IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Himmler Benitez
April, 2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

Methodologies

- Data was obtained from different sources and methods:
  - Directly from the Space X API
  - Web Scraping from Wikipedia
- After the Exploratory Data Analysis, the raw data got processing and standardization, and the application of this methodologies:
  - Data wrangling,
  - Data visualization
  - Interactive dashboard
- Finally, a Machine Learning Prediction model was implemented

Results

- ES-L1, GEO, HEO and SSO orbits were the ones with the highest success rates.
- SO orbit has a success rate of zero
- There is a correlation between PayloadMass and success rate which seems to favor the heavier payloadMass
- Decision tree method was the most adequate model to predict the outcomes of the landings

Introduction

Space Y is a new company that intends to compete with space X and is investigating to answers the following problems:

- What are the launches with the best successful rate based in the types of boosters and payloads mass?
- What is the launch site with the best successful rate?
- What is the best approach and method to predict the outcomes of the landings using Machine Learning?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- 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

- Data was collected through 2 methods:
  - API Request: https://api.spacexdata.com/v4/rockets/
  - Web Scraping: https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches

Data Collection – SpaceX API

The data was collected from the SpaceX public API by doing direct request to the public API.

- Link

Data Collection - Scraping

- SpaceX launches data is obtained from Wikipedia through web scraping.

- GitHub URL

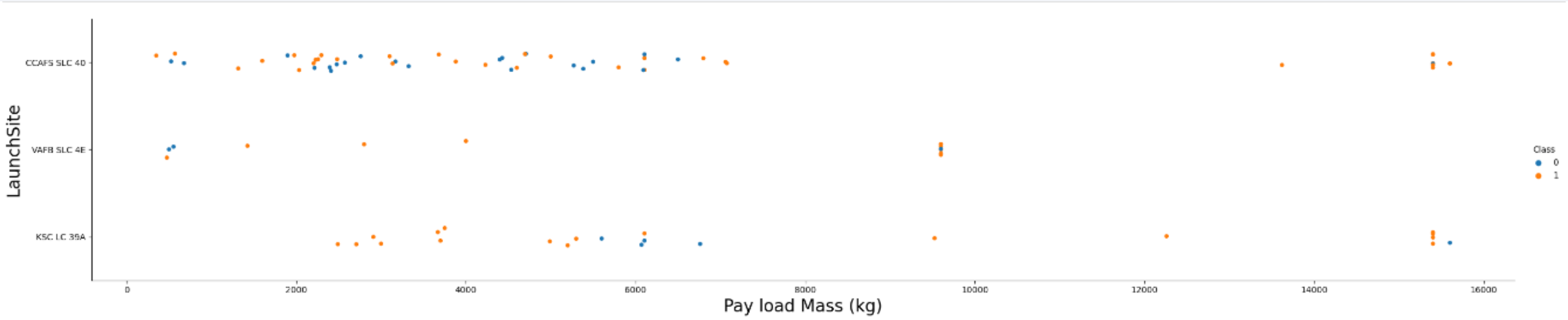
Data Wrangling

- Data was obtained though a CSV file and then analyzed.
- Data Wrangling process:

- GitHub URL

EDA with Data Visualization

- The scatterplots and barplots charts plotted were used to explore the relationship between features.
  - FlightNumber vs. PayloadMass
  - FlightNumber vs LaunchSite



- [GitHub URL]

EDA with SQL

- Performed SQL queries:

  - The names of the launch sites
  - Top 5 records where launch sites begin with the string 'CCA'
  - The total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - The date when the first successful landing outcome in ground pad was achieved
  - The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - The total number of successful and failure mission outcomes
  - The names of the booster_versions which have carried the maximum payload mass
  - The records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015
  - The count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.
  - GitHub URL

Build an Interactive Map with Folium

- Added map objects such as markers, circles, and lines in the folium map
  - Markers to indicate the launch sites
  - Circles to highlight coordinates of the launch sites
  - Color markers to indicate success/failed launches
  - Lines to indicate distance to the nearest shore and coast

- GitHub URL

Build a Dashboard with Plotly Dash

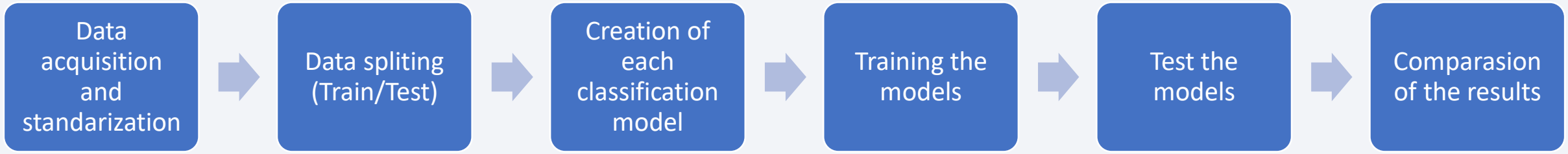
- Graphs and plots were used to visualize and explore the date
  - Percentage of launches per site
  - Payload range

With those resources, the data can be analyzed in an easy and relevant way to identify the most appropriate launch site based in the payloads.

GitHub URL

Predictive Analysis (Classification)

- Data was standardized and split to use it for training and testing the four classification models (Logistic regression, support vector machine, decision tree and K nearest neighbors).



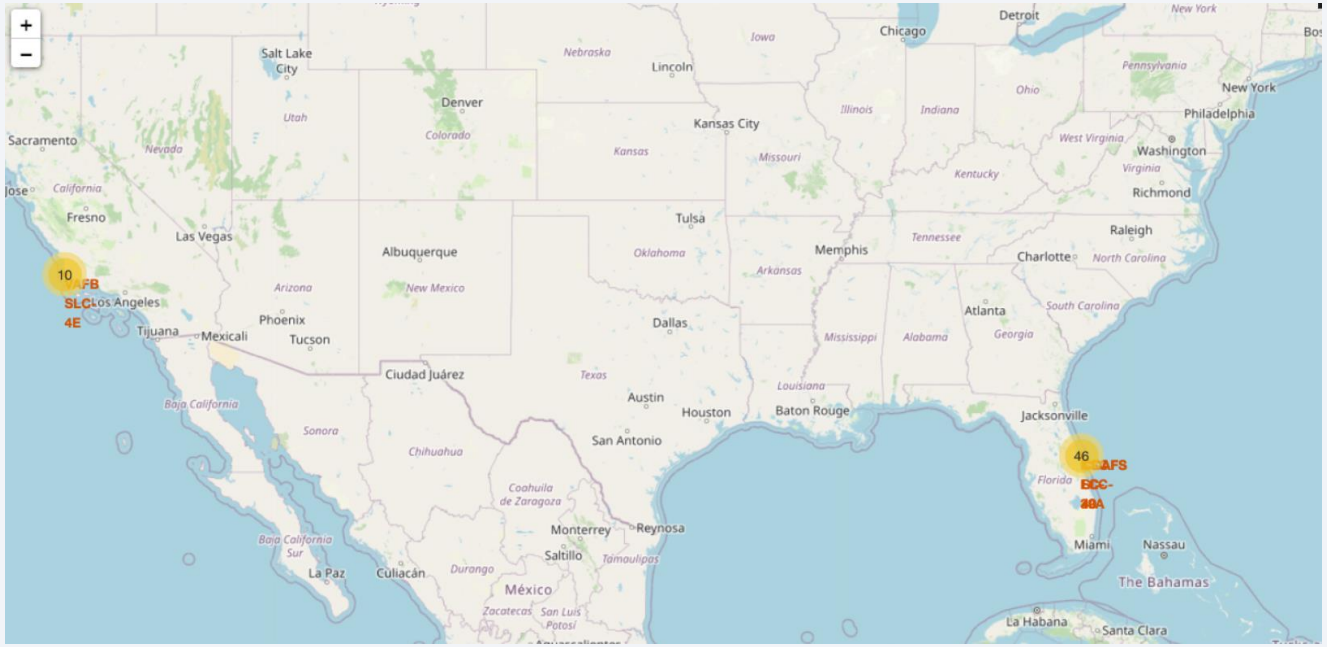
- GitHub URL

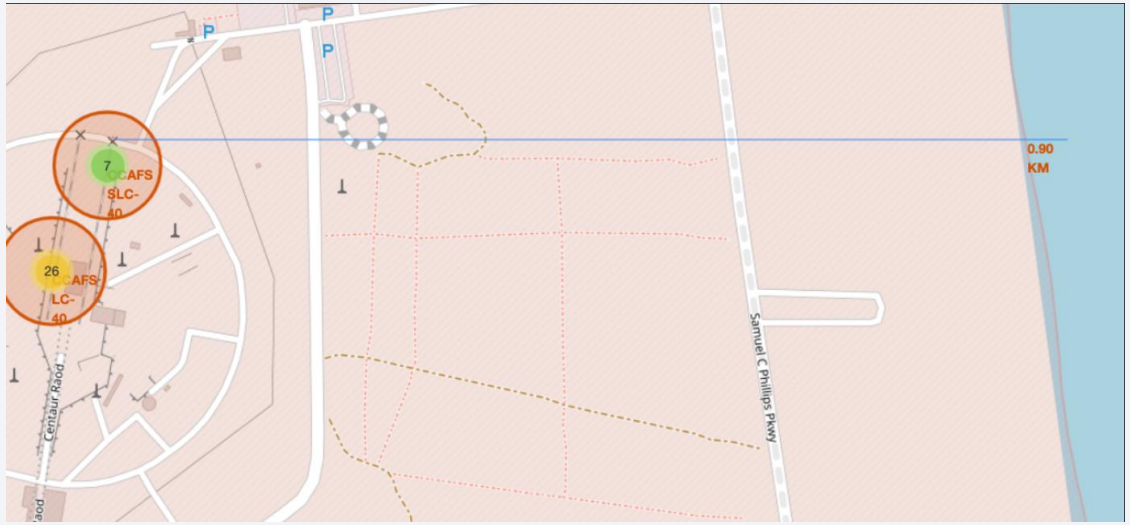
Results

- Exploratory data analysis results
  - There are 4 different launch sites for Space X launches
  - CCAFS SLC 40 is the most used launch site
  - The first success landing was in 2015
  - The average payload of F9 v1.1 booster is 2,928 kg
  - The total payload mass carried by boosters launched by NASA (CRS) was 111268 Kg
  - 98 of out 101 launches were successful

Results

- Launch sites were located in the maps with interactive analytics. Finding that the majority of the launches were in the east cost.

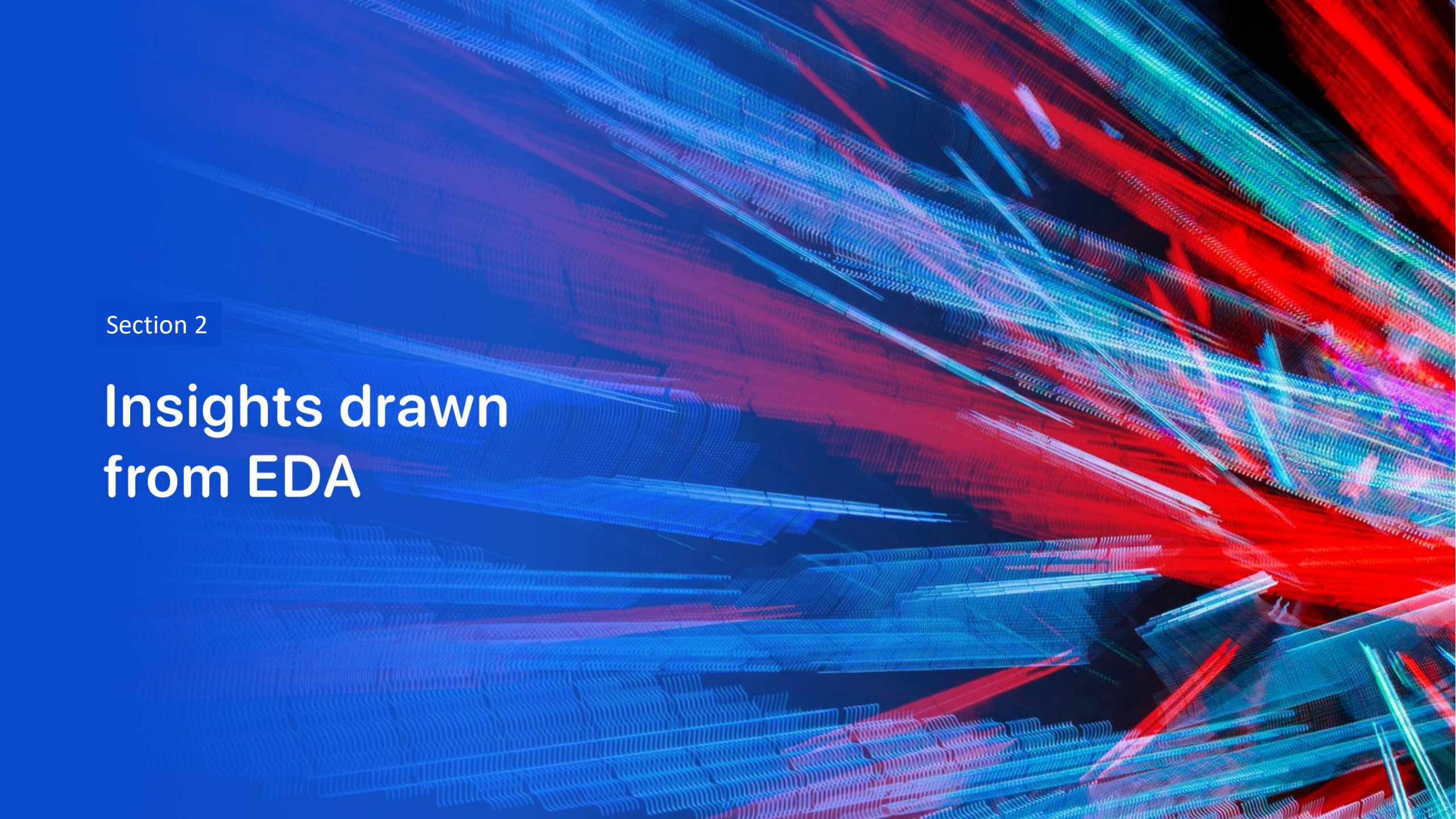




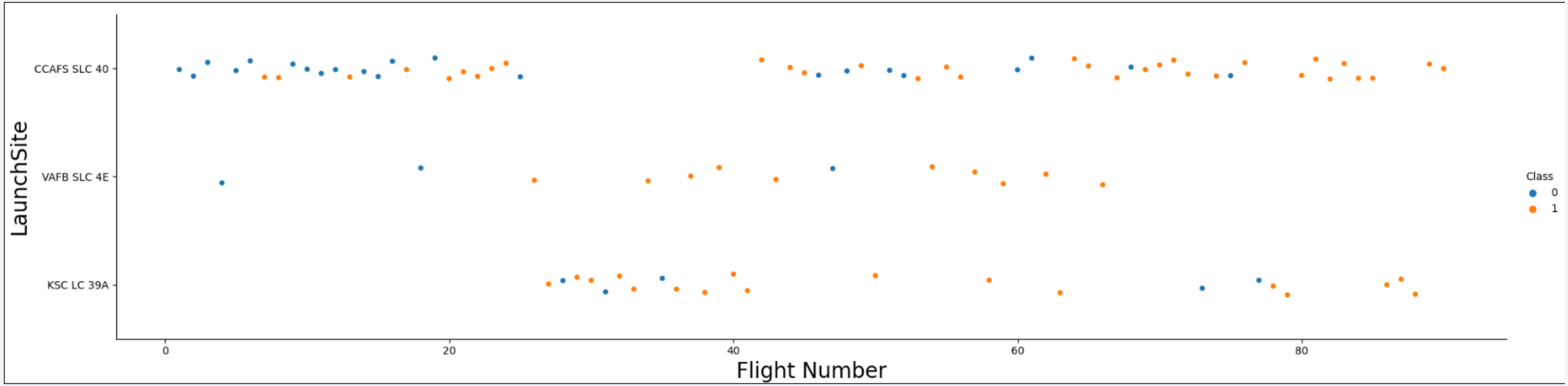
Results

- Predictive analysis results

Decision tree model was the best model of the 4 with an accuracy of 88%.

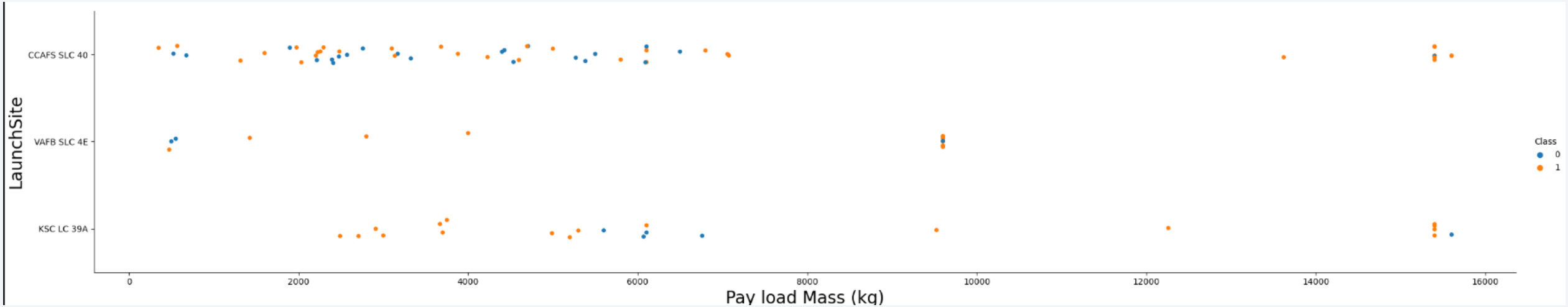

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



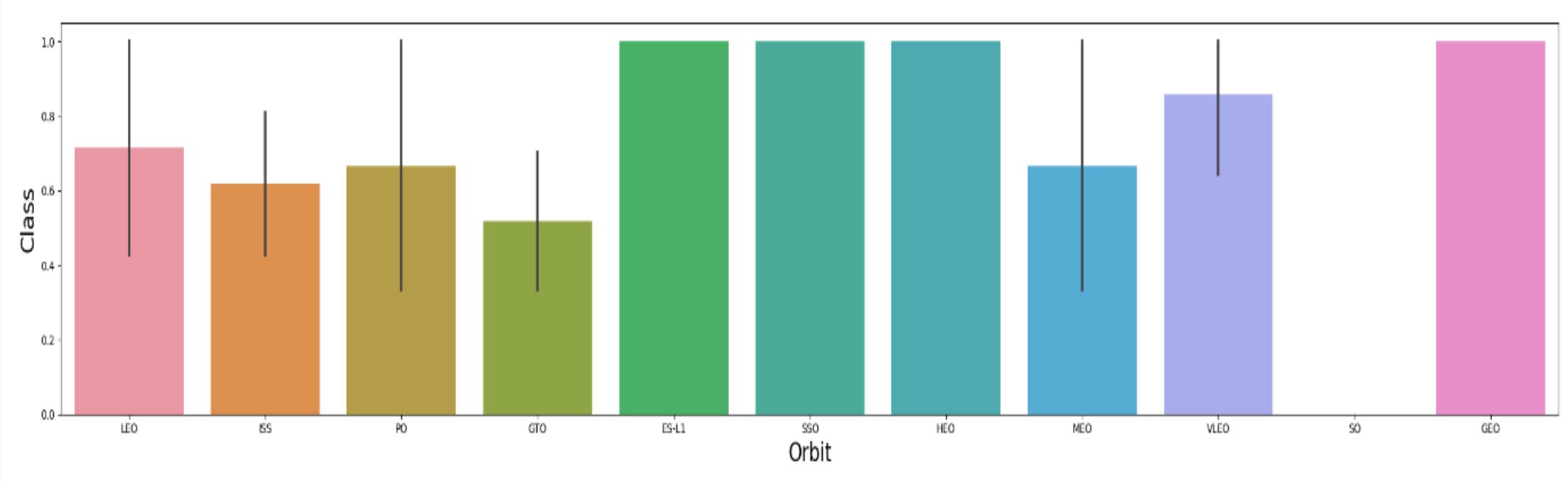
- Based on the plot, CCAF5 SLC 40 launch site is the most used for the launches and the one with the major successful launches.
- VAFB SLC 4E is the launch site with the best success ratio based in the total of launches in there

Payload vs. Launch Site



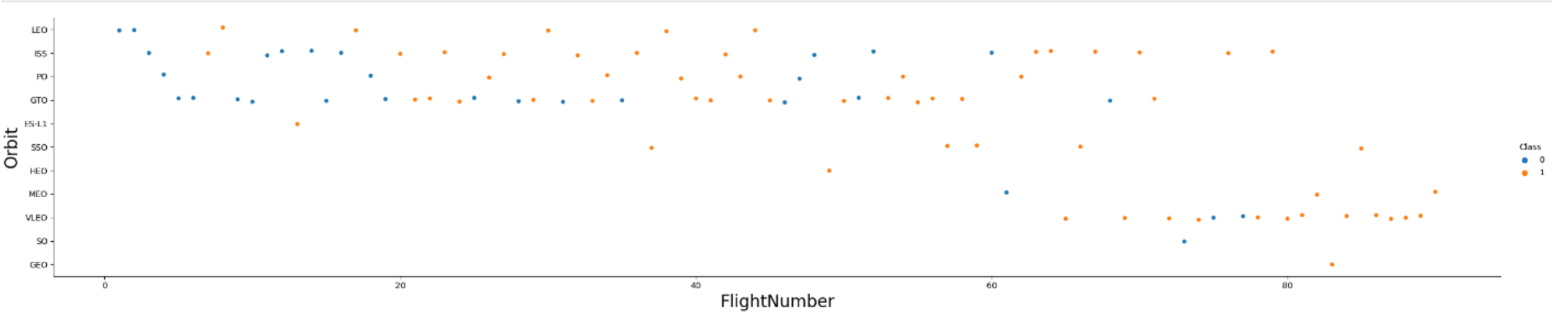
- The success ratio for payloads over 8000kg is outstanding
- VAFB SLC 4E seems not being suitable, or at least the last option, for heavy payloads
- < 8000 kg payloads were the most frequent launches

Success Rate vs. Orbit Type



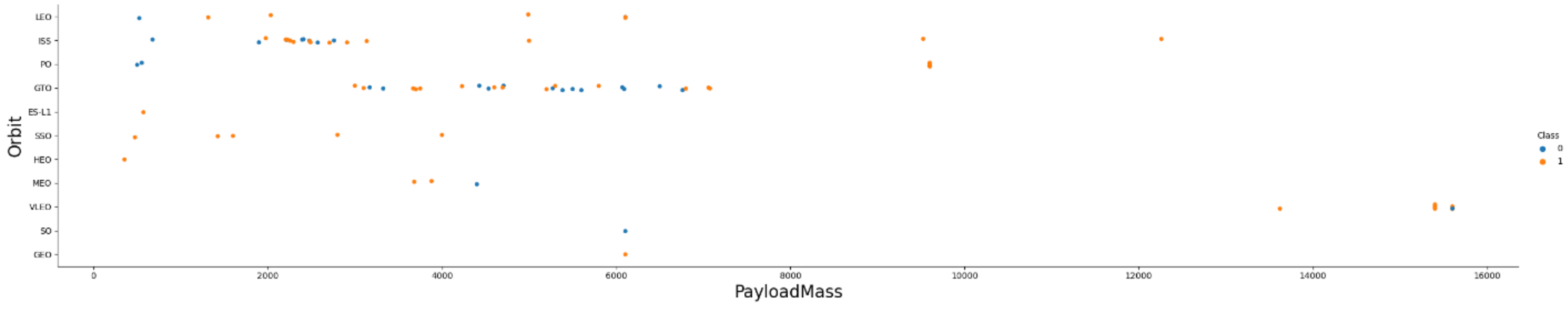
- ES-L1, GEO, HEO and SSO orbits were the ones with the highest success rates.
- SO orbit has a success rate of zero

Flight Number vs. Orbit Type



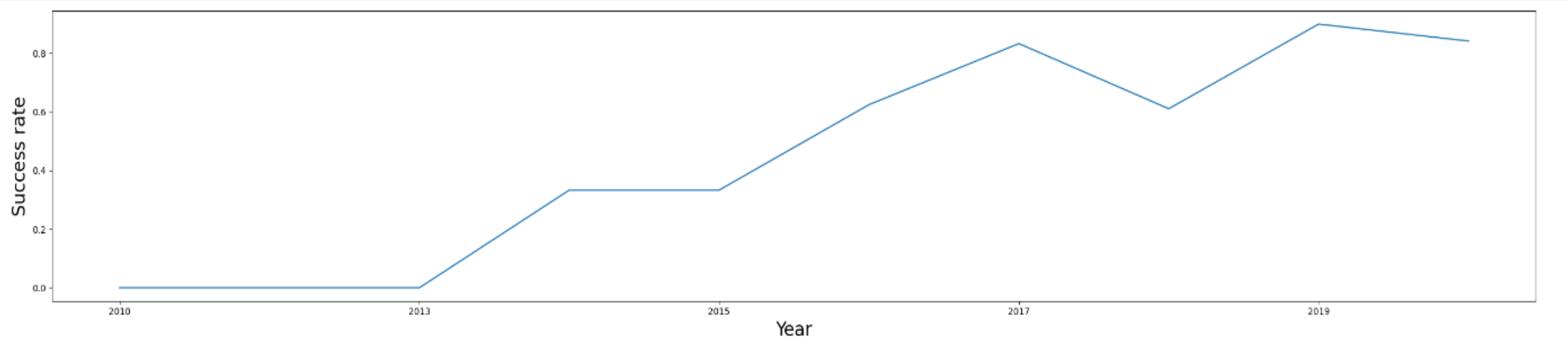
- Success rate has improved over time
- SO failed its unique flight

Payload vs. Orbit Type



- VLEO orbits has only been flight by high payload mass

Launch Success Yearly Trend



- Success rate improved over time
- In 2018, the success rate tendency went down but it recovered next year

All Launch Site Names

- The query allowed to get all the unique launch site names

```
In [22]: %sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL
 * sqlite:///my_data1.db
Done.
```

Out[22]:

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

Launch Site Names Begin with 'CCA'

Task 2

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

```
[9]: %sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

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

[9]:

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

Five simples of CCAFS LC-40 launch site

Total Payload Mass

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

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE payload LIKE '%CRS%'
```

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

SUM(PAYLOAD_MASS__KG_)
111268

The keyword to sum all the payloads is “CRS” giving a result of 111,268 Kg

Average Payload Mass by F9 v1.1

Task 4

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

```
[11]: %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version LIKE '%F9 v1.1%'
```

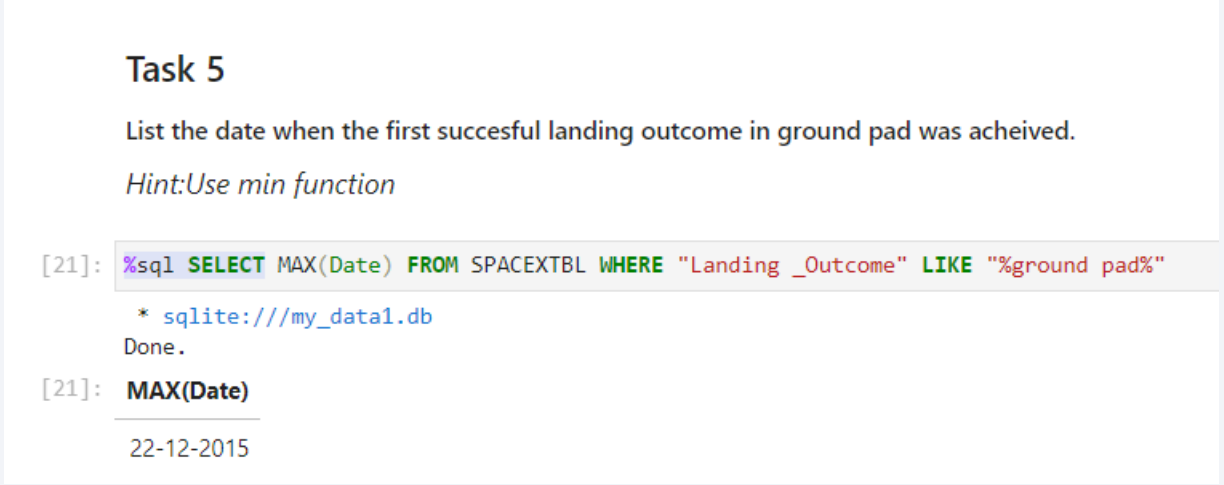
```
 * sqlite:///my_data1.db
Done.
```

[11]:

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

- The query consult all the record where the Booster Version contains F9 v1.1 and calculate the Average for the payload

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

```
[21]: %sql SELECT MAX(Date) FROM SPACEXTBL WHERE "Landing _Outcome" LIKE "%ground pad%"
```

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

[21]:

MAX(Date)
22-12-2015

- The first landing was in Dec 2015. The query filters the data with the function MAX

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

```
3]: %sql SELECT Booster_Version FROM SPACEXTBL WHERE "Landing _Outcome" = "Success (drone ship)" and PAYLOAD_MASS__KG_ >4000 and PAYLOAD_MASS__KG_<6000
```

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

3]:

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

- All the boosters with a success outcome and payload between 4000 and 6000 are displayed
- Two conditions are stablished

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

```
[4]: %sql SELECT COUNT(Mission_Outcome) as "Total" , Mission_Outcome FROM SPACEXTBL group by Mission_Outcome
```

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

[4]:

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

- To calculate the total number of the outcomes, the data was grouped

Boosters Carried Maximum Payload

Task 8

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

```
5]: %sql SELECT DISTINCT(Booster_Version) FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

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

5]:

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

- All these boosters carried the maximum payload mass

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.

```
%sql SELECT substr(Date,4,2) as "Month", "Landing _Outcome", Booster_Version, Launch_Site FROM SPACEXTBL WHERE "Landing _Outcome" = "Failure (drone ship)" and substr(Date,7,4)
```

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

Month	Landing _Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Only two landing failed in 2015

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.

```
17]: %sql select COUNT("Landing _Outcome"), "Landing _Outcome", Date from SPACEXTBL GROUP BY "Landing _Outcome" HAVING "Landing _Outcome" LIKE '%Success%' AND Date between '04-06-2
```

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

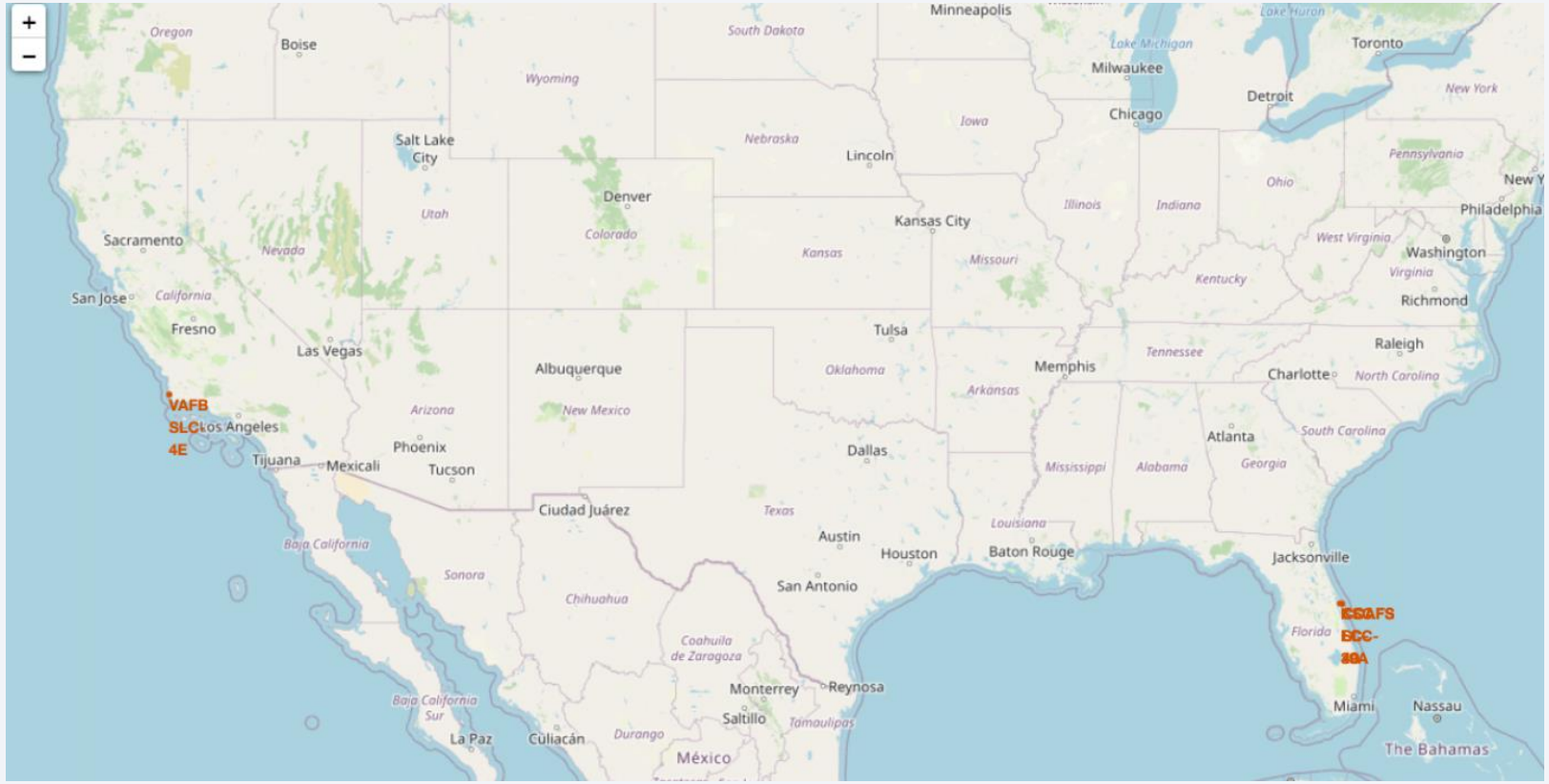
17]:

COUNT("Landing _Outcome")	Landing _Outcome	Date
14	Success (drone ship)	08-04-2016

- 14 successful landing outcomes were registed in the given timeframe

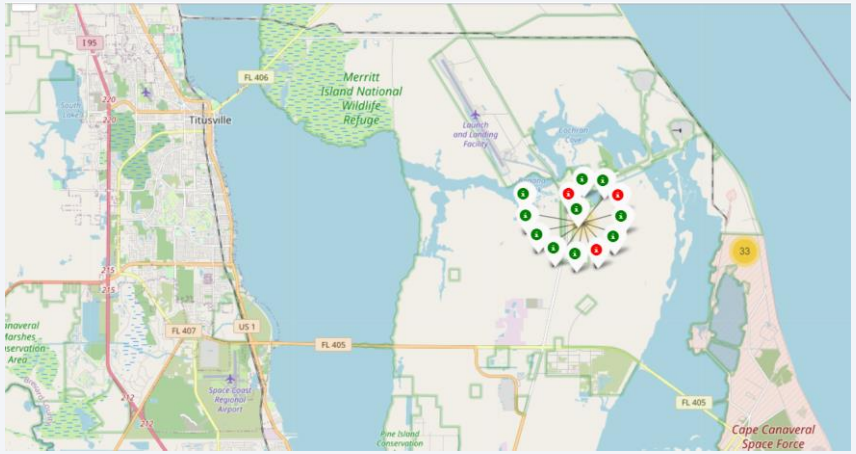

Section 3
Launch Sites
Proximities Analysis

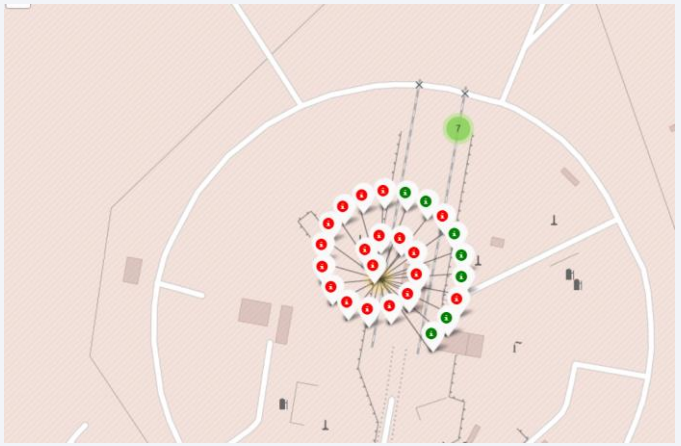
All launch sites in the global map



- The launch sites are place on the west and east coast of US

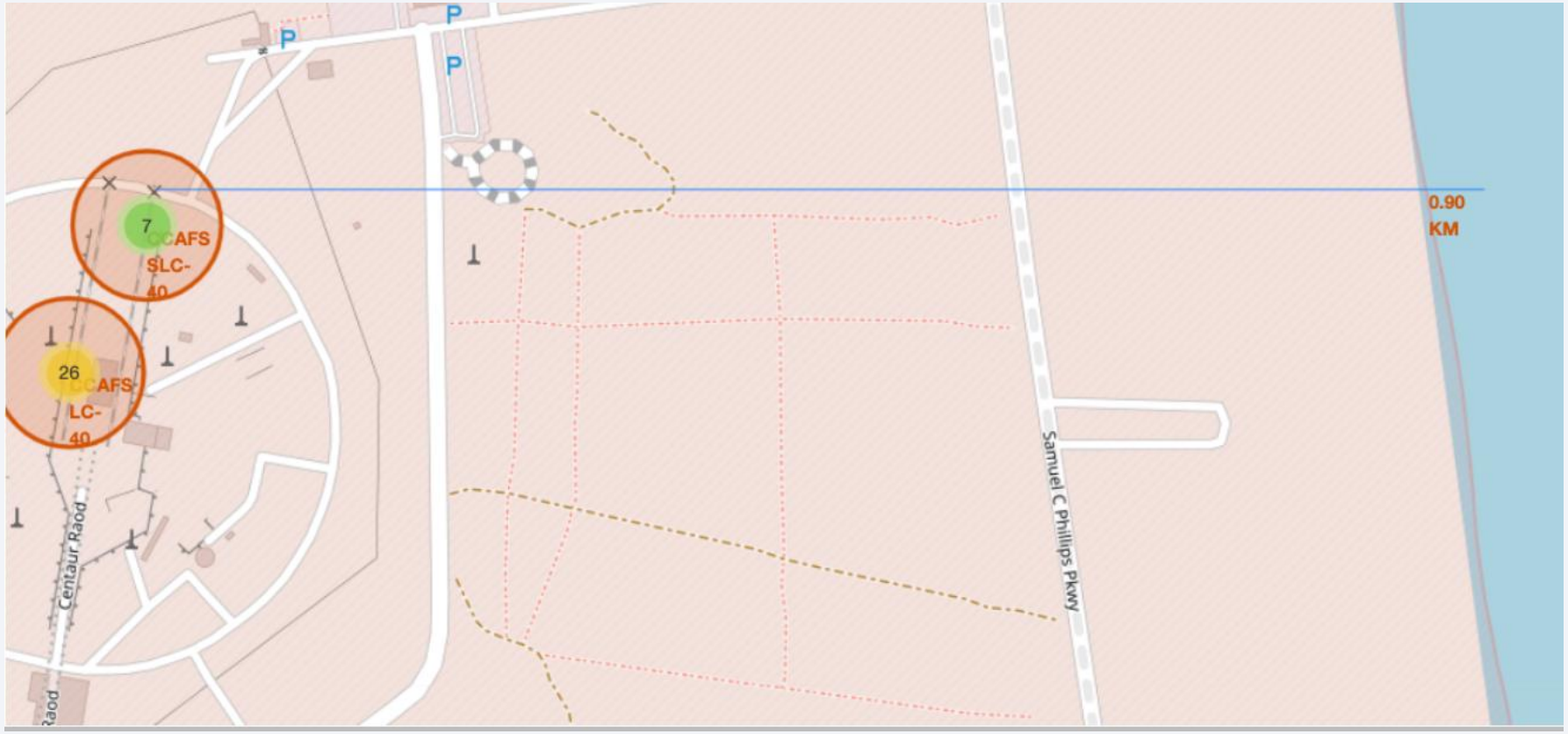
Color-labeled by launch outcomes



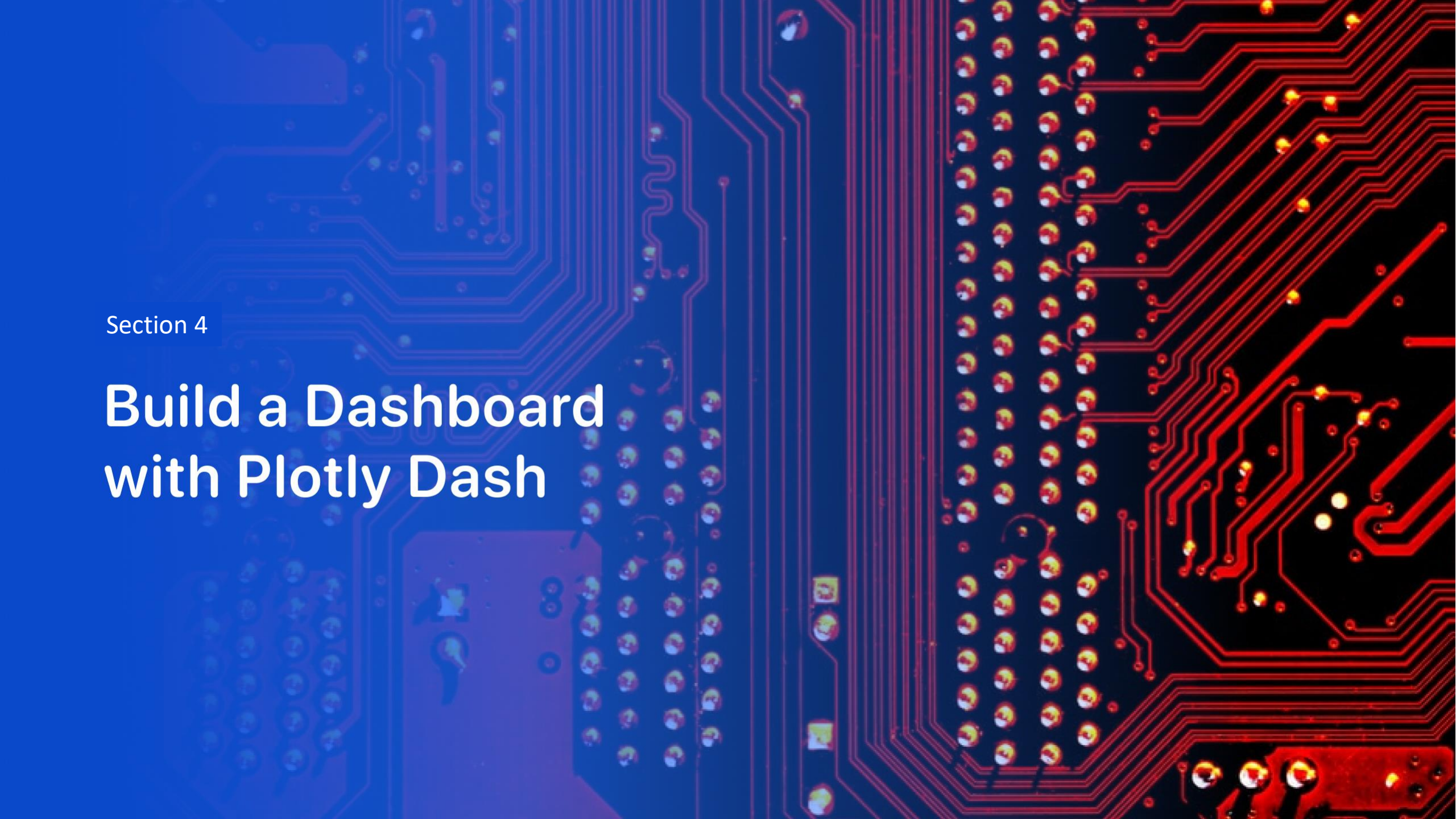


- The green labels are successful launches
- The red labels are unsuccessful launches

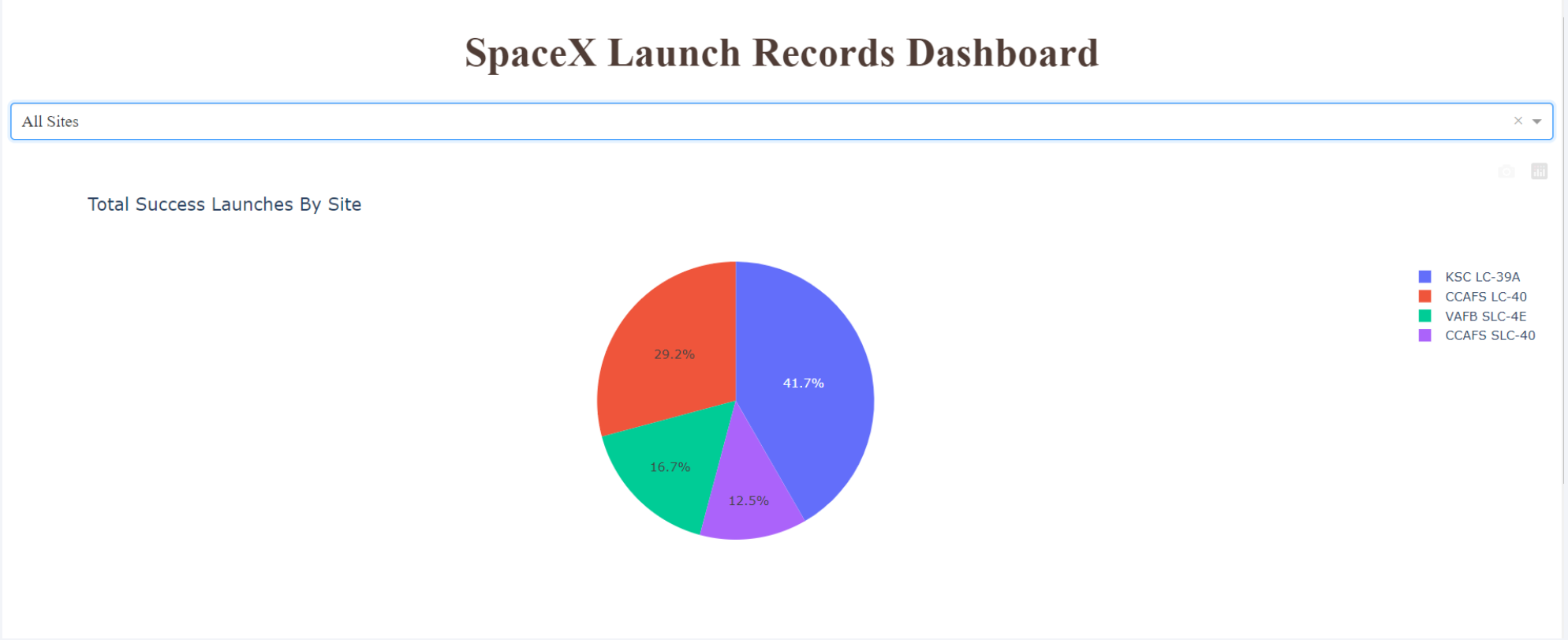
<Folium Map Screenshot 3>



- CCAFS SLC-40 is less than 1 km near the beach

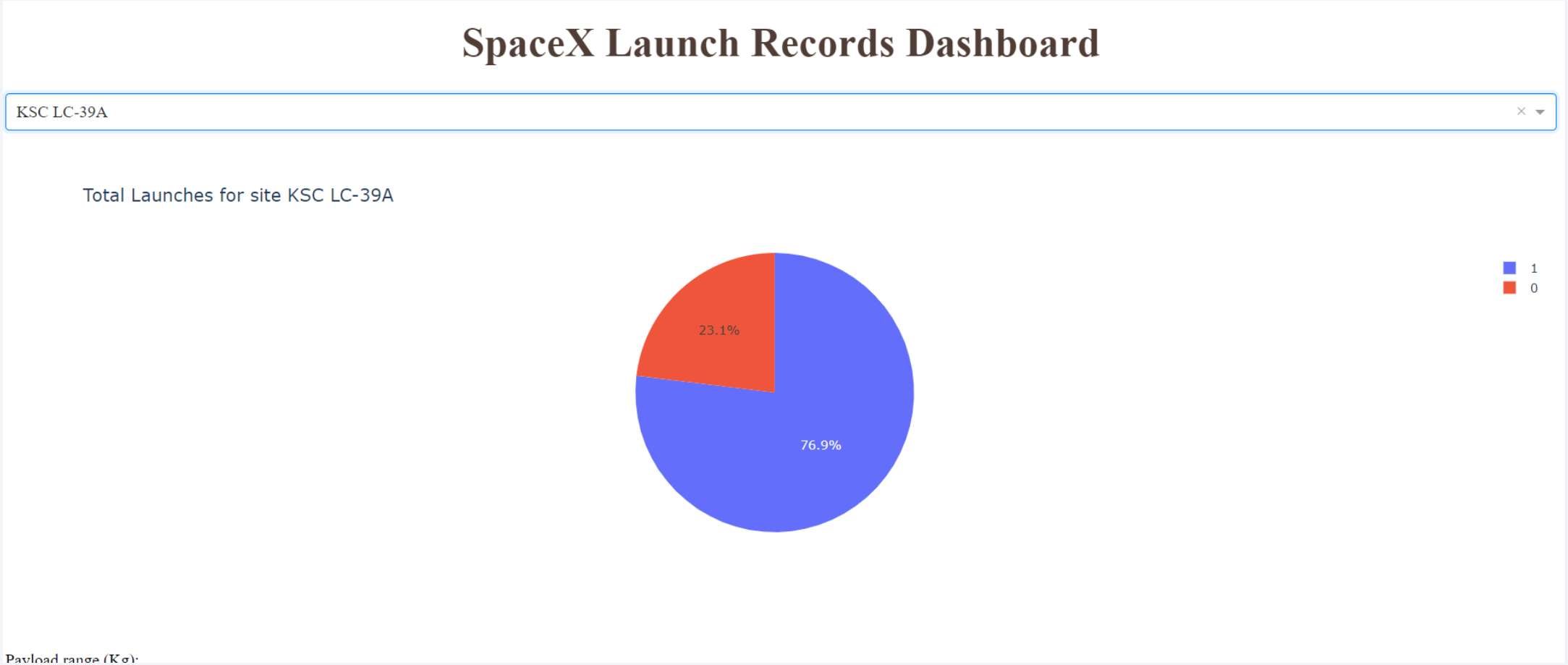

Section 4
Build a Dashboard with Plotly Dash

Successful launches ratio by sites



- KSC LC-39A has the best success ratio for launches of all the launches sites

Launch site with the highest success ratio



- This launch site has a 76.9% success ratio

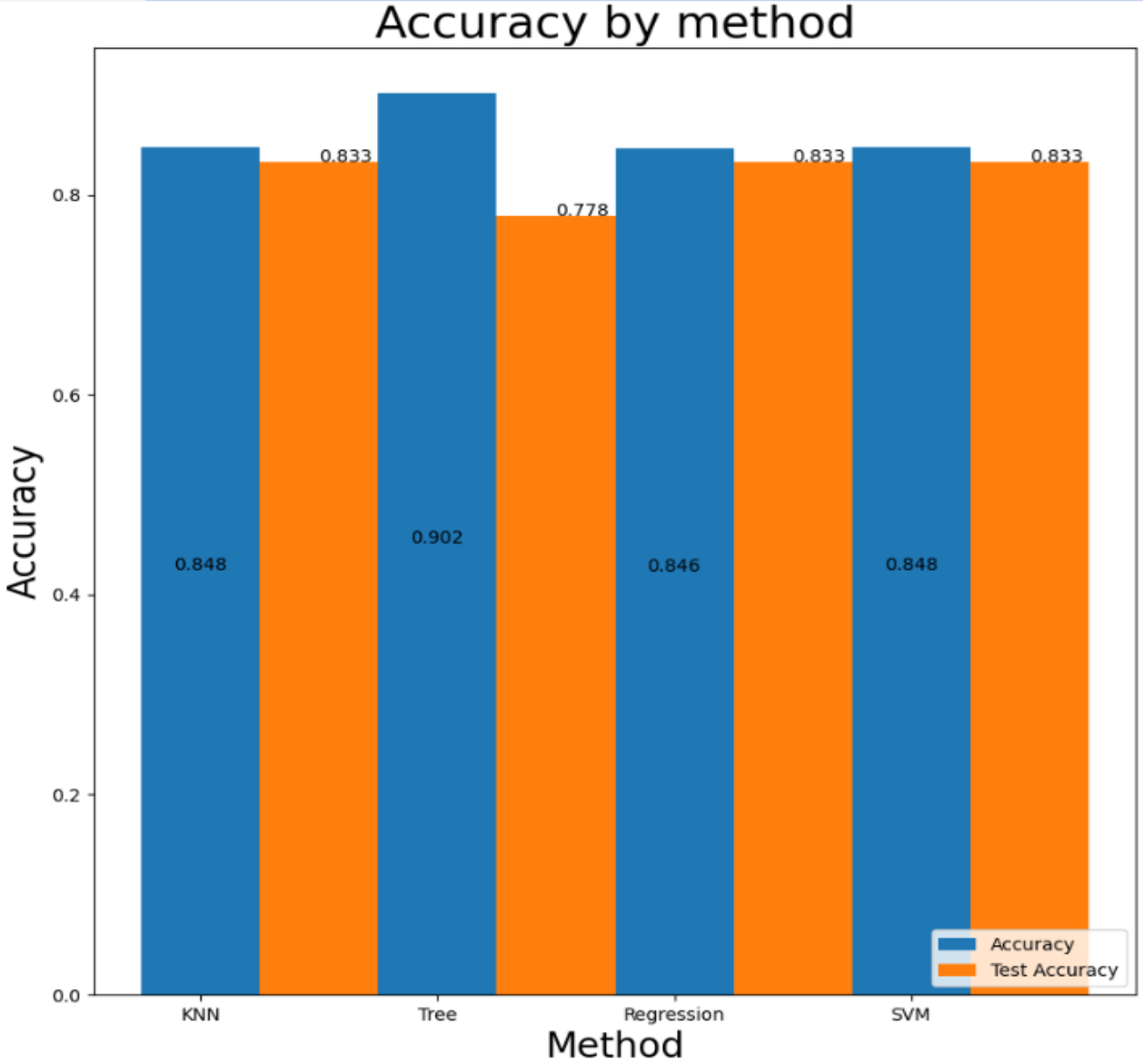
Successful launches per payload mass and booster version

- The payload mass between 2000 and 5000 are the most launched and the most successful

Section 5

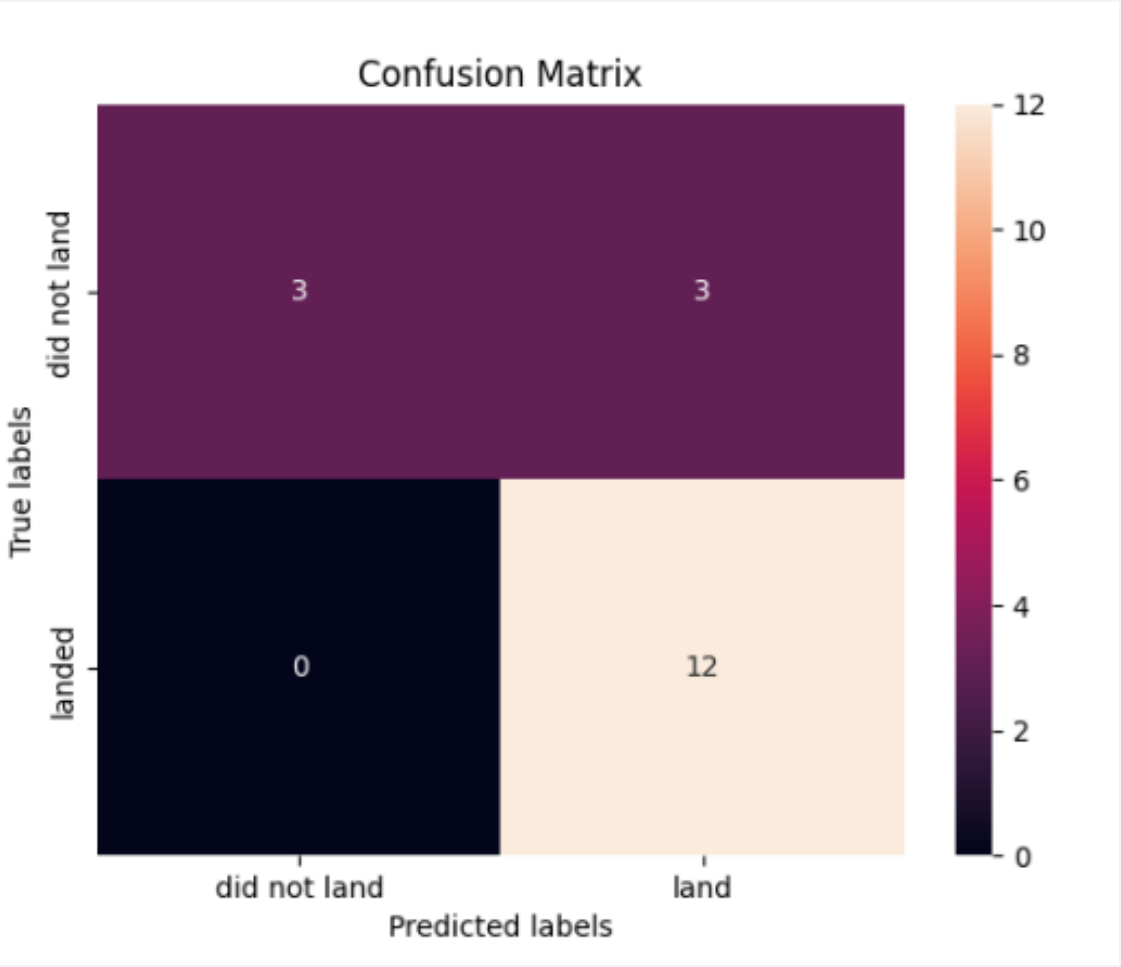
Predictive Analysis (Classification)

Classification Accuracy



- Decision Tree got the best accuracy with 90%

Confusion Matrix



- The confusion matrix shows that the decision tree method has predicted correctly the majority of the predictions

Conclusions

- Decision tree method was the most adequate model to predict the outcomes of the landings
- Data was obtained from different sources and methods
  - Directly from the Space X API
  - Web Scraping from wikipedia
- ES-L1, GEO, HEO and SSO orbits were the ones with the highest success rates.
- SO orbit has a success rate of zero
- There is a correlation between PayloadMass and success rate which seems to favor the heavier payloadMass

Thank you!