IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mehdi Shishehbor
10/3/2022

Outline

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

Executive Summary

Summary of methodologies

- Collected data from public SpaceX API and SpaceX Wikipedia page.
- Created labels column ‘class’ which classifies successful landings.
- Explored data using SQL
- Visualization, folium maps, and dashboards.
- Predictive analysis: one hot encoding, Standardized data, parameters optimization and calculate the accuracy

Summary of the results

- Four machine learning models were produced:
- All produced similar results with accuracy rate of about 83.33%.
- More data is needed for better model determination and accuracy.

Introduction

Project background and context

- Commercial Space Age is Here
- Space X has best pricing ($62 million vs. $165 million USD)
- Largely due to ability to recover part of rocket (Stage 1)
- Space Y wants to compete with Space X

Problems you want to find answers

- Space Y tasks us to train a machine learning model to predict successful Stage 1 recovery

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Rest API and web Scraping
- Perform data wrangling
  - Classifying true landings as successful and unsuccessful otherwise
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Use RL, KNN, SVM, DT models have been built and evaluated for the best classifier

Data Collection

Describe how data sets were collected.

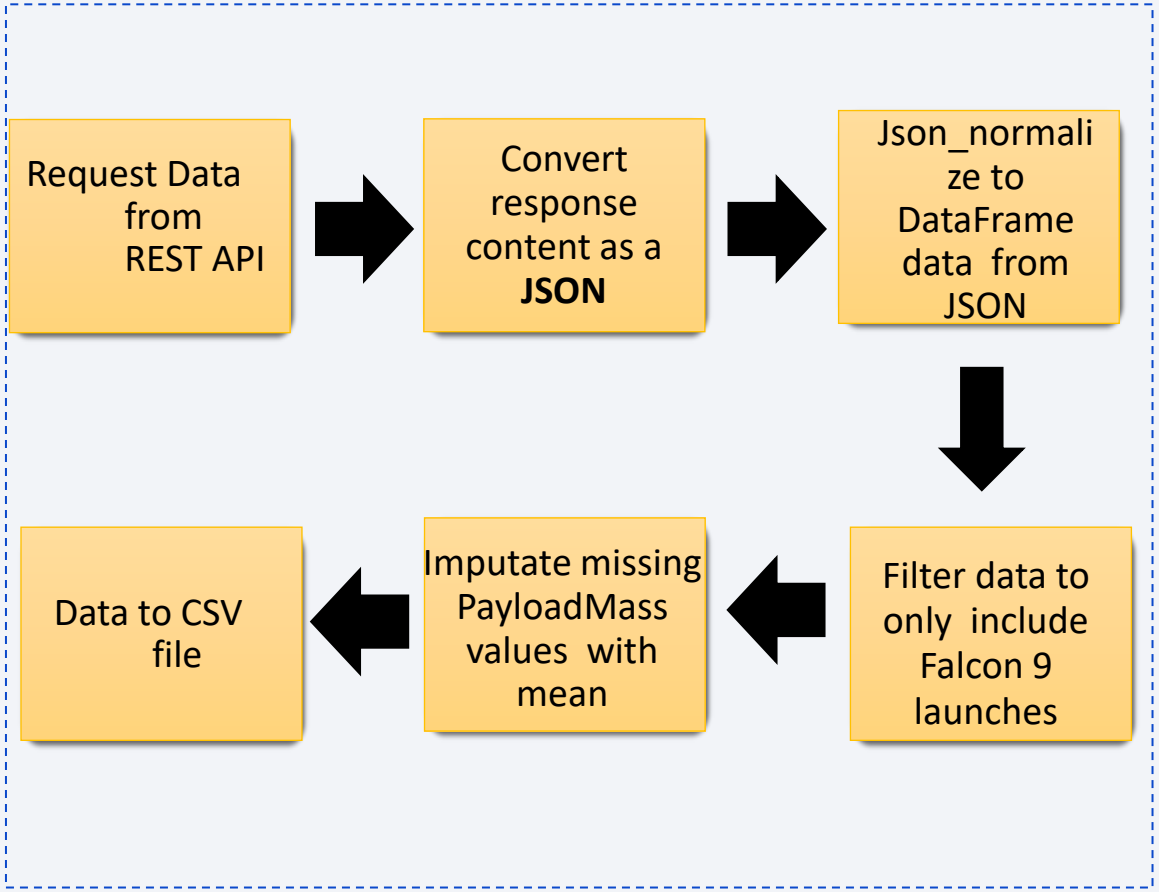
- SpaceX REST API

You need to present your data collection process use key phrases and flowcharts

Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

- Add the GitHub URL of the completed SpaceX API calls notebook (must include completed code cell and outcome cell), as an external reference and peer-review purpose

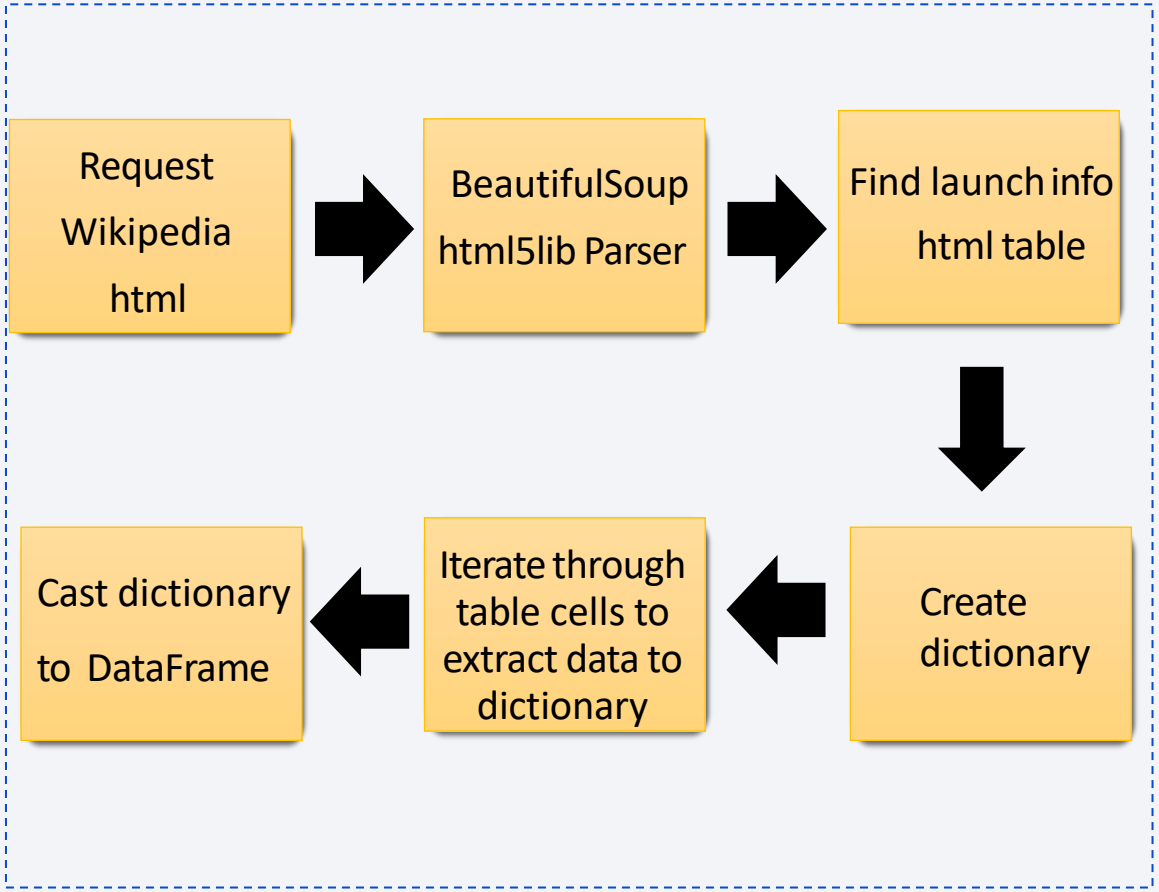


https://github.com/mshisheh/IBM_data_science/blob/main/Data-collection-api.ipynb

Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts

- Add the GitHub URL of the completed web scraping notebook, as an external reference and peer-review purpose

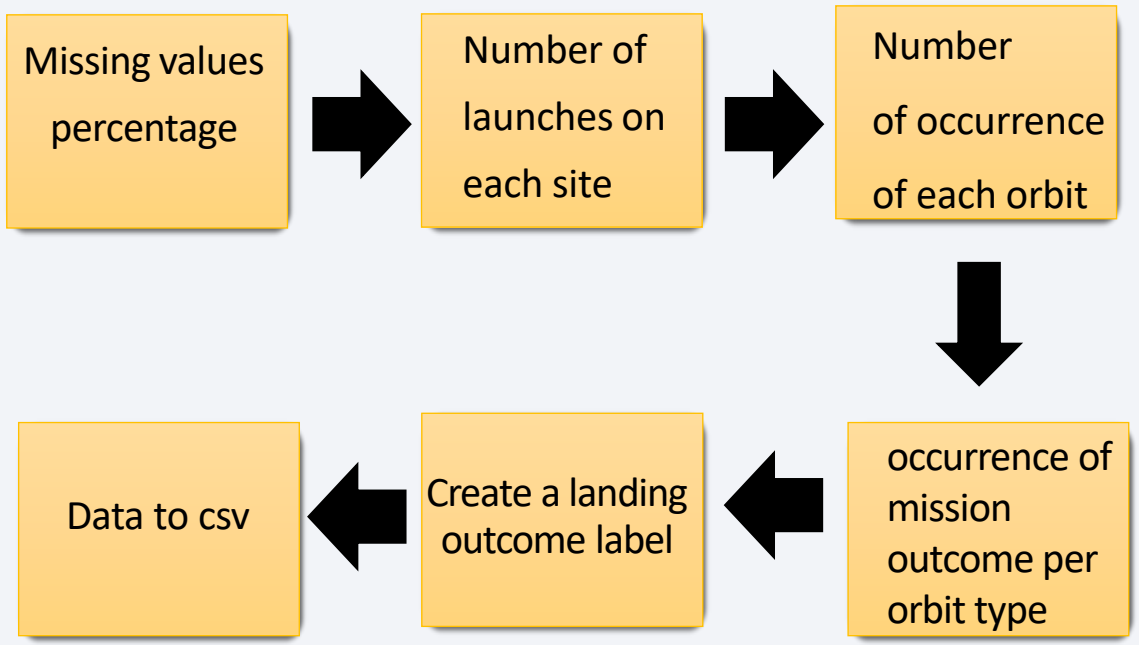


https://github.com/mshisheh/IBM_data_science/blob/main/web_scraping.ipynb

Data Wrangling

- Describe how data were processed
- You need to present your data wrangling process using key phrases and flowcharts
- Add the GitHub URL of your completed data wrangling related notebooks, as an external reference and peer-review purpose

https://github.com/mshisheh/IBM_data_science/blob/main/Data%20wrangling.ipynb



EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
- Add the GitHub URL of your completed EDA with data visualization notebook, as an external reference and peer-review purpose

Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit vs. Success Rate, Flight Number vs. Orbit, Payload vs Orbit, and Success Yearly Trend

Scatter plots, line charts, and bar plots were used to compare relationships between variables to decide if a relationship exists so that they could be used in training the machine learning model

https://github.com/mshisheh/IBM_data_science/blob/main/EDA%20with%20Visualization.ipynb

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
- Add the GitHub URL of your completed EDA with SQL notebook, as an external reference and peer-review purpose

Loaded data set into IBM DB2 Database.

Queried using SQL Python integration.

Queries were made to get a better understanding of the dataset.

Queried information about launch site names, mission outcomes, various pay load sizes of customers and booster versions, and landing outcomes

https://github.com/mshisheh/IBM_data_science/blob/main/eda-sql.ipynb

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
- Explain why you added those objects
- Add the GitHub URL of your completed interactive map with Folium map, as an external reference and peer-review purpose

Folium maps mark Launch Sites, successful and unsuccessful landings, and a proximity example to key locations: Railway, Highway, Coast, and City.

This allows us to understand why launch sites may be located where they are. Also visualizes successful landings relative to location.

https://github.com/mshisheh/IBM_data_science/blob/main/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

Dashboard includes a pie chart and a scatter plot.

Pie chart can be selected to show distribution of successful landings across all launch sites and can be selected to show individual launch site success rates.

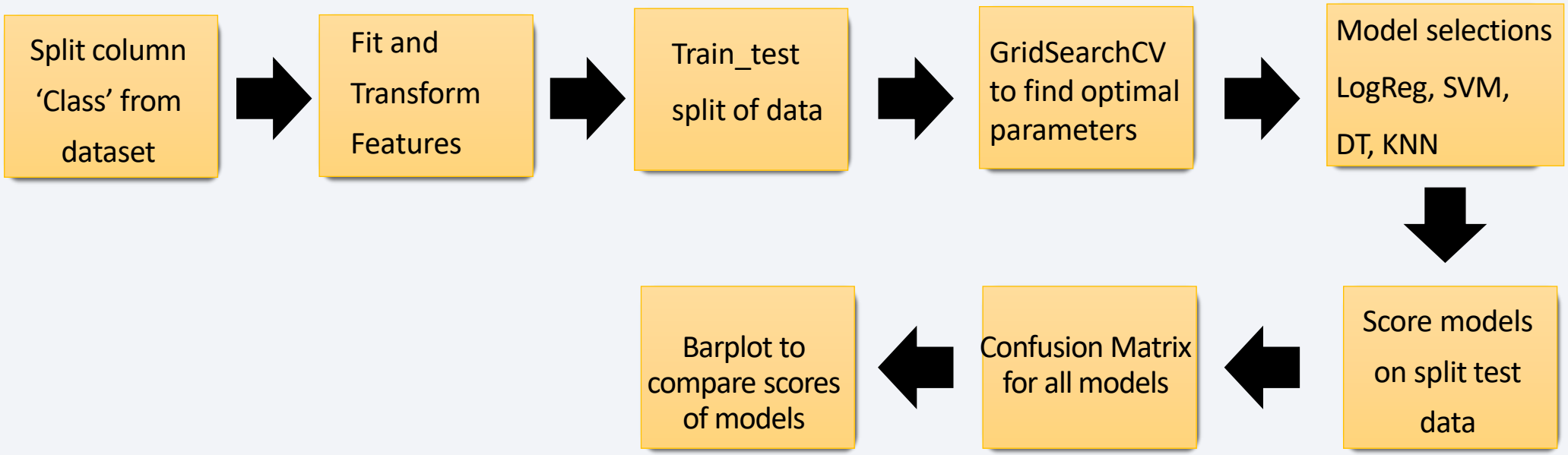
Scatter plot takes two inputs: All sites or individual site and payload mass on a slider between 0 and 10000 kg.

The pie chart is used to visualize launch site success rate.

The scatter plot can help us see how success varies across launch sites, payload mass, and booster version category.

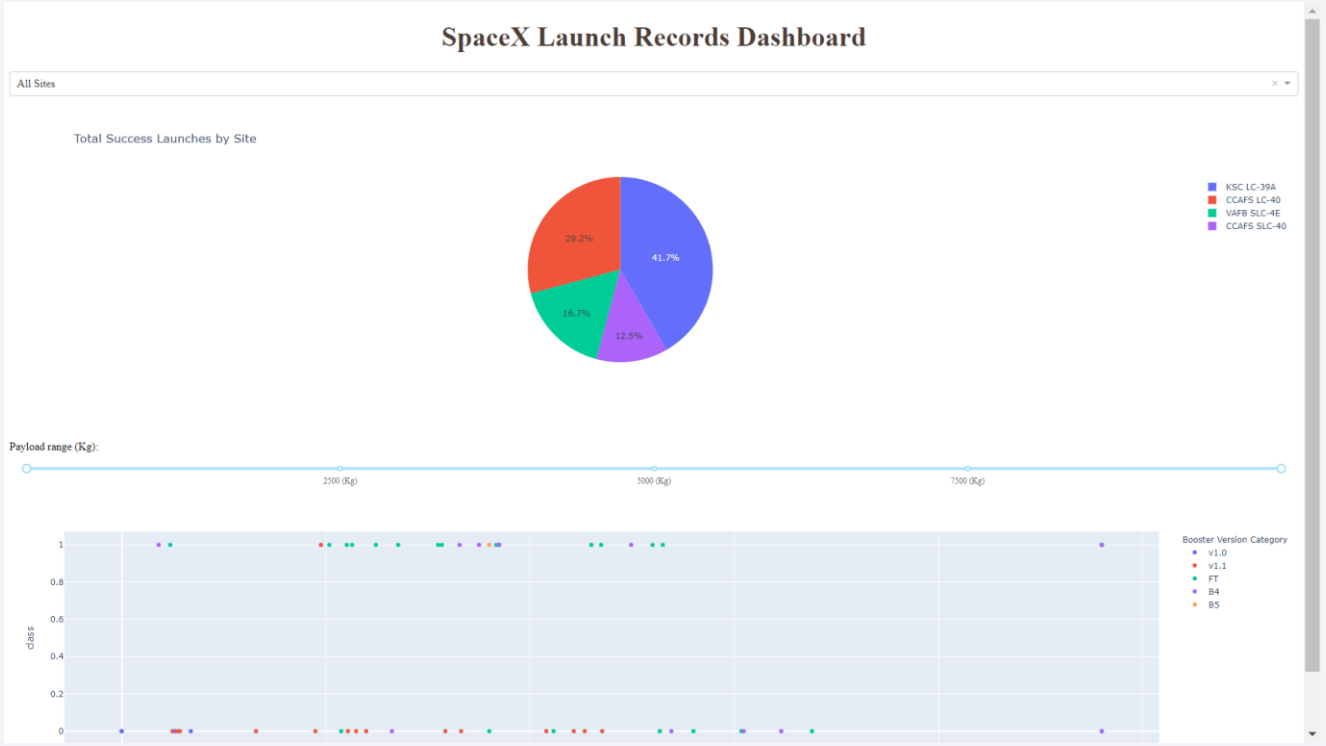
https://github.com/mshisheh/IBM_data_science/blob/main/Dashboard%20with%20Plotly%20Dash.py

Predictive Analysis (Classification)



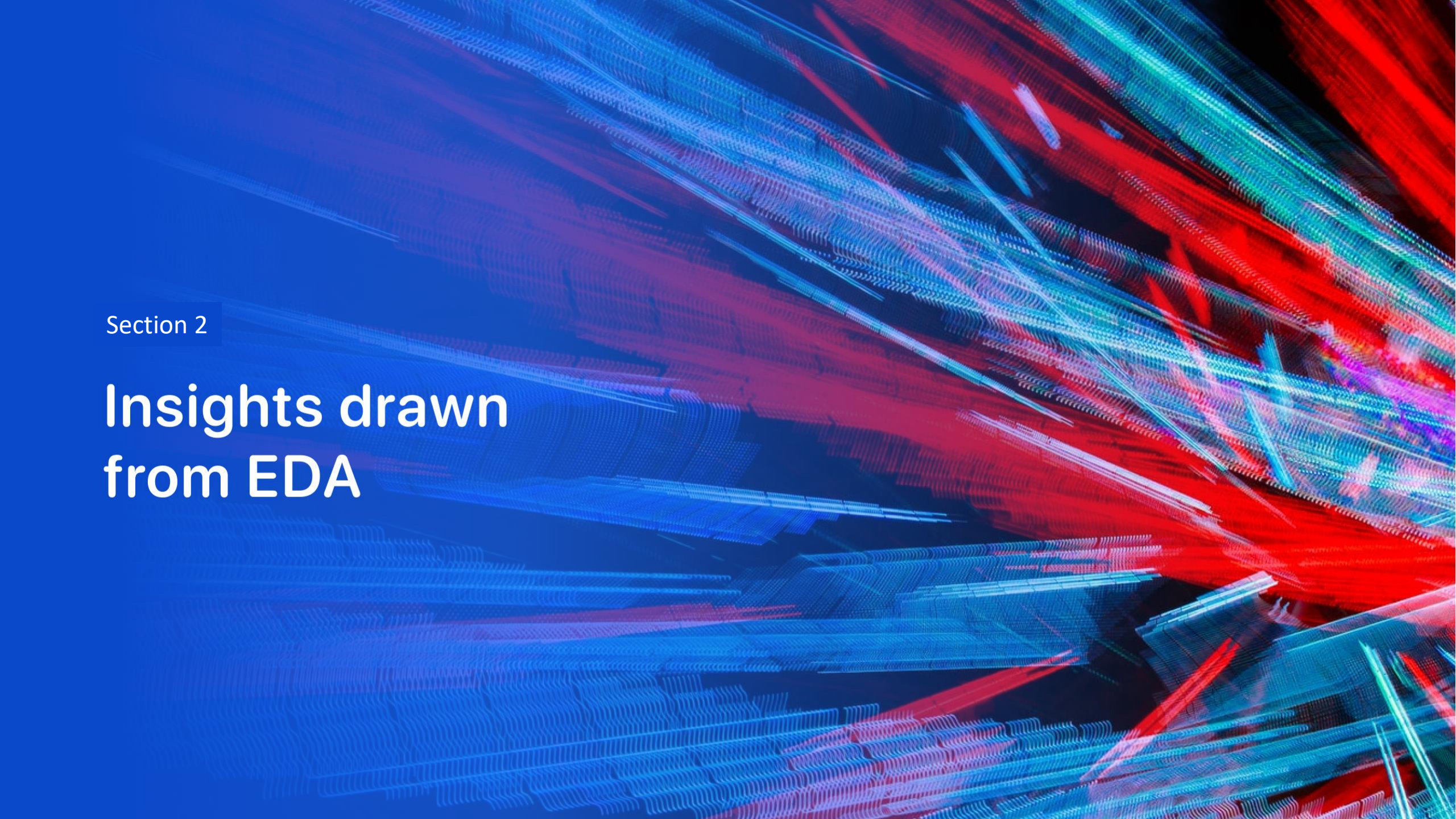
https://github.com/mshisheh/IBM_data_science/blob/main/Machine_Learning_Prediction.ipynb

Results



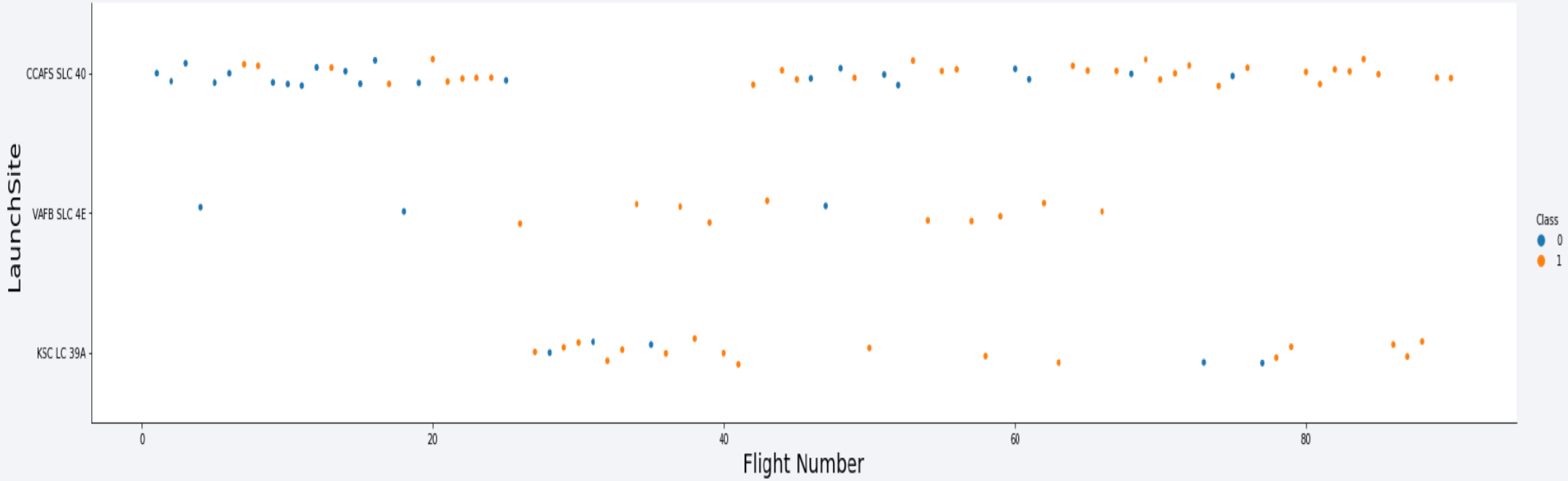
- This is a preview of the Plotly dashboard. The following sides will show the results of EDA with visualization, EDA with SQL, Interactive Map with Folium, and finally the results of our model with about 83% accuracy.

- Decision Tree model are the best in terms of prediction accuracy for this dataset.


Section 2
Insights drawn from EDA

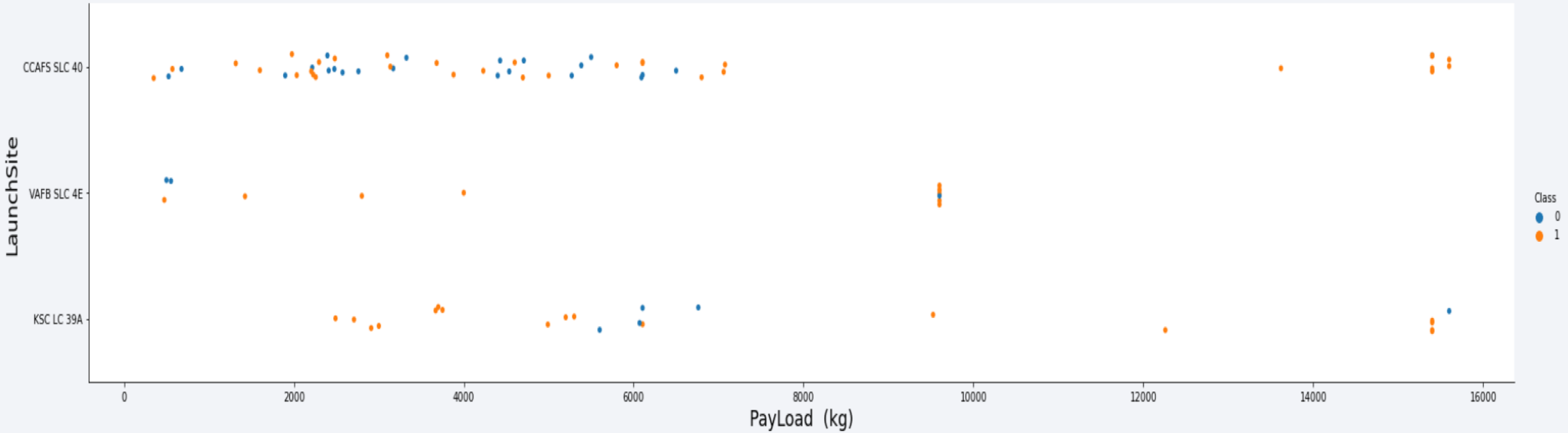
Flight Number vs. Launch Site

Different launch sites have different success rates. But as we increase the number of flights the success rate increase.

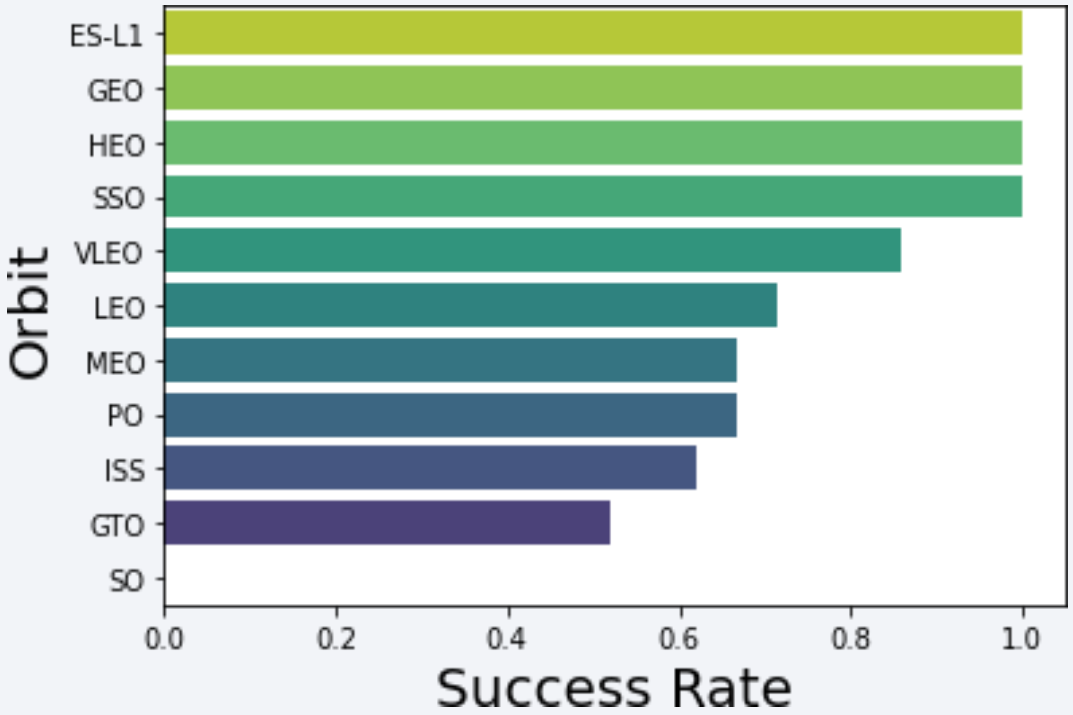


Payload vs. Launch Site

As well, if increase the number of we Pay Load Mass (kg) the success rate increase.



Success Rate vs. Orbit Type



ES-L1 (1), GEO (1), HEO (1) have 100% success rate (sample sizes in parenthesis) SSO (5) has 100% success rate

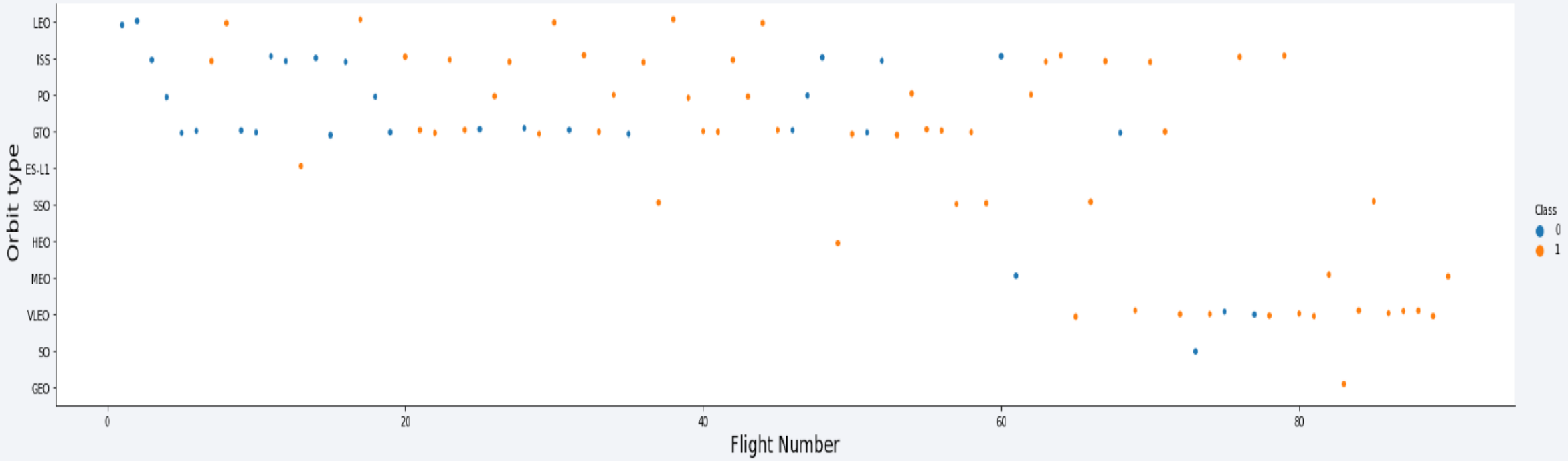
VLEO (14) has decent success rate and attempts

SO (1) has 0% success rate

GTO (27) has the around 50% success rate but largest sample

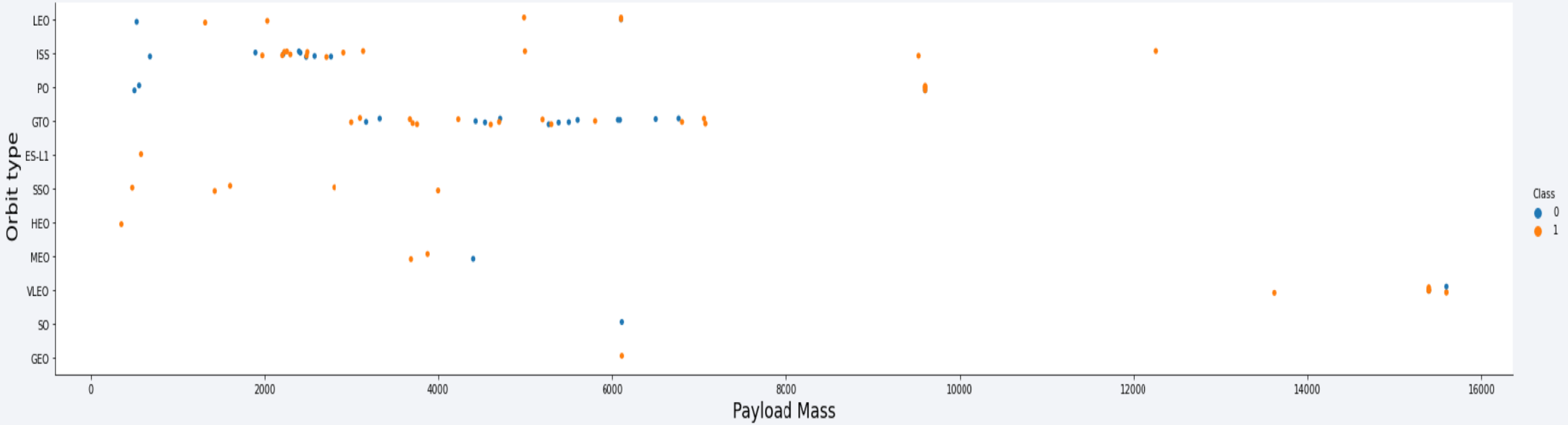
Success Rate Scale with
0 as 0%
0.6 as 60%
1 as 100%

Flight Number vs. Orbit Type

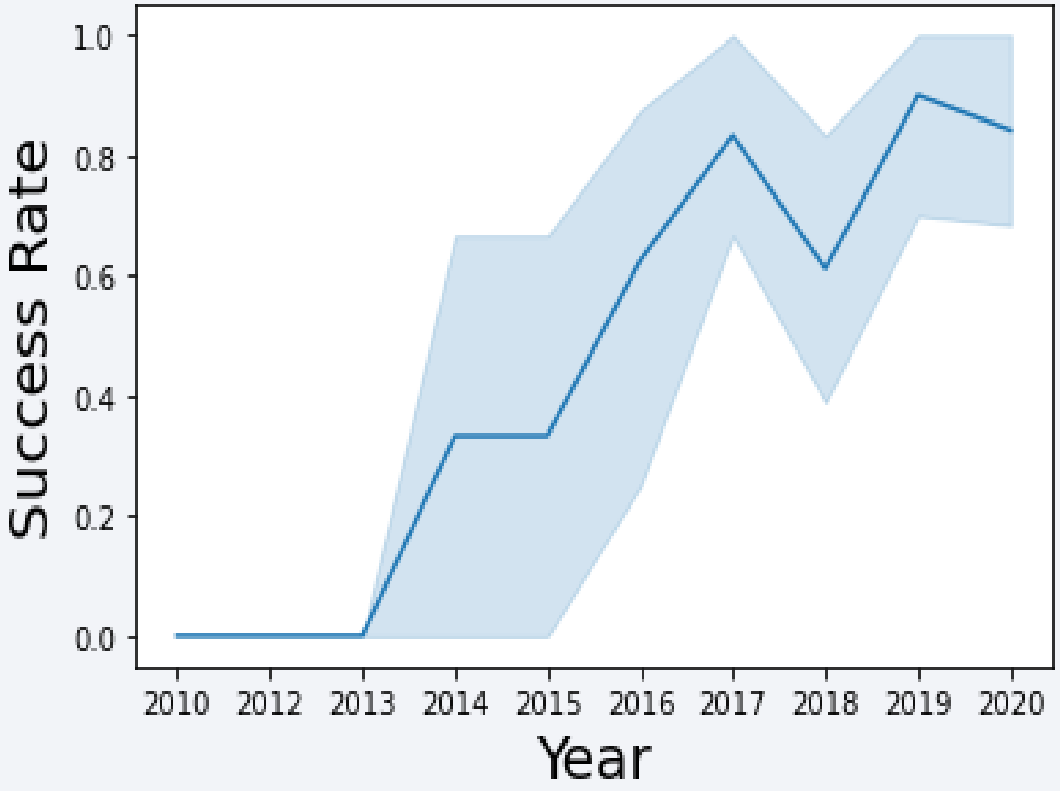


Payload vs. Orbit Type

There is a connection between ISS and payload in the range of 2000 to 3000. Also between GTE and Payload at 4000 to 8000.



Launch Success Yearly Trend



Success generally increases over time since 2013 with a slight dip in 2018

Success in recent years at around 80%

All Launch Site Names

%sqlSELECT DISTINCT(launch_site) FROM Spacex

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

Launch Site Names Begin with 'CCA'

%sqlSELECT * FROM SpacexWHERE launch_siteLIKE'CCA%' LIMIT 5

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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)
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-12	22:41:00	F9 v1.1	CCAFS LC-40	SES-8	3170	GTO	SES	Success	No attempt

Total Payload Mass

%sqlSELECT SUM(payload_mass__kg_) FROM SpacexWHERE customer='NASA (CRS)'

1
45596

Average Payload Mass by F9 v1.1

%sqlSELECT AVG(payload_mass__kg_) FROM SpacexWHERE booster_version='F9 v1.1'

1
2928

First Successful Ground Landing Date

%sqlSELECT MIN(DATE) FROM SpacexWHERE landing__outcome='Success (ground pad)'

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sqlSELECT booster_versionFROM Spacex
WHERE landing__outcome='Success (drone ship)' AND payload_mass__kg_ BETWEEN 4000 AND 6000
```

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

Total Number of Successful and Failure Mission Outcomes

%sqlSELECT COUNT(MISSION_OUTCOME) AS missionoutcomesFROM SpacexWHERE mission_outcomeLIKE 'Success%’

missionoutcomes
100

%sqlSELECT COUNT(MISSION_OUTCOME) AS missionoutcomesFROM SpacexWHERE mission_outcomeLIKE 'Failure

missionoutcomes
1

Boosters Carried Maximum Payload

%sqlSELECT booster_versionAS MaxboosterversionFROM SpacexWHERE payload_mass__kg_=(SELECT MAX(payload_mass__kg_) FROM Spacex)

maxboosterversion
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

%sqlSELECT landing__outcome,booster_version,launch_site,DATEFROM SpacexWHERE landing__outcome='Failure (drone ship)' AND EXTRACT(YEAR FROM DATE)='2015'

landing__outcome	booster_version	launch_site	DATE
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015-10-01
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015-04-14

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

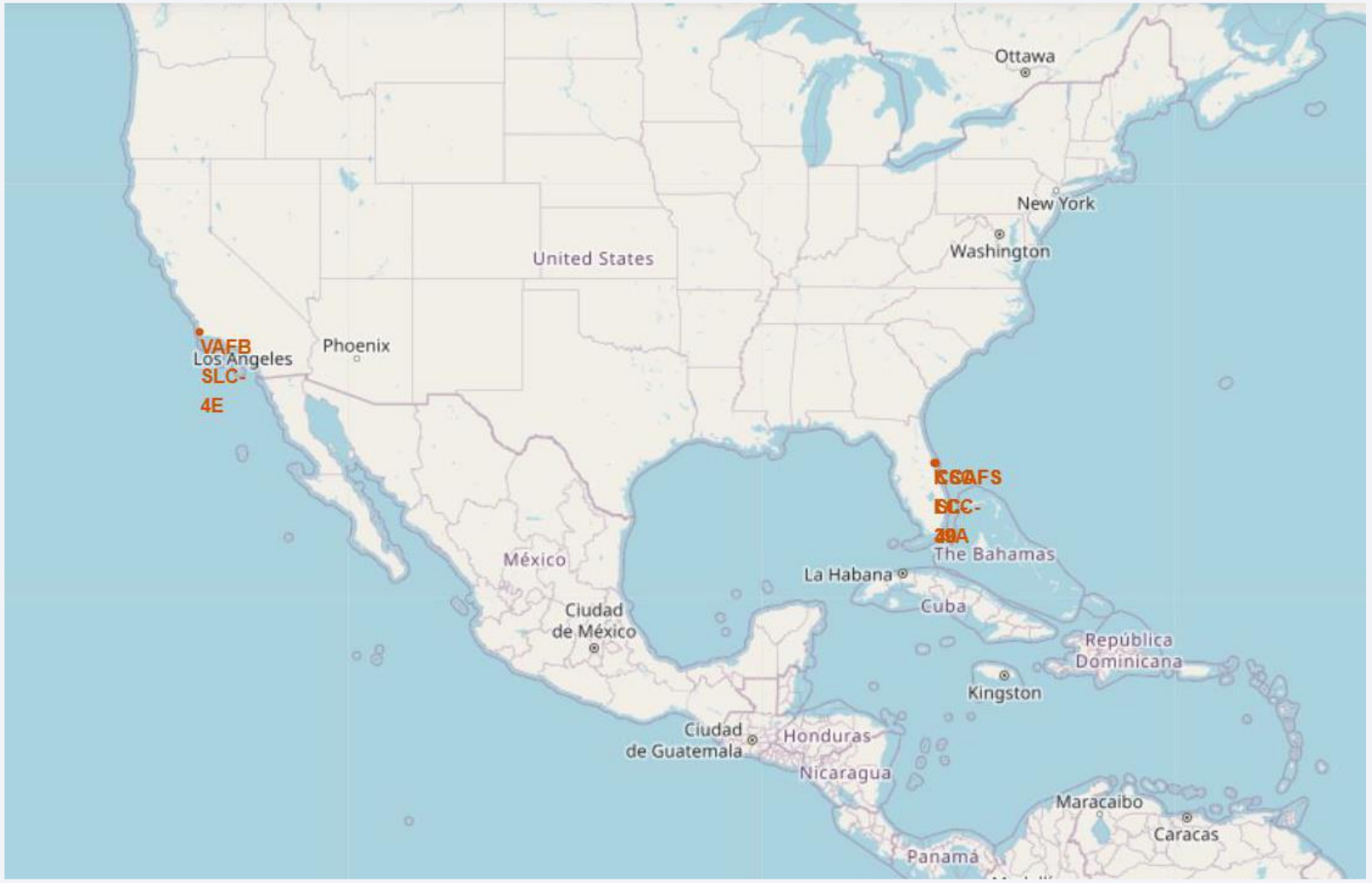
%sqlSELECT landing__outcome, COUNT(landing__outcome) FROM SpacexWHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY landing__outcomeORDER BY COUNT(landing__outcome) DESC

landing__outcome	2
No attempt	10
Success (drone ship)	6
Failure (drone ship)	5
Success (ground pad)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	1
Precluded (drone ship)	1

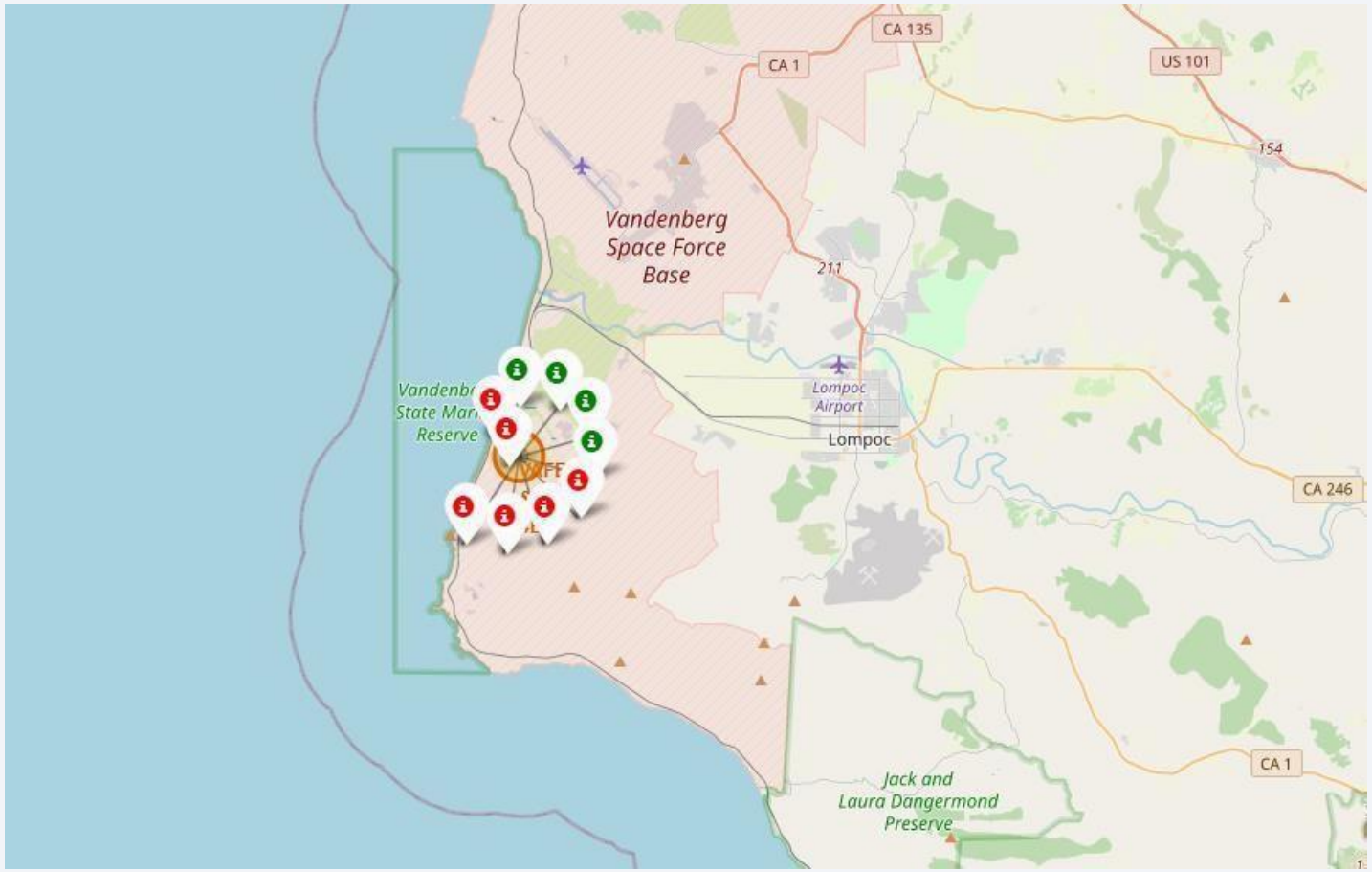
Section 3

Launch Sites Proximities Analysis

<Folium Map Screenshot 1>

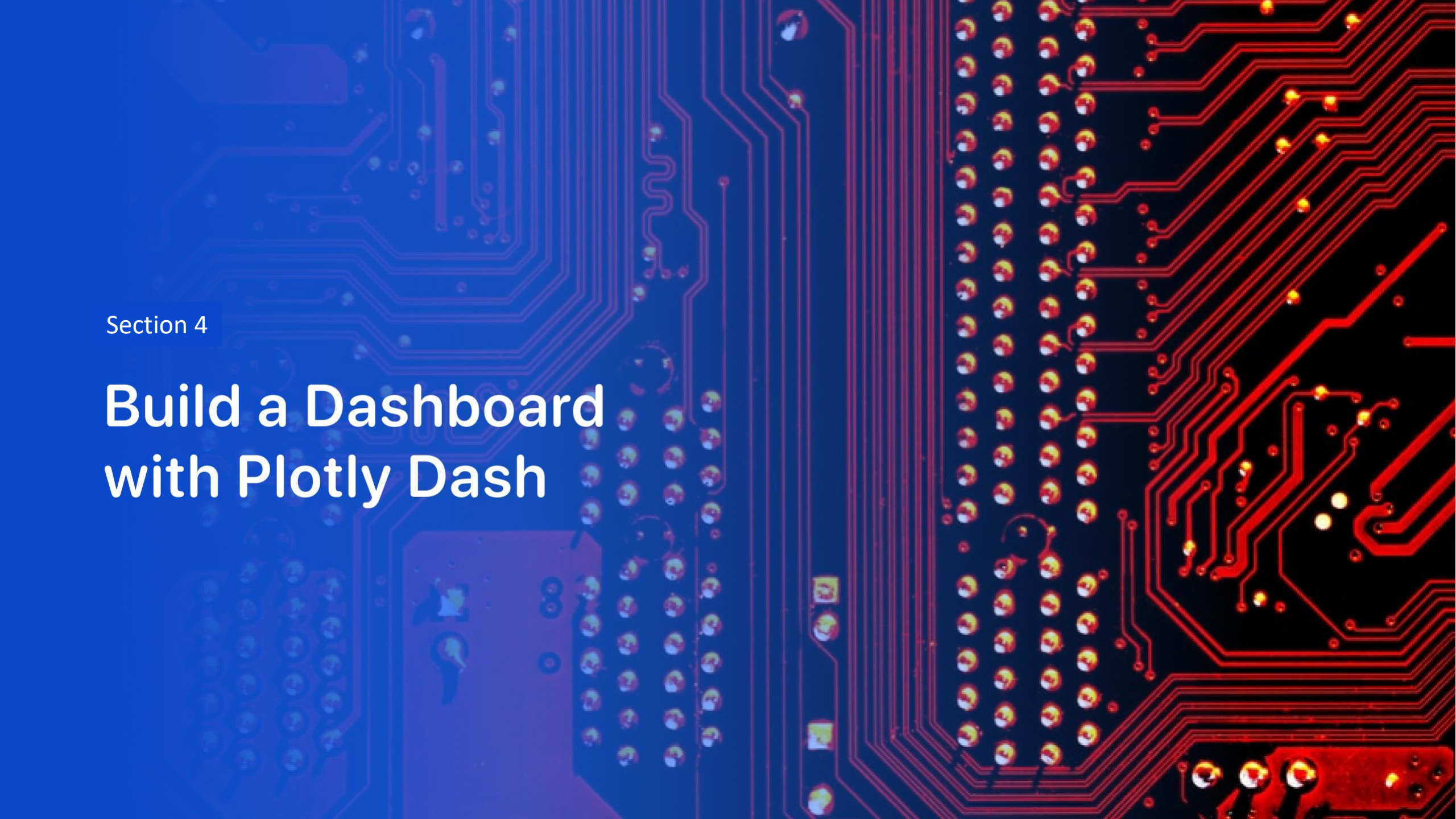


<Folium Map Screenshot 2>

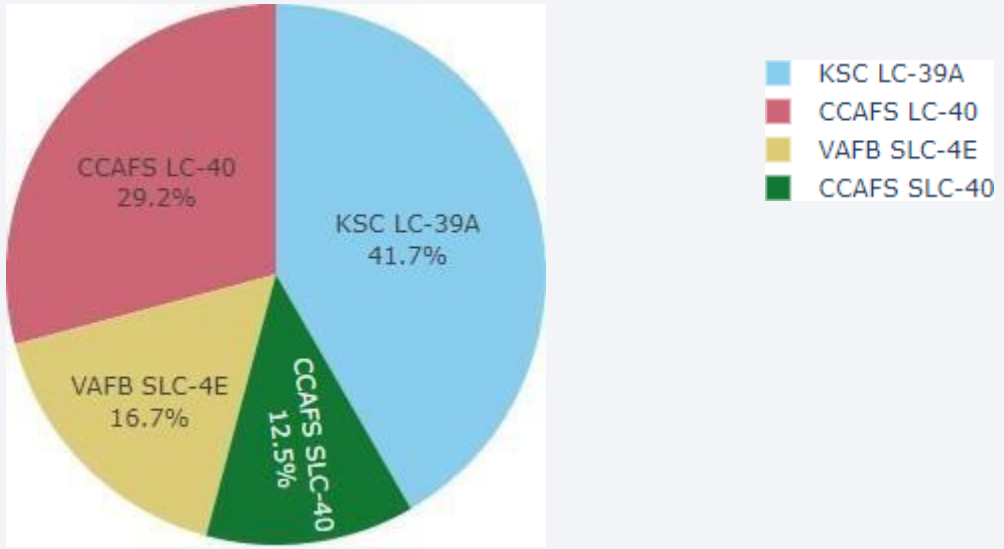


Clusters on Folium map can be clicked on to display each successful landing (green icon) and failed
landing (red icon). In this example VAFB SLC-4E shows 4 successful landings and 6 failed landings.

<Folium Map Screenshot 3>

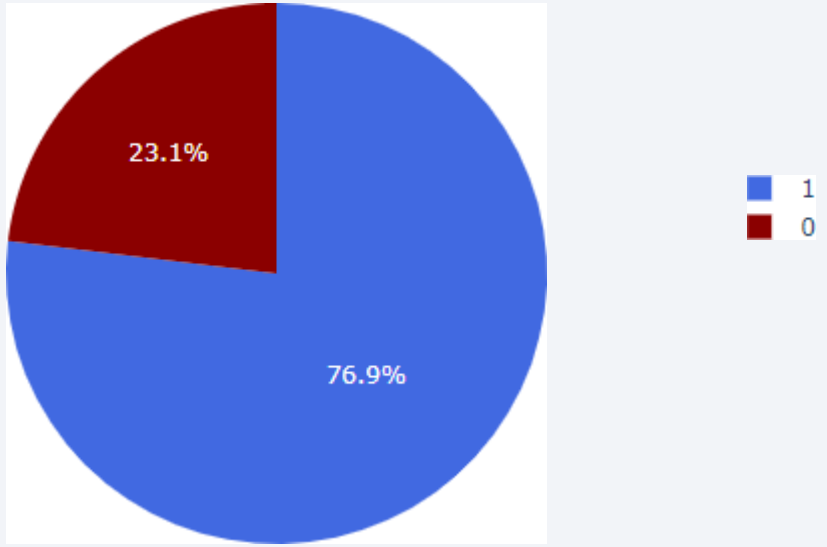

Section 4
Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>



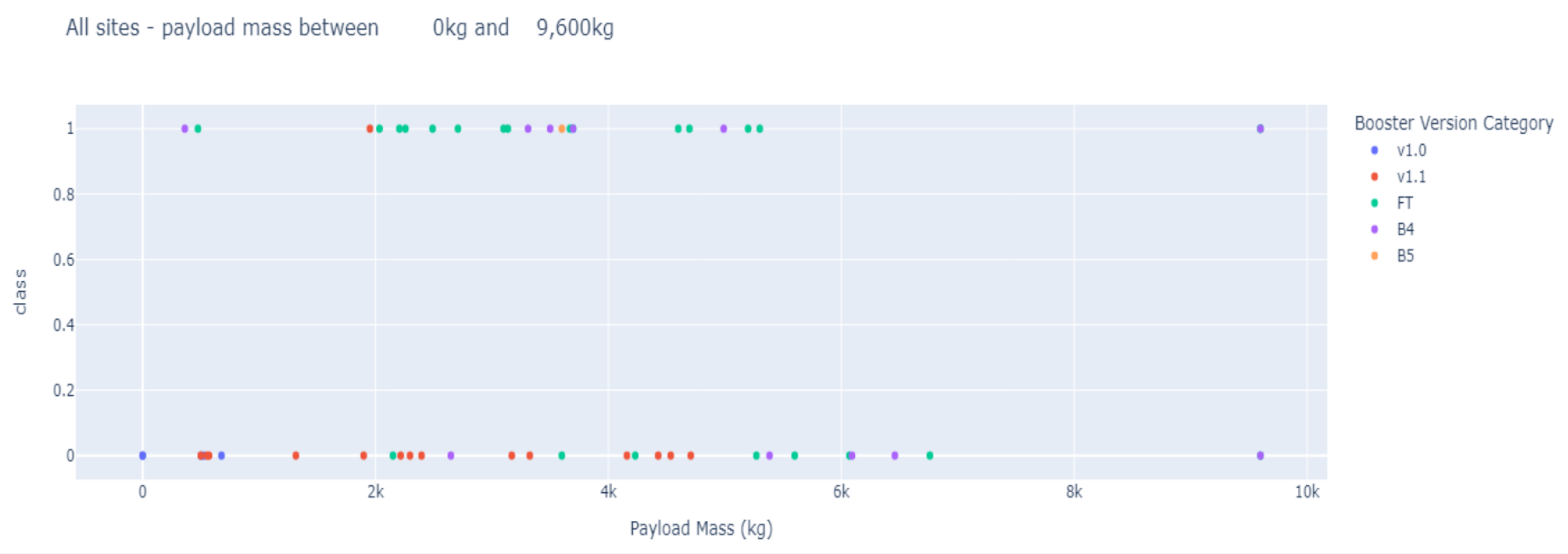
This is the distribution of successful landings across all launch sites. CCAFS LC-40 is the old name of CCAFS SLC-40 so CCAFS and KSC have the same amount of successful landings, but a majority of the successful landings where performed before the name change. VAFB has the smallest share of successful landings. This may be due to smaller sample and increase in difficulty of launching in the west coast.

<Dashboard Screenshot 2>



KSC LC-39A has the highest success rate with 10 successful landings and 3 failed landings.

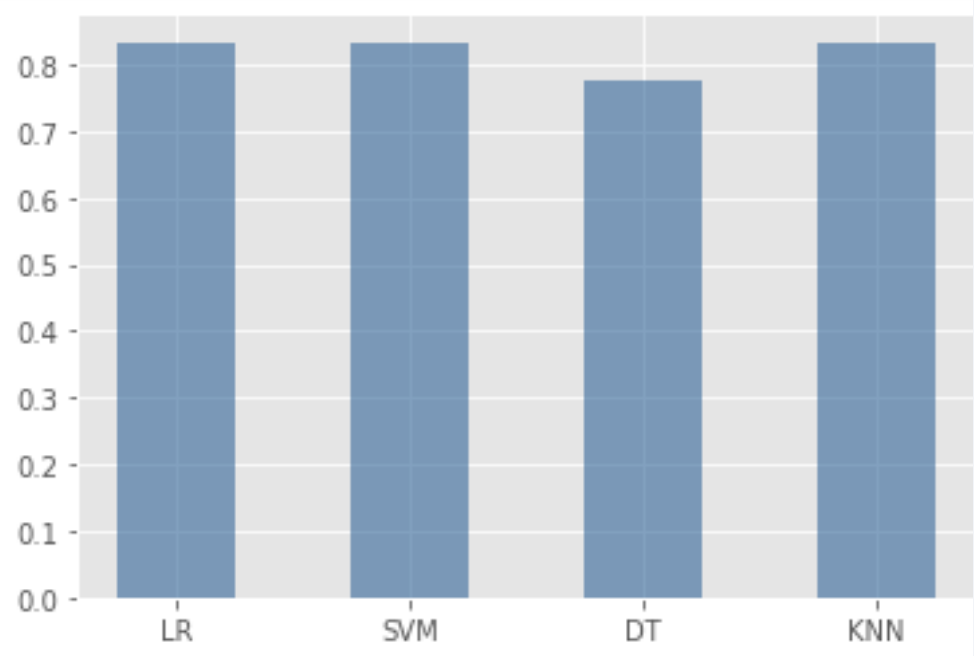
<Dashboard Screenshot 3>



Section 5

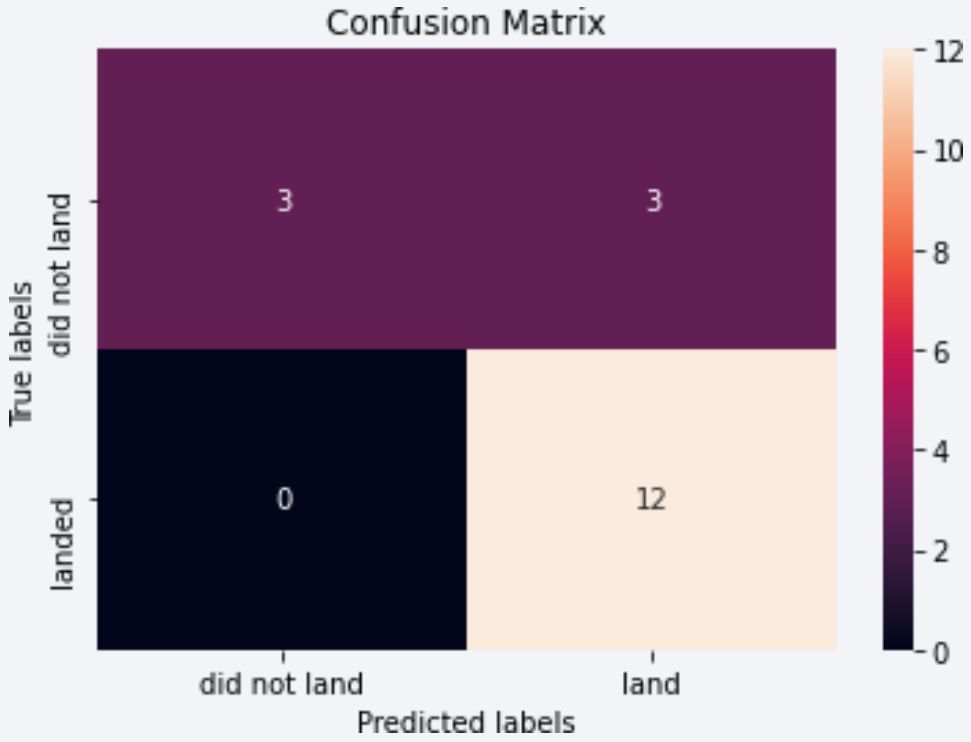
Predictive Analysis (Classification)

Classification Accuracy



As we can see Decision Tree has the highest accuracy with almost 0.89, then comes the remaining models with almost the same accuracy of 0.84.

Confusion Matrix



Measure	Derivations	Result
Precision	PPV = TP / (TP + FP)	0.67
Accuracy	ACC = (TP + TN) / (P + N)	0.89
F1 Score	F1 = 2TP / (2TP + FP + FN)	0.80

Conclusions

- After analysis data the CCAFS SLC-40 site and KSC LC-39A site are has most successful launches from all the sites.

•Orbit GEO,HEO,SSOES L1 has the best Success Rate.

•The payload of 0 kg to 5000 kg was more diverse than 6000 kg to 10000

•The Decision Tree model is the best in terms of prediction accuracy for this dataset.

Appendix

Special Thanks to All Instructors:

https://www.coursera.org/professional-certificates/ibm-data-science?#instructors

Thank you!