IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Andriy Savka
September 10, 2022

Outline

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

Executive Summary

- Data was collected from two sources: SpaceX Rest API and Web scraping Falcon 9 launch records from Wikipedia. Data wrangling consists of the following steps. Data from SpaceX Rest API was retrieved in JSON format and converted into Pandas data frame. Web scraping data from Wikipedia was processed and converted into Pandas data frame. Exploratory data analysis (EDA) using visualization and SQL revealed potential predictors. Interactive visual analytics using Folium and Plotly Dash further facilitated data exploration. Predictive analysis was performed using the following classification models: logistic regression, SVM, decision-tree classifier, and k-nearest-neighbors classifier.
- Data analysis revealed the following. Earlier launches had lower success rate than the later ones. Launces with higher payload mass had higher success rate. Launches for orbits ES-L, GEO, HEO, and SSO has outstanding success rate, while launches to the orbit SO were highly unsuccessful. Over time launches switched to different orbits. The highest percentage of successful launches happened on site KSC LC-39A. 76.9% launches on this site were successful. Overall success rate since 2013 kept increasing till 2020. Total payload mass of all launches is 45596 kg. Four ML models were built to predict outcomes a prospective lunch. K-nearest neighbor model has the highest predictive accuracy both for training and test sets (85% and 83%).

Introduction

- SpaceX is one of the most efficient companies that launches cargo, satellites, and maned missions into space. SpaceX was able to decrease the cost of a launch from average 165 to 62 million dollars by utilizing rockets Falcon 9 with reusable first stage.
- Cost of a launch depends on whether a rocket can successfully land and, thus, can be reused for future launches. The goal of this project is to determine the cost of launch by predicting if the first stage can be reused relying on public information available via various online sources.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology
  - Data was collected from two sources: SpaceX Rest API and Web scraping Falcon 9 launch records from Wikipedia
- Data wrangling
  - Data from SpaceX Rest API was retrieved in JSON format and converted into Pandas data frame. Web scraping data from Wikipedia was processed and converted into Pandas data frame.
- Exploratory data analysis (EDA) using visualization and SQL
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classification models
  - Dataset was standardized and then split between training and test sets in 0.8-0.2 proportion. Cross-validation grid search was performed to choose the most appropriate parameters for four ML models (logistic regression, SVM, decision-tree classifier, k-nearest-neighbors classifier). Model evaluation was based on R-squared parameter.

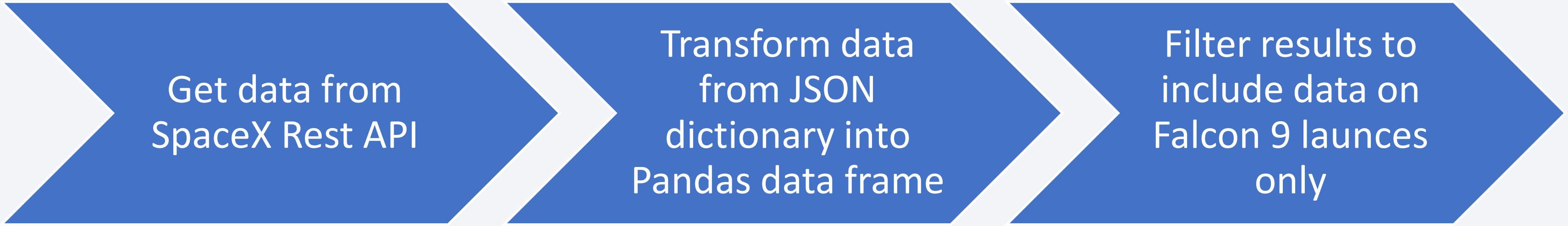
Data Collection

- Data was collected from two sources: SpaceX Rest API and Web scraping Falcon 9 launch records from Wikipedia
- In order to get data from SpaceX Rest API, we used ‘requests’ package
- Web scraping Falcon 9 launch records from Wikipedia was done utilizing ‘BeautifulSoap’ package

Data Collection – SpaceX API

- ‘requests’ package was used to get data from SpaceX Rest API
- Data was transformed from JSON dictionary into Pandas data frame
- Data frame was filtered to include Falcon 9 launches only

- GitHub URL of the completed SpaceX API calls notebook:
  - https://github.com/andriy-savka/IBM-Applied-Data-Science-Capstone/blob/953d95ec850e538f7c763bddb8e046a0a7d1d9bd/jupyter-labs-spacex-data-collection-api.ipynb



Get data from SpaceX Rest API

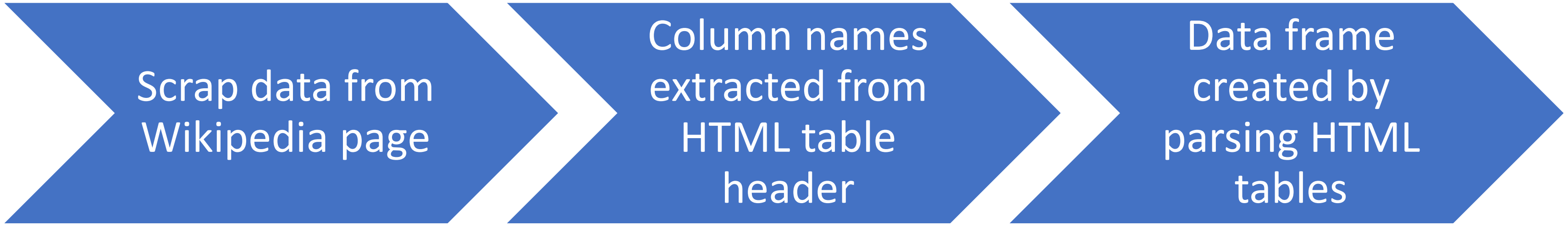
Transform data from JSON dictionary into Pandas data frame

Filter results to include data on Falcon 9 launces only

Data Collection - Scraping

- ‘requests’ and ‘BeautifulSoap’ packages were used to get data from Wikipedia page “List of Falcon 9 and Falcon Heavy launches”
- Column names were extracted from the HTML table header
- A data frame was created by parsing the launch HTML tables

- GitHub URL of the completed web scrapping from Wikipedia page notebook:
  - https://github.com/andriy-savka/IBM-Applied-Data-Science-Capstone/blob/953d95ec850e538f7c763bddb8e046a0a7d1d9bd/jupyter-labs-webscraping.ipynb



Scrap data from Wikipedia page

Column names extracted from HTML table header

Data frame created by parsing HTML tables

Data Wrangling

- Missing values in the variable ‘PayloadMass’ were replaced by the mean of ‘PayloadMass’
- Missing values in the variable ‘LandingPad’ will become a separate category
- A landing outcome label from ‘Outcome’ column was created

- GitHub URL of completed data wrangling related notebooks:
    - https://github.com/andriy-savka/IBM-Applied-Data-Science-Capstone/blob/953d95ec850e538f7c763bddb8e046a0a7d1d9bd/jupyter-labs-spacex-data-collection-api.ipynb
    - https://github.com/andriy-savka/IBM-Applied-Data-Science-Capstone/blob/953d95ec850e538f7c763bddb8e046a0a7d1d9bd/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
- In order to understand how various predictors correspond with launch outcomes (target variable) and between each other, we plotted a series of scatter plots: Flight Number against Payload Mass, Flight Number and Launch Site, Payload and Launch Site, Flight Number and Orbit Type, Payload and Orbit Type
- We also visualized the launch success trend to understand dynamic of target variable
- Success rate of each orbit type was plotted to learn about differences correlation between launch outcomes and orbit type

- GitHub URL of completed EDA with data visualization notebook:
  - https://github.com/andriy-savka/IBM-Applied-Data-Science-Capstone/blob/953d95ec850e538f7c763bddb8e046a0a7d1d9bd/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- The following aspects of the daset were studied via performing SQL queries as a part of EDA phase:
  - The names of the unique launch sites in the space mission
  - 5 records where launch sites begin with the string 'CCA'
  - The total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - A list the date when the first successful landing outcome in ground pad was achieved.
  - A list the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - A list the total number of successful and failure mission outcomes
  - A list the names of the booster versions which have carried the maximum payload mass
  - A list the records which displayed the month names, failure landing outcomes in drone ship, booster versions, launch site for the months in year 2015
  - A rank the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order

- GitHub URL of completed EDA with SQL notebook:
  - https://github.com/andriy-savka/IBM-Applied-Data-Science-Capstone/blob/953d95ec850e538f7c763bddb8e046a0a7d1d9bd/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- The following objects were added to the Folium map:
  - Using ‘folium.Circle’ and ‘folium.Marker’, we added circles to the map for each launch site in data frame
  - Via ‘market_claster’, we visualized successful and failed launches for each launch site
  - Utilizing ‘folium.PolyLine’, we visualized the shortest path from a launch side to the closest coastline

- GitHub URL of completed interactive map with Folium map:
  - https://github.com/andriy-savka/IBM-Applied-Data-Science-Capstone/blob/953d95ec850e538f7c763bddb8e046a0a7d1d9bd/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- We build an interactive Plotly Dashboard with the following graphs and interaction elements:
    - Dropdown list to enable Launch Site selection and a slider to select payload range
    - A pie chart to show the total successful launches count for all sites and a scatter chart to show the correlation between payload and launch success

- GitHub URL of completed Plotly Dash dashboards:
    - https://github.com/andriy-savka/IBM-Applied-Data-Science-Capstone/blob/953d95ec850e538f7c763bddb8e046a0a7d1d9bd/spacex_dash_app.py

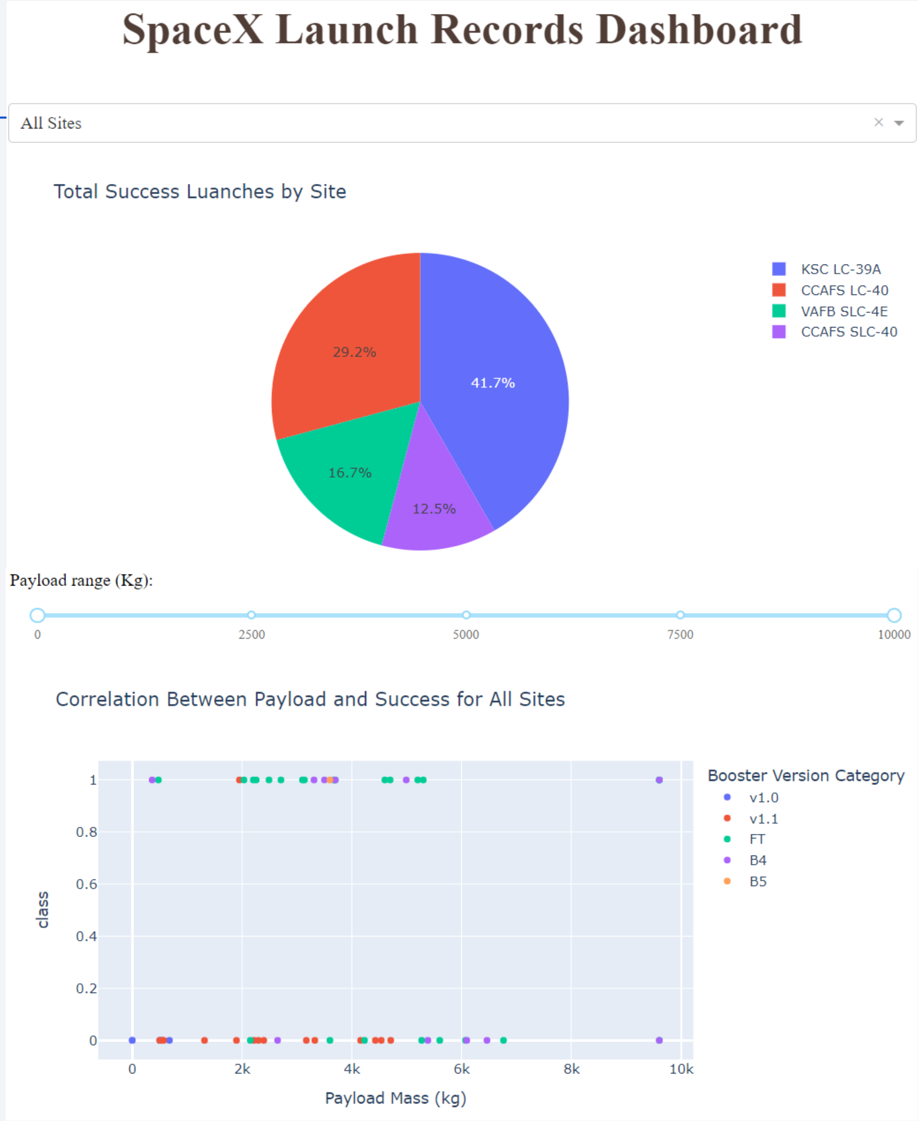
Predictive Analysis (Classification)

- Dataset was standardized and then split between training and test sets in 0.8-0.2 proportion
- For each model, cross-validated grid search was performed to choose the most appropriate parameters
- The following models were built:
  - Logistic regression
  - SVM
  - Decision-tree classifier
  - K-nearest-neighbors classifier
- Model evaluation parameter is R-squared (default parameter for Scikit-learn method '.score()')

- GitHub URL of completed predictive analysis:
  - https://github.com/andriy-savka/IBM-Applied-Data-Science-Capstone/blob/953d95ec850e538f7c763bddb8e046a0a7d1d9bd/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

- Exploratory data analysis, including interactive dashboard allowed to discover variables that could be candidates for strong predictors.
- Predictive analysis results revealed that 3 models have similar performance results on test set: logistic regression, SVM, and k-nearest-neighbors classifier, while k-nearest-neighbors classifier has the best performance on train set.



SpaceX Launch Records Dashboard

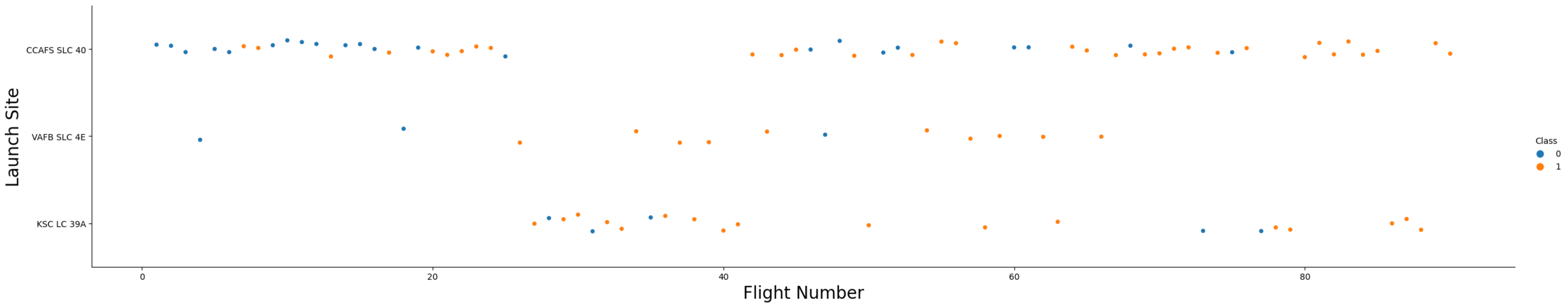
All Sites

Payload range (Kg):

Section 2

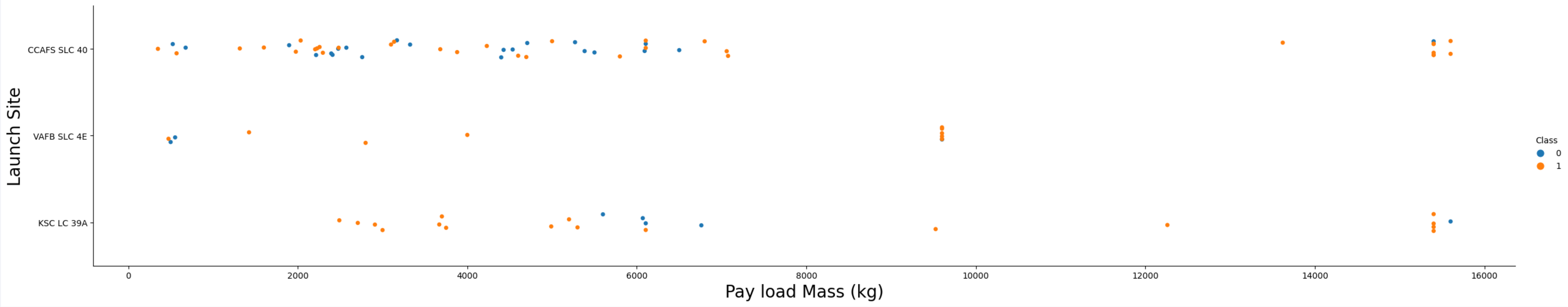
Insights drawn from EDA

Flight Number vs. Launch Site



- Site CCAFS SLC 40 is the most used site with mixed results. Site VAFB SLC 4E has been used the least. The scatter plot reveals that earlier launches had lower success rate than the later ones.

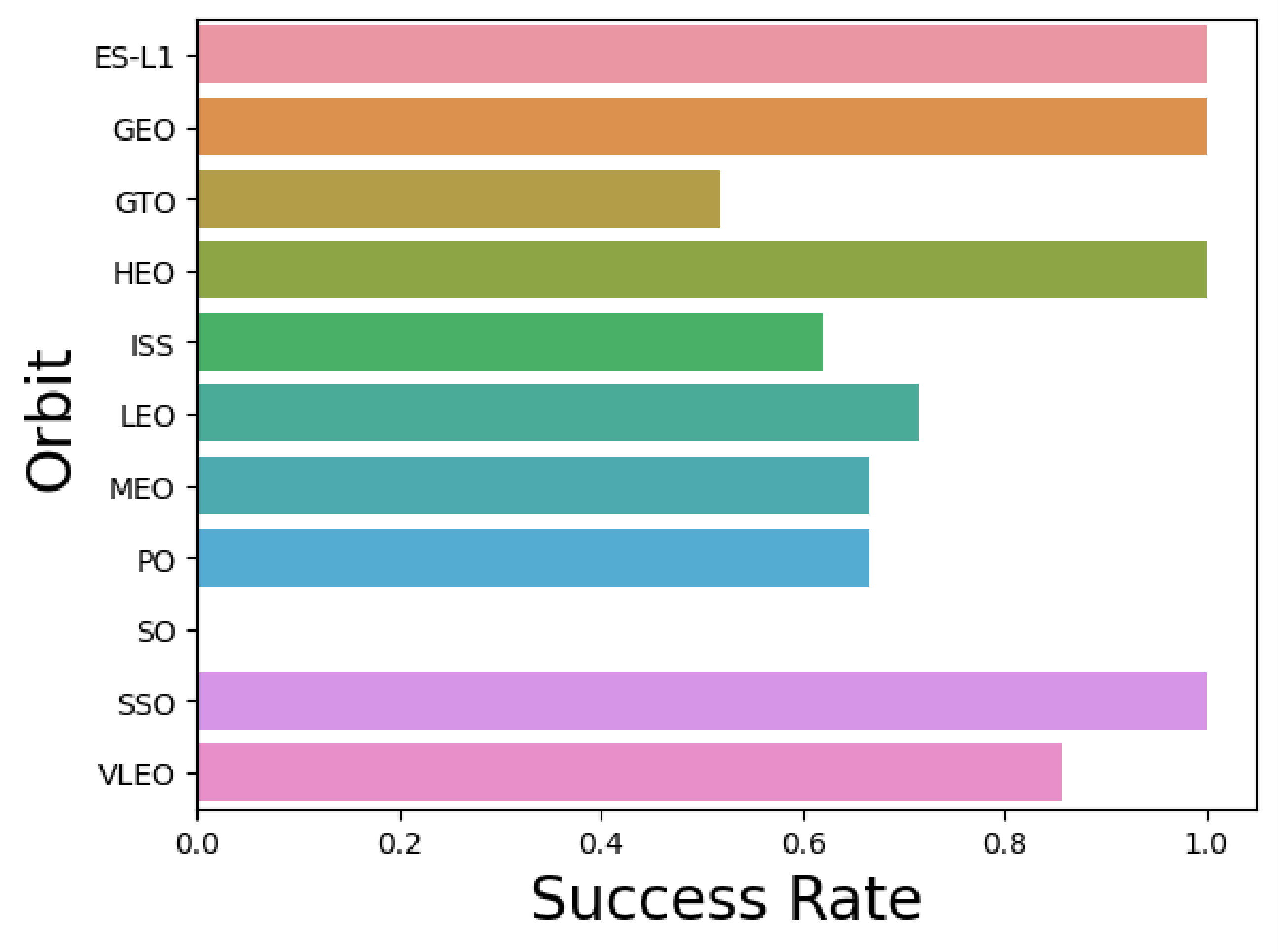
Payload vs. Launch Site



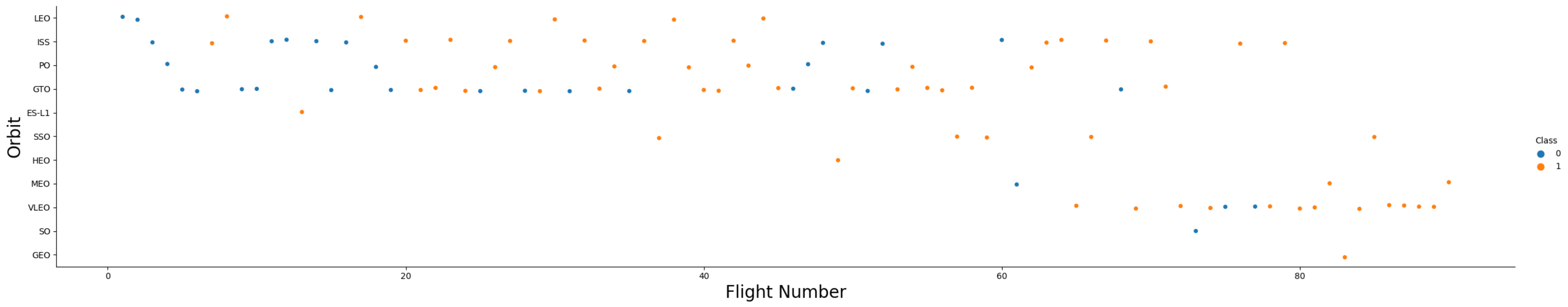
- The graph demonstrates higher success rate for launces with higher payload mass.

Success Rate vs. Orbit Type

- The graph demonstrates outstanding success rate for orbits ES-L, GEO, HEO, and SSO, while launches to the orbit SO were highly unsuccessful.

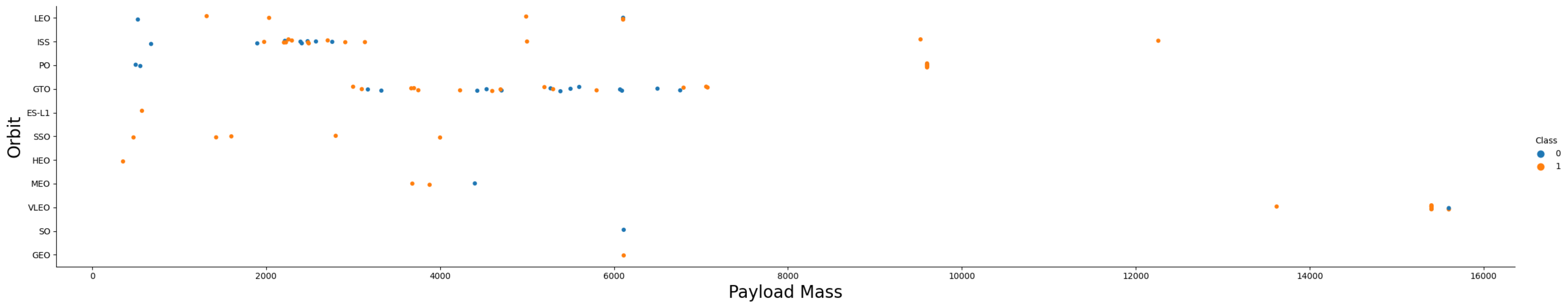


Flight Number vs. Orbit Type



- The graph demonstrates that the launches switched to different orbits.

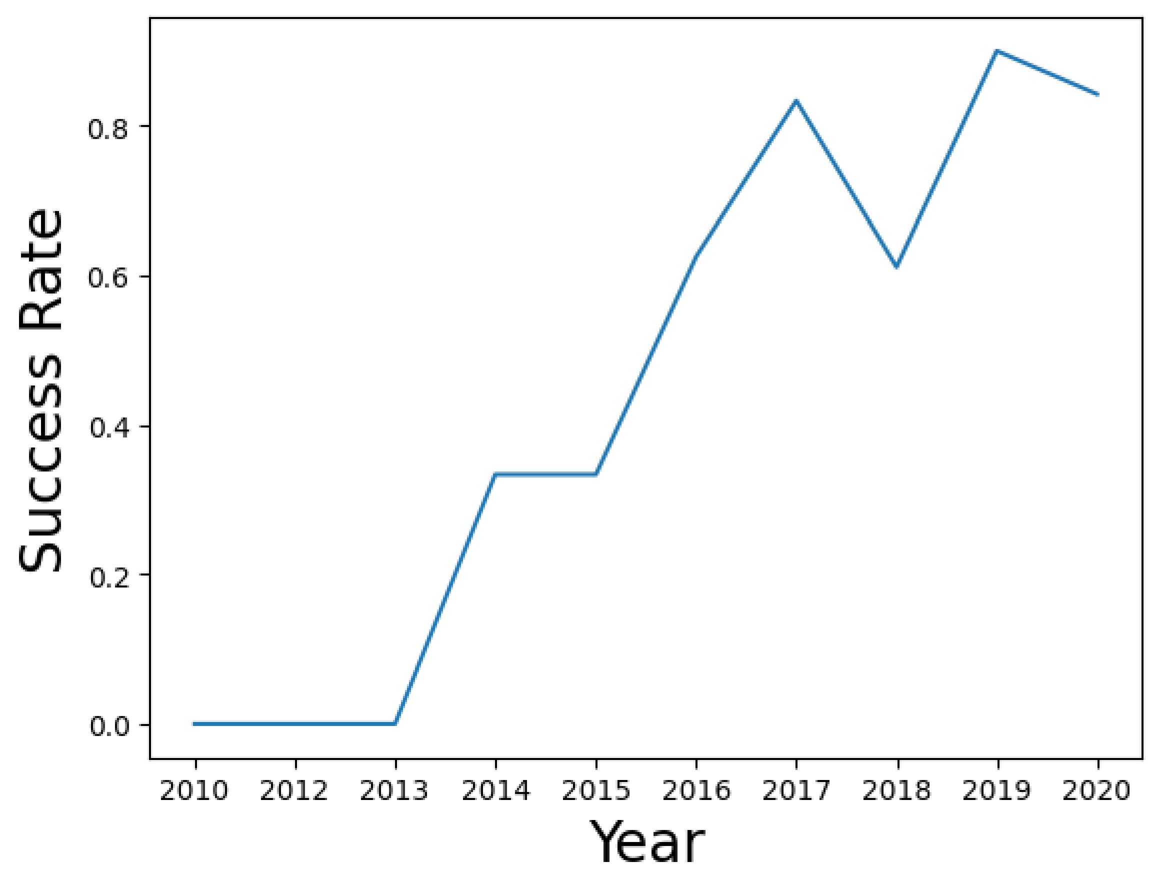
Payload vs. Orbit Type



- The graph demonstrates that with heavy payloads the successful landing or positive landing rate are more for PO, LEO and ISS.

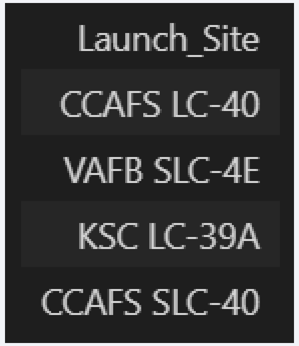
Launch Success Yearly Trend

- The graph demonstrates that the success rate since 2013 kept increasing till 2020.



All Launch Site Names

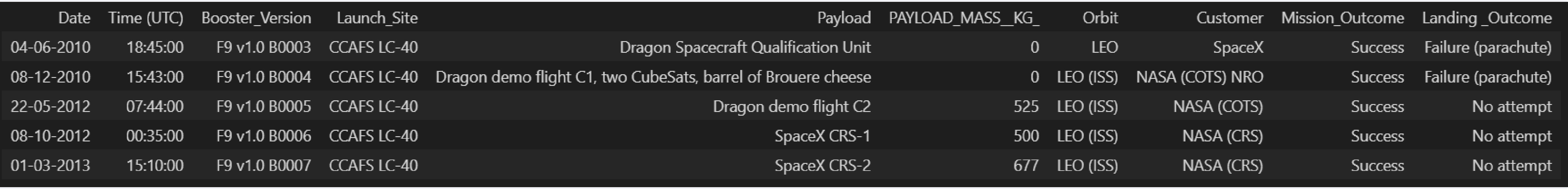
- The names of the unique launch sites:



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

Launch Site Names Begin with 'CCA'

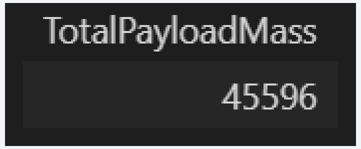
- 5 records where launch sites begin with `CCA`:



Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

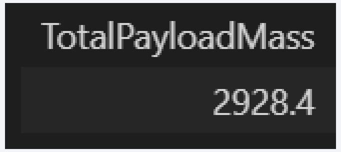
- The total payload carried by boosters from NASA:



TotalPayloadMass
45596

Average Payload Mass by F9 v1.1

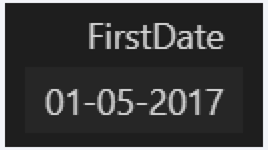
- The average payload mass carried by booster version F9 v1.1:



TotalPayloadMass
2928.4

First Successful Ground Landing Date

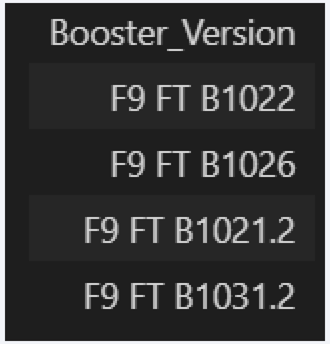
- The dates of the first successful landing outcome on ground pad:



FirstDate
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

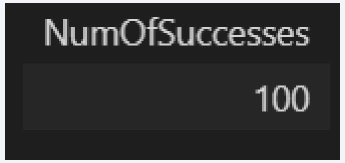
- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000:



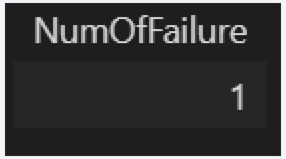
Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes:



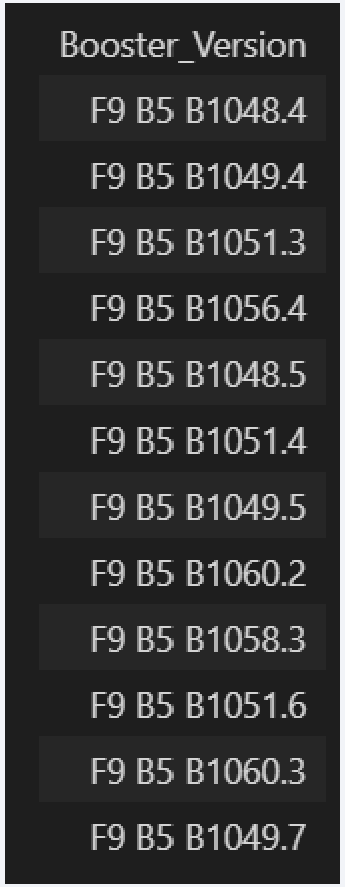
NumOfSuccesses
100



NumOfFailure
1

Boosters Carried Maximum Payload

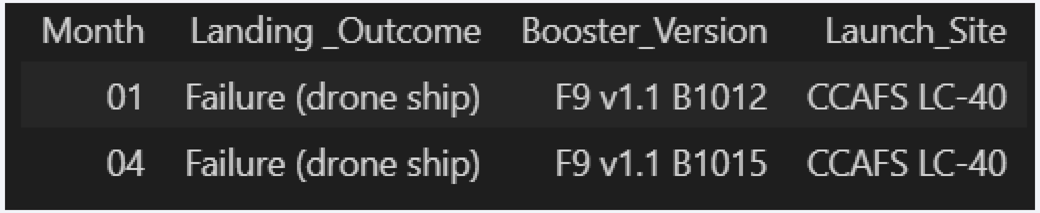
- The names of the booster which have carried the maximum payload mass:



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

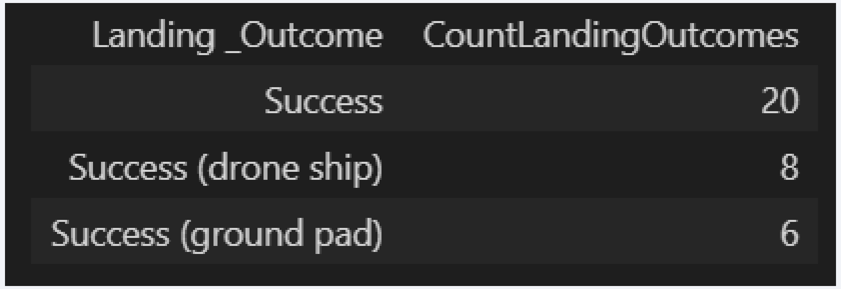
- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015:



Month	Landing _Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- The 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:

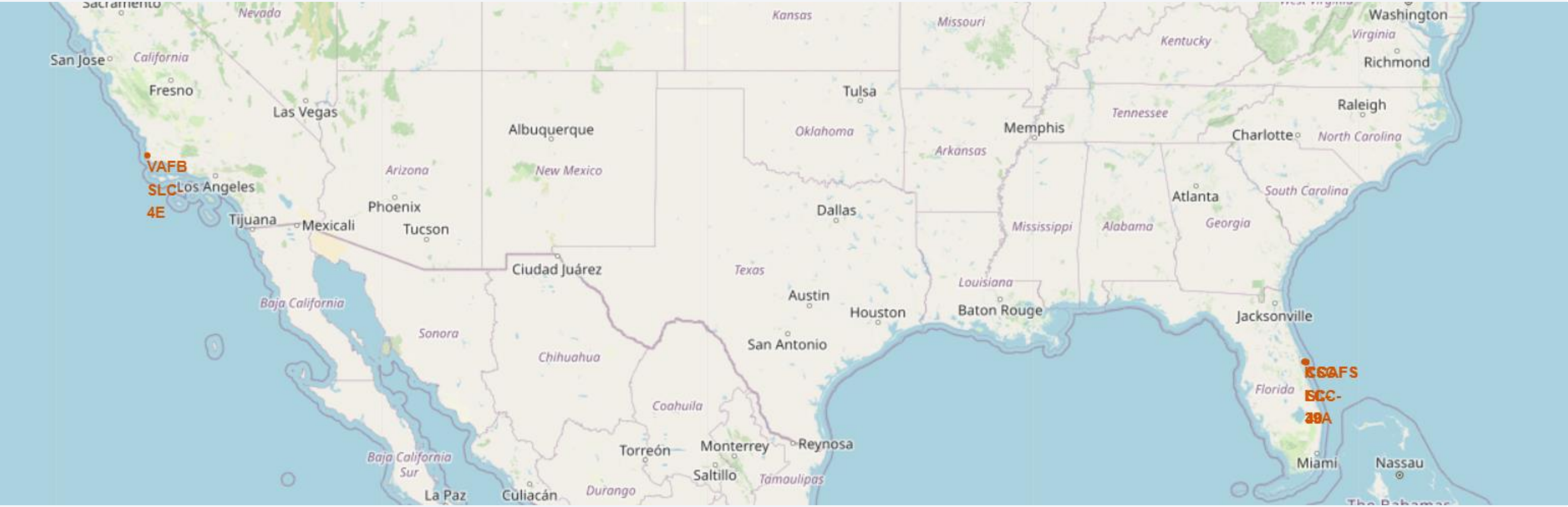


Landing _Outcome	CountLandingOutcomes
Success	20
Success (drone ship)	8
Success (ground pad)	6

Section 3

Launch Sites Proximities Analysis

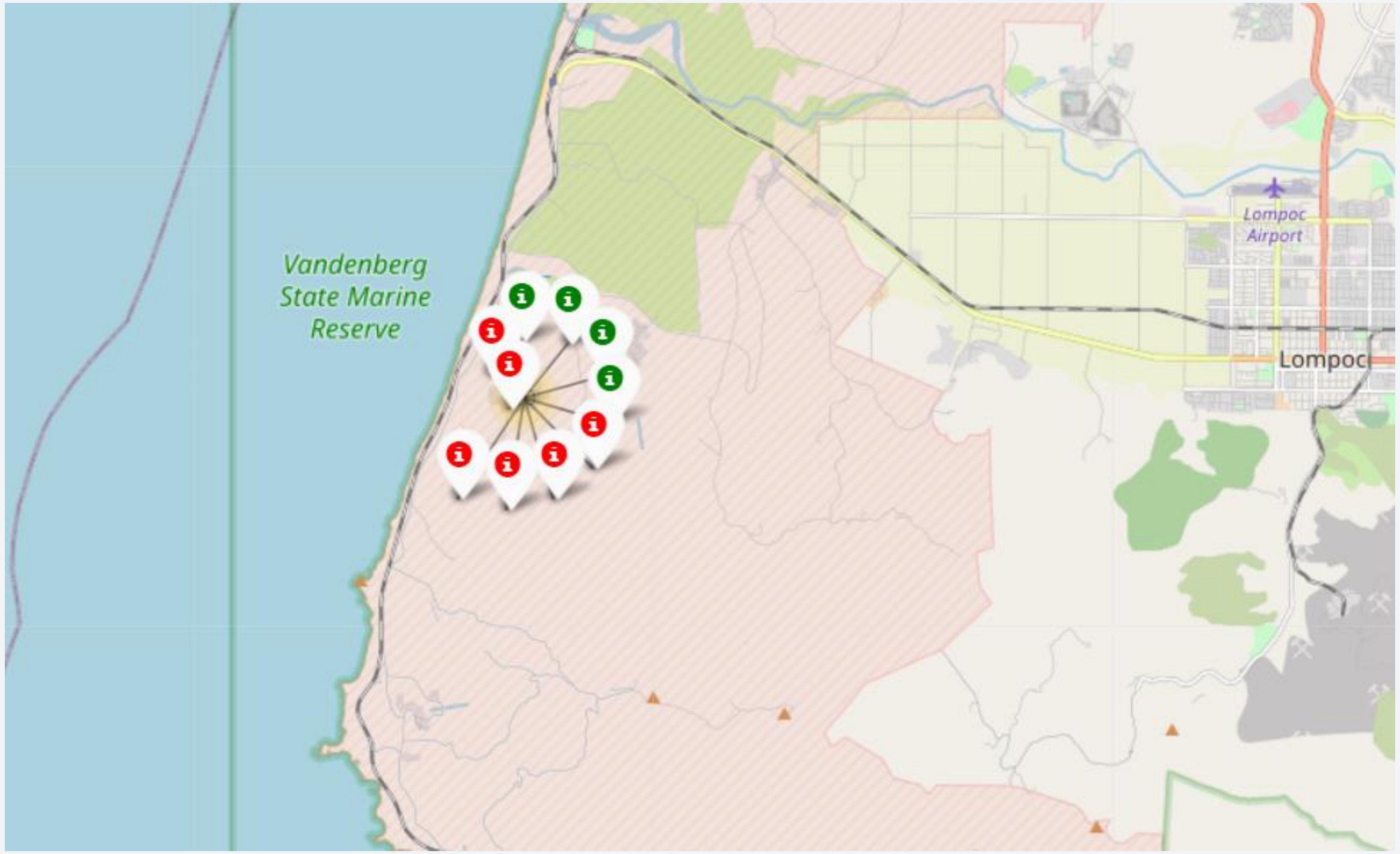
Map of all launch sites



- This map shows locations of all launch sites

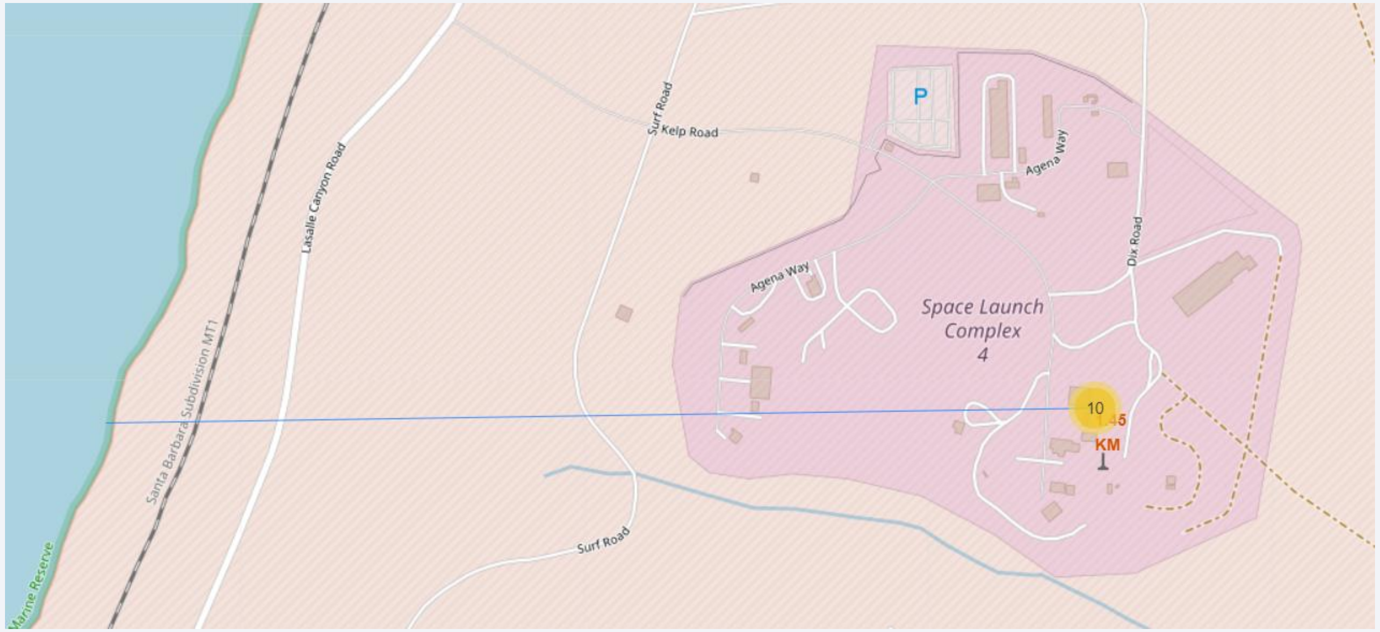
Launch outcomes for site VAFB SLC-4E

- This map demonstrates successful and failed launches for the launch site VAFB SLC-4E



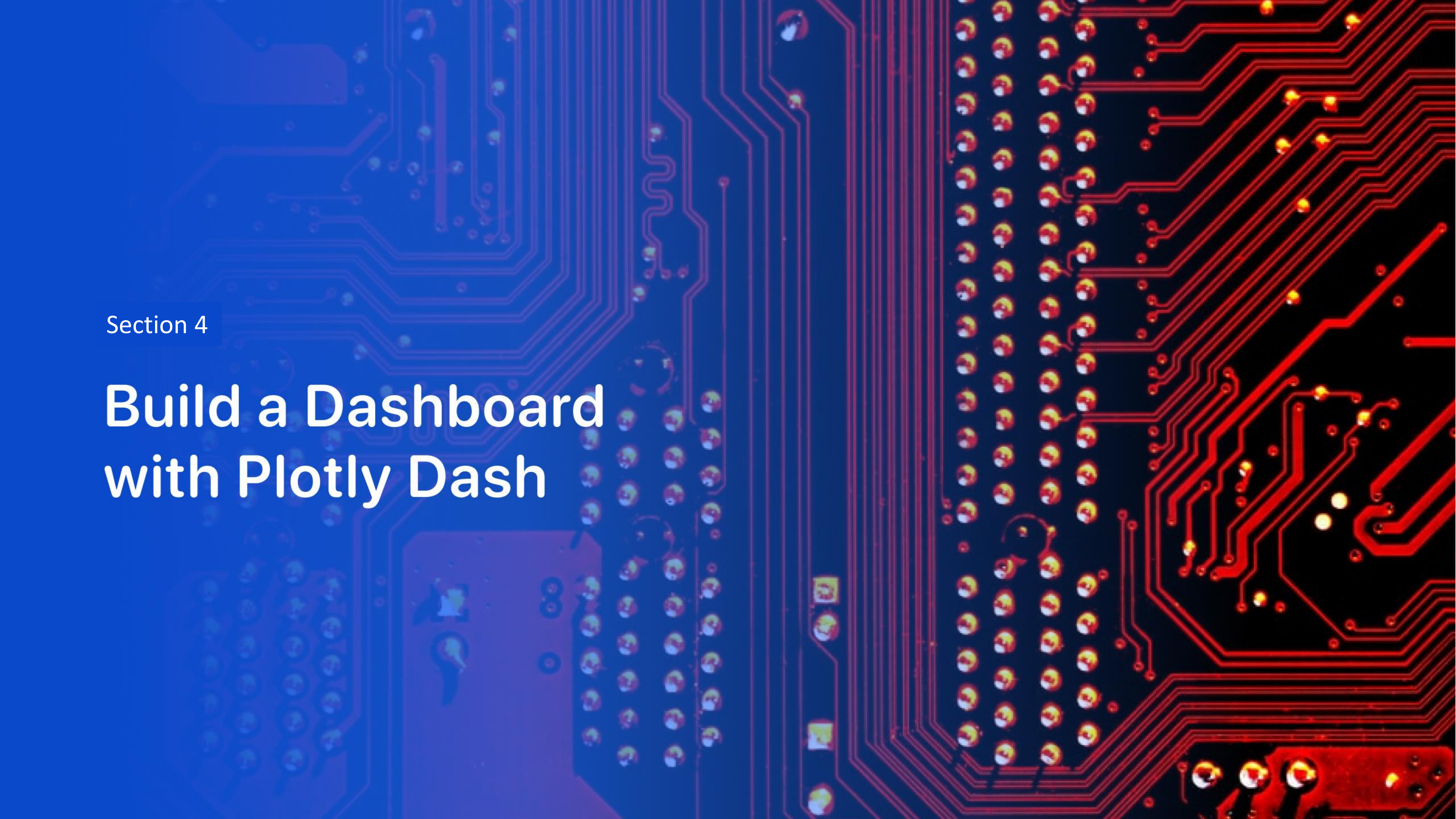
Proximity of the launch site VAFB SLC-4E to coastline

- This map shows proximity of launch site VAFB SLC-4E to the coastline with calculated distance on km.



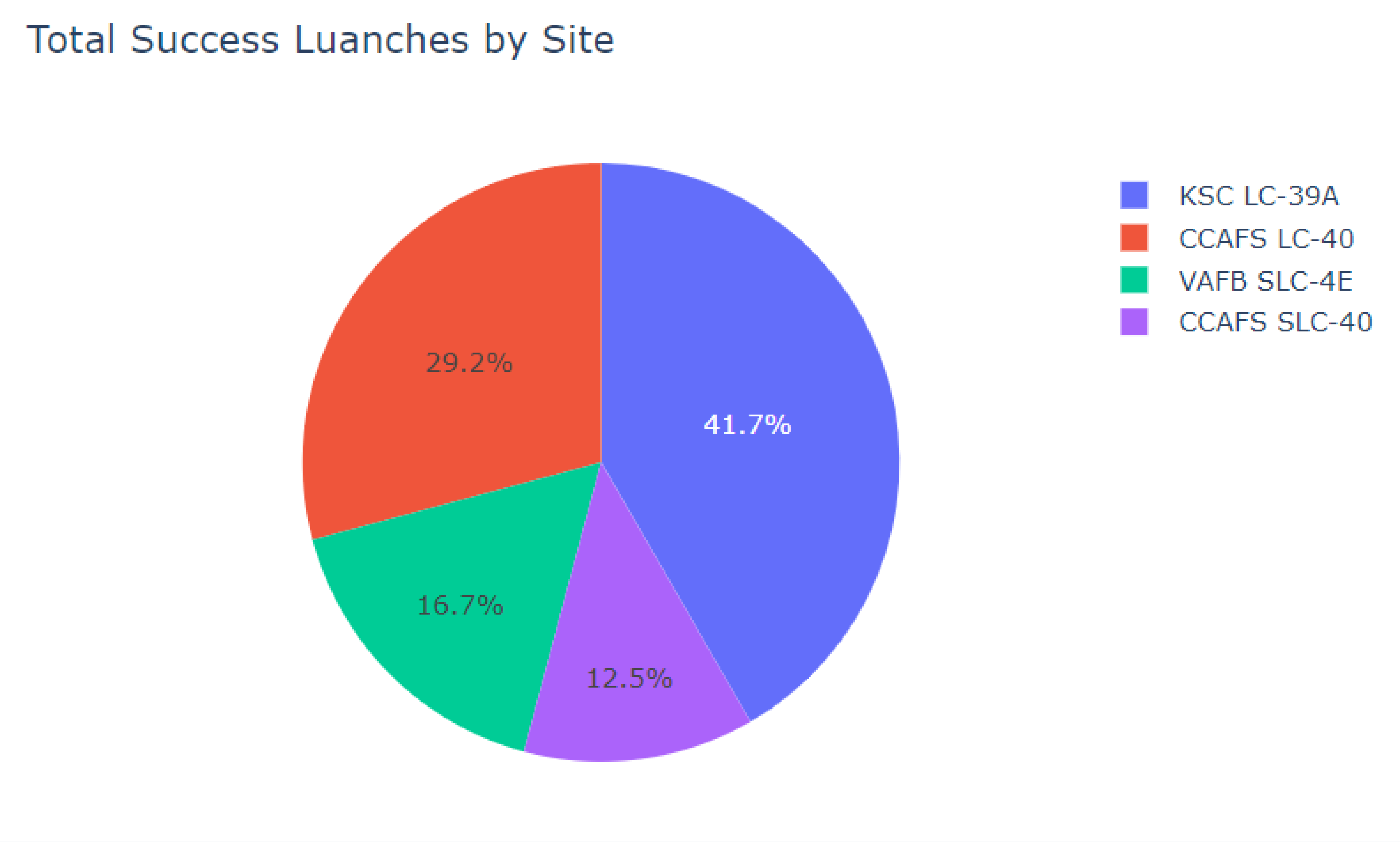
Section 4

Build a Dashboard with Plotly Dash



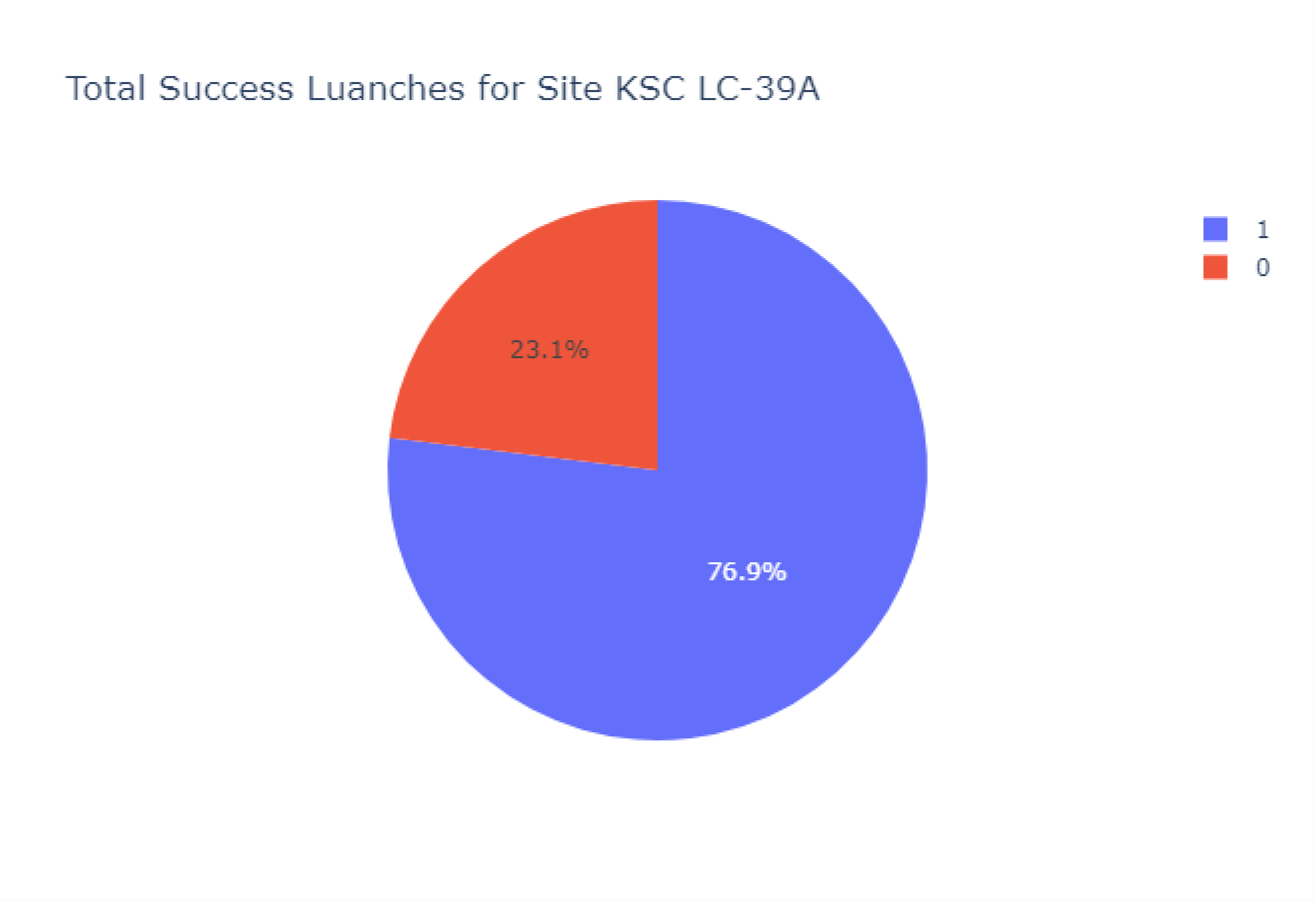
Launch success count for all sites

- The chart demonstrates the launch success count for all sites in percentage of total successful launches
- The highest percentage of successful launches happened on site KSC LC-39A

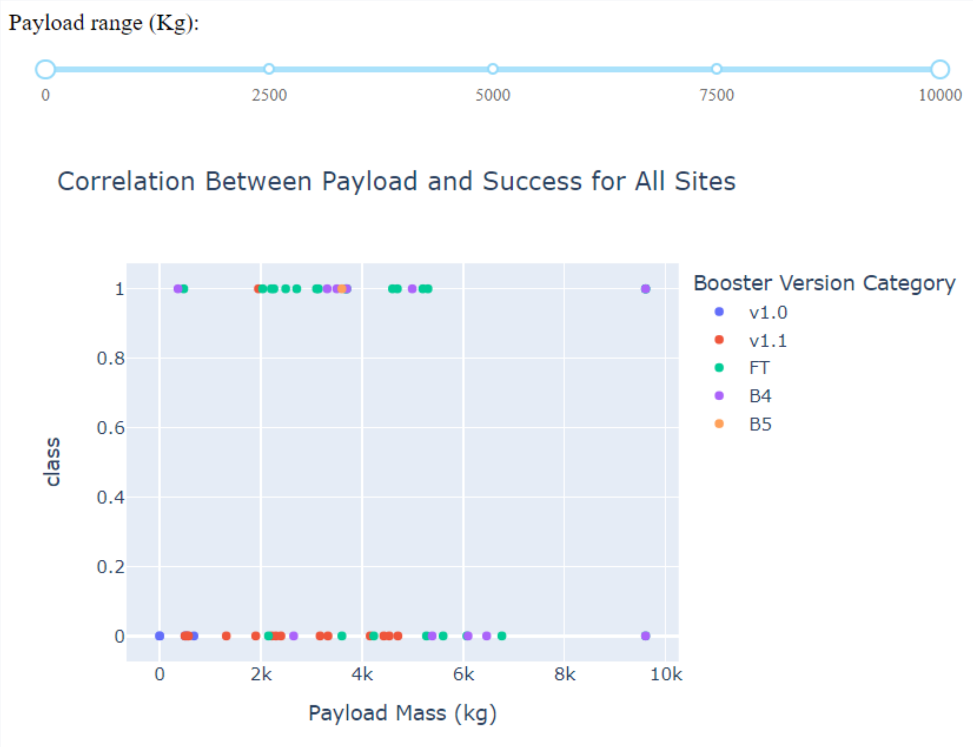


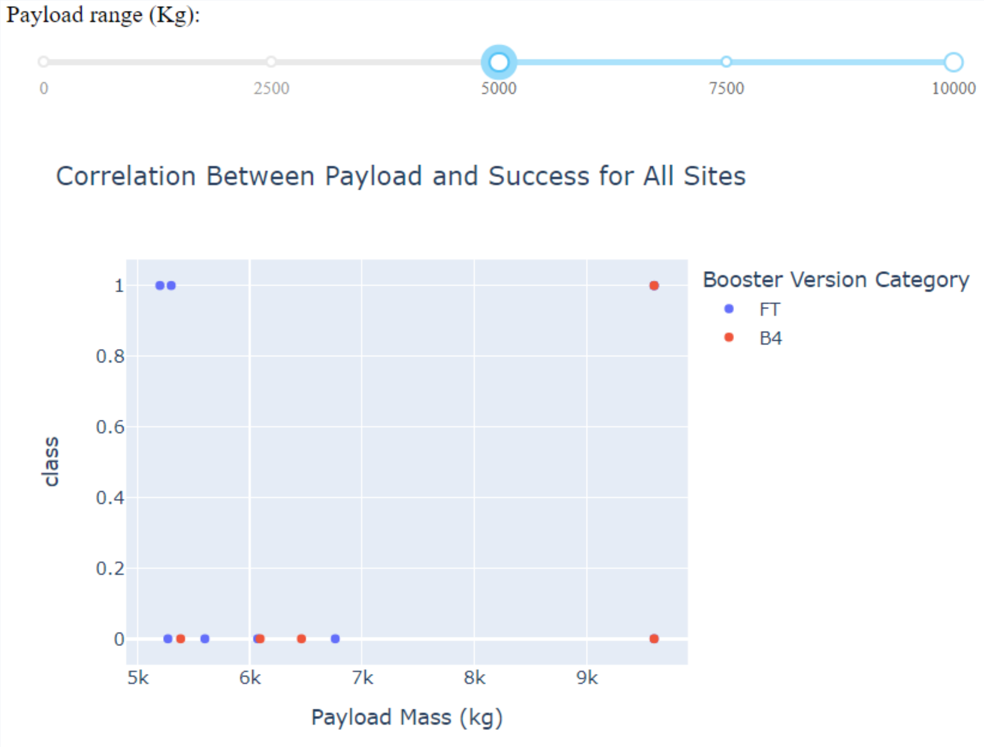
Launch success ratio for the site KSC LC-39A

- The graph demonstrates launch success ratio for the site KSC LC-39A
- The percentage of successful launches (in blue) is 76.9%, while the percentage of failures (red) is 23.1%



<Dashboard Screenshot 3>





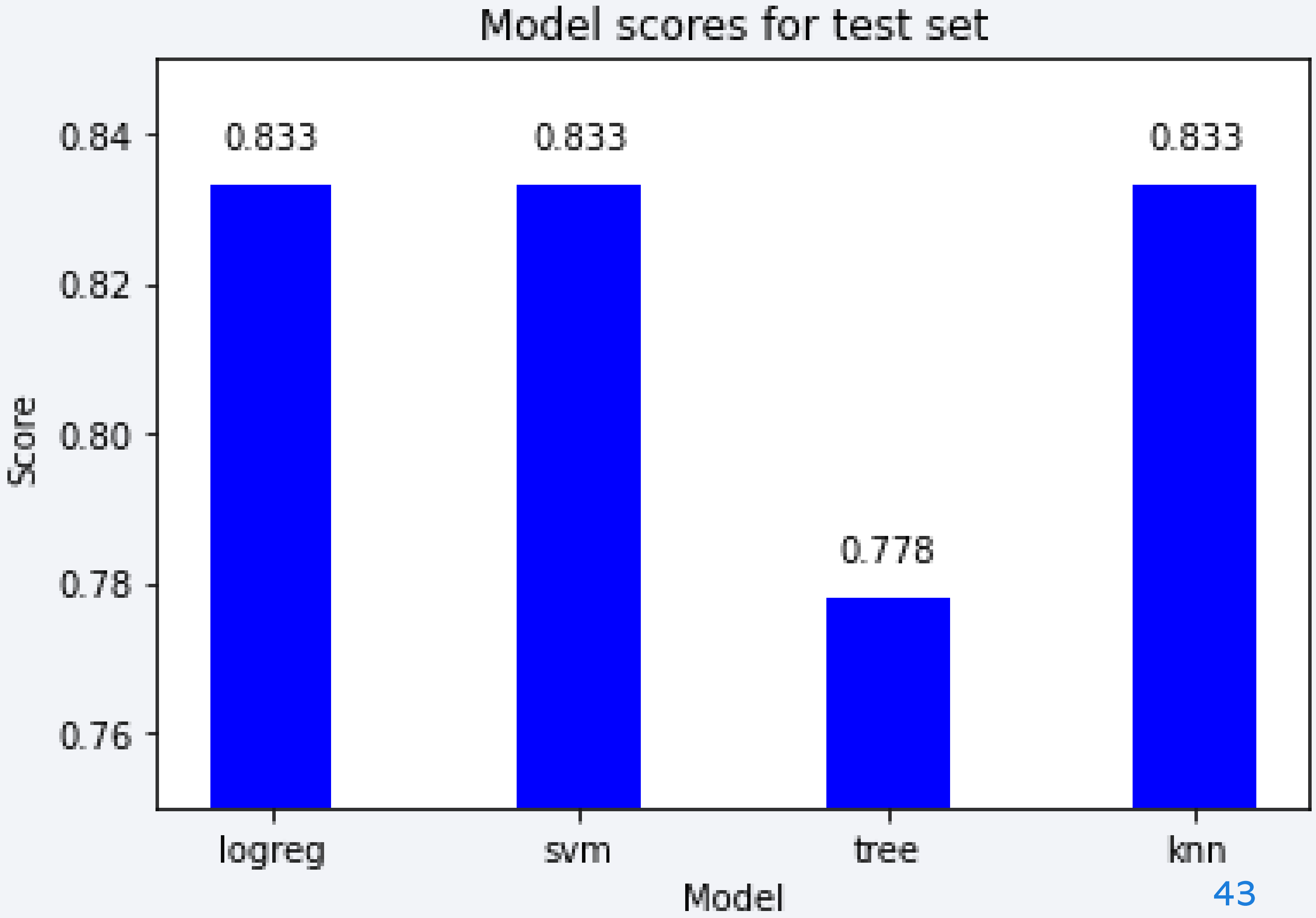
- The graphs demonstrate correspondence of launch outcomes to payload mass and booster version. Chart of the left visualizes data for all sites and the full range of payloads, while the chart of the right is focused on payload over 5000 Kg.

Section 5

Predictive Analysis (Classification)

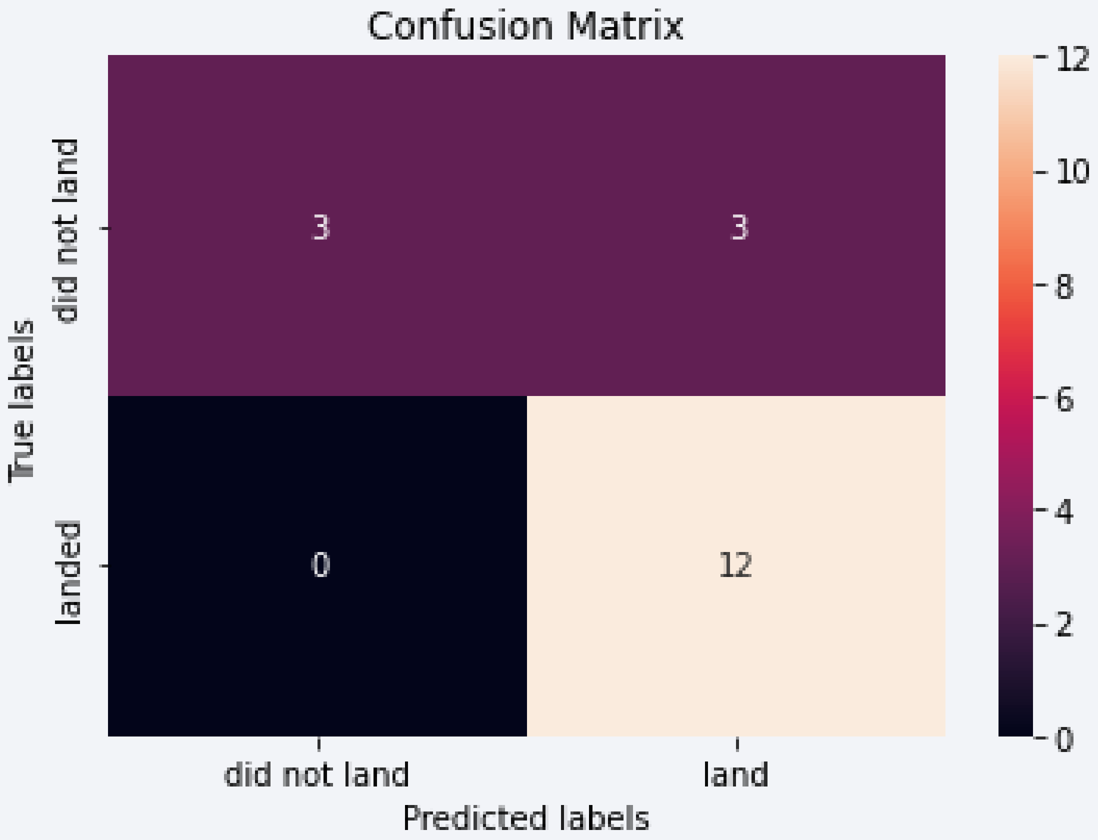
Classification Accuracy

- 3 models have similar performance results on test set: logreg, svm, and knn.
- Out of these 3 options, knn has the best performance on train set.



Confusion Matrix

- Confusion matrix form k-nearest neighbor model
- The matrix is divided into four quantiles:
- The first (upper left) and the third (lower right) quantiles demonstrate correct classification (15 out of 18 cases)
- The second (upper left) and the fourth (lower left) quantiles demonstrate incorrect classification (3 out of 18 cases)



Conclusions

- Earlier launches had lower success rate than the later ones.
- Launces with higher payload mass had higher success rate. Launches for orbits ES-L, GEO, HEO, and SSO has outstanding success rate, while launches to the orbit SO were highly unsuccessful. Over time launches switched to different orbits.
- The highest percentage of successful launches happened on site KSC LC-39A. 76.9% launches on this site were successful.
- Overall success rate since 2013 kept increasing till 2020.
- Four ML models were built to predict outcomes a prospective lunch. K-nearest neighbor model has the highest predictive accuracy both for training and test sets (85% and 83%).

Appendix

- The project Git repository: https://github.com/andriy-savka/IBM-Applied-Data-Science-Capstone.git

Thank you!