IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Lopez V., Jesus A.
11/20/2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection with SpaceX API
  - Data Collection using Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Maps with Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis results
  - Interactive Dashboard analytics in screenshots
  - Predicting and Classification models Evaluation and Analysis

Introduction

In the currents times commercial space travel is an option that is very close for regular people to experience, so companies are looking to make that possible and affordable for everyone. SpaceY, a rocket company funded by Allon Musk, would like to compete with SpaceX for the most successful company in space flights and in sending manned missions to space. One reason SpaceX can do this is the rocket launches are relatively inexpensive, since of the two stages a rocket launch have they can recover and reuse the first stage for the Falcon 9 rocket launcher. So for this SpaceY project the objective is to determine the cost of this each launch using information gathered about SpaceX launches, and since not every 1st stage could not land it is needed to determine the rate of success for every landing.

Section 1

Methodology

Methodology

Executive Summary

SpaceX launch data will be gathered from an API, specifically the SpaceX REST API, some other information will be gathered with web scrapping. Then with the data gathered, it will be wrangled to extract landing outcome information that will be converted into two classes: 0 for not landing outcome, and 1 for landing outcome. The next step is to examine using Exploratory Data Analysis (EDA) with graphs visualization and SQL queries, if the data could be used to determine if the Falcon 9 first stage will land. If yes it will be performed and interactive visual analytics of the data using Folium and Plotly Dash. Next an implementation of predictive analysis using classification models will be performed to evaluate the best machine learning model to predict the successful landing of the Falcon 9 1st stage.

Data Collection

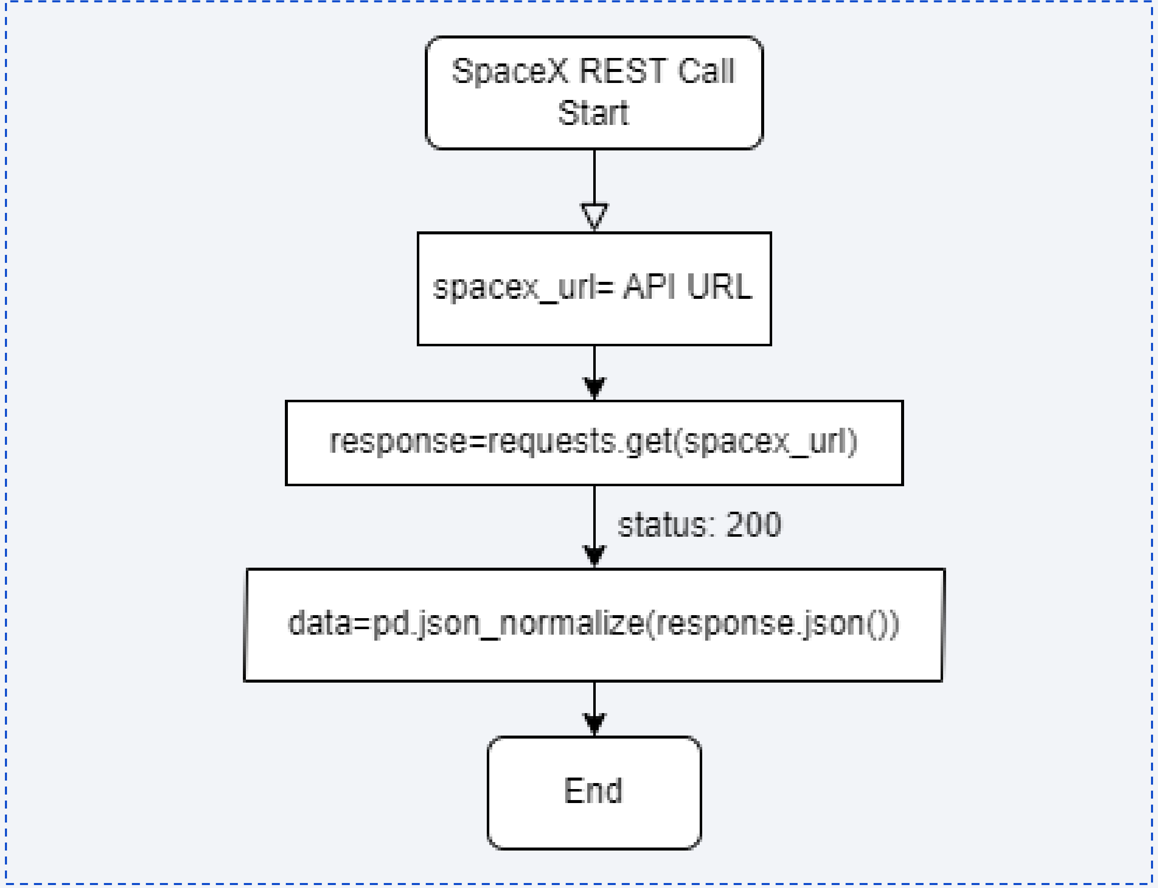
SpaceX launch data will be gathered from an API, specifically the SpaceX REST API. This API will give data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome. A get request will be performed using the requests library to obtain the launch data, which will be used to get the data from the API. The response will be in the form of a JSON, specifically a list of JSON objects that will be later converted into a data frame.

Another data source for obtaining Falcon 9 Launch data is web scraping the related Wikipedia page. The Python BeautifulSoup package will be used for web scrapping some HTML tables that contain valuable Falcon 9 launch records.

Data Collection – SpaceX API

The data obtained with the SpaceX REST API is in a list of JSON objects, each being a rocket launch, were stored in a pandas dataframe from which meaningful information were extracted to create a new dataframe with the data that will be used in the project.

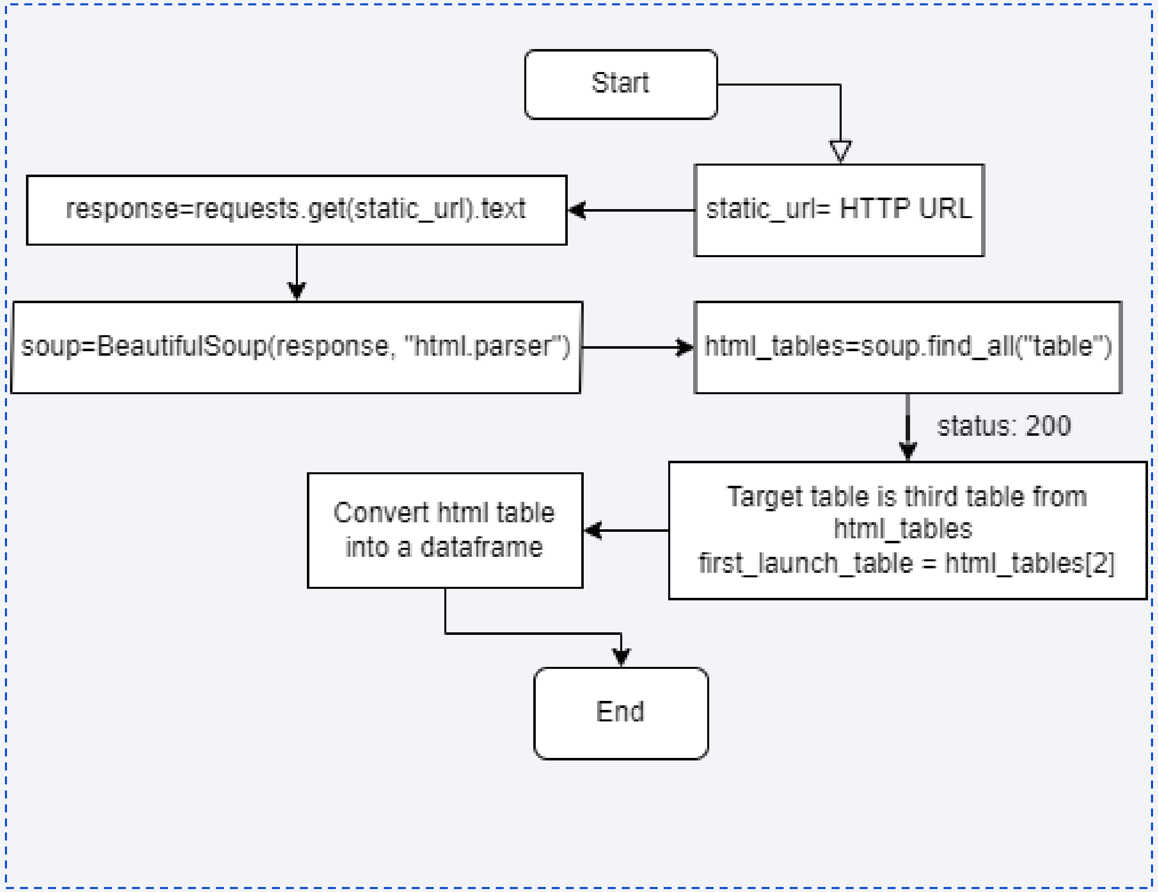
- [GitHub URL]



Data Collection - Scraping

From the HTML table "List of Falcon 9 and Falcon Heavy launches" from a Wikipage, all data on the table were scrapped as a text data BeautifulSoup object, to be stored in a dataframe.

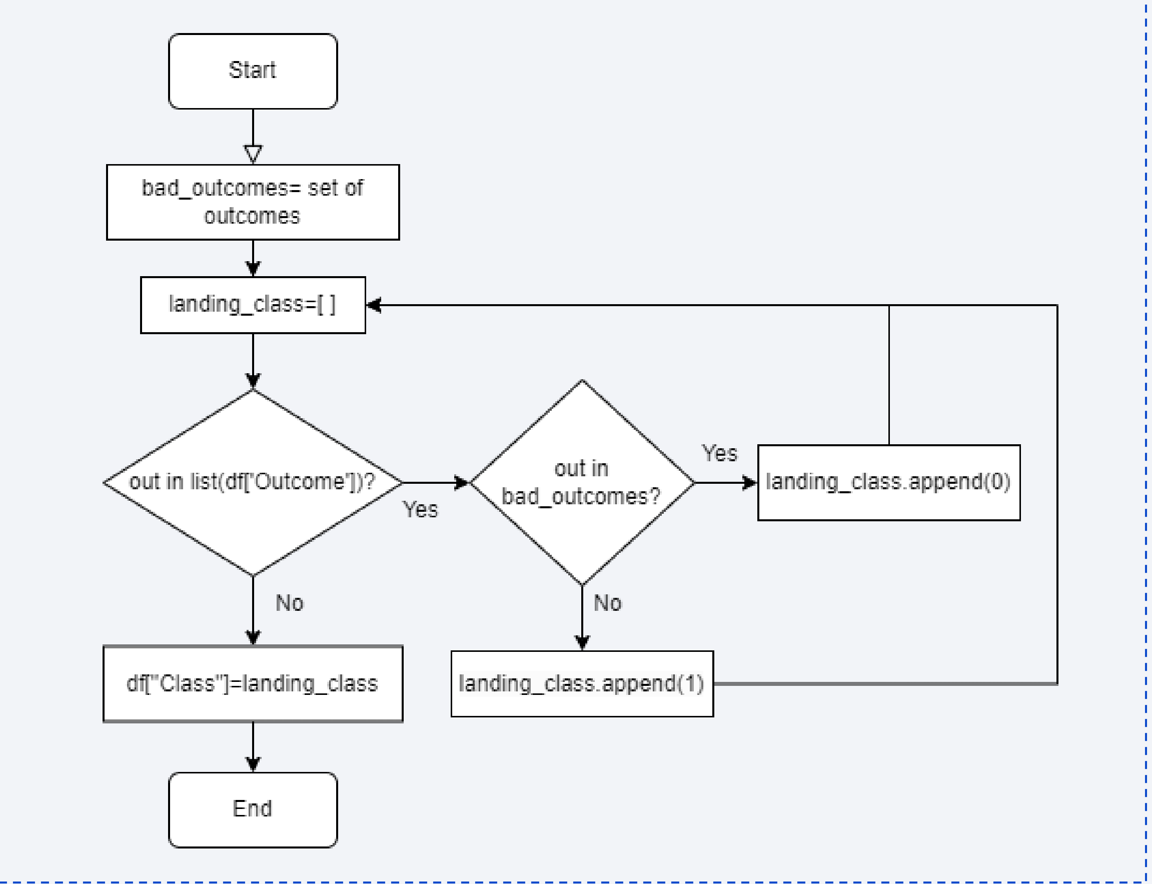
- [GitHub URL]()



Data Wrangling

From the “Outcome” column, is it specified the different landing outcomes of every Falcon 9 mission, so from this information it was created a new column called “class” as a classification variable. If the value is zero, the first stage did not land successfully; one means the first stage landed successfully. This class column was added to the project dataframe.

- [GitHub URL]



EDA with Data Visualization

- For the Exploratory Data Analysis stage of this project a series plots were created. Scatter plots were used to visualize the success outcome progress of every rocket launches using the Flight Number feature, and its relation with other features like Payload Mass (kg), Launch Site, and Orbit type. Similarly, scatter plots for Payload Mass (kg) vs Launch Site and Payload Mass (kg) vs Orbit Type, were plotted to reveal the relationship between those features.
- A bar chart were used to visualize the relationship between every orbit type and the success rate of the landing. Conclusions about the higher success rate between the orbits from this plot can be made.
- A line chart were plotted to get the average launch success trend by year.

GitHub URL

EDA with SQL

- Some queries were used to extract unique information from the database, like the total payload mass carried by an unique Booster in certain launch site, or the average payload mass carried by an specific booster version.
- Queries to extract all records about specific launch sites; a query for the booster records that comply with certain criteria, like the ones that carried the maximum payload mass.
- More complex queries give more detailed information, like the names of the booster versions which have carried the maximum payload mass; and a list of the records displaying the month names, failure landing outcomes in drone ship, booster versions, and launch sites, during 2015.

GitHub URL

Build an Interactive Map with Folium

- Using the interactive maps, three circles with three markers were added to the map to indicate each rocket launch site in the data frame.
- A cluster of markers with different colors were added to each launch site to indicate the success outcome of every launch on that site.
- The MousePosition feature were added over the map to see the coordinates of the most closest coastline, railway, highway, and a city, to one of the launch sites. Then polylines between the launch site and these points were added, next to a marker to indicate the distance in kilometers. That way it's visible how the launch site is placed in relation to its proximities.

GitHub URL

Build a Dashboard with Plotly Dash

- For the dashboard application it was added a dropdown menu to select each launch site name, even an option “All Sites” was added to represent the data of all launch sites in the same graph.
- Next is the pie chart plot to visualize the percentage of success landing on every site, or success landing rate of an individual launch site.
- Finally, a scatter plot next to a range-slider which function is to select the payload mass range, shows every Booster version that succeeded or failed to land in the launch site selected in the dropdown menu. This way it can be shown which Booster version succeeded or failed the most, and the payload range it was carrying in any case.

GitHub URL

Predictive Analysis (Classification)

- The models trained and tested were Logistic Regression, Support Vector Machines, Decision Tree Classifier, and K-nearest neighbors. To build a machine learning pipeline to predict if the first stage of the Falcon 9 lands successfully, the first step is the preprocessing of the data to get it standardized. Then the Train-Test split of the data is performed.
- Next, to evaluate a machine learning model, the model is trained with the training data and later the GridSearchCV object is performed to find the parameters that allow the model to perform best.
- Now using these parameters to get the accuracy value of the model, it can be determine which of the models perform best.
- Finally with the testing data, the prediction of the landing is performed, and the confusion matrix for every model is plotted using the real classification vs the predicted classification.

GitHub URL

Results

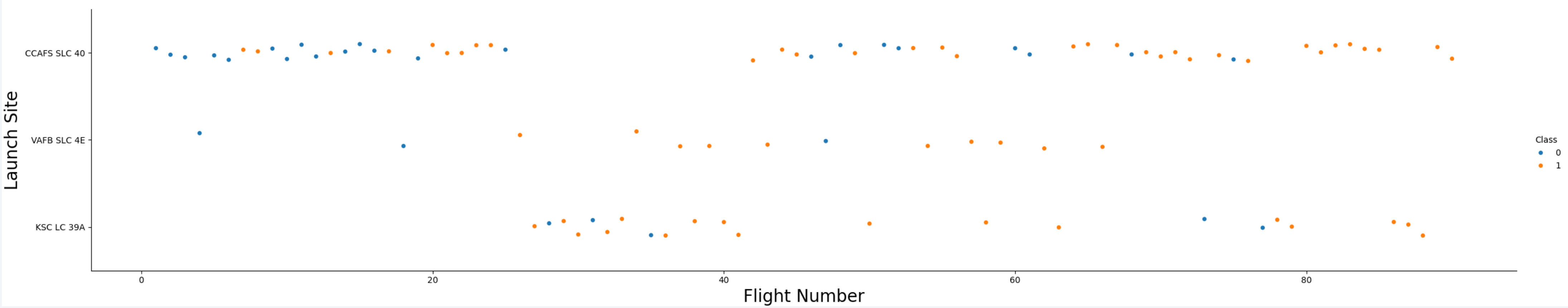
In the next slides the next content will be presented:

- Exploratory data analysis results with plots.
- Interactive visual analytics plots in dashboard.
- Predictive analysis results.

Section 2

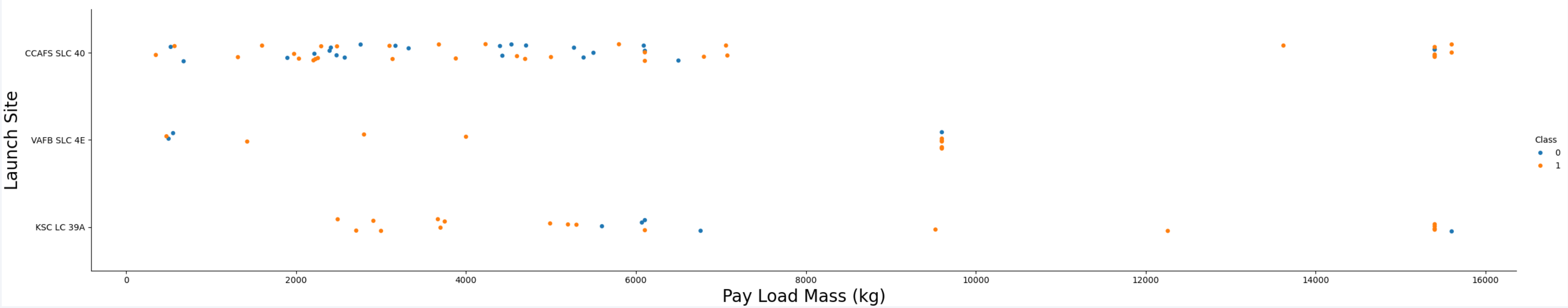
Insights drawn from EDA

Flight Number vs. Launch Site



- The site with major rocket launches is CCAFS SLC-40, and the site with less flights was VAFB SLC 4E. Major success landing was shown in the latter in comparison with other sites but keep in count the low quantity of launches there. Major success landing was shown in CCAFS SLC-40 in its latest flights.
- No launches were made in KSC LC-39A site before the 26 launch, and between the 25 and 42 rocket launch no launches were made in CCAFS SLC-40 site.

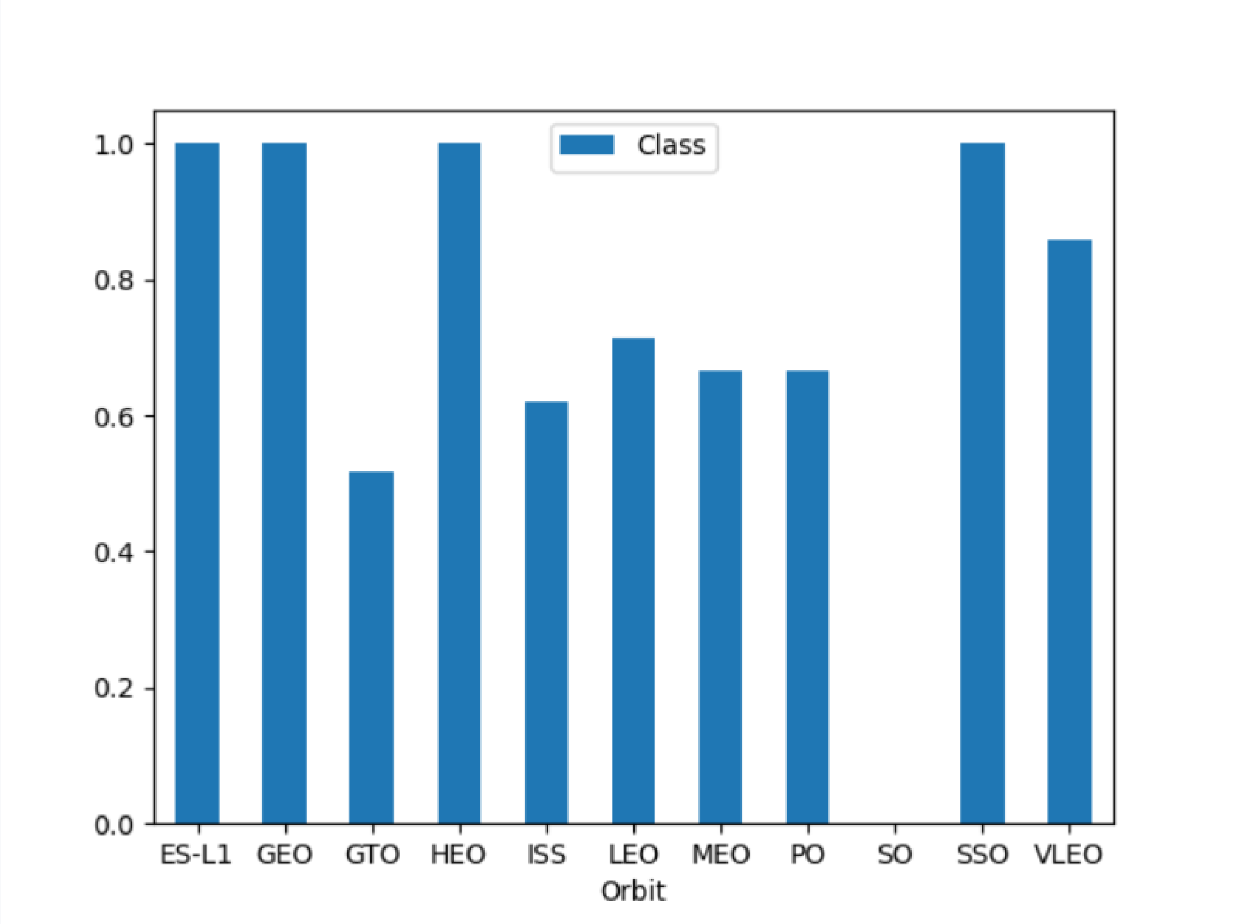
Payload vs. Launch Site



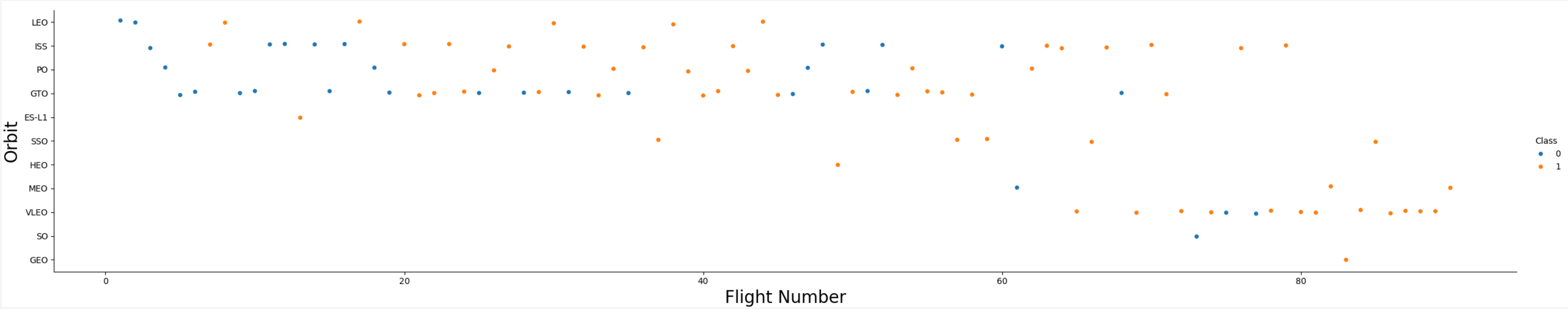
- For VAFB SLC 4E site there are no rocket launches with a pay load mass greater than 10000 kg.
- Most of CCAFS SLC-40 site's launches carried a payload less than 8000 kg.

Success Rate vs. Orbit Type

- The orbits with high success rate are ES-L1, GEO, HEO, SSO, and VLEO.

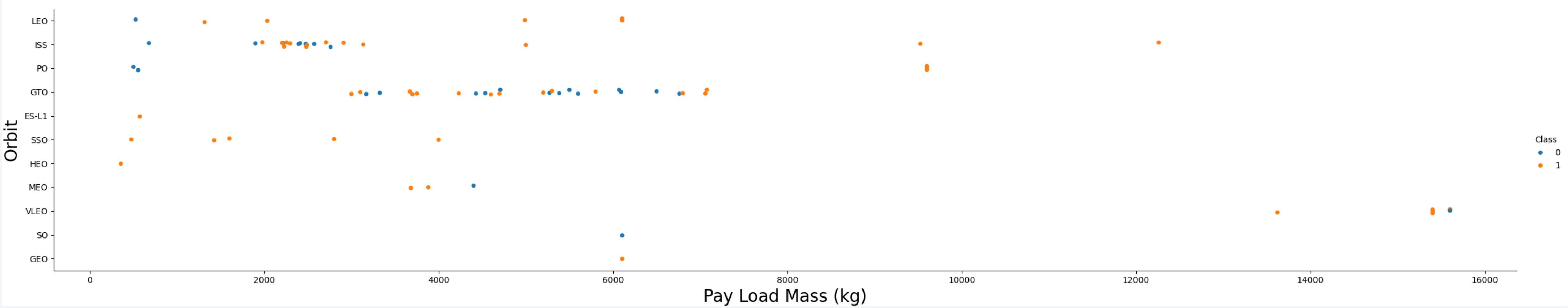


Flight Number vs. Orbit Type



- The only orbit to show some relationship between flight number and success landig

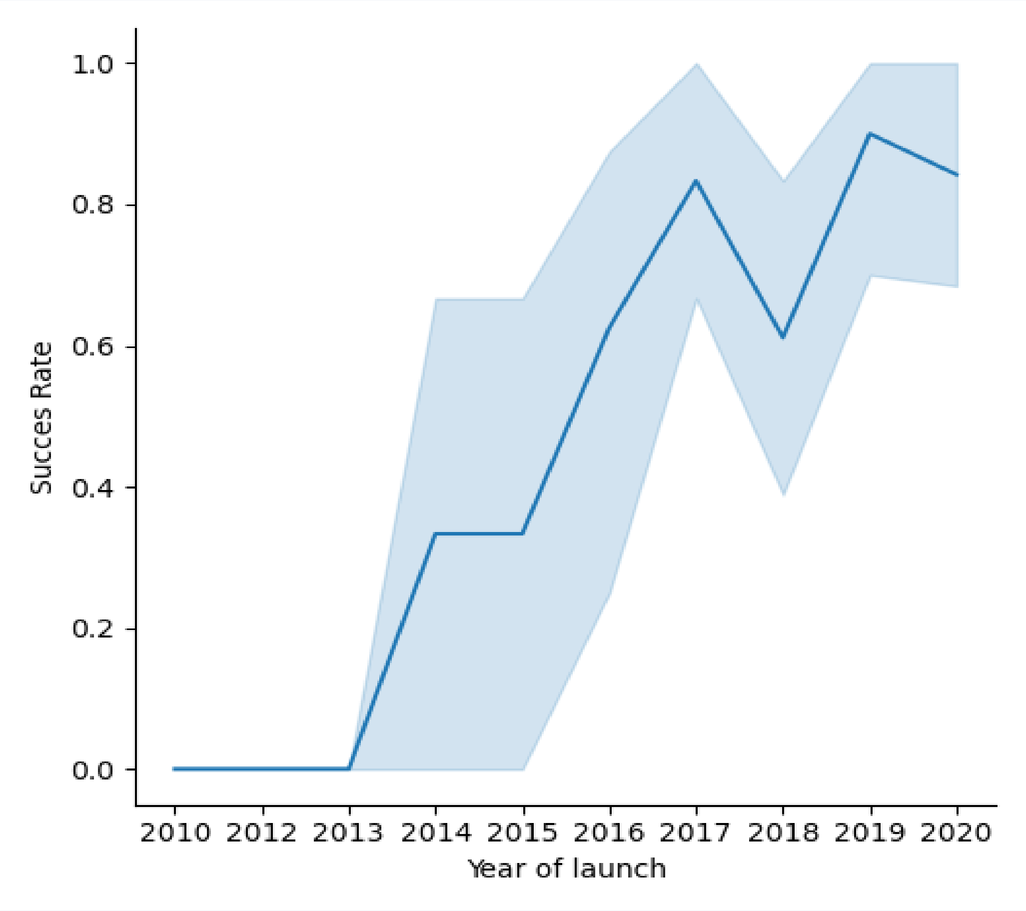
Payload vs. Orbit Type



- The successful landing rate is high for ISS, PO, and VLEO with higher pay load mass values.
- The GTO orbit seems to have equal amount of successful and unsuccessful landing for similar payloads.

Launch Success Yearly Trend

- The success rate since 2013 kept increasing until 2020



All Launch Site Names

- The SpaceX database only present four sites of rocket launches. Further in this report the locations will be presented.

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

Launch Site Names 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

- 5 records where launch sites begin with “CCA”.

Total Payload Mass

- The total payload carried by boosters from NASA is 45956 kg.

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS total_payload_kg FROM SPACEXTBL WHERE Customer == 'NASA (CRS)'
```

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

total_payload_kg
45596

Average Payload Mass by F9 v1.1

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

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS avg_payload_kg FROM SPACEXTBL WHERE Booster_Version == 'F9 v1.1'
```

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

avg_payload_kg
2928.4

First Successful Ground Landing Date

- On 12/22/2015 was the first successful landing outcome on ground pad.

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

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

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

Booster_Version	PAYLOAD_MASS__KG_	Booster_Version	PAYLOAD_MASS__KG_
F9 v1.1	4535	F9 FT B1031.2	5200
F9 v1.1 B1011	4428	F9 FT B1032.2	4230
F9 v1.1 B1014	4159	F9 B4 B1040.2	5384
F9 v1.1 B1016	4707	F9 B5 B1046.2	5800
F9 FT B1020	5271	F9 B5 B1047.2	5300
F9 FT B1022	4696	F9 B5 B1046.3	4000
F9 FT B1026	4600	F9 B5 B1048.3	4850
F9 FT B1030	5600	F9 B5 B1051.2	4200
F9 FT B1021.2	5300	F9 B5B1060.1	4311
F9 FT B1032.1	5300	F9 B5 B1058.2	5500
F9 B4 B1040.1	4990	F9 B5B1062.1	4311

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes taken from the Spacex dataset. Only one mission was a failure.

```
%sql SELECT "Mission_Outcome", 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

Boosters Carried Maximum Payload

- The next table presents the names of the Booster Versions which have carried the maximum payload mass: 15600 kg.

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

- This table presents the failed landing outcomes in drone ship, their booster versions, and launch site names on 2015. Both occurred in the same landing site:

month	Booster_Version	Launch_Site	Landing_Outcome
10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

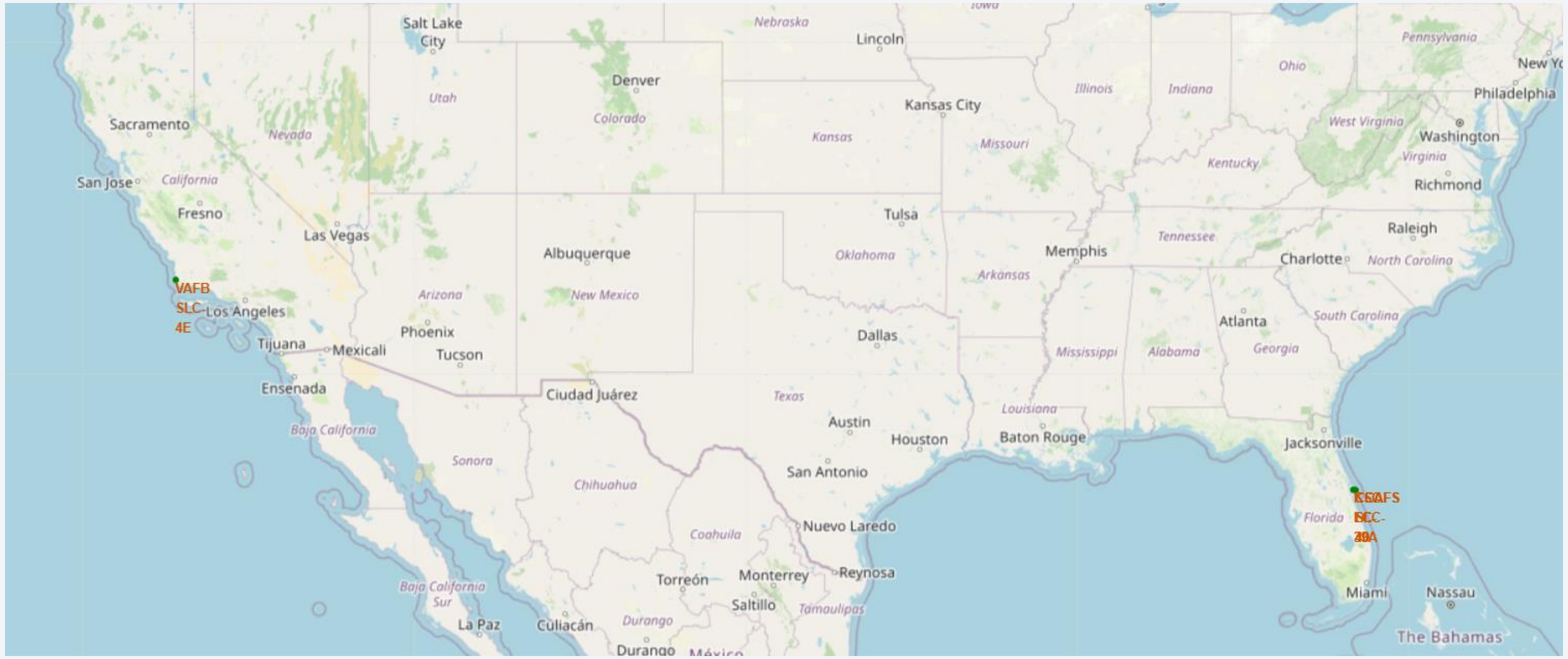
- Next is the rank of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order:

Landing_Outcome	TOTAL
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

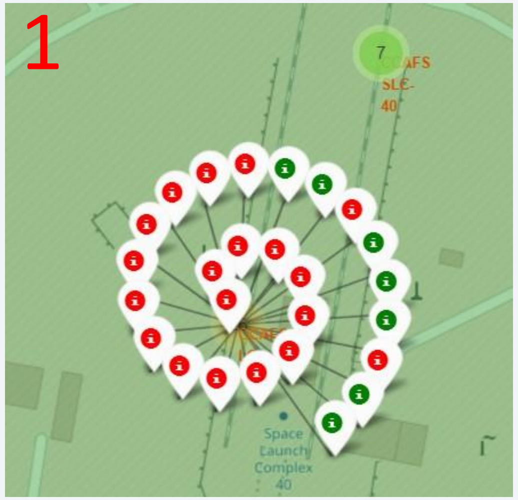
Launch Sites Proximities Analysis

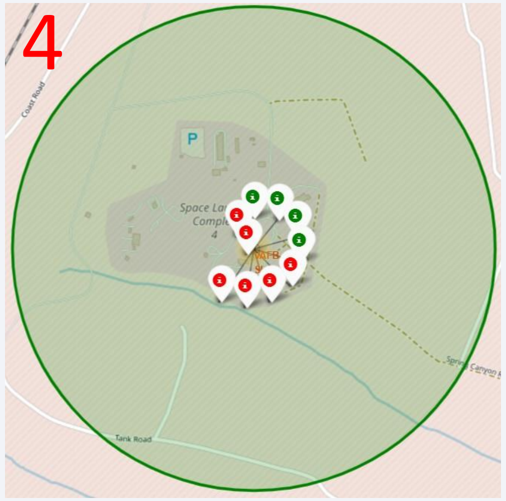
Lunch Sites Location



As the map shows, the location of the launching sites are placed very close to the coast and very south on the territory.

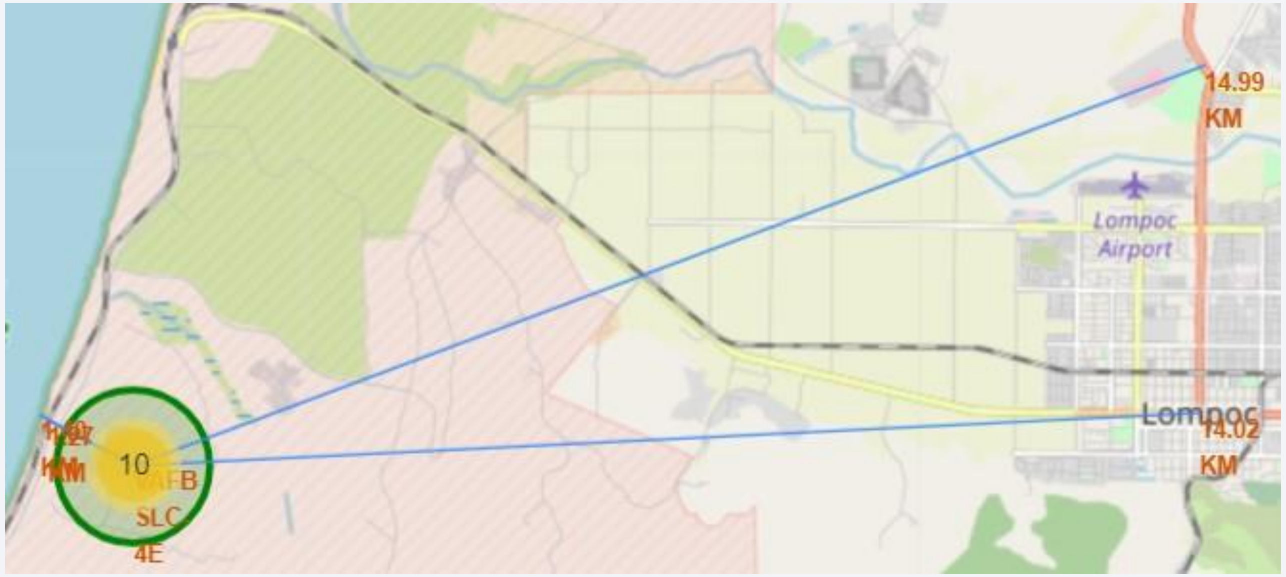
Success/Failed launches shown in the map





- CCAFS SLC-40 (2) and CCAFS LC-40 (1) are next to each other but SLC-40 presents a very low landing success rate. CCAFS SLC-40 (2) and VAFB SLC-4E (4) show more failures than success. KSC LC-39A (3) is the only one showing a high rate of success landings.

Launch site and its distance from proximities

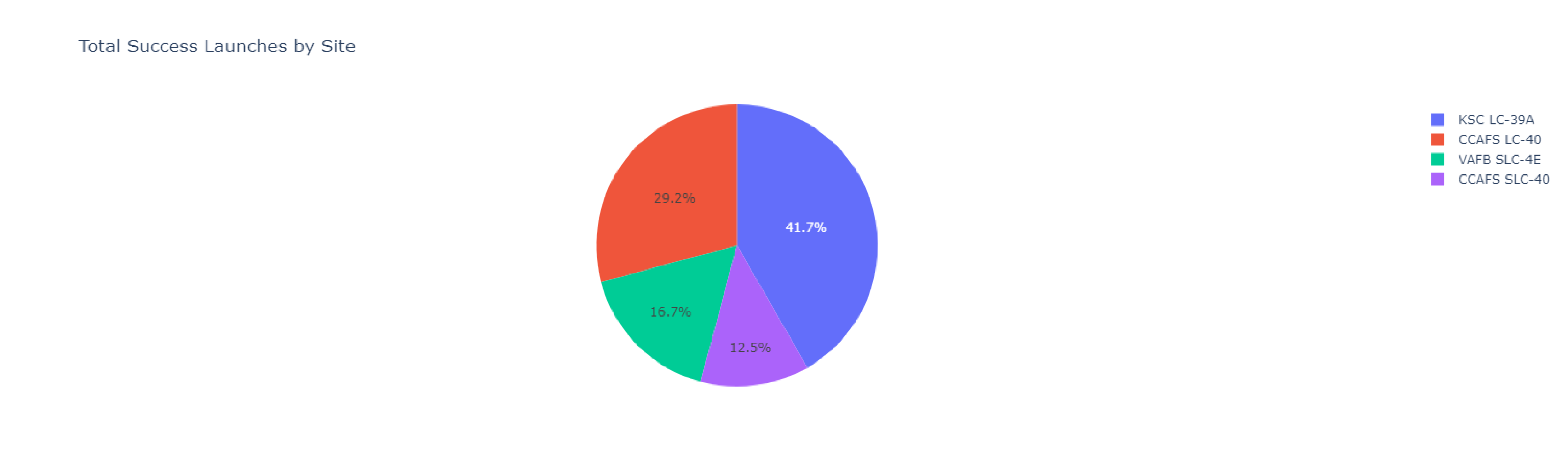


- The launch sites seems to be far away from highways and cities, but very close to railways and to the coast line. The first could be explained as a safety hazard, and in the case of the railway it could be explained as a way of transporting material required for the launching site's work.

Section 4

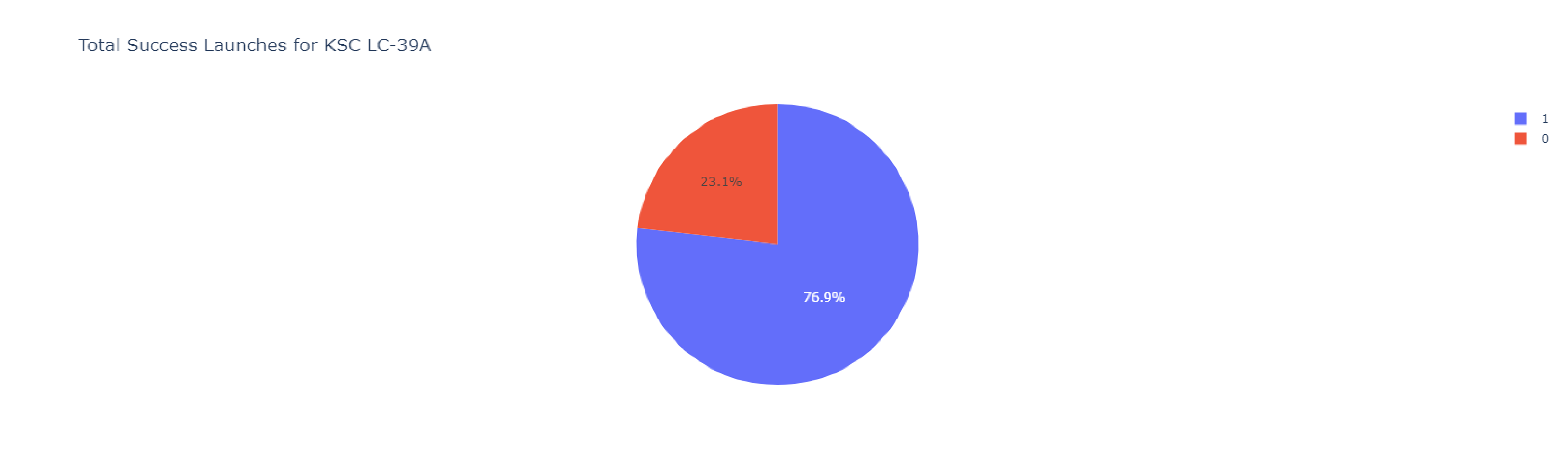
Build a Dashboard with Plotly Dash

Launch success count for all sites



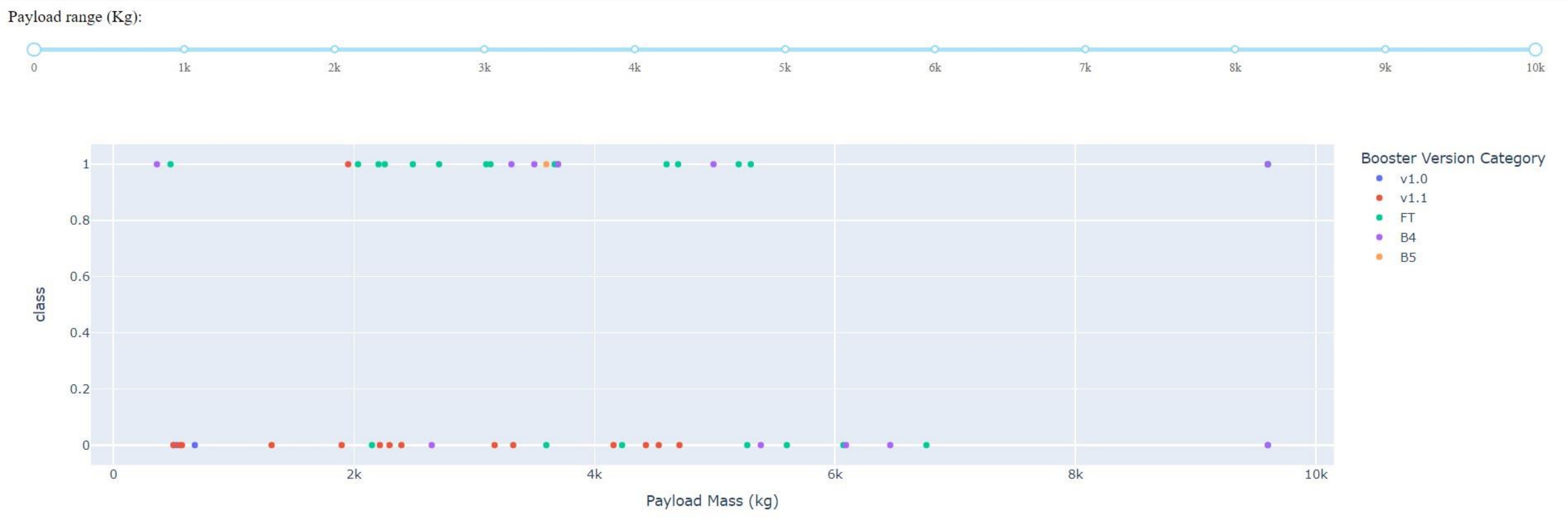
- As shown in the pie chart, the launch site with higher success landing is KSC LC-39A, and the lowest success landing is from CCAFS SLC-40.

Launch site with highest launch success ratio



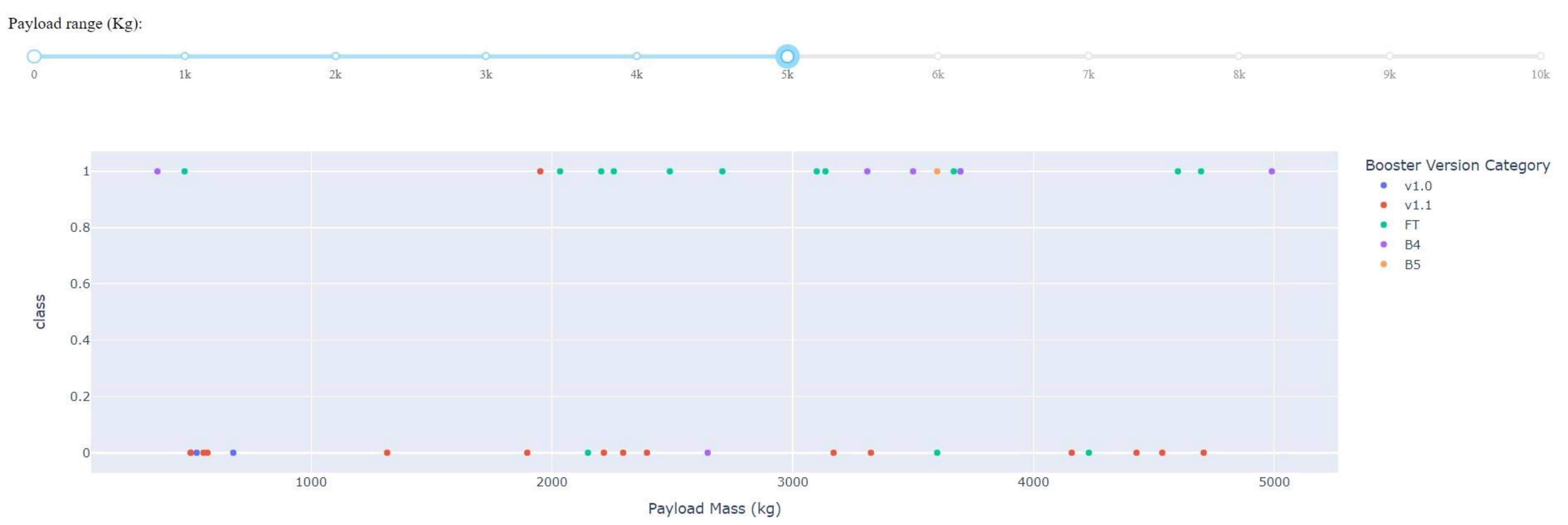
- The launch site with higher success landing is KSC LC-39A, with 76.9% of success landing.

Payload vs. Launch Outcome scatter plot for all sites (1)



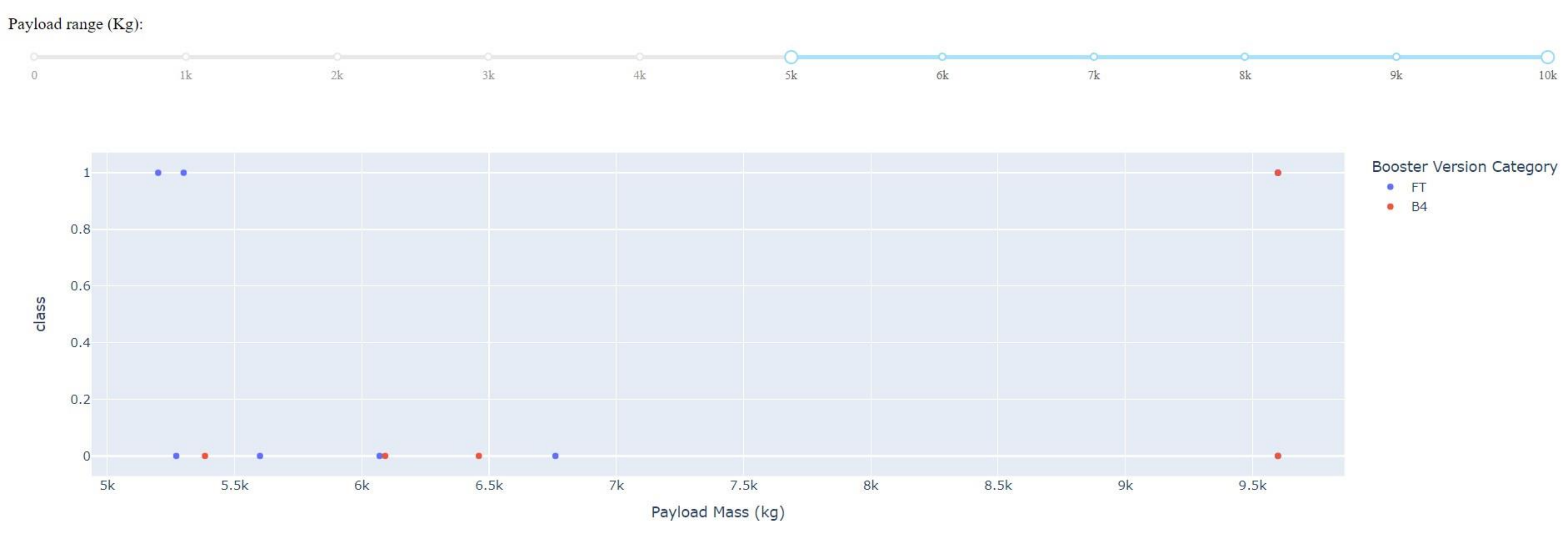
- This is scatter plot of Payload Mass vs Class for all sites in every Booster version category and the whole Payload range.

Payload vs. Launch Outcome scatter plot for all sites (2)



- This is scatter plot of Payload Mass vs Class for all sites in every Booster version category and the Payload range is less than 5000 kg.

Payload vs. Launch Outcome scatter plot for all sites (3)



- This is scatter plot of Payload Mass vs Class for all sites in every Booster version category and the Payload range between 5000 kg and 10000 kg.

Payload vs. Launch Outcome scatter plot for all sites

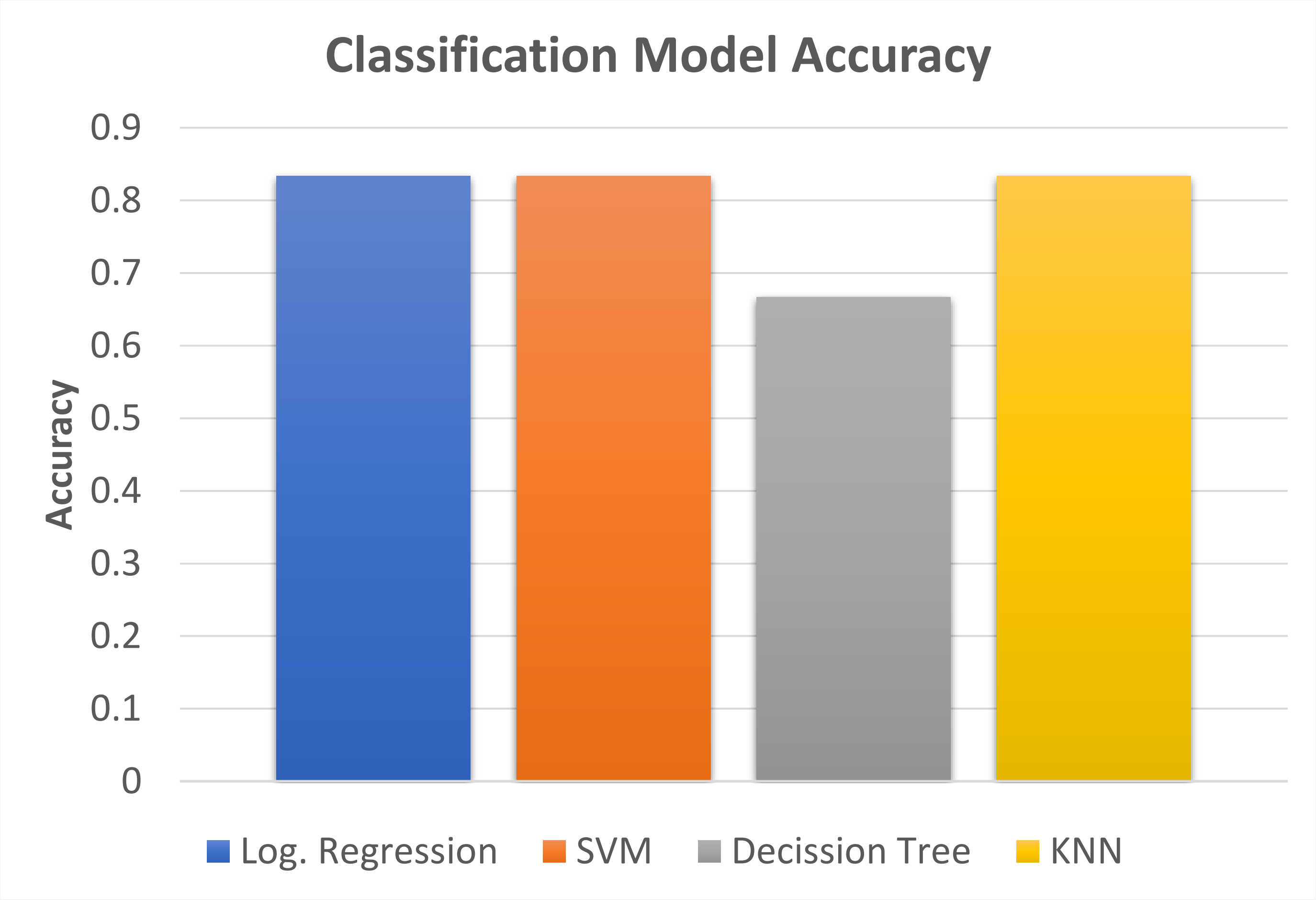
- In these scatter plots of Payload Mass vs Class for all sites indicating every Booster version category and the whole Payload range, it can be seen that the Booster version with most failed landings is the v1.1 with payload less than 5000 kg.
- Only Booster version B4 and FT had launched payloads greater than 8000 kg.
- Booster v1.0 launched payloads of less than 1000 kg and all of them failed in the landing.
- Most of successful landings carried payloads below 6000 kg and were carried by FT and B4 Boosters version.

Section 5

Predictive Analysis (Classification)

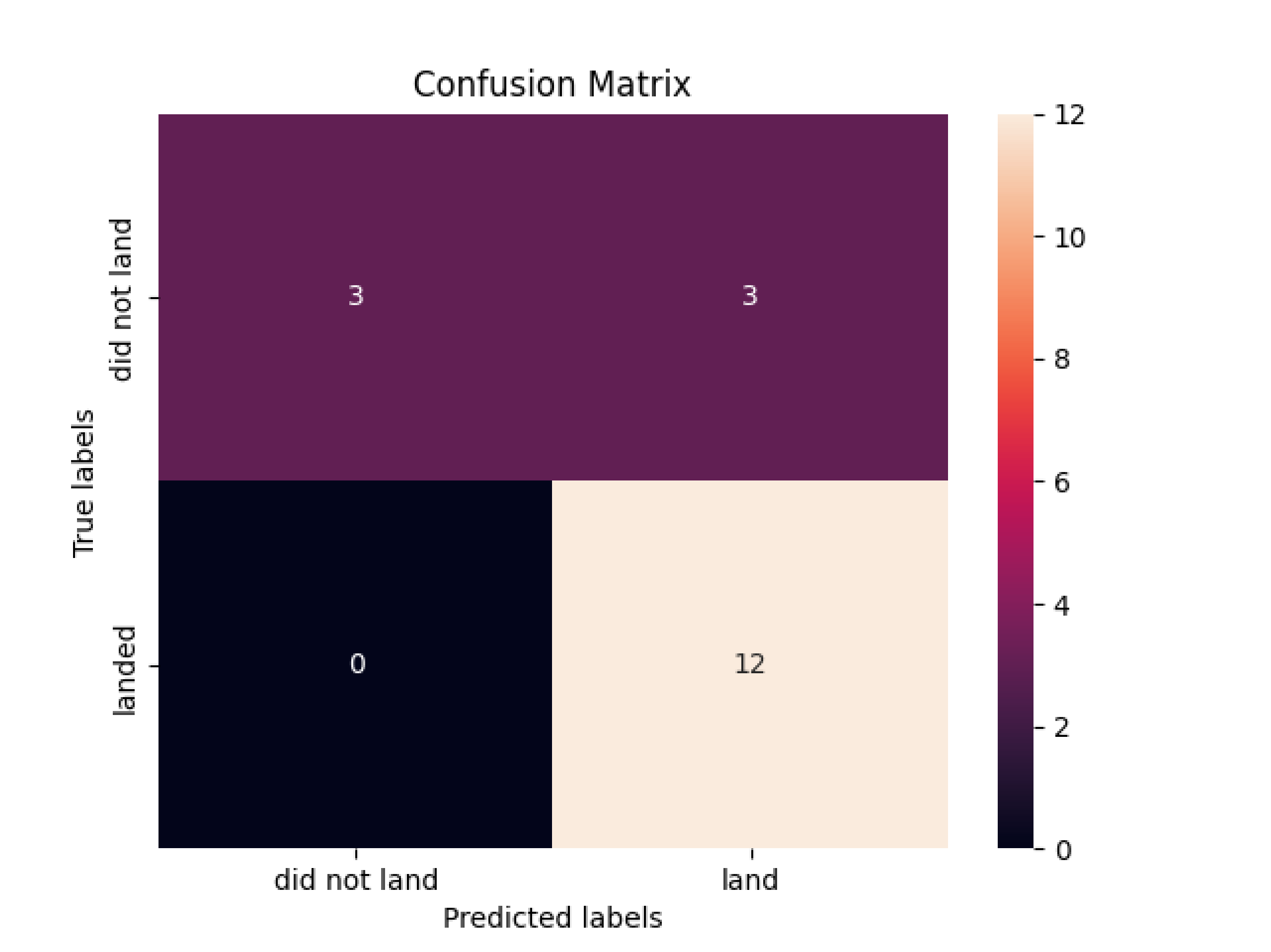
Classification Accuracy

- The Support Vector Machine (SVM), Logistic Regression and Key Nearest Neighbors (KNN) models have great performance with the test data, and the three of them give an equal score accuracy of 0.8333. Although, in Training data Logistic Regression gave maximum accuracy.



Confusion Matrix

- Practically all three of the algorithms gave the same result, including its confusion matrix. It can be seen that there is major True Positives with zero False Negatives, only problem seen is a 50% of False Positives from the True did not land data.



Conclusions

- Most of the models evaluated gave great accuracy values, but the one giving the most accuracy including the training dataset is the Logistic Regression algorithm. So for the provided dataset this models was choose to perform the predictions for this project.
- Boosters that carried light payloads (below 5000 kg) were most successful in landing, except for Booster v1.0 and v1.1.
- KSC LC-39A is the launch site with higher success rate but all launchings with payload over 5500 kg have failed.
- Successful rocket launches have been increasing significatively over time since 2013.

Thank you!