IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Apurva Mehrotra
May 3, 2023

Outline

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

Executive Summary

Summary of methodologies

The following methodologies were used to inform the findings contained in this presentation:

- Data collection; data wrangling; EDA with data visualization; EDA with SQL; Interactive maps with Folium; Interactive dashboards with Plotly Dash; Predictive analysis (using machine learning)

This presentation includes the results from the methodologies outlined above:

- Results of exploratory data analysis
- Interactive dashboards
- Predictive analysis

Introduction

- Project background and context

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 if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

- Problems you want to find answers

I want to determine the price of different rocket launches and understand what parameters most influence whether or not different launch types will result in a reusable first stage.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX REST API and web-scraping from Wikipedia
- Perform data wrangling
  - Data was processed using one-hot encoding for categorical features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Create a machine learning pipeline to predict if the first stage will land
  - Train the best performing model to make accurate predictions

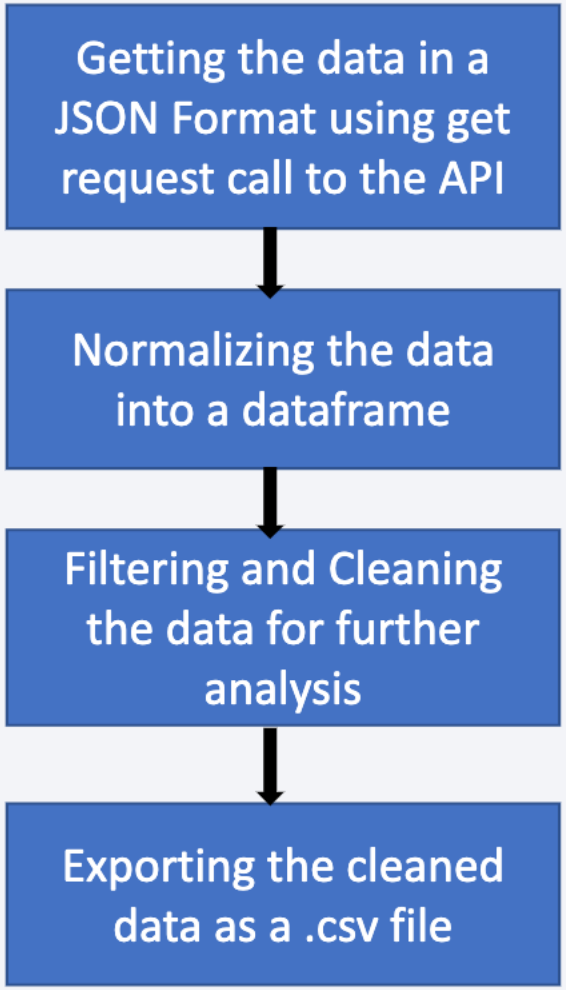
Data Collection

We collected SpaceX launch data so that we could visualize and model data to predict if a launch would be successful. One data set was collected using the SpaceX API. The other data set was used by web scraping the Falcon 9 Wikipedia page.

Data Collection – SpaceX API

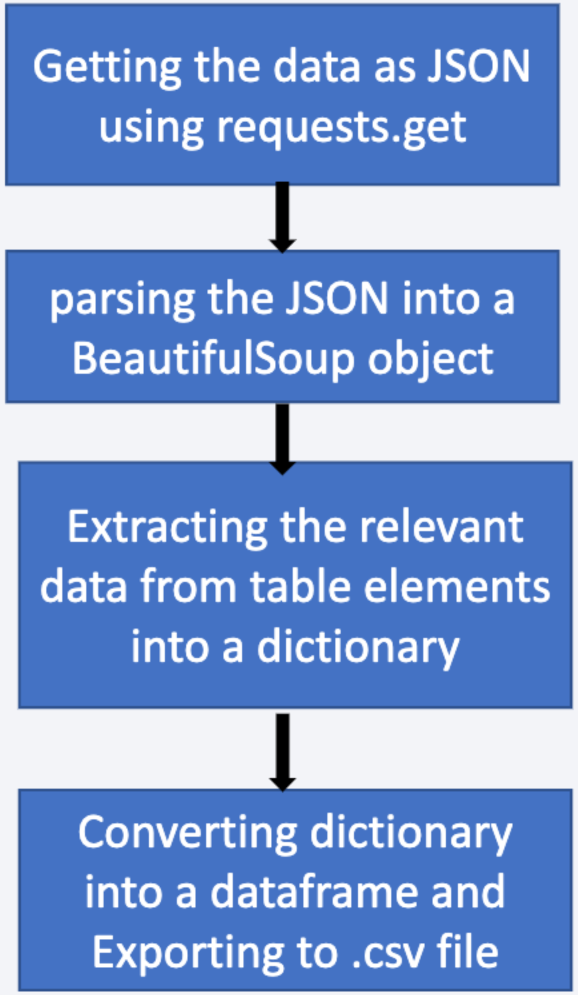
The first dataset is collected with the SpaceX API using the “get” request in Python. We normalize the JSON contents and use pandas to extract the information we need and clean the data.

See data collection notebook in GitHub here.



Data Collection - Scraping

- We performed web-scraping to collect historical launch records of the Falcon 9 from Wikipedia. After requesting the page we used Beautiful Soup to parse and extract the relevant data.
- See the web-scraping notebook in Github here.



Data Wrangling

Exploratory Data Analysis (EDA) was performed on the data in order to calculate the number of launches at each site; calculate the number of orbits for which each launch was aimed; and calculate the number of mission outcomes by orbit type.

An Outcome column was created in which a 0 meant failure and a 1 signaled success.

See the Data Wrangling and EDA notebook on Github here.

EDA with Data Visualization

We analyzed and visualized the relationship between a number of variables including flight number, launch site, payload, and orbit type. We used scatterplots to observe relationships between variables.

See the notebook for EDA with data viz on Github here.

EDA with SQL

After connecting our database to the IBM Db2 API, we performed data analysis using SQL to get the following information:

- Unique launch sites
- Launch sites beginning with CCA
- Total payload mass carried by boosters launched by NASA
- Average payload mass carried by booster version F9v1.1
- Date of first successful landing outcome
- Total number of successful and failed outcomes
- Names of booster versions carrying maximum payload mass
- Rank of the count of landing outcomes for specified date ranges

See the EDA with SQL notebook on Github here

Build an Interactive Map with Folium

The following objects were added to the interactive map using Folium:

Circles: to highlight a particular area on a specific site

Markers: to mark the site

Marker clusters: To show several sites around the same coordinates

Mouse position: to get coordinates for where the mouse is pointing on a map

Polyline: to draw a line from different geographic locations

See the Interactive Map with Folium notebook on Github here.

Build a Dashboard with Plotly Dash

We built an interactive dashboard using Plotly Dash with the following features:

- A dropdown for selecting sites with a pie chart to show success/failure for each site
- A range slider for the payload range to view scatterplots between payload and class
- This interactivity allow us to quickly analyze the relationship between payload and launch sites

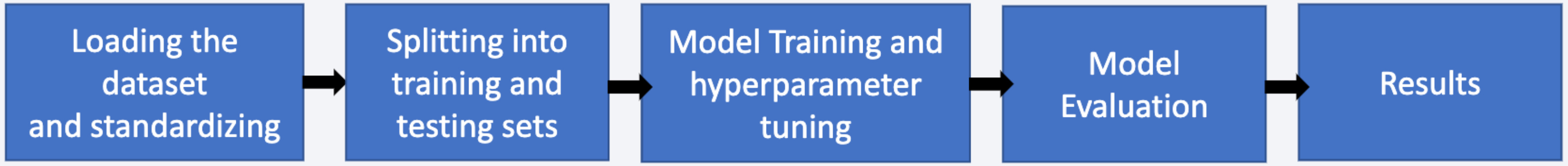
See the Plotly Dash interactive notebook on Github here

Predictive Analysis (Classification)

The features (or predictor variables) and the target variables were loaded into dataframes.

The data was split into training and test data. We used classification models such as logistic regression, SVM, decision trees, and KNN. We also plotted the confusion matrices for each model.

We use the test data to determine which machine learning model performs best.



See the notebook for predictive analysis on Github here:

Results

- The exploratory data analysis has shown us that successful landing outcomes are somewhat correlated with flight number. It was also apparent that successful landing outcomes have had a significant increase since the year 2015.
- All launch sites are located near the coast line. Perhaps, this makes it easier to test rocket landings in the water.
- Furthermore, the sites are also located near highways and railways. This may facilitate transportation of equipment and research material.
- The machine learning models that were built, were able to predict the landing success of rockets with an accuracy score of 83.33%. This accuracy can be increased in future projects with more data.

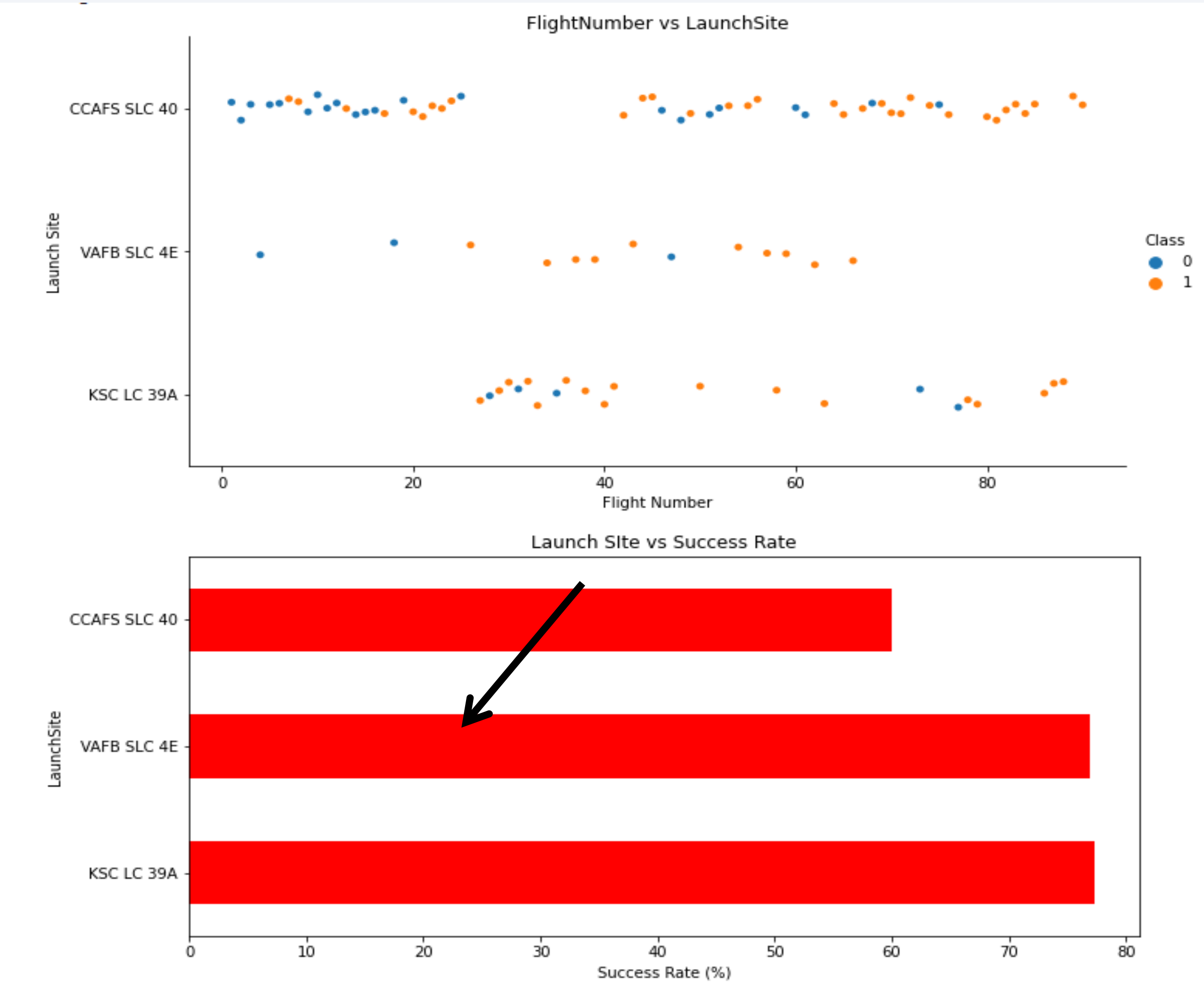
Section 2

Insights drawn from EDA

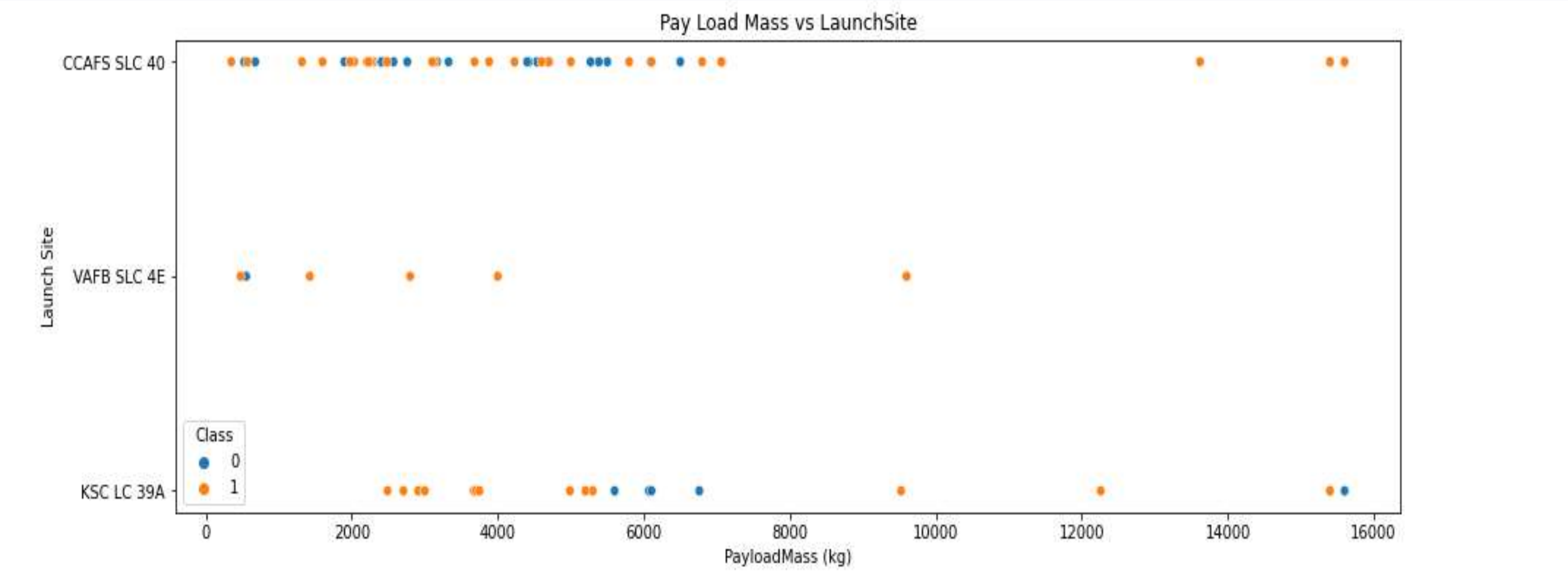
Flight Number vs. Launch Site

- It appears that there were more successful landings as the flight numbers increased. It also seems that launch site **CCAFS SLC 40** had the most number of landing attempts while the site **VAFB SLC 4E** had the least number of attempts.

- Looking at the second chart, we can see that there is no Launch Site with a success rate below 60%.



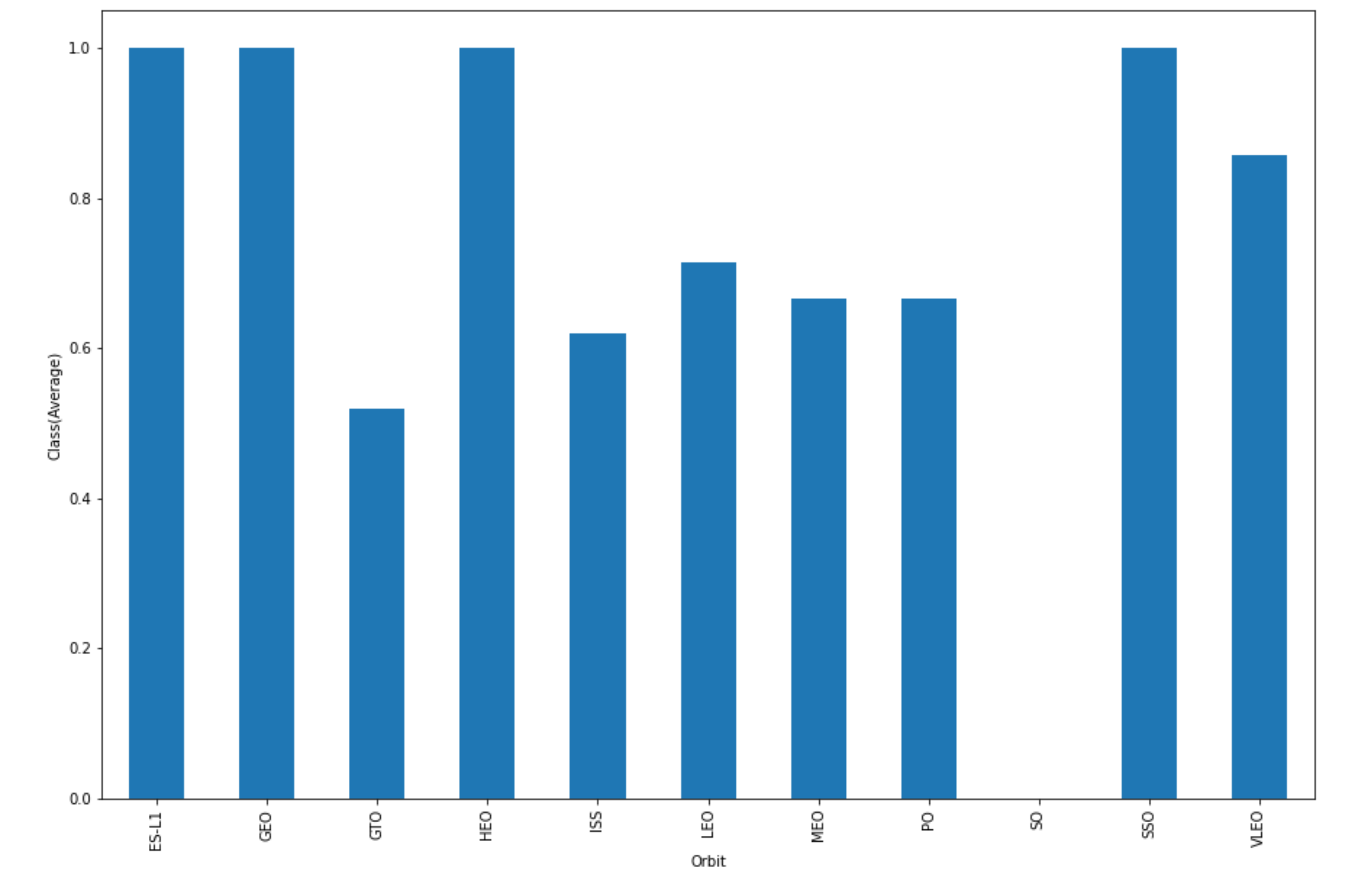
Payload vs. Launch Site



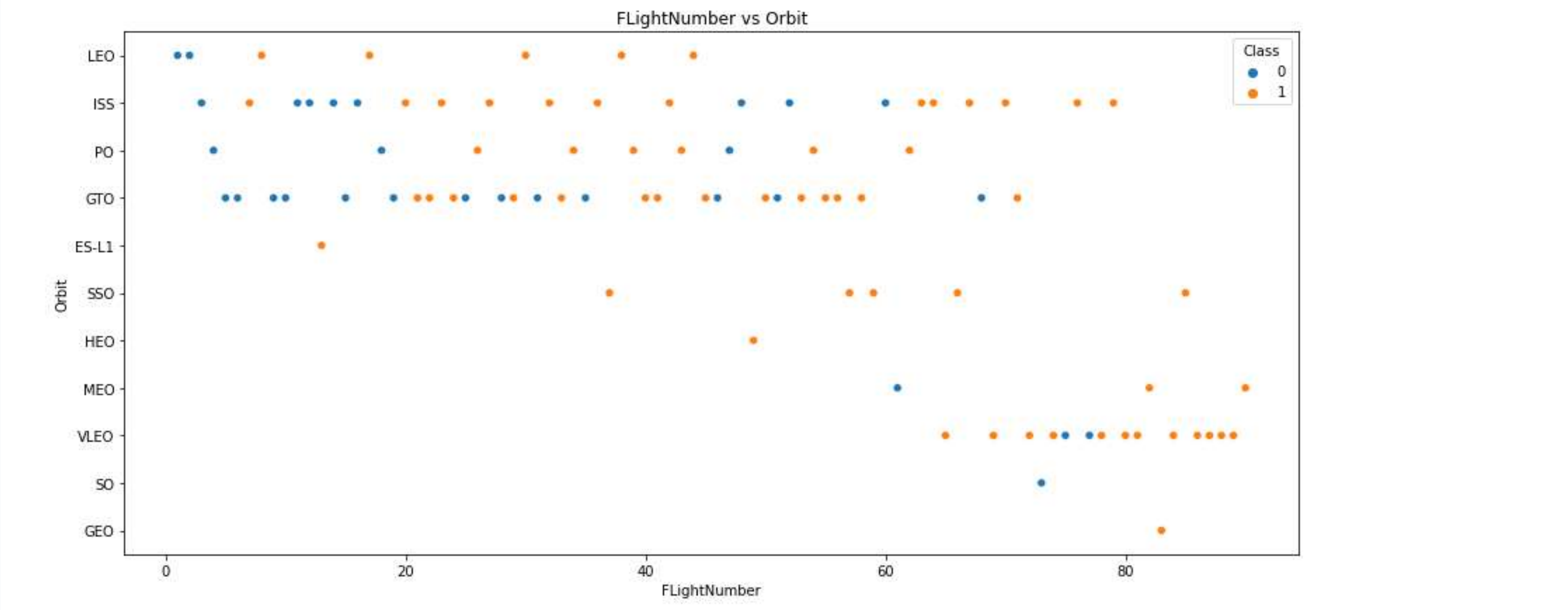
- Now if you observe the scatter point chart, you will find for the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).

Success Rate vs. Orbit Type

The orbit types **SSO**, **HEO**, **GEO** and **ES-L1** had the highest success rate.

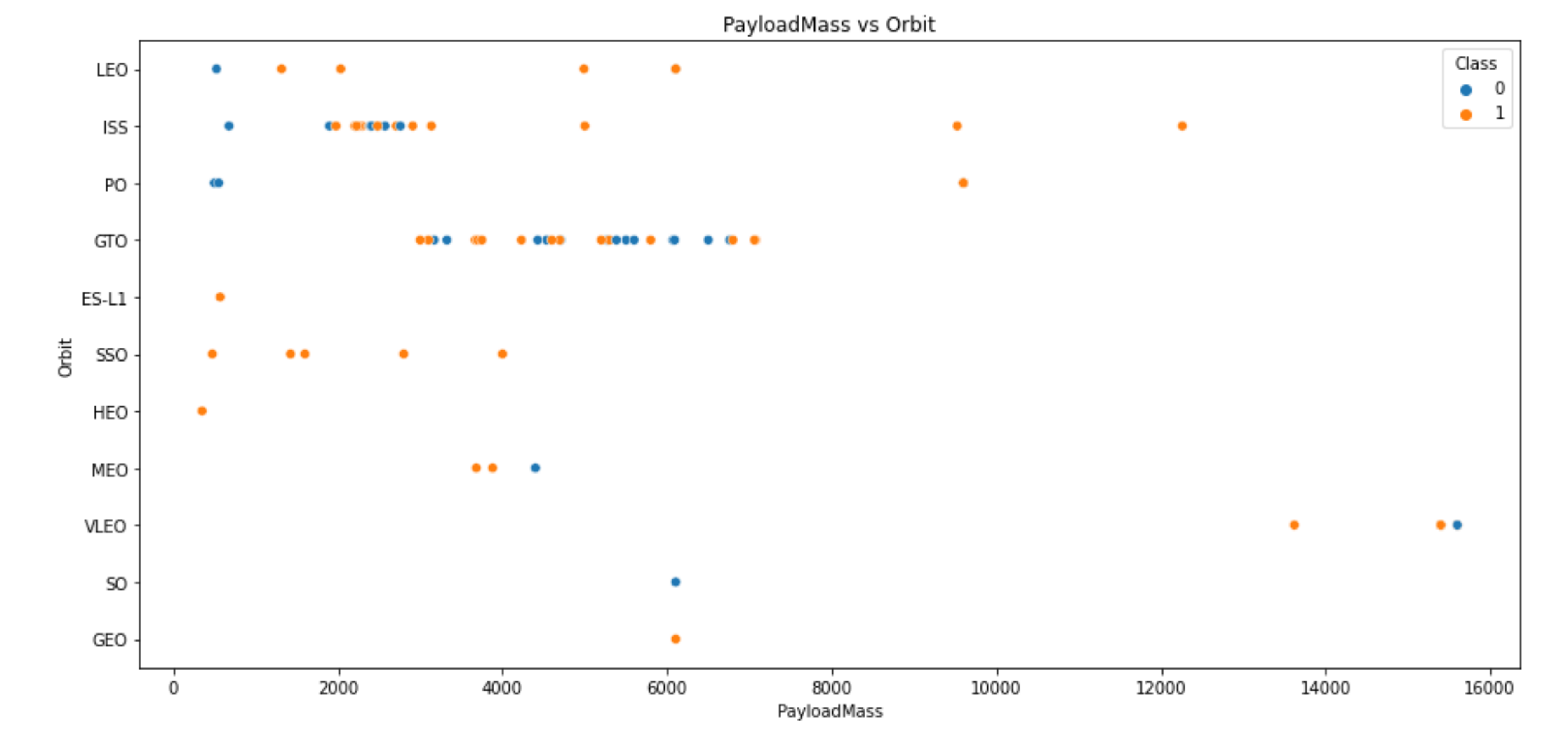


Flight Number vs. Orbit Type



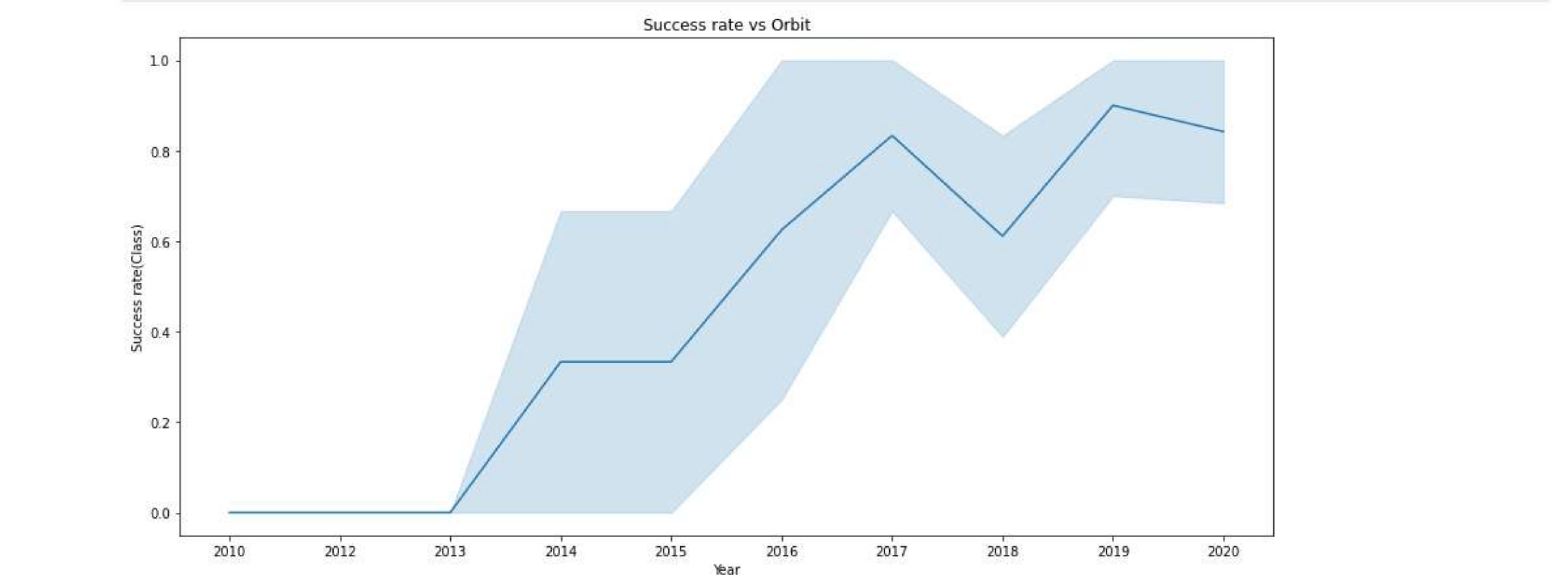
You can see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



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

Launch Success Yearly Trend



It is apparent that the success rate has significantly increased from 2013 to 2020.

All Launch Site Names

Given the data, these are the names of the launch sites where different rocket landings where attempted:

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

Launch Site Names Beginning with 'CCA'

Date	Launch_Site	Orbit	Customer	Mission_Outcome	Landing_Outcome
04-06-2010	CCAFS LC-40	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	CCAFS LC-40	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22-05-2012	CCAFS LC-40	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	CCAFS LC-40	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	CCAFS LC-40	LEO (ISS)	NASA (CRS)	Success	No attempt

These are 5 records where launch sites begin with the letters 'CCA'. As we can see, there are other organizations besides SpaceX that were testing their rockets.

Total Payload Mass

- The information in the table displays the total payload mass carried by boosters launched by NASA .
- It seems that *NASA (CRS)* had a significantly higher total payload mass compared to the rest.

Customer	Total_Payload_Mass
NASA (CRS)	45596
NASA (CCDev)	12530
NASA (CCP)	12500
NASA (CCD)	12055
NASA (CTS)	12050
NASA (CRS), Kacific 1	2617
NASA / NOAA / ESA / EUMETSAT	1192
NASA (LSP) NOAA CNES	553
NASA (COTS)	525
NASA (LSP)	362
NASA (COTS) NRO	0

Average Payload Mass by F9 v1.1

Average_Payload_Mass (kg)	Booster_Version
2928.4	F9 v1.1

- The average payload mass carried by F9 v1.1 was 2928.4 kg.

First Successful Ground Landing Date

Date	Landing_Outcome
22-12-2015	Success (ground pad)

- The first successful ground pad landing took place in December 2015. This was a historic reusable-rocket milestone for both SpaceX and the world.
- Prior to this, no one had ever brought an orbital class booster back intact.

Successful Drone Ship Landing with Payload between 4000 and 6000

Booster_Version	PAYLOAD_MASS__KG_	Landing_Outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

- It appears that there only 4 Boosters with a payload mass between 4000 and 6000.
- It is interesting to see that they all had successful landing outcomes.

Total Number of Successful and Failure Mission Outcomes

Mission_Outcome	Outcomes
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- It appears that missions generally tend to be successful with the exception of one failure.

Boosters That Carried the Maximum Payload Mass

- 12 boosters have carried the maximum payload mass of 15600 kg.
- Since the version names are similar, they might be from the same manufactures.

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 - Failed Landing Outcomes

Date	Launch_Site	Booster_Version	Landing_Outcome
10-01-2015	CCAFS LC-40	F9 v1.1 B1012	Failure (drone ship)
14-04-2015	CCAFS LC-40	F9 v1.1 B1015	Failure (drone ship)

- It appears that 2 boosters failed to land at the beginning of the year..
- The first successful landing took place later that year in December as we saw earlier.

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

- If we observe the table, it is apparent that the number of successful landings have increased since 2015.
- Before 2013, it seems that there were no attempts to land the boosters.

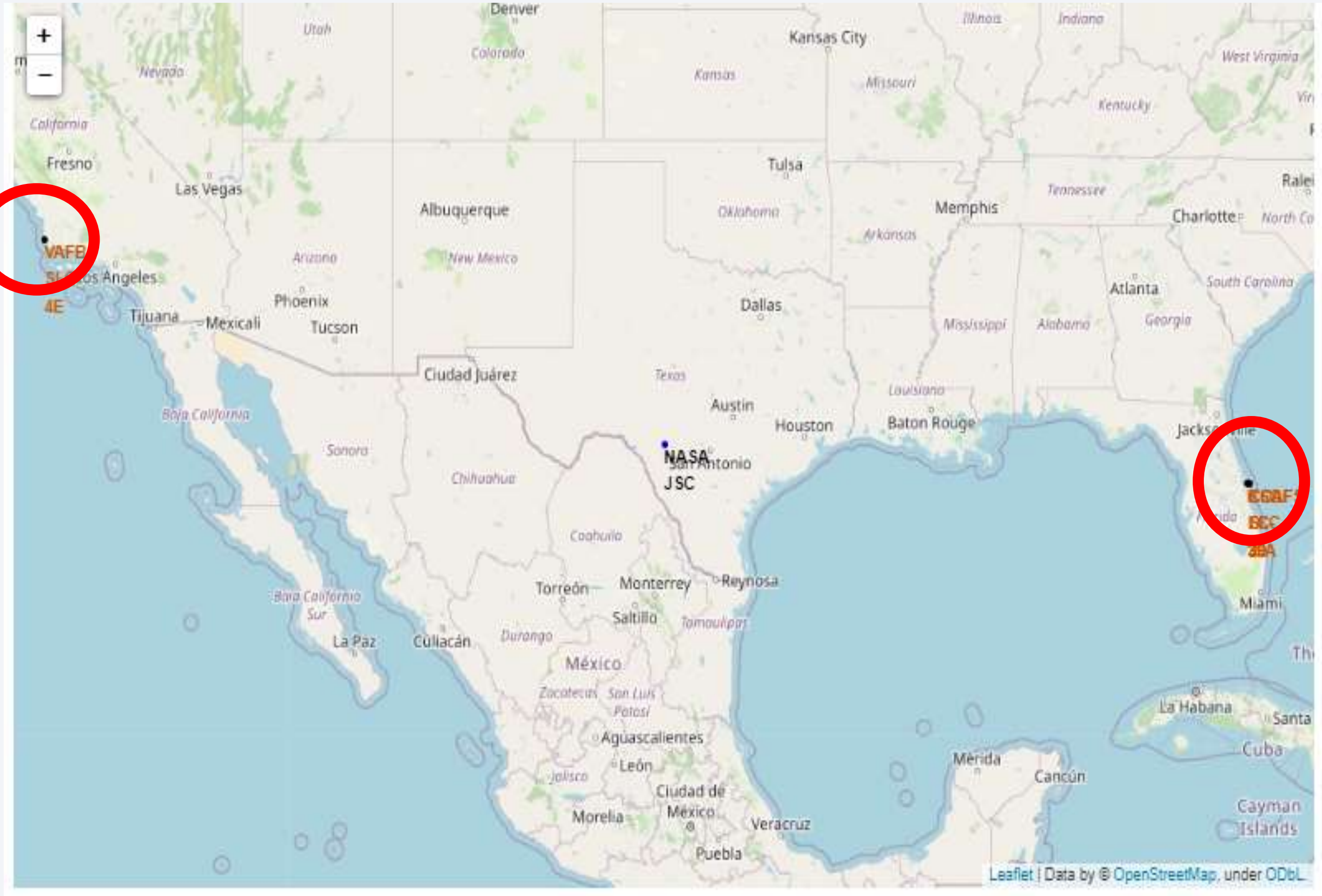
date	Landing_Outcome	Outcomes
2016-04-08	Success (drone ship)	14
2015-12-22	Success (ground pad)	9
2015-06-28	Precluded (drone ship)	1
2015-01-10	Failure (drone ship)	5
2014-04-18	Controlled (ocean)	5
2013-09-29	Uncontrolled (ocean)	2
2012-05-22	No attempt	22

Section 3

Launch Sites Proximities Analysis

Launch Site Locations

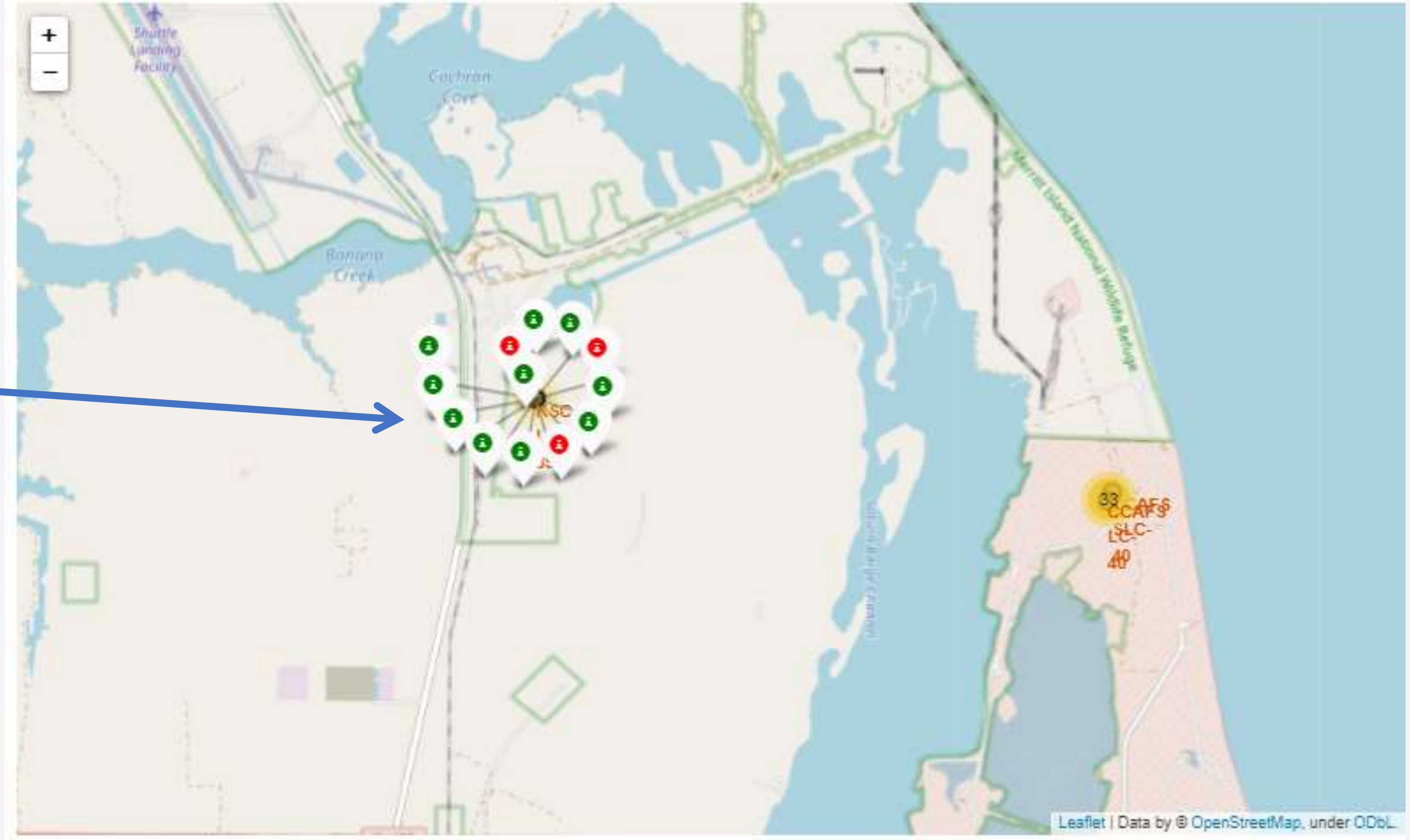
- We can see that all launch sites are in very close proximity to the coast and they are also a couple thousand kilometers away from the equator line.
- It is interesting to see that most launch sites are concentrated near Miami.



Success Rate of Rocket Launches

- The successful launches are represented by a green marker while the red marker represents failed rocket launches.

- It appears that **KSC LC-39A** had the highest success rate of rocket launches compared to other launch sites.



Surrounding Landmarks

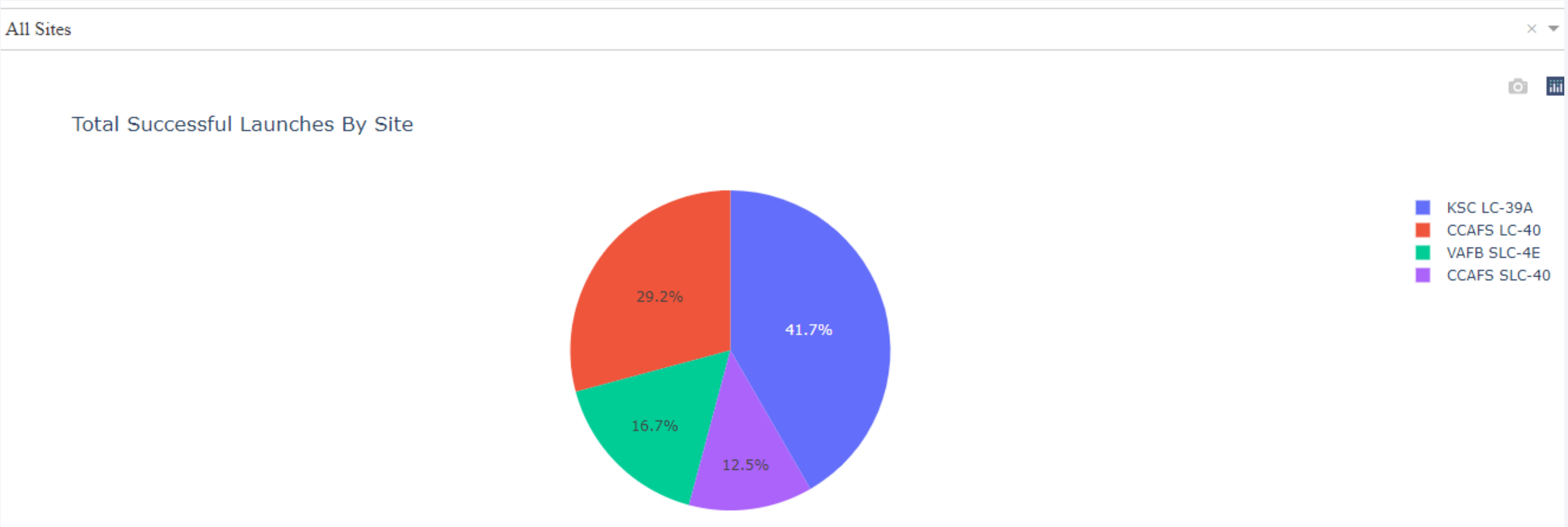
- It appears that launch sites are usually set up at least 18 km away from cities. This may be because of the desire to prevent any crashes near populated areas.
- It is also apparent that launch sites are in very close proximity to railways and highways. Perhaps, due to the necessary transportation requirements for rocket parts.
- The sites are close the coast line. This is evident with the many rocket landing tests on water bodies like the ocean.

Map Object	Colour
Nearest Highway	Green
Nearest Railway	Purple
Nearest City	Crimson
Nearest Coastline	Dark Blue

Section 4

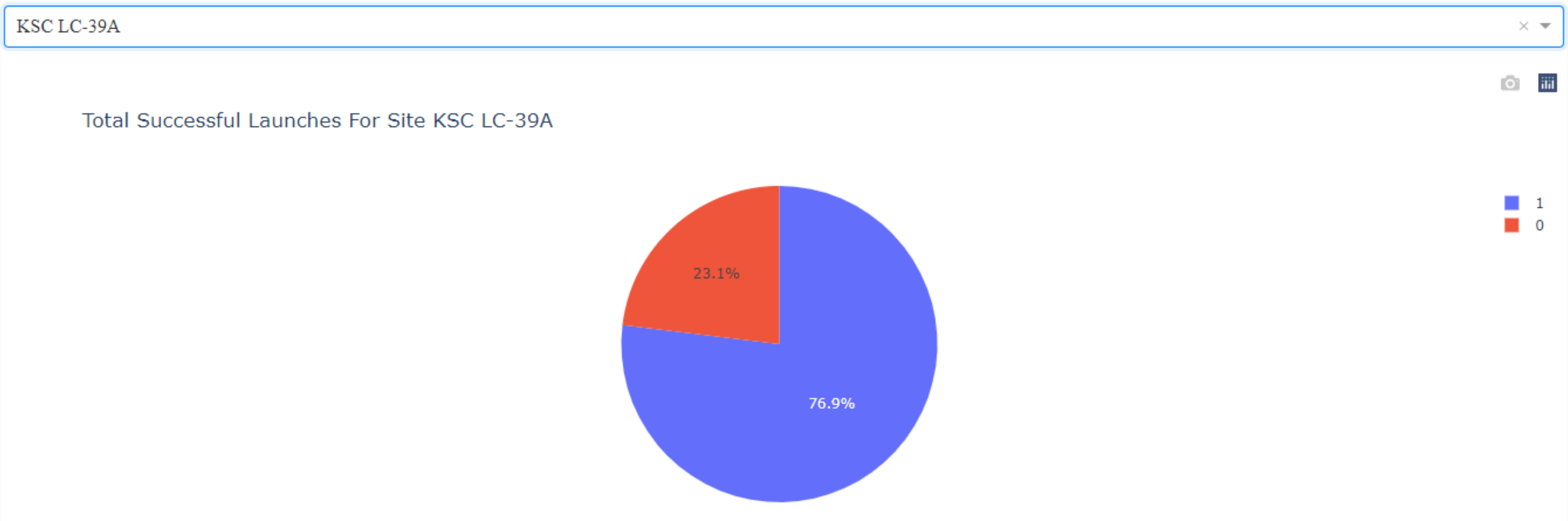
Build a Dashboard with Plotly Dash

Successful Launches by Site



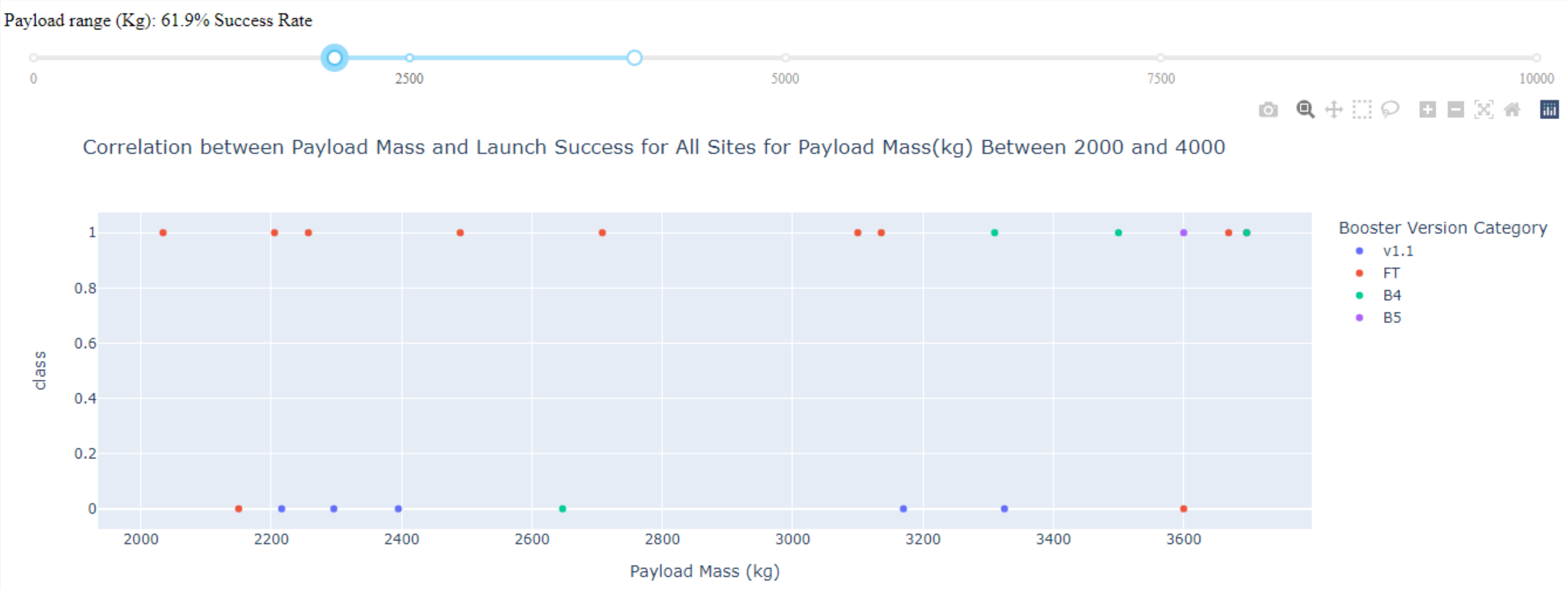
- Site KSC LC-39A has the largest successful launches as well the highest launch success rate.
- More investigation may be needed to determine why KSC LC-39A is the preferred launch site.

Total Successful Launches for Site KSC LC-39A



- As we can see, 76.9% of the total launches at site KSC LC-39A were successful. This is a the highest success rate of all the different launch sites.
- However, this success rate was only around 3% higher than the runner up; site CCAFS LC-40.

Payload Mass vs. Launch Success for All Sites



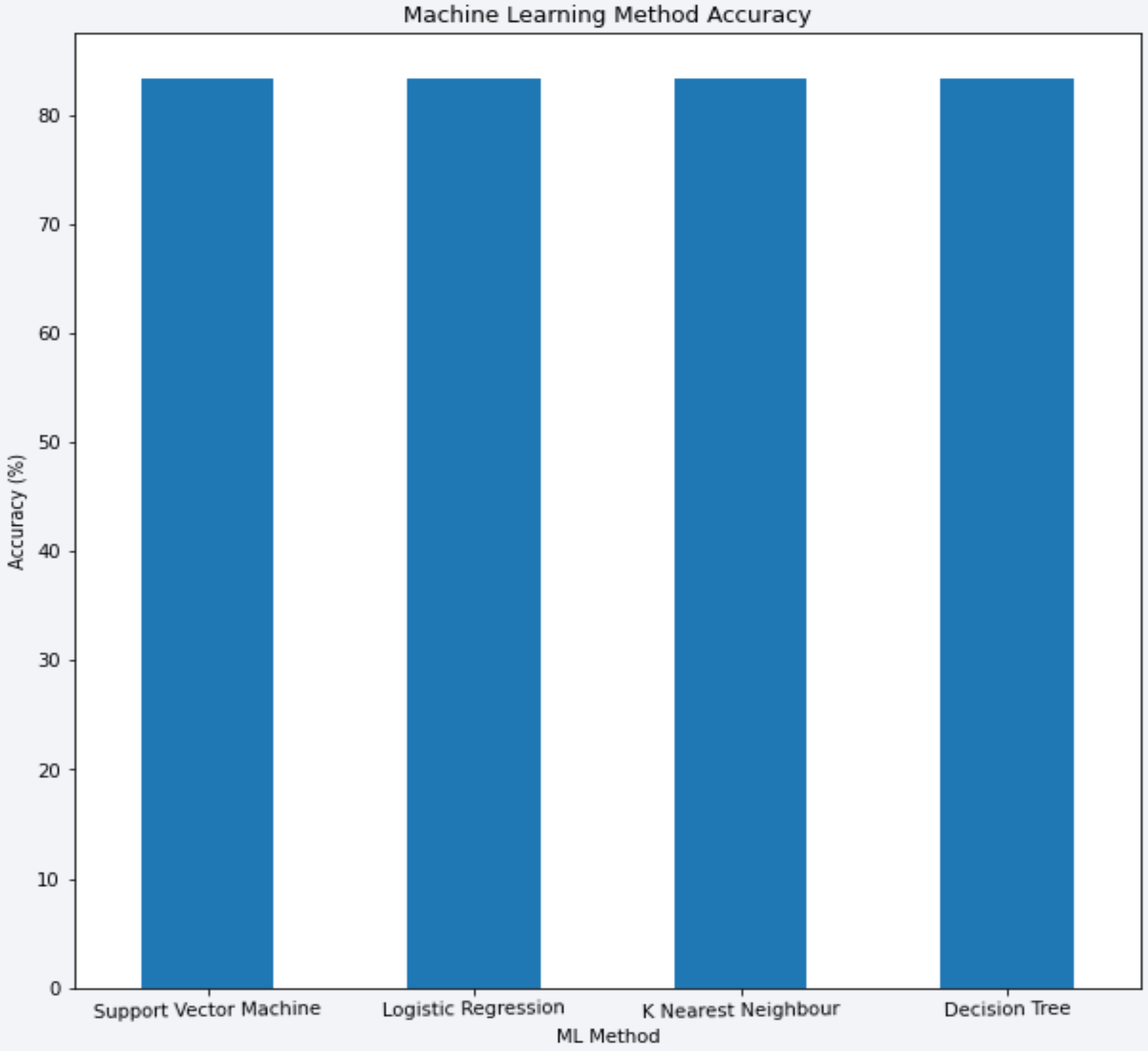
- It appears that the payload range between 2000 kg and 4000 kg has the highest success rate.
- The launch success rate was also dramatically low between the payload range of 0kg and 2500kg. Perhaps very low masses decrease launch success.
- The booster version **FT**, seems to have a higher success rate than other booster versions

Section 5

Predictive Analysis (Classification)

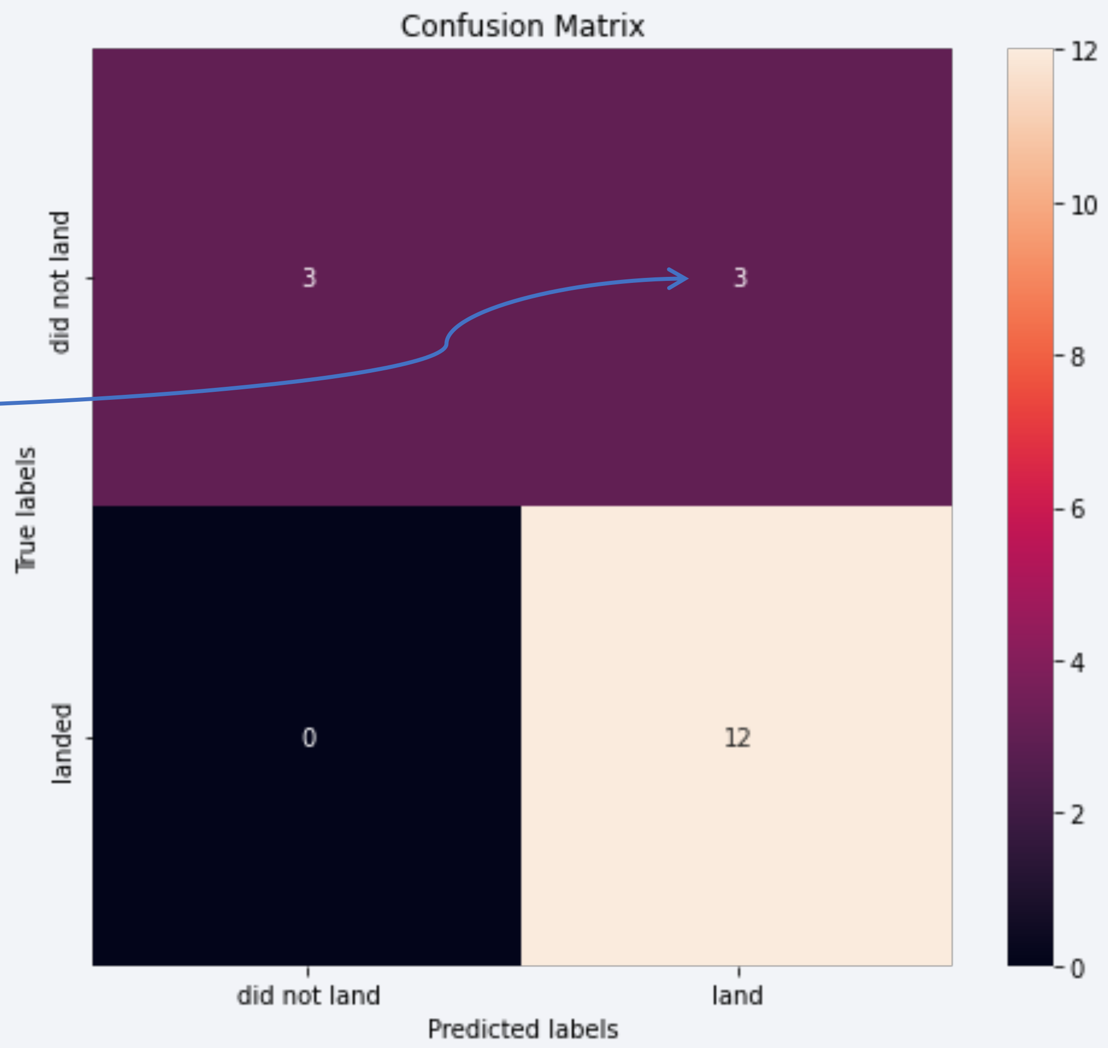
Classification Accuracy

- Since all the methods have an identical accuracy score of 83.33%, we decided to use Logistic Regression for the classification



Confusion Matrix

- The chart shows the confusion matrix of the Logistic Regression model that was chosen.
- The model only failed to accurately predict 3 labels.



Conclusions

- The most successful launches seem to originate from site KSC LC-39A and have payloads of over 8,000 kg.
- VLEO orbit appears to be the best choice with a demonstrated high success rate
- More recent launches (since 2013) have low failure rates
- Launch sites tend to be clustered close to the equator and close to railway lines and highways

Thank you!