IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jonas Ziegler
02.02.2024

Outline

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

Executive Summary

- Publicly available data from SpaceX API and Wikipedia was gathered
- Explorative Data Analysis shows
  - different launch sites mostly in Florida were used
  - different landing sites and methods (drone ship, ground pad, parachute) were used
  - Strong learning curve over several years, first evidence of possible outcome influence by payload mass and orbit
- Machine Learning models can be trained and achieve upto 88,8% accuracy in predictions

Introduction

- Space launches are very expensive, to reduce costs companies like SpaceX try to land and reuse the first stage of the booster
- To predict a landing outcome and thereby the associated cost of a launch we want to train machine learning on publicly available data.

Section 1

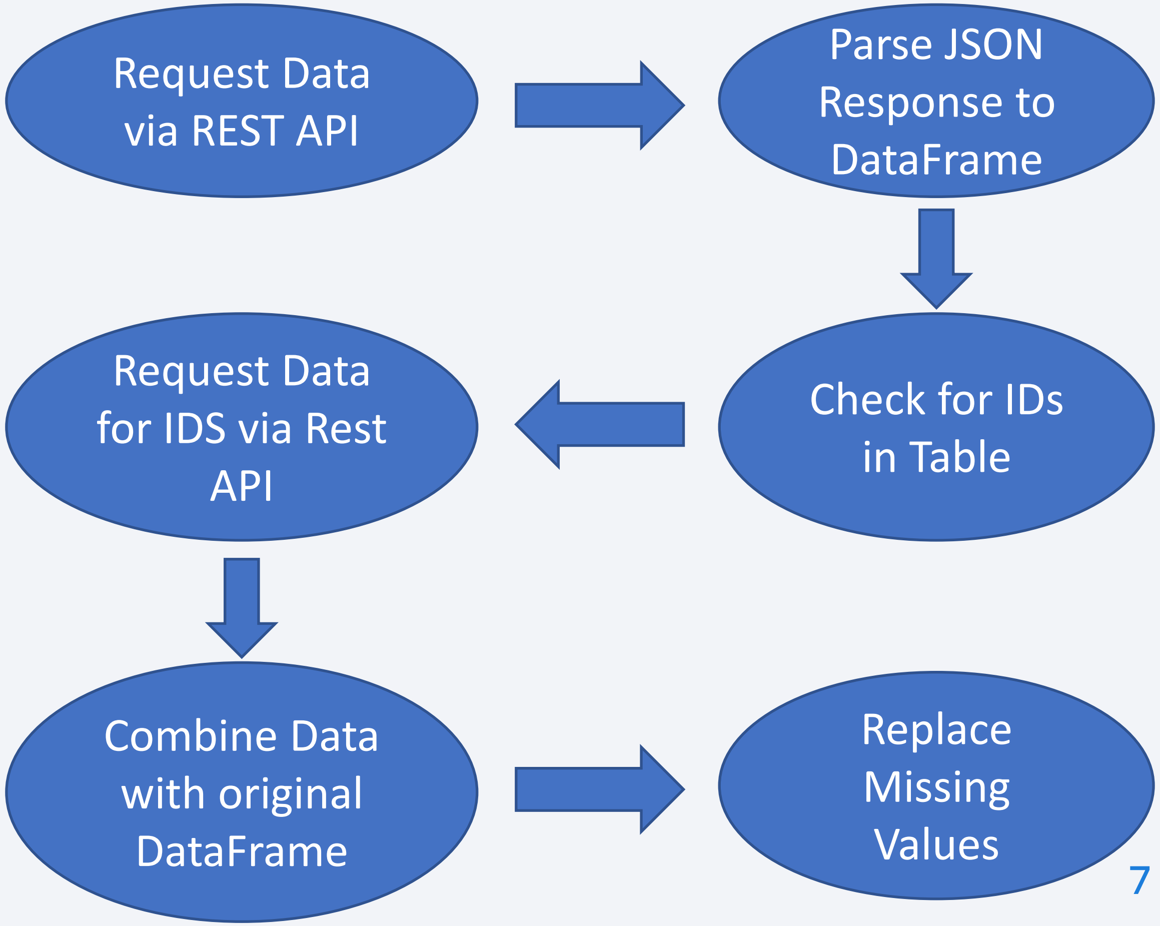
Methodology

Data Collection

- To gather data we use two different sources
  - SpaceX Rest API as JSON
  - Wikipedia Article as HTML
  - Convert Data to DataFrame

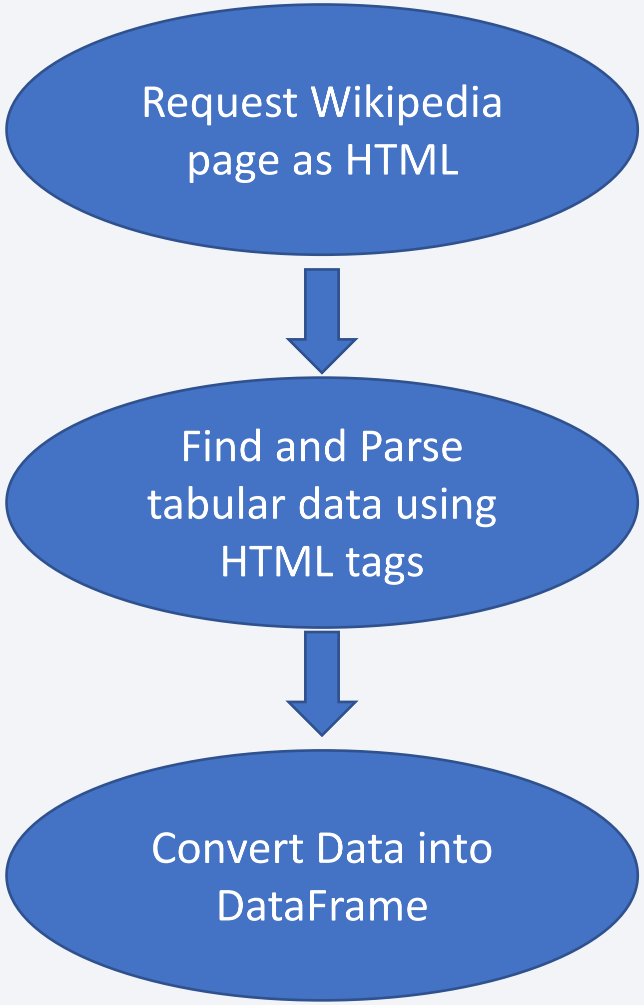
Data Collection – SpaceX API

- Via the SpaceX REST API we can access historic launch data
- We parse the JSON response to create a pandas DataFrame
- A lot of the information about booster versions, launch sites etc. are not directly included, instead IDs are used as a point of reference
- After combining the data for these IDs we replace missing data where necessary
- GitHub: [ADSC/jupyter-labs-spacex-data-collection-api.ipynb at main · iceboy910447/ADSC (github.com)]



Data Collection - Scraping

- Using the http Requests we scrap a Wikipedia page with a table containing information about SpaceX launches
- Using Beautiful soup we find the the table by its HTML tags and sequentially extract headers and row by row each cell info
- Convert Data into DataFrame
- GitHub: [ADSC/jupyter-labs-webscraping.ipynb at main · iceboy910447/ADSC (github.com)]



Data Wrangling

- After the data collection the data was preprocessed
- Categorical value counts were calculated for
  - launch sites
  - Orbits
  - Landing Outcomes
- Additionally a new parent category for the landing outcome was generated
- GitHub:ADSC/labs-jupyter-spacex-Data wrangling.ipynb at main · iceboy910447/ADSC (github.com)

EDA with Data Visualization

- To allow a better visual understanding of the data, several types of graphs were used
  - line plots for trends
  - scatter plots to search for correlations
  - bar charts to compare differences between groups or categories
- Using found correlations allowed to define features for data preparation
- GitHub: ADSC/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb at main · iceboy910447/ADSC (github.com)

EDA with SQL

- To get a better understanding of the data descriptive SQL-commands were used, to get unique values for landing sites, the sum of payload and or the first (minimum) date of a successful ground pad landing.
- To get further insight into details filters using the WHERE clause, BETWEEN and wildcard phrases were used to get better insight
- For the most complex insights subquries and unions of results were used to combine results.
- GitHub: [ADSC/jupyter-labs-eda-sql-coursera_sqllite.ipynb at main · iceboy910447/ADSC (github.com)]

Build an Interactive Map with Folium

- To explore the geospatial information for all launches, a folium map was build
- The positions of each launch site was visualized on a map using circles
- To visualize each start and landing outcome, green and red markers were used. Due to the number of landings on prevent overcrowding on the map, a marker cluster was used
- Additional lines to show the distance to important points of interests, like oceans, railways and streets as well as cities were added.
- GitHub: [ADSC/lab_jupyter_launch_site_location.jupyterlite.ipynb at main · iceboy910447/ADSC (github.com)]

Build a Dashboard with Plotly Dash

- To allow better visualization of percentages of successful outcomes, depending on the launch site, payload and booster version a piechart and scatter plot was created
- The possibility for the user to define a payload mass range allows for a detailed analysis of specific aspects in the scatter plot
- Statistics for all launch sites can be further explored for a specific launch site in the set
- GitHub: ADSC/spacex_dash_app.py at main · iceboy910447/ADSC (github.com)

Predictive Analysis (Classification)

- Using scikit-learn 4 different models were generated and trained
- The available data was splitted into training and test data (80%/20%)
- To optimize the models, a hyperparameter optimization was performed using gridsearch with crossvalidation
- To compare the models, the accuracy of each modell was calculated on the test data and compared between all models
- Confusion matrices were generated to inspect underlying performance values like false positive and false negative percentages
- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose:
- ADSC/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite(1).ipynb at main · iceboy910447/ADSC (github.com)

Results

- Visual and numerical exploration of the data
- Interactive tool for exploration in more detail to uncover hidden influences
- Machine Learning models to predict the outcome
- Evaluation of the usability of machine learning on such a problem

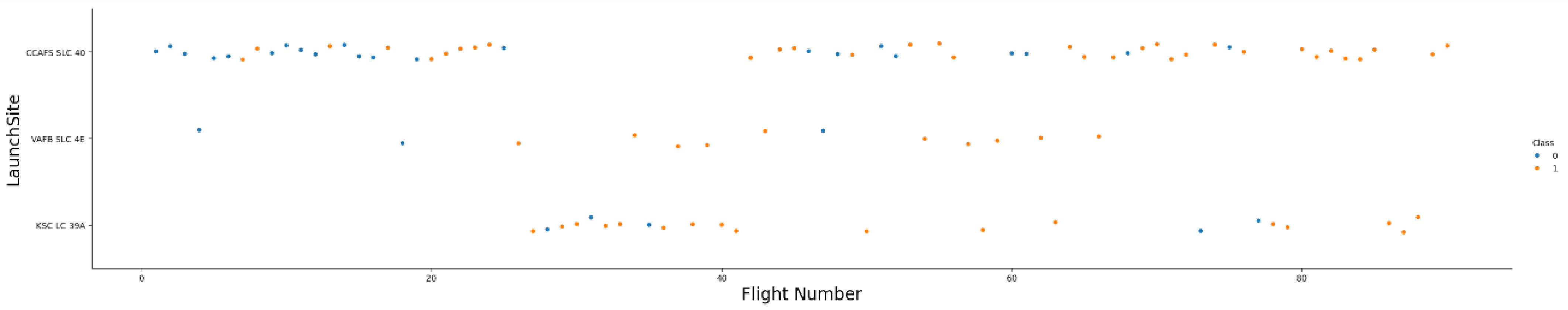
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

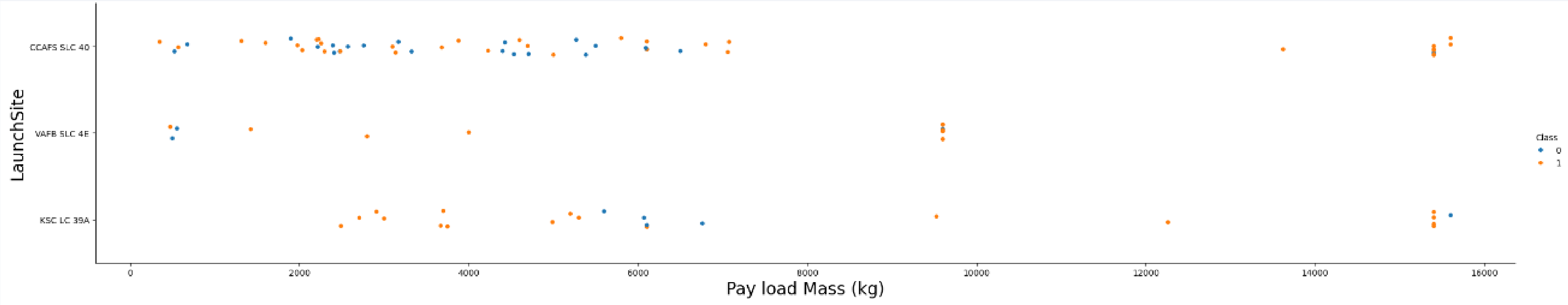
Looking at the usage of different launch sites over all flights shows 2 key results

- Some launch sites are used preferred
- Some launch sites were not used right from the beginning, only later in the history of the company



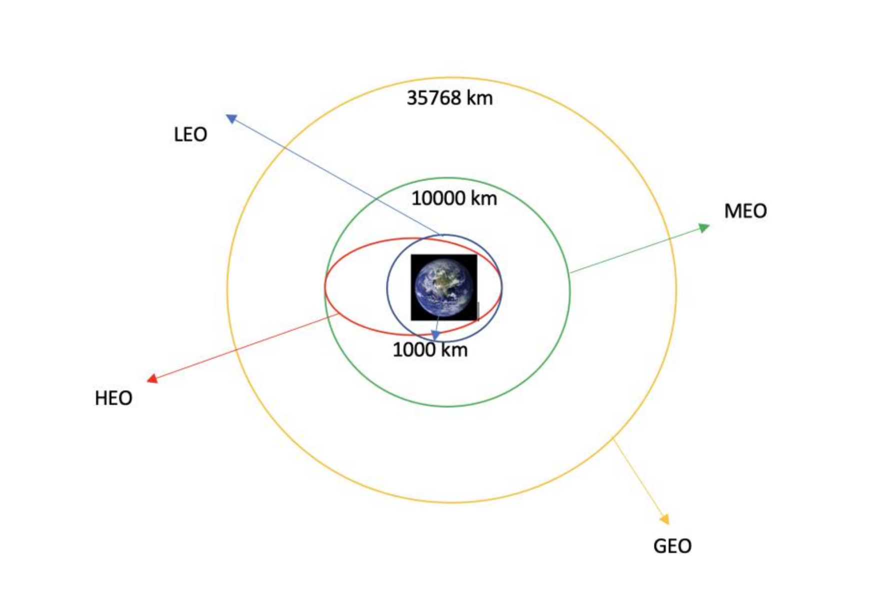
Payload vs. Launch Site

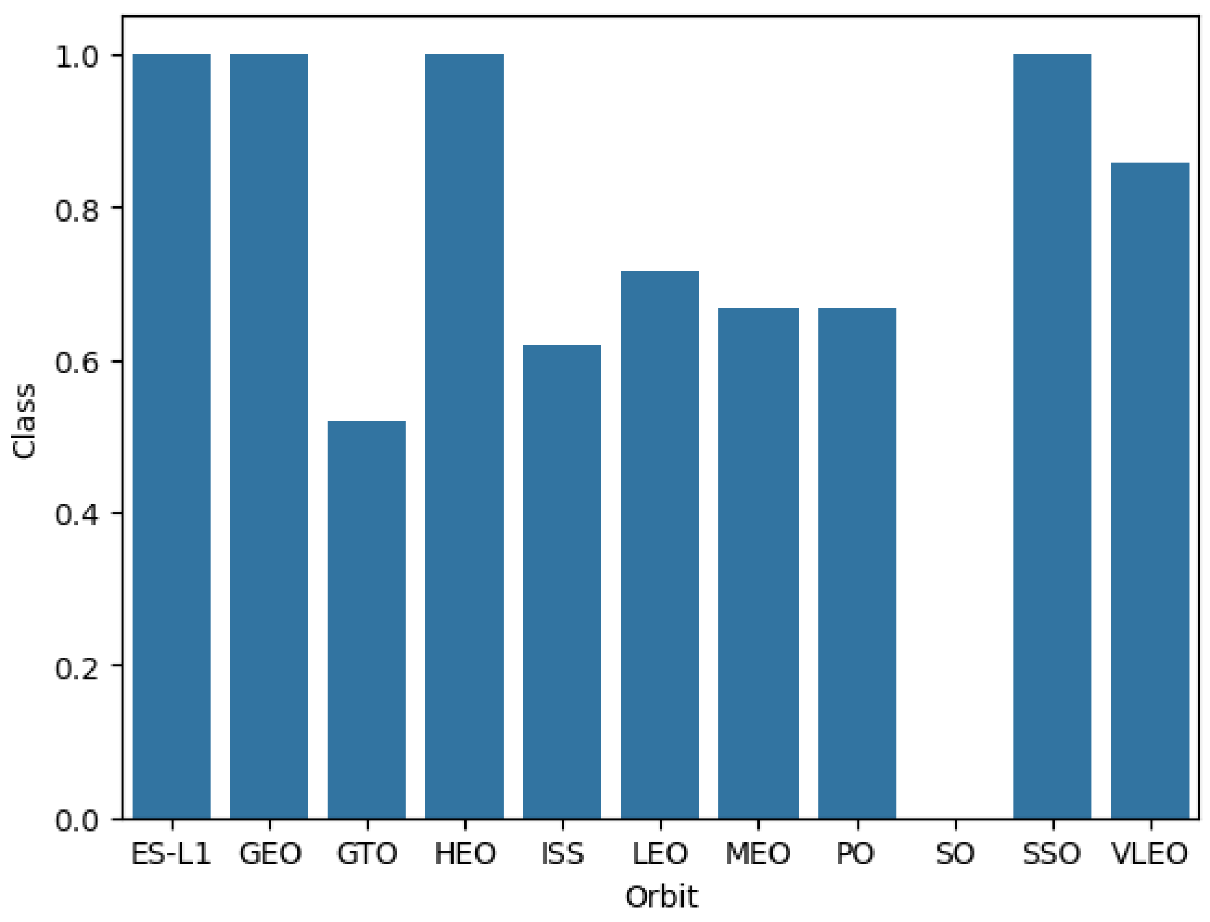
- Visual comparison of the payload mass for different launch sites shows that not only were most starts conducted in Cape Canaveral and the Kennedy Space Center, only few at the Vandenberg Air Force Base, but also that the maximum payload at the Vandenberg Air Force Base is significantly lower in comparison (less than 10000kg)



Success Rate vs. Orbit Type

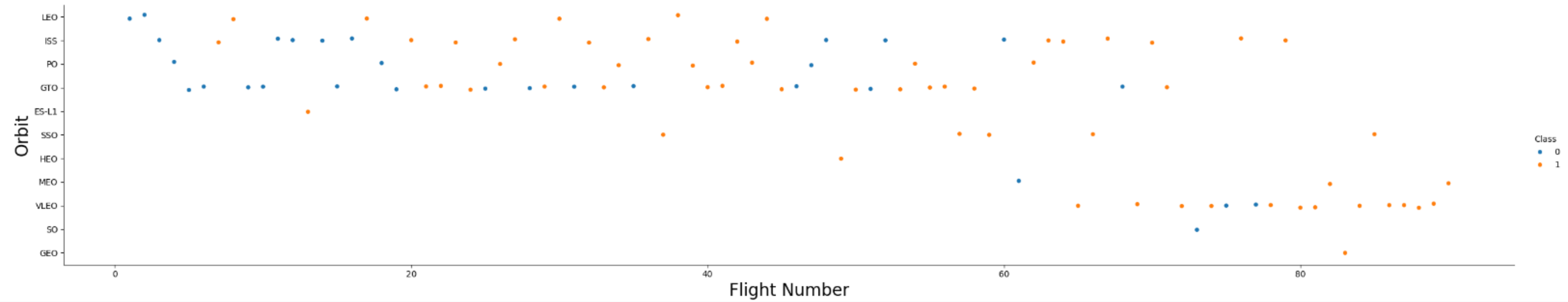
- Success rate differs between orbits, but no clear proof of causal relationsship





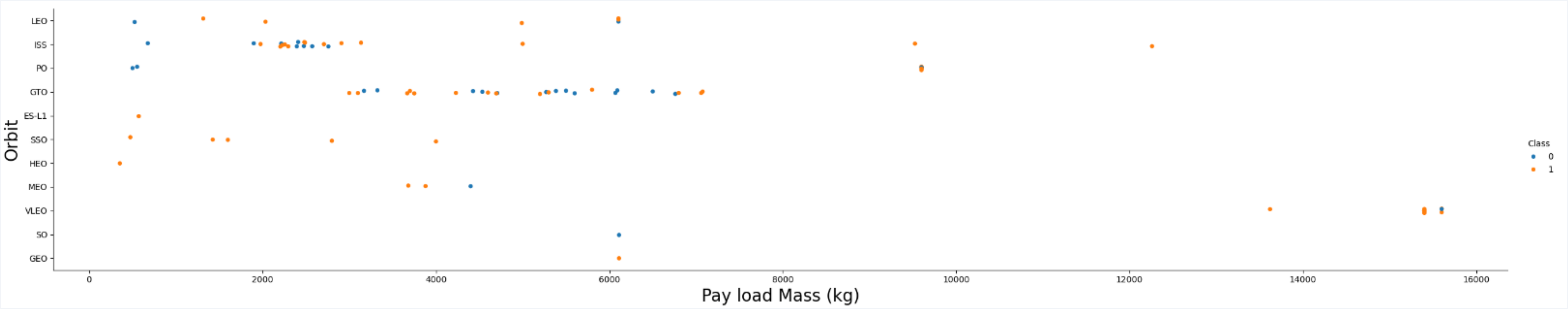
Flight Number vs. Orbit Type

Plotting the different outcomes for different orbits over all flights shows that SpaceX started with a few orbits, had considerably more failures in the first phase (blue dots) and later expanded to other orbits like GEO, which is considerably higher than LEO. Therefore a learning curve over all flights can be assumed, weakening the evidence for a causal relationship between orbit and outcome, as suggested in the bar chart



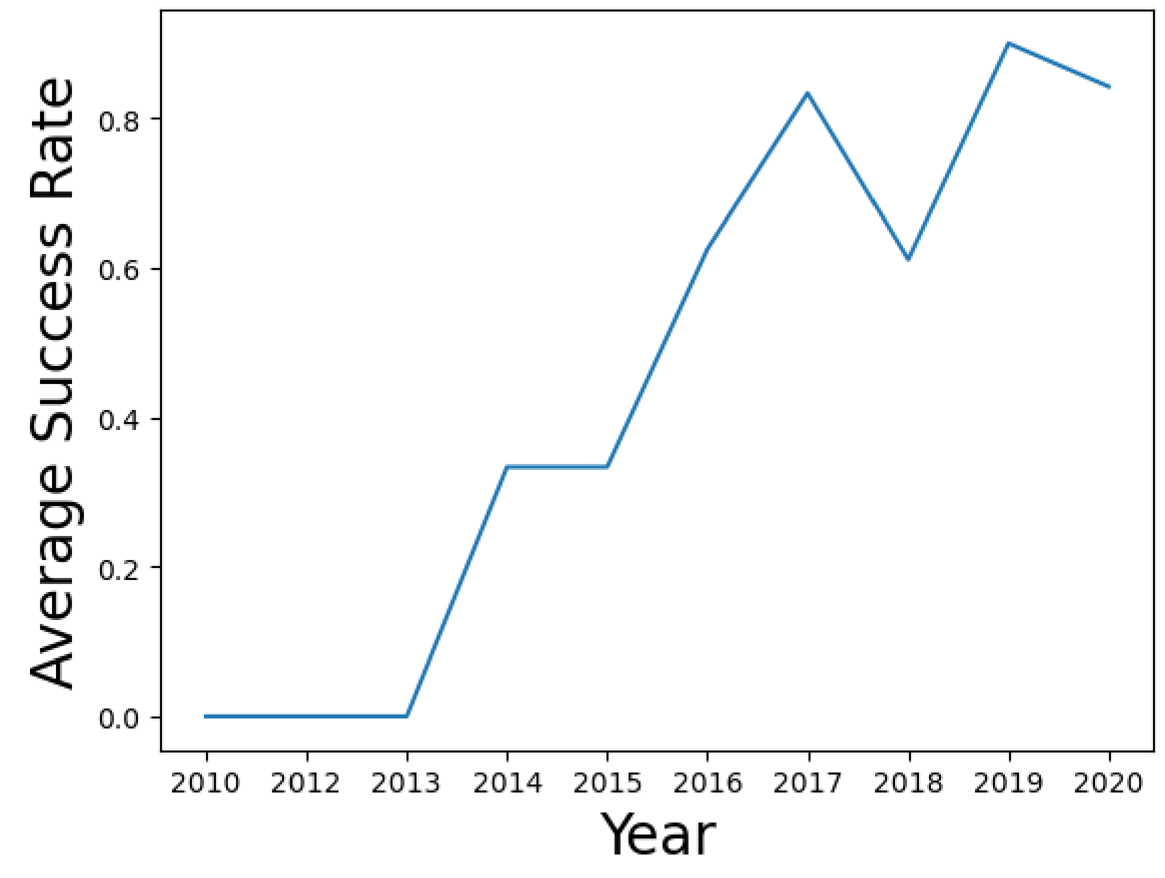
Payload vs. Orbit Type

The scatter plot between Orbit and Payload Mass also shows, that different orbits were used with different average and maximum weights for the payload.



Launch Success Yearly Trend

- The yearly average success rate shows a clear upwards trend (learning curve)
- Between 2010 and 2013 the average success rate was close to 0%. After a steady increase between 2013 and 2017 the average success rate dropped from over 80% to about 60% in 2018.
- The average success rate peaked in 2019 at around 85% and stablized above 80% in 2020.



All Launch Site Names

- The data collected shows 4 different launch sites
  - Two at Cape Canaveral
  - One at Vandenberg Air Force Base
  - One at Kennedy Space Center

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

Launch Site Names Begin with 'CCA'

- The 2 launchsites at Cape Canaveral also share a similarity in their name. Filtering the name by using wildcard characters allows getting combined results for the facility

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

- In total SpaceX launches carried a total payload of 45596 kg

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Given the total payload over all launches, the average weight amounts to 2534kg
- As shown in slide 21 and 22 there is a clear increase of the maximum payload over time, as well as a clear evidence for different payload mass averages for different orbits like ISS and GTO

SUM(PAYLOAD_MASS__KG_)/COUNT(PAYLOAD_MASS__KG_)
2534

First Successful Ground Landing Date

- The first successful ground landing by SpaceX was achieved in December 2015

MIN("Date")
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- SpaceX did not only attempt to land on ground, but also on drone ships in the ocean.
- Different Booster Versions of the F9 Falcon landed successfully on a drone ship after delivering a payload between 4000kg and 6000kg into space

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

Total Number of Successful and Failure Mission Outcomes

- Over all mission launches, 100 missions were officially a success, only 1 mission
- Successful missions don‘t require a successful landing

Outcome	Count
Failure	1
Success	100

Boosters Carried Maximum Payload

- Overall multiple booster versions as shown on the right side achieved to deliver the maximum recorded payload mass of ~16000kg since it’s first launch

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

- In 2015 two landing on drone ships failed, one in January and on in April. Both started from the same launch site

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

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

- Between June 2010 and March 2017 in 21 cases of 31 starts attempted a landing of the booster
- Attempted landings were tried on 3 different types of targets
  - drone ships
  - ground pad
  - ocean
  - parachute
- The most attempts tried to land on a drone ship, of which 50% ended successful (5 of 10 attempts)

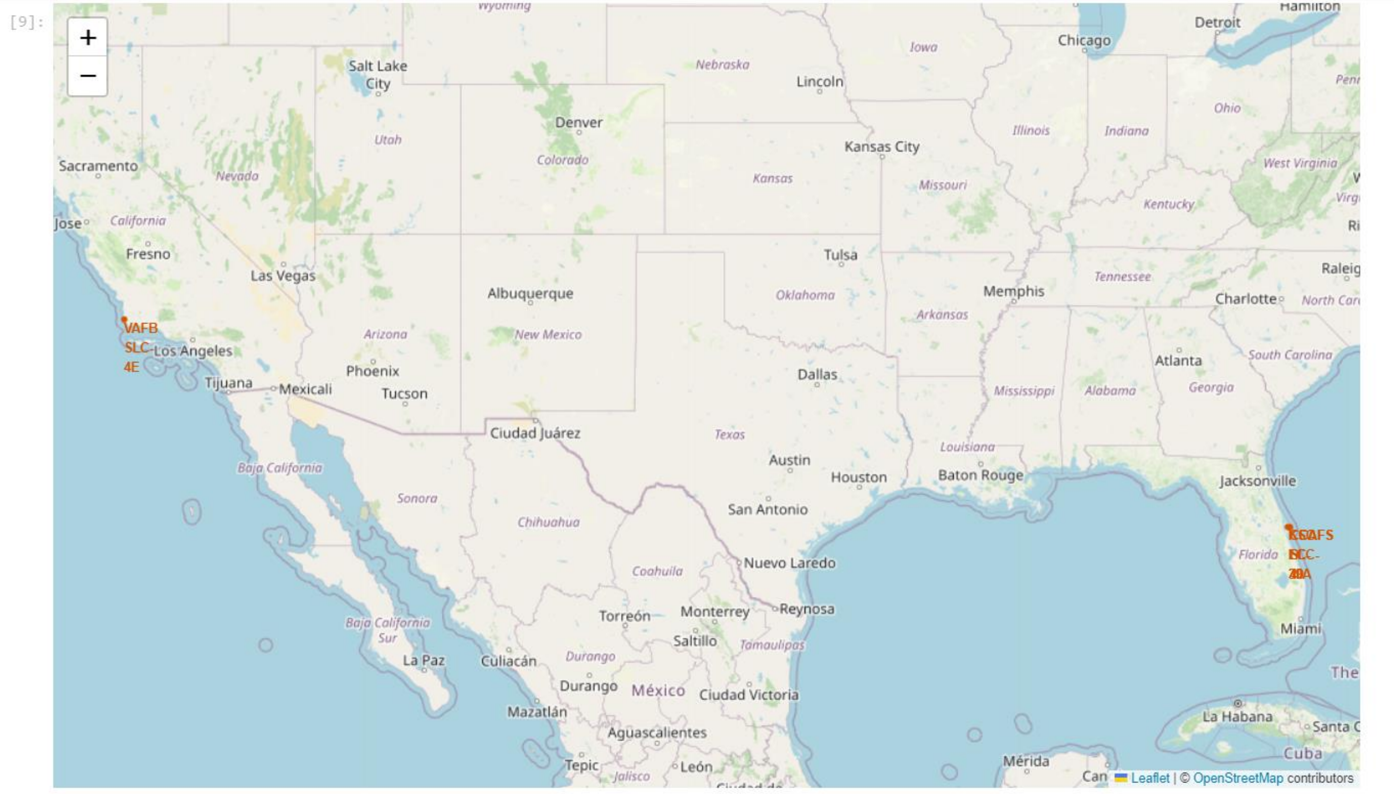
COUNT	Landing_Outcome
10	No attempt
5	Success (drone ship)
5	Failure (drone ship)
3	Success (ground pad)
3	Controlled (ocean)
2	Uncontrolled (ocean)
2	Failure (parachute)
1	Precluded (drone ship)

Section 3

Launch Sites Proximities Analysis

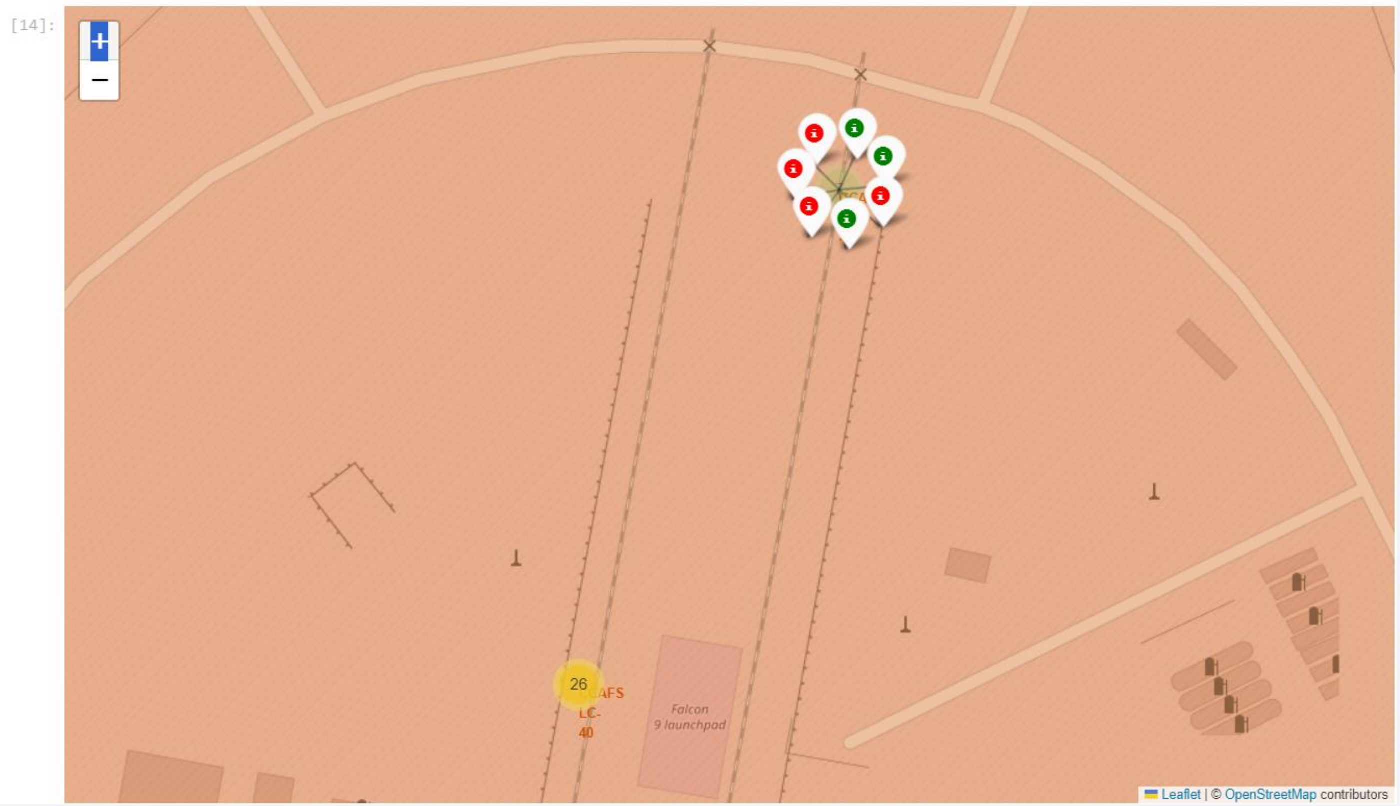
Exploring Launch Site Locations on a Map

- 3 of 4 launch sites are in Florida, one in California, all close to the ocean



Mapping Outcomes

- To show the outcome of each launch at each site, colorcoded markers were set, and combined into clusters, to prevent overcrowding. No clear relationship between outcome and site was found



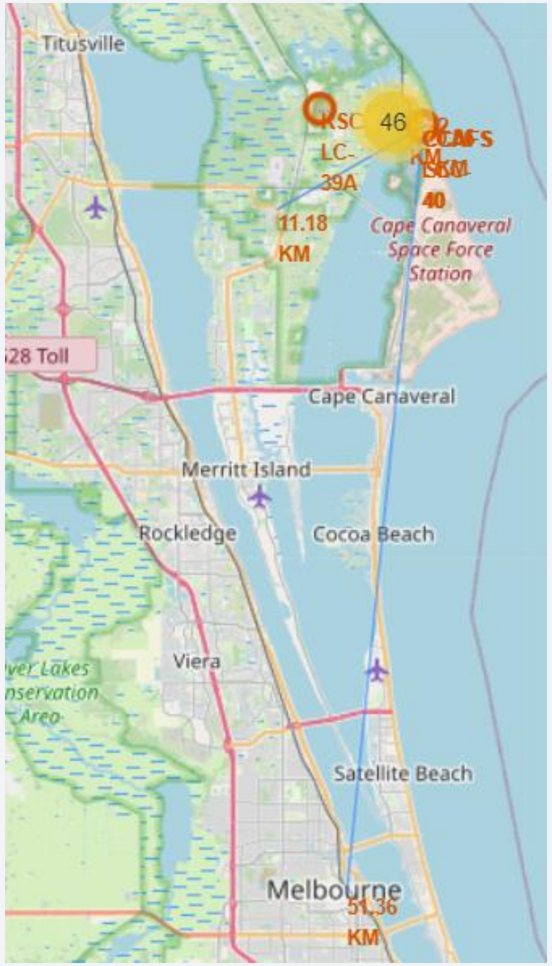
Visualizing Distance to Infrastructure

Launch sites like CCA seem to be build close to

- the ocean (~900Meters)
- railways (1.2km),

but far from

- Highways(~11km)
- airports and cities (~51km)

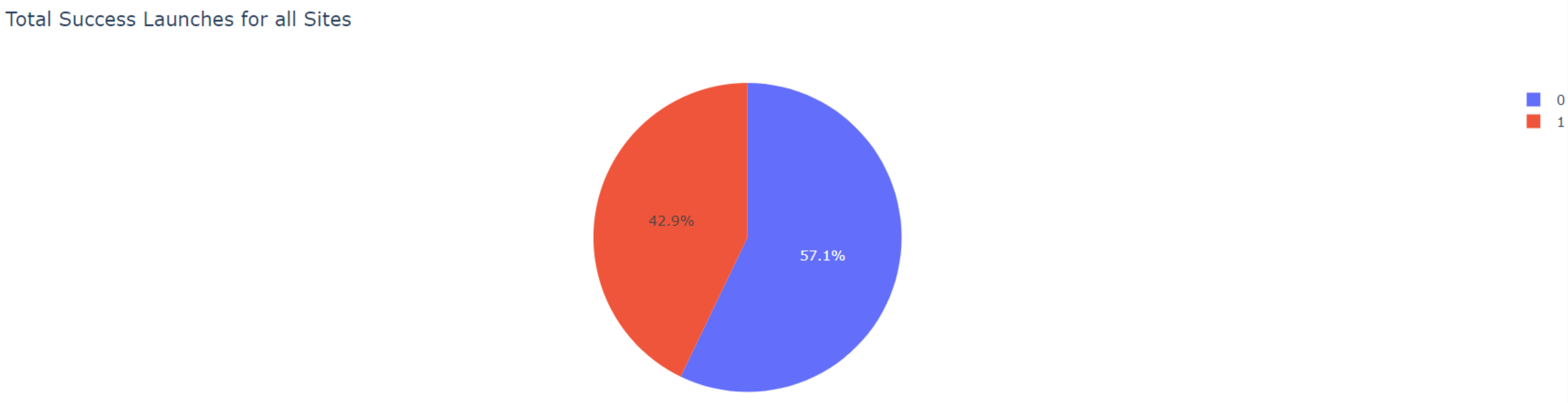


Section 4

Build a Dashboard with Plotly Dash

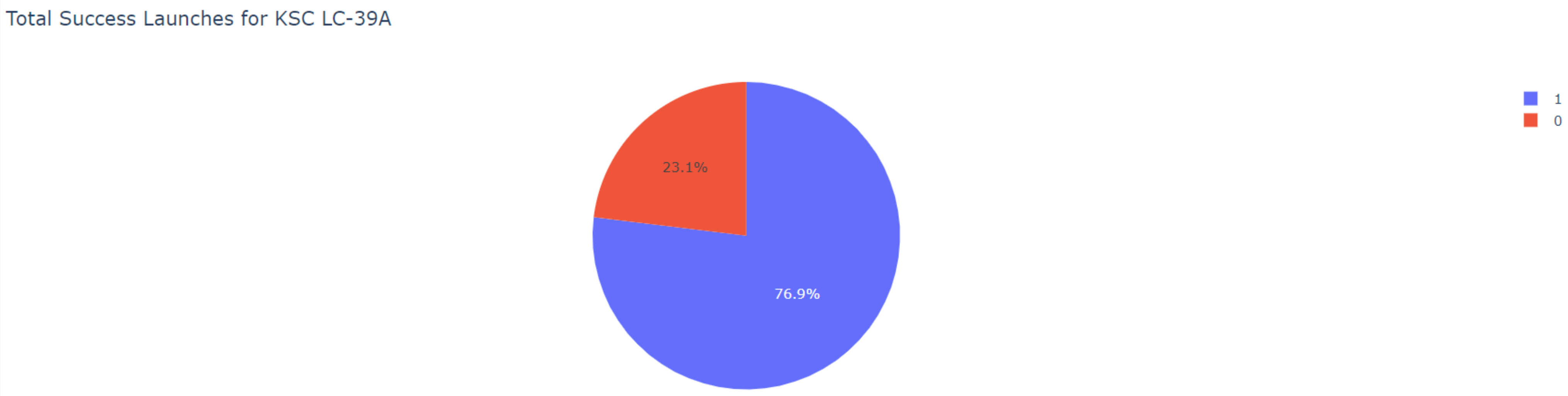
Interactive exploration of success for different launch sites

- Over all launches the total success rate is 57,1%



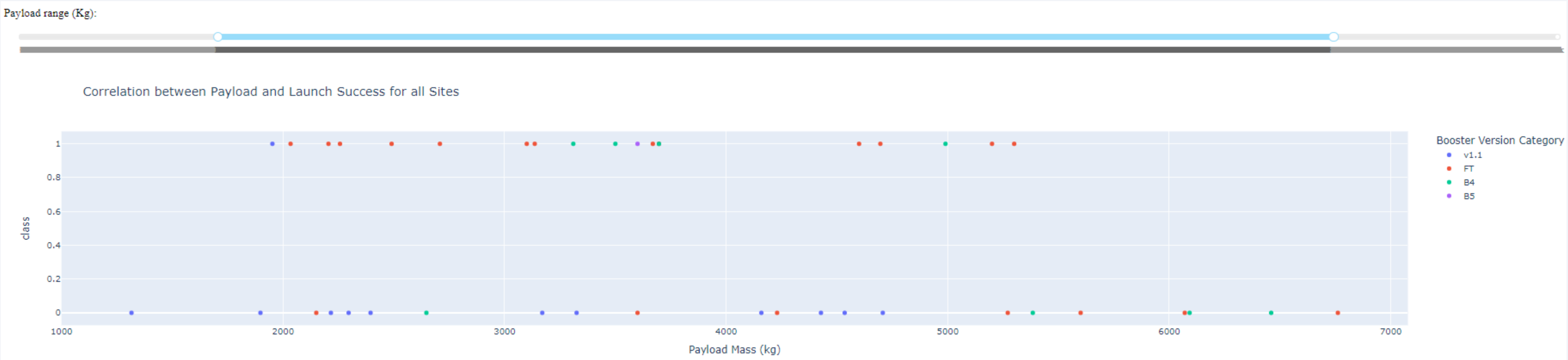
Landing site with best outcome statistic

- The best overall success rate was achieved for launches at the Kennedy Space Center with 76,9%



Influence of Booster Version and Payload Mass on Outcome

- While the v1.1 Booster version has a low success rate for landings, the FT Booster version not only achieves a higher success rate for payloads between 1000-5000kg, but also can deliver payloads with a mass greater 5000kg

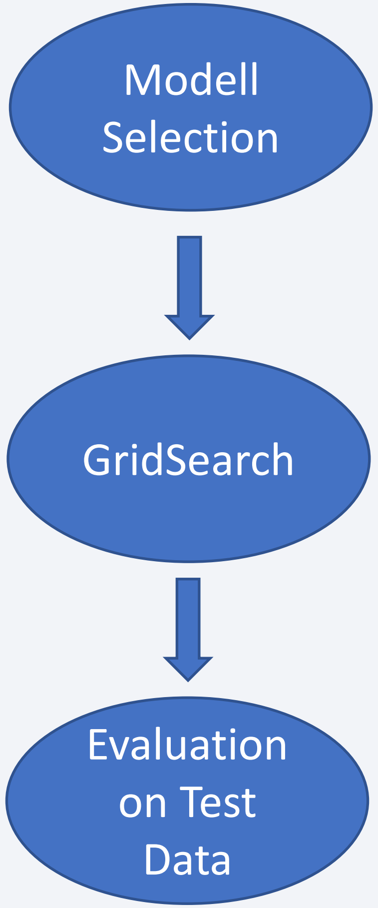


Section 5

Predictive Analysis (Classification)

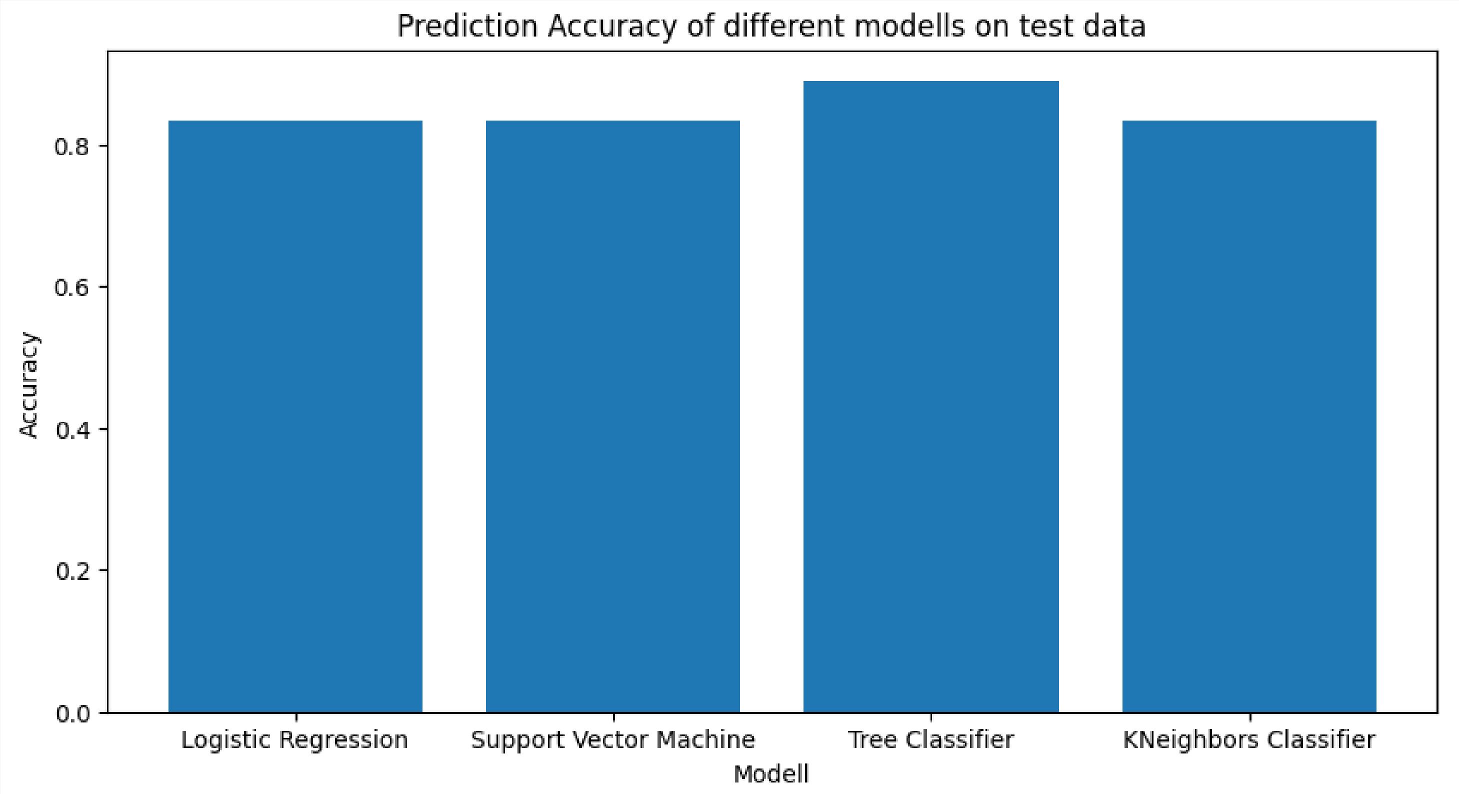
Modell training

- 4 different machine learning models for classification were compared:
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree Classifier
  - K-Nearest Neighbors Classifier
- Available data splitted:
  - 80% for modell training
  - 20% for modell testing
- The Hyperparameters of all modells were optimized:
  - GridSearch of Hyperparameters
  - Crossvalidation using 10 folds



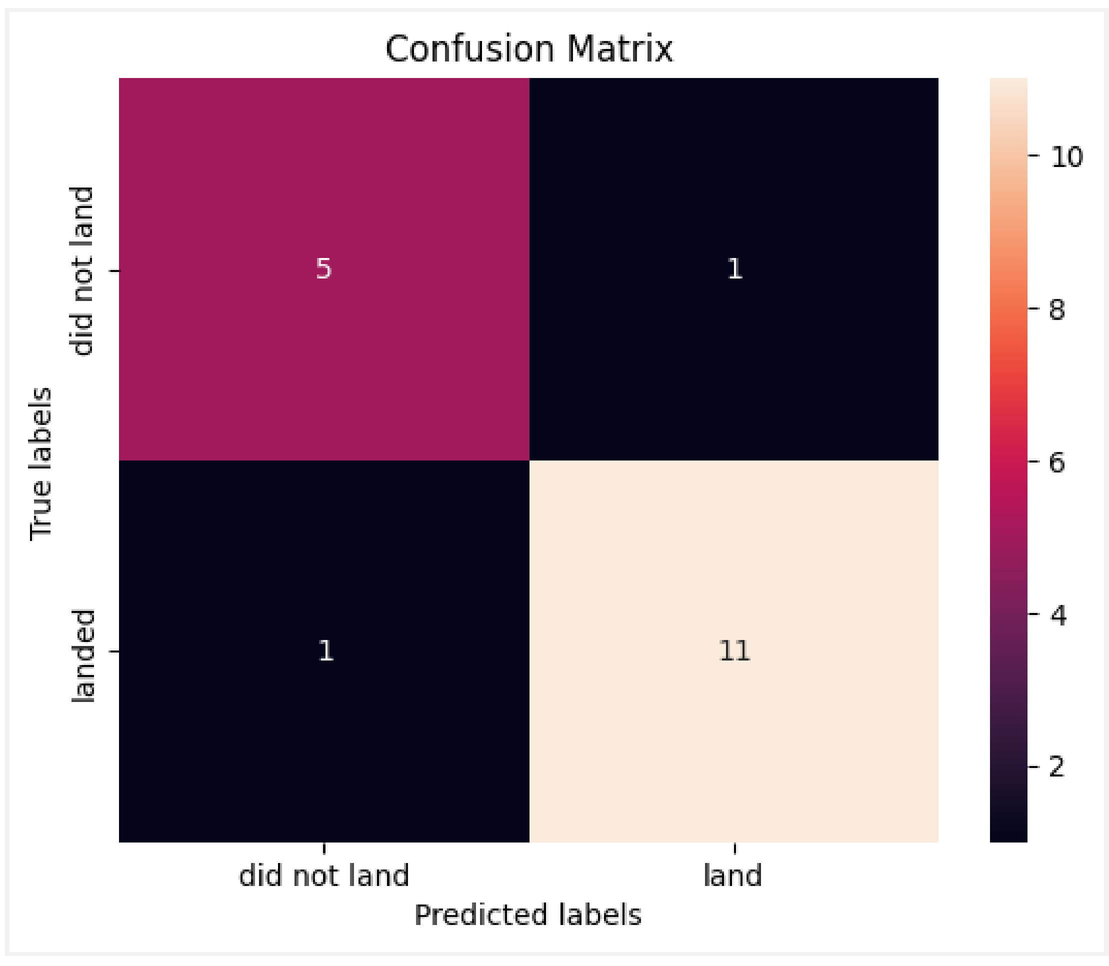
Classification Accuracy

- Logistic Regression, Support Vector Machine and KNeighbors Classifier performed identical on the test data (83,3% accuracy)
- Decision Tree Classifier performed best on the test data
(88,8% accuracy)



Confusion Matrix

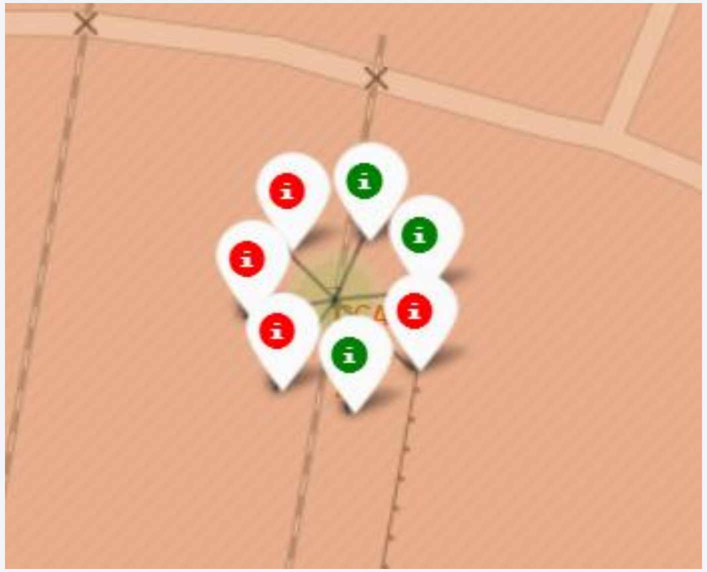
- The best modell predicted 88% of the cases correctly.
- Only in 1 of 12 cases where the landing ended successfully the modell would predict a failure (false negative)
- Only in 1 of 6 cases where the landing failed, the modell would predict a successful landing (false positive)



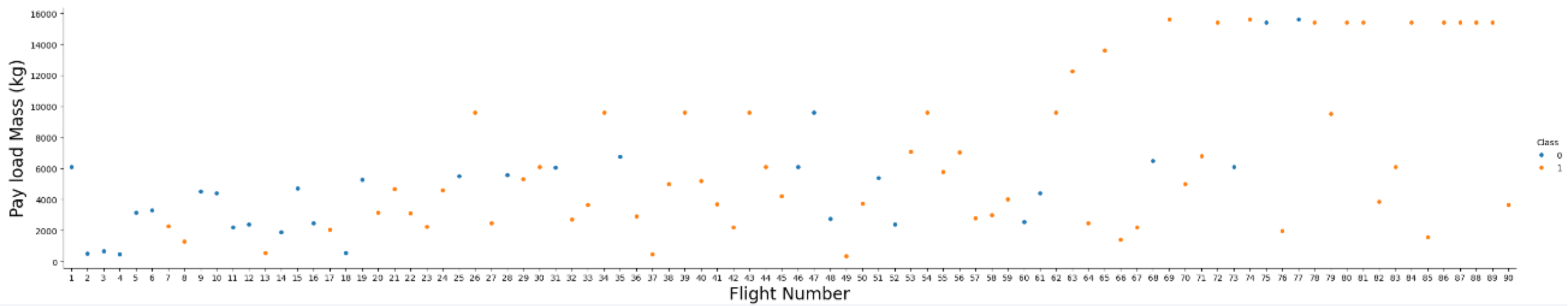
Conclusions

- Machine Learning can predict the outcome of a landing attempt with over 80% accuracy
- More influences (or features) like weather condition, wind etc. should be considered
- Underrepresentation of certain orbits and launch sites can be problematic for the result

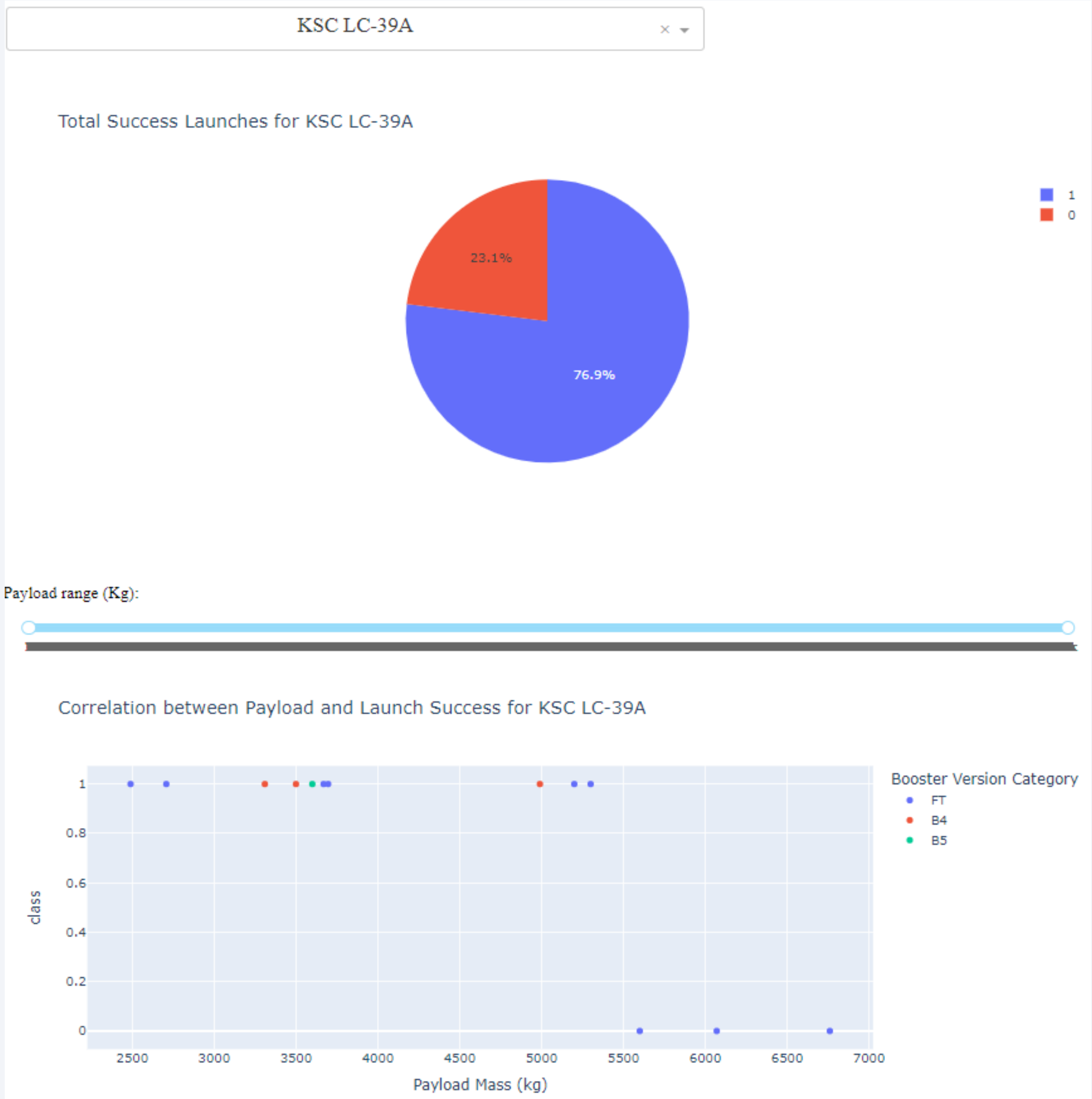
Appendix



Appendix

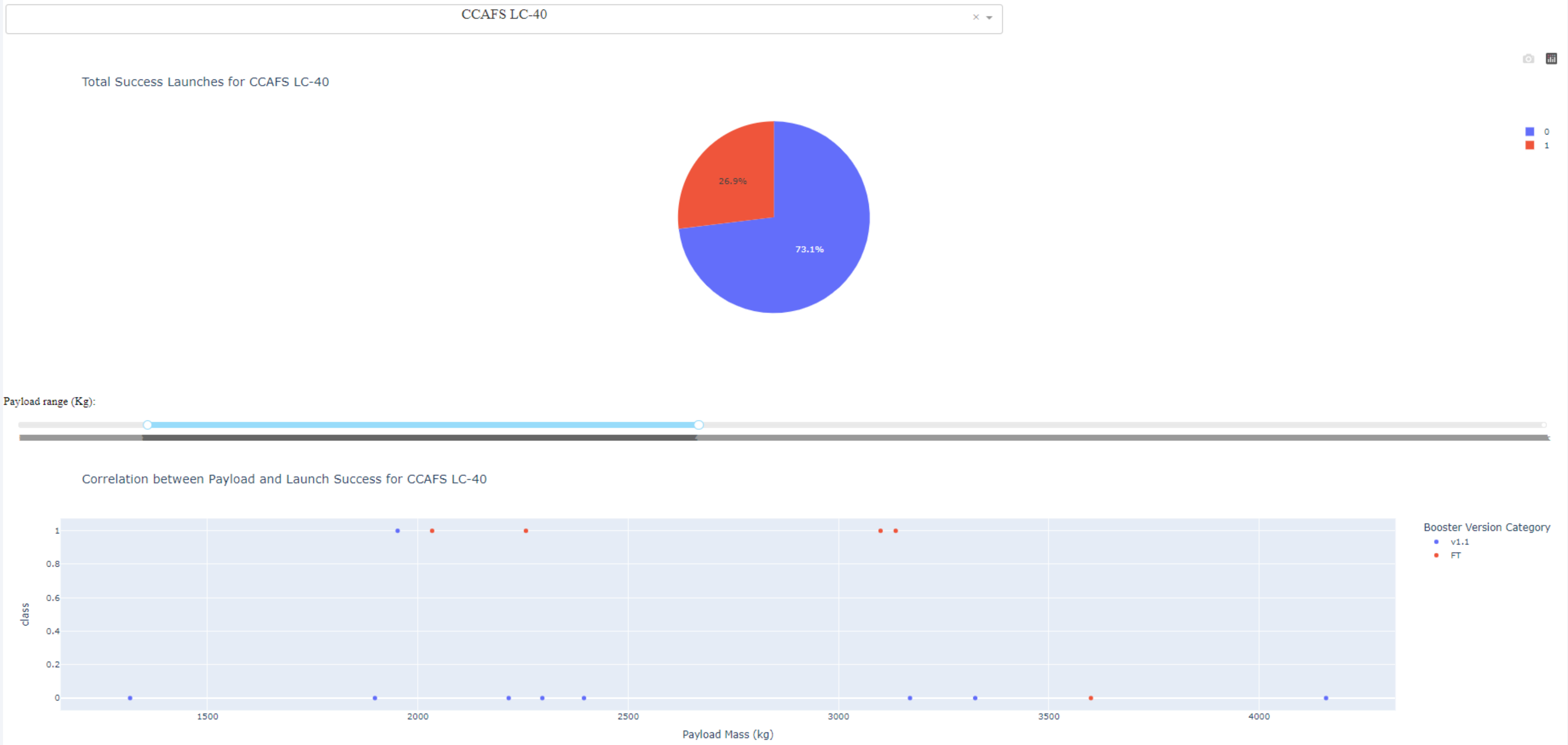


Appendix



KSC LC-39A
Total Success Launches for KSC LC-39A
23.1%
76.9%
1
0
Payload range (Kg):
Correlation between Payload and Launch Success for KSC LC-39A
Booster Version Category
FT
B4
B5
class
1
0.8
0.6
0.4
0.2
0
2500
3000
3500
4000
4500
5000
5500
6000
6500
7000
Payload Mass (kg)

Appendix


CCAFS LC-40
Total Success Launches for CCAFS LC-40
26.9%
73.1%
0
1
Payload range (Kg):
Correlation between Payload and Launch Success for CCAFS LC-40
class
Payload Mass (kg)
Booster Version Category
v1.1
FT

Thank you!