IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Toch Davy
11/10/2024

Outline

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

Executive Summary

- Methodology
  - Data collection :
    - Web scraped of public Wikipedia pages related to SpaceX launch history
    - Consulted SpaceX public REST API
  - Data wrangling:
    - Only one added Feature (‘Class’) added indicating landing outcome (0 or 1)
  - Exploratory data analysis:
    - Historic data of SpaceX launches was loaded in database for SQL query exploration
    - Different visualizations performed (scatter, bar, line plots, interactive dashboards, geolocation maps)
  - Predictive analysis:
    - Different models were evaluated (Logistic Regression, k-Nearest Neighbor, Decision Tree, Support Vector Machine)
- Conclusion
  - Decision Tree Model is the most accurate model to predict landing outcome (‘Class’ Feature)

Introduction

- SpaceX is one of the first rocket launch projects that has commercial success
- Important factors to be able to outcompete other programs in the future:
  - Know which factors (location, weather, rocket type, …) contribute to highest launch/landing rate
  - Predict future landing success rate rockets to maximize reuse for multiple launches

Section 1

Methodology

Methodology

- Data collection :
  - Web scraped of public Wikipedia pages related to SpaceX launch history
  - Consulted SpaceX public REST API
- Data wrangling:
  - Only one added Feature (‘Class’) added indicating landing outcome (0 or 1)
- Exploratory data analysis:
  - Historic data of SpaceX launches was loaded in database for SQL query exploration
  - Different visualizations performed (scatter, bar, line plots, interactive dashboards, geolocation maps)
- Predictive analysis:
  - Different models were evaluated (Logistic Regression, k-Nearest Neighbor, Decision Tree, Support Vector Machine)

Data Collection

- Data collection was performed by
  - Web scraping public Wikipedia pages related to SpaceX launch history
  - Consuming SpaceX public REST APIs

Data Collection – SpaceX API

- API calls performed:
  - Get historic general data of launches (GET https://api.spacexdata.com/v4/launches/past)
  - For each general data record, get details of :
    - Used Rocket Booster Version (GET https://api.spacexdata.com/v4/rockets)
    - Used Launch pad (GET https://api.spacexdata.com/v4/launchpads)
    - Payload details (GET https://api.spacexdata.com/v4/payloads)
    - Rocket core details (GET https://api.spacexdata.com/v4/cores)
- Extracted information:

Flight Number	Flight Date	Rocket Booster Version	Payload Mass	Orbit	Launch Site
Outcome	Flights	Grid Fins	Reused	Legs	Landing Pad
Block	Reused Count	Serial	Longitude	Latitude	

- GitHub URL : https://github.com/sirasen/sandbox/blob/main/data_science_lab/Capstone_01_Data_Collection.ipynb

Data Collection - Scraping

- Used BeautifulSoup library to web scrape tables with launch history from https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- Extracted information:

Flight Number	Flight Date	Flight Time	Rocket Booster Version
Launch Site	Payload Type	Payload Mass	Orbit
Customer	Launch outcome	Booster landing	

- GitHub URL : https://github.com/sirasen/sandbox/blob/main/data_science_lab/Capstone_01_Data_Collection.ipynb

Data Wrangling

- Data wrangling steps:
  - Only one additional Feature ‘Class’ was added based on Feature ‘Outcome’
    - 0 : unsuccessful landing outcome
    - 1 : successful landing outcome
- GitHub URL : https://github.com/sirasen/sandbox/blob/main/data_science_lab/Capstone_02_Data_Wrangling.ipynb

EDA with Data Visualization

- Charts that were plotted:
  - Flight Number vs Launch Site (scatter plot)
  - Payload mass vs Launch Site (scatter plot)
  - Orbit Type vs Success Rate (bar plot)
  - Flight number vs Orbit Type (scatter plot)
  - Payload mass vs Orbit Type (scatter plot)
  - Launch success yearly trend (line plot)
- GitHub URL : https://github.com/sirasen/sandbox/blob/main/data_science_lab/Capstone_03_Exploratory_Analysis.ipynb

EDA with SQL

- Performed SQL queries:
  - List of all launch sites
  - List of 5 launch sites whose name starts with ‘CCA’
  - Total payload of rockets for NASA (CRS) as customer
  - Average payload of F9 v1.1 rockets
  - Date when the first rocket successfully landed on a ground pad
  - Rocket booster versions that landed on a drone ship with payload of 4 to 6 metric tons
  - Total number of successful and failed landings
  - Rocket booster versions that carried the maximum payload
  - For 2015, month, booster version and launch site of the failed landings on a drone ship
  - Ranking of landing outcomes in period 04/06/2010 – 20/03/2017
- GitHub URL: https://github.com/sirasen/sandbox/blob/main/data_science_lab/Capstone_03_Exploratory_Analysis.ipynb

Build an Interactive Map with Folium

- Created Folium map of North America with:
  - Markers indicating locations of the launch sites
  - Circles indicating the launch success count for each launch site
- GitHub URL : https://github.com/sirasen/sandbox/blob/main/data_science_lab/Capstone_04_Interactive_Map.ipynb

Build a Dashboard with Plotly Dash

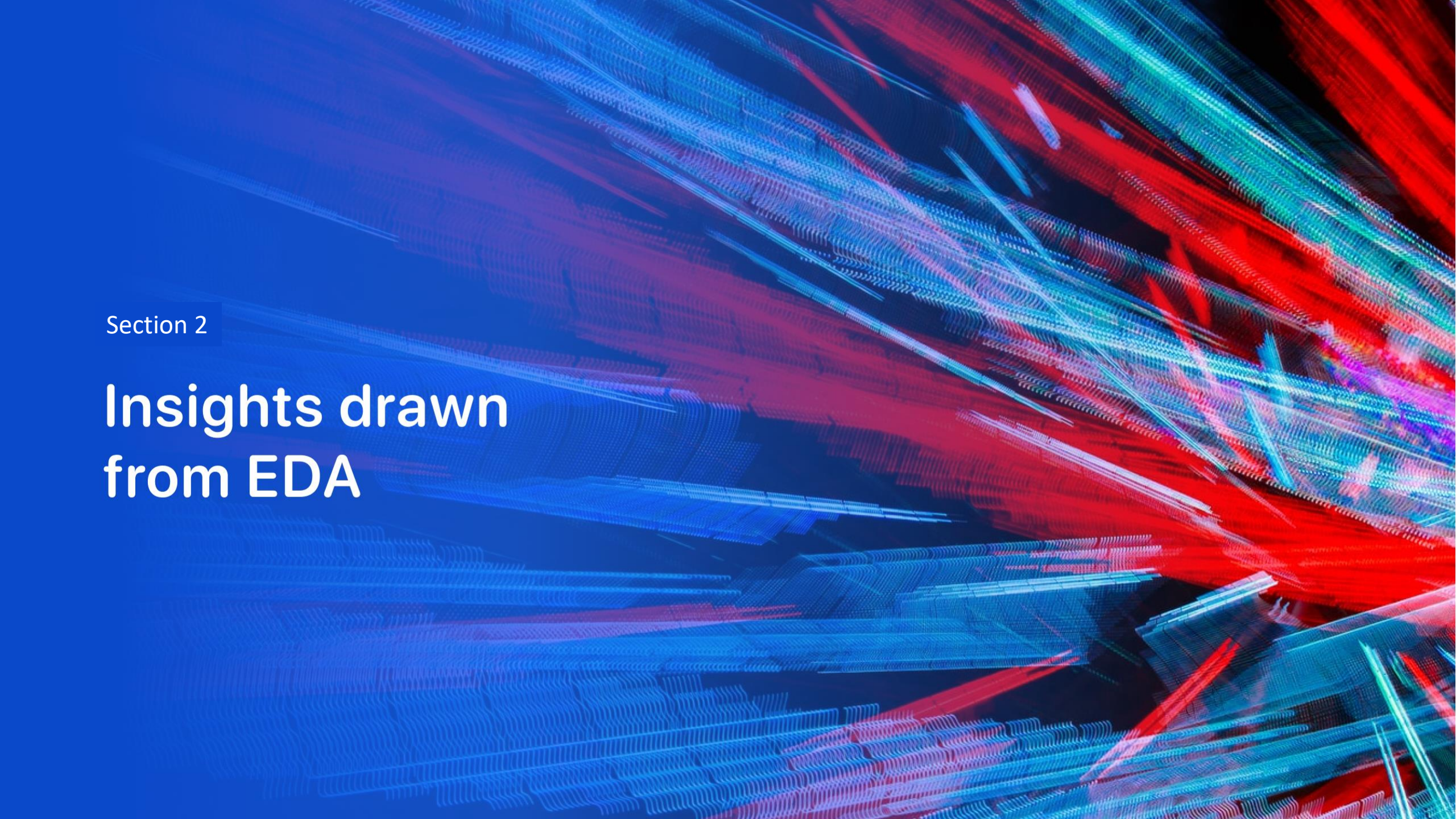
- Created a Dashboard with:
  - For all launch sites:
    - A pie chart showing the ratio of launches for each launch site
    - A scatter plot showing the correlation between payload mass and launch success rate
  - For each launch site:
    - A pie chart showing the ratio between successful and failed launches
    - A scatter plot showing the correlation between payload mass and launch success rate
- GitHub URL : https://github.com/sirasen/sandbox/blob/main/data_science_lab/Capstone_04_Interactive_Dashboard.py

Predictive Analysis (Classification)

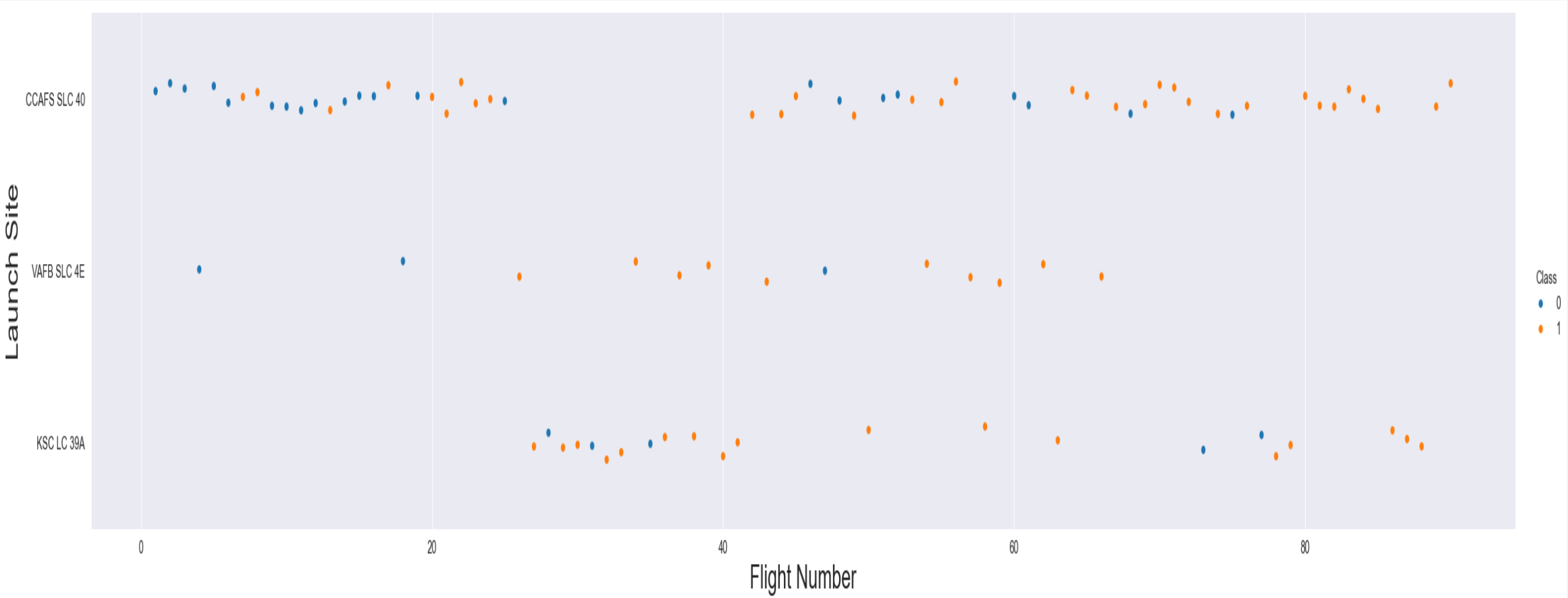
- Evaluated the following models:
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree
  - k-Nearest Neighbor
- Scikit-Learn GridSearchCV used to find optimal hyperparameters
- Decision Tree model :
  - Is the most accurate (86.2 % accuracy)
  - Optimal hyperparameters:
    - Criterion : gini
    - Maximum depth = 4
    - Minimum samples leaf = 1
    - Minimum samples split : 2
    - Spllitter = random
- GitHub URL : https://github.com/sirasen/sandbox/blob/main/data_science_lab/Capstone_05_Predictive_Analysis.ipynb

Results

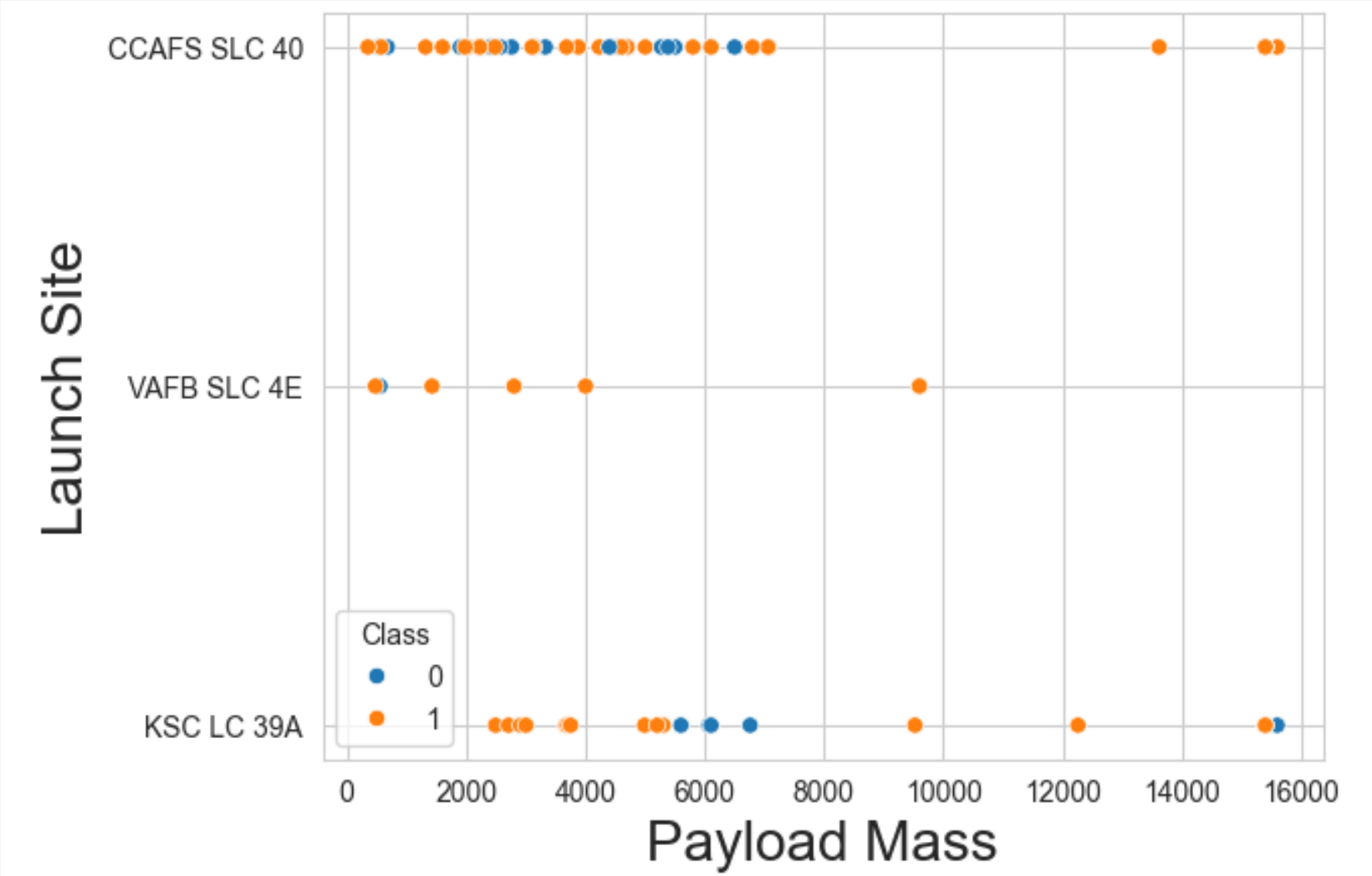
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn from EDA

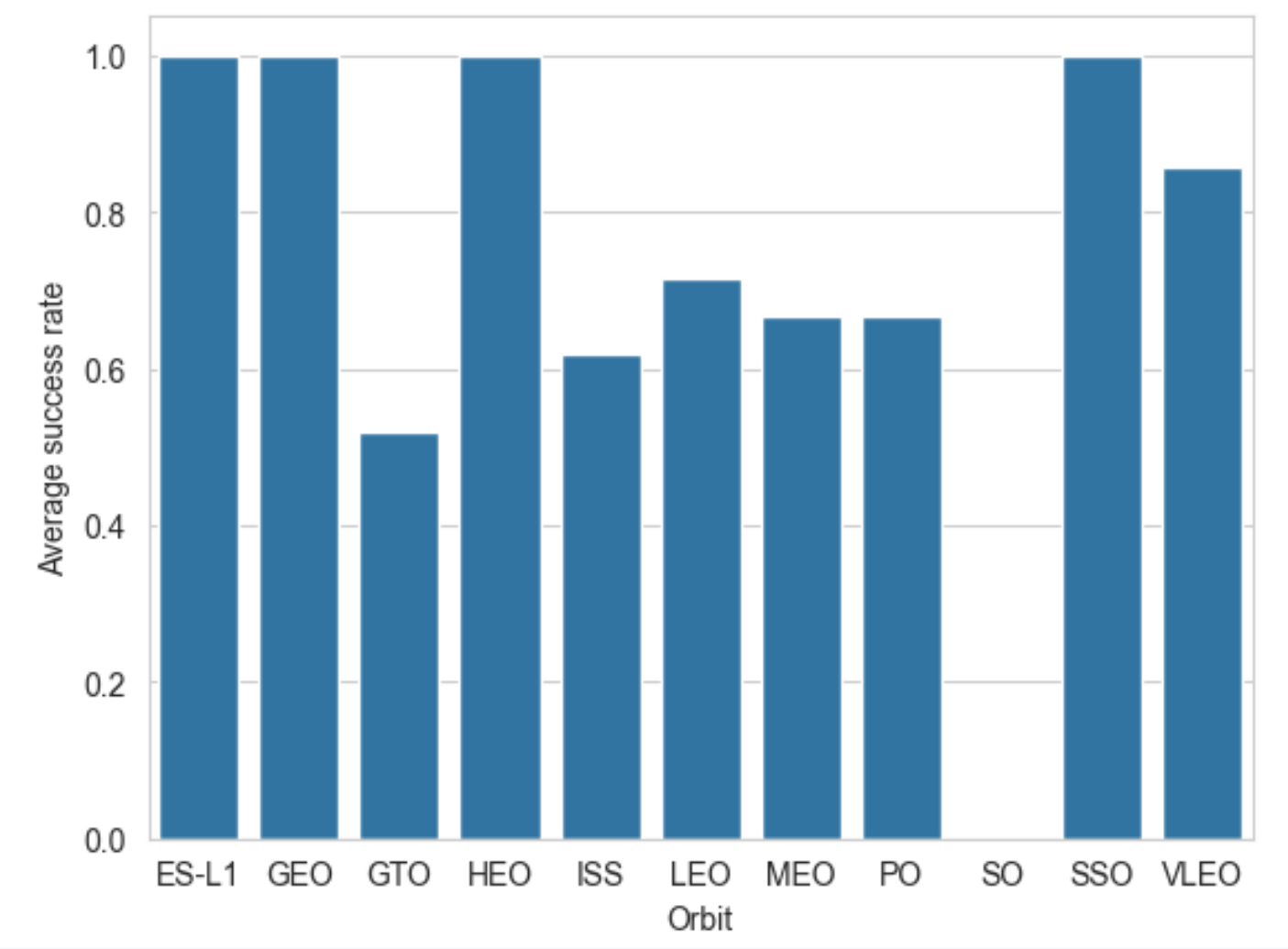
Flight Number vs. Launch Site



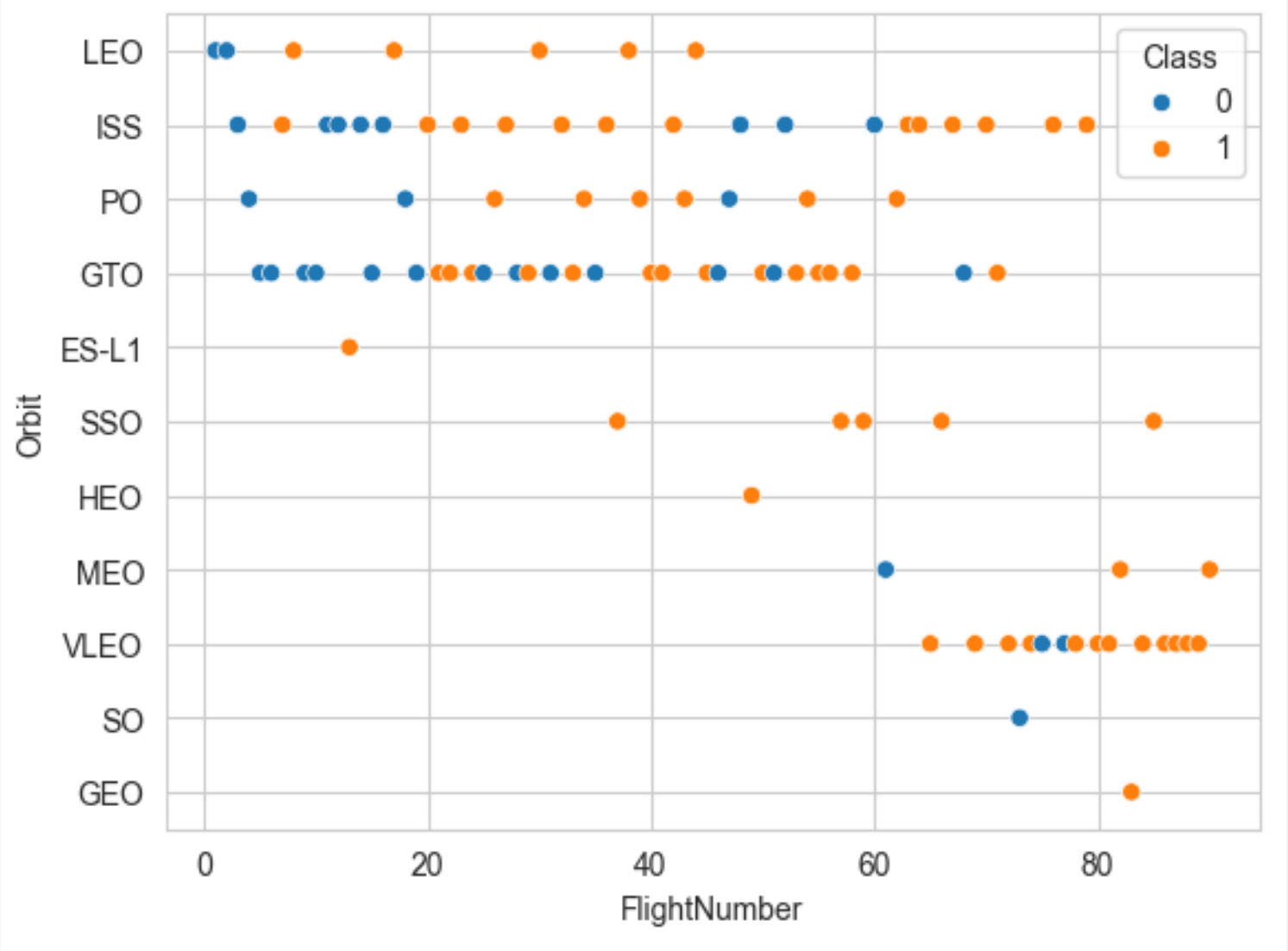
Payload vs. Launch Site



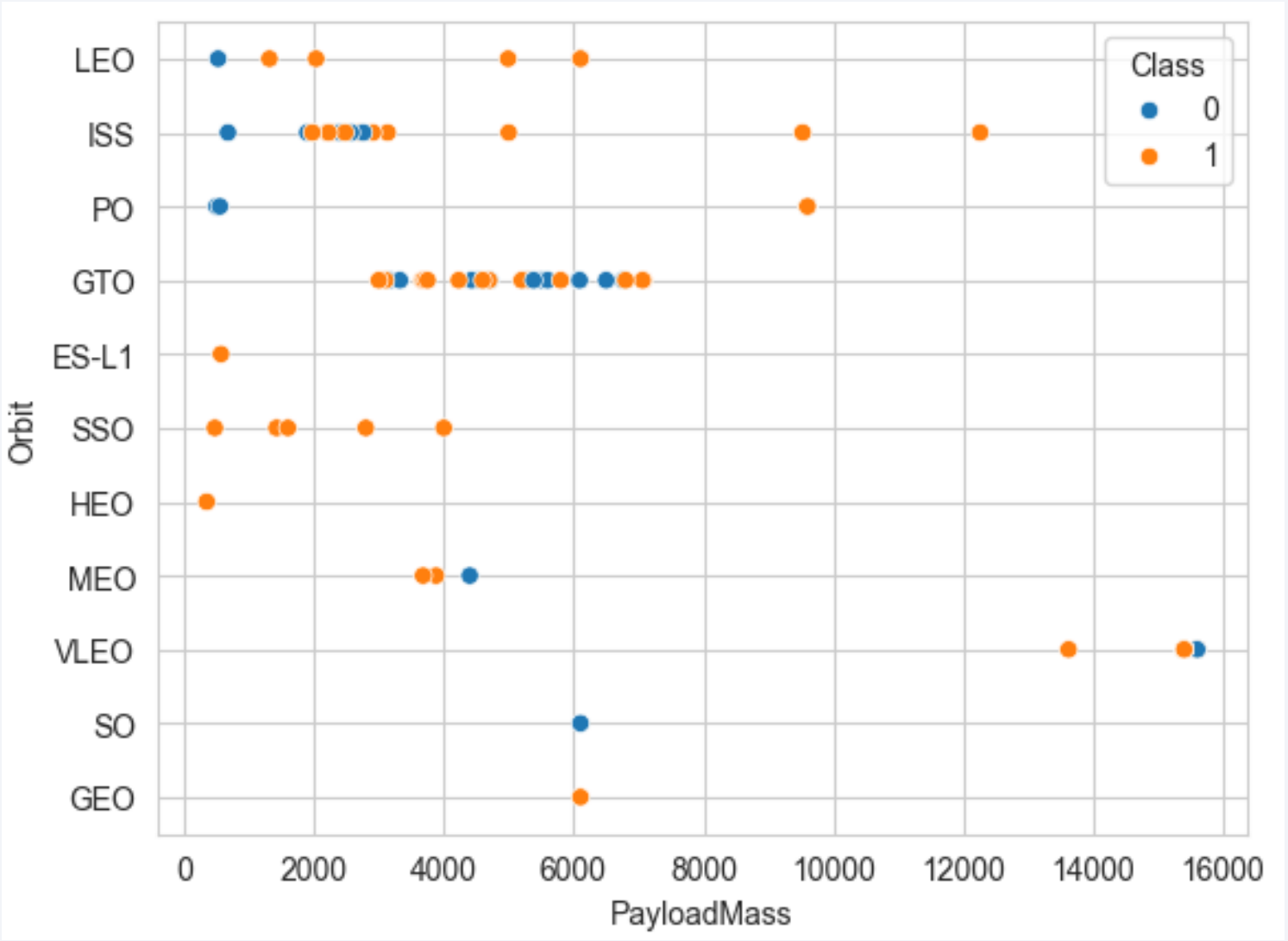
Orbit Type vs Success Rate



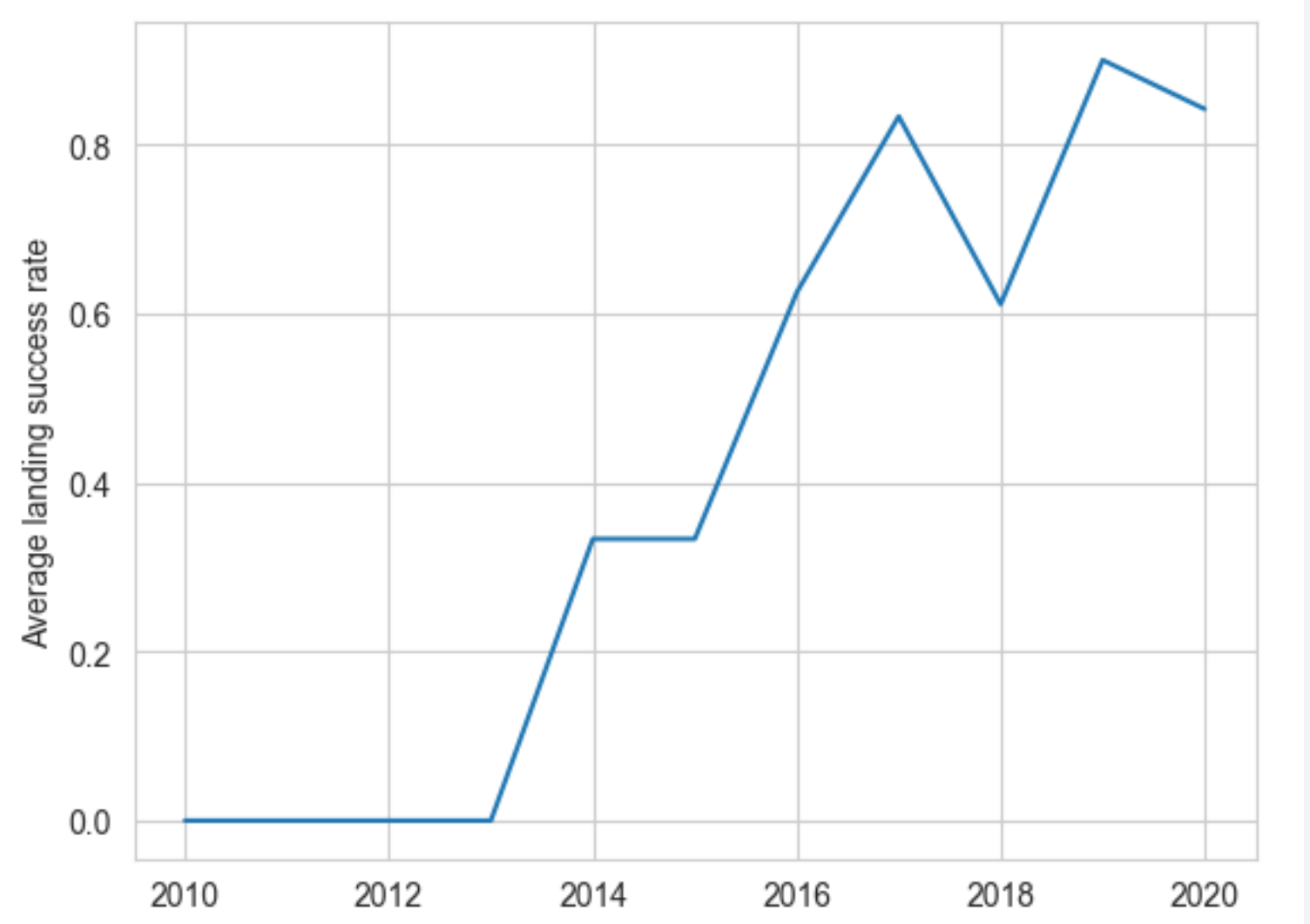
Flight Number vs. Orbit Type



Payload vs. Orbit Type



Launch Success Yearly Trend



All Launch Site Names

```
%sql SELECT DISTINCT(Launch_Site) FROM SPACEXTABLE;
 [12] < 10 ms
   * sqlite:///my_data.db
  Done.
```

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

Launch Site Names Begin with 'CCA'

```
%sql SELECT DISTINCT(Launch_Site) FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
```

✓ [15] < 10 ms

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

Launch_Site
CCAFS LC-40
CCAFS SLC-40

Total Payload Mass of boosters from NASA

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer LIKE '%NASA (CRS)';
✓ [17] < 10 ms
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

```
%sql SELECT ROUND(AVG(PAYLOAD_MASS__KG_), 1) FROM SPACEXTABLE WHERE Booster_Version like 'F9 v1.1%'
```

✓ [20] < 10 ms

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

ROUND(AVG(PAYLOAD_MASS__KG_),1)
2534.7

First Successful Ground Landing Date

```
%sql SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)';
✓ [21] < 10 ms

 * sqlite:///my_data.db
Done.

 MIN(Date)

2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ >
 4000 AND PAYLOAD_MASS__KG_ < 6000;
✓ [22] < 10 ms
```

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

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

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT COUNT(Landing_Outcome) FROM SPACEXTABLE WHERE Landing_Outcome LIKE '%Failure%' OR Landing_Outcome LIKE ;
'%Success%';
✓ [25] < 10 ms

 * sqlite:///my_data.db
Done.

COUNT(Landing_Outcome)
                    71
```

Boosters Carried Maximum Payload

```
%sql SELECT DISTINCT(Booster_Version) FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ IN (SELECT MAX(PAYLOAD_MASS__KG_) FROM
  SPACEXTABLE);
✓ [26] 10ms
  * sqlite:///my_data.db
 Done.
```

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

```
%sql SELECT substr(Date, 6, 2) AS Month, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE Landing_Outcome =
 'Failure (drone ship)' AND substr(Date, 0, 5) = '2015';
✓ [27] < 10 ms
```

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

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

```
%sql SELECT Landing_Outcome, COUNT(Landing_Outcome) FROM SPACEXTABLE WHERE Date >= '2010-06-04' AND Date <=
 '2017-03-20' GROUP BY Landing_Outcome  ORDER BY COUNT(Landing_Outcome) DESC;
✓ [44] < 10 ms
```

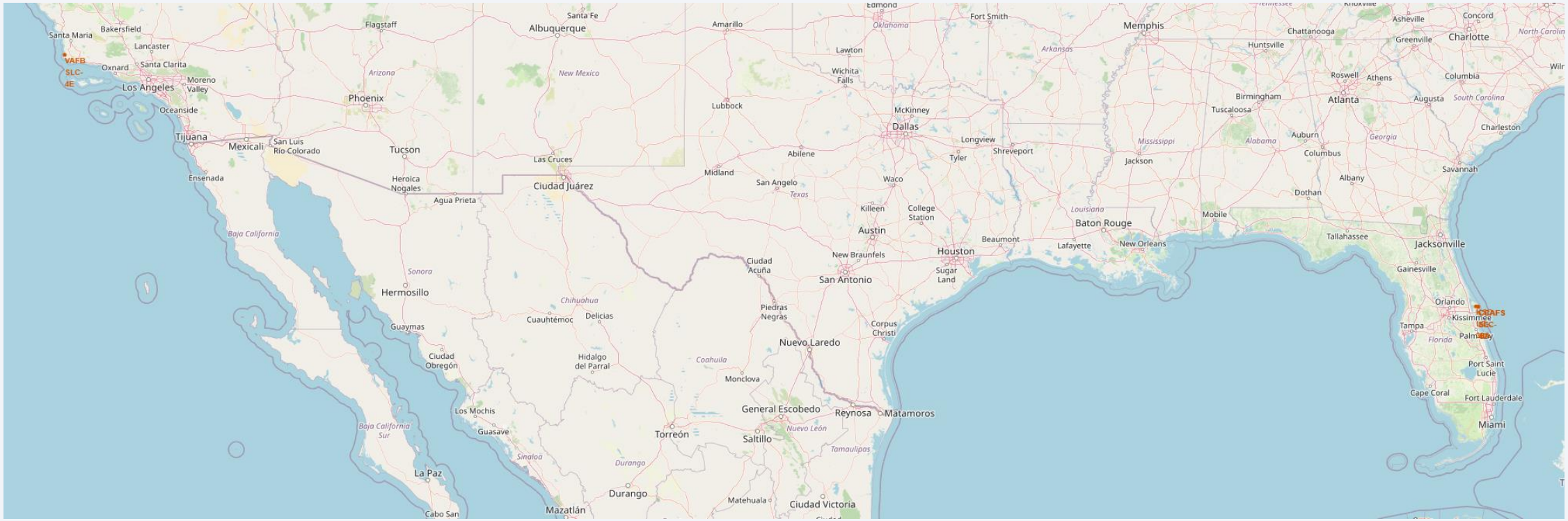
```
 * sqlite:///my_data.db
Done.
```

Landing_Outcome	COUNT(Landing_Outcome)
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

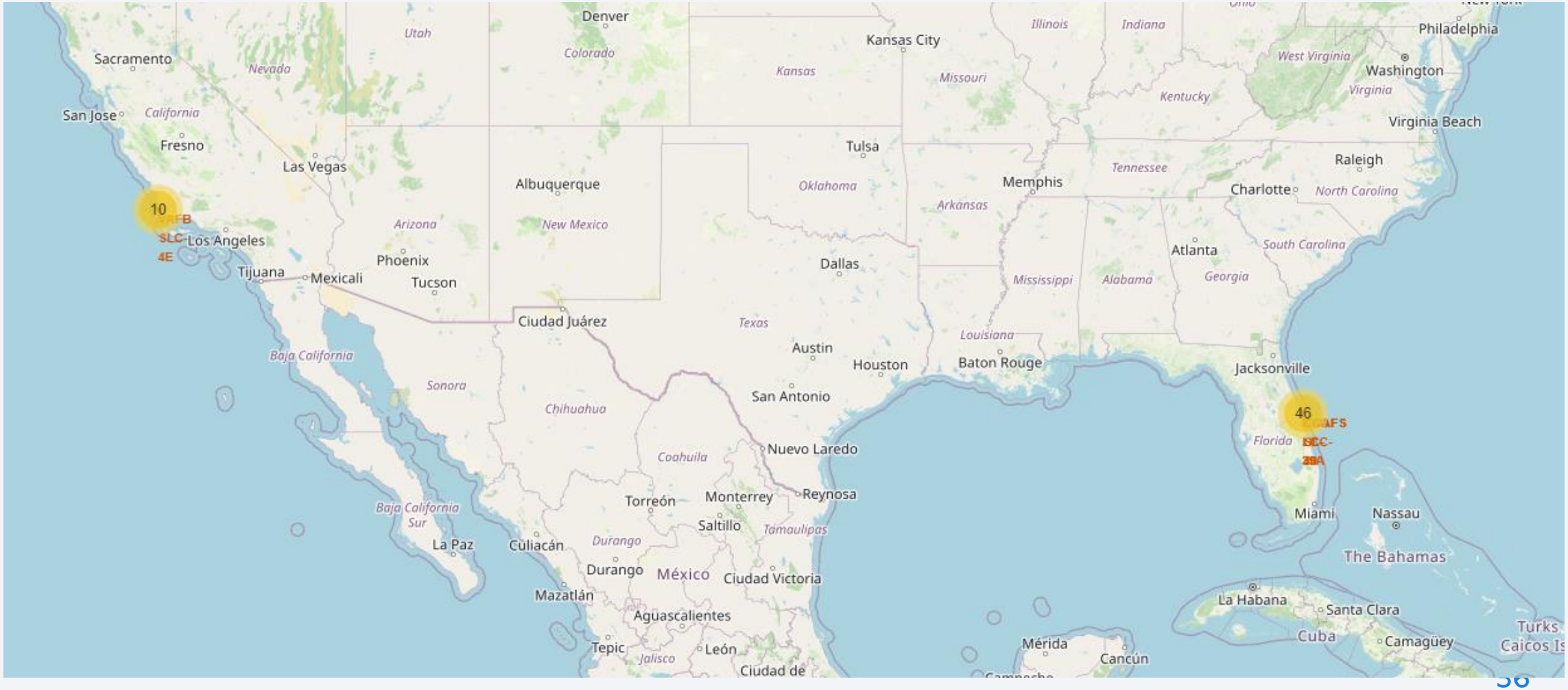
Section 3

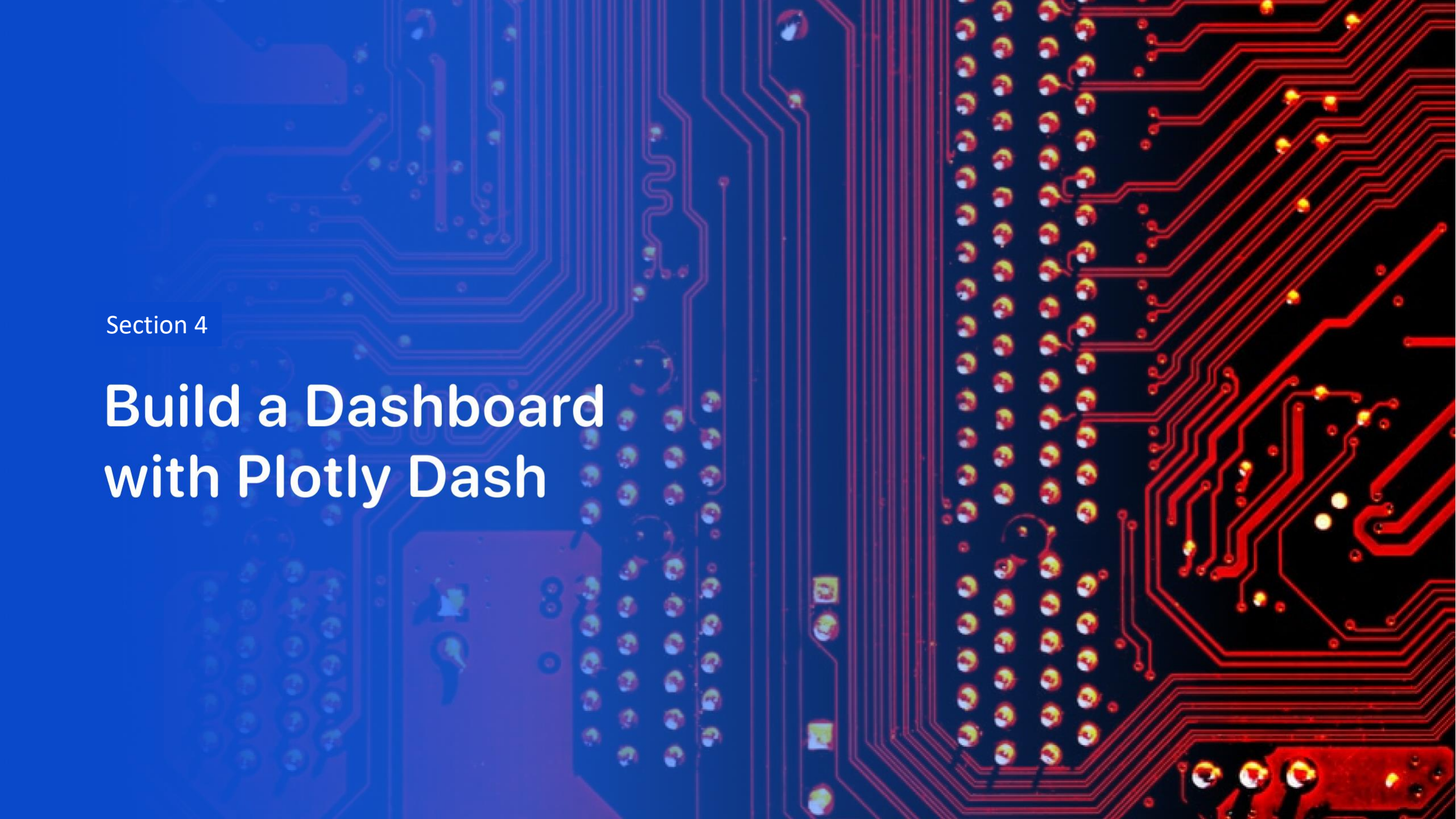
Launch Sites Proximities Analysis

Main launch sites

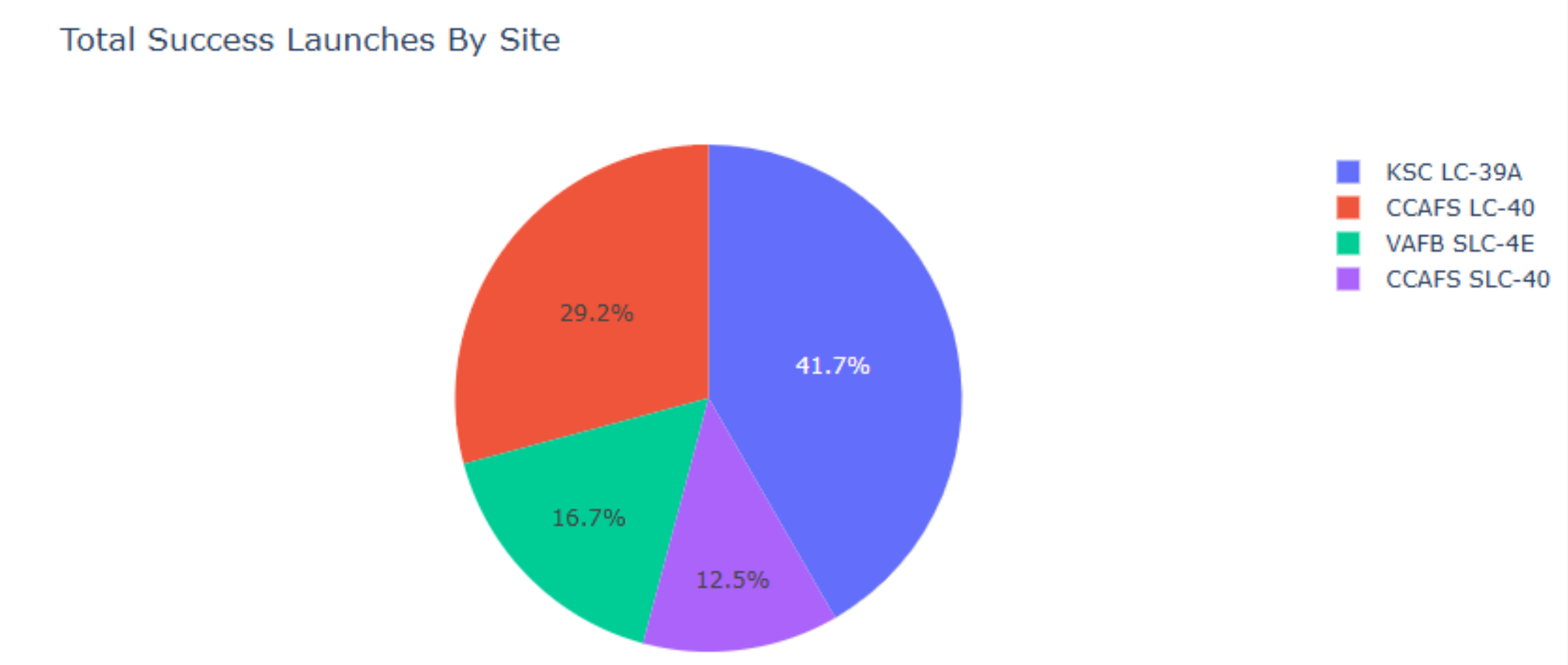


Main launch sites – launch outcomes

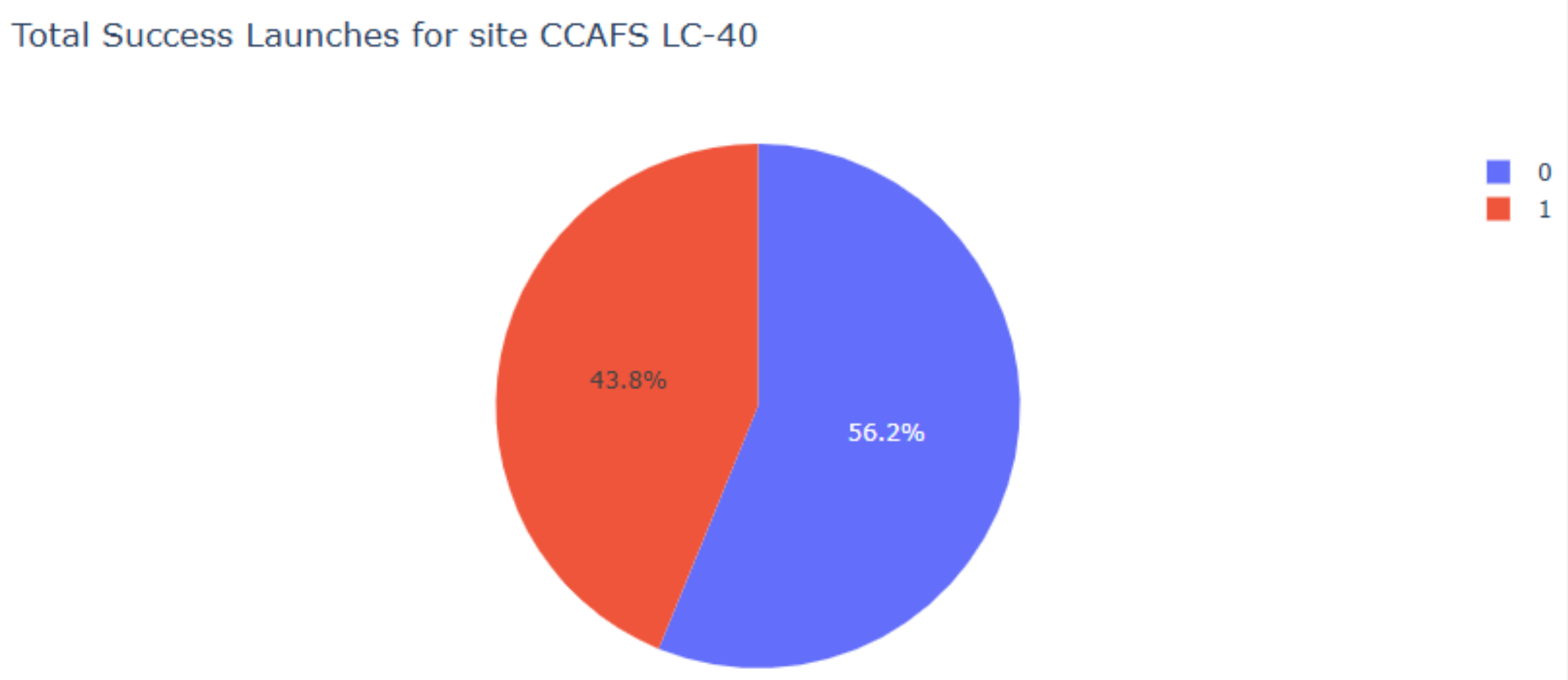



Section 4
Build a Dashboard with Plotly Dash

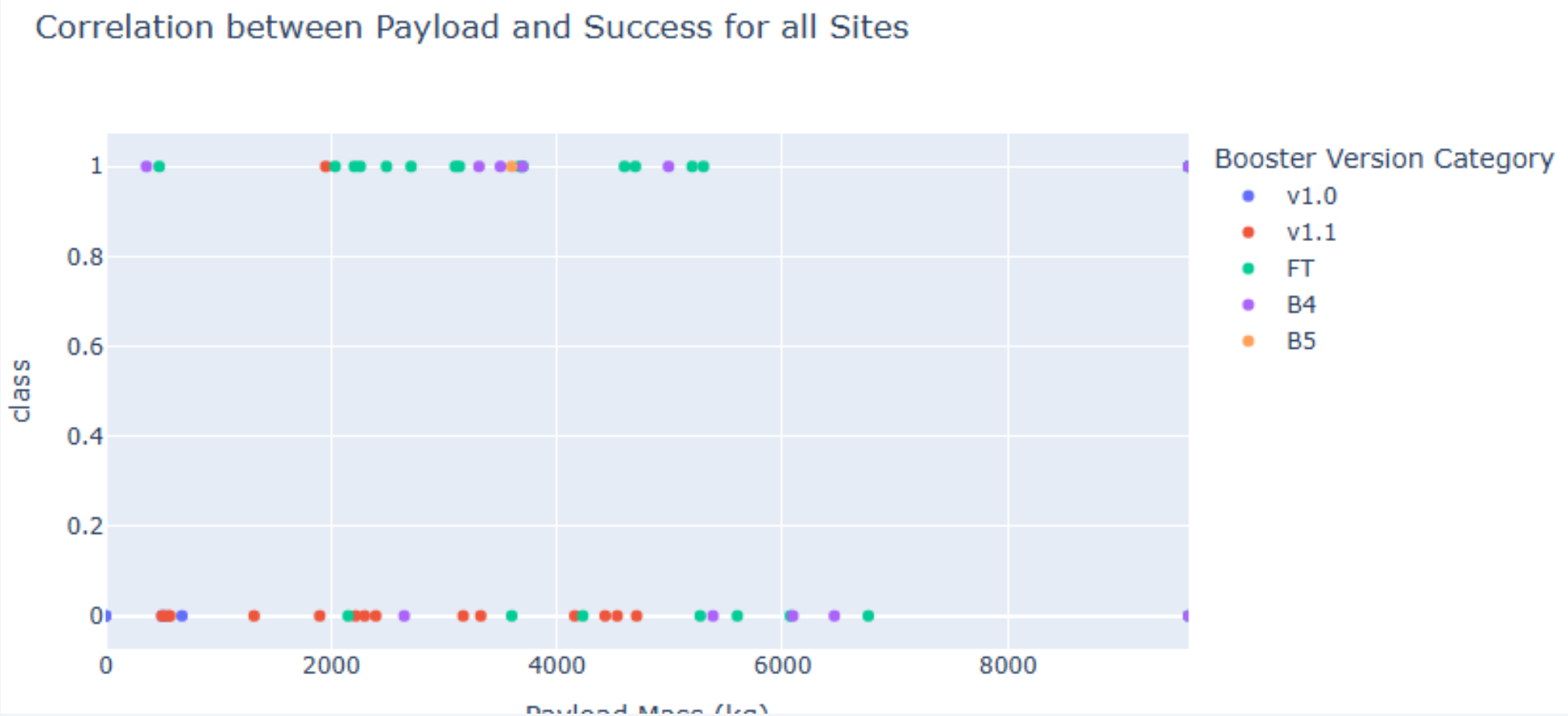
Launch success count for all sites



Launch site with highest launch success



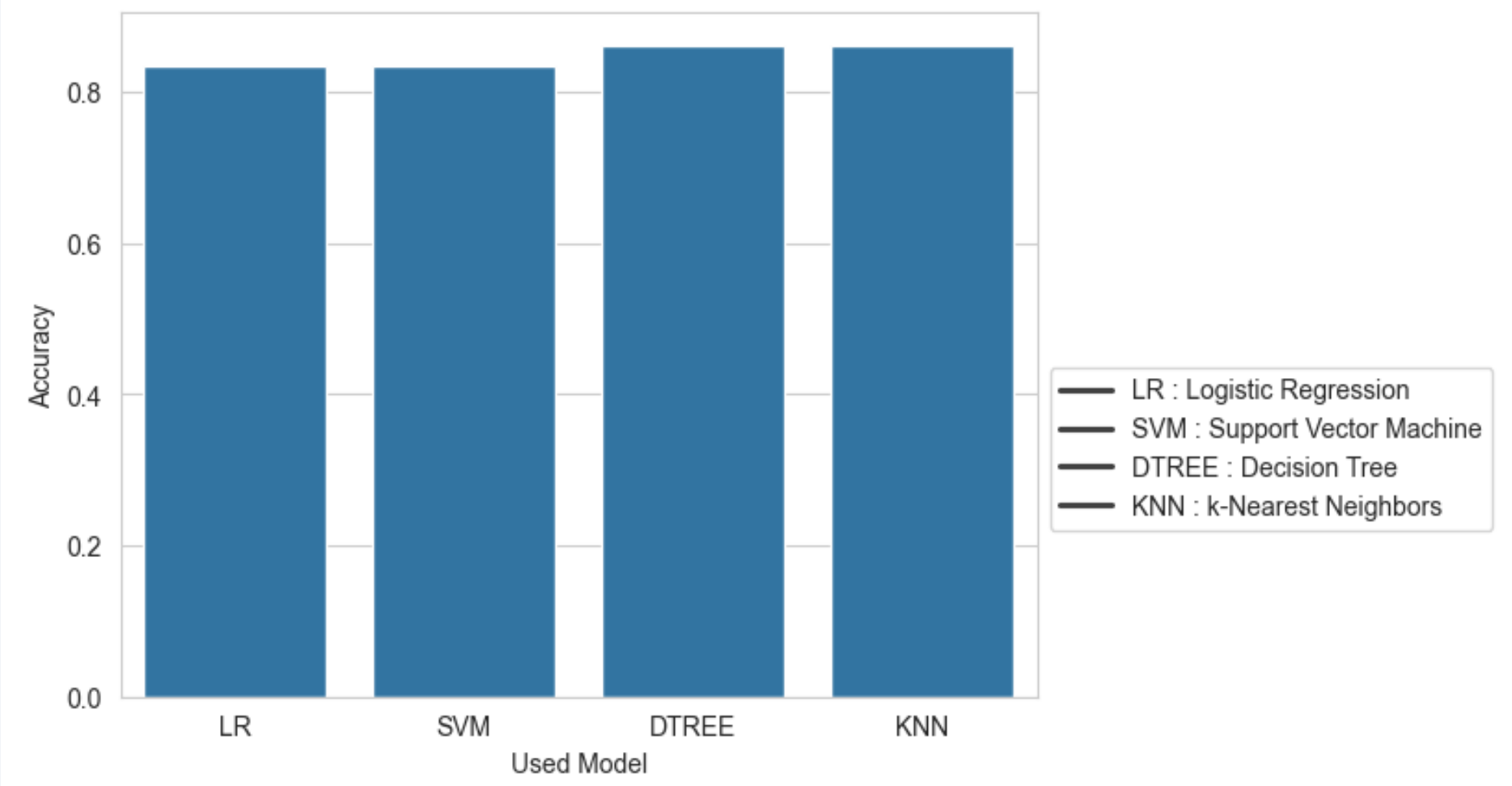
Payload vs Launch Success Rate



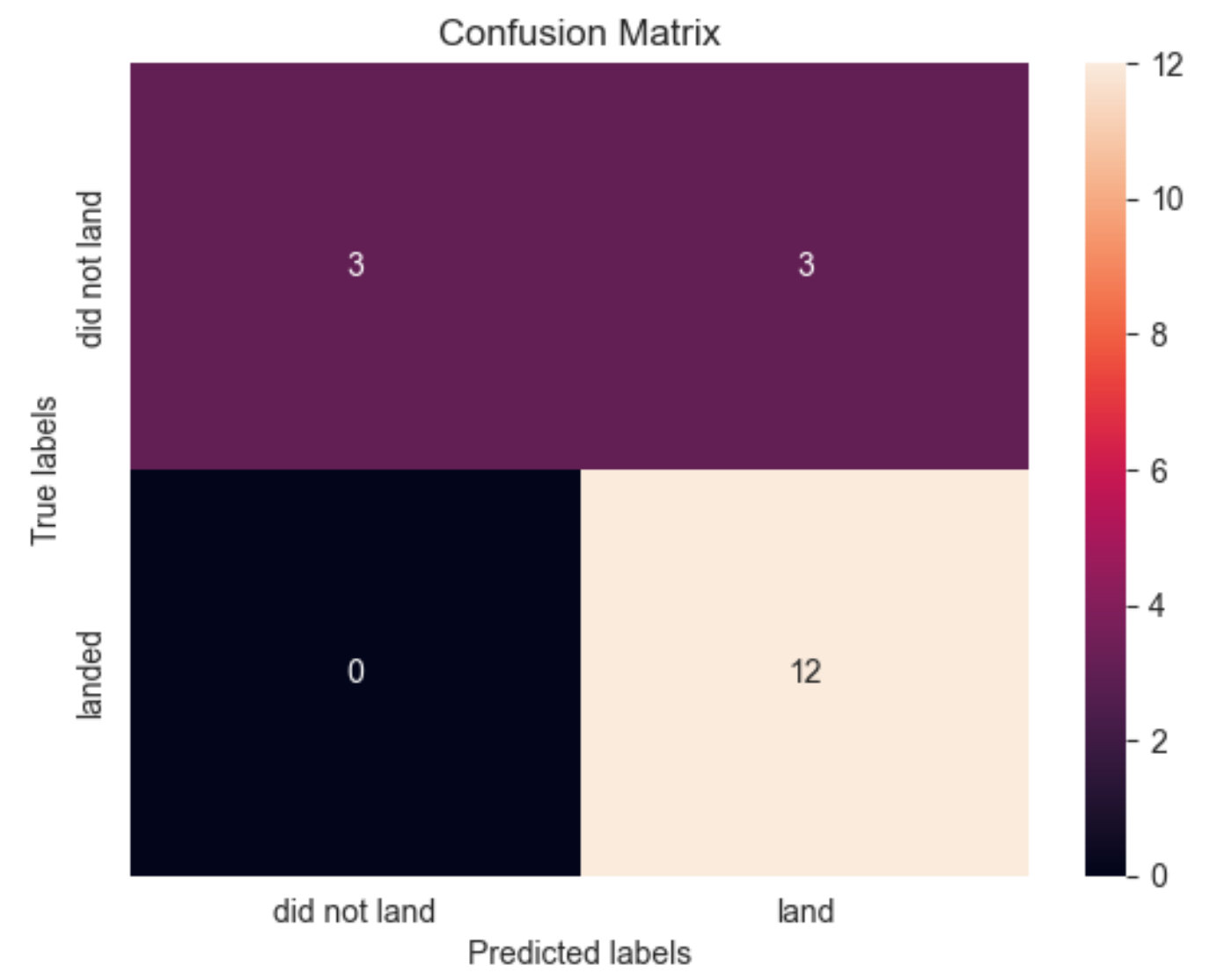
Section 5

Predictive Analysis (Classification)

Classification Accuracy



Confusion Matrix of most accurate model (Decision Tree)



Conclusions

- Launch and landing success rate has increased with the years
- More launches were performed on the West-Coast launch sites
- Landing success rate can be predicted with > 86% accuracy using Decision Tree ML

Appendix

- Python Notebooks and Scripts : see https://github.com/sirasen/sandbox/tree/main/data_science_lab
- Used datasets :
  - https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/dataset_part_1.csv
  - https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/dataset_part_2.csv
  - https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/dataset_part_3.csv
  - https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json
  - https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/spacex_launch_geo.csv
  - https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/labs/module_2/data/Spacex.csv
  - https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
  - https://github.com/sirasen/sandbox/blob/main/data_science_lab/spacex_launch_dash.csv
- Used REST API:
  - https://api.spacexdata.com/v4

Thank you!