IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Hans W. Hiser
June 28, 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection using web scraping and Space X API
  - Exploratory Data Analysis (EDA) including Data Wrangling, Data Visualization, and interactive visual analytics
  - Prediction using Machine Learning
- Summary of all results
  - Data was collected from public sources.
  - EDA was used to determine which features were the best predictors of success of launches and landings.
  - Machine Learning was used to create a model that used key factors to predict the probability of a successful landing.

Introduction

- The objective is to determine the viability of creating a new company SpaceY to compete with SpaceX
- Desirable outcomes
  - Determine the key factors which lead to a successful landing
  - Determine the factors which will allow SpaceY to out compete SpaceX.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data from Space X was obtained from 2 sources:
    - Space X API (https://api.spacexdata.com/v4/rockets/)
    - WebScraping

    (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL

Methodology

Executive Summary

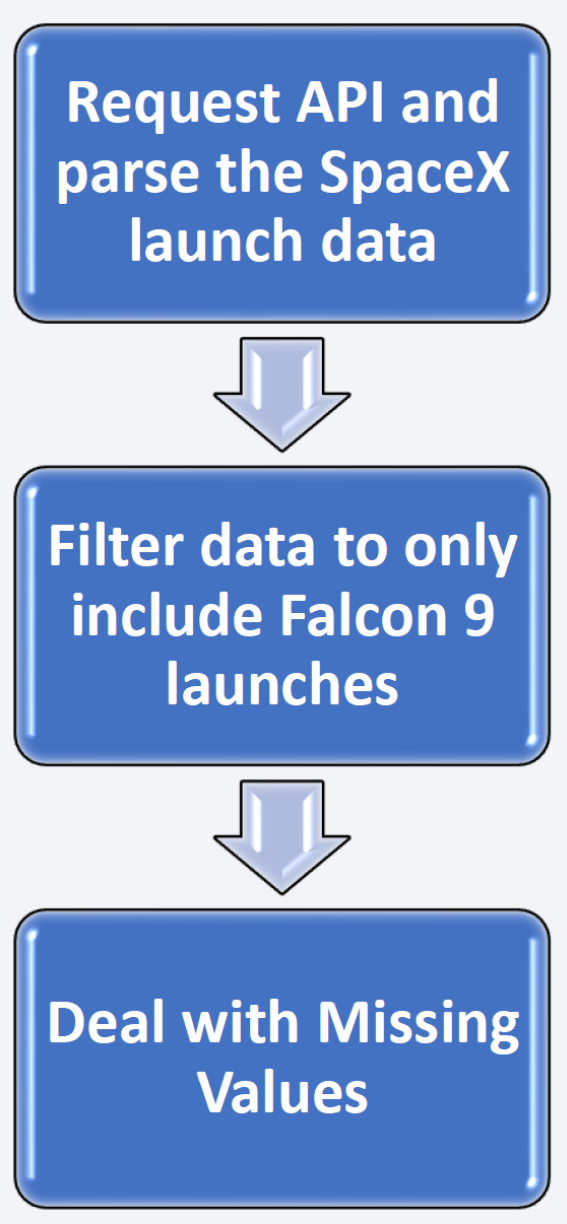
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Data that was collected until this step were normalized, divided in training
    - and test data sets and evaluated by four different classification models, being
    - the accuracy of each model evaluated using different combinations of
    - parameters.

Data Collection

Data sets were collected from Space X API (https://api.spacexdata.com/v4/rockets/) and from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches), using web scraping technics.

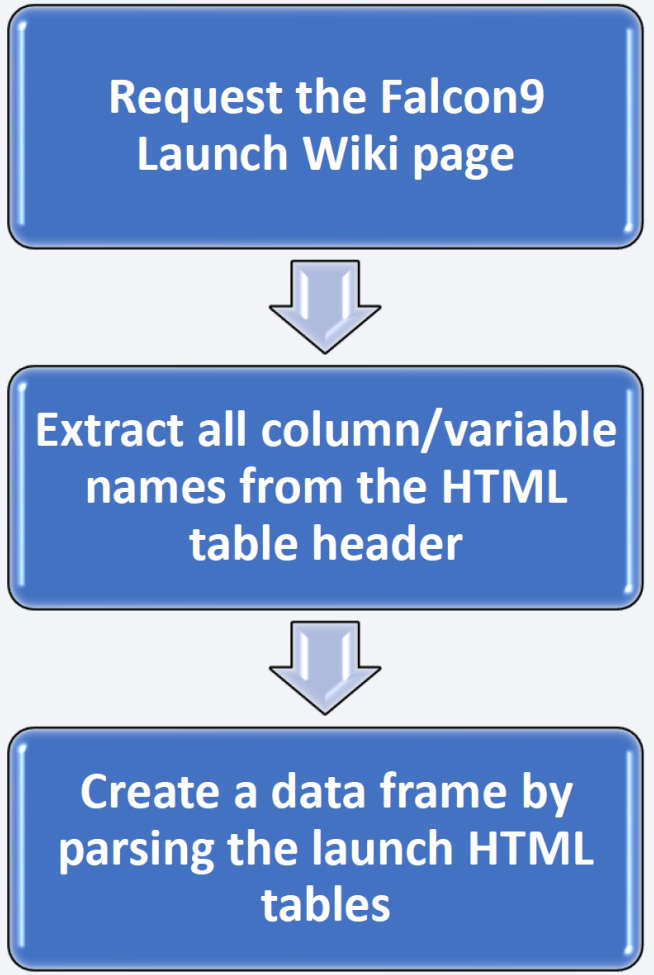
Data Collection

- SpaceX offers a public API from where data can be obtained and then used;
- This API was used according to the flowchart beside and then data is persisted.
- Source Code: https://github.com/hwhiser/IBMPCFinalProject/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Data from SpaceX launches can also be obtained from Wikipedia;
- Data are downloaded from Wikipedia according to the flowchart.
- Source Code: https://github.com/hwhiser/IBMPCFinalProject/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

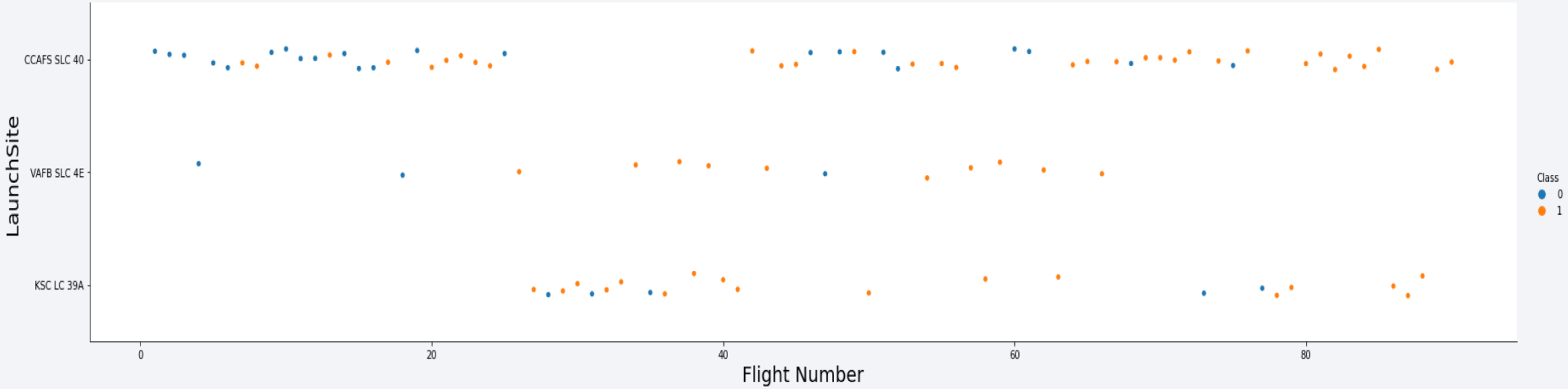


Data Wrangling

- Initially some Exploratory Data Analysis (EDA) was performed on the dataset.
- Then summaries launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated.
- Finally, the landing outcome labels were created from the Outcome column.
- Source Code: https://github.com/hwhiser/IBMPCFinalProject/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with Data Visualization

- Data exploration was conducted by compiling scatterplots and barplots to visualize the data and find trends,
    - Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, Payload and Orbit
- Source Code: https://github.com/hwhiser/IBMPCFinalProject/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb



EDA with SQL

- The following SQL queries were performed:
  - Names of the unique launch sites in the space mission;
  - Top 5 launch sites whose name begin with the string 'CCA';
  - Total payload mass carried by boosters launched by NASA (CRS);
  - Average payload mass carried by booster version F9 v1.1;
  - Date when the first successful landing outcome in ground pad was achieved;
  - Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg;
  - Total number of successful and failure mission outcomes;
  - Names of the booster versions which have carried the maximum payload mass;
  - Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015; and
  - Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.
- Source Code: https://github.com/hwhiser/IBMPCFinalProject/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

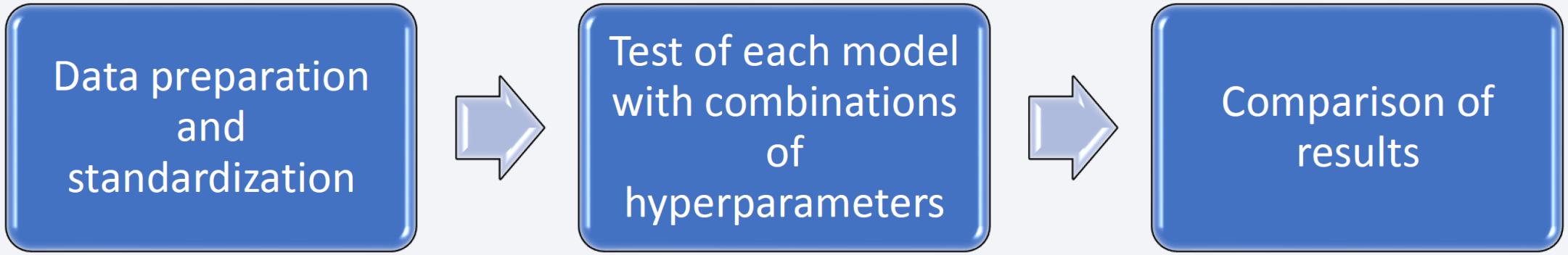
- Markers, circles, lines and marker clusters were used with Folium Maps
- Markers indicate points like launch sites;
- Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center;
- Marker clusters indicates groups of events in each coordinate, like launches in a launch site; and
- Lines are used to indicate distances between two coordinates.
- Source Code: https://github.com/hwhiser/IBMPCFinalProject/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- The following graphs and plots were used to visualize data
  - Percentage of launches by site
  - Payload range
  - This combination allowed to quickly analyze the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads.
- Source Code:
  https://github.com/hwhiser/IBMPCFinalProject/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Four classification models were compared: logistic regression, support vector machine, decision tree, and k nearest neighbors.
- Source Code: https://github.com/hwhiser/IBMPCFinalProject/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

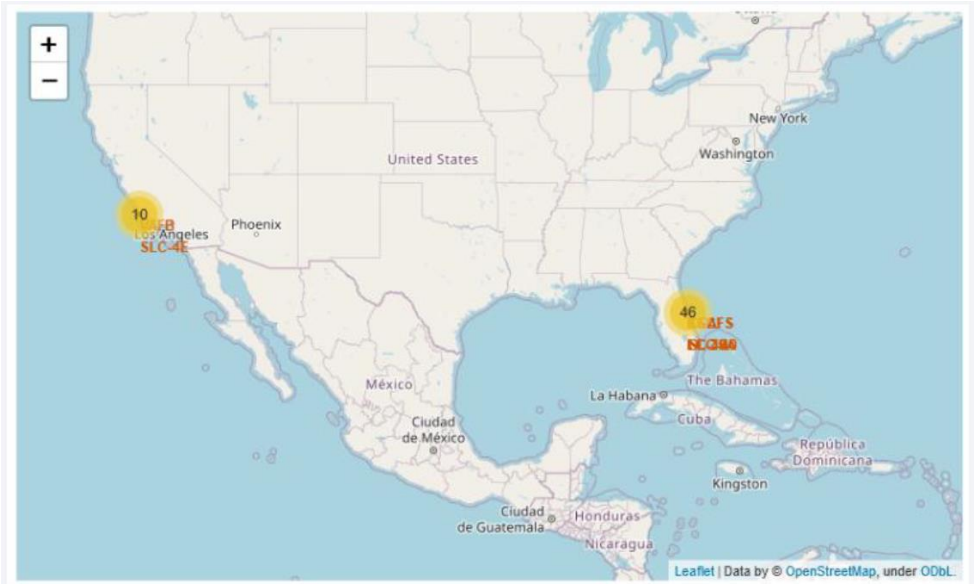


Results

- Exploratory data analysis results
  - Space X uses 4 different launch sites;
  - The first launches were done to Space X itself and NASA;
  - The average payload of F9 v1.1 booster is 2,928 kg;
  - The first success landing outcome happened in 2015 fiver year after the first launch;
  - Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average;
  - Almost 100% of mission outcomes were successful;
  - Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015;
  - The number of landing outcomes became as better as years passed.

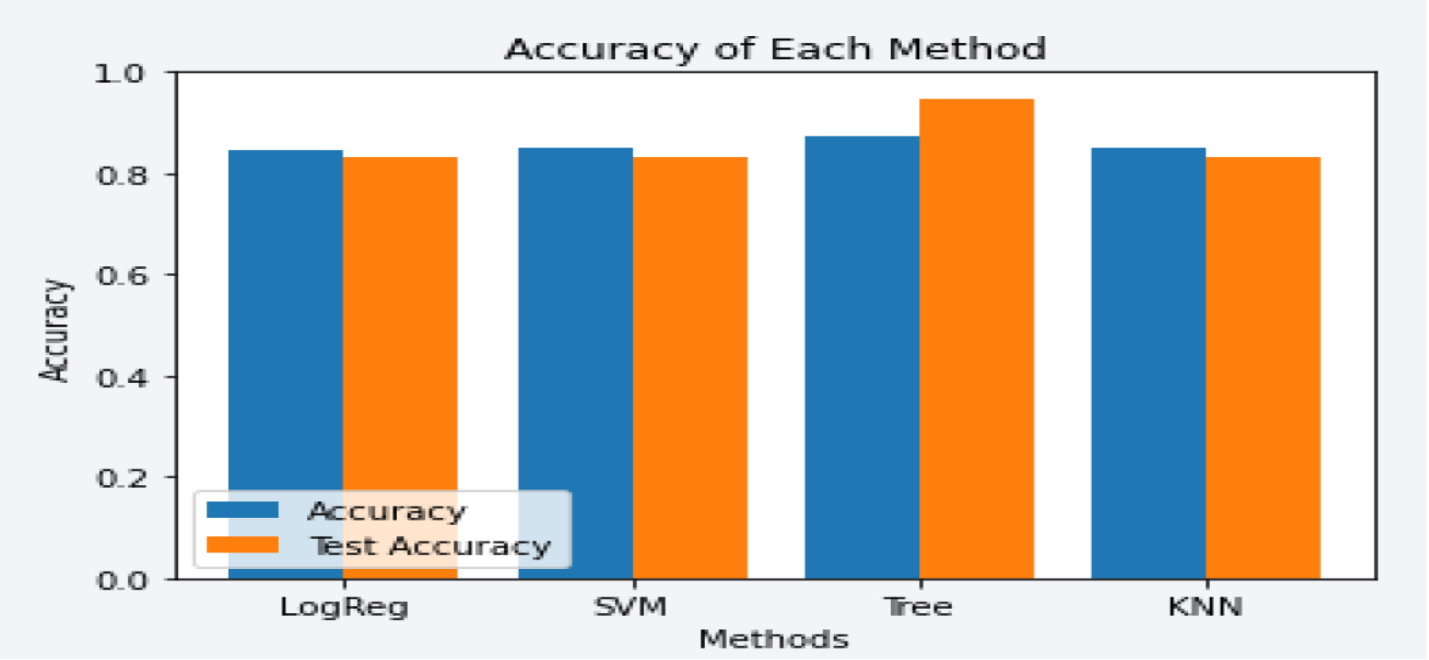
Results

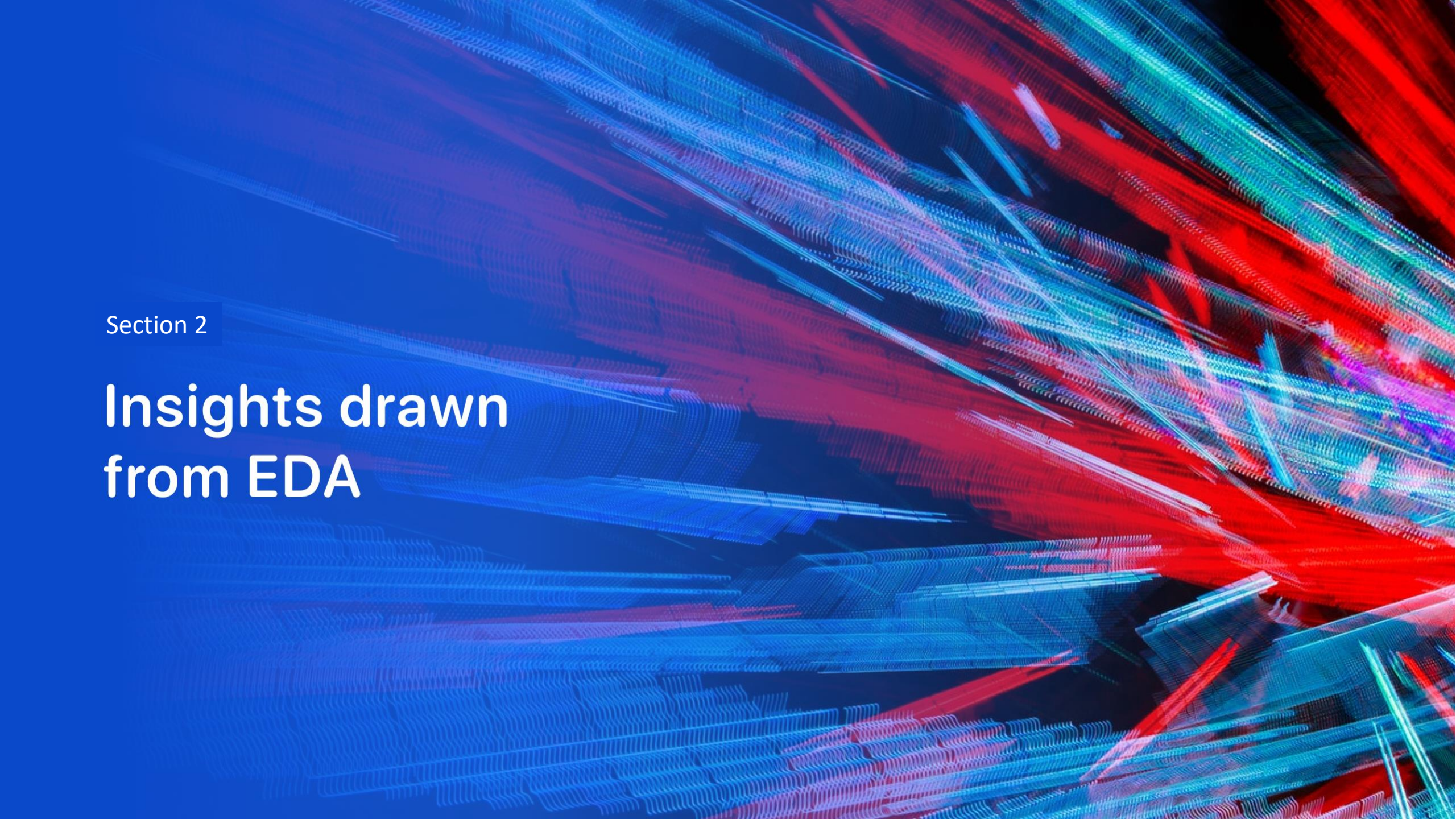
- Interactive analytics
  - Using interactive analytics was possible to identify the launch sites are located in safe places, near the sea, and have a good logistic infrastructure around.
  - Most launches happens on the east cost.



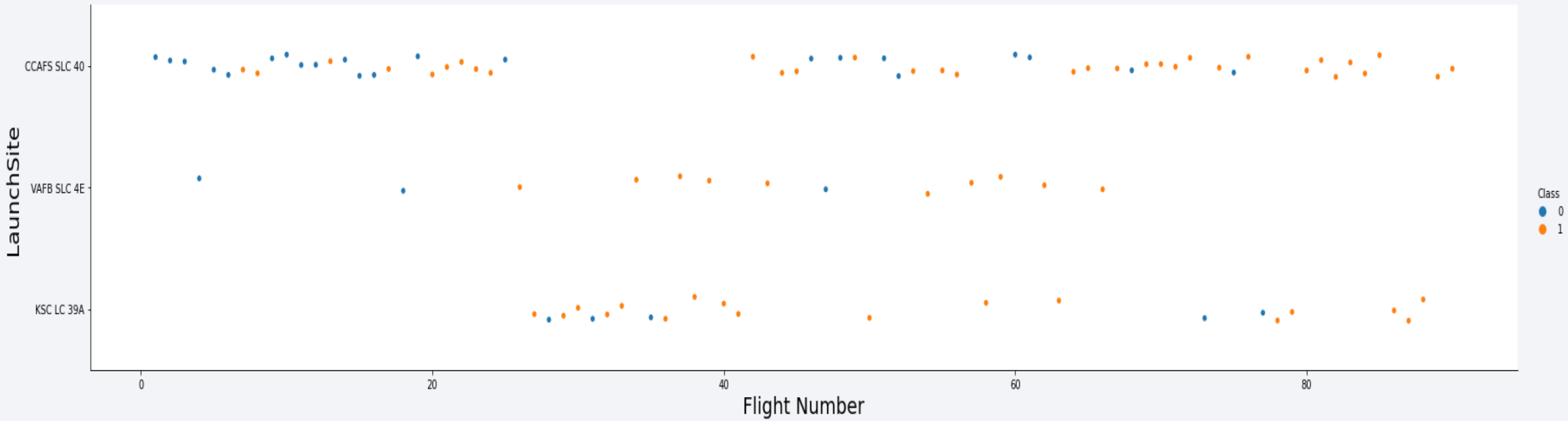
Results

- Predictive analysis results
  - Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings, having accuracy over 87% and accuracy for test data over 94%.



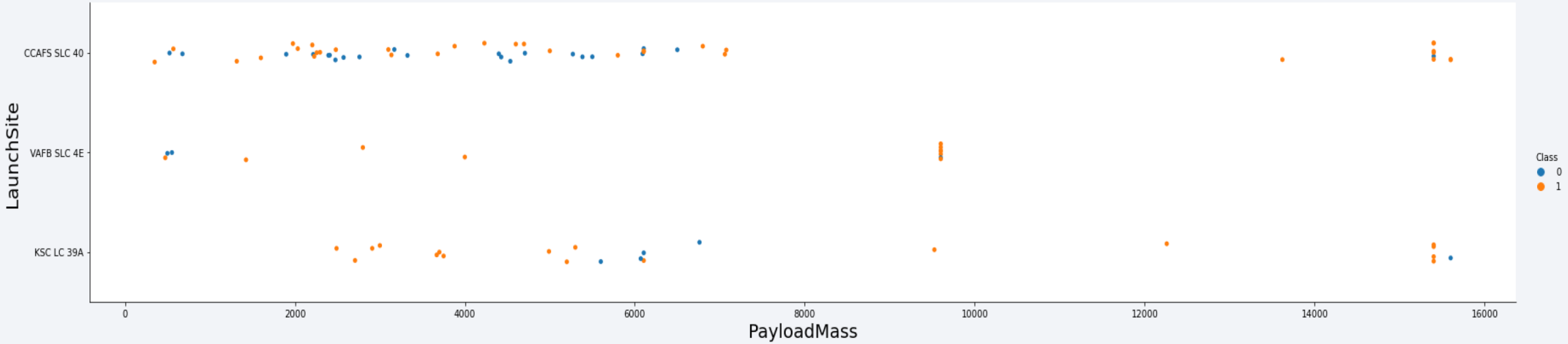

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- According to the plot above, it's possible to verify that the best launch site nowadays is CCAF5 SLC 40, where most of recent launches were successful;
- In second place VAFB SLC 4E and third place KSC LC 39A;
- It's also possible to see that the general success rate improved over time.

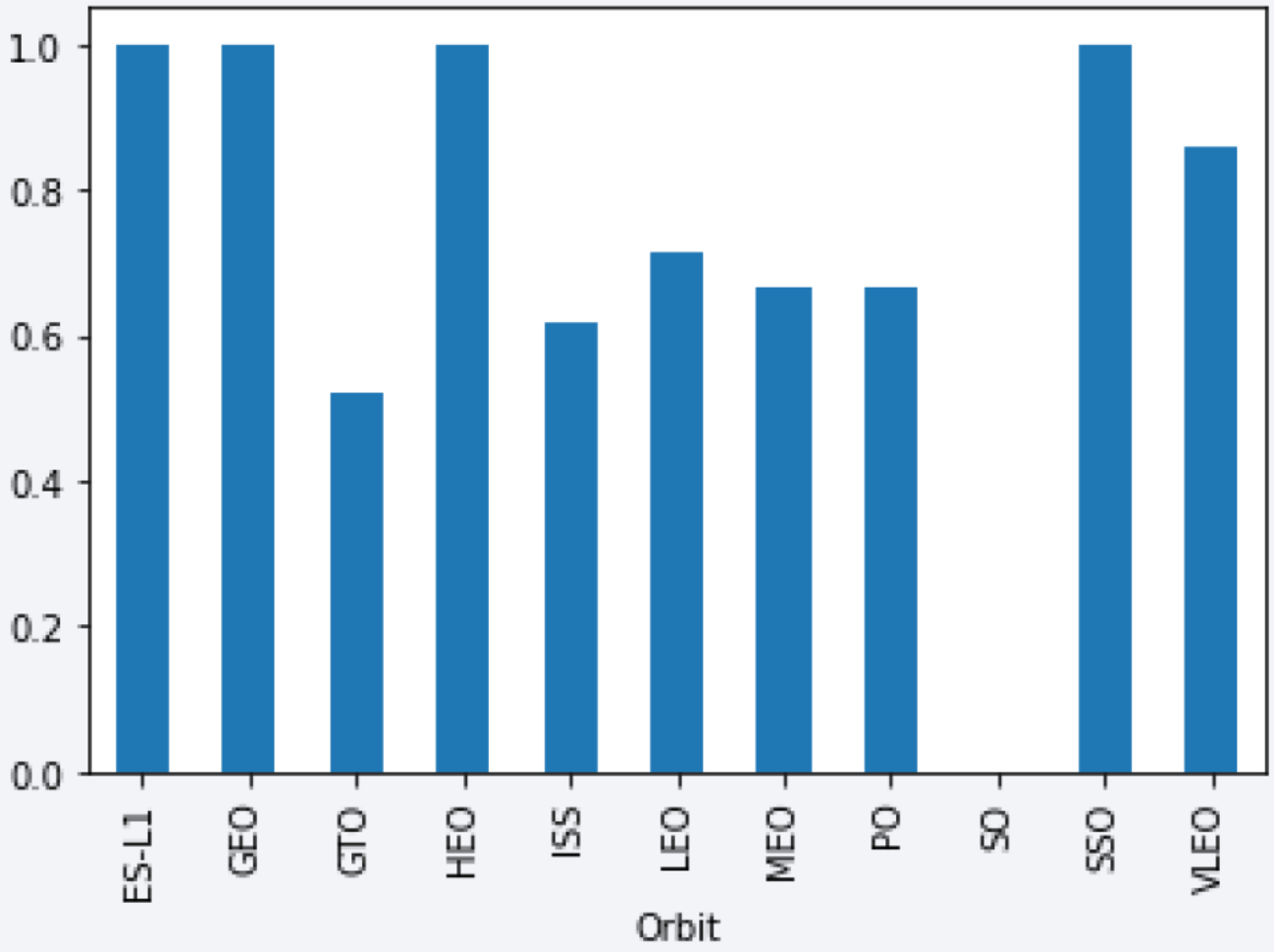
Payload vs. Launch Site



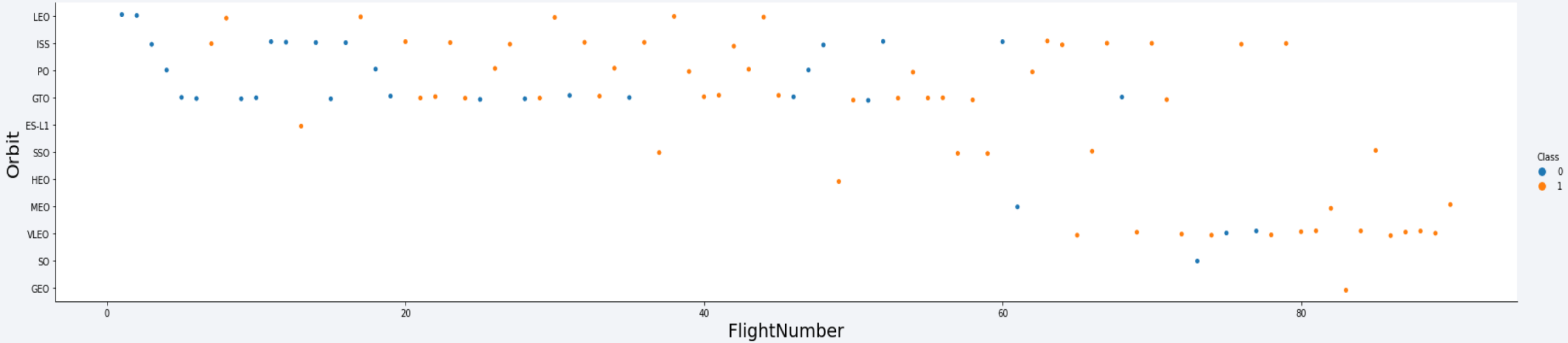
- Payloads over 9,000kg (about the weight of a school bus) have excellent success rate;
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.

Success Rate vs. Orbit Type

- The biggest success rates happens to orbits:
  - ES-L1;
  - GEO;
  - HEO; and
  - SSO.
- Followed by:
  - VLEO (above 80%); and
  - LFO (above 70%).

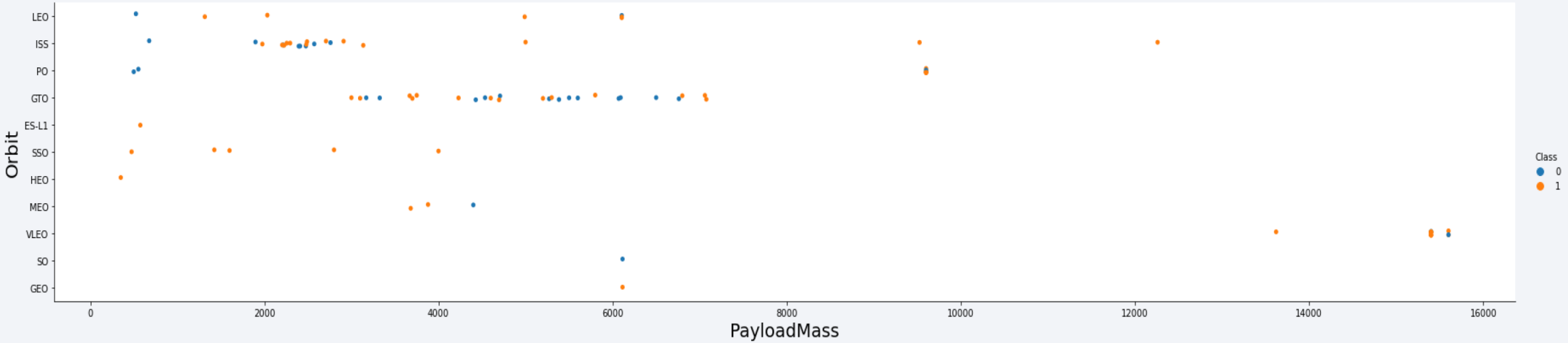


Flight Number vs. Orbit Type



- Apparently, success rate improved over time to all orbits;
- VLEO orbit seems a new business opportunity, due to recent increase of its frequency.

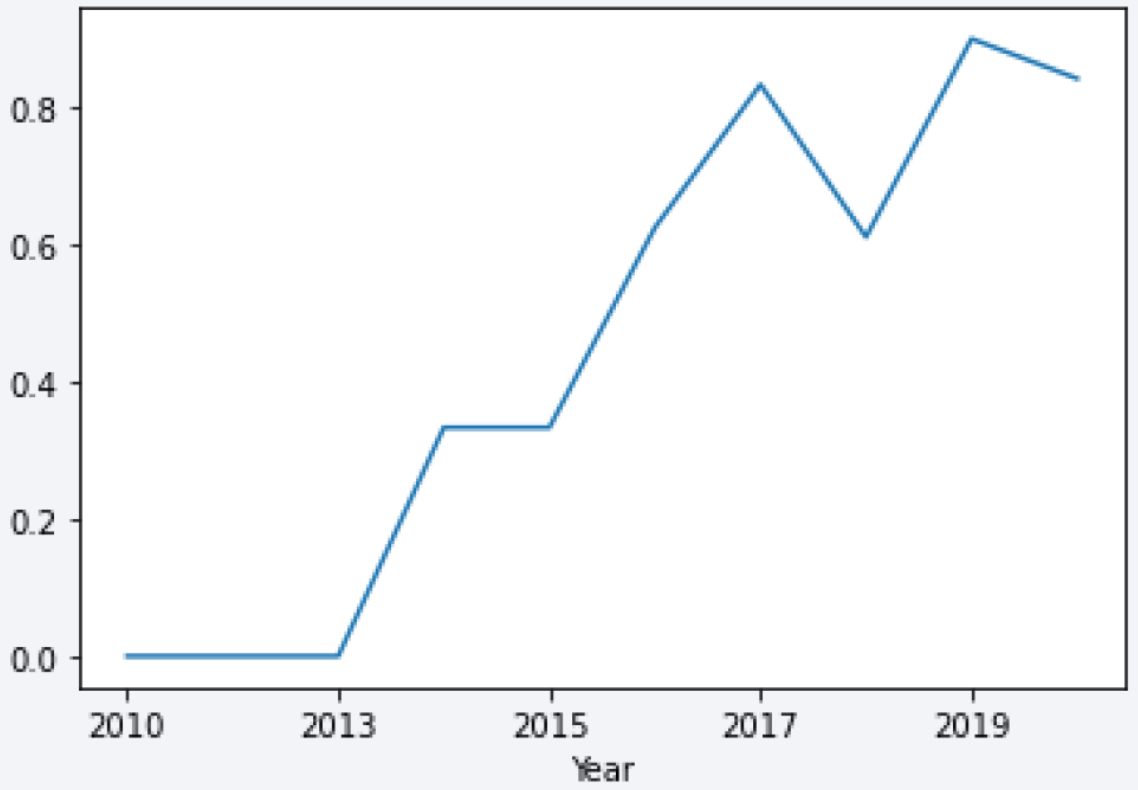
Payload vs. Orbit Type



- Apparently, there is no relation between payload and success rate to orbit GTO;
- ISS orbit has the widest range of payload and a good rate of success;
- There are few launches to the orbits SO and GEO.

Launch Success Yearly Trend

- Success rate started increasing in 2013 and kept until 2020;
- •It seems that the first three years were a period of adjusts and improvement of technology.



All Launch Site Names

- According to data, there are four launch sites:
- They are obtained by selecting unique occurrences of “launch_site” values from the dataset.

Launch Site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

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
2010-06-04	18:45:00	F9 v1.0 B0003	**CCA**FS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	**CCA**FS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	**CCA**FS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	**CCA**FS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	**CCA**FS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attemp

The only launch site that began with CCA was Cape Canaveral launch site.

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

Total Payload (kg)
111.268

- Total payload of launches using NASA boosters was done by summing all payloads that contained CRS, which coresponds to NASA.

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1

Avg Payload (kg)
2.928

- First, we filtered the data by querying the data and returning only launches using the booster version F9 v1.1. Then we calculated the average payload of the launches in our new table.

First Successful Ground Landing Date

- First successful landing outcome on ground pad:

Min Date
2015-12-22

- We used a query that looked for the earliest date that had a successful landing of Stage 1 rocket.

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

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

- We filtered the data using the range of payloads above and the criterion of a successful landing of the Stage 1 rocket. Then we looked at only distinct booster versions to compile the list above.

Total Number of Successful and Failure Mission Outcomes

- Number of successful and failure mission outcomes

Mission Outcome	Occurrences
Success	99
Success (payload status unclear)	1
Failure (in flight)	1

- We used the SQL count function to determine the number of outcomes in each of the categories.

Boosters Carried Maximum Payload

- Booster which have carried the maximum payload mass

Booster Version (...)
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3

Booster Version
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

- The above boosters were used to carry the maximum payload.

2015 Launch Records

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

Booster Version	Launch Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- There were only 2 failures in 2015, per the list above.

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

- Rank of the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

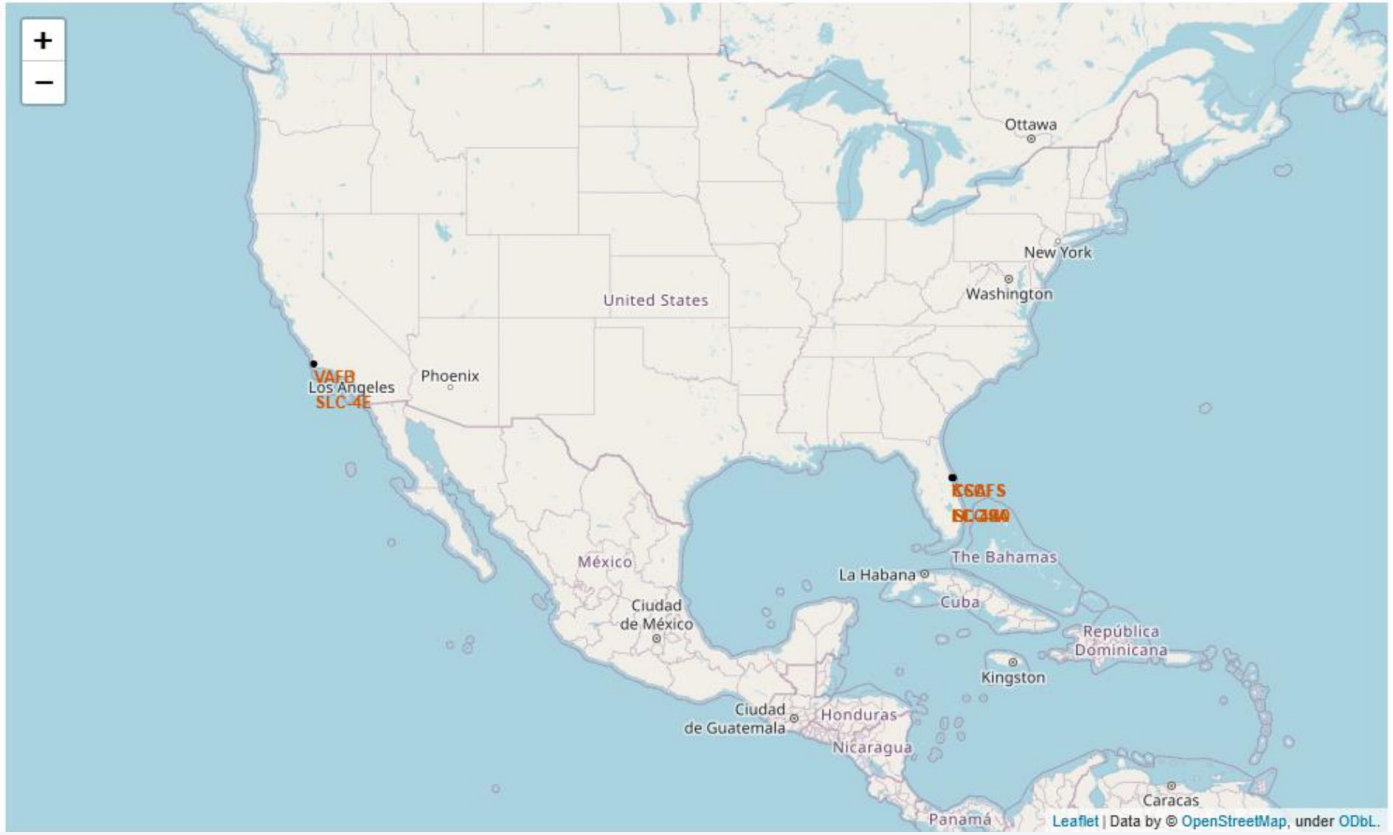
Landing Outcome	Occurrences
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

- Above we see a list of landing outcomes and occurrences. Of note NO Attempt means that there was no attempt of landing the 1st stage of the rocket.

Section 3

Launch Sites Proximities Analysis

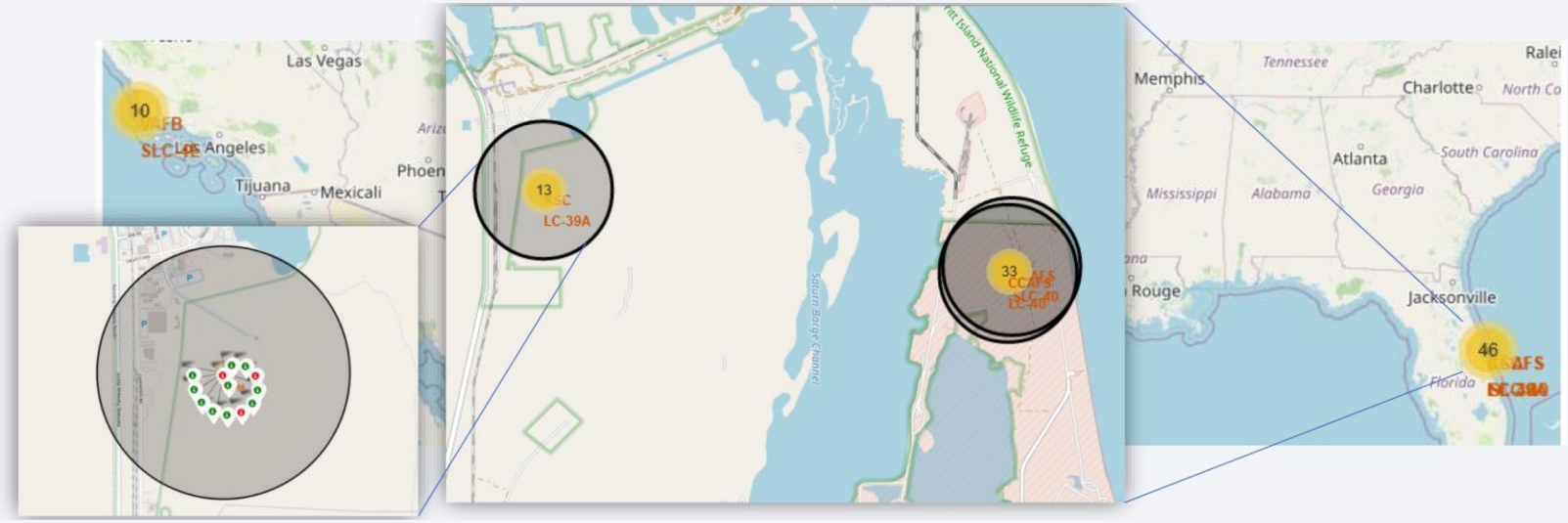
Map of all launch sites used by the SpaceX program



- Launch sites are located near the ocean so the recovery of stage 1 of the rocket will be easily recovered. Also, the infrastructure near the sites needs to be very well developed.

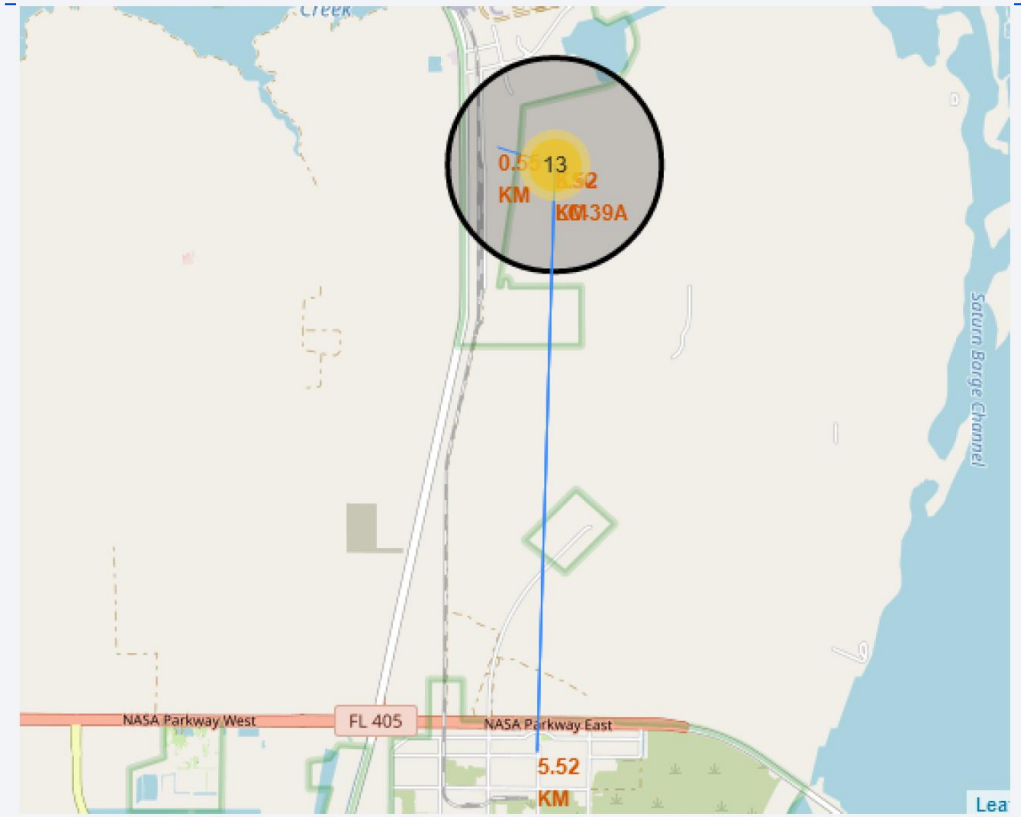
Launch outcome by site

- Example using the KSC LC-39A launch site

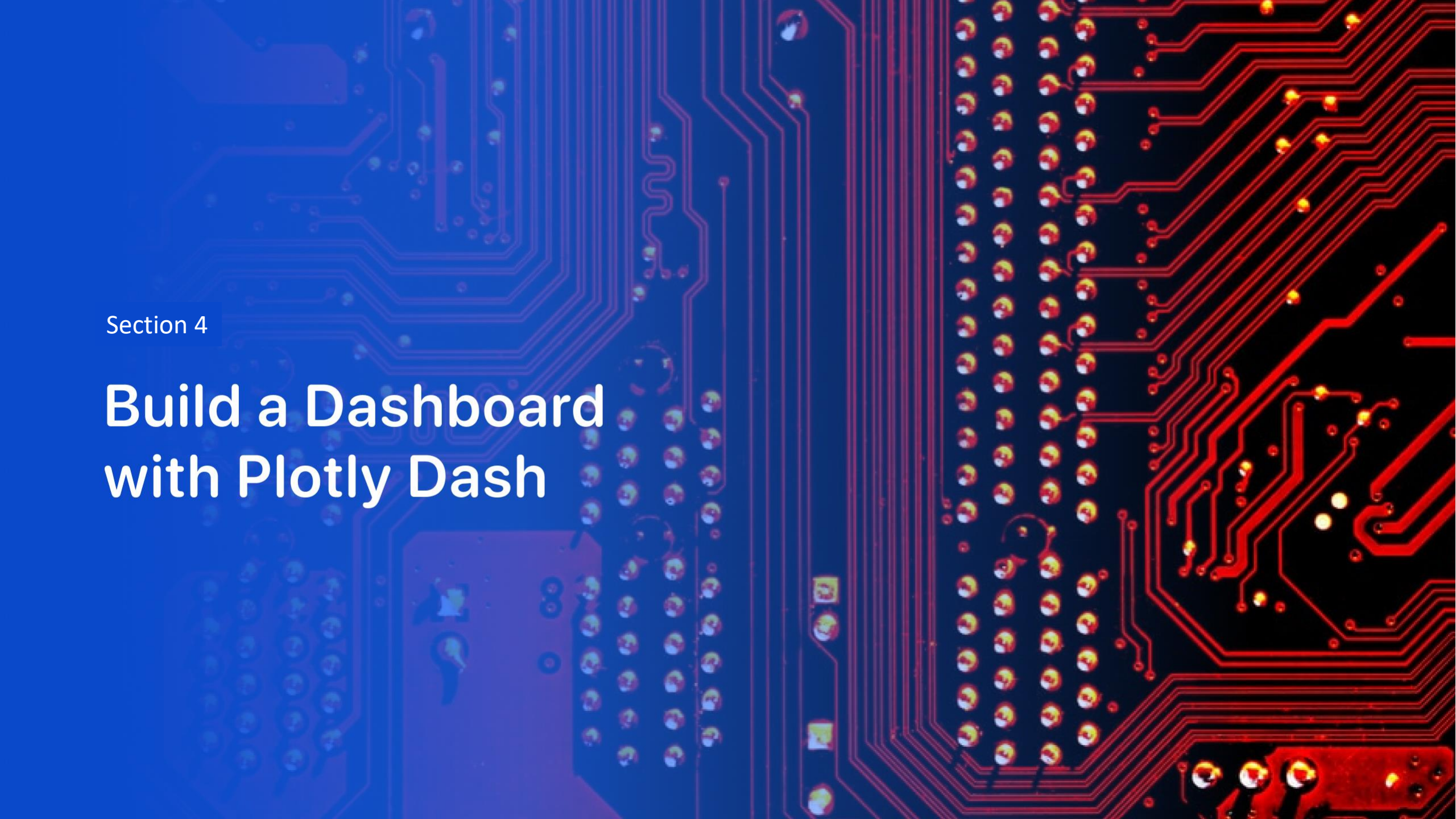


- Green markers represent successful stage 1 landing. Red indicates unsuccessful stage 1 landings.

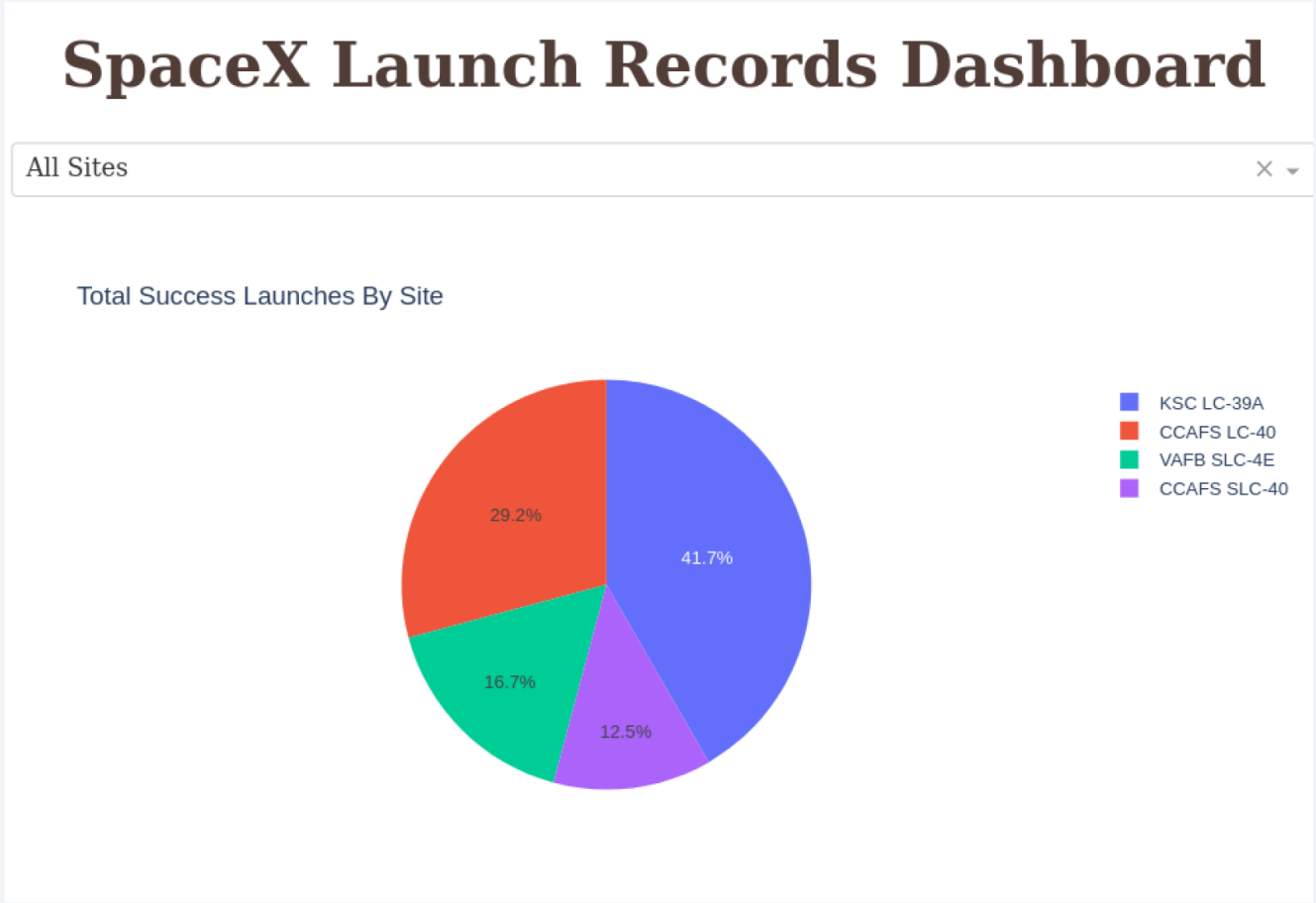
Logistics and infrastructure



- Launch site KSC LC-39A has good logistics and infrastructure, being near railroad and road and relatively far from inhabited areas.

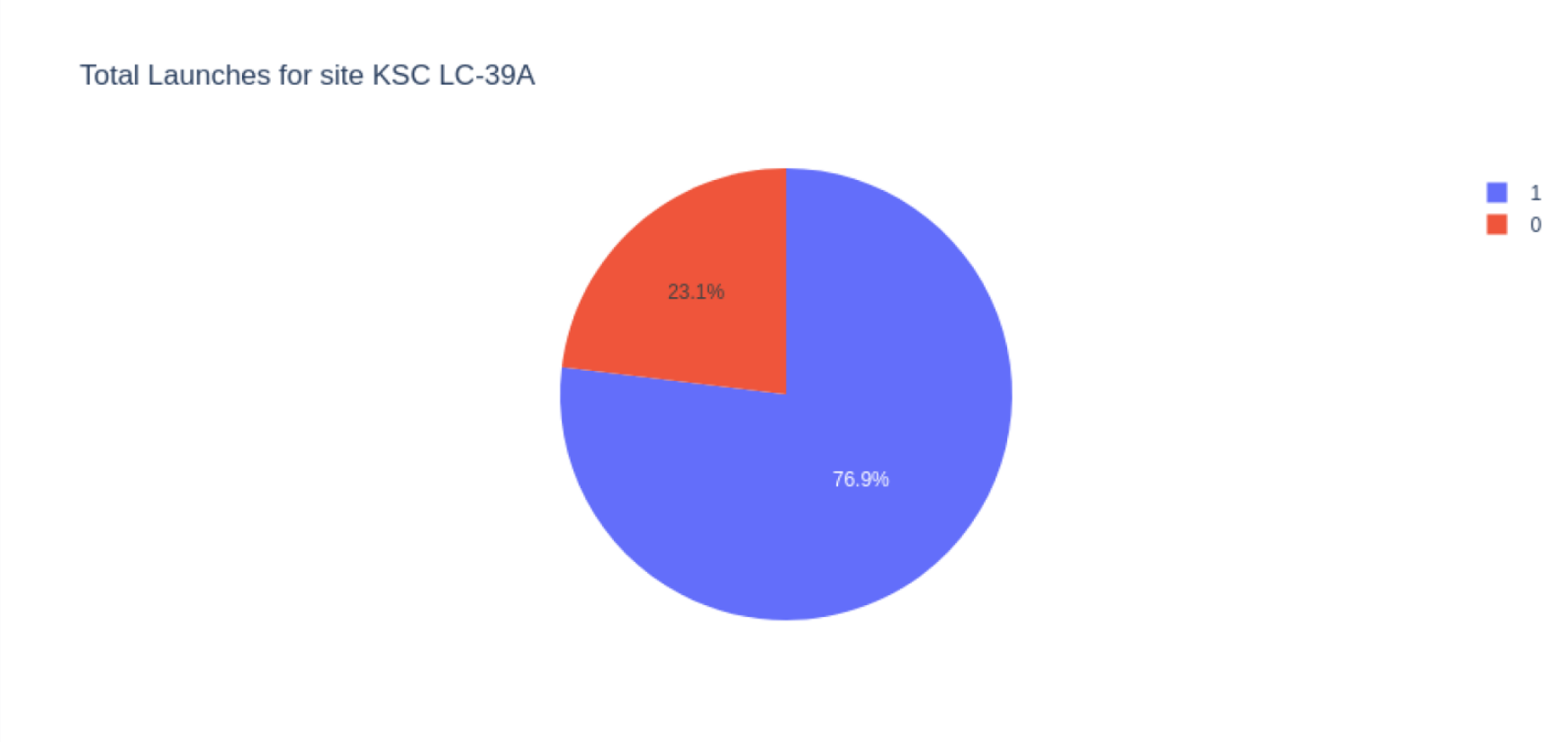

Section 4
Build a Dashboard with Plotly Dash

Successful Launches by Site



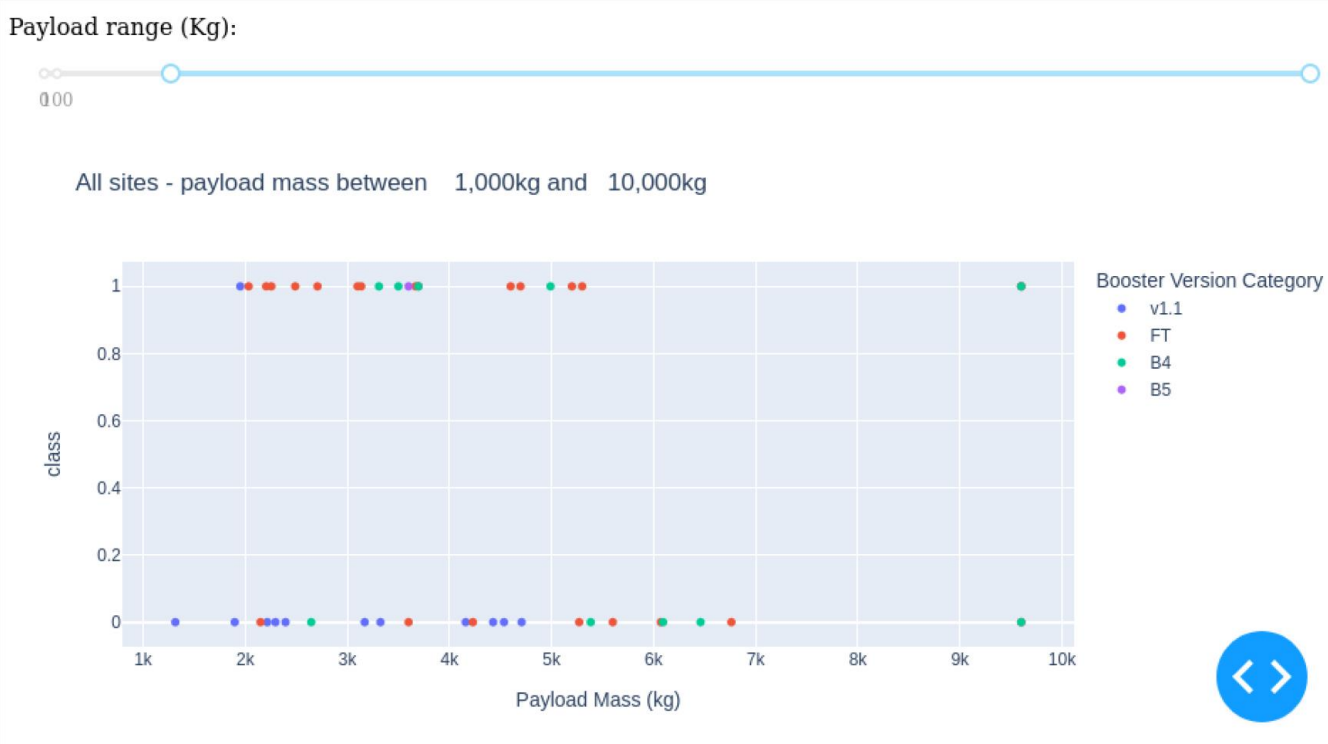
- Launch site is an important determinant of successful Stage 1 landing.

Launch success percent KSC LC-39A



- 76.9% of launches were successful.

Payload v Launch Outcomes



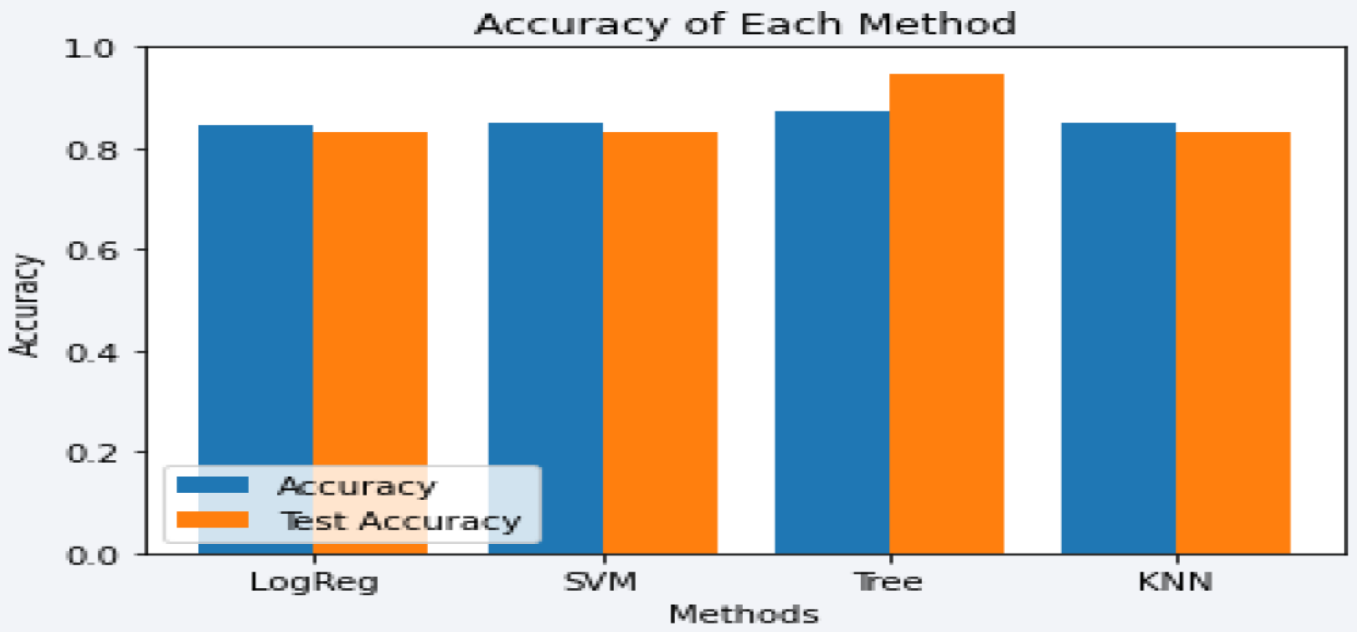
- A combination of a payload under 6,000 kg and using FT boosters is the best determinate of launch success.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Four classification models were tested, and their accuracies are plotted
- The model with the highest classification accuracy is Decision Tree Classifier, which has accuracies over than 87%.



Confusion Matrix

- Confusion matrix of Decision Tree Classifier proves its accuracy by showing the big numbers of true positive and true negative compared to the false positive and negative outcomes.

Conclusions

- Launch sites need to located close to the ocean and have access to well developed infrastructure.
- The launch site with the highest success rate was KSC0- LC-39A.
- Payload mass should be constrained to 6,000 kg and under.
- FT boosters displayed the highest rate of success.
- Although most missions have been successful, success rate has increased over time.
- Decision trees produced the best model.

Appendix

- As an improvement for model tests, it's important to set a value to np.random.seed variable.
- All code and figures can be accessed at through the following URL.
  - https://github.com/hwhiser/IBMPCFinalProject
- Folium didn't show maps on Github , so I took screenshots.

Thank you!