IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Benjamin Reale
1/12/2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection via the SpaceX API and web scraping through BeautifulSoup form Wikipedia;
  - Data Wrangling and Exploratory Data Analysis (EDA) through Visualization and SQL;
  - Interactive Location Analysis through Folium, Dashboard creation through Plotly Dash, and ML Predictive Analysis.
- Summary of all results
  - It is possible to predict if a new launch will land successfully or not. The data collected, either from a database or through web scraping, provided insight into the relation between multiple variables and the chance of a successful landing, and led to the creation of an accurate model for predicting landing success.

Introduction

- Project background and context
  - SpaceX has been successful in reducing the cost of space travel through the reuse of the first stage of their rockets. This reuse is only possible, however, if the first stage successfully lands. Knowing the chance of a successful landing before launch and the attributes that contribute to a successful launch would assist SpaceX, or a rival space travel company, in successfully reusing more first stage rockets.
- Questions to be answered.
  - How do variables such as launch site, payload mass, target orbit, and more affect the success of the landing of the first stage?
  - What kind of classification model will most accurately predict future landings?

Section 1

Methodology

Methodology

- Data collection methodology:
  - Via the SpaceX API
  - Via Web Scraping from Wikipedia
- Perform data wrangling
  - Creating a landing outcome label from landing data, replacing missing values, and One Hot Encoding categorical variables to binary classification.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building, fitting, and comparing the accuracy of four popular classification models to determine most accurate for future predictions.

Data Collection

Data collection was completed using GET requests from the SpaceX API and web scraping the html data from the Falcon 9 Wikipedia page. After downloading and converting the datasets to Pandas dataframes, the desired variables were selected and combined into a single dataset.

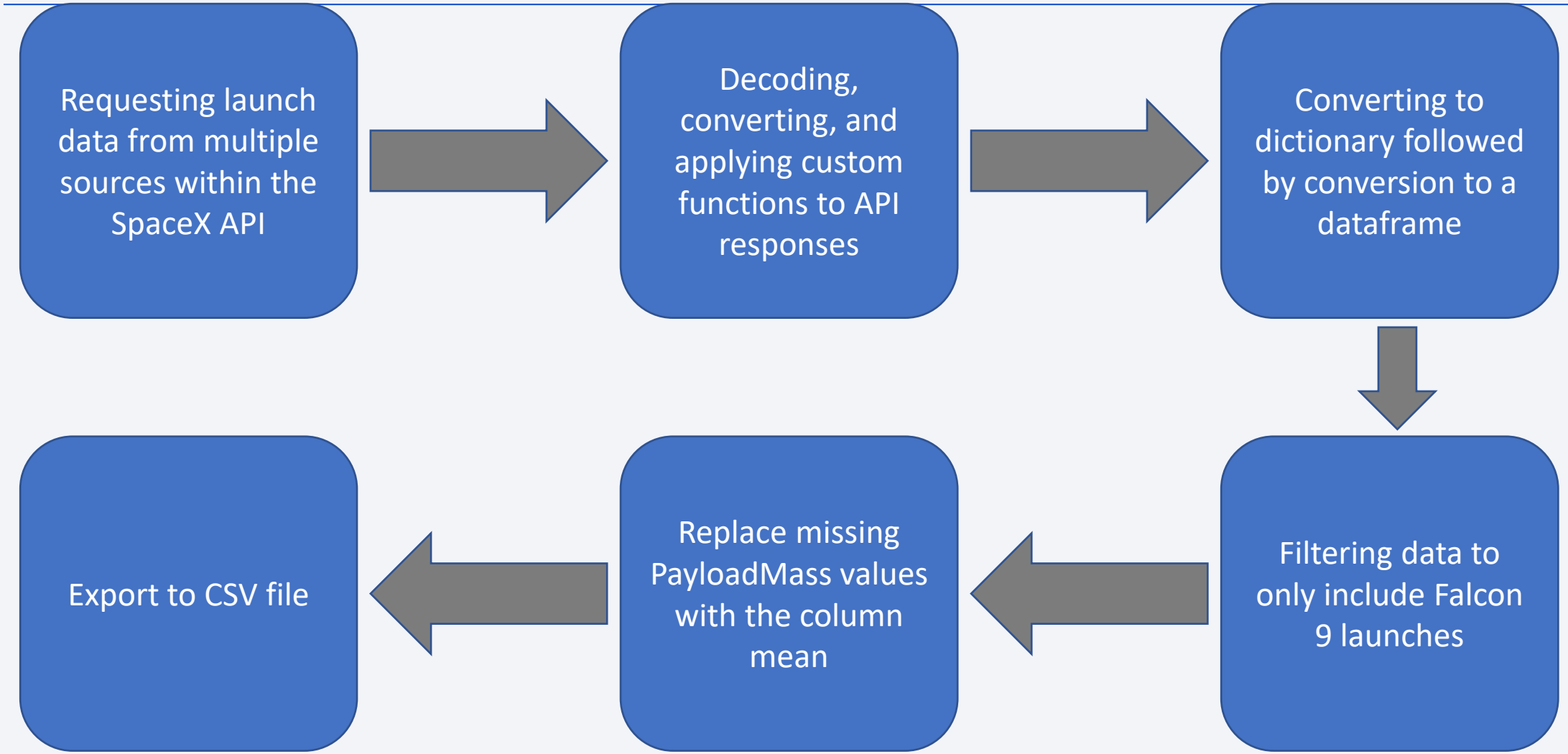
Data obtained from SpaceX API:

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Latitude, Longitude

Data obtained from Wikipedia web scraping:

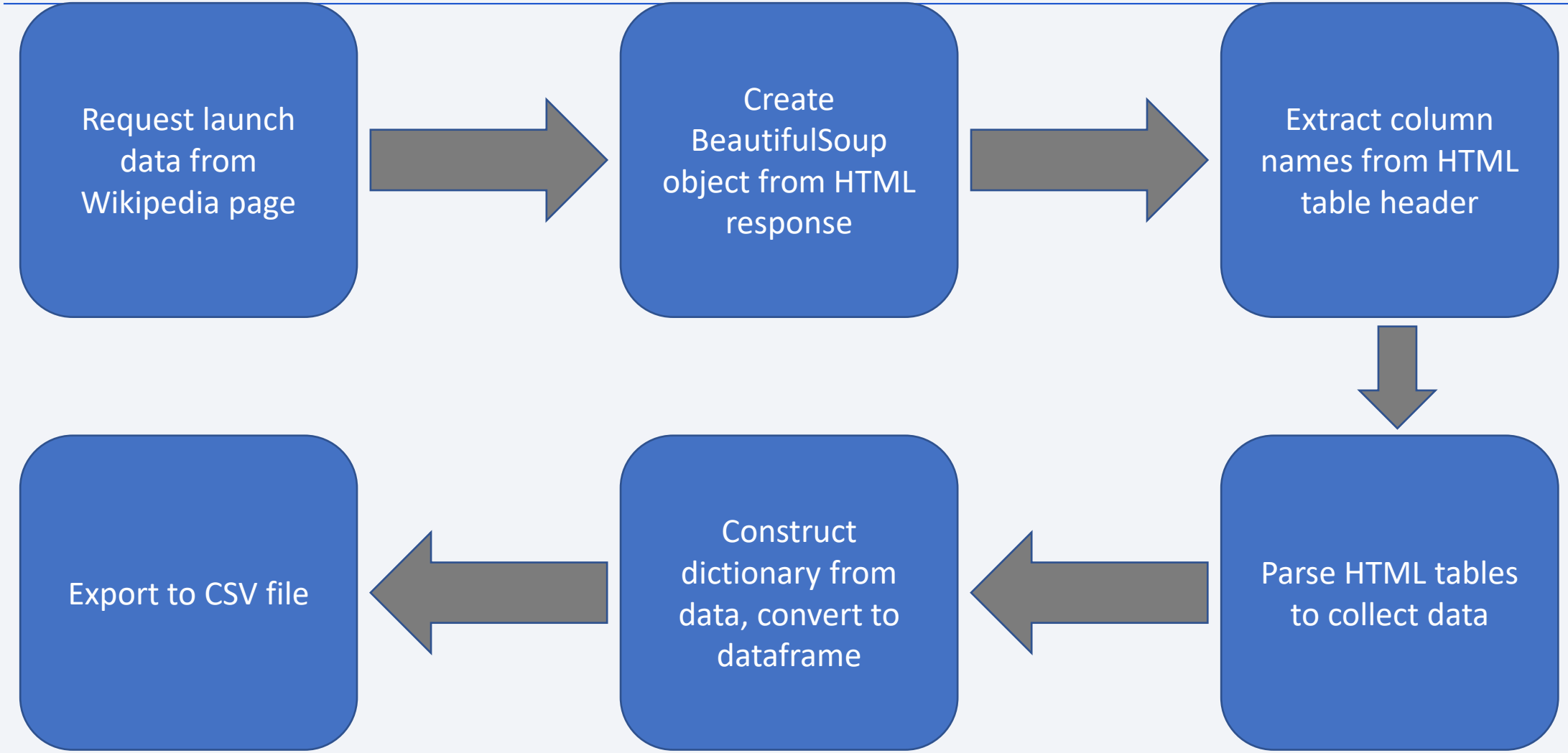
Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



GitHub link

Data Collection - Scraping

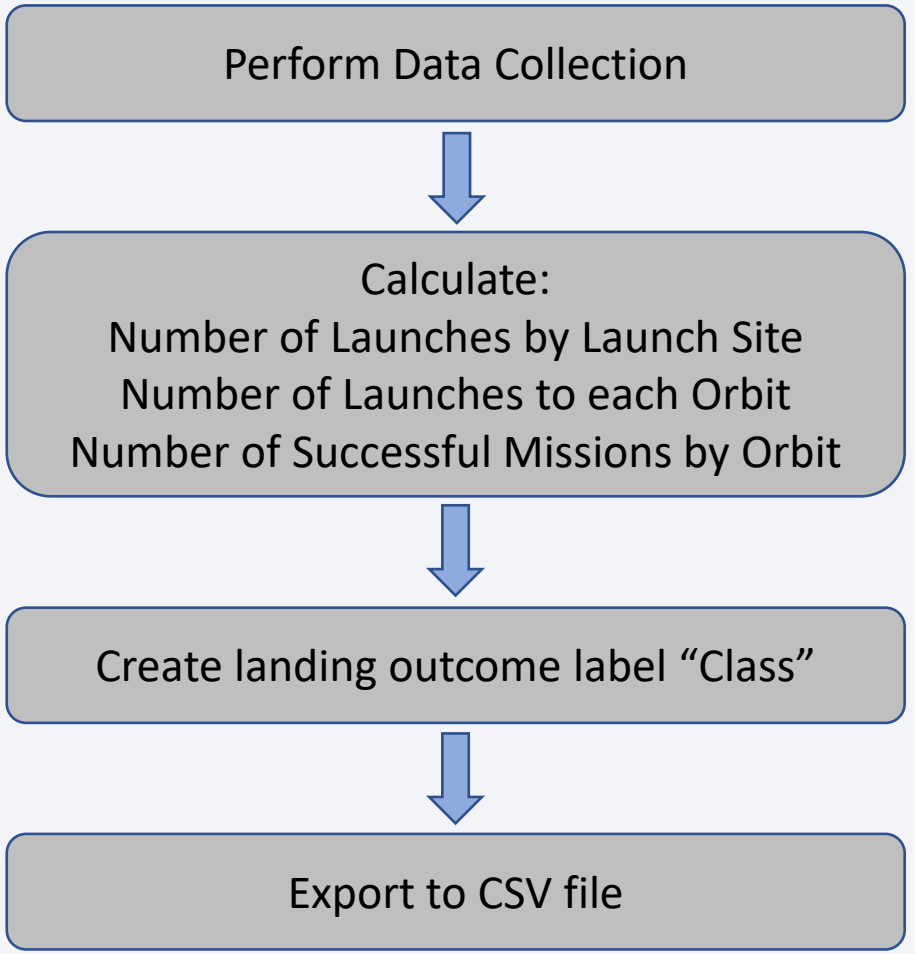


GitHub link

Data Wrangling

In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS and True ASDS means the rocket successfully landed on a ground pad or drone ship while False RTLS and False ASDS means the rocket failed to land on a ground pad or drone ship.

All various successful landings were converted to “1” and all failures were converted to “0”.



GitHub Link

EDA with Data Visualization

To explore the data, scatterplots, bar graphs, and line graphs were created to visualize the relationship between two variables, with the scatterplot data colored to show the successful and unsuccessful landings of the first stage.

Scatterplots:
- Payload Mass vs. Orbit
- Flight Number vs. Orbit
- Flight Number vs. Payload Mass
- Flight Number vs. Launch Site
- Payload Mass vs. Launch Site

Bar Graph: Success Rate by Orbit

Line Graph: Average Success Rate over Years

GitHub Link

EDA with SQL

SQL Queries Performed:

- Display the names of unique landing outcomes
- Display the first 5 records at launch sites with the string ‘CCA’
- Display the payload masses carried by boosters launched by NASA (CRS)
- Display the average payload mass carried by F9 1.1 booster
- Display the date when the first successful landing outcome on a ground pad was achieved
- Display the names of boosters that had success on a drone ship and had payload mass between 4000 and 6000 kg
- Display total number of successful and failed mission outcomes
- Display names of booster version that have carried maximum payload mass
- Display month, year, landing outcomes, booster version, and launch site for launches in 2015 that failed on a drone ship
- Display the count of successful landings between 04/06/2010 and 20/3/2017, ranked.

GitHub Link

Build an Interactive Map with Folium

Objects added to Folium map:

- All launch sites are marked with circle marker, label, and popup label.
- Each launch site has a marker cluster for each launch at that site, with each marker in the cluster either colored red for failed landings or green for successful landings.
- Four more places are marked around the CCAFS SLC-40 launch site, and each location has a line connecting it to the launch site and a label showing how far it is from the launch site. These locations are the closest coastline, railroad, highway, and city to the launch site.

GitHub Link

Build a Dashboard with Plotly Dash

Launch Site Dropdown:

- A dropdown list to select all or one specific launch site.

Pie Chart:

- Shows number of successful landings at all launch sites or proportion of successful and failed landings at one specific launch site.

Payload Mass Slider:

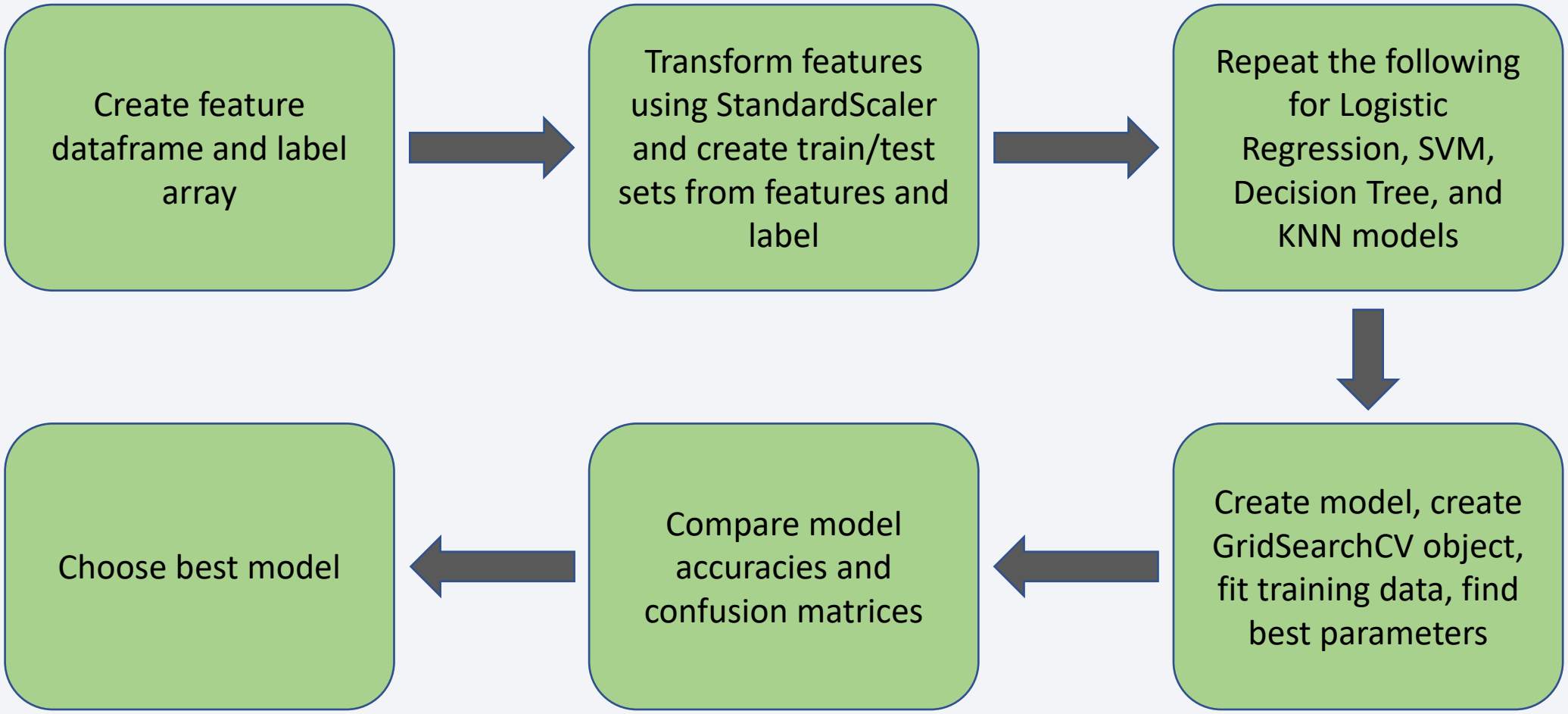
- A range slider to select the range of payload mass you wish to display.

Scatter Plot:

- Graphs payload mass against successful landings.

GitHub Link

Predictive Analysis (Classification)



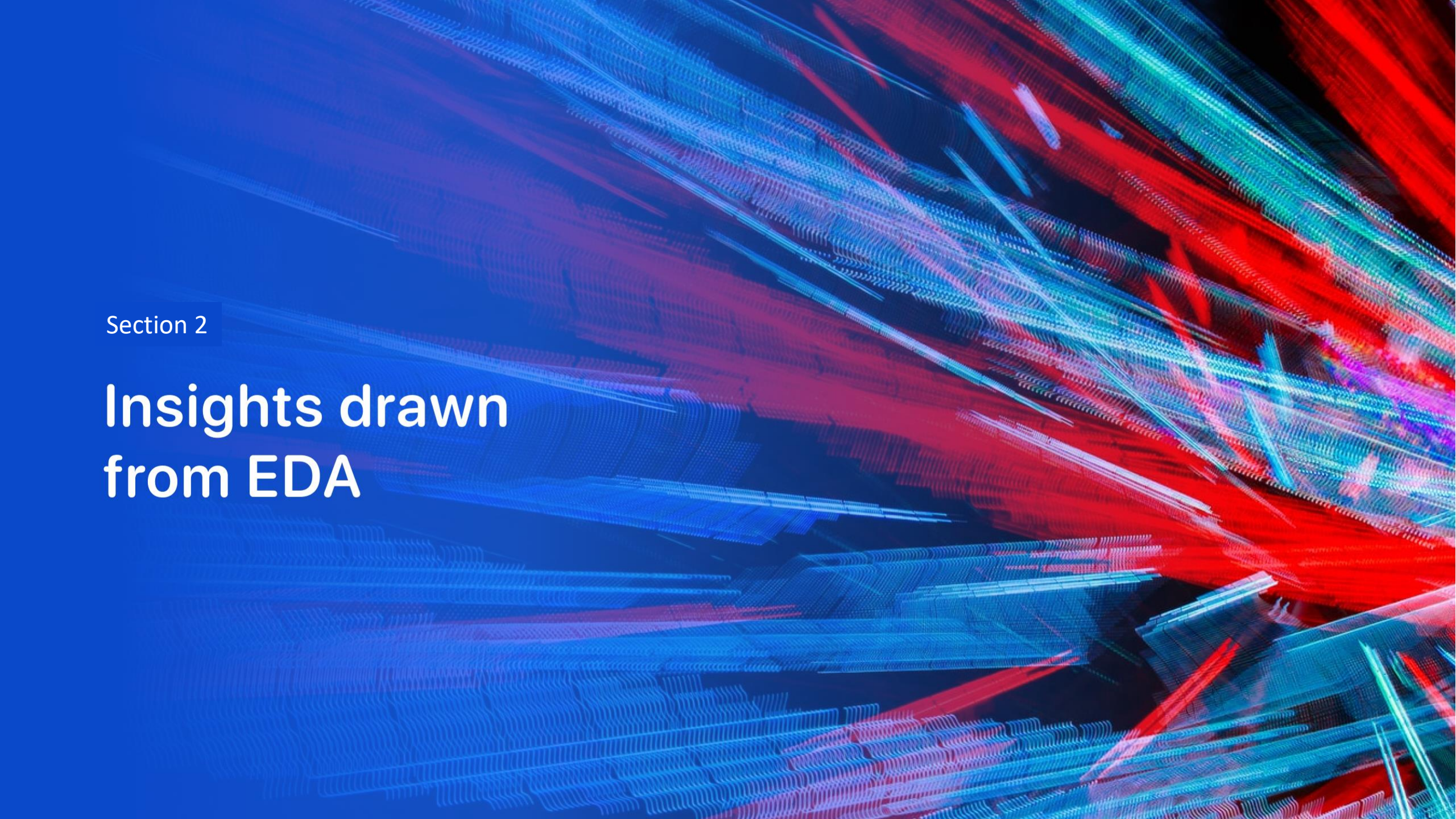
Results

EDA results

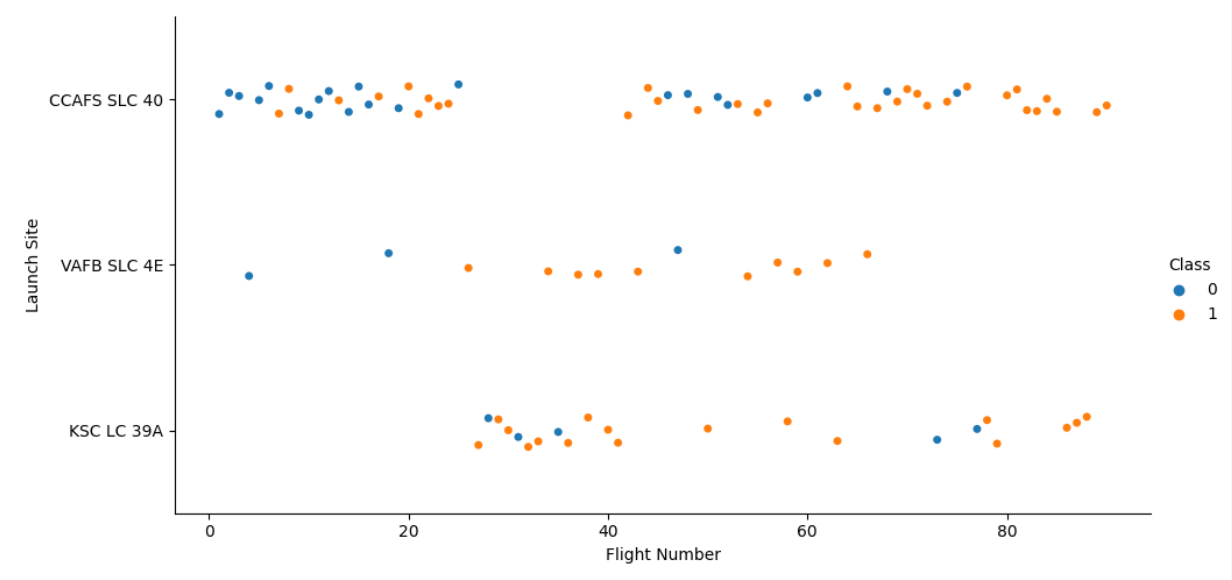
- Landings became more frequently successful over time
- One launch site has a lower landing success rate than the other two
- Most launches have low payload mass, if there is a heavy payload it is usually the max mass
- Four orbits have perfect success rate, and all but one are above 50%, one launch to SO had a failed landing

Predictive analysis results

- Logistic Regression, Support Vector Machine, and K-Nearest Neighbors all had the same high accuracy, and the Decision Tree had a slightly lower accuracy


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

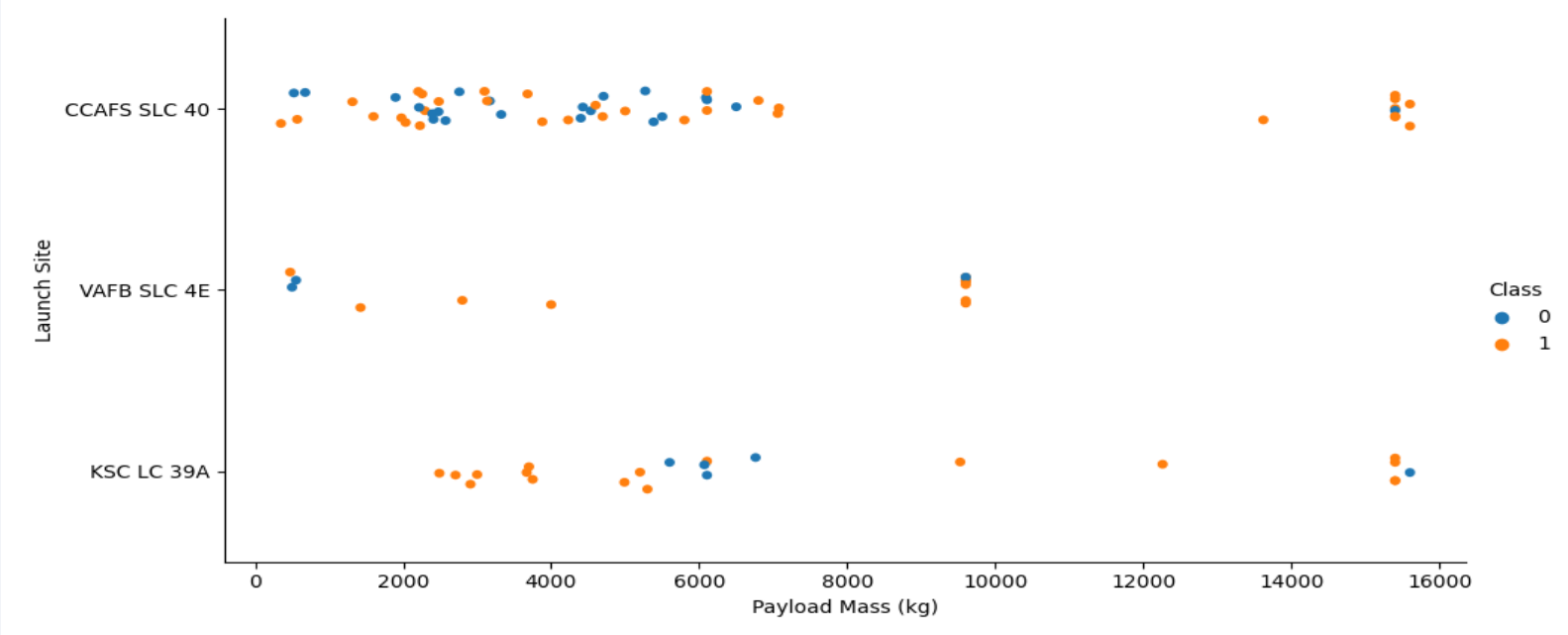


Orange indicates a successful landing and blue indicates unsuccessful landing.

This graph suggests the success rate has increased over time, with some kind of change occurring around flight 20 which increased success rate.

Most launches being at the CCAFS site.

Payload vs. Launch Site

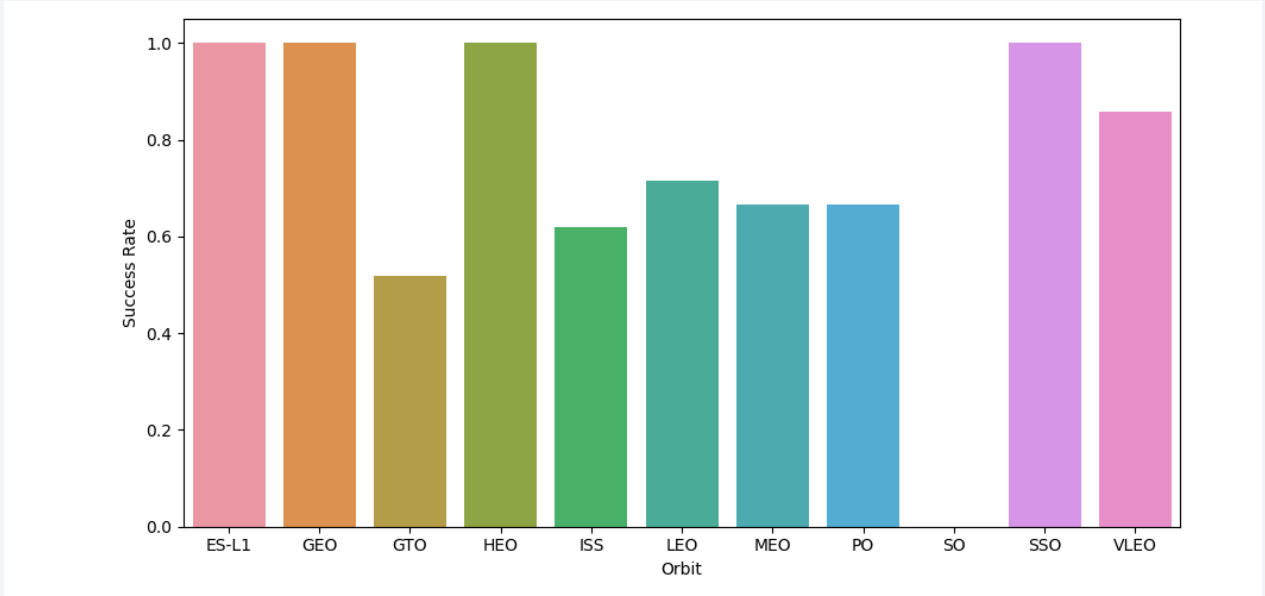


Orange indicates a successful landing and blue indicates unsuccessful landing.

Most launches have payloads less than 7000 kg.

It appears launch site may be partially determined by payload mass.

Success Rate vs. Orbit Type

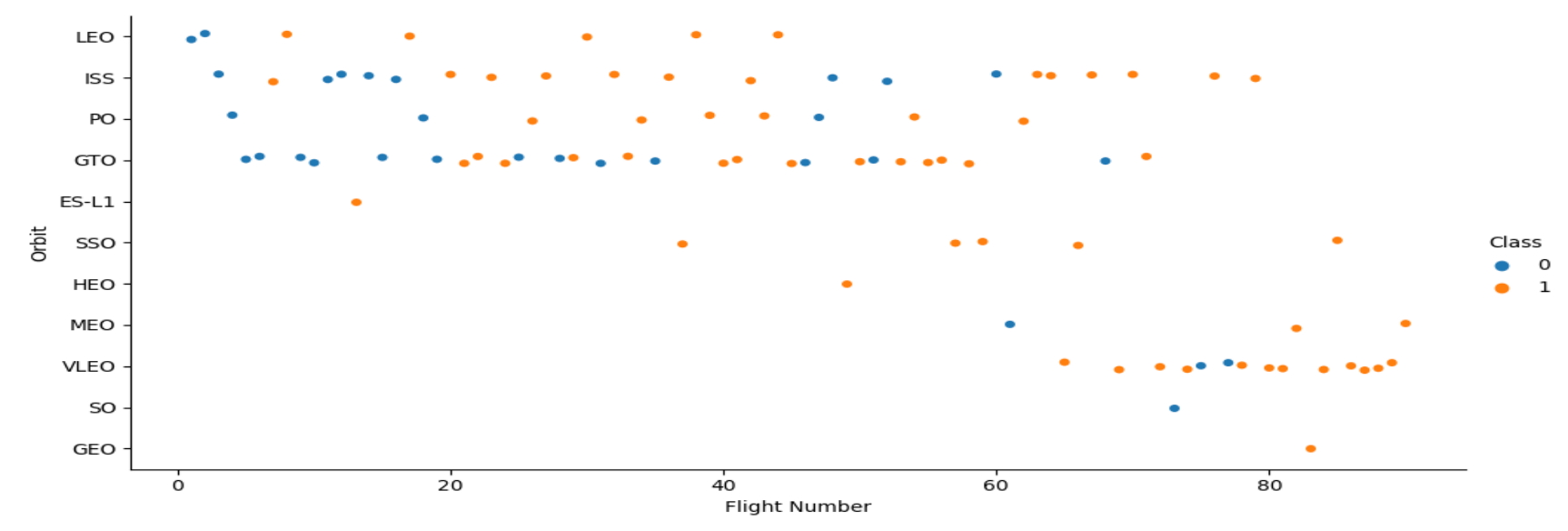


Success Rate is a percentage as a decimal.

While ES-L1, GEO, and HEO have 100% success and SO has 0%, they each only have one launch. SSO has 5 launches with 100% success.

GTO has 27 launches with approximately 50% success

Flight Number vs. Orbit Type

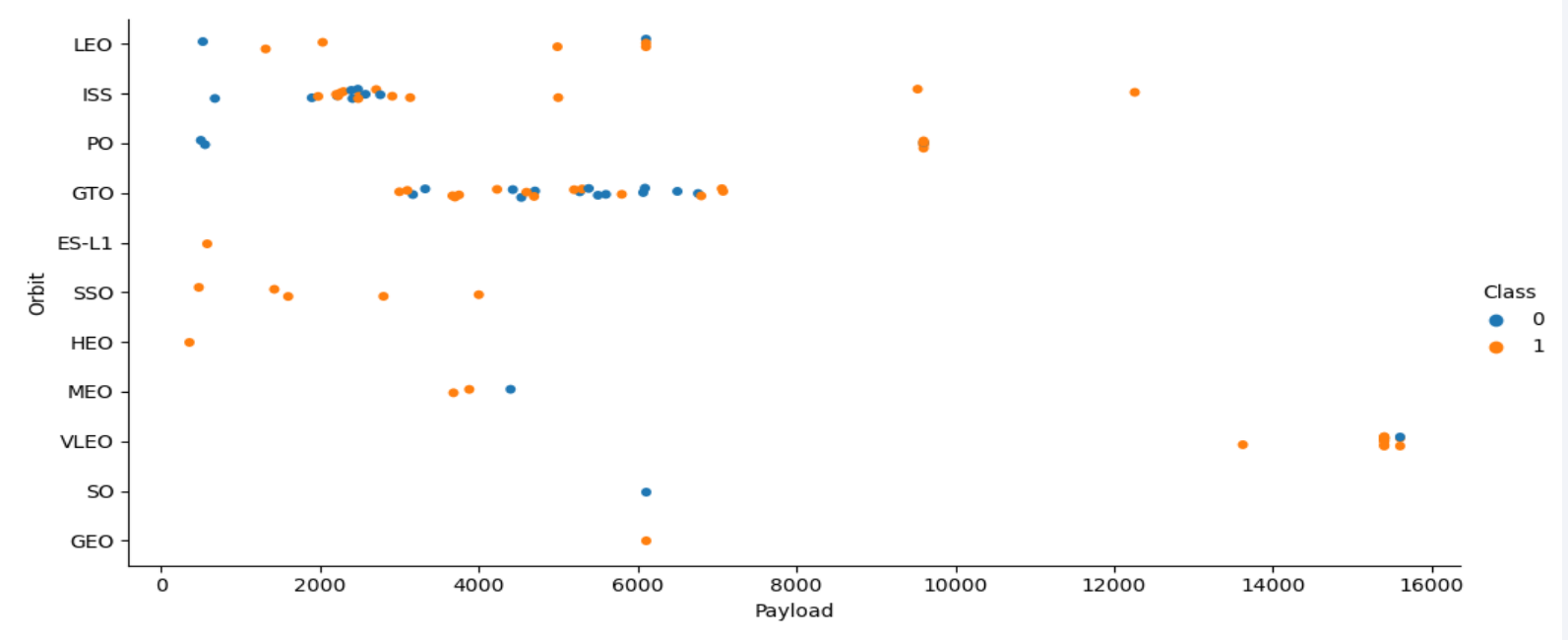


Orange indicates successful landing, blue indicates failed.

Again, success rate appears to increase over time.

Target Orbit seems to have shifted over time, away from LEO towards VLEO.

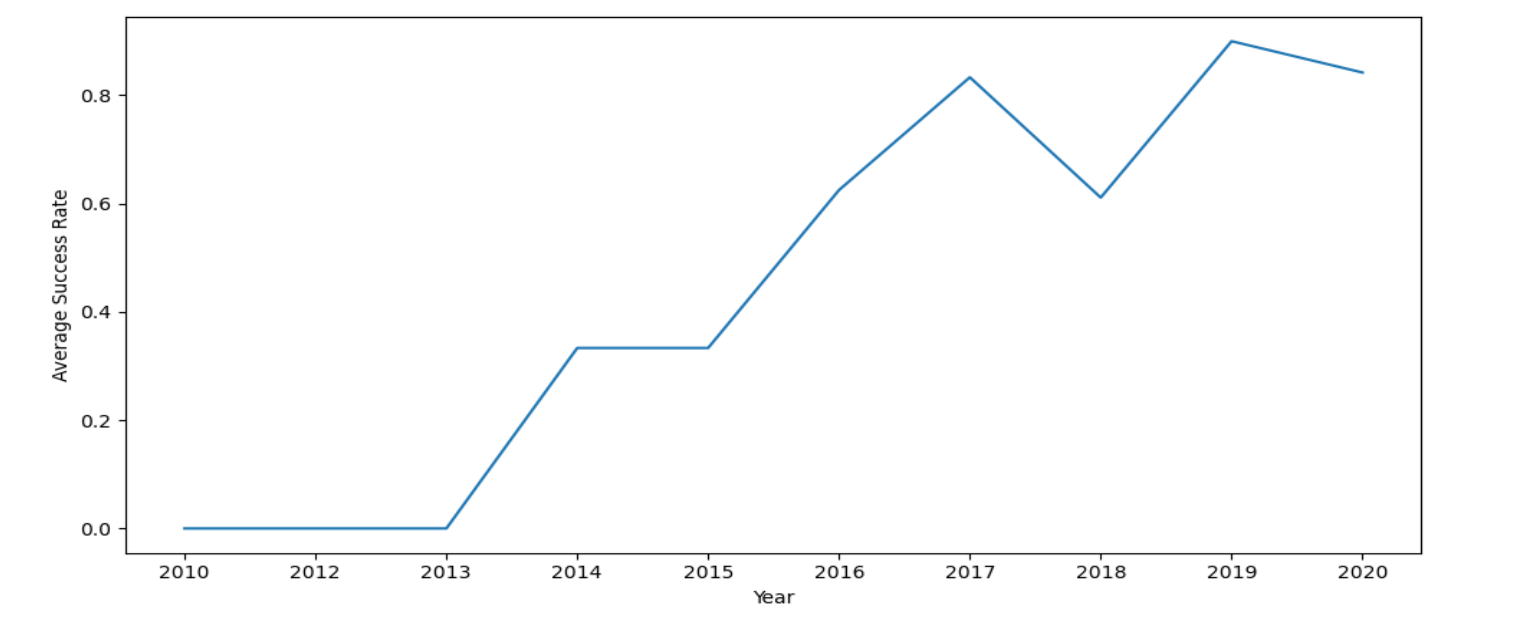
Payload vs. Orbit Type



Orange indicates successful landing, blue indicates failed.

The ISS seems to have the most low-mass launches, GTO has the most medium-mass, and VLEO has the highest-mass launches.

Launch Success Yearly Trend



Success rate increased over time with a dip in 2018, with a peak success rate around 85%.

All Launch Site Names

```
%sql select DISTINCT LAUNCH_SITE from SPACEXDATASET
```

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

SQL Query for unique launch site names from the database.

CCAFS LC-40 was renamed CCAFS SLC-40 so the data is treated as the same launch site.

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	Payload_Mass_KG	Orbit	Customer	Mission_Outcome	Landing_Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

First five entries with launch site beginning with ‘CCA’. As noted on the previous slide, despite the change in launch site name, both CCAFS SLC-40 and CCAFS LC-40 are the same launch site.

Total Payload Mass

```
%sql select sum(payload_mass__kg_) as sum from SPACEXDATASET where customer like 'NASA (CRS)'
```

SUM
45596

This is the total payload mass in kg that SpaceX launch for NASA.

CRS stands for Commercial Resupply Services, meaning these launches resupplied the International Space Station

Average Payload Mass by F9 v1.1

```
%sql select avg(payload_mass__kg_) as Average from SPACEXDATASET where booster_version like 'F9 v1.1%'
```

average
2534

This query calculated the average payload mass carried by launches using the F9 v1.1 booster. It should be noted that this payload mass is in the lower portion of the payload mass range.

First Successful Ground Pad Landing Date

```
%sql select "Date" from SPACEXTBL where "Landing_Outcome" like "%Ground Pad%" limit 1;
```

Date
22-12-2015

This returns the first successful landing on a ground pad on December 22, 2015.

The first successful landing was in 2014.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select booster_version from SPACEXDATASET where (mission_outcome like 'Success')
AND (payload_mass__kg_ BETWEEN 4000 AND 6000) AND (landing__outcome like 'Success (drone ship)')
```

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

This query shows the only four versions of SpaceX boosters that have successfully landing on a drone ship after launching with a payload between 4000 and 6000 kg.

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT mission_outcome, count(*) as Count FROM SPACEXDATASET GROUP by mission_outcome ORDER BY mission_outcome
```

mission_outcome	COUNT
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

This query shows the count of each of the three mission outcomes. It is important to note that this is the mission outcome, not the landing outcome.

SpaceX has a mission success rate of approximately 99%.

Boosters Carried Maximum Payload

```
%sql select "Booster_Version" from SPACEXTBL where "Payload_Mass_KG" = (select max("Payload_Mass_KG") from SPACEXTBL);
```

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

This query shows the different booster versions that launched carrying the max payload mass of 15,600 kg.

Since all booster versions are F9 B5 B10, we can infer that booster version and payload mass are correlated.

2015 Launch Records

```
%%sql select substr("Date",4,2) as "Month", substr("Date",7,4) as "Year", "Landing_Outcome", "Booster_Version", "Launch_Site"
from SPACEXTBL where ("Landing_Outcome" like "Failure%Drone%") and (substr("Date",7,4) = '2015');
```

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

This query shows the month, year, landing outcome, booster version, and launch site for launches that failed to land on a drone ship in 2015.

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

```
%%sql select "Landing_Outcome", count("Landing_Outcome") as "Number of Landings"
from SPACEXTBL where ("Landing_Outcome" like "Success (%") and ("Date" between '04,06,2010' and '20-03-2017')
group by "Landing_Outcome" order by count("Landing_Outcome") desc;
```

Landing_Outcome	Number of Landings
Success (drone ship)	8
Success (ground pad)	6

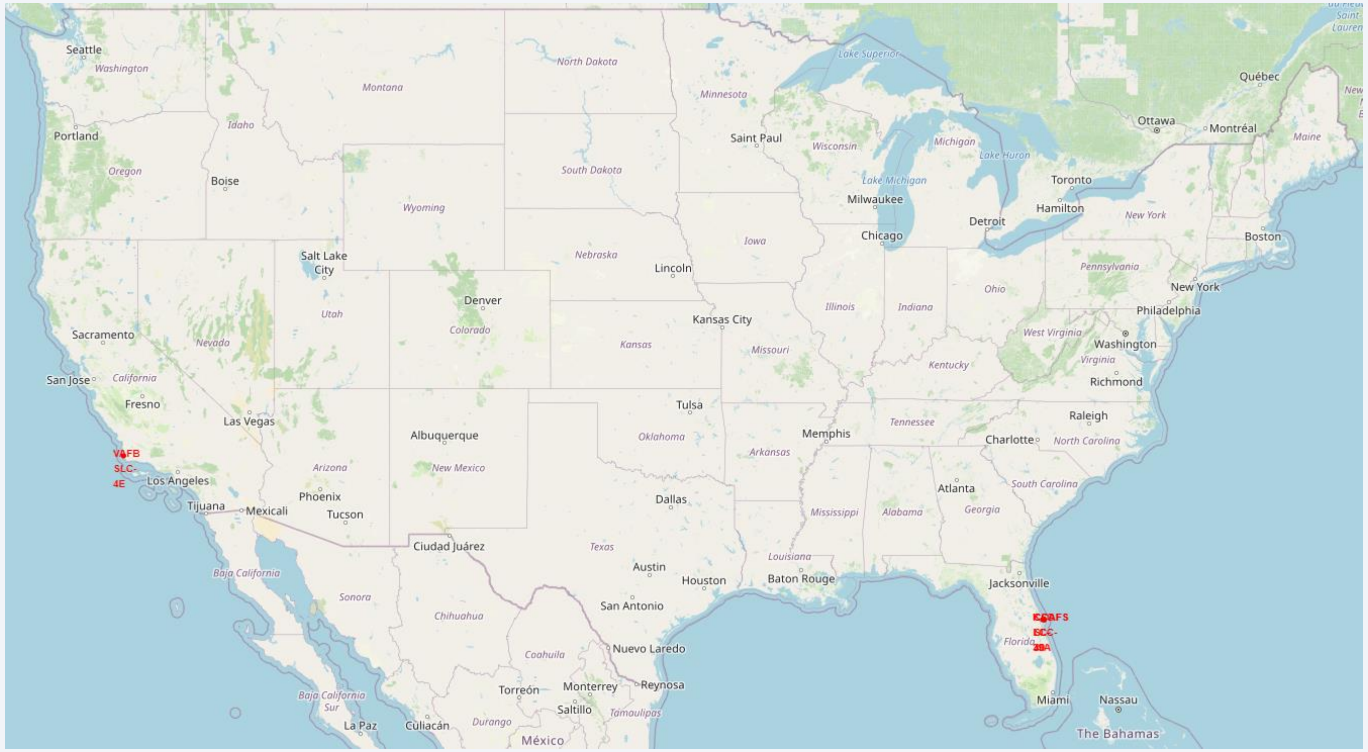
This query shows the count of successful landings and type of landing between 06/04/2010 and 03/20/2017.

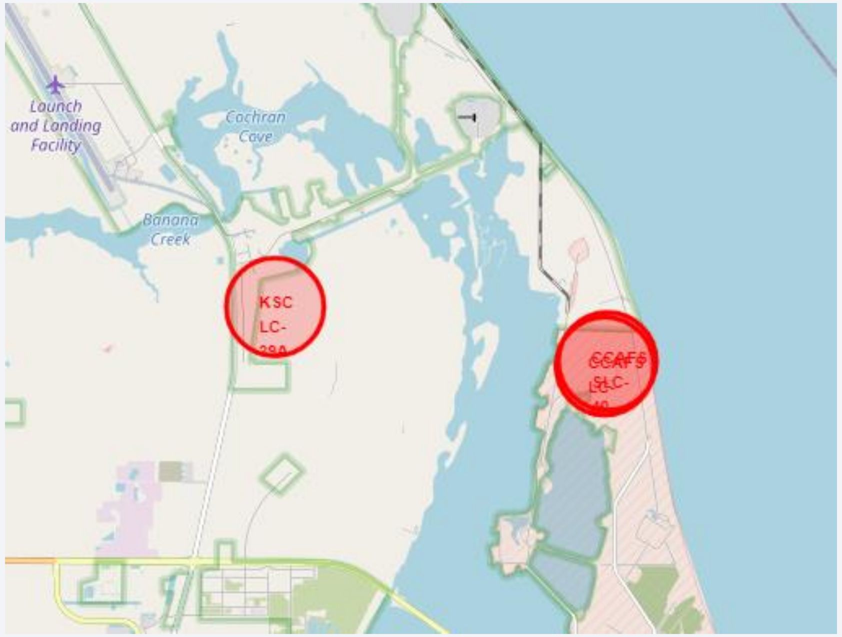
A total of 14 successful landings between drone ships and ground pads.

Section 3

Launch Sites Proximities Analysis

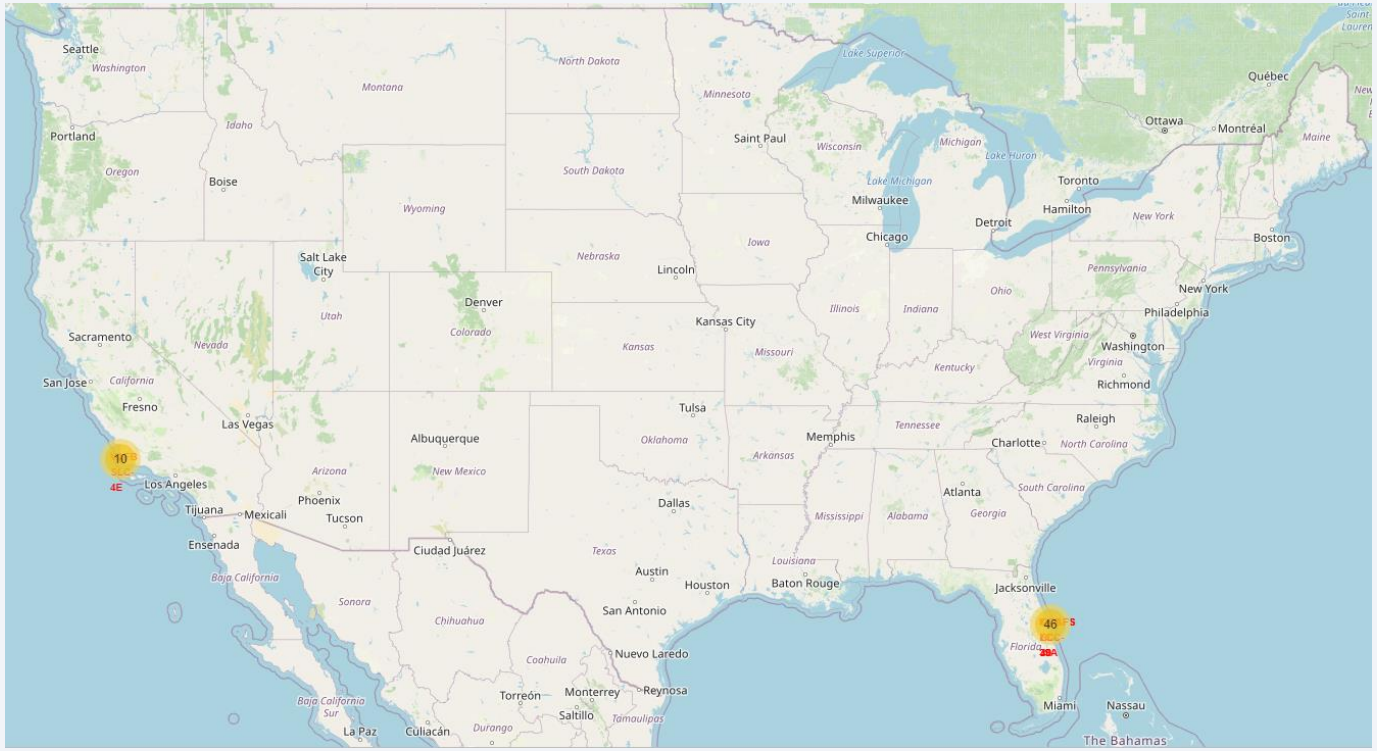
Launch Site Locations

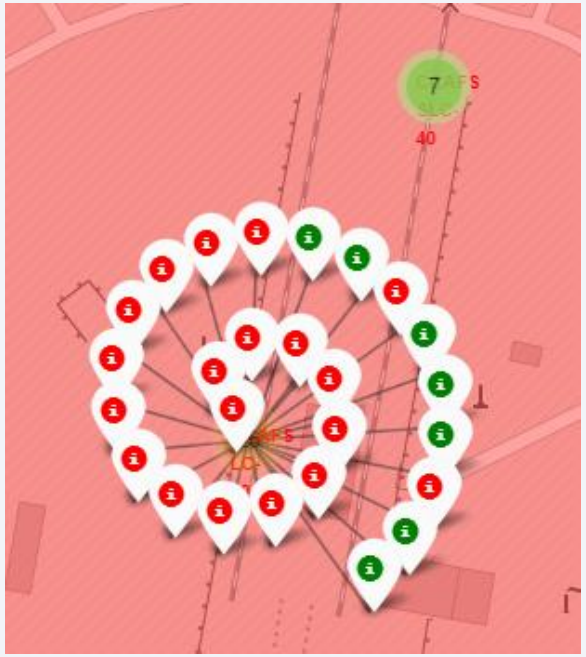




The left map shows all three launch sites in the US. The right map shows the two launch sites on the coast of Florida given their proximity to each other. The only notable characteristics about the launch sites is their distance to an ocean, likely a safety feature in case a launch falls back to Earth.

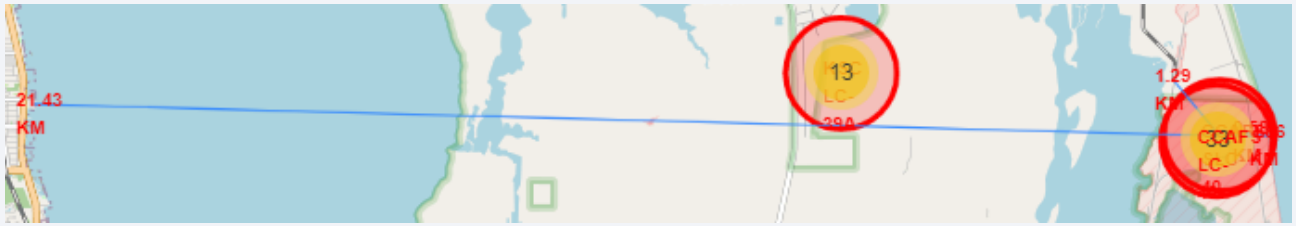
Color-Coded Launch Markers



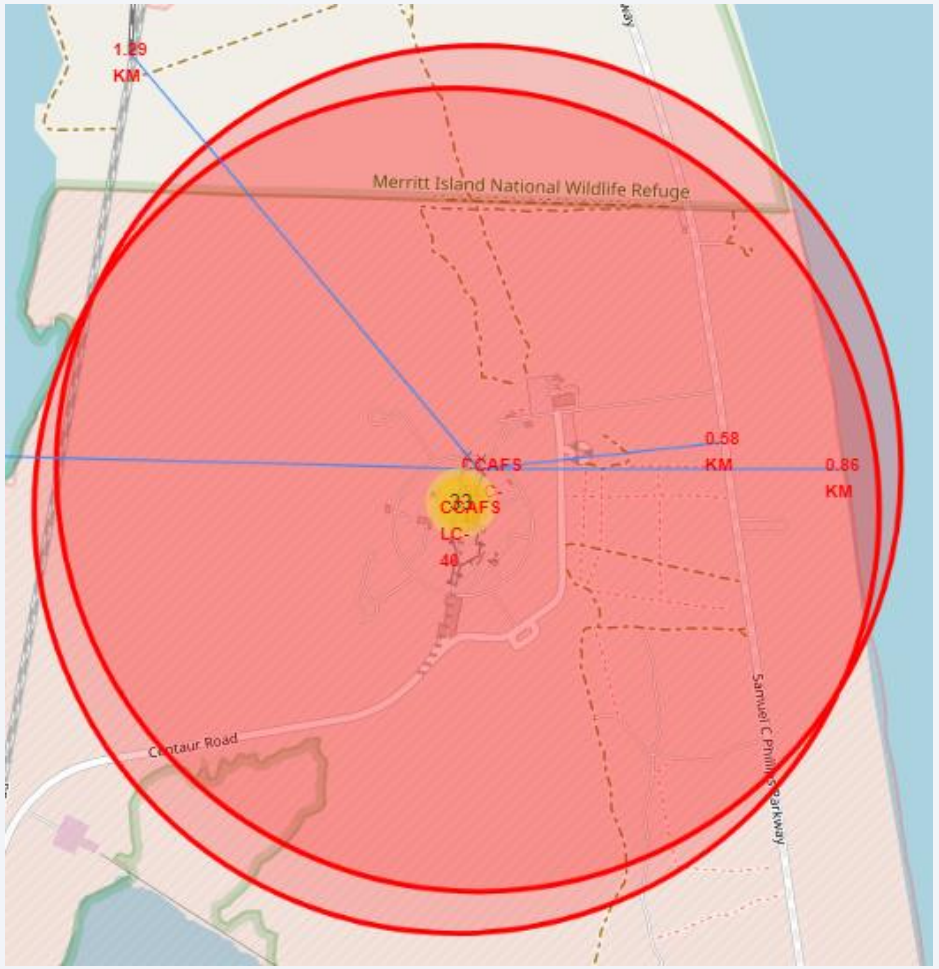


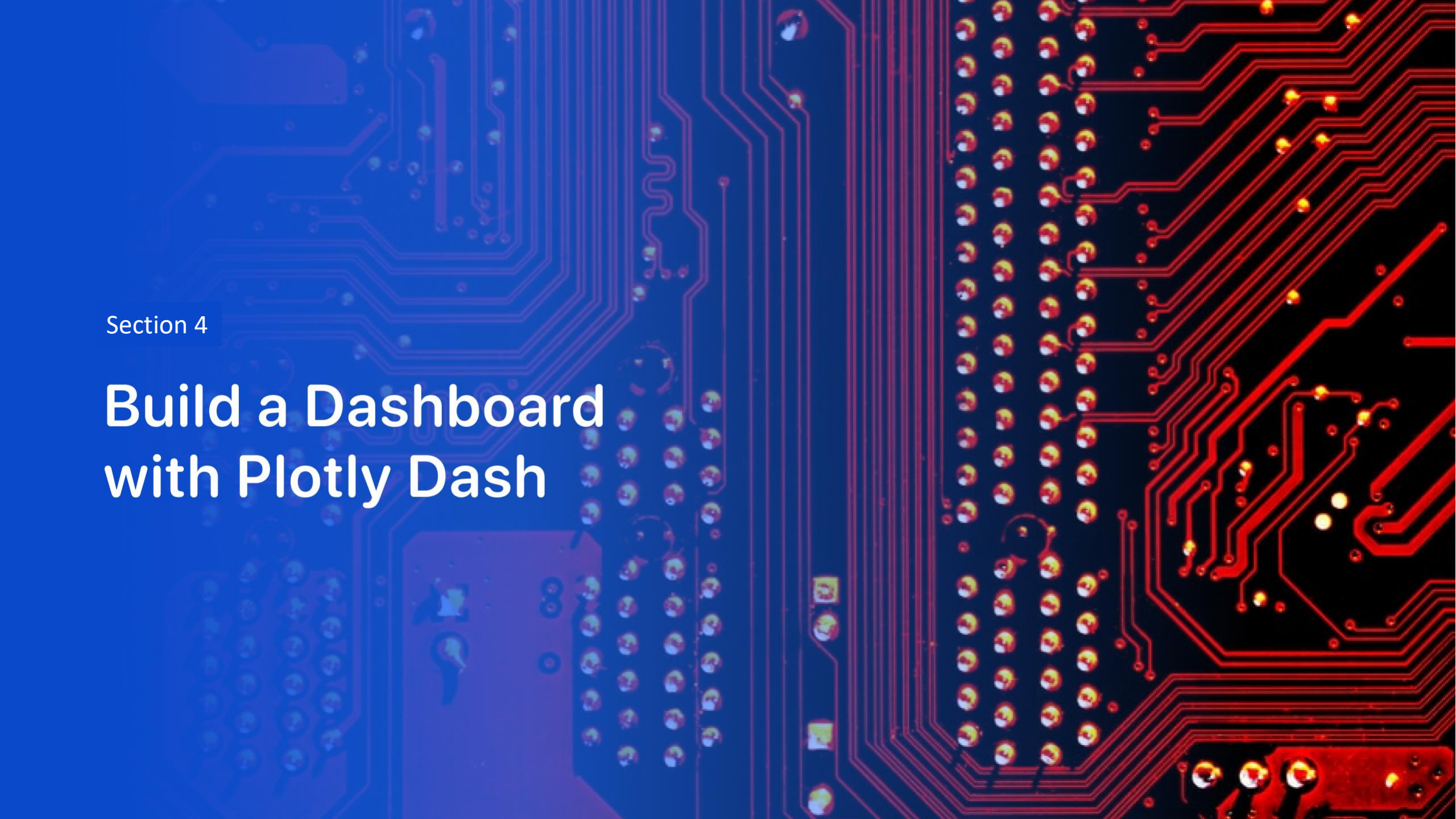
The left map shows all three launch sites in the US with their new marker clusters. The right map shows a close up of the marker cluster around the pre-renaming of CCAFS LC-40. The red markers signify a failed launch and green signifies a successful launch.

Proximities to CCAFS SLC-40

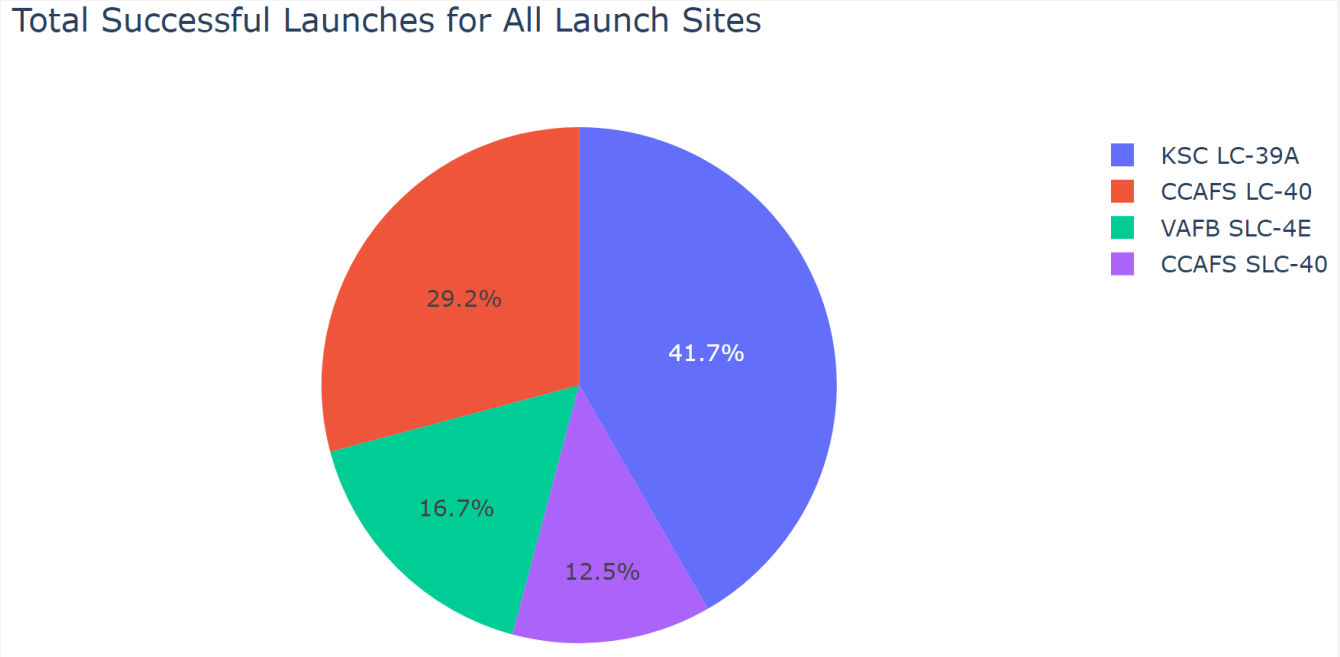


Using CCAFS SLC-40 as an example, we can see the large distance between launch sites and cities, but shorter distances to highways, coasts, and railroads. Highways and railroads are close to provide easy access for large parts and personnel. The ocean is near, and the closest city is far, in case a launch fails and can crash into the ocean to avoid populated areas.



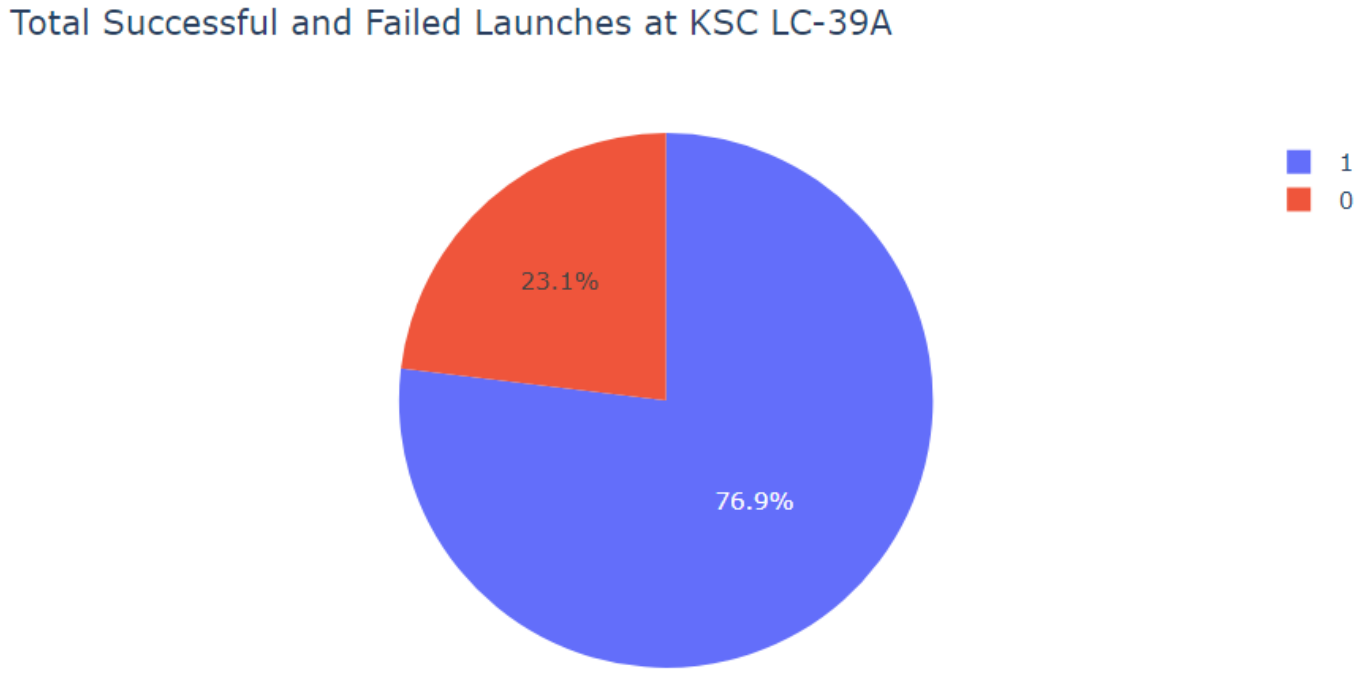

Section 4
Build a Dashboard with Plotly Dash

Successful Launches by Launch Site



This pie chart shows the proportion of all successful launches attributed to each launch site. It is important to note that CCAFS has the same percent of successful launches as KSC, since the name change splits the data. VAFB has the smallest percent of successful launches, likely due to a smaller number of launches and launches on the west coast are more difficult than the east coast.

Launches at KSC LC-39A



Purple signifies a successful launch and red signifies a failed launch.

KSC LC-39A has the highest rate of successful launches with 10 of the 13 launches being successful.

Payload Mass vs. Success with Booster Version Category

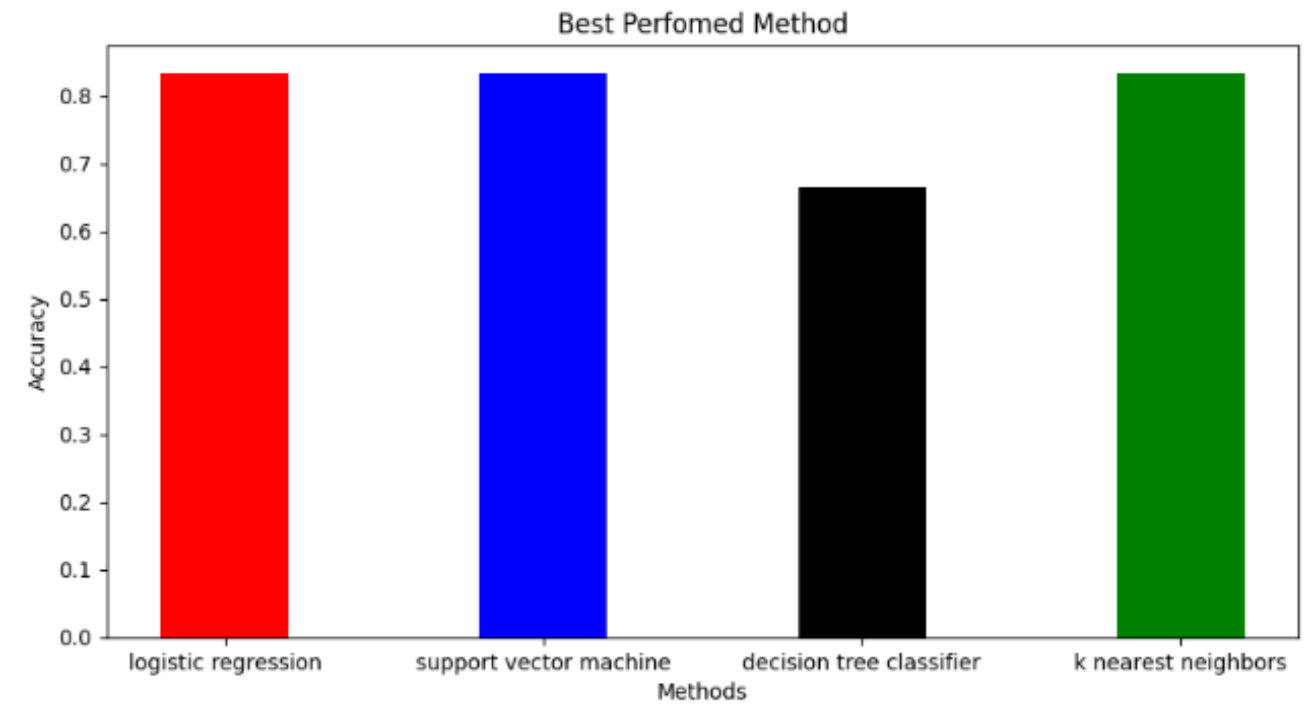
Class 1 signifies a successful launch and class 0 signifies a failed launch.

This scatter plot shows payload mass against launch success, with the color of the data signifying the category of the booster version used on the launch. Most success is found below 6000 kg. However, there are very few launches with payload mass greater than 6000 kg.

Section 5

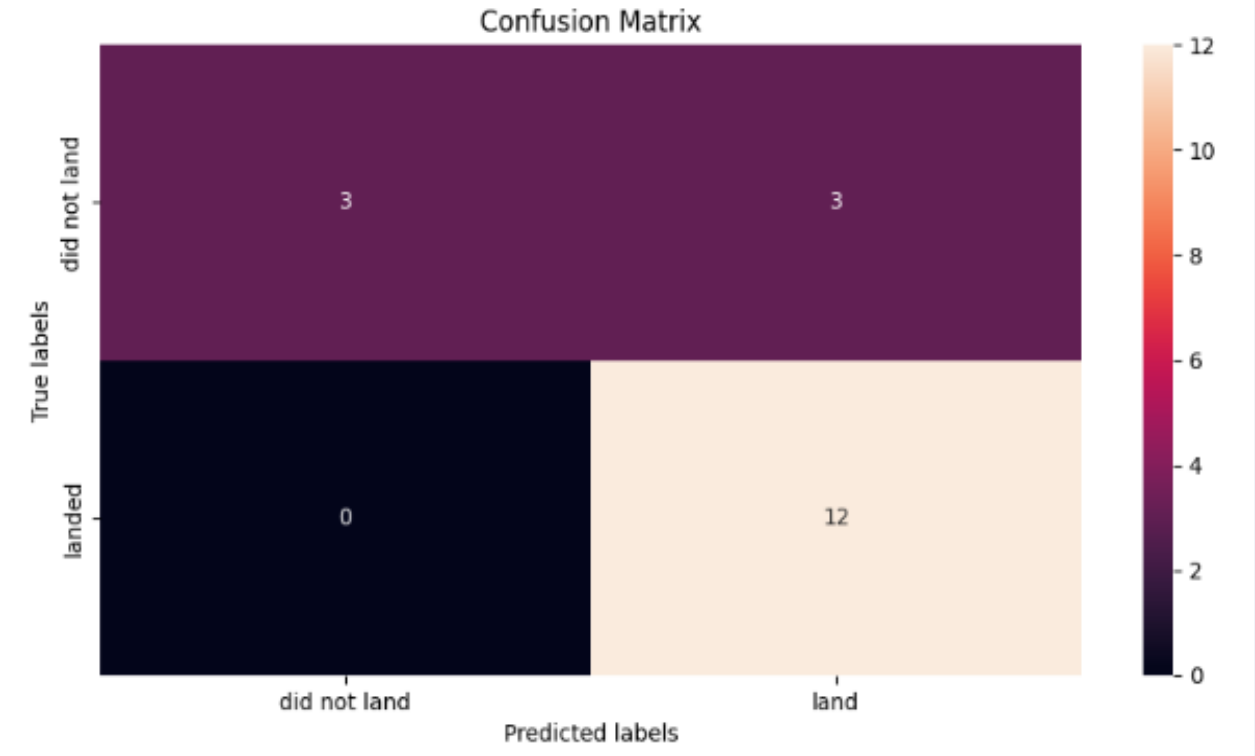
Predictive Analysis (Classification)

Classification Accuracy



The decision tree had the lowest accuracy at 66.67% while logistic regression, support vector machine, and k-nearest neighbors all had an accuracy at 83.34%. However, the sample size is small, only 18 launches. This causes a large variance in results if the decision tree classifier were to be run repeatedly. More data is needed to determine the best model.

Confusion Matrix



The correct predictions are the top left and bottom right.

Logistic regression, support vector machine, and k-nearest neighbors all have the same confusion matrix. They all correctly predicted the 12 successful landings but labeled 3 of the 6 failed landings as successful landings. This tells us the models over-predict successful landings.

Conclusions

- Task: Develop a machine learning model to predict successful landings of SpaceX's Stage 1 rockets.
- Collected data from the SpaceX API and web scraped from the SpaceX Wikipedia page.
- Created and stored data features and label in DB2 SQL database.
- Created a visualization dashboard using Plotly Dash.
- Created three ML models that have 83% accuracy.
- This model can be used to accurately predict if a SpaceX launch will end with a successful Stage 1 landing before launch.
- More data should be collected to better develop the models, improve accuracy, and better determine the best model.

Appendix

- GitHub Repository

Thank you!