IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ezra Hsieh

Jan. 6th, 2024

Outline

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

Executive Summary

- In this data science project, I aim to model and predict the success of SpaceX's rocket stage one landing.
- Datasets were obtained from SpaceX's API and related Wikipedia pages.
- Datasets were cleaned, then exploratory data analysis were peformed using pandas, matplotlib, and SQL.
- Interactive dashboards using Folium and Plotly Dash were created to assist visual analysis.
- Data analytical algorithms, including logistic regression, SVM, decision tree, and k-Nearest neighbor were created and optimized using machine learning techniques to model and predict stage one landing success.
- Certain attributes including the ordered number of launches, launching site, booster version, and payload mass seem to correlate with landing success.
- Predictive analysis algorithms with machine learning have an accuracy of 83.3% on test data.

Introduction

- SpaceX has achieved significant milestones, including sending spacecraft to the International Space Station, providing satellite internet access through Starlink, and sending manned missions to space.
- SpaceX's Falcon 9 rocket launches are relatively inexpensive compared to other providers, mainly because they can reuse the first stage.
- From the perspective of a hypothetical competitor, I seek, among other analysis, to predict the success of each launch based on whether SpaceX will reuse its first stage, using public information and machine learning.

Section 1

Methodology

Methodology

Summary

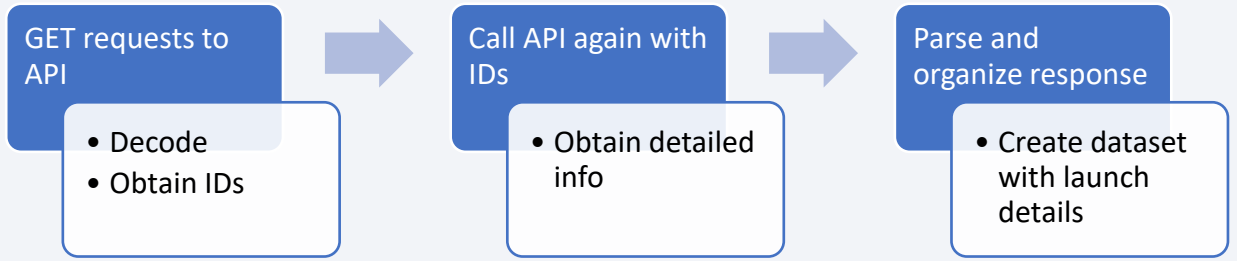
- Data collection methodology:
  - Rocket Launch data were collected from the SpaceX API and scraped from the Wikipedia page on SpaceX Falcon 9 launch page.
- Perform data wrangling
  - Only Falcon 9 launches were considered, missing values were filled with the average values, and irrelevant attributes were removed.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Built and tested using scikit-learn, with model parameters selected through Grid Search.

Data Collection

- First group of data were collected from SpaceX's publicly available API.
  - API requests
  - Obtain data including rocket booster, payload, launchpad, and cores, then organizing them into the columns of a pandas dataset.
- Second group of data were scraped from the Wikipedia page "List of Falcon 9 and Falcon Heavy launches"
  - Requests
  - Created BeautfulSoup object on HTML responses.
  - Extracted column and variable names.
  - Parsed through HTML tables and extracted necessary data.

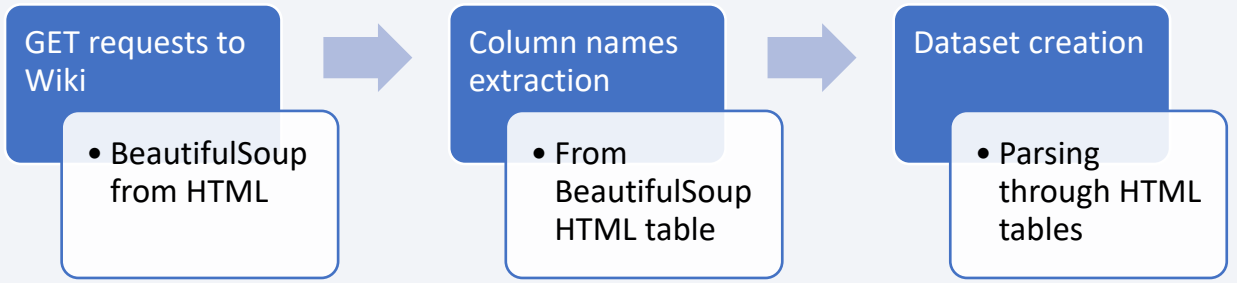
Data Collection – SpaceX API

- GET requests to on SpaceX API
- Decoded response as a Json then converting to Pandas Data Frame.
- Called the API again using IDs obtained for detailed information of each launch.
  - Rocket, payloads, launchpads, and cores.
- Parsed through responses and created another Data Frame
- https://github.com/EzraHsieh/SpaceX-Launches-Data-Science-Capstone/blob/8ba46d71bdcb0a3181026779876f17f0dbe3ad1e/jupyter-labs-spacex-data-collection-api.ipynb



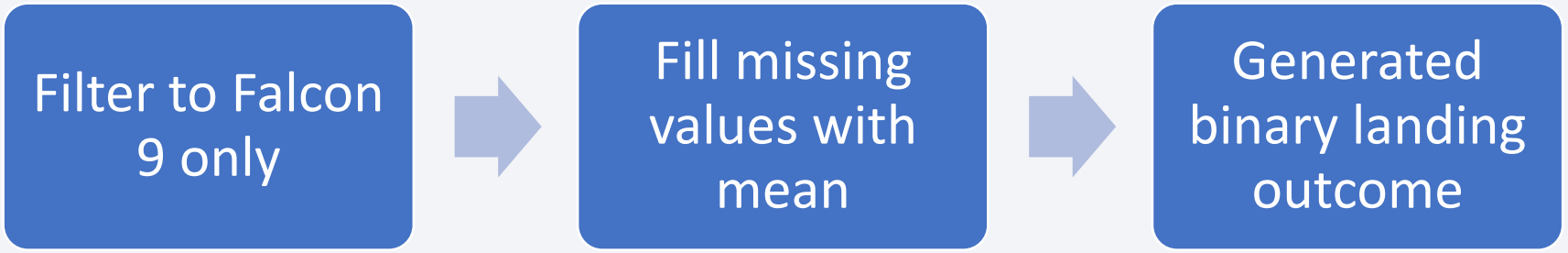
Data Collection - Scraping

- Requested Falcon 9 launch wiki page
  - Created BeautifulSoup object
- Extracted column names from HTML table header
- Created data frame by parsing HTML tables.
- https://github.com/EzraHsieh/SpaceX-Launches-Data-Science-Capstone/blob/8ba46d71bdcb0a3181026779876f17f0dbe3ad1e/jupyter-labs-webscraping.ipynb



Data Wrangling

- First, data frames were filtered to include only Falcon 9 launches.
- Second, managed null and missing values.
  - Replaced missing payload mass with the mean value.
- Then, created binary landing outcome 'Class' variable by combining different 'bad outcomes' whose second stage failed
- https://github.com/EzraHsieh/SpaceX-Launches-Data-Science-Capstone/blob/8ba46d71bdcb0a3181026779876f17f0dbe3ad1e/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- Scatterplot of flight number vs payload mass, coloring each point based on ‘class’ success result.
  - See if flight number correlates with success and payload.
- Scatterplot between flight number and launch site, colored by ‘class.’
  - Identify correlation between flight number, launch sites, and success.
- Scatterplot of payload and launch site, colored by class.
- Bar graph of the success rate of each orbit type to see relationship between orbit and success.
- Scatterplot of flight number and orbit type, colored by ‘class.’
- Scatterplot of payload and orbit types.
- Line plot that shows relationship between year and landing success rate
- https://github.com/EzraHsieh/SpaceX-Launches-Data-Science-Capstone/blob/8ba46d71bdcb0a3181026779876f17f0dbe3ad1e/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- SQL query on the dataset, to display:
  - Unique launch sites
  - Total payload masses launched by NASA (CRS).
  - Average payload of booster version F9 v1.1
  - Date of first successful landing outcome in ground pad
  - Total number of successful and failure outcomes
  - Booster versions that carried the max payload
  - Count of all different landing outcomes
- https://github.com/EzraHsieh/SpaceX-Launches-Data-Science-Capstone/blob/8ba46d71bdcb0a3181026779876f17f0dbe3ad1e/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

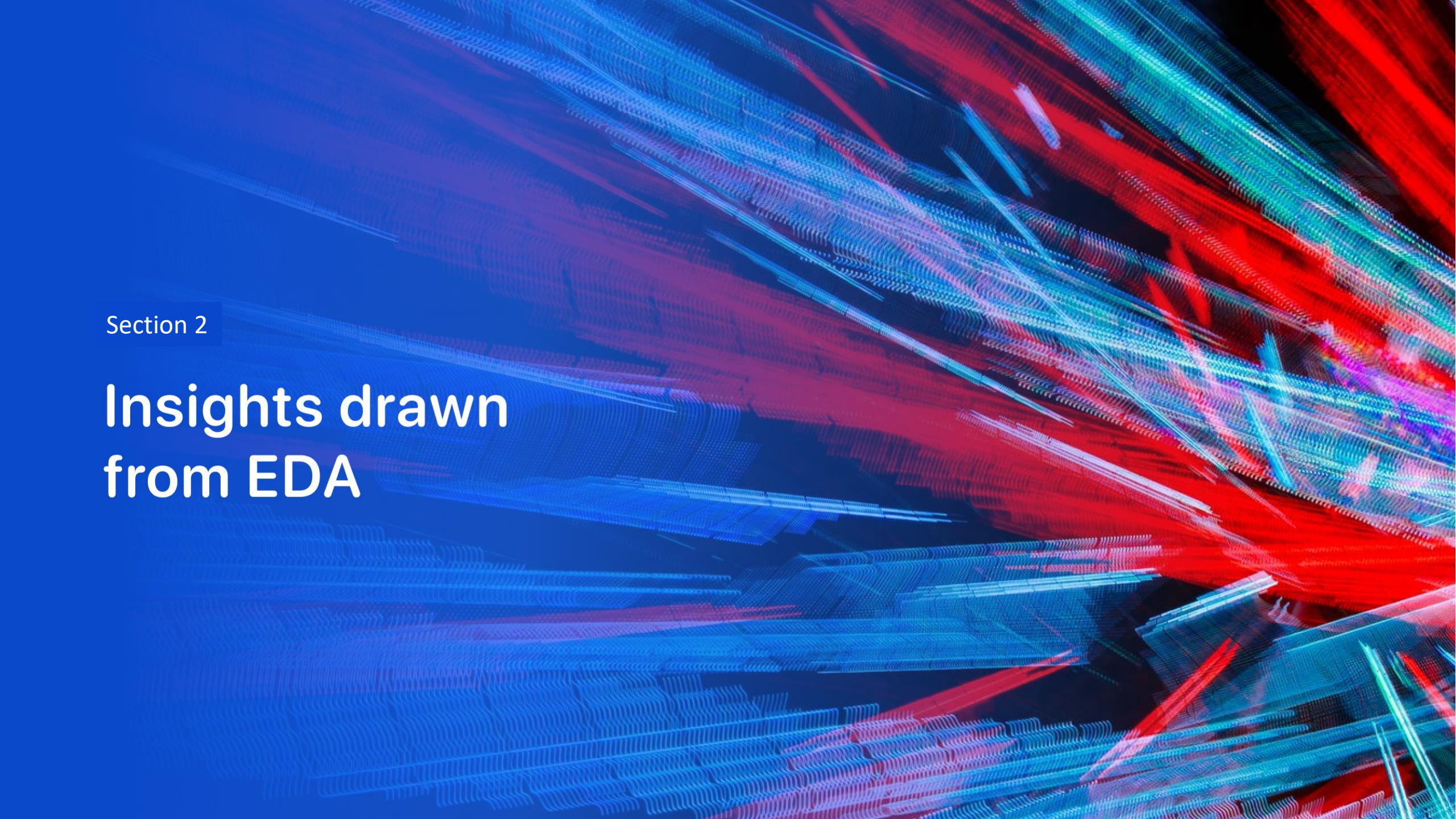
- Created interactive map with each launching site labeled as circle and marker.
- Each launch is labeled with colored maker as clusters on the location of their launch site.
  - Colored to show the success or failure outcome of each launch.
- Lines are drawn between each site and the nearest coastline, highway, railway, and city.
  - Distance was calculated and shown as markers to demonstrate the characteristics of each site's location.
- https://github.com/EzraHsieh/SpaceX-Launches-Data-Science-Capstone/blob/8ba46d71bdcb0a3181026779876f17f0dbe3ad1e/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Interactive Dashboard to assist visual analysis.
- Built pie chart that shows either successful launces count for each launch site.
  - Or success or failure of each site.
  - To visualize the how different launch site relates to landing outcome.
- Built interactive scatterplot between payload and launch success.
  - Also colored each point by booster version.
  - Can also customize with launch site option.
  - Assist visual analysis on payload, booster, launch site, and landing outcome.
- https://github.com/EzraHsieh/SpaceX-Launches-Data-Science-Capstone/blob/8ba46d71bdcb0a3181026779876f17f0dbe3ad1e/spacex_dash_app.py

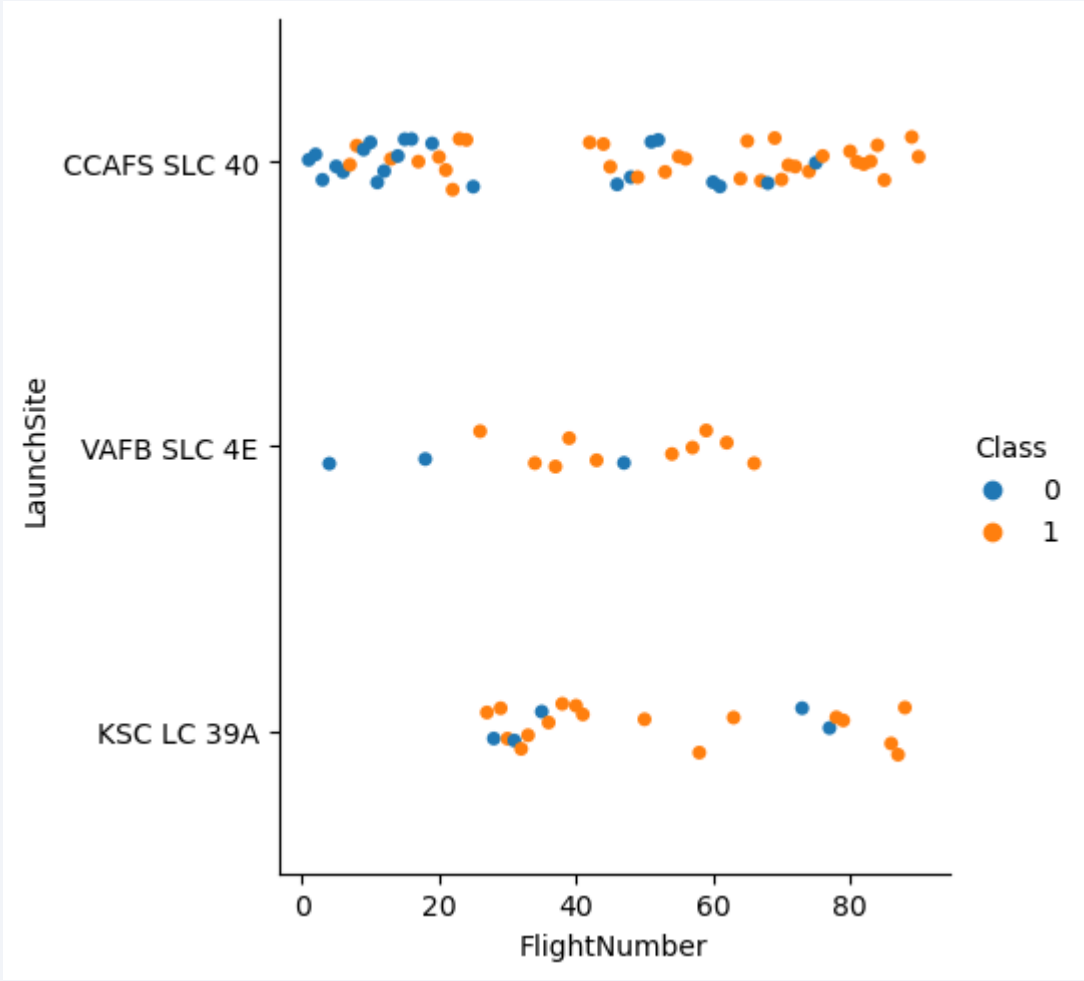
Predictive Analysis (Classification)

- Built, evaluated, and improved classification models with machine learning via scikit-learn.
  - Logistic regression, SVM, decision tree, and k-nearest neighbor.
- Organized datasets with predictor variables: flight number, date, booster version, payload mass, orbit type, launch site, flights, block, reused, landing pad, Grid fins, legs, and serial.
- Standardized all predictor variables.
- Divided data into 80% training and 20% test set.
- Identified best parameters for each model with Grid Search.
- Fitted and checked accuracy scores of each model on training data.
- Calculated the accuracy scores on test data, select models with best scores.
- https://github.com/EzraHsieh/SpaceX-Launches-Data-Science-Capstone/blob/8ba46d71bdcb0a3181026779876f17f0dbe3ad1e/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb


Section 2
Insights drawn from EDA

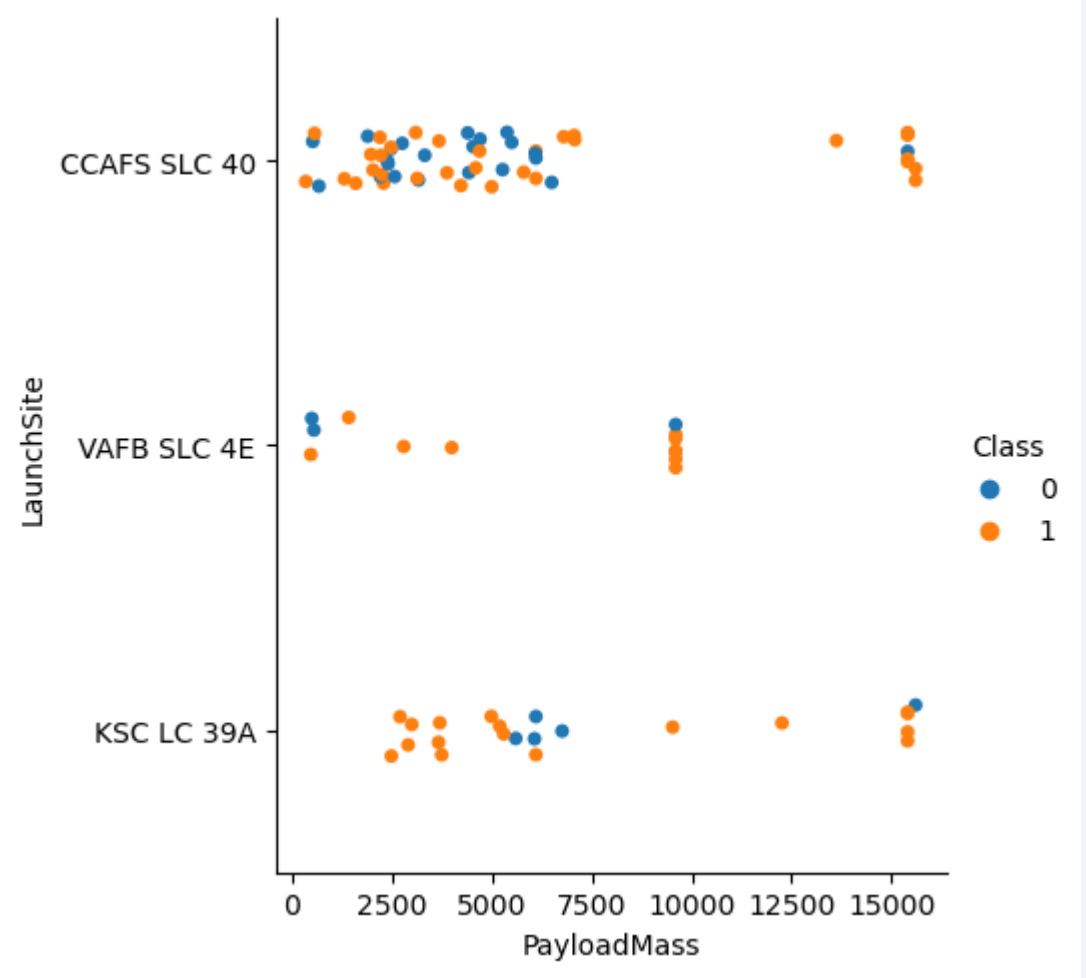
Flight Number vs. Launch Site

- CCAFS has the most rocket launches, while KSC's rocket launches are more recent.
- KSC has the highest rate of landing success.



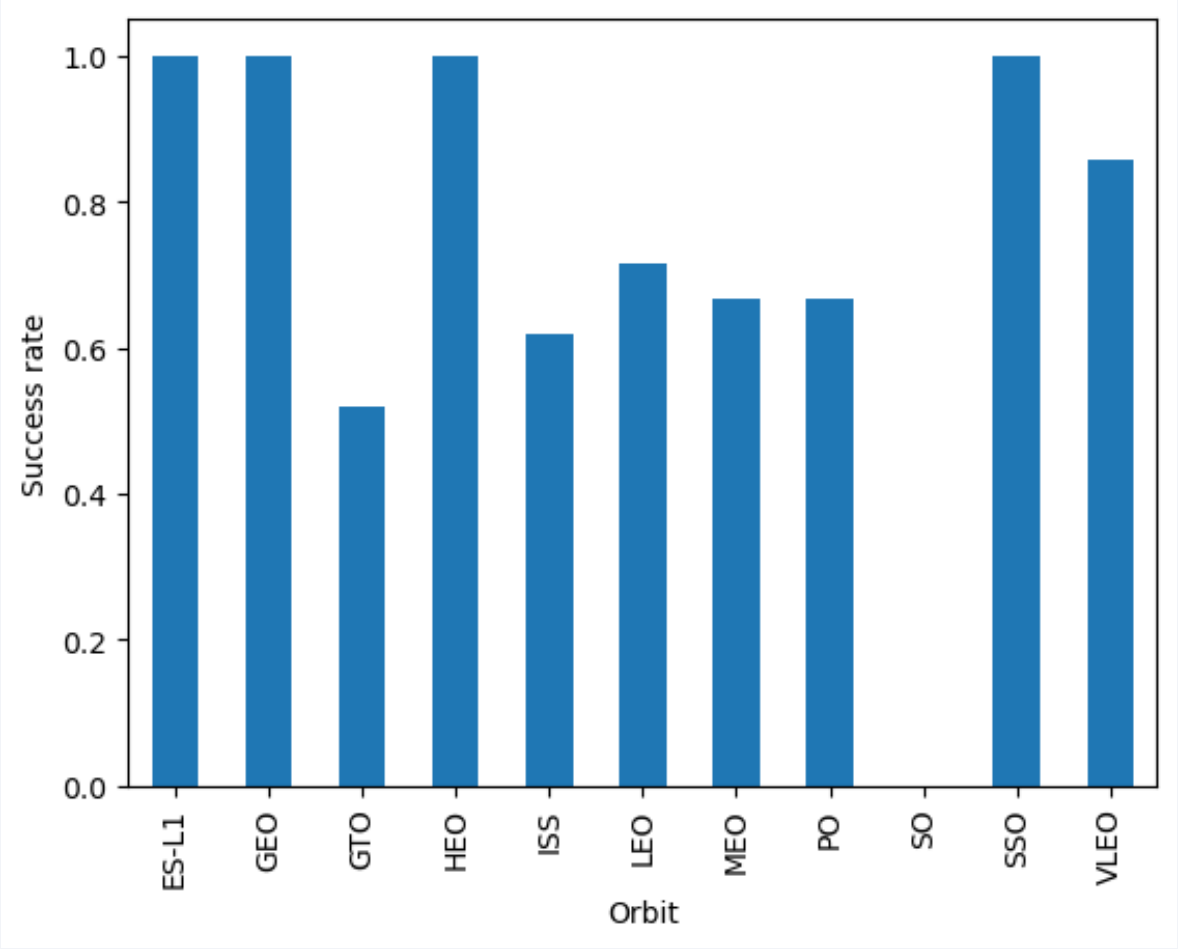
Payload vs. Launch Site

- Both CCAFS and KSC sites have heavy and light payloads.
- VAFB has no payload greater than 10000 kg.



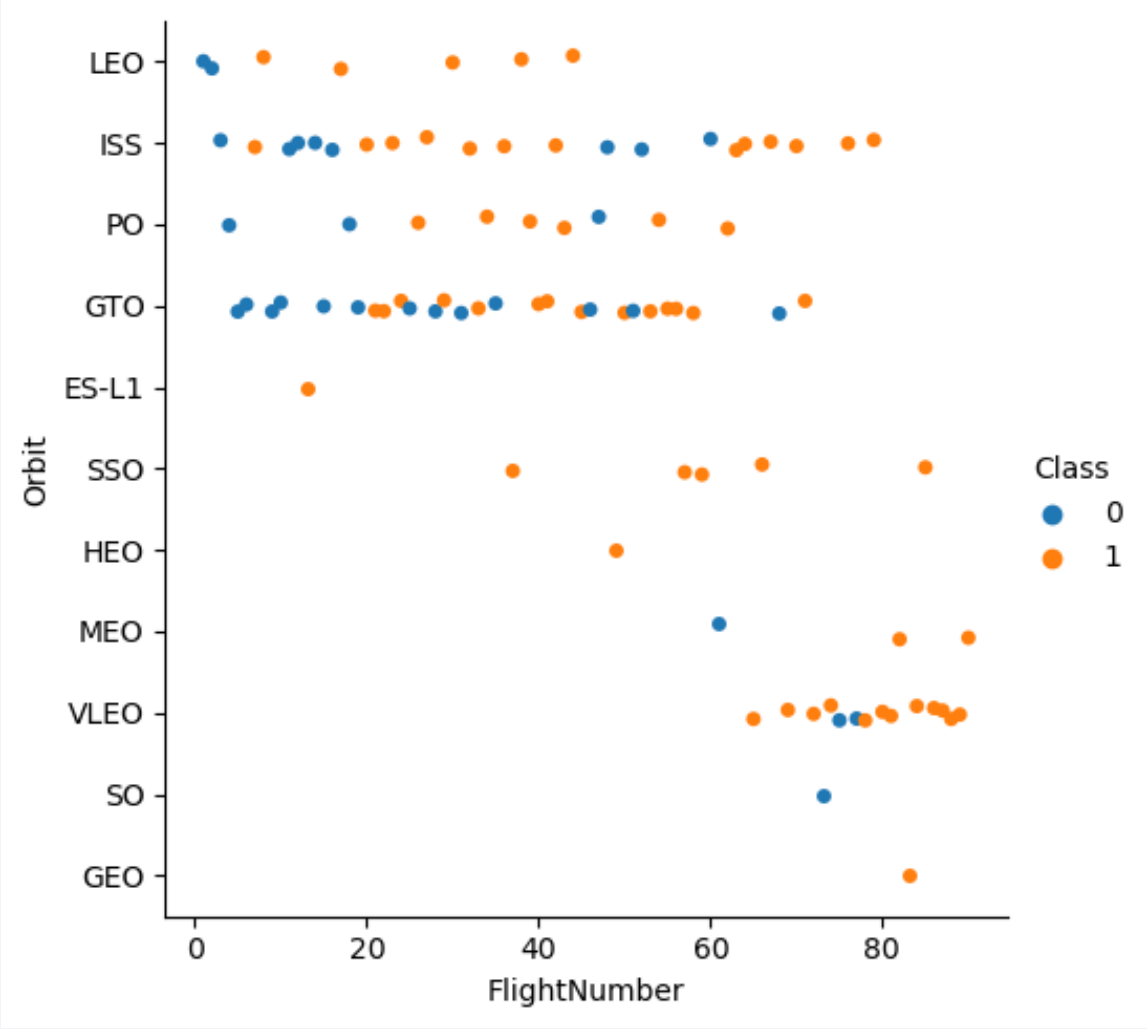
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO have perfect success rates.
- The rest orbit types have success rates ranging between 50 and 90%.
- Except for SO, which has a success rate of 0.



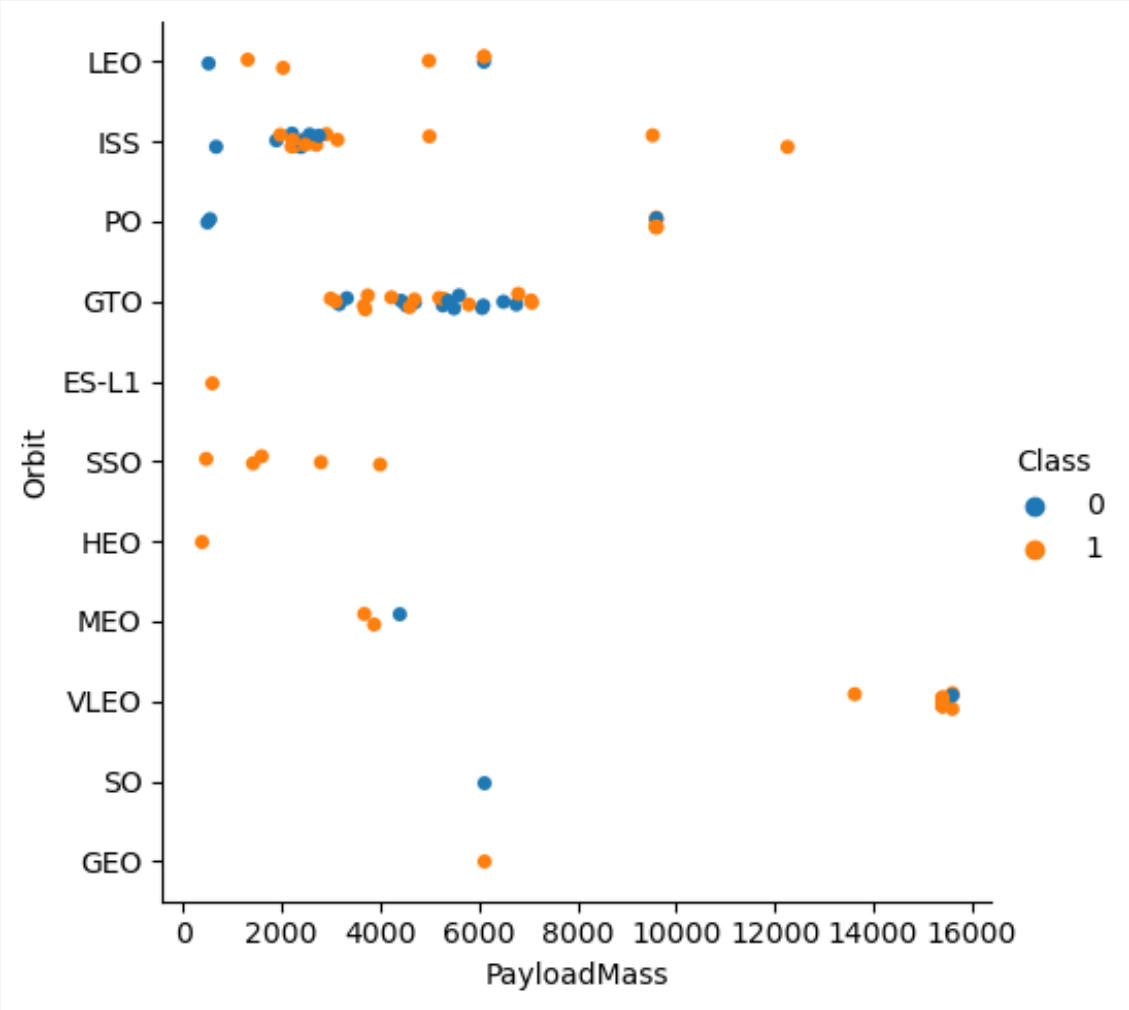
Flight Number vs. Orbit Type

- Success seems to increase as flight number increases for LEO orbit.
- No relationship between flight number and success for GTO orbit.



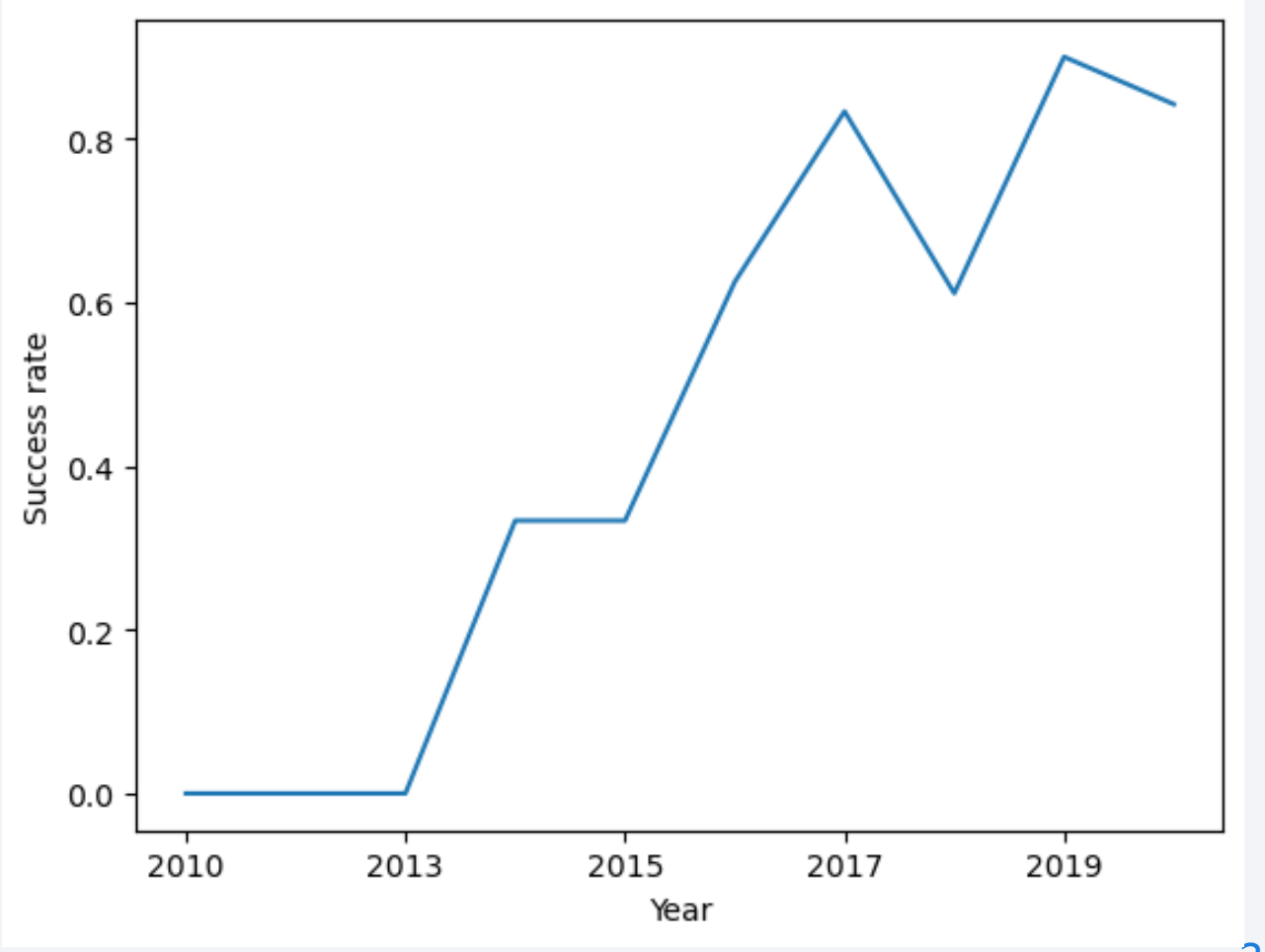
Payload vs. Orbit Type

- Heavier payloads seem to relate to greater success for LEO, PO, and ISS orbits.
- No relationship between payload and success for GTO.



Launch Success Yearly Trend

- Success rate have increased over the years overall.
- Some drop in 2017.



All Launch Site Names

- Find the names of the unique launch sites
- SQL query output:

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- SQL query result:

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- SQL query result:
- 45598 kg

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Query result:
- 2928.4 kg

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Query result:
- 2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- Query result:

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Query result of the mission outcomes and their count:

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Query result of all the booster versions with the highest payload:
  - All variations of B5

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

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Query results:

MONTH_NAME	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1017	VAFB SLC-4E
03	Failure (drone ship)	F9 FT B1020	CCAFS LC-40
06	Failure (drone ship)	F9 FT B1024	CCAFS LC-40

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

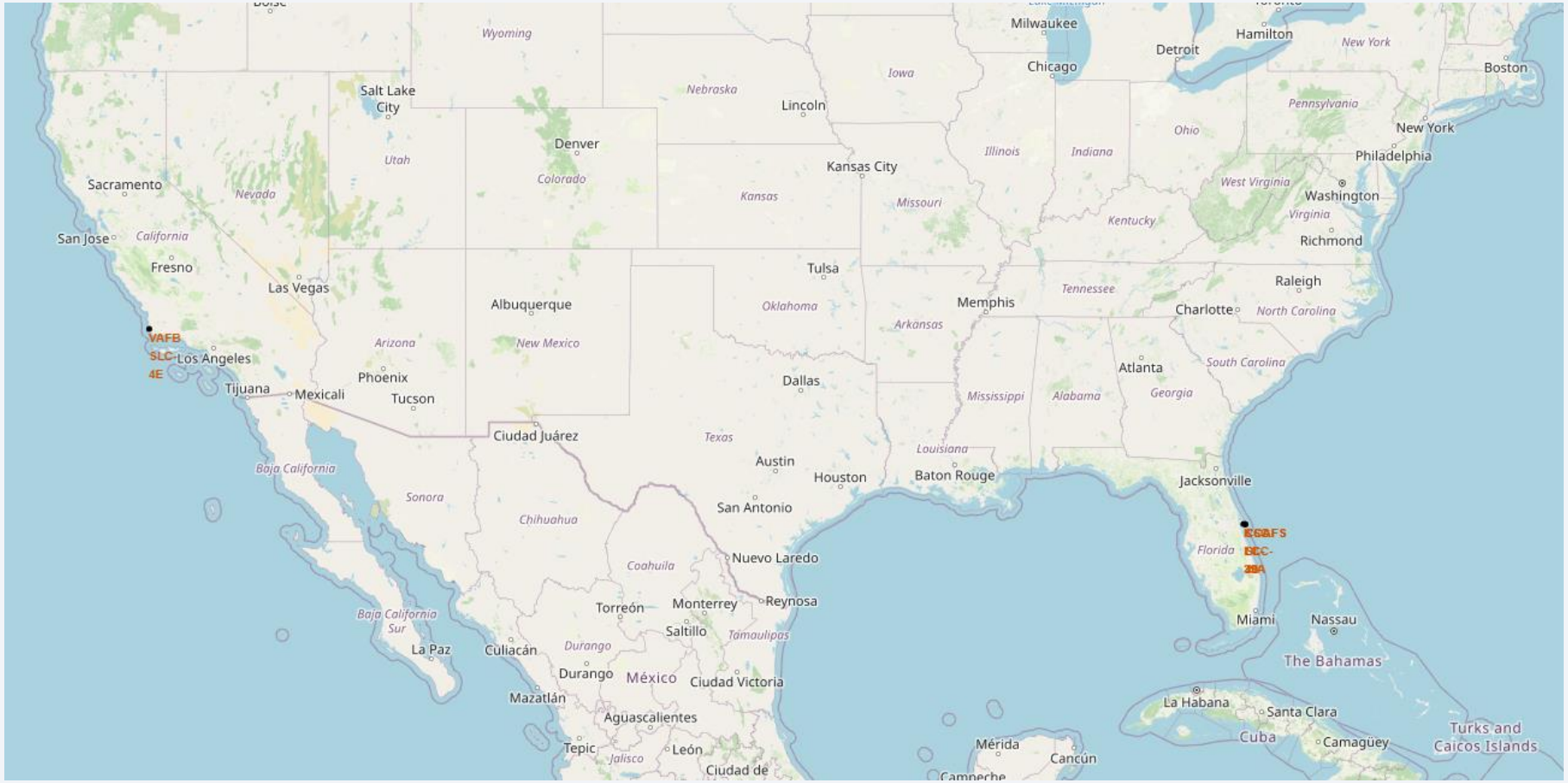
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order
- Query result of ranked count of each landing outcome:

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

Section 3

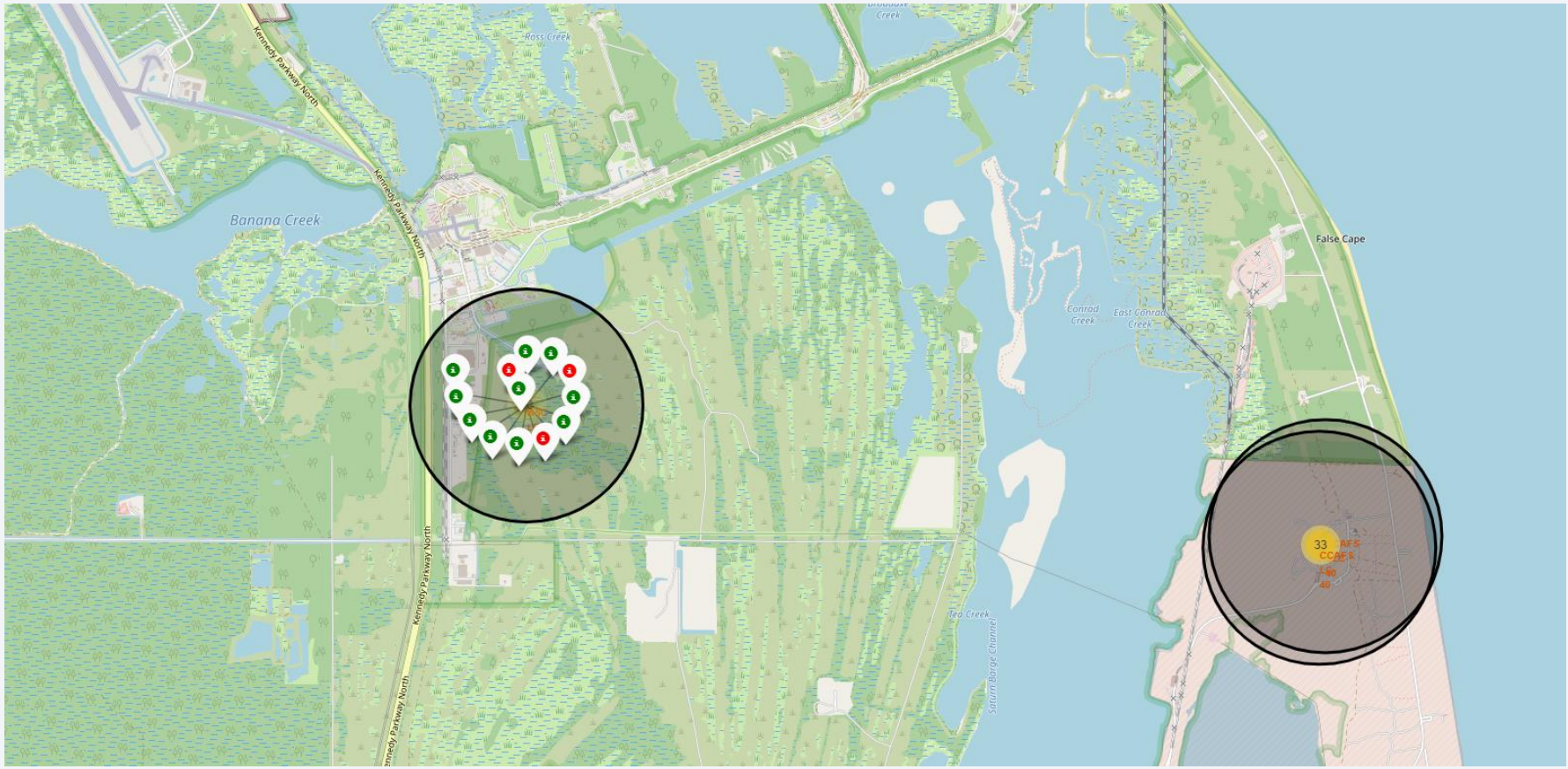
Launch Sites Proximities Analysis

Folium Map: Launch Sites



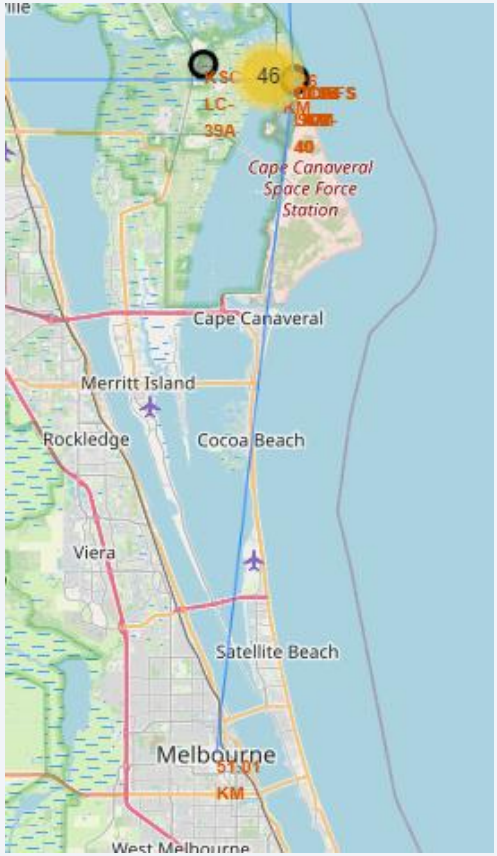
- All launch site are relatively close to the equatorial line and the coast.

Folium Map: Launch Outcomes

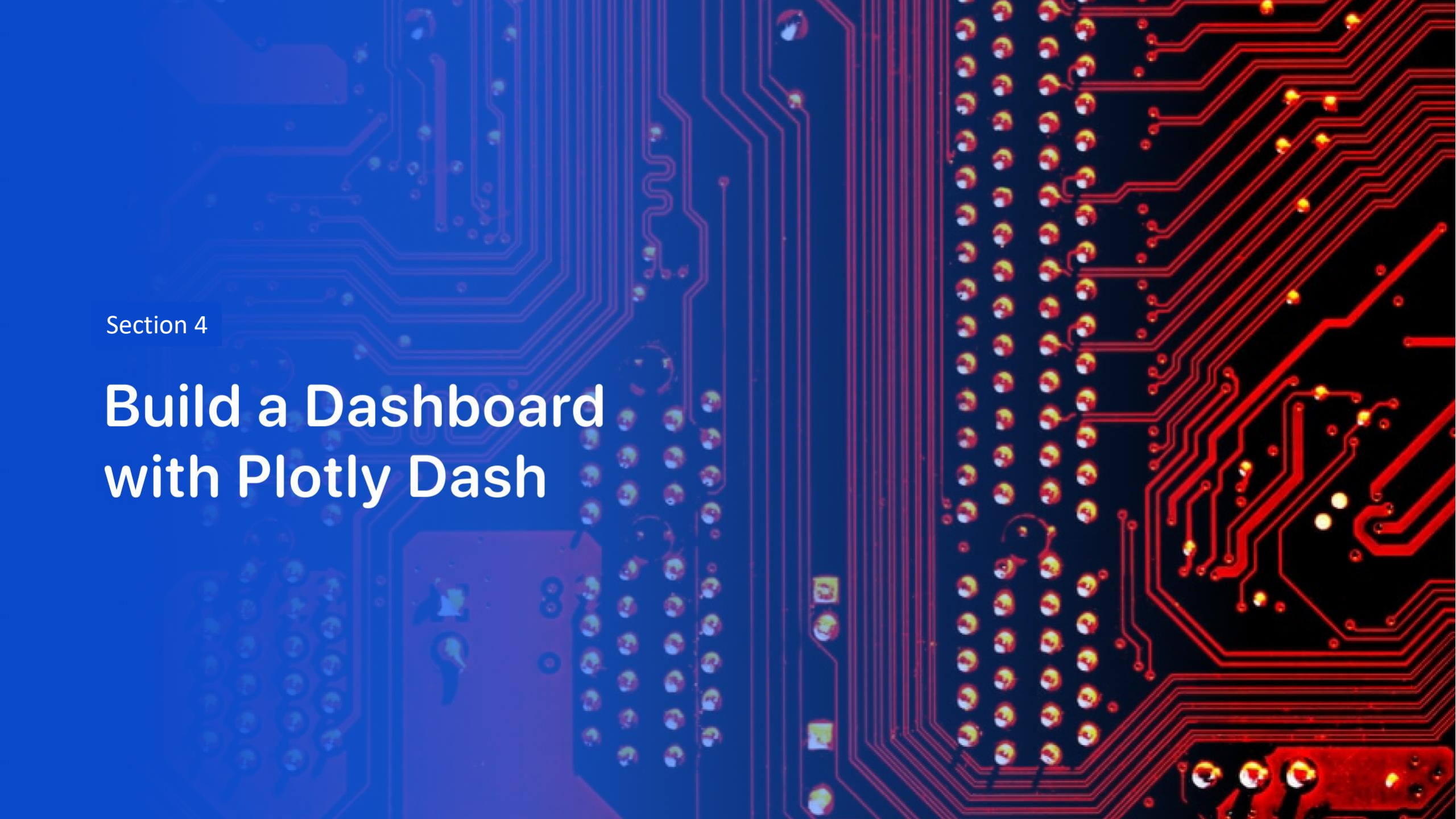


- KSC launch site has the highest success rate of launch outcomes.

Folium Map: Launch Site Proximities

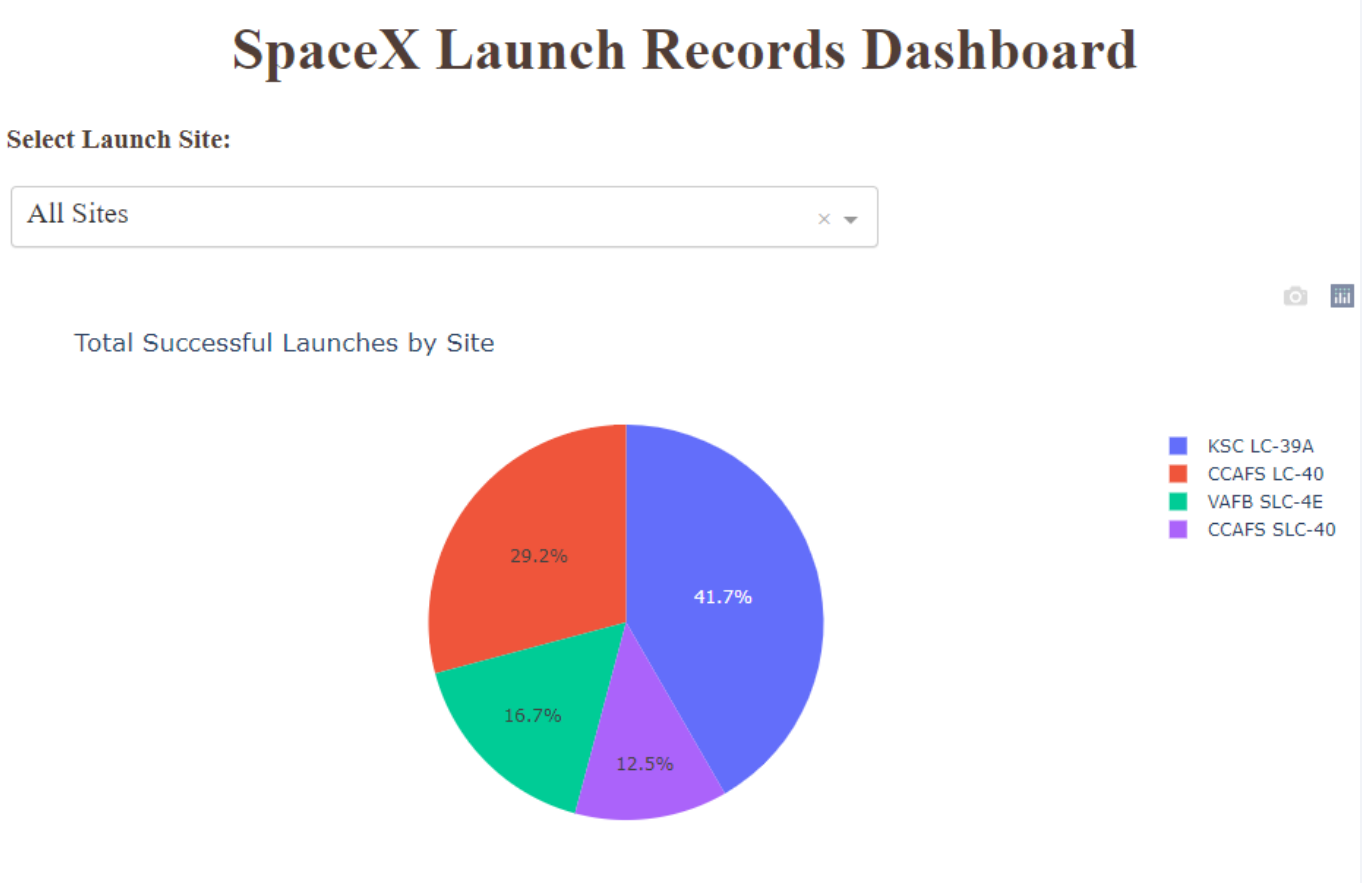


- All launch sites are within close proximities to coastline, railway, and highway while being relatively far away from major cities.


Section 4
Build a Dashboard with Plotly Dash

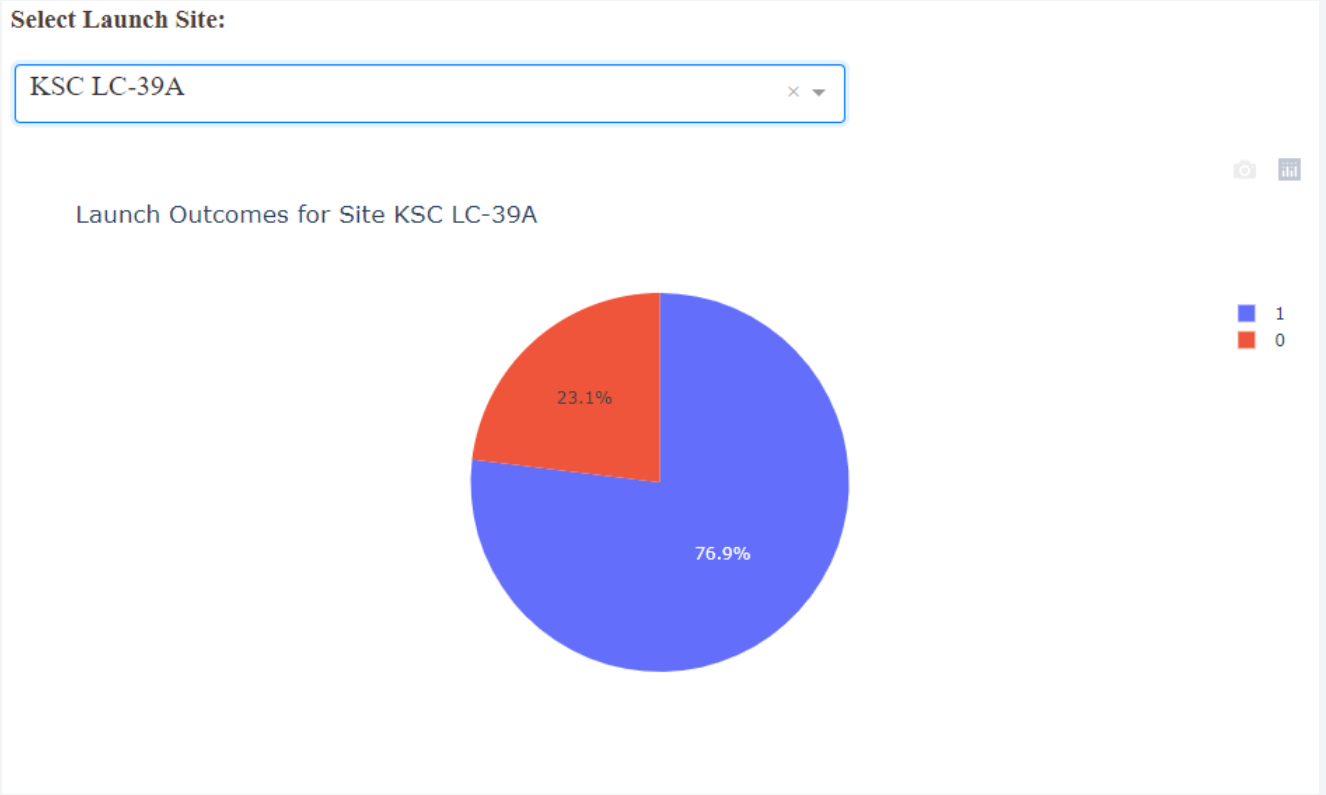
Dashboard: All Sites Pie Chart

- Total Successful Landing of Different Launch sites
- KSC has the highest number of successes.



Dashboard: KSC Pie Chart

- KSC has a high landing success rate too.
- 76.9%



Dashboard: Payload vs Launch Outcome Scatterplot

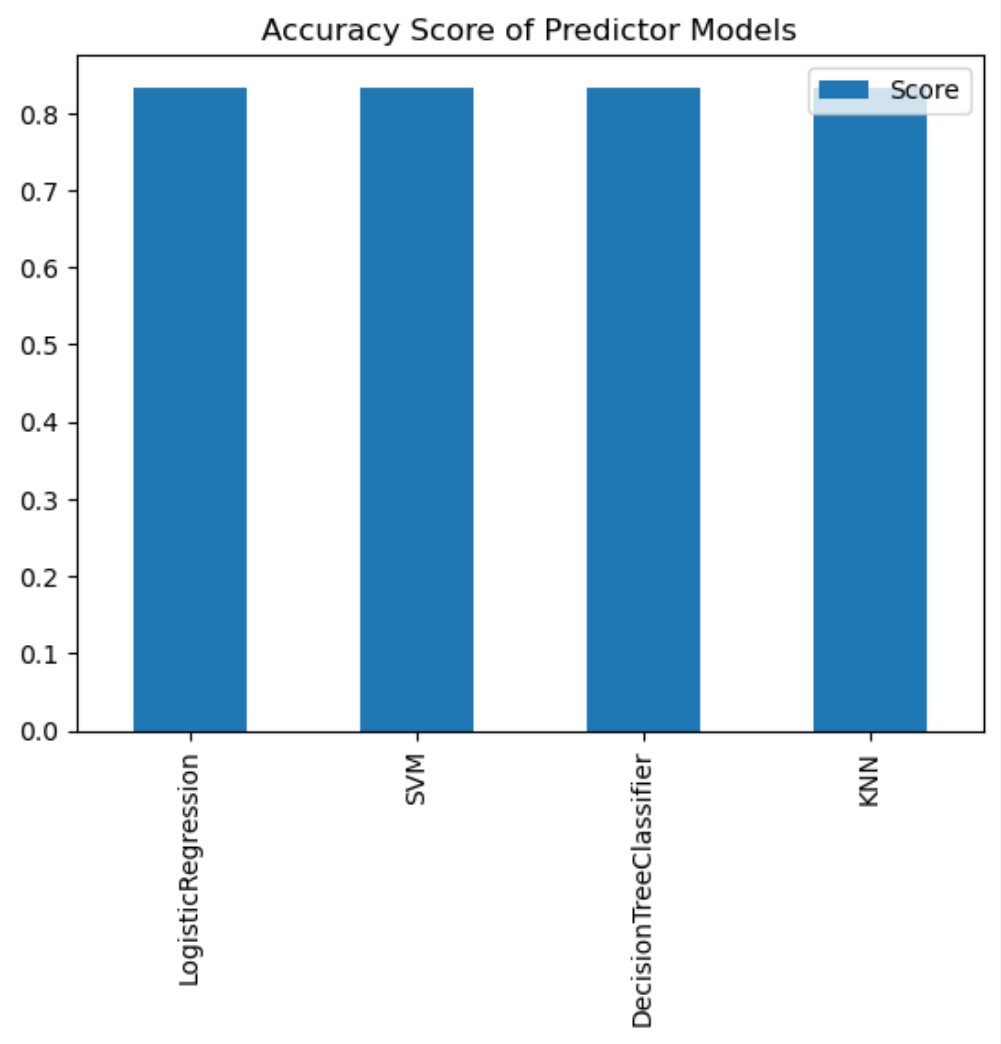
- Scatterplot shows payload mass (kg) vs launch outcome (1 for success), with booster version representing different colors.
- Boosters v1.0 and v1.1 have very low success rate.
- Payload mass between 2000 and 4000 kg have relatively high success rate.

Section 5

Predictive Analysis (Classification)

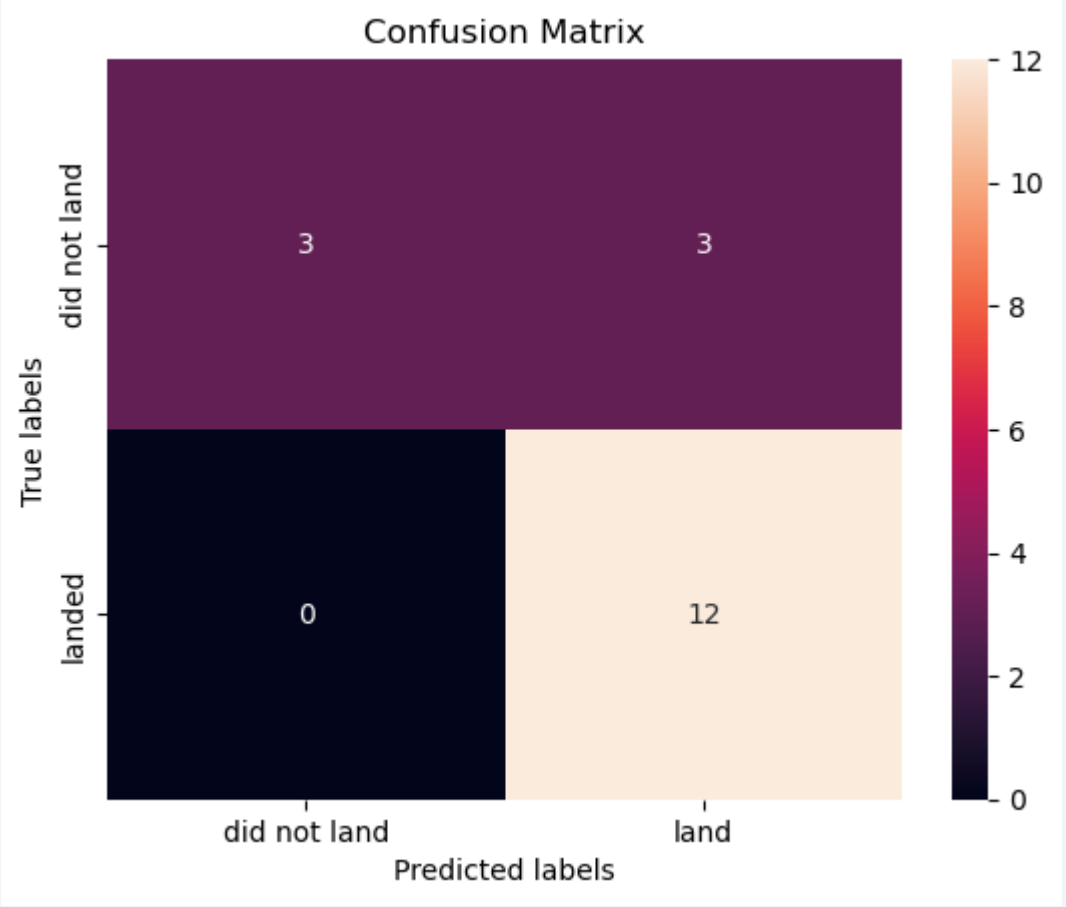
Classification Accuracy

- All classification models appear to have the same accuracy score.
- 83.3%



Confusion Matrix

- All models have the same confusion matrix.
- 100% precision for landing failure.
- 12/15 precision for landing success.
- 100% recall for landing success.
- 50% recall for landing failure.
- High over all F1 score: 88.9%
- Problem with false postives.



Results

- EDA
  - SpaceX success rates increase over the years.
- Interactive analytics
  - Launching sites, booster versions, payload masses all are predictors that have relationship with launching success.
- Predictive analysis
  - Logistic regression, SVM, decision tree, and k-nearest neighbor are all solid classification models predicting launching outcome.

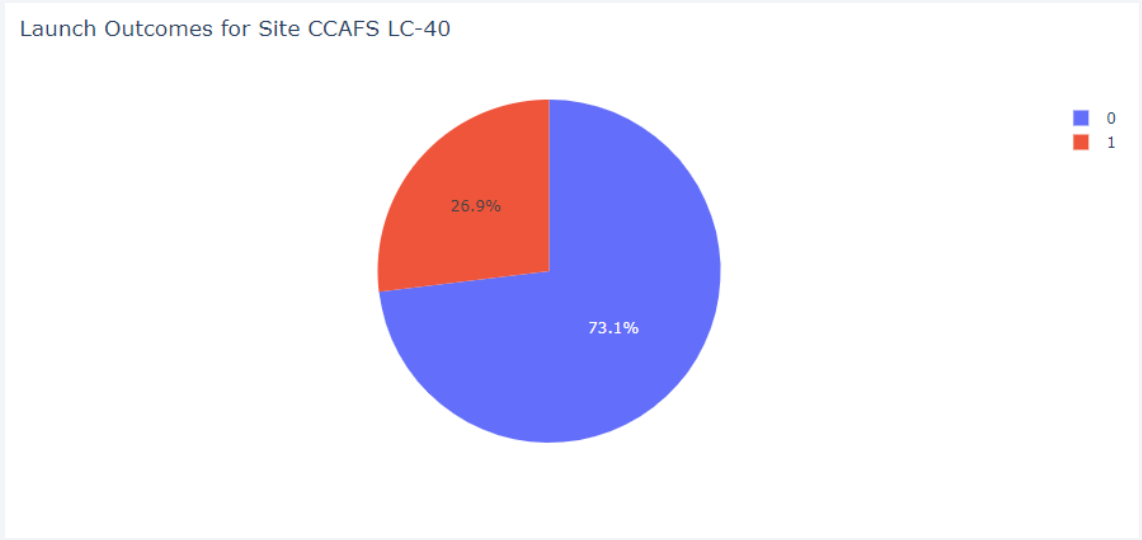
Conclusions

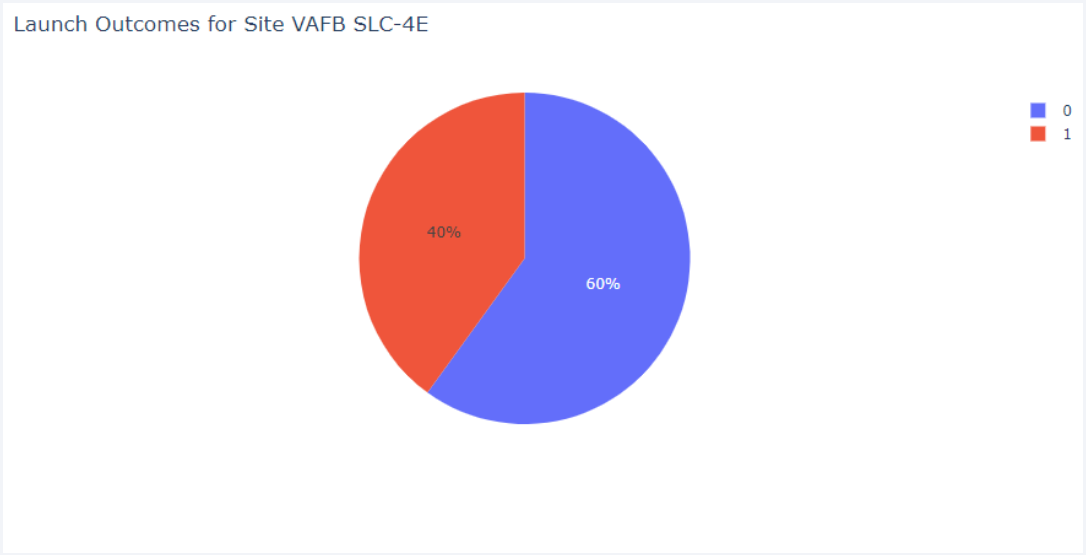
- Since SpaceX landing success increases over time, the cost of their launches will decrease as more stage one boosters can be reused.
  - Tough for competitors.
- Newer booster versions and sites also correlate with higher landing successes.
- More recent data after 2021 can be collected for training and testing of classifier models, in order to select the best model with lower false positive rate.

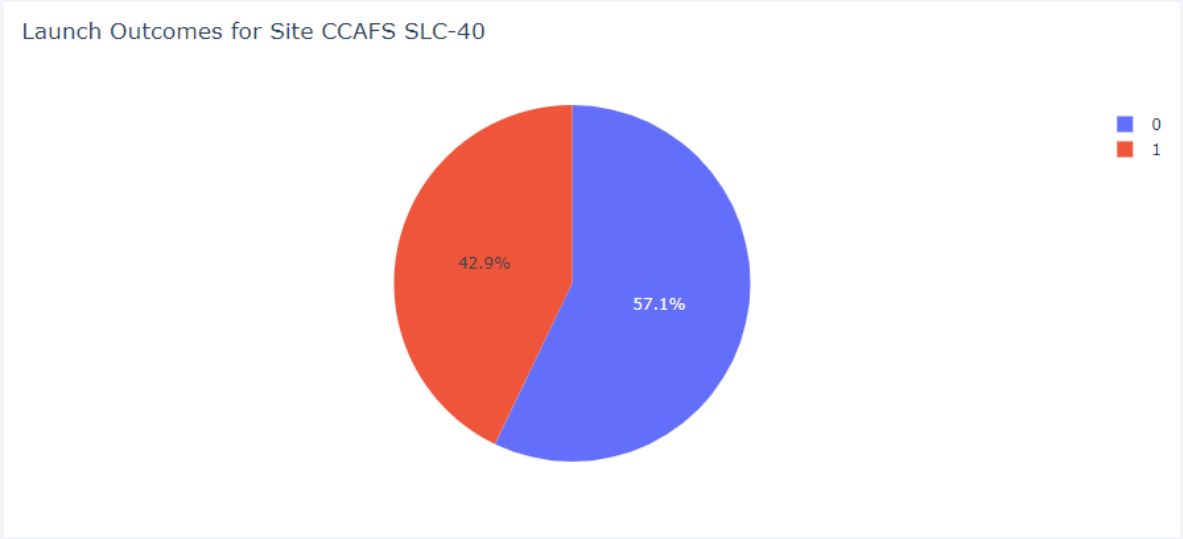
Acknowledgments

- Thanks to IBM Skills Network for the Data Science Professional Certificate and this capstone project.
- Special thanks to instructors Joseph Santarcangelo and Yan Luo
- And thanks to Rav Ahuja, Lakshmi Holla, and Azim Hirjani for creating the lab instructions, project guidelines, and skeleton codes for this project.

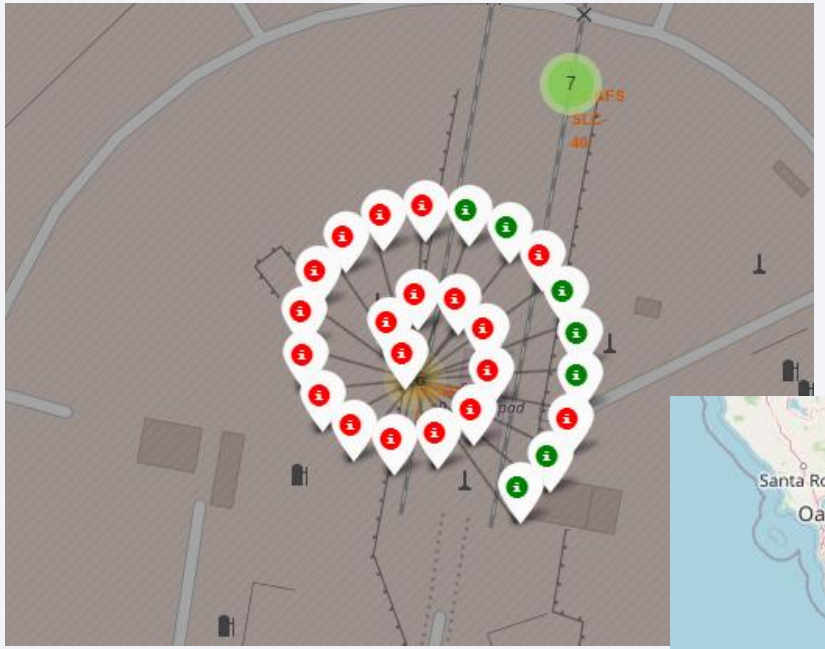
Appendix: Pie Charts for Other Launch Sites

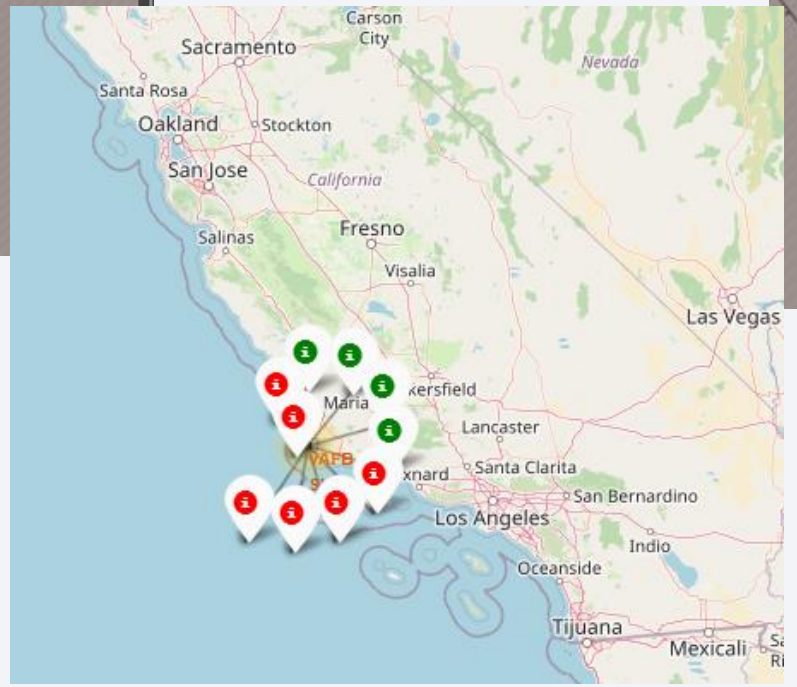






Appendix: Launch Outcomes at Other Sites





Thank you!