Winning Space Race with Data Science

Natalie Johns
January 26, 2023

Outline

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

Executive Summary

- Summary of methodologies
    - Data collection methodology:
        - Data from SpaceX API & Wikipedia Webscraping were pulled
  - Perform data wrangling
  - Perform exploratory data analysis (EDA) using visualization and SQL
  - Perform interactive visual analytics using Folium and Plotly Dash
  - Perform predictive analysis using classification models

- Summary of all results
  - Exploratory data analysis results
    - No clear correlation between successful missions & payload mass
    - Generally, success rates improved over time
  - Predictive analysis results
    - Reasonably accurate predictions of landing success can be determined

Introduction

Project background and context

- SpaceX has gained worldwide attention for a series of historic milestones
- It is the only private company ever to return a spacecraft from low-earth orbit
- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars whereas other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage.
- If we can determine if the first stage will land, we can determine the cost of a launch

Problem: Is it possible to predict if the first stage will successfully land?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API was used to pull boosters, payloads, orbits, flight numbers, results of launch
  - Wikipedia Webscraping was used to pull additional details of the flights (like customer)
- Perform data wrangling
  - Wrangled the columns into proper formats & accounted for missing values
  - Create additional columns (i.e. outcome column)
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

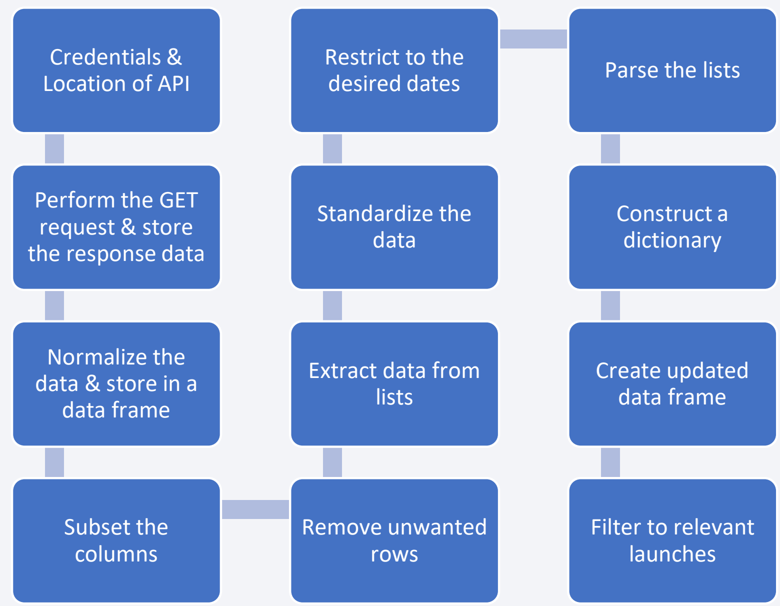
Data Collection

- Two datasets were pulled:
  - SpaceX API
  - Webscraping of Wikipedia page

Data Collection – SpaceX API

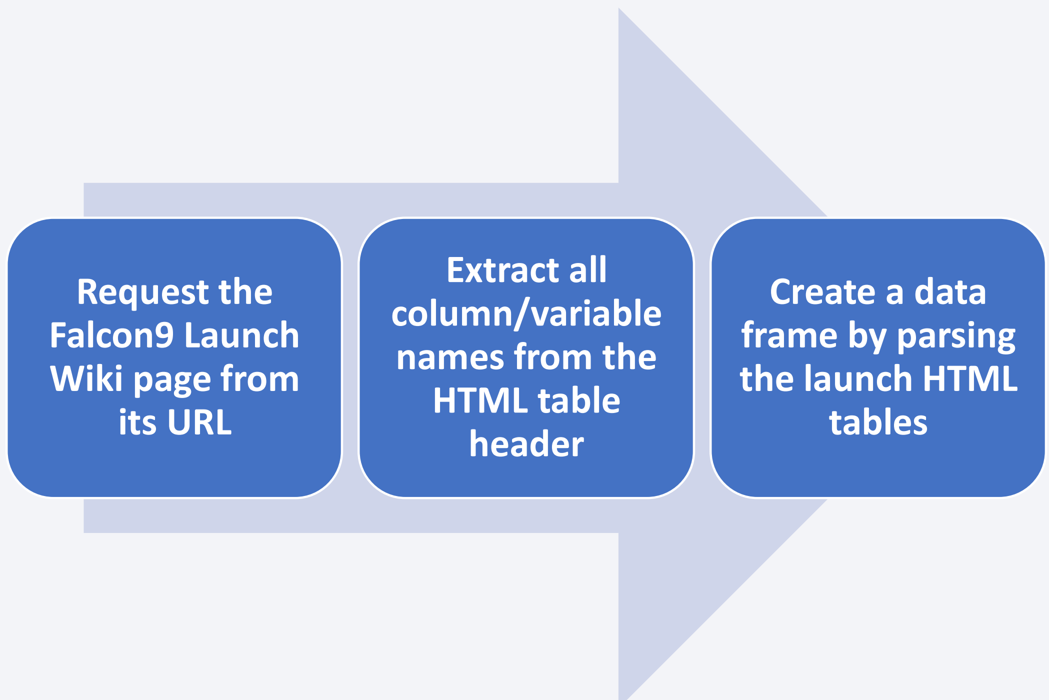
- Utilized the SpaceX REST API to get the raw data
- Normalized the JSON response
- Wrangled the data to get to the relevant data

- https://github.com/NJ5671/Coursera-Data-Science-Capstone-Project/blob/main/SpaceX%20Data.ipynb



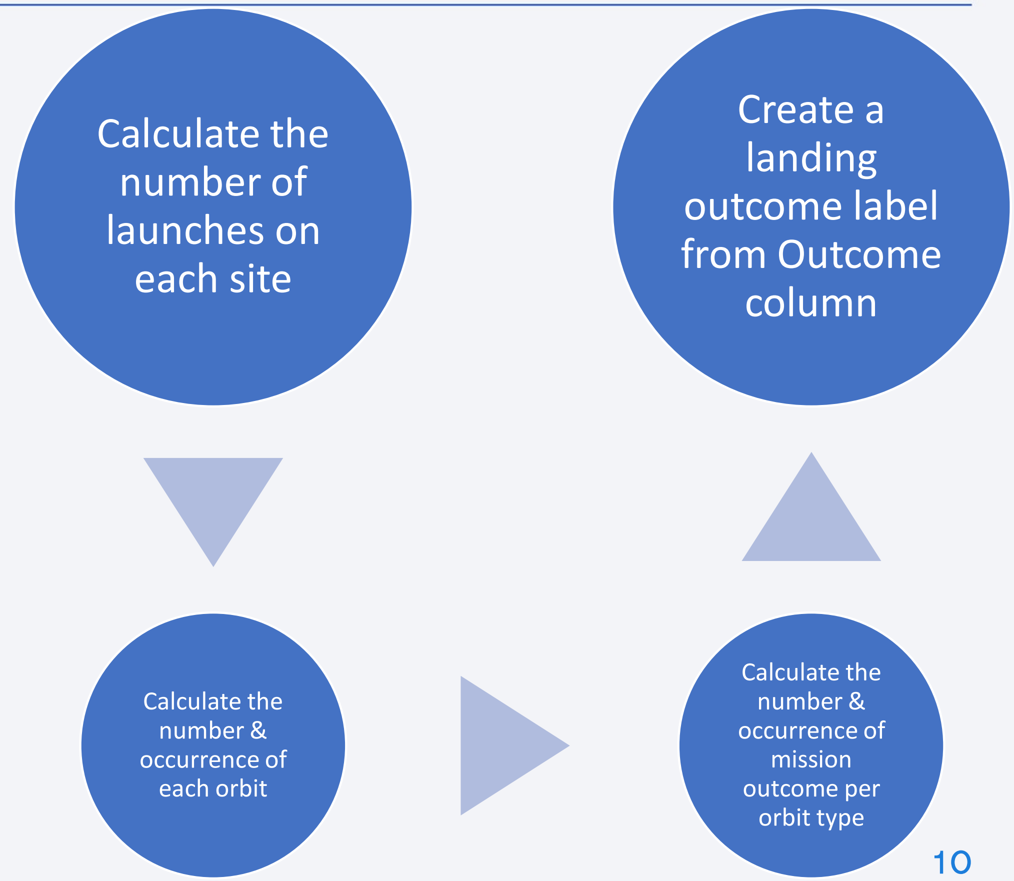
Data Collection - Scraping

- Use an HTTP GET to pull the HTML for the Wikipedia page
- Utilized BeautifulSoup to extract the Falcon 9 launch records HTML table
- Parsed the table and convert it into a Pandas data frame
- https://github.com/NJ5671/Coursera-Data-Science-Capstone-Project/blob/main/SpaceX%20Data%20--%20Web%20Scraping.ipynb



Data Wrangling

- Validated the data in each column was appropriate
- Identified which launches were successful
- Identified which launches were unsuccessful
- Created and outcome column & validated the success rate
- https://github.com/NJ5671/Coursera-Data-Science-Capstone-Project/blob/main/SpaceX%20Data%20--%20Data%20Wrangling.ipynb



EDA with Data Visualization

- Scatterplots with the outcome of the launch overlayed
  - Flight Number vs Payload Mass
  - Flight Number vs Launch Site
  - Flight Number vs Orbit
  - Payload Mass vs Orbit
- Bar chart of Success Rate over each type of Orbit
- Line Chart of yearly launch Success Rate
- https://github.com/NJ5671/Coursera-Data-Science-Capstone-Project/blob/main/SpaceX%20Data%20--%20Exploratory%20Data%20Analysis%20w%20Viz.ipynb

EDA with SQL

- Find the distinct launch sites
- Display five records where launch sites begin with 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass
- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- 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
- https://github.com/NJ5671/Coursera-Data-Science-Capstone-Project/blob/main/SpaceX%20Data%20--%20Exploratory%20Data%20Analysis.ipynb

Build an Interactive Map with Folium

- Map objects in interactive Map
  - Markers to flag launch sites, mark which were successful, calculate distance
  - Circle Marker to highlight the NASA Johnson Space Center in Houston, Texas
  - PolyLine to identify distance to closest city, railway & highway

- https://github.com/NJ5671/Coursera-Data-Science-Capstone-Project/blob/main/SpaceX%20Data%20--%20Interactive%20Dashboard.ipynb

Build a Dashboard with Plotly Dash

- Interactions within the dashboard
  - Dropdown list to enable Launch Site selection
  - Slider to select payload range
- Plots
  - Pie chart to show the total successful launches count for all sites
  - Scatter plot to show the correlation between payload and launch success

- https://github.com/NJ5671/Coursera-Data-Science-Capstone-Project/blob/main/SpaceX_interactive_dash.py

Predictive Analysis (Classification)

Prepare the data for modeling

- Create a Numpy array with the Class (success) field
- Standardize the data
- Split the data into test & train data sets

Model, tuning to best parameters

- Logistic Regression
- Support Vector Machine
- Decision Tree Classifier
- (GridSearchCV)
- KNN

Evaluate the Models

- Accuracy scores
- Confusion matrices

- https://github.com/NJ5671/Coursera-Data-Science-Capstone-Project/blob/main/SpaceX%20Data%20--%20Predictive%20Analysis.ipynb

Results

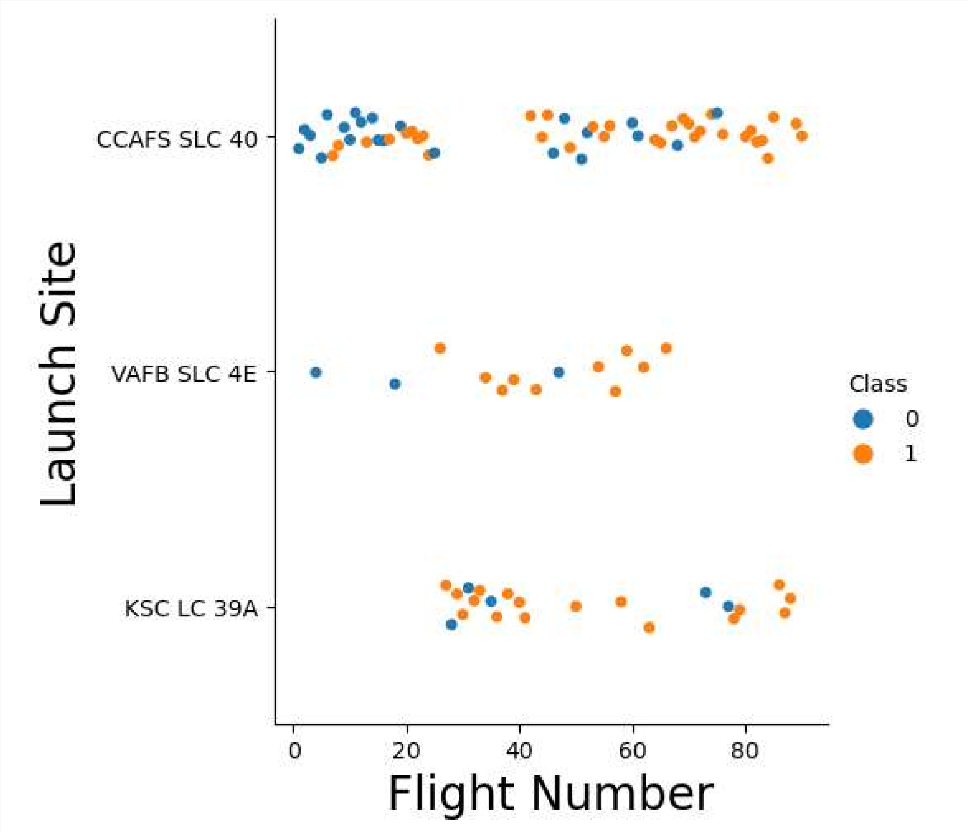
- Exploratory data analysis results
  - No clear correlation between successful missions & payload mass
  - Generally, success rates improved over time
- Interactive analytics demo in screenshots
- Predictive analysis results
  - Reasonably accurate predictions of landing success can be determined

Section 2

Insights drawn from EDA

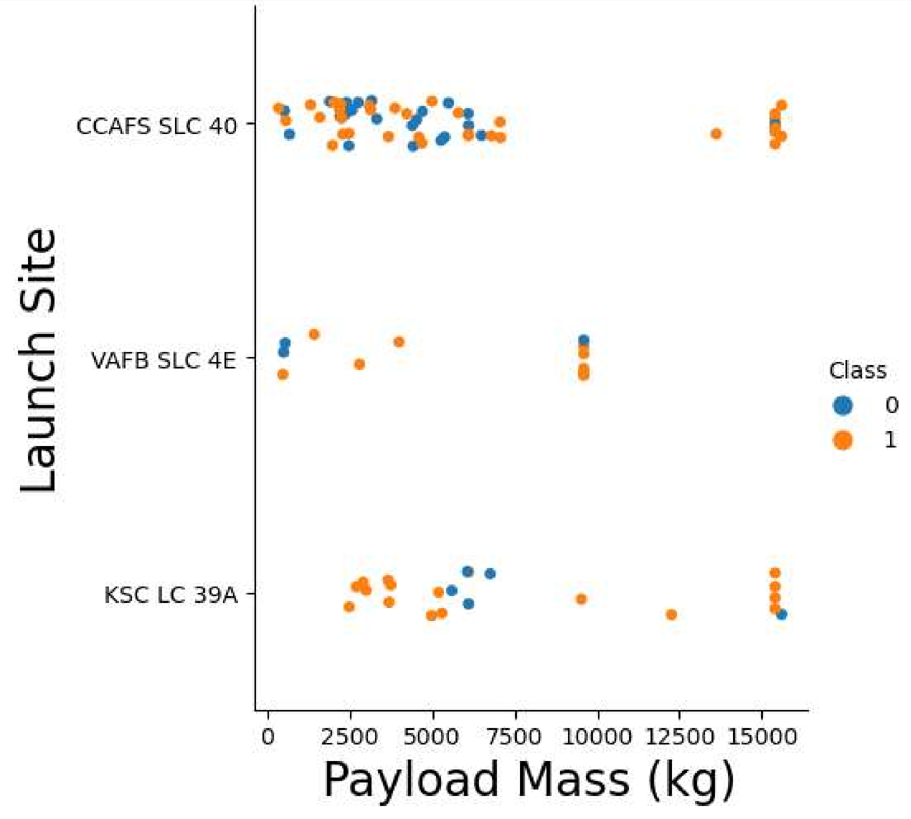
Flight Number vs. Launch Site

- Launch sites KSC LC 39A & VAFB SLC 4E had proportionally more successful flights than CCAFS SLC 40
- As flight number increased, success of the flights increased



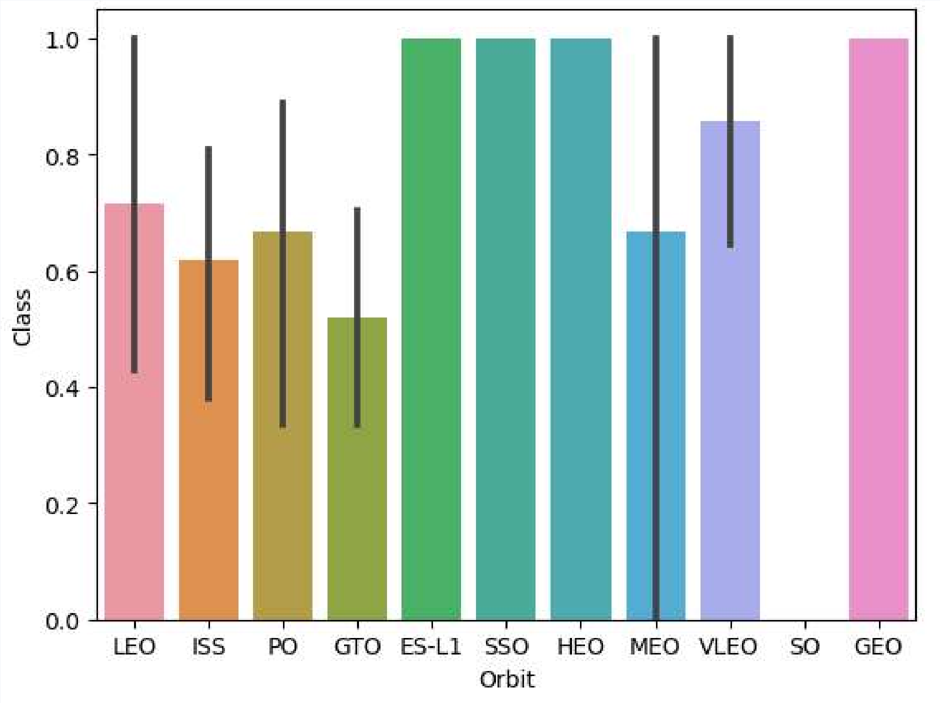
Payload vs. Launch Site

- Not many flights were attempted between 7,500 kg & 15,000 kg
- Mixed results from Launch Site CCAFS SLC 40—no clear correlation between successful missions & payload mass



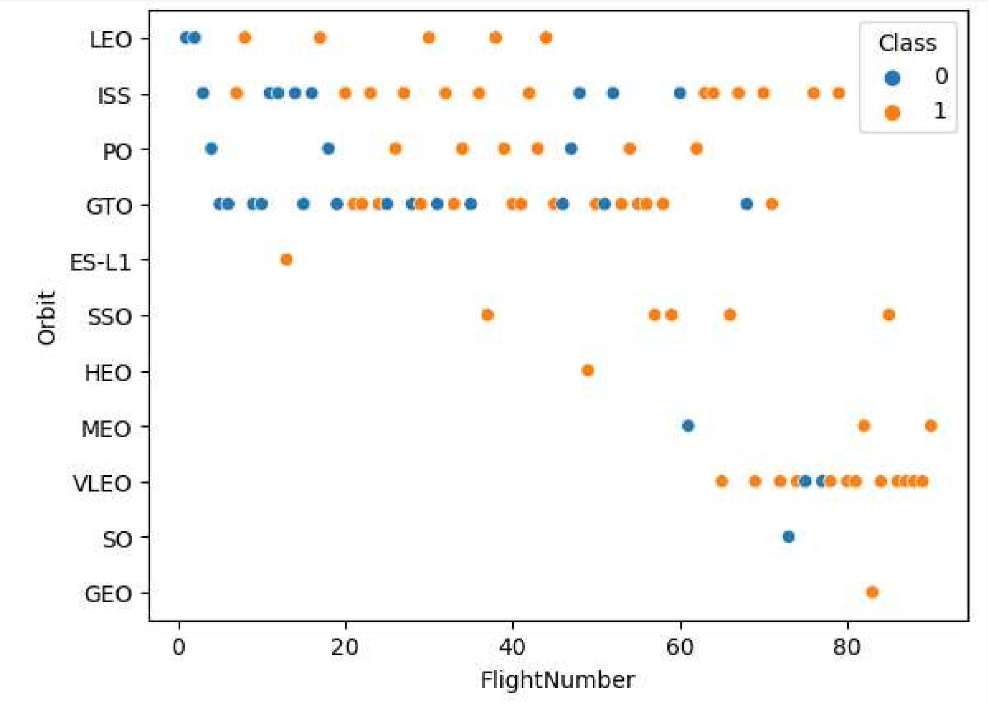
Success Rate vs. Orbit Type

- ES-L1, SSO, HEO, GEO had perfect successful launches
- GTO had the worst percent of successful launches



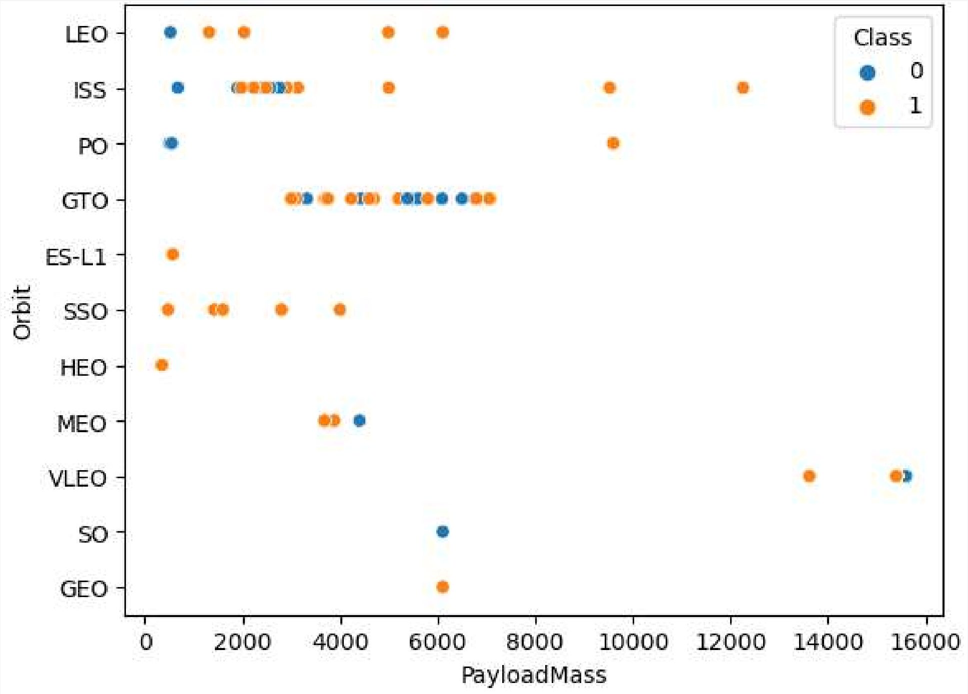
Flight Number vs. Orbit Type

- As flight number increased, success rate increased
- Orbits tested only with low flight numbers had a lower success rate



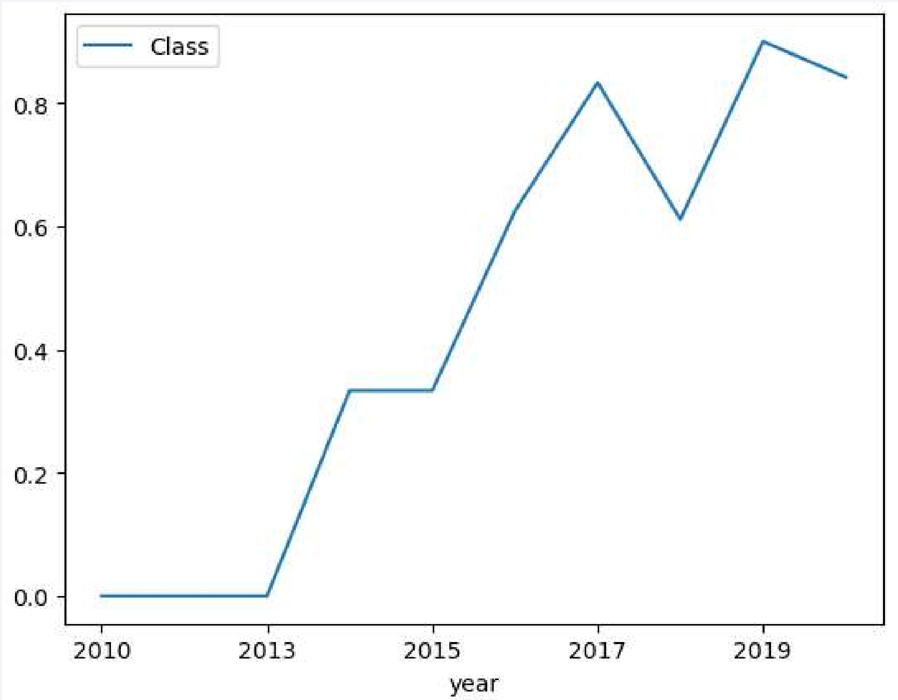
Payload vs. Orbit Type

- Most orbits did not have higher payload masses tested
- SSO was successful regardless of payload mass



Launch Success Yearly Trend

- Generally, success rates improved over time
- Exceptions are 2018 & 2020



All Launch Site Names

- Four launch sites were used

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`
- First two indicate no payload mass was tested
- All landing outcomes either failed or were not attempted

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

Total Payload Mass

- Over 45k kg was carried by boosters from NASA

1
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 was just over 2,900 kg

1
2928

First Successful Ground Landing Date

- The first successful landing outcome on ground pad occurred at the end of 2015
- This was a big step forward in the continued optimization of landing outcomes

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- 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

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes
- The vast majority of mission outcomes were successful

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- The boosters which have carried the maximum payload mass
- This allows us to see which boosters have been tested with the maximum payload mass

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

- The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

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

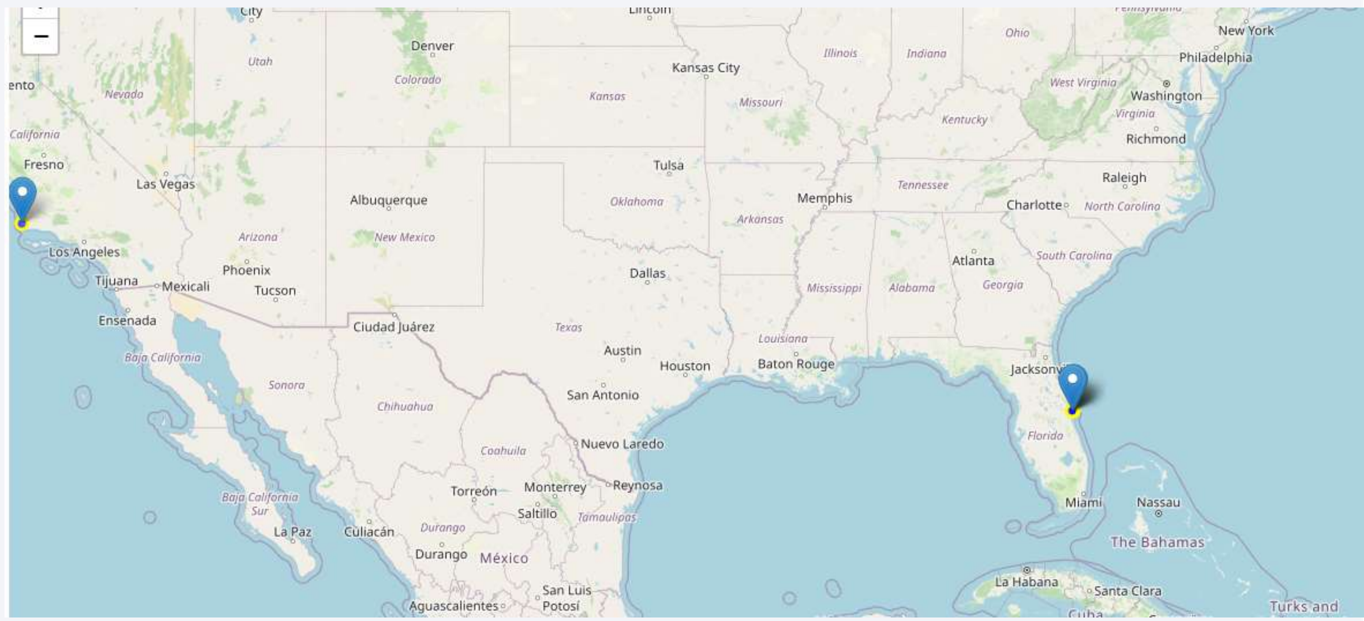
- Rank of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order
- Roughly 1/3 of all landing outcomes were not attempted during this date range

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


Section 3
Launch Sites
Proximities Analysis

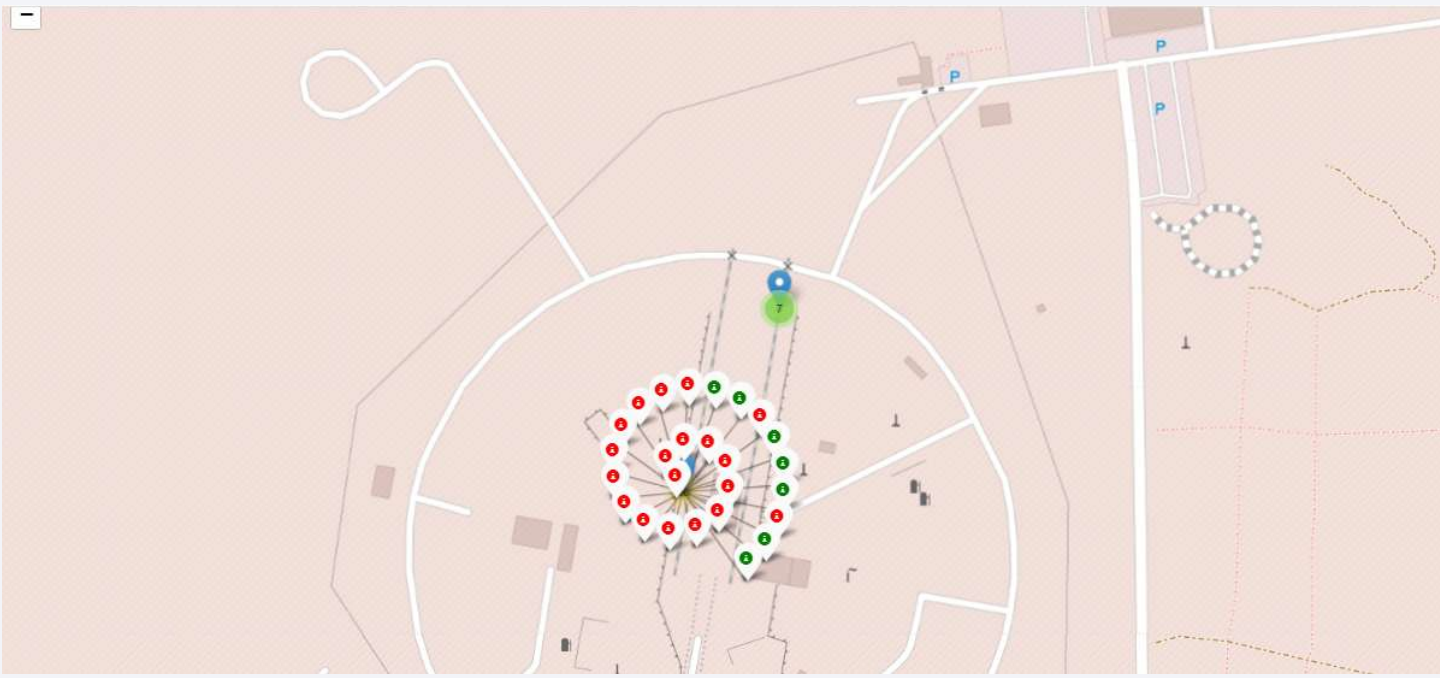
Launch Site Locations

- Launch sites were placed on each coasts
- Coastlines tend to be more flat & very near water
- Launch sites are towards the southern borders, where climates are warmer

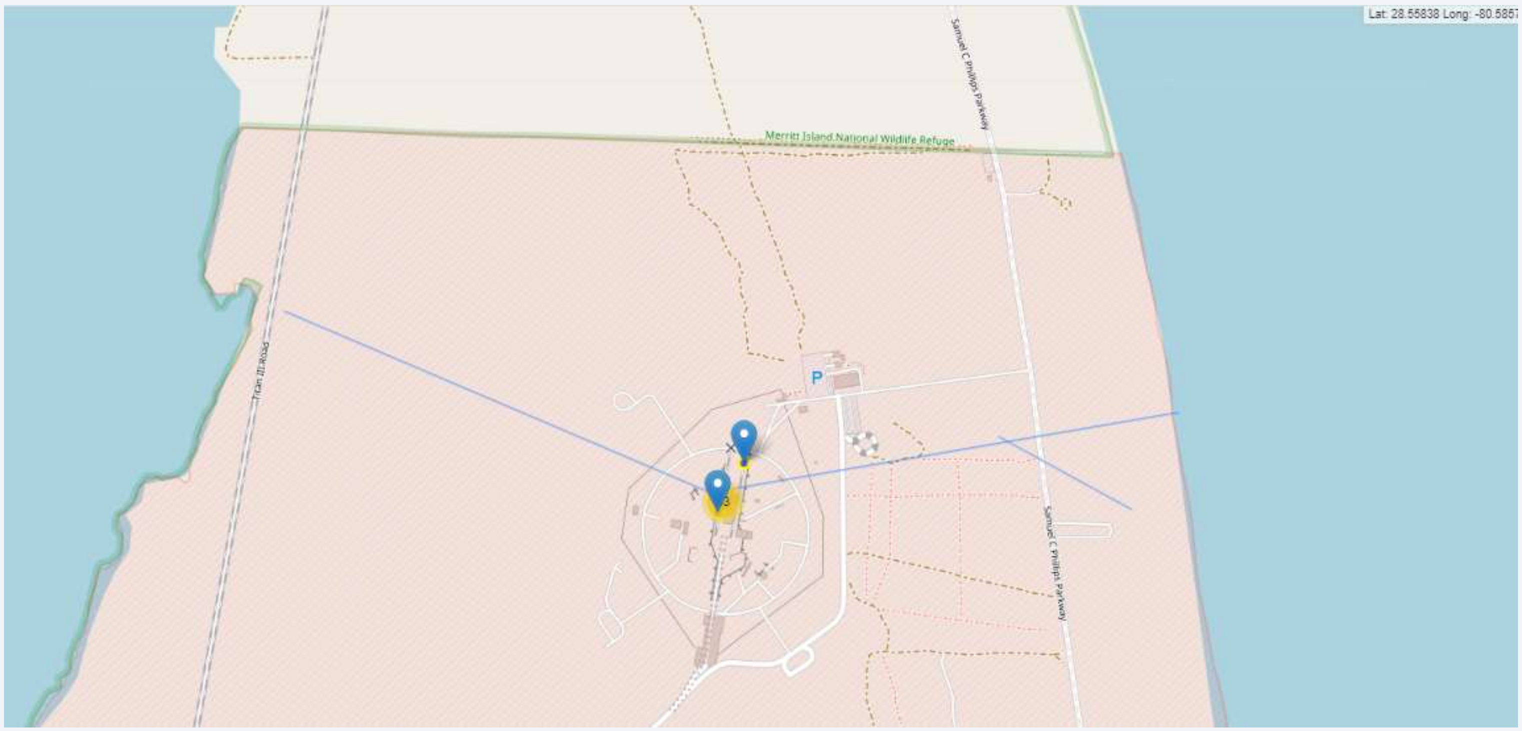


Launch Outcomes Map

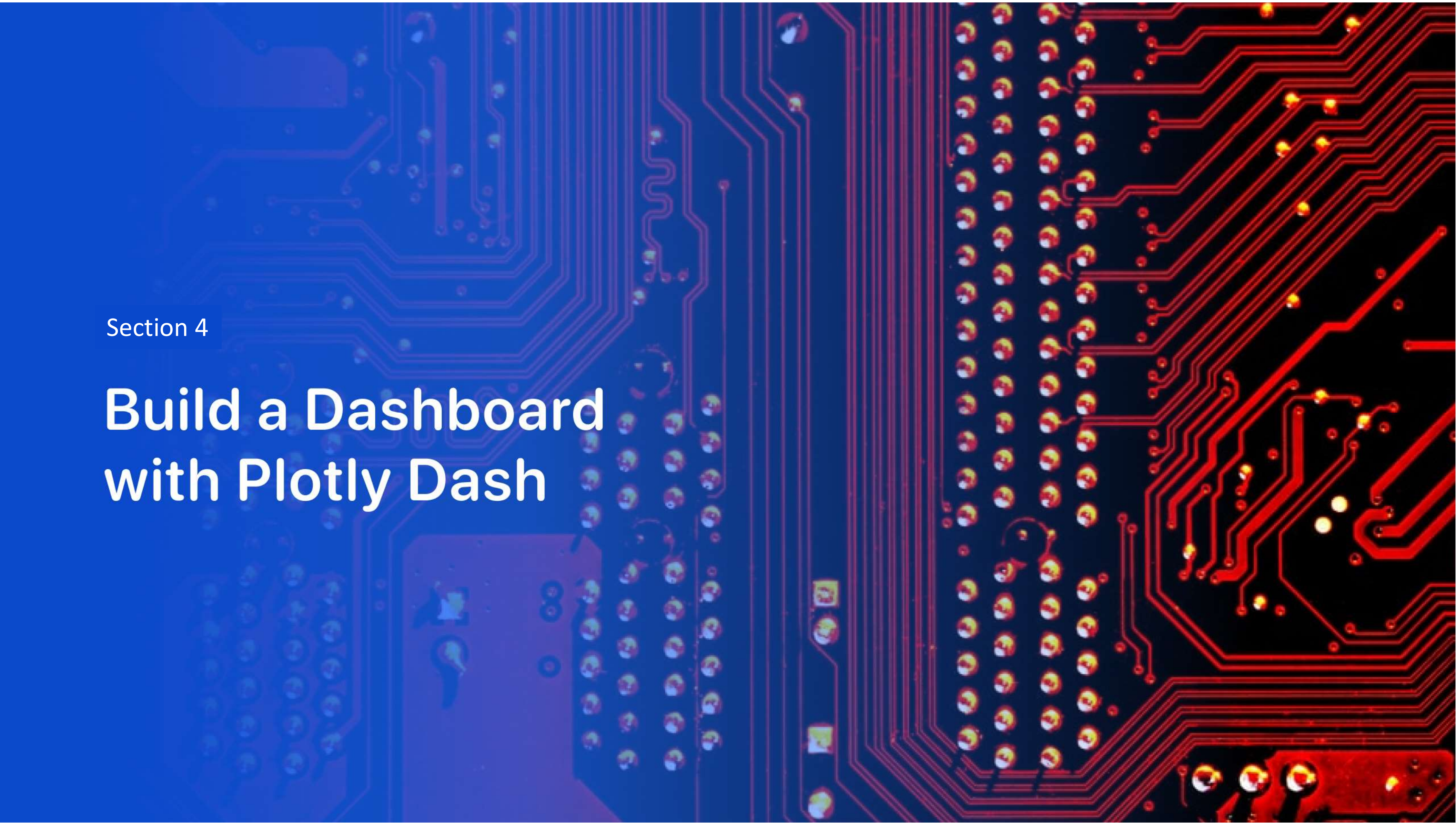
- Launch sites in Florida showing more failures than successes at this location



Launch Site Proximities to Key Resources



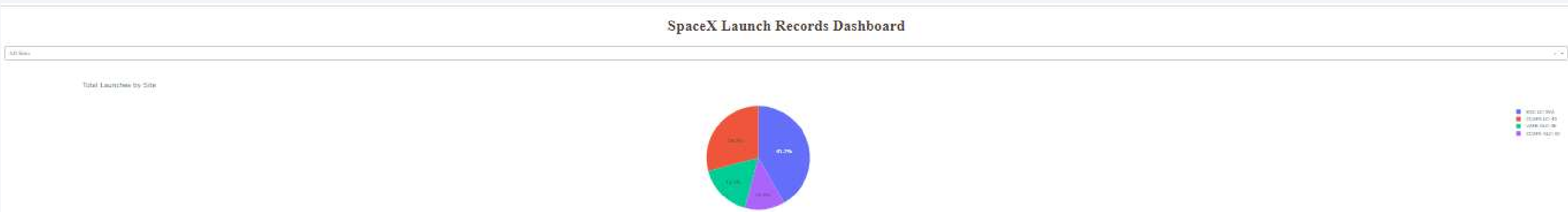
- Launch sites were placed far enough from resources to not endanger them
- Launch sites were placed close enough to resources to make use of them



Section 4

Build a Dashboard with Plotly Dash

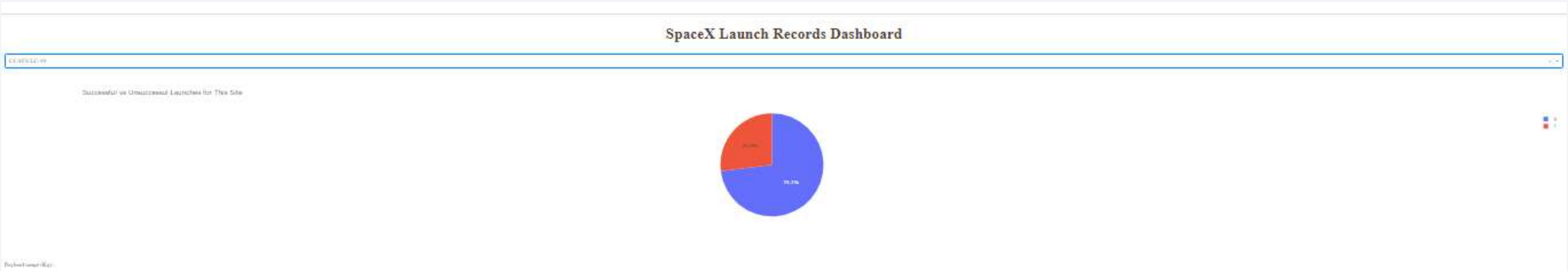
Launch Site Counts



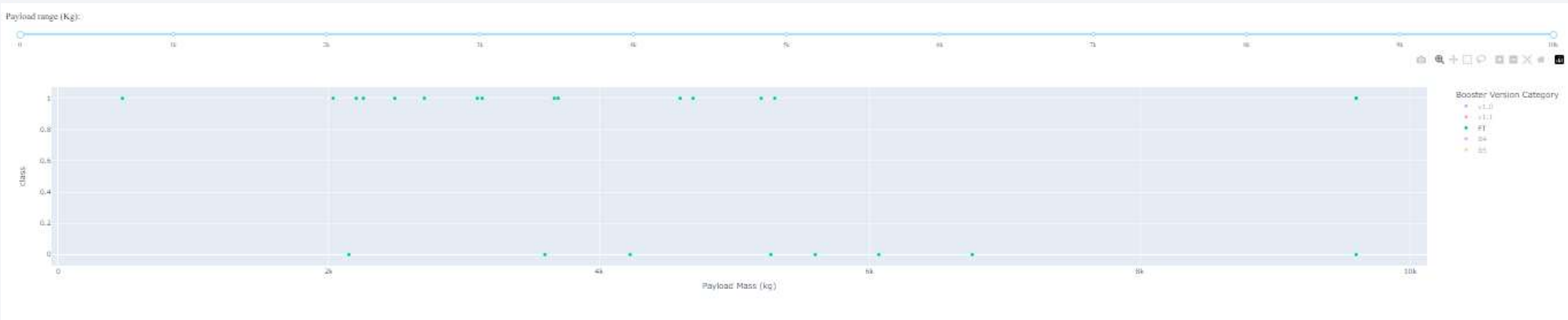
The launches were disproportionately CCAFS SLC 40

Launch site with highest success rate

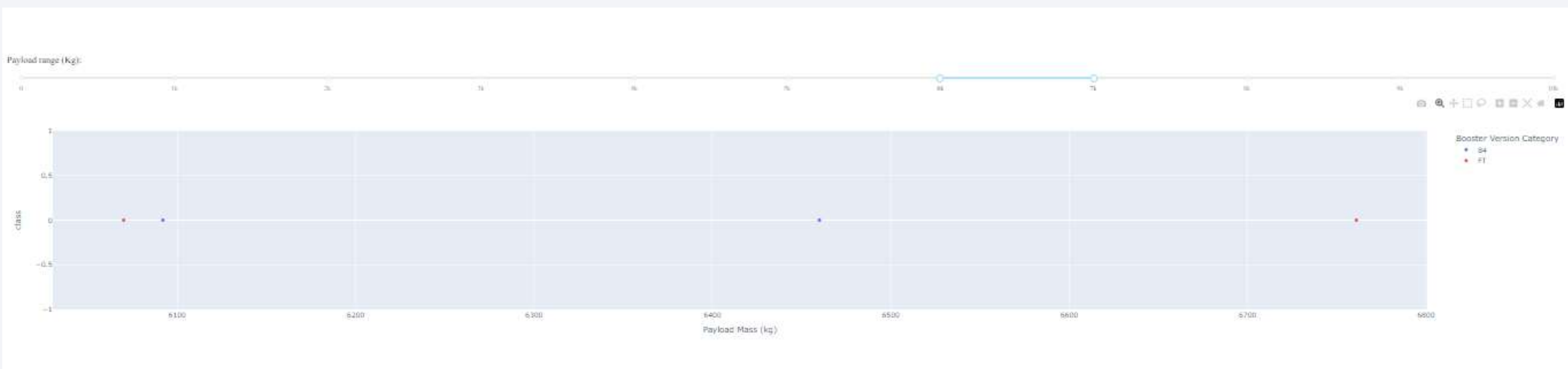
- CCAFS SLC 40 had the highest success rate with nearly ¾ of all landings being successful



Payload vs Launch Outcome



- By adjusting the slider, we can see that differences (some correlation) between payload & launch outcome



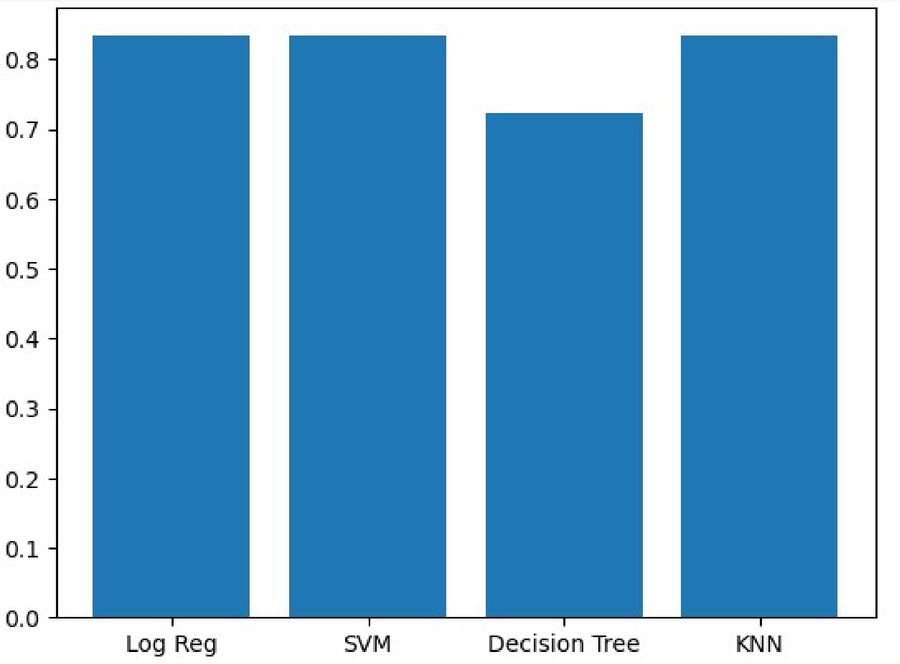
Section 5

Predictive Analysis (Classification)

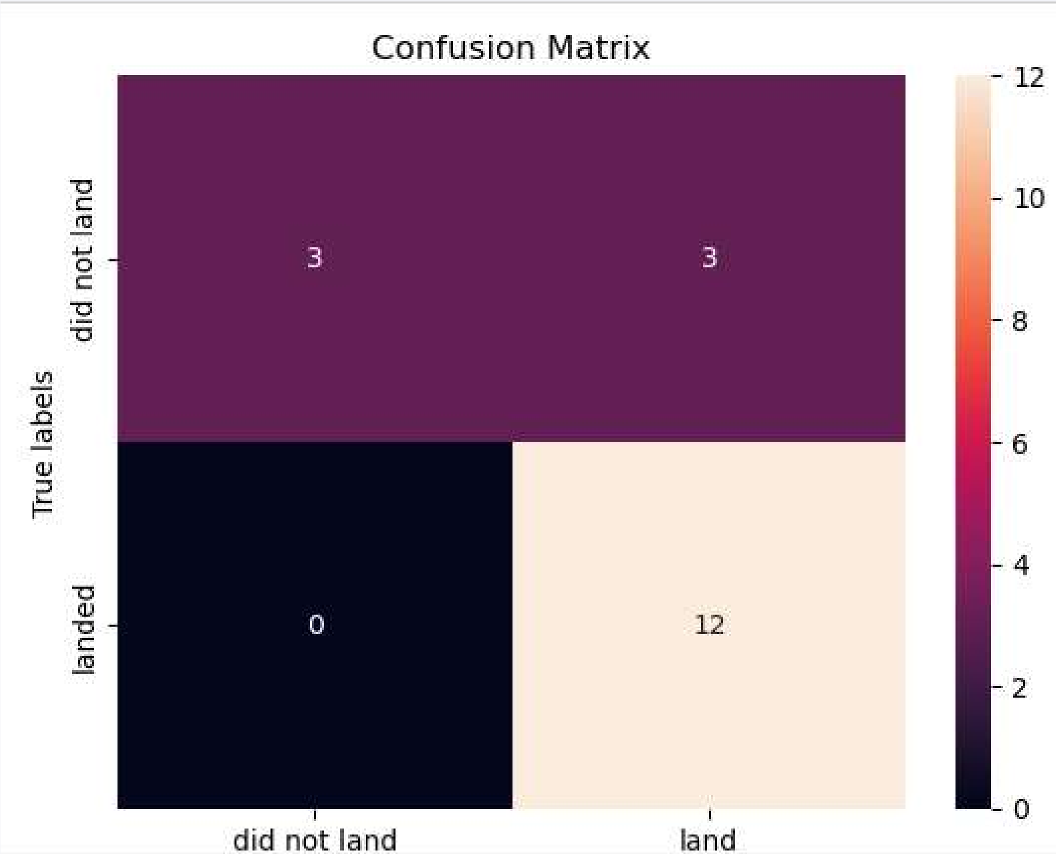
Classification Accuracy

Log Reg, SVM & KNN had the same accuracy rate of 0.833

Choosing Log Reg due to the simplicity of the model



Confusion Matrix



The Confusion Matrix is showing that the model is doing a pretty good job accurately predicting those that are going to land.

However, it is not doing as good a job accurately predicting those that did not land.

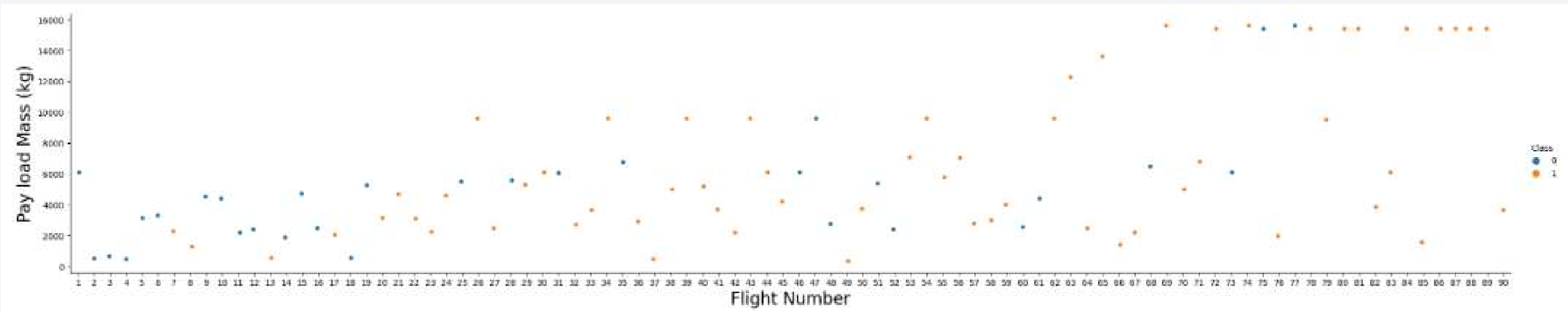
Conclusions

- Similar results were seen between the different models
- Logistic regression has a reasonable result & is much easier to interpret by non-data scientists
- Reasonably accurate predictions of landing success can be determined

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Appendix



As flight number increased, success rate improved. Advances are seen as missions were iterated. These improvements were seen, regardless of pay load mass.


Thank you!