IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Emily Brehmer
August 8th, 2022

Outline

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

Executive Summary

To assist in winning the space race, this project is focused on determining if SpaceX will successfully recover the most expensive part of a Rocket Launch, the Stage1 Component.

Using historical SpaceX data, we have determined the factors that lead to a successful recovery of the Stage1 component. These factors include the details on the launch, the mission, and the customer.

Finally using a decision tree model, we are able to predict if a new launch will successfully recover the Stage1 component with an accuracy of approximately 94%.

Introduction

Our company, SpaceY, is currently competing against several others to secure Rocket Launch Missions from potential clients. SpaceX is our leading competitor, offering the lowest prices. This is mainly due to SpaceX being able to recover the most expensive component of a launch sequence, the Stage1. *More information on the Stage1 component is provided in the following slide.* Therefore, if we can determine the chance the Stage1 will be recovered, we can determine the possible cost of a SpaceX mission.

To assist SpaceY in offering the most competitive bid, this project is designed to answer the question:

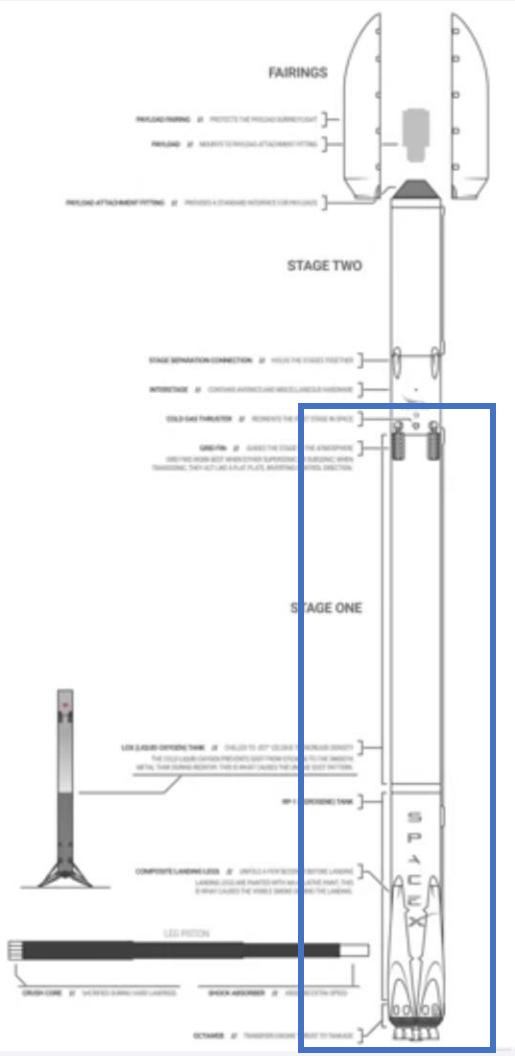
Will SpaceX recover the Stage1 component for the requested Launch Mission?

What are the main features that determine if the Stage1 component will be recovered?

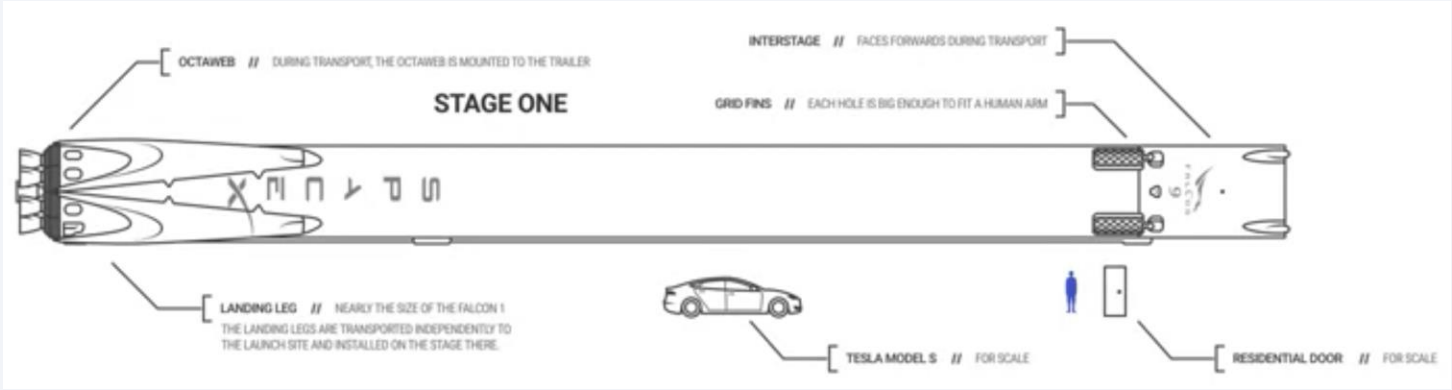
The Hypothesis :

SpaceX historical details of previous launches, missions, and the customers, will allow the creation of a model that can accurately predict if a Stage 1 component will be recovered in an upcoming launch

Stage1 Review



- The image on the left is of a Falcon 9 Rocket, with the Stage1 component outlined in blue. This is the largest component of the launch.
- The below image farther demonstrates the size of the Stage1. This stage does most of the work to get the mission into space.
- Recovery of the Stage1 component is rare in across other companies.



Section 1

Methodology

Methodology

Executive Summary

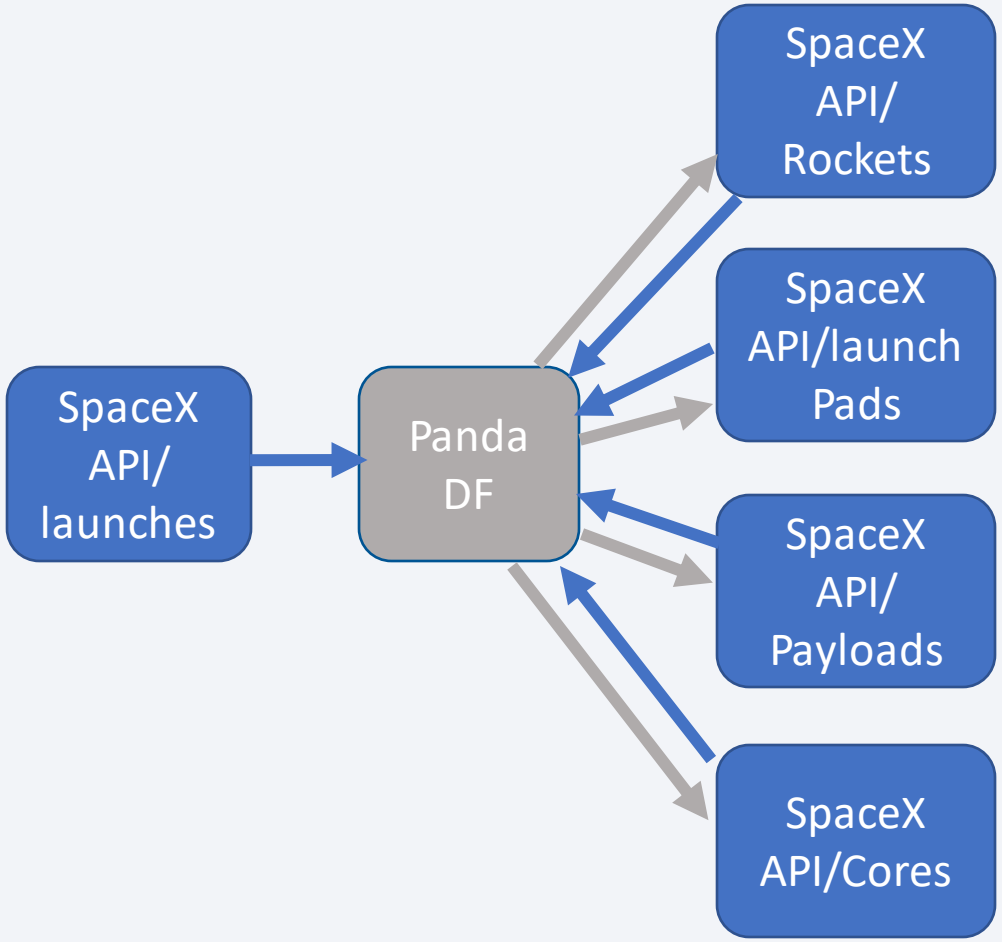
- Data collection methodology:
  - The launch data is collected from two sources : SpaceX API and the Wiki page : List of Falcon9 and Falcon Heavy Launches
- Perform data wrangling
  - Incomplete launch data was collected from SpaceX API, Null Values were converted to mean where applicable, hot encoding converted text data to numerical values, and the outcome was converted to a distinguishable bool.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistic Regression, KNN, SVM, and Decision Tree models were tested with Grid Search CV to find the most accurate model.

Data Collection

- Two sources were used for Data Collection of historical launch data
  - SpaceX API URL : https://api.spacexdata.com/v4/launches/past
  - Wiki page : List of Falcon9 and Falcon Heavy Launches : https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetwork-Channel-SkillsNetworkCoursesIBMDS0321ENSkillsNetwork26802033-2022-01-01
- SpaceX API was converted from Json to fit into a Pandas DataFrame for further processing
- Falcon9 and Falcon Heavy Launches data was processed using Beautiful Soup to web scrape the data, which was then stored into a Pandas DataFrame for further processing.

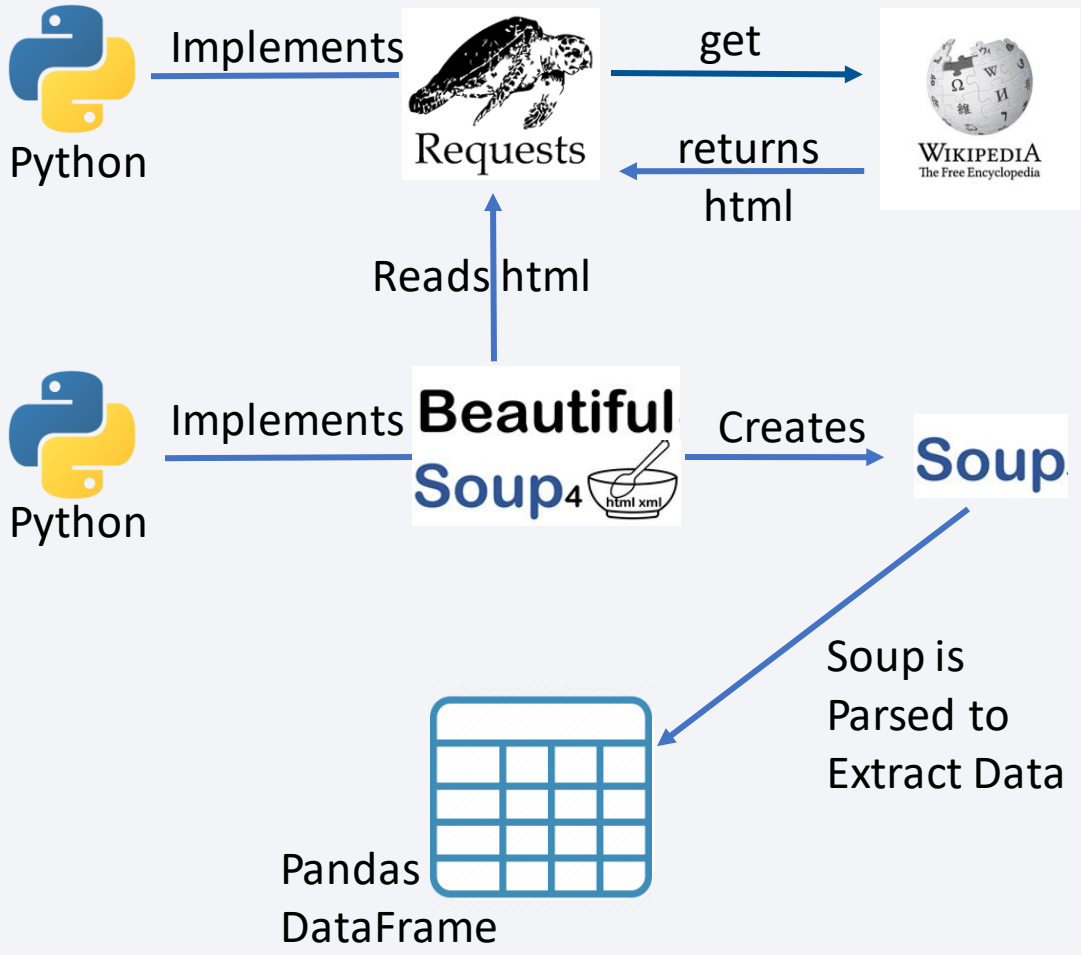
Data Collection – SpaceX API

- SpaceX Data was collected through an API call to retrieve past launch data.
- The launch data contained reference codes for the Rockets, LaunchPads, Payloads, and Core data, which required multiple calls to the API to gather remaining the information.
- All data was stored within a Pandas DataFrame for further processing.
- GitHub URL: API Data Collection



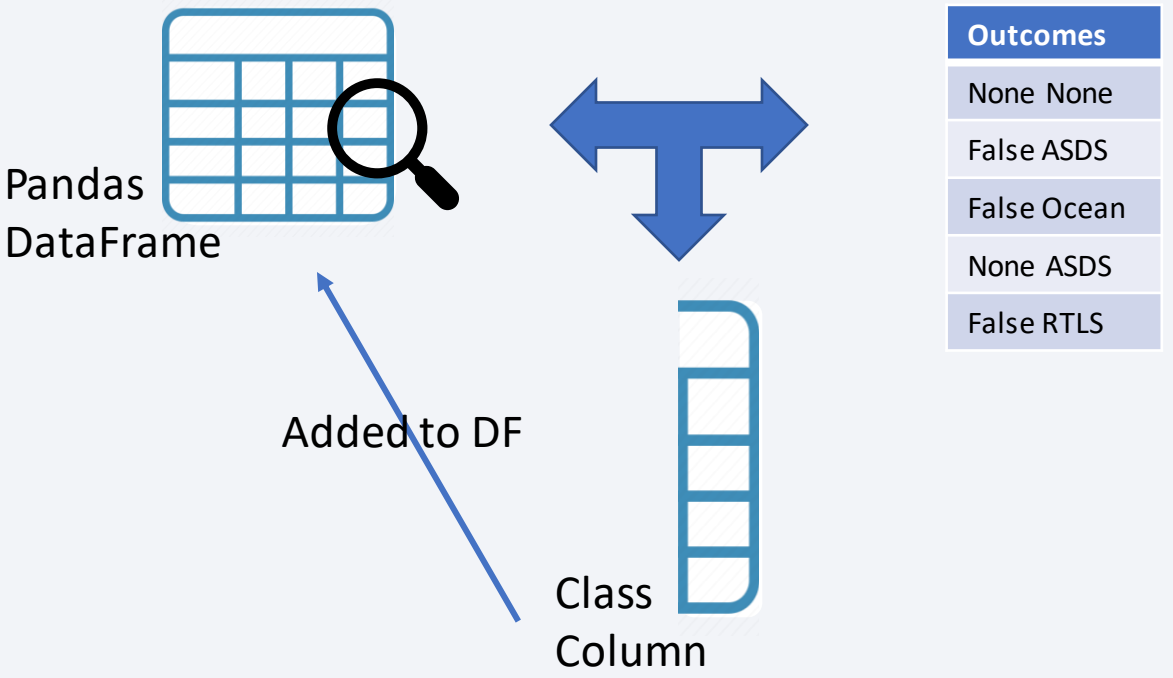
Data Collection - Scraping

- Extract a Falcon 9 launch records HTML table from Wikipedia
- Parse the table and convert it into a Pandas data frame
- GitHub URL: Webscraping Data Collection



Data Wrangling

- The dataset created from SpaceX and the Wiki page were examined to determine a method that would distinguish between a successful Stage1 recovery and a failure.
- A list of failed outcomes was created to separate the data into two classes.
- 1 – Success
- 0 – Failure
- GitHub URL: Data Wrangling



EDA with Data Visualization

- Scatter Plots show the correlation between the Flight Number, Launch Site, Payload, Orbit, and Outcome/ Class
- A Bar Chart demonstrates the various success rates of different Orbit Missions
- Finally another scatter plot was used to demonstrate the yearly trend of successful Stage1 recovery missions.
- Outcome: Mission success correlates to the Flight number, Launch Site, Payload, and Orbit of the mission.
- GitHub URL: Data Visualization

EDA with SQL

- SQL Queries Performed:
  - Retrieved a list of all Launch Site names
  - Displayed 5 records for Launch sites beginning with CCA
  - Provided total payload mass carried by boosters launched by NASA (CRS)
  - Provided Average payload mass carried by booster version "F9 v1.1"
  - Retrieved Date of first successful landing outcome in Ground Pad
  - Names of the boosters with a success in Drone ship with payload mass between 4000 and 6000
  - Totaled the success and failure missions
  - Delivered Booster names that carried the max payload mass
  - Booster Version and Launch Site for failed drone ship from 2015
  - Ranked the count of different landing outcomes between 2010-06-04 and 2017-03-20 in descending order
- GitHub URL: Exploratory Data Analysis with SQL

Build an Interactive Map with Folium

- Map Contents :
  - Marker Cluster at each launch site – this provided the user with the count of launches at each site, as well as visual breakdown of the success and failures.
  - Coastal Line - indicated to the user the distance from the launch sites to the coast to demonstrate that all launch sites are located in close proximity to a coast.
  - Railroad Line - Line connecting the launch site to the nearest railroad track. The distance is included so the user can visually see that railroad tracks are within a mile of each launch site.
  - City Line - Line connecting the launch site to the nearest city. The distance is included as well so the user can see that cities are often 50km away from a launch site. Exclusion is in California, where the launch site is within 13km of a small city.
  - Highway Line - Line connecting the launch sites to the nearest highway, with a distance measurement. In most cases the highways are near the launch site. There is one exclusion in California, where the highway is 13km away.
- GitHub URL: Interactive Map

Build a Dashboard with Plotly Dash

- Dashboard Contents:
  - Drop down menu to select launch sites – defaulted to select all. Created to allow the user to interact with the charts and graphs on the page, providing further insight into Launch Sites. This menu is also searchable, so a user can begin typing to filter the launch site list.
  - Pie Chart - Is the first chart on the page that shows the percentage each launch site contributes to the overall success. When a launch site is selected, the pie chart shows the success and failure percentage for that site.
  - Range Slider for Payload Mass – defaulted to the min and max payload of the dataset. Created to allow the user to interact with the Scatter plot to provide more insight into payload mass correlation to successful recoveries.
  - Scatter Plot – A color coded plot of the payload mass and the Stage1 recovery success. The colors correspond to the different booster versions in the dataset. This plot is also connected to the dropdown menu and defaults to show the data for all launch sites, but will update to the show only selected launch site data.
- GitHub URL: Plotly Dash Python File

Predictive Analysis (Classification)

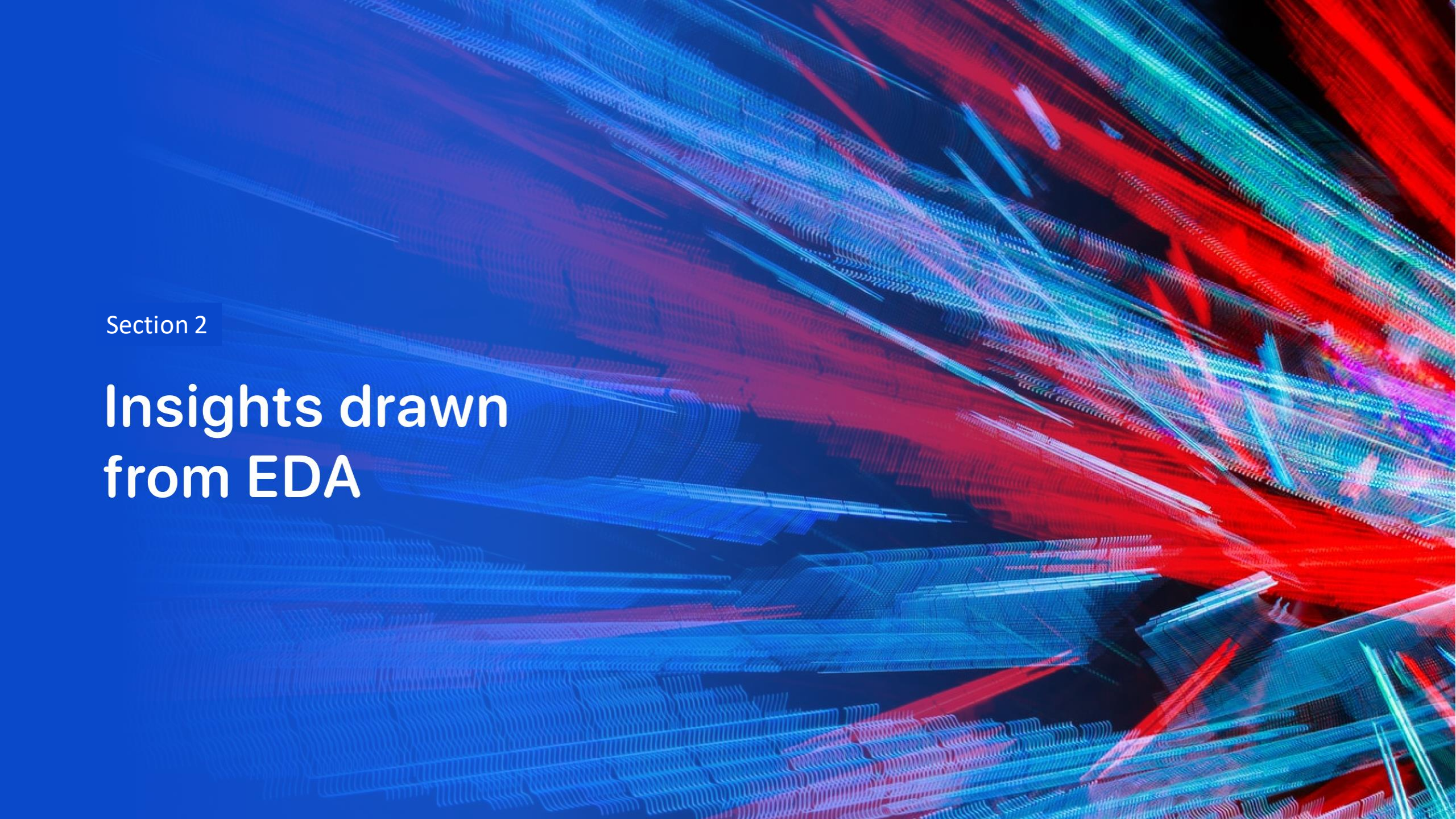
- After determining the features that correlate to the Stage1 landing outcome, the data is pulled into a Pandas DataFrame to create a training model.
- These features were then standardized and transformed to provide an accurate outcome.
- 80% of the standardized data was used as the training set and 20% made up the testing set.
- With the data setup complete, four classification methods were modeled by setting up a set of hyperparameters that would be used along with the GridSearchCV method. (Logistic Regression, KNN, SVM, and Decision Trees)
- The classification models were generated and tested with the GridSearchCV, determining the best hyperparameters for each model.
- Accuracy was measured among each model and a confusion matrix was generated to determine the model with the highest performance.
- GitHub URL: Classification Analysis

Results

- Correlation was identified amongst several launch features during Exploratory data Analysis:
  - Flight Number showed a linear increase in success rate as the flight number increased, which is expected as lessons are learned from each mission.
  - The orbit type also showed a strong correlation with some types reaching 100% accuracy and others having 0%.
- From our interactive report:
  - The highest success rate can be seen at KSC LC-39A launch site
  - The payload range 2000 kg – 5000 kg also produces the highest success rate at this time
- Classification Model Outcome:
  - The Decision Tree model may offer the best accuracy but can be very volatile, so we should continue to monitor accuracy until stability is achieved.
  - Alternatively, we may have better luck by exploring feature engineering further to produce a higher accuracy amongst the other models.

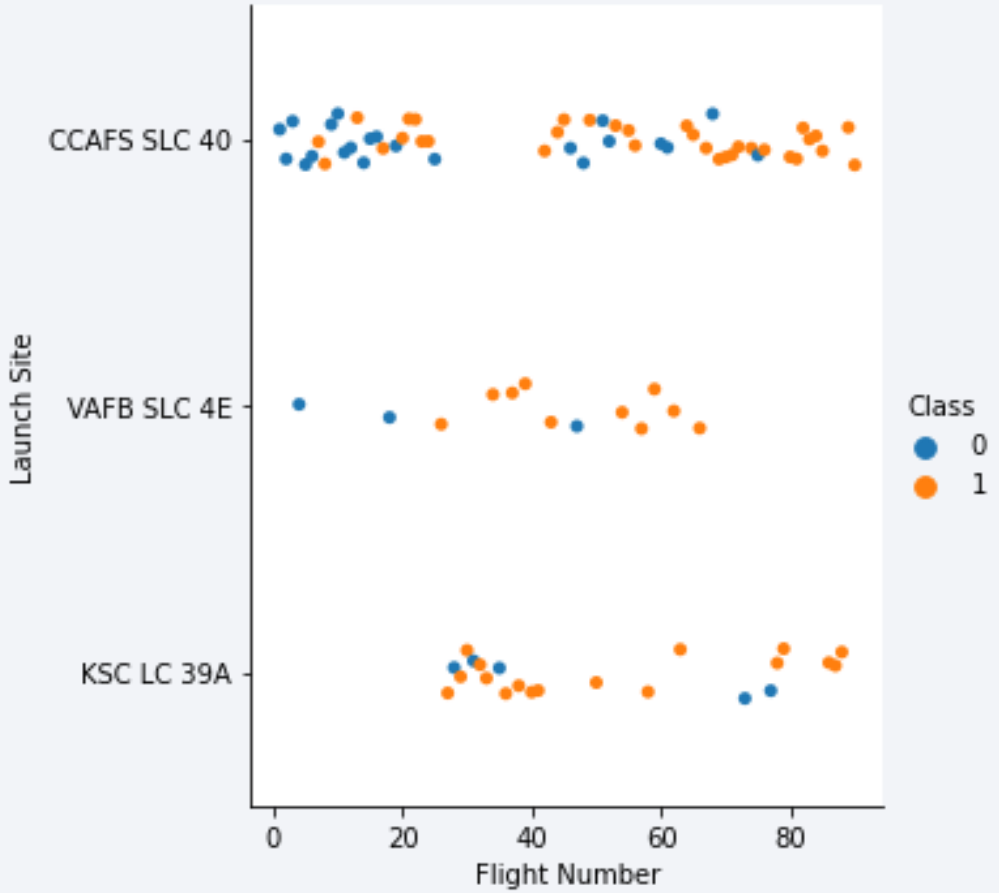
Section 2

Insights drawn from EDA



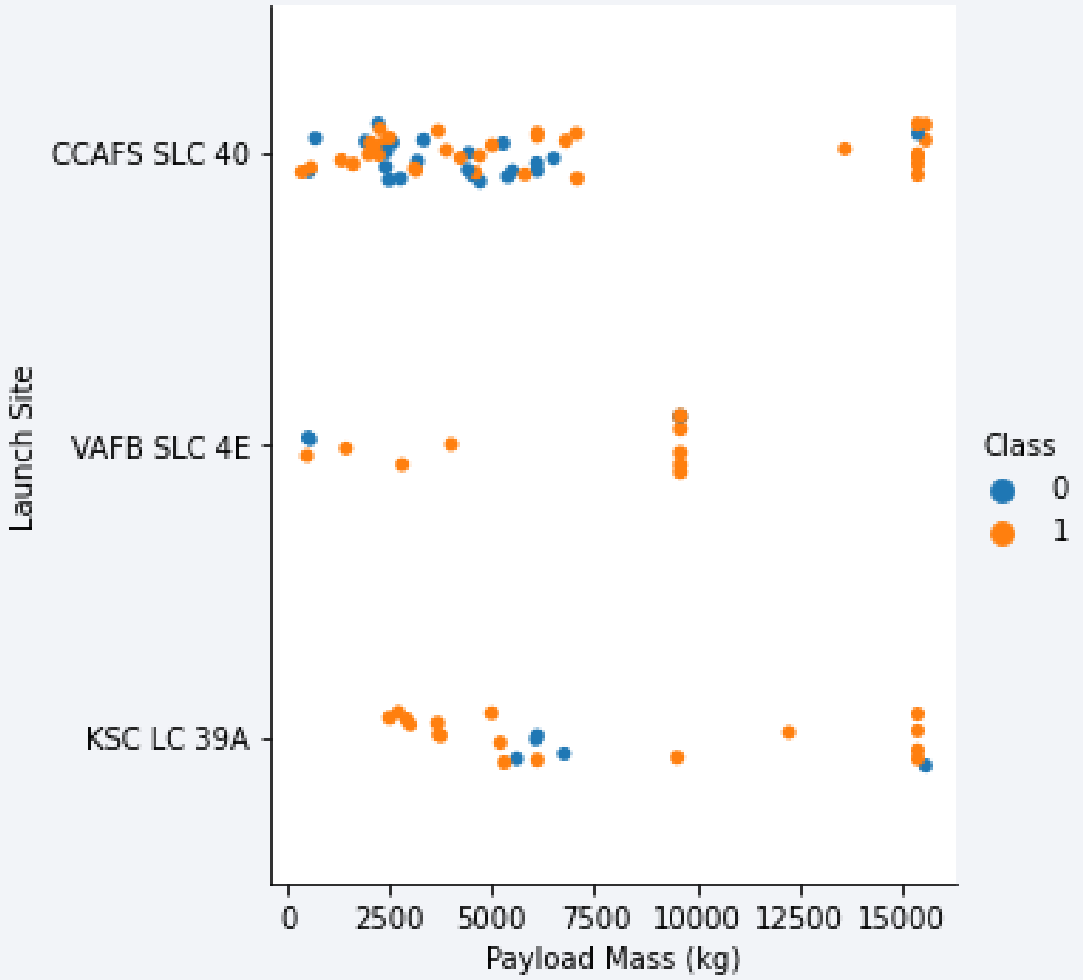
Flight Number vs. Launch Site

- From this scatter plot, we can see that as the flight number increases for each launch site, the success rate also increase.



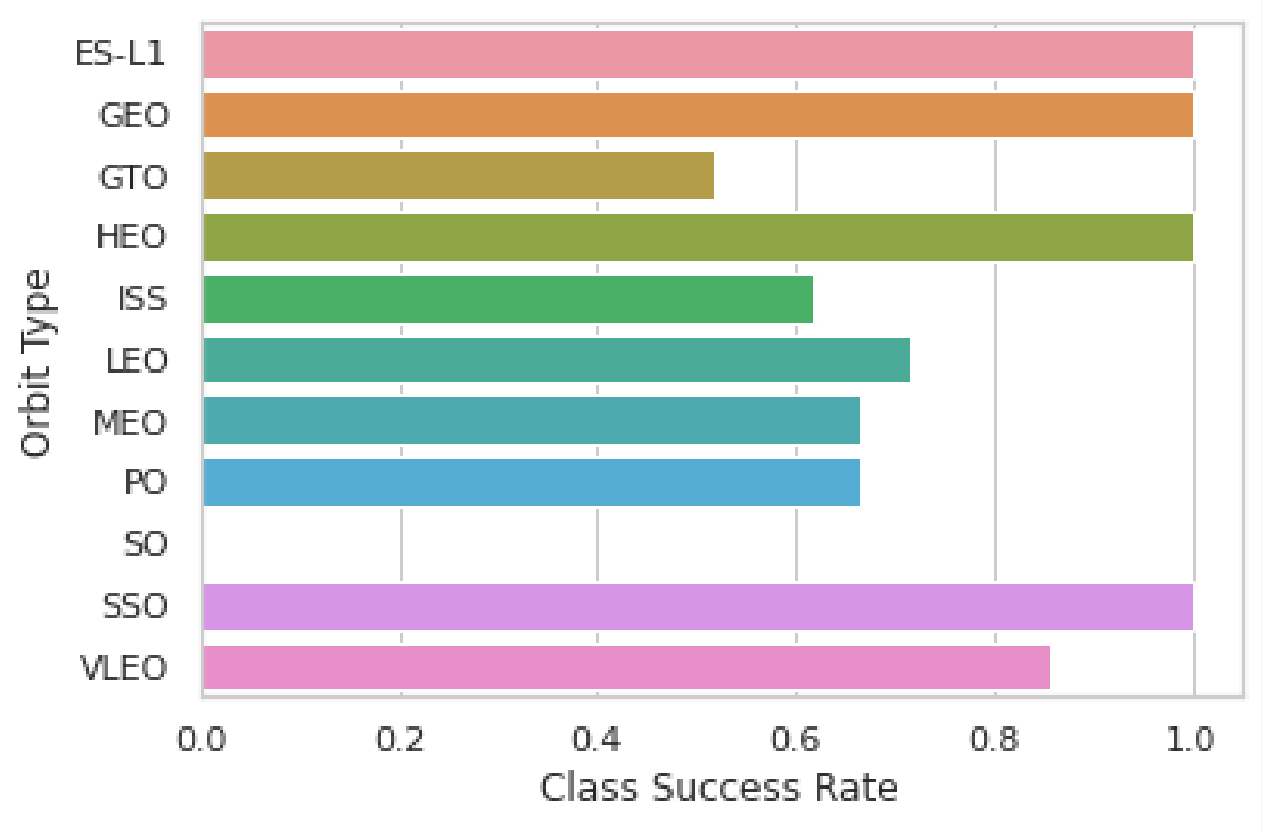
Payload vs. Launch Site

- From the scatter plot, we see that the pay load mass has similar outcomes at all 3 sites.
- When the pay load mass is between 5000 and 7000, we see a lower success rate at both CCAFS SCL 40 and KSC LC 39A



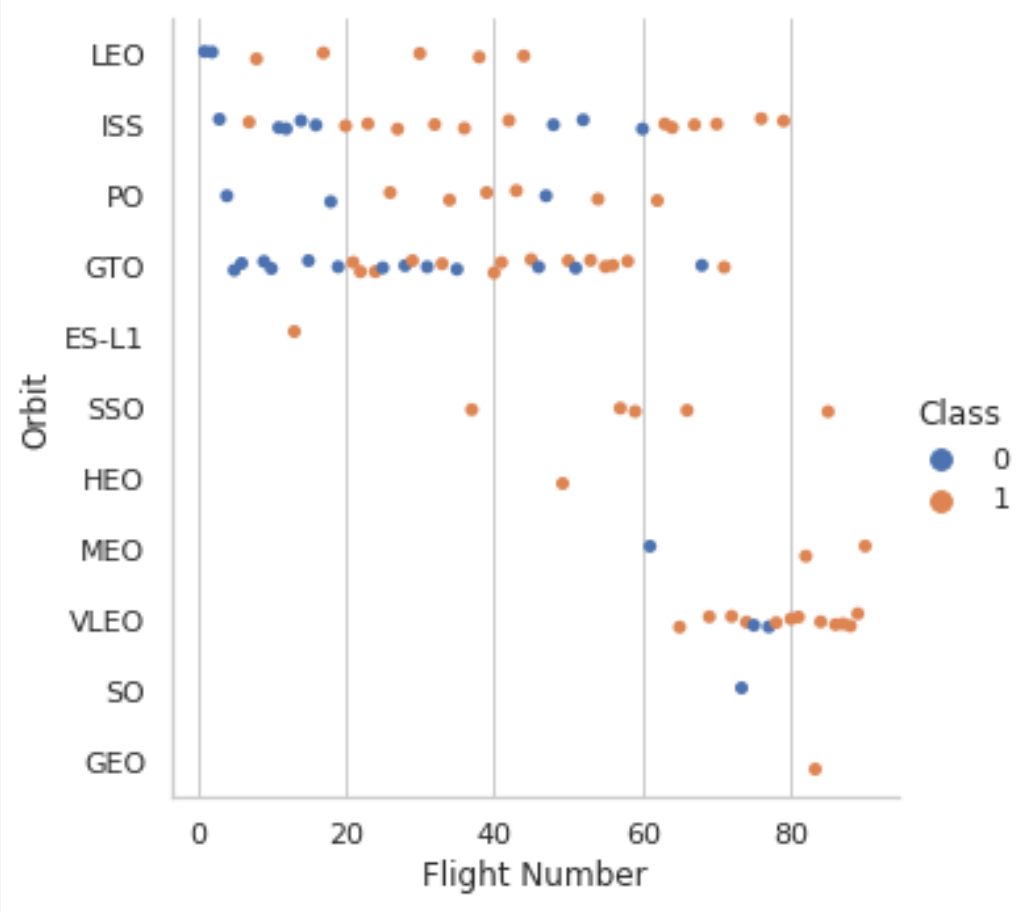
Success Rate vs. Orbit Type

- Orbit type shows a strong correlation to success rate.
- With multiple orbits reaching 100% and one orbit type with 0%.
- The remaining orbit types are also close to or above 50%.



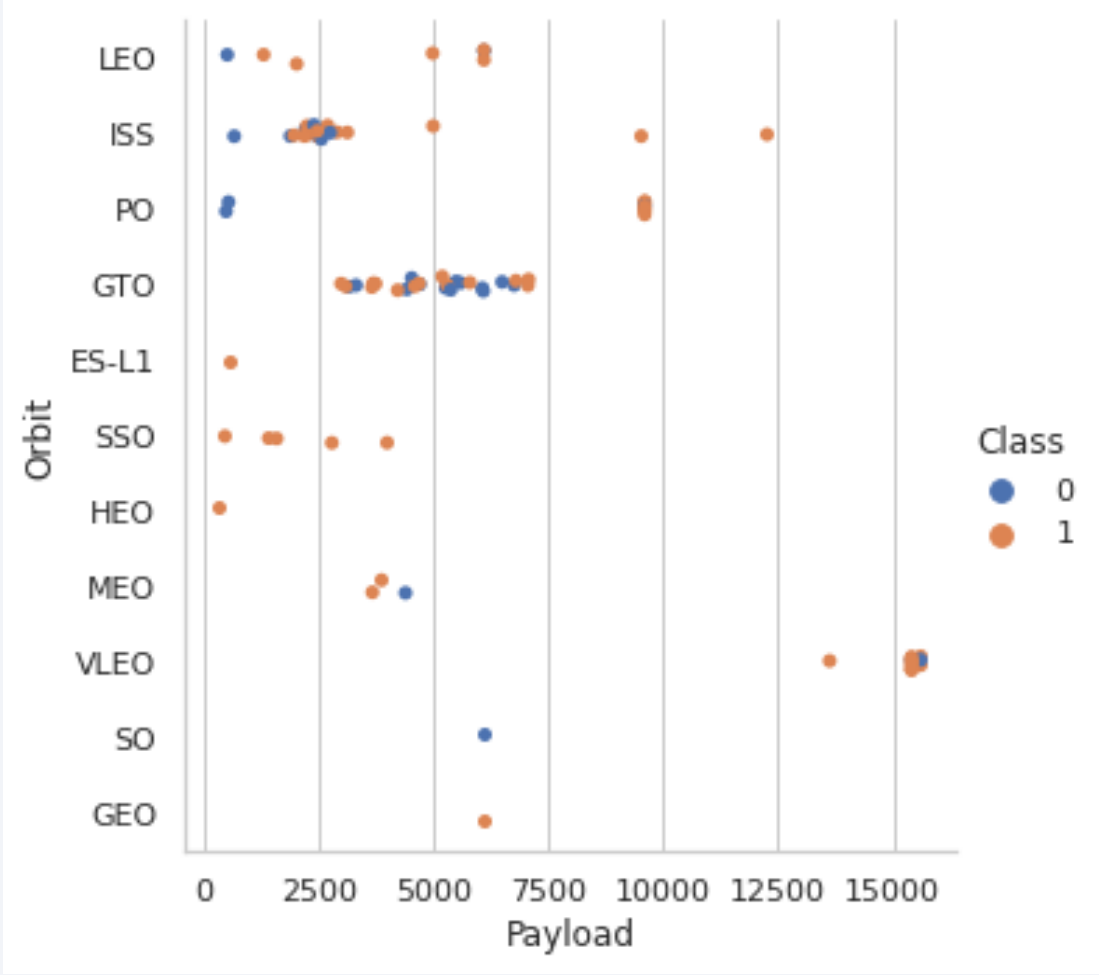
Flight Number vs. Orbit Type

- This scatter plot continues to demonstrate that as flight numbers increase, success rate also increases.
- With the LEO, we see that the early flights were unsuccessful, but all later flights were successful.
- However with orbit GTO, the flight number does not seem to impact the success rate as much.



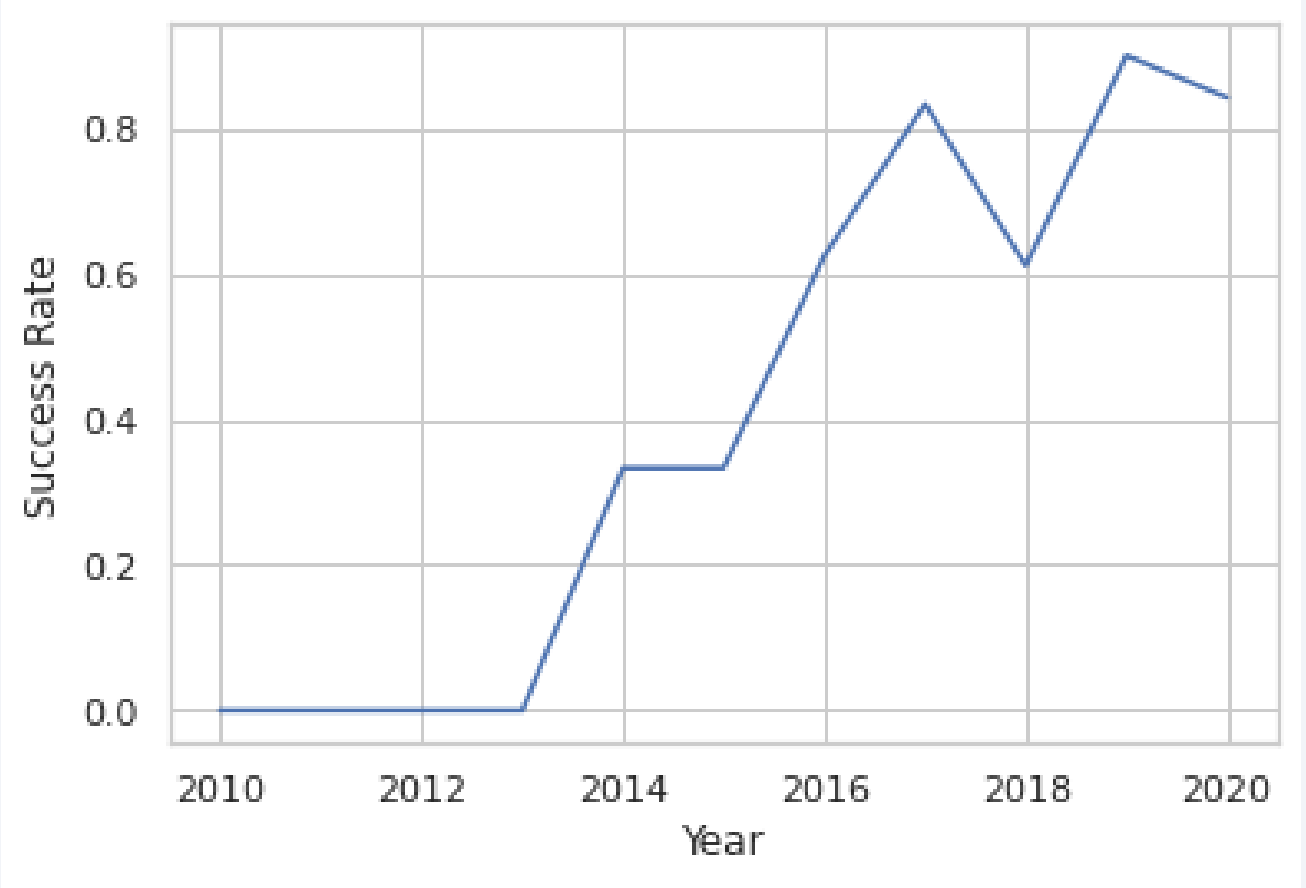
Payload vs. Orbit Type

- This graph demonstrates that the payload mass does not fluctuate much within an orbit type. Meaning most types stay within a designated payload size.
- We can also see where some orbits types are more likely to fail based on the payload. LEO, PO, and ISS show lighter payloads are more often to fail.



Launch Success Yearly Trend

- From the line chart, we can see that the success rate increased over the years.
- However there were two instances where the success rate decreased, 2018 and 2020.
- In general, we can expect the success rate to stabilize around the 80% area.



All Launch Site Names

- Achieved by using distinct on the launch_site column, another option would be to group by launch_site.
- 4 results unique results returned

Task 1

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

```
]: %%sql SELECT DISTINCT(launch_site)
     FROM SPACEXDATASET
```

```
 * ibm_db_sa://rdt82430:***@ba99a9e6-d59e-4883-8fc0-d6a8c
Done.
```

15]:

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

Launch Site Names Begin with 'CCA'

- The results show that data is contained for SpaceX and NASA, as well as a variety of different payloads and booster_versions.
- We also see that some launches did not make a landing attempt.

Task 2

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

```
%%sql SELECT *
        FROM SPACEXDATASET
        WHERE launch_site like 'CCA%'
        LIMIT 5
```

```
 * ibm_db_sa://rdt82430:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

]:

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	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)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	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 mass launched by NASA, is 45,596 kg, based on the data source.

Task 3

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

```
%%sql SELECT SUM(payload_mass__kg_)
        FROM SPACEXDATASET
        where customer like 'NASA%(CRS)'
```

```
 * ibm_db_sa://rdt82430:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a0
Done.
```

]:

1
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 : 2,534 kg

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

```
%%sql SELECT AVG(payload_mass__kg_)
        FROM SPACEXDATASET
        where booster_version like 'F9 v1.1%'
```

```
 * ibm_db_sa://rdt82430:***@ba99a9e6-d59e-4883-8fc0-d
Done.
```

3]:

1
2534

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad : 12-22-2015

Task 5

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

Hint:Use min function

```
%%sql SELECT MIN(DATE) as Date
        FROM SPACEXDATASET
        where landing__outcome = 'Success (ground pad)'
```

```
 * ibm_db_sa://rdt82430:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1o
Done.
```

]:

DATE
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- 4 boosters were retrieved which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

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 SPACEXDATASET
        WHERE landing__outcome = 'Success (drone ship)'
            AND payload_mass__kg_ between 4000 and 6000
        GROUP BY booster_version
```

```
 * ibm_db_sa://rdt82430:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdo
Done.
```

]:

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 and failure mission outcomes is 61: 40

Task 7

List the total number of successful and failure mission outcomes

```
%%sql SELECT SUM(CASE WHEN landing__outcome like 'Success%' THEN 1 ELSE 0 END) as success,
      SUM(CASE WHEN landing__outcome not like 'Success%' THEN 1 ELSE 0 END) as failures
      FROM SPACEXDATASET
```

```
 * ibm_db_sa://rdt82430:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.data
Done.
```

1]:

success	failures
61	40

Boosters Carried Maximum Payload

- Several names retrieved for booster that have carried the maximum payload mass
- The results may be changed if we wanted to look at only successful outcomes.

Task 8

List the names of the booster_versions which have carried the maximu

```
%%sql SELECT booster_version
      FROM spacexdataset
      WHERE payload_mass__kg_ =
      (SELECT max(payload_mass__kg_) FROM SPACEXDATASET)
      GROUP BY BOOSTER_VERSION
```

```
 * ibm_db_sa://rdt82430:***@ba99a9e6-d59e-4883-8fc0-d6a8c
Done.
```

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

- Booster versions and launch site names for the year 2015, that have failed landing_outcomes in drone ship.

```
%%sql SELECT booster_version, launch_site
        FROM SPACEXDATASET
        WHERE YEAR(DATE) = 2015
            AND landing__outcome = 'Failure (drone ship)'
        GROUP BY booster_version, launch_site
```

```
 * ibm_db_sa://rdt82430:***@ba99a9e6-d59e-4883-8fc0-d6a8
Done.
```

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

- The rank of 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
- The most common outcome for the provided data is "No Attempt".
- However, we can also see from the result set, that there are only
  2 outcomes clearly stated as
  successes and 2 clearly stated as failures.

```
%%sql SELECT landing__outcome,
             count(1) cnt_outcome,
             RANK() OVER(order by COUNT(1) DESC) rnk
        FROM SPACEXDATASET
        WHERE "DATE" BETWEEN '2010-06-04' and '2017-03-20'
        GROUP BY landing__outcome
```

```
 * ibm_db_sa://rdt82430:***@ba99a9e6-d59e-4883-8fc0-d6a8
Done.
```

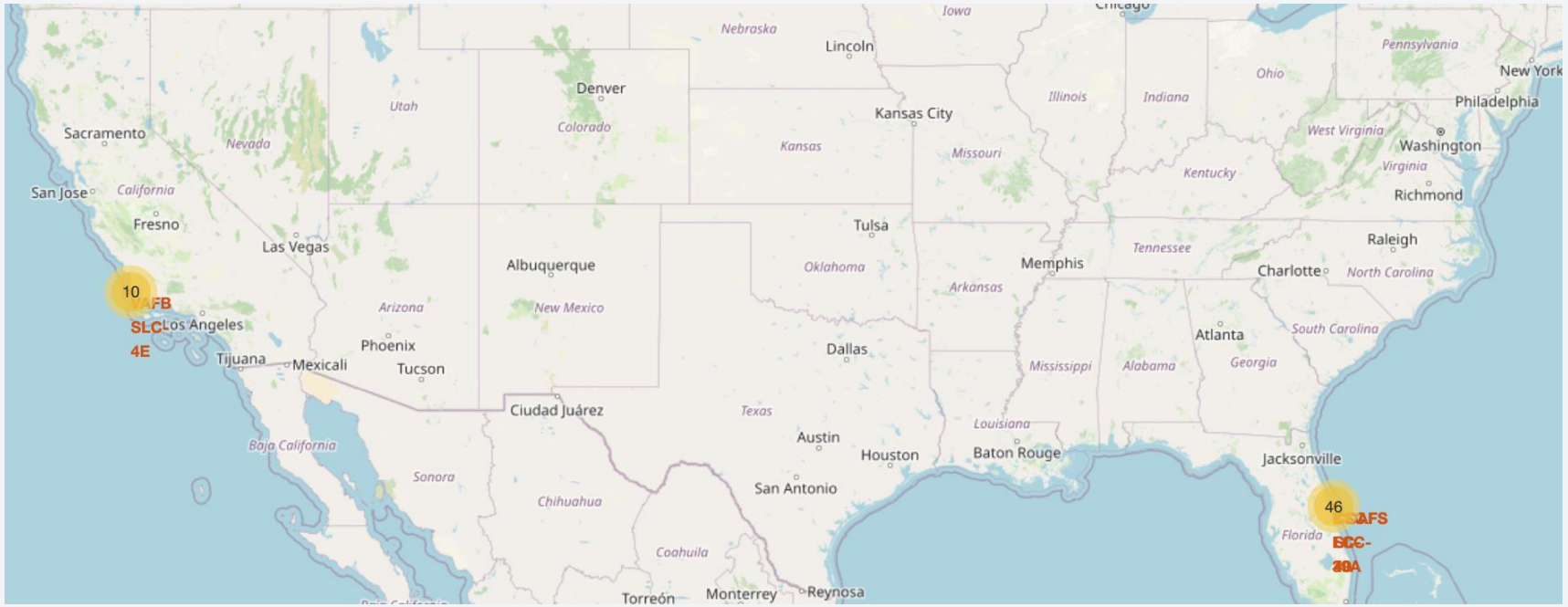
]:

landing__outcome	cnt_outcome	rnk
No attempt	10	1
Failure (drone ship)	5	2
Success (drone ship)	5	2
Controlled (ocean)	3	4
Success (ground pad)	3	4
Failure (parachute)	2	6
Uncontrolled (ocean)	2	6
Precluded (drone ship)	1	8


Section 3
Launch Sites
Proximities Analysis

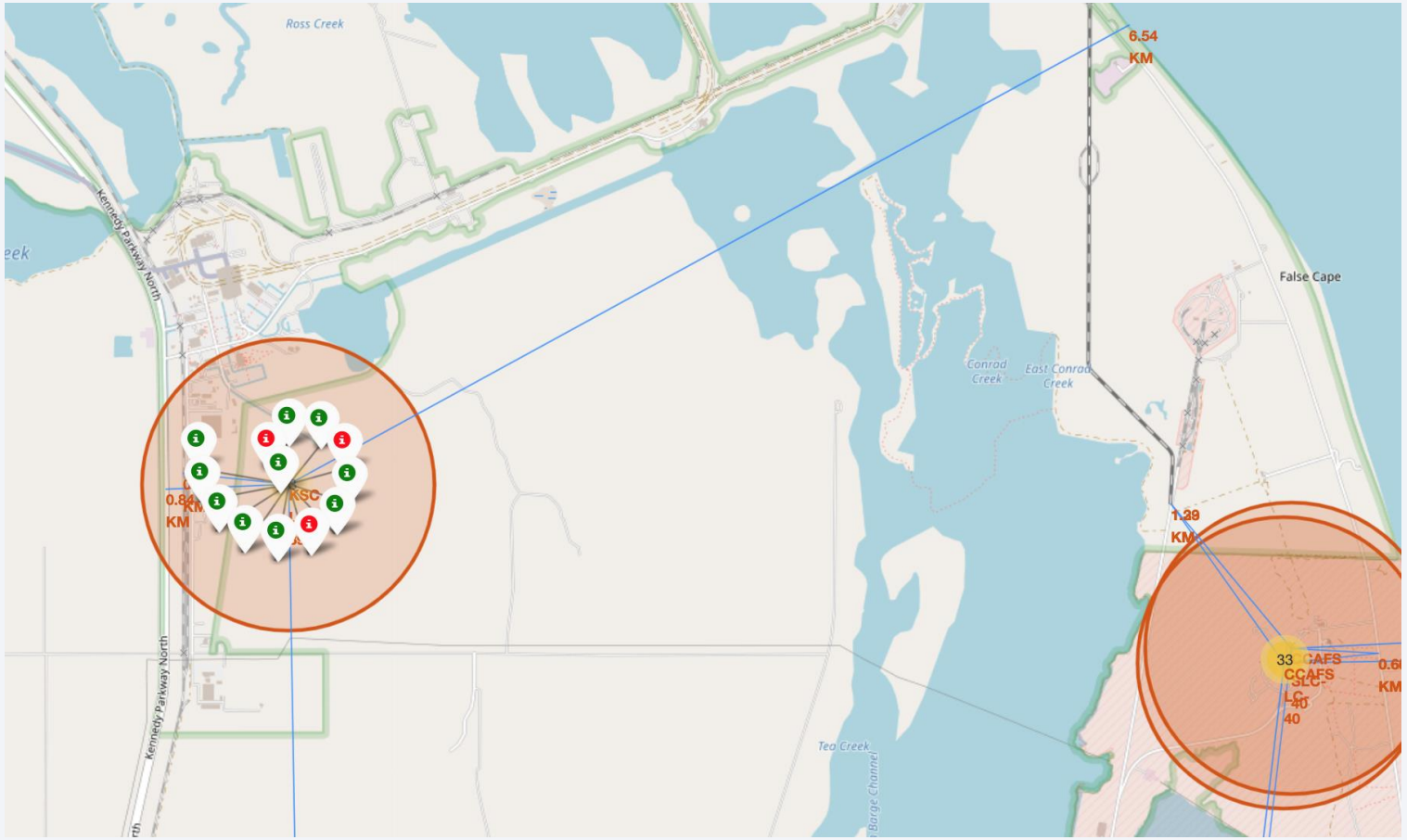
Launch Site Locations

- Out of the 4 launch sites identified in the data set, the zoomed out map appears to only show 2 locations.
- This is due to 3 locations being in close proximity to each other in the southern east coast region of Florida.



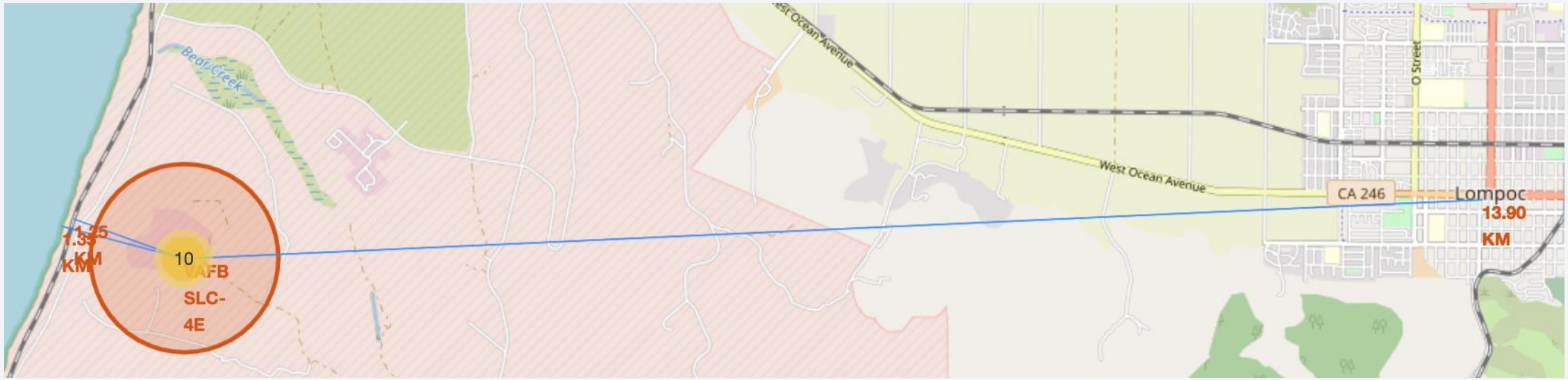
KSC LC-39a Success

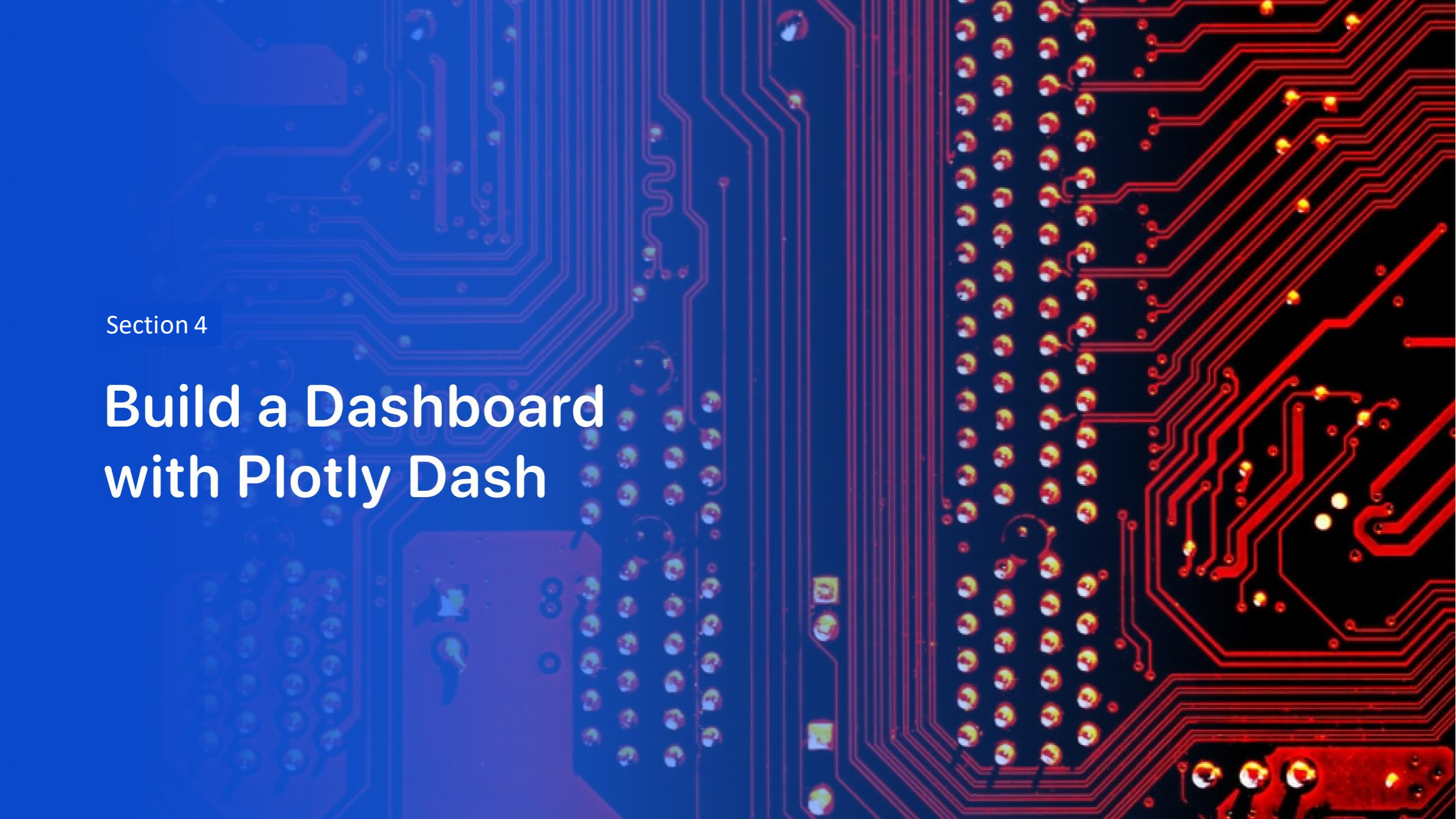
- The selected cluster is for the most successful launch site.
- Here we can clearly see the amount of successful launch recoveries vs failures.
- Each site produces a similar result when selected.



VAFB SLC-4E Site Distance

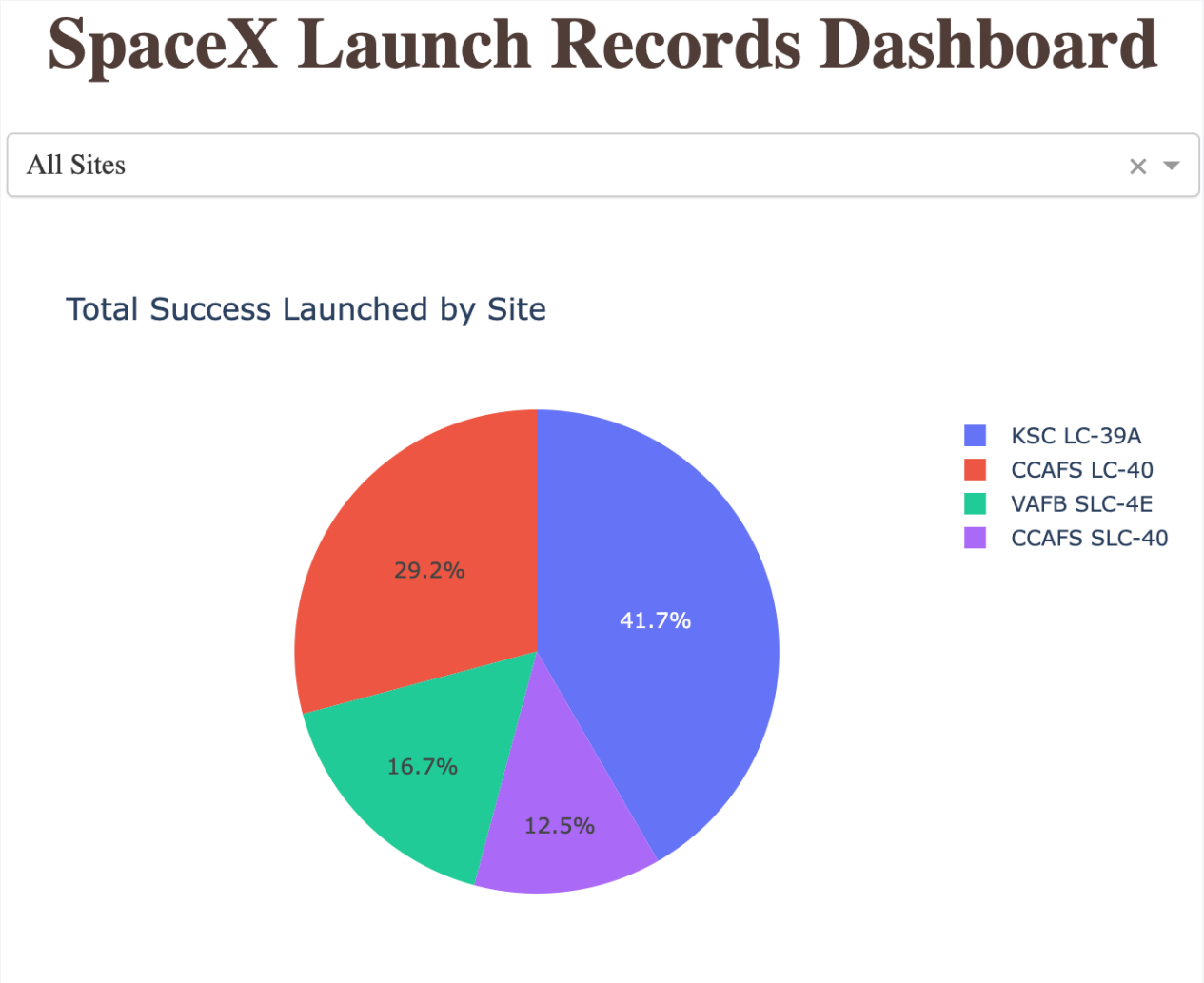
- This screen shot demonstrates how the distance is displayed for various location points.
- The railway is 1.25km away and the coastline is 1.35km
- VAFB has the farthest distance to a highway of 13.90km, but the closest distance to a city center, with 13.90km as well.




Section 4
Build a Dashboard with Plotly Dash

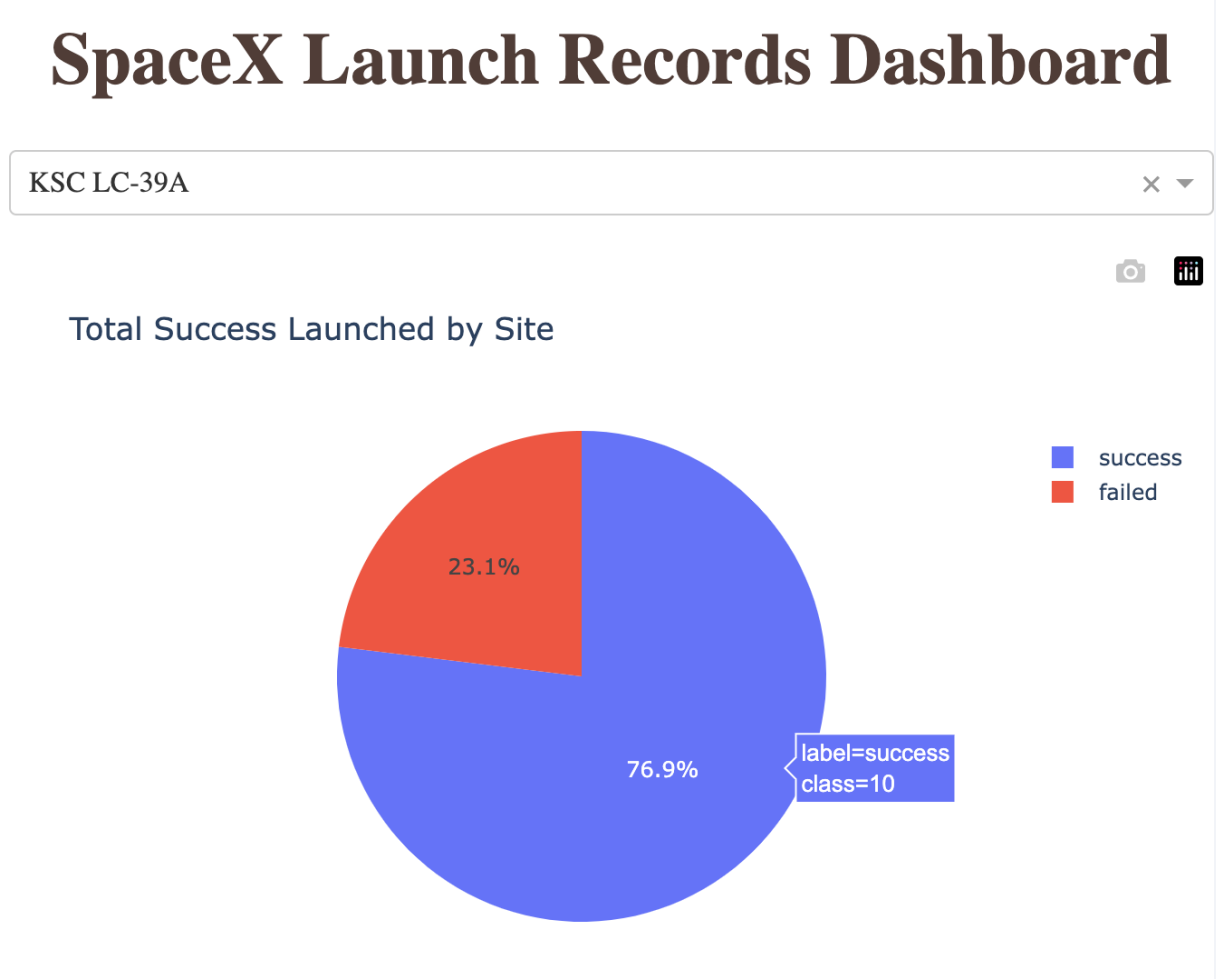
Launch Sites Success

- The pie chart figure shows that launch site KSC LC-39A is the largest contributor to the overall success of SpaceX Launches.
- The figure also shows that CCAFS SLC-40 is the launch site with the lowest successful recoveries.
- With this data we can look further into what types of missions each site encounters to get a better understanding of why there is such a difference.



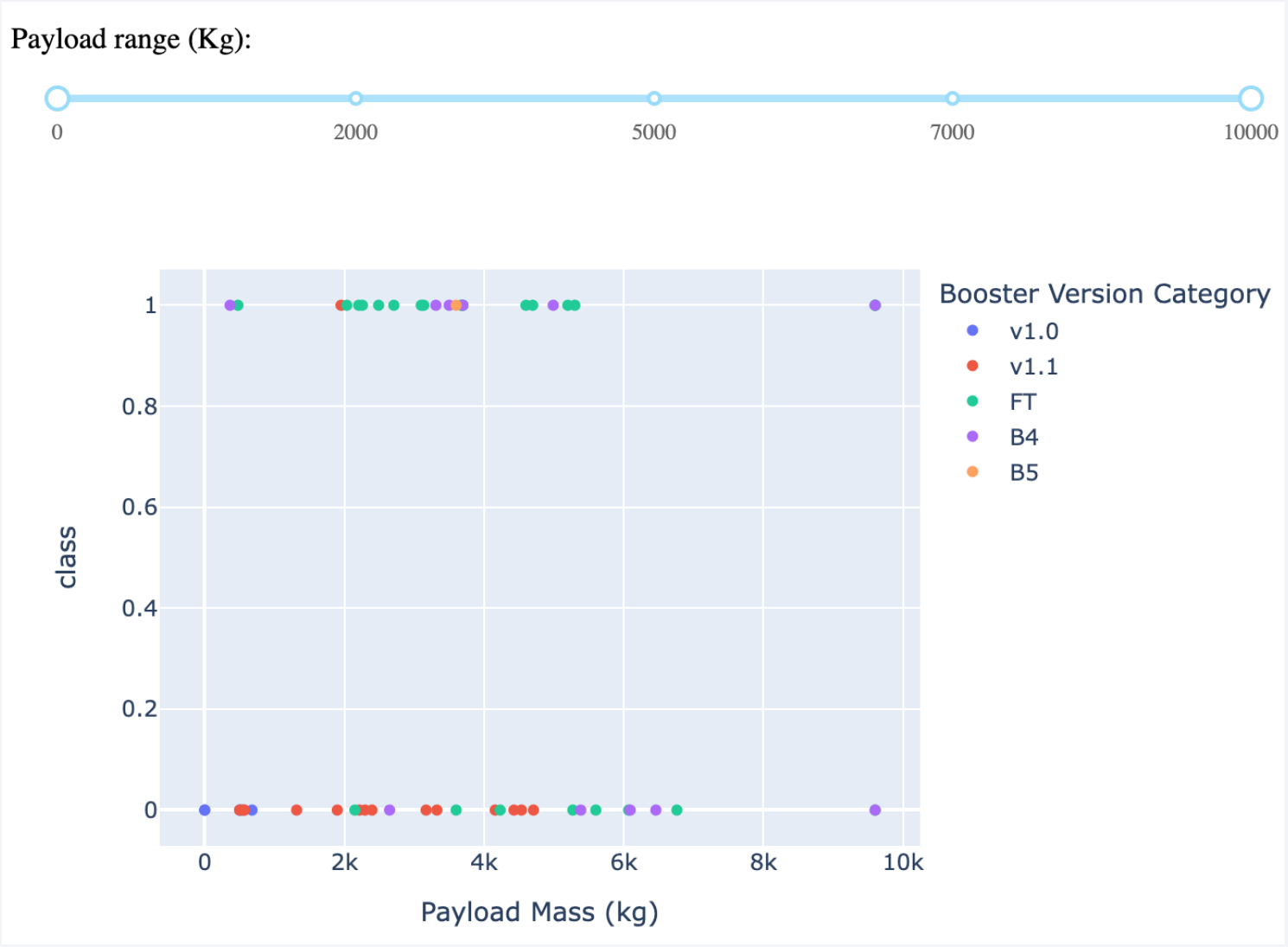
Launch Site KSC LC-39A Mission Success

- Looking at the highest contributing launch site, we're able to see the large success rate of the site.
- Hovering over a pie section provides the associated count.



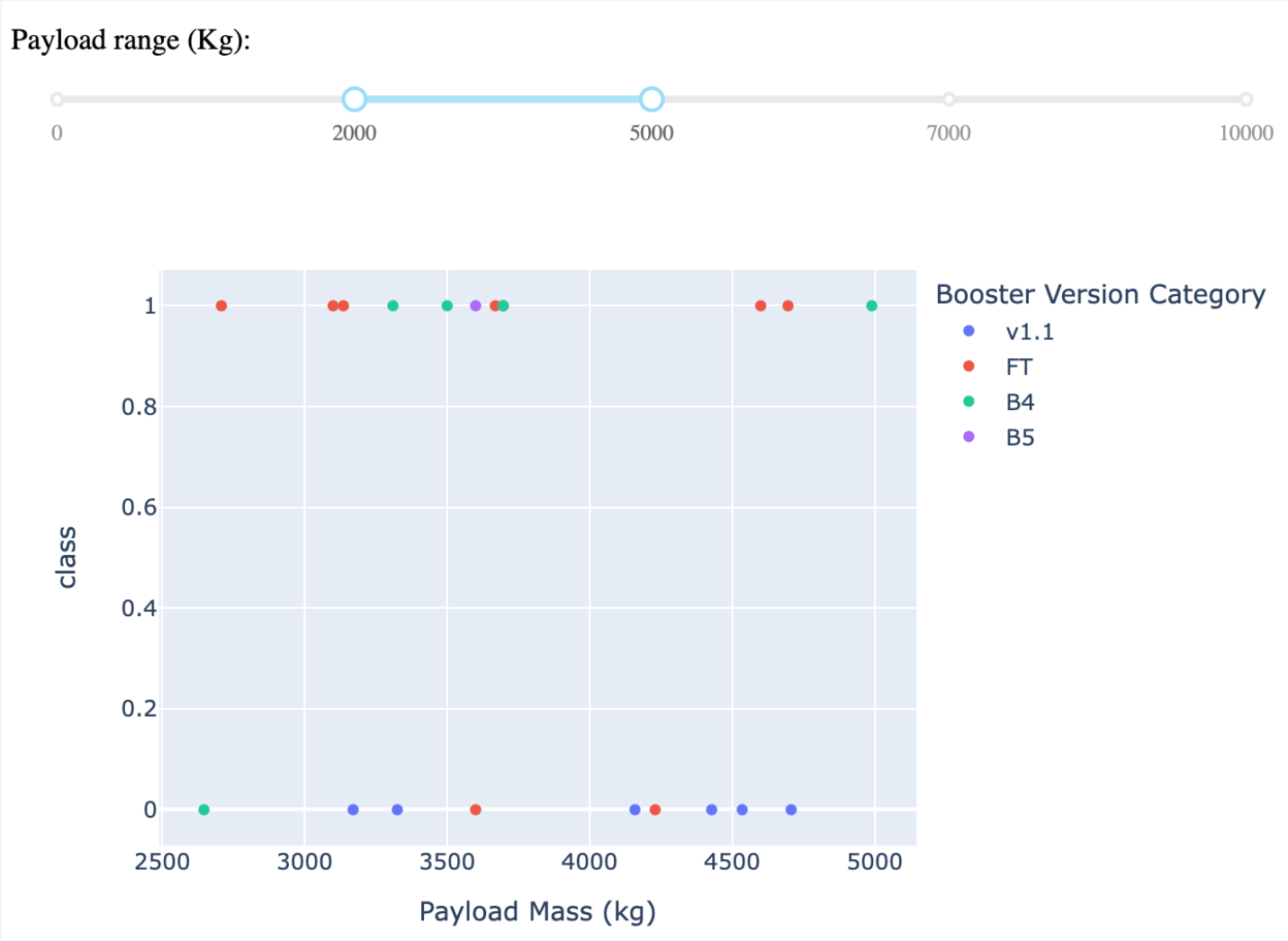
Payload vs. Launch Outcome Overall

- Data shown is for all launch sites.
- The scatter plot helps recognize the payload mass that has the highest success.
- The color coded booster version shows that the FT booster is also associated with successful recoveries.
- Booster v1.1 is also shows a high association with failed recoveries.



Payload vs. Launch Outcome Overall

- Data shown is for all launch sites with the range slider adjusted to highlight the most successful recoveries.
- Looking at the selected range of 2000Kg to 5000Kg, we can see that the amount of successes outweighs the failures in this range.

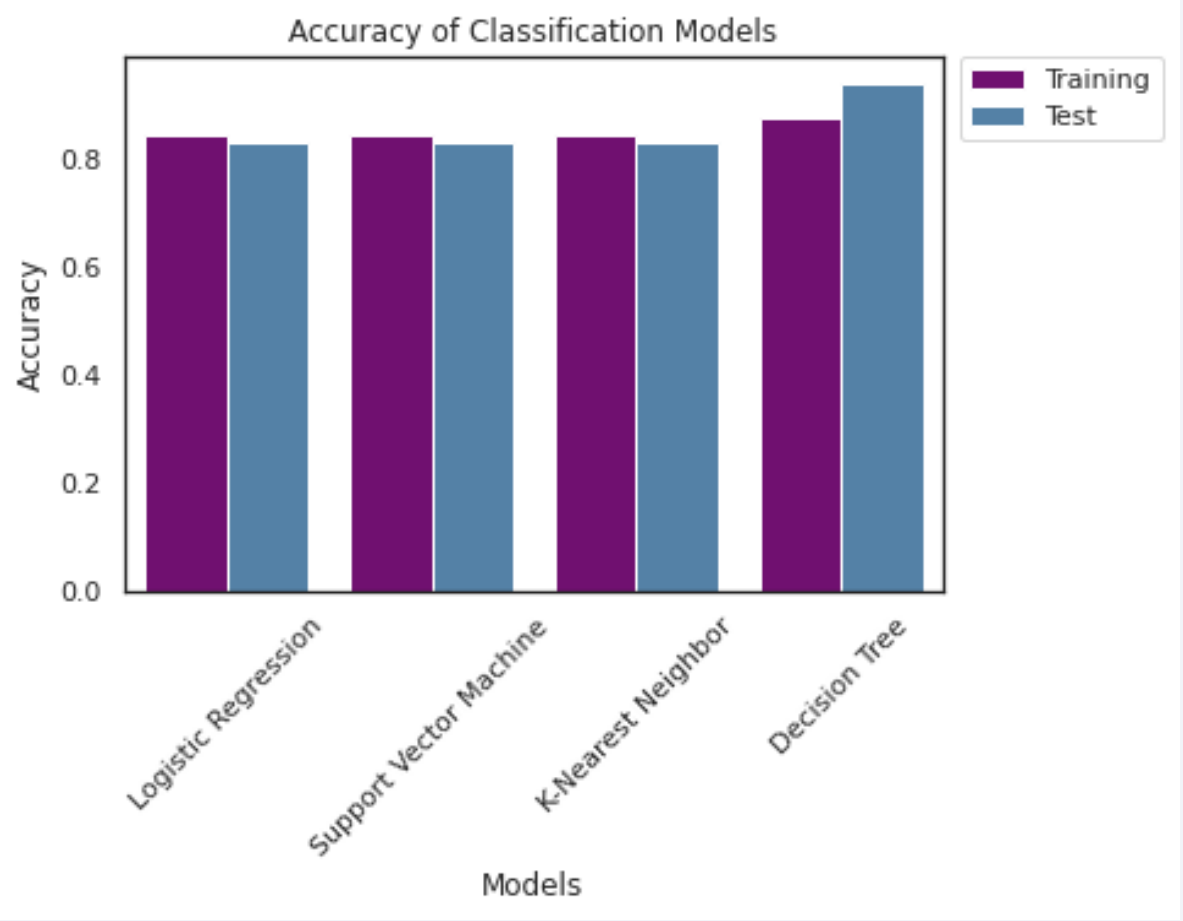


Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The decision tree has the highest classification accuracy, compared to the other 3 models tested.
- However, I discovered that this outcome is highly volatile. The accuracy of the decision tree can vary on each run of the GridSearchCV, and also with each train/test split.
- After discovering the high accuracy once, I only needed to continue running the cells to find it again. Each execution produced different hyperparameters as well.



Confusion Matrix – Best performance

- The Confusion matrix to the right is the outcome form the decision tree model.
- With close to 100% accuracy, only one false positive was detected.
- In all other models, false positives are the commonality.

Confusion Matrixes – All Others

- The Confusion matrix to the right is the outcome form the logistic regression, k-nearest neighbor, and support vector machine models.
- The matrix shows that the largest concern is with false positives.

Conclusions

- Due to the highly volatile outcome of the decision tree, reliability is low so monitoring accuracy will be important when using this model.
- Other models may be improved if we spend more time looking at feature engineering
  - One possibility would be to add weights to the features so more discerning features are highlighted in the models.
- Overall, predicting the outcome of the successful recovery of the Stage 1 component on an upcoming launch has an accuracy between 83% to 94%.
  - With a predictive outcome of False Positives being the most likely scenario.

Thank you!