IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Li Ma

May 27, 2023

Outline

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

Executive Summary

- We have used the following methods:
    - Data Collection using API and Web Scrapping
    - Data Wrangling
    - Explorative Data Analysis using SQL and Data Visualization
    - Interactive Visual Analytics and Dashboards
    - Machine Learning
- We have obtained the following results:
    - We have collected data from public resources
    - We have found the best features to predict launching result
    - We have established models for prediction using Machine Learning

Introduction

- The company SpaceX is very successful and our object is to help the company SpaceY to compete with SpaceX
- Our tasks:
    - Determine the price of each launch
    - Predict whether SpaceX will reuse the first stage

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - From the API of SpaceX and scrapping websites
- Perform data wrangling
  - Fill in missing data and add new columns to the data frame using original data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Divide data into training and test sets, train models on the training set and evaluate them on the test set.

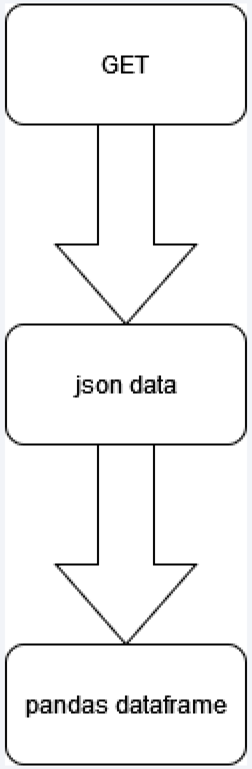
Data Collection

- We collect data from the SpaceX API
- We also collect data from scrapping webpages on Wikipedia

Data Collection – SpaceX API

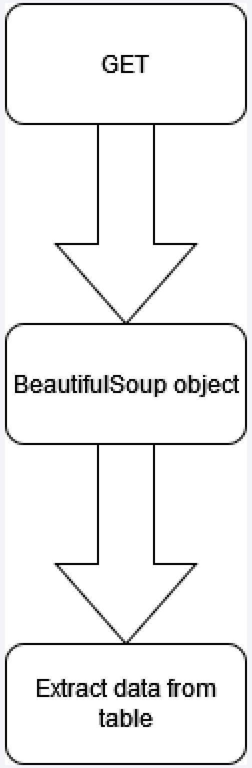
- Use 'get' command to download data from the SpaceX API in json format, then convert it to pandas data frame

- GitHub URL: https://github.com/limabielefeld/coursera_homework/blob/8e5640468a5a13dba316ba4f07d9c40c6be65bbe/Applied%20Data%20Science%20Capstone/1.jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Download webpage from wikipedia to a BeautifulSoup object, then extract data from the tables inside
- GitHub URL: https://github.com/limabielefeld/coursera_homework/blob/8e5640468a5a13dba316ba4f07d9c40c6be65bbe/Applied%20Data%20Science%20Capstone/2.jupyter-labs-webscraping.ipynb



Data Wrangling

- Missing data are fixed by filling in average values
- Landing outcome labels are created using outcome data
- GitHub
  URL: https://github.com/limabielefeld/coursera_homework/blob/8e5640468a5a13dba316ba4f07d9c40c6be65bbe/Applied%20Data%20Science%20Capstone/3.IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

- The charts plotted are:
  - Scatter point plot, to visualize the relations: FlightNumber vs. PayloadMass, PayloadMass vs. LaunchSite, etc.
  - Bar chart, to visualize success rate of each orbit
  - Line chart, to visualize success rate of each year
- GitHub URL: https://github.com/limabielefeld/coursera_homework/blob/8e5640468a5a13dba316ba4f07d9c40c6be65bbe/Applied%20Data%20Science%20Capstone/5.IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- SQL queries performed:
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string '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 succesful landing outcome in ground pad was acheived.
  - 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

EDA with SQL

- SQL queries performed (continued):
  - List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
  - List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
  - Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.
- GitHub URL: https://github.com/limabielefeld/coursera_homework/blob/8e5640468a5a13dba316ba4f07d9c40c6be65bbe/Applied%20Data%20Science%20Capstone/4.jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

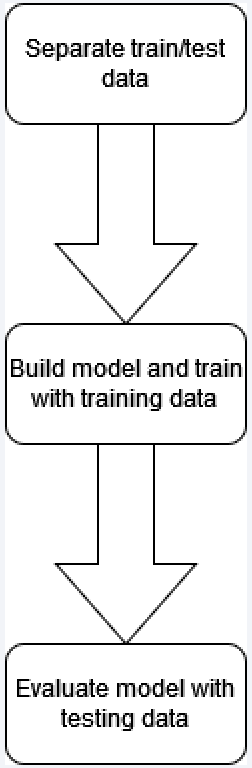
- Objects added to a folium map:
  - Markers for locations such as launch sites
  - Circles to highlight an area
  - Lines to mark distance between two points
- GitHub URL: https://github.com/limabielefeld/coursera_homework/blob/8e5640468a5a13dba316ba4f07d9c40c6be65bbe/Applied%20Data%20Science%20Capstone/6.IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- Plots/graphs and interactions added to a dashboard:
  - Launch success rate per site
  - Payload range
- These are added so that one can easily interact with the dashboard to visualize the relations among different factors
- GitHub URL: https://github.com/limabielefeld/coursera_homework/blob/8e5640468a5a13dba316ba4f07d9c40c6be65bbe/Applied%20Data%20Science%20Capstone/7.spacex_dash_app.py

Predictive Analysis (Classification)

- We built logistic regression model, support vector machine, decision tree and k nearest neighbors models.
- We train the models using training data and then test their performances using test data
- GitHub URL: https://github.com/limabielefeld/coursera_homework/blob/8e5640468a5a13dba316ba4f07d9c40c6be65bbe/Applied%20Data%20Science%20Capstone/8.IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb



Separate train/test data

Build model and train with training data

Evaluate model with testing data

Results

- Exploratory data analysis results:
  - SpaceX uses four launch sites, the average payload of F9 v1.1 booster is 2,928 kg, the F9 v1.1 B1012 and F9 v1.1 B1015 boosters failed in 2015
- Interactive analytics demo in screenshots
- Predictive analysis results
  - After selecting the best hyperparameters for the decision tree classifier using the validation data, it achieves 83.3% accuracy on the test data

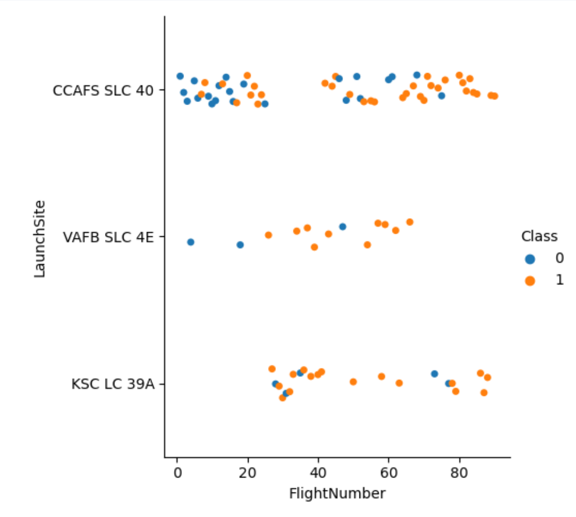
P
P
P
7
AFS
SLC-
40
26
AFS
LC-
40
0.90
KM
Centaur Road
Samuel C Phillips Pkwy

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- A scatter plot of Flight Number vs. Launch Site

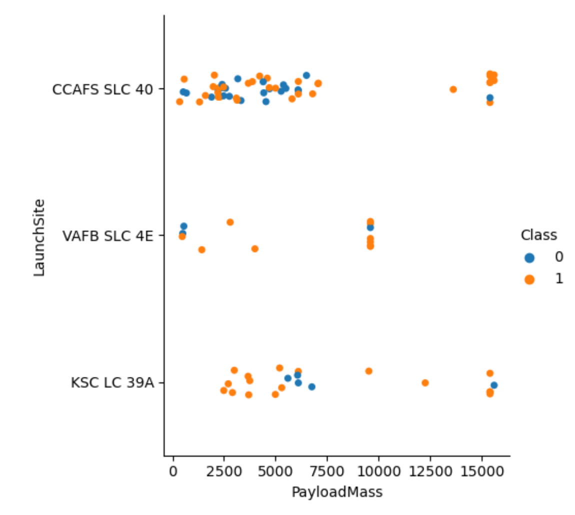


- Explanations:
    - CCAFS SLC 40 has most Class 1 Launches in higher Flight Numbers
    - KSC LC 39A has the highest percentage of Class 1 Launches

Payload vs. Launch Site

- A scatter plot of Payload vs. Launch Site

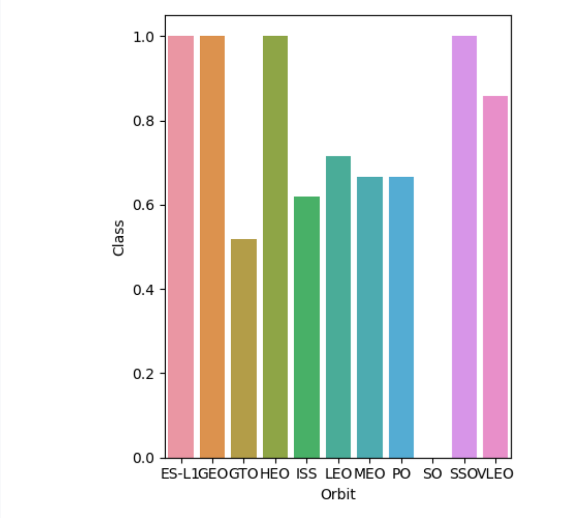
- Explanations:
  - KSC LC 39A has most Class 1 Launches with low Payload Mass
  - VAFB SLC 4E doesn't have very high Payload Mass Launches



Success Rate vs. Orbit Type

- A bar chart for the success rate of each orbit type

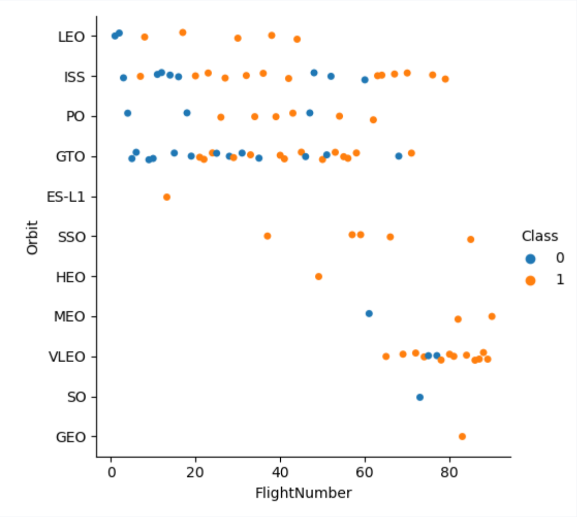
- Explanations:
  - The orbit types with high success rates are: ES-L1, GEO, HEO, SSOV, LEO
  - The orbit types with low success rates are: GTO, ISS, LEO, MEO, PO, SO



Flight Number vs. Orbit Type

- A scatter point of Flight number vs. Orbit type

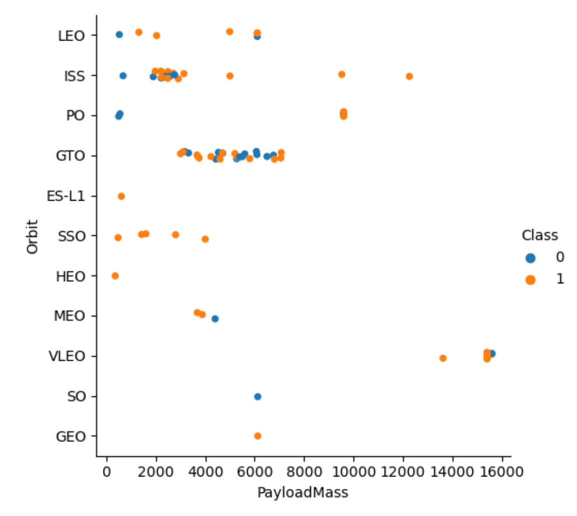
- Explanations:
  - For each Orbit type, Class 1 Launches become more frequent in higher Flight Numbers
  - VLEO has the most Class 1 Launches in higher Flight Numbers



Payload vs. Orbit Type

- A scatter point of payload vs. orbit type

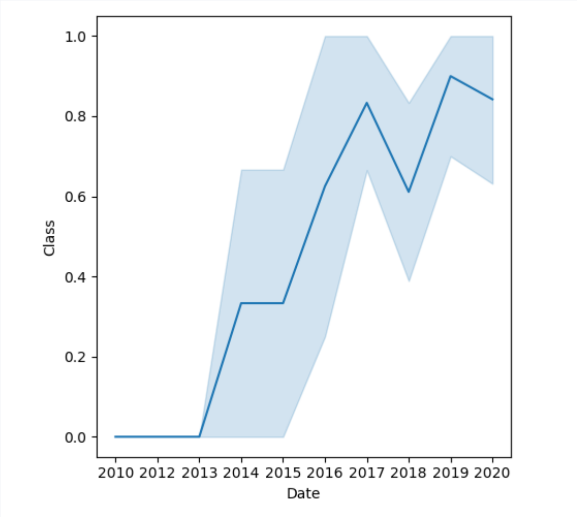
- Explanations:
  - Only VLEO has very high Payload Mass
  - Class 0 Launches happen more frequently with low Payload Mass



Launch Success Yearly Trend

- A line chart of yearly average success rate

- Explanations:
  - In general, success rate increases with respect to time
  - There is a big decrease near the year 2018

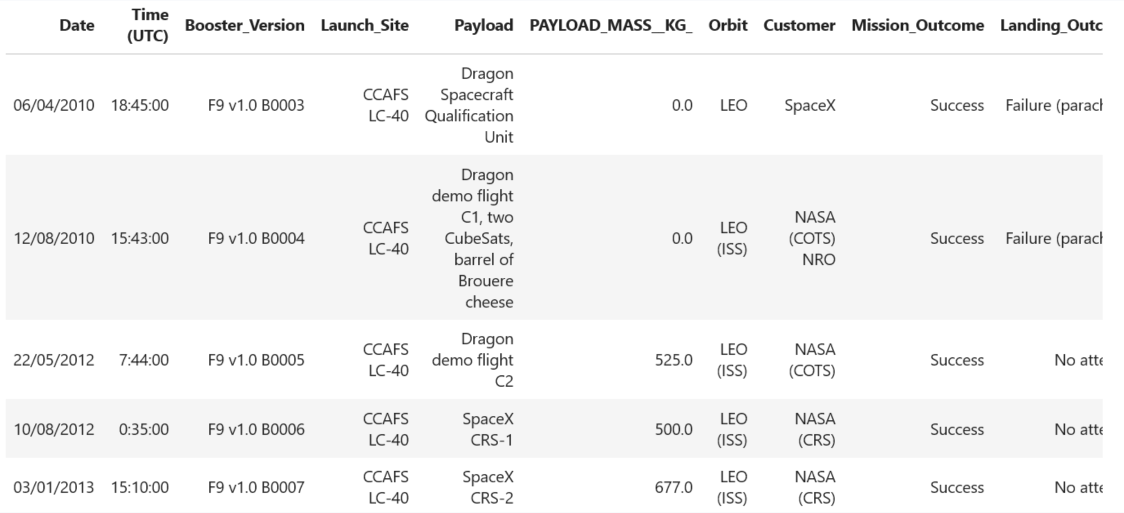


All Launch Site Names

- The names of the unique launch sites
  - CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40
- Explanation:
  - select unique launch sites from the data

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`
- Explanation:
  - Select launch sites that matches pattern 'CCA%'
  - Limit to 5 results



Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outc
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parach
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parach
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No atte
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No atte
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No atte

Total Payload Mass

- The total payload carried by boosters from NASA
    - 45596.0
- Explanation:
    - Calculate sum of payload mass where customer is NASA (CRS)

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1
  - 2928.4
- Explanation:
  - Calculate the average payload mass where booster version is F9 v1.1

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad
  - 01/08/2018
- Explanation:
  - Find the min of date where the outcome is `'Success (ground pad)'`

Successful Drone Ship Landing with Payload between 4000 and 6000

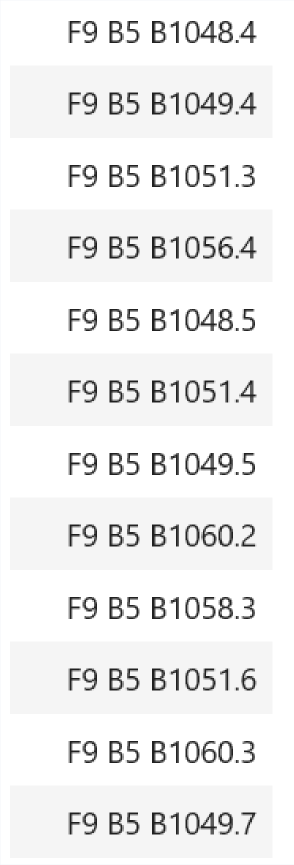
- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
  - F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2
- Explanation:
  - Select boosters where payload mass > 4000 and < 6000 and landing outcome is 'Success (drone ship)'

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes
  - Success: 100
  - Failure: 1
- Explanation:
  - Group mission outcomes that match 'Success'

Boosters Carried Maximum Payload

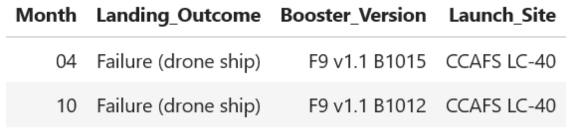
- The names of the booster which have carried the maximum payload mass
- Explanation:
  - Select distinct names of boosters where payload mass equals maximum payload mass (calculated from a subquery)



```
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



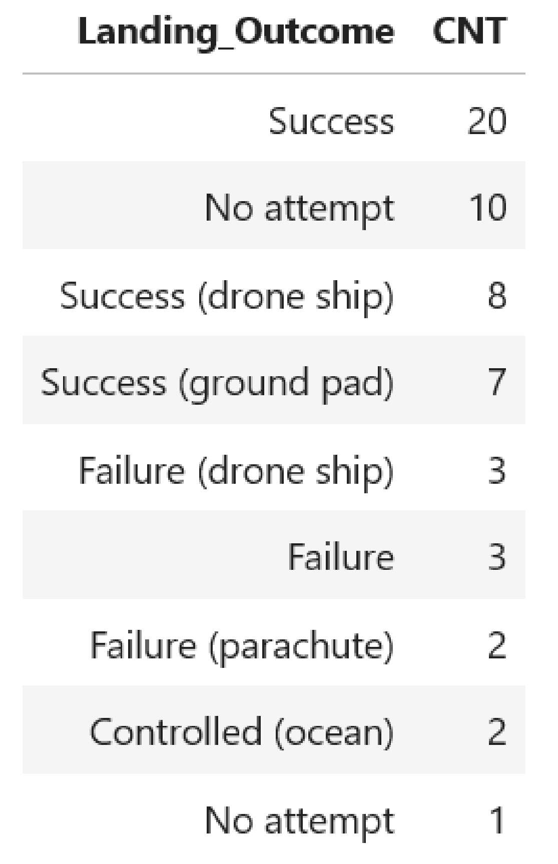
Month	Landing_Outcome	Booster_Version	Launch_Site
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40

- Explanation:
    - Select with landing outcome 'Failure (drone ship)' and year is 2015

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

- Explanation:
  - Group by landing outcomes and sort by their count in descending order

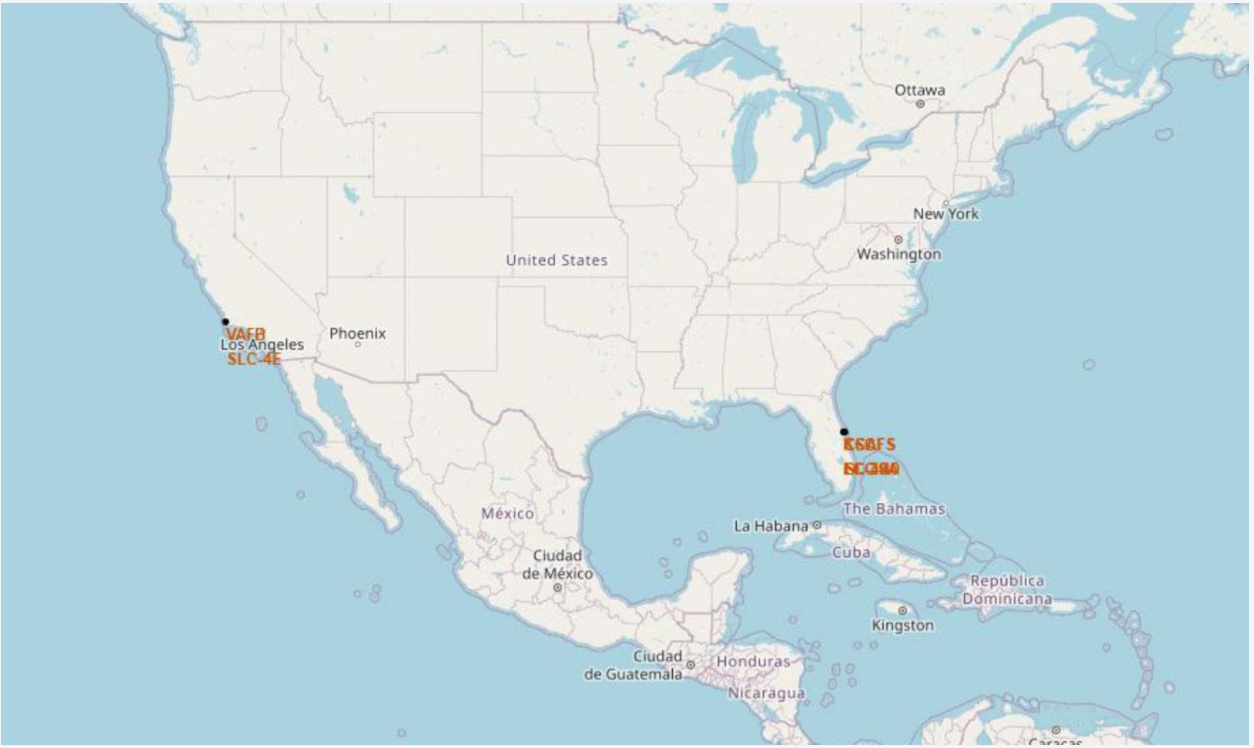


Landing_Outcome	CNT
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

Section 3

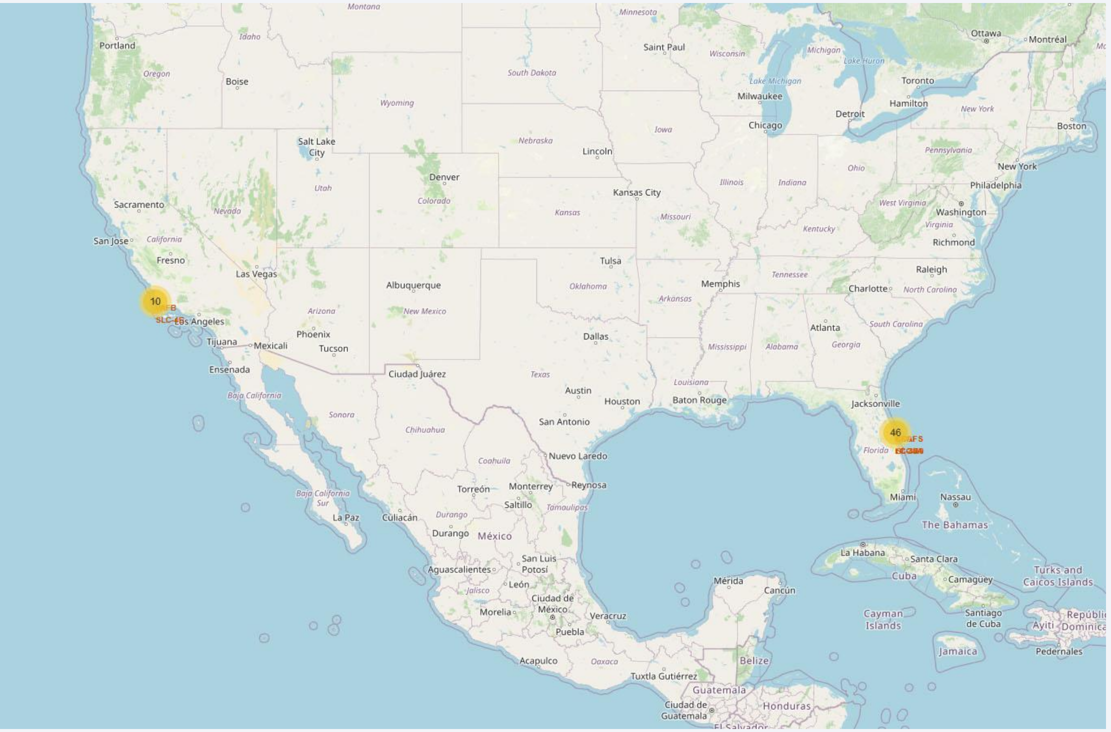
Launch Sites Proximities Analysis

All Launch Sites



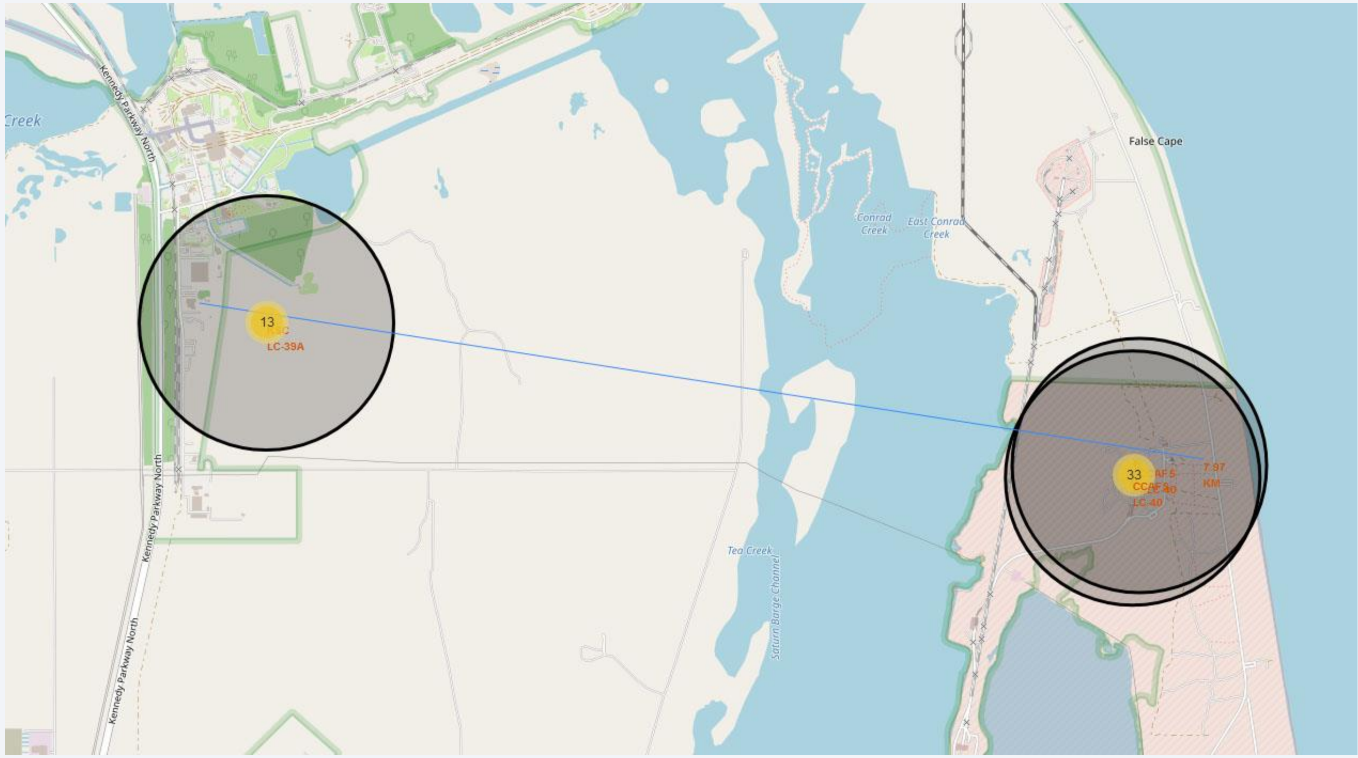
- The launch sites are near the two coast lines.

Launch Outcomes



- The markers indicate success or not

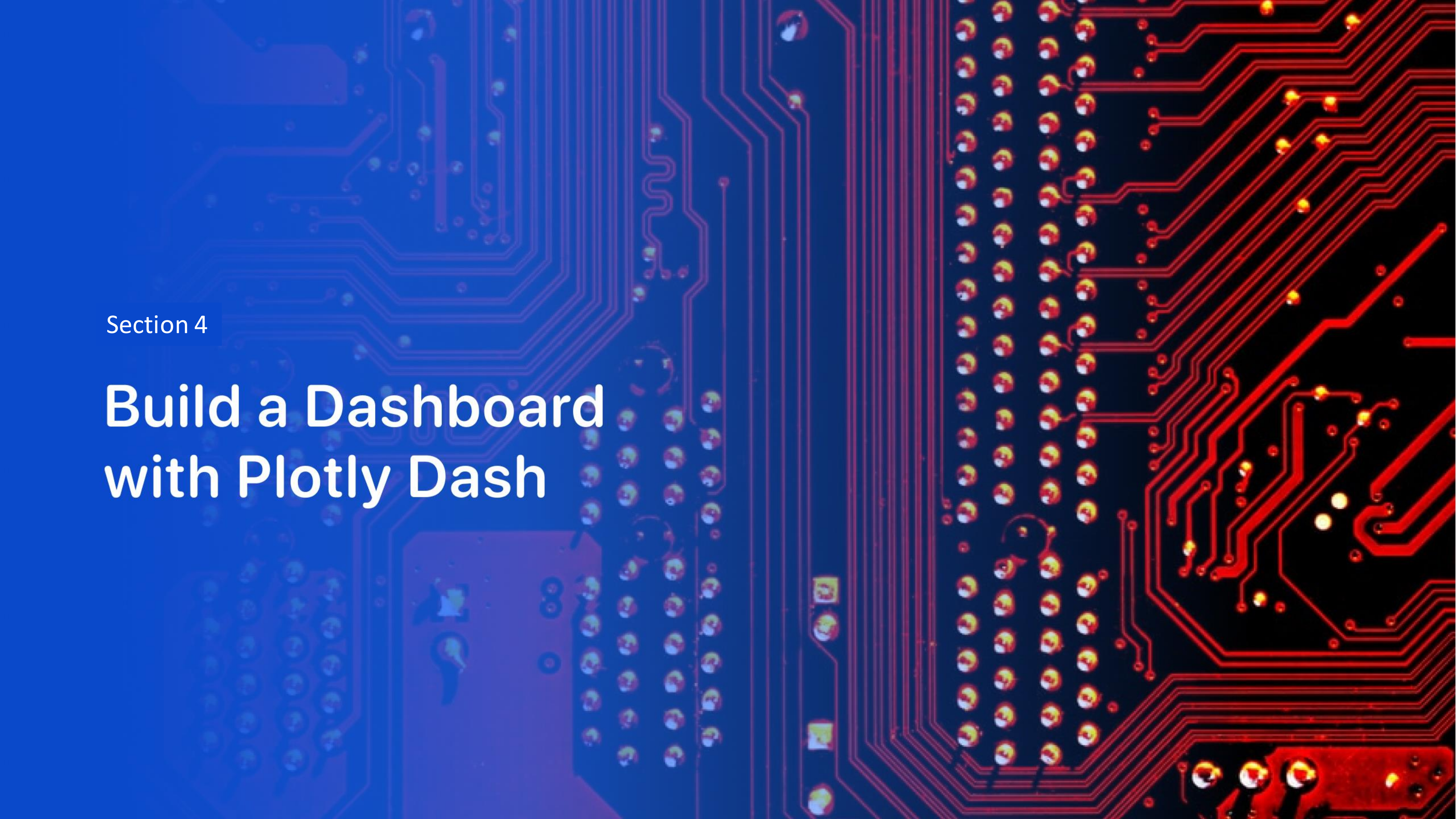
Distance to Coast Line



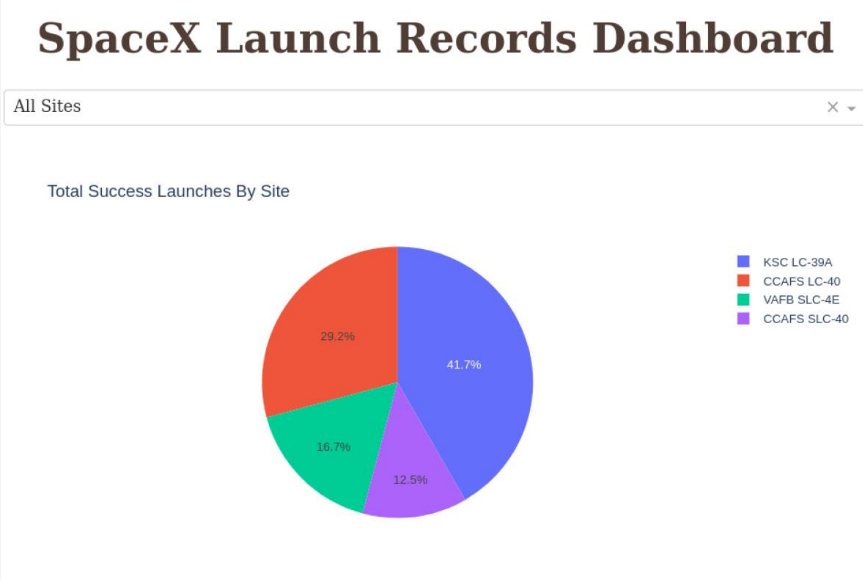
- The Launch site is 7.97 KM from coast line

Section 4

Build a Dashboard with Plotly Dash



Distribution of Success Launches

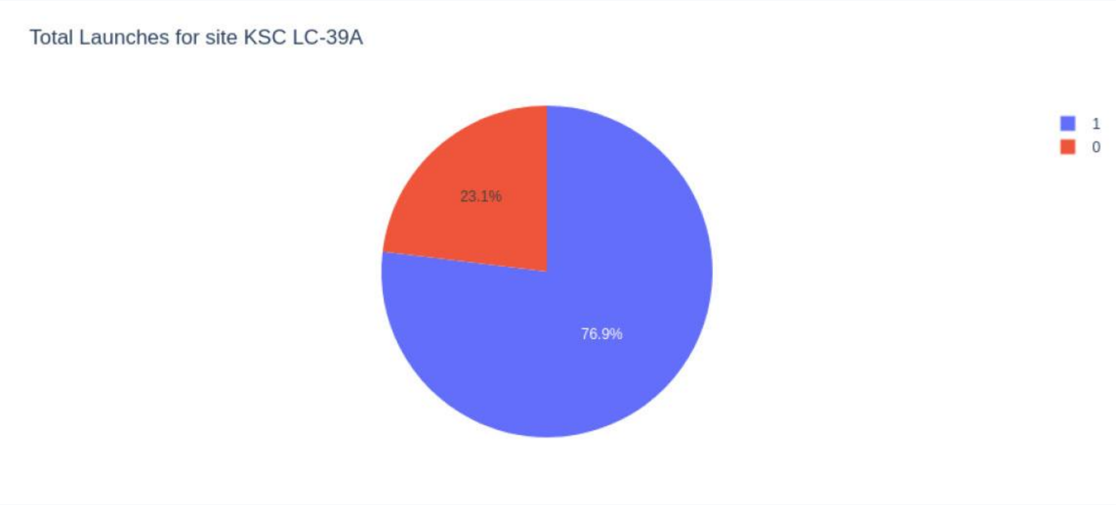


SpaceX Launch Records Dashboard

All Sites

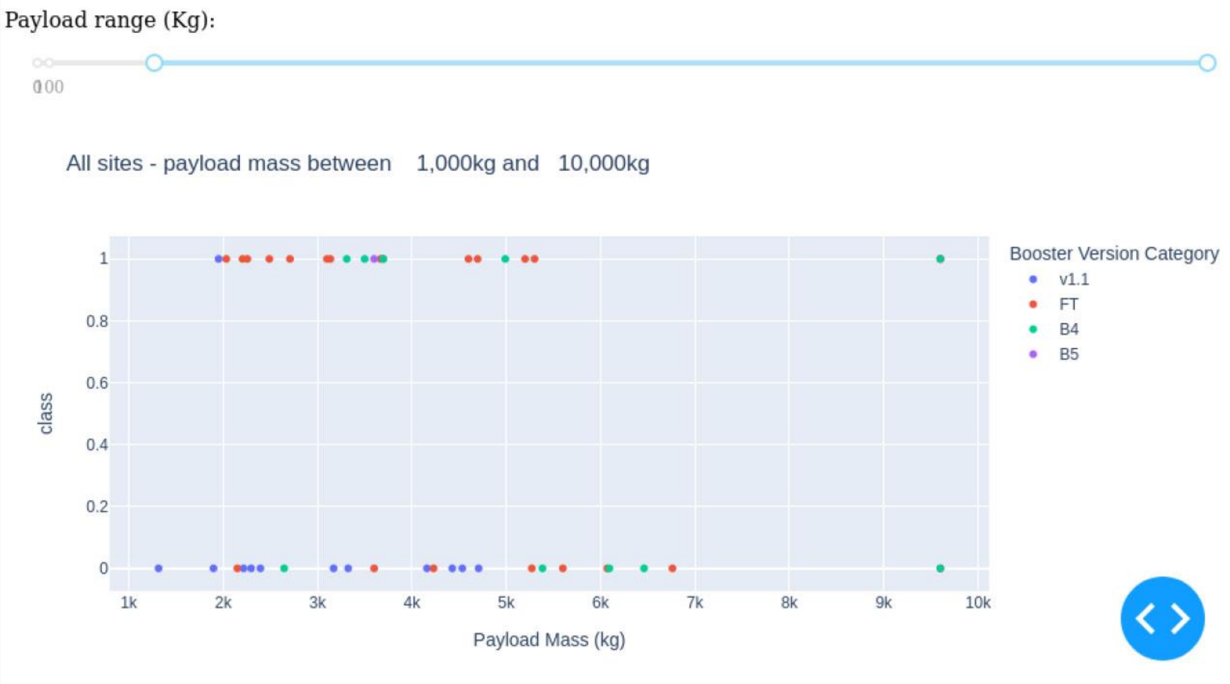
- The chart shows the distributions of success launches in the four sites

Success Rate for One Site



- The chart shows the success rate on the site KSC LC-39A

<Dashboard Screenshot 3>



- Most Payloads are under 7k.

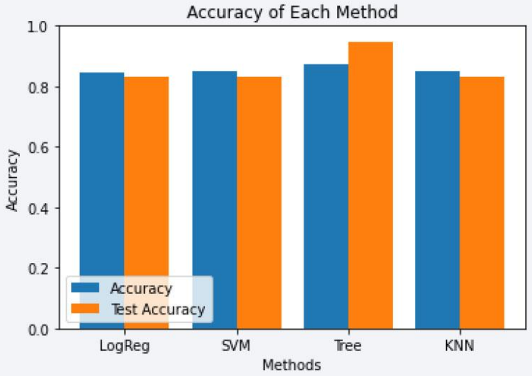
Section 5

Predictive Analysis (Classification)

Classification Accuracy

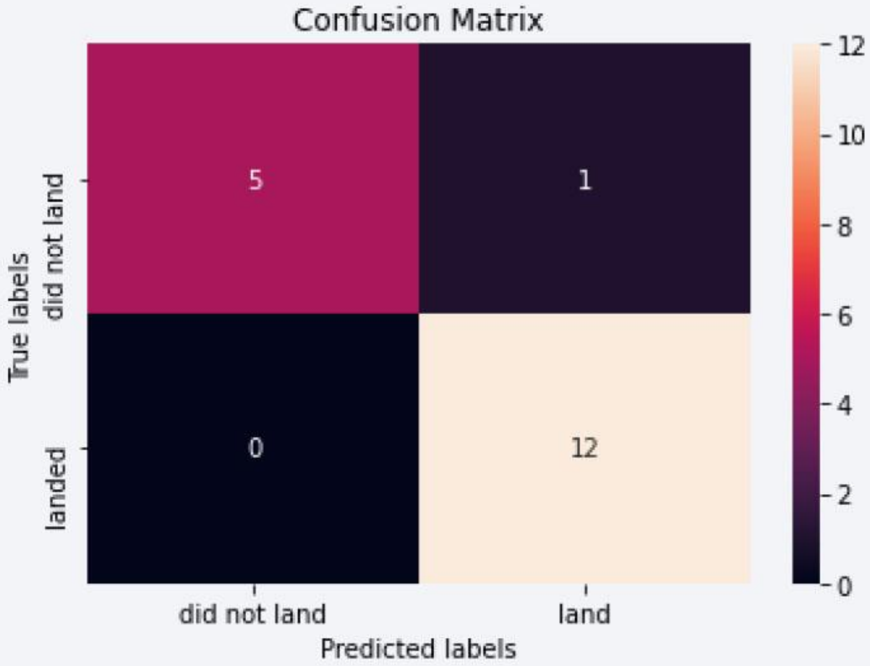
- Visualize the built model accuracy for all built classification models, in a bar chart

- The decision tree model has the highest classification accuracy



Confusion Matrix

- Explanation:
  - The model has no false negative prediction but one false positive prediction



Conclusions

- We have collected data from different sources
- We have wrangled the collected data to suit our usage
- We have used explorative data analysis to find out which factors are important
- We have used interactive visual analytics to explore our data
- We have trained several models and found that the decision tree model predicts landing the best

Appendix

- The source files are available on GitHub: https://github.com/limabielefeld/coursera_homework/tree/8e5640468a5a13dba316ba4f07d9c40c6be65bbe/Applied%20Data%20Science%20Capstone

Thank you!