IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Gonzalo Sainz de Baranda Garrido
December 14, 2022

Outline

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

Executive Summary

Summary of methodologies:

- Collect data through SpaceX API
- Data Collection with Web Scraping
- Convert the data into a dataframe and then perform some data wrangling
- EDA with SQL
- EDA with visualization tools from the seaborn library
- Analyze the proximity of launch locations with Folium
- Analyze launch records interactively with Plotly Dash
- Machine Learning Prediction

Summary of all results

As a result we got several satisfactory models with an accuracy level of 83.3%.

Introduction

In this capstone, as a Data Scientist of the company Space Y, we will predict if the Falcon 9 first stage will land successfully. 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.

We are going to:

- Determine if the first stage will land

- Determine the cost of a launch

- Use this information for an alternate company that wants to bid against SpaceX

Section 1

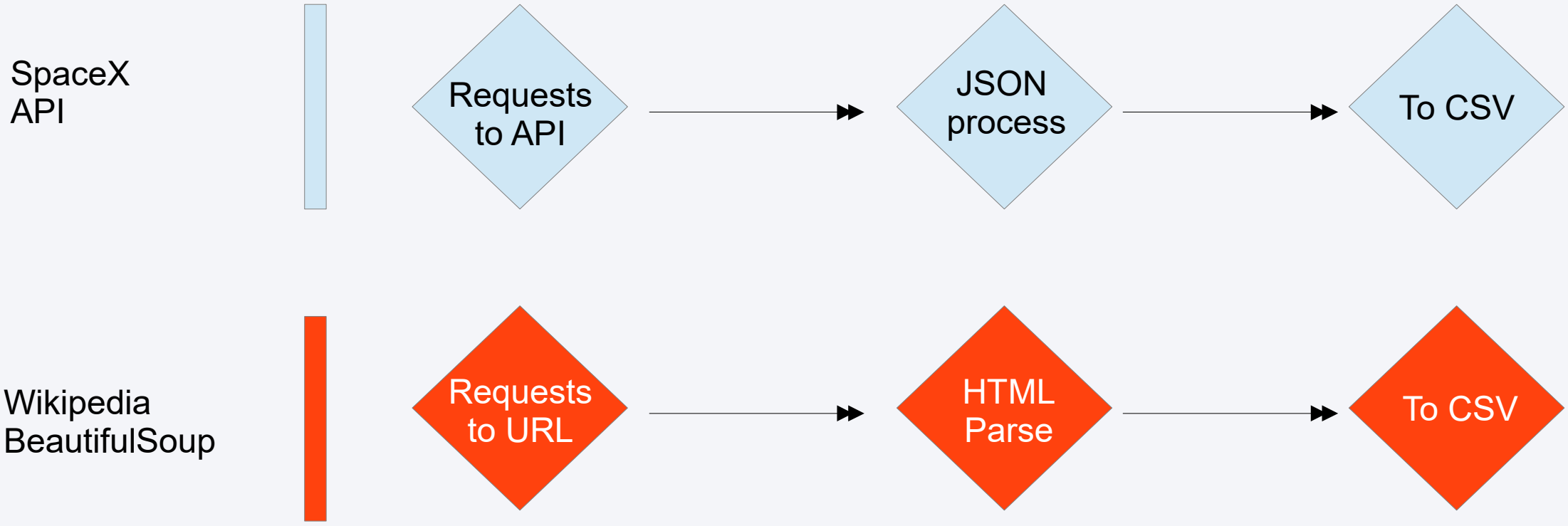
Methodology

Methodology

Executive Summary

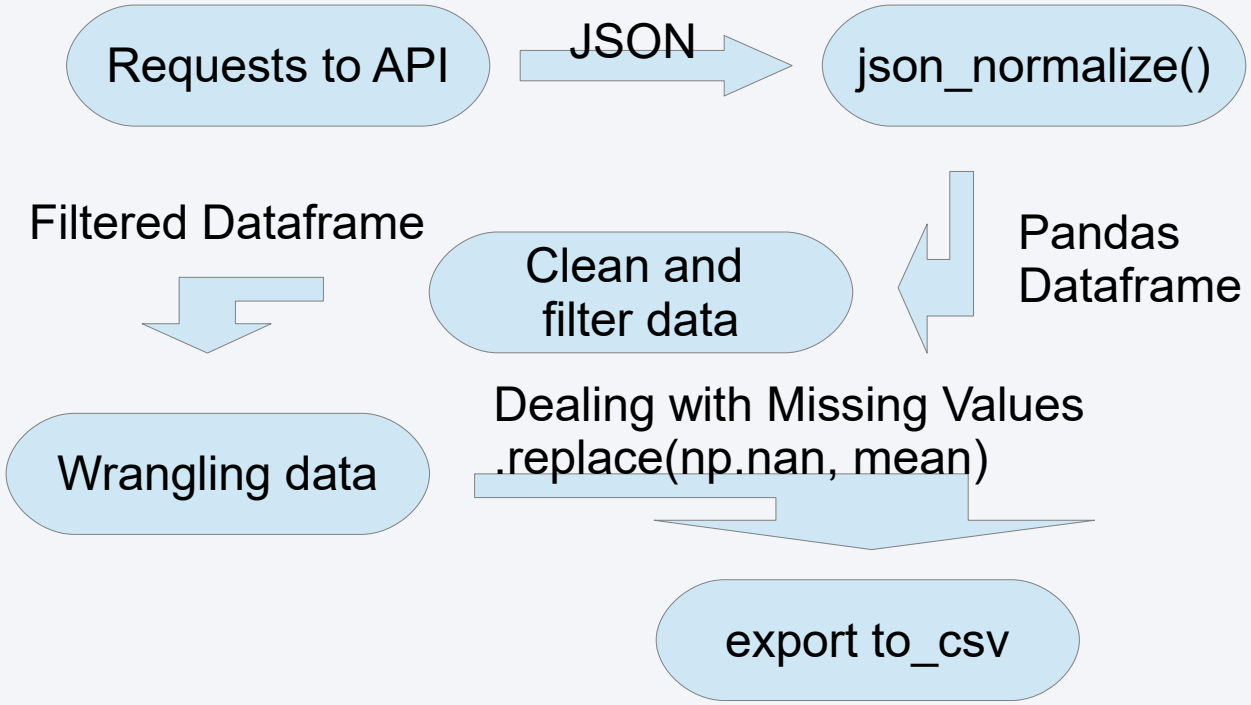
- Data collect methodology:
  - Collect data through SpaceX API
  - Scraping Wikipedia by using BeautifulSoup library
- Perform data wrangling
  - Find some patterns in the data and convert those outcomes into Training Labels
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Test four different classification models and find their best parameters using GridSearchCV

Data Collection



Data Collection – SpaceX API

Process by which using the Space X API we get the data collection in JSON format and process it to CSV after cleaning the requested data.



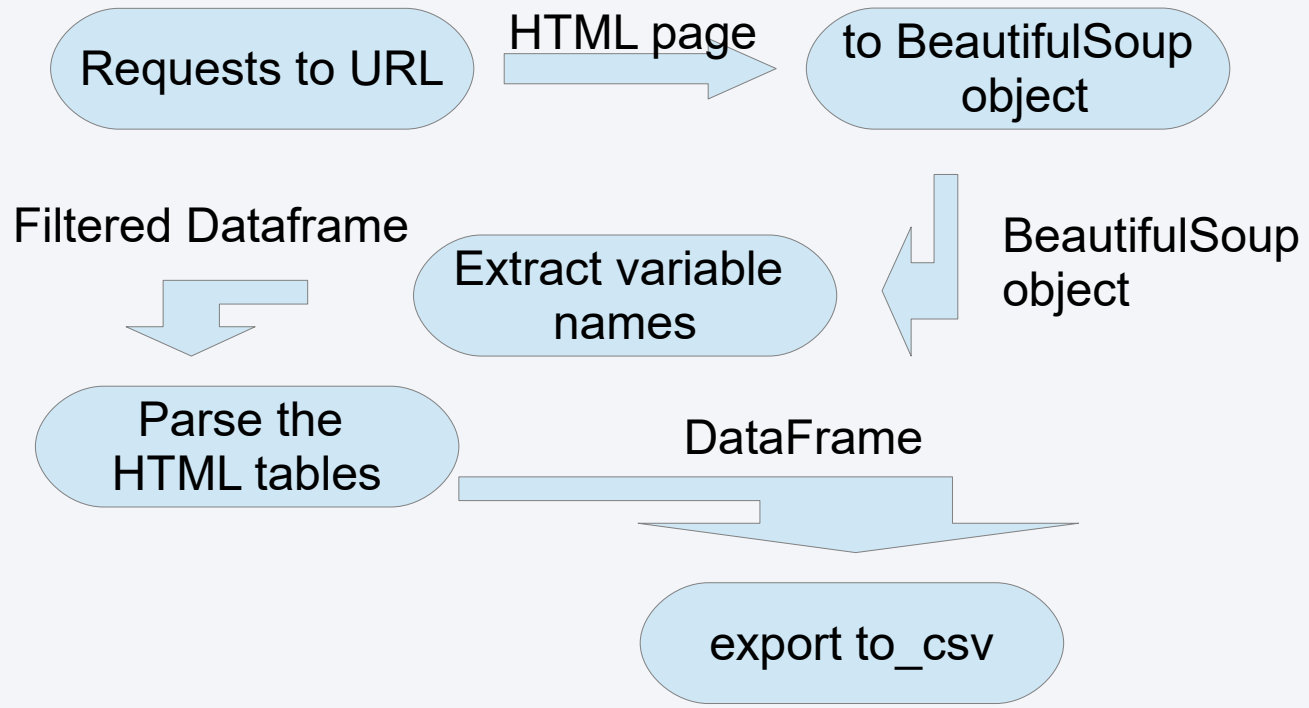
GitHub URL

Spacex data collection api

Data Collection - Scraping

Scraping Wikipedia by using BeautifulSoup library, collecting all relevant column names from the HTML table header.

Create a data frame by parsing the launch HTML tables then export it to a CSV.



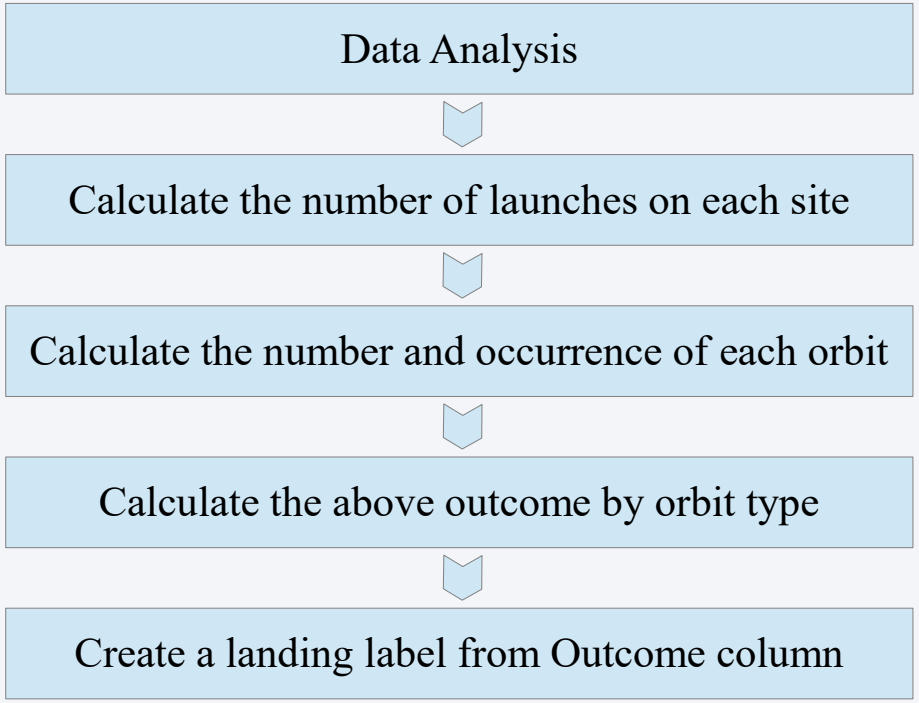
GitHub URL

Web scraping from Wikipedia

Data Wrangling

Perform an 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, we mainly convert those cases into training labels with 1 means the booster landed successfully and 0 means it was unsuccessful.



GitHub URL

Data wrangling

EDA with Data Visualization

Visualization charts and reason for the choice of the chart:

- Scatter plot: A scatter plot is a type of chart used to show the relationship between two variables. In this project we analyze the relationship between number of flights, Launch Sites and Payload.
- Bar chart: Used to display data that is organized into categories, orbits in this case and we analyzed their success rate
- Line chart: Used to show how data changes over time, success rate here.

GitHub URL

EDA with Data Visualization

EDA with SQL

SQL queries performed:

1. Display the names of the unique launch sites in the space mission
2. Display 5 records where launch sites begin with the string 'CCA'
3. Display the total payload mass carried by boosters launched by NASA (CRS)
4. Display average payload mass carried by booster version F9 v1.1
5. List the date when the first successful landing outcome in ground pad was achieved
6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
7. List the total number of successful and failure mission outcomes
8. List the names of the booster versions which have carried the maximum payload mass. Use a subquery
9. List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
10. 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, in descending order

GitHub URL

EDA with SQL

Build an Interactive Map with Folium

Using the Folium library we draw objects on a map to perform interactive visual analytics.

The map objects include:

- Markers: To mark all the launch records.
- Circles: To add a highlighted area with a text label with the name of each Launch site.
- MarkerCluster: to group all Markers in a more clear way when many are grouped according to their place of launch in a MarkerCluster.
- Lines: To highlight measured distances to key points.

GitHub URL

Interactive Map with Folium

Build a Dashboard with Plotly Dash

The dashboard application contains two charts:

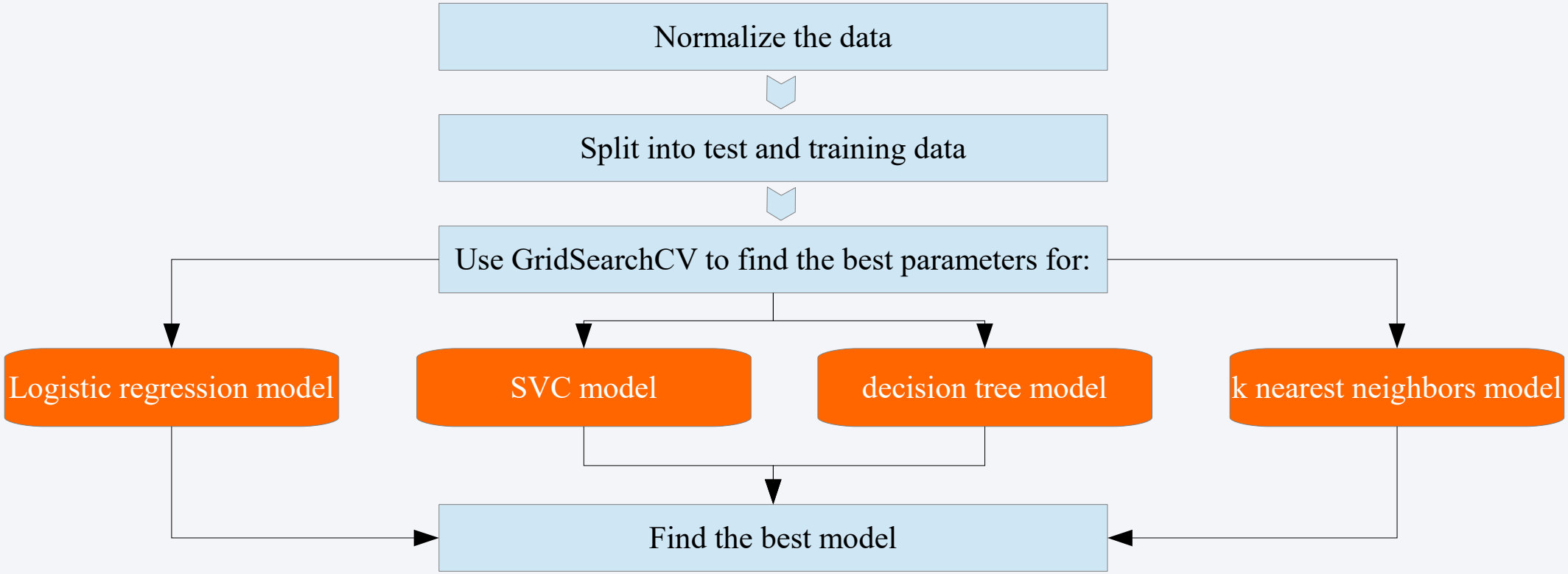
- A pie chart to show the total successful launches count for all sites, if a specific launch site was selected, show the Success vs. Failed counts for the site

- Add a scatter chart to show the correlation between payload and launch success

GitHub URL

Dash application

Predictive Analysis (Classification)



GitHub URL

Machine Learning Prediction

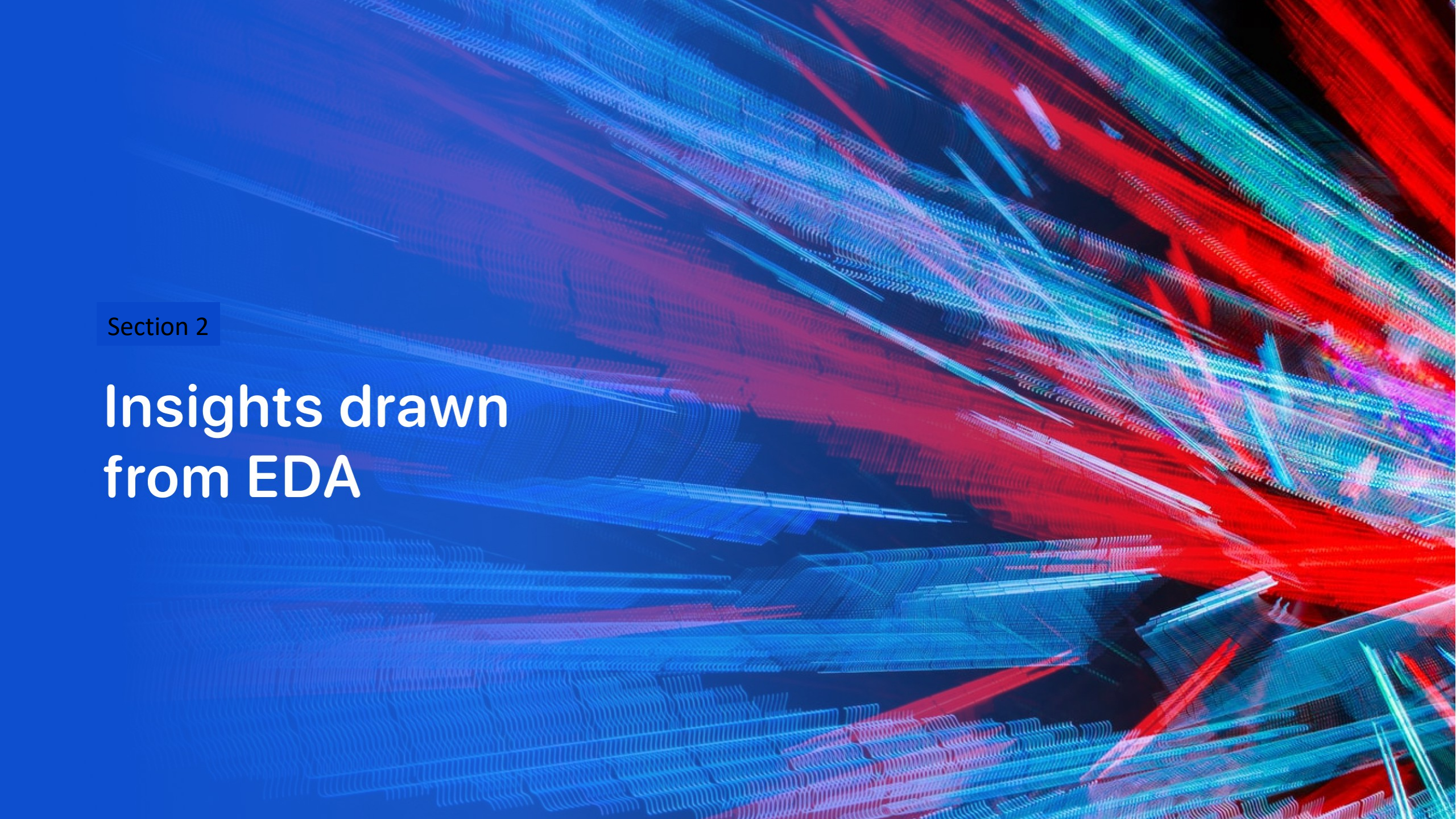
Results

The EDA has shown us that there is a relationship between the number of flights and successful landings, currently standing at around 80% success.

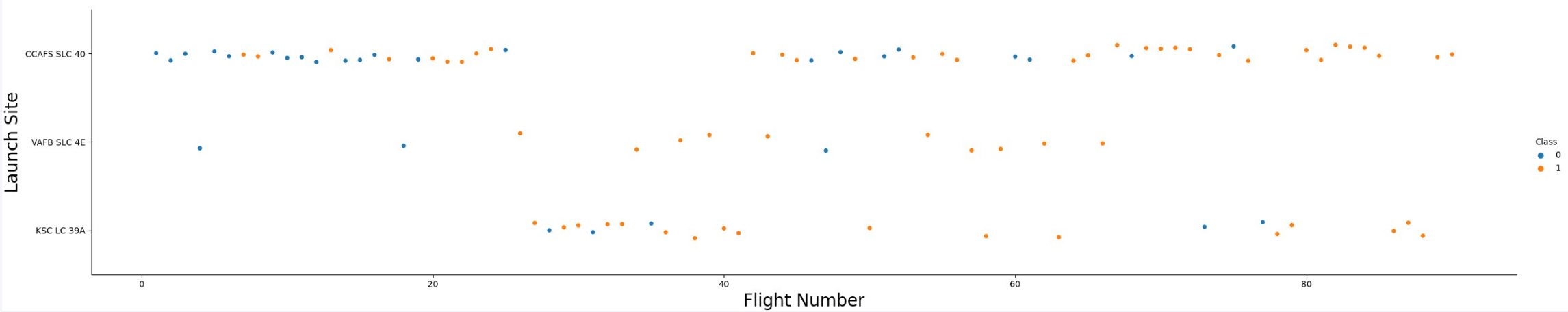
There is also a relationship with the PayLoad, being easier to land successfully the less load it has.

With the interactive visual analytics we have been able to see that the landing sites for security reasons are near the sea and to facilitate their use they have nearby train lines and highways.

3 of the 4 machine learning models were able to predict the success of a launch with an accuracy score of 83.3%


Section 2
Insights drawn from EDA

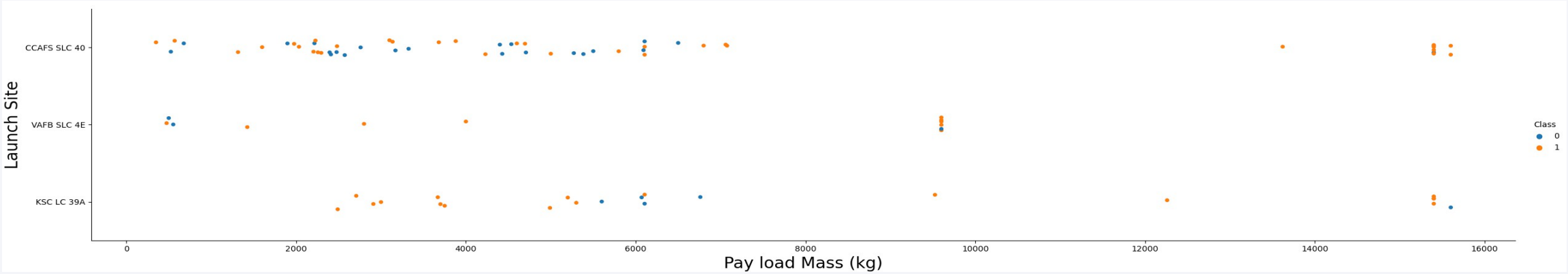
Flight Number vs. Launch Site



The blue dots represent the successful launches while the red dot represent unsuccessful luanches.

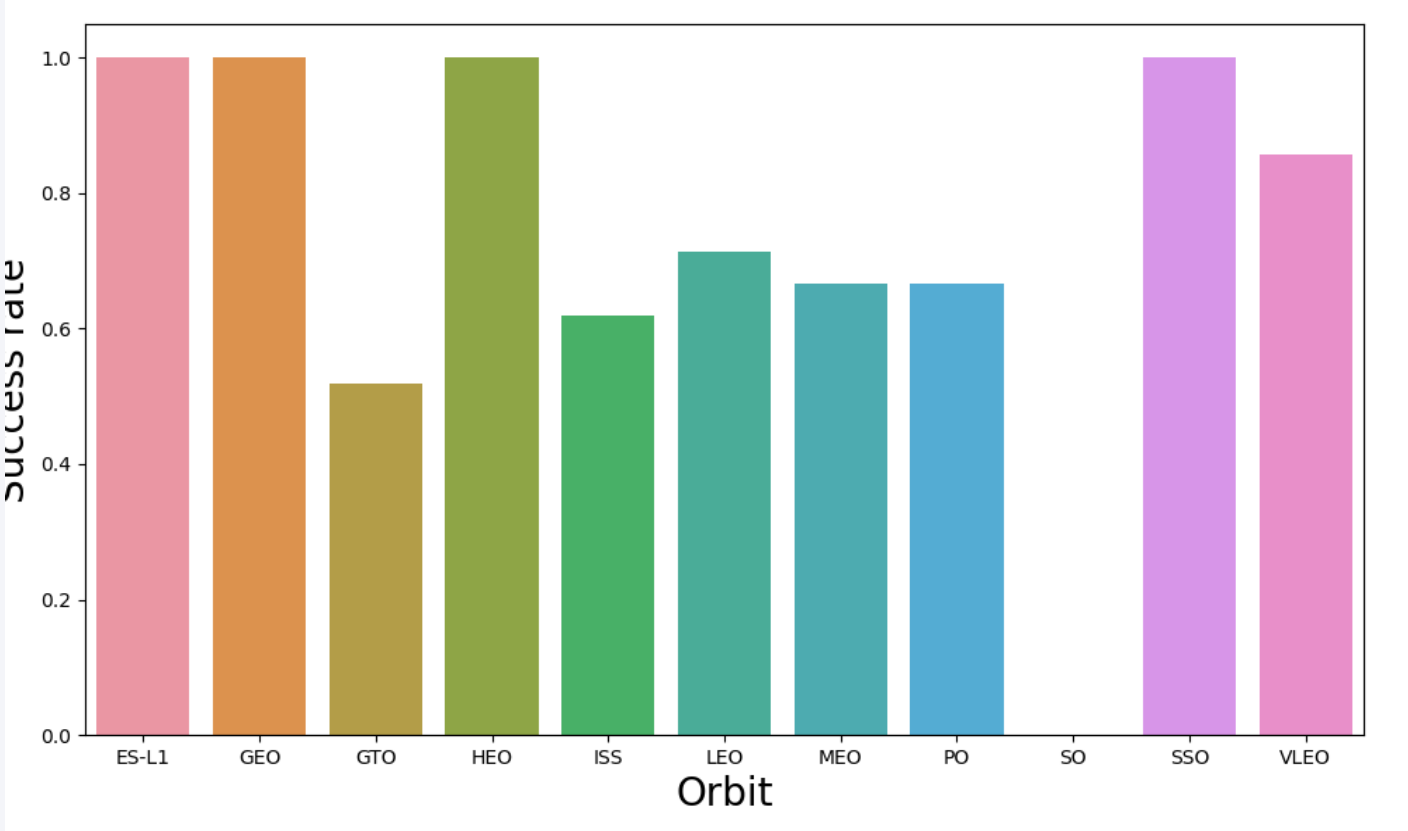
The graph shows us how the more launches there are, the more likely a successful landing seems.

Payload vs. Launch Site



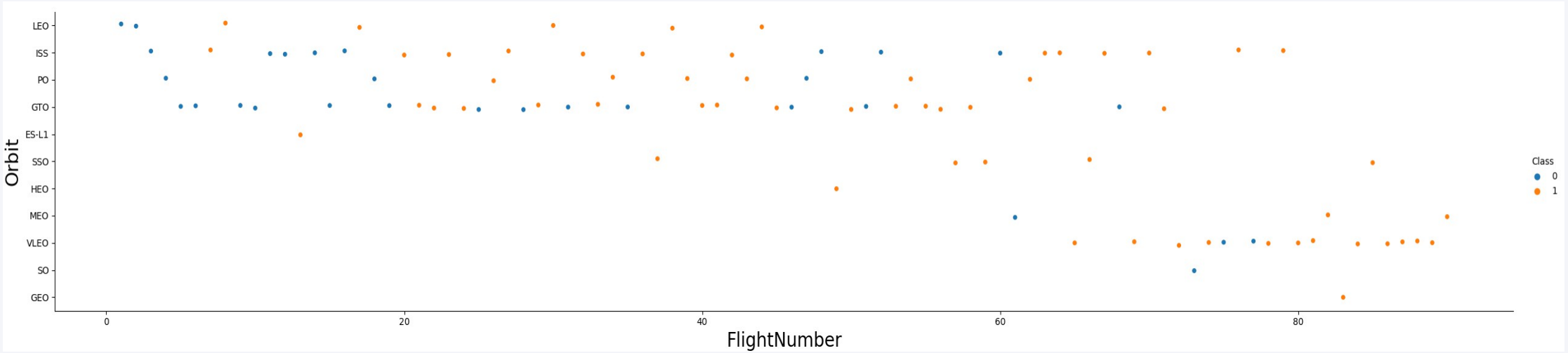
The graph shows us how it seems that the "launch site" VAFB-SLC 4E does not seem to have the ability to launch rockets with a heavy pay load (above 10000 kg).

Success Rate vs. Orbit Type



The graph seems to show that there is an important relationship between success and orbit, being 100% in the best cases (ES-L1, GEO, HEO and SSO)

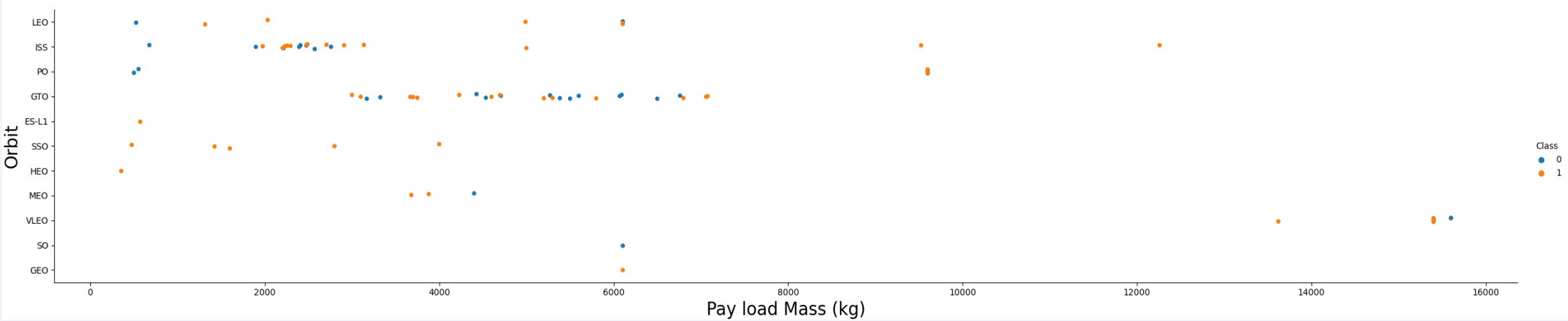
Flight Number vs. Orbit Type



We can see that in the 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.

The SSO orbit has a 100% success rate but perhaps with too few samples to be relevant data.

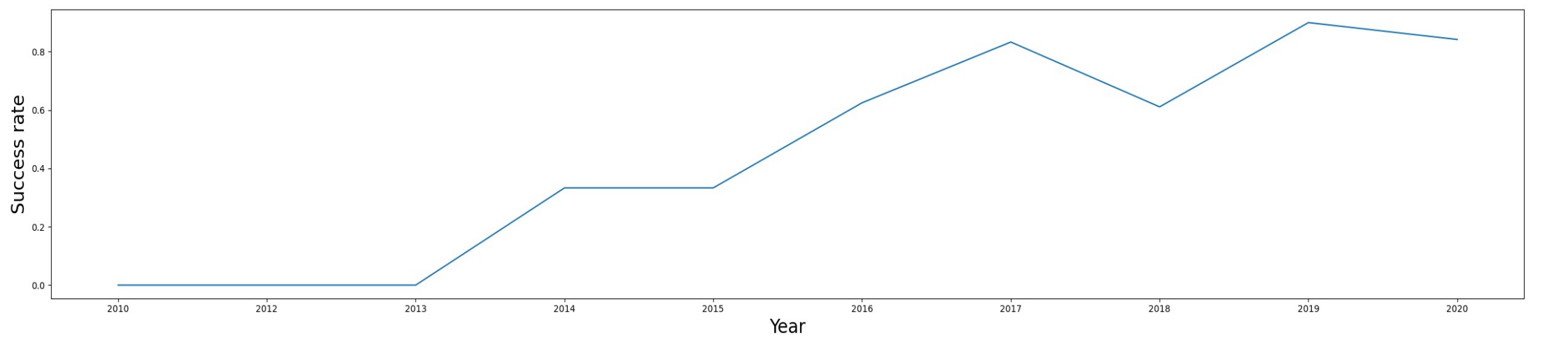
Payload vs. Orbit Type



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



We can see how the success rate has not stopped increasing since 2013.

All Launch Site Names

```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEX;
```

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

We ask that it show us the values of the Launch_Site column of the SPACEX table, the DISTINCT clause asks that it not repeat results.

Launch Site Names Begin with 'CCA'

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

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

We ask that it show us the values of the SPACEX table, the LIKE 'CCA%' clause at the end of the sentence indicates that we only want results where 'CCA' is found.

“LIMIT 5” as it clearly indicates asks to only return 5 rows.

Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS__KG_)FROM SPACEX\
    where CUSTOMER = 'NASA (CRS)'
```

1
45596

We ask that it show us the SUM of all the values of the PAYLOAD_MASS_KG_ column of the SPACEX table WHERE the value of the CUSTOMER column is = "NASA (CRS)"

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_)FROM SPACEX\
    where BOOSTER_VERSION = 'F9 v1.1'
```

1
2928

We ask that it show us the average (AVG) of all the values of the PAYLOAD_MASS_KG_ column of the SPACEX table WHERE the value of the BOOSTER_VERSION column is = "F9 v1.1"

First Successful Ground Landing Date

```
%sql SELECT Min(DATE) FROM SPACEX\
    where LANDING__OUTCOME LIKE 'Success%'
```

1
2015-12-22

We ask that it show us the minimum value (Min) of the column DATE from the SPACEX table WHERE the value of the LANDING_OUTCOME contains (LIKE) the word “Success”

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT BOOSTER_VERSION, PAYLOAD_MASS__KG_ FROM SPACEX\
    where LANDING__OUTCOME = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000
```

booster_version	payload_mass__kg_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

We ask that it show us the values of the BOOSTER_VERSION and PAYLOAD_MASS_KG_ columns of the SPACEX table, where the value of the column LANDING_OUTCOME was = “success (drone ship)” and the value of the column PAYLOAD_MAS_KG_ was between 4000 and 6000.

Total Number of Successful and Failure Mission Outcomes

```
%sql select count(MISSION_OUTCOME), MISSION_OUTCOME \
   from SPACEX \
   group by MISSION_OUTCOME
```

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

We ask that it show us the the amount (count) of the MISSION_OUTCOME column of the SPACEX table, the final group by clause asks that it show us the answers grouped by their possible types.

Boosters Carried Maximum Payload

```
%sql SELECT BOOSTER_VERSION, PAYLOAD_MASS__KG_ from SPACEX where  PAYLOAD_MASS__KG_ = (SELECT Max(PAYLOAD_MASS__KG_) from SPACEX)
```

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

We ask that it show us the values of the BOOSTER_VERSION and PAYLOAD_MASS_KG_ columns of the SPACEX table, only where the value of the column PAYLOAD_MAS_KG_ was = the maximum (Max) value of the column.

2015 Launch Records

```
%sql SELECT DATE, LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEX where LANDING__OUTCOME = 'Failure (drone ship)' AND DATE LIKE '2015%';
```

DATE	landing__outcome	booster_version	launch_site
2015-01-10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

We ask that it show us the values of the DATE, LANDING_OUTCOME, BOOSTER_VERSION and LAUNCH_SITE columns of the SPACEX table, only where the value of the column LANDING_OUTCOME was = “Failure (drone ship)” AND the column DATE contains (LIKE) “2015”

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

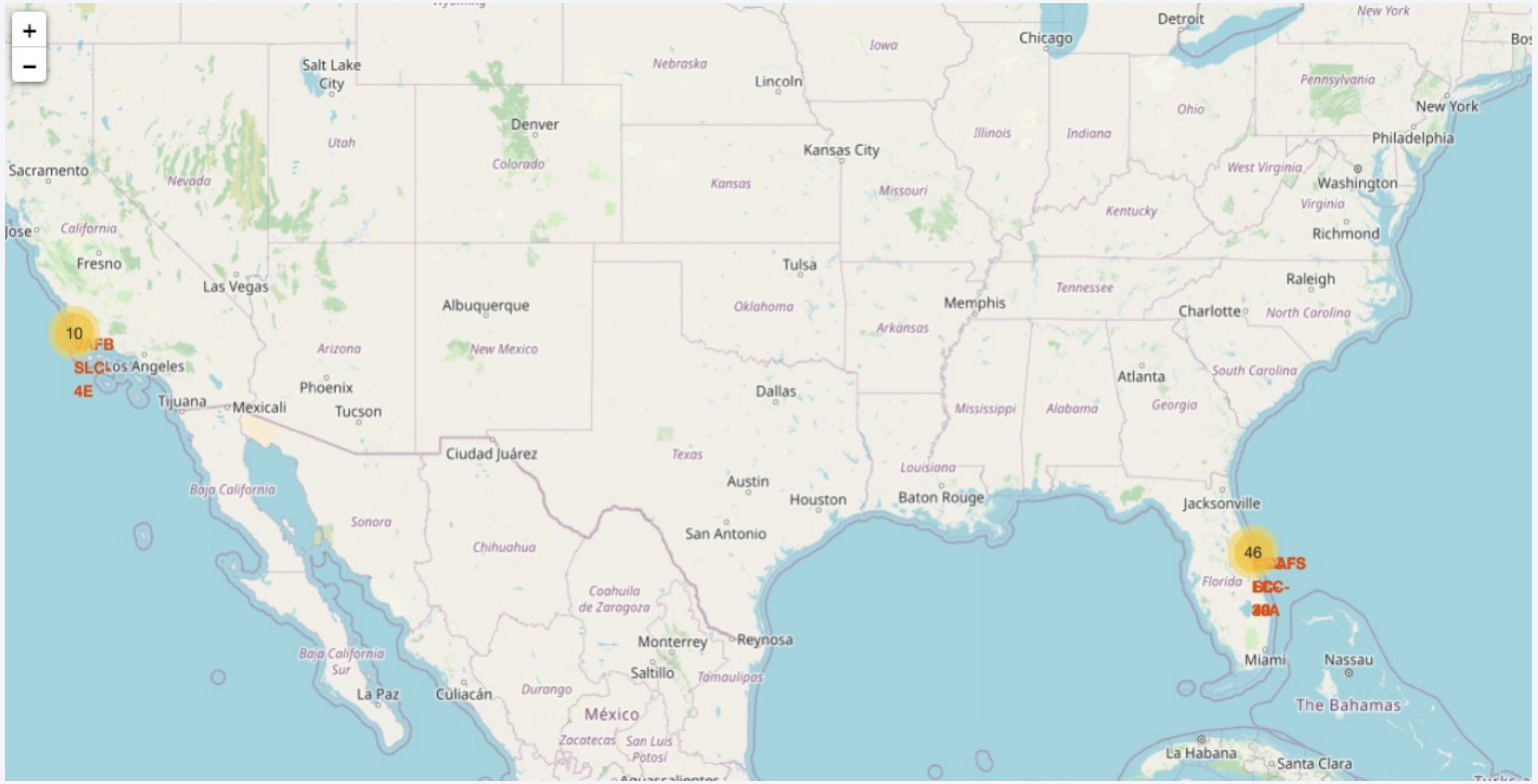
```
%sql select LANDING__OUTCOME, count(LANDING__OUTCOME) \
   from SPACEX \
   where DATE BETWEEN '2010-06-04' AND '2017-03-20'\
   group by LANDING__OUTCOME order by count(LANDING__OUTCOME) desc
```

landing__outcome	2
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

We ask that it show us the the amount (count) of the LANDING_OUTCOME column of the SPACEX table, where the value of the column DATE was between “2000-06-04” AND “2017-03-20”.
The final statement order by count (LANDING_OUTCOME) desc asks to return the values sorted in descending order based on the value of column LANDING_OUTCOME.

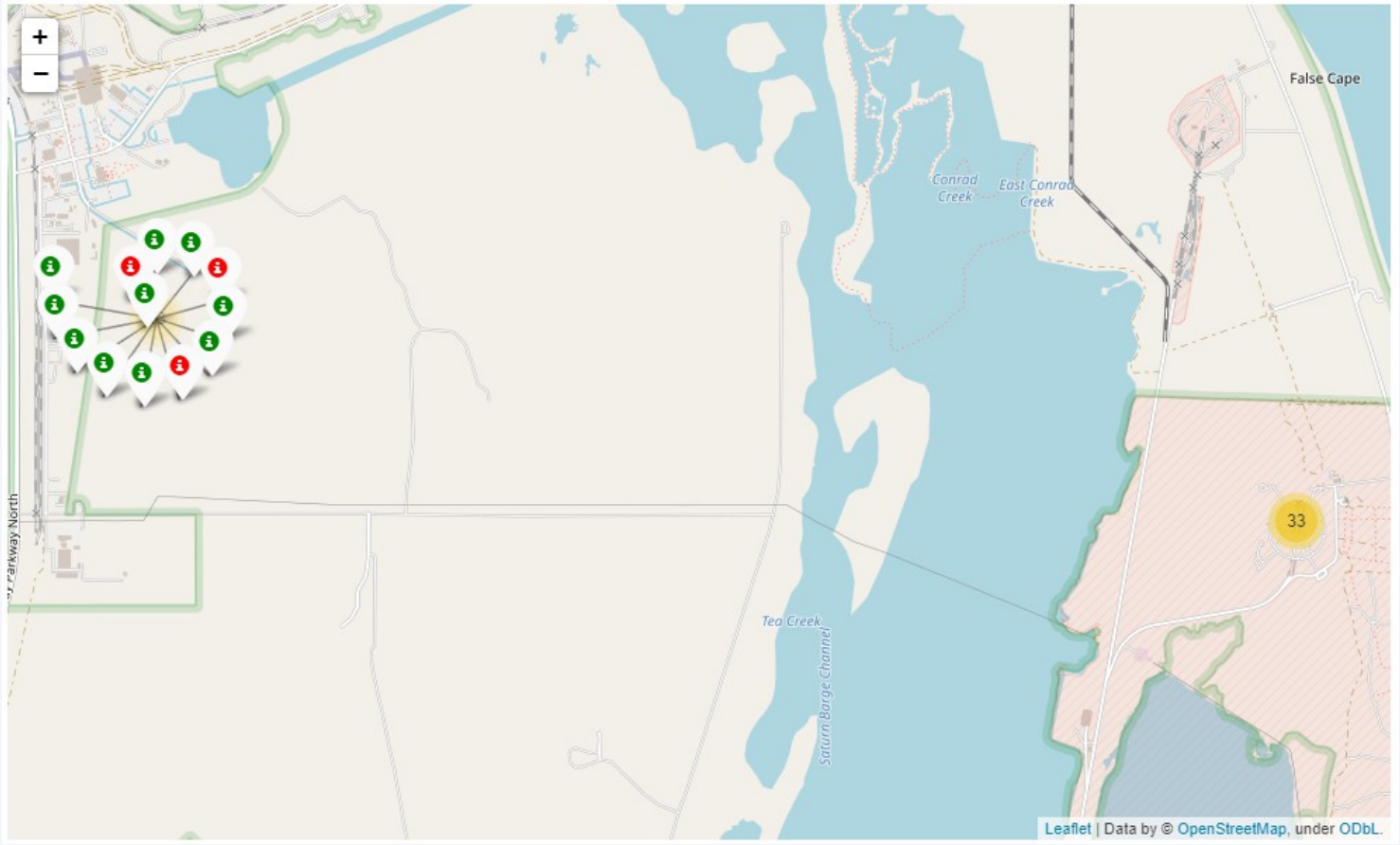

Section 3
Launch Sites
Proximities Analysis

SpaceX Launch Sites Locations



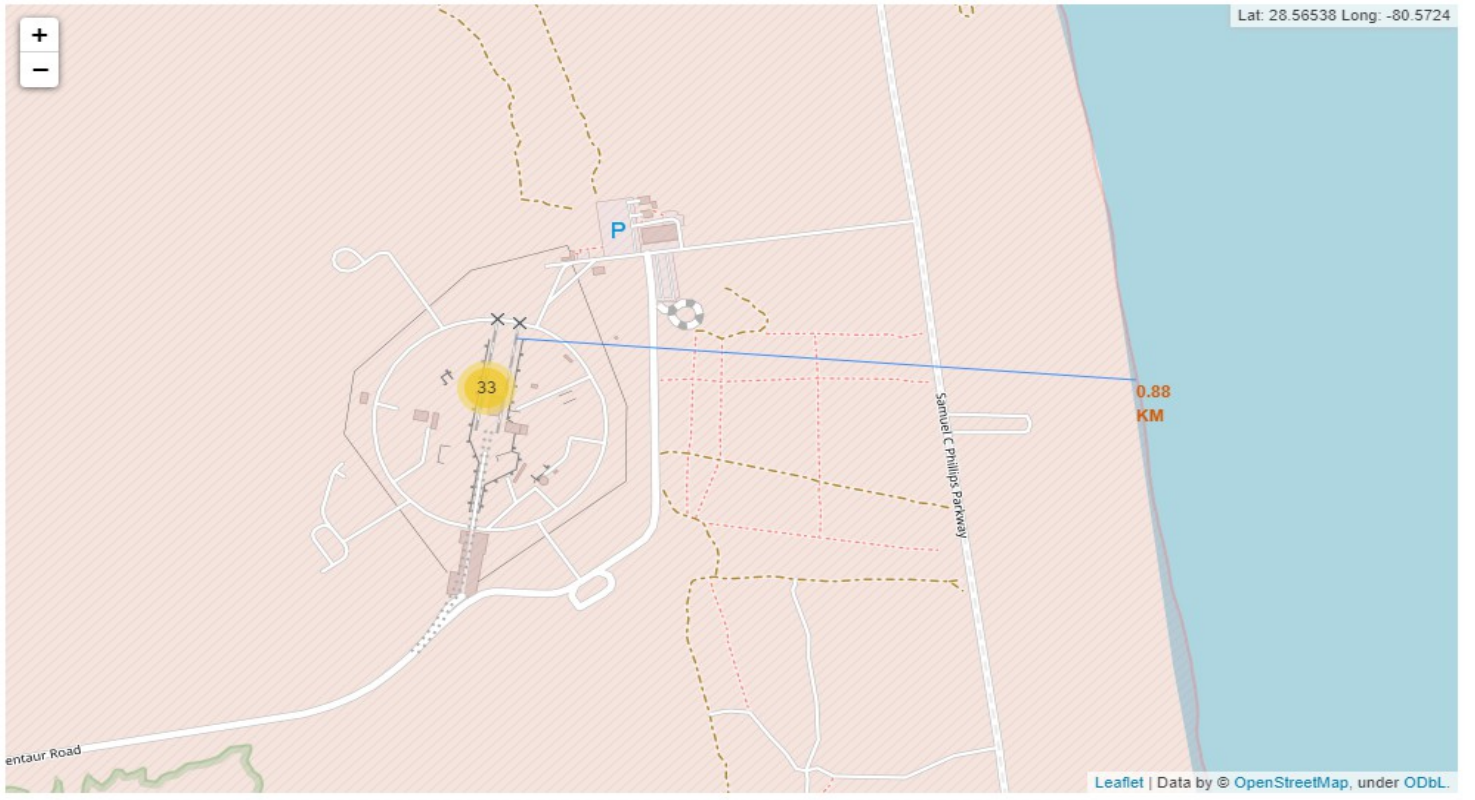
We can see how the location of the launches is near the sea.

SpaceX Launch Result

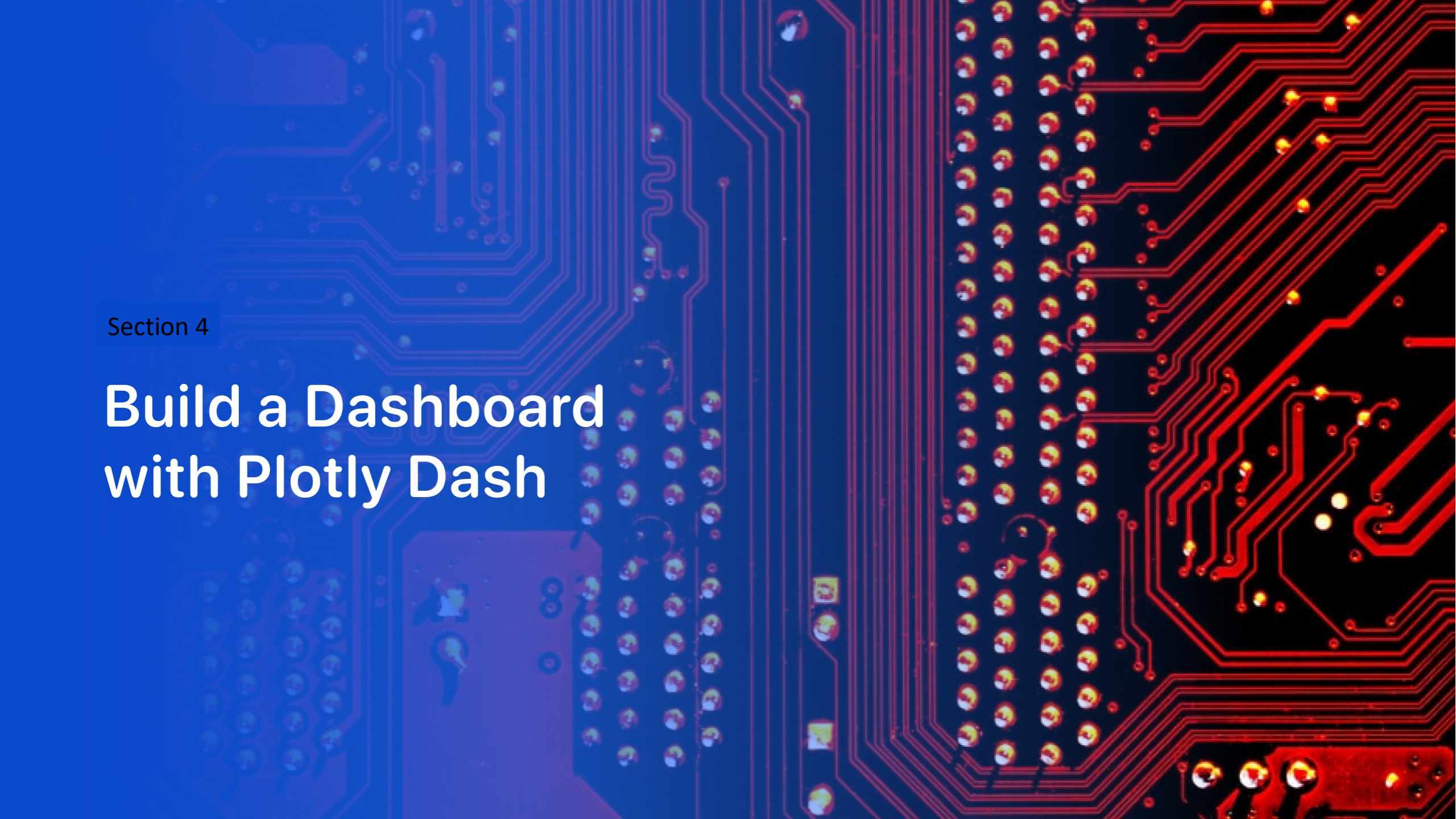


The successful launches are represented by a **green** marker while the **red** marker represents the failed ones.

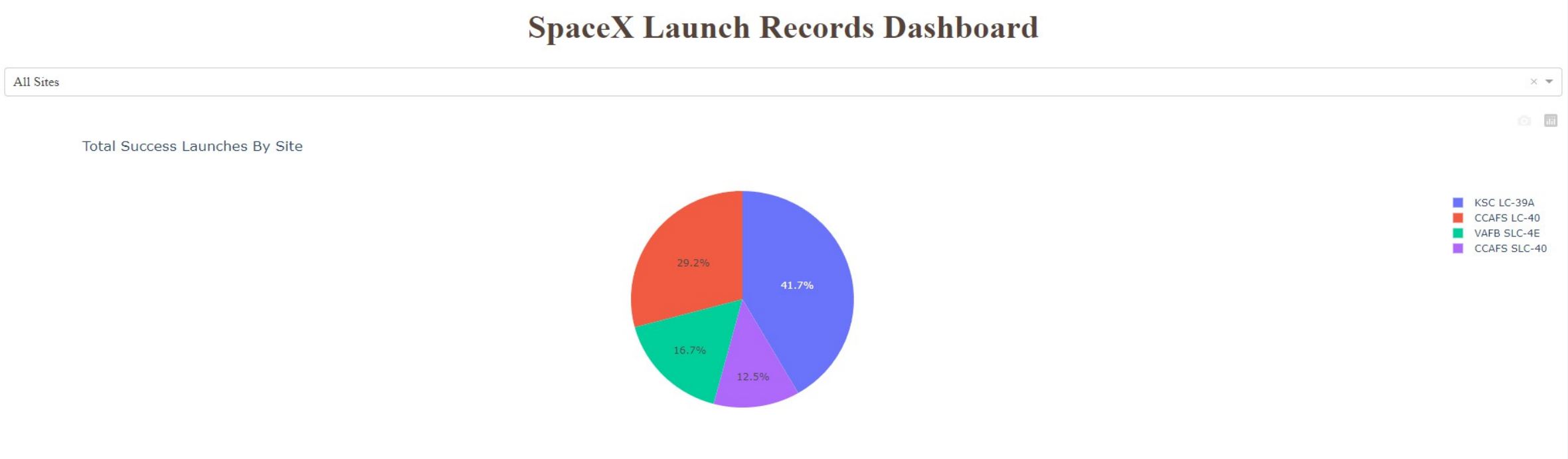
Launch Site Proximities



we have been able to see that the landing sites for security reasons are near the sea and to facilitate their use they have nearby train lines and highways.

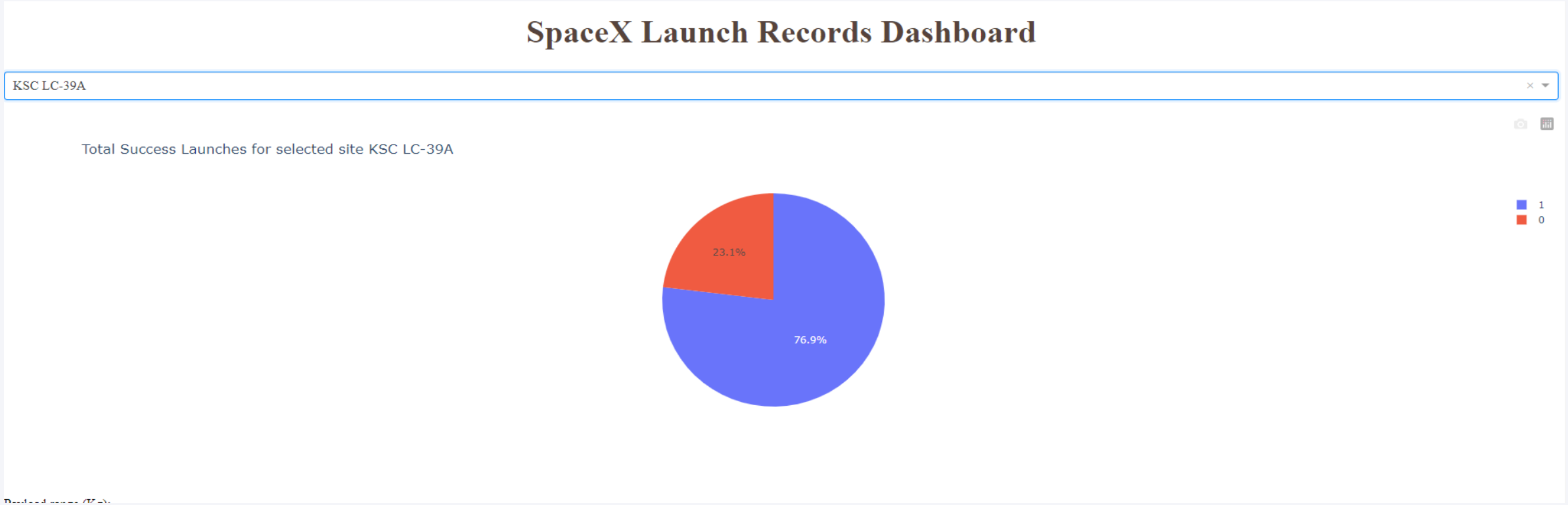

Section 4
Build a Dashboard with Plotly Dash

Total Success Launches By Site



The KSC LC-39A is the launch point with the most successful launches with 41.7% of the total successful launches

Most Successful Launch Site



The KSC LC-39A was the most successful launch site with a successful rate of 76.9%

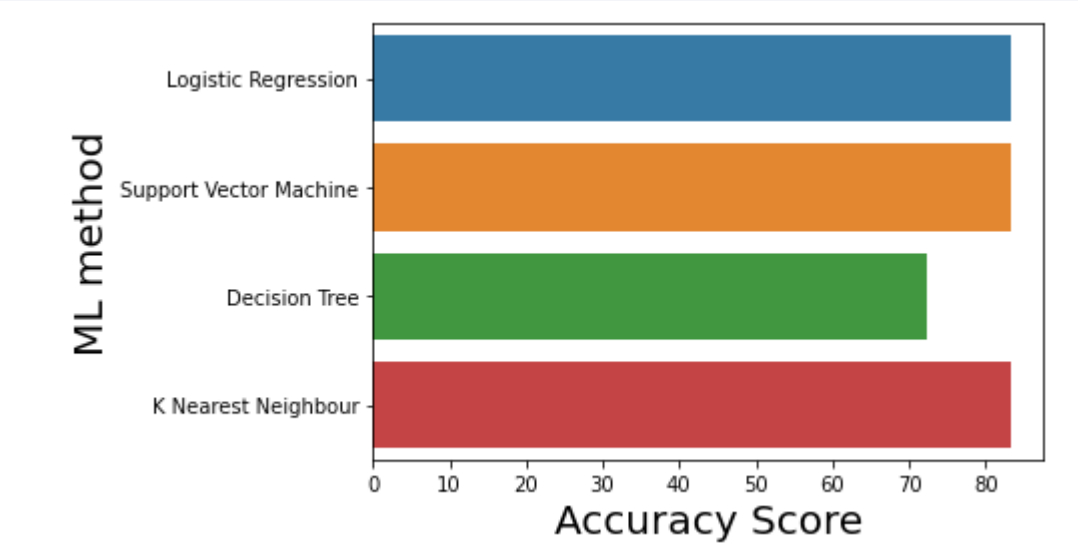
Payload Vs. Launch Outcome

Below a Payload of 6000 kg we can see that a higher Success rate is achieved

Section 5

Predictive Analysis (Classification)

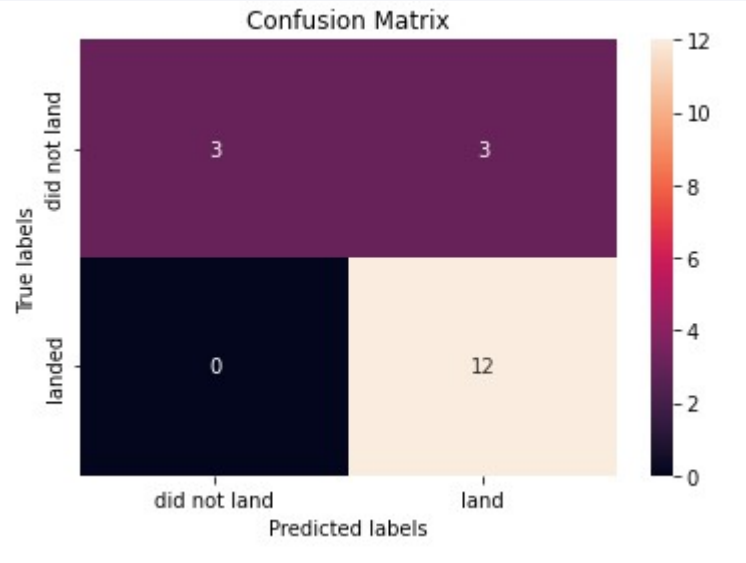
Classification Accuracy



3 of the methods have an accuracy score of 83.3%, the only one that fails to reach it is the Decision Tree, which is left with a slightly lower result of 72%.

The models were tested on 3 occasions, all the models repeated the results, except for the Decision Tree, which managed to reach the same value as the rest in the second attempt 83.3%, but in the other 2 it remained below

Confusion Matrix



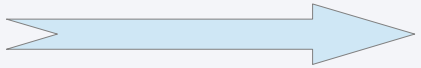
The model predicted 15 times that it would land guessing 12 times.

The confusion matrix is the same in the 3 models.

Conclusions

- The launch sites are strategically located near highways and railways for the transport of personnel and cargo, but also away from cities and near the sea for safety.
- With loads below 6000 kg the successful rate increases a lot.
- SpaceX has been achieving a higher success rate every year, considering that we have only calculated up to 2020 the data today could be somewhat out of date.
- We can predict the successful rate with 83.3% accuracy.

Appendix

GitHub repository  https://github.com/GonzaloSBG/IBM-Data-Science

Thank you!