IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Kiley Junker
31 August 2023

Outline

- EXECUTIVE SUMMARY
- INTRODUCTION
- METHODOLOGY
- RESULTS
- CONCLUSION
- APPENDIX

Executive Summary

- Determining if the first stage will land allows for an estimation of the cost of the launch.
- Data was collected from the SpaceX API and webscraping of Wikipedia.
- Performed exploratory data analysis (EDA) using visualization and SQL.
- Created interactive visual analytics using Folium and Plotly Dash.
- Performed predictive analysis using four different classification models.
- Determined that the decision tree model best classifies the data.

Introduction

The first stage of a rocket does most of the work for the launch, making it quite large and expensive.

SpaceX rocket launches are relatively inexpensive due to their ability to reuse the first stage of a rocket.

Using SpaceX data, we can train a model to determine if the first stage is able to be reused.

Determining if the first stage will land allows for an estimation of the cost of the launch.

Section 1

Methodology

Methodology

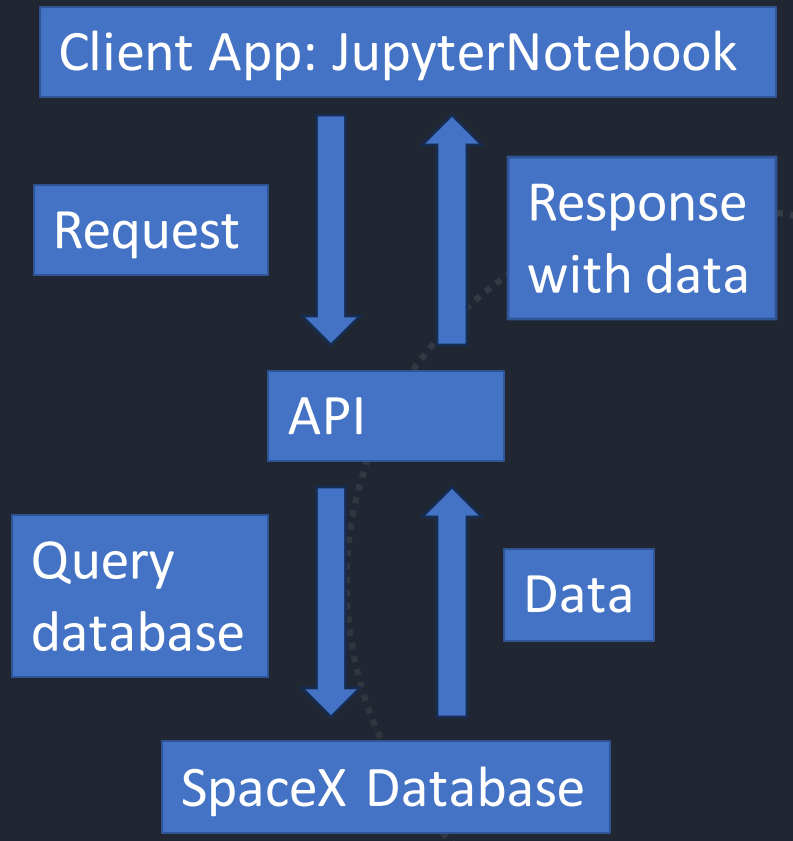
Executive Summary

- Data collection methodology
  - SpaceX API and Webscraping of Wikipedia
- Perform data wrangling
  - Created dummy variables for categorical variables
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Using GridSearch, tuned models and selected based on accuracy metrics.

Data Collection

- The SpaceX API was used to gather launch data.
- Webscraping of Wikipedia was used to gather Falcon9 specific launch data.

Data Collection- SpaceX API



Appendix: Data Collection SpaceX API

Data Collection- Scraping

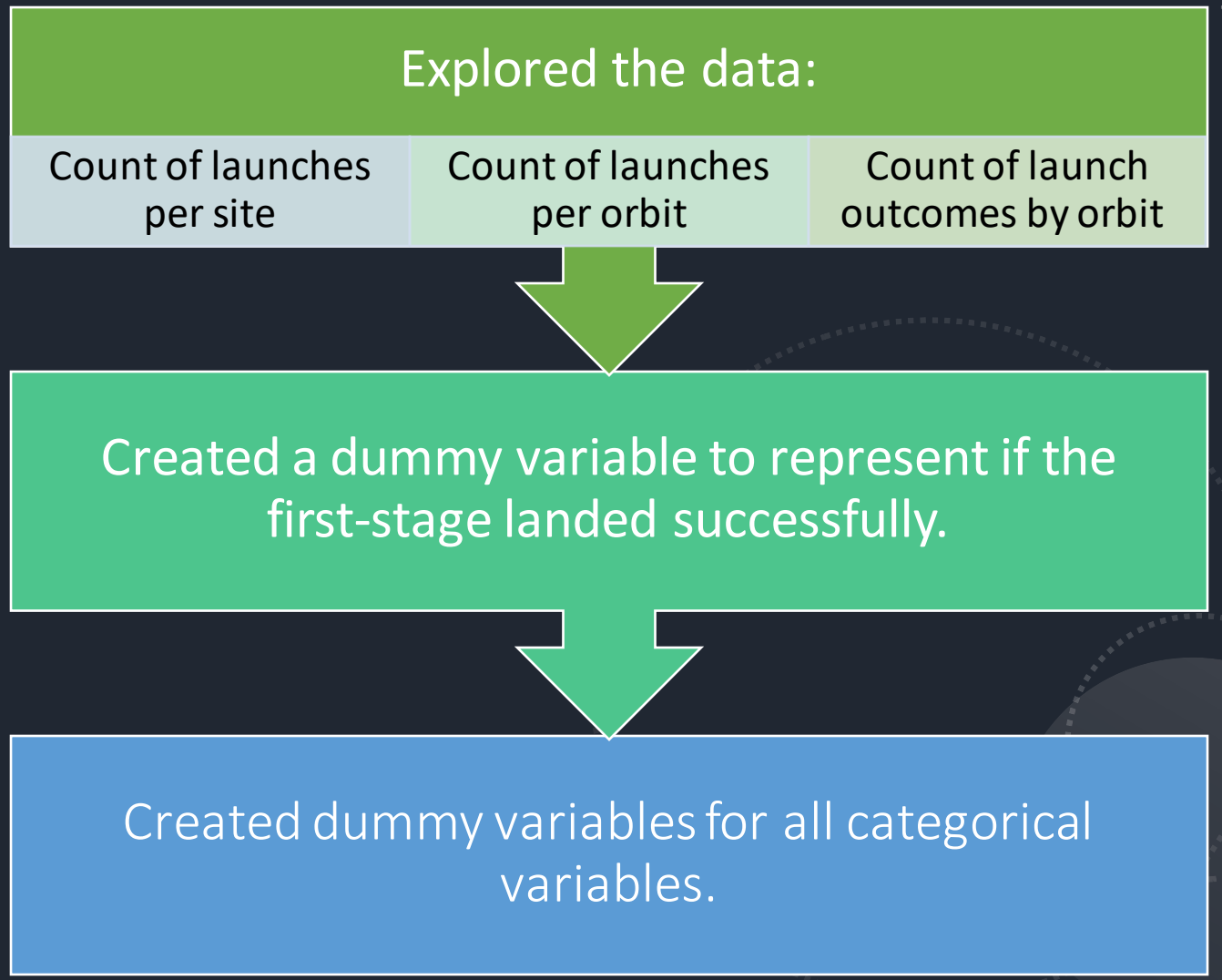
Request Falcon9 Launch Wiki page from its URL.

Extract all variable names from the HTML table header.

Create data frame by parsing the launch HTML tables.

Appendix: Data Collection Webscraping

Data Wrangling



Appendix: Data Wrangling

EDA with Data Visualization

Flight Number v. Launch Site

Payload v. Launch Site

Success Rate by Orbit Type

Flight Number v. Orbit Type

Payload v. Orbit Type

Launch Success Yearly Trend

Appendix: EDA with Data Visualization

EDA with SQL

Unique Launch Sites

Records with Launch Site CAA...

Total Payload Mass by NASA (CRS)

Average Payload Mass by booster F9 v1.1

Date of first successful ground pad landing

Boosters successful in drone ship landings, mass between 4000 and 6000

Total successful and failed missions

Booster Version to have carried the maximum payload mass

Date, Month, Year, Booster, Launch site of failed drone ship landings

Count of landing outcomes 6/4/2010-3/20/2017

Appendix: EDA with SQL

Build an Interactive Map with Folium

- Marked all launches on the map to see distribution of launches across sites.
- Created Circle and Marker objects for each launch site.
- Mark the success/failed status for each launch.
- Calculated and annotated distances between launch sites and coastlines to judge proximity to the ocean.

Appendix: Folium Map

Build a Dashboard with Plotly Dash

- A dropdown menu is used to choose a specific site or all sites for analysis.
- If all sites is selected:
  - A pie chart is displayed showing successful launches by site
  - A scatter plot is displayed showing payload v. launch outcome by site
- If one site is selected:
  - A pie chart is displayed showing the launch success of the selected site
  - A scatter plot is displayed showing payload v. launch outcome of the selected site

Appendix: Plotly Dash Dashboard

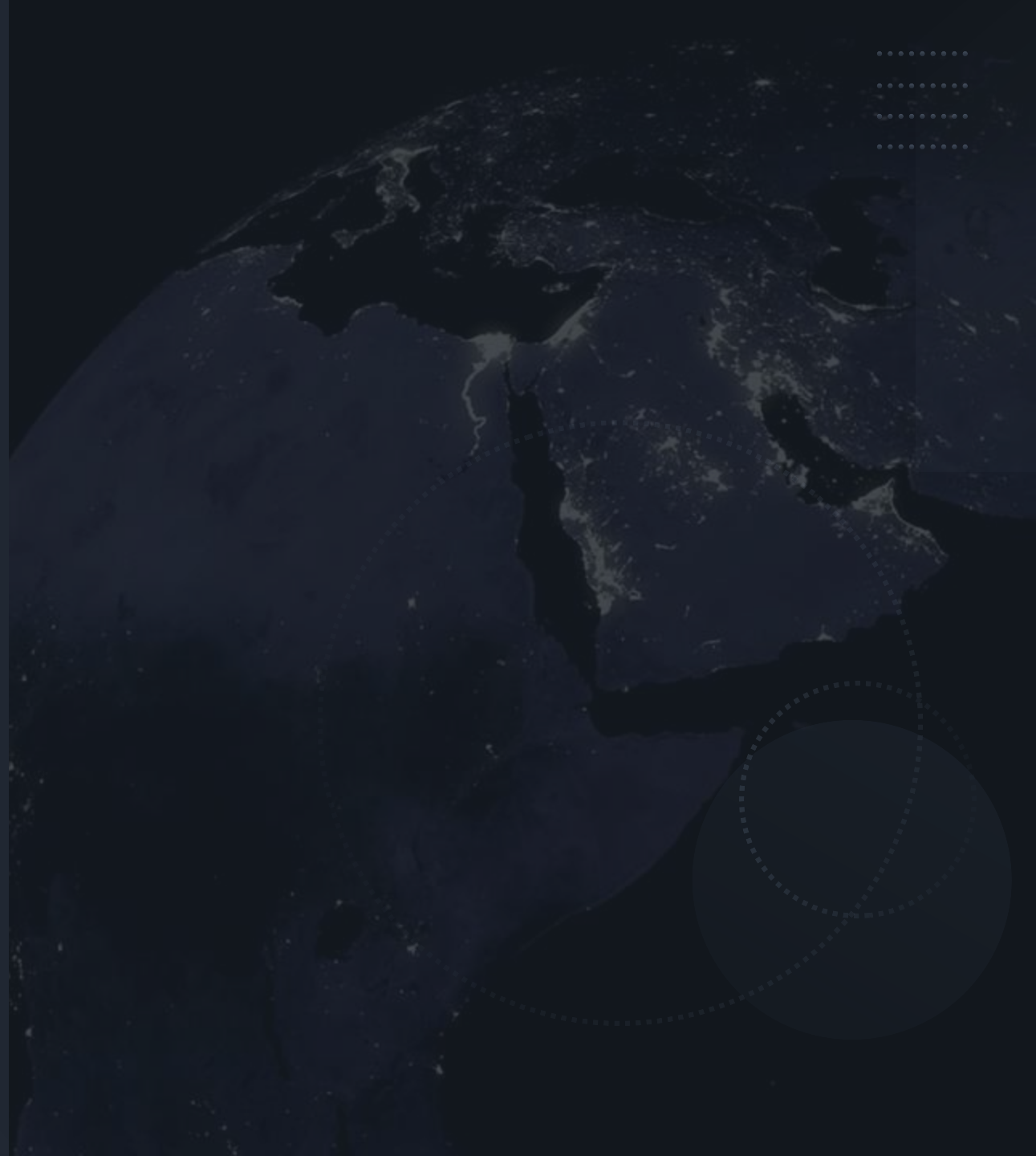
Predictive Analysis (Classification)

- Imported SpaceX Launch Data.
- Preprocess and test/train split 20/80.
- Define Logistic Regression, SVM, Decision Tree, and KNN objects.
- Use Grid Search with 10-fold cross validation to find best parameters for each model.
- For each model, fit with training data and evaluate on testing data.
- Compare accuracy metrics to determine model with best fit.

Appendix: Machine Learning Predictions

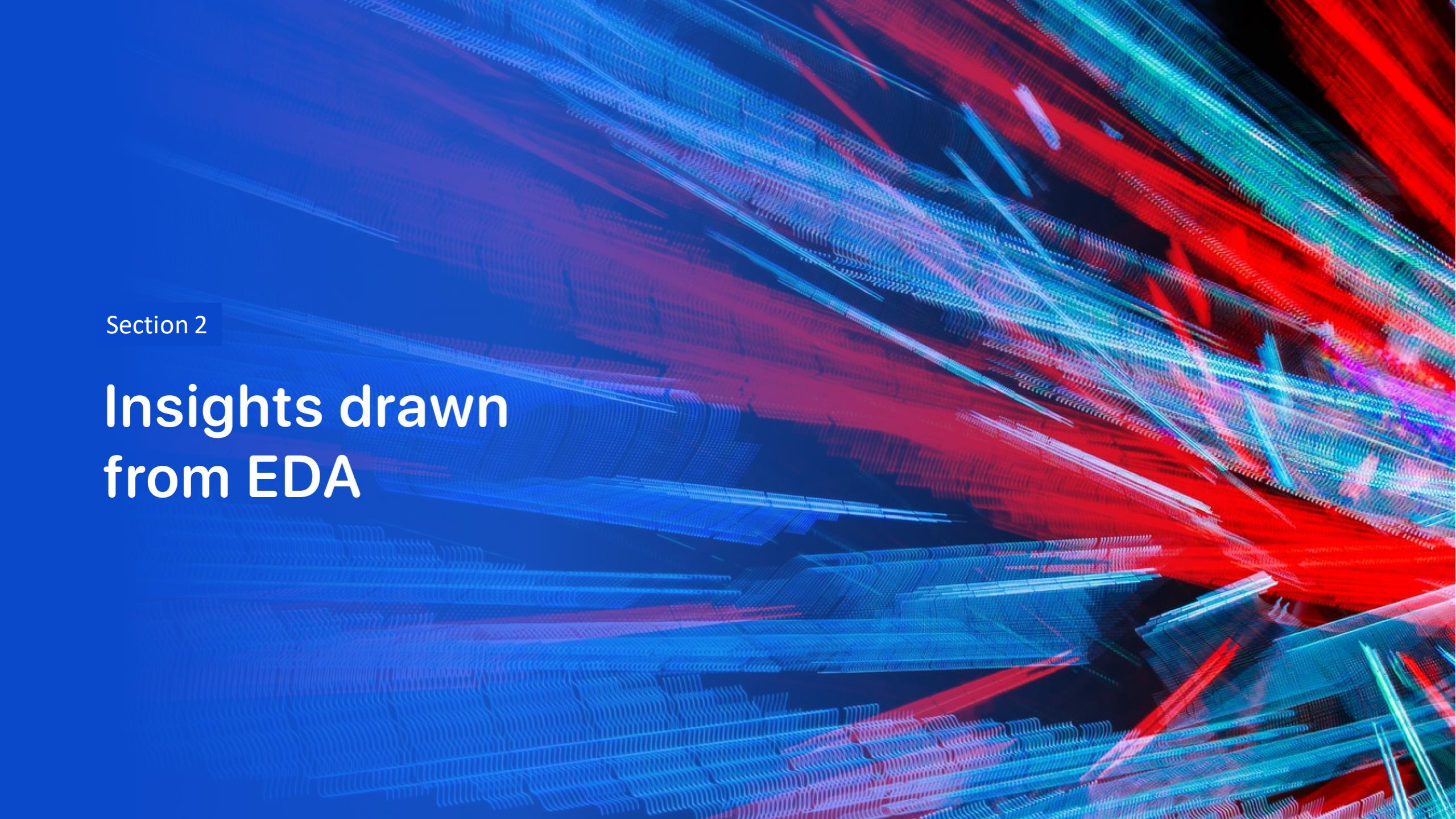
Results

- Sites tend to have fewer failed landings as flight number and time increase.
- All launch sites are within 10km of a large ocean, with three of the sites being in Florida within 8km of each other.
- The site KSC LC-39A has the highest landing success rate of 76.9%, using boosters B4, B5, and FT, with total success for payload mass less than 5,500kg.
- Based on accuracy metrics, the Decision Tree model is best at predicting the outcome of a rocket's first-stage landing with around 90% accuracy.



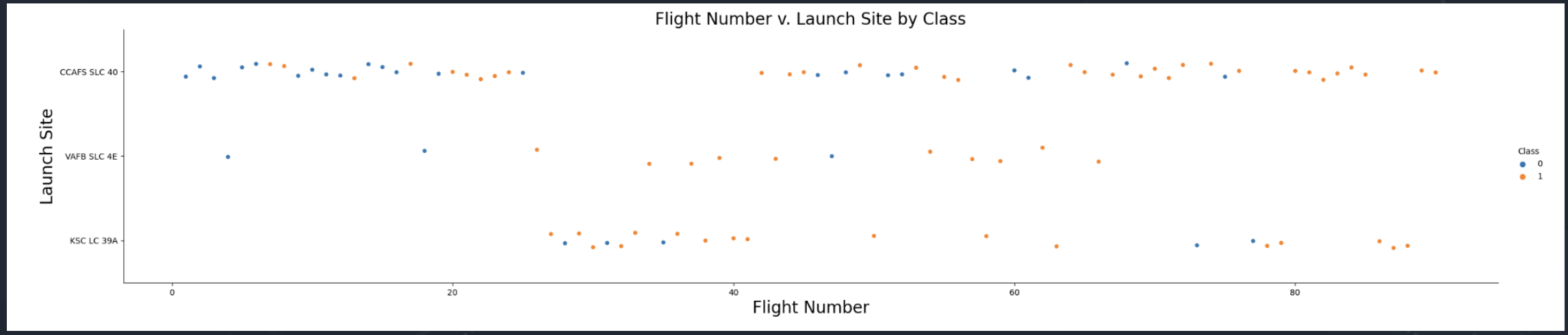
Section 2

Insights drawn from EDA



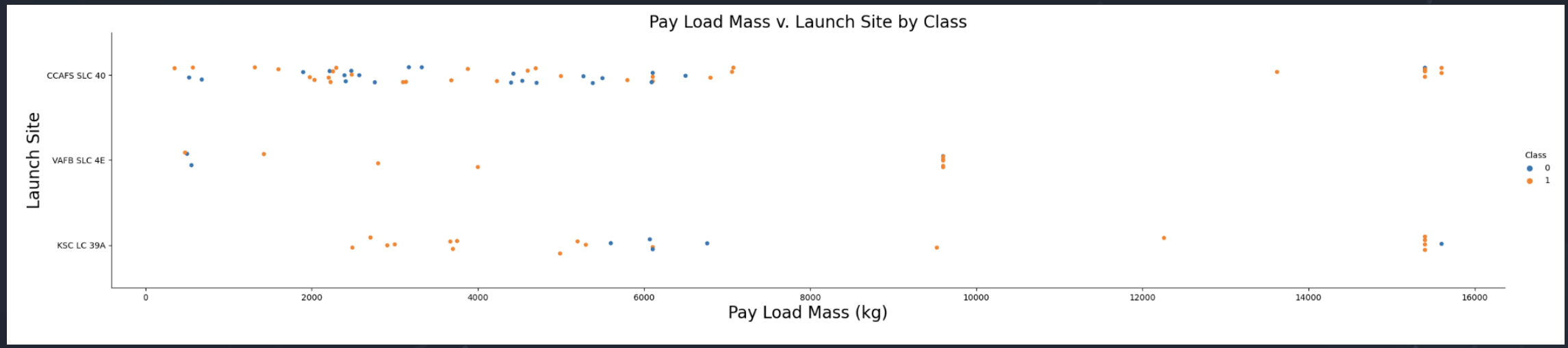
Flight Number vs. Launch Site

- VAFB SLC 4E launched the fewest rockets, with flight numbers under 20 being unsuccessful.
- CCAFS SLC-40 launched the most rockets with higher flight numbers having a better chance of success.



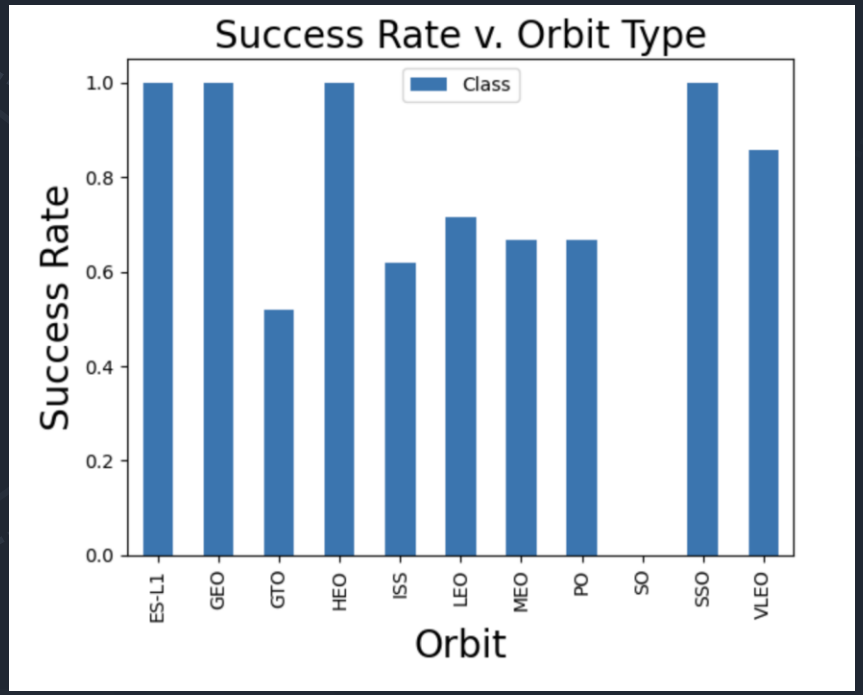
Payload vs. Launch Site

- Rockets with higher payloads have a greater chance of landing successfully across all sites.
- VAFB SLC 4E did not launch any rockets with a payload greater than 10,000.



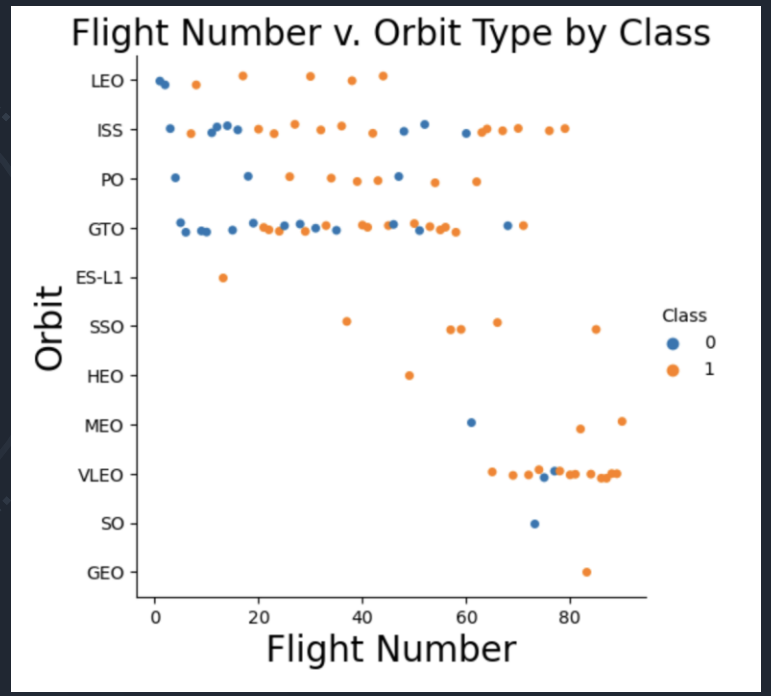
Success Rate vs. Orbit Type

- The orbits with no failed landings are ES-L1, GEO, HEO, and SSO.
- No launches with the SO orbit landed successfully.



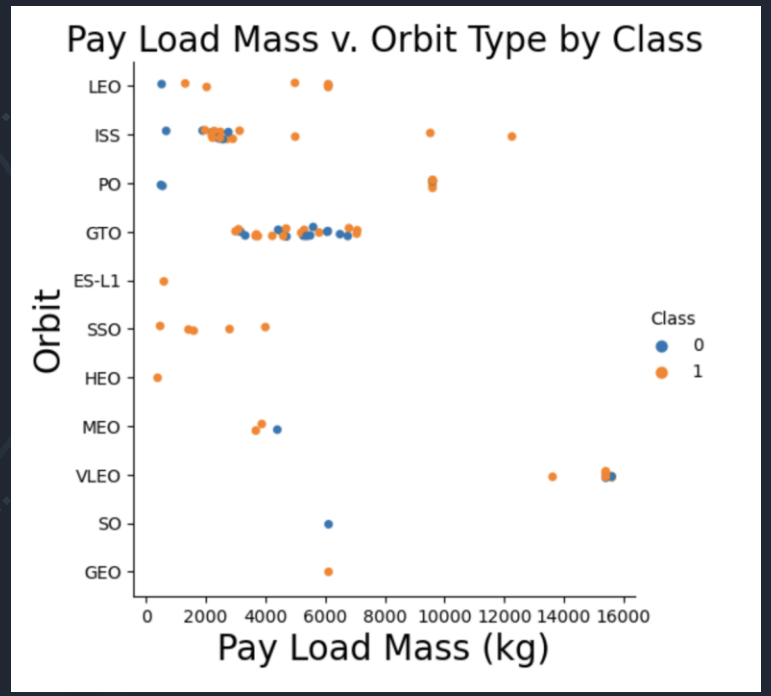
Flight Number vs. Orbit Type

- A site tends to have fewer failed landings as flight number increases.
- The sites with only one launch are ES-L1, HEO, SO, and GEO



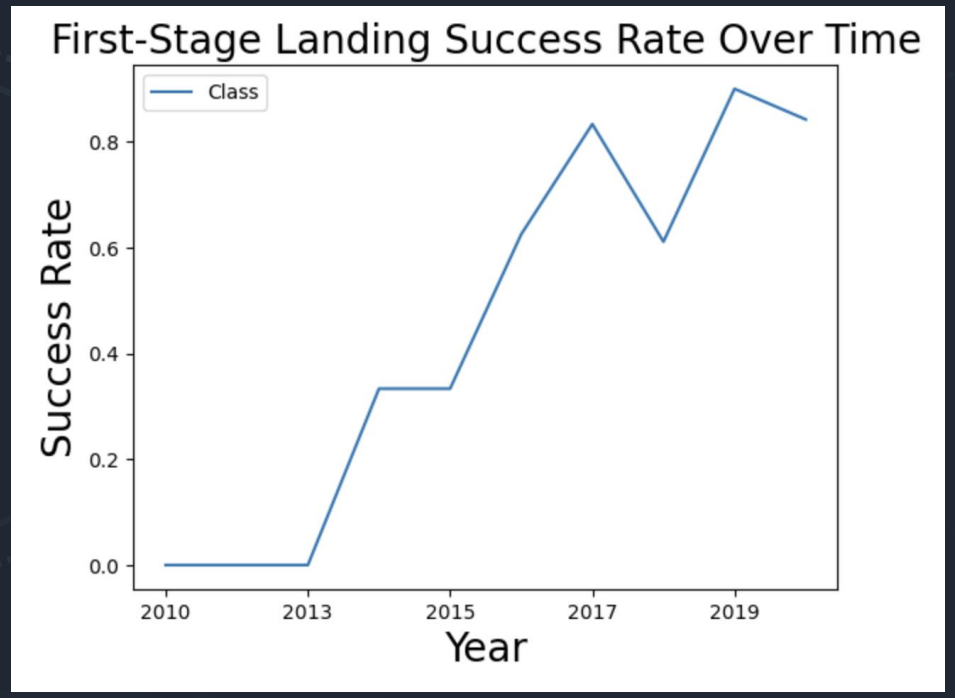
Payload vs. Orbit Type

- GTO launches rockets with a payload between 3000 and 800kg.
- LEO, ISS, and PO had more successful landings with higher payloads.



Launch Success Yearly Trend

- There has generally been an increase in the first-stage landing success rate over time.
- There was no increase in success rate during 2010-2013 and a decrease from 2017-2018.



All Launch Site Names

- These are the distinct launch sites included in this study.

```
%sql select distinct "Launch_Site" from SPACEXTABLE
```

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

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

Launch Site Names Begin with 'CCA'

- These are 5 records whose launch sites begin with CCA.

```
%sql select * from SPACEXTABLE where "Launch_Site" like "CCA%"  limit 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcon
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachut
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 (parachut
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attem
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attem
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attem

Total Payload Mass

- This is the total payload mass in kilograms carried by boosters from NASA (CRS).

```
%sql select sum("PAYLOAD_MASS__KG_"), "Customer" from SPACEXTABLE where "Customer"="NASA (CRS)"
```

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

sum("PAYLOAD_MASS__KG_")	Customer
45596	NASA (CRS)

Average Payload Mass by F9 v1.1

- This is the average payload mass carried by booster version F9 v1.1.

```
%sql select avg("PAYLOAD_MASS__KG_") from SPACEXTABLE where "Booster_Version" like "F9 v1.1%"
```

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

avg("PAYLOAD_MASS__KG_")
2534.6666666666665

First Successful Ground Landing Date

- This is the date of the first successful landing outcome on ground pad.

```
%sql select min("Date"),"Landing_Outcome" from SPACEXTABLE where "Landing_Outcome" == "Success (ground pad)"
```

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

min("Date")	Landing_Outcome
2015-12-22	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

- These are the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000kg but less than 6000kg.

```
%sql select distinct("Booster_Version"),"Landing_Outcome","PAYLOAD_MASS__KG_" from \
    (select * from SPACEXTABLE where "Landing_Outcome" == "Success (drone ship)") \
    where "PAYLOAD_MASS__KG_" > 4000 and "PAYLOAD_MASS__KG_" < 6000
```

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

Booster_Version	Landing_Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

Total Number of Successful and Failure Mission Outcomes

- These are the total numbers of successful and failure mission outcomes.

List the total number of successful and failure mission outcomes

```
%sql SELECT "Mission_Outcome", count(*) as Count FROM SPACEXTABLE GROUP BY 1
```

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

Mission_Outcome	Count
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- These are the names of the boosters which have carried the maximum payload mass.

```
%sql select "Booster_Version","PAYLOAD_MASS__KG_" from SPACEXTABLE \
    where "PAYLOAD_MASS__KG_" = (select max("PAYLOAD_MASS__KG_") from SPACEXTABLE)
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

- These are the failed landing outcomes in drone ship, their booster versions, and launch site names for the year 2015.

```
%sql select "Date",substr("Date",-1,3) as Month, substr("Date",0,5) as Year, \
    "Landing_Outcome", "Booster_Version", "Launch_Site" from \
    (select * from SPACEXTABLE where "Landing_Outcome" == "Failure (drone ship)") \
    where Year == "2015"
```

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

Date	Month	Year	Landing_Outcome	Booster_Version	Launch_Site
2015-10-01	1	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	4	2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- This is a rank of the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.

```
%sql select count("Landing_Outcome"),"Landing_Outcome"  from SPACEXTABLE \
    where "Date">="2010-06-04" and "Date"<="2017-03-20" \
    group by "Landing_Outcome" \
    order by count("Landing_Outcome") desc
```

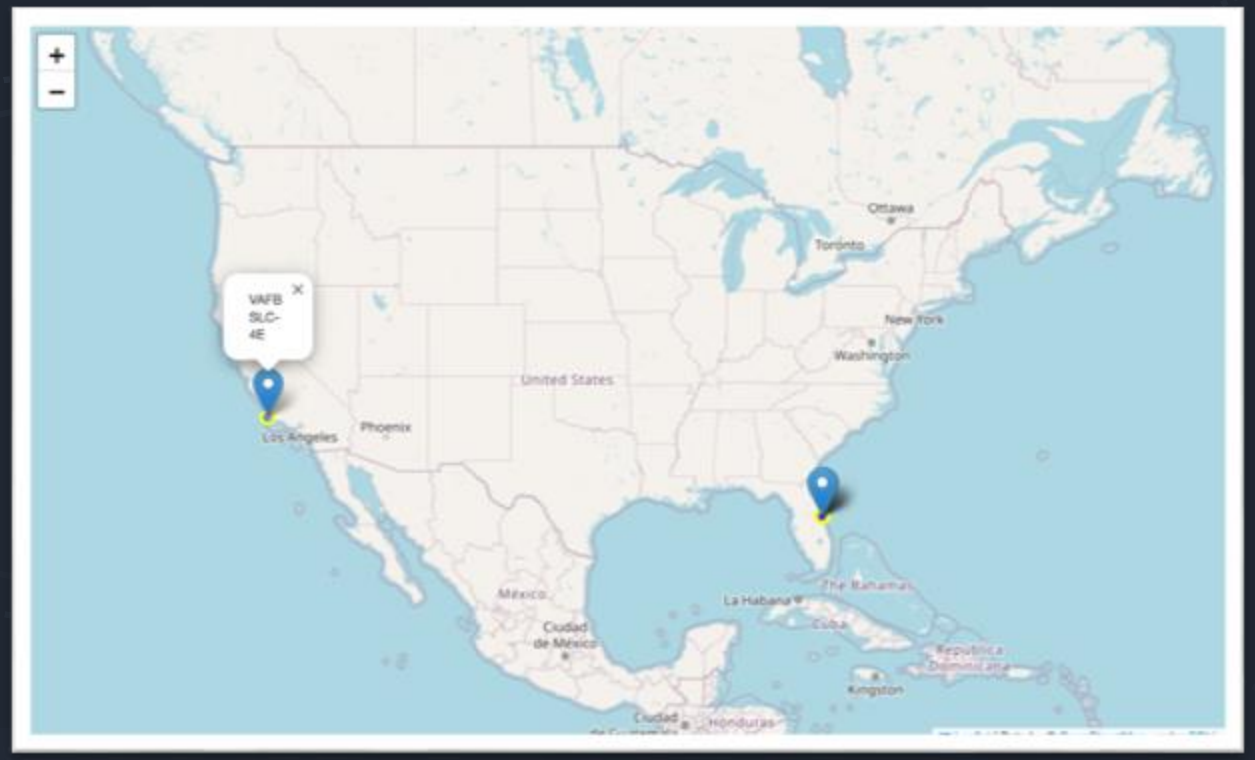
```
 * sqlite:///my_data1.db
Done.
```

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


Section 3
Launch Sites
Proximities Analysis

Map of Launch Locations

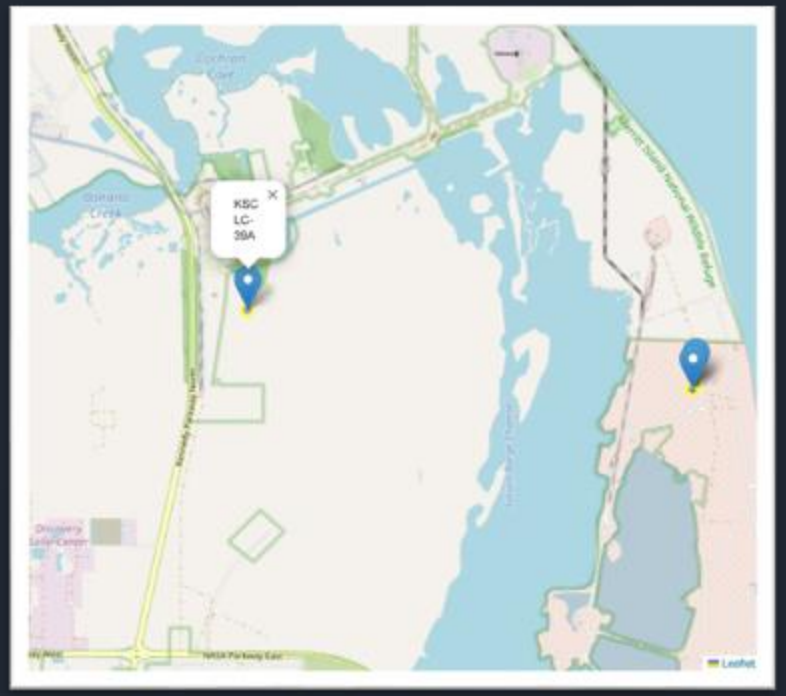
- All of the launch sites are located along coastlines, with VAFB SLC-4E in California and the others in Florida.
- All are along the southernmost coast of the United States near large oceans.

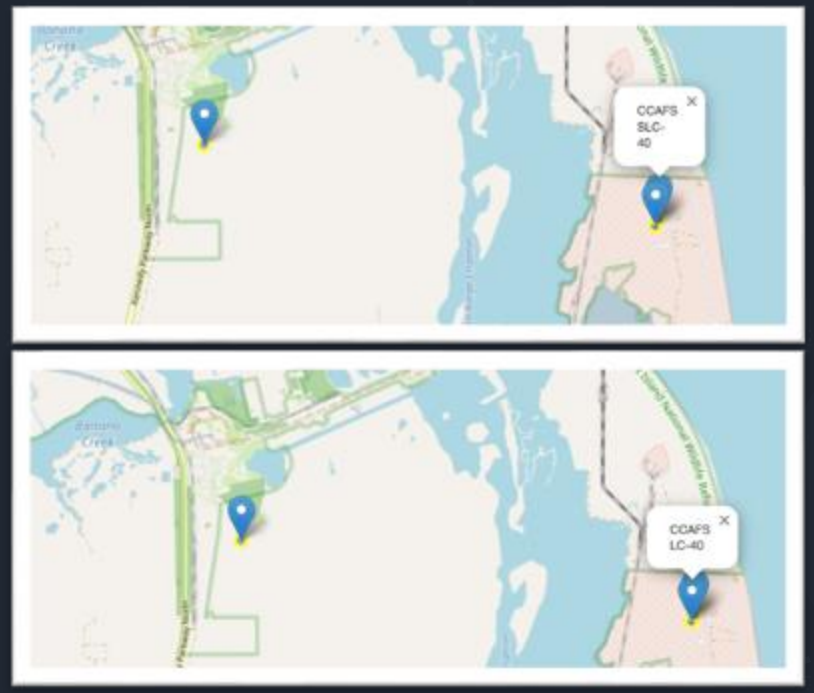


Appendix: Folium Map

Map of Launch Locations Cont.

- KSC LC-39A is set very close to the other two launch sites, which share land.
- CCAFS SLC-40 and CCAFS LC-40 both have direct access to the Atlantic Ocean, which is useful for recovering rockets that have landed in the ocean.

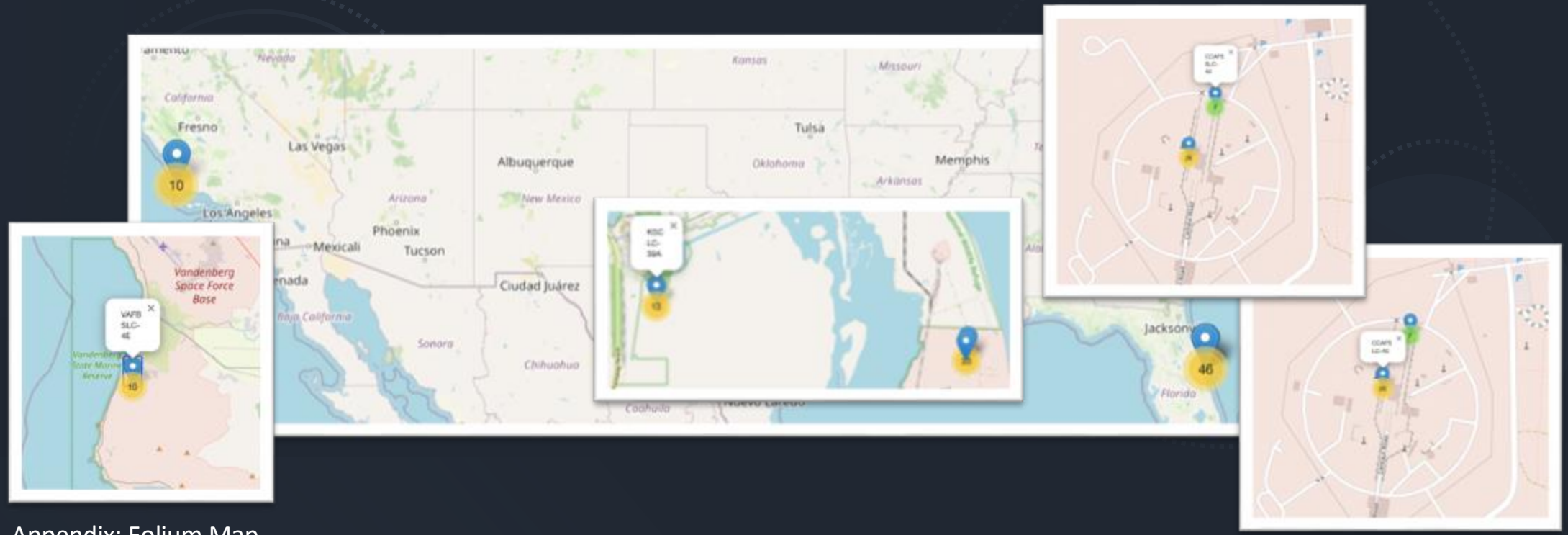




Appendix: Folium Map

Map of Launch Totals by Site

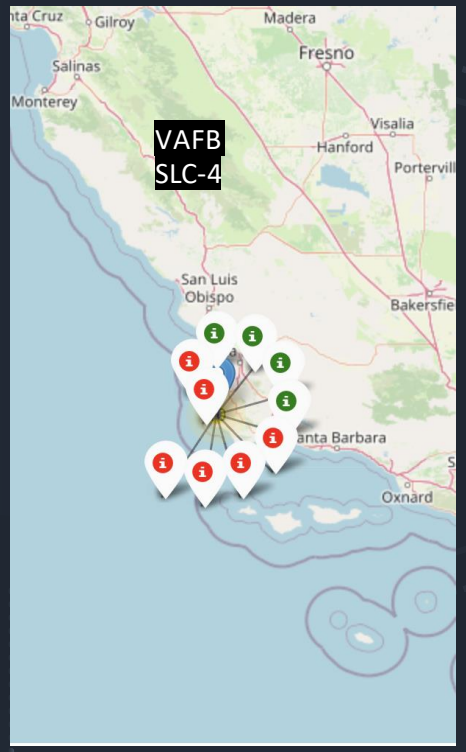
- Of the 56 launches in the dataset, the two launch sites with 33 combined launches were the most used site , CCAFS LC-40 with 26 launches, and the least used site, CCAFS SLC-40 with 7 launches.
- The California site, VAFB SLC-4, had 10 launches, with the three Florida sites totaling 46 launches.

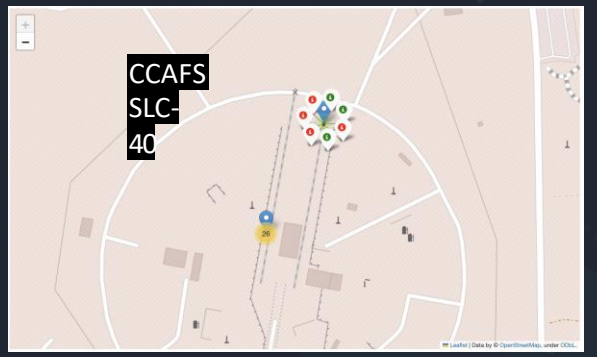


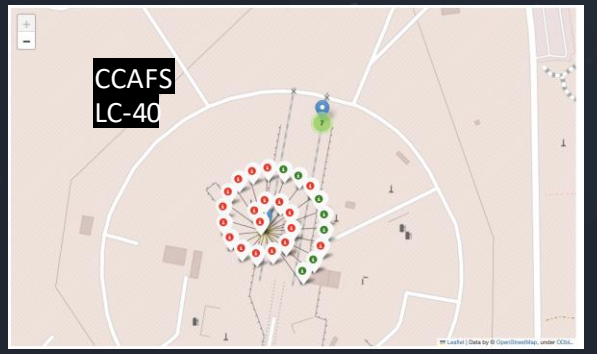
Appendix: Folium Map

Maps of Launch Successes by Site

- KSC LC-39A and CCAFS SLC-40 have a first-stage landing success rate of over 50%.
- CCAFS LC-40 has the most launches by far, but it also has the worst first-stage landing success rate.



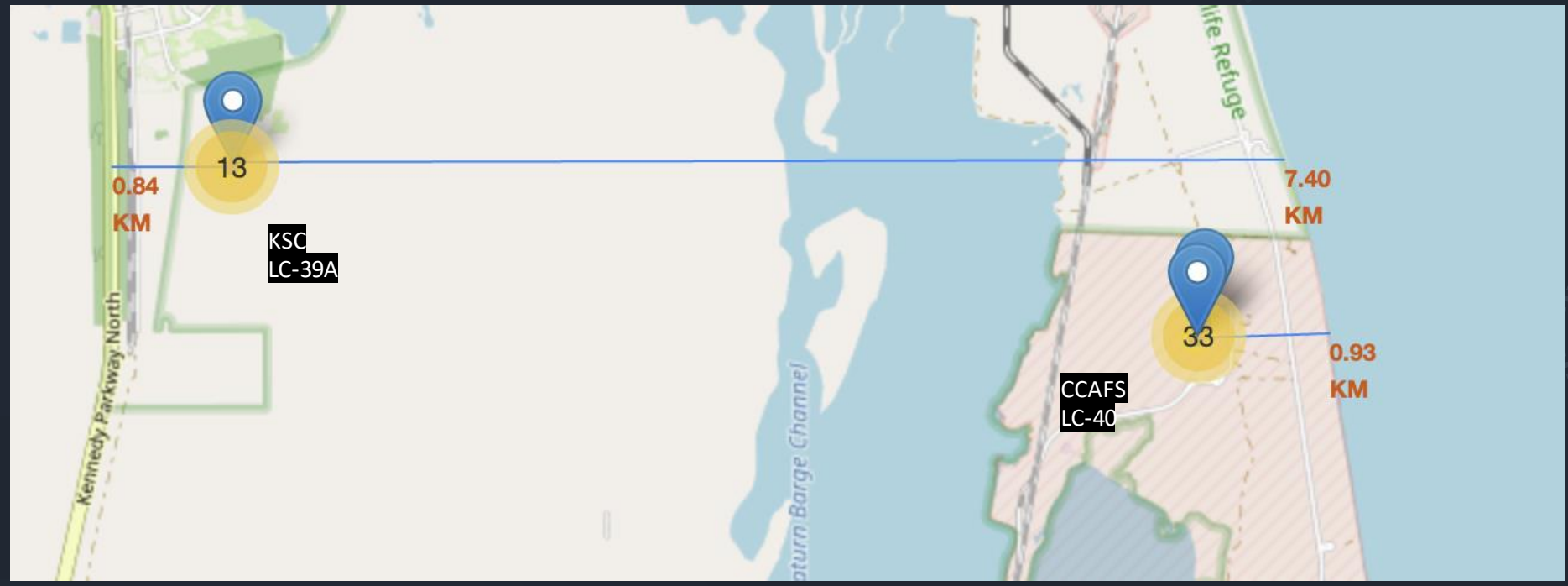




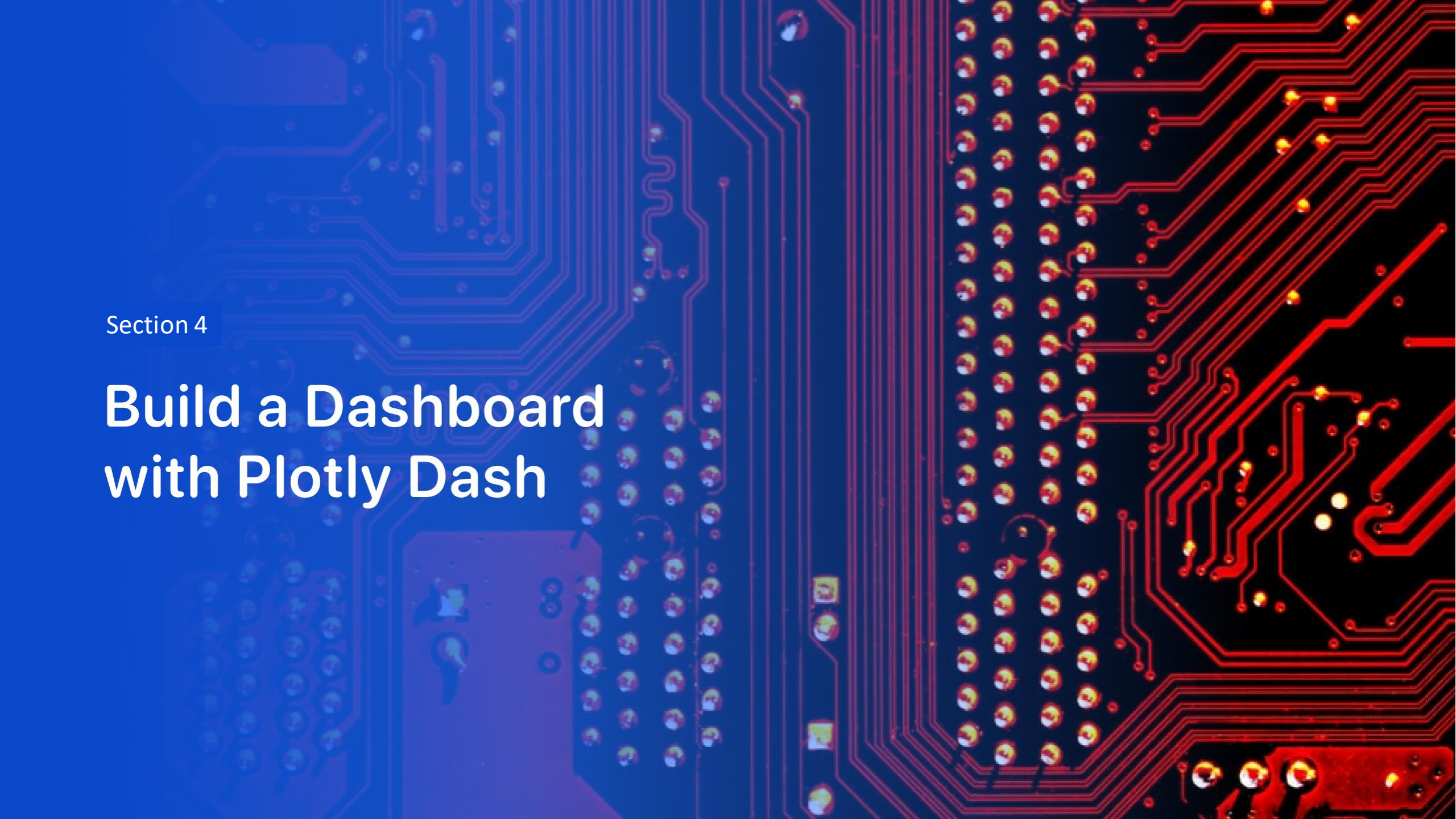
Appendix: Folium Map

Map of Launch Site Distances to Places of Interest

- KSC LC-39A is 6.47km further inland than the two closest sites, but it is much closer to a road.

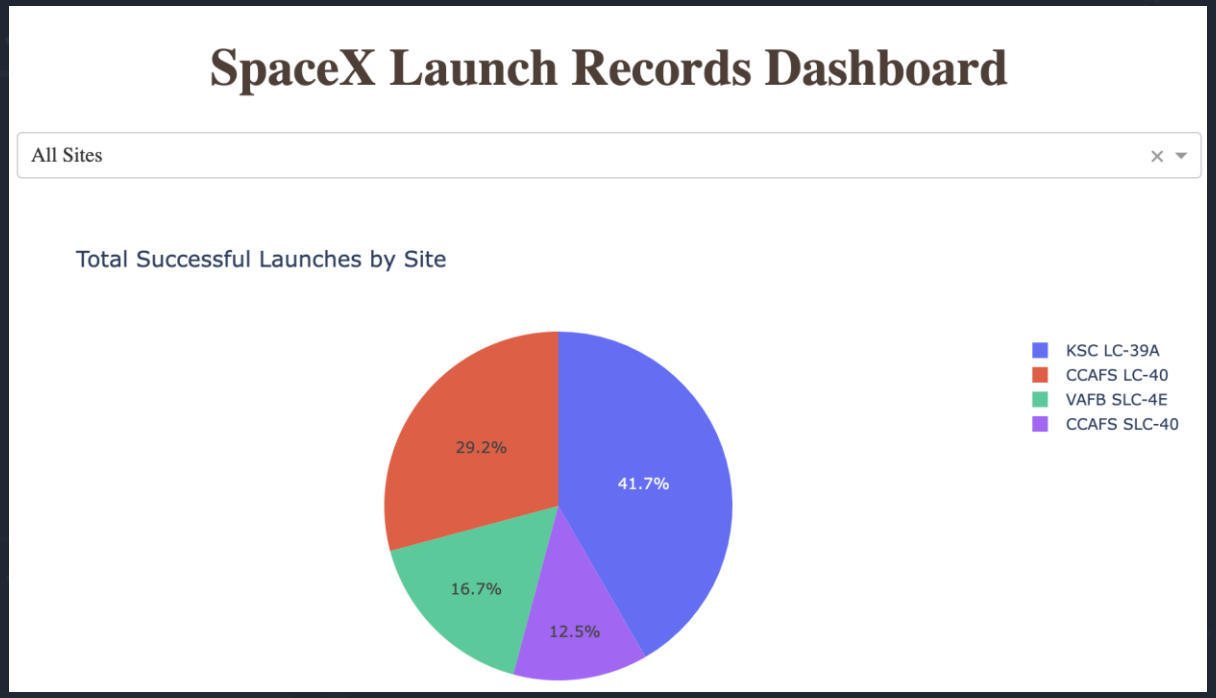


Appendix: Folium Map


Section 4
Build a Dashboard with Plotly Dash

Pie Chart of Successful Launches by Site

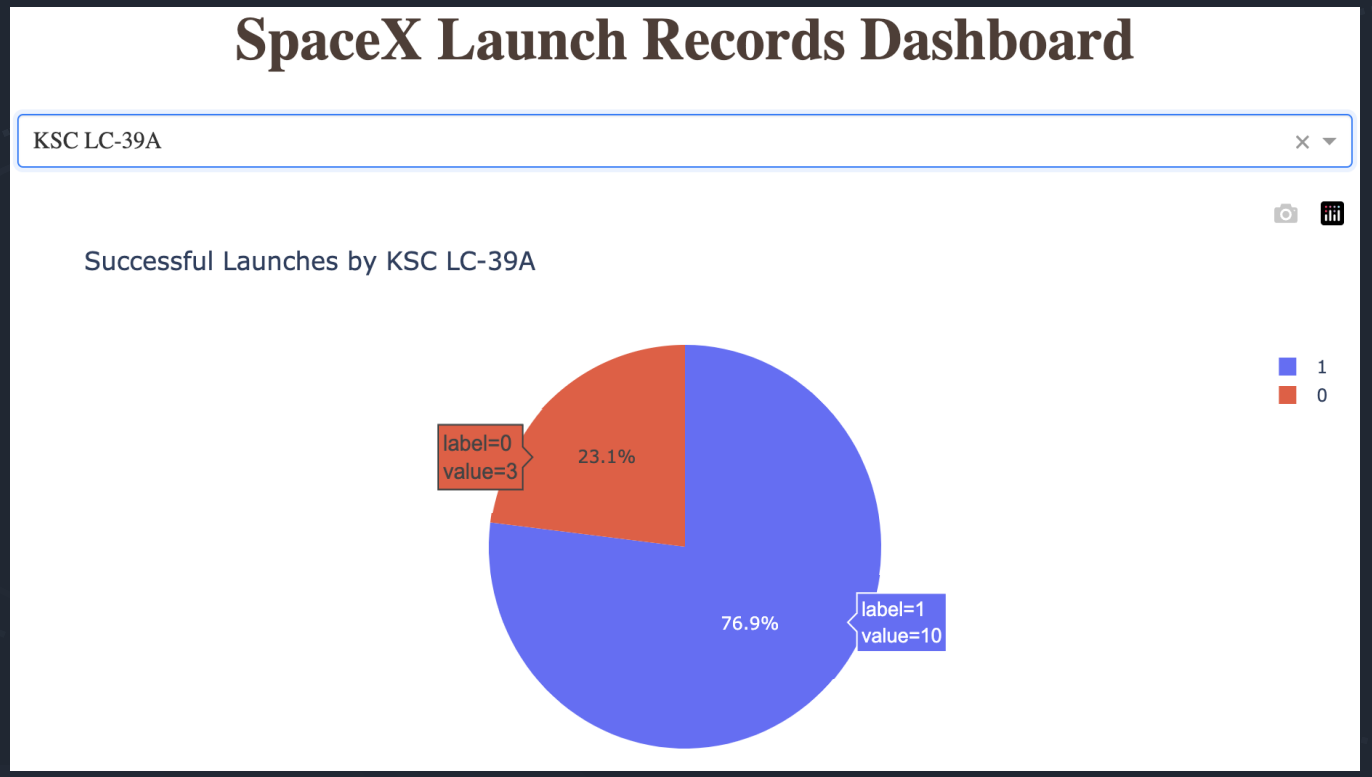
- Of the successful launches, KSC LC-39A had the most at 41.7% and CCAFS SLC-40 had the least at 12.5%.



Appendix: Plotly Dash Dashboard

Rates of Success for Launches

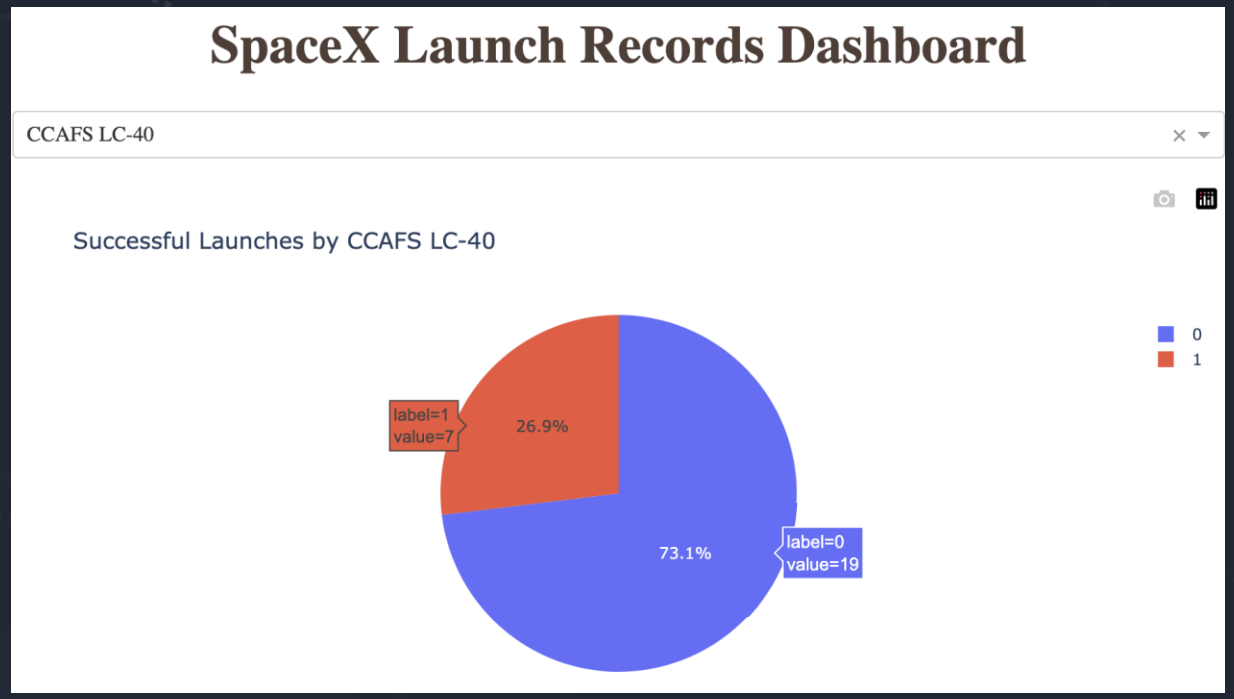
- KSC LC-39A has the highest success rate of 76.9% with 10 successful launches out of 13 total.



Appendix: Plotly Dash Dashboard

Rates of Success for Launches

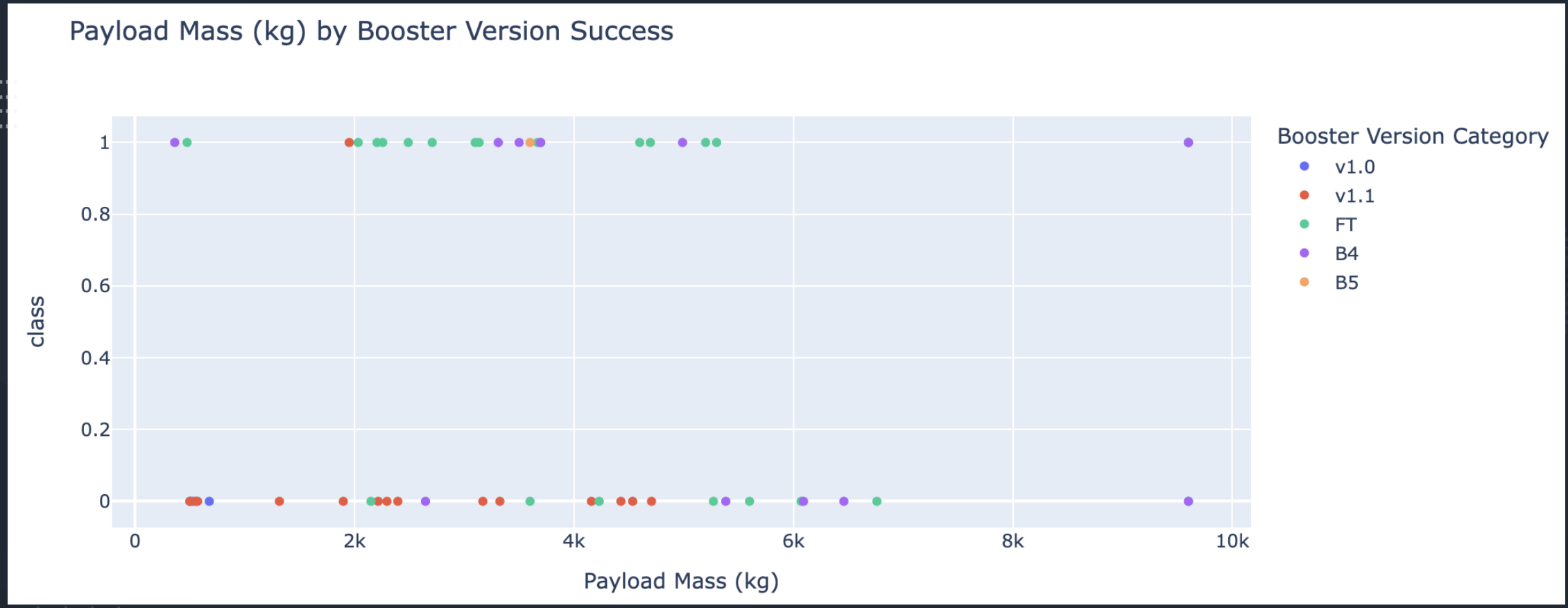
- CCAFS LC-40 had the most launches of all the sites at 26 total, but it had the worst success rate, with 73.1% (or 19) of their launches failing.
- CCAFS LC-40 makes up 29.2% of all successful launches with 7 successes, but this translates to a mere 26.9% success rate.



Appendix: Plotly Dash Dashboard

Payload vs. Launch Outcome by Site

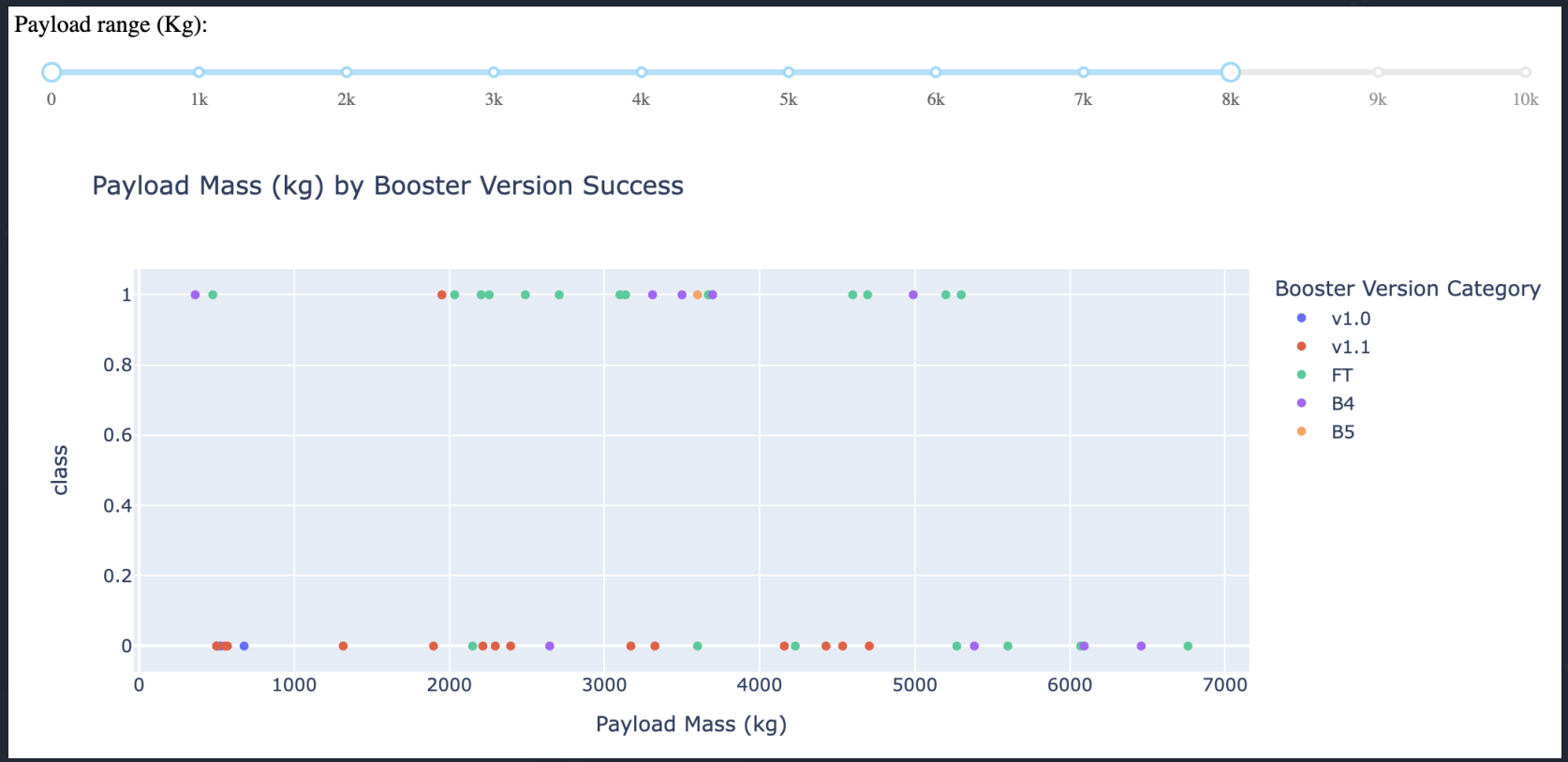
- There are few observations over 8,000kg, so the map can be limited.



Appendix: Plotly Dash Dashboard

Payload vs. Launch Outcome by Site

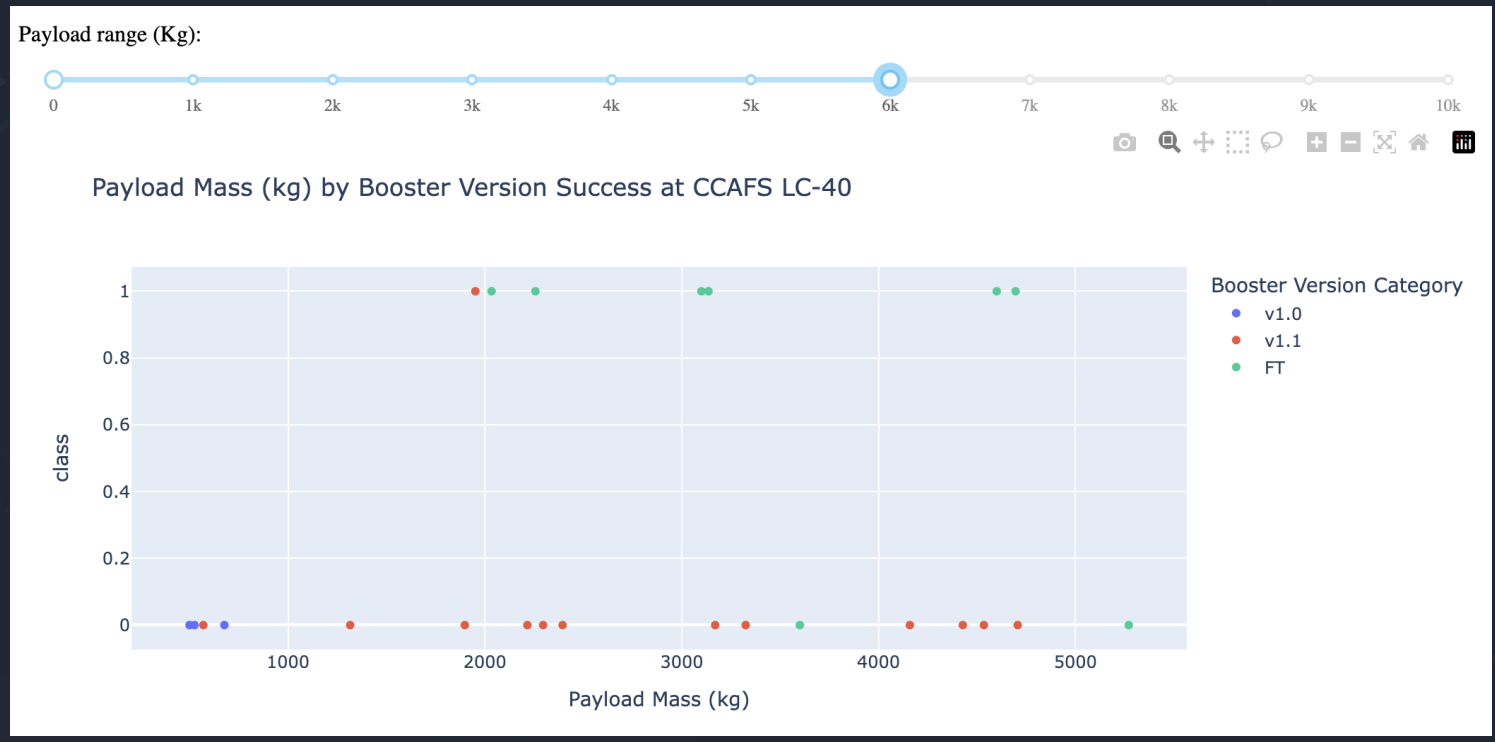
- There are few observations over 8,000kg, so the map can be limited.
- Initially, the V1.1 booster looks to be performing poorly.



Appendix: Plotly Dash Dashboard

Payload vs. Launch Outcome: CCAFS LC-40

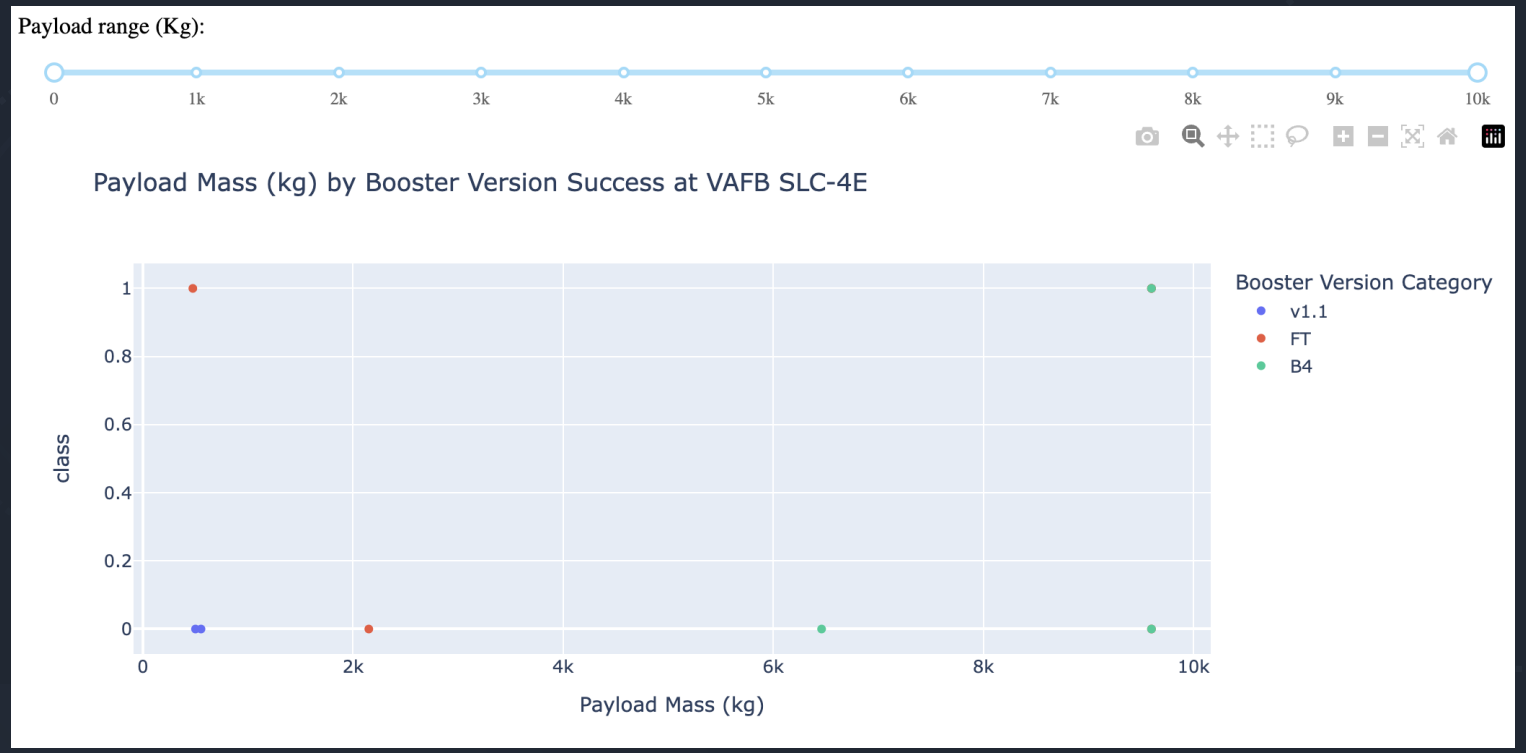
- There are no observations over 6,000kg, so the map can be limited.
- The V1.1 and v1.0 boosters look to be performing poorly.
- The FT booster is performing a little better than average.



Appendix: Plotly Dash Dashboard

Payload vs. Launch Outcome: VAFB SLC-4E

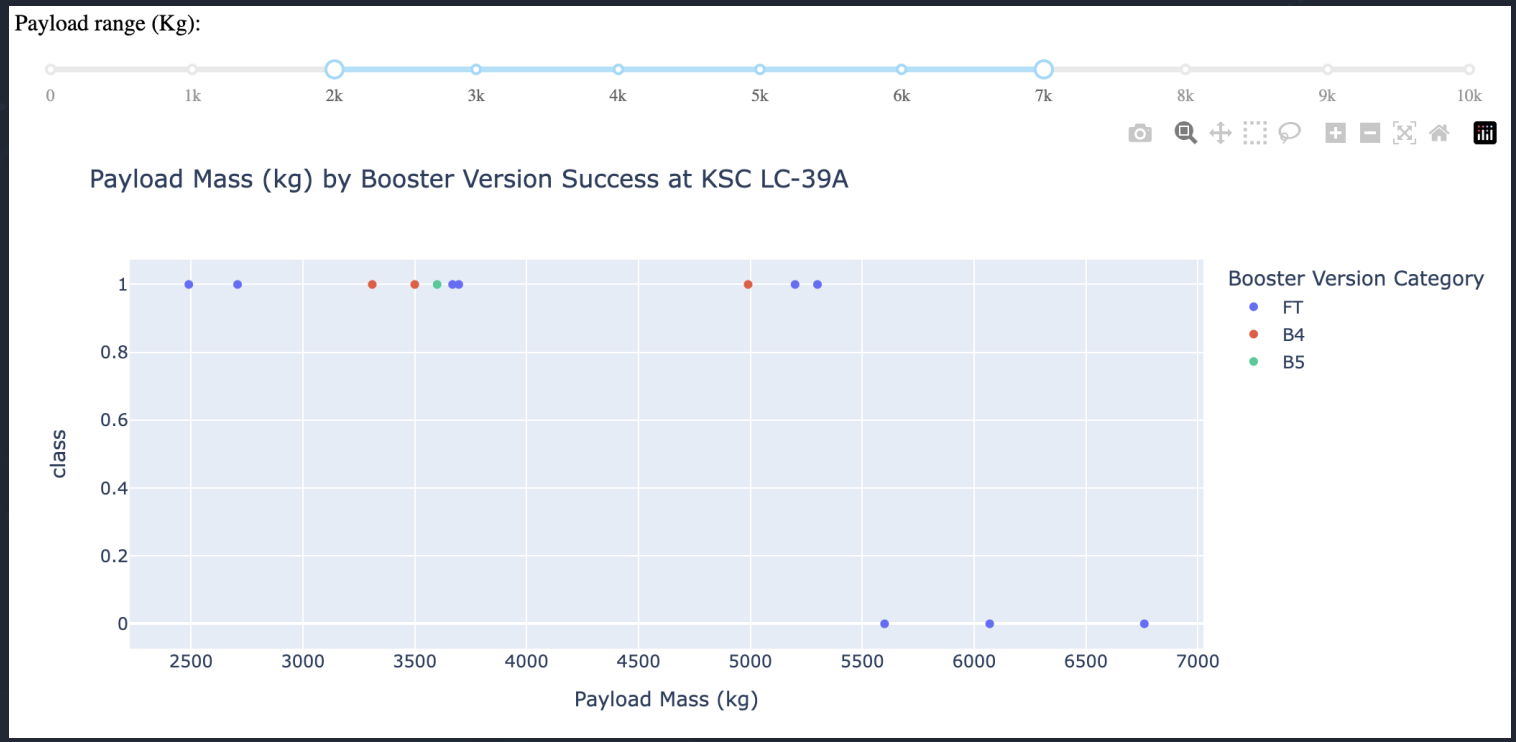
- There are no observations with payload between 3,000kg and 6,000kg.
- The V1.1 booster is still performing poorly.
- The FT booster is average for payload under 3,000kg and the B4 is below average for payload over 6,000kg.



Appendix: Plotly Dash Dashboard

Payload vs. Launch Outcome: KSC LC-39A

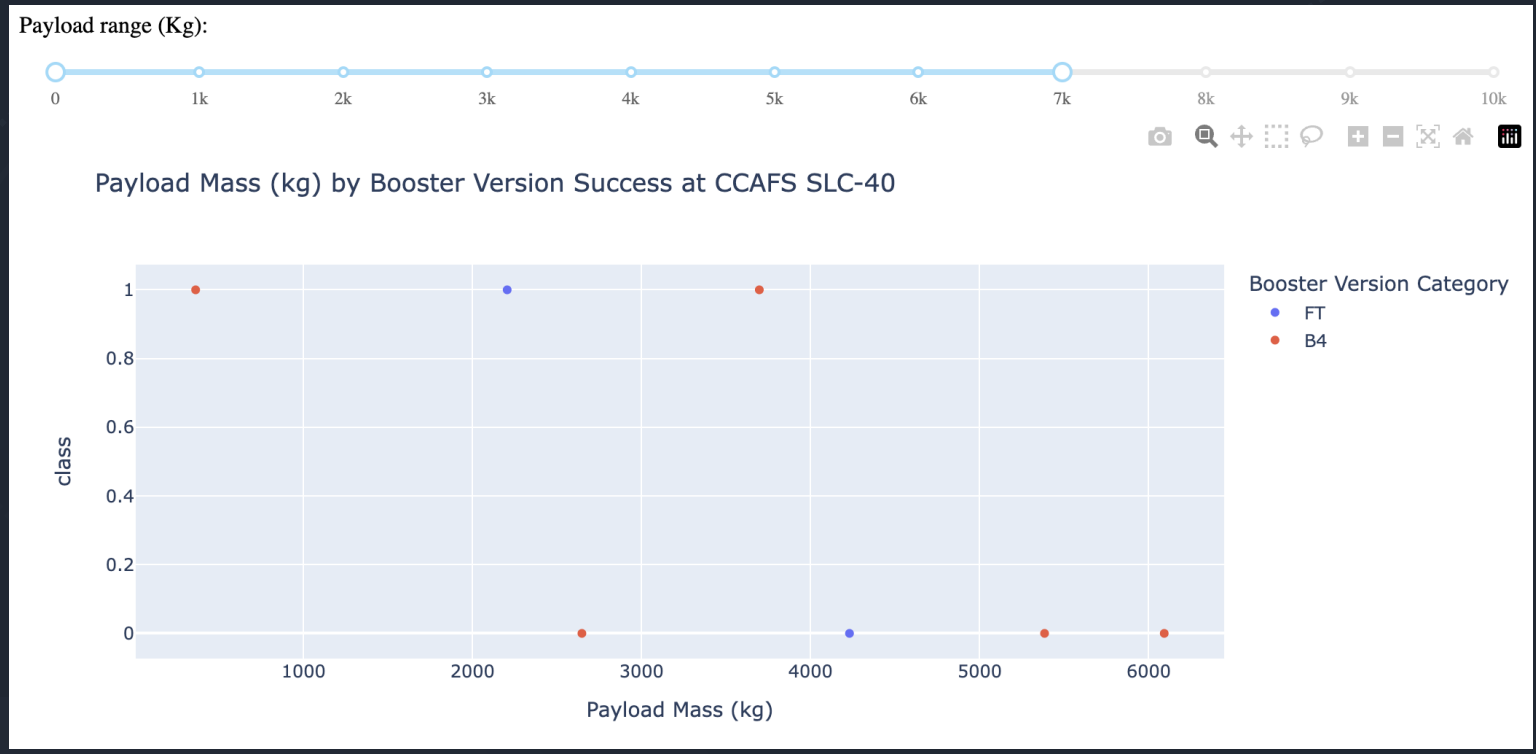
- There are no observations under 2,000 or over 7,000kg, so the map can be limited.
- KSC LC-39A had great success with all boosters for a payload under 5,500kg.
- Launches over 5,500kg used the FT booster and failed.



Appendix: Plotly Dash Dashboard

Payload vs. Launch Outcome: CCAFS SLC-40

- There are no observations over 7,000kg, so the map can be limited.
- Only the FT and B4 boosters were used with success less likely as the payload increased.



Appendix: Plotly Dash Dashboard

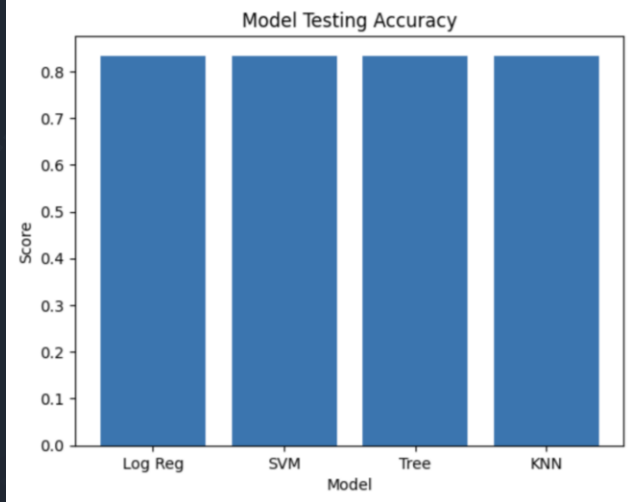
Section 5

Predictive Analysis (Classification)

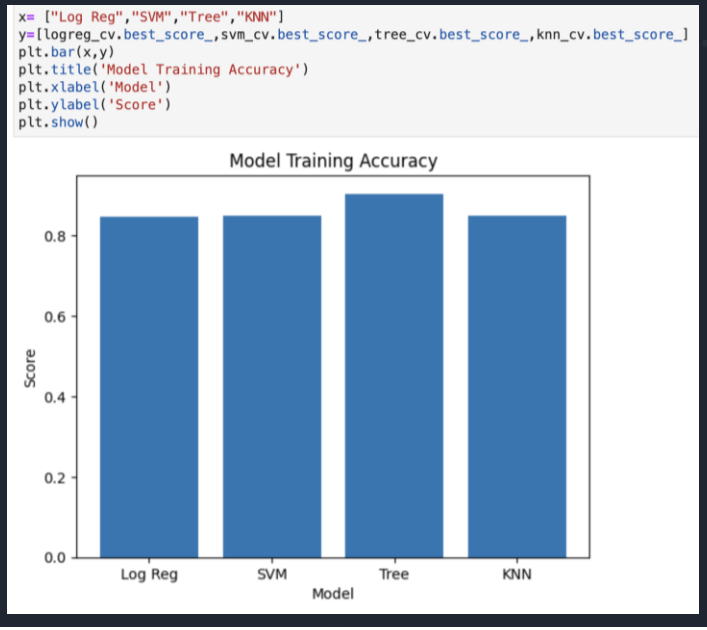
Classification Accuracy

- The Decision Tree model has the best training accuracy at 90.4%.
- KNN and SVM had the worst training accuracy of 84.8%.

```
x= ["Log Reg","SVM","Tree","KNN"]
y=[logregTS,svmTS,treeTS,knnTS]
plt.bar(x,y)
plt.title('Model Testing Accuracy')
plt.xlabel('Model')
plt.ylabel('Score')
plt.show()
```




```
x= ["Log Reg","SVM","Tree","KNN"]
y=[logreg_cv.best_score_,svm_cv.best_score_,tree_cv.best_score_,knn_cv.best_score_]
plt.bar(x,y)
plt.title('Model Training Accuracy')
plt.xlabel('Model')
plt.ylabel('Score')
plt.show()
```

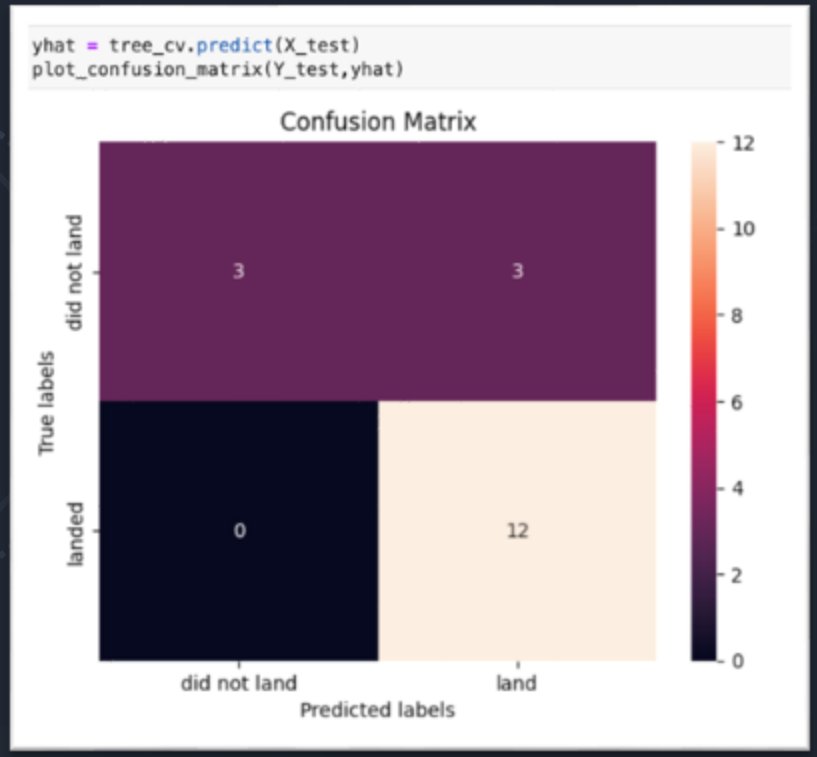


	Logistic Regression	Support Vector Machine	Decision Tree	K-Nearest Neighbors
Train Accuracy	0.846429	0.848214	0.903571	0.848214
Test Score	0.833333	0.833333	0.833333	0.833333

Appendix: Machine Learning Predictions

Confusion Matrix

- For the tree model, 15 observations were correctly identified, but 3 launches whose first stage failed to land were misclassified.



Conclusions

- Sites tend to have fewer failed landings as flight number and time increase.
- The orbits with no failed landings are ES-L1, GEO, HEO, and SSO.
- All launch sites are within 10km of a large ocean, with three of the sites being in Florida within 8km of each other.
- The site KSC LC-39A has the highest landing success rate of 76.9%, using boosters B4, B5, and FT, with total success for payload mass less than 5,500kg.
- The v1.0 and v1.1 boosters were highly unsuccessful across all sites.
- Based on accuracy metrics, the Decision Tree model is best at predicting the outcome of a rocket's first-stage landing with around 90% accuracy.

Appendix

All Code for Visualizations is on GitHub:

Data Collection API

Data Collection Webscraping

Data Wrangling

EDA with Data Visualization

EDA with SQL

Folium Map

Plotly Dash Dashboard

Machine Learning Predictions

Thank you!