IBM Developer
SKILLS NETWORK

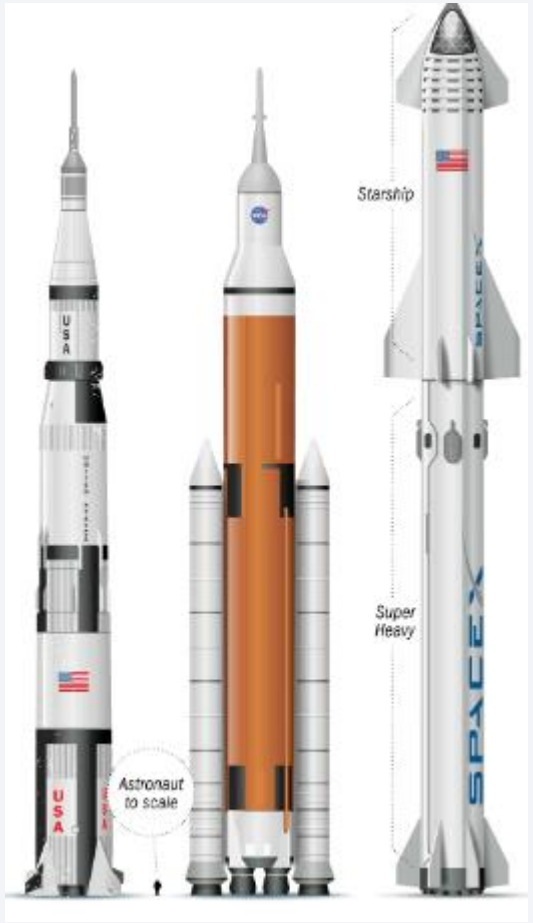
Winning Space Race with Data Science

Claudio Castro
15/10/2023

Outline

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

Executive Summary



This research project focuses on developing a **predictive model for the successful landing of the Falcon 9 rocket's first stage**, a crucial factor in SpaceX's ability to offer cost-effective launches. With the aim of assisting other space companies in improving their bidding strategies, we employed data science techniques from data collection to machine learning.

The data was collected through **web scraping** from publicly available sources such as Wikipedia as well as SpaceX API, and **extensive data wrangling and exploration were performed using SQL** to build a clean and structured dataset. We also employed **data visualization** to understand the correlations between various factors, including the intended orbit of launch, rocket model, and payload mass.

Machine learning techniques, including decision trees, k-nearest neighbors, logistic regression, and support vector machines, were implemented to create predictive models. Our final model achieved an **93% accuracy rate in predicting the success of the first stage's landing for the test set.**

This research provides valuable insights into the factors influencing the successful landing of the Falcon 9, potentially aiding other companies in optimizing their rocket launch bids.

Introduction

SpaceX's Falcon 9 rocket has revolutionized the space launch industry by dramatically reducing the cost of access to space. One of the key reasons for this cost reduction is SpaceX's pioneering approach of reusing the first stage of the Falcon 9 rocket, which significantly lowers the overall launch expenses. Other space launch providers typically spend upwards of $165 million for each launch, while SpaceX offers its services at a mere $62 million per launch, thanks to the reusability factor. **A crucial aspect of this reusability is the successful landing of the Falcon 9's first stage**, which we set out to predict in this research.

The ability **to predict the success of a rocket's first stage landing** is of great importance for both SpaceX and other aerospace companies looking to compete with SpaceX's cost-efficient offerings. By developing a predictive model that can assess the likelihood of a successful landing, other companies can fine-tune their bidding strategies, aiming to win contracts by offering reliable and cost-effective launch services.
Our research project encompasses a comprehensive data science approach.

We started by collecting data from freely available sources, particularly Wikipedia, and underwent an extensive data wrangling process to create a clean and well-structured dataset. Leveraging SQL, we performed data exploration and employed data visualization tools to unravel correlations between various variables such as the intended orbit of launch, the model of the rocket used, and the payload mass.

To create a robust predictive model, we experimented with different machine learning techniques, including decision trees, k-nearest neighbors, logistic regression, and support vector machines. After rigorous testing and evaluation, we identified a final model that achieved an impressive accuracy rate of 93% in predicting the success of the Falcon 9's first stage landing.

In this paper, we present the results of our research, offering valuable insights that can benefit not only SpaceX but also other aerospace companies looking to enhance their competitiveness in the rocket launch industry. Our findings can guide these companies in optimizing their bidding strategies and ensuring the reliability of their launch services.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Collecting data using the Space-X API and data scraping from Wikipedia.
- Perform data wrangling
  - Applying data wrangling with python.
- 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

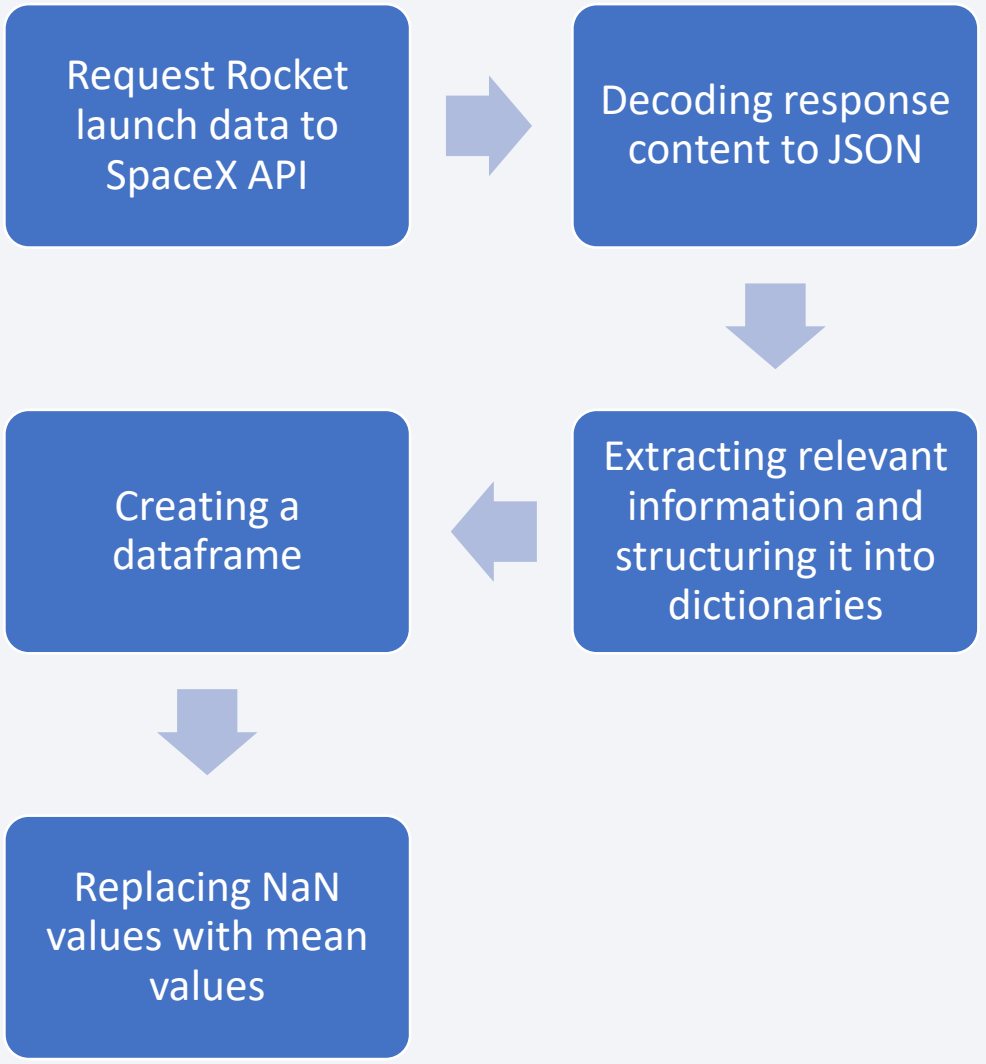
Data Collection

All the data used in this project was obtained from publicly available sources. The first one is directly from the SpaceX REST API and the second one is the information of launches by spaceX available from Wikipedia. From this data we selected the information relevant to our research which is information about rockets, payloads, launchpads, and cores.

Data Collection – SpaceX API

- Data collection from the SpaceX API was done on python using a get.request() command on the API to get a JSON file which was then arranged into a dataset.

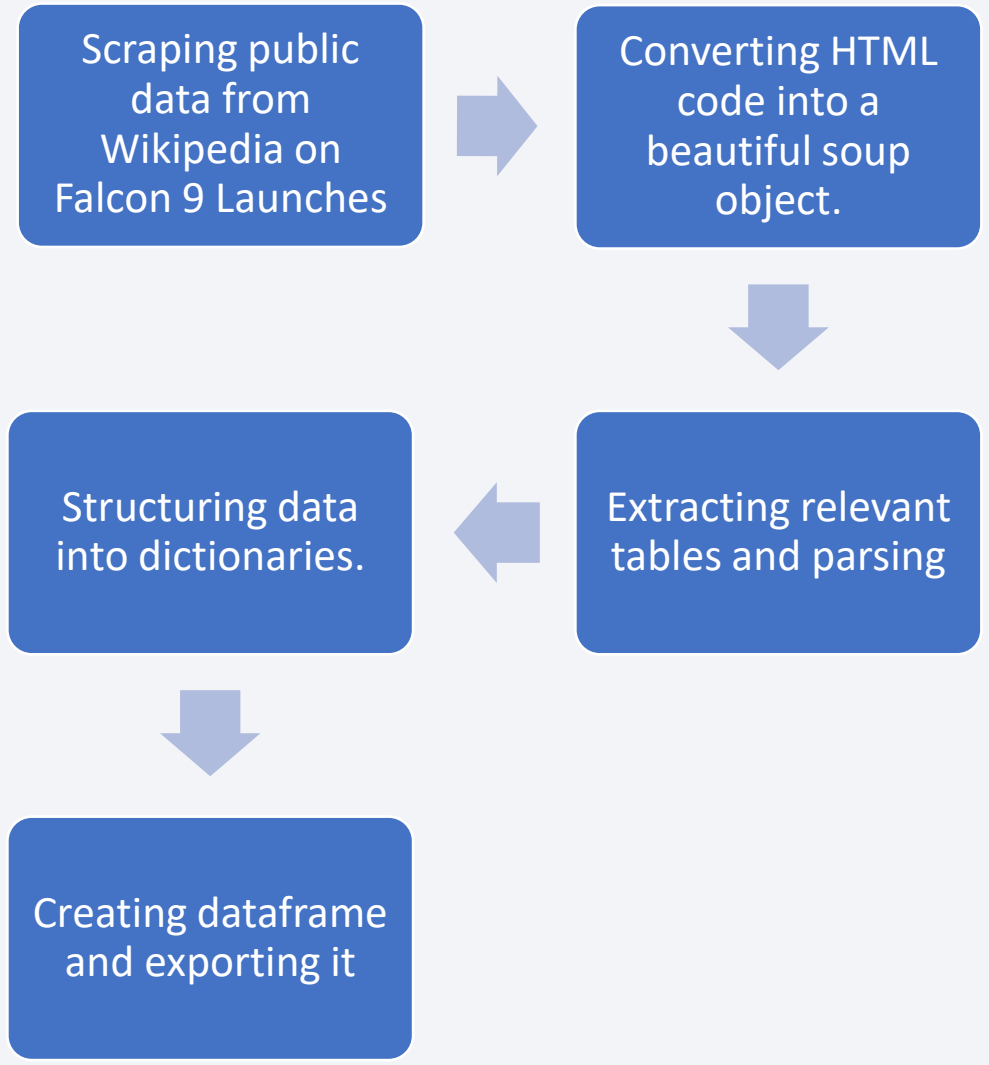
- The Jupiter notebook where this process was done can be found in this link.



Data Collection - Scraping

- Data was also collecting using public available data from Wikipedia by webscrapping using python's beautiful soup.

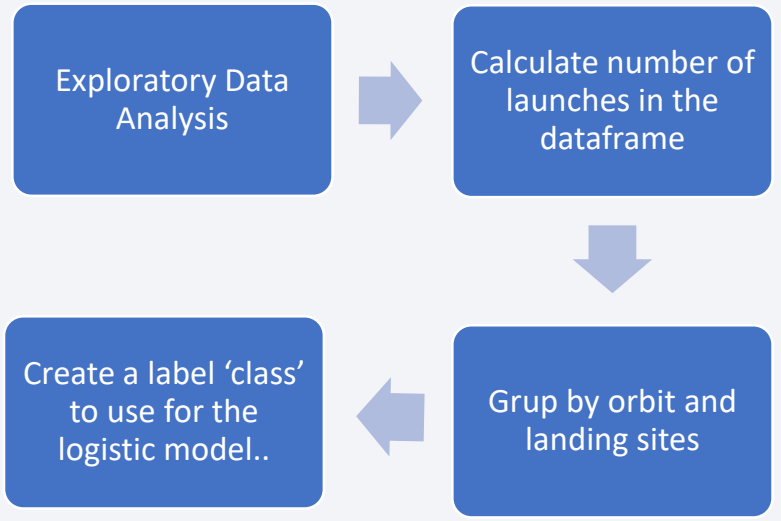
- The Jupiter notebook where this process was done can be found in this link.



Data Wrangling

In the data wrangling process. The data set created in the previous step is reduced to better represented the information needed for our model. This is, the success or failure of each launch. In this sense, data is going to be classified by success or failure creating by grouping by relevant variables consider to have a correlation with the success or failure og launches such as launch sites and orbits. A label called ‘class’ is the created to represent fail/success which is binary. This also allow us to calculate the success rate which ended up being 66,6%.

The notebook detailing this process can be found in the following link.



EDA with Data Visualization

We used Data Visualization in order to easily identify the relationship between the many variables that could have an incidence on the 'class' of each launch. In particular we explored the relationships:

- Launch Site vs Flight Number (scatter plot)
- Launch Site vs Payload Mass (scatter plot)
- Rate of success for each orbit (bar chart)
- Orbit vs Flight Number (scatter plot)
- Orbit vs Payload mass (scatter plot)
- Succes Rate vs Time (line plot)

This process can be detailed in the Jupiter notebook found in this link.

EDA with SQL

We used SQL to perform exploratory data analysis by which be applied the following queries.

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

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

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

```
%sql SELECT * FROM SPACEXTBL WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5
```

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

```
%sql SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "Booster_Version" LIKE 'F9 v1.1%'
```

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

```
%sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "Customer" = 'NASA (CRS)'
```

- List the date when the first successful landing outcome in ground pad was achieved.

```
%sql SELECT DISTINCT "Landing_Outcome" FROM SPACEXTBL
```

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

```
%sql SELECT DISTINCT("Booster_Version")  FROM SPACEXTBL WHERE "Landing_Outcome" = 'Success (drone ship)' AND "PAYLOAD_MASS__KG_" < 6000 AND "PAYLOAD_MASS__KG_" > 4000
```

- List the total number of successful and failure mission outcomes

```
%sql SELECT DISTINCT "Mission_Outcome" FROM SPACEXTBL
```

- List the names of the booster_versions which have carried the maximum payload mass.

```
%sql SELECT DISTINCT ("Booster_Version")   FROM SPACEXTBL WHERE "PAYLOAD_MASS__KG_" = 15600
```

- 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

```
%sql SELECT MONTH(DATE),MISSION_OUTCOME,BOOSTER_VERSION,LAUNCH_SITE FROM SPACEXTBL where EXTRACT(YEAR FROM DATE)='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, in descending order.

```
%sql SELECT LANDING_OUTCOME FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' ORDER BY DATE DESC;
```

This process can be detailed in the Jupiter notebook found in this link.

Build an Interactive Map with Folium

Using the Folium package we proceeded to create a visualization on the physical locations of each landing site in the world's map, To the map we added information on the location of each launch with labels to identify successful from failed launches.

This objects where added with the purpose to illustrate the relation between the class of each launch with its physical location. In addition to this, we also included annotations to illustrate how close where the landing sites to locations like roadways, railroads and cities.

This process can be detailed in the Jupiter notebook found in this link.

Build a Dashboard with Plotly Dash

We used Plotly Dash to add further insight on the relationship between our variables of interest by:

Our visualization displays the proportion of success launches for the different launching sites using a dropdown menu.

We also included an interactive range slider to visualize the relationship between payload and launch success.

Predictive Analysis (Classification)

For the predictive analysis stage we used the data prepared from the previous state to design a model to accurately predict success or fail for launces.

The first step was to divide the data set into a test and training set. Afterwards we proceeded to test different machine learning paradigms:

- decision trees
- k-nearest neighbors
- Logistic regression
- support vector machines,

To test each model, we measured its accuracy to determine which one was the best performing model.

This process can be detailed in the Jupiter notebook found in this link

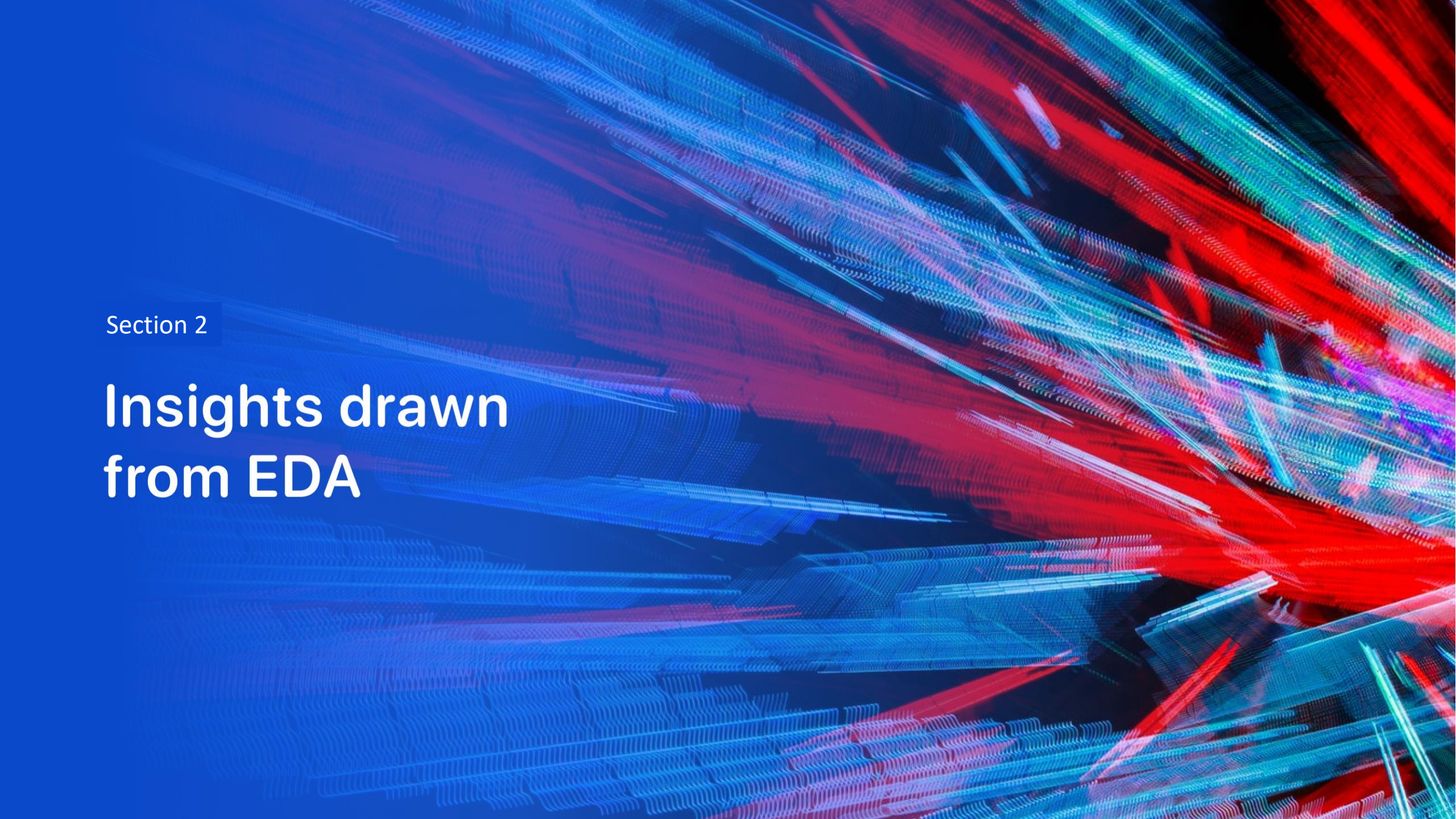
Results

The exploratory data Analysis allowed us to find some interesting insight:

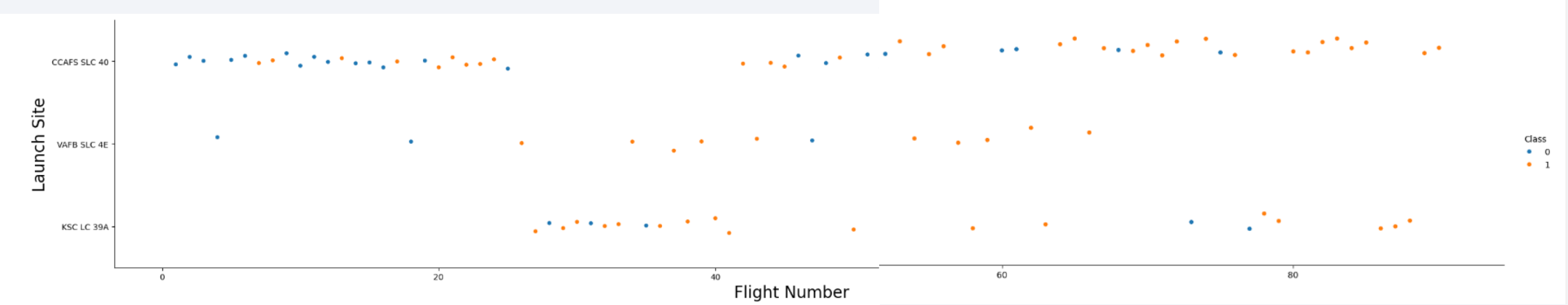
- Launches were carried out from 4 different landing sites been the most common CCAFS
- The average payload is 2435kg
- Most missions resulted in success (it's important to state that success is not the same in the context of our research, success means that the payload was delivered)
- The success rate for each landing has improved with the years and continues to do so.
- There is a different success ratio from some orbits compared to others
- Even when there is variance between the class of landings depending on landing site, payload and orbit, the strongest predictor of landing success is how recent the landing is considering that the trend is for landings to be more successful proportionally.

Predictive analysis results:

For the predictive analysis the model that perform the best was with a test accuracy of **0,94** is **the decision trees** model

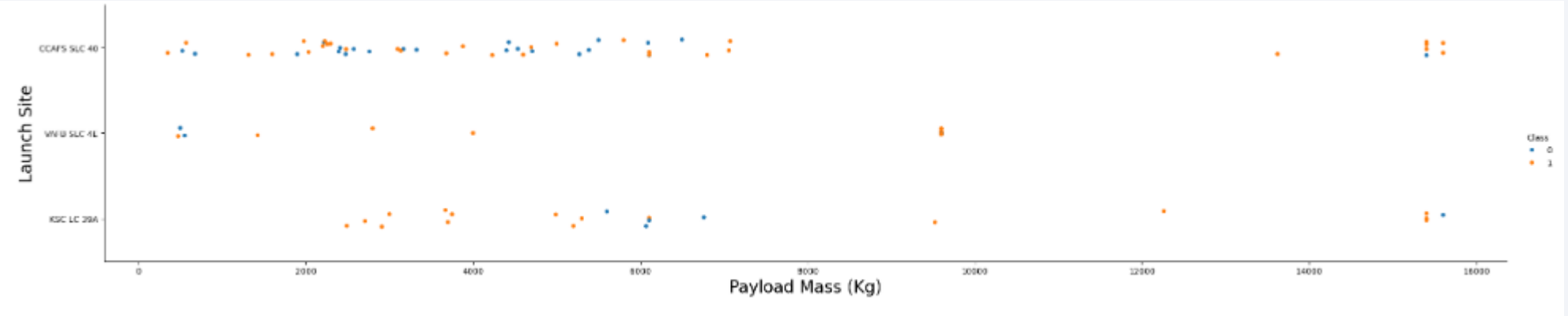

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



This plot represents a visualization of the relationship between flight number and launch site. This plot confirms the fact that with each new launch the probability of a successful landing increases. We can also see that there are many more launches on the first launch site and that the second one seems to be the most successful one by proportion.

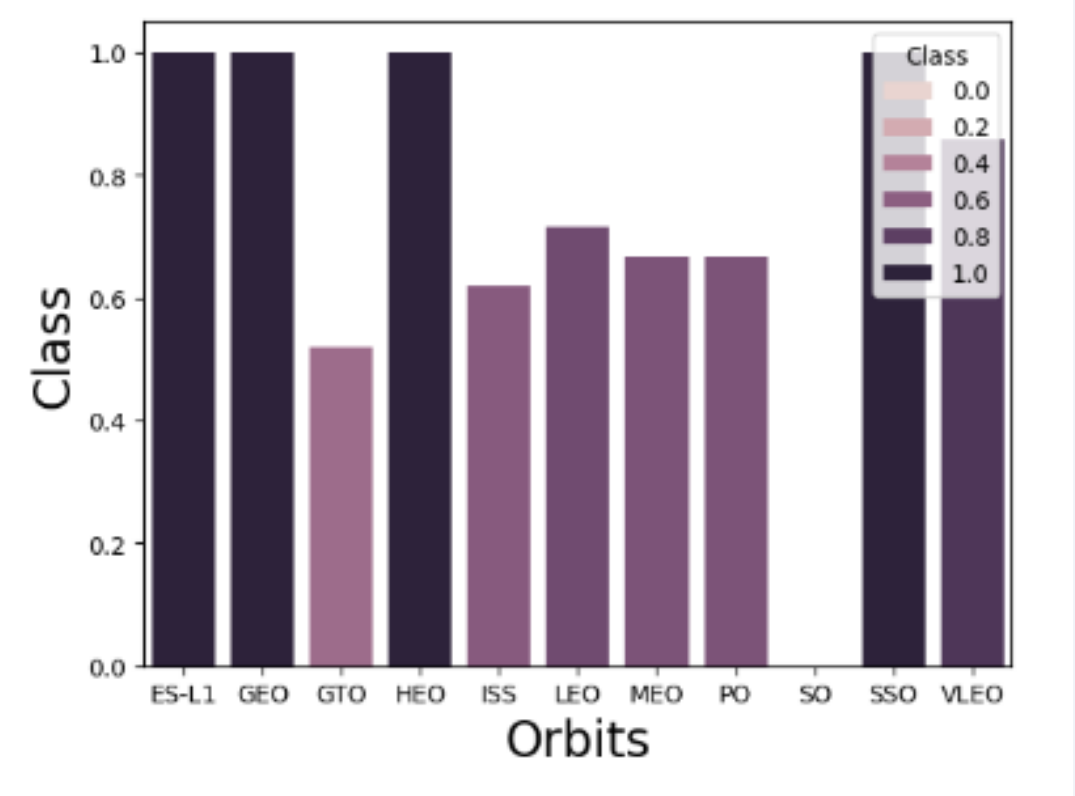
Payload vs. Launch Site



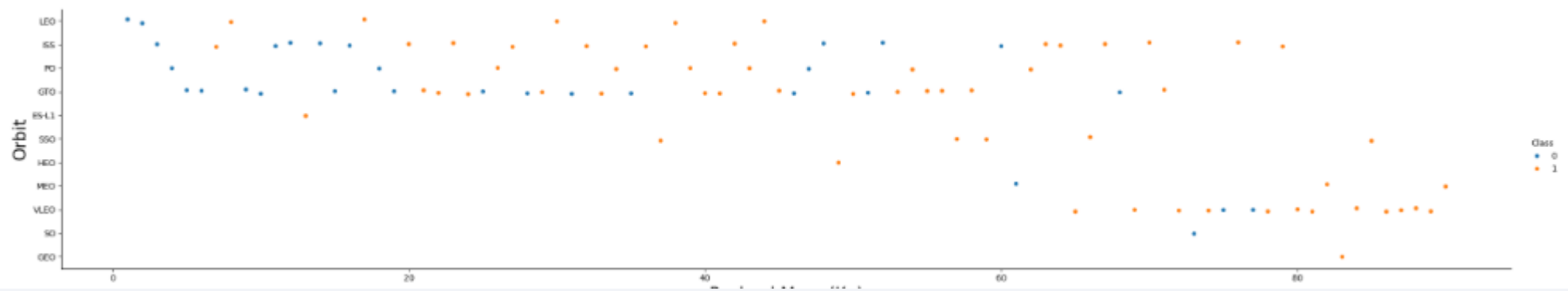
This plot represents a visualization of the relationship between flight number and payload. Higher payloads are les common but have a greater likelihood of succesful landing. We can also see that there is a gap for payloads in the mid-range.

Success Rate vs. Orbit Type

Orbits have a different average rate of success. It is worth noting that the orbits with the highest score are also the ones with with less launches so perhaps this information needs to be complemented with the number of flights which is the next graph's purpose.

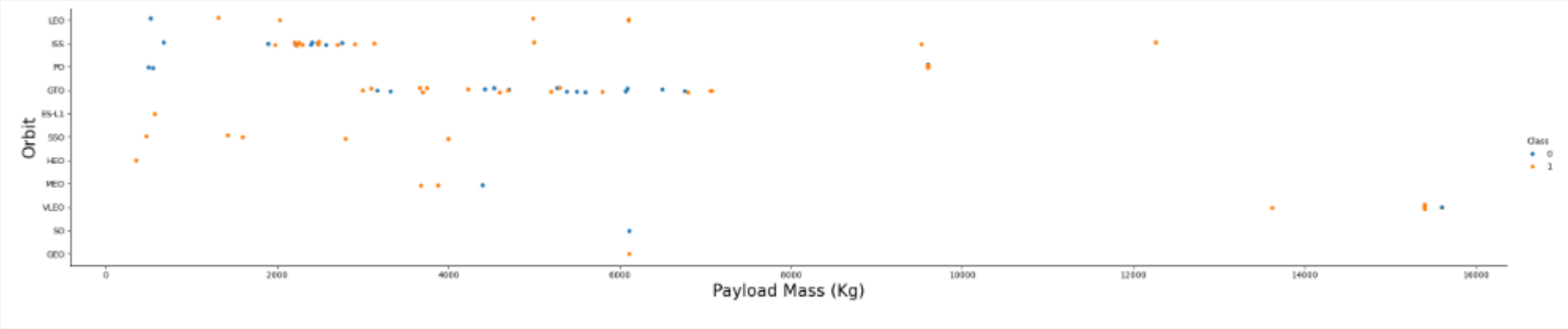


Flight Number vs. Orbit Type



This chart seems to confirm what was previously accerted. Most orbits with a high succes rate are “one time” successes. The orbits ISS, LEO, GTO, PO and VLEO are the ones with the most number of flights. All of them having similar success rates with the execption of VLEO which is the highest with .8. In adittion to this, for the LEO orbit there seems to be a Good correlation between Flight number and success.

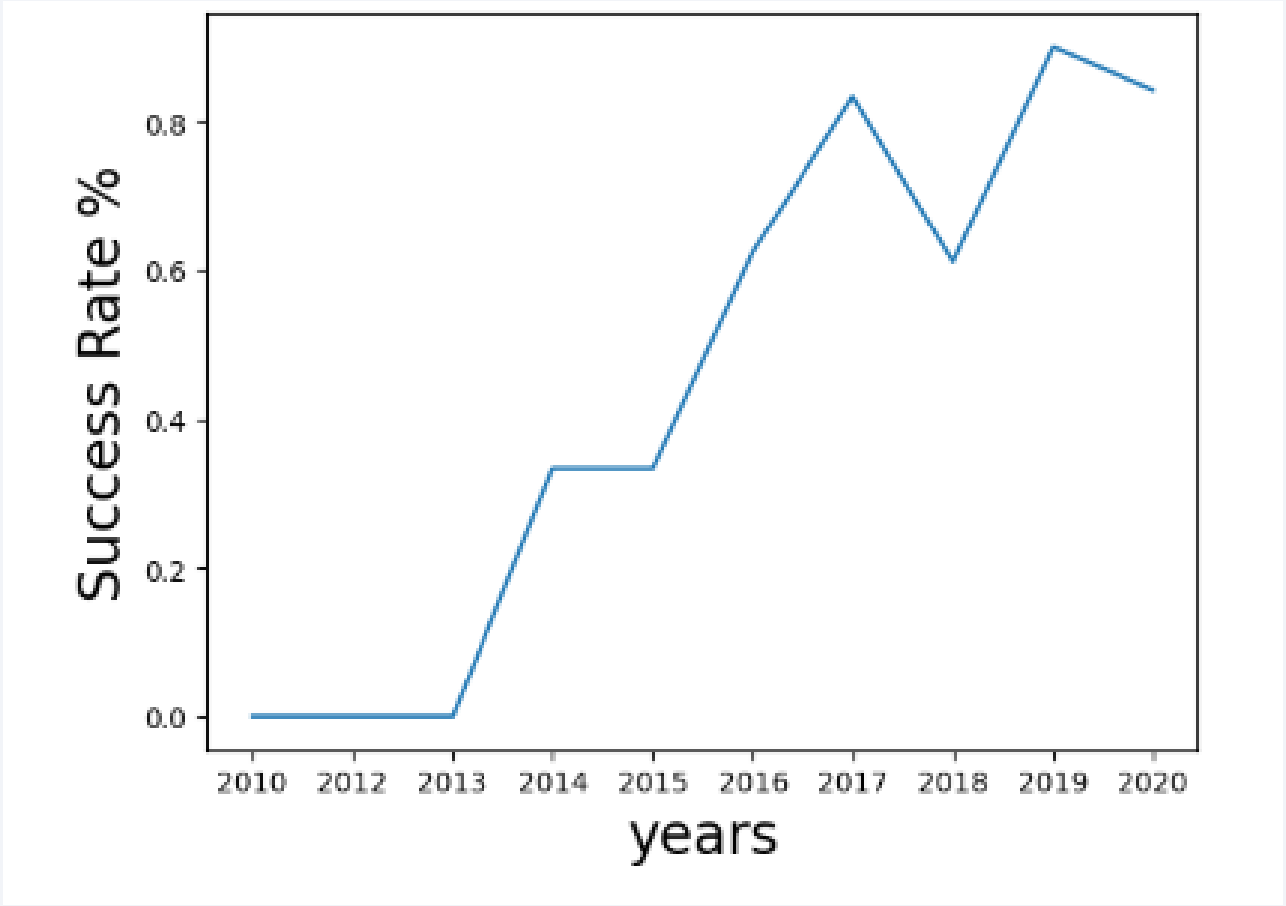
Payload vs. Orbit Type



This graph is consistent with the previous payload comparison. Heavier payloads tend to have a better likelihood of successful landing. This also explains the high success rate of the VLEO orbit considering that is mostly used for heavier payloads. With most orbits this seems to be the trend with the exception of GTO which seems to be unaffected by payload.

Launch Success Yearly Trend

This plot provides a valuable insight and the confirmation of a previous observation. The strongest variable contributing to successful landings is how recent the launch was as there is a clear trend towards better success rate.



All Launch Site Names

Identifying unique landing sites using SQL queries.

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

Launch Site Names Begin with 'CCA'

Every launch on a launch site starting with CCA

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

```
%sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "Customer" = 'NASA (CRS)'
```

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

SUM("PAYLOAD_MASS__KG_")
45596

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "Booster_Version" LIKE 'F9 v1.1%'
```

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

AVG("PAYLOAD_MASS__KG_")
2534.6666666666665

First Successful Ground Landing Date

```
%sql SELECT MIN("Date") FROM SPACEXTBL WHERE "Landing_Outcome" = 'Success (ground pad)'
```

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

MIN("Date")
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT DISTINCT("Booster_Version")  FROM SPACEXTBL WHERE "Landing_Outcome" = 'Success (drone ship)' AND "PAYLOAD_MASS__KG_" < 6000 AND "PAYLOAD_MASS__KG_" > 4000
```

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

Total Number of Successful and Failure Mission Outcomes

Mission_Outcome
Success
Failure (in flight)
Success (payload status unclear)
Success

```
%sql SELECT COUNT(*) FROM SPACEXTBL WHERE "Mission_Outcome" LIKE 'Success%'
```

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

COUNT(*)
100

```
%sql SELECT COUNT(*) FROM SPACEXTBL WHERE "Mission_Outcome" LIKE 'Failure%'
```

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

COUNT(*)
1

Boosters Carried Maximum Payload

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

substr(Date,7,4)	substr(Date, 4, 2)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Mission_Outcome	Landing _Outcome
2015	01	F9 v1.1 B1012	CCAFS LC-40	SpaceX CRS-5	2395	Success	Failure (drone ship)
2015	04	F9 v1.1 B1015	CCAFS LC-40	SpaceX CRS-6	1898	Success	Failure (drone ship)

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

Landing_Outcome
No attempt
Success (ground pad)
Success (ground pad)
Success (drone ship)
Success (ground pad)
Success (drone ship)
Success (drone ship)
Success (ground pad)
Failure (drone ship)
Success (drone ship)
Success (drone ship)
Failure (drone ship)
Failure (drone ship)
Success (ground pad)
Controlled (ocean)
Failure (drone ship)
Precluded (drone ship)
No attempt
Failure (drone ship)

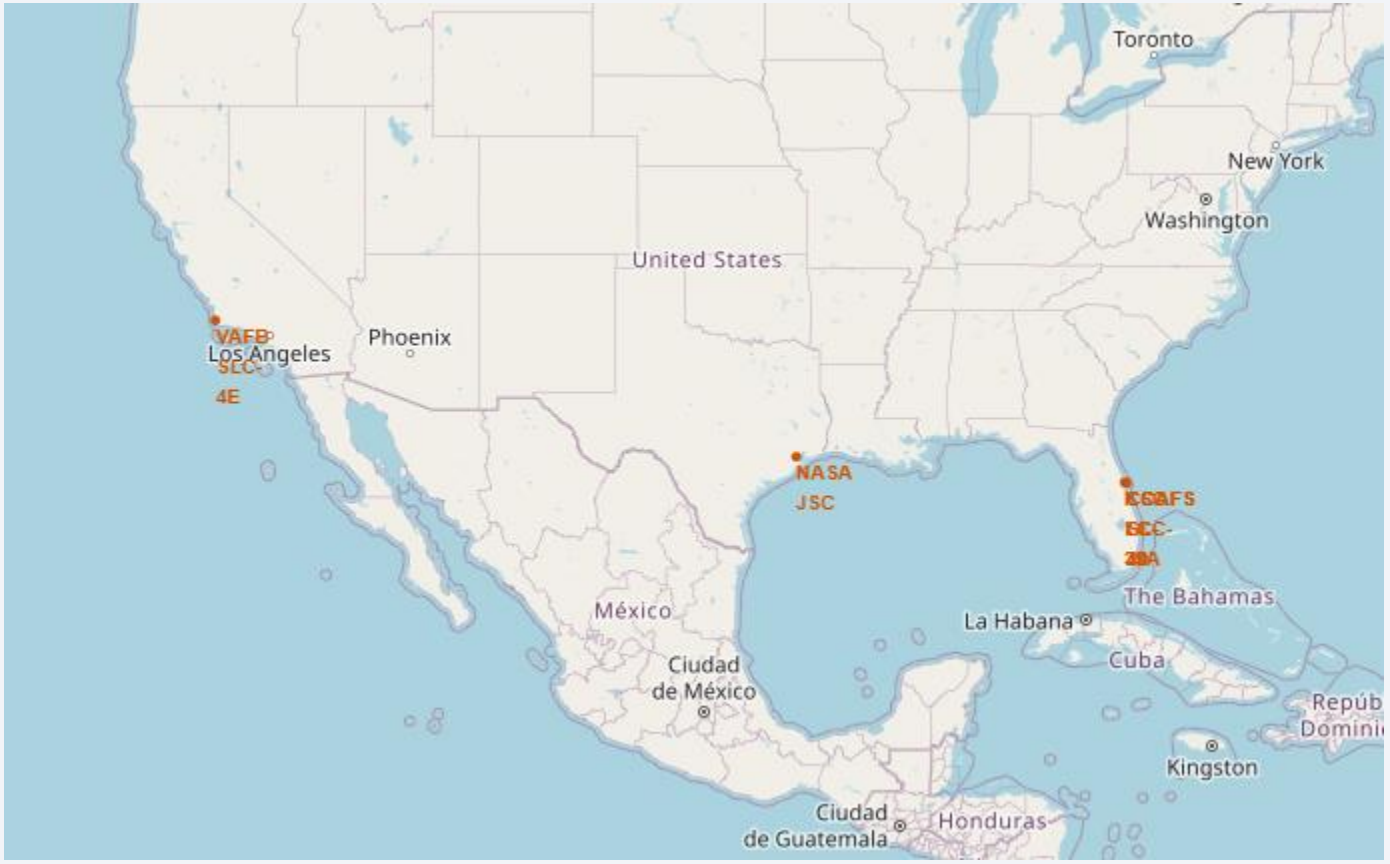
No attempt
Uncontrolled (ocean)
Controlled (ocean)
No attempt
No attempt
No attempt
Controlled (ocean)
Uncontrolled (ocean)
No attempt
No attempt
No attempt
No attempt
Failure (parachute)

Section 3

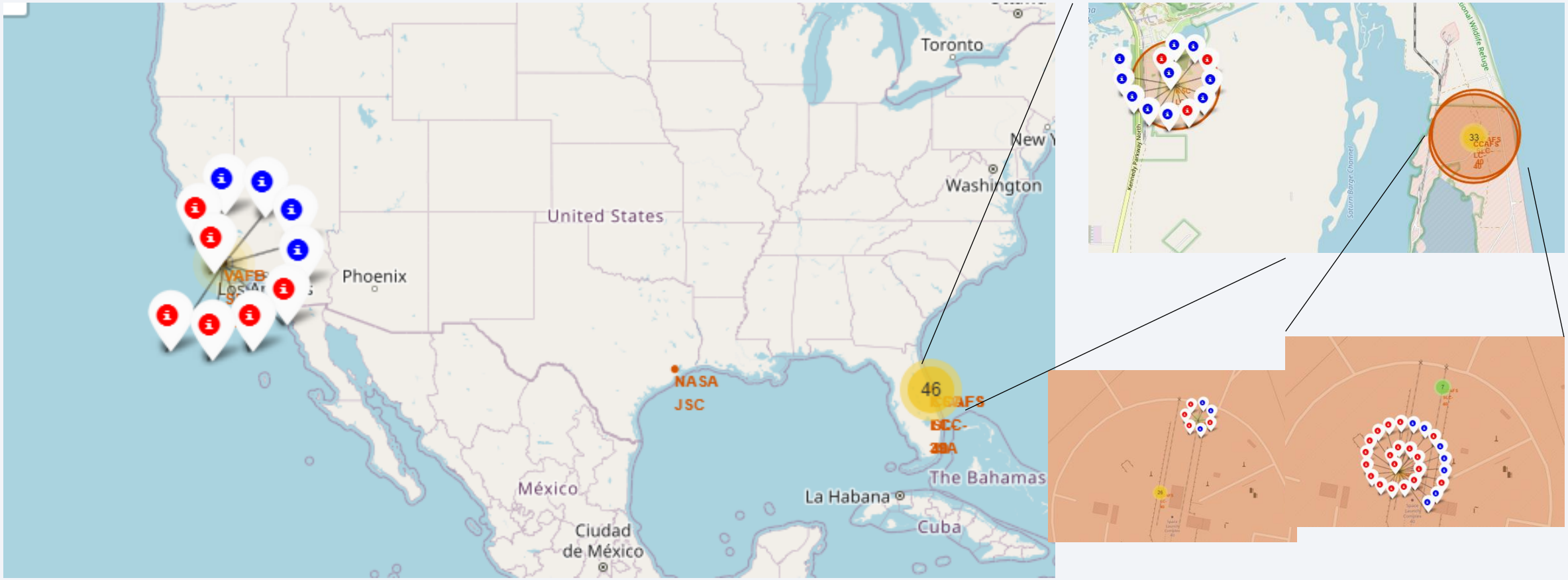
Launch Sites Proximities Analysis

Landing sites map.

The map shows the geographical locations of each of the unique landing sites obtained previously in the EDA with SQL

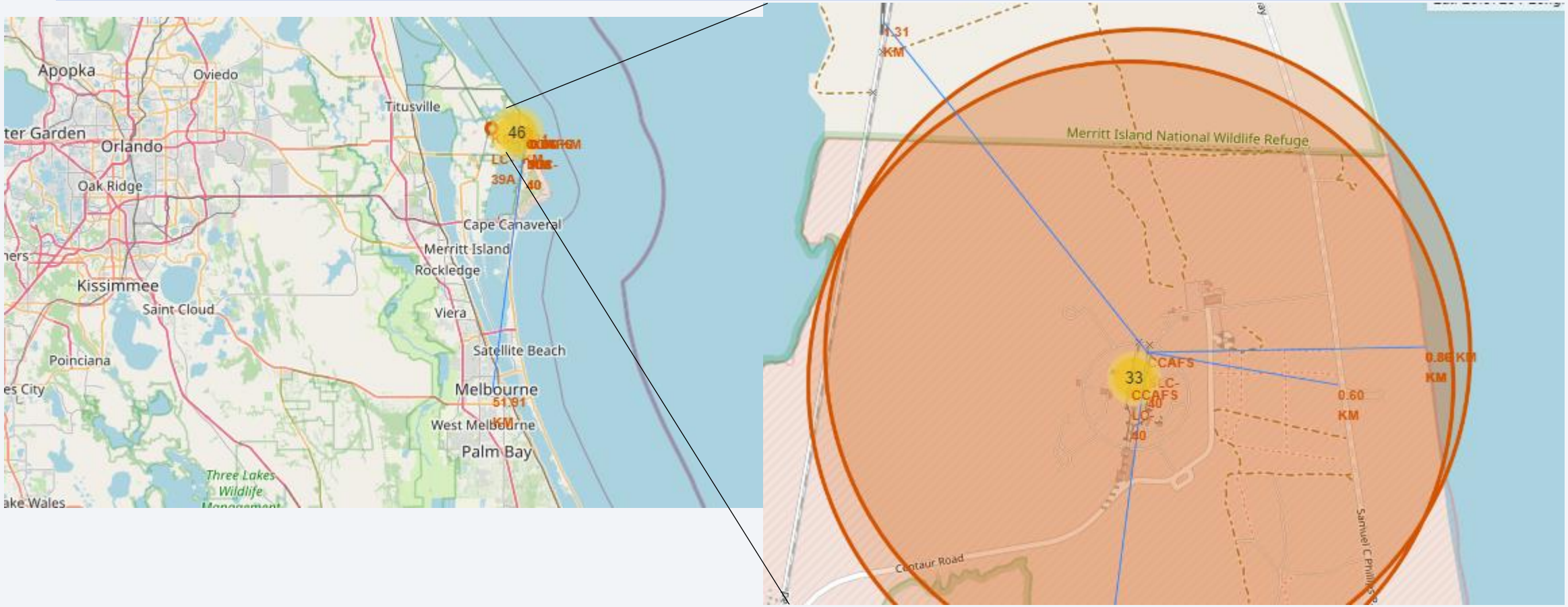


Map labels for each landing and outcome

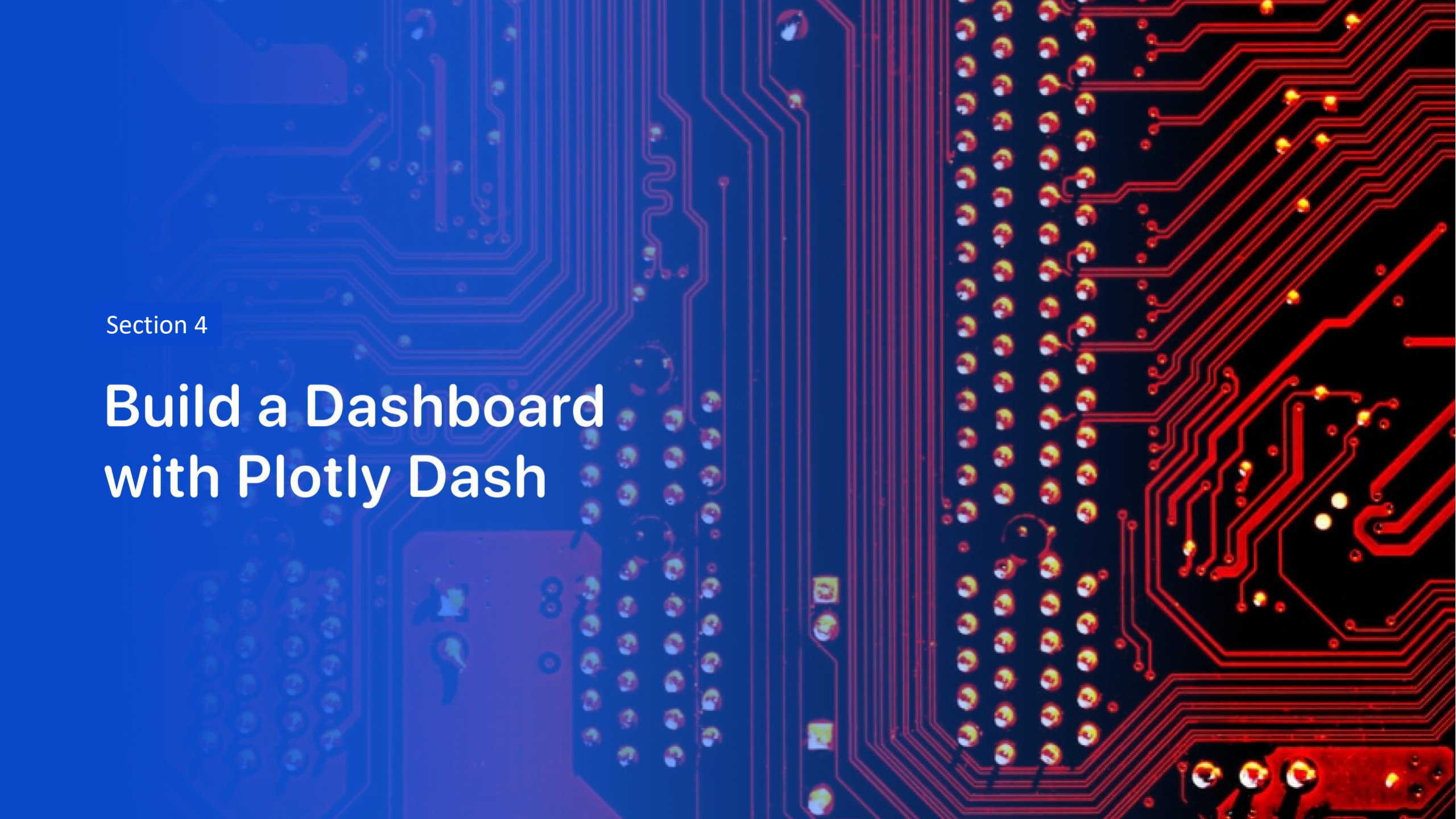


This map shows each landing labeled (blue label means success/red label means failure)

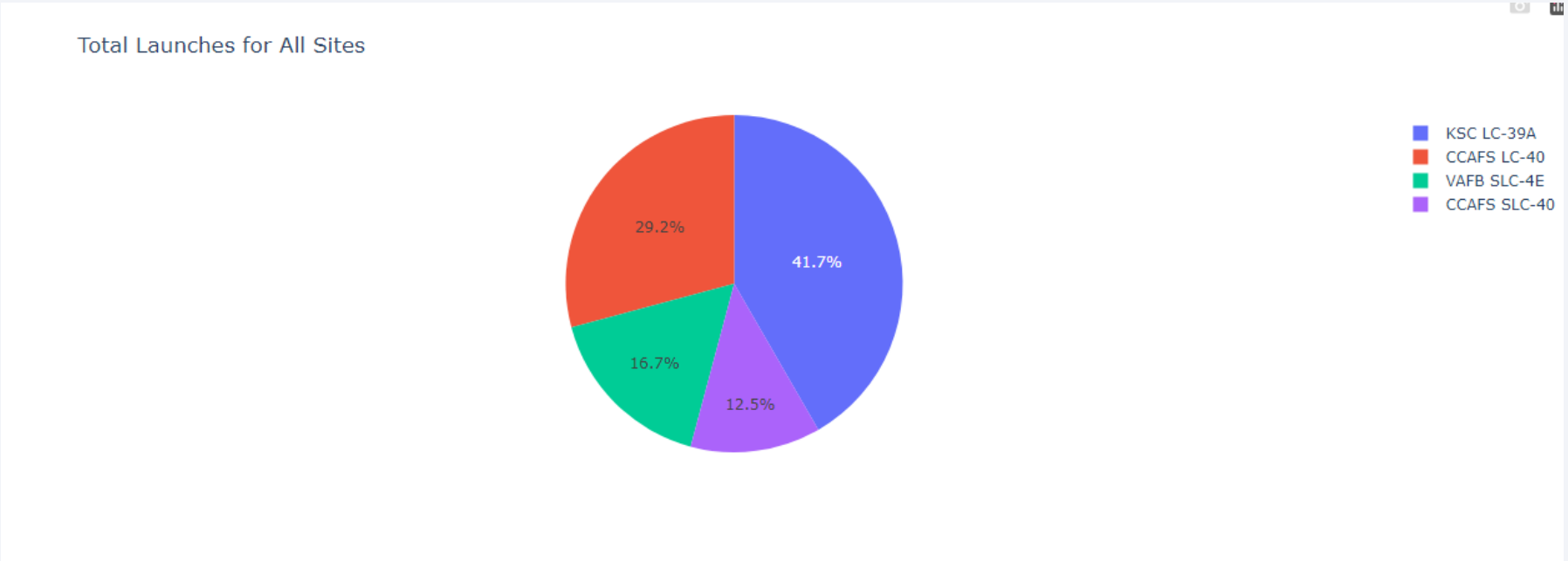
Proximity to places of interest



This map shows the distance of this particular landing sites (CCA) to the nearest road, railway and city.

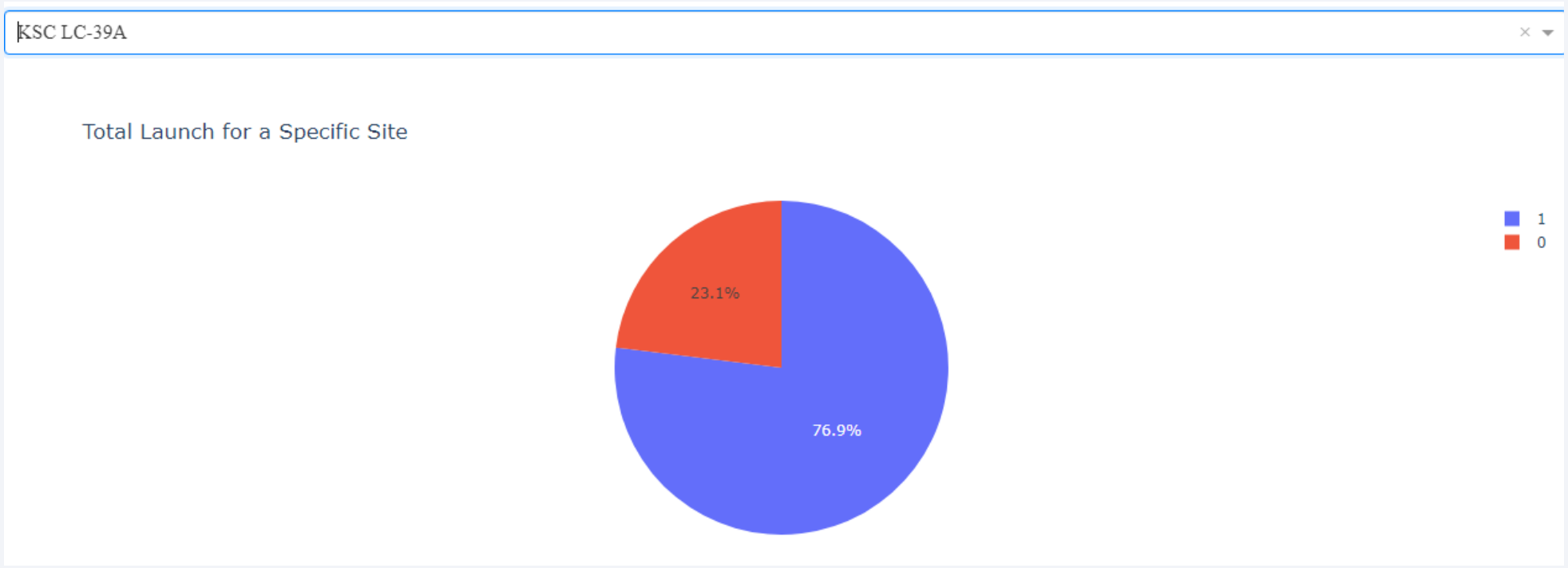

Section 4
Build a Dashboard with Plotly Dash

Dahsboard pie chart: Total launches



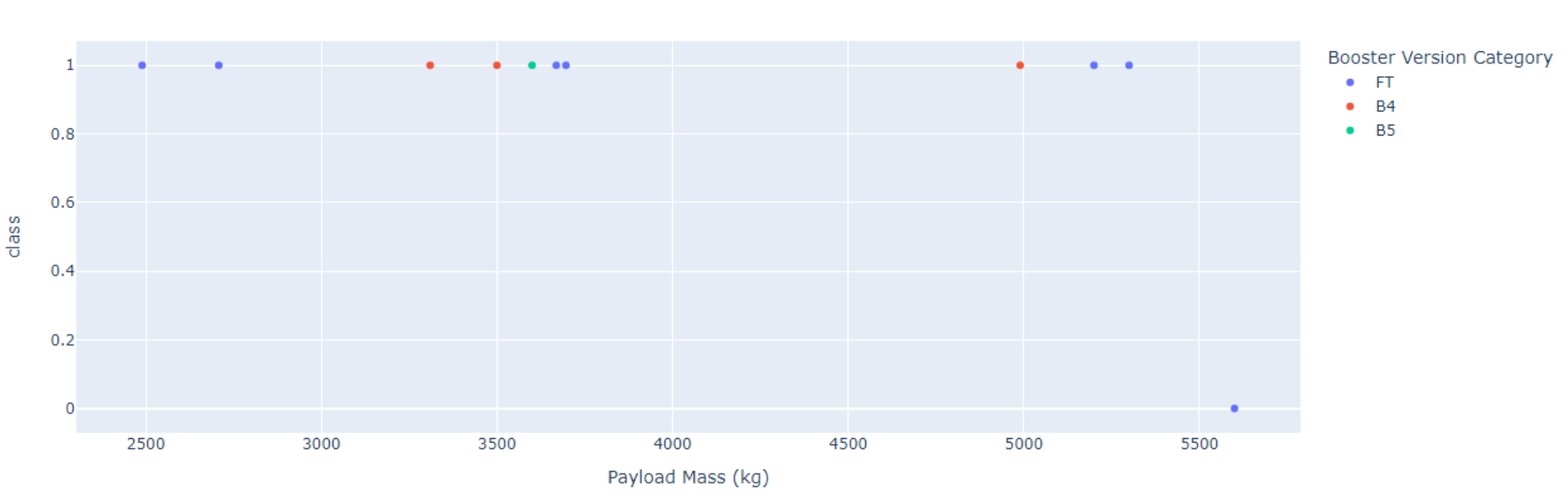
This dashboard created with plotly dash shows the proportion of launches for each site. We can see that the most frequent landing site is KSC LC while the least frequent CCAFS SLC-40

Dashboard pie chart: Highest success rate site



This plot created using a dropdown menu shows the success rate for different sites. In this screenshot we can see the highest rate of success for a site which corresponds to KSC LC which is also the site with most launches as seen previously

Slider chart: Payload vs Class vs Booster Version



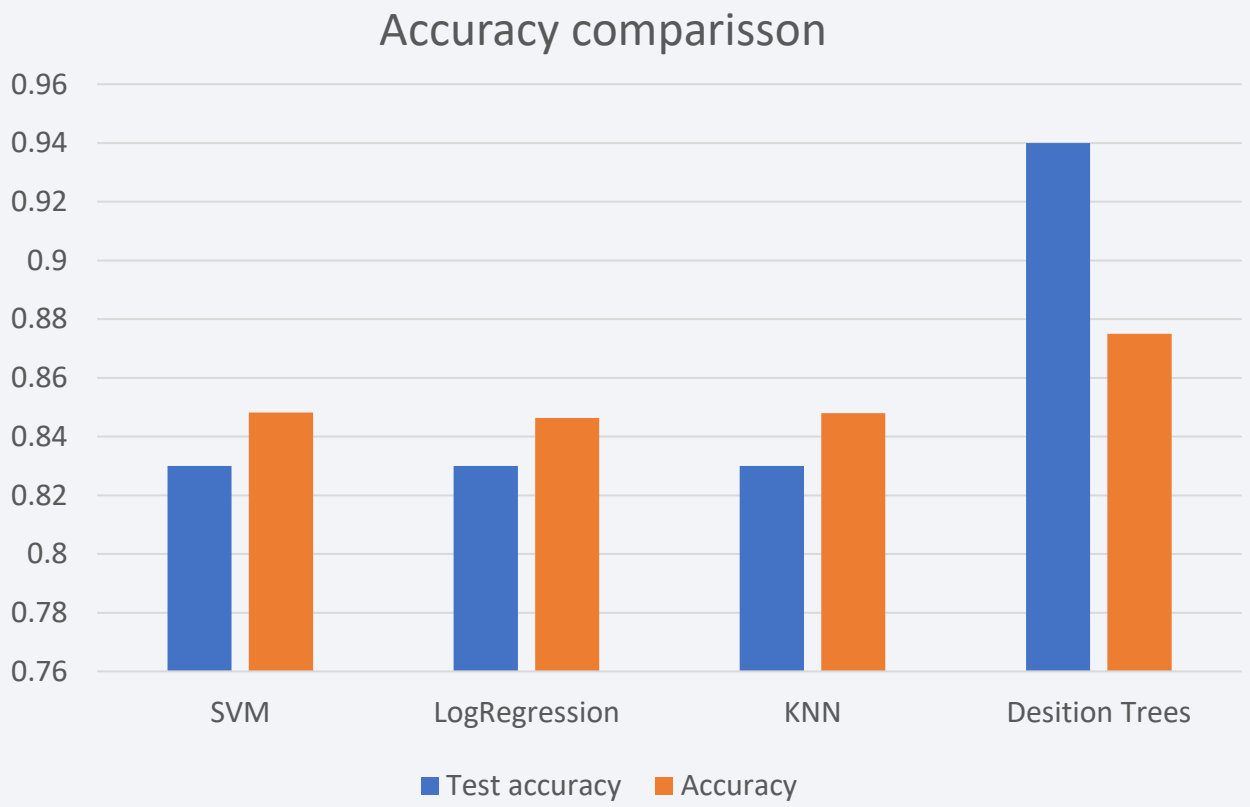
This chart shows information about the relation between Class (success/failure) and Payload for the highest rated site KSC LC. We can see that most successful launces where in the range of 2500-5500 kg payload and the booster FT the most common

Section 5

Predictive Analysis (Classification)

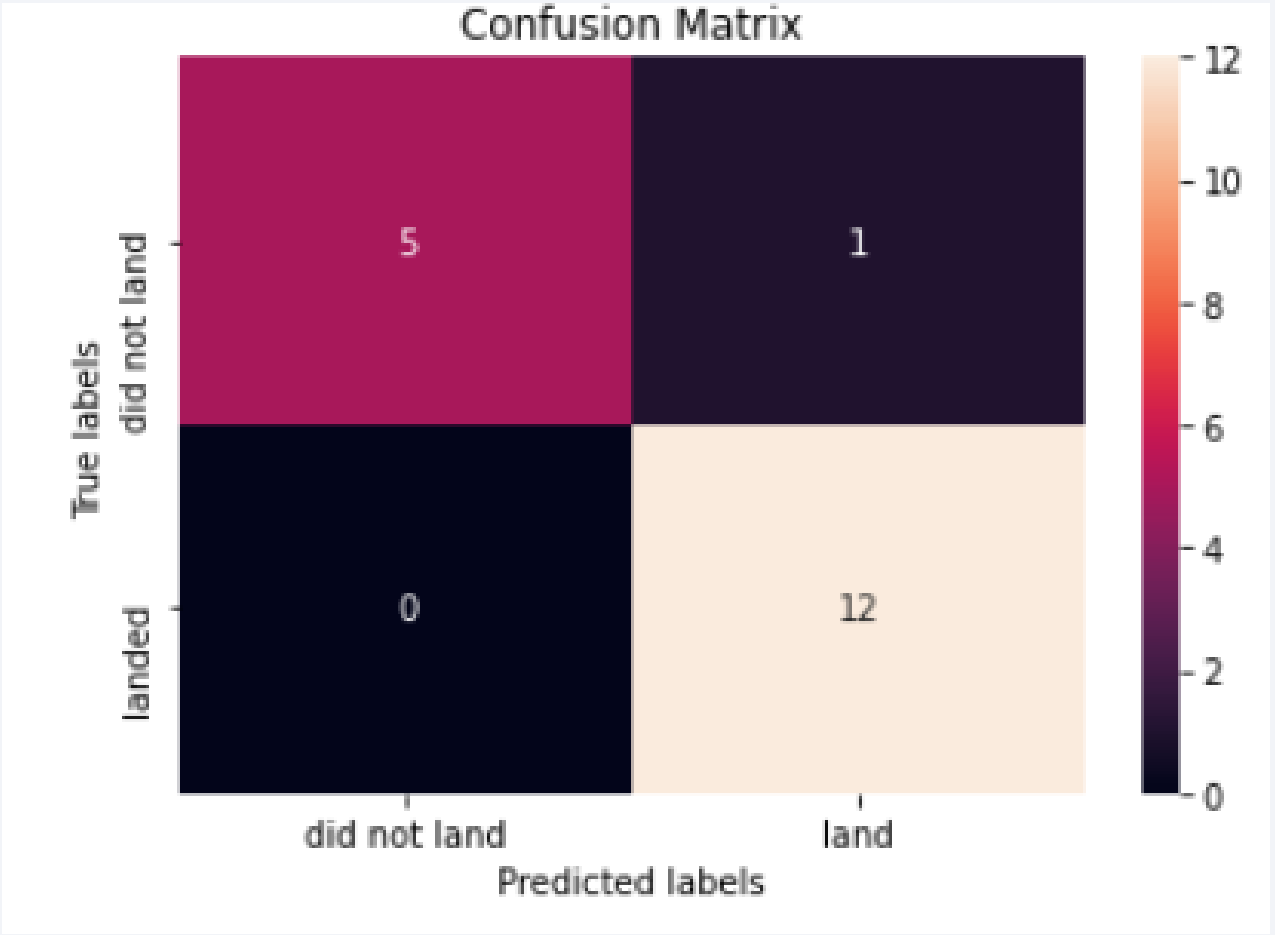
Classification Accuracy

- For the predictive analysis we tried different models to test for accuracy in prediction. Performance among them was pretty much similar with the exception of the model using desition trees which performed better in both the test and train set



Confusion Matrix

In the confusion Matrix for the decision trees model shows that there is a slight tendency for false positives. This is something that was also present and was worse for the other models.



Conclusions

- Returning to the original question posed at the beginning we found that successful landing of the rockets can be reasonably be predicted with the available data.
- Success rate for landings have been improving with the years.
- Heavier payloads tend to yield more successful landings
- Some orbits have better success rate but this is due to them being used little compared to others. For the orbits that were used the most, success rate seems to be regular. Which means that orbit might not be such a strong predictor.
- KSC LC landing site has the highest proportion of launches and also the highest success rate.

Appendix

- The code used for the dashboards application can be found in this link

Thank you!