IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Dorota Mularczyk
02.10.2023

Outline

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

Executive Summary

- Summary of methodologies

The data was collected from two sources: wikipedia and SpaceX API, using Python. Exploratory Data Analysis was conducted with the use of Python and SQL, visualizations were created with Seaborn library and Folium. A dashboard was created in order to present the results in an interactive manner and facilitate drawing conclusions. A model was built, with the use of scikit-learn, in order to predict the launch outcomes.

- Summary of all results

EDA enabled better understanding of the data, including factors affecting the outcome of launches. A created model allows predict the outcome with over 90% accuracy.

Introduction

- Project background and context

The space race is on. One of the top players is SpaceX. Their strategy is to reuse part of the spaceship in order to minimize the cost of a launch. In order to compete with them, it is necessary to estimate the cost of launches, which can only work if the successful landing of the reusable part be predicted. Data science can help in this task, as well as in understanding better the factors that affect the landing outcome.

- Problems you want to find answers

What factors contribute to the success of a launch?

How to predict whether a launch will be successful or not?

What is a realistic cost estimation of a launch?

Section 1

Methodology

Methodology

Executive Summary

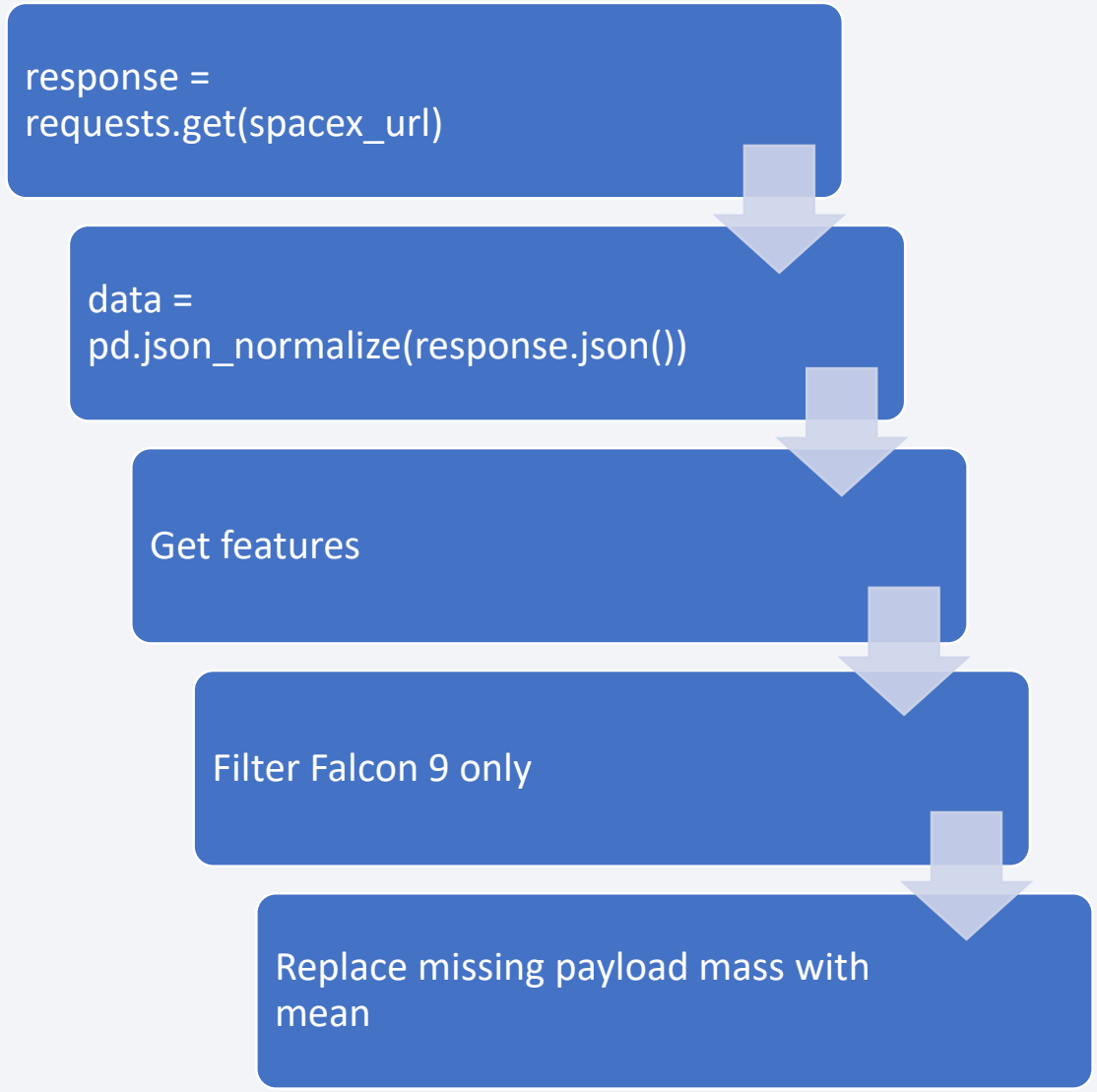
- Data collection methodology
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

The datasets for this project were sourced from two sites:

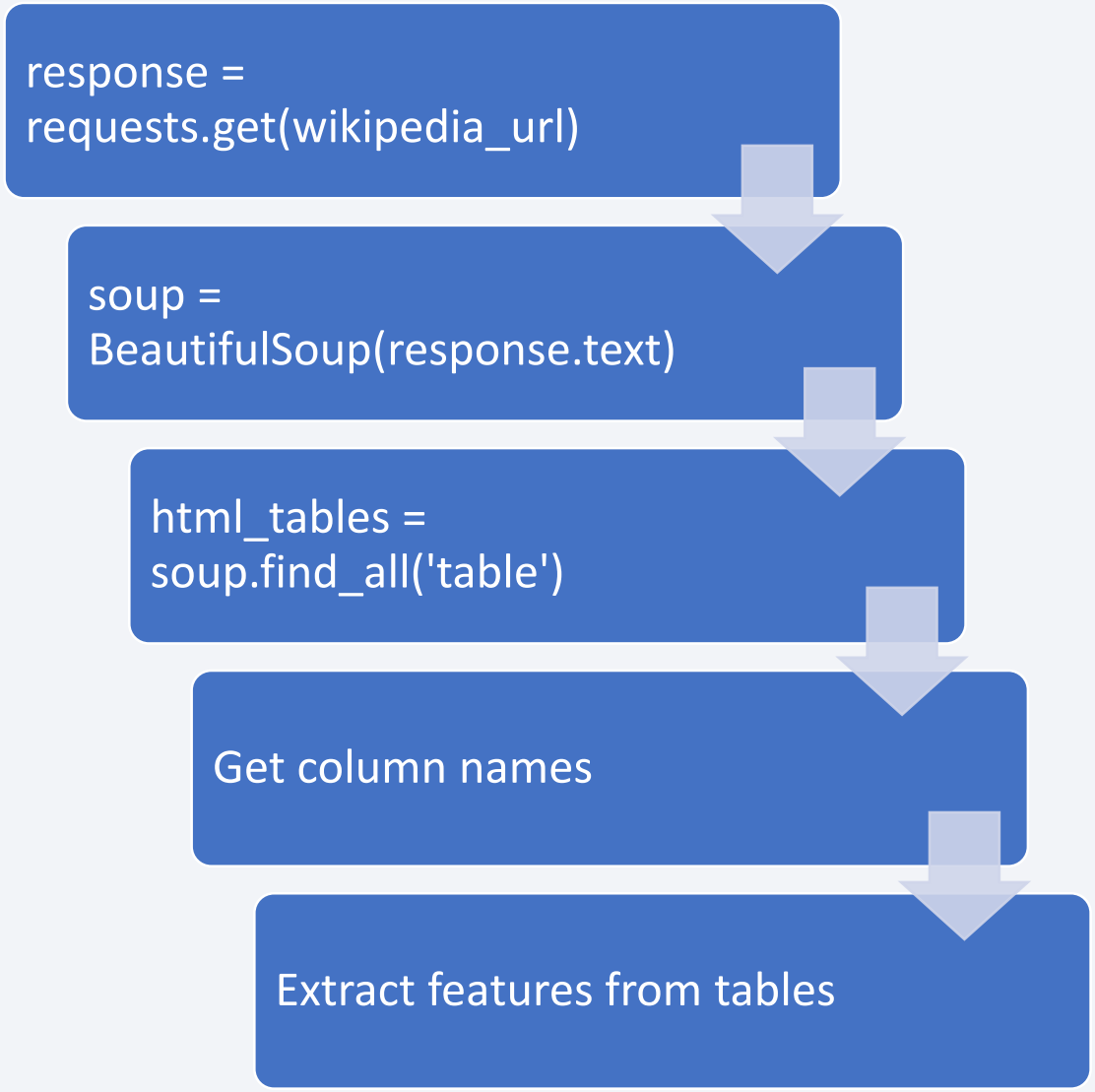
- SpaceX API
- Wikipedia (using webscraping).

Data Collection – SpaceX API



https://github.com/ontitoa/CourseraDS/blob/main/capstone1_jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

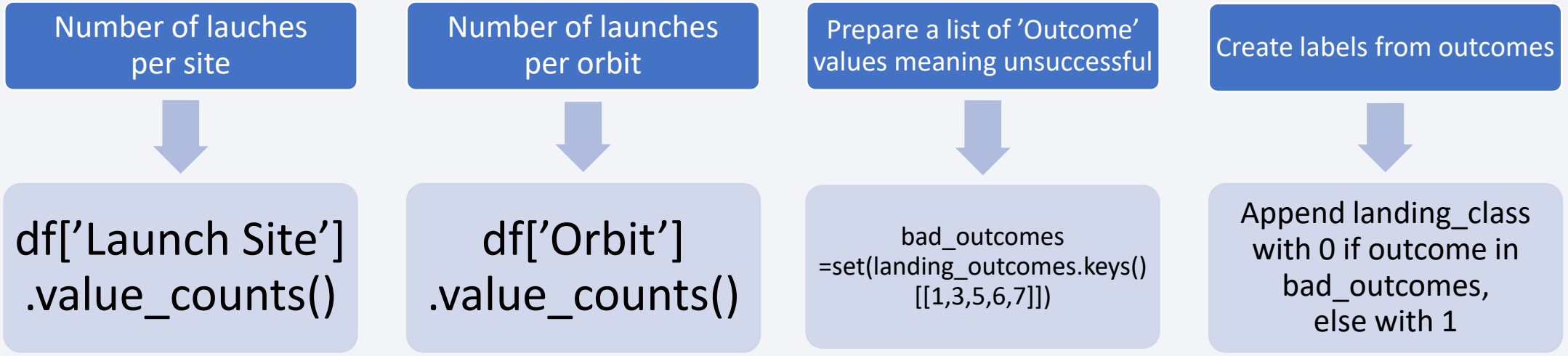


https://github.com/ontitoa/CourseraDS/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

First, to get more familiar with the data, the numer of launches per site and per orbit was determined. Later binary labels (0: failure, 1: success) were prepared (the original data included more details about the outcome and the outcome was given as string)

https://github.com/ontitoa/CourseraDS/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb



EDA with Data Visualization

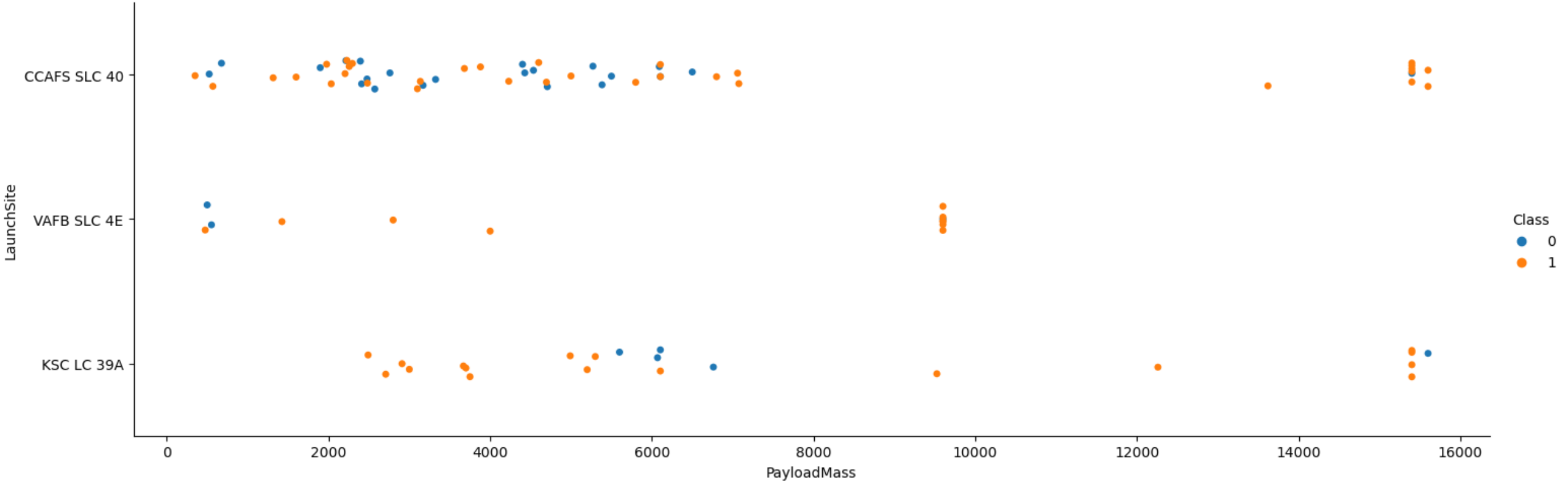
The exploratory data analysis was performed by preparing different types of plots with seaborn library:

- Scatter plot – to examine relationships between payload mass, flight numer, location and orbit together with the outcome encoded as color
- Bar chart – to present success rates at different orbits
- Line plot – to show yearly trends of launch success

https://github.com/ontitoa/CourseraDS/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with Data Visualization – Chart Examples

Is there any relationship between launch sites and their payload mass, and outcome?

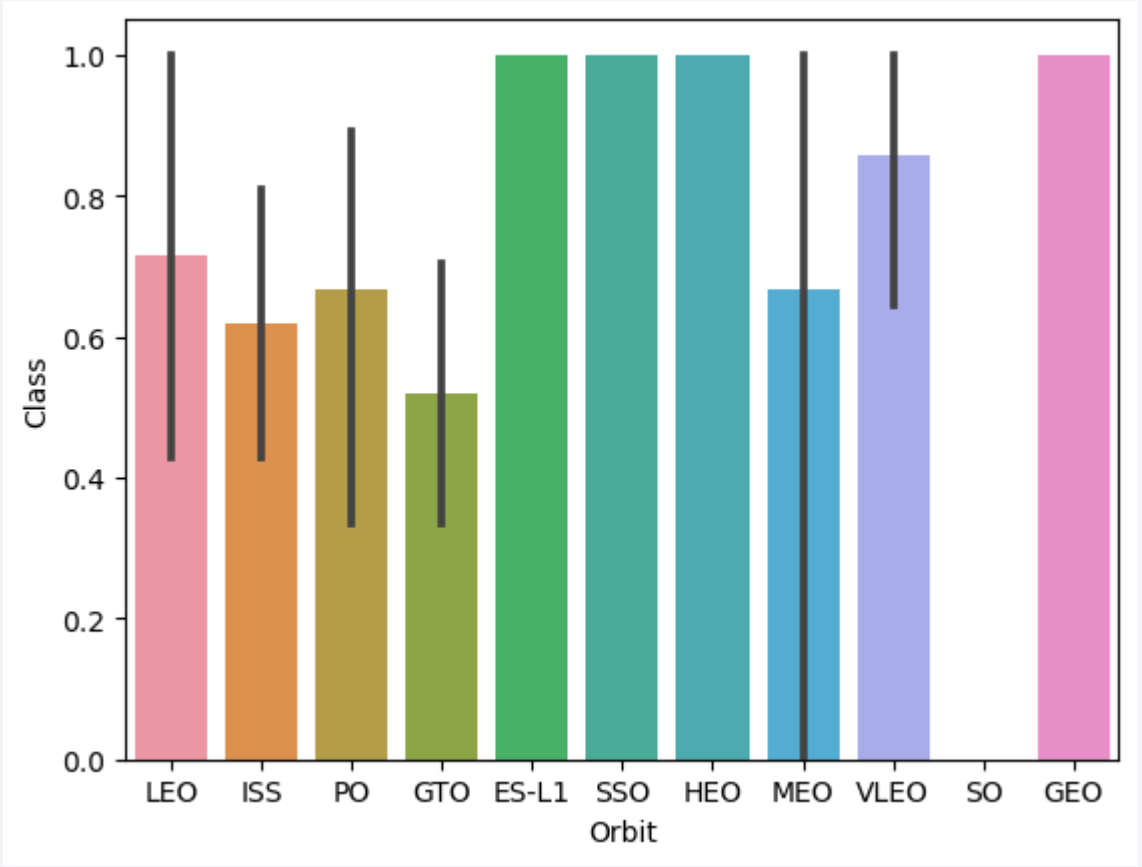


- In VAFB SLC 4E mostly lower payloads are launched.
- In KSC LC 39A no lightest payloads were launched.
- There is no clear correlation between payload mass and success of a launch.

EDA with Data Visualization – Chart Examples

What are success rates for different orbits?

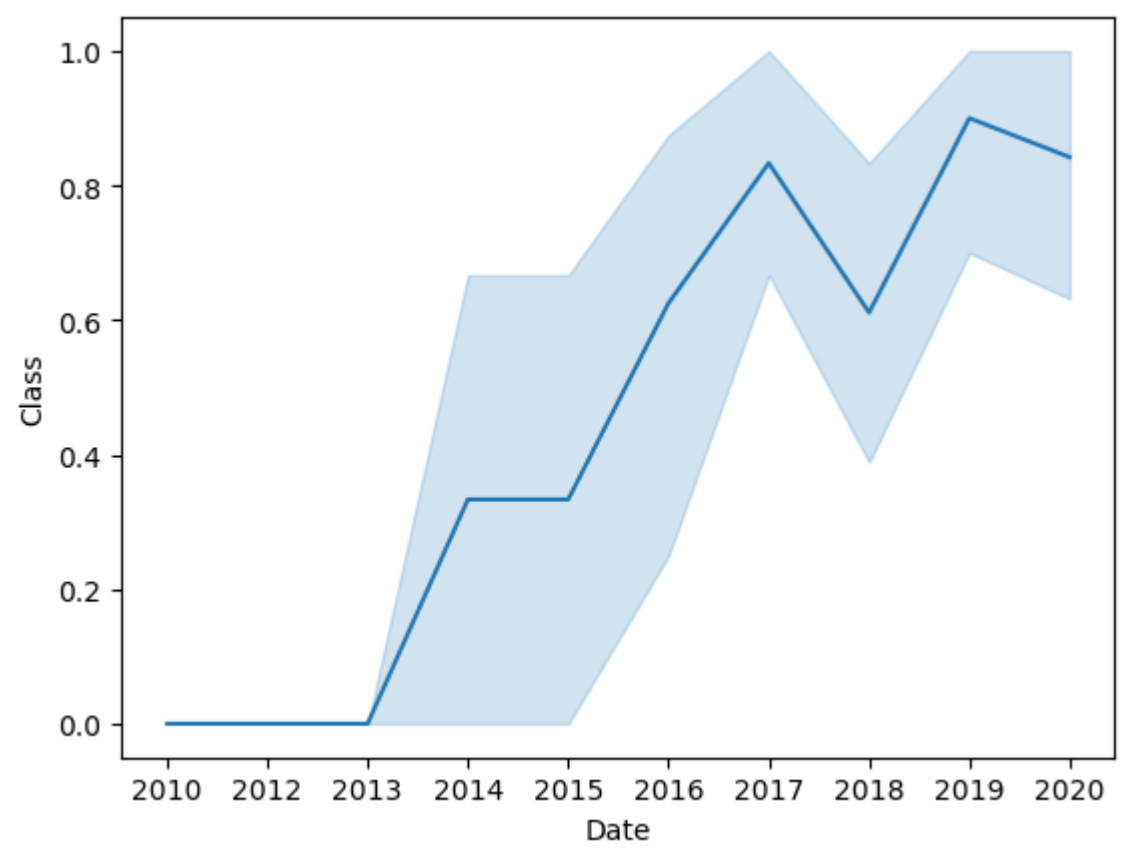
- There were no successful launches for SO
- Second-lowest success rate is for GTO
- There is 100% success rate for:
  ES-L1, SSO, HEO and GEO



EDA with Data Visualization – Chart Examples

What is the yearly trend of launch success?

The success rate increases with time. The biggest change was observed between 2013 and 2017, since then the growth has been slower.



EDA with SQL

Exploratory Data Analysis with SQL delivers the following insights from the data:

- There are **4 unique launch sites** in the space mission (CCAFS LC-40, VAFB SLC-4E, KSC LC-39A and CCAFS SLC-40)
- See examples of records where launch sites begin with the string 'CCA’
- The **total payload** mass carried by boosters launched by **NASA (CRS)** equals **45,596 kg**.
- The **average payload** mass carried by booster version **F9 v1.1** is **2,928.4 kg**.
- The **first succesful landing outcome in ground pad** was achieved on the **22nd of December 2015**.

EDA with SQL

Exploratory Data Analysis with SQL delivers the following insights from the data:

- There are **4 boosters** which have **success in drone ship** and have **payload mass greater than 4000 but less than 6000** (list available, all names start with F9 FT B10xx)
- In total, **100** missions ended with a **success** and only **1** with **failure**.
- There are **12 the booster_versions** which have carried the **maximum payload mass** (list available, all names start with F9 B5 B10xx)
- There were **two failure landings in drone ship in 2015**, one in April and one in Octobre (list including the months, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015 available)
- The **most numerous landing outcome** between the date 2010-06-04 and 2017-03-20 is **„No attempt”** (full ranked list available)

https://github.com/ontitoa/CourseraDS/blob/main/jupyter-labs-eda-sql-coursera_sqllite%20(1).ipynb

Build an Interactive Map with Folium

On the Folium map the locations of Launch Centres were marked with circles and appropriately labelled. Markers of different colours were also used to show the outcome of launches at each site. The distance to proximities was marked with lines (and labelled with length in km).

Adding these objects enabled getting an insight into characteristics of specific sites. Allowed to draw conclusions about good locations for a launch centre and see where the launchces are most numerous and most successful.

https://github.com/ontitoa/CourseraDS/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

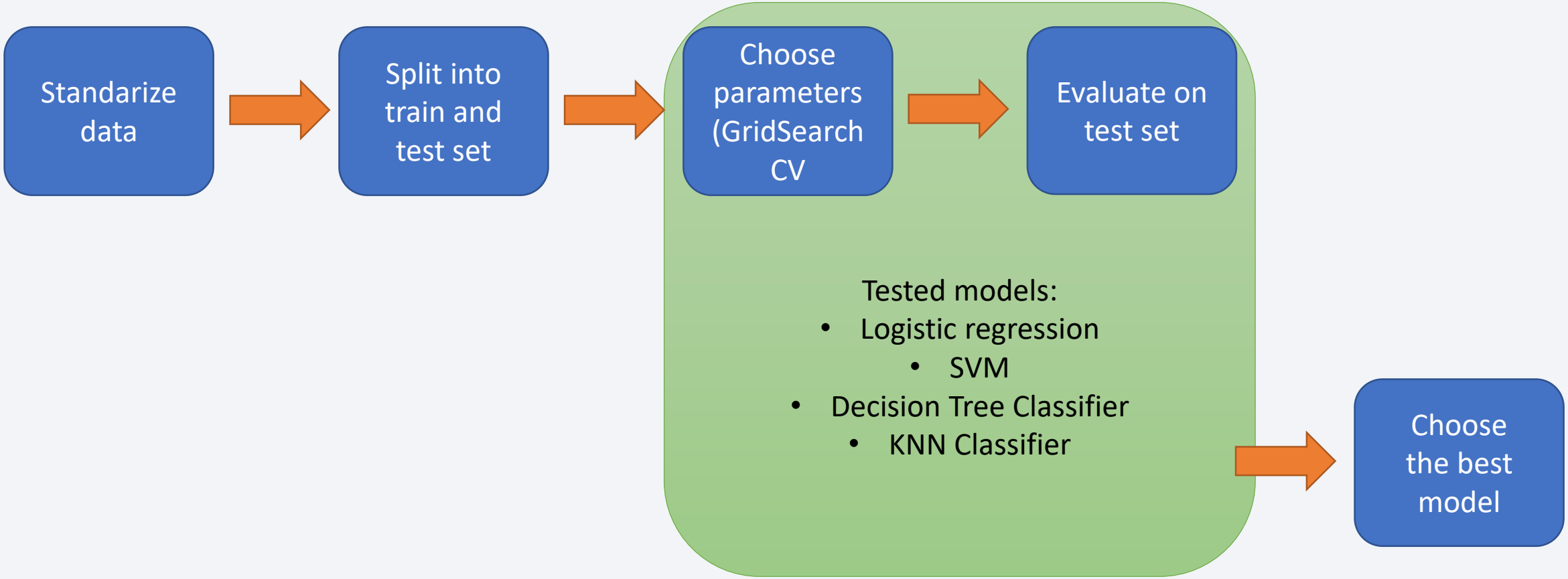
Build a Dashboard with Plotly Dash

A dashboard presents two types of plots: a pie plot showing success rates and scatter plot showing outcome of launches for different booster versions. In both cases it's possible to choose the site (or all sites) for which the data should be shown. For scatter plot it's possible to adjust the payload mass range of interest.

These plots, presented in an interactive manner, deliver a great deal of information helpful in analysis. The interactivity lets the user see the data clearly, focusing on chosen aspects. Dashboard can be helpful in drawing conclusions about different sites and booster versions, such as where the launches were most successful or which booster versions are better for use with different payload masses.

https://github.com/ontitoa/CourseraDS/blob/main/dash_spacey.py

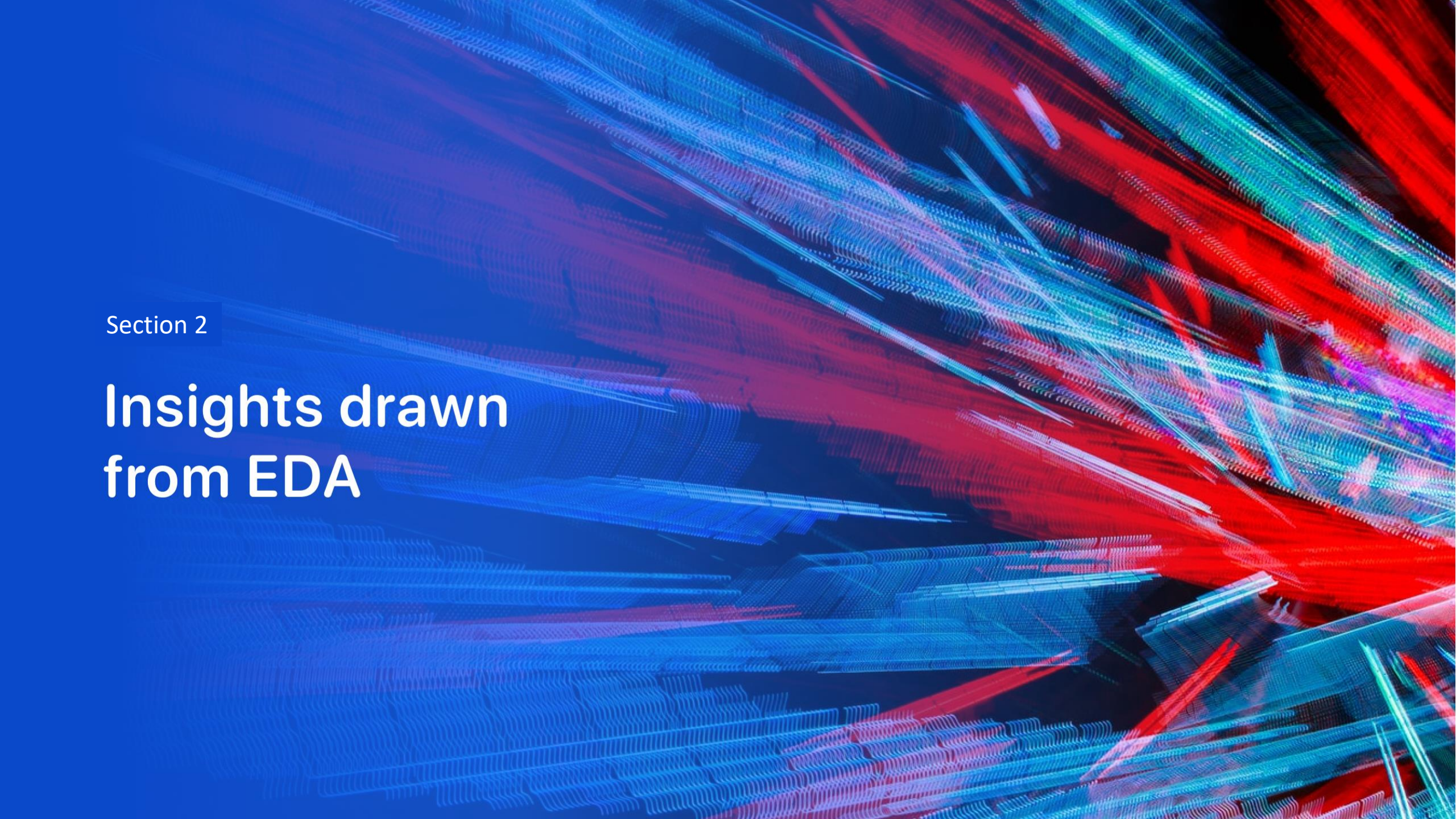
Predictive Analysis (Classification)



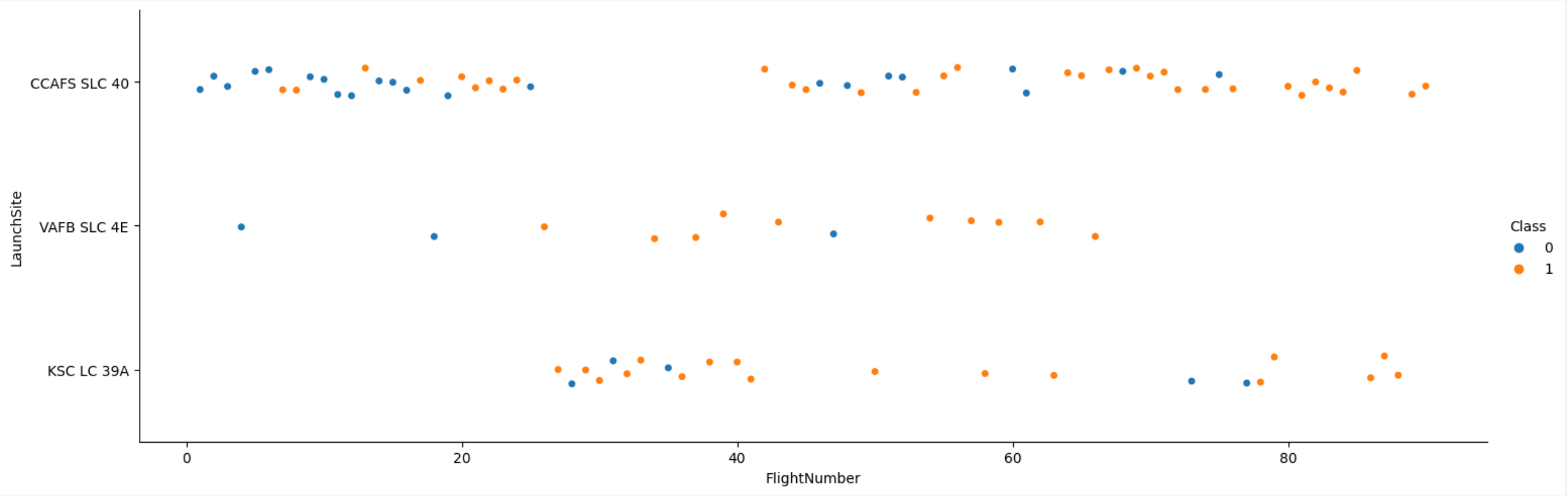
https://github.com/ontitoa/CourseraDS/blob/main/Pred_correct_IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite%20(1).ipynb

Results

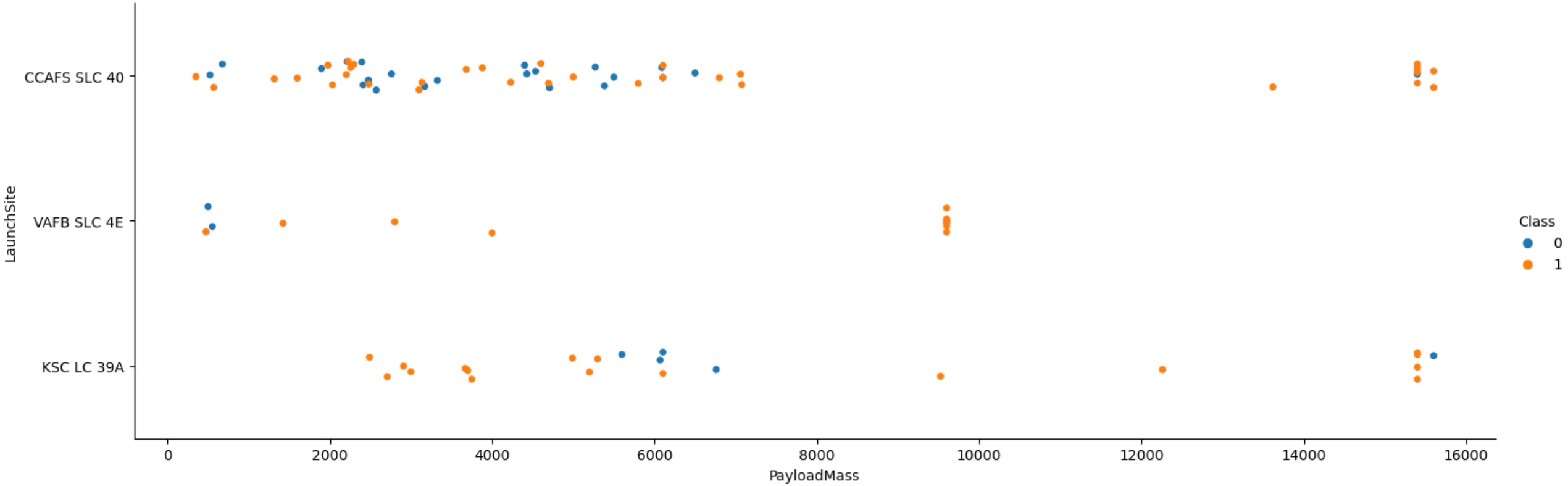
- Exploratory data analysis showed that the success ratio depends on many factors, such as target orbit or payload mass. Also some orbits normally take heavier loads than others, therefore it's hard to judge if the success is affected more by the orbit or the mass. Overall success rate increases in time. The analysis also showed what good launch site locations are like.
- The predictive analysis proved that it's possible to build a simple model that will be able to predict the launch outcome based on collected data.


Section 2
Insights drawn from EDA

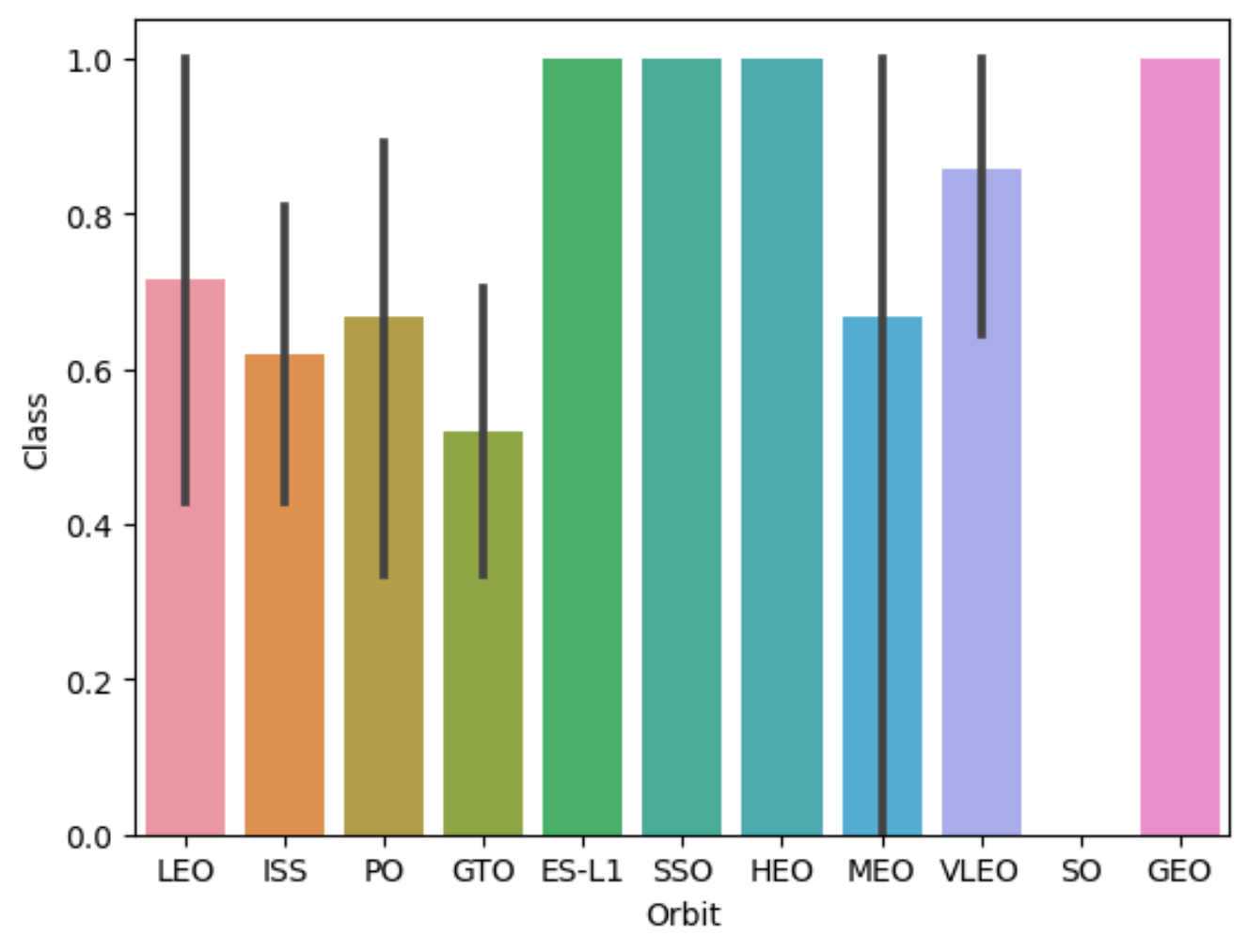
Flight Number vs. Launch Site



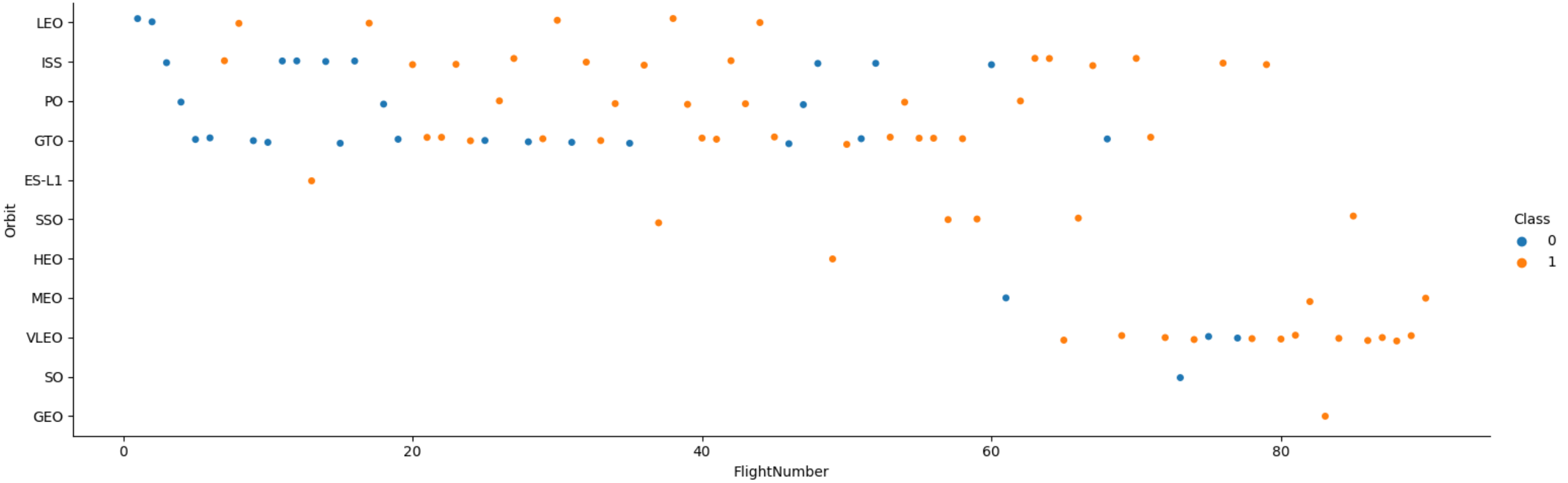
Payload vs. Launch Site



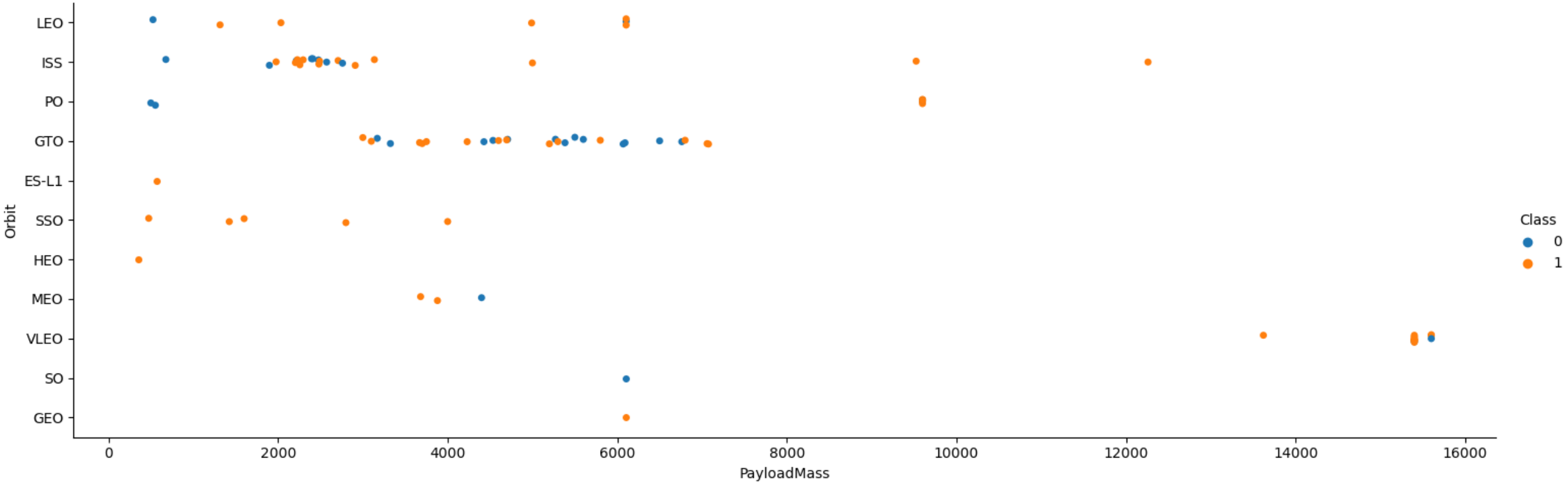
Success Rate vs. Orbit Type



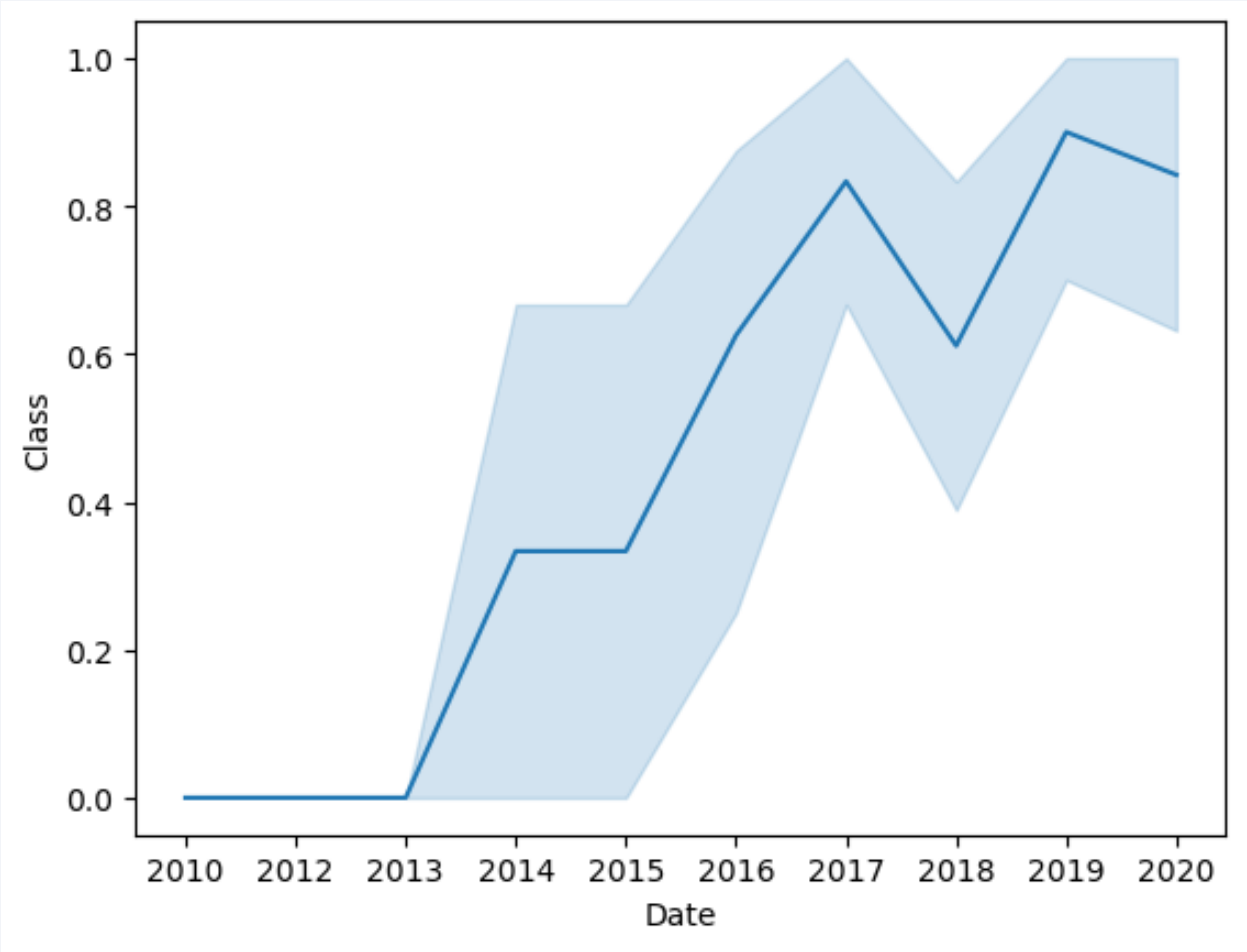
Flight Number vs. Orbit Type



Payload vs. Orbit Type



Launch Success Yearly Trend



All Launch Site Names

- There are **4 unique launch sites** in the space mission (CCAFS LC-40, VAFB SLC-4E, KSC LC-39A and CCAFS SLC-40)

Task 1

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

```
[14]: %sql SELECT DISTINCT "Launch_Site" FROM SPACEXTBL;
```

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

[14]:

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

Launch Site Names Begin with 'CCA'

By using option LIMIT the results were limited to 5 records. Option „LIKE” allowed to include all launching sites starting with „CCA.”

Task 2

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

[26]:
```
%sql SELECT * FROM SPACEXTBL WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5;
```

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

[26]:

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

- The **total payload** mass carried by boosters launched by **NASA (CRS)** equals **45,596 kg**.

Task 3

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

```
[34]: %sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE Customer="NASA (CRS)";
```

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

[34]:

SUM("PAYLOAD_MASS__KG_")
45596

Average Payload Mass by F9 v1.1

- The **average payload** mass carried by booster version **F9 v1.1** is **2,928.4 kg**.

Task 4

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

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

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

AVG("PAYLOAD_MASS__KG_")
2928.4

First Successful Ground Landing Date

- The **first succesful landing outcome in ground pad** was achieved on the **22nd of December 2015**.

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
[38]: %sql SELECT MIN(Date) from SPACEXTBL WHERE "Landing_Outcome"="Success (ground pad)"
```

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

[38]:

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- There are **4 boosters** which have **success in drone ship** and have **payload mass greater than 4000 but less than 6000** (all names start with F9 FT B10xx)

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

[39]:
```
%sql SELECT DISTINCT "Booster_Version" FROM SPACEXTBL WHERE "Landing_Outcome"="Success (drone ship)" AND "PAYLOAD_MASS__KG_" BETWEEN 4000 AND 6000;
```

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

[39]:

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 total, **100** missions ended with a **success** and only **1** with **failure**.

```
[49]: %sql SELECT COUNT(*) AS SUCCESS_NUMBER FROM SPACEXTBL WHERE "Mission_Outcome" LIKE "Success%";
```

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

[49]:

SUCCESS_NUMBER
100

```
[50]: %sql SELECT COUNT(*) AS FAILURE_NUMBER FROM SPACEXTBL WHERE "Mission_Outcome" LIKE "Failure%";
```

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

[50]:

FAILURE_NUMBER
1

Boosters Carried Maximum Payload

- There are **12 the booster_versions** which have carried the **maximum payload mass** (all names start with F9 B5 B10xx)

Task 8

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

[52]:
```
%sql SELECT DISTINCT "Booster_Version" FROM SPACEXTBL WHERE "PAYLOAD_MASS__KG_"=(SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTBL);
```

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

[52]:

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

- There were **two failure landings in drone ship in 2015**, one in April and one in Octobre (list including the months, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015 available)

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.

[60]:
```
%sql SELECT substr(Date, 6, 2) as Month, "Landing_Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTBL WHERE "Landing_Outcome" = "Failure (drone ship)" AND substr(Date, 1,4)='2015';
```

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

[60]:

Month	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- The **most numerous landing outcome** between the date 2010-06-04 and 2017-03-20 is **„No attempt”**

Task 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.

[67]:
```
%sql SELECT "Landing_Outcome", COUNT(*) AS "Outcome_Count" FROM SPACEXTBL WHERE "Date" BETWEEN "2010-06-04" AND "2017-03-20" GROUP BY "Landing_Outcome" ORDER BY "Outcome_Count" DESC;
```

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

[67]:

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

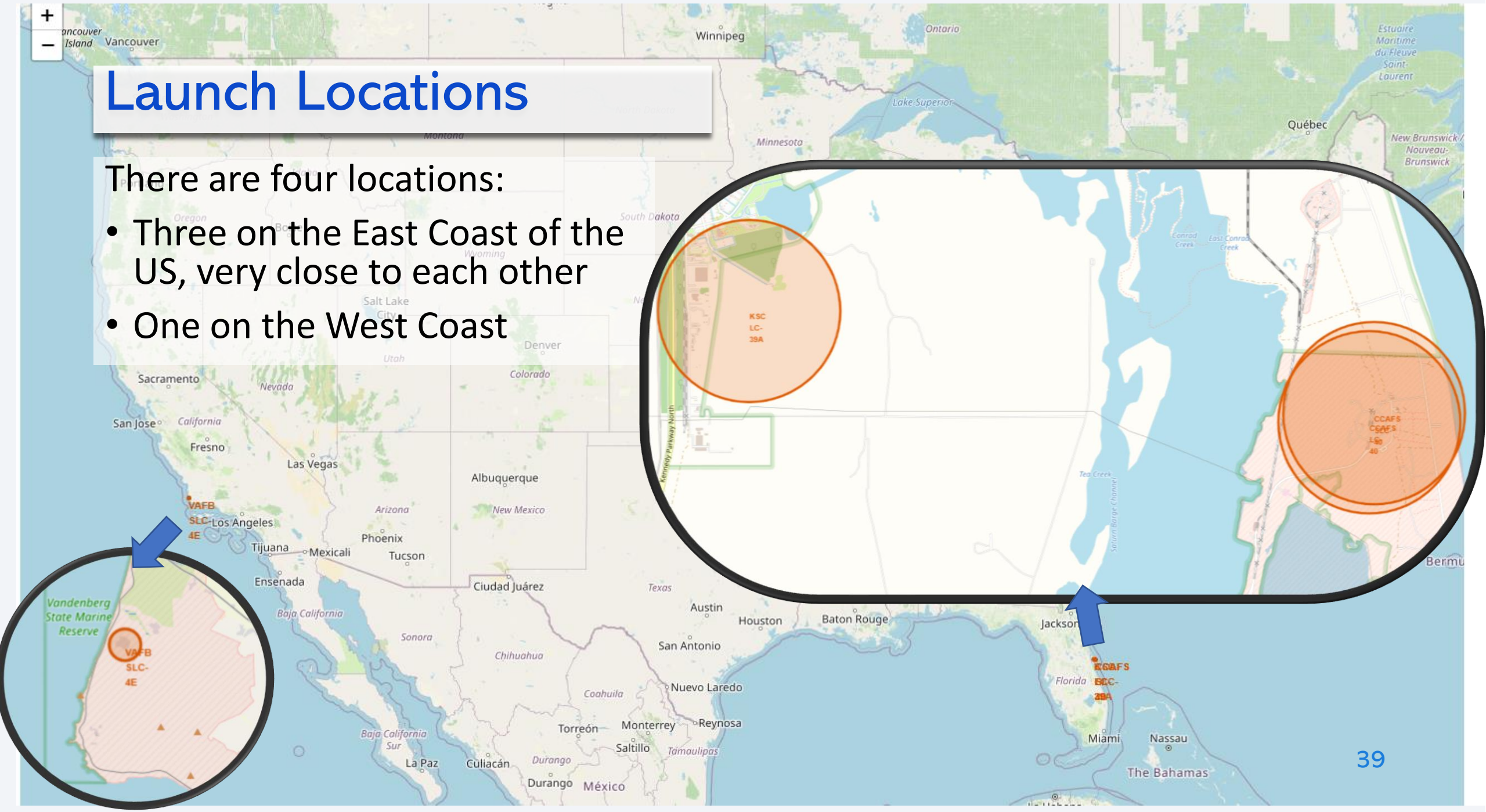
Section 3

Launch Sites Proximities Analysis

Launch Locations

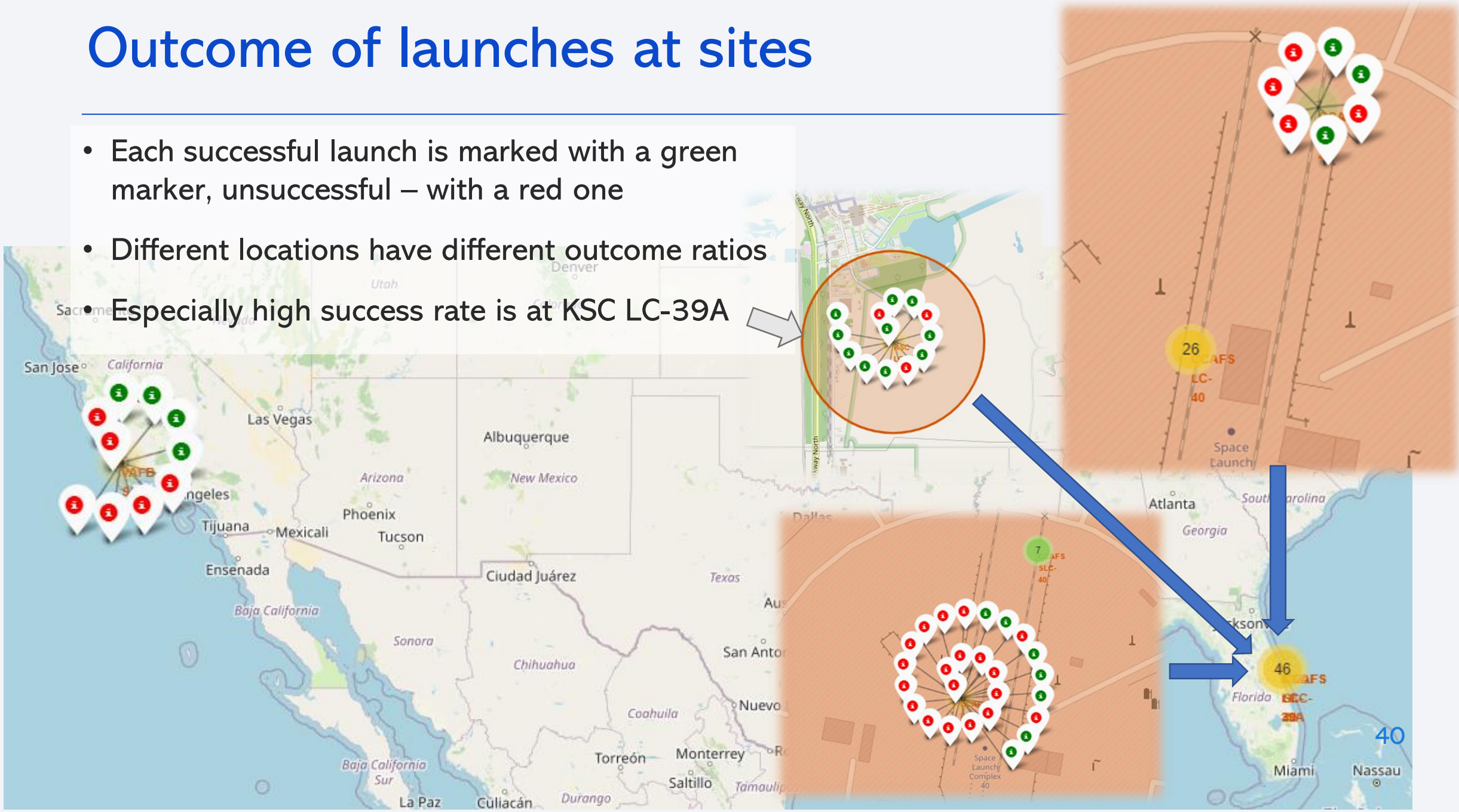
There are four locations:

- Three on the East Coast of the US, very close to each other
- One on the West Coast



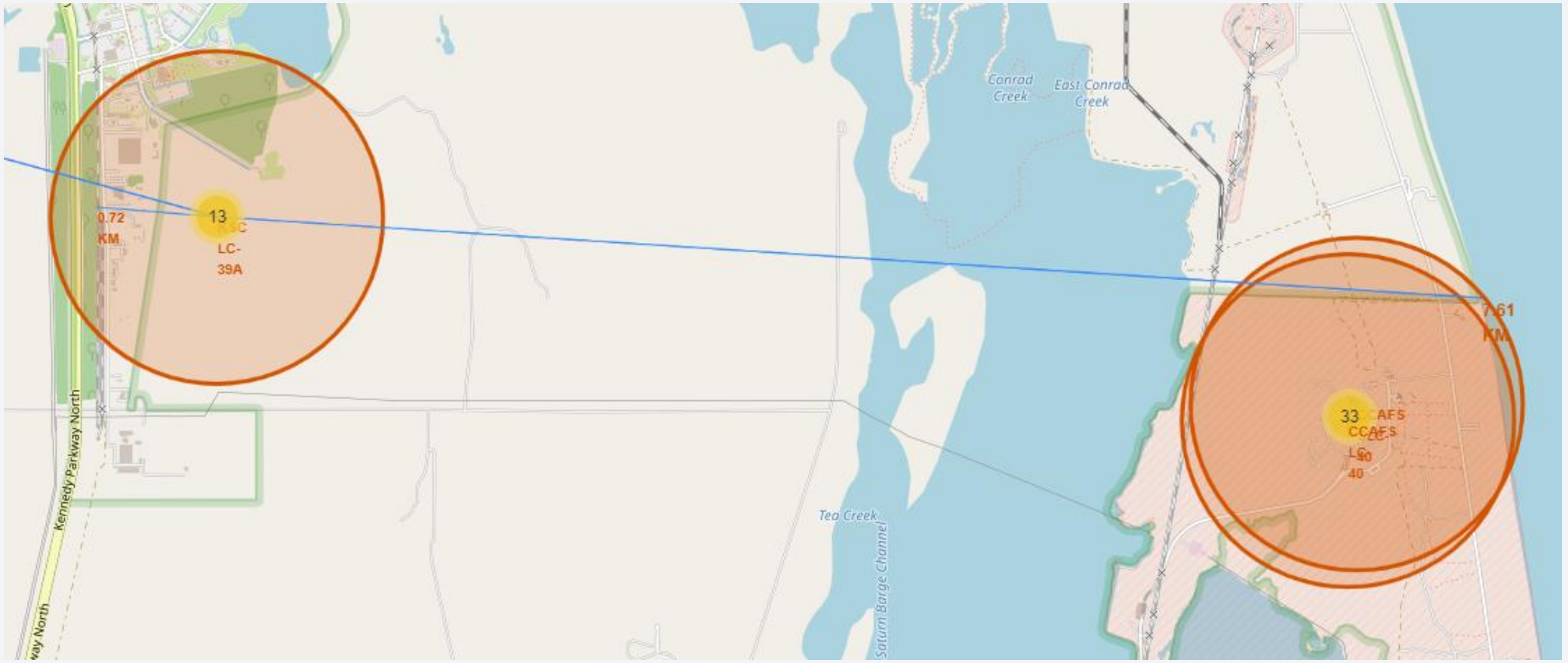
Outcome of launches at sites

- Each successful launch is marked with a green marker, unsuccessful – with a red one
- Different locations have different outcome ratios
- Especially high success rate is at KSC LC-39A



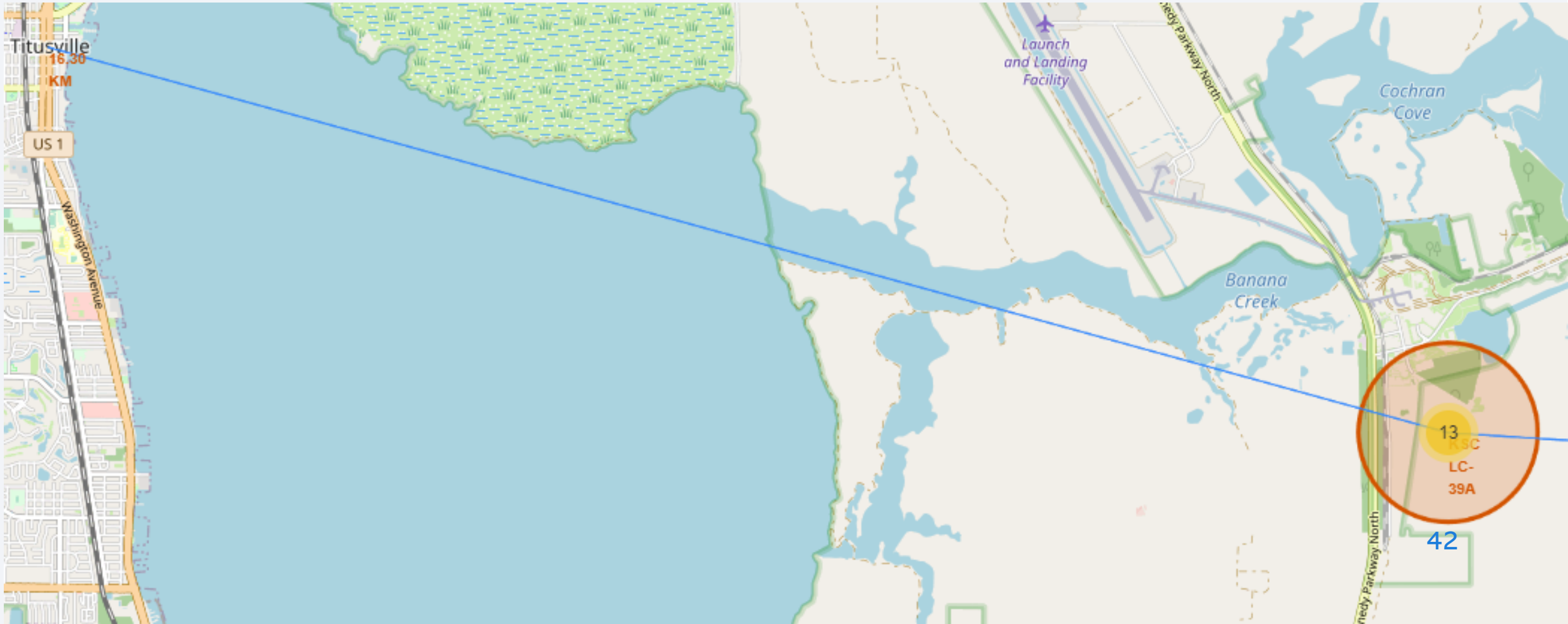
Distance to coastline

KSC LC-39A is located 7.61 km away from the nearest coastline.



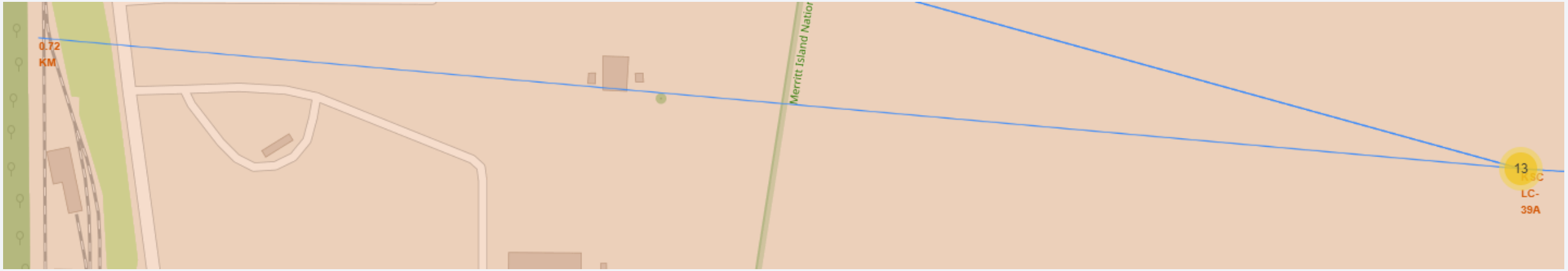
Distance to city

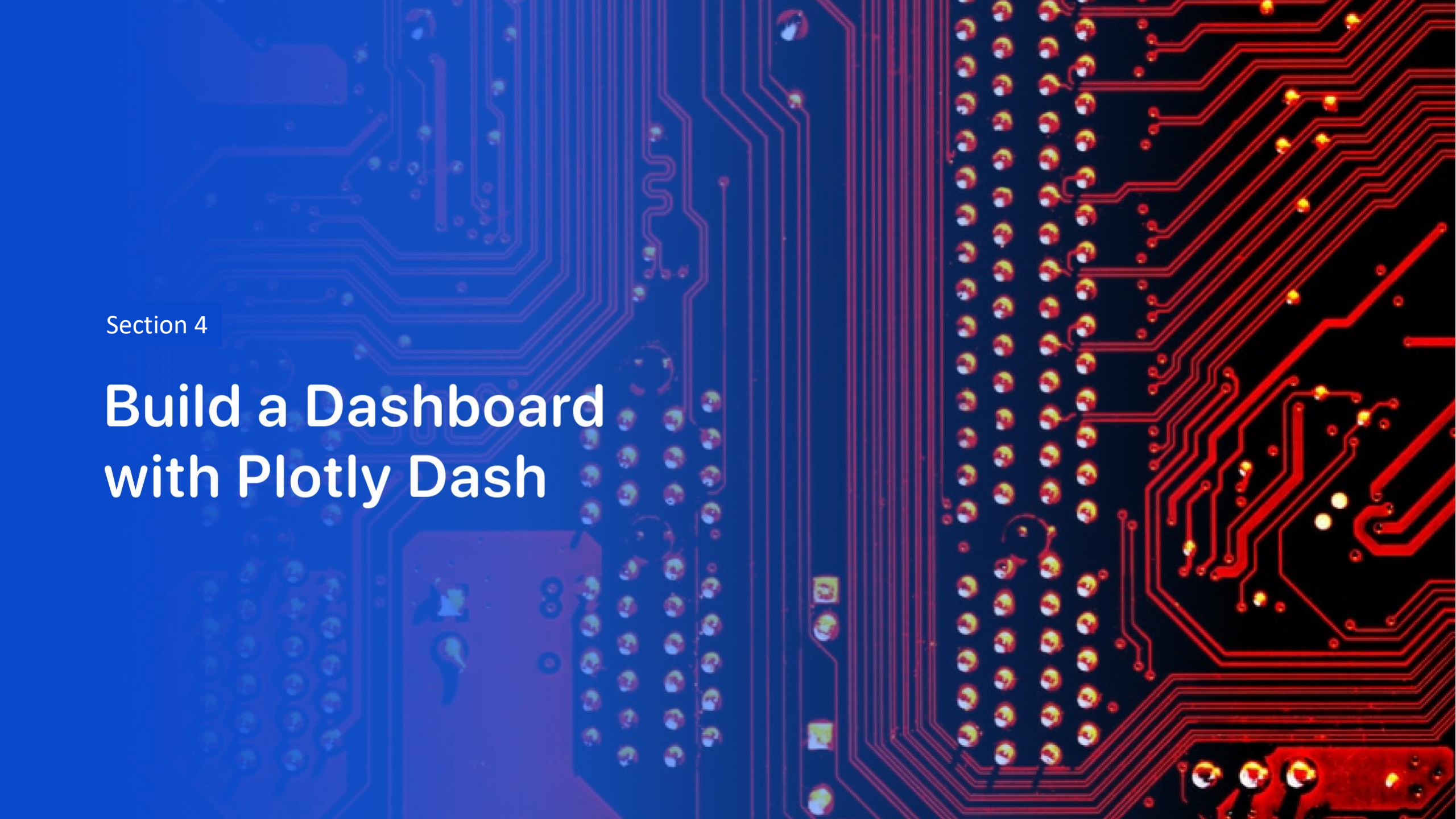
KSC LC-39A is located 16.3 km away from the nearest city.



Distance to railway

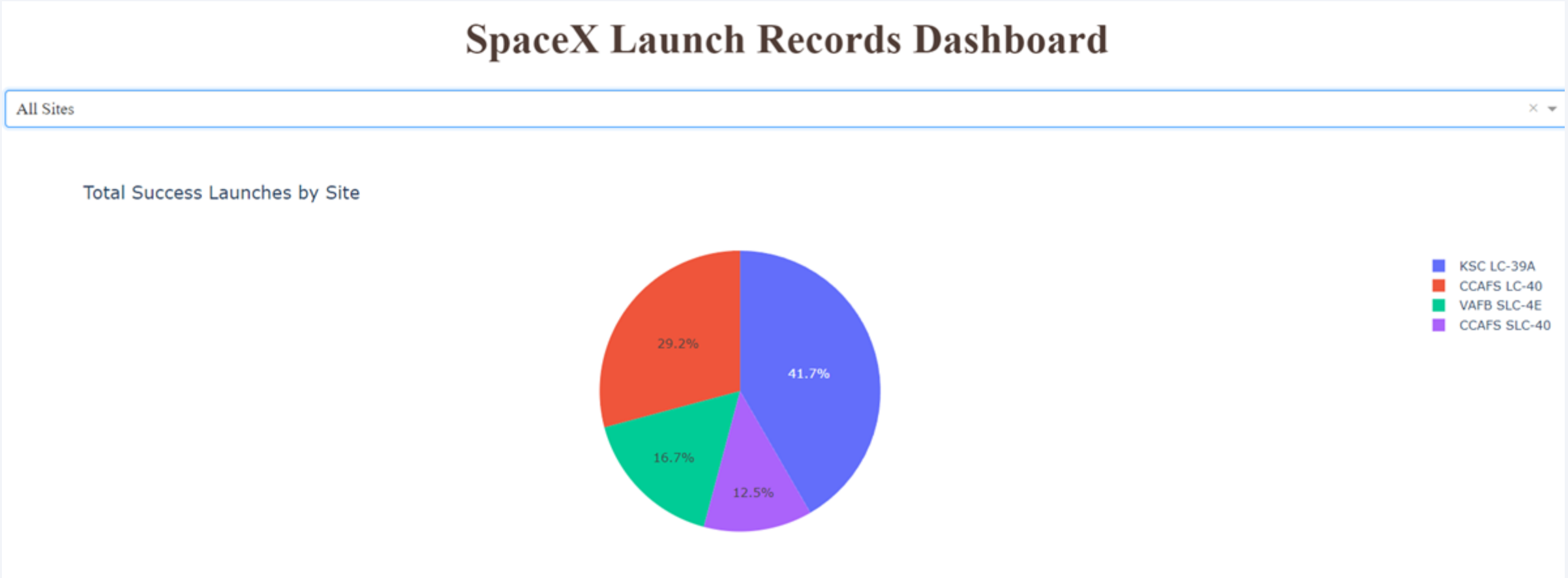
KSC LC-39A is located 0.72 km away from the nearest railway.




Section 4
Build a Dashboard with Plotly Dash

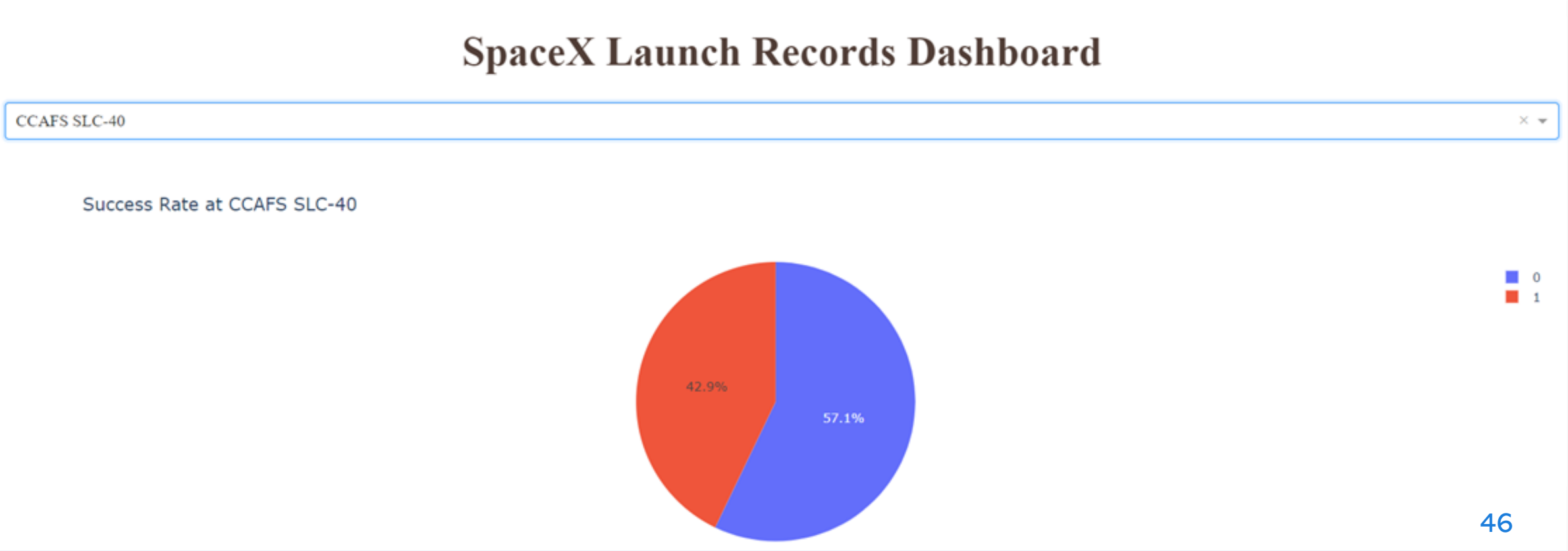
Total Successful Launches by Site

- The biggest number of successful launches happened at KSC LC-39A (41.7%)
- The smallest number of successful launches happened at CCAFS SCL-40 (12.5%)



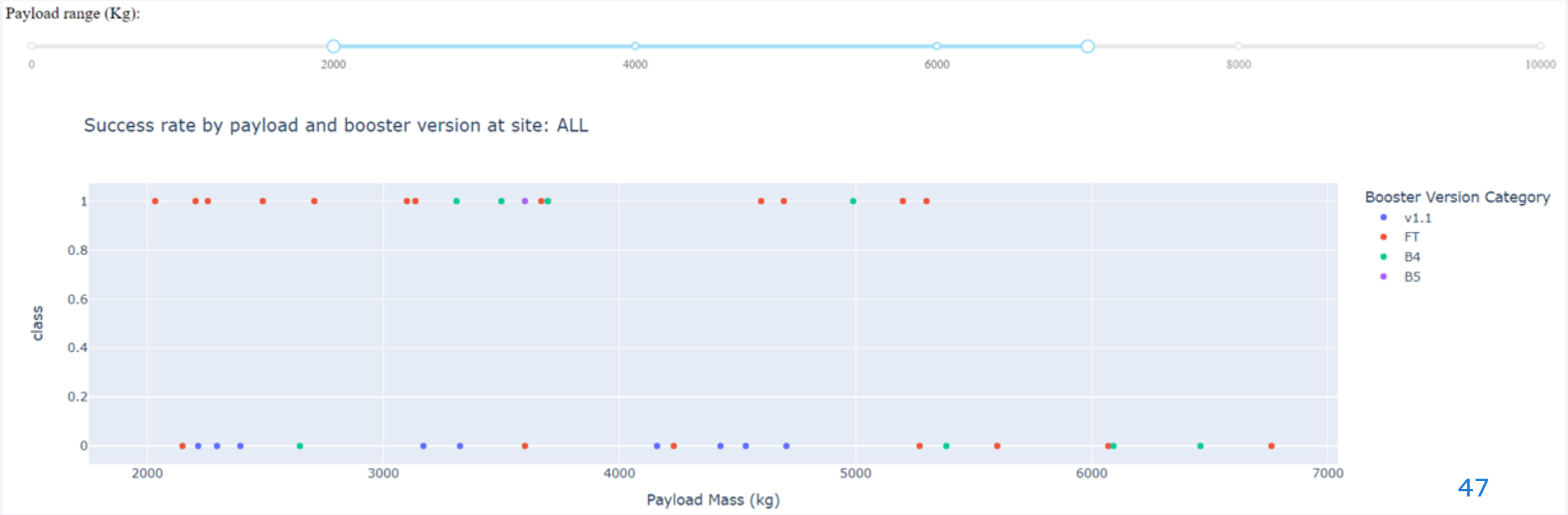
Success ratio at the most successful site

The highest success ratio was achieved at CCAFS SLC-40 launch site (42.9%)



Outcome dependence on payload mass and booster version

In the 2000-7000 kg payload mass range the booster version v1.1 is very unsuccessful, while FT performs decently well.

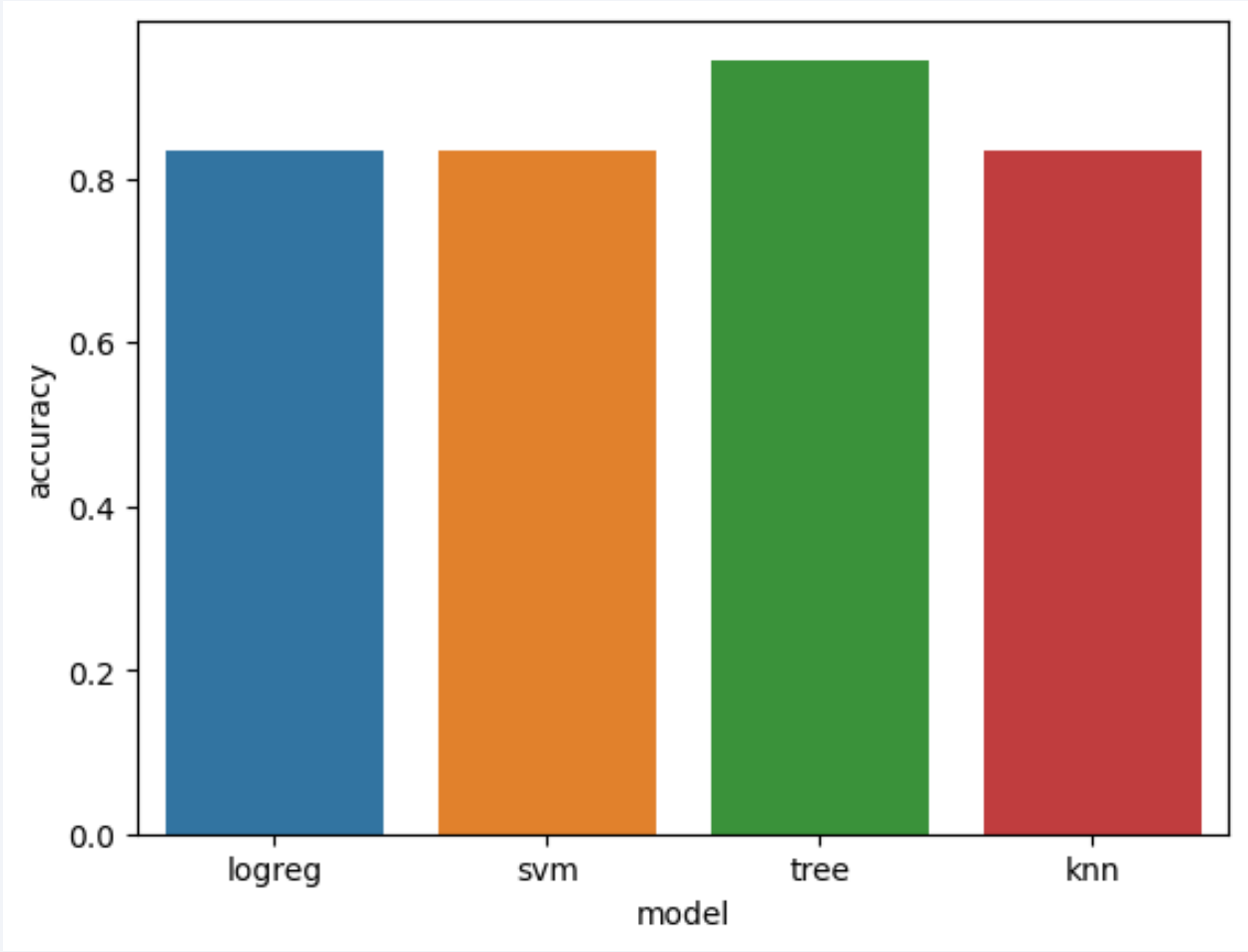


Section 5

Predictive Analysis (Classification)

Classification Accuracy

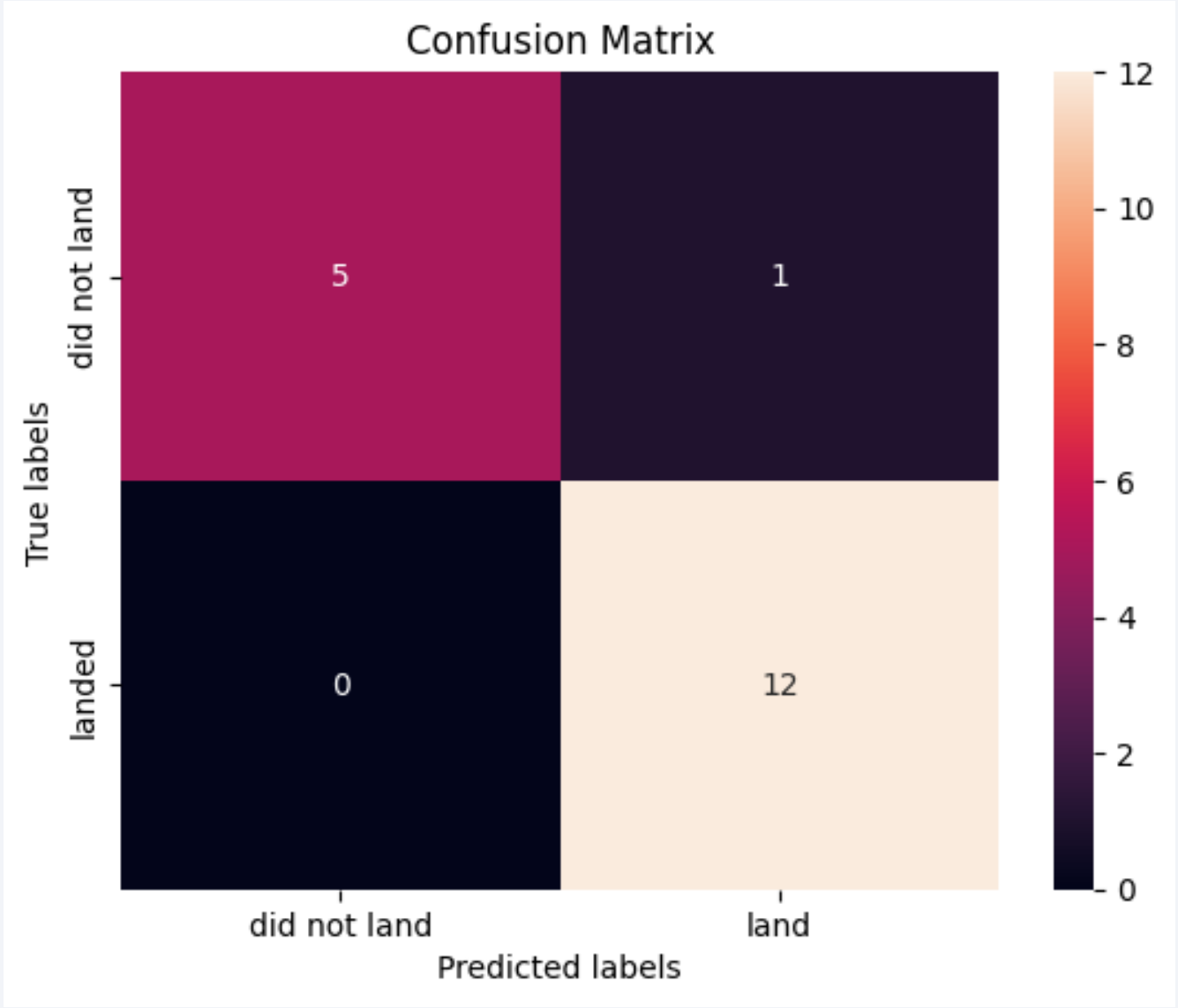
The highest accuracy was achieved by the tree classifier model (accuracy = 94,4%).



Confusion Matrix

Using the tree model:

- Out of 6 test examples which did not land successfully, 5 were correctly classified as „did not land” and one was incorrectly classified as „landed.”
- Out of 12 test examples which landed successfully, all were classified correctly.



Conclusions

- At site VAFB SLC 4E no heavy payloads are launched
- There were no successful launches to SO
- Success rate is 100% for orbits: ES-L1, SSO, HEO, GEO
- The target orbit changes together the flight number, which might also suggest changes in time
- The heaviest payloads were launched to VEO
- Ideal launch locations are located close to highways or railways and outside of cities (but not too far away)
- The biggest number of successful launches happened at KSC LC-39A (41.7%)
- The highest success ratio was achieved at CCAFS SLC-40 launch site (42.9%)
- In the 2000-7000 kg payload mass range the booster version v1.1 is very unsuccessful, while FT performs decently well.
- A predictive model for launch outcomes was successfully built using a decision tree.

Thank you!