IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

S. Colton Crowther
01 MAY 2024
coltoncrowther@gmail.com

Outline

- Abstract
- Introduction
- Methodology
- Results
- Conclusion

Abstract

Getting ahead in the private sector space race will provide a large financial and popular boost. SpaceX is one of the leading private space programs due to their ability to reduce costs by reusing the first stage booster. In order to do that they must recover this booster by controlling its landing on the ground or on a ship-borne landing pad.

SpaceX's launch data was collected from the SpaceX API as well as from Wikipedia. The information was cleaned and formatted for best use and explored numerically and graphically. Launch results were mapped and labeled if successful or not. And finally predictive methods were used to determine the likelihood of future success rates.

The first 3 years (2010-2012) of launches were completely unsuccessful while the project improved from year to year starting in 2013 up until 2020. Most flights were launched from the KSC LC-39A and CCAFS SLC-40 sites in Florida. All launch sites are in coastal regions in the most southern latitudes of the United States with close access to the coast and railroads and safely distanced from more public locations. The v1.1 booster was very unsuccessful while the booster FT was much more successful. All predictive methods resulted in a R^2 score of 0.833 except for the decision tree method which resulted in a score of 0.778. Using classification methods a model was developed with a high level of sensitivity of 100% with a lower level of specificity.

Introduction

Background

Getting ahead in the private sector space race will provide a large financial and popular boost. The main players are SpaceX, Blue Origin, Virgin Galatic, and Rocket Lab. SpaceX is leading this competition due to their ability to reduce costs by reusing the first stage booster. In order to do that they must recover this booster by controlling its landing on the ground or on a ship-borne, drone controlled landing pad.

The booster is programmed to reach the landing pad and carefully set down. This strategy has improved with time but catastrophic landings still occur. This exploration of SpaceX's data attempts to show SpaceX's booster recovery success rate, explores the most successful launch sites, and tries to predict the outcome of future attempted landings.

Section 1

Methodology

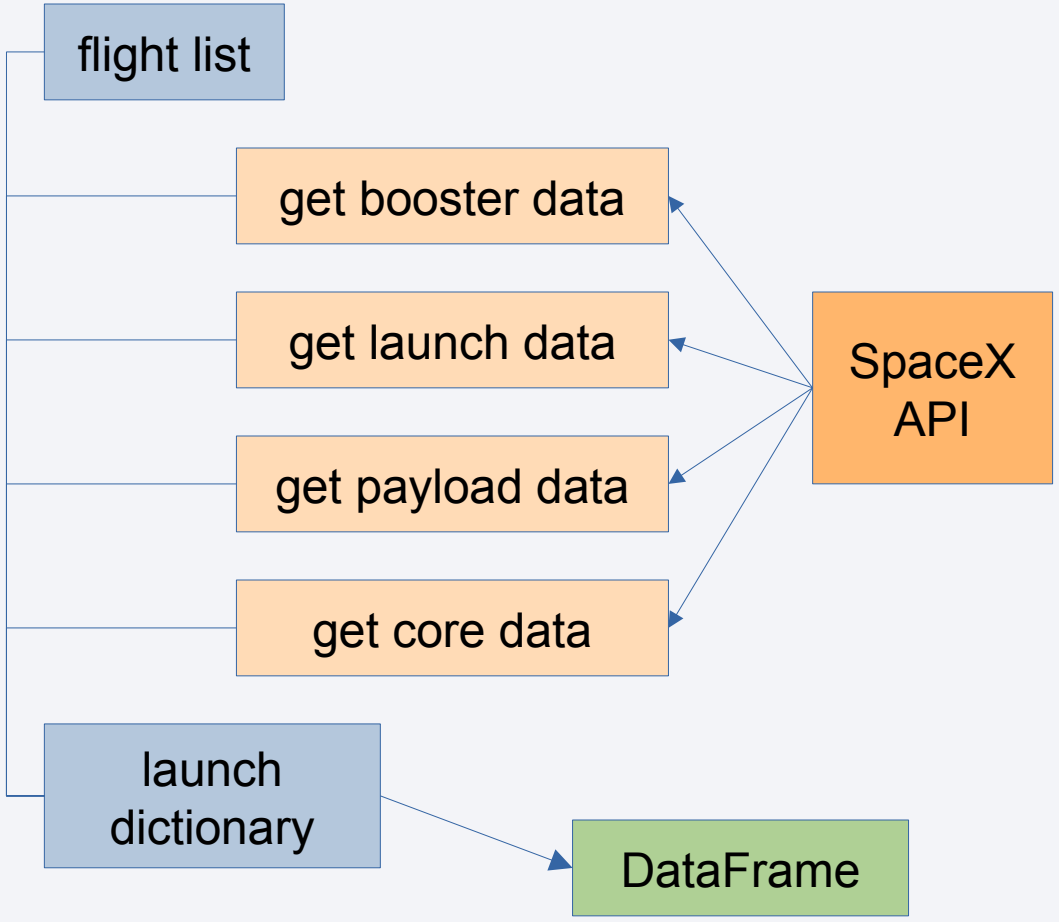
Methodology

Executive Summary

- Data collection
  - The launch and outcome data was collected from the SpaceX API as well as web-scrapping a Wikipedia page that contains pertinent data.
- Data wrangling
  - The data set was limited to Falcon 9 boosters. Missing data was filled in when needed and a data frame was created to handle the data.
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classification models
  - A summary of the predictive methods and their accuracies will be presented.

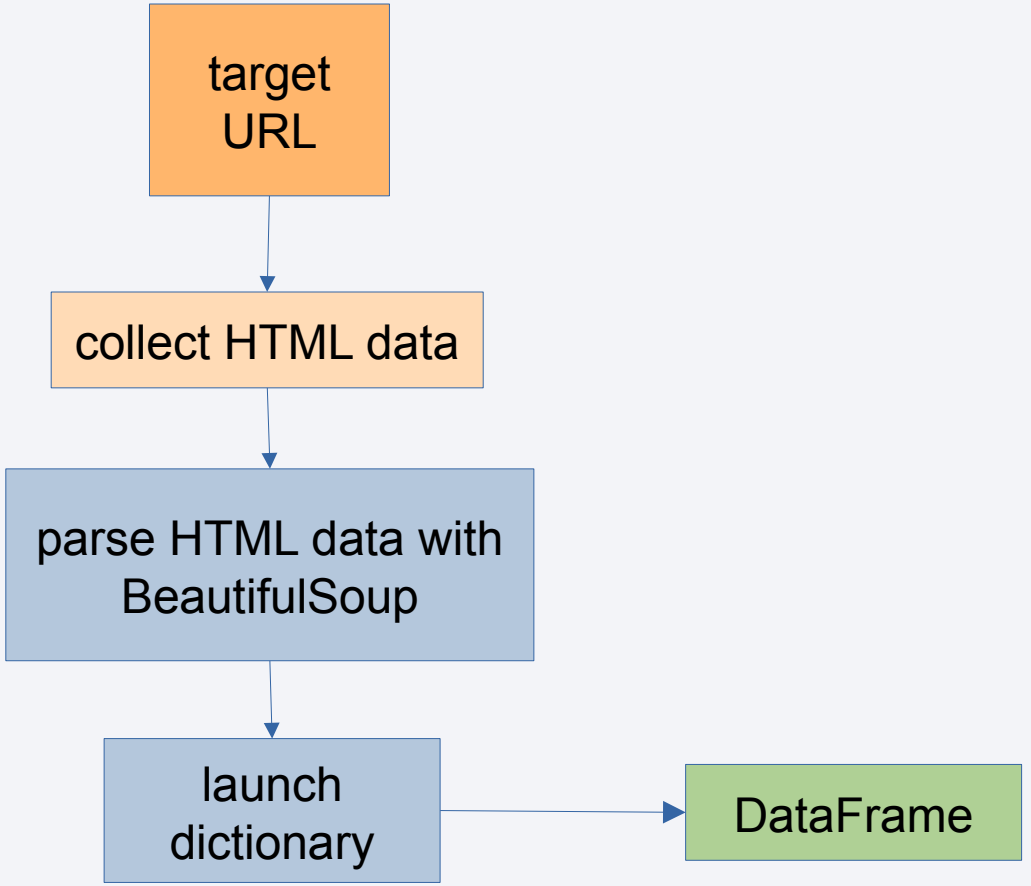
Data Collection – SpaceX API

- Started with a list of SpaceX flights provided by IBM's Skill Network
- Used helper functions to collect the data from the API as lists then combined into a launch dictionary
  - getBoosterVersion: collect booster type
  - getLaunchSite: collect launch site name and location (Lat, Long)
  - getPayloadData: collect payload mass and orbit data
  - getCoreData: collect flight number, landing, and outcomes
- API target for launch data
  - https://api.spacexdata.com/v4/launches/past
- Convert the launch dictionary into a pandas DataFrame
- GitHub notebook URL
  - https://github.com/ojo4f3/Final-IBM-data-science-project/blob/main/spacex-api-data-collection.ipynb



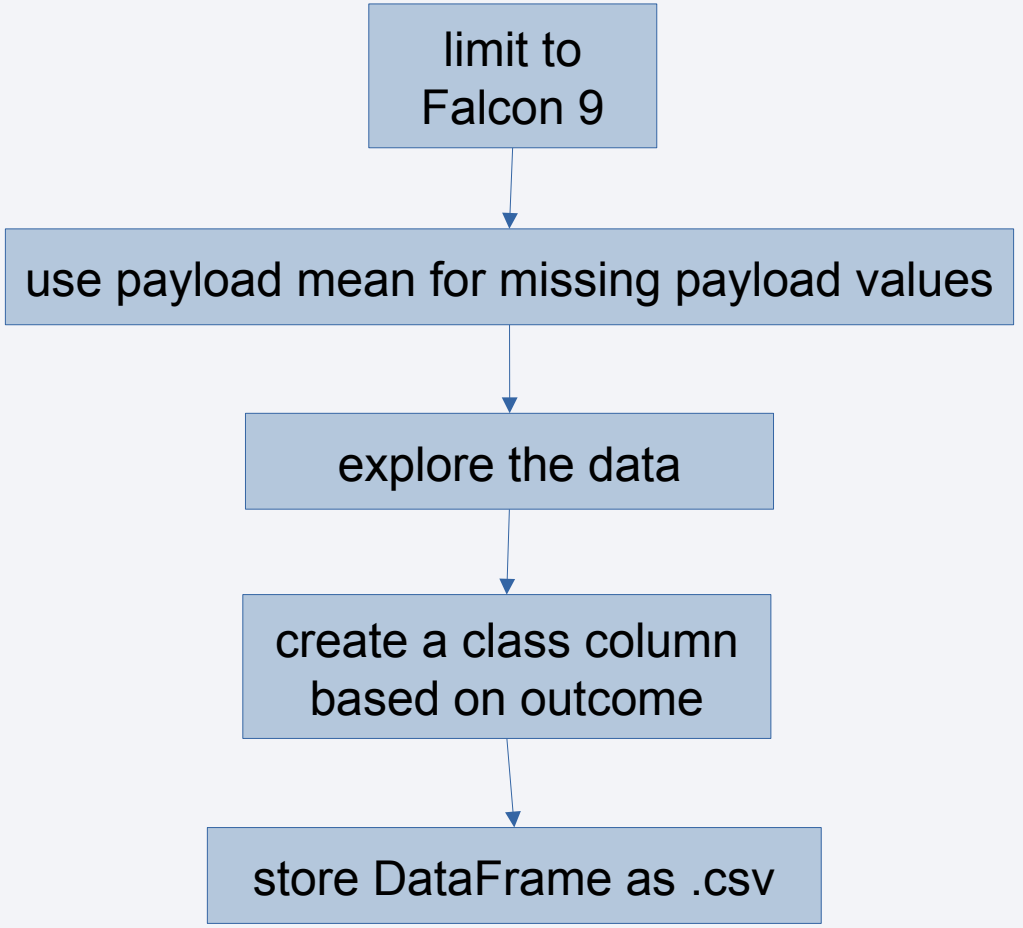
Data Collection - Scraping

- Source website
- https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Used the BeautifulSoup library to parse the HTML data
- Collected launch data from a table
  - Created a list of column names
  - Extracted each row as an entry into a launch dictionary
- Stored the launch dictionary as a pandas DataFrame
- GitHub notebook URL
  - https://github.com/ojo4f3/Final-IBM-data-science-project/blob/main/spacex-web-scraping-data-collection.ipynb



Data Wrangling

- Flights were limited to Falcon 9 launches only
- There were 5 missing Payload Mass values which were assigned the mean Payload Mass of all other flights
- Identified the data type of each column
- Determined the number of launches per:
  - Launch site
  - Orbit
- Tallied number of launch outcomes
- Classified launch outcomes as either a success or failure
- Created a Class column where 1 signifies a successful landing,0 a failed landing
- Average Class outcome of all launches was 0.667
- Stored the data as a csv file
- GitHub notebook URL
  - https://github.com/ojo4f3/Final-IBM-data-science-project/blob/main/spacex-data-wrangling.ipynb



EDA with Data Visualization

- Charts
  - Flight Number vs Launch Site
    - Scatter plot with points colored based on outcome. Shows early vs late flights as well as launch site success.
  - Payload Mass vs Launch Site
    - Scatter plot with points colored based on outcome. Illustrates whether payload mass affected the success of the launch.
  - Orbit Type vs Success Rate
    - Bar graph that shows which orbit types had a higher or lower success rate.
  - Flight Number vs Orbit Type
    - Scatter plot with points colored based on outcome. Shows where early and late launches were sent and how successful different orbit types missions were.
  - Payload vs Orbit Type
    - Scatter plot with points colored based on outcome. Demonstrates the success of specific orbits and if payload mass was a strong factor in a successful outcome.
  - Launch Success Trend
    - Illustrates the success rate over time.
- GitHub notebook URL
  - https://github.com/ojo4f3/Final-IBM-data-science-project/blob/main/spacex-eda-visualization.ipynb

EDA with SQL

- Explored the data to determine:
  - All launch site names
  - First five launches starting with “CCA”
  - Total payload mass launched by NASA
  - Average payload mass for F9 v1.1 boosters
  - First successful ground landing date
  - Names of boosters with successful drone ship landings within a specific mass range
  - Total successes and failures
  - Boosters that carried the maximum payload mass
  - Launches with failed landings during 2015
  - Total landing outcomes by landing site within in a specific date range
- GitHub notebook URL
  - https://github.com/ojo4f3/Final-IBM-data-science-project/blob/main/spacex-eda-sql.ipynb

Build an Interactive Map with Folium

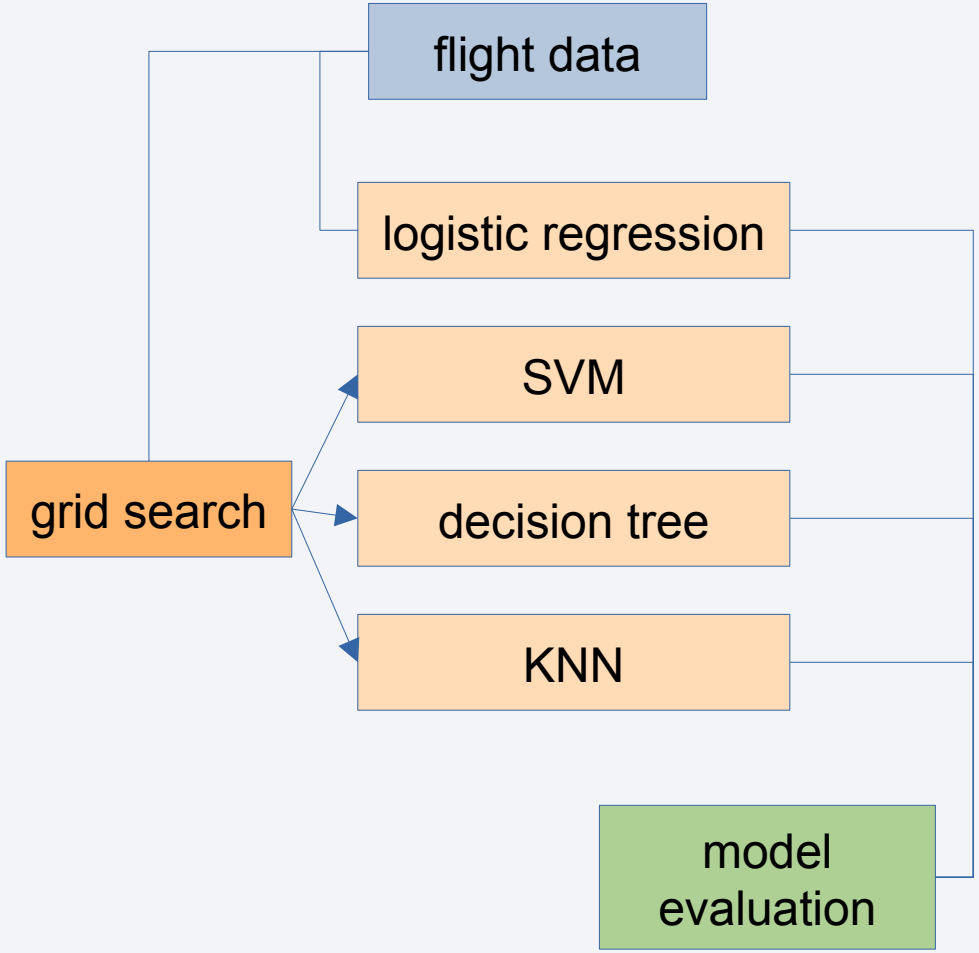
- All launch sites were marked on an interactive map.
  - Visually understand where in the United States the launches took place.
- Added labeled outcomes at each launch site.
  - Quickly get an idea of the success rate at each site as well as which sites had more launches.
- Finally distance lines were added to demonstrate the distance the launch sites tend to have from highways, railroads, and cities.
  - This is to get an idea of what kind of location is suitable for a launch site.
- GitHub notebook URL
  - https://github.com/ojo4f3/Final-IBM-data-science-project/blob/main/spacex-launch-site-map.ipynb

Build a Dashboard with Plotly Dash

- Created a pie chart that shows the success rate for each launch site.
  - Easily see which launch sites are successful and which tend to have more failures.
- Added interactivity to the pie chart.
  - This allows the user to select specific site to visualize the success and failure rate of that specific site.
- Finally, a scatter plot was created that shows payload mass compared to landing outcome. A slider line is included to limit which launches to include given their payload mass.
  - The user can limit launches based on payload mass to get a more focused picture of launch outcomes.
- GitHub URL
  - https://github.com/ojo4f3/Final-IBM-data-science-project/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Multiple classification strategies were used to find the most accurate model to predict future launch outcomes.
- Grid search was used with the support vector machine, decision tree, and K-nearest neighbor methods.
- Confusion matrices and R^2 scores were used to determine which model was the most accurate.
- GitHub notebook URL
  - https://github.com/ojo4f3/Final-IBM-data-science-project/blob/main/spacex-landing-prediction.ipynb



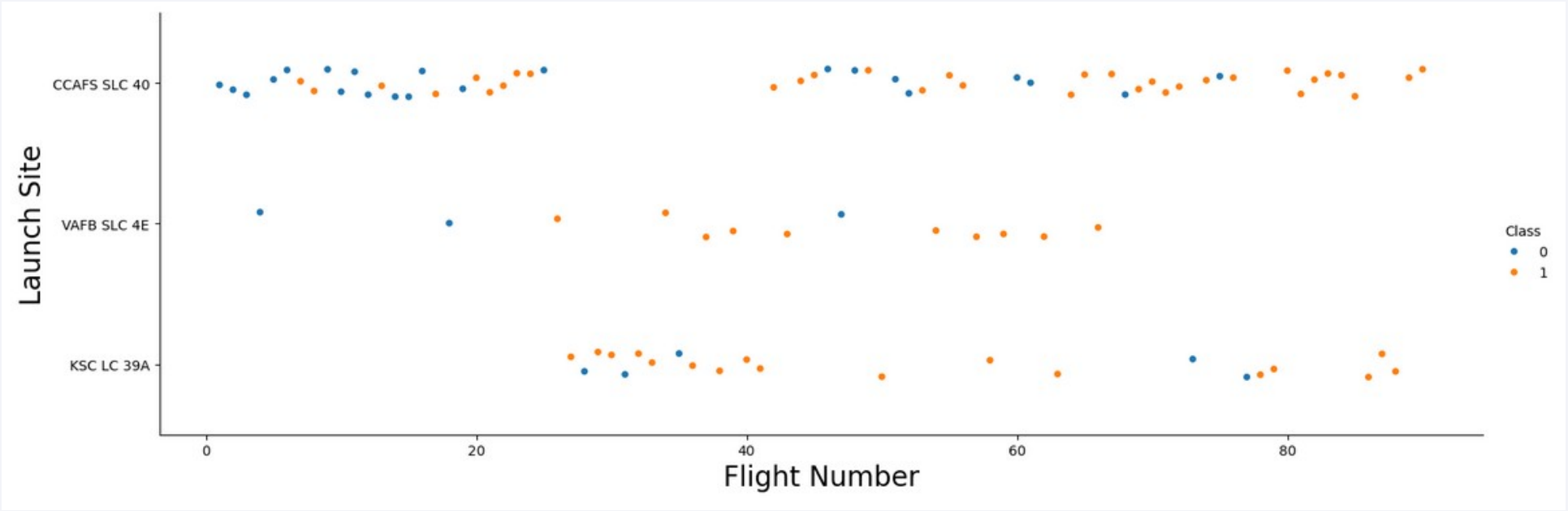
Results

- The first 3 years (2010-2012) of launches were completely unsuccessful
- The SpaceX project improved from year to year starting in 2013 up until 2020
  - There was a significant dip in success during 2018
- Most flights were launched from the KSC LC-39A and CCAFS SLC-40 sites in Florida
  - KSC LC-39A is the most successful launch site, in terms of recovering the booster rocket
  - CCAFS SLC-40 has the lowest success rate of all launch sites
- Launches with higher payload mass had greater success compared to lower mass launches
  - Not many launches between 8,000 – 14,000 kg range
- All launch sites are in coastal regions in the most southern latitudes of the United States
  - The VAFB SLC-4E site is 1.39 km away from the coast
  - The site is even closer to the railroad which it likely uses to move rocket components
  - Public structures are farther from the site likely for safety like highways and cities
- The v1.1 booster was very unsuccessful while the booster FT was much more successful
- All predictive methods resulted in a R^2 score of 0.833 except for the decision tree method which resulted in a score of 0.778
  - The model accurately labels each successful landing with a high sensitivity of 100%
  - The model labeled 3 launches that did not land as landing successfully which are false positives
    - The model's specificity is 50%

Section 2

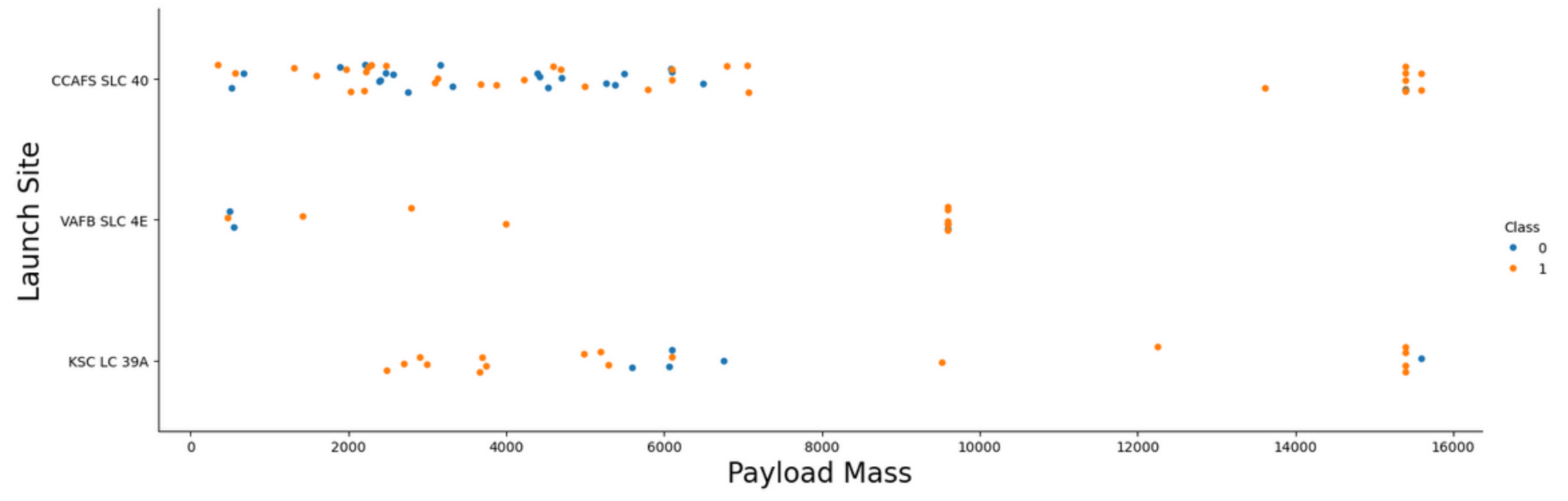
Insights drawn from EDA

Flight Number vs. Launch Site



The scatter plot indicates that most flights were at the CCAFS SLC 40 and KSC LC 39A sites. It also demonstrates that later launches were more successful than earlier launches.

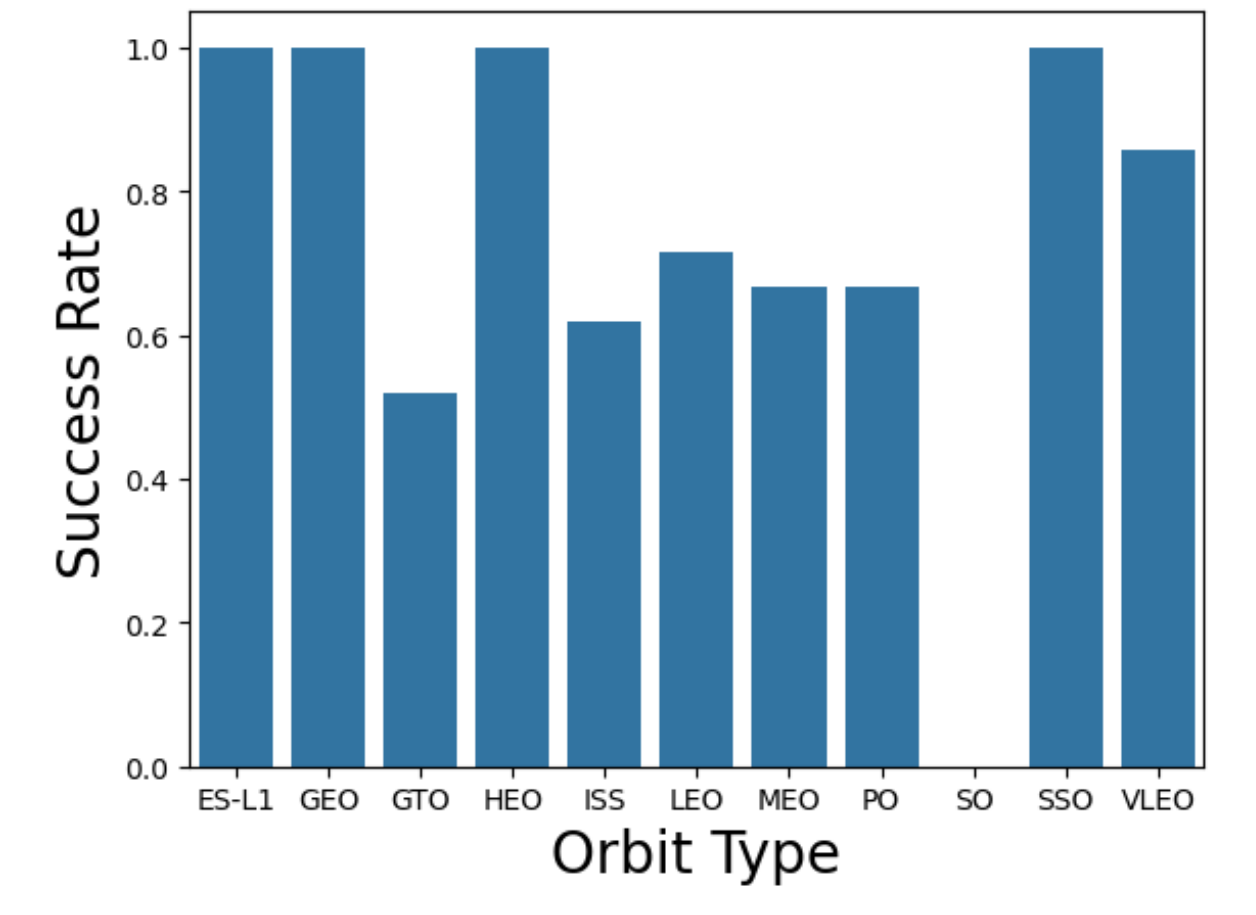
Payload vs. Launch Site



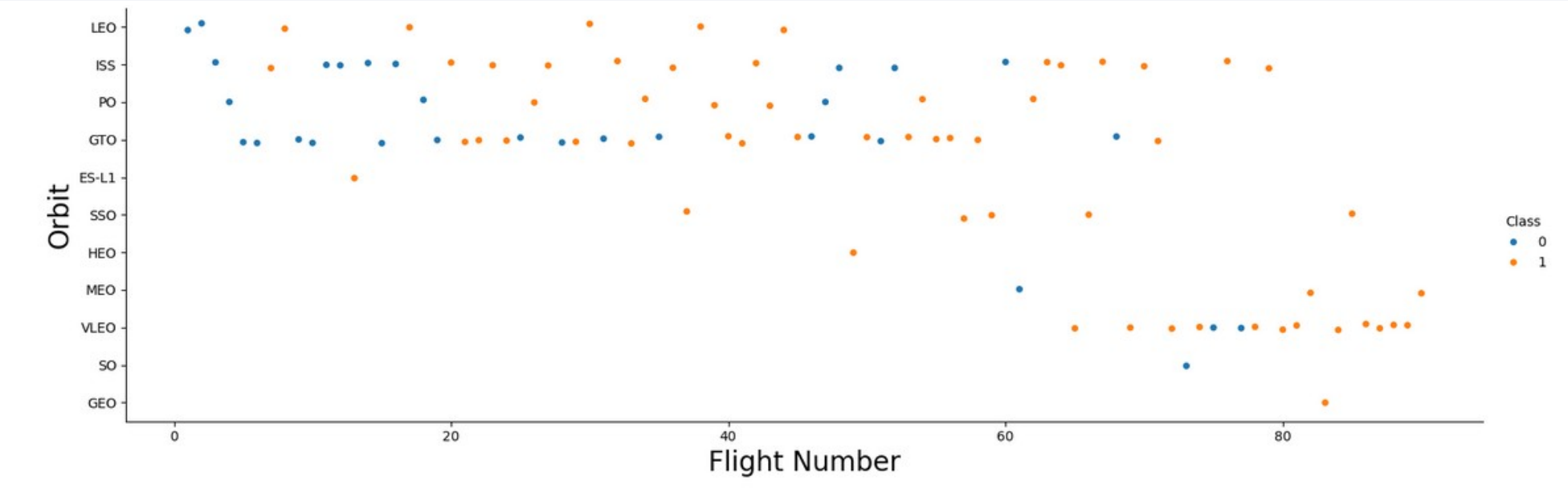
There were very few launches around the 8,000-14,000 kg mass range. Although there are less of them the higher mass launches had a higher success rate than lower mass launches.

Orbit Type vs. Success Rate

- SpaceX has sent many missions to GTO and ISS. Their success rate is better than 50%.
- LEO and MEO are commonly used for telecommunication satellites Their likely the most used for commercial projects and have a mixed success rate.
- GEO launches were completely successful and can be used for telecommunications and observation. If the satellite can function at a GEO orbit instead of LEO or MEO, it may be advantageous for SpaceX to launch to GEO since their booster recovery rate is 100%

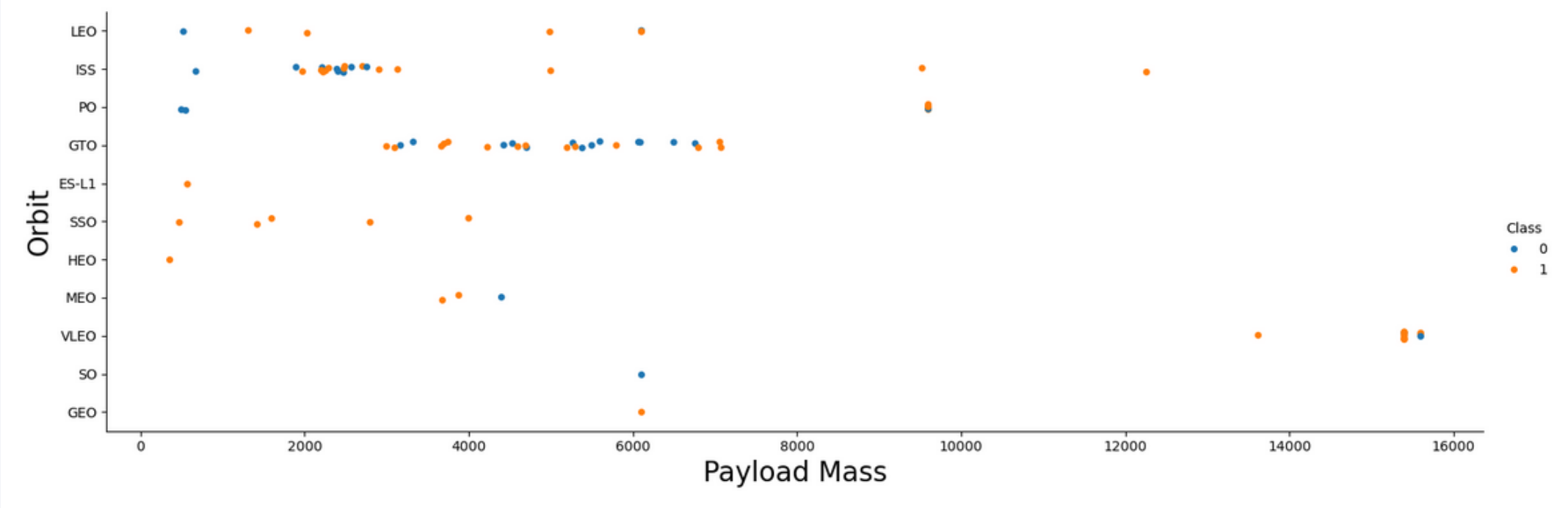


Flight Number vs. Orbit Type



- This figure here shows that although the GEO orbit has a 100% success rate, it has only had one launch.
- This figure also shows that early flights were not sent to HEO, MEO, VLEO, SO, and GEO.

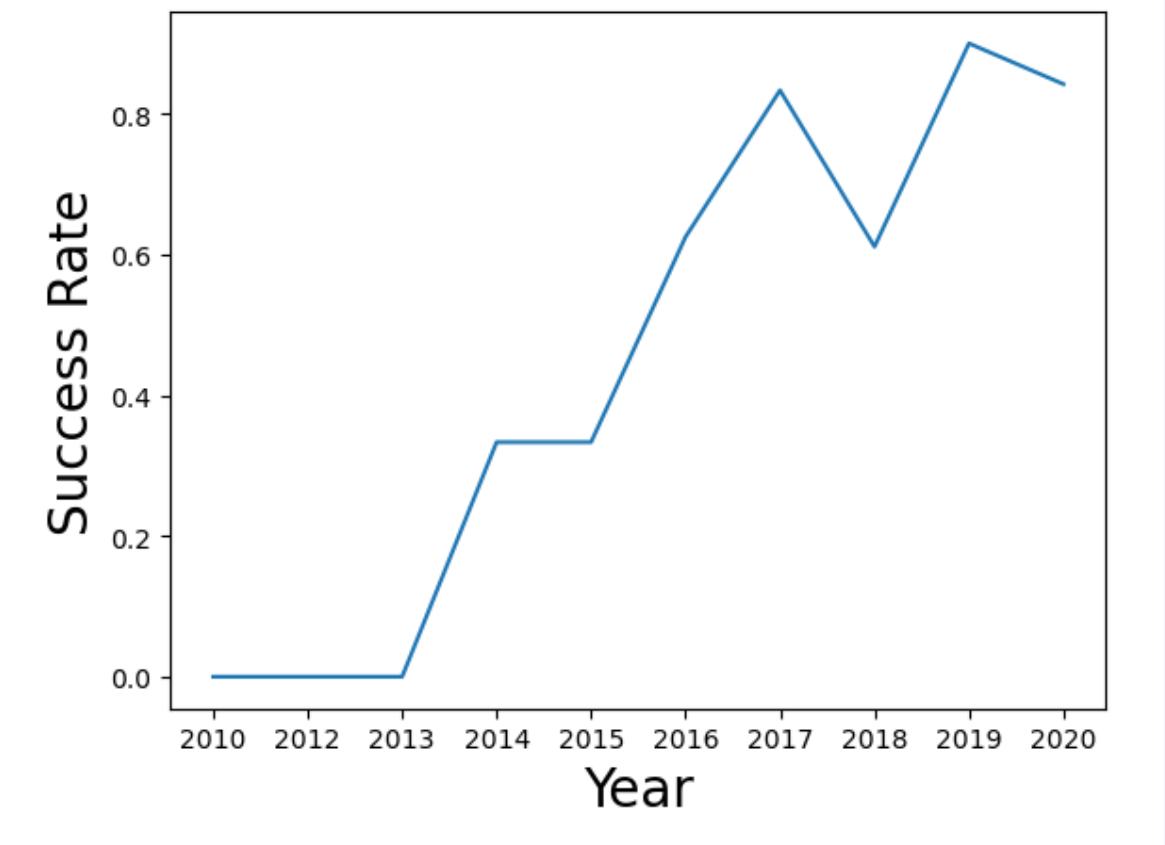
Payload vs. Orbit Type



- Payloads of similar mass were sent to various orbits, especially on the lower payload mass end.
- It also demonstrates there are few heavy payloads compared to lighter payloads.

Launch Success Yearly Trend

- The SpaceX project improved from year to year starting in 2013 up until 2020.
- The first 3 years of launches were completely unsuccessful.
- There is a significant dip in result during 2018.



All Launch Site Names

- Query
  - SELECT DISTINCT("Launch_Site") FROM SPACEXTABLE
- Results
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40
- This collects all launch site names.

Launch Site Names Begin with 'CCA'

- Query
  - SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE "CCA%" LIMIT 5
- Results
  - [('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 query finds the first 5 launches from site starting with ‘CCA’.

Total Payload Mass

- Query
  - SELECT SUM(PAYLOAD_MASS__KG_) as 'Total NASA (CRC) Mass' FROM SPACEXTABLE WHERE Customer == 'NASA (CRS)
- Result
  - 45596
- NASA used SpaceX to carry an accumulated payload of 45,596 kg.

Average Payload Mass by F9 v1.1

- Query
  - SELECT ROUND(AVG(PAYLOAD_MASS__KG_), 2) FROM SPACEXTABLE WHERE Booster_Version LIKE 'F9 v1.1%'
- Result
  - 2534.67
- The average payload mass for a F9 v1.1 booster rocket is 2,534.67 kg.

First Successful Ground Landing Date

- Query
  - SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome == 'Success (ground pad)'
- Result
  - 2015-12-22
- The first successful ground pad landing was on 2015-12-22.

Successful Drone Ship Landing with Payload between 4000 and 6000

- Query
  - SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000
- Results
  - [('F9 v1.1',), ('F9 v1.1 B1011',), ('F9 v1.1 B1014',), ('F9 v1.1 B1016',), ('F9 FT B1020',), ('F9 FT B1022',), ('F9 FT B1026',), ('F9 FT B1030',), ('F9 FT B1021.2',), ('F9 FT B1032.1',), ('F9 B4 B1040.1',), ('F9 FT B1031.2',), ('F9 B4 B1043.1',), ('F9 FT B1032.2',), ('F9 B4 B1040.2',), ('F9 B5 B1046.2',), ('F9 B5 B1047.2',), ('F9 B5B1054',), ('F9 B5 B1048.3',), ('F9 B5 B1051.2 ',), ('F9 B5B1060.1',), ('F9 B5 B1058.2 ',), ('F9 B5B1062.1',)]
- The listed boosters are the boosters that had successful drone ship landings and had a payload between 4,000 and 6,000 kg.

Total Number of Successful and Failure Mission Outcomes

- Query
  - SELECT SUM(Mission_Outcome LIKE 'Success%'), SUM(Mission_Outcome LIKE 'Failure %') FROM SPACEXTABLE
- Result
  - [(100, 1)]
- There was a total of 100 successful mission outcomes and only 1 failed mission.

Boosters Carried Maximum Payload

- Query
  - SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ == (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
- Results
  - [('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 ',)]
- The maximum payload was carried by the above 12 booster rockets.

2015 Launch Records

- Query
  - SELECT SUBSTR(Date,6,2), Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE Landing_Outcome == 'Failure (drone ship)' AND SUBSTR(Date,0,5) == '2015'
- Results
  - ('January', ('Failure (drone ship)', 'F9 v1.1 B1012', 'CCAFS LC-40')),
  - ('April', ('Failure (drone ship)', 'F9 v1.1 B1015', 'CCAFS LC-40'))
- The two launches were the only drone ship failed launches during 2015.

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

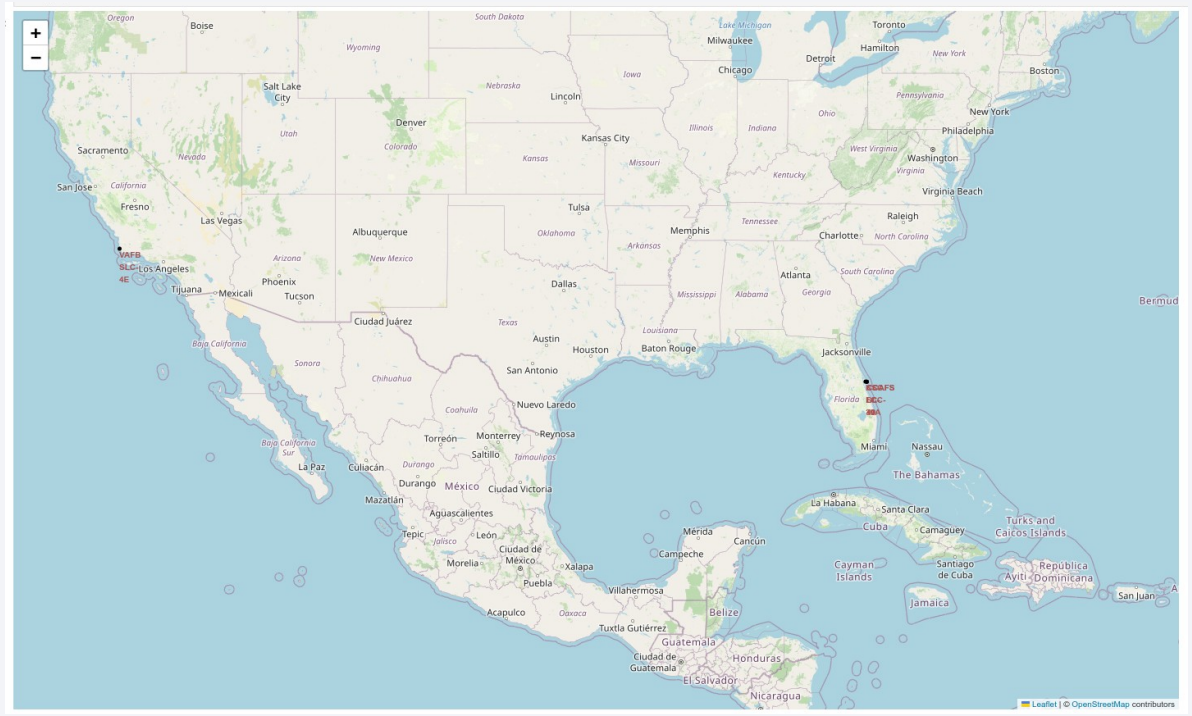
- Query
  - SELECT Landing_Outcome, COUNT(*) AS Count FROM SPACEXTABLE WHERE Date > '2010-06-04' AND Date < '2017-03-20' GROUP BY Landing_Outcome ORDER BY Count DESC
- Results
  - ('No attempt', 10),
  - ('Success (drone ship)', 5),
  - ('Failure (drone ship)', 5),
  - ('Success (ground pad)', 3),
  - ('Controlled (ocean)', 3),
  - ('Uncontrolled (ocean)', 2),
  - ('Precluded (drone ship)', 1),
  - ('Failure (parachute)', 1)]
- The list above demonstrates the number of outcomes for launches between 2010-06-04 and 2017-03-20 with the most common result being ‘No attempt’.

Section 3

Launch Sites Proximities Analysis

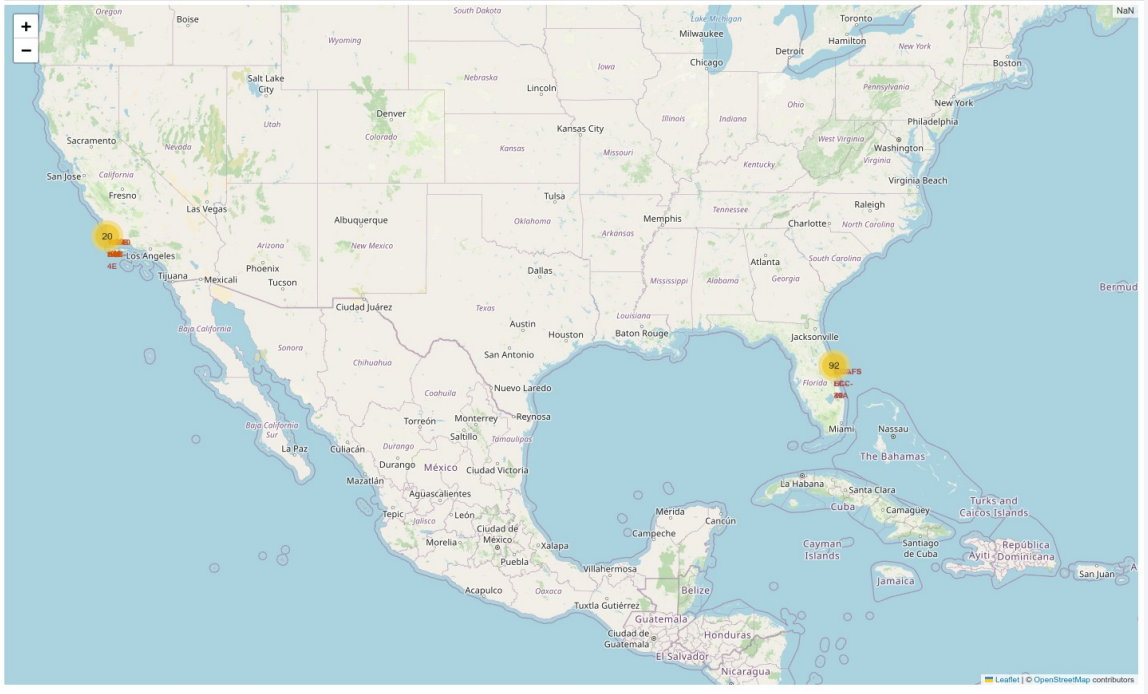
Launch Site Locations

- The image shows each launch location.
- All launch sites are in coastal regions in the most southern latitudes of the United States.



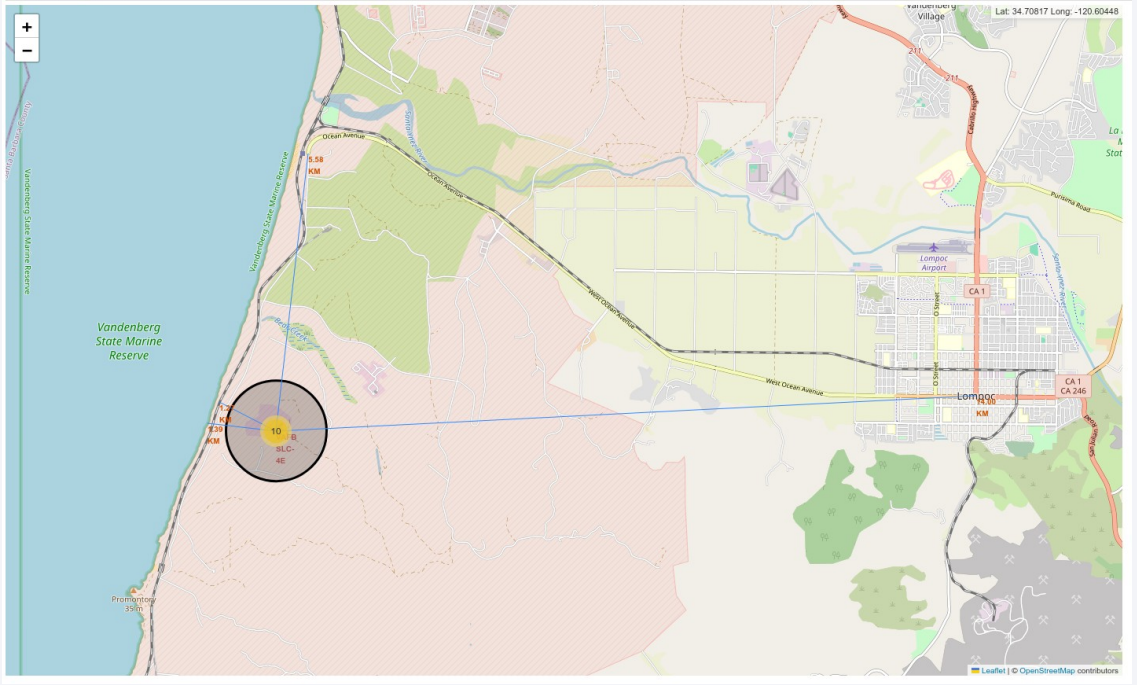
Number & Outcome of Launches

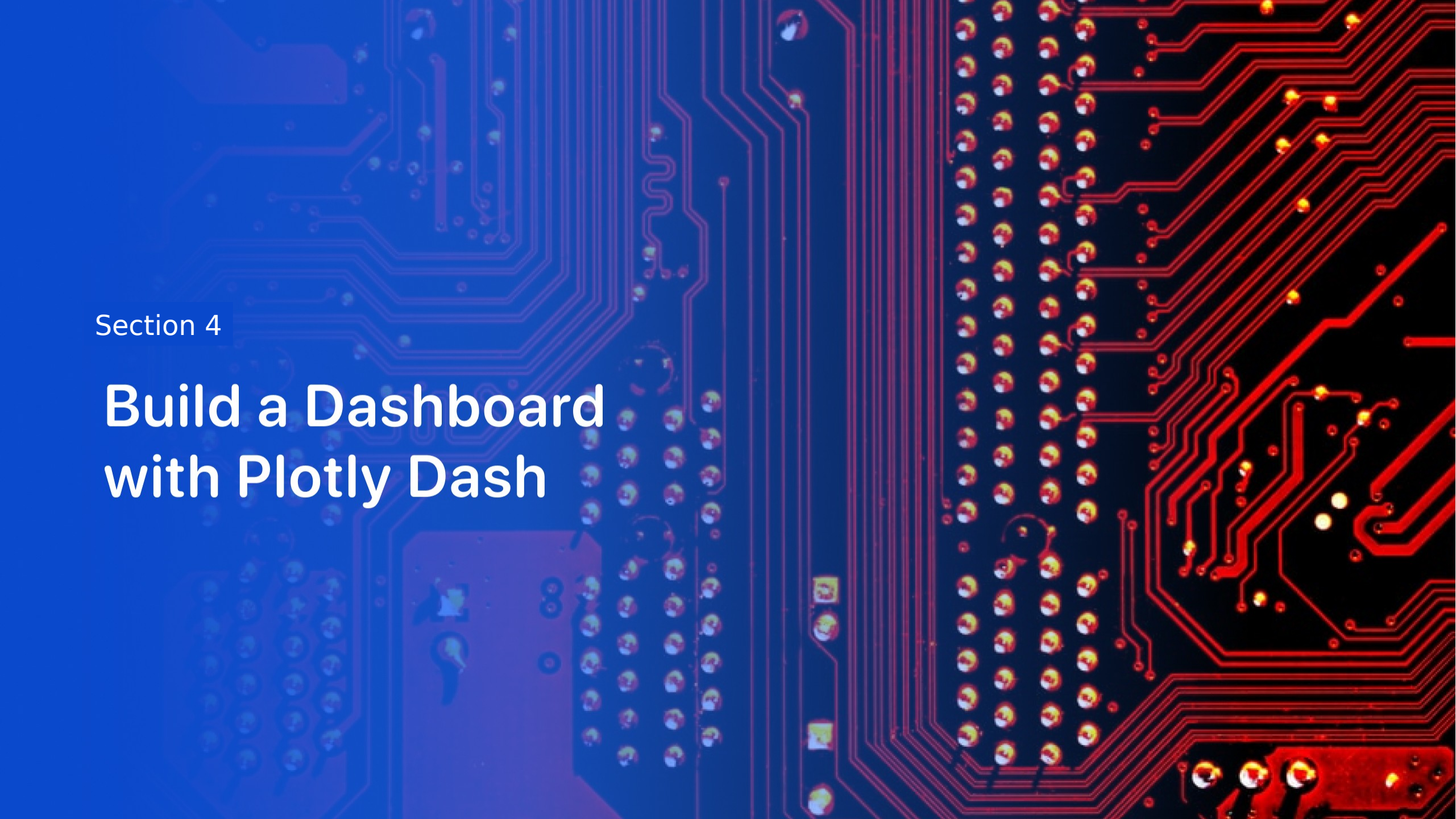
- This image shows the relative number of launches by the size of the circle. The interactive map allows for viewing a pin for each launch and whether it had a successful landing or not.
- A basic understanding of how successful a launch site is can be quickly seen when each pin is viewed.



Launch Site Location

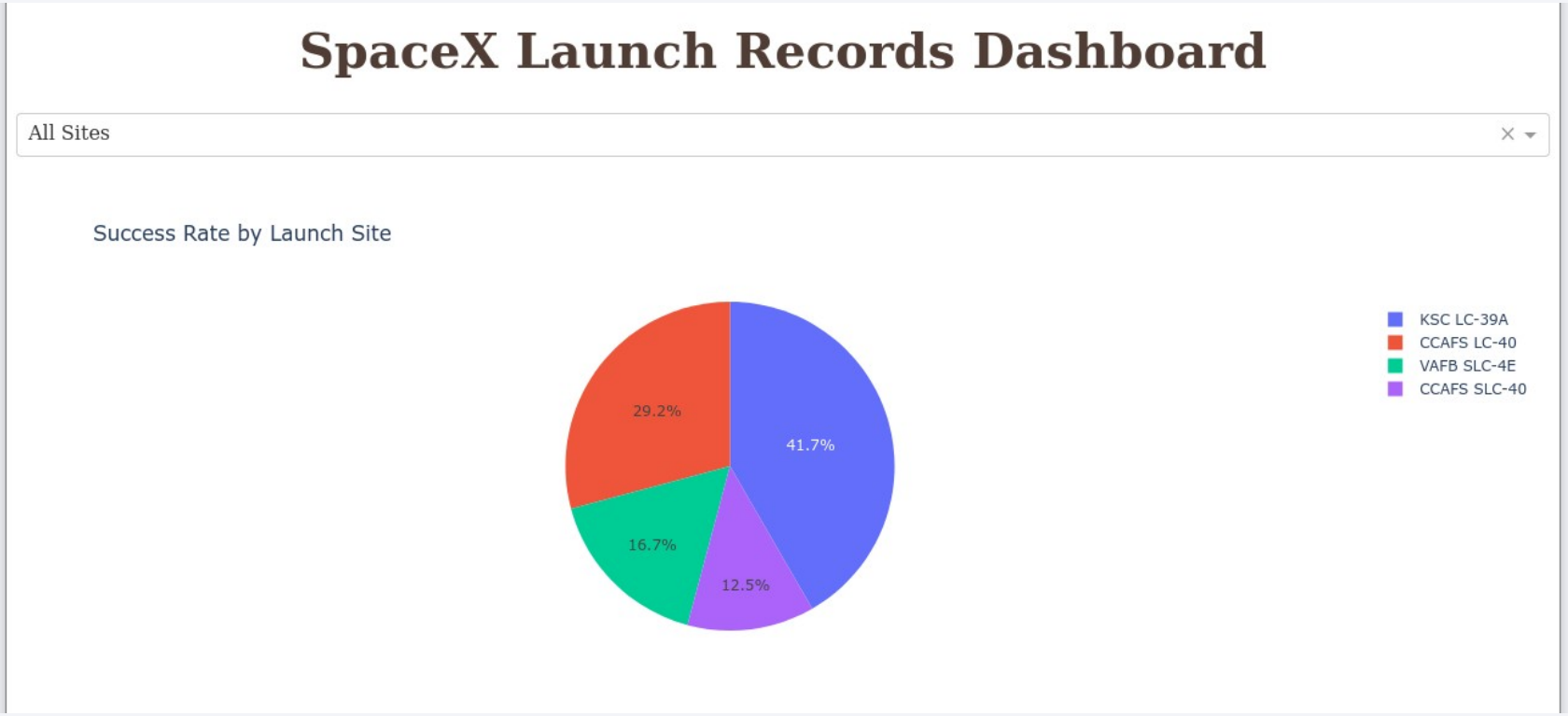
- The VAFB SLC-4E site is very close to the coast, only 1.39 km away.
- It is even closer to the nearest railroad at 1.27 km.
- Farther features include the nearest highway at 5.58 km and the city of Lompoc at 14.0 km.




Section 4
Build a Dashboard with Plotly Dash

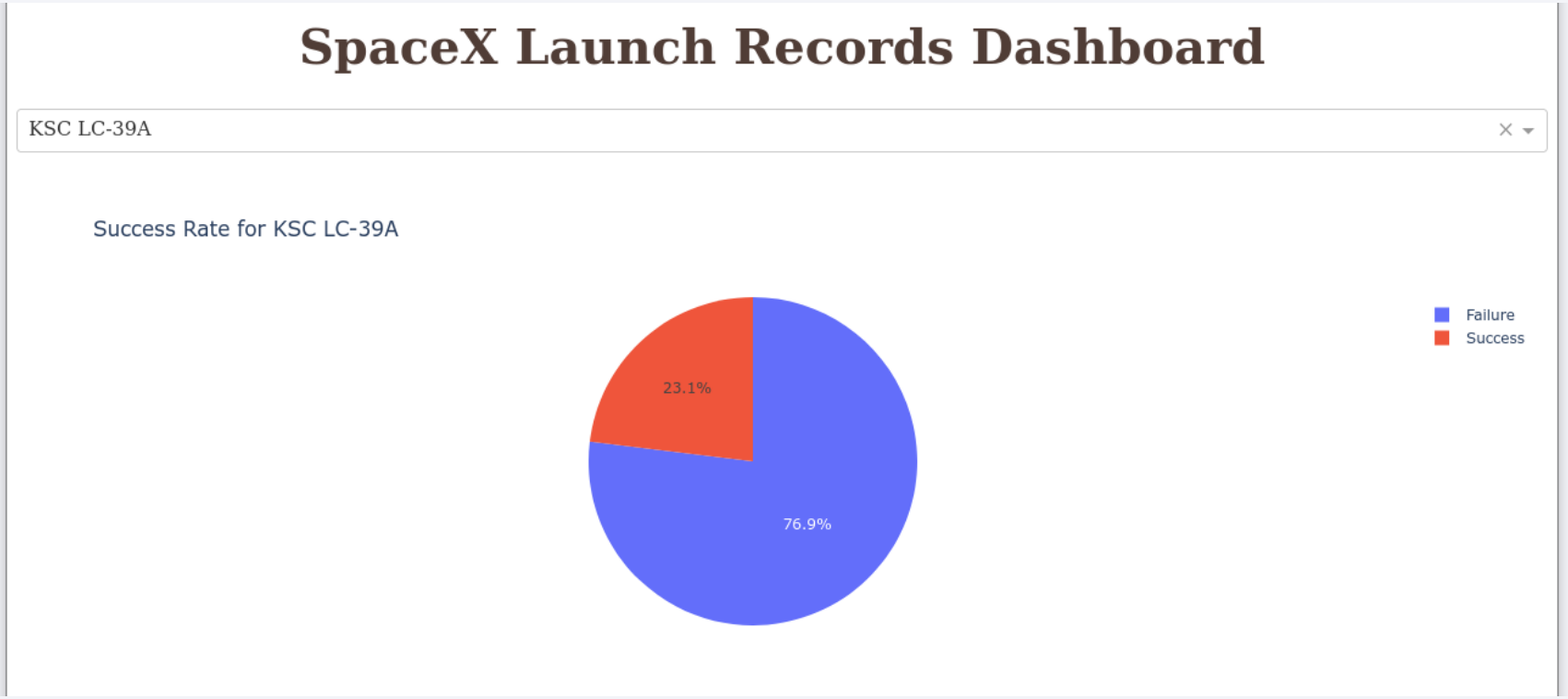
Launch Site Success

- The chart shows that KSC LC-39A is the most successful launch site, in terms of recovering the booster rocket, and CCAFS SLC-40 is the least successful.



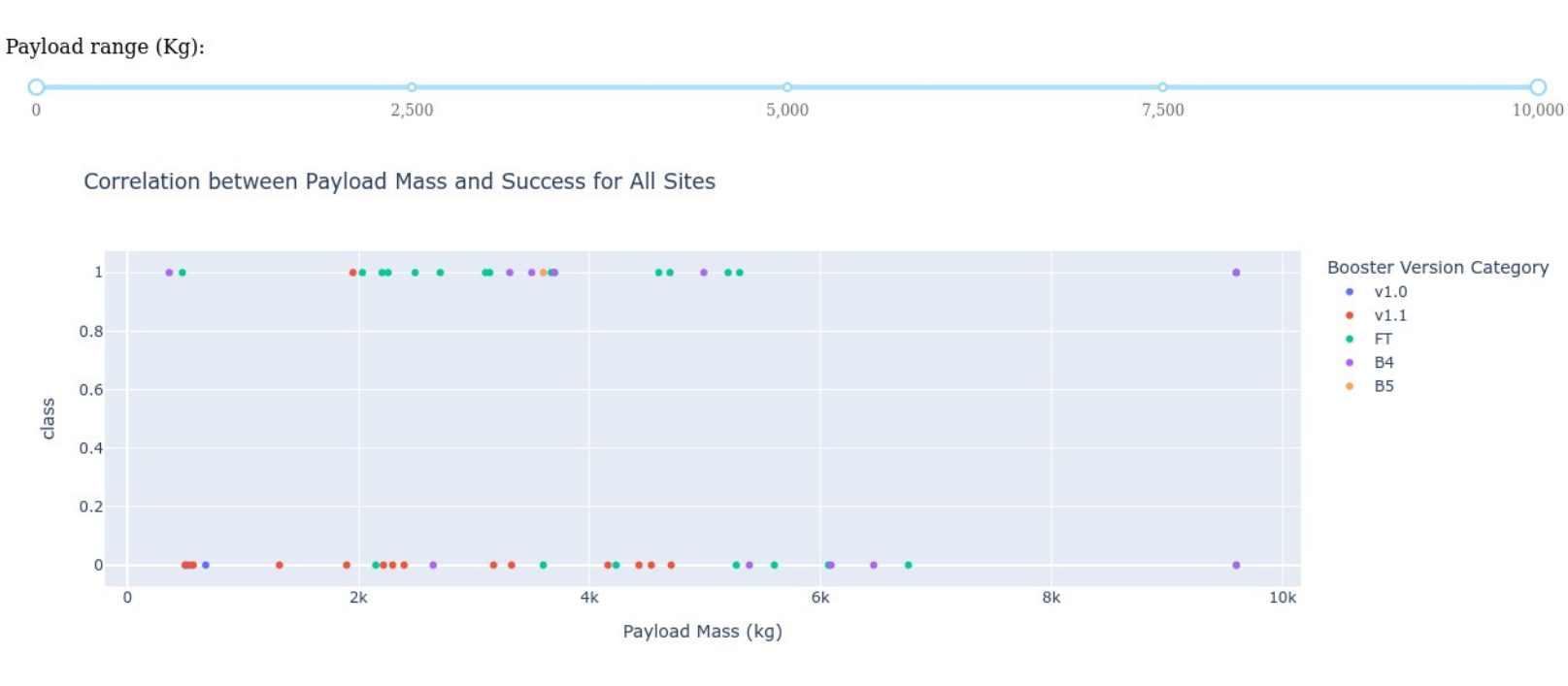
Individual Launch Site

- The user can choose which launch site to explore.
- KSC LC-39A is show below with a success rate of 23.1%.



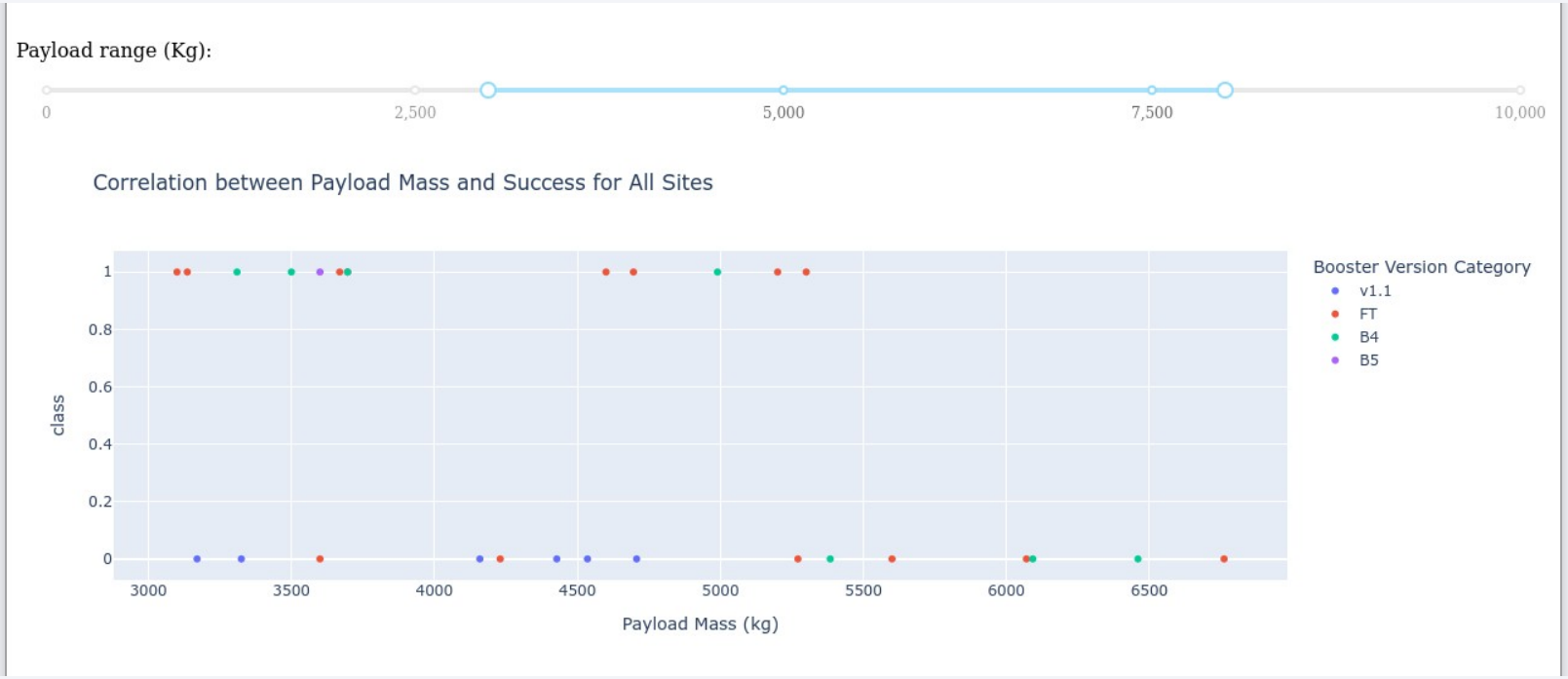
Payload vs. Launch Outcome

- This scatter plot shows the payload mass compared to the launch outcome (1= success, 0= failure). Each booster version is labeled with a different color. It clearly shows that v1.1 was very unsuccessful while the booster FT was much more successful.
- Currently it shows the full range of payloads from 0 to 10,000 kg.



Payload vs. Launch Outcome Cont.

- Here the scatter plot is limited to launches with payloads between 3,000 and 8,000 kg.
- Within this range the FT booster is the most successful while the v1.1 booster does not have a single success.

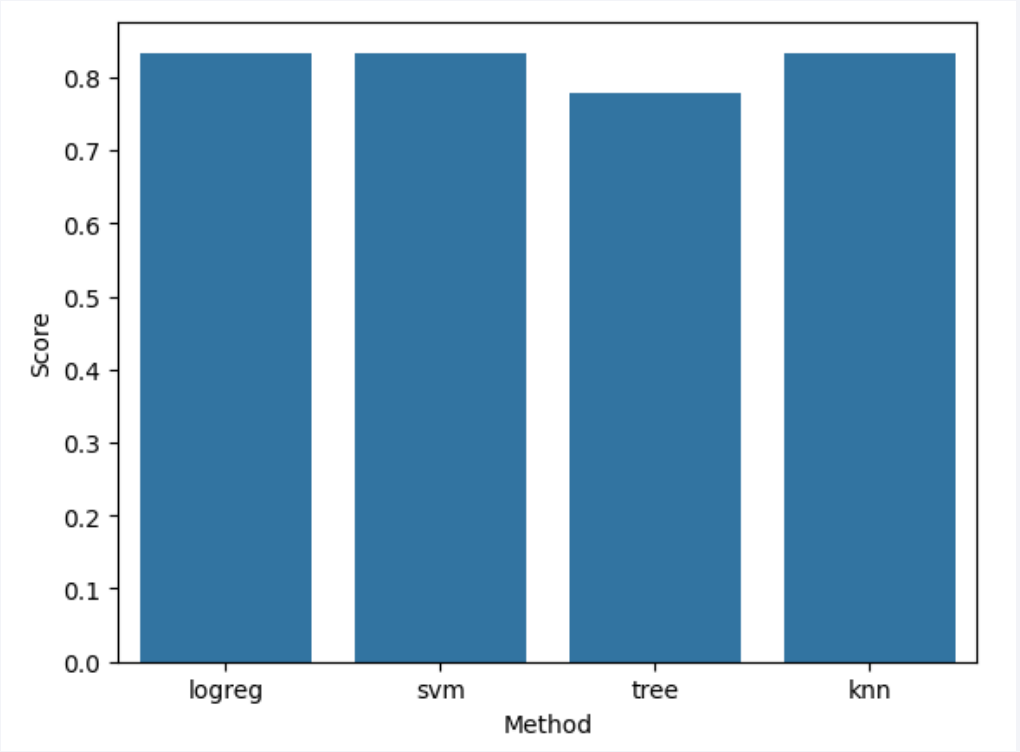


Section 5

Predictive Analysis (Classification)

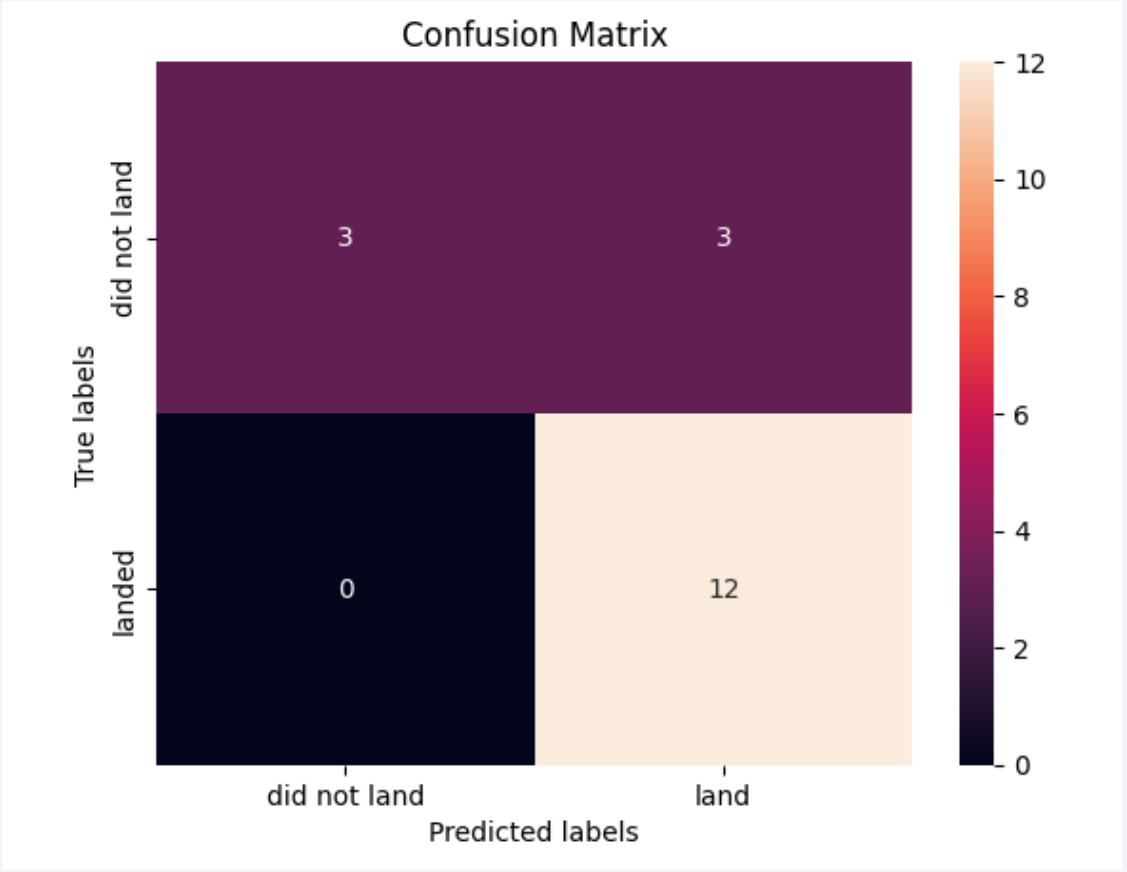
Classification Accuracy

- The bar chart shows the R^2 score of each classification method. All methods resulted in a score of 0.833 except for the decision tree method which resulted in a score of 0.778.



Confusion Matrix

- Again the most successful methods were the logistic regression, k-nearest neighbor, and support vector machine methods. Each had the same confusion matrix shown here.
- The model accurately labels each successful landing. It has a high sensitivity (100%) in that there are no false negatives.
- There is an issue with false positives. The model labeled 3 launches that did not land as landing successfully which is false. This shows there is a lower level of specificity (50%).



Conclusions

The SpaceX project improved from year to year starting in 2013 up until 2020. There was a significant dip in success during 2018. It would be worthwhile to determine the slump in performance in order to mitigate any possible sources of weakness in the landing outcomes.

Florida has the most launch sites with most flights originating from the KSC LC-39A and CCAFS SLC-40 launch sites. KSC LC-39A stands out as the best launch site to recover the booster rocket intact.

Launches with higher payload mass had greater success compared to lower mass launches; furthermore, they were launched later in the program so this may just reflect the projects success and not specifically heavier payloads success.

Launch sites are situated in coastal regions as far south as possible. The ocean is used as a crash landing location so being near the water is required. Railroads are close to launch sites to facilitate rocket transport. Due to the inherent danger of launching rockets, launch sites are located a safe distance from populated areas.

Logistic regression, support vector machine, and k-nearest neighbor methods were the most accurate at classifying landing success. The decision tree model was less accurate.

- Using the more accurate methods, a model can predict the outcome of a launch with perfect sensitivity and low specificity

Thank you!