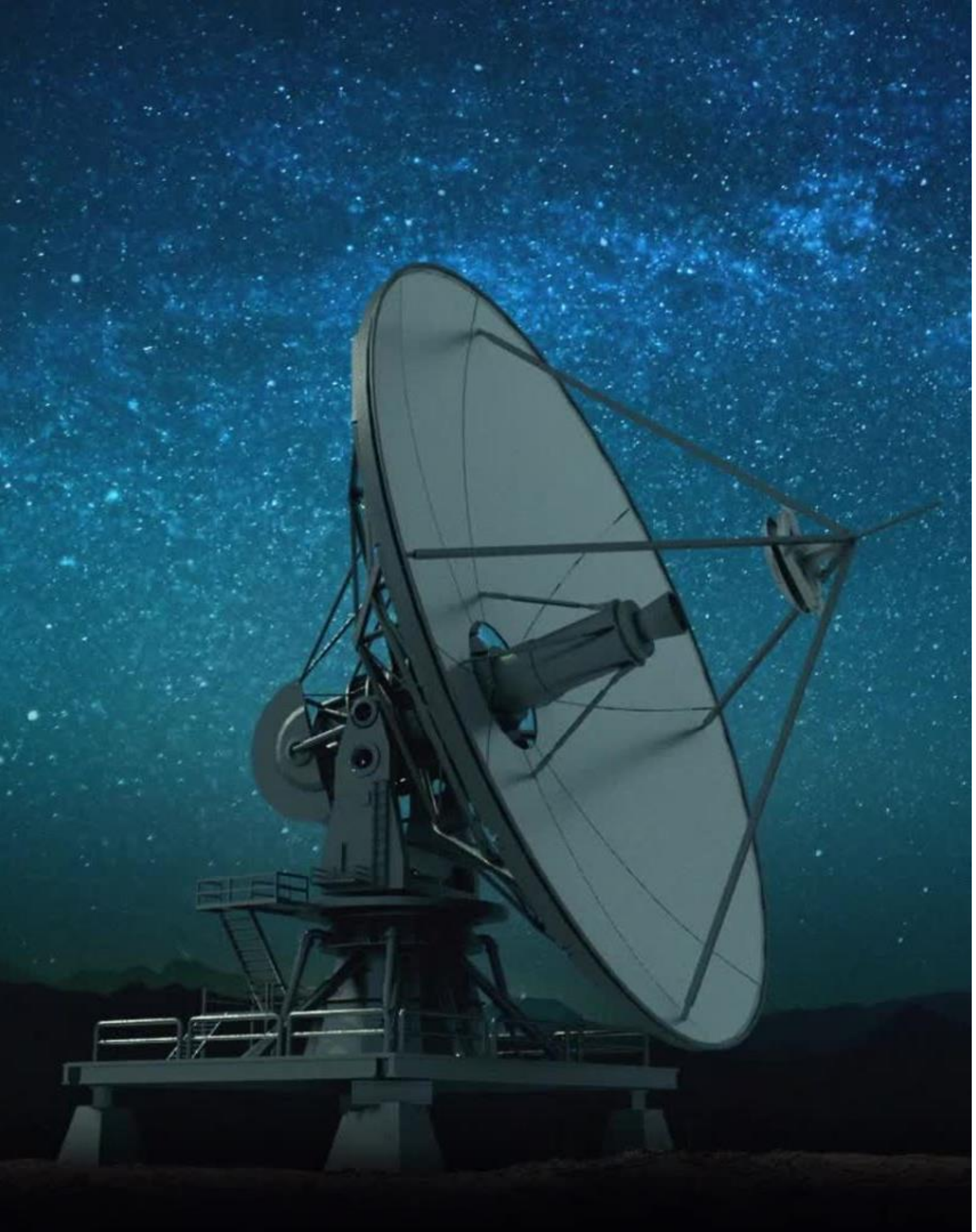
IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Julia Kalothi
12/28/2023

# Outline

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



# Introduction

**Background**

SpaceX, an industry leader in space travel has achieved significant milestones like ISS missions, satellite constellations, and manned spaceflights all while aiming for affordability. This was achieved by reusing the first stage of its Falcon 9 rocket, significantly reducing launch costs to $62 million. This research attempts to identify the factors for a successful rocking landing that is critical for SpaceX's success.

**Explore**

- How payload mass. Launch site, number of flights, and orbits affect first-stage landing success
- Rate of successful landings over time
- Best predictive model for a successful landing using binary classification models.

# Executive Summary

- Launch success has improved over time with KSC LC-39A having the highest success rate among landing sites
- Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate
- Most launches are near the equator and are close to the coast
- All of the models used in this analysis performed similarly on test and validation samples. SVM slightly outperformed other model types

Section 1

# Methodology

# Methodology

## Executive Summary

- Data collection methodology:
    - Data was collected using the SpaceX APT and web scraping from Wikipedia
- Perform data wrangling
    - One-hot encoding was applied to categorical 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
    - How to build, tune, evaluate classification models

# Data Collection

The object of this exercise was to extract the launch records as an HTML table, parse the table, convert it into a pandas data frame and clean the data for further analysis

- Data collection was completed by using a request to the SpaceX API
- Parsed the response using the .json() function and created a pandas data frame
- Cleaned the data, checked for missing values and imputed the missing where necessary
- Performed web scraping from Wikipedia for Falcon 9 launch records with BeautifulSoup

# Data Collection – SpaceX API

- Used the get request to the SpaceX API to collect the data and then cleaned and formatted the data for further analysis

- GitHub URL: https://github.com/stephejulia/IBM-Data-Science-Capstone-Project/blob/ccec8ae617b422acac71d1f52ae4d4c2c02fbc51/1_Data%20Collection%20with%20API.ipynb

1. Request rocket launch data from SpaceX API
2. Receive Response Data
3. Request and parse the SpaceX data using GET request
4. Decode the response content using .json() and turn it into a pandas dataframe using .json_normalize()

# Data Collection - Scraping

- Web scrapped Falcon 9 lunch records using BeautifulSoup
- Parsed the table and converted it into a pandas dataframe
- GitHub URL: https://github.com/stephejulia/IBM-Data-Science-Capstone-Project/blob/e0f464628ef33e7879cff82b7ad3cd5431e72dba/2_Data%20Collection%20with%20Web%20Scraping.ipynb

1. Perform a HTTP Get method to request the Falcon9 Launch HML page
2. Create a BeautifulSoup object from the HTML response
3. Extract the launch records using extract_column_from_header()
4. Create a dataframe by parsing the launch HTML tables

# Data Wrangling

- Describe how data were processed
- You need to present your data wrangling process using key phrases and flowcharts
- GitHub URL: https://github.com/stephejulia/IBM-Data-Science-Capstone-Project/blob/e0f464628ef33e7879cff82b7ad3cd5431e72dba/3_Data%20Wrangling.ipynb

# EDA with SQL

- Loaded the SpaceX dataset into a Posgress SQL database within jupyter notebook
- Applied EDA with SQL gain insights on the data. Here is an example of the queries:
    - Names of unique launch sites if the space mission
    - The total payload mass carried by boosters launched by NASA (CRS)
    - The average payload mass carried by booster version F9 v1.1
    - The total number of successful and failure mission outcomes
    - The failed landing outcomes in drone ship, their booster version and launce site names
- GitHub URL: https://github.com/stephejulia/IBM-Data-Science-Capstone-Project/blob/e0f464628ef33e7879cff82b7ad3cd5431e72dba/4_EDA%20with%20SQL.ipynb

# EDA with Data Visualization

- View relationship by using scatterplots. The variables could be useful for modeling in a relationship exists
- Show comparisons among discrete categories with bar charts

Explored the data by visualizing the relationship between the following:

- Flight Number the Launch Site
- Payload and Launch Site
- Success Rate of each Orbit Type
- Flight Number and Orbit Type
- Launch success yearly trend

GitHub URL: https://github.com/stephejulia/IBM-Data-Science-Capstone-Project/blob/e0f464628ef33e7879cff82b7ad3cd5431e72dba/5_EDA%20with%20Data%20Visualization.ipynb

# Build an Interactive Map with Folium

- NASA Johnson Space Center's coordinate with a popup label showing its name using latitude and longitude
- Added red circles on all launch coordinate sites with a pop-up label showing its name using latitude and longitude
- Added colored markers of successful (green) and unsuccessful (red) launches
- Used color-labeled marker clusters to identify which launce sites have had a relatively high success rate
- Calculated the distances between the launch site to its proximities. We answered the following:
  - How close are those launch sites to coastlines, railways, and highways
  - Do launch sites keep certain distances from cities
- GitHub URL: https://github.com/stephejulia/IBM-Data-Science-Capstone-Project/blob/e0f464628ef33e7879cff82b7ad3cd5431e72dba/6_Launch%20Site%20Location%20with%20Folium.ipynb

# Build a Dashboard with Plotly Dash

- Built an interactive dashboard with Plotly Dash
- Created a dropdown list with launch sites allowing the user to select all launch sites or a certain launch site
- Created a pie chart showing successful launches which allows the user to see successful and unsuccessful launches as a percent of the total
- Slider of Payload Mass Range allows the user to select payload mass range
- Plotted scatter graph showing the relationship with the Outcome (Success/Fail) and Payload Mass (Kg) for the different booster version.
- GitHub URL: https://github.com/stephejulia/IBM-Data-Science-Capstone-Project/blob/e0f464628ef33e7879cff82b7ad3cd5431e72dba/7_%20Interactive%20Visual%20Analytics%20Plotly.py

# Predictive Analysis (Classification)

- Loaded the data using numpy and pandas, transformed the data, and split the data into train and test
- Fit logistic regression, SVM, decision trees, and KNN models and tuned the hyperparameters using GridSearchCV
- Used accuracy, F1, and Jaccard to identify the best performing model
- GitHub: https://github.com/stephejulia/IBM-Data-Science-Capstone-Project/blob/e0f464628ef33e7879cff82b7ad3cd5431e72dba/8_Machine%20Learning%20Prediction.ipynb

# Results

**Exploratory Data Analysis**

- Launch success has improved over time
- KSC LC-39A has the highest success rate among landing sites
- Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate

**Visualization**

- Most launch sites are near the equator and are close to the coast
- Launch sites are far enough away from to not threaten cities, highways and railways but are close enough to bring people and materials to support launch activities

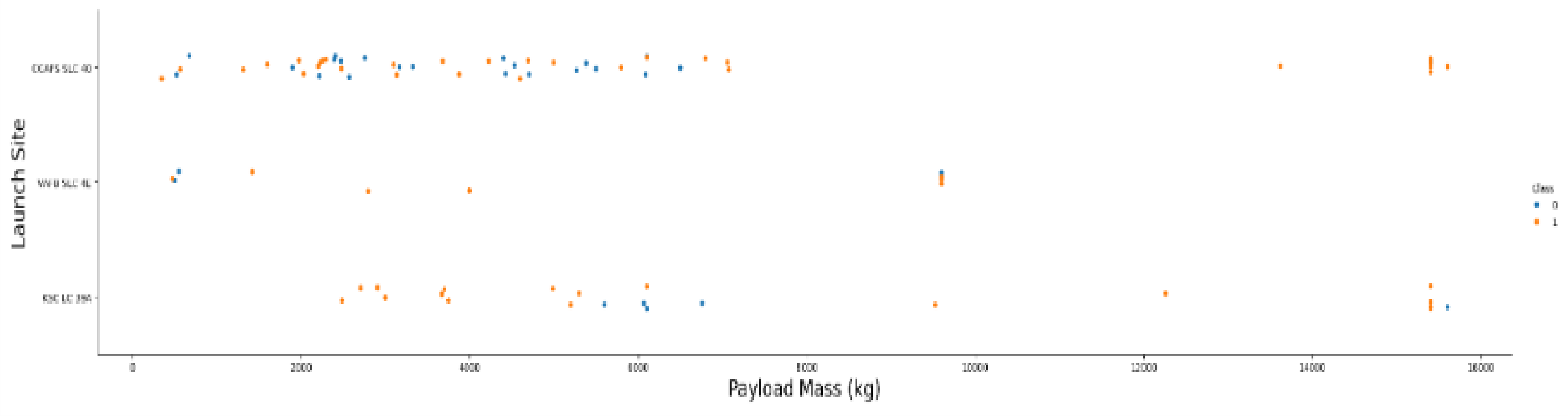
**Predictive Analysis**

- The decision tree model is the best predictive model for the data provided

Section 2

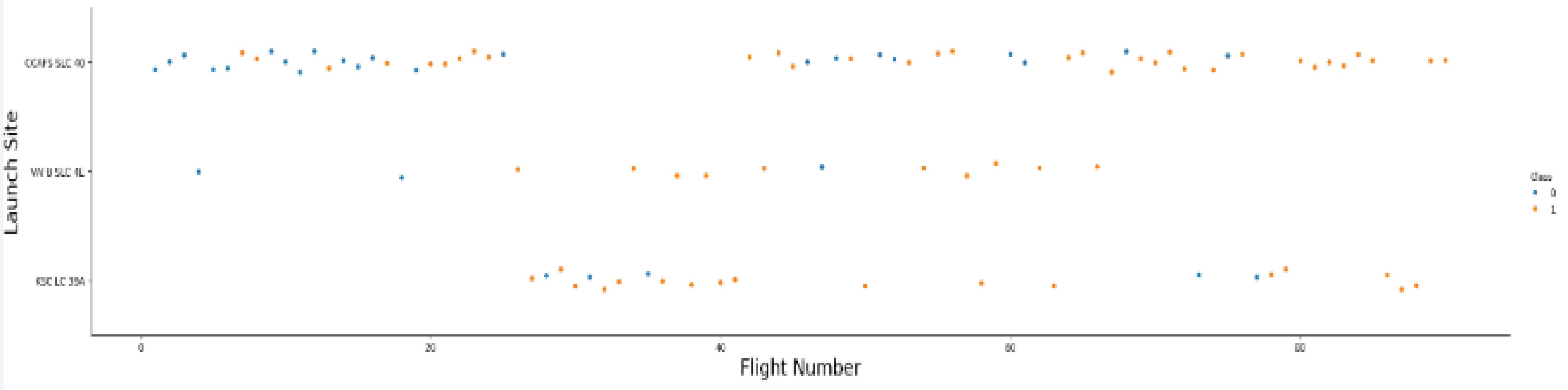
# Insights drawn from EDA

# Payload vs. Launch Site




- Earlier fights had lower success rates (blue=fail)
- Later flights had higher success rates (orange=success)
- Around half of the launches were from CCAFS SLC 40 launch site
- VAFB SLC 4E and KSC LC 39A have higher success rates
- We can infer that new launches have a higher success rate

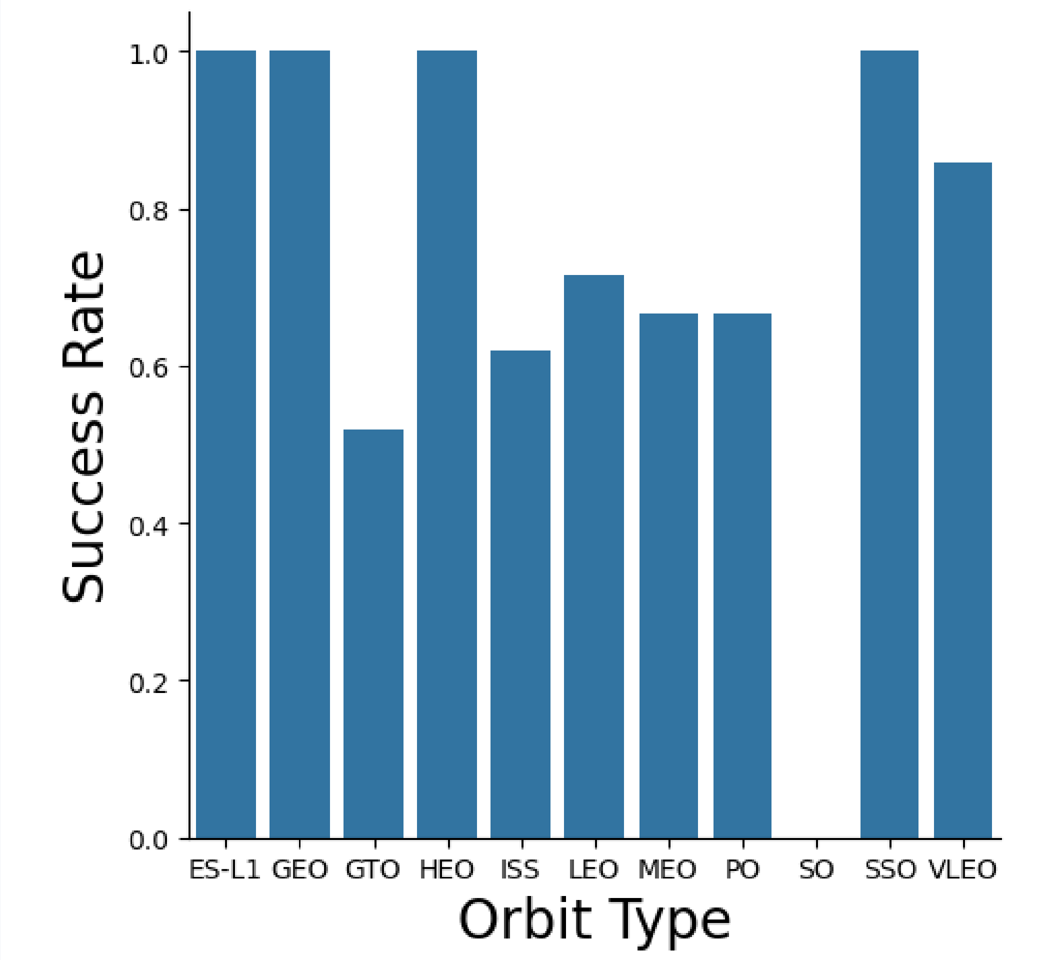
# Flight Number vs. Launch Site



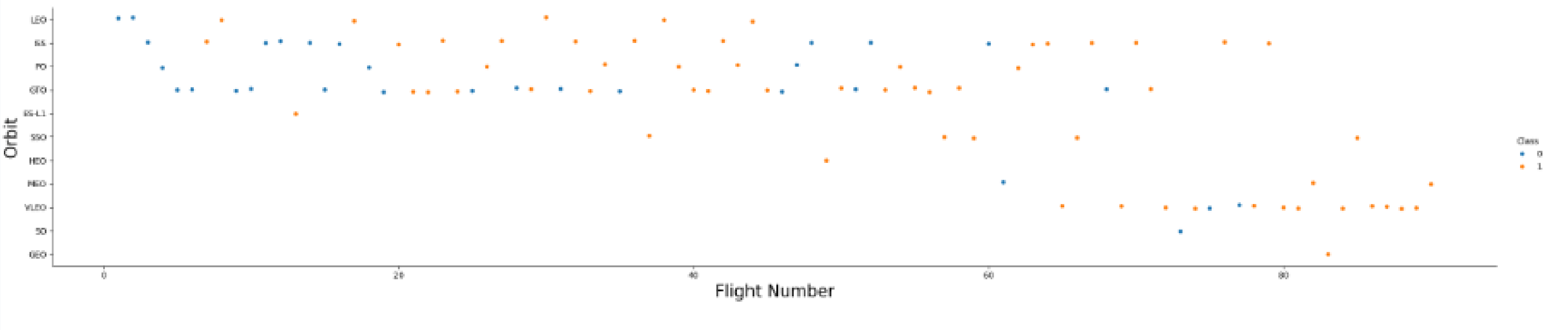

- Typically, the higher the Payload Mass (Kg), the higher the success rate
- Most launches with a payload greater than 7,000 kg were successful
- KSC LC 39A has a 100% success rate for launches with less than 5,500 kg
- VAFB SKC 4E has not launched anything greater than ~10,000 kg

# Success Rate vs. Orbit Type

- **100% Success Rate:** ES-L1, GEO, HEO, and SSO
- **50%-80% Success Rate:** GTO, ISS, LEO, MEO, POT
- **0% Success Rate:** SO

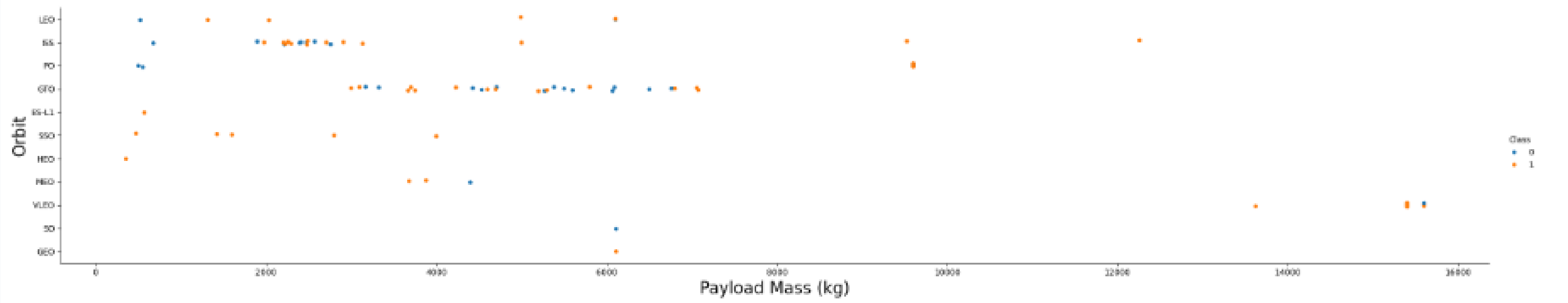


# Flight Number vs. Orbit Type




- The success rate typically increases with the number of lights for each orbit
- This relationship is highly apparently for the LEO orbit
- The GTO orbit, however, does not follow this trend

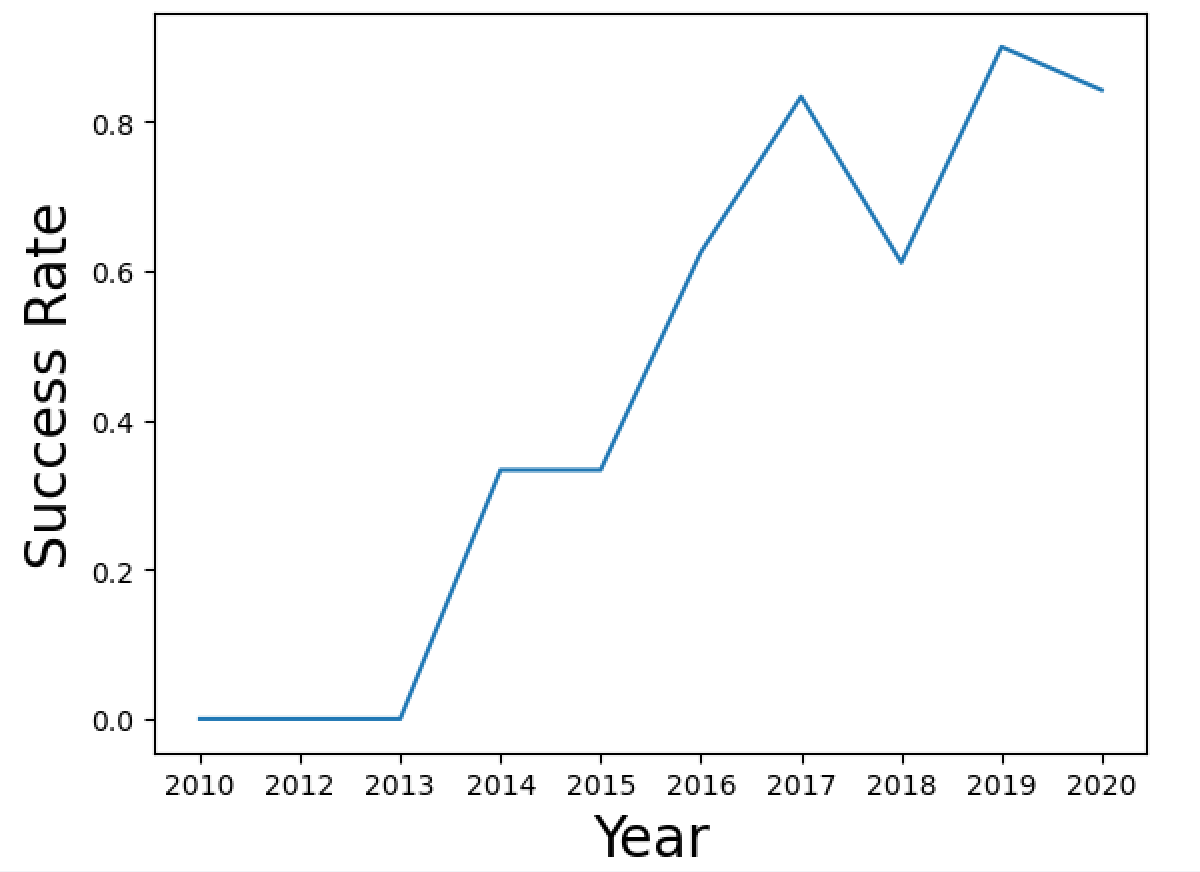
# Payload vs. Orbit Type




- Heavy payloads are better with LEO, ISS, and PT orbits
- The GTO orbit has mixed success with heavier payloads

# Launch Success Yearly Trend

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



# All Launch Site Names

Launch Site Names

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

### Task 1

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

In [12]:
```
%sql select distinct launch_site from SPACEXTABLE;
```

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

Out[12]:

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

# Launch Site Names Begin with 'CCA'

- Records starting with CCA – Displaying 5 records below

## Task 2

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

In [13]:
```
%sql select * from SPACEXTABLE where launch_site like 'CCA%' limit 5;
```

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

Out[13]:

| 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

- Total Payload Mass – 45,596 kg carried by boosters launched by NASA (CRS)

### Task 3

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

In [14]:
```
%sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXTABLE where customer = 'NASA (CRS)';
```

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

Out[14]:

| total_payload_mass |
|---|
| 45596 |

# Average Payload Mass by F9 v1.1

- 2,534.7 kg average payload mass carried by booster version F9 v1.1

### Task 4

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

In [15]:
```
%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXTABLE where booster_version like '%F9 v1.1%';
```

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

Out[15]:

| average_payload_mass |
|---|
| 2534.6666666666665 |

# First Successful Ground Landing Date

- The first successful landing in Ground Pad was 12/22/2015

### Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

*Hint:Use min function*

In [19]:
```
%sql select min(date) as first_successful_landing from SPACEXTABLE where landing_outcome = 'Success (ground pad)';
```

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

Out[19]:

| first_successful_landing |
|---|
| 2015-12-22 |

# Successful Drone Ship Landing with Payload between 4000 and 6000

- Booster Drone Ship Landing with Payload between 4000-6000
  - JSCAT-14, JSCAT-16, SES-10, SES-11/EchoStar 105

### Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [21]:
```
%sql select booster_version from SPACEXTABLE where landing_outcome = 'Success (drone ship)' and payload_mass__kg_ between 40
```

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

Out[21]:

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

# Total Number of Successful and Failure Mission Outcomes

- Total Number Sucess and Failure Missions
  - 99 Success
  - 1 Flight Failure (in flight)
  - 1 Success (Payload Status unclear)

### Task 7

List the total number of successful and failure mission outcomes

In [22]:
```
%sql select mission_outcome, count(*) as total_number from SPACEXTABLE group by mission_outcome;
```

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

Out[22]:

| Mission_Outcome | total_number |
|---|---|
| Failure (in flight) | 1 |
| Success | 98 |
| Success | 1 |
| Success (payload status unclear) | 1 |

# Boosters Carried Maximum Payload

- Below is a query and a list of the names of the boosters which have carried the maximum payload mass:

- 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

## Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [24]:
```
%sql select booster_version from SPACEXTABLE where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTABLE);
```

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

Out[24]:

| 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

Booster Versions F9 v1.1 B1012 and F9 v1.1 B1015 Failed due to drone ship in 2015

## Task 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.

**Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.**

21]:
```
%%sql select substr(Date,6,2) as month, date, booster_version, launch_site, landing_outcome from SPACEXTABLE
    where landing_outcome = 'Failure (drone ship)' and substr(Date,0,5)='2015';
```

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

21]:

| month | Date | Booster_Version | Launch_Site | Landing_Outcome |
|---|---|---|---|---|
| 01 | 2015-01-10 | F9 v1.1 B1012 | CCAFS LC-40 | Failure (drone ship) |
| 04 | 2015-04-14 | F9 v1.1 B1015 | CCAFS LC-40 | Failure (drone ship) |

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

- Rank in descending order the landing outcomes between 2010-06-04 and 2017-03-20

### Task 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.

[18]:
```
%%sql select landing_outcome, count(*) as count_outcomes from SPACEXTABLE
      where date between '2010-06-04' and '2017-03-20'
      group by landing_outcome
      order by count_outcomes desc;
```

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

[18]:

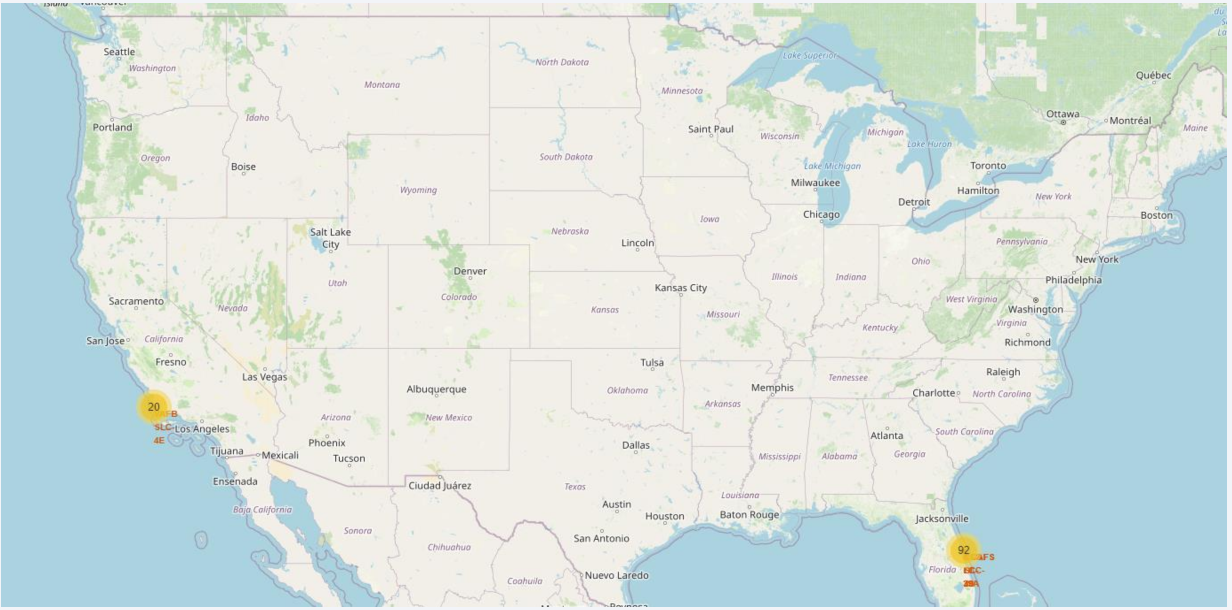
| Landing_Outcome | count_outcomes |
|---|---|
| No attempt | 10 |
| Success (drone ship) | 5 |
| Failure (drone ship) | 5 |
| Success (ground pad) | 3 |
| Controlled (ocean) | 3 |
| Uncontrolled (ocean) | 2 |
| Failure (parachute) | 2 |
| Precluded (drone ship) | 1 |

Section 3

# Launch Sites Proximities Analysis

# Launch Site Locations

- **Near the Equator:** The closer the launch site is to the equator, the easier it is to launch to equatorial orbit, and the more help you get from the Earth's rotation for a prograde orbit.



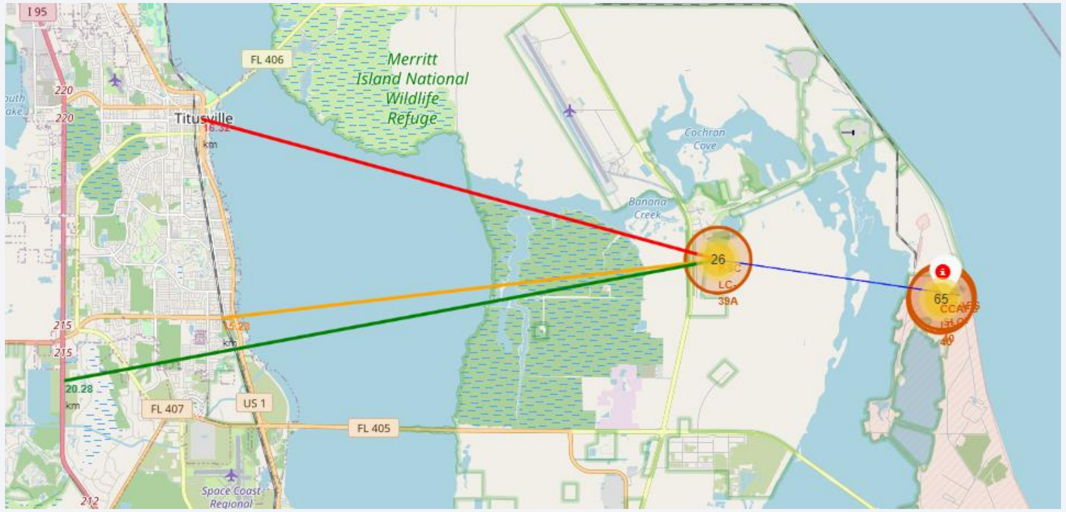
# Launch Site Outcomes

Launch Outcomes at Space Launch Complex 40

- 14 Green markers for successful launches
- 38 Red markers for unsuccessful launches
- Overall success rate of 27%

# Launch Site Proximities

- Coasts help ensure that the spent stages drop along the launch path or failed launces do not fall on people or property
- The need to be in exclusion zones to keep unauthorized people away
- These sites need to be away from anything a failed launch could damage, but close enough to roads, rails, and docks to be able to bring people and materials to and from the launch site



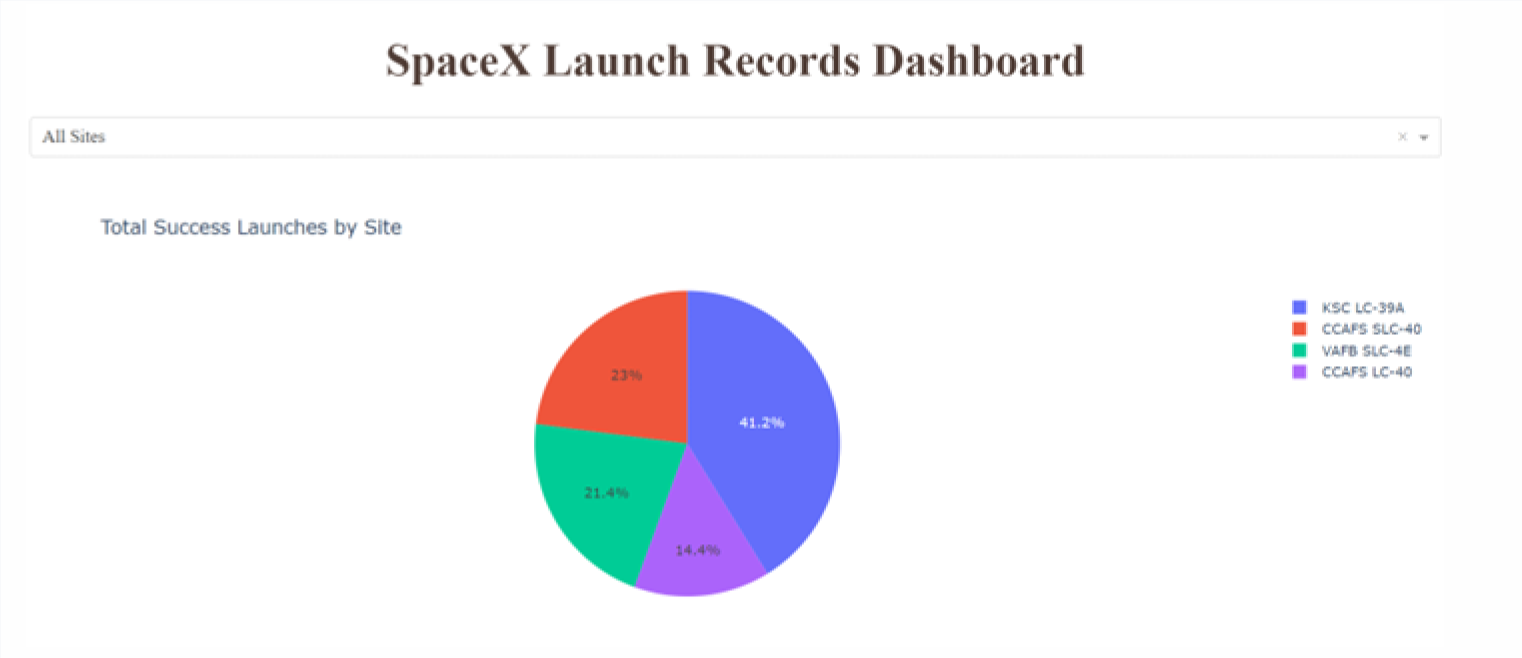
Section 4

# Build a Dashboard with Plotly Dash

# Launch Success by Site

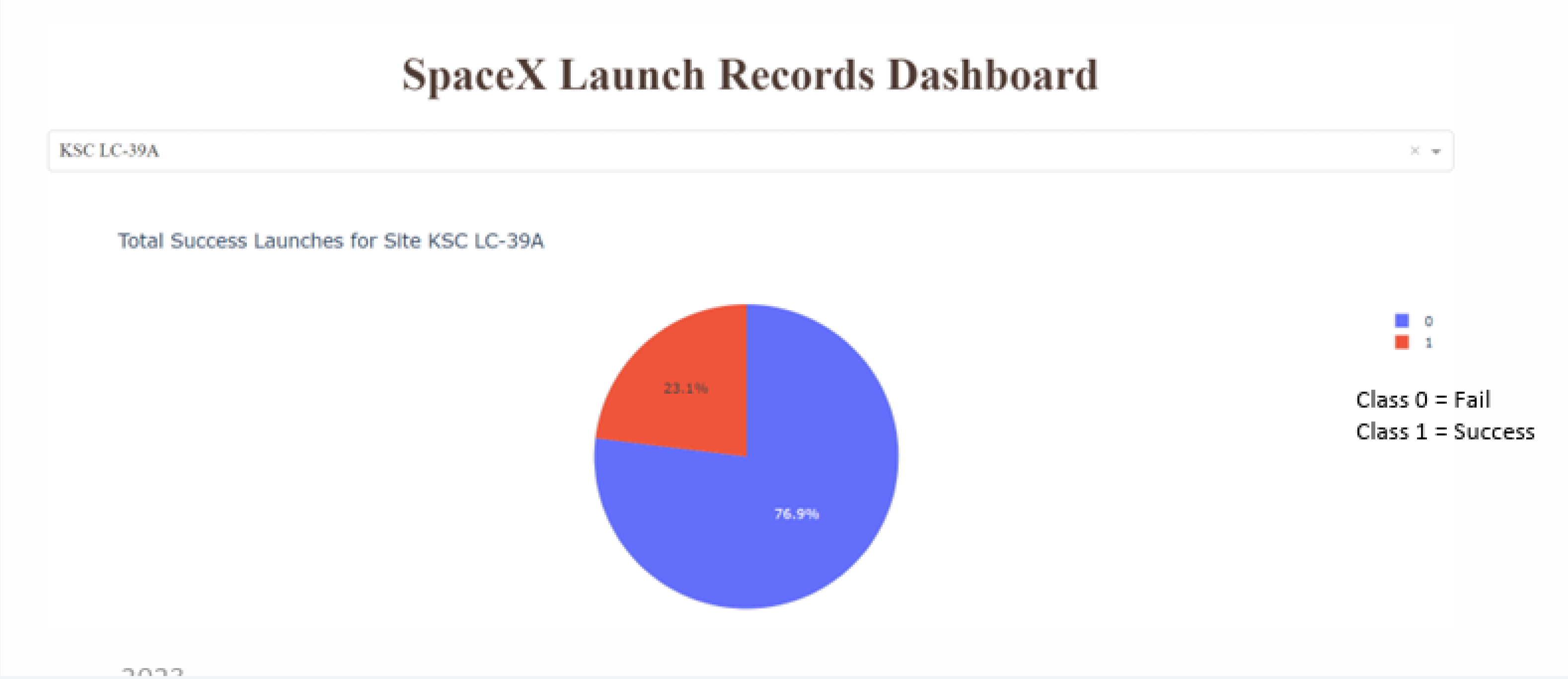
Success as a Percent of Total

KSC LC-39A has the most successful launches amongst launch sites (41.2%)



# Launch Success (KSC LC-39A)

- KSC LC-39A has the highest success rate amongst launch sites (76.9%)
- 10 successful launches and 3 failed launches



# Payload vs Launch Outcome

- Payloads between 2,000kg and 5,000 kg have the highest success rate
- 1= success, 0=unsuccessful

Section 5

# Predictive Analysis (Classification)

# Classification Accuracy

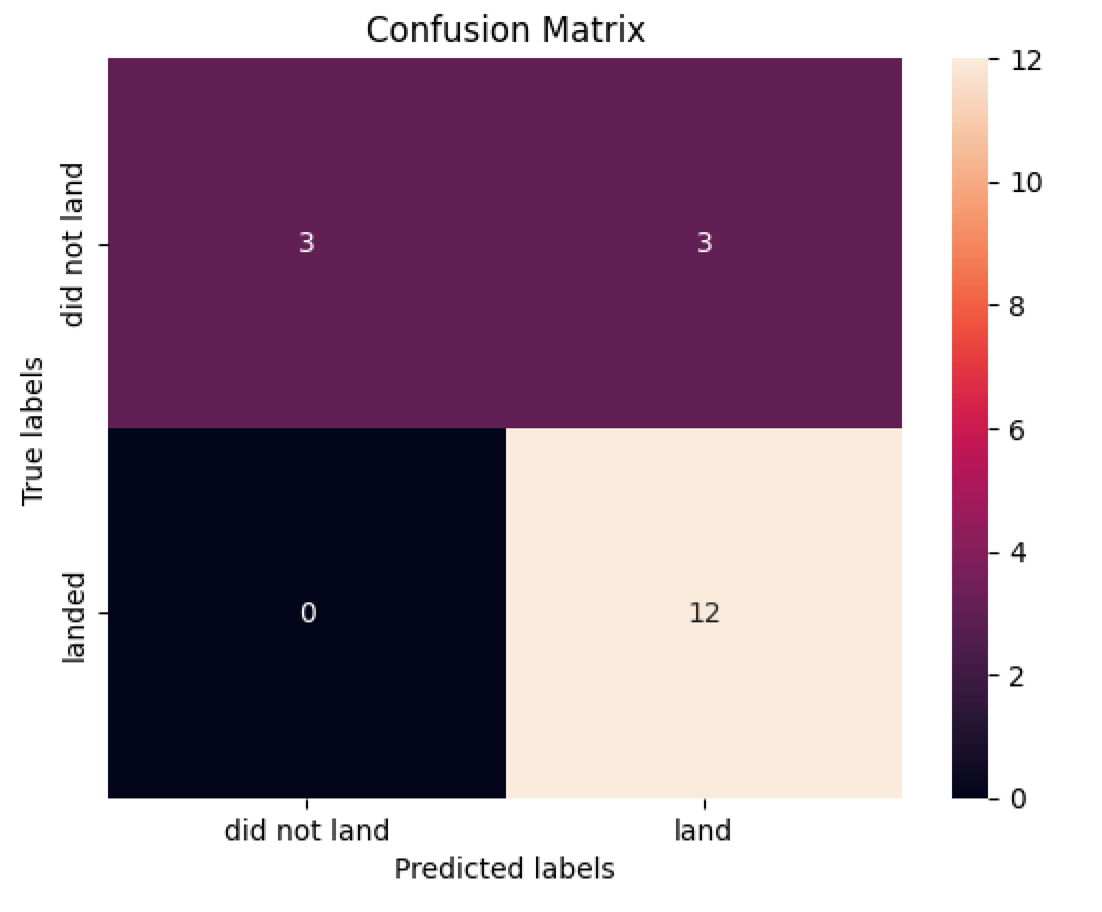
- All models performed at or about the same level. The decision tree had an error I could not debug so, based on my results, the SVM out performed all other model types

Out[33]:

| | LogReg | SVM | Tree | KNN |
|---|---|---|---|---|
| **Jaccard_Score** | 0.833333 | 0.845070 | 0.746667 | 0.819444 |
| **F1_Score** | 0.909091 | 0.916031 | 0.854962 | 0.900763 |
| **Accuracy** | 0.866667 | 0.877778 | 0.788889 | 0.855556 |

# Confusion Matrix

- A confusion matrix summarizes the performance of a classification algorithm
- True Positive=12
- True Negative=3
- False Positive=3
- False Negative=0
- There are 3 false positives that are a Type 1 error which is not good



# Conclusions

- Most of the launch sites are near the equator for the additional natural boost and all launch sites are near the coast for safety reasons
- SpaceX launch success has increased overtime
- KSC LC-39A has the highest success rate among launch sites. It has a 100% success rate for launches less than 5,500kg
- Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate
- Across all launch sites, the higher the payload mass (kg), the higher the success rate
- The model performance of the SVM model outperformed other model types on this data

# Thank you!