IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Alex Maier
5/20/2025

Outline

Executive Summary

- Summary of methodologies
  - Data collection using the SpaceX API and webscraping
  - Exploratory Data Analysis (EDA) using data wrangling, data visualization, and interactive visual analytics
  - Machine Learning Prediction
- Summary of all results
  - EDA enabled us to identify which features are the best to predict successful launches
  - Machine Learning Prediction showed the best model to predict which characteristics are indicative of a successful launch, according to the data collected

Introduction

SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

Our Mission: use Falcon 9 rocket launch data to predict if the first stage will land successfully

Section 1

Methodology

Methodology

- Executive Summary
- Data collection methodology:
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

- GitHub Repository URL: https://github.com/alexgmaier11/Data-Science-Capstone
- Data sets came from two sources:
  - SpaceX API: https://api.spacexdata.com/v4/launches/past
  - Wikipedia: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- SpaceX API data was obtained using a GET request and Wikipedia data came from web scraping

Data Collection – SpaceX API

- SpaceX has a publicly available API that holds launch records for their rockets
- The below flowchart demonstrates the order of operations applied to the API data
- GitHub URL of SpaceX API calls lab for external reference and peer-review: https://github.com/alexgmaier11/Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

GET Request to SpaceX API → Parse data from GET Request → Filter data to only include Falcon 9 launches → Handle Missing Values

Data Collection - Scraping

- More SpaceX launch data is publicly available on Wikipedia
- The below flowchart demonstrates the order of operations applied to the Wikipedia data
- GitHub URL of Webscraping lab for external reference and peer-review: https://github.com/alexgmaier11/Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb

Extract Falcon 9 records as HTML table → Extract column and variable names from HTML header → Convert into a Pandas data frame

Data Wrangling

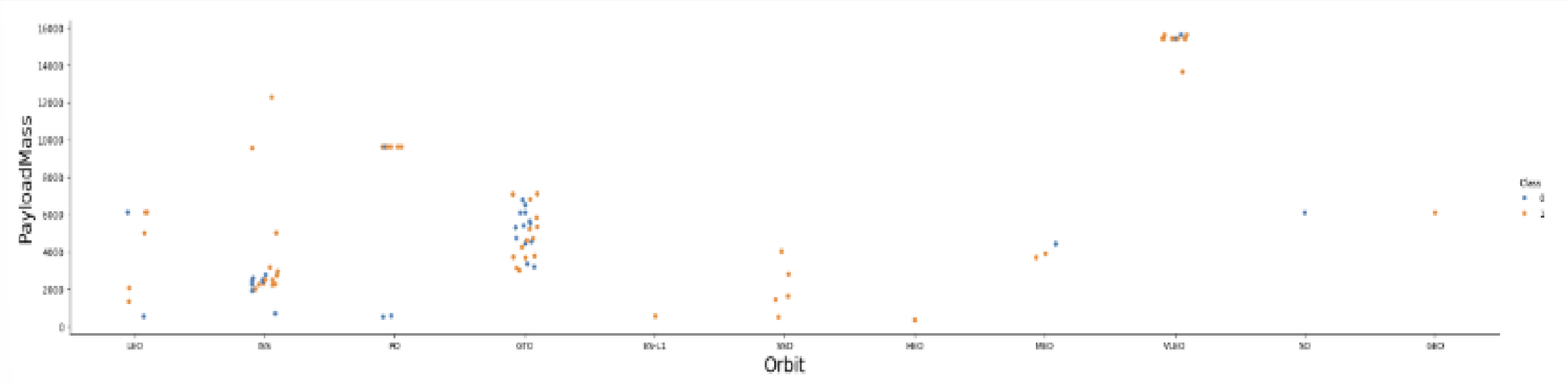
- Some Exploratory Data Analysis (EDA) was performed on the dataset
- The below flowchart shows the order of EDA performed on the dataset
- GitHub URL of Data Wrangling lab for external reference and peer-review: https://github.com/alexgmaier11/Data-Science-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

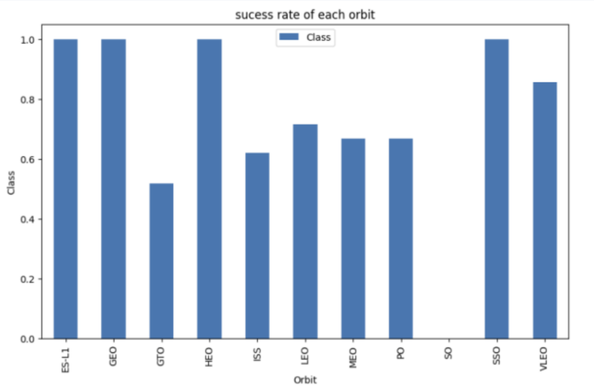
Calculate number of launches per site → Calculate number and occurrence of each orbit → Calculate number and occurrence of mission outcome per orbit → Create a landing outcome label

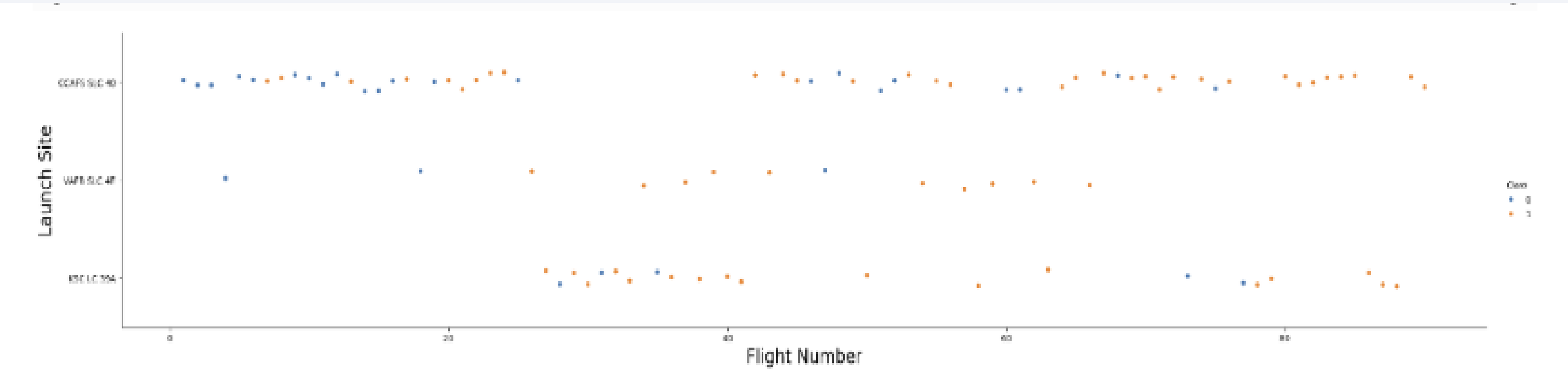
EDA with Data Visualization

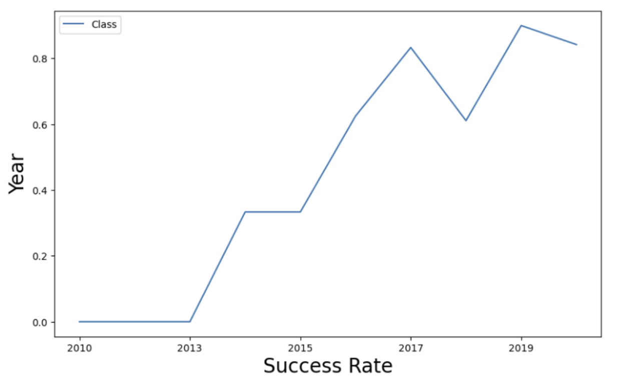
- Exploratory Data Analysis and Feature Engineering were performed using Matplotlib and Pandas in the below order:
  - Exploratory Data Analysis
  - Scatter Plot: Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Flight Number vs. Orbit Type, Payload Mass vs. Orbit Type
  - Bar Chart: Success Rate of Each Orbit Type
  - Line Chart: Yearly Success Trends
  - Feature Engineering
- GitHub URL of EDA with data visualization lab for external reference and peer-review: https://github.com/alexgmaier11/Data-Science-Capstone/blob/main/edadataviz.ipynb

EDA with Data Visualization Charts:









EDA with SQL

1. **Display the names of the unique launch sites in the space mission:**
   - `%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE`
2. **Display 5 records where launch sites begin with the string 'CCA'**
   - `%sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE "CCA%" LIMIT 5`
3. **Display the total payload mass carried by boosters launched by NASA (CRS)**
   - `%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE "Customer" LIKE "NASA (CRS)"`
4. **Display average payload mass carried by booster version F9 v1.1**
   - `%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE "Booster_Version" LIKE "%F9 v1.1%"`
5. **List the date when the first successful landing outcome in ground pad was acheived.**
   - `%sql SELECT MIN(Date) FROM SPACEXTABLE`

GitHub URL of EDA with SQL lab for external reference and peer-review: https://github.com/alexgmaier11/Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

EDA with SQL

6. **List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000**
   - %sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE ("PAYLOAD_MASS__KG_" < 6000 AND "PAYLOAD_MASS__KG_" > 4000 AND Landing_Outcome LIKE "%Success (drone ship)%")
7. **List the total number of successful and failure mission outcomes**
   - %sql SELECT Mission_Outcome, COUNT(*) FROM SPACEXTABLE GROUP BY Mission_Outcome
8. **List all the booster_versions that have carried the maximum payload mass. Use a subquery.**
   - %sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX (PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
9. **List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.**
   - %sql SELECT substr(Date, 6,2), Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE (substr(Date,0,5)='2015' AND Landing_Outcome LIKE "%Failure%")
10. **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.**
    - %sql SELECT Landing_Outcome, COUNT(*) FROM SPACEXTABLE WHERE (Date BETWEEN '2010-06-04' AND '2017-03-20') GROUP BY Landing_Outcome ORDER BY COUNT(*) DESC

GitHub URL of EDA with SQL lab for external reference and peer-review: https://github.com/alexgmaier11/Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

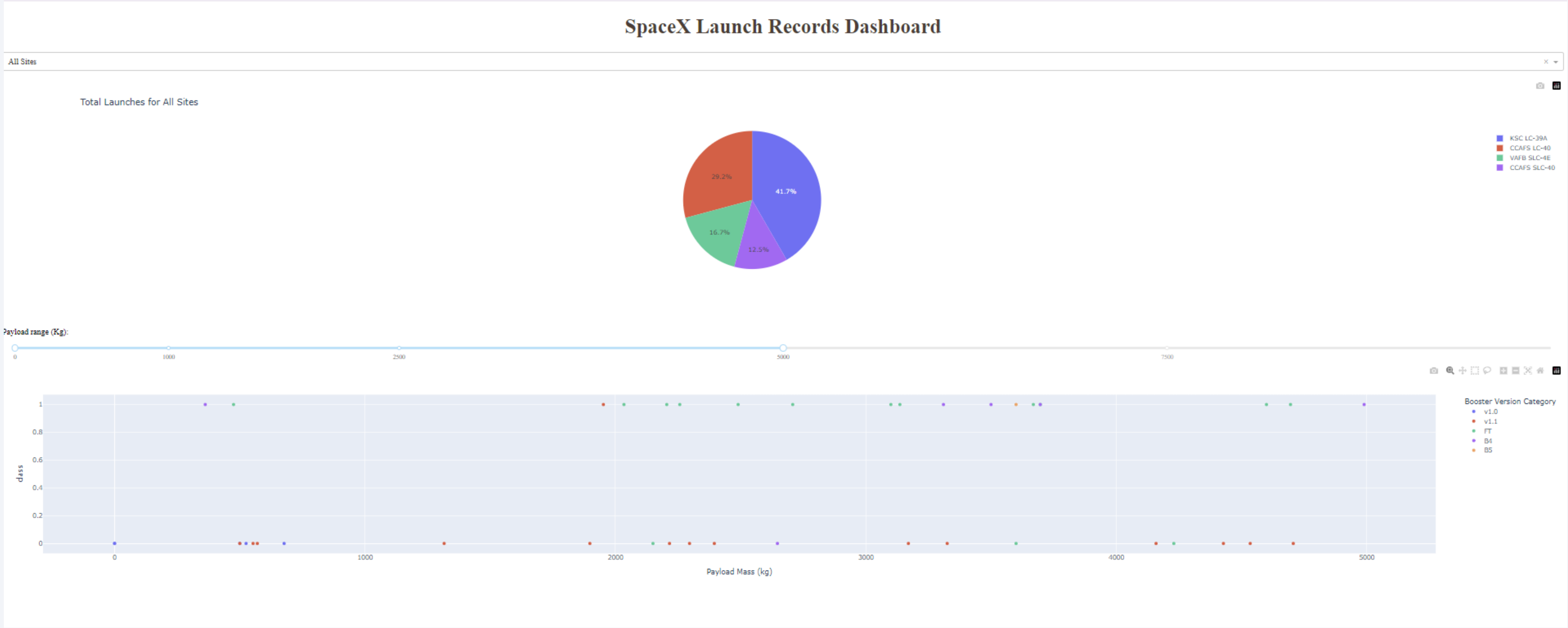
Build an Interactive Map with Folium

- Markers, circles, lines, and marker clusters were added to Folium Maps
  - Markers indicate points, like launch sites
  - Circles indicate highlighted areas around specific coordinates
  - Lines indicate distances between two coordinates
  - Marker clusters are used to indicate groups of events in each coordinate
- GitHub URL of Launch Site Locations lab for external reference and peer-review: https://github.com/alexgmaier11/Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- 🚀 Interactive dashboard application was built using Plotly dash:
  - 🚀 Adding a Launch Site Drop-Down Input Component
  - 🚀 Adding a callback function to render success-pie-chart based on site dropdown
  - 🚀 Adding a Range Slider to select Payload
  - 🚀 Adding a callback function to render success-payload-scatter-chart scatter plot
- 🚀 GitHub URL of Plotly Dash lab for external reference and peer-review: https://github.com/alexgmaier11/Data-Science-Capstone/blob/main/spacex-dash-app.py

SpaceX Launch Records Dashboard:

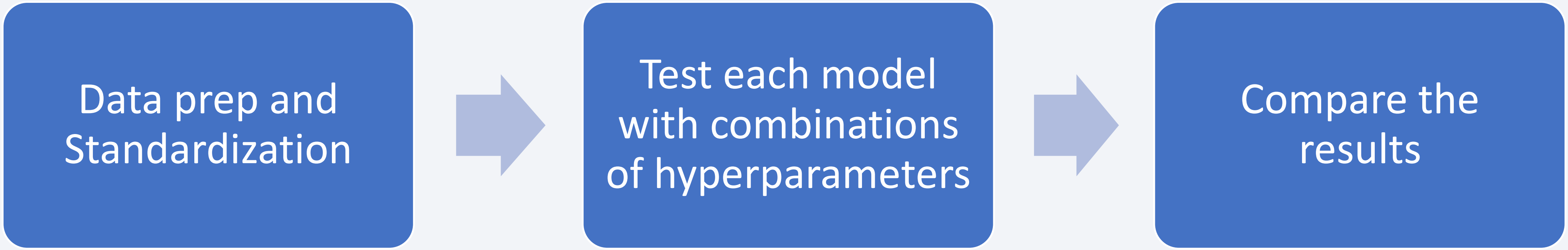


Predictive Analysis (Classification)

- After data is loaded into a Pandas Dataframe, perform Exploratory Data Analysis and use the following to determine Training Labels:
  - Create a NumPy array from the column Class by applying the method to numpy(), then assign it to the variable Y as the outcome variable
  - Next, the StandardScaler() function from sklearn is used to standardize the feature dataset
  - Finally, the data is split into training and testing sets via the function train_test_split from sklearn_model selection, using test size parameter set to 0.2 and random state set to 2
- GitHub URL for the predictive analysis lab for external reference and peer-review: https://github.com/alexgmaier11/Data-Science-Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Predictive Analysis (Classification)

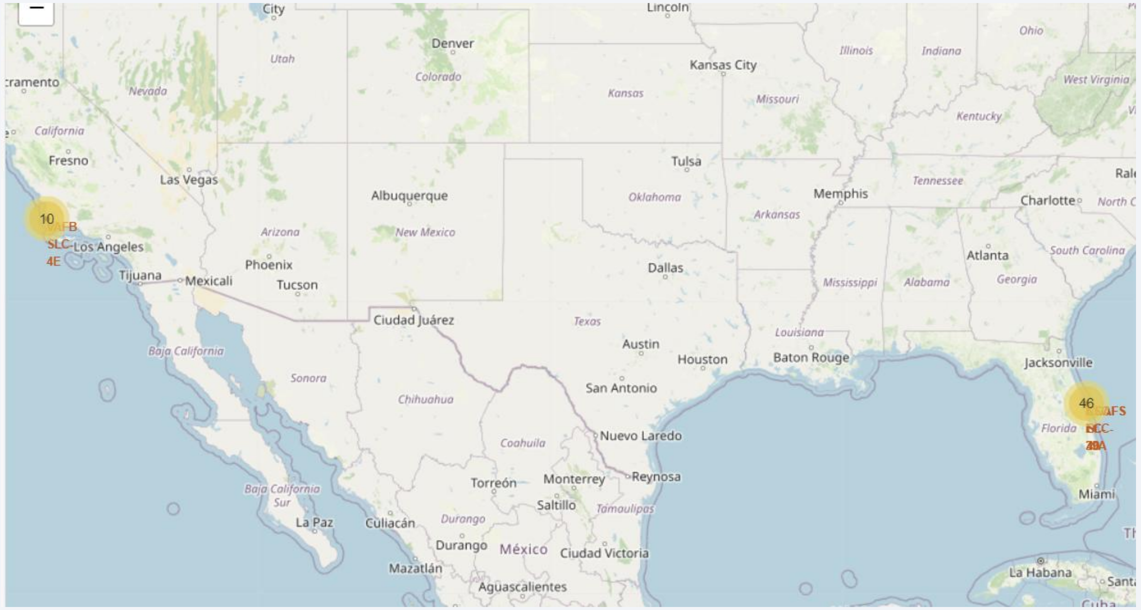
Our goal: to find the best Machine Learning model based on the test data between Logistic Regression, SVM, Classification Tree, and K-Nearest Neighbors:



Interactive Analytics Results

Observations from Folium Maps:

- All of the launch sites are on the coast, either in California or Florida, and are close enough to population centers that workers can commute every day, but are far enough away from those population centers to avoid damage in case of a failed launch



Predictive Analysis Results

- The table on the right shows the accuracy score if test data for each method, comparing them to show which method performed the best:
- Methods being compared are:
  - Logistic Regression
  - SVM
  - Decision Tree
  - K-Nearest Neighbor

Method	Logistic_Reg	SVM	Decision Tree	KNN
Test Data Accuracy	0.833333	0.833333	0.888889	0.833333

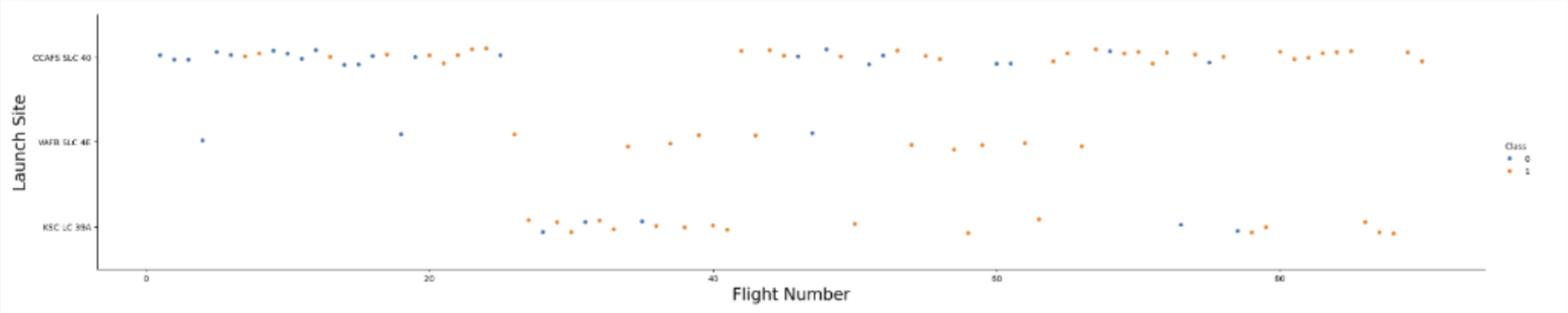
- Decision Tree method was the most successful of the four methods, with an accuracy of 0.88

Section 2

Insights drawn from EDA

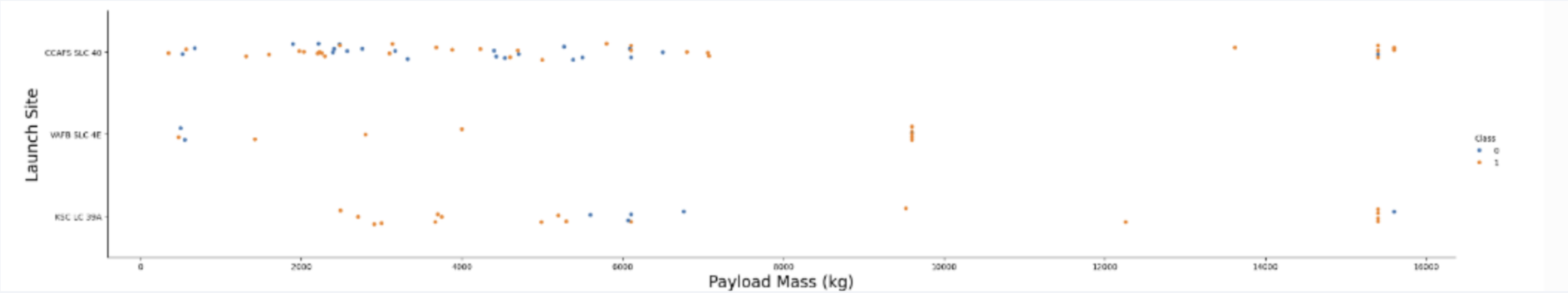
Flight Number vs. Launch Site

- According to the below plot, we can very that the current best launch site is CCAFS SLC-40, since most of its recent launches have been successful
- Second most recent successful launches is VAFB SLC-4E, and third is KSC LC-39A
- We can see an increase in the success rate over time



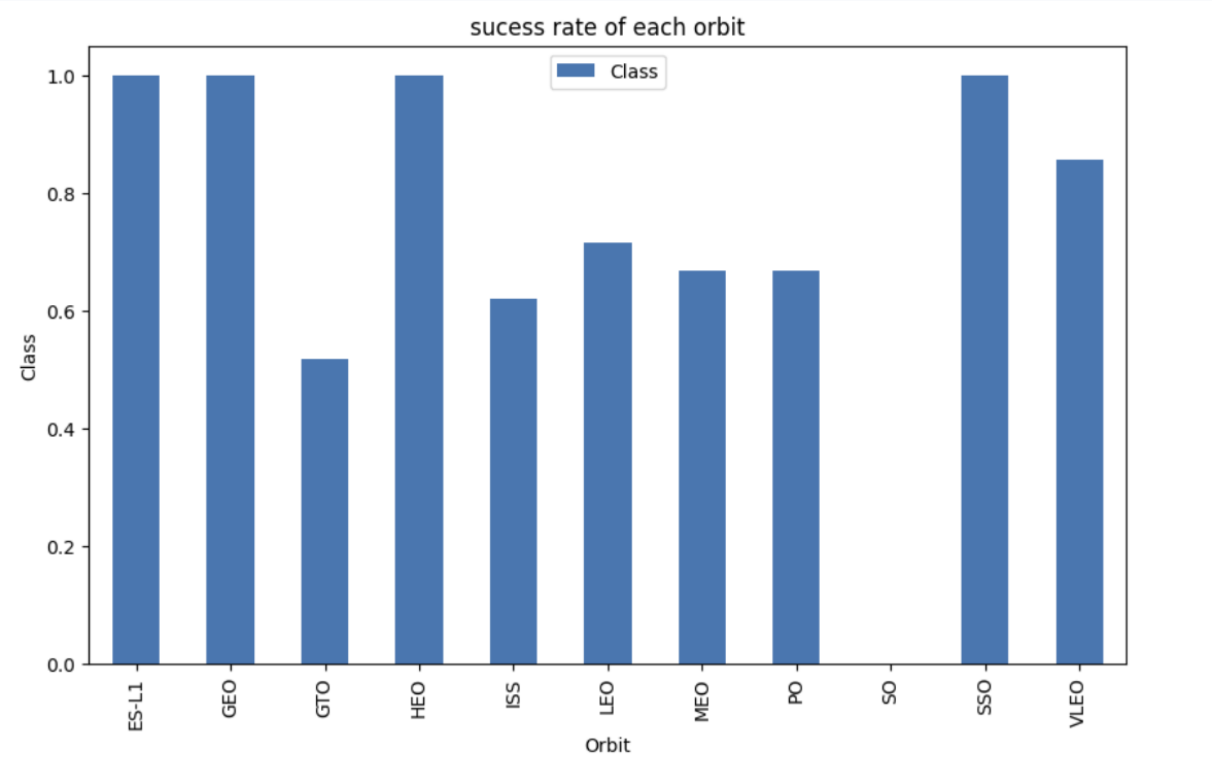
Payload vs. Launch Site

- Heavy payloads seem to only be possible at CCAFS SLC-40 and KSC LC-39A. The heaviest payload recorded at VAFB SLC-4E is around 9,500 kg, while the other two top out around 15,8000 kg.
- Payloads over 9,000 kg have an incredible success rate



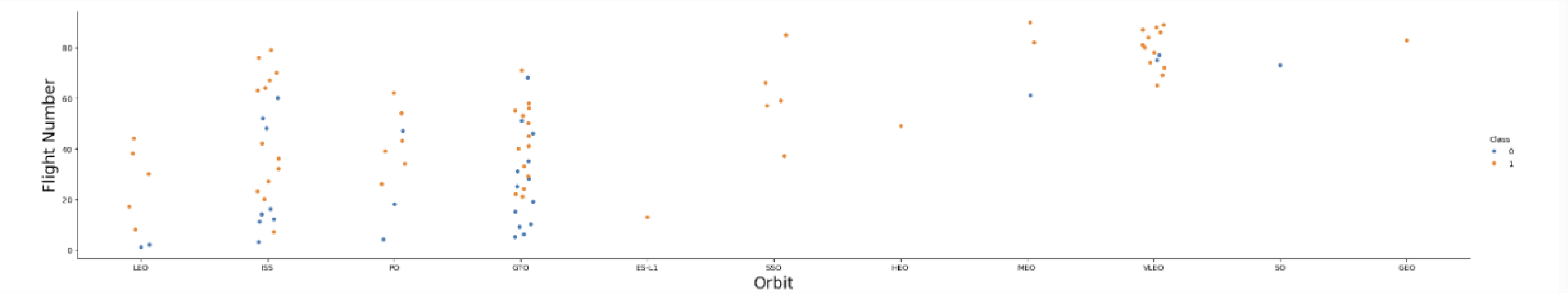
Success Rate vs. Orbit Type

- The orbits with a success rate of 100%:
  - ES-L1
  - GEO
  - HEO
  - SSO
- Followed by:
  - VLEO (>80%)
  - LFO (>70%)



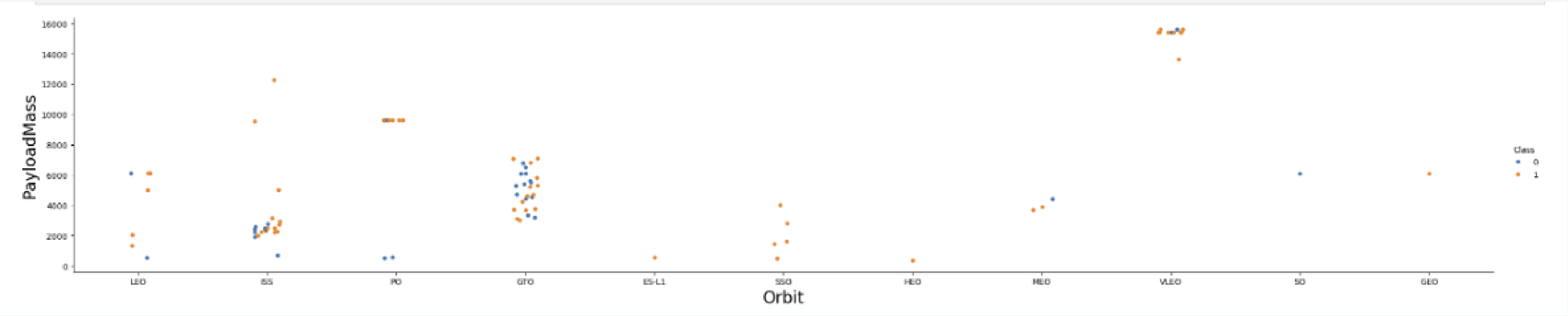
Flight Number vs. Orbit Type

- It seems that success rates improved over time for all orbits
- VLEO orbit seems to be a newer venture, due to the recent increase of its frequency



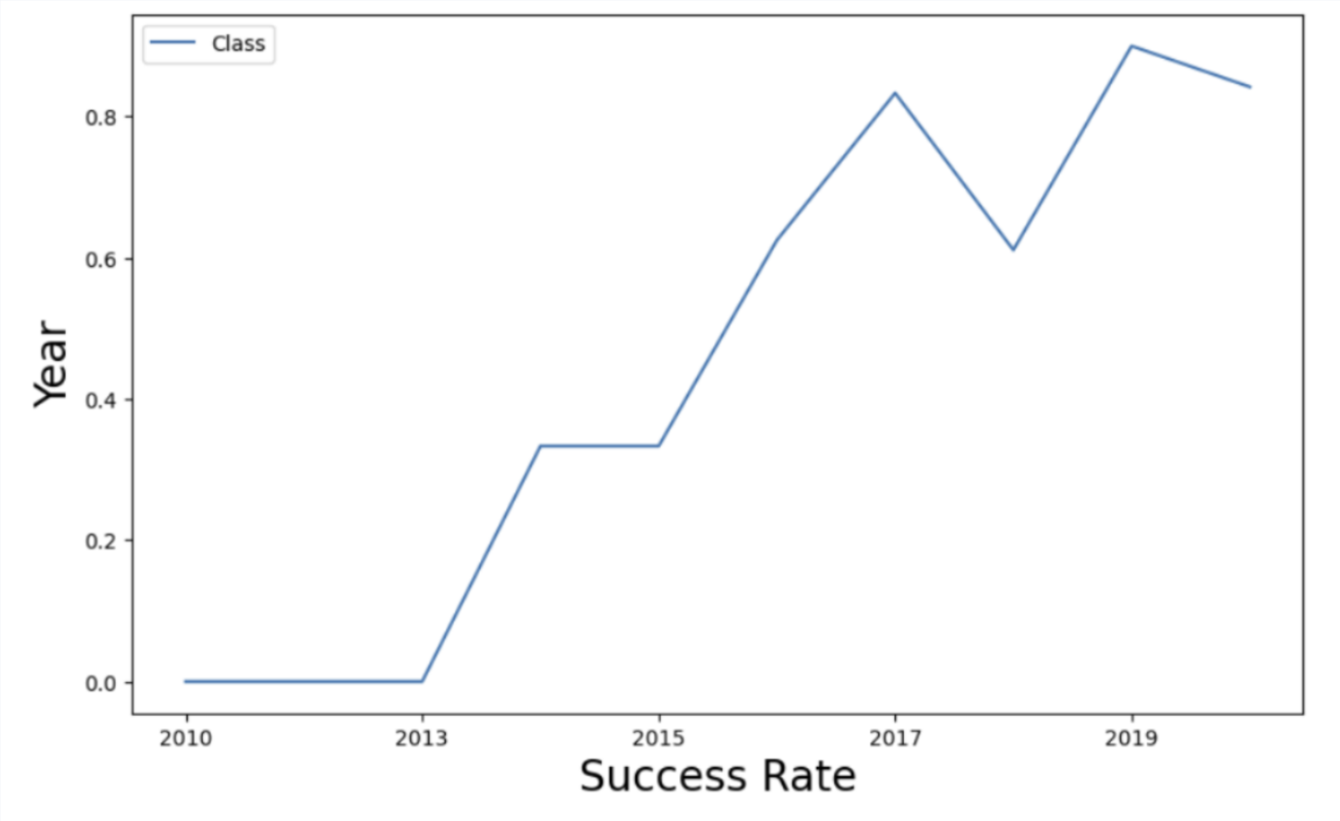
Payload vs. Orbit Type

- There is no direct relationship between payload and success rate for orbit GTO
- The ISS orbit has a large range of payload and a good success rate
- Orbits SO and GEO only have one launch recorded for each



Launch Success Yearly Trend

- The success rate began to increase in 2013 and mostly climbed until 2020
- The first three years, 2010-2013, show a success rate of 0% until the climb began in 2013



All Launch Site Names

- Find the names of the unique launch sites
- Using DISTINCT, we get all the unique values in "Launch_Site"

```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE
```

```
* 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'

- Find 5 records where launch sites begin with \`CCA\`
- We can compare a text string to cell value using WHERE “Column_Name” LIKE “text%”

```
%sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE "CCA%" LIMIT 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- Using the SUM(Column_Name) query, we can add up the contents of a column

```
%sql SELECT SUM(PAYLOAD_MASS__KG
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Using the AVG(Column_Name) query, we can get the mean value of a column

```
%sql SELECT AVG(PAYLOAD_MASS__KG
```

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

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Using the MIN(Column_Name) query, we get the minimum value of that column, which in this case returns the minimum date

```
%sql SELECT MIN(Date) FROM SPACEXTABLE
```

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

MIN(Date)
2010-06-04

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
- Using the Greater Than (>) and Less Than (<) operators, we can compare column contents to values to rule out outliers

```
%sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE ("PAYLOAD_MASS__KG_" < 6000 AND "PAYLOAD_MASS__KG_" > 4000 AND L
```

```
 * 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

- Calculate the total number of successful and failure mission outcomes
- Using the GROUP BY Column_Name query, we can group each record by their value in the specified column

```
%sql SELECT Mission_Outcome, COUNT(*) FROM SPACEXTABLE GROUP BY Mission_Outcome
```

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

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Using the MAX query, we get the maximum value in that column
- This lets us use MAX in the subquery to define the maximum payload

```
%sql SELECT DISTINCT Booster_Versi
```

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

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

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Using the substr(Column_Name, x, y) query, we get a substring from the column given

```
%sql SELECT substr(Date, 6,2), Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE (substr(Date,0,5)='2015'
```

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

substr(Date, 6,2)	Landing_Outcome	Booster_Version	Launch_Site
01	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

- 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
- Using the ORDER BY Column_Name query, we can sort by column values

```
%sql SELECT Landing_Outcome, COUNT(*) FROM
```

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

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

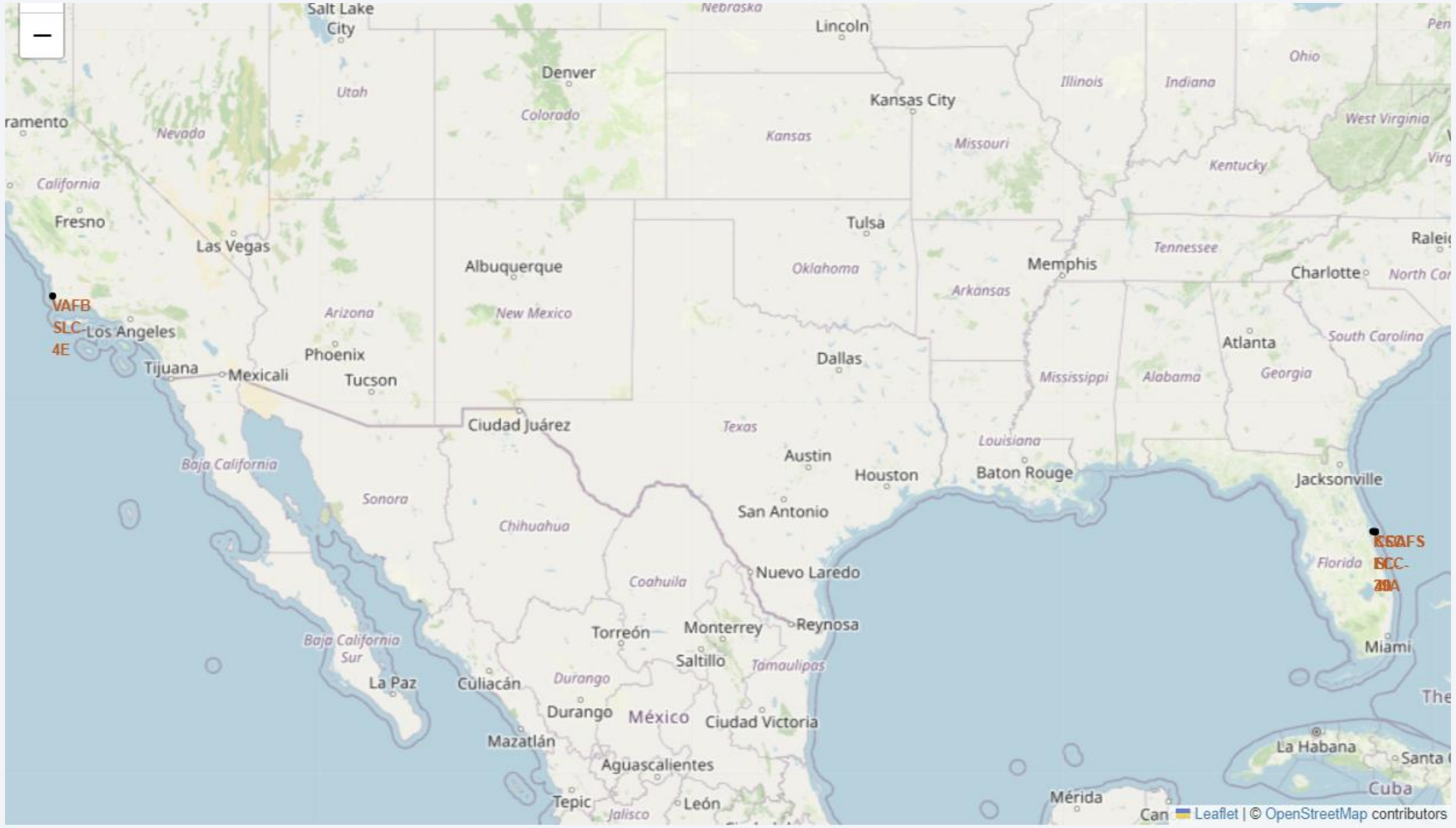
Section 3

Launch Sites Proximities Analysis

SpaceX Launch Sites Map

Four SpaceX Launch Sites:

- VAFB SLC-4E:
  - Vandenberg Space Launch Complex 4 (CA)
- KSC-LC29A:
  - Kennedy Space Center – Merritt Island (FL)
- CCAFS-LC40:
  - Cape Canaveral Launch Complex 40 (FL)
- CCAFS-SLC40:
  - Cape Canaveral Space Launch Complex 40 (FL)



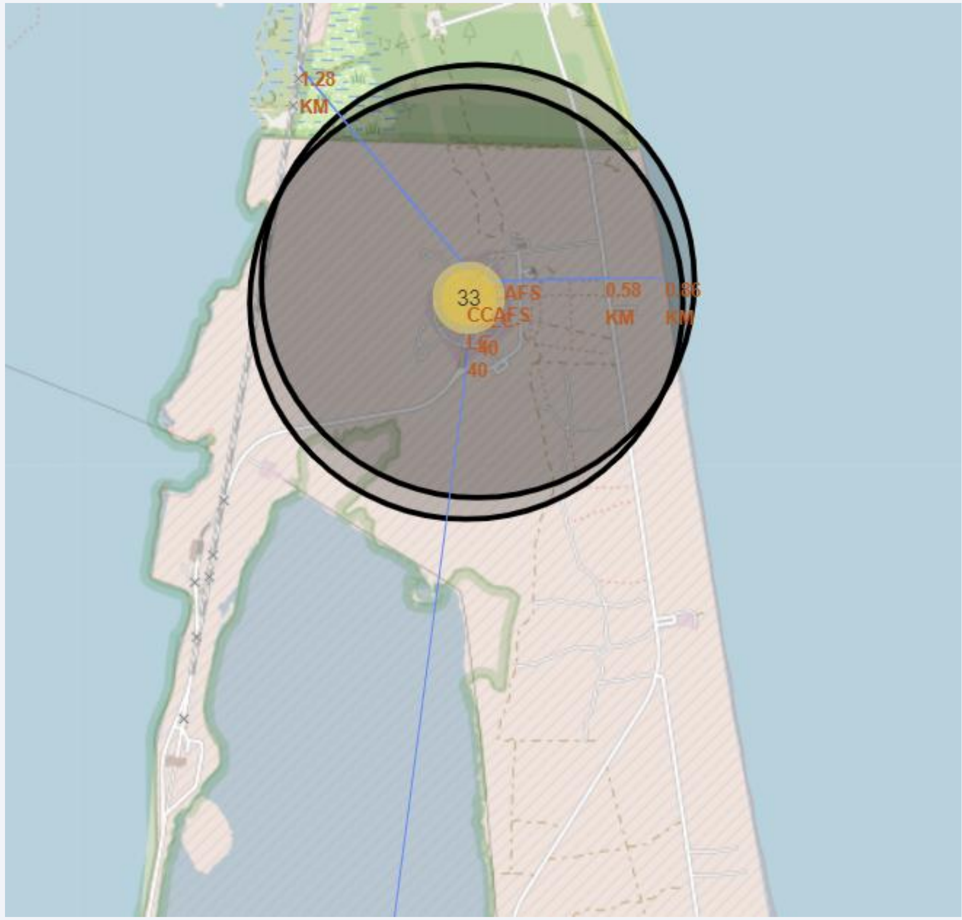
VAFB SLC-4E Site Map

Example of VAFB SLC-4E:

- Red Pins indicate a failed launch
- Green Pins indicate a successful launch
- Vandenberg AFB had 10 total launch attempts:
  - 4 Successful
  - 6 Unsuccessful

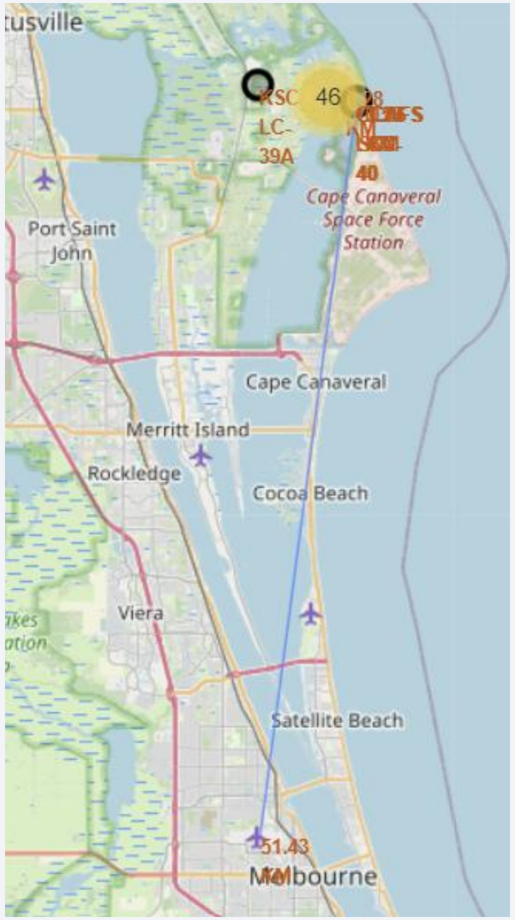
Cape Canaveral AFB SLC-40

- 🚀CCAFB is 0.58 kilometers from the nearest highway
- 🚀CCAFB is 0.86 kilometers from the nearest coastline
- 🚀CCAFB is 1.28 kilometers from the nearest railroad line



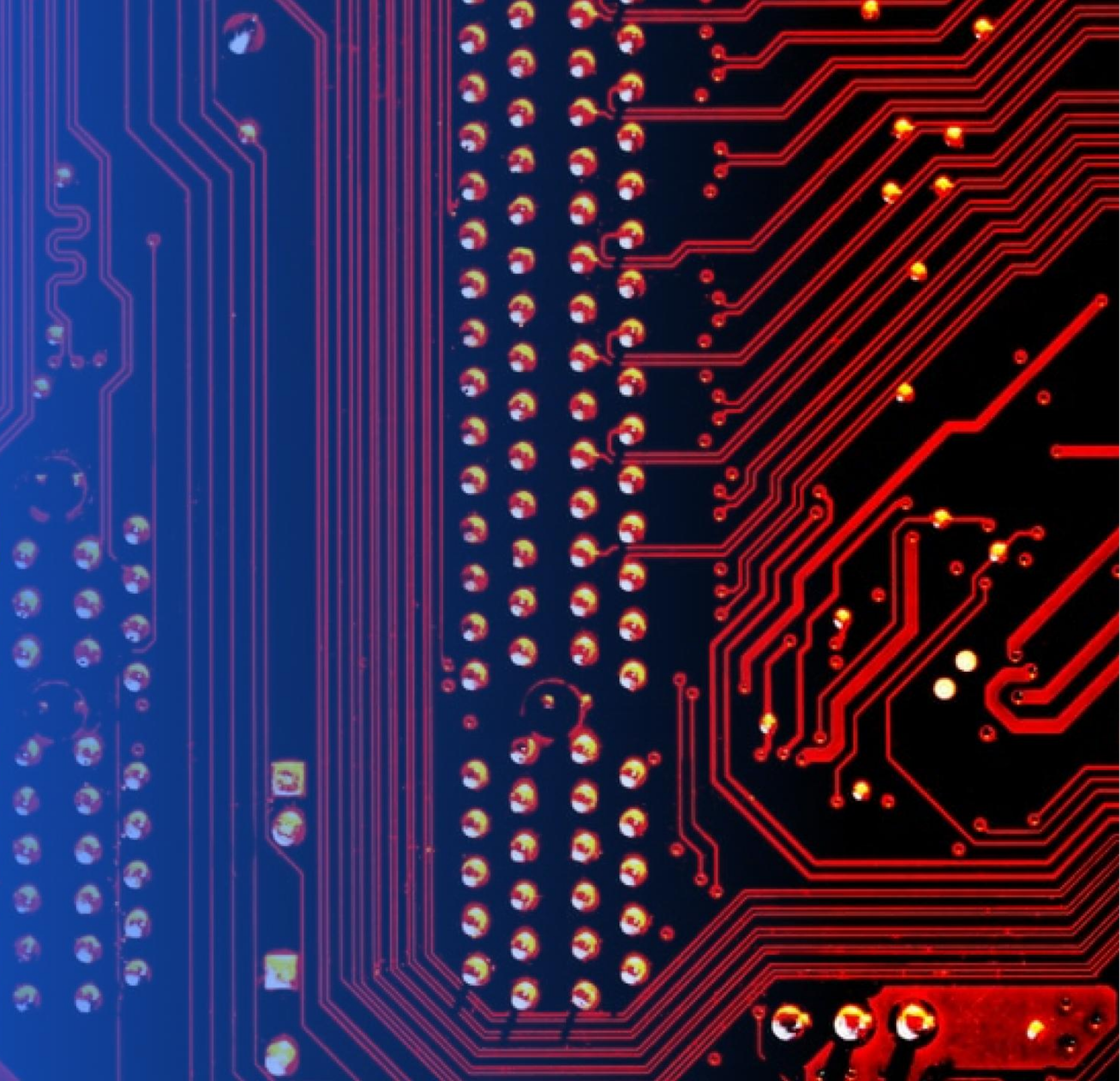
Cape Canaveral AFB SLC-40

- CCAFB is approximately 51.43 kilometers from the nearest city of Melbourne, FL



Section 4

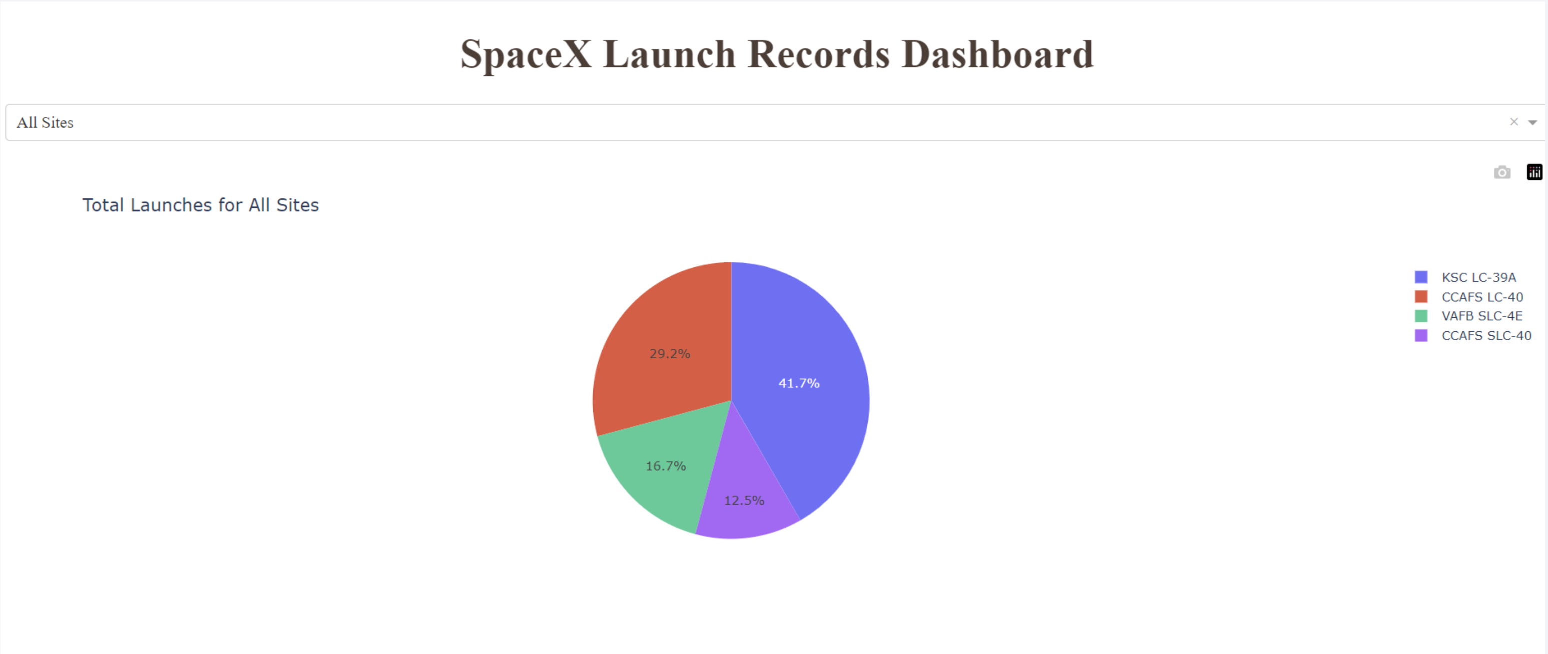
Build a Dashboard with Plotly Dash



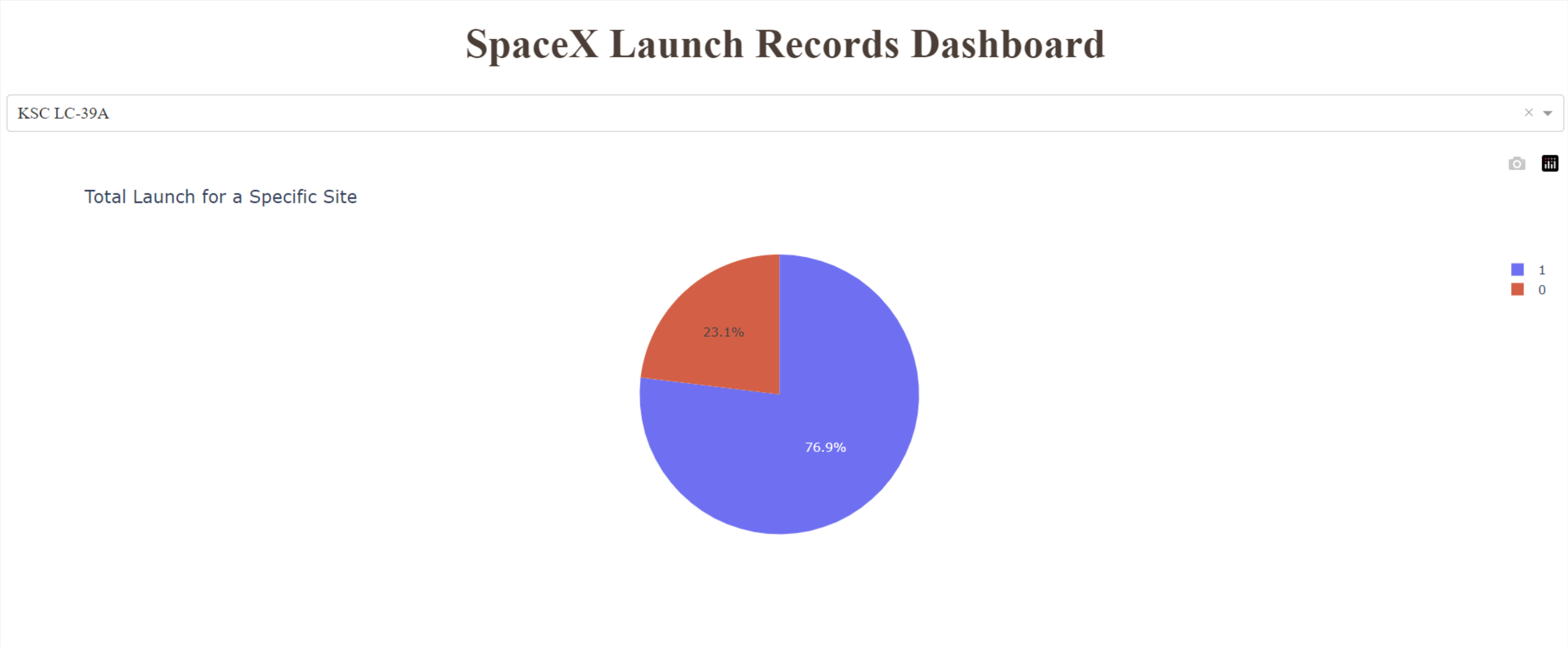
All Sites Launch Success Count

This pie chart shows the launch success rate for each of the four sites in the dataset:

1. KSC LC-39A came in first for number of total successful launches with 41.7% of total successful launches.
2. CCAFS LC-40 came in second with 29.2% of total successful launches.
3. VAFB SLC-4E came in third, with 16.7% of total successful launches.
4. CCAFS SLC-40 came in fourth, with 12.5% of total successful launches.

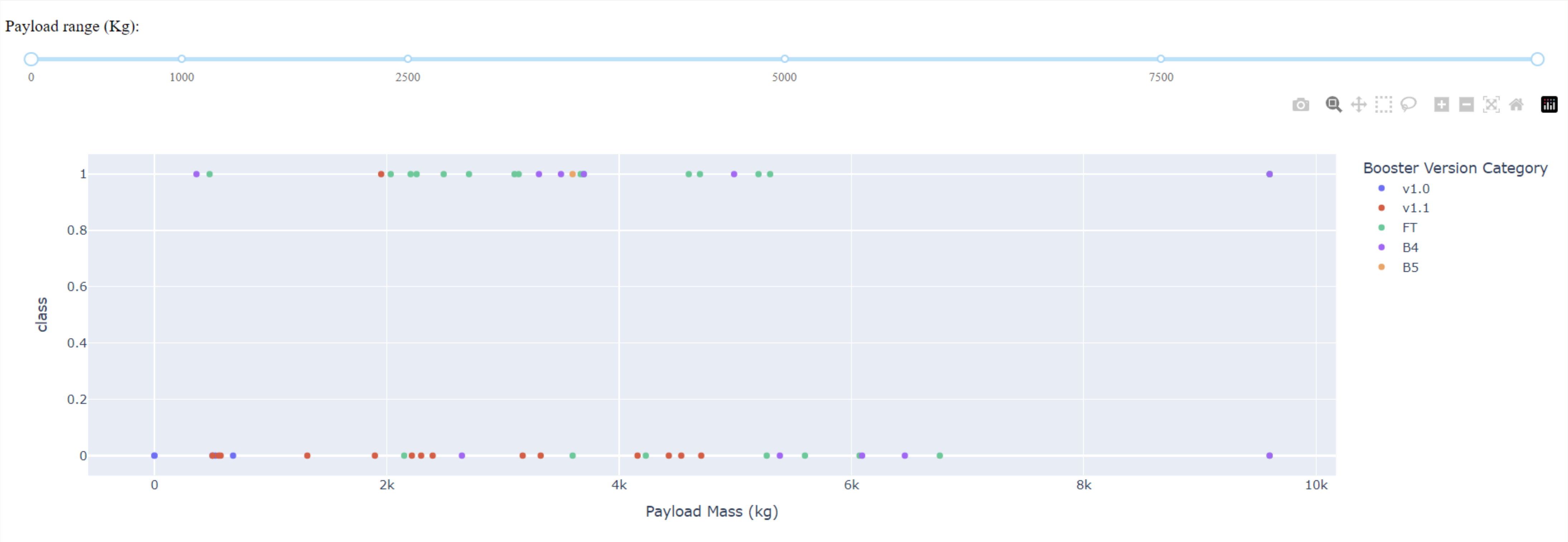


Launch Success Ratio for KSC LC-39A



76.9% of launches at KSC LC-39A are successful, which is the highest ratio for any launch site

Payload vs. Launch Outcome Scatter Plot



The most successful combination includes payloads under 6,000 kg and FT boosters

Section 5

Predictive Analysis (Classification)

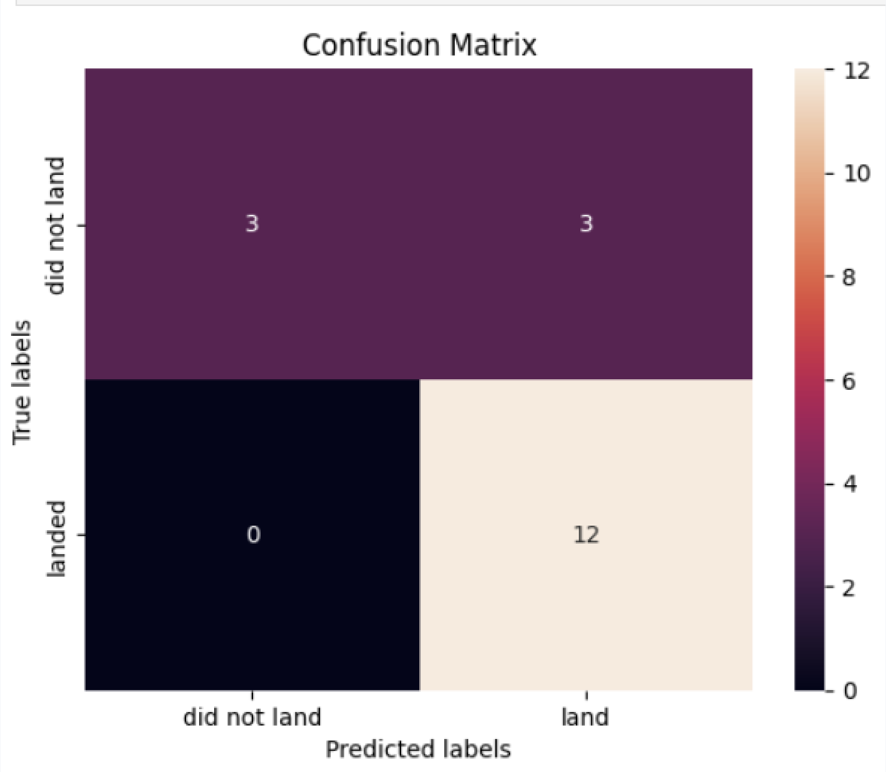
Classification Accuracy

- The table below shows each classification method and its test data accuracy
- Logistic Regression, SVM, and K-Nearest Neighbor each showed 0.83, while Decision Tree showed 0.88 accuracy

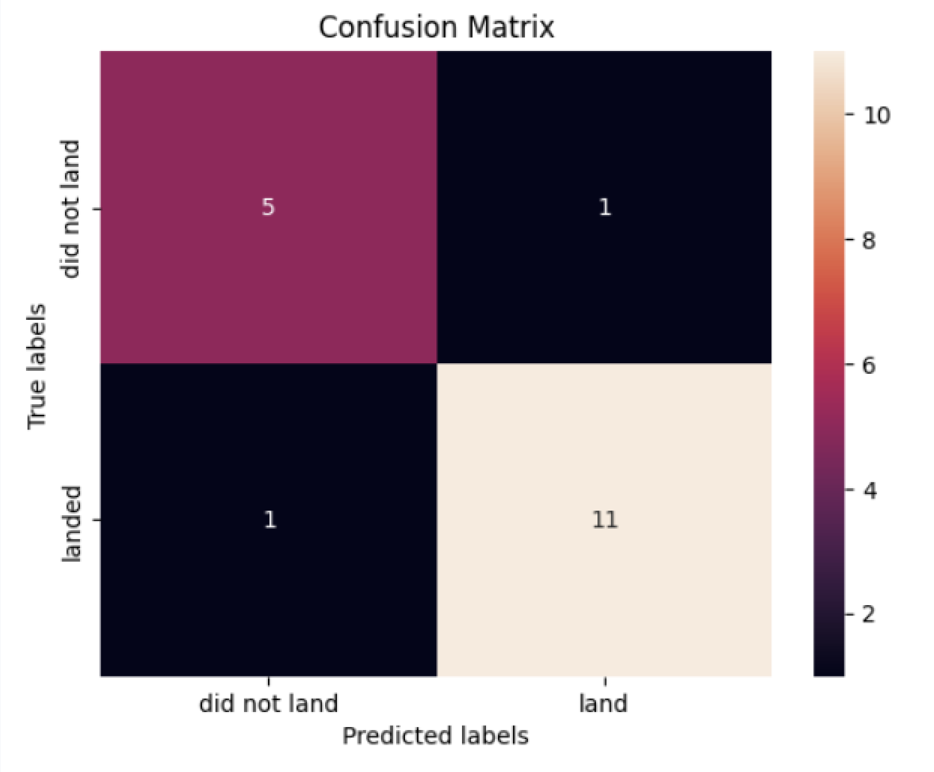
Method	Logistic_Reg	SVM	Decision Tree	KNN
Test Data Accuracy	0.833333	0.833333	0.888889	0.833333

Confusion Matrix

Confusion Matrix of Logistic Regression, K-Nearest Neighbor, and SAM



Confusion Matrix of Decision Tree



This shows that the Decision Tree Method is more accurate. It not only can predict the successes accurately, but can predict the failures with accuracy.

Conclusions

- 🚀We analyzed different data sources and refined conclusions throughout the process
- 🚀The most successful launch site is KSC LC-39A
- 🚀Launches above 7,000 kg have less risk attached
- 🚀As time passed, successful landings seemed to improve over time. This suggests evolution in the development and deployment of SpaceX rockets
- 🚀Decision Tree Classifiers were the most accurate prediction tools for successful landings and failed landings

Thank you!