IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Dr. Samuel P. Wilcock

17 July 2024

# Outline

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

# Executive Summary

- Summary of methodologies:
  - Data collection, scraping, and preprocessing
  - Exploratory Data Analysis to investigate the data, including graphically
  - Making an interactive dashboard
  - Fitting predictive models to seek the best way to classify launches
- Summary of all results:
  - Success has become more likely over time.
  - The KSC LC-39A site had the highest percentage of successful landings.
  - The number of orbits SpaceX has offered has increased over time.
  - Launch sites tend to be near coastlines, and have access to transportation, but are not generally close to their nearest town.
  - The Decision Tree Classification is the best at predicting the landing outcome, but has some difficulty with predicting when we will have a failure, while generally being able to correctly predict successful landings.

# Introduction

- The goal of this project is to model SpaceX launches and see what variables may be useful in predicting whether a particular launch has a successful landing of the booster.
- To do this, we need to:
    - Obtain data pertinent to the setting
    - Clean any data issues and decide which variables are most helpful
    - Visualize the relationship between the variables and successful booster landings
    - Build a model to predict the outcome of the booster landing attempt for a launch

Section 1

# Methodology

# Methodology

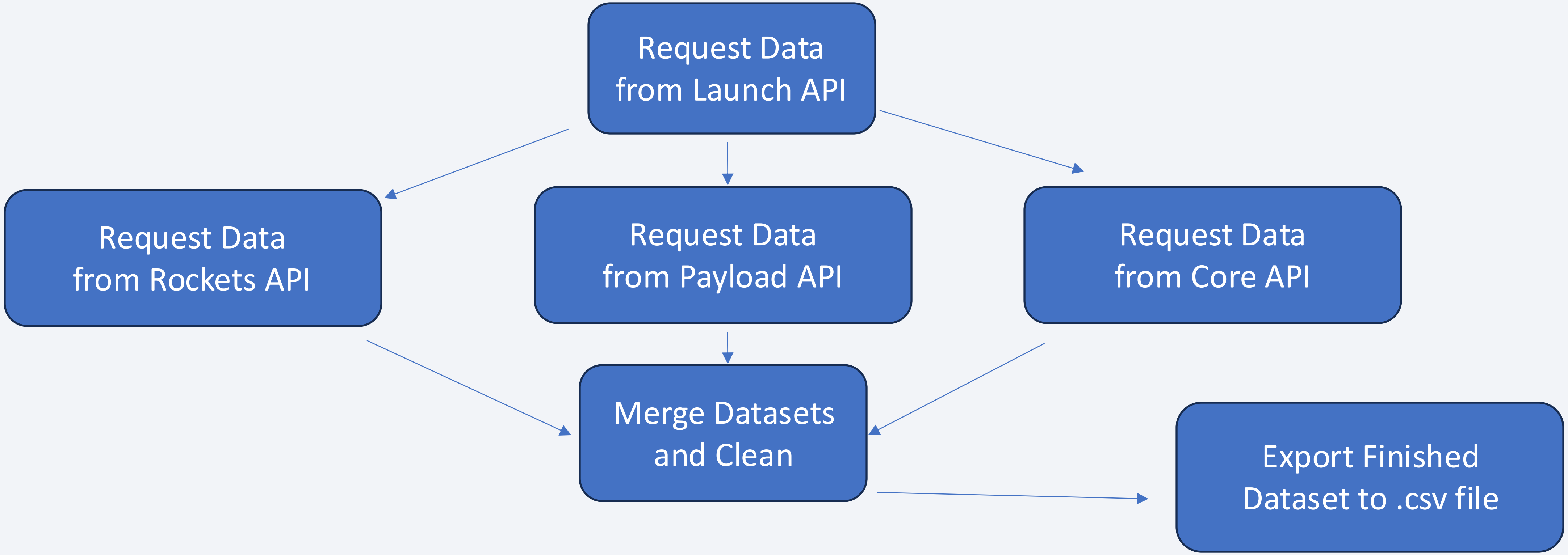
## Executive Summary

- Data collection methodology:
  - Data was gathered from several online sources, including combining data from several places in the publicly available SpaceX API
- Perform data wrangling
  - The data was processed by normalizing variables for more consistent comparison.
  - We also reduced the size of the dataset to focus only on Falcon 9 Boosters.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - We tested four different classification models to find the best predictor of success.

# Data Collection

- The needed data were housed in various locations within the SpaceX API.
- The needed datasets were downloaded from the API and then merged into one coherent dataset that could be used to accurately evaluate important factors.
- Falcon 9 launches were the only ones selected so that we can accurately predict success for only those launches which are comparable to our plans.
- It was noted that some payload information was missing, so this was addressed.

# Data Collection – SpaceX API



Request Data from Launch API

Request Data from Rockets API

Request Data from Payload API

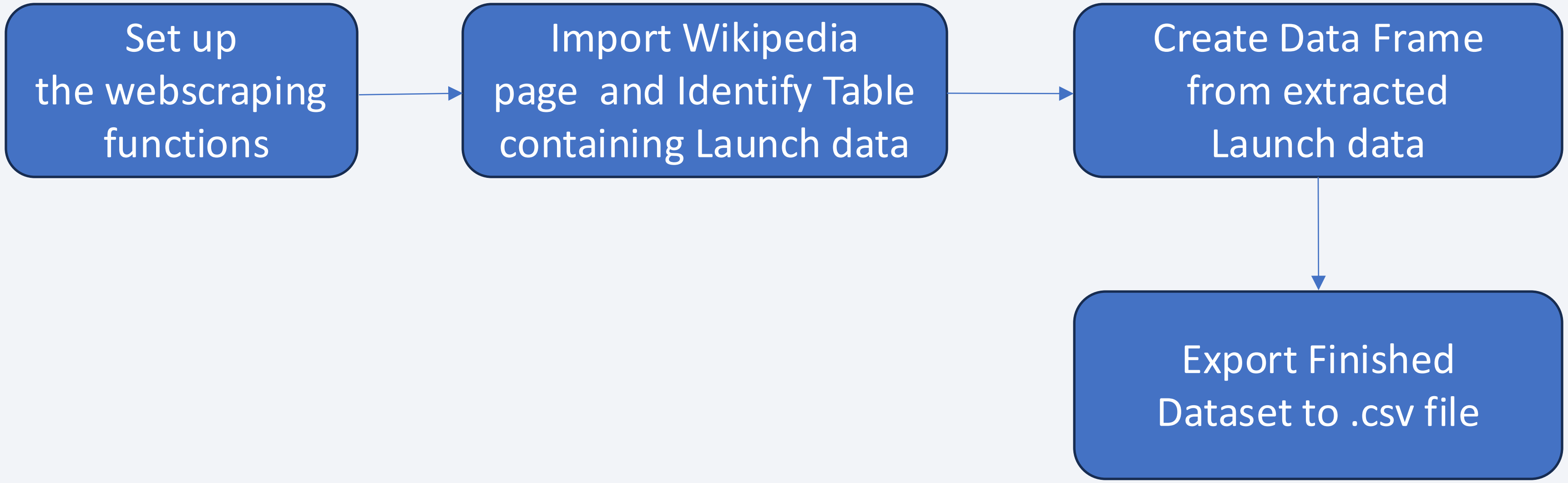
Request Data from Core API

Merge Datasets and Clean

Export Finished Dataset to .csv file

- https://github.com/Dr-Wilcock/DataScienceCapstone/blob/de201a3c7be7f9dde6b30a787c3906b7090ae0c8/jupyter-labs-spacex-data-collection-api.ipynb

# Data Collection - Scraping



Set up the webscraping functions

Import Wikipedia page and Identify Table containing Launch data

Create Data Frame from extracted Launch data

Export Finished Dataset to .csv file

- https://github.com/Dr-Wilcock/DataScienceCapstone/blob/de201a3c7be7f9dde6b30a787c3906b7090ae0c8/jupyter-labs-webscraping.ipynb

# Data Wrangling

- After importing the data from our previous work, we did the following:
  1. We then identified missing values and the types of data.
  2. We identified how many launches occurred at each site, how many launches targeted each orbit, and how many landing outcomes there were in the original coding
  3. We then converted these landing outcomes to a new binary variable based on success/failure

- https://github.com/Dr-Wilcock/DataScienceCapstone/blob/de201a3c7be7f9dde6b30a787c3906b7090ae0c8/labs-jupyter-spacex-Data%20wrangling%20(1).ipynb

# EDA with Data Visualization

- The following charts and graphs were constructed:
  - Scatterplots comparing various values including Payload, Flight Number, Launch Site, and Orbit
  - A bar chart of the success rate by targeted orbit
  - A line graph of success over by year

- https://github.com/Dr-Wilcock/DataScienceCapstone/blob/de201a3c7be7f9dde6b30a787c3906b7090ae0c8/edadataviz.ipynb

# EDA with SQL

- SQL was used to complete some exploratory data analysis
  - The unique names of the launch sites were obtained
  - We looked over the first 5 records from the two CCA launch sites
  - We found the total payload carried by SpaceX in the dataset
  - We found the average payload mass for all F9 v1.1 boosters
  - We found the date of the earliest successful ground pad landing
  - We found the name of the boosters with successful drone ship landings and payloads between 4000 and 6000 kg
  - We found the number of each landing outcome type: "success", "failure", and "no attempt"\
  - We discovered the name of each booster version that carried the maximum payload
  - We looked up the months of the two failed landings in 2015.
  - Finally, we ranked the landing outcomes by how many of each there were, from greatest to least
- https://github.com/Dr-Wilcock/DataScienceCapstone/blob/de201a3c7be7f9dde6b30a787c3906b7090ae0c8/jupyter-labs-eda-sql-coursera_sqllite%20(1).ipynb

# Build an Interactive Map with Folium

- To better visualize location and landing related data, Folium was used to find launch sites on a map of the US and success data was added to each site.
- To better understand an example site, distance to important landmarks such as coastline, railroads, highways, and towns were added to a representative site
- This data can help to visualize whether location, or facts about it are important in predicting successful recapture of the first rocket stage
- https://github.com/Dr-Wilcock/DataScienceCapstone/blob/de201a3c7be7f9dde6b30a787c3906b7090ae0c8/lab_jupyter_launch_site_location%20(1).ipynb

# Build a Dashboard with Plotly Dash

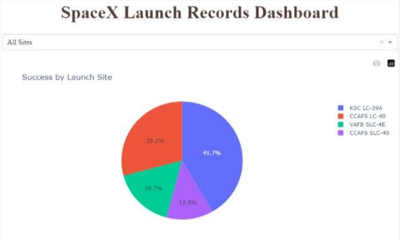
- We used Plotly Dash to create a Dashboard with the following options/plots:
  - For all sites combined:
    - A Pie chart of the percentage of overall successful landings attributable to each site
    - A scatterplot of values of payload mass and landing classification with a slider to select payload range
  - For each site:
    - A Pie chart of the landings by success/failure
    - A scatterplot of values of payload mass and landing classification with a slider to select payload range
- These plots further allow us to analyze site related effects on success
- https://github.com/Dr-Wilcock/DataScienceCapstone/blob/de201a3c7be7f9dde6b30a787c3906b7090ae0c8/spacex_dash_app.py

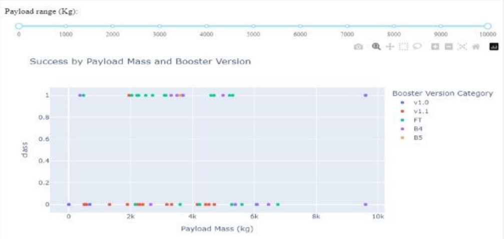
# Predictive Analysis (Classification)

- Four types of classification models were tested for their predictive ability to suggest whether a particular launch would result in a successful landing of the booster:
    - Logistic Regression
    - Support Vector Machines
    - Decision Tree Classification
    - K-Nearest Neighbors
- Each model was fitted across a grid of parameters and the best possible fit selected
- Each model was then evaluated using standard measures of accuracy and scoring
- https://github.com/Dr-Wilcock/DataScienceCapstone/blob/de201a3c7be7f9dde6b30a787c3906b7090ae0c8/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

# Results

- After exploration of the data, the following variables were kept:
    - Flight Number, Launch site, Payload, Payload mass, Orbit, Customer, Launch outcome, Version booster, Booster landing, Date, and Time
- We found that the first successful ground pad landing occurred in 2015. Only two drone ship landings failed that year, one in January, and the other in April.
- We were able to discover graphically, that several of the variables are related and that success has generally increased over time.
- An interactive dashboard was created to allow further insight. An example can be seen here (more details provided in a later slide):



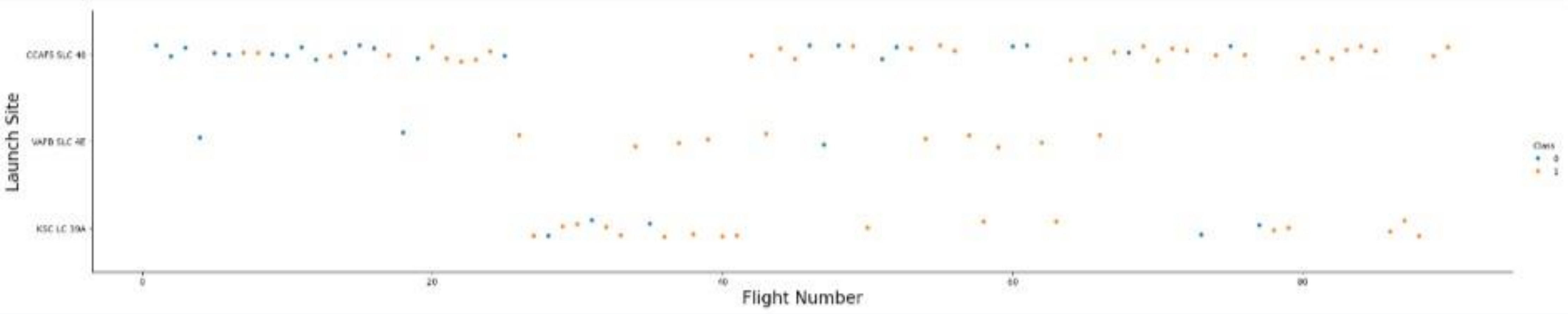


- Overall, the Decision Tree method was the best predictive model.

Section 2

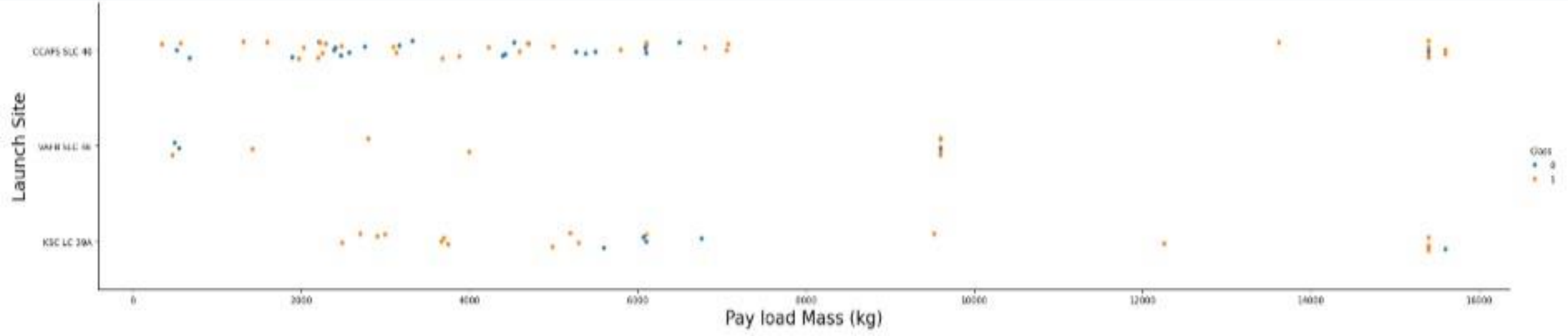
# Insights drawn from EDA

# Flight Number vs. Launch Site



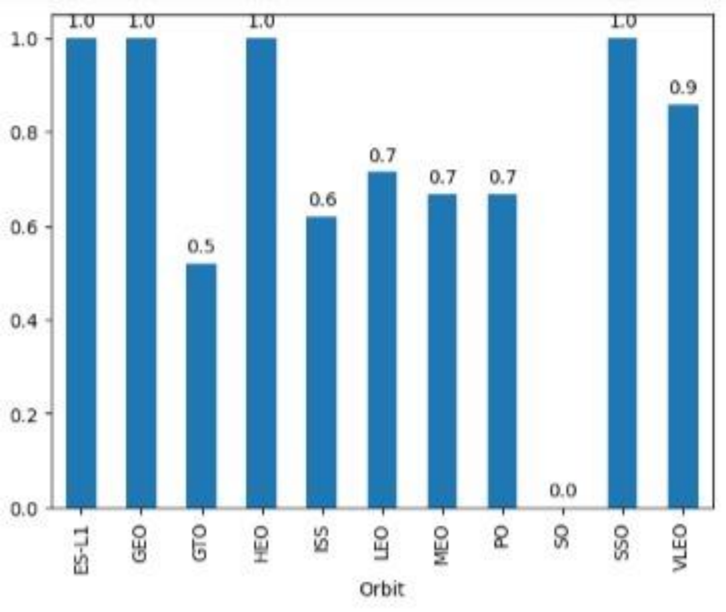
- Orange dots indicate successful landings.
- Note that for later flights, all launch sites show an increased success rate.

# Payload vs. Launch Site



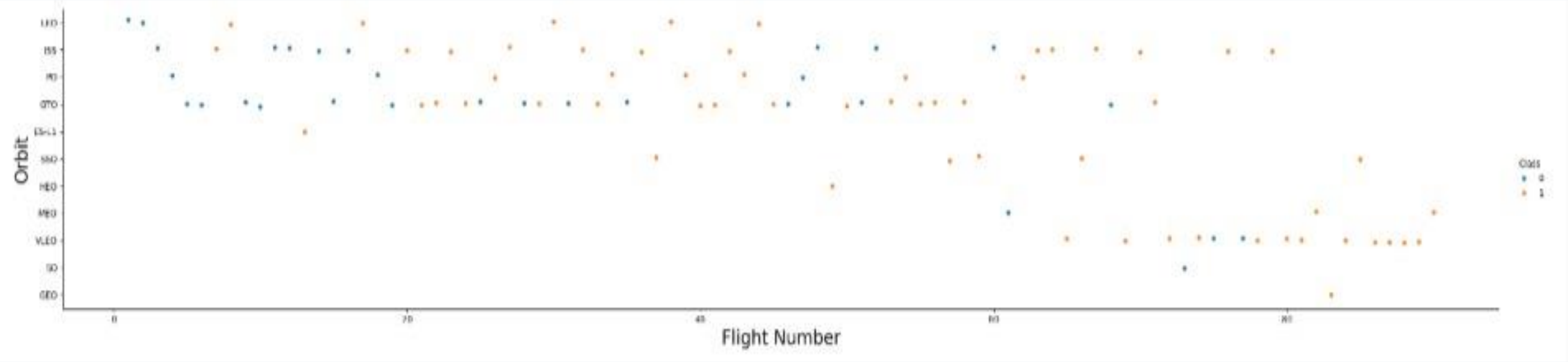
- Orange dots indicate successful landings.
- Note that there are gaps in payload size, and that larger payloads seem to be more successful

# Success Rate vs. Orbit Type



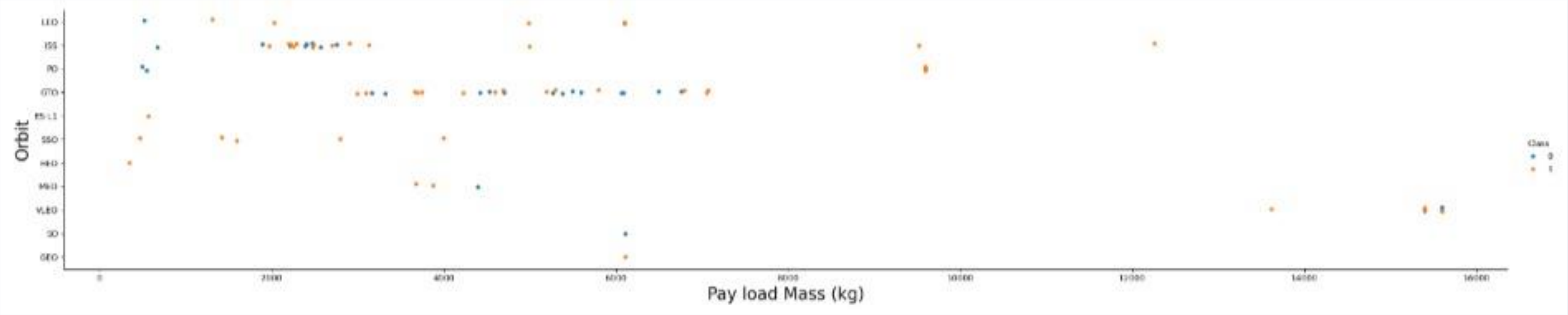
- Note that there were no successful landings for SO (Sun Synchronous Orbit)
- Note that there are four orbit types with no failed landings:
  - o ES-L1 (Earth-Sun Lagrange point 1)
  - o GEO (Geosynchronous)
  - o HEO (High earth)
  - o SSO (another type of Sun Synchronous Orbit)

# Flight Number vs. Orbit Type



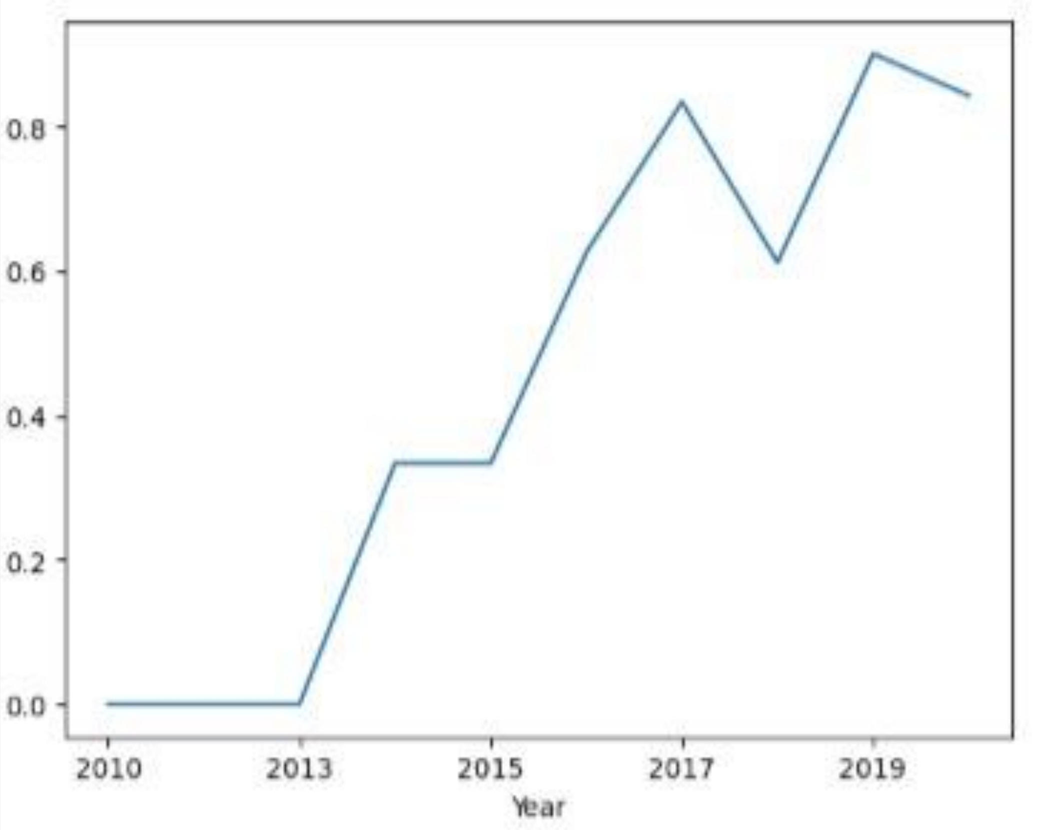
- Orange dots indicate successful landings
- It can be noted that the variety of orbit targets has grown over time, and that success has increased over time

# Payload vs. Orbit Type



- Orange dots indicate a successful landing
- Higher payloads are more likely to be successful, but there are only a few orbit targets that have used larger payloads.

# Launch Success Yearly Trend



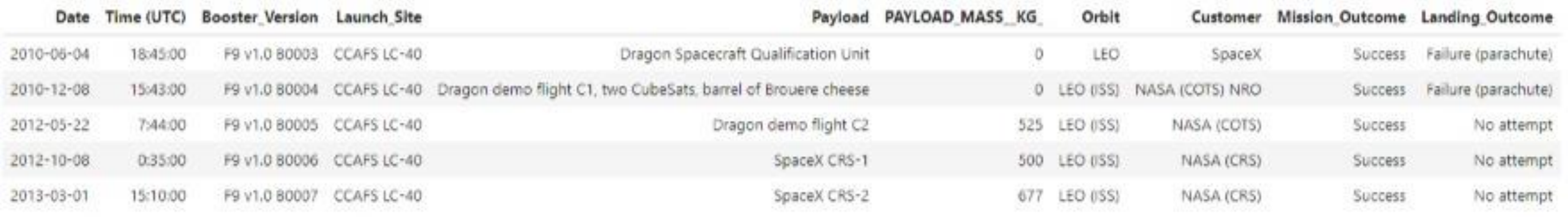
- Note that there is a general increasing trend.
- There was a downturn in 2018, and then again in 2020 (the last year of data in this dataset)
- This reinforces the observation that landings have increased in success, since flight number is corr elated with time.

# All Launch Site Names

- The unique launch site names are as follows:
  - CCAFS LS-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E
- This is the four unique sites SpaceX has used. These will later be seen on a map for comparison. It is interesting to note that these are all near a coast.
- Code: %sql select distinct(launch_site) from SPACEXTABLE

# Launch Site Names Begin with 'CCA'

- The first 5 records can be seen here:



| 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 |

- Note that these are all from the same CCAFS site, but have differing payloads and landing outcomes
- Code: %sql select * from SPACEXTABLE where launch_site like 'CCA%' limit 5

# Total Payload Mass

- The total payload launched for NASA (CRS) was 45,596 kg
- Code: %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer='Nasa (CRS)'

# Average Payload Mass by F9 v1.1

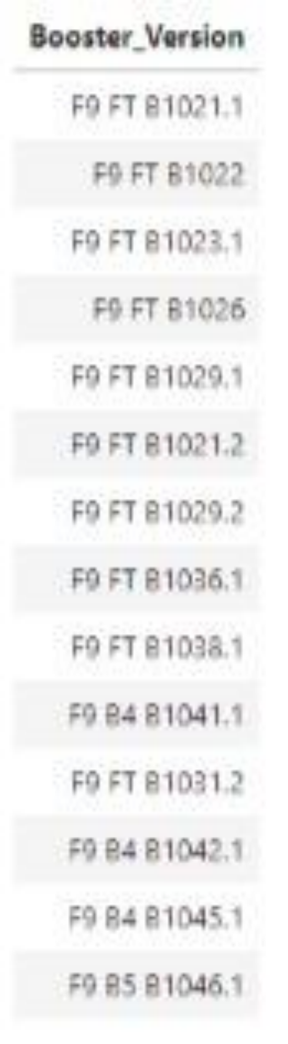
- The average payload for flights with booster version F9 v1.1 was 2,534.67
- Code: %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version like 'F9 v1.1%'

# First Successful Ground Landing Date

- The first successful ground pad landing occurred on 22 Dec 2015
- Code: %sql select min(Date) from SPACEXTABLE where Landing_Outcome like 'Success%'

# Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of the boosters with successful ship landings and payload in the desired range can be seen at right.
- Code: %sql select distinct(Booster_Version) from SPACEXTABLE where Landing_Outcome='Success (drone ship)'  and 4000<PAYLOAD_MASS__KG_<6000



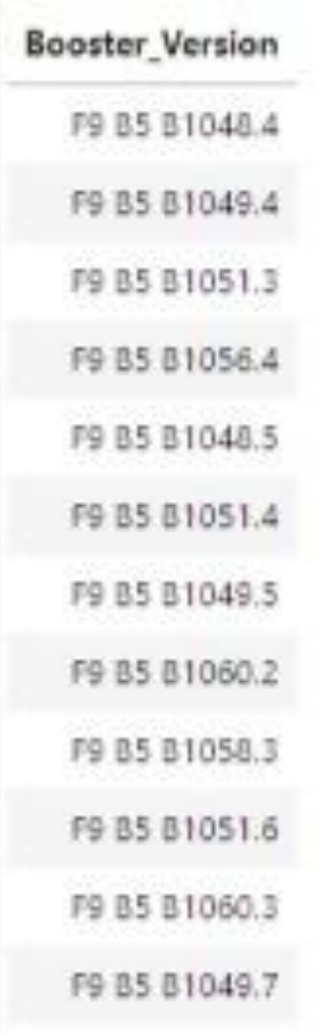
| Booster_Version |
|---|
| F9 FT B1021.1 |
| F9 FT B1022 |
| F9 FT B1023.1 |
| F9 FT B1026 |
| F9 FT B1029.1 |
| F9 FT B1021.2 |
| F9 FT B1029.2 |
| F9 FT B1036.1 |
| F9 FT B1038.1 |
| F9 B4 B1041.1 |
| F9 FT B1031.2 |
| F9 B4 B1042.1 |
| F9 B4 B1045.1 |
| F9 B5 B1046.1 |

# Total Number of Successful and Failure Mission Outcomes

- In total:
  - 61 missions were successful.
  - 10 missions were failures.
  - 22 Missions were labeled "No Attempt"
- Three code snippets:
  - %sql select count(Landing_Outcome) from SPACEXTABLE where Landing_Outcome like 'Success%'
  - %sql select count(Landing_Outcome) from SPACEXTABLE where Landing_Outcome like 'Failure%'
  - %sql select count(Landing_Outcome) from SPACEXTABLE where Landing_Outcome like 'No Attempt%'

# Boosters Carried Maximum Payload

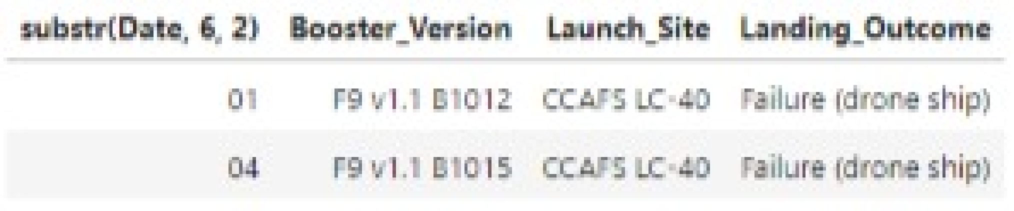
- The boosters which carried the maximum payload can be seen at the right.
- Code: %sql select distinct(Booster_version) from SPACEXTABLE where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)



| 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 landings on a drone ship can be seen here, with their booster versions, and launch site names for the year 2015

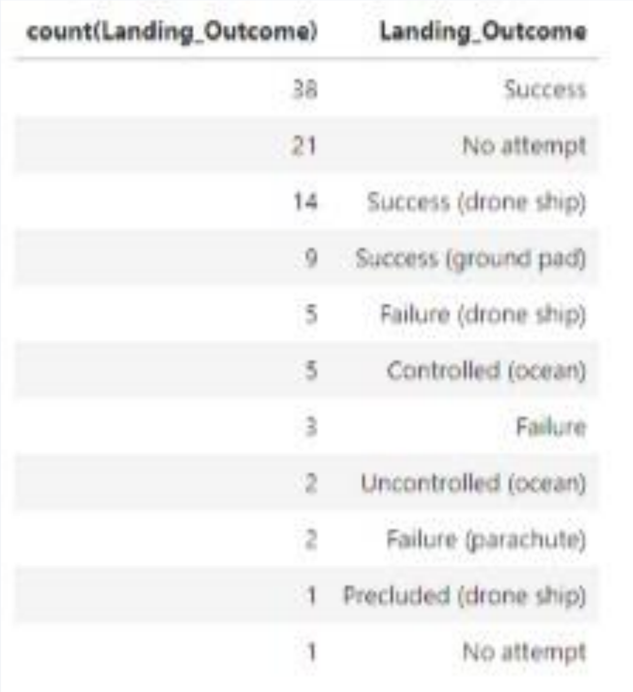


| substr(Date, 6, 2) | Booster_Version | Launch_Site | Landing_Outcome |
|---|---|---|---|
| 01 | F9 v1.1 B1012 | CCAFS LC-40 | Failure (drone ship) |
| 04 | F9 v1.1 B1015 | CCAFS LC-40 | Failure (drone ship) |

- Code: %sql select substr(Date, 6, 2), Booster_version, Launch_site, Landing_Outcome from SPACEXTABLE where Landing_Outcome='Failure (drone ship)' and substr(Date,0,5)='2015'

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

- The count of landing outcomes between the date 2010-06-04 and 2017-03-20 are shown at the right, in descending order from the most common
- Code: %sql select count(Landing_Outcome), Landing_Outcome from SPACEXTABLE where '2010-06-04'<=Date<='2017-03-20' group by Landing_Outcome Order by count(Landing_Outcome) desc

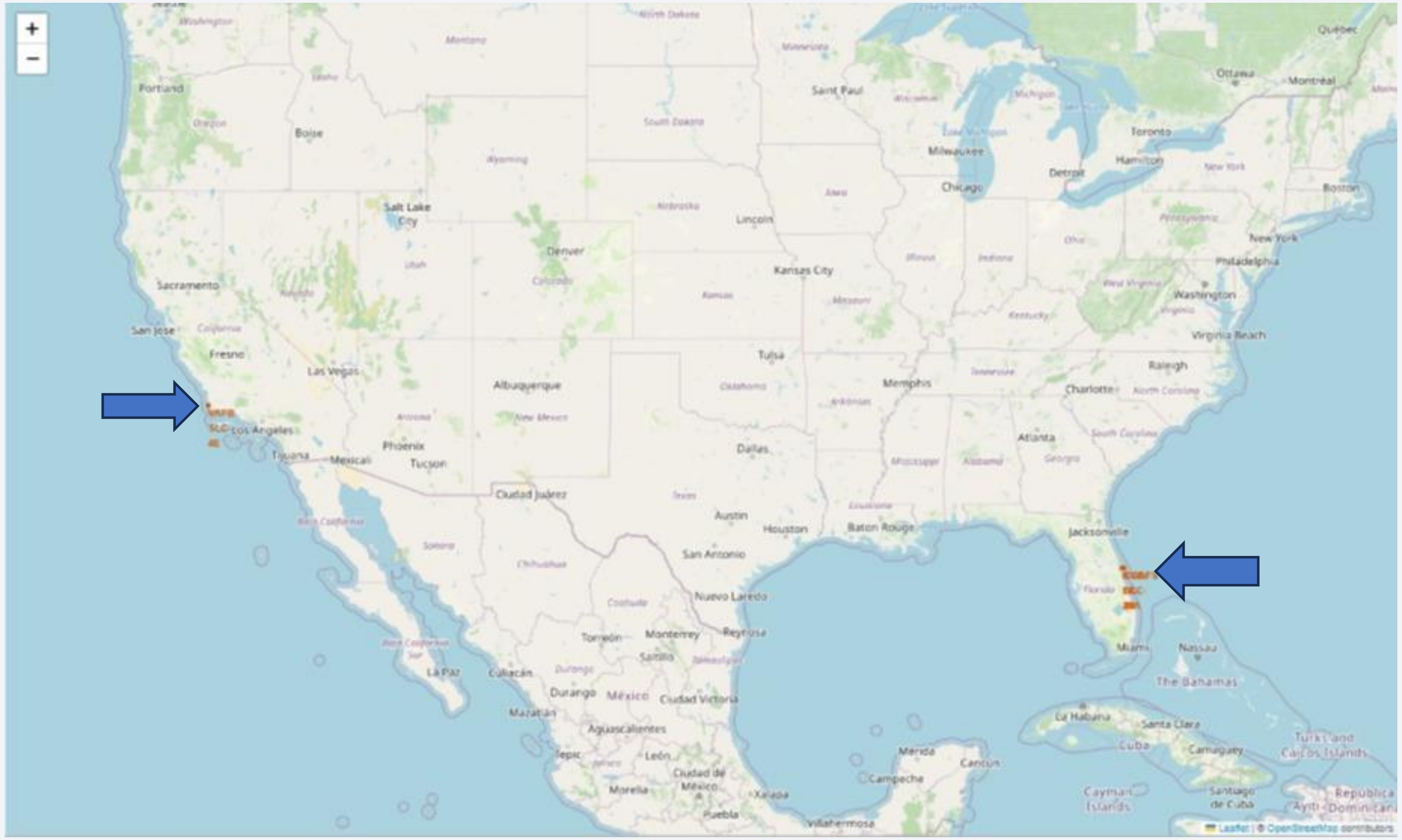


| count(Landing_Outcome) | Landing_Outcome |
|---:|---:|
| 38 | Success |
| 21 | No attempt |
| 14 | Success (drone ship) |
| 9 | Success (ground pad) |
| 5 | Failure (drone ship) |
| 5 | Controlled (ocean) |
| 3 | Failure |
| 2 | Uncontrolled (ocean) |
| 2 | Failure (parachute) |
| 1 | Precluded (drone ship) |
| 1 | No attempt |

Section 3

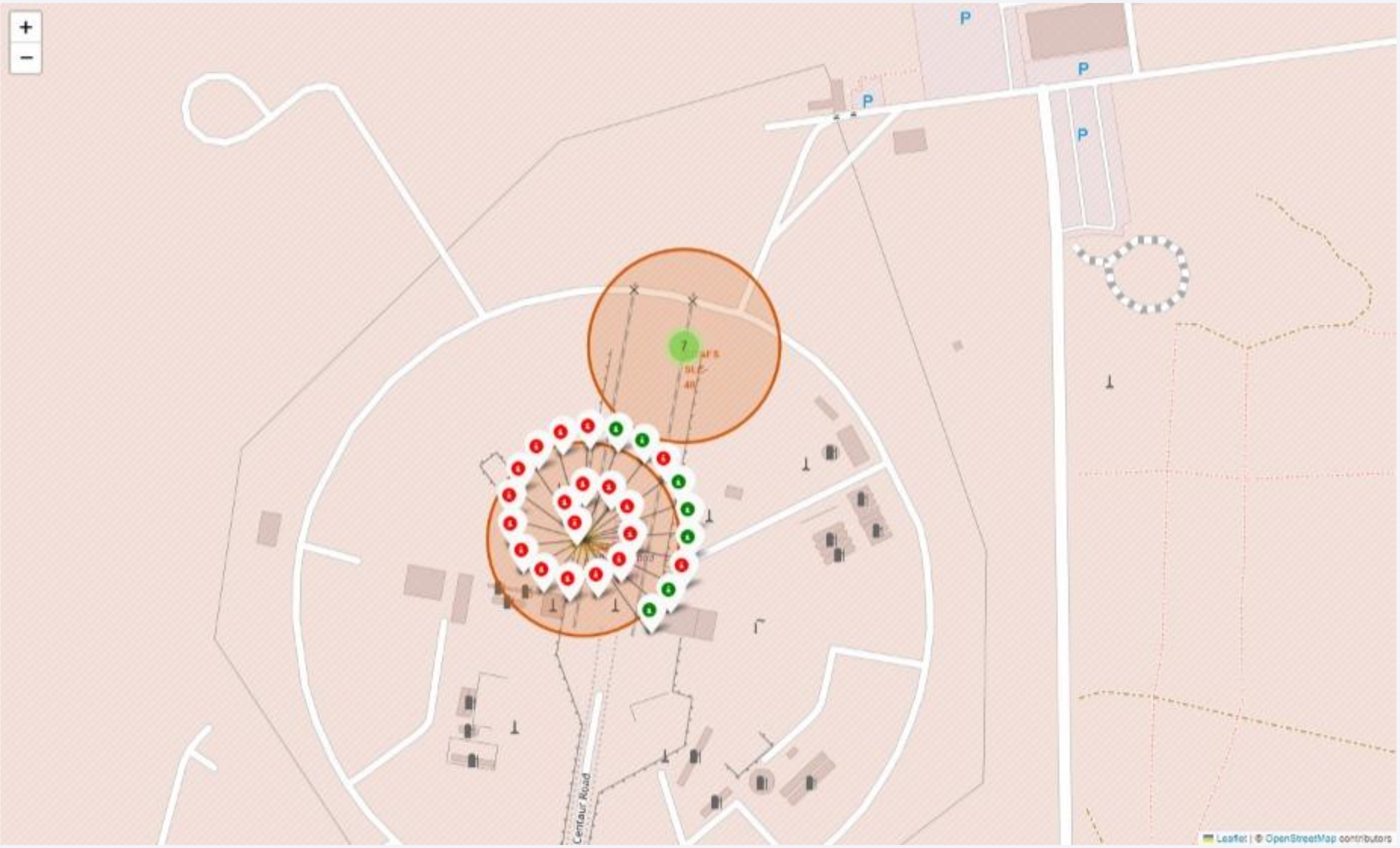
# Launch Sites Proximities Analysis

# Map of Launch Sites



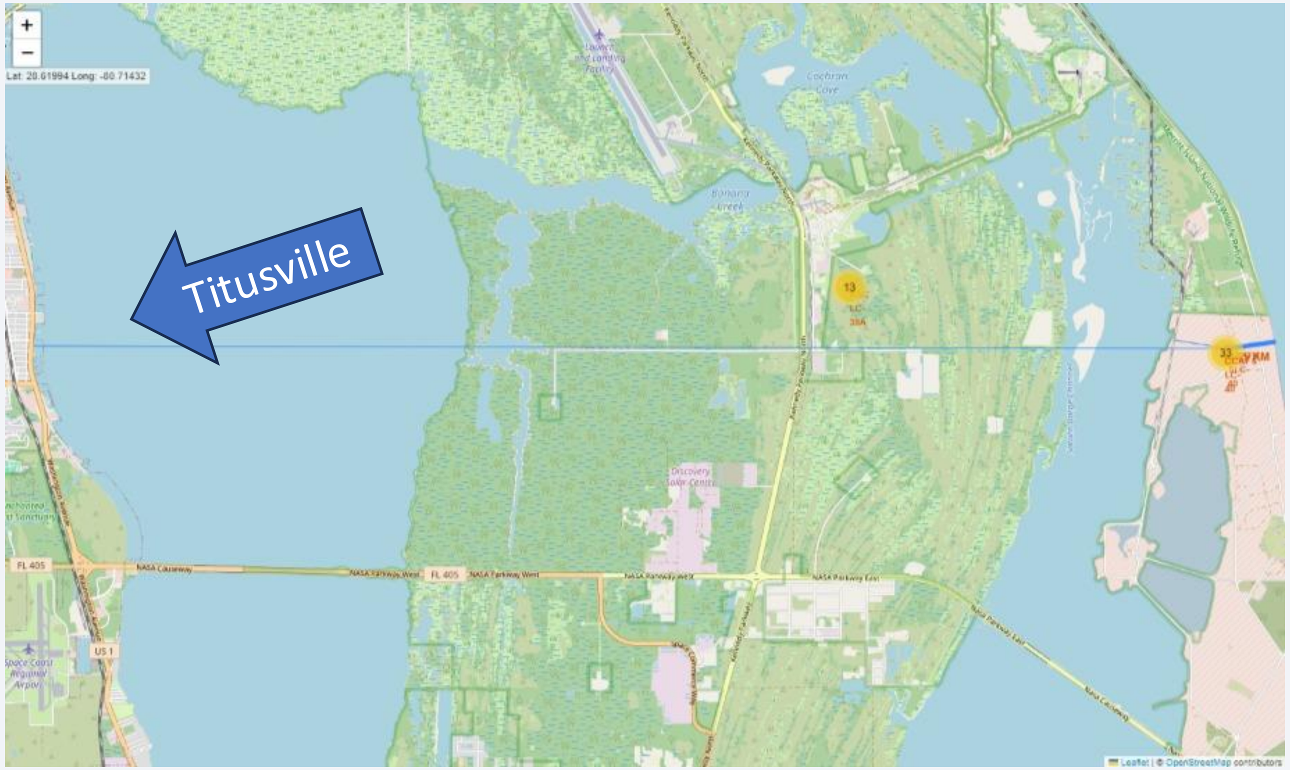
- Note that all indicated sites are close to a coastline.

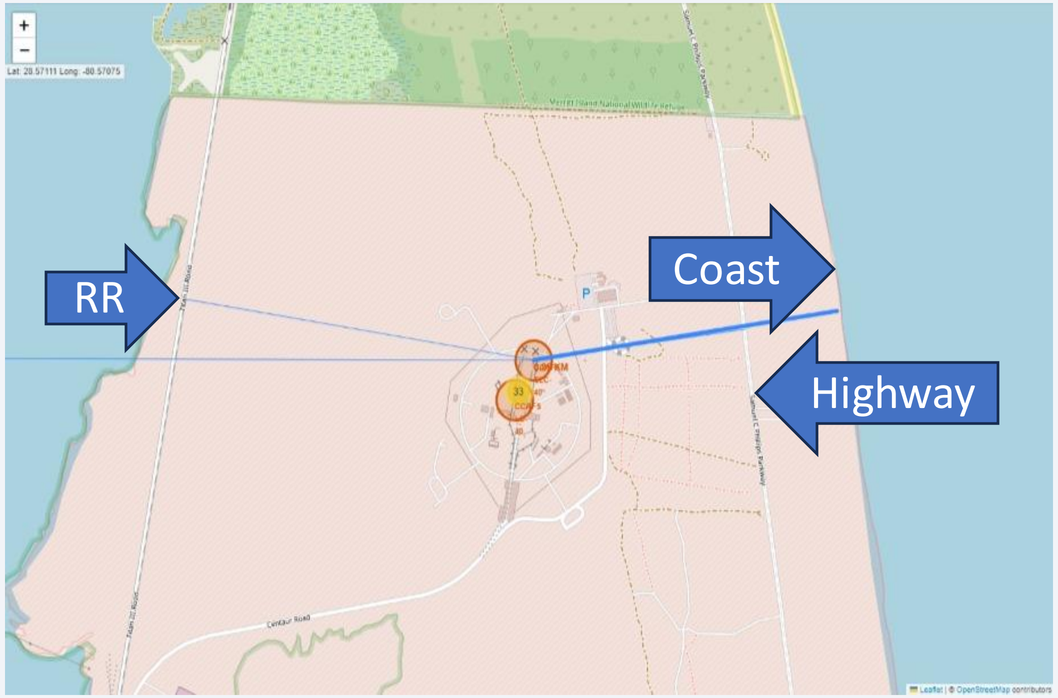
# Launch Outcomes at a Representative Site



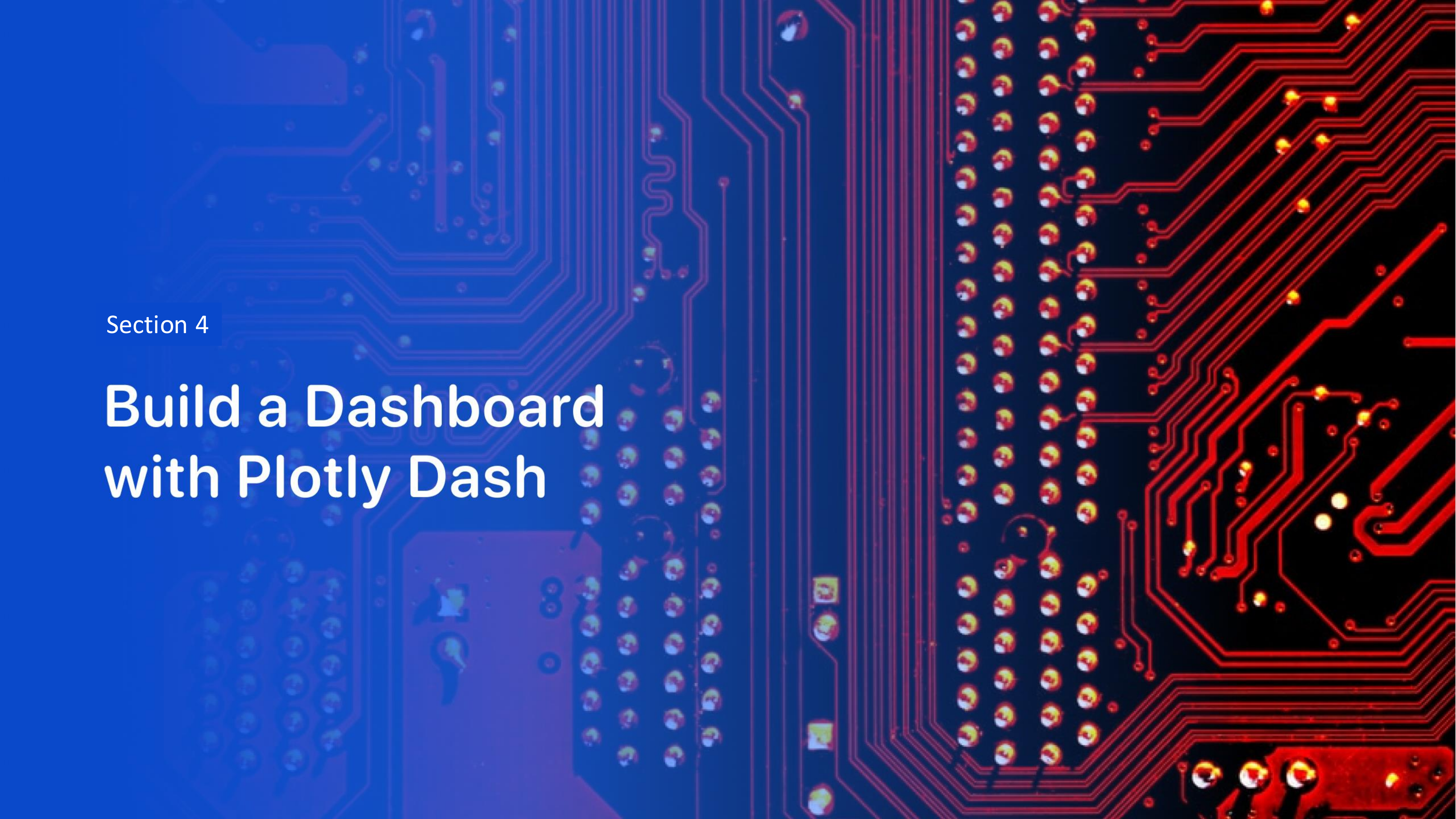
- Note that initially there are many failures, but that there are more successes later for this site.

# Nearby Points of Interest for CCAFS SLC-40


+
−
Lat: 28.61994 Long: -80.71432
Titusville
Leaflet | © OpenStreetMap contributors


+
−
Lat: 28.57111 Long: -80.57075
RR
Coast
Highway
Leaflet | © OpenStreetMap contributors

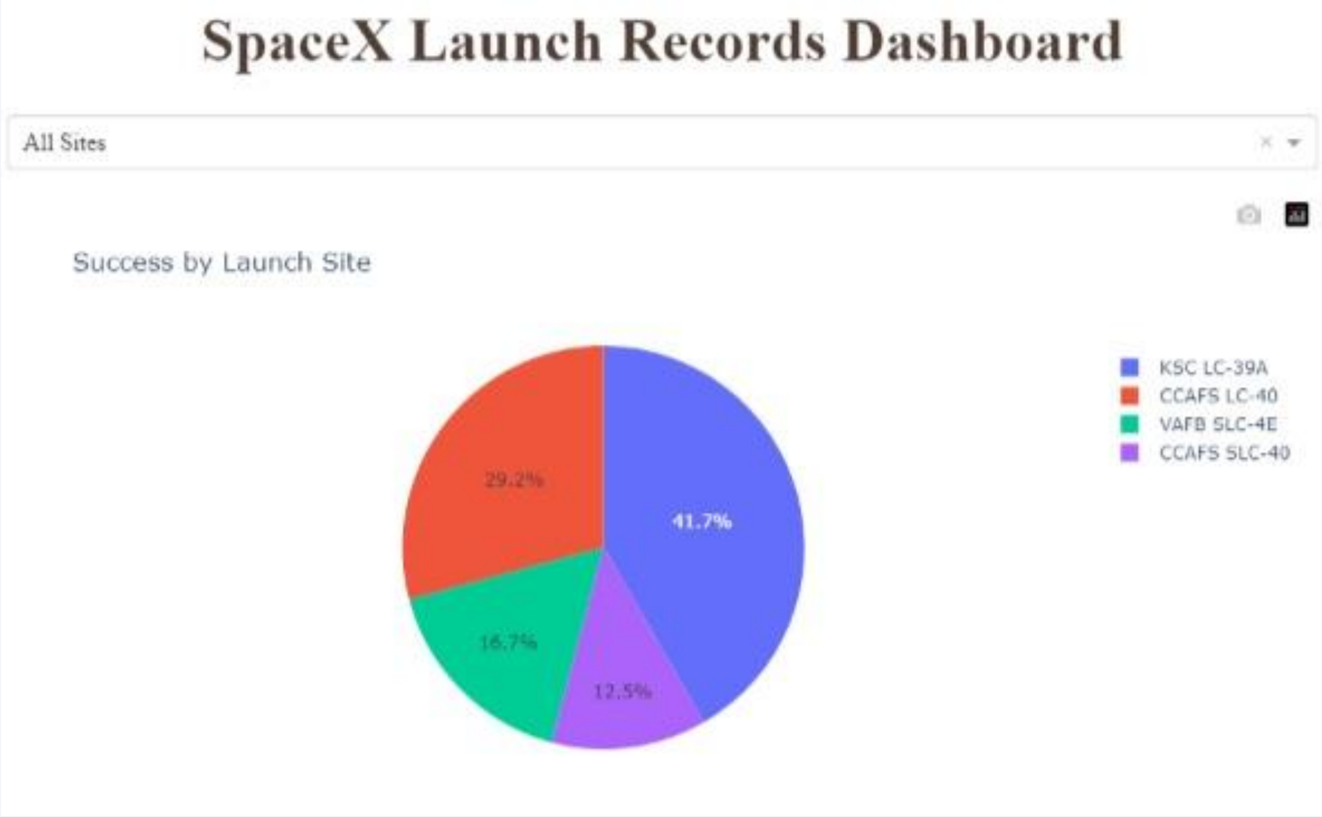
- Maps show that the nearest town is some distance away, but railroads, roads, and the coastline are quite close. Distances are not visible from this distance but can be viewed by zooming in on the site more closely.



Section 4

# Build a Dashboard with Plotly Dash

# Which Site has the Most Successes?



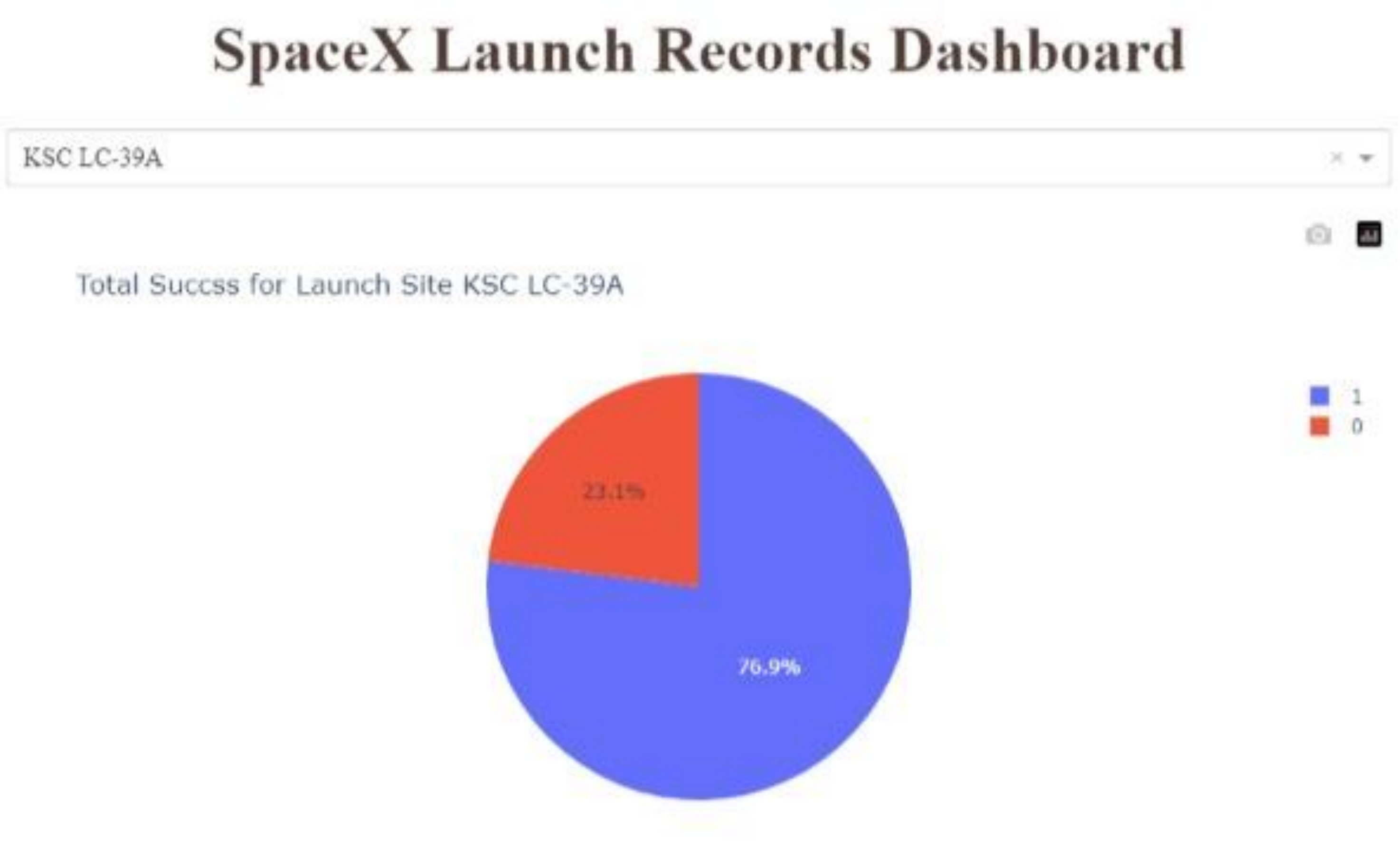
SpaceX Launch Records Dashboard

All Sites

Success by Launch Site

- This is a screenshot of the piechart of the percent of overall launch success that can be attributed to each of the sites
- Note that KSC LC-39A accounts for the most, at 41.7% of all successful launches.

# Percent Successful for the Most Successful Site

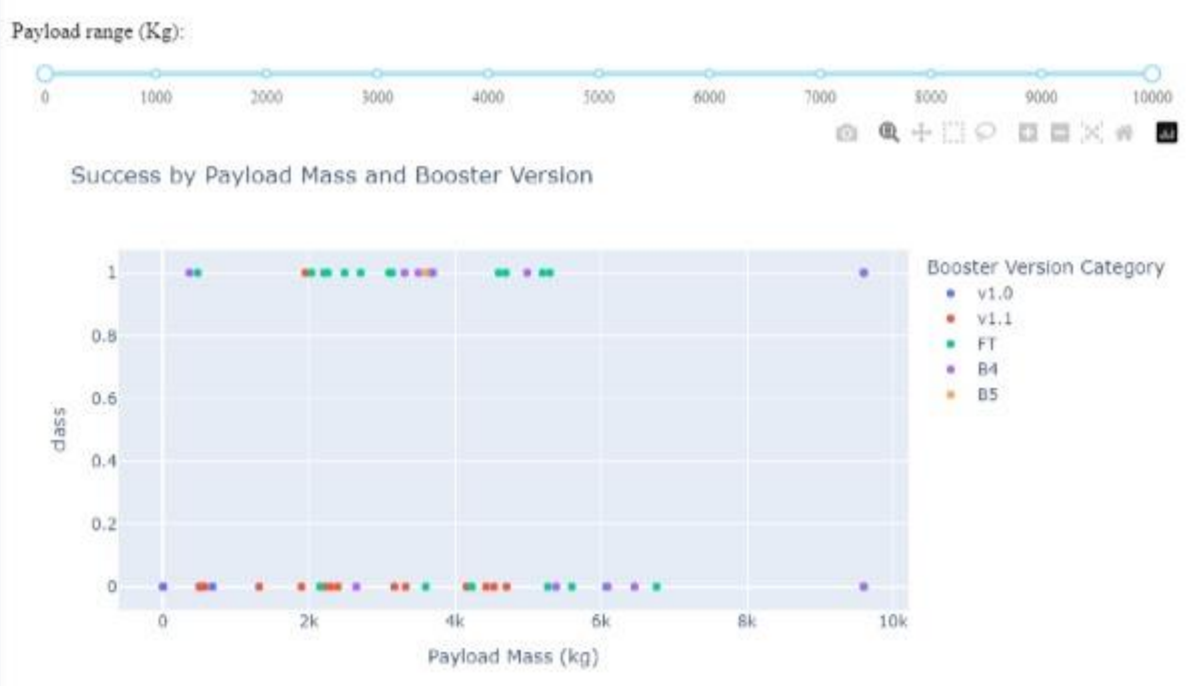


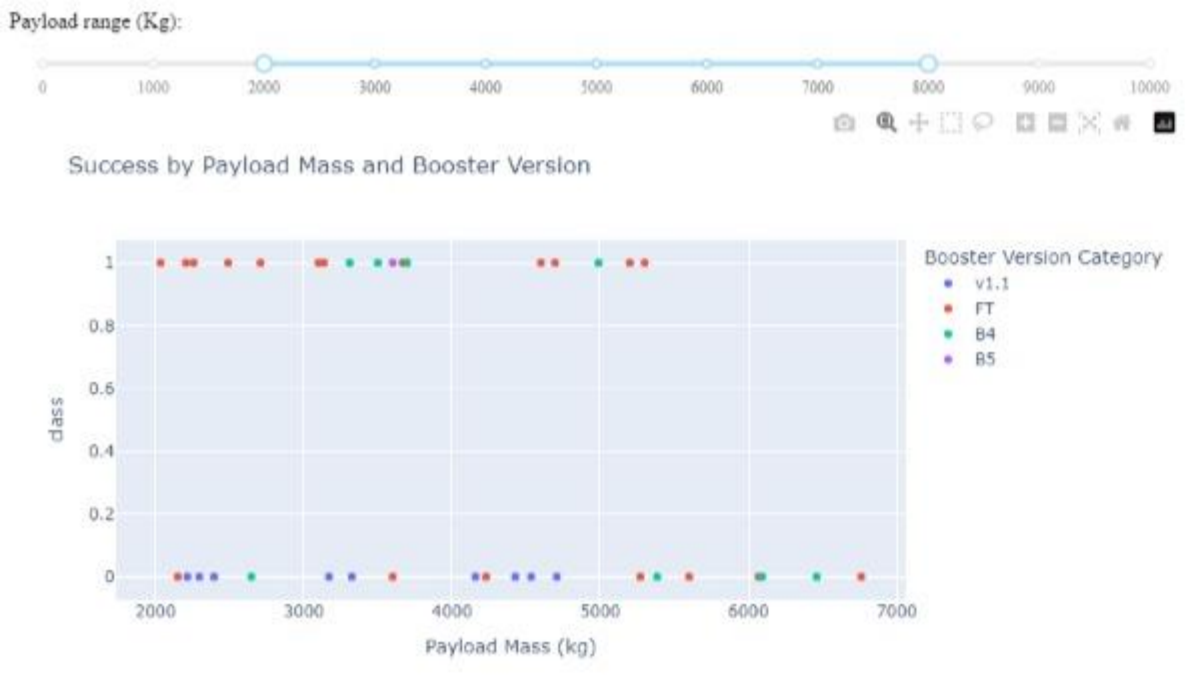
SpaceX Launch Records Dashboard

KSC LC-39A

- The most successful site, by percent, is KSC LC-39A
- Pictured here is the pie chart of success/failure for this site.
- 76.9% of launches from this site are successful

# Payload versus Launch Outcome for All Sites





- Payload vs. Launch Outcome scatter plot for all sites. The left plot shows all payloads, while the right shows only payloads between 2,000 and 8,000 kg (removing extremes).
- Note that payload alone does not overall seem to be a good predictor, though the second plot seems to show 5,000 to 7,000 to be a problematic range.

Section 5

# Predictive Analysis (Classification)

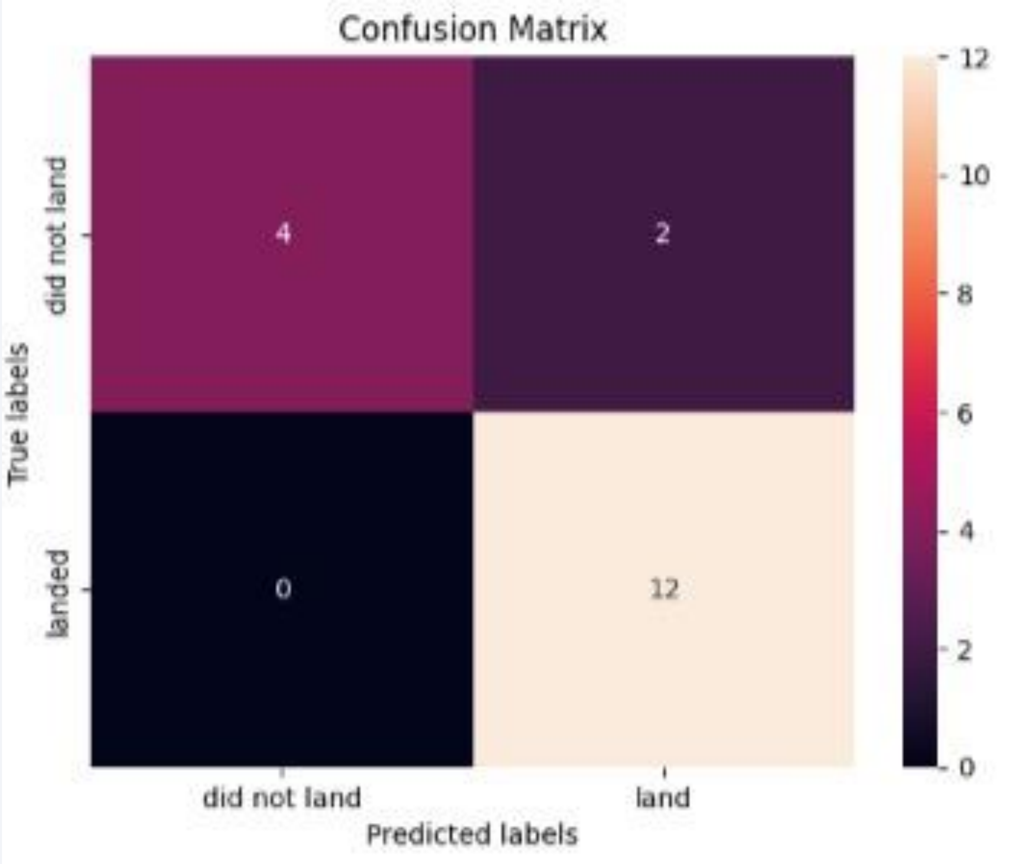
# Classification Accuracy

- The four models had similar accuracy, but did differ slightly. This can be seen in the following table:

| Classification Model | Accuracy | Score |
|---|---|---|
| Logistic Regression | 0.8464285714285713 | 0.8333333333333334 |
| Support Vector Machine | 0.8482142859142856 | 0.8333333333333334 |
| Decision Tree Classification | 0.8767857142857143 | 0.8888888888888888 |
| K-Nearest Neighbors | 0.8482142857142858 | 0.8333333333333334 |

- The Decision Tree Classification model is the best model of those tested and trained on the split data.

# Confusion Matrix



- For the Decision Tree Classification model indicated on the previous slide, this is the confusion matrix.
- For successful landings, the model correctly predicted all 12 of these in the testing dataset.
- For failures, the model only correctly predicted 4 of 6 in the testing dataset.
- There were no false negatives, but two false positives from the model.

# Conclusions

- Success has become more likely over time.
- The KSC LC-39A site had the highest percentage of successful landings.
- The number of orbits SpaceX has offered has increased over time.
- Launch sites tend to be near coastlines, and have access to transportation, but are not generally close to their nearest town.
- The Decision Tree Classification is the best at predicting the landing outcome, but has some difficulty with predicting when we will have a failure, while generally being able to correctly predict successful landings.

# Appendix

- Full copies of all code, notebooks, and output considered in this report, and this presentation, can be found on GitHub at the following link: https://github.com/Dr-Wilcock/DataScienceCapstone
- Guidance for this project was done through the Coursera Course.

# Thank you!