IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<ABRAHAM MAANE>
<12-06-2024>

Outline

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

Executive Summary

- SpaceX has achieved an impressive overall launch success rate of 84% and continues to improve.
- Launch success has consistently increased year over year, showcasing a commitment to refining processes and technology.
- Strategic launch site selection based on historical data enhances mission success.
- Newer booster versions consistently outperform older models, driving increased reliability.
- SpaceX leverages data analysis to identify factors influencing success and make informed decisions, ensuring continued leadership in the commercial space industry.
- SpaceX demonstrates a high success rate across various launch scenarios.
- The Support Vector Machine (SVM) model outperforms Logistic Regression, Decision Trees, and K-Nearest Neighbors in predicting launch success.
- SpaceX can leverage machine learning to continuously refine its launch predictions and decision-making processes.
- While the SVM model shows promise, additional analysis with a larger dataset and a broader range of features could provide more robust conclusions.
- Exploring other evaluation metrics and fine-tuning hyperparameters could further enhance the predictive power of the models

Introduction

SpaceX's revolutionary reusable rockets are transforming spaceflight. Landing site selection is critical for mission success, cost reduction, and environmental sustainability. This analysis leverages historical launch data to identify patterns in successful landings and optimize future operations. By understanding the factors influencing landing success, we aim to support SpaceX's continued innovation and leadership in the space industry. This data-driven approach ensures that each landing is as safe, efficient, and environmentally conscious as possible.

Introduction

Optimal Landing Site Selection: Which landing sites consistently yield the highest success rates, taking into account factors like weather patterns, geographical location, and infrastructure?

Risk Mitigation: Are there specific landing sites associated with higher risks of failure due to environmental factors or technical challenges? Identifying these can inform risk mitigation strategies.

Cost Optimization: Do certain landing sites offer cost advantages due to proximity to launch sites, fuel efficiency, or reduced recovery efforts? Optimizing landing site choices can lead to significant cost savings.

Predictive Modeling: Can we develop predictive models that accurately forecast landing success based on various factors such as weather conditions, booster type, payload, and landing site? Such models can aid in pre-flight decision-making.

Continuous Improvement: By identifying patterns and trends in successful landings, can we glean insights that guide improvements in landing technology, procedures, and overall mission planning?

Section 1

Methodology

Methodology

Executive Summary

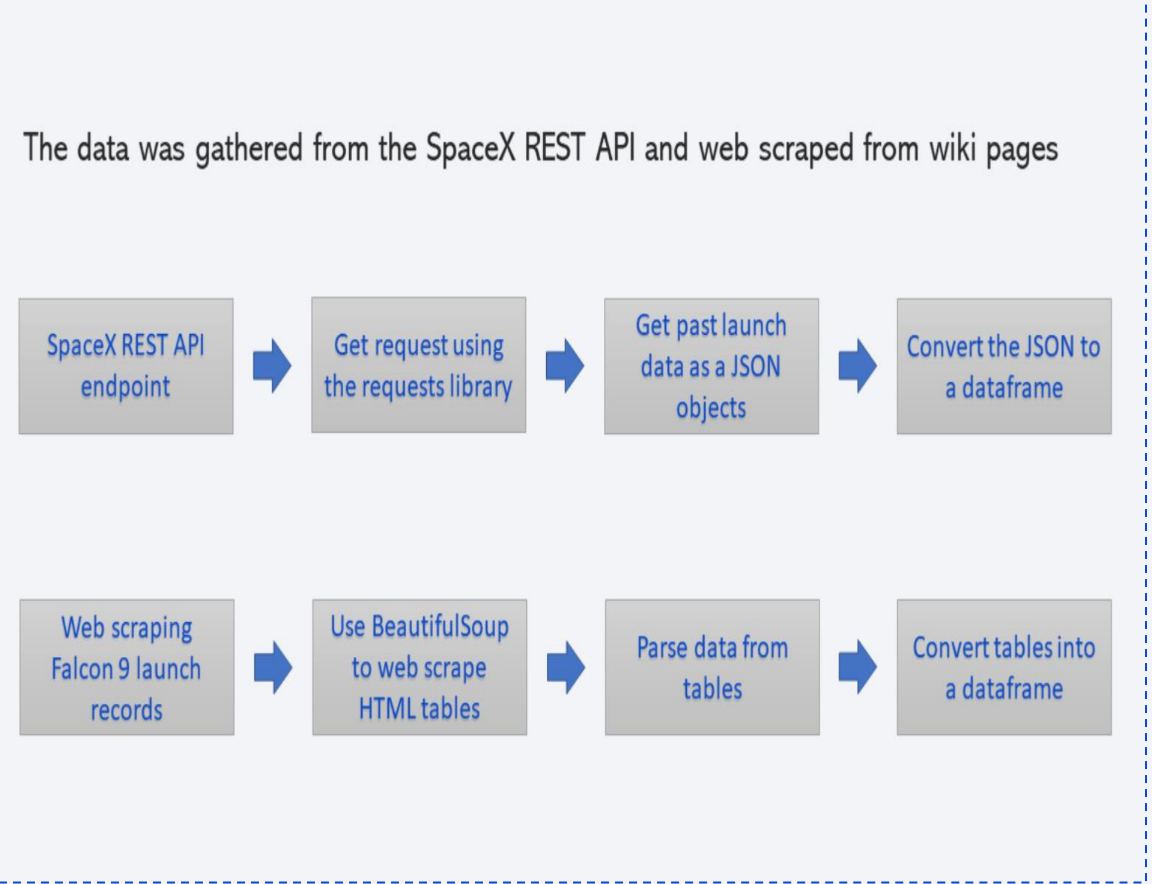
- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- SpaceX's open API provides detailed records of each launch, including the landing site and whether the landing was successful or not.
- By analyzing this data, we can gain valuable insights into the effectiveness of different landing sites and identify patterns that could lead to improvements in future missions.
- This analysis could also help optimize the selection of landing sites for specific missions, balancing factors such as safety, efficiency, and cost-effectiveness.

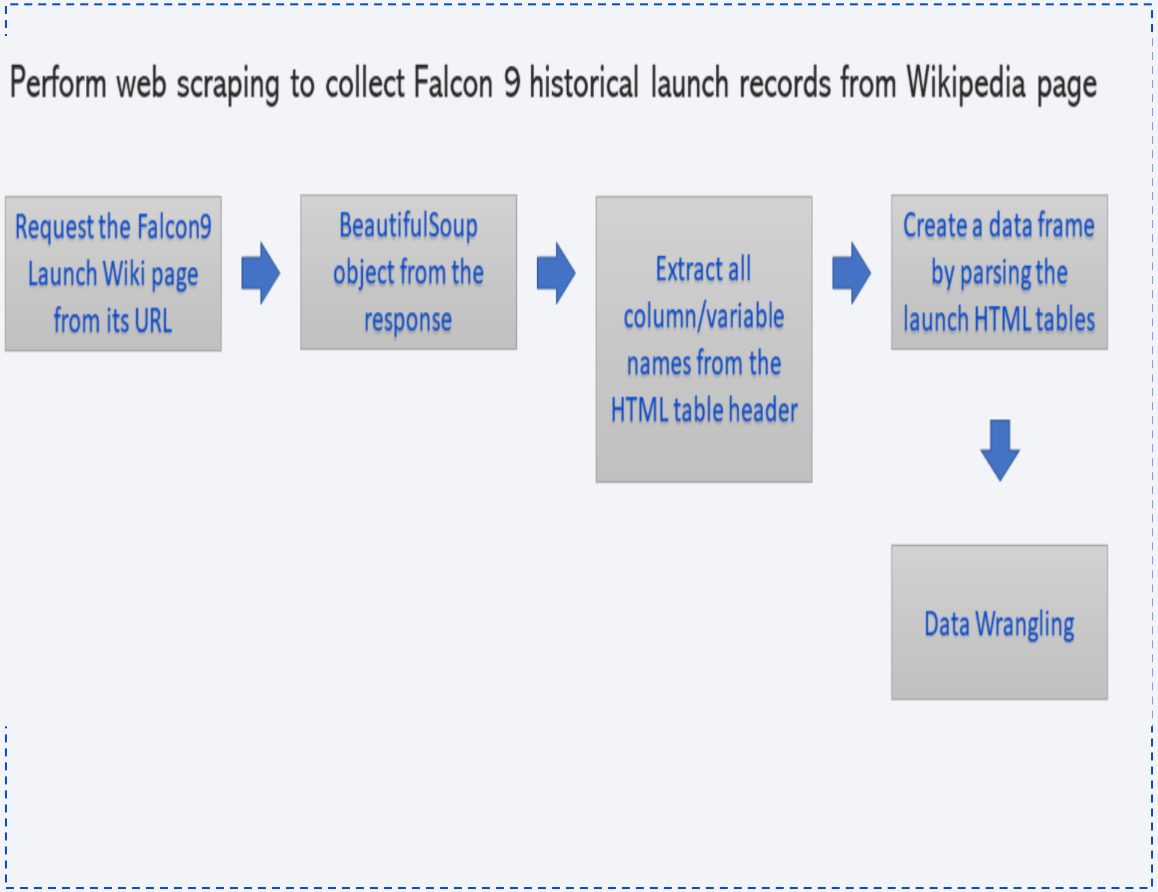
Data Collection – SpaceX API

- https://github.com/abraham1907/DATA-SCIENCE-IBM/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

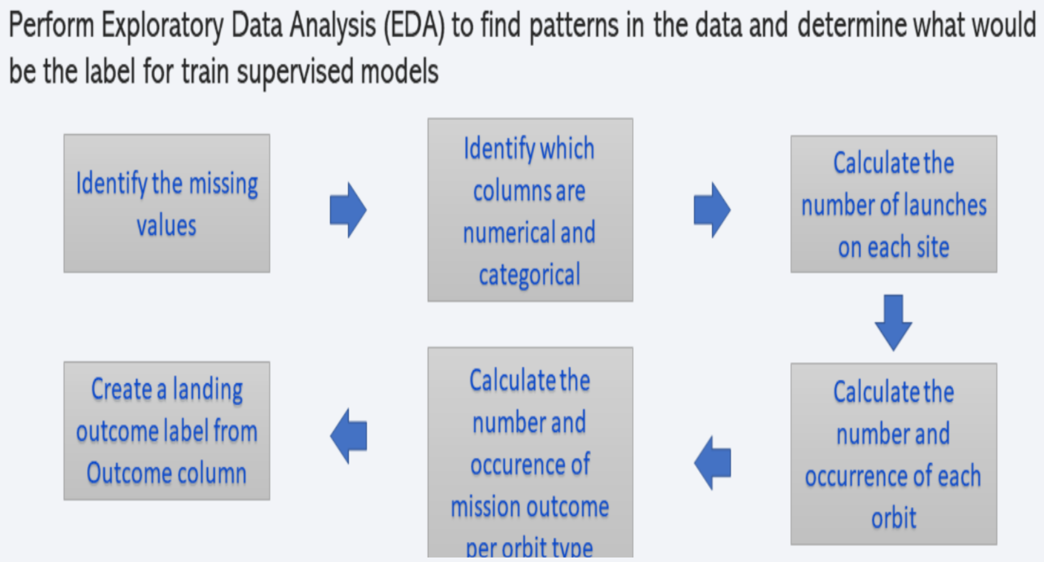
- https://github.com/abraham1907/DATA-SCIENCE-IBM/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Fetched structured launch data from SpaceX's API.
- Cleaned data by handling missing values and transforming features.
- Split data into training and testing sets for model development.
- Used cross-validation for model evaluation and hyperparameter tuning.
- Trained and compared various machine learning models to predict landing success.

https://github.com/abraham1907/DATA-SCIENCE-IBM/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- **Summary of charts that were plotted:**
- Catplot to visualize the relationship between Flight Number and Payload.
- Catplot to visualize the relationship between Flight Number and Launch Site.
- Catplot to visualize the relationship between Payload and Launch Site.
- Bar chart to visualize the relationship between success rate of each Orbit type.
- Catplot to visualize the relationship between Flight Number and Orbit type.
- Catplot to visualize the relationship between Payload and Orbit type.
- Line chart to visualize the launch success yearly trend.
- https://github.com/abraham1907/DATA-SCIENCE-IBM/blob/main/edadataviz.ipynb

EDA with SQL

Display the names of the unique launch sites in the space mission:

- SELECT DISTINCT(launch_site) FROM SPACEXTBL;

Display 5 records where launch sites begin with the string 'CCA':

- SELECT * FROM SPACEXTBL WHERE launch_site LIKE 'CCA%' LIMIT 5;

Display the total payload mass carried by boosters launched by NASA (CRS):

- SELECT SUM(payload_mass_kg) AS TOTAL_PAYLOAD_MASS FROM SPACEXTBL WHERE customer='NASA (CRS)';

Display average payload mass carried by booster version F9 v1.1:

- SELECT AVG(payload_mass_kg) AS AVG_PAYLOAD_MASS FROM SPACEXTBL WHERE booster_version='F9 v1.1';

List the date when the first successful landing outcome in ground pad was achieved:

- SELECT MIN(DATE) AS first_successful_landing FROM SPACEXTBL WHERE (landing_outcome)='Success (ground pad)';

https://github.com/abraham1907/DATA-SCIENCE-IBM/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- **Summary of map objects that were created and added to the Folium map:**
- **folium.Circle and folium.Marker:** To add a highlighted circle area with a text label on a specific coordinate for each launch site on the site map.
- **MarkerCluster object:** For simplifying a map containing many markers having the same coordinate.
- **MousePosition:** On the map to get the coordinates for a mouse over a point on the map.
- **folium.PolyLine object:** To draw a line between a launch site to its closest city, railway, and highway.
- https://github.com/abraham1907/DATA-SCIENCE-IBM/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- **Summary of map objects that were created and added to the Folium map:**
- **folium.Circle and folium.Marker:** To add a highlighted circle area with a text label on a specific coordinate for each launch site on the site map.
- **MarkerCluster object:** For simplifying a map containing many markers having the same coordinate.
- **MousePosition:** On the map to get the coordinates for a mouse over a point on the map.
- **folium.PolyLine object:** To draw a line between a launch site to its closest city, railway, and highway.
- https://github.com/abraham1907/DATA-SCIENCE-IBM/blob/main/lab_jupyter_launch_site_location.ipynb

Predictive Analysis (Classification)

- Summary of the model development process used to predict if the first stage will land given the data from the preceding labs:

- Creation of a NumPy array from the column Class in data.
- Data standardization.
- Use of the function train_test_split to split the data X and Y into training and test data.
- Searching for the best Hyperparameters for Logistic Regression, SVM, Decision Tree and KNN classifiers.
- Searching for the method that performs best using test data.Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose
- https://github.com/abraham1907/DATA-SCIENCE-IBM/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Results

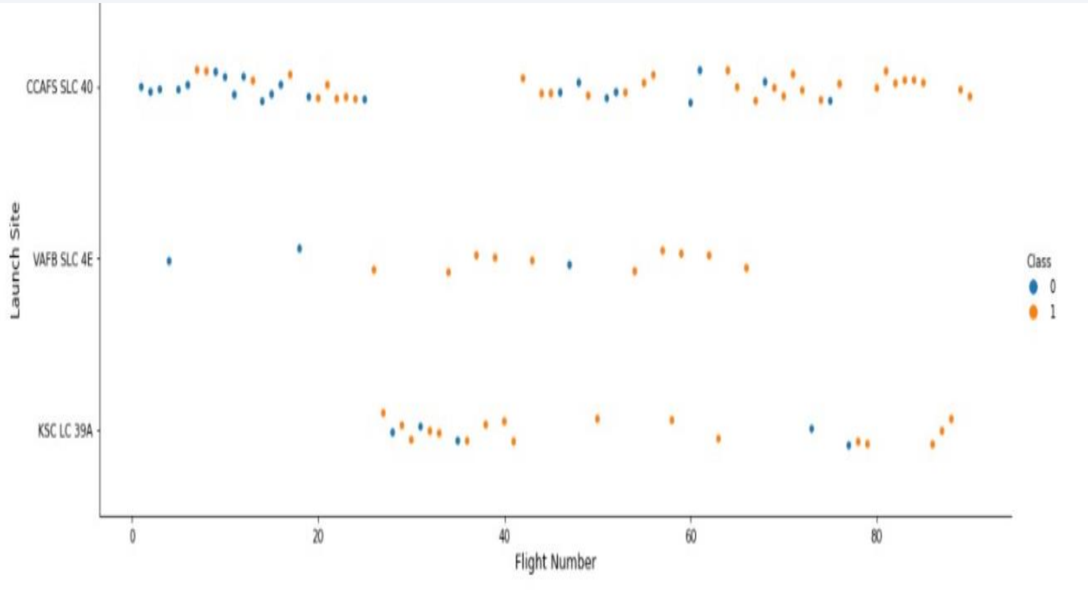
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

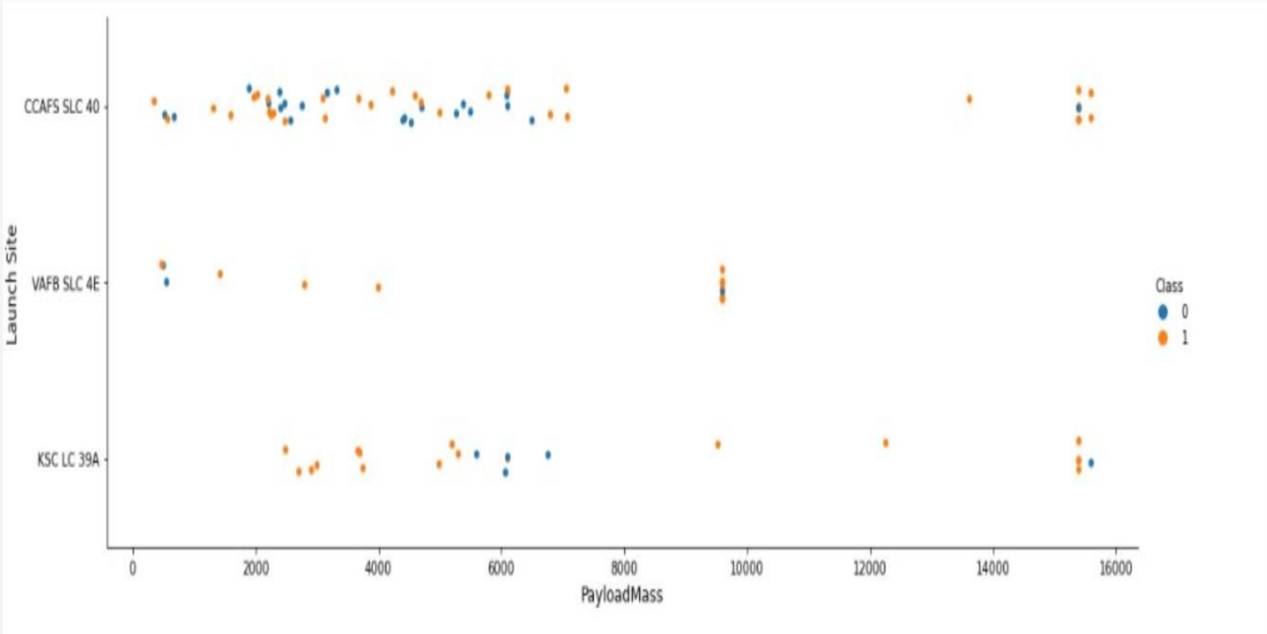
Flight Number vs. Launch Site

- Show a scatter plot of Flight Number vs. Launch Site



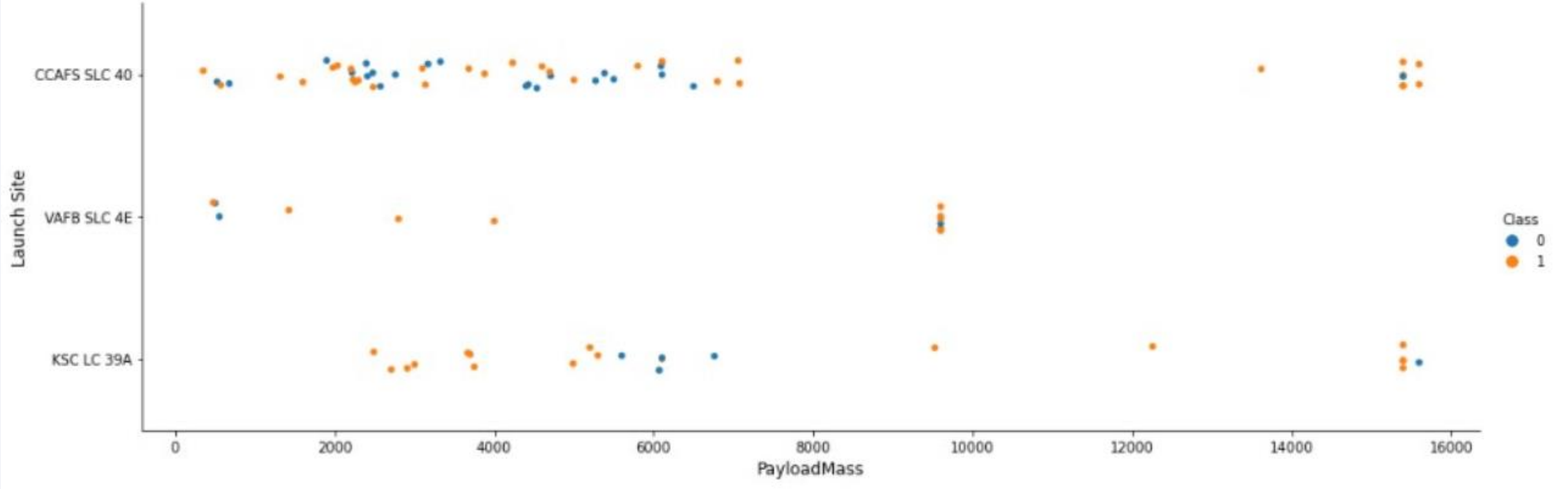
- With time, the successful rate has increased for every Launch Site, especially for CCAFS SLC 40, where the majority of launches are concentrated.
- VAFB SLC 4E and KSC LC 39A have a higher successful rate but represent one-third of the total launches.

- Show the screenshot of the scatter plot with



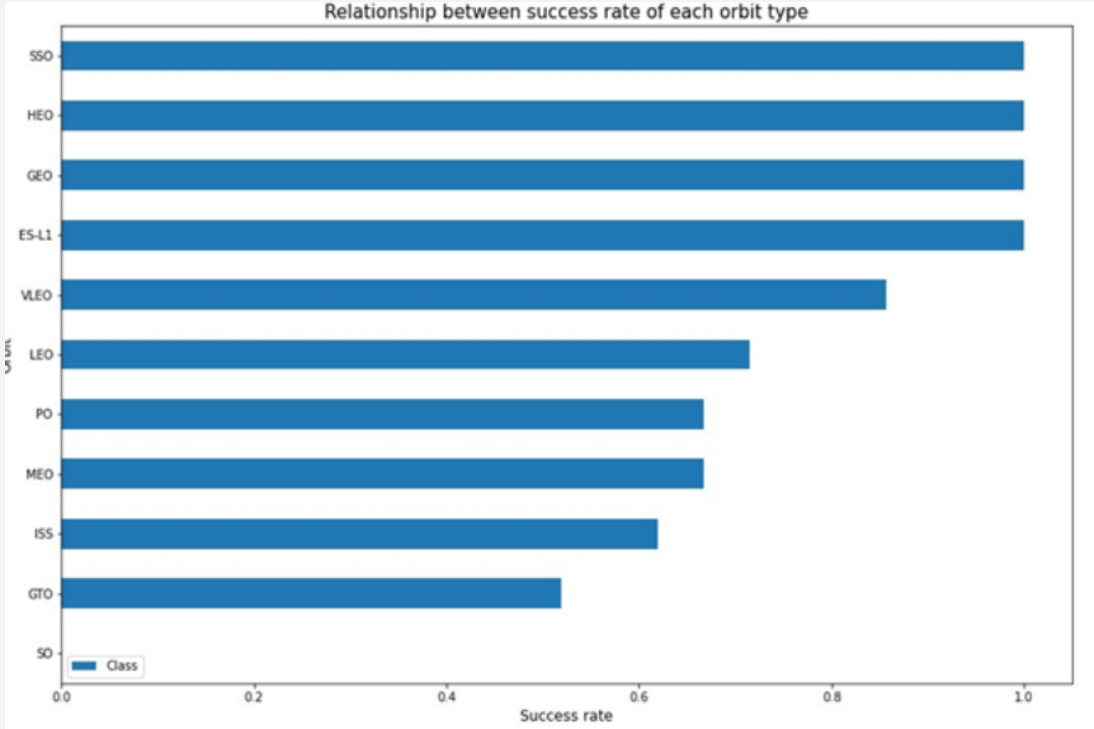
- In VAFB SLC launch site there are no rockets launched for heavy payload mass (greater than 10000kg).
- In KSC LC launch site there are no rockets launched for lower payload mass (less than 2500kg).
- CCAFS SLC has launched rockets less than 7500kg and more than 13000kg payload mass but not in between.

Payload vs. Launch Site



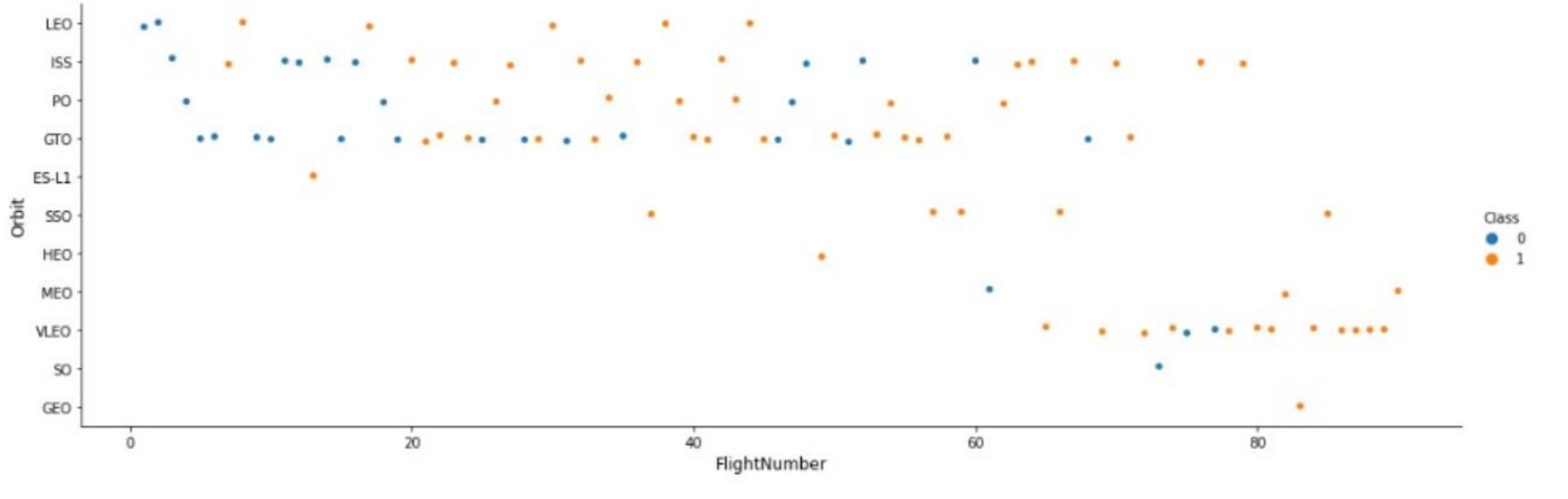
- VAFB SLC: No rockets have been launched from this site with heavy payloads exceeding 10,000 kg.
- KSC LC: This site has not launched any rockets with payloads below 2,500 kg.
- CCAFS SLC: Rockets launched from this site have either had payloads less than 7,500 kg or more than 13,000 kg, with no launches falling in the range between these values.

Success Rate vs. Orbit Type



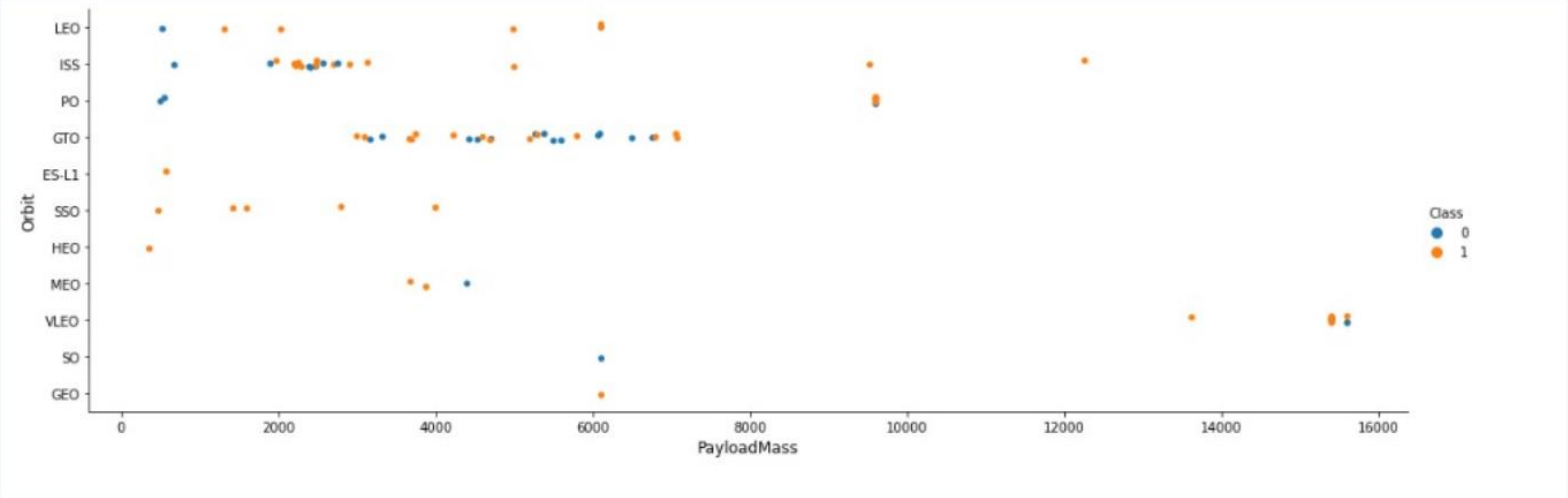
SSO, HEO,GEO and E5-11 Have the highest success rates

Flight Number vs. Orbit Type



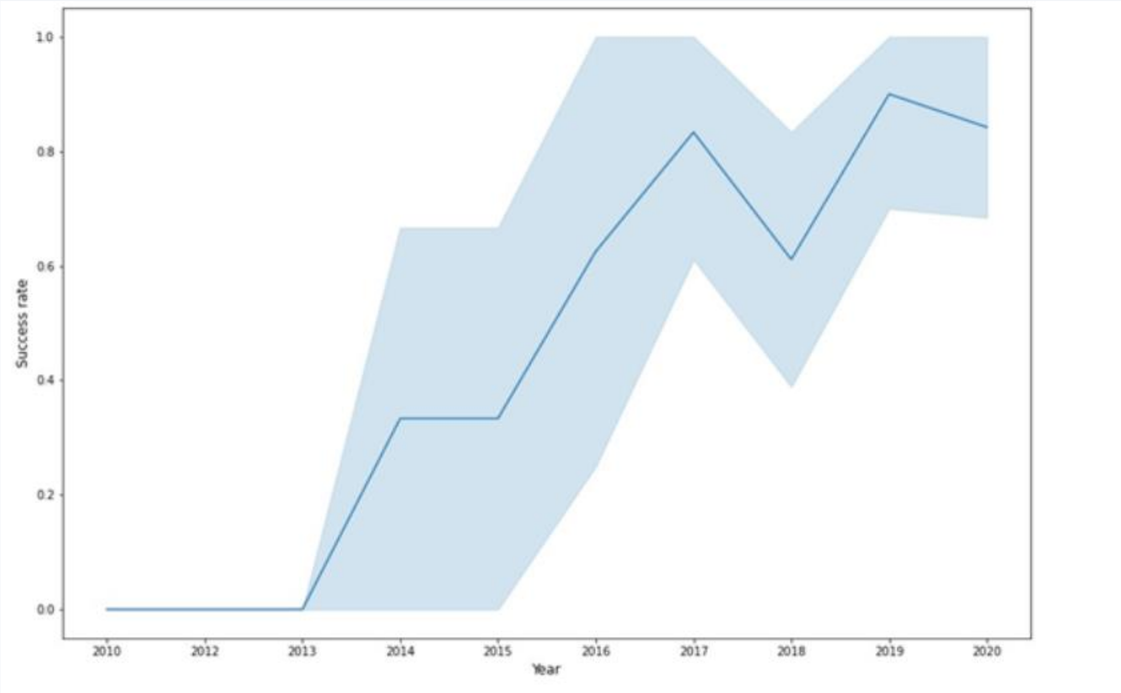
- As expected, there are more failures at the beginning of the series of launches, but after the first 40 launches, the ratio improves by reducing the 50 percent of unsuccessful landings.
- GTO and ISS orbits have the higher concentration of launches with the lowest ratio of successful landings.
- The orbits with higher success rates have one or just a few numbers of launches.

Payload vs. Orbit Type



- The image provides these conclusions:
- There is a visible limit of payload around 7600 kg. Less than 10 launches exceed that limit.
- With heavy payloads, the successful landing rate is higher for Polar, LEO, and ISS orbits.
- For GTO orbits, it is difficult to determine the successful landing rate as both successful and failed landings have occurred.

Launch Success Yearly Trend



- The image above shows a line chart of yearly average success rate

All Launch Site Names

The query below returns the names of the unique launch sites

Query

```
%sql SELECT DISTINCT(launch_site) FROM SPACEXTBL;
```

Output

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

Launch Site Names Begin with 'CCA'

- The sql query below returns 5 records where launch sites begin with `CCA`
- Query

```
%sql SELECT * FROM SPACEXTBL WHERE launch_site LIKE 'CCA%' LIMIT 5;
```

- Output

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

- The query below calculates the total payload carried by boosters from NASA

- Query

```
%sql SELECT SUM(payload_mass__kg_) AS TOTAL_PAYLOAD_MASS FROM SPACEXTBL WHERE customer='NASA (CRS)';
```

- output

total_payload_mass
45596

Average Payload Mass by F9 v1.1

- The query below calculates the average payload mass carried by booster version F9 v1.1
- Query

```
%sql SELECT AVG(payload_mass__kg_) AS AVG_PAYLOAD_MASS FROM SPACEXTBL WHERE booster_version='F9 v1.1';
```

- output

avg_payload_mass
2928

First Successful Ground Landing Date

- The query below shows the dates of the first successful landing outcome on ground pad.
- Query

```
%sql SELECT MIN(DATE) AS first_successful_landing FROM SPACEXTBL WHERE (landing_outcome)='Success (ground pad)';
```

- Output

first_successful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Below is a query that lists the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- Query

```
%sql SELECT booster_version, payload_mass__kg_, landing_outcome FROM SPACEXTBL \
      WHERE landing_outcome='Success (drone ship)' AND (payload_mass__kg_ BETWEEN 4000 AND 6000)
```

- output

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)

Total Number of Successful and Failure Mission Outcomes

- The quey below shows the total number of successful and failure mission outcomes
- Query

```
%sql SELECT mission_outcome, COUNT(mission_outcome) AS TOTAL FROM SPACEXTBL GROUP BY mission_outcome;
```

- Output

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

Boosters Carried Maximum Payload

- Below is a query that lists the names of the booster which have carried the maximum payload mass

- Query

```
%sql SELECT DISTINCT(booster_version), (SELECT MAX(payload_mass__kg_) AS "maximum_payload_mass" FROM SPACEXTBL) FROM SPACEXTBL
```

- Output

booster_version	maximum_payload_mass
F9 B4 B1039.2	15600
F9 B4 B1040.2	15600
F9 B4 B1041.2	15600
F9 B4 B1043.2	15600
F9 B4 B1039.1	15600

2015 Launch Records

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

Query:

```
%sql SELECT landing_outcome, booster_version, launch_site, DATE FROM SPACEXTBL WHERE landing_outcome LIKE '%Failure (drone ship)%' /
```

Outcome:

landing_outcome	booster_version	launch_site	DATE
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015-01-10
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015-04-14

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

- The below query shows th landing outcomes arranged in descending order

```
%sql SELECT landing_outcome, COUNT(landing_outcome) AS "total" FROM SPACEXTBL WHERE (DATE BETWEEN '2010-06-04' AND '2017-03-20')
```

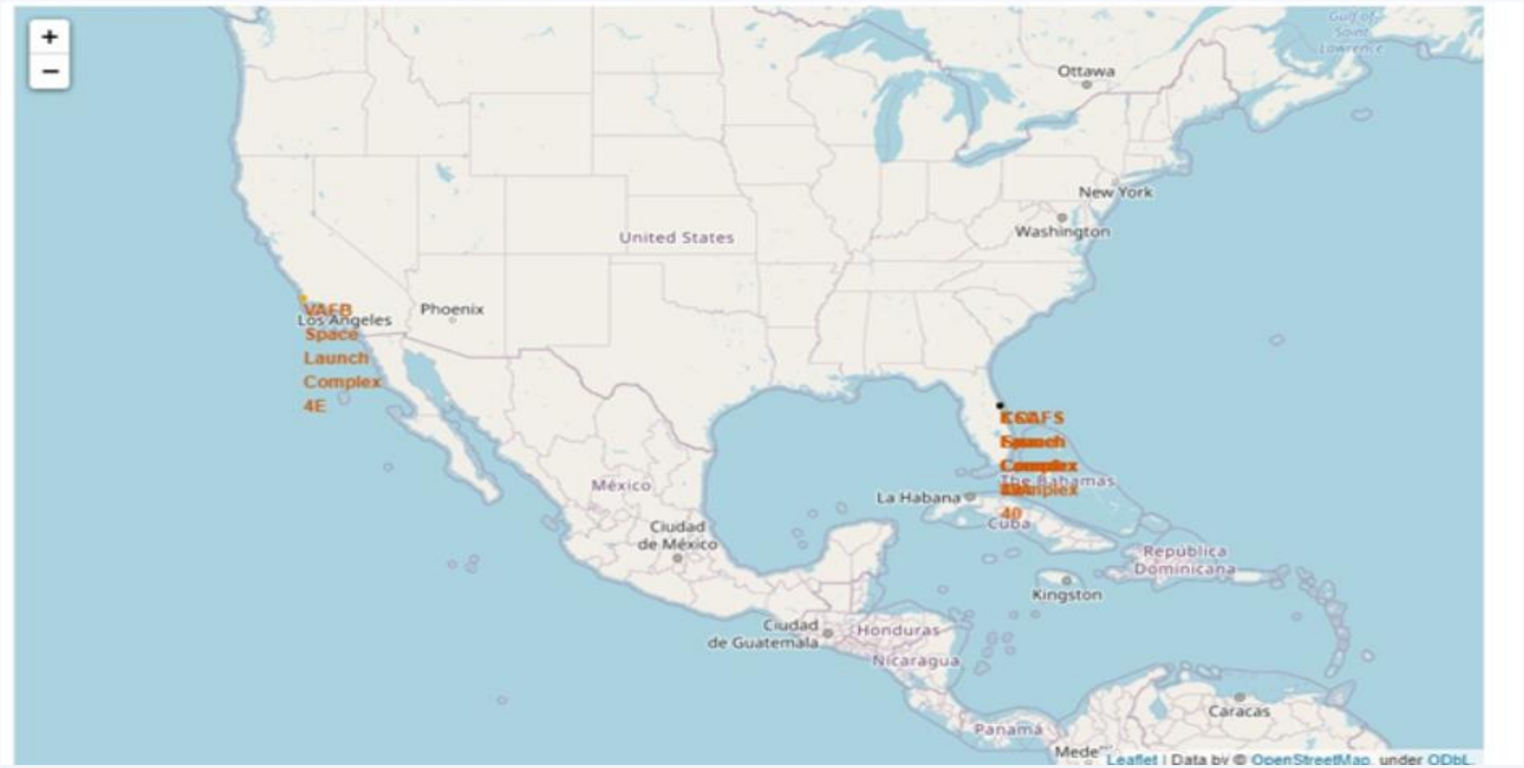
Done.

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

Section 3

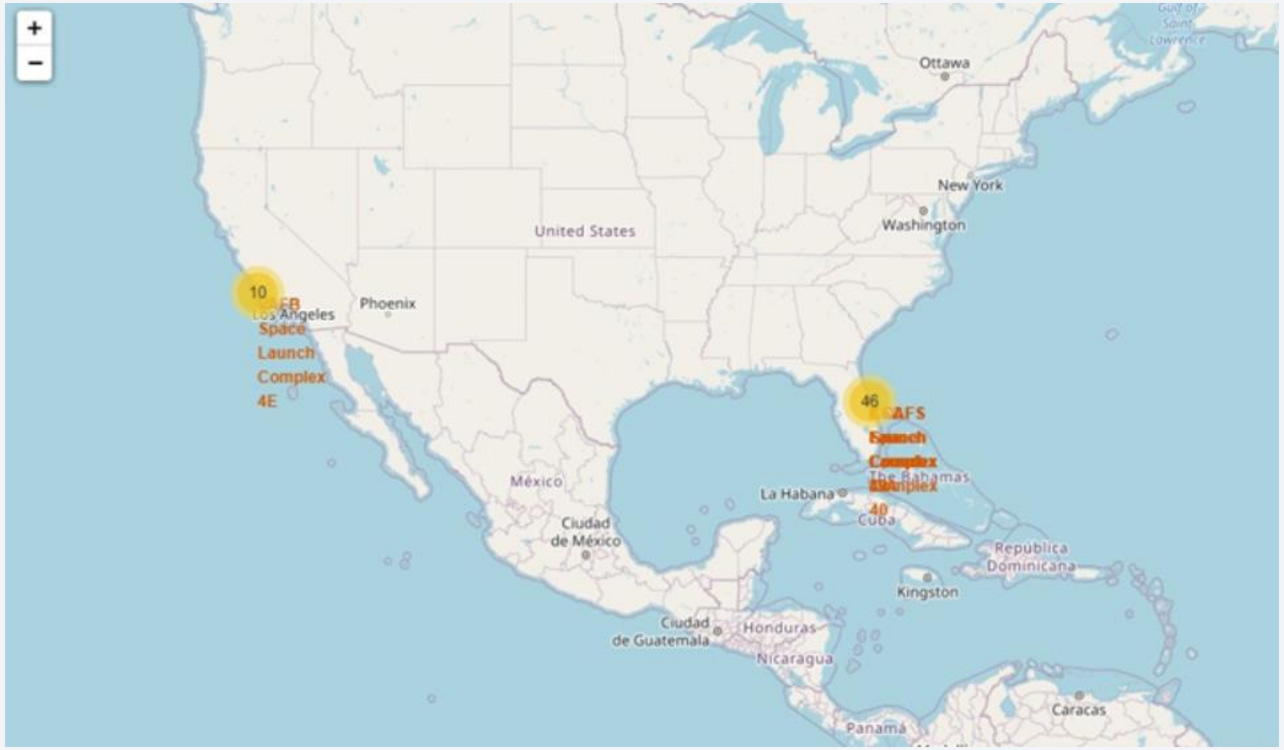
Launch Sites Proximities Analysis

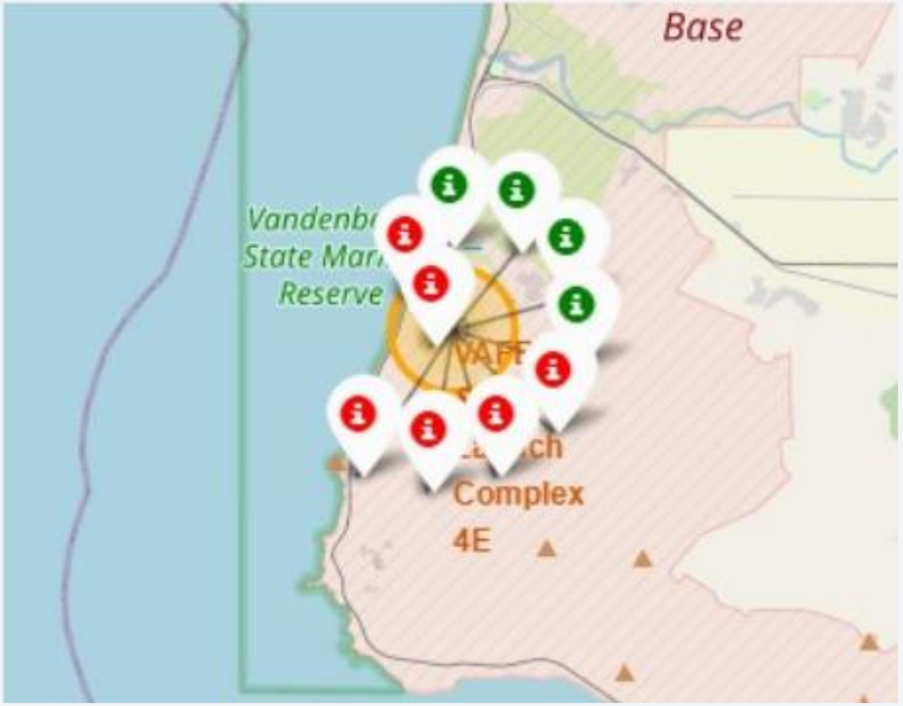
All Launch Sites



- All launch sites anre near the coast in a restricted area

Successful and failed launches for each site





- The first map displays the density and distribution of launch sites, using clusters to represent multiple launches from the same location.
- The second map provides a visual representation of the success or failure of each launch. Green markers indicate successful launches, while red markers indicate failed launches.

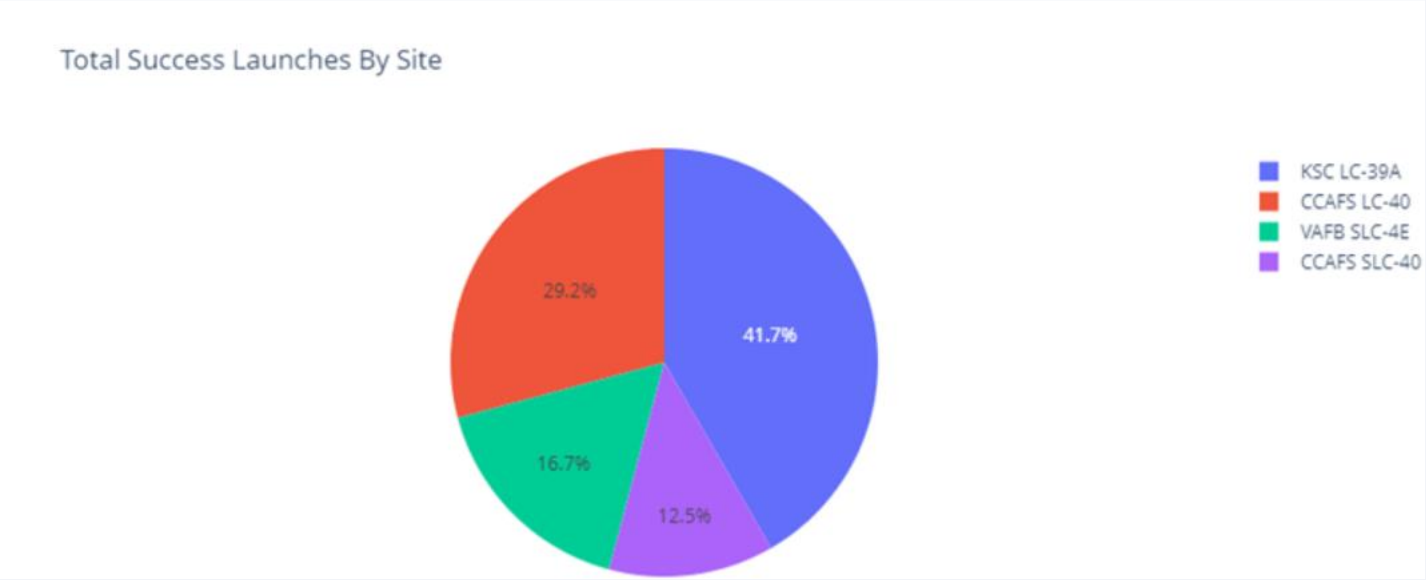
Launch site and proximity

- Launch site is near rail way line, raods and coast. City is fairly far

Section 4

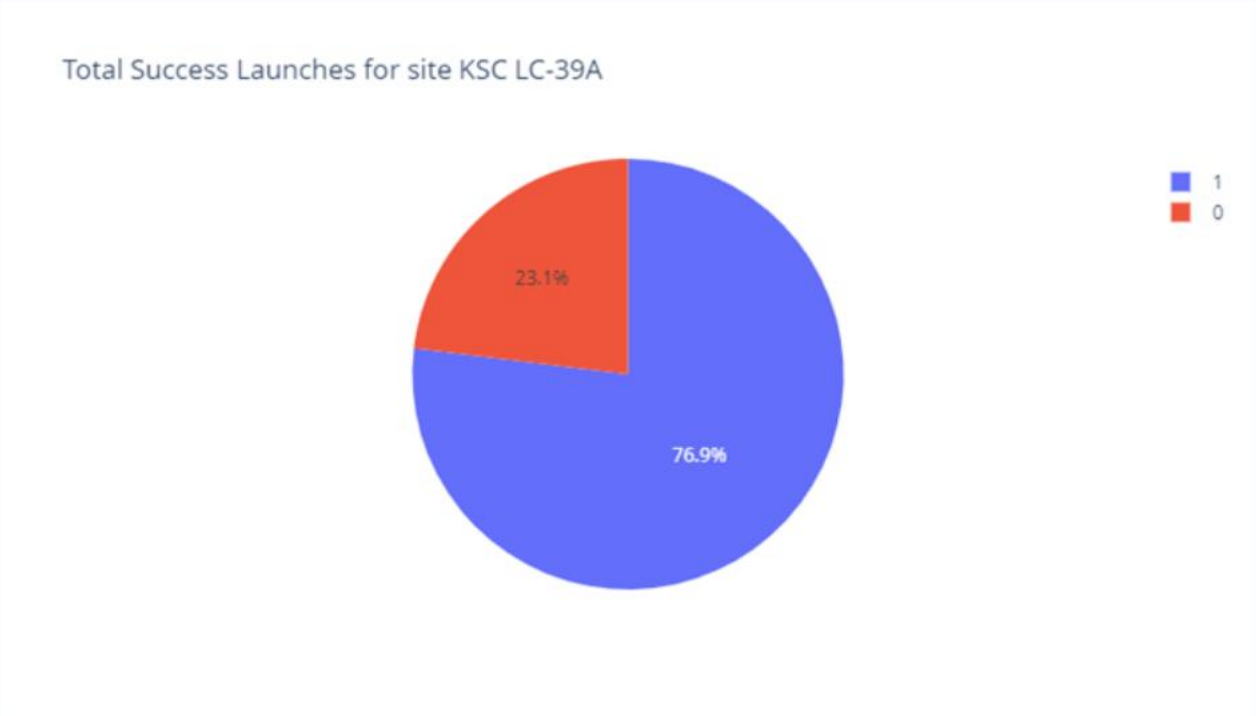
Build a Dashboard with Plotly Dash

Total Success by launch Sites



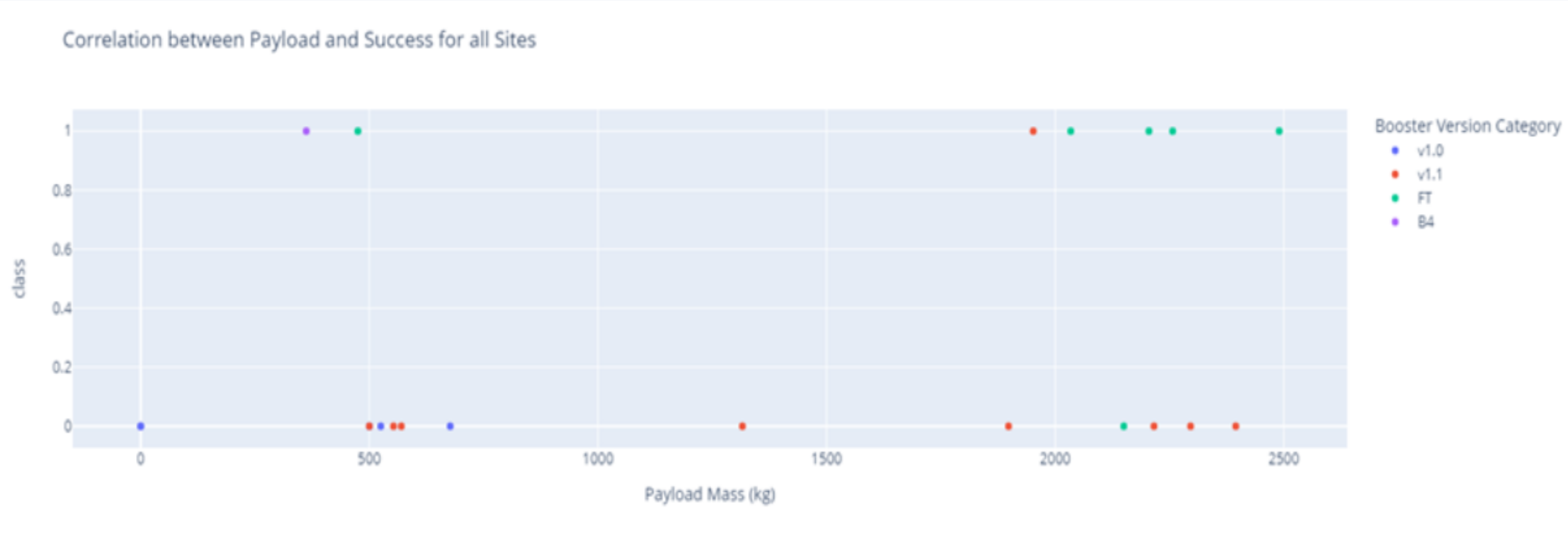
- KSC LC-39A has a highest launch success rate

KSC LC-39A



- KSC LC-39A had a high launch success rate compared to the failure rate

Payload vs Launch Outcome

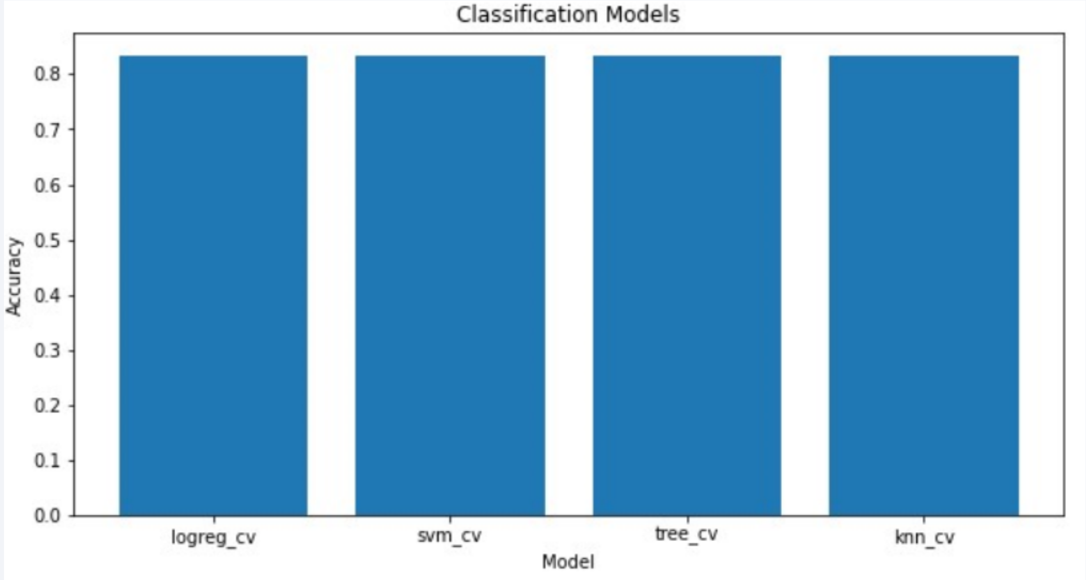


- The scatter plot shows the distribution of launch outcomes (success vs. failure) across different payload mass ranges for all launch sites.
- 2500-5000 kg Payload Range: This range has the highest concentration of successful launches.
- 0-2500 kg Payload Range: While it has the most failed launches, the overall difference in success rates between the three payload ranges (0-2500 kg, 2500-5000 kg, and 10000+ kg) is relatively small.

Section 5

Predictive Analysis (Classification)

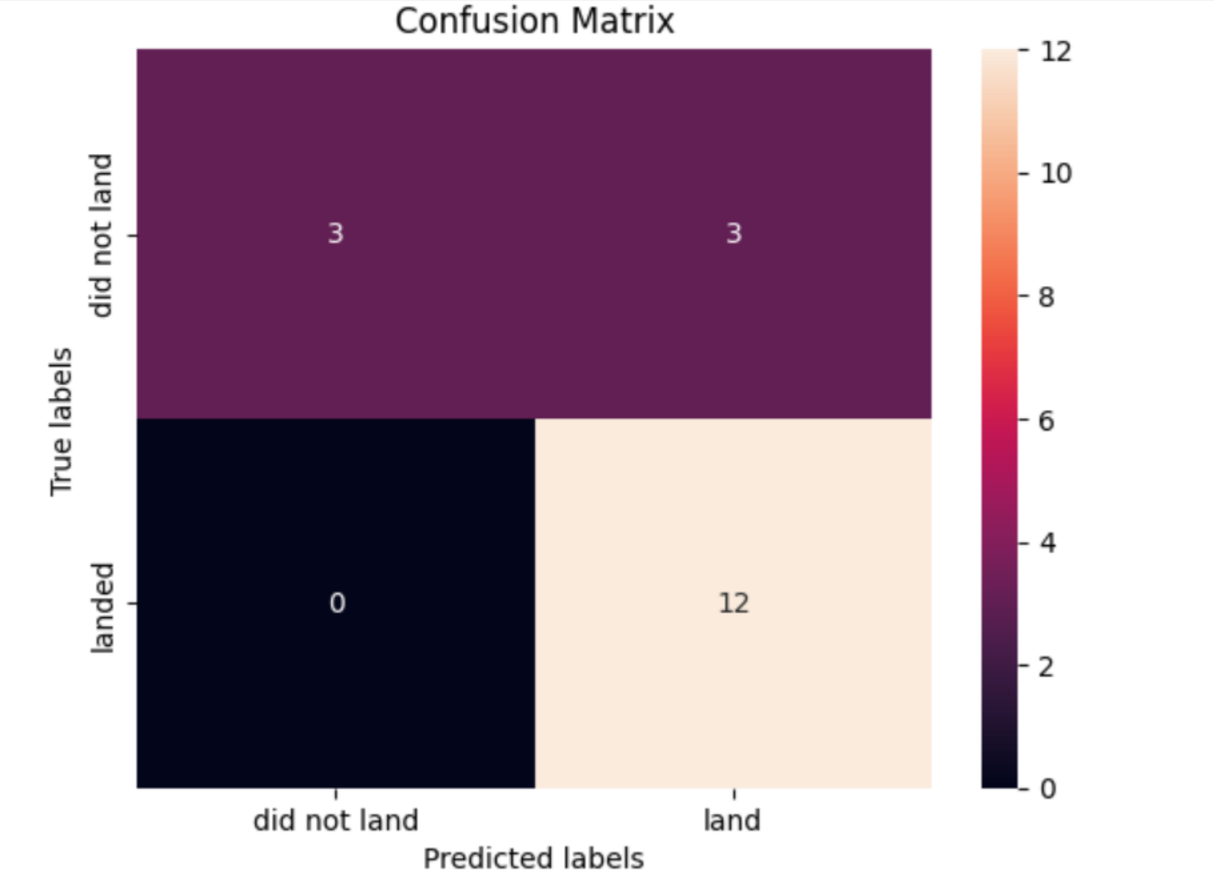
Classification Accuracy



- Almost all models have the same acuracy
- The decision tree has the highest classification accuracy

Confusion Matrix

- All models have a simila confusion matrix



Conclusions

- All the algorithms are giving the similar accuracy, they all perform practically the same.
- By using our machine learning model, we can predict if the first stage of our competitor will land and determine the cost of a launch….

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!