IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Yunus Emre Gündüz
November 18, 2022

Outline

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

Executive Summary

- The methodologies used to analyze data:
  - Data collection using web scraping from Wikipedia [1] and using SpaceX api [2]
  - Data wrangling including filling null spaces and merging several data resources
  - Exploratory data analysis (EDA) by
    - visualizing the relationships between different features of the data
    - SQL queries
    - Locating launch sites on the map using folium
    - Creating interactive dashboard to enable data to be inspected easily

Executive Summary

- The following results are acquired:
  - Exploratory data analysis gives us insights about the rocket launches and relationship of different features in the dataset.
  - Interactive dashboard can be used to inspect the correlation between success rate and payload mass.
  - SpaceX's first stage landing can be predicted that if it will fail or succeed.
  - Predictions which made by different classification algorithms result in close accuracy rate which is about %85.

Introduction

- A new commercial rocket launcher company
    - wants to bid against SpaceX
    - needs to advertise their rocket launches with a lower-cost
    - decided to analyze Falcon 9 launches to find a way

- Problems looking for answers:
    - The best way to predict whether the first stage lands successfully or not, which directly impacts on the cost
    - The best place to execute rocket launches

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - web scrapping from Wikipedia's "List of Falcon 9 and Falcon Heavy Launches" page [1]
  - requesting data using SpaceX's API [2]
- Perform data wrangling
  - The data is enriched by creating a landing outcome label
  - Orbits and launch sites are analyzed
- Perform exploratory data analysis (EDA) using visualization and SQL

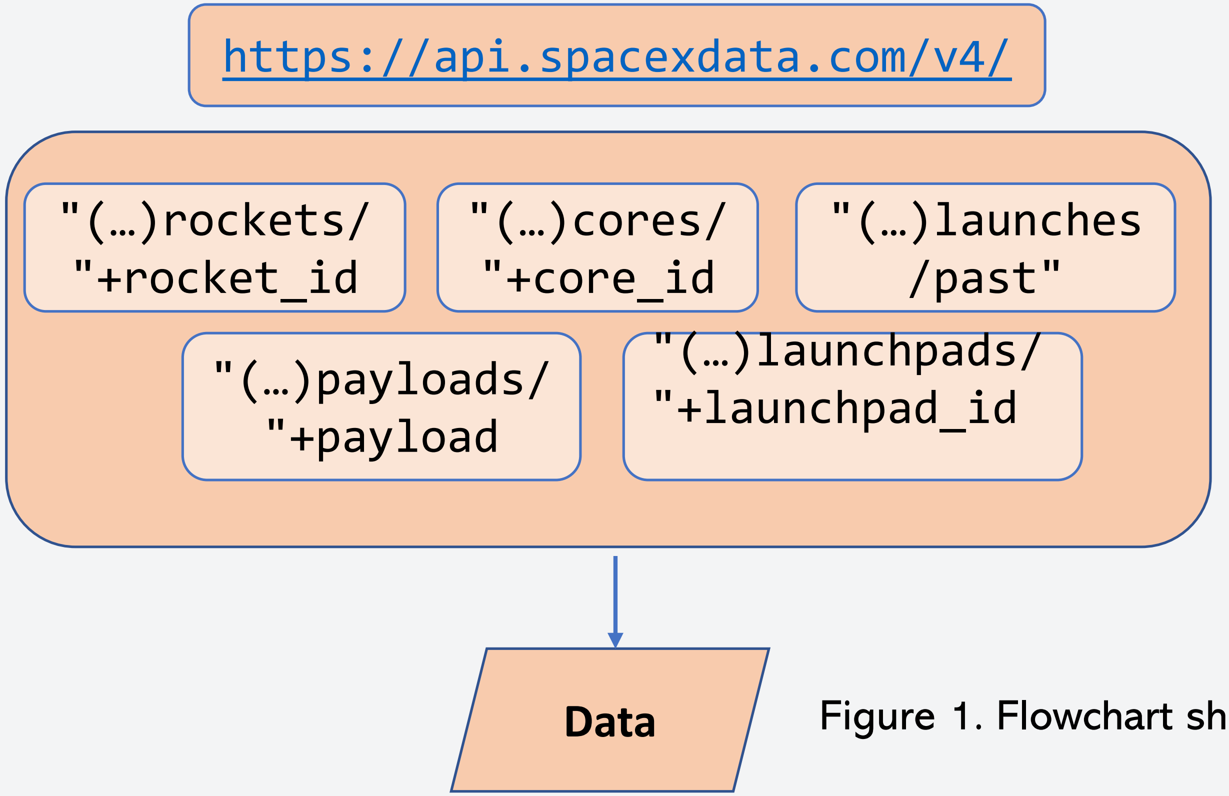
Methodology

Executive Summary

- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The data is standardized
  - Train-test split approach is used
  - The best hyperparameter for support vector machine (SVM), logistic regression, classification trees and k-nearest neighbors (KNN) is found
  - The best-performing method is obtained

Data Collection

- SpaceX's API [2] is called by requesting data from five different websites, which are combined:



https://api.spacexdata.com/v4/

"(…)rockets/ "+rocket_id

"(…)cores/ "+core_id

"(…)launches /past"

"(…)payloads/ "+payload

"(…)launchpads/ "+launchpad_id

Data

Figure 1. Flowchart showing SpaceX REST call process

- Wikipedia's "List of Falcon 9 and Heavy Launches" page is web scrapped

(https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922)

Note: Because the page is updated, to maintain consistency in the assignment, a snapshot is used.

Data Collection – SpaceX API

- The data contains 90 rows after filtering

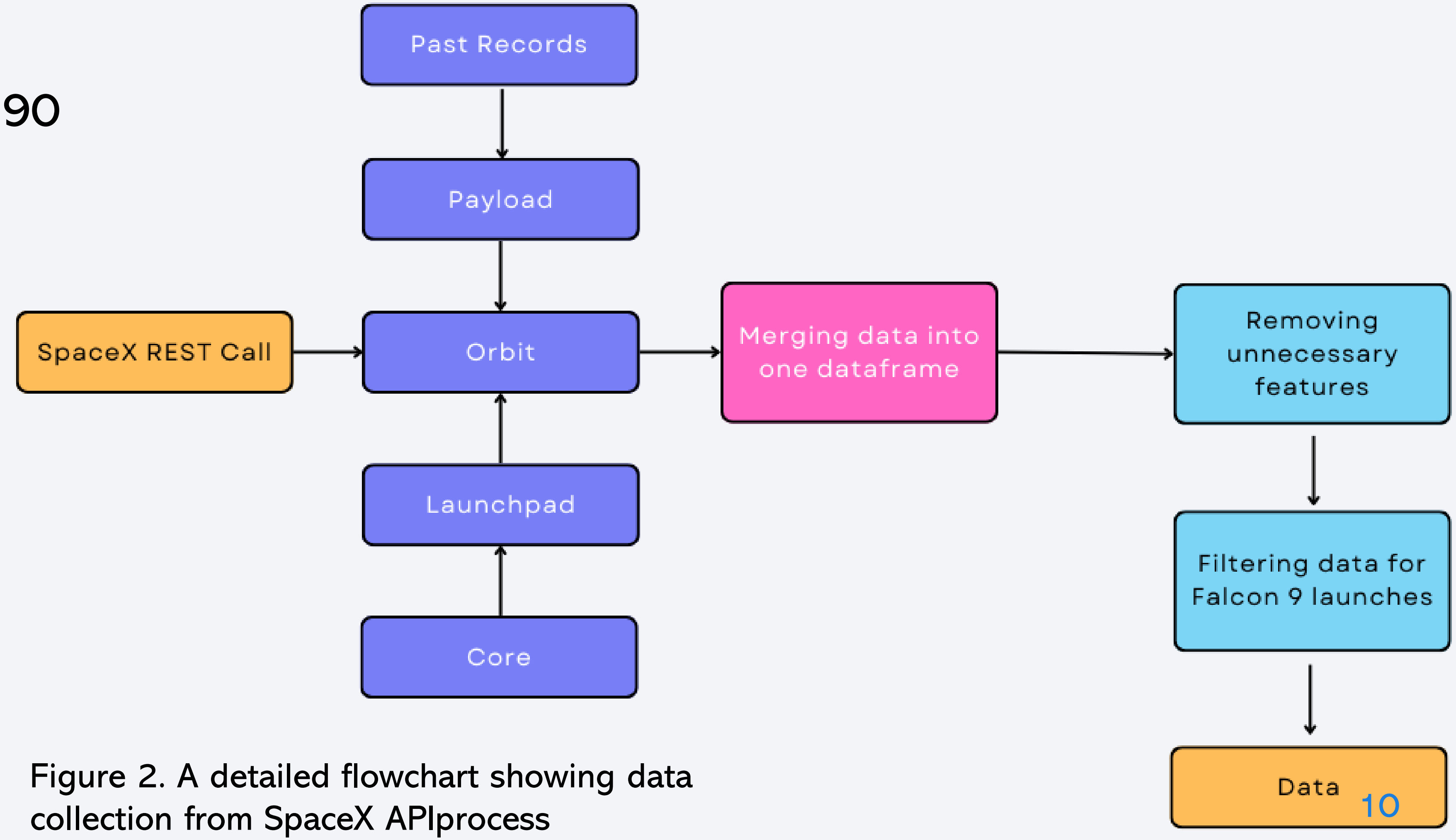

Past Records
Payload
SpaceX REST Call
Orbit
Launchpad
Core
Merging data into one dataframe
Removing unnecessary features
Filtering data for Falcon 9 launches
Data

Figure 2. A detailed flowchart showing data collection from SpaceX APIprocess

Link for code

Data Collection - Scraping

- The data contains 121 rows and 11 columns at the end

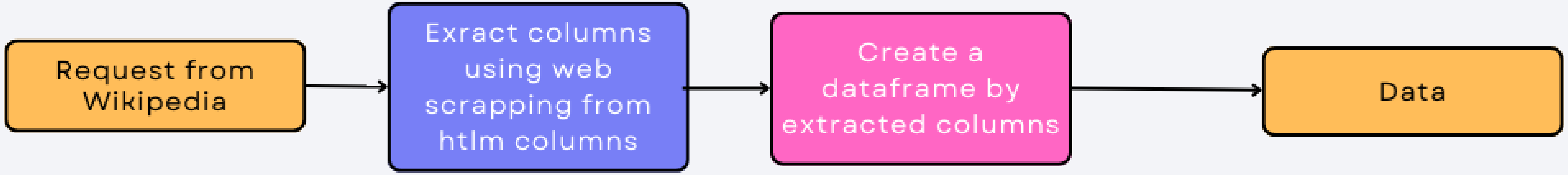

Request from Wikipedia
Exract columns using web scrapping from htlm columns
Create a dataframe by extracted columns
Data

Figure 3. Flowchart showing web scrapping from Wikipedia

Link for code

Data Wrangling

- The data, collected by SpaceX API, is used
- Orbits, launch sites, and mission outcomes are summarized
- New feature, called «Class» is created based on mission outcomes as shown in figure 1

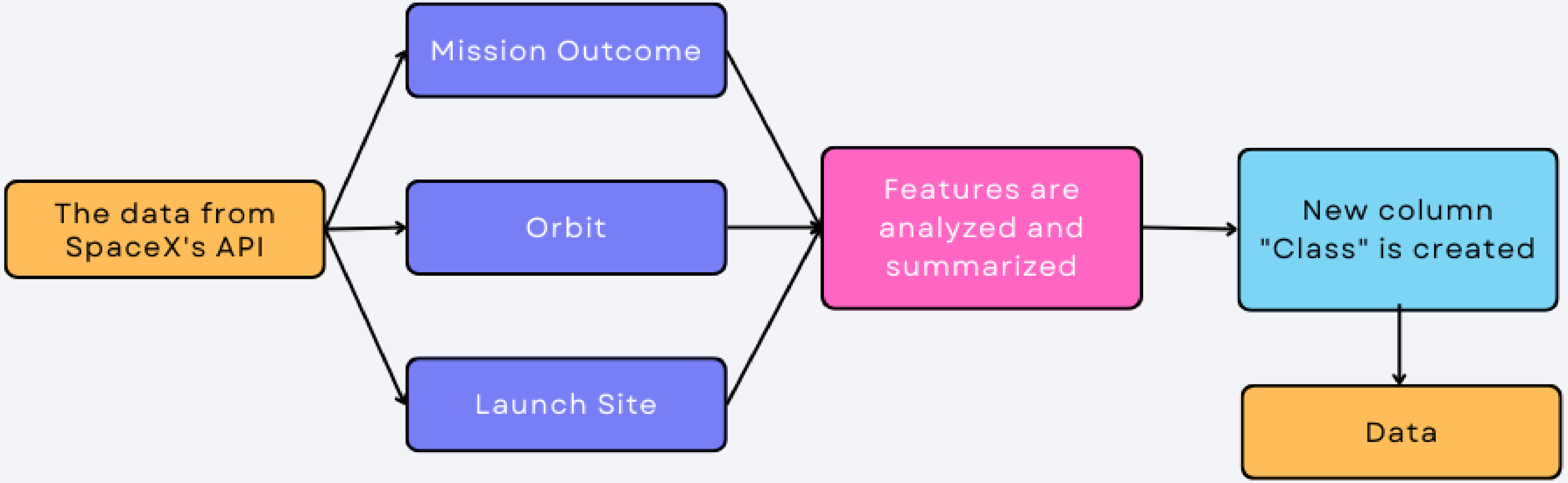
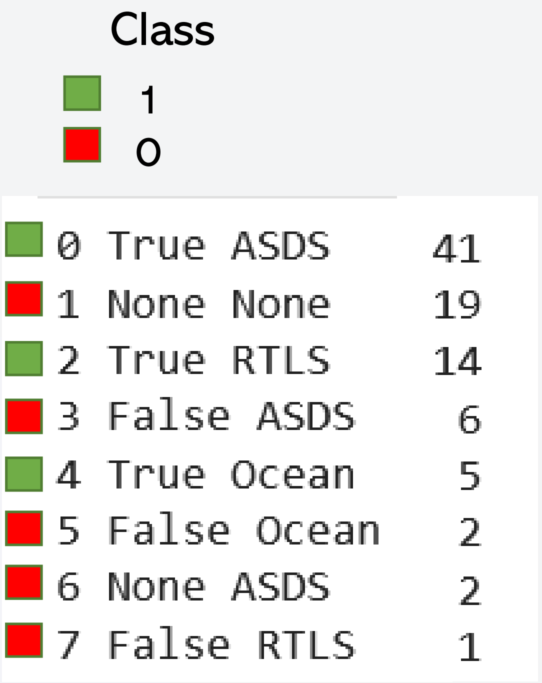

The data from SpaceX's API
Mission Outcome
Orbit
Launch Site
Features are analyzed and summarized
New column "Class" is created
Data

Figure 4. Flowchart showing data wrangling process



Class

🟩 1
🟥 0

```
🟩 0 True ASDS      41
🟥 1 None None      19
🟩 2 True RTLS      14
🟥 3 False ASDS      6
🟩 4 True Ocean      5
🟥 5 False Ocean     2
🟥 6 None ASDS       2
🟥 7 False RTLS      1
```

Figure 5. Mission Outcomes and their counts

Link for code

EDA with Data Visualization

- Different types of graphs and charts are created to analyze the relationships between the data as written on the right

- At the end, categorical columns are converted into dummy variables to enable prediction models to consider them

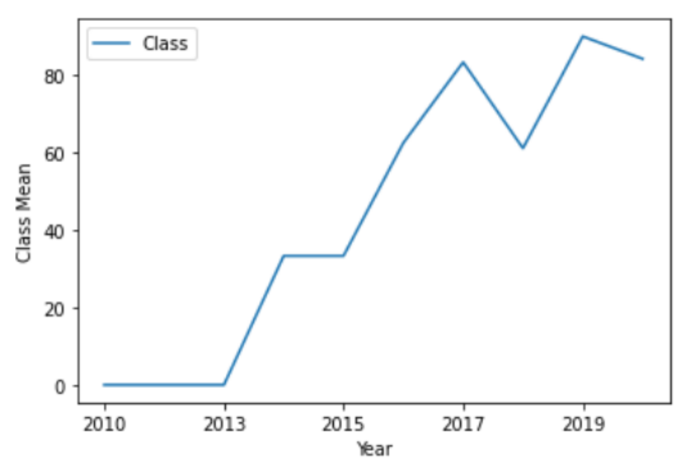


Figure 6. Success rate yearly trend

- Scatter plots:
  - Payload mass vs. Flight number
  - Launch sites vs. Flight number
  - Payload mass vs Launch site
  - Orbit type vs Flight number
  - Payload mass vs Orbit type
- Bar graph:
  - Success rate of different orbit types
- Line graph:
  - Launch success rate yearly trend

Link for code

EDA with SQL

- The data, fetched from Wikipedia’s page, is used
- All launch sites for Falcon 9 are detected
- 5 records where launch sites begin with the string 'CCA’ are displayed
- Total payload mass carried by boosters launched by NASA (CRS) is printed
- Average payload mass carried by booster version F9 v1.1 is calculated
- The date of first successful landing for ground pad is obtained
- The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000 are listed
- Mission outcomes are summarized
- The names of the booster versions which have carried the maximum payload mass are listed
- The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015 are extracted
- Landing outcomes and the number of them within specified dates are shown

Link for code

Build an Interactive Map with Folium

- Different launch sites are shown on the map with markers and circles
- The number of launches are displayed by clustering them with marker clusters in each launch sites
- The proximity of each launch site to the nearest coastline, highway, railway and city is calculated and shown with a polyline on the map
- Launch sites and their distances from mentioned locations are displayed on the map in order to:
    - To identify if they keep some certain distances to specific locations
    - To have an understanding of characteristics of an rocket launch site

Link for code

Build a Dashboard with Plotly Dash

- A pie chart is created
  - The success rate at each launch site is visualized
  - The launch sites are compared according to their successful first stage landings
- A scatter plot is added
  - The number of successful and failing landings within a specified payload mass is plotted with their booster versions

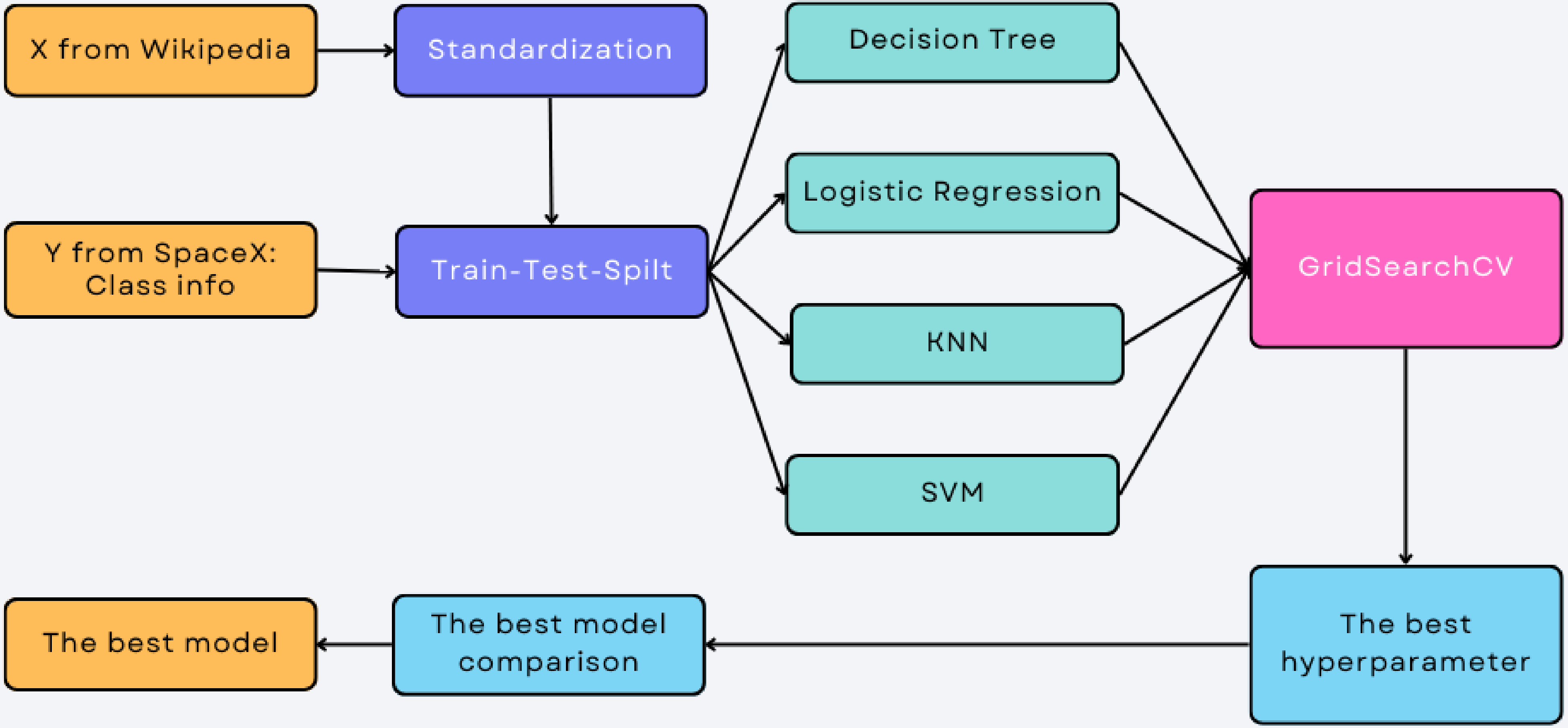
Link for code

Predictive Analysis (Classification)

- X contains standardized version of the data fetched from Wikipedia
- Y contains the class information if first stage landed succesfully from the data collected by calling SpaceX API
- 4 different classification models are built with SVM, KNN, logistic regression and classification trees.
- The best hyperparameters are found by grid searching
- Different models' accuracy on the test data are compared among each other to find the best model

Link for code

Predictive Analysis (Classification)



Link for code

Figure 7. Flowchart showing prediction proccess

Results

Exploratory data analysis results:

- SpaceX uses 4 different launch sites
- The average payload of F9 v1.1 is 2534 kg
- The success rate rises over years
- Payload mass has different impacts on the outcome for different ranges
- Only 1 mission outcome is in failure
- Orbits do not affect the launch outcome

Results

Interactive analytics

- An interactive dashboard is created

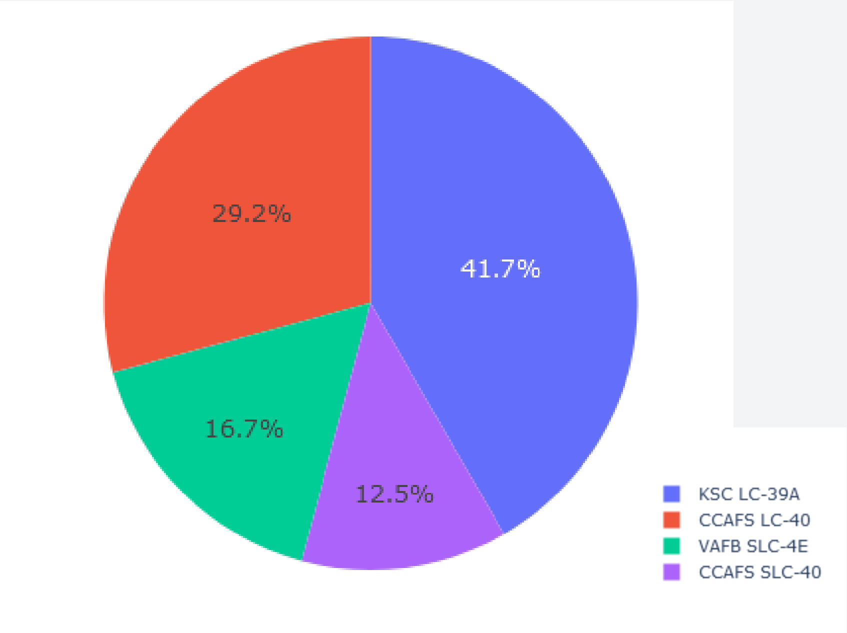


Figure 8. Success ratio for all launch sites

- Characteristics of launch sites are detected

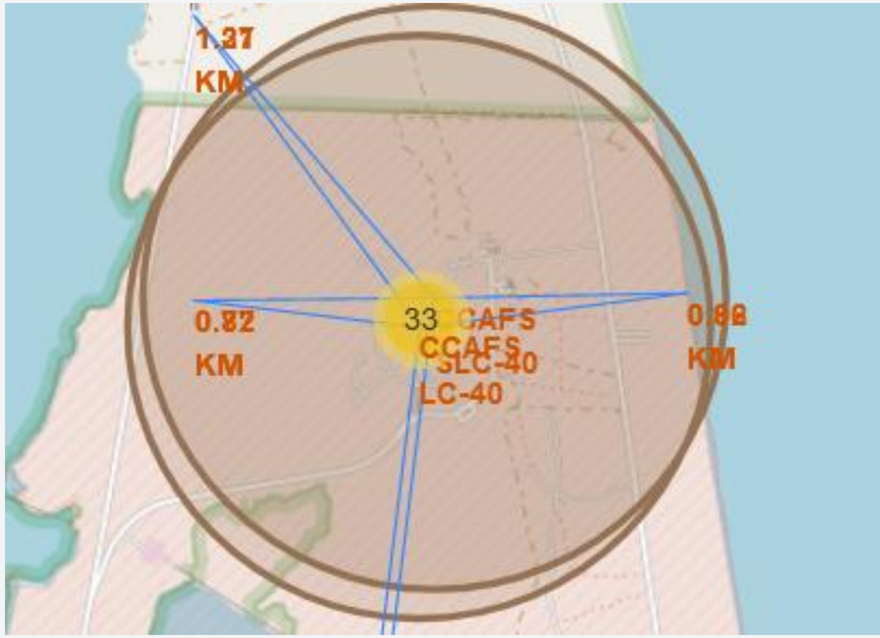


Figure 9. Distance from CCAFS SLC-40 to certain locations

Results

Predictive Analysis

- Decision Tree is found as the best model performing 94% accuracy on the test data

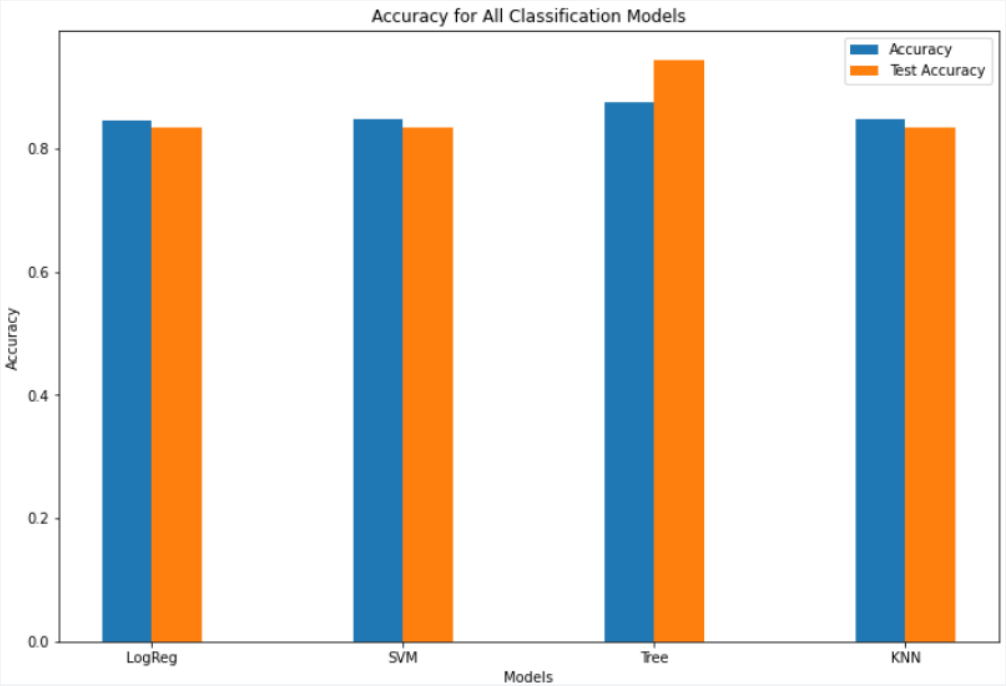


Figure 10. Accuracy of all models

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- CCAFS SLC 40 has the success rate of %60, while VAFB SLC 4E and KSC LC 39A has the success rate of %77 in total
- The general success rate rises over time
- Launches from CCAFS SLC 40 had a considerable imporements compared to its first flights

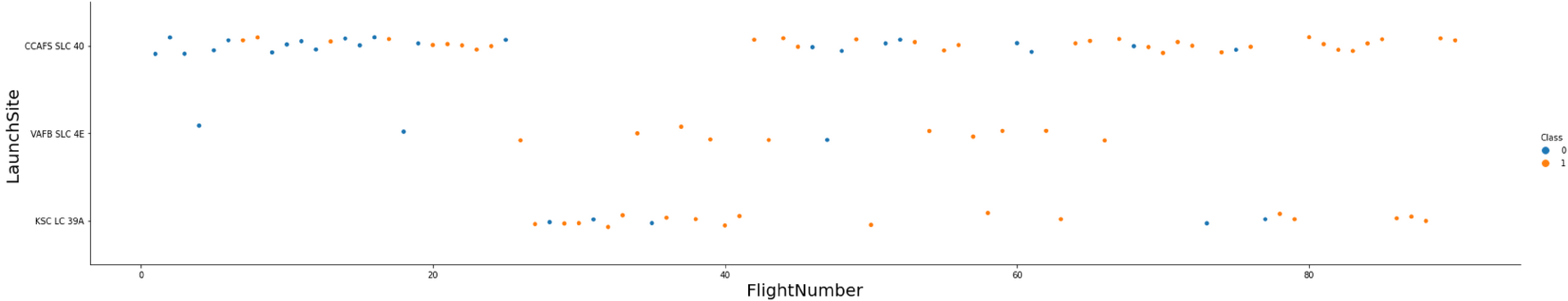


Figure 11. Flight Number vs. Launch Site

Payload vs. Launch Site

- No launches more massive then 10000 kg is launched from VAFB SLC 4E
- Common payload mass of rockets is gathered between 1000 and 8000
- Launches having the payload mass between 8000 and 15000 is considerably rare
- Launches massive then 7000 are less risky

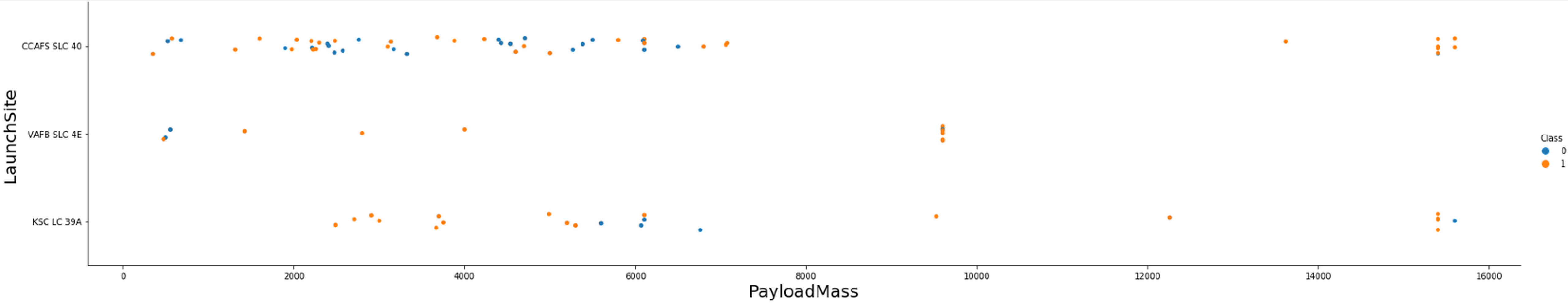


Figure 11. Payload Mass vs. Launch Site

Success Rate vs. Orbit Type

- The first stage of all rockets landed successfully, which are sent to ES-L1, GEO, HEO and SSO
- However, when the number of launches to each orbit is visualized, it can be seen that the most successful orbits has the lowest number of launches.
- The most commonly rocket-sent orbits have a success rate around 60%

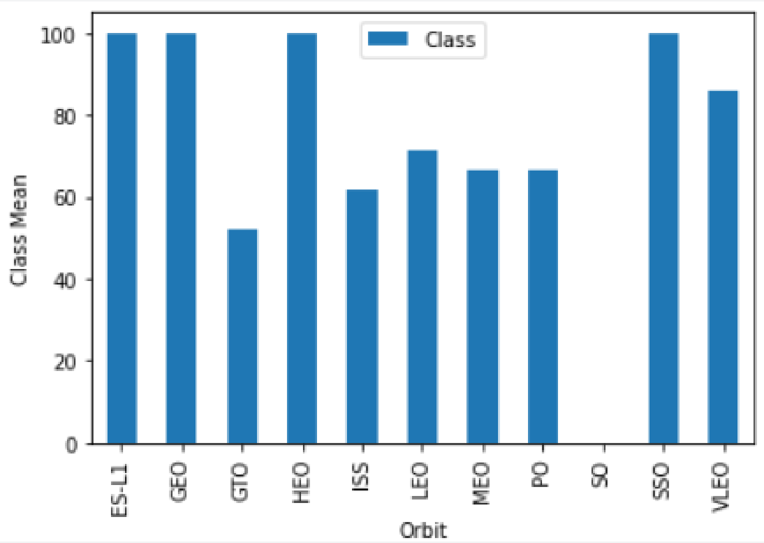


Figure 12. Success Rate vs. Orbit Type

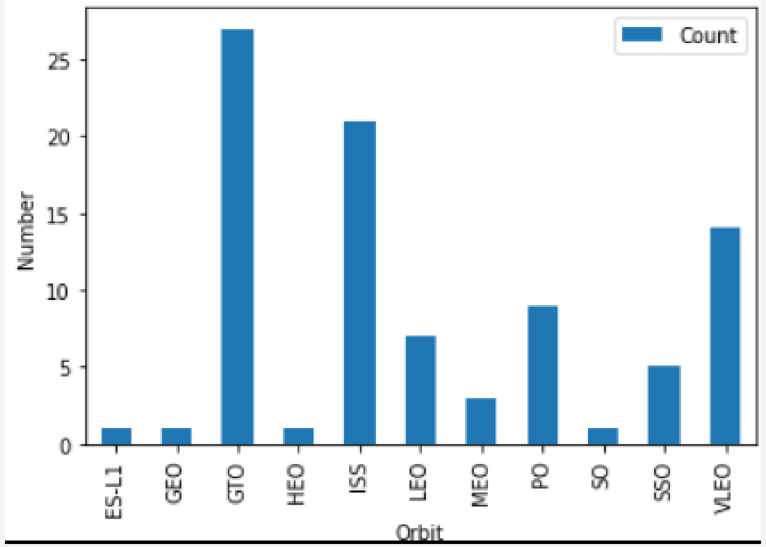


Figure 13. Orbit Type vs. Their Counts

Flight Number vs. Orbit Type

- VLEO, SO, GEO, MEO, HEO, SSO and ES-L1 orbits do not have any records in the early times of SpaceX Falcon9 launches
- VLEO is the one where launches are preferred to be sent lately
- The most of the Falcon 9 launches were sent o GTO, PO, ISS and LEO

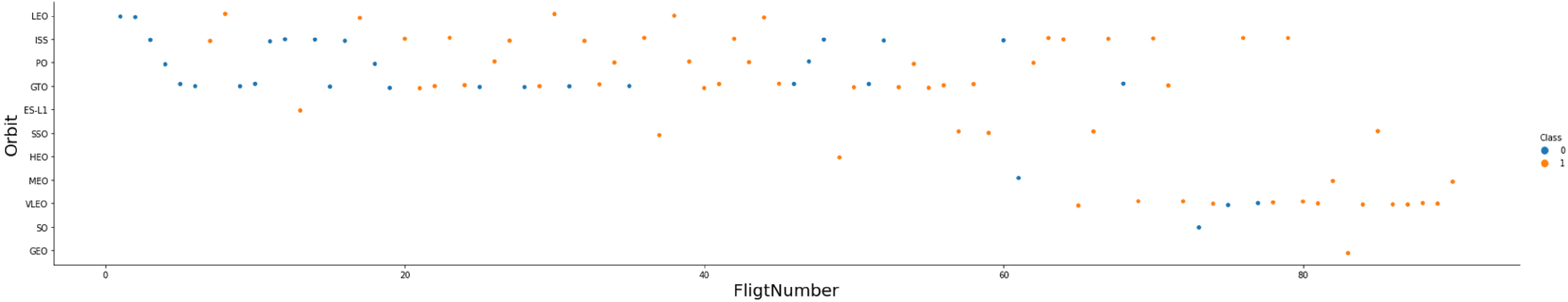


Figure 14. Flight Number vs. Orbit Type

Payload vs. Orbit Type

- VLEO receives the most massive launches
- GTO, the most used orbit, receives launches between 3000 and 9000
- There is no relation between the success rate and payload for GTO orbit
- ISS has the largest payload range among other orbits

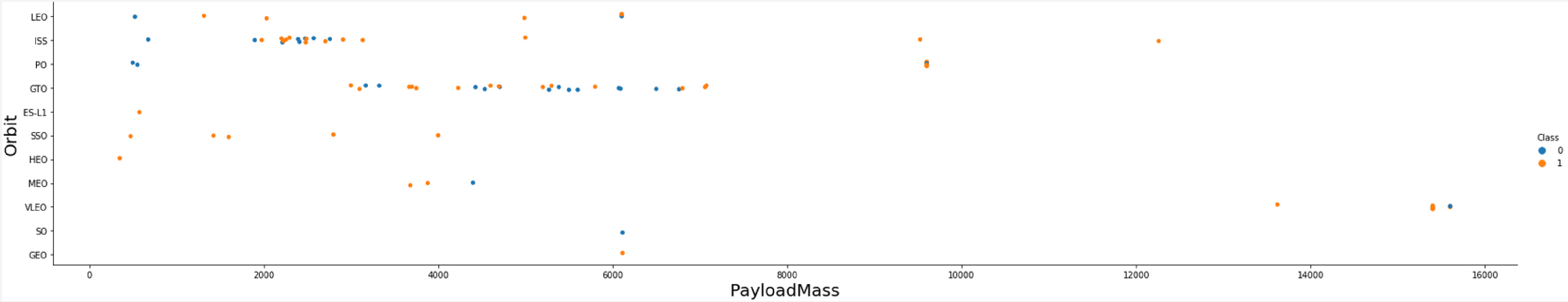


Figure 15. Payload Mass vs. Orbit Type

Launch Success Yearly Trend

- Apparently, the first stage landing success rises over the years
- The first years of first stage landing tries encountered with failure

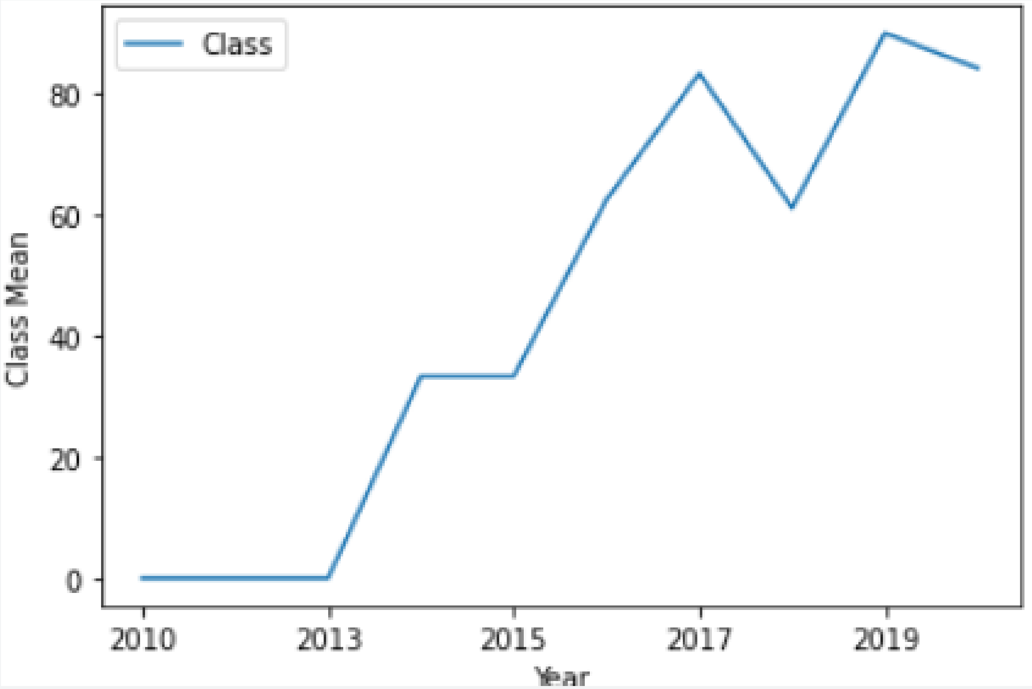
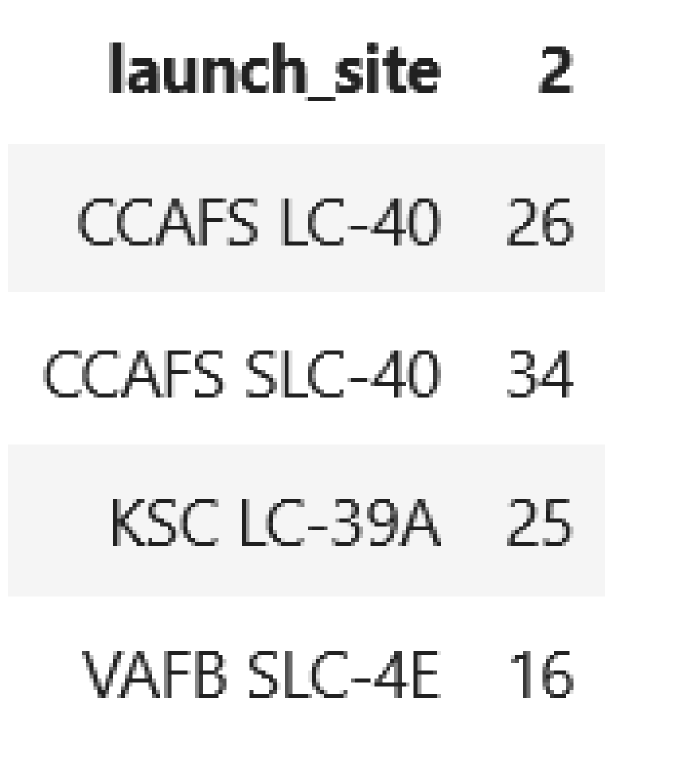


Figure 6. Launch Success Yearly Trend

All Launch Site Names

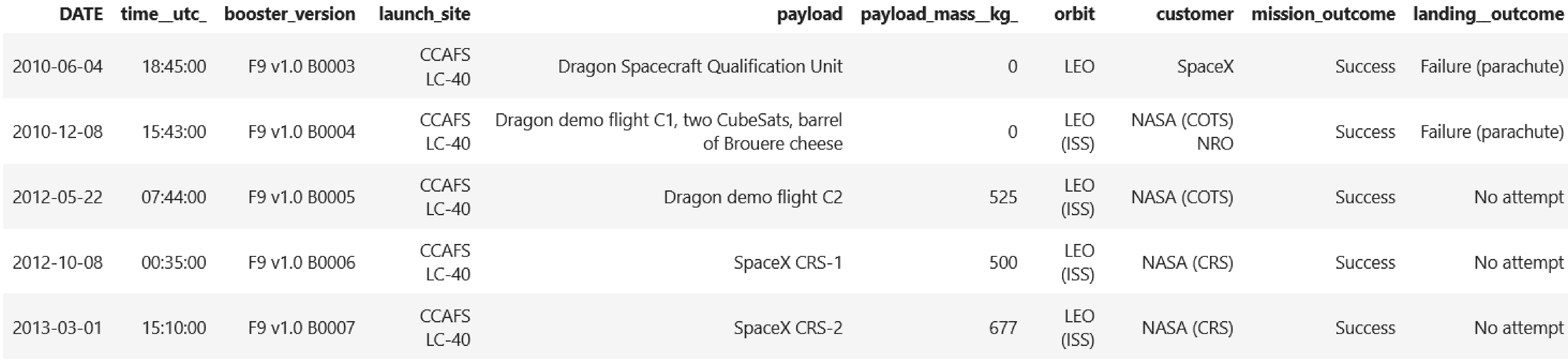
- CCAFS LC-40 and CCAFS SLC-40 indicates the same location, known with it a formal name as Cape Canaveral Air Force Station Space Launch Complex 40
- The most commonly used launch site is Cape Canaveral Space Launch Complex 40, that has 60 launches in total for Falcon 9 heavy launches
- KSC LC-39A come after CCAFS SLC-40 in the number of launches, followed by VAFB SLC-4E



launch_site	2
CCAFS LC-40	26
CCAFS SLC-40	34
KSC LC-39A	25
VAFB SLC-4E	16

Launch Site Names Begin with 'CCA'

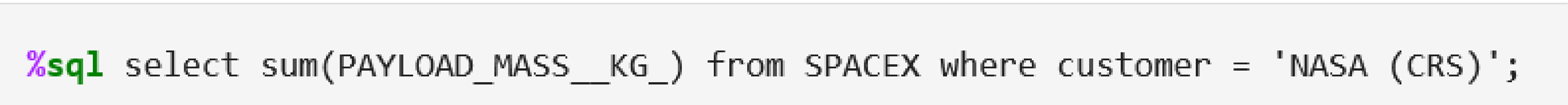
- Here we can see 5 samples of Cape Canaveral Air Force Station launches



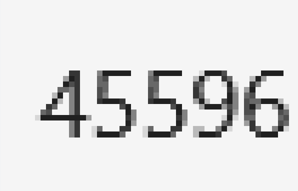
DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight 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 is calculated as below:



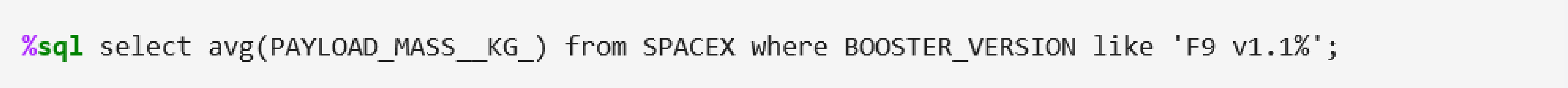
```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEX where customer = 'NASA (CRS)';
```



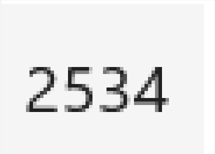
```
45596
```

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is calculated as below:



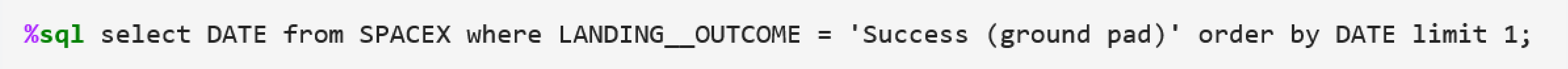
```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEX where BOOSTER_VERSION like 'F9 v1.1%';
```



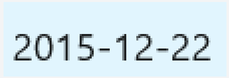
```
2534
```

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad is found as below:



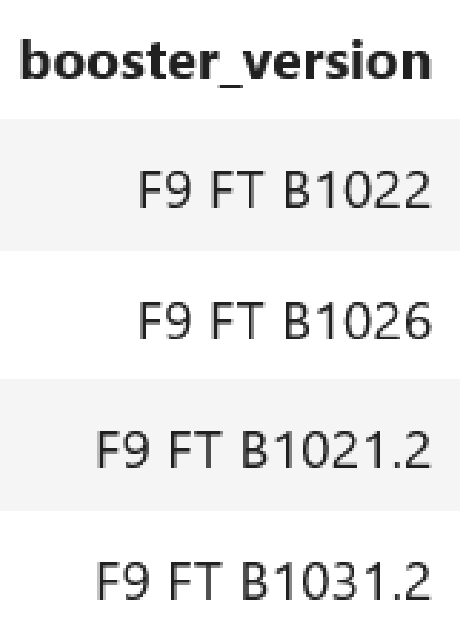
```
%sql select DATE from SPACEX where LANDING__OUTCOME = 'Success (ground pad)' order by DATE limit 1;
```



```
2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

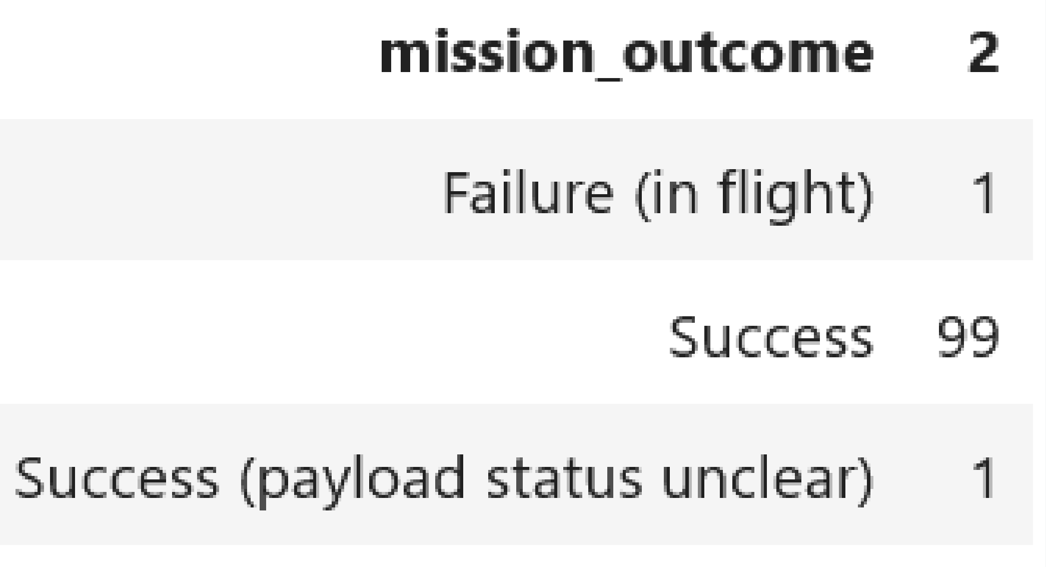
- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000:



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

Total Number of Successful and Failure Mission Outcomes

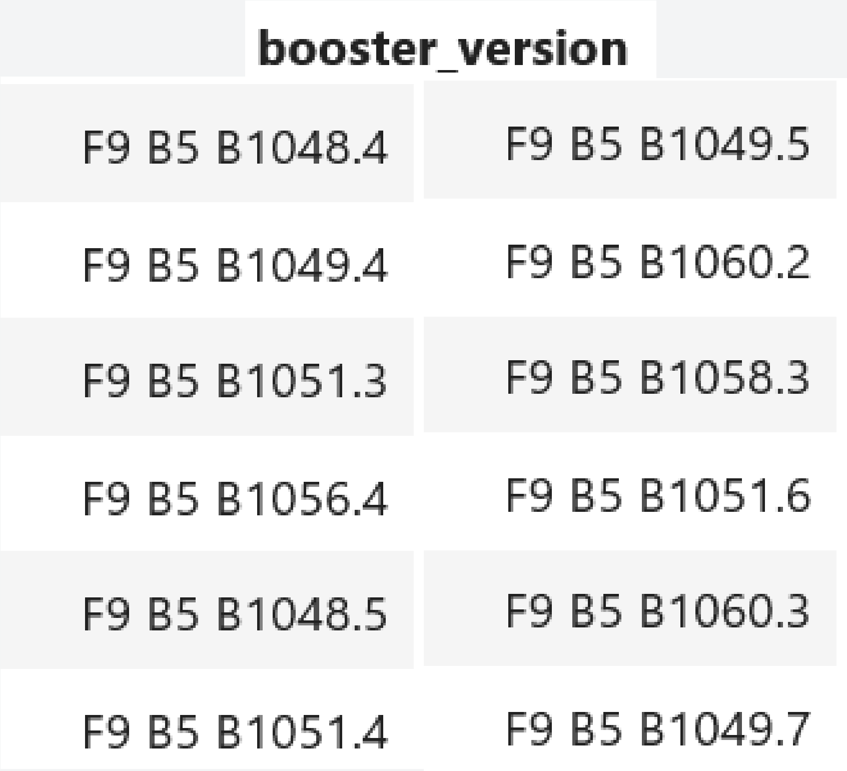
- The total number of successful and failure mission outcomes is calculated:



mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

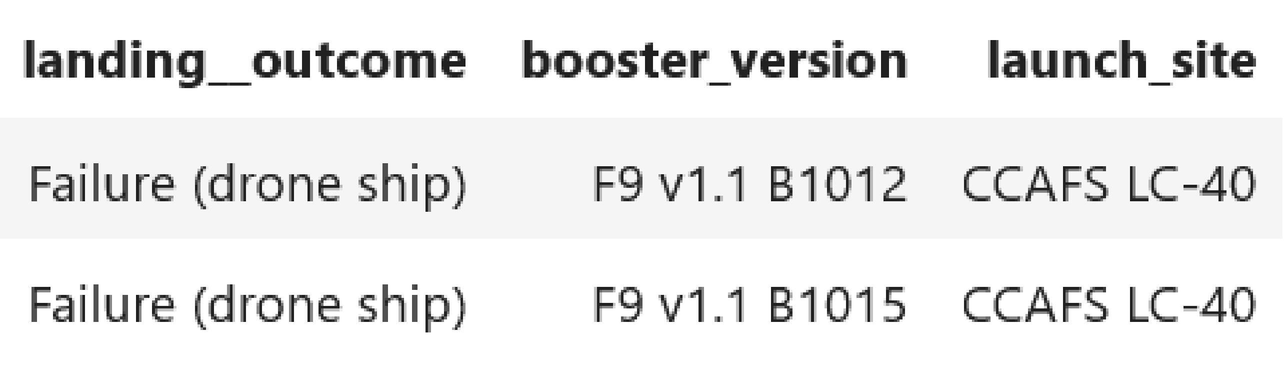
- The names of the booster which have carried the maximum payload mass:



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

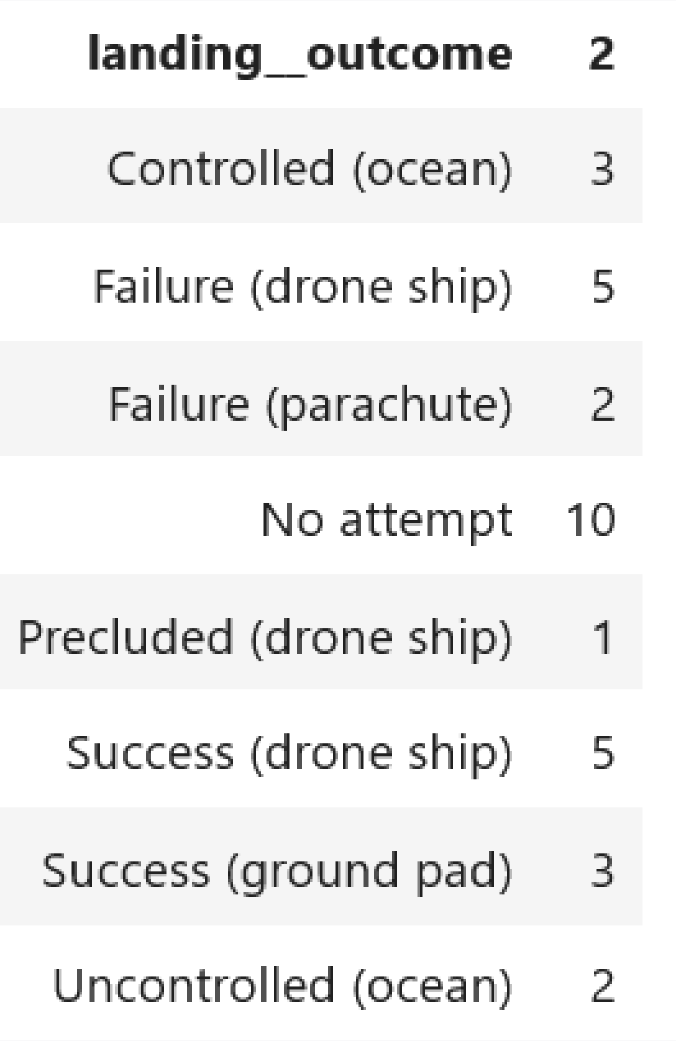
- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015:



landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- The count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20 is ranked in descending order:



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

Section 3

Launch Sites Proximities Analysis

Location of All Launch Sites on the Map

- VAFB SLC-4E, CCASF (S)LC-40, and KSC LC-39A are displayed on the map
- While KSC LC-39A and CCASF (S)LC-40 are in close proximity to each other, VAFB SLC-4E is located on the other side of the country

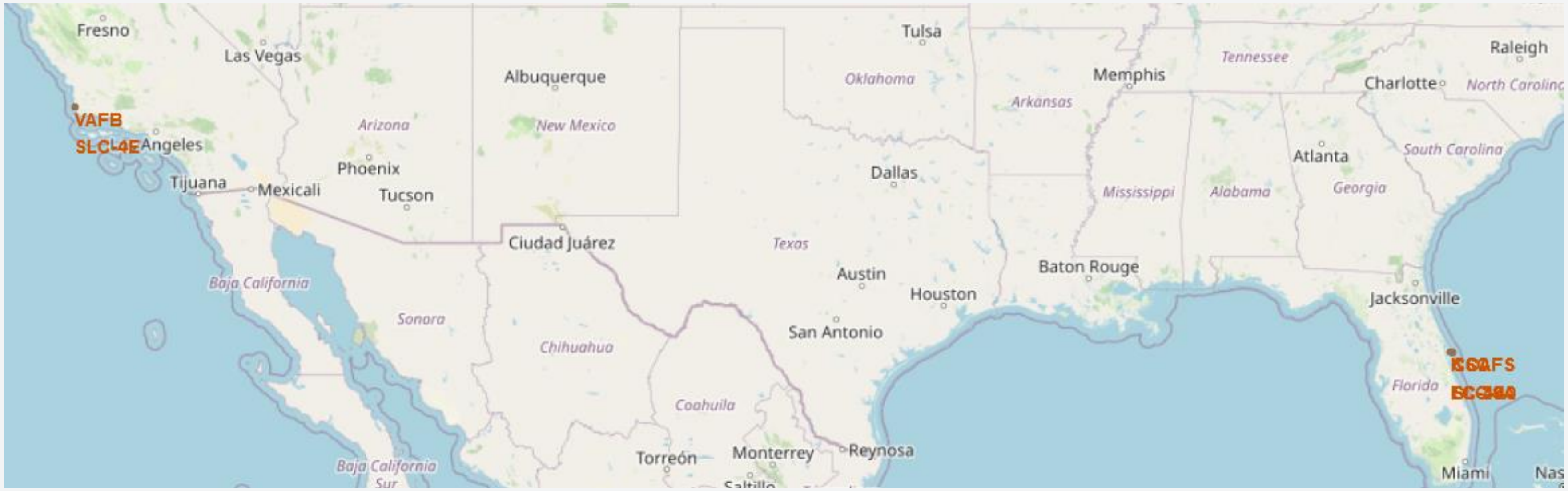


Figure 16. Map of launch sites

The Number of Launches on Each Launch Site

- The most of the launches are made in the East side of the USA, namely 46.
- CCAFS SLC-40 has the hightest number of launches, which is 33
- KSC LC-39A and VAFB SLC-4E have 13 and 10 launches respectively

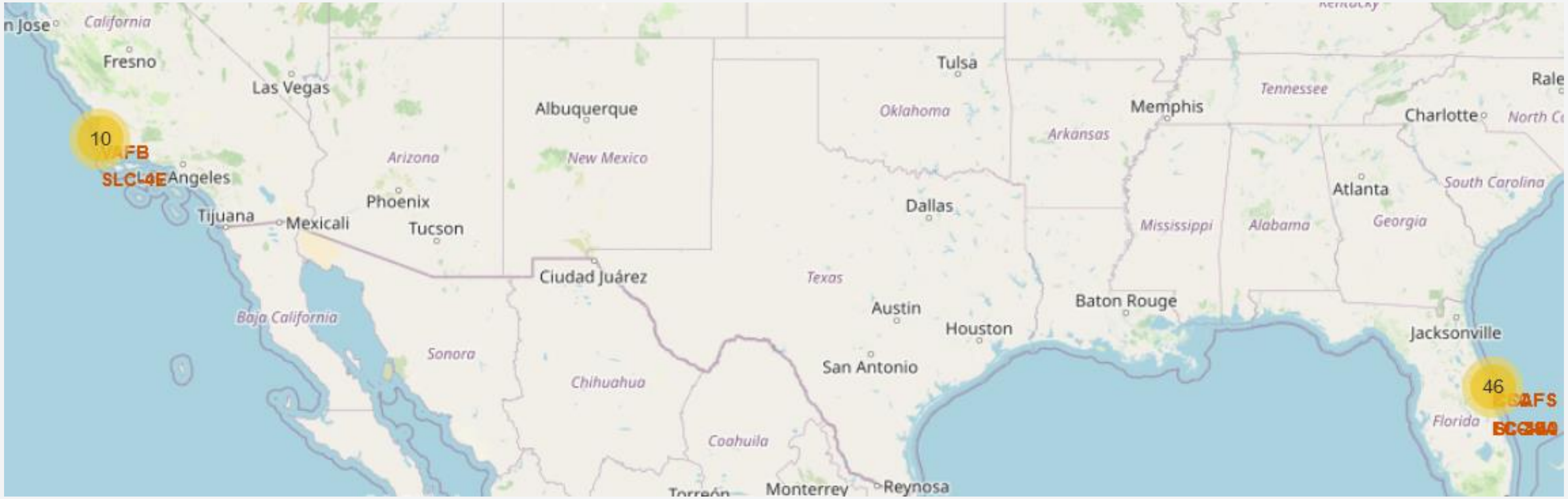


Figure 17. Map of clustered launches in different launch sites

Distances between Launch Sites and Certain Locations

- All of launch sites are in nearly 1KM proximity to coastline and railways
- All of launch sites except VAFB SLC-4E are in nearly 1KM proximity to highways
- All them keep at least 15KM distance from cities

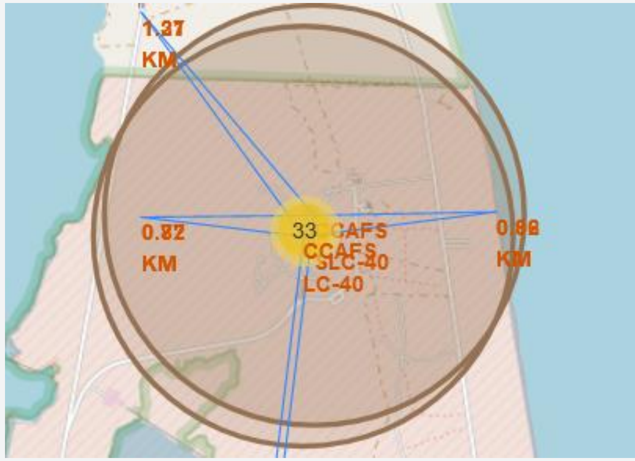


Figure 9. Distance from CCAFS SLC-40 to certain locations

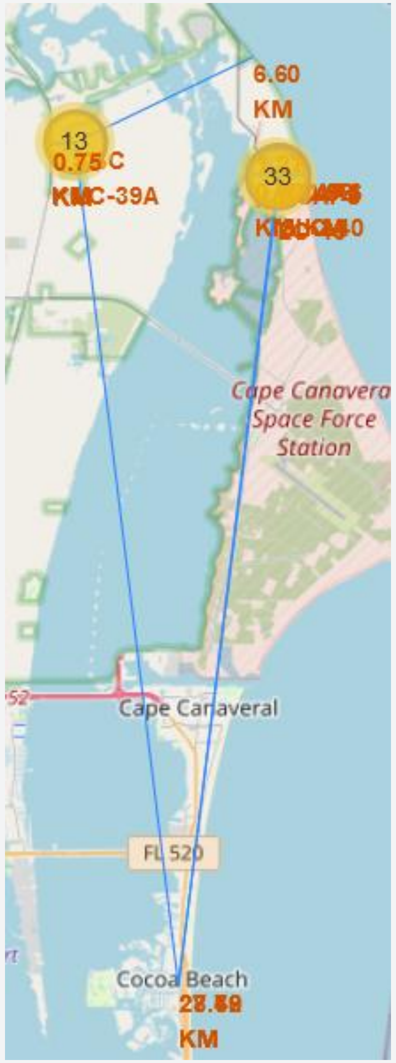
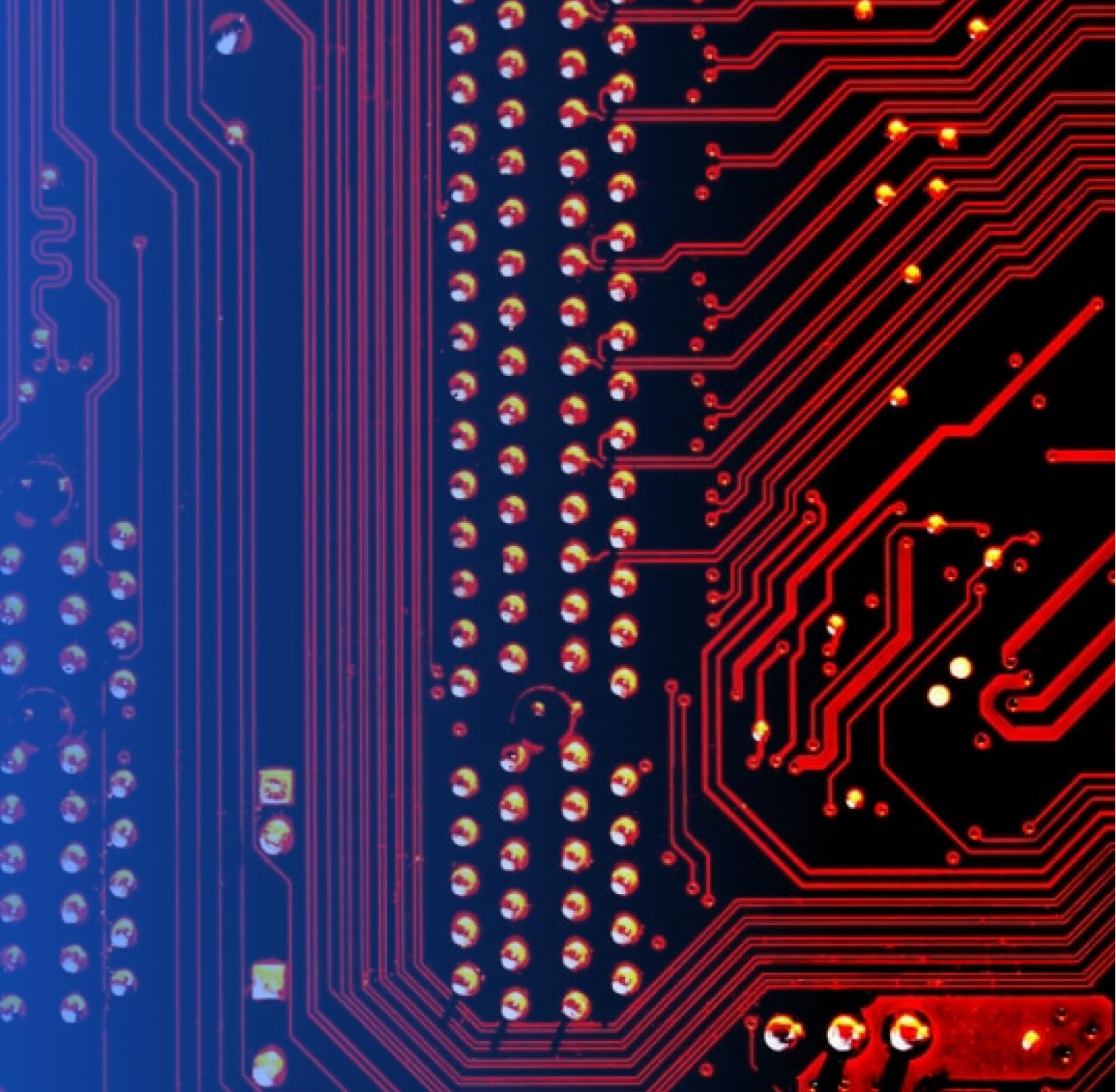


Figure 18. Distance from launch sites to certain locations

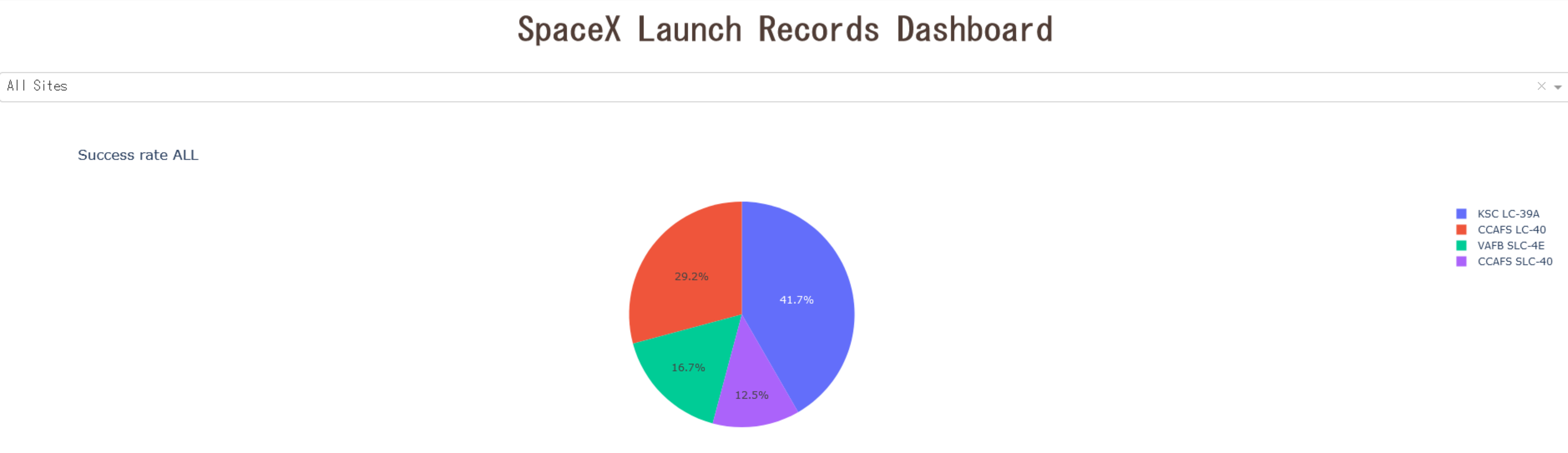
Section 4

Build a Dashboard with Plotly Dash



Success Ratio for All Launch Sites

- KSC LC-39A has the highest success ratio among others, followed by CCAFS SLC-40



SpaceX Launch Records Dashboard

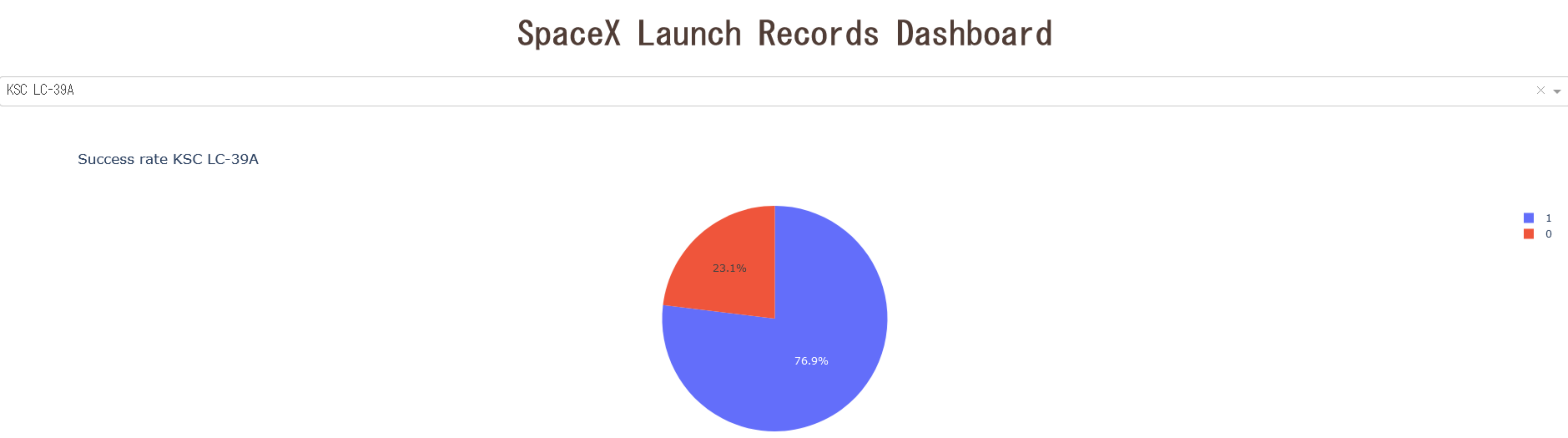
All Sites

Success rate ALL

Figure 19. Success ratio for all launch sites

The Launch Site with Highest Launch Success Ratio

- KSC LC-39A has 76.9% success ratio



SpaceX Launch Records Dashboard

KSC LC-39A

Success rate KSC LC-39A

Figure 20. The launch site with the highest launch success ratio

Success Count on Payload Mass for All Sites

- Payload mass vs. launch outcome (Class) is visualized on a scatter plot, with booster version category and selectable payload range option

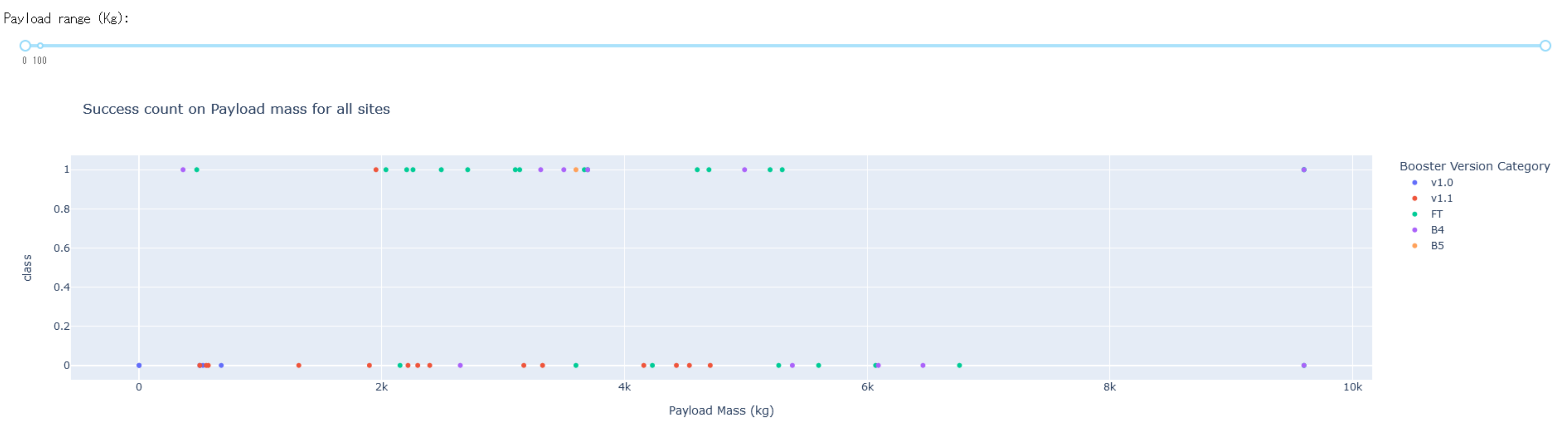


Figure 21. Success count on payload mass for all sites

Success Count on Payload Mass for All Sites

- Within the most commonly used payload range (2000-7000), it seems that successful launches are distributed equally among failing launches

- Within this range FT booster version is the mostly used one

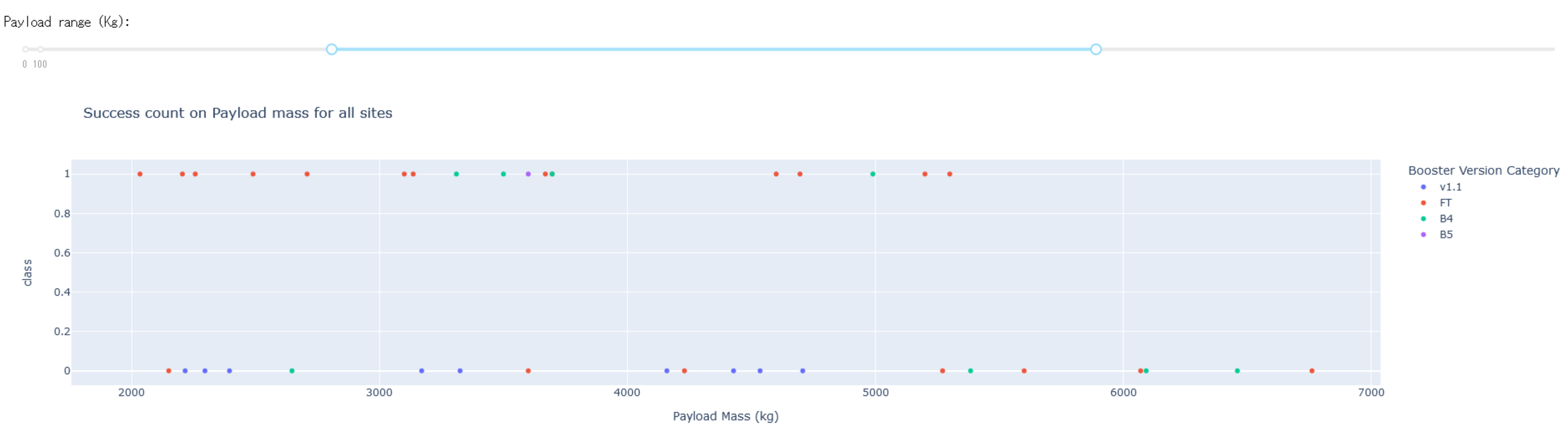


Figure 22. Success count on payload mass for all sites within a specified payload range

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The best hyperparameter for each model is found by using GridSearchCV

- The accuracies on the training data is the same for all models, namely 83%

- Decision Tree model is the best-performing one on the test data, reaching 94% accuracy

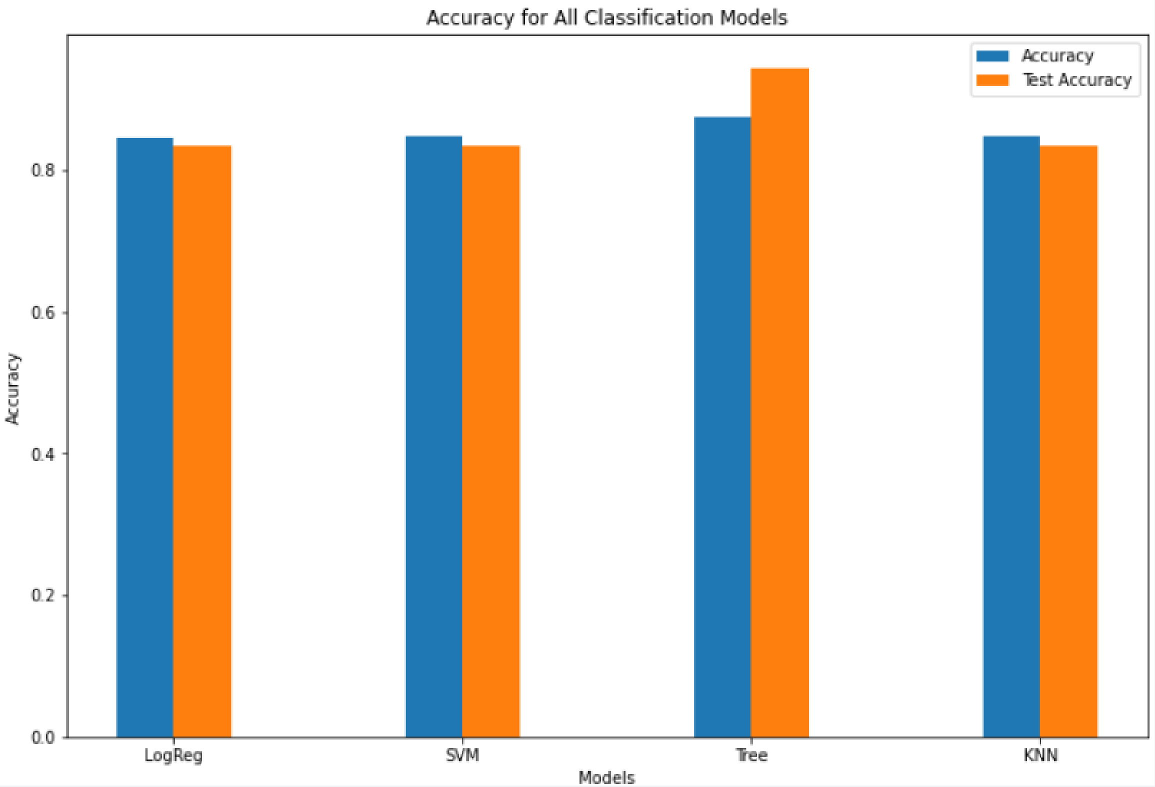


Figure 10. Accuracy of all models

Confusion Matrix

- On the test data which contains 18 rows, the built model predicted all of first stage landing results correctly except 1

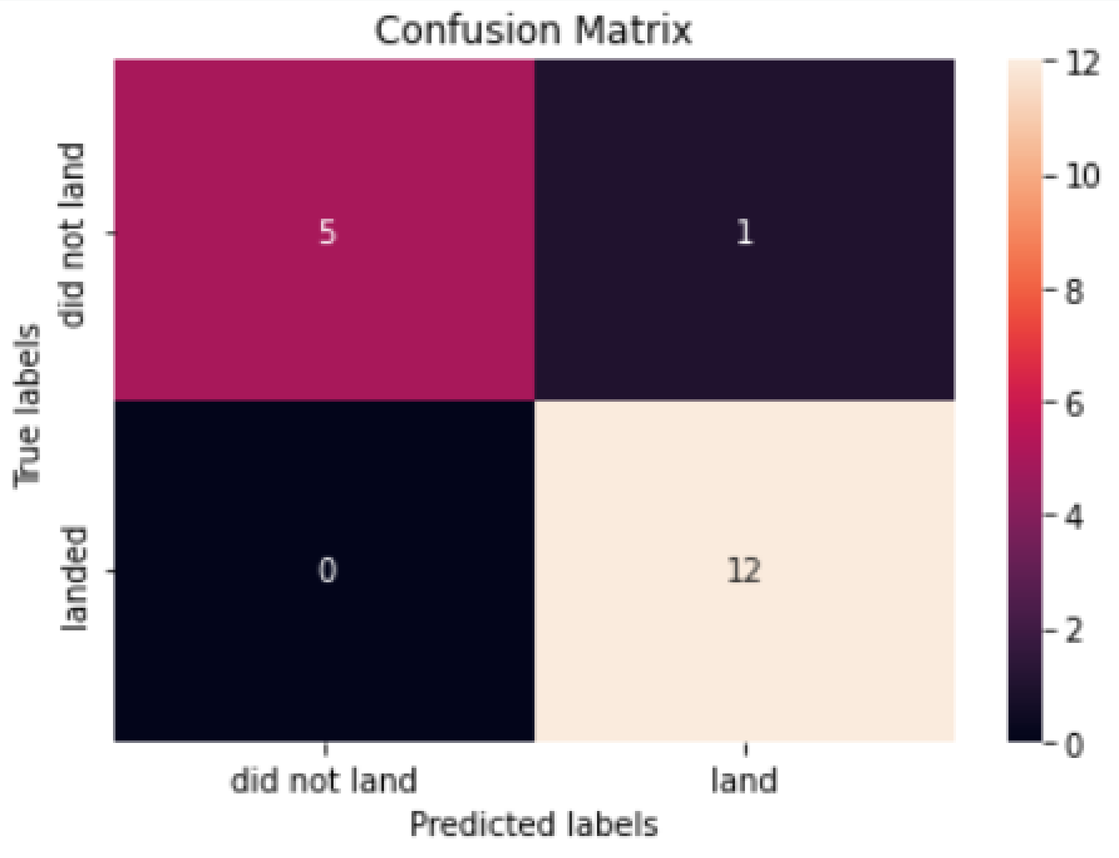


Figure 23. Confusion matrix of decision tree model

Conclusions

- The data is collected from different sourced, merged, and filtered for Falcon 9
- The data is visualized and queried by SQL to make an exploratory data analysis
- Launch sites are located on the map to inspect launch site characteristics
- An interactive dashboard is built to have an understanding on the relationship between launch site, payload, and success rate
- Different models are built to find the best performing-one, which is the classification tree model

Appendix

- Folium do not show maps on Github in files having ipynb extension. The maps are only available through this report presentation file
- Whereas overall success rate increases as payload mass increase, it have a falling period between 3000 and 6000 payload mass in kg. Charts presented below.

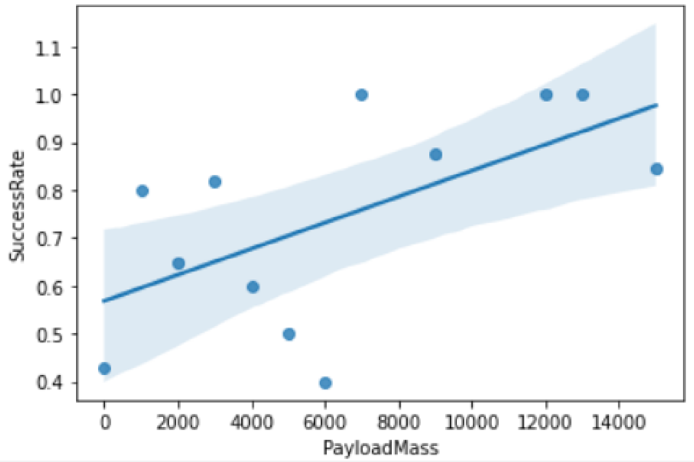


Figure 24. Success Rate vs. Payload Mass regression plot

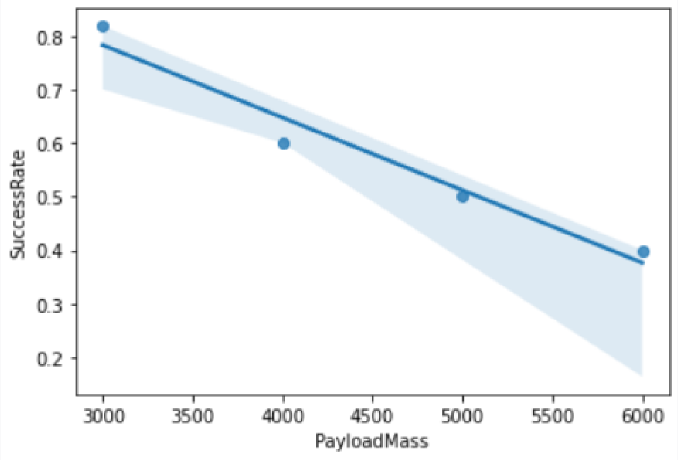


Figure 24. Success Rate vs. Payload Mass regression plot, payload between 3000 and 6000

Resources

[1] https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

[2] https://api.spacexdata.com/v4/launches

Thank you!