Winning Space Race with Data Science

Wolfgang Huang
29.11.2024

Outline

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

Executive Summary

The project's aim was to identify and predict factors that are beneficial to a successful launch of Falcon 9 rockets.

Methodologies used:

- Data collection via requests to the SpaceX API and web-scraping from Wikipedia
- Data Preparation
- Exploratory Data Analysis using SQL
- Rocket data visualization as well as geographical maps
- Launch success prediction using various ML algorithms

Results are presented as

- Selected visualizations
- Interactive Dashboard
- Prediction results

Introduction

Background

Commercialized space exploration has opened a new era: Whereas delivering payload to space was in a price range of several hundreds of millions USD in the past, costs have now gone down to around 65 million USD. This makes space (or rather earth orbits) much more accessible, allowing, for example, satellites and satellite networks for earth observation, climate protection, high-bandwidth communication, and more.

One of the key components of such a reduced price is the reusability of parts of rockets. This means that the reused parts need to return to earth safely.

Problem

Ideally, all rocket stages intended to return to earth would actually do so – so far, only certain percentage of it have successfully returned. What are the decisive factors for a safe return? This data analysis will explore a broad variety of features and try to identify their interplay and effect on landing success.

Section 1

Methodology

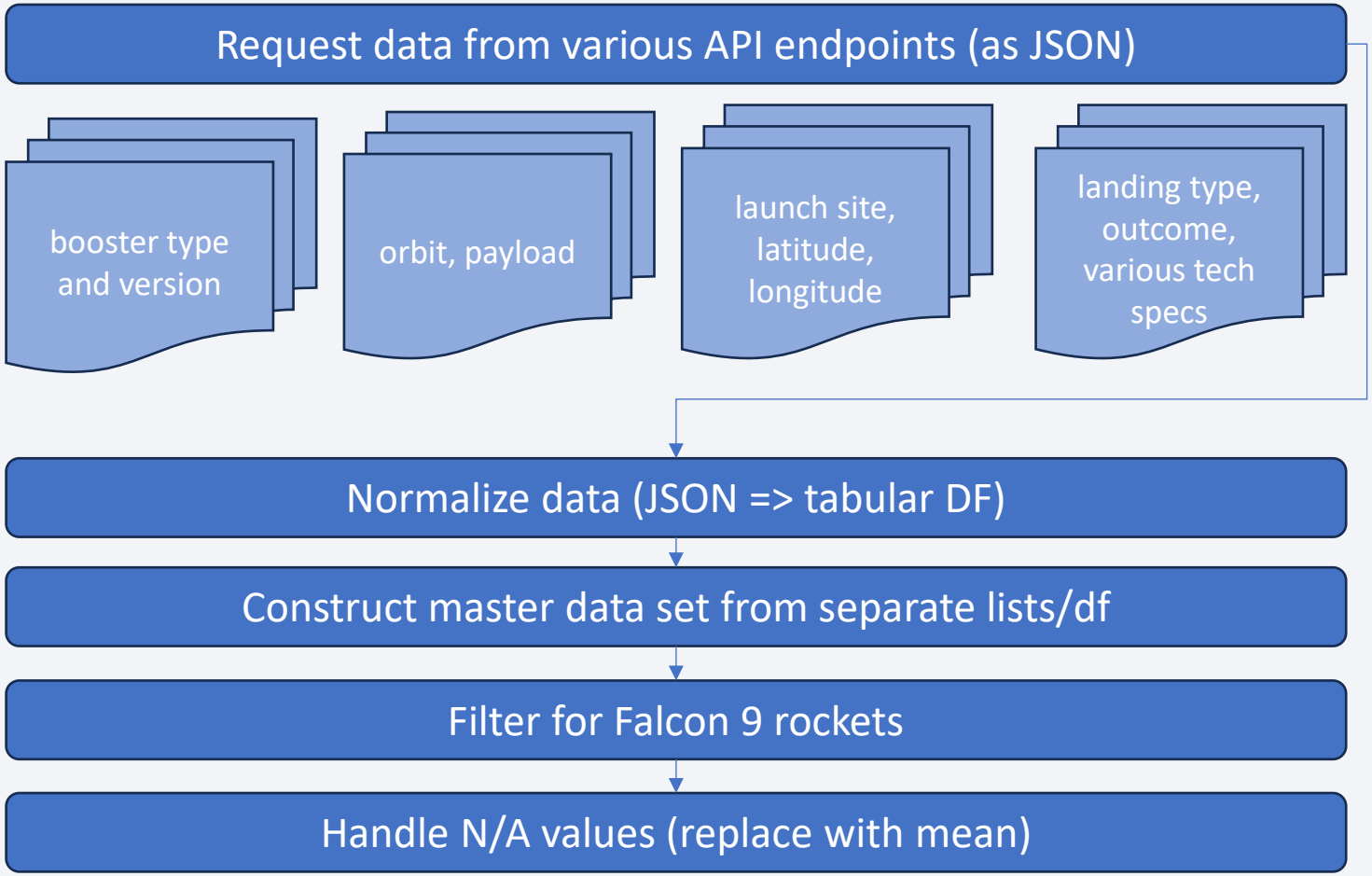
Methodology

Executive Summary

- Data collection methodology:
  - API requests, web-scraping
- Perform data wrangling
  - Cleaning, encoding, filtering
- 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 – SpaceX API

The flow chart to the right shows the main steps of data collection.

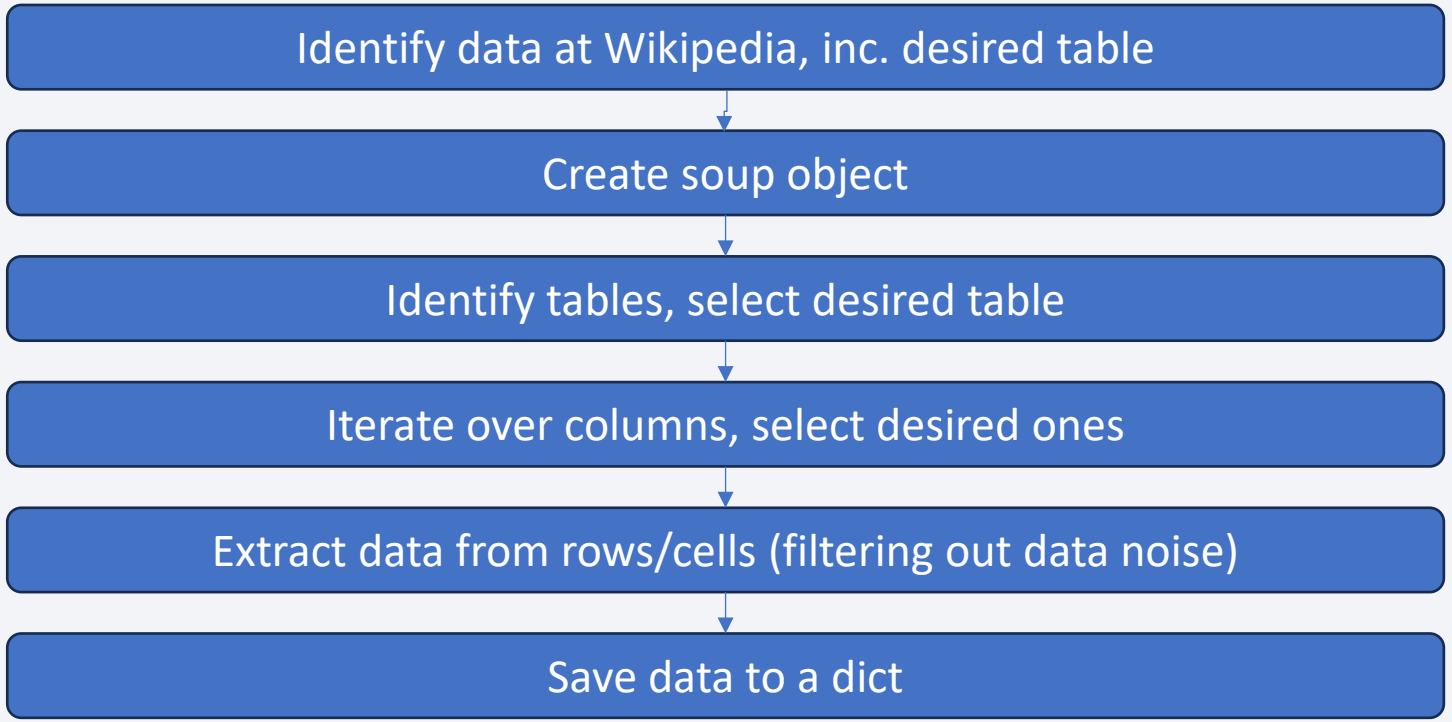


Code Link

https://github.com/WolfgangHuang/ibmdatasciencecourse/blob/main/capstone1-jupyter-labs-spacex-data-collection-api-v2_WH.ipynb

Data Collection - Scraping

The flow chart to the right shows the main steps of data scraping.

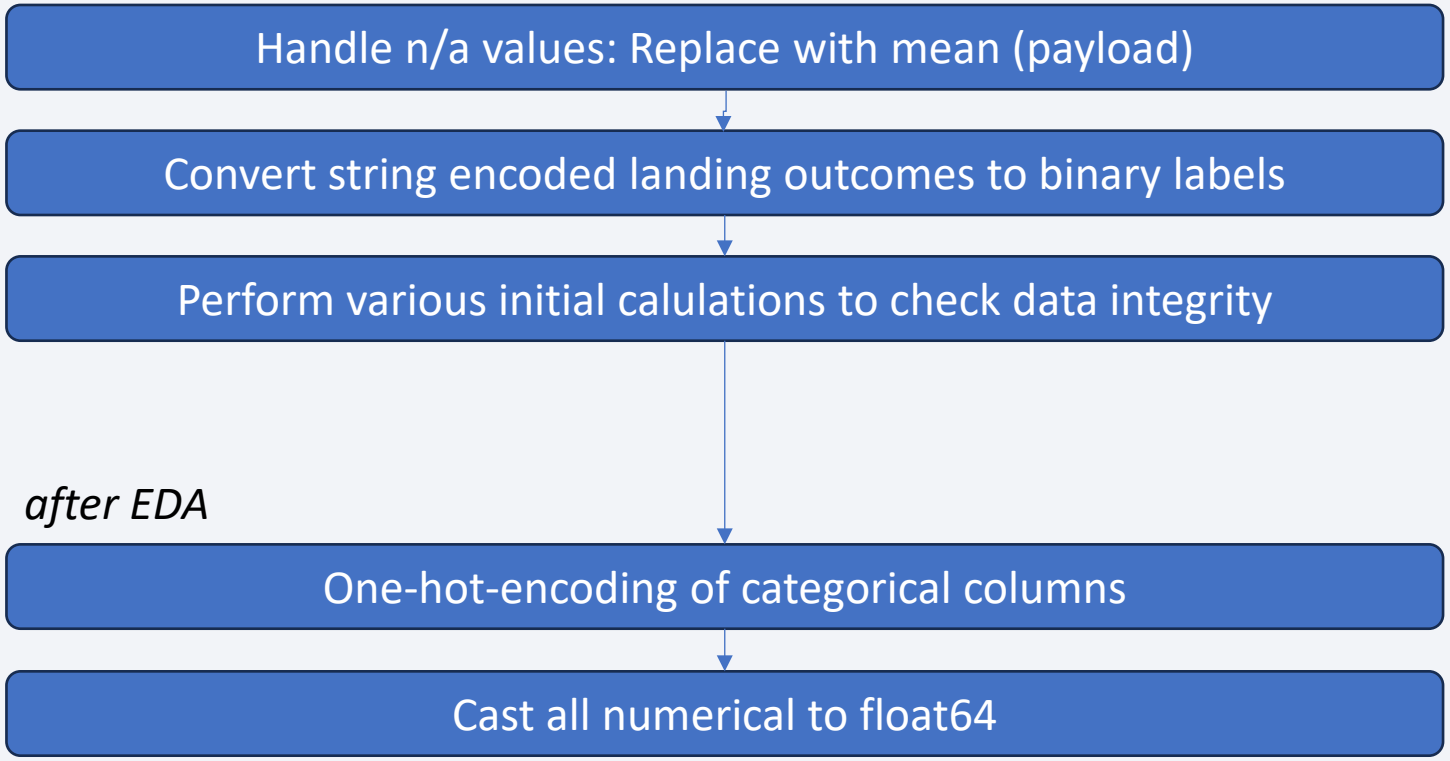


Code Link

https://github.com/WolfgangHuang/ibmdatasciencecourse/blob/main/capstone2-jupyter-labs-webscraping_WH.ipynb

Data Wrangling

The flow chart to the right shows the main steps of data wrangling.



Code Link

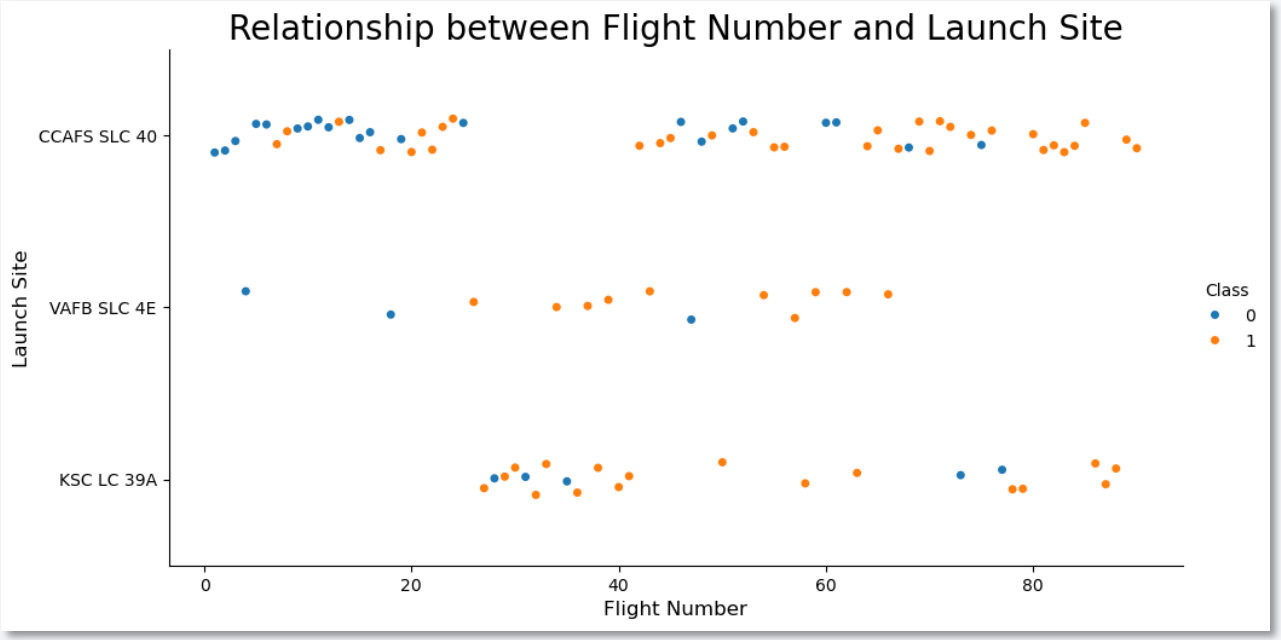
https://github.com/WolfgangHuang/ibmdatasciencecourse/blob/main/capstone3-labs-jupyter-spacex-Data%20wrangling_WH.ipynb

EDA with Data Visualization

EDA visualizations to identify significant success factors included the following relationships:

1) flight number vs. launch site
2) payload vs. launch site
3) success rate vs. orbit type
4) flight number vs. orbit type
5) payload vs. orbit type
6) launch success yearly trend

Code Link
https://github.com/WolfgangHuang/ibmdatasciencecourse/blob/main/capstone4-jupyter-labs-eda-dataviz_WH.ipynb



EDA with SQL

The following SQL EDA queries were performed:

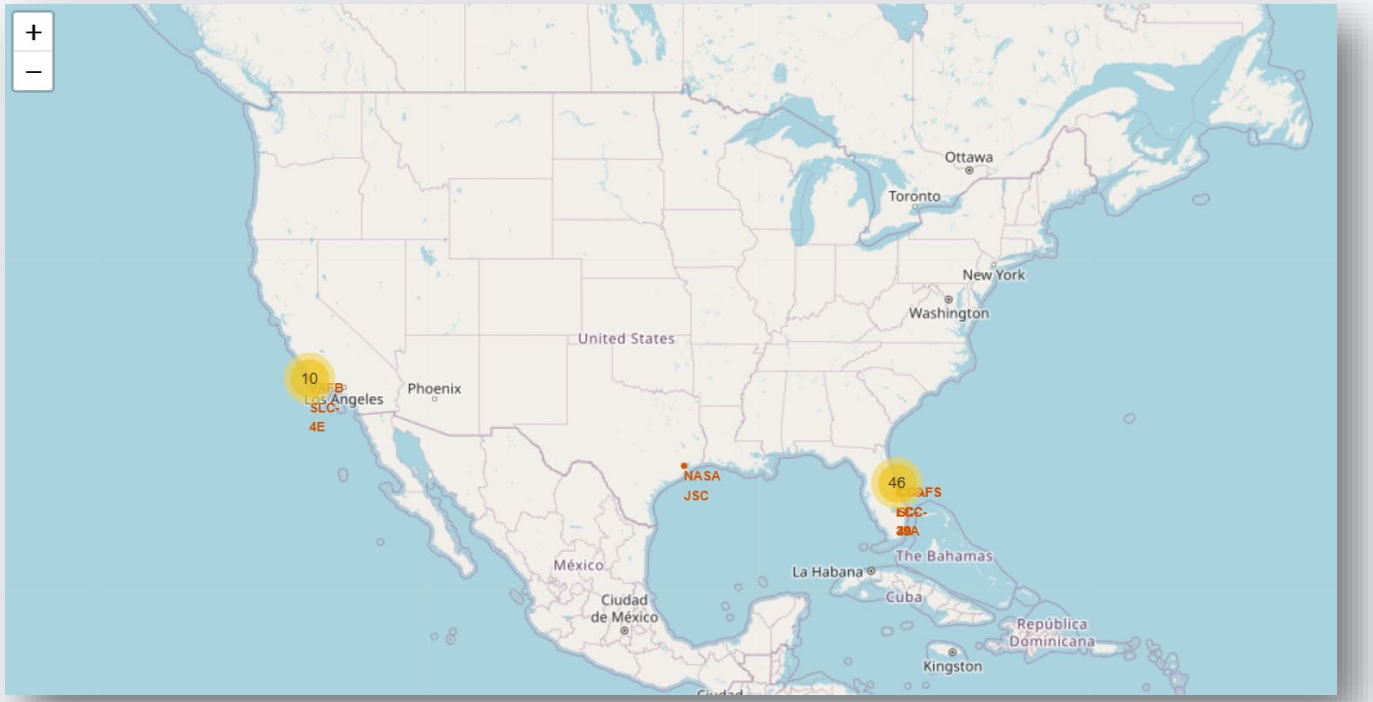
- Unique launch site names
- Top 5 records with launch sites starting with “CCA”
- Total payload mass carried by boosters launched by NASA
- Average payload mass carried by booster version F9 v1.1
- Date of first successful landing on ground pad
- Boosters with successful drone ship landing and payload between 4000K and 6000Kg
- Total number of successful and failed missions
- Boosters with max payload
- Months with failed landings on drone ships
- Ranked landing outcome between 2010-06-04 and 2017-03-20

Code Link

https://github.com/WolfgangHuang/ibmdatasciencecourse/blob/main/capstone5-jupyter-labs-eda-sql-coursera_sqllite_WH.ipynb

Build an Interactive Map with Folium

Folium has been used to visualize launch sites on an interactive map. Circles were used to mark areas, icons to label the areas, and colored markers in marker clusters to show launch outcomes. Polylines were used to show distances.



Code Link

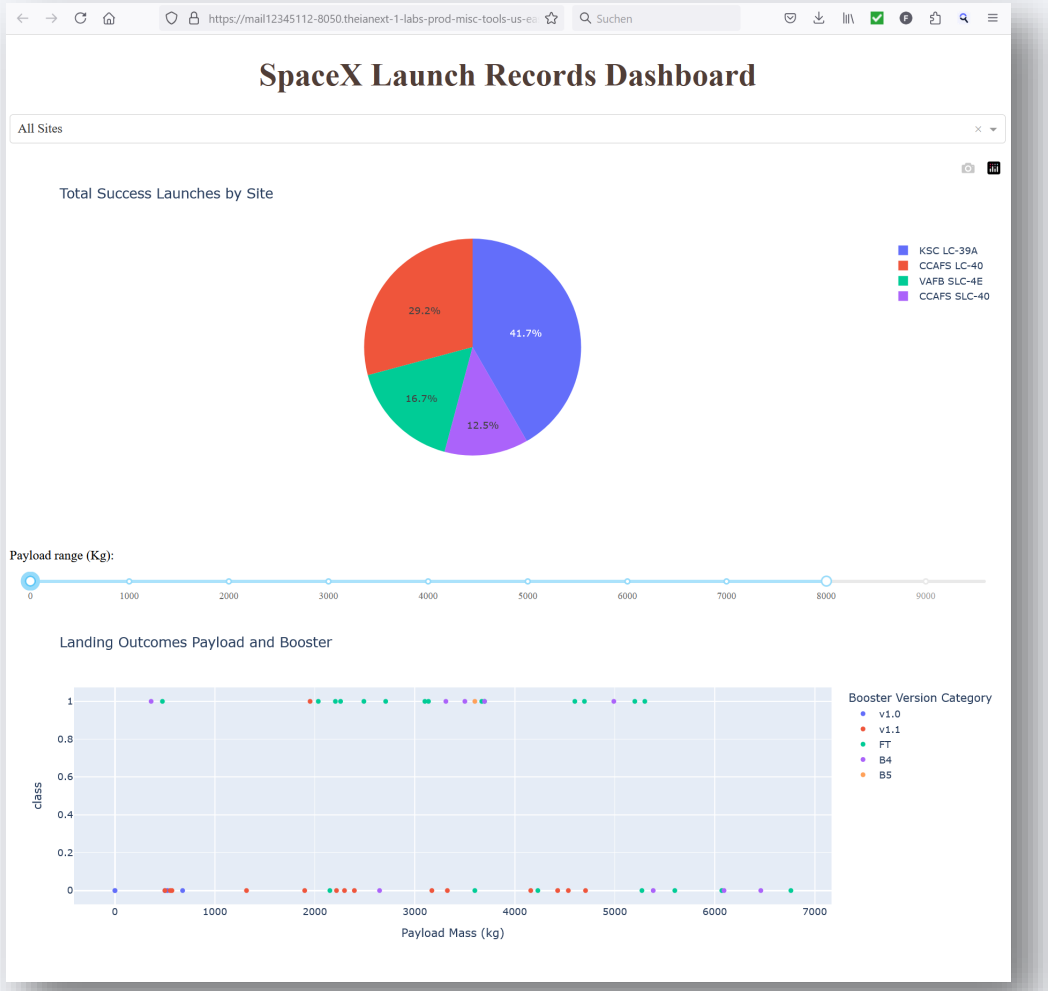
https://github.com/WolfgangHuang/ibmdatasciencecourse/blob/main/capstone6-lab-jupyter-launch-site-location_WH.ipynb

Build a Dashboard with Plotly Dash

The dashboard shows two plots:

1) total successful launches by site
2) landing outcome by payload and booster.

These plots – apart from their insights – nicely demonstrate the interactive features of Dash dashboards.



Code Link
https://github.com/WolfgangHuang/ibmdatasciencecourse/blob/main/capstone7-spacex_dash_app_WH.py

Predictive Analysis (Classification)

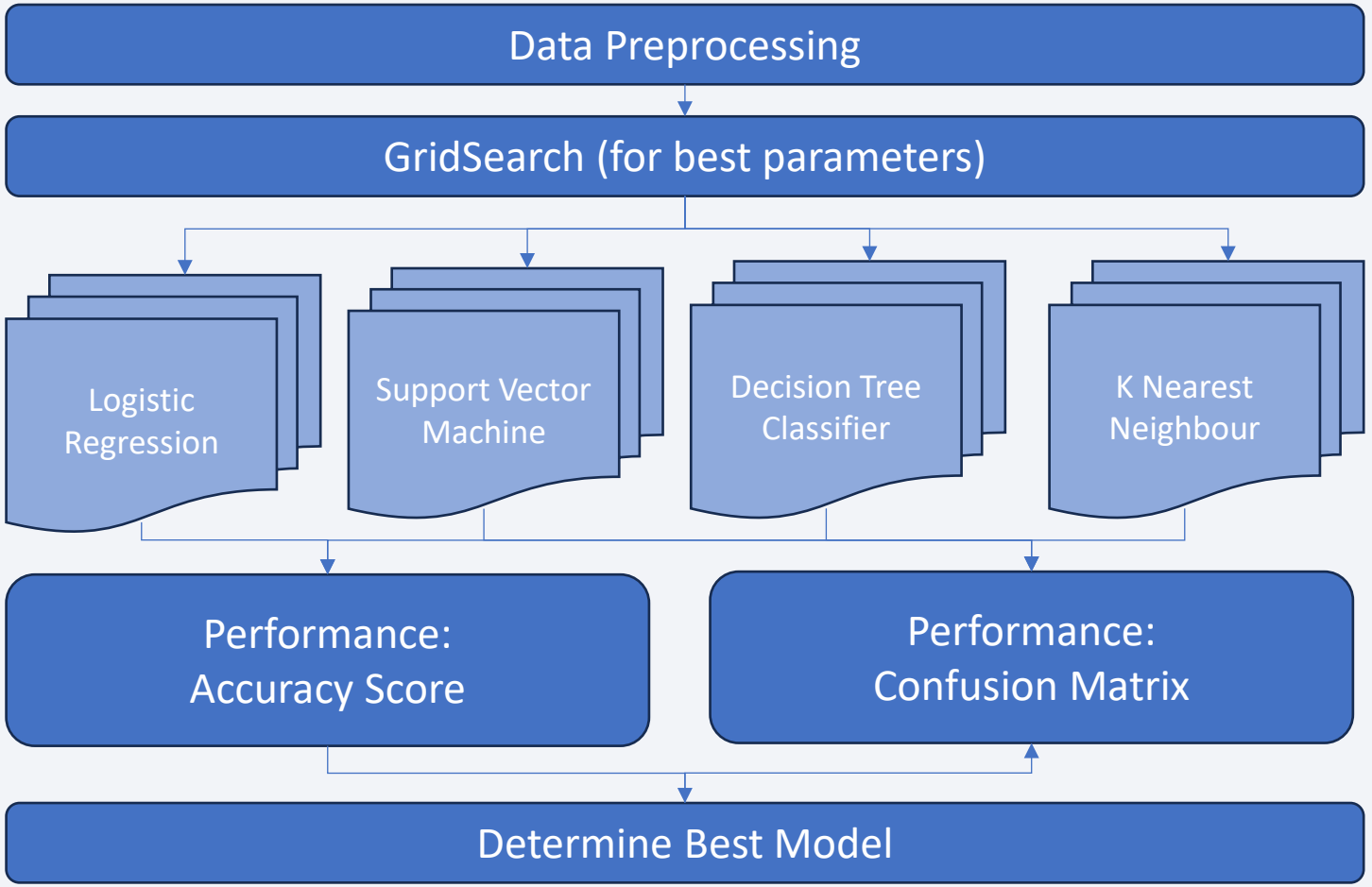
Before running the classification algorithms, data was preprocessed:

1. Separate features (X) and targets (Y)
2. Scale features
3. Divide X and Y into training and test sets.

Next, GridSearch was used to identify the best parameters for several models.

Training and test accuracy was determined for all algorithms using best parameters.

Code Link
https://github.com/WolfgangHuang/ibmdatasciencecourse/blob/main/capstone8-SpaceX-Machine-Learning-Prediction-Part-5_WH.ipynb



Section 2

Insights drawn from EDA

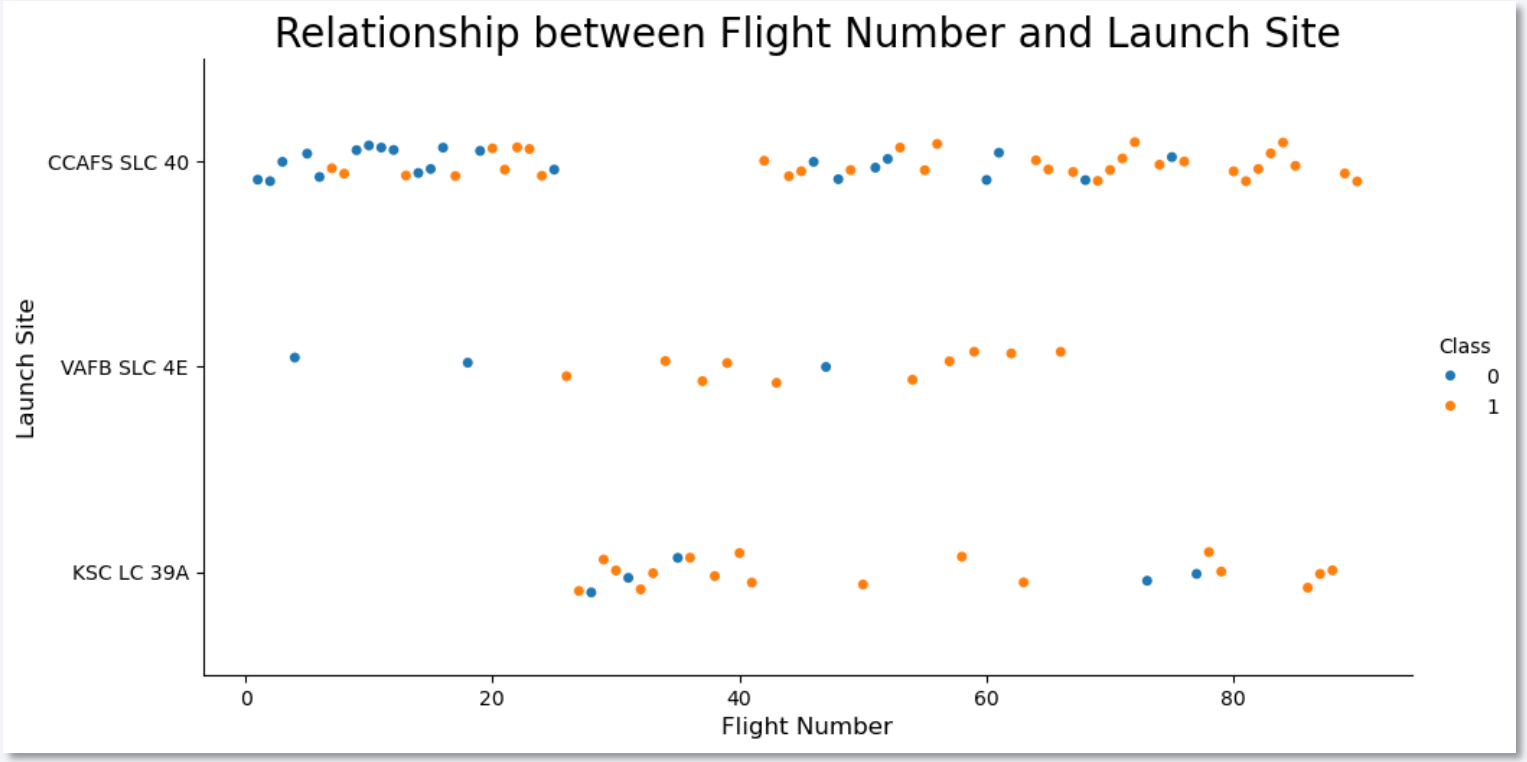
Flight Number vs. Launch Site

Insights

As the flight number correlates with time, we can conclude that mission become more and more successful over time, most likely due to increasing experience.

All past 5 missions at all sites have been successes.

Overall, the Cape Canaveral site has most starts, earliest missions, and thereby also most failures.

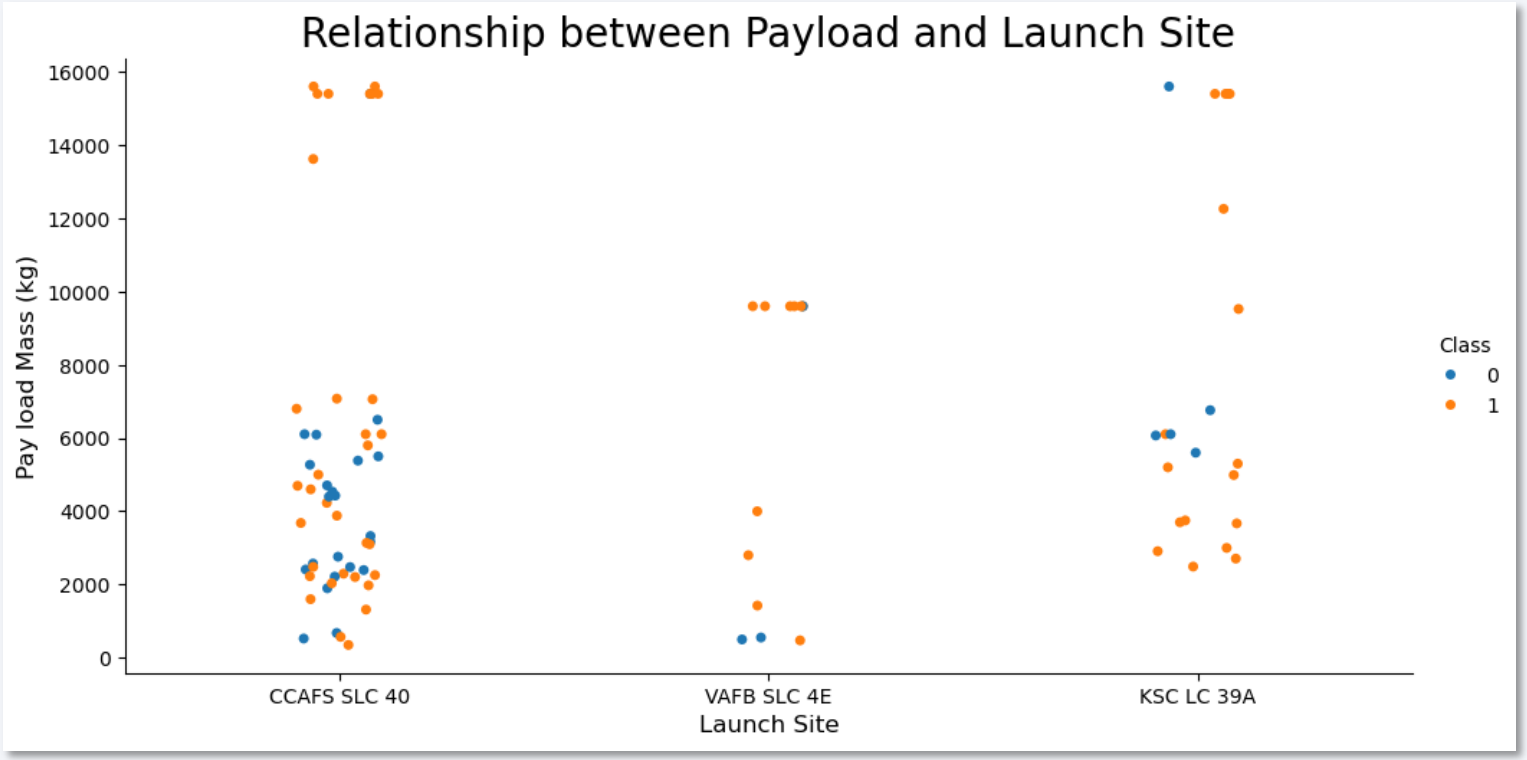


Payload vs. Launch Site

Insights

We can see that certain payload missions are preferably launched at specific sites: VAFB is particularly successful for medium payload amounts, while CCAFS has the highest success rate for very high payloads.

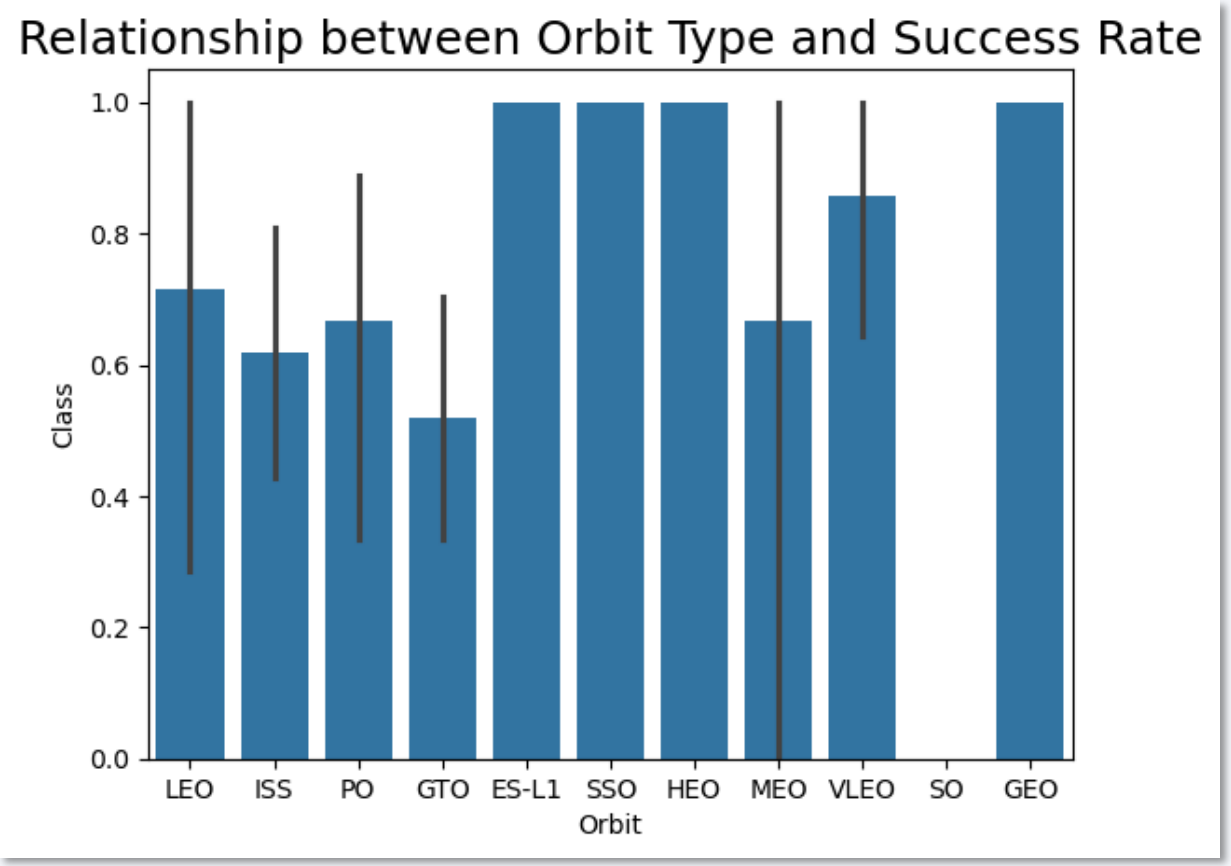
VAFB does not seem to be equipped for very high payloads at all, as there are not starts recorded.



Success Rate vs. Orbit Type

Insights

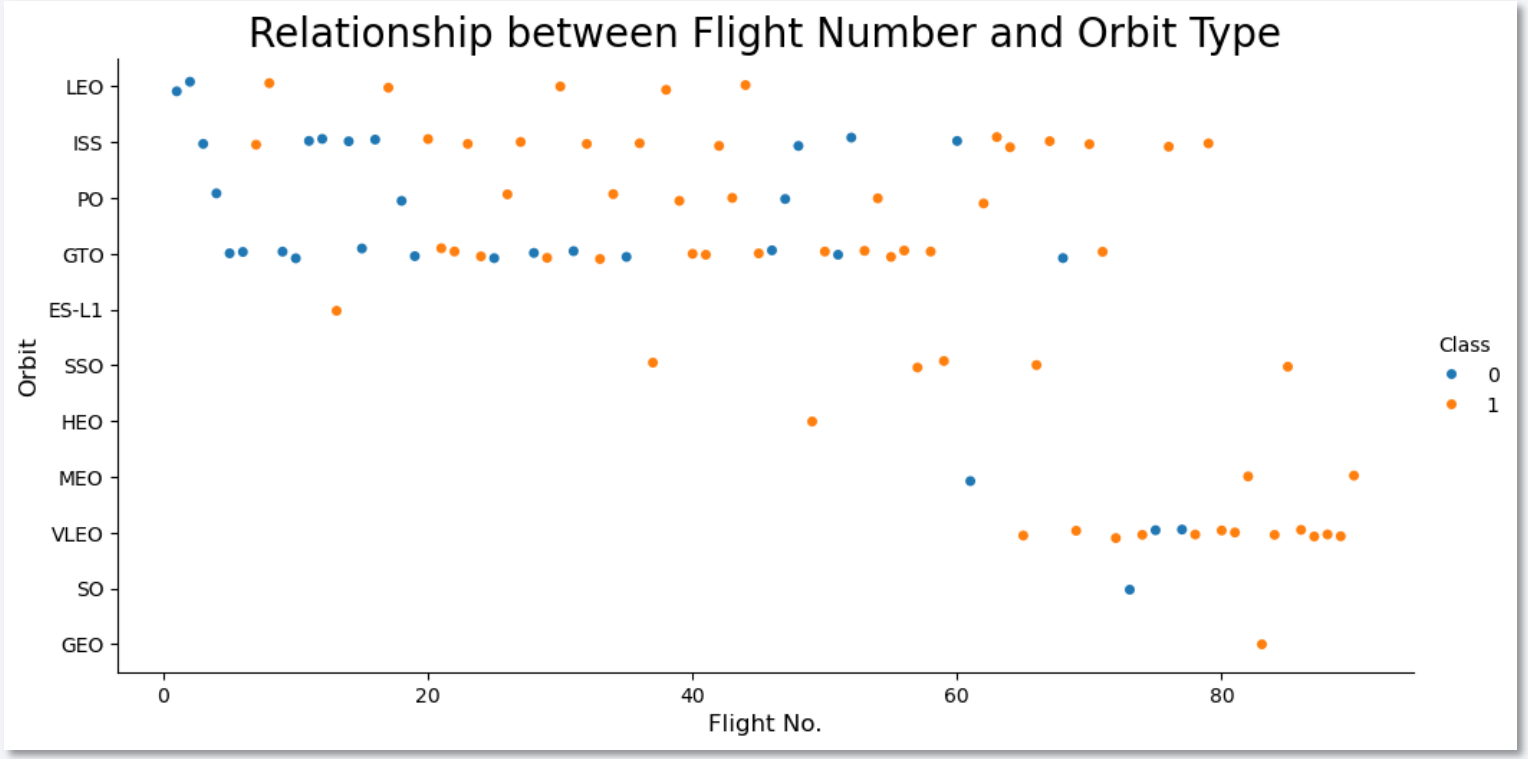
Some orbits seem to have a 100% success rate, such as ES-L1, SSO, HEO and GEO. It is, however, not clear from this plot why that is. Also, it can be suspected that there are confounding variables.



Flight Number vs. Orbit Type

Insights

In addition to earlier findings, it appears that the most recent missions mostly have been to rather low orbits, such as the ISS and VLEO orbits.

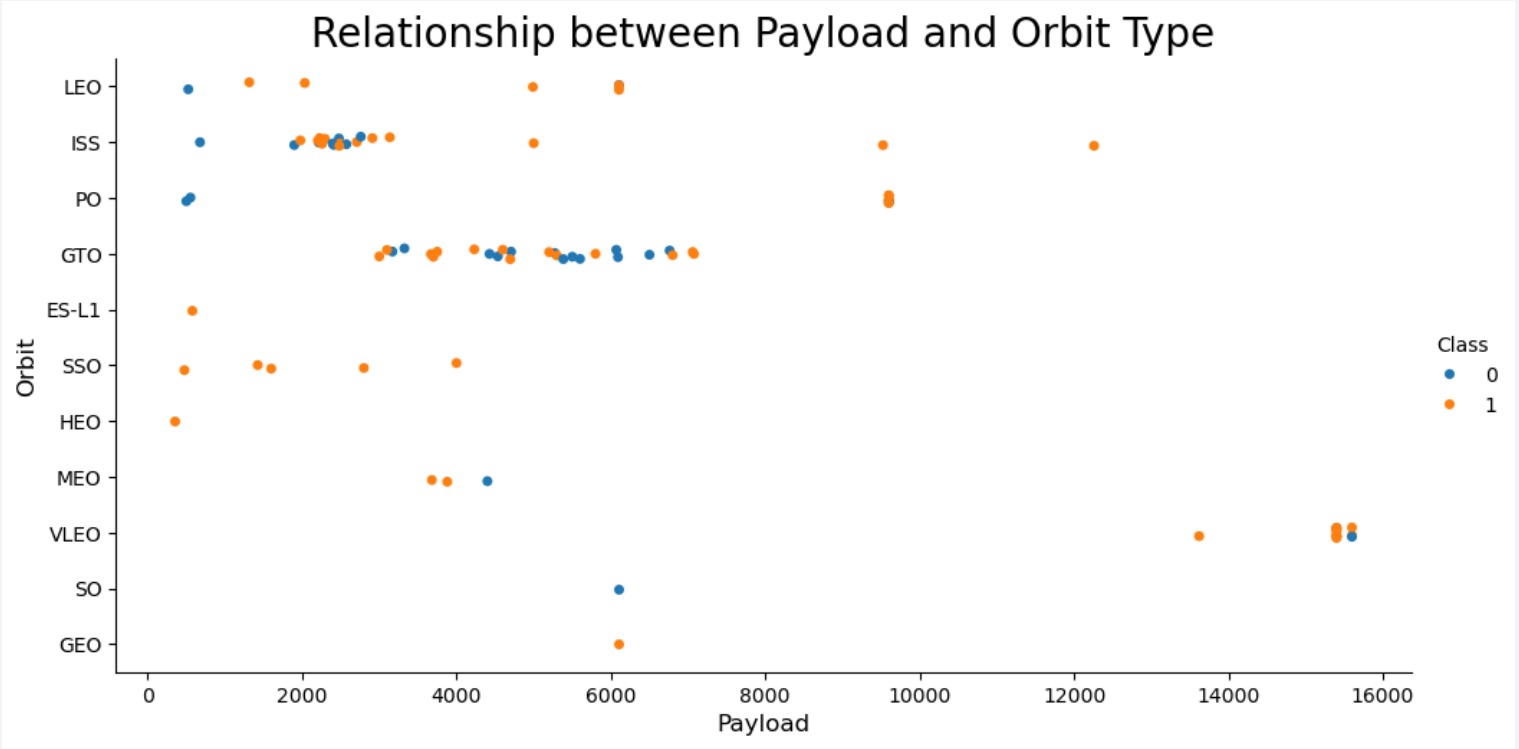


Payload vs. Orbit Type

Insights

This plot shows an interesting cluster structure. There is a clear correlation between certain payload ranges and specific orbits. This would hint a very specific (and repeated) mission types.

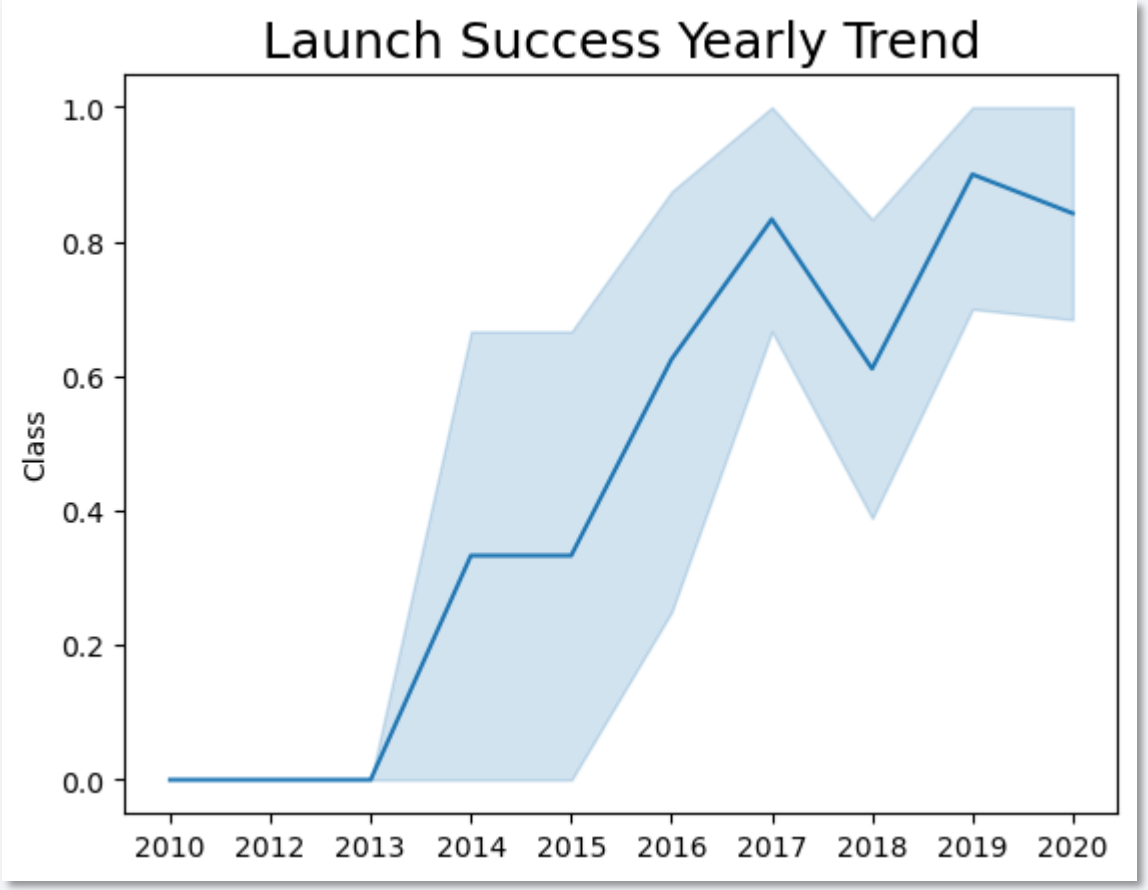
For example, the mission to the ISS orbit may refer to ISS delivery missions, which have clearly defined mission characteristics.



Launch Success Yearly Trend

Insights

Obviously, the missions (landing outcome) become more successful over time, with slight setbacks in 2018 and 2020.



All Launch Site Names


```
%sql SELECT DISTINCT [Launch_Site] FROM SPACEXTBL
```

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

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

There are four unique launch site names, as shown in the table above.

Launch Site Names Begin with 'CCA'


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

Python

```
 * sqlite:///my_data1.db
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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

Filter results for 5 records where launch sites begin with the string 'CCA'

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

The total payload mass carried by boosters launched by NASA is 45,596 KG.

Average Payload Mass by 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

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

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

The date of the first successful landing outcome on ground pad is: 2015-12-22.

Successful Drone Ship Landing with Payload between 4000 and 6000


```
%%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.
```

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

The table above shows the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.

Total Number of Successful and Failure Mission Outcomes


```
%sql SELECT Mission_Outcome, COUNT(*) AS Count FROM SPACEXTBL GROUP BY Mission_Outcome
```

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

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

The table above shows the total number of successful and failure mission outcomes.

Boosters Carried Maximum Payload


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

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

List of names of boosters which have carried the maximum payload mass.

2015 Launch Records


```
%%sql
SELECT substr(Date, 6,2), Landing_Outcome, Booster_version, Launch_Site FROM SPACEXTBL
WHERE substr(Date, 0,5) = '2015' AND Landing_Outcome = 'Failure (drone ship)'
```

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

substr(Date, 6,2)	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

Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

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


```
%%sql
SELECT Landing_Outcome, COUNT(*) AS Count FROM SPACEXTBL
WHERE Date > '2010-06-04' AND Date < '2017-03-20' GROUP BY Landing_Outcome ORDER BY Count DESC
```

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

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

Ranked count of landing outcomes with various criteria

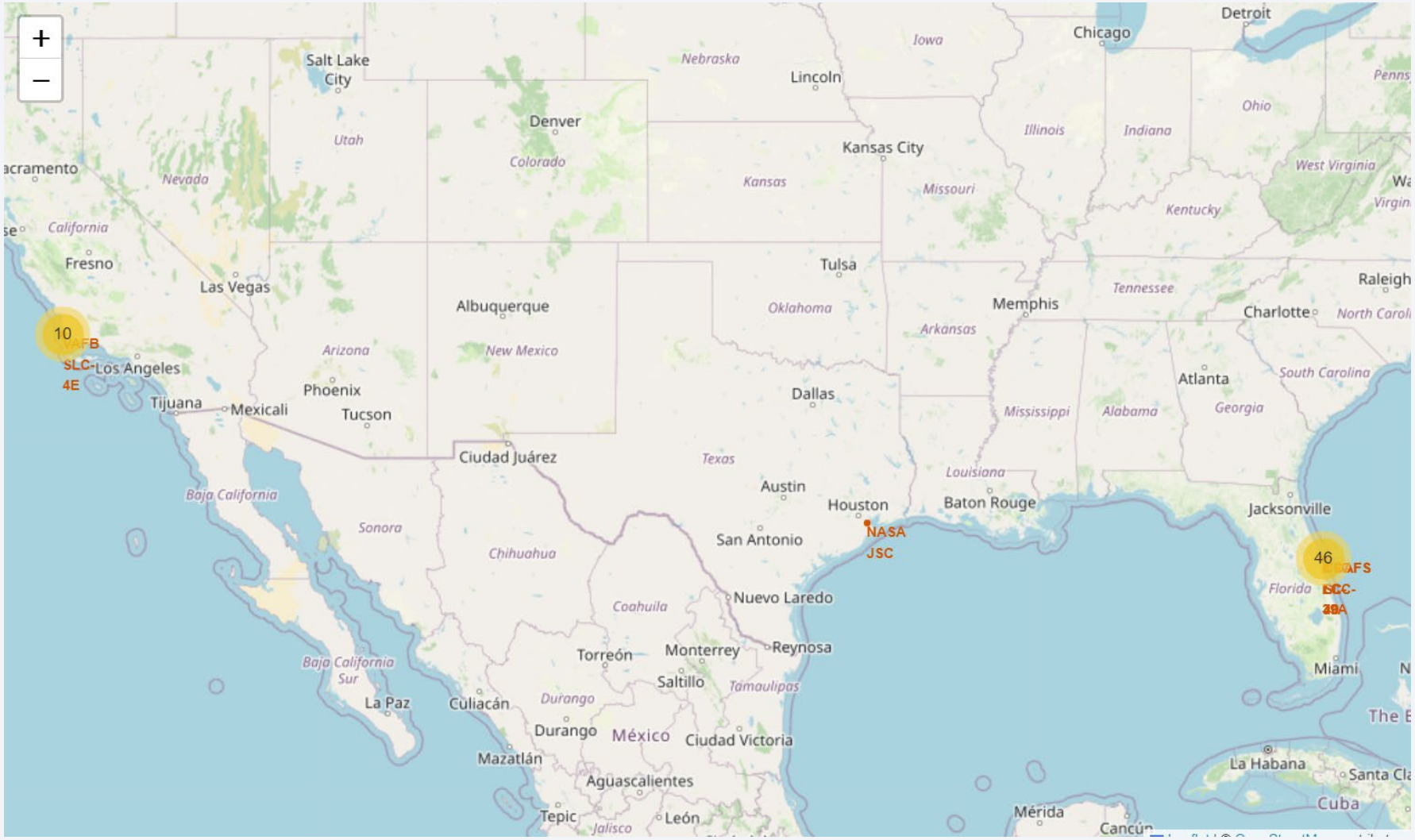

Section 3
Launch Sites
Proximities Analysis

Map of SpaceX Launch Sites

Insights

The map shows all launch sites used by SpaceX.

Besides the fact that most of these already existed before SpaceX started launching rockets, it is easy to see that all are close to the coast (probably to safely direct failed starts to the open sea) as well as relatively close to the equator.



Landing Outcomes for CCAFS LC-40

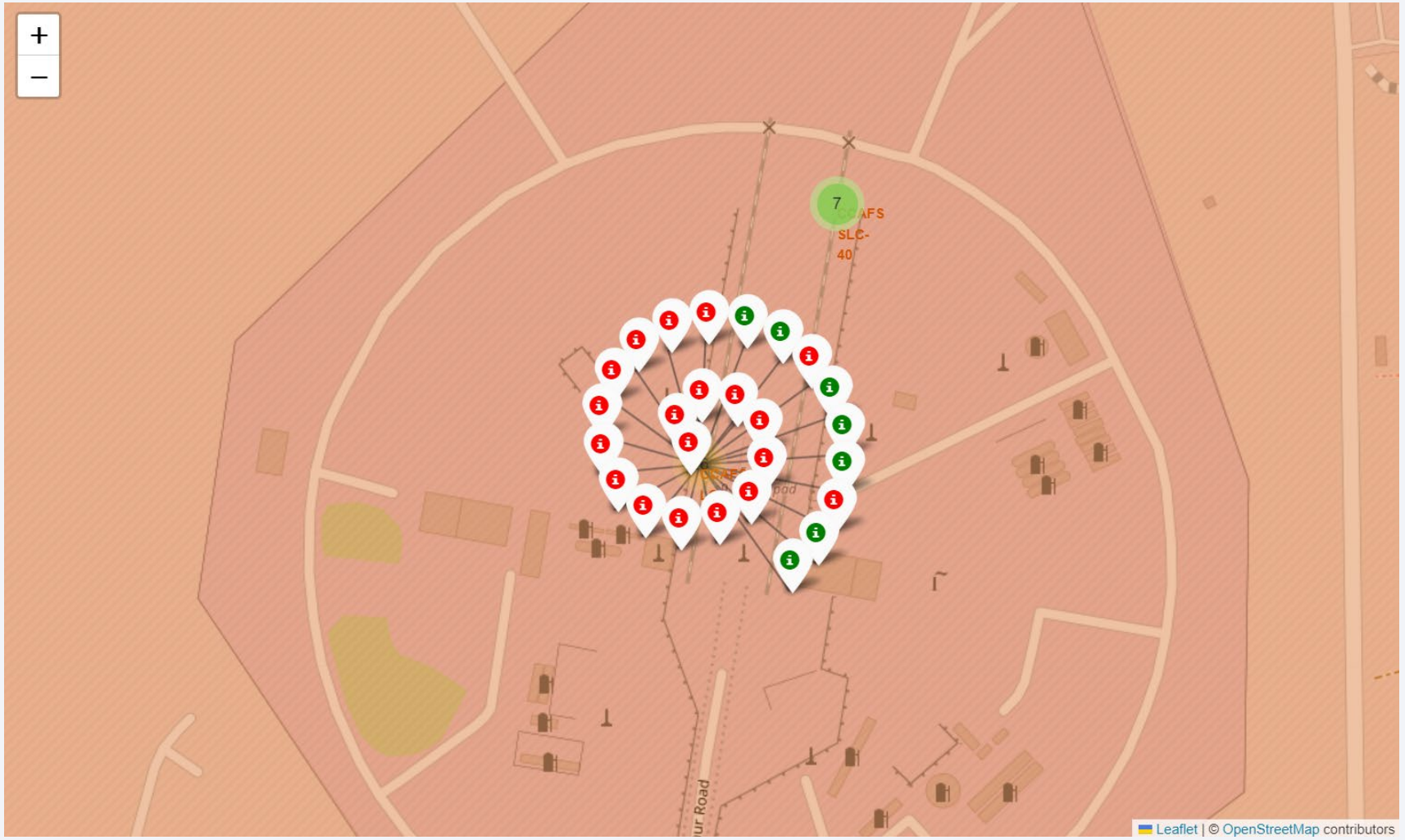
Insights

This graph shows the landing outcomes for the Cape Canaveral site CCAFS LC-40.

The launches are arranged in a spiral, with the earliest being the innermost.

It is easy to see that the missions have become more successful over time.

The marker “7” refers to another site.

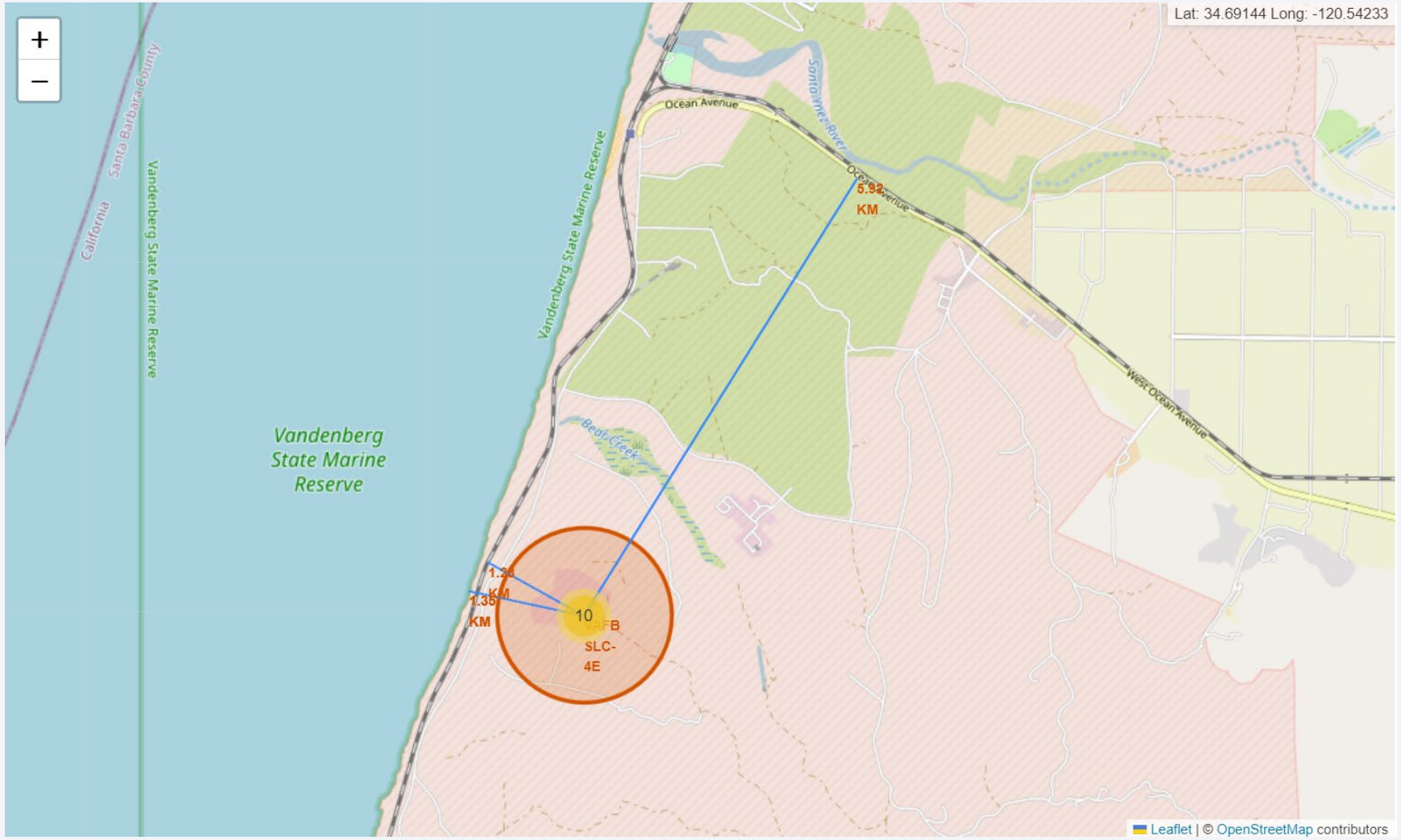


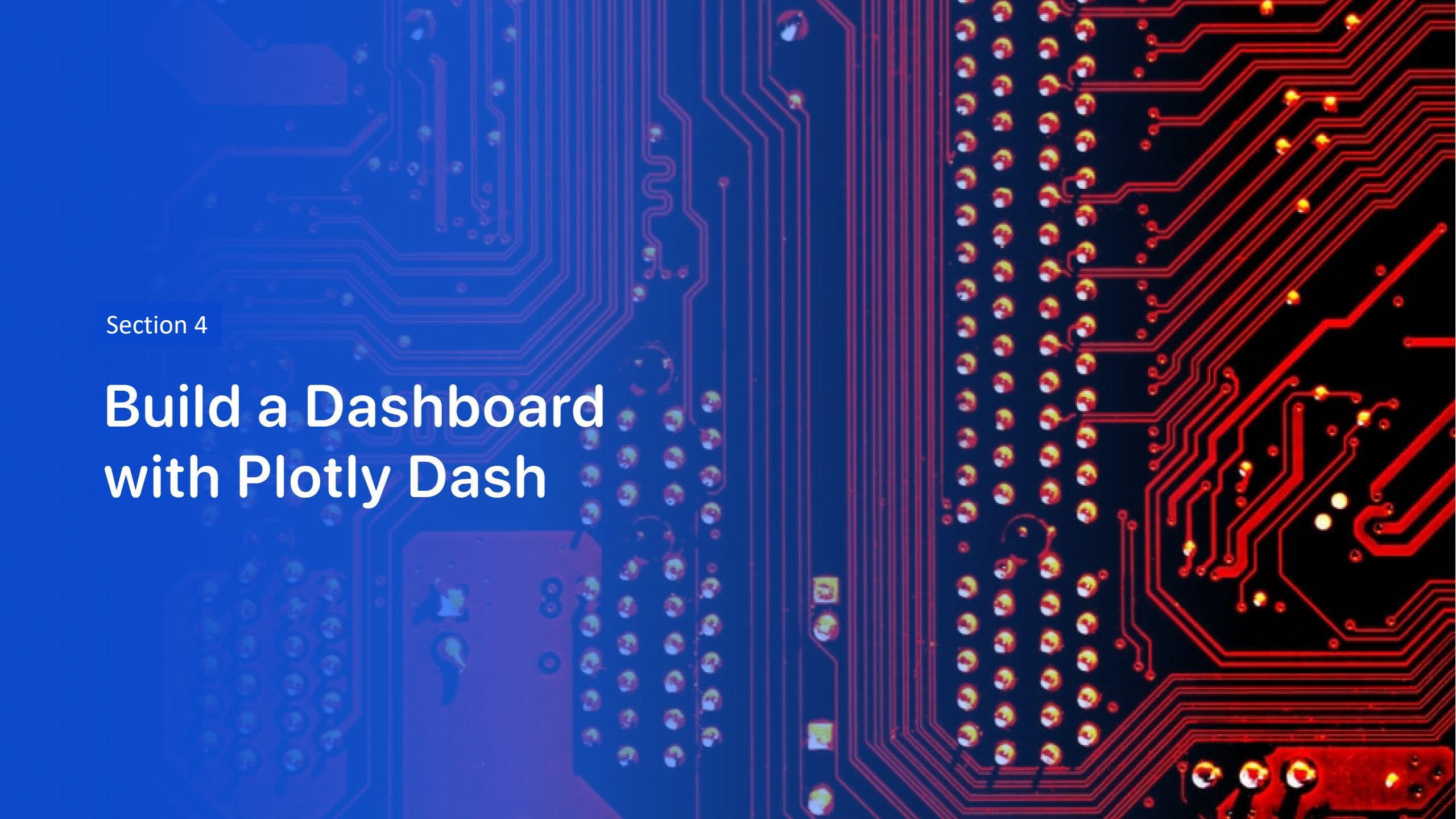
Launch Site Proximities

Insights

This graph shows nearby landmarks and structures such as the coastline, railway, and highway (major road).

To no surprise, the launch site is well connected, allowing easy transfer of people and goods.




Section 4
Build a Dashboard with Plotly Dash

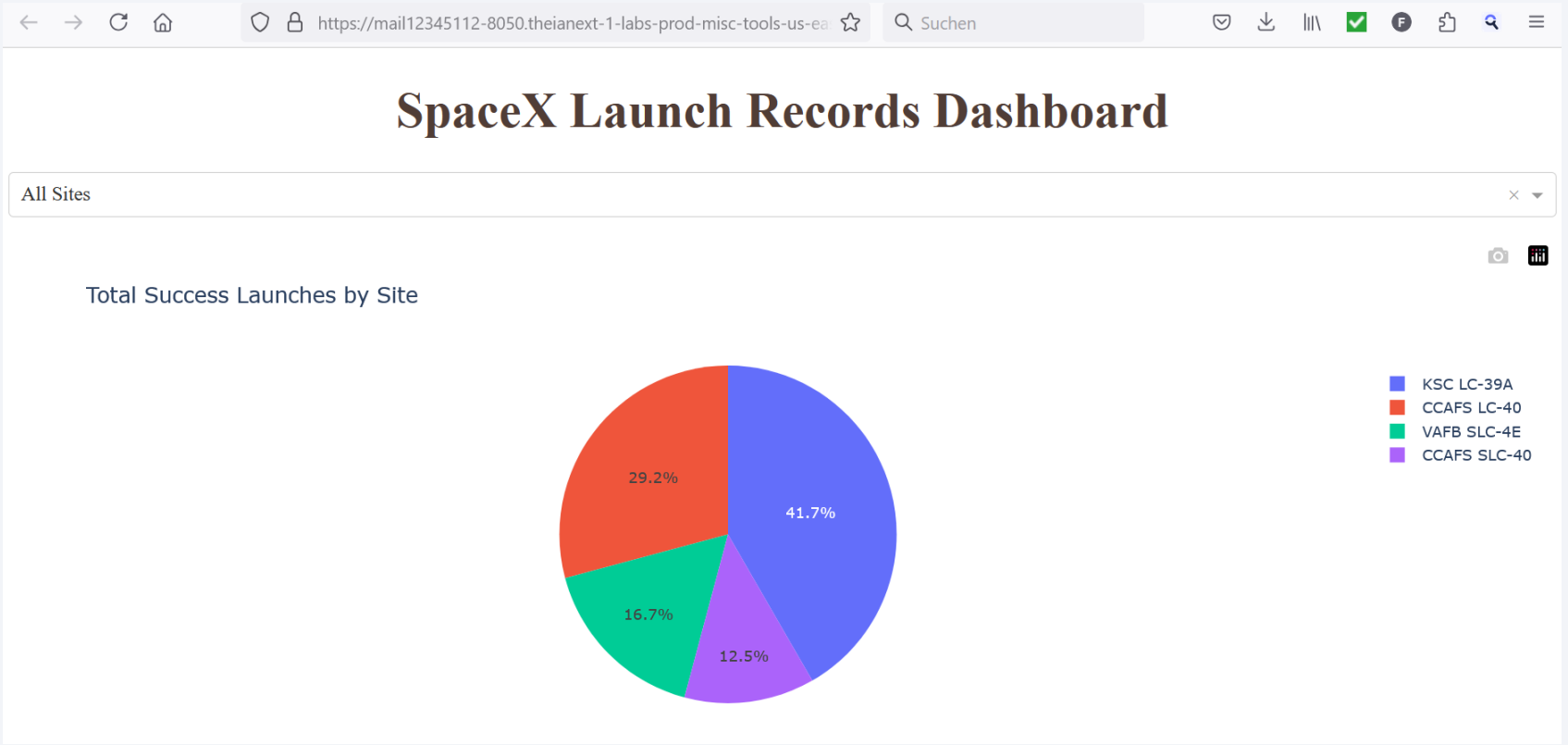
Dashboard: All Sites

Insights

This screenshot shows the dash app with the “Total Success Launches by Site” pie chart.

The site dropdown is placed on top of it.

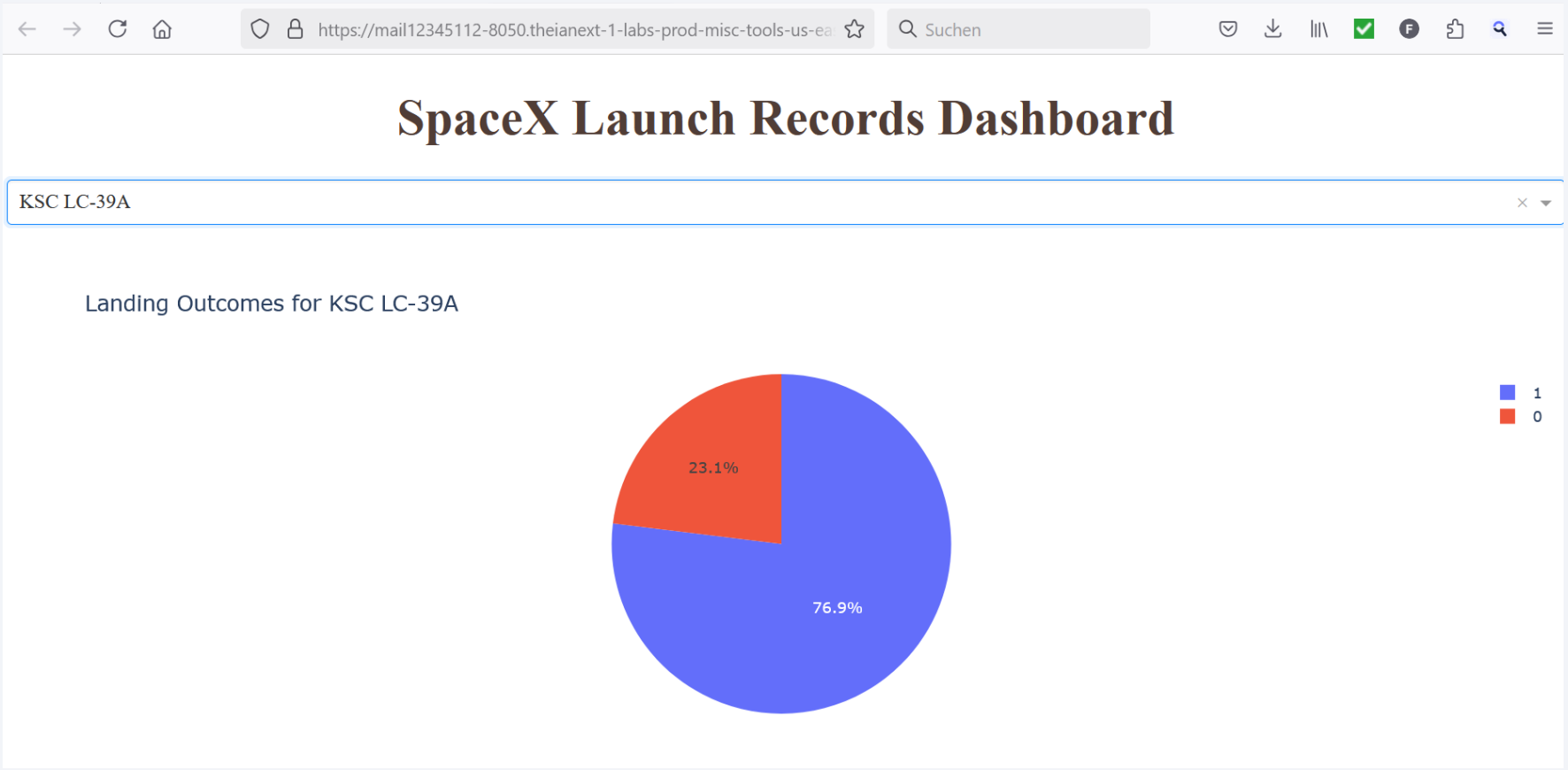
We can see that KSC LC-39A has the highest proportion of successful launches.



Dashboard: Results for KSC LC-39A

Insights

This screenshot shows the dash app with the results for KSC LC-39A, which has 76.9% successful landing outcomes.



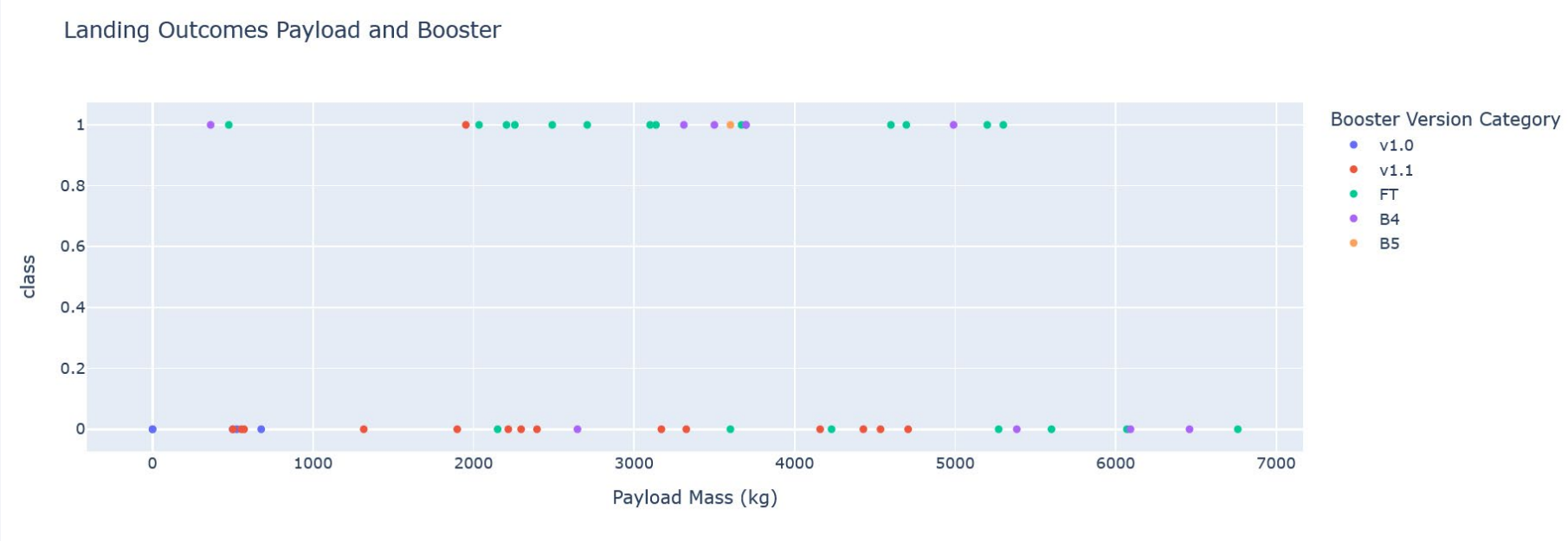
Dashboard: Landing Outcomes by payload

Insights

Most successful landing occurs in the mid-range payloads: between 2000-4000 kg, and also around 5000 kg.

The FT booster seems to be most reliable, while v1.1 performs poorly.

Very low and very high payload missions are associated to failures, the earlier category most likely being test missions.



Section 5

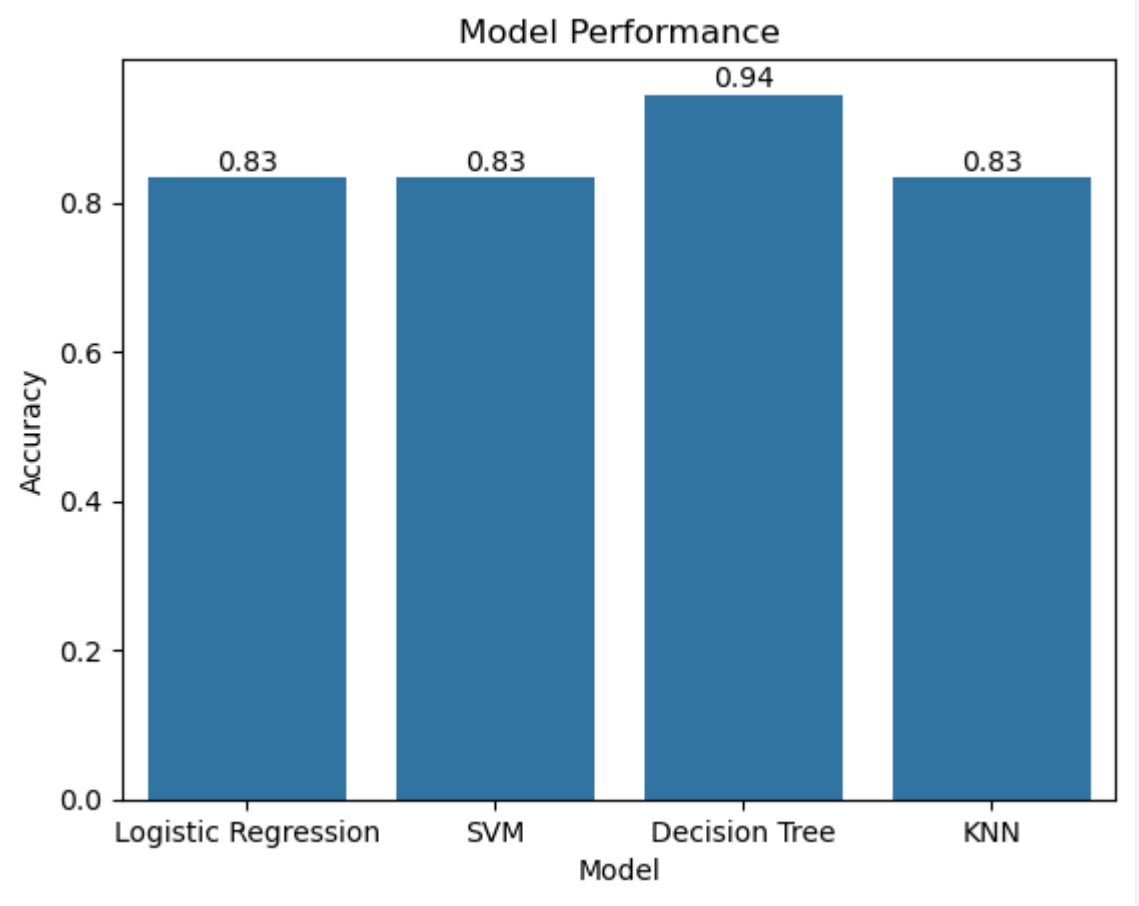
Predictive Analysis (Classification)

Classification Accuracy

Insights

As it turns out, the decision tree classifier performs best with a test accuracy score of 0.94.

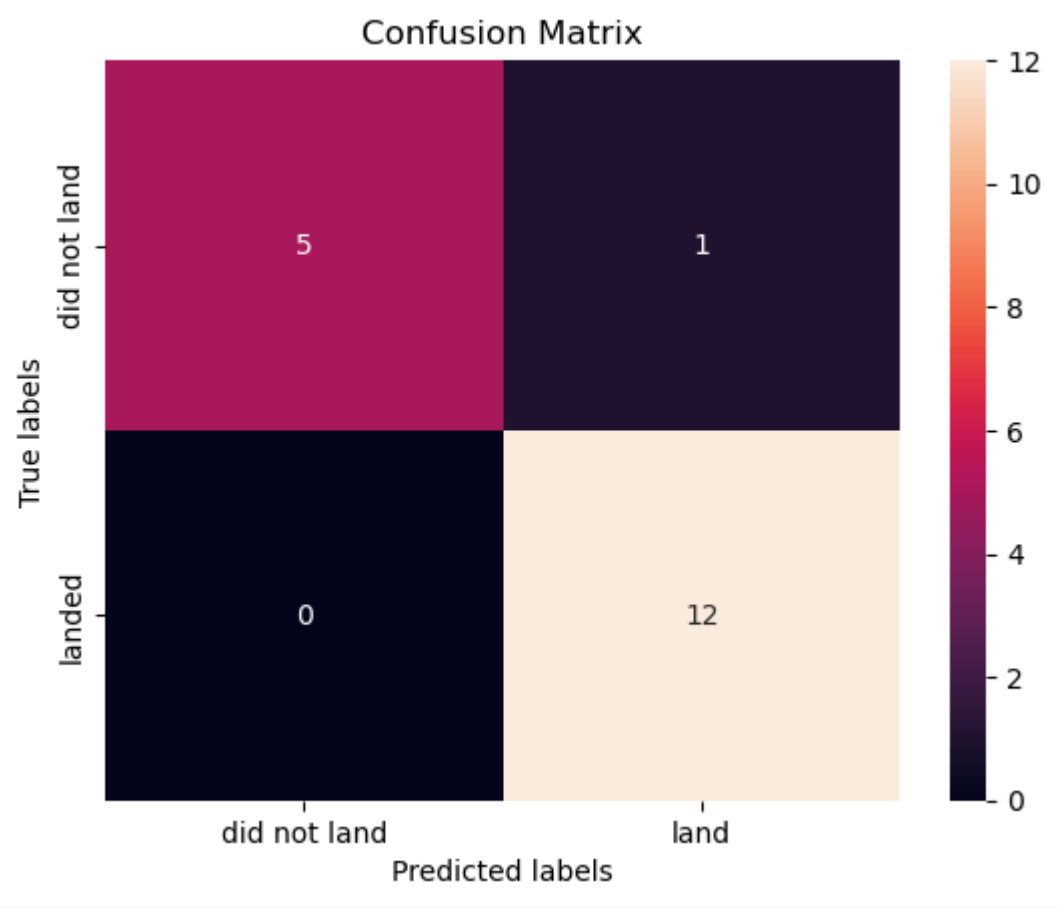
All other three models have identical (lower) results on their test accuracy.



Confusion Matrix

Insights

Whereas the other contestants have values 3/3 in the first row, meaning they only have a 50% prediction accuracy for the actual “did not land” missions, the decision tree classifier has 5/1, meaning only 1 out of 6 predictions is wrong.



Conclusions

If we further analyze the feature importance as determined by the Decision Tree Classifier, quite surprisingly, there are only three features that are important:

- `Legs_True 0.924813`
- `FlightNumber 0.055131`
- `Reused_False 0.020056`

In essence, we can say – and this is almost funny – that boosters need to have legs to successfully land. Furthermore, more modern versions work better, and also the success rate is higher if the booster is new instead of reused.

All other factors may play a role, and they may be co-dependent with other features, but the decision tree classifier only uses these three features for predictions.

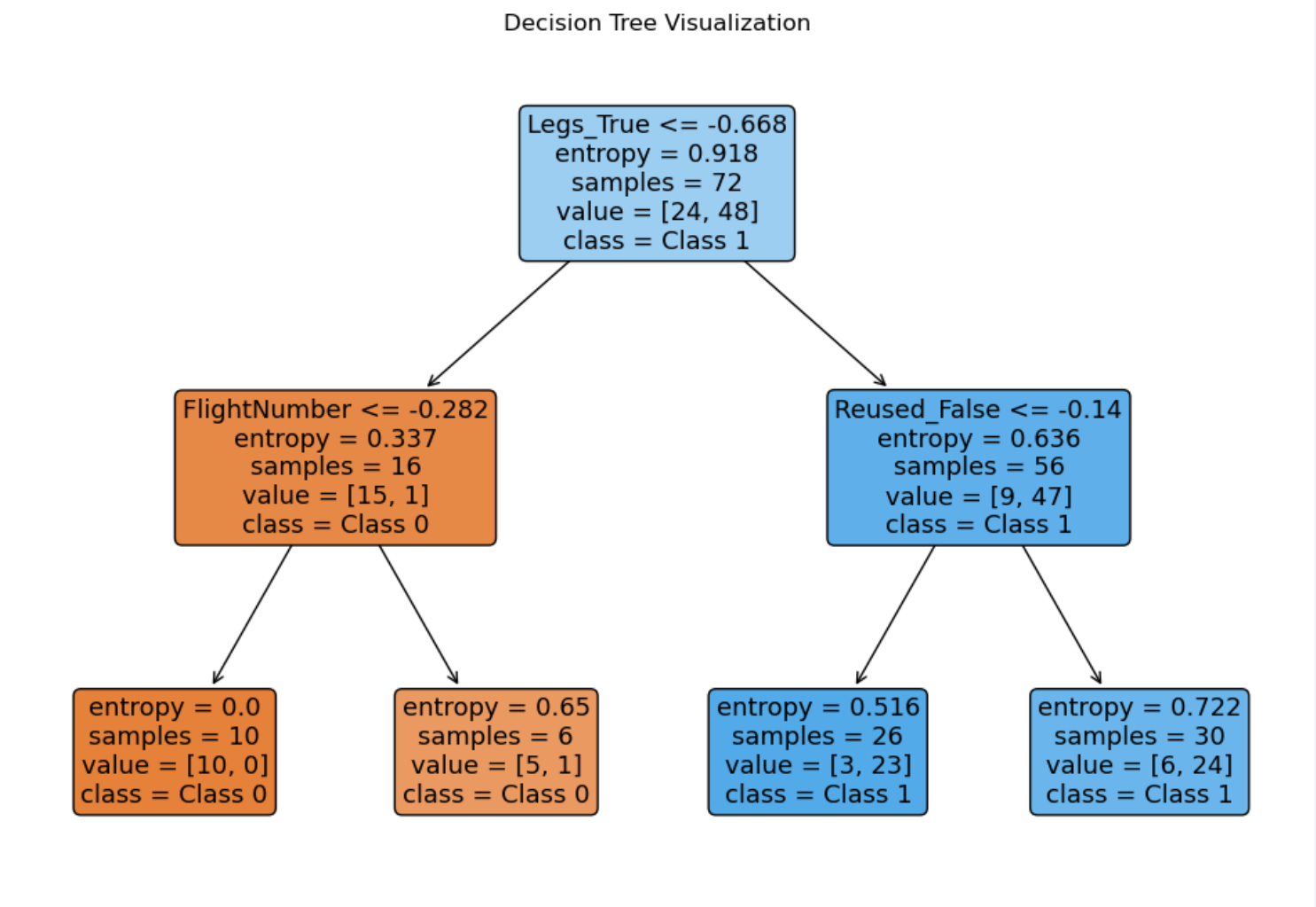
Conclusions

The decision tree can also be shown graphically.

Important note: Both the training and particularly the test sample had very small numbers of samples, which may negatively affect the prediction quality (and also the accuracy of identifying important features).

Also note: Although the subtrees split result in the same class, they reduce entropy, which is why they occur.

“Legs” are actually a binary feature, but after scaling, False corresponds to -0.668 (instead of 0).



Thank you!