IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Cho Hyunwoo
01/29/2024

Outline

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

Executive Summary

- By collecting data from the “SpaceX” API and Wiki, visualizing with charts, maps, dashboards and exploring the correlations with the Stage 1 success-fullness, several models were created and tested to estimate Falcon 9 Stage 1 success.
- 4 machine learning models where train and tested. SVM showed a fair accuracy of 79.6/100 on training stage and 15/16 on the testing stage. Nevertheless, more data is needed to increase the prediction score.

Introduction

- “SpaceX” has privately funded the development of orbital launch systems that can be reused many times, similar to the re-usability of aircraft
- Project benefits:
  - make spacial travel more accessible and reliable
  - cut costs of the project drastically
- The problem is to identify whether the rocket succeed on the Stage 1

Section 1

Methodology

Methodology

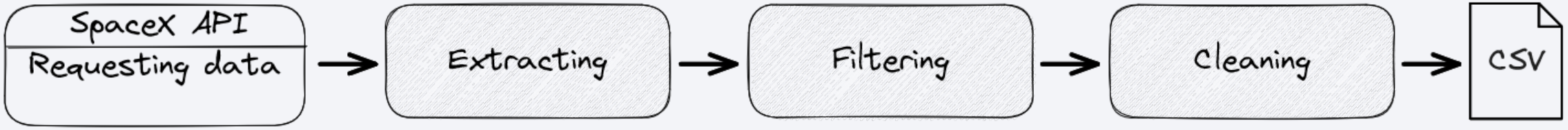
Executive Summary

- Data collection methodology:
  - official “SpaceX” API
- Perform data wrangling
  - features encoding
  - missing values handling
- 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

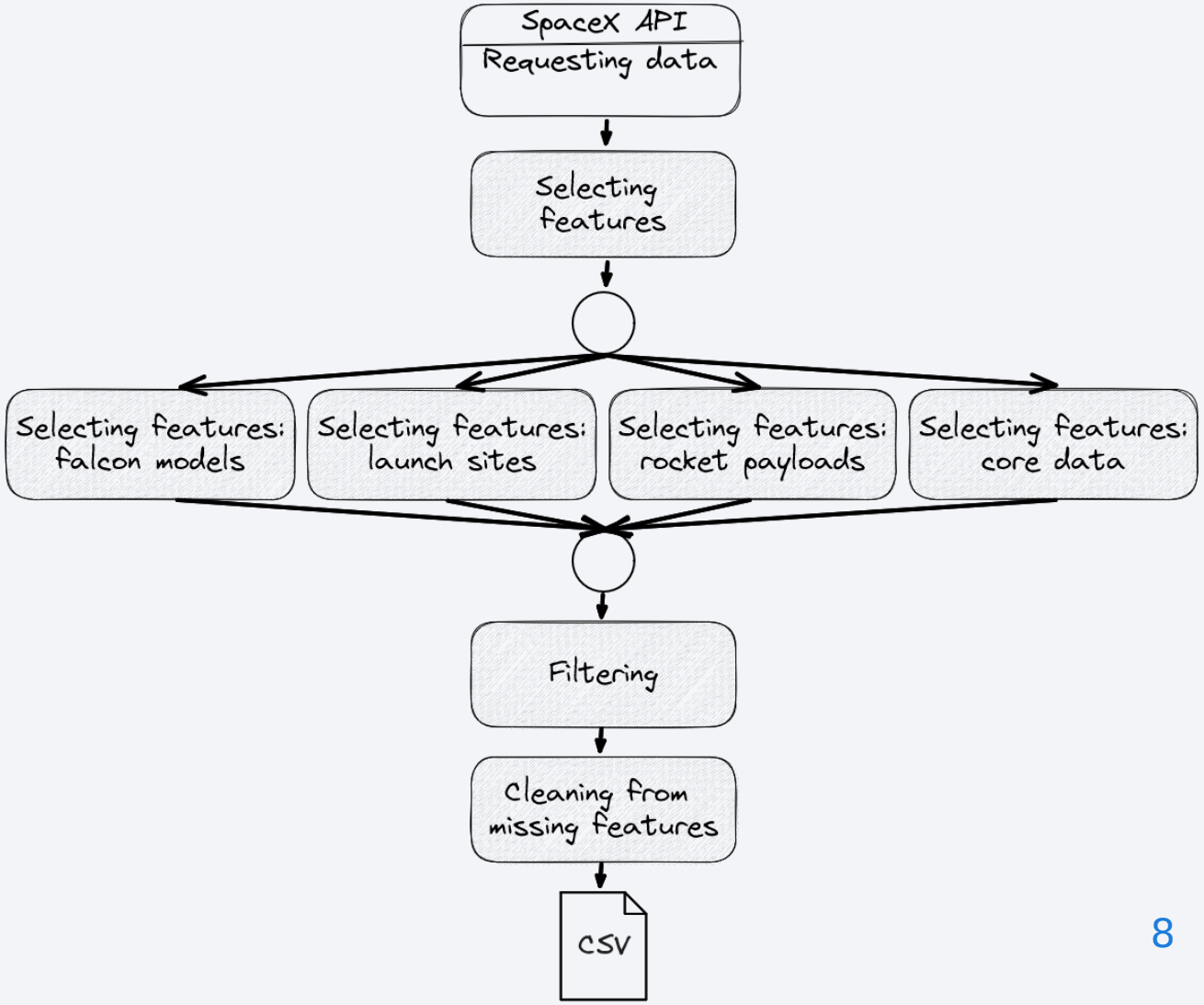
All the data was collected from SpaceX API by requesting them through REST API.

You need to present your data collection process use key phrases and flowcharts



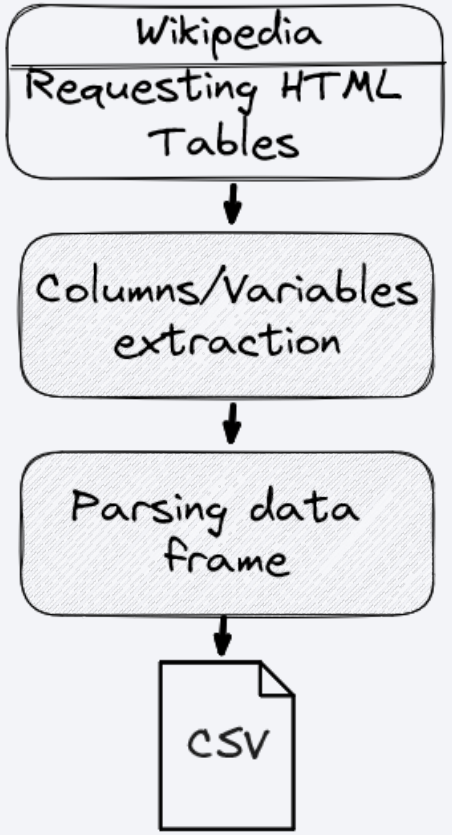
Data Collection – SpaceX API

- Requesting data
- Selecting features
  - Falcon models
  - Launch sites
  - Rocket payloads
  - Core data
- Filtering
- Cleaning from missing features
- Resulting CSV



Data Collection - Scraping

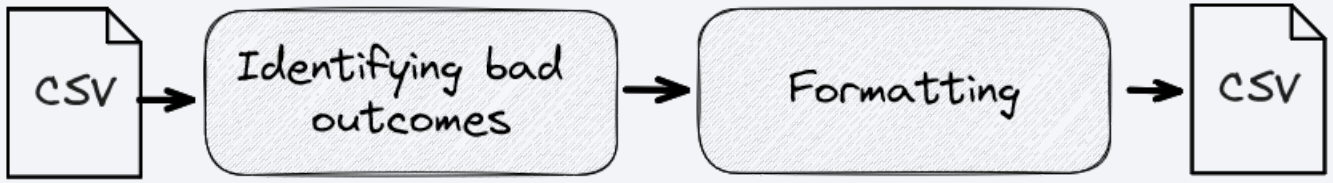
- Wiki request
- Columns/variables extraction
- Parsing data frame
- Resulting CSV



Data Wrangling

The only data wrangling applied is grouping several types of outcomes into 2 essential: success or fail.

- CSV data set from the data collection pipeline
- Identifing bad outcomes
- Formatting
- Resulting CSV



EDA with Data Visualization

The following visualization charts were used:

- Scatter plot (to identify correlating features against success rate)
- Bar chart (to distinguish success rate contributors)
- Line chart (to visualize the average success rate trend line)

EDA with SQL

SQL queries:

- Selecting 5 launch sites the name of which start with CCA
- Selecting total payload mass in kilograms for NASA (CRS)
- Selecting average payload mass for Falcon 9 v1.1 rocket
- Selecting the first successful landing outcome for “SpaceX”
- Selecting the total amount of outcomes for each type of outcome
- Selecting Falcon 9 models with the highest payload mass

Build an Interactive Map with Folium

- All launch sites:
  - Markers: unique launch sites (to see the locations of them)
  - Circles: launch sites areas (to see the approximate impact of launches)
- Distance between launch sites and coastlines:
  - Marker: VAFB SLC-4E launch site (to highlight the launch site)
  - PolyLine: distance between the marker and the closest coastline (to highlight the distance)

Build a Dashboard with Plotly Dash

Dashboard charts (dashboard code):

- Pie chart: displays the proportion of launches for each launch site for all launch sites and the proportion of success launches for a single launch site
- Scatter plot: displays the correlation between payload mass and success rate for each Falcon 9 model

Filters:

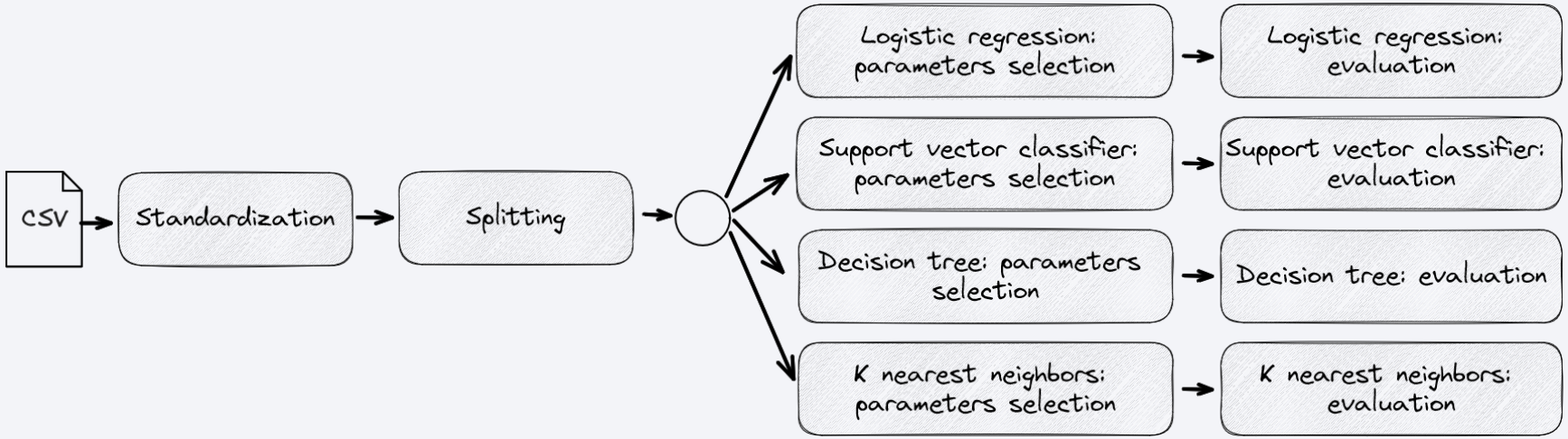
- Launch sites drop-down: specifies a single launch site or all of them
- Payload mass range slider: specifies the payload mass range for scatter plot

Predictive Analysis (Classification)

Models training and evaluation:

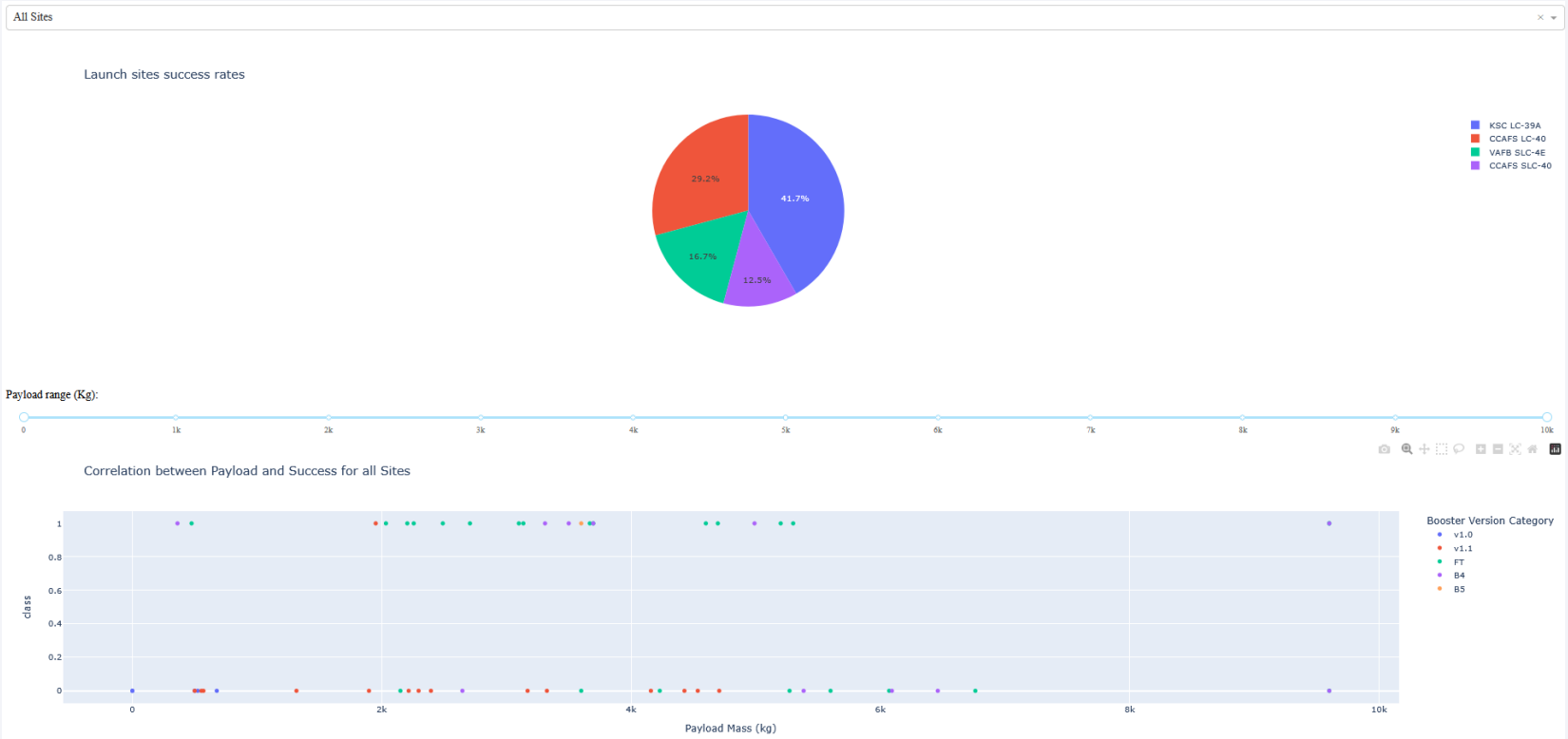
- Loading CSV
- Standardizing data
- Splitting for training and testing

- Logistic regression (parameters selection; evaluation)
- Support vector classifier (parameters selection; evaluation)
- Decision tree (parameters selection; evaluation)
- K nearest neighbors (parameters selection; evaluation)



Results

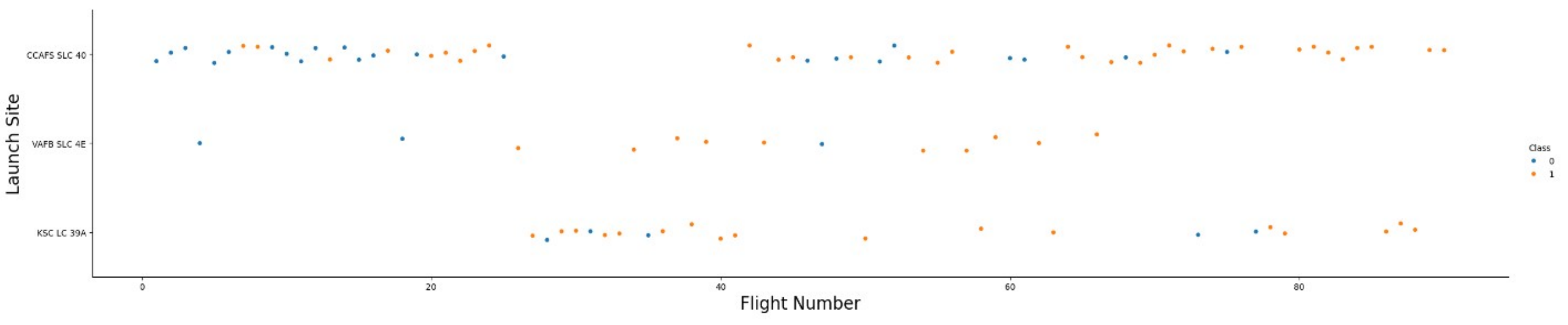
Through data analysis no significant correlations has been found, except the fact that the success rate increases over time. The model with the best training and testing scores (79.6/100 and 15/16) is SVC.



Section 2

Insights drawn from EDA

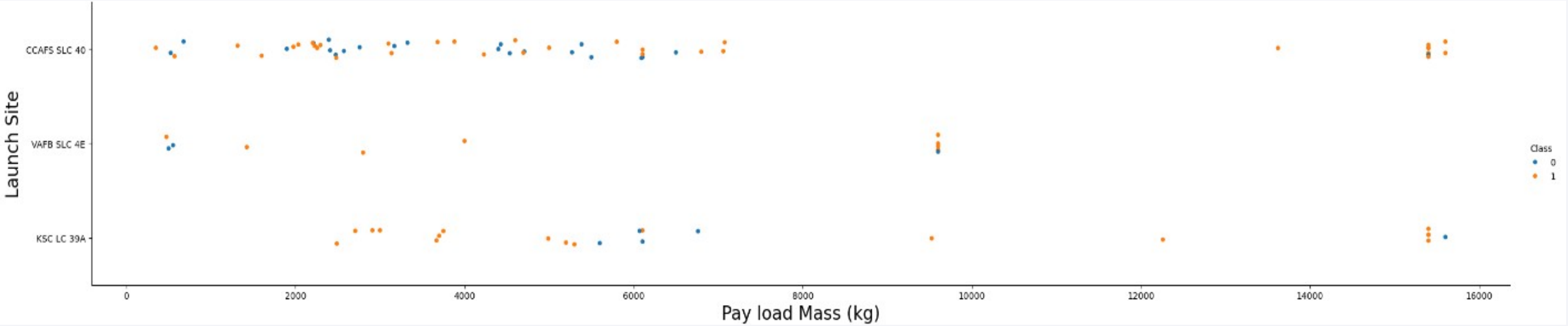
Flight Number vs. Launch Site



Note: blue dots are unsuccessful launches, while orange dots are successful ones.

The amount of successful launches increased over time, where the rapid increase happened after the 18-20th launch. It also seems that CCAFS SLC-40 was the most common launch site basing on its amount of observations compared to other launch sites.

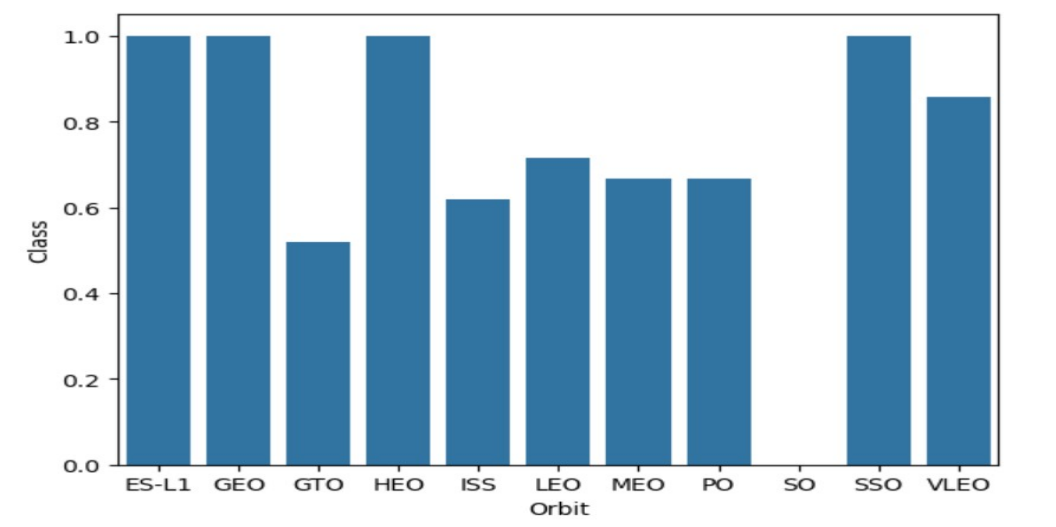
Payload vs. Launch Site



Note: blue dots are unsuccessful launches, while orange dots are successful ones.

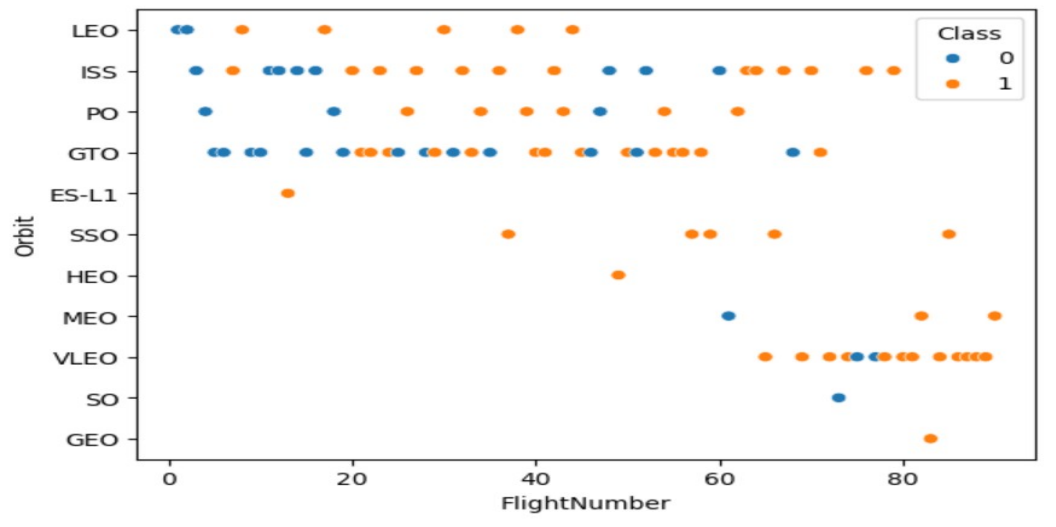
The payload mass varies depending on a launch site. Generally, most Falcon 9 models that were launched has a payload mass in the range of 0 – 7500 kilograms. The launched rockets with the payload mass greater than 8500 kilograms have a greater success rate in comparison to the previously mentioned ranged, but less volume on the other hand.

Success Rate vs. Orbit Type



The SO orbit seems has no successful Stage 1 executions. The ES-L1, GEO, HEO and SSO has the perfect Stage 1 execution. The other types of orbits, including SO, contribute to the failed exectuions.

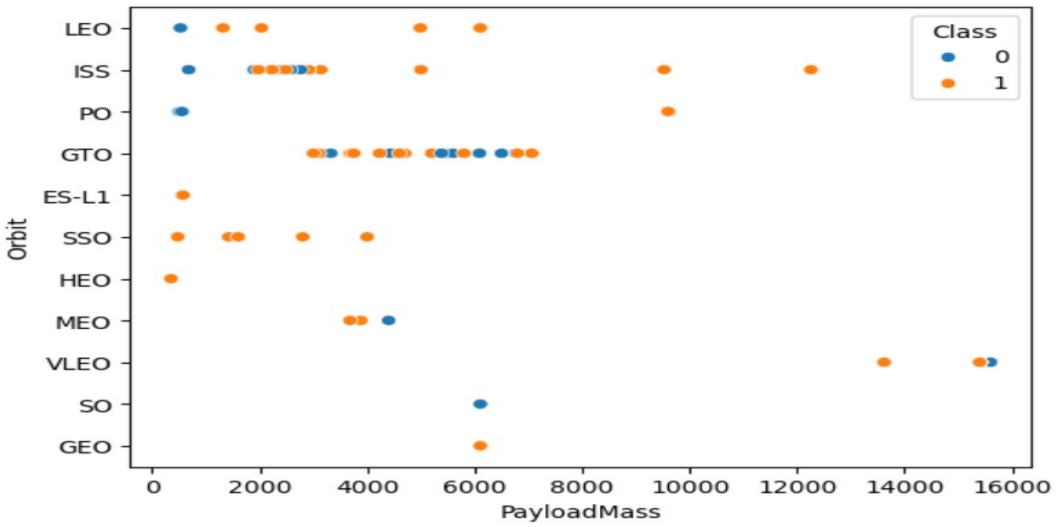
Flight Number vs. Orbit Type



Note: blue dots are unsuccessful launches, while orange dots are successful ones.

The ES-L1, SO, GEO and HEO have each a single Stage 1 execution attempts, which is not enough to judge about them. The MEO orbit should be excluded as well, considering the amount of launches. The GTO, ISS, VLEO and PO are considered to be the most common orbits used. The preference shifted from the ISS, LEO, PO and GTO torwards VLEO over the amount of flight numbers. Correlation is evident.

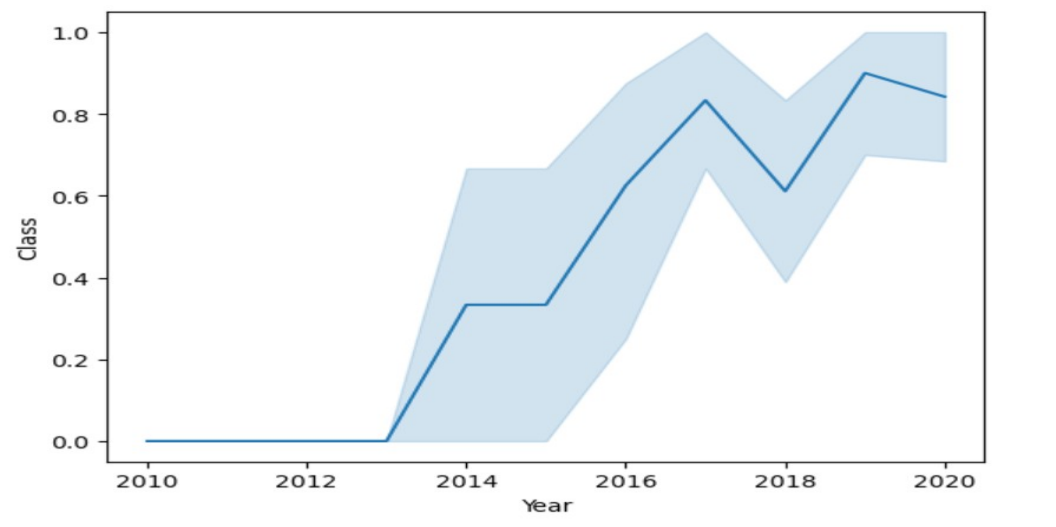
Payload vs. Orbit Type



Note: blue dots are unsuccessful launches, while orange dots are successful ones.

The Falcon 9 models targeted to VLEO had the highest payload masses. The correlation is evident.

Launch Success Yearly Trend



The success rate increases over the years. It increased from 35% in 2014 to 80% in 2020.

All Launch Site Names

```
helper_request("SELECT DISTINCT Launch_Site FROM SPACEXTABLE")
```

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

All unique launch site titles are presented. Launch sites CCAFS LC-40 and CCAFS SLC-40 are the same launch sites, but with the different names, hence presenting 3 unique launch sites overall.

Launch Site Names Begin with 'CCA'

```
helper_request("SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5")
```

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

5 launch sites with with the names CCAFS LC-40 and CCAFS SLC-40 are presented.

Total Payload Mass

```
helper_request("SELECT SUM(PAYLOAD_MASS__KG_) AS NASA_CRS_Total_Payload_Mass_KG FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)' GROUP BY Customer")
```

```
-----
45596
-----
```

The total amount of payload mass in kilograms with, as the role of a customer, NASA (CRS). CRS stands for Commercial Resupply Services, whose payloads were mainly sent to ISS orbit.

Average Payload Mass by F9 v1.1

```
helper_request("SELECT AVG(PAYLOAD_MASS__KG_) AS Booster_F9_v11_Average_Payload_Mass_KG FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1' GROUP BY Booster_Version")
```

```
------
2928.4
------
```

The average payload mass in kilograms loaded for Falcon 9 models version 1.1, which is common in comparison to other loads.

First Successful Ground Landing Date

```
helper_request("SELECT MIN(Date) as Date_Of_First_Succesful_Landing_Outcome FROM SPACEXTABLE WHERE Mission_Outcome = 'Success'")
```

```
----------
2010-06-04
----------
```

The date of the first successful landing outcome is presented.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
helper_request("SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000")
```

```
--------------
F9 FT B1022
F9 FT B1026
F9 FT  B1021.2
F9 FT  B1031.2
--------------
```

4 distinct Falcon 9 models with the successful drone ship and payload mass between 4000 and 6000 kilograms are presented.

Total Number of Successful and Failure Mission Outcomes

```
helper_request("SELECT COUNT(*) AS Total, Mission_Outcome FROM SPACEXTABLE GROUP BY Mission_Outcome")
```

```
--  -------------------------------
 1  Failure (in flight)
98  Success
 1  Success
 1  Success (payload status unclear)
--  -------------------------------
```

The quantities of cases for each type of mission outcome are presented. The success rate for mission outcomes is high.

Boosters Carried Maximum Payload

```
helper_request("SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)")
```

```
-------------
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
-------------
```

12 model versions of Falcon 9 with the maximum amount of payload mass loaded are presented.

2015 Launch Records

```
helper_request("SELECT SUBSTR(Date, 6, 2) AS MONTH, Landing_Outcome, Booster_Version, Launch_Site \
FROM SPACEXTABLE \
WHERE Landing_Outcome = 'Failure (drone ship)' AND SUBSTR(Date, 0, 5) = '2015'")
```

```
--  --------------------  -------------  -----------
01  Failure (drone ship)  F9 v1.1 B1012  CCAFS LC-40
04  Failure (drone ship)  F9 v1.1 B1015  CCAFS LC-40
--  --------------------  -------------  -----------
```

Launch records that include the Falcon 9 models failing drone ship from 2015 are presented.

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

```
helper_request("SELECT COUNT(Landing_Outcome) AS Total, Landing_Outcome FROM SPACEXTABLE \
WHERE Landing_Outcome IN ('Failure (drone ship)', 'Success (ground pad)') AND Date BETWEEN '2010-06-04' AND '2017-03-20' ORDER BY Date DESC")
```

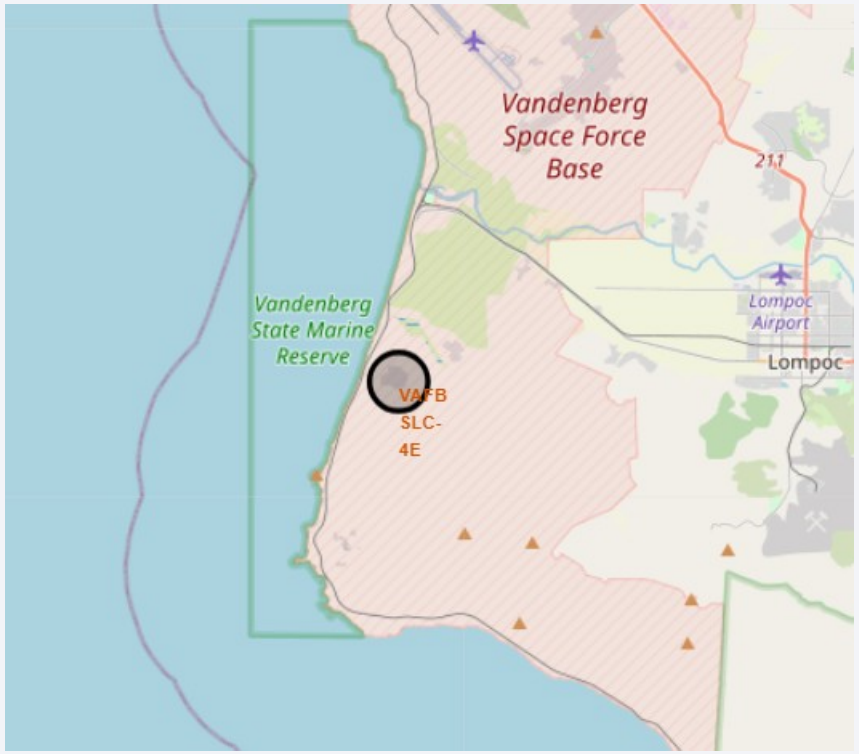
```
-  --------------------
8  Failure (drone ship)
-  --------------------
```

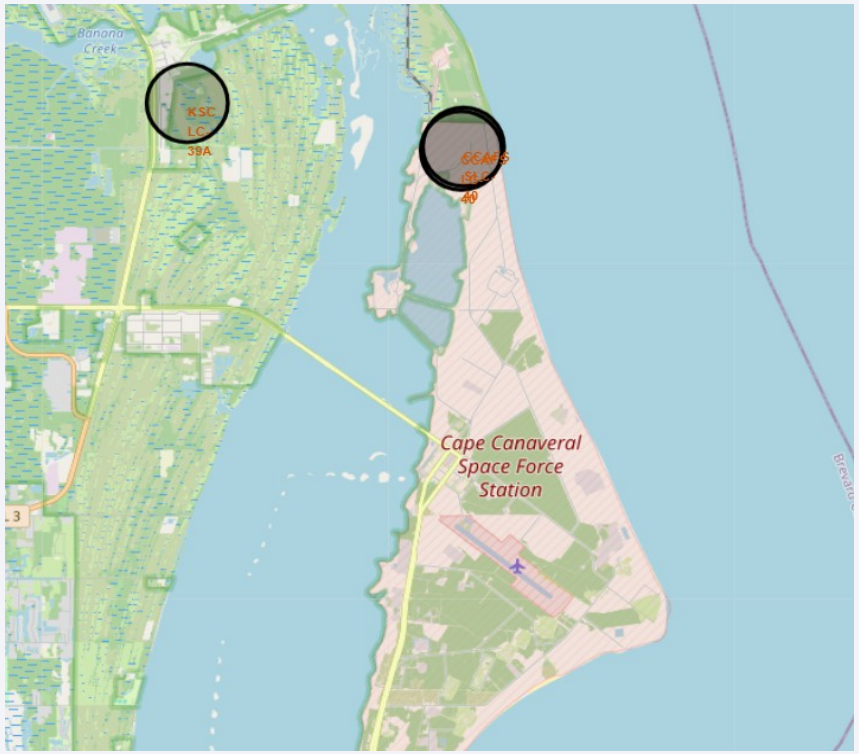
The lading outcomes consisting of cases with failure drone ship and successful ground pad between 4th June 2010 and 20th March 2017 are presented. No successful ground pad observed during this period.

Section 3

Launch Sites Proximities Analysis

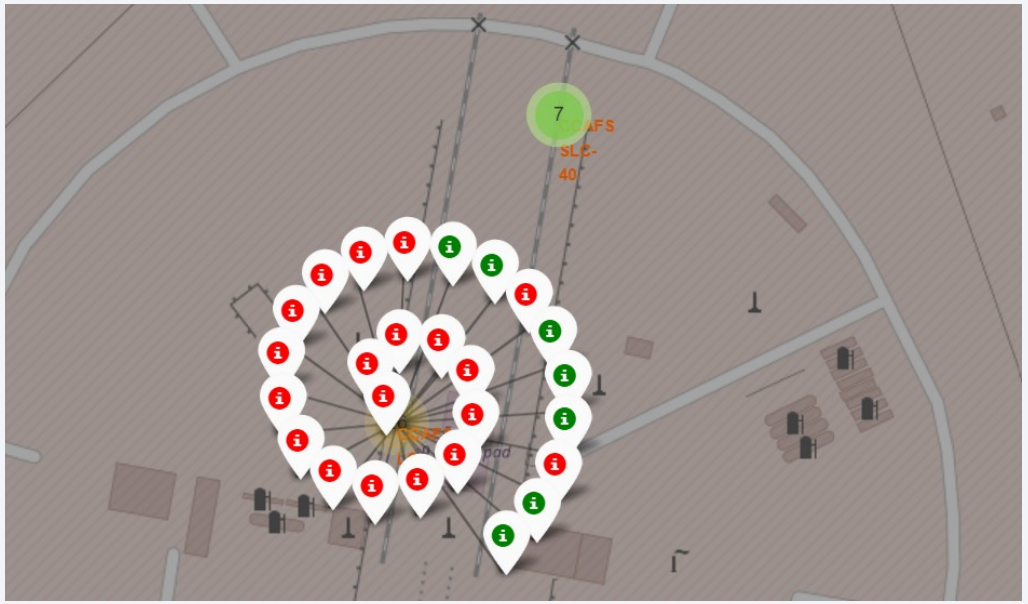
Launch Sites on a Map





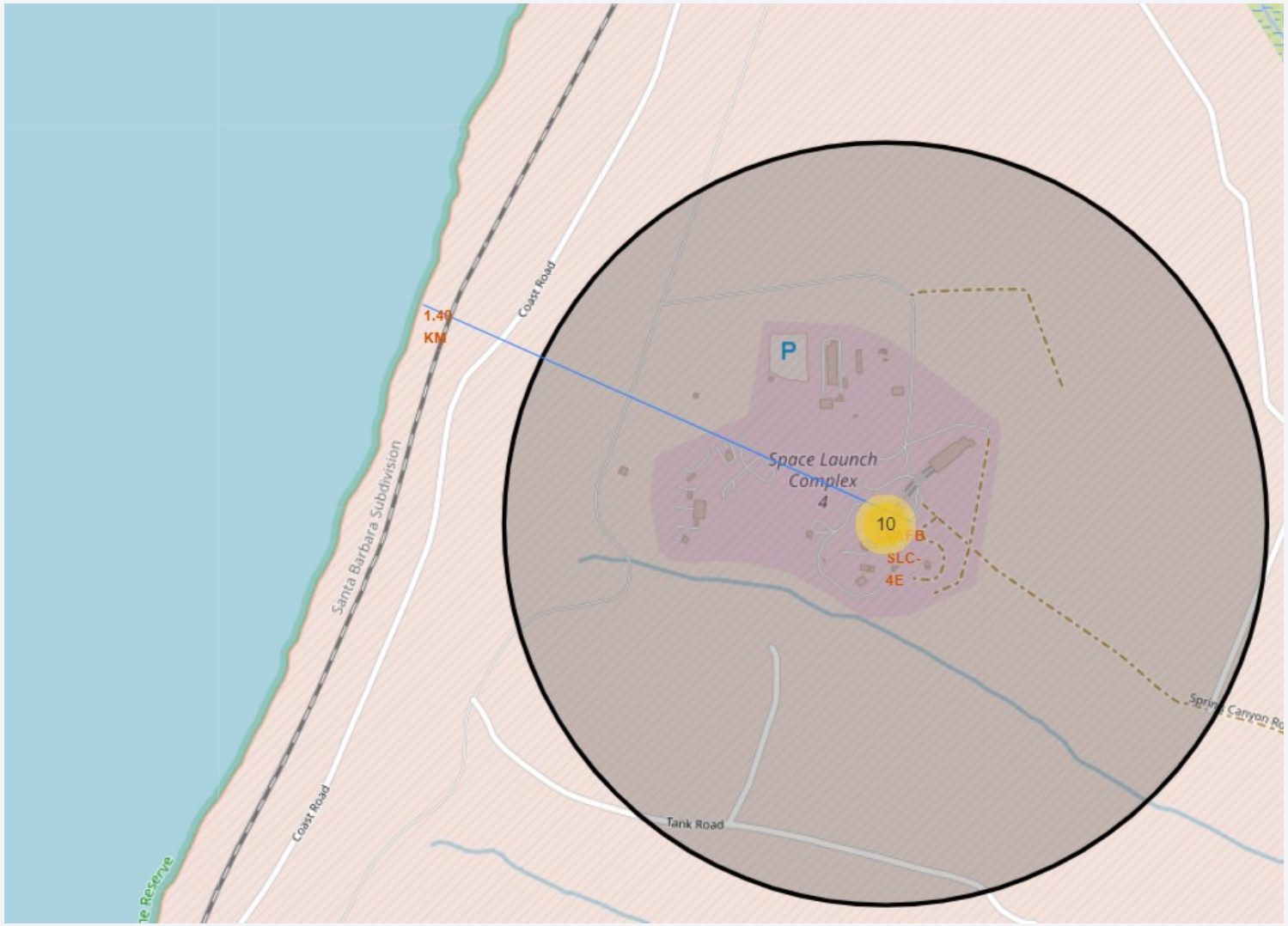
All the launch sites located alongside the shores and the airports for the purpose of safety.

Successful and Failed Launches for Each Site



Note: red markers present the failed launches, while green ones the successful launches.

Launch Sites in contrast of Its Proximity

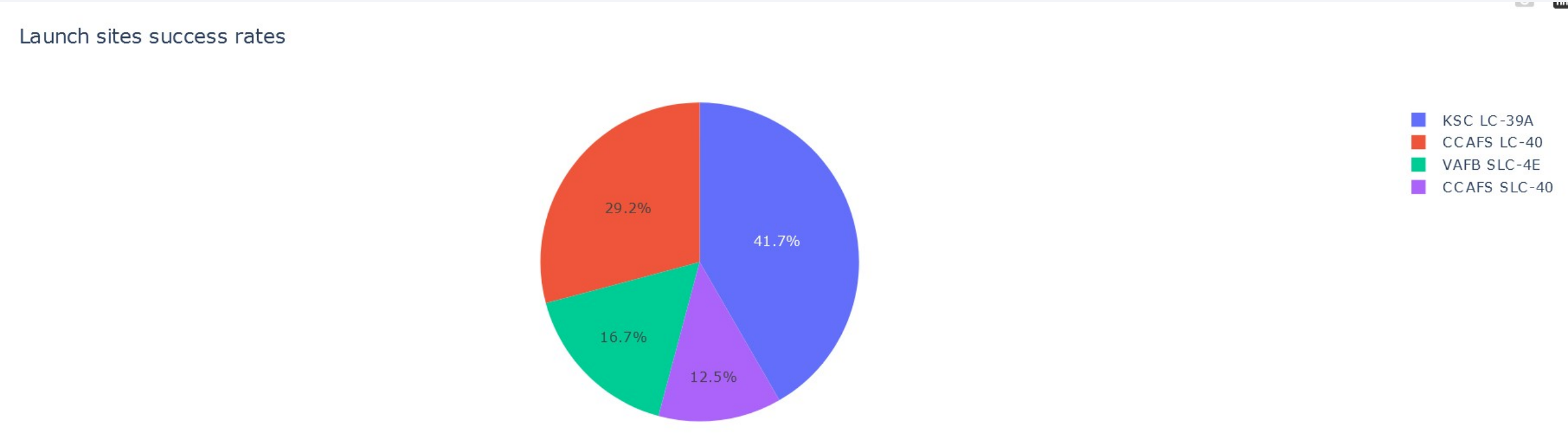


The VAFB SLC-4E launch site is about 1.5 kilometers away from the coastline.

Section 4

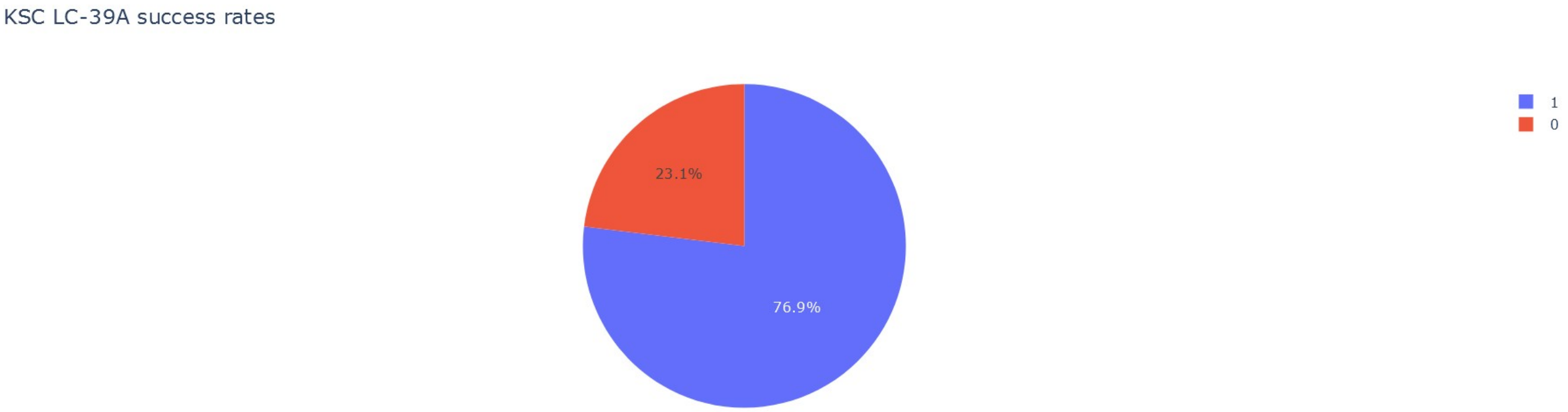
Build a Dashboard with Plotly Dash

Launch Sites Success Rates



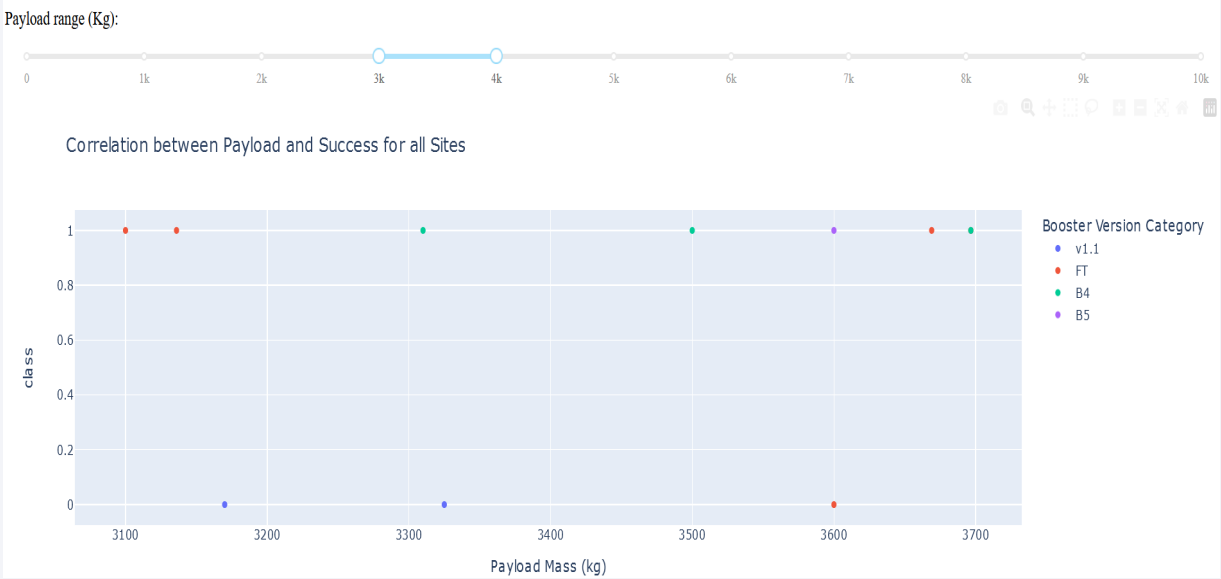
Launch sites CCAFS LC-40 with the CCAFS SLC-40 and KSC LC-39A has the same amount of successful stages 1, while VAFB SLC-4E has the least amount.

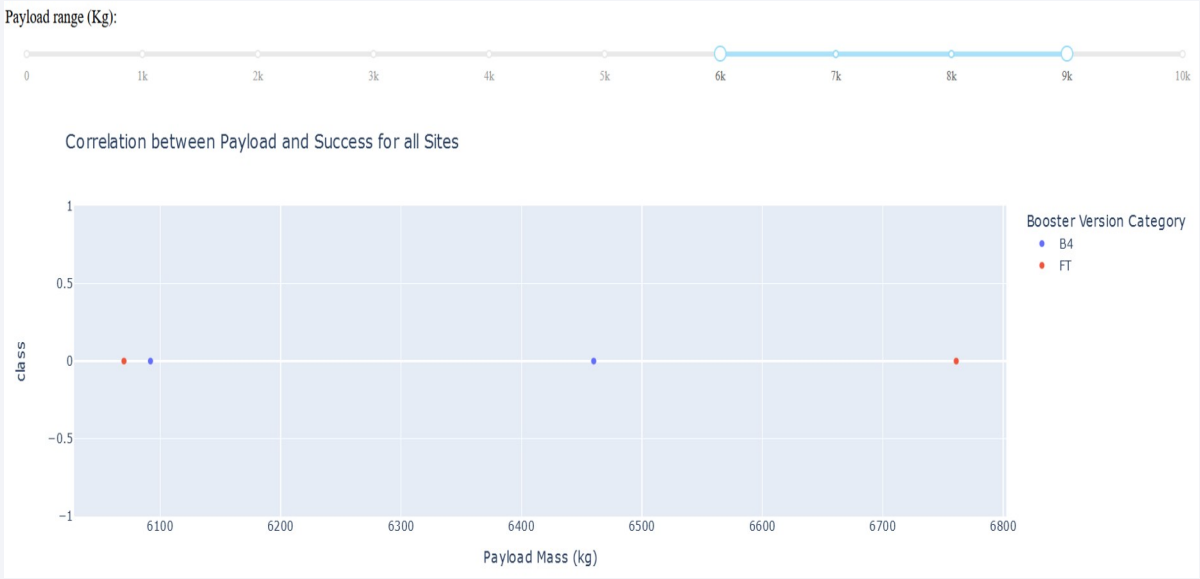
KSC LC-39A Success Rates



KSC LC-39A has the greatest success rates in contrast to other launch sites.

Correlation between Payload and Success for All Sites



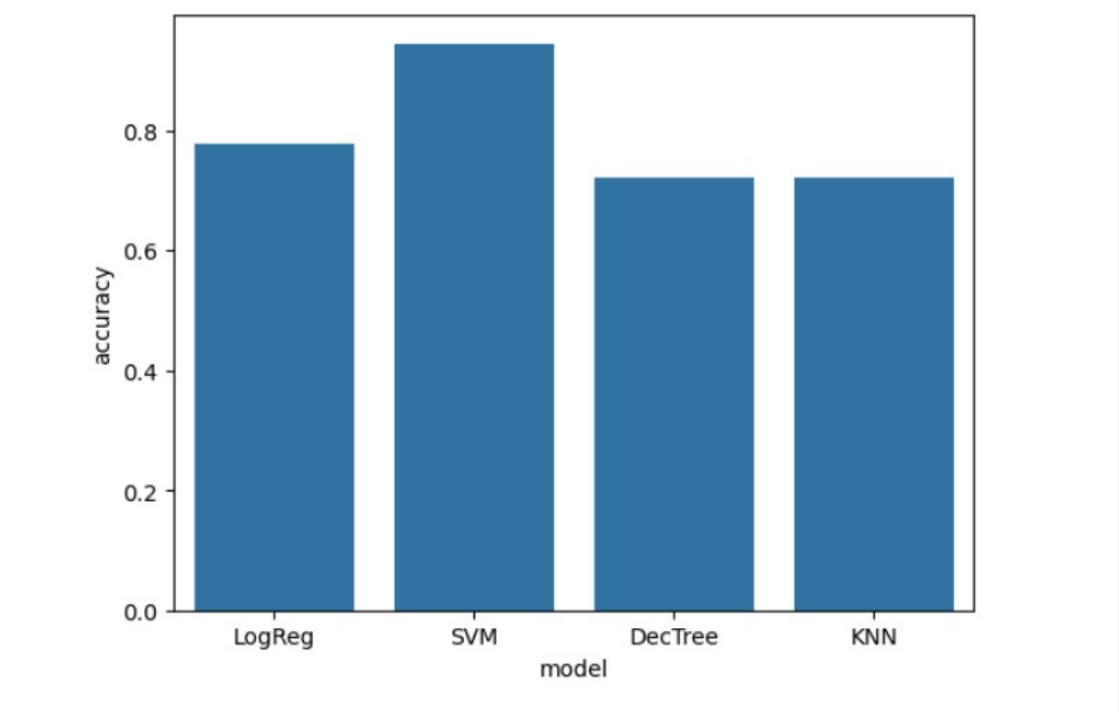


The payload mass ranges from the 3000 to 4000 kg and from the 6000 to 9000 kg have the largest success rates.

Section 5

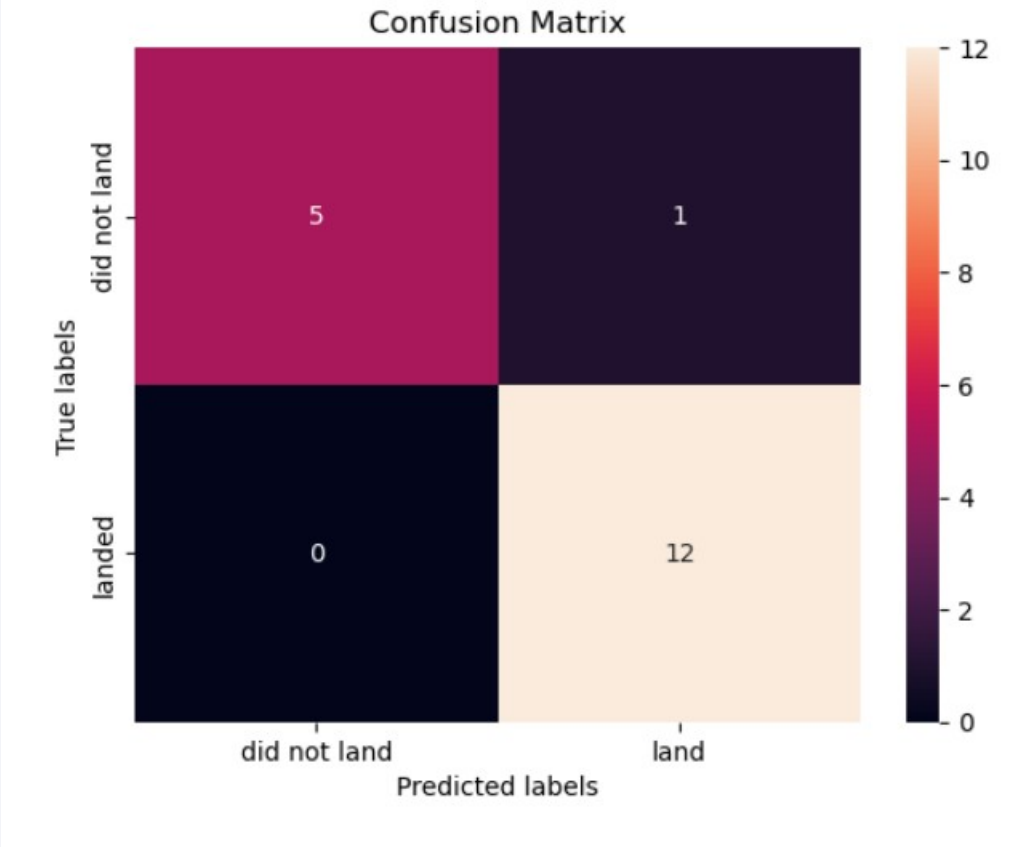
Predictive Analysis (Classification)

Classification Accuracy



Support vector classifier performs the best both for training and testing data sets with relative accuracies

Confusion Matrix



The confusion matrix of the Support vector classifier is presented. It estimated correctly 12 successful and 5 failed cases.

Conclusions

- The data used in the project were taken from SpaceX API and Wikipedia
- Visualization and dashboards were presented
- The model was created with the highest accuracy of 79.6% on training and 94.4% on testing data set
- More data is needed to draw more reliable insights

Appendix

- All the datasets used in the lab were taken from the IBM cloud: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/
- Repository: https://github.com/Dimonoy/IBM-Course-Projects
- Notebooks: https://github.com/Dimonoy/IBM-Course-Projects/tree/master/Final-project

Thank you!