



IBM Developer  
SKILLS NETWORK

# Winning Space Race with Data Science

Nguyen Giang Son  
October 2021



# Outline

---

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

# Executive Summary

---

- In this project, I have:
  - Collected and cleaned SpaceX's launch data
  - Explored the data to draw some conclusion about the launches
  - Visualized the data on a map and a dashboard
  - Built predictive models on whether a launch will succeed or fail
- Summary of all results:
  - The project is rather successful and all predictive models give encouraging results.

# Introduction

---

- Project background and context:
  - SpaceX is a company that provides commercial space company that boast inexpensive rocket launches.
  - SpaceX's Falcon 9 launches cost \$62M, others cost at least \$165M.
  - To keep costs low, SpaceX need to reuse the first stage of launches.
- Problems you want to find answers:
  - Predict if the Falcon 9 first stage will land successfully.
  - Using that information (successful / unsuccessful landing of the first stage) to determine the price of a launch.



Section 1

# Methodology

# Methodology

---

## Executive Summary

- Data collection methodology:
  - Data was collected from SpaceX's API and from Wikipedia
- Perform data wrangling
  - Missing values were filled
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Multiple models were experimented and tuned to maximize their accuracy.

# Data Collection

---

- The data is collected using 2 methods:
  - Using [SpaceX's REST API](#).
  - Scraping the Wikipedia page: [List of Falcon 9 and Falcon Heavy launches](#)

# Data Collection – SpaceX API

---

- Present your data collection with SpaceX REST calls:
  - The REST API was used to collect data of booster version, launch pad, payload mass, orbit, ...
- Github link:

[Collecting data using SpaceX API notebook](#)

Place your flowchart of SpaceX API calls here



# Data Collection - Scraping

---

- Present your web scraping process:
  - Using BeautifulSoup, I scraped the Wikipedia page for the list of Falcon 9 launches

- Github link:

[Webscraping notebook](#)

Place your flowchart of web scraping here

# Data Wrangling

---

- Describe how data were processed:
  - filling in missing values
  - Adding outcome label
- Github link: [Data wrangling notebook](#)

# EDA with Data Visualization

---

- The types of charts included in this exercises are
  - scatter plot: to show the correlation between 2 variables
  - bar chart: to compare values of different categories
  - and line chart: to explore how a metric changed over time
- Github link: [EDA with Data Visualization notebook](#)

# EDA with SQL

---

- Using bullet point format, summarize the SQL queries you performed:
  - List the names of launch sites
  - 5 records where launch sites begins with “CCA”
  - Total payload mass carried by boosters launched by NASA
  - AVG payload mass carried by booster version F9 v1.1
  - The date when the first successful landing outcome in ground pad was achieved
  - The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - The total number of successful and failure mission outcomes
  - The names of the booster\_versions which have carried the maximum payload mass
  - failed landing\_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order
- Github link: [EDA with SQL notebook](#)

# Build an Interactive Map with Folium

---

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
  - Markers: to mark launch sites
  - Circles: to mark the NASA Johnson Space Center
  - Lines: to visualize the distance between a launch site and the nearest city, railway, ...
- Github link: [Folium notebook](#)

# Build a Dashboard with Plotly Dash

---

- Summarize what plots/graphs and interactions you have added to a dashboard
- Explain why you added those plots and interactions
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose



# Predictive Analysis (Classification)

---

- Summarize how you built, evaluated, improved, and found the best performing classification model:
  - Multiple models were trained on the train set: logistic regression, SVM and decision tree
  - They were tuned with hyperparameter tuning, then evaluated using their accuracy score on the test set
  - In conclusion, all models perform equally well.
- Github link: [Machine Learning notebook](#)

# 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 solid blue area on the left side, which transitions into a dynamic pattern of diagonal streaks in shades of blue and red on the right. These streaks are layered over a faint, light-blue grid pattern, creating a sense of depth and movement.

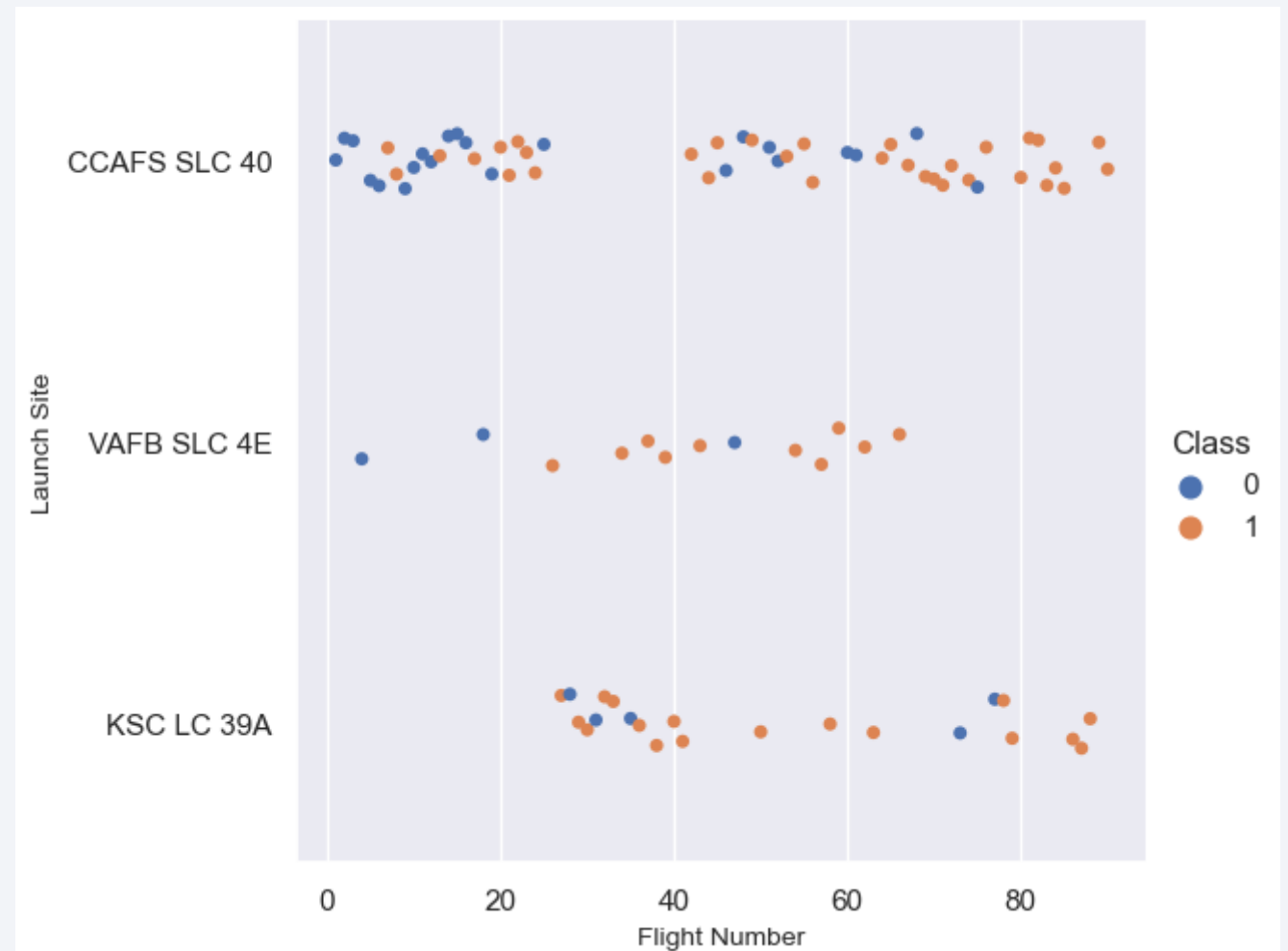
Section 2

# Insights drawn from EDA



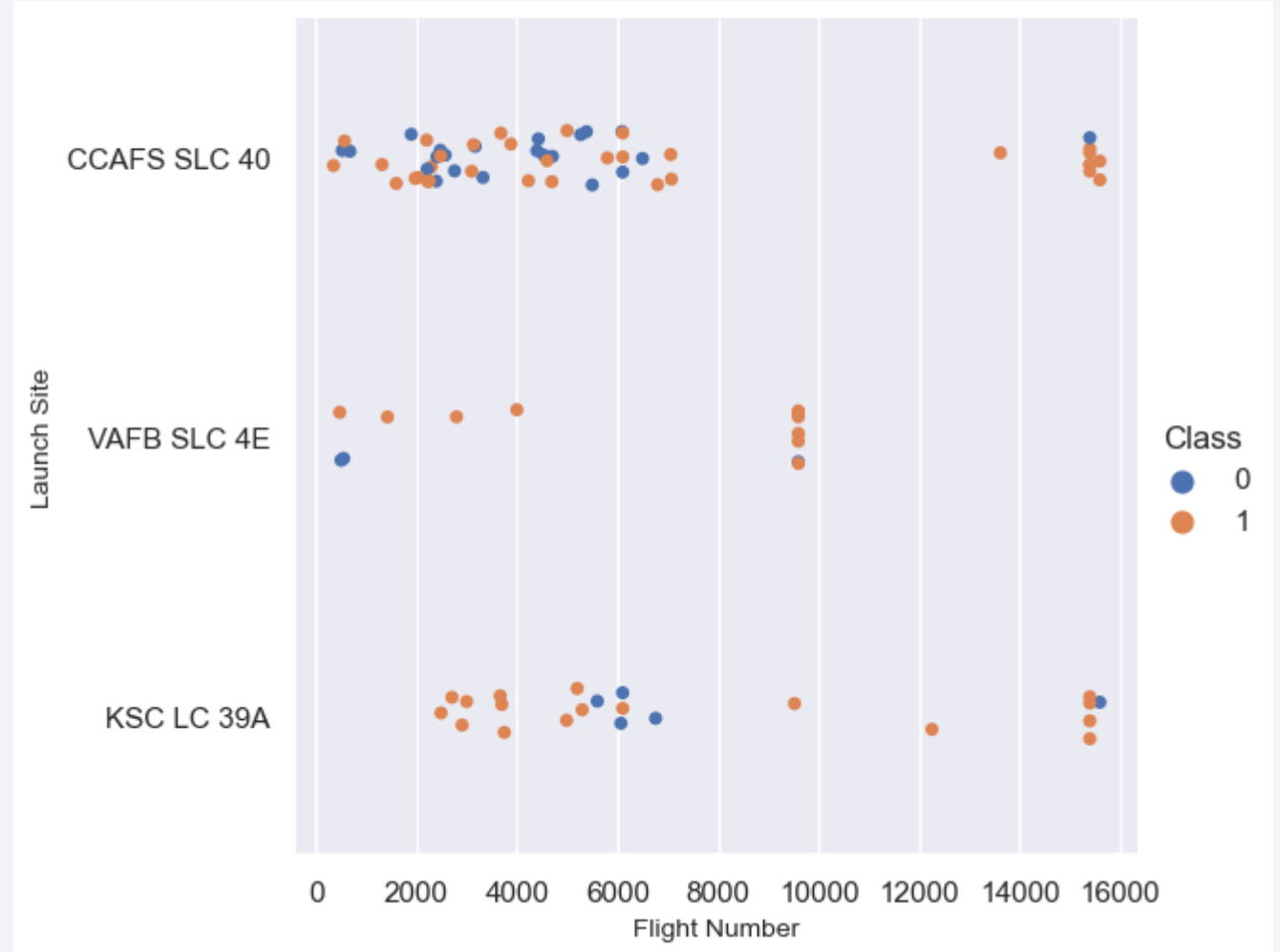
# Flight Number vs. Launch Site

- CCAFS SLC 40 site hosted the most flights.
- The majority of flights at VAFB SLC 4E AND KSC LC 39A were from number 20+
- The most recent flight at VAFB SLC 4E was flight no. 64.



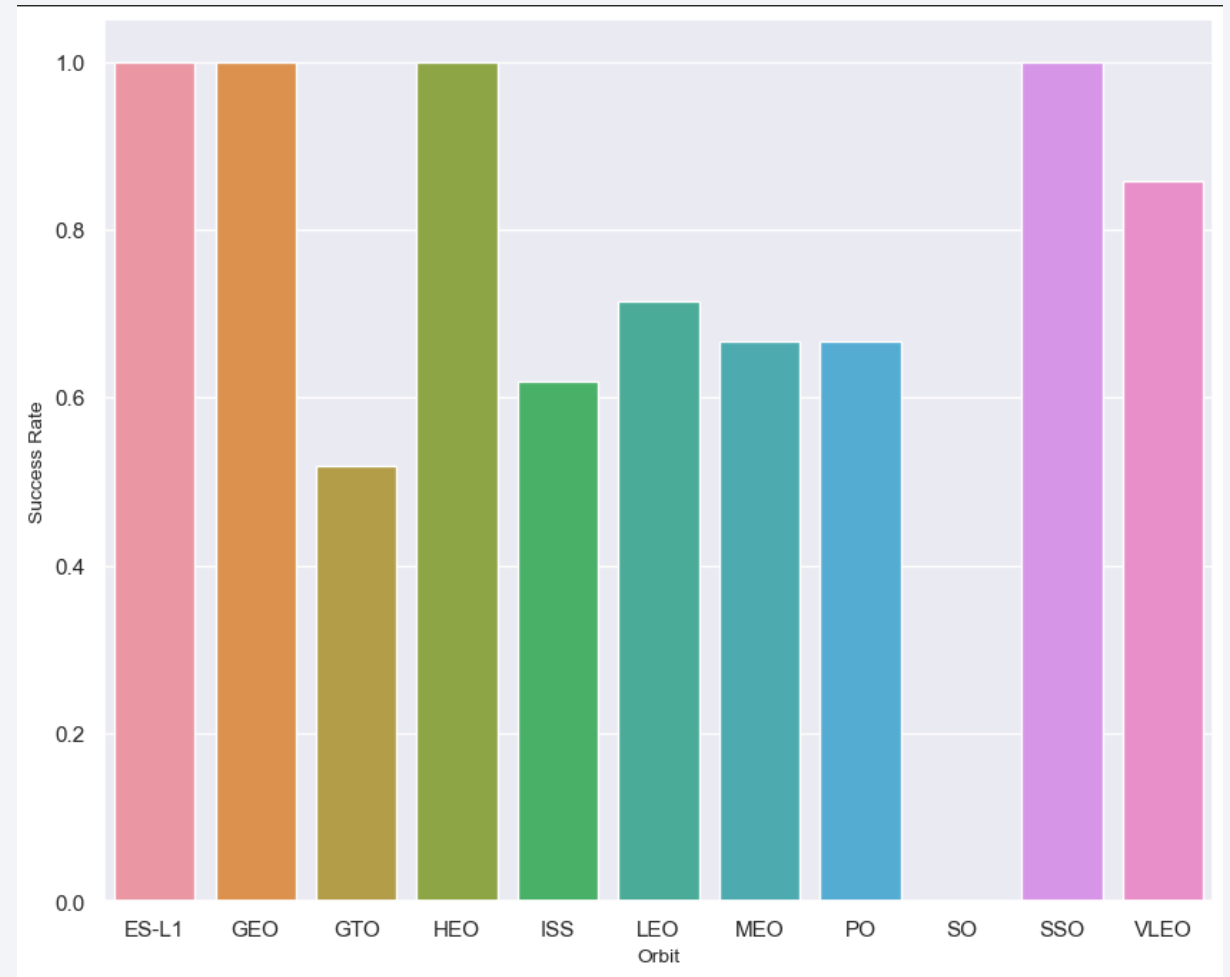
# Payload vs. Launch Site

- For the VAFB-SLC launchsite there are no rockets launched for heavy payload mass ( $> 10000\text{kg}$ ).
- For flights that carried heavy payload ( $> 10000\text{kg}$ ), the failure rate were quite low (only 2 flights were unsuccessful).



# Success Rate vs. Orbit Type

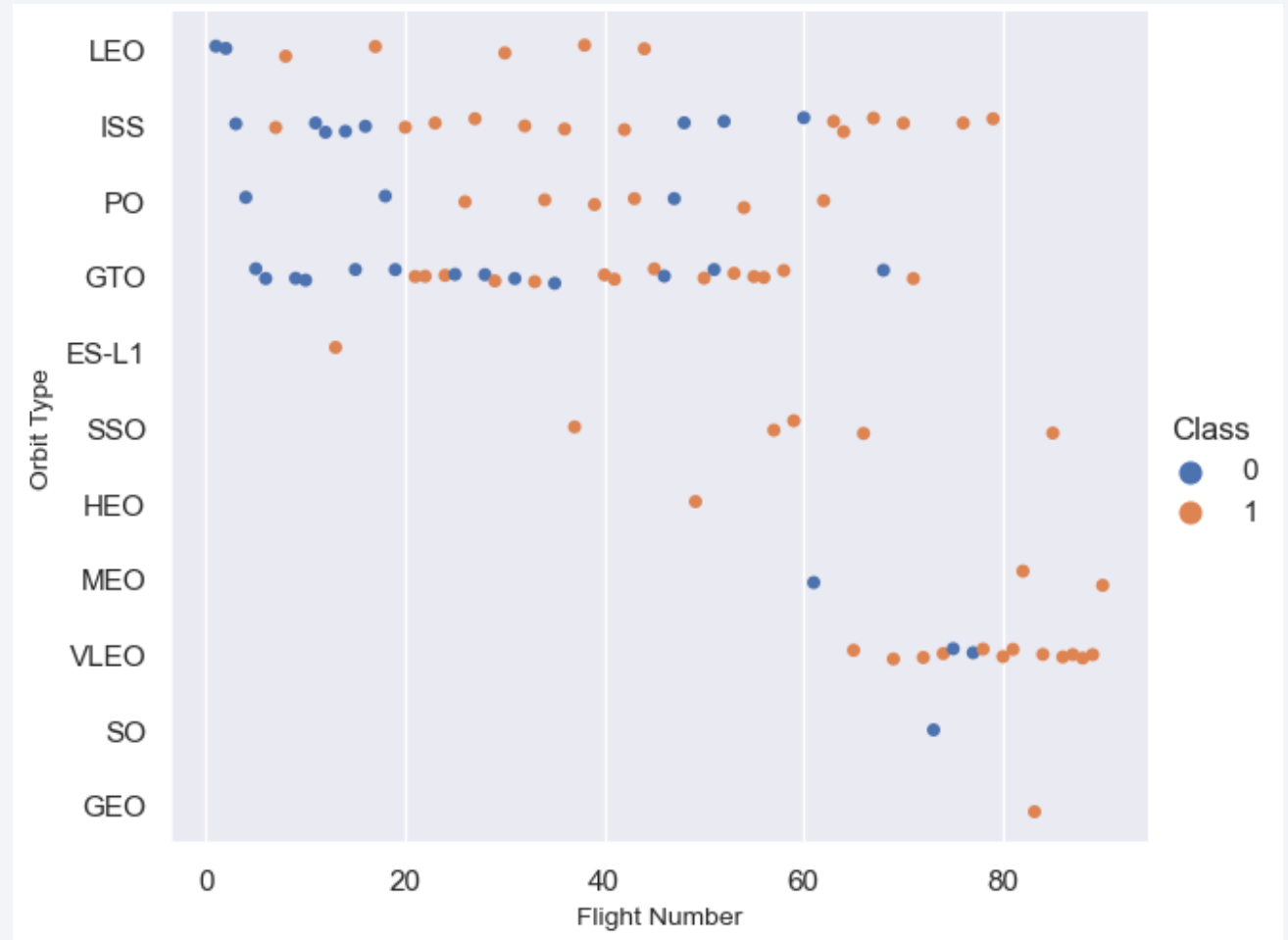
- For 4 orbit types (i.e: ES-L1, GEO, HEO, SSO), the success rates were 100%.
- 5 of 10 orbit types only had success rates between 50% and 70%.





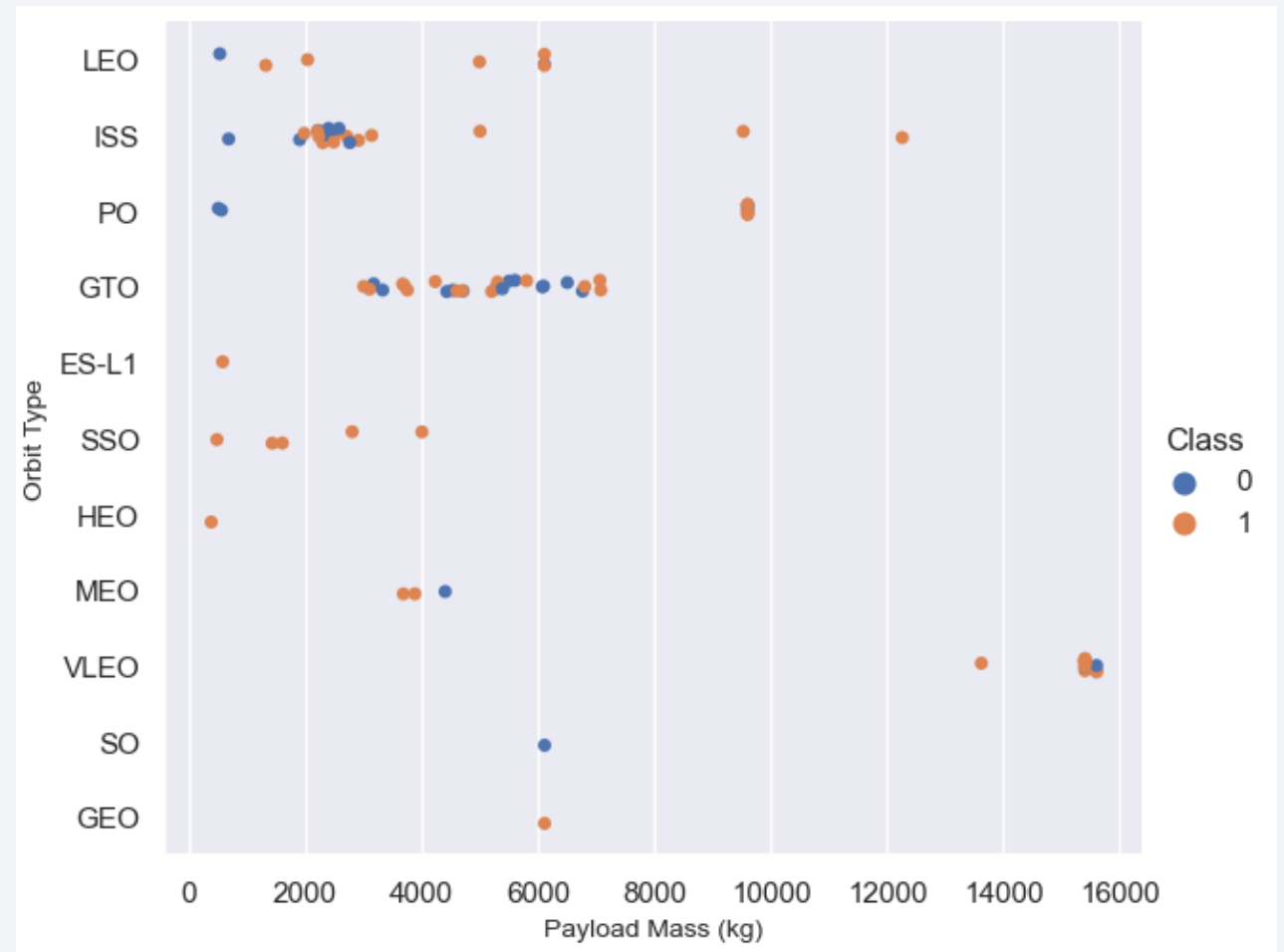
# Flight Number vs. Orbit Type

- In the LEO orbit the Success appears related to the number of flights;
- There seems to be no relationship between flight number when in GTO orbit.



# Payload vs. Orbit Type

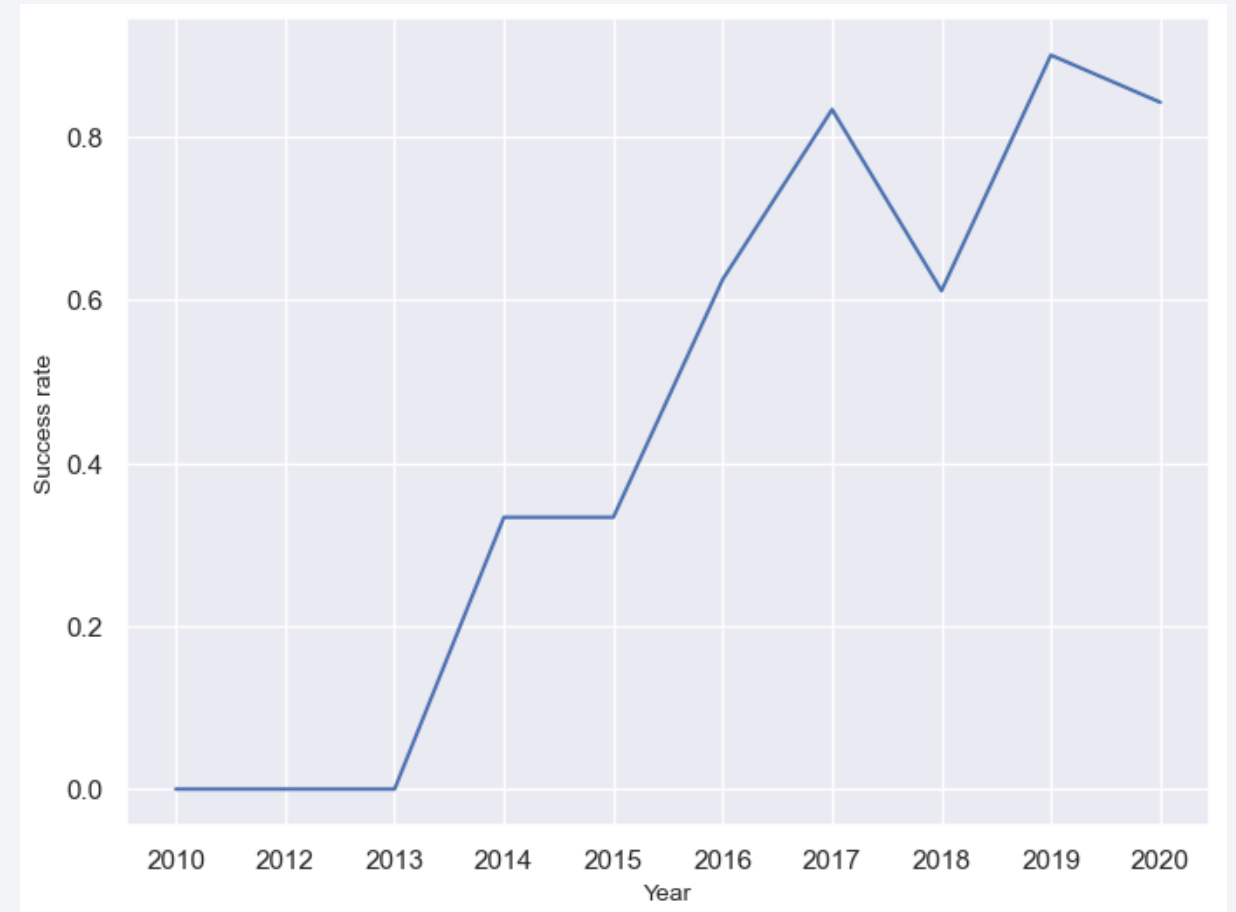
- With heavy payloads the positive landing rate are more for Polar, LEO and ISS.
- However, for GTO we cannot distinguish this well as both positive landing rate and negative landing are both there here.



# Launch Success Yearly Trend

---

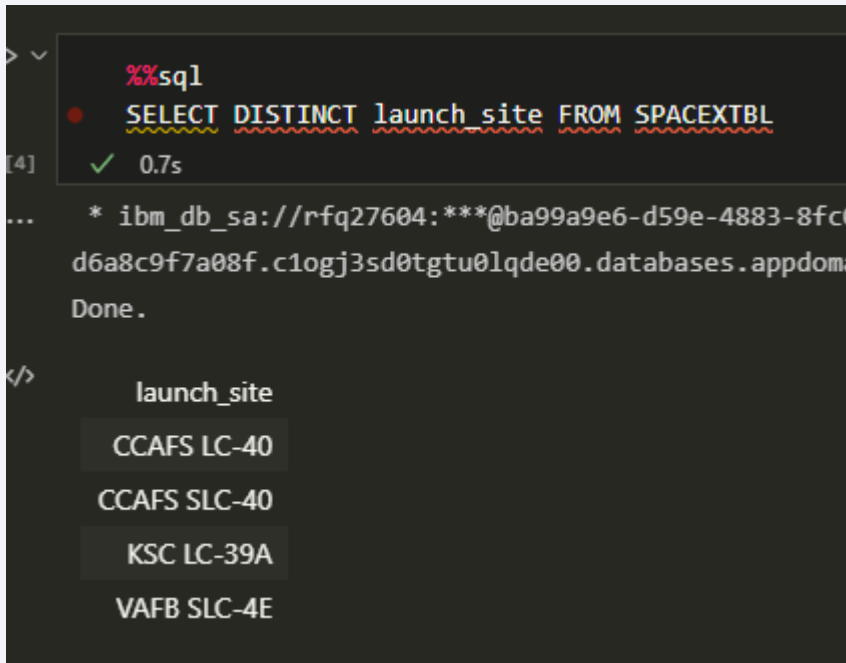
- From 2013 to 2020, annual success rates were continuously increasing (except in 2018).



# All Launch Site Names

---

- Find the names of the unique launch sites
- There are 4 unique launch sites



```
> %sql
SELECT DISTINCT launch_site FROM SPACEXTBL
[4] ✓ 0.7s
... * ibm_db_sa://rfq27604:***@ba99a9e6-d59e-4883-8fcd6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdom
Done.
</>
launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E
```

# Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Below are 5 records where the launch sites begin with 'CCA'

```
%%sql
SELECT *
FROM SPACEXTBL
WHERE launch_site LIKE 'CCA%'
LIMIT 5
```

✓ 0.7s

\* ibm\_db\_sa://rfq27604:\*\*\*@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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
- The total payload mass carried by boosters from NASA is 45596 kg

```
> %%sql
SELECT SUM(payload_mass_kg) AS total_payload_mass
FROM SPACEXTBL
WHERE customer = 'NASA (CRS)'
GROUP BY customer
[6] ✓ 0.7s
... * ibm_db_sa://rfq27604:***@ba99a9e6-d59e-4883-8fc0-d6a8
Done.
</>
total_payload_mass
45596
```



# Average Payload Mass by F9 v1.1

---

- Calculate the average payload mass carried by booster version F9 v1.1
- The average payload mass carried by booster version F9 v1.1 is 2928 kg

```
> %%sql
SELECT AVG(payload_mass_kg) AS avg_payload_mass
FROM SPACEXTBL
WHERE booster_version = 'F9 v1.1'
GROUP BY booster_version

[7] ✓ 0.7s
... * ibm_db_sa://rfq27604:***@ba99a9e6-d59e-4883-8fc0-d6
Done.

</>
avg_payload_mass
2928
```

# First Successful Ground Landing Date

---

- Find the dates of the first successful landing outcome on ground pad
- The date of the first successful landing outcome on ground pad is 22/12/2015

```
%%sql
SELECT MIN(date) AS first_successful_landing_date
FROM SPACEXTBL
WHERE landing_outcome = 'Success (ground pad)'
```

✓ 0.7s

\* ibm\_db\_sa://rfq27604:\*\*\*@ba99a9e6-d59e-4883-8fc0-d6a8

Done.

first_successful_landing_date
2015-12-22

## 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
- 4 boosters have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
%%sql
SELECT DISTINCT booster_version
FROM SPACEXTBL
WHERE landing_outcome = 'Success (drone ship)' AND payload_mass_kg BETWEEN 4000 AND 6000
```

[9] ✓ 0.7s

... \* ibm\_db\_sa://rfq27604:\*\*\*@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdoma

Done.

</>

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

# Total Number of Successful and Failure Mission Outcomes

---

- Calculate the total number of successful and failure mission outcomes
- There are 100 successful mission and 1 failed mission

```
%%sql
SELECT
    mission_outcome,
    COUNT(mission_outcome) AS count_mission_outcome
FROM (SELECT
    CASE
        WHEN mission_outcome LIKE 'Success%' THEN 'Success'
        ELSE 'Failure'
    END AS mission_outcome FROM SPACEXTBL)
GROUP BY mission_outcome
```

✓ 0.7s

\* ibm\_db\_sa://rfq27604:\*\*\*@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3s

Done.

mission_outcome	count_mission_outcome
Failure	1
Success	100

# Boosters Carried Maximum Payload

---

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

```
%%sql
SELECT DISTINCT booster_version
FROM SPACEXTBL
WHERE payload_mass_kg_ = (SELECT MAX(payload_mass_kg_) FROM SPACEXTBL)
✓ 0.7s

* ibm_db_sa://rfq27604:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0
Done.

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3
```

# 2015 Launch Records

---

- List the failed landing\_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- 2 launches failed in 2015

```
%%sql
SELECT booster_version, launch_site, landing_outcome
FROM SPACEXTBL
WHERE LEFT(date, 4) = '2015'
      AND landing_outcome = 'Failure (drone ship)'
```

✓ 0.7s

\* ibm\_db\_sa://rfq27604:\*\*\*@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1log

Done.

booster_version	launch_site	landing_outcome
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)



# 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.
- 10 landings were not attempted between the dates above.

```
%%sql
SELECT landing_outcome, COUNT(landing_outcome) AS count_landing_outcomes
FROM SPACEXTBL
WHERE date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY landing_outcome
ORDER BY count_landing_outcomes DESC
```

✓ 0.7s

\* ibm\_db\_sa://rfq27604:\*\*\*@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu01

Done.

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

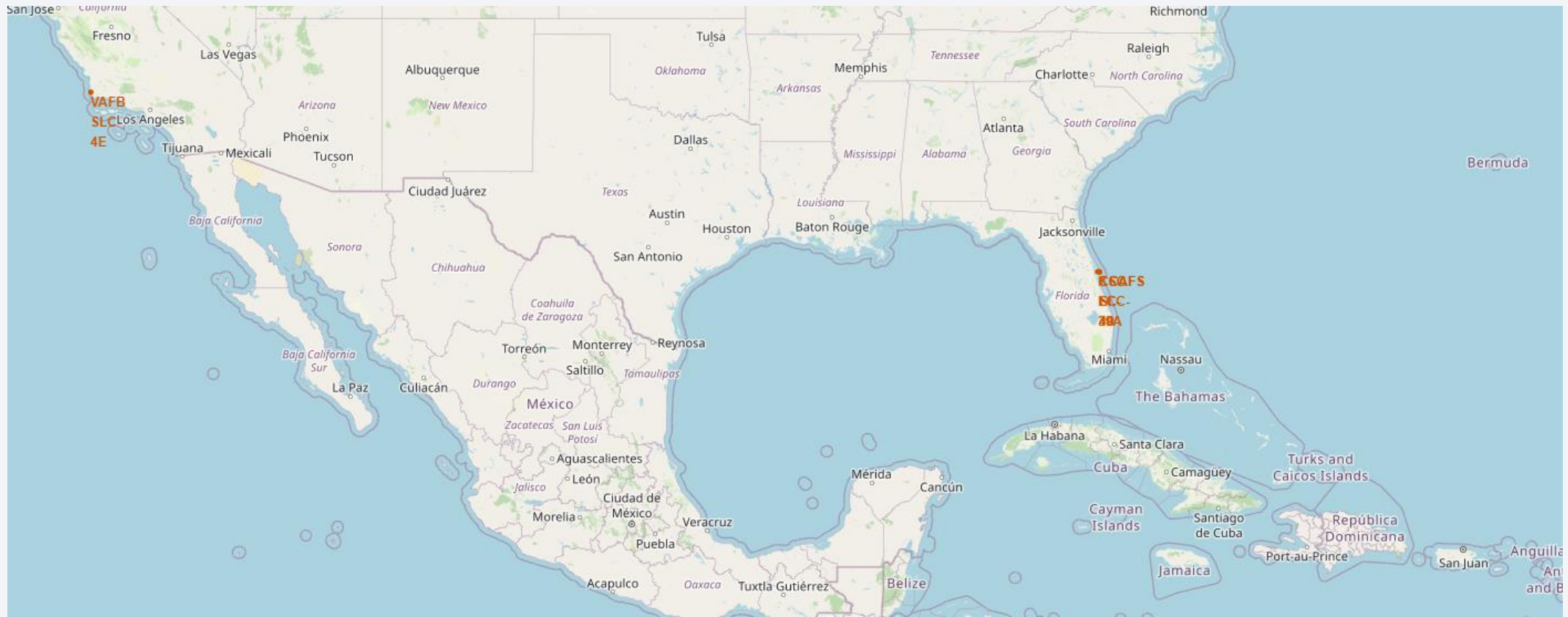
Section 4

# Launch Sites Proximities Analysis



# <Folium Map Screenshot 1>

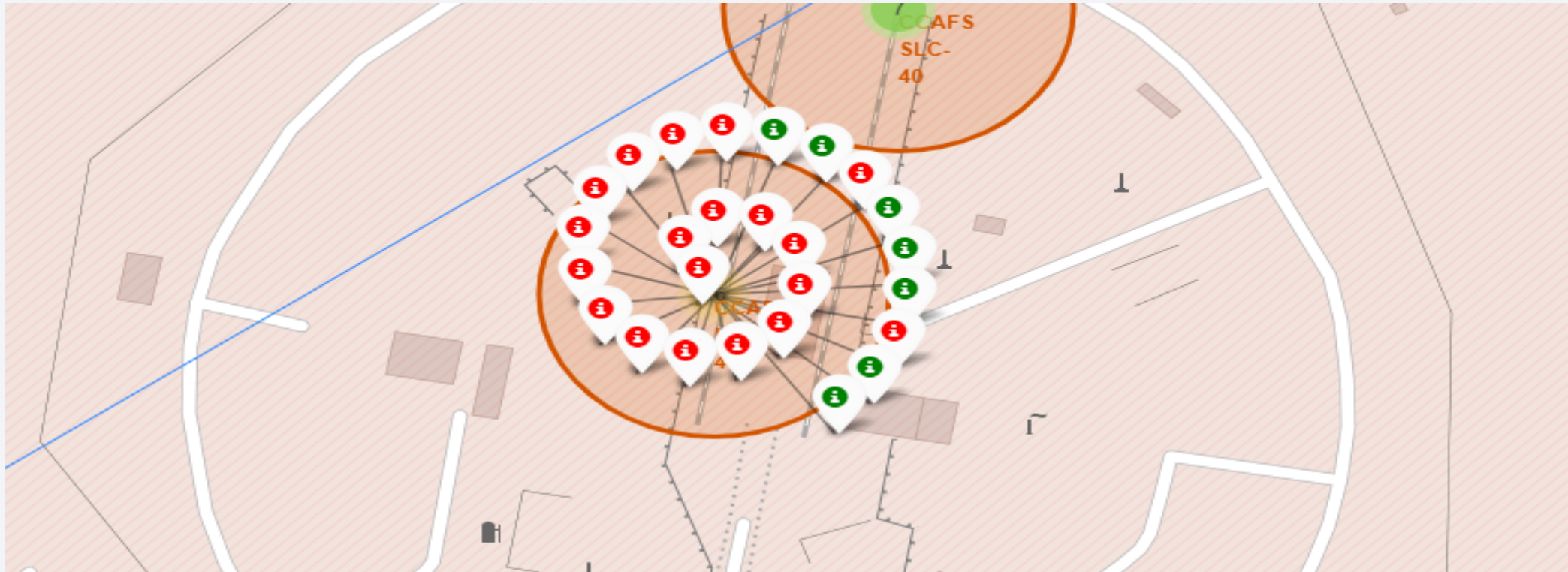
- This map marks the location of all 4 launch sites.



## <Folium Map Screenshot 2>

---

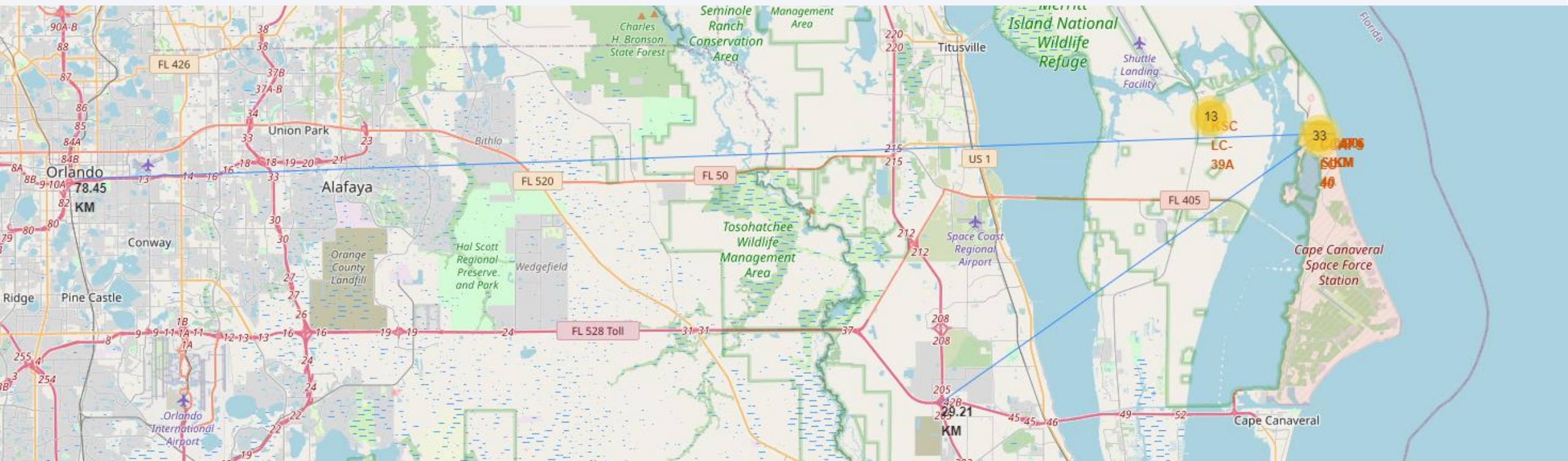
- This maps shows the launch outcomes as indicated by color (green = success, red = failure)





# <Folium Map Screenshot 3>

- This map shows the distance of a selected launch site to the nearest railroad, coastline, and city.







Section 5

# Build a Dashboard with Plotly Dash

# <Dashboard Screenshot 1>

---

- Replace <Dashboard screenshot 1> title with an appropriate title
- Show the screenshot of launch success count for all sites, in a piechart
- Explain the important elements and findings on the screenshot

## <Dashboard Screenshot 2>

---

- Replace <Dashboard screenshot 2> title with an appropriate title
- Show the screenshot of the piechart for the launch site with highest launch success ratio
- Explain the important elements and findings on the screenshot



## <Dashboard Screenshot 3>

---

- Replace <Dashboard screenshot 3> title with an appropriate title
- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider
- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.

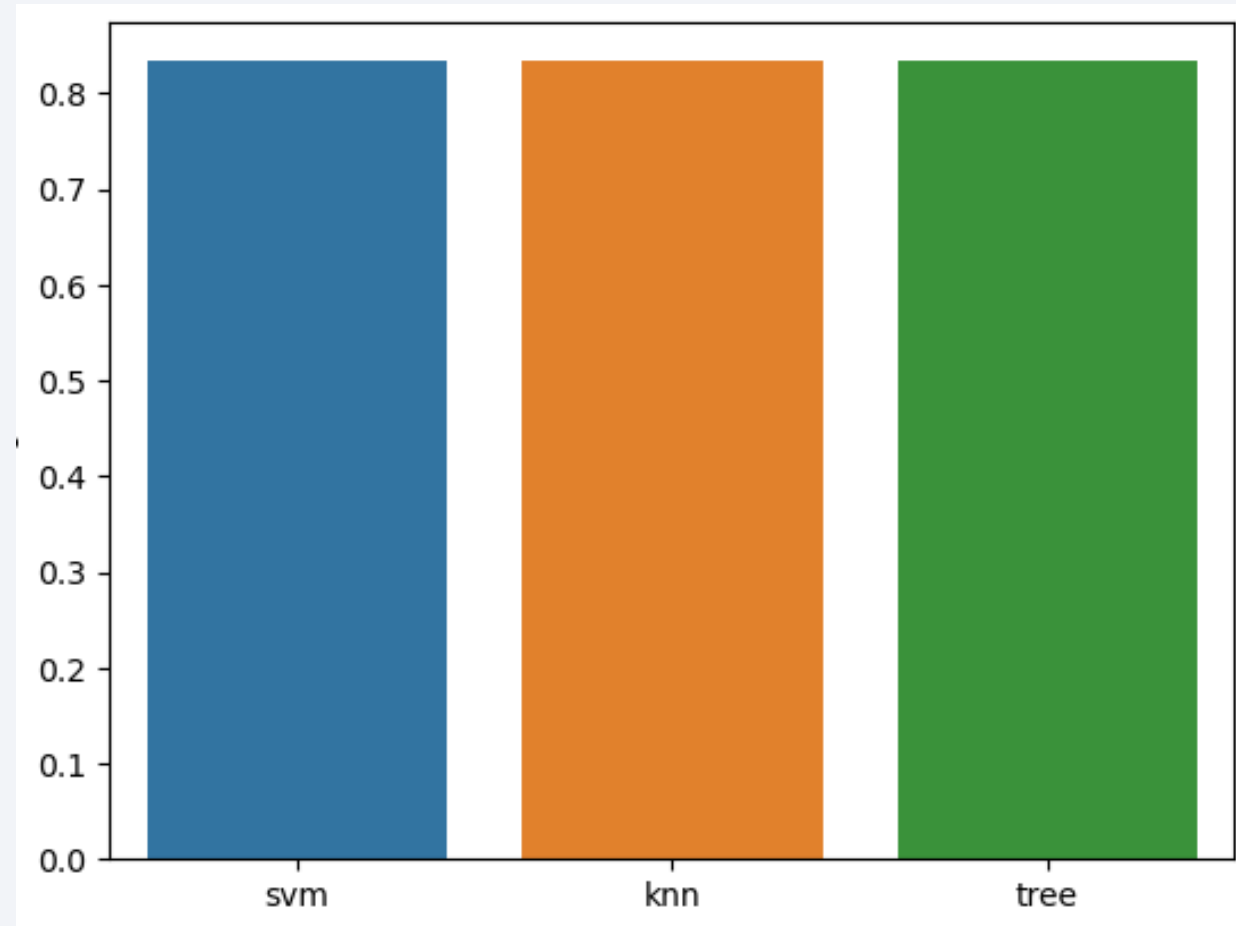
Section 6

# Predictive Analysis (Classification)

# Classification Accuracy

---

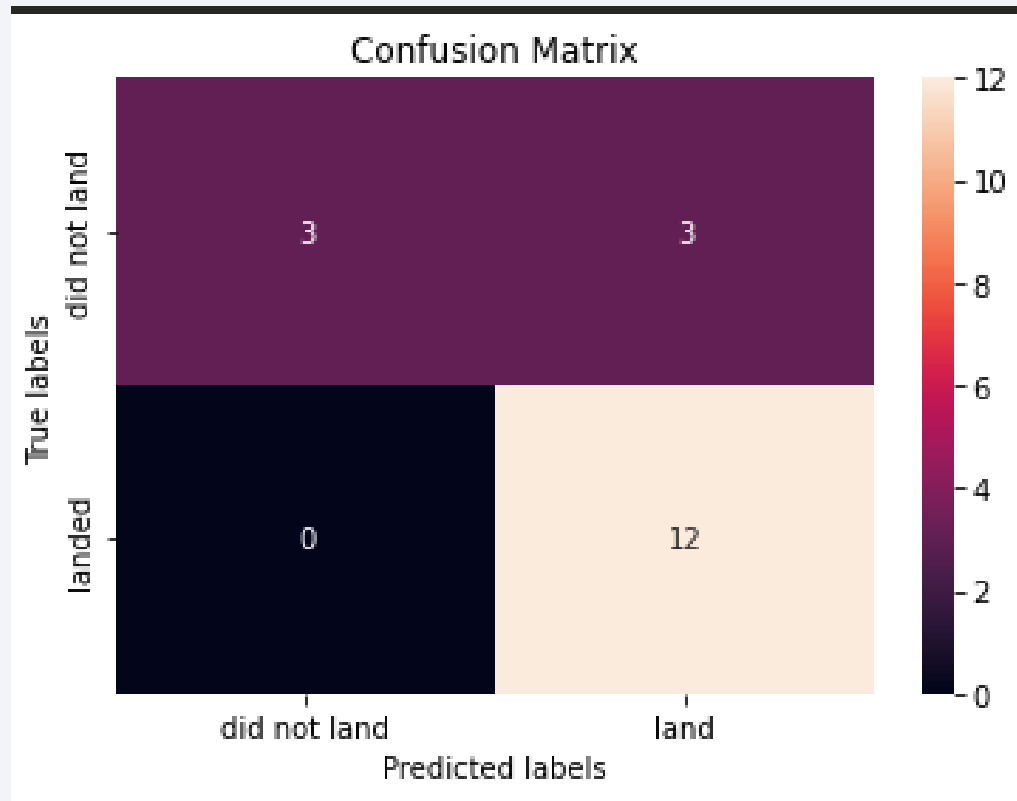
- All models (SVM, KNN and Decision Tree) perform similarly and have equal accuracy score (0.83)



# Confusion Matrix

---

- All models give the same confusion matrix as follows:



# Conclusions

---

- All ML models perform equally well
- Their accuracies are the same (0.83)

# Appendix

---

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!

