Winning Space Race with Data Science

Tatiana Miklashevich
25.08.2024

Outline

Executive Summary

This paper analyzes data on rocket launches and analyzes success depending on various factors. The methodology includes obtaining and collecting data from various sources, data pre-processing, statistical and visual analysis data, model building and estimation accuracy.

As a result, it was possible to analyze the data, build several models and visualize the dynamics on Dachboad.

Introduction

Space has always been an intriguing topic for human scientific research. In the modern world, space flights have not only become a reality, but are also carried out by both public and private companies. At the same time, the cost of the flight is reduced due to the development of science.

The purpose of the work is to analyze the dependence of the success of space flights depending on various factors based on real data on completed flights.

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology:

- We used the SpaceX API

Perform data wrangling

- We normalized the data, converted the data to a dataframe, and filtered out the zeros

Perform exploratory data analysis (EDA) using visualization and SQL

Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models

- We used Python libraries to create models, split the data into training and test data, trained the models, tested them, and obtained the corresponding accuracy estimates.

Data Collection

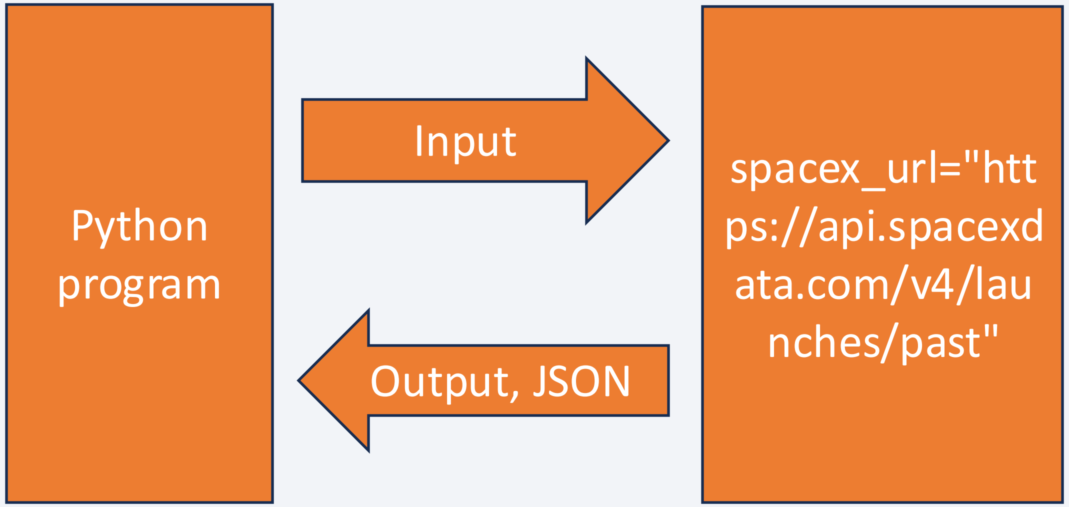
We used a URL to target the API to get data about past runs. We'll issue a get request to get the launch data. The response was in the form of JSON, namely a list of JSON objects, each of which represents a run. To convert this JSON into a data frame, we used the json_normalize function..

This API will provide us with launch data, including information about the rocket used, payload delivered, launch specifications, landing characteristics, and landing results.

Data Collection – SpaceX API

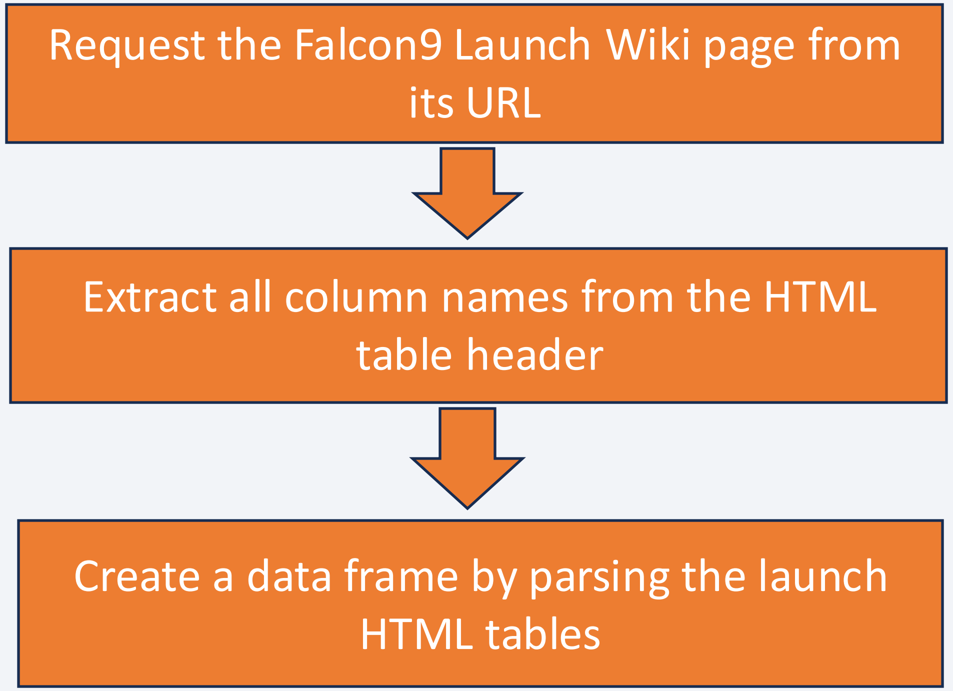
- This API will provide us with launch data, including information about the rocket used, payload delivered, launch specifications, landing characteristics, and landing results.

- https://github.com/TatianaMikl/testrepo/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- We performed web scraping to collect historical records of Falcon 9 launches from the Wikipedia page titled “List of Falcon 9 and Falcon Heavy launches.” https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- https://github.com/TatianaMikl/testrepo/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

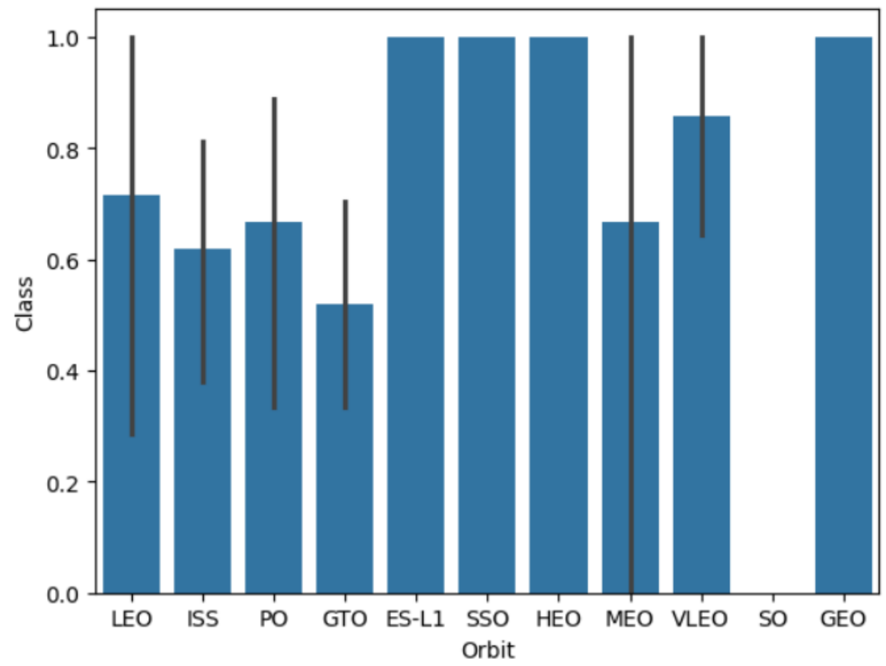
We conducted exploratory data analysis (EDA) to find some patterns in the data and determine what we would call training supervised models. We converted these results into training labels, where 1 means the booster landed successfully, 0 means it failed.

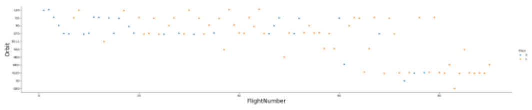
We converted these results into training labels, where 1 means the booster landed successfully, 0 means it failed.

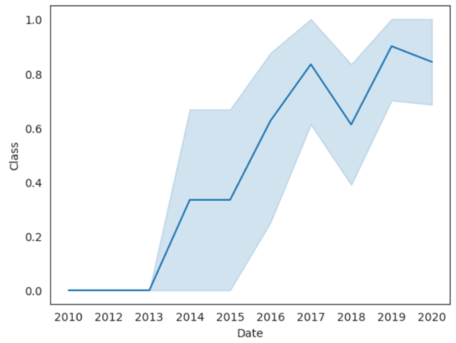
https://github.com/TatianaMikl/testrepo/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- We visualized:
- the relationship between Flight Number and Launch Site
- the relationship between Payload Mass and Launch Site
- the relationship between success rate of each orbit type
- the relationship between FlightNumber and Orbit type
- the relationship between Payload Mass and Orbit type
- the launch success yearly trend
- https://github.com/TatianaMikl/testrepo/blob/main/edadataviz.ipynb







EDA with SQL

SQL queries :

the names of the unique launch sites in the space mission

launch sites begin with the string 'CCA'

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

average payload mass carried by booster version F9 v1.1

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

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

total number of successful and failure mission outcomes

names of the booster_versions which have carried the maximum payload mass

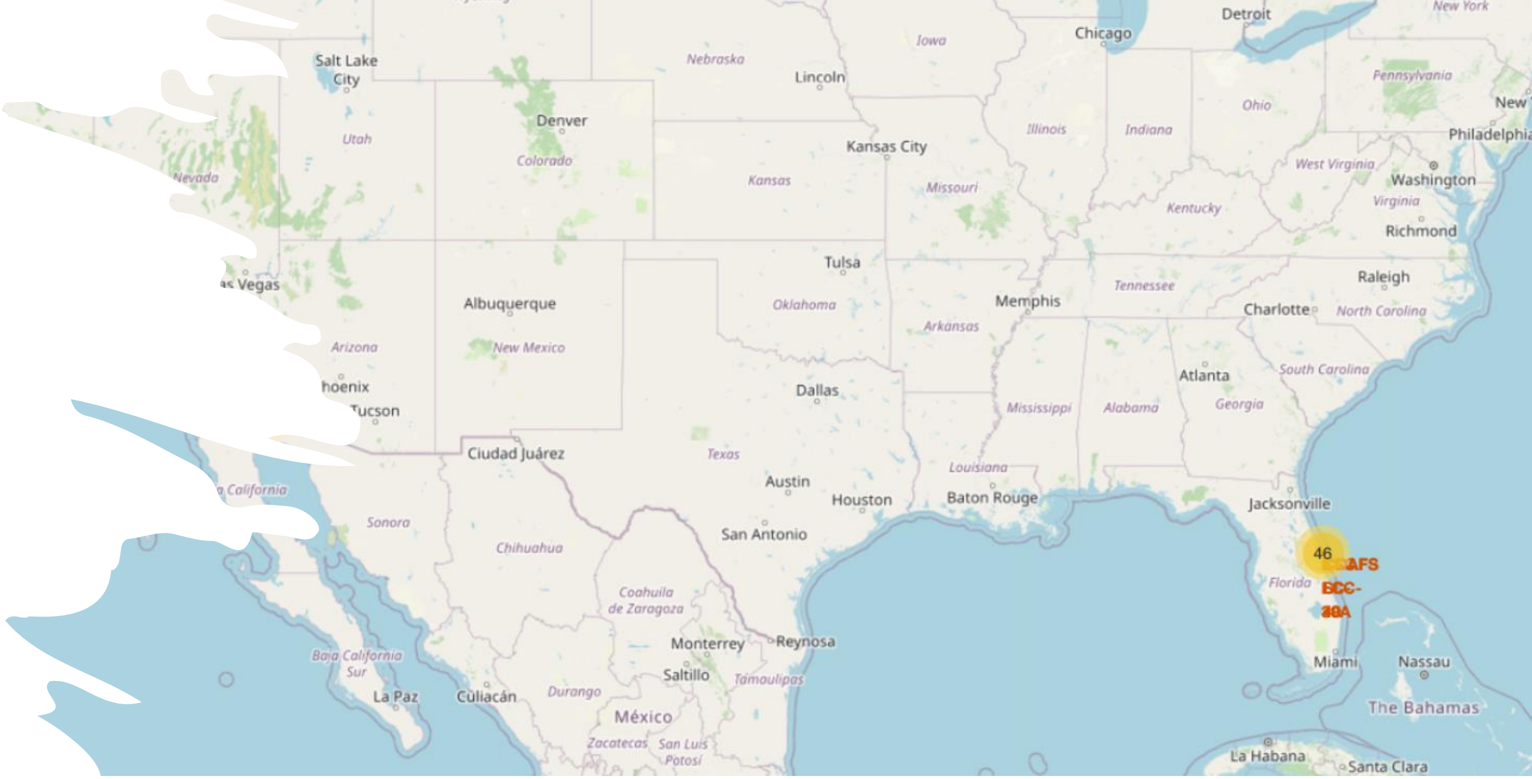
failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015

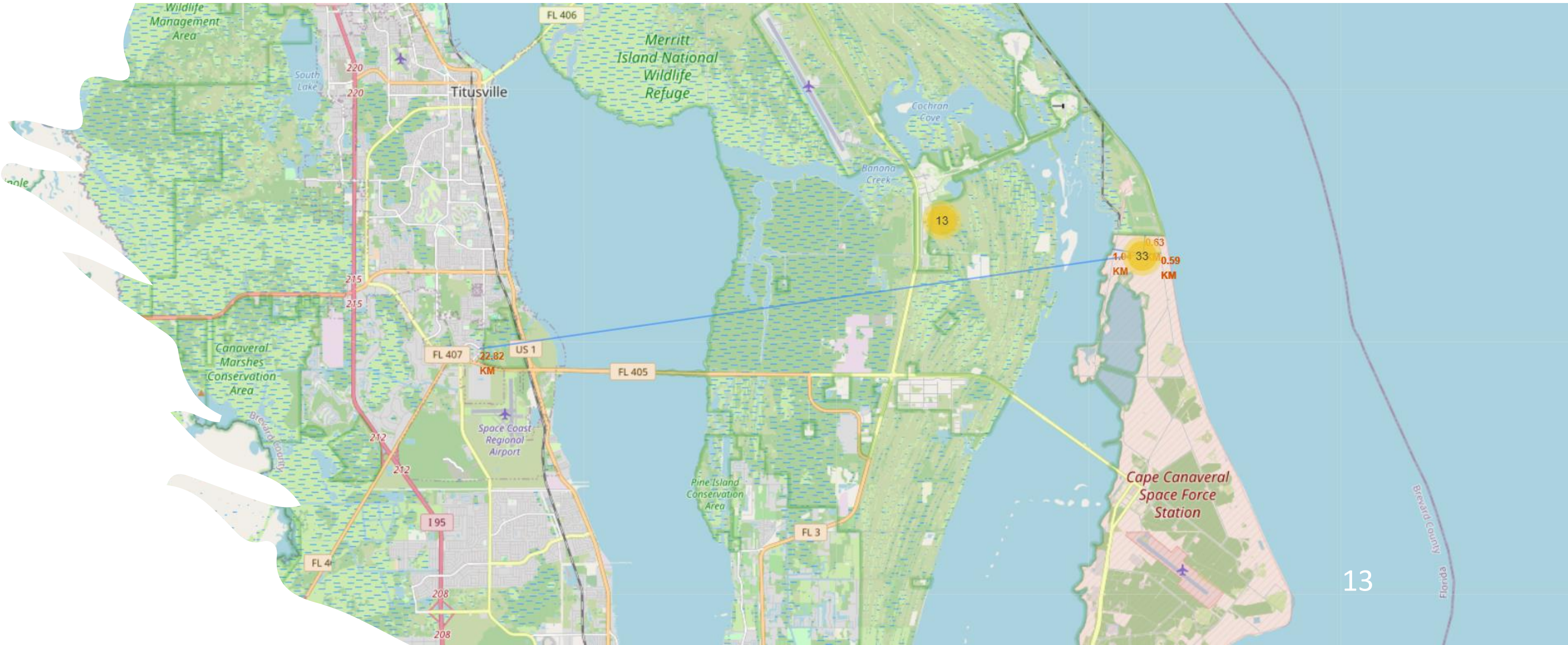
count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

https://github.com/TatianaMikl/testrepo/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

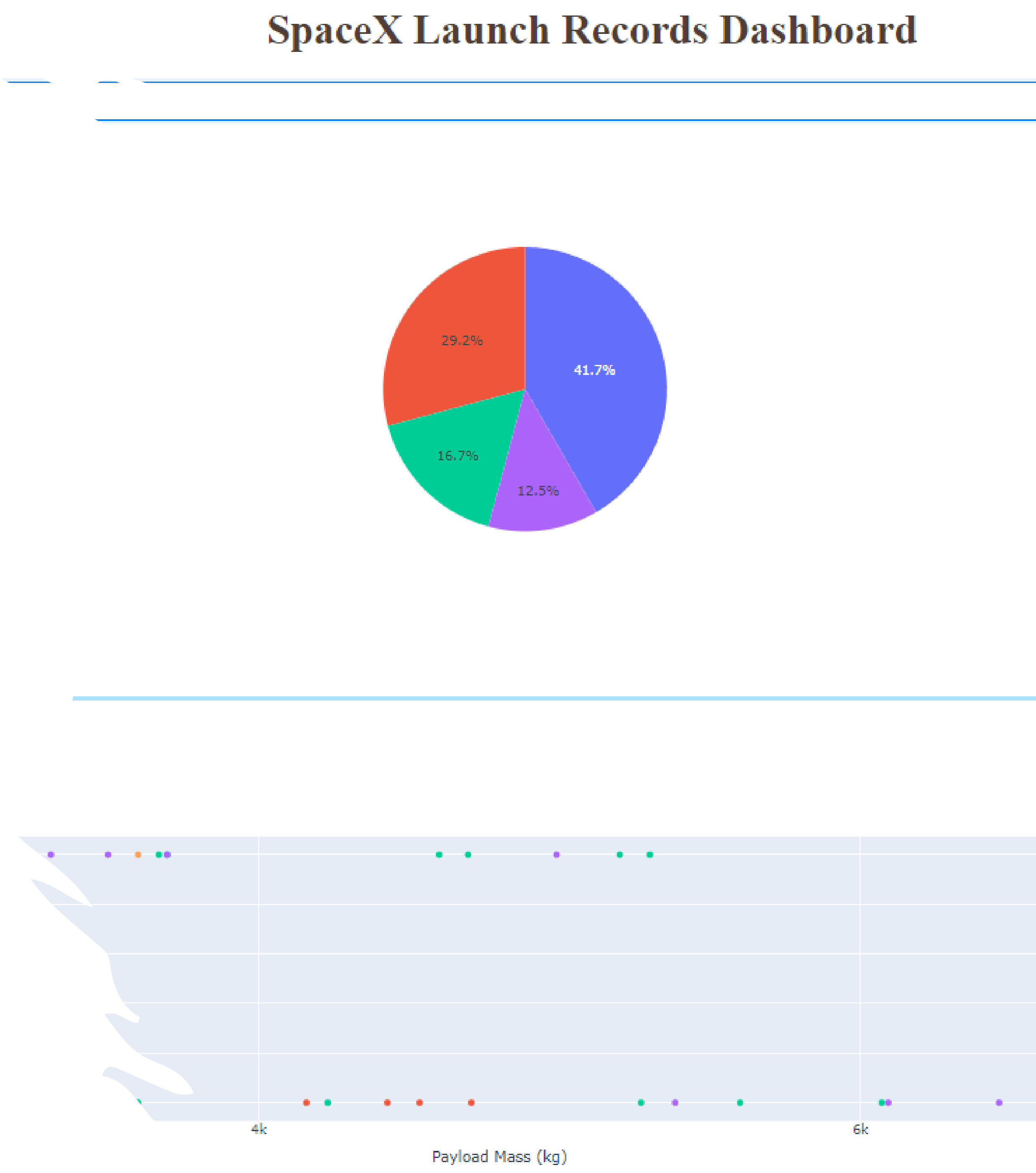
- We analyzed existing launch site locations to determine the factors that influence their successful location. For this we created and added to a folium map markers, circles and lines
- We marked all the cosmodromes on the map, successful/unsuccessful launches for each site and calculated the distances between the cosmodrome and its surroundings.
- https://github.com/TatianaMikl/testrepo/blob/main/lab_jupyter_launch_site_location.ipynb





Build a Dashboard with Plotly Dash

- The dashboard application contains input components such as a dropdown list and a range slider to interact with a pie chart and a scatter point chart.
- We use it to analyze SpaceX launch data to understand where the most successful launches occurred, which version of F9 Booster (v1.0, v1.1, FT, B4, B5, etc.) has the highest
- percentage of successful launches
- which payload range has the highest launch success rate and which payload range has the lowest launch success rate.
- https://github.com/TatianaMikl/testrepo/blob/main/spacex_dash_app.py



Predictive Analysis (Classification)

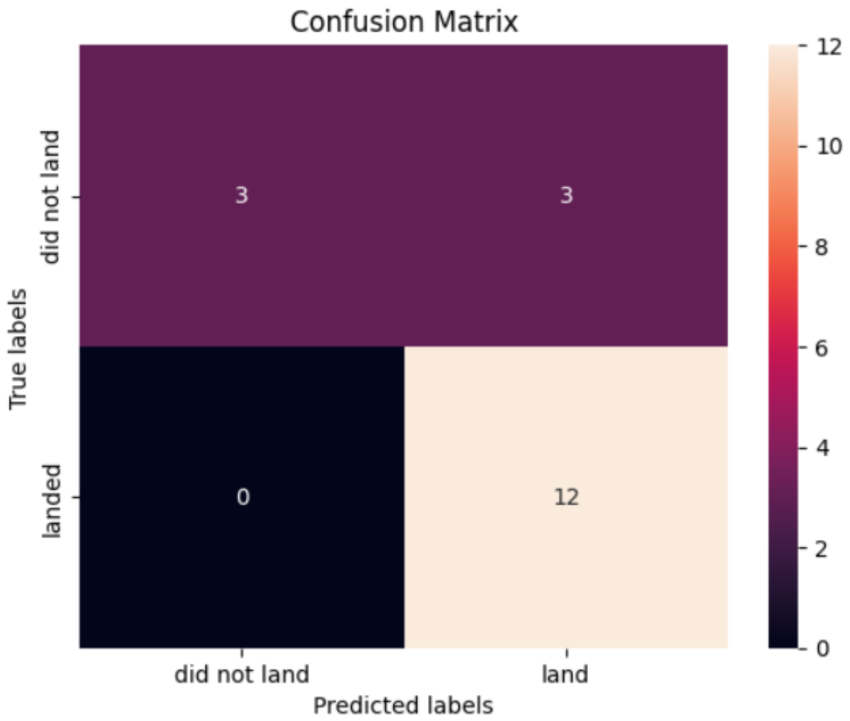
We built a machine learning pipeline to predict whether the first stage would complete, given data from previous labs.

We prepared the data, divided it into training and testing data. We tuned parameters for the following models: logistic regression, vector machine, tree classifier, k nearest neighbors. We calculated the accuracy of these models on test data. We constructed confusion matrices for each model.

https://github.com/TatianaMikl/testrepo/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

- All four models performed well on the test data set. If you use precision when training data, you can give preference to the method tree classifier
- We see the most successful launches on the KSC LC-40. But the percentage of successful launches is higher on the CCAFS SLS-40. The following payload ranges characteristic of successful launches can be distinguished: 362-475, 1952-3696, 4600-5300. And, accordingly, for unsuccessful ones 5384-6761. The F9 Booster FT version has the highest percentage of successful launches.



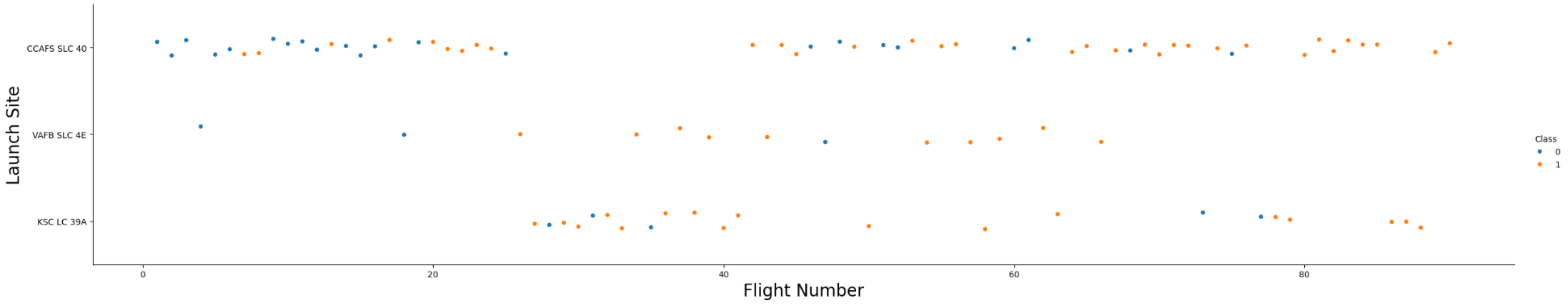
model	Train_accuracy	Test_accuracy
Logistic Regression	0.846	0.833
SVC	0.848	0.833
KNN	0.848	0.833
Tree	0.862	0.833

Section 2

Insights drawn from EDA

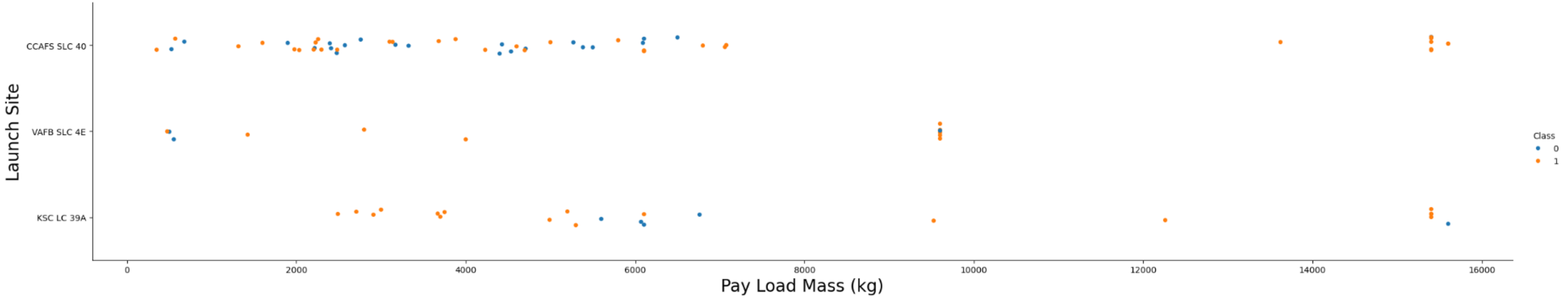
Flight Number vs. Launch Site

- We are seeing an overall increase in the frequency of successful launches
- The launch success rate improves significantly for the CCAFS SLC 40 from 63 flights
- The launch success rate improves significantly for the VAFB SLC 4E from 26 flights
- The launch success rate improves significantly for the KSC LC 39A from 36 flights



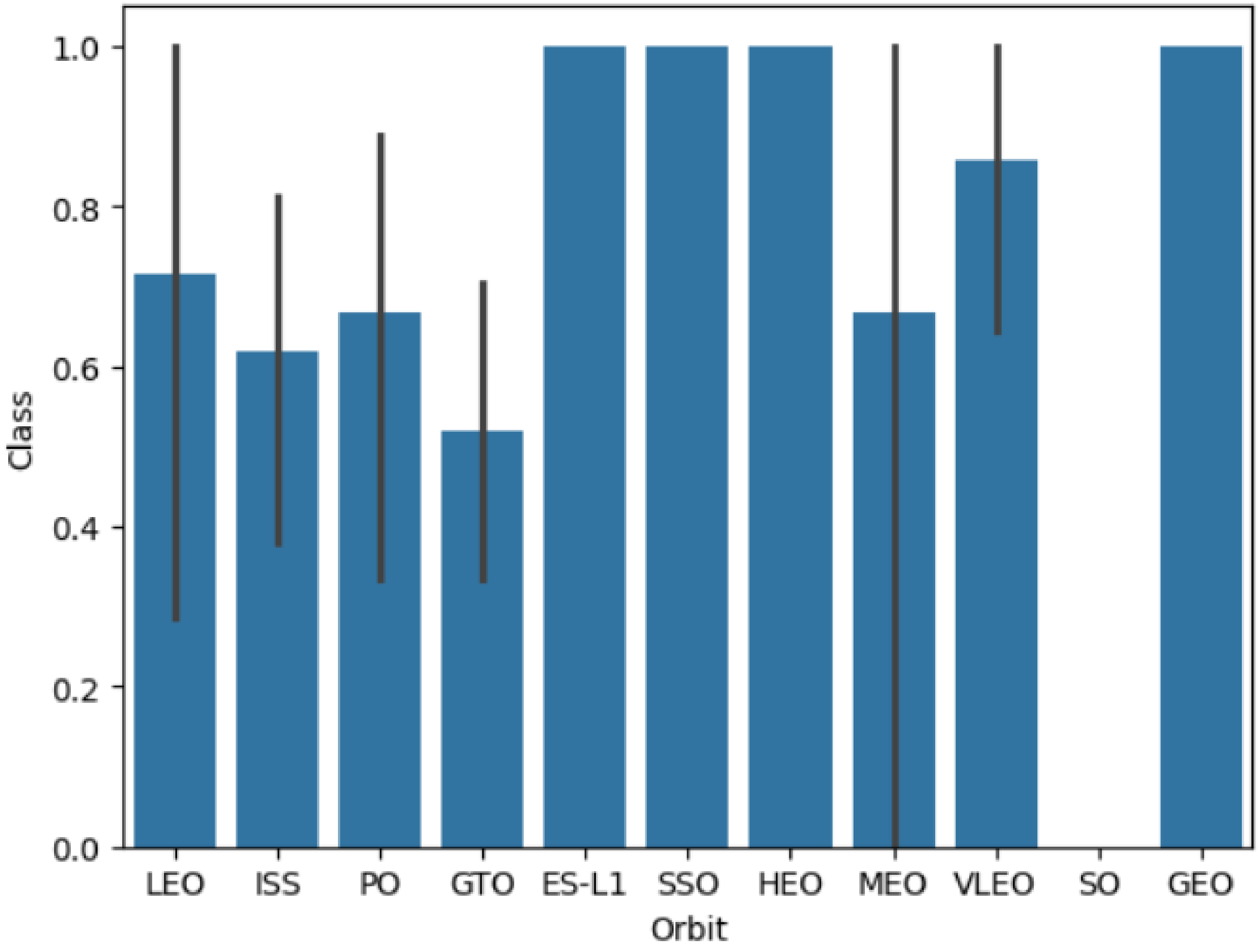
Payload vs. Launch Site

- For CCAFS SLC 40 and VAFB SLC 4E, high Payload values have a good effect
- For KSC LC 39A we observe a range of unfavorable Payload values 6000



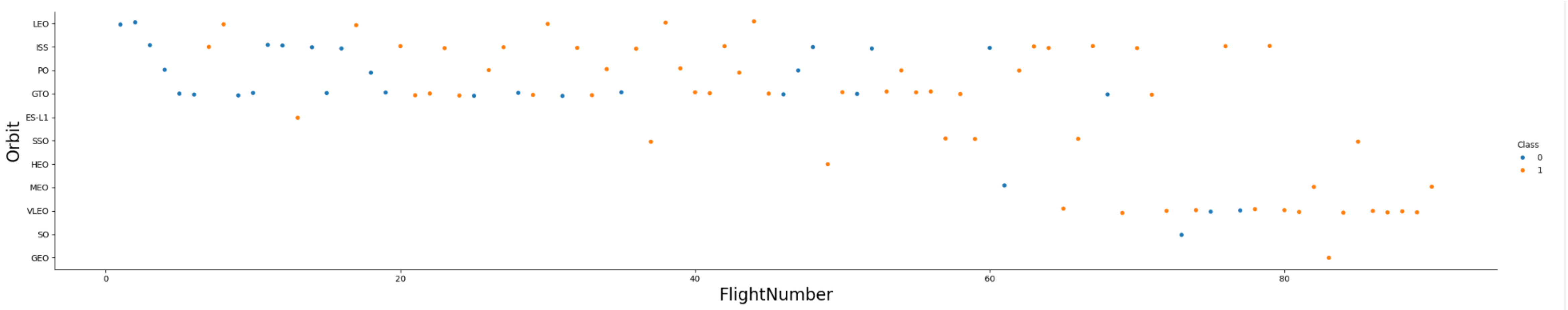
Success Rate vs. Orbit Type

- Orbits with complete success:
- ES-L1
- SSO
- HEO
- GEO



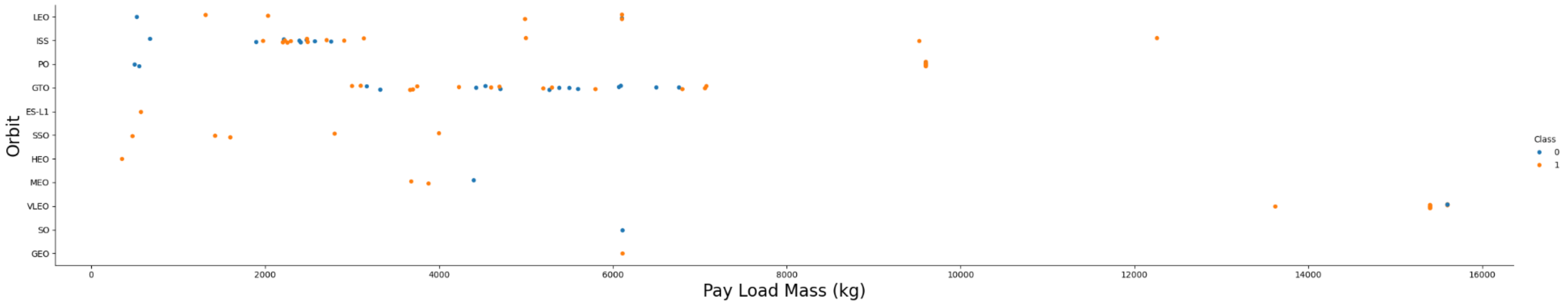
Flight Number vs. Orbit Type

- LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.



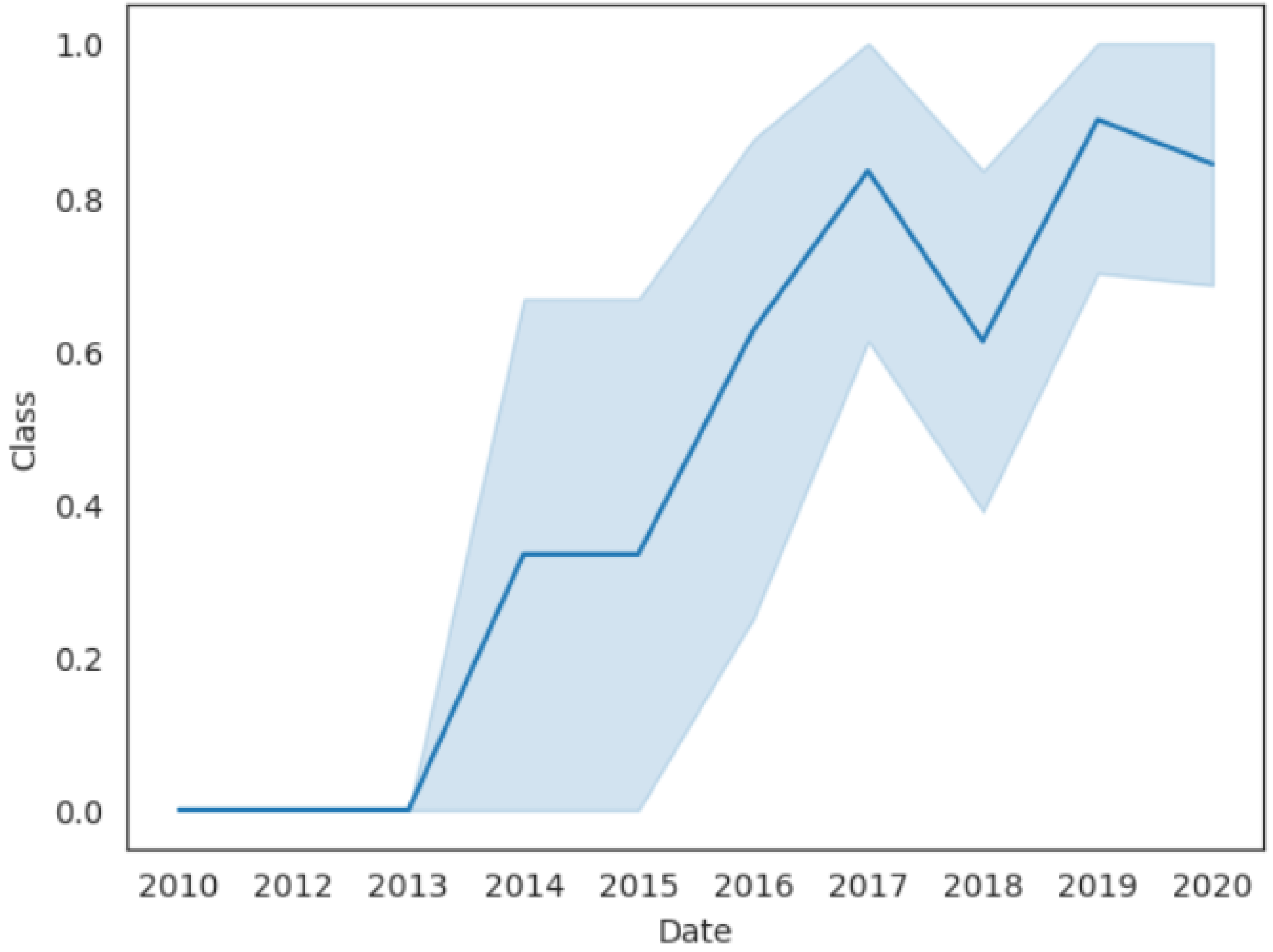
Payload vs. Orbit Type

- For heavy payloads, successful landing or positive landing speed is more suitable for Polar, LEO and ISS
- For GTO, it is difficult to distinguish between a successful landing and an unsuccessful one, since both outcomes are present.



Launch Success Yearly Trend

Sucess rate since 2013 kept increasing till 2020



All Launch Site Names

- Here are the coordinates of launch sites

Launch Site	Lat	Long
CCAFS LC-40	28.562302	-80.577356
CCAFS SLC-40	28.563197	-80.576820
KSC LC-39A	28.573255	-80.646895
VAFB SLC-4E	34.632834	-120.610745

Launch Site Names Begin with 'CCA'

5 records where launch sites begin with `CCA`

```
%sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" like 'CCA%' limit 5
```

```
 * 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

Total Payload Mass

- The total payload carried by boosters from NASA

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

```
%sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "Customer" like 'NASA%'
```

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

SUM("PAYLOAD_MASS__KG_")
99980

Average Payload Mass by F9 v1.1

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

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

```
%sql SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "Booster_Version" like
```

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

AVG("PAYLOAD_MASS__KG_")
2534.6666666666665

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad

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

Hint:Use min function

```
%sql SELECT MIN("Date") FROM SPACEXTABLE
```

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

MIN("Date")
2010-06-04

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- Present your query result with a short explanation here

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 "Booster_Version" FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" IN (4000, 6000) AND "Mission_Outcome" like "Success"
```

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

Booster_Version
F9 B5 B1046.3

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

List the total number of successful and failure mission outcomes

```
%sql SELECT "Mission_Outcome",COUNT("Mission_Outcome") FROM SPACEXTABLE GROUP BY "Mission_Outcome"
```

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

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

Boosters Carried Maximum Payload

List the names of the booster which have carried the maximum payload mass

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

```
%sql SELECT "Booster_Version","PAYLOAD_MASS__KG_" FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_"=(SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTABLE)
```

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

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

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Present your query result with a short explanation here

```
%sql SELECT substr(Date, 6,2), "Landing_Outcome", "Booster_Version", "Launch_Site"  FROM SPACEXTABLE WHERE substr(Date,0,5)='2015'
```

```
 * 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
02	Controlled (ocean)	F9 v1.1 B1013	CCAFS LC-40
03	No attempt	F9 v1.1 B1014	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
04	No attempt	F9 v1.1 B1016	CCAFS LC-40
06	Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40

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

- 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
- Present your query result with a short explanation here

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 count("Landing_Outcome"), "Date" FROM SPACEXTABLE WHERE "Landing_Outcome" like 'Failure%' AND "Date" in ("2010-06-04","2017-03-20")
```

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

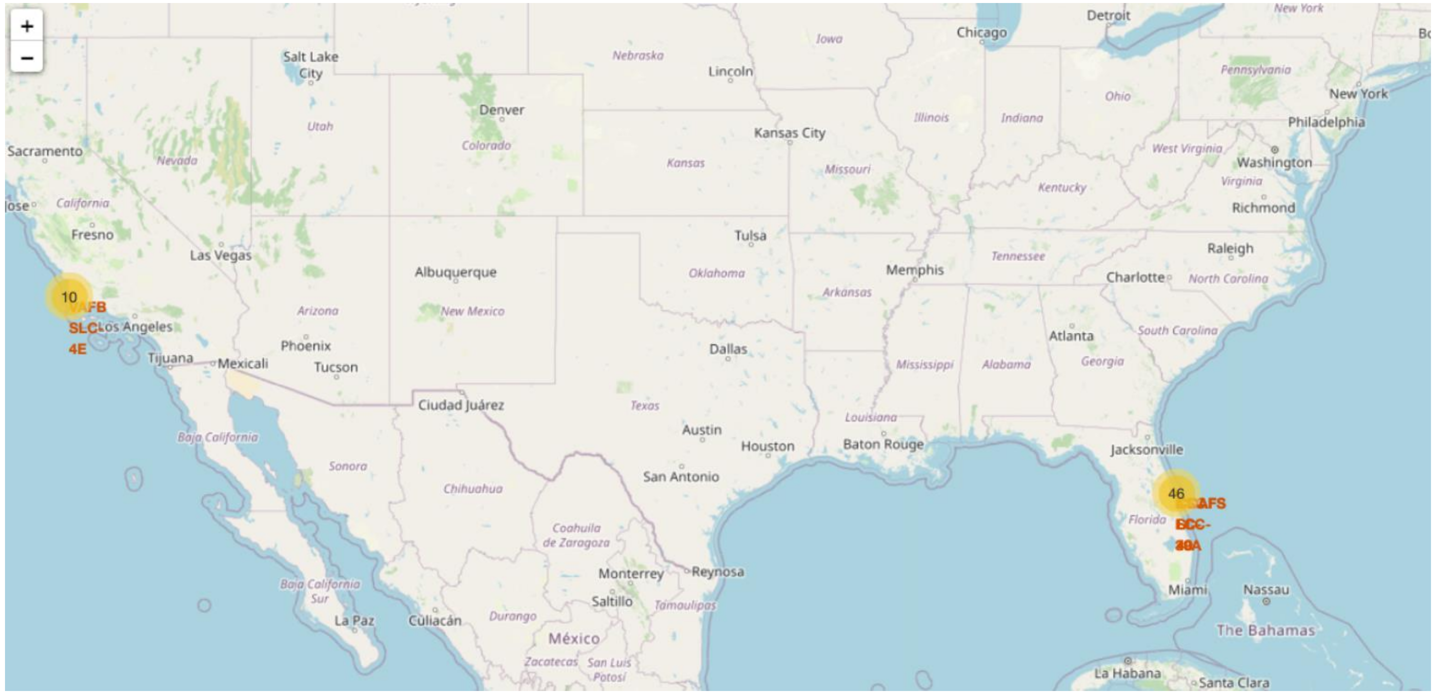
count("Landing_Outcome")	Date
1	2010-06-04

Section 3

Launch Sites Proximities Analysis

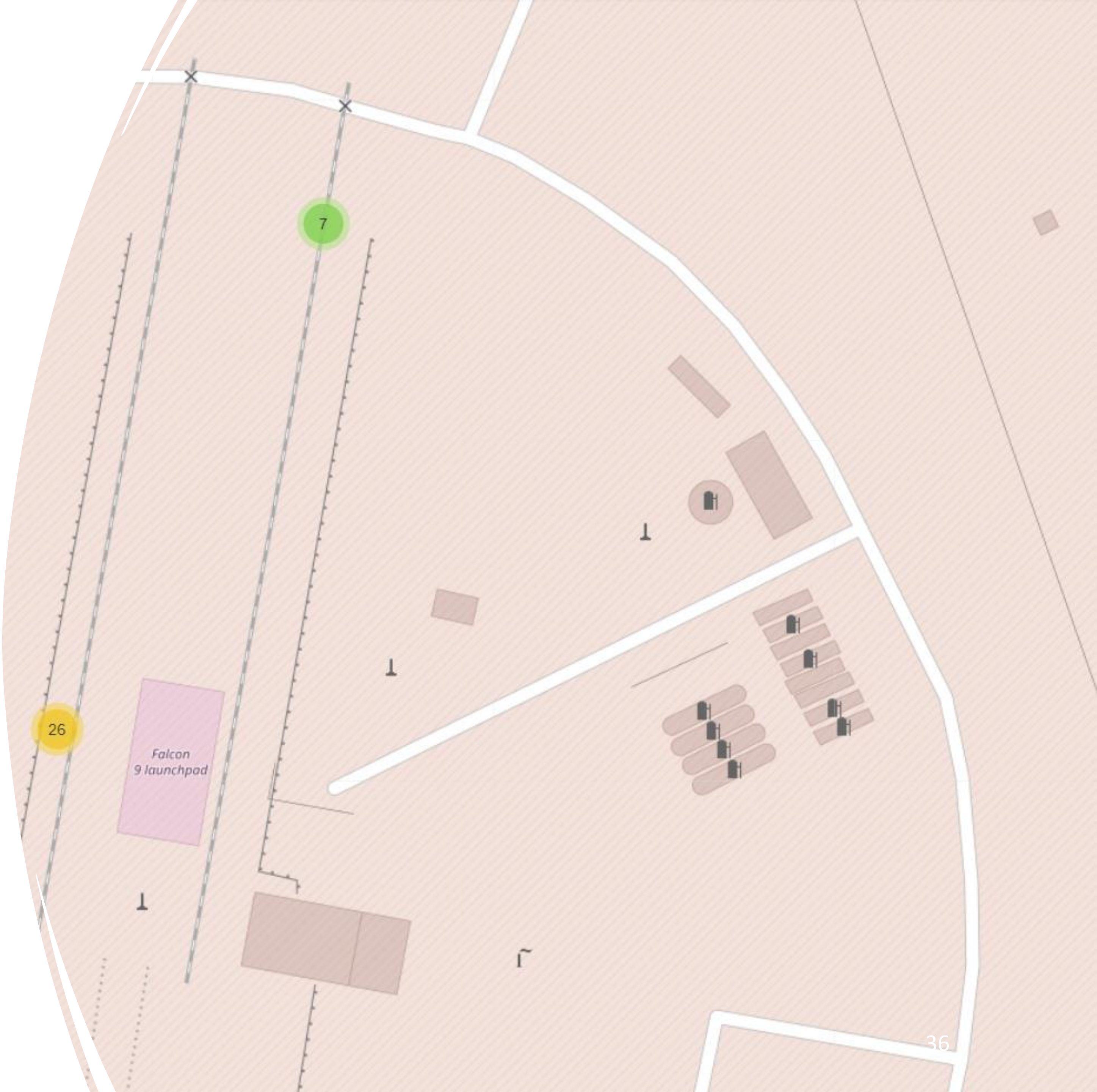
Launch sites on the map

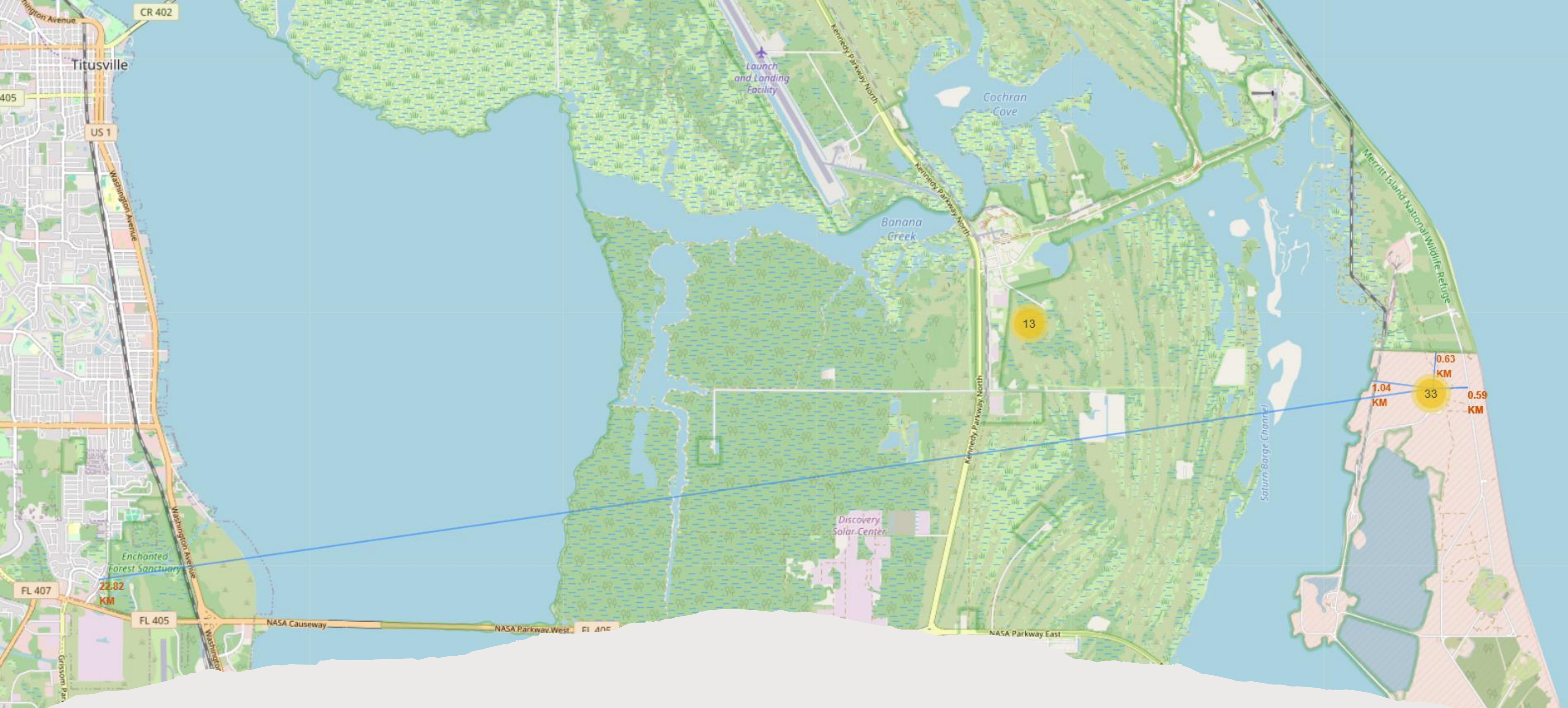
- On the map you can see a specific location of sites near the ocean coast.



Color-labeled markers in marker clusters

Color-labeled markers in marker clusters





Map with distance line

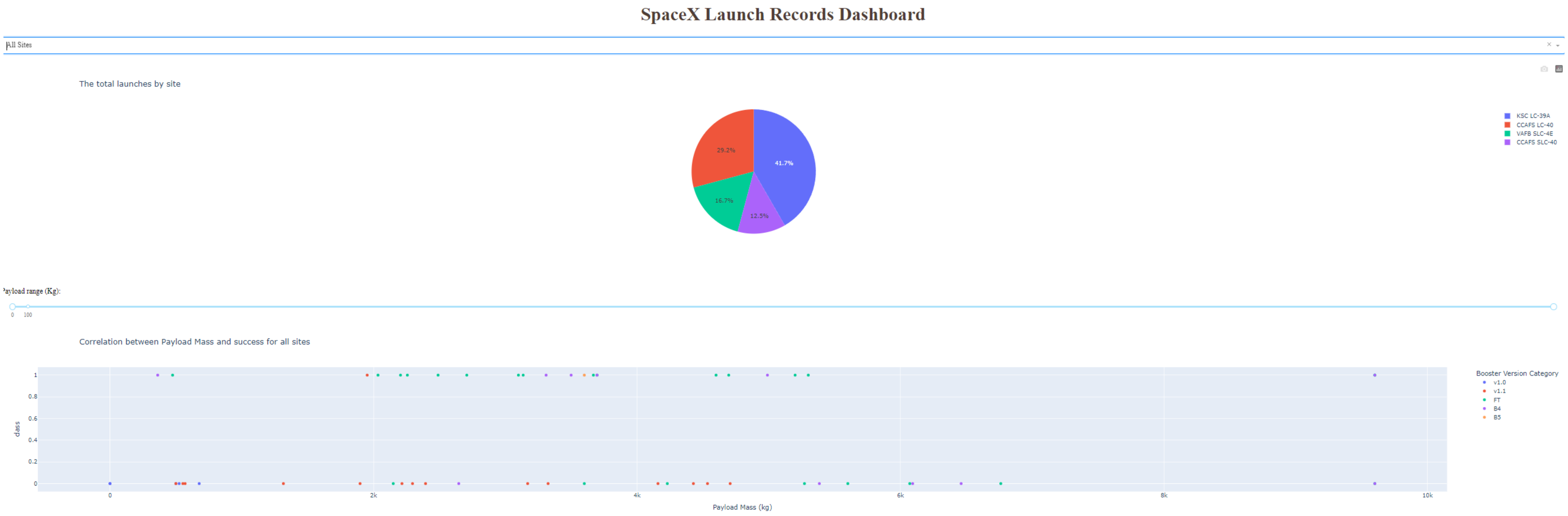
Lines between a launch site to its closest city, railway, highway, etc.

Section 4

Build a Dashboard with Plotly Dash

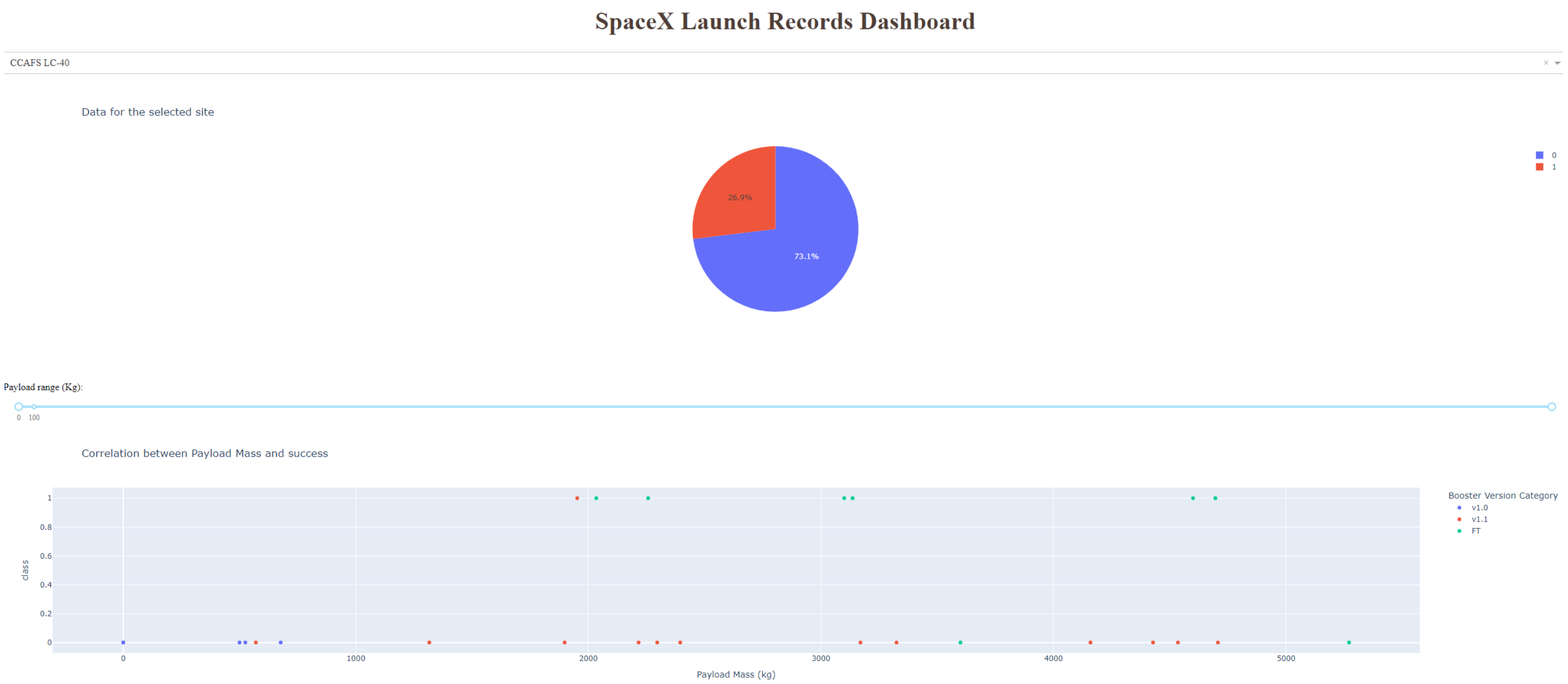
Dashboard

The KSC LC-40 site has the most successful launches.

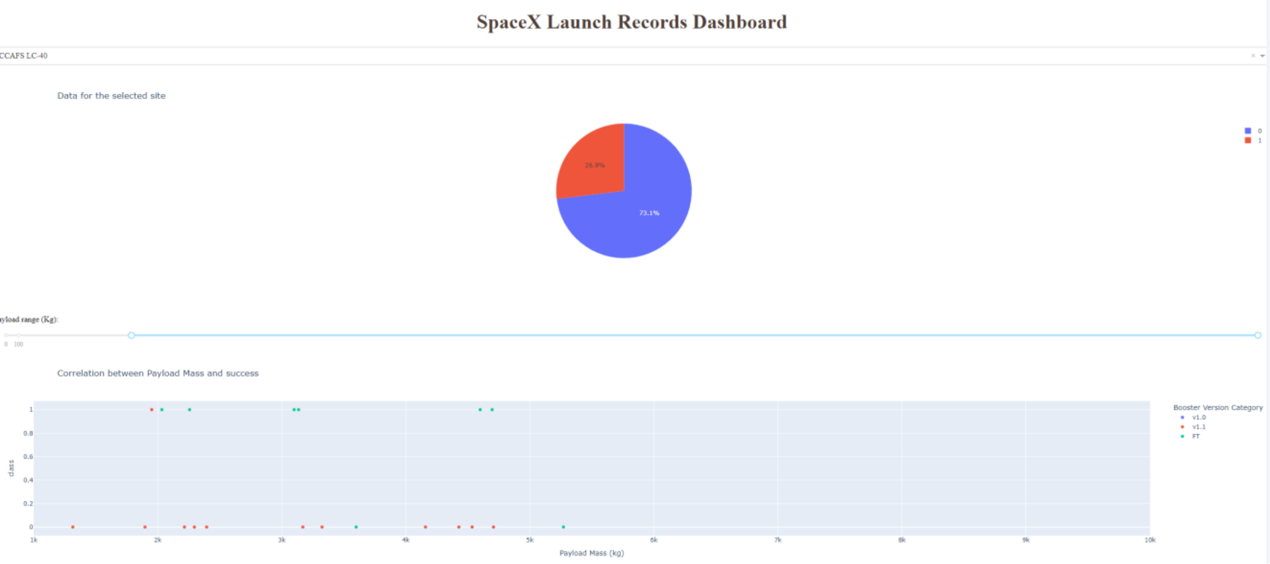


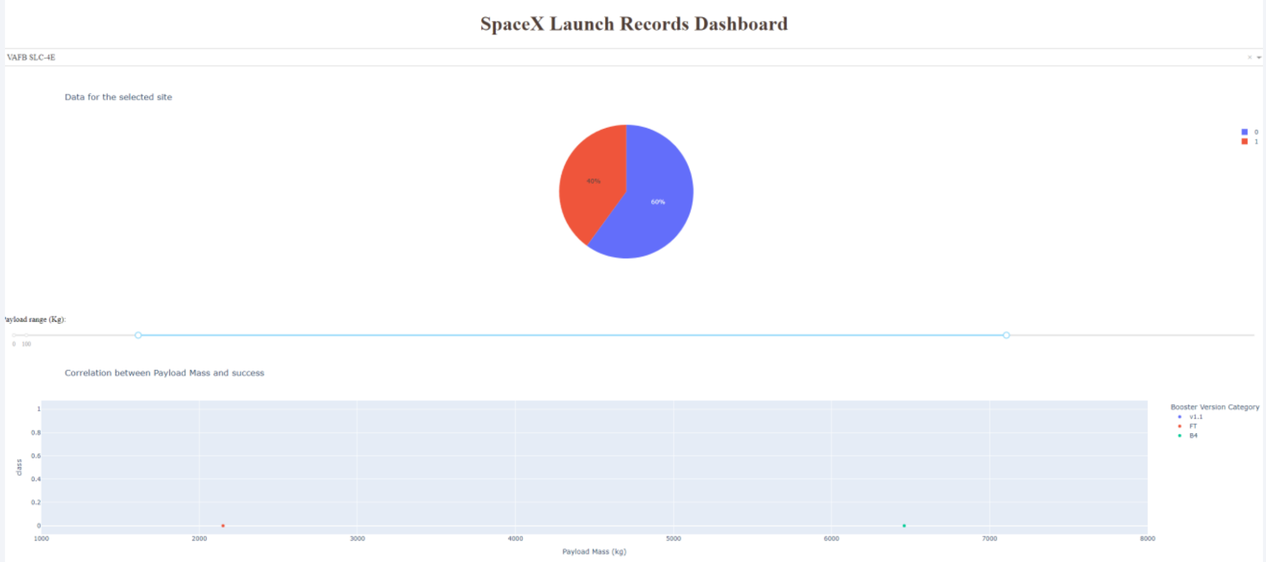
Dashboard

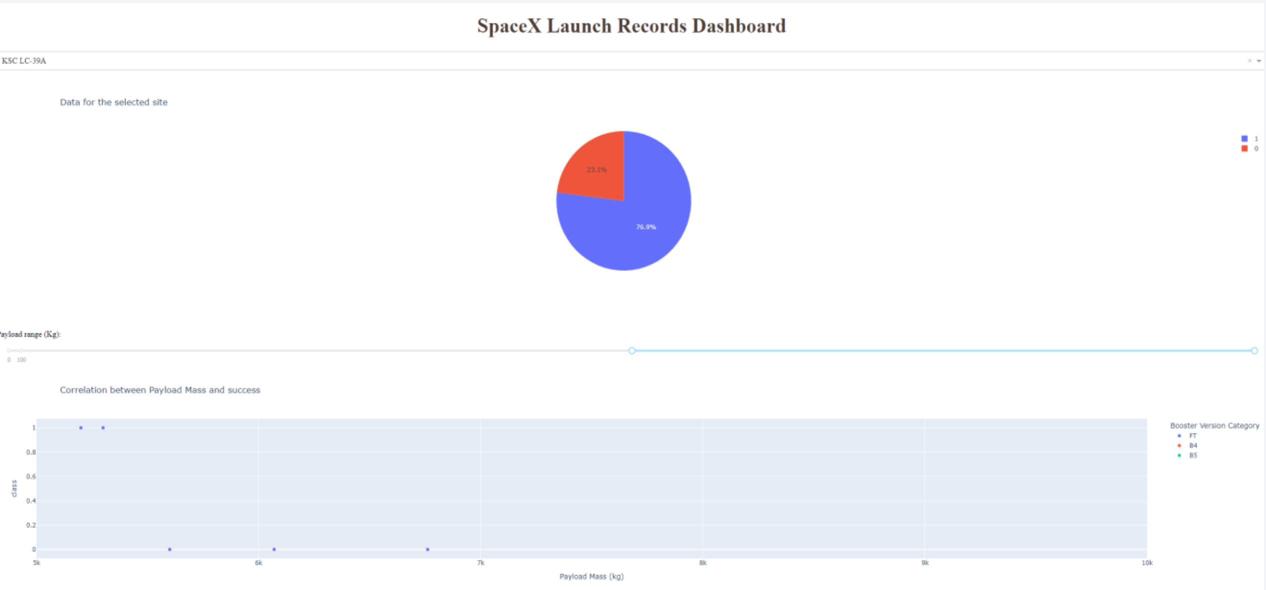
CCAFS SLS-40 is the launch site with highest launch success ratio

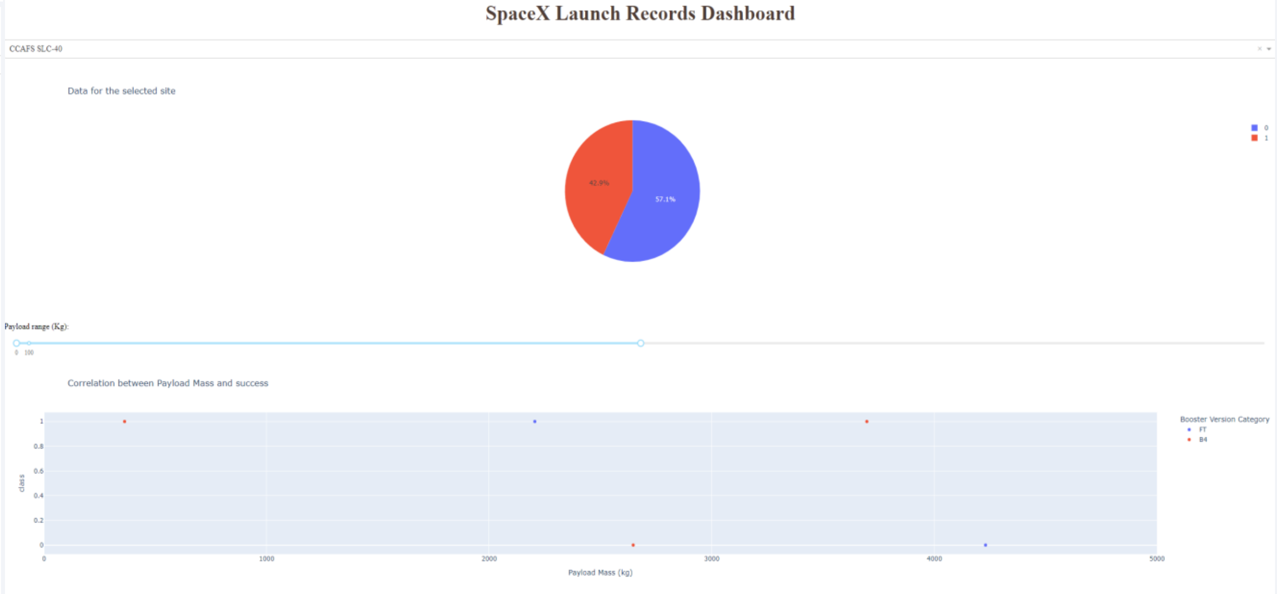


Dashboard for different sites







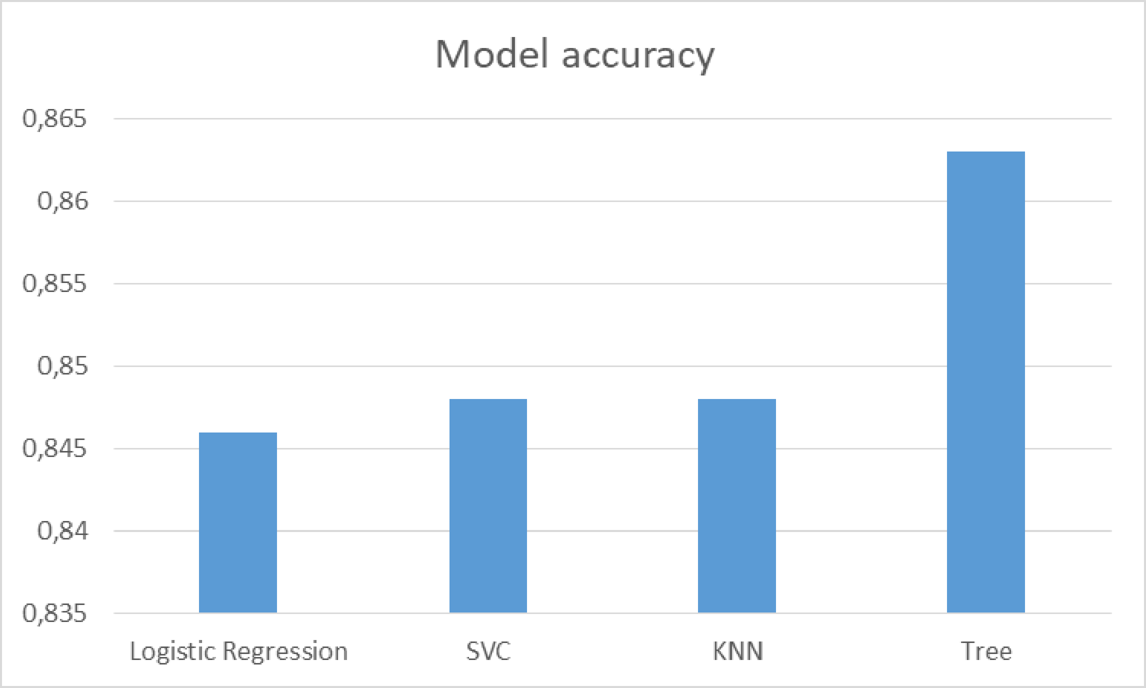


Section 5

Predictive Analysis (Classification)

Classification Accuracy

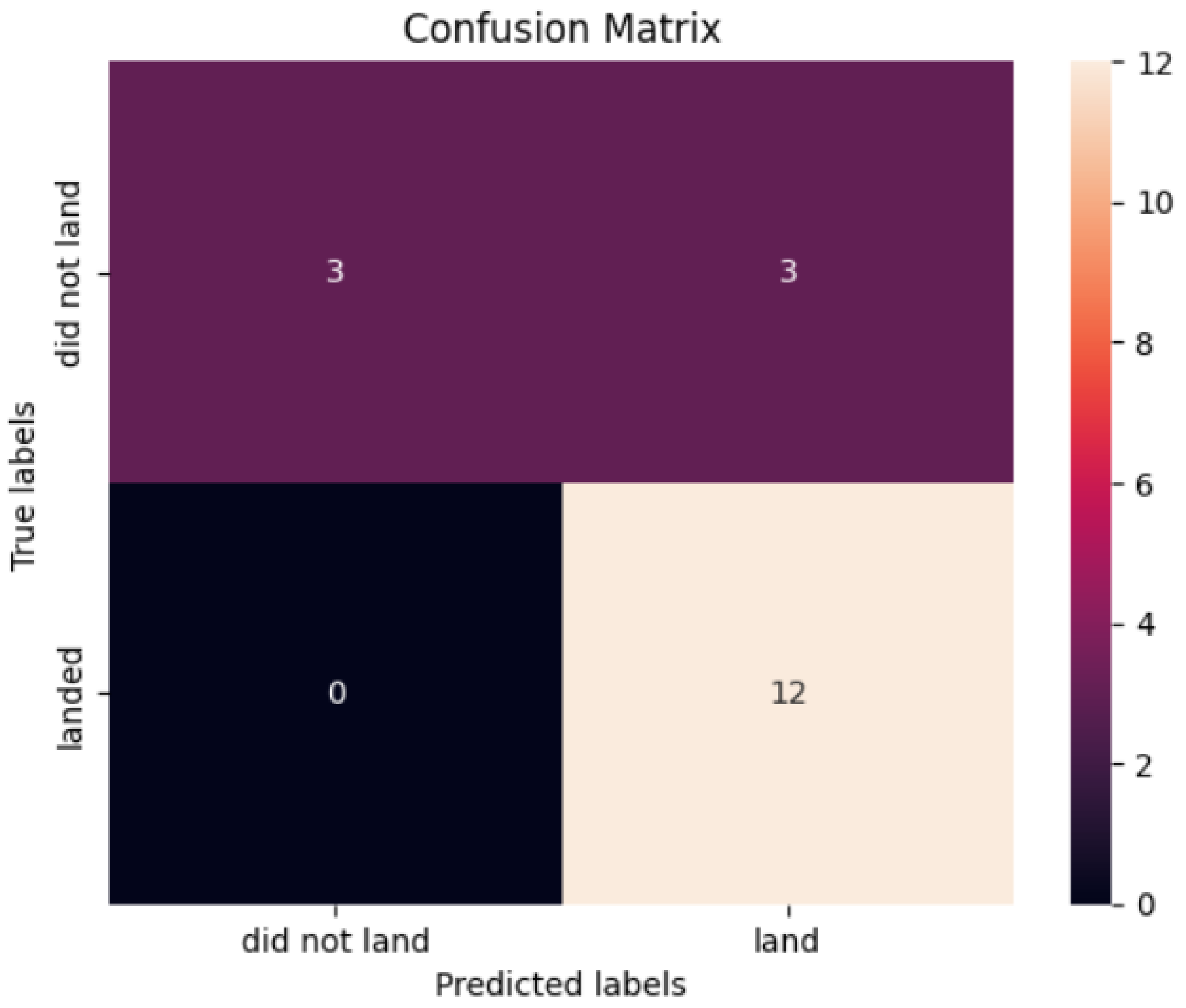
- All four models performed well on the test data set. If you use precision when training data, you can give preference to the method tree classifier



model	Train_accuracy	Test_accuracy
Logistic Regression	0.846	0.833
SVC	0.848	0.833
KNN	0.848	0.833
Tree	0.862	0.833

Confusion Matrix

- Examining the confusion matrix, we see that Tree classifer can distinguish between the different classes. We see that the problem is false positives.
- Overview:
- True Postive - 12 (True label is landed,
- Predicted label is also landed)
- False Postive - 3 (True label is not landed,
- Predicted label is landed)



Conclusions

- As a result of the analysis, we can conclude that the number of successful launches increases over time, which means the possibility of reusing the first stage, and, accordingly, reducing the cost of space launches in the future.
- Launch success may depend on launch site, orbit, payload and many other factors
- Different mathematical models were applied to the data used, and all showed high accuracy

Appendix

- https://github.com/TatianaMikl/testrepo
- All works are posted here

Thank you!