IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Dhruba Jyoti Seal
August 07, 2022

Outline

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

Executive Summary

- Methodologies
  - Data collection: Collecting the data from API using JSON and GET request
  - Perform data wrangling
  - Perform exploratory data analysis (EDA) using visualization and SQL
  - Perform interactive visual analytics using Folium and Plotly Dash
  - Perform predictive analysis using classification models
- Summary of all results
  - Some conclusions can be drawn from EDA and Plotly Dash
  - It does not matter which model is used.
  - The size of the test data is very small. In order to make a better statement about which model fits best, a larger test set must be provided.

Introduction

- SpaceX has gained worldwide attention for a series of historic milestones. It is the only private company ever to return a spacecraft from low-earth orbit, which it first accomplished in December 2010. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars wheras other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage.
- Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch. This dataset includes a record for each payload carried during a SpaceX mission into outer space.

Section 1

Methodology

Methodology

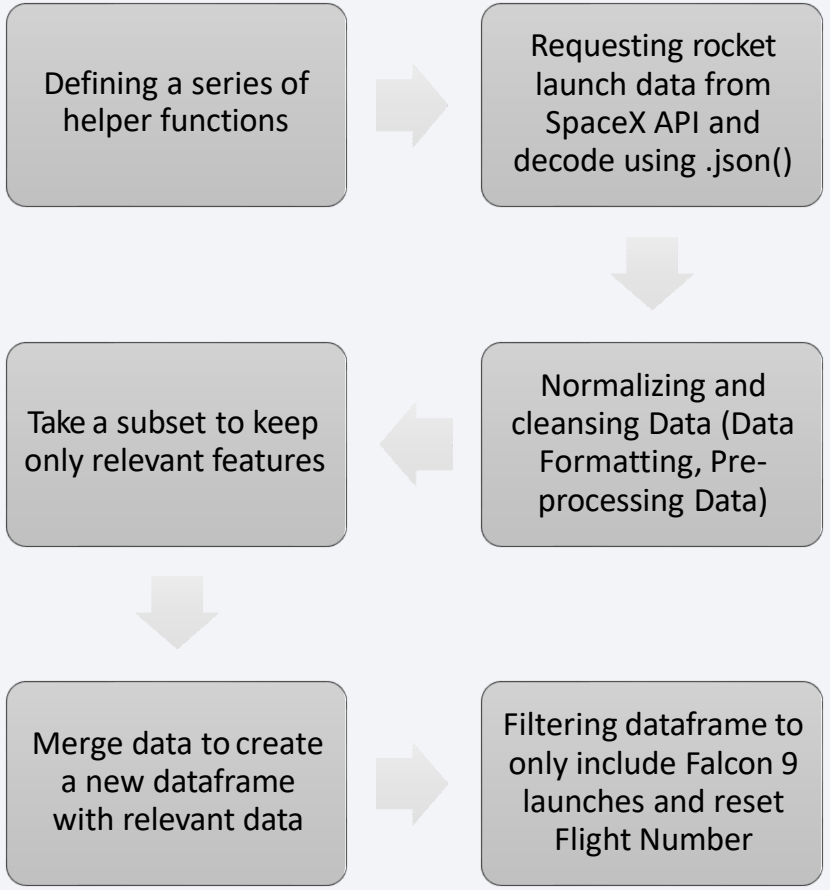
Executive Summary

- Data collection methodology:
  - Collecting the data from API using JSON and GET request
- Perform data wrangling
  - Dealing with missing values and finding some patterns in the data to determine what would be the label for training supervised models
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Standardize, split into training and test data, find best hyperparameters with GridSearchCV

Data Collection – SpaceX API

- Request to the SpaceX API and clean the requested data

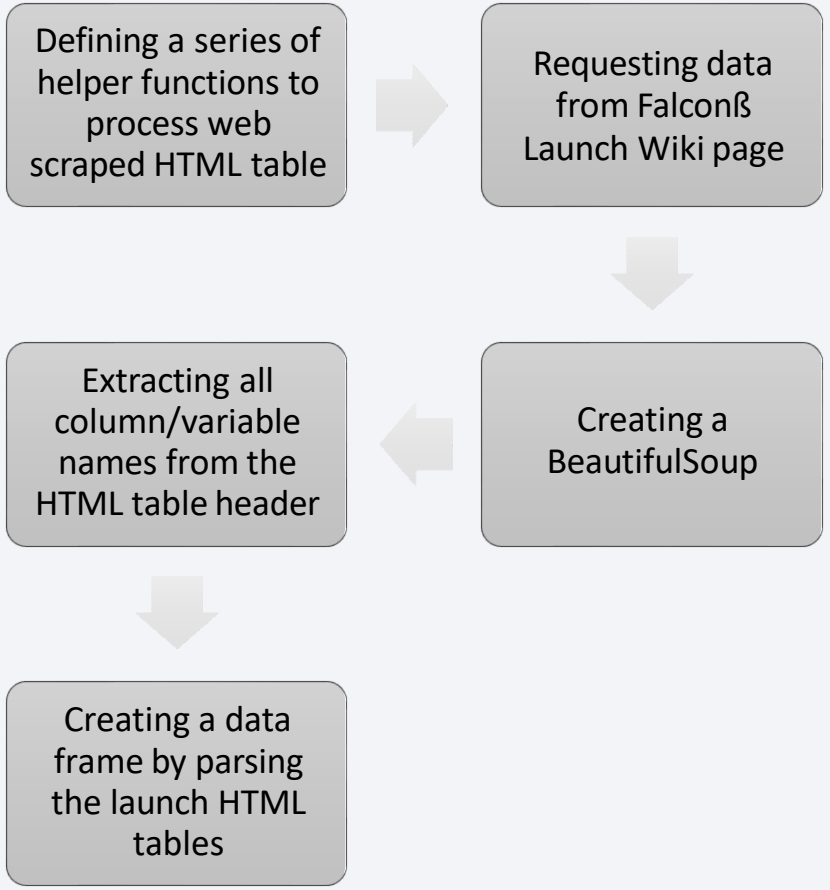
- Notebook for Data Collection
  IBM_Capstone_project/jupyter-labs-spacex-data-collection-api.ipynb at main · LuciLul/IBM_Capstone_project · GitHub



Data Collection - Scraping

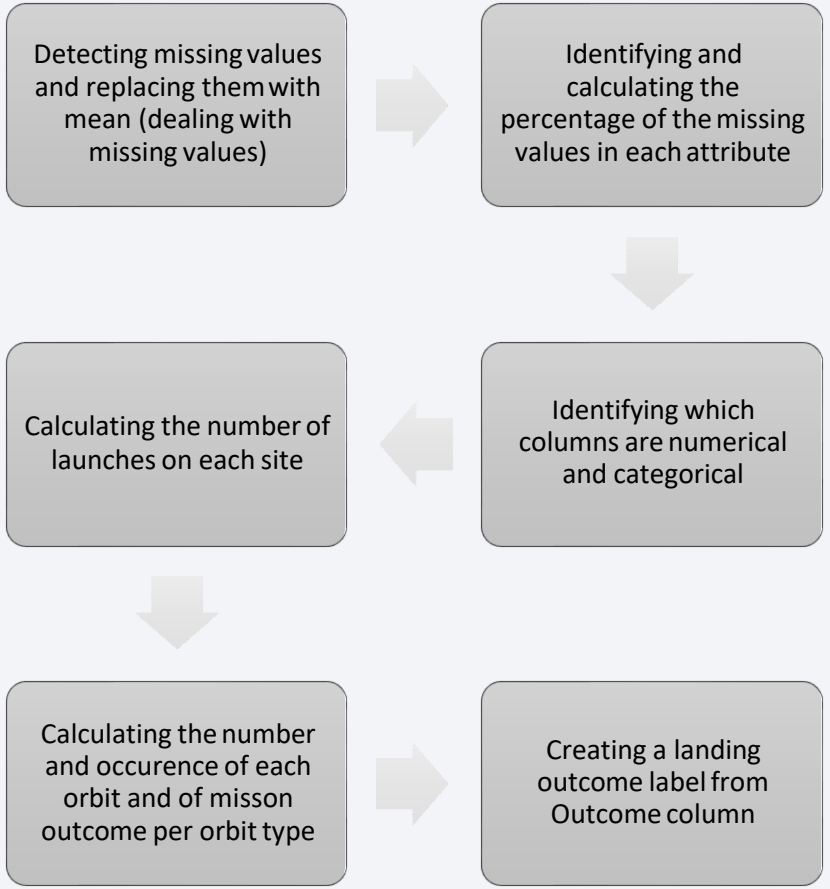
- Perfoming web scraping to collect Falcon 9 historical launch records from a Wikipedia page titled “List of Falcon 9 and Falcon Heavy launches”

- Notebook for Web Scraping: IBM_Capstone_project/IBM_Capstone_project-webscraping.ipynb at main · LuciLul/IBM_Capstone_project · GitHub



Data Wrangling

- Dealing with missing values and finding some patterns in the data to determine what would be the label for training supervised models
- Notebook for Data Wrangling: IBM_Capstone_project/IBM Capstone Project_Data Wrangling.ipynb at main · LuciLul/IBM_Capstone_project · GitHub



EDA with Data Visualization

Exploratory Data Analysis and Preparing Data Feature Engineering

- Scatter plot Flight Number vs. Launch Site
  to visualize the relationship between Flight Number and Launch Site
- Scatter plot Payload vs. Launch Site
  to observe if there is any relationship between launch sites and their payload mass
- Bar plot success rate of each orbit type
  try to find which orbits have high sucess rate and check if relationship between success rate and orbit type
- Scatter plot Payload vs. Orbit type
  to reveal the relationship between Payload and Orbit type
- Scatter plot Flight Number vs. Orbit type
  to reveal the relationship between Flight Number and Orbit type
- Line chart launch success yearly trend
  to get the average launch success trend

- Notebook for EDA with Data Visualization:
  IBM_Capstone_project/IBM_Capstone_project_Exploratory Analysis using pandas and Matpolib.ipynb at main · LuciLul/IBM_Capstone_project (github.com)

EDA with SQL

- SQL queries to understand the Spacex DataSet and to know how to prepare for further actions
  - All Launch Site Names
  - Launch Site Names Begin with 'CCA'
  - Total Payload Mass
  - Average Payload Mass by F9 v1.1
  - First Successful Ground Landing Date
  - Successful Drone Ship Landing with Payload between 4000 and 6000
  - Total Number of Successful and Failure Mission Outcomes
  - Boosters Carried Maximum Payload
  - 2015 Launch Records
  - Rank Landing Outcomes Between 2010-06-04 and 2017-03-20
- Notebook SQL queries:
  IBM_Capstone_project/IBM_Capstone_project_Exploratory Analysis using SQL.ipynb at main · LuciLul/IBM_Capstone_project (github.com)

Build an Interactive Map with Folium

- Created interactive Maps
  - To see how successful the launch sites on the west and east coach are
  - To realize the distances to railways, coast and cities to assess possible dangers

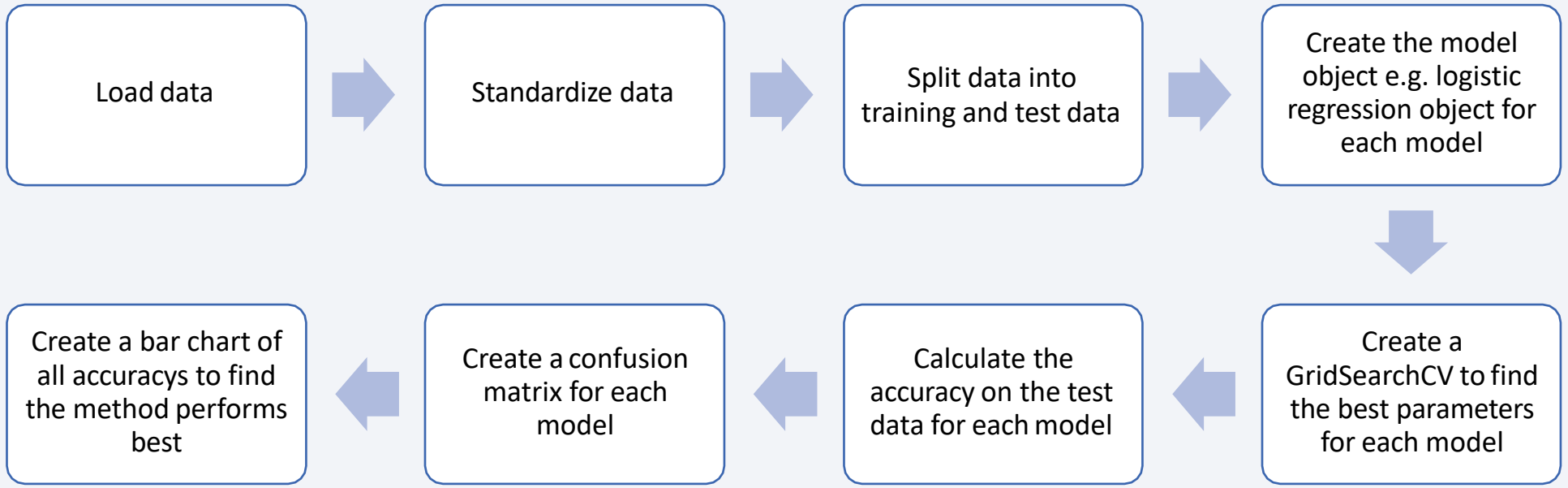
- Notebook Interactive Map with Folium:
  IBM_Capstone_project/IBM_Capstone_project_Interactive Viusal Analytics and Dashboard with Folium.ipynb at main · LuciLul/IBM_Capstone_project · GitHub

Build a Dashboard with Plotly Dash

- plots/graphs
  - Total Success Launches By Site (to see the most successful launch site)
  - Total Success Launches for site KSC LC-39A (most successful launch site, to see the distribution)
  - Correlation between Payload mass and Success for All sites (to see the payload ranges with the highest/lowest success rate and to see the booster version with the highest success rate)
- Interactions
  - Possibility to select the launch site to see the different successful landing quotes
  - Possibility to select a payload mass range to realize the relationship between Payload mass and success
- Plotly Dash lab:
  IBM_Capstone_project/Dash_Spacex_Launch.py at main · LuciLul/IBM_Capstone_project · GitHub

Predictive Analysis (Classification)

- Find best Hyperparameter for SVM, Classification Trees, K Nearest Neigbors and Logistic Regression and find the method performs best using test data



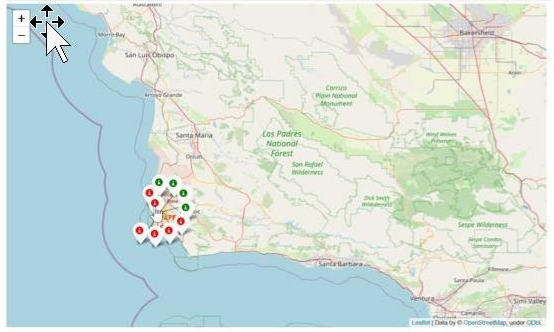
- Notebook Machine Learning Prediction:
  IBM_Capstone_project/IBM_Capstone_project_Machine Learning Prediction.ipynb at main · LuciLul/IBM_Capstone_project · GitHub

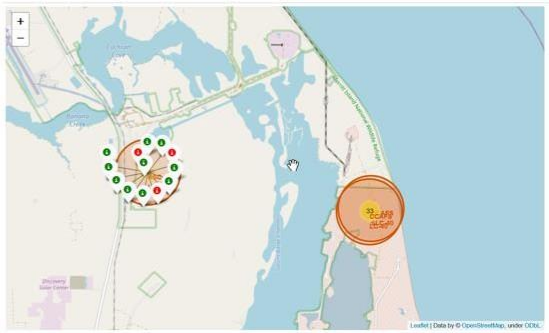
Results

- Exploratory data analysis results
  - As the flight number increases, the first stage is more likely to land successfully.
  - The Launch Site is also important; it seems that CCAFS SLC 40 is the launch site with most successful landings of first stage
  - For the VAFB-SLC launch site there are no rockets launched for heavy payload mass
  - ES-L1, GEO, HEO and SSO are the most successful orbit types
  - The LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.
  - With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS. However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.
  - The sucessrate since 2013 kept increasing till 2020

Results

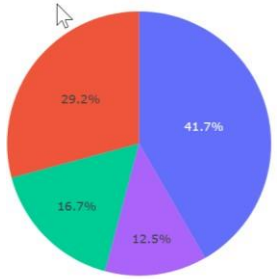
- Interactive analytics demo in screenshots
  - The launch sites are very close to the coasts.

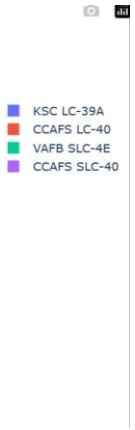


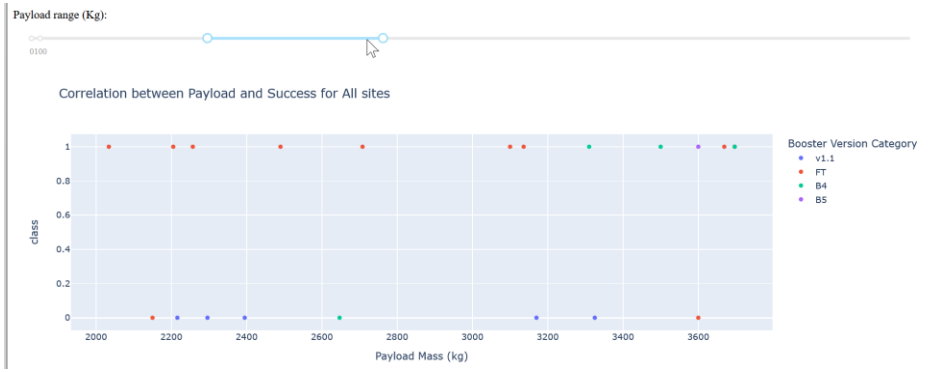


Results

- Launch Site KSC LC 39A has the most successful launches
- The payload range between 2k and 4k has the highest launch success rate.
- The booster version FT has the highest launch success rate.

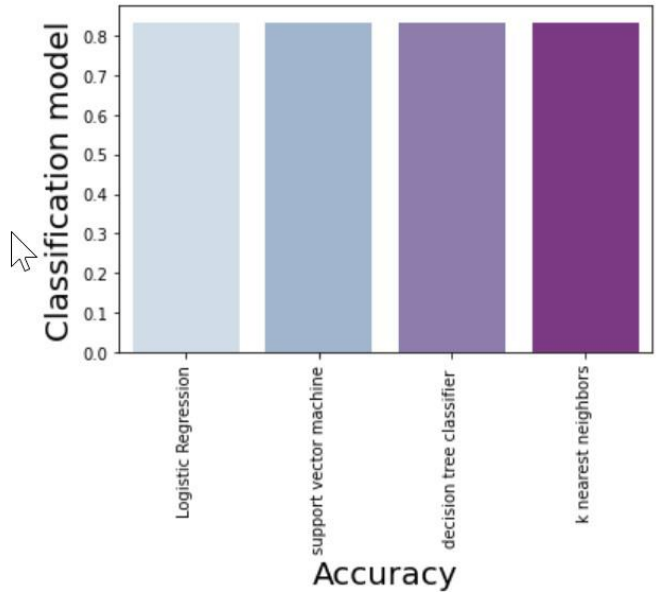






Results

- Predictive analysis results
  - All Models have the same accuracys
  - The size of the test data is very small. In order to make a better statement about which model fits best, a larger test set must be provided.

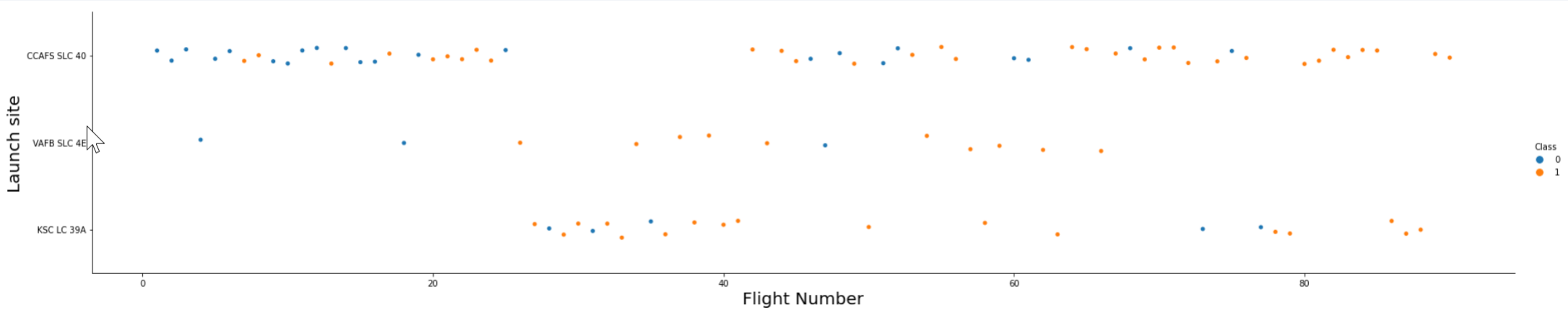


Section 2

Insights drawn from EDA

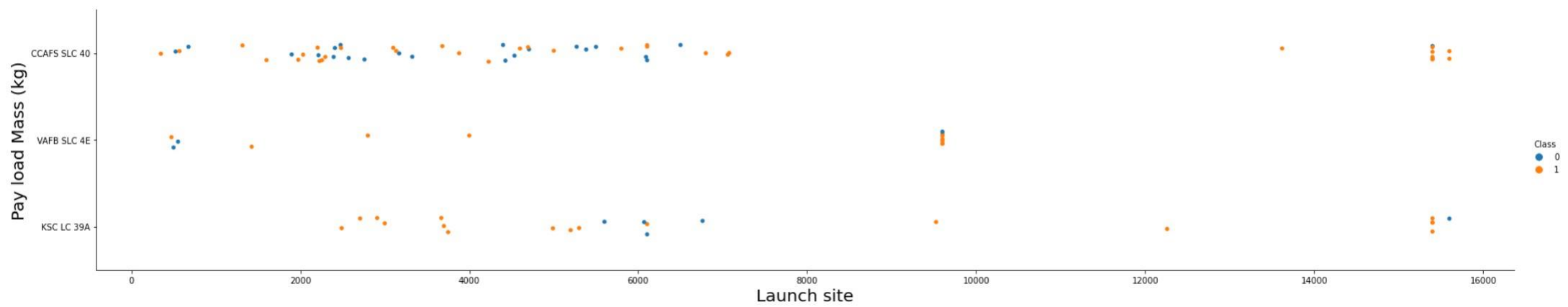
Flight Number vs. Launch Site

- As the flight number increases, the first stage is more likely to land successfully.
- The Launch Site is also important; it seems that CCAFS SLC 40 is the launch site with most successful landings of first stage.



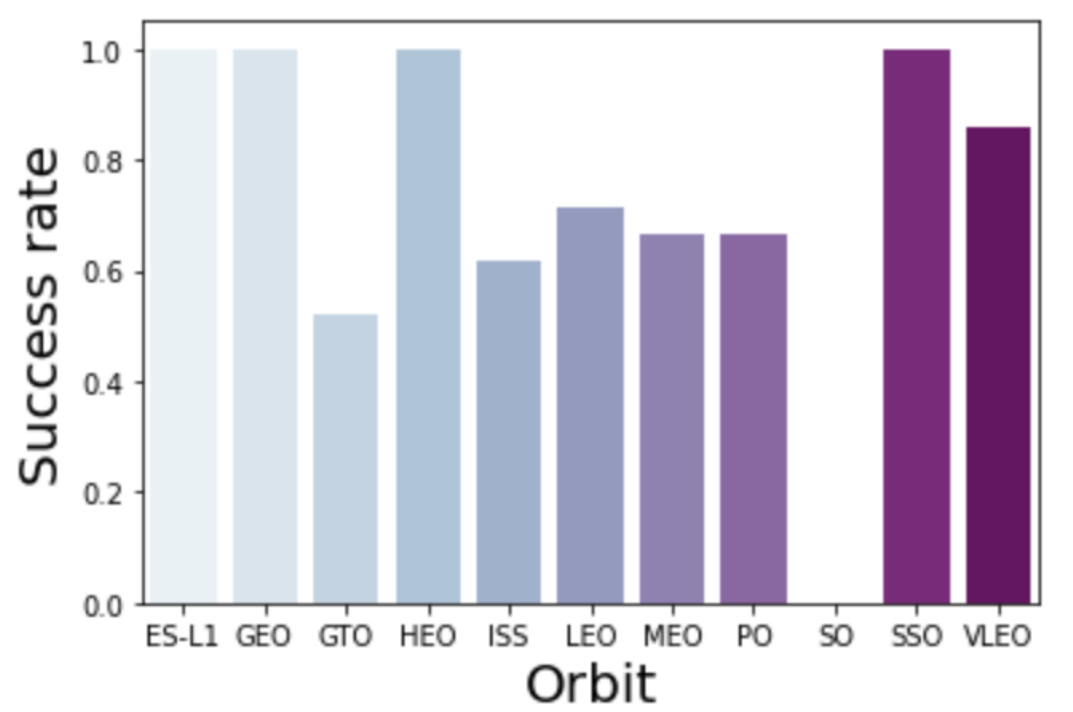
Payload vs. Launch Site

- For the VAFB-SLC launchsite there are no rockets launched for heavypayload mass (greater than 10000).



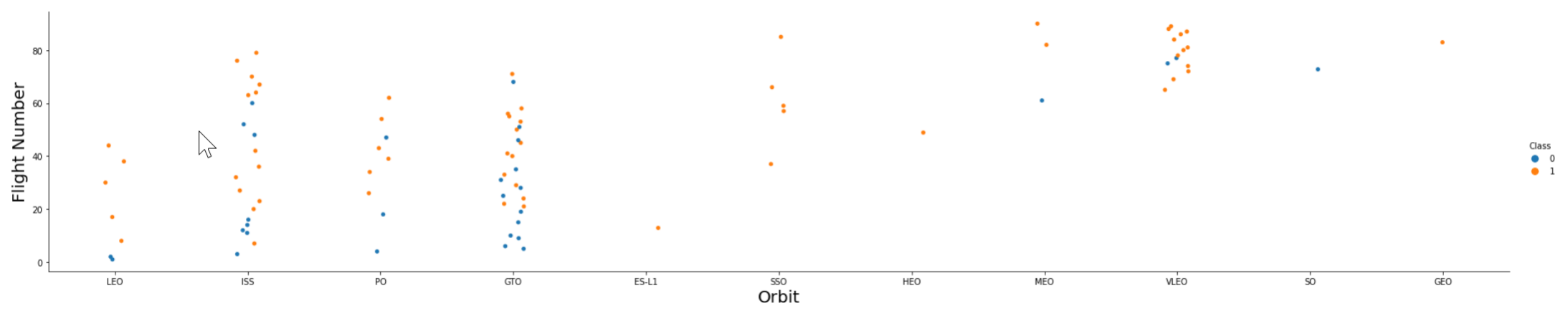
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO and SSO are the most successful orbit types.



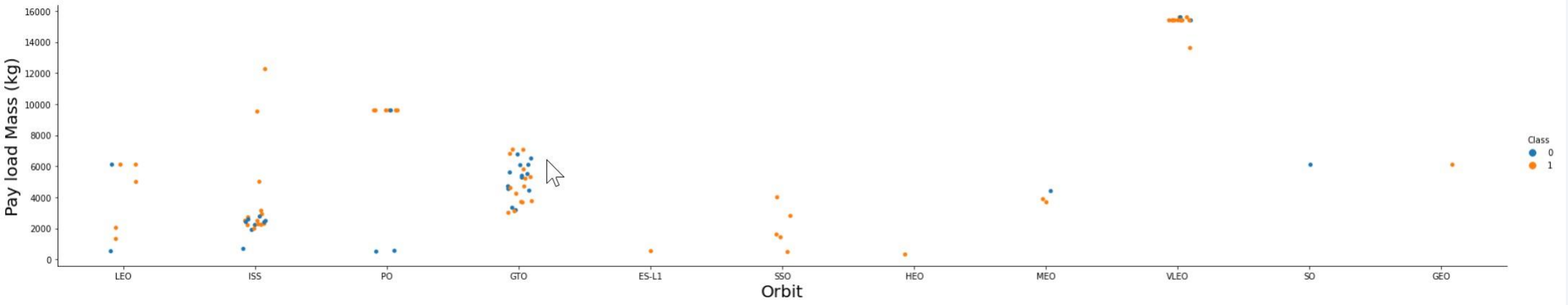
Flight Number vs. Orbit Type

- the LEO orbit the Success appears related to the number of flights;
- on the other hand, there seems to be no relationship between flight number when in GTO orbit.



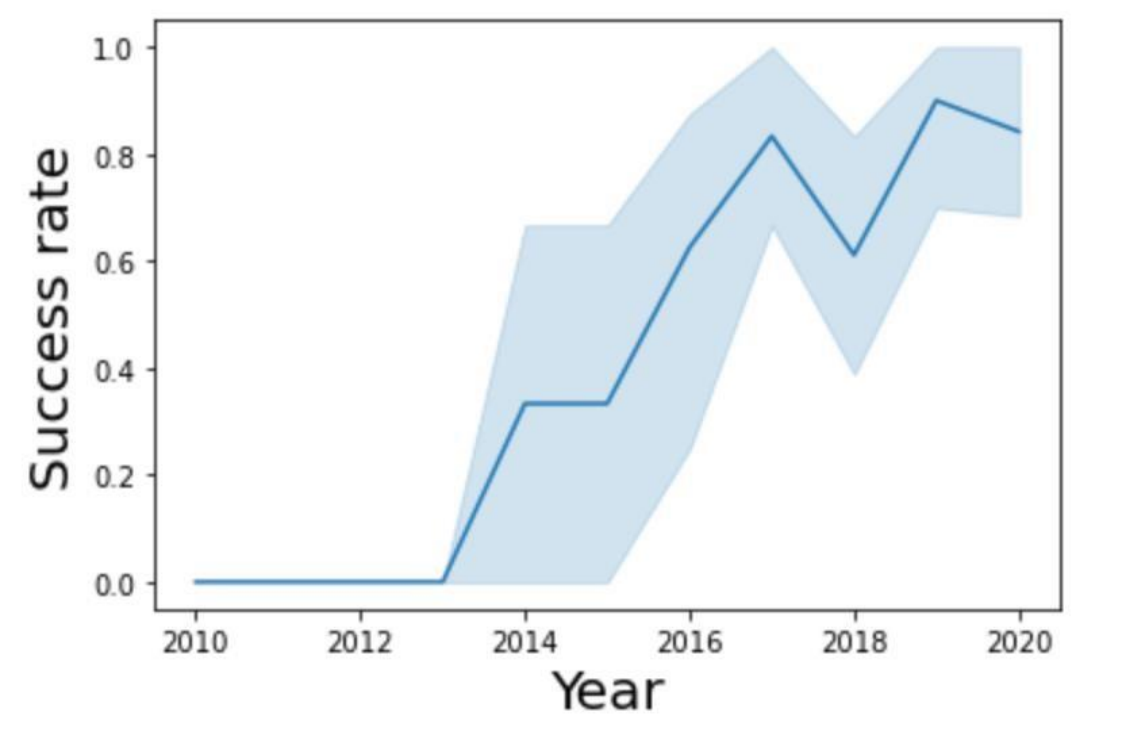
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.



Launch Success Yearly Trend

- the sucess rate since 2013 kept increasing till 2020



All Launch Site Names

- The data contains several Space X launch facilities, the location is placed in the column LaunchSite

LaunchSite	LaunchSite in dataframe
Cape Canaveral AFS Launch Complex 40	CCAFS LC-40
Cape Canaveral Space Launch Complex 40	CCAFS SLC-40
Vandenberg Air Force Base Space Launch Complex 4E	VAFB SLC-4E
Kennedy Space Center Launch Complex 39A	KSC LC-39A

Launch Site Names Begin with 'CCA'

- For example, 5 records where launch sites is like CCAFS LC-40 or CCAFS SLC-40

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	None	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	None	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	None	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	None	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	None	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

Total payload mass carried by boosters from NASA

45.596

Average Payload Mass by F9 v1.1

Average payload mass carried by booster version F9 v1.1

2.928

First Successful Ground Landing Date

First successful landing outcome in ground pad

December 22, 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

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

Total Number of Successful and Failure Mission Outcomes

Most of the missions are successful.

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- 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

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

booster_version	launch_site	landing__outcome
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

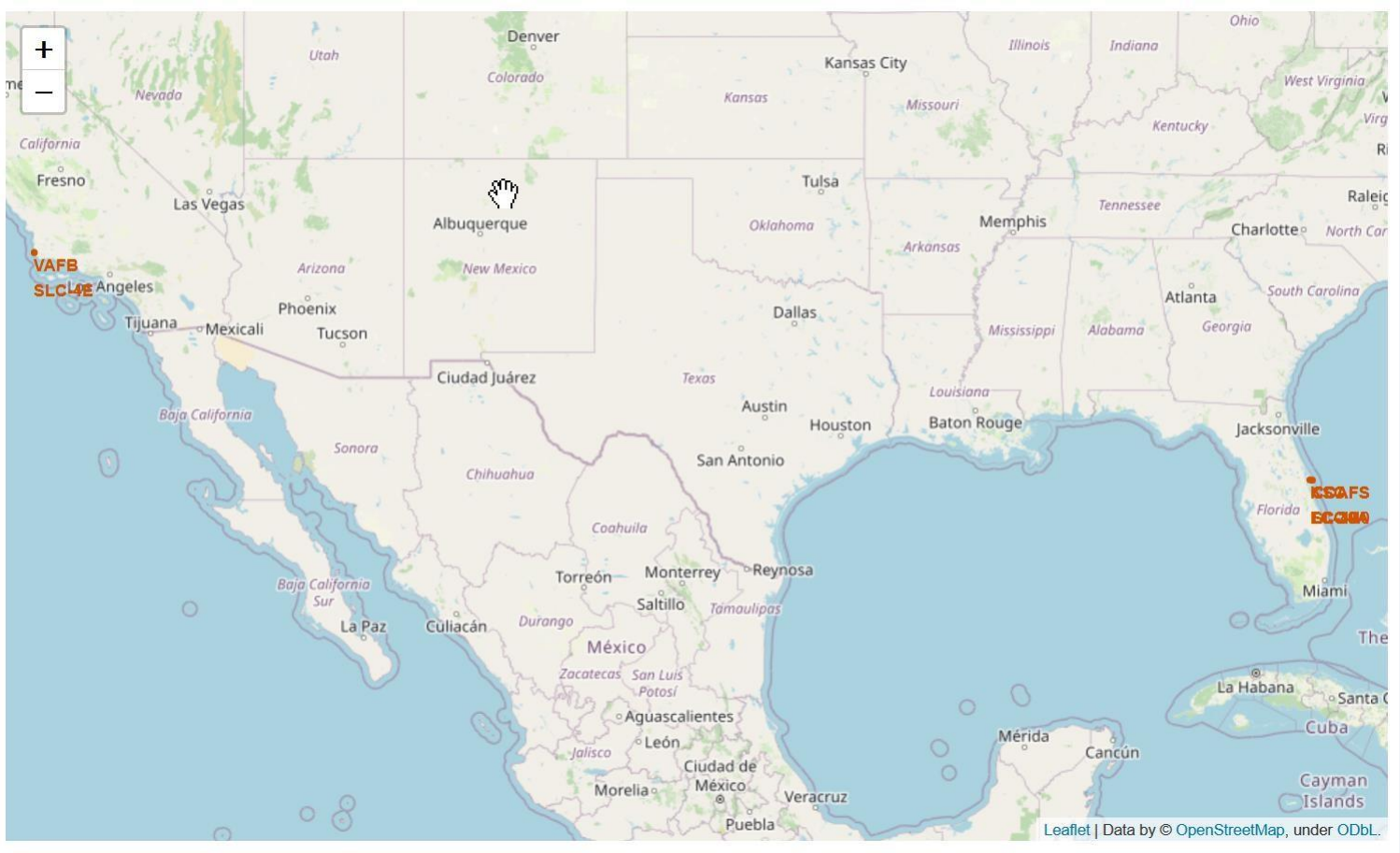
- Count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

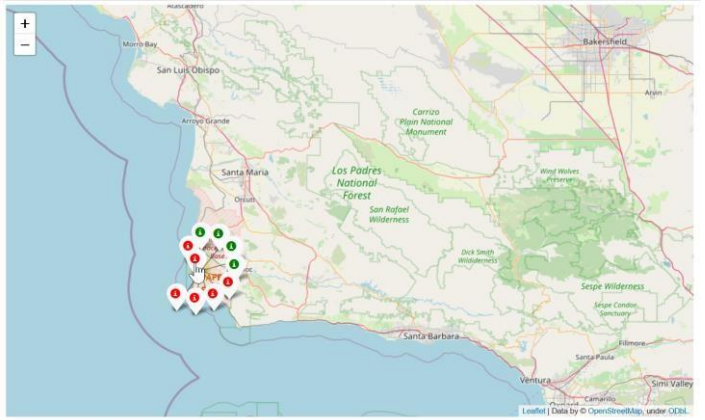
1	landing__outcome
10	No attempt
5	Failure (drone ship)
5	Success (drone ship)
3	Controlled (ocean)
3	Success (ground pad)
2	Failure (parachute)
2	Uncontrolled (ocean)
1	Precluded (drone ship)

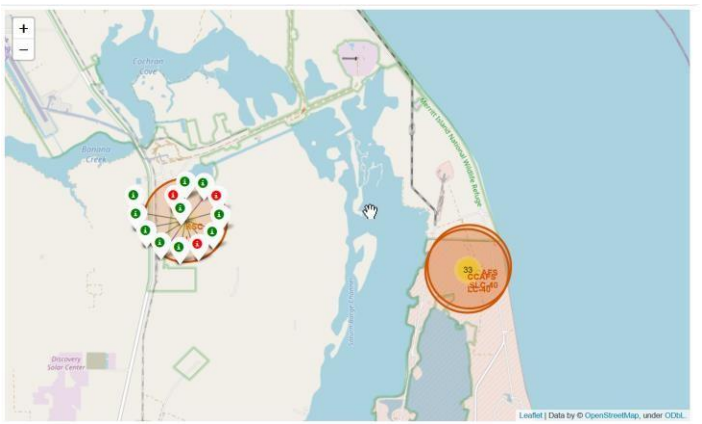

Section 3
Launch Sites
Proximities Analysis

Locations of all launch sites

- Two of the launch sites are located on the west coast and two others on the east coast.





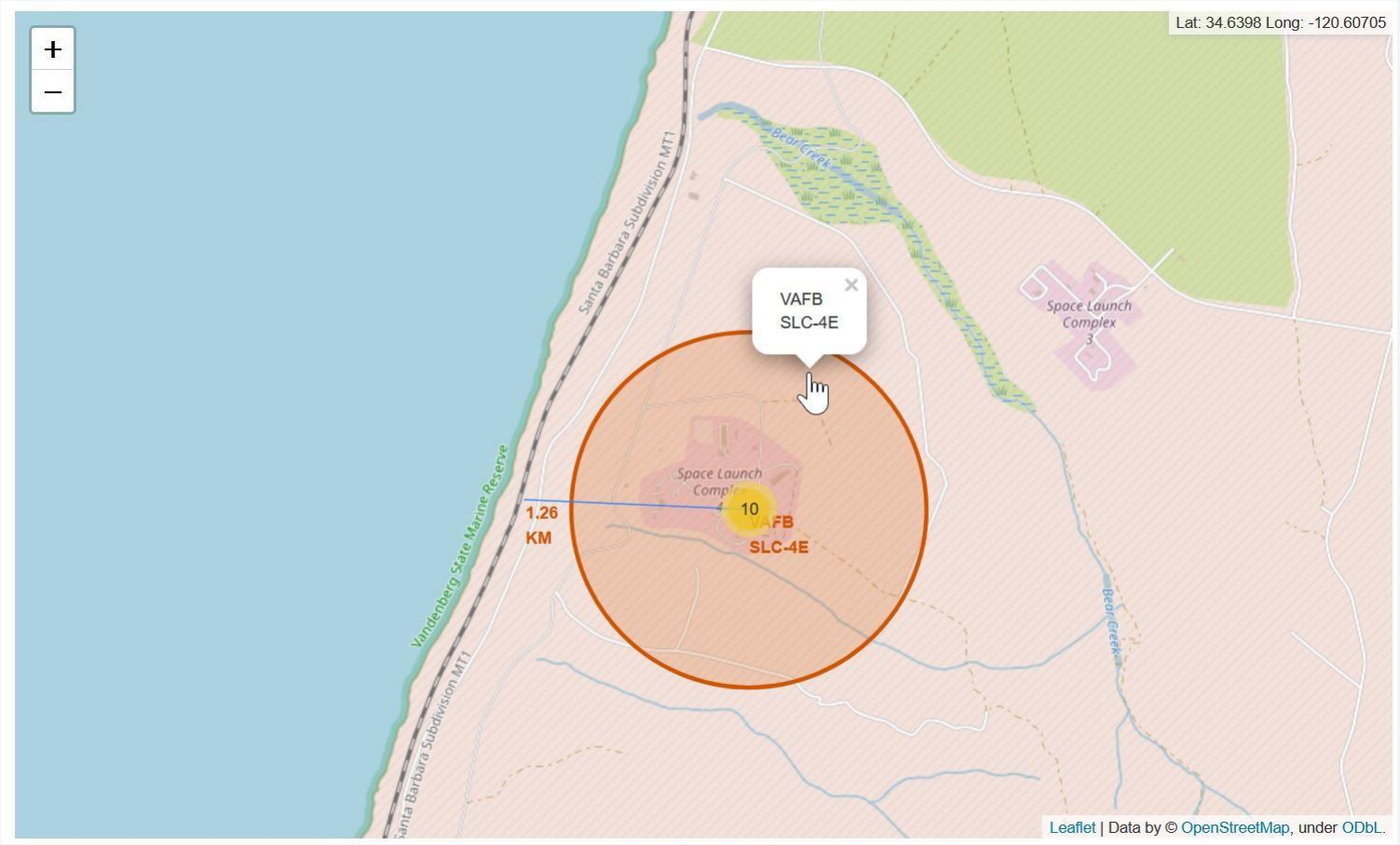


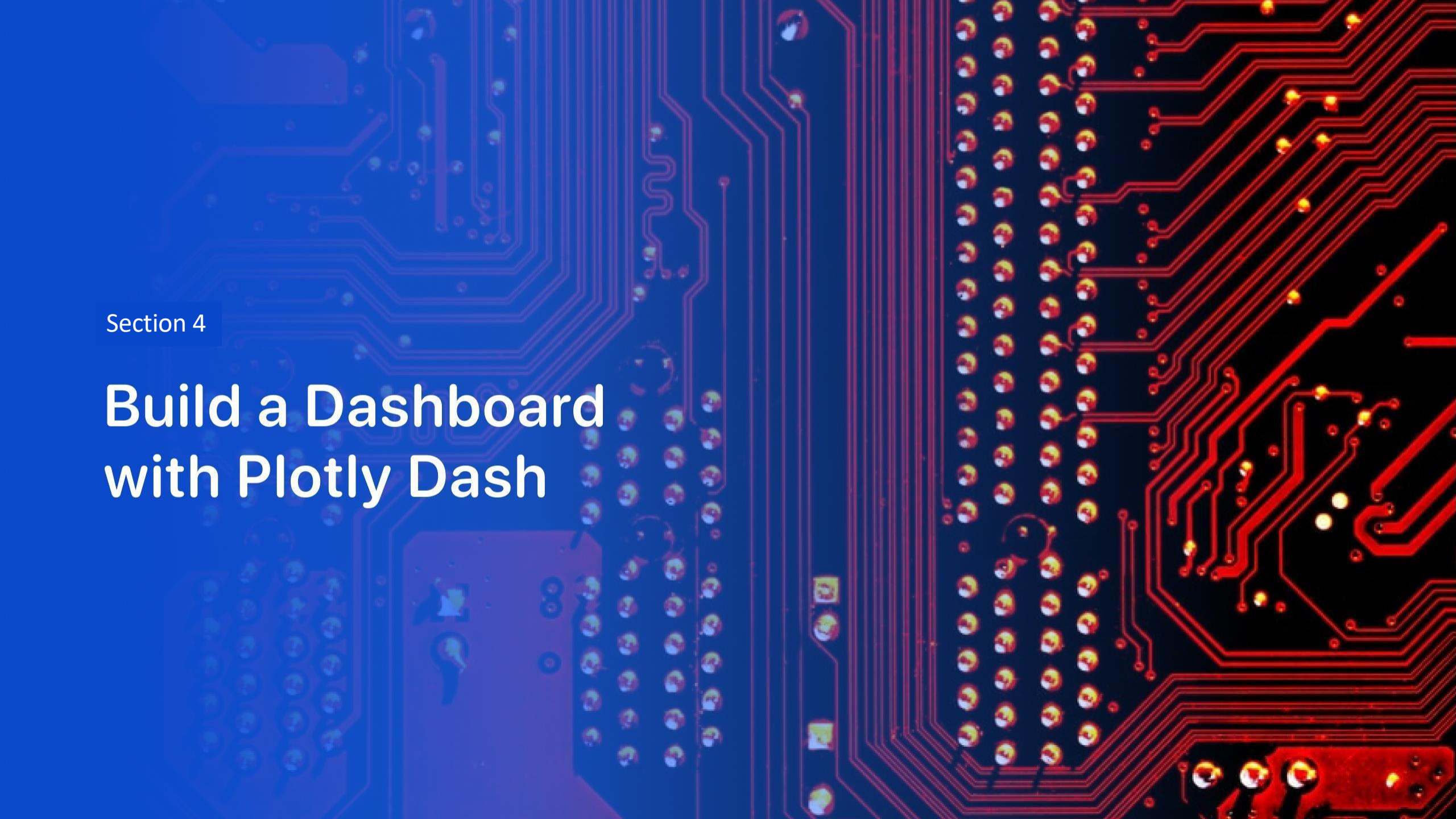
Success and failed launches for each site on the map

- Interactive Map: Click on the launch site to see all launch outcomes
- Green is successful, red not successful => many successful launches on the east coast

Distance between launch site VAFB SLC-4E and railway

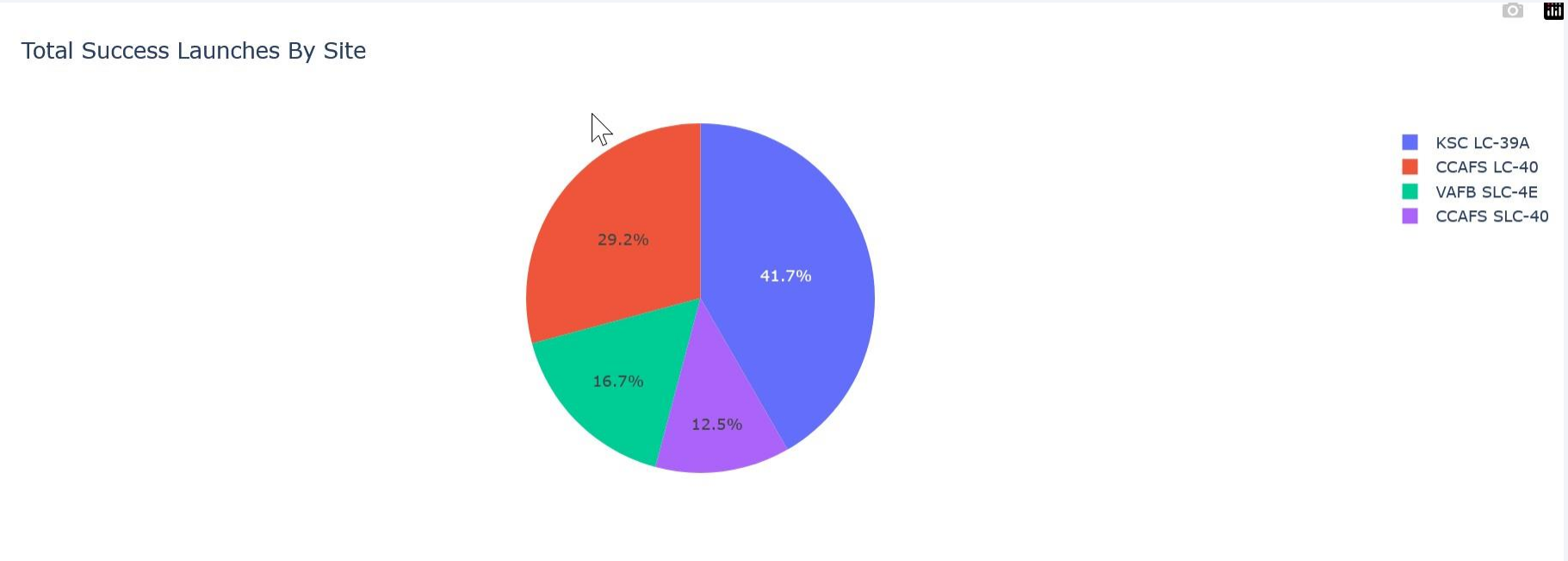
- The distance between launch site VAFB SLC-4E and railway is only 1.26 km




Section 4
Build a Dashboard with Plotly Dash

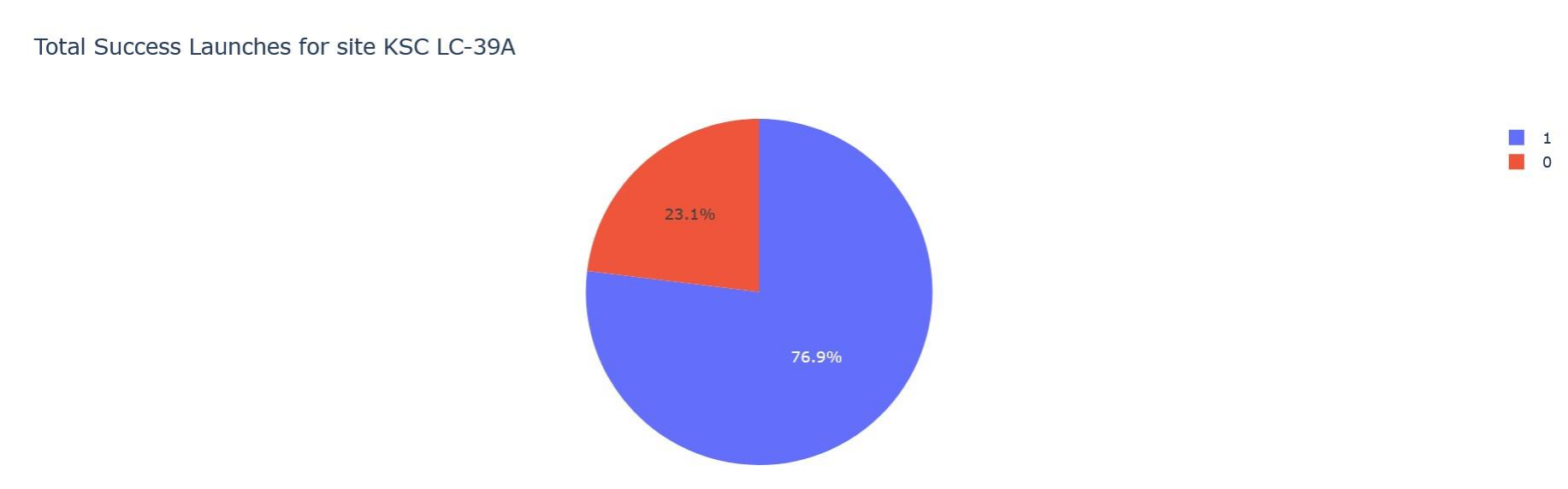
Total Success Launches By Site

- Launch Site KSC LC-39A has the most successful launches



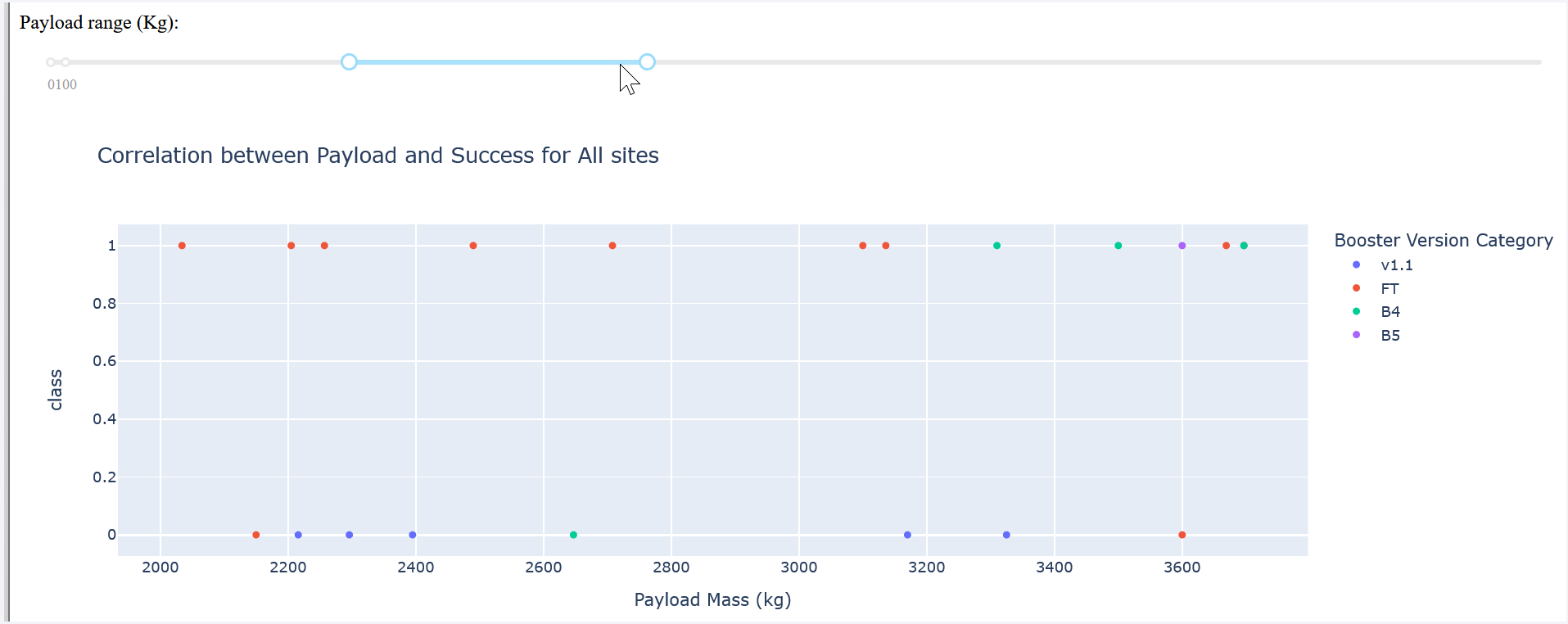
Total Success Launches for site KSC LC-39A

- 76.9 % of all Launces for site KSC LC-39A were sucessful



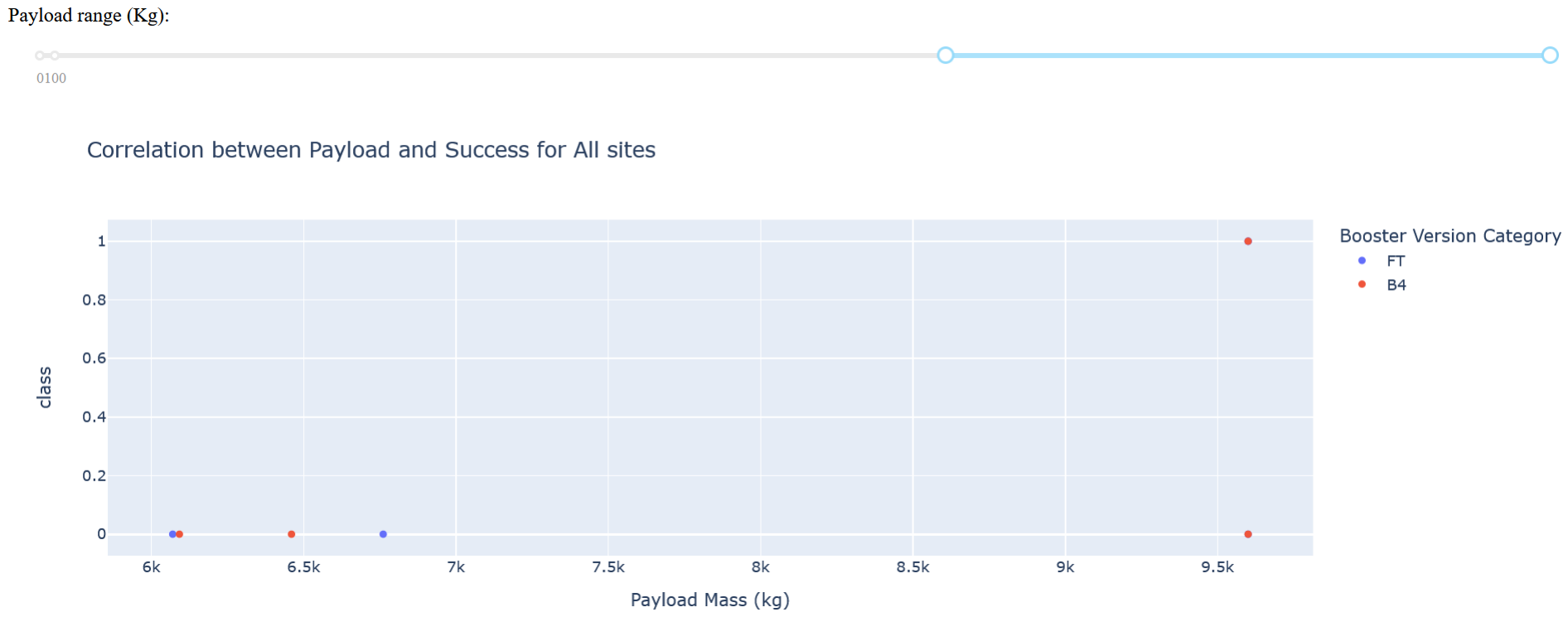
Correlation between Payload mass and Success for All sites

- The payload range between 2k and 4k has the highest launch success rate.



Correlation between Payload mass and Success for All sites

- The payload range between 6k and 8k has the lowest launch success rate.



Correlation between Payload mass and Success for All sites

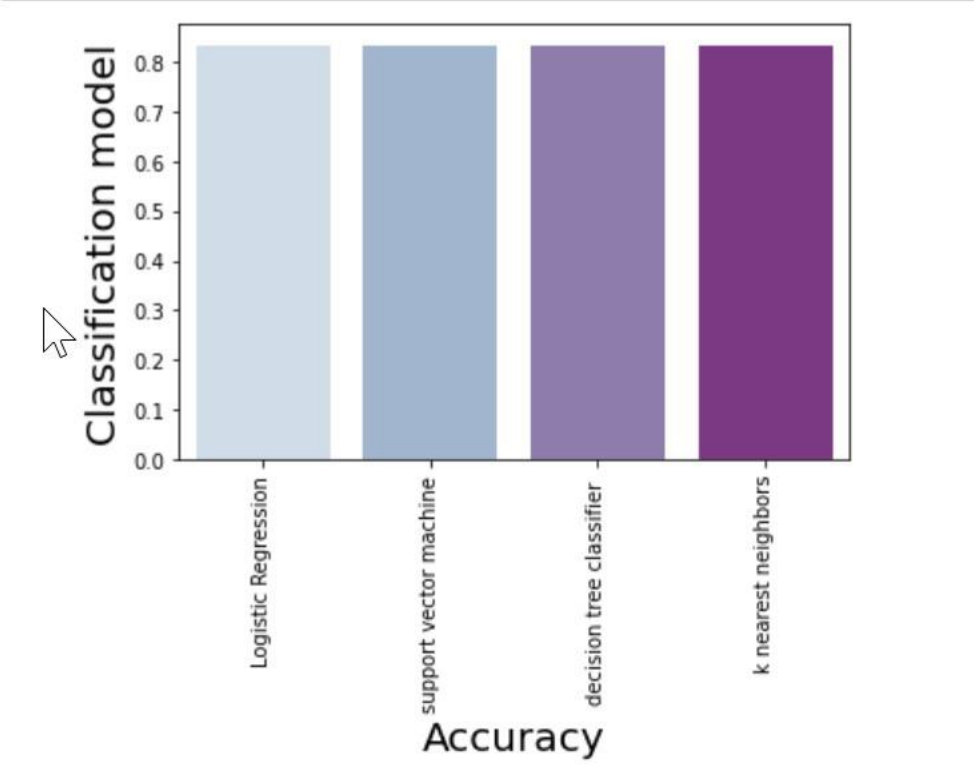
- The booster version FT has the highest launch success rate.

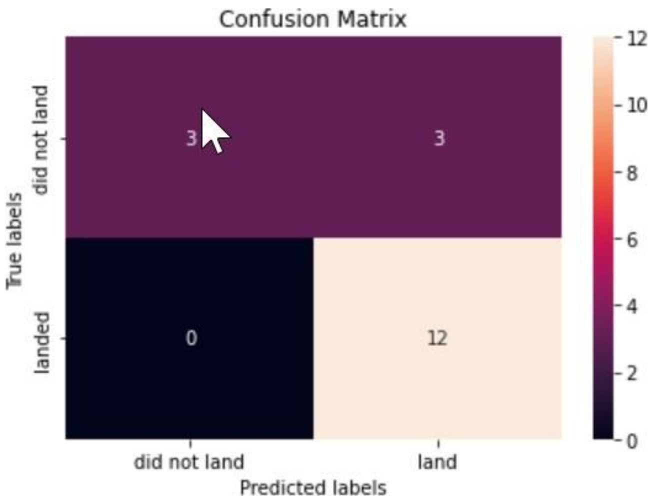
Section 5

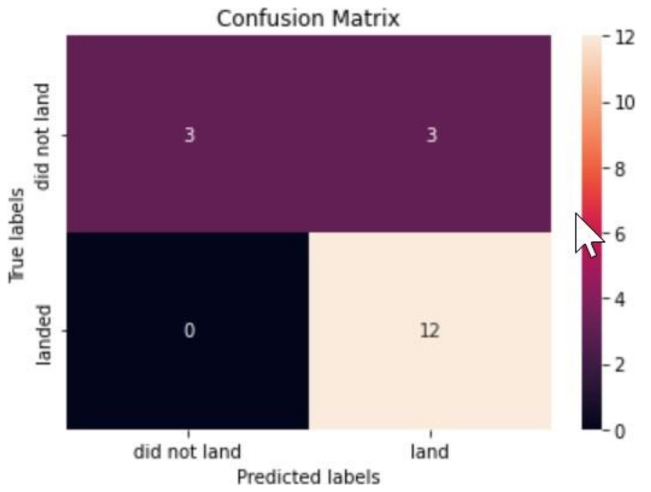
Predictive Analysis (Classification)

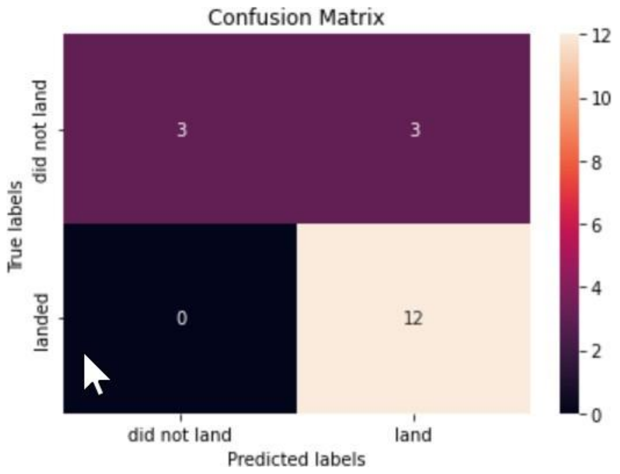
Classification Accuracy

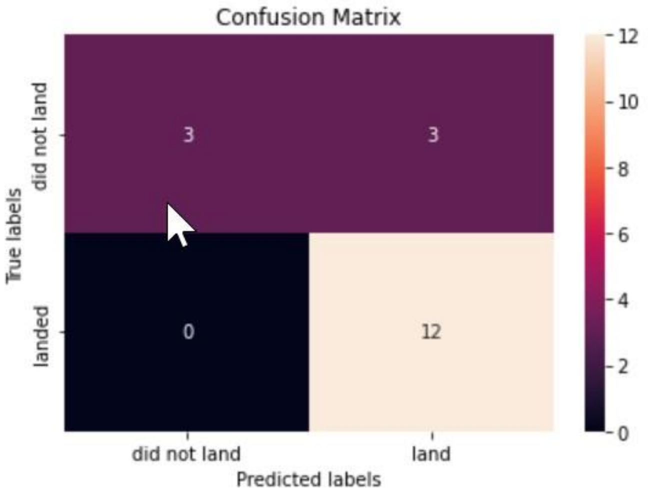
- All Models have the same accuracys











Confusion Matrix

- All Models have the same confusion matrix

Conclusions

- Some conclusions can be drawn from EDA and Plotly Dash, such as which launch site is best chosen. (cf. slide no. 17 -18)
- All Models have the same accuracys and confusion matrix.
- Therefore, it does not matter which model is used.
- The size of the test data is very small. In order to make a better statement about which model fits best, a larger test set must be provided.
- In order to be able to make an even better prediction, the model should always be adjusted or extended with new data and, ideally, additional data from other competitors should be added.

Appendix

- Code snippets, SQL queries, charts and Notebook outputs is to find under
  GitHub - LuciLul/IBM_Capstone_project

Thank you!