IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

YiHao Chen
2025–01–10

Outline

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

Executive Summary

- Summary of methodologies
  - Collecting and wrangling data using SpaceX api and SpaceX wikipedia
  - Exploring Data Analytics with SQL and Visualization
  - Interactive Maps with Folium and Dashboards with Plotly Dash
  - Predict the dataset with different models to arrive at the best model
- Summary of all results
  - Orbit type and launch sites are very important and have a significant relationship with the Flights Number and Payloads
  - Different launch sites have commonalities
  - A better predictive model can be derived using GridSearchCV

Introduction

- Project background and context
  - SpaceX has attracted many companies for its very low rocket launch prices, which are $100 million less than other providers.
  - It can achieve such a low price because it can recycle its first stage rockets, which greatly reduces launch costs.
  - As a data scientist at SpaceY, it is necessary to find patterns of recoverable first–stage rockets and design a model to predict the success rate of recovering first–stage rockets.

- Problems you want to find answers
  - How to find the best place to launch
  - Hoping to find a model that best predicts the success of a rocket retrieval

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data from SpaceX API and SpaceX Wikipedia page
- Perform data wrangling
  - Filtering the required data, checking for missing data, and binary categorization of predicted data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Standardize the organized data into training and test sets, set parameters for GridSearchCV, get the best parameters, and use the test set to calculate the accuracy and find the best model

Data Collection

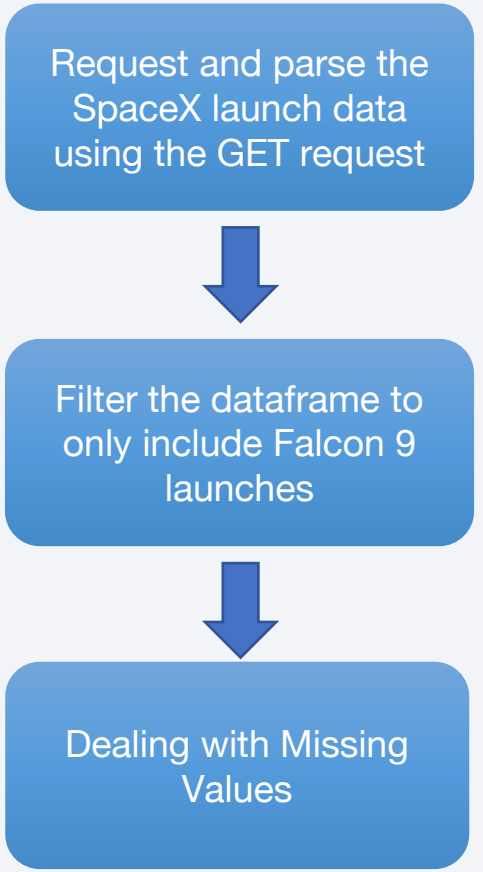
- First, we need to retrieve the SpaceX API information using the requests.get function. And we decode the content into json format and turn it into a dataframe using json_normalize.
- Then we get the information we want from the dataframe and recreate a dataframe with only the information we need.
- When collecting information on the List of Falcon 9 and Falcon Heavy launches, we used beautifulsoup to do web scraping, and then extracted the data we needed to make a dataframe.

Data Collection – SpaceX API

- For the SpaceX API use requests.get function to get url, and decode content as a Json, finally turn json into a dataframe.

- GitHub Link

- (https://github.com/SeanLearningAccount/IBM-Applied-Data-Science-Capstone/blob/main/1.%20spacex-data-collection-api.ipynb)

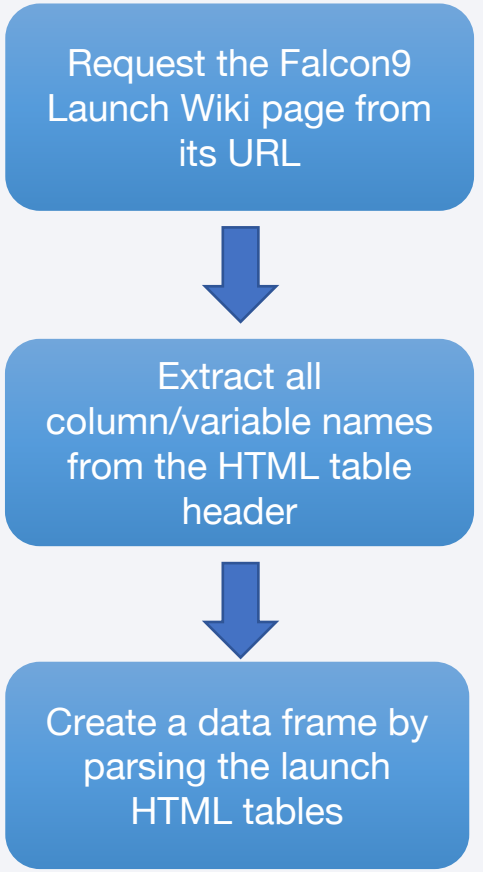


Data Collection - Scraping

- Make a request to the falcon9 launch weiki page, then extract the table using BeautifulSoup, collect the data by parsing the HTML table, and form a dataframe.

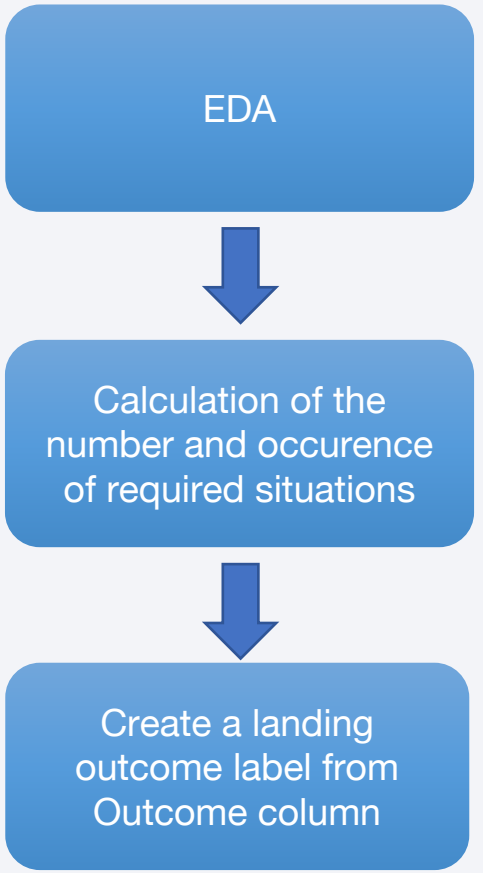
- GitHub Link

- (https://github.com/SeanLearningAccount/IBM-Applied-Data-Science-Capstone/blob/main/2.%20webscraping.ipynb)



Data Wrangling

- Filter booster version with only falcon9, check for missing data and data type.Missing values for ‘PayloadMass’ are replaced with average values.
- Check and count launch sites and orbit types.
- Binary the launch outcomes for subsequent analysis.



- GitHub Link
- (https://github.com/SeanLearningAccount/IBM-Applied-Data-Science-Capstone/blob/main/3.%20spacex-Data%20wrangling.ipynb)

EDA with Data Visualization

- To visualize the relationship, used a scatter plot to visualize Flight Number and Launch Site, Payload Mass and Launch Site, FlightNumber and Orbit type, Payload Mass and Orbit type.
- To check if there are any relationship between success rate and orbit type, used a bar chart.
- In order to get a trend of successful launches per year, used a line chart.

- GitHub Link
- (https://github.com/SeanLearningAccount/IBM-Applied-Data-Science-Capstone/blob/main/5.%20edadataviz.ipynb)

EDA with SQL

- Display the names of the unique launch sites in the space mission

```
%sql select distinct Launch_Site from SPACEXTBL
```

- Display 5 records where launch sites begin with the string 'CCA'

```
%sql select * from SPACEXTBL where Launch_Site LIKE 'CCA%' limit 5
```

- Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql select SUM(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer LIKE 'NASA%(CRS)%'
```

- Display average payload mass carried by booster version F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version LIKE 'F9%v1.1'
```

- GitHub Link
- (https://github.com/SeanLearningAccount/IBM-Applied-Data-Science-Capstone/blob/main/4.%20eda-sql-coursera_sqllite.ipynb)

Build an Interactive Map with Folium

- Markers and circles were used to mark the launch sites on the folium map.
- Markers and MarkerCluster were used to mark the success/failed launches for each site on the folium map.
- Markers, mouse position and line were used to calculate the distances between a launch site to its proximities on the folium map.
- GitHub Link
- (https://github.com/SeanLearningAccount/IBM-Applied-Data-Science-Capstone/blob/main/6.%20launch_site_location.ipynb)

Build a Dashboard with Plotly Dash

- Added a dropdown list to enable Launch Site selection to a dashboard.
- Added a pie chart to show the total successful launches count for all sites.
- Added a a slider to select payload range.
- Added a scatter chart to show the correlation between payload and launch success.

- GitHub Link
- (https://github.com/SeanLearningAccount/IBM-Applied-Data-Science-Capstone/blob/main/7.%20Build%20an%20Interactive%20Dashboard%20with%20Ploty%20Dash)

Predictive Analysis (Classification)

- First, we allocate the data to X and Y, standardize the data in X, split the data into training and test data.
- Then we need to test Logistic Regression, SVM, Decision Tree and KNN.
- When testing the models, we need to set the parameters first, run GridSearchCV object to find the most appropriate parameters for each model to be applied to the test set.
- Finally, the model with the highest accuracy will be the output model.

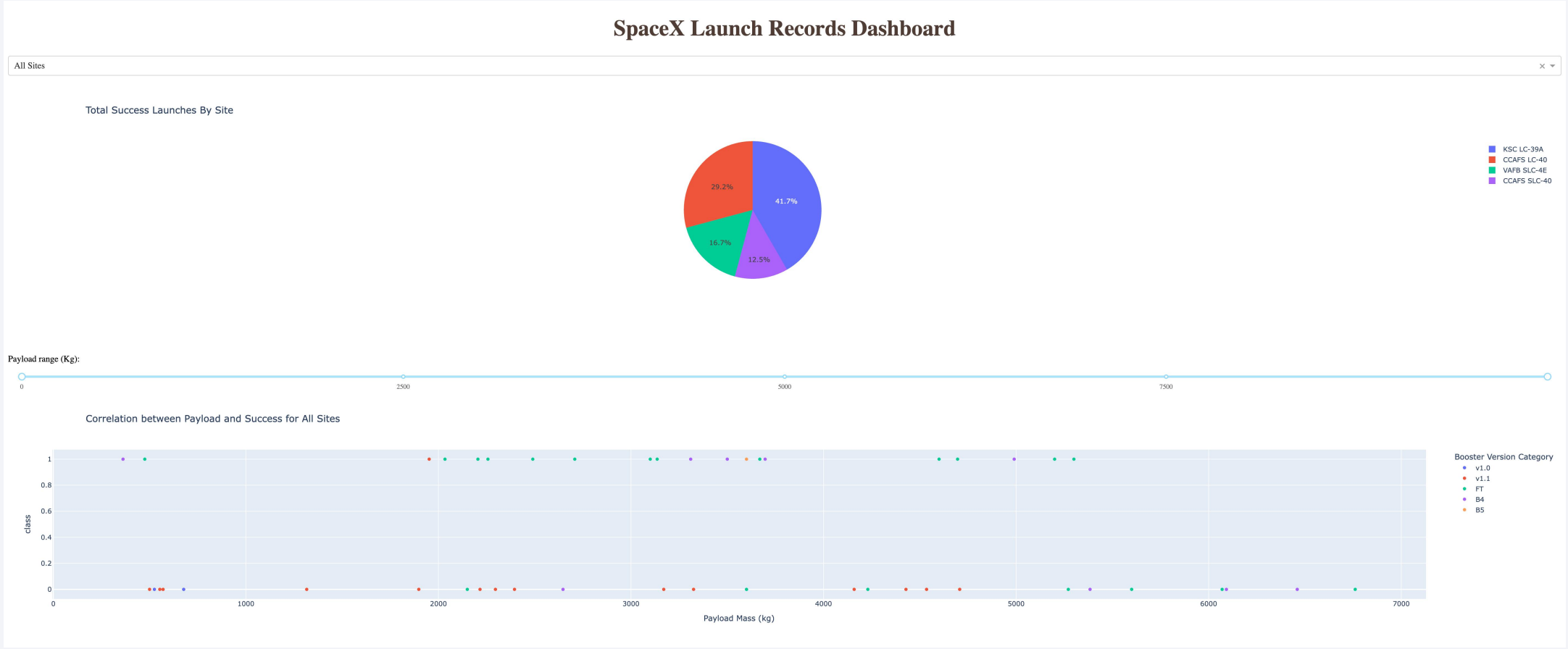
- GitHub Link
- (https://github.com/SeanLearningAccount/IBM-Applied-Data-Science-Capstone/blob/main/8.%20SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb)

Results

- After exploratory data analysis, there is a clear relationship between success rate in terms of year and orbit.
- Orbit types have something to do with the flights number and the payload mass.
- Launch Site have something to do with the flights number and the payload mass.

Results

- Interactive analytics demo in screenshot

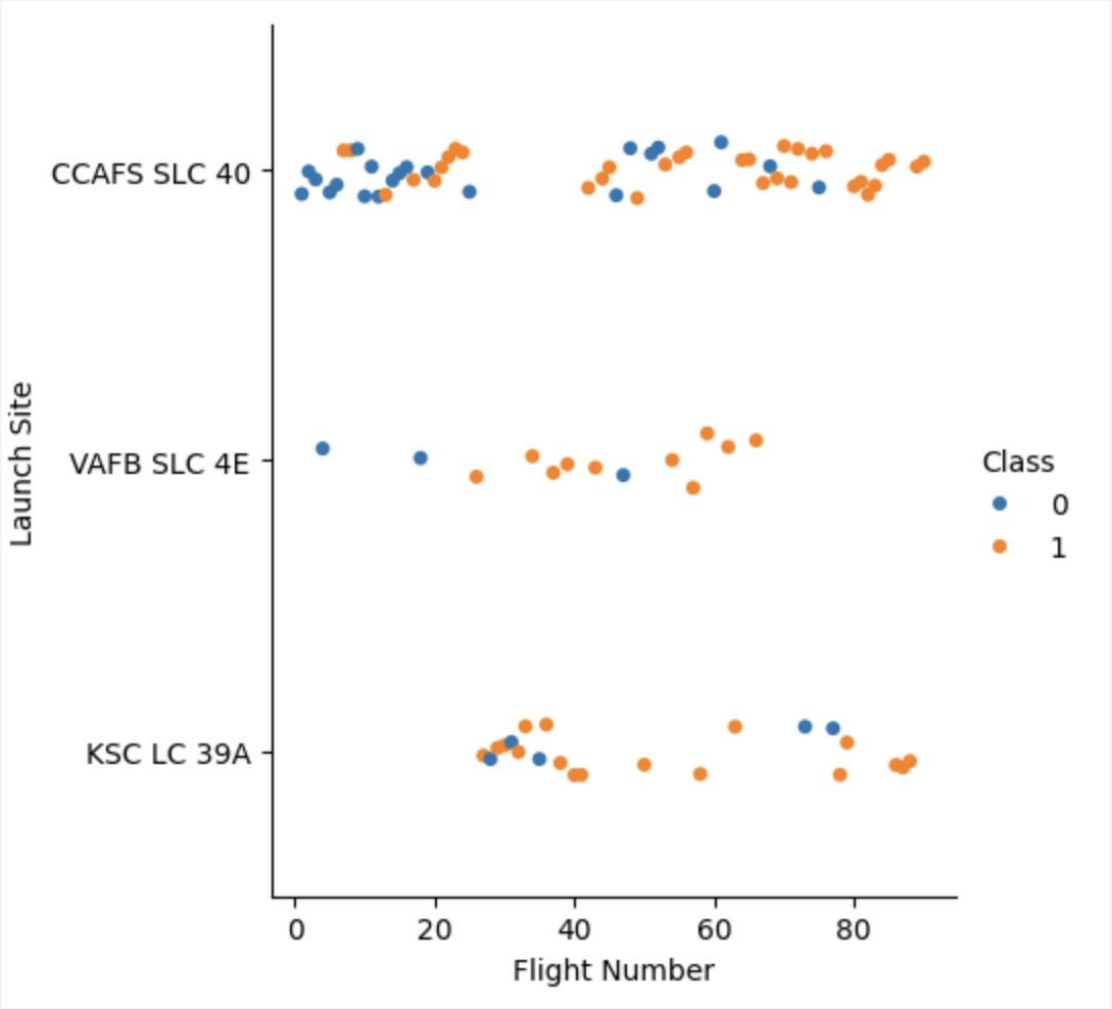


Section 2

Insights drawn from EDA

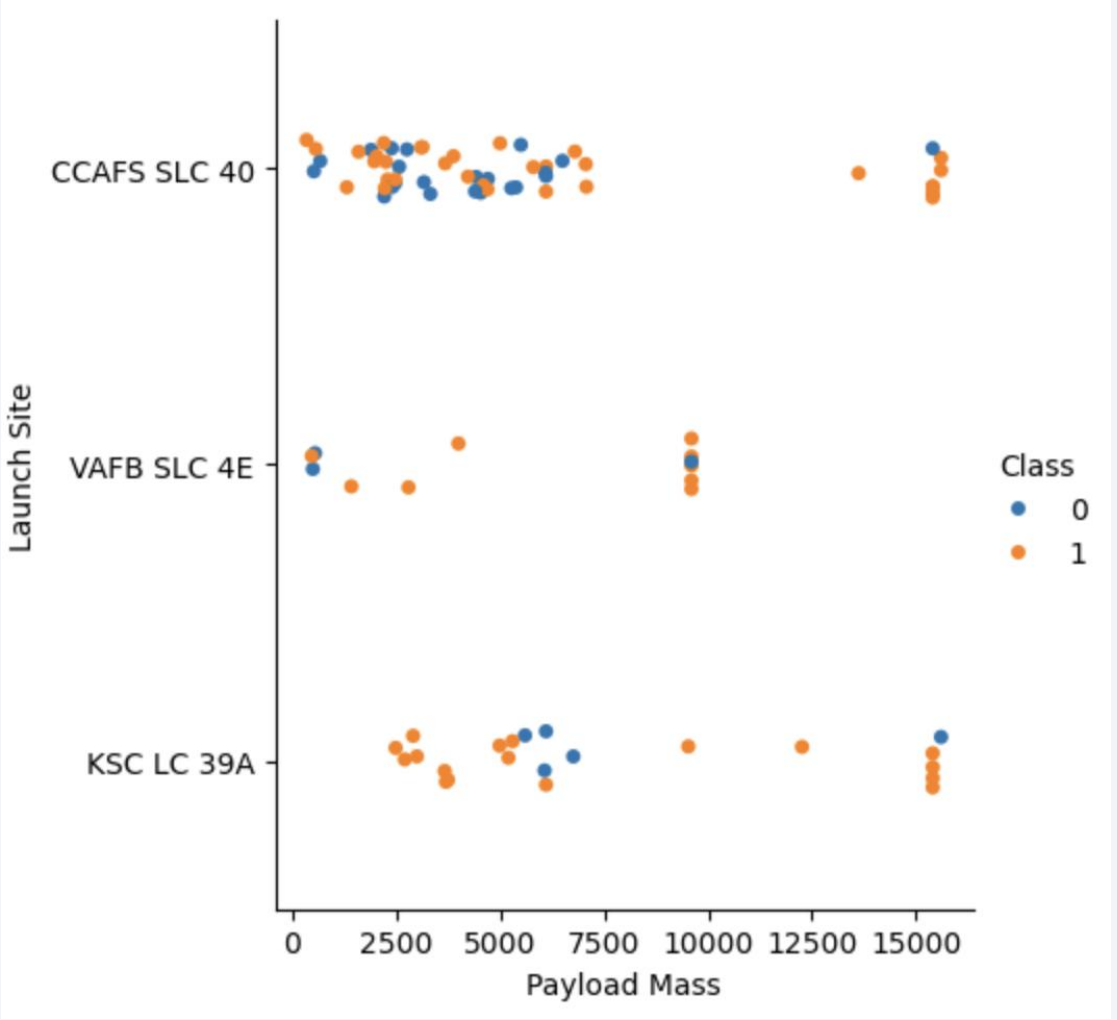
Flight Number vs. Launch Site

- The CCAFS SLC 40 launch site was the first to be used and has gone from a relatively high failure rate in the early days to a relatively high success rate today.
- VAFB SLC 4E was also used as a launch site in the early days, but has not been used since.
- Probably due to the high failure rate of the CCAFS SLC 40 in the early days, it is speculated that the KSC LC 39A started as a backup launch site for the CCAFS SLC 40, and that was the time when the KSC LC 39A launched the most.
- Now the launch site is primarily CCAFS SLC 40, with a few launches at KSC LC 39A.



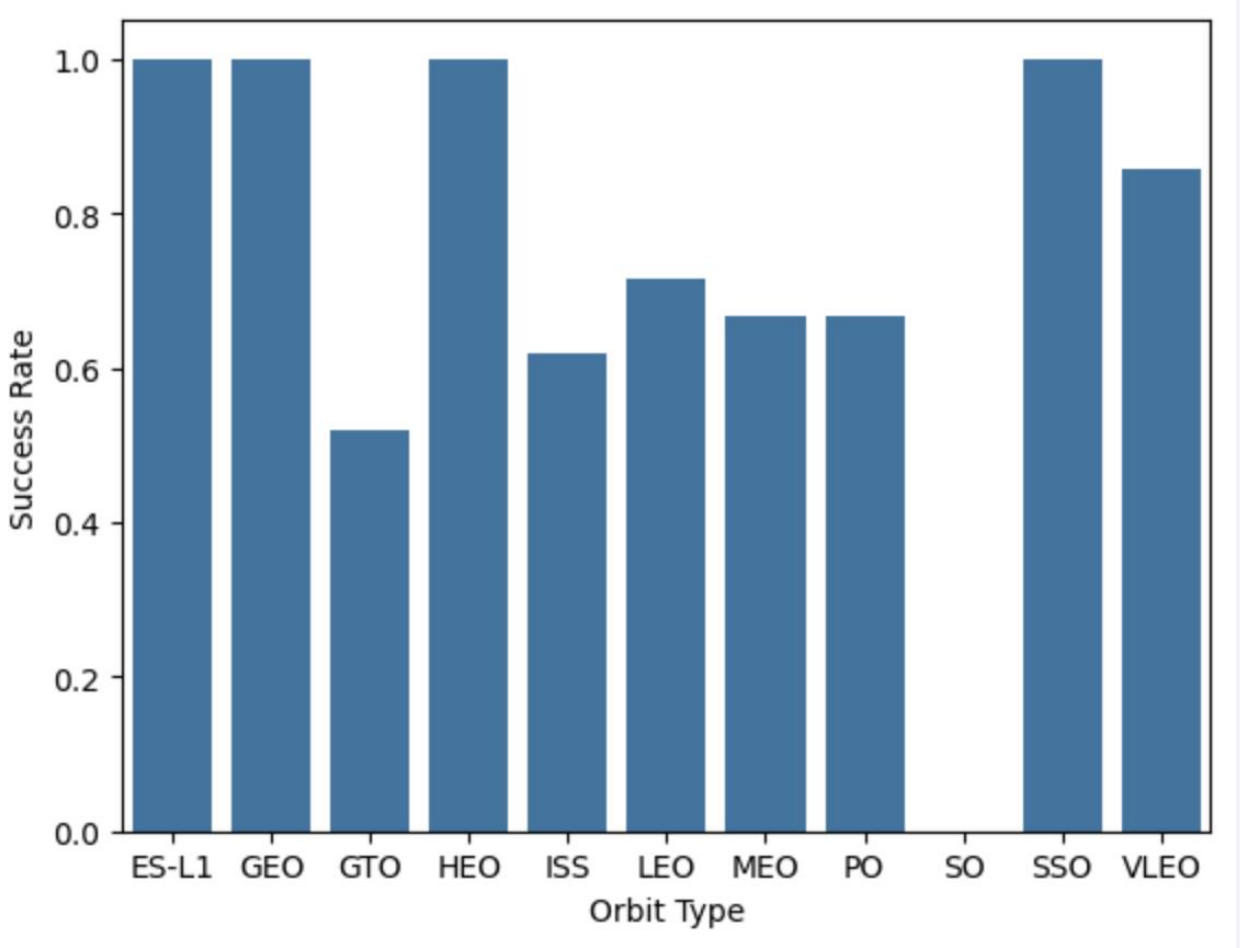
Payload vs. Launch Site

- Obviously, as the payload mass increases, the success rate increases significantly.
- CCAFS SLC 40 seems to have a high failure rate at low payload mass, but this could be a misconception because this launch site is the one used by SpaceX in the early days, so the failure rate is normal, not a problem with this launch site.
- Currently, SpaceX launches the most low and medium mass payloads, high mass payloads are currently launched less frequently, but with a high success rate.



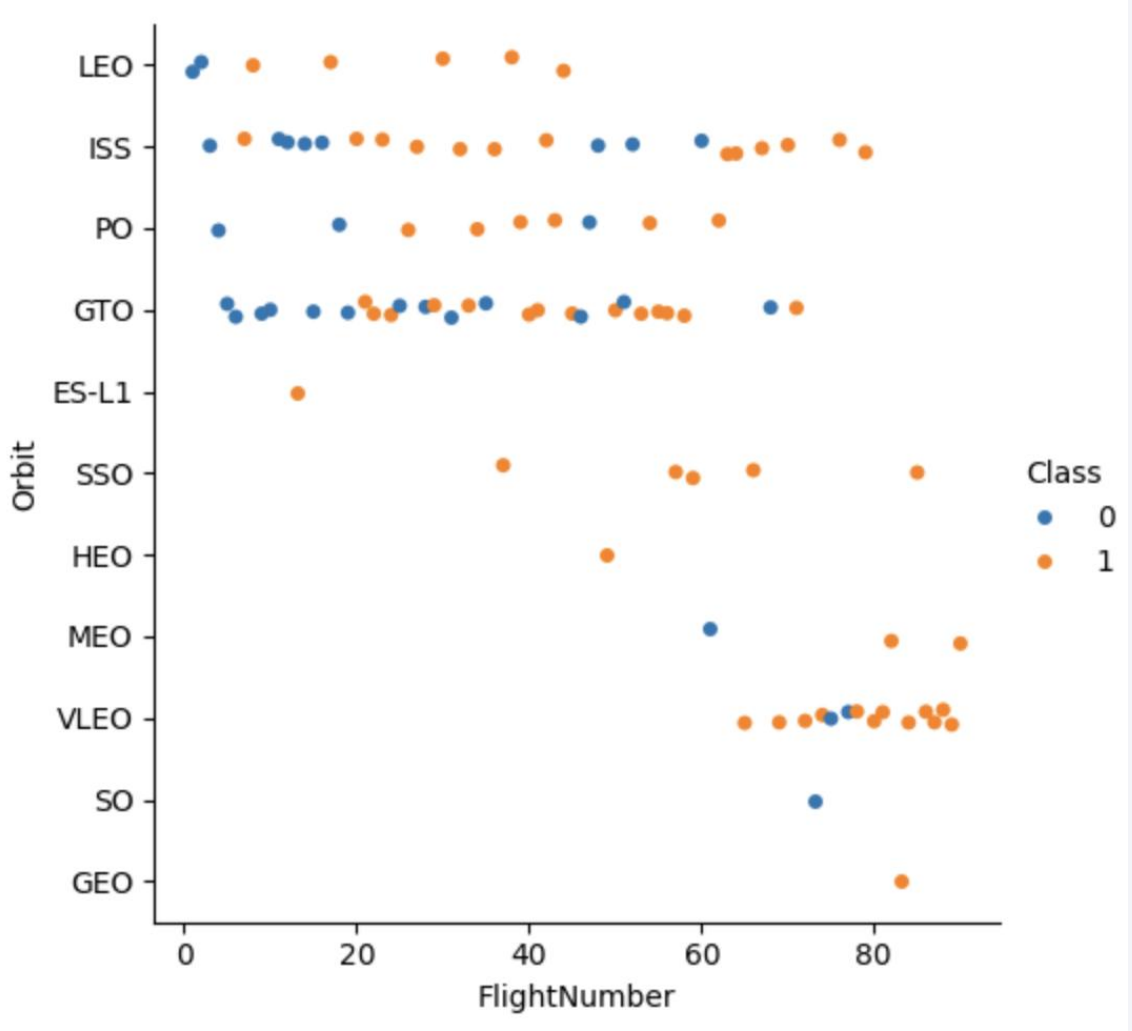
Success Rate vs. Orbit Type

- The highest success rates are for these orbit types:
  - ES-L1
  - GEO
  - HEO
  - SSO
- The missions in these orbits are more oriented towards specific scientific research, observation of the polar regions, or specific mission needs.



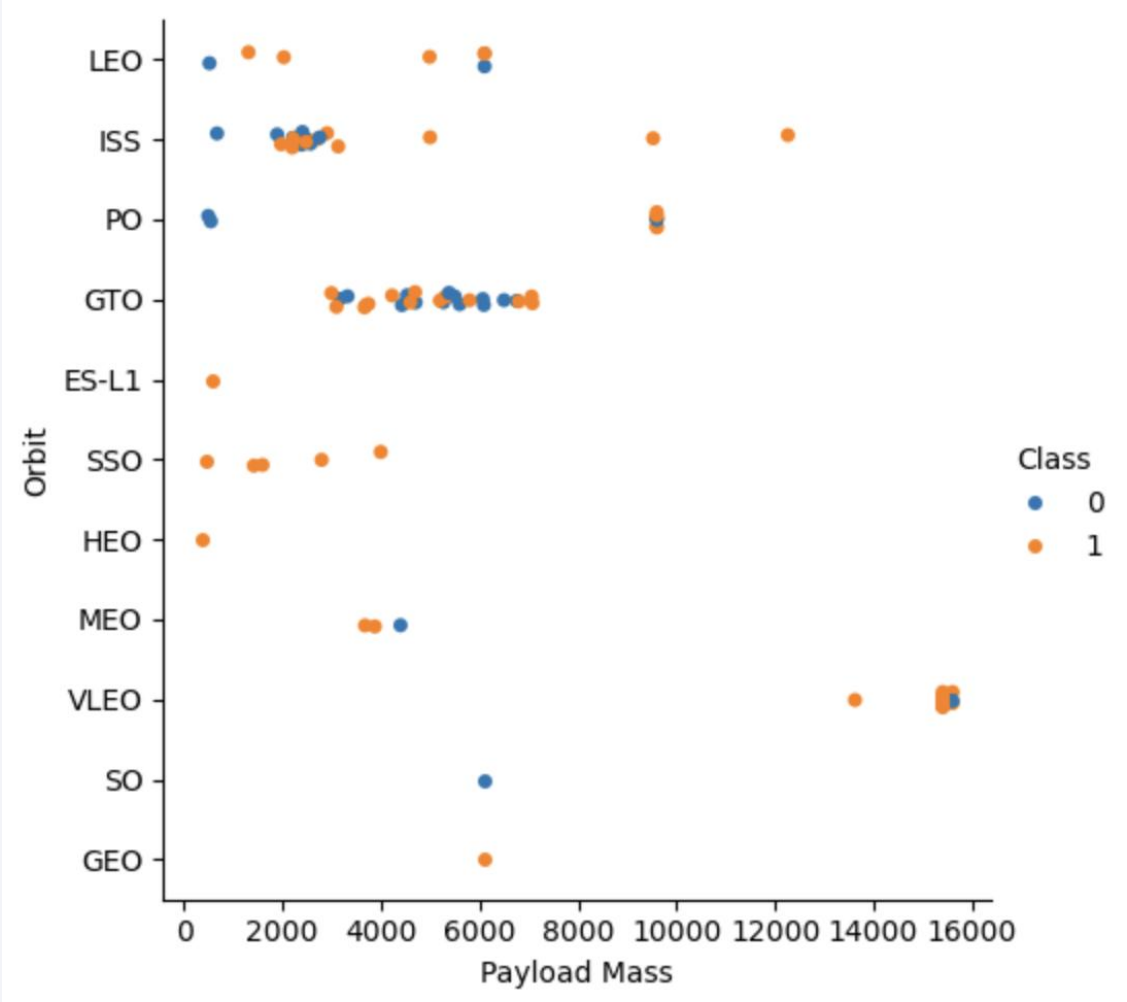
Flight Number vs. Orbit Type

- Later launches into VLEO orbits are more frequent and have high success rates.This orbit has very high commercial potential
- It was mentioned earlier that ES-L1, HEO, and GEO have a high success rate, but from this chart we can see that all of these orbits have only been launched once and lack the amount of data. Only SSO has been launched more times and has never failed.



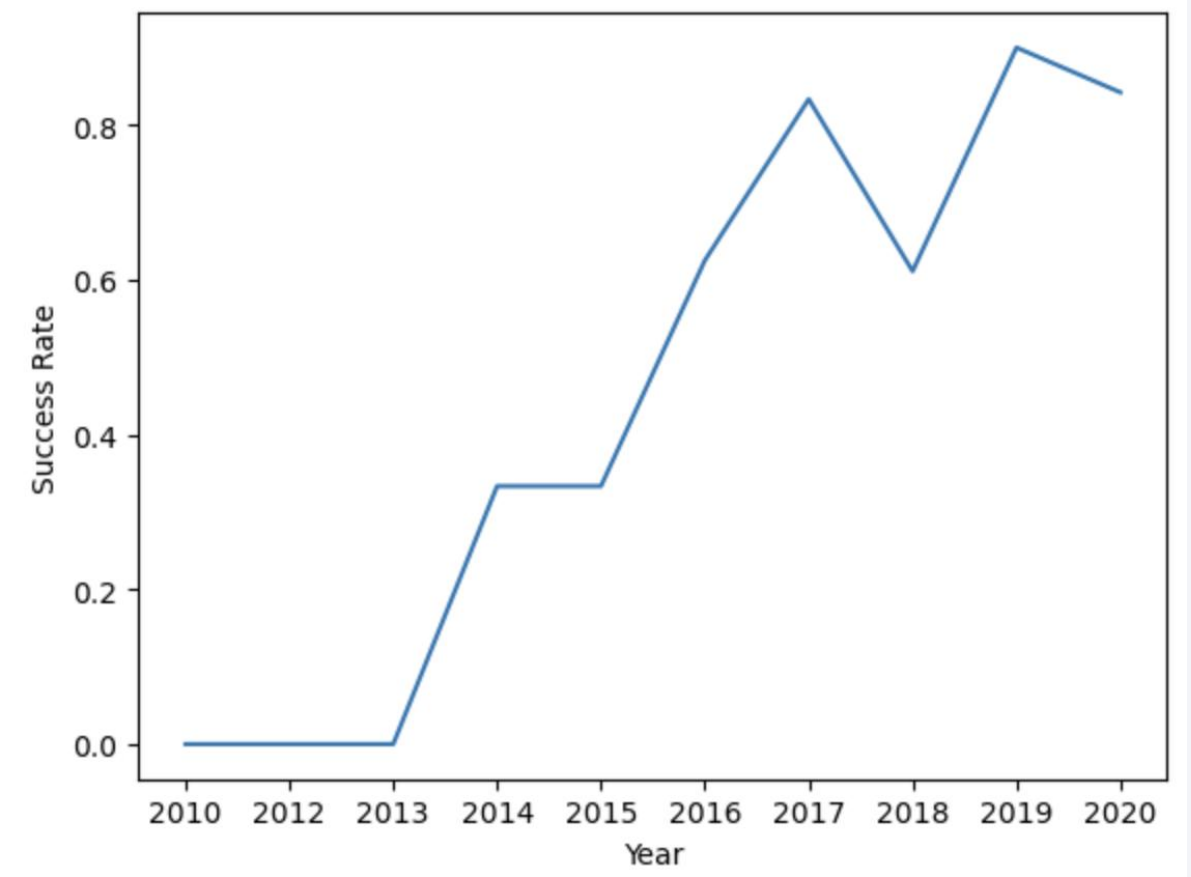
Payload vs. Orbit Type

- VLEO is a near-Earth orbit, but the payload mass is the largest.
- The payload mass range of the GTO orbit is more concentrated and the number of launches is greater, indicating that this orbit is more mature and widely used.
- The ISS has a very wide payload mass range and the success rate increases as the payload mass increases.



Launch Success Yearly Trend

- The launch success rate increased year by year, and finally the success rate was maintained at over 80%.
- At the beginning of 2010-2013, the launch success rate was 0%. 2014 was the first turning point when successful rockets were launched.
- The subsequent turning points were in 2016 and 2017, both of which doubled the launch success rate from 2014.
- Although the launch success rate declined in 2018, it peaked again in 2019.



All Launch Site Names

- There are four different launch sites in total.
- They are obtained by retrieving a unique value from the database.

Launch Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

- There are 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
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

- The total payload carried by boosters from NASA (CRS).
- Calculate the total payload mass after selecting NASA (CRS) from the database Customer.

Total payload mass carried by NASA (CRS)
48213 KG

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1.

- Calculate the average payload mass after selecting F9 v1.1 from the booster version of the database.

Average Payload Mass by F9 v1.1 (KG)
2928.4

First Successful Ground Landing Date

- This is the first successful landing outcome on ground pad.
- After selecting Success (ground pad) from the landing result, find the minimum value.

First Date
2015-12-22

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:

- Finding Success (drone ship) from the landing outcome and specifying payload mass greater than 4000 but less than 6000 yields this list.

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:
- Use subquery to calculate the number of successes and failures and then form a table.

Mission Outcome	Count
Success	100
Failure (in flight)	1

Boosters Carried Maximum Payload

- The booster which have carried the maximum payload mass.
- Use subquery to find the maximum payload mass and then find the matching booster version.

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

- This is the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015.

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

- First specify the Failure (drone ship) and 2015, then extract the desired information from the table.

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

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

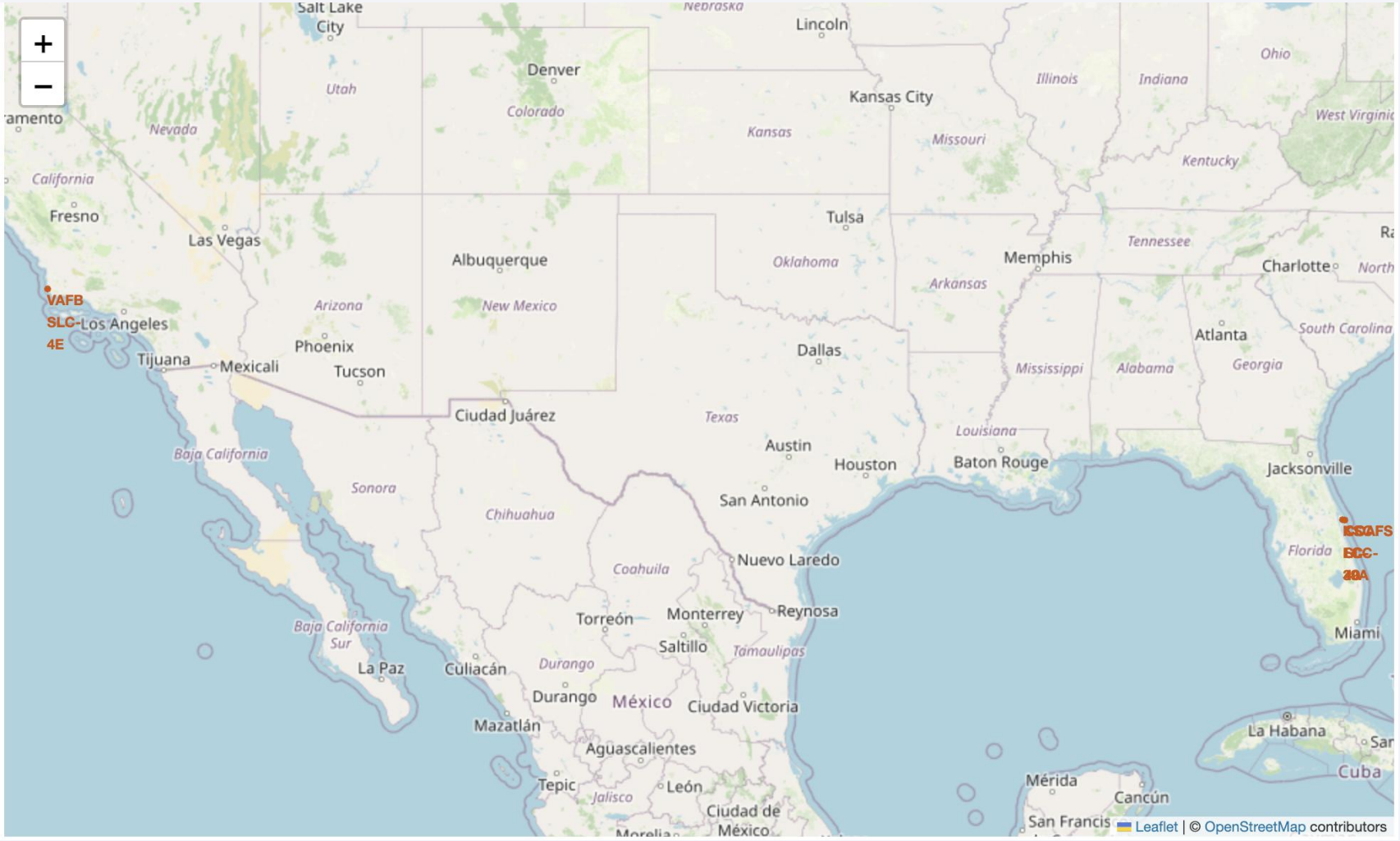
- The main thing to note here is the use of Rank Over for sorting.

Landing Outcome	Outcome Count	Rank
No attempt	10	1
Success (drone ship)	5	2
Failure (drone ship)	5	2
Success (ground pad)	3	4
Controlled (ocean)	3	4
Uncontrolled (ocean)	2	6
Precluded (drone ship)	1	7
Failure (parachute)	1	7


Section 3
Launch Sites
Proximities Analysis

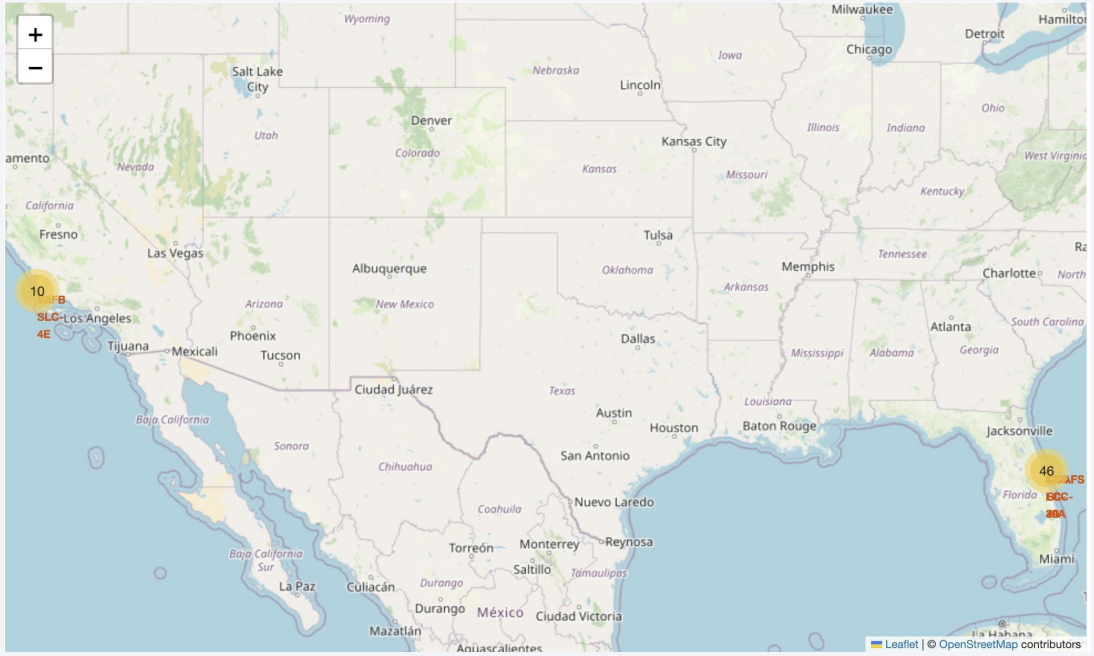
Discovery of the launch site

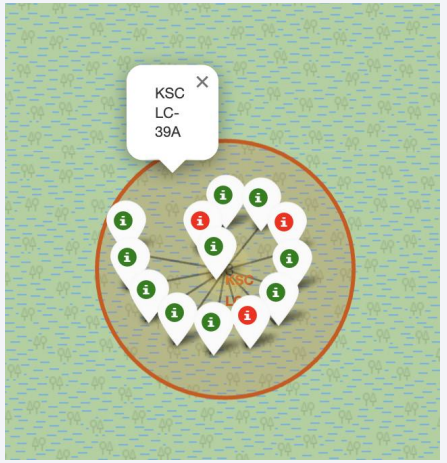
- The launch sites are all at relatively low latitudes.
- And they're all by the sea.



Launch Site and Launch Outcomes

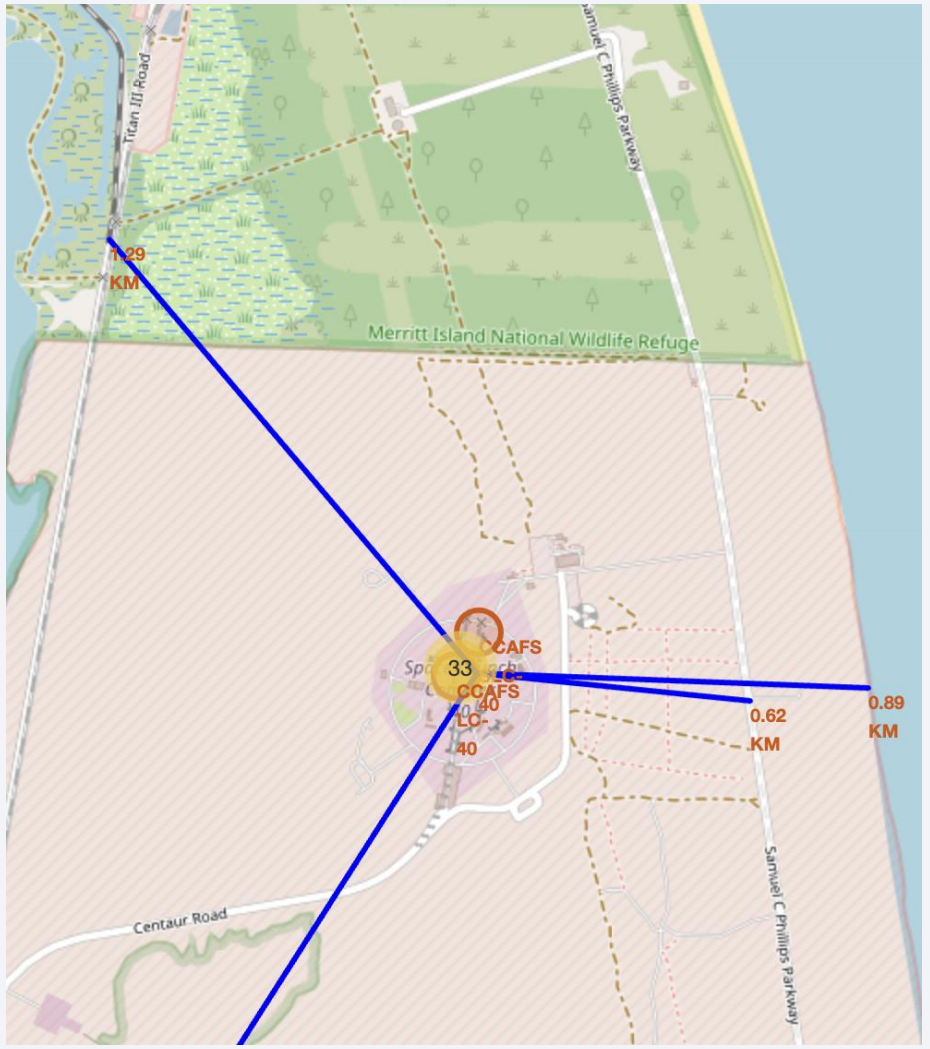
- There were more launches from launch sites at lower latitudes.
- The launch site with the highest launch success rate is KSC LC-39A.
- Green markers the number of successes.

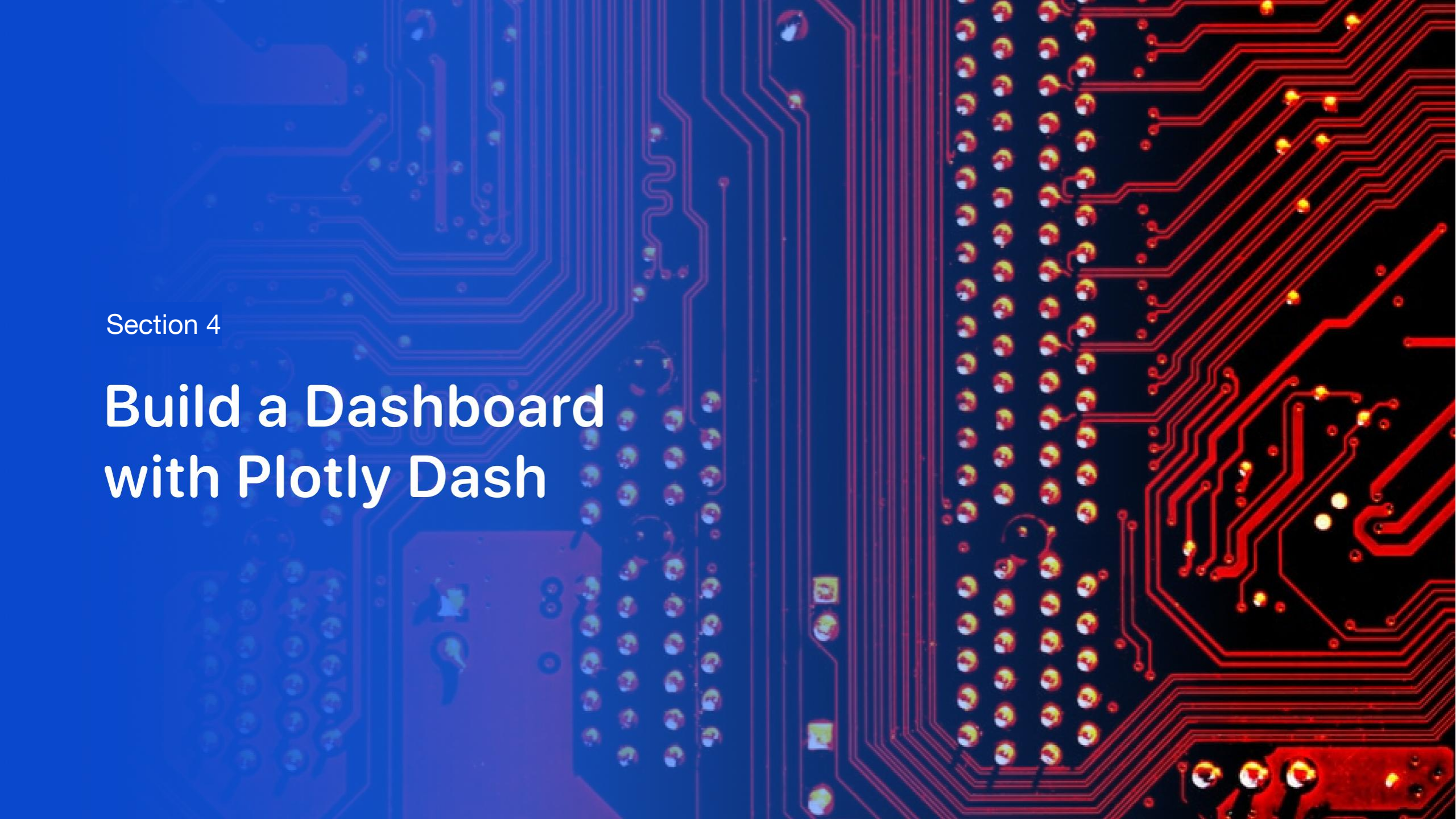




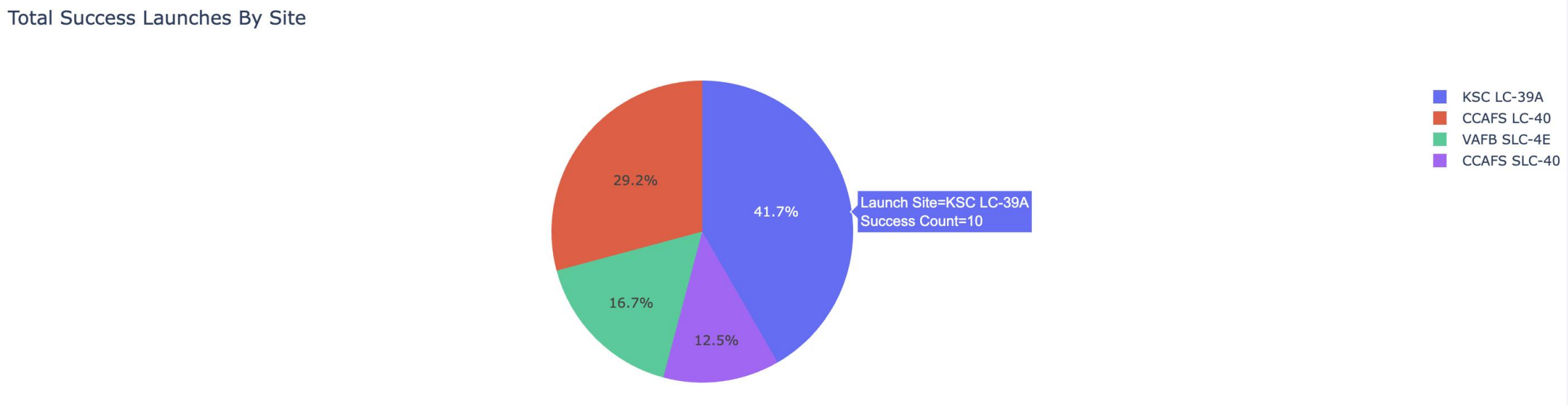
Security Consideration

- The launch site is far from railroad, highway, city. It's for the safety of the residents and the safety of the public infrastructure.



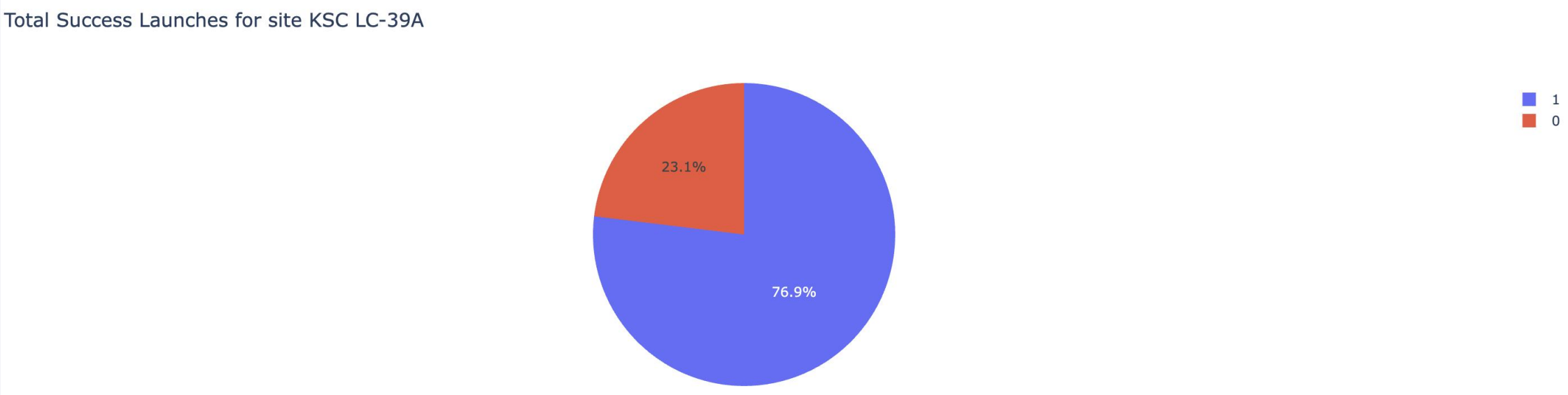

Section 4
Build a Dashboard with Plotly Dash

Total Success Launches by Site



- The launch site with the highest number of launches is KSC LC-39A, it's over 40%.
- The launch site with the lowest number of launches is CCAFS SLC-40, it's only 12.5%.
- The choice of rocket launch site is very important.

Success Rate at KSC LC-39A



- At KSC LC-39A, 76.9% of launches are successful.
- No launch site has ever exceeded an 80% success rate.

Which booster has a 100% success rate?

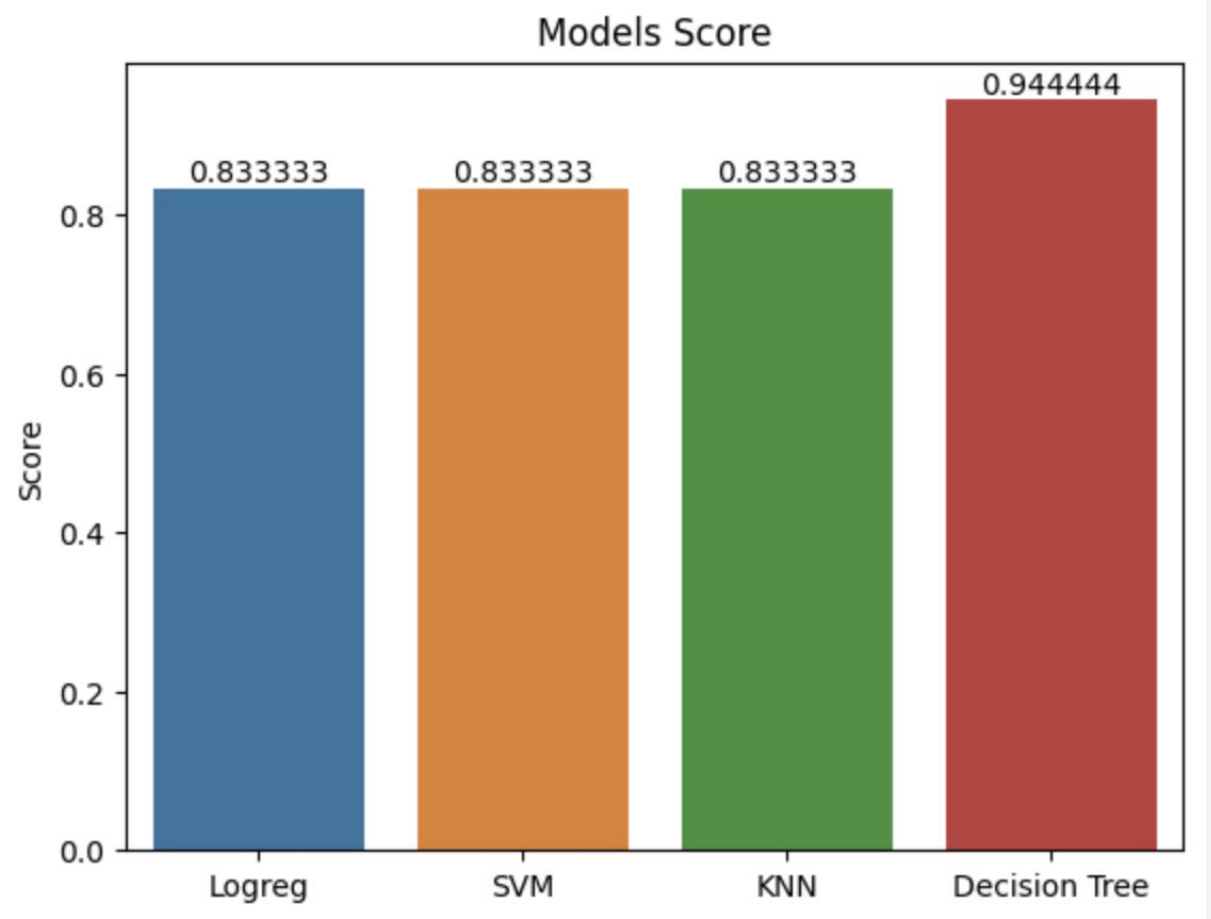
- B4 Booster has a 100% success rate in the 3000-5000 payload range.

Section 5

Predictive Analysis (Classification)

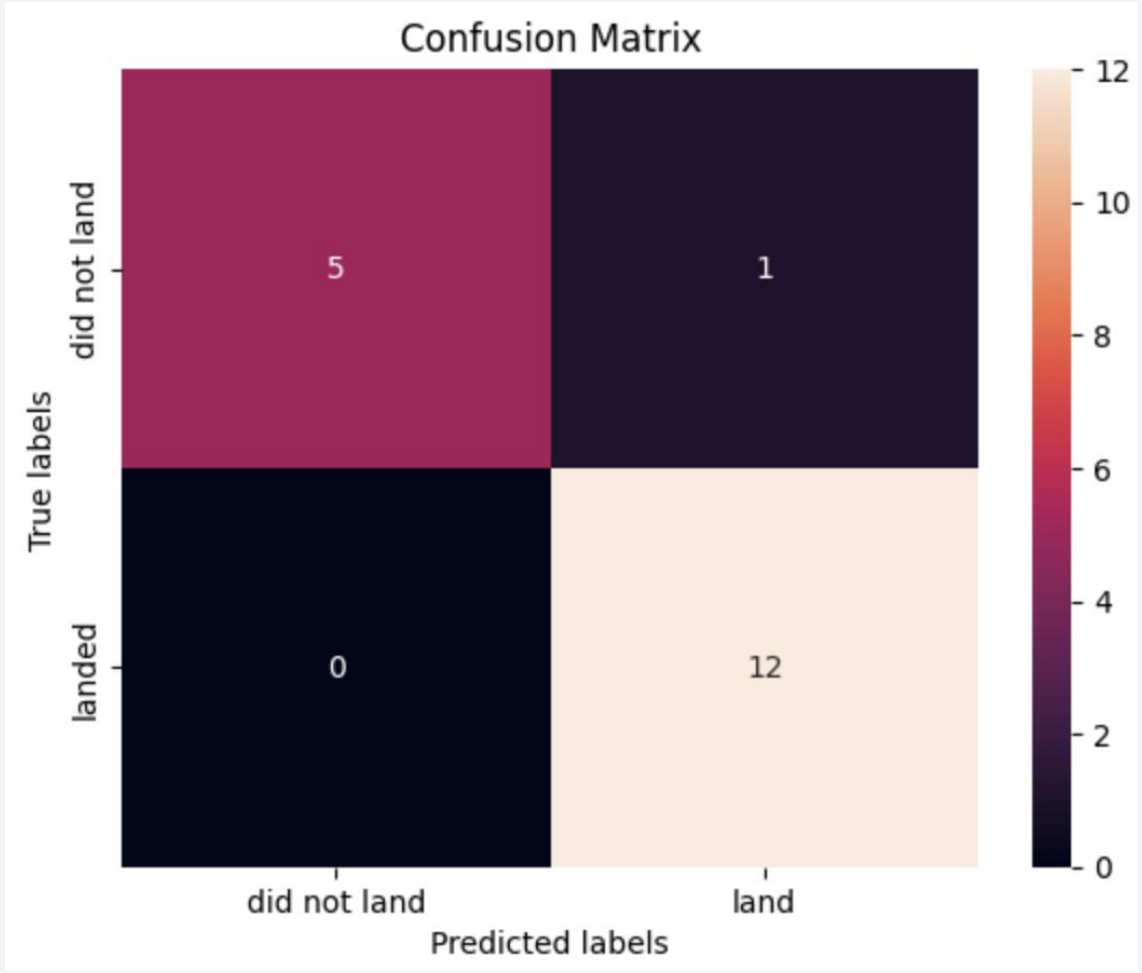
Classification Accuracy

- The model with the highest accuracy is the Decision Tree Classifier with an accuracy of 0.94.
- Best parameters:
    - criterion: entropy
    - max_depth: 4
    - max_features: sqrt
    - min_samples_leaf: 1
    - min_samples_split: 10
    - splitter: best



Confusion Matrix

- This is confusion matrix of decision tree classifier.
- True Postive is 12, same as the other models. But it's False Postive only 1, this result is much better than any other model.



Conclusions

- KSC LC 39A is the launch site with the highest launch success rate and the highest launch rate in most payload ranges. Best launch site in terms of overall capability.
- SSO is the orbit with the highest launch success rate.
- VLEO is the key launch orbit of the near future and the orbit with the largest payload, an orbit with potential but also risk.
- After the first successful recovery of the first stage rocket on 2015-12-22, the launch success rate increased exponentially.
- The latitude of the launch site is chosen to be low-dimensionally near the sea, for safety reasons, away from the railroad, highway, especially away from the city.
- Decision trees are the most accurate models and can later be used to predict whether or not a first stage rocket will be successfully recovered.

Appendix

- The dataset is mainly taken from the official SpaceX API.
- And the SpaceX Wikipedia page.
- Dashboard with Plotly Dash.

Thank you!