Predicting the success of SpaceX Falcon 9's 1st stage landing using Data Science

Syed Ahmed Mahmood

1st July 2023

Outline

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

Executive Summary

- Methodology

1. Data Collection: Data was gathered using the SpaceX REST API and by web scraping the Wikipedia page for Falcon 9 launches.
2. Data Wrangling: The collected data was cleaned and processed to create a target variable that indicates if the 1st stage landed successfully or not.
3. Data Exploration: Various data visualization techniques were utilized to explore the data, focusing on features such as payload mass, launch site, flight number. We explored the success rates of different launch sites and orbit types, then we also observed the yearly trends of launch successes.
4. Data Analysis: SQL and python were used to analyze the data, this involved calculating statistics such as total payload, payload range for successful launches, and the total number of successful and failed outcomes.
5. Data Visualization: Success rates of launch sites were visualized on a map with markers indicating the successes and failures and we also visualized the proximity to landmarks such as coastlines, highways, railways and cities.
6. Predictive Modeling: Models, including logistic regression, support vector machine (SVM), decision tree, and K-nearest neighbor (KNN), were trained to predict the outcome of the 1st stage landing.

- Results

1. The overall launch success rate has improved over the years.
2. KSC LC-39A is the launch site with the highest success rate.
3. ES-L1, GEO, HEO, and SSO are orbits that have a 100% success rate.
4. Most of the launch sites are near the equator with all of them being very close to a coastline.
5. All trained models performed similarly on the test set with decision tree being the best performing.

Introduction

Producing and launching a rocket is an expensive process, with most providers charging upwards of $165 million for each launch. However, SpaceX stands out by offering Falcon 9 rocket launches at a significantly lower cost of $62 million.

SpaceX achieves cost savings by implementing a reusable rocket design. They are able to reuse the first stage of their rockets, which is the lower part of the rocket that disconnects from the second stage containing the payload.

The successful landing of the first stage is crucial in determining the overall cost of a launch. By predicting whether the Falcon 9 first stage will land successfully, valuable information can be obtained for a rival company that wishes to compete with SpaceX for a rocket launch contract.

In this project, our objective is to develop a predictive model that can determine the likelihood of a successful landing for the Falcon 9 first stage.

Goals:

- Explore which features affect the landing success of the 1st stage.
- Analyze the pattern of landing success rate over the years.
- Find the best model to predict the success of the 1st stage landing.

Methodology

Methodology

- Data collection methodology:
  - Data was collected using the SpaceX REST API and by web scraping the Wikipedia page.
- Perform data wrangling
  - Filtered the data, handled missing values and applied one hot encoding to convert categorical features into numerical 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
  - Split the data into train and test sets and trained the model.
  - Tuned the hyperparameters and evaluated the classification models

Data Collection - API

1. Request the launch data from the SpaceX REST API.
2. Convert the JSON response into a dataframe.
3. The launch data only contains IDs for the features we require, so we create functions to collect the actual data using these IDs and the SpaceX REST API.
4. We store the data we need from the original dataframe, and the data collected in the previous step into a dict, which we then convert into a new dataframe.
5. Filter the dataframe to only include Falcon 9 launches.
6. Replace missing values of Payload Mass using the mean of the column.
7. Export the dataframe as a csv file.

GitHub URL: https://github.com/AhmedMahmood19/Space-X/blob/main/spacex-data-collection-from-api.ipynb

Data Collection - Scraping

1. Request the launch data from the Wikipedia page for Falcon 9 launches.
2. Create a BeautifulSoup object from the HTML response.
3. Extract column names from the HTML table headers.
4. Parse the HTML tables, extract launch records and store them in a dict.
5. Convert the dict into a dataframe.
6. Export the dataframe as a csv file.

GitHub URL: https://github.com/AhmedMahmood19/Space-X/blob/main/spacex-web-scraping-from-wikipedia.ipynb

Data Wrangling

1. Calculate the number of launches for each site to find that *CCAFS SLC 40* had the most launches(55).
2. Calculate the number of launches for each orbit to find that *GTO* had the most launches(27).
3. Calculate the number of launches for each mission outcome to find that True ASDS(successfully landed on a drone ship) had the most launches(41).
4. The column "Outcome" has several types of successful and unsuccessful landings, but since we only care about successful or unsuccessful landing, we convert these values to 1 (successful landing) or 0 (unsuccessful landing) and store them in the target variable "Class".
5. We find the mean of "Class" to get a success rate of 0.66.
6. Export the dataframe as a csv file.

GitHub URL: https://github.com/AhmedMahmood19/Space-X/blob/main/spacex-data-wrangling.ipynb

EDA with Data Visualization

- The following charts were made:
  1. Flight Number vs. Payload Mass (kg) (scatter plot)
  2. Flight Number vs. Launch Site (scatter plot)
  3. Payload Mass (kg) vs. Launch Site (scatter plot)
  4. Orbit type vs. Success rate (bar chart)
  5. Flight Number vs. Orbit type (scatter plot)
  6. Payload Mass (kg) vs. Orbit type (scatter plot)
  7. Year vs. Average Success Rate (line chart)
- Observe a relationship between variables by using scatter plots, these variables may be useful for machine learning if a relationship exists.
- Compare categories of a categorical variable with bar charts, it shows the relationships among the categories and a measured value.
- Observe trends over time using line charts.

GitHub URL: https://github.com/AhmedMahmood19/Space-X/blob/main/spacex-eda-and-sns-viz.ipynb

EDA with SQL

- Display names of unique launch sites.
- Display 5 records where launch site begins with 'CCA'.
- Display total payload mass carried by boosters launched by NASA (CRS).
- Display average payload mass carried by booster version F9 v1.1.
- List the date of first successful landing on ground pad.
- List the names of boosters which had success landing on drone ship and have payload mass between 4,000 kg and 6,000 kg.
- List the total number of successful and failed missions.
- List the names of booster versions which have carried the max payload mass.
- List the failed landing outcomes on drone ship, their booster version and launch site for the months in the year 2015
- Rank the count of landing outcomes between 2010-06-04 and 2017-03-20 in descending order.

GitHub URL: https://github.com/AhmedMahmood19/Space-X/blob/main/spacex-eda-with-sql.ipynb

Build an Interactive Map with Folium

- Red circle and marker at NASA Johnson Space Center with a popup label showing its name.
- Red circles and markers at all launch sites with a popup label showing their names, using their coordinates from the dataframe.
- Marker cluster at each launch site with green markers for successful launches and red markers for unsuccessful launches to visualize each launch site's success rate.
- PolyLines from the launch site CCAFS SLC40 to the nearest coastline, railway, highway, and city. Markers at each of these proximities labelled with the distance from the marker to the launch site.

GitHub URL: https://github.com/AhmedMahmood19/Space-X/blob/main/spacex_launch_site_locations_folium_viz.ipynb

Note: Use the following link instead since folium maps do not render on GitHub.

https://nbviewer.org/github/AhmedMahmood19/Space-X/blob/main/spacex_launch_site_locations_folium_viz.ipynb

Build a Dashboard with Plotly Dash

- Dropdown List with Launch Sites
  - Allow user to select all launch sites or a certain launch site
- Slider of Payload Mass Range
  - Allow user to select a payload mass range
- Pie Chart Showing Successful Launches
  - Allow user to see successful and unsuccessful launches as a percent of the total
- Scatter Plot Showing Payload Mass vs. Success Rate by Booster Version
  - Allow user to see the correlation between Payload and Launch Success

GitHub URL: https://github.com/AhmedMahmood19/Space-X/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

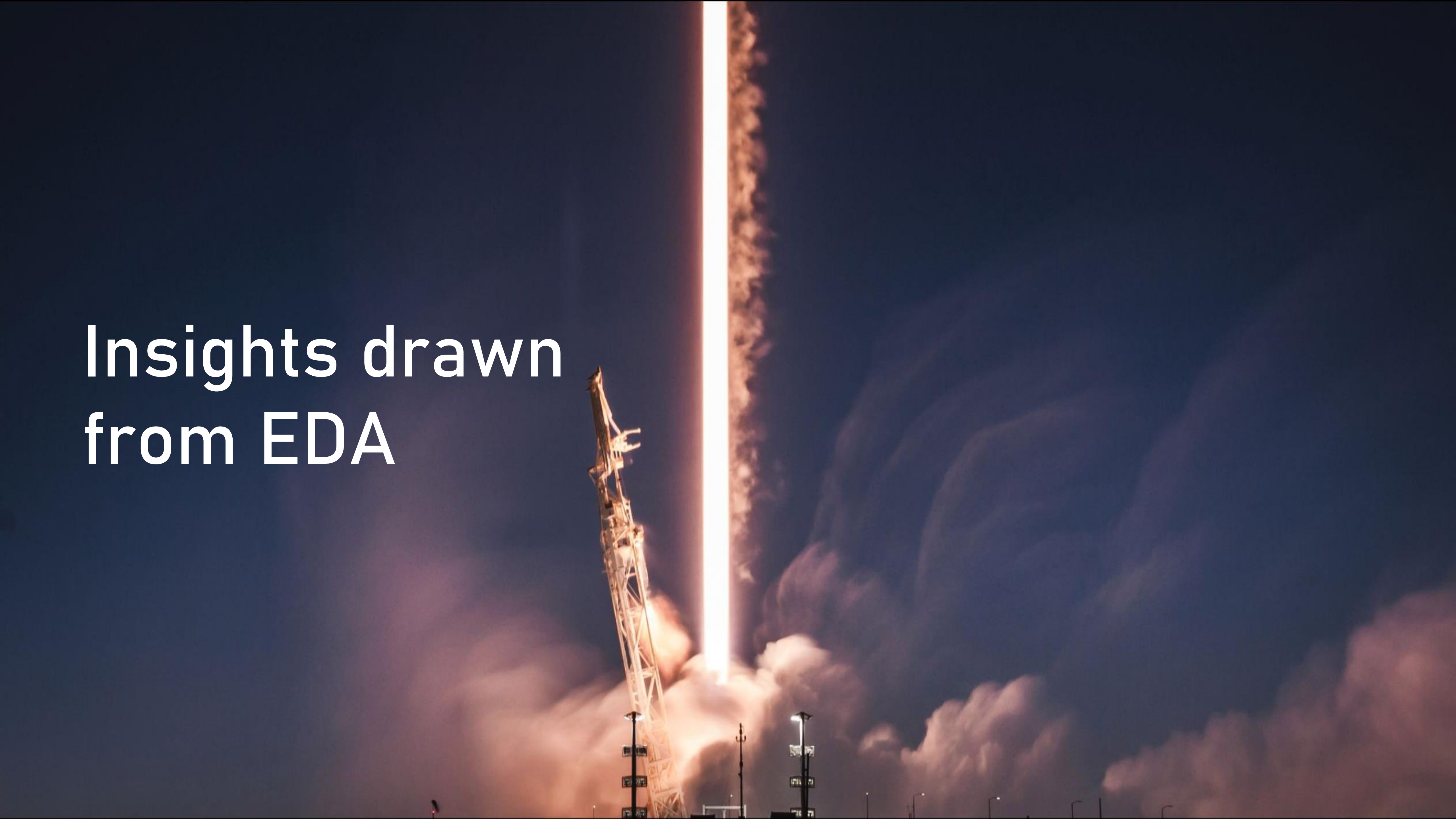
- Create a NumPy array Y from the column “Class” in the dataframe.
- Standardize the dataframe X with *StandardScaler*, then fit and transform it.
- Split X and Y into train and test sets with *train_test_split*.
- Use *GridSearchCV* with cv=10 to train several models and select the best hyperparameters.
- The models that we trained are Logistic Regression, Decision Tree, Support Vector Machine and K-Nearest Neighbors.
- For each model, calculate the accuracy on the test data using *.score()* and create a confusion matrix too.
- Using the evaluation metrics calculated above, we identify that Decision Tree slightly outperforms the rest of the models with an accuracy of 0.888.

GitHub URL: https://github.com/AhmedMahmood19/Space-X/blob/main/spacex_machine_learning_compared_4_models.ipynb

Results

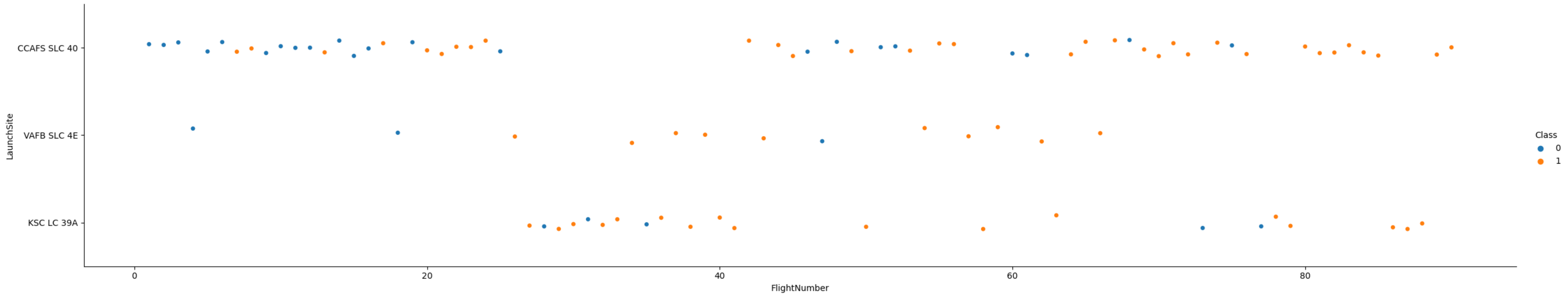
- EDA results:
    - Average Success Rate has improved over the years
    - KSC LC-39A has the highest success rate among the launch sites
    - Orbits ES-L1, GEO, HEO and SSO have a 100% success rate
- Visual Analysis results:
    - Most launch sites are near the equator, and all are close to the coast.
    - Launch sites are kept far away from cities to prevent damage due to a failed launch.
    - It is also kept at an adequate distance from highways and railways while still being close enough to transport equipment, materials and employees to the launch site.
- Predictive Analysis results
    - Decision Tree is the best performing predictive model for this dataset.

Insights drawn from EDA



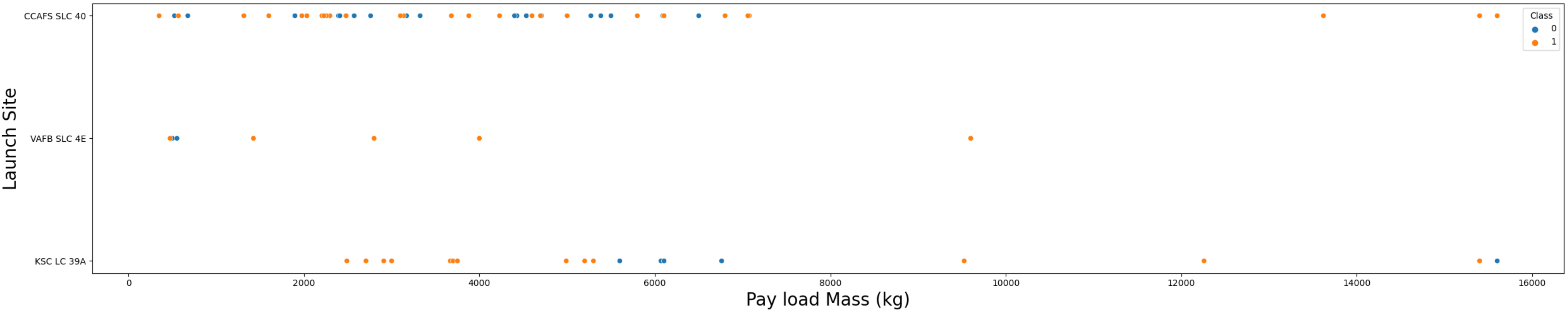
Flight Number vs. Launch Site

- The majority of the launches were from CCAFS SLC 40.
- Later flights were more likely to land successfully compared to earlier flights.
- Flights launched from VAFB SLC 4E and KSC LC 39A were more likely to succeed than those launched from CCAFS SLC 40.



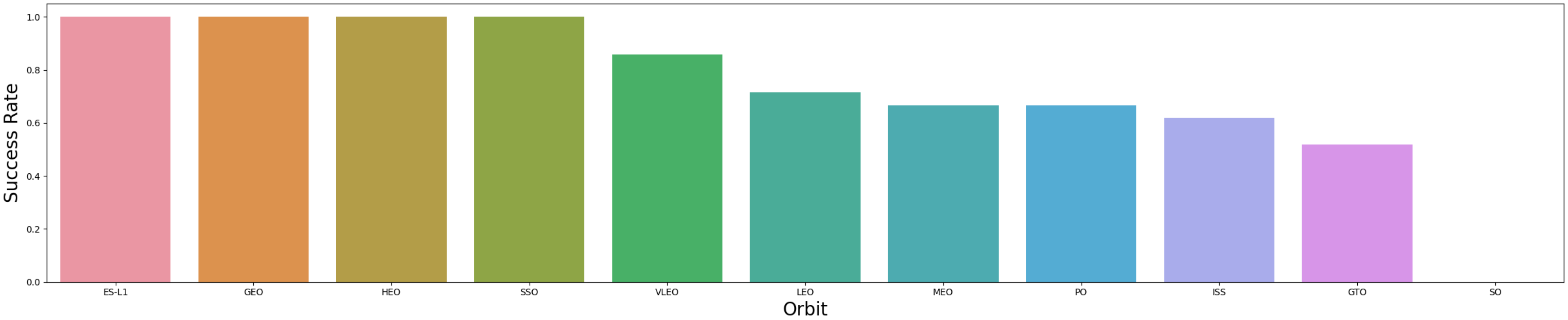
Payload vs. Launch Site

- Most launches with a payload greater than 7,000 kg successfully landed.
- Most launches have a payload mass lesser than 7000 kg.
- KSC LC 39A has a 100% success rate for launches less than 5,500 kg.
- Payload mass is always lesser than 10,000 kg for VAFB SKC 4E.
- Higher payload mass results in a higher success rate.



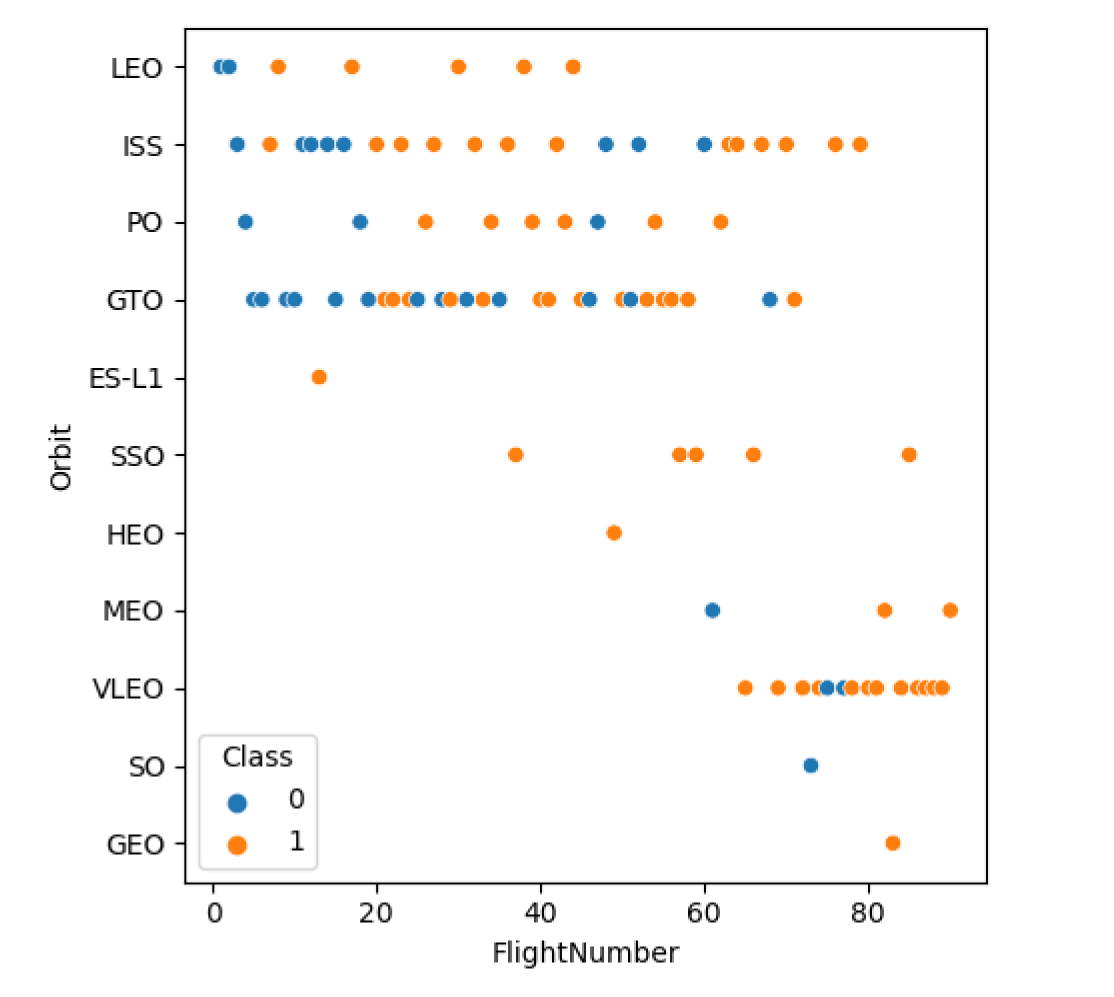
Success Rate vs. Orbit Type

- 100% Success Rate: ES-L1, GEO, HEO and SSO
- 50%-85% Success Rate: GTO, ISS, LEO, MEO, PO
- 0% Success Rate: SO



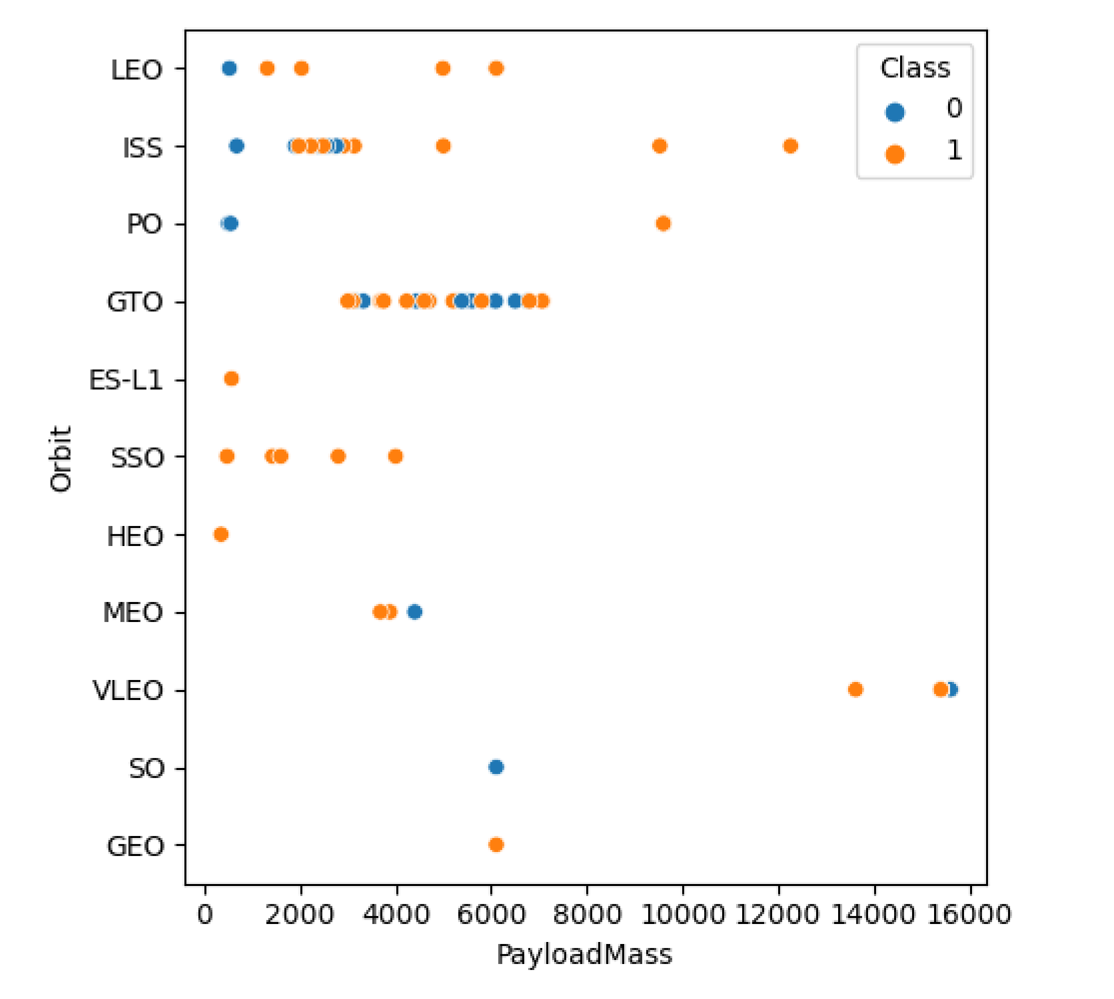
Flight Number vs. Orbit Type

- The success rate generally increases with the number of flights for all the orbits, especially in the case of LEO.
- The trend above is not followed by GTO.



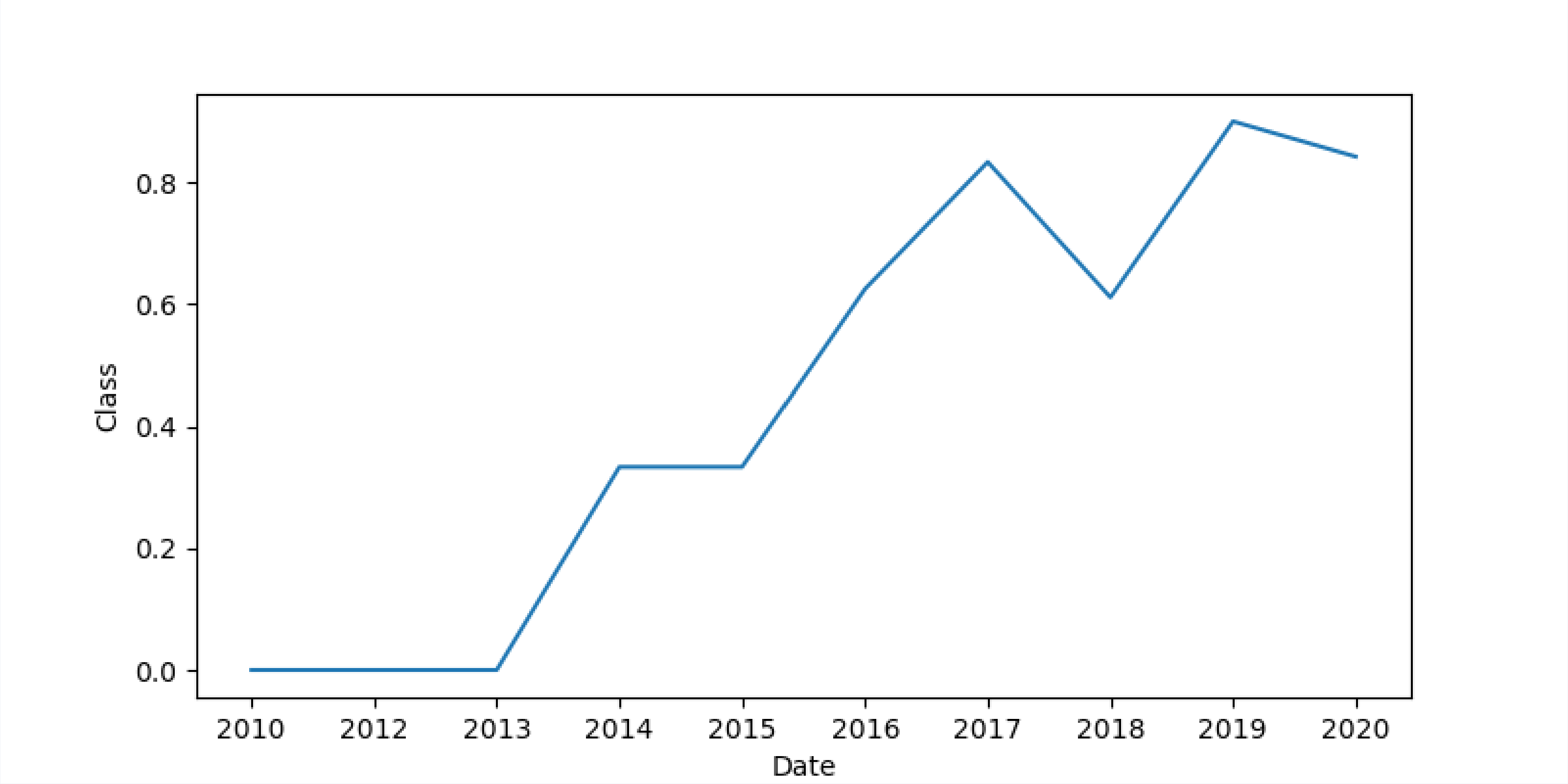
Payload vs. Orbit Type

- As the payload mass increases, the success rate increases for LEO, ISS and PO.



Launch Success Yearly Trend

- The success rate improved from 2013-2017 and 2018-2019.
- The success rate decreased from 2017-2018 and from 2019-2020.
- Overall, the success rate has vastly improved since 2013.



All Launch Site Names

- There are 4 launch sites, and their names are listed below in the query result.

```
%sql SELECT DISTINCT(LAUNCH_SITE) FROM SPACEXTBL WHERE LAUNCH_SITE IS NOT NULL;
```

```
 * 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'

- Shown below are 5 records where the launch site's name begins 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
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- The total payload mass carried by NASA boosters is 45596.0 kg.

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

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

SUM(PAYLOAD_MASS__KG_)
45596.0

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2534.66 kg.

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

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

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad is 22nd December 2015.

```
%sql SELECT MIN(DATE(substr(date,7,4)||'-'||substr(date,4,2)||'-'||substr(date,1,2))) as first_success_gro
```

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

first_success_groundpad
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Listed below are the names of boosters which have successfully landed on a drone ship and had a payload mass greater than 4000 but less than 6000.

```
%sql SELECT Booster_Version FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

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

- Shown below are the total number of successful and failure mission outcomes.

```
%sql SELECT mission_outcome, count(*) FROM SPACEXTBL GROUP BY MISSION_OUTCOME;
```

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

Mission_Outcome	count(*)
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Listed below are the names of the boosters which have carried the maximum payload mass.

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

```
 * 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

- Listed below are the records of all the failed landings on drone ships in 2015 shown with the booster versions and launch site names.

```
%sql SELECT substr(DATE, 4, 2) AS MONTH_NAME, LANDING_OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE LANDING_OUTCOME='Failure (drone ship)'
```

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

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

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

- Listed below are all the counts, in descending order, of each type of landing outcome between 2010-06-04 and 2017-03-20.

```
%sql SELECT LANDING_OUTCOME, COUNT(*) FROM SPACEXTBL WHERE DATE(substr(date,7,4)||'-'||substr(date,4,2)||'-'||substr(date,1,2)) BETWEEN DATE('2010-06-0
```

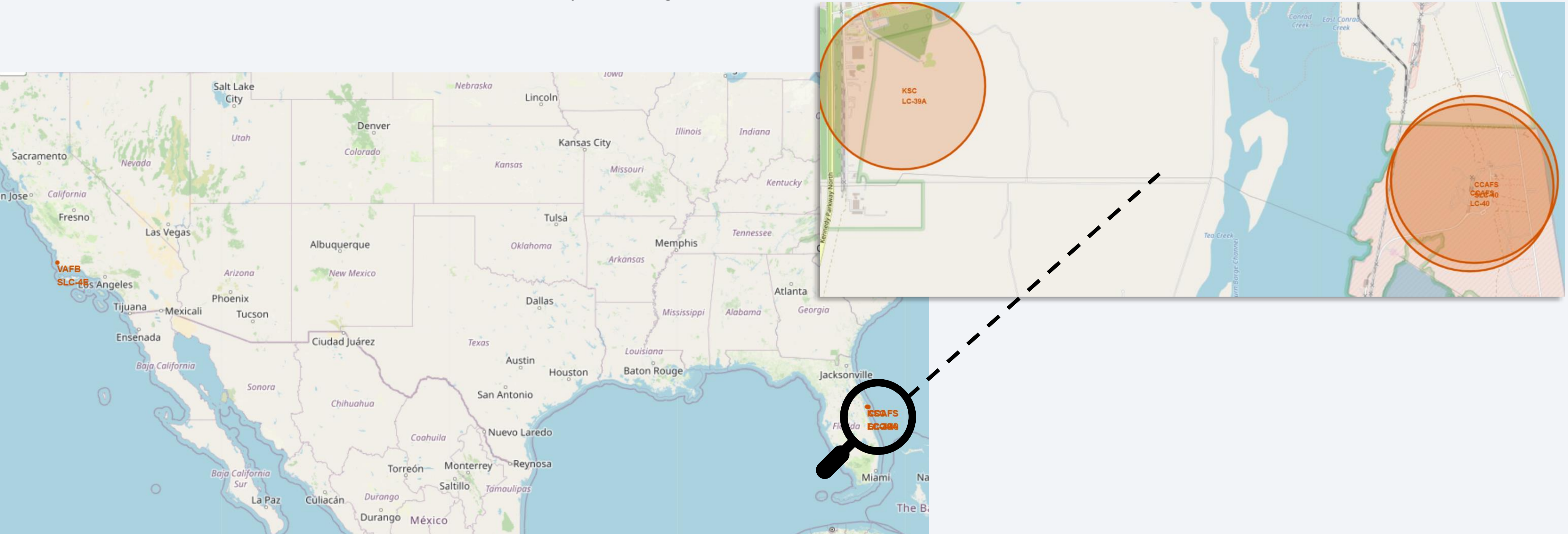
```
 * sqlite:///my_data1.db
Done.
```

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

Launch Sites
Proximities Analysis

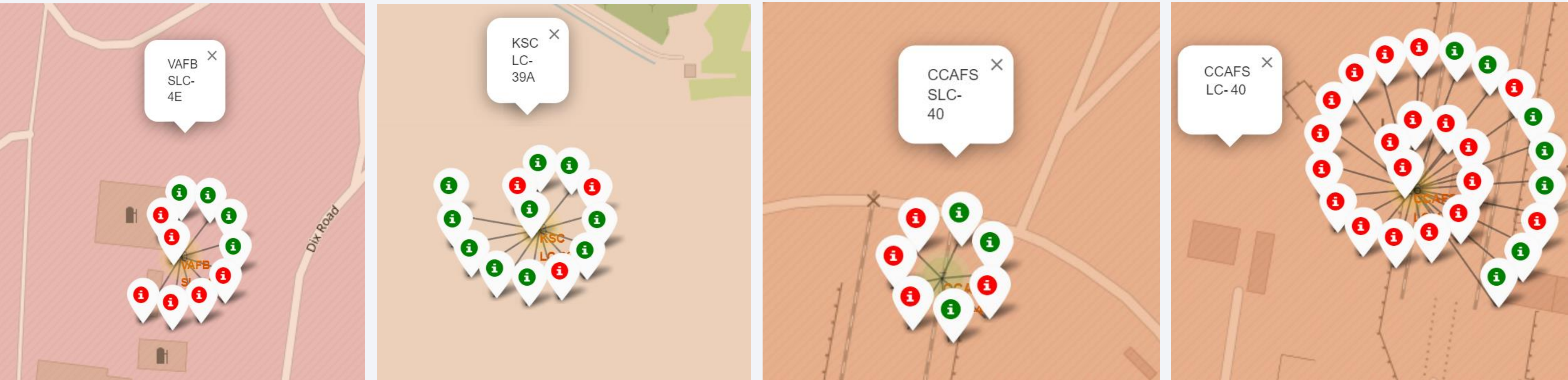
Launch Sites

- Proximity of the launch sites to the Equator provides several advantages for launches:
  - It is easier to launch into equatorial orbit by leveraging Earth's rotational speed.
  - The rotational speed also provides a natural boost to the rockets, reducing the need for additional fuel and boosters and thereby saving costs.



Launch Outcomes

- The colored markers represent **successful** and **unsuccessful** launches at each launch site and thus makes it easy to visualize the success rate for each site.
- We can tell that KSC LC-39A has the highest success rate due to the green markers greatly outnumbering the red markers at that site.



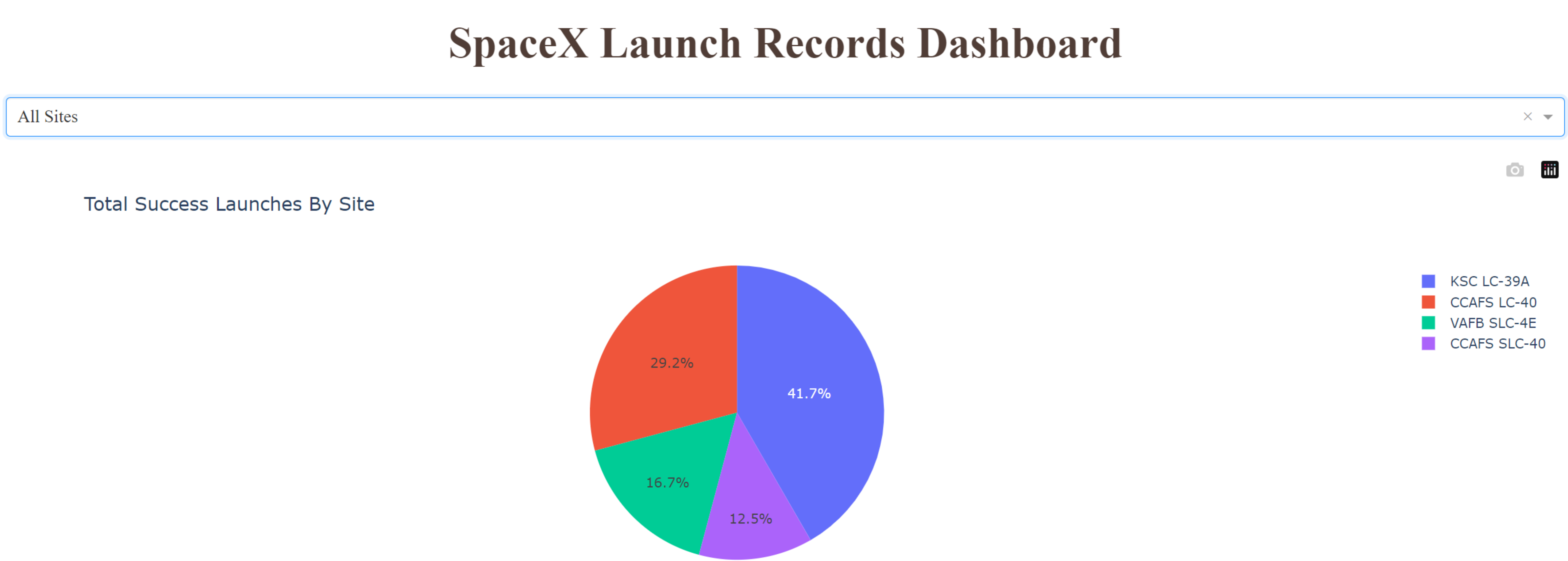
Launch Site Proximities

- Distance from CCAFS SLC-40 to :
  - The nearest coastline = 78.44 km
  - The nearest railway = 1.72 km
  - The nearest city = 0.86 km
  - The nearest highway = 1.07 km
- It is far from the city to minimize damage due to crashes.
- It is near the coast so that debris falls into the ocean and to easily recover the 1st stage from drone ships.
- It is near enough to highways and railways for easy transportation of materials and employees.


Build a Dashboard
with Plotly Dash

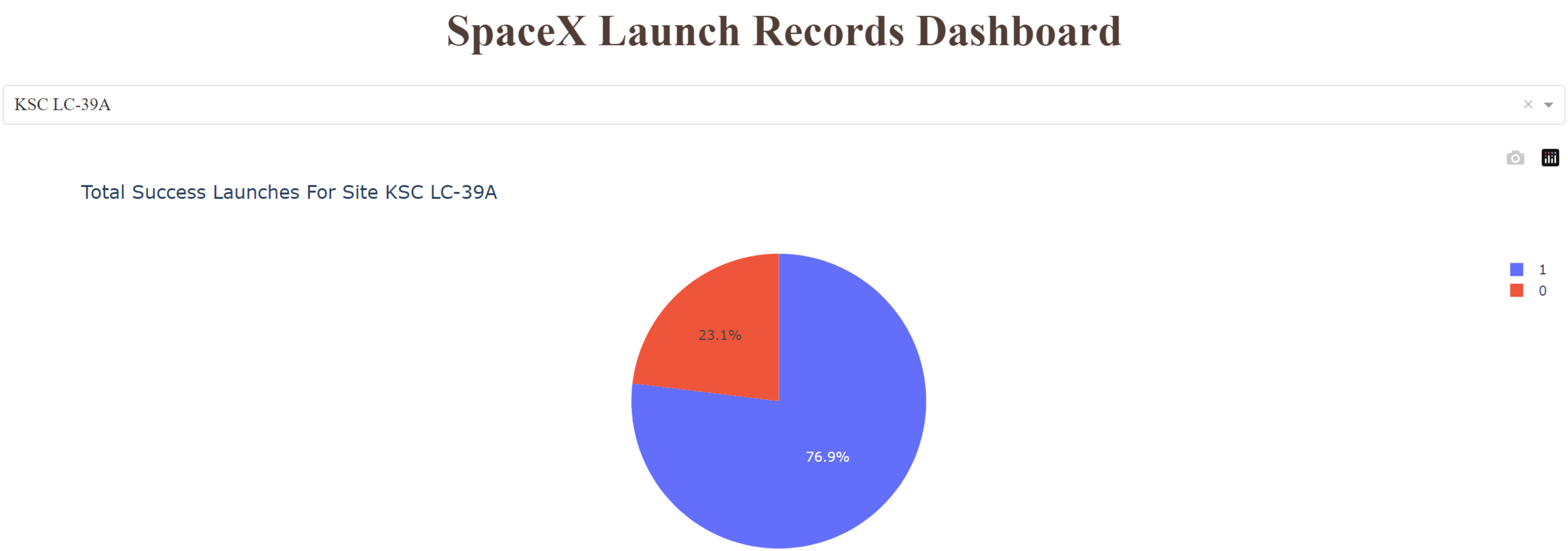
Launch Success Count for All Sites

- This pie chart visualizes the no. of successful launches at each site.
- We can clearly see that KSC LC-39A had the most successful launches out of all the sites.



Launch Site with Highest Launch Success Ratio

- This pie chart visualizes the no. of successful and failed launches for a launch site.
- We know that KSC LC-39A has the highest launch success ratio out of all the sites.



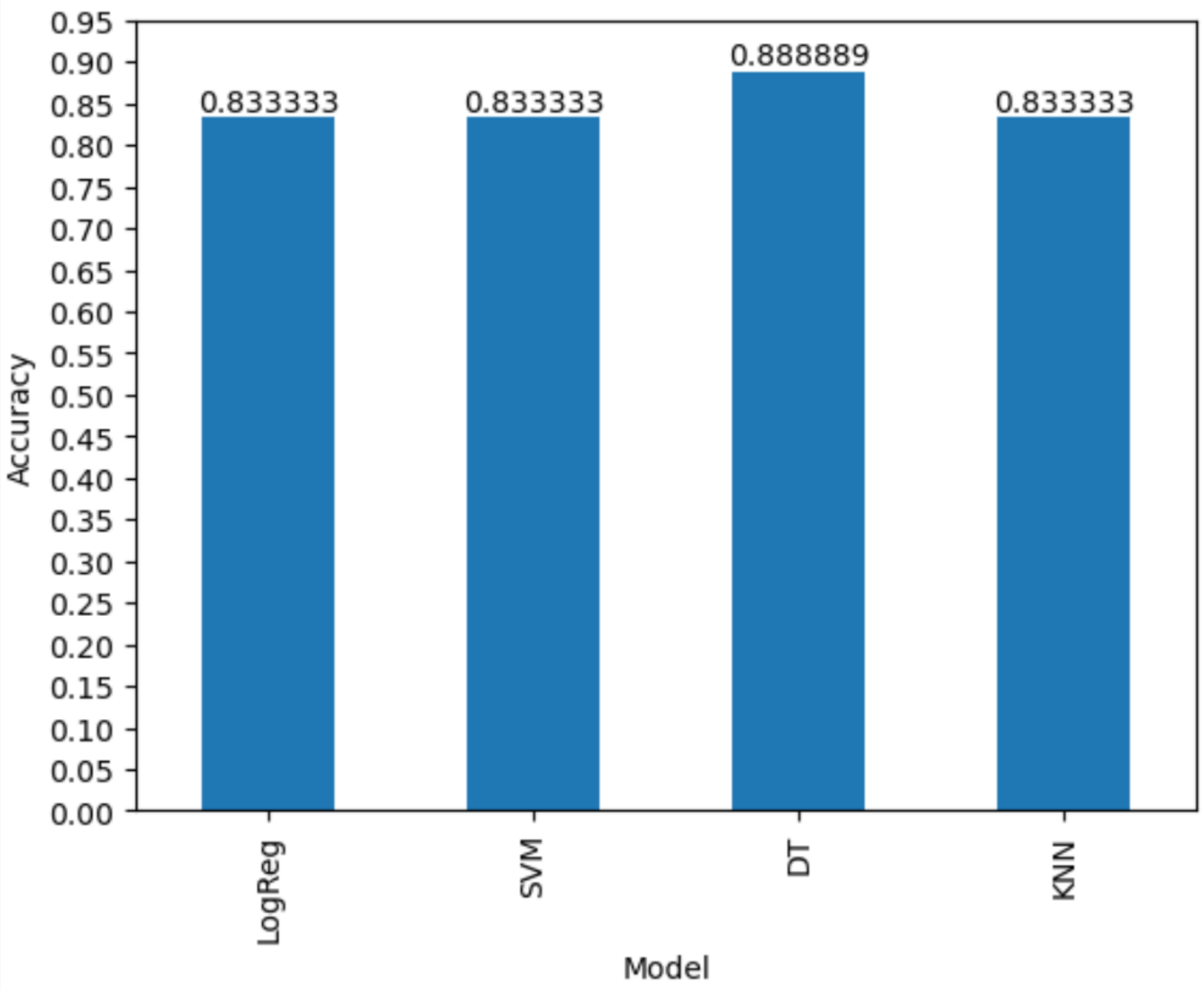
Payload vs. Launch Outcome Scatter Plot for All Sites

- This scatter plot visualizes the successful and failed launches at all the sites, their payload mass and their booster version.
- We can tell that the largest success rate is for payload masses between 2000 kg and 5000 kg, and the booster version *FT.*

Predictive Analysis (Classification)

Classification Accuracy

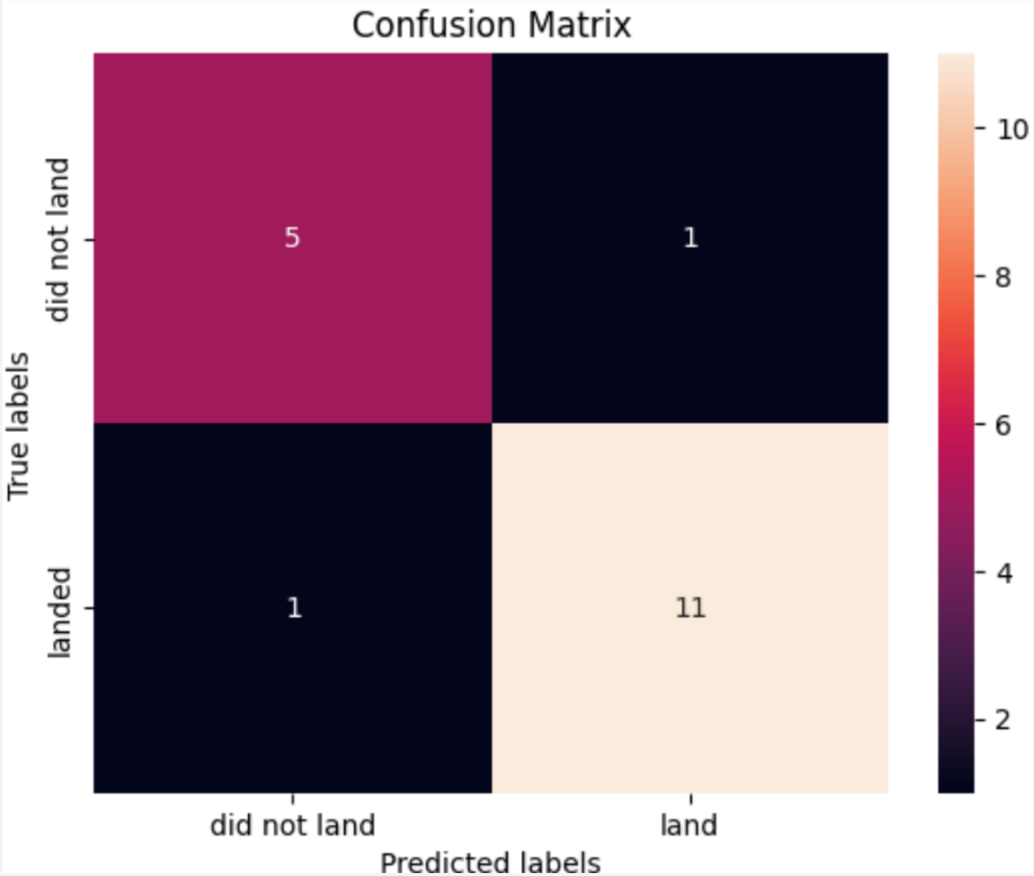
- This is a bar chart that shows the accuracy of each trained model.
- We can tell that the Decision Tree(DT) model had the highest classification accuracy of approximately 0.89, which very slightly outperforms the rest of the models.



Confusion Matrix

- This is the confusion matrix of the best performing model, the Decision Tree model.
- This model only has 1 False Positive and 1 False Negative, which is better compared to the confusion matrices of the other models that all have 3 False Positives.

Measure	Value
Sensitivity	0.9167
Specificity	0.8333
Precision	0.9167
Accuracy	0.8889
F1 Score	0.9167



Conclusions

- Most launch sites are located near the equator, near the coast, near highways and railways, but far from cities since it is a strategic position for rocket launching companies that has many advantages.
- The average launch success rate had rapidly increased over the years.
- As the payload mass (kg) increased, the launch success rate also increased.
- The launch site KSC LC-39A had the highest success rate among the launch sites we analyzed.
- The following orbits had a 100% success rate: ES-L1, GEO, HEO.
- The Decision Tree model only slightly outperformed the other classification models we tried.


SPACEX
Thank you!