



IBM Developer  
SKILLS NETWORK

# Winning Space Race with Data Science

Hotien Huang  
2024.05.28



# Outline

---

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

# Executive Summary

---

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

# Introduction

---

- Project background

- Space X has best pricing (\$62million, others \$165million USD)
- Largely due to ability to recover part of rocket (Stage 1)
- Space Y company wants to compete with Space X

- Problems

- In order to reuse the part of rocket, Space Y has to training model to analysis the key factor lead success



Section 1

# Methodology

# Methodology

---

## Executive Summary

- Data collection methodology:
  - SpaceX Data collected from 2 resources
    - SpaceX API (<https://api.spacexdata.com/v4/rockets/>)
    - Web Scraping (Wikipedia)
- Perform data wrangling
  - Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features

# Methodology

## Executive Summary

- 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 that was collected until this step were normalized, divided in training and test data sets and evaluated by four different classification models, being the accuracy of each model evaluated using different combinations of parameters.

# Data Collection

---

## How data sets were collected

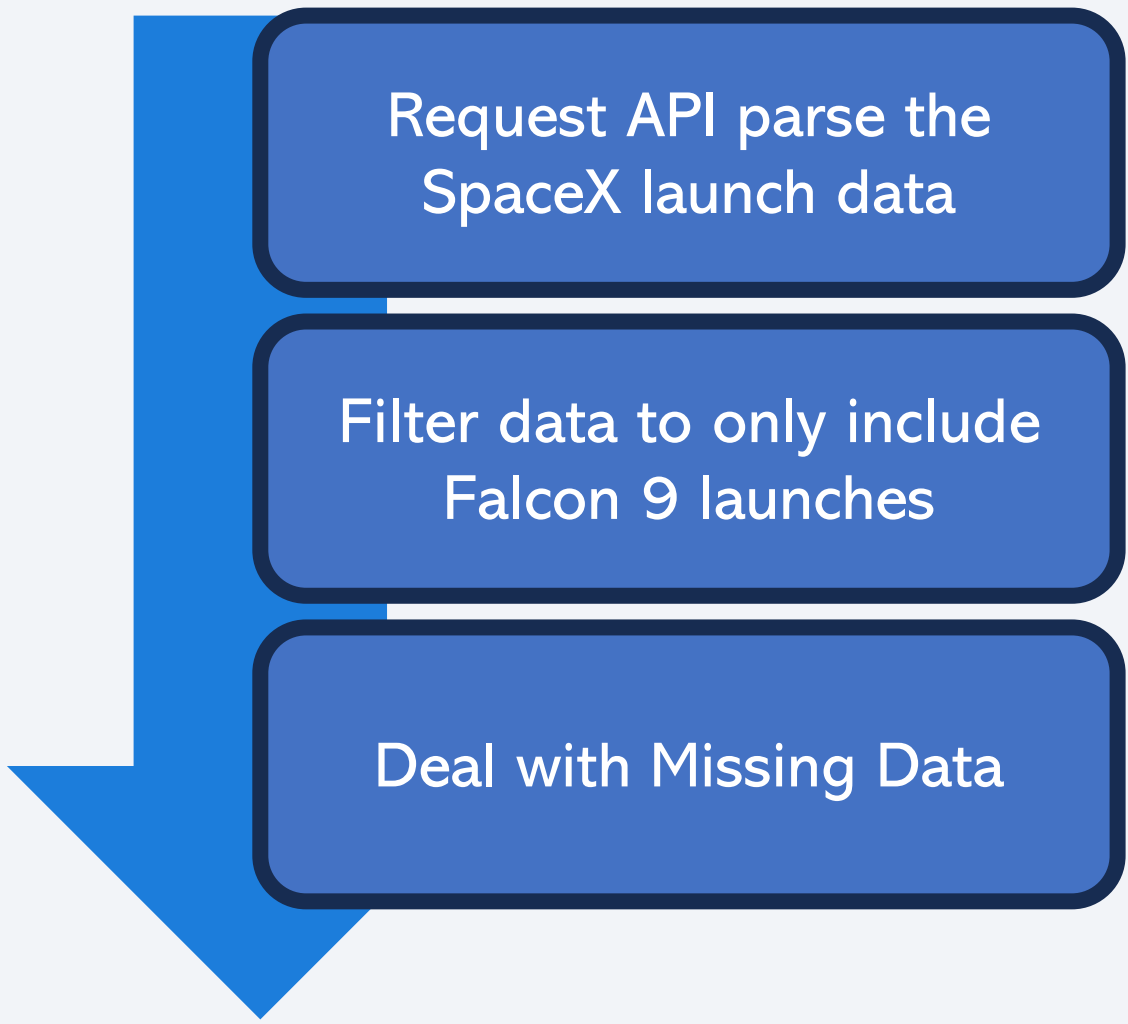
- Datasets were collected from SpaceX API  
(<https://api.spacexdata.com/v4/rockets/>)
- Wikipedia  
([https://en.wikipedia.org/wiki/List\\_of\\_Falcon/\\_9/\\_and\\_Falcon\\_Heavy\\_launches](https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)),  
using web scraping technics.



# Data Collection – SpaceX API

---

- SpaceX offers a public API from where data can be obtained and then used;
- This API was used according to the flowchart beside and then data is persisted.
- GitHub URL of the completed SpaceX API calls notebook :  
[https://github.com/TonyHuang1688/Course10\\_Final\\_Assignment/blob/31ffd71676cef768ca95400d042e92de2859bdf3/Week1%20spacex-data-collection-api.ipynb](https://github.com/TonyHuang1688/Course10_Final_Assignment/blob/31ffd71676cef768ca95400d042e92de2859bdf3/Week1%20spacex-data-collection-api.ipynb)



Request API parse the SpaceX launch data

Filter data to only include Falcon 9 launches

Deal with Missing Data

# Data Collection - Scraping

---

- Data from SpaceX launches can also be obtained from Wikipedia;
- Data are downloaded from Wikipedia according to the flowchart and then persisted.
- GitHub URL of the completed web scraping notebook :  
[https://github.com/TonyHuang1688/Course10\\_Final\\_Assignment/blob/31ffd71676cef768ca95400d042e92de2859bdf3/Week1%20spacex-data-collection-api.ipynb](https://github.com/TonyHuang1688/Course10_Final_Assignment/blob/31ffd71676cef768ca95400d042e92de2859bdf3/Week1%20spacex-data-collection-api.ipynb)



Request the Falcon9 Launch Wiki page

Extract all column/variable names from HTML table header

Create a data frame by parsing the launch HTML tables

# Data Wrangling

---

- Initially some Exploratory Data Analysis (EDA) was performed on the dataset.
- Then the summary launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated.
- Finally, the landing outcome label was created from Outcome column.
- GitHub URL of the completed data wrangling related notebooks :  
[https://github.com/TonyHuang1688/Course10\\_Final\\_Assignment/blob/31ffd71676cef768ca95400d042e92de2859bdf3/Week1%20spacex-Data%20wrangling.ipynb](https://github.com/TonyHuang1688/Course10_Final_Assignment/blob/31ffd71676cef768ca95400d042e92de2859bdf3/Week1%20spacex-Data%20wrangling.ipynb)



Exploratory Data Analysis(EDA)

Summarizations

Creation of Landing Outcome Label

# EDA with Data Visualization

---

- Scatter Point Chart : It can intuitively display the relationship between two variables, especially numerical variables.

TASK 1: Visualize the relationship between Flight Number and Launch Site

TASK 2: Visualize the relationship between Payload and Launch Site

TASK 4: Visualize the relationship between FlightNumber and Orbit type

TASK 5: Visualize the relationship between Payload and Orbit type

- Bar Chart : Suitable for displaying comparisons between different categories, clearly showing the differences in values.

TASK 3: Visualize the relationship between success rate of each orbit type

- Line Plot : Suitable for displaying time series data or trends over time.

TASK 6: Visualize the launch success yearly trend

- Categorical Columns : Can be combined with various chart types such as bar charts and pie charts.

TASK 7: Create dummy variables to categorical columns

TASK 8: Cast all numeric columns to float64

- GitHub URL of the completed EDA with data visualization notebook :

[https://github.com/TonyHuang1688/Course10\\_Final\\_Assignment/blob/31ffd71676cef768ca95400d042e92de2859bdf3/Week2%20edadataviz.ipynb](https://github.com/TonyHuang1688/Course10_Final_Assignment/blob/31ffd71676cef768ca95400d042e92de2859bdf3/Week2%20edadataviz.ipynb)

# EDA with SQL

---

- To explore data, scatterplots and bar plots were used to visualize the relationship between pair of features:
- Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, Payload and Orbit
- GitHub URL of the completed EDA with SQL notebook :  
[https://github.com/TonyHuang1688/Course10\\_Final\\_Assignment/blob/35e77e90ab3f14b649d68258a4215f8d1150ba92/Week2%20eda-sql-coursera\\_sqlite.ipynb](https://github.com/TonyHuang1688/Course10_Final_Assignment/blob/35e77e90ab3f14b649d68258a4215f8d1150ba92/Week2%20eda-sql-coursera_sqlite.ipynb)



# Build an Interactive Map with Folium

---

- Markers, circles, lines and marker clusters were used with Folium Maps
- Markers indicate points like launch sites;
- Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center;
- Marker clusters indicates groups of events in each coordinate, like launches in a launch site; and
- Lines are used to indicate distances between two coordinates.
- GitHub URL of the completed interactive map with Folium map :  
[https://github.com/TonyHuang1688/Course10\\_Final\\_Assignment/blob/35e77e90ab3f14b649d68258a4215f8d1150ba92/Week3%20launch\\_site\\_location.ipynb](https://github.com/TonyHuang1688/Course10_Final_Assignment/blob/35e77e90ab3f14b649d68258a4215f8d1150ba92/Week3%20launch_site_location.ipynb)

# Build a Dashboard with Plotly Dash

---

- We built an interactive dashboard with Plotly dash
- We plotted pie charts showing the total launches by a certain sites
- We plotted scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster version.
- GitHub URL of the completed Plotly Dash lab :  
[https://github.com/TonyHuang1688/Course10\\_Final\\_Assignment/blob/35e77e90ab3f14b649d68258a4215f8d1150ba92/Week3%20spacex\\_dash\\_app.py](https://github.com/TonyHuang1688/Course10_Final_Assignment/blob/35e77e90ab3f14b649d68258a4215f8d1150ba92/Week3%20spacex_dash_app.py)

# Predictive Analysis (Classification)

---

- We loaded the data using numpy and pandas, transformed the data, split our data into training and testing.
- We built different machine learning models and tune different hyperparameters using GridSearchCV.
- We used accuracy as the metric for our model, improved the model using feature engineering and algorithm tuning.
- We found the best performing classification model.
- GitHub URL of the completed predictive analysis lab :  
[https://github.com/TonyHuang1688/Course10\\_Final\\_Assignment/blob/35e77e90ab3f14b649d68258a4215f8d1150ba92/Week4%20SpaceX\\_Machine%20Learning%20Prediction.ipynb](https://github.com/TonyHuang1688/Course10_Final_Assignment/blob/35e77e90ab3f14b649d68258a4215f8d1150ba92/Week4%20SpaceX_Machine%20Learning%20Prediction.ipynb)

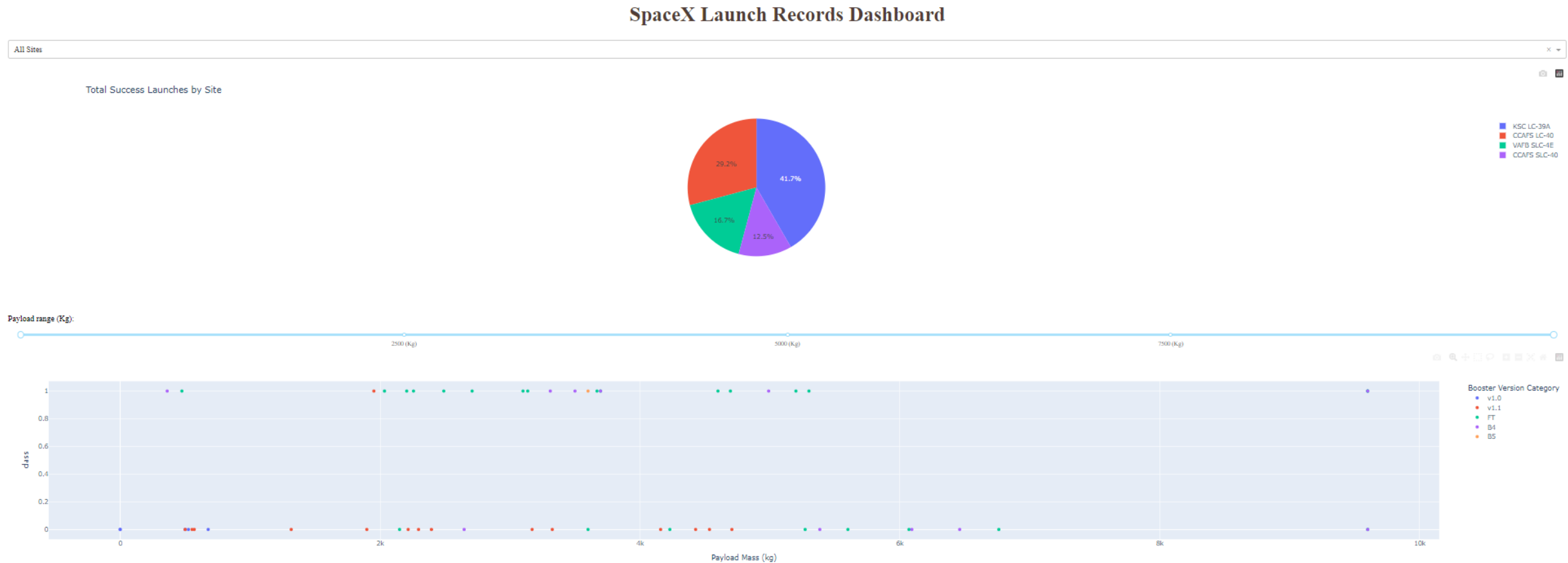
# Results

---

- Exploratory data analysis results

- Space X uses 4 different launch sites;
- The first launches were done to Space X itself and NASA;
- The average payload of F9 v1.1 booster is 2,928 kg;
- The first success landing outcome happened in 2015 five year after the first launch;
- Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average;
- Almost 100% of mission outcomes were successful;
- Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015;
- The number of landing outcomes became as better as years passed.

# Results



- This is a preview of the Plotly dashboard. The following slides will show the results of EDA with visualization, EDA with SQL, Interactive Map with Folium, and finally the results of our model with about 83% accuracy.



The background of the slide is an abstract composition. It features a dark blue base color. Overlaid on this are numerous diagonal streaks in shades of red and cyan. A faint, light blue grid pattern is also visible, particularly in the lower-left quadrant. The overall effect is dynamic and technological.

Section 2

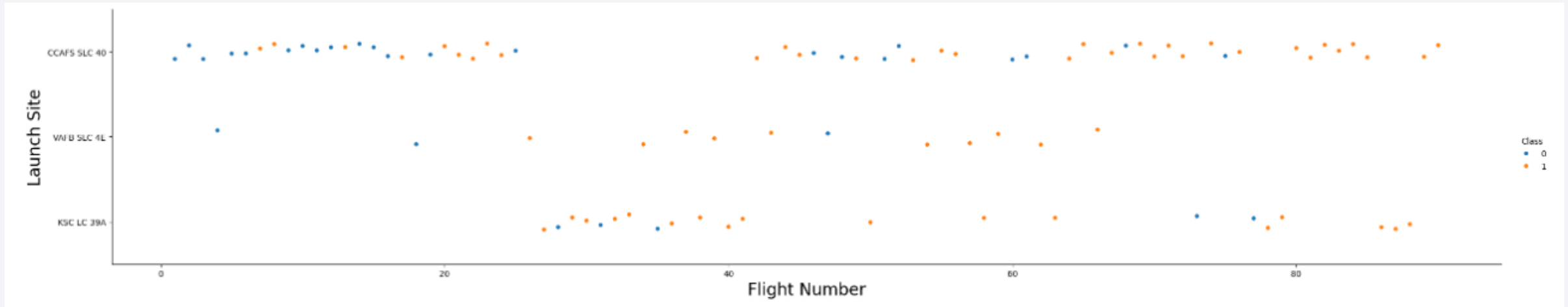
# Insights drawn from EDA



# Flight Number vs. Launch Site

---

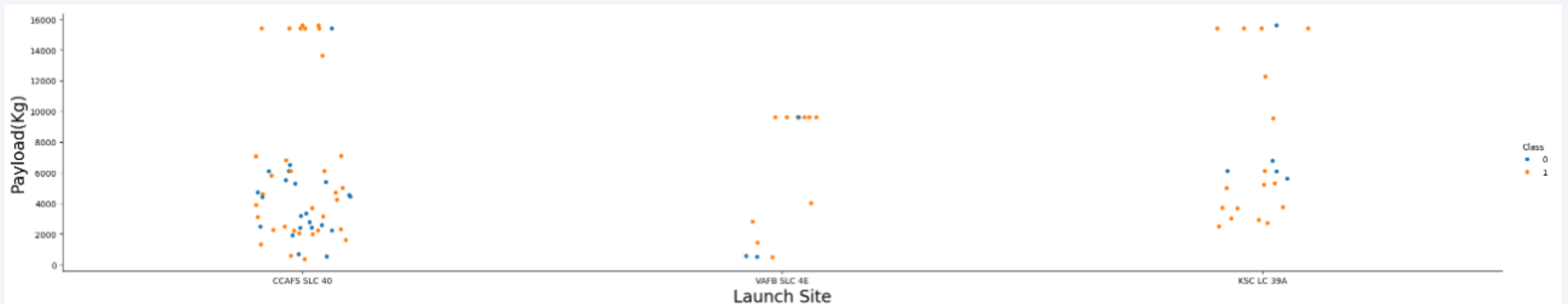
- From the plot, we found that the larger the flight amount at a launch site, the greater the success rate at a launch site.



# Payload vs. Launch Site

---

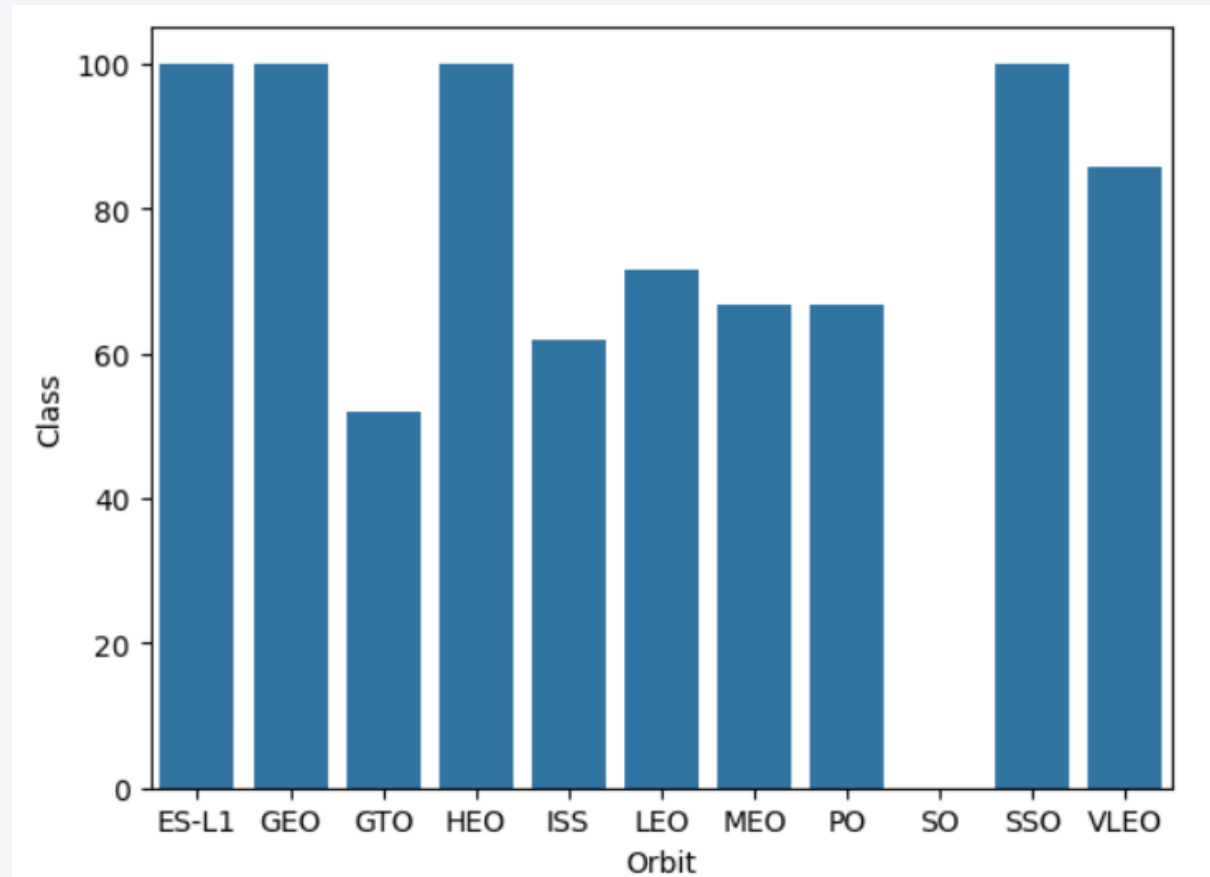
- The greater the payload mass for launch site CCAFS SLC 40 the higher the success rate for the rocket.



# Success Rate vs. Orbit Type

---

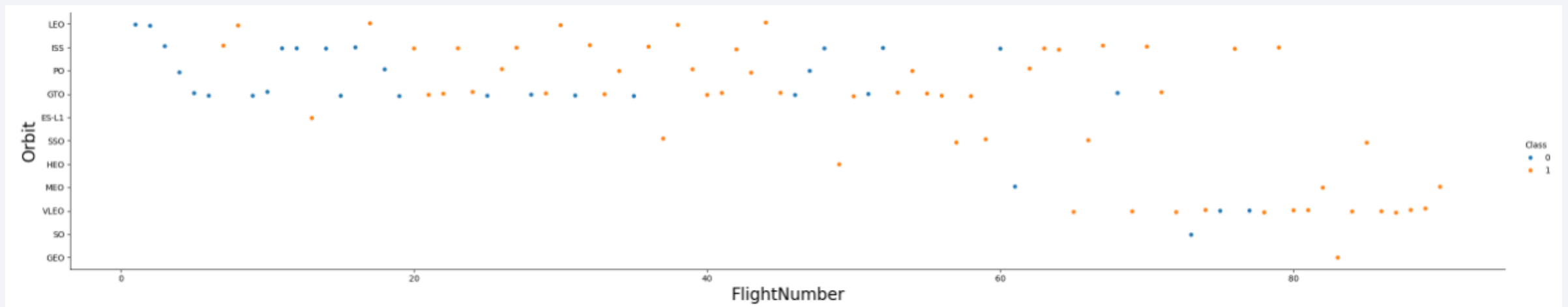
- From the plot, we can see that ES-L1, GEO, HEO, SSO, VLEO had the most rate.



# Flight Number vs. Orbit Type

---

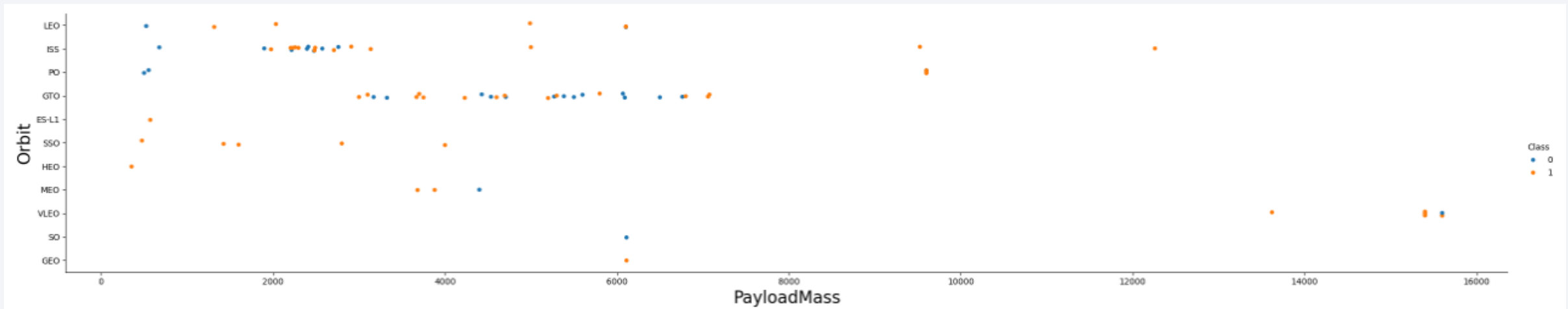
- The plot below shows the Flight Number vs. Orbit type. We observe that in the LEO orbit, success is related to the number of flights whereas in the GTO orbit, there is no relationship between flight number and the orbit.





# Payload vs. Orbit Type

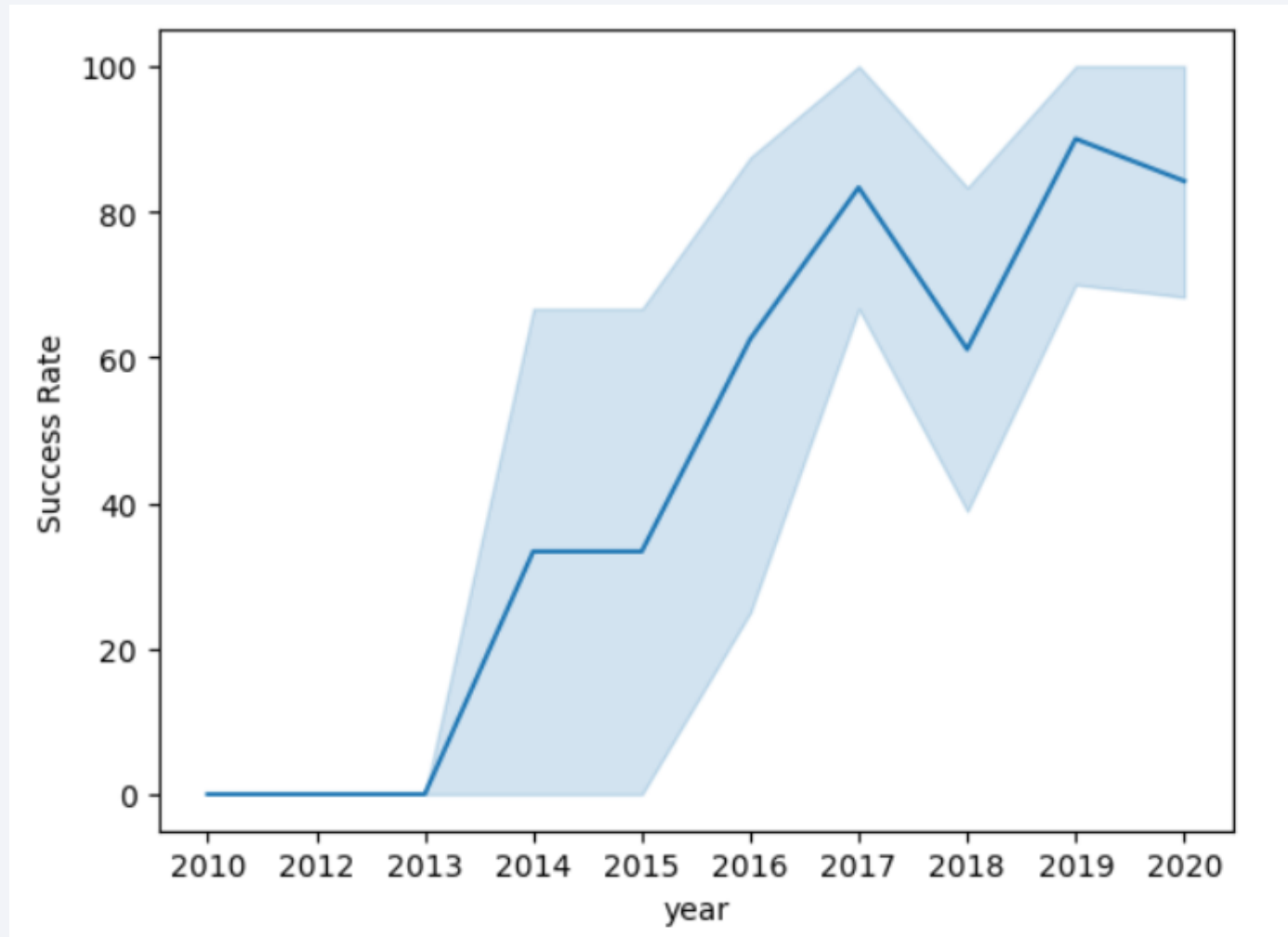
- We can observe that with heavy payloads, the successful landing are more for PO, LEO and ISS orbits.



# Launch Success Yearly Trend

---

- From the plot, we can observe that success rate since 2013 kept on increasing till 2020.



# All Launch Site Names

---

- We used the key word DISTINCT to show only unique launch sites from the SpaceX data.

```
In [8]: %sql SELECT Distinct LAUNCH_SITE FROM SPACEXTBL
```

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

```
Done.
```

```
Out[8]: Launch_Site
```

```
CCAFS LC-40
```

```
VAFB SLC-4E
```

```
KSC LC-39A
```

```
CCAFS SLC-40
```

# Launch Site Names Begin with 'CCA'

- We used the query above to display 5 records where launch sites like 'CCA%'

```
In [9]: %sql SELECT* FROM SPACEXTBL WHERE launch_site like "CCA%" limit 5
```

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

```
Out[9]:
```

| Date       | Time (UTC) | Booster_Version | Launch_Site | Payload   | PAYLOAD_MASS_KG_ | Orbit     | Customer        | Mission_Outcome | Landing_Outc        |
|------------|------------|-----------------|-------------|---|------------------|-----------|-----------------|-----------------|---------------------|
| 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

---

- We calculated the total payload carried by boosters from NASA as 45596 using the query below

```
In [11]: %sql SELECT sum(payload_mass__kg_) as sum from SPACEXTBL where customer like 'NASA (CRS)'
```

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

```
Out[11]:  sum  
         sum  
         45596
```



# Average Payload Mass by F9 v1.1

---

- We calculated the average payload mass carried by booster version F9 v1.1 as 2928.4

```
In [15]: %sql select avg(payload_mass__kg_) as Average from SPACEXTBL where booster_version = 'F9 v1.1'
```

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

```
Done.
```

```
Out[15]: Average
```

```
2928.4
```

# First Successful Ground Landing Date

---

- We observed that the dates of the first successful landing outcome on ground pad was 22nd December 2015

```
[31]: %sql SELECT min(Date) AS First_Success FROM SPACEXTBL WHERE Landing_Outcome like 'Success (ground pad)'
```

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

```
Done.
```

```
[31]: First_Success
```

| <b>First_Success</b> |
|----------------------|
| 2015-12-22           |

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

---

- We used the WHERE clause to filter for boosters which have successfully landed on drone ship and applied the AND condition to determine successful landing with payload mass greater than 4000 but less than 6000

```
[17]: %sql SELECT Booster_Version From SPACEXTBL WHERE (Mission_Outcome like 'success') AND (payload_mass__kg_ BETWEEN 4000 AND 6000) AND (landing_outcome = 'successful')
```

\* sqlite:///my\_data1.db  
Done.

```
[17]: Booster_Version
```

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

# Total Number of Successful and Failure Mission Outcomes

---

- We used “Count” and “Group BY” to filter for WHERE MissionOutcome was a success or a failure.

```
[18]: %sql SELECT mission_outcome, count(*) FROM SPACEXTBL GROUP BY mission_outcome ORDER BY mission_outcome
```

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

```
Done.
```

```
[18]:
```

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

# Boosters Carried Maximum Payload

- We determined the booster that have carried the maximum payload using a subquery in the WHERE clause and the MAX() function.

```
[19]: maxm = %sql select max(payload_mass__kg_) from SPACEXTBL
      maxv = maxm[0][0]
      %sql select booster_version from SPACEXTBL where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTBL)

      * sqlite:///my_data1.db
      Done.
      * sqlite:///my_data1.db
      Done.
[19]: 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

---

- We used a combinations of the WHERE clause, LIKE, AND, and BETWEEN conditions to filter for failed landing outcomes in drone ship, their booster versions, and launch site names for year 2015

```
[20]: %sql select substr(Date, 6,2) as Month, Landing_Outcome, booster_version, launch_site from SPACEXTBL where DATE like '2015%' AND Landing_Outcome
```

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

```
Done.
```

```
[20]:
```

| Month | 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

---

- We selected Landing outcomes and the COUNT of landing outcomes from the data and used the WHERE clause to filter for landing outcomes BETWEEN 2010-06-04 to 2017-03-20.
- We applied the GROUP BY clause to group the landing outcomes and the ORDER BY clause to order the grouped landing outcome in descending order.

```
[21]: %sql select Landing_Outcome, count(*) as Count from SPACEXTBL where Date >= '2010-06-04' AND Date <= '2017-03-20' GROUP BY Landing_Outcome ORDER BY Count DESC
```

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

```
Done.
```

```
[21]:
```

| 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     |

A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The background is a deep blue gradient.

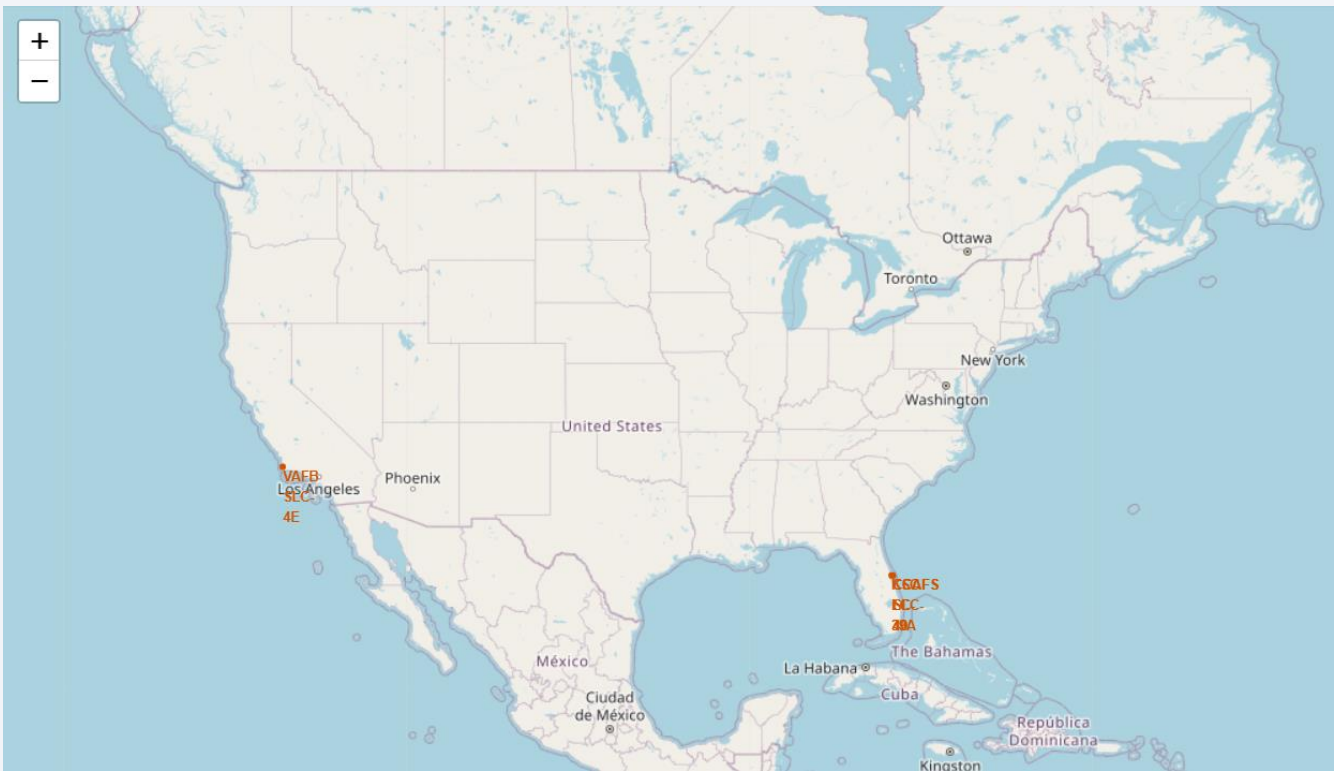
Section 3

# Launch Sites Proximities Analysis



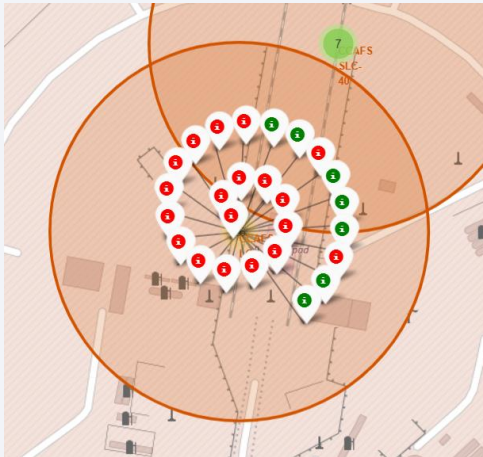
# Launch Site Locations

- We can see the launch sites are located in the coast of America

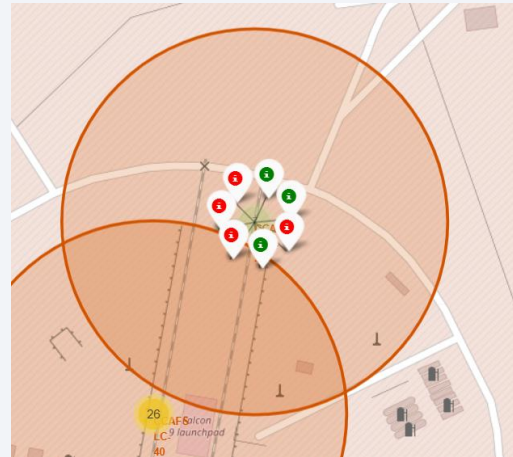


# Markers showing launch sites with color labels

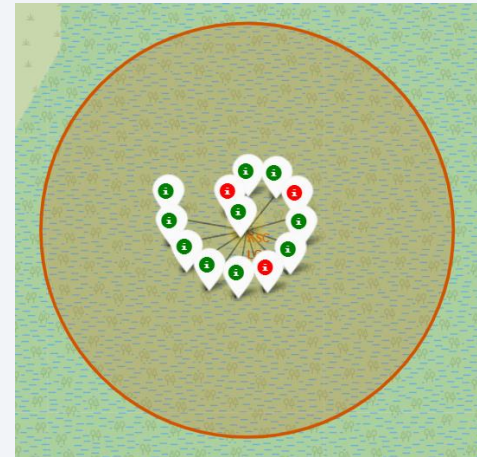
- Green Marker shows successful launches, and red marker shows failure.



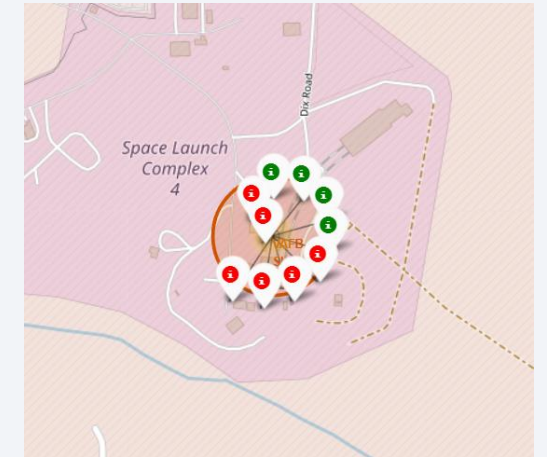
CCAFS LC-40



CCAFS SLC-40



KSC LC-39A



VAFB SLC-4E

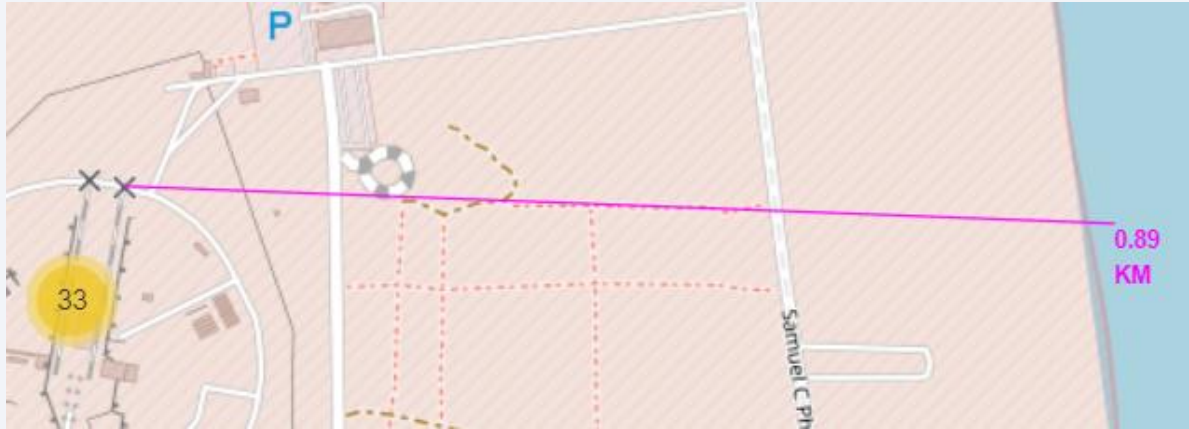
**FLORIDA**

**California**

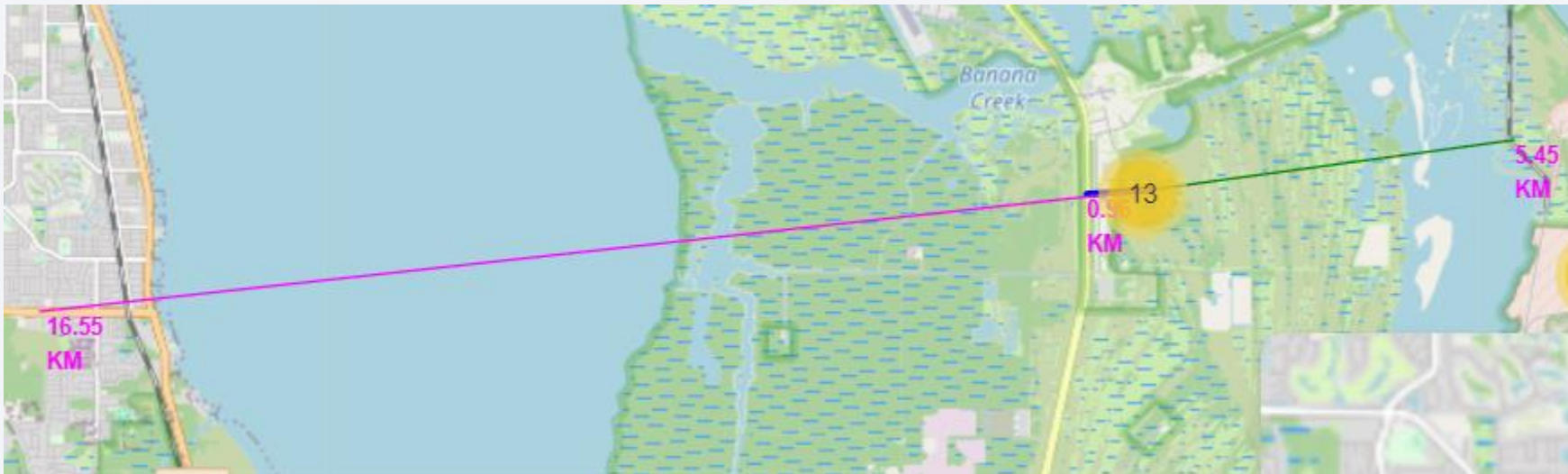


# Launch Site distance to landmarks

To the coast : 0.89km



To highway : 0.96km



To the city : 16.55km

To railway : 5.45km

The background of the slide is a close-up, artistic photograph of a printed circuit board (PCB). The board is dark, and the intricate circuit traces are highlighted in a vibrant, glowing red. Numerous small, circular components, likely solder joints or micro-components, are visible along the traces, some of which also appear to be glowing. The overall effect is a high-tech, digital aesthetic.

Section 4

# Build a Dashboard with Plotly Dash

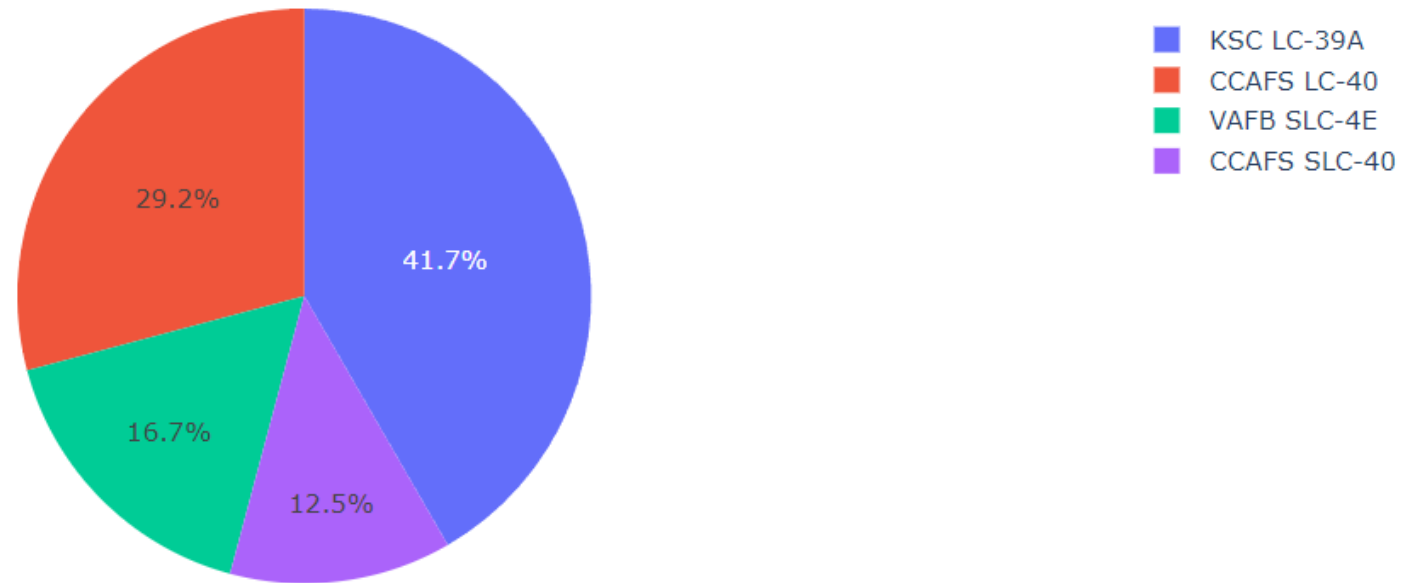


# The success percentage achieved by each launch site

---

- KSC LC-39A had the most successful launches

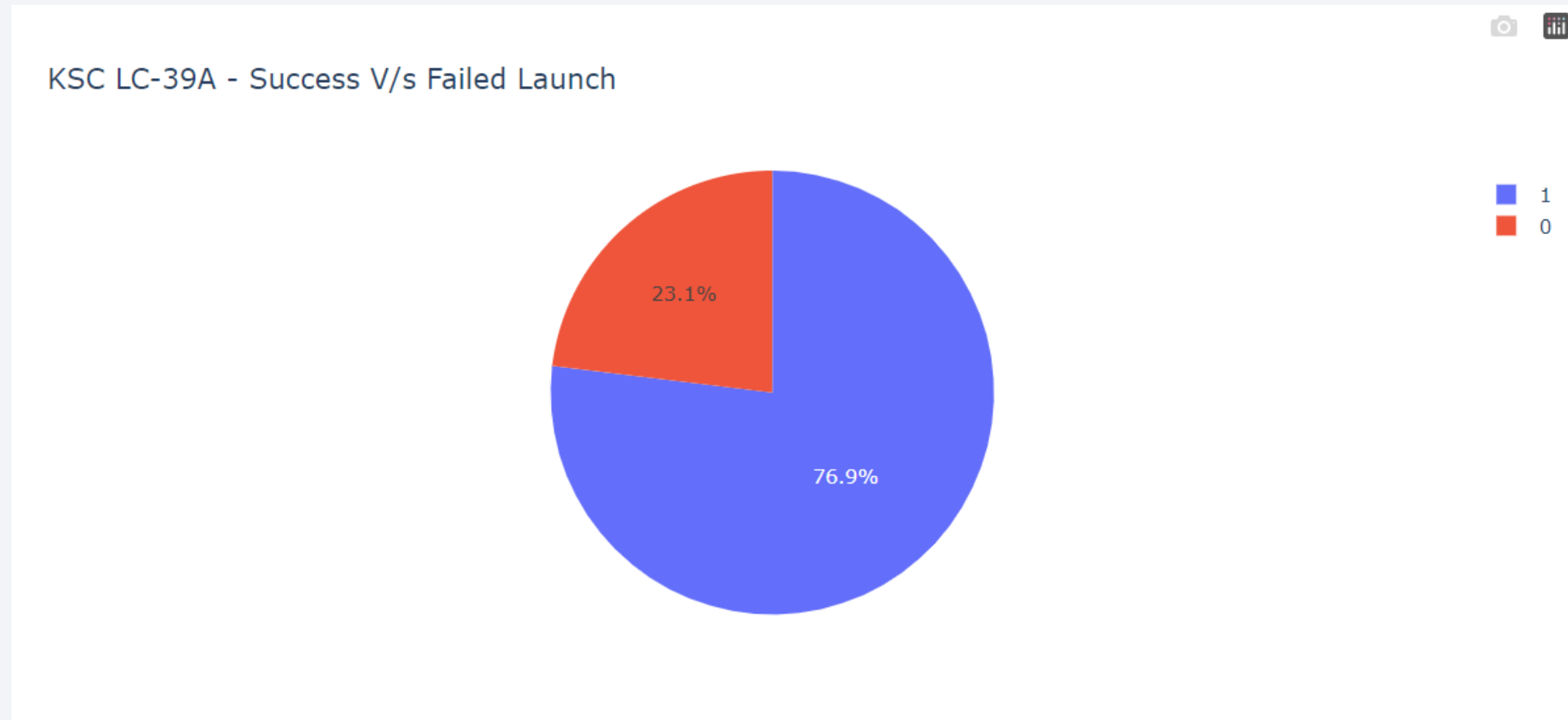
All Sites



# Launch site with the highest launch success ratio

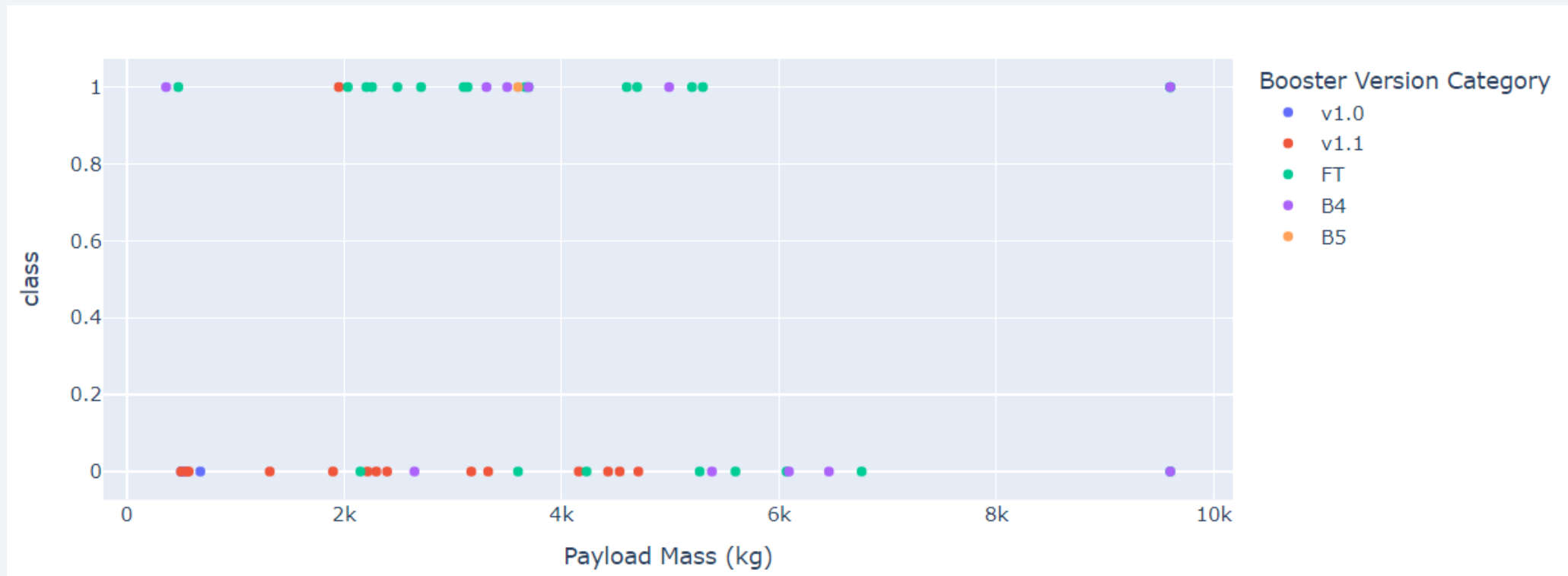
---

- KSC LC-39A Success rate reached 76.9%



## Payload vs Launch Outcome for all sites, with different payload selected in the range slider

- We can see the payload weight under 4000kg had higher success rate than which payload is over 4000kg.



Section 5

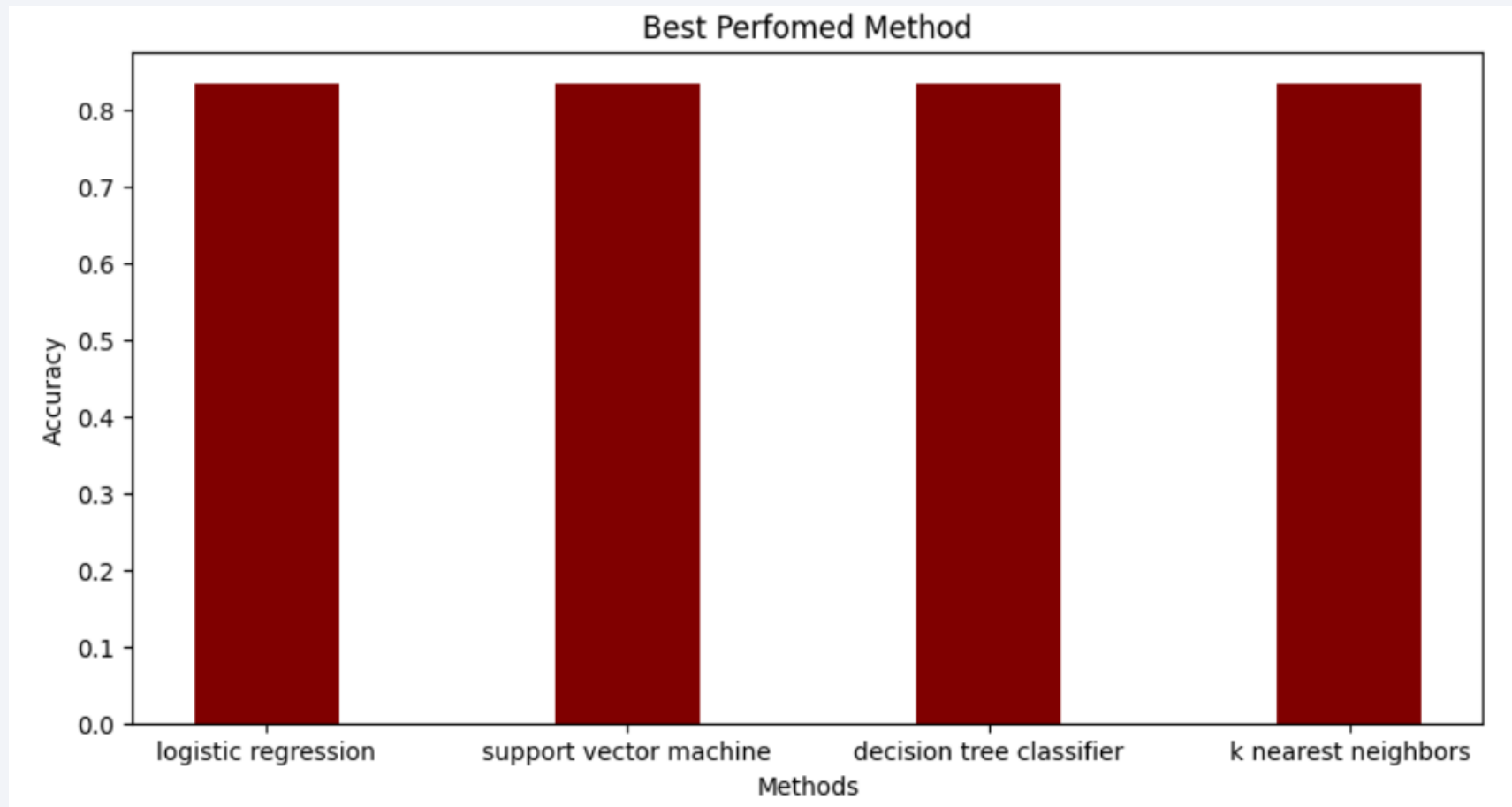
# Predictive Analysis (Classification)



# Classification Accuracy

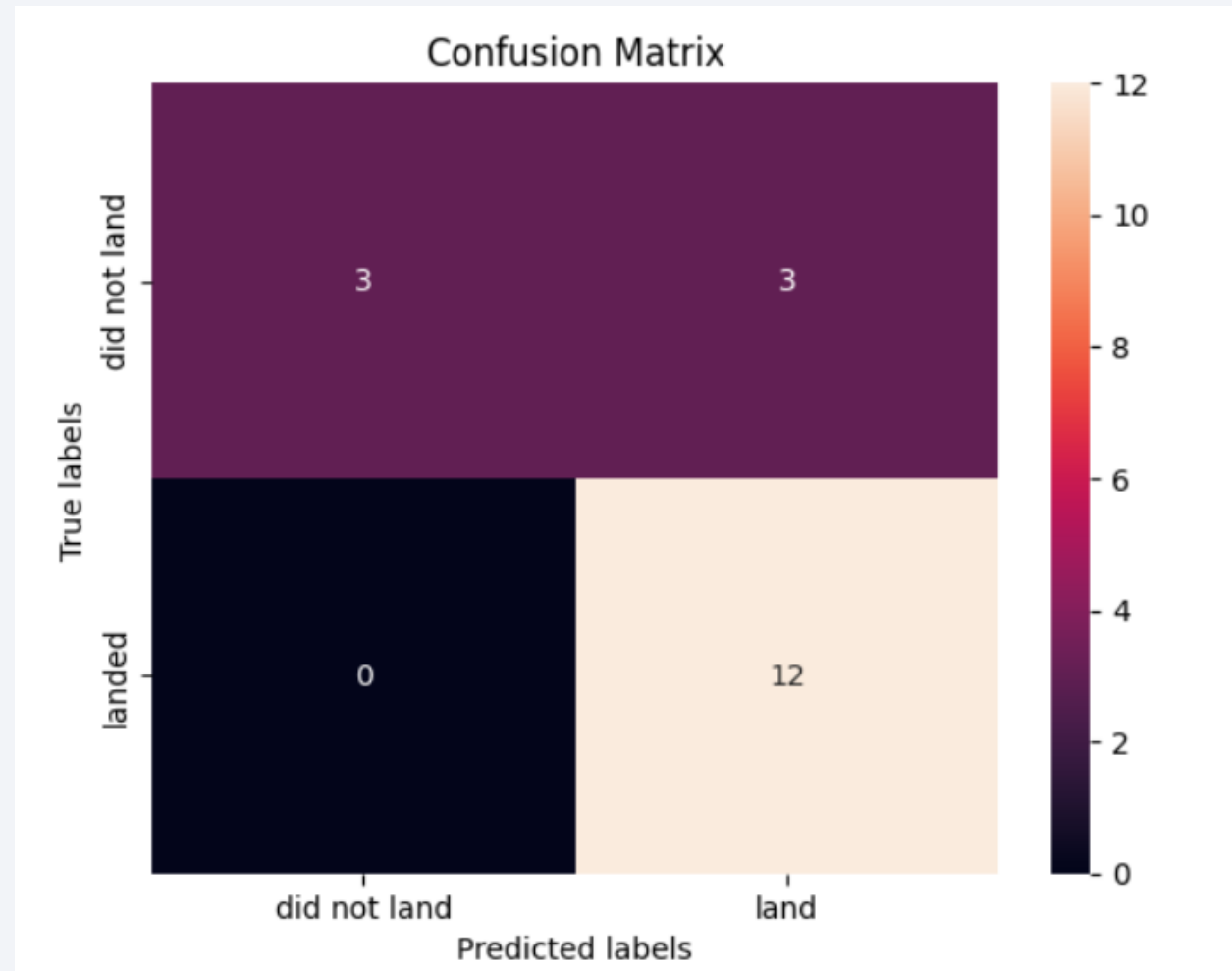
---

- All models had virtually the same accuracy on the test set at 83.33% accuracy.



# Confusion Matrix

- All models had same result in confusion matrix.
- TP(True Positive) : 12
- TF(False Positive) : 0
- TN (True Negative) : 3
- FN(False Negative) : 3



# Conclusions

---

We can conclude that:

- The larger the flight amount at a launch site, the greater the success rate at a launch site.
- Launch success rate started to increase in 2013 till 2020.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- KSC LC-39A had the most successful launches of any sites.
- The machine learning algorithm are same accuracy for this task.

# Appendix

---

- GitHub repository url :

[https://github.com/TonyHuang1688/Course10\\_Final\\_Assignment.git](https://github.com/TonyHuang1688/Course10_Final_Assignment.git)

Thank you!

