IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mansi Verma
07-07-2023

Outline

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

Executive Summary

- The data collected included information about SpaceX missions, particularly focusing on the success or failure of landings. A new column called 'class' was created to classify successful landings.
- The analysis began with exploring the data using SQL queries to gain insights. Visualization techniques were applied to understand the data better, including the use of folium maps and dashboards.
- To prepare the data for machine learning, relevant columns were selected as features. Categorical variables were transformed into binary values using one-hot encoding. The data was standardized to ensure consistent scaling across features.
- GridSearchCV, a method for hyperparameter tuning, was employed to find the best parameters for four machine learning models: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors.
- After training the models, it was found that all four models achieved similar results with an accuracy rate of around 83.33%. However, it was observed that all the models tended to overpredict successful landings.
- It was concluded that more data is needed to improve the model's determination and accuracy. Obtaining additional data points could potentially help in achieving more accurate predictions regarding the success or failure of SpaceX landings.

Introduction

The background context of the problem revolves around the commercial space industry, with a specific focus on two companies: SpaceX and Space Y. The commercial space age is described as being in full swing, and SpaceX is noted for having the best pricing compared to its competitors, such as Space Y. The pricing advantage is attributed to SpaceX's ability to recover a part of the rocket, specifically Stage 1, which helps reduce costs.

The problem at hand is that Space Y wants to compete with SpaceX in terms of successfully recovering Stage 1 of their rockets. To achieve this, they have tasked the team with training a machine learning model that can predict the success of Stage 1 recovery. Presumably, this prediction can assist in optimizing and improving the recovery process for Space Y's rockets, potentially reducing costs and enhancing competitiveness.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Combined data from SpaceX public API and SpaceX Wikipedia page
- Perform data wrangling
  - Classifying true landings as successful and unsuccessful otherwise
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Tuned models using GridSearchCV

Data Collection

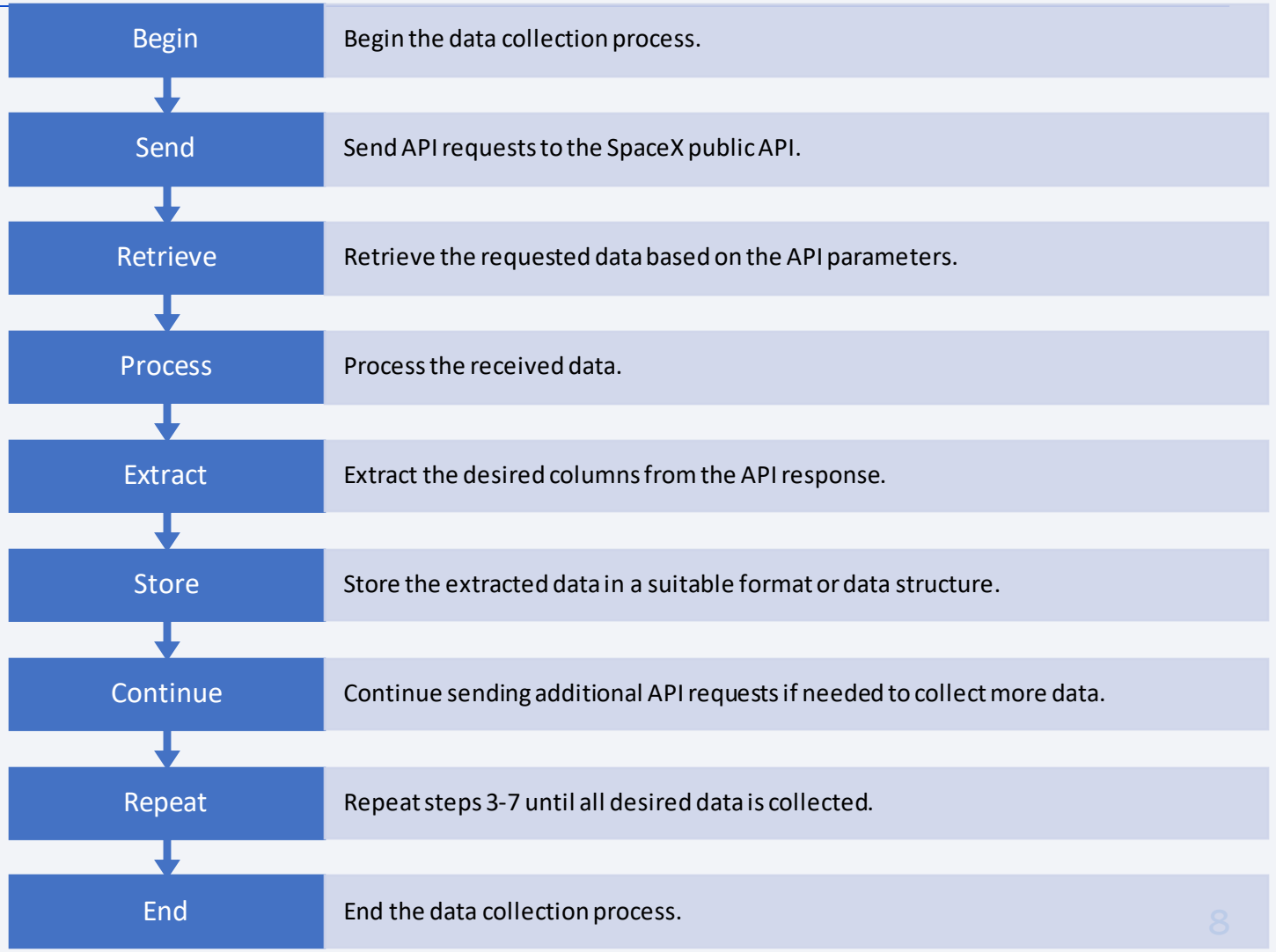
The data collection process for this project involved a combination of API requests from SpaceX's public API and web scraping data from a table in SpaceX's Wikipedia entry.

For the API data collection, the following columns were retrieved: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, and Latitude. These columns provide information about the flight details, booster characteristics, payload, launch outcome, and landing information.

Regarding the web scraping data collection, a table from SpaceX's Wikipedia entry was scraped to obtain the following columns: Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, and Time. These columns provide similar information to the API data, including flight details, payload, orbit, launch outcome, booster version, and landing information.

Data Collection – SpaceX API

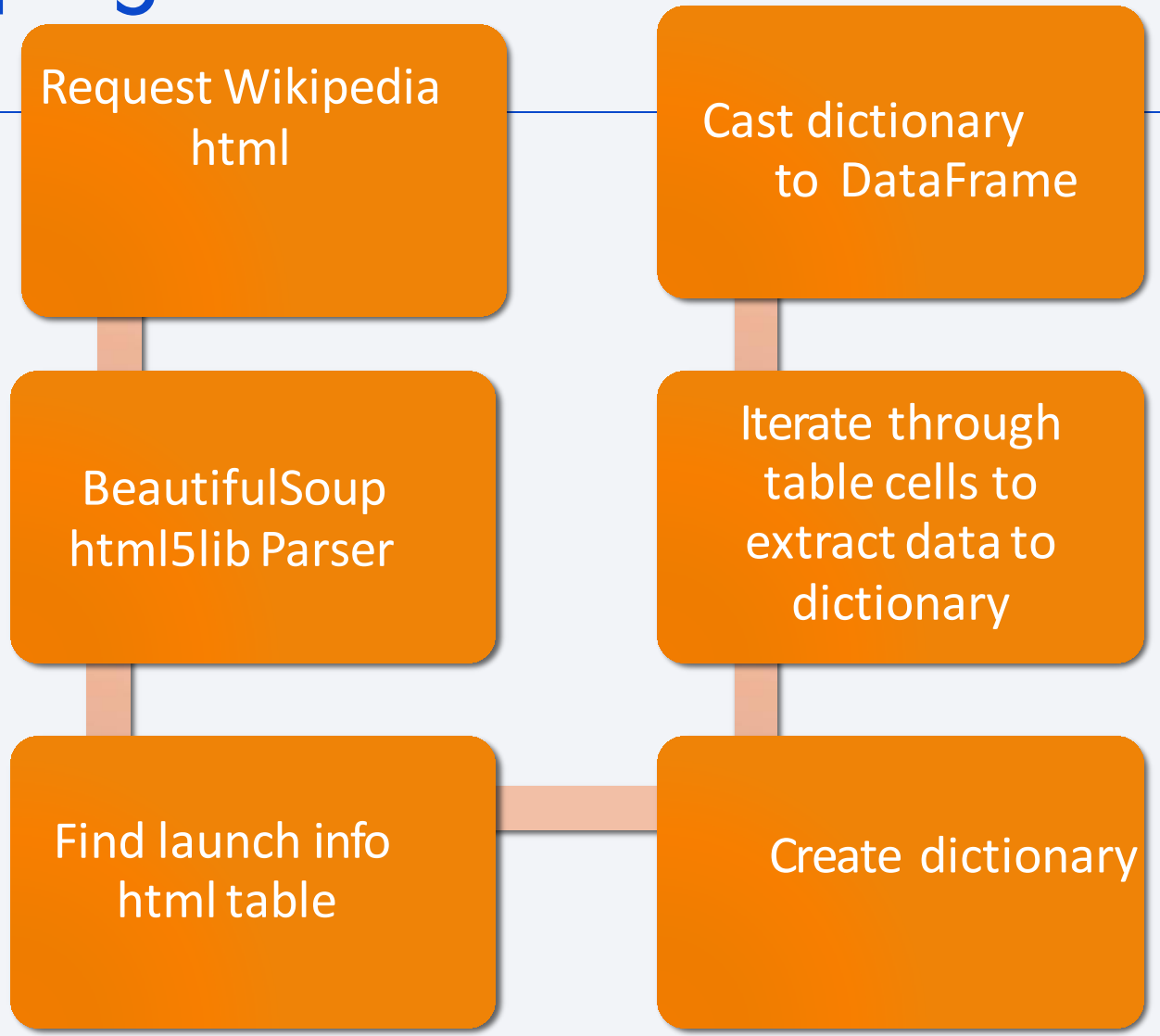
- Add the GitHub URL of the completed SpaceX API calls notebook (must include completed code cell and outcome cell), as an external reference and peer-review purpose



Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts

- Add the GitHub URL of the completed web scraping notebook, as an external reference and peer-review purpose



Data Wrangling

- Create a training label with landing outcomes where successful = 1 & failure = 0.
- Outcome column has two components: ‘Mission Outcome’ ‘Landing Locatio n’
- New training label column ‘class’ with a value of 1 if ‘Mission Outcome’

  is True and 0 otherwise. Value Mapping:
- True ASDS, True RTLS, & True Ocean – set to -> 1
- None None, False ASDS, None ASDS, False Ocean, False RTLS – set to -> 0

EDA with Data Visualization

The exploratory data analysis (EDA) focused on several variables: Flight Number, Payload Mass, Launch Site, Orbit, Class (success/failure), and Year. The goal was to investigate the relationships between these variables and determine if any patterns or trends could be identified for potential use in training the machine learning model.

- Various plots were used to visualize these relationships. The plots included:
- Flight Number vs. Payload Mass: This scatter plot aimed to observe any correlation between the flight number and payload mass, examining if there were any trends or patterns.
- Flight Number vs. Launch Site: This scatter plot visualized the relationship between the flight number and launch site, allowing for a comparison of launches from different sites.
- Payload Mass vs. Launch Site: This scatter plot depicted the payload mass in relation to the launch site, providing insights into the distribution of payload masses across different launch sites.
- Orbit vs. Success Rate: A bar plot or similar visualization was used to show the success rate for different orbits. This analysis aimed to determine if the success rate varied based on the target orbit.
- Flight Number vs. Orbit: This line chart showcased the flight number over time, highlighting the associated orbit. It helped identify any patterns or trends in the selection of orbits for different flights.
- Payload vs. Orbit: This scatter plot examined the relationship between the payload and the target orbit, providing insights into the distribution and ranges of payloads for different orbits.
- Success Yearly Trend: A line chart was used to visualize the yearly trend of successful launches, helping identify any significant changes or patterns over time.

.

EDA with SQL

- The loaded dataset was stored in an IBM DB2 Database, which provided a structured and efficient storage system for further analysis. The SQL Python integration was utilized to query the database and gain a better understanding of the dataset.
- Various queries were executed to extract specific information and insights from the dataset. Some of the areas of interest included:
- Launch Site Names: Queries were made to retrieve the names of the launch sites used by SpaceX for their missions. This information could provide insights into the distribution and frequency of launches across different sites.
- Mission Outcomes: Queries were performed to gather information about the outcomes of the missions, aiming to analyze the success rates and identify any patterns or trends in mission success or failure.
- Payload Sizes of Customers: Queries were used to extract data related to the payload sizes of different customers. This analysis could shed light on the distribution of payload sizes and potentially identify any variations based on customer requirements.
- Booster Versions: Queries were executed to obtain information about the different versions of boosters used by SpaceX. This analysis could provide insights into the evolution of booster technology over time.
- Landing: Queries were made to retrieve data related to the landing of boosters, focusing on factors such as landing success rates, landing pad usage, and any other relevant information.

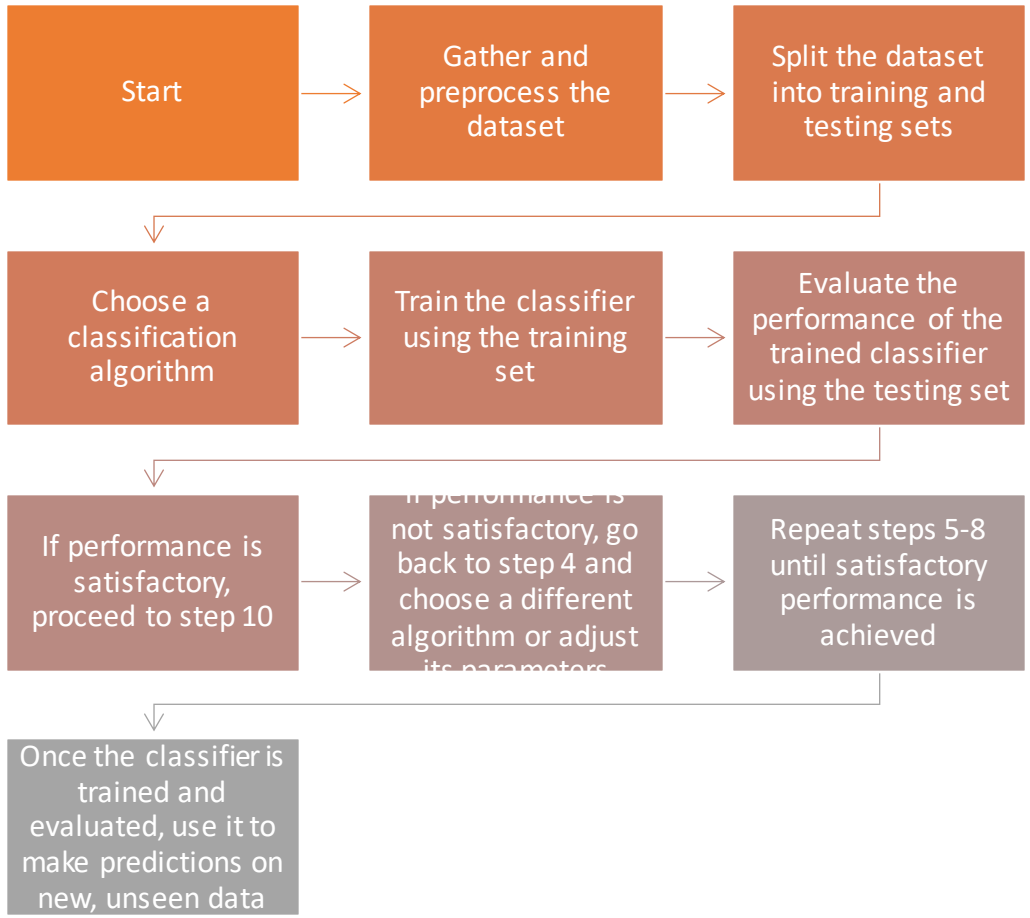
Build an Interactive Map with Folium

- Folium maps were utilized to visualize various aspects related to launch sites, successful and unsuccessful landings, as well as proximity to key locations such as Railway, Highway, Coast, and City. The objective was to gain insights into the reasons behind the selection of specific launch site locations and to visualize the distribution of successful landings in relation to geographical features.
- The following elements were visualized using Folium maps:
- Launch Sites: The maps marked the locations of the launch sites, providing a visual representation of their distribution. This allowed for an understanding of the geographic spread of SpaceX's launch infrastructure.
- Successful and Unsuccessful Landings: The maps differentiated between successful and unsuccessful landings, using markers or color codes to represent the outcomes. This visualization helped identify any spatial patterns or clusters related to the success or failure of landings.
- Proximity to Key Locations: The maps included markers or overlays for key locations such as Railway, Highway, Coast, and City. This demonstrated the proximity of launch sites to these features, providing insights into the factors that influenced the selection of launch site locations.

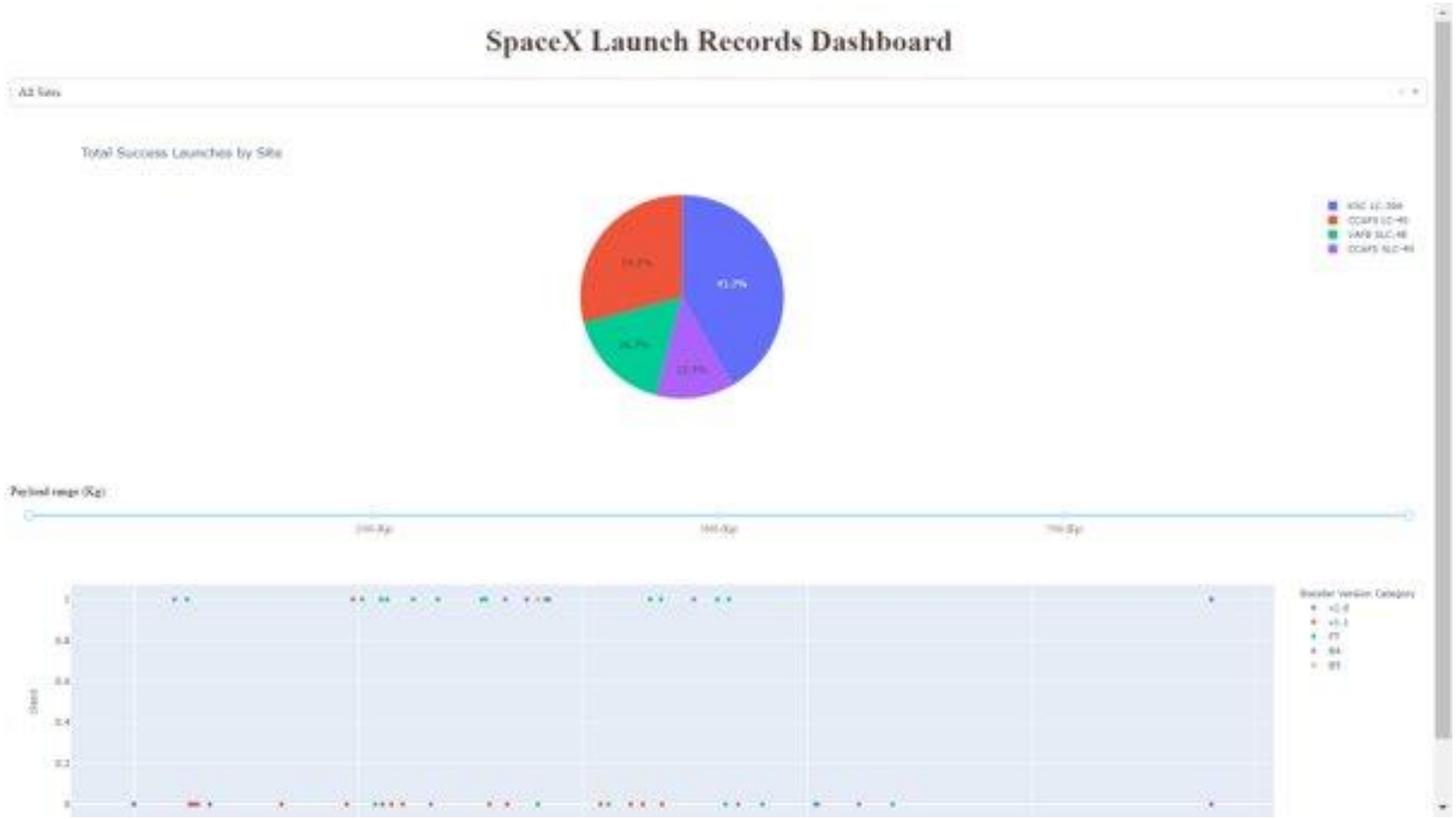
Build a Dashboard with Plotly Dash

- Dashboard includes a pie chart and a scatter plot.
- Pie chart can be selected to show distribution of successful landings across all launch sites and can be selected to show individual launch site success rates.
- Scatter plot takes two inputs: All sites or individual site and payload mass on a slider between 0 and 10000 kg.
- The pie chart is used to visualize launch site success rate.
- The scatter plot can help us see how success varies across launch sites, payload mass, and
- booster version category.

Predictive Analysis (Classification)

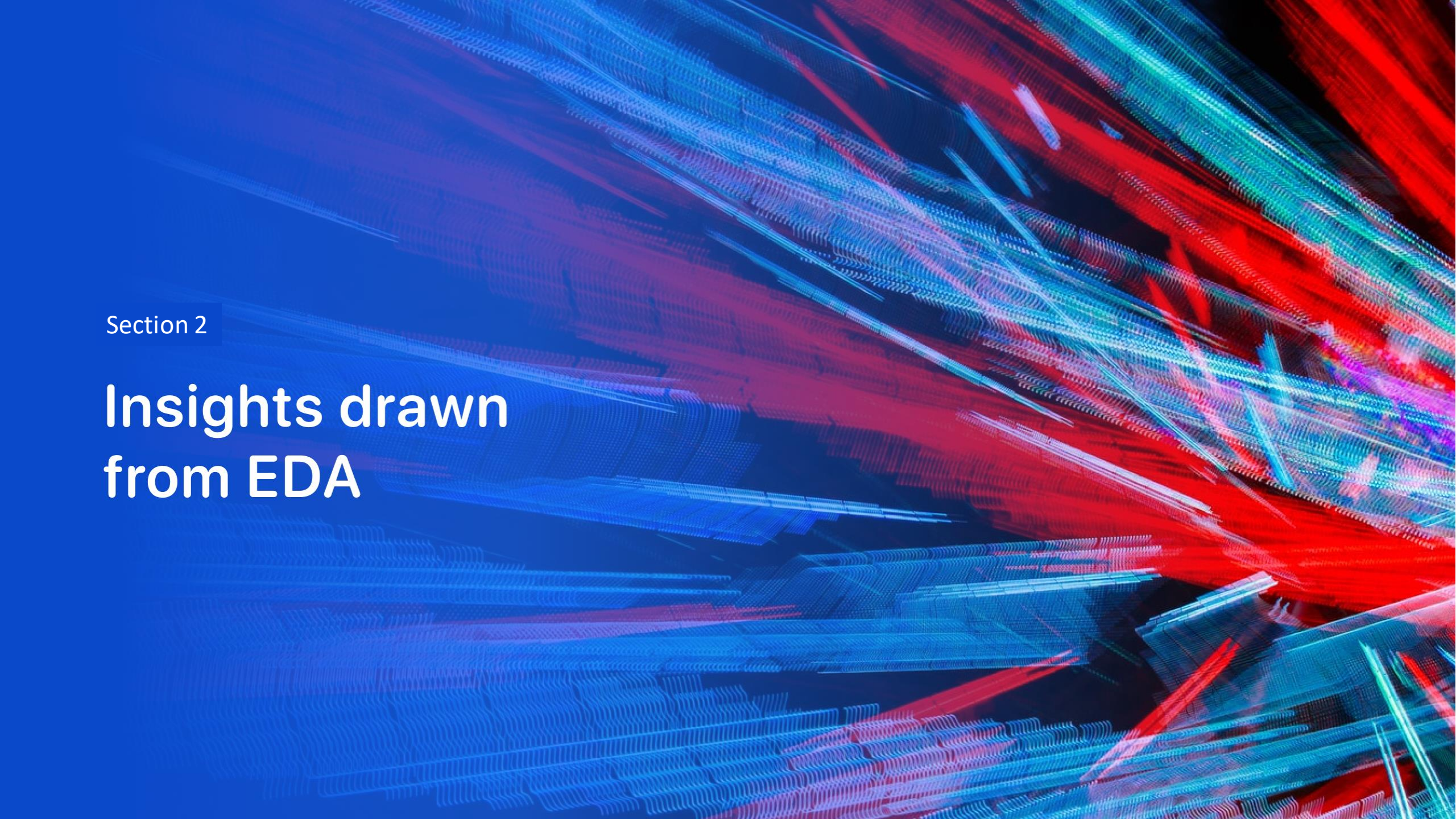


Results

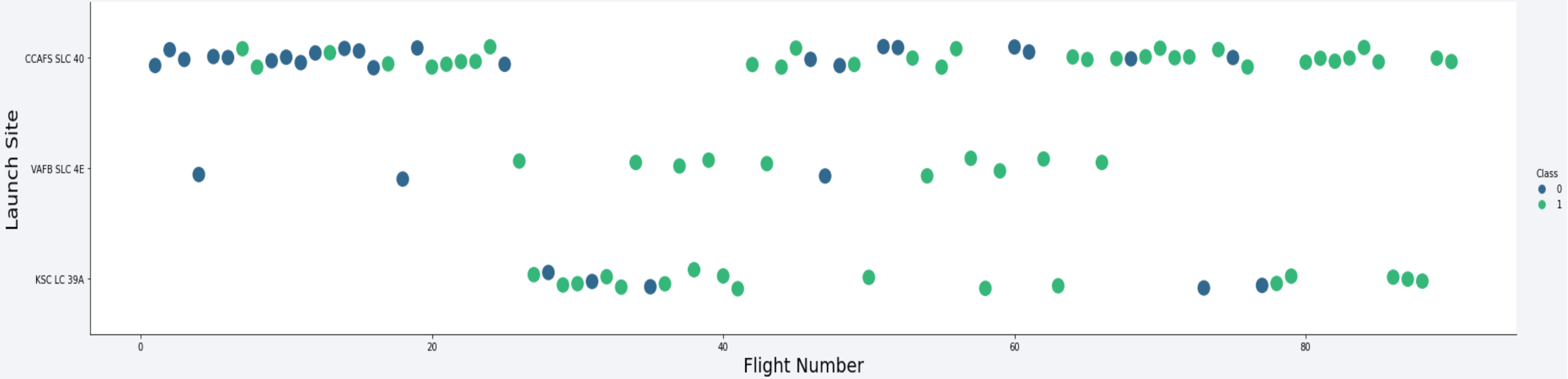

SpaceX Launch Records Dashboard
All Sites
Total Success Launches by Site
Payload range (Kg)

Section 2

Insights drawn from EDA



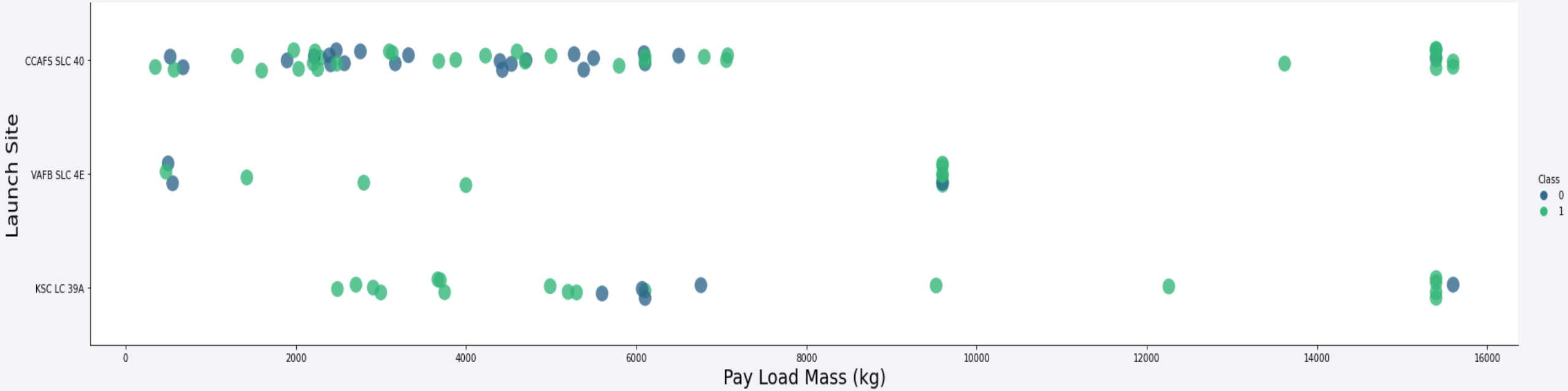
Flight Number vs. Launch Site



Green indicates successful launch; Purple indicates unsuccessful launch.

Graphic suggests an increase in success rate over time (indicated in Flight Number). Likely a big breakthrough around flight 20 which significantly increased success rate. CCAFS appears to be the main launch site as it has the most volume.

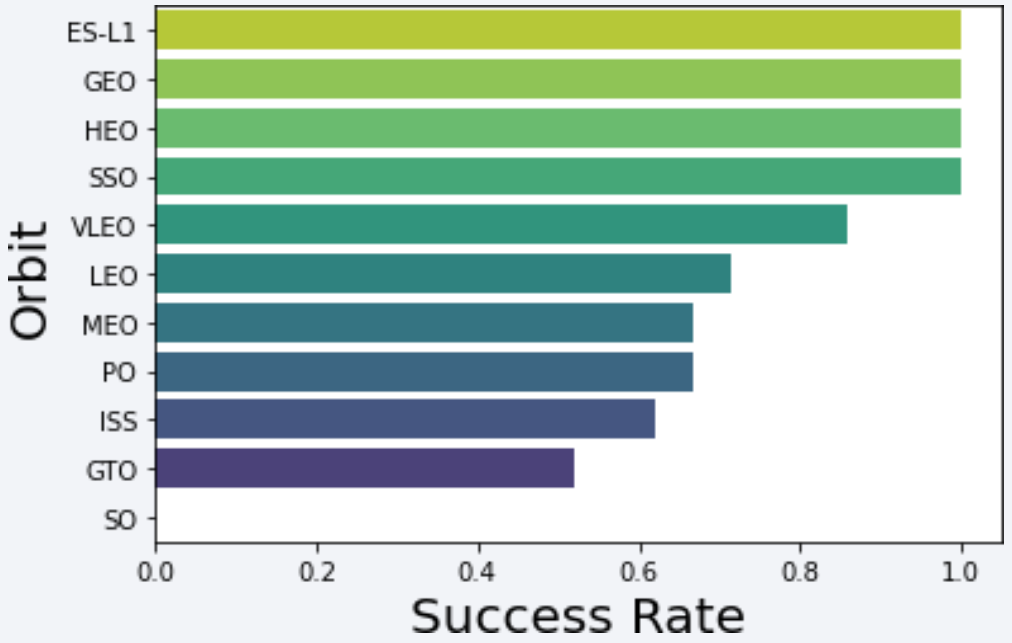
Payload vs. Launch Site



Green indicates successful launch; Purple 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



Success Rate Scale with 0 as 0% 0.6 as 60% 1 as 100%.

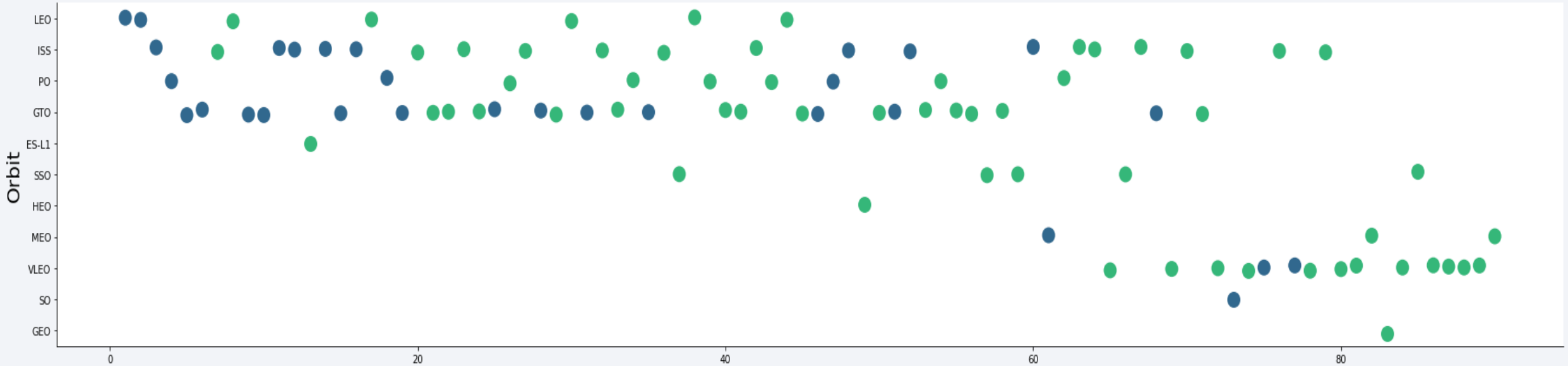
ES-L1 (1), GEO (1), HEO (1) have 100% success rate (sample sizes in parenthesis) SSO (5) has 100% success rate

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

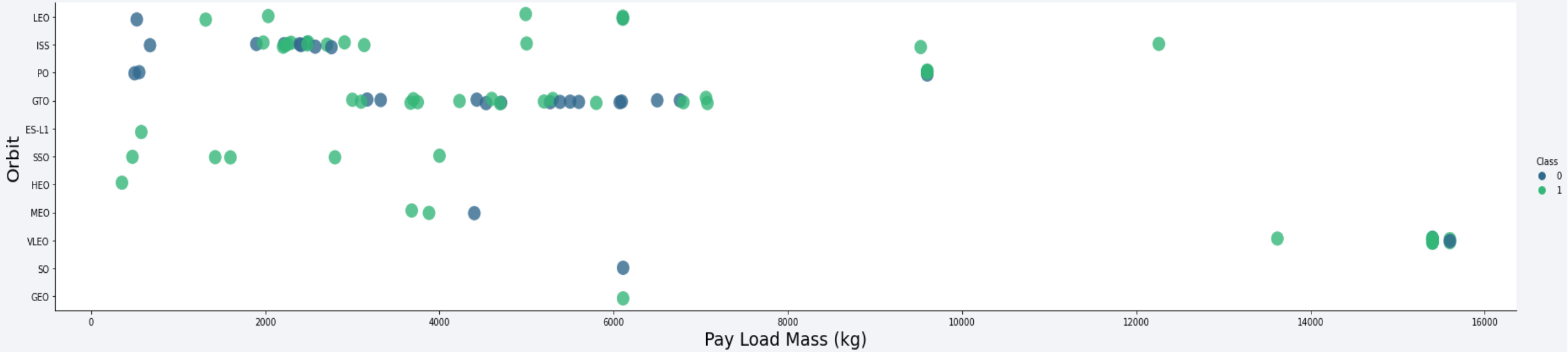
Flight Number vs. Orbit Type



- Green indicates successful launch; Purple 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

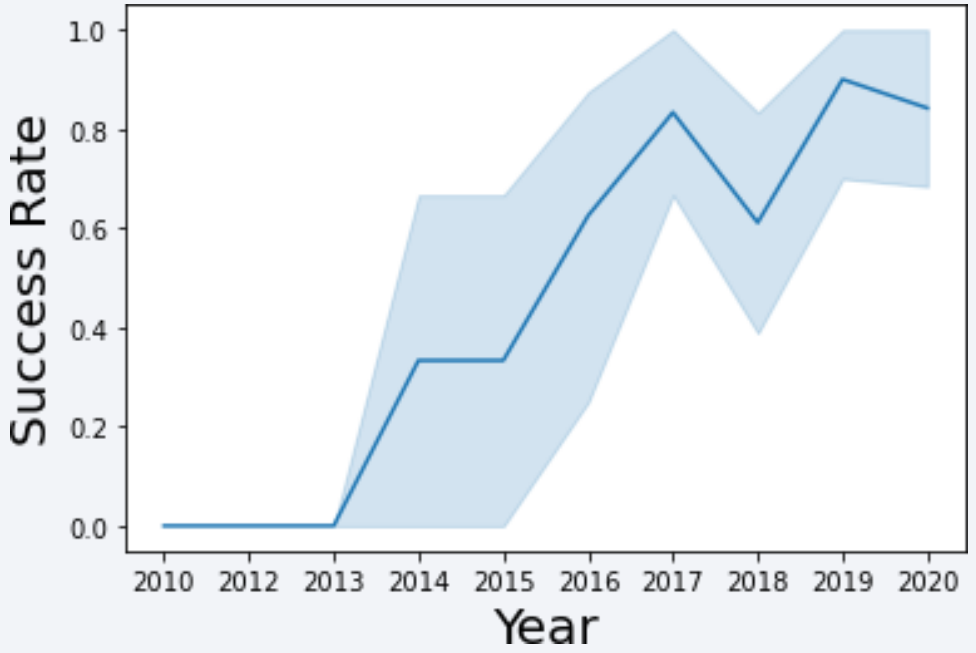
Payload vs. Orbit Type



Green indicates successful launch; Purple 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



95% confidence interval (light blue shading)

Success generally increases over time since 2013 with a slight dip in 2018

Success in recent years at around 80%

All Launch Site Names

- Query unique launch site names from database.
- CCAFS SLC-40 and CCAFSSLC-40 likely all represent the same
- launch site with data entry errors.
- CCAFS LC-40 was the previous name. Likely only 3 unique launch_site values: CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E

```
In [4]: %%sql
        SELECT UNIQUE LAUNCH_SITE
        FROM SPACEXDATASET;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2
Done.
```

Out[4]:

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

Launch Site Names Begin with 'CCA'

First five entries in database with Launch Site name beginning with CCA.

In [5]:
```
%%sql
SELECT *
FROM SPACEXDATASET
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[5]:

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

- This query sums the total payload mass in kg where NASA was the customer.
- CRS stands for Commercial Resupply Services which indicates that these payloads were sent to the International Space Station (ISS).

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS SUM_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE CUSTOMER = 'NASA (CRS)';
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86
Done.
```

sum_payload_mass_kg
45596

Average Payload Mass by F9 v1.1

- This query calculates the average payload mass or launches which used booster version F9 v1.1
- Average payload mass of F9 1.1 is on the low end of our payload mass range

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE booster_version = 'F9 v1.1'
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-8(
Done.
```

avg_payload_mass_kg
2928

First Successful Ground Landing Date

This query returns the first successful ground pad landing date.

First ground pad landing wasn't until the end of 2015.

Successful landings in general appear starting 2014.

```
%%sql
SELECT MIN(DATE) AS FIRST_SUCCESS
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (ground pad)';
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81
Done.
```

first_success
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- This query returns the four booster versions that had successful drone ship landings and a payload mass between 4000 and 6000 noninclusively.

```
%%sql
SELECT booster_version
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4001 AND 5999;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.database
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- This query returns a count of each mission outcome.
- SpaceX appears to achieve its mission outcome nearly 99% of the time.
- This means that most of the landing failures are intended.
- Interestingly, one launch has an unclear payload status and unfortunately one f ailed in flight

```
%%sql
SELECT mission_outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
GROUP BY mission_outcome;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-
Done.
```

mission_outcome	no_outcome
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- This query returns the booster versions that carried the hig hest payload mass of 15600 kg.
- These booster versions are ve ry similar and all are of the F9 B5 B10xx.x variety.
- This likely indicates payload mass correlates with the boo ster version that is used.

```
%%sql
SELECT booster_version, PAYLOAD_MASS__KG_
FROM SPACEXDATASET
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXDATASET);
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1
Done.
```

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 query returns the Month, Landing Outcome, Booster Version, Payload Mass (kg), and Launch site of 2015 launches where stage 1 fail ed to land on a drone ship.
- There were two such occurrences.

```
%%sql
SELECT MONTHNAME(DATE) AS MONTH, landing__outcome, booster_version, PAYLOAD_MASS__KG_, launch_site
FROM SPACEXDATASET
WHERE landing__outcome = 'Failure (drone ship)' AND YEAR(DATE) = 2015;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.app
Done.
```

MONTH	landing__outcome	booster_version	payload_mass__kg_	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	2395	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	1898	CCAFS LC-40

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

- This query returns a list of successful landings and between 2010-06-04 and 2017-03-20 inclusively.
- There are two types of successful landing outcomes: drone ship and ground pad landings.
- There were 8 successful landings in total during this time period

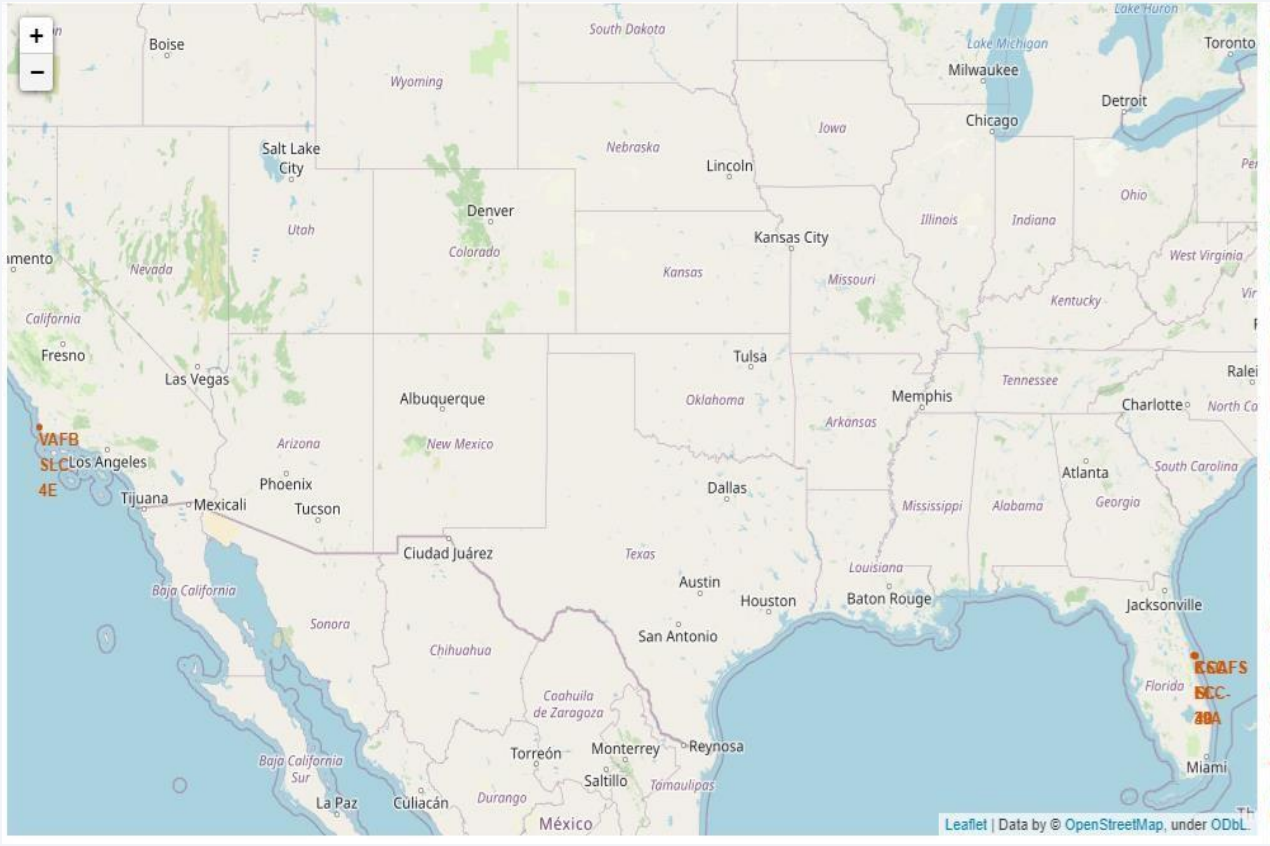
```
%%sql
SELECT landing__outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
WHERE landing__outcome LIKE 'Succes%' AND DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY landing__outcome
ORDER BY no_outcome DESC;
```

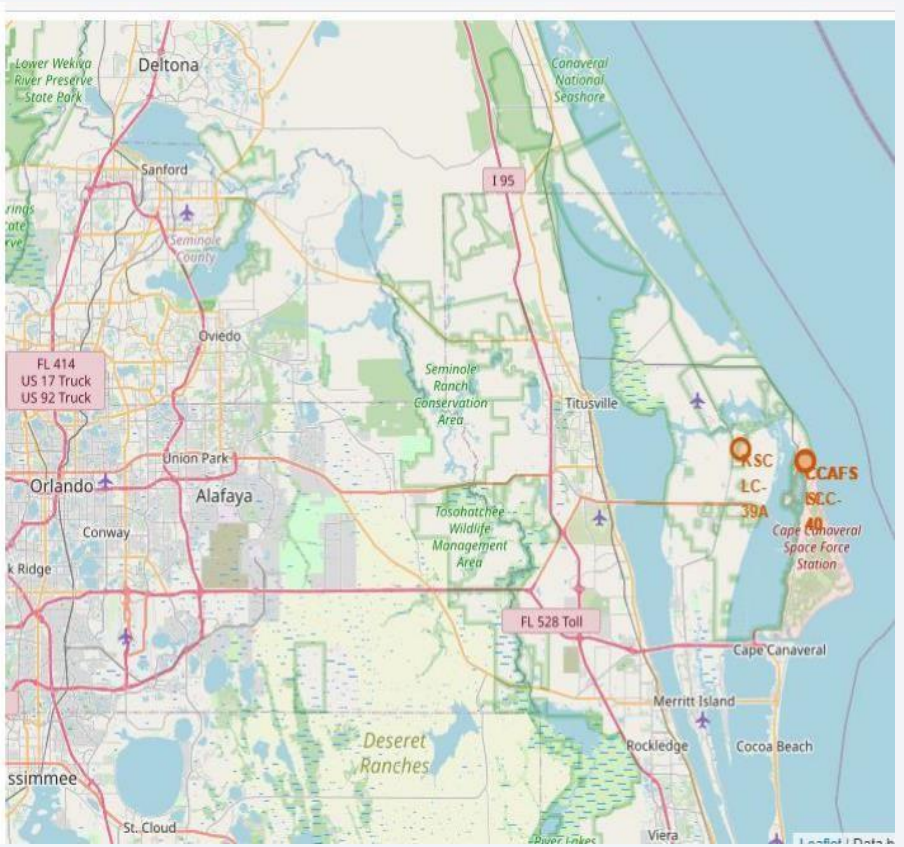
```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg
Done.
```

landing__outcome	no_outcome
Success (drone ship)	5
Success (ground pad)	3


Section 3
Launch Sites
Proximities Analysis

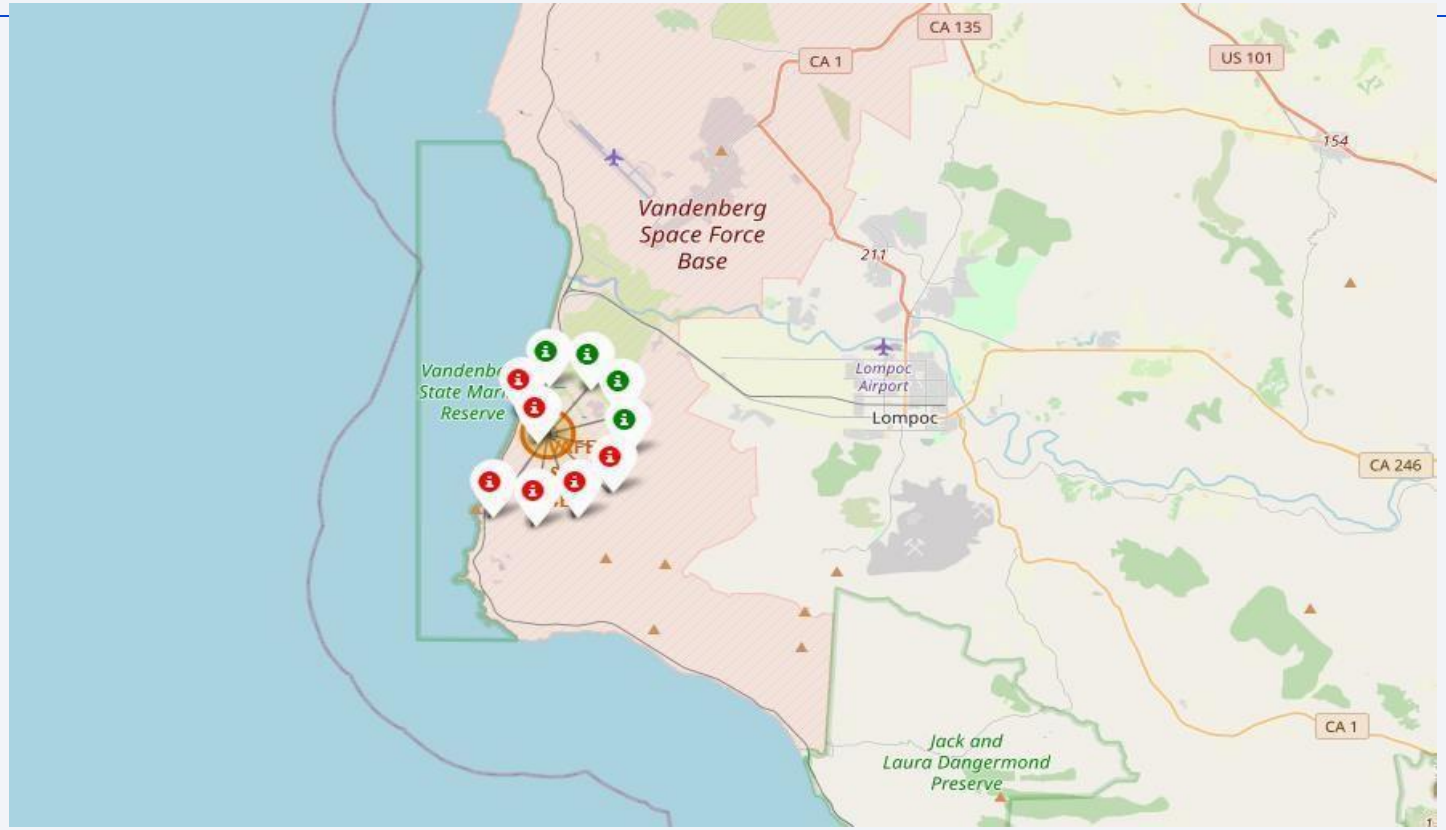
Launch Site Locations





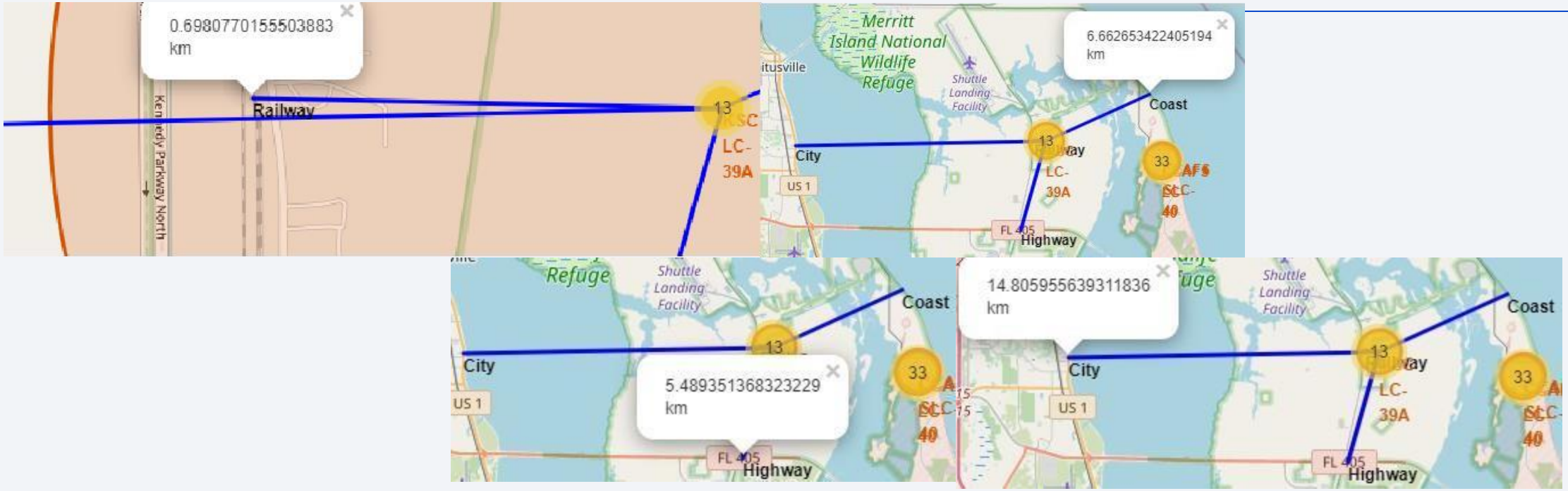
The left map shows all launch sites relative US map. The right map shows the two Florida launch sites since they are very close to each other. All launch sites are near the ocean.

Color-Coded Launch Markers

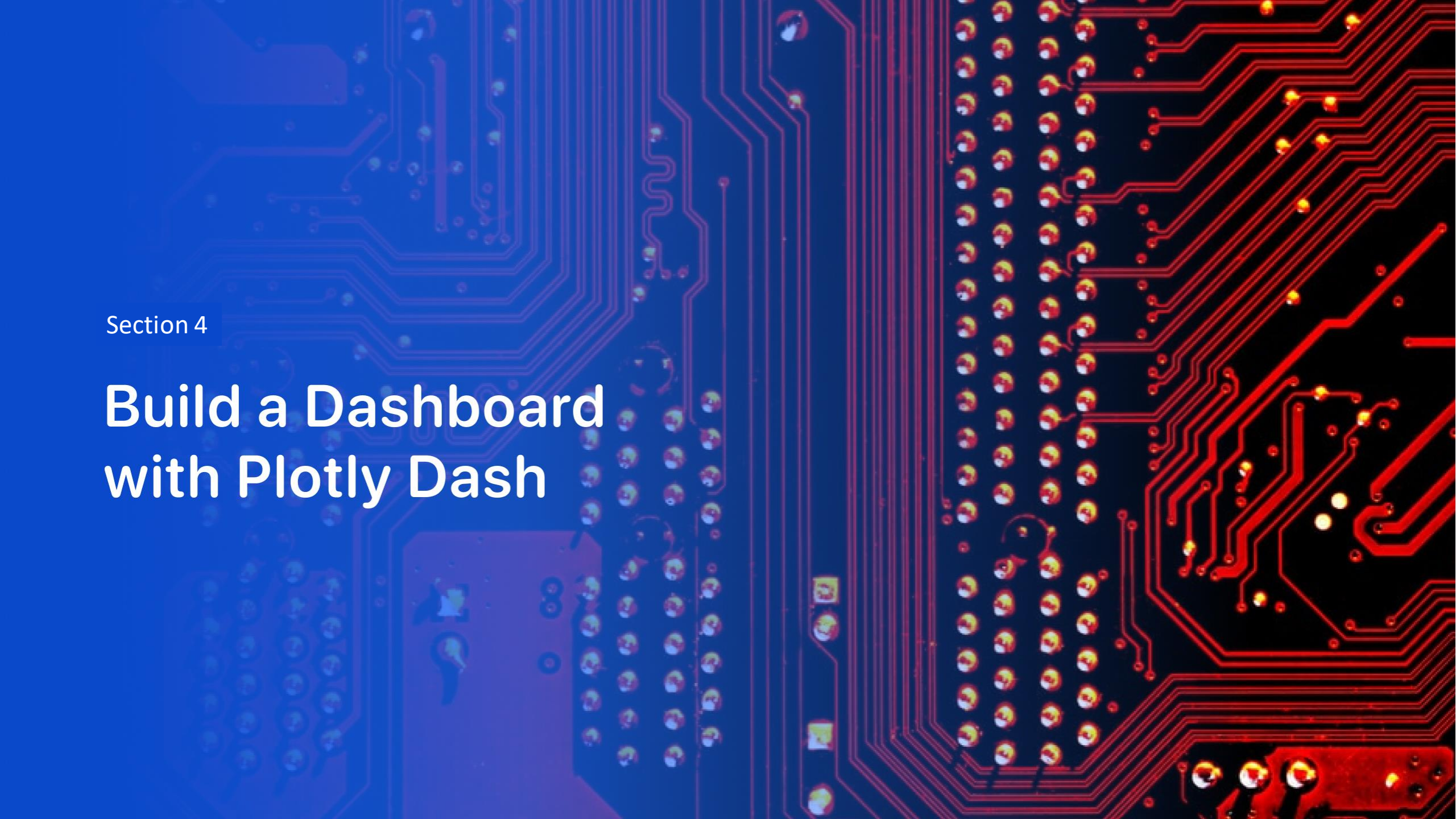


Clusters on Folium map can be clicked on to display each successful landing (green icon) and failed
landing (red icon). In this example VAFB SLC-4E shows 4 successful landings and 6 failed landings.

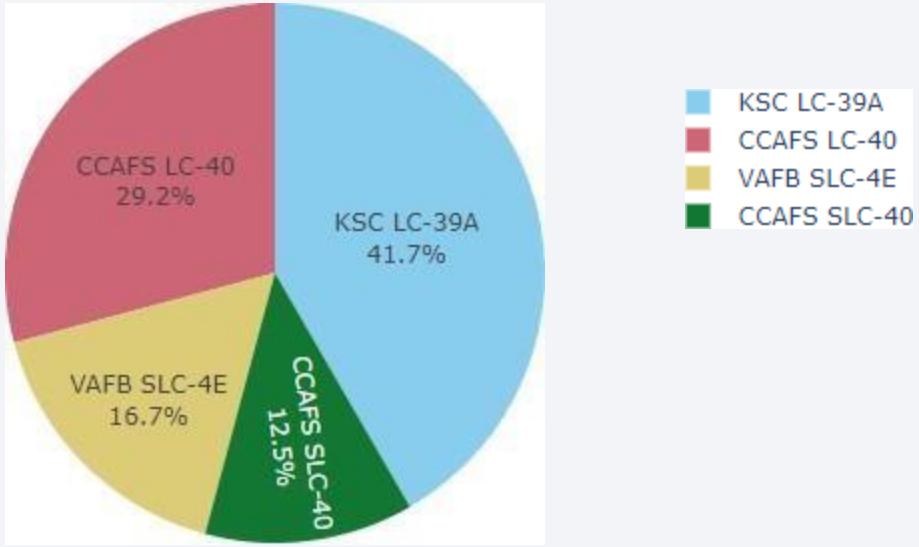
Key Location Proximities



Using KSC LC-39A as an example, launch sites are very close to railways for large part and supply transportation. Launch sites are close to highways for human and supply transport. Launch sites are also close to coasts and relatively far from cities so that launch failures can land in the sea to avoid rockets falling on densely populated areas.


Section 4
Build a Dashboard with Plotly Dash

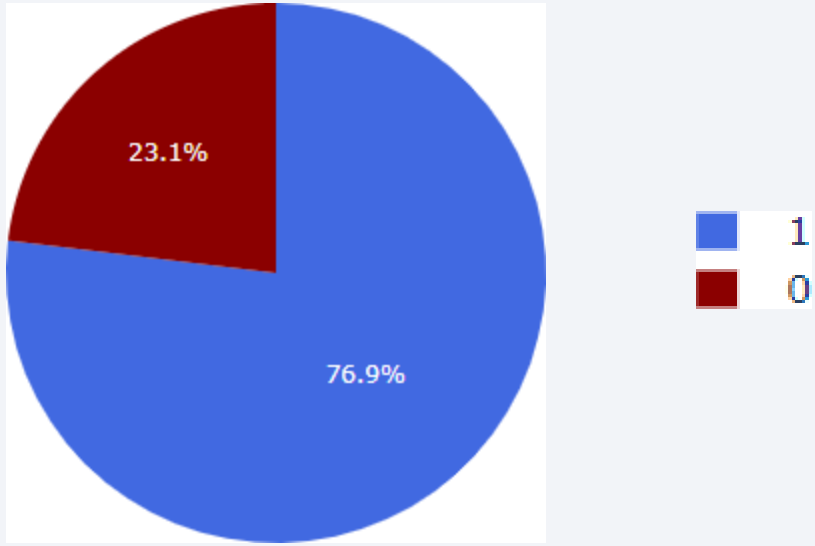
Successful Launches Across Launch Sites



This is the distribution of successful landings across all launch sites. CCAFS LC-40 is the old name of CCAFS SLC-40 so CCAFS and KSC have the same amount of successful landings, but a majority of the successful landings where performed before the name change. VAFB has the smallest share of successful landings. This may be due to smaller sample and increase in difficulty of launching in the west coast.

Highest Success Rate Launch Site

KSC LC-39A Success Rate (blue=success)



KSC LC-39A has the highest success rate with 10 successful landings and 3 failed landings.

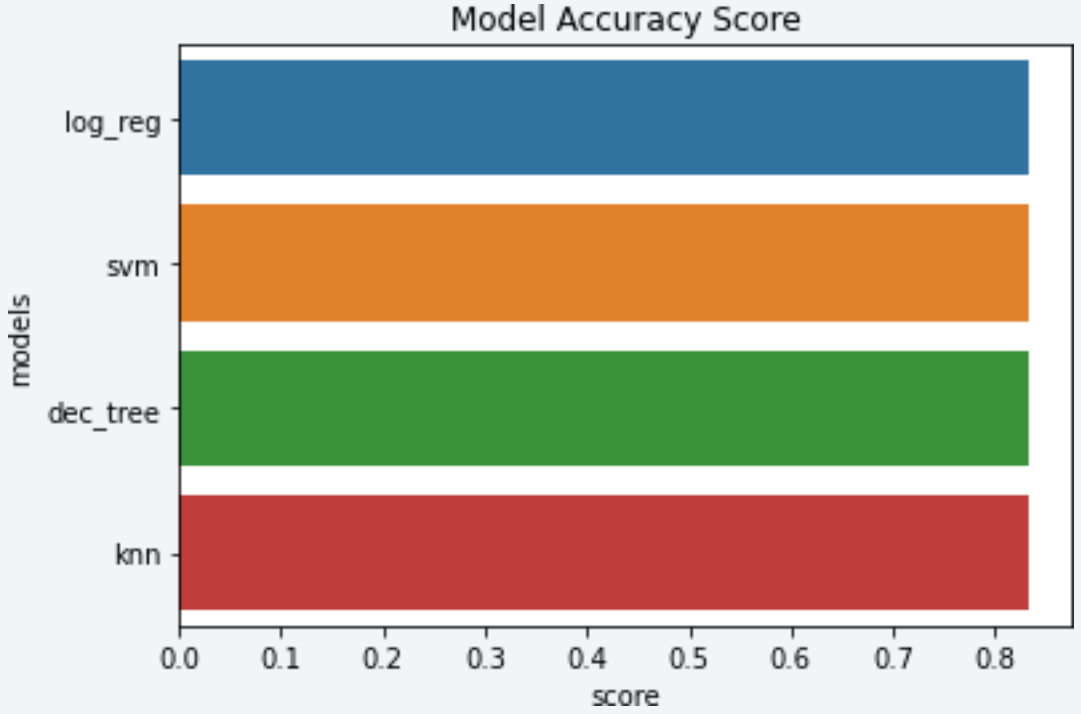
Payload Mass vs. Success vs. Booster Version Category

Plotly dashboard has a Payload range selector. However, this is set from 0-10000 instead of the max Payload of 15600. Class indicates 1 for successful landing and 0 for failure. Scatter plot also accounts for booster version category in color and number of launches in point size. In this particular range of 0-6000, interestingly there are two failed landings with payloads of zero kg.

Section 5

Predictive Analysis (Classification)

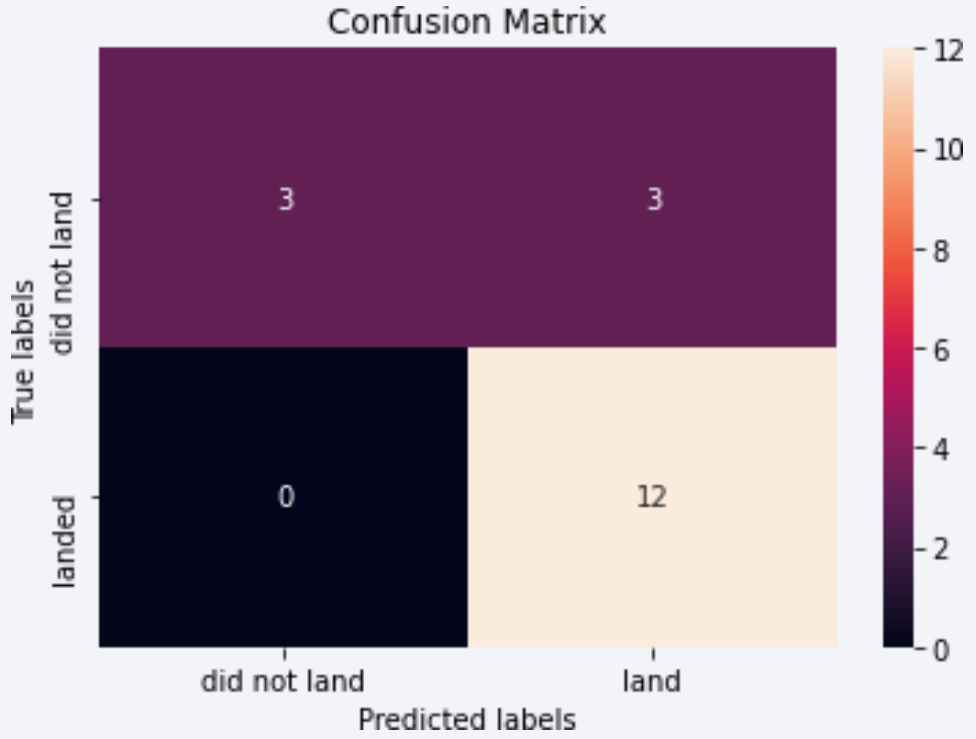
Classification Accuracy



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



Correct predictions are on a diagonal from top left to bottom right.

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

- The task at hand was to develop a machine learning model for Space Y, a company bidding against SpaceX in the commercial space industry. The objective of the model was to predict the successful landing of Stage 1 rockets, which could result in significant cost savings of approximately $100 million USD.
- To accomplish this task, data was collected from both the public SpaceX API and web scraping the SpaceX Wikipedia page. The collected data was labeled and stored in an IBM DB2 SQL database for further analysis and modeling.
- A dashboard was created to provide visualizations of the data, enabling better understanding and exploration of the dataset.
- A machine learning model was developed, achieving an accuracy rate of 83%. This model can be utilized by Space Y, specifically Allon Mask, to predict the success of Stage 1 landings before the launch. This prediction can assist in decision-making regarding whether to proceed with a launch or not, based on the likelihood of a successful landing.

Appendix

GitHub repository url:

Instructors:
Instructors: Rav Ahuja, Alex Aklson, Aije Egwaikhide, Svetlana Levitan, Romeo Kienzler, Polong Lin, Joseph Santarcangelo, Azim Hirjani, Hima Vasudevan, Saishruthi Swaminathan, Saeed Aghabozorgi, Yan Luo

Special Thanks to All Instructors:

https://www.coursera.org/professional-certificates/ibm-data-science?#instructors

Thank you!