IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

M Shahzad Ashraf
April 24, 2024

Outline

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

Github URL:
- Shazzy3204/Applied-Data-Science-Capstone (github.com)
- Please Note all the URLs in this presentation are clickable.

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

Introduction

Background and Context

- In this capstone, we will predict if the Falcon 9 first stage will land successfully. 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. In this module, you will be provided with an overview of the problem and the tools you need to complete the course. Space Y wants to compete with Space X

Problem at Hand

- Space Y tasks us to train a machine learning model to predict successful Stage 1 recovery

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- 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 using various methods
- Data collection was done using get request to the SpaceX API.
- Next, we decoded the response content as a Json using .json() function call and turn it into a pandas dataframe using .json_normalize().
- We then cleaned the data, checked for missing values and fill in missing values where necessary.
- In addition, we performed web scraping from Wikipedia for Falcon 9 launch records with BeautifulSoup.
- The objective was to extract the launch records as HTML table, parse the table and convert it to a pandas dataframe for future analysis.

Data Collection – SpaceX API

Request rocket launch data from SpaceX API using URL → **Decode the response content as a Json using and convert it into a pandas dataframe** → **Fixed information in correct columns (using previously defined functions)** → **Construct dataset using the data we have obtained**

Github URL: Applied-Data-Science-Capstone/jupyter-labs-spacex-data-collection-api.ipynb at main · Shazzy3204/Applied-Data-Science-Capstone (github.com)

Data Collection - Scraping

Request the Falcon9 Launch Wiki page from its URL

→

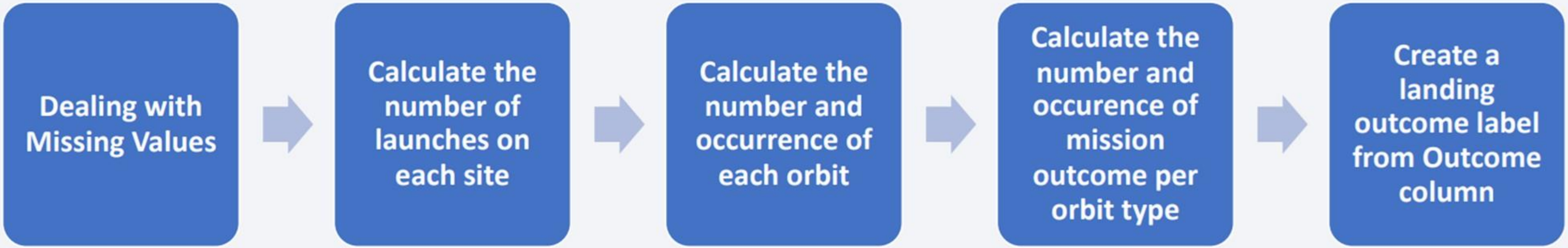
Extract all column/variable names from the HTML table header

→

Create a data frame by parsing the launch HTML tables

Github URL: Applied-Data-Science-Capstone/jupyter-labs-webscraping.ipynb at main · Shazzy3204/Applied-Data-Science-Capstone (github.com)

Data Wrangling



[Github URL: Applied-Data-Science-Capstone/labs-jupyter-spacex-Data wrangling.ipynb at main · Shazzy3204/Applied-Data-Science-Capstone (github.com)]

EDA with Data Visualization

- The graphs used are
  - Scatter Plot
  - Bar chart
  - Line chart

  because they are the ones that best highlight the relationships between the variables considered

Github URL:
[Applied-Data-Science-Capstone/edadataviz.ipynb at main · Shazzy3204/Applied-Data-Science-Capstone (github.com)]

EDA with SQL

- **SQL queries performed**
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string ‘CCA’
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in the ground pad was achieved .
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
  - List of total number of successful and failure mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass.
  - List the failed landing_outcomes in drone ship,their booster versions, and launch site names for in year 2015
- GITHUB URL:

[Applied-Data-Science-Capstone/jupyter-labs-eda-sql-coursera_sqllite.ipynb at main · Shazzy3204/Applied-Data-Science-Capstone (github.com)]

Build an Interactive Map with Folium

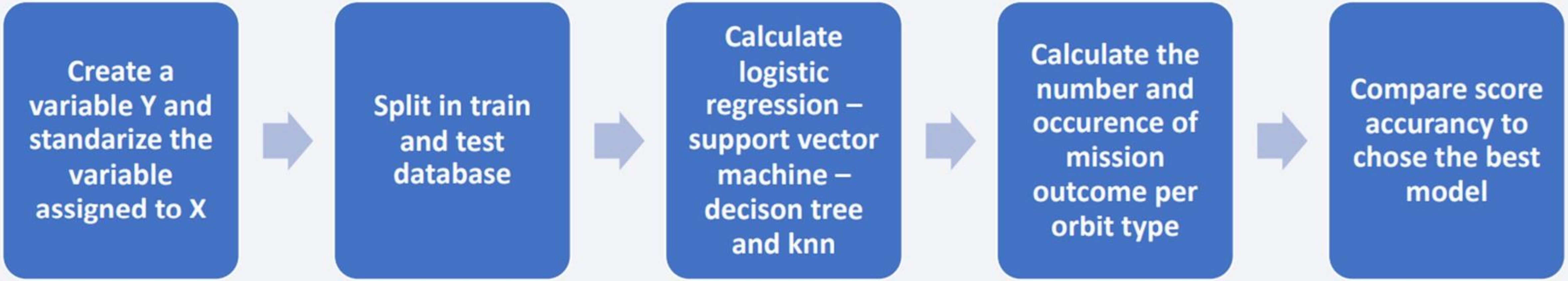
- Summary of map objects:
- Markers: Show a geo location from latitude and longitude data
- Cluster: Show a group of markers
- Circles: Show a single location
- Lines: Show distance between two

- GITHUB URL: Applied-Data-Science-Capstone/lab_jupyter_launch_site_location.ipynb at main · Shazzy3204/Applied-Data-Science-Capstone (github.com)

Build a Dashboard with Plotly Dash

- Summary of plots:
- Bar: Show categories differences
- Line: Reports time series changes
- Pie: Shows the percentage of events
- Tree: Shows complex relationship of variables in interactive way
- Map: Shows variables of states on a map

Predictive Analysis (Classification)



GITHUB URL: [Applied-Data-Science-Capstone/SpaceX_Machine Learning Prediction.ipynb at main · Shazzy3204/Applied-Data-Science-Capstone (github.com)]

Results

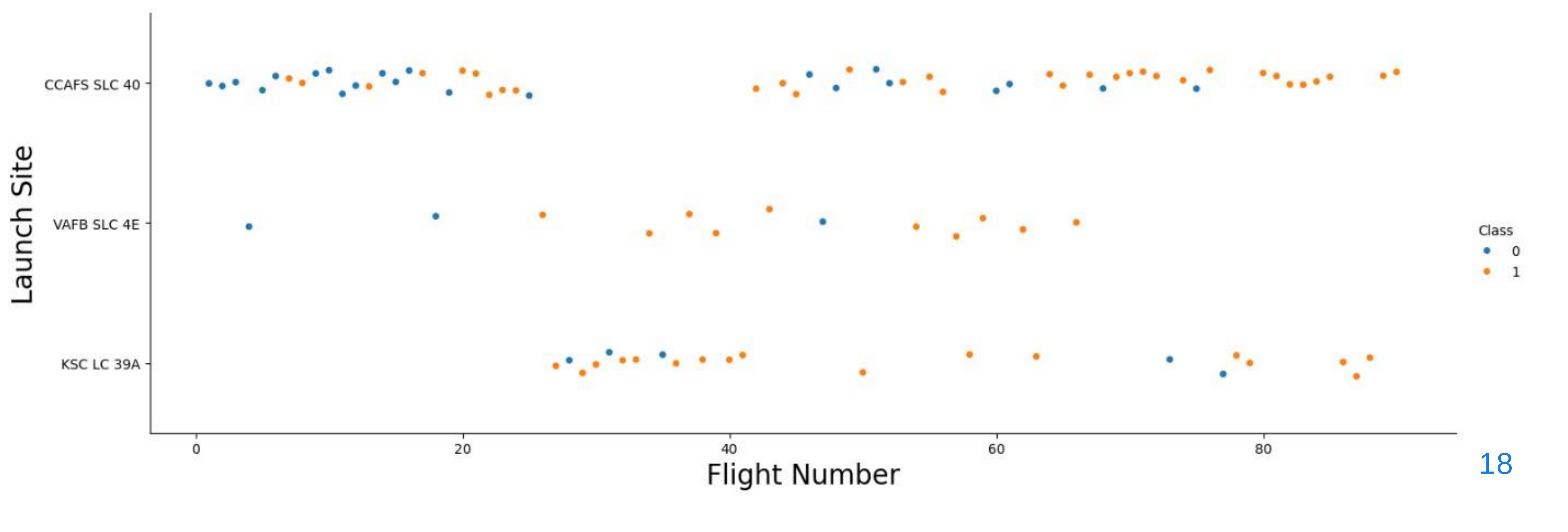
- Exploratory data analysis results
  - Both API and web scraping are capable to collect Xspace data
- Interactive analytics demo in screenshots
  - EDA with SQL is effective for data filtering
  - EDA with interactive visualization provides informative information
  - Plotly Dash is powerful to show instant data change
- Predictive analysis results
  - Decision Tree Classifier Algorithm has the best accuracy of predicting.

Section 2

Insights drawn from EDA

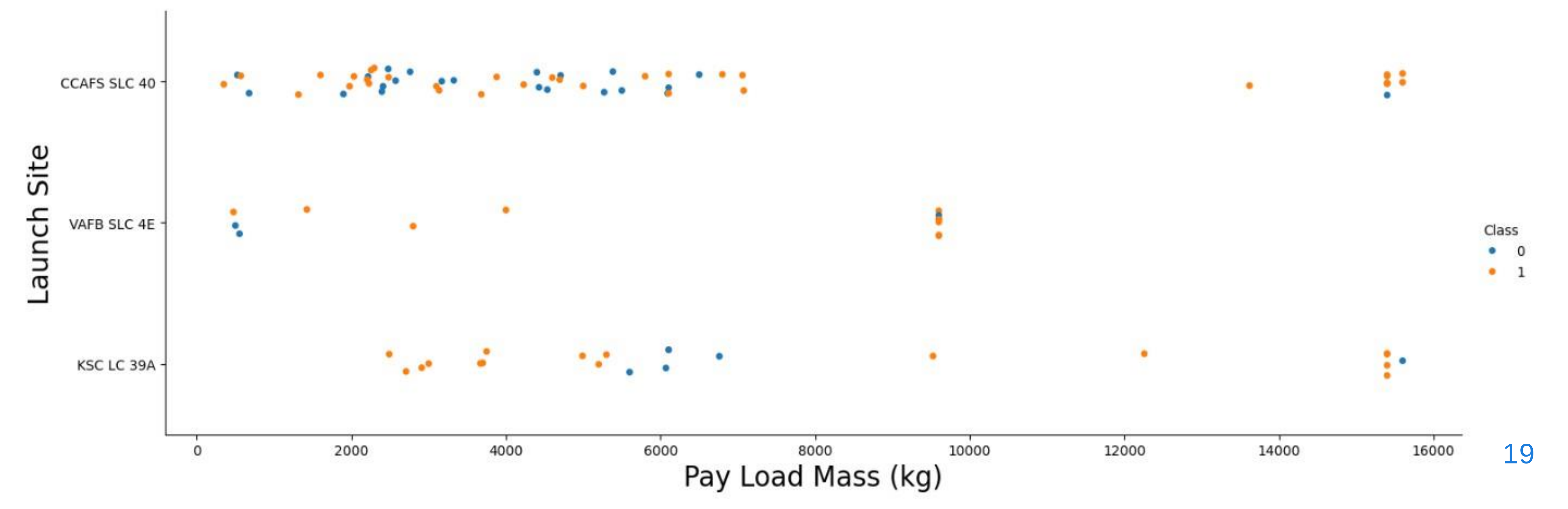
Flight Number vs. Launch Site

- From the plot, we found that the larger the flight amount at a launch site, the greater the success rate at a launch site.



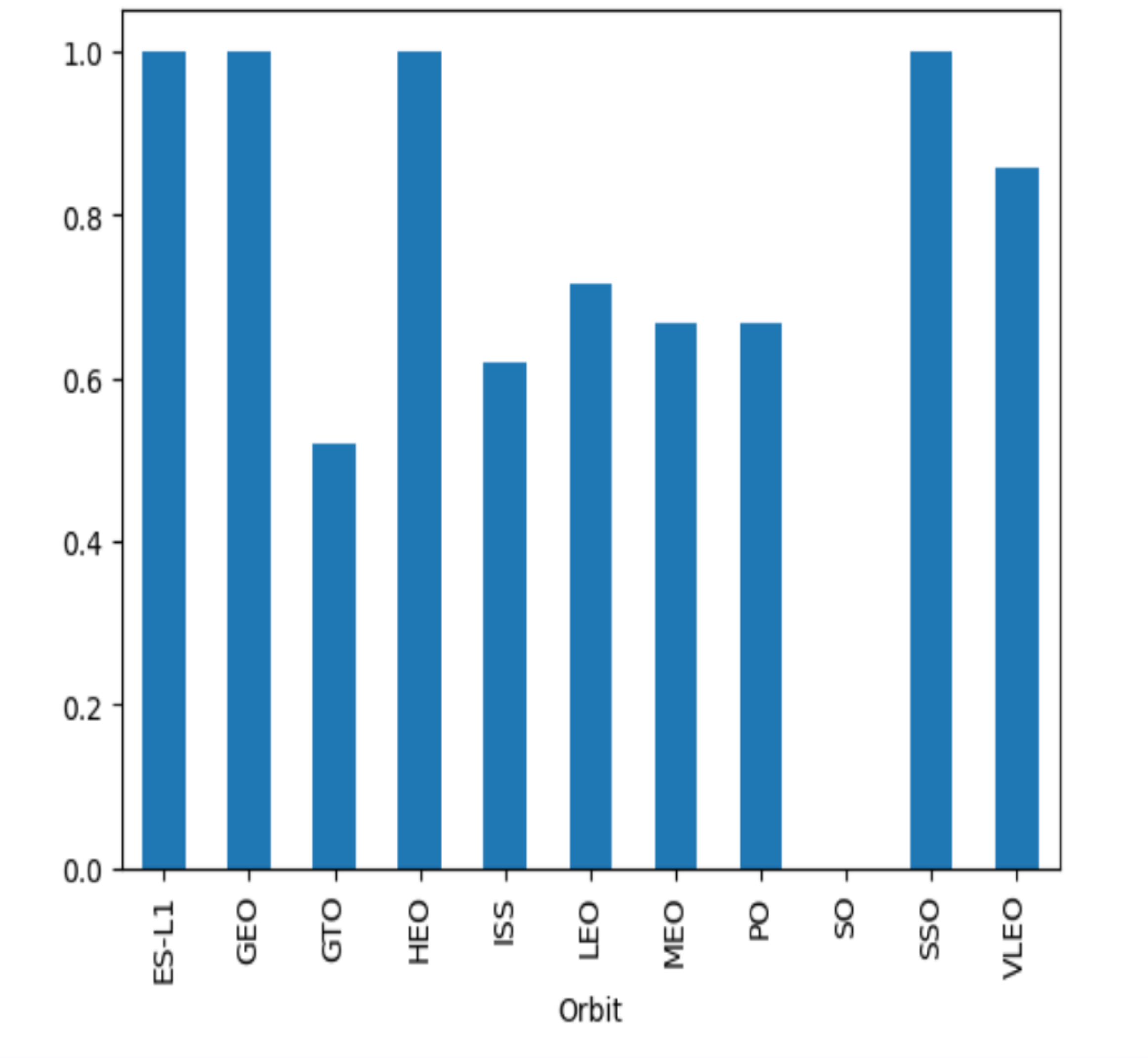
Payload vs. Launch Site

Blue indicates successful launch and orange indicates unsuccessful launch. Payload mass appears to fall mostly between 0-6000 kg.   Different launch sites also seem to use different payload mass.



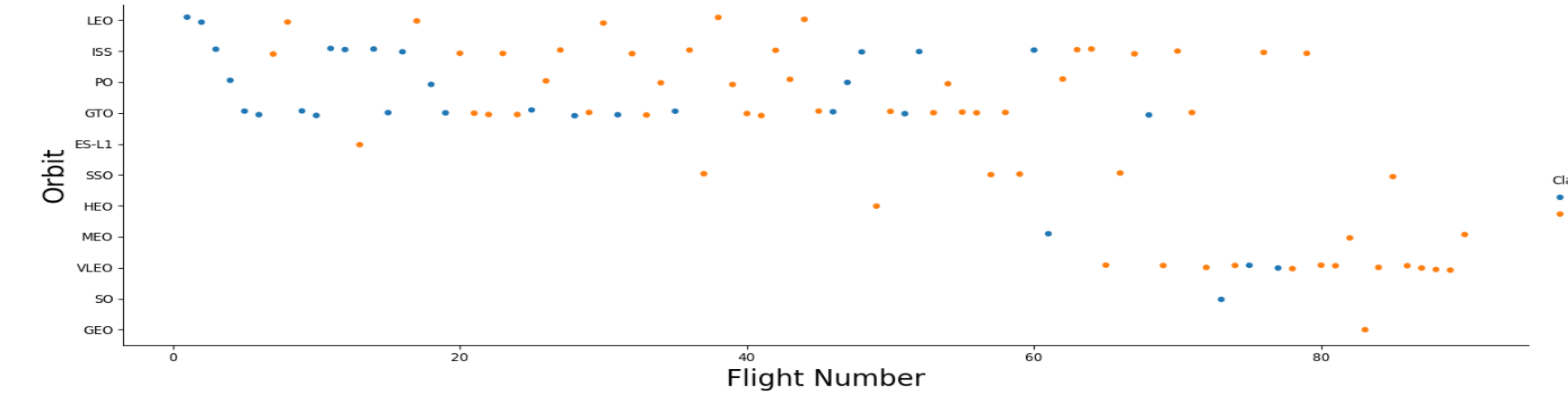
Success Rate vs. Orbit Type

- ES-L1 (1), GEO (1), HEO (1) have 100% success rate (sample sizes in parenthesis) SSO (5) has 100% success rate
- VLEO (14) has decent success rate and attempts
- SO (1) has 0% success rate
- GTO (27) has the around 50% success rate but largest sample



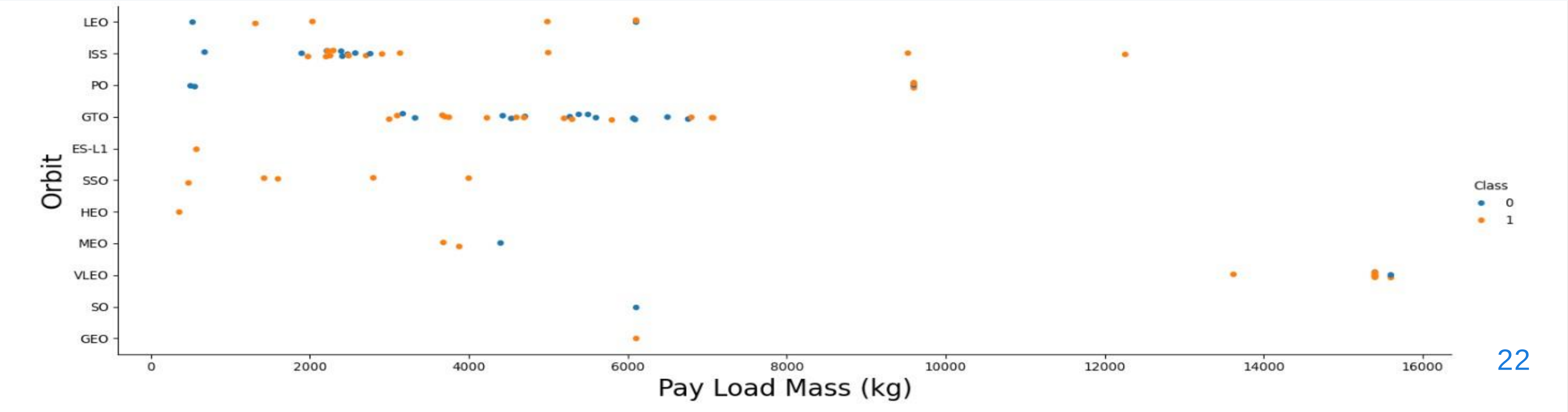
Flight Number vs. Orbit Type

- Blue indicates successful launch and orange indicates unsuccessful launch.
- Launch Orbit preferences changed over Flight Number.
- Launch Outcome seems to correlate with this preference.
- SpaceX started with LEO orbits which saw moderate success LEO and returned to VLEO in recent launches
- SpaceX appears to perform better in lower orbits or Sun-synchronous orbits



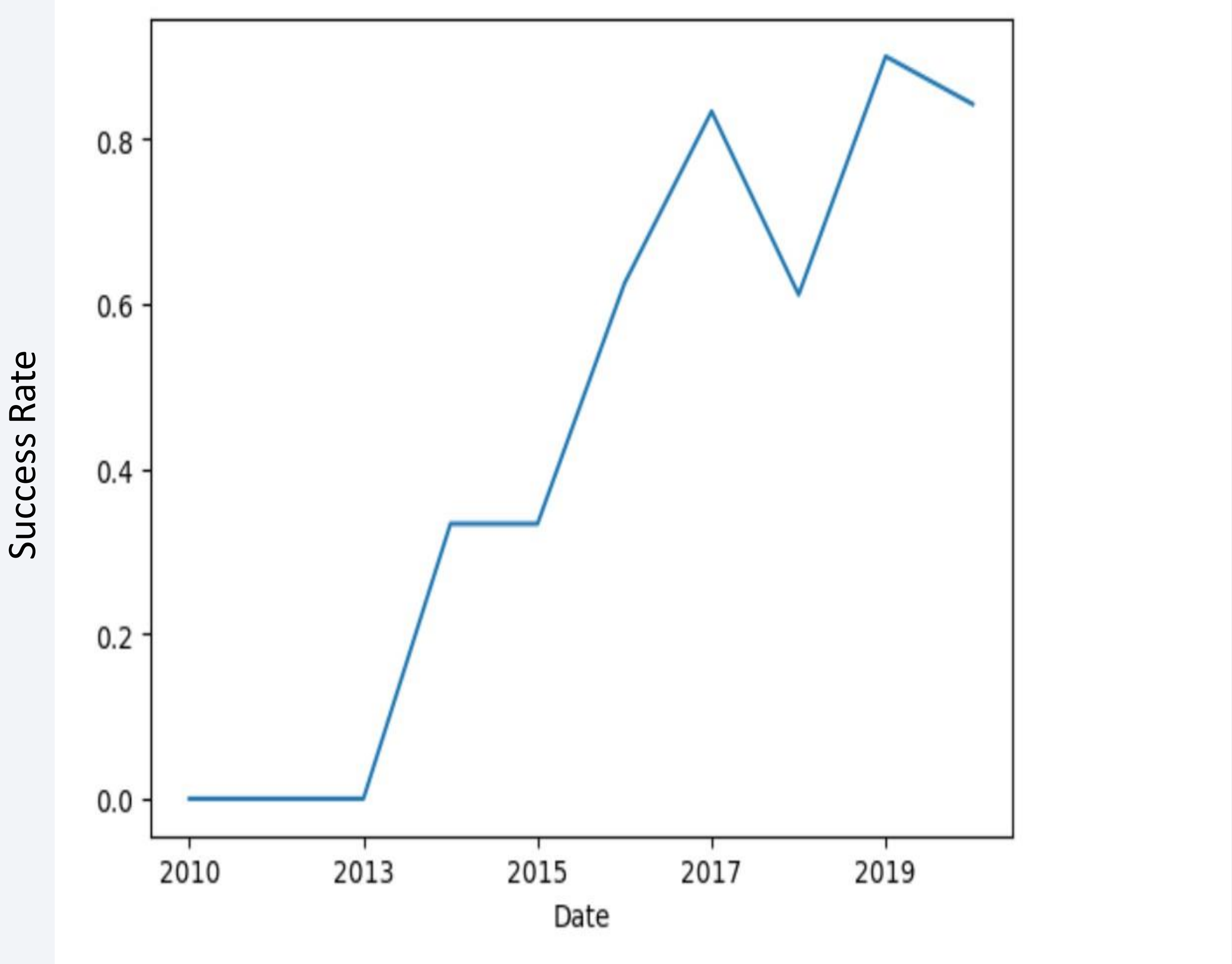
Payload vs. Orbit Type

- Blue indicates successful launch and orange indicates unsuccessful launch.
- Payload mass seems to correlate with orbit
- LEO and SSO seem to have relatively low payload mass
- The other most successful orbit VLEO only has payload mass values in the higher end of the range



Launch Success Yearly Trend

- Success generally increases over time since 2013 with a slight dip in 2018
- Success in recent years at around 80%



All Launch Site Names

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

```
%sql select distinct launch_site from SPACEXTBL;
```

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

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

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXTBL 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

Total Payload Mass

```
%sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';
```

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

total_payload_mass
45596

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
```

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

AVG_PAYLOAD
2928.4

First Successful Ground Landing Date

```
%sql select min(date) as first_successful_landing from SPACEXDATASET where Landing_Outcome = 'Success (ground pad)';
```

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

first_successful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- %sql select booster_version from SPACEXDATASET where Landing_Outcome = 'Success (drone ship)' and payload_mass_kg_ between 4000 and 6000;

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

Total Number of Successful and Failure Mission Outcomes

```
%sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;
```

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

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

Boosters Carried Maximum Payload

```
%sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET);
```

```
 * 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

2015 Launch Records

%sql SELECT date, substr(Date, 6, 2) AS Month, BOOSTER_VERSION, LAUNCH_SITE
 FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Failure (drone ship)' AND
substr(Date,0,5)='2015';

Date	Month	Booster_Version	Launch_Site
2015-01-10	01	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	04	F9 v1.1 B1015	CCAFS LC-40

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

```
%%sql select Landing_Outcome, count(*) as count_outcomes from SPACEXDATASET
      where date between '2010-06-04' and '2017-03-20'
      group by Landing_Outcome
      order by count_outcomes desc;
```

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

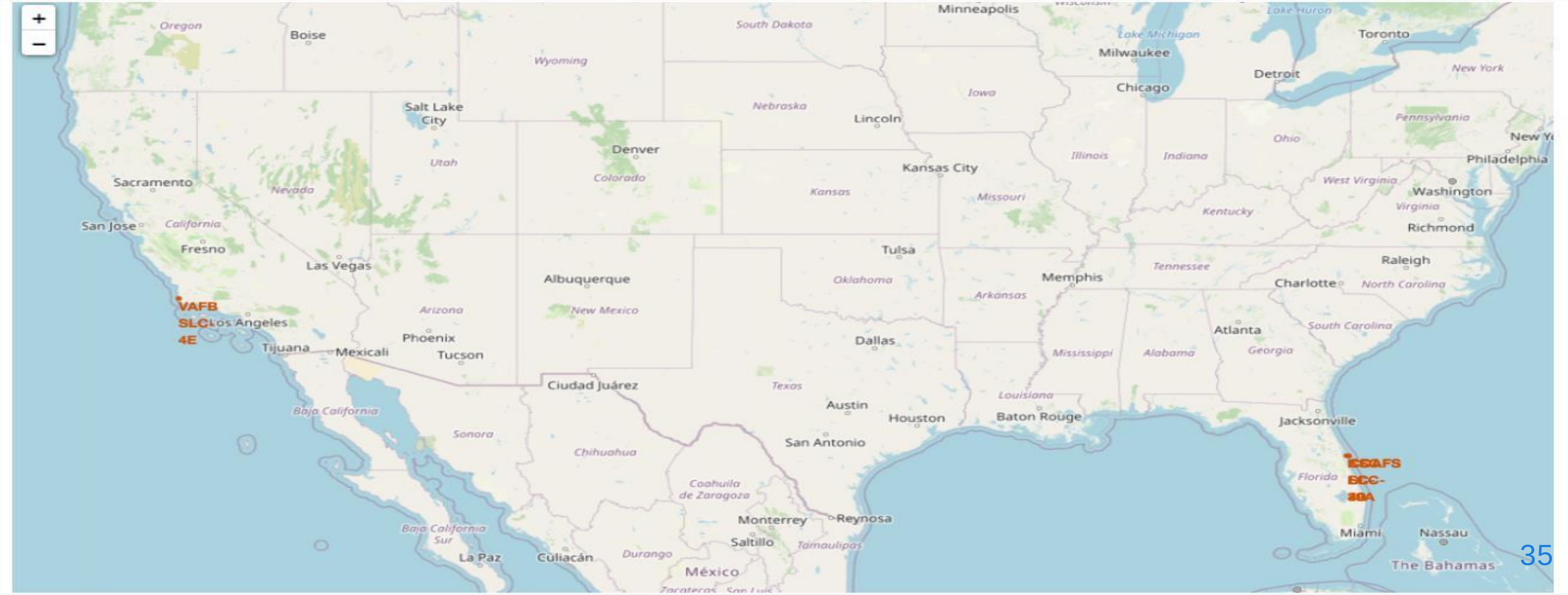
Landing_Outcome	count_outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

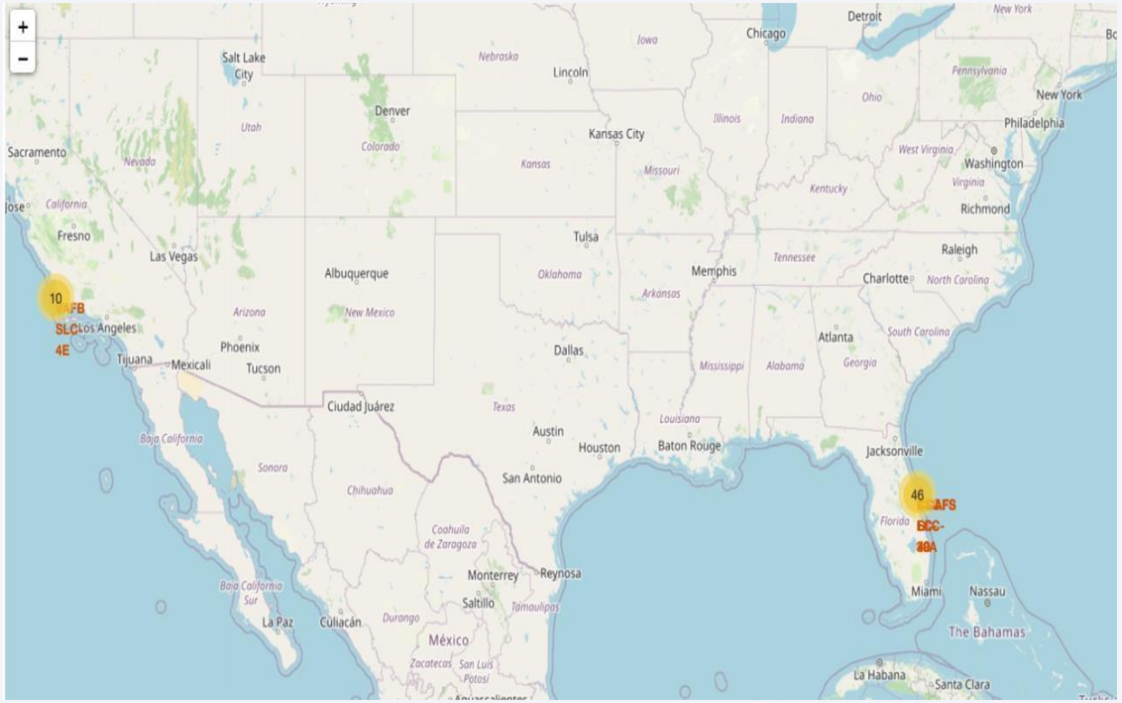
Launch Sites Proximities Analysis

All Launch Sites’ Location Markers

All the launches are near USA, Florida, and California

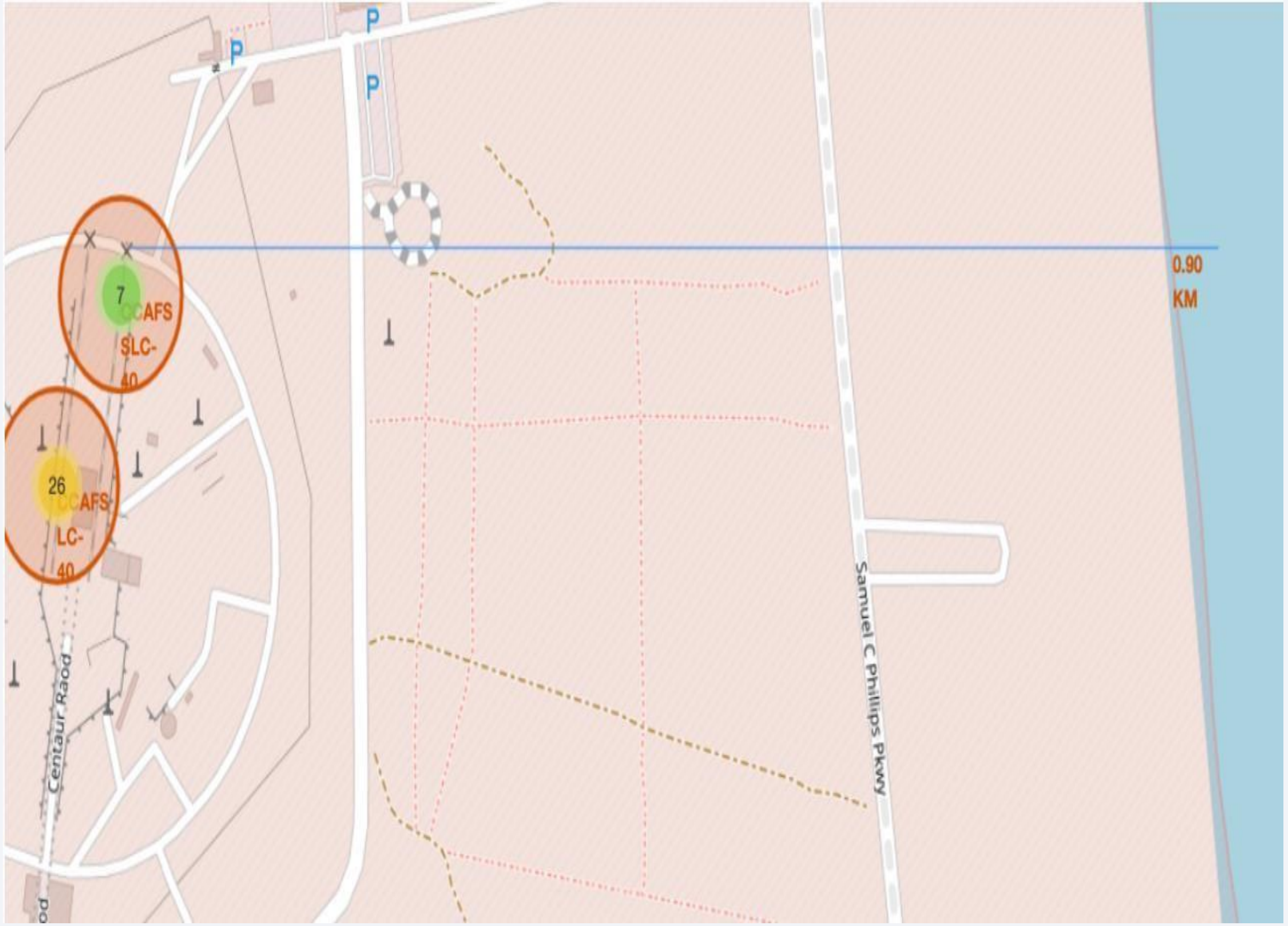


Color-labeled Launch Outcomes



Green means successful
Red means Failure

Launch Sites to its Proximities

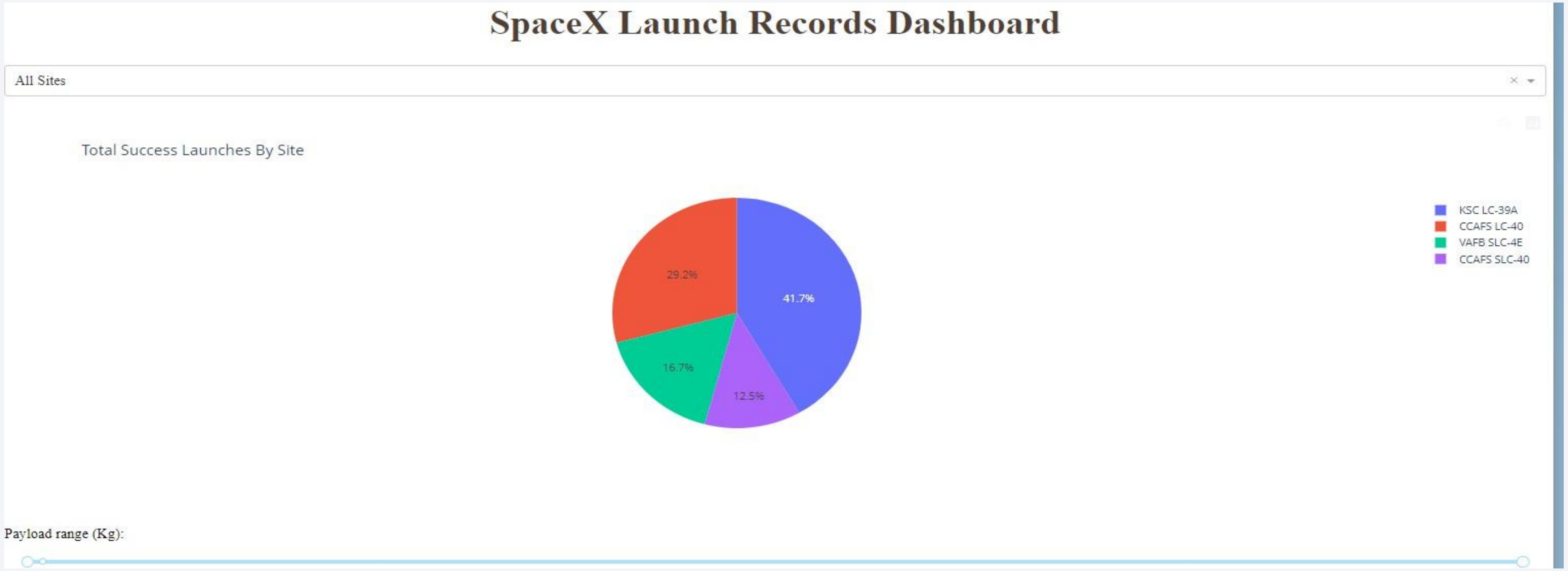


All distances from launch sites to its proximities, they weren't far from railway tracks.

Section 4

Build a Dashboard with Plotly Dash

Launch Success Count

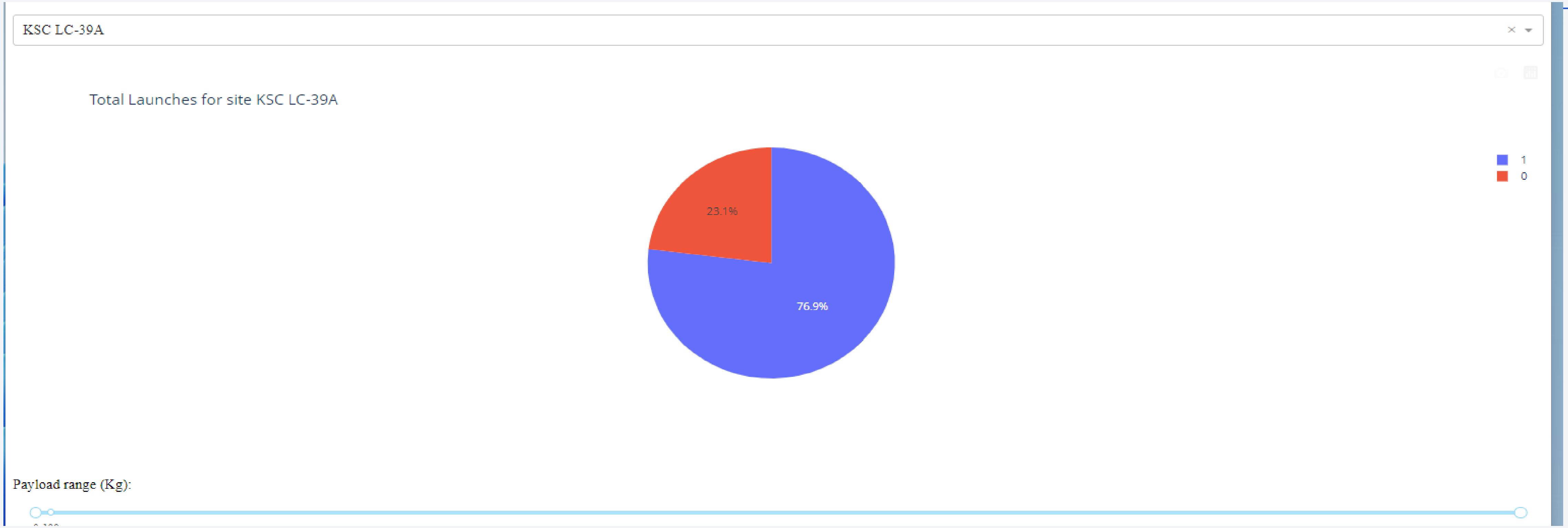


KSC LC-39A has the highest success score with 41.7%

CCAFS LC-40 comes next with 29.2%

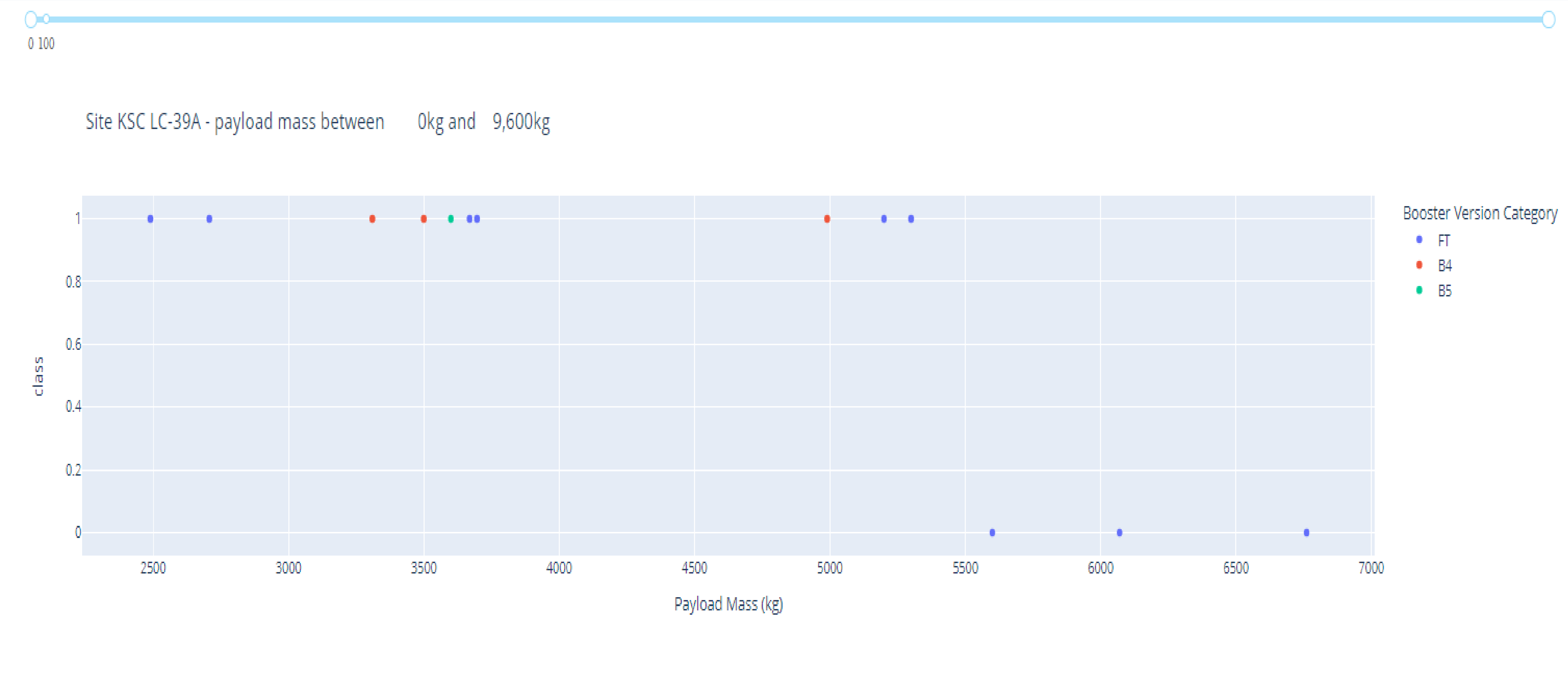
Finally, VAFB SLC-4E and CCAFS SLC-40 with 16.7% and 12.5% respectively

Launch Site with Highest Score



KSC LC-39A has the highest score with 76.9% with payload range of 2000 kg – 10000 kg, and FT booster version has the highest score

Payload vs. Launch Outcome



Section 5

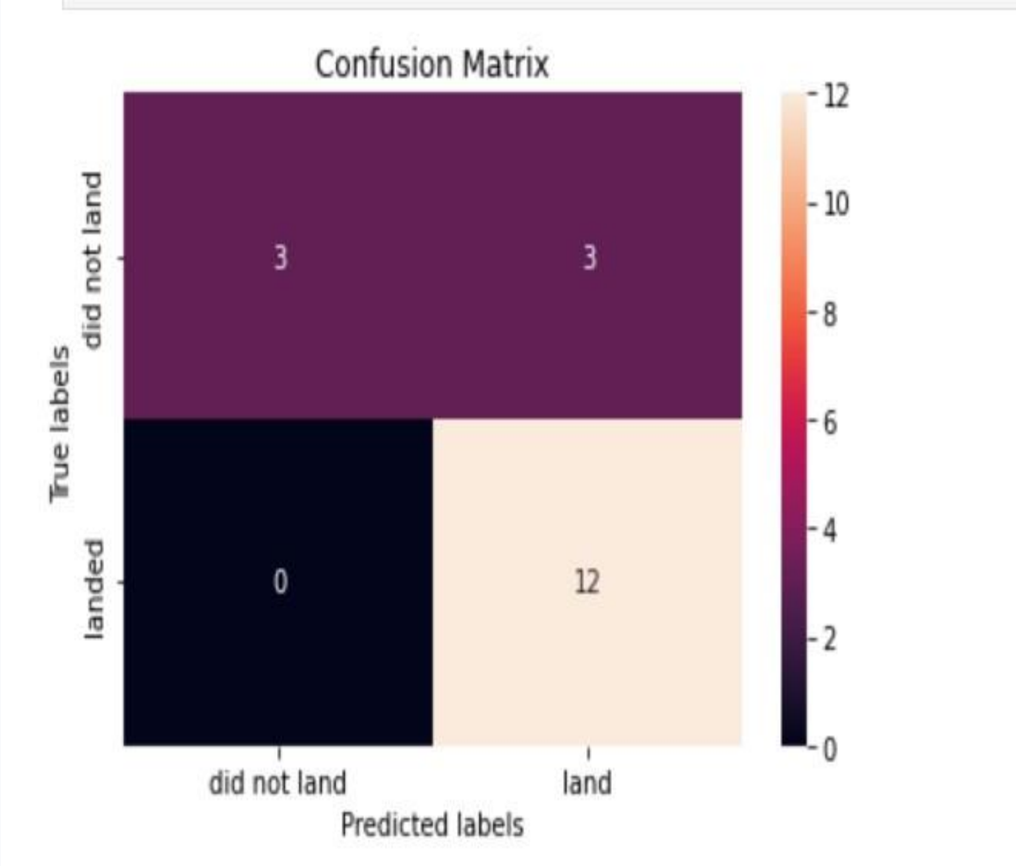
Predictive Analysis (Classification)

Classification Accuracy

All models had virtually the same accuracy on the test set at 83.33% accuracy. It should be noted that test size is small at only sample size of 18.

This can cause large variance in accuracy results, such as those in Decision Tree Classifier model in repeated runs.
We likely need more data to determine the best model.

Confusion Matrix



- Since all models performed the same for the test set, the confusion matrix is the same across all models. The models predicted 12 successful landings when the true label was successful landing.
- The models predicted 3 unsuccessful landings when the true label was unsuccessful landing.
- The models predicted 3 successful landings when the true label was unsuccessful landings (false positives). Our models over predict successful landings.

Conclusions

We can conclude that:

- The larger the flight amount at a launch site, the greater the success rate at a launch site.
- Launch success rate started to increase in 2013 till 2020.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- KSC LC-39A had the most successful launches of any sites.
- The Decision tree classifier is the best machine learning algorithm for this task.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!