IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Grayson P. Ostermeyer, MSc
January 4, 2022

Outline

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

Executive Summary

- The goal of this project is to analyze the determinants of successful launch outcomes of SpaceX Falcon 9 rockets so that we may predict whether the first stage will land successfully so that we at SpaceY can make our enterprise cost-effective
- The data science tools used in this project were data collection using REST API calls and web scraping using the BeautifulSoup module, data wrangling and data labeling, EDA using visualization and SQL, interactive visual analytics using Folium and Plotly Dash, and predictive analysis using classification models with model evaluation
- Key findings
  - Light payload missions have higher launch outcome success
  - The FT booster version has the highest launch success rate, while v1.1 booster has the lowest success rate
  - Flights associated with the ES-L1, GEO, HEO, and SSO orbits have high launch success rates
  - The KSC LC-39A launch site has a high launch outcome success rate, especially for light payload missions
  - Launch sites are located far from highways and especially cities, but located near coastlines
- While all classifiers were robust at predicting the test data after hyperparameters were tuned, the Decision Tree classifier performed slightly better than SVM, KNN, and Logistic Regression

Introduction

- SpaceY seeks to be a leader in the commercial space age
- Enterprise and innovation at our competitor, SpaceX, has guided their business-industrial model to make rocket launches as efficient and cost-effective as possible
  - Falcon 9 rocket launches cost $62 million each, whereas other competitors' launches cost as much as $165 million
  - SpaceX's first stage is reusable, which leads to massive savings
- If we can determine whether the first stage will land, we can determine the cost of each launch
- Data Science tools that we will use in our analysis:
  - Exploratory data analysis (EDA) using visualization and SQL
  - Interactive visual analytics using Folium and Plotly Dash
  - Predictive analysis using classification models

Section 1

Methodology

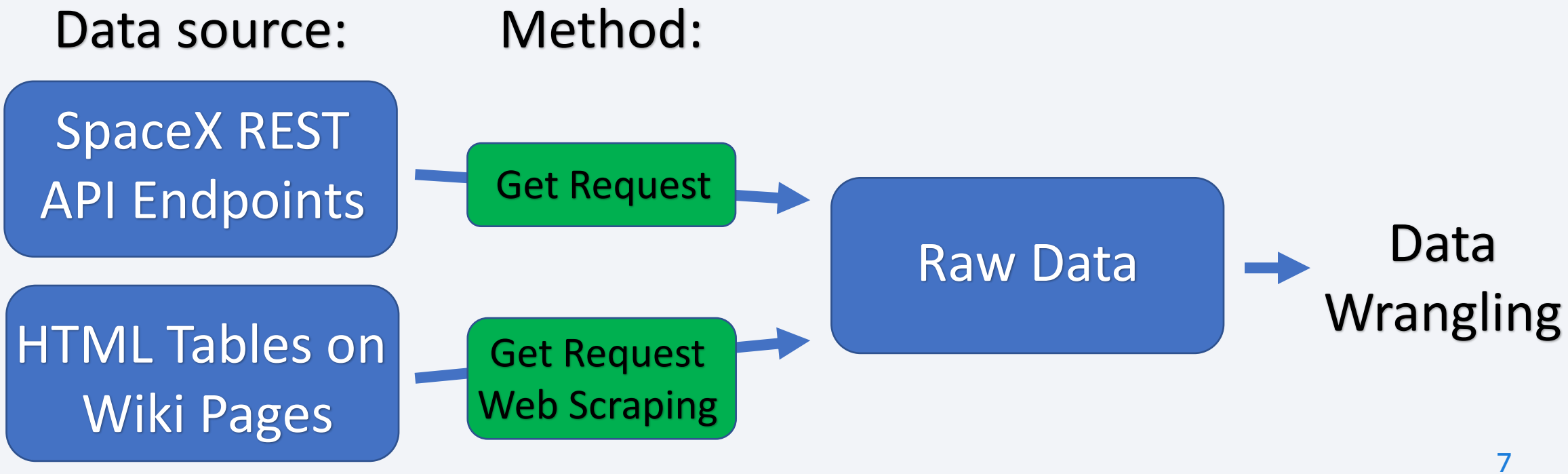
Methodology

Executive Summary

- Data collection methodology:
  - Data collection by Get Requests using REST API endpoints and web scraping
- Perform data wrangling
  - Data was filtered, null values replaced with column means, and training labels were generated
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Machine learning models generated: KNN, Decision Tree, SVM, Logistic Regression
  - GridSearchCV objects created for each classifier to find optimal hyperparameter values
  - Model evaluation by confusion matrices, classification accuracies, Jaccard indices, F1 scores

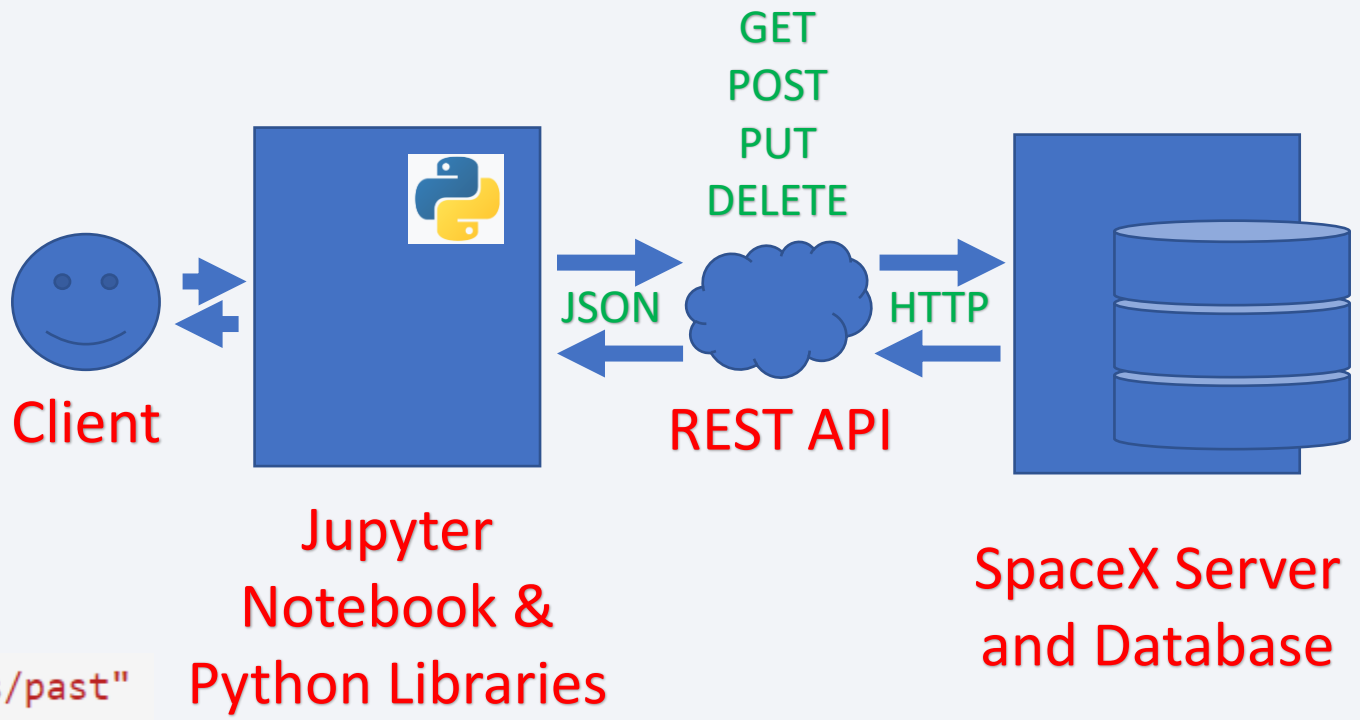
Data Collection

- Raw datasets were collected from REST API endpoints and HTML Tables on Wiki Pages via Get Requests



Data Collection – SpaceX API

- Data collection with SpaceX REST API calls
- https://github.com/GPOstermeyer/Coursera-SpaceX-Capstone-Project-Assets/blob/master/jupyter-labs-spacex-data-collection-api%20(1).ipynb
- Code cells below demonstrate how to retrieve and view data using a RESTful API



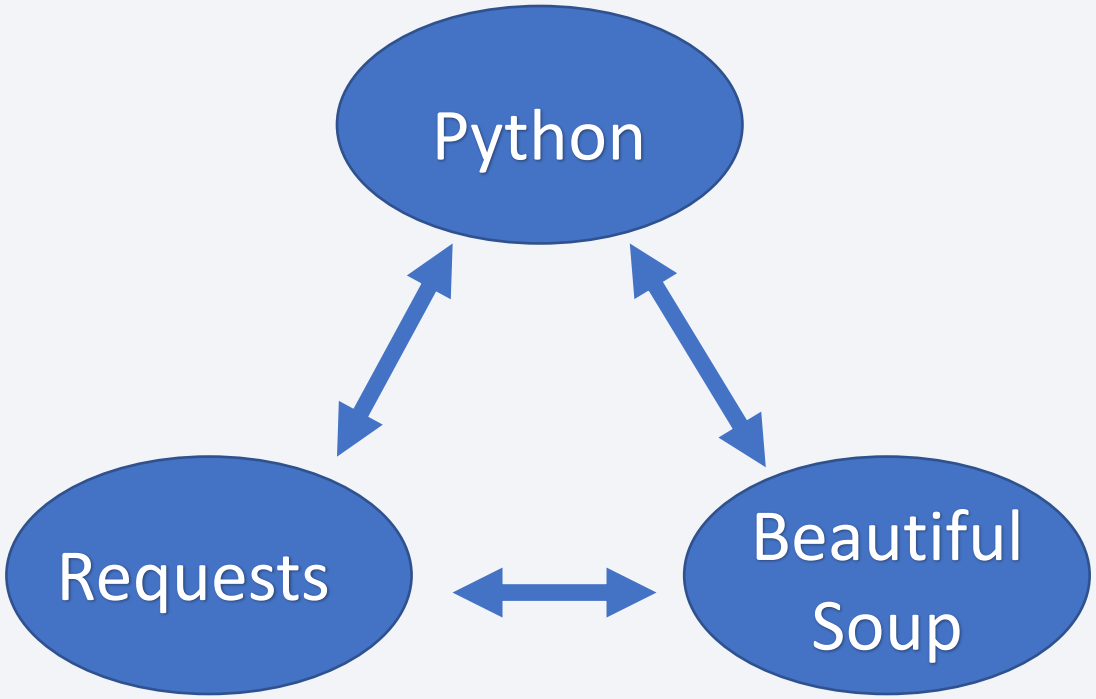

```
spacex_url="https://api.spacexdata.com/v4/launches/past"

response = requests.get(spacex_url)

print(response.content)
```

Data Collection - Scraping

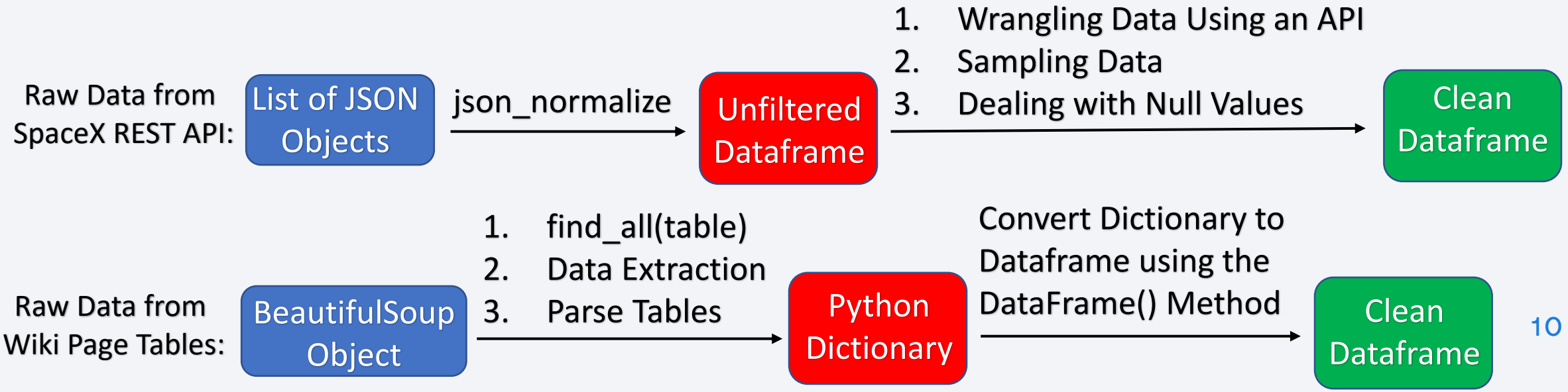
- Web scraping process using the BeautifulSoup module
- https://github.com/GPOstermeyer/Coursera-SpaceX-Capstone-Project-Assets/blob/master/jupyter-labs-webscraping%20(1).ipynb
- Code cells below demonstrate how to use the HTTP GET method from a webpage as an HTTP response, then create a BeautifulSoup object from which data may be extracted and parsed




```
static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"
response = requests.get(static_url).text
soup = BeautifulSoup(response,'html5lib')
```

Data Wrangling

- Data Wrangling of JSON Objects and Beautiful Soup Objects
- Processed dataframes and training labels generated
- https://github.com/GPOstermeyer/Coursera-SpaceX-Capstone-Project-Assets/blob/master/labs-jupyter-spacex-Data%20wrangling%20(1).ipynb



EDA with Data Visualization

- Scatter point charts to visualize the relationship between…
  - Flight Number and Launch Site
  - Payload and Launch Site
  - Flight Number and Orbit Type targeted
  - Payload and Orbit Type targeted
- Bar chart to visualize the relationship between…
  - Launch Success Rate and Orbit Type targeted
- Line chart to visualize…
  - Yearly Trend of Launch Success
- https://github.com/GPOstermeyer/Coursera-SpaceX-Capstone-Project-Assets/blob/master/jupyter-labs-eda-dataviz%20(1).ipynb

EDA with SQL

- SQL queries performed on Falcon 9 launch data are presented in the following slides
- https://github.com/GPOstermeyer/Coursera-SpaceX-Capstone-Project-Assets/blob/master/jupyter-labs-eda-sql-coursera%20(2).ipynb

EDA with SQL

Task 1

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

- %%sql
- SELECT UNIQUE(LAUNCH_SITE)
- FROM SPACEXTBL
- ;

EDA with SQL

Task 2

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

- %%sql
- SELECT LAUNCH_SITE
- FROM SPACEXTBL
- WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
- ;

EDA with SQL

Task 3

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

- %%sql
- SELECT SUM(PAYLOAD_MASS__KG_)
- FROM SPACEXTBL
- WHERE CUSTOMER = 'NASA (CRS)'
- ;

EDA with SQL

Task 4

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

- %%sql
- SELECT AVG(PAYLOAD_MASS__KG_)
- FROM SPACEXTBL
- WHERE BOOSTER_VERSION = 'F9 v1.1'
- ;

EDA with SQL

Task 5

List the date when the first successful landing outcome in ground pad was acheived.

- %%sql
- SELECT MIN(DATE)
- FROM SPACEXTBL
- WHERE LANDING__OUTCOME LIKE 'Success (ground pad)'
- ;

EDA with SQL

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

- %%sql
- SELECT BOOSTER_VERSION
- FROM SPACEXTBL
- WHERE LANDING__OUTCOME LIKE '%Success%' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
- ;

EDA with SQL

Task 7

List the total number of successful and failure mission outcomes

- %%sql
- SELECT
  COUNT(MISSION_OUTCOME)
- FROM SPACEXTBL
- ;

EDA with SQL

Task 8

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

- %%sql
- SELECT BOOSTER_VERSION,PAYLOAD_MASS__KG_
- FROM SPACEXTBL
- WHERE PAYLOAD_MASS__KG_ =
- (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
- ;

EDA with SQL

Task 9

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

- %%sql
- SELECT LANDING__OUTCOME,BOOSTER_VERSION,LAUNCH_SITE
- FROM SPACEXTBL
- WHERE LANDING__OUTCOME LIKE '%Failure%' AND DATE LIKE '2015%'
- ;

EDA with SQL

Task 10

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

- %%sql
- SELECT LANDING__OUTCOME
- FROM SPACEXTBL
- WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
- GROUP BY LANDING__OUTCOME
- ORDER BY COUNT(LANDING__OUTCOME) DESC
- ;

Build an Interactive Map with Folium

- Map objects added to folium map:
  - Blue circle at the NASA Johnson Space Center with a popup label showing its name
  - Marker at the NASA Johnson Space Center with an icon showing its name
  - Red circle at the CCAFS LC-40 Launch Site with a popup label showing its name
  - Marker at the CCAFS LC-40 Launch Site with an icon showing its name
  - Red circle at the CCAFS SLC-40 Launch Site with a popup label showing its name
  - Marker at the CCAFS SLC-40 Launch Site with an icon showing its name
  - Red circle at the KSC LC-39A Launch Site with a popup label showing its name
  - Marker at the KSC LC-39A Launch Site with an icon showing its name
  - Red circle at the VAFB SLC-4E Launch Site with a popup label showing its name
  - Marker at the VAFB SLC-4E Launch Site with an icon showing its name
  - Marker clusters at each launch site showing launch successes (green) and failures (red)
  - Line and distance marker from the CCAFS LC-40 Launch Site to a nearby coastline to demonstrate its distance
  - Line and distance marker from the CCAFS LC-40 Launch Site to a nearby railroad to demonstrate its distance
  - Line and distance marker from the CCAFS LC-40 Launch Site to a nearby highway to demonstrate its distance
  - Line and distance marker from the CCAFS LC-40 Launch Site to a nearby city to demonstrate its distance
- https://github.com/GPOstermeyer/Coursera-SpaceX-Capstone-Project-Assets/blob/master/lab_jupyter_launch_site_location%20(2).ipynb

Build a Dashboard with Plotly Dash

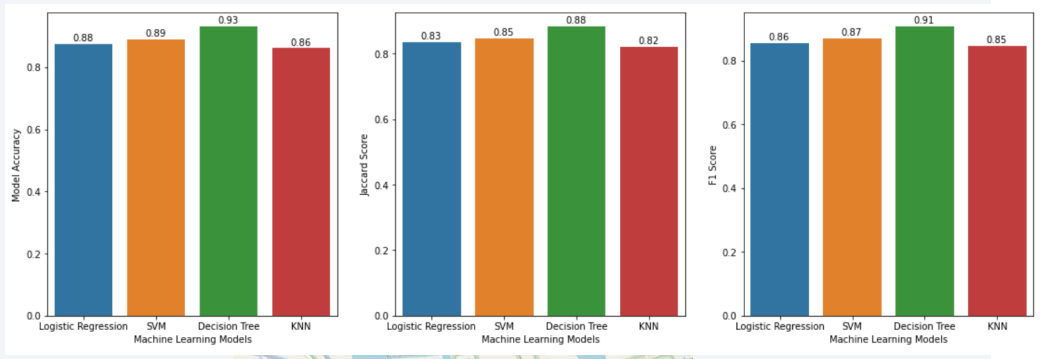
- Dashboard application for user to view SpaceX launch data in real time
- Dashboard features dropdown lists and range slider input components for user to interact with a pie chart and a scatter point chart
- Dropdown gives user the option to select a launch site
- Pie chart shows user the relative success of launches at the selected site
- Range slider allows user to limit selection of rockets carrying a given payload
- Scatter point chart allows user to visualize relative launch success at a selected site and payload
- https://github.com/GPOstermeyer/Coursera-SpaceX-Capstone-Project-Assets/blob/master/spacex_dash_app.py.ipynb

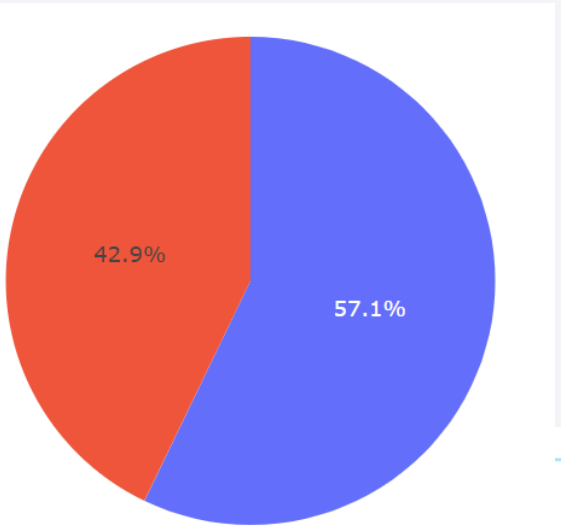
Predictive Analysis (Classification)

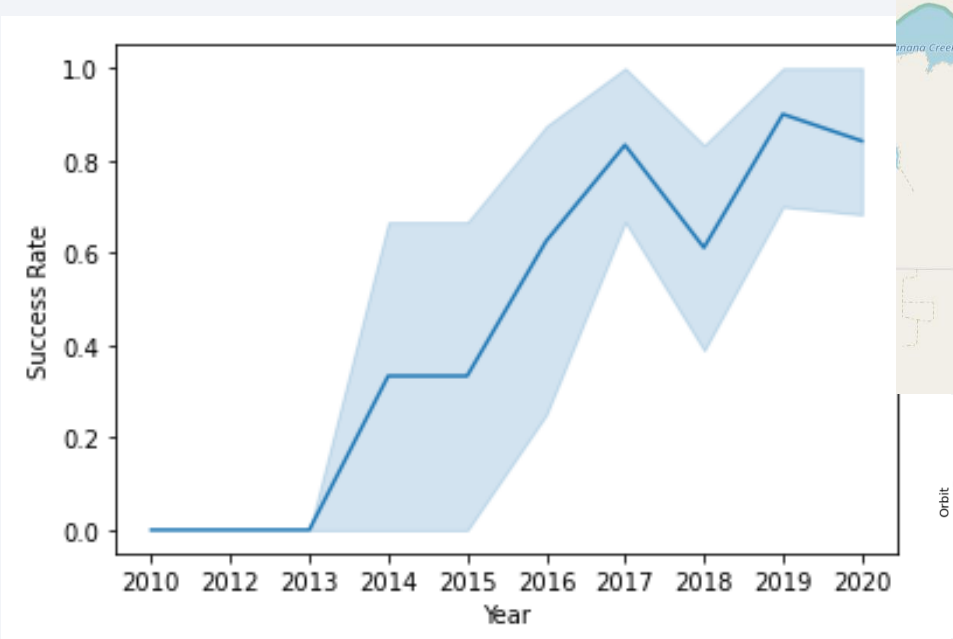
- The categorical variables used for classification were orbit type, launch site, landing pad, and serial
  - These features were one-hot encoded into labels of 0 (launch failure) and 1 (launch success)
  - The classifier should predict the launch outcome (label) of test data after it is fit on train data
- The train_test_split function was used to split standardized launch outcome labels into training and testing data
- Logistic Regression, SVM, Decision Tree, and KNN classification models were created and GridSearchCV objects were fit to them to find optimal hyperparameter values
- After training, validating, and testing the classifiers, confusion matrices, accuracy scores, Jaccard index scores, and F1-scores were calculated for each model to determine the best performing classification model
- https://github.com/GPOstermeyer/Coursera-SpaceX-Capstone-Project-Assets/blob/master/SpaceX_Machine%20Learning%20Prediction_Part_5%20(2).ipynb

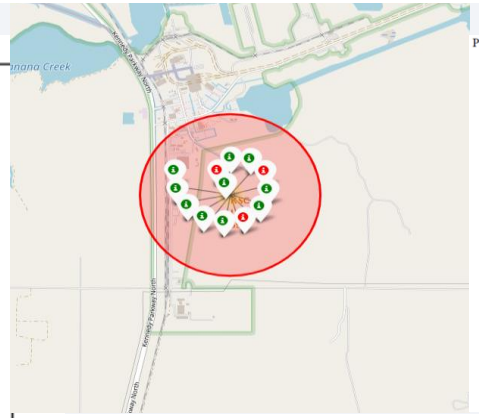
Results

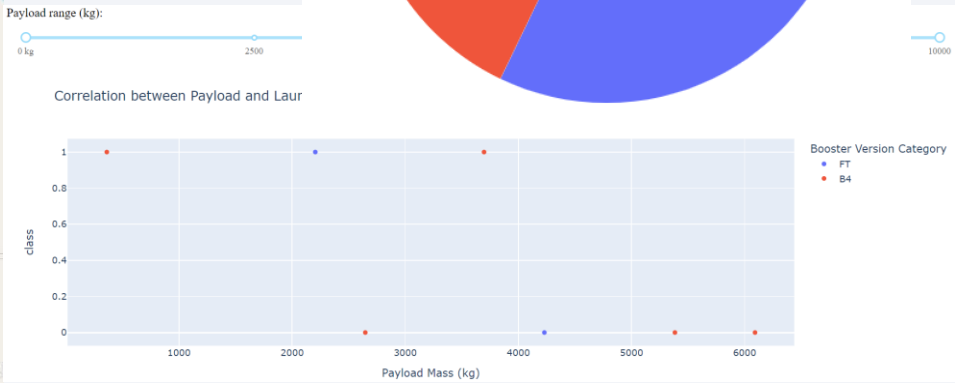
- Exploratory data analysis
- Interactive analytics
- Predictive analysis

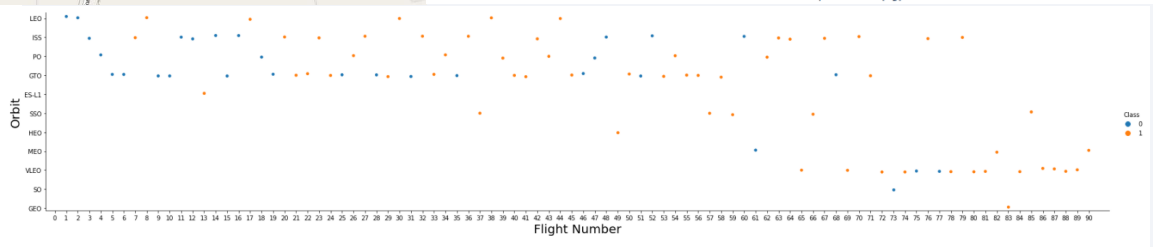






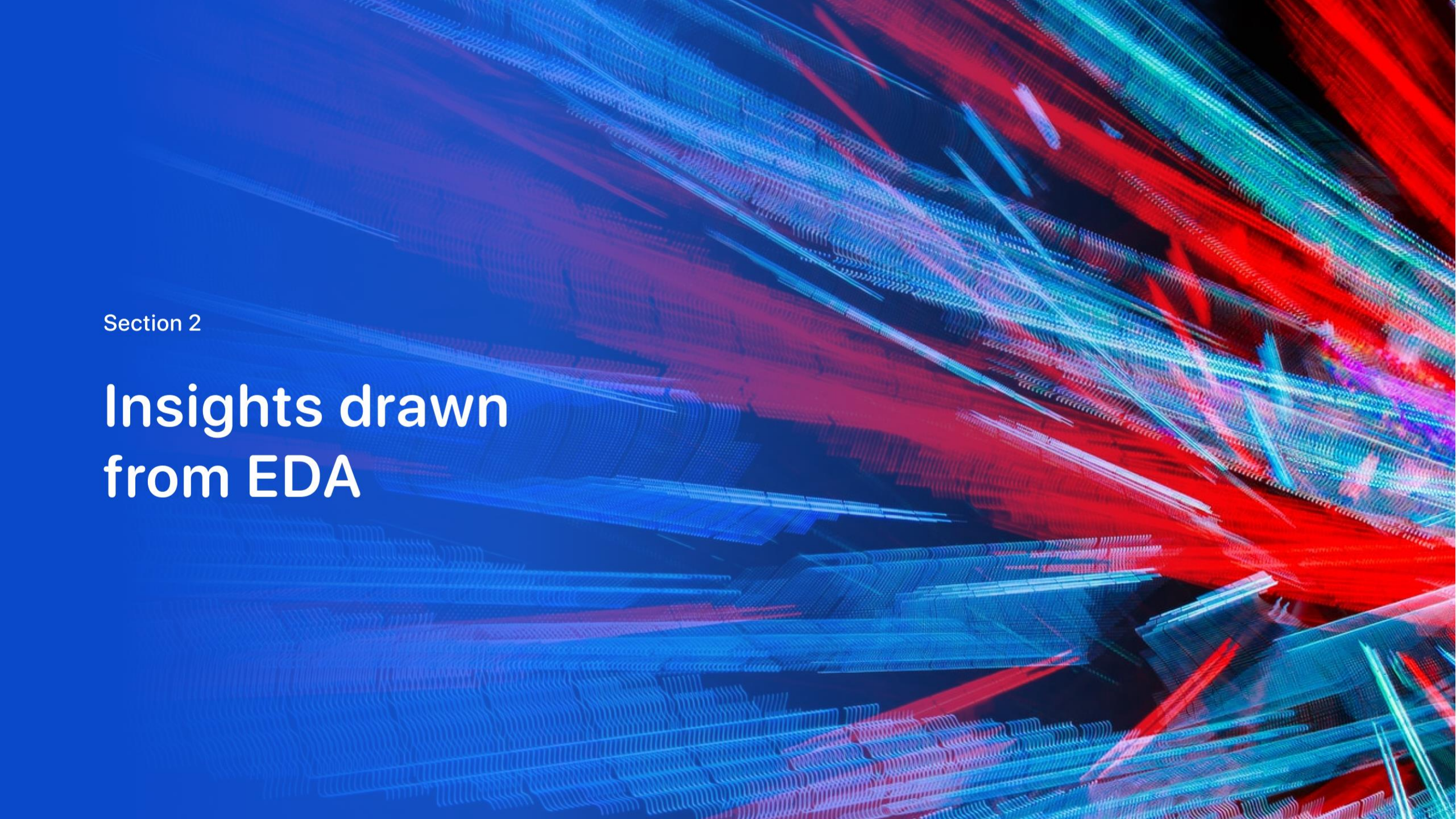






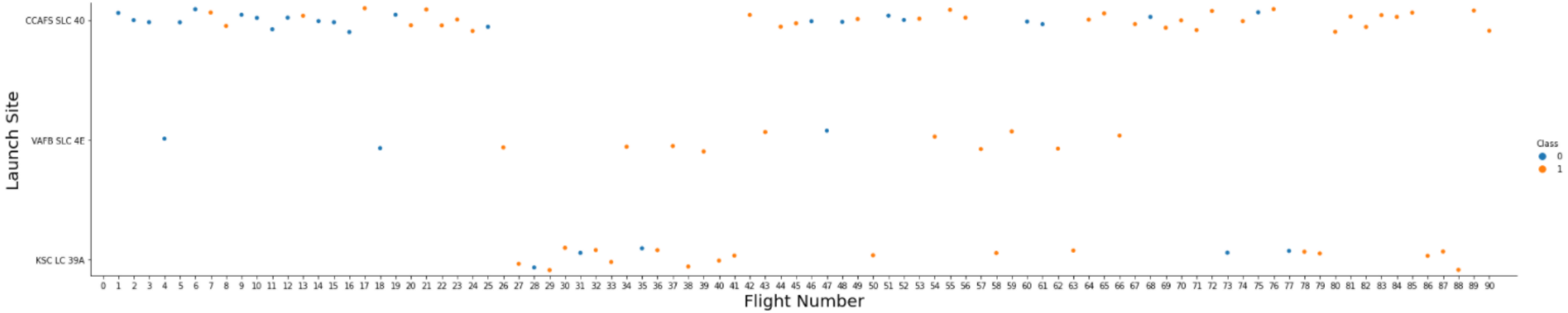
Section 2

Insights drawn from EDA



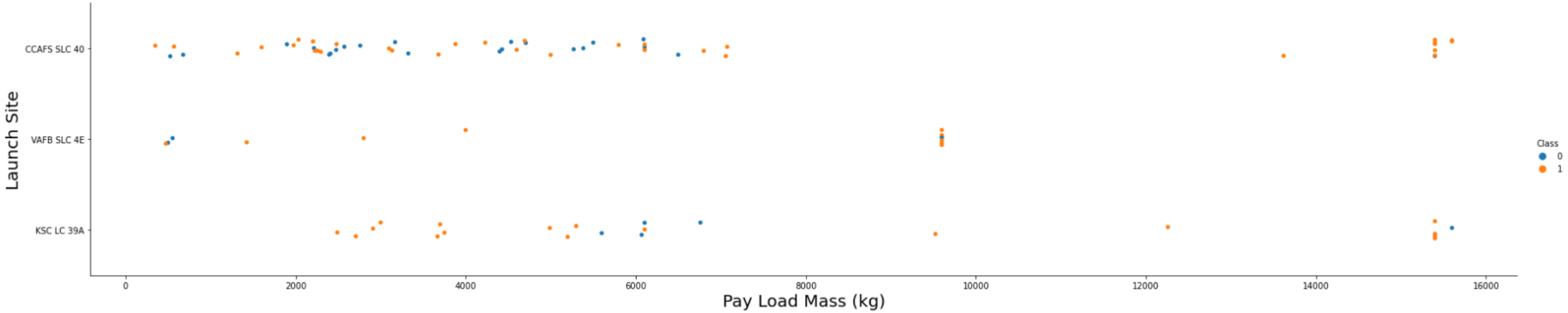
Flight Number vs. Launch Site

- Flight Number is an index of the launch order, where “0” occurred on 4 June 2010 and “91” occurred on 8 August 2020
- Class refers to the launch outcome, where 0 (blue) points indicate a failed outcome, and 1 (orange) points indicate a successful outcome
- CCAFS SLC 40 had the highest frequency of launches, with improvement in launch success over time
- A hiatus in launches at CCAFS SLC 40 coincided a high frequency of new, mostly successful launches at KSC LC 39A between Flight Numbers 27 and 41



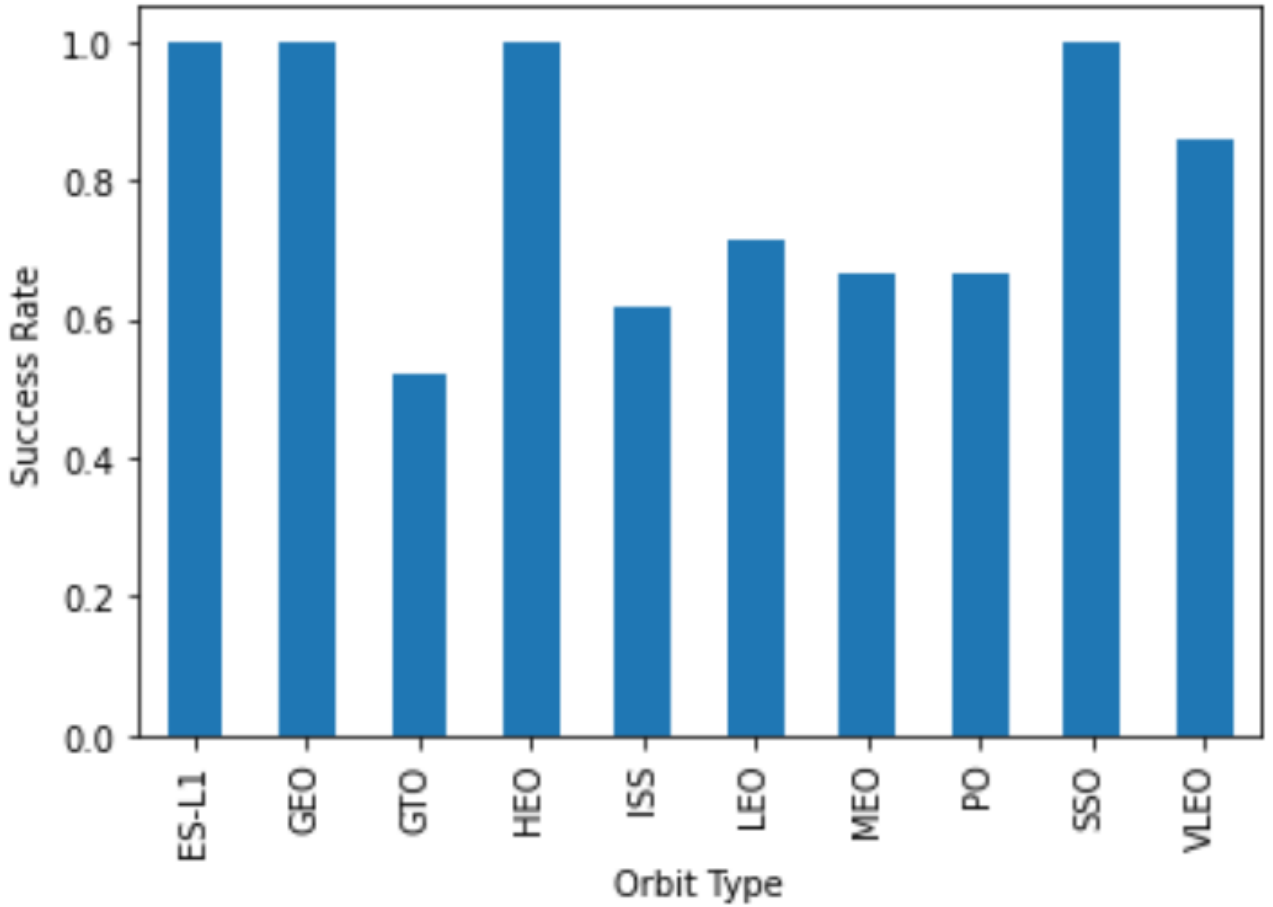
Payload vs. Launch Site

- Class refers to the launch outcome, where 0 (blue) points indicate a failed outcome, and 1 (orange) points indicate a successful outcome
- CCAFS SLC 40 had the highest frequency of launches, hosting mostly light payloads (≤ 7000 kg), but also very heavy payloads (≥ 13000 kg) which were mostly successful launches
- There were no rockets launched for heavy payload masses (≥ 10000 kg) at the VAFB-SLC launch site
- Very heavy payload launches at the KSC LC 39A launch site were mostly successful
- There is not a clear relationship between launch outcome and payload mass for any of the launch sites based on this plot, however it is notable that all launches at the KSC LC 39A launch site with light payload masses (≤ 5000 kg) were all successful



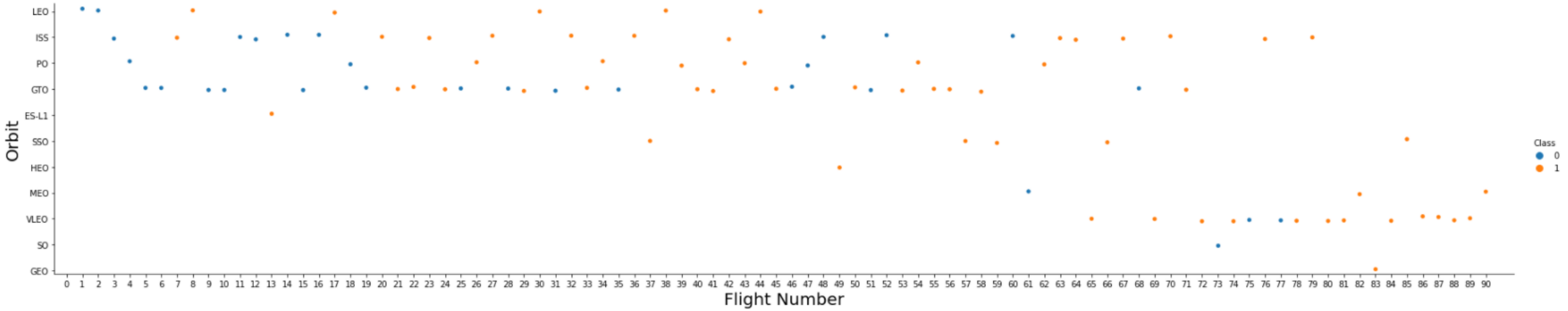
Success Rate vs. Orbit Type

- Flights associated with the ES-L1, GEO, HEO, and SSO orbits all had successful launch outcomes
- Only half of the launches to the GTO orbit were successful
- Sub-orbital launches are excluded



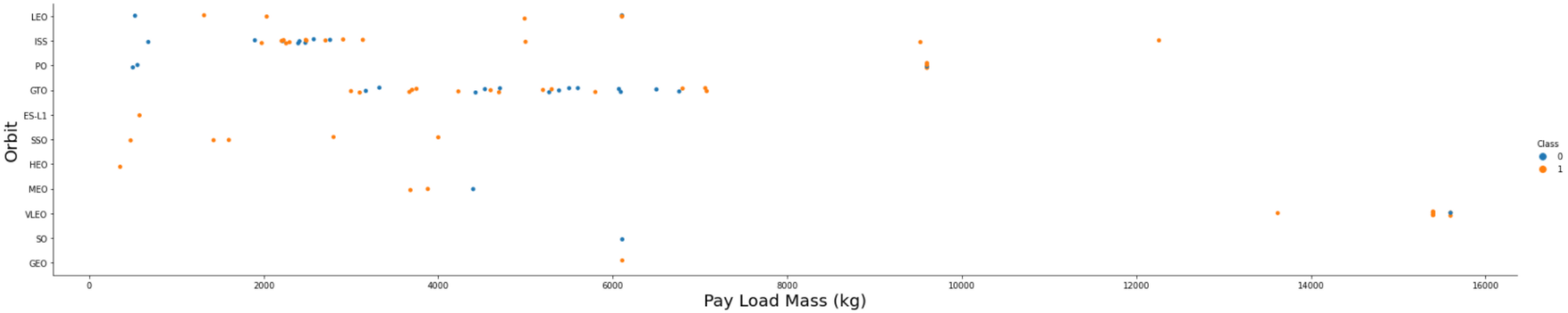
Flight Number vs. Orbit Type

- Flight Number is an index of the launch order, where “0” occurred on 4 June 2010 and “91” occurred on 8 August 2020
- Class refers to the launch outcome, where 0 (blue) points indicate a failed outcome, and 1 (orange) points indicate a successful outcome
- The LEO, ISS, PO, and GTO orbits have been targeted since early rocket launches
- The VLEO orbit has been the most frequent orbital target since the 65th flight, which occurred on 5 December 2018
- A successful launch outcome appears to be related to flight number for the LEO orbit, but this relationship does not appear to be true for GTO orbit launches



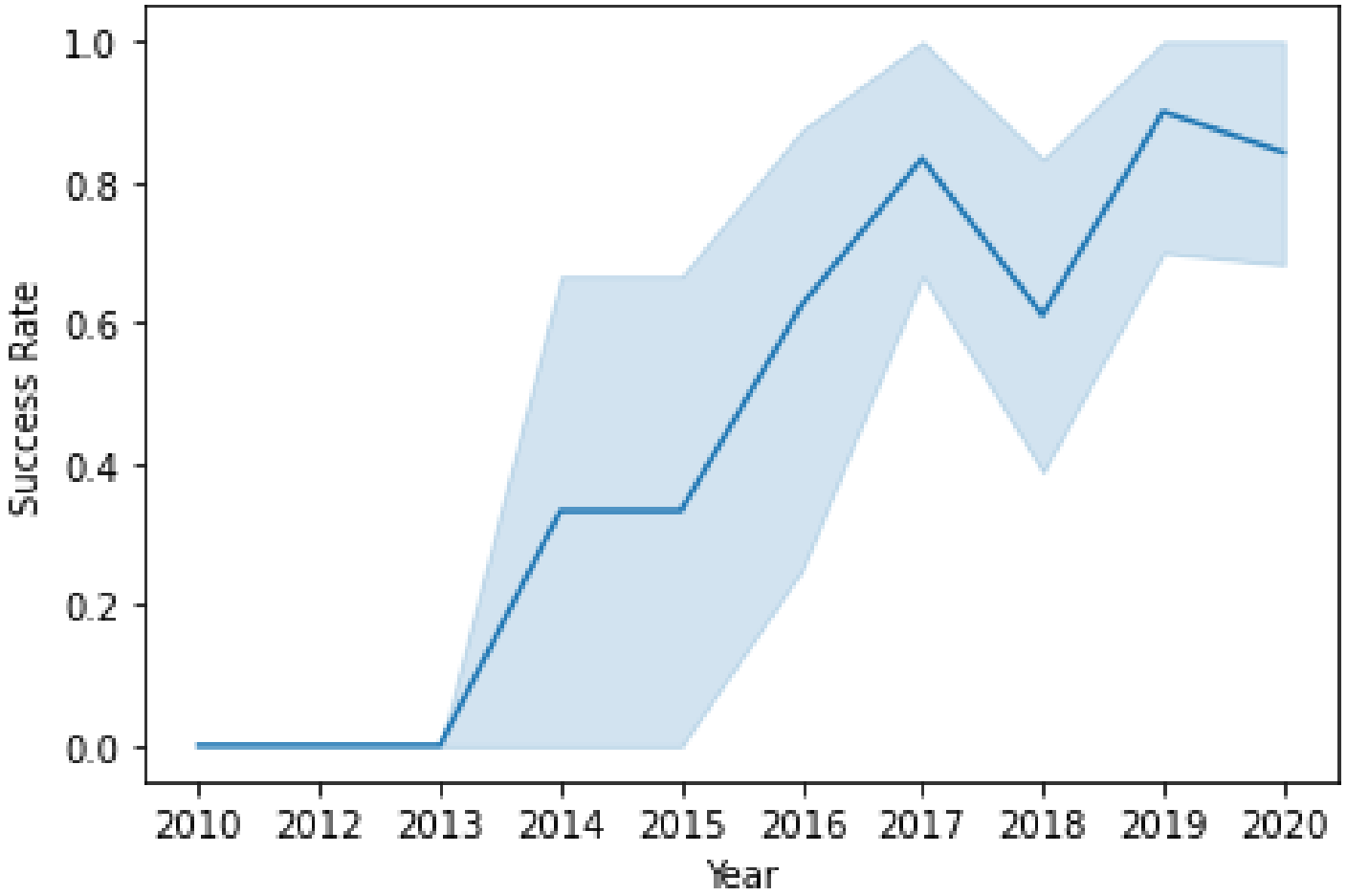
Payload vs. Orbit Type

- Class refers to the launch outcome, where 0 (blue) points indicate a failed outcome, and 1 (orange) points indicate a successful outcome
- Flights to most orbits are associated with payloads of < 7000 kg
- Flights to the ISS and PO orbits were capable of carrying payloads well above the average payload
- The heaviest payloads of > 13000 kg targeted the VLEO orbit
- A relationship between launch success and payload mass cannot be easily discerned for the GTO orbit
- Higher successful launch outcomes on heavy payload flights are associated with flights targeting the PO, LEO, and ISS orbits



Launch Success Yearly Trend

- The average success rate of Falcon 9 launch outcomes increased steadily from 2013 to 2020



All Launch Site Names

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

- This dataframe lists each of the launch sites used to launch Falcon 9 rockets.

Launch Site Names Begin with 'CCA'

launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

- This dataframe lists five records of launch sites that begin with the string ‘CCA’.

Total Payload Mass

	1
	45596

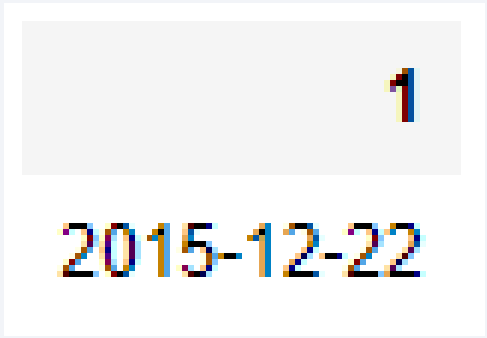
- This dataframe lists the total payload in kilograms carried by boosters launched by NASA (CRS).

Average Payload Mass by F9 v1.1

	1
2928	

- This dataframe lists the average payload mass carried by booster version F9 v1.1v.

First Successful Ground Landing Date



	1
0	2015-12-22

- This dataframe lists the date of the first successful ground pad landing.

Successful Drone Ship Landing with Payload between 4000 kg and 6000 kg

booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 B4 B1043.1
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1046.3
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

- This dataframe lists the booster versions that both successfully landed on drone ship and carried payload masses between 4000 and 6000 kg.

Total Number of Successful and Failed Mission Outcomes

- This dataframe lists the total number of mission outcomes, both successful and failed.

Boosters Carried Maximum Payload

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

- This dataframe lists the names of boosters carrying the maximum payload mass.

2015 Launch Records

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- This dataframe lists the failed landing outcomes in drone ship, their booster versions, and the launch sites used for those launches in the year 2015.

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

landing__outcome
No attempt
Failure (drone ship)
Success (drone ship)
Controlled (ocean)
Success (ground pad)
Failure (parachute)
Uncontrolled (ocean)
Precluded (drone ship)

- This dataframe lists the landing outcomes between 2010-06-04 and 2017-03-20, ranked in descending order.

Section 4

Launch Sites Proximities Analysis

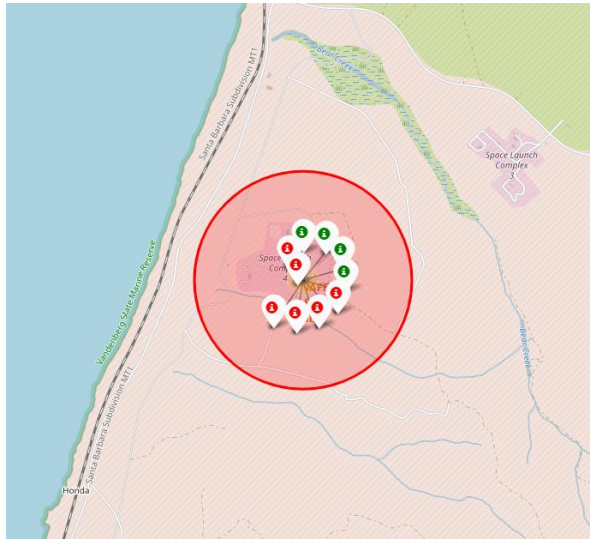
Falcon 9 Launch Sites

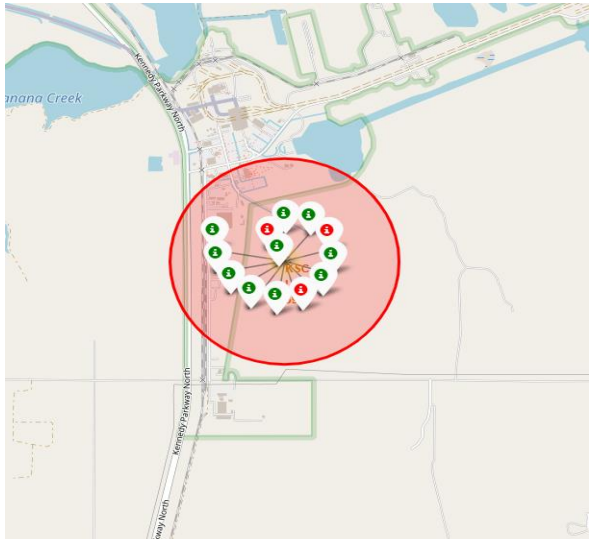
- Locations of Falcon 9 launch sites designated by circles around their coordinates and their respective launch site names listed as markers.

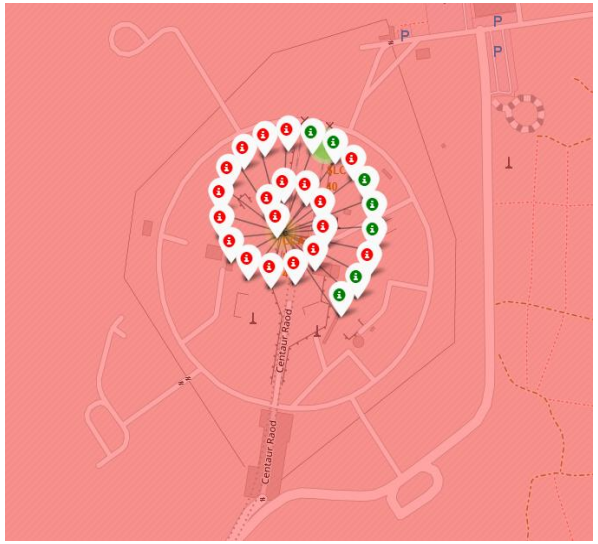
	Launch Site	Lat	Long
0	CCAFS LC-40	28.562302	-80.577356
1	CCAFS SLC-40	28.563197	-80.576820
2	KSC LC-39A	28.573255	-80.646895
3	VAFB SLC-4E	34.632834	-120.610746

Relative Success of Launches by Launch Site

- Color-labeled launch outcomes overlaid on each launch site indicating the number of successful and failed Falcon 9 launches. Green markers indicate a successful launch, whereas red markers indicate failed launches.



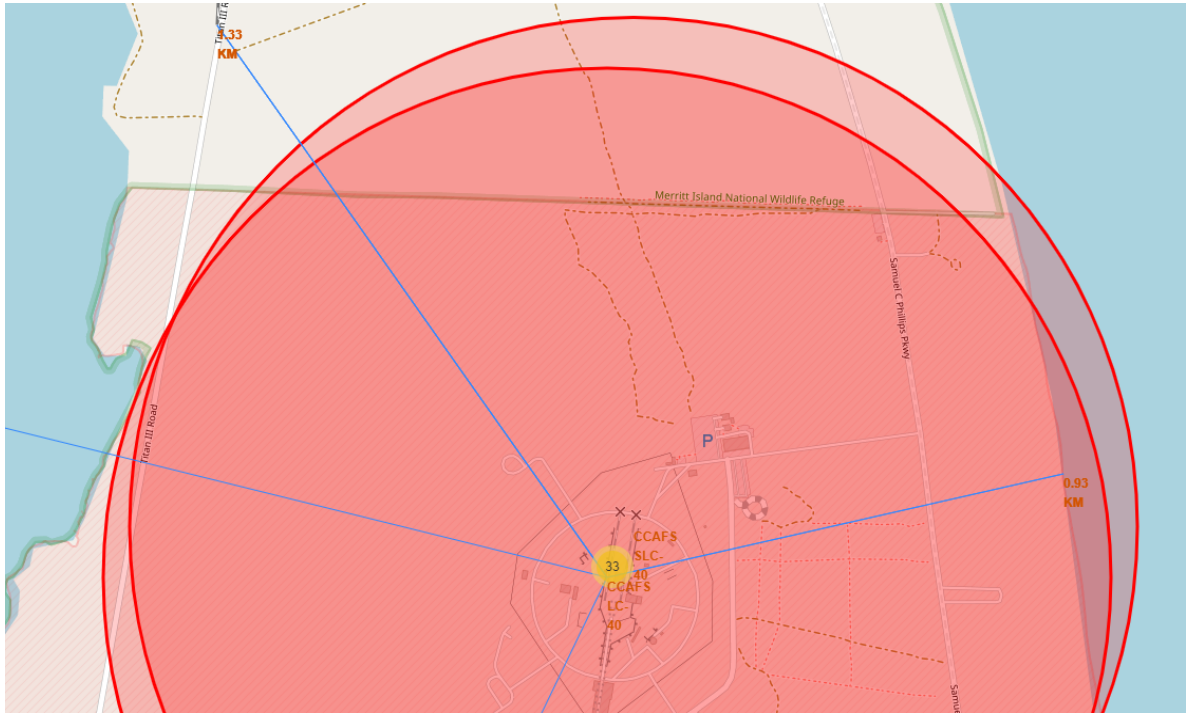


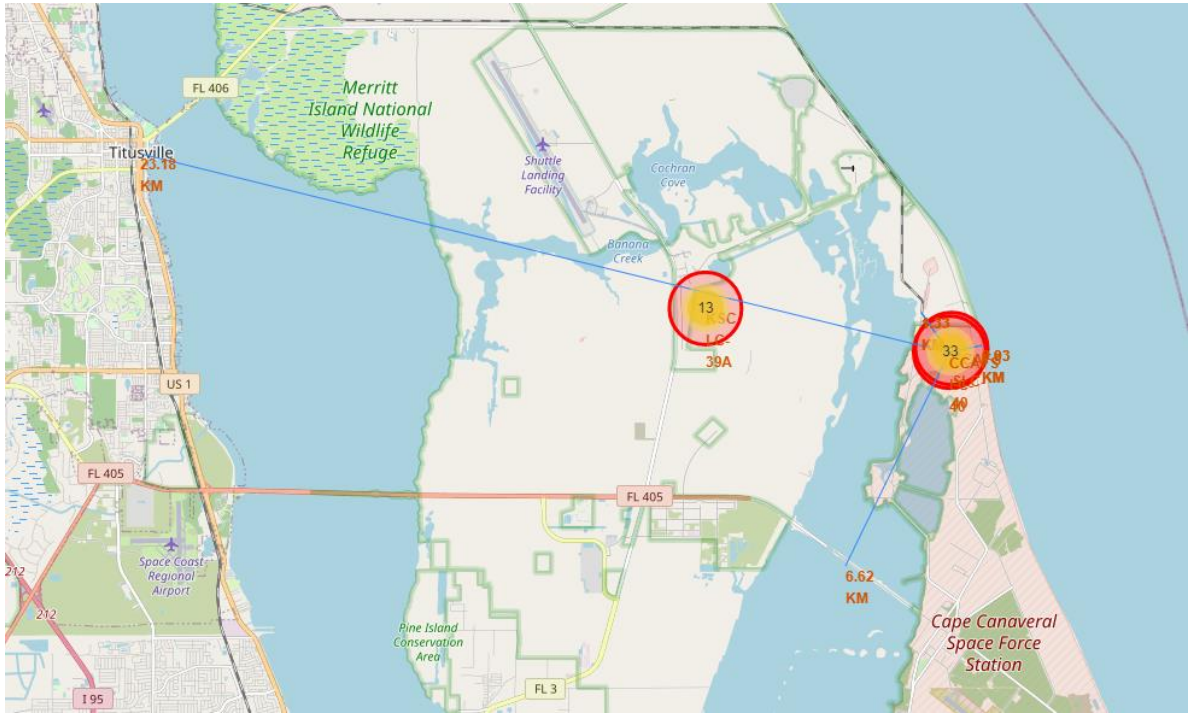


	VAFB SLC-4E	CCAFS SLC-40	CCAFS LC-40	CCAFS SLC-40
Total Success:	4	10	7	3
Success Rate:	40%	77%	27%	43%

Proximity of the CCAFS-LC 40 Launch Site to Nearby Points of Interest

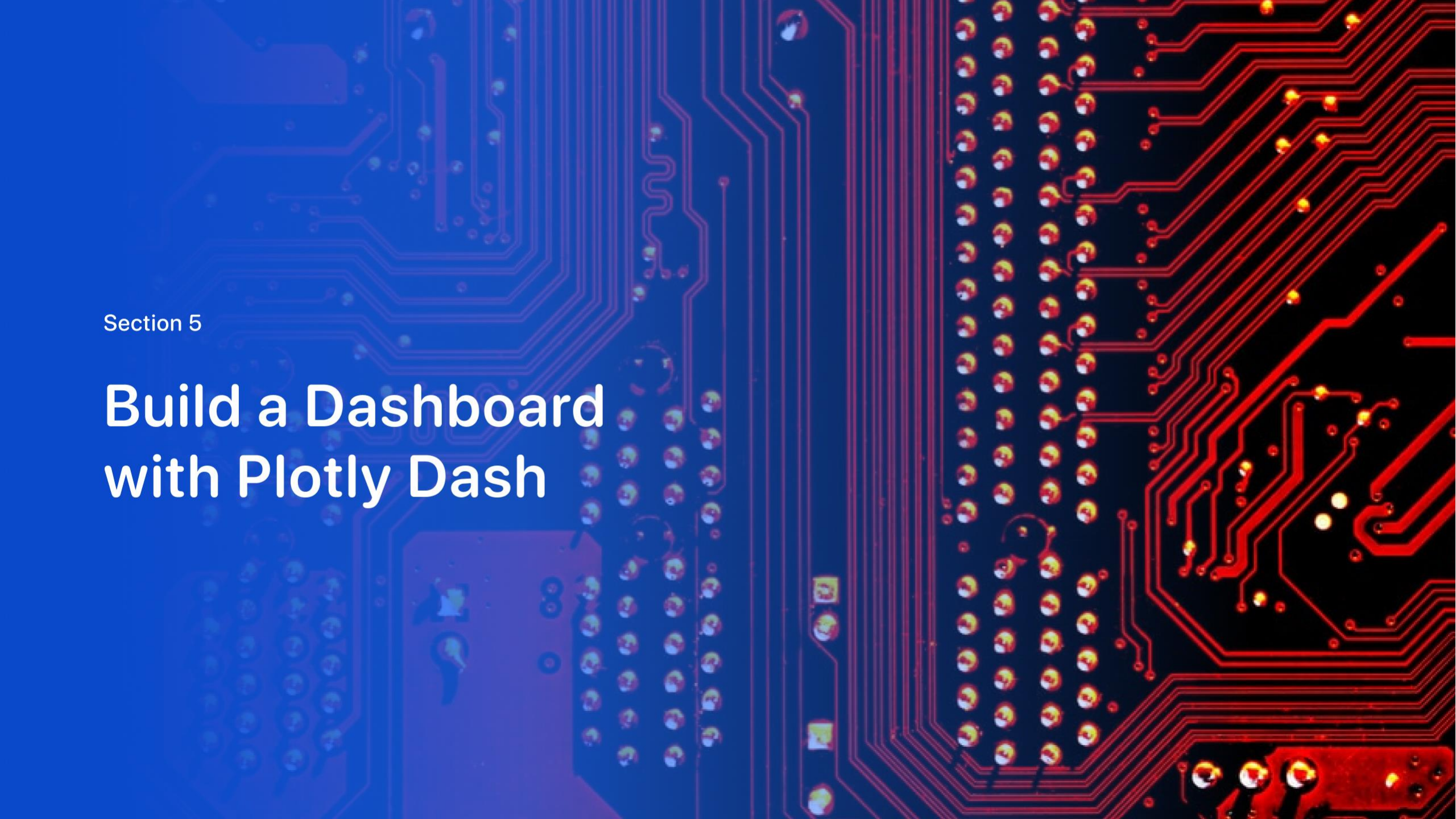
- Distance of the CCAFS-LC 40 Launch Site to the nearest coastline (0.93 km), railway (1.33 km), highway (6.62 km), and city (23.18 km). Distances between the launch site and points of interest are marked by blue lines.





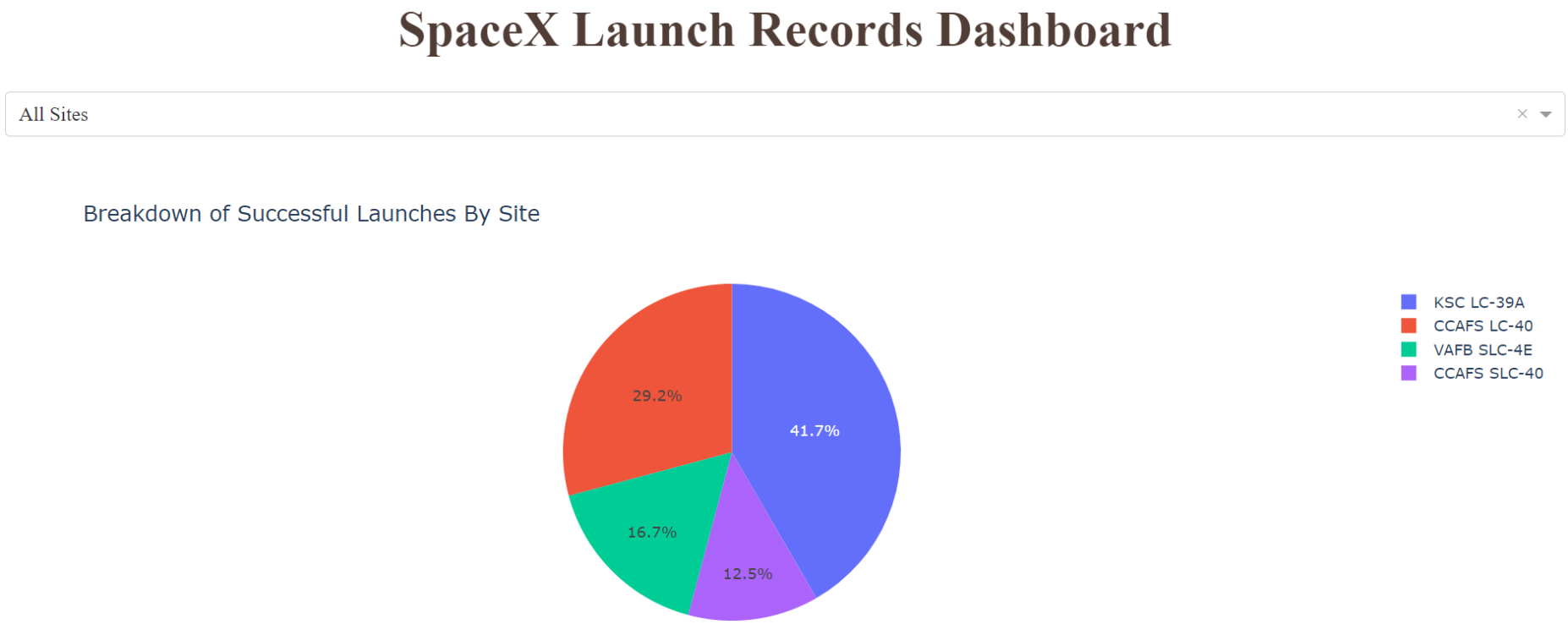
Section 5

Build a Dashboard with Plotly Dash



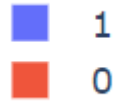
Breakdown of Successful Launches At All Sites

- The ratio of successful-to-failed launch outcomes is different at each site
- The largest percentage of the total successful launch outcomes occurred at KSC LC-39A
- The lowest percentage occurred at CCAFS SLC-40

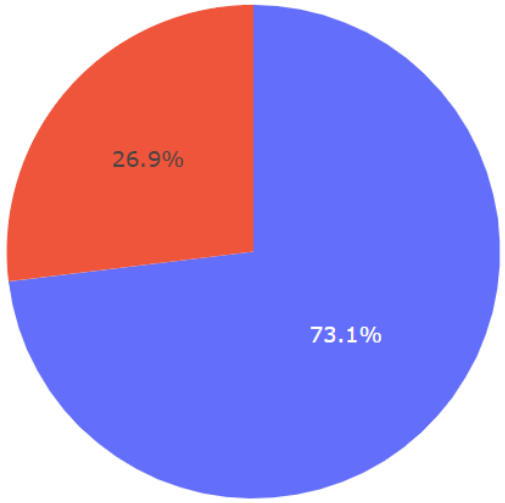


Breakdown of Successful Launches At Each Site

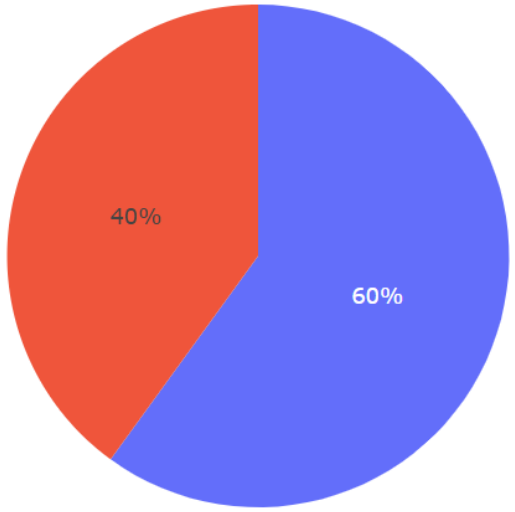
- Blue (1) indicates a successful outcome, whereas red (0) indicates a failed outcome
- The launch site with the highest percent of successful launch outcomes was KSC LC-39A
- The launch site with the lowest percent of successful launch outcomes was CCAFS SLC-40



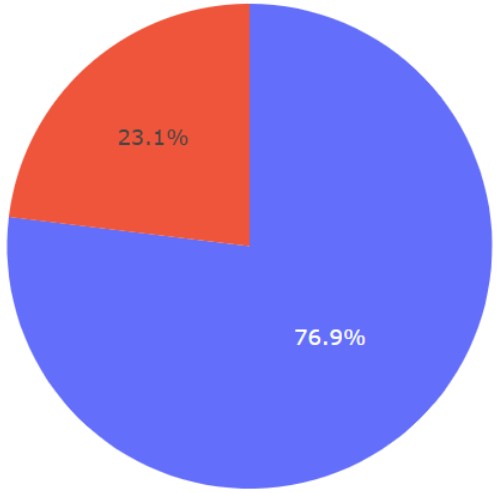
Breakdown of Successful Launches at CCAFS LC-40



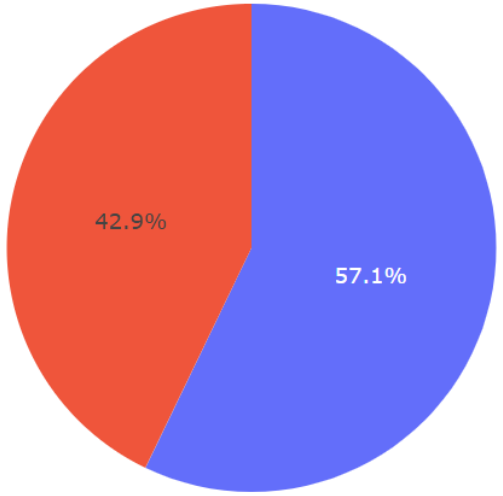
Breakdown of Successful Launches at VAFB SLC-4E



Breakdown of Successful Launches at KSC LC-39A



Breakdown of Successful Launches at CCAFS SLC-40

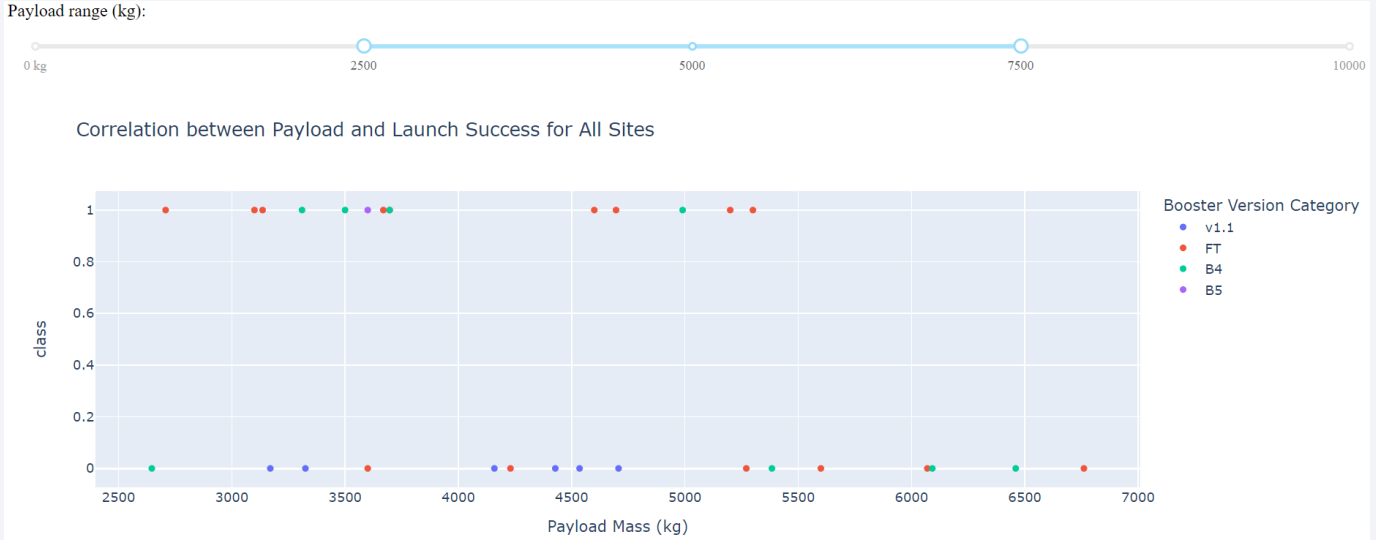


Payload vs. Launch Outcome – All Sites

Payload range (kg):

0 - 10000

2500 - 7500

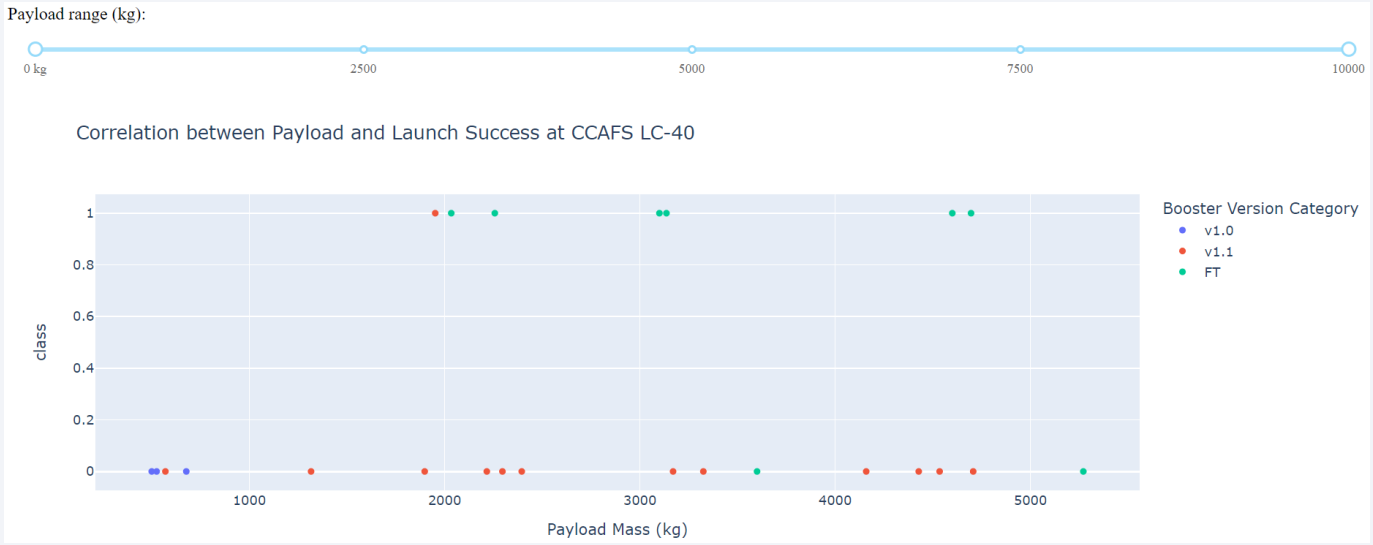


- Rockets of all booster versions at all launch sites have higher launch success rates when carrying light payloads
- At all launch sites, the FT booster version is associated with the largest percentage of successful launch outcomes
- At all launch sites, the v1.1 booster version is associated with the largest percentage of failed launch outcomes

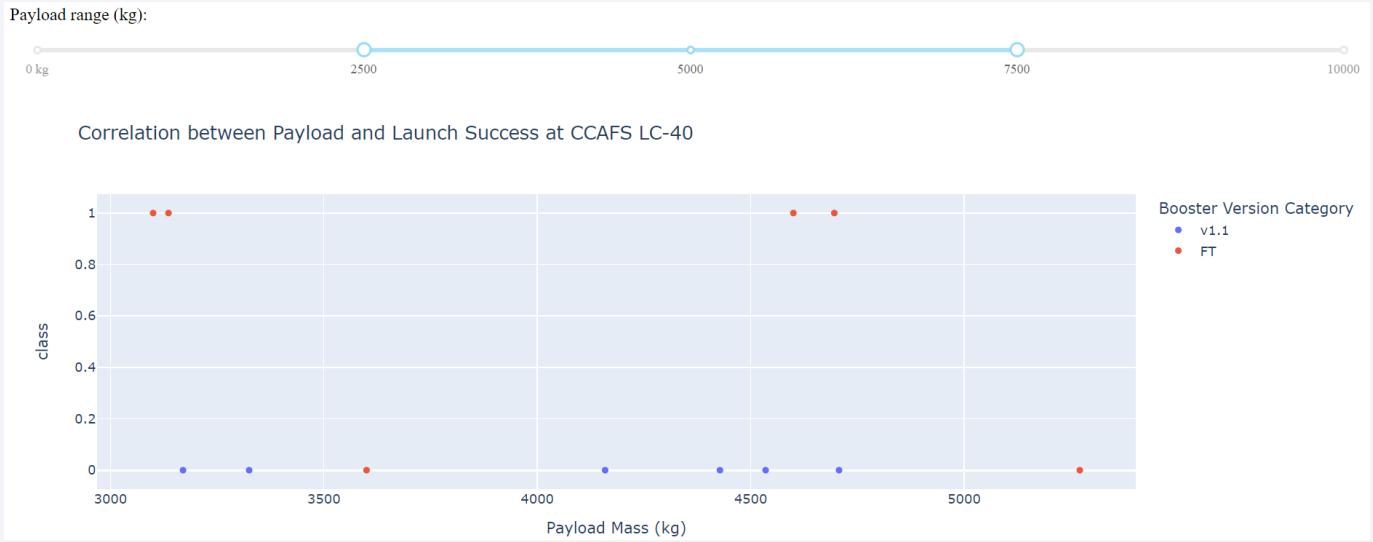
Payload vs. Launch Outcome – CCAFS LC-40

Payload range (kg):

0 - 10000



2500 - 7500

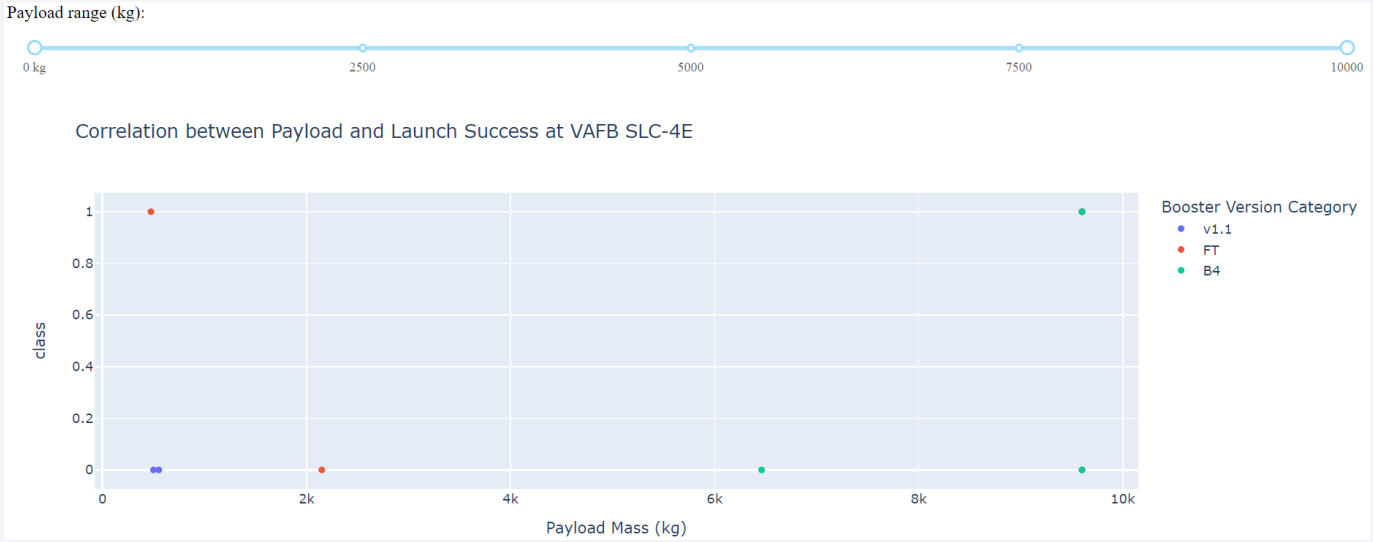


- At the CCAFS LC-40 site, all v1.1 booster version rockets had unsuccessful launch outcomes when carrying payloads above 2000 kg

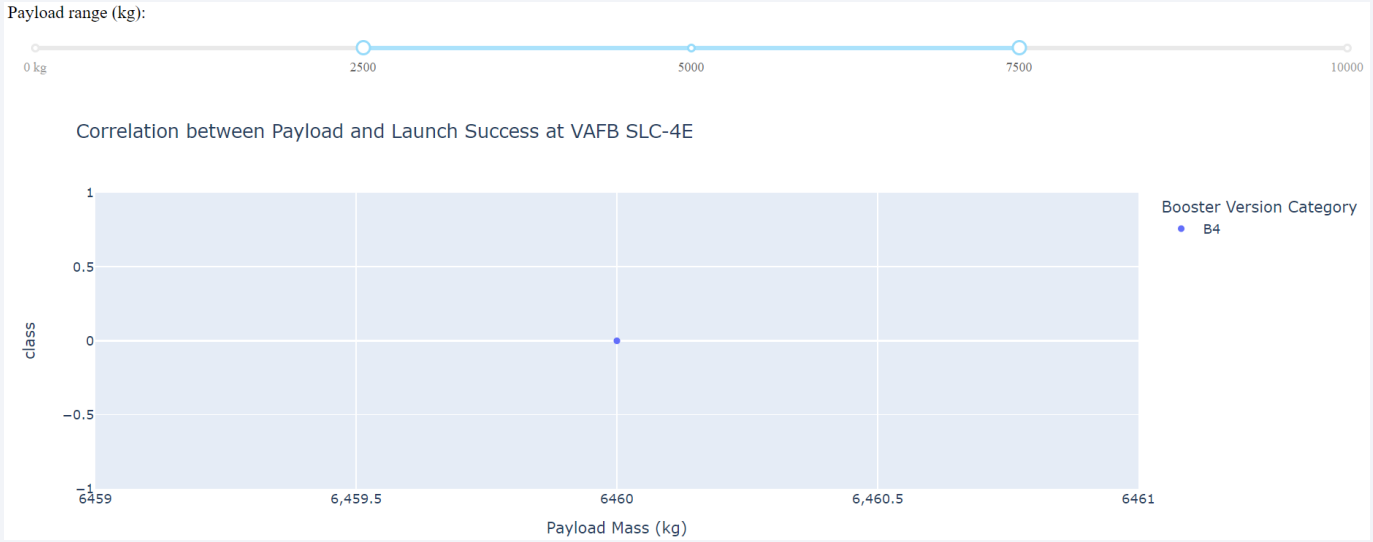
Payload vs. Launch Outcome – VAFB SLC-4E

Payload range (kg):

0 - 10000



2500 - 7500

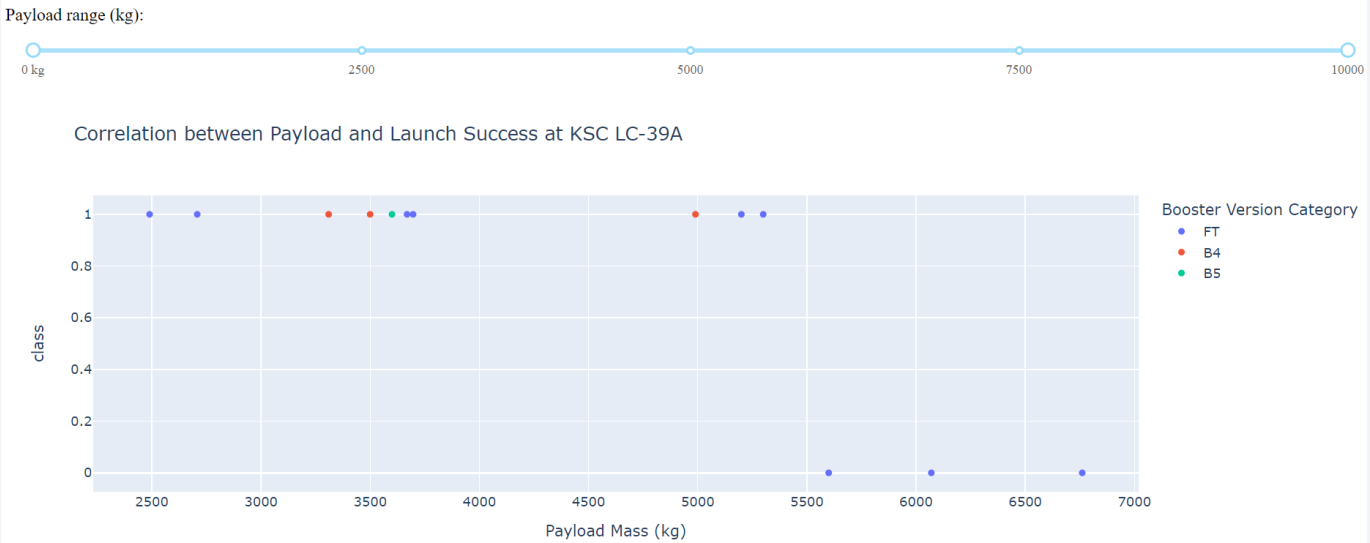


- Too few launches were attempted at the VAFB SLC-4E site to draw meaningful conclusions
- It is notable that booster version B4 was used for heavy payloads (> 6000 kg), whereas booster versions v1.1 and FT were used for light payload (< 3000 kg) missions

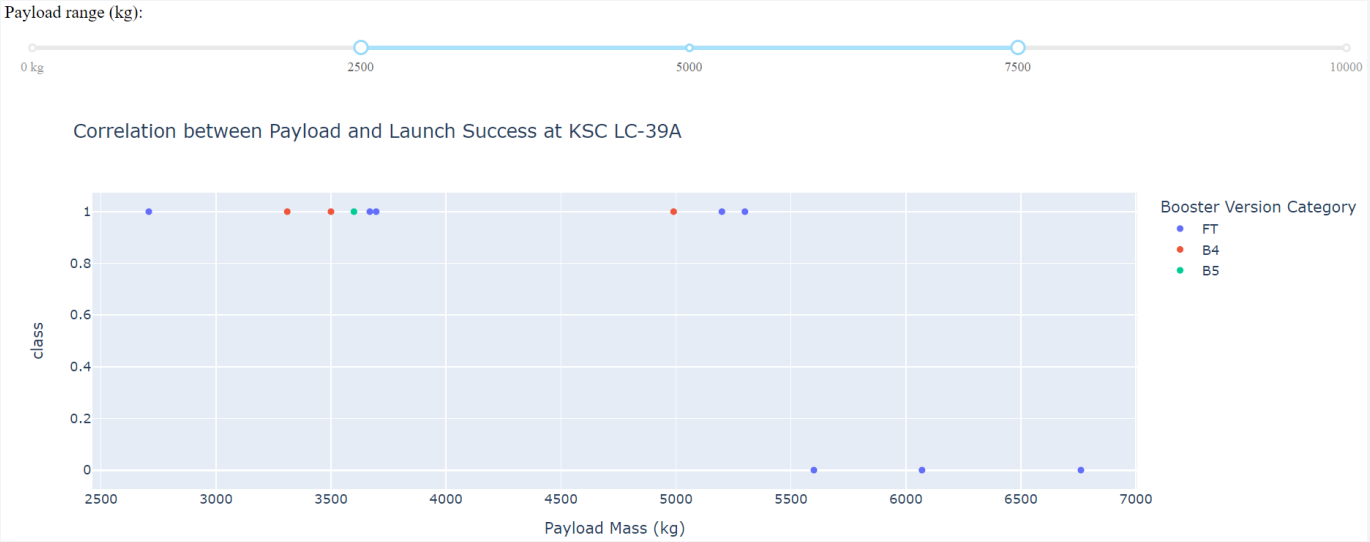
Payload vs. Launch Outcome – KSC LC-39A

Payload range (kg):

0 - 10000



2500 - 7500

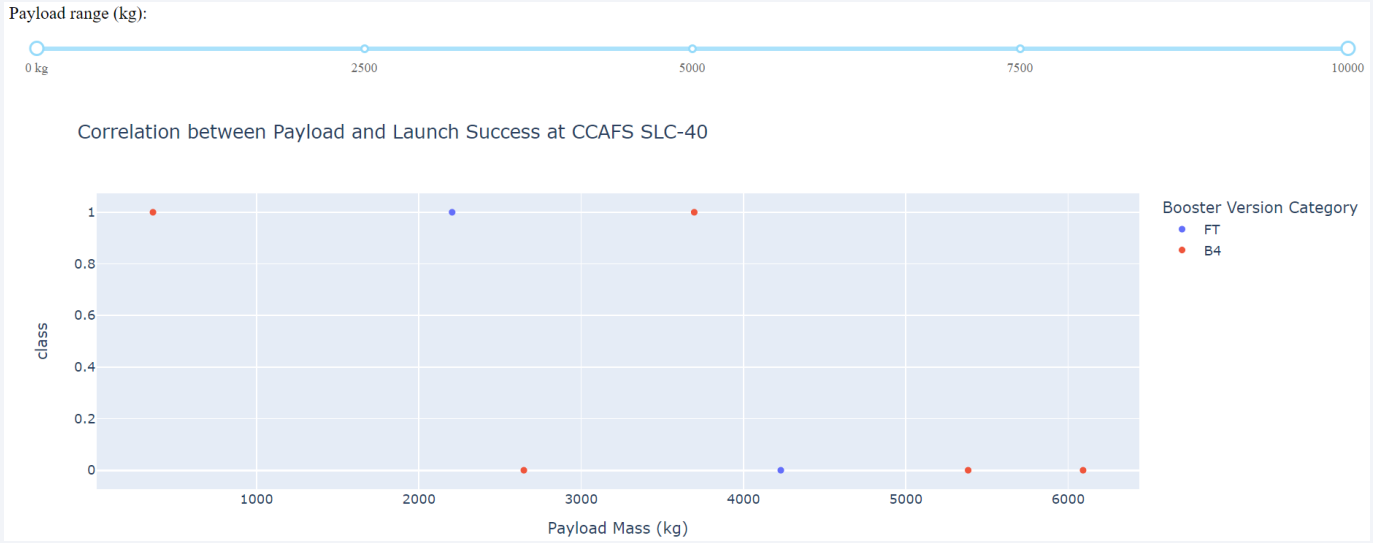


- All launch outcomes associated with the B4 booster version were successful when payloads were below 5500 kg, but failed when carrying heavier payloads
- All booster versions used at this site were associated with successful missions when the payload was below 5500 kg

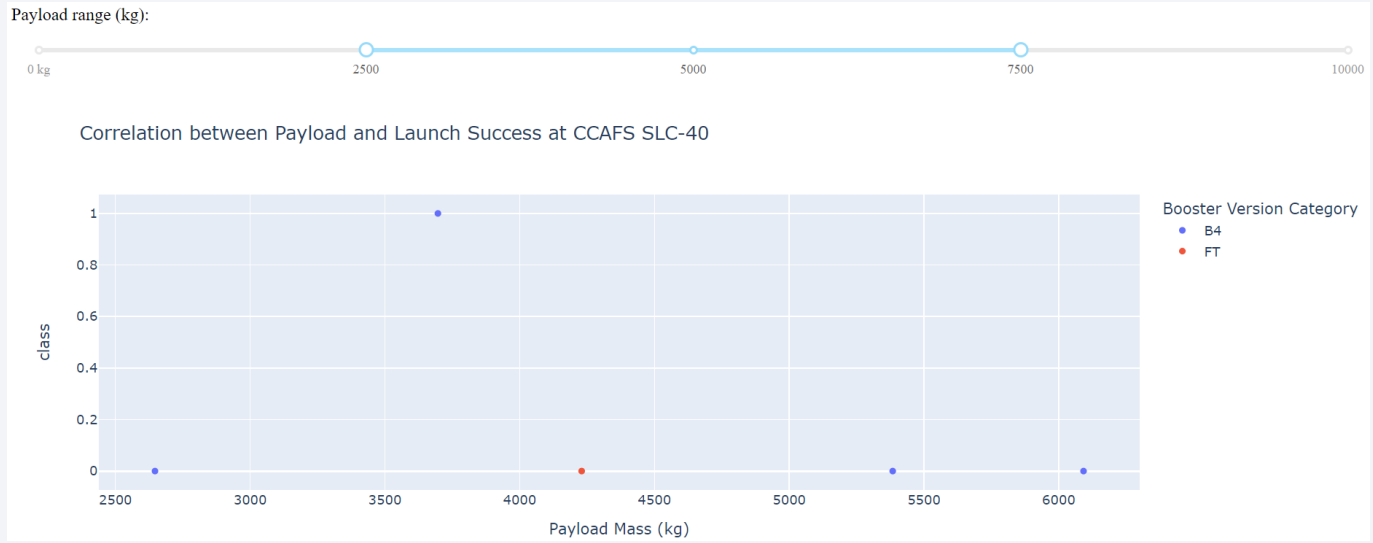
Payload vs. Launch Outcome – CCAFS SLC-40

Payload range (kg):

0 - 10000



2500 - 7500



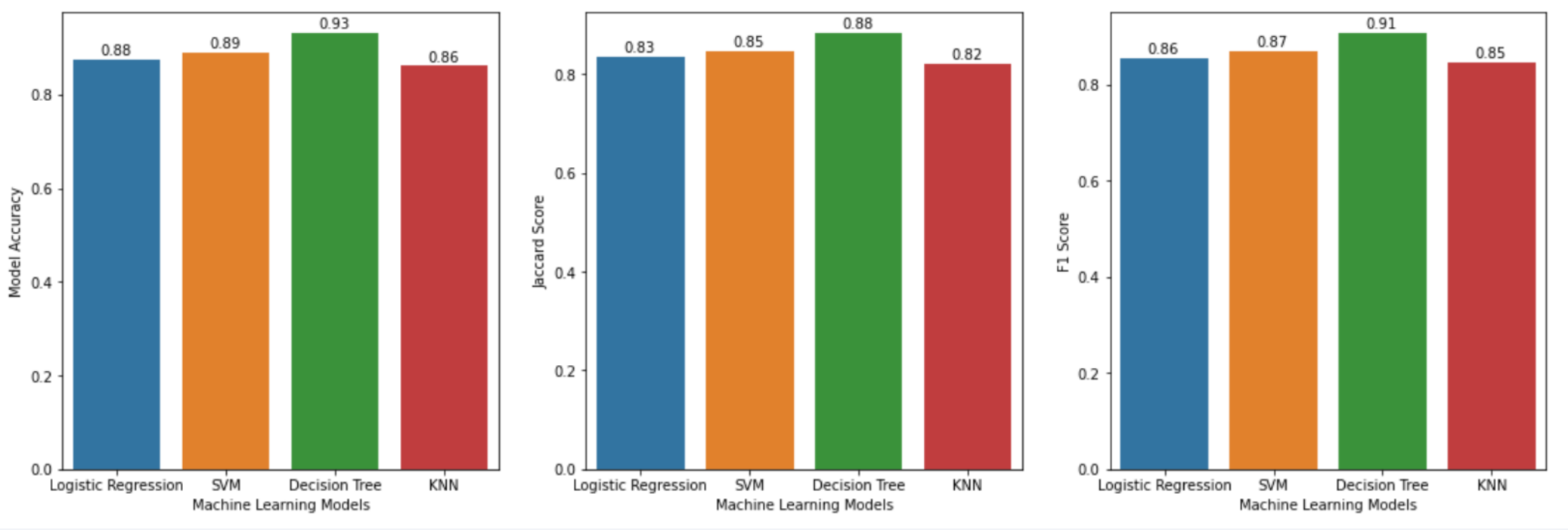
- Not many launches were attempted at the CCAFS SLC-40 site, but all successful launch outcomes for both booster versions occurred when the payload was below 4000 kg, but failed above that payload mass

Section 6

Predictive Analysis (Classification)

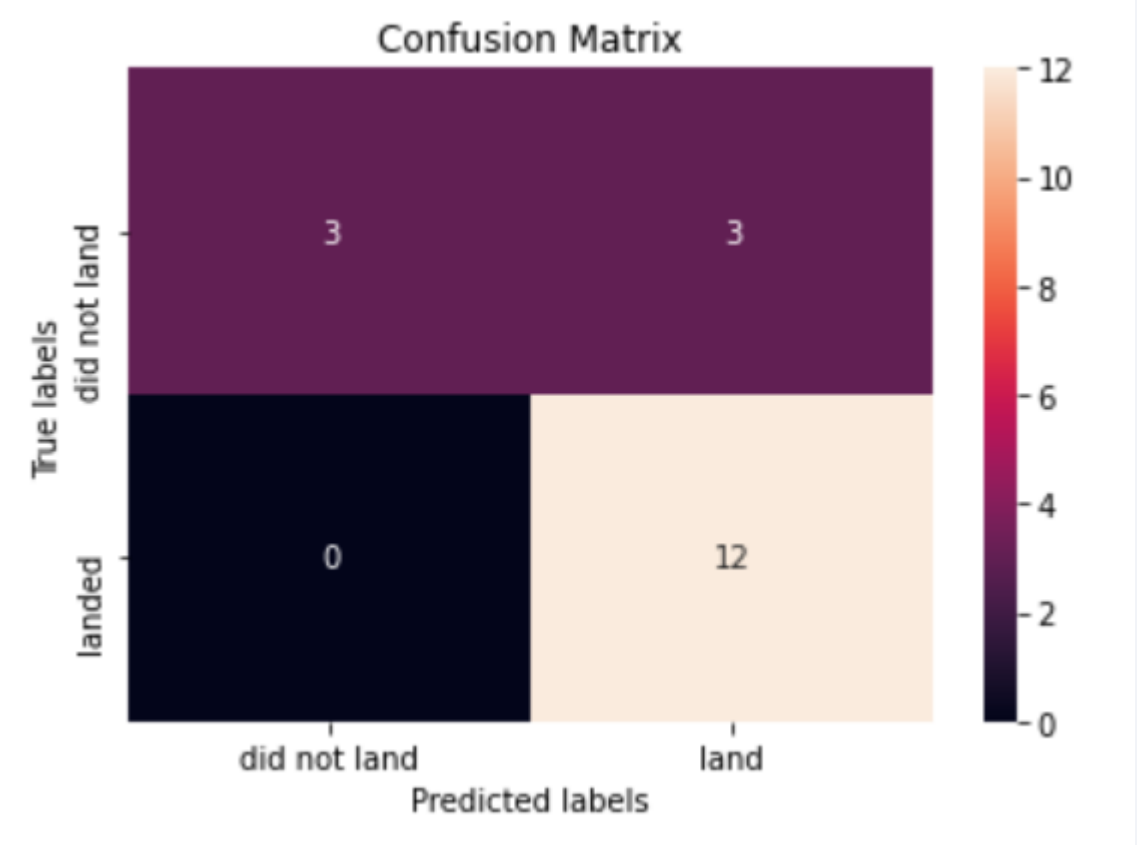
Classification Accuracy

- While all classifiers scored highly, the Decision Tree model had the highest classification accuracy (93%), as well as the highest Jaccard index score (0.88) and F1-score (0.91)



Confusion Matrix

- Confusion matrix of the Decision Tree classifier
  - True positives: 12
  - True negatives : 3
  - False positives: 3
  - False negatives: 0
- Precision: 0.80
- Recall: 1.00



Conclusions

- The average launch success rate of Falcon 9 rockets has been steadily improving since 2013
- Light payload missions (e.g. <5500 kg) tend to have higher successful launch outcomes, however the booster version, launch site, and orbit targeted are also important success determinants
  - The FT booster version is associated with the highest launch outcome success rate, however it tends to be utilized in light payload missions, whereas the B4 version tends to be used in heavy payload missions
  - The v1.1 booster versions was associated with the largest rate of failed launch outcomes
  - Flights associated with the ES-L1, GEO, HEO, and SSO orbits have high launch success rates
- The B5 booster version was designed for maximum payload missions
- The KSC LC-3A9 launch site has a high launch success rate, especially for light payload missions
- Launch sites are located far from highways and especially far from cities, but are near coastlines
- All classifiers whose hyperparameters were optimized scored highly during model evaluation
  - The Decision Tree classifier ranked highest in predicting the test data
  - The confusion matrices of each classifier indicated that an area of weakness was in the False Positive rate

Appendix

- Optimal hyperparameters used in classifiers
  - Logistic Regression:

```
tuned hyperparameters :(best parameters)  {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}
```

  - SVM:

```
tuned hyperparameters :(best parameters)  {'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}
```

  - Decision Tree:

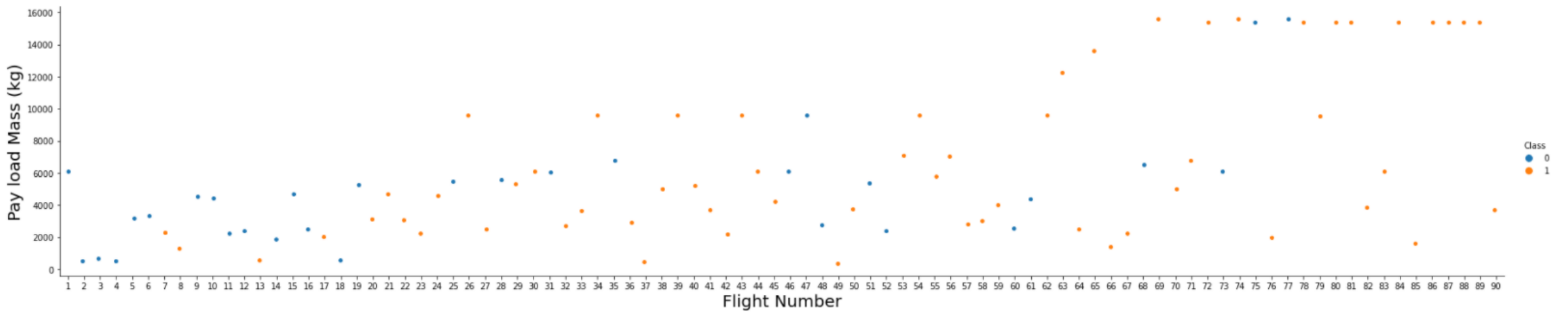
```
tuned hyperparameters :(best parameters)  {'criterion': 'entropy', 'max_depth': 18, 'max_features': 'auto', 'min_samples_leaf': 4, 'min_samples_spli
t': 5, 'splitter': 'best'}
```

  - KNN:

```
tuned hyperparameters :(best parameters)  {'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}
```

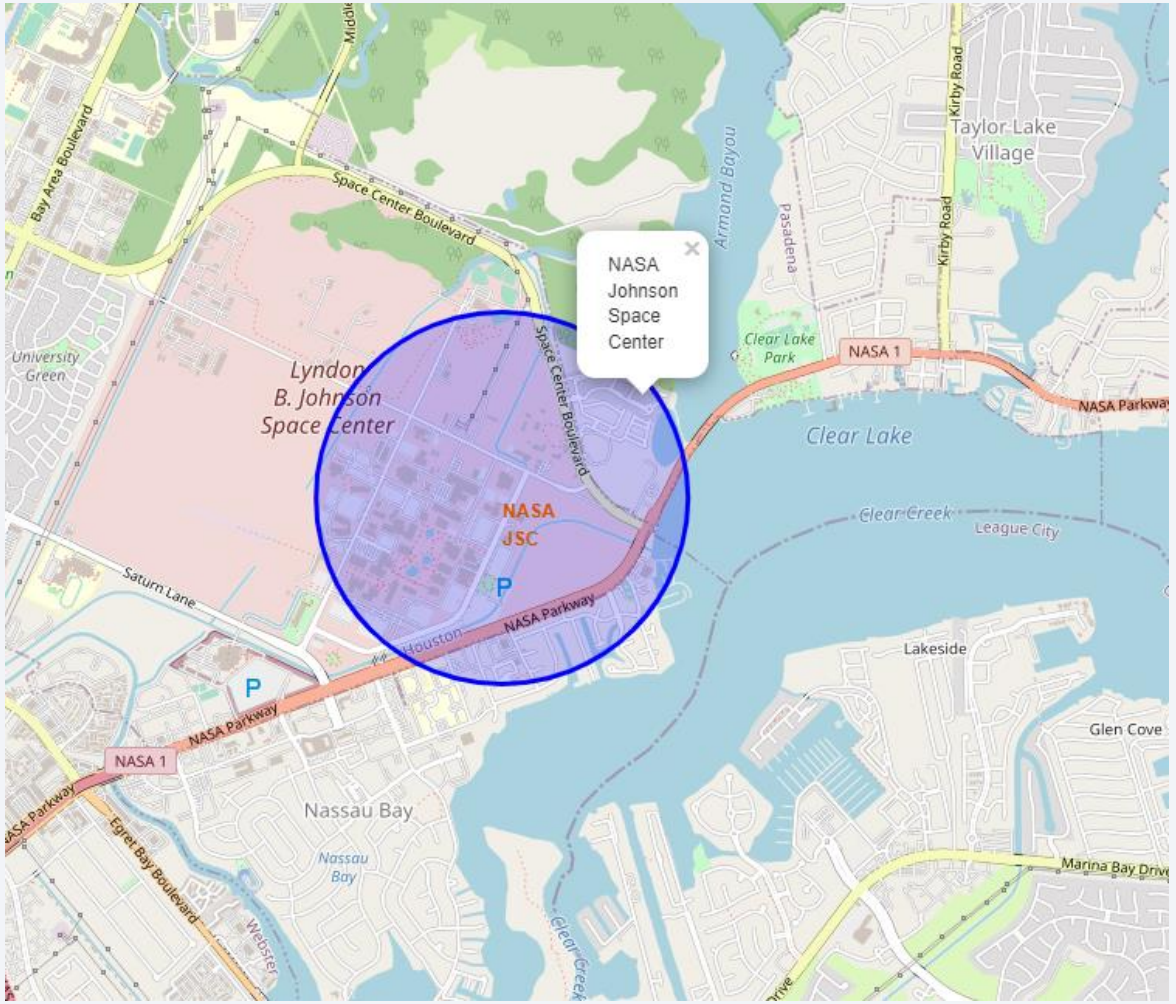
Appendix

- Flight Number vs. Payload Mass (kg)
- KSC LC-39A and VAFB SLC 4E had launch success rates of 77%, whereas CCAFS LC-40 had a success rate of 60%
- Heavier payloads have been carried successfully into orbit over time



Appendix

- Visualizing the NASA Johnson Space Center in folium with circle and marker labels



Thank you!