



IBM Developer  
SKILLS NETWORK

# Winning Space Race with Data Science

Nicholas Judice  
June 29, 2022



# Outline

---

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

# Executive Summary

---

- Methodologies Used:
  - Data Collection by Web Scraping and SpaceX API
  - Data Wrangling, Exploratory Data Analysis (EDA) with Data Visualization and SQL,
  - Interactive Visual Analytics via Folium maps and Plotly Dash
  - Machine Learning and Predictive Analysis
- Summary of all results
  - Public Sources contained useful data
  - EDA in conjunction with Machine Learning help predict which features are most vital for successful launches.

# Introduction

---

- Project background:
  - In order for our new company to compete with SpaceX, we must take a look at how they are able to effectively cut costs compared to their competition. The ability to reuse materials due to successful launches and landings provides a major advantage.
- Problems you want to find answers:
  - Which conditions allow for the greatest success rates for launches?
  - Which rocket variables have greatest impact on success rates?



Section 1

# Methodology

# Methodology

---

## Executive Summary

- Data collection methodology:
  - SpaceX API
  - Web Scraping SpaceX Wikipedia tables.
- Perform data wrangling
  - Simplifying landing outcomes was achieved by analyzing collected data categories.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Dividing data into training and test sets, and evaluating said data through 4 different methods.

# Data Collection

---

- Data was Scraped from Wikipedia  
([https://en.wikipedia.org/wiki/List\\_of\\_Falcon\\_9\\_and\\_Falcon\\_Heavy\\_launches](https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches))
- SpaceX API was also used (<https://api.spacexdata.com/v4/rockets/>)

# Data Collection – SpaceX API

---

- SpaceX public API is used.
- Source code:  
<https://github.com/njudice/Data-Science-Capstone/blob/master/Data%20Collection%20API%20lab.ipynb>





# Data Collection - Scraping

---

- Web Scrape Wikipedia Page titled "List of Falcon 9 and Falcon Heavy Launches"
- Source  
Code: <https://github.com/njudice/Data-Science-Capstone/blob/master/Complete%20Data%20Collection.ipynb>

Request Falcon9 Launch Wiki from its URL and use BeautifulSoup



Extract all column/variable names from HTML table header

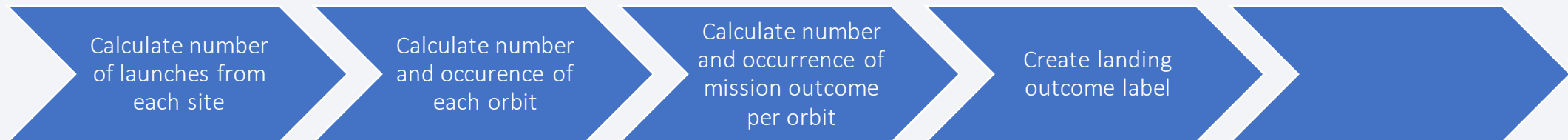


Create DF by parsing HTML launch tables

# Data Wrangling

---

- I calculated number and occurrence of launches per site, occurrence per orbit, and occurrence of mission outcome per orbit. This yielded 8 possible outcomes.
- Once outcomes were assigned integer values, success rate can be calculated.
- Source Code: <https://github.com/njudice/Data-Science-Capstone/blob/master/EDA.ipynb>



# EDA with Data Visualization

---

- Scatter Point Charts to represent relationships between: Flight Number and Launch Site, Launch Site and Payload Mass, Orbit and Flight Number, Payload Mass and Orbit
- Bar Graph to show Success Rate by Orbit
- Line Graph of Success Rate by Year
- Source Code: <https://github.com/njudice/Data-Science-Capstone/blob/master/EDA%20with%20Data%20Visualization.ipynb>

# EDA with SQL

---

- Display names of unique Launch Sites
- Display 5 records where Launch Sites begin with "CCA"
- Display total Payload Mass carried by Boosters launched by NASA
- Display average Payload Mass carried by Booster version F9 v1.1
- List date when first successful landing outcome on ground pad was achieved
- List names of Boosters which have success in drone ship and have payload mass 4000-6000
- List total number of successful and failure mission outcomes
- List names of booster versions which have carried max payload mass
- List records to display month names, failure landing outcomes in drone ship, booster versions, and launch site for months in year 2015
- Rank count of successful landing outcomes from 4/6/2010 to 3/20/2017
- Source Code: [https://github.com/njudice/Data-Science-Capstone/blob/master/jupyter-labs-eda-sql-coursera\\_sqlite%20\(1\)%20\(1\).ipynb](https://github.com/njudice/Data-Science-Capstone/blob/master/jupyter-labs-eda-sql-coursera_sqlite%20(1)%20(1).ipynb)

# Build an Interactive Map with Folium

---

- Created and added Circles, Markers, Lines, and Marker Clusters
  - Circles show highlighted areas around objects such as Launch Sites
  - Markers show specific points or instances
  - Lines help measure distance between to points
  - Marker Clusters indicate multiple events in an area.
- Source Code: <https://github.com/njudice/Data-Science-Capstone/blob/master/Folium%20Lab.ipynb>

# Build a Dashboard with Plotly Dash

---

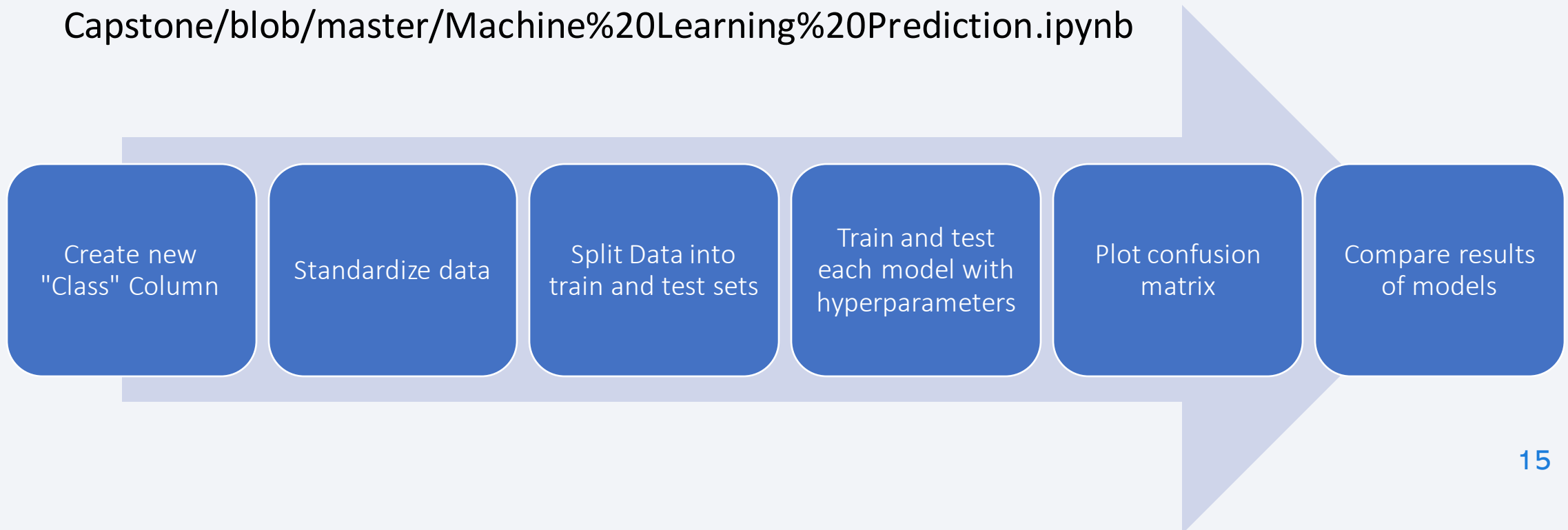
- Added dropdown menus to choose Launch Site or All Sites. The selection shows success rates for site of the user's choosing.
- Scatter plot shows chosen Site along with relation to Payload Mass, which can also be adjusted via sliding scale.
- Adjusting of Site and Payload Mass gives more clear picture of combinations that yield successful launches.
- Source Code: [https://github.com/njudice/Data-Science-Capstone/blob/master/spacex\\_plotly\\_dash](https://github.com/njudice/Data-Science-Capstone/blob/master/spacex_plotly_dash)



# Predictive Analysis (Classification)

---

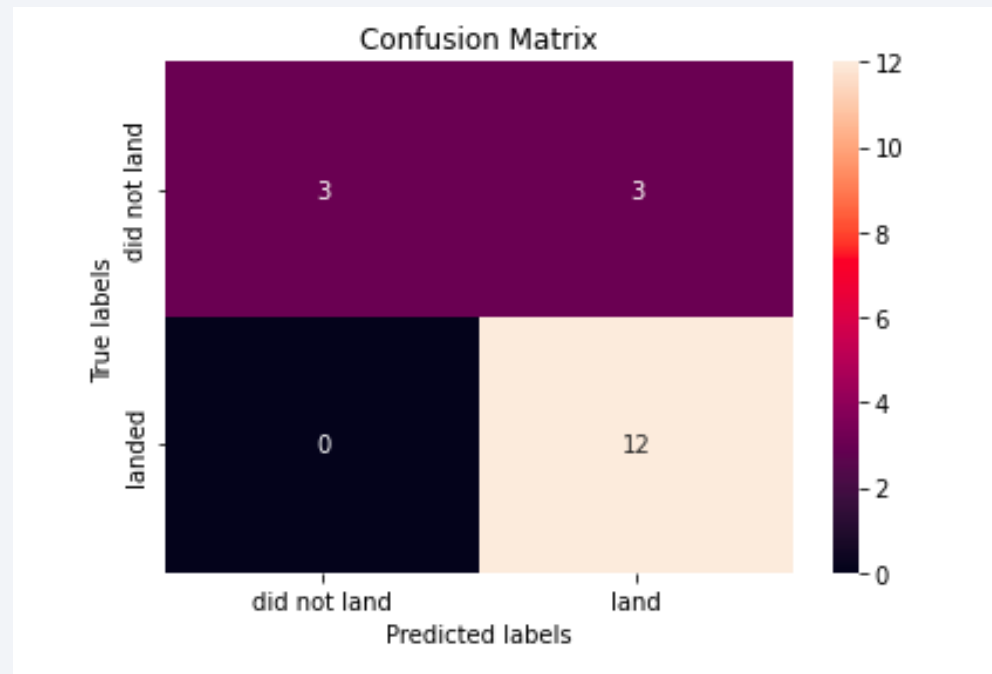
- Prepared and split data before undergoing training and testing using Logistic Regression, SVM, Decision Tree, and K-nearest neighbor.
- Source Code: <https://github.com/njudice/Data-Science-Capstone/blob/master/Machine%20Learning%20Prediction.ipynb>



# Results

---

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





The background of the slide is an abstract composition. It features a dark blue field on the left side, which transitions into a complex pattern of diagonal streaks in shades of blue, red, and teal on the right. These streaks have a textured, almost woven appearance. Overlaid on this pattern is a faint, light blue grid that recedes into the distance, creating a sense of depth and perspective.

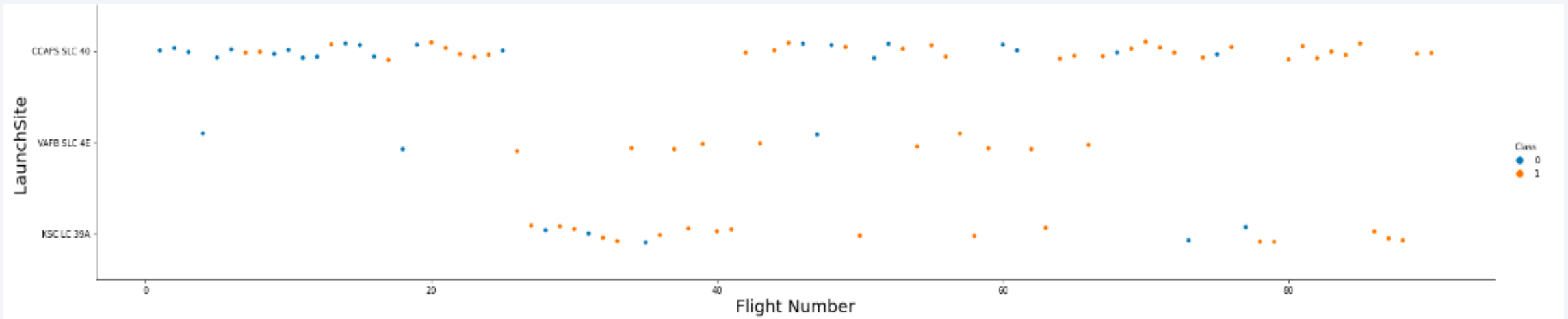
Section 2

# Insights drawn from EDA



# Flight Number vs. Launch Site

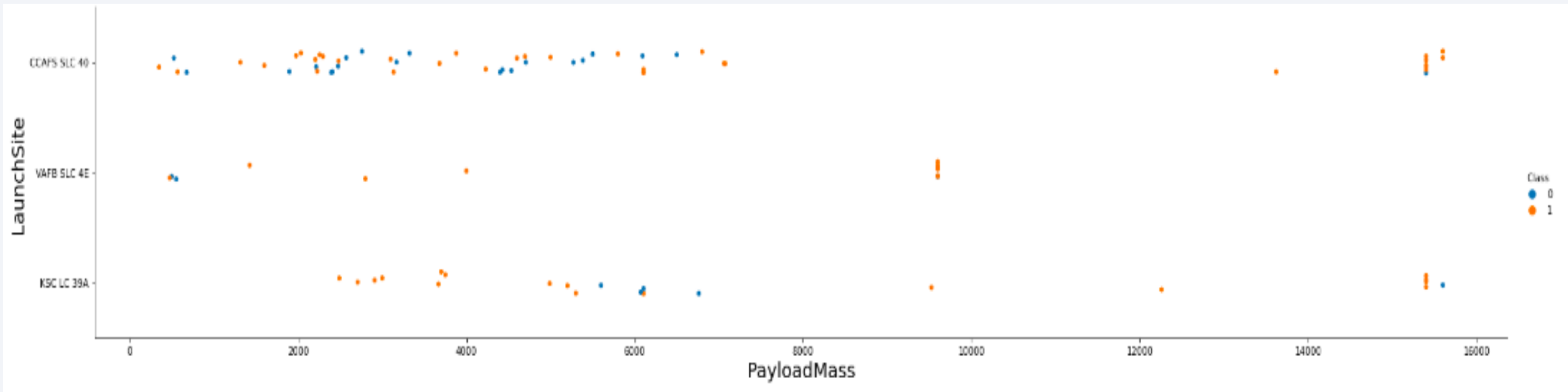
- Overall: as time goes on and more launches are performed (Flight Number going up), the success rate increases as seen by the stronger concentration of red towards the right side of the figure.



# Payload vs. Launch Site

---

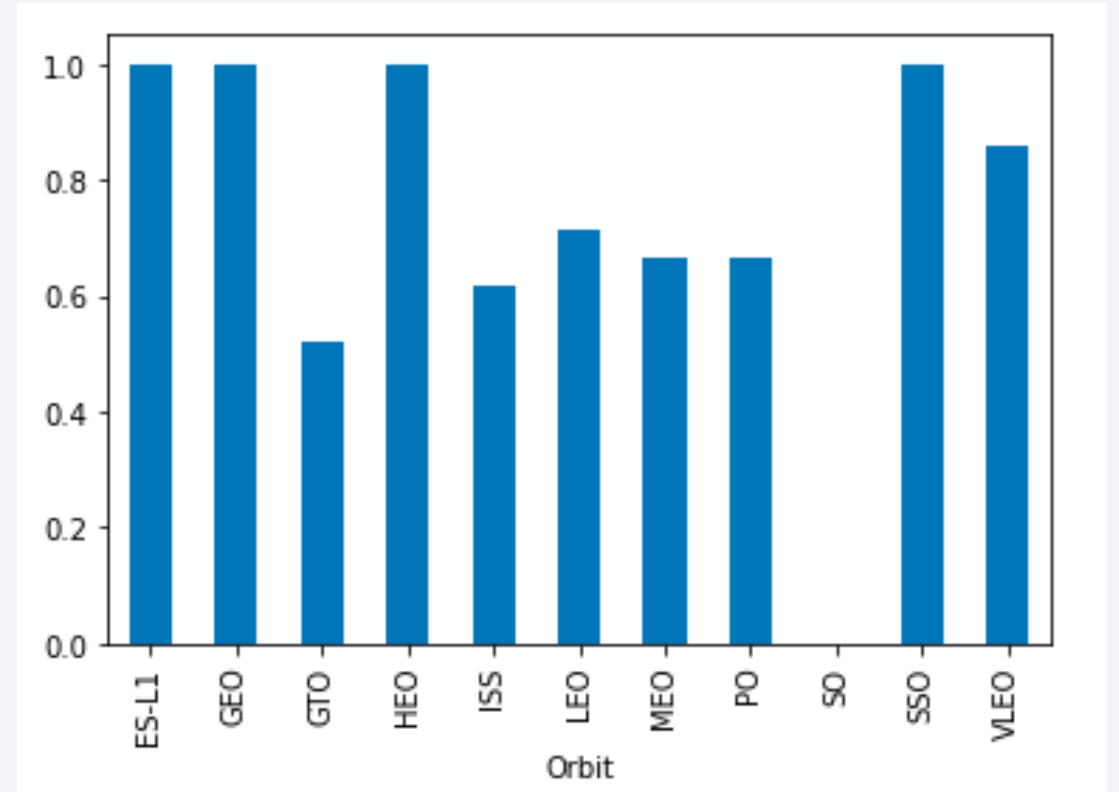
- A Payload over 8,000kg has a very strong success rate indicating more weight helps the rocket land no matter which Launch Site we use. This is true until we approach weights over 15,000kg.



# Success Rate vs. Orbit Type

---

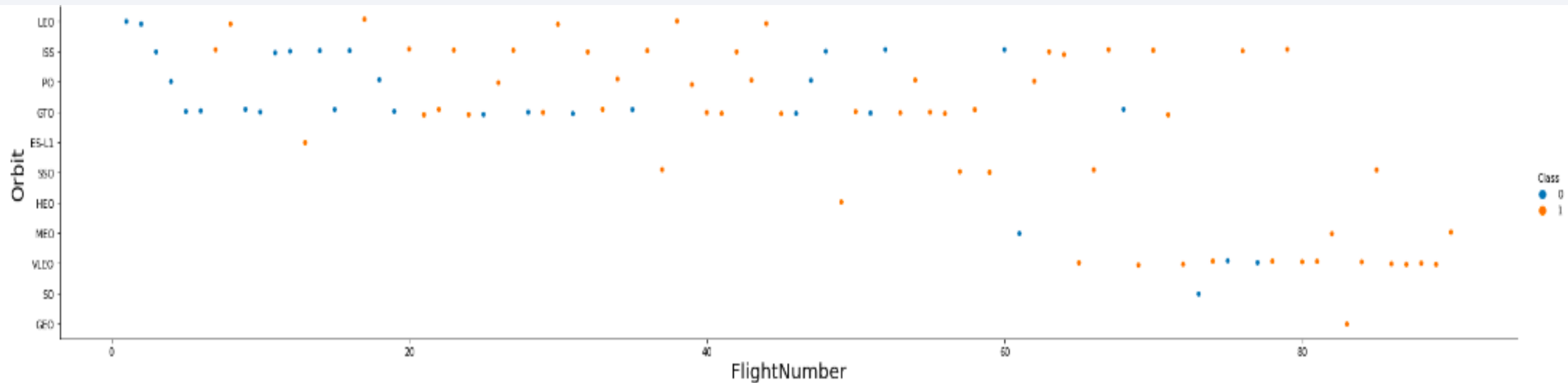
- Here we can see the average success rate broken down by orbit type. We see the greatest success with: ES-L1, GEO, HEO, and SSO.





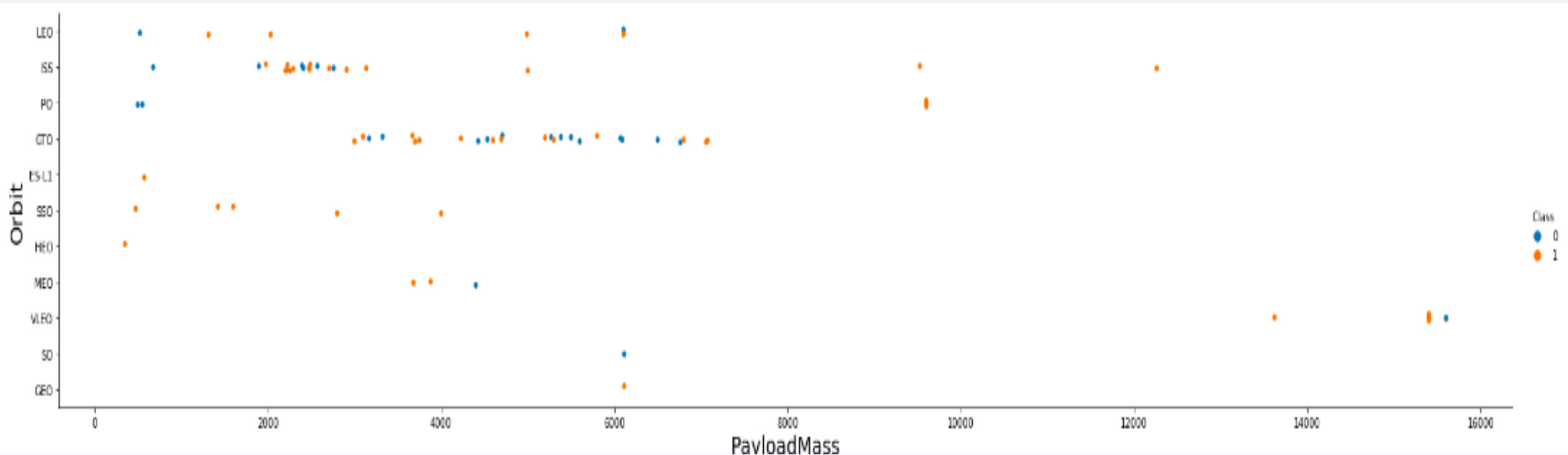
# Flight Number vs. Orbit Type

- Again, we see success rate improve as Flight Number increases, meaning that over time the proper adjustments are being made.
- There is a migration over time towards different Orbit Types, most recently VLEO.



# Payload vs. Orbit Type

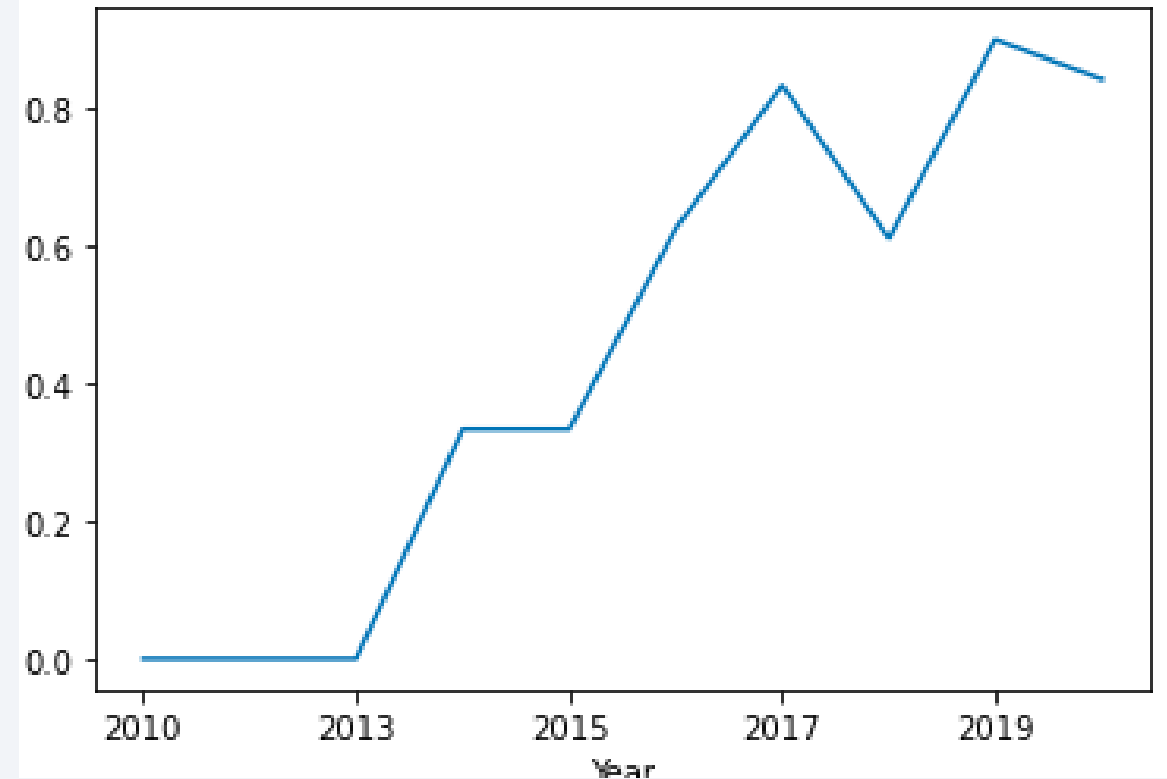
- Payload has a different effect on each orbit type. Some orbits, such as MEO and SSO, seem to be successful with a lower payload. However, LEO and ISS clearly perform better on a more consistent basis as you increase the payload.



# Launch Success Yearly Trend

---

- Over the long term, we see a general upward success trajectory as time passes.
- The Years 2018 and 2020 may have included some changes that were less than favorable as we see dips in these years.



# All Launch Site Names

---

- There are 4 unique Launch Sites: CCAFS LC-40, CCAFS SLC-40, VAFB SLC-4E, KSC LC-39A.

```
%sql Select distinct "Launch_Site" from SPACEXTBL;
```

```
* 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'

- Here are 5 Records with Launch Site names beginning with 'CCA'.
- Select \* from SPACEXTBL where Launch\_Site>= "CCA" limit 5;

| 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

---

- To find total payload we use the function sum() on the Payload\_Mass\_KG column.

```
%sql SELECT sum(PAYLOAD_MASS_KG_) as total_payload from SPACEXTBL where customer like 'NASA (CRS)%';
```

| total_payload |
|---------------|
|---------------|

|       |
|-------|
| 48213 |
|-------|



# Average Payload Mass by F9 v1.1

---

- We use the function avg() on the Payload\_Mass\_KG column and the 'where' and 'like' commands to limit results to F9 v1.1

```
%sql select avg(PAYLOAD_MASS_KG_) from SPACEXTBL where Booster_Version like 'F9 v1.1';
```

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

```
Done.
```

```
avg(PAYLOAD_MASS_KG_)
```

---

```
2928.4
```

# First Successful Ground Landing Date

---

- Using the min() function on the date column, we are able to gather the first year with the results of "Success (ground pad).

```
%sql select min(date) from SPACEXTBL where "Landing _Outcome" like "Success (ground pad)";
```

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

```
Done.
```

```
min(date)
```

```
01-05-2017
```

## Successful Drone Ship Landing with Payload between 4000 and 6000

---

- These Booster Versions have successfully performed drone ship landings with a payload between 4000-6000kg.

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

# Total Number of Successful and Failure Mission Outcomes

---

- From the table in the Mission Outcome Column we have a record count of 100 Successes and 1 Failure.

```
%sql select (select count("MISSION_OUTCOME") from SPACEXTBL where "MISSION_OUTCOME" like '%Success%') as Success, \
(select count("MISSION_OUTCOME") from SPACEXTBL where "MISSION_OUTCOME" like '%Failure%') as Failure
```

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

```
Done.
```

| Success | Failure |
|---------|---------|
| 100     | 1       |

# Boosters Carried Maximum Payload

---

- By using the max() function on the Payload\_Mass\_KG column we are able to obtain the max Payload possible. Then we are able to obtain the Booster Version that includes this max amount.

## 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

---

- Failed Landing outcomes in the year 2015, the month the occurred, Booster Version, and Launch Site

| MONTH | Booster_Version | Launch_Site |
|-------|-----------------|-------------|
| 01    | F9 v1.1 B1012   | CCAFS LC-40 |
| 04    | 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

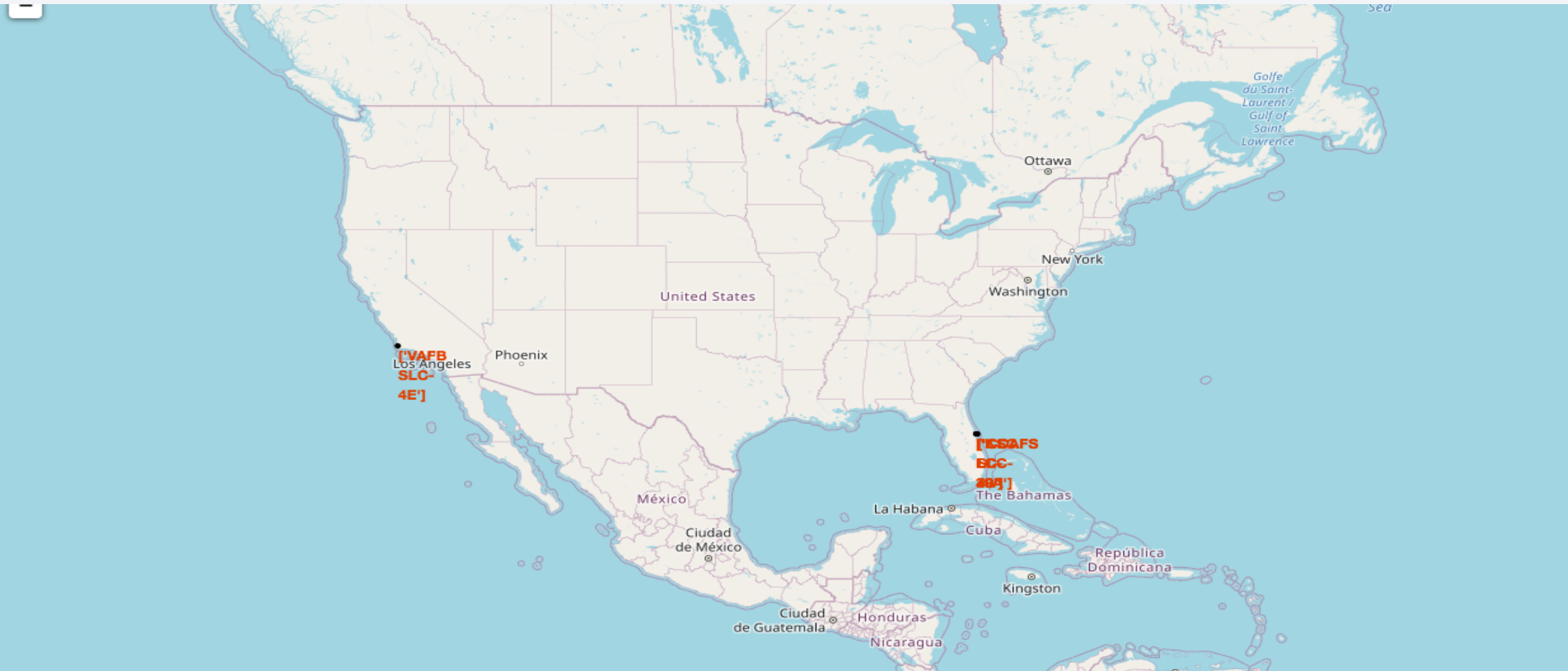
A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The image is a composite of a solid blue background on the left and a satellite photograph of Earth on the right. The Earth's surface is dark blue, with numerous bright yellow and orange lights representing cities and urban areas. The horizon line of the Earth is visible, separating the dark surface from the blackness of space.

Section 3

# Launch Sites Proximities Analysis

# Folium Map- Launch Sites

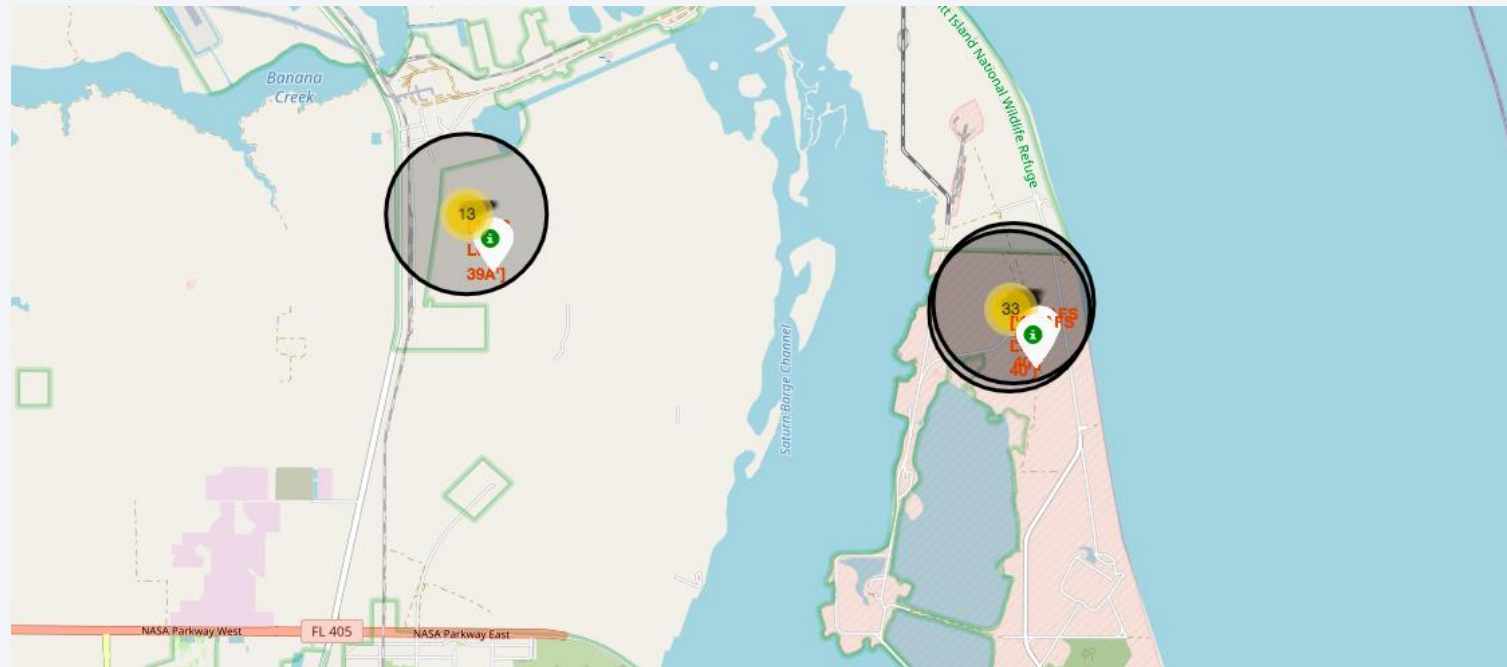
- The SpaceX Launch sites are shown here. There are located in Florida and California.




# Folium Map- Marker Clusters

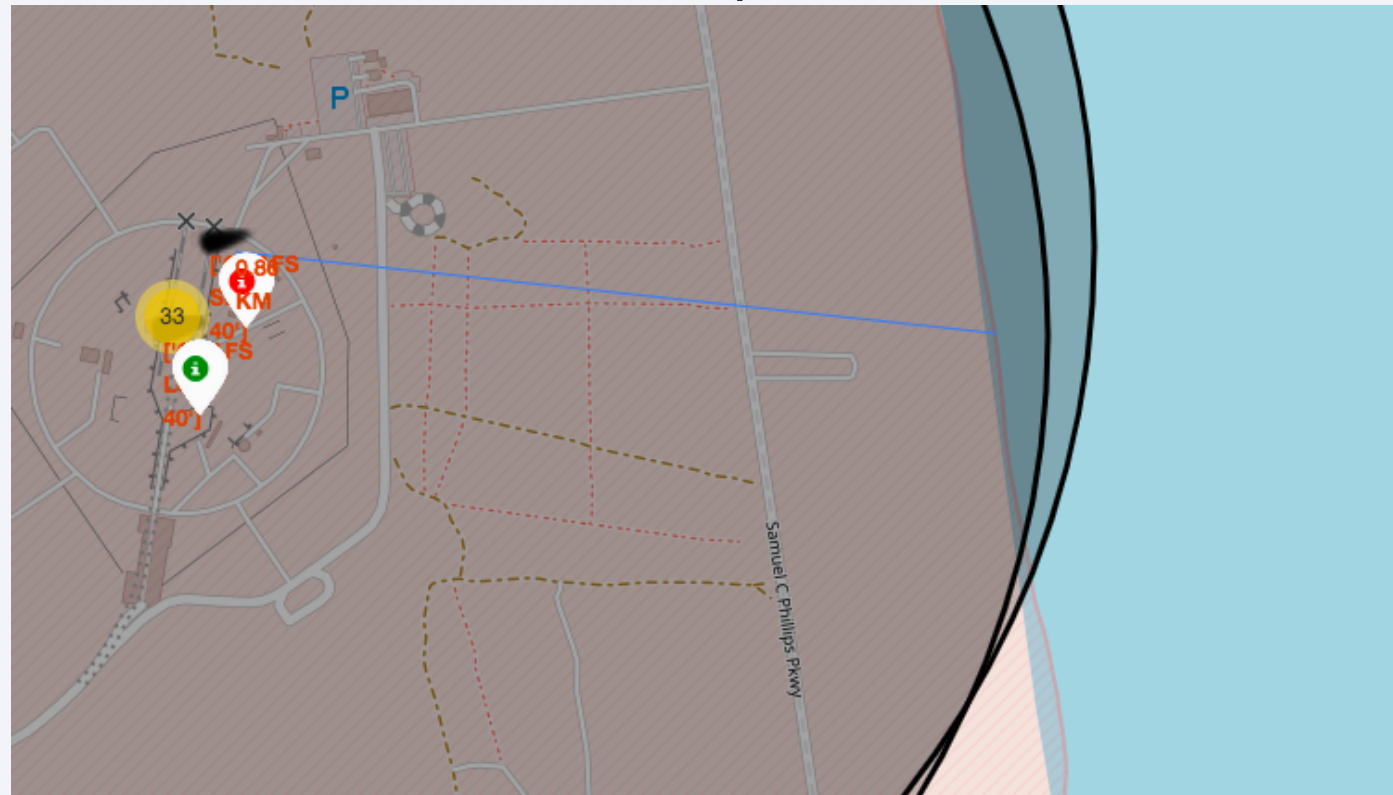
---

- Here we see clusters of Launches from a few of the Launch Sites. Successful launches are showcased in green, with Failed Launches in red.



# Folium Map- Distance Line

- We are now able to hover over a certain point with our Mouse to gather coordinates. This allows us to plot a line across the map which in turn helps us to measure distance between two places or landmarks.
- 





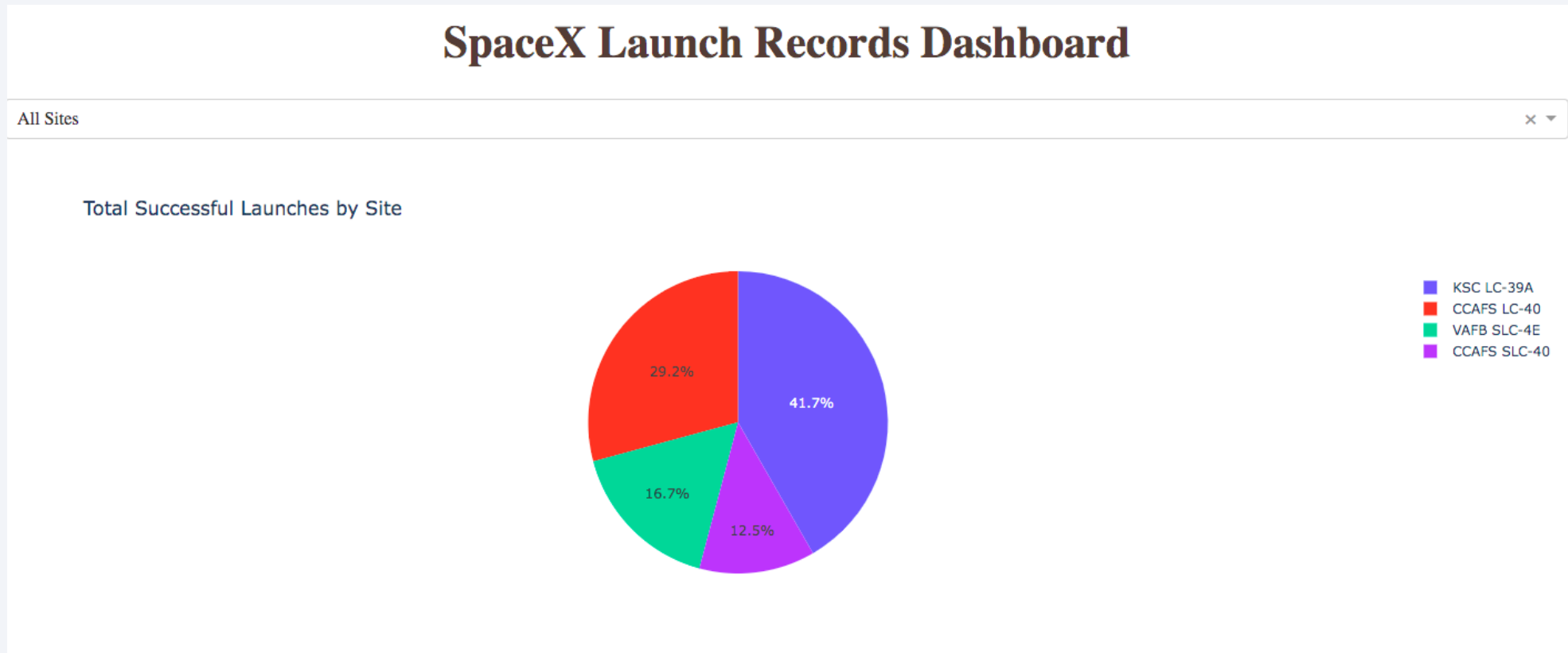


Section 4

# Build a Dashboard with Plotly Dash

# SpaceX Launch Records for All Sites- Plotly

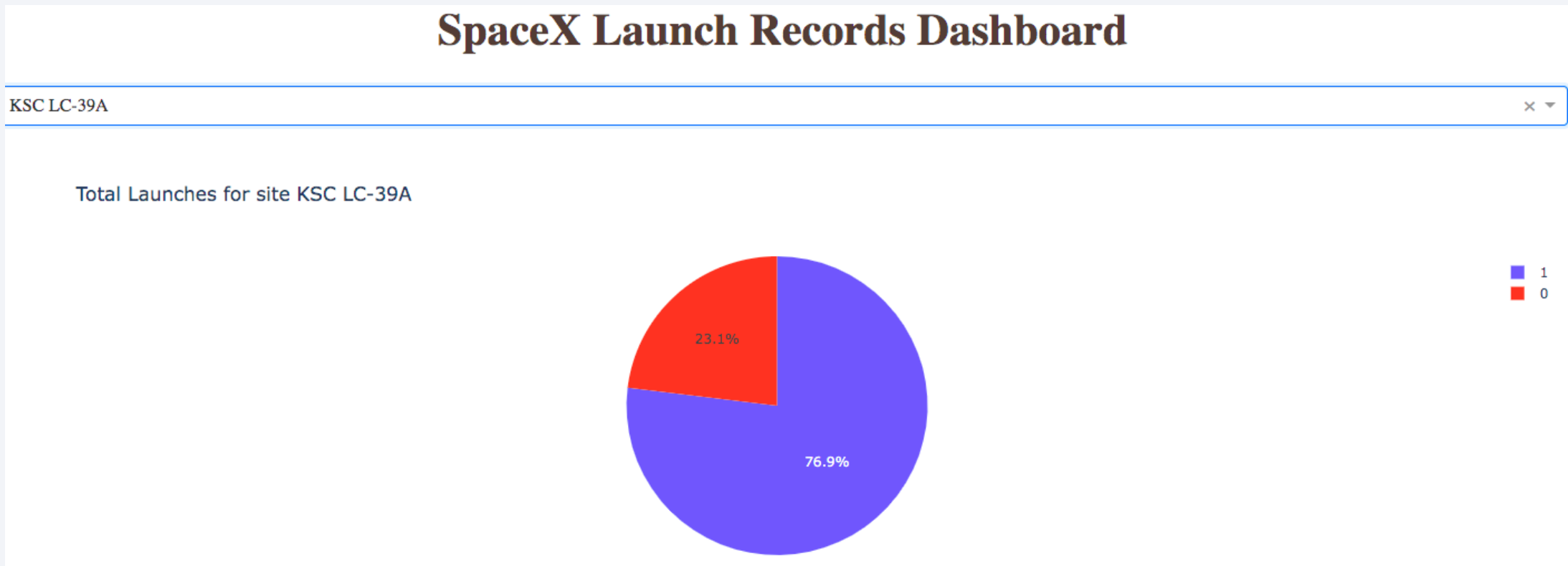
- Here we have a pie chart which compares the various Launch sites by their success rates. We see here that KSC LC-39A has the highest success rate.



# SpaceX Plotly Success by Launch Site

---

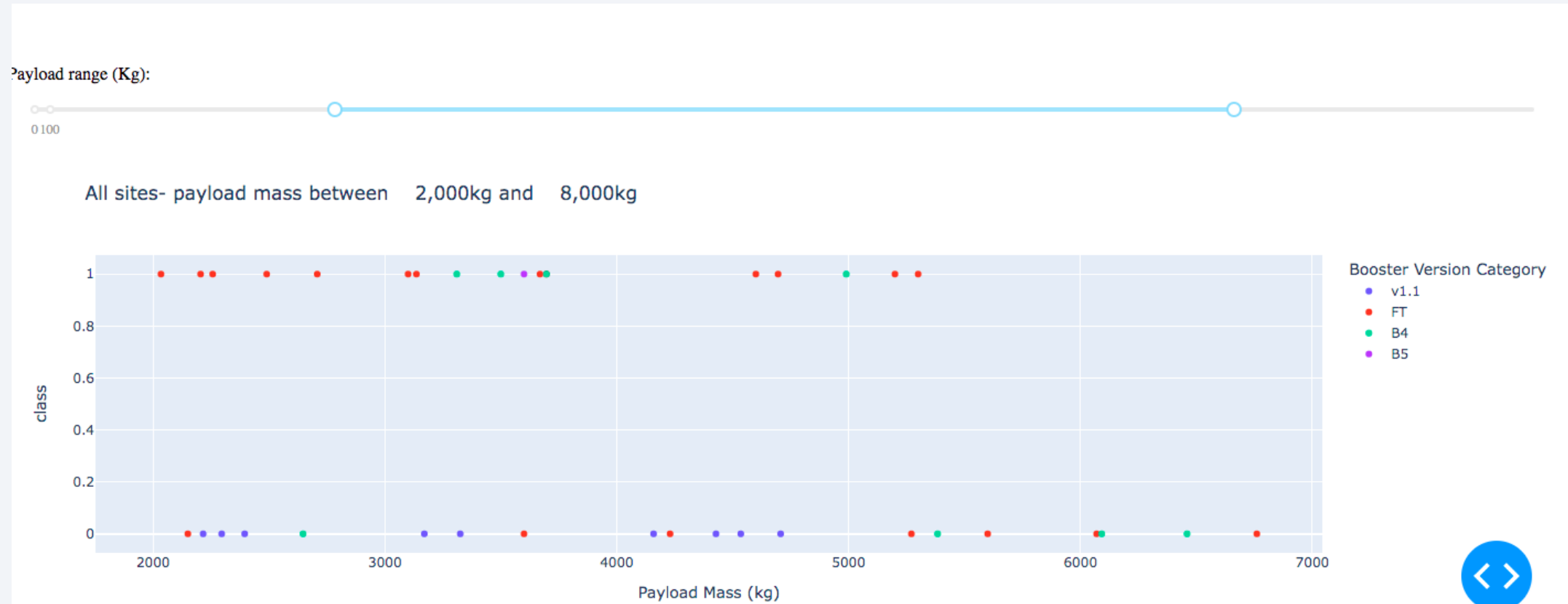
- Here we gain a closer look at the success rate of the KSC LC-39A site.





# SpaceX Plotly- Success rate by Payload

- The scatter plot with an interactive sliding bar allows us to see success rate among all sites when different payload ranges are selected.





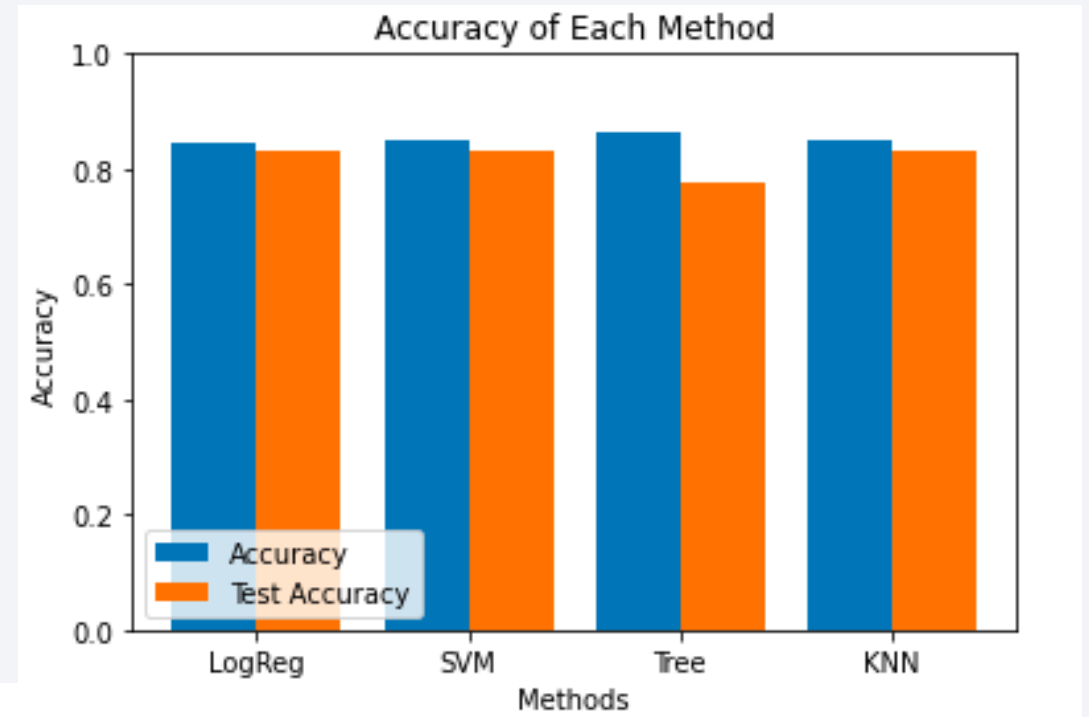
Section 5

# Predictive Analysis (Classification)

# Classification Accuracy

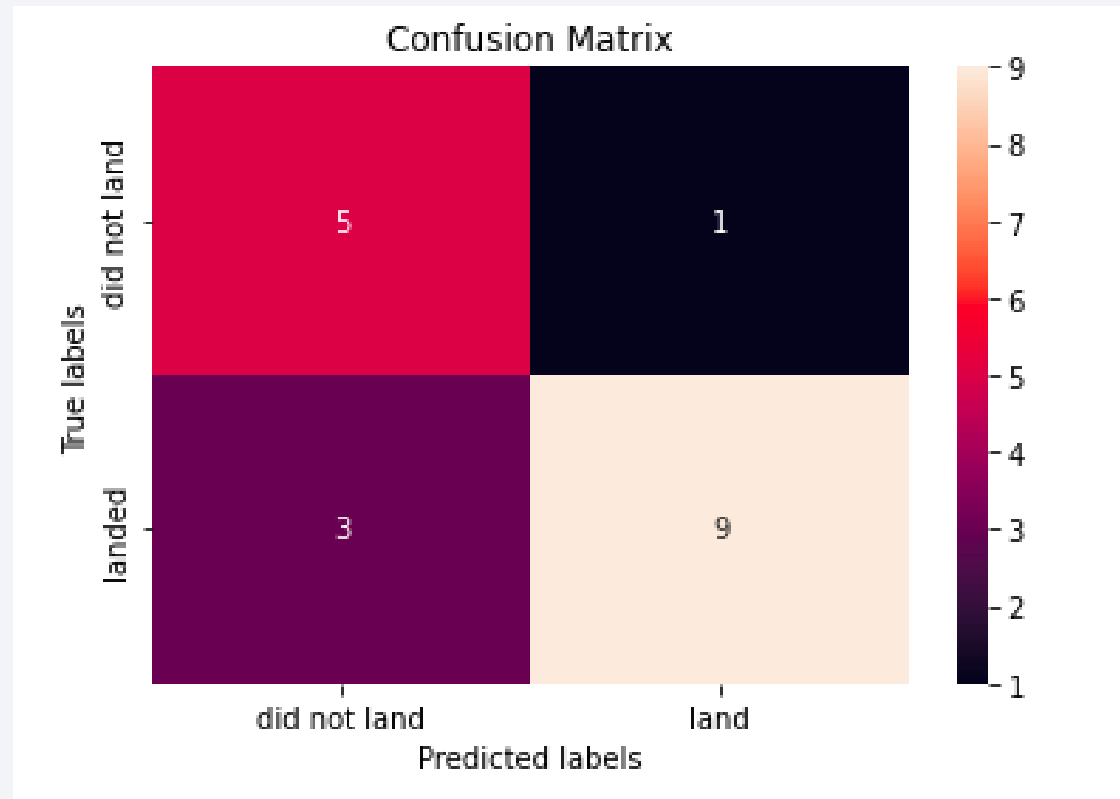
- Side-by-side comparison of Accuracy vs Test Accuracy for each classification model.
- The Decision Tree Classifier gives the highest Accuracy.

| Model  | Accuracy | TestAccuracy |
|--------|----------|--------------|
| LogReg | 0.84643  | 0.83333      |
| SVM    | 0.84821  | 0.83333      |
| Tree   | 0.86429  | 0.77778      |
| KNN    | 0.84821  | 0.83333      |



# Confusion Matrix

- Decision Tree confusion matrix. The top-left and bottom-right regions are the True Positive and True Negative regions. This confusion matrix contains the highest values in these areas.



# Conclusions

---

- The most important and obvious conclusion from the data is that over time, we are having higher concentrations of success.
- KSC LC-39A is the Launch Site with the highest success rate.
- A higher Payload Mass seems to be beneficial in most cases.
- The Decision Tree Classifier returned the best training results even though the data in each classifier was identical.

# Appendix

---

- All records of procedures can be viewed here: <https://github.com/njudice/Data-Science-Capstone/tree/master>



Thank you!

