IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Konstantinos Zygogiannis
25-04-2023

Outline

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

Executive Summary

- Summary of methodologies

❖ Data collection

❖ Data wrangling

❖ Exploratory Data Analysis (EDA) and data visualization

❖ Predictive analysis with Machine Learning

❖ Interactive map (Folium)

❖ Interactive dashboard (Plotly)

- Summary of all results

❖ EDA showed the correlation between different features and the success of launchings

❖ Predictive analysis showed the most suitable model to predict successful landings , using collected data

Introduction

- *Project background*

SpaceX is a rocket company that launches satellites at significantly lower prices compared to their competition. This is possible , mainly because the company can reuse a part of their rocket for future launches.

- *Problem*

We would like to use the available data , in order to estimate the probability of the rocket landing back to the launching pad. We would also like to see how different variables, such as launch site, payload mass, orbit, landing pad location and booster version affect the successful landing of the rocket.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
    - Data collected from:
    - SpaceX API (https://api.spacexdata.com/v4/rockets)
    - WebScrapping (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
- Perform data wrangling
    - Extracted data from the sources above and added an extra column referring to the landing outcome

Methodology

- Perform exploratory data analysis (EDA) using visualization and SQL

- Perform interactive visual analytics using Folium and Plotly Dash

- Perform predictive analysis using classification models
  - We begin by normalizing the data collected from the sources above and split the data in training and test sets. Afterwards, we build four different machine learning classification models, which are then compared based on their accuracy of prediction.

Data Collection

- Data sets were collected from the following sources:
  - ✓SpaceX API (https://api.spacexdata.com/v4/rockets)
  - ✓Wikipedia (through web scrapping) (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)

Data Collection – SpaceX API

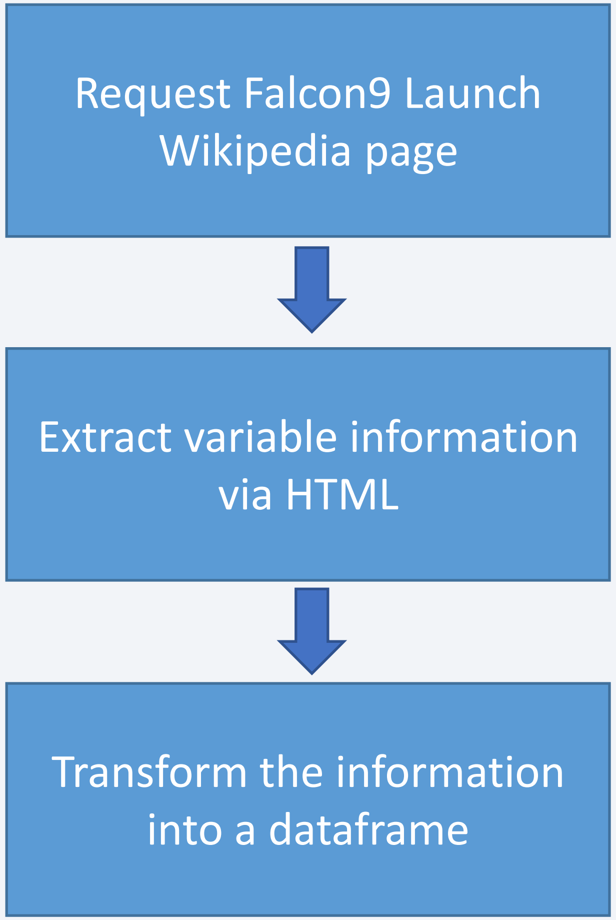
- Data was obtained from a public SpaceX API

- Data was manipulated according to the flowchart on the right side of the slide.

- GitHub link: https://github.com/Konszygo/Coursera_Data_Science_Capstone/blob/main/jupyter-labs-spacex-data-collection-api(4).ipynb

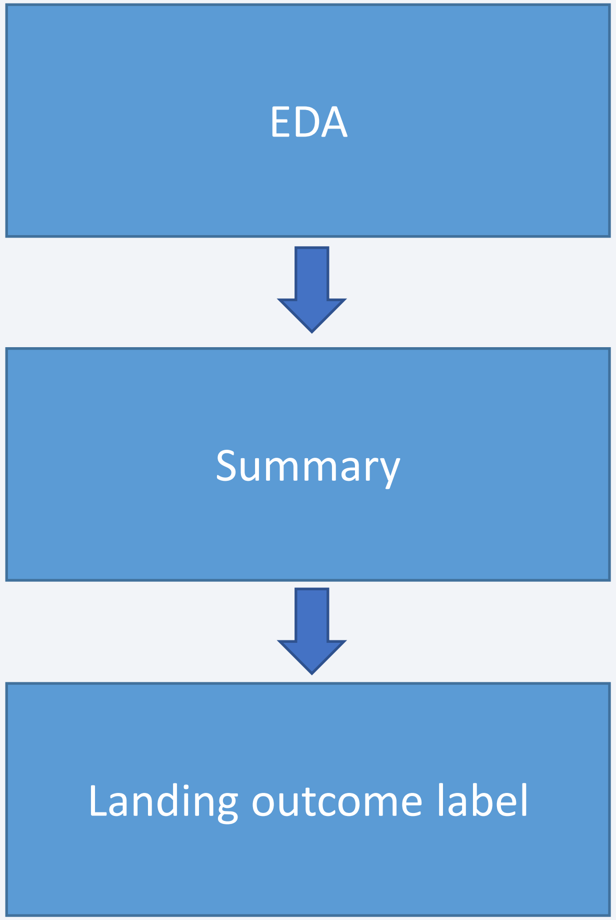
Data Collection - Scraping

- Data was obtained from Wikipedia
- Data was manipulated according to the flowchart on the right side of the slide.
- GitHub link: https://github.com/Konszygo/Coursera_Data_Science_Capstone/blob/main/jupyter-labs-webscraping(1).ipynb



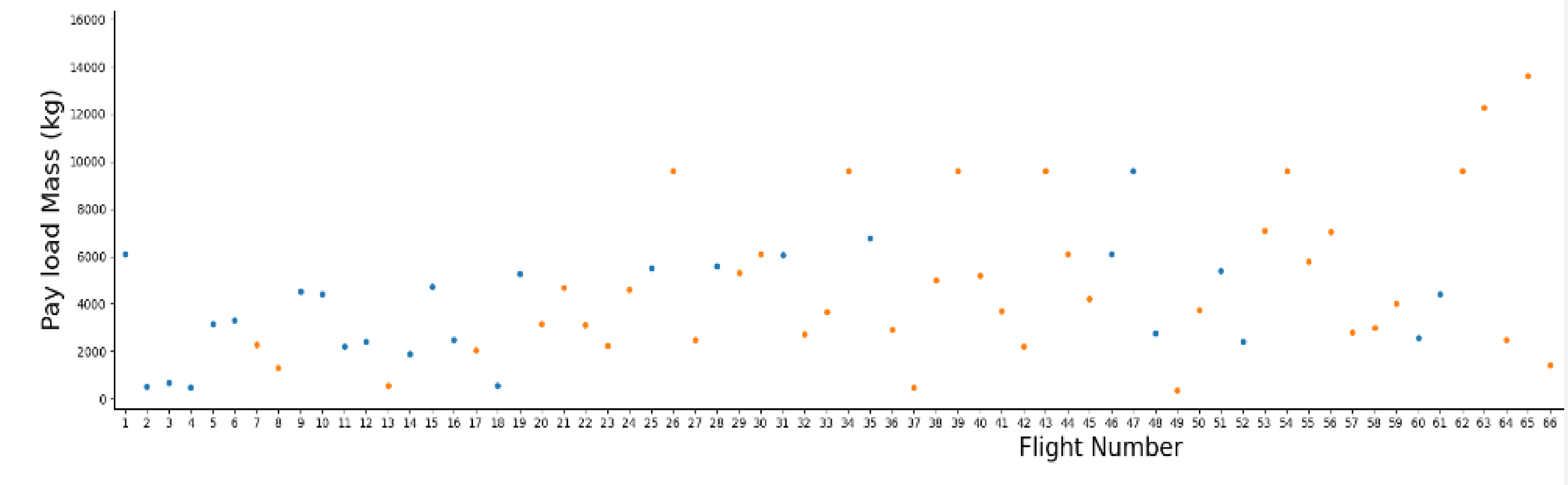
Data Wrangling

- We did Exploratory Data Analysis (EDA) on the dataset, calculating features such as : specific orbit appearances and mission outcome
- We, then, created an extra label, the outcome label, according to the flowchart.
- GitHub link: https://github.com/Konszygo/Coursera_Data_Science_Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb



EDA with Data Visualization

- Scatter plots and bar plots were used to visualize the relationship between different features of the dataset. One of the plots is given below:



- GitHub link: https://github.com/Konszygo/Coursera_Data_Science_Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- Summary of the SQL queries performed:
    - ✓ Names of the unique launch sites in the mission
    - ✓ Top 5 launch sites whose name begins 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 of the first successful landing outcome in ground pad
    - ✓ Names of the boosters which have success in drone ship and their payload mass is between 4000 and 6000 kilograms
    - ✓ Total number of successful and failure mission outcomes
    - ✓ Names of the booster versions which have carried the maximum payload mass
    - ✓ Records of month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015
    - ✓ Rank of the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order

- GitHub link: https://github.com/Konszygo/Coursera_Data_Science_Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite(1).ipynb

Build an Interactive Map with Folium

- We used map objects such as markers, circles, lines and marker clusters in the folium map created.
  - *Markers* indicate points such as launch sites
  - *Circles* indicate areas around some specific coordinates
  - *Lines* indicate the distance between two specific points in the map
  - *Marker clusters* indicate groups of events within a specific area

- GitHub link: https://github.com/Konszygo/Coursera_Data_Science_Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

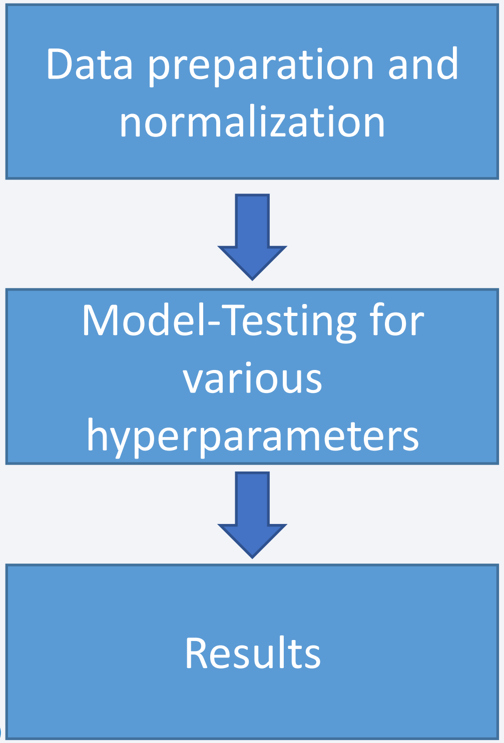
Build a Dashboard with Plotly Dash

- We added the following plots and graphs to the dashboard:
  - Percentage of launches by site
  - Payload range
- The plots mentioned above helped us recognize what is the best place to launch the rocket according to payload mass.

- GitHub link: https://github.com/Konszygo/Coursera_Data_Science_Capstone/blob/main/Ploty_dash.txt

Predictive Analysis (Classification)

- Regarding classification algorithms, we built four different models, evaluated their parameters and compared the corresponding results. The models used are the following:
  - ❖ Logistic regression
  - ❖ Support vector machine
  - ❖ Decision tree
  - ❖ K nearest neighbors



- GitHub link:

https://github.com/Konszygo/Coursera_Data_Science_Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

- Exploratory data analysis results:
  - Space X has 4 different launch sites
  - The average payload for the F9 v1.1 booster is 2,2928 kg
  - The first successful landing happened in 2015
  - Falcon 9 booster versions were successful in drone ships landing
  - Most of the missions had successful outcomes
  - The number of landing outcomes increased with the passage of years

Results

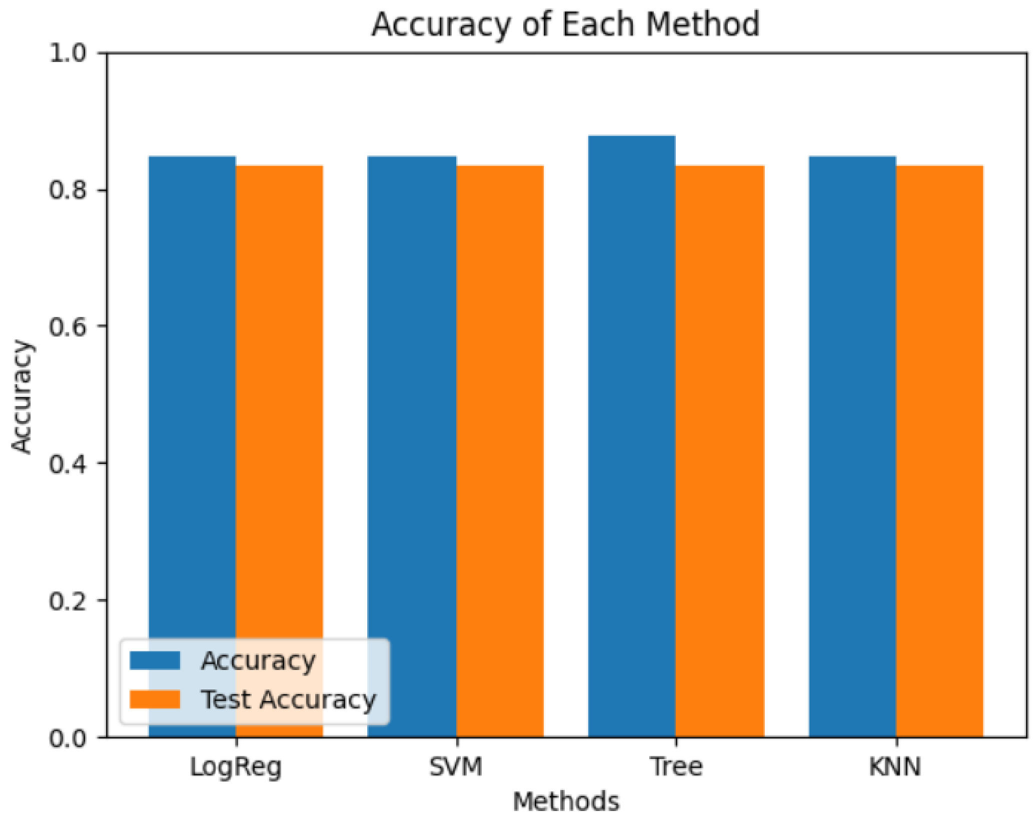
- Interactive analytics showed that most launch sites are in safe places, for example near the sea. Below, we share a screenshot depicting this fact.

Results

- Predictive analysis results

The analysis showed that the best model to predict successful landings is the Decision Tree classifier, with an accuracy of ~88% and a test accuracy of ~83%.

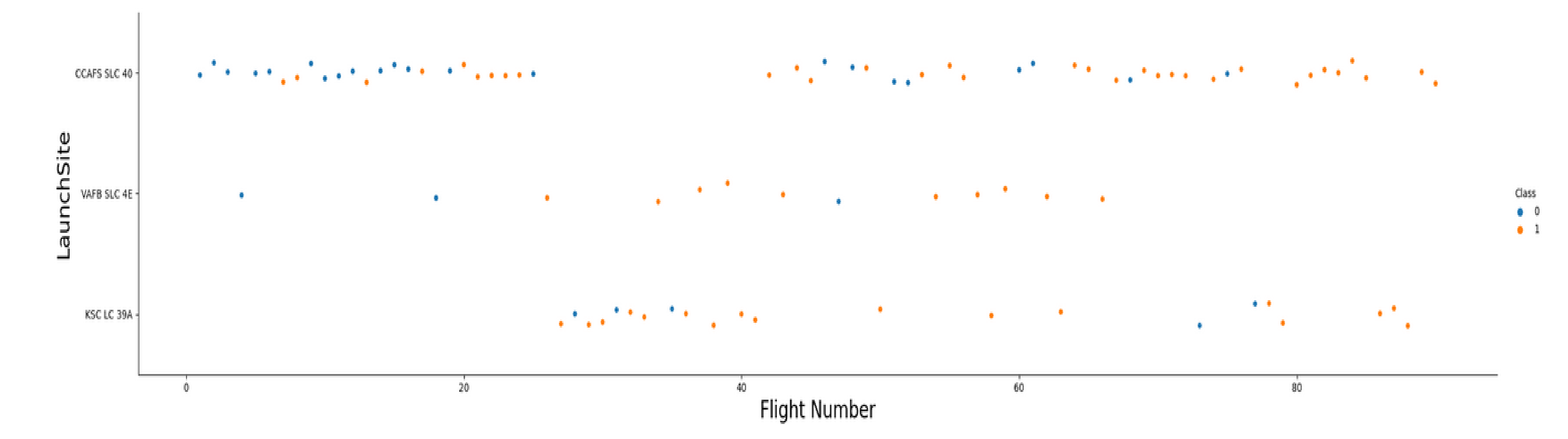
Model	Accuracy	TestAccuracy
LogReg	0.84643	0.83333
SVM	0.84821	0.83333
Tree	0.87679	0.83333
KNN	0.84821	0.83333



Section 2

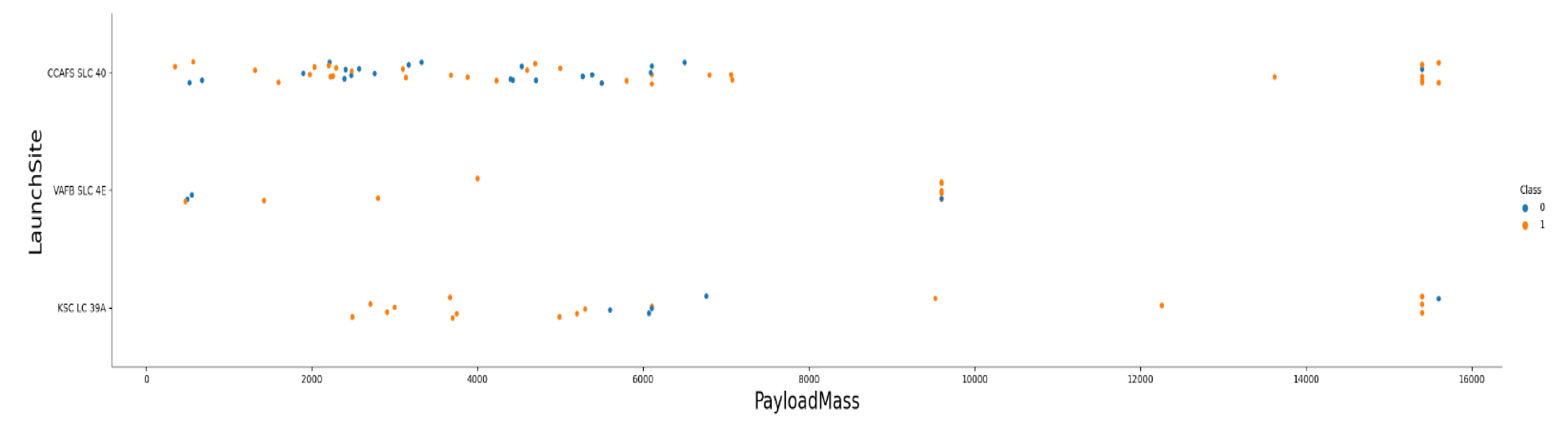
Insights drawn from EDA

Flight Number vs. Launch Site



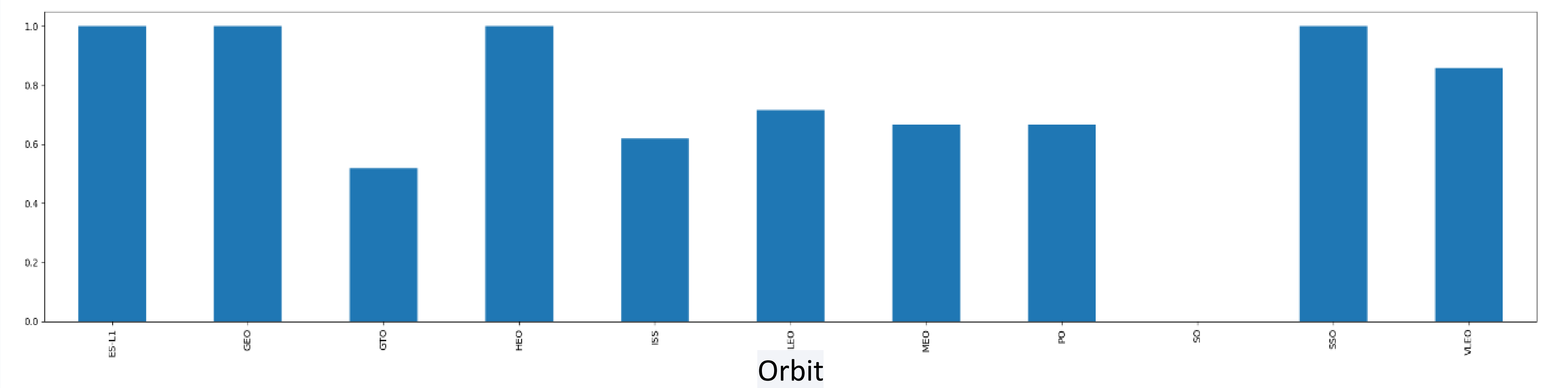
- In general, we can observe that success rate improved over time, as most of the recent launches belong to class 1.
- According to the plot above, the best launch site in our days is CCAF5 SLC 40 , followed by VAFB SLC 4E , as far as successful launches are concerned.

Payload vs. Launch Site



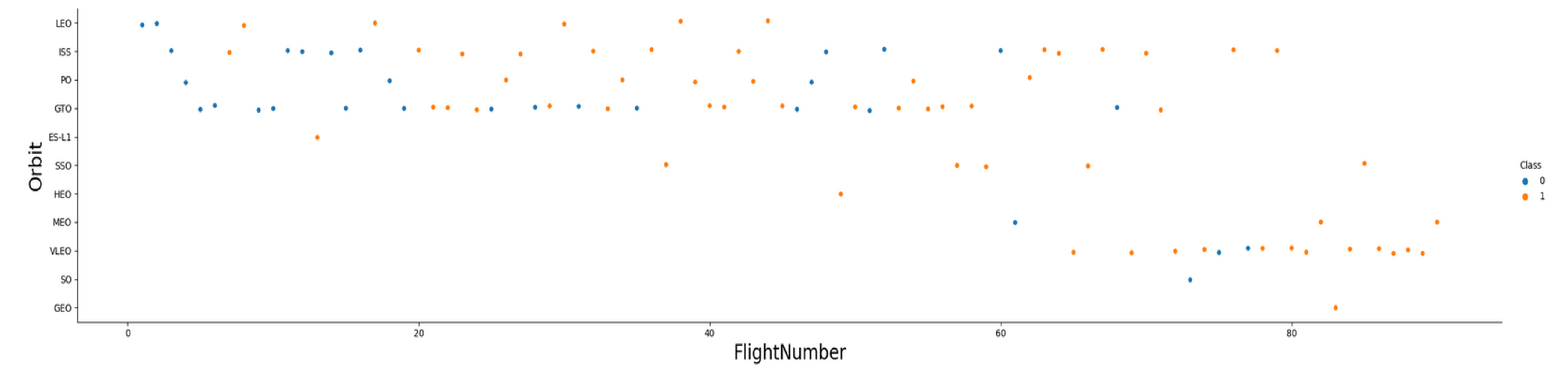
- From the plot above, we can observe that success rate greatly increases for payload mass over 8,000kg.
- Heavy payloads (over 12,000kg) is possible only on KSC LC 39A and CCAFS SLC40.
- Launch site KSC LC 39A seems the most effective one on low payload masses.

Success Rate vs. Orbit Type



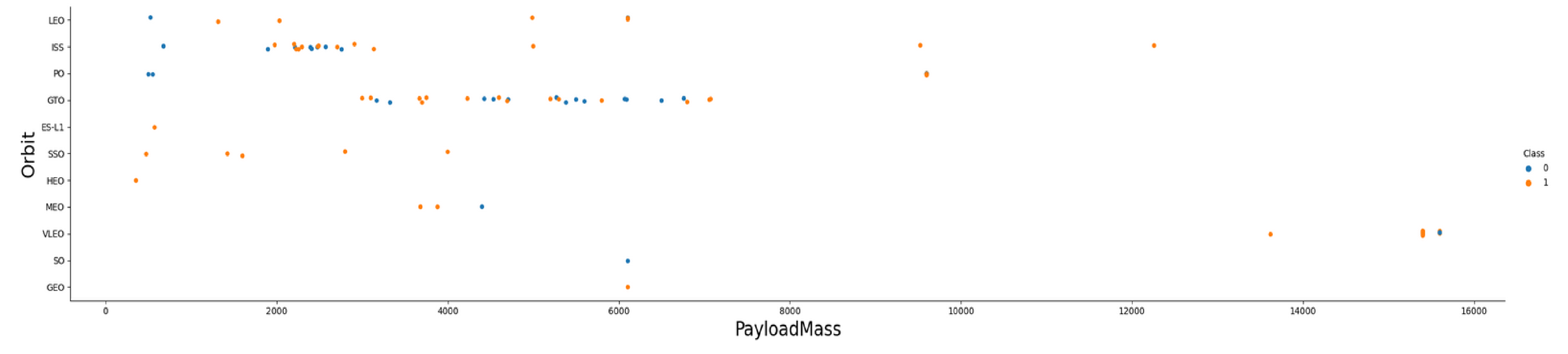
- From the bar chart above, we can see that the leftmost and rightmost orbits are those which have the highest success rate. Namely, those orbits are:
  ES-L1, GEO, HEO, SSO and VLEO

Flight Number vs. Orbit Type



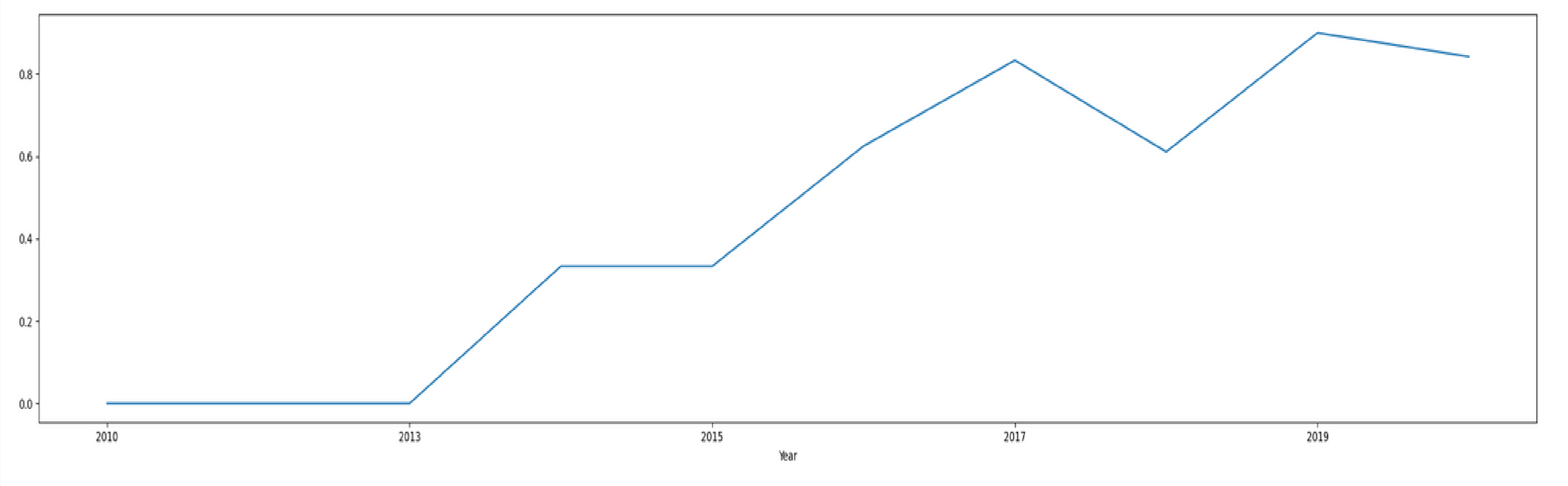
- It is clear that success rate improved for all orbits over time, as most of the class 0 launches belong to the first Flight Numbers.
- VLEO seems to be a new type of orbit, as it exists after Flight Number 65 and it also seems to have a high success rate.

Payload vs. Orbit Type



- As we can see, we cannot make any clear observations about the success rate from this kind of plot, where payload mass is plotted against orbit type.
- We observe that ISS and GTO are the two most common orbits.

Launch Success Yearly Trend



- Success rate greatly increased in years 2013, 2015 and 2016. There was a decrease in success rate in 2017, while it kept going upwards after that.

All Launch Site Names

- In our dataset, there are four launch sites:

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

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
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Total payload carried by boosters from NASA:

TOTAL_PAYLOAD
111268

- The total payload for NASA boosters is calculated by summing all payloads that contain ‘CRS’ in their names, counted in kilograms.

Average Payload Mass by F9 v1.1

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

AVG_PAYLOAD
2928.4

- The average payload mass calculated for launches with booster version F9 v1.1 is shown above, counted in kilograms.

First Successful Ground Landing Date

- Date of the first successful landing outcome on ground pad:

Date of 1st landing
2015-12-22

- The date of 1st successful landing outcome on ground pad shown above, is calculated after taking the minimum value in the date column.

Successful Drone Ship Landing with Payload between 4000 and 6000

- 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

- As shown above, there are four distinct booster versions.

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes:

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

- The table above shows us the different mission outcomes and their occurrences(QTY).

Boosters Carried Maximum Payload

- Booster versions 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
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 boosters shown above are those which have carried the maximum payload mass.

2015 Launch Records

- Failed landing outcomes in drone ship, booster versions and launch site names in year 2015:

Booster version	Launch site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- There are two booster versions for the above filters.

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

- Ranking of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20:

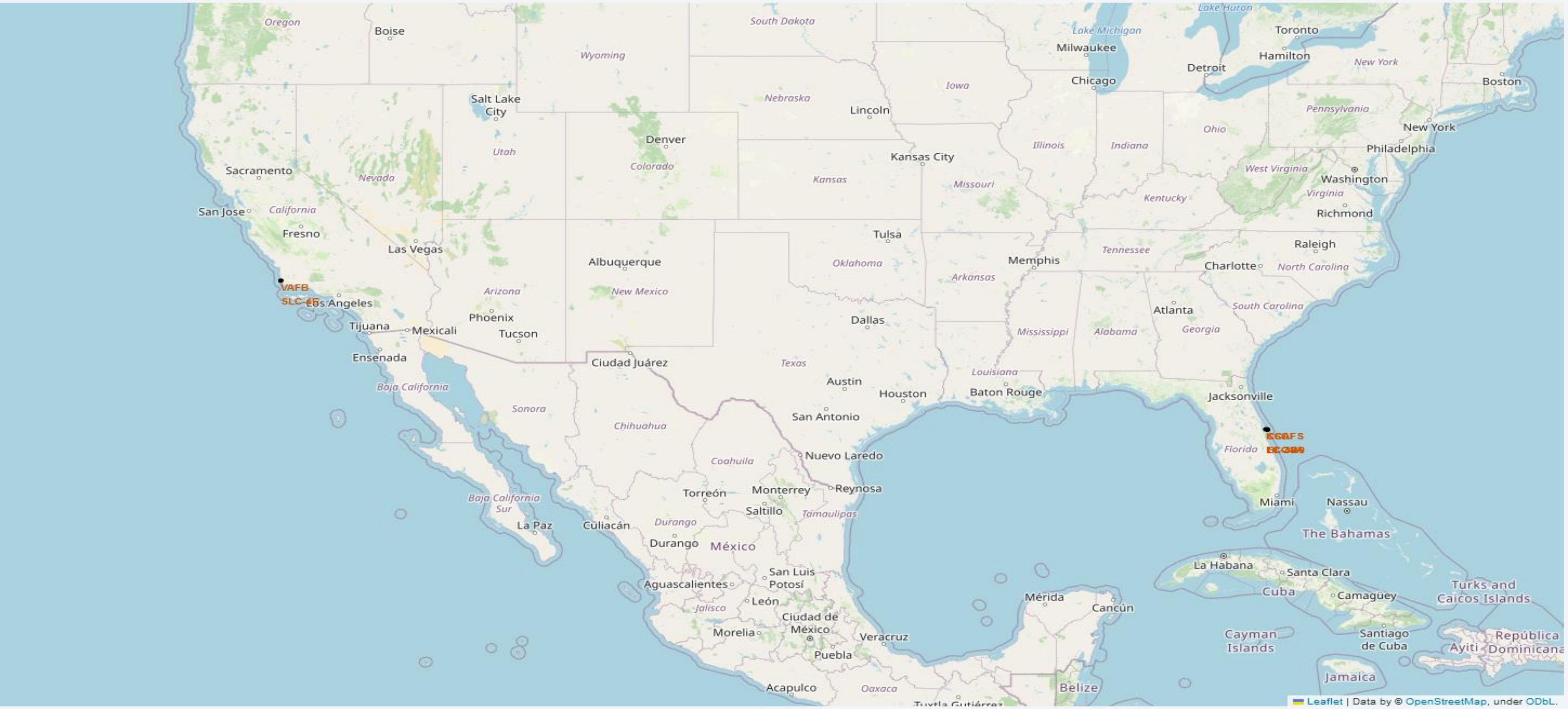
Landing outcome	Quantity
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (drone ship)	2
Precluded (drone ship)	1

- From the table shown above, we can observe that many of the landing outcomes between the requested dates were not attempted at all.

Section 3

Launch Sites Proximities Analysis

Launch Sites

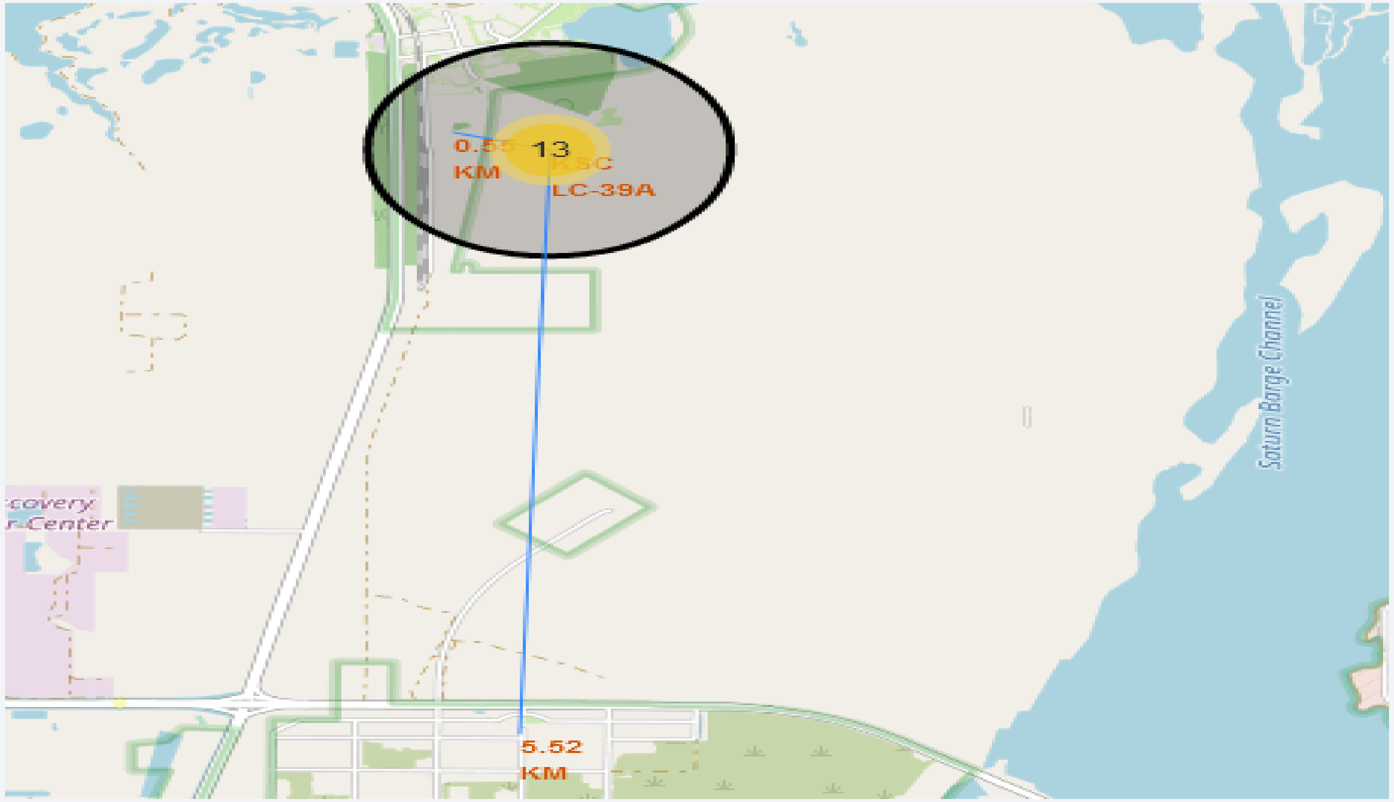


- As we can see, all launch sites are placed close to the sea, which is probably done for safety reasons, in case a rocket launch does not have a successful outcome.

Color labeled launch outcomes

- The screenshots above indicate zooming to the picture as we're going to the right.
- The green marks show successful landings while the red ones show failure landings.

Launch site proximity to its environment

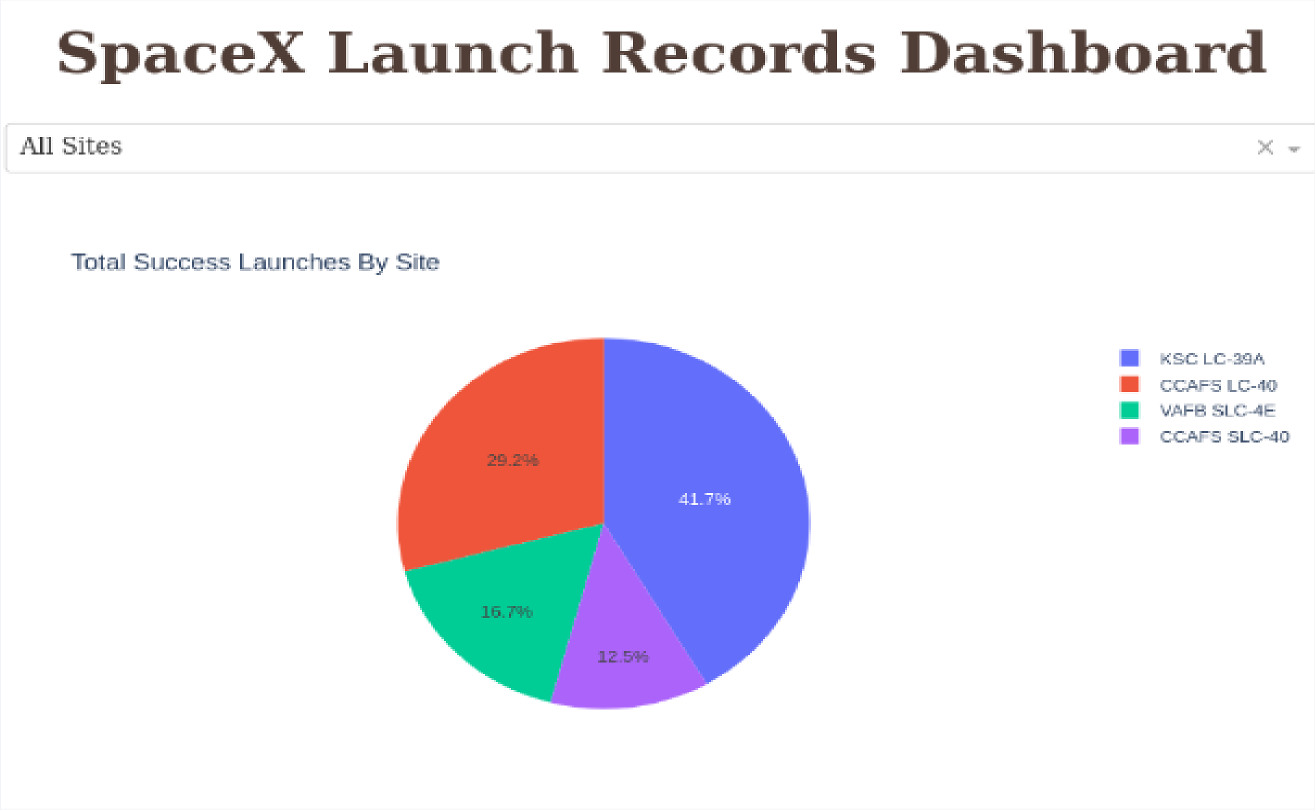


- In the Folium map shown above, we calculated a specific launch site's distance from the closest railway (0.55km) and from the closest inhabited area (5.52km).

Section 4

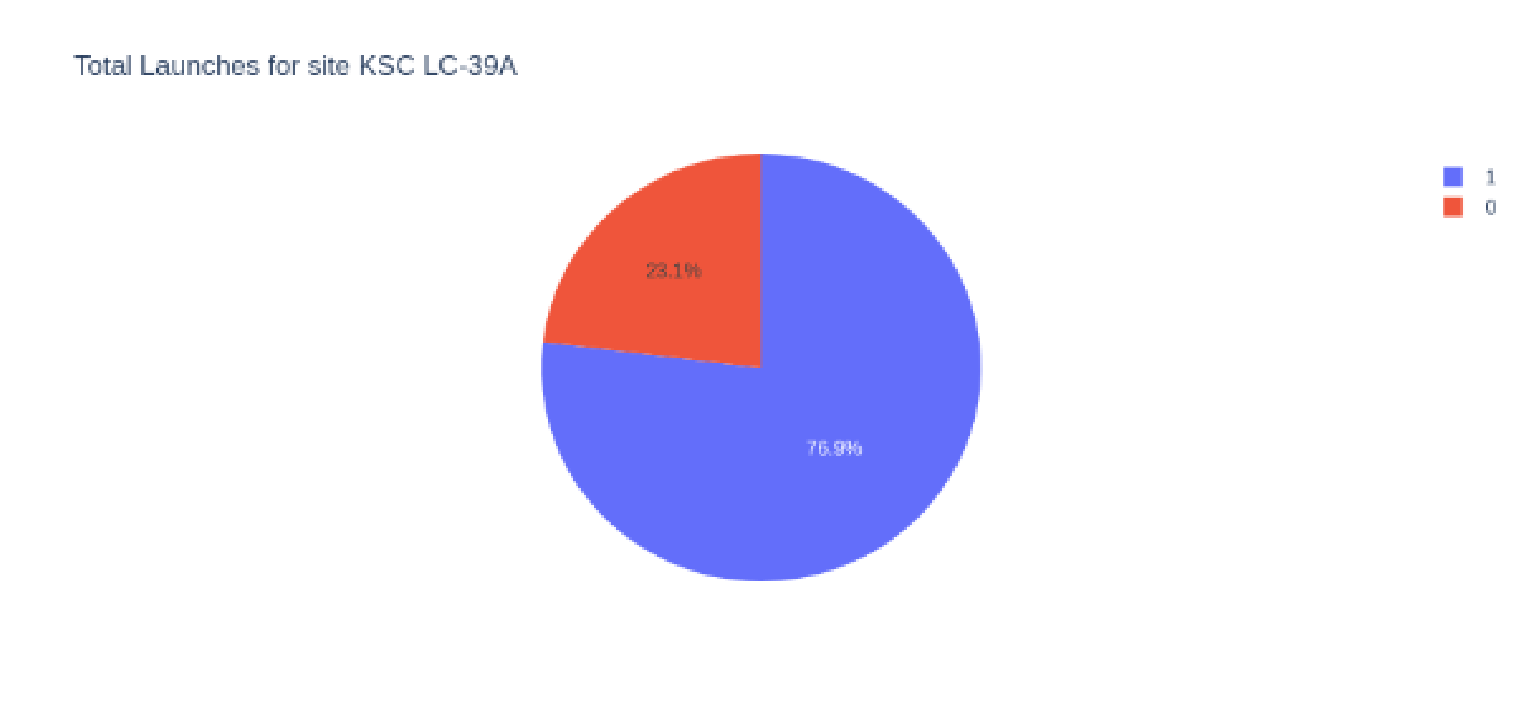
Build a Dashboard with Plotly Dash

Launch success count – Pie chart



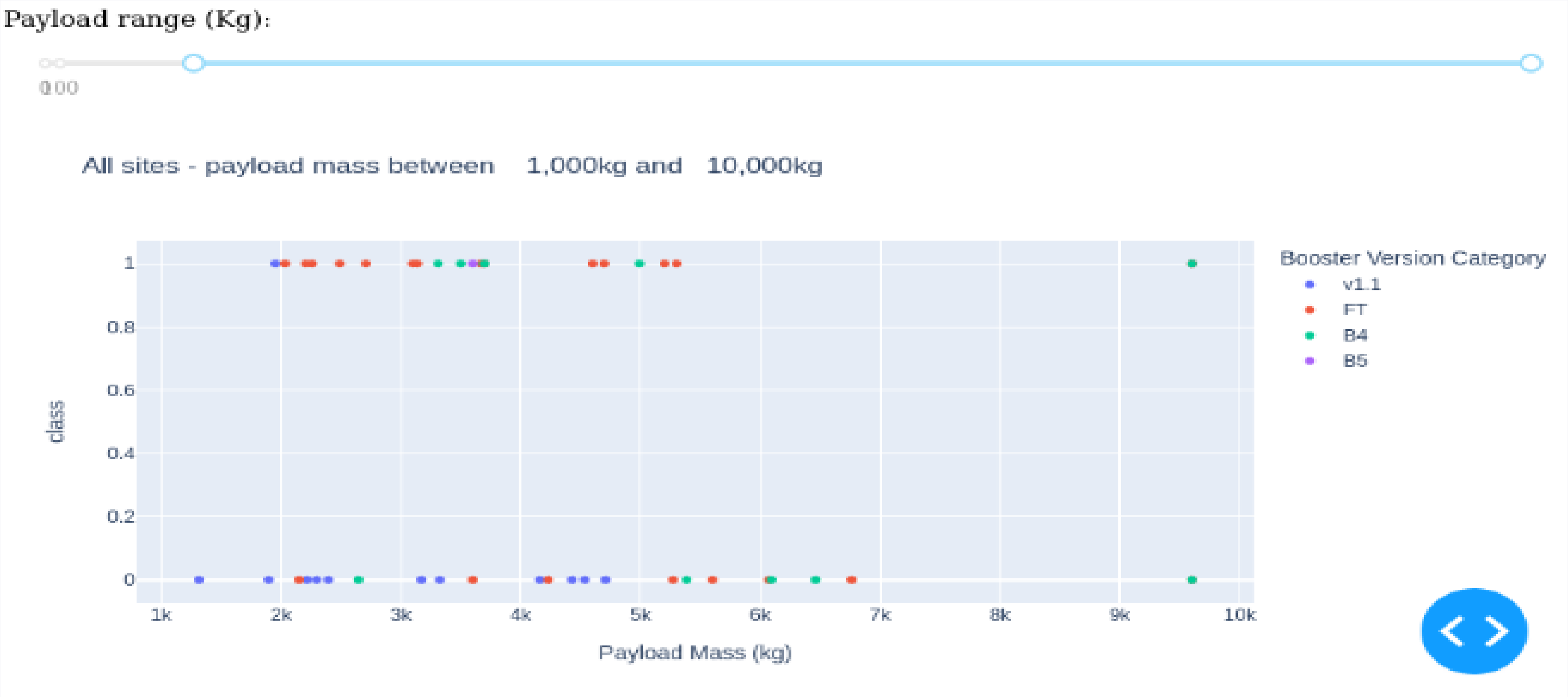
- Pie chart showing successful launches' percentage for various launch sites
- There seems to be a relationship between successful launches and launch site.

KSC LC-39A launch success pie chart



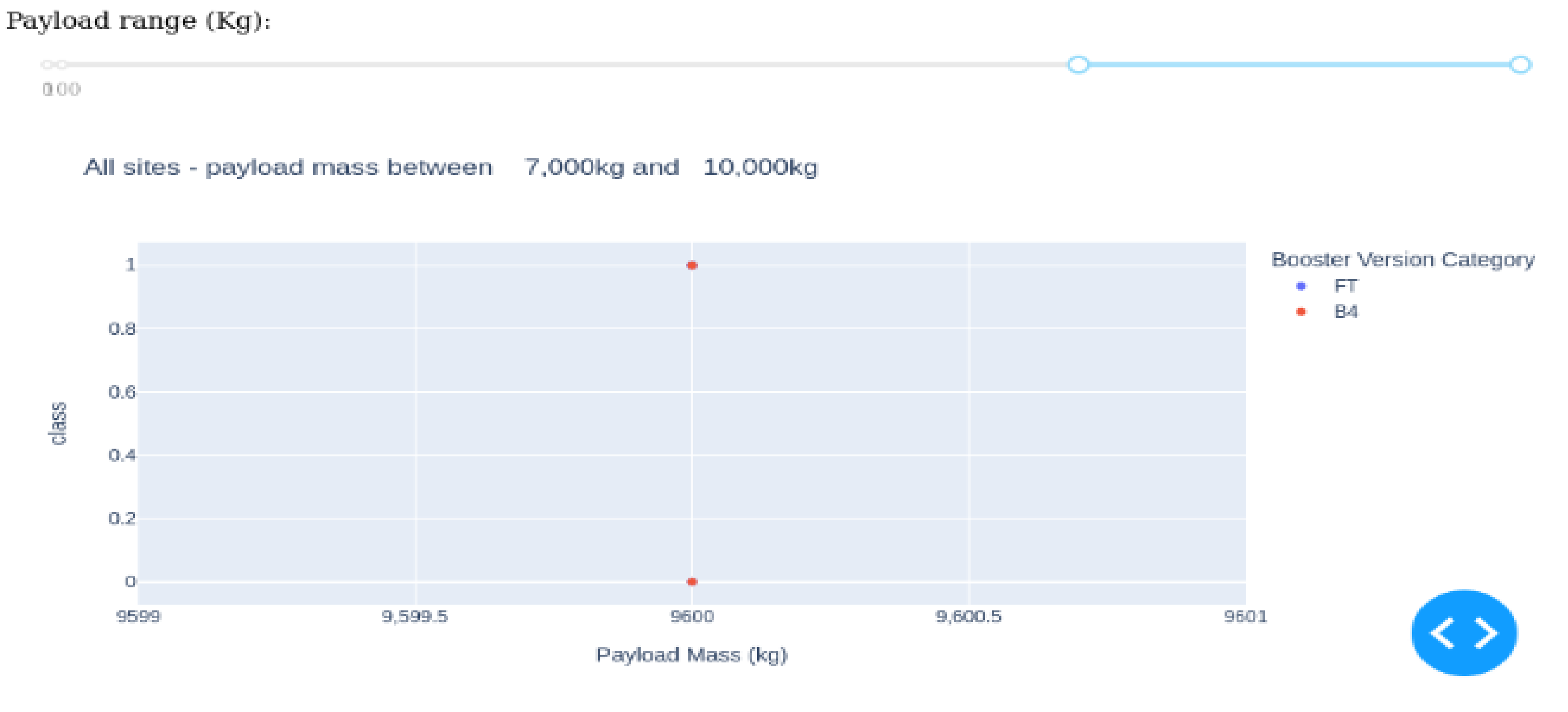
- The previous pie chart showed that KSC LC-39A was by far the most successful launch site.
- We can see that , for this site, 76.9% of the total launches were successful.

Payload vs. Launch Outcome scatter plot (1)



- From the scatter plot above, we conclude that FT booster version with payload mass under 5,500kg seems to be the most successful combination of booster version plus payload mass.

Payload vs. Launch Outcome scatter plot (2)

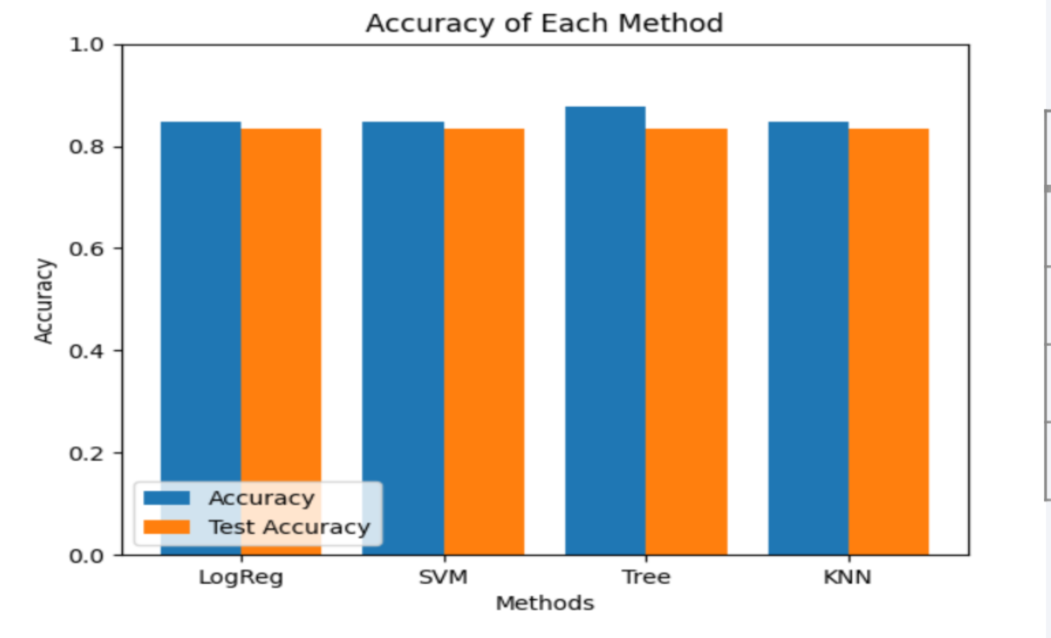


- It is apparent from this scatter plot, that there are not enough data for payload mass greater than 7,000kg , so we cannot make any reasonable conclusions for these cases.

Section 5

Predictive Analysis (Classification)

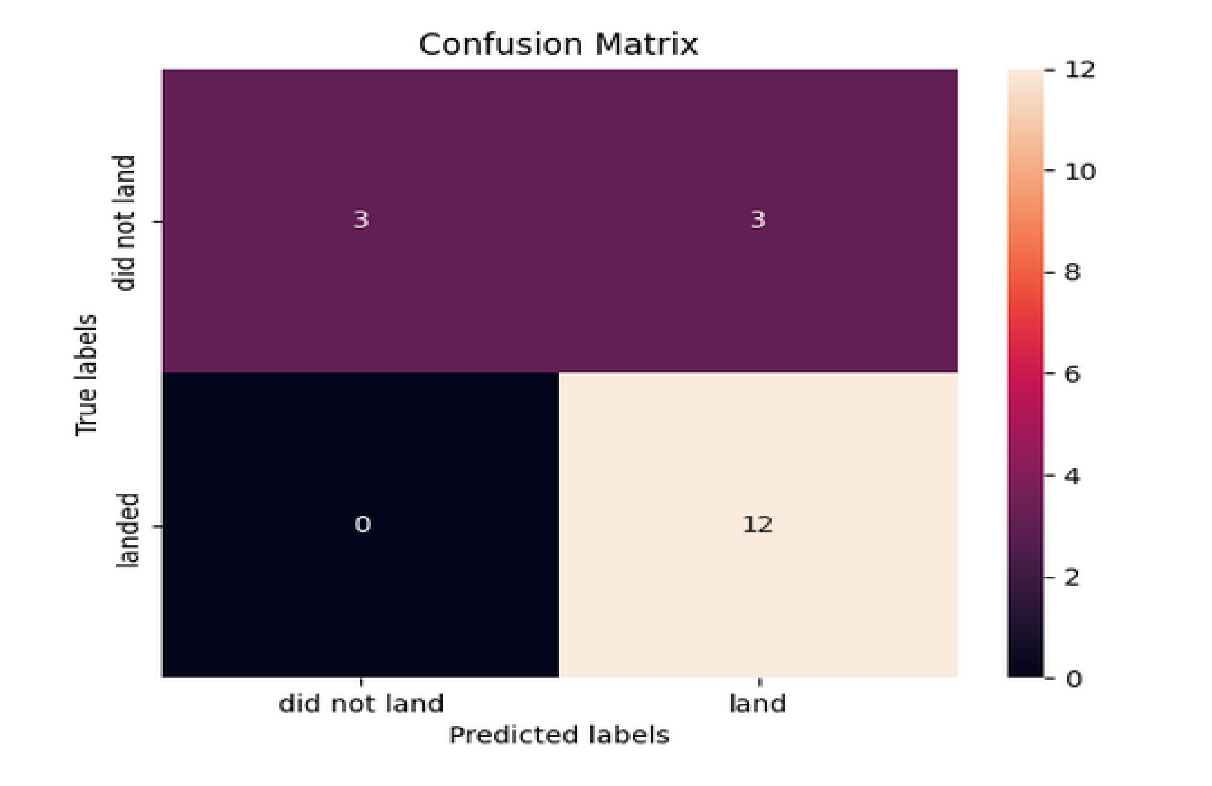
Classification Accuracy



Model	Accuracy	Test Accuracy
LogReg	0.84643	0.83333
SVM	0.84821	0.83333
Tree	0.87679	0.83333
KNN	0.84821	0.83333

- In the plot and corresponding table above, we see for each of our four classification models, the Decision Tree classifier was the one with the highest accuracy on the training set. (that is, 87.679%)
- On the other side, all four of our models had the same accuracy on the test set. (83.3%)

Confusion Matrix



- The confusion matrix above corresponds to the decision tree classifier, our best performing model, as shown before. As it is clear, while there are no false negative results, however there are some false positive results, which means that there is still a margin for improvement for our model.

Conclusions

- We consider KSC LC-39A to be the best launch site.
- There seems to be a great improvement over time regarding launch outcomes, which is a reasonable result, as the process and quality of the rockets are getting better and better.
- FT booster version with payload mass under 5,500kg is a good combination of booster version and payload mass.
- The best launch sites are those which are close to the sea and isolated from inhabited areas.
- Decision Tree classifier is the best model to predict successful landings.

Appendix

- Ploty Dash lab environment was not working properly, so there is a txt file of the code used, uploaded on GitHub. The code was run in PyCharm IDE environment.

Thank you!