IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Luis Gil Guerrero
22/05/2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection from API and Web Scrapping
  - Data wrangling (Prepare data for Machine Learning)
  - Exploratory data analysis (EDA) through Data visualization and SQL
  - Visual analysis using interactive maps with Folium
  - Visual analysis through interactive dashboard with Plotly
  - Predictive analysis by Machine Learning (Classification)
- Summary of all results
  - Analysis from EDA results
  - Analysis from visual results
  - Predictions results from Machine Learning

Introduction

- Project background and context

  SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars, making it the cheapest compared with other companies because they can reuse the first stage in future landings.

  If we can determine if the first stage will land, we can determine the cost of the next launch, information that can be used if an alternate company wants to bid against SpaceX for a rocket launch.

- Problems to solve
  - Which factors determines a successful landing?
  - What are the characteristics of the rocket that have the most influence on landing?
  - What are the optimal conditions necessary to ensure a correct landing?

VS

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from SpaceX Rest API and through web scrapping from Wikipedia
- Perform data wrangling
  - Dropping irrelevant data and applying One-hot encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Graphs and SQL queries.
- Perform interactive visual analytics using Folium and Plotly Dash
  - Interactive maps and dashboards
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

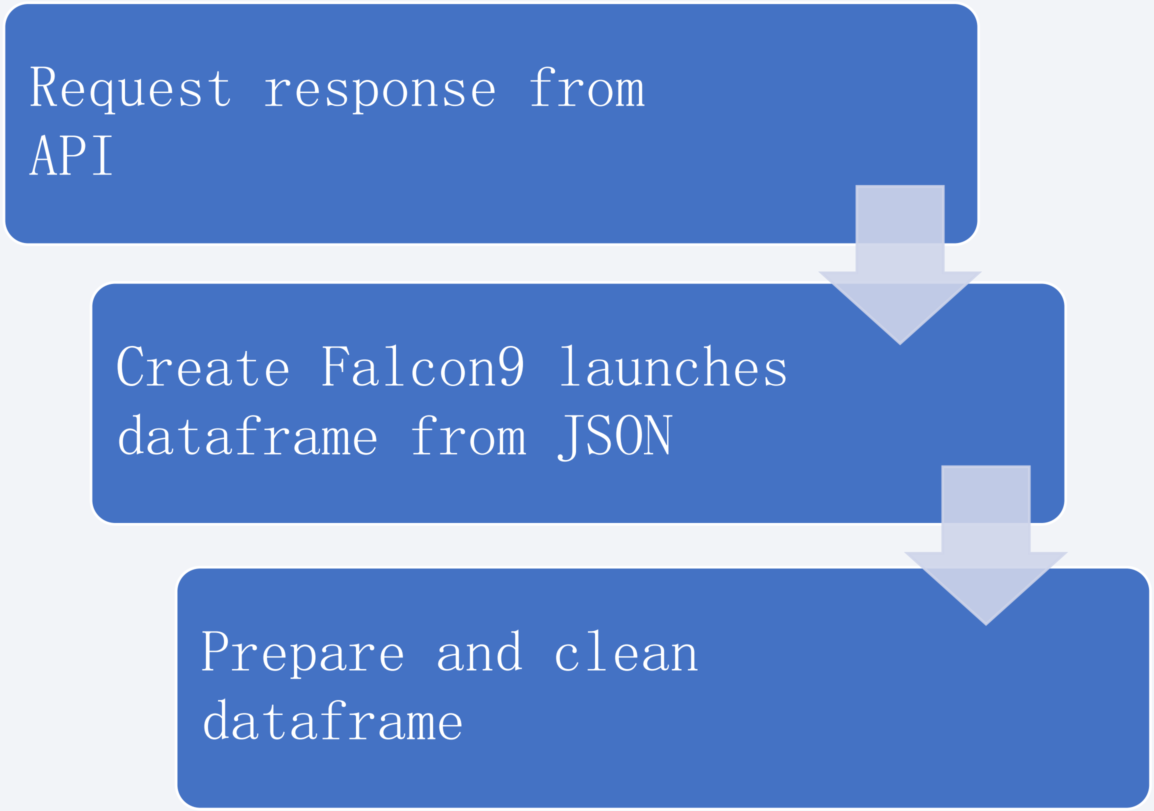
Data Collection

- SpaceX API
  - This rest API will provide us data about the rockets, launchpads and payloads used by SpaceX in their launches in form of a JSON file.
  - Our first step will be normalizing this files and transform them into dataframes that we can clean, drop and prepare to use in our analysis.
- Wikipedia
  - The launch records are stored in a HTML table
  - We can get this data through web scrapping and apply the same methods.

Data Collection – SpaceX API

- Get a response from the API
- Save response into a JSON file
- Prepare the data

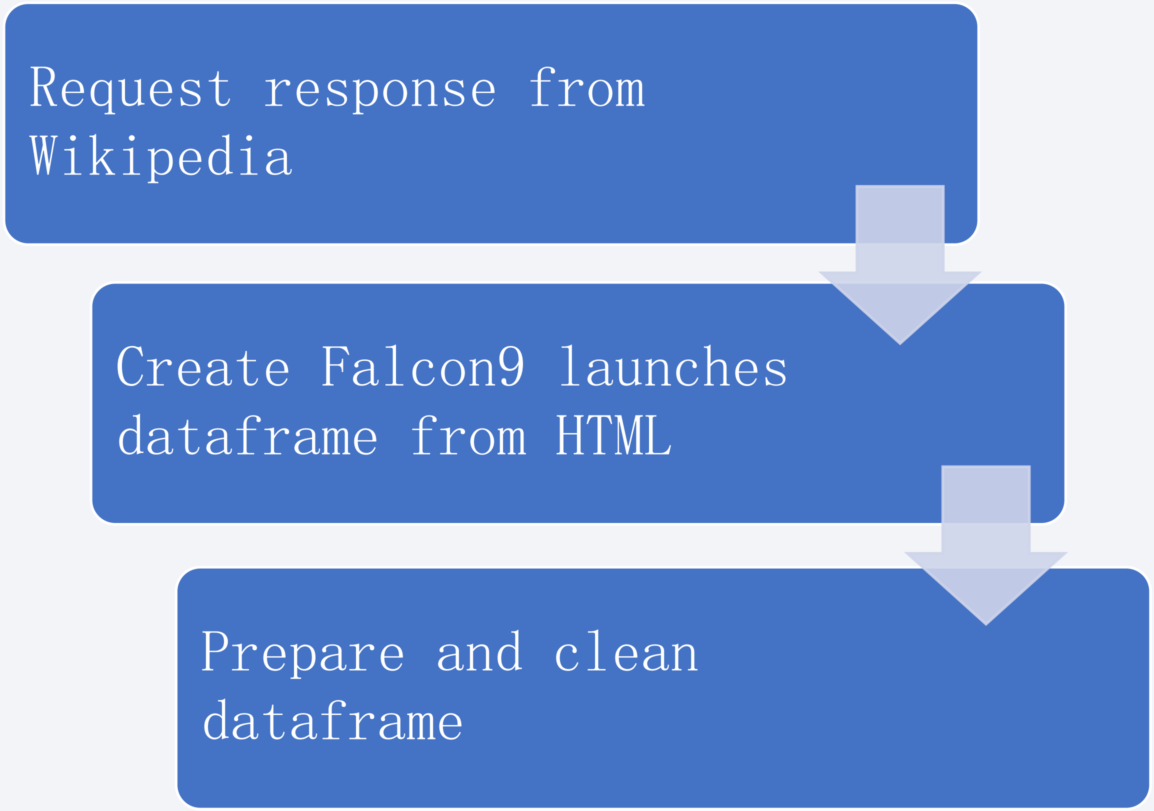
- [Github code]



Data Collection - Scraping

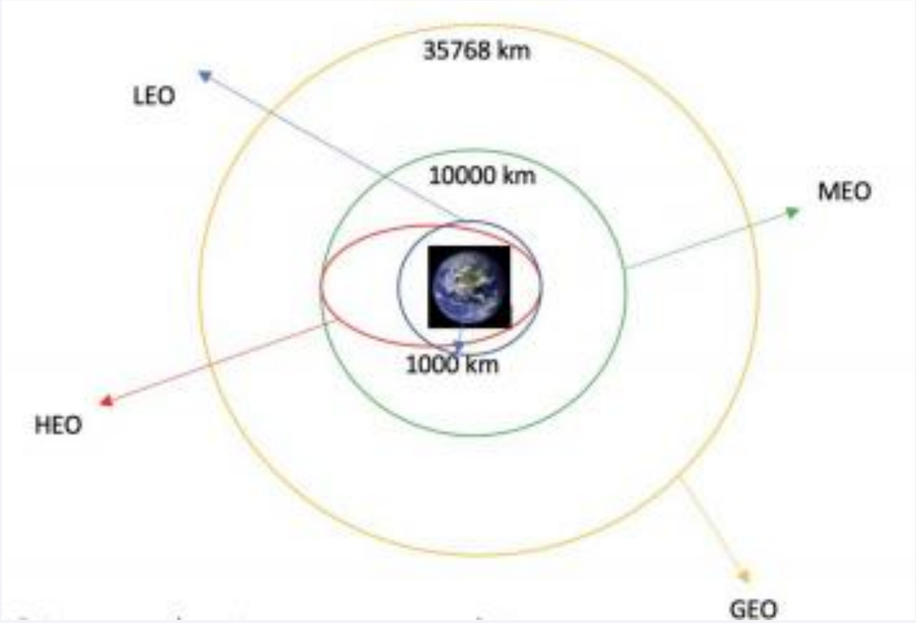
- Get the HTML from the list of launches from Wikipedia
- Select the valuable data
- Prepare the data

- [Github code]



Data Wrangling

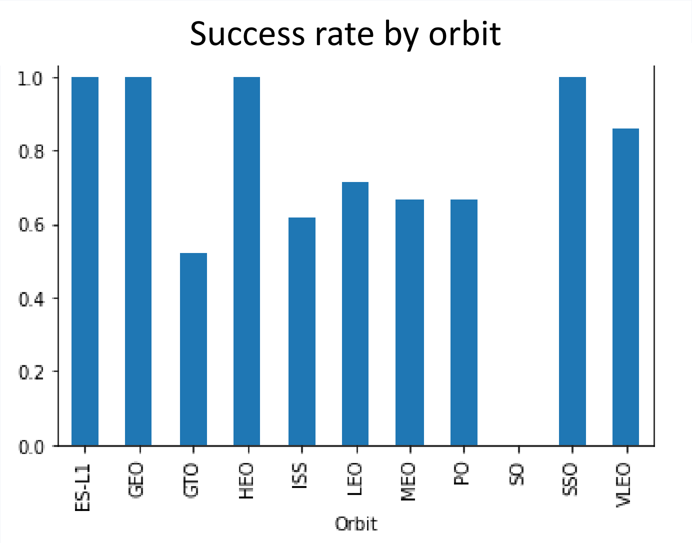
- First we did some exploratory data analysis on the dataset, saving only the relevant data and applying One-hot encoding.
- Then we calculated the number of launches at each site, the number and occurrence of each orbit and the number and occurrence of mission outcome per orbit type.
- We exported the dataset as a CSV file.
- We created a landing outcome label from the Outcome column and find success rate for every landing.
- Github code

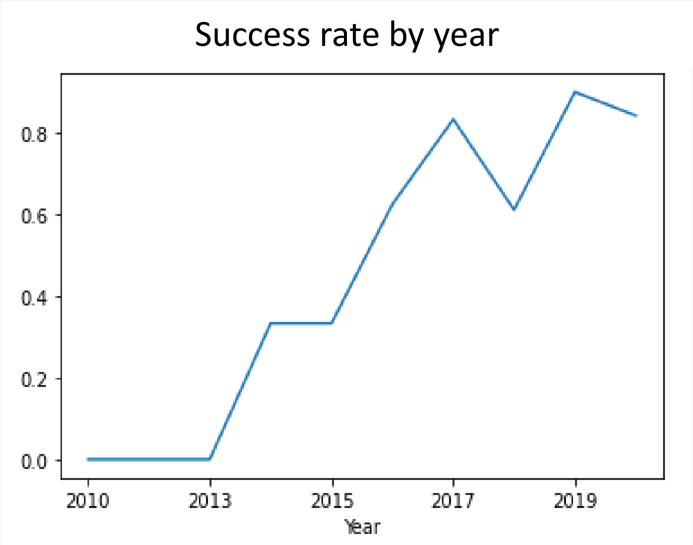


Exploratory data análisis → Calculate launches data → Save outcome

EDA with Data Visualization

- We analyzed the relationship of key variables like launch sites, flight number, payload and orbit type through visual data in form of bar charts, bar plots, scatter plots.
- [Github code]





EDA with SQL

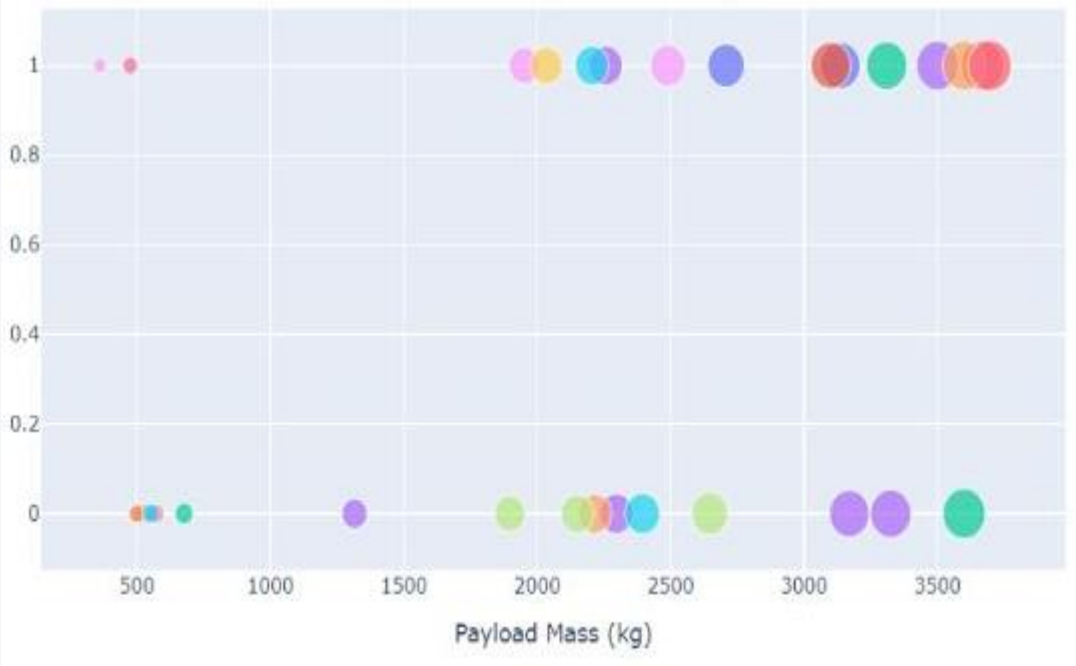
- Data was analyzed through SQL queries.
- Examples:
  - Names of the unique launch sites in the space mission
  - Total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - Date when the first successful landing outcome in ground pad was achieved
  - Total number of successful and failure mission outcomes
  - Names of the booster versions which have carried the maximum payload mass
- Github code

Build an Interactive Map with Folium

- Launch sites were localized and marked in an interactive map for better visualization.
- Markers indicate launch sites, circles highlight important areas, clusters contains groups of data and success rate and lines indicate distance between coordinates
- All these features helps us to identify possible problems and causes of success rate by the location of the launch site.
- [Github code]()

Build a Dashboard with Plotly Dash

- A dashboard created with Plotly Dash gives us relevant information in an interactive way
- We could visualize easily the percentage of launches by site and the payload range.
- We could analyze the relationship between this two factors for every booster versions.
- [Github code]()



Predictive Analysis (Classification)

- We build our model, loading the dataset using Numpy and Pandas and dividing it into training and test.
- Then we implemented different machine learning models to the data selecting the best hyperparameters.
- Lastly we evaluated the models accuracy and plotted confusion matrixes to find the best classification model for this case.
- Github code

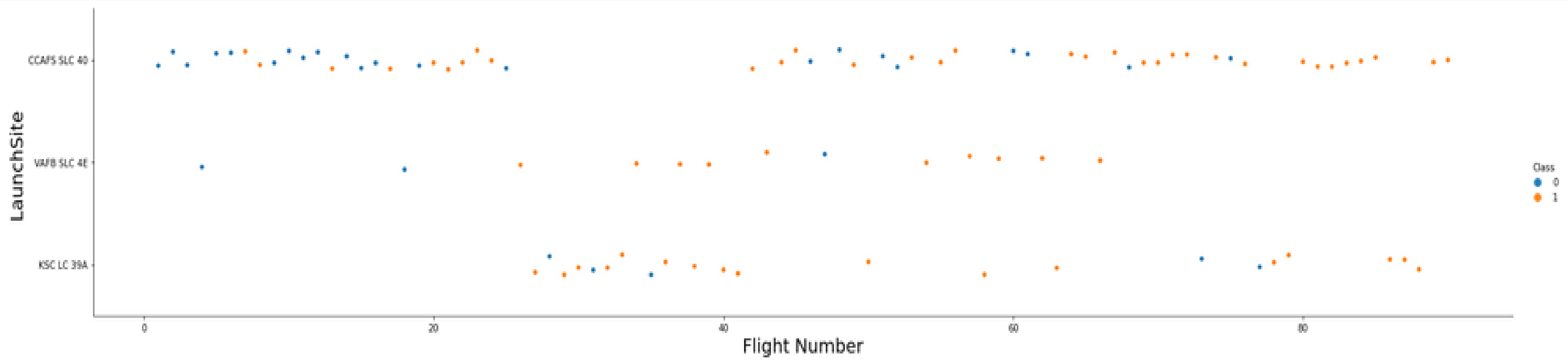
Results

- Exploratory data analysis results
  - The first success landing was in 2015
  - The number of successful landings increases as years passed
- Interactive analytics demo in screenshots
  - Most launches were on east coast
  - Launches take place near the sea and need multiple forms of transport
- Predictive analysis results
  - Decision tree classifier is the best model to predict a successful landing

Section 2

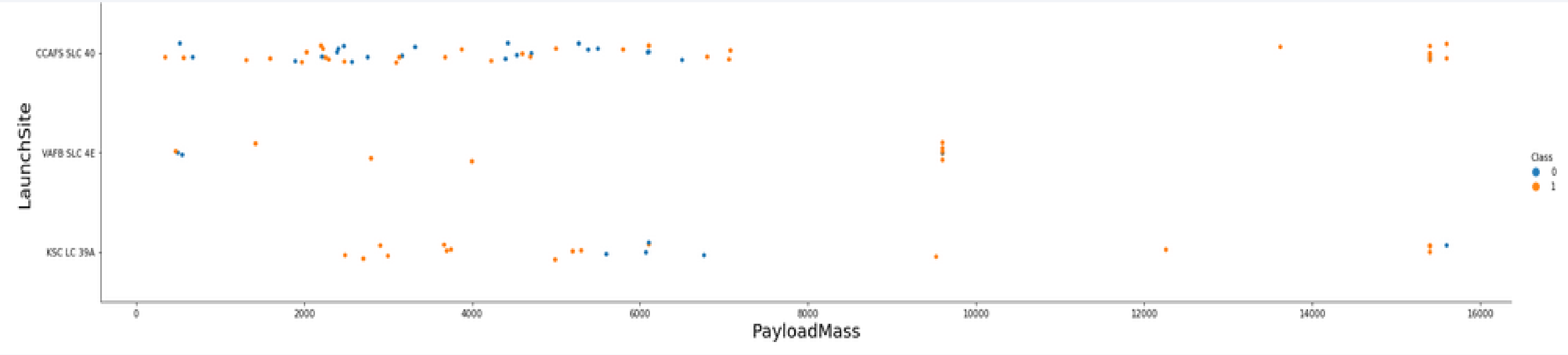
Insights drawn from EDA

Flight Number vs. Launch Site



- As we can see, as the flight amount increases in a launch site, the success rate goes higher.
- The best launch site is CCAF5 SLC 40.

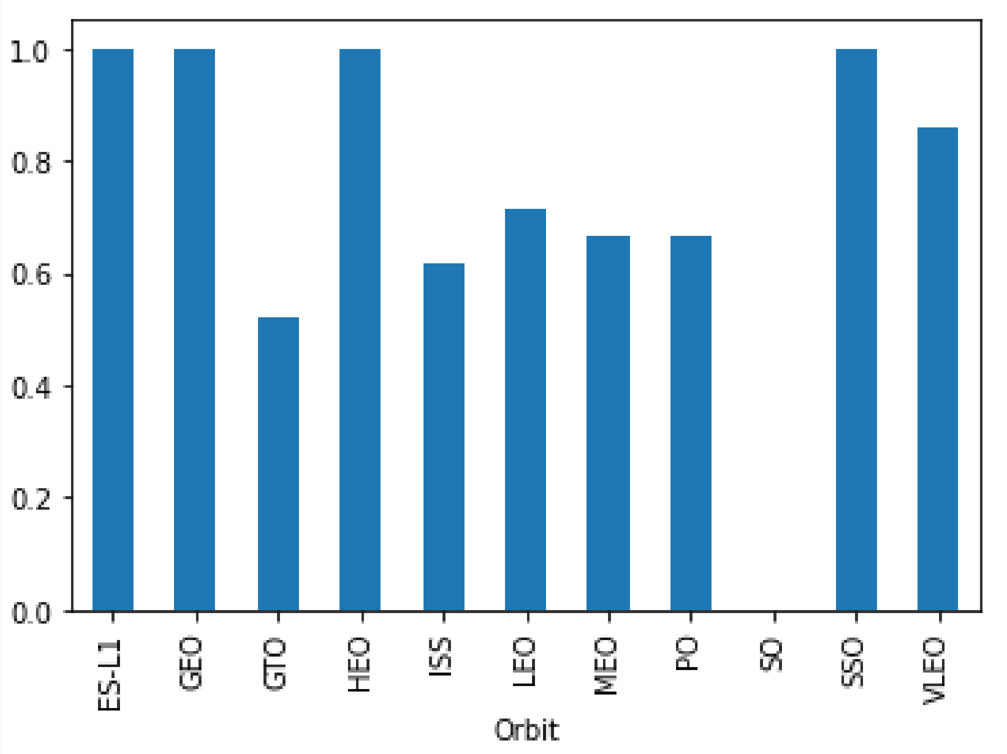
Payload vs. Launch Site



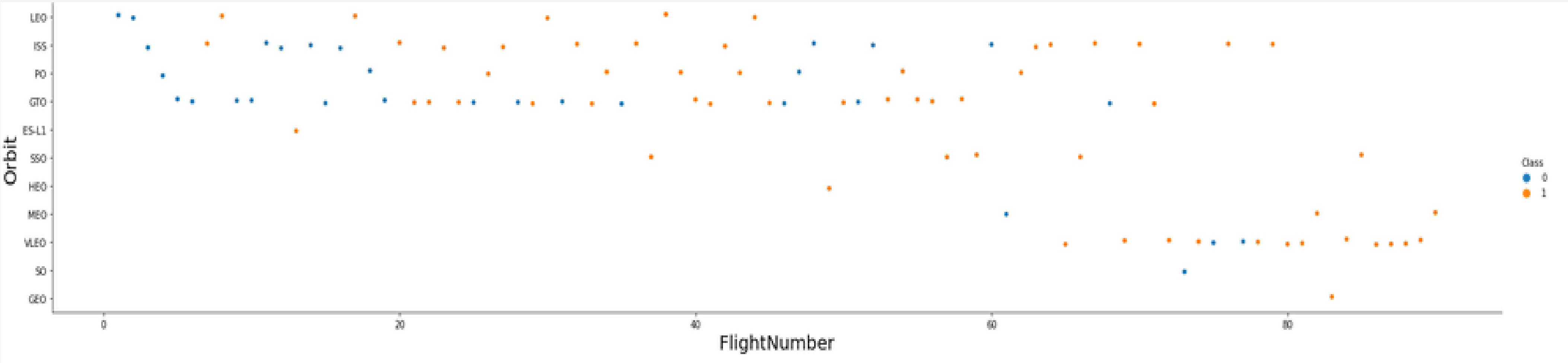
- The greater the payload mass, the higher the success rate.
- In VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).

Success Rate vs. Orbit Type

- SSO, HEO, GEO and ES-L1 have the highest success rate.
- GTO is in the mean.
- There are no success for SO.

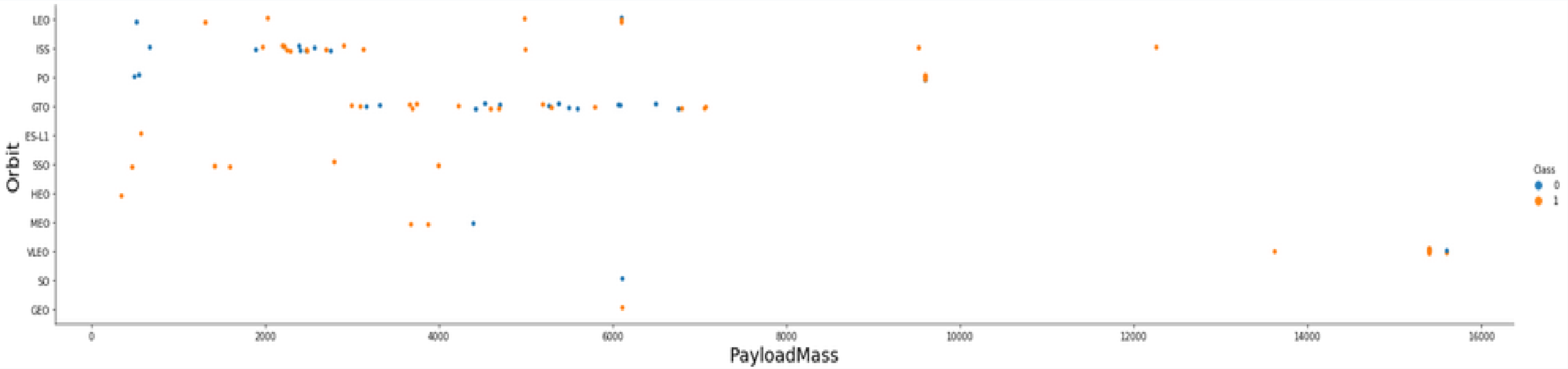


Flight Number vs. Orbit Type



- Success rate improved over the years passed for every orbit.
- LEO orbit success is related to the number of flights.
- GTO orbit have no relationship between flight number and the orbit.

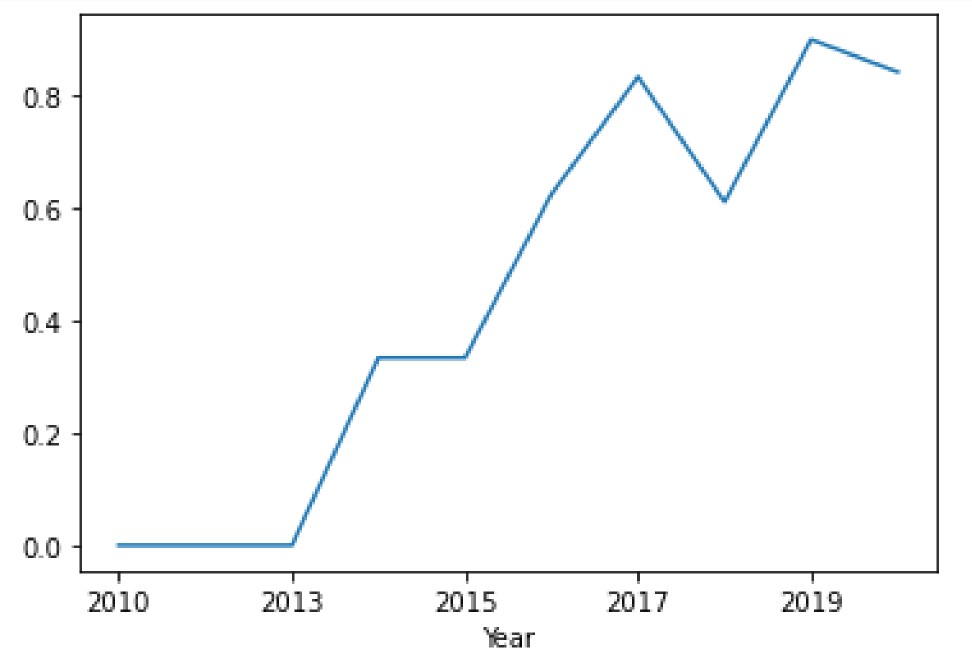
Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- 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.
- There was no success the first three years.



All Launch Site Names

- There are four unique launch sites in the data.

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

Launch Site Names Begin with 'CCA'

- The following launches where in `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	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	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)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	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 was 111.268,00 KG
- It was calculated summing all payloads from nasa.

total_payload
111268

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 was 2.928 KG
- We obtained this data filtering the booster version and calculating the avg payload.

avg_payload
2928

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad was the 22 of December of 2015.
- It was obtained by getting the successful landings and getting the minimum value.

first_success_gp
2015-12-22

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 were:
  - F9 FT B1021.2
  - F9 FT B1031.2
  - F9 FT B1022
  - F9 FT B1026
- It was obtained selecting distinct booster versions that were successful with a range of payload between 4 a 6 tons.

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

Total Number of Successful and Failure Mission Outcomes

- There was 100 successful missions
- There was 1 failed mission.
- It was obtained by grouping mission outcomes.

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

Boosters Carried Maximum Payload

- These are the names of the booster which have carried the maximum payload mass.
- It was obtained selecting the distinct booster versions where the payload was the maximum.

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

2015 Launch Records

- These were the failed landing_outcomes in drone ship with their booster versions and launch site names for year 2015.

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

- Ranking of all landing outcomes between 4/06/2010 and 20/03/2017

landing_outcome	qty
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

Section 3

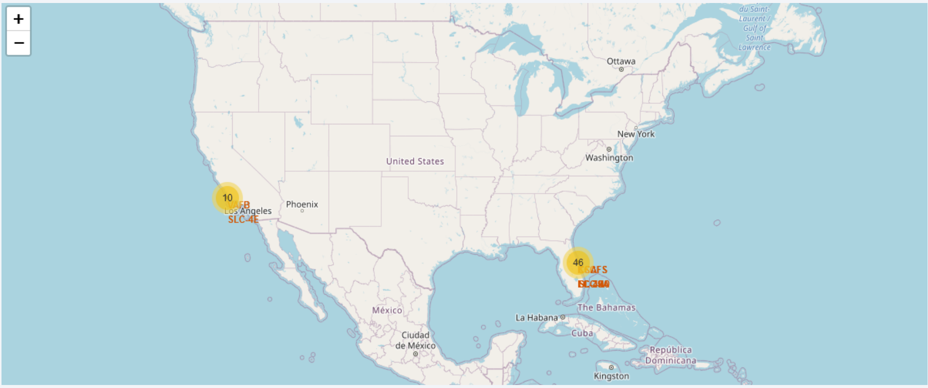
Launch Sites Proximities Analysis

Analyzed launch sites

- There are five launch sites, all in North America.
- Launch sites are located in Florida and California.
- Launch sites are near sea.

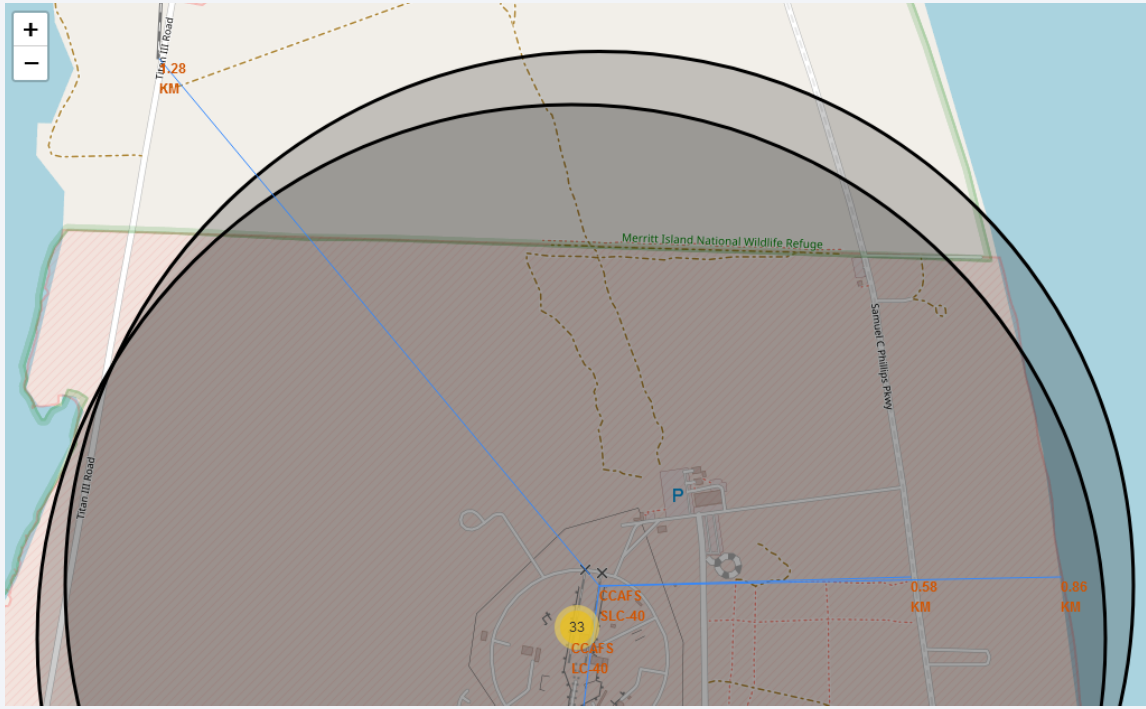
Launch sites markers and clusters

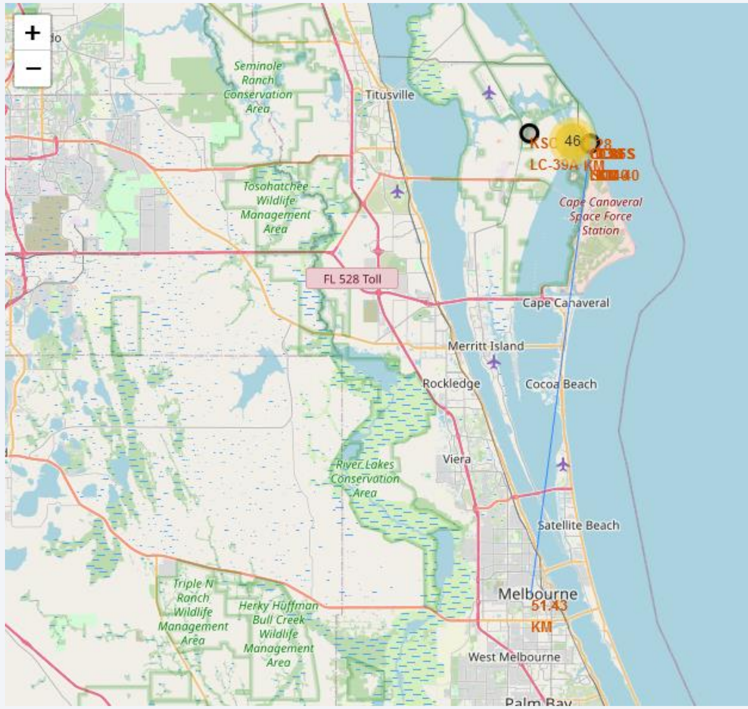
- The numbers inside the clusters are the number of launches.
- The red markers in the clusters indicate failed launches, green markers indicate successful launches.



Launch site distance to logistics

- Launchsites are far from highways or habitated areas.
- Launchsites are near coast and multiple logistics like railways or airports.

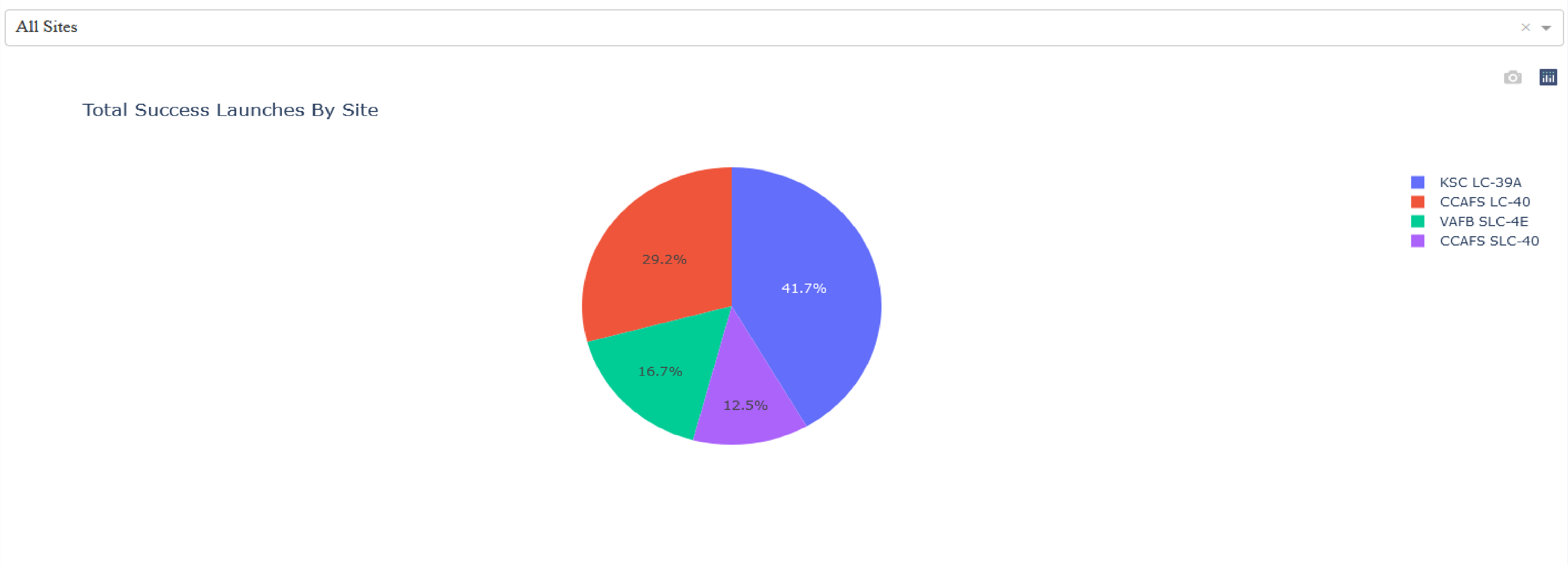




Section 4

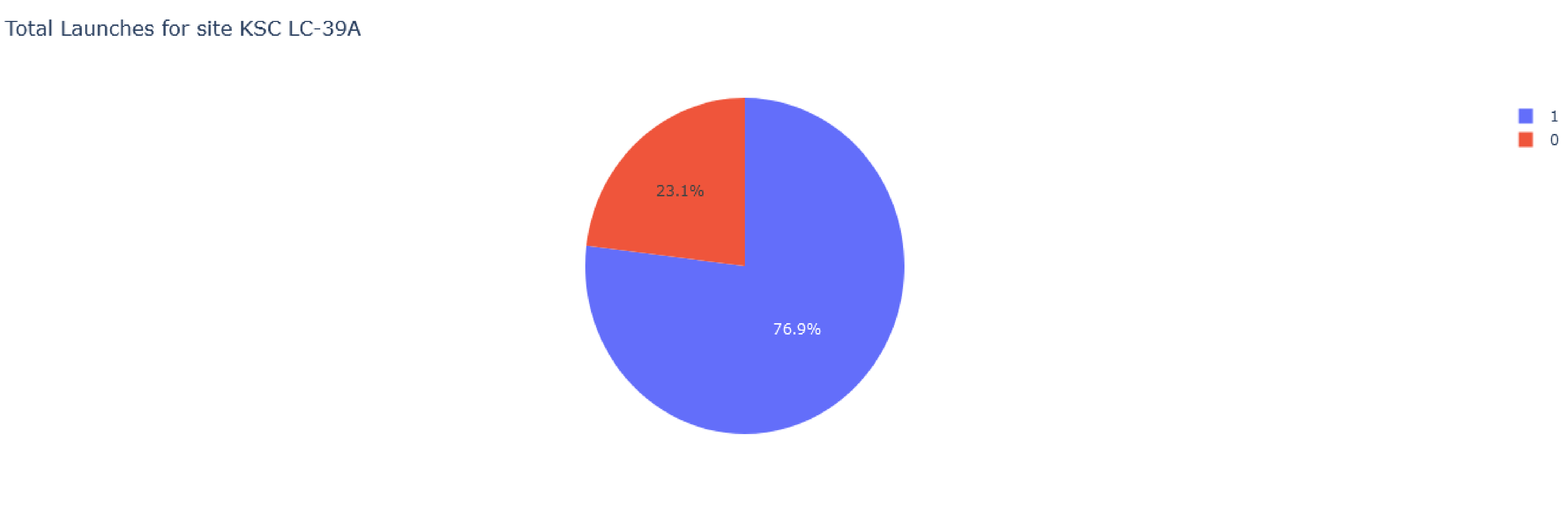
Build a Dashboard with Plotly Dash

Total success launches by all sites



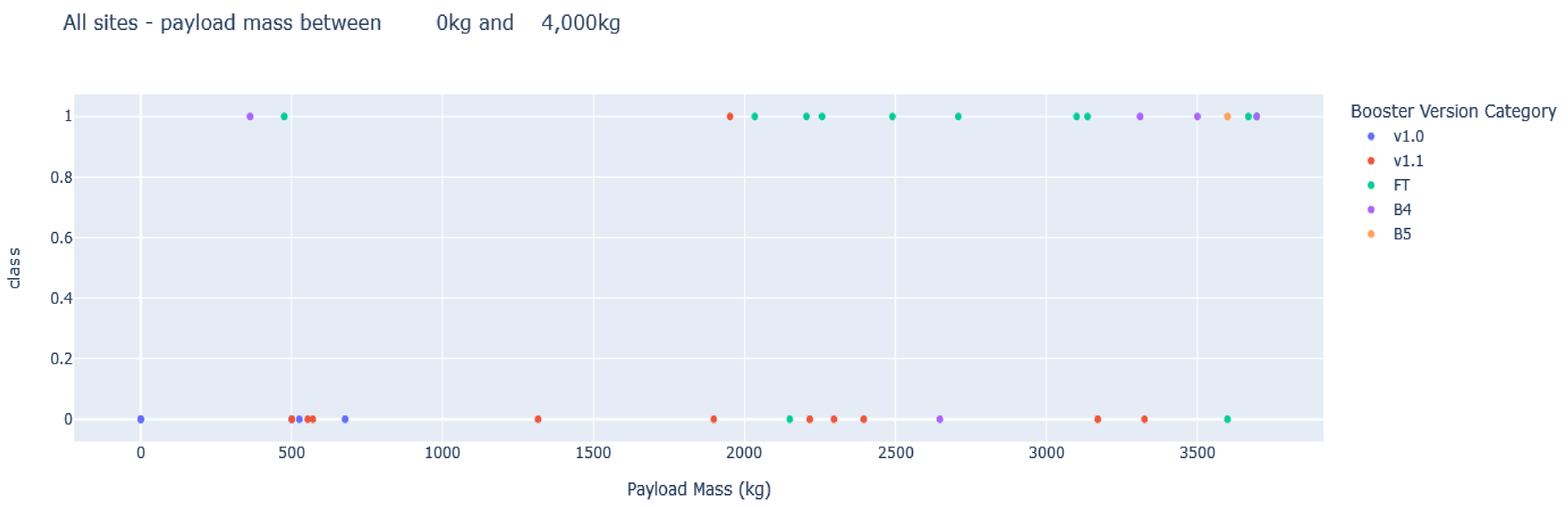
- KSC LC-39A has the most successful launches.

Total launches success for KSC LC-39A

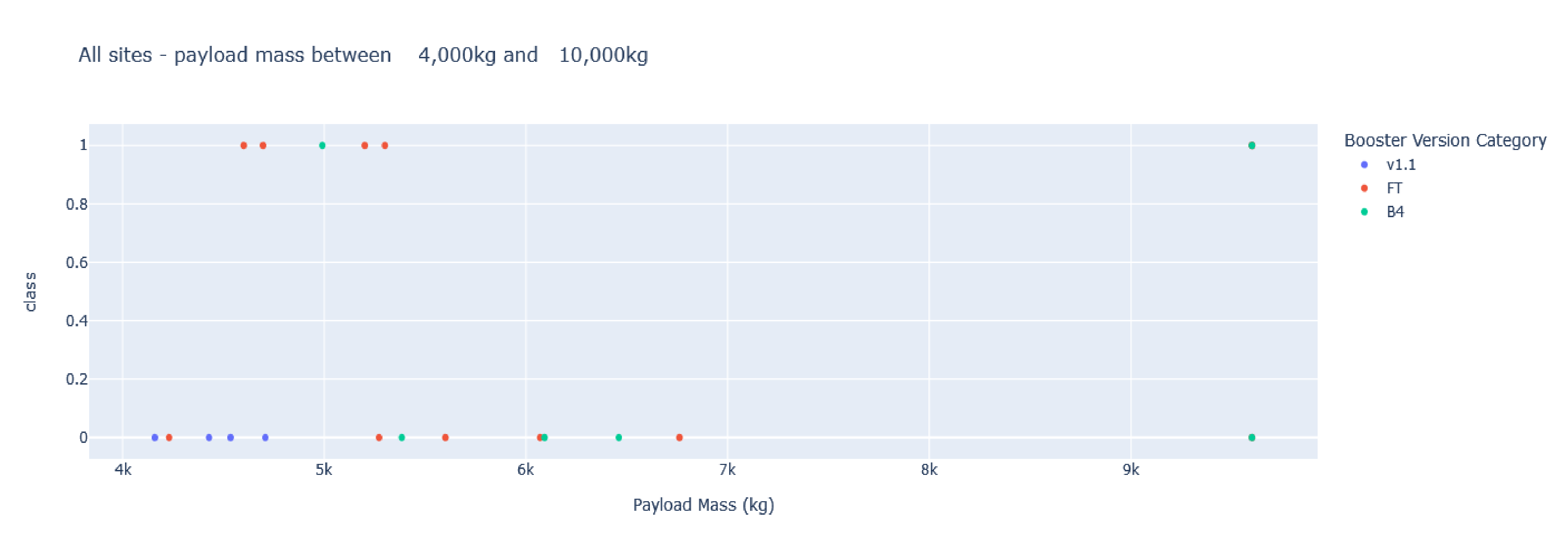


- KSC LC-39A has a 76,9% success

Low Weighted Payload



Heavy Weighted Payload



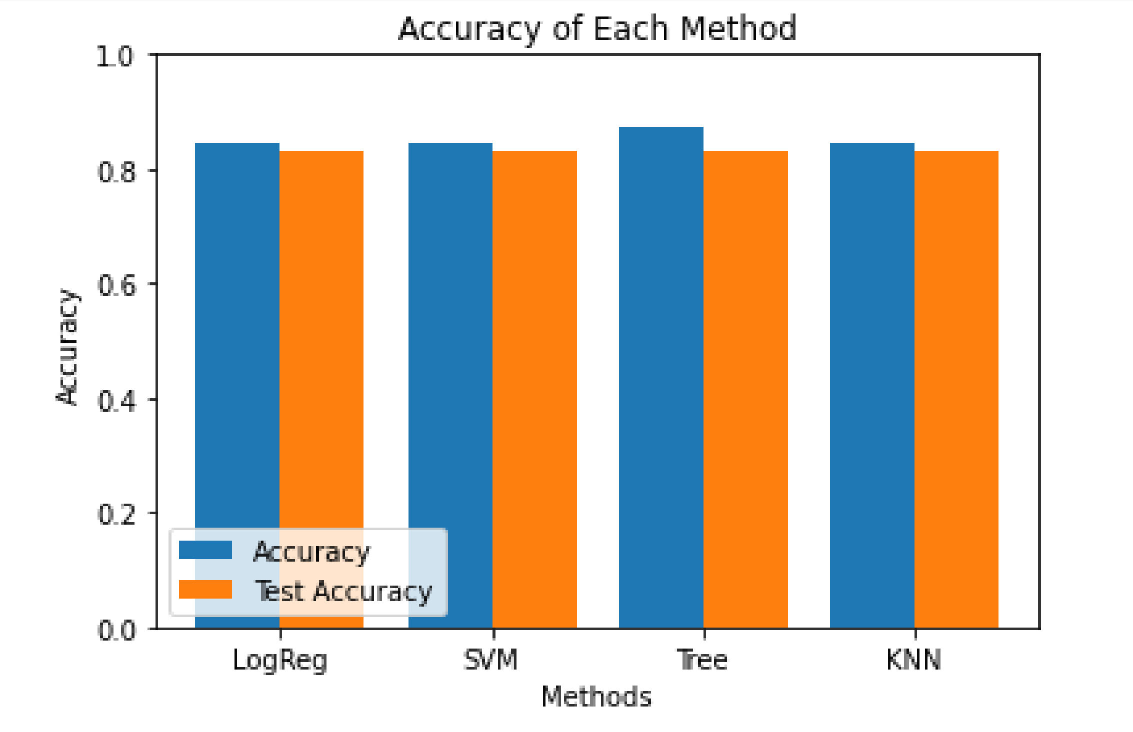
- The success rate for low weighted payloads is higher than the heavy ones.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The classifiers used were: Logistic regression, KNN, Support Vector Machine and Decision tree.
- The highest accuracy is obtained by Decision Tree Classifier, which 87,5%.



Confusion Matrix

- The confusion shows that the classifier has no problems determining False negatives and True positives, but it has problems determining False positives.

Conclusions

- The landings success increased through time as more flights were produced and they were perfectioning their methods.
- The best launch site Is KSC LC-39A.
- The success rate for low weighted payloads is higher than the heavy ones
- Training a machine learning model with Decision Tree Classifier is the best option to predict a successful landing in the next mission.

Thank you!