IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Dawid Osiecki
22 October 2023

Outline

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

Executive Summary

In this project, we embarked on a comprehensive analysis to determine the pricing of rocket launch based on the reusability of the first stage of launch. The analysis is based on the history of the SpaceX Falcon 9 launches. The methodologies employed can be summarized as follows:

- Data Gathering – We gathered extensive public information about SpaceX, including historical rocket launch data, mission parameters, payload details, and outcomes of previous launches.
- Machine Learning Model – to predict the reusability of Falcon 9's first stage, we developed a machine learning model. This model was trained on a diverse dataset that included mission-specific variables and historical success rates.

The analysis yielded noteworthy results. The machine learning model demonstrated high accuracy in predicting the recoverability of the first stage, providing Space Y with a valuable tool for competitive strategy.

Introduction

In the ever-evolving landscape of space exploration and commercialization, our project stands at the intersection of innovation and competition. The commercial space age has dawned, and companies like SpaceX have made space travel more accessible than ever before. As we embark on this endeavor, it's essential to understand the dynamic context that has shaped our mission.

SpaceX, under the visionary leadership of Elon Musk, has achieved remarkable feats, from sending spacecraft to the International Space Station to establishing the revolutionary Starlink satellite internet constellation. Their cost-effective approach to space launches, epitomized by the Falcon 9 rocket, has disrupted the industry. SpaceX's ability to recover and reuse the first stage of their rockets has been a game-changer, significantly reducing launch costs.

However, in the midst of this competition, Space Y, a new entrant into the market, seeks to challenge SpaceX's dominance. To do so, we must delve into the heart of the matter. Our project seeks to answer crucial questions: What drives the pricing of space launches, and how can we offer competitive pricing? Can we predict the reusability of the first stage, allowing for strategic decisions? These questions are central to our mission, and as data scientists, our task is to leverage data, methodologies, and insights to provide solutions that propel Space Y into the forefront of the commercial space launch industry.

Section 1

Methodology

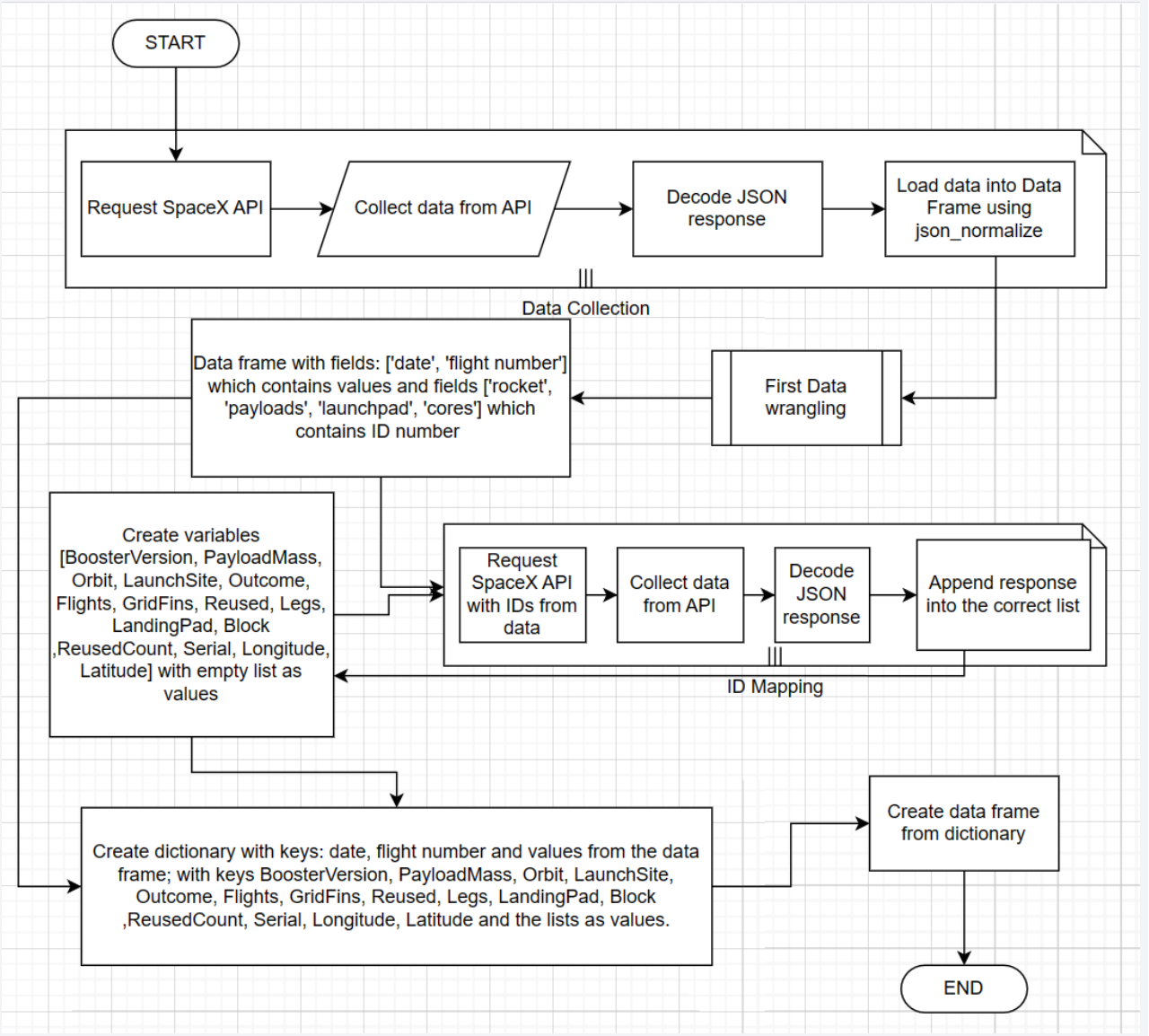
Methodology

Executive Summary

- Data collection methodology:
  - Collection data from public accessible SpaceX API as well as from the wikipedia.com website.
- Perform data wrangling
  - Cleaning and filtering data to ensure its quality and suitability for analysis.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Train various models, find their best parameters, calculating metrics and selecting the best algorithm.

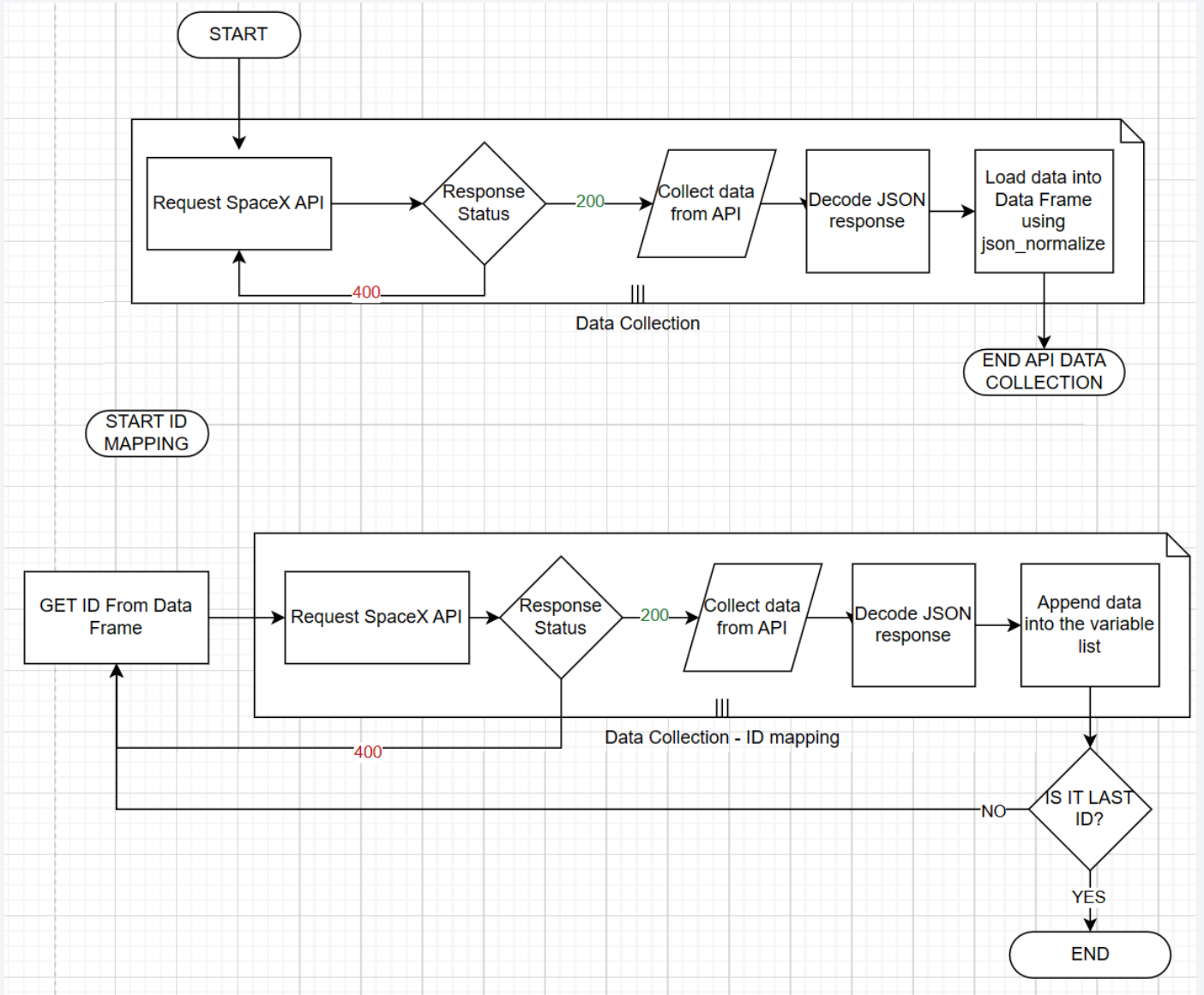
Data Collection

- Data sources: SpaceX API, Wikipedia page with Falcon 9 information.
- Data collected via methods: request SpaceX API and web scrapping techniques.
- Carefully selected relevant data fields.
- Data validation - check what values data is presenting - many fields contains lists with ID instead of values.
- Perform first data wrangling to extract ID from lists.
- Request SpaceX API again to get information based on ID.
- Create data frame.



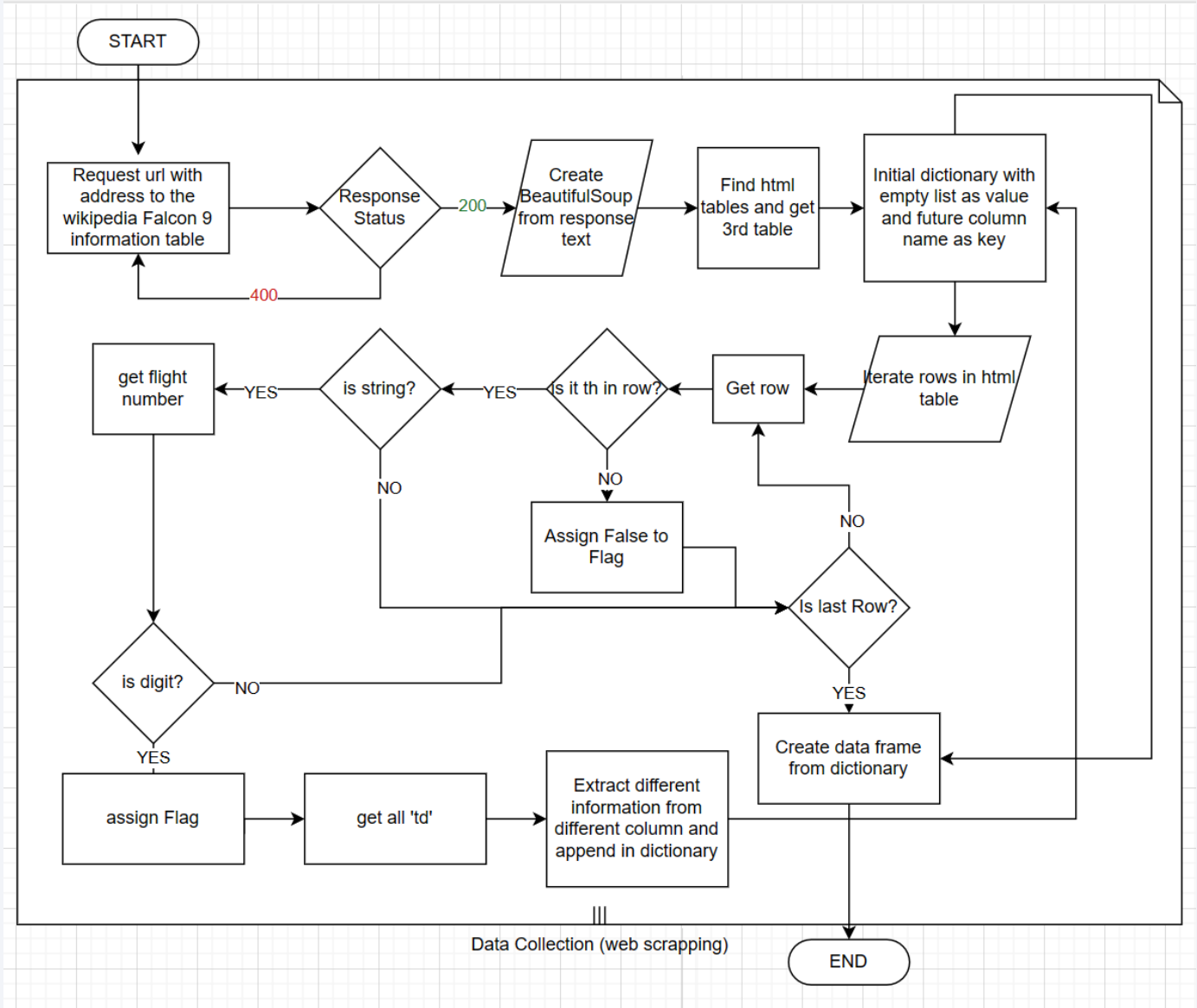
Data Collection – SpaceX API

- Data Collection from the SpaceX API has been split into 2 parts.
- One was to get all data with historical launches and load into the data frame.
- Second was to request specific information from API based on its ID. Those were rockets, launchpads, payloads, cores and store in the list, from which later on the dictionary were created and data frame.
- Jupyter Notebook with full code (link)



Data Collection - Scraping

- Request html text from the url of Wikipedia page with Falcon 9 launches historical data. Create Beautiful soup from request response. Find correct table and iterate table rows to extract information about launches.
- At the end make data frame from all those information and export to csv file.
- Jupyter Notebook with full code (link)



Data Wrangling

The data fetched from the SpaceX API contains a lot of lists with ID instead of the fields names. Those included rocket, payloads, launchpad and cores. Therefore the first data wrangling was done in half of the data collection. Those were:

- Filtering the rocket launch dates to earlier or equal than 13 November 2020.
- Extracting the ID value from the list.

Once whole collection were done we take only the Falcon 9 launches by filter BoosterVersion != 'Falcon 1' and we checked null values in data frame. There were 2 fields with null values – PayloadMass and Landing Pad.

For Payload Mass we fill nulls with the value of average of payload mass and the Landing pad left with null values.

The last and main part of the data wrangling was performing exploratory data analysis and determine training labels.

We checked different launch outcome and define which doesn't mean succesffull land of the second stage. Then we assign to those class = 0 and to the succesfull lands class = 1. Average of the class is 0.666 which means that the success rate of all launches is 66%.

Jupyter notebook with code used to data wrangling

Jupyter notebook with code used for the first data wrangling (done during data collection)

EDA with Data Visualization

Charts plotted during EDA:

- Scatter plot with Flight number vs payload mass and launch outcome.
- Scatter plot with Flight number vs Launch site and launch outcome.
- Scatter pot with Payload vs Launch site and launch outcome.
- Bar chart with success rate of each orbit.
- Scatter plot with Flight number vs Orbit and launch outcome.
- Scatter plot with Palyoad vs Orbit type and launch outcome.
- Line plot with year and success rate.

Scatter plot charts were chosen to find correlations between different fields.

Bar plot shows which orbit has the best success rate, and line plot shows that the success trend over the years is growing.

The full EDA analysis jupyter notebook.

EDA with SQL

- Find unique Launch sites.
- Find 5 launch sites which starts with KSC
- Find total payload mass for the Nasa customer
- Find average payload mass carried by booster version F9 v1.1
- Find all dates were the drone ship was landed successfully.
- Find the booster names which successfully landed in ground and which carried payload mass between 4000 and 6000 Kg
- Find total number of successful and failure mission outcomes.
- Find names of the booster versions which carried the 15600Kg payload mass which is the max of the payload carried on
- Find months of the year 2017, booster version and launch site for which the landing outcome was success in ground pad.
- Create ranking of number of launches with different landing outcome.

Jupyter notebook with SQL EDA analysis.

Build an Interactive Map with Folium

We build the interactive map which show the Launch sites as blue circles. Inside each site there are green or yellow circle with number of launches. By clicking on it you can see green or red icons which represent successful or failure launch.

You can find also the blue lines with text which represent number of kilometers between the launch site and closest city, railway or highway.

Those objects were added to easily see if launch sites are close to railways, highways, coastline and cities.

Map also nicely visualize that the launch sites are on east and west of the US and shows very clearly how many launches were from each site.

Jupyter notebook with the full code of plotting maps.

In case of not seeing the interactive maps on github, please use the nbviewer tool:

Jupyter Notebook Viewer (nbviewer.org)

Build a Dashboard with Plotly Dash

Dashboard build with plotly contains 2 plots and 2 options to choose by user.

User can choose the launch site and range of the payload.

Based on the selected options the pie chart will show successful and failure lands and the scatter plot will show correlation between payload and launch success.

Code of the dashboard in the github.

Predictive Analysis (Classification)

Once the EDA were done and we could find the best fields for the features of our model, we replaced categorical values by numeric ones and load those as the features of our model. The prediction values of model are the column class which contains the 1 or 0 as success or failure land.

Next the class were loaded to numpy array Y and the features were normalized with array X as result.

We split the data into the train and test model with ratio of 80/20.

Then we create different models – logistic regression, support vector machine, decision tree, and k-nearest neighbors; Then we find the best parameters with grid search and fit the model with the train data.

For each model we plot the confusion matrix, and calculate accuracy and other metrics. The confusion matrix plot and metrics table gives the possibility to compare all models and found the best performing one.

Jupyter notebook with full code on github.

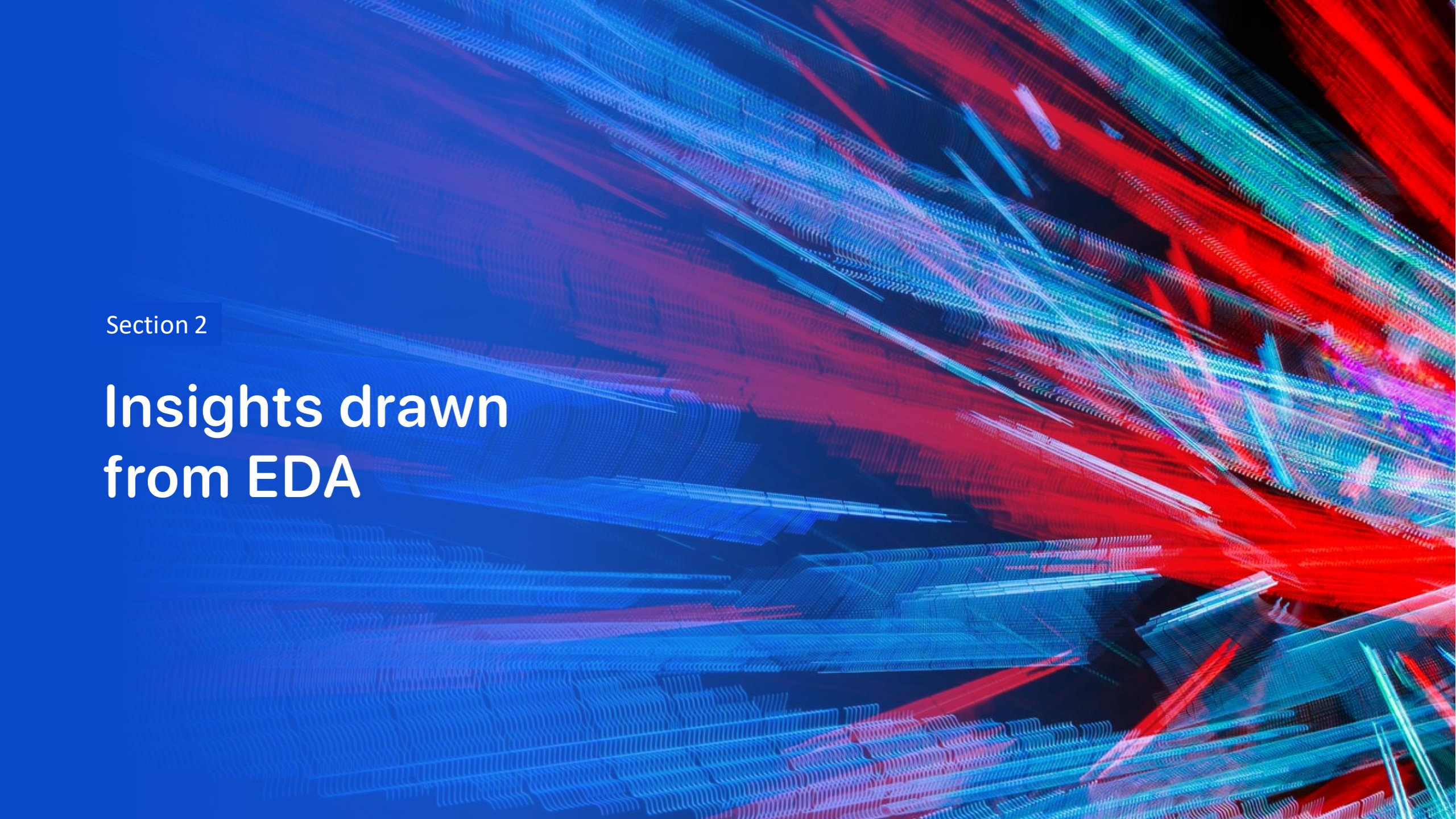
Jupyter notebook in nbviewer.org in case not all is rendered on github.

Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

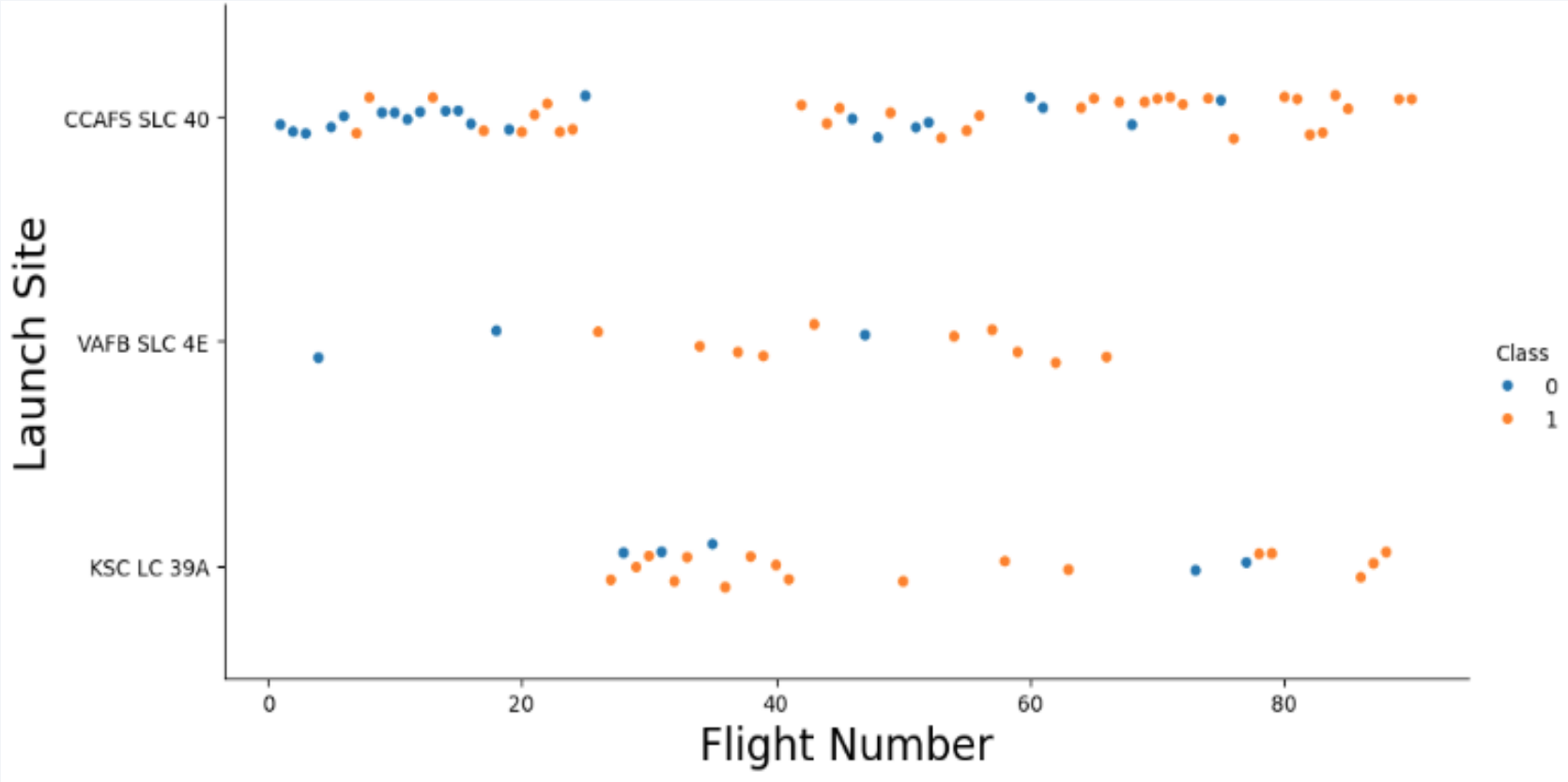


Flight Number vs. Launch Site

The CCAFS SLC 40 site has the biggest amount of flights and have very low success rate(35%) in the first 20 flights, but after that improve and achieved success rate 85% in the last 20 flights.

All sites has better success rate at the end than at the beginning.

The VAFB SLC 4E site launching not many flights but regular ones, when the KSC LC 39A launched a lot of flights in time when the CCAFS has break.

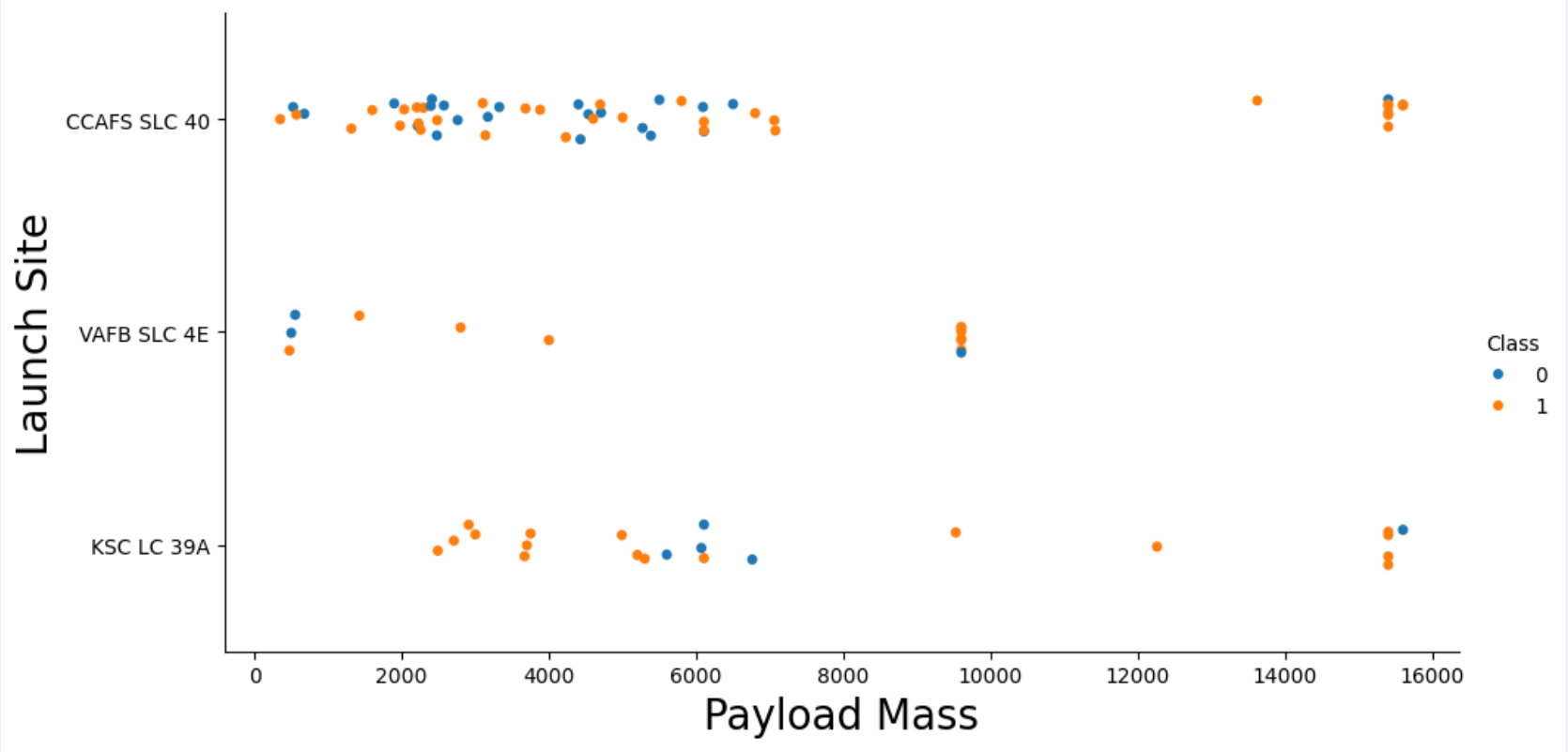


Payload vs. Launch Site

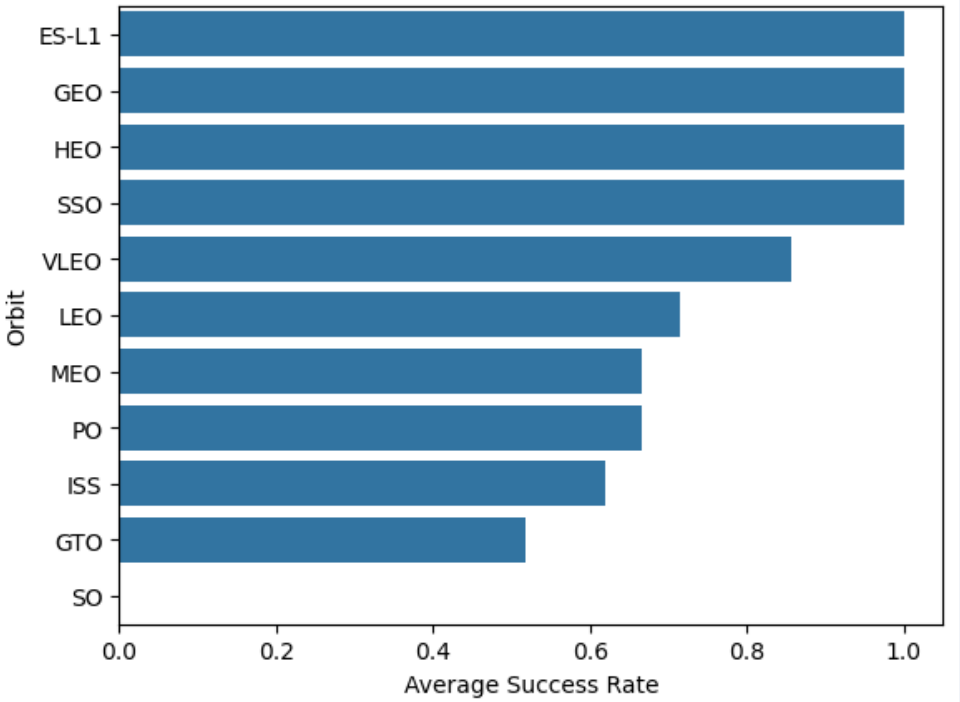
The VAFB SLC site doesn't launched rockets with greater than 10000 kg and the KSC LC site doesn't launched rockets lighter than 2000 kg.
The CCAFS site has very random success rate with the loads lighter than 8000 kg and very high for the loads heavier than 14000 kg. As well as it don't launch rockets between 8000 and 14 000 kg.
KSC LC site fail only with rockets weighted between 5500 and 7000kg and once with rocket weight approx. 15500 kg.

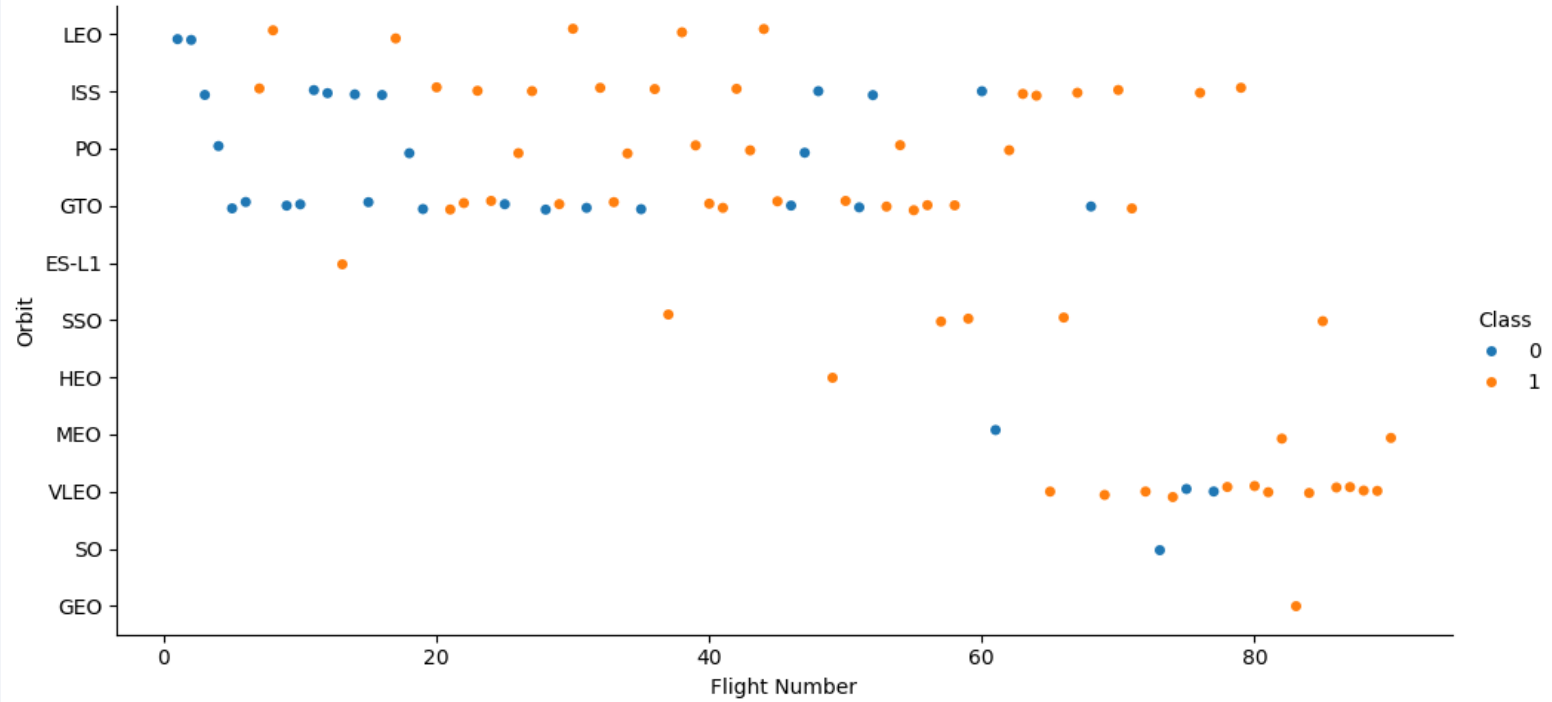
There were some launches with exactly the same payload from CCAFS SLC and KSC LC site with weight of +/-15300 kg.

Also the VAFB site has launched few times the same load with weight of +/-9100 kg



Success Rate vs. Orbit Type

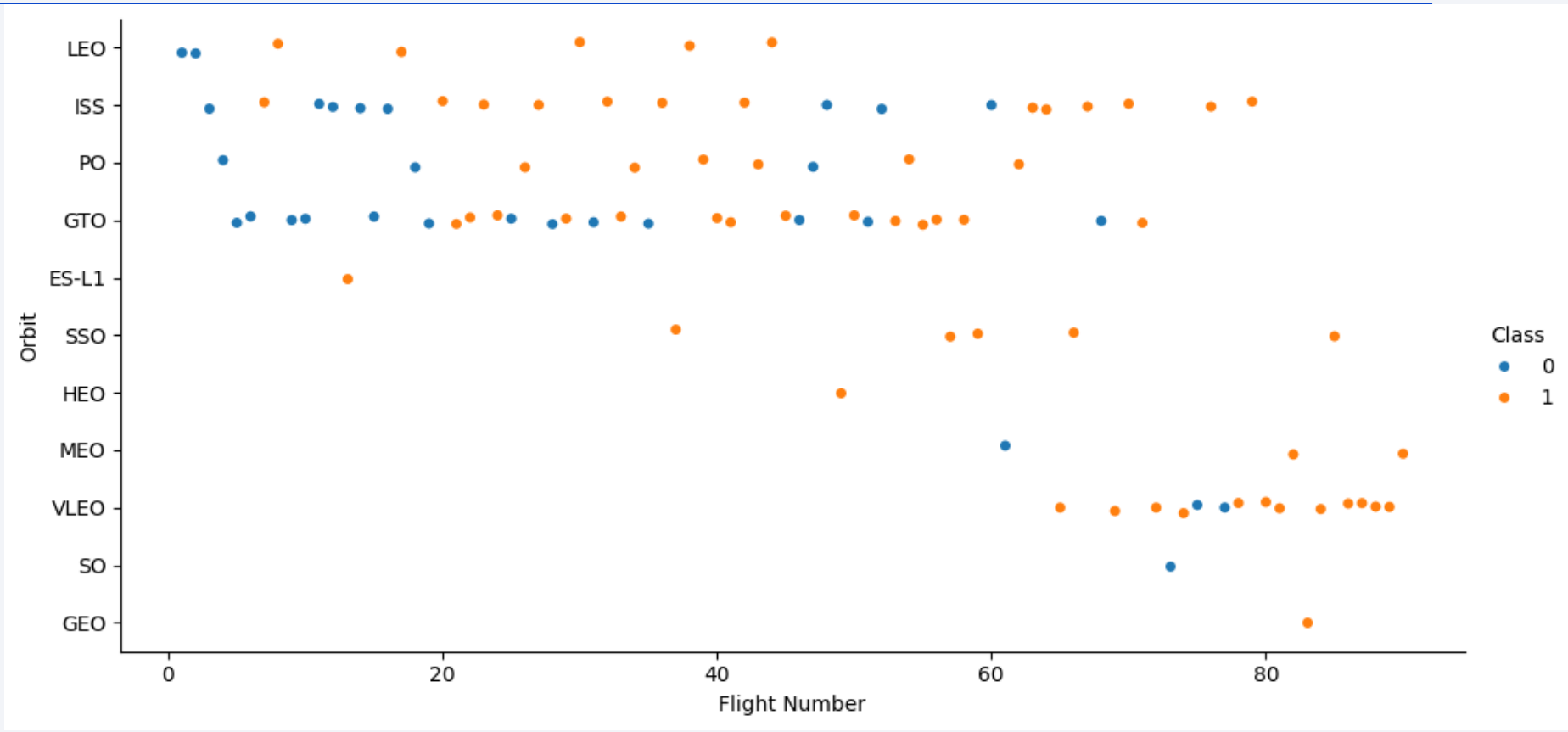




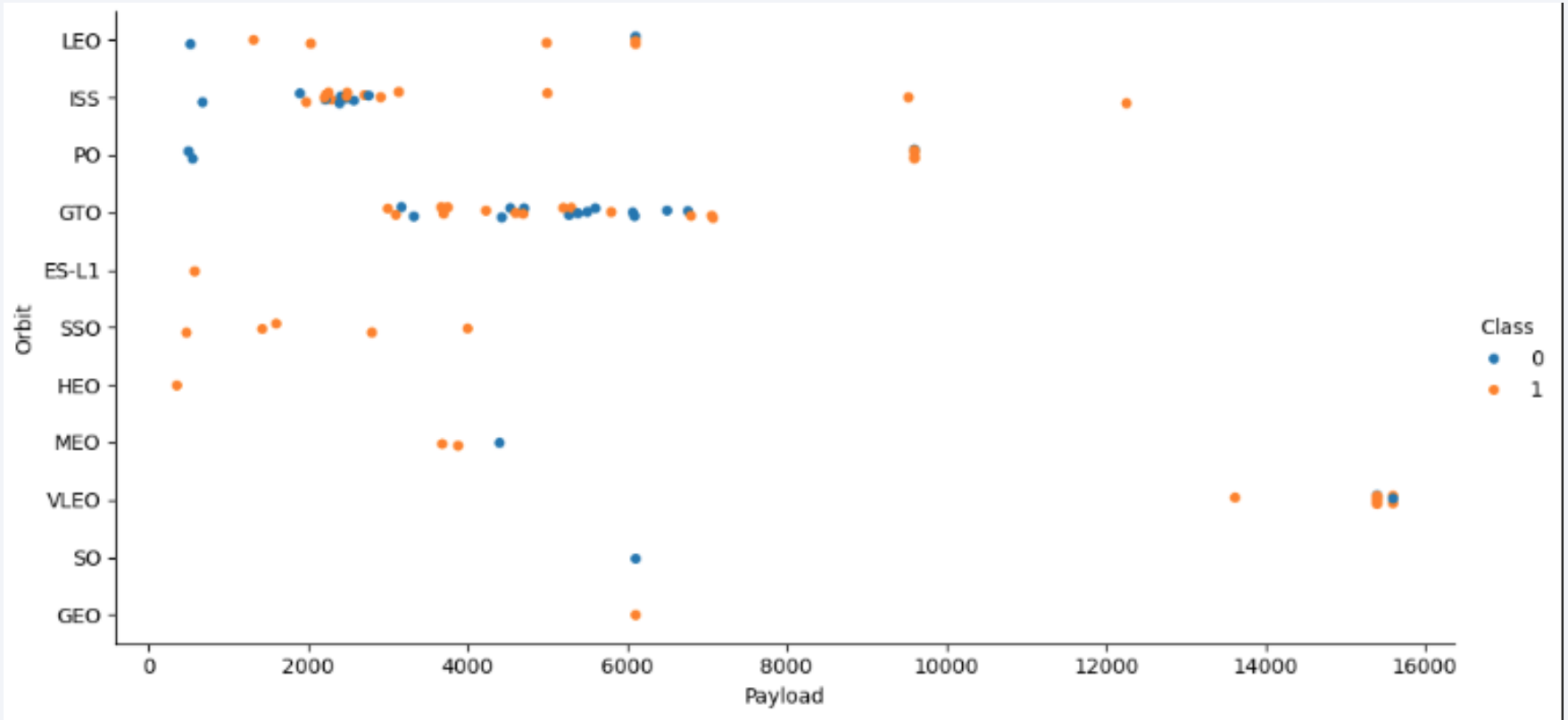
While bar plot shows that there are some orbits with very high success rate, the scatter plot gives a more insides of it and shows that the first 3 top orbit has only single launch, therefore high success rate of those cannot be compared with the success rate of GTO orbit to were much more launches.

Flight Number vs. Orbit Type

There is visible that in the LEO orbit success appears related to the number of flights when there seems to be no relationship between flight number in GTO orbit.



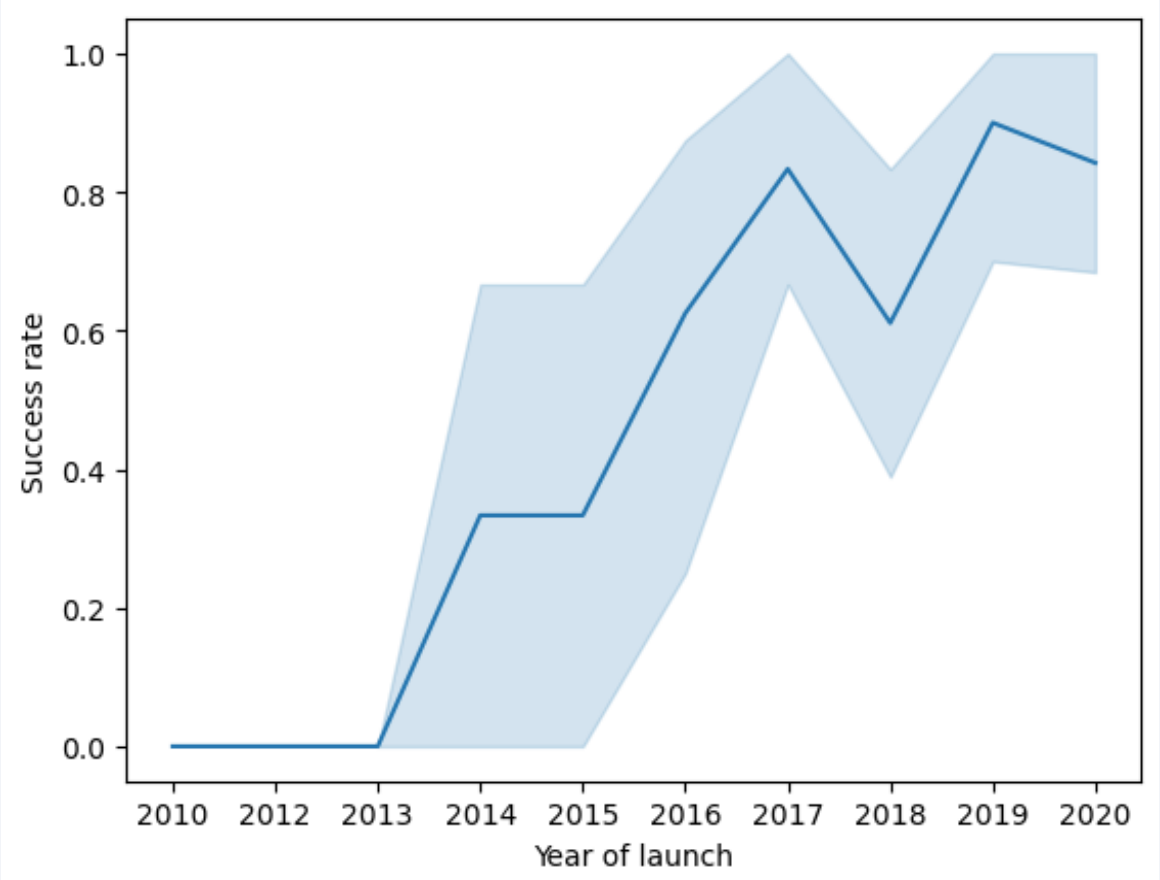
Payload vs. Orbit Type



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



you can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

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

- Result of query which gives the distinct names of the column launch site in the spacextbl table.

Launch Site Names Begin with 'KSC'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-03-16	06:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-01-05	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2017-05-15	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

Result of querying the 5 records of which Launch site is start with "KSC"

Total Payload Mass

Customer	SUM("PAYLOAD_MASS__KG_")
NASA (CRS)	45596

Sum of the total payload mass for which the customer was NASA (CRS)

Average Payload Mass by F9 v1.1

Booster_Version	AVG("PAYLOAD_MASS__KG_")
F9 v1.1 B1003	2534.6666666666665

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

First Successful Ground Landing Date

Date	Mission_Outcome	Landing_Outcome
2016-05-27	Success	Success (drone ship)

- Date of the first successful landing outcome in drone ship

Successful Drone Ship Landing with Payload between 4000 and 6000

Booster_Version	PAYLOAD_MASS__KG_	Landing_Outcome
F9 FT B1032.1	5300	Success (ground pad)
F9 B4 B1040.1	4990	Success (ground pad)
F9 B4 B1043.1	5000	Success (ground pad)

The Booster versions which have success in ground pad and have payload mass greater than 4000 but less than 6000 kg.

Total Number of Successful and Failure Mission Outcomes

Mission_Outcome	COUNT(*)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- Total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

- All Boster versions which have carried the maximum payload mass.

2015 Launch Records

Month	Year	Landing_Outcome	Booster_Version	Launch_Site
02	2017	Success (ground pad)	F9 FT B1031.1	KSC LC-39A
01	2017	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
03	2017	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
08	2017	Success (ground pad)	F9 B4 B1039.1	KSC LC-39A
07	2017	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A
12	2017	Success (ground pad)	F9 FT B1035.2	CCAFS SLC-40

- Month, year, landing outcome, booster version and launch site for the months in year 2017

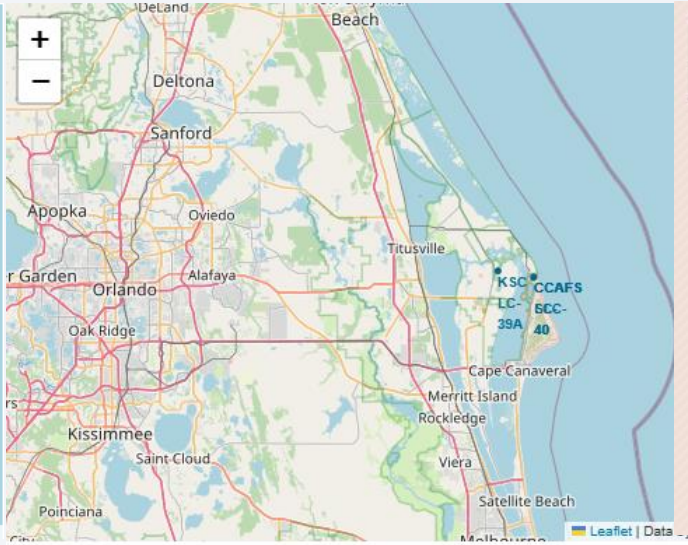
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

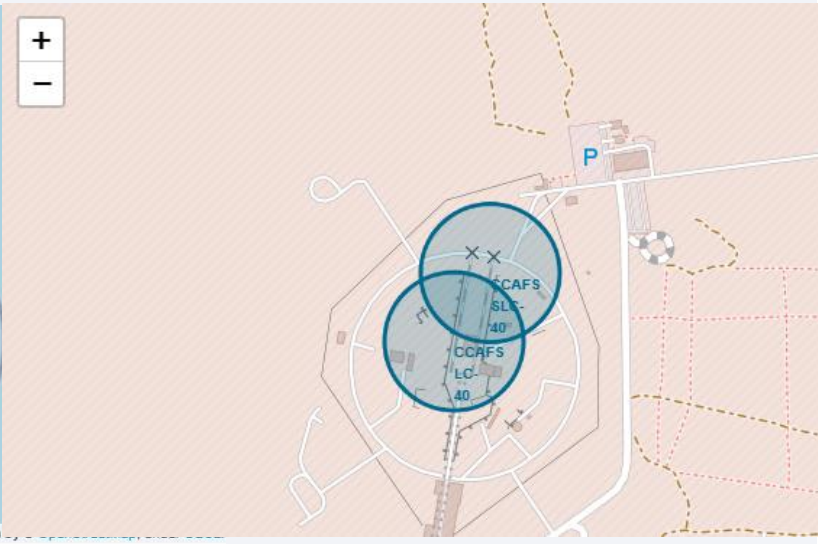
Landing_Outcome	launches	Ranking
Success	38	1
No attempt	21	2
Success (drone ship)	14	3
Success (ground pad)	9	4
Failure (drone ship)	5	5
Controlled (ocean)	5	5
Failure	3	7
Uncontrolled (ocean)	2	8
Failure (parachute)	2	8
Precluded (drone ship)	1	10
No attempt	1	10

- Rank the count of landing outcomes.
- First place has the Success outcome with 38 launches.


Section 3
Launch Sites
Proximities Analysis

Launch sites locations

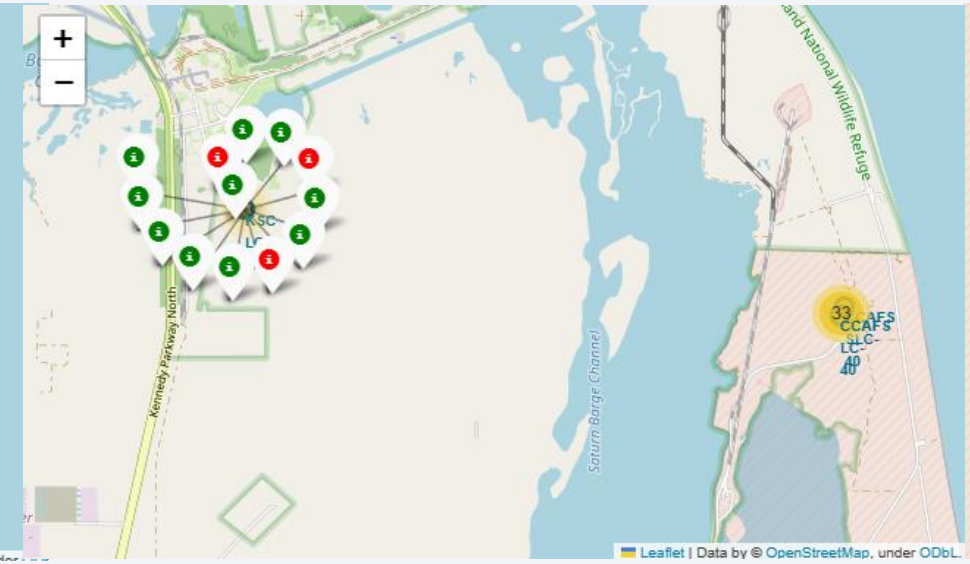


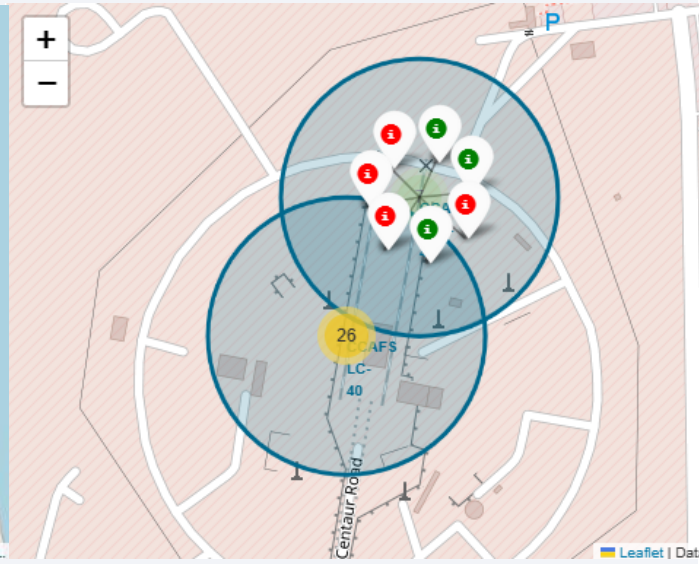


All four Launch sites are situated in the US. All launch sites are in proximity to the Equator line and all are very close to the coastline. Three of them are on the opposite coastline to the fourth one.

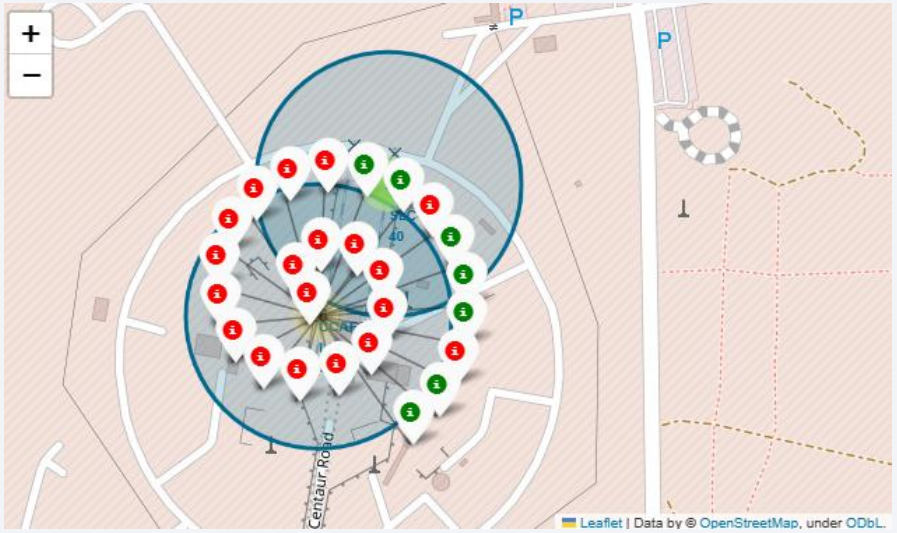
Three launch sites on the east coastline are so close to each other that they are not visible correctly without zooming in.

Successes and failures of each location

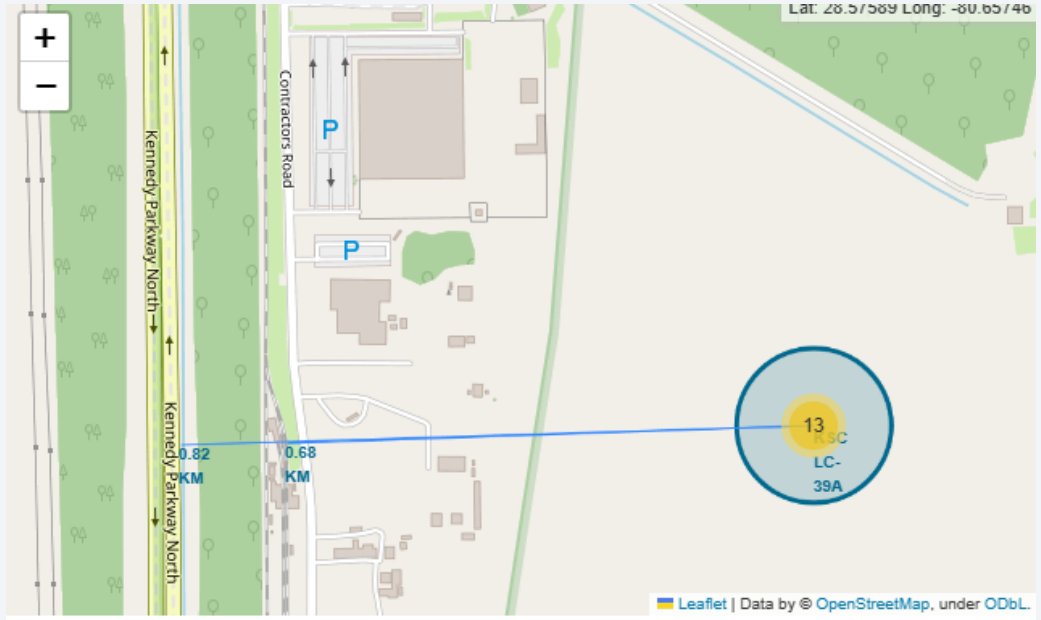


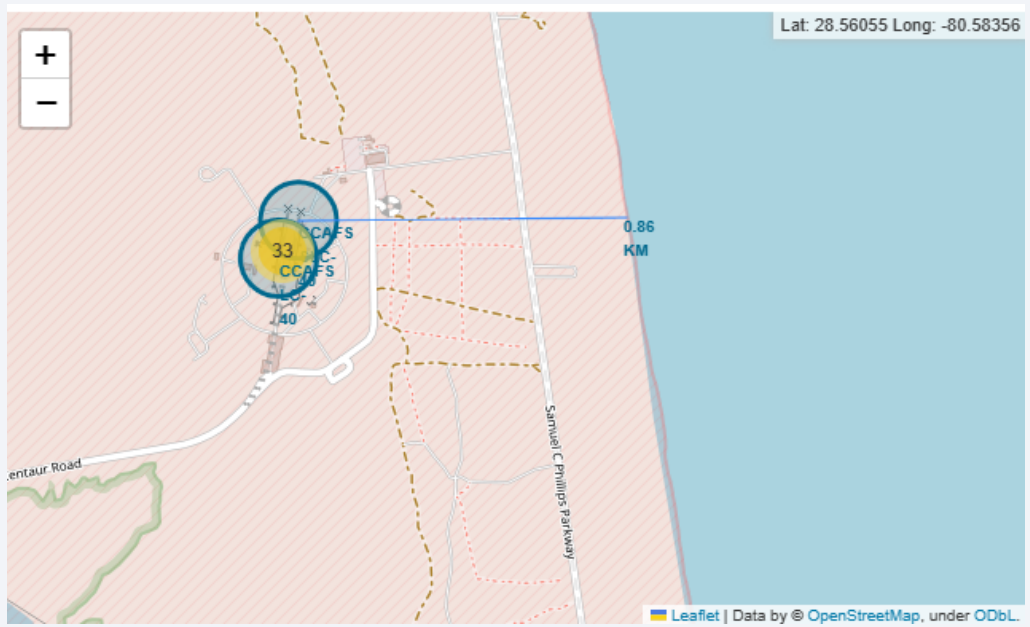


The interactive map displays a number of the launches. After zooming in you can see the color indication for which green means the launch had a successful landing and red failure.



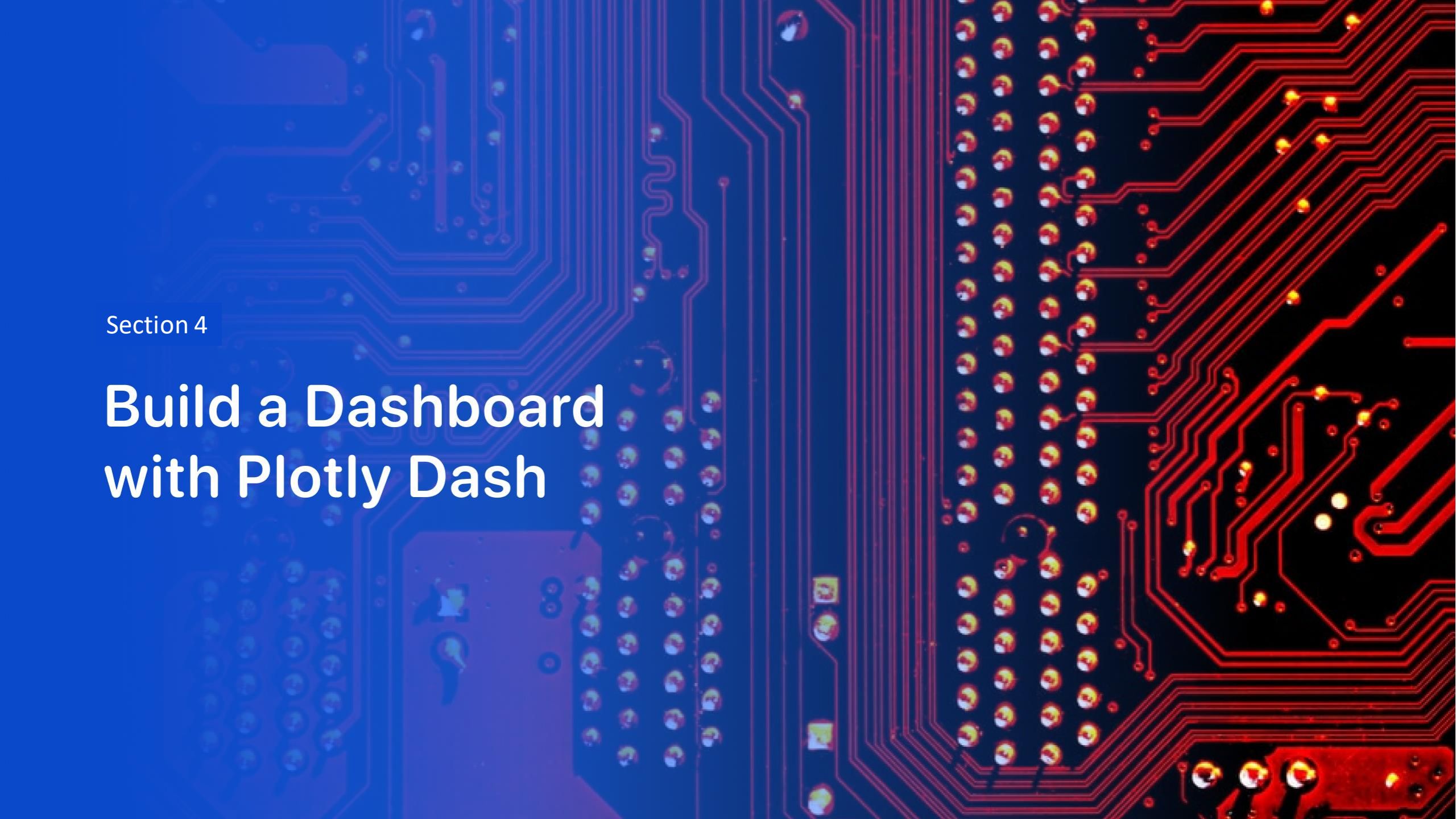
Distance from Launch sites to the strategic points.





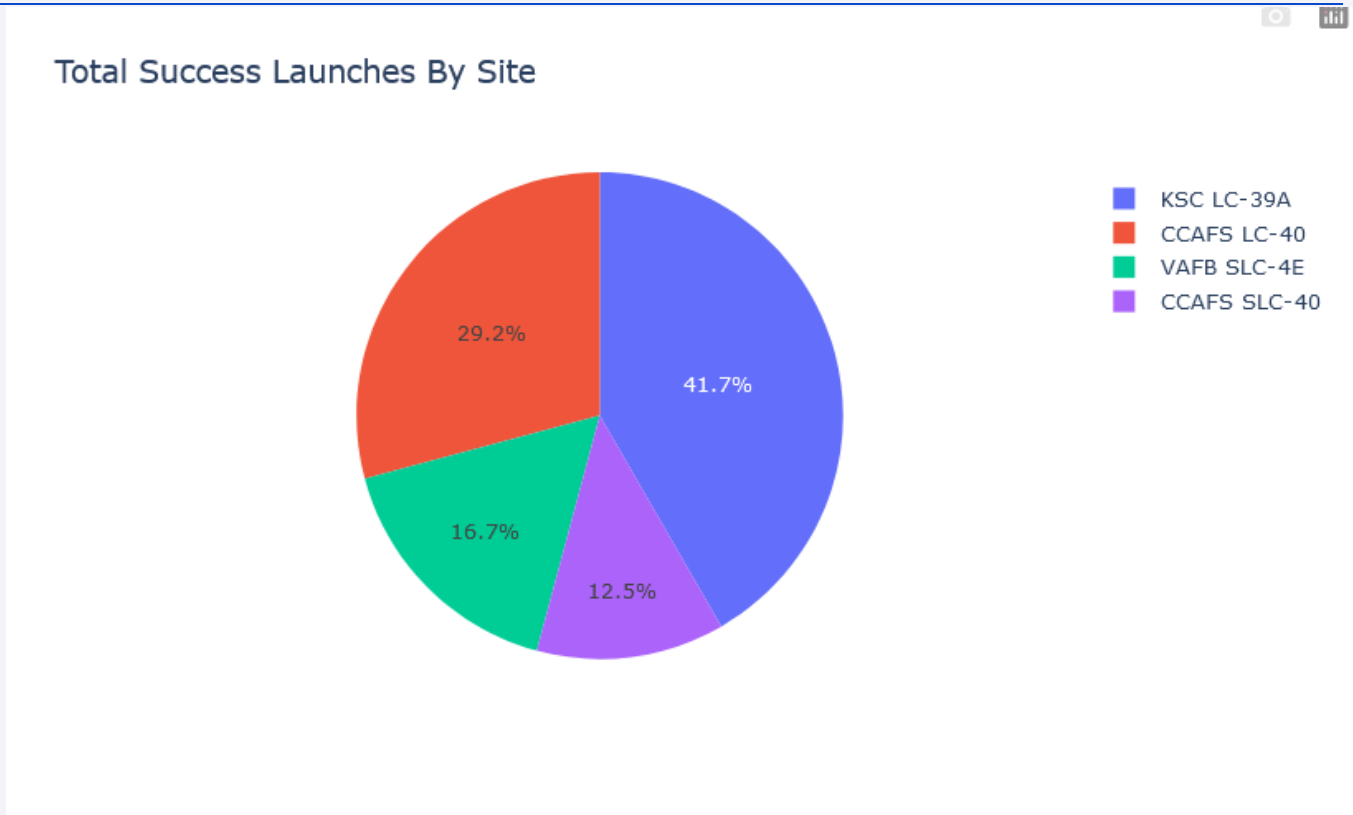
Blue lines on the map have marked the distance between the launch site and the closest coastline, railway, highway, and city.

Most of those places are not farther than 1 km from the Launch site, except for the city which is approximately 14 km away.


Section 4
Build a Dashboard with Plotly Dash

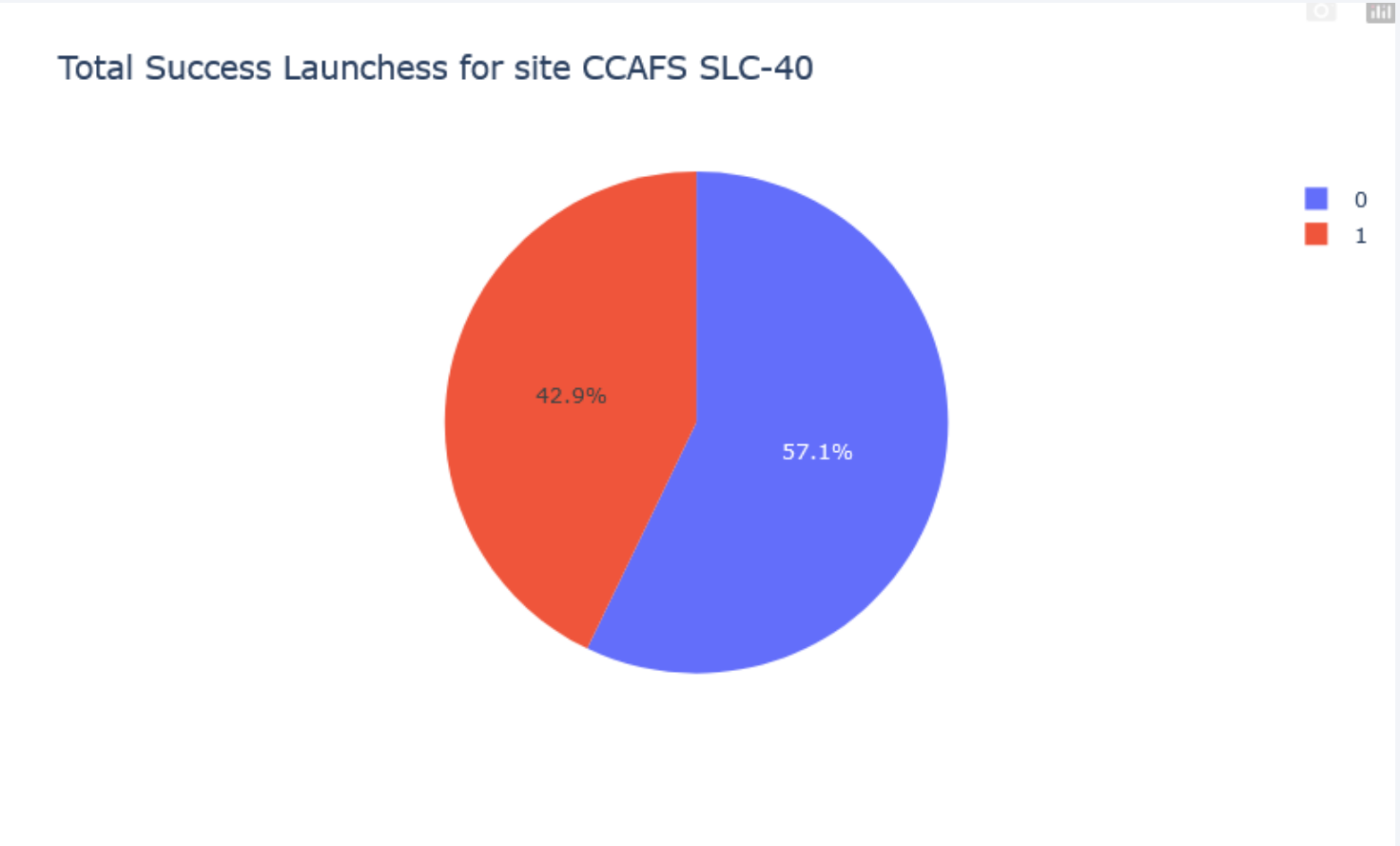
Success Launch count for all sites

- Pie chart shows total success launches by site. The KSC LC site have the biggest amount of success launches, which is 41.7% of all success launches.



Launch site with highest launch success ratio

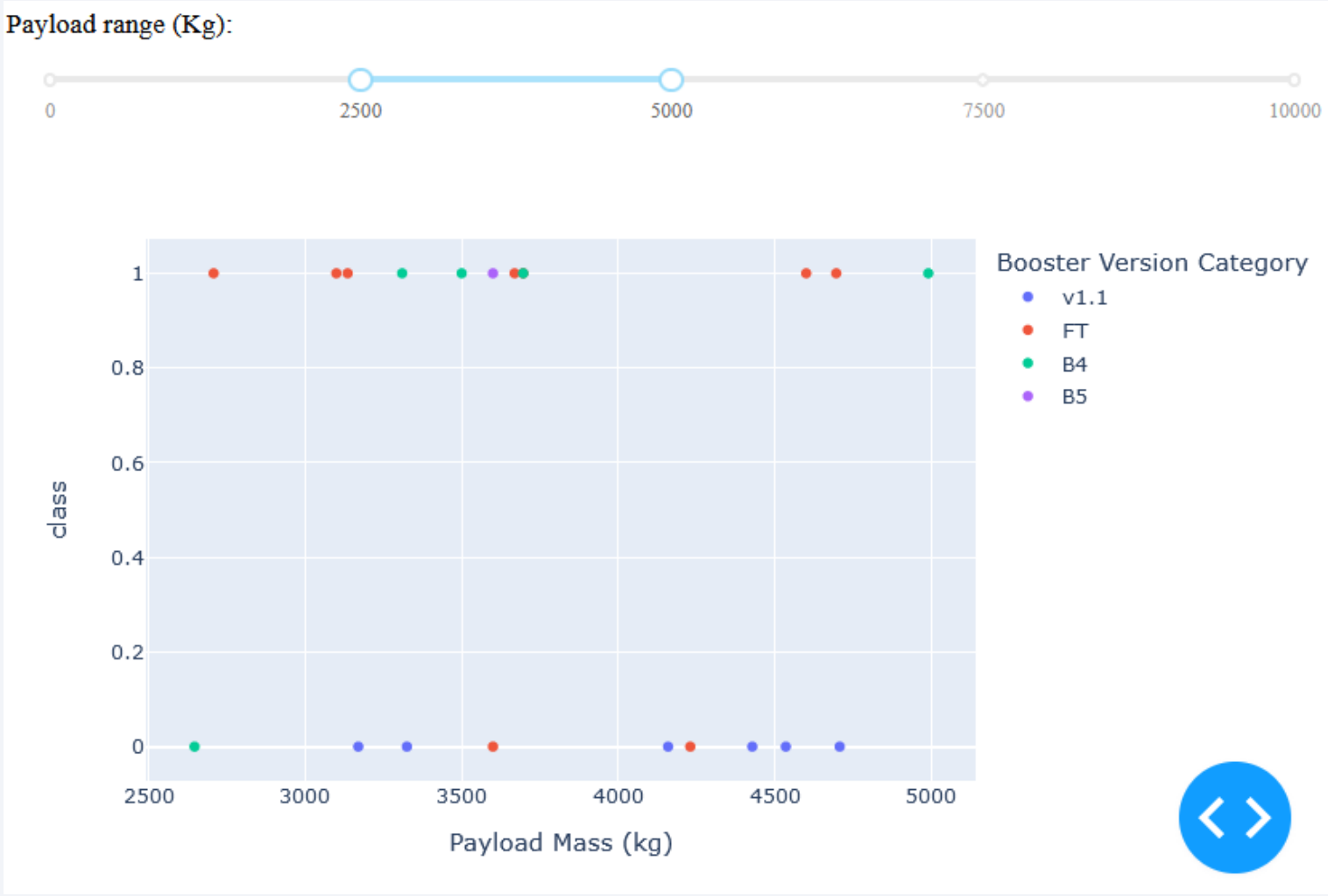
- Launch site CCAFS SLC-40 have 42.9 % of success landing launches which is the highest ratio of success versus failure.



Launch result vs Payload for launches with payload between 2500 and 5000 kg

Dashboard allow to select different payload mass range to see the successes and failures of the launches with marked by color the Booster Version Category.

Selected range from 2500 to 5000 kg shows that the FT booster have 7 successes and 2 failures when Booster v1.1 in the same payload range have only 6 failures.

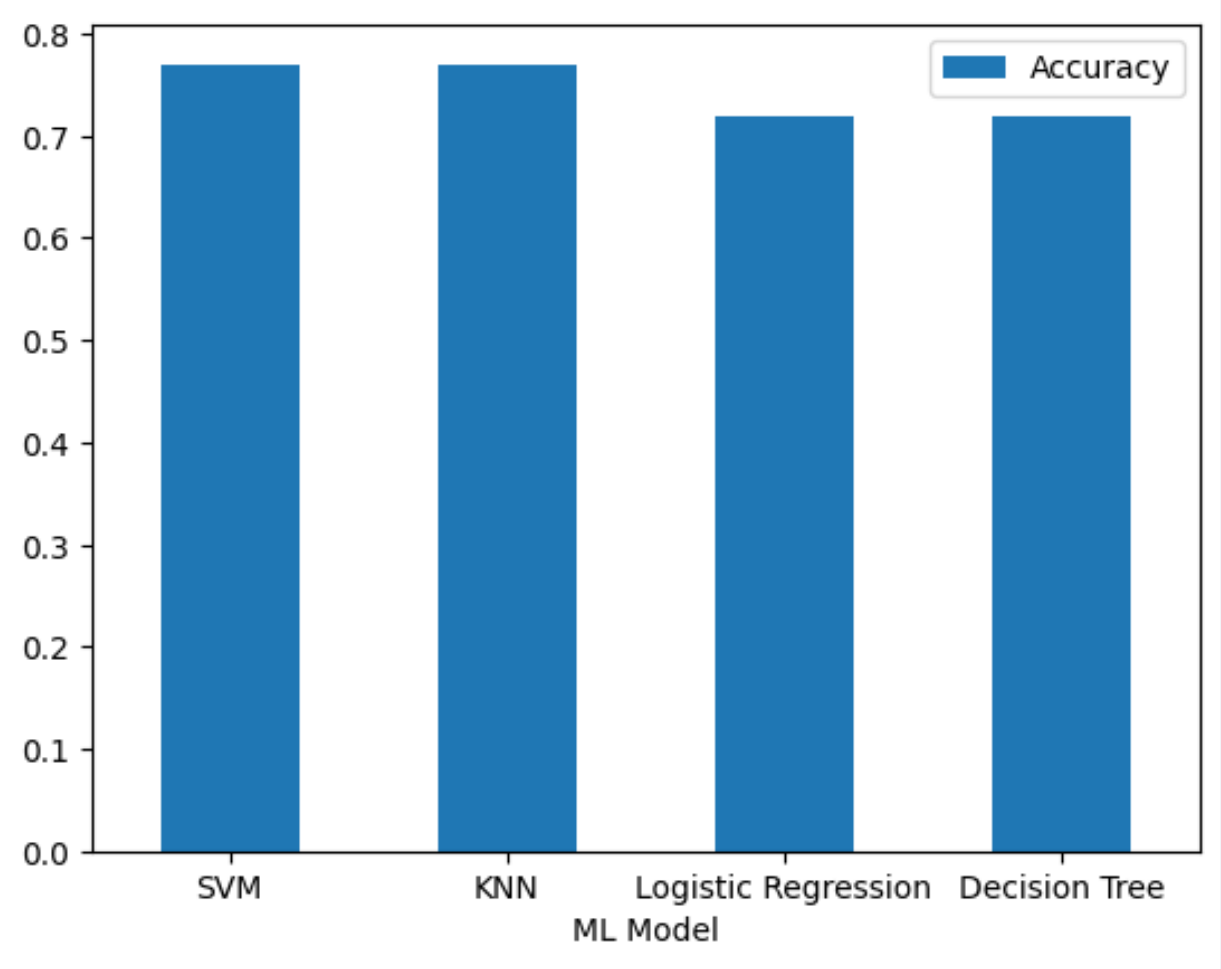


Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The classification accuracy doesn't give clearly one the best model – both the SVM and KNN has the highest accuracy 0.77

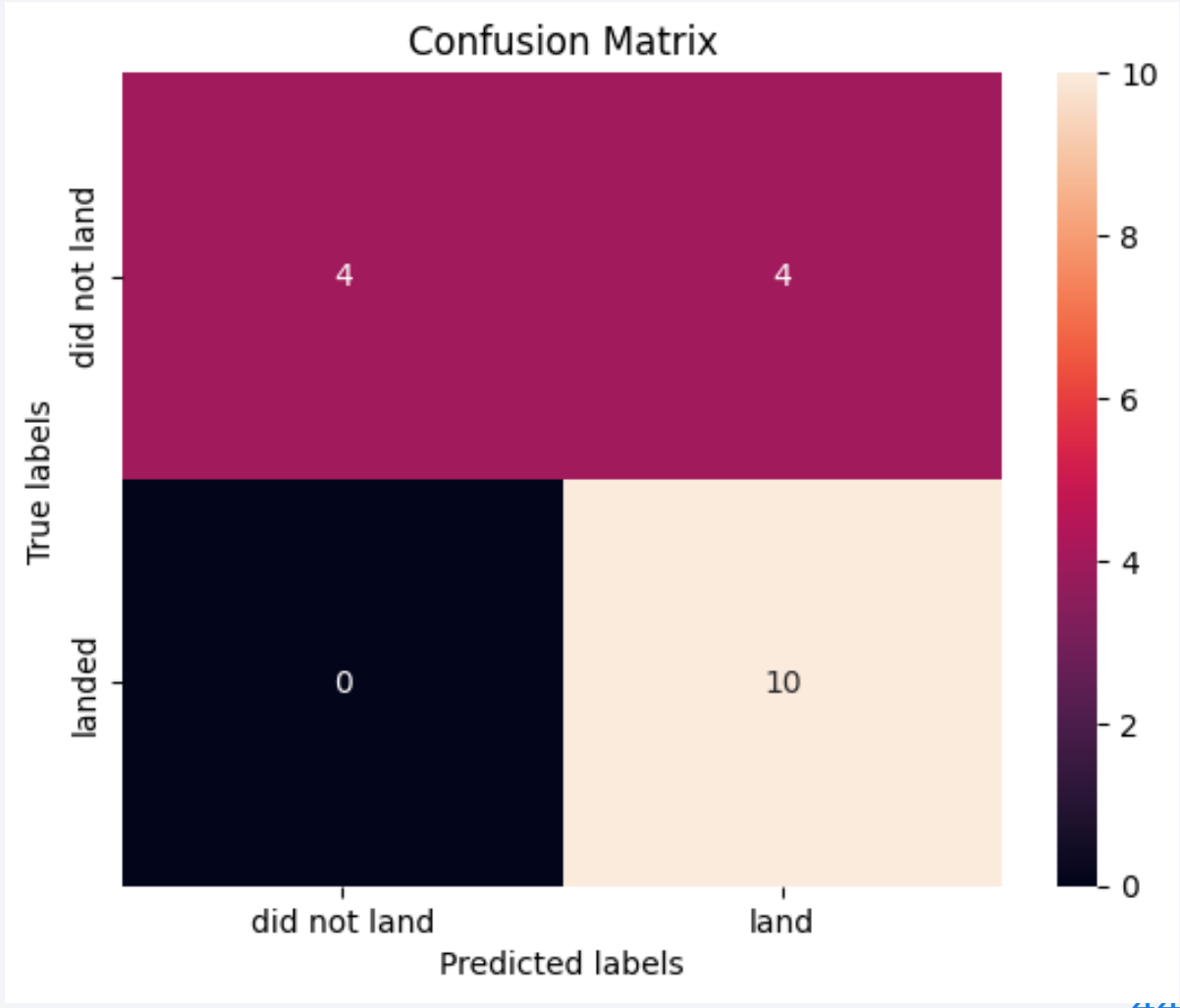


Confusion Matrix

Confusion Matrix of the best performing model shows that model predicted correctly 10 on 10 lands of the second part of the rocket.

It also predict correctly 4 on 4 launches were the second part of the rocket didn't land.

Model was wrong in for cases, were predicted that the rocket will not land, but it was landed.



Conclusions

- Training four different models gives the same two results. The SVM and KNN models gives slightly better results than Logic regression and decision tree.
- The similar results could be caused by small size of training data.
- The accuracy on the level of 0.77 is not the best and could be improved by bigger size of data.
- Still with trained model we could correctly predicted 4 launches which did not land and 10 launches which land.
- Model make mistake with 4 launches for which predict success, but it was failure in landing.
- Trained model together with right budget management can give already some savings for the Y company and can be improved when more launches will be done.

Appendix

Everything what was used in this project is included in the GitHub repositorium under the following address:

Thank you!