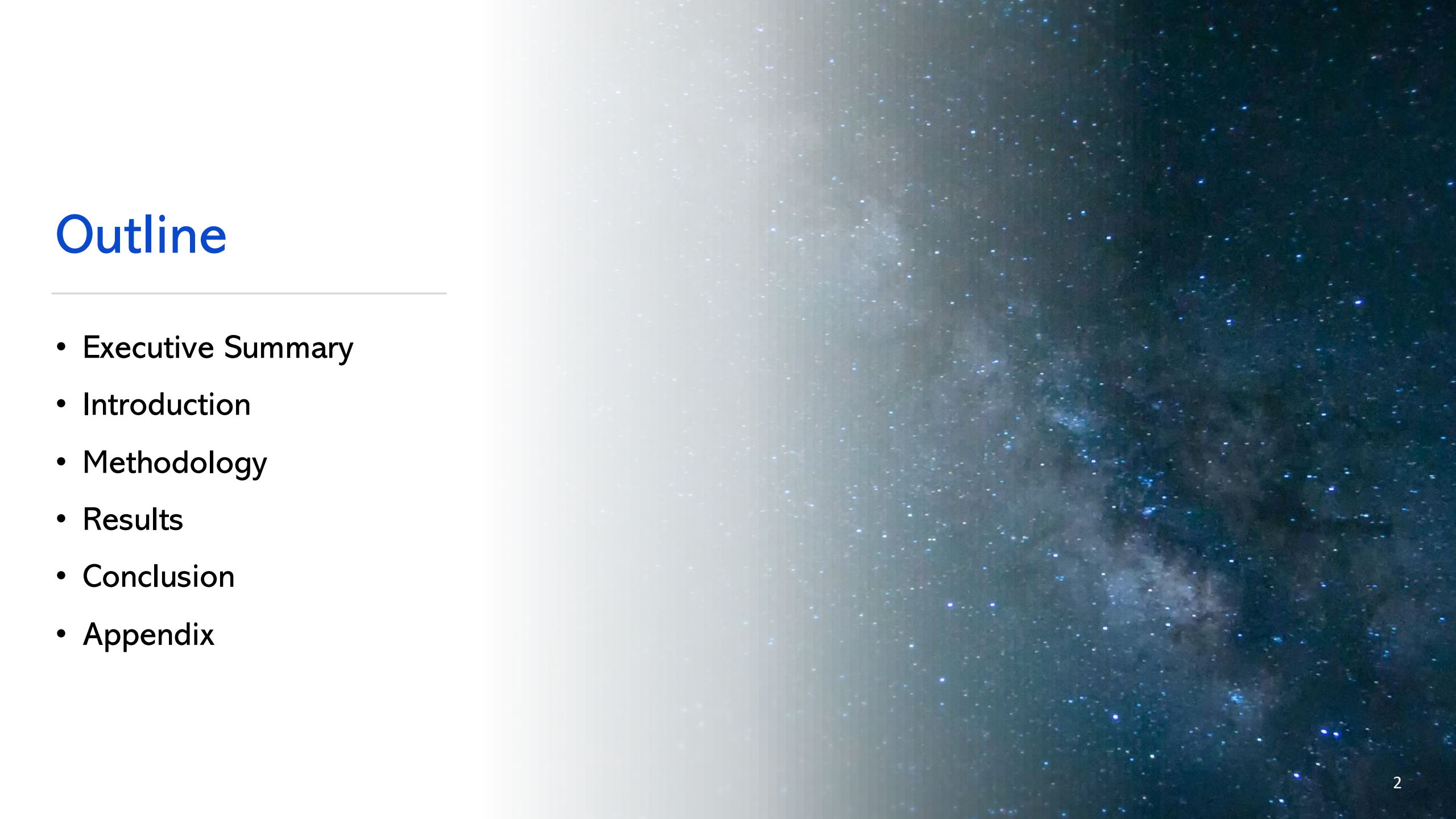
IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Benjamin Kon
9/12/2023

Outline

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



Executive Summary

Using data obtained via the Spacex API and through webscraping techniques, this analysis seeks to gain insight into the success of Spacex missions, particularly its introduction of reusable first stage rockets.

The data was cleaned and pre-processed using Pandas and Numpy to locate null values in the data; obtain a count of missions by launch site, orbit, and landing outcomes; and finally convert landing outcomes to a binary field for later use in predictive models.

Exploratory analyses using SQL queries and Python libraries including Seaborn, Folium, and Dash begin to answer questions and visualize Spacex missions, launch site locations, orbits followed, payload sizes, and whether the rocket was successfully landed for reuse.

Finally, machine learning models are trained on the data and evaluated to assess whether such models could be used to predict the success of future missions.

Findings show that the highest rate of successful landings was achieved by Kennedy Space Center (76.9%), while Cape Canaveral Space Launch Complex is notable for having successfully landed the largest payload. Generally, payloads in the range of 2,000-4,000 Kg had the greatest number successful landings, while the booster type with the most successful landings was the FT booster rocket.

Of the predictive models trained on the data, a decision tree classifier with optimized hyperparameters was found to have the highest measure for accuracy at 88.8%. However, when run on the test data all four models performed equally well, each with 83.3% correct predictions when compared against the actual target values. The confusion matrices for all four models were also equivalent with errors tending toward false positives rather than false negatives.

Introduction

SpaceX aims to be the most successful and cost effective company for future space travel. Its claim of being able to conduct spaceflight at $62 million per launch, just over 1/3 the price of competitors, largely hinges on the ability to land and reuse its first stage rockets. Numerous successes using this technique, particularly with the deployment of its newest Falcon 9 rockets, indicate significant progress toward these aims.

However, what are the details of these successes? Which launch sites have had the most successful landings? Which orbits did successful missions follow? Is there an optimal payload size? Furthermore, based on the data collected, can predictive models be used to determine whether future launches will be successful?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Spacex API.
  - Webscraping.
- Perform data wrangling
  - Convert collected data to generate binary target variable.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Split data into training and test sets.
  - Determine optimal parameters and test models for accuracy.

Data Collection

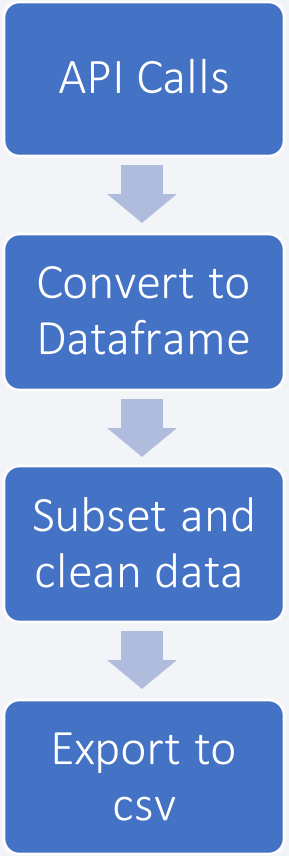
- Data was collected using a connection to the Spacex API.
- Additional data was collected using web-scraping techniques on a Wikipedia page with detailed information about Spacex launches.

Data Collection – SpaceX API

- The API connection was established using Python requests.
- The response was then converted to JSON and then to a Pandas Dataframe using json_normalize().
- Finally, a subset of the dataframe was created containing only the desired fields and custom functions were used to further extract the required information from the raw data.
- Cleaned data was exported to csv.

Completed Jupyter Notebook showing API calls and results.

CSV file containing the extracted data.

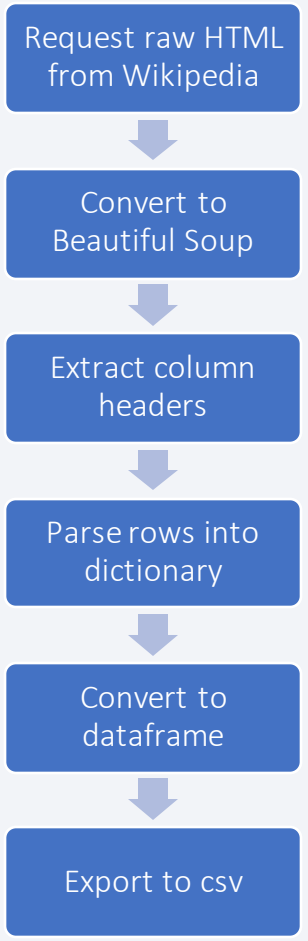


Data Collection - Scraping

- Python get request was initiated to Wikipedia page on Falcon 9 and Falcon Heavy Launches.
- Response was converted to a Beautiful Soup object and column names were extracted from HTML table headers.
- Table rows were parsed into a dictionary using custom functions.
- Dictionary was converted to a Pandas Dataframe and then exported to csv.

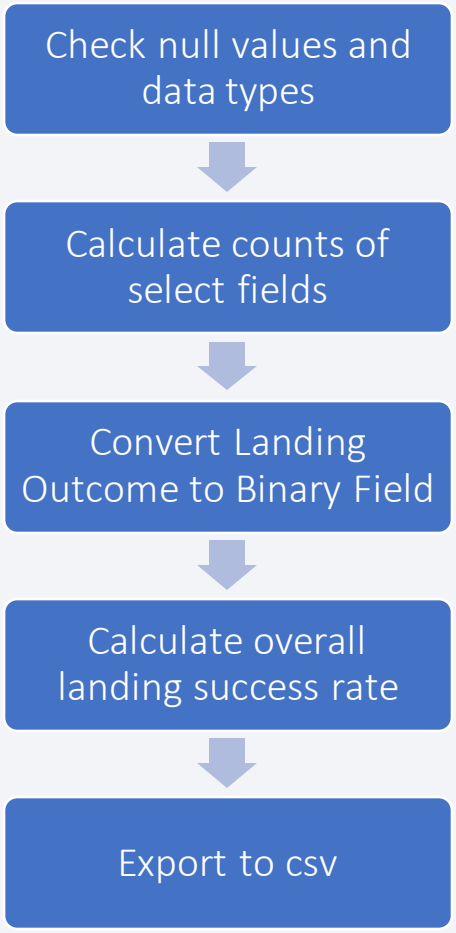
Completed notebook showing webscraping process.

CSV file with scraped data.



Data Wrangling

- Dataset was examined for null values and checked for correct data types.
- Counts were taken of each value in the fields Launch Site, Orbit, and Landing Outcome.
- Data was pre-processed to convert the success or failure of each launch to a binary variable for later use in training predictive models.
- The overall landing success rate was calculated to be 66.7%
- Export resulting data set to csv.
- Jupyter Notebook on Data Wrangling.
- CSV file with pre-processed and cleaned data.



EDA with Data Visualization

Plots used for analysis:

- Scatter plot of Launch Number versus Launch Site, color-coded by landing success value.
- Scatter plot of Payload versus Launch Site, color-coded by landing success value.
- Bar plot of success rate for each orbit type.
- Scatter plot of Launch Number versus Orbit, color-coded by landing success value.
- Scatter plot of Payload versus Orbit, color-coded by landing success value.
- Line plot showing change in success rate over time.

Purpose:

- Visually examine for correlation between variables.
- Visually examine for correlation between variables.
- Find orbits with best landing success rates.
- Visually examine for correlation between variables.
- Visually examine for correlation between variables.
- Examine the overall trend in landing success through out the Spacex program.

Jupyter Notebook – Exploratory Data Analysis with Data Visualization.

CSV containing one-hot variables in preparation for predictive modeling.

EDA with SQL

SQL queries were performed to find:

- Names of all launch sites.
- Data from five launches, where the launch site begins with 'CCA'.
- Total mass of payloads carried by boosters from NASA.
- Average payload mass carried by booster version F9 v1.1.
- Date of the first successful landing outcome on ground pad.
- Boosters that successfully landed on drone ship with payload between 4,000 and 6,000 Kg.
- Number of successful and failure mission outcomes.
- Boosters which have carried the maximum payload mass.
- Launches with failed landings in drone ship during 2015.
- Ranked count of landing outcomes between June 4, 2010 and March 3, 2017.
- Jupyter Notebook containing Exploratory Data Analysis with SQL.

Interactive Map with Folium

Map Features:

- Blue circular markers corresponding to latitude and longitude of launch sites.
- Red and green rocket markers with clustered together by launch site.
- Blue lines labeled with measured distances in red.

Purpose:

- Locate launch sites on the map and examine surrounding geographical features.
- Visually display successful and failed landings for each launch site.
- Show distance between launch sites and geographic features of interest.

Jupyter Notebook containing interactive Folium Map.

Dashboard with Plotly Dash

Dashboard Features:

- Pie Charts
- Scatter Plots
- Dropdown Menu
- Slider Menu

Purpose:

- Show overall share of success and individual rate of success for all launch sites.
- Visualize landing successes/failures for different launch sites and payload ranges.
- Select and update which launch sites are shown.
- Select and update the range of payloads to include.

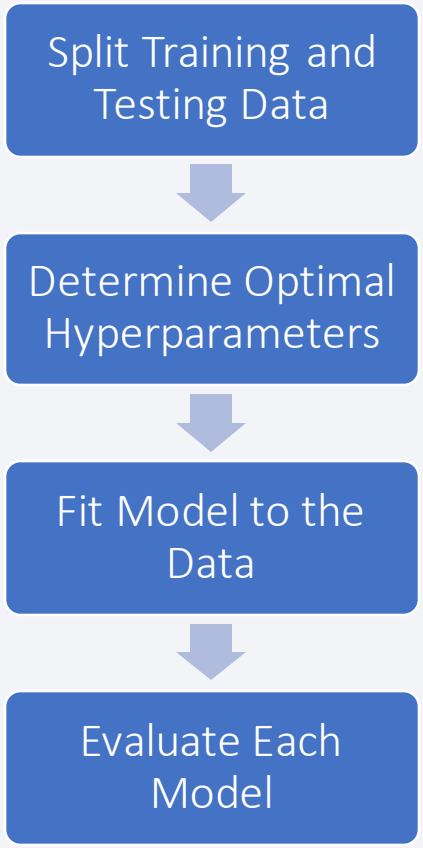
Github Link to Python Code for Dashboard

Provided Dataset Used for Dashboard

Predictive Analysis (Classification)

- Data was split into training and test sets to be run using four models.
- GridSearchCV was used to determine optimal hyperparameters for each model.
- Each model was fit to the training data set.
- Each model was tested for accuracy and the output was used to generate a confusion matrix.

- Jupyter Notebook containing complete predictive analysis.

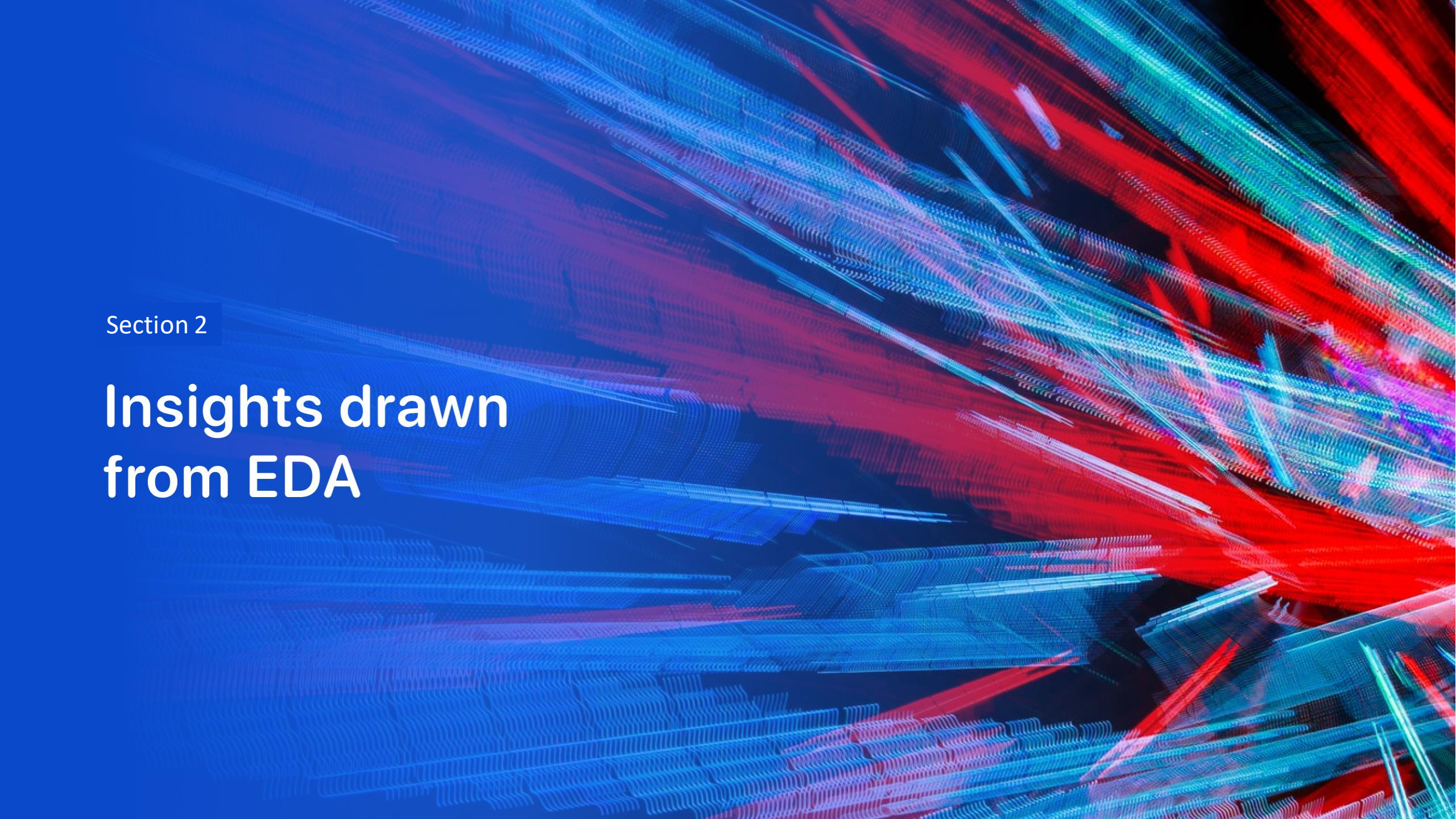


Results

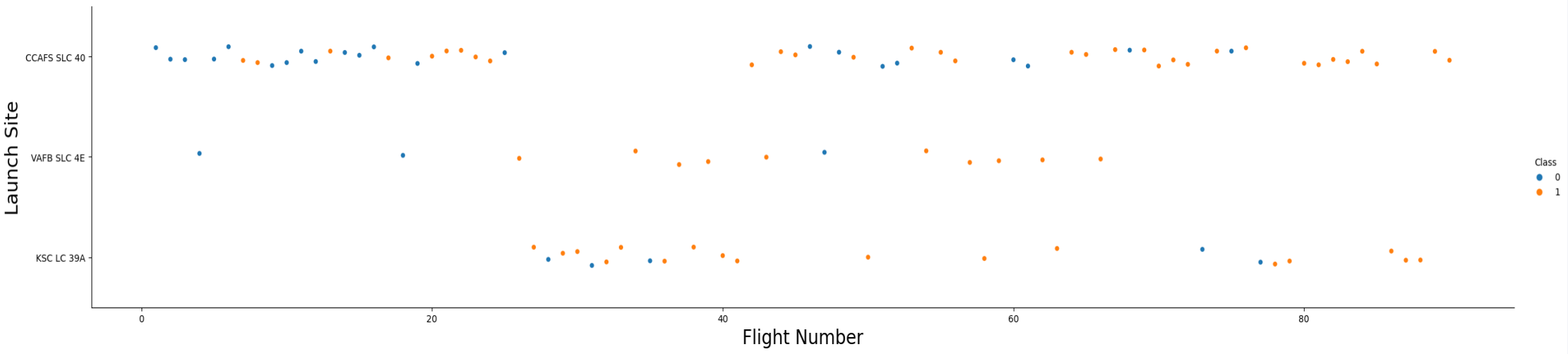
- Exploratory data analysis results
- Launch Sites Proximities Map
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

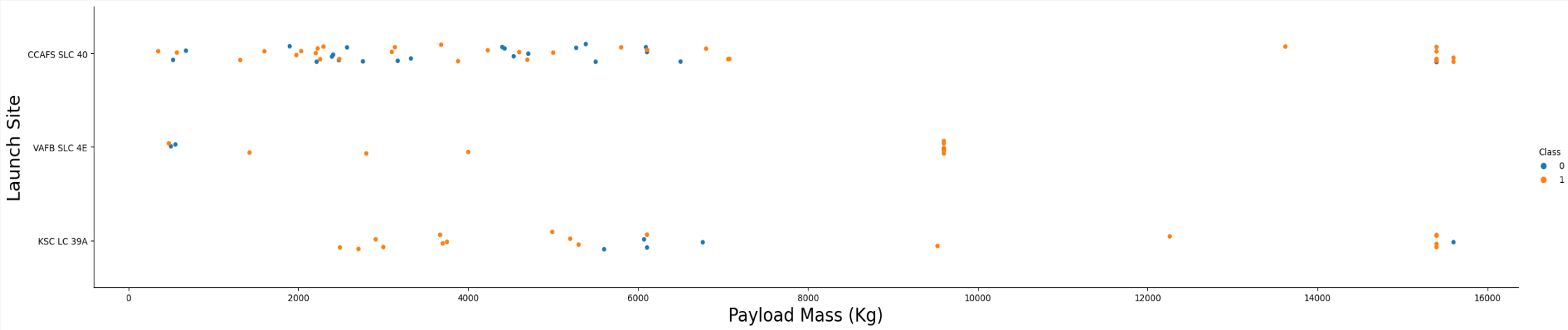


Flight Number vs. Launch Site



- Early flight numbers, corresponding with early years of the Spacex program, are primarily launched from Cape Canaveral Space Launch Complex and show mostly failures to land (class=0).
- Kennedy Space Center, with the highest success rate of all the launch sites, also seems to have had the advantage of beginning its launches after the initial failures at other launch sites.

Payload vs. Launch Site



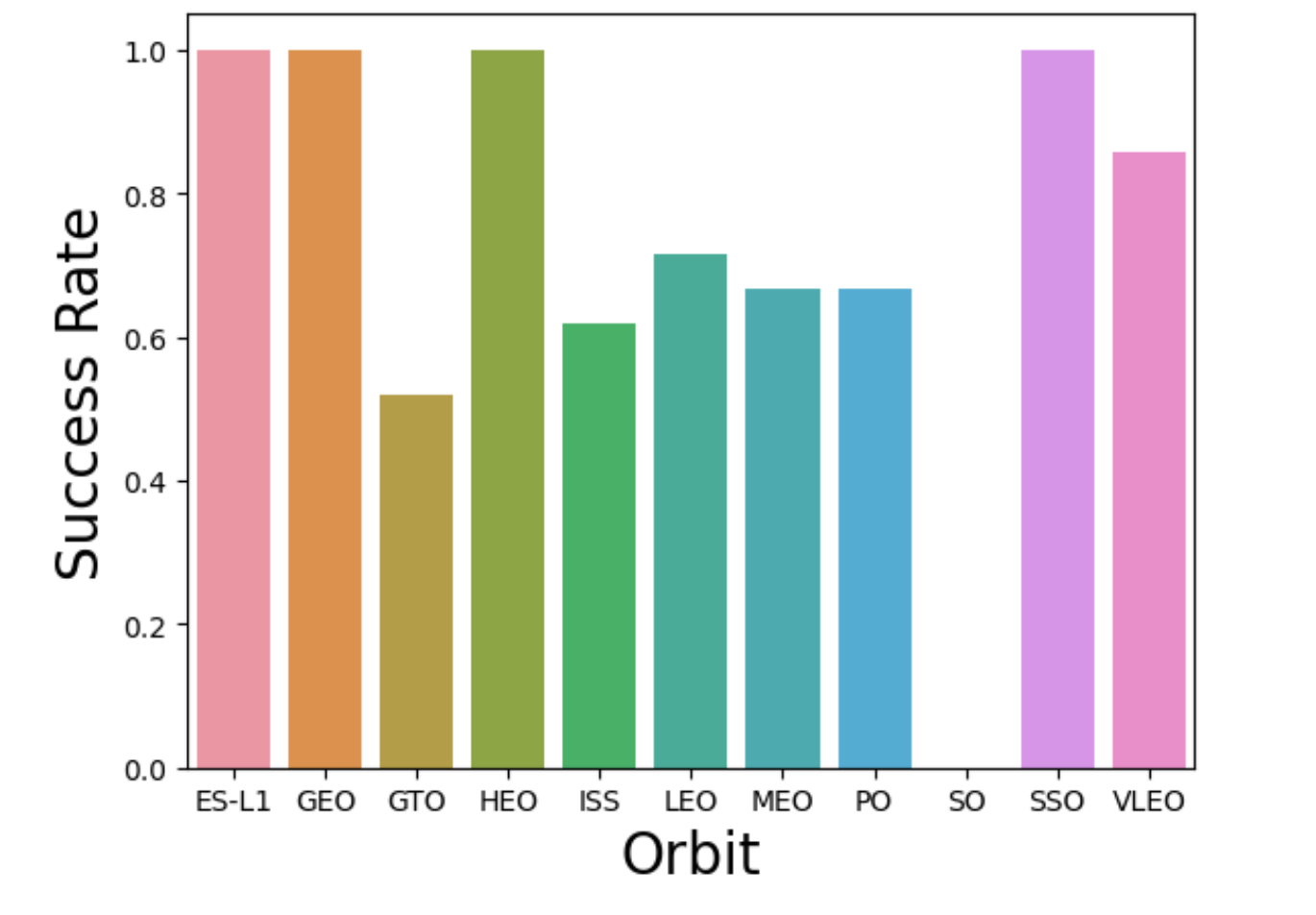
- Kennedy Space Center and Cape Canaveral Space Launch Complex were the launch sites using payloads greater than 10,000 Kg. Most of these landed successfully.
- Cape Canaveral Space Launch Complex appears to have successfully landed the largest payloads (15,600 Kg, *see Boosters Carrying Maximum Payload*).
- Significantly more attempts were made using payloads less than 8,000 Kg with landing results that were generally more mixed.

Success Rate vs. Orbit Type

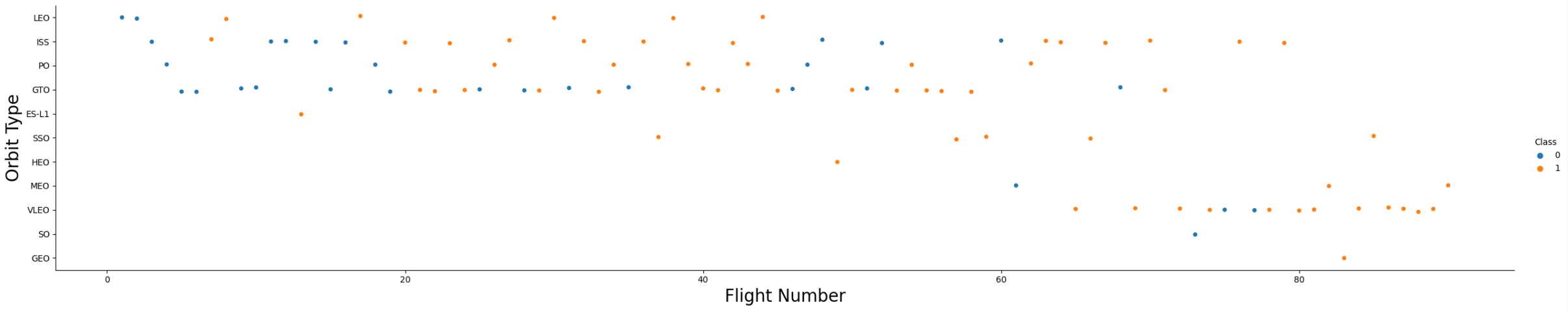
- ES-L1, GEO, HEO, and SSO orbits all have perfect, 100% success rates. However, three of those four orbits have only been tried once. VLEO, with success rate of 86% over 14 missions is notable.

Counts of each orbit from initial data wrangling:

GTO	27
ISS	21
VLEO	14
PO	9
LEO	7
SSO	5
MEO	3
ES-L1	1

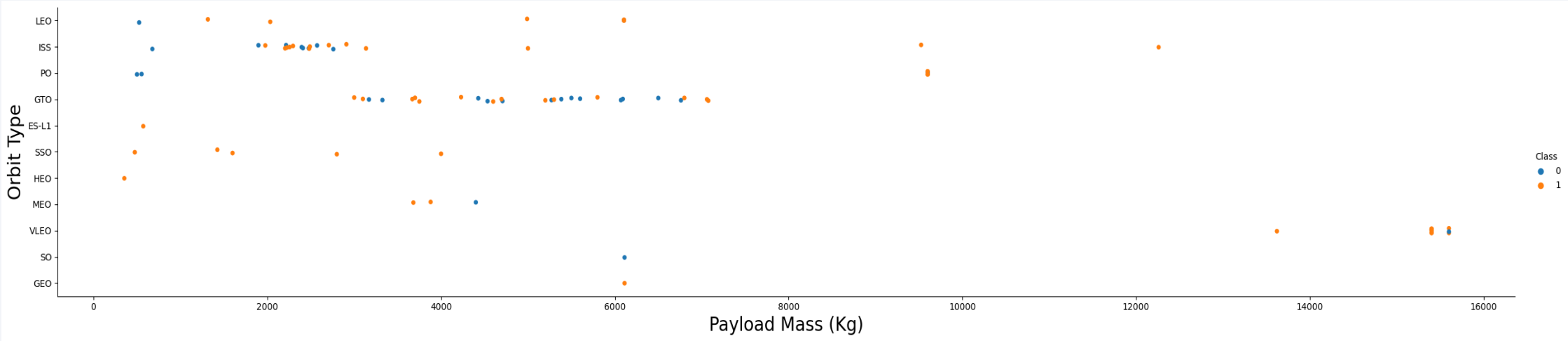


Flight Number vs. Orbit Type



- There is a noticeable cluster of failures (class=0) for lower flight numbers, corresponding to the early years of the Spacex program from 2010-2013 (see *Launch Success Yearly Trend*).
- The LEO orbit showed consistent success as flight numbers increased.

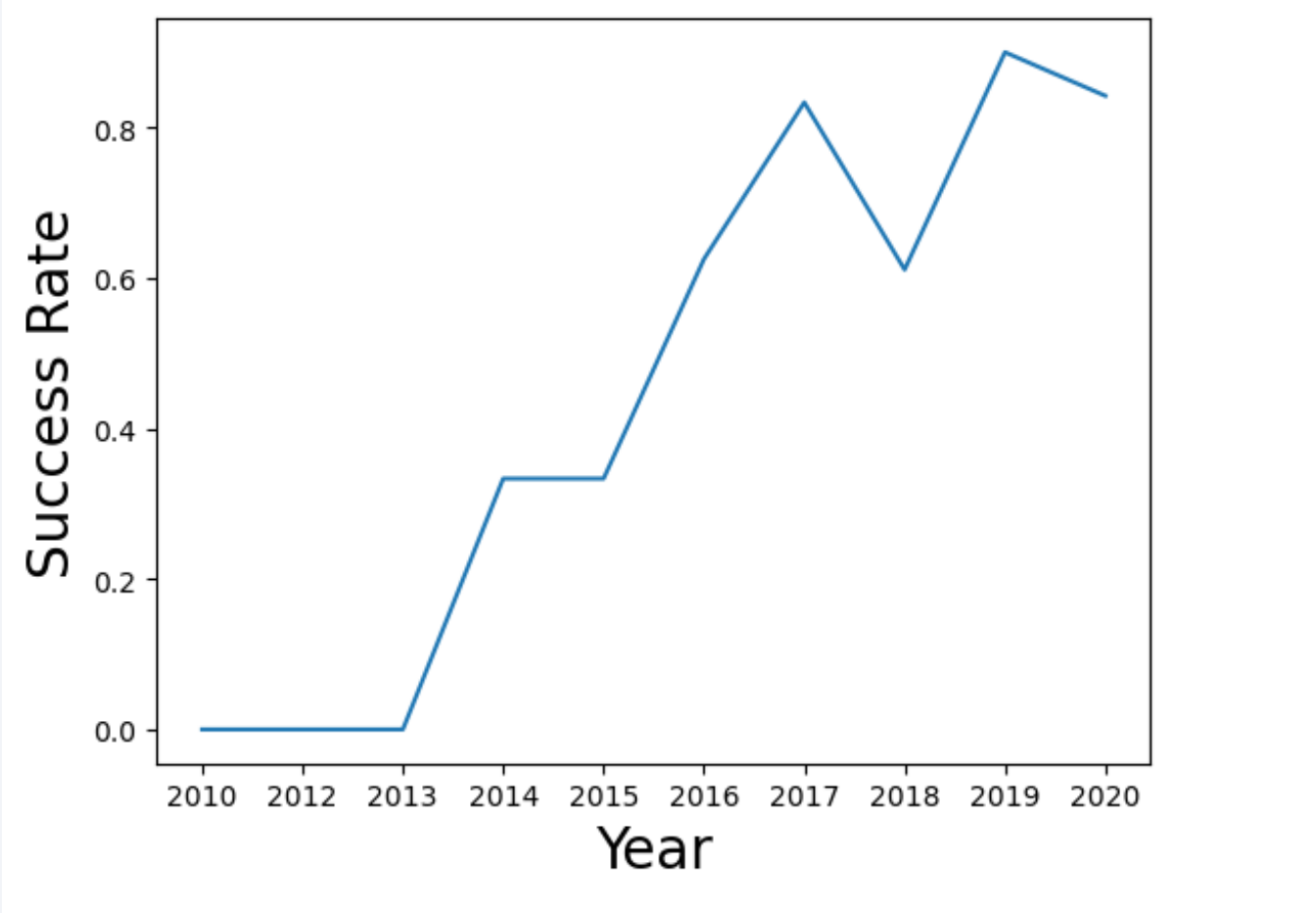
Payload vs. Orbit Type



- Heavy payloads had the most successful landings using the ISS, PO, and VLEO orbits, however, no other orbits had payloads above 8,000 Kg.
- All five successful landings following the SSO orbit had payloads of roughly 4,000 Kg or less.

Launch Success Yearly Trend

- The rate of successful landings has generally increased after 2013, reaching a maximum value of 90% in 2019.



All Launch Site Names

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

Cape Canaveral Launch Complex
Vandenberg Space Force Base
Kennedy Space Center
Cape Canaveral Space Launch Complex



Launch Site Names Begin with 'CCA'

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

- First five launches from Cape Canaveral. All landings failed or were not attempted.

Total Payload Mass

Customer	SUM(PAYLOAD_MASS__KG_)
NASA (CRS)	45596

- The total mass of payloads carried by boosters from NASA was 45,596 Kg.

Average Payload Mass by F9 v1.1

Booster_Version	AVG(PAYLOAD_MASS__KG_)
F9 v1.1	2928.4

- The average payload mass carried by booster version F9 v1.1 was 2928.4 Kg.

First Successful Ground Landing Date

MIN(Date)
2015-12-22

- The date of the first successful landing outcome on ground pad was December 22, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

Booster_Version	Landing_Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

- List showing the names of boosters which have successfully landed on drone ship with payload mass between 4,000 and 6,000 Kg.

Total Number of Successful and Failure Mission Outcomes

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

- Table showing the number of successful and failure mission outcomes*.

*Mission success does not necessarily imply and is not always dependent on a successful landing of the rocket.

Boosters Carrying Maximum Payload

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

- List of the names of boosters which have carried the maximum payload mass (15,600 Kg).

2015 Launch Records

Month	Year	Landing_Outcome	Booster_Version	Launch_Site
10	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- List of launches with failed landings in drone ship during 2015, including their booster versions and launch site names.

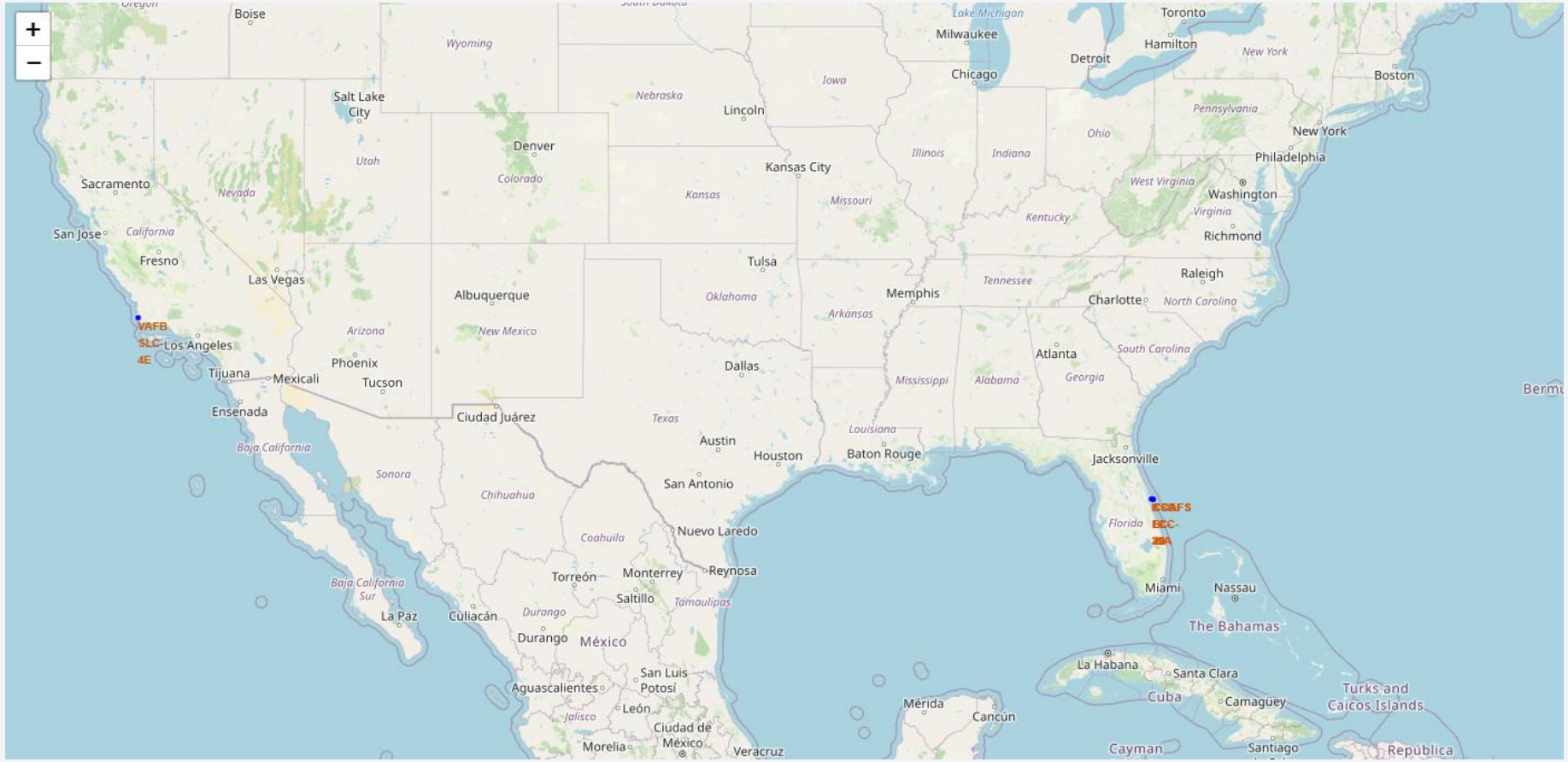
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- Ranking of the count of landing outcomes between June 4, 2010 and March 3, 2017, in descending order

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


Section 3
Launch Sites
Proximities Analysis

Launch Site Locations

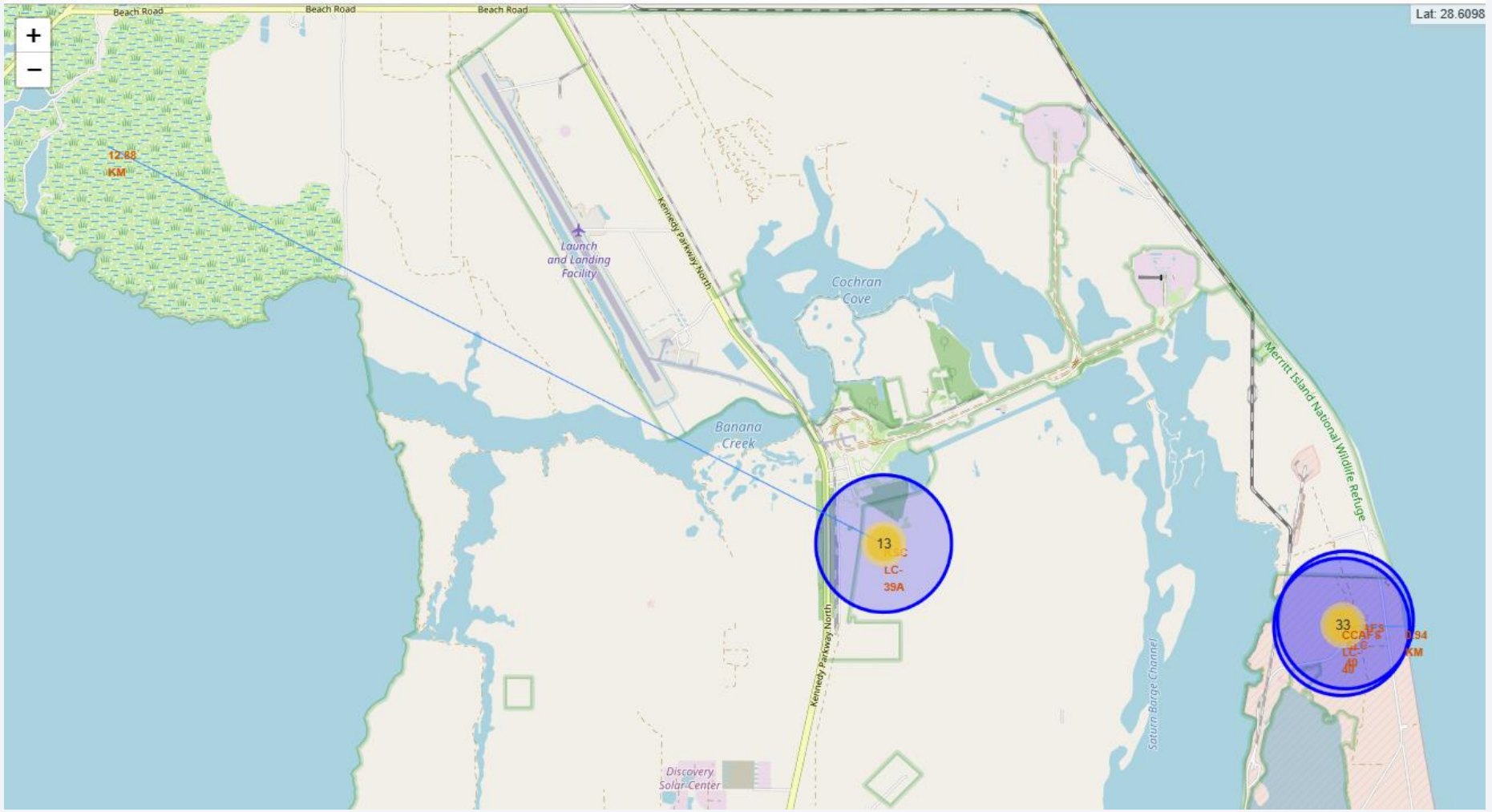


- All launch sites are located near coastal areas.
- Only Vandenberg Space Force Base is located on the U.S. Pacific West Coast, in the state of California.
- All other launch sites are found in close proximity on the U.S. Altantic East Coast, in the state of Florida.

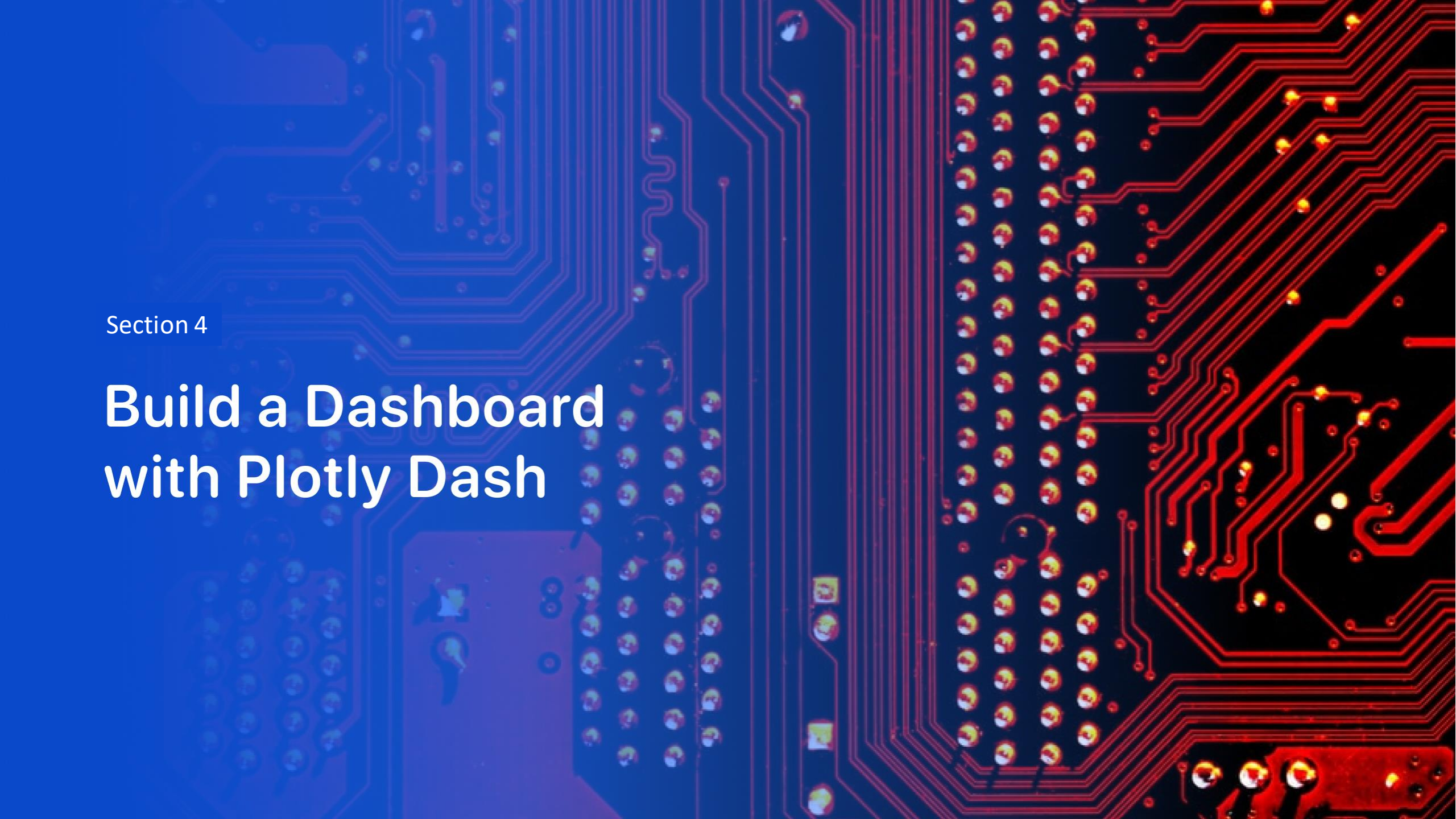
<Folium Map Screenshot 2>

- Map of Cape Canaveral Launch Complex with color-coded launch icons indicating success or failure of landing.

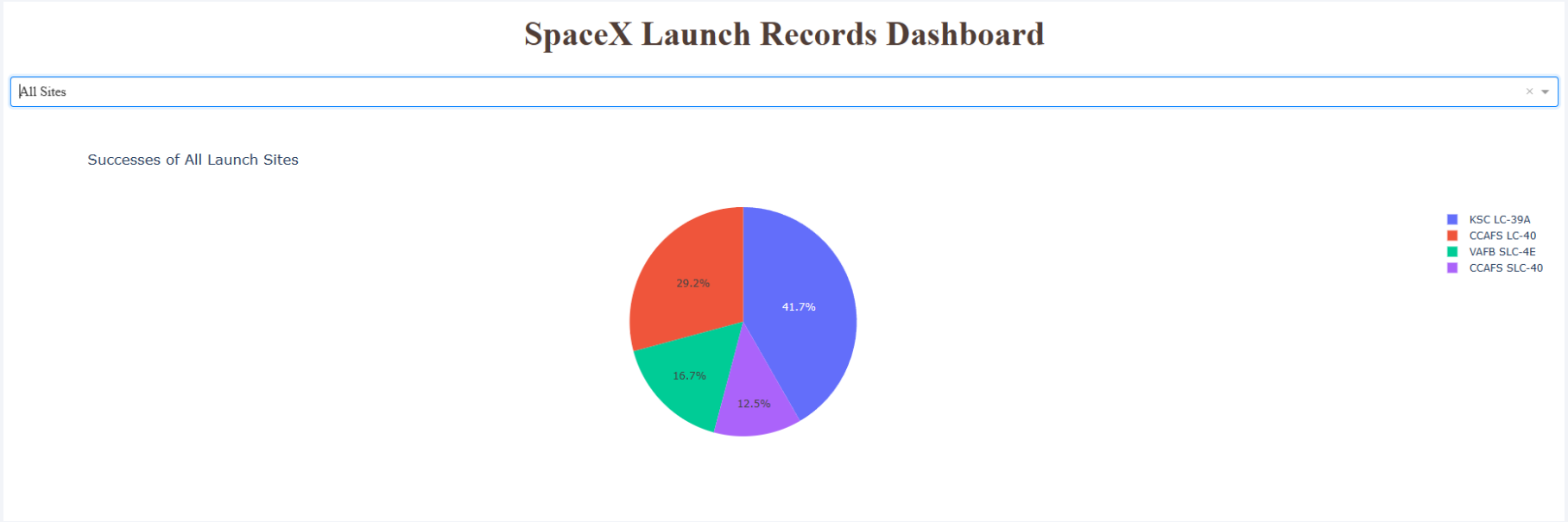
<Folium Map Screenshot 3>



- Map showing distance of Cape Canaveral Launch Complex to the coast (0.94 km) as well as proximity of Kennedy Space Center to the Merritt Island National Wildlife Refuge (12.88 km).

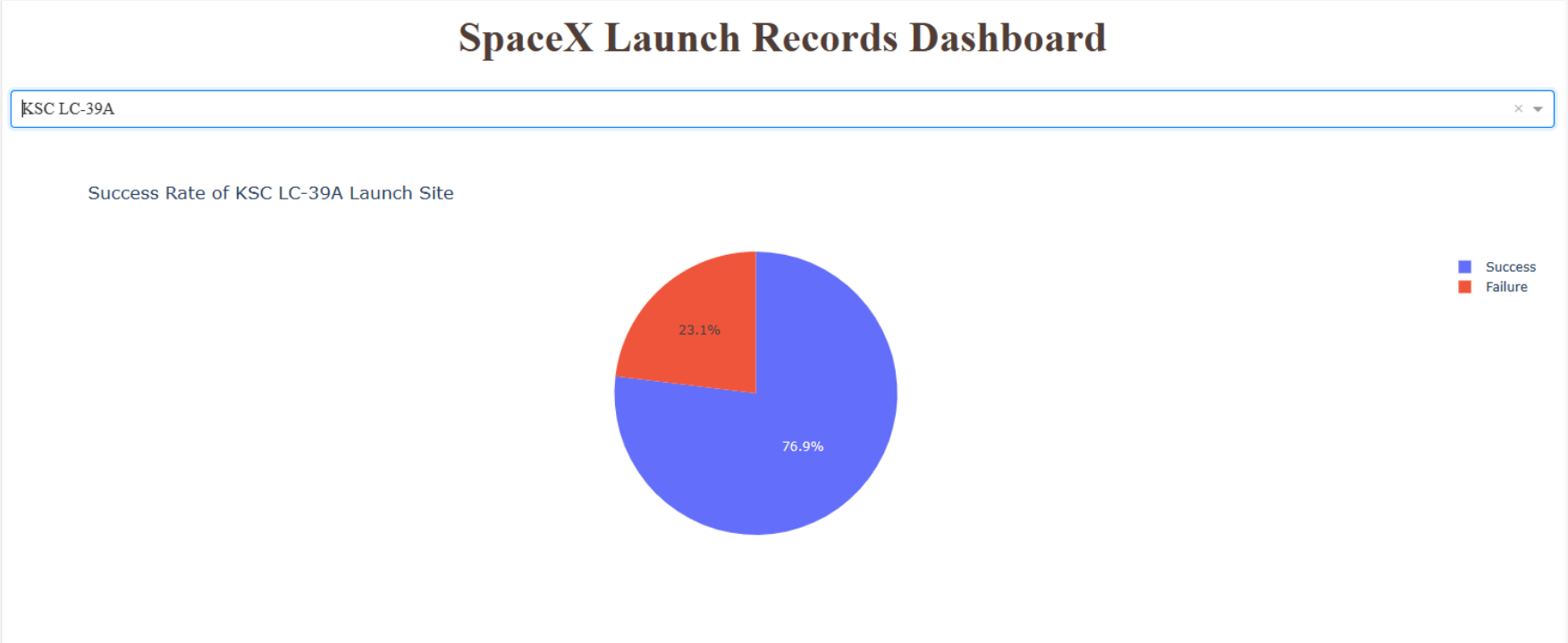

Section 4
Build a Dashboard with Plotly Dash

Success Rates of All Launch Sites



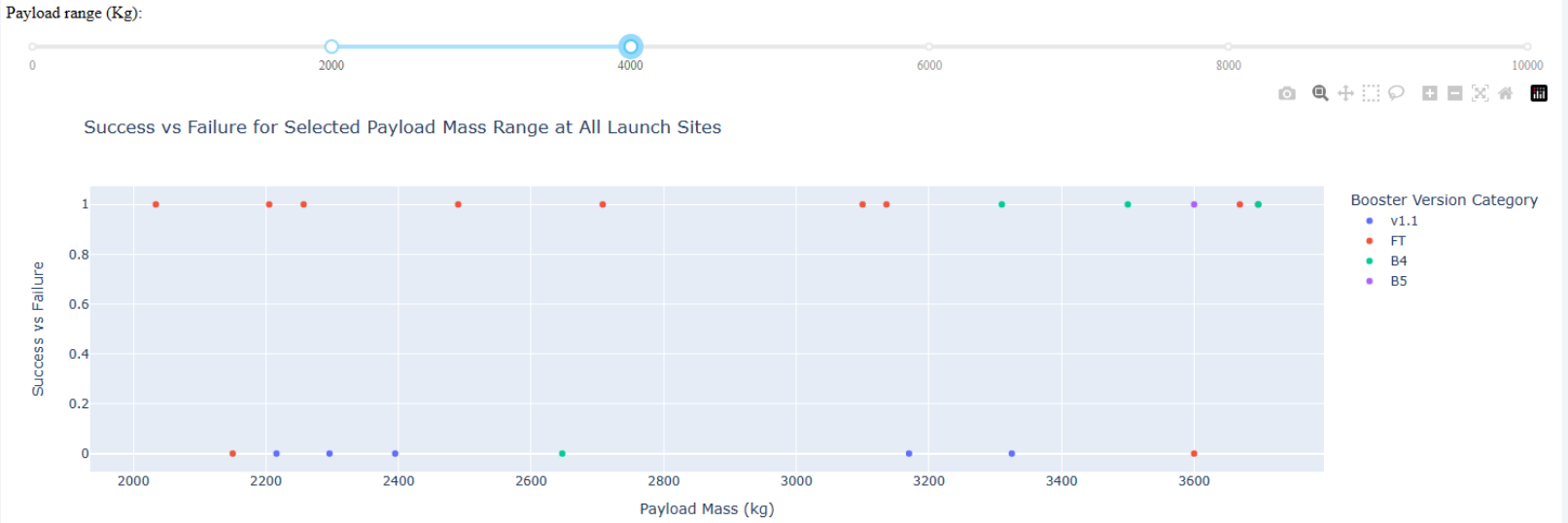
- Kennedy Space Center had the highest share of successful landings at 41.7% of the total for all landing successes.
- It is interesting to note that Cape Canaveral had an equivalent share of successes to KSC, when looking at the combined successes of the earlier CCAFS LC-40 site and the newer CCAFS SLC-40 site.

Launch Site with Highest Rate of Successful Landings



- Kennedy Space Center not only has the highest percentage of successes overall, but also the highest rate of success with 76.9% of its launches resulting in a successful landing.

Payload Mass vs Launch Outcome



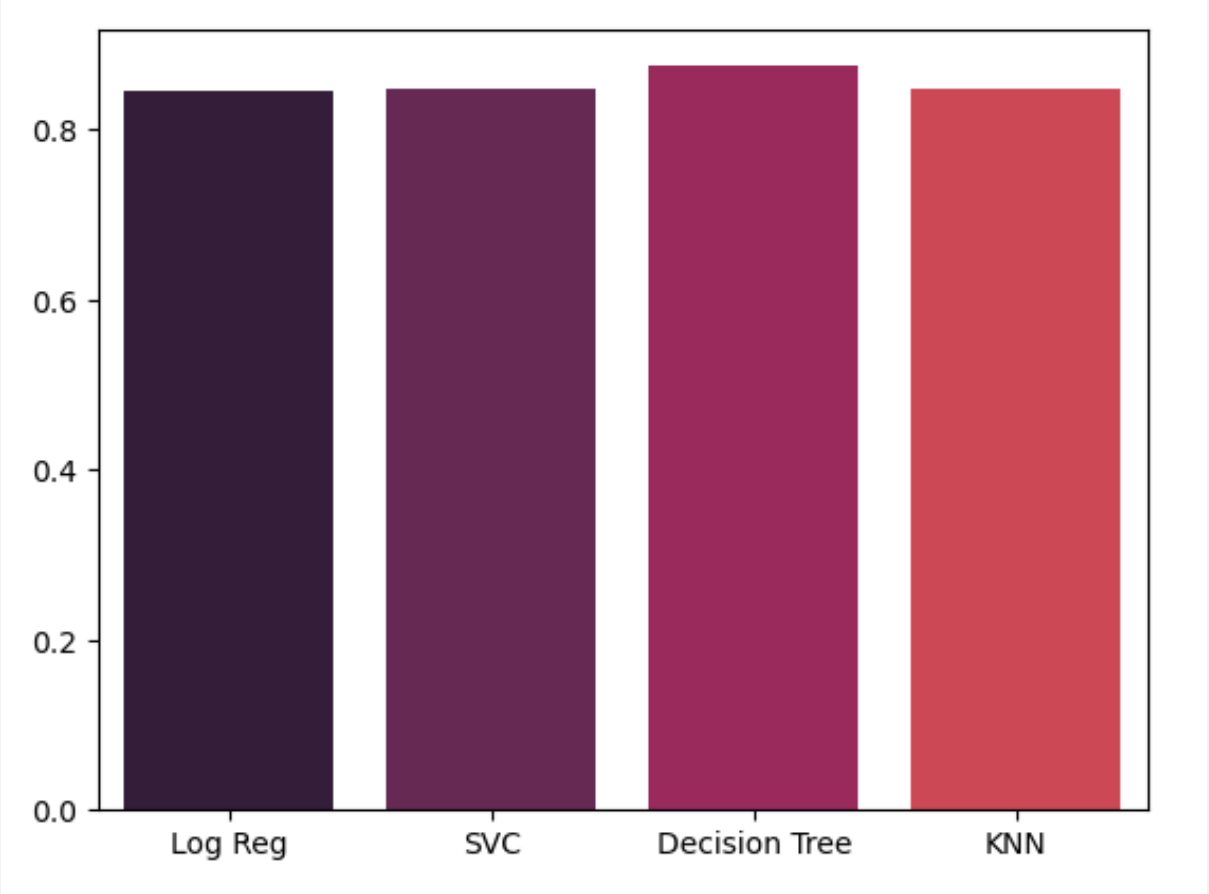
- From the scatter plots on the left it can be ascertained that the FT booster had a higher rate of success than other booster types.
- This is particularly evident with payloads between 2,000 to 4,000 Kg, a range notable for having the highest number of successful landings when examining the distribution of payload mass.

Section 5

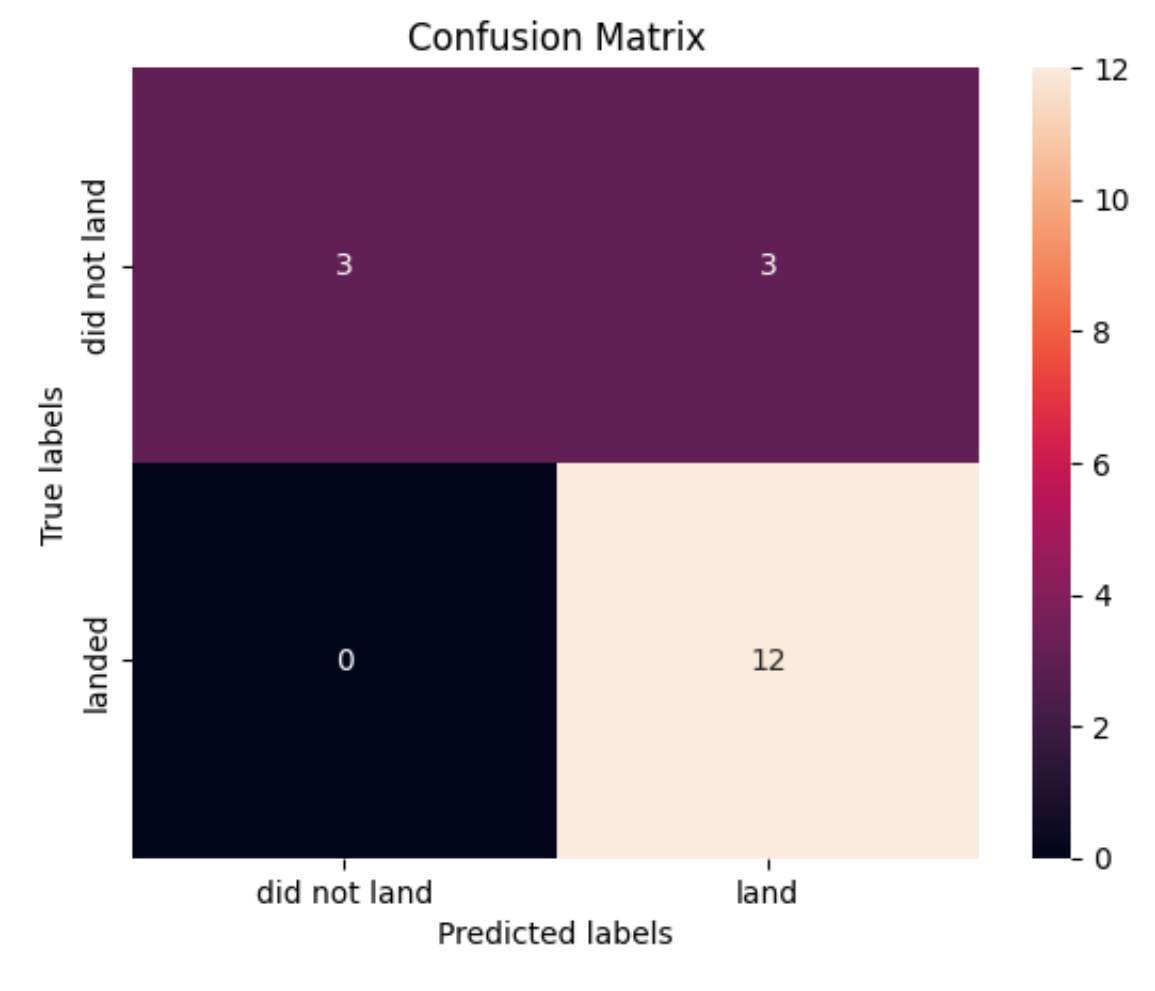
Predictive Analysis (Classification)

Classification Accuracy

- The bar chart at the right shows a comparison of the built model accuracy for the four predictive models tested.
- The decision tree classifier returned the highest accuracy val ue at 87.5%.



Confusion Matrix



- When used with the test data, all four models produced an equivalent confusion matrix , showing 3 false positives and 0 false negatives, with an accuracy measure of 83.33%.
- A larger test data set may help differentiate between the models.

Conclusions

- Spacex is making significant progress in its goal of creating a more cost-effective form of space travel, with the rate of successfully landed first stage rockets increasing steadily from the program's inception and reaching a level above 90% since 2019.
- This includes successful landings with increasingly heavier payloads, although payloads between 2,000-4,000 have been the most used and most successful to date.
- Successful landings have been made on 10 of 11 of the attempted orbits, although several of them have only been tried once. VLEO had the highest success rate of the orbits with more than a dozen attempts.
- Of the four predictive models, a decision tree classifier had the highest built model accuracy rate at 87.5%. Accuracy measured against the test data was equivalent (83.3%) for all models.
- It will be exciting to see results of new Spacex launches as additional data could bring about further insights and improved model training in future analyses.

Appendix

- Completed Jupyter Notebook showing API calls and results.
- CSV file containing the extracted data.
- Completed notebook showing webscraping process.
- CSV file with scraped data.
- Jupyter Notebook on Data Wrangling.
- CSV file with pre-processed and cleaned data.
- Jupyter Notebook – Exploratory Data Analysis with Data Visualization.
- CSV containing one-hot variables in preparation for predictive modeling.
- Jupyter Notebook containing Exploratory Data Analysis with SQL.
- Jupyter Notebook containing interactive Folium Map.
- Github Link to Python Code for Dashboard
- Provided Dataset Used for Dashboard
- Jupyter Notebook containing complete predictive analysis.
- Full Project Link - Github

Thank you!