IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Johannes Keck
February 2024

Outline

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

Executive Summary

- Summary of methodologies
  - Collected data from public SpaceX API and SpaceX Wikipedia page.
  - Created labels column ‘class’ which classifies successful landings.
  - Explored data using SQL, visualization, folium maps, and dashboards. Gathered relevant columns to be used as features.
  - Changed all categorical variables to binary using one hot encoding.
  - Standardized data and used GridSearchCV to find best parameters for machine learning models.
  - Visualize accuracy score of all models.
- Summary of all results
  - Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors.
  - All produced similar results with accuracy rate of about 83.33%.
  - All models over predicted successful landings. More data is needed for better model determination and accuracy.

Introduction

- Project background and context
  - Space X has best pricing ($62 million vs. $165 million USD)
  - Largely due to ability to recover part of rocket (Stage 1)
  - Space Y wants to compete with Space X

- Problems you want to find answers
  - Space Y tasks us to train a machine learning model to predict successful Stage 1 recovery

Section 1

Methodology

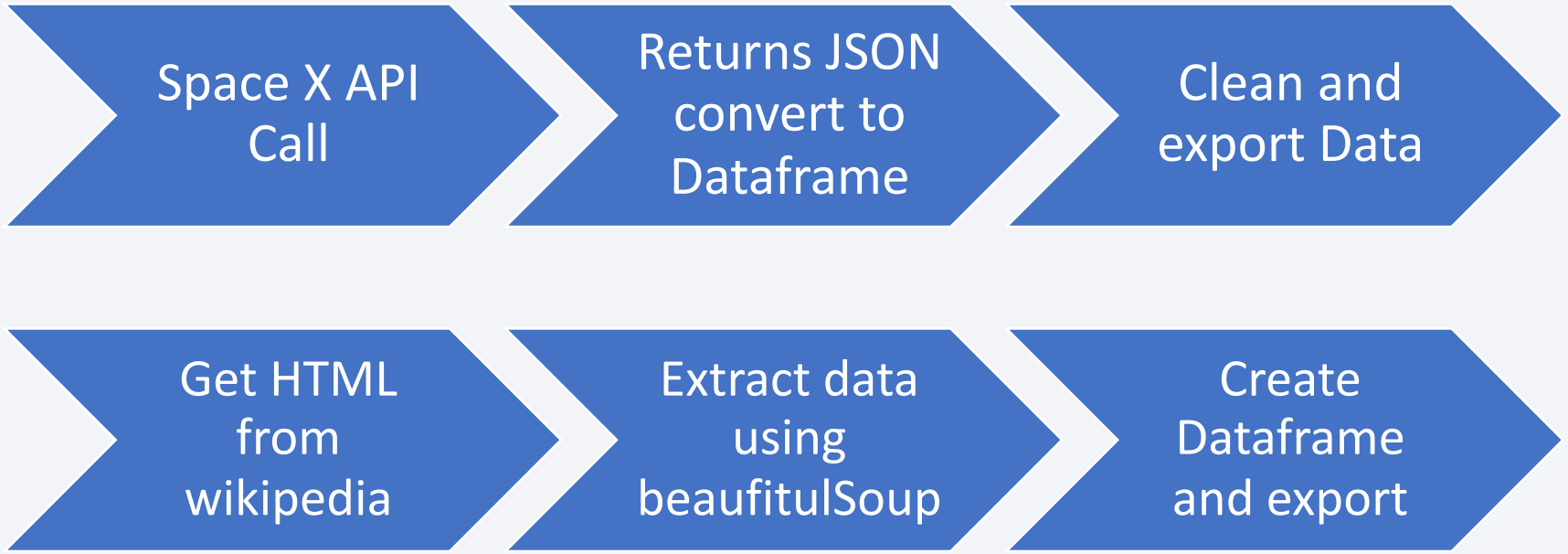
Methodology

Executive Summary

- Data collection methodology:
  - Combined data from SpaceX public API and SpaceX Wikipedia page
- Perform data wrangling
  - Classifying true landings as successful and unsuccessful otherwise
  - One encoding for classification models
- 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
  - Build, tune, evaluate

Data Collection

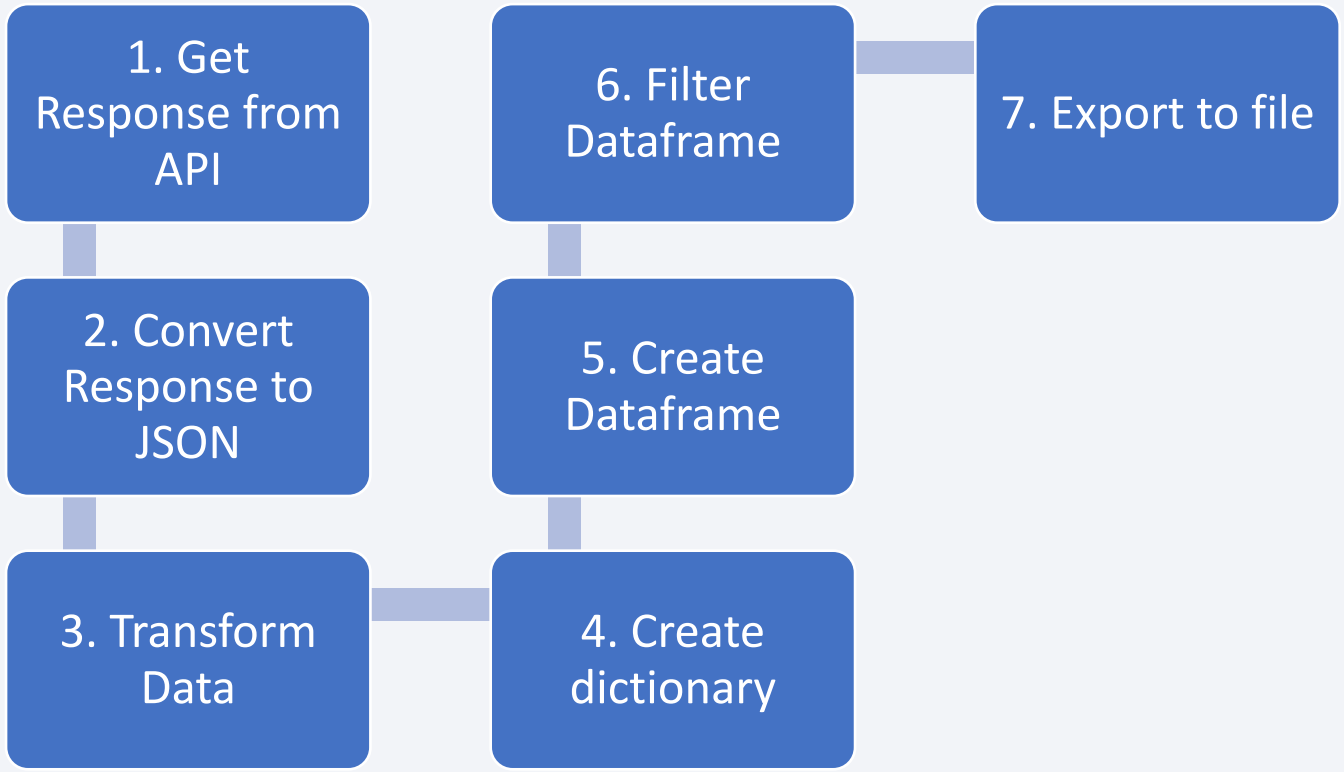
- Describe how data sets were collected.
  - Datasets are collected from Rest SpaceX API and webscrapping Wikipedia
    - API = rocket, launches, payload information
    - Wikipedia = launches, landing, payload information



Data Collection – SpaceX API

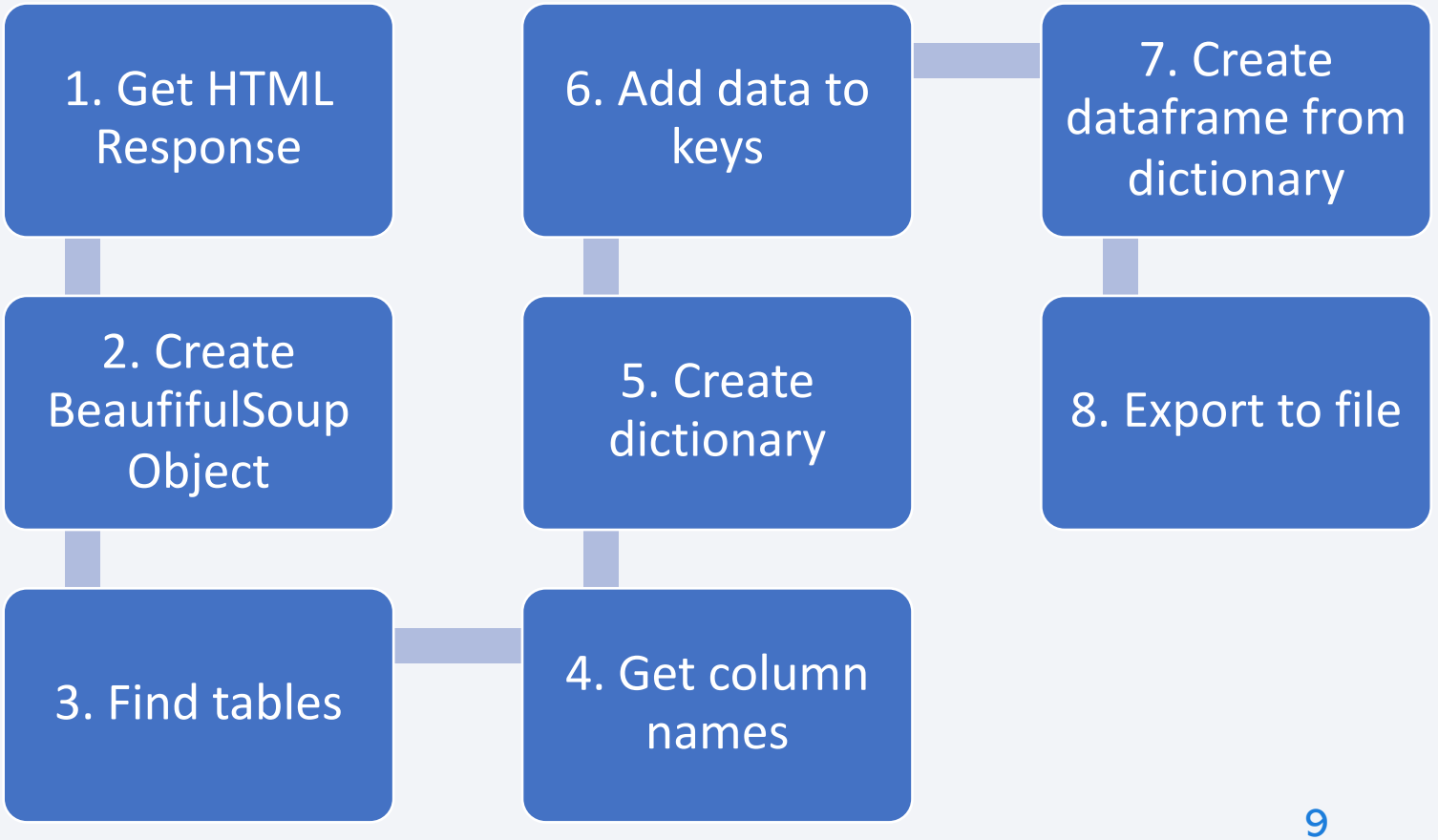
- Present your data collection with SpaceX REST calls using key phrases and flowcharts

- API Calls



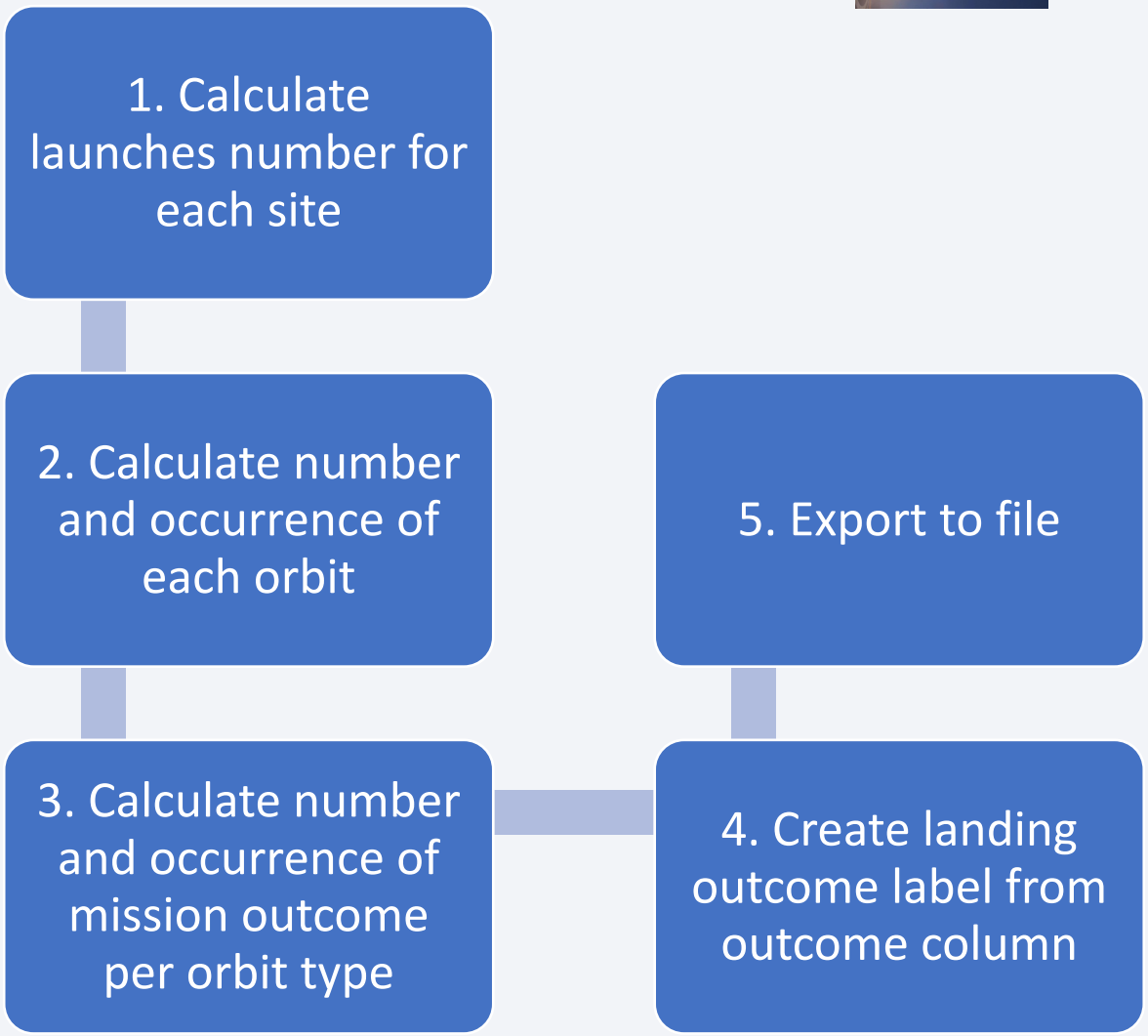
Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts
- Webscrapping



Data Wrangling

- Describe how data were processed
- You need to present your data wrangling process using key phrases and flowcharts
- Data Wrangling



EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
  - Scatter Plots, Bar Charts, Line Graphs
  - Investigate and deepen the understanding of the relationships between variables in order to make informed decisions about data processing and machine learning models
- Add the GitHub URL of your completed EDA with data visualization notebook, as an external reference and peer-review purpose
  - Data Visualization

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
  - Loaded data set into IBM DB2 Database.
  - Queried using SQL Python integration.
  - Queries were made to get a better understanding of the dataset.
  - Queried information about launch site names, mission outcomes, various pay load sizes of customers and booster versions, and landing outcomes
- Add the GitHub URL of your completed EDA with SQL notebook, as an external reference and peer-review purpose
  - EDA with SQL

Build an Interactive Map with Folium

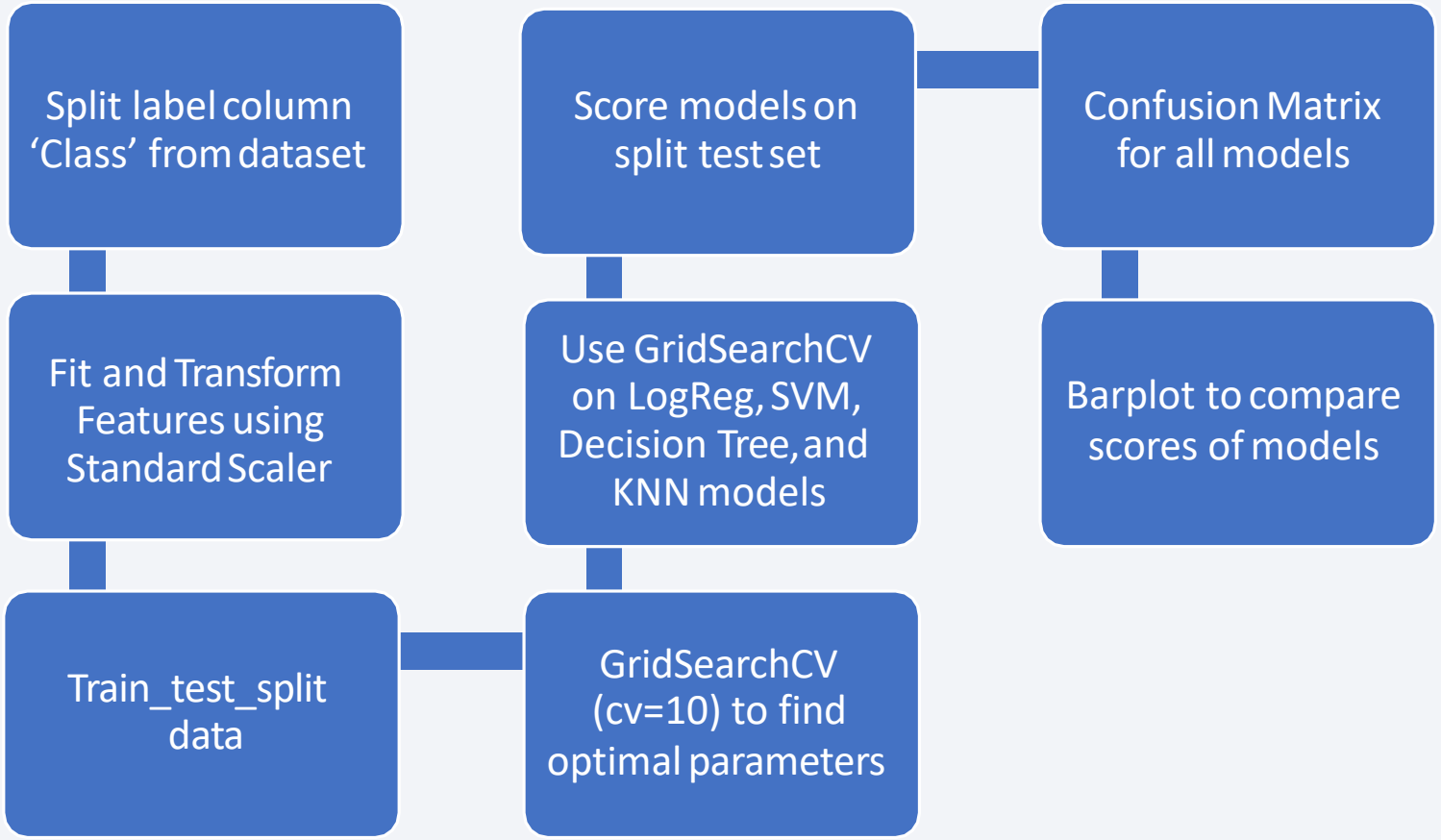
- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
  - Folium maps mark
    - Launch Sites,
    - successful and unsuccessful landings,
    - Proximity to important locations of interest
- Explain why you added those objects
    - Understand the importance of the launch sites location
    - success vs. non-successful launches and the impact of location
- Add the GitHub URL of your completed interactive map with Folium map, as an external reference and peer-review purpose
    - Launch Sites

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboardDashboard includes a pie chart and a scatter plot. / Why did I add those plots and interactions?
  - 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.
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose
  - Python Dash Lab

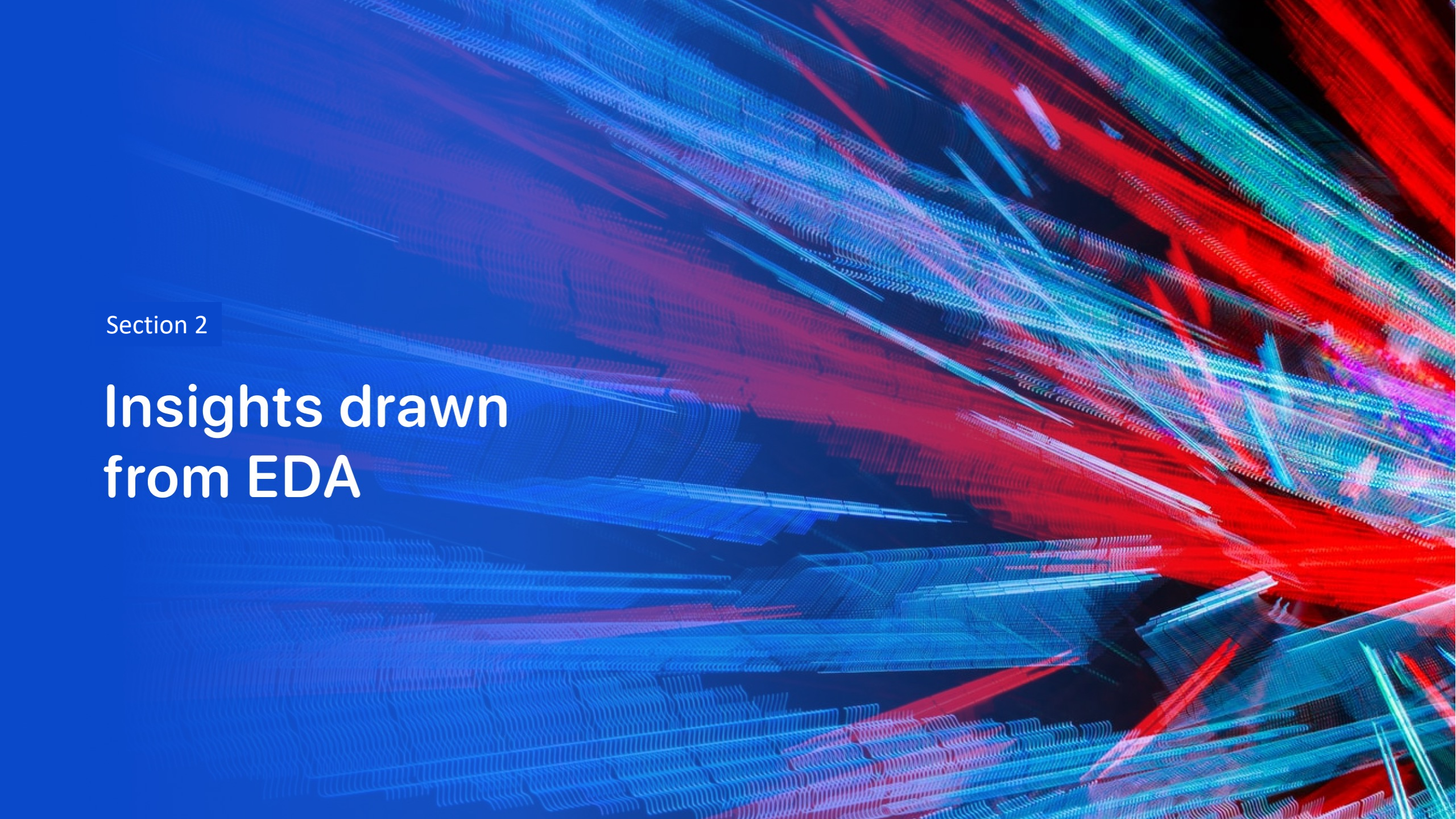
Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
  - Load, normalize, split
  - Selection of ML algorithm, parameters for GridSearchCV, Training
  - HyperparameterTuning, Accuracies, CM
  - Compare Models
- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose
- ML

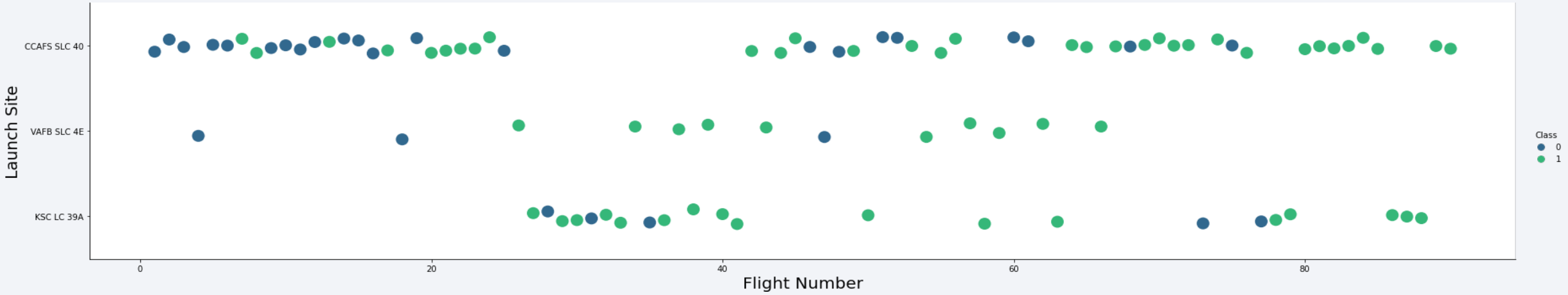


Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

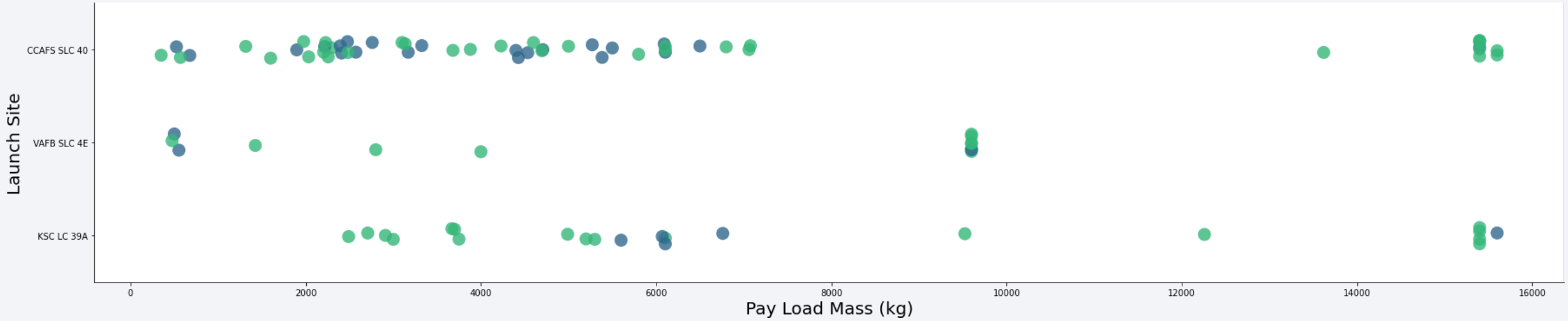

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



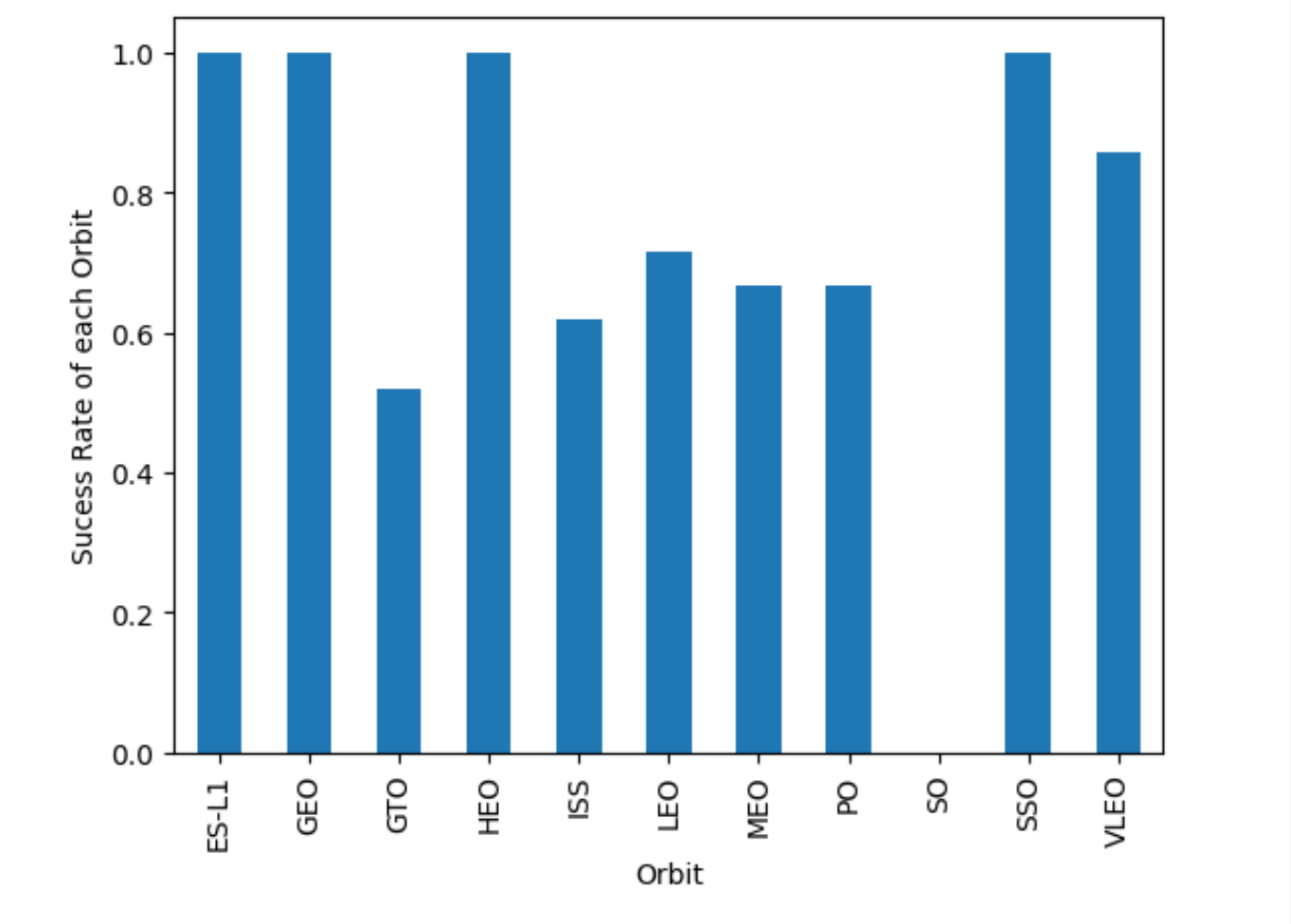
- The graphic illustrates a progressive increase in success rate over time, as denoted by Flight Number.
- A significant breakthrough is apparent around the 20th flight, resulting in a notable surge in success rate.
- CCAFS emerges as the predominant launch site, evident from its higher launch volume.

Payload vs. Launch Site



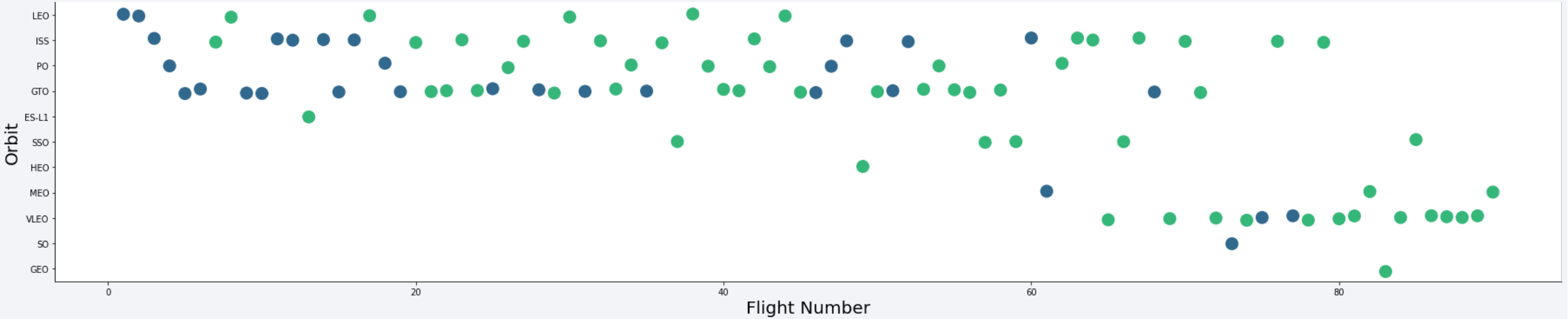
- Payload masses predominantly range from 0 to 6000 kg.
- Various launch sites exhibit differences in payload masses, with distinct ranges observed.

Success Rate vs. Orbit Type



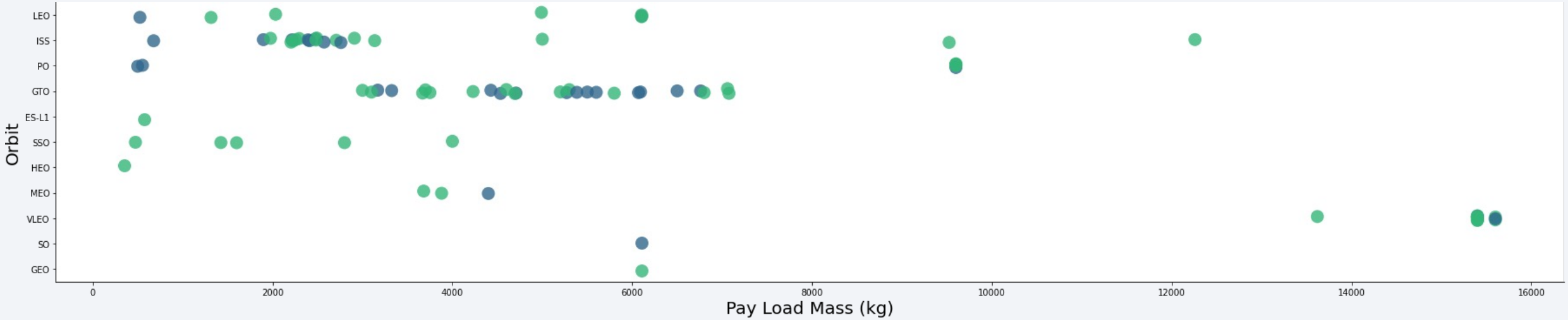
- ES-L1 (1), GEO (1), HEO (1) boast a 100% success rate (sample sizes in parentheses), while SSO (5) also maintains a perfect success rate.
- VLEO (14) demonstrates a commendable success rate with multiple attempts.
- SO (1) records a 0% success rate.
- GTO (27) showcases approximately a 50% success rate, representing the largest sample size.

Flight Number vs. Orbit Type



- There has been a shift in Launch Orbit preferences over Flight Number, with Launch Outcome seemingly correlating with these preferences.
- SpaceX initially favored LEO orbits, which exhibited moderate success, then transitioned back to VLEO in recent launches.
- SpaceX appears to achieve better performance in lower orbits or Sun-synchronous orbits.

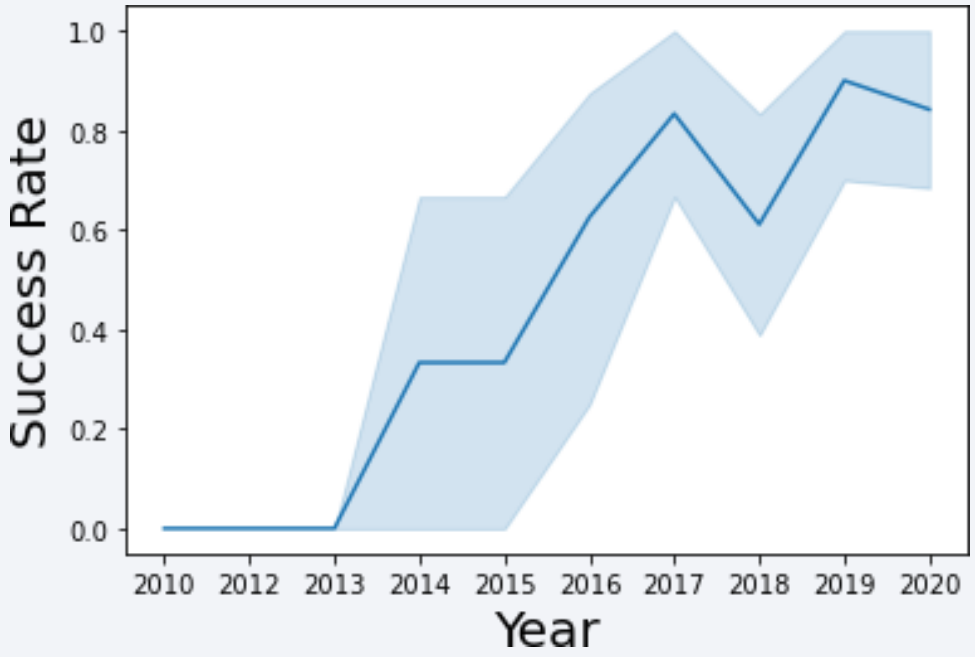
Payload vs. Orbit Type



- Payload mass appears to correlate with orbit type, with LEO and SSO orbits generally associated with lower payload masses.

- VLEO, another successful orbit, typically features payload mass values at the higher end of the range compared to other orbits.

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

- Find the names of the unique launch sites
- Present your query result with a short explanation here

```
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

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Present your query result with a short explanation here

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

- Calculate the total payload carried by boosters from NASA
- Present your query result with a short explanation here

```
%%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

- 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).

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Present your query result with a short explanation here

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

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

avg_payload_mass_kg
2928

- 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

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Present your query result with a short explanation here

```
%%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

- This query returns the first successful ground pad landing date.
- First ground pad landing wasn't until the end of 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

- Present your query result with a short explanation here

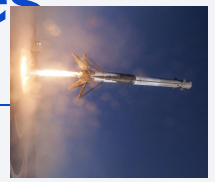
```
%%sql
SELECT booster_version
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4001 AND 5999;
```

```
 * 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

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

Total Number of Successful and Failure Mission Outcomes



- Calculate the total number of successful and failure mission outcomes
- Present your query result with a short explanation here

```
%%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

- 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 failed in flight.

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Present your query result with a short explanation here

```
%%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

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

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

- Present your query result with a short explanation here

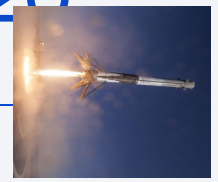
```
%%sql
SELECT MONTHNAME(DATE) AS MONTH, landing__outcome, booster_version, PAYLOAD_MASS__KG_, launch_site
FROM SPACEXDATASET
WHERE landing__outcome = 'Failure (drone ship)' AND YEAR(DATE) = 2015;
```

```
 * 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

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

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



- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order
- Present your query result with a short explanation here

```
%%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;
```

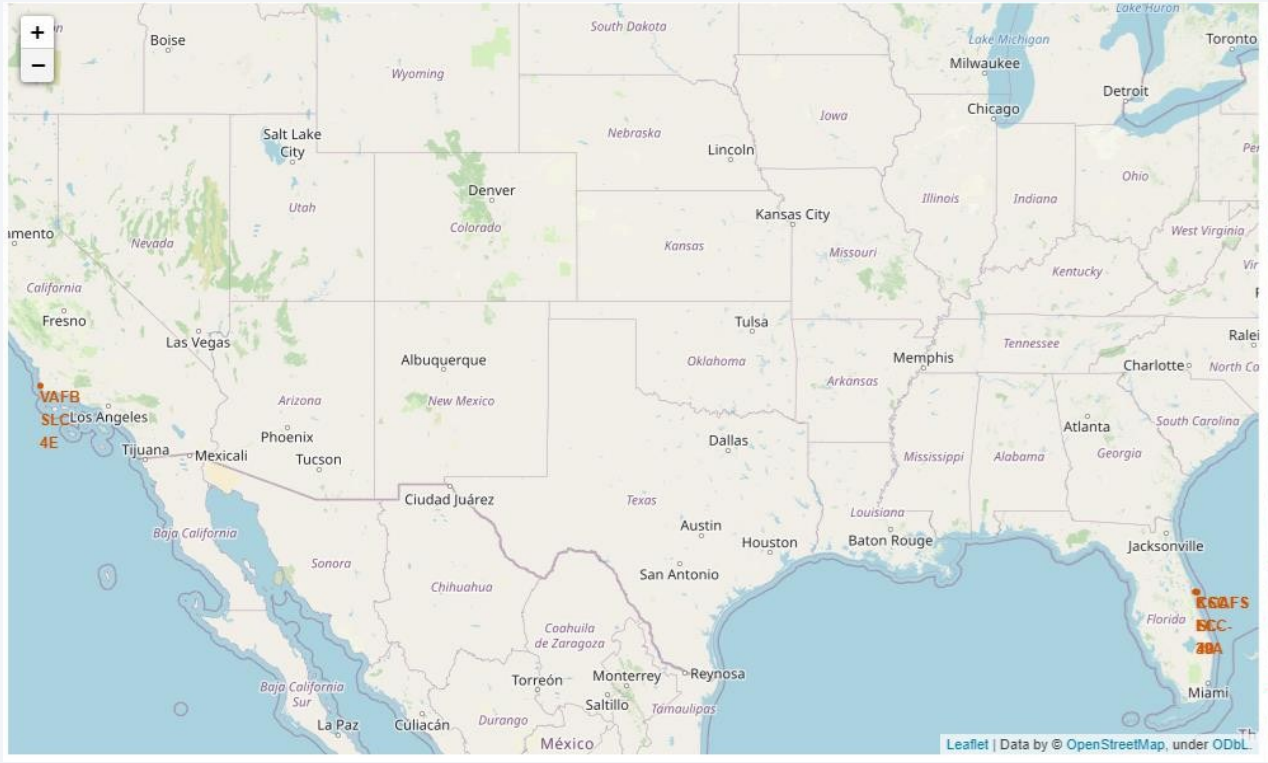
```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg
Done.
```

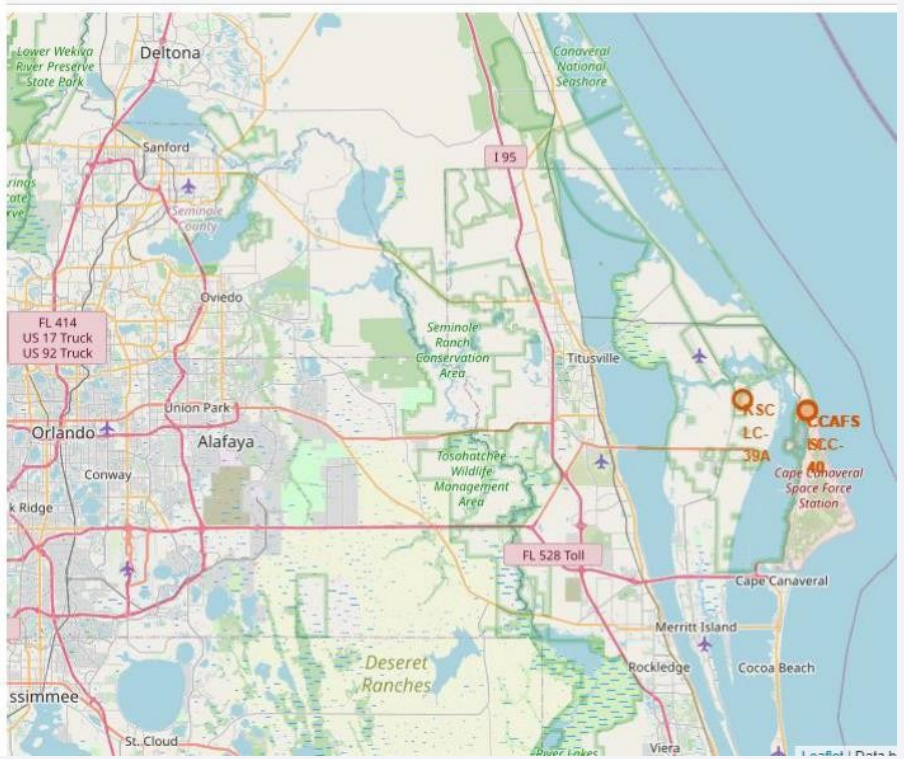
landing__outcome	no_outcome
Success (drone ship)	5
Success (ground pad)	3

- 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


Section 3
Launch Sites
Proximities Analysis

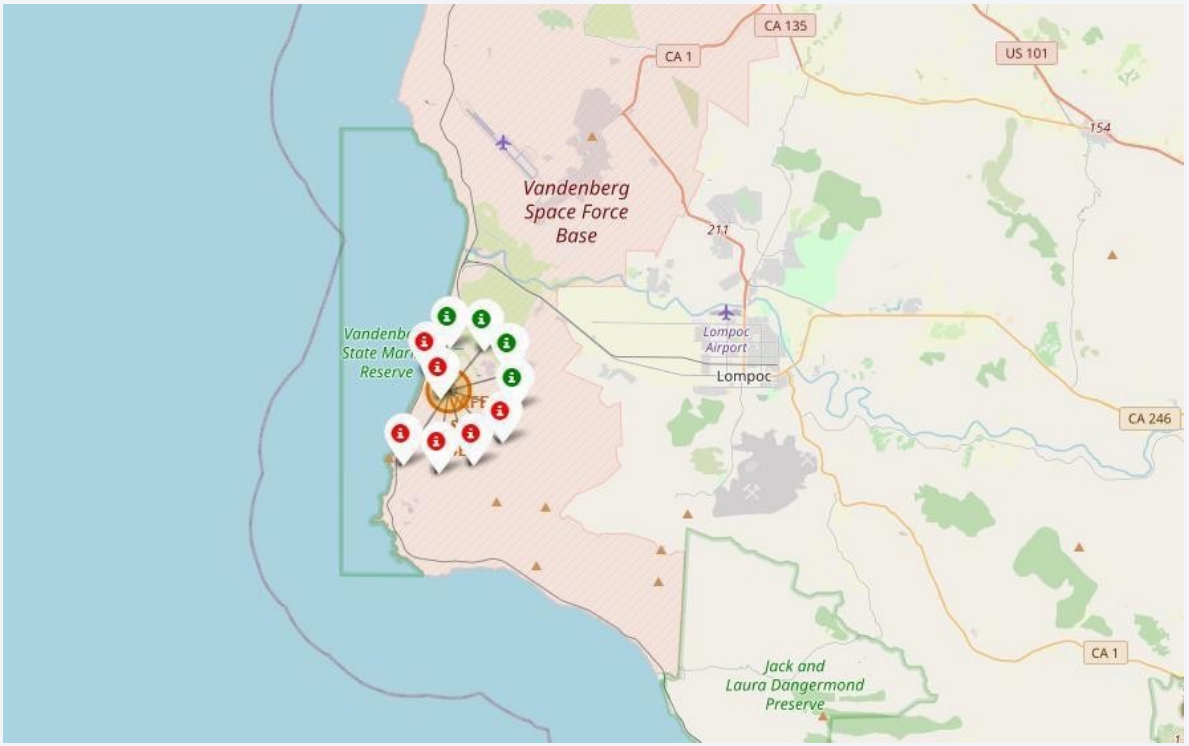
Launch Site Locations





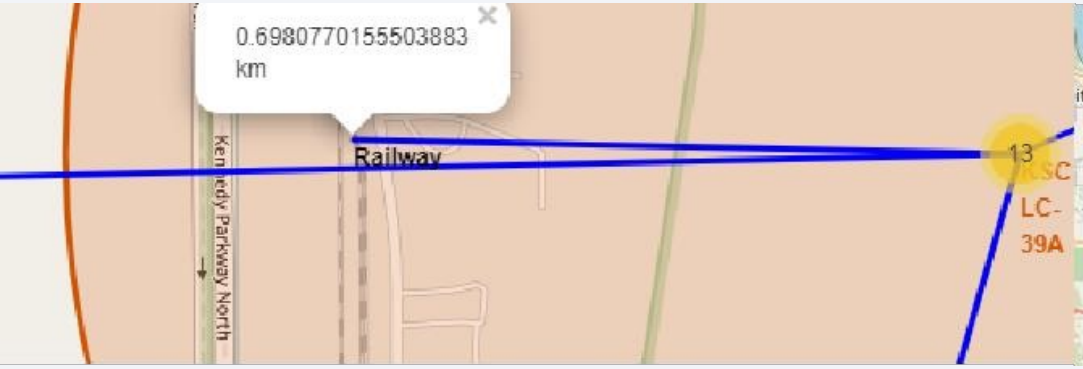
- The left map depicts all launch sites in relation to the US map.
- The right map focuses on the two Florida launch sites, which are in close proximity to each other.
- All launch sites are located near the ocean.

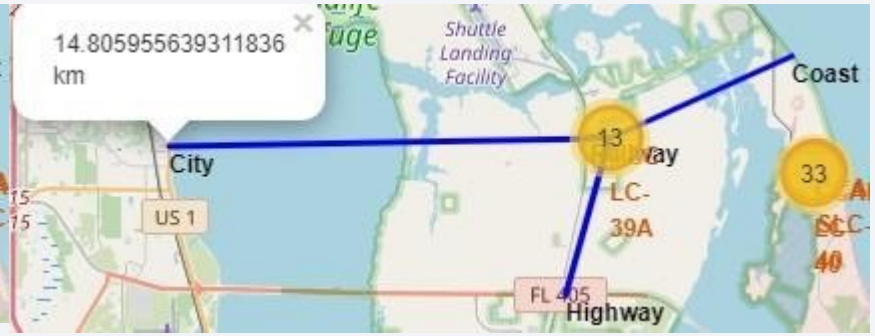
Color Coded Launch Markers



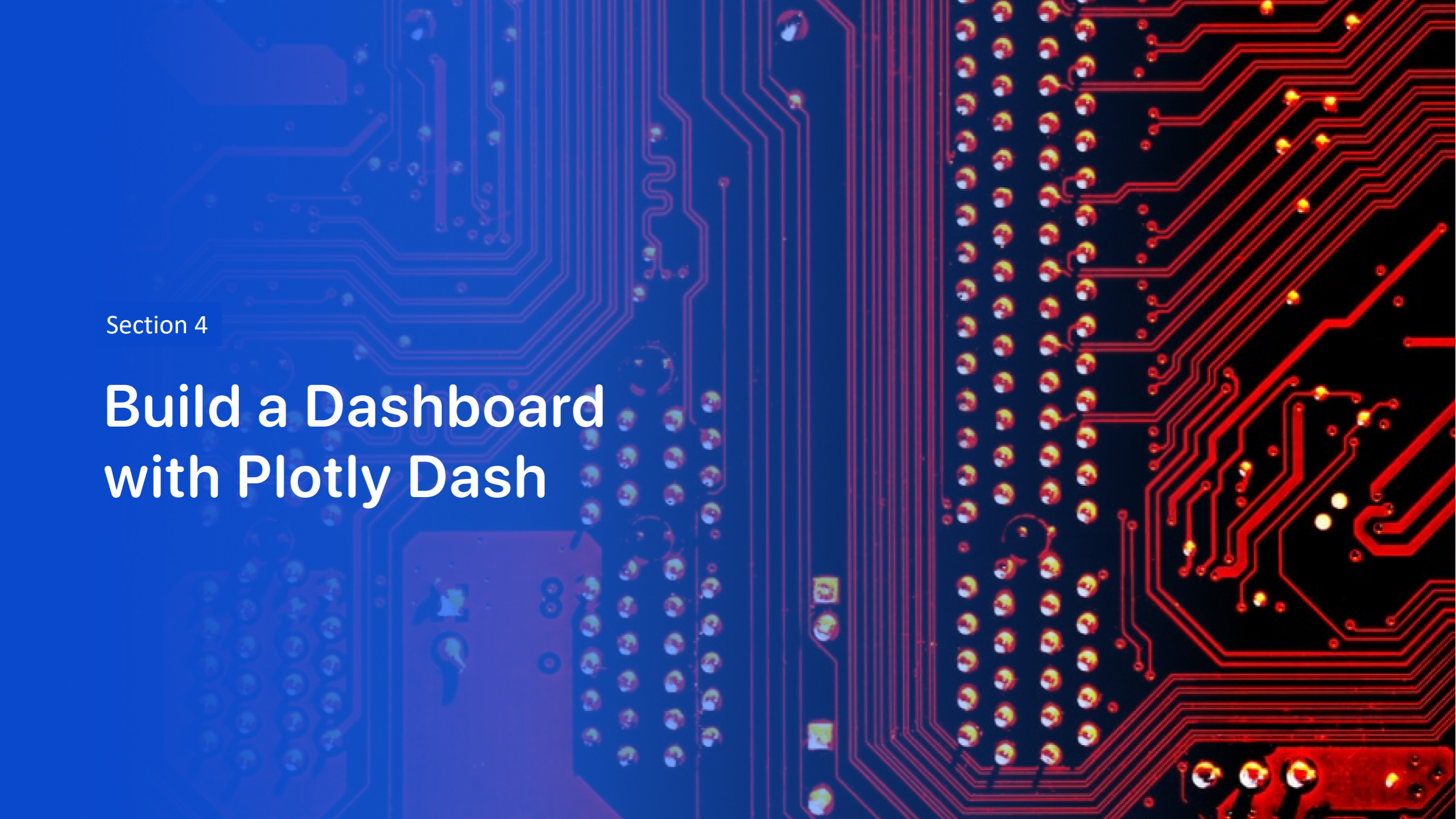
- Clusters on the Folium map are interactive and can be clicked on.
- Each cluster reveals successful landings with green icons and failed landings with red icons upon clicking.
- For instance, at VAFB SLC-4E, there have been 4 successful landings and 6 failed landings according to this example.

Key Location Proximities





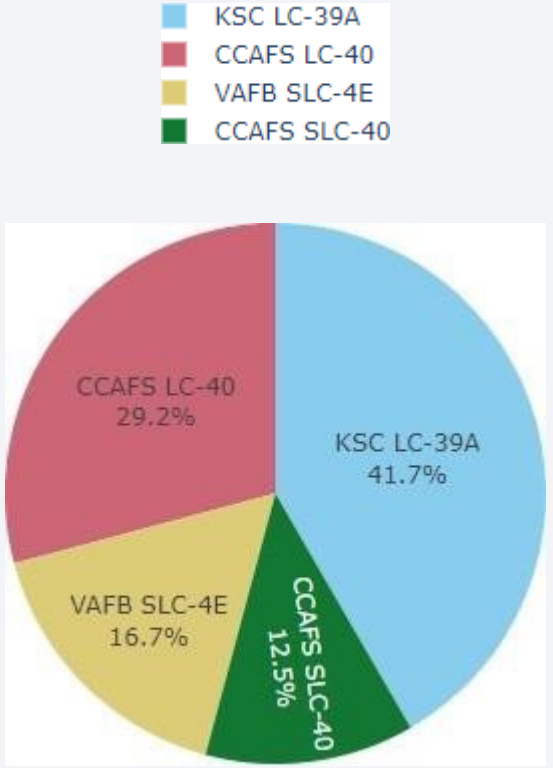
- Using KSC LC-39A as an example, launch sites are strategically positioned near railways, facilitating the transportation of large supplies.
- Additionally, launch sites are situated in close proximity to highways, enabling efficient transportation of personnel and supplies.
- Furthermore, launch sites are strategically located near coasts and are relatively distant from densely populated areas, reducing the risk of launch failures impacting heavily populated regions. This ensures that any potential rocket failures can safely land in the sea, mitigating the risk of rockets falling on densely populated areas.


Section 4
Build a Dashboard with Plotly Dash

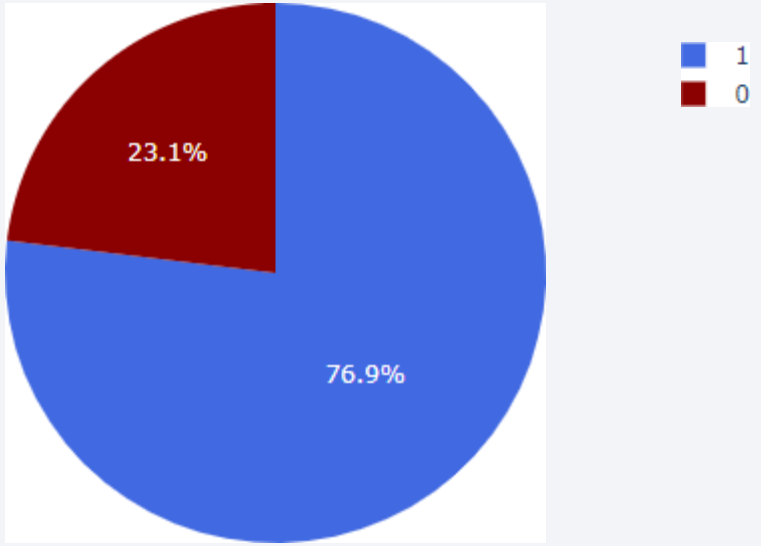
Successful Launches

- The distribution of successful landings across all launch sites indicates that CCAFS LC-40, previously known as CCAFS SLC-40, and KSC have an equal number of successful landings. However, a majority of these successful landings occurred before the name change.

- On the other hand, VAFB has the smallest share of successful landings. This could be attributed to a smaller sample size and potentially increased difficulty in launching from the west coast.

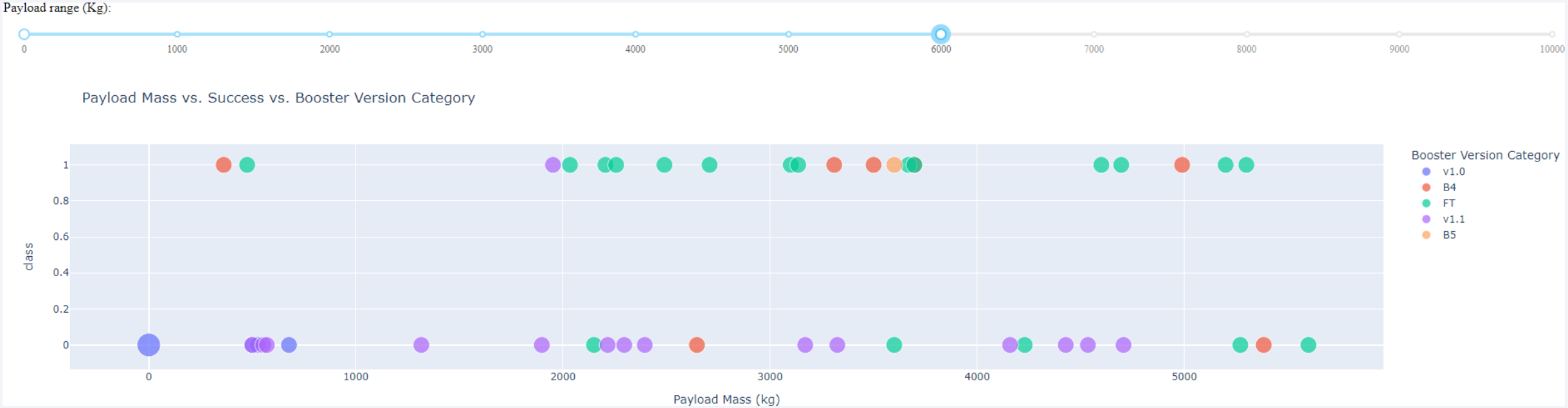


Highest Success Rate



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

Payload Mass vs. Success. Vs. Booster Version Category

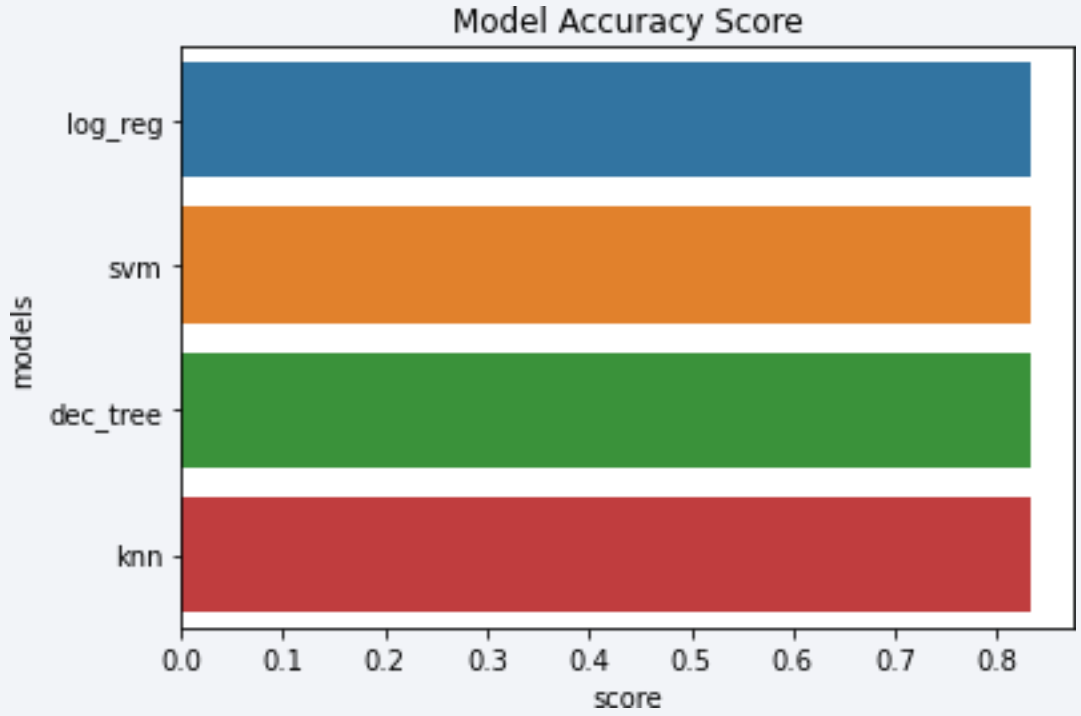


- The Plotly dashboard features a Payload range selector, which is currently set from 0 to 10000 instead of the maximum payload of 15600.
- The "Class" parameter indicates 1 for successful landings and 0 for failures.
- Additionally, the scatter plot incorporates the booster version category in color and the number of launches in point size.
- Interestingly, within the range of 0 to 6000 kg, there are two failed landings with payloads of zero kg.

Section 5

Predictive Analysis (Classification)

Classification Accuracy



	Accuracy Train	Accuracy Test
Logreg	0.846429	0.833333
Svm	0.848214	0.833333
Tree	0.875000	0.777778
Knn	0.848214	0.833333

Confusion Matrix

Since all models performed similarly on the test set, the confusion matrix remains consistent across all models. Specifically:

- The models correctly predicted 12 successful landings when the true label was a successful landing.
- They also correctly predicted 3 unsuccessful landings when the true label was an unsuccessful landing.
- However, the models incorrectly predicted 3 successful landings when the true label was an unsuccessful landing (false positives).

Overall, the models tend to overpredict successful landings.

Conclusions

- - Developed a machine learning model for Space Y to compete with SpaceX.
- - Objective: Predict successful Stage 1 landings to potentially save around $100 million USD.
- - Utilized data from a public SpaceX API and web scraping of SpaceX Wikipedia page.
- - Curated data labels and stored information into a DB2 SQL database.
- - Created a visualization dashboard for data analysis.
- - Achieved a model accuracy of 83%.
- - Allon Mask of SpaceY can use the model to forecast successful Stage 1 landings with relatively high accuracy.
- - Suggested gathering more data to enhance model performance and accuracy.

Thank you!