IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Yueren Xing
9/28/2023

Outline

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

Executive Summary

- Summary of methodologies

- Collecting Data -> Data Wrangling -> Explorative Data Analysis -> Interactive Data Visualization -> Predictive Analysis by Machine Learning

- Summary of all results

- Launch site KSC LC-39A has the most successful launches and CCAFS LC-40 has the highest launch success rate

- All 4 launch sites are in close proximity to railroads, highways, airports, and coastline while keeping a distance of at least 10 kilometers from the nearest city

- Payload weighing 2k-4k kg has the highest launch success rate and payload weighing over 4k kg has the lowest launch success rate

- Rocket models Falcon 9 FT and B5 have the highest launch success rate

- Machine learning models reached the highest accuracy of 83.33% on predicting if the first stage successfully lands

Introduction

- **Background**

SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, information on whether the first stages successfully land can be used to determine the cost of a launch if an alternate company wants to bid against SpaceX for a rocket launch.

- **Problems you want to find answers**

Given the data about the technical details of a rocket launch, i.e. variables such as orbit type, payload mass, booster version, etc., can one predict whether the first-stage booster will successfully land?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

  Request to the SpaceX API and Web scrape Falcon 9 launch records from Wikipedia

- Perform data wrangling

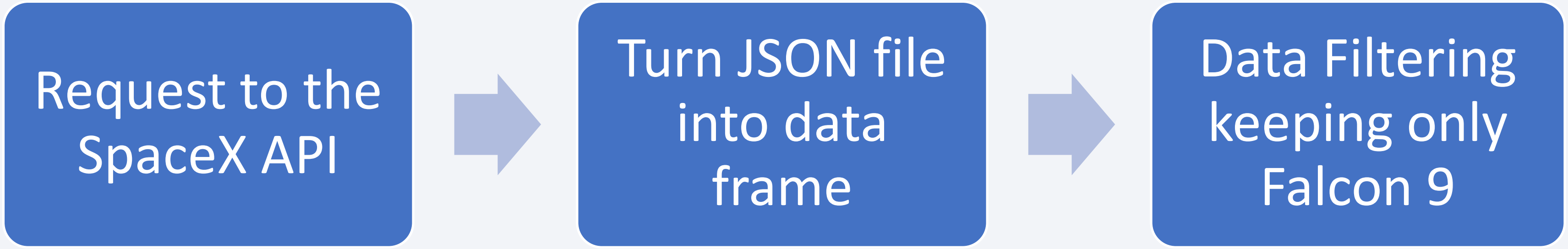
  Missing values are filled and HTML tables are parsed and turned into a data frame

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
- Build various machine learning models (logistics regression, decision tree, SVM, KNN) and then find the best hyperparameters by doing grid search over multiple groups of parameters

Data Collection

- Data are collected by two means:

1. Request to the SpaceX API

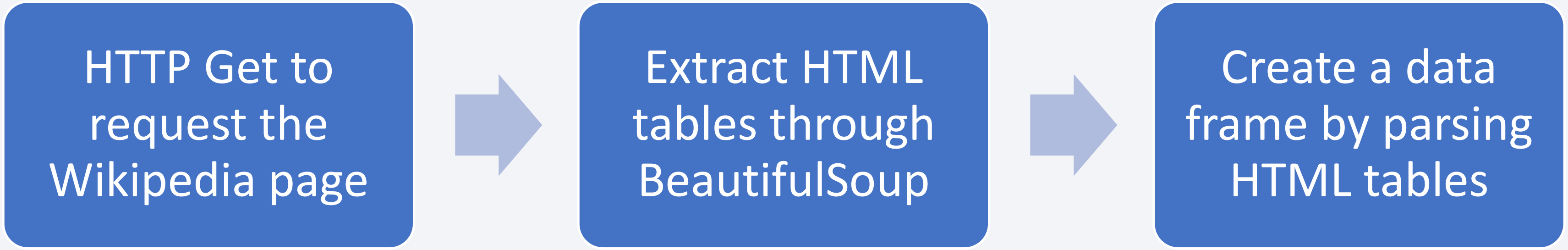


Request to the SpaceX API

Turn JSON file into data frame

Data Filtering keeping only Falcon 9

2. Web scrape Falcon 9 launch records from Wikipedia in the form of HTML tables



HTTP Get to request the Wikipedia page

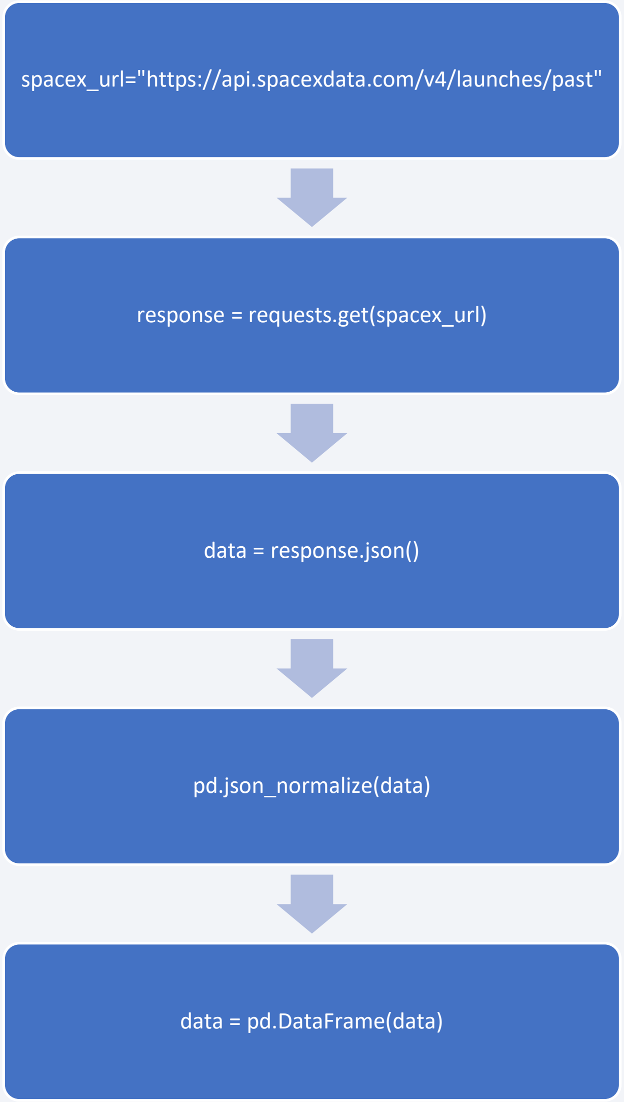
Extract HTML tables through BeautifulSoup

Create a data frame by parsing HTML tables

Data Collection – SpaceX API

- Make the call to the SpaceX API using REQUESTS, decode the response as a JSON and then turn it into a Pandas data frame

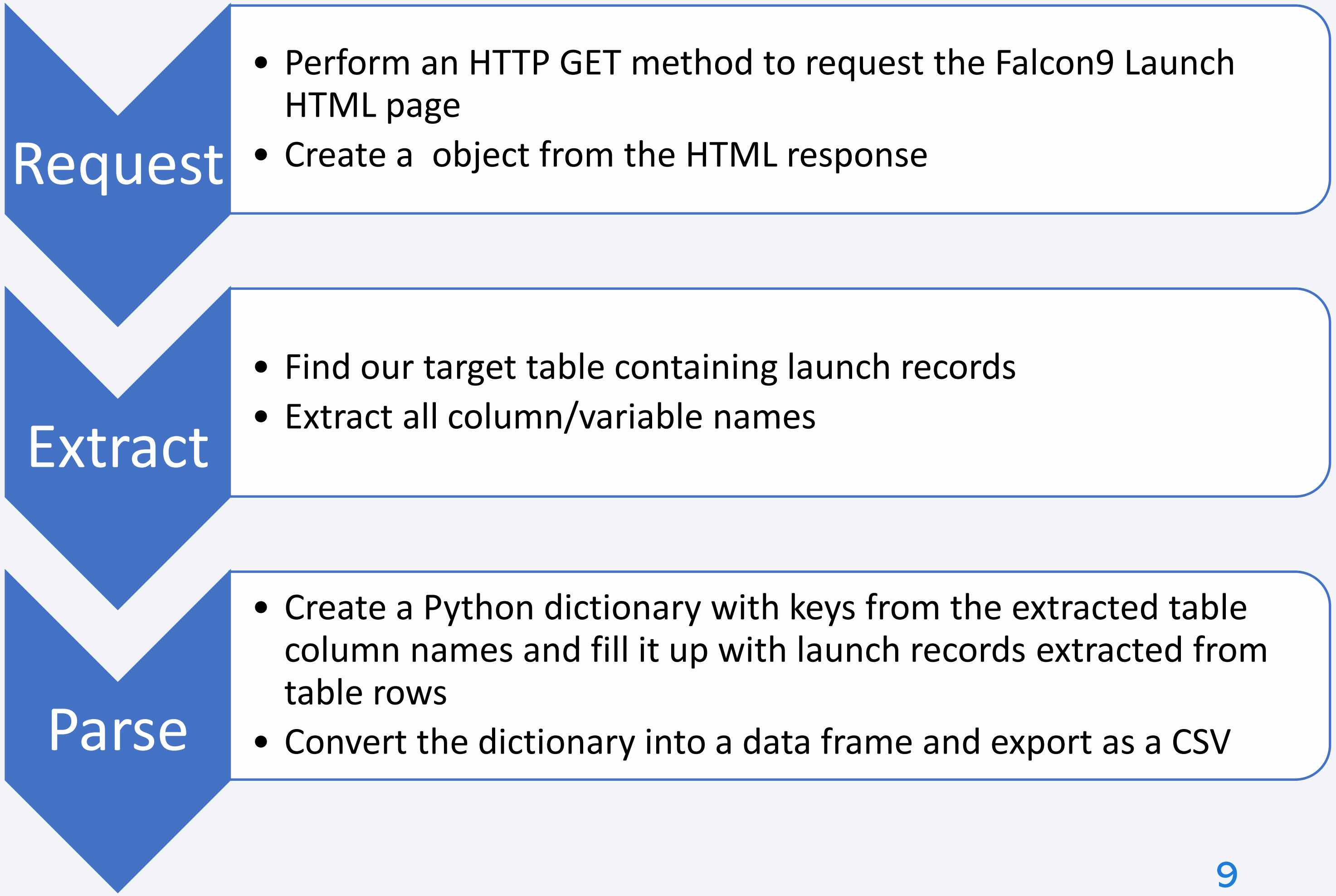
- https://github.com/YMaXing/IBM-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



```
spacex_url="https://api.spacexdata.com/v4/launches/past"
response = requests.get(spacex_url)
data = response.json()
pd.json_normalize(data)
data = pd.DataFrame(data)
```

Data Collection - Scraping

- https://github.com/YMaXing/IBM-Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb



Request

- Perform an HTTP GET method to request the Falcon9 Launch HTML page
- Create a object from the HTML response

Extract

- Find our target table containing launch records
- Extract all column/variable names

Parse

- Create a Python dictionary with keys from the extracted table column names and fill it up with launch records extracted from table rows
- Convert the dictionary into a data frame and export as a CSV

Data Wrangling

- The data frame is filtered to include only Falcon 9 launches
- The missing values in the column “PayloadMass” are replaced with mean payload mass.
- Further, all categorical features are then one-hot encoded.
- Finally, a landing outcome label is created from the “Outcome” column, where 0s correspond to landing failures and 1s correspond to successes.

- https://github.com/YMaXing/IBM-Data-Science-Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

Charts plotted:

1. “Flight Number vs. Payload Mass” scatter plot
2. “Flight Number vs. Launch Sites” scatter plot
3. “Payload Mass vs. Launch Sites” scatter plot
4. “Orbit Type vs. Success Rate” bar chart
5. “Flight Number Mass vs. Orbit Type ” scatter plot
6. “Payload Mass vs. Orbit Type” scatter plot
7. “Year vs. Success Rate” line chart

The above charts are plotted to see how different variables affect the landing outcomes and the correlations among themselves.

https://github.com/YMaXing/IBM-Data-Science-Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

SQL queries performed

- select distinct "Launch_Site" from SPACEXTBL
- select * from SPACEXTBL where "Launch_Site" like 'CCA%' limit 5
- select sum("PAYLOAD_MASS__KG_") from SPACEXTBL where "Customer" = "NASA (CRS)“
- select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where "Booster_Version" = "F9 v1.1“
- select min(Date) from SPACEXTBL where Landing_Outcome = "Success (ground pad)“
- select Booster_Version from SPACEXTBL where Landing_Outcome = "Success (drone ship)" and (PAYLOAD_MASS__KG_ between 4000 and 6000)
- select count(*) from SPACEXTBL where lower(Mission_Outcome) like '%success%’ (‘%failure%’)
- select distinct Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
- select substr(Date, 6, 2) as Month, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTBL where substr(Date,1,4) ='2015' and Landing_Outcome = "Failure (drone ship)“
- select Landing_Outcome, count(Landing_Outcome) from SPACEXTBL where Date between '2010-06-04' and '2017-03-20' group by Landing_Outcome order by count(Landing_Outcome) desc

https://github.com/YMaXing/IBM-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- A circle and a marker for each launch site on the site map

  - show the locations of launch sites on the map

- A marker for each launch record

  - identify which sites have relatively high success rates

- Distance lines between each launch site to its proximities

  - see if each launch site is in close proximity to railroads, highways, coastline, etc., and if each launch site keeps a certain distance away from the nearest cities

http://localhost:8888/lab/tree/IBM%20Data%20Science/Untitled%20Folder/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- **A dropdown menu to select different launch sites**
- **Pie charts for all sites together and each site via a callback function based on the dropdown menu**

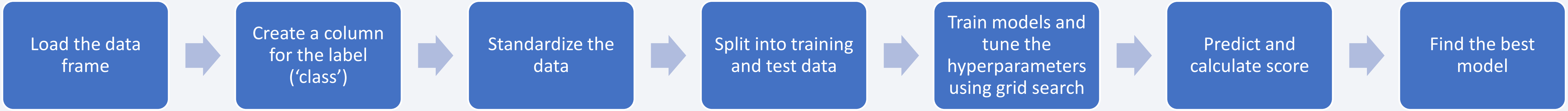
- To see which launch site has the largest success count, then select one specific site and to check its detailed success rate

- **A range slider to select payload**
- **A scatter plot of payload vs. class(landing outcome) via a callback function based on both the dropdown menu and the range slider**

- To easily select different payload ranges to see how the payload may be correlated to landing outcomes for selected sites.

https://github.com/YMaXing/IBM-Data-Science-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)



Load the data frame

Create a column for the label (‘class’)

Standardize the data

Split into training and test data

Train models and tune the hyperparameters using grid search

Predict and calculate score

Find the best model

Models used in the training:

- Logistic Regression
- Support Vector Machine (SVM)
- Decision Tree
- K Nearest Neighbors (KNN)

Scores used in the test:

- Accuracy
- Confusion Matrix

http://localhost:8888/lab/tree/IBM%20Data%20Science/Untitled%20Folder/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

- Exploratory data analysis results

As the flight number increases, the first stage is more likely to land successfully.

The more massive the payload, the less likely the first stage will return.

Different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E have a success rate of 77%.

For the VAFB-SLC launch site, there are no rockets launched for heavy payload mass(greater than 10000)

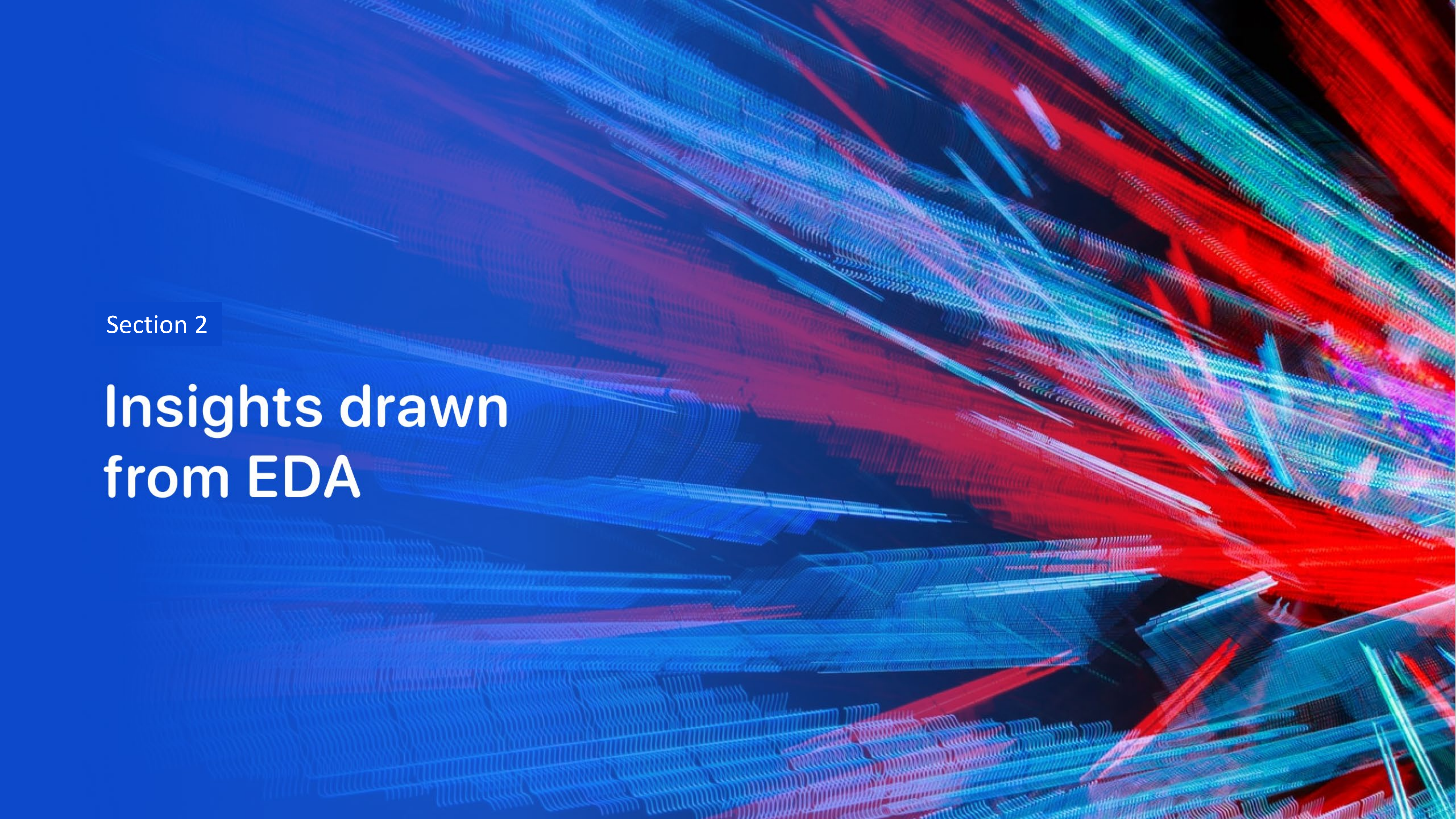
For some orbits, the success appears related to the number of flights; for others, there is no clear relationship.

With heavy payloads, the successful landing or positive landing rate is higher for Polar, VLEO, and ISS. However, for GTO we cannot distinguish this well as both positive landing rate and negative landing (unsuccessful mission) are both there here.

The success rate since 2013 kept increasing till 2020

- Interactive analytics demo in screenshots
- Predictive analysis results

All 4 models achieved the same performance on the test set with an accuracy of 83.33%

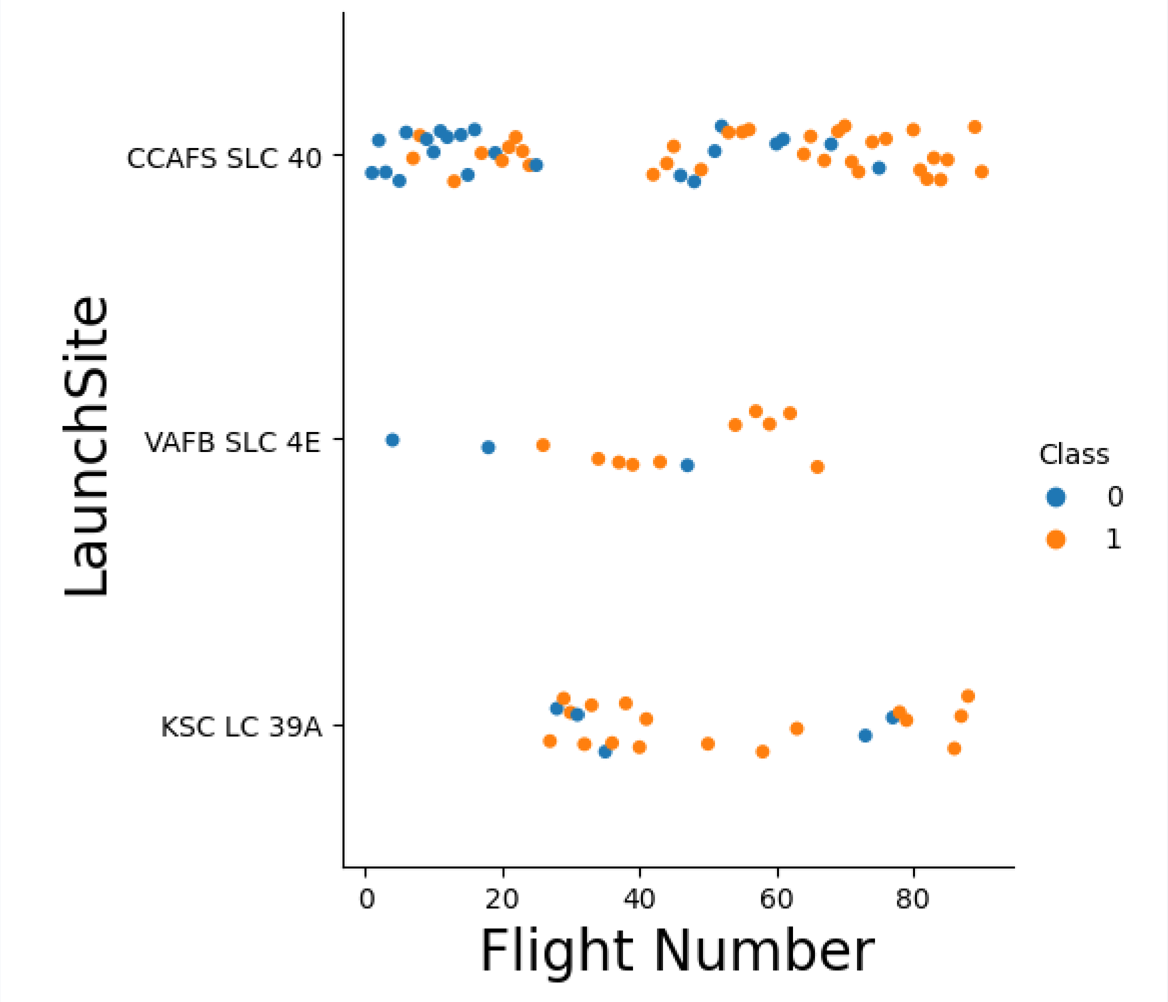


Section 2

Insights drawn from EDA

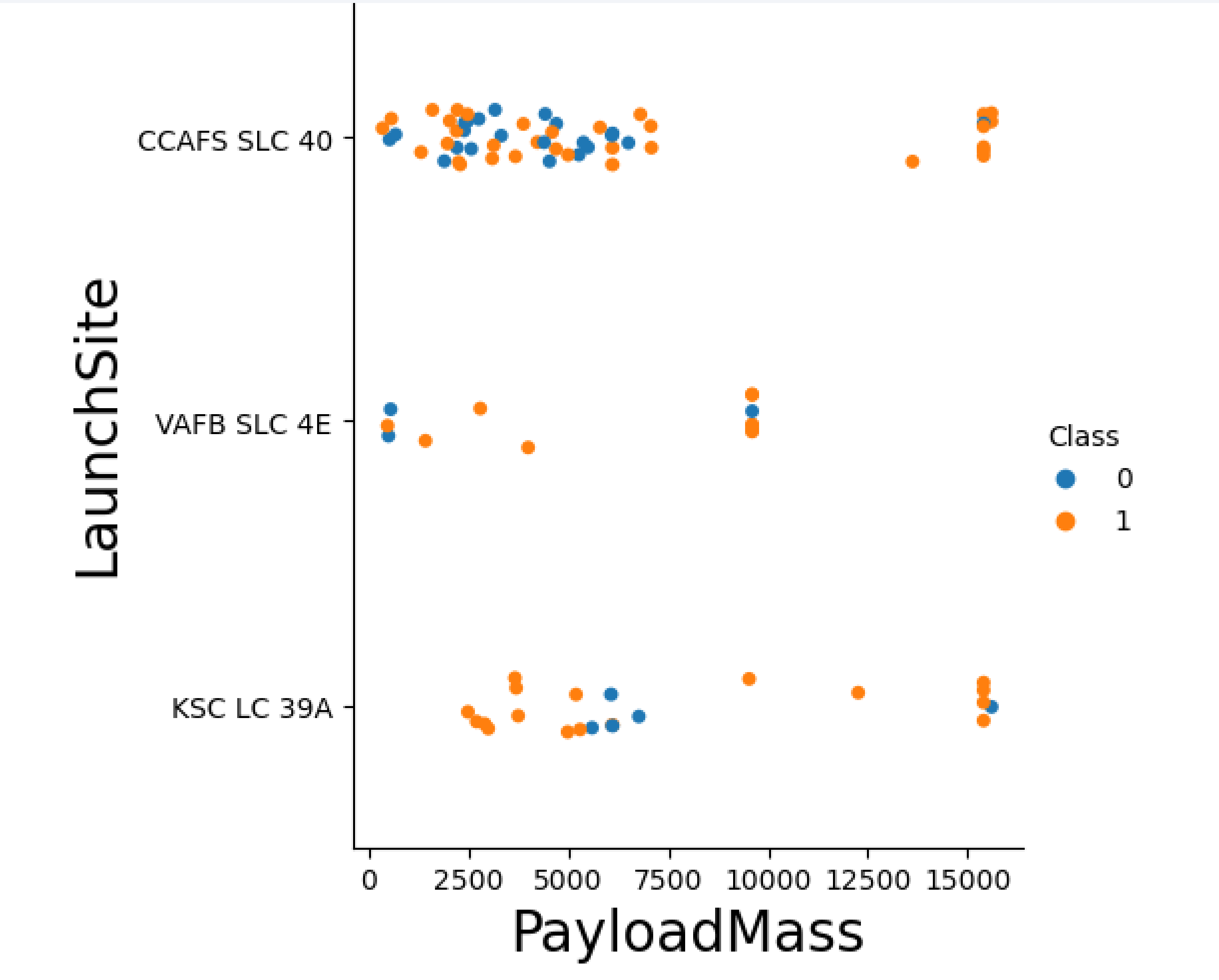
Flight Number vs. Launch Site

- Early launches (flight number < 25) were mostly at CCAFS SLC 40 with mixed outcomes
- CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%
- Success rates at all three sites increased significantly after flight number 50



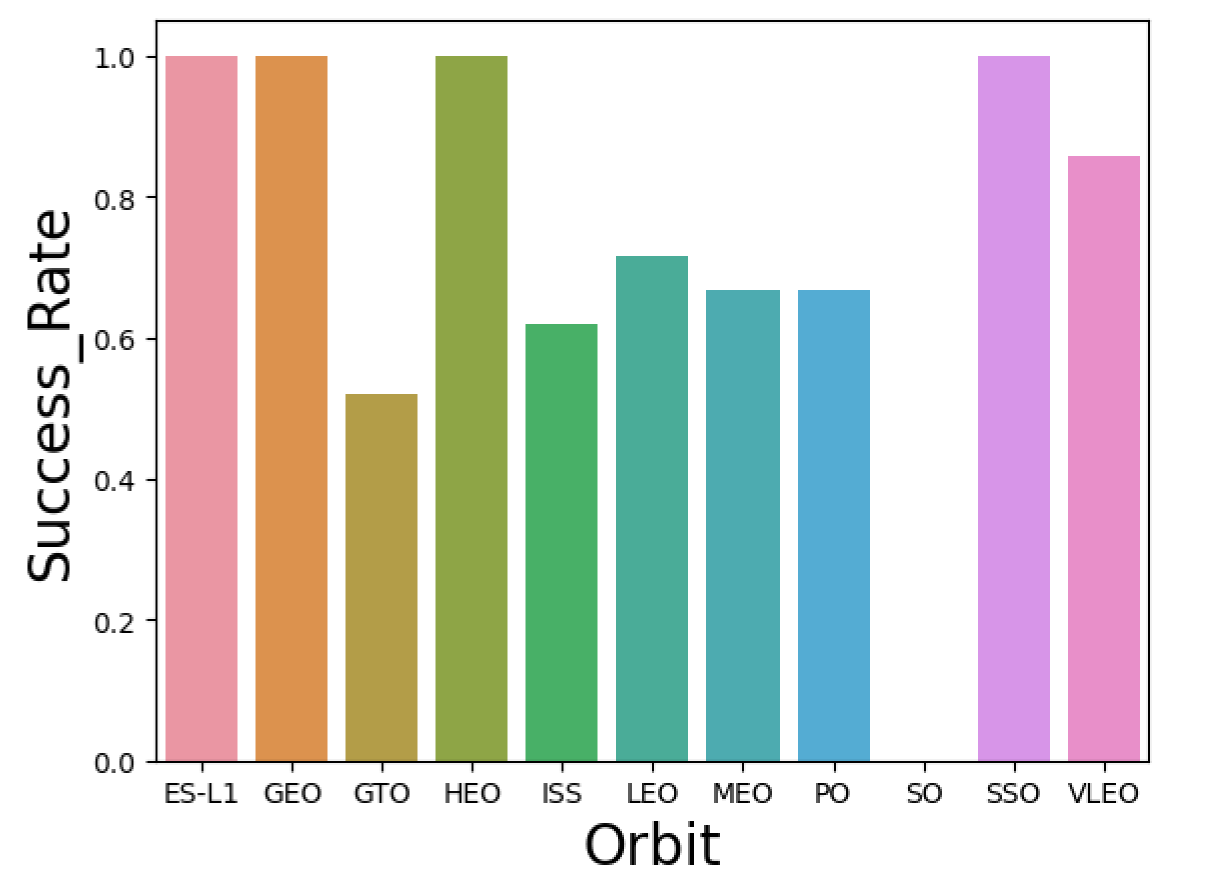
Payload vs. Launch Site

- For the VAFB-SLC 4E launch site there are no rockets launched for heavy payload mass(greater than 10000).
- A majority of rockets with lighter payload mass (< 7500kg) were launched from CCAFS SLC 40, some were launched from KSC LC 39A, few were launched from VAFB-SLC.



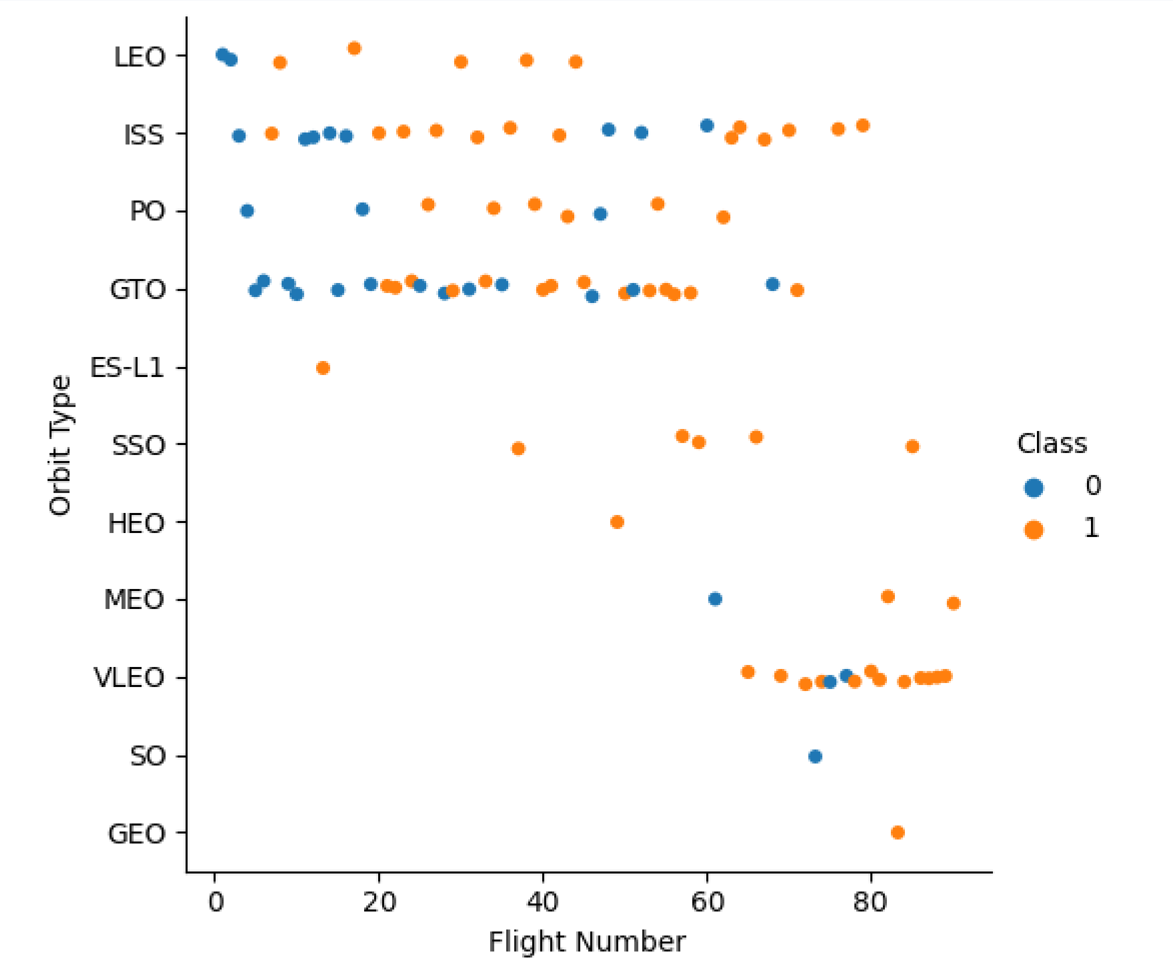
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO are the four orbits with the highest success rate (100%)
- Other orbits each have a success rate between 50% - 85% except for orbit SO where the success rate is 0%



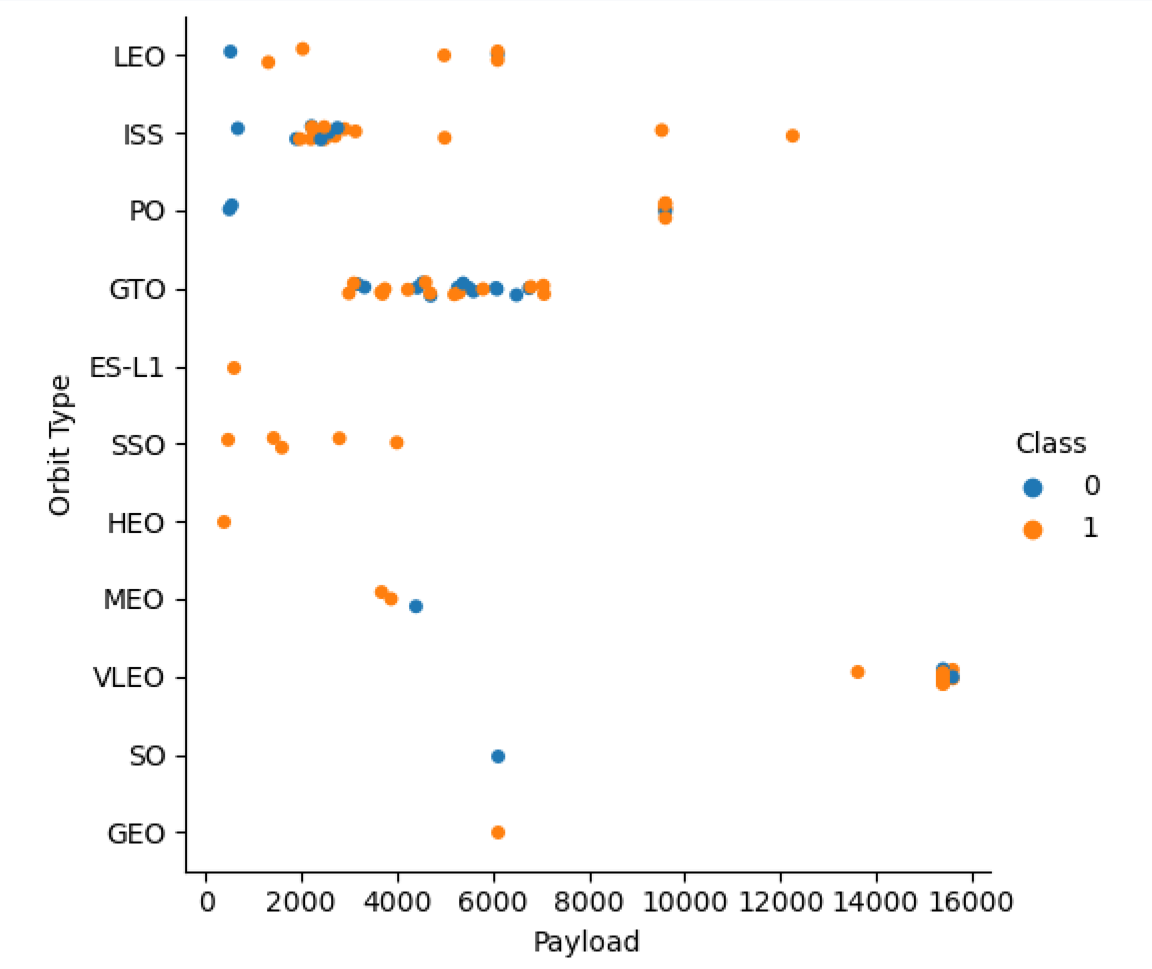
Flight Number vs. Orbit Type

- For the LEO orbit, the success rate appears related to the flight number. On the other hand, there seems to be no relationship between the flight number when in GTO orbit.



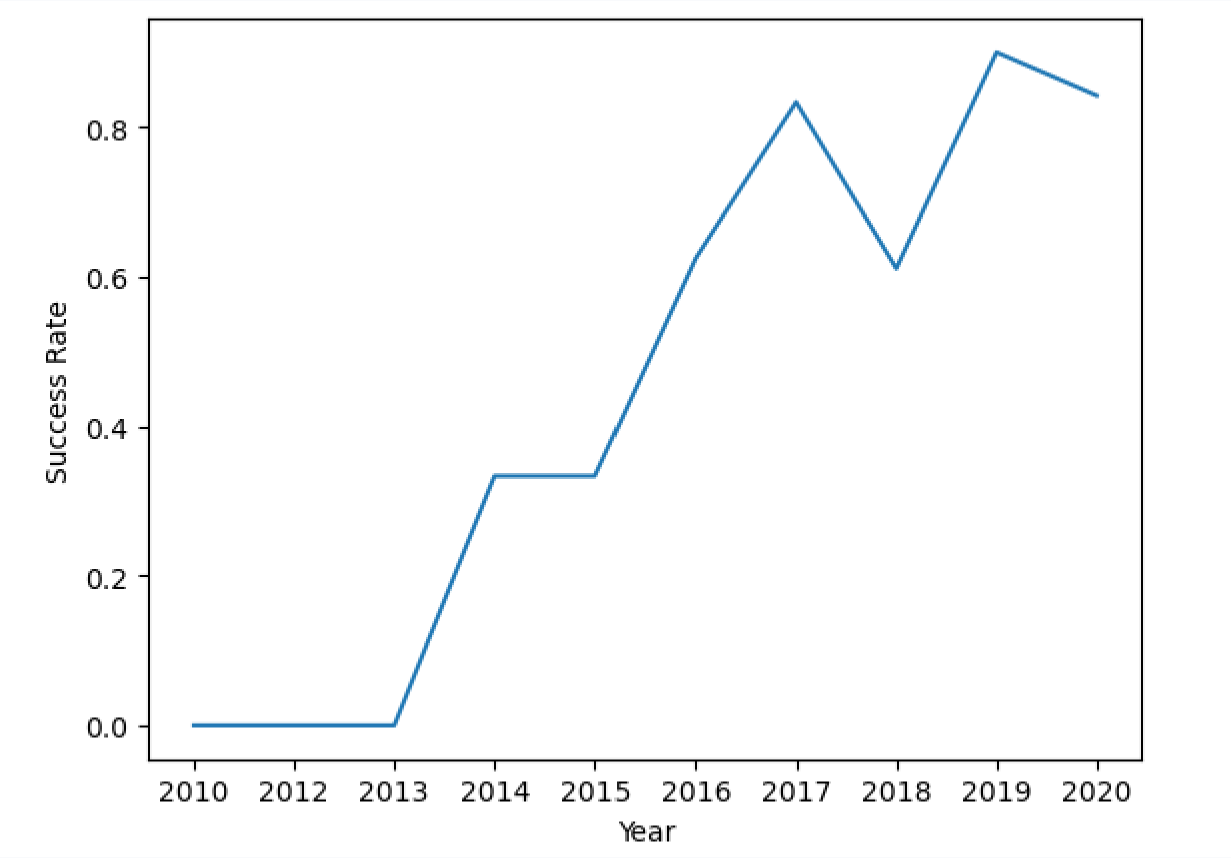
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are higher for the orbits Polar, VLEO, PO and ISS.
- However for GTO it is not clear-cut since both success and failure spread all over the payload range from 3000-7000 kg.



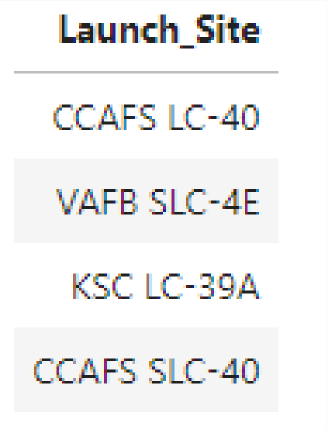
Launch Success Yearly Trend

- The success rate since 2013 has mostly been on a rising trend till 2020.



All Launch Site Names

The names of the unique launch sites:

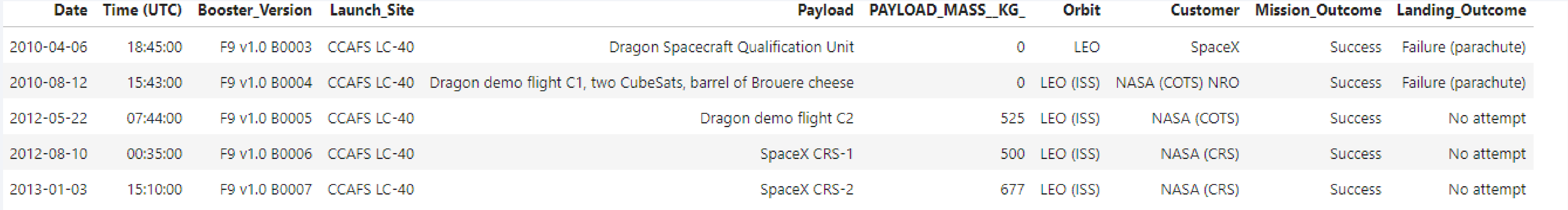


Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

A total of 4 different launch sites

Launch Site Names Begin with 'CCA'

The 5 records where launch sites begin with “CCA”

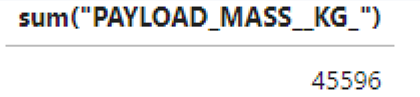


	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
	2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
	2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo 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

The records of the 5 earliest launch records at launch sites CCAFS LC-40

Total Payload Mass

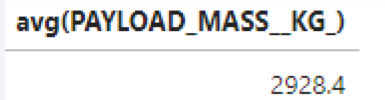
The total payload carried by boosters from NASA: 45596 kg



sum("PAYLOAD_MASS__KG_")
45596

Average Payload Mass by F9 v1.1

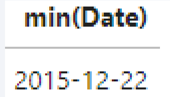
The average payload mass carried by booster version F9 v1.1: 2928.4 kg



avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

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

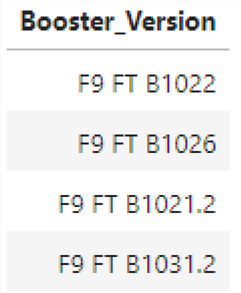


min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters that have successfully landed on drone ships and had payload mass greater than 4000 but less than 6000:

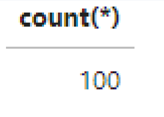
B1022, B1026, B1021.2, B1031.2



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

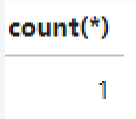
Total Number of Successful and Failure Mission Outcomes

The total number of successful mission outcomes: 100

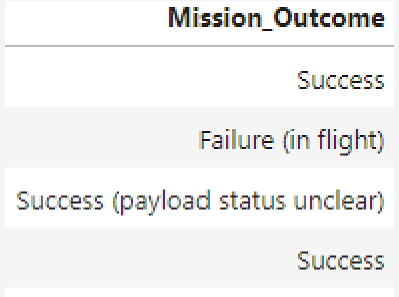


count(*)
100

The total number of failed mission outcomes: 1



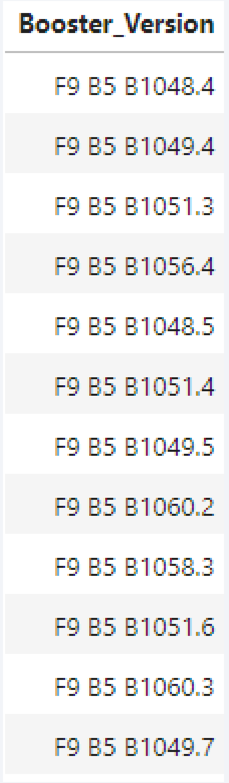
count(*)
1



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

Boosters Carried Maximum Payload

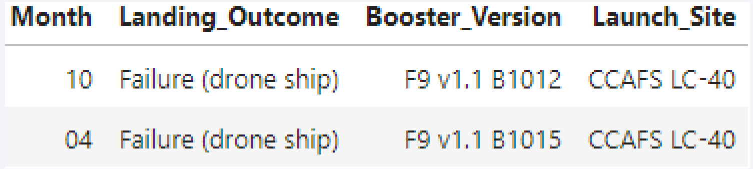
- The names of the booster that have carried the maximum payload mass



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

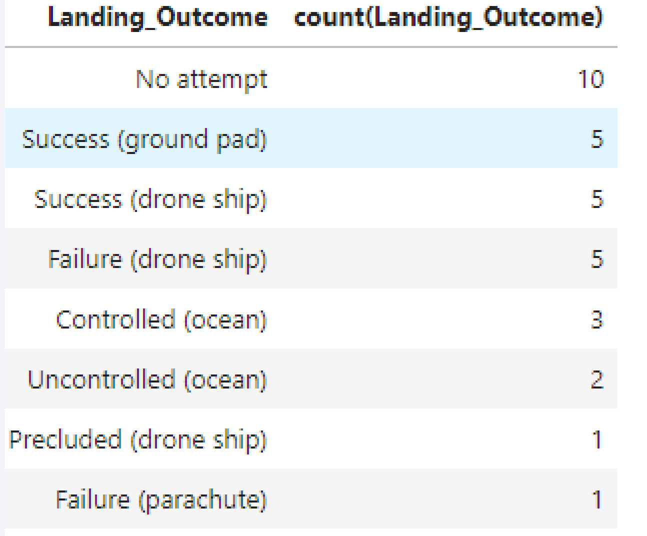
List the failed landing outcomes in drone ships, their booster versions, and launch site names for in year 2015



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

- 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



Landing_Outcome	count(Landing_Outcome)
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

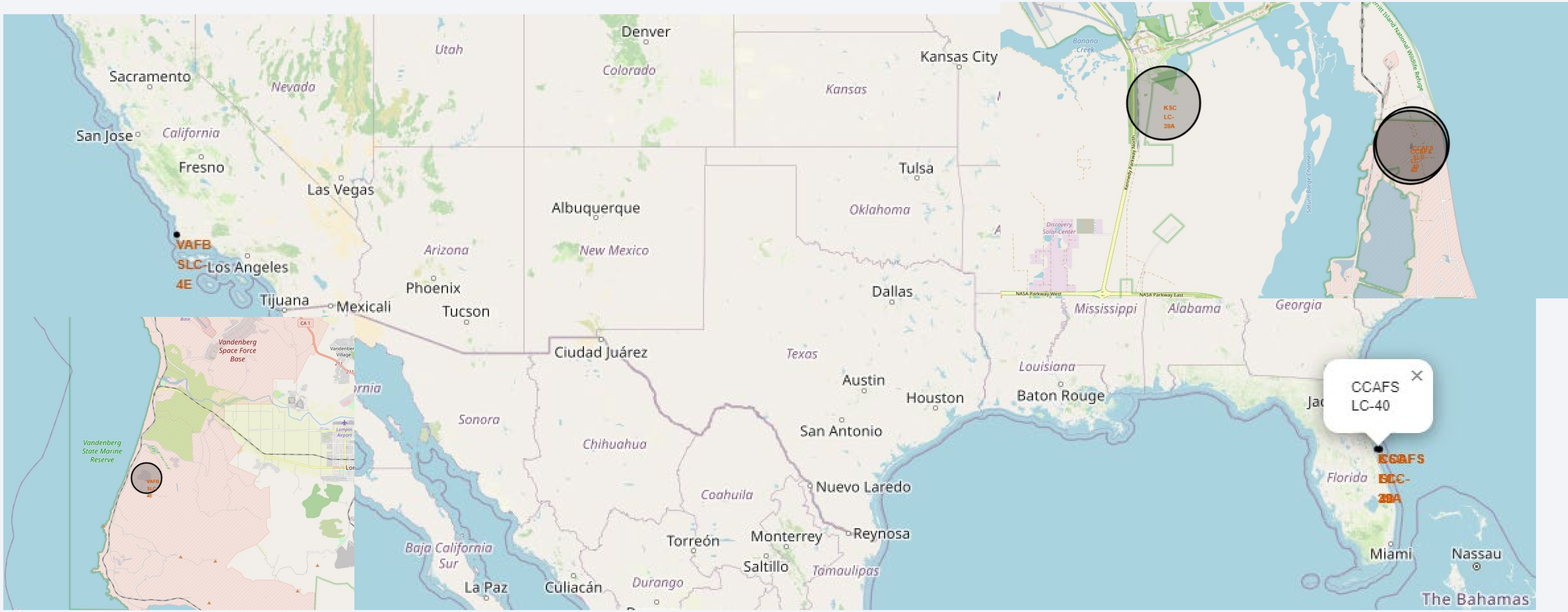
The locations of all launch sites on a map

- All four launch sites (shown as markers and circles)are in low-latitude areas

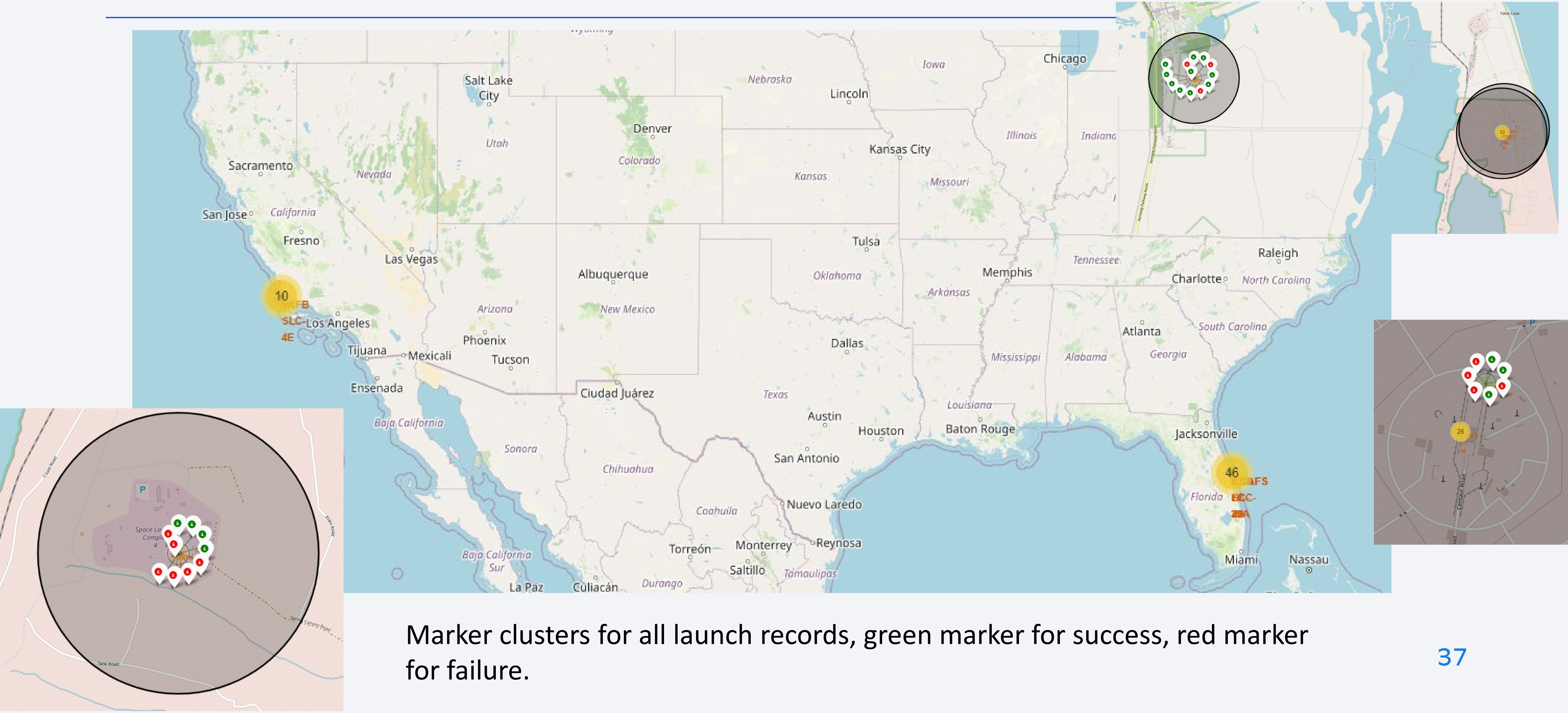
- Three of the four launch sites: KSC LC-39A, CCAFS LC-40, and CCAFS SLC-40 are located very close to each other on either Merritt Island or Cape Canaveral in Florida.

- One launch site VAFB SLC-4E is located near Vandenberg Space Force Base in California.

The locations of all launch sites on a map

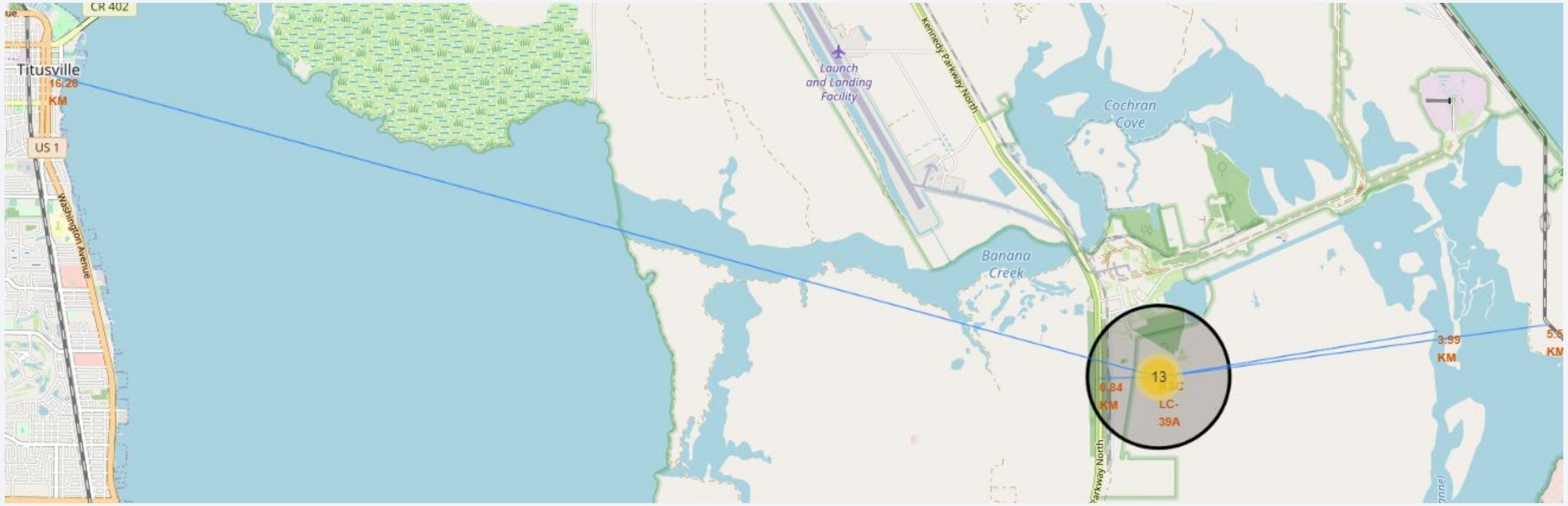


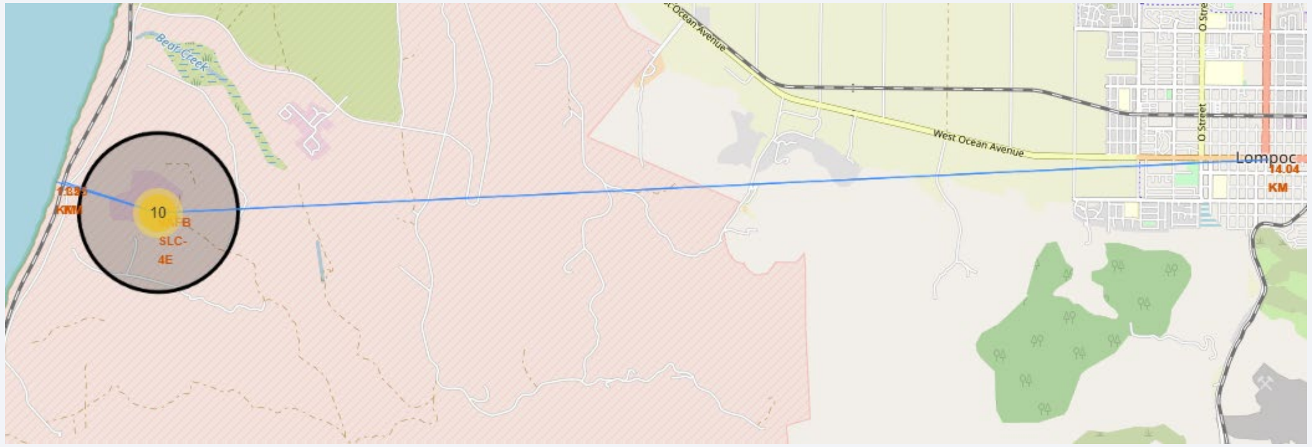
All launch records on the site map

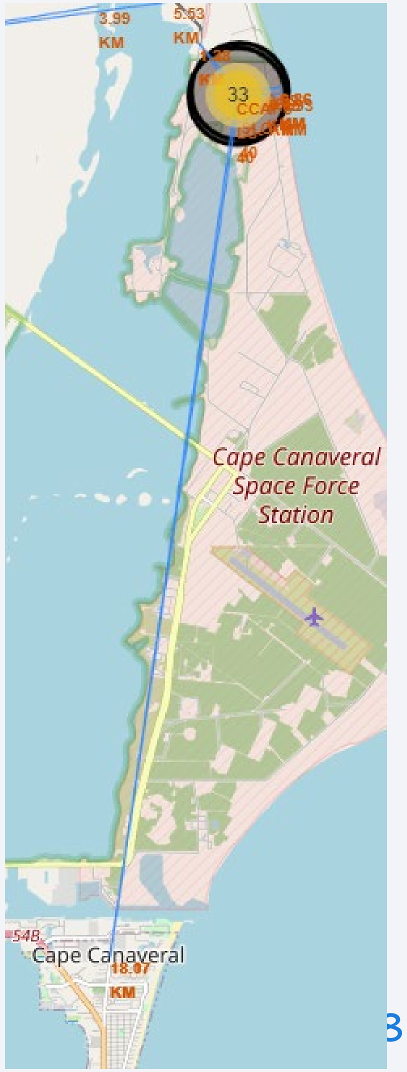


Marker clusters for all launch records, green marker for success, red marker for failure.

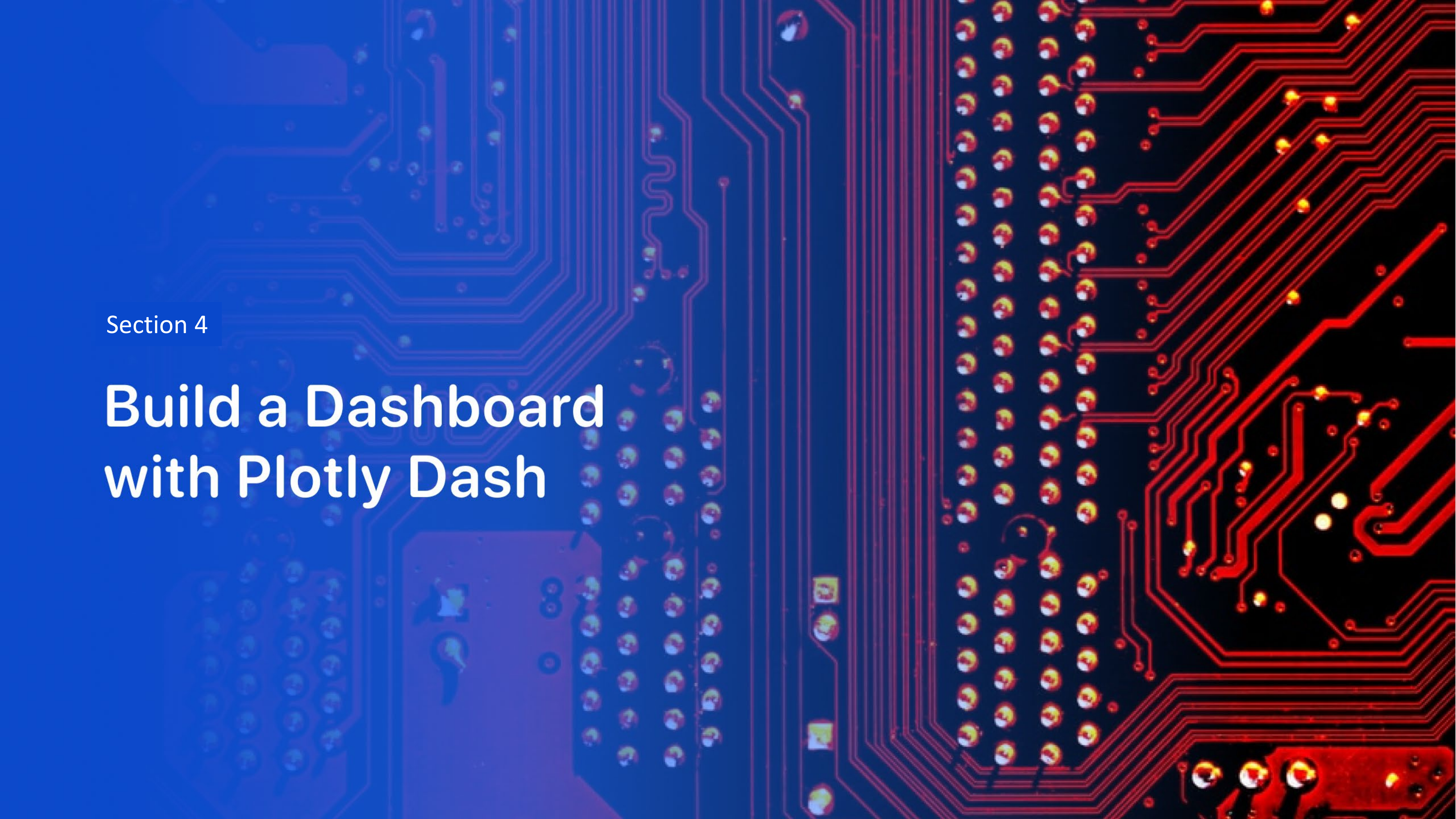
The proximities of launch sites







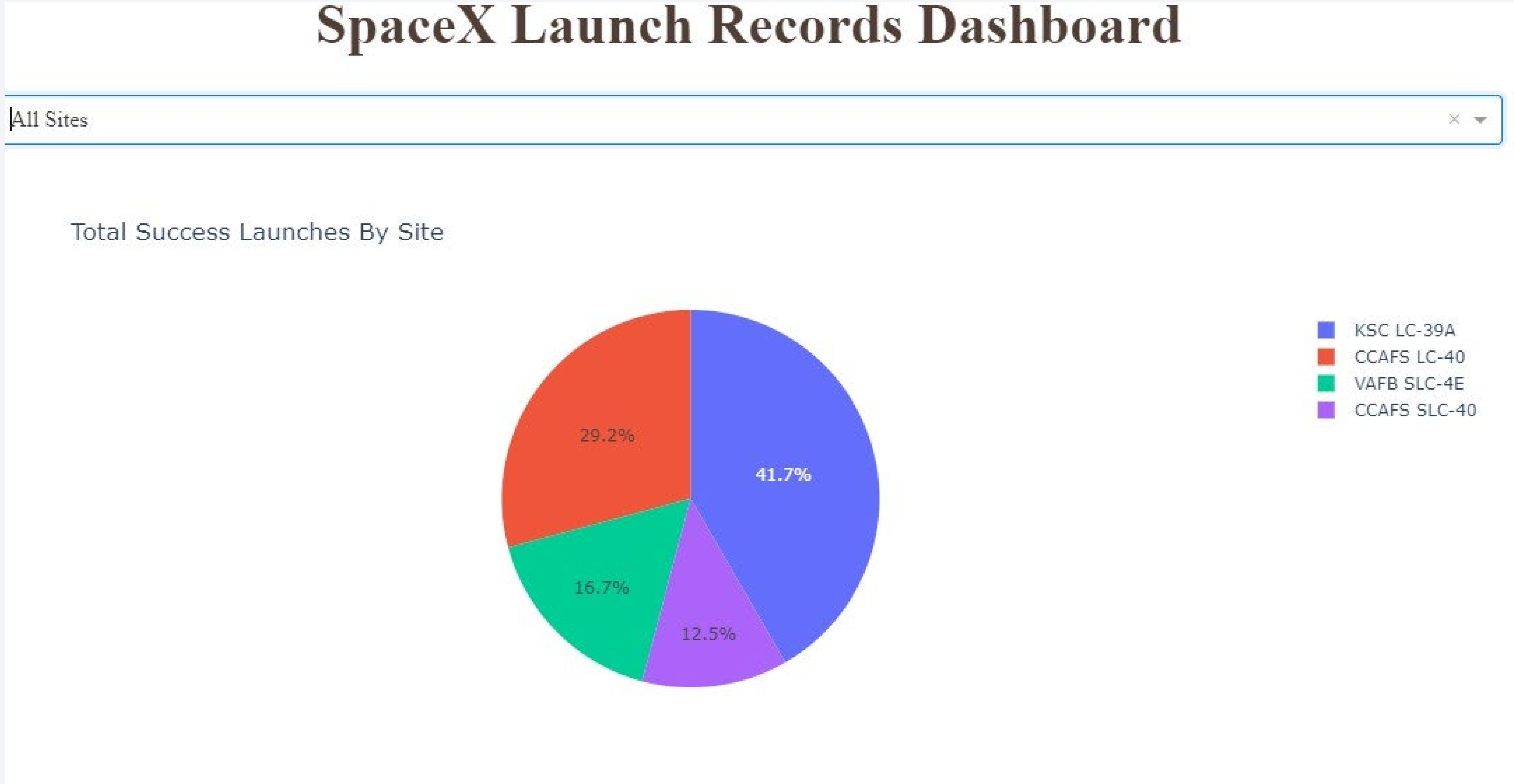
All 4 launch sites are in close proximity to railroads, highways, and coastline while keeping a distance of at least 10 kilometers from the nearest city



Section 4

Build a Dashboard with Plotly Dash

Total Success Launches In Percentage by Site

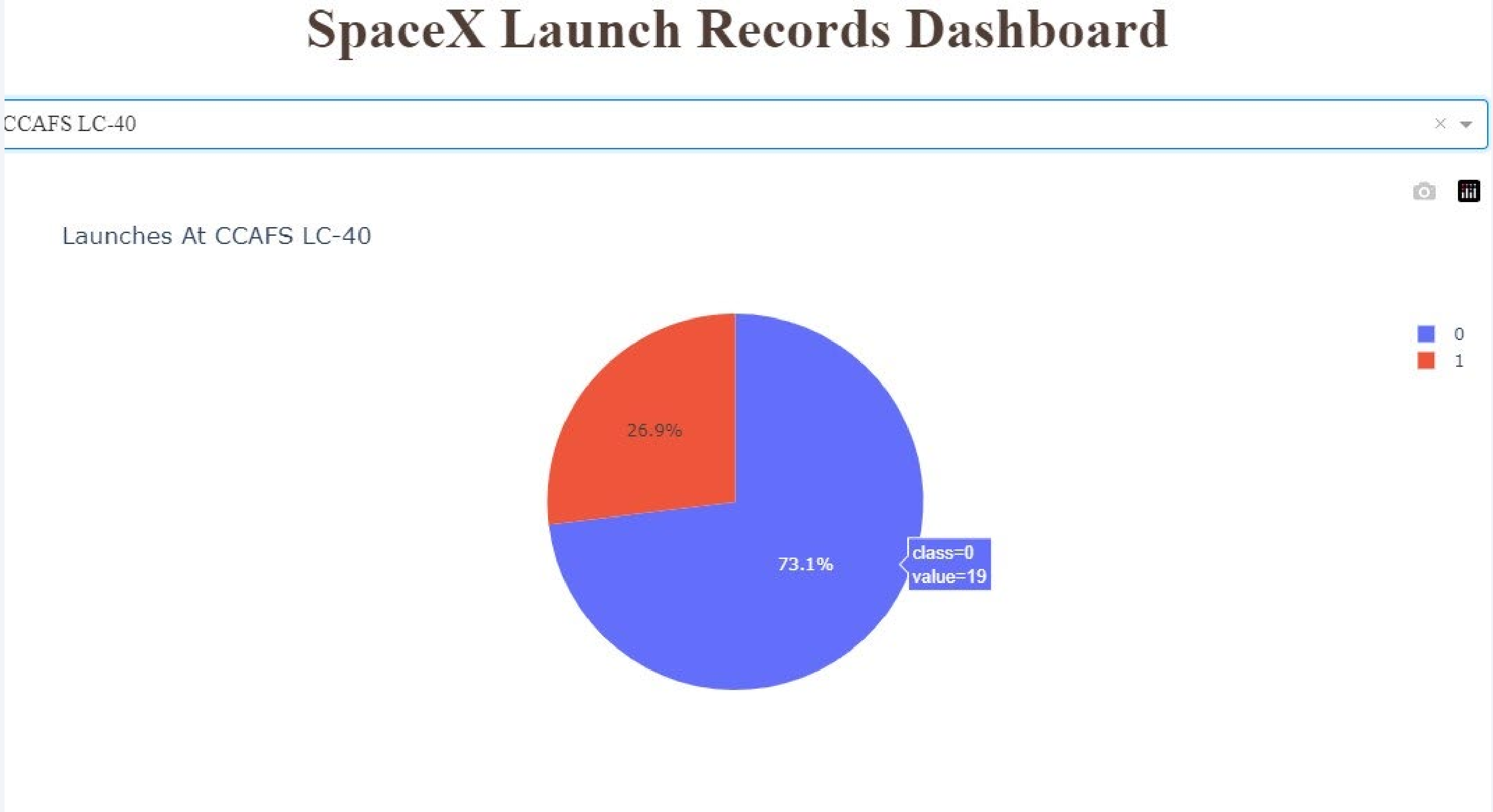


SpaceX Launch Records Dashboard

All Sites

The site KSC LC-39A has the most successful rocket launches (41.7%)

The Launch Site with the Highest Success Rate



SpaceX Launch Records Dashboard

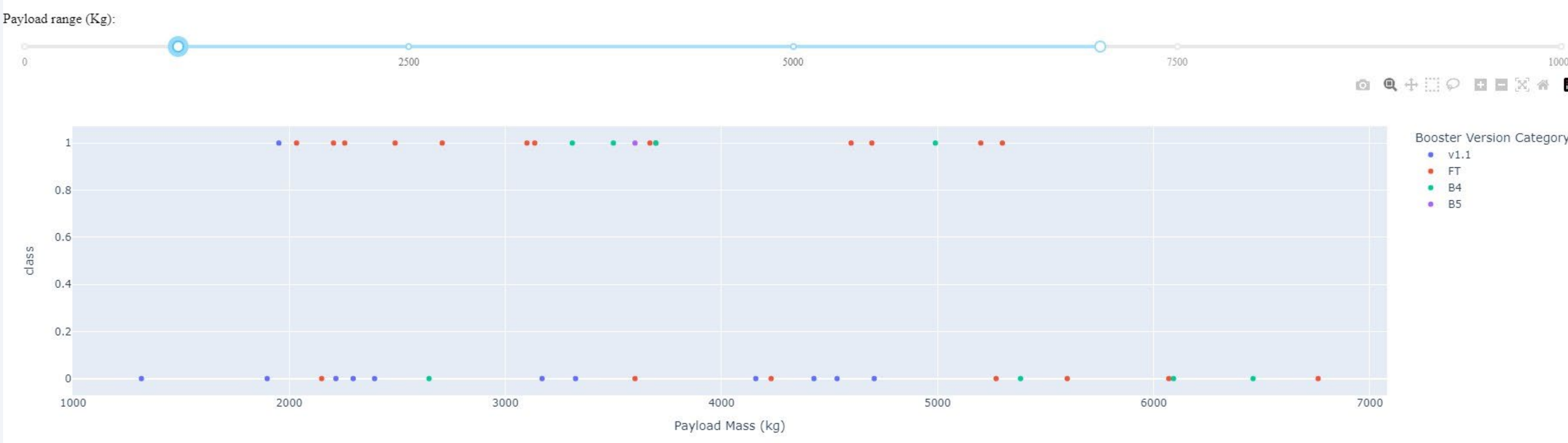
CCAFS LC-40

The launch site CCAFS LC-40 has the highest rate of successful launches at 73.1%

Payload vs. Launch Outcome

- Payload weighing 2k-4k kg has the highest launch success rate and payload weighing over 4k kg has the lowest launch success rate

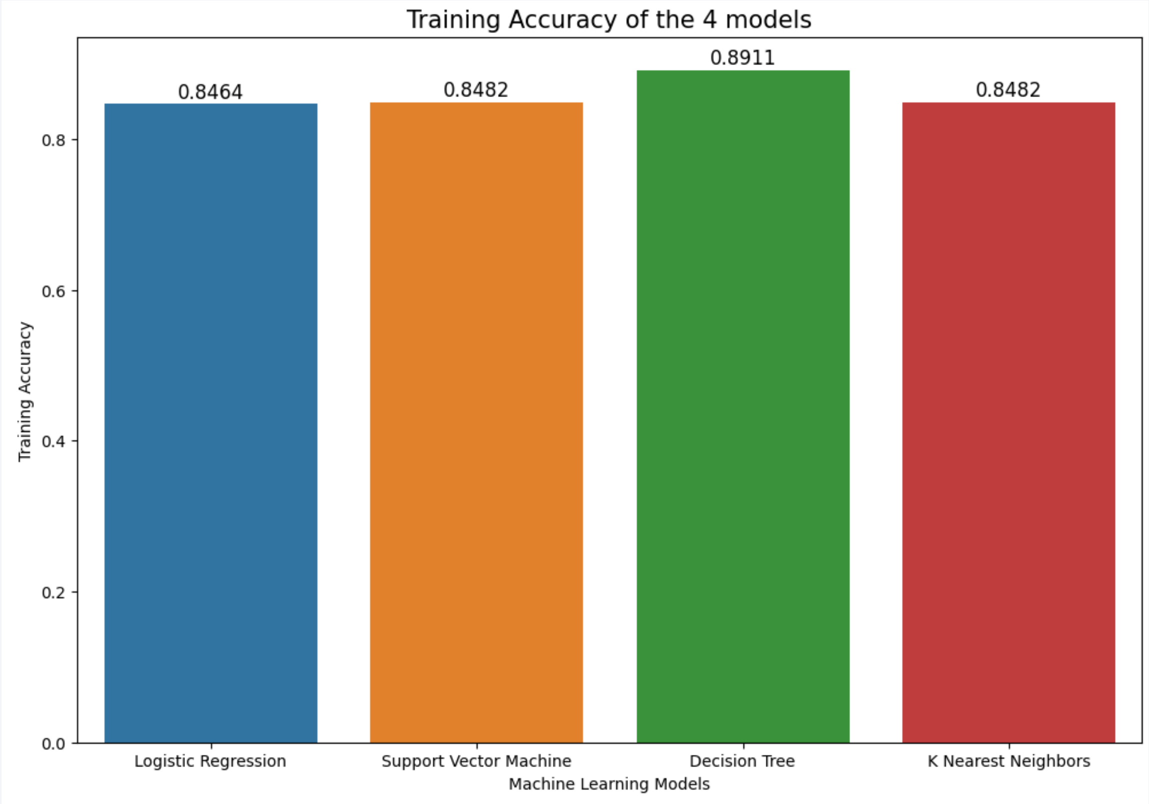
- Rocket models Falcon 9 FT and B5 have the highest launch success rate

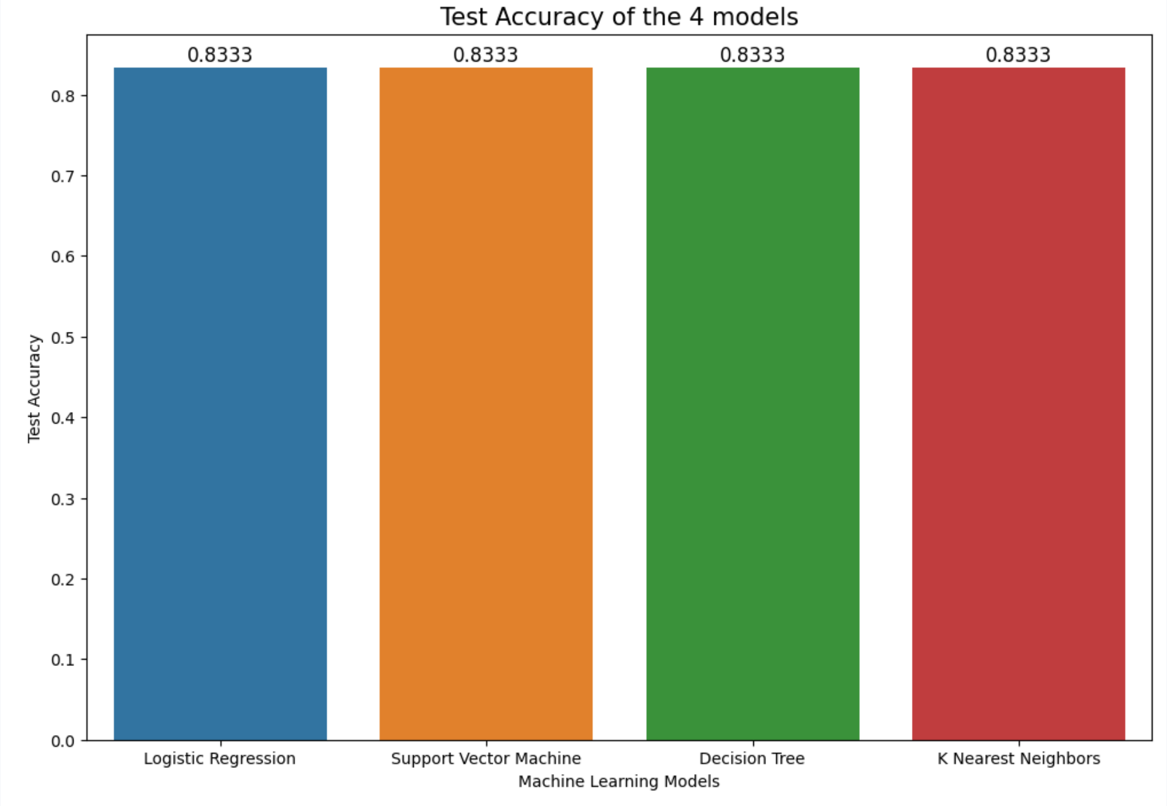


Section 5

Predictive Analysis (Classification)

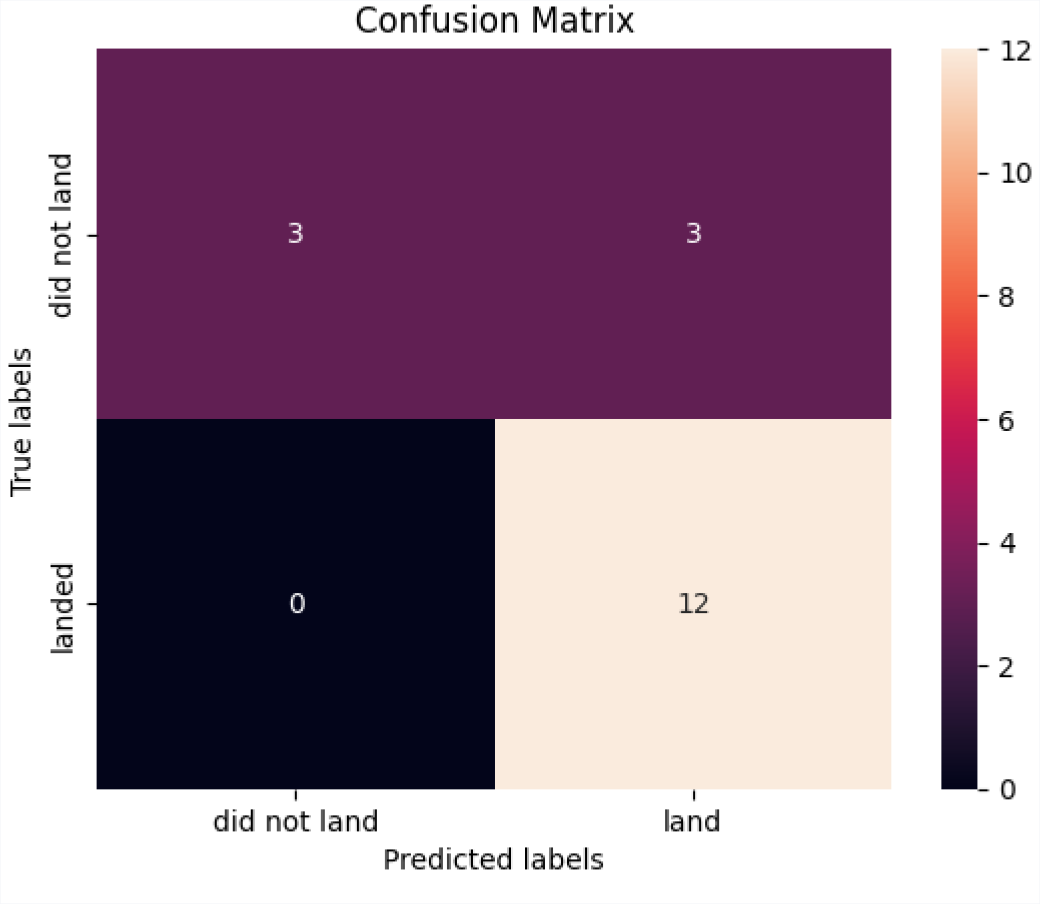
Classification Accuracy





- All the models (Logistic regression, SVM, Decision Tree and KNN) achieved the same accuracy of 83.33% on the test set.
- Decision Tree had the highest training accuracy of 89.11%

Confusion Matrix



All 4 models have the same confusion matrix.

We see the models can distinguish between the two classes (‘land’ and ‘did not land’) and there are no false positives (predicting ‘did not land’ but the rocket landed), but there are 3 false negatives (predicting ‘land’ but the rocket did not land), which is a major problem of the model.

Conclusions

- While the models have different training accuracies, they all have the same test accuracy, which is somewhat lower than their respective training accuracy.

Especially for the Decision Tree model, the test accuracy (83.33%) is significantly lower than the training accuracy (89.11%) suggesting a high variance problem. Regularization methods can be introduced to mitigate the overfitting problem and may achieve better performance on test data.

- False negatives are a major problem of all the models according to the confusion matrix, potential measures to address this problem may be collecting more data, grid searching more extensively over the hyperparameter space, using domain knowledge of rocket science to select more relevant features in the data set, and so on.

Appendix

For any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets, see https://github.com/YMaXing/IBM-Data-Science-Capstone

Thank you!