IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Peiyong Yu>
<2022.4.5>

Outline

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

Executive Summary

- The project uses machine learning models and public information from SpaceX to predict if a rocket launch will land successfully.
- Data are obtained from SpaceX REST API and web scrapping of Falcon 9 Wikipage.
- SQL techniques and Python visualizations are used to explore the data.
- The data is also examined through an interactive dashboard.
- It is discovered that the success rate of rocket launch are closely related to payload, launch site, year, orbit type, and flight number.
- Four classification models are tested to predict the status of a rocket launch. The decision tree model performs the best with the highest accuracy rate.

Introduction

- **Project background and context**
- A new commercial space company, Space Y, would like to compete with SpaceX. The cost of rocket launching is the key in the space industry.
- The aim of this project is to determine the average price of each rocket launch for Space Y. High cost of rocket launching occurs when the launching or the recycling fails. That being said, reusing of the rockets from various stages would decrease the cost dramatically.
- This project aims to determine if SpaceX will reuse the first stage. Instead of using rocket science, we will train a machine learning model and use public information to predict if the first stage will land successfully.

Section 1

Methodology

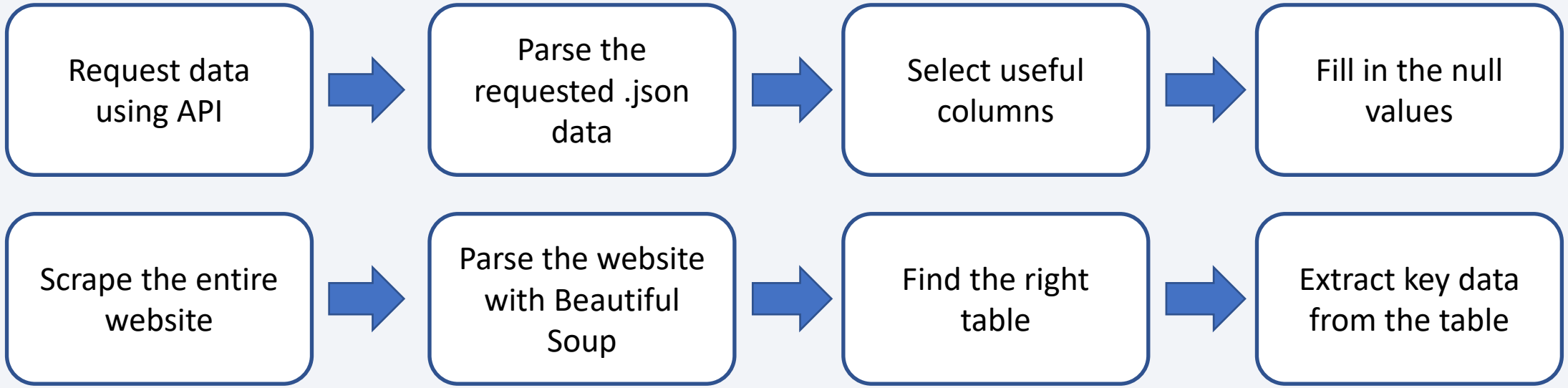
Methodology

Executive Summary

- Data collection methodology:
  - SpaceX REST API and Web Scraping from Falcon 9 Wikipedia page
- Perform data wrangling
  - Fill in null values, filters, and create dummies
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Normalization, data split, tuning parameters, and comparing accuracy

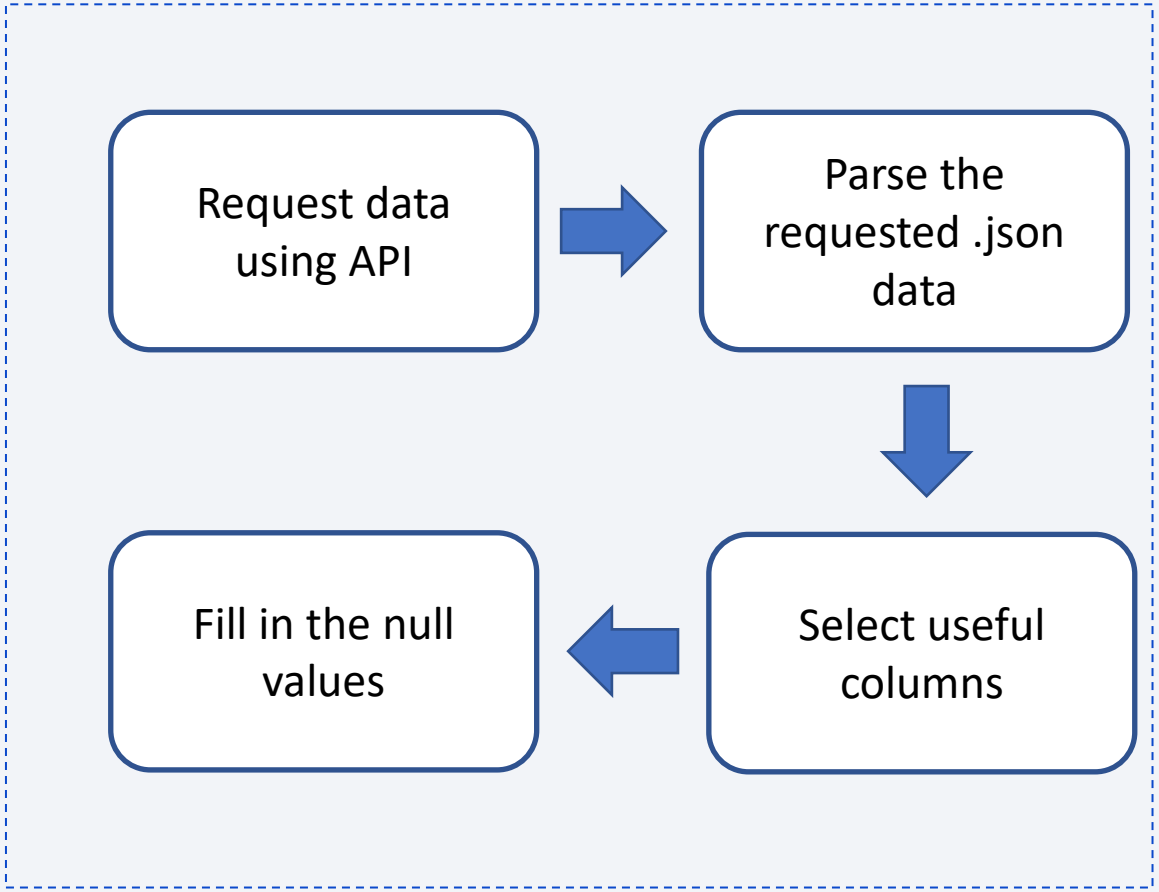
Data Collection

- Data was collected in two ways:
  - From SpaceX public data using SpaceX REST API
  - Web scraping to collect Falcon 9 historical launch records from a Wikipedia page



Data Collection – SpaceX API

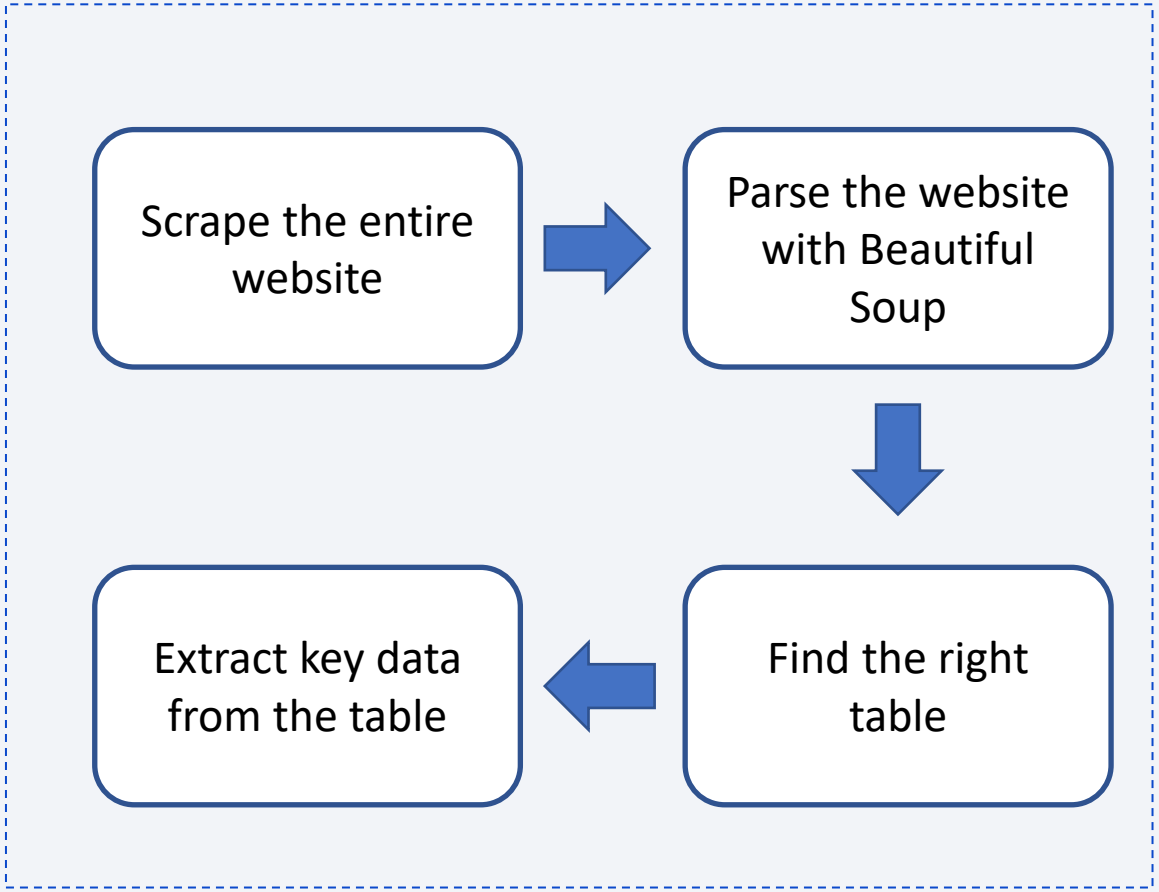
- Data collection with SpaceX REST API
  - Source: https://api.spacexdata.com/v4/launches/past
  - Use requests library to get the content
  - Use the pandas library to parse the requested .json file to a table
  - Select columns and use additional APIs to fill in null values



- https://github.com/pyu999/Capstone_IBM_DS/blob/master/ML%20Final%20Assignment.ipynb

Data Collection - Scraping

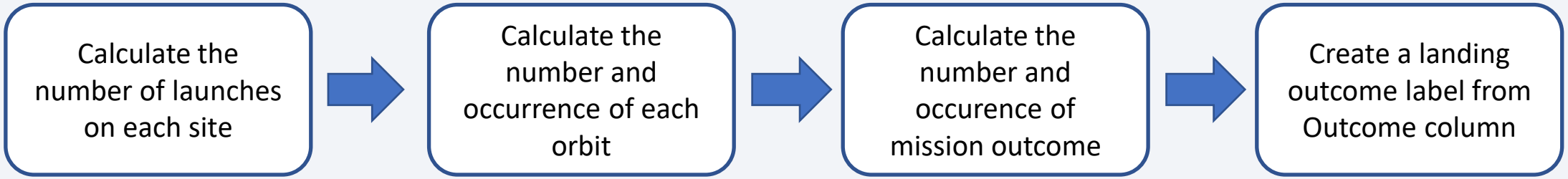
- Web scraping process:
  - Source: List of Falcon 9 and Falcon Heavy launches Wikipage
  - Using requests library to get the content
  - Using beautiful soup library to parse the result
  - Using find() and find_all() functions to extract content
- https://github.com/pyu999/Capstone_IBM_DS/blob/master/Web%20Scraping%20lab.ipynb



Data Wrangling

- Initial data screening
  - Calculate the number of launches on each site
  - Calculate the number and occurrence of each orbit
  - Calculate the number and occurence of mission outcome
  - Create a landing outcome label from Outcome column

- https://github.com/pyu999/Capstone_IBM_DS/blob/master/EDA%20Lab.ipynb



EDA with Data Visualization

- The charts plotted aim to find which factors are related to the launch success rate. So, successfulness or success rate is plotted against the following factors:
  - Flight Number
  - Pay load mass
  - Launch site
  - Orbit
  - Year

- https://github.com/pyu999/Capstone_IBM_DS/blob/master/EDA%20with%20Visualization.ipynb

EDA with SQL

- SQL queries performed to investigate:
  - The names of the unique launch sites in 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
  - Dates when the first successful landing outcome in ground pad was achieved
  - Names of the boosters which have success in drone ship
  - The successful and failure mission outcomes with features like drone ship, their booster versions, and launch site
  - Find the booster versions which have carried the maximum payload mass
- https://github.com/pyu999/Capstone_IBM_DS/blob/master/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

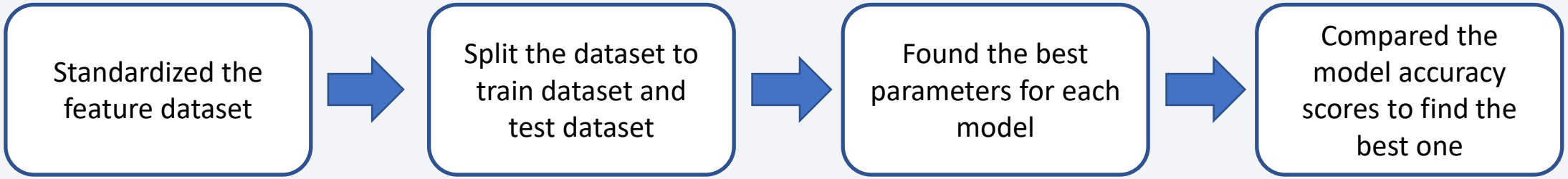
- The Folium map displayed the following objects:
  - Locations and name labels of the four launch sites
  - Launch attempts at each of the four sites
  - Successful attempts are labeled green, failed one are labeled red
  - Distance between California Site VAFB SLC-4E and its nearest coastline
- This exercise helps to understand the locations of the launch sites, and their respective success rates
- https://github.com/pyu999/Capstone_IBM_DS/blob/master/Interactive%20Visual%20Analytics.ipynb
-

Build a Dashboard with Plotly Dash

- An interactive dashboard were built with Plotly and Dash libraries
- The dashboard visualized:
  - The ratio of success launches from each launch site
  - The success rate of each site upon selection
  - A scatter plot shows the success/failure status of each launch against payload classified by Booster type
- https://github.com/pyu999/Capstone_IBM_DS/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Classification models built, evaluated, improved, and found the best performing one
  - Four models are selected to predict if a rocket launch will success or fail, namely logistic regression, support vector machine, decision tree, and K nearest neighbors
  - The GridSearchCV function was used to select parameters for each model
  - The overall accuracy was used to determine a best model
- https://github.com/pyu999/Capstone_IBM_DS/blob/master/ML%20Prediction.ipynb



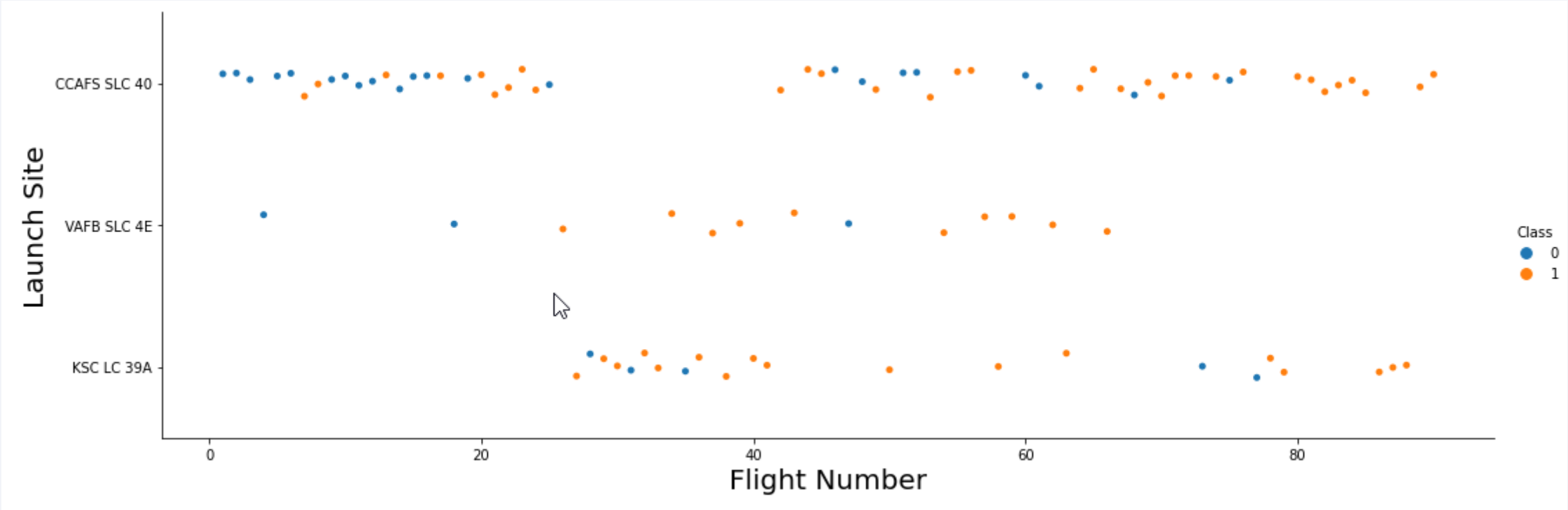
Results

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

Section 2

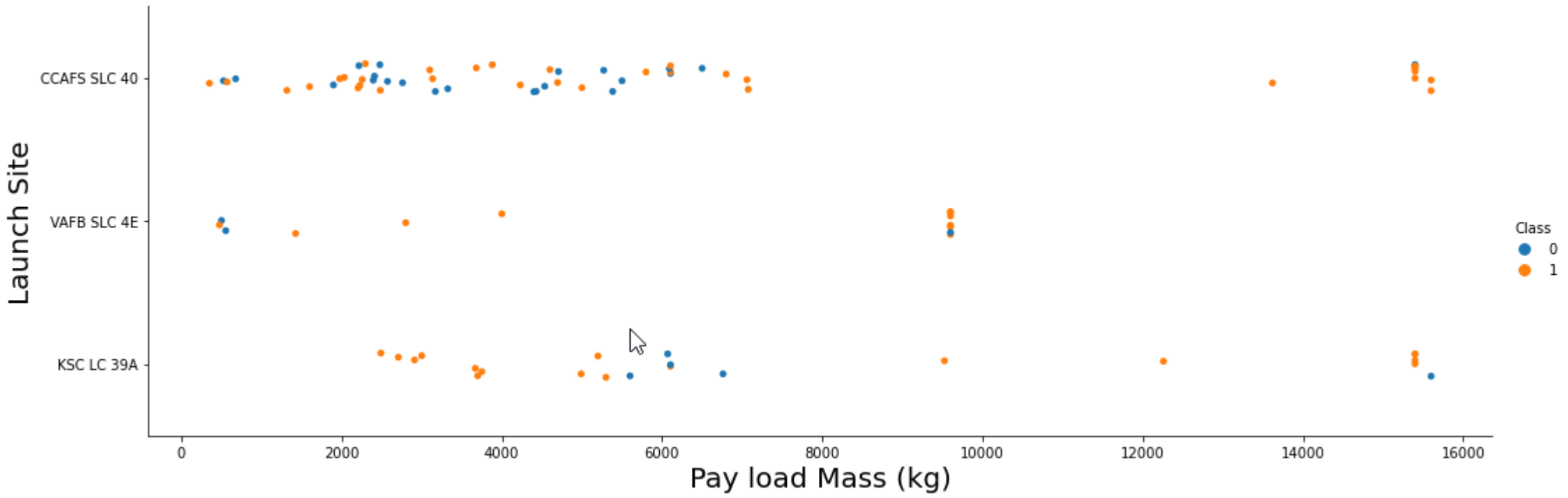
Insights drawn from EDA

Flight Number vs. Launch Site



- The success rate increased as the flight number increased (experienced gain from testing)
- CCAFS SLC 40 held most of the recent launches, which have been successful.

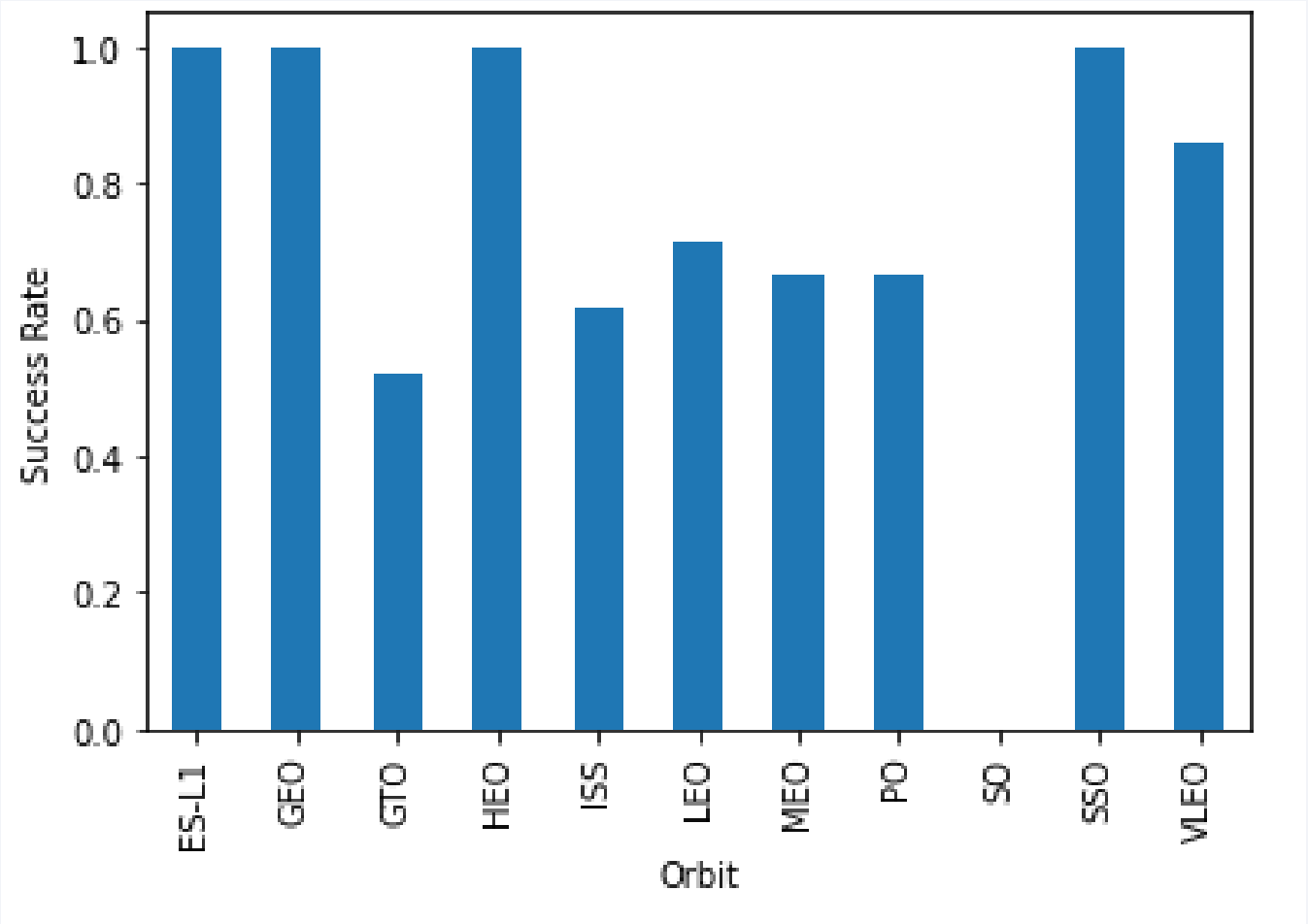
Payload vs. Launch Site



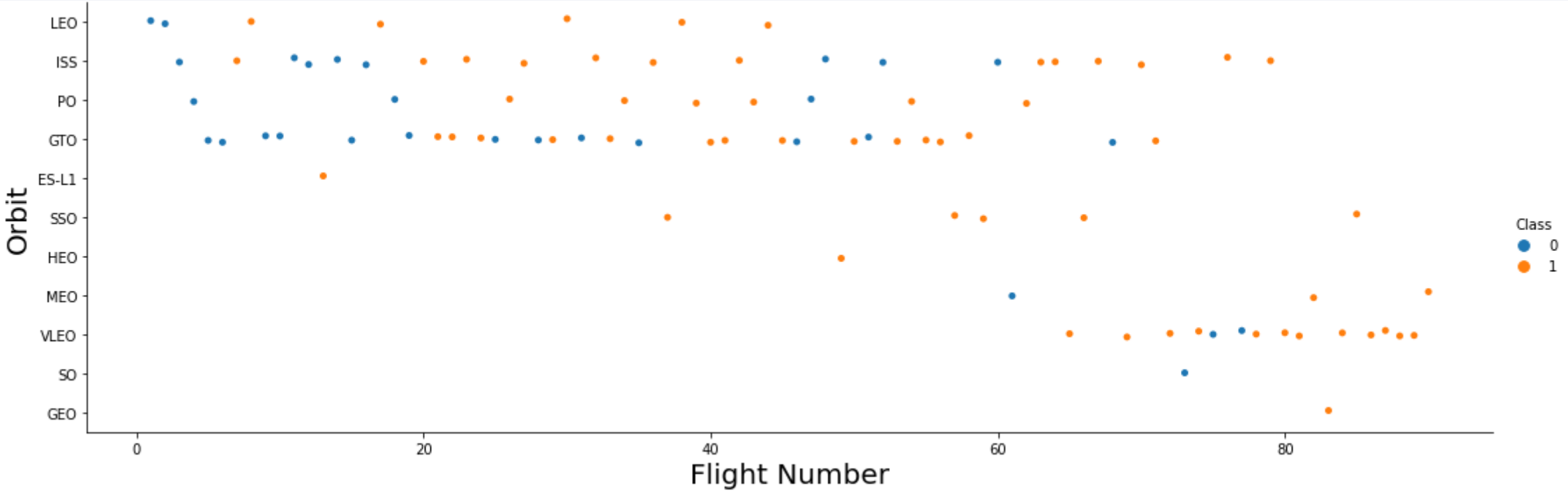
- Rockets launched with higher pay load mass tend to have a higher success rate.

Success Rate vs. Orbit Type

- Orbit types of ESL1, GEO, HEO, and SSO have the highest success rate of 100 percent.

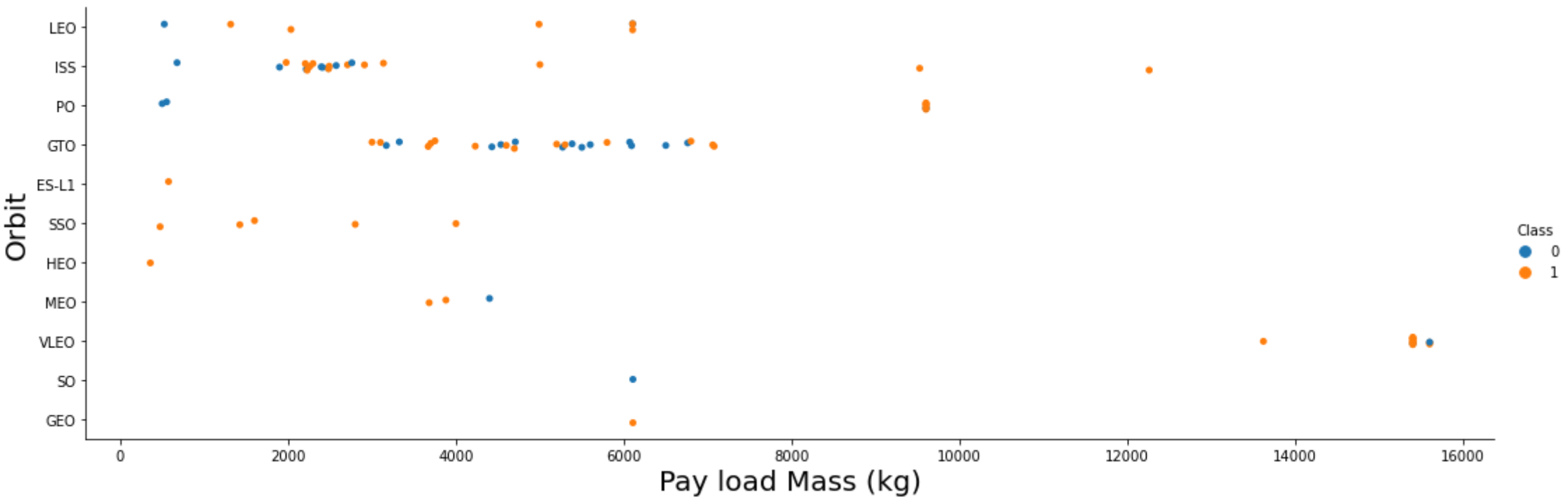


Flight Number vs. Orbit Type



- In the LEO orbit, the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

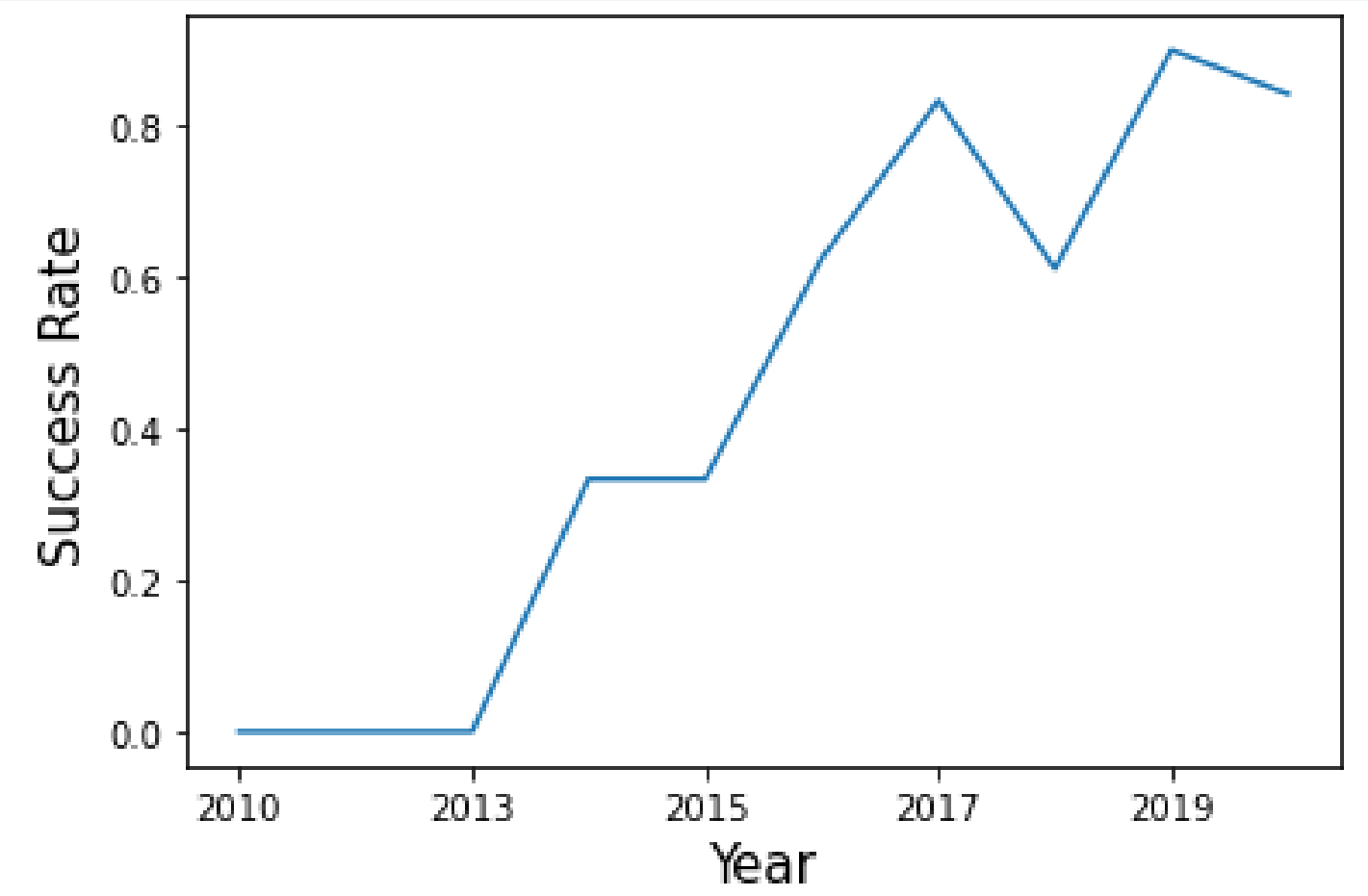
Payload vs. Orbit Type



- With heavy payloads the successful landing or 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(unsuccessful mission) are both there here.

Launch Success Yearly Trend

- The success rate since 2013 kept increasing till 2020.



All Launch Site Names

```
%%sql
select distinct launch_site from spacex
```

```
 * ibm_db_sa://zmn69768:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

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

Launch Site Names Begin with 'CCA'

```
%%sql
select * from spacex where launch_site like 'CCA%' limit 5
```

```
 * ibm_db_sa://zmn69768:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/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

```
%%sql
select sum(payload_mass__kg_) from spacex
```

```
 * ibm_db_sa://zmn69768:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

1
619967

- The total payload carried by boosters from NASA is 619,967 kg

Average Payload Mass by F9 v1.1

```
%%sql
select avg(payload_mass__kg_) from spacex where booster_version = 'F9 v1.1'
```

```
 * ibm_db_sa://zmn69768:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

1
2928

- The average payload mass carried by booster version F9 v1.1 is 2,928 kg.

First Successful Ground Landing Date

```
%%sql
select min(date) from spacex where landing__outcome = 'Success (ground pad)'
```

```
 * ibm_db_sa://zmn69768:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

1
2015-12-22

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

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
select distinct booster_version from spacex where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ >4000 and payl
oad_mass__kg_ <6000
```

```
 * ibm_db_sa://zmn69768:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

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

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are F9 FT B1021.2, F9 FT B1031.2, F9 FT B1022, and F9 FT B1026.

Total Number of Successful and Failure Mission Outcomes

```
%%sql
select 'Success' as Outcomes, count(*) as Count from spacex where landing__outcome like 'Success%'
union
select 'Failure' as Outcomes, count(*) as Count from spacex where landing__outcome like 'Failure%'
```

```
 * ibm_db_sa://zmn69768:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

outcomes	COUNT
Failure	10
Success	61

- With all missions in the dataset with a record, 61 was successful and 10 has failed.

Boosters Carried Maximum Payload

```
%%sql
select distinct booster_version from spacex where payload_mass__kg_ = (select max(payload_mass__kg_) from spacex)
```

```
 * ibm_db_sa://zmn69768:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
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

- The above list shows all the names of the booster which have carried the maximum payload mass

2015 Launch Records

```
%%sql
select booster_version, launch_site from spacex where landing__outcome = 'Failure (drone ship)' and year(date) = 2015
```

```
 * ibm_db_sa://zmn69768:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- F9 v1.1 B1012 and F9 v1.1 B1012 failed landing in drone ship in CCAFS LC-40 launch site in 2015.

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

```
%%sql
select landing__outcome, count(*) as count
from spacex
where (date > to_date('2010-06-04','YYYY-MM-DD')) and (date < to_date('2017-03-20','YYYY-MM-DD'))
group by landing__outcome
order by count desc
```

```
 * ibm_db_sa://zmn69768:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

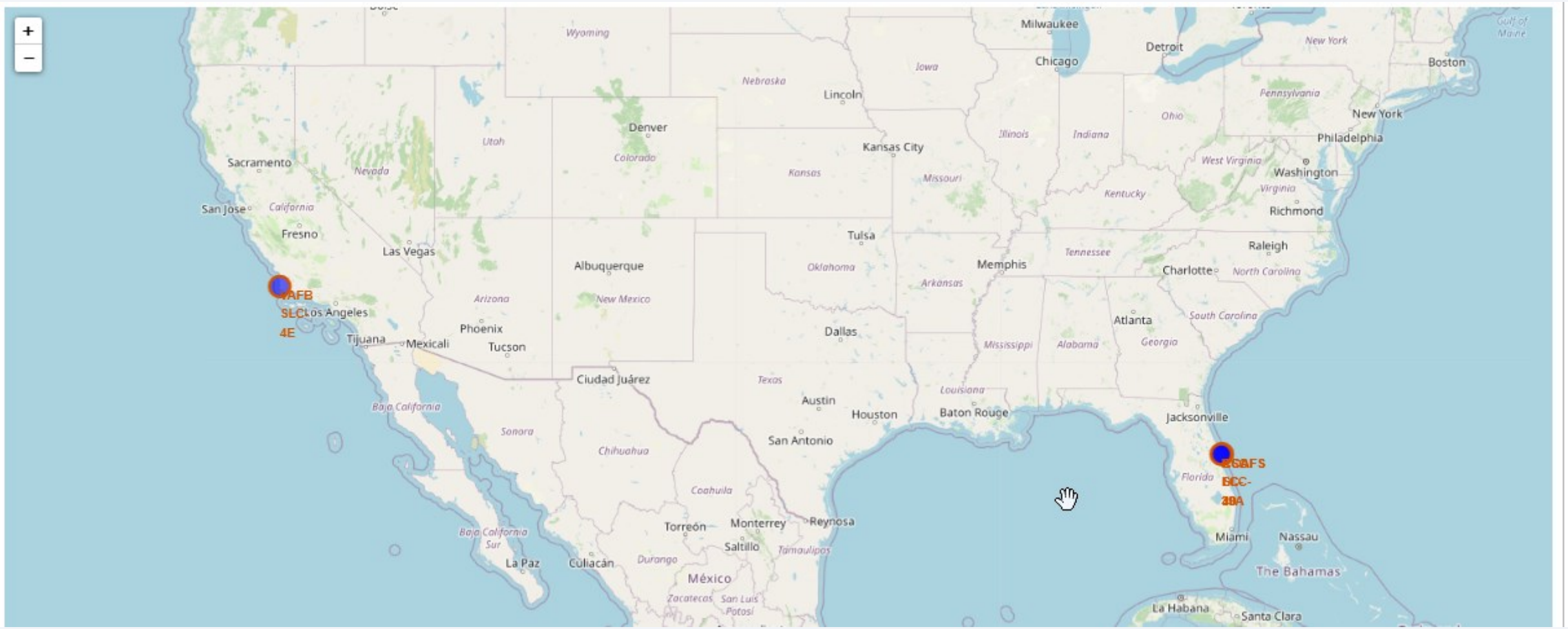
landing__outcome	COUNT
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Uncontrolled (ocean)	2
Failure (parachute)	1
Precluded (drone ship)	1

Landing on drone ships generates the highest number of successes as well as failures.

Section 3

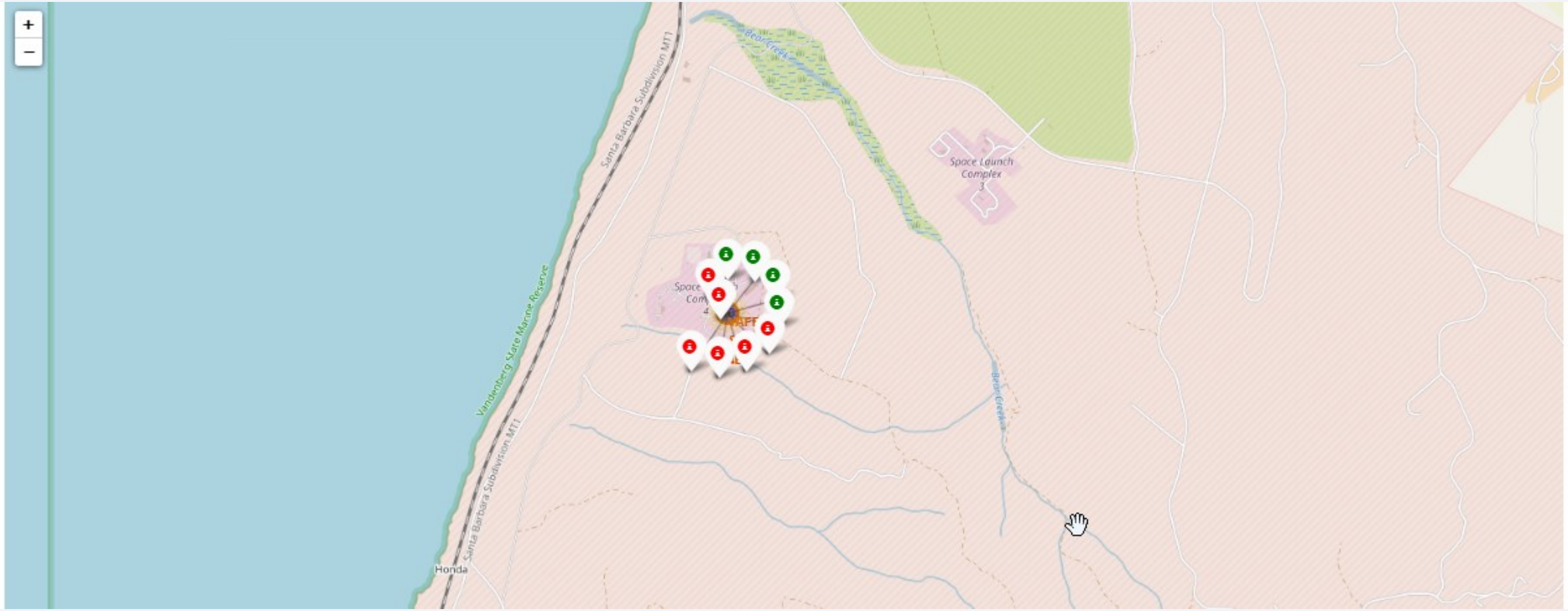
Launch Sites Proximities Analysis

The Launch Sites



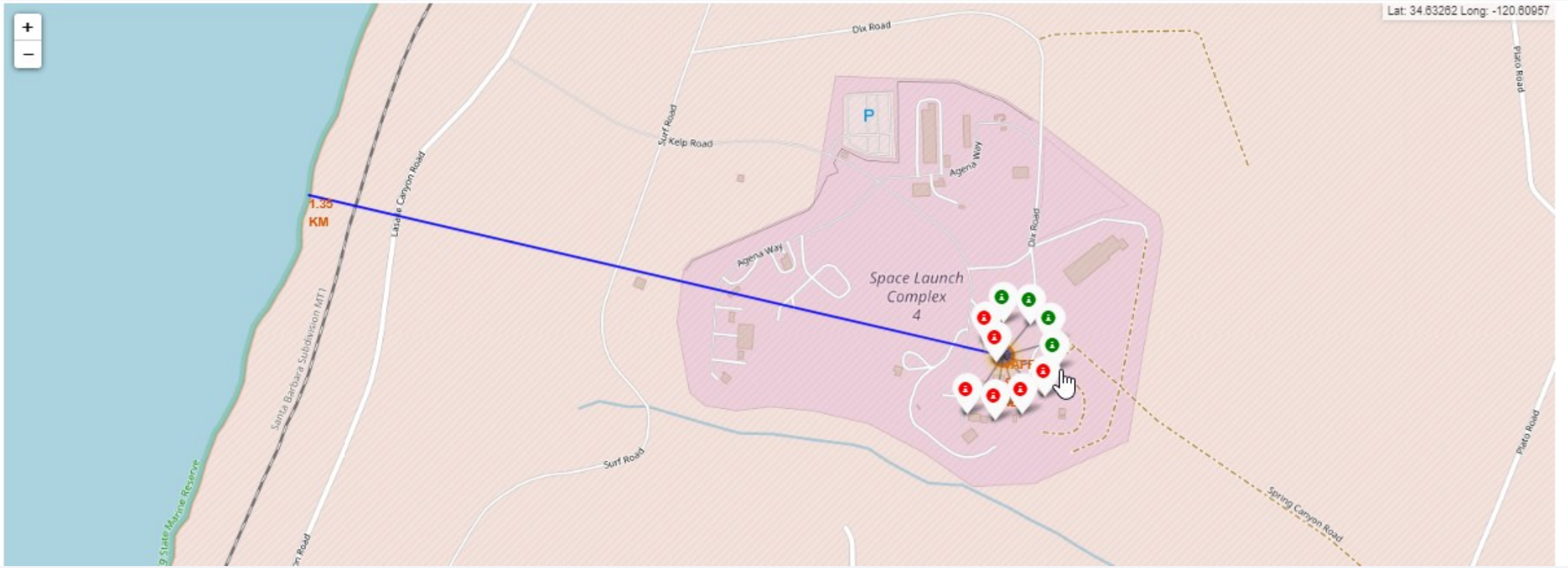
- Out of four launch sites displayed, one is in California, and three are in Florida.

Launch Outcomes at Site VAFB SLC-4E

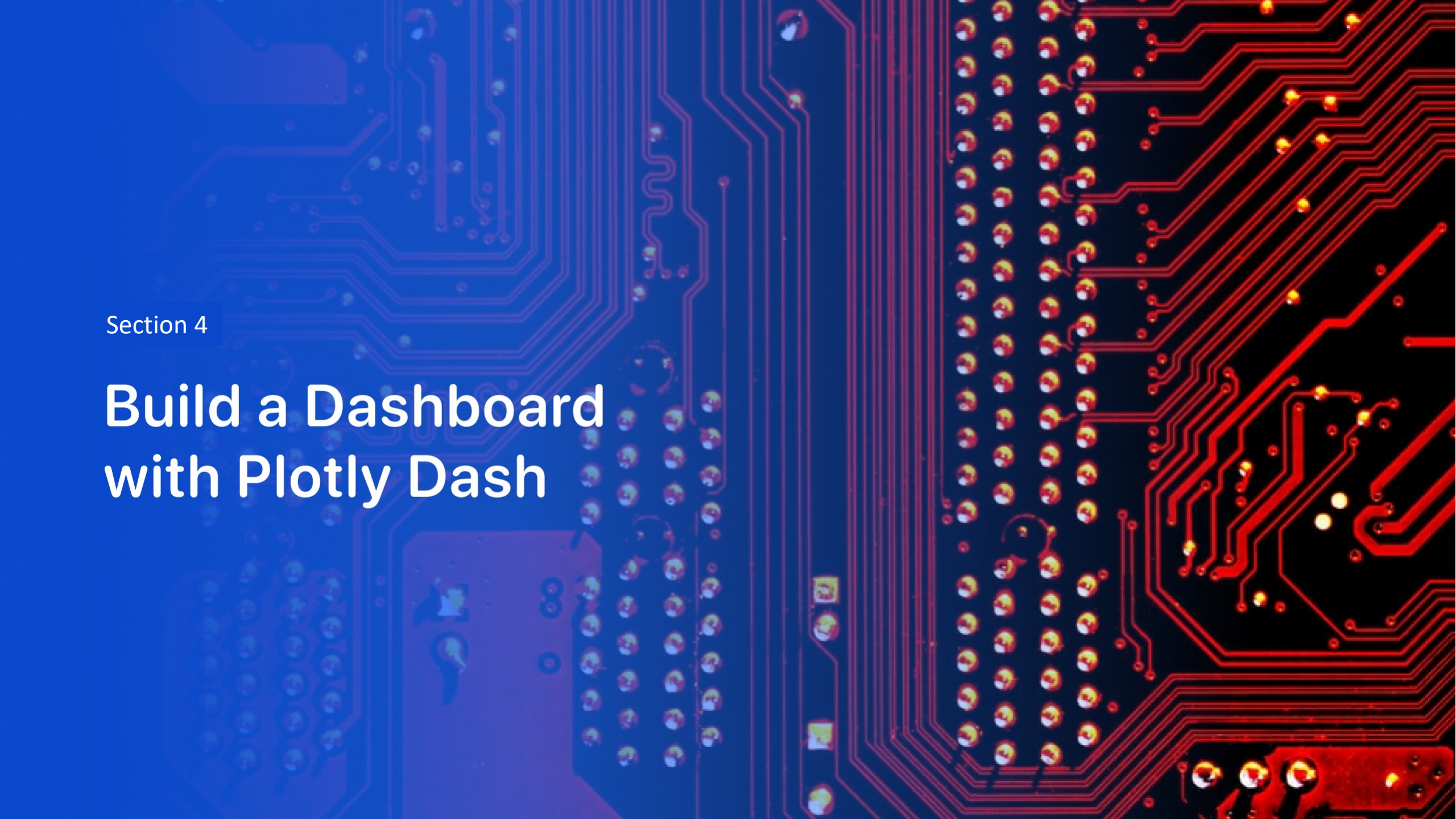


- At the California Site VAFB SLC-4E, out of 10 launch attempts, four were successful (green)

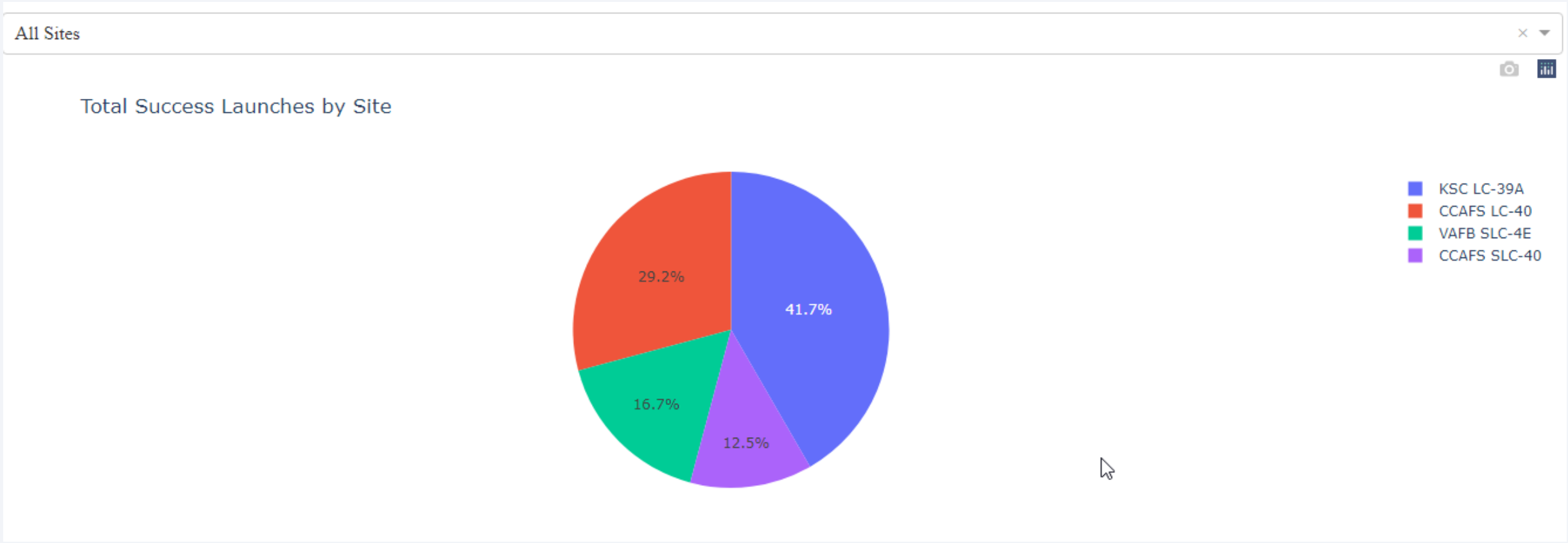
Distance Between Site VAFB SLC-4E and its Closest Coastline



- The California Site VAFB SLC-4E is 1.35 km away from its nearest coastline.

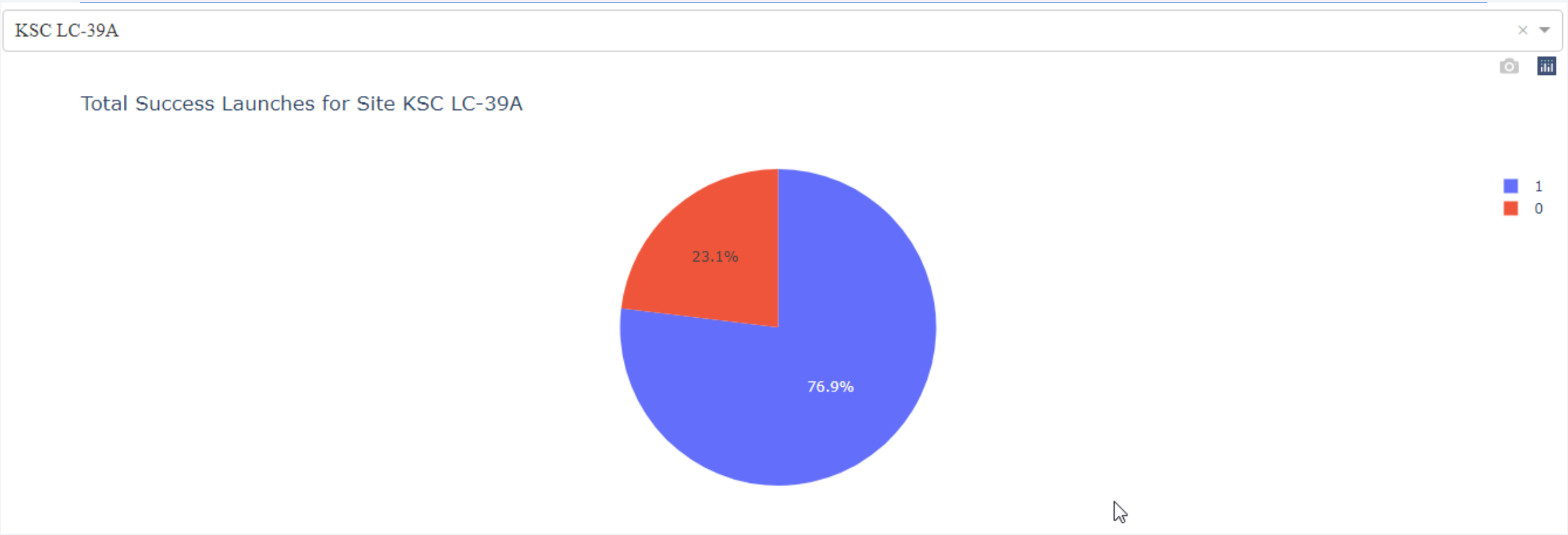

Section 4
Build a Dashboard with Plotly Dash

Success Launches by Sites



- 41.7 percent of success launches are from Site KSC LC-39A, highest among all sites

Success Rate of Site KSC LC-39A

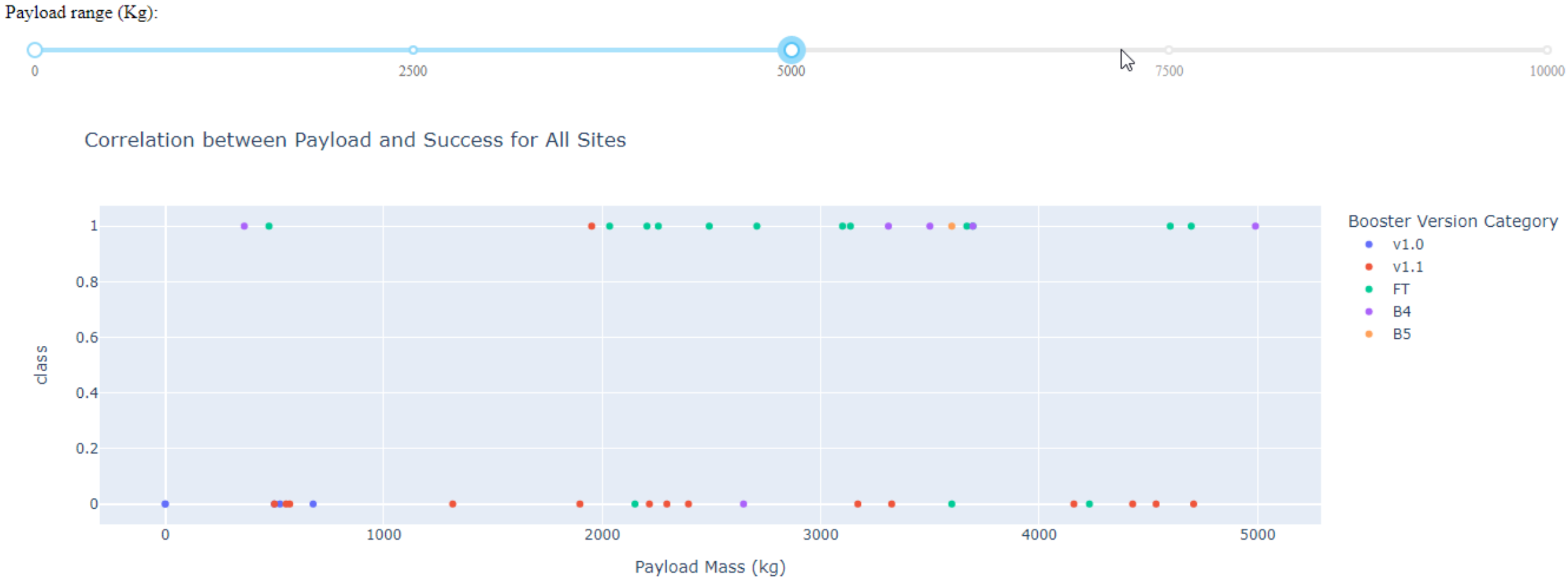


- 76.9 percent of launches from KSC LC-39A were successful.

Payload vs. Launch Outcome scatter plot 1

- Most successful launches are with a payload between 2,000 kg and 6,000 kg

Payload vs. Launch Outcome scatter plot 2



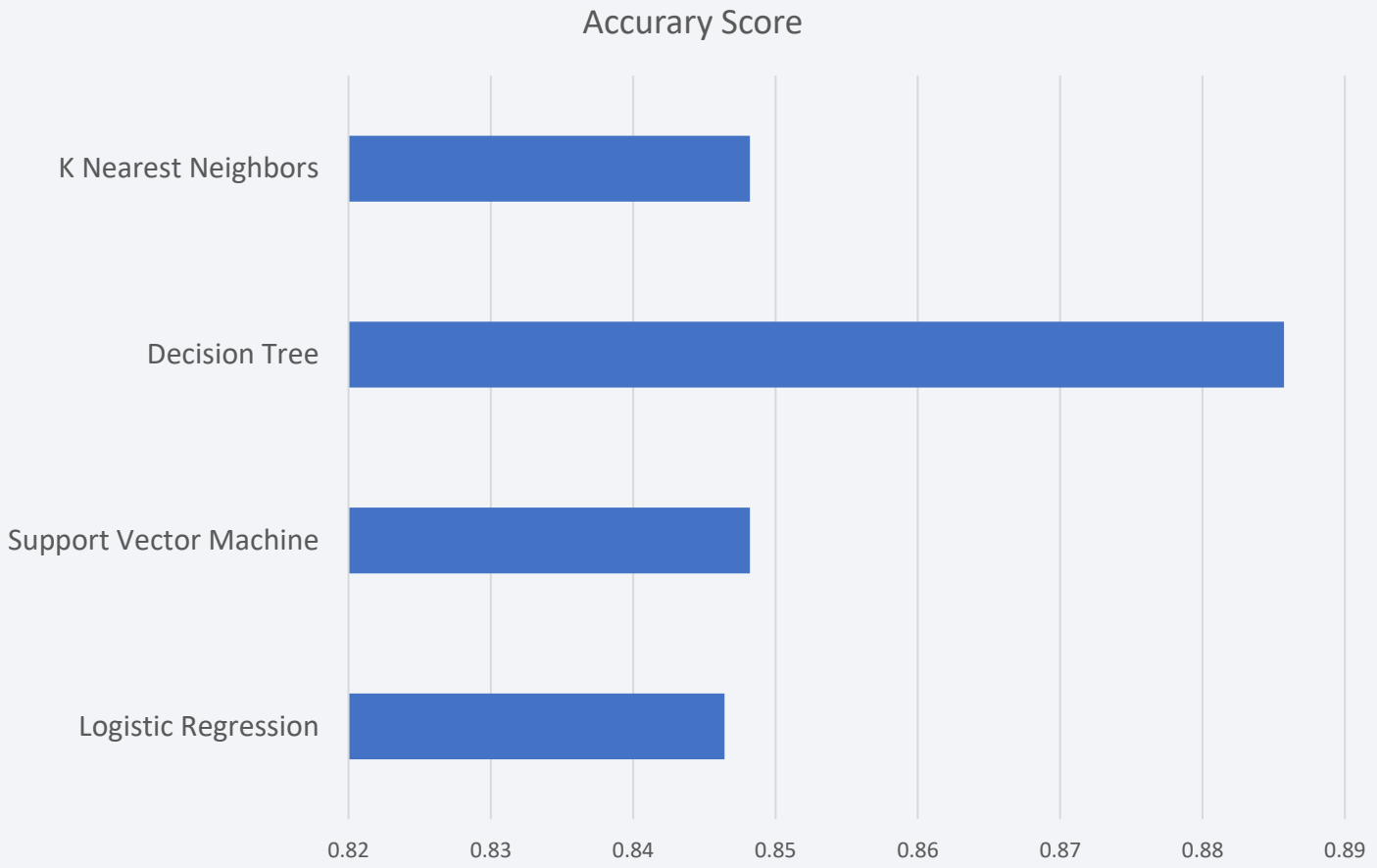
- Of the launches with lighter payloads (0 – 5,000 kg), the v1.1 Booster has a relatively low success rate.

Section 5

Predictive Analysis (Classification)

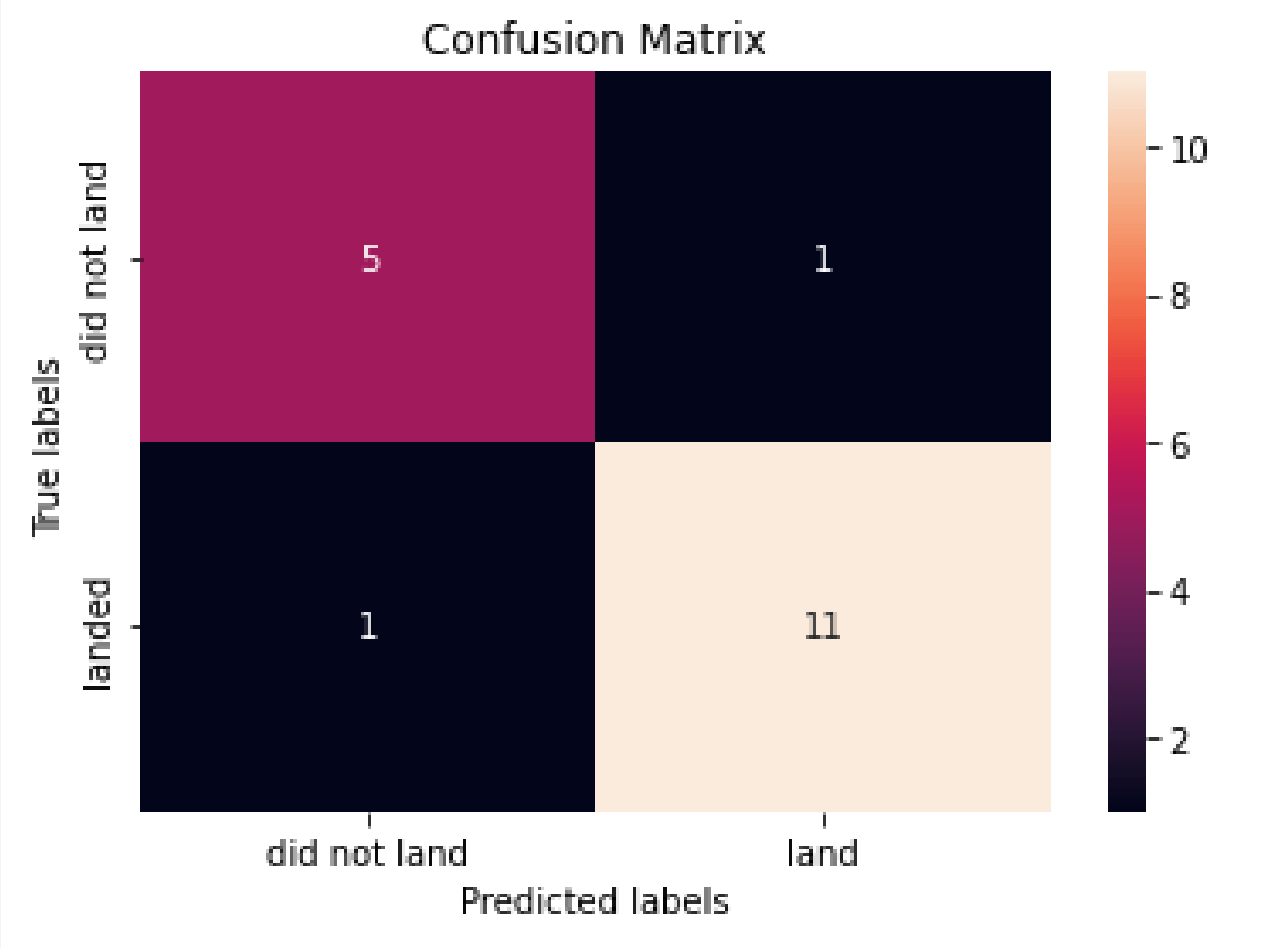
Classification Accuracy

- Decision tree is the most model for predicting the successfulness of rocket launch with an accuracy score of 0.88.



Confusion Matrix

- The confusion matrix shows that out of 18 records in the test dataset, the decision tree produces the best prediction, with 16 correct predictions, one false positive and one false negative.



Conclusions

- It is discovered that the success rate of rocket launch are closely related to payload, launch site, year, orbit type, and flight number.
- The success rate increased as the flight number increased and over time (experienced gain from testing).
- CCAFS SLC 40 held most of the recent launches, which have been successful.
- Rockets launched with higher pay load mass tend to have a higher success rate.
- Orbit types of ESL1, GEO, HEO, and SSO have the highest success rate of 100 percent.
- With heavy payloads the successful landing or 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(unsuccessful mission) are both there here.
- Of the launches with lighter payloads (0 – 5,000 kg), the v1.1 Booster has a relatively low success rate.
- Logistic regression, support vector machine, decision tree, and K nearest neighbors are selected to test the perform. Decision tree stands out being the most accurate model with an accuracy rate of 88 percent.

Appendix

- Dashboard: https://github.com/pyu999/Capstone_IBM_DS/blob/main/spacex_dash_app.py
- Data collection API: https://github.com/pyu999/Capstone_IBM_DS/blob/master/ML%20Final%20Assignment.ipynb
- Web scrapping: https://github.com/pyu999/Capstone_IBM_DS/blob/master/Web%20Scraping%20lab.ipynb
- Data Wrangling EDA: https://github.com/pyu999/Capstone_IBM_DS/blob/master/EDA%20Lab.ipynb
- EDA with visualization: https://github.com/pyu999/Capstone_IBM_DS/blob/master/EDA%20with%20Visualization.ipynb
- EDA with SQL: https://github.com/pyu999/Capstone_IBM_DS/blob/master/EDA%20with%20SQL.ipynb
- Map visualization: https://github.com/pyu999/Capstone_IBM_DS/blob/master/Interactive%20Visual%20Analytics.ipynb
- ML predictions: https://github.com/pyu999/Capstone_IBM_DS/blob/master/ML%20Prediction.ipynb

Thank you!