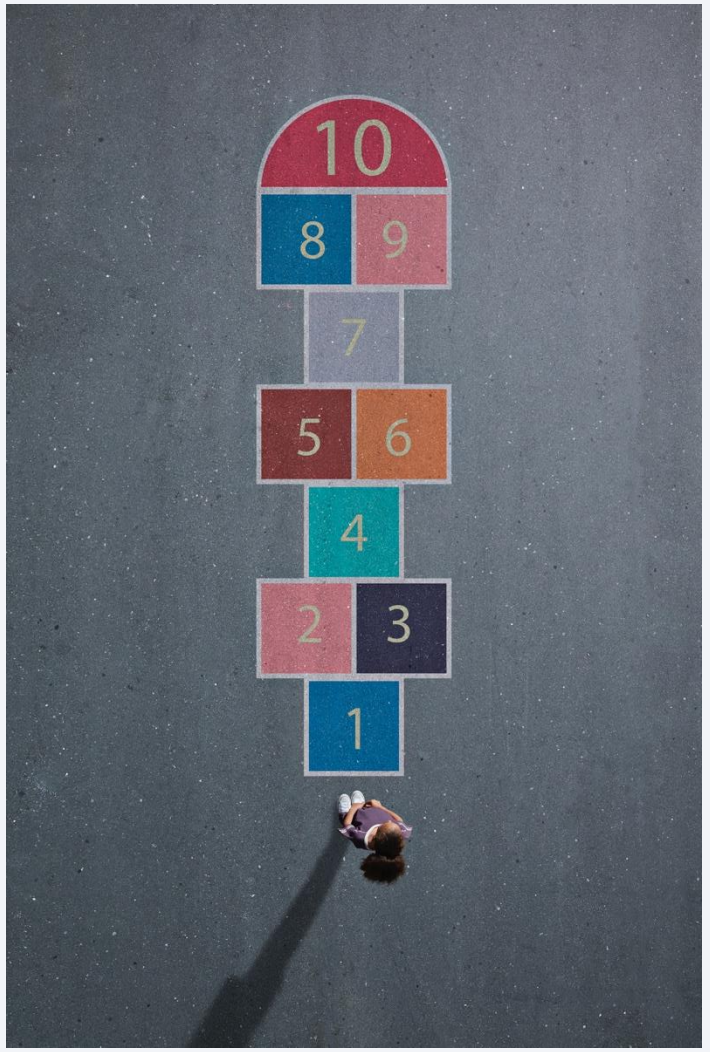
IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Bernard van Middendorp
February 18, 2025

Outline



Executive Summary

- Summary of methodologies
  - Data collection
  - Data Wrangling
  - EDA with Data Visualization
  - EDA with SQL
  - Build an Interactive Map with Folium
  - Build a Dashboard with Plotly Dash
  - Predictive Analysis (Classification)
- Summary of all results
  - Exploratory data analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Introduction

Project Background and Context

In this capstone project, our goal is to predict the successful landing of the Falcon 9 first stage. SpaceX promotes its Falcon 9 rocket launches on its website, priced at 62 million dollars, while other providers charge over 165 million dollars per launch. A significant part of the cost savings comes from SpaceX's ability to reuse the first stage. Thus, if we can ascertain whether the first stage will land successfully, we can estimate the overall launch cost. This analysis could be valuable for any competing company interested in bidding against SpaceX for a rocket launch.

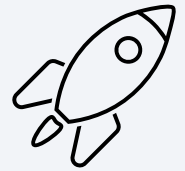
Questions to be answered

- On which variables is the mission-outcome dependable.
- Can we predict the mission outcome and if so, what model is the best

Section 1

Methodology

Methodology



- Data collection methodology:

  The data was collected by accessing the SpaceX data API and web scraping of the SpaceX Wikipedia webpage

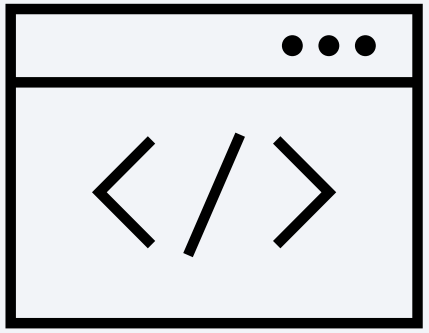
- Perform data wrangling

  The mission outcome was transformed in numerical inputs (binary encoding)

- 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



The data was collected by accessing the SpaceX Data API and by web scraping of Wikipedia page of SpaceX

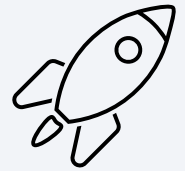
- SpaceX Data API: The data was directly turned into a Panda's data frame using the json-normalize function

  `https://api.spacexdata.com/v4/launches/past`

- The Wikipedia page of SpaceX contains a table with Falcon9 flights and was scraped by BeautifulSoup

  `https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922`

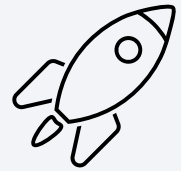
Data Collection – SpaceX API

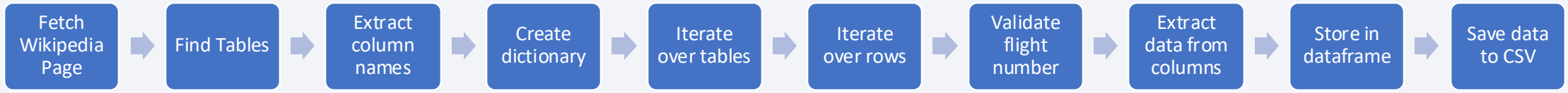


1. Fetch past launch Data: https://api.spacexdata.com/v4/launches/past
2. Loop through rocket column: https://api.spacexdata.com/v4/rockets/{rocket_id}
3. Loop trough launchpad column: https://api.spacexdata.com/v4/launchpads/{launchpad_id}
4. Loop trough payloads column: https://api.spacexdata.com/v4/payloads/{payload_id}
5. Loop trough cores column: https://api.spacexdata.com/v4/cores/{core_id}

GitHub: Data Collection - SpaceX API

Data Collection - Scraping



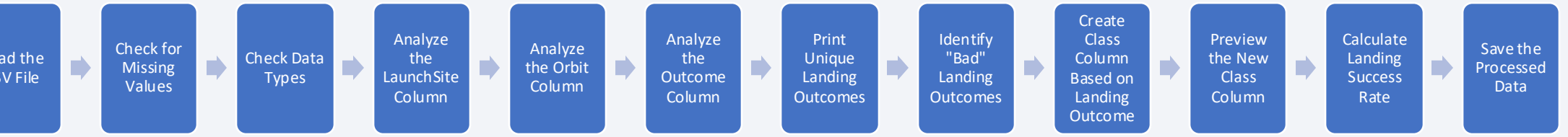


- Fetch Wikipedia Page by BeautifulSoup
- Find Tables with class "wikitable"
- Extract column names (th elements)
- Create dictionary (init dictionary)
- Iterate over each table in HTML
- Iterate over rows (tr elements)
- Validate flight number
- Extract data from columns

  Date and Time, Booster Version, Launch Site, Payload, Payload Mass, Orbit, Customer, Launch Outcome, Booster Landing
- Store in Pandas DataFrame
- Save data to spacex_webscraped.csv

GitHub: Data Collection - Scraping

Data Wrangling



The main purpose was adding the 'Class' column to the dataset. The column has a value of

- 0: if the mission outcome was unsuccessful
- 1: if the mission outcome was successful

The mission outcome was considered successful if the column 'Outcome' contains 'True' otherwise not.

The successful rate could now easy be calculated by the mean of 'Class' times 100. It was 66.67%

GitHub: Data Wrangling

EDA with Data Visualization

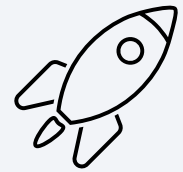
Used the following charts:

- Scatter plots to see relationships between variables:*
  - Flight-number vs Payload Mass
  - Flight-number vs Launch Site
  - Payload mass vs Launch Site
  - Flight-number vs Orbit type
  - Payload mass vs Orbit type
- Bar chart for the success-rate of orbit types
- Line plot for the success-rate over time.

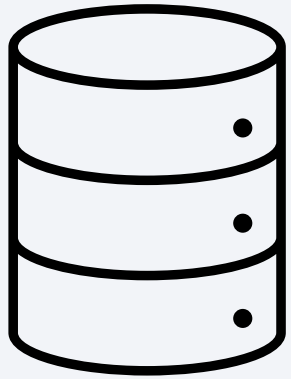
* the color of the dot depends on the mission outcome

GitHub: EDA with Data Visualization

EDA with SQL



Executed SQL queries:



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

```
SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
```

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

```
SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
```

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

```
SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)';
```

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

```
SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1';
```

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

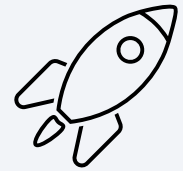
```
SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)'
```

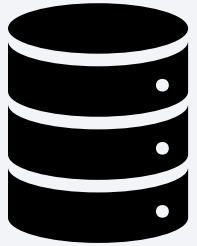
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' AND
PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000
```

GitHub: EDA with SQL

EDA with SQL





Executed SQL queries:

- List the total number of successful and failure mission outcomes

```
Select Mission_Outcome, COUNT(*) AS count FROM SPACEXTABLE GROUP BY Mission_Outcome;
```

- List the names of the booster_versions which have carried the maximum payload mass

```
SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
```

- 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

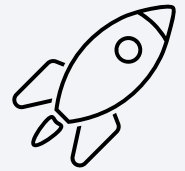
```
SELECT substr(Date, 6,2) as month, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE substr(Date,0,5)='2015' AND Landing_Outcome = 'Failure (drone ship)';
```

- 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

```
SELECT Landing_Outcome, Count(Landing_Outcome) as count FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY count DESC;
```

GitHub: EDA with SQL

Build an Interactive Map with Folium

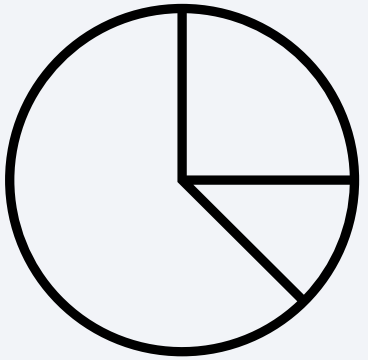


Used Folium objects:

- Circle: to highlight the launch-sites
- Label: to add the name of the launch site
- Marker:
  - To show the outcome of the mission on the launchsite (multiple markers on the same location can use MarkerCluster) by color.
  - To show the distance to nearby object
- Lines: To draw lines to nearby objects (such as coastline, roads, cities etc)

GitHub: Interacive Map with Folium

Build a Dashboard with Plotly Dash



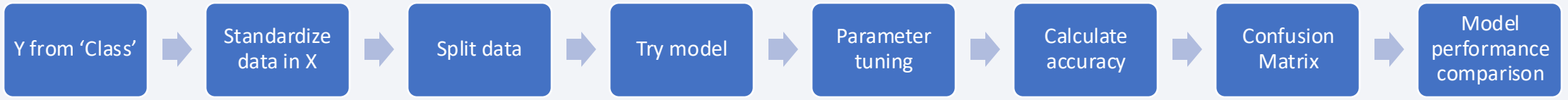
Used components

- Input
  - Dropdown: To select the launch site (or ALL)
  - Slider: Select the payload mass range
- Output:
  - Pie chart: The pie chart shows the success-rate for the selected launch site (or all, if ALL is selected)
  - Scatter plot: Mission Outcome vs Payload Mass for different booster versions

With this dashboard it was possible to see the relationship between payload mass, booster version, launch site and the mission outcome

GitHub: Dashboard with Plotly Dash

Predictive Analysis (Classification)

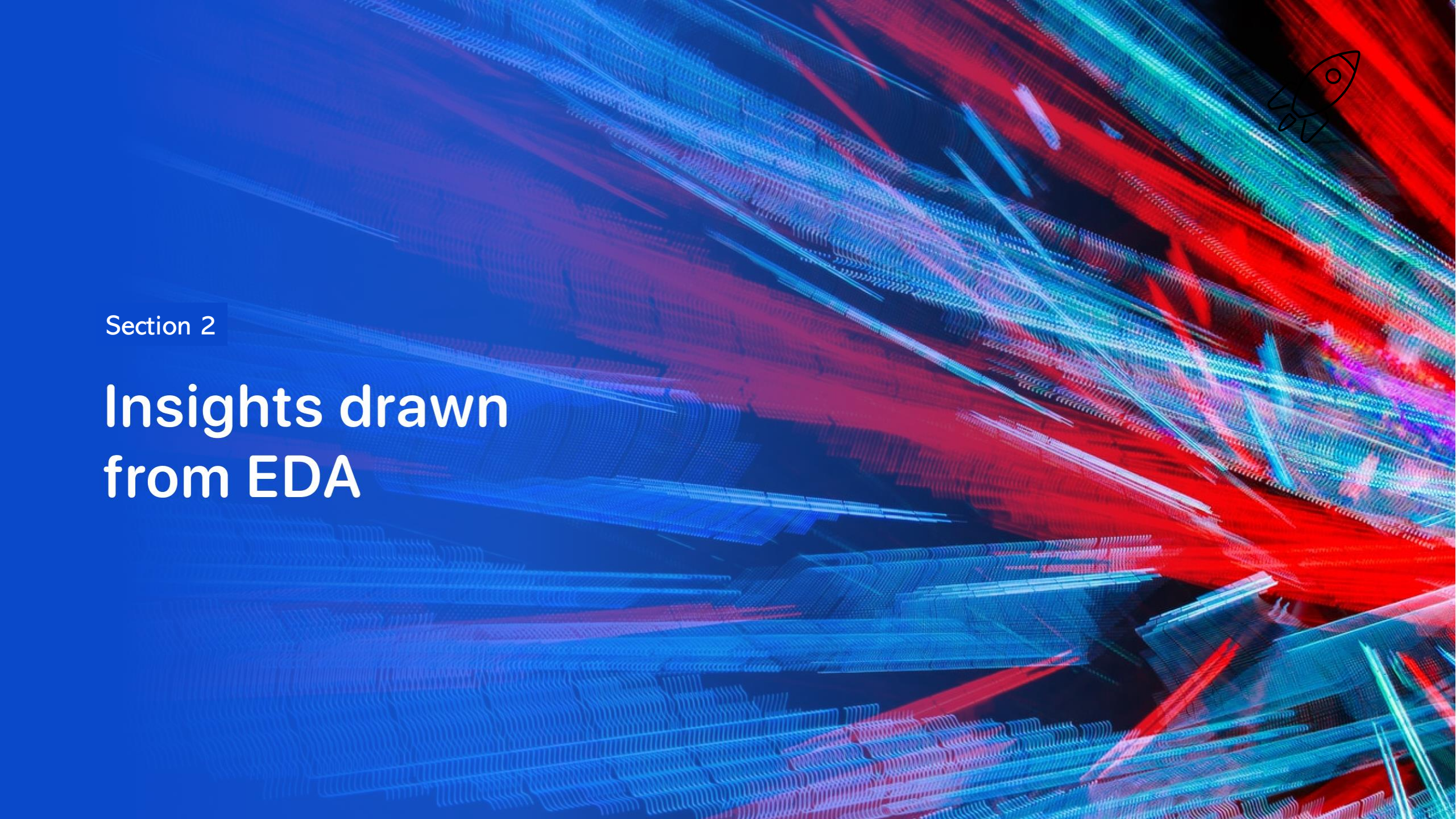


- Y from 'Class': create Numpy array from 'Class'
- Standardize data in X
- Split data: Split data in to training and testing data
- Try model: Fit the following models (Logistic Regression, Support Vector Machine, Decision Tree and K-Nearest Neighbors)
- Parameter tuning: For each model tune the parameters by GridSearchCV
- Calculate accuracy: Calculate the accuracy per model
- Confusion Matrix: Look per model at the confusion matrix. They were the same, except for Decision Tree
- Model performance comparison: compare the models by Accuracy, Precision, Recall, F1, F0.5 score and ROC-AUC and select the best model

GitHub: Predictive Analysis

Results

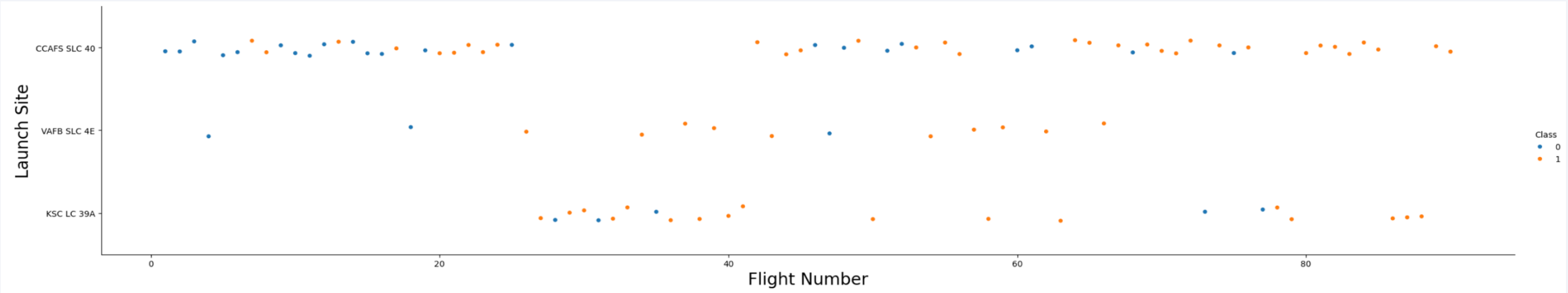
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



Section 2

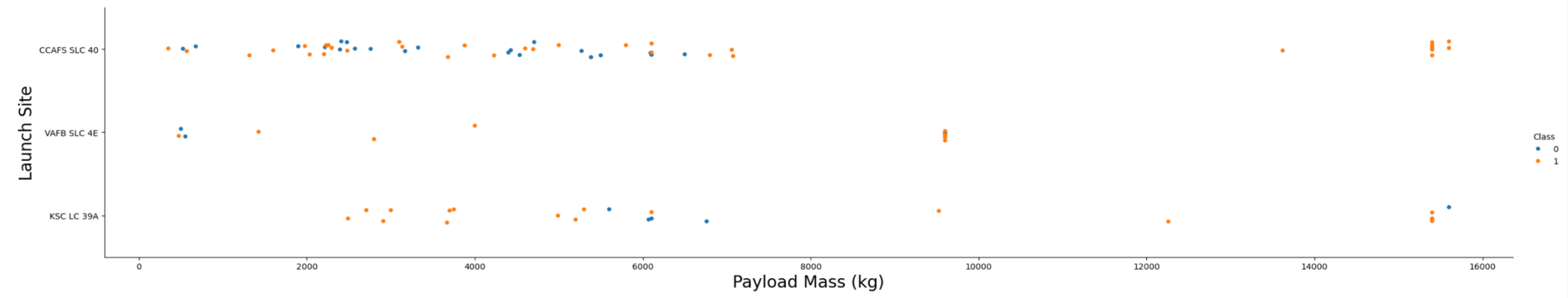
Insights drawn from EDA

Flight Number vs. Launch Site



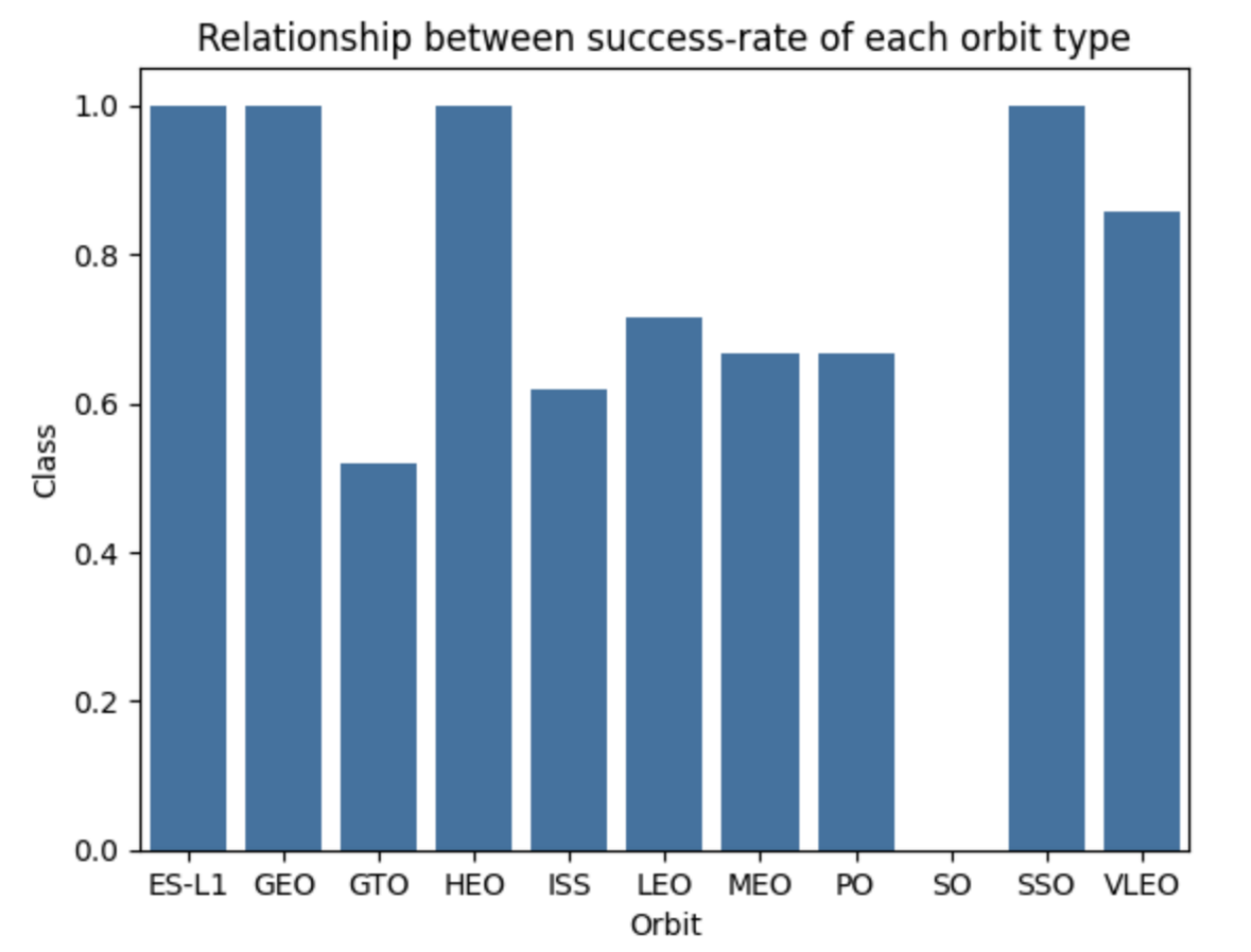
- Steady use of CCAFS SLC 40 over time
- VAFB SLC 4E seems not used anymore
- KSC LC 39A start from around flight number 20 and still in use

Payload vs. Launch Site



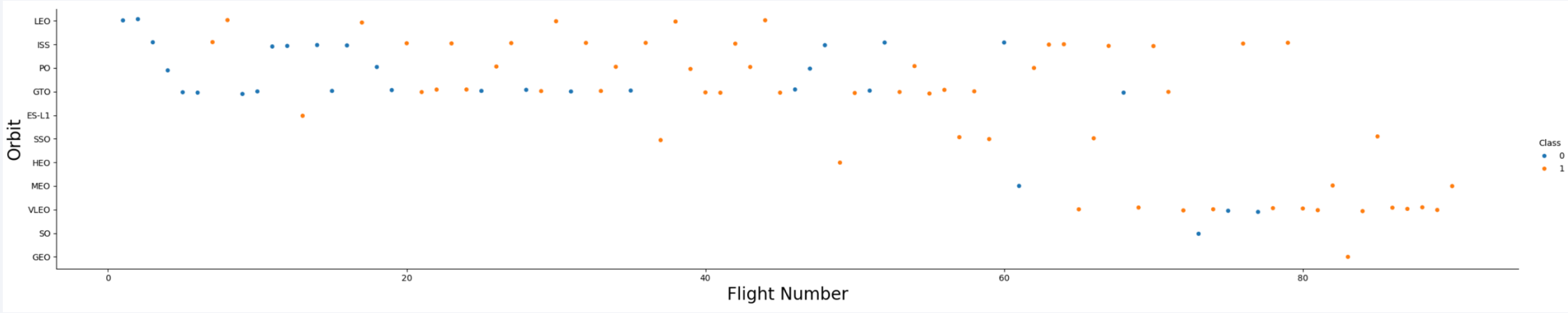
- 100% success-rate for heavy payloads at CCAFS SCL 40 launch site
- VAFB SCL 4E doesn't launch rockets with heavy payloads (> 10000)
- KSC LC 39A doesn't launch rockets with light payloads (< 2000 kg)

Success Rate vs. Orbit Type



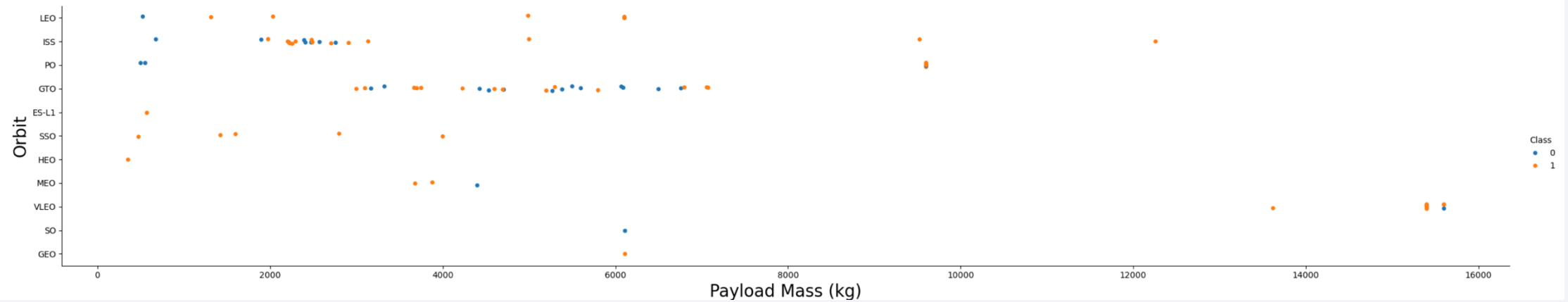
- SO: 0% success-rate
- ES-L1, GEO, HEO and SSO: 100% success-rate
- Others above 50%

Flight Number vs. Orbit Type



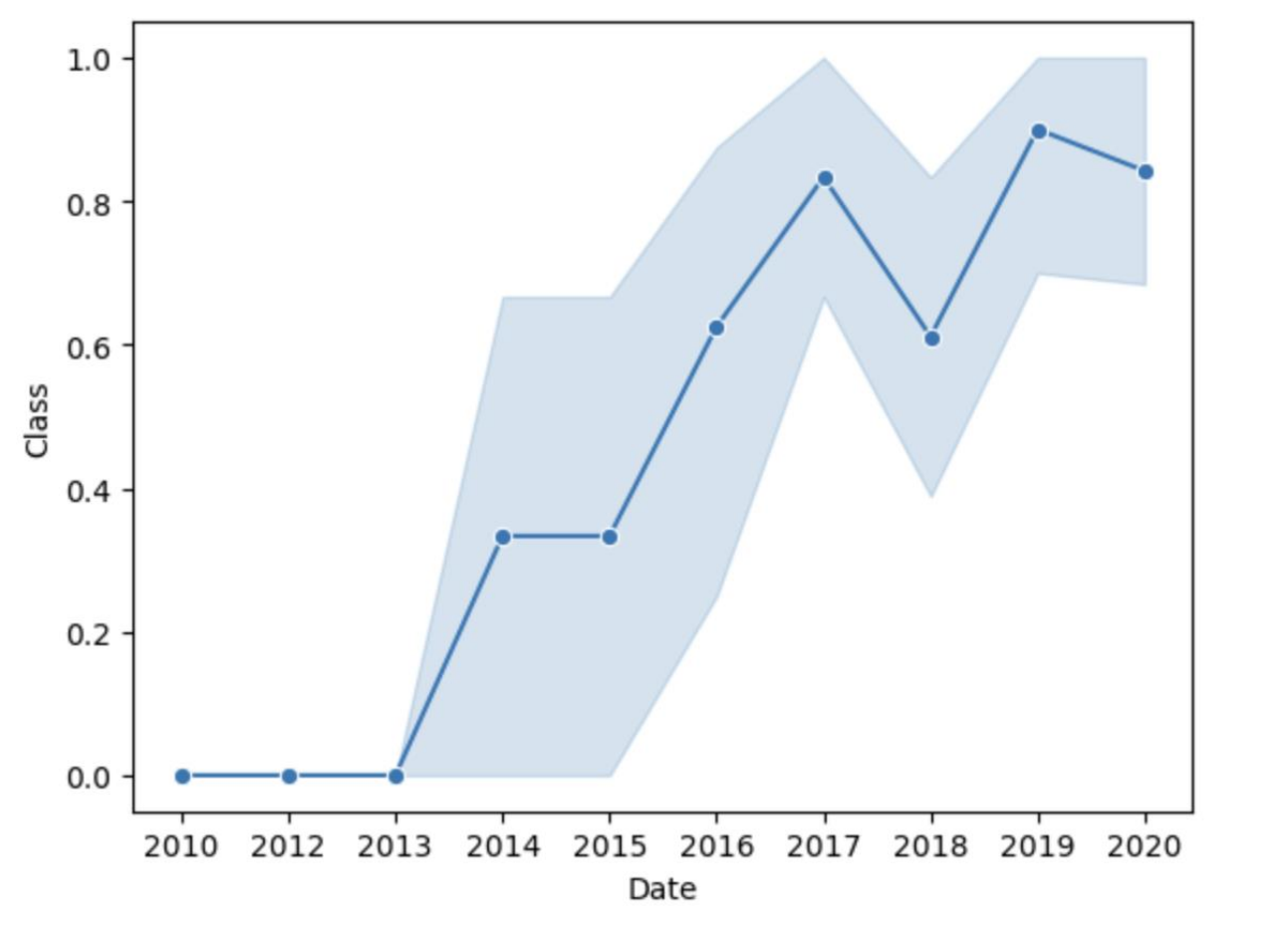
- Started with LEO orbit, but not used anymore
- Steady use of ISS over time
- Some use of SSO, HEO, MEO, SO, GEO orbits, but not huge and started late.
- Start using VLEO from around flight 60 and keeps used (propably because it has a high success-rate)

Payload vs. Orbit Type



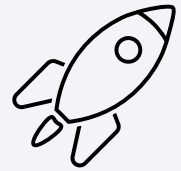
- The payload mass for the ISS typically ranges from 2000 to 3000 kg.
- For GTO, the mass usually falls between 3000 and 7000 kg.
- The payloads sent into VLEO orbit are very heavy (> 13000 kg)

Launch Success Yearly Trend



- The success rate improves as time progresses.
- This trend was also evident in the scatterplots that included flight numbers.

All Launch Site Names



```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
```

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

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

The table contains 4 unique launch sites

Launch Site Names Begin with 'CCA'

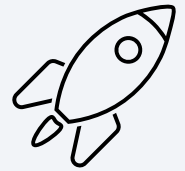
```
%sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
```

```
 * 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

Above the first 5 records where the launch sites begin with `CCA`

Total Payload Mass



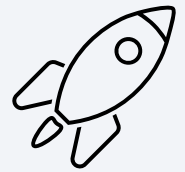
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)';
```

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

SUM(PAYLOAD_MASS__KG_)
45596

The total payload mass carried by boosters from NASA is 45596

Average Payload Mass by F9 v1.1



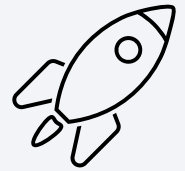
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1';
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

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

First Successful Ground Landing Date



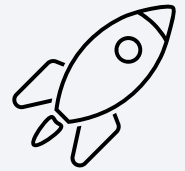
```
%sql SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)'
```

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

MIN(Date)
2015-12-22

The first successful landing outcome on ground pad was on 2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000



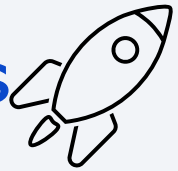
```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000
```

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

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

The query gives a list 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



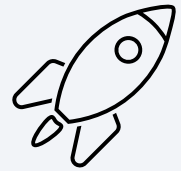
```
%sql Select Mission_Outcome, COUNT(*) AS count FROM SPACEXTABLE
```

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

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

- Success: 99 (98+1, probably because of an extra whitespace)
- Failure (in flight): 1
- Success (payload status unclear): 1

Boosters Carried Maximum Payload



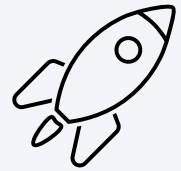
```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

```
 * 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

Above a list of booster versions which carried the maximum payload mass

2015 Launch Records



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

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

month	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

Above a list of booster versions, launch sites and month numbers of missions which failed landing on a drone ship in the year 2015

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

```
%sql SELECT Landing_Outcome, Count(Landing_Outcome) as count FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'  GROUP BY Landing_Outcome ORDER BY count DESC
```

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

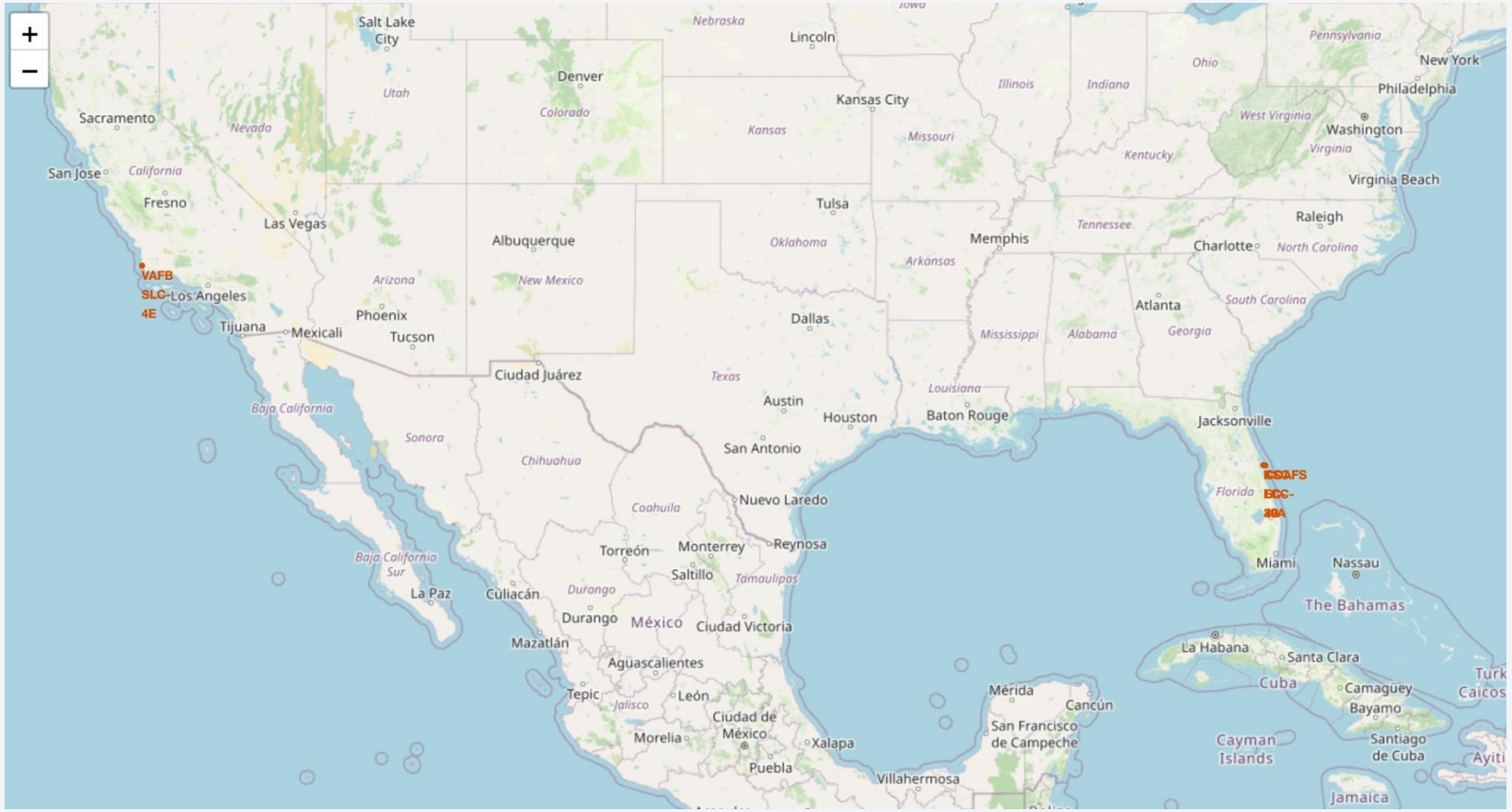
Landing_Outcome	count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Above 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

Section 3

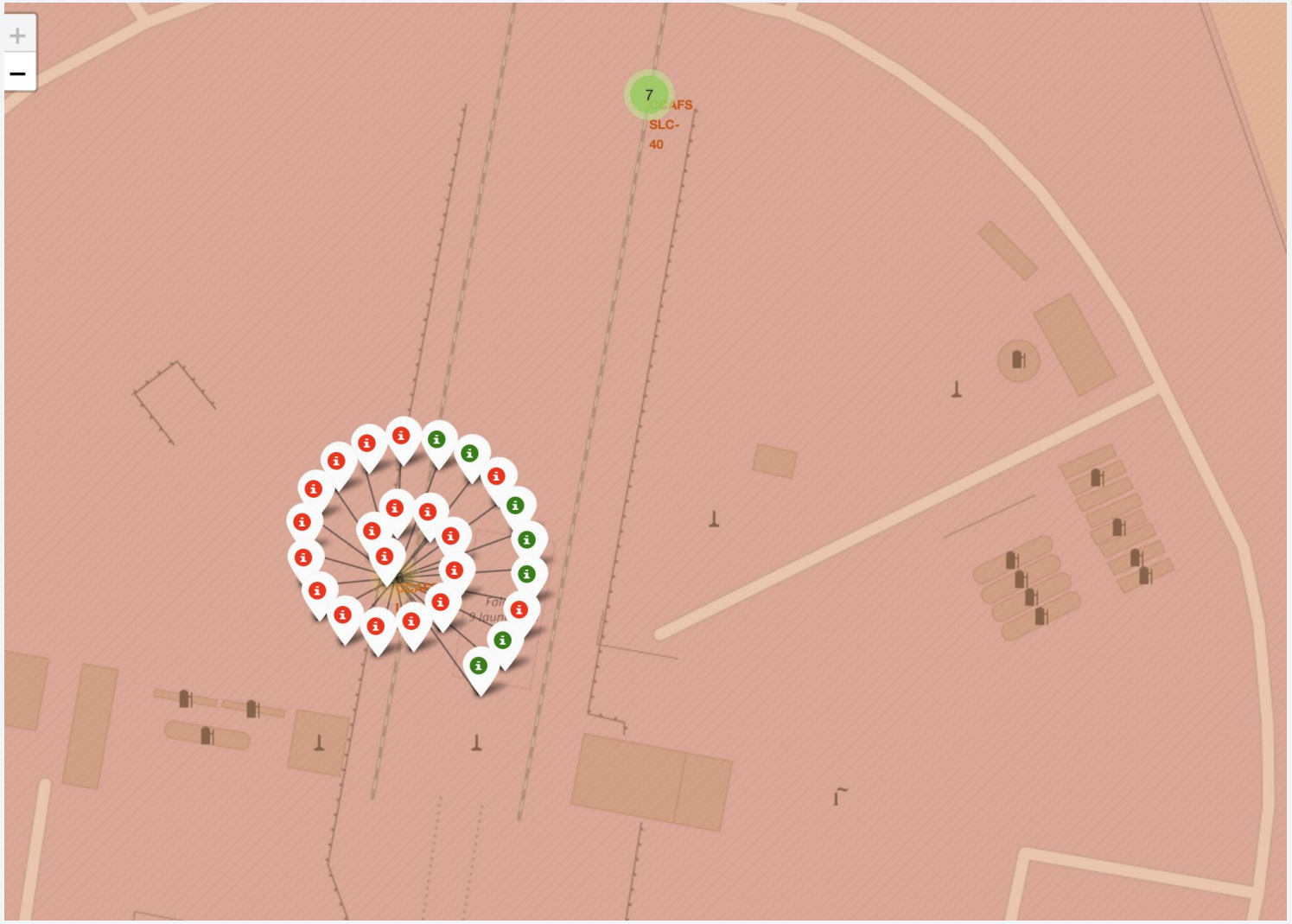
Launch Sites Proximities Analysis

Locations of the launch sites



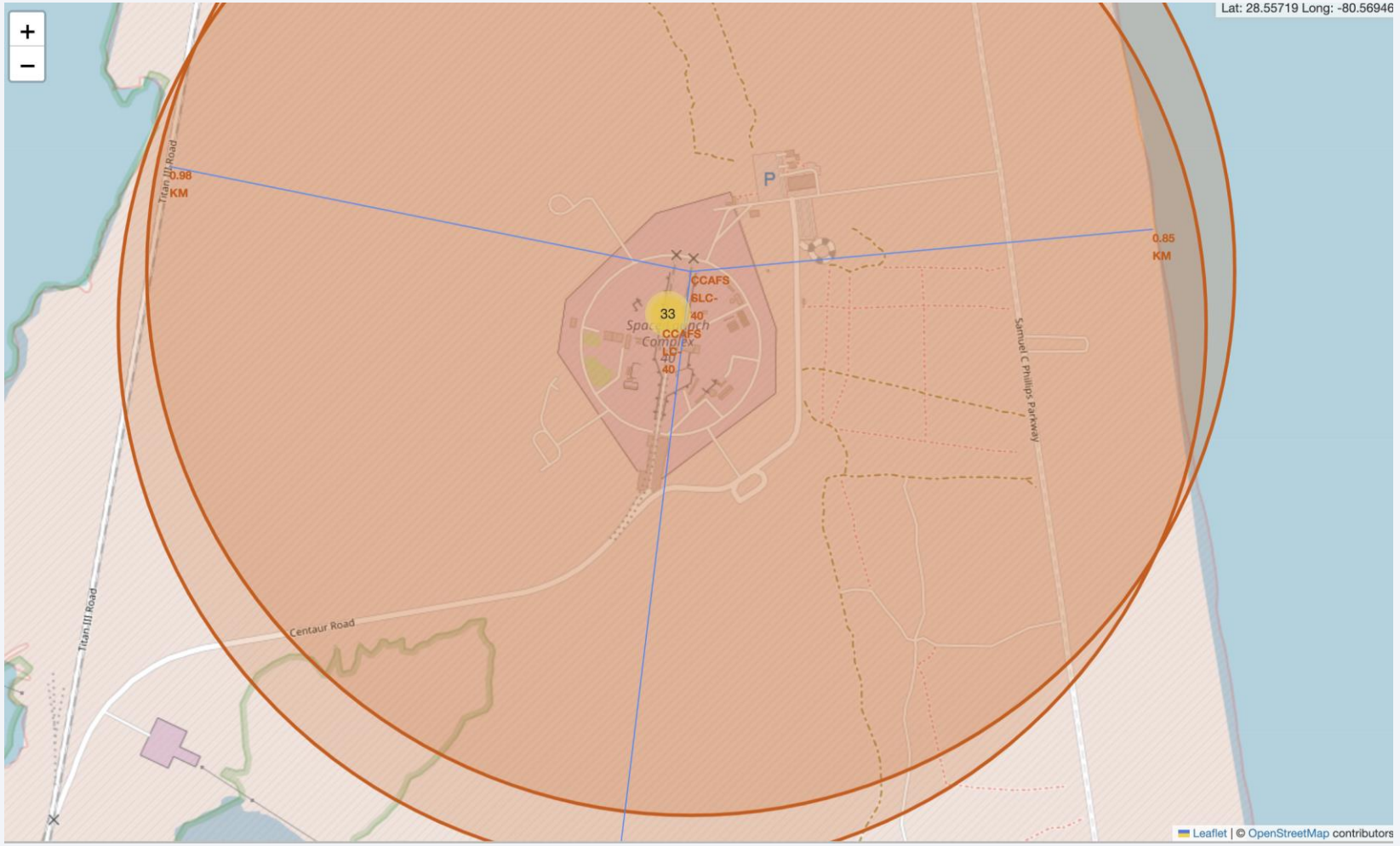
- Launch sites are near the equator (extra velocity boost)
- Launch sites are close to the ocean. (for safety reasons)

Mission outcome per launch site



- Mission outcome depends on launch site.
- For example:
  - CCAFS SLC-40 has 19 failed missions and 7 successful
  - KSC LC-39A has only 3 failed and 10 successful missions

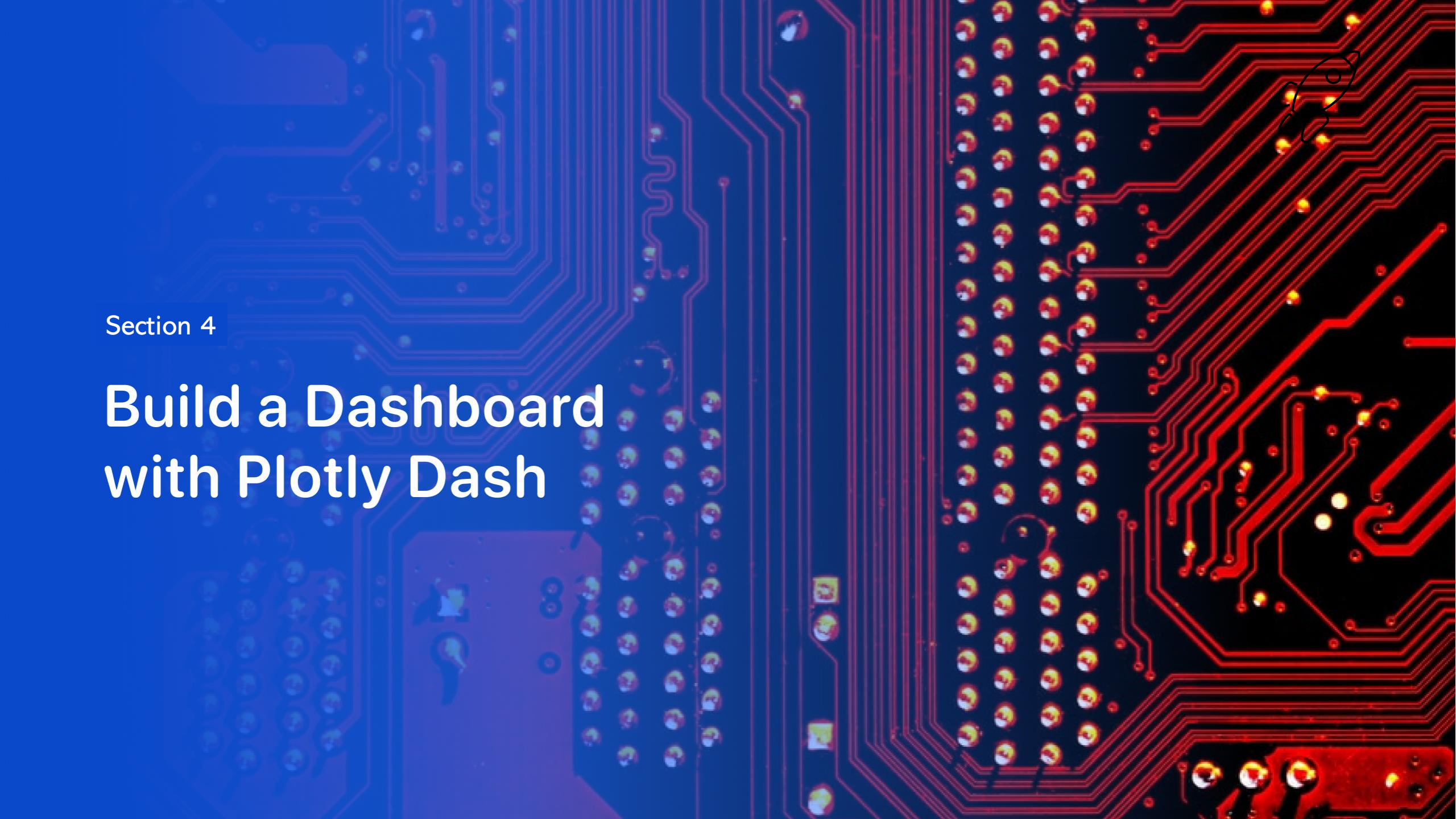
Proximities to launch sites



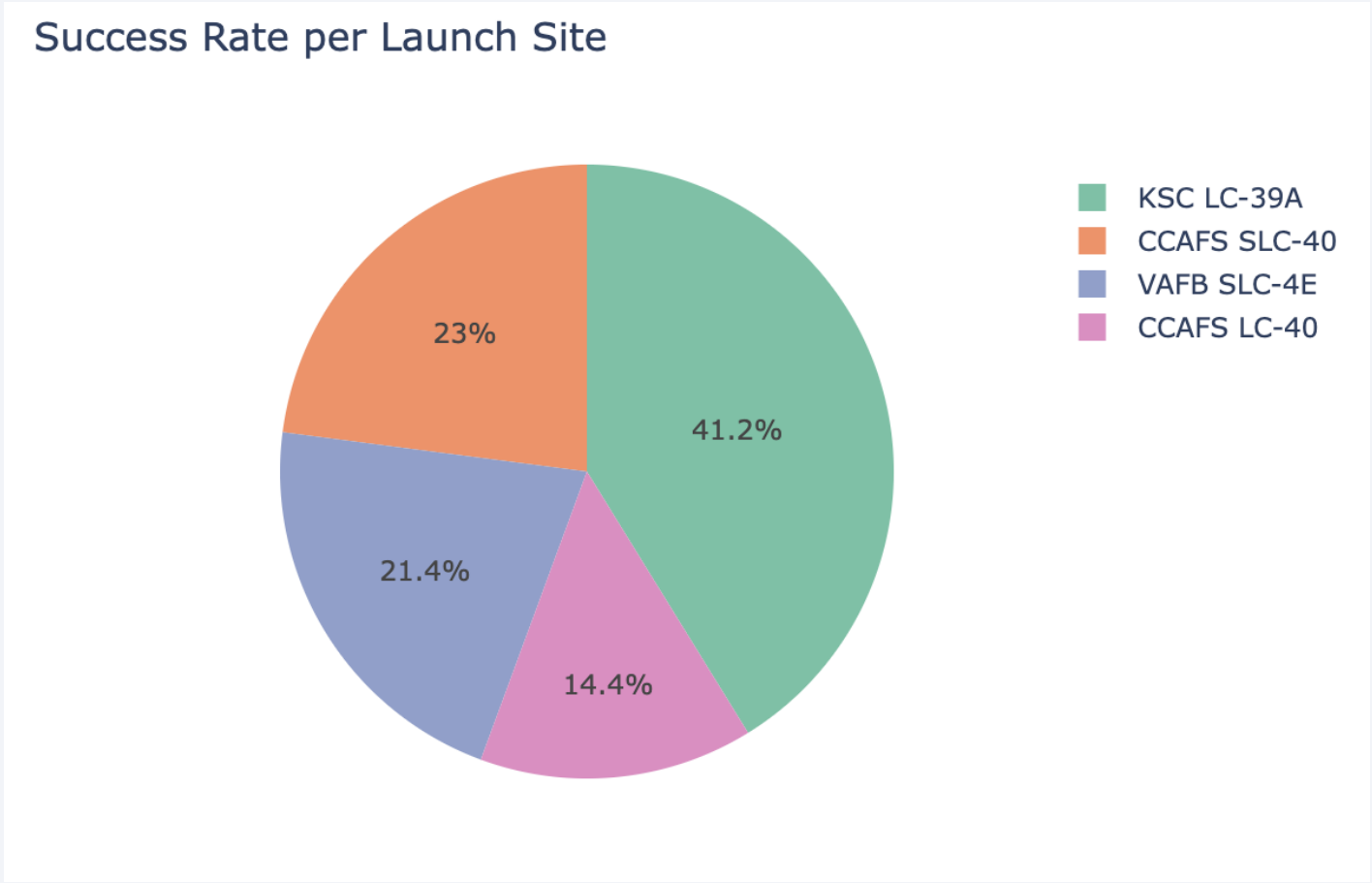
- The launch sites are close to the coast (850m) (use of drone-boats possible)
- Somewhat further to railroads (1km)
- More distance to cities. The closest city for the selected launch-site CCAFS SLC-40 is 17.4 km (because to reduce the risk of falling of a rocket on a city, but not to far because of workers)

Section 4

Build a Dashboard with Plotly Dash

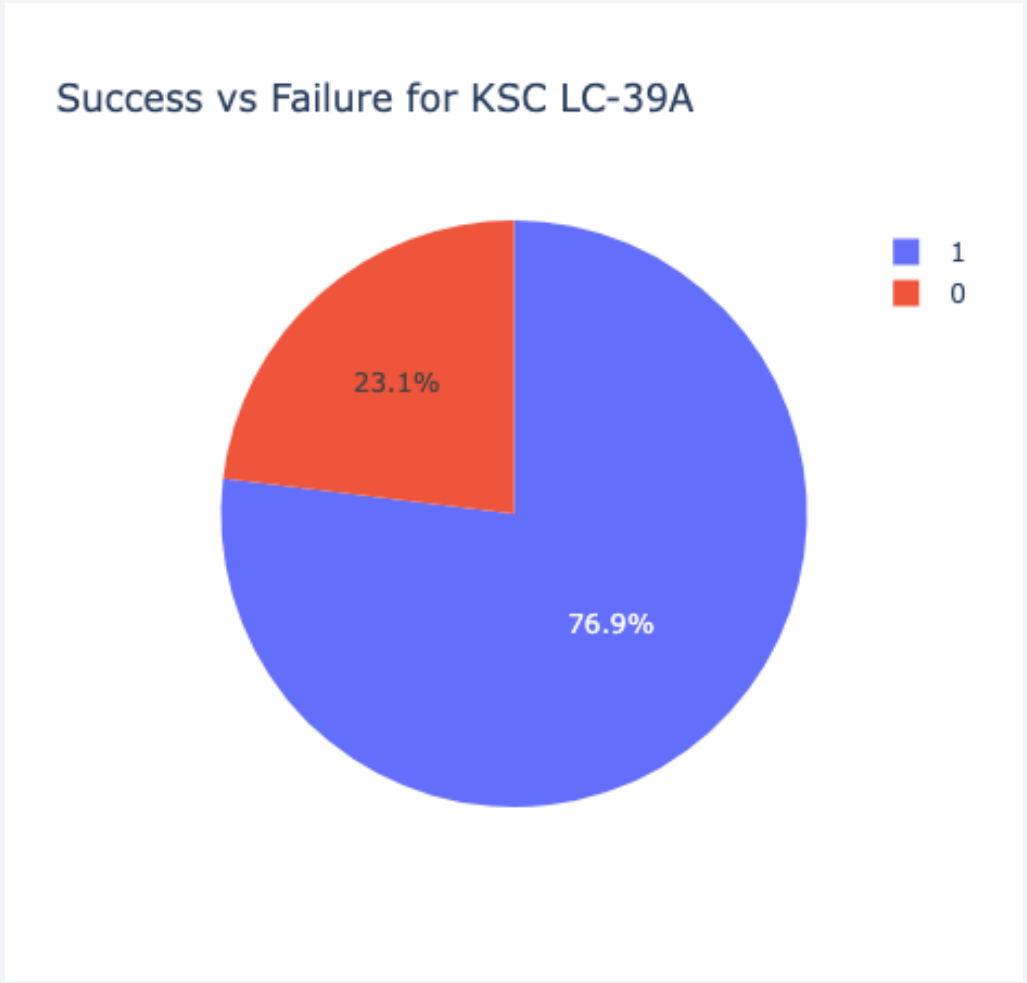


Success Rate per launch site



Launch site KSC LC-39 has the highest success rate

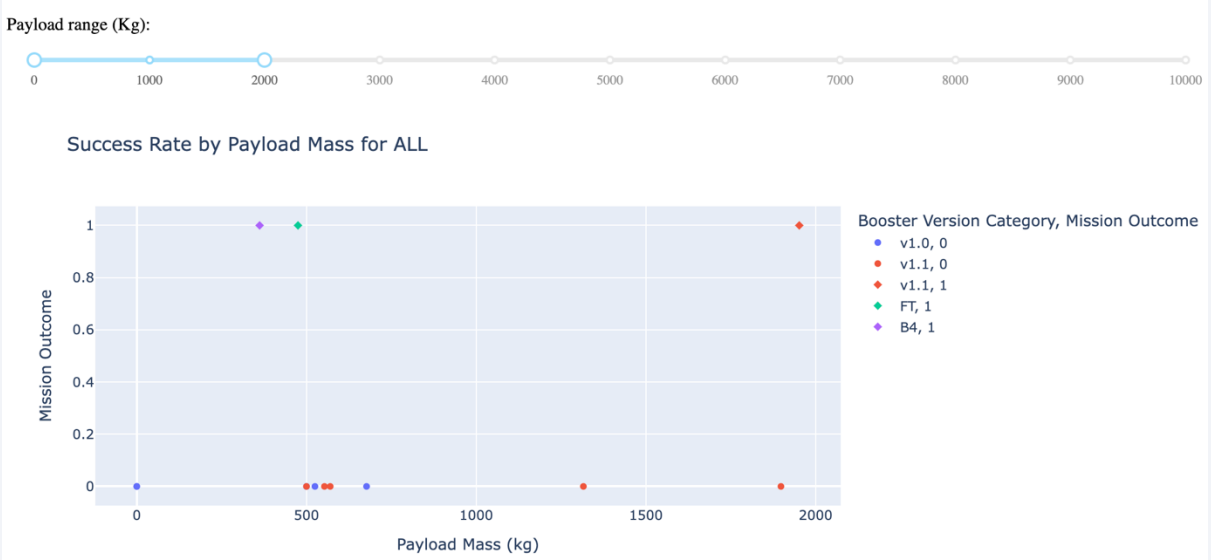
Success vs Failure for launch site KSC LC39A



- 10 missions are successful (76.9%)
- 3 missions failed (23.1 %)

Payload vs success-rate

- Payloads with a low mass (< 2000 kg) have a very low success-rate (see image at top-right)
- Payloads with booster FT,1 have a 100% success-rate (image at bottom-right)

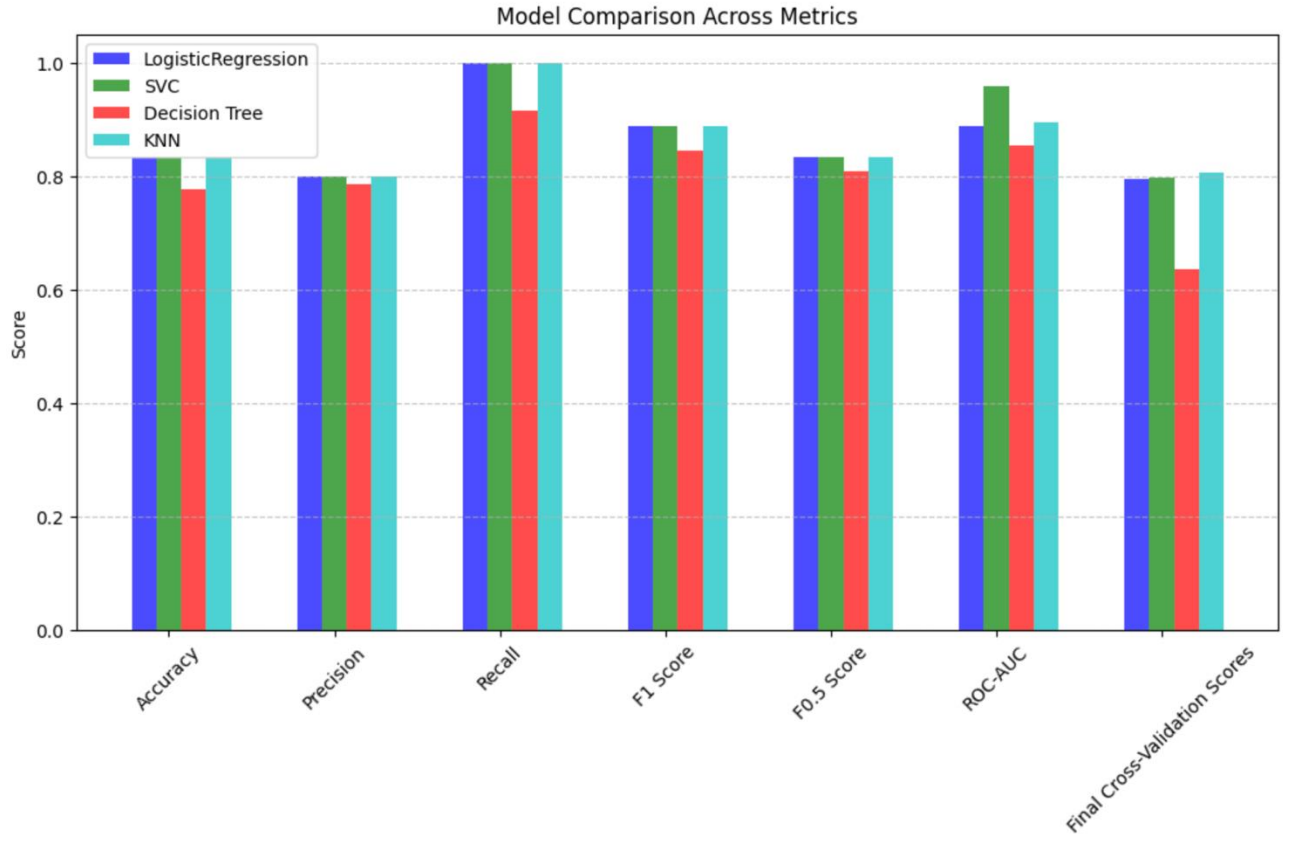


Section 5

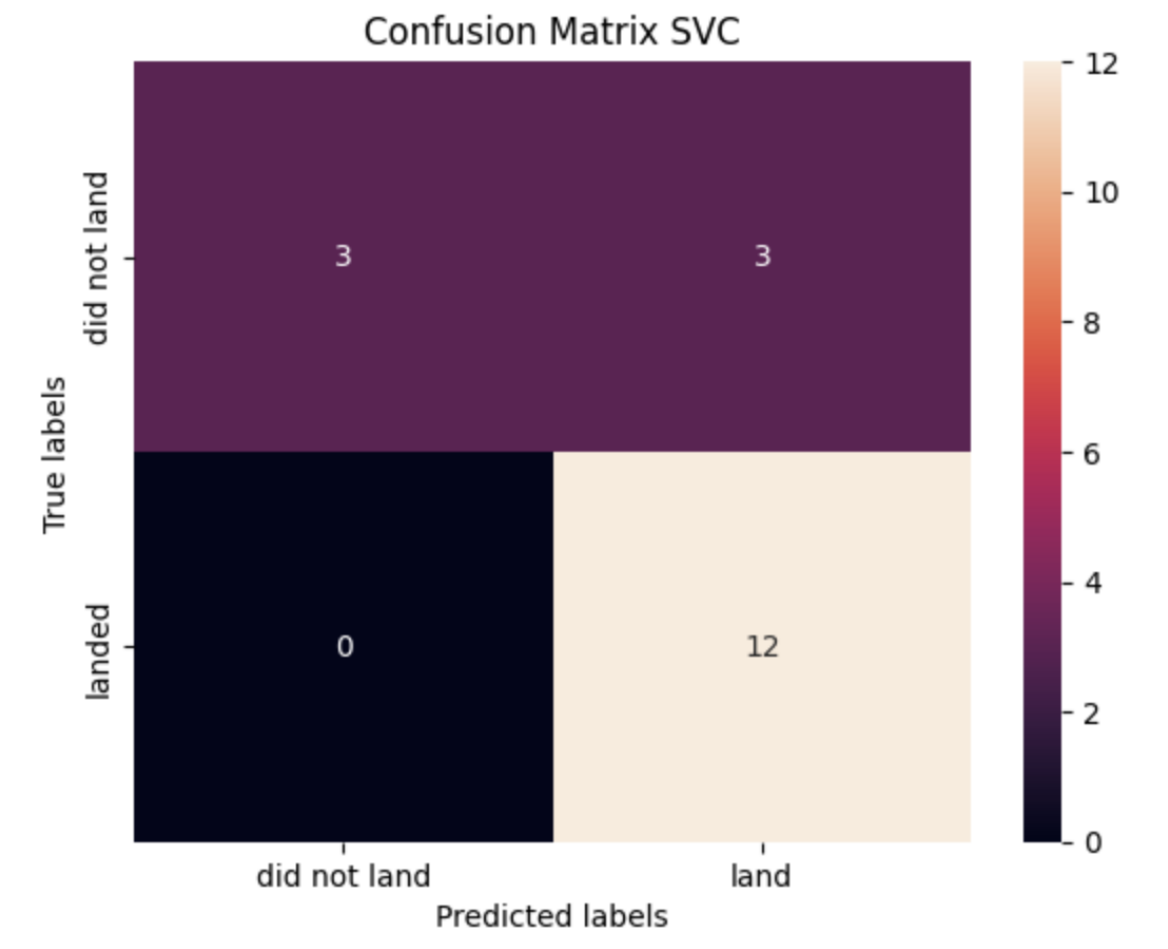
Predictive Analysis (Classification)

Model Comparison

- Decision tree is performing low on all metrics
- For other models are accuracy, precision, recall, F1 and F0.5 scores the same
- SVC has the best ROC-AUC
- KNN is the best on Final Cross validation
- SVC is the overall best model



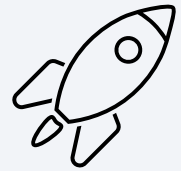
Confusion Matrix for the SVC model



- Top Left: True Negative (TN): 3
- Top Right: False Positive (FP): 3
- Bottom Left: False Negative (FN): 0
- Bottom Right: True Positive (TP): 12

False Positive is the most important because it is costly to lose rockets unintended

Conclusions

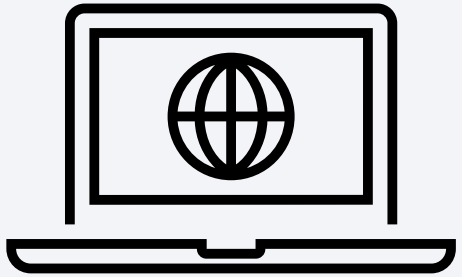


- The likelihood of success improves over time.
- Payloads with minimal mass tend to have lower success rates.
- The booster FT,1 have a success rate of 100%.
- The use of orbit type did change over time.
- The success rate is influenced by the launch site.
- Three out of four models exhibit nearly identical performance metrics.
- The decision tree model performs the least well, showing a low cross-validation score.
- The SVC model stands out as the top performer:
    - It has the highest ROC_AUC at 0.9583.
    - It also shows a strong cross-validation score of 0.7969.

Appendix

All the data is available on the following GitHub repository:

https://github.com/bernard-van-middendorp/applied-data-science-capstone/tree/main



Thank you!