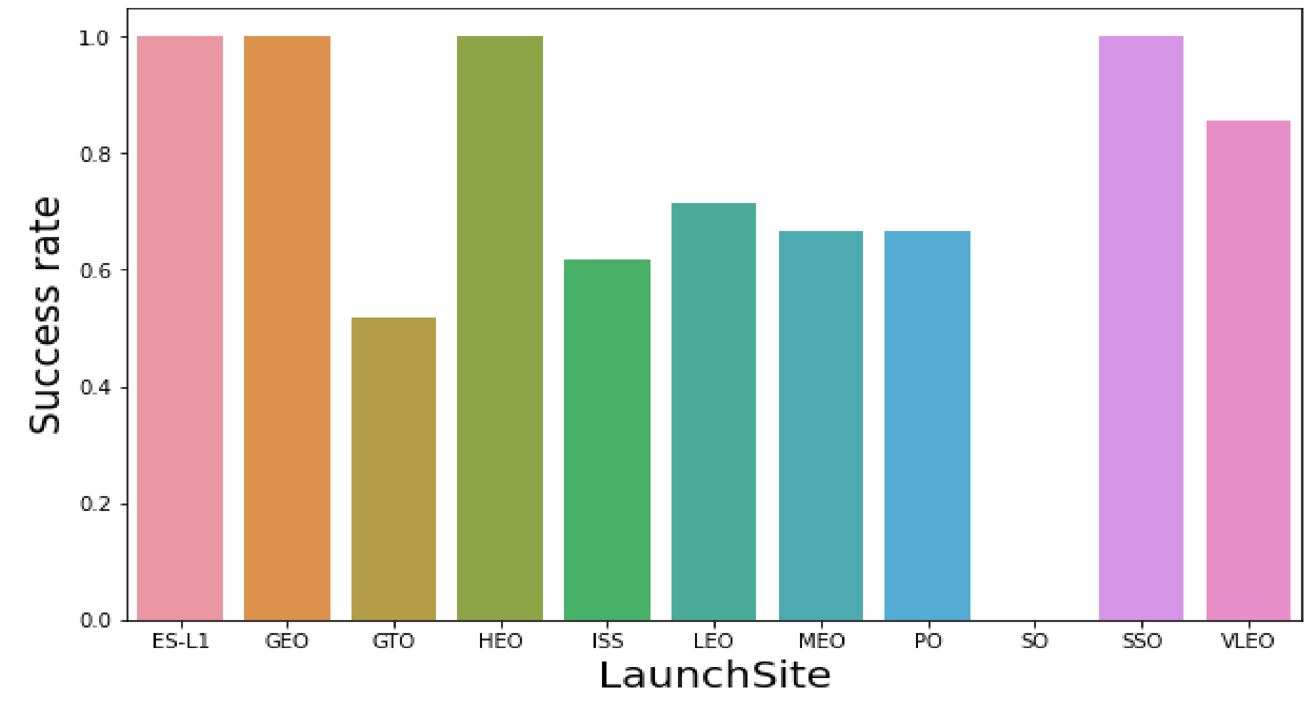
Payload vs. Launch Site

- A scatter plot for the relationship between the payload of the rocket and the launch site and showing if the launch landed successfully.



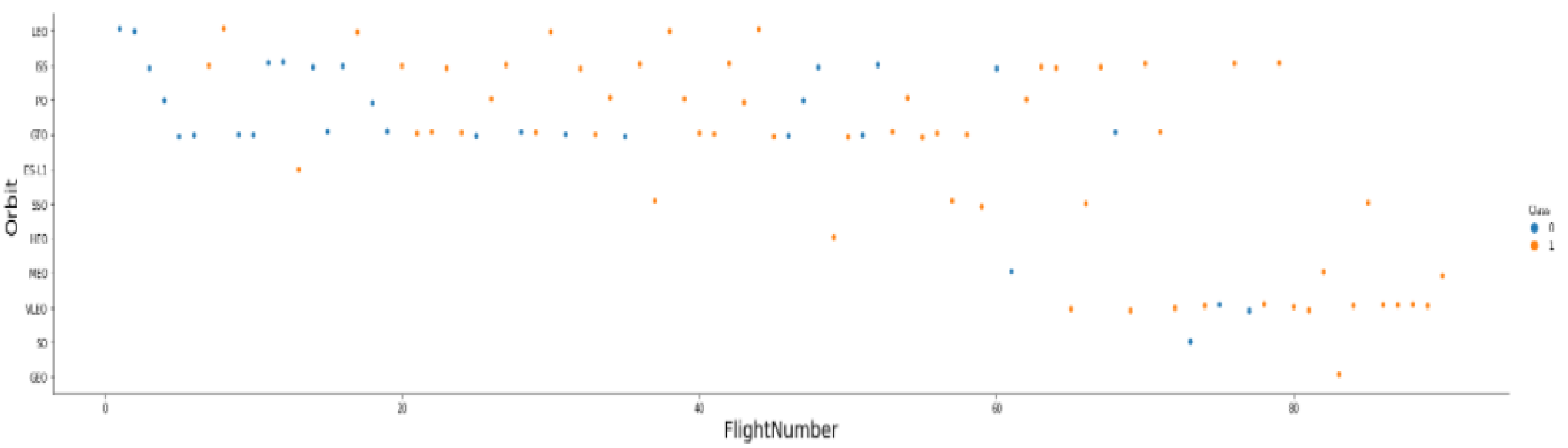
Success Rate vs. Orbit Type

- A bar chart for the success rate of each orbit type



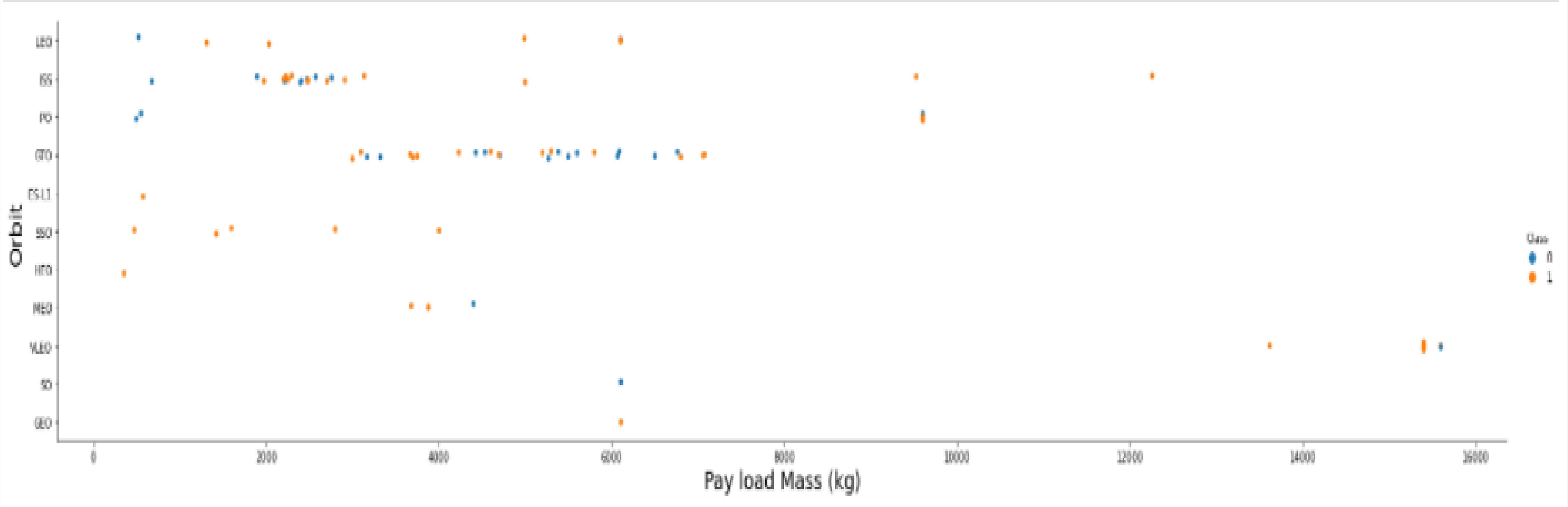
Flight Number vs. Orbit Type

- A scatter point of Flight number vs. Orbit type and showing if the landing was successful.



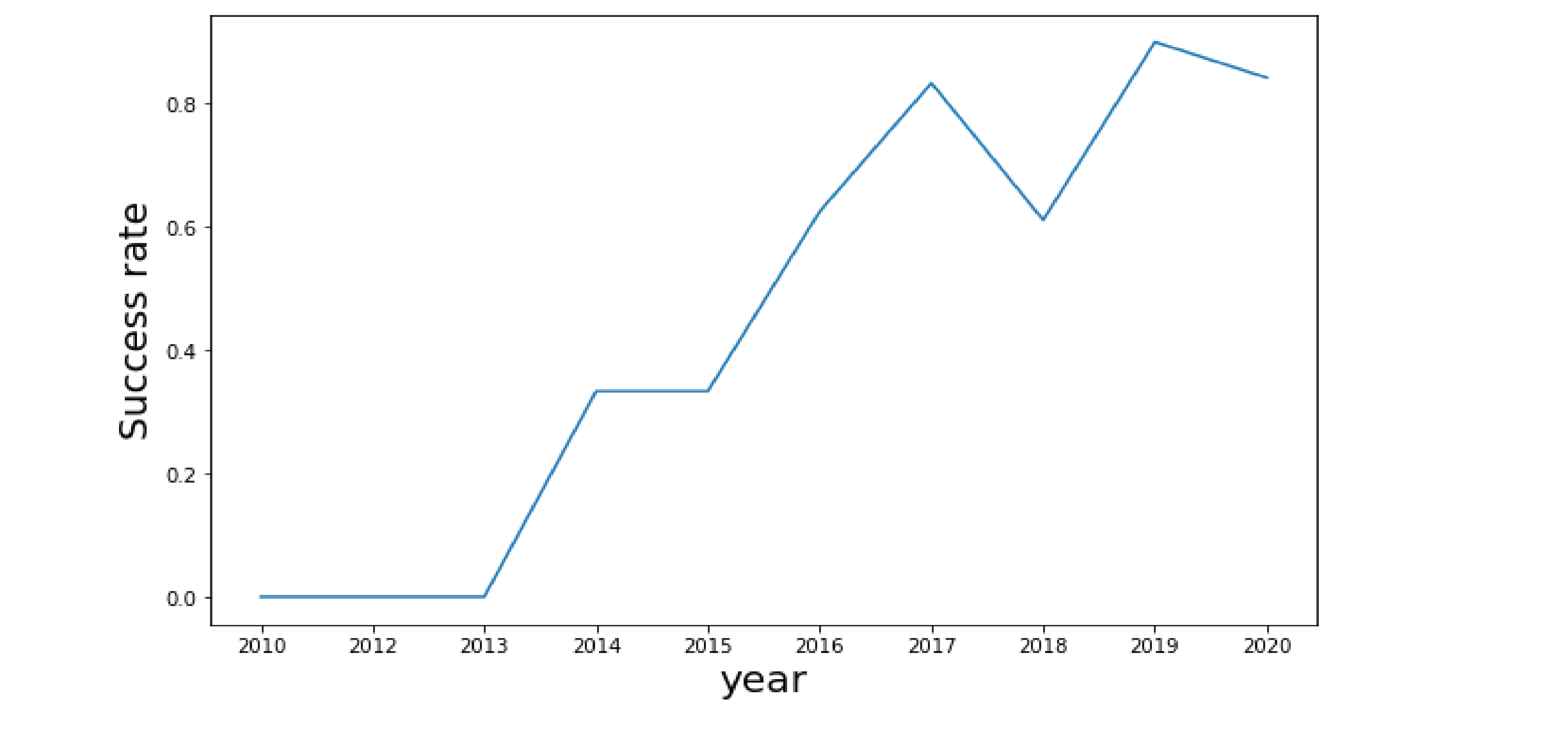
Payload vs. Orbit Type

- A scatter point of payload vs. orbit type



Launch Success Yearly Trend

- A line chart of yearly average success rate



All Launch Site Names

The names of the unique launch sites:

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

```
%sql select unique(launch_site) from spacexdataset;
```

```
 * ibm_db_sa://qjn14301:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`:

```
%%sql
select * from spacexdataset
where launch_site like 'CCA%'
limit 5
```

```
 * ibm_db_sa://qjn14301:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/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 carried by boosters from NASA

```
%%sql
select sum(payload_mass__kg_) as total_mass from spacexdataset
where payload like '%CRS%'
```

```
 * ibm_db_sa://qjn14301:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

total_mass
111268

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1

```
%%sql
select avg(payload_mass__kg_) from spacexdataset
where booster_version like 'F9 v1.1%'
```

```
 * ibm_db_sa://qjn14301:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

1
2534

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad

```
%%sql
select min(date) as first_succe
where landing__outcome ='Success (ground pad)'
```

noun: نجاح, فوز, توفيق, ظفر, عمل ناجح...
translated from: English

```
 * ibm_db_sa://qjn14301:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

first_successful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
%%sql
select booster_version from spacexdataset
where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000
```

```
 * ibm_db_sa://qjn14301:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

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 total number of successful and failure mission outcomes

```
%%sql
select mission_outcome, count(*) from spacexdataset
group by mission_outcome
```

```
 * ibm_db_sa://qjn14301:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

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

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass

```
%%sql
select booster_version from spacexdataset
where payload_mass__kg_ = (select max(payload_mass__kg_) from spacexdataset)
```

```
 * ibm_db_sa://qjn14301:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
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

2015 Launch Records

- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%%sql
select booster_version, launch_site from spacexdataset
where landing__outcome = 'Failure (drone ship)' and date like '2015%'
```

```
 * ibm_db_sa://qjn14301:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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 date 2010-06-04 and 2017-03-20, in descending order

```
%%sql
select landing__outcome, count(*) as landing_count from spacexdataset
where date between '20100604' and '20170320'
group by landing__outcome
order by landing_count desc
```

```
 * ibm_db_sa://qjn14301:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

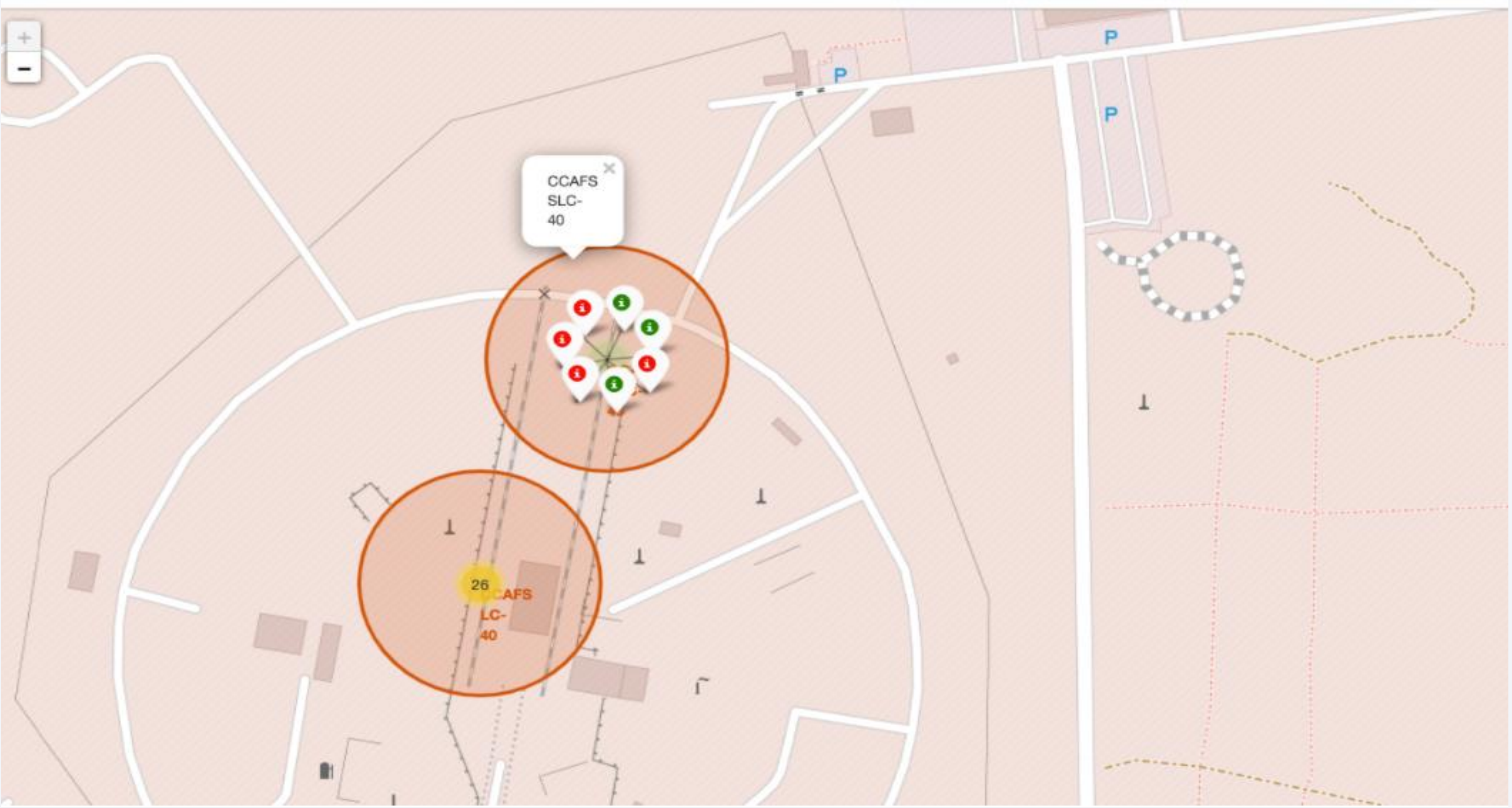
landing__outcome	landing_count
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

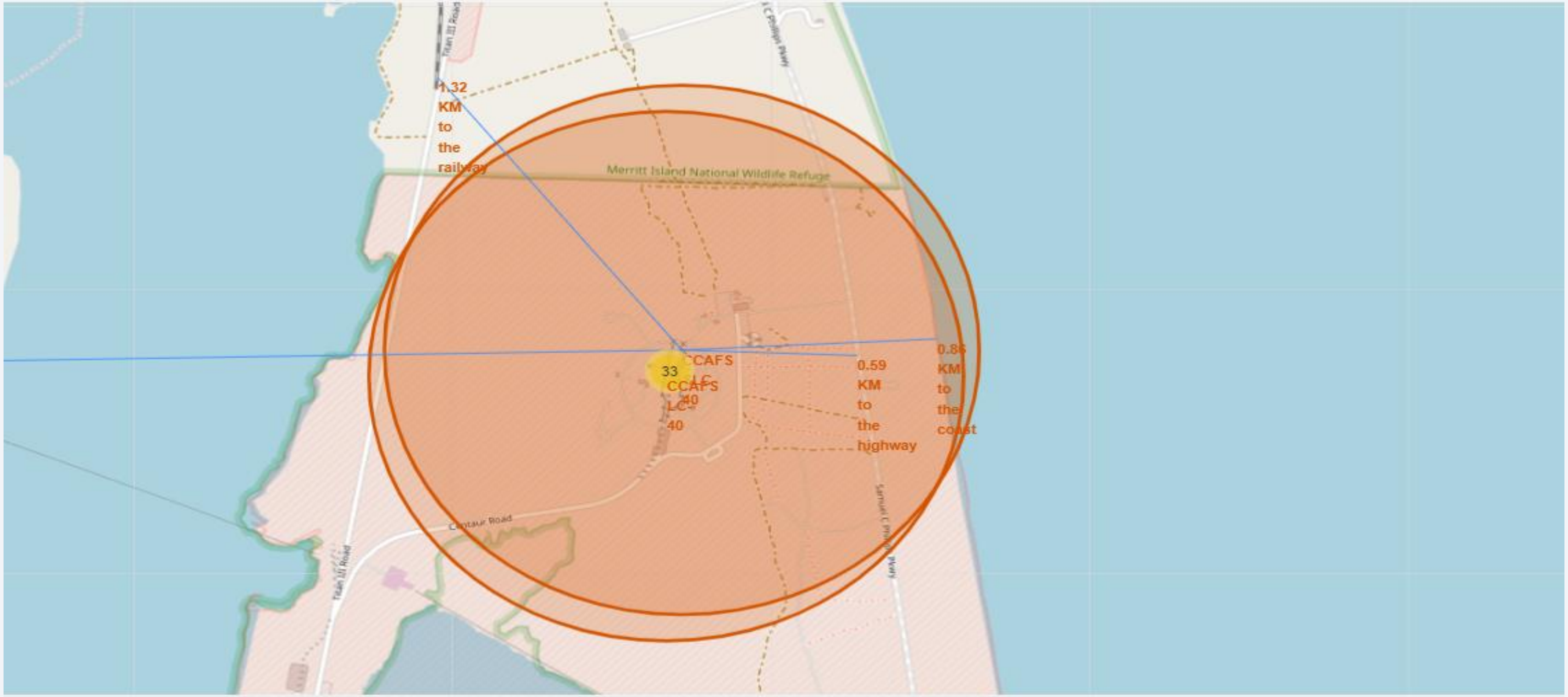
Launch Sites Proximities Analysis

Launch Sites Locations

The Color-labeled Launch Outcomes



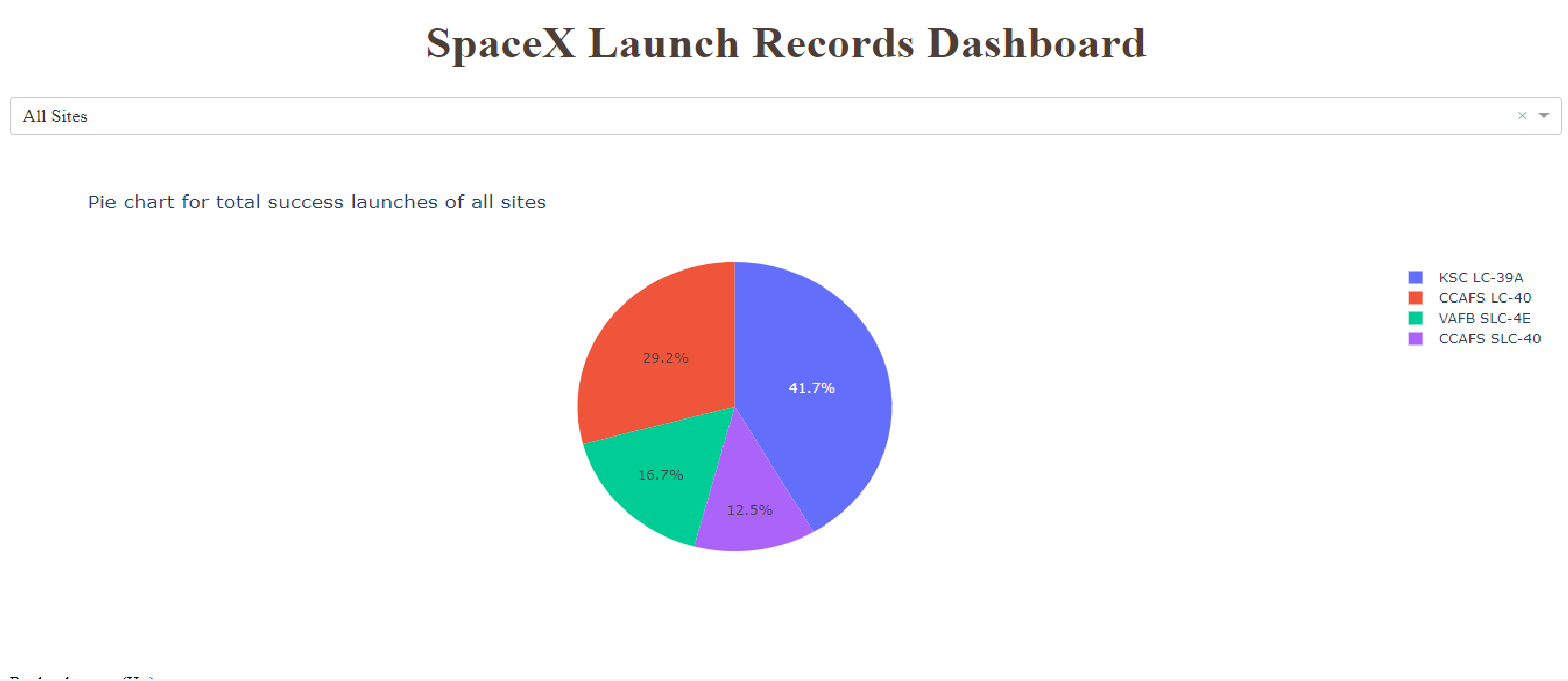
CCAFS SLC-40 Site to its Proximities



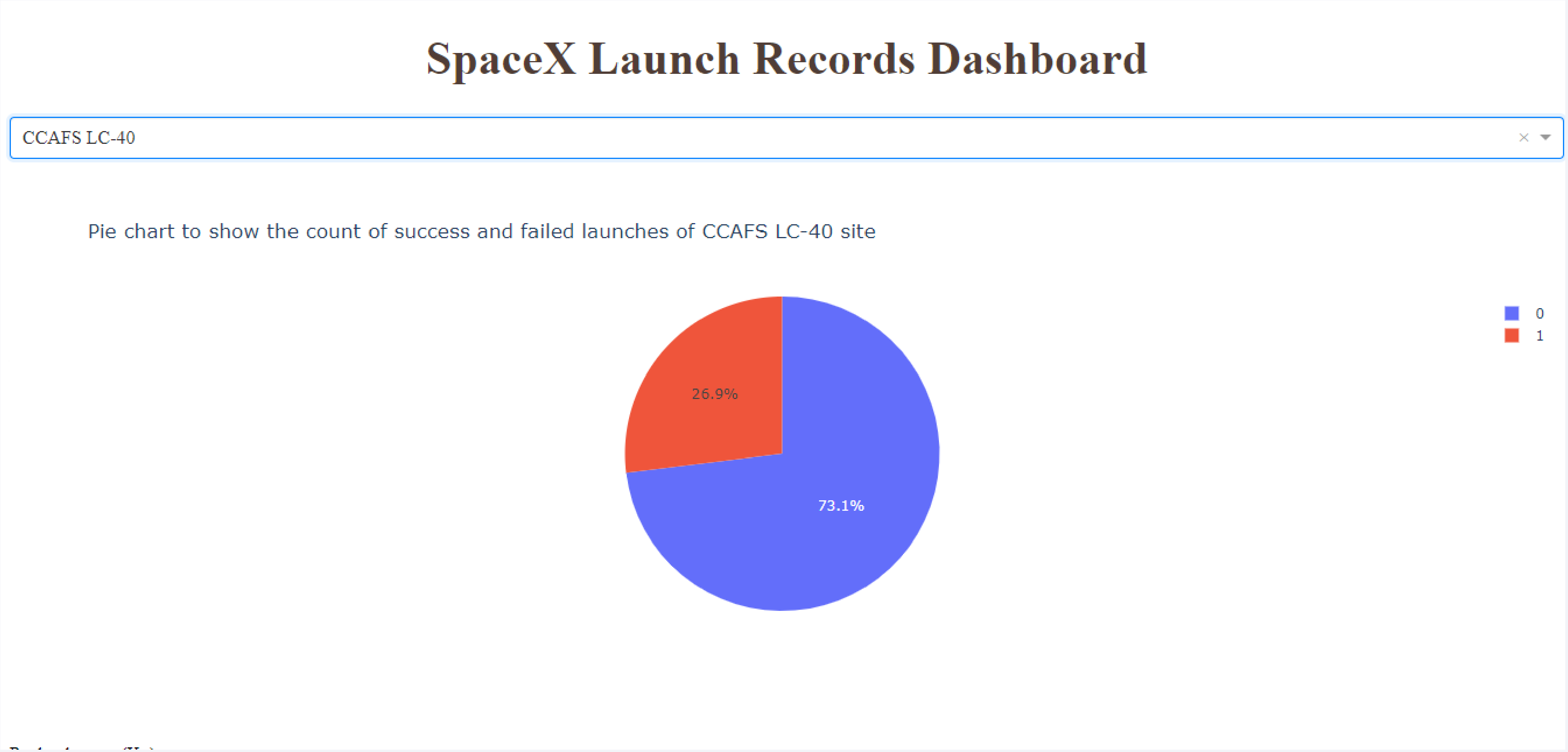
Section 4

Build a Dashboard with Plotly Dash

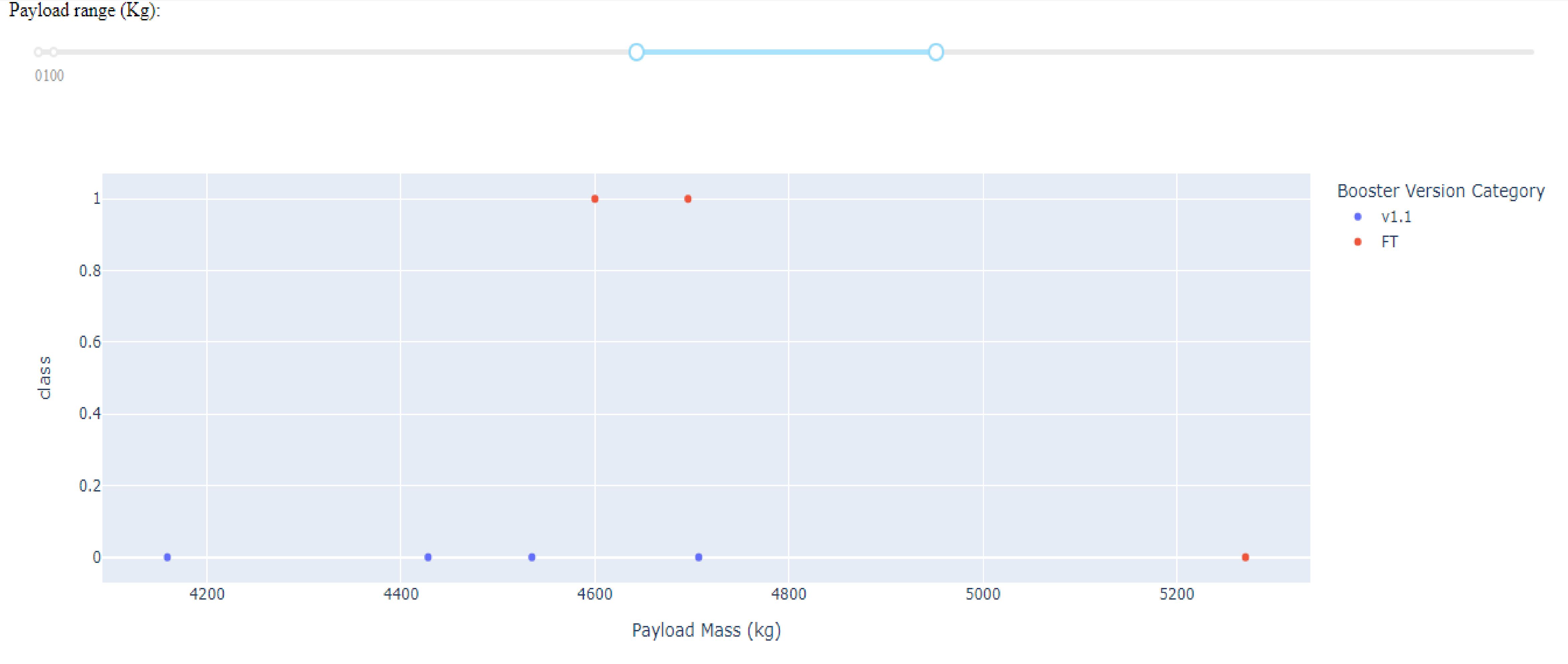
Launch Success Count for All Sites



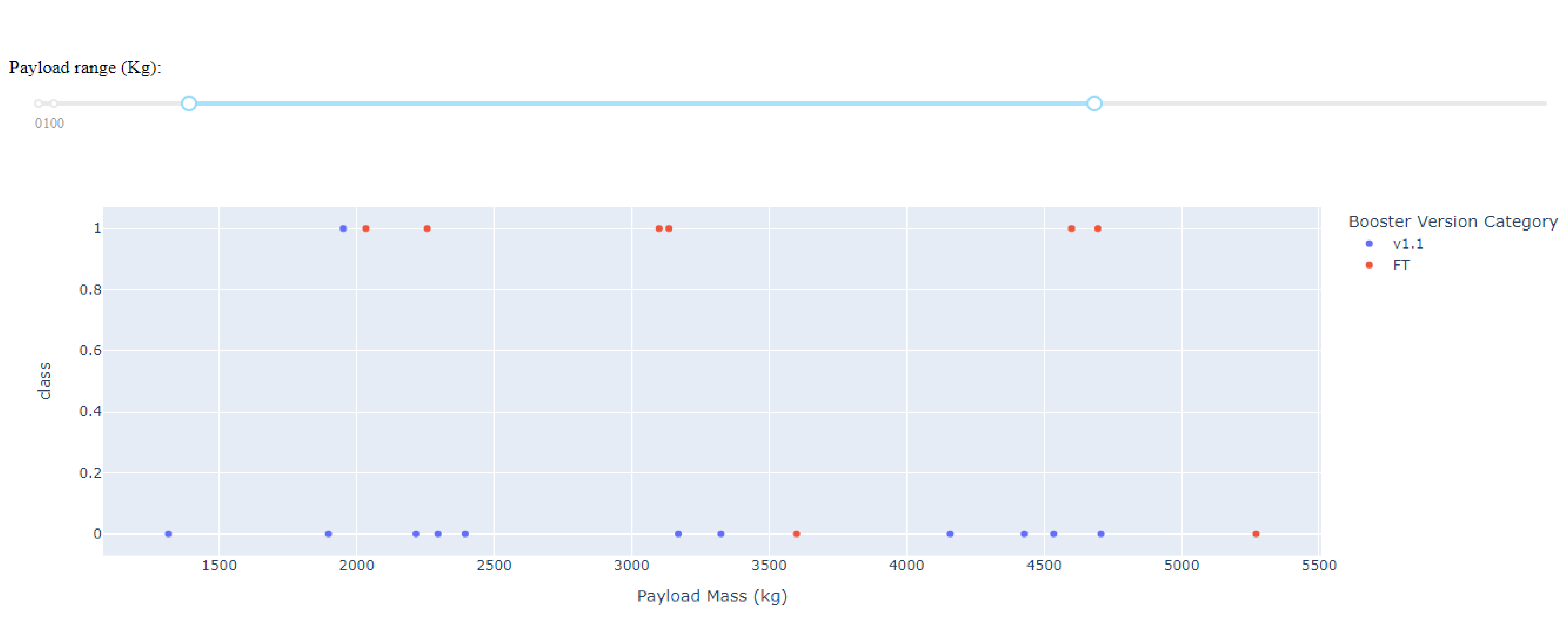
The Launch Site with Highest Success Rate



Payload vs. Launch Outcome on Dashboard



Payload vs. Launch Outcome on Dashboard

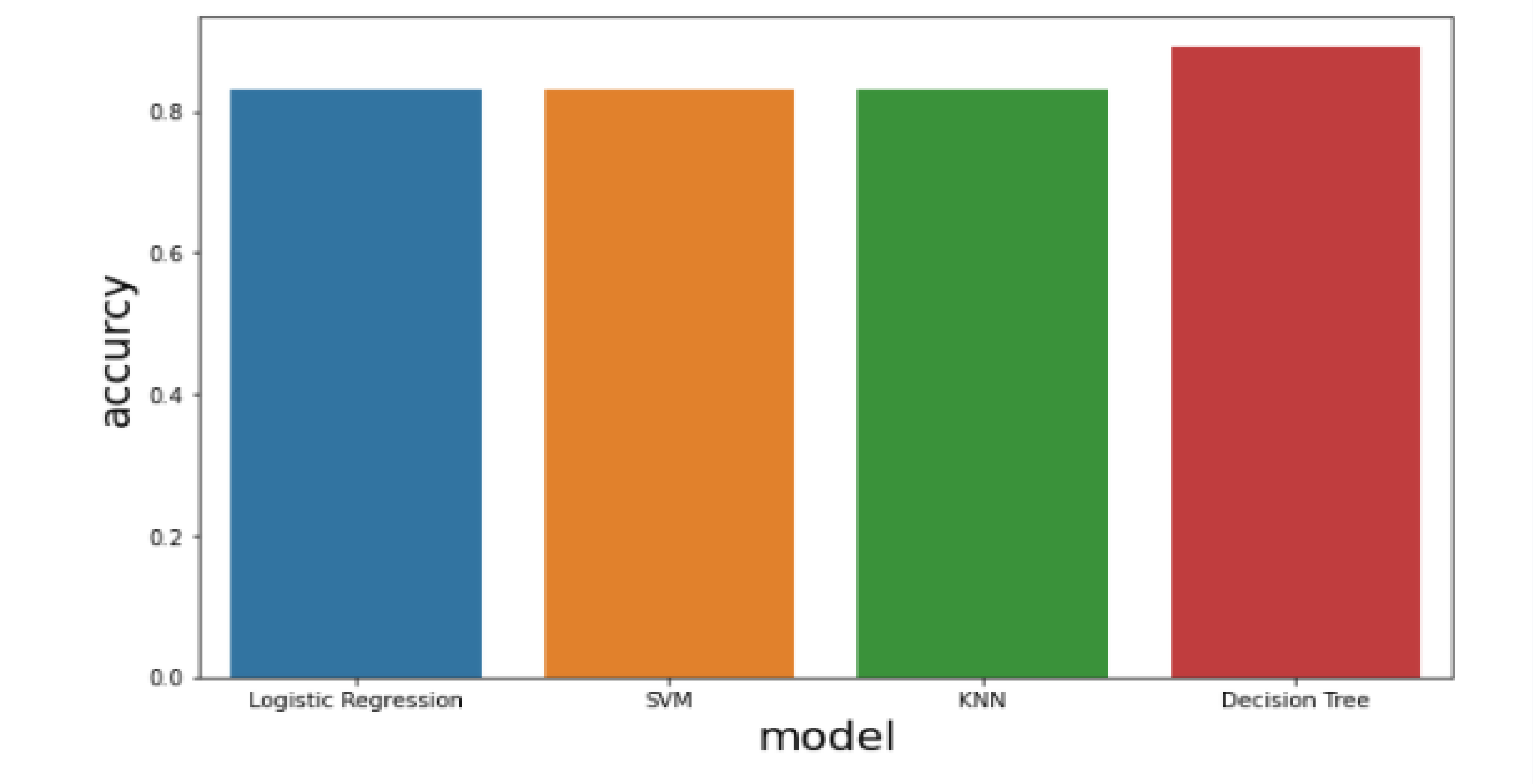


Section 5

Predictive Analysis (Classification)

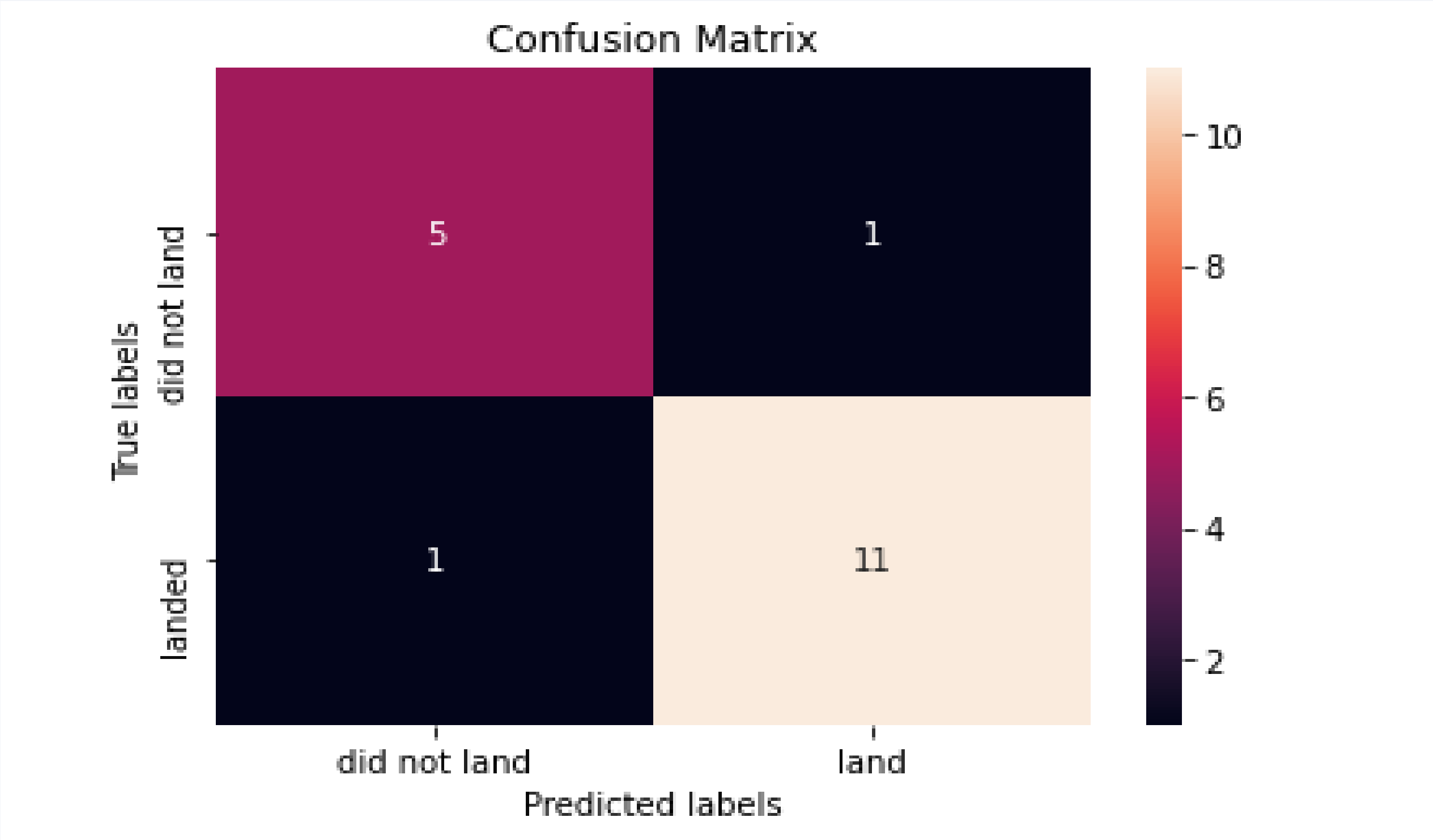
Classification Accuracy

- model accuracy for all built classification models, in a bar chart



Confusion Matrix

- The confusion matrix of the best performing model.



Conclusions

During this journey we passed with different process to achieve the of predicting if a rocket first stage will land successfully therefore, we can predict the cost of the launch.

- ✓ We collected data using tow different ways by using SpaceX API, and web scraping by BeautifulSoup
- ✓ EDA on the dataset, processing, standardizing, cleaning, and encoding categorical values.
- ✓ We create interactive analytics dashboard using Plotly.
- ✓ We visualized and collecting another information on maps and location of launch sites using Folium.
- ✓ We trained some different classification methods ('Logistic Regression', 'SVM', 'KNN', 'Decision Tree').
- ✓ Finally, we tested these methods on the test dataset and decided which method is the most accurate.

Appendix

- Dataset after collecting :

FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitud
1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.57736
2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.57736
3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.57736
4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.61082
5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.57736
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	.
86	2020-09-03	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	2	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	9	B1060	-80.60395
87	2020-10-06	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	3	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	9	B1058	-80.60395
88	2020-10-18	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	6	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	10	B1051	-80.60395
89	2020-10-24	Falcon 9	15600.0	VLEO	CCSFS SLC 40	True ASDS	3	True	True	True	5e9e3033383ecbb9e534e7cc	5.0	9	B1060	-80.57736
90	2020-11-05	Falcon 9	3681.0	MEO	CCSFS SLC 40	True ASDS	1	True	False	True	5e9e3032383ecb6bb234e7ca	5.0	3	B1062	-80.57736

ws × 17 columns

Thank you!