Winning Space Race with Data Science

Abagael Barba
09/23/2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection via Web Scrapping and API
  - Data Wrangling
  - EDA with SQL and Data Visualization
  - Interactive Maping using Folium
  - Interactive Dashboard using Plotly Dash
  - Predictive Analysis
- Summary of all results
  - EDA results
  - Interactive Results
  - Predictive Analysis Results

Introduction

- Project background and context
  - We will be looking at the overall data from SpaceX's Falcon 9 rocket to see how well or poorly this rocket preforms when compared against itself
- Problems you want to find answers
  - What factors determine if the rocket will land successfully?
  - The interaction amongst various features that determine the success rate of a successful landing.
  - What operating conditions needs to be in place to ensure a successful landing program.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected via web scraping from Wikipedia
  - Data was collected via API's using SpaceX database
- Perform data wrangling
  - One Hot Encoding data fields are applied for categorical features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

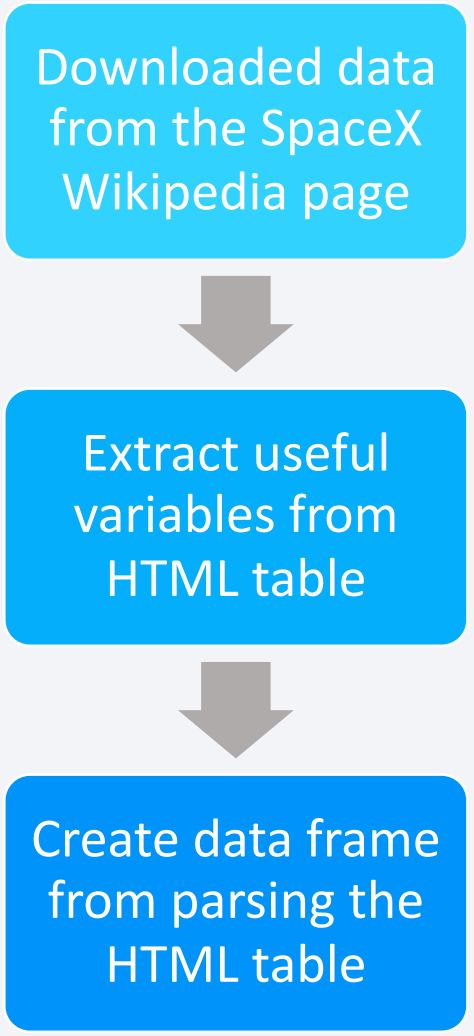
- Describe how data sets were collected.
  - Data collection was done using get request to the SpaceX API.
  - We then cleaned the data, checked for missing values and fill in missing values where necessary.
  - In addition, we performed web scraping from Wikipedia for Falcon 9 launch records with BeautifulSoup.
  - The objective was to extract the launch records as HTML table, parse the table and convert it to a pandas dataframe for future analysis.

Data Collection – SpaceX API

- Data was collected using the “get” request on the SpaceX API
- Data was then cleaned and reformatted in preparation for visualization
- https://github.com/Barba-Abby/Data-Science-Capstone

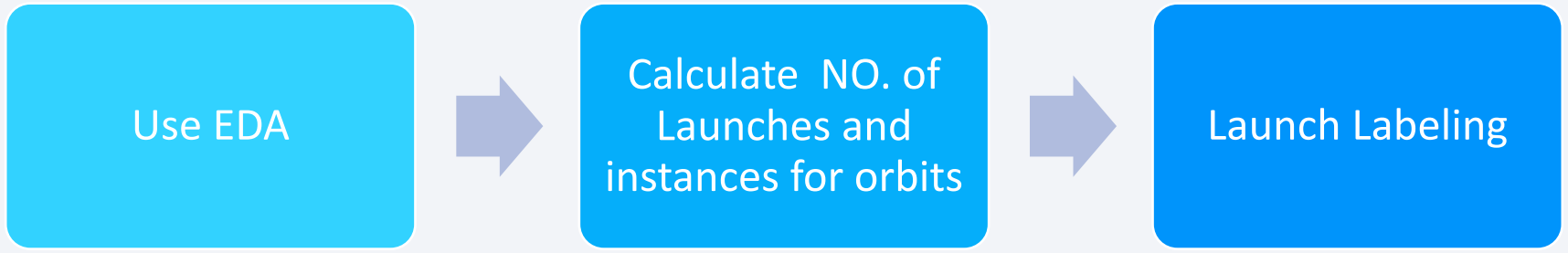
Data Collection - Scraping

- Web Scraping was done using BeautifulSoup to gather the data for launch records
- Data was converted to a pandas dataframe usable for later analysis
- https://github.com/Barba-Abby/Data-Science-Capstone



Data Wrangling

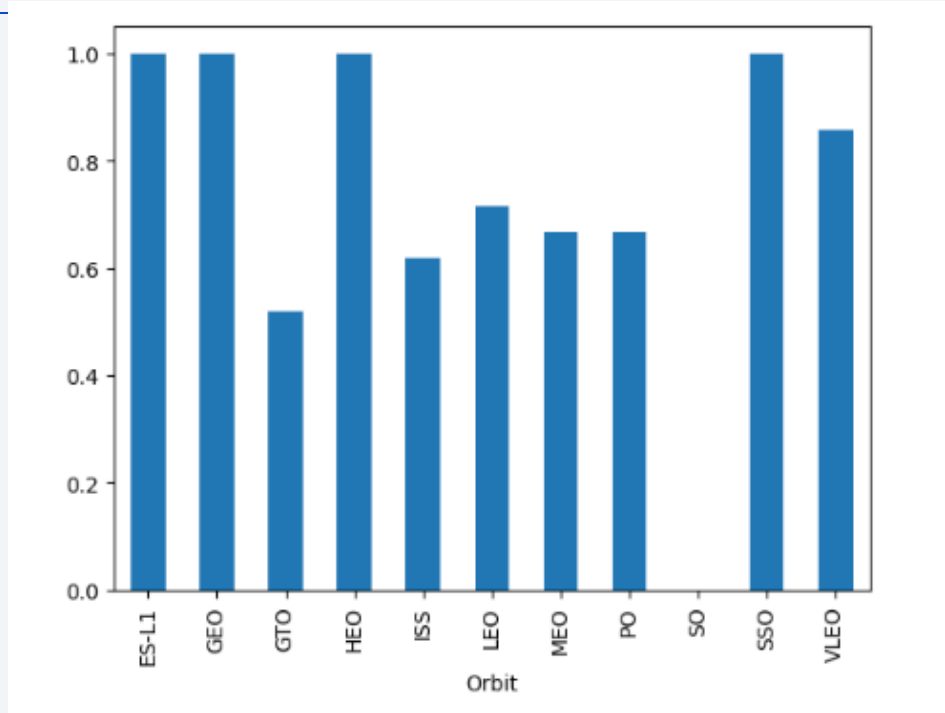
- Exploratory data analysis was done to the data frame to determine what information was critical to the Falcon 9's statistics
- Using data analytics, the number of orbits as well as launch site locations for each occurrence

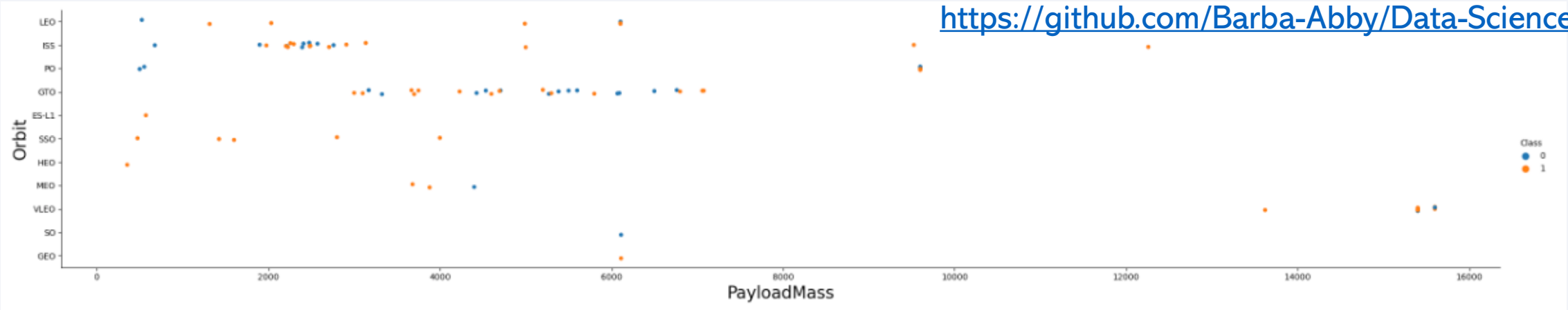


- https://github.com/Barba-Abby/Data-Science-Capstone

EDA with Data Visualization

- Scatter, column, and line charts were used to best visualize the relationship between the following:
  - flight number and launch site
  - payload and launch site, success rate of each orbit type
  - flight number and orbit type
  - the launch success yearly trend





https://github.com/Barba-Abby/Data-Science-Capstone

EDA with SQL

- EDA with SQL was used to gather statistical insight from the data.
- The following queries were used:
  - The names of unique launch sites in the space mission.
  - Top 5 launch sites beginning with “CCA”
  - The total payload mass carried by boosters launched by NASA (CRS)
  - The average payload mass carried by booster version F9 v1.1
  - The date of the first successful mission of Falcon 9
  - The total number of successful and failure mission outcomes
  - The failed landing outcomes in drone, booster version, and launch site
- https://github.com/Barba-Abby/Data-Science-Capstone

Build an Interactive Map with Folium

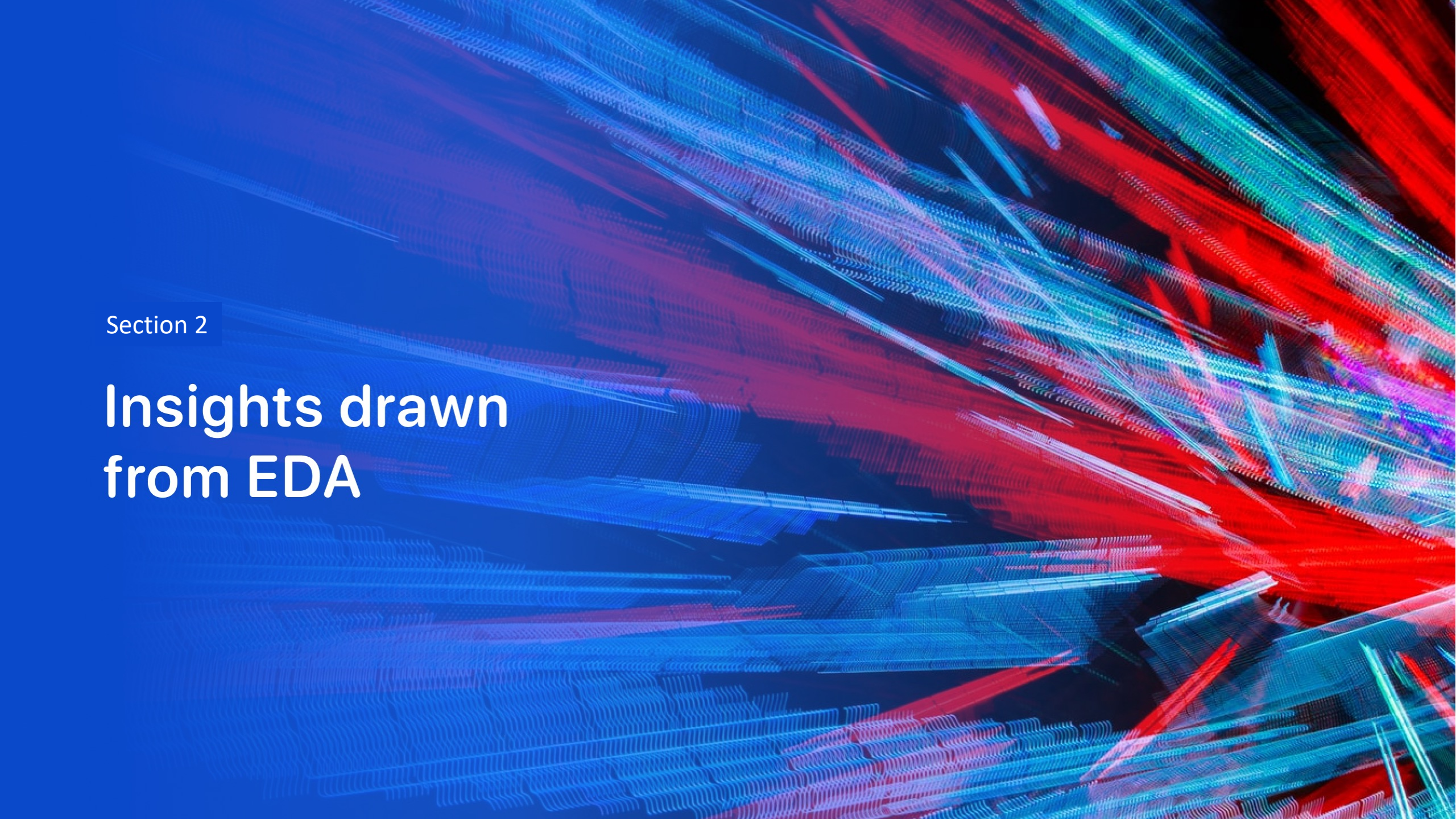
- Using markers and circles, launch sites were added to a map
- Lines showed distances between two or more sites
- Using a binary failure/success mode (0= failure, 1 = success), all launch sites were clearly marked accordingly
    - Using color-coded clustering, higher yield success sites are clearly marked
- https://github.com/Barba-Abby/Data-Science-Capstone

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
- https://github.com/Barba-Abby/Data-Science-Capstone

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
- You need present your model development process using key phrases and flowchart
- https://github.com/Barba-Abby/Data-Science-Capstone


Section 2
Insights drawn from EDA

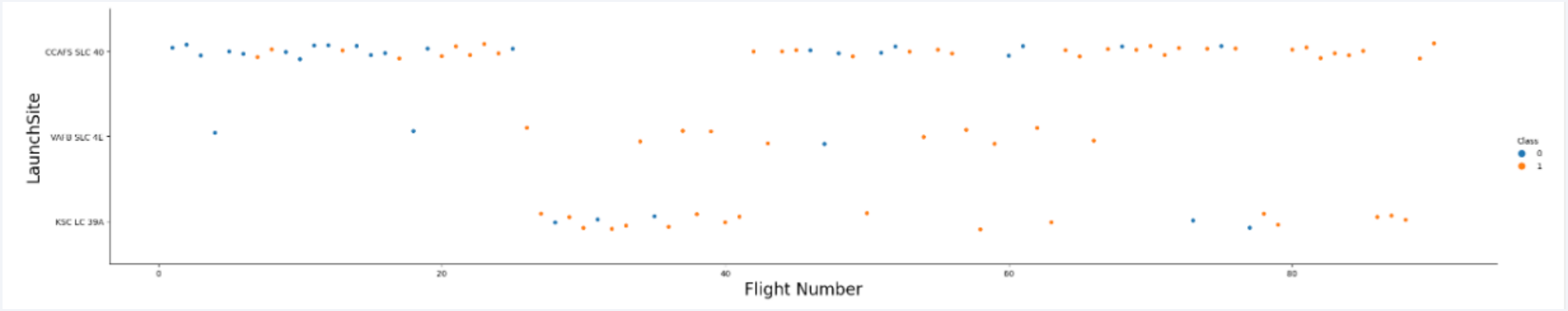
Flight Number vs. Launch Site

022 5:22 PM

TASK 1: Visualize the relationship between Flight Number and Launch Site

Use the function `catplot` to plot `FlightNumber` vs `LaunchSite`, set the parameter `x` parameter to `FlightNumber`,set the `y` to `Launch Site` and set the parameter `hue` to `'class'`

[4]:
```
# Plot a scatter point chart with x axis to be Flight Number and y axis to be the launch site, and hue to be the class value
sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("LaunchSite",fontsize=20)
plt.show()
```

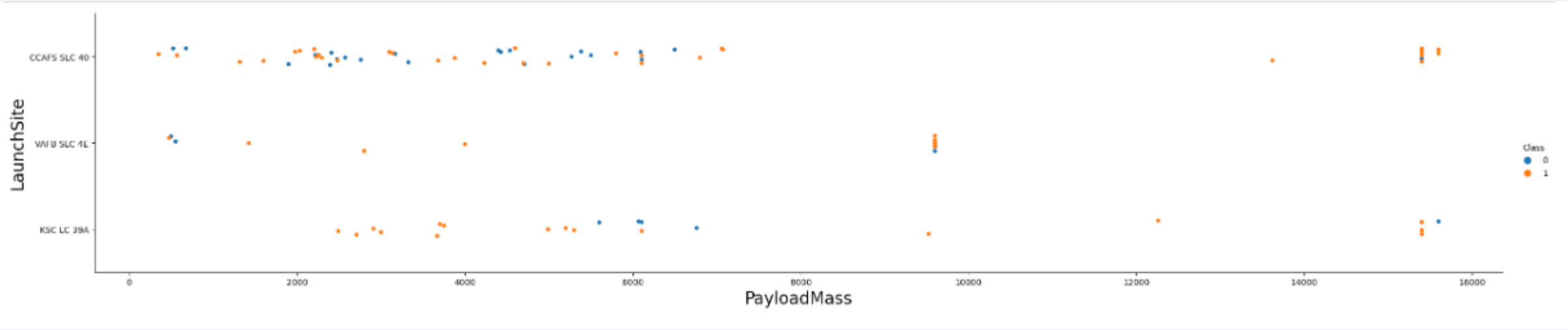


Payload vs. Launch Site

TASK 2: Visualize the relationship between Payload and Launch Site

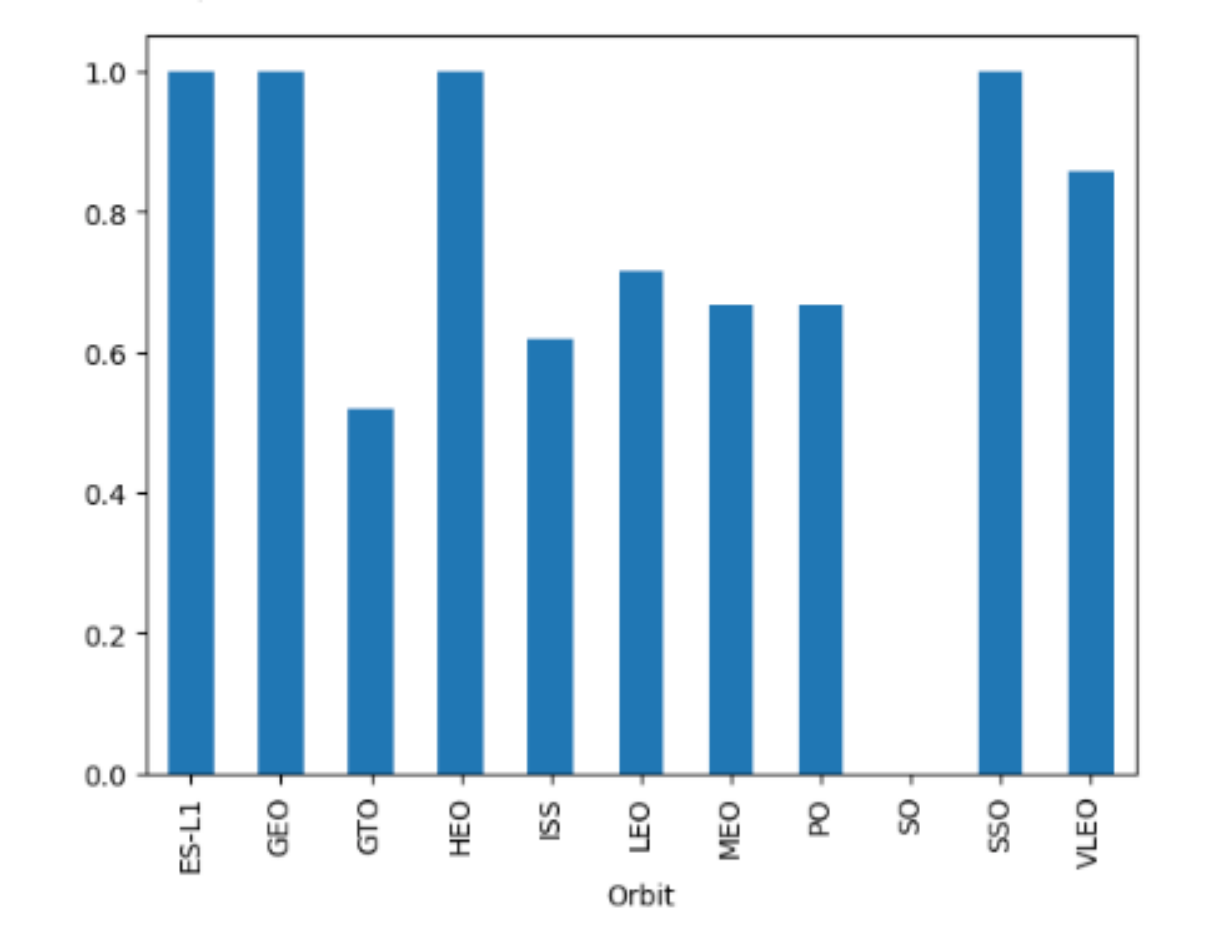
We also want to observe if there is any relationship between launch sites and their payload mass.

[5]:
```
# Plot a scatter point chart with x axis to be Pay Load Mass (kg) and y axis to be the launch site, and hue to be the class value
sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("PayloadMass",fontsize=20)
plt.ylabel("LaunchSite",fontsize=20)
plt.show()
```



Success Rate vs. Orbit Type

- Most Successful
  - ES-L1
  - GEO
  - HEO
  - SSO
- Least Successful
  - GTO
  - ISS
  - PO
  - SO

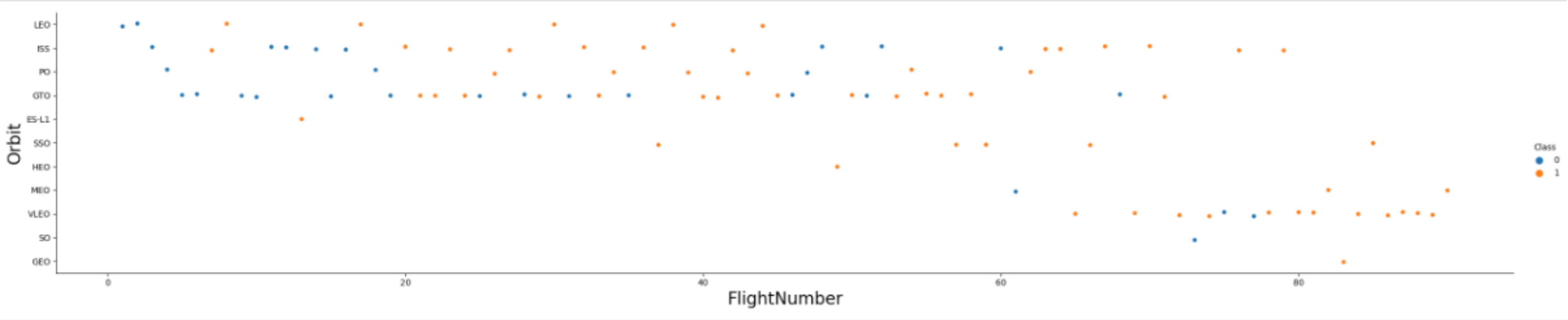


Flight Number vs. Orbit Type

TASK 4: Visualize the relationship between FlightNumber and Orbit type

For each orbit, we want to see if there is any relationship between FlightNumber and Orbit type.

```
[7]: # Plot a scatter point chart with x axis to be FlightNumber and y axis to be the Orbit, and hue to be the class value
sns.catplot(y="Orbit", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("FlightNumber",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```

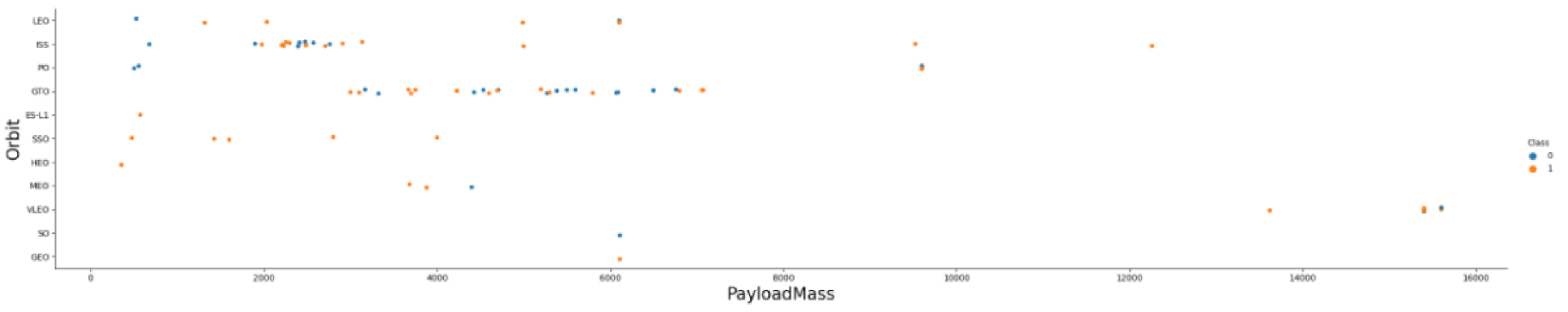


Payload vs. Orbit Type

TASK 5: Visualize the relationship between Payload and Orbit type

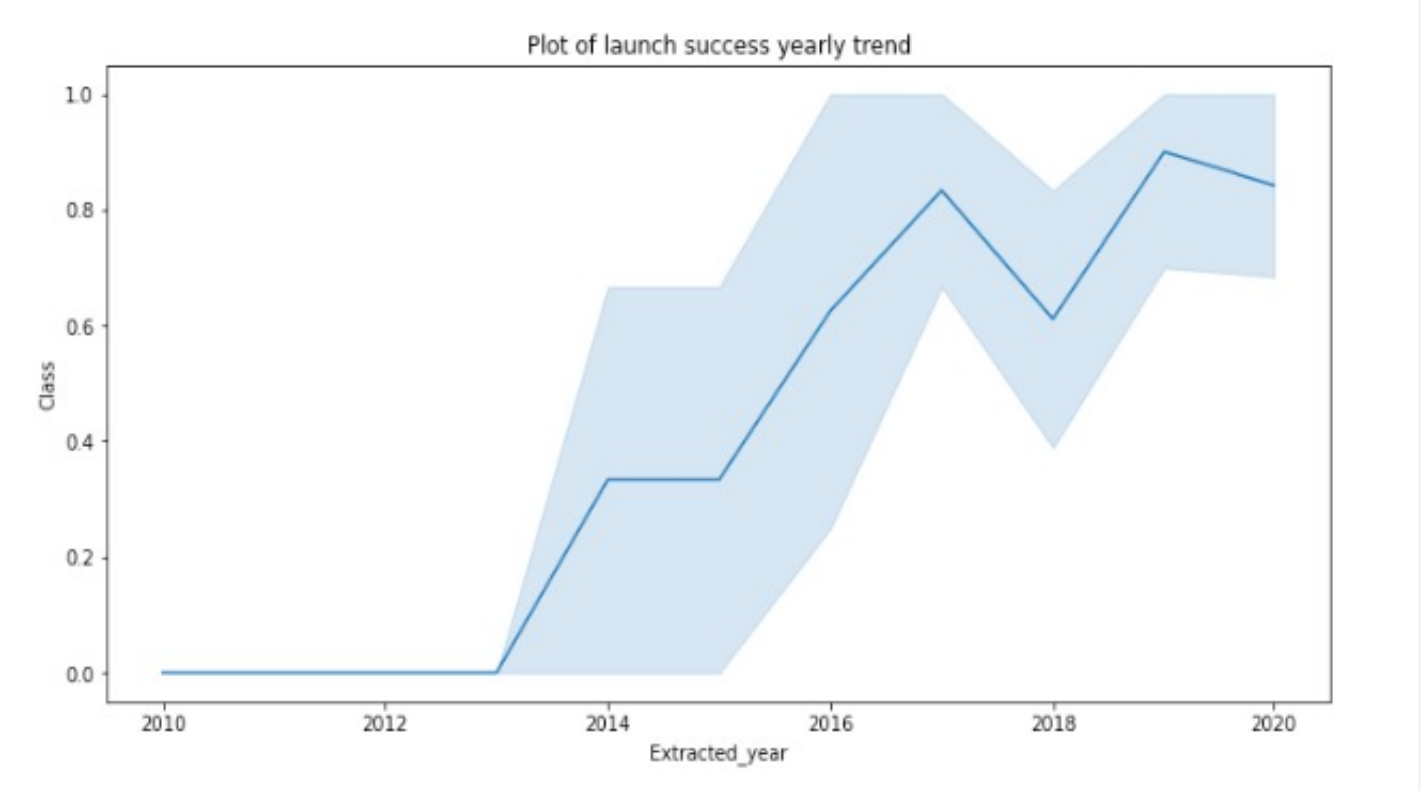
Similarly, we can plot the Payload vs. Orbit scatter point charts to reveal the relationship between Payload and Orbit type

[8]:
```
# Plot a scatter point chart with x axis to be Payload and y axis to be the Orbit, and hue to be the class value
sns.catplot(y="Orbit", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("PayloadMass",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```



Launch Success Yearly Trend

- Overall success of the Falcon 9 launces steadily increases from 2013 to 2019



All Launch Site Names

- There are 4 distinct launch sites in total

Task 1

Display the names of the unique launch sites in the space mission

```
[6]: sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
 * sqlite:///my_data1.db
Done.
```

[6]:

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

Launch Site Names Begin with 'CCA'

Display 5 records where launch sites begin with the string 'CCA'

[7]:
```
sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

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

[7]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- The total payload carried by boosters from NASA was 111,268 kilograms

Display the total payload mass carried by boosters launched by NASA (CRS)

```
[8]: sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE PAYLOAD LIKE '%CRS%';
 * sqlite:///my_data1.db
Done.
[8]: TOTAL_PAYLOAD
     111268
```

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2928.4 kilograms

Display average payload mass carried by booster version F9 v1.1

```
[9]: sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
```

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

[9]:

AVG_PAYLOAD
2928.4

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad was December 22, 2015

In [13]:
```
sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Success (ground pad)';
```

```
 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.clogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[13]:

first_success_gp
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
In [14]: sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND LANDING__OUTCOME = 'Success (drone
 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.clogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[14]:

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

Total Number of Successful and Failure Mission Outcomes

- The total number of successful missions:
  - 98
- The total number of failure mission:
  - 1

List the total number of successful and failure mission outcomes

[12]:
```
sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;
```

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

[12]:

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

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

[13]:
```
sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL) ORDER BY BOOSTER_VERSION;
```

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

[13]:

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

In [24]:
```
sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND DATE_PART('YEAR', DATE) = 201
```

```
 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.clogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[24]:

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

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

In [25]:
```
sql SELECT LANDING__OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING__OUTCOME (
```

```
 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.clogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

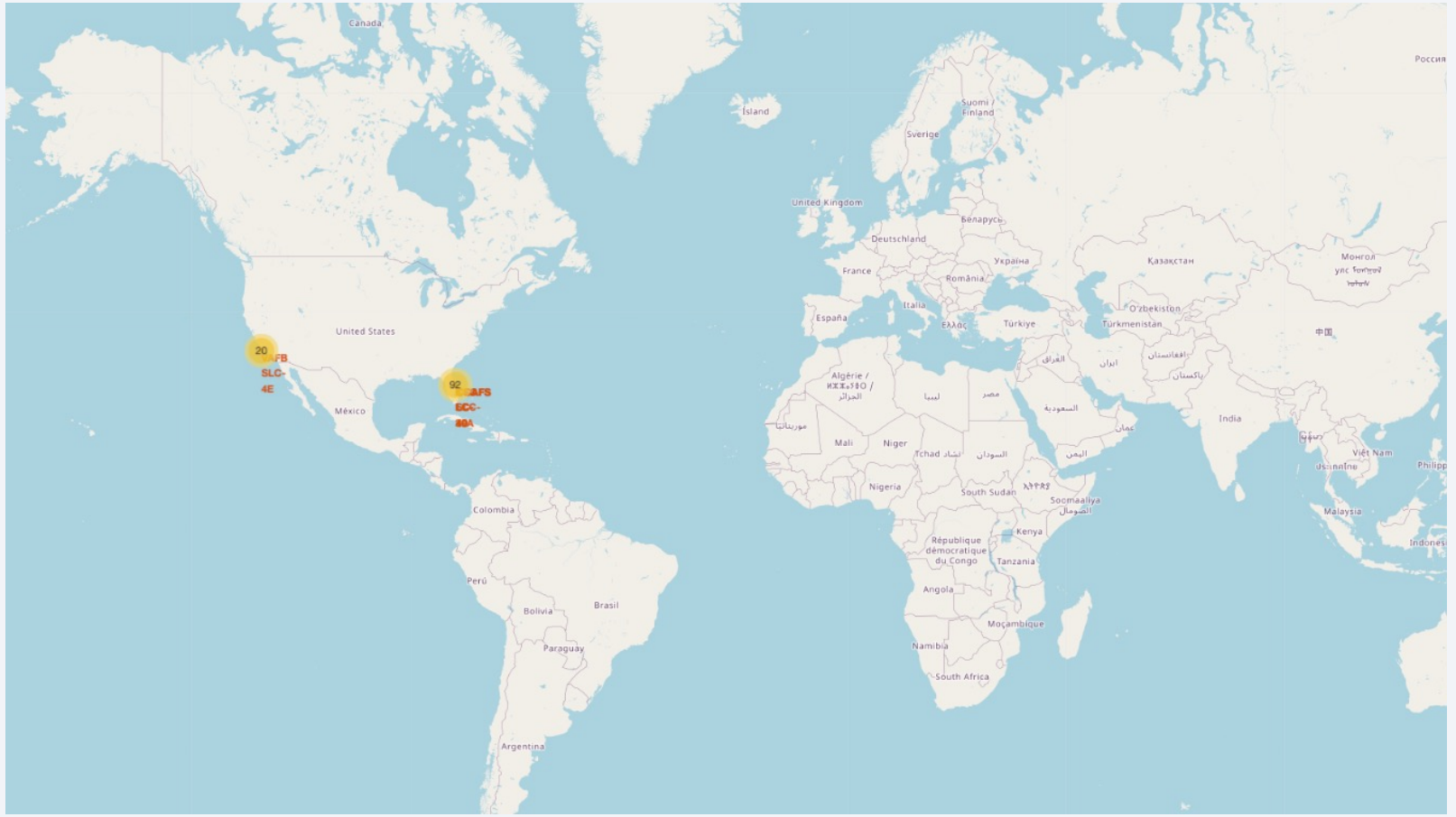
Out[25]:

landing__outcome	qty
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 3
Launch Sites
Proximities Analysis

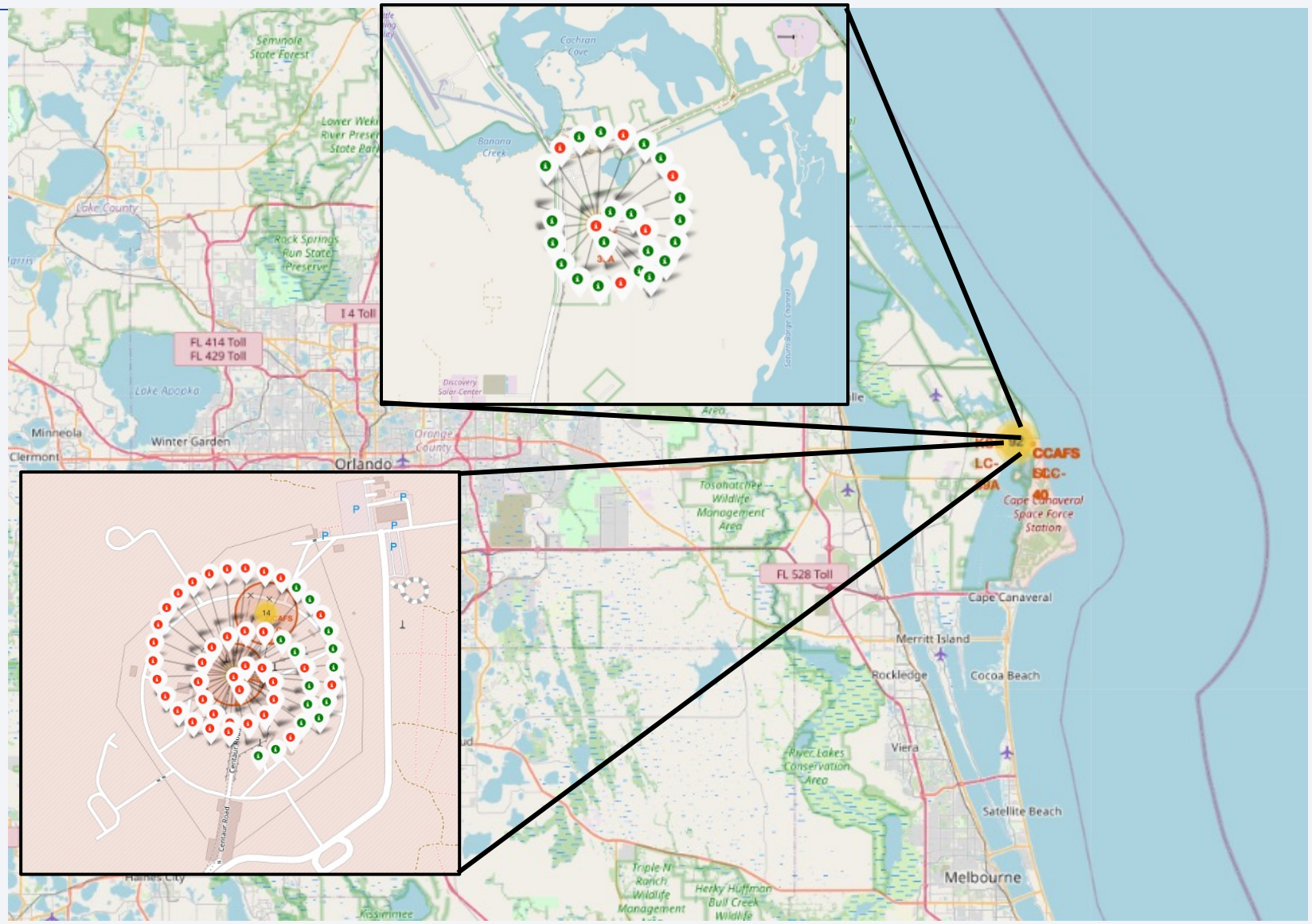
All SpaceX Launch Sites

- All launch sites used by SpaceX are on the East and West Coasts of the United States.
  - This is reasonable since SpaceX is an America-based company



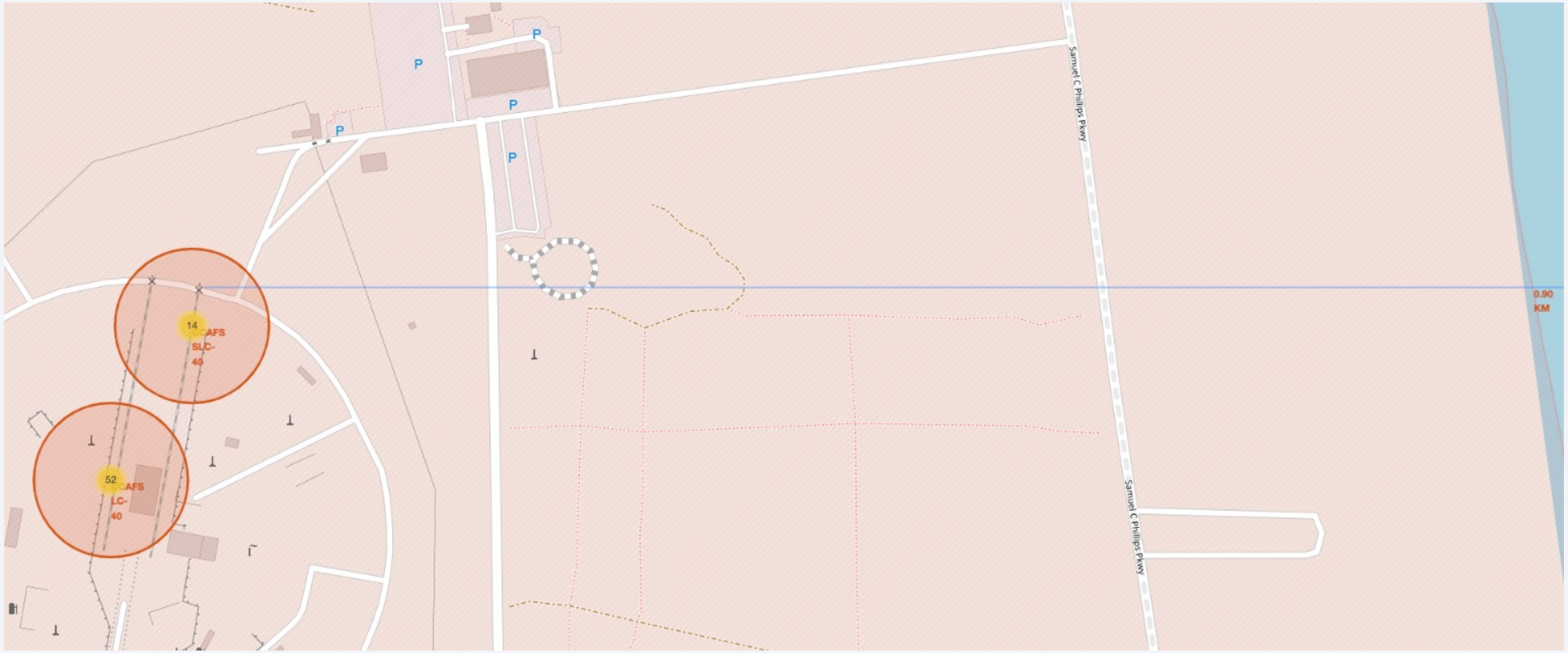
Pinpoint Launch Sites Located in Florida

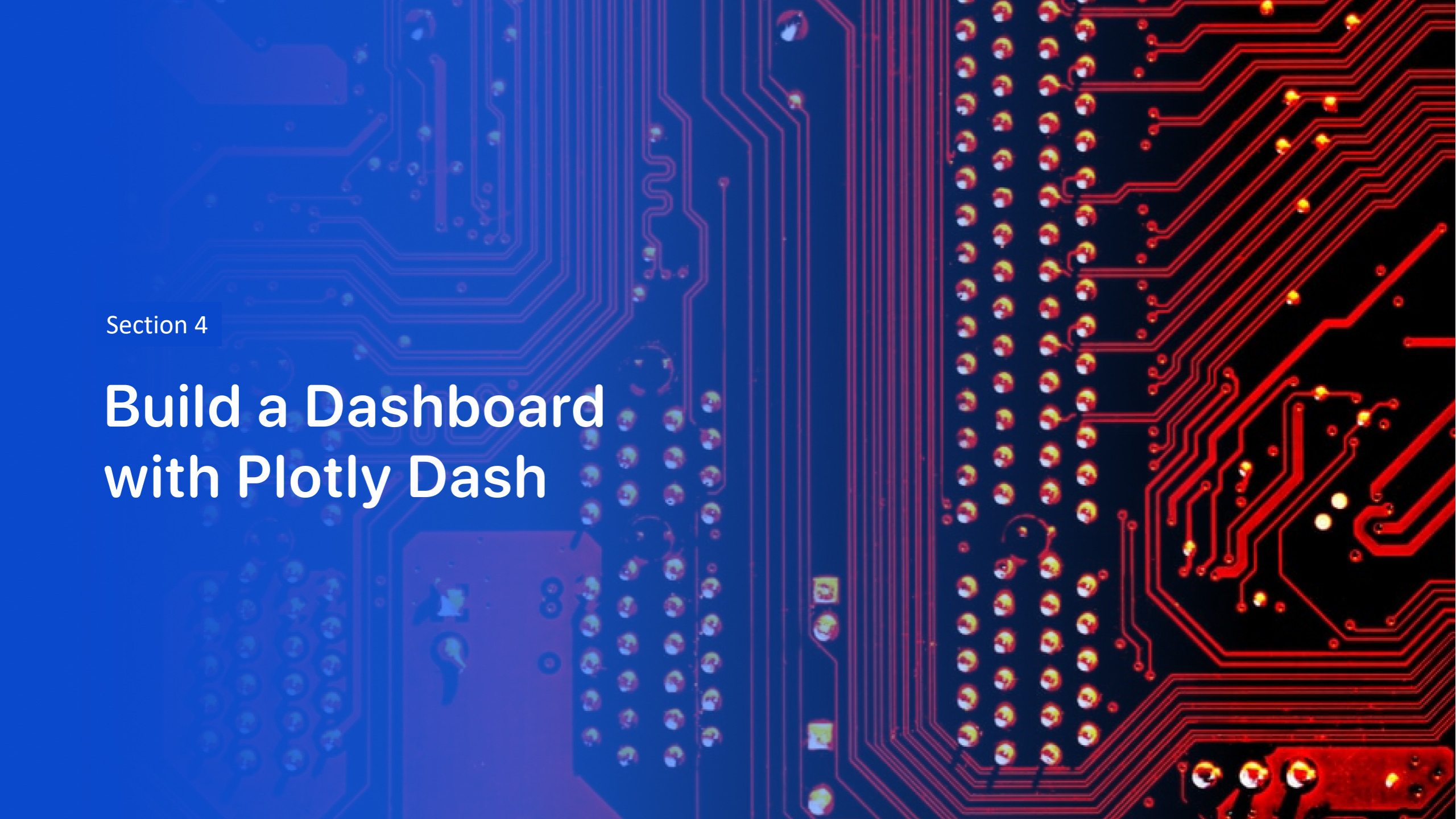
- The Folium Interactive mapping allows color coding of each failed and successful launch at each site



Coastline Distancing

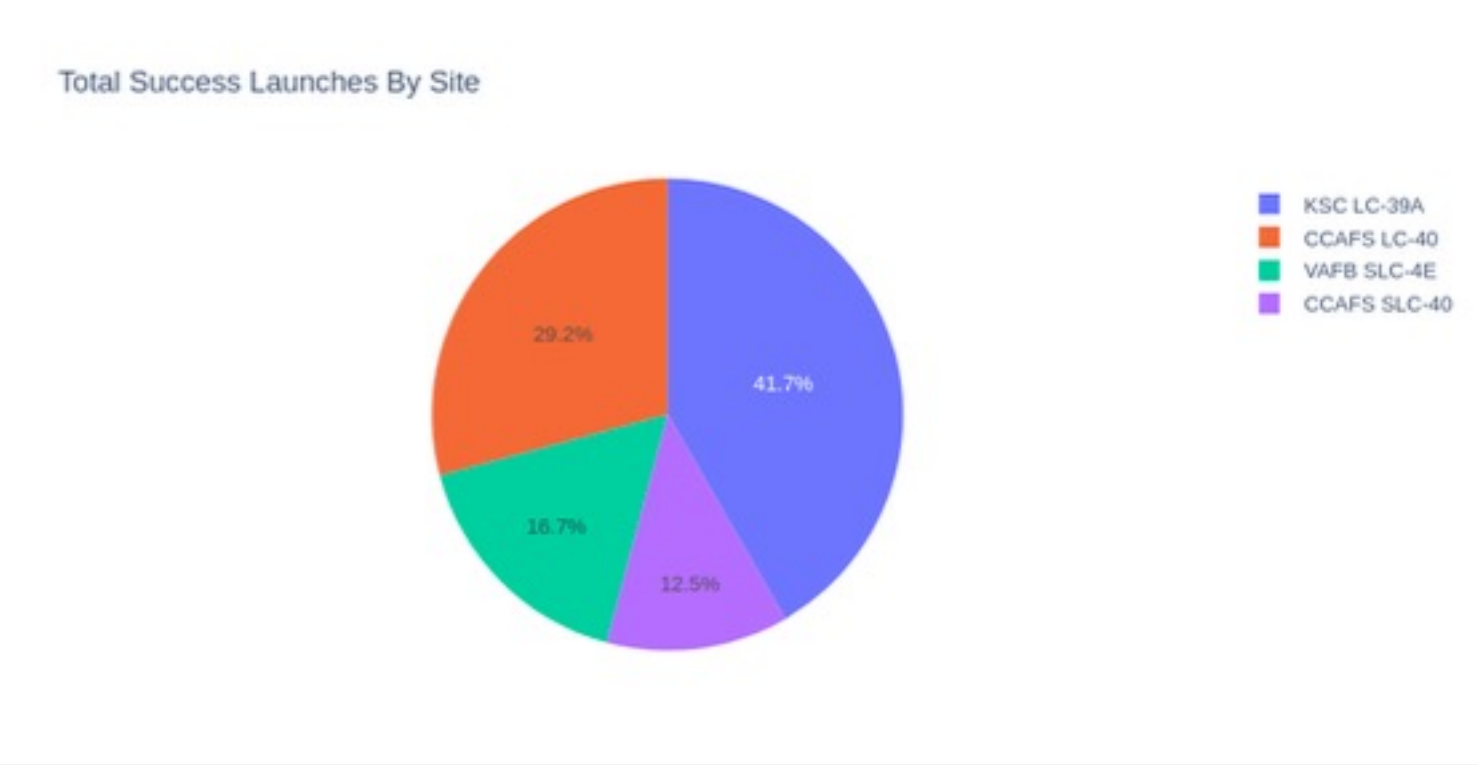
- CCAFS-SLC-40 is 90 km from the Florida shore putting it at minimal risk of failed rockets impacting local highways and railroads




Section 4
Build a Dashboard with Plotly Dash

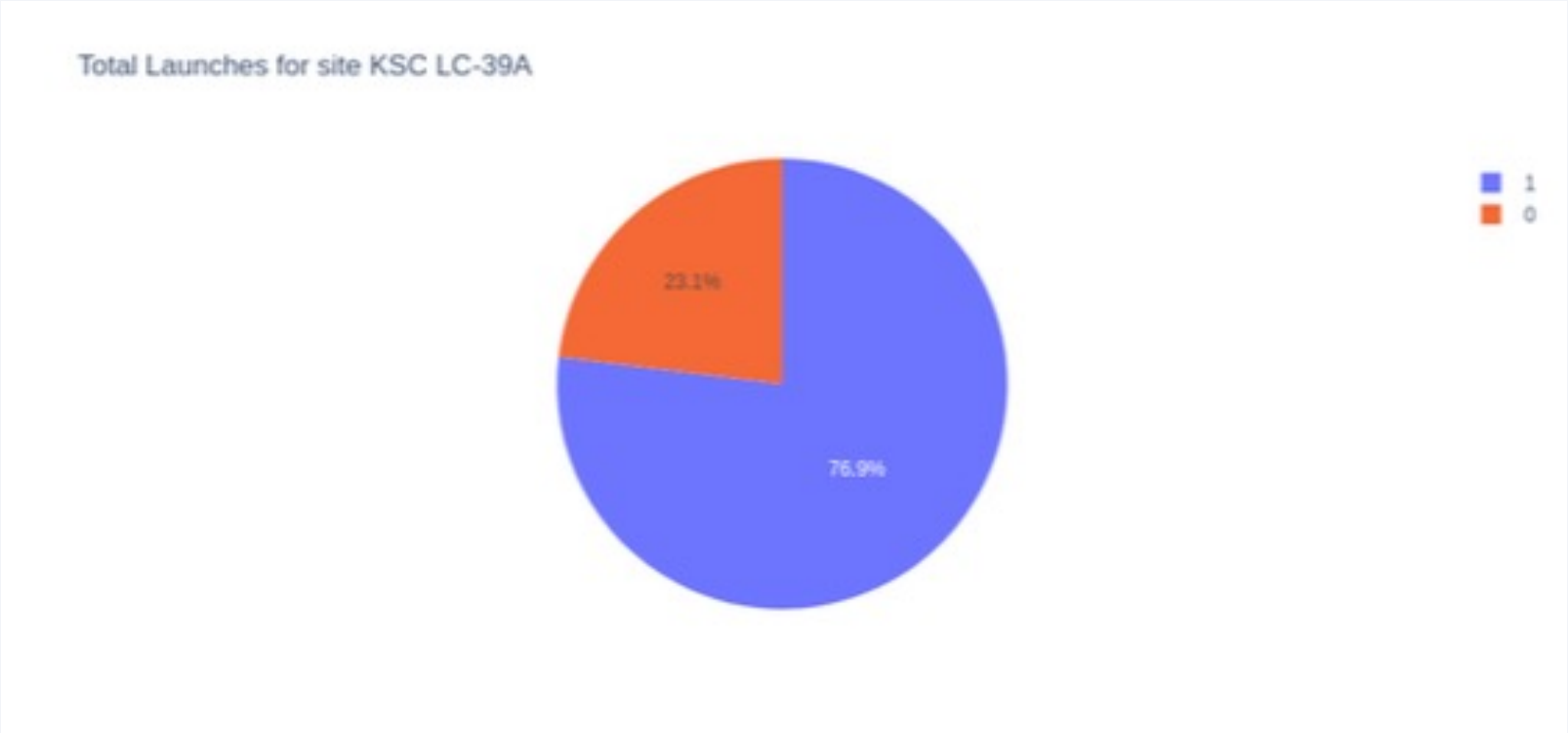
Successful Launches by Site

- As shown KSC LC-39A followed by CCAFS LC-40 have the highest success, proving that location is key when launching new Falcon 9's.



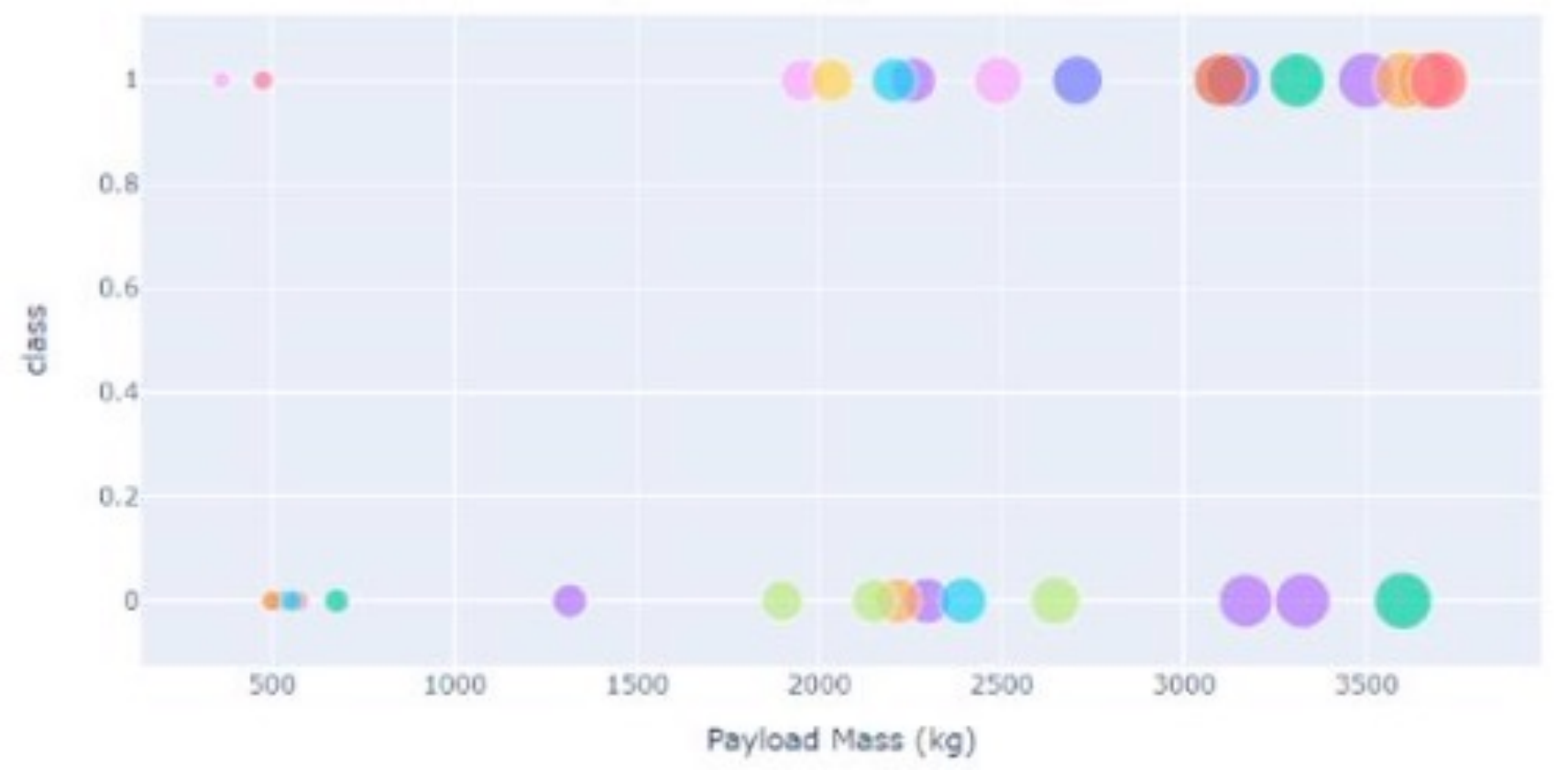
Further Look at KSC LC-39A's Overall Success

- As shown in our binary success/failure modes, KSC LC-39A has an overall success rate of 76.9%



Distribution of Payload to Launch Outcome Ratios

- As shown, the success of the upper weights of payloads has a higher success rate than lower



Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Using a comparison between all models the best model to use for this test would be the decision tree classifier

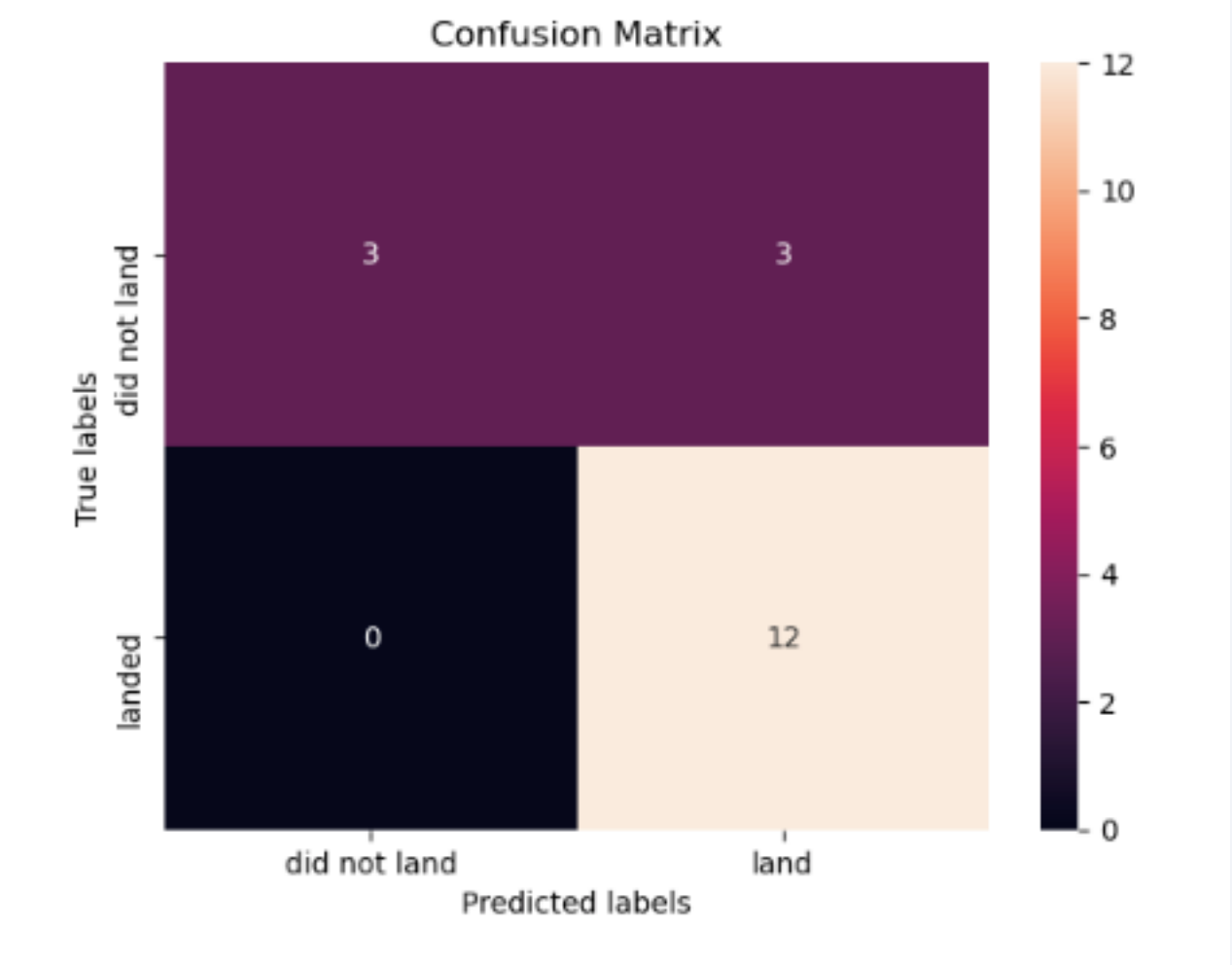
```
models = {'KNeighbors':knn_cv.best_score_,
              'DecisionTree':tree_cv.best_score_,
              'LogisticRegression':logreg_cv.best_score_,
              'SupportVector': svm_cv.best_score_}

bestalgorithm = max(models, key=models.get)
print('Best model is', bestalgorithm,'with a score of', models[bestalgorithm])
if bestalgorithm == 'DecisionTree':
    print('Best params is :', tree_cv.best_params_)
if bestalgorithm == 'KNeighbors':
    print('Best params is :', knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best params is :', logreg_cv.best_params_)
if bestalgorithm == 'SupportVector':
    print('Best params is :', svm_cv.best_params_)
```

```
Best model is DecisionTree with a score of 0.8732142857142856
Best params is : {'criterion': 'gini', 'max_depth': 6, 'max_features': 'auto', 'min_samples_leaf': 2, 'min_samples_split': 5, 'splitter': 'random'}
```

Confusion Matrix

- Shown is the Decision Tree Classifier comparing which classes of True statements correlate to the Predicted Labels



Conclusions

- Overall, the more launches completed at a given site, proved to show higher yield of success in the end
- From 2013 through 2019 there was a steady increase of launch suceess
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- KSC LC-39A had the highest yield of successful launches when compared to all other sites.
- The Decision Tree Classifier is the most useful machine learning algorithm.

Thank you!