IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Lawrence Liu
9/10/2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection: SpaceX-API and Webscraping of SpaceX Wikipedia page
  - Data Wrangling: Replacing missing Values by mean values
  - Exploratory Data Analysis:
    - Analyze outcome by orbit type
    - Analyze outcome by payload mass and booster versions with SQL
    - Visual Analysis with charts by payload mass, time, orbit type and launch site
    - Visual Analysis with map by site
  - Interactive Dashboard: Web Application of Analysis by Site, Payload and Booster version
  - Predictive Analysis Using Classification: Logistic Regression, SVM, Decision Tree, KNN
- Summary of all results
  - Launch success rate increases over time
  - Higher success rate for higher orbits
  - Higher success rate for higher payload mass
  - Low success rate for booster versions v1.0, v1.1, high success rate for FT, B4, B5
  - Higher success rate for Kennedy Space center and recent starts at Cape Canaveral

Introduction

- Project background and context
  - SpaceX advertises low-cost Falcon 9 rocket launches have an average of $62m comparing to $165m from its competitors.
  - This success is because of the reusability of the first stage
- Problems you want to find answers
  - If we can determine if the first stage will land, we can determine the cost of a launch and prepare for the future tasks accordingly.

Section 1

Methodology

Methodology

Executive Summary

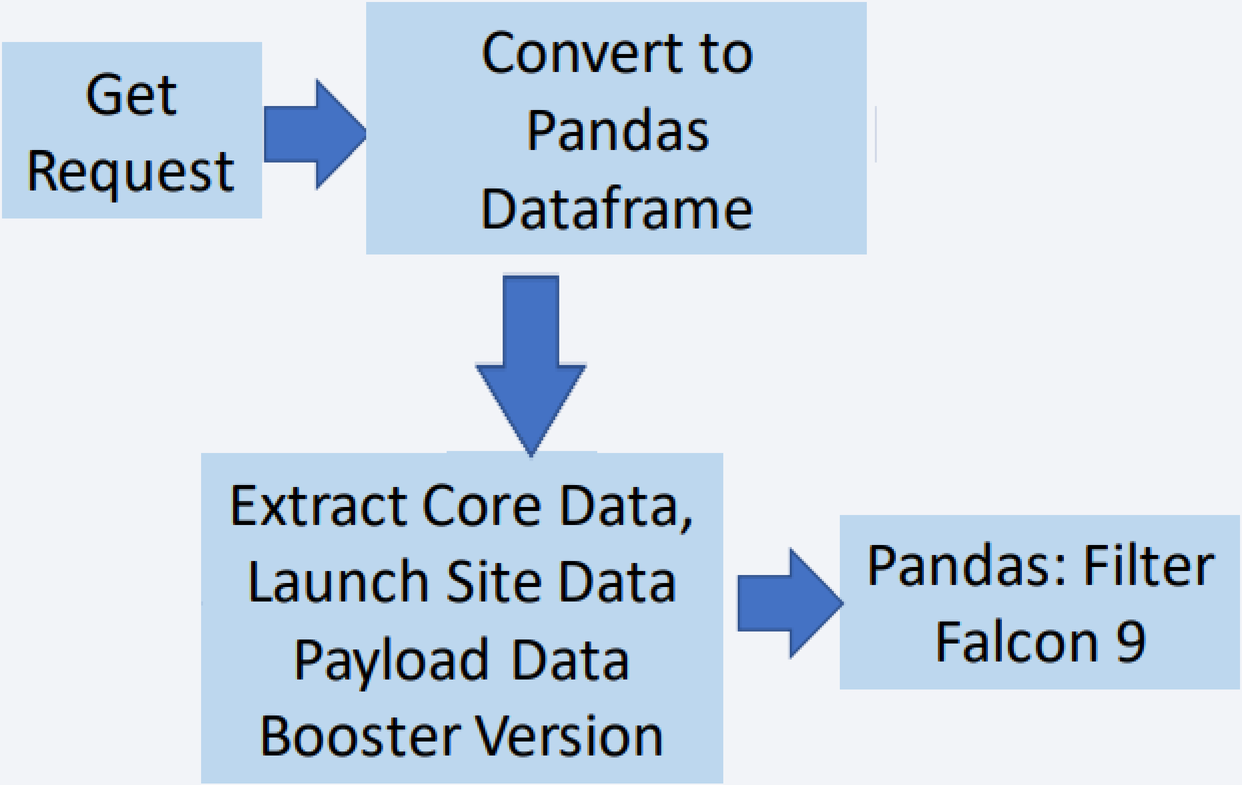
- Data collection methodology:
  - SpaceX-API and Webscraping were used for data collection.
- Perform data wrangling
  - Missing Values of Payload Mass were replaced by mean values of the Payload Mass.
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Analyze outcome by orbit type
  - Analyze outcome by payload mass and booster versions with SQL
  - Visual Analysis with charts by payload mass, time, orbit type and launch site
- Perform interactive visual analytics using Folium and Plotly Dash
  - Visual Analysis with map by launch site
  - Interactive Dashboard: Analysis by Site, Payload and booster version in dropdowns and callbacks
- Perform predictive analysis using classification models
  - Logistic Regression, SVM, Decision Tree, KNN
  - Visual Analysis of Confusion Table

Data Collection

- Describe how data sets were collected.
  - SpaceX REST API and Webscraping of SpaceX Wikipedia Page were used to collect data. As for SpaceX REST API, it is a RESTful Interface, which was used to get Core Data, Booster Version, Launch Site Data and Payload Data. Webscraping of SpaceX Wikipedia Page used HTML Requests (HTTP-Get) and Python / BeautifulSoup (Package for Webscraping) to extract column names from HTML table header in the webpage.
- Data Collection Jupyter Notebook

Data Collection – SpaceX API

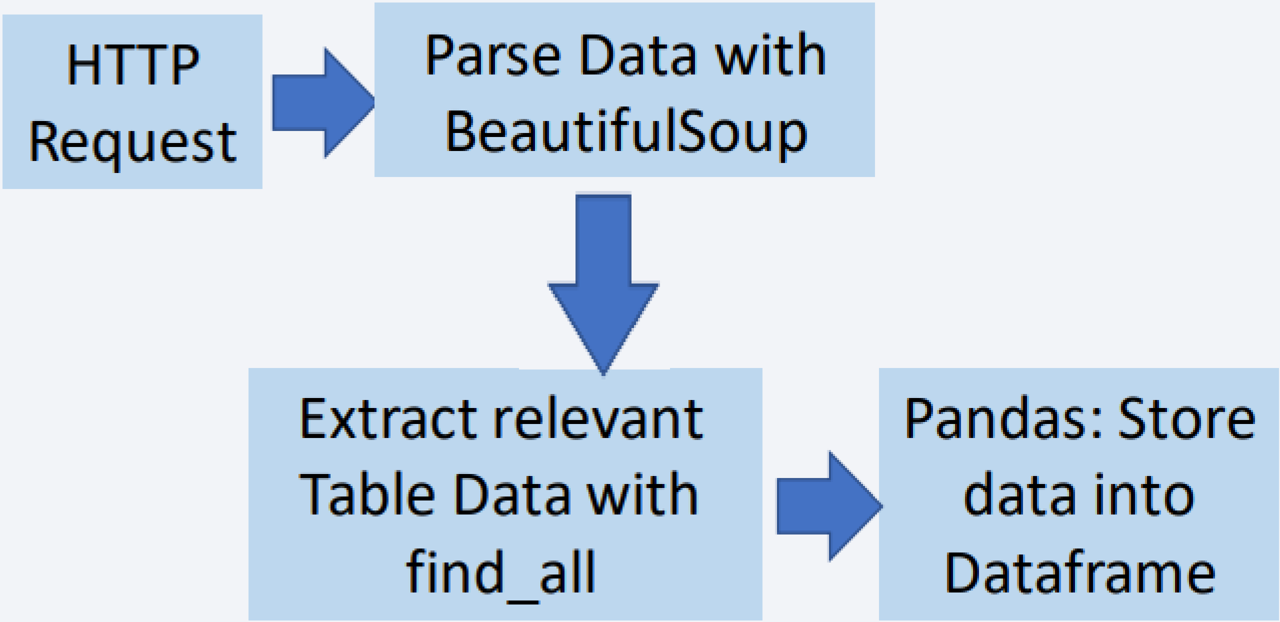
- Data collection with SpaceX REST calls using key phrases and flowcharts
  - Send Get Request to SpaceX API interface website
  - Parse data into Pandas dataframe
  - Extract data with specific functions for:
  - Core data
  - Launch Site Data
  - Payload Mass
  - Booster Version
  - Since Data contains other than Falcon 9 data, we filter for Falcon 9 data only
- [GitHub URL of the completed SpaceX API calls notebook]



Data Collection - Scraping

- Web scraping process
  - Send HTTP Request to SpaceX Wikipedia website
  - Parse data into Pandas dataframe with BeautifulSoup Webscraper
  - Extract data with find_all method
  - Store data into Pandas dataframe for further use

- [GitHub URL of the completed Data Scraping notebook]



Data Wrangling

- The data were processed mainly by replacing the missing payload mass with the mean value of the payload values.

- [GitHub URL of the completed Data Wrangling notebook]

EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
  - Payload mass vs. Flight number vs. Success rate: This shows us the development of the payload mass and the success rate over time
  - Launch site vs. Flight number vs. Success rate: This shows us the success rate of each launch site over time
  - Launch site vs. Payload mass vs. Success rate: This shows us which payload is best to have success at a specific launch site
  - Orbit type vs. Success rate: This can give us a hint which orbit types have the highest success rates
  - Orbit type vs. Flight number vs. Success rate: This shows us the development of orbit types over time
  - Orbit type vs. Payload mass vs. Success rate: Shows us the success rate for specific orbit type / payload mass clusters
  - Success rate vs. Year: Shows the success development over time
- GitHub URL of the completed EDA with Data Visualization

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
    - Extract a list of all launch sites
    - Display 5 records where the name of launch sites starts 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 carried the maximum payload mass
    - List the failed landing_outcomesin 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

- GitHub URL of the completed EDA with SQL

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
  - Edged Circles (radius 1000m): Space launch sites
  - Markers: for labeling all objects
  - MarkerCluster: for creating a bunch of markers around space launch sites to indicate success (green) or failure (red) of the landing of the rocket's first stage
  - Lines: Measure the distance between the launch site and the next coast or next city
- Explain why you added those objects
  - These objects were added to have a better data presentation.
- GitHub URL of Interactive Map with Folium

Build a Dashboard with Plotly Dash

- Input Elements:
  - Dropdown list for the launch site
  - RangeSlider for selecting the payload mass
- Output Elements:
  - PieChart: for showing the success rate of each launch site, or showing the number of successful landing outcomes
  - Scatterplot: Show success/failure by payload and booster version

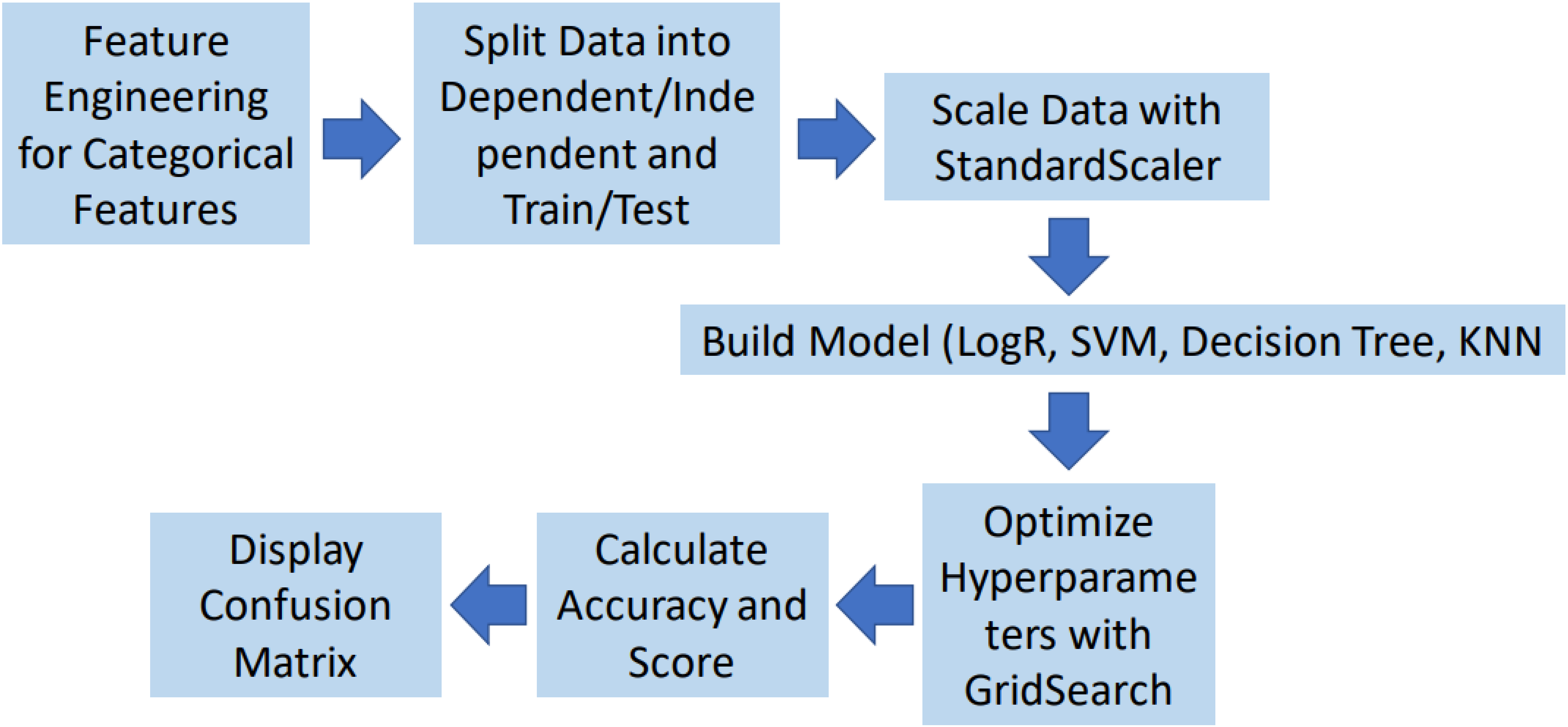
GitHub URL of a Dashboard with Plotly Dash

Predictive Analysis (Classification)

- • Preprocessing
  - One-Hot-Encoding for Categorical Features
  - Split data into dependent/independent variables and train/test data
  - Scale Data with StandardScaler
- • Model Building for each Method
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree
  - K-Nearest Neighbor
- • Optimization
  - Use Gridsearch for optimizing the models based on their hyperparameters
- • Evaluation
  - Use Accuracy of Gridsearch for selecting the best parameter
  - Use Score to compare each classification method

GitHub URL of a Dashboard with Plotly Dash

Predictive Analysis (Classification)



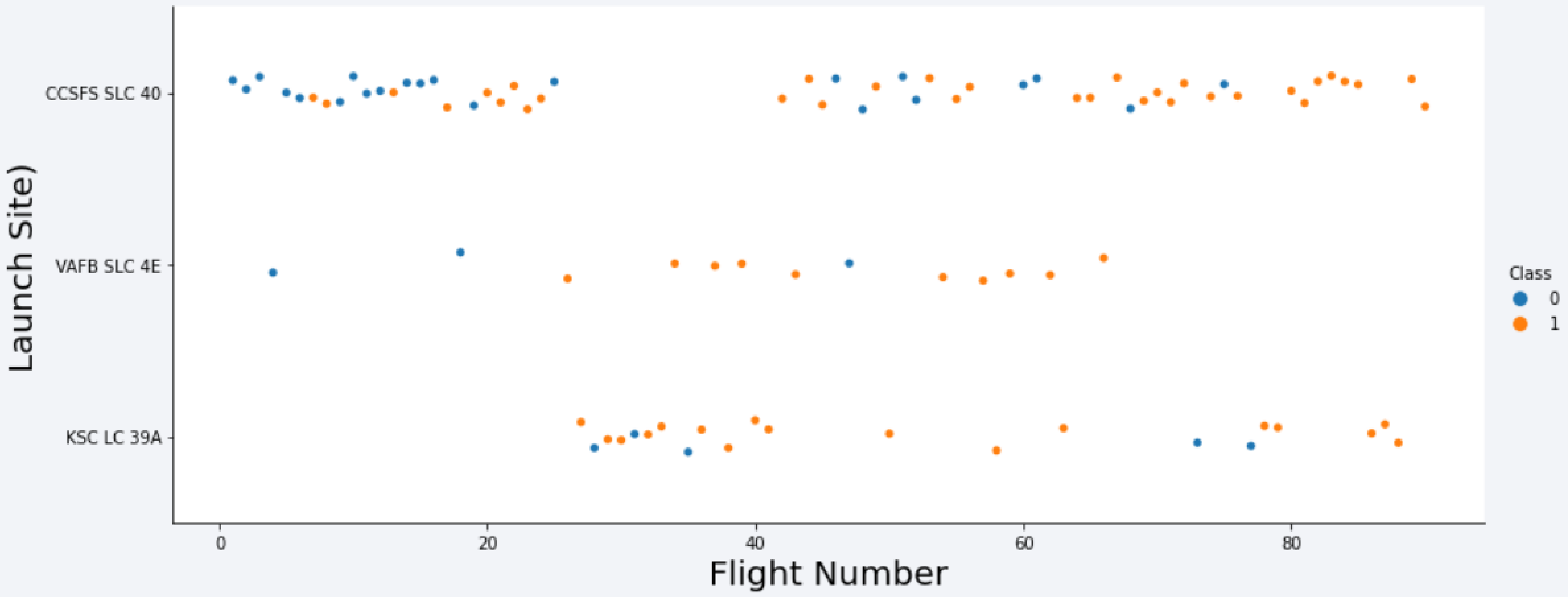
Results

- Exploratory data analysis results
  - Launch success rate increases over time
  - Higher success rate for higher orbits
- Interactive analytics demo in screenshots
  - Higher success rate for higher payload mass
  - Low success rate for booster versions v1.0, v1.1, high success rate for FT, B4, B5
  - Higher success rate for Kennedy Space center and recent starts at Cape Canaveral
- Predictive analysis results
  - Best prediction results with Logistic Regression and Support Vector Machine

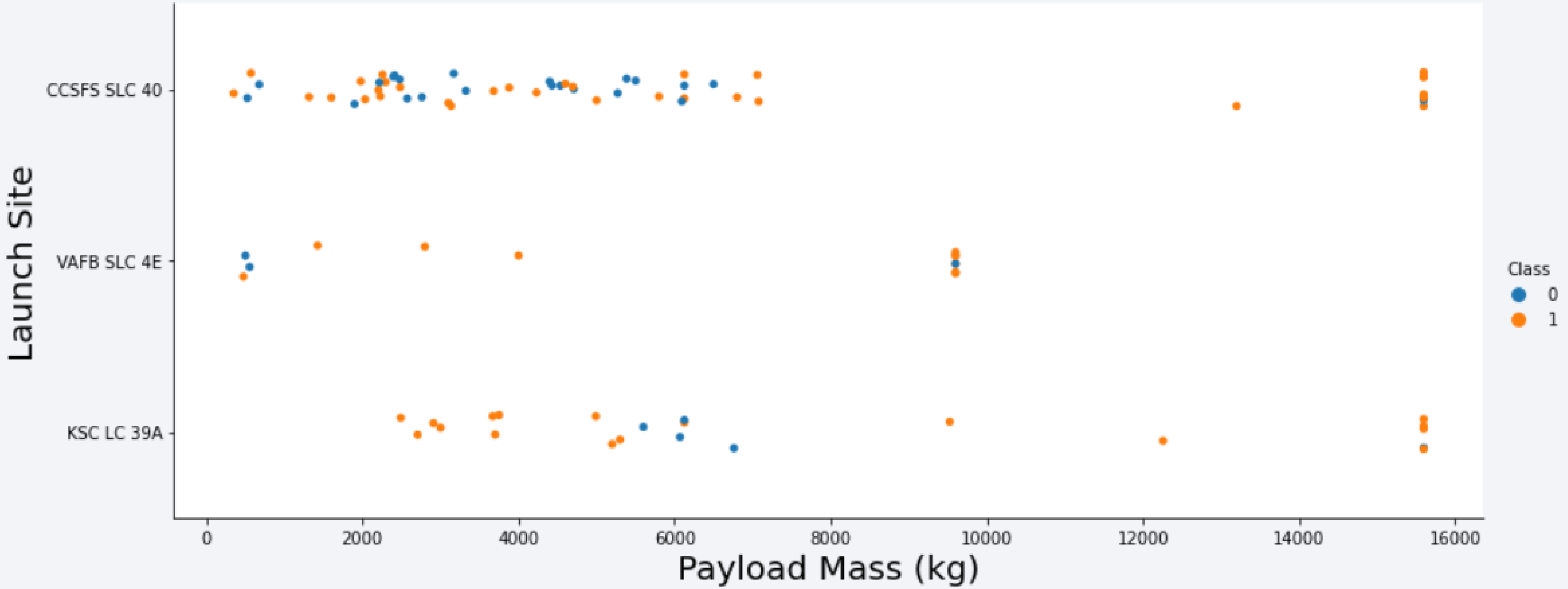
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site



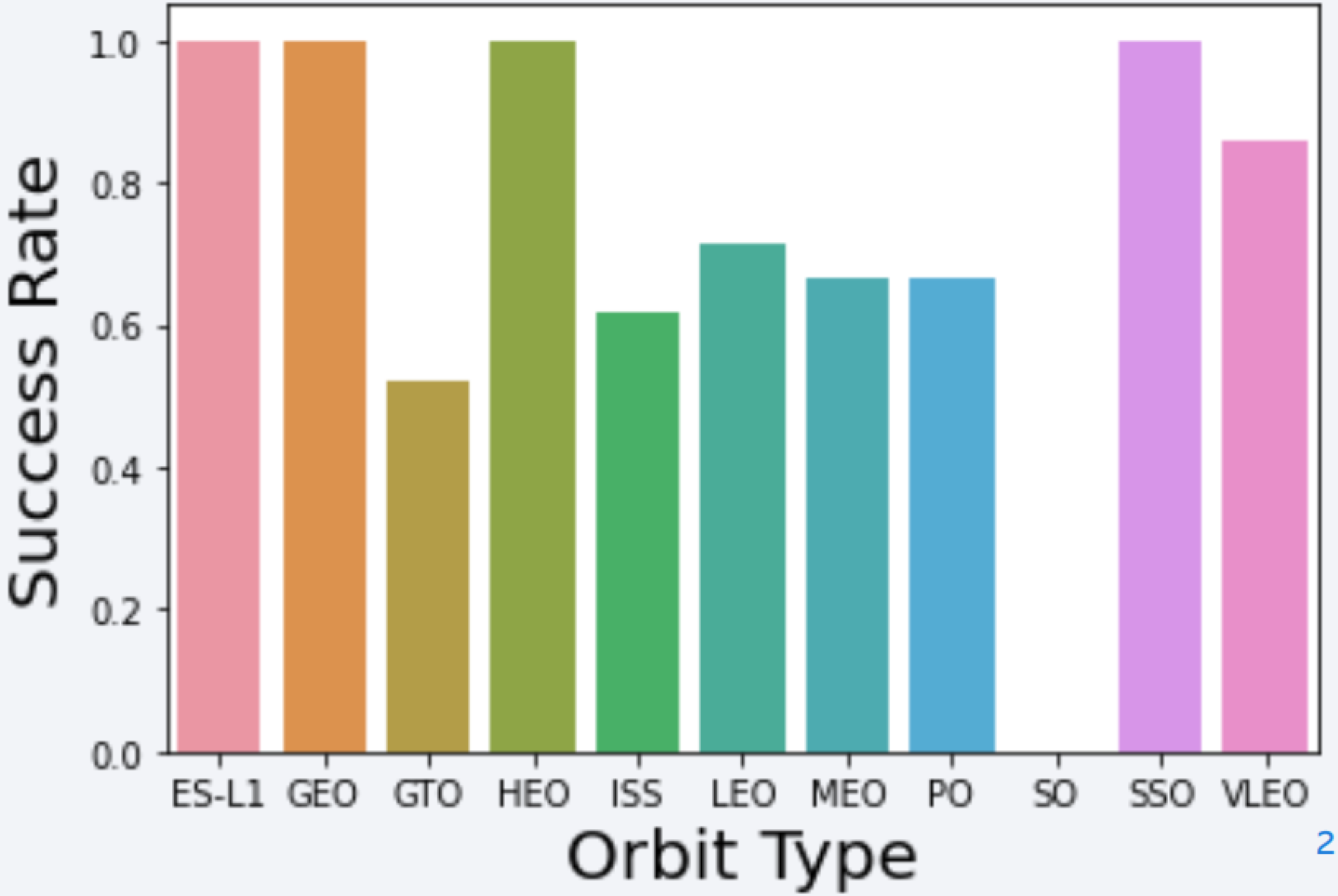
Payload vs. Launch Site



Success Rate vs. Orbit Type

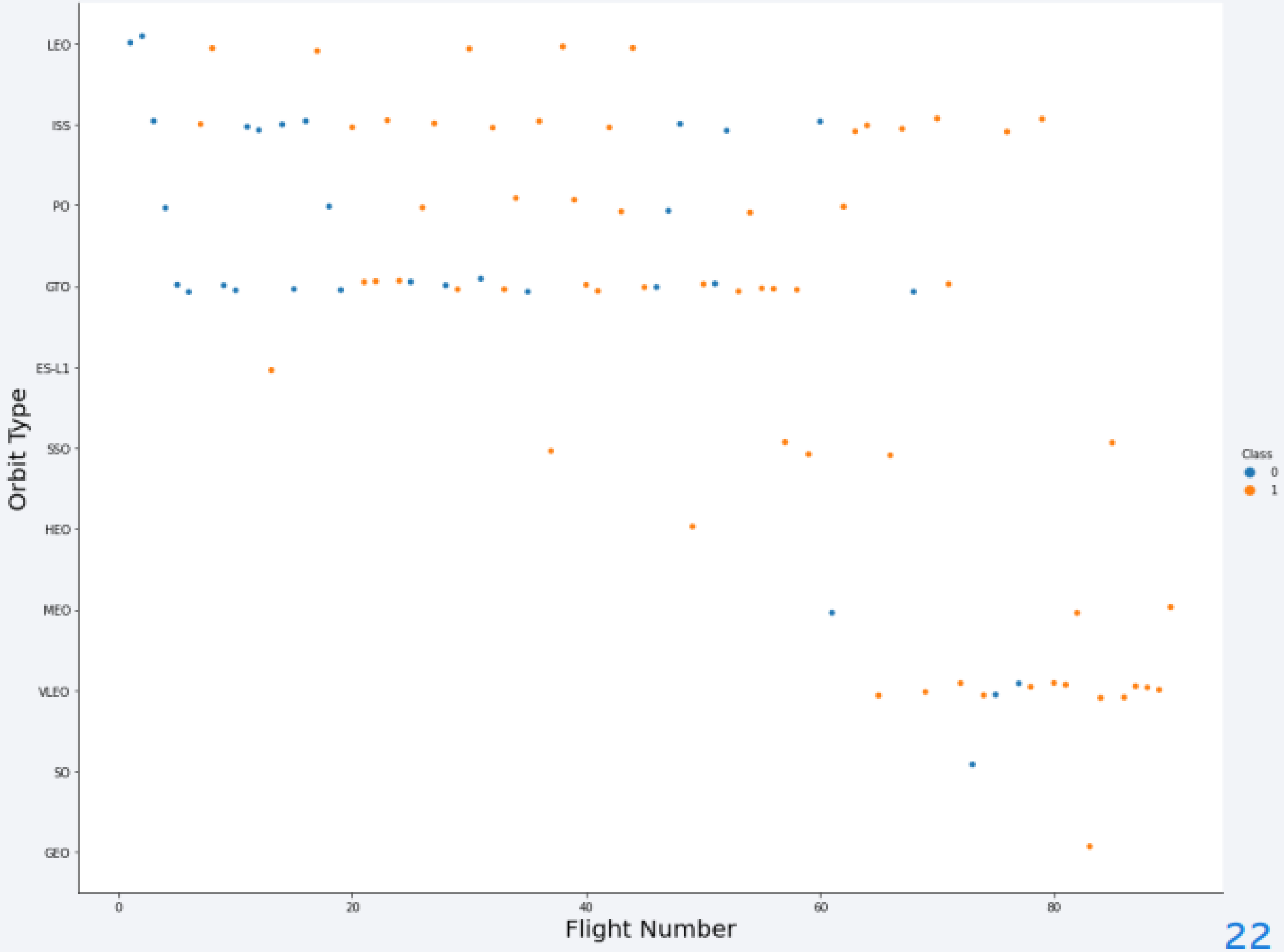
Low Earth Orbits: GTO • ISS • LEO • MEO • PO • VLEO

High Earth Orbits: ES-L1 • GEO • HEO • SSO

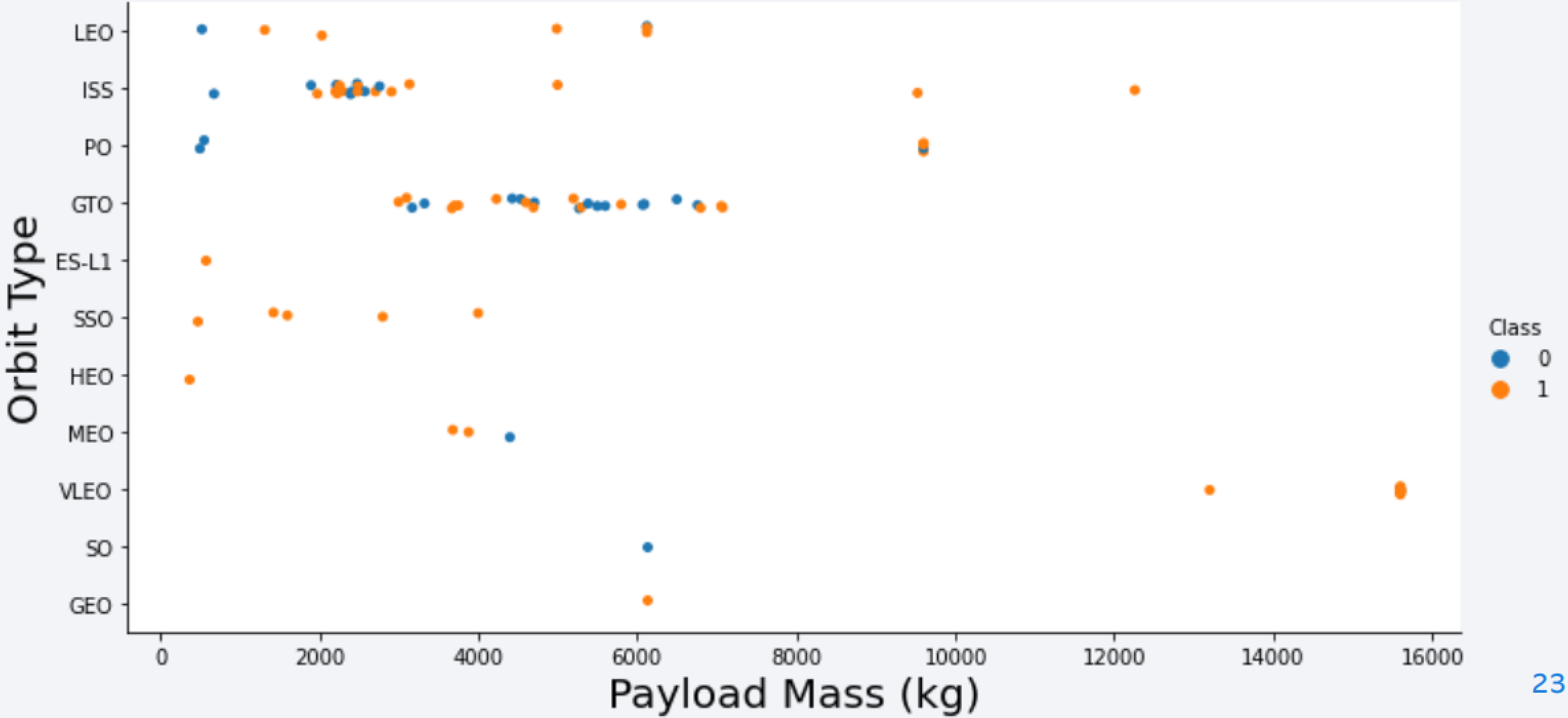


Flight Number vs. Orbit Type

- The orbit types are changing over time. Success rate has increased over time for all orbit types.

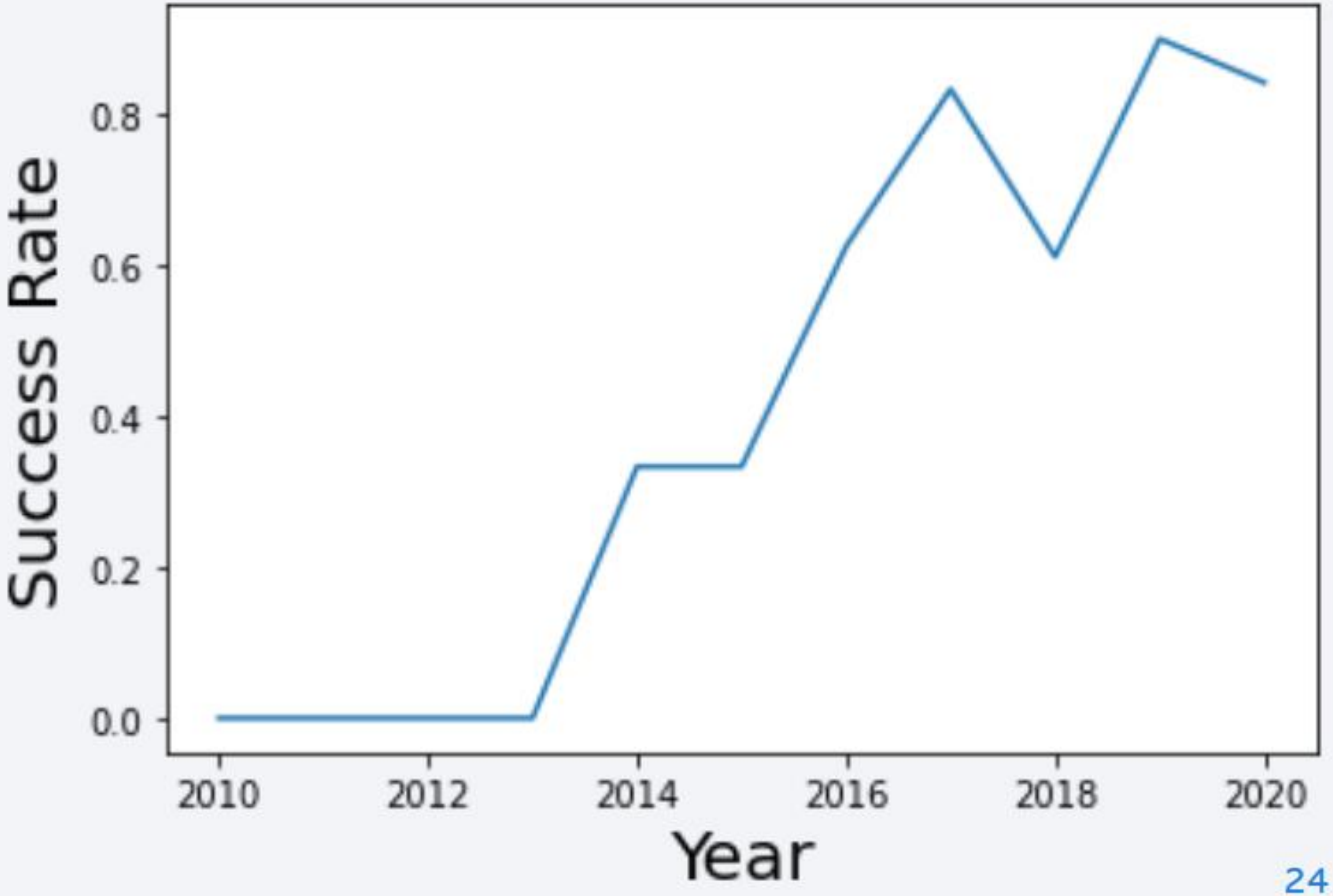


Payload vs. Orbit Type



Launch Success Yearly Trend

Launch success is increasing over the years



All Launch Site Names

- KSC: Kennedy Space Center
- CCA?: Cape Canaveral Launch Center
- VAFB: Vandenburg Air Force Base

	Launch_Site
0	CCAFS LC-40
1	VAFB SLC-4E
2	KSC LC-39A
3	CCAFS SLC-40

Launch Site Names Begin with 'CCA'

Date	Time_(UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_C
2010-06-04 00:00:00	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success
2010-12-08 00:00:00	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
2012-05-22 00:00:00	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success
2012-10-08 00:00:00	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success
2013-03-01 00:00:00	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success

Total Payload Mass

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

	Booster_Version
0	F9 FT B1022
1	F9 FT B1026
2	F9 FT B1021.2
3	F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

Mission_Outcome	count(*)
Failure	1
Success	100

Boosters Carried Maximum Payload

- Names of the booster with the maximum payload mass

	Booster_Version
0	F9 B5 B1048.4
1	F9 B5 B1049.4
2	F9 B5 B1051.3
3	F9 B5 B1056.4
4	F9 B5 B1048.5
5	F9 B5 B1051.4
6	F9 B5 B1049.5
7	F9 B5 B1060.2
8	F9 B5 B1058.3
9	F9 B5 B1051.6
10	F9 B5 B1060.3
11	F9 B5 B1049.7

2015 Launch Records

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
Failure (drone ship)	F9 v1.1 B1017	VAFB SLC-4E
Failure (drone ship)	F9 FT B1020	CCAFS LC-40
Failure (drone ship)	F9 FT B1024	CCAFS LC-40

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

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

Section 3

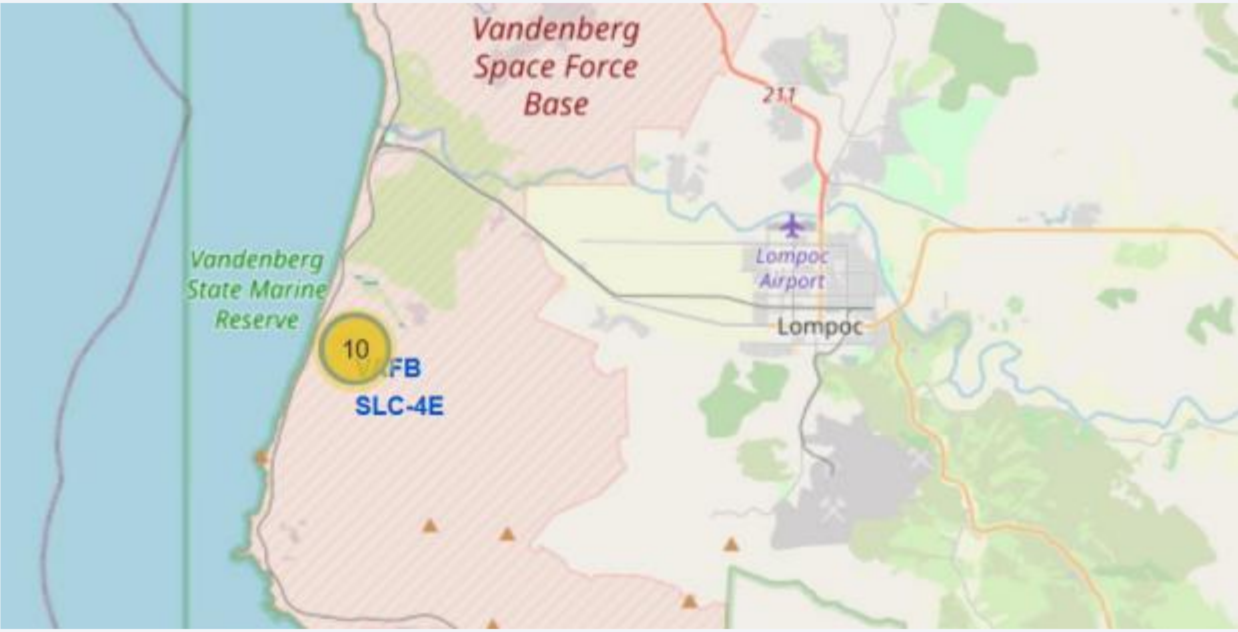
Launch Sites Proximities Analysis

Folium Map: Launch Sites

- Launch sites are at the East and West coast, near the southernmost U.S. mainland area, which is Florida and California

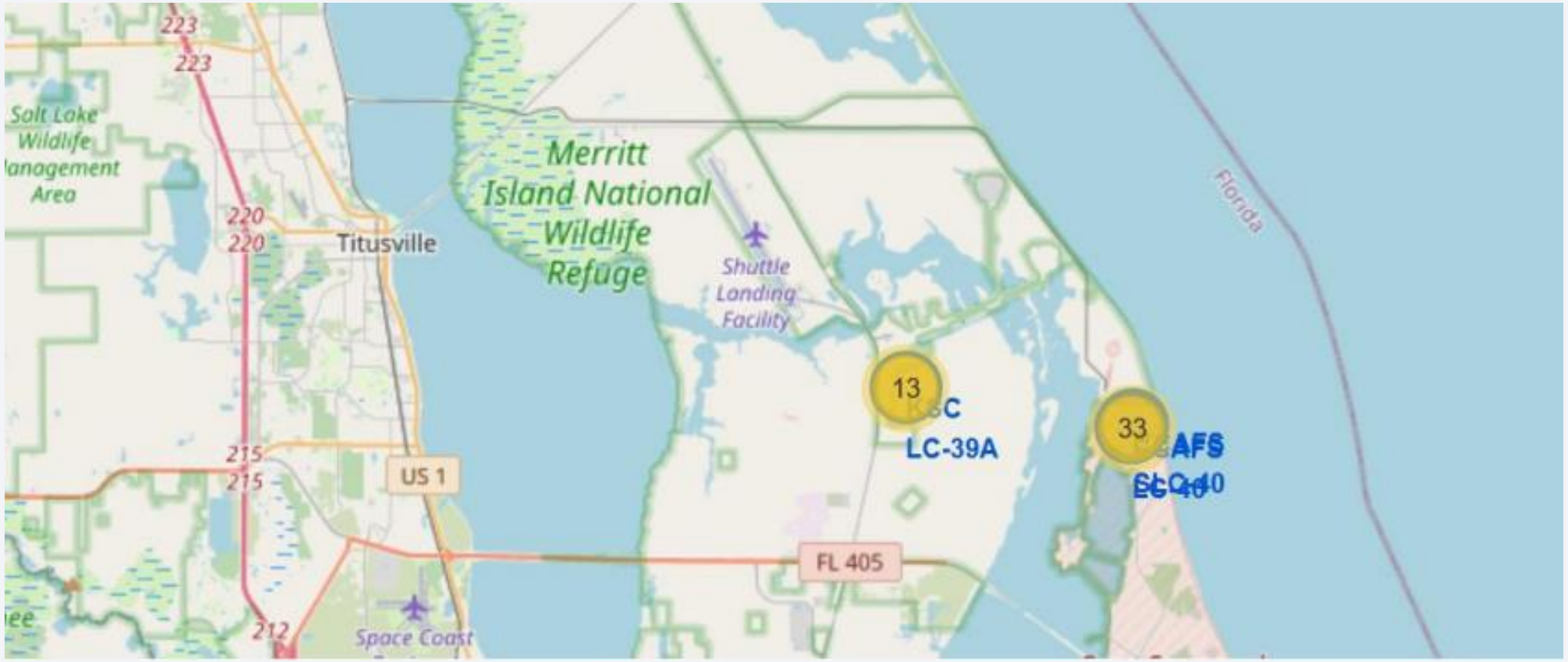
Folium Map: Proximity Vandenburg AFB

- As shown in the map on the right, the town of Lompoc is right next to the Vandenburg AFB. This could be an issue if the stage-1 landing cannot be controlled.



Folium Map: Proximity Kennedy Space Center (KSC)/ Cape Canaveral

- As shown in the map on the right, KSC is isolated to itself.

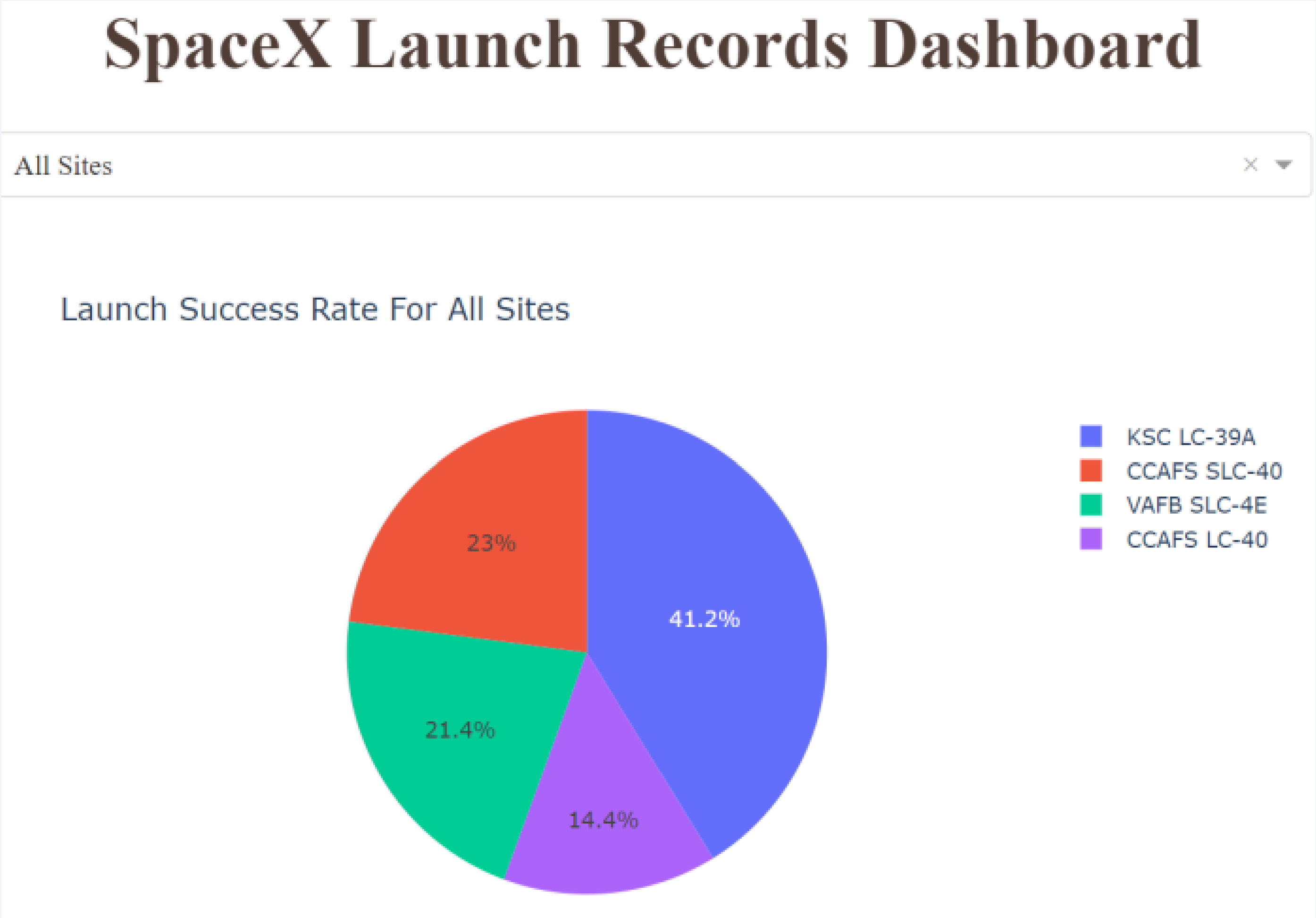


Section 4

Build a Dashboard with Plotly Dash

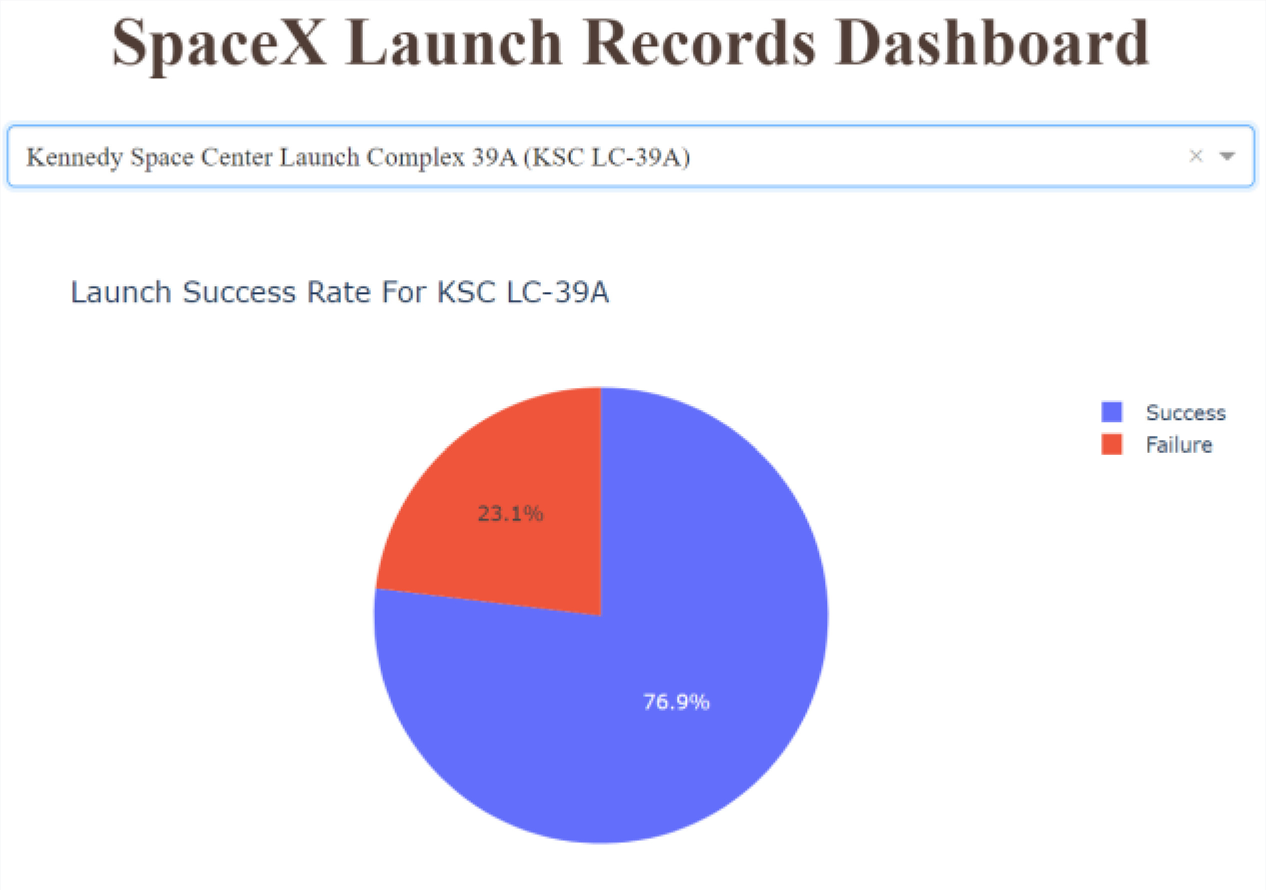
Dashboard: Launch Success Count For All Sites

- KSC hast the most successful stage-1 landings
- Vandenberg Air Force Base has the least number of successful stage-1 landings



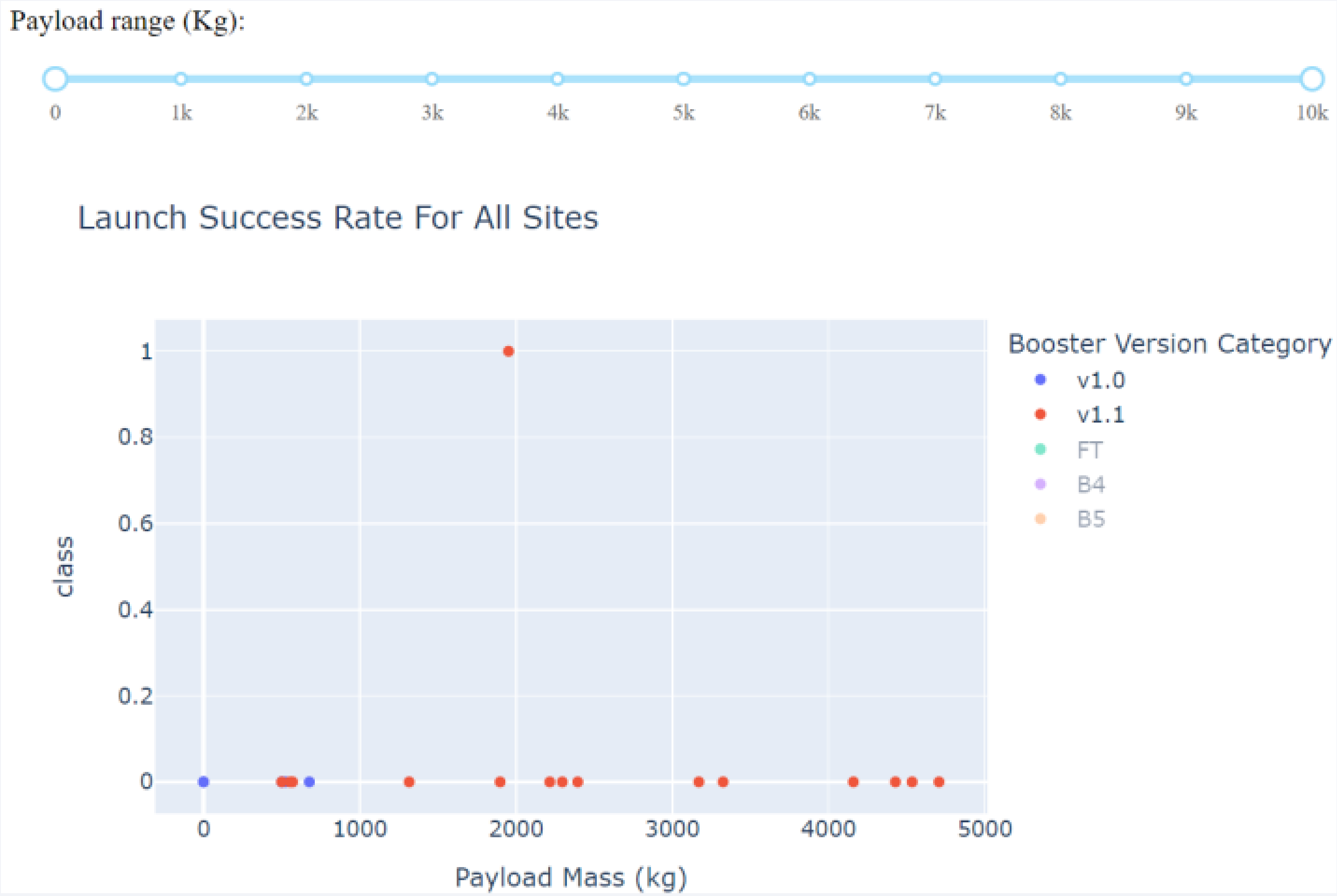
Dashboard: Success Rate at KSC

- Almost 8 out of 10 landings are successful at KSC



Dashboard: Booster Version V 1.0 and V1.1

- This graph basically can not tell anything about the characteristics of Booster Version V1.0 and V1.1, since there is only 1 failed launch with V1.1 and all the rest are successful.

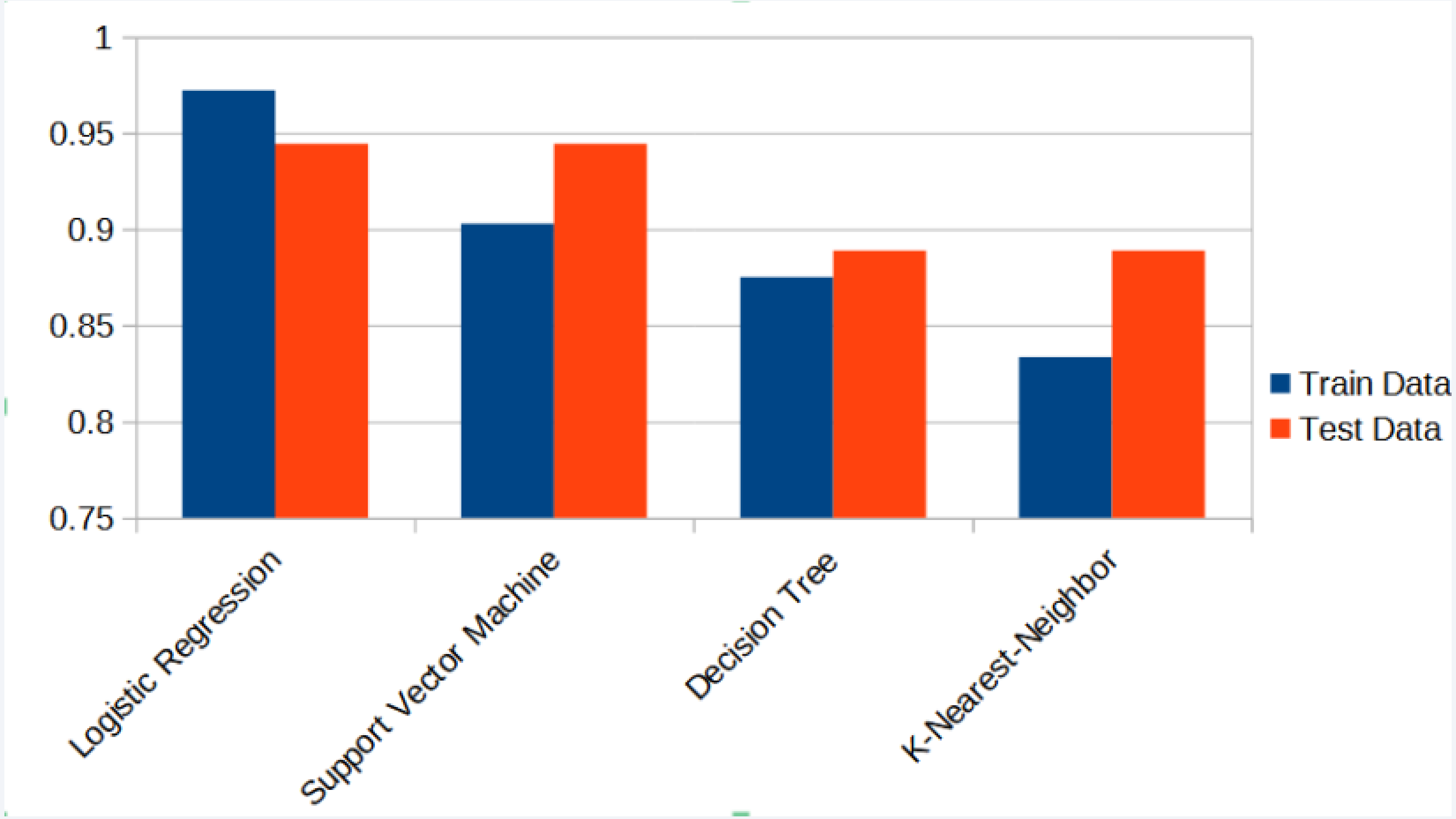


Section 5

Predictive Analysis (Classification)

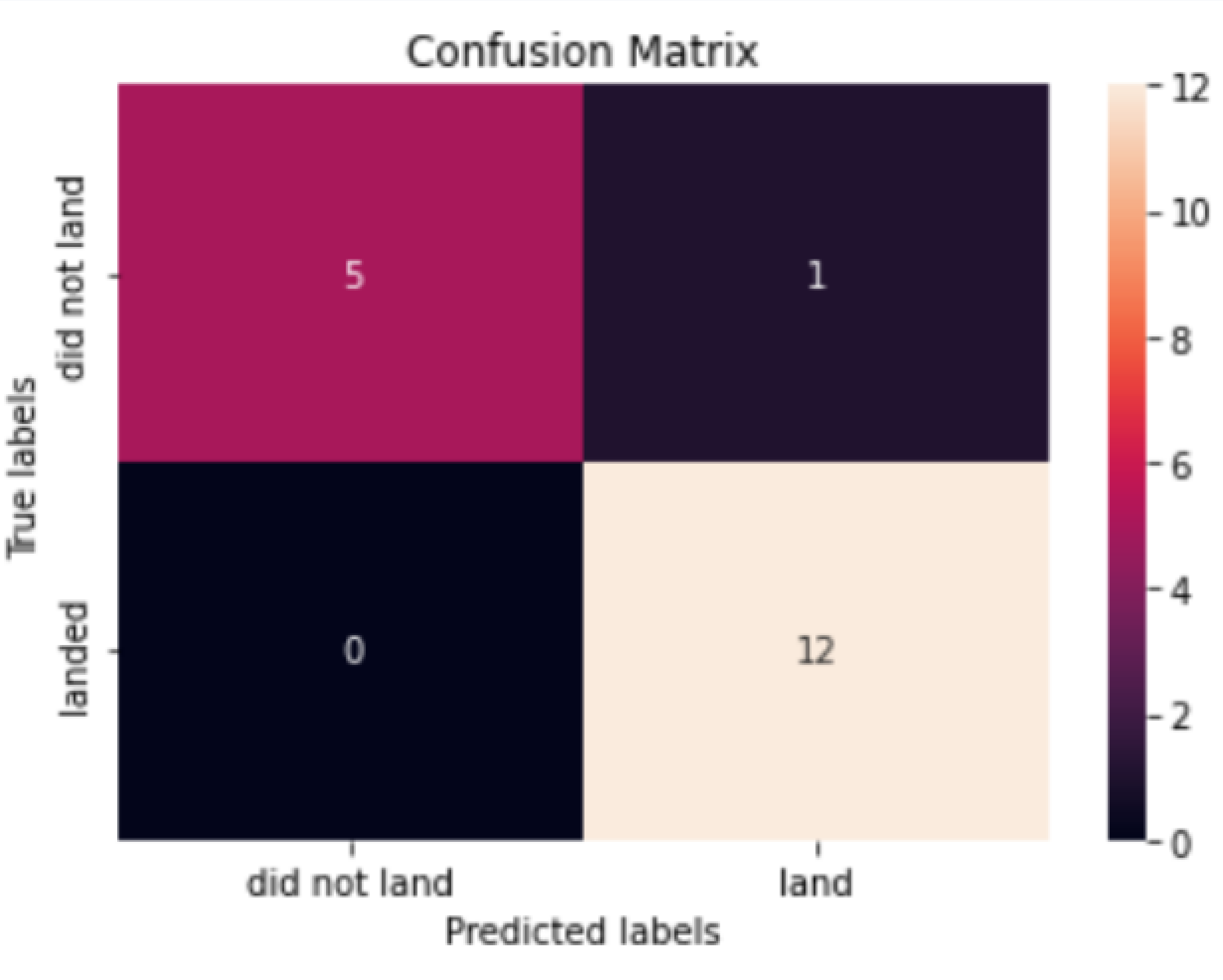
Classification Accuracy

- Logistic Regression has the best result for train data
- Logistic Regression and Support Vector Machines have the best results on test data



Confusion Matrix

- True Positives: 12
- True Negatives: 5
- False Positives: 1
- False Negatives: 0



Conclusions

- None of the prediction is perfectly matching the test data.
- Prediction with Logistic Regression is quite accurate.
- Support Vector Machine also provide a good result for predicting the landing outcome.
- None of the predictions from the Machine Learning (ML) models had false negatives.
- All predictions from the ML models had at least one false positive.

Appendix

Python code, Notebook and SQL are available at:

- Jupyter Notebook
- Plotly Dashboard

The current version of this document is available at:

- Analysis Presentation

Thank you!