IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Adam Cunningham
October 10, 2025

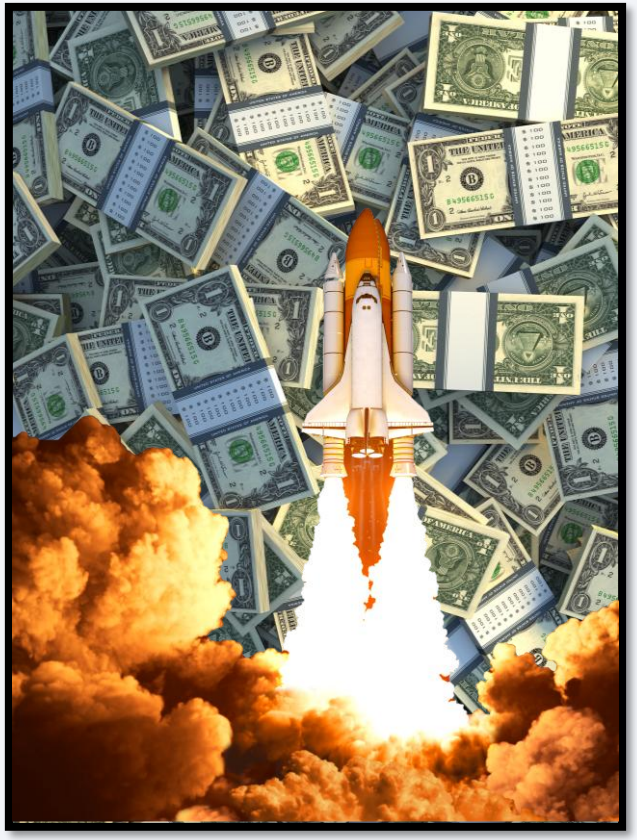
Table of Contents

Executive Summary

- We performed Data Collection using API and webscraping, performed Data Wrangling and EDA using SQL Lite and Seaborn, represented relationships in the data through Folium interactive maps, Plotly interactive dashboards, and the Seaborn library for static graphs. Finally, we used the data and EDA to build and evaluate 4 classification models for predicting booster launch successful recoveries.
- Our results showed us that flight number, orbit, launch site location, payload mass, and booster version can all be used to effectively predict with an 83.34% accuracy whether the recovery of the booster will be successful.

Introduction

- The commercial space age is here, companies are making space travel affordable for everyone. Virgin Galactic is providing suborbital spaceflights. Rocket Lab is a small satellite provider. Blue Origin manufactures sub-orbital and orbital reusable rockets. Perhaps the most successful is SpaceX.
- One reason SpaceX is so successful is the rocket launches are relatively inexpensive. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch.
- Space Y that would like to compete with SpaceX founded by Billionaire industrialist Allon Musk. Your job is to determine the price of each launch. You will do this by gathering information about Space X and creating dashboards for your team. You will also determine if SpaceX will reuse the first stage. Instead of using rocket science to determine if the first stage will land successfully, you will train a machine learning model and use public information to predict if SpaceX will reuse the first stage.



Section 1

Methodology

Methodology

Data collection methodology:

Past SpaceX launch data was secured via API requests to SpaceX REST API as well as web scraping related Wiki pages using BeautifulSoup.

Perform data wrangling

Data was processed using SQL Lite to sample data, review attributes and relationships, and transform and simplify the target variable.

Perform exploratory data analysis (EDA) using visualization and SQL

Using SQL Lite and the Seaborn Python library, we identified key features, their relationships with one another, and plotted them visually to help identify patterns.

Perform interactive visual analytics using Folium and Plotly Dash

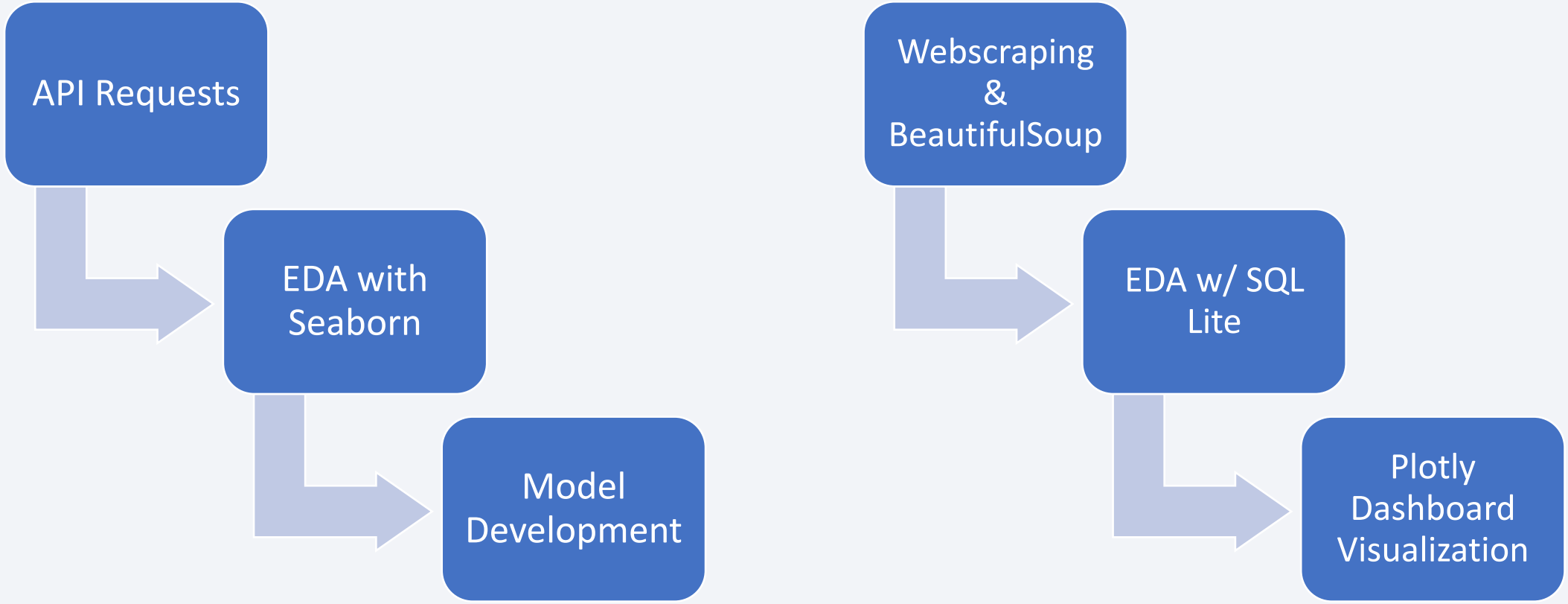
Using Folium, we explored success and failure rates by site visually, and built a dashboard for visualizing relationships interactively through pie charts and scatterplots.

Perform predictive analysis using classification models

We built, tuned, and evaluated 4 different predictive classification models to aid in predicting the success or failure of future launches.

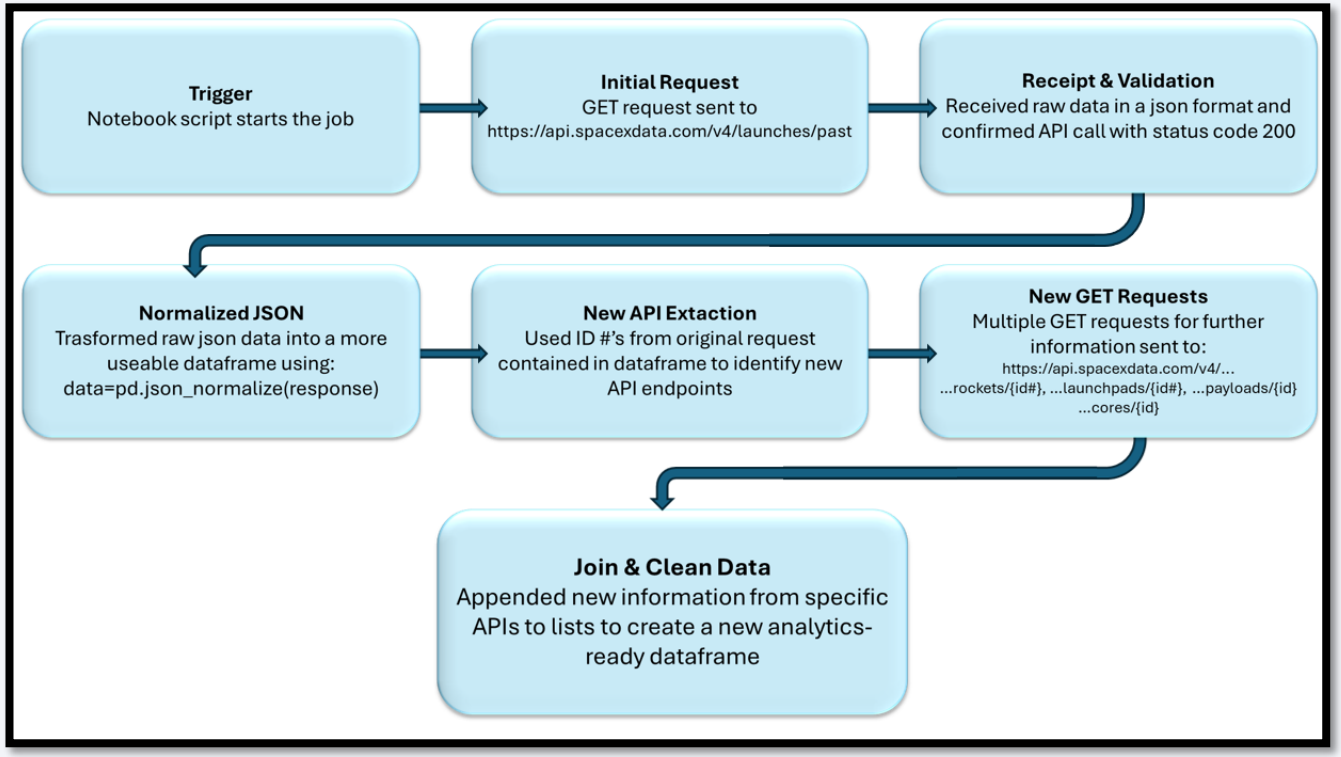
Data Collection

We took a two-pronged approach to data collection through both API's GET requests and Webscraping. Each data set was then passed along separate routes to aid in predicting and visualization.



Data Collection – SpaceX API

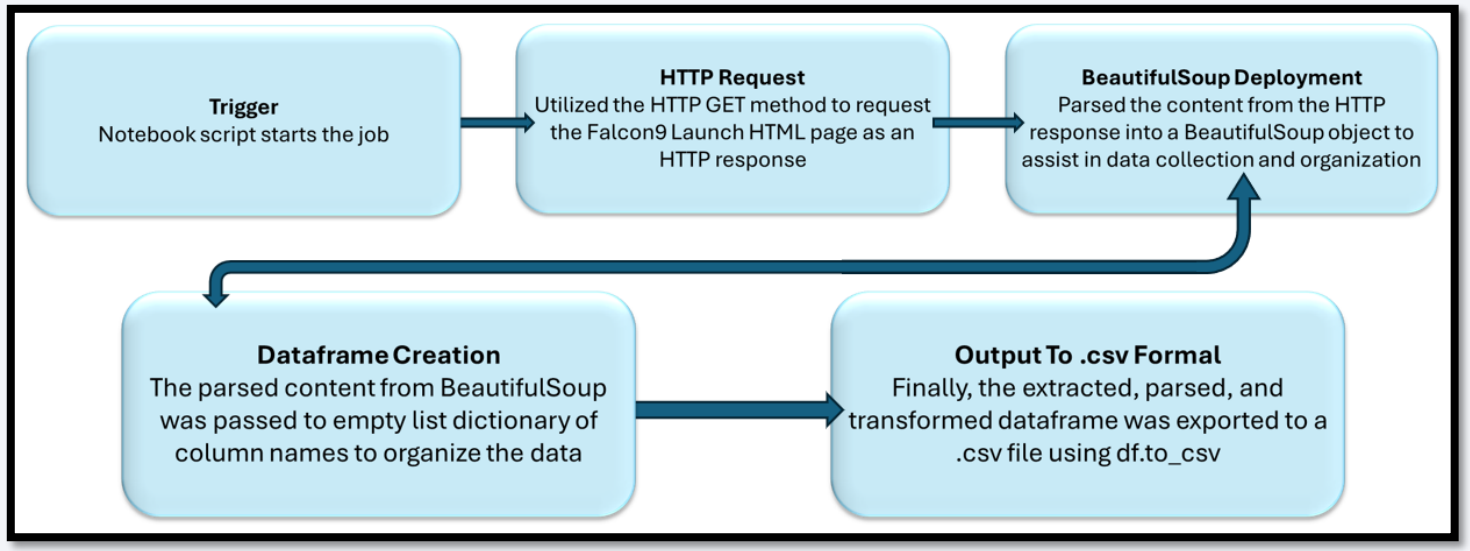
- Initial raw launch data was acquired via GET request sent to a high-level SpaceX REST API. ID numbers in the original data was in turn used to identify specific API endpoints to extract further, more detailed information on launch data which was then filtered to remove Falcon1 launch data, replaced missing data in the PayloadMass column with the mean PayloadMass, and organized into an analytics-ready dataframe for model predictions.



API Notebook GitHub link: [IBM-Data-Science-Capstone-Project/jupyter-labs-spacex-data-collection-api.ipynb at main · agcunning25-byte/IBM-Data-Science-Capstone-Project]

Data Collection - Scraping

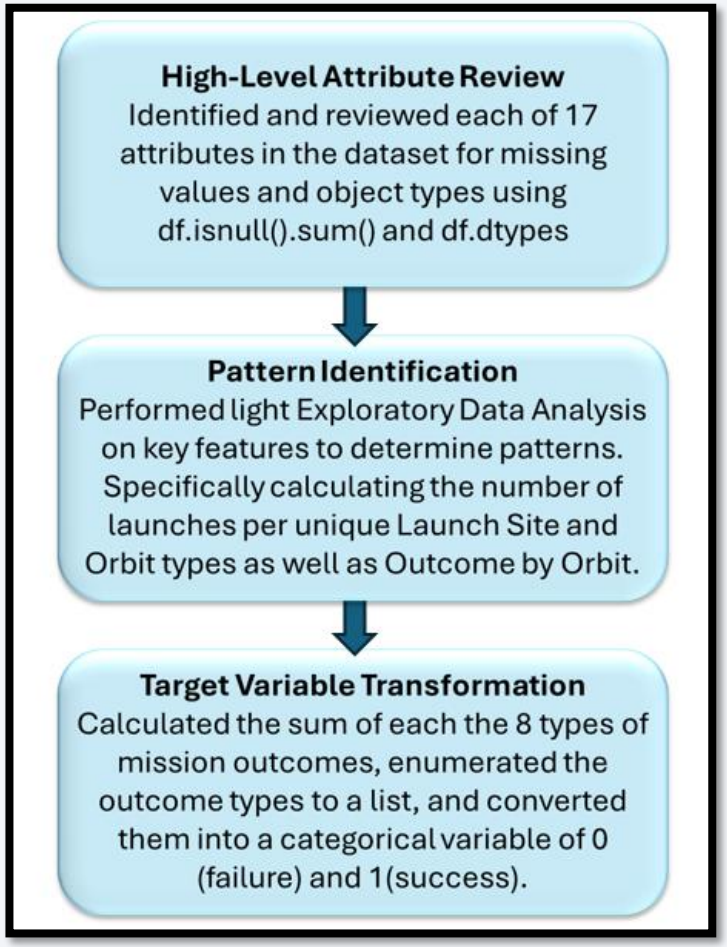
- A secondary method of data collection was used via webscraping Wiki pages using BeautifulSoup. HTML tables were viewed and parsed to capture and transform the data into clean dataframe ready for analysis with SQL.



Webscraping Notebook GitHub link: [IBM-Data-Science-Capstone-Project/jupyter-labs-webscraping.ipynb at main · agcunning25-byte/IBM-Data-Science-Capstone-Project]

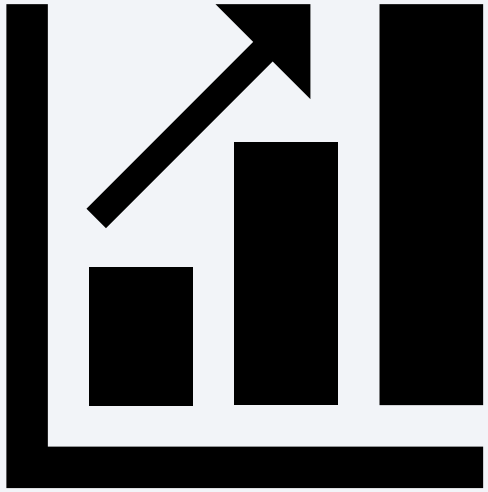
Data Wrangling

- Utilizing domain knowledge coupled with light EDA techniques, we identified key, high-level patterns for important features such as the number of Launch Sites, Orbits, and Outcomes. We then transformed and simplified our target variable, Outcomes, into a categorical variable describing the successful landing of the booster or failure to land the booster.



Data Wrangling Notebook GitHub link: [IBM-Data-Science-Capstone-Project/labs-jupyter-spacex-Data wrangling.ipynb at main · agcunning25-byte/IBM-Data-Science-Capstone-Project]

EDA with Data Visualization



- We plotted a total of 7 graphs to help in visualizing the data.
- Flight # v. Payload showed slight positive correlation
- Flight # v. Launch Site showed activity differences
- Payload v. Launch Site showed the workhorses
- Success v. Orbit displayed some are more successful than others
- Flight # v. Orbit type showed some orbits were not attempted until later flights
- Payload v. Orbit showed some orbits to be more successful than others
- Success rate over Year showed a positive trend in success rates over time

EDA Notebook GitHub link: [IBM-Data-Science-Capstone-Project/edadataviz.ipynb at main · agcunning25-byte/IBM-Data-Science-Capstone-Project]

EDA with SQL

Utilizing SQL Lite in a jupyter notebook, we performed various queries to extact the following information:

- ✓ Extracted the unique names of the 4 launch sites
- ✓ Displayed 5 launch records from launch sites beginning with “CCA”
- ✓ Determined the total payload mass in kg from boosters launched by NASA
- ✓ Calculated the average payload mass carried by F9 v1.1 boosters
- ✓ Acquired the date of the first successful ground pad landing was achieved
- ✓ Uncovered which boosters carrying a payload mass between 4000 and 6000 kg had successful drone ship landings
- ✓ Catalogued the total number of successful and failure mission outcomes
- ✓ Listed all booster versions that have carried the maximum payload
- ✓ Discovered what month, booster version, and launch site all drone ship failure landings were recorded in 2015
- ✓ Recorded and ranked the landing outcomes of all launches between June 4, 2010 and March 20, 2017.

SQL Lite Notebook GitHub link: IBM-Data-Science-Capstone-Project/jupyter-labs-eda-sql-coursera_sqllite (1).ipynb at main · agcunning25-byte/IBM-Data-Science-Capstone-Project

Build an Interactive Map with Folium

- Using Folium, we analyzed launch site locations geography on an interactive map to discover patterns visually.
- We plotted launch site locations using circle objects and learned they are all located on the coasts.
- We plotted each of the launches at the separate launch sites using marker clusters, color coded by success or failure, and learned that some sites are much more active than others, and some sites have higher success rates.
- Finally, we plotted the location of nearby features such as railways, highways, coasts, and cities to find similarities between launch sites.

Folium Notebook GitHub link: IBM-Data-Science-Capstone-Project/lab_jupyter_launch_site_location (1).ipynb at main · agcunning25-byte/IBM-Data-Science-Capstone-Project

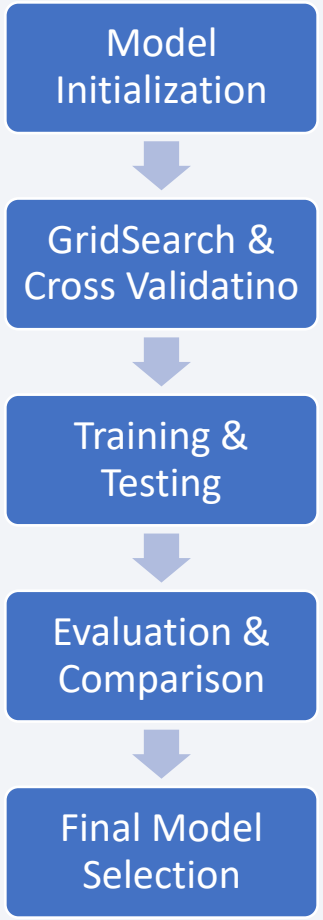
Build a Dashboard with Plotly Dash

- Using Plotly Dash, we created an interactive dashboard to help us in finding insights more easily than with static graphs.
- We included a dropdown list for launch sites as well as a payload range slider to further sample and aid in visualizing the data to easily draw insights.
- We a pie chart for % success and failure rate by selected site to draw inferences at a glance.
- We incorporated an interactive scatterplot tied to the payload range slider to assist in visualizing success rate by payload range.

Dashboard Notebook GitHub link: IBM-Data-Science-Capstone-Project/spacex_dash_app_py.ipynb at main · agcunning25-byte/IBM-Data-Science-Capstone-Project

Predictive Analysis (Classification)

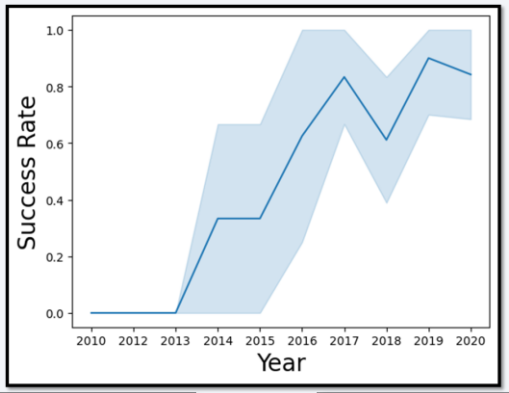
- We built 4 separate Classification models: Logistic Regression, Support Vector Machine, Decision Tree Classifier and K-Nearest Neighbor classifier.
- Each of the 4 models was tuned and cross-validated using GridSearchCV to find best parameters.
- Each model was trained on an isolated training set consisting of 80% of the processed data and tested on a test set consisting of 20% of isolated unseen data.
- Each of the 4 models was evaluated using accuracy score and their results plotted using a confusion matrix to identify the areas of model success and failure.
- The models test, training, and differences in accuracy scores was then entered into a data frame and compared visually with bar charts to aid in choosing the final model.

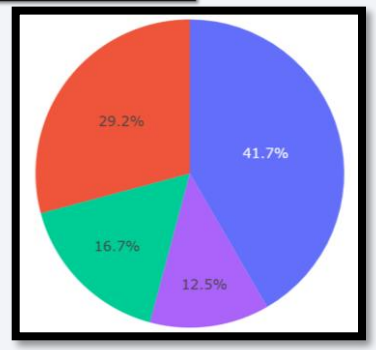


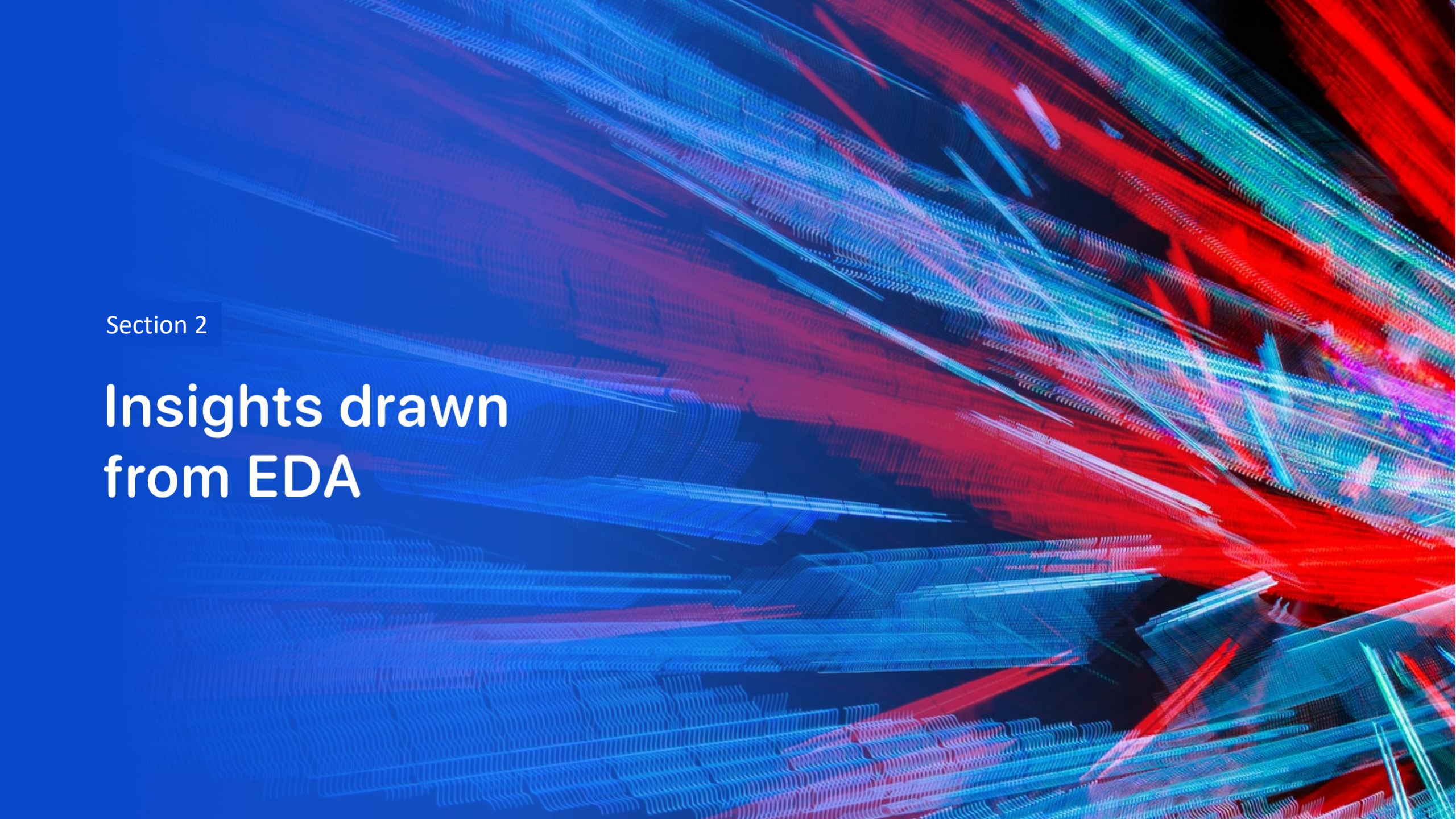
Machine Learning Notebook GitHub link: [IBM-Data-Science-Capstone-Project/SpaceX_Machine Learning Prediction_Part_5.ipynb at main · agcunning25-byte/IBM-Data-Science-Capstone-Project]

Results

- Through Exploratory Data Analysis we found that the success rate over time (by flight number) is positively correlated, that the selected orbit can help determine success rates as some orbits, like GEO, HEO, and SSO, have higher success rates, and we found that higher payloads may be positively located to success rates as well.
- Through our interactive dashboards and Folio map, we were able to determine that all launch sites are located near the coast, highways, and railways, but are intentionally distant from nearby cities. We also aided in visualizing that Kennedy Space Center has the highest percentage of all successful recoveries, CCAFS SLC-40 has the highest success rate of all sites, the FT booster version has the highest success rate, and payloads between 3,000 and 4,000 kg have the highest success rate.
- Of the 4 models we built, 3 had a prediction accuracy of 83.34% on the test set. The Decision Tree Classifier showed promise on the training set, but it was extremely overfit and had the lowest accuracy score on the test set. Due to the least margin between training and testing accuracy, the KNN model was chosen as our final model selection.

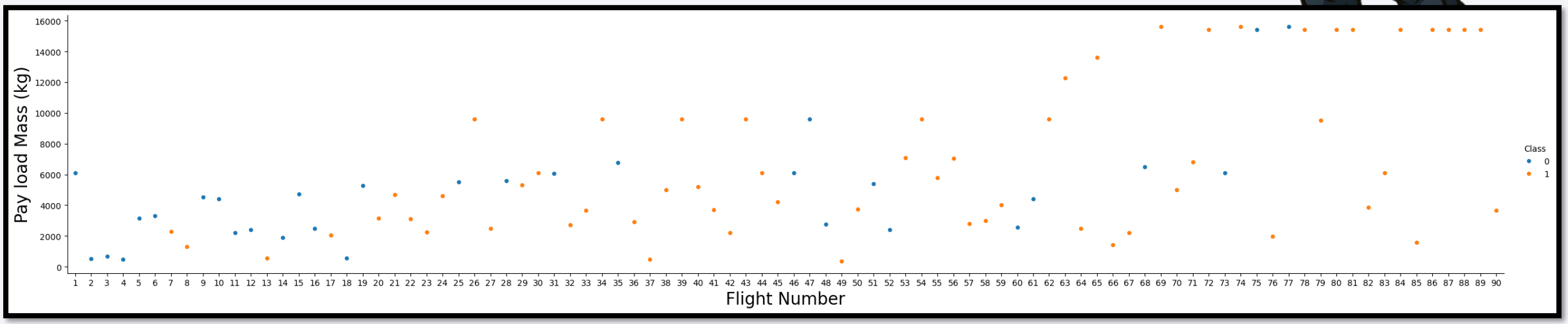





Section 2
Insights drawn from EDA

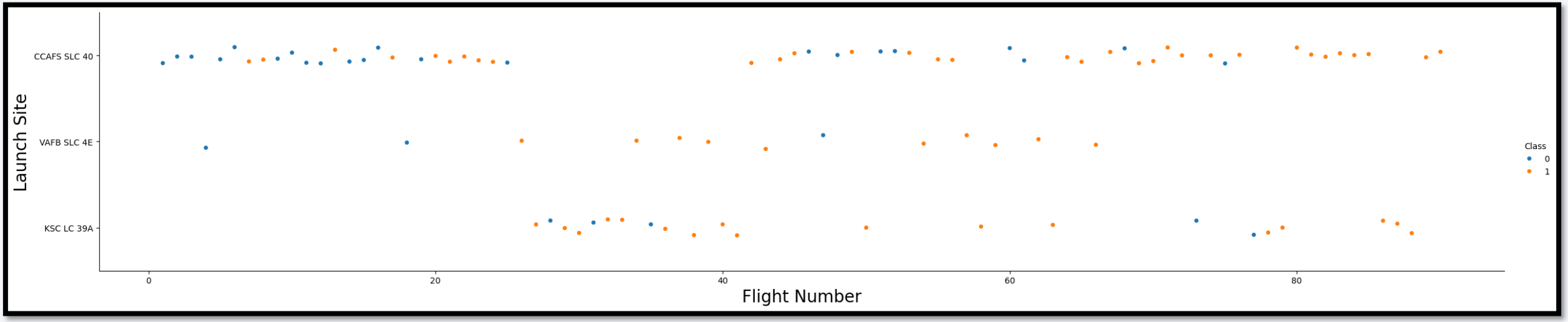
Flight Number vs. Payload

- This was our first plot showing the relationship between flight numbers and payload mass.
- As flight numbers increased, heavier payloads were introduced.
- There were no payload heavier than 10,000 kg in the first 25 flights, and **max payloads weren't attempted until around the 70th flight.**



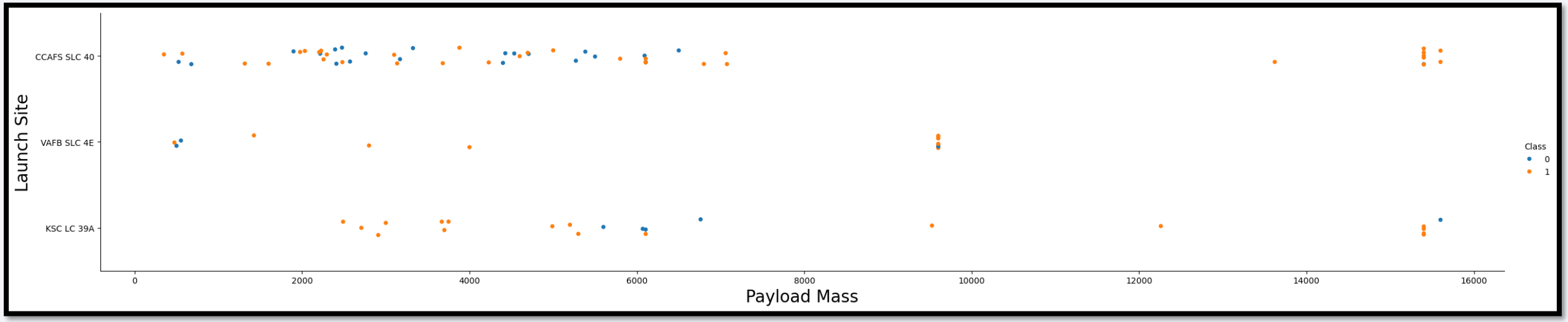
Flight Number vs. Launch Site

- As flight number increases, so does the likelihood of the outcome being successful.
- CCAFS SLC 40 has the most flights and VAFB SLC 4E has the fewest
- There were only 4 unsuccessful flights after around flight # 62.
- CCAFS had the most unsuccessful flights, followed by KSC and VAFB.



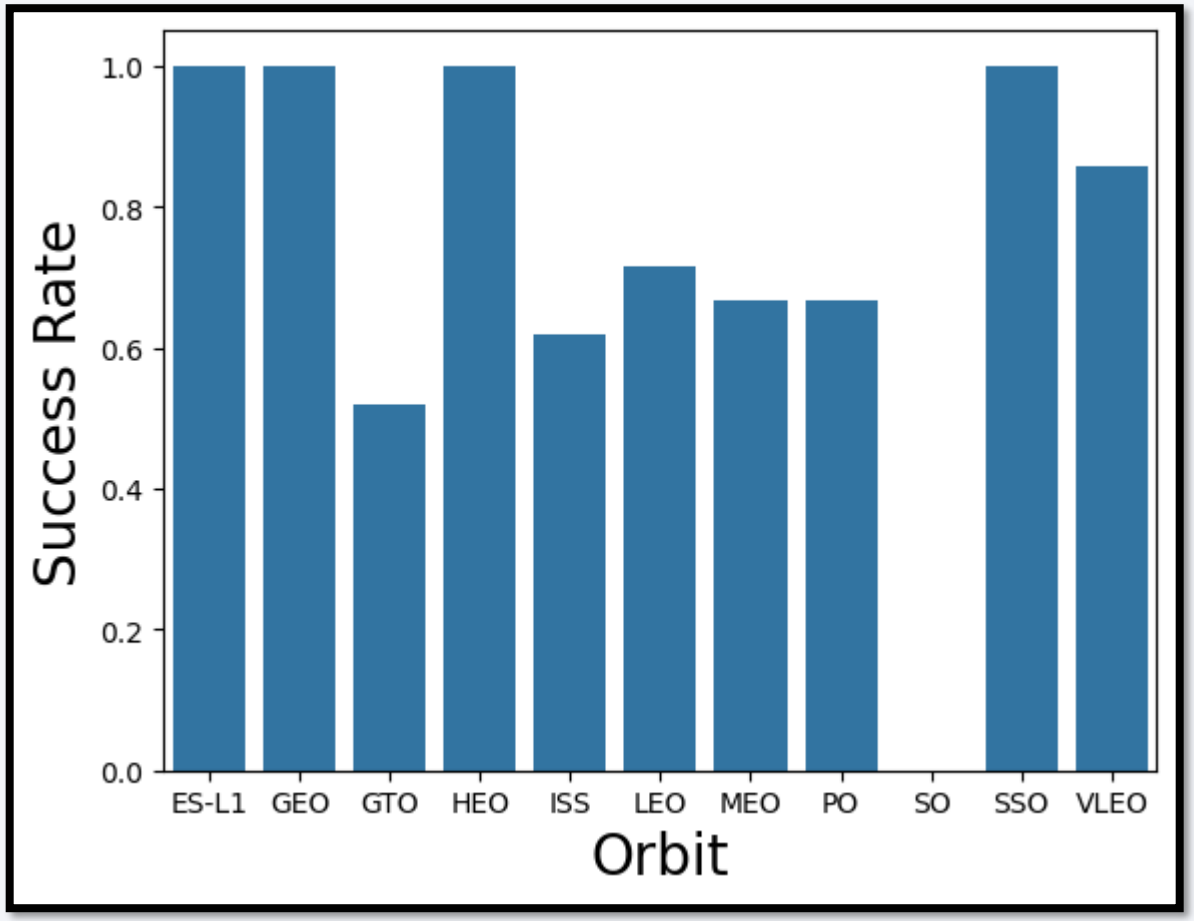
Payload vs. Launch Site

- CCAFS SLC-40 launched many more payloads under ~7,500 kg than heavier payloads.
- VAFB-SLC did not launch any payloads over 10,000 kg.
- KSC LC 39-A did not launch any payloads less than 2,000 kg.



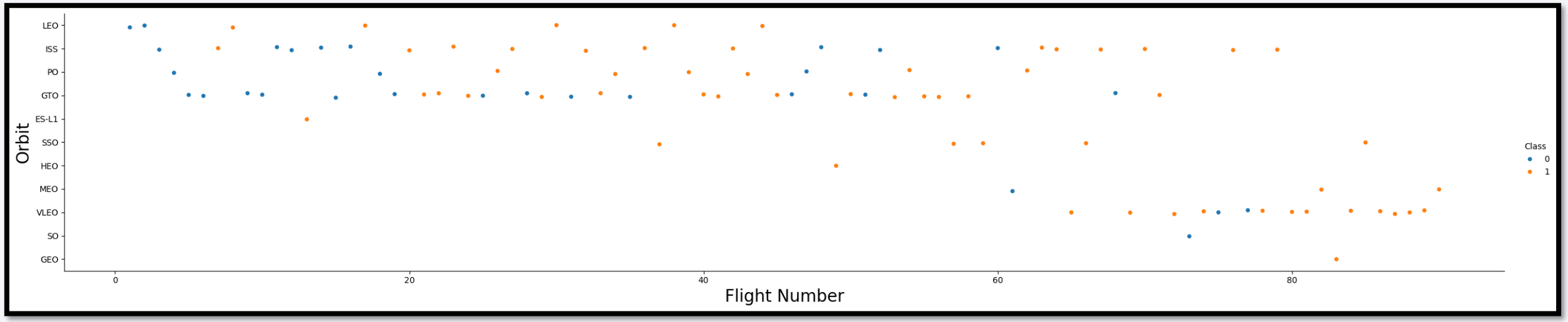
Success Rate vs. Orbit Type

- GTO orbit had the lowest success rate at around 55%.
- ES-L1, GEO, HEO, and SSO all had perfect success rates.



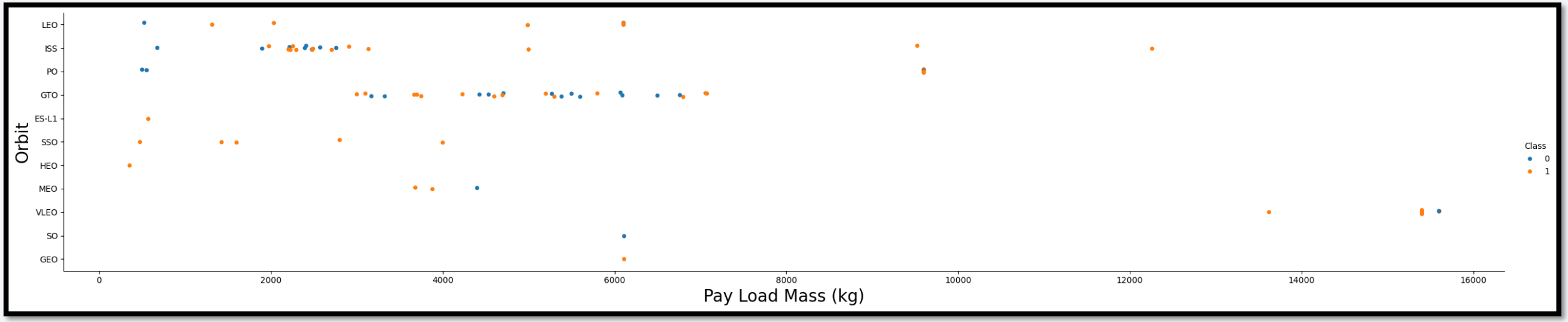
Flight Number vs. Orbit Type

- Overall, the success rate tends to climb with the number of flights.
- This is especially true for LEO orbit.
- GTO orbit doesn't show a clear relationship between success and flight numbers.
- HEO, MEO, VLEO, SO, and GEO were not even attempted until after flight 50.



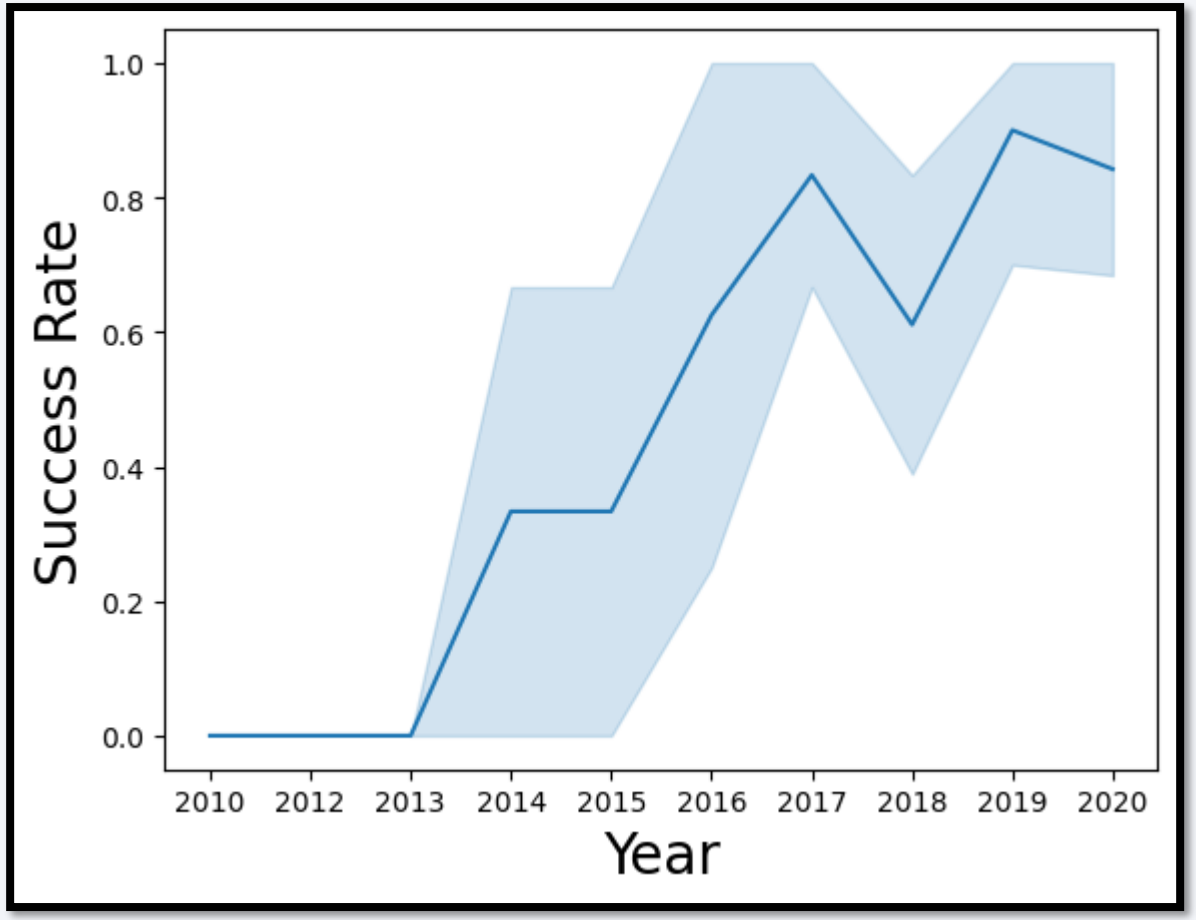
Payload vs. Orbit Type

- Heavier payloads increase the success rate in LEO, ISS, and Polar orbits
- ES-L1, SSO, and HEO all had 100% success rates, but all carried lighter payloads.
- No clear relationship is visible for GTO Orbit.



Launch Success Yearly Trend

- There were no successful booster recoveries from 2010 – 2013 in our data
- Success rate climbed drastically from 2013 – 2017, going from 0 to ~80% average success rate.
- There was a sharp decline in average success rate between 2017 – 2018.
- Success rate peaked in 2019 at around 85% average success rate.



All Launch Site Names

- We were able to extract the unique names of each launch site

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

- The query results show two launch sites at Cape Canaveral, one at Vandenberg Air Force Base, and one at Kennedy Space Center.

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

- The first 5 launches that begin with CCA in the dataset all came from launch site 40 at Cape Canaveral, all sent to LEO orbit, and most were sent by NASA.

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

SUM(PAYLOAD_MASS__KG_)
45596

- Through this query, we were able to find that the sum of total payload mass in kg carried by boosters launched by NASA to be 45,596 kg.

Average Payload Mass by F9 v1.1

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

AVG(PAYLOAD_MASS__KG_)
2928.4

- Through this query, we were able to find that the average payload mass carried by version F9 v1.1 boosters to be 2,928.4 kg.

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

MIN(Date)
2015-12-22

- By selecting the MIN(Date) from the dataset where the Landing Outcome to a ground pad resulted in success, we were able to find that the first successful landing outcome was December 22, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

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

- From this query, we were able to find that booster versions F9 FT B1022, F9 FT B1026, F9 FT B1021.2, and F9 FT B1031.2 all successfully landed on a drone ship and carried a payload mass between 4000 and 6000kg.

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

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

- Through this query, we were able to find an astonishingly high mission success rate with only one failure, but it is important to note the difference that the mission success does not infer a successful landing outcome.

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- There were a total of 12 booster versions that carried the maximum payload mass varying from F9 B5 B1048.4 to F9 B5 1060.3.

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

2015 Launch Records

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

Month	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- There were a total of two launches that had failed landing outcomes to drone ships in 2015. Both came from launch site CCAFS LC-40. One in January, and one in April.

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

- 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

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

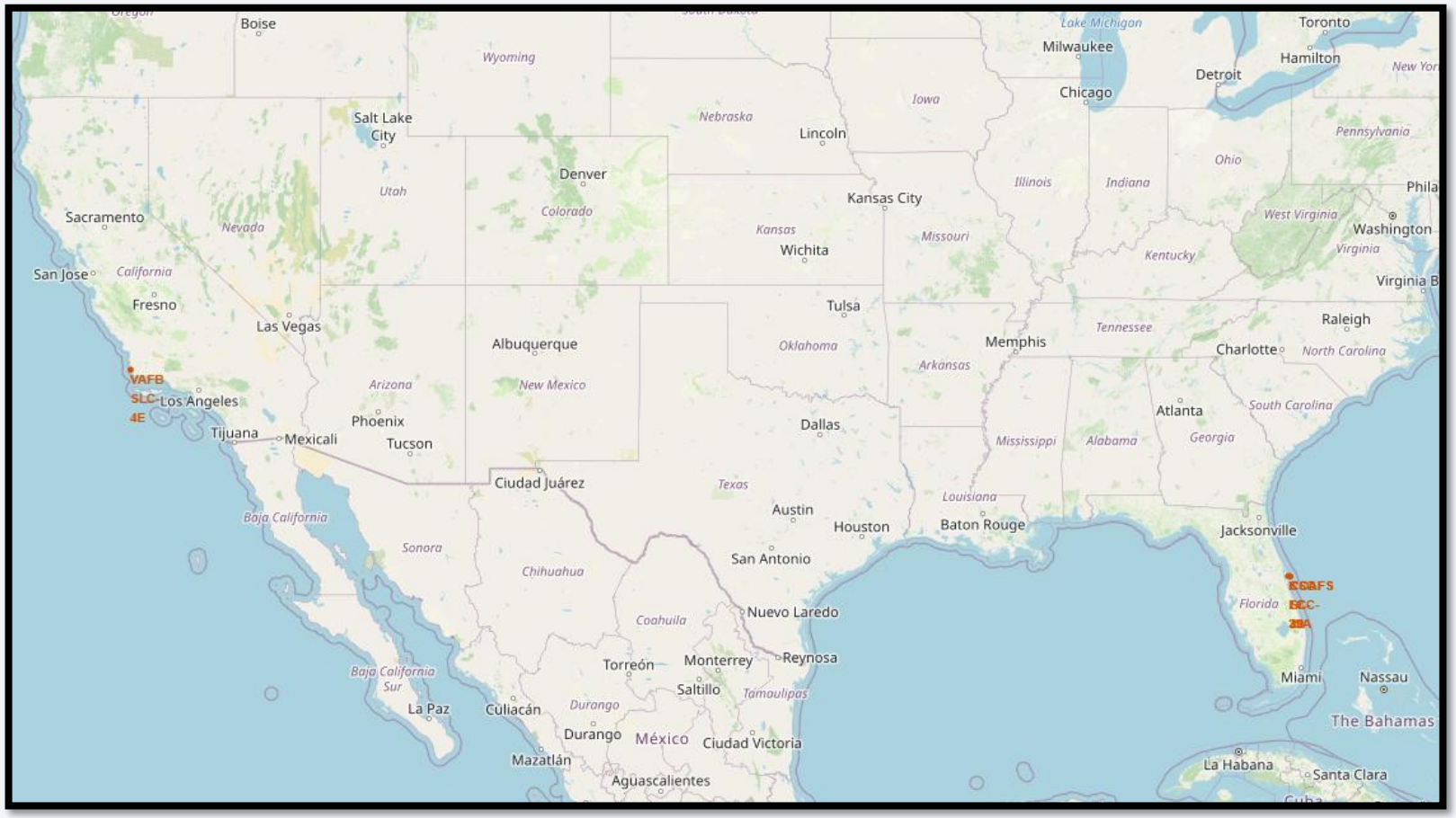
- From this query, we can see that there were a total of 31 launches within this time frame. No attempt was made to recover 10 of the boosters, and most of the recovery attempts were to drone ships with a 50% success rate.

Section 3

Launch Sites Proximities Analysis

Launch Locations – High Level

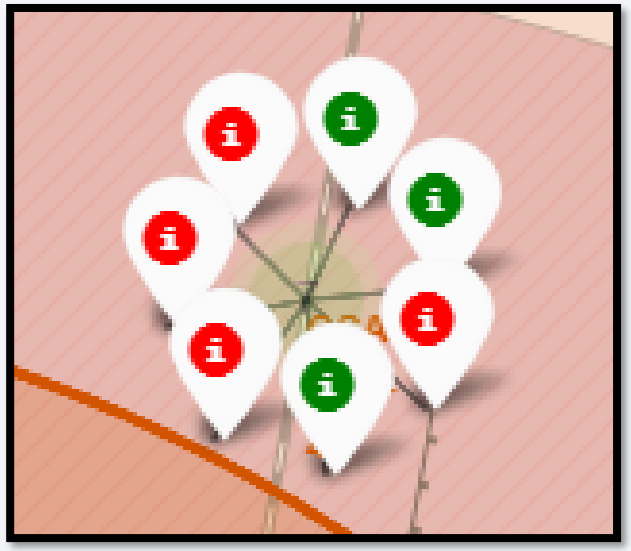
Here we can see the launch site locations on a map. There is only one on the west coast of the US, and 3 clustered tightly together in FL. It's important to note that all launch sites are located on the coasts.



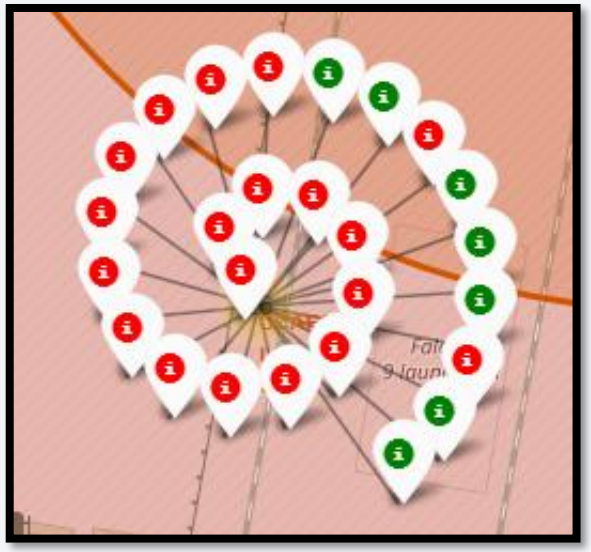
Launch Sites – Success vs Failures

From the screenshots below, we can visually see that CCAFS LC-40 had the most launches. We can also see that the most successful launches came from KSC LC-39A. CCAFS SLC-40 had the fewest launches, and CCAFS LC-4 had the most failures.

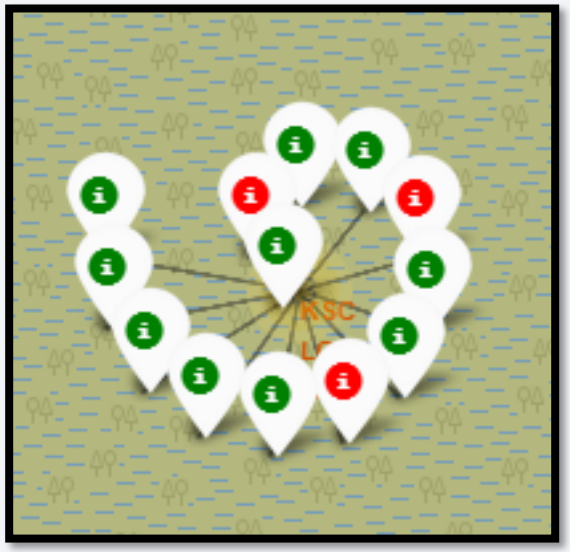
CCAFS SLC-40



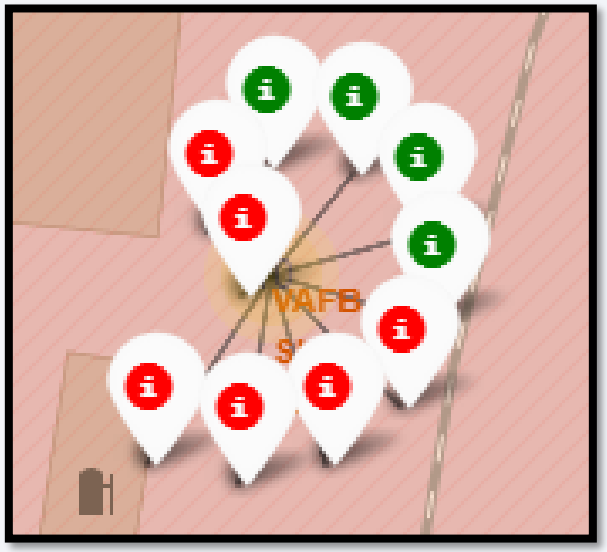
CCAFS LC-40



KSC LC-39A

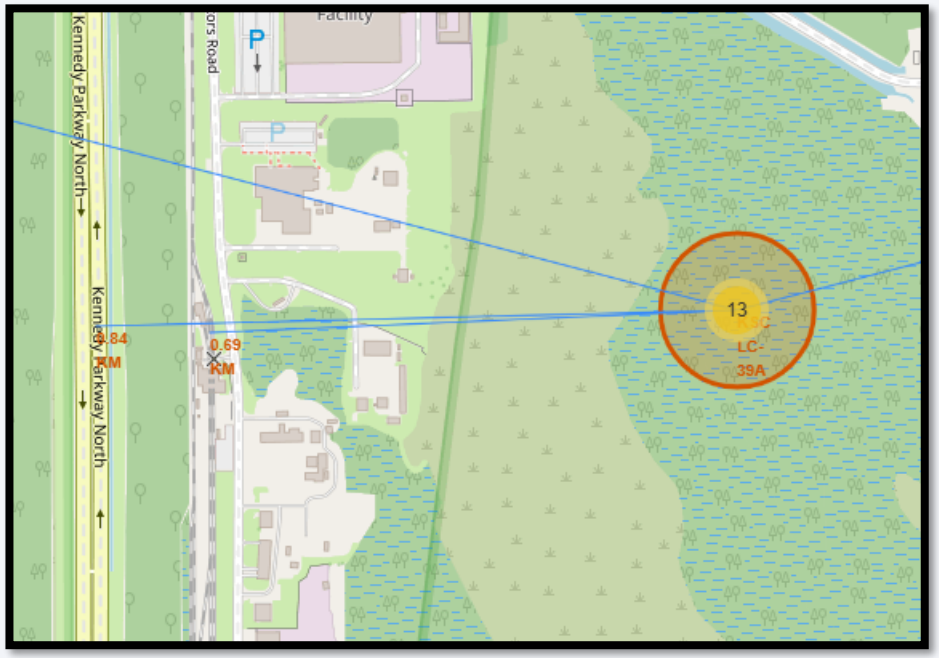


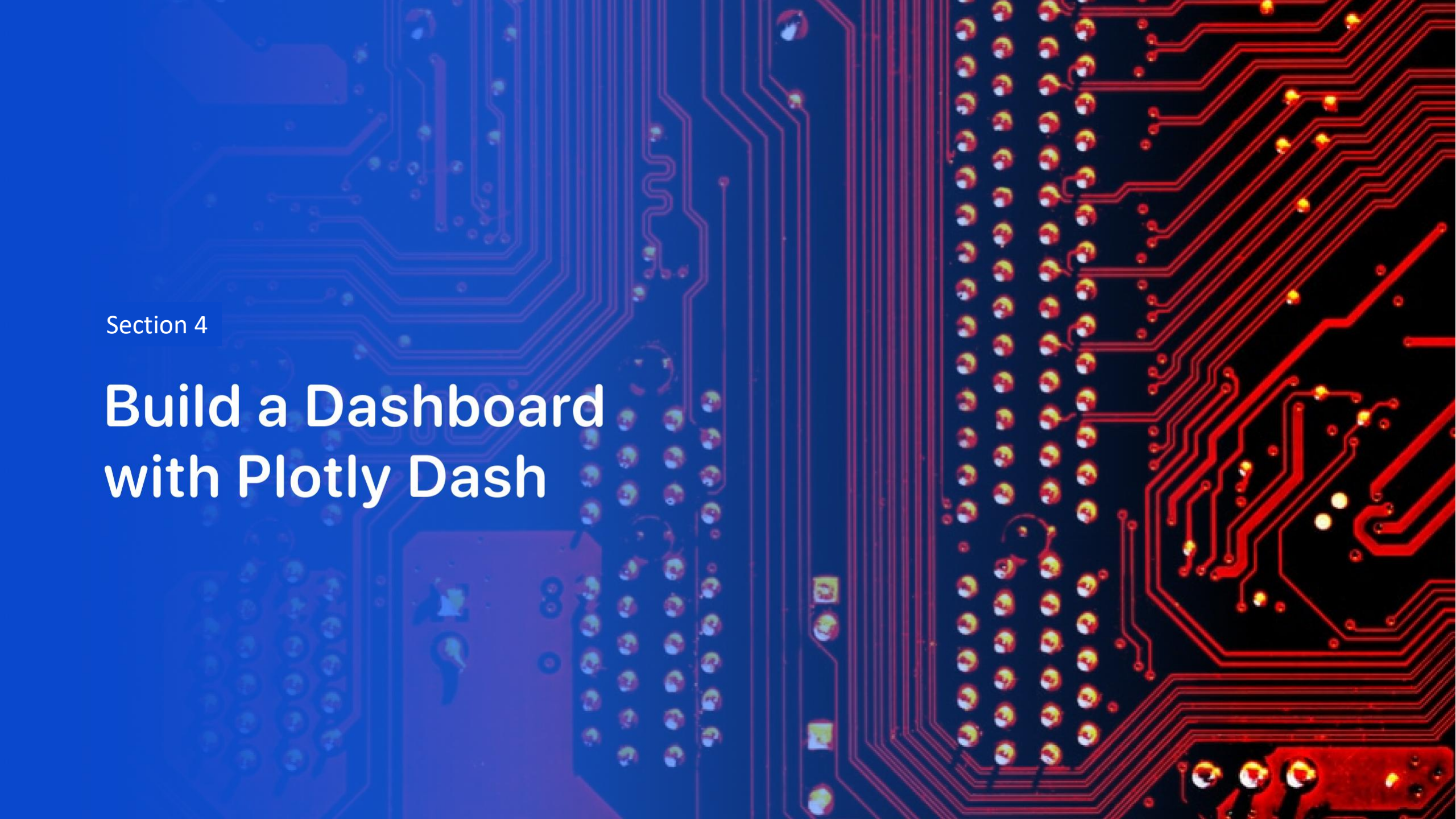
VAFB SLC-4E



<Folium Map Screenshot 3>

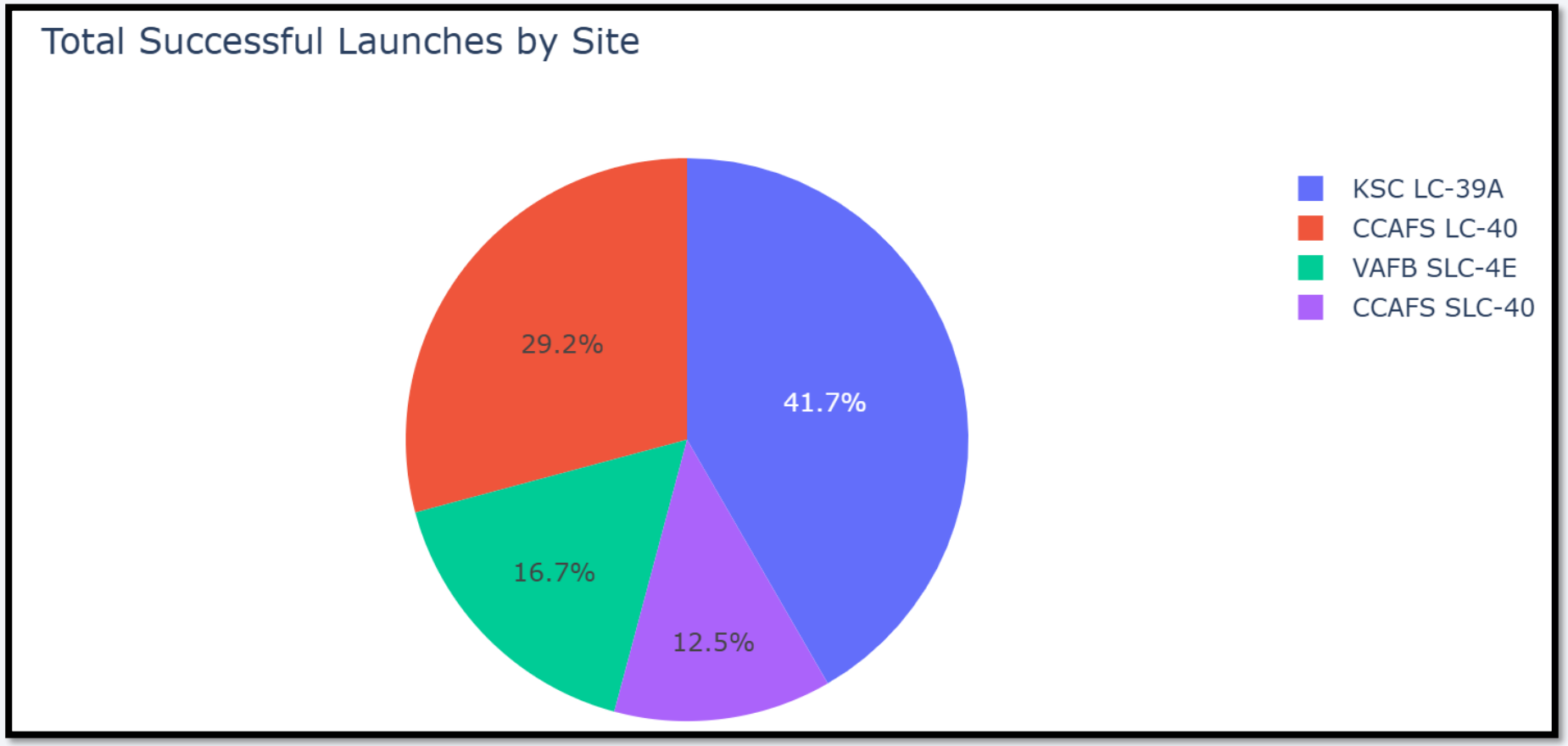
One thing that all launch sites had in common was the relationship between their location and proximities to nearby map features. As an example, we can see below that KSC, like the other launch sites, is situated close to coastlines, railroads, and highways, but intentionally distant from nearest towns or cities.




Section 4
Build a Dashboard with Plotly Dash

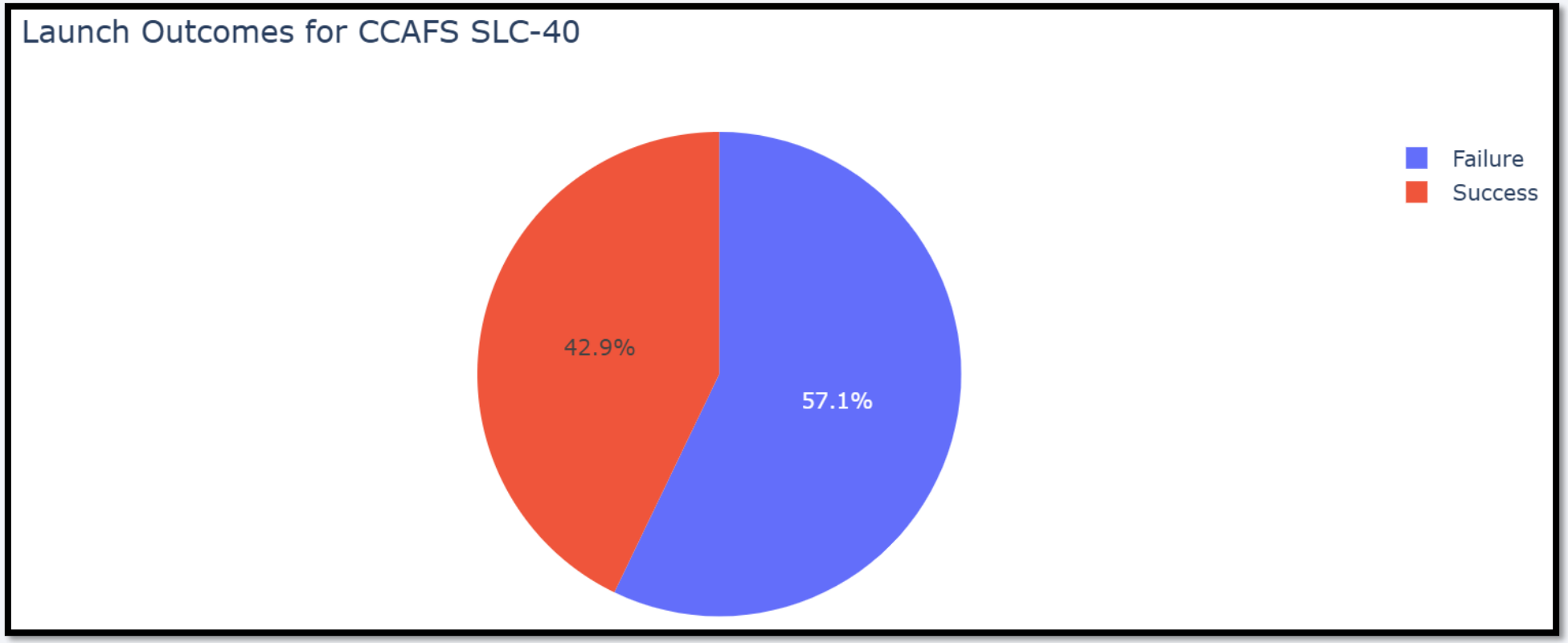
Total Successful Launches by Site

Using our dashboard, we can see that Kennedy Space Center had the most successful launches with 41.7% success rate of all launches from all sites.



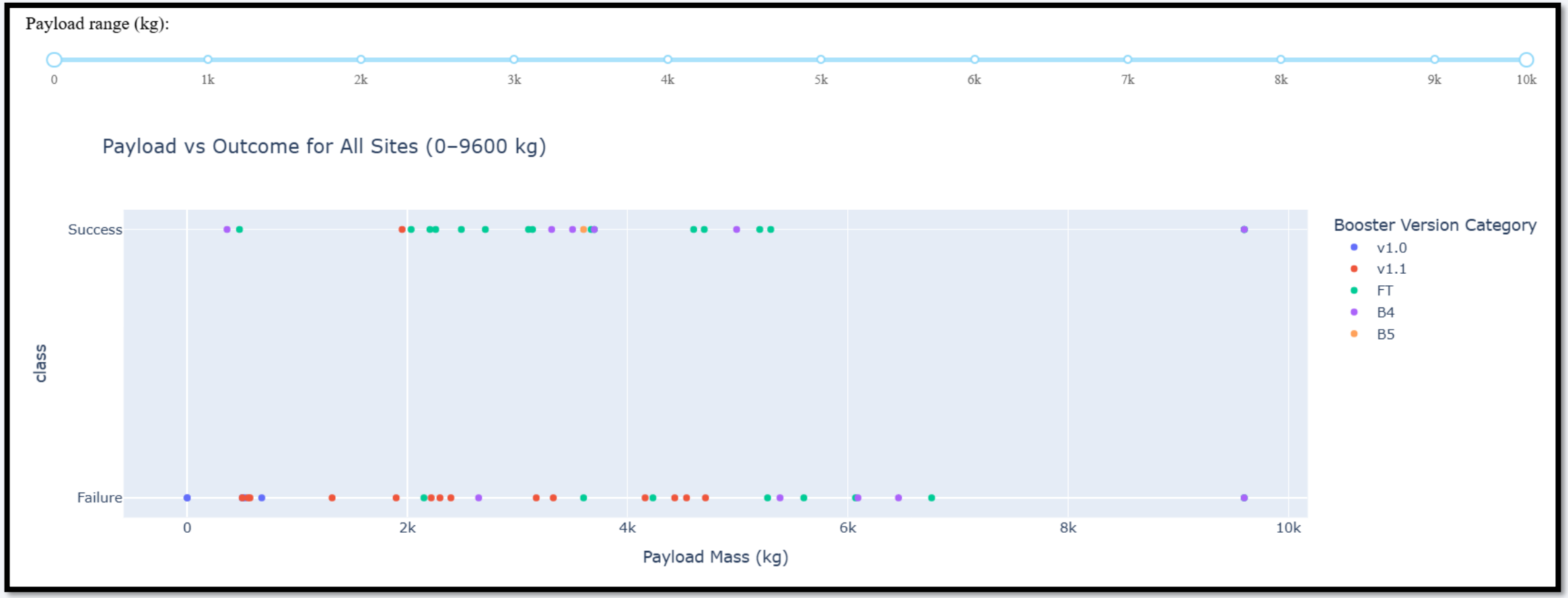
Highest Success Rate by Site

CCAFS SLC-40 had the highest success rate of all sites, with 42.9% of all launches ending with the successful recovery of the Falcon9 boosters.



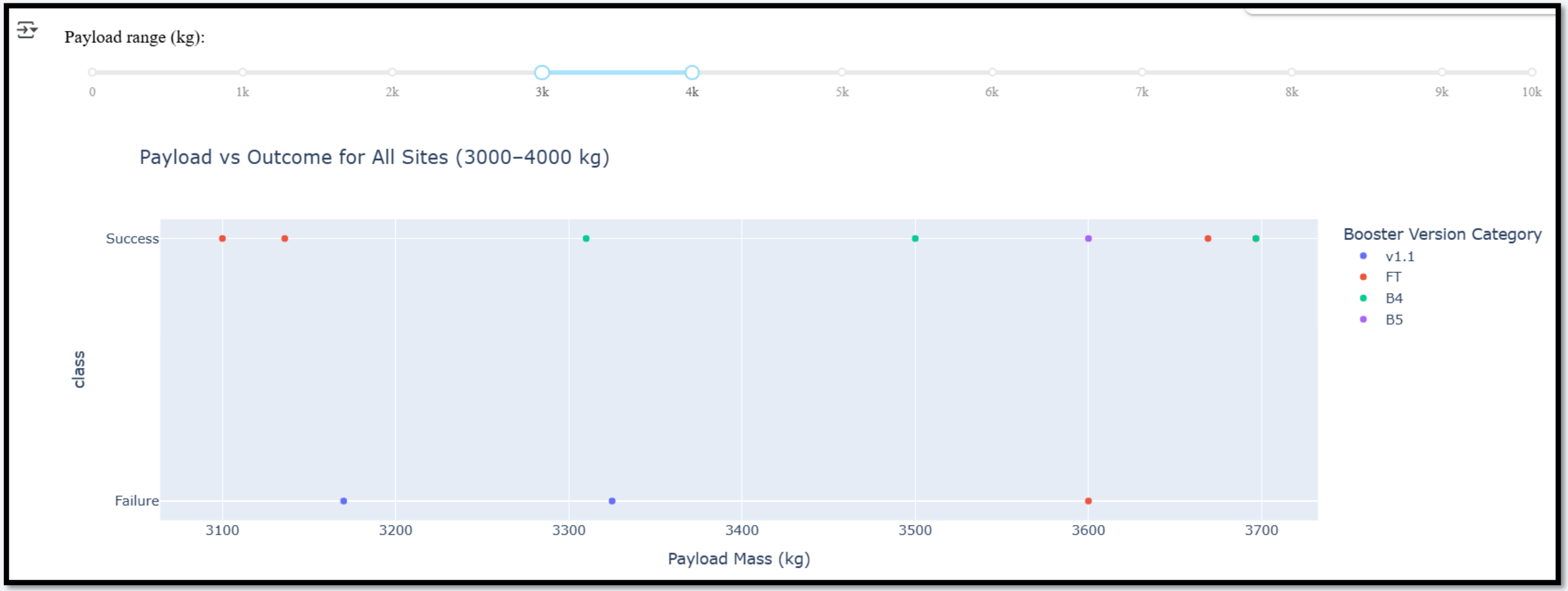
Payload vs. Outcome Relationships

The screenshot below shows that of all the booster versions, B5 had the highest success rate of 100%. However, it only had one launch. FT had the second highest success rate at ~68% with a total of 19 launches.



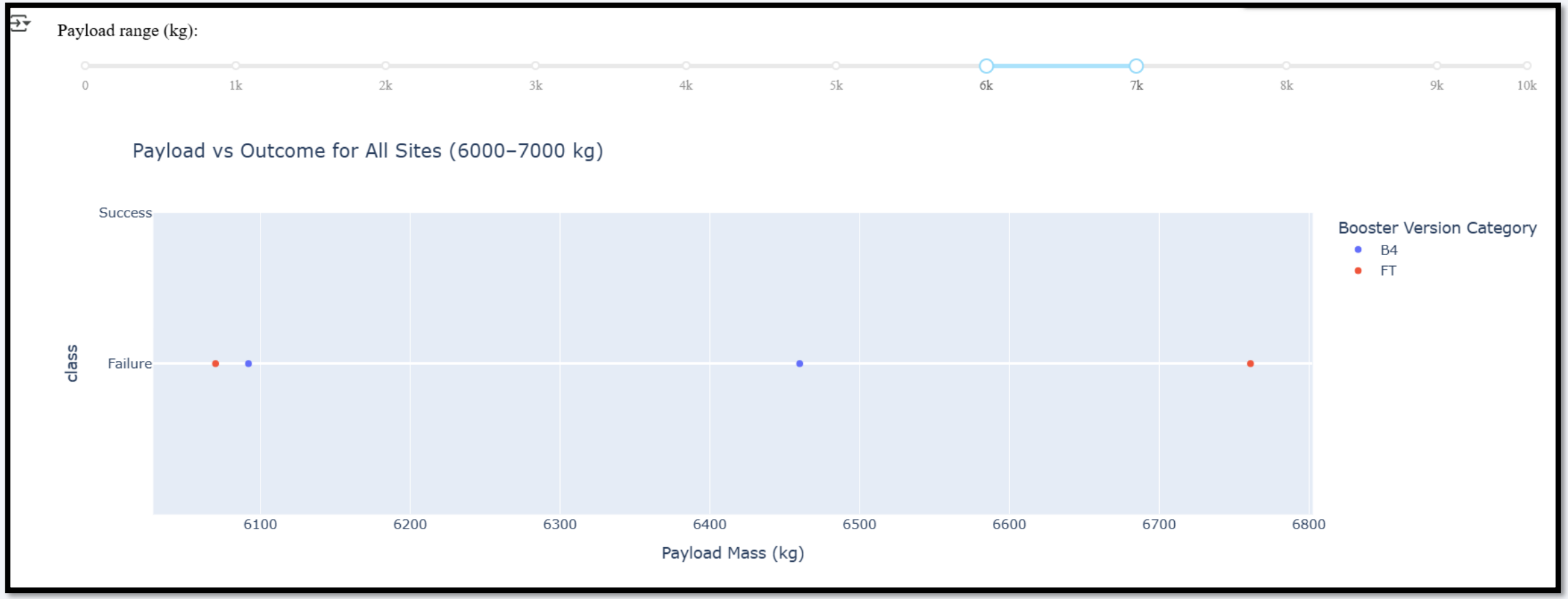
Payload vs. Outcome Relationships (cont.)

The screenshot below shows that payloads between 3000-4000 kg had the highest success rate at around 70% success rate.



Payload vs. Outcome Relationships (cont.)

The screenshot below shows that payloads between 6000-7000 kg had the lowest success rate with no successful recoveries and 0% success rate.

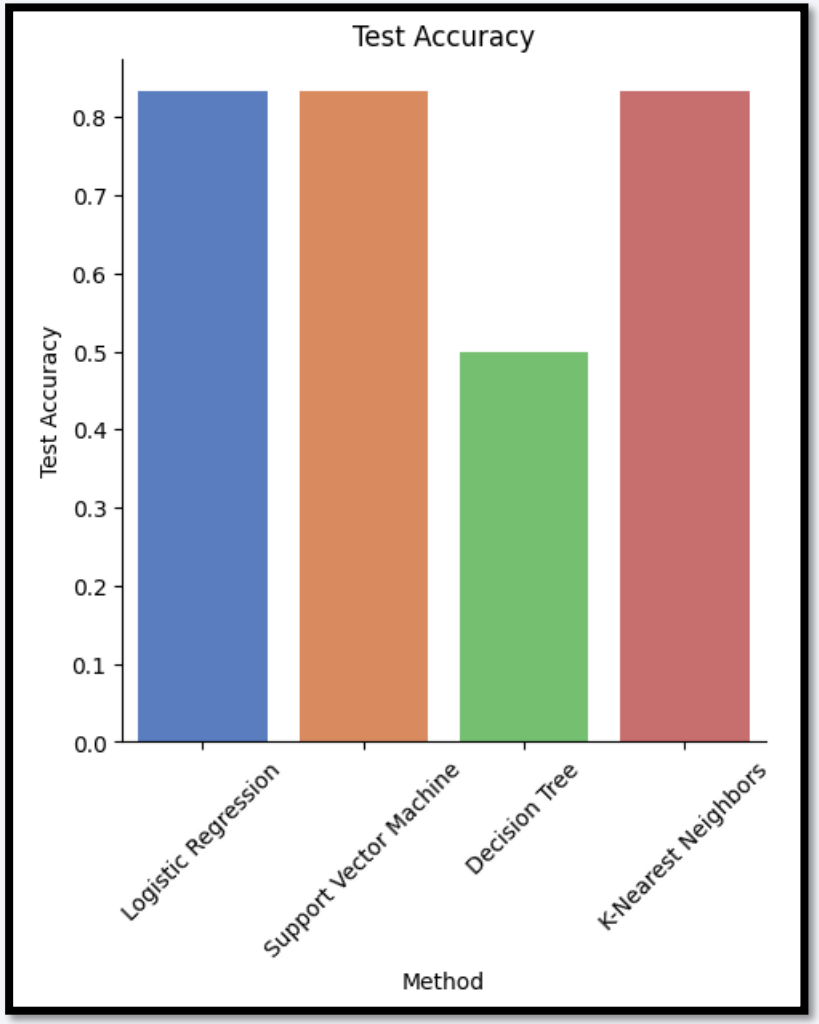


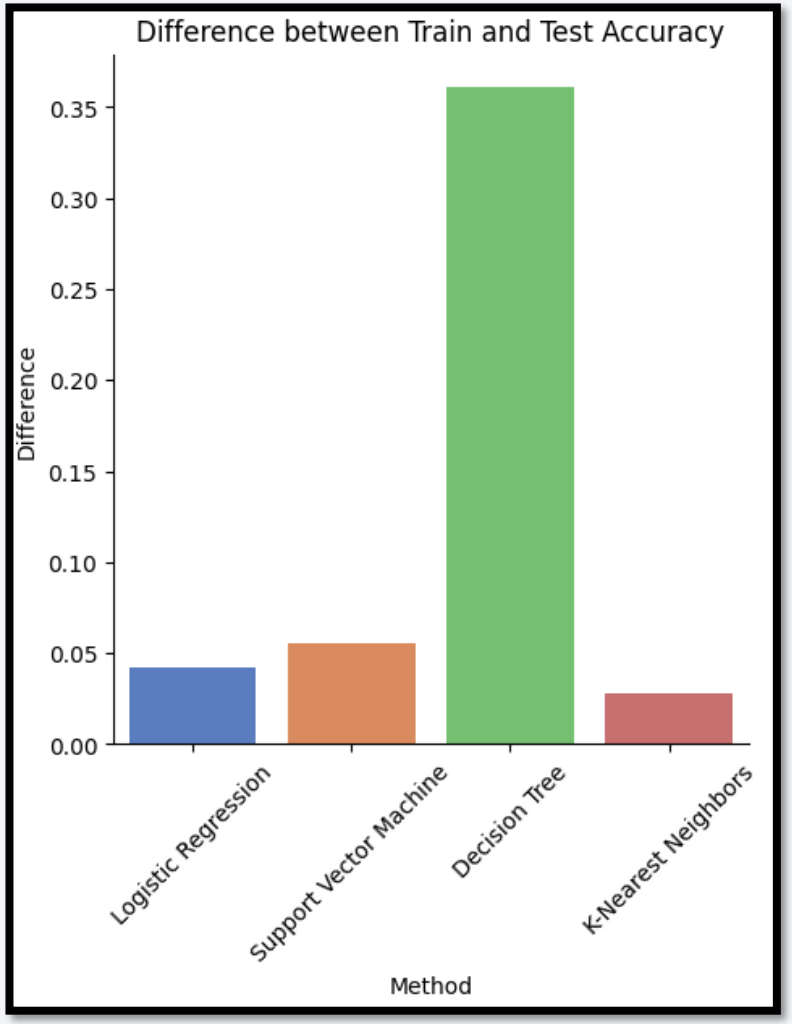
Section 5

Predictive Analysis (Classification)

Classification Accuracy

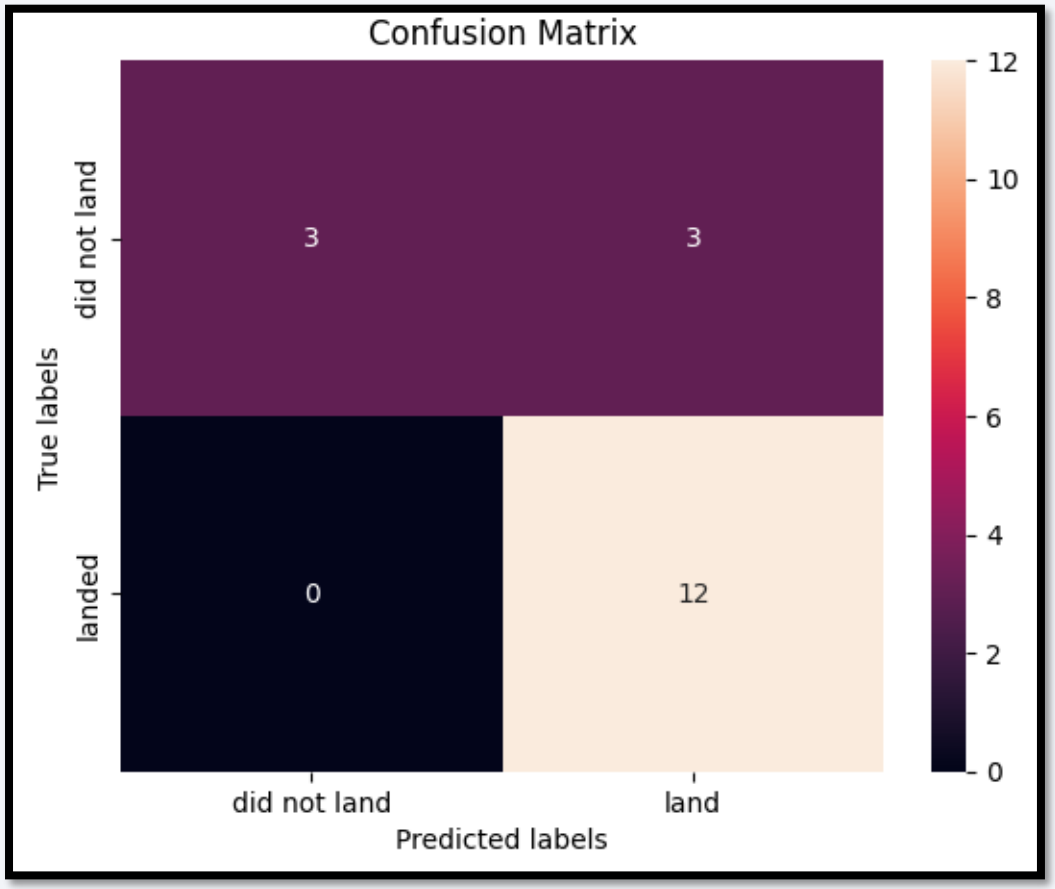
- Three of the models had the exact same performance on the test set at ~83.34% accuracy. However, since KNN had the least difference between training and testing at ~.027%, KNN is my selection for final model.





Confusion Matrix

- The KNN model had perfect accuracy on the majority class (successful landing), but struggled with failures, only getting 50% correct.



Conclusions

- We could potentially use the Logistic Regression, K-Nearest Neighbor, or Support Vector Machine models to make future predictions on the probability of successful returns of the Falcon9 boosters with around an 83% accuracy. However, since the actual success rates of current missions is around the same % success rate, none of the models are doing much better than simply predicting all launches will be successful.
- If we must go with a single model now from this research, since the difference in the training and test scores were lowest on the KNN model, I would suggest that model as our final model.
- It is important to note though, that the Decision Tree Classifier showed the most promise with the highest training accuracy, and more tuning should be done before eliminating this as an option.
- Given the imbalance of the target variable, we should go back and stratify the datasets when splitting, as well as assign weights to the target variables to help in overcoming the class imbalance and improve recall by reducing false negatives before landing on a final model.

Thank you!