IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Brittany Williams
11/20/2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- In this report, we will be analyzing a database of past SpaceX launches to determine the likelihood of success of future launches.
- The methods used include data wrangling, exploratory data analysis, interactive data visualization, and predictive analysis with multiple machine learning models.
- Results showed that launch site, orbit, payload mass, and booster version were all predictors of a launch success.

Introduction

- The goal of this project is to predict if the first stage of a Falcon 9 rocket will land successfully.
- If the rocket lands successfully and can be reused, it will greatly reduce the overall cost of the launch.

Section 1

Methodology

Methodology

Executive Summary

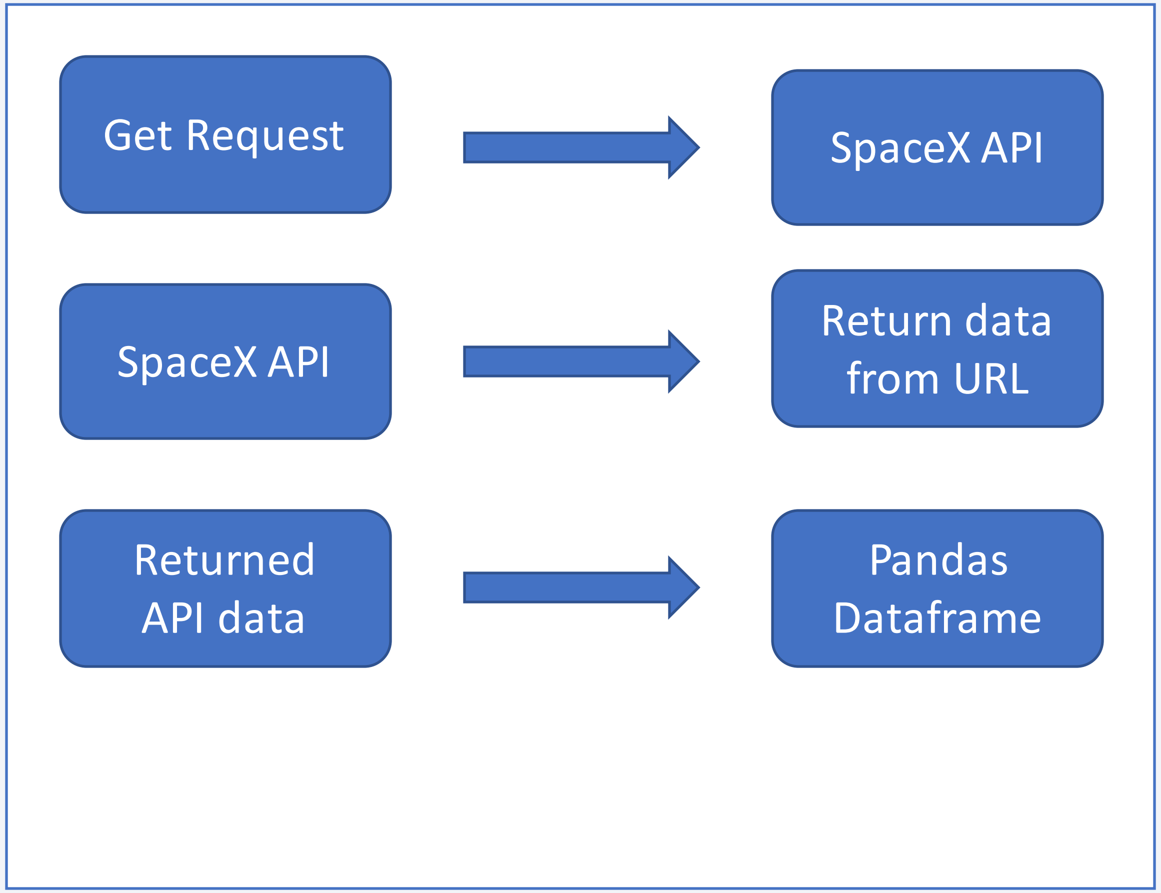
- Data collection methodology:
  - Data was retrieved via SpaceX API and by webscraping
- Perform data wrangling
  - Data was cleaned to include only relevant information
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Machine learning models tested include Logistic Regression, K Nearest Neighbors, Decision Tree, and SVM

Data Collection

- Data was collected by using a get request to the SpaceX API
- Data was also acquired from Wikipedia via webscraping

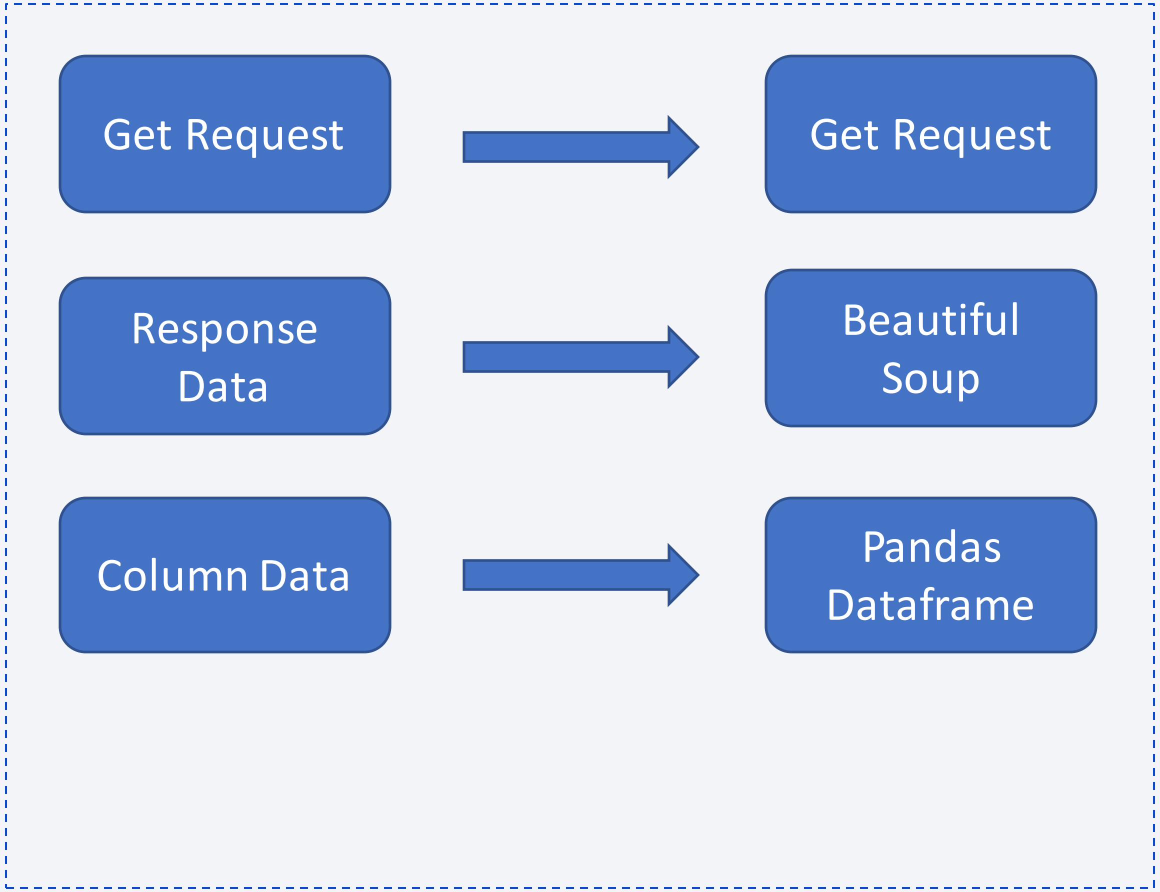
Data Collection – SpaceX API

- Several functions were defined to get data from the API and append it to a new dataframe.
- Get request was sent to the SpaceX API to retrieve data. It was then made into a pandas dataframe.
- https://github.com/baw08d/IBM-Data-Science-Capstone/blob/fc1bbbefc246db9dcf4ee7a896e06fb81c3aeb9a/jupyter-labs-spacex-data-collection-api%20(1).ipynb



Data Collection - Scraping

- A get response was sent to the Wikipedia URL to obtain data
- Response content was converted to a Beautiful Soup object
- The correct table was identified, and the columns were extracted to a dataframe
- https://github.com/baw08d/IBM-Data-Science-Capstone/blob/fc1bbbefc246db9dcf4ee7a896e06fb81c3aeb9a/jupyter-labs-webscraping.ipynb



Data Wrangling

- Number of launches per site was calculated with the value_counts method
- Number of launches to each orbit was found with the value_counts method
- A new dataframe of outcomes was created based on the count of each result
- A new column "Class" was added and any launches with bad outcomes were set to 0, and all else set to 1
- The average success rate was determined to be 66%
- https://github.com/baw08d/IBM-Data-Science-Capstone/blob/fc1bbbefc246db9dcf4ee7a896e06fb81c3aeb9a/labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

- A scatter plot of Flight Number vs. Payload Mass, colored by Class, was created to observe the change in success rate over time
- A scatter plot of Launch Site vs. Flight Number, colored by Class, showed which sites had the most successful launches
- A bar chart of class and orbit showed which orbits had the highest success rate
- A scatter plot of Flight Number vs. Orbit, colored by Class, showed that orbits changed over time and influenced the success of the launch

EDA with Data Visualization

- A scatterplot of Payload vs. Orbit, colored by Class, showed which orbit types worked best for heavy payloads
- A line chart of Date and Class is evidence that launch success has increased over time
- https://github.com/baw08d/IBM-Data-Science-Capstone/blob/fc1bbbefc246db9dcf4ee7a896e06fb81c3aeb9a/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- Select Distinct Launch Sites
- Select 5 Launch Sites beginning with CCA
- Select total Payload Mass
- Select Date of first successful landing
- Select Booster Versions with successful drone ship landings and Payload Mass between 4000 and 6000kg
- Select total number of successful and failed missions
- Select which Booster Versions have carried maximum Payload
- Select month of failed drone ship missions in 2015
- Rank the count of successful landings between 04/66/10 and 20/03/17

https://github.com/baw08d/IBM-Data-Science-Capstone/blob/fc1bbbefc246db9dcf4ee7a896e06fb81c3aeb9a/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Markers and labels for all launch sites added to map
- Pop icons added to each site to show the number of successful and failed missions
- Distance from VAFB SLC-4E to coast calculated and marked
- Distance from VAFB SLC-4E to Santa Barbara calculated and marked
- https://github.com/baw08d/IBM-Data-Science-Capstone/blob/fc1bbbefc246db9dcf4ee7a896e06fb81c3aeb9a/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- Added pie chart to show success rates for each site
- Added site dropdown to pie chart to show percentage of successful / failed missions at selected site
- Added scatter chart to show Payload vs. Launch Outcome
- Added range slider for the scatter plot to control which payload masses are displayed
- https://github.com/baw08d/IBM-Data-Science-Capstone/blob/c4096040808a4e0644daf758bf591cfb0719a7b4/spacex_dash_app.py

Predictive Analysis (Classification)

- Preprocessed and normalized data for X and Y values
- Split data into training and testing splits
- Created model objects for Logistic Regression, SVM, Decision Tree, and KNN
- Used GridSearchCV to choose best hyperparameters and fit the models
- Obtained score and confusion matrix for each model to determine which performed best
- https://github.com/baw08d/IBM-Data-Science-Capstone/blob/fc1bbbefc246db9dcf4ee7a896e06fb81c3aeb9a/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite%20(1).ipynb

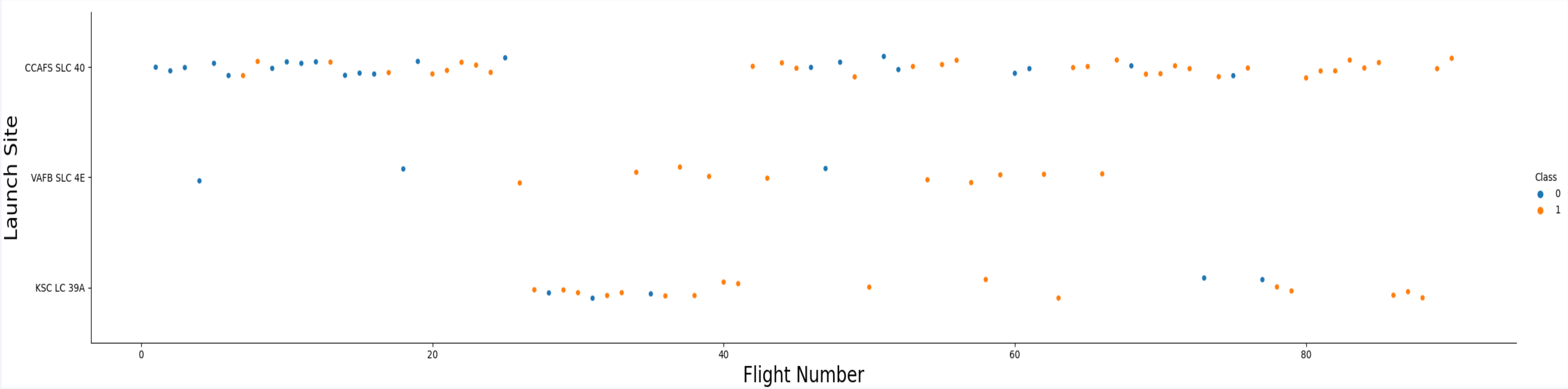
Results

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

Section 2

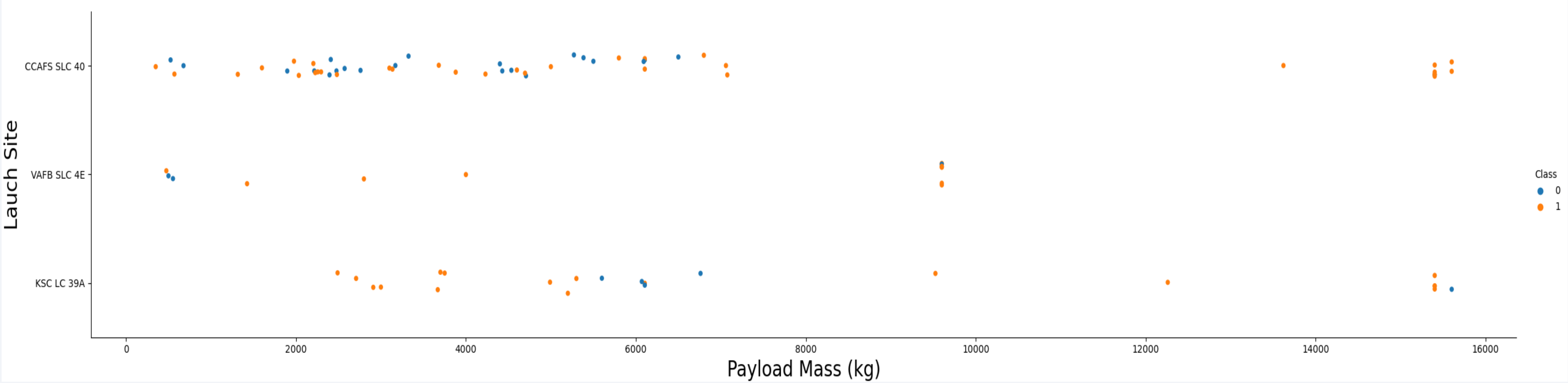
Insights drawn from EDA

Flight Number vs. Launch Site



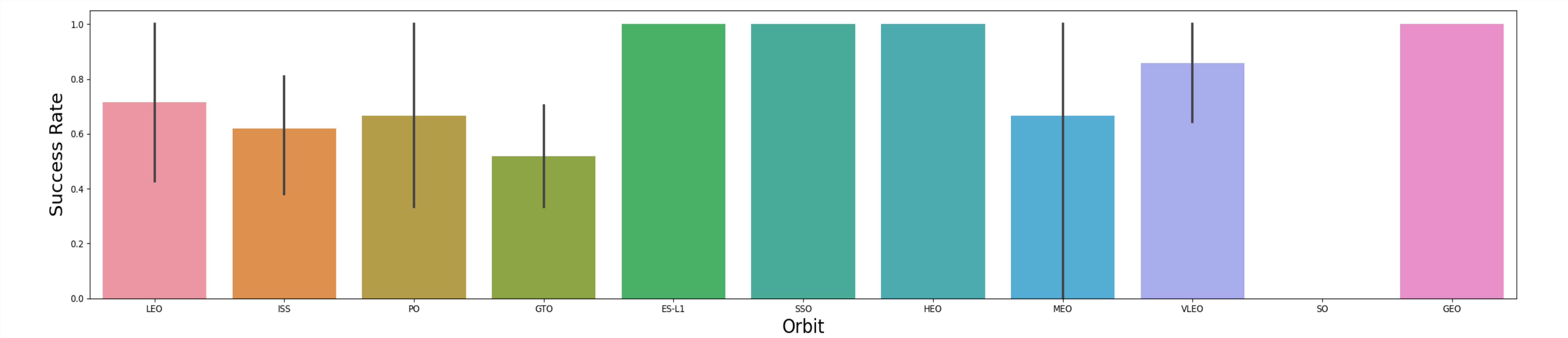
- At site VAFB SLC 4E, later flight numbers have a higher success rate
- Site CCAFS SLC 40 also has somewhat more successes as flight number increases

Payload vs. Launch Site



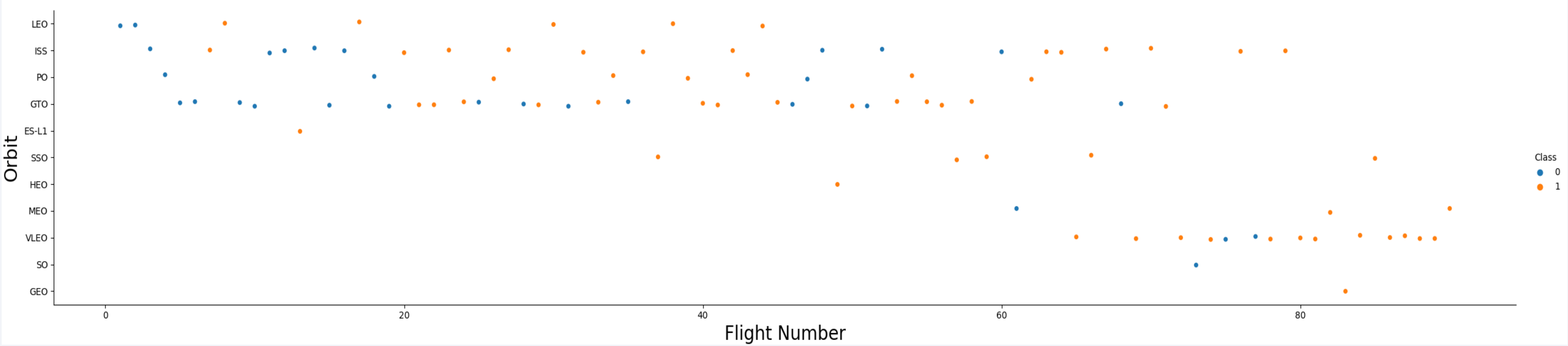
- The highest payload mass rockets were only launched from KSC LS 39A and CCAFS SLC 40
- These launches also have a high success rate
- Launches in the 6000kg range from KSC LS 39A had very low success rate

Success Rate vs. Orbit Type



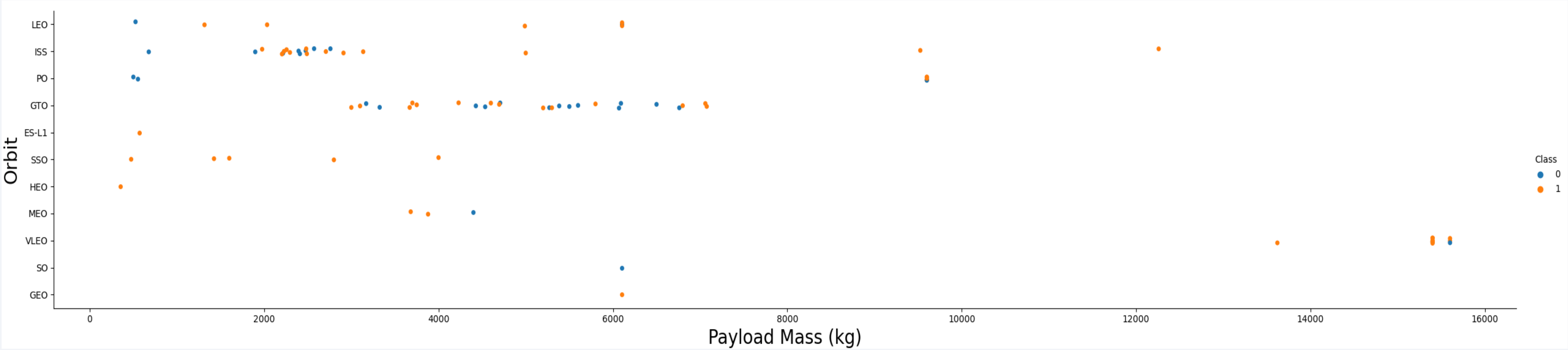
- ES-L1, SSO, HEO, and GEO orbit types all have 100% success rates
- There were no successful launches in the SO orbit
- All others have mixed success

Flight Number vs. Orbit Type



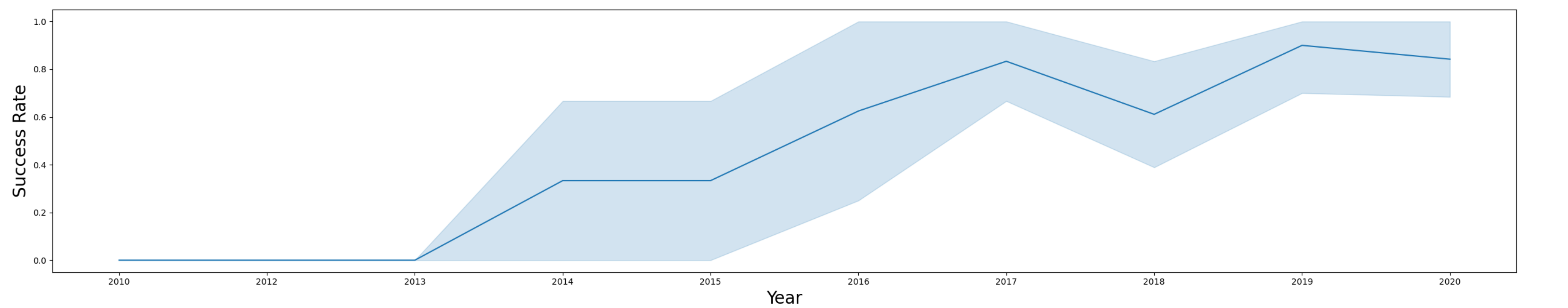
- Orbits VLEO, SO, and GEO are only the target of later Flight Number missions
- Orbit type LEO has a higher success rate in later missions
- As seen on the previous slide, SSO, HEO, GEO, and ES-L1 all have 100% success rate, however, only SSO has multiple launches for a good sample size

Payload vs. Orbit Type



- Missions to LEO orbit have a higher success rate with increased Payload Mass
- The highest Payload Masses are at VLEO orbit, and have a high success rate

Launch Success Yearly Trend



- There were no successful launches in 2010, 2011, or 2012
- Since 2013, there has been an overall improvement in success rate over time
- 2018 has a dip in successful launches compared to the surrounding year

All Launch Site Names

- Find the names of the unique launch sites:

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

```
%%sql

SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL;
```

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`:

In [8]:
```
%%sql
SELECT * FROM SPACEXTBL
WHERE LAUNCH_SITE LIKE 'CCA%'
limit 5;
```

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

Out[8]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	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:

SUM(PAYLOAD_MASS__KG_)
107010

- `%%sql`
- `SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL`
- `WHERE CUSTOMER LIKE'%NASA%';`

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1:

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

- **%%sql**
- **SELECT AVG**(PAYLOAD_MASS__KG_) **FROM** SPACEXTBL
- **WHERE** "Booster_Version" **LIKE** '%F9 v1.1%';

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad:

MIN("Date")
01-05-2017

- **%%sql**
- **SELECT MIN**("Date") **FROM** SPACEXTBL
- **WHERE** "Landing _Outcome" **LIKE** 'Success (ground pad)';

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:

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

- **%%sql**
- **SELECT** "Booster_Version" **FROM** SPACEXTBL
- **WHERE** "Landing _Outcome" **LIKE** 'Success (drone ship)'
- **and** PAYLOAD_MASS__KG_ **BETWEEN** 4000 **AND** 6000 ;

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes:

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

- **%%sql**
- **SELECT** "Mission_Outcome", **COUNT**(*) **FROM** SPACEXTBL
- **GROUP BY** "Mission_Outcome";

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass:

Booster_Version
F9 v1.1 B1018

- **%%sql**
- **SELECT** "Booster_Version" **FROM** SPACEXTBL
- **WHERE** "Booster_Version" **IN** (**SELECT MAX**("Booster_Version") **FROM** SPACEXTBL);

2015 Launch Records

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

substr("Date",4,2)	Booster_Version	Launch_Site	Landing _Outcome
01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- `%%sql`
- `SELECT substr("Date",4,2), "Booster_Version","Launch_Site","Landing _Outcome" FROM SPACEXTBL`
- `WHERE "Landing _Outcome" LIKE 'Failure (drone ship)'`
- `and substr("Date",7,4) = '2015';`

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

- Rank the count of successful landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order:

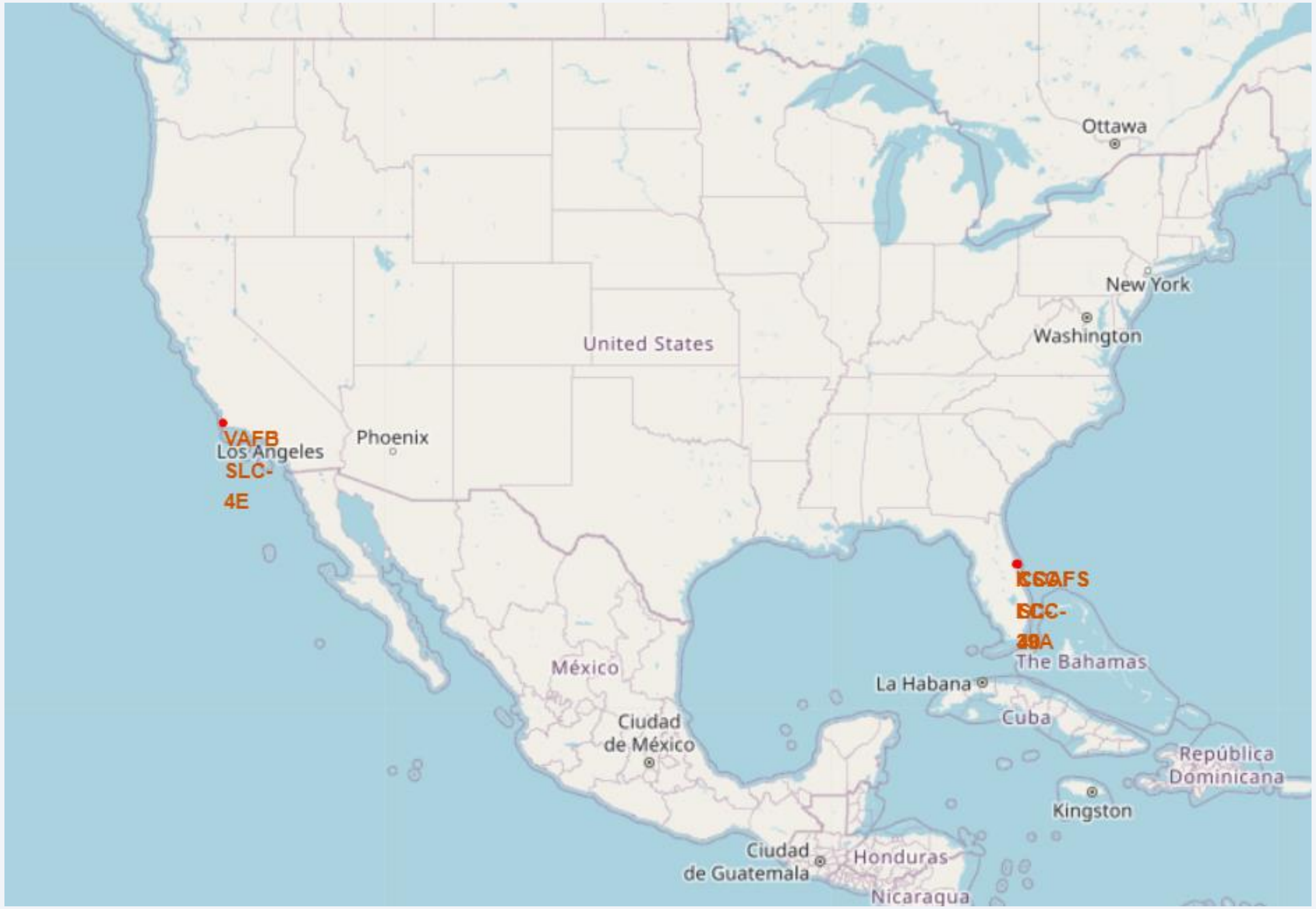
Landing _Outcome	COUNT("Landing _Outcome")
Success	20
Success (drone ship)	8
Success (ground pad)	6

- `%%sql`
- `SELECT "Landing _Outcome",COUNT("Landing _Outcome") FROM SPACEXTBL`
- `WHERE "Landing _Outcome" LIKE '%Success%'`
- `and "Date" BETWEEN "04-06-2010" AND "20-03-2017"`
- `GROUP BY "Landing _Outcome"`

Section 3

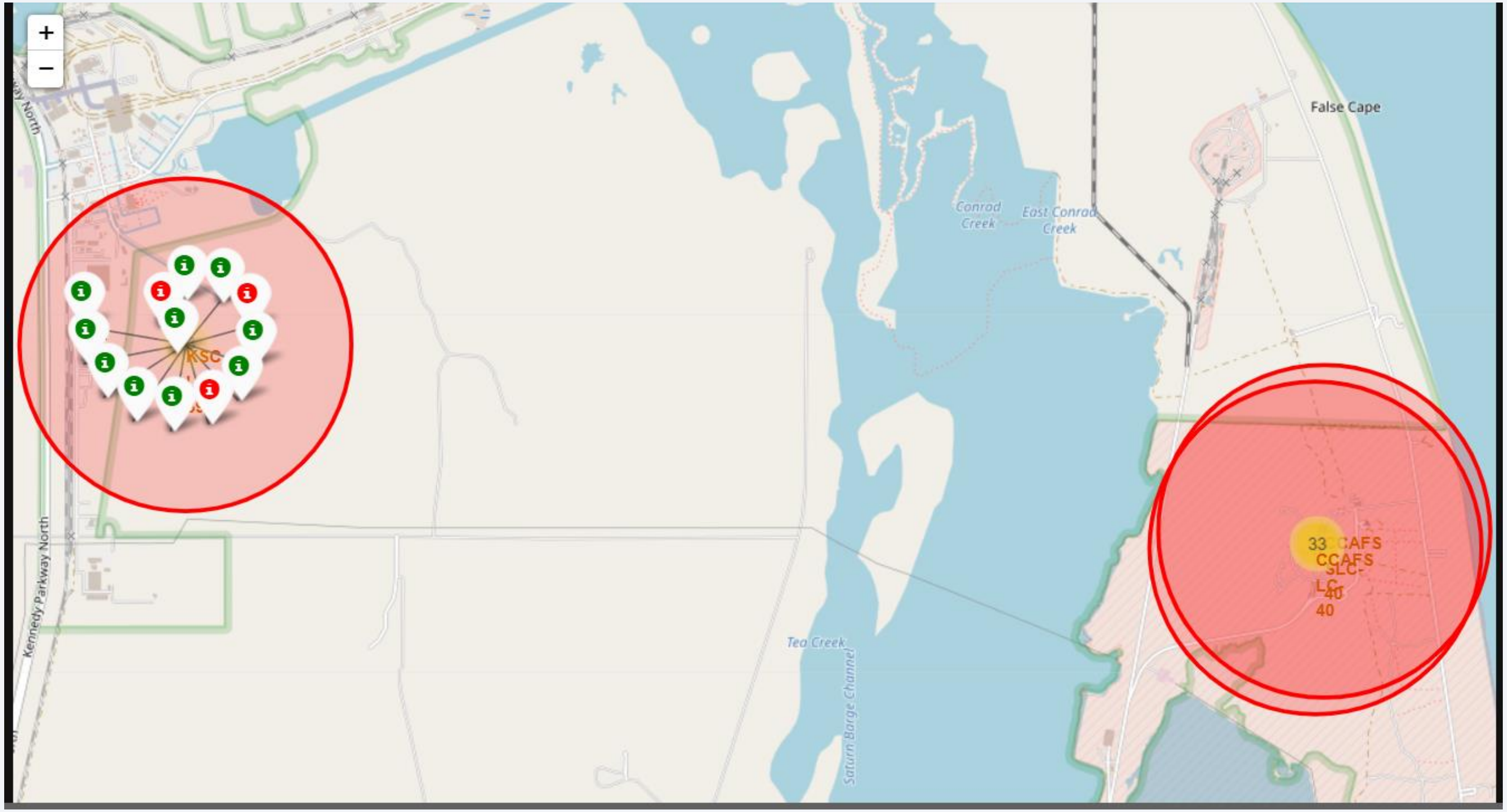
Launch Sites Proximities Analysis

Launch Site Locations on Folium Map



- Launch Site locations marked and labeled on Folium map.
- Three launch sites in Florida are located so closely that the labels overlap (KSC LC-39A, CCAFS SLC-40, CCAFS LC-40)
- The fourth site is on the opposite side of the country (VAFB SLC-4E)

Mission Outcomes at Each Site



- This map has pop up labels added to every site to display how many successful and failed missions were sent.
- Successful missions are green and failed are in red
- Screenshot shows KSC LC-39A, which has significantly more successful launches than failures

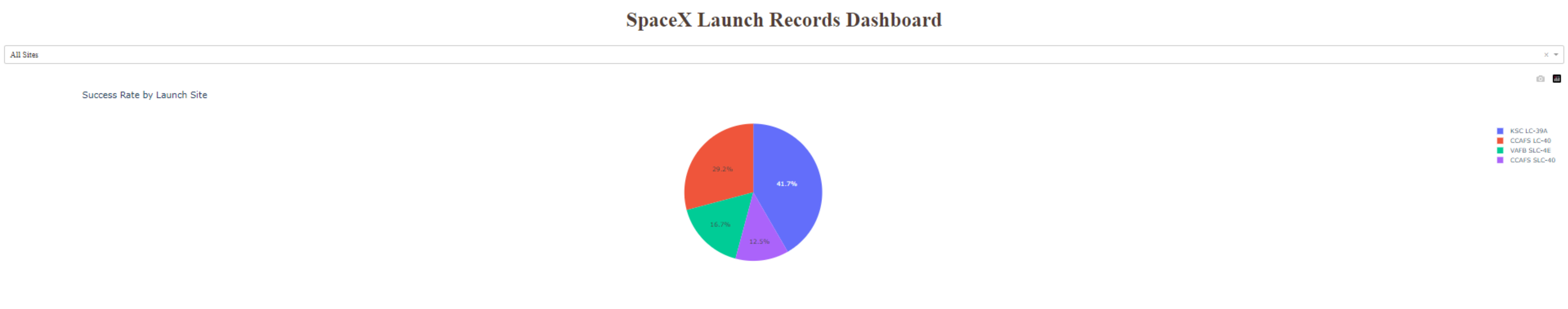
Distance from Launch Site to Nearest City

- On this map, the distance from VAFB SLC-4E to Santa Barbara has been calculated and marked
- The nearest city to this site is approximately 85.61 KM

Section 4

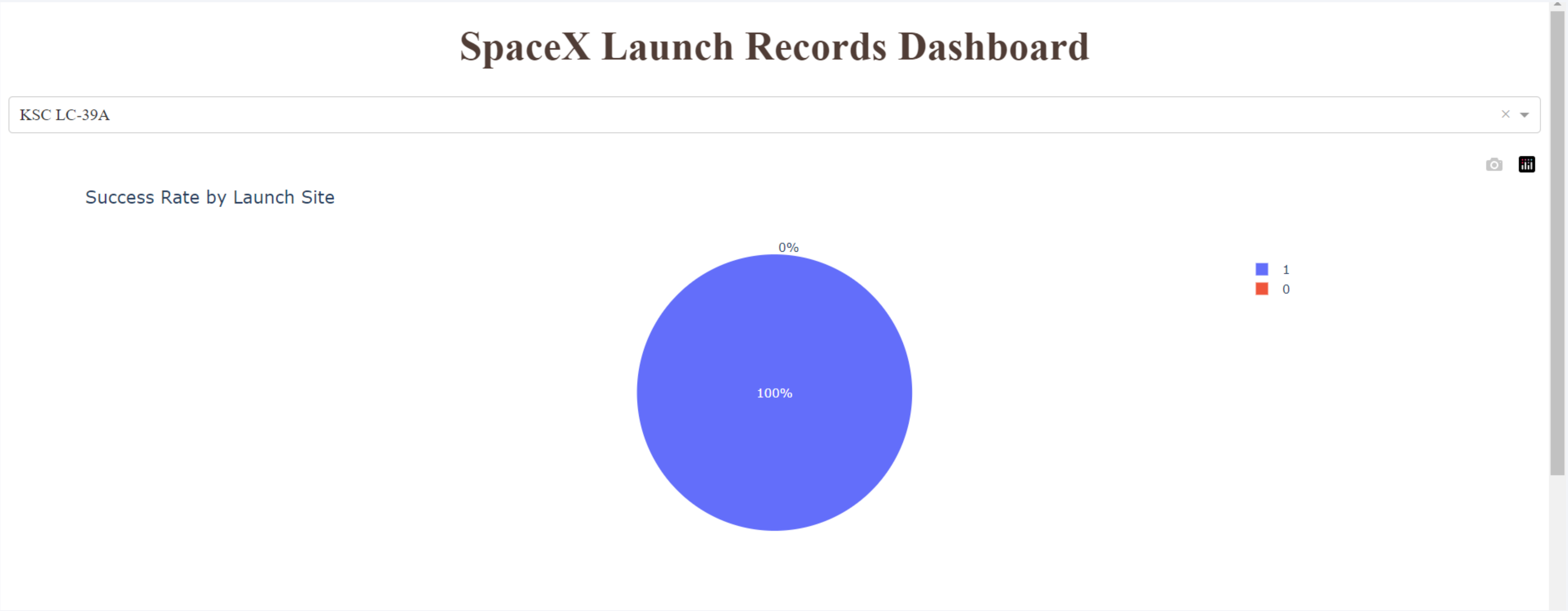
Build a Dashboard with Plotly Dash

Count of Launch Successes for All Sites



- This is a pie chart showing the success rate for all sites (All sites is selected in the dropdown)
- Site KSC LC-39A has the highest success rate at 41.7%, followed by CCAFS LC-40 at 29.2%

Single Launch Site Pie Chart



- Pie chart of success rate for one selected site

Payload vs. Launch Outcome Scatter Plot

- Payload vs. Landing Outcome with range slider above, colors show different booster versions
- V1.1 booster types have a low success rate, while FT boosters have a high success rate
- Overall, Payload Mass does not seem to have a consistent effect on success rate

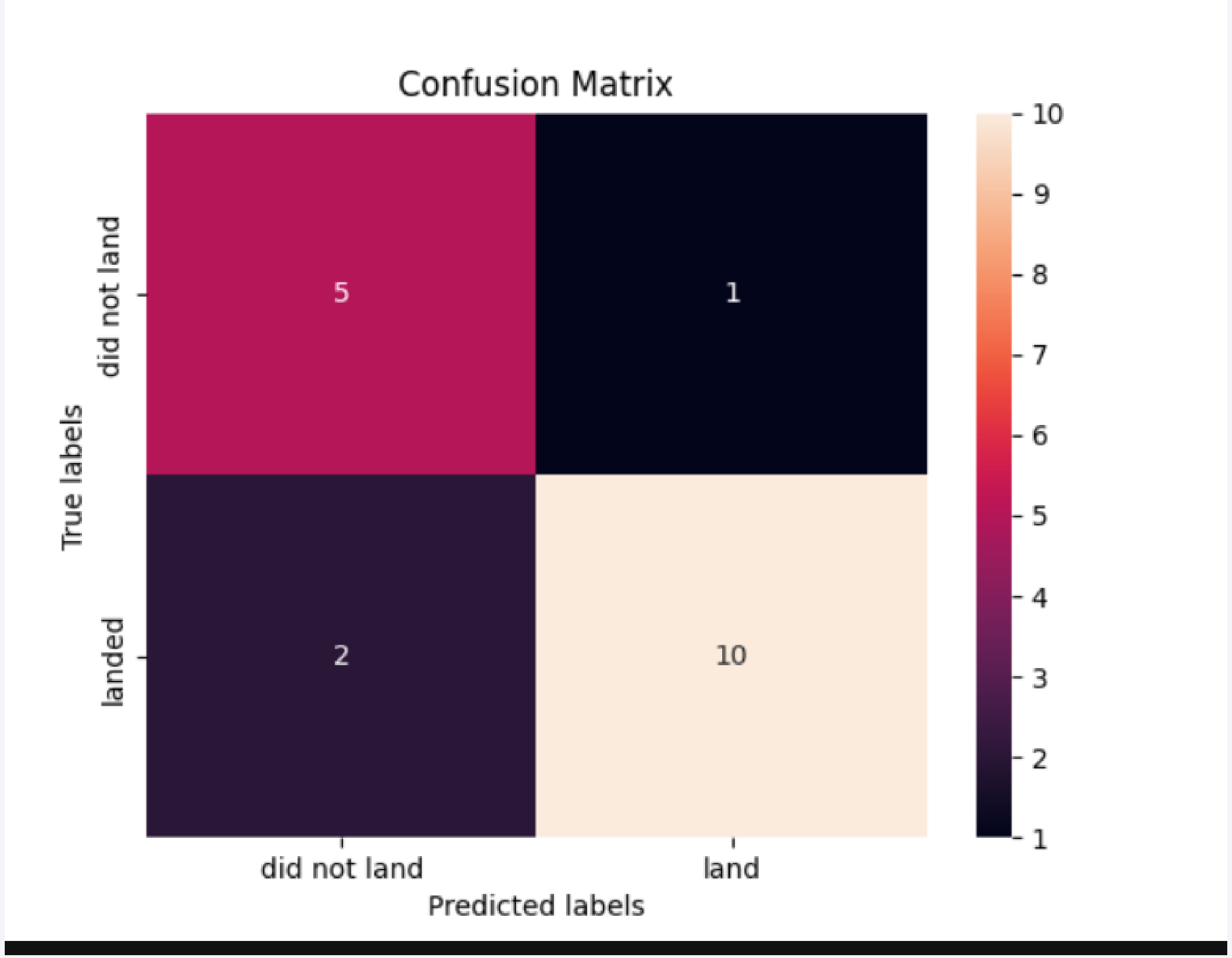
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- After fitting the best hyperparameters, all four model types correctly predicted 15 out of the 18 test samples
- The low sample size makes it difficult to know which model worked best
- KNN, SVM, and LR all had three false positives
- Decision Tree had one false positive and two false negatives
- Decision Tree was chosen as the best model due to the aims of the project; false negatives will be less harmful to our goal than false positives

Confusion Matrix



- Confusion Matrix for the Decision Tree Model
- There is one false positive in the upper right-hand corner, and two false negatives in the lower left-hand corner
- The other 15 outcomes were predicted correctly, for a model accuracy of 83.34%

Conclusions

- There is an overall upward trend in success rate over time, which is to be expected as SpaceX continues to learn from past launches and improve upon any issues
- Orbit type SSO has 100% success rate, even with multiple launches and would be a good target for future launches
- Certain Payload Masses have higher success with a specific Booster Version or Orbit type
- Launch site KSC LC-39A has the highest success rate

Thank you!