IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Onat Pehlivan
13.09.2024

Outline

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

Executive Summary

Methodologies

- Data collection from API and open source web scraping
- Data cleaning and wrangling
- EDA with SQL and pandas
- Creating dashboards
- ML prediction

Results

Through the methodologies employed we determined what ML models offer the best predictive performance for the dataset, along with the ideal hyperparameters.

Introduction

As a corporation that aims to stage rocket launches, our aim was to extract insights in this context using the data of another firm that operates in the same field.

We collected data related to our competitor’s operations in order to achieve this.

Section 1

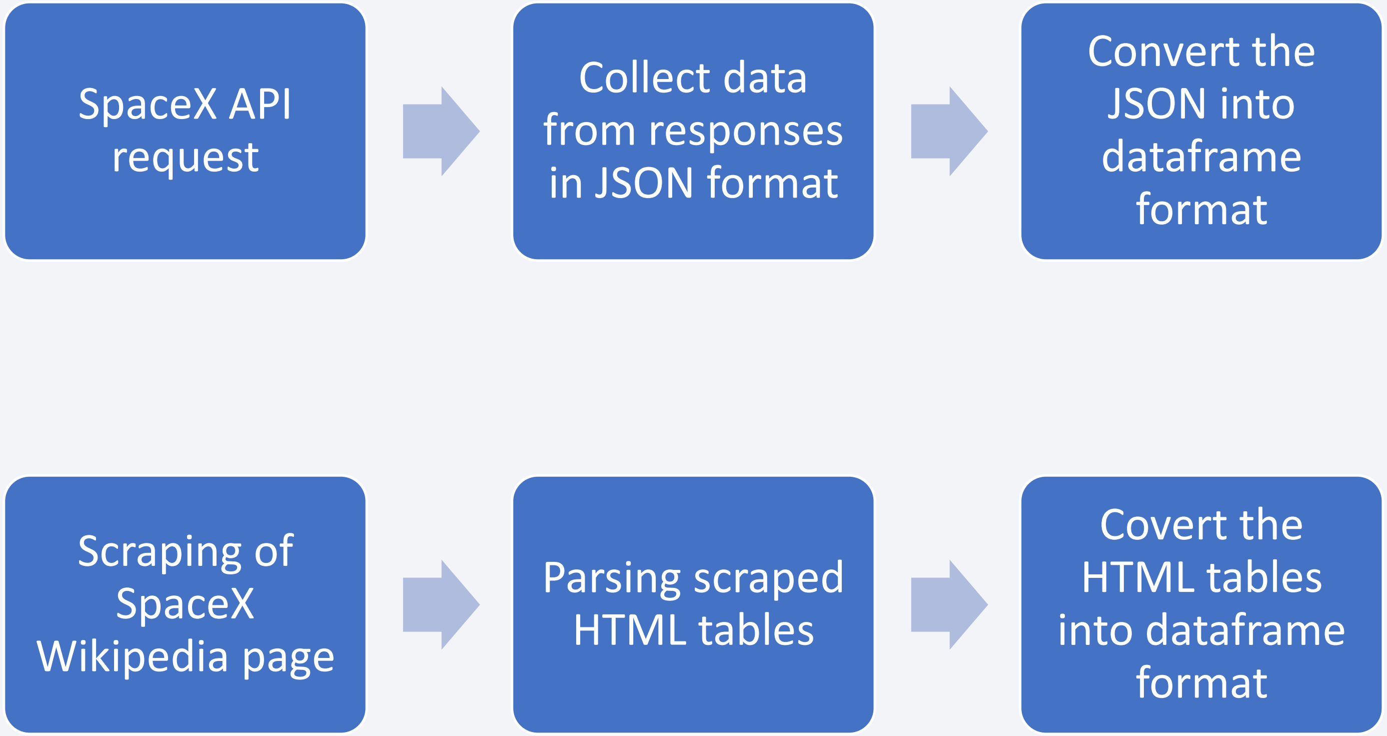
Methodology

Methodology

Executive Summary

- Data collection methodology:
  - We collected data through SpaceX API and web scraping of the Wikipedia page related to SpaceX's rocket launches.
- Perform data wrangling
  - API data was collected as JSON and scraped data was saved as HTML. We converted both the data into a pandas dataframe.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Different ML Classification models were tuned and results were compared regarding their predictive performance on whether a rocket will land successfully or not.

Data Collection



SpaceX API request

Collect data from responses in JSON format

Convert the JSON into dataframe format

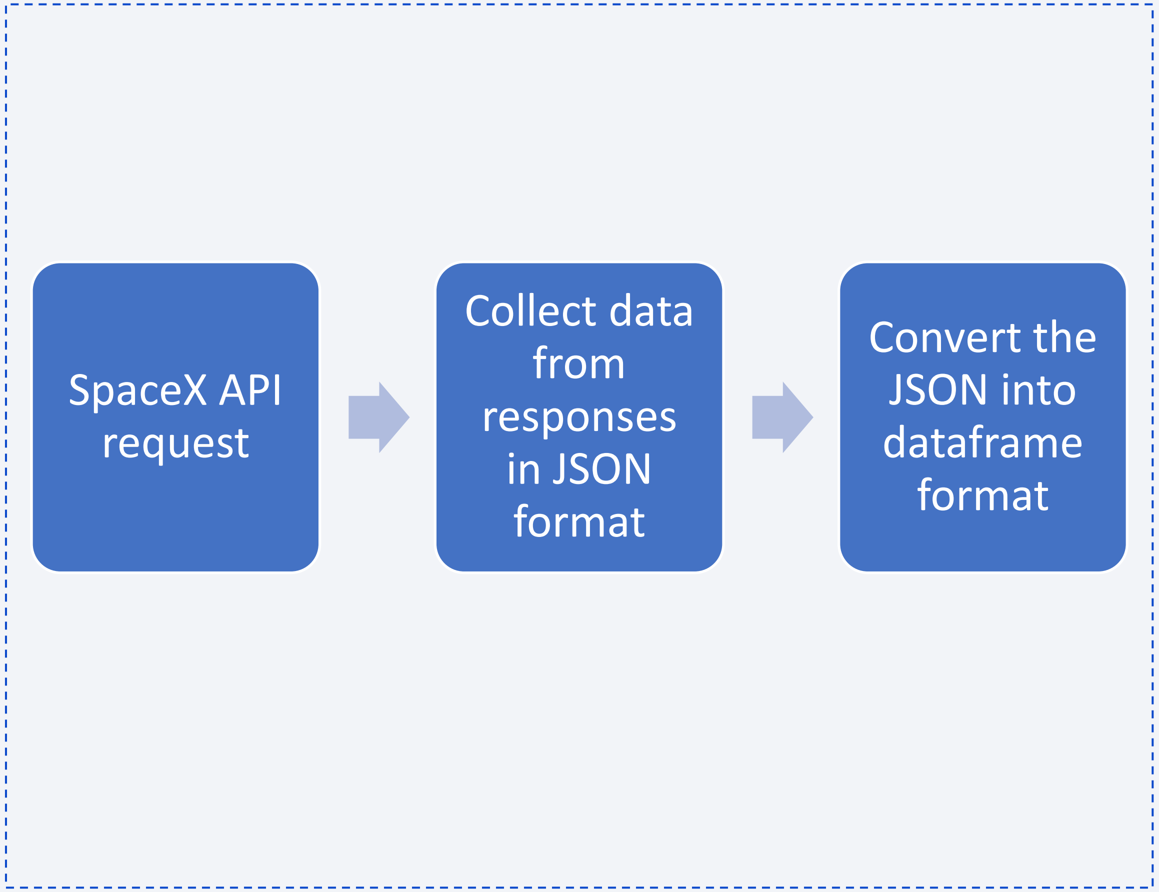
Scraping of SpaceX Wikipedia page

Parsing scraped HTML tables

Covert the HTML tables into dataframe format

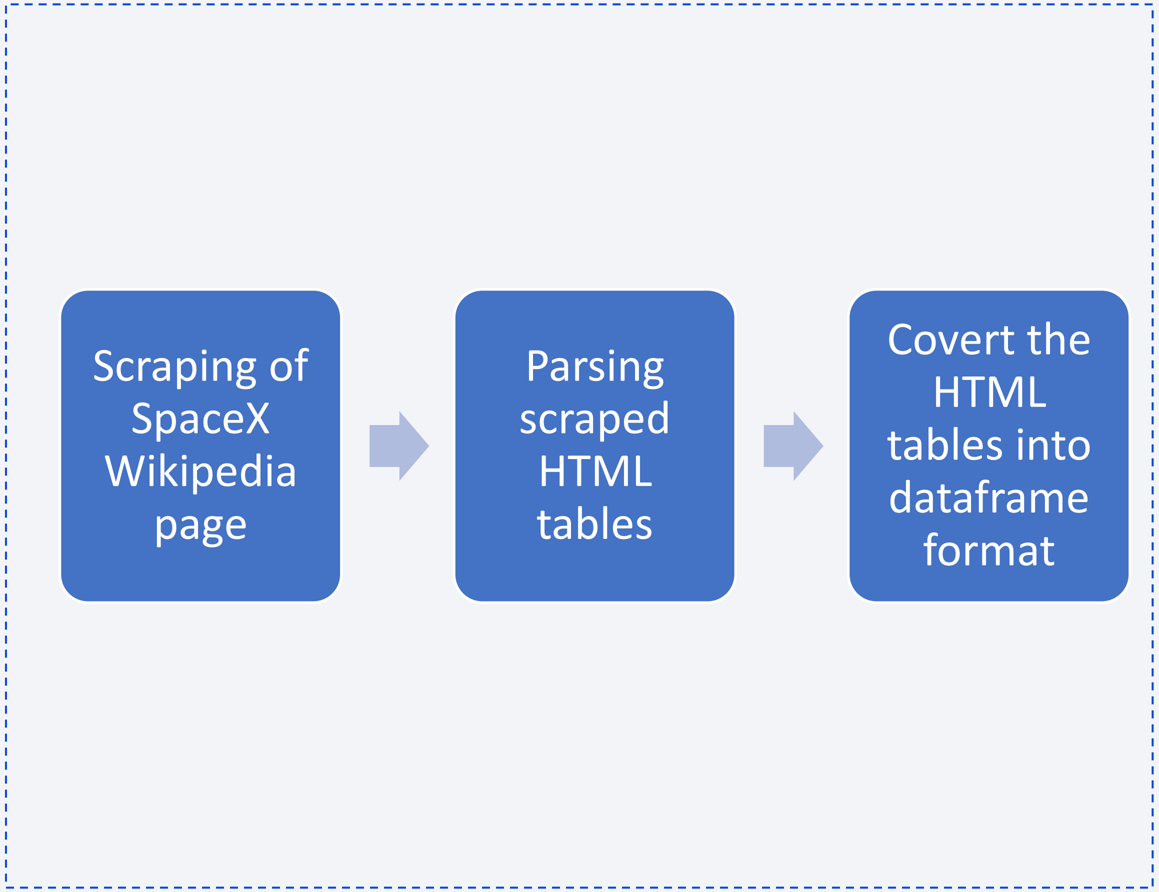
Data Collection – SpaceX API

API GitHub Page

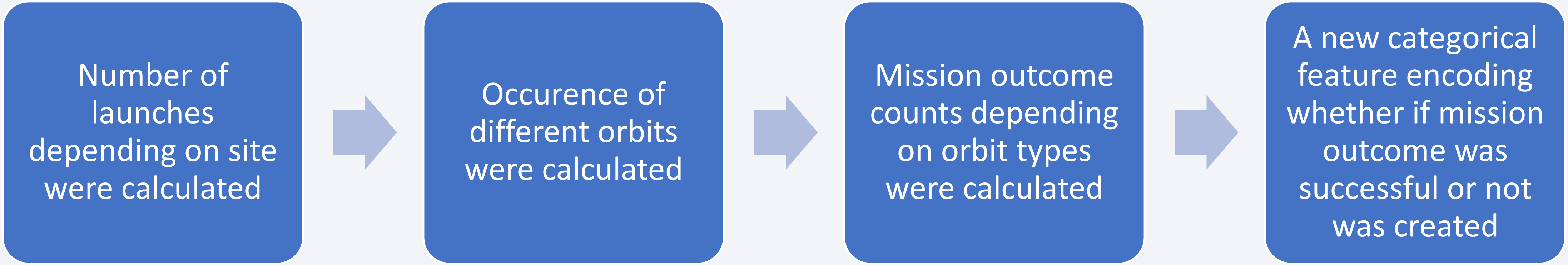

SpaceX API request
Collect data from responses in JSON format
Convert the JSON into dataframe format

Data Collection - Scraping

[Web Scraping GitHub]


Scraping of SpaceX Wikipedia page
Parsing scraped HTML tables
Covert the HTML tables into dataframe format

Data Wrangling



Number of launches depending on site were calculated

Occurence of different orbits were calculated

Mission outcome counts depending on orbit types were calculated

A new categorical feature encoding whether if mission outcome was successful or not was created

Data Wrangling GitHub

EDA with Data Visualization

- Catplots were used to visualize relationships between:
  - Flight Numbers and Payloads
  - Flight Numbers and Launch Sites
  - Payloads and Launch Sites
  - Flight Numbers and Orbit Types
  - Payloads and Orbit Types
- Line graph was used to draw the launch success trend depending on date
- Bar plot was used to visualize the success rates of launches depending on orbit types.

EDA GitHub

EDA with SQL

- Display the names of the unique launch sites in the space mission
  - ***%sql SELECT DISTINCT("Landing_Outcome") FROM SPACEXTABLE;***
- Display 5 records where launch sites begin with the string 'CCA’
  - ***%sql SELECT * FROM SPACEXTABLE 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 SPACEXTABLE WHERE "Customer" LIKE 'NASA (CRS)'***
- Display average payload mass carried by booster version F9 v1.1
  - ***%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE "Booster_Version" LIKE 'F9 v1.1%'***
- List the date when the first succesful landing outcome in ground pad was acheived.
  - ***%sql SELECT MIN(Date) FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE 'Success (ground pad)’***
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - ***%sql SELECT DISTINCT(Booster_Version) FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000 AND "Landing_Outcome" LIKE 'Success (drone ship)'***
- List the total number of successful and failure mission outcomes
  - **%sql SELECT COUNT(Mission_Outcome) AS Success_Outcome FROM SPACEXTABLE WHERE Mission_Outcome LIKE 'Success%'**
  - **%sql SELECT COUNT(Mission_Outcome) AS Failure_Outcome FROM SPACEXTABLE WHERE Mission_Outcome LIKE 'Failure%'**
- List the names of the booster_versions which have carried the maximum payload mass.
  - ***%sql SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)***
- List the records which will display the month names, failure landing_outcomes in drone ship, booster versions, launch_site for the months in year 2015.
  - ***%sql SELECT Booster_Version, Launch_Site, Landing_Outcome FROM SPACEXTABLE WHERE Landing_Outcome LIKE 'Failure (drone ship)' AND Date BETWEEN '2015-01-01' AND '2015-12-31'***
- 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.
  - ***%sql SELECT Landing_Outcome, COUNT(Landing_Outcome) FROM SPACEXTABLE WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY COUNT(Landing_Outcome) DESC***

Build an Interactive Map with Folium

- Circle and Marker objects were introduced using folium. These were used to highlight specific areas of launch sites.
- Marker clusters to show that there are multiple markers present at the same coordinate.
- Mouse position object to increase interactivity by providing coordinates of the location where the cursor is.
- Poly line object to draw lines between launch sites and the closest city, railroad or highway to it.

Predictive Analysis (Classification)

- The «Class» column was separated from the data as the feature to be predicted by the model. Rest of the features that would be used as input were standardized.
- Data was split into training (80%) and test data (20%).
- Logistic Regression, SVM, Decision Tree and KNN models were used.
- Grid searching and cross validation was employed to tune model hyperparameters and increase generalization.
- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose

Results

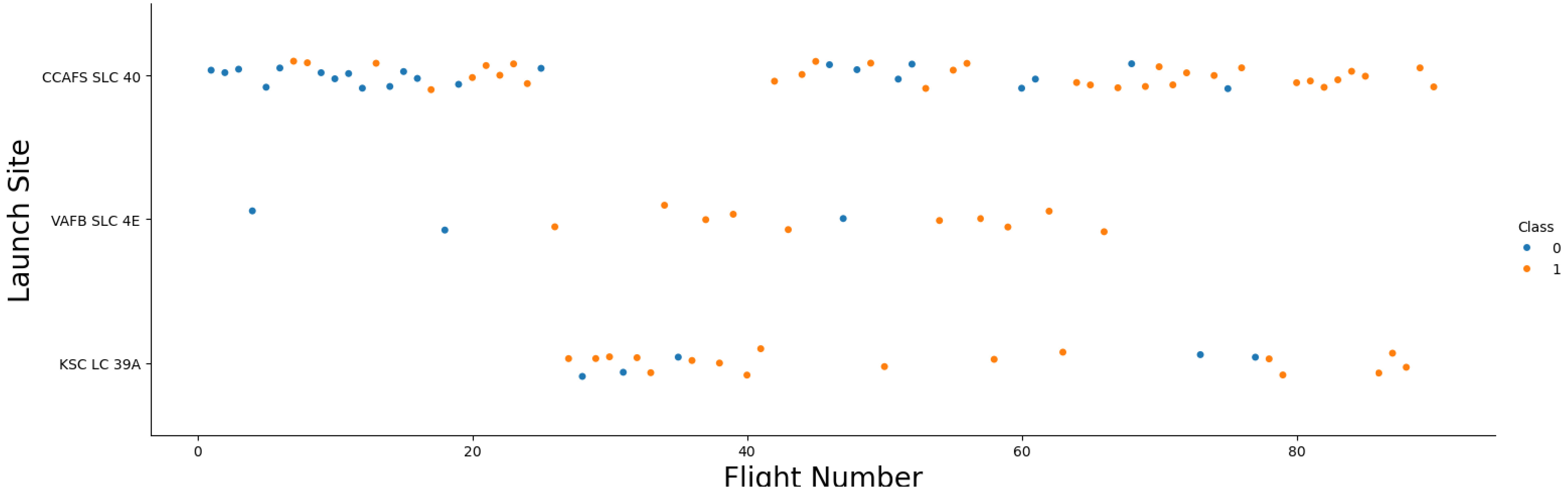
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

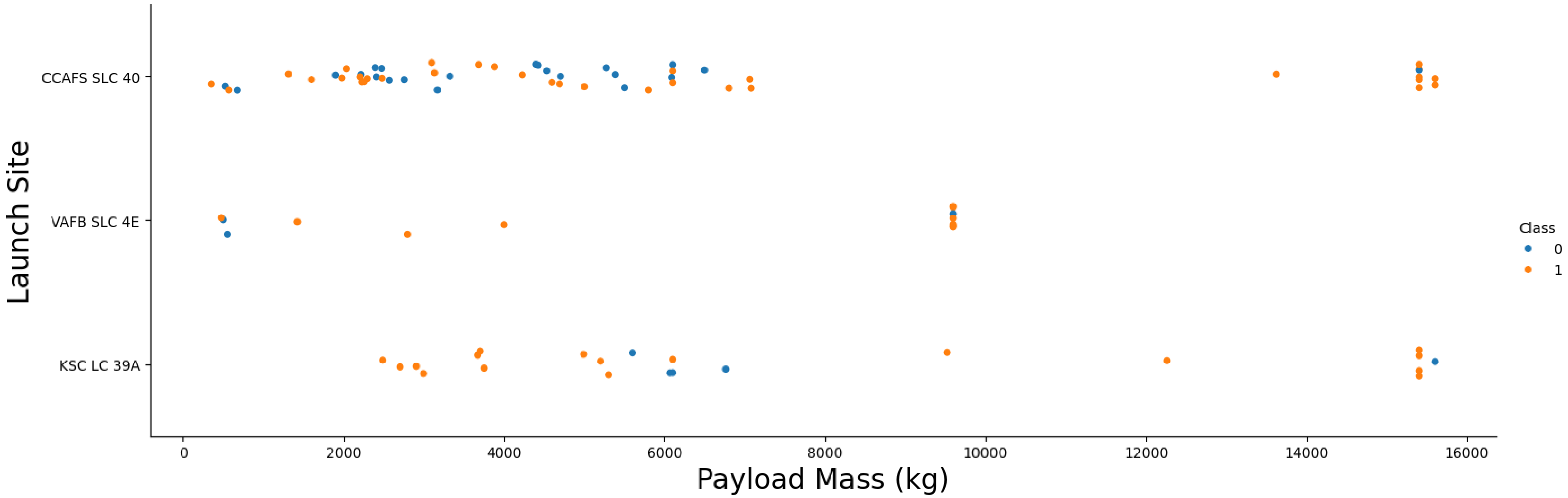
Flight Number vs. Launch Site

- It’s observed that as SpaceX was able to achieve success in more launches as they gained experience. Their launch success rate increased as they launched more rockets.



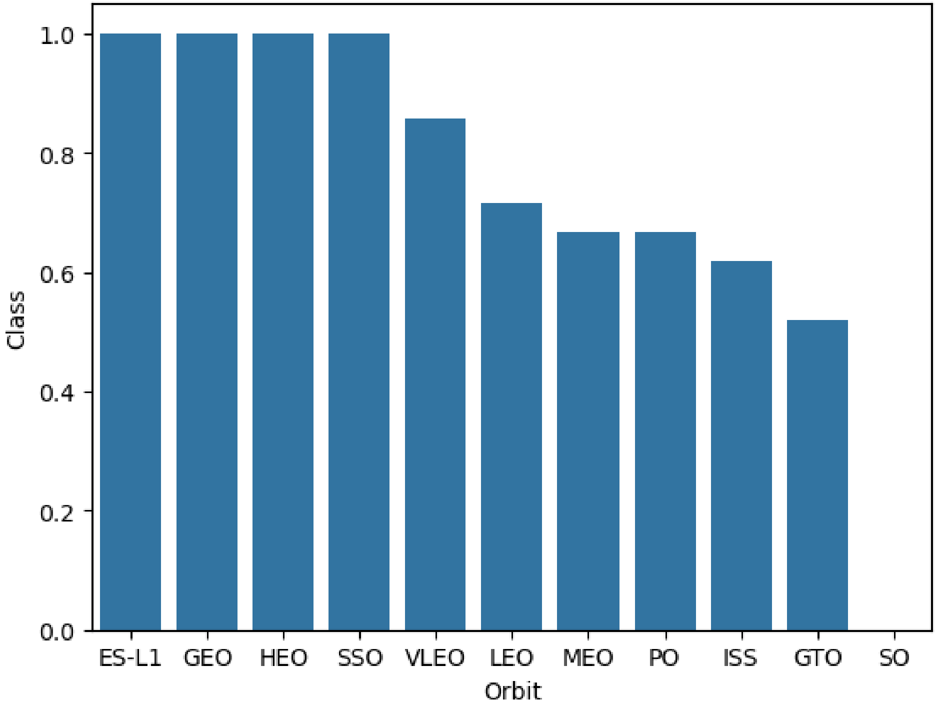
Payload vs. Launch Site

- Although SpaceX appears to have focused on launches with lighter loads, their success rate is higher with heavier launches.



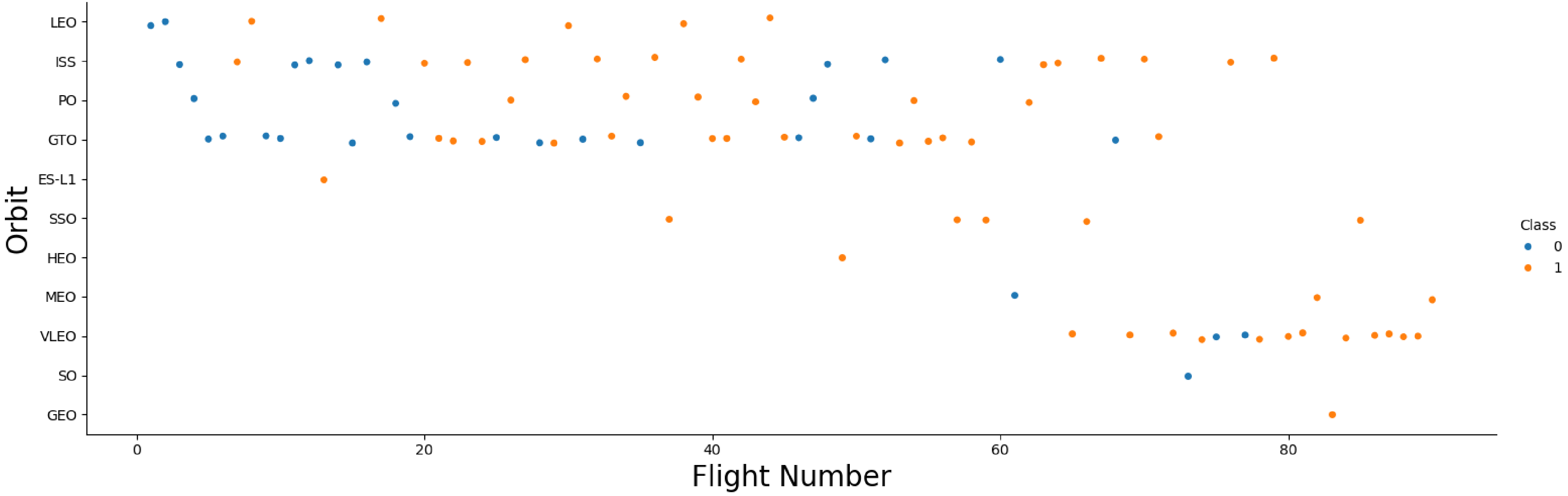
Success Rate vs. Orbit Type

- The plot shows all launches to orbits ES-L1, GEO, HEO and SSO were a success.



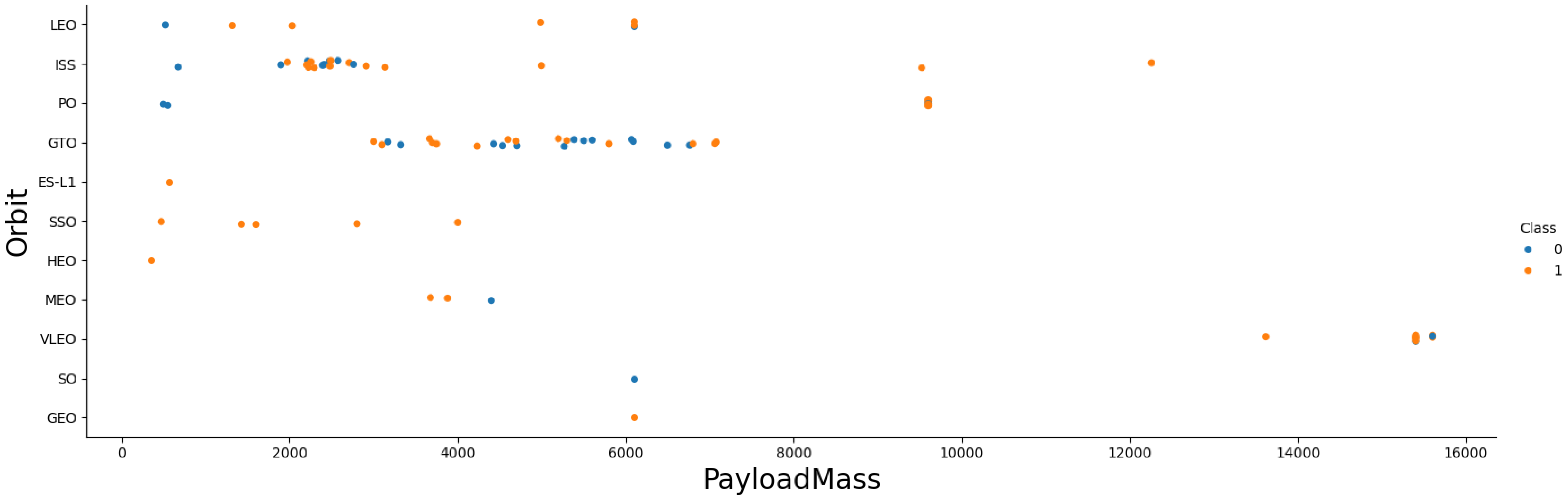
Flight Number vs. Orbit Type

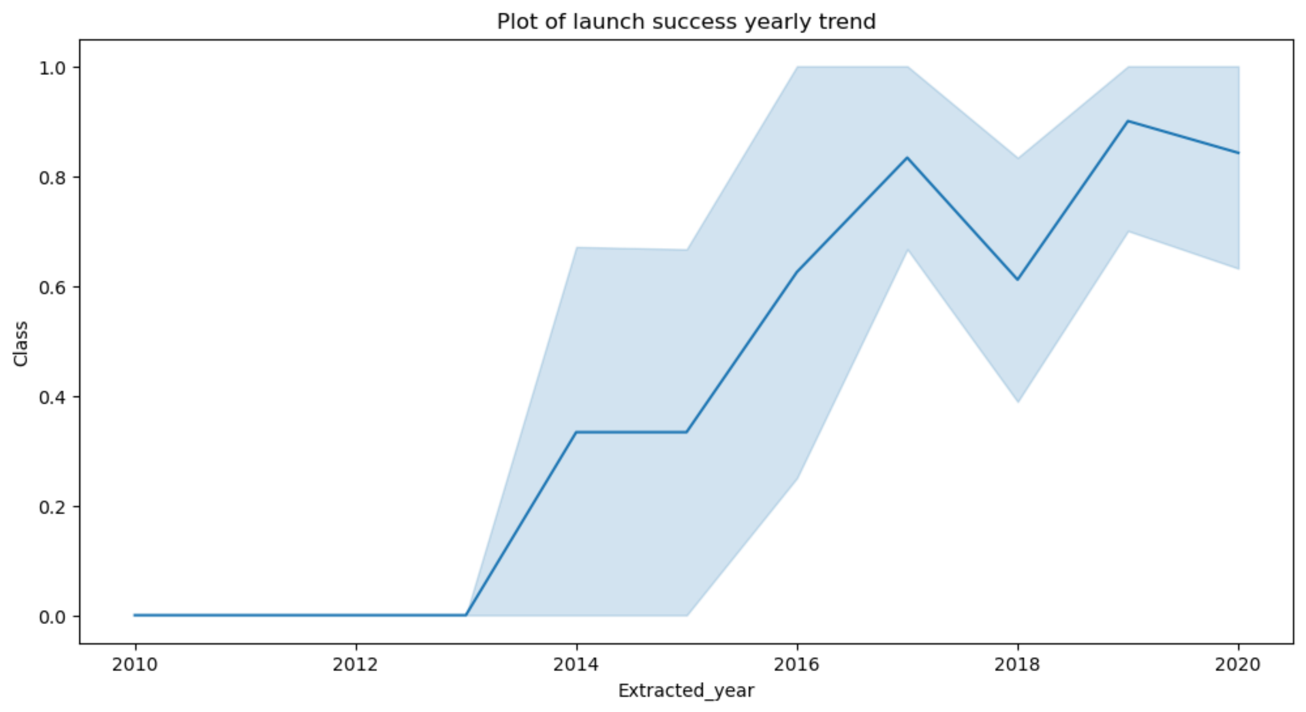
- However as seen on this scatter plot, orbits ES-L1, GEO, HEO and SSO had significantly fewer launches compared to other orbits, and therefore the 100% success rate must be evaluated in this context.



Payload vs. Orbit Type

- GTO has the most payload variation in launch attempts made.
- The range of payload is relatively smaller for ISS orbit.
- SSO launches are promising for lighter payloads, however lack samples with heavier loads.



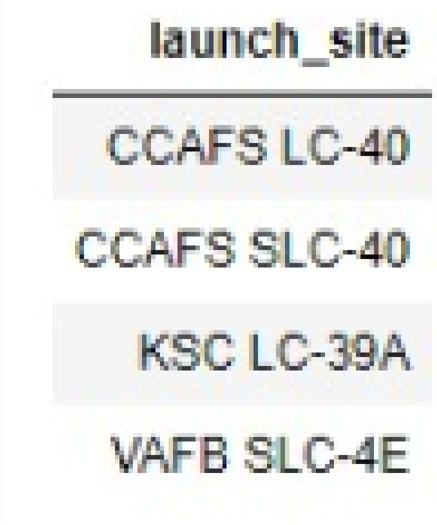


Launch Success Yearly Trend

- Launch success rate appears to rise through time in general.
- However it’s visible that there’s a drop of success rates between 2017-2018.

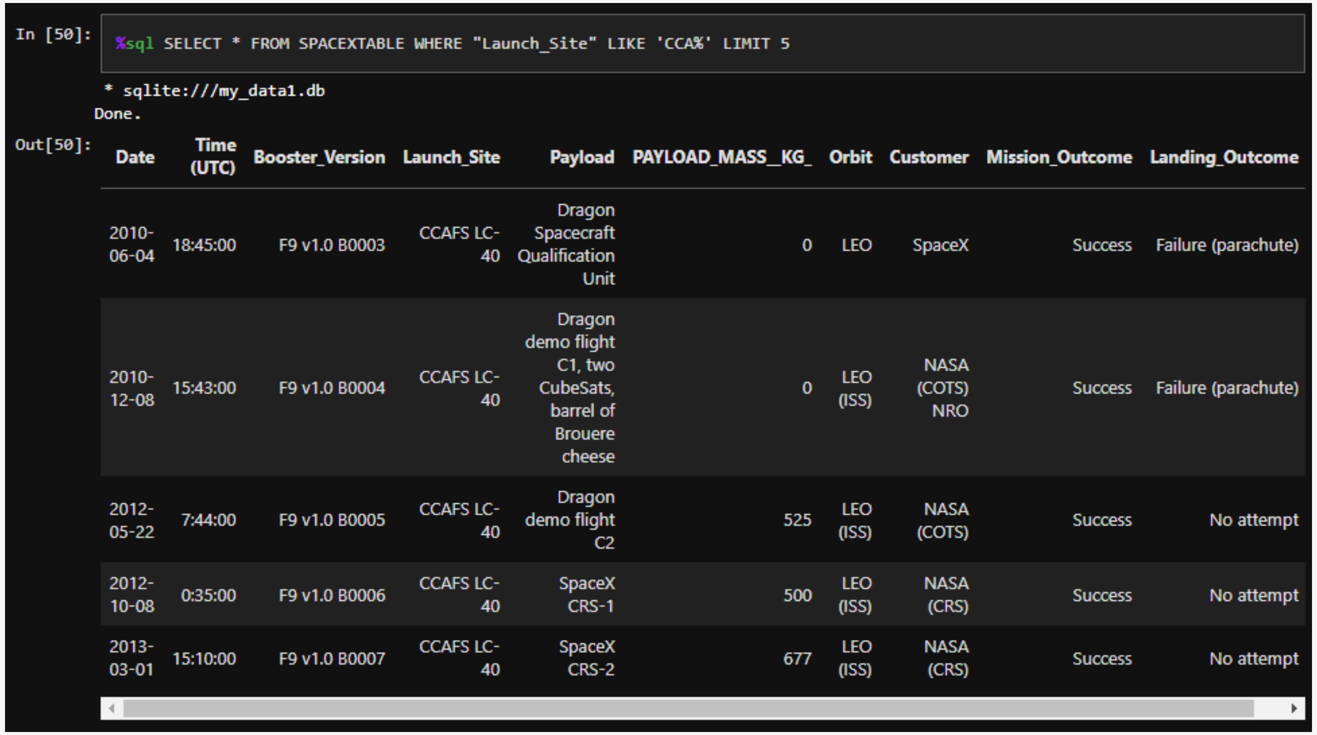
All Launch Site Names

- There are 4 different launch sites in total.
- SQL’s DISTINCT statement was used to extract this table.



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

Launch Site Names Begin with 'CCA'



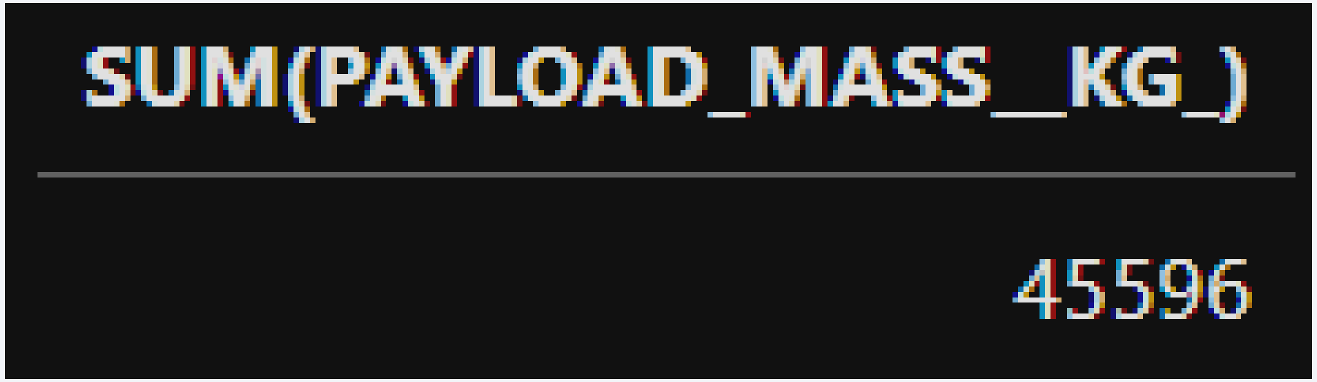
```
In [50]: %sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5

 * sqlite:///my_data1.db
Done.
Out[50]:
```

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

- SQL’s SELECT, WHERE, LIKE and LIMIT statements were used in the query to extract the table.

Total Payload Mass

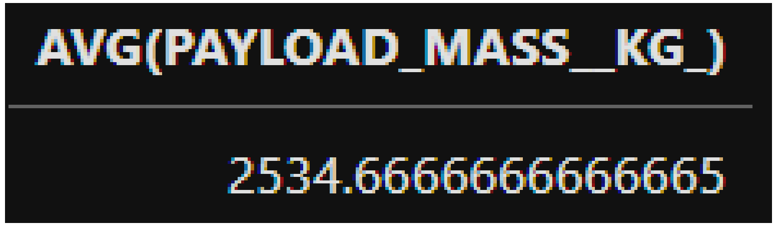
- SQL’s SELECT, WHERE, LIKE and SUM statements were used in the query to extract the table.



SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

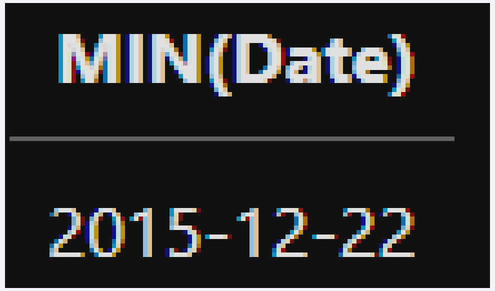
SQL's SELECT, WHERE, LIKE and AVG statements were used in the query to extract the table.



AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

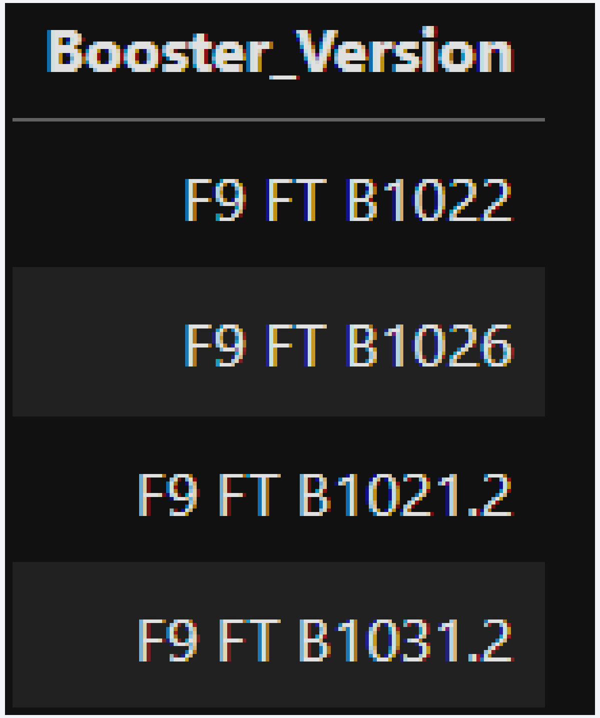
SQL’s SELECT, WHERE, LIKE and MIN statements were used in the query to extract the table.



MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

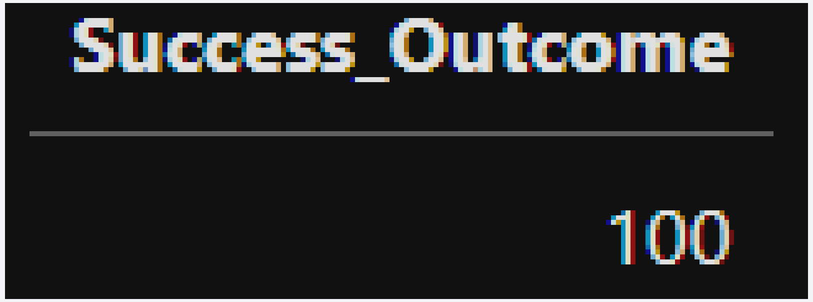
SQL’s SELECT, WHERE, AND, DISTINCT, and operators < and > were used in the query to extract the table.



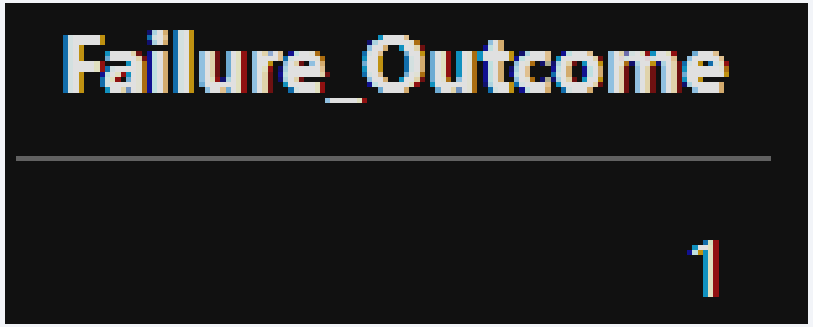
Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

SQL's SELECT, COUNT, AS and WHERE statements were used in the query to extract the table.



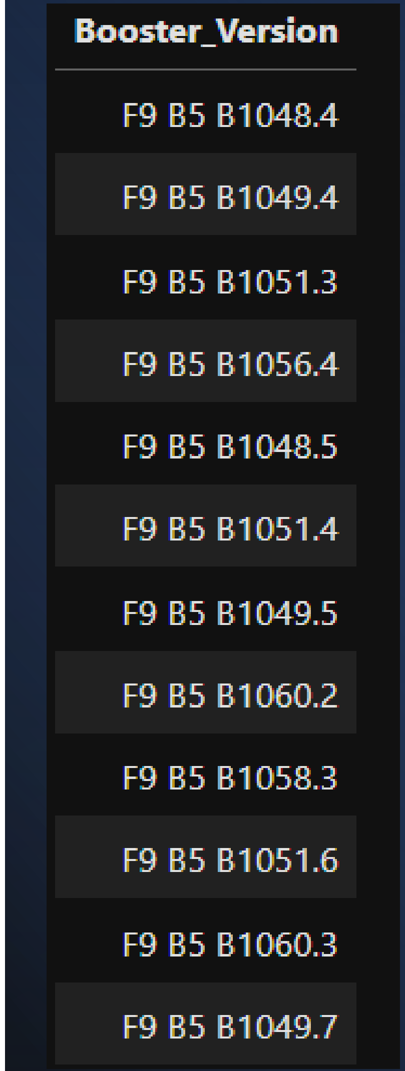
Success_Outcome
100



Failure_Outcome
1

Boosters Carried Maximum Payload

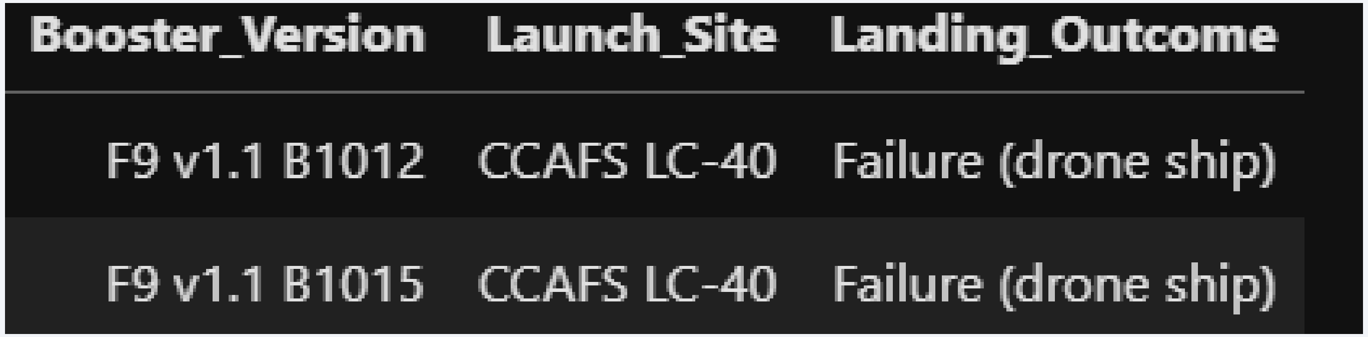
- SQL’s SELECT, WHERE, and MAX statements were used in the query, along with a subquery to extract the table.



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

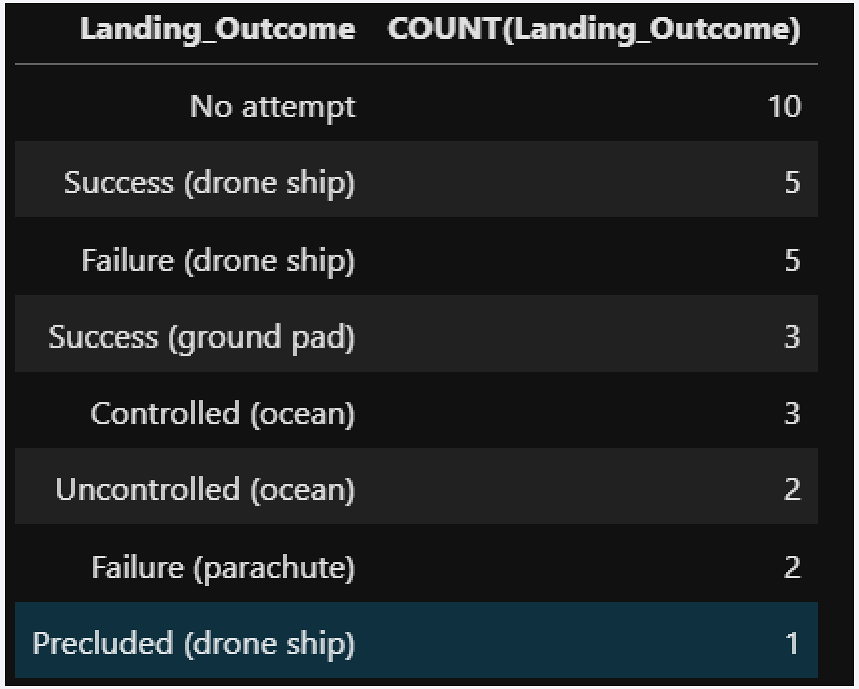
- SQL’s SELECT, WHERE, AND and BETWEEN statements were used in the query to extract the table.



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

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

- SQL’s SELECT, COUNT, DATE, BETWEEN, AND, GROUP BY, ORDER BY AND DESC statements were used in the query to extract the table.



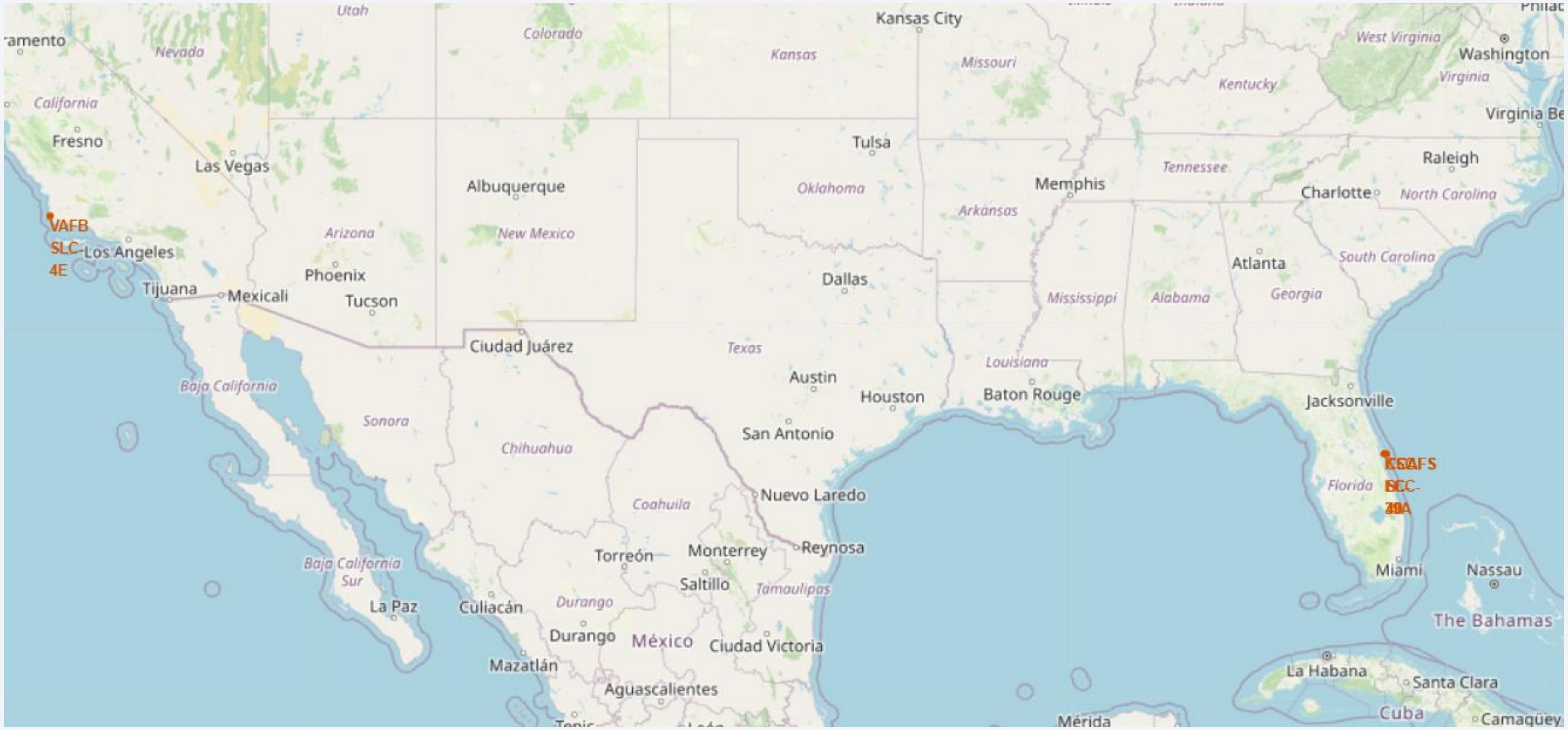
Landing_Outcome	COUNT(Landing_Outcome)
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

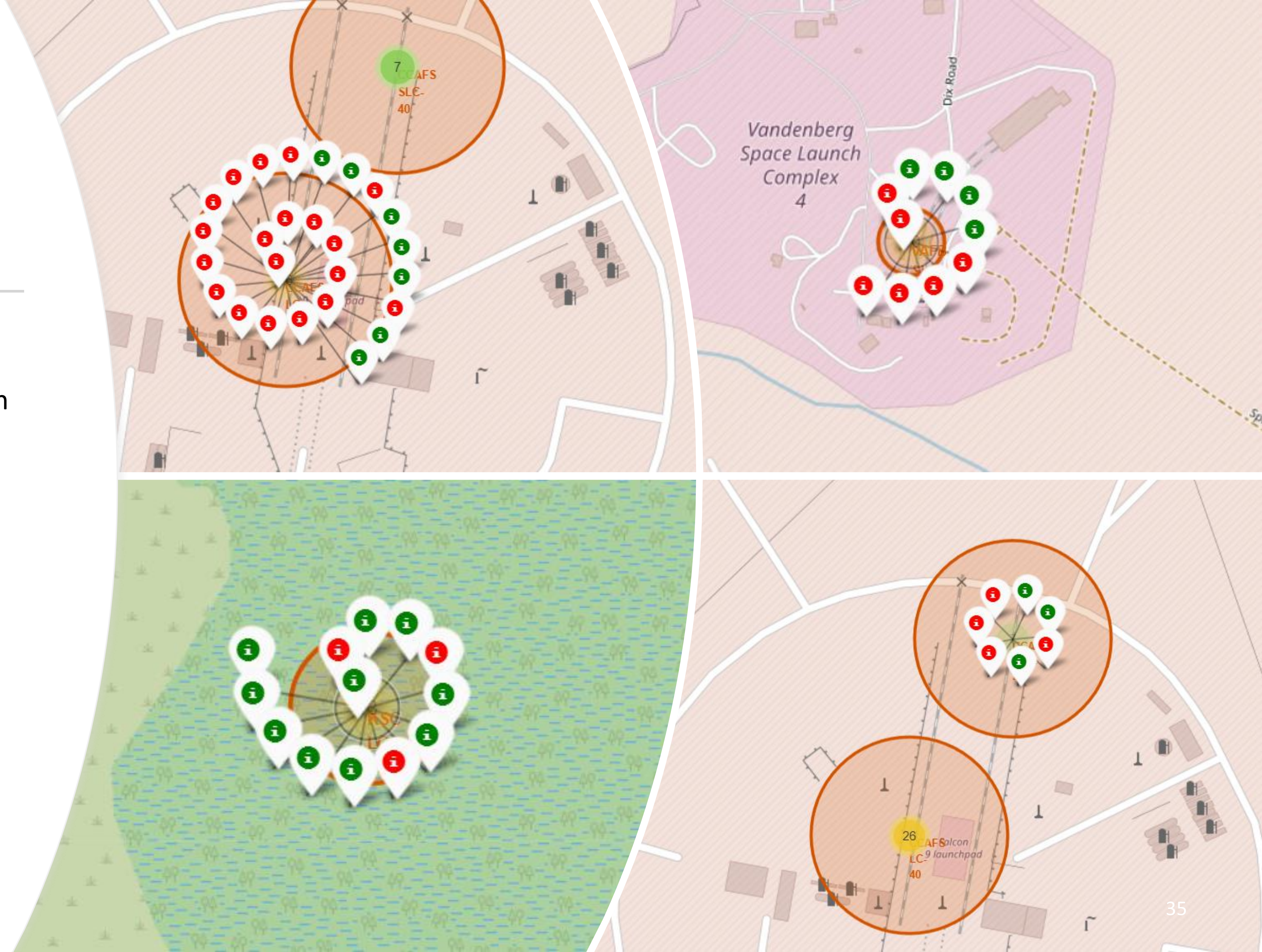
Location of Launch Sites

- It’s observed that SpaceX mainly operates on two lauch sites, one in California, and the other Florida.



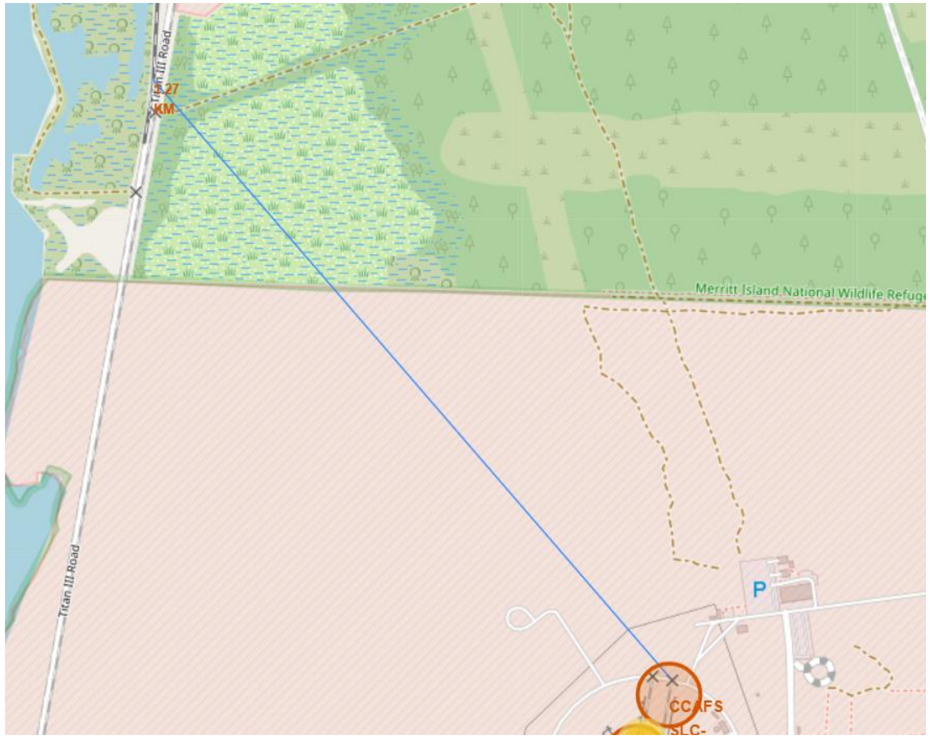
Launch Outcomes on Map

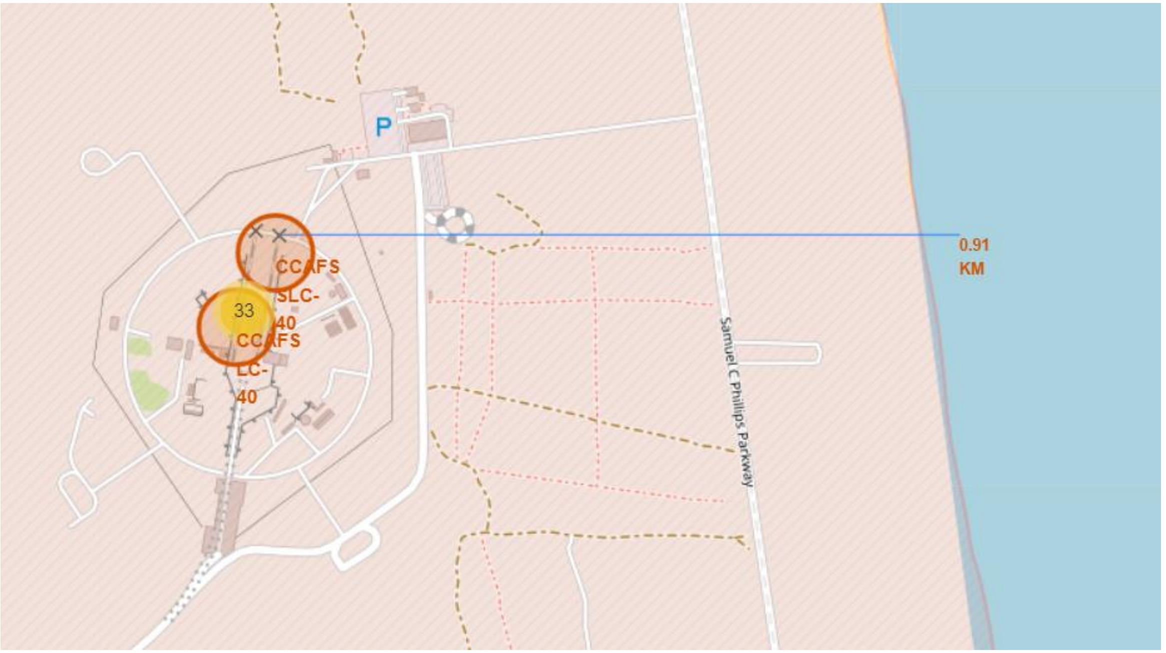
- There were a total of four separate launch sites with varying launch outcomes in terms of success. Launch site KSC LC-39A had the most success rate in terms of launch outcomes.



Connections of Launch Sites

It appears that various launch sites have various connnections to mainstream transit channels such as coastlines, railways or highways. The means tend to differ due to geographical circumstances.



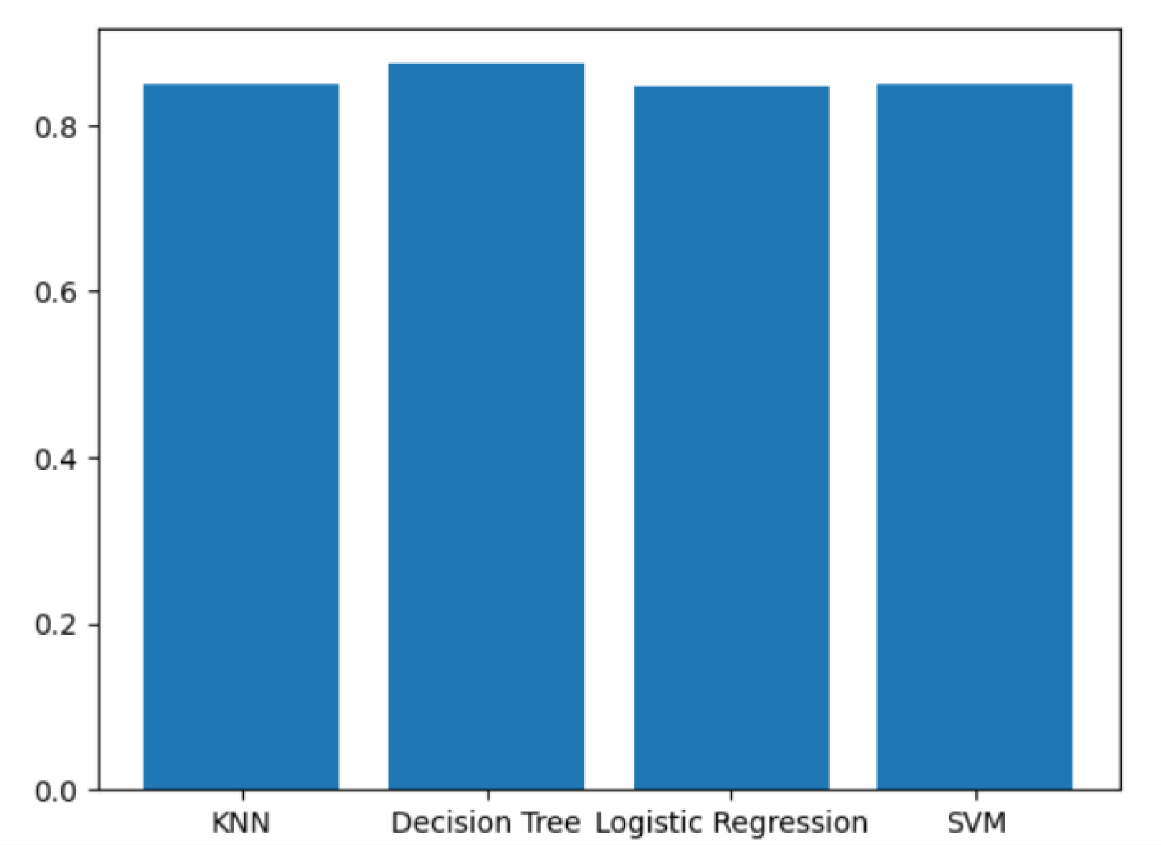


Section 4

Predictive Analysis (Classification)

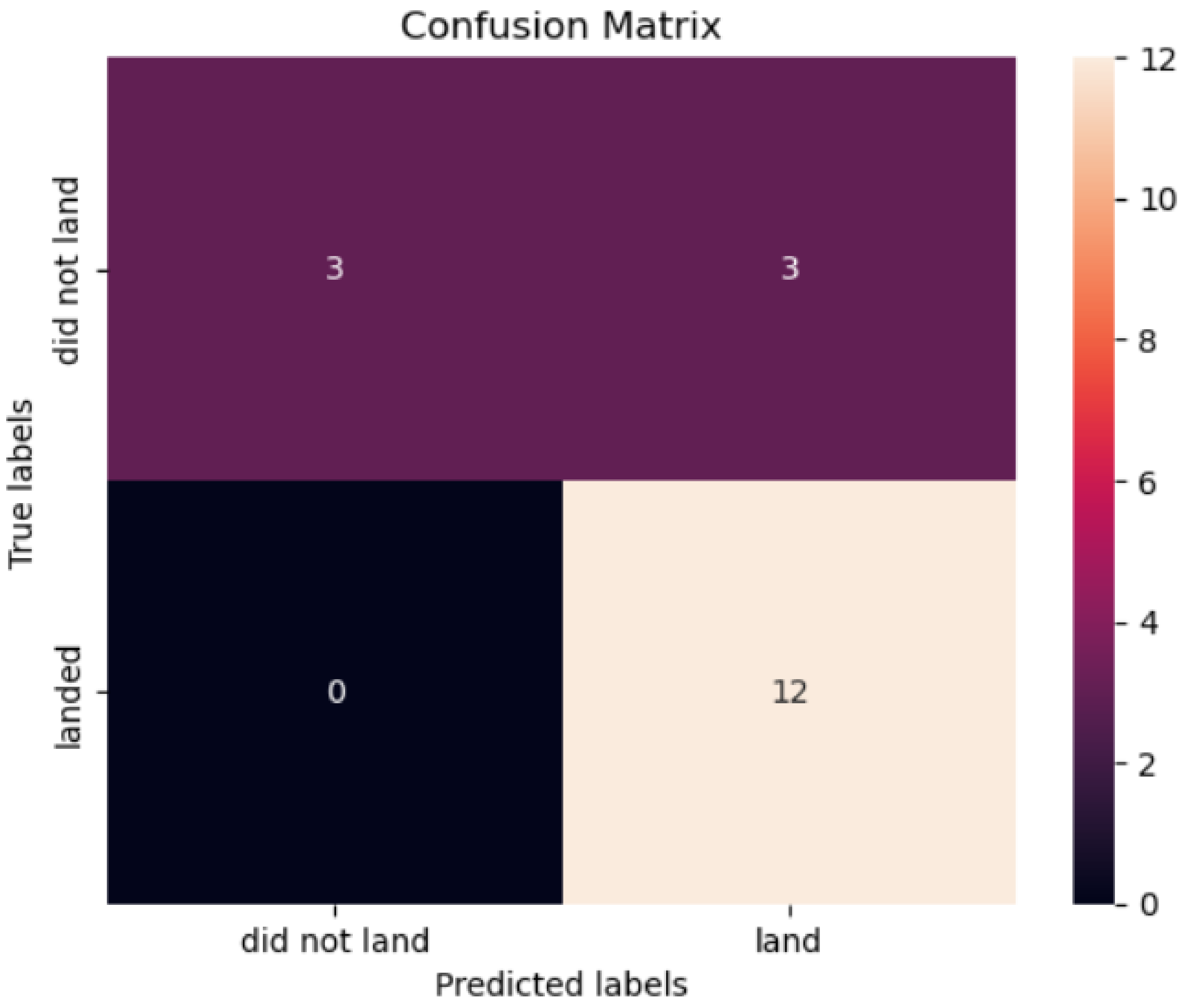
Classification Accuracy

- The best accuracy was provided by Decision Tree Model.



Confusion Matrix

Decision tree model performed the best using the data provided.



Conclusions

- It’s possible to predict launch outcomes with up to ~87% accuracy using Decision Tree Classifiers.
- All models provided the same classification distribution in practice, due to the limited amount of data.
- As useful a method open source data collection may be (web scraping, API calls), the data may provide insufficient in context of volume or features to create a deployable model.

Thank you!