IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

M. Zhao
May 4, 2023

Outline

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

Executive Summary

- This study aims to identify the factors influencing the successful landing of the first stage of Space X Falcon 9 rocket and develop classification models to predict the launch outcome using data publicly available. The data sets were collected from Space X API and Wikipedia. Exploratory data analysis was performed using visualization and SQL query. Interactive Folium map and Plotly dashboard were created to discover the patterns between launch outcome and different factors. Predictive analysis was conducted using four models including logistic regression, support vector machine, decision trees, and k-nearest neighbors.
- The results showed the launch site, payload mass, orbit, and the number of previous launches may affect the launch outcome. Depending on the orbit, the payload mass can affect the success rate as some orbits better fit for a light or heavy payload. Most heavy payload (>10,000 kg) launches were successful; for payload below 10,000 kg, the highest success rate was observed for payload between 3,000 and 4,000 kg. Orbits ES-L1, GEO, HEO and SSO had 100% success rate. All launch sites were in close proximate to coastline. The success rates varied by site with KSC LC-39A having the highest success rate (76.9%). The success rates showed a generally increasing trend over the years since 2013, with about 80% for most recent years. For the data studied, decision trees model performed best at predicting launch outcome in terms of accuracy, precision, F1 score, and ROC-AUC score.

Introduction

Background and Context

- SpaceX has gained worldwide attention for a series of historic milestones. The company advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars, other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information could be helpful if an alternative company wants to bid against SpaceX for a rocket launch. Using machine learning models and data publicly available, this project aims to predict if Falcon 9 first stage will land successfully.

Study Questions

- How do variables like launch site, payload mass, number of flights and orbits affect the success of the first stage landing?
- What is the trend of successful landing over the years?
- What is the best model to predict the landing outcome?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data collection using API and web scrapping
- Perform data wrangling
  - Data cleaning and transforming
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis rm using classification models
  - Model building, validation, and evaluation

Data Collection

- Data collection process involved a combination of API requests from Space X REST API and web scraping data from a table in Space X's Wikipedia entry.

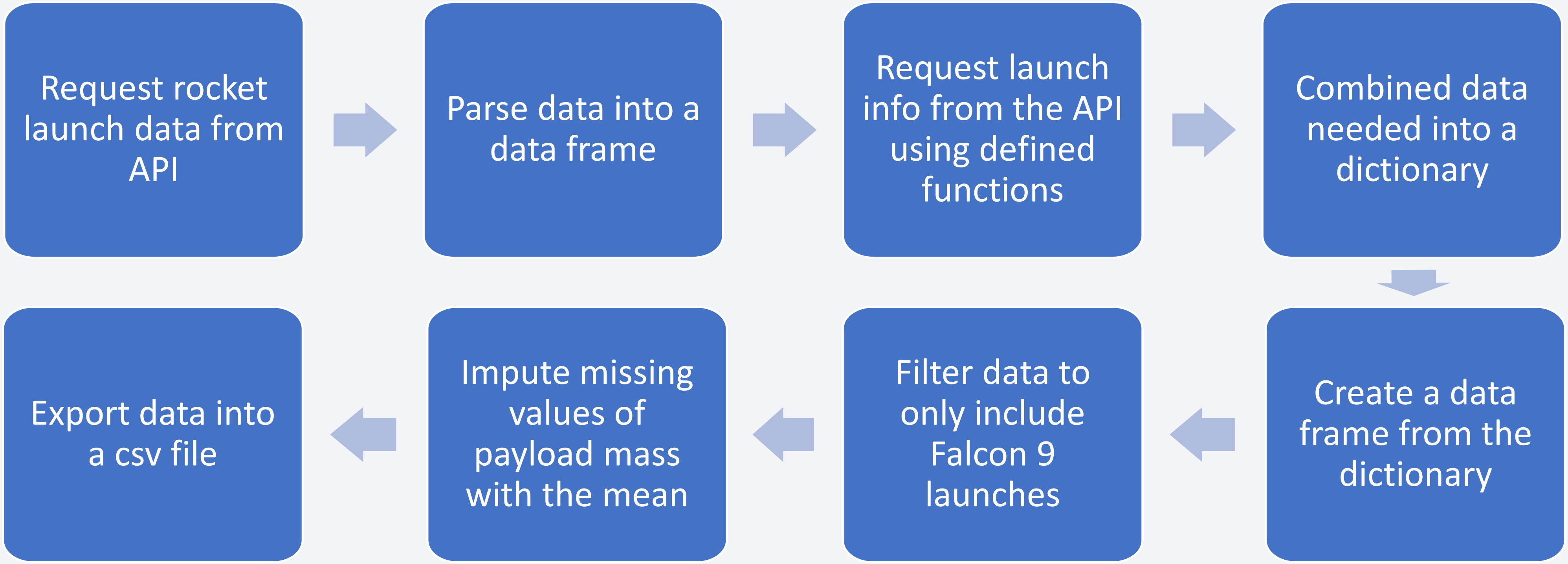
  - Variables collected from Space X REST API:

  FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Latitude, Longtitude

  - Variables Collected from Wikipedia:

  Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



Request rocket launch data from API

Parse data into a data frame

Request launch info from the API using defined functions

Combined data needed into a dictionary

Create a data frame from the dictionary

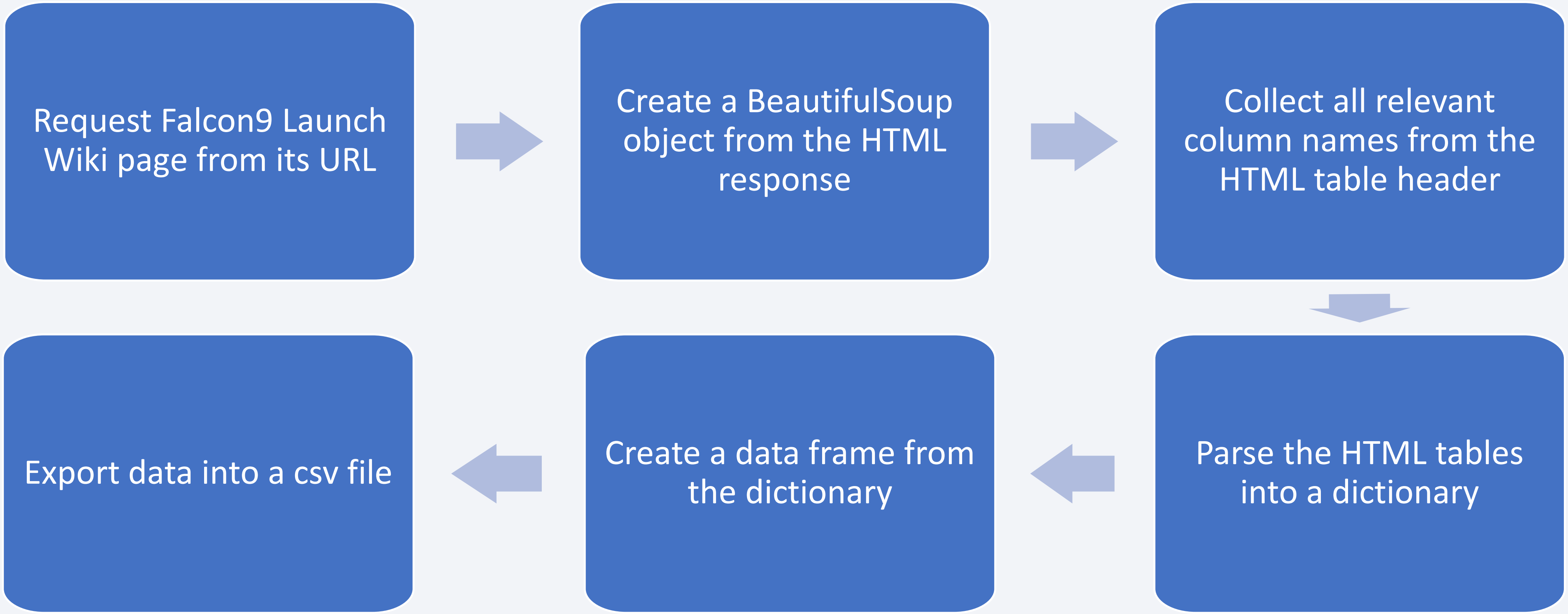
Filter data to only include Falcon 9 launches

Impute missing values of payload mass with the mean

Export data into a csv file

Link to codes: https://github.com/hikarikk/Final-Presentation/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping



Request Falcon9 Launch Wiki page from its URL

Create a BeautifulSoup object from the HTML response

Collect all relevant column names from the HTML table header

Parse the HTML tables into a dictionary

Create a data frame from the dictionary

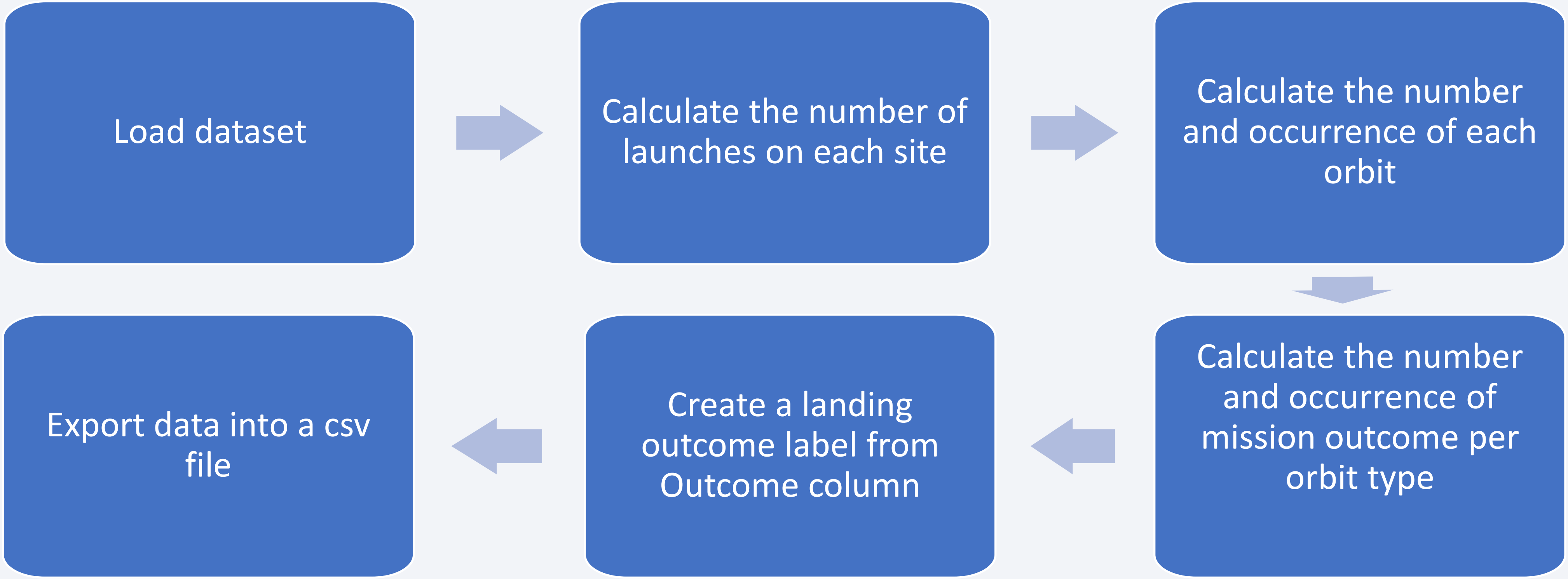
Export data into a csv file

Link to codes: https://github.com/hikarikk/Final-Presentation/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

- In the data set, there are several cases where the booster did not land successfully.
  - True Ocean, True RTLS, True ASDS means the booster was successfully landed.
  - False Ocean, False RTLS, False ASDS means the booster was unsuccessfully landed.
- This task converted the landing outcomes into Training Labels with 1 means the booster successfully landed and 0 means it was unsuccessful

Data Wrangling



Load dataset

Calculate the number of launches on each site

Calculate the number and occurrence of each orbit

Calculate the number and occurrence of mission outcome per orbit type

Create a landing outcome label from Outcome column

Export data into a csv file

Link to codes: https://github.com/hikarikk/Final-Presentation/blob/main/jupyter-labs-spacex-data_wrangling.ipynb

EDA with Data Visualization

- Scatter plots to discover relationship between variables
  - Flight Number vs. Payload Mass
  - Flight Number vs. Launch Site
  - Payload Mass vs. Launch Site
  - Flight Number vs. Orbit Type
  - Payload Mass vs. Orbit Type
- Bar chart to visualize the success rate of each orbit type
- Line plot to visualize the yearly trend of success rate

Link to codes: https://github.com/hikarikk/Final-Presentation/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- Perform SQL queries to understand the data
  1. Display the names of the unique launch sites in space mission
  2. Display 5 records when launch sites begin with the string "CCA"
  3. Display the total payload mass carried by boosters launched by NASA(CRS)
  4. Display average payload mass carried by booster version F9 v1.1
  5. List the date when the first successful landing outcome in ground pad was achieved
  6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
  7. List the total number of successful and failure mission outcomes.
  8. List the names of the booster_versions which have carried the maximum payload mass.
  9. 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.
  10. Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017in descending order.

Link to codes: https://github.com/hikarikk/Final-Presentation/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Mark all launch sites on a map
  - Add a highlighted circle area (circles, markers) for each site with a text label showing site name
- Mark the success/failed launches for each site
  - Add colored markers of successful (green) and failed (red) launches using marker cluster
- Calculate and visualize the distances between a launch site to its proximities
  - Create and add a marker for a selected proximity
  - Draw a polyline between a launch site and the selected point
- As success rate may depend on location and proximities of a launch site, those map objects are created to help understand factors that affect launch site selection by analyzing existing launch site locations.

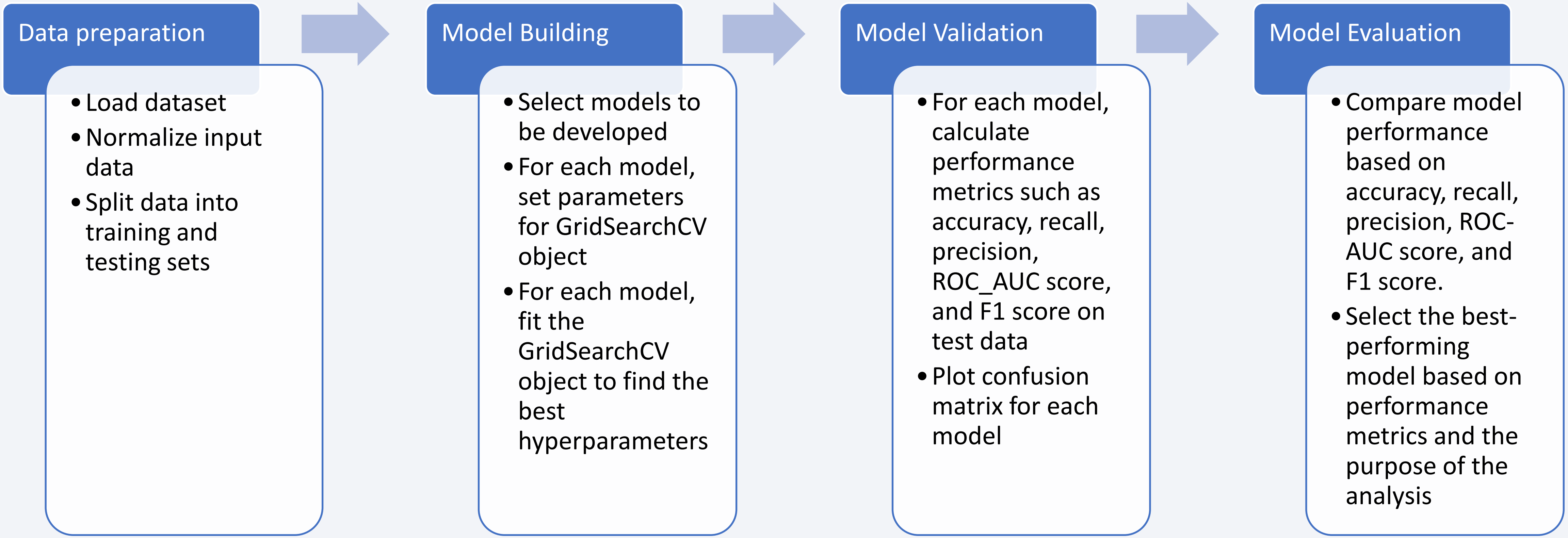
Link to codes: https://github.com/hikarikk/Final-Presentation/blob/main/jupyter-labs-launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- A dashboard application was created to perform interactive visual analytics on SpaceX launch data. Components of the dashboard included:
  - A dropdown list to select a launch site
  - A range slider to select payload mass
  - A pie chart to show the total counts of success/failures for all launch site, and the proportional of success/failure at a specific selected from the dropdown
  - A scatter chart to show the relationship between variables, e.g., launch success vs payload mass, launch success of different booster versions

Link to codes: https://github.com/hikarikk/Final-Presentation/blob/main/spacex_launch_dash.py

Predictive Analysis (Classification)



Data preparation

- Load dataset
- Normalize input data
- Split data into training and testing sets

Model Building

- Select models to be developed
- For each model, set parameters for GridSearchCV object
- For each model, fit the GridSearchCV object to find the best hyperparameters

Model Validation

- For each model, calculate performance metrics such as accuracy, recall, precision, ROC_AUC score, and F1 score on test data
- Plot confusion matrix for each model

Model Evaluation

- Compare model performance based on accuracy, recall, precision, ROC-AUC score, and F1 score.
- Select the best-performing model based on performance metrics and the purpose of the analysis

Link to codes: https://github.com/hikarikk/Final-Presentation/blob/main/jupyterlite-SpaceX_Machine_Learning_Prediction.ipynb

Results

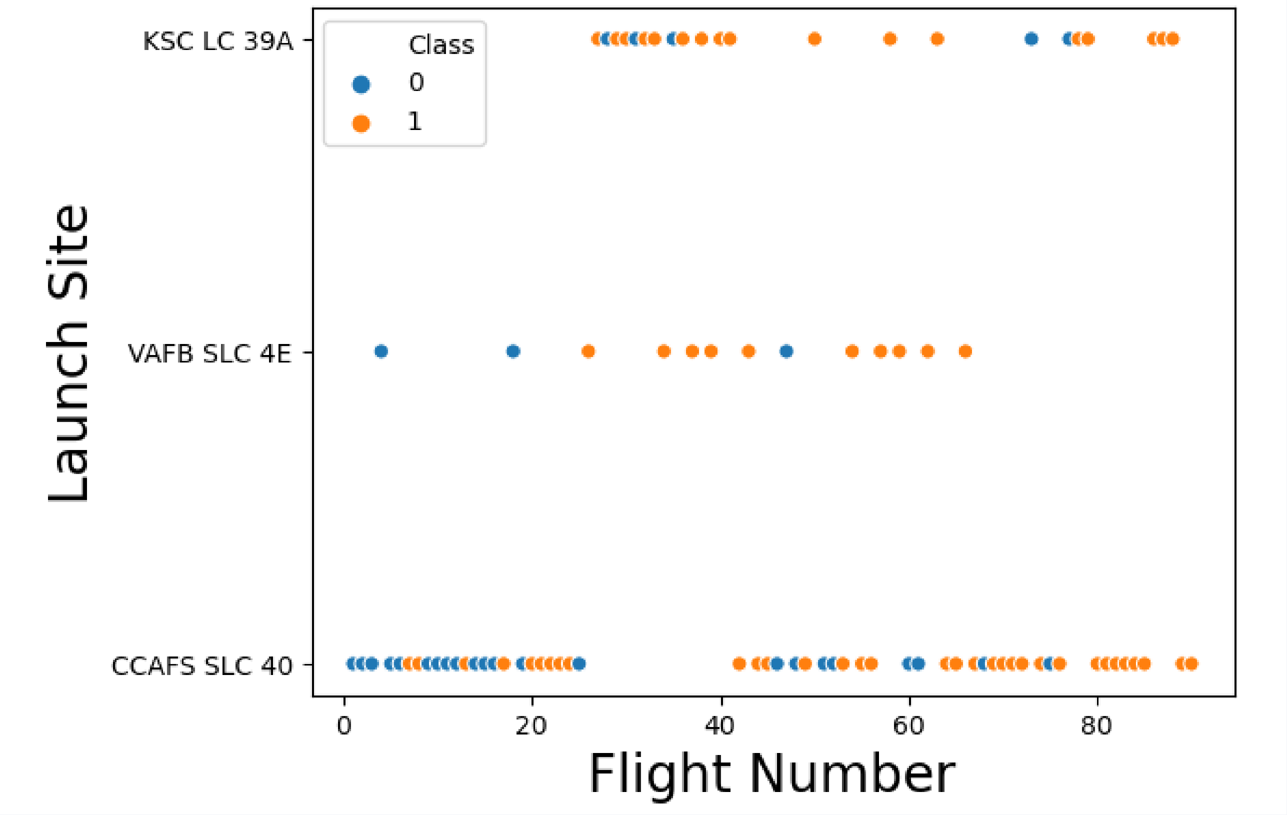
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

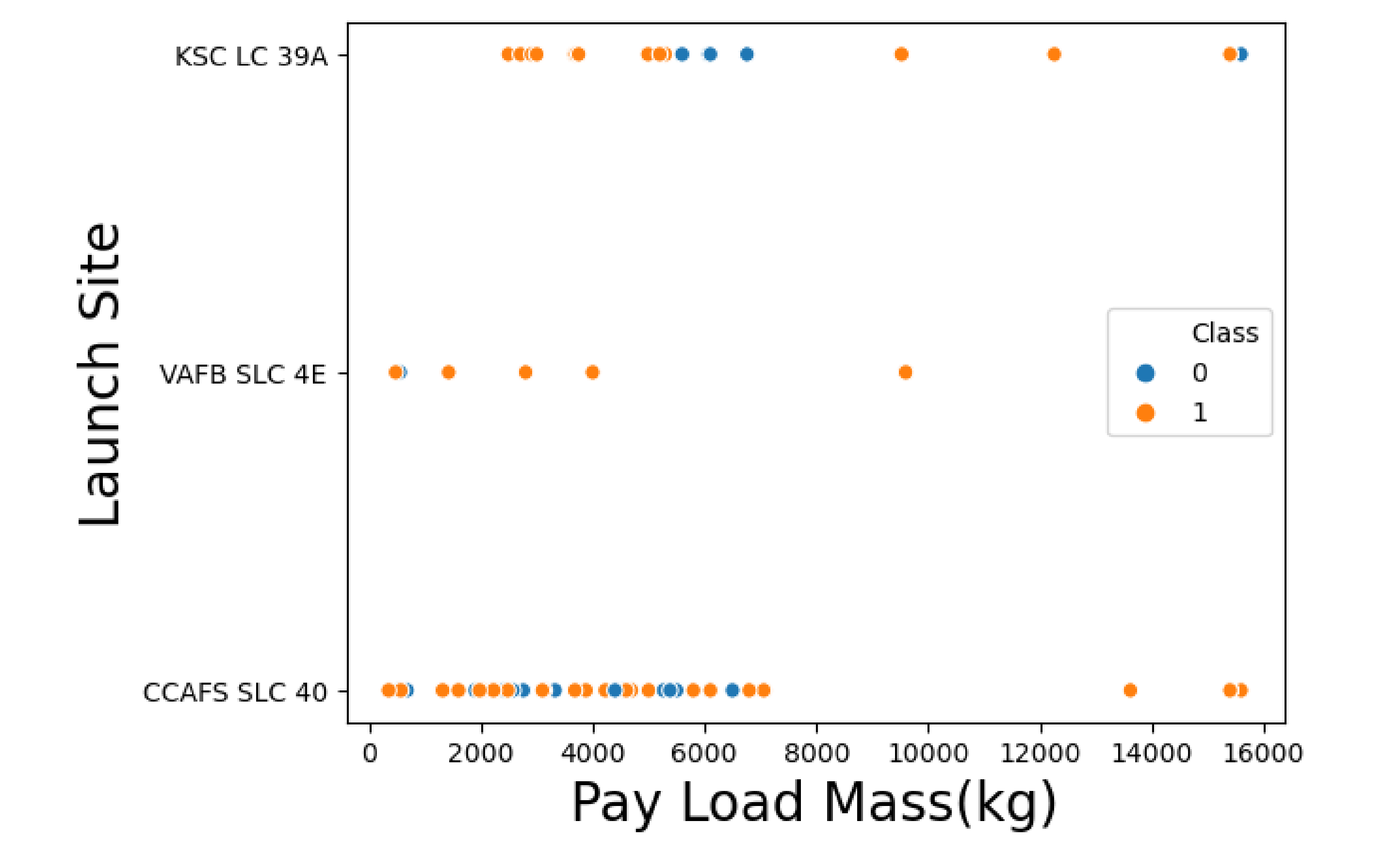
Flight Number vs. Launch Site

- CCAFS SLC 40 had more launches than the other two sites
- KSC LA39A and WAFB SLC 4E had higher success rate than CCAFS SLC 40
- As the fight number increased, success launches increased, that is, the most recent launches were more likely to be successful, probably due to knowledge learned from previous launches



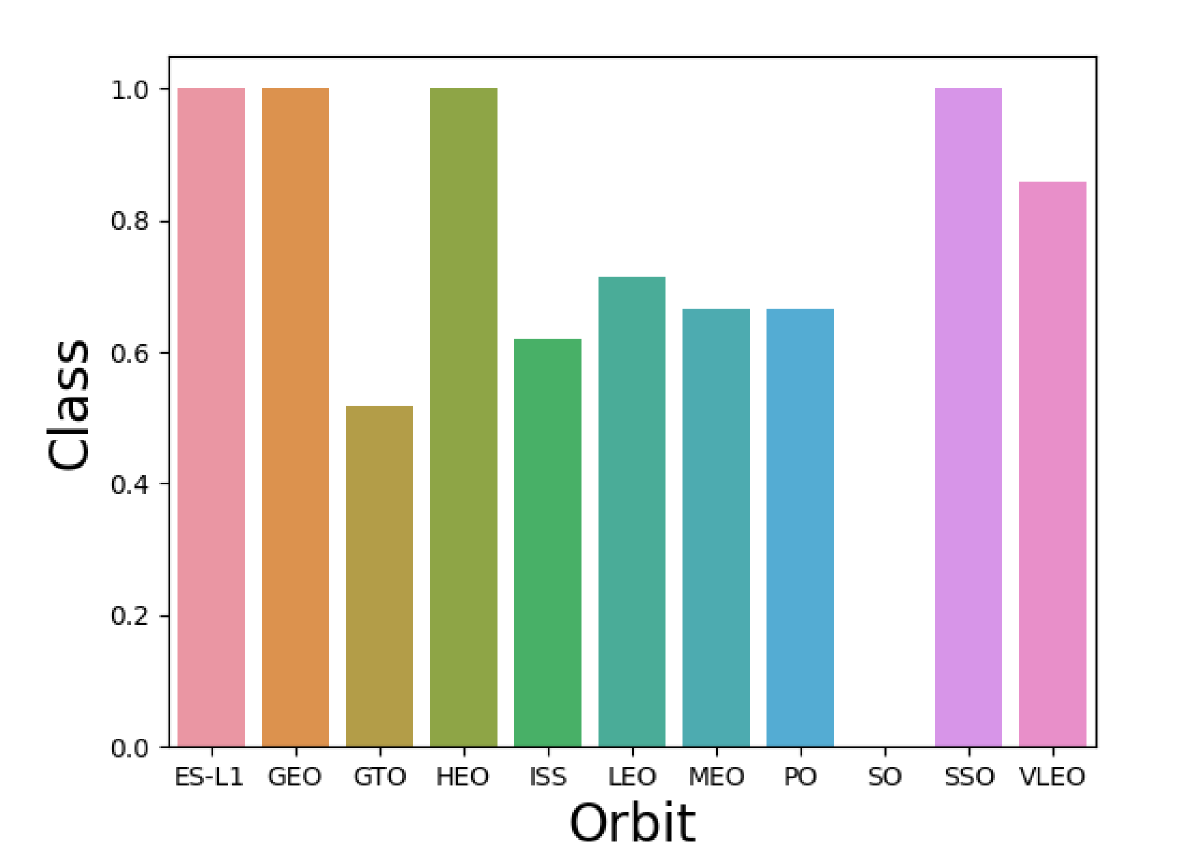
Payload vs. Launch Site

- There were no rockets launched for heavy payload mass (greater than 10,000 kg) at VAFB-SLC
- Success rate for heavy payload mass (greater than 10,000 kg) were higher, but the total number of launches were smaller than those with payload below 10,000 kg
- Launches with payload mass between 2,000 kg and 5,500 kg also had relatively higher success rate
- KSC LC 39A had 100% success rate for payload mass less than 5,500 kg



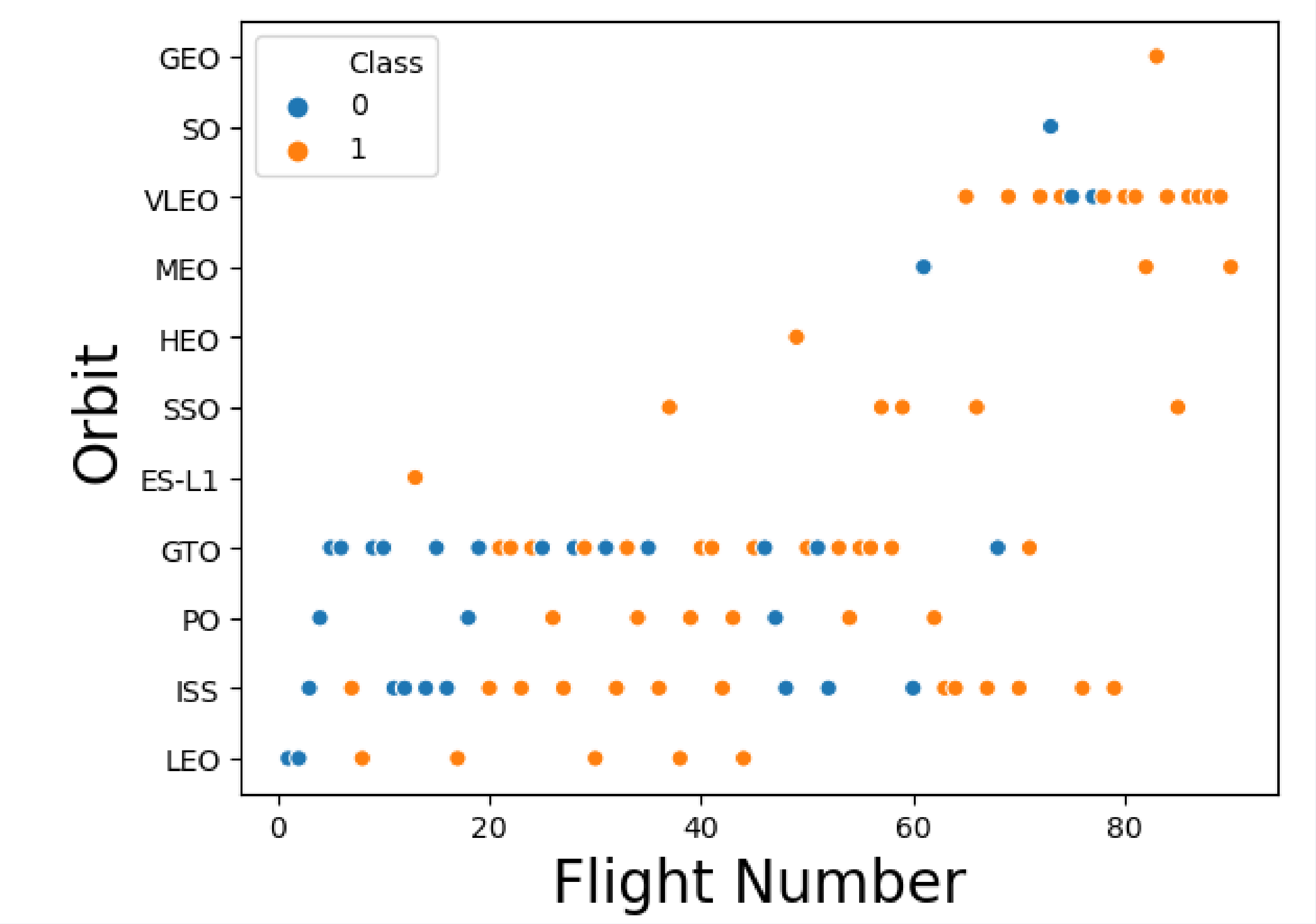
Success Rate vs. Orbit Type

- The success rate was 100% for ES-L1, GEO, HEO, and SSO
- The success rate was 0 for SO
- Success rates of ISS, LEO, MEO, and PO were similar, between 60% - 70%
- Orbit GTO had the second lowest success rate, about 50%



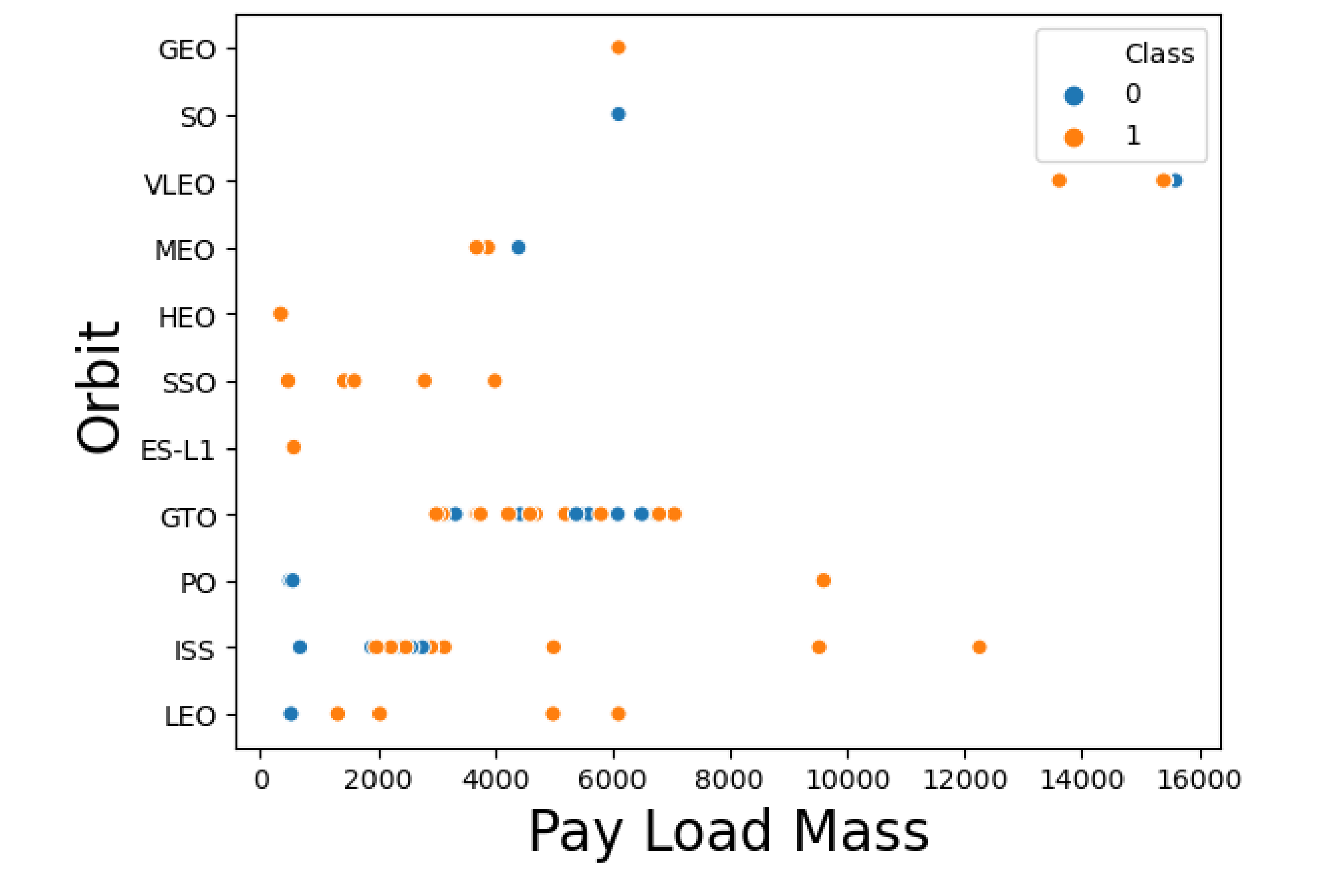
Flight Number vs. Orbit Type

- For LEO, the success rate of was increased with the increase of flight number
- For GTO, no clear relationship between orbit and flight number
- The higher success rate of SSO, HEO, and VLEO may be related to the knowledge learned from previous launches for all other orbits
- GTO had the greatest number of launches



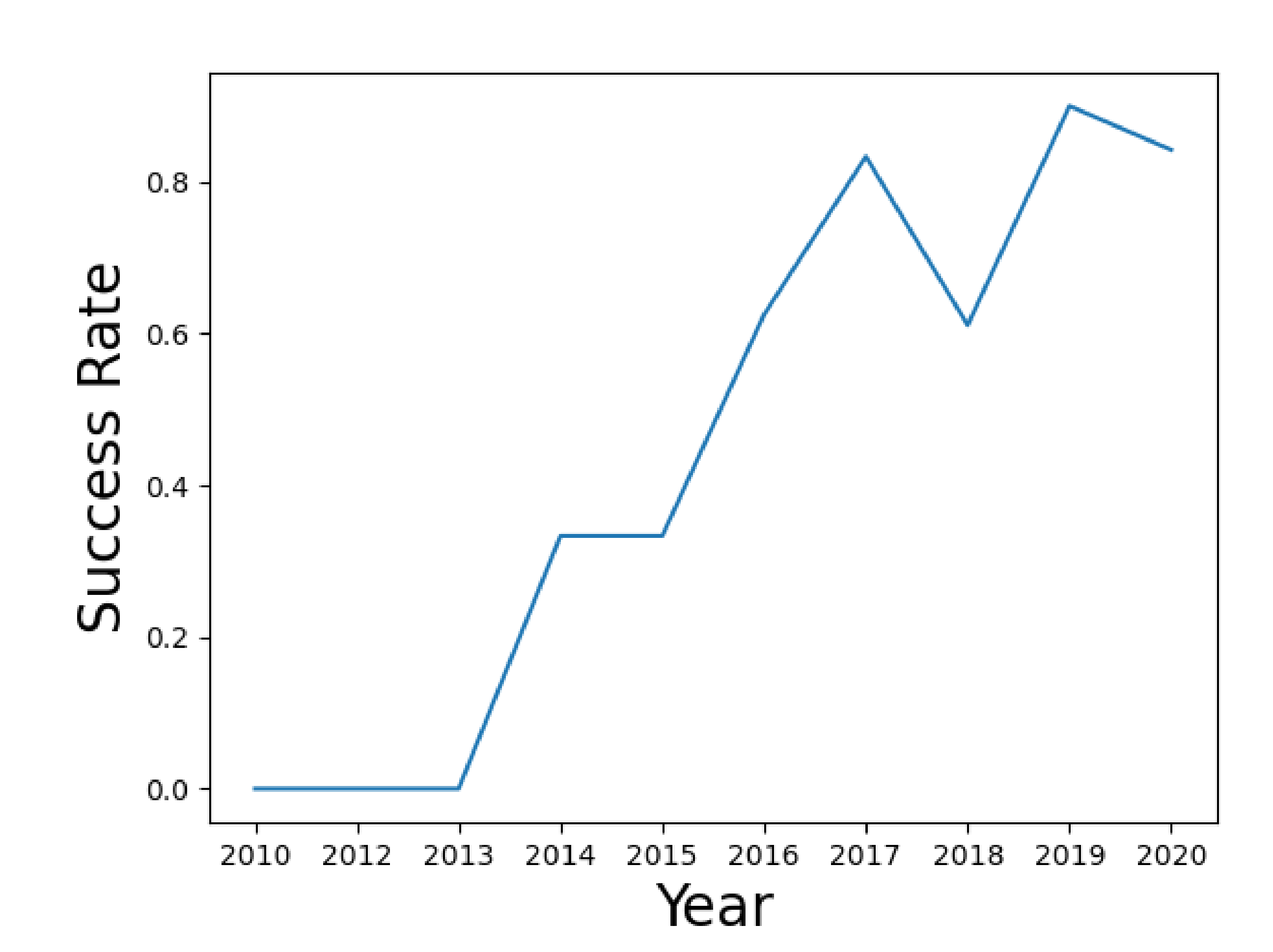
Payload vs. Orbit Type

- LEO, PO, and ISS had higher success rate for heavy payload
- For GTO, it was hard to distinguish successful and unsuccessful landings using payload



Launch Success Yearly Trend

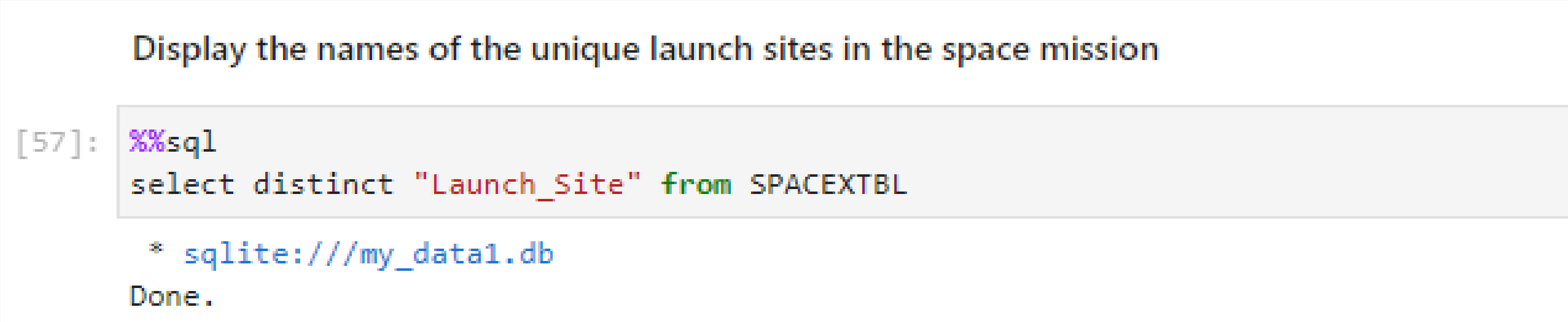
- There was an increasing trend of success since 2013 to 2020, as the technology advance and knowledge learned from previous launches
- Success rate was around 80% in recent years
- There was a drop of 20% in 2018 compared to 2017



All Launch Site Names

- Use “distinct” function to find unique values of launch site names

SQL query

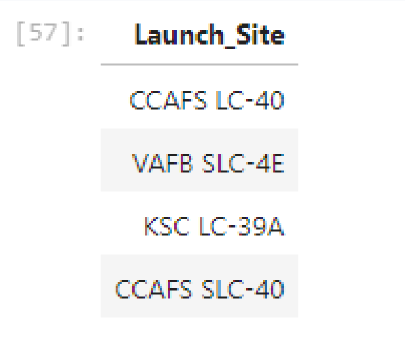


Display the names of the unique launch sites in the space mission

```
[57]: %%sql
      select distinct "Launch_Site" from SPACEXTBL

 * sqlite:///my_data1.db
Done.
```

Result



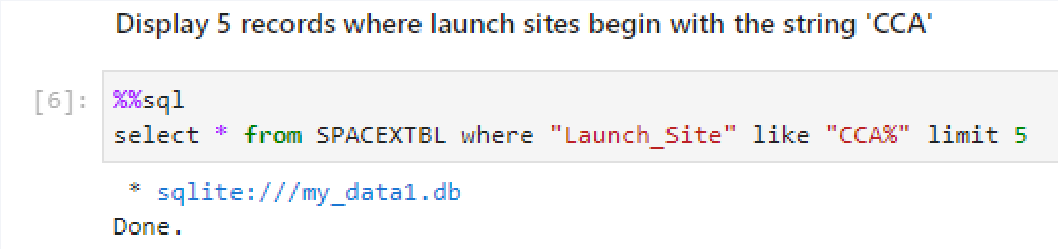
[57]:

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

Launch Site Names Begin with 'CCA'

- Use “like” with “where” clause to filter launch sites whose name beginning with ‘CCA’

SQL query

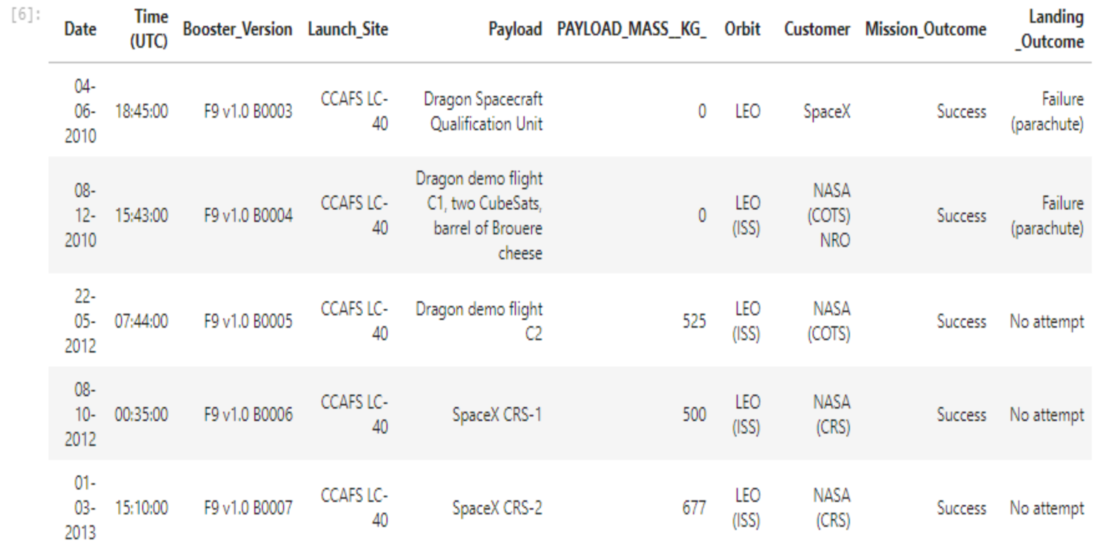


Display 5 records where launch sites begin with the string 'CCA'

```
[6]: %%sql
     select * from SPACEXTBL where "Launch_Site" like "CCA%" limit 5

      * sqlite:///my_data1.db
     Done.
```

Result



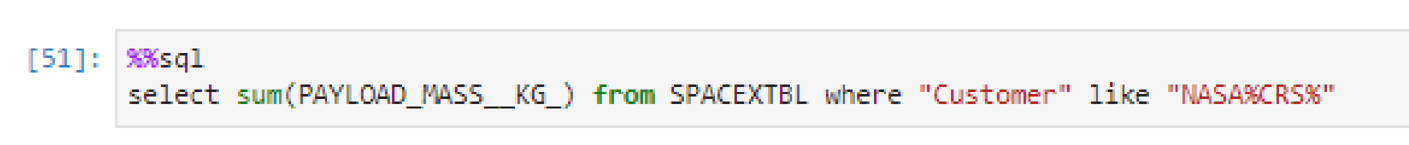
[6]:

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

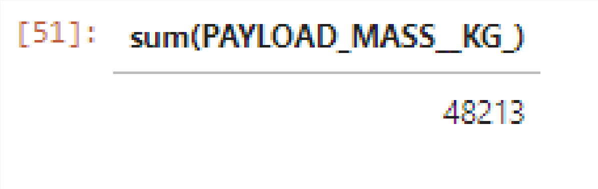
- Calculate the total payload carried by boosters from NASA (CRS)

SQL query



```
[51]: %%sql
      select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where "Customer" like "NASA%CRS%"
```

Result

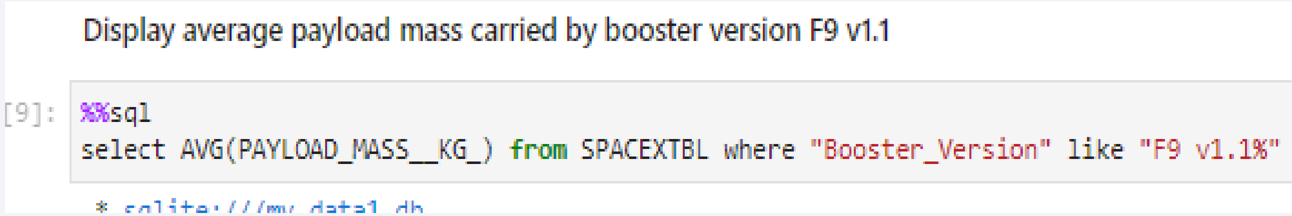


```
[51]:  sum(PAYLOAD_MASS__KG_)
                        48213
```

Average Payload Mass by F9 v1.1

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

SQL query

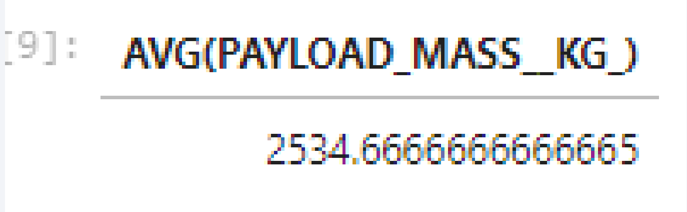


Display average payload mass carried by booster version F9 v1.1

```
[9]: %%sql
     select AVG(PAYLOAD_MASS__KG_) from SPACEXTBL where "Booster_Version" like "F9 v1.1%"

      * sqlite:///my_data1.db
```

Result

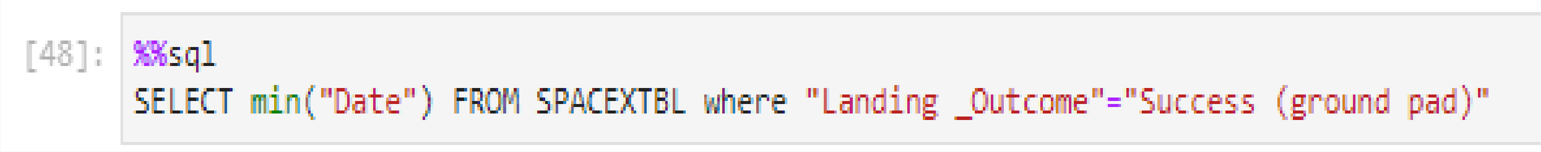


```
[9]:  AVG(PAYLOAD_MASS__KG_)
      2534.6666666666665
```

First Successful Ground Landing Date

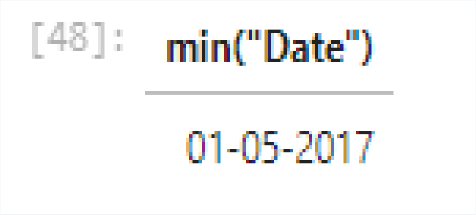
- Find the dates of the first successful landing outcome on ground pad
- Use the “min” function to find the date of firs successful landing

SQL query



```
[48]: %%sql
      SELECT min("Date") FROM SPACEXTBL where "Landing _Outcome"="Success (ground pad)"
```

Result

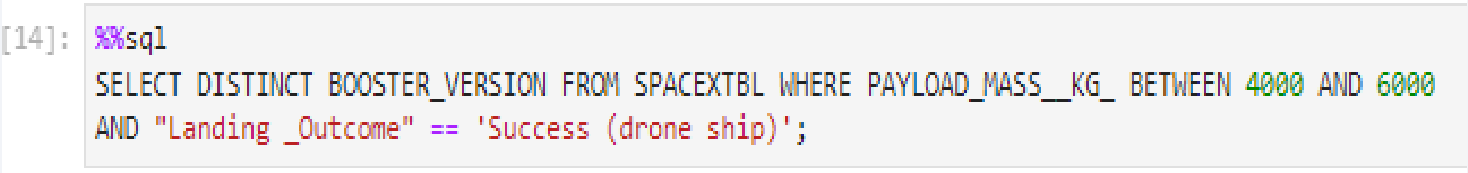


```
[48]:  min("Date")
       01-05-2017
```

Successful Drone Ship Landing with Payload between 4,000 and 6,000 kg

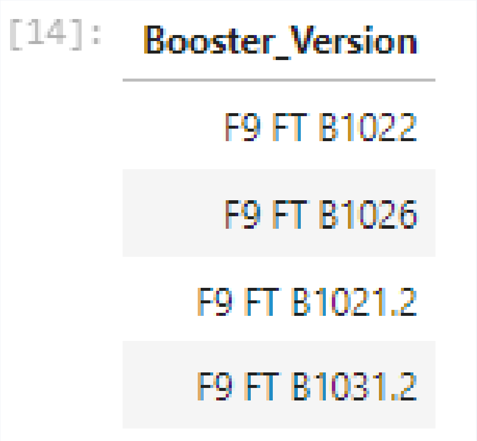
- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4,000 but less than 6,000 kg
- Use “where” and “and” to filter the data

SQL query



```
[14]: %%sql
      SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
      AND "Landing _Outcome" == 'Success (drone ship)';
```

Result



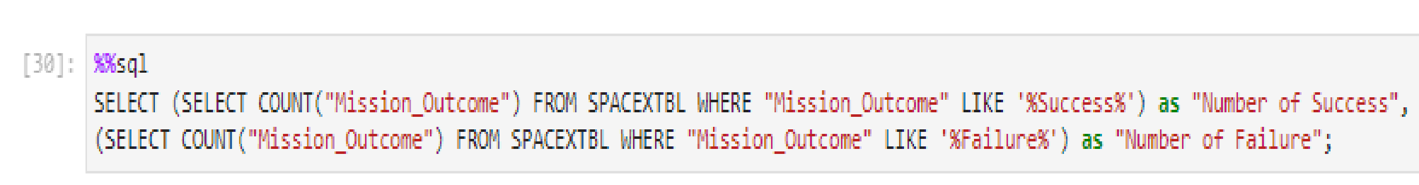
[14]:

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

Total Number of Successful and Failure Mission Outcomes

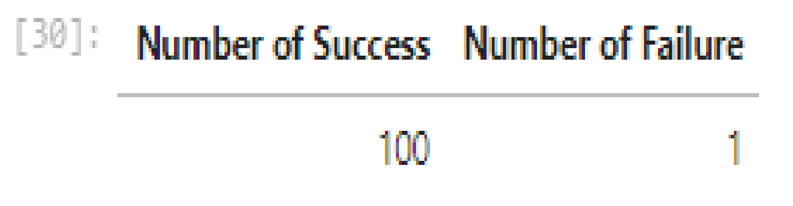
- Use subquery to calculate the numbers of successful and failure mission outcomes

SQL query



```
[30]: %%sql
SELECT (SELECT COUNT("Mission_Outcome") FROM SPACEXTBL WHERE "Mission_Outcome" LIKE '%Success%') as "Number of Success",
(SELECT COUNT("Mission_Outcome") FROM SPACEXTBL WHERE "Mission_Outcome" LIKE '%Failure%') as "Number of Failure";
```

Result



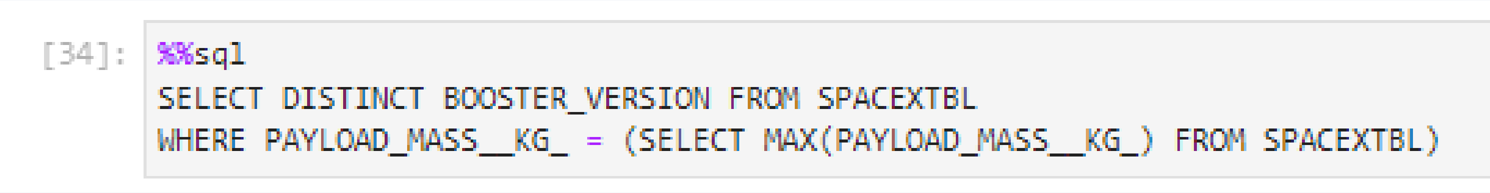
[30]:

Number of Success	Number of Failure
100	1

Boosters Carried Maximum Payload

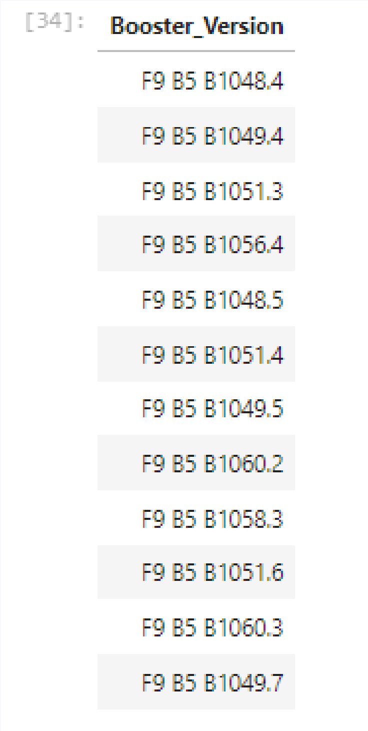
- Use a subquery to filter data and output the names of the booster which have carried the maximum payload mass

SQL query



```
[34]: %%sql
SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

Result



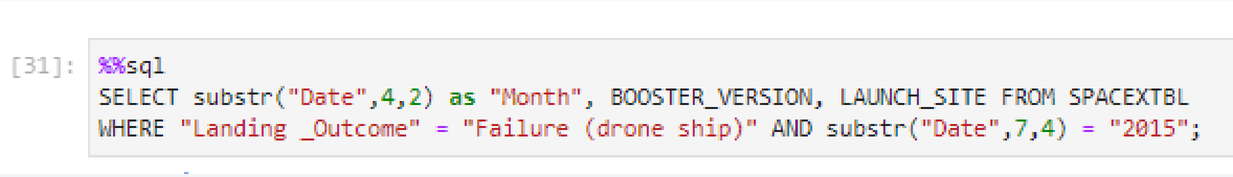
[34]:

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

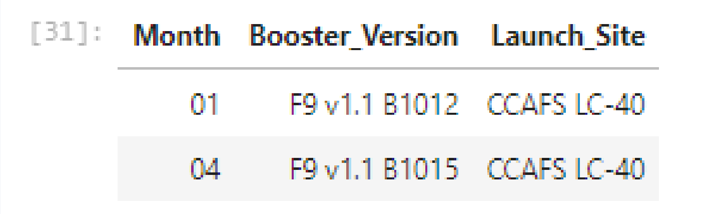
- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Because SQLLite does not support month names, use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

SQL query



```
[31]: %%sql
      SELECT substr("Date",4,2) as "Month", BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL
      WHERE "Landing _Outcome" = "Failure (drone ship)" AND substr("Date",7,4) = "2015";
```

Result



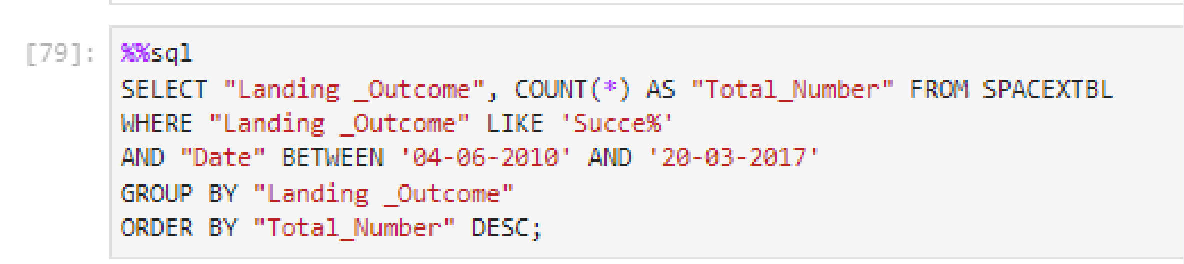
[31]:

Month	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

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

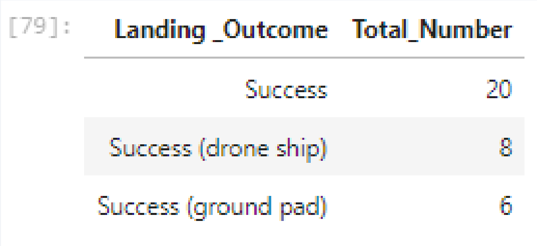
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20 in descending order
- Use “GROUP BY” clause groups results by landing outcome, and use “ORDER BY” and “COUNT DESC” to show results in decreasing order

SQL query



```
[79]: %%sql
      SELECT "Landing _Outcome", COUNT(*) AS "Total_Number" FROM SPACEXTBL
      WHERE "Landing _Outcome" LIKE 'Succe%'
      AND "Date" BETWEEN '04-06-2010' AND '20-03-2017'
      GROUP BY "Landing _Outcome"
      ORDER BY "Total_Number" DESC;
```

Result



[79]:

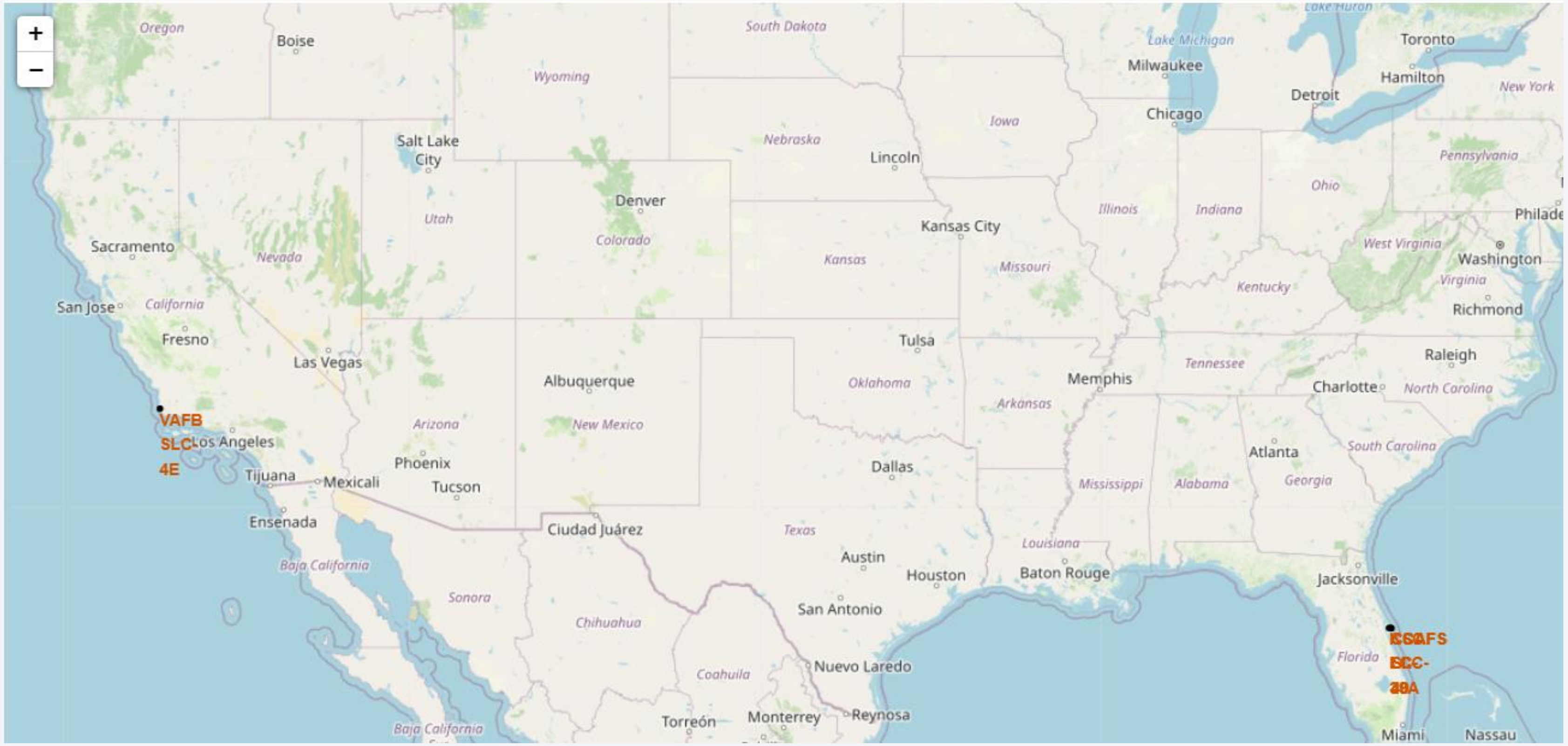
Landing _Outcome	Total_Number
Success	20
Success (drone ship)	8
Success (ground pad)	6

Section 3

Launch Sites Proximities Analysis

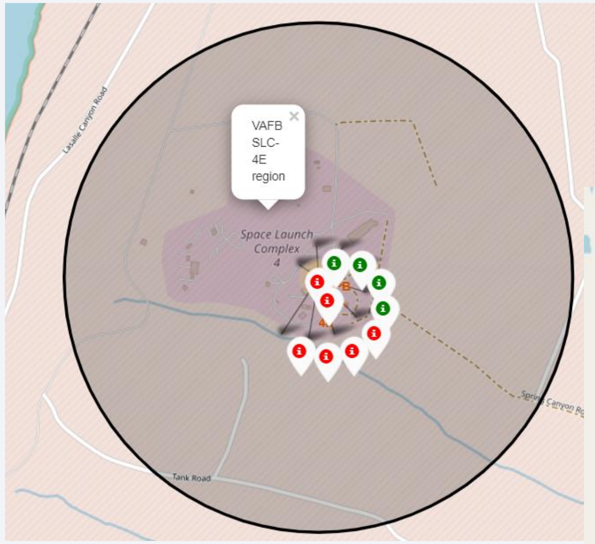
Folium Map – Launch Site Locations

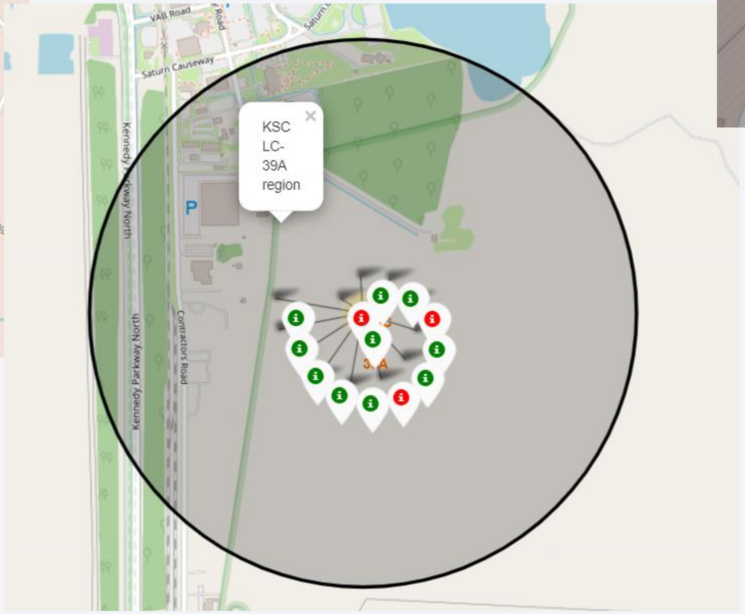
- Space X launch sites are near the coastline
- One site in south California and the others in Florida

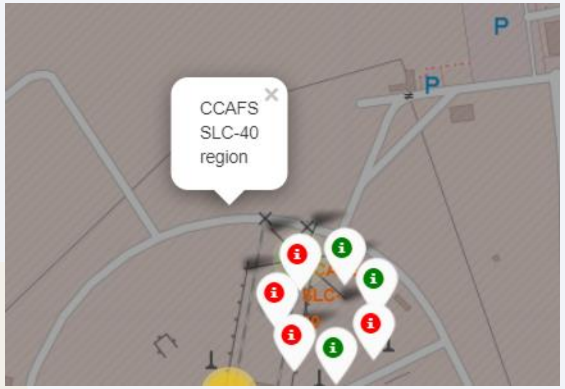


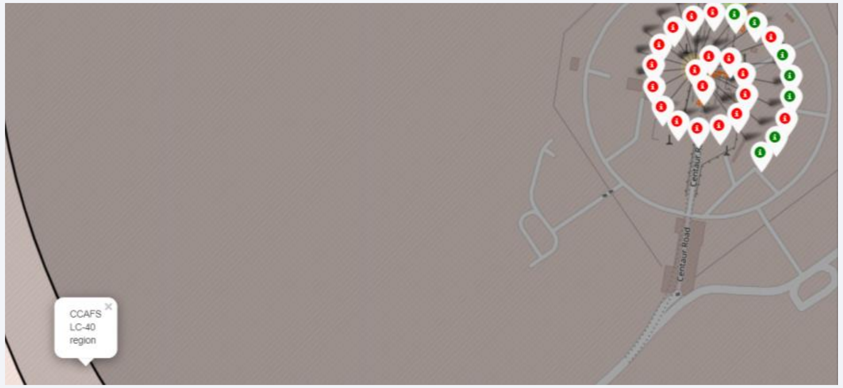
Folium Map – Colored Labeled Markers for Launch Outcomes

- Green marker represents success and red marker represents failure
- KSC LC 39A had a higher success rate



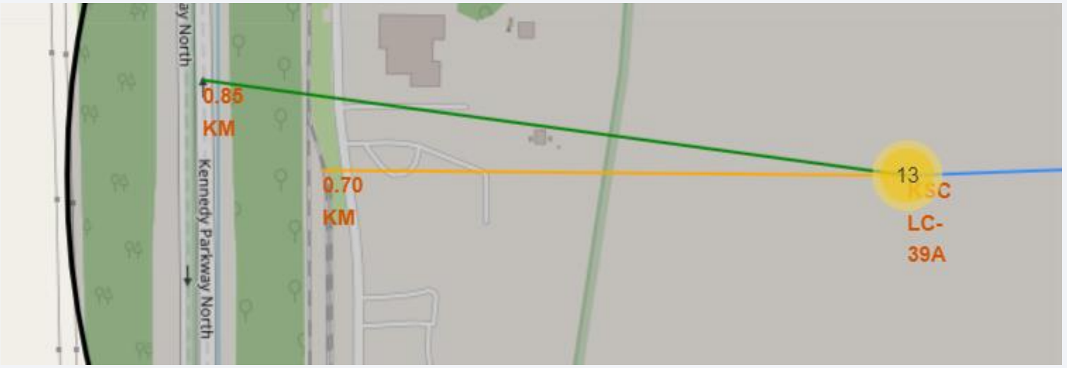


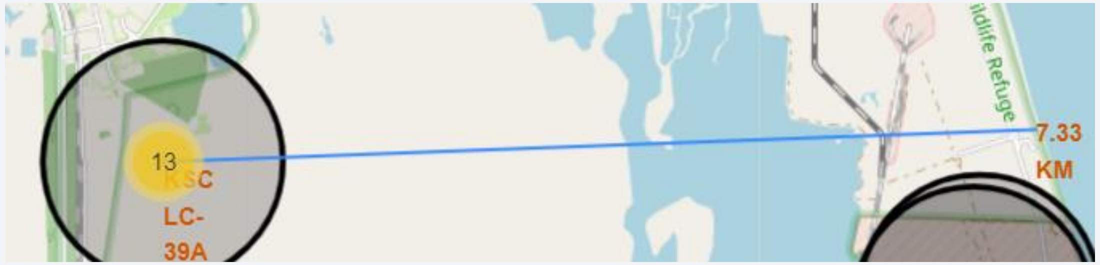


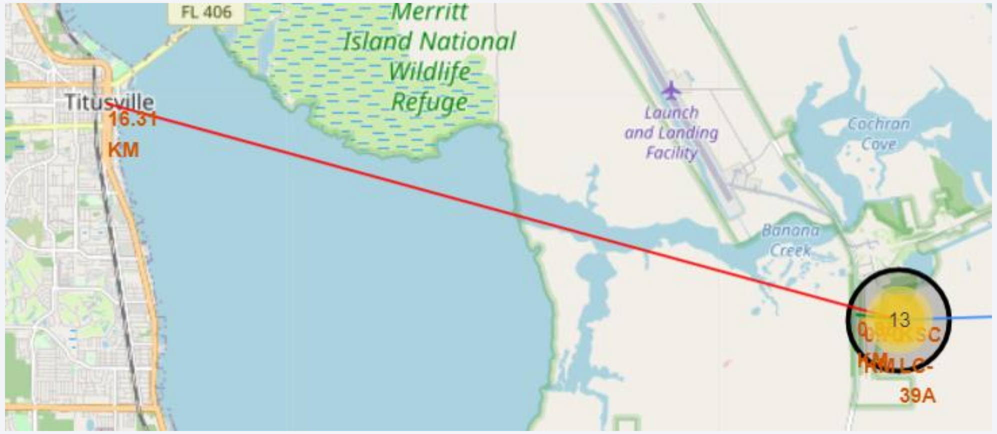


Folium Map – Distance between KSC LC-39A and its proximities

- KSC LC-39A is in close proximity to coastline, railway, and highway
- The closest city, Titusville, is about 16 km away, as the launch site should be far away from populated areas for public safety purpose





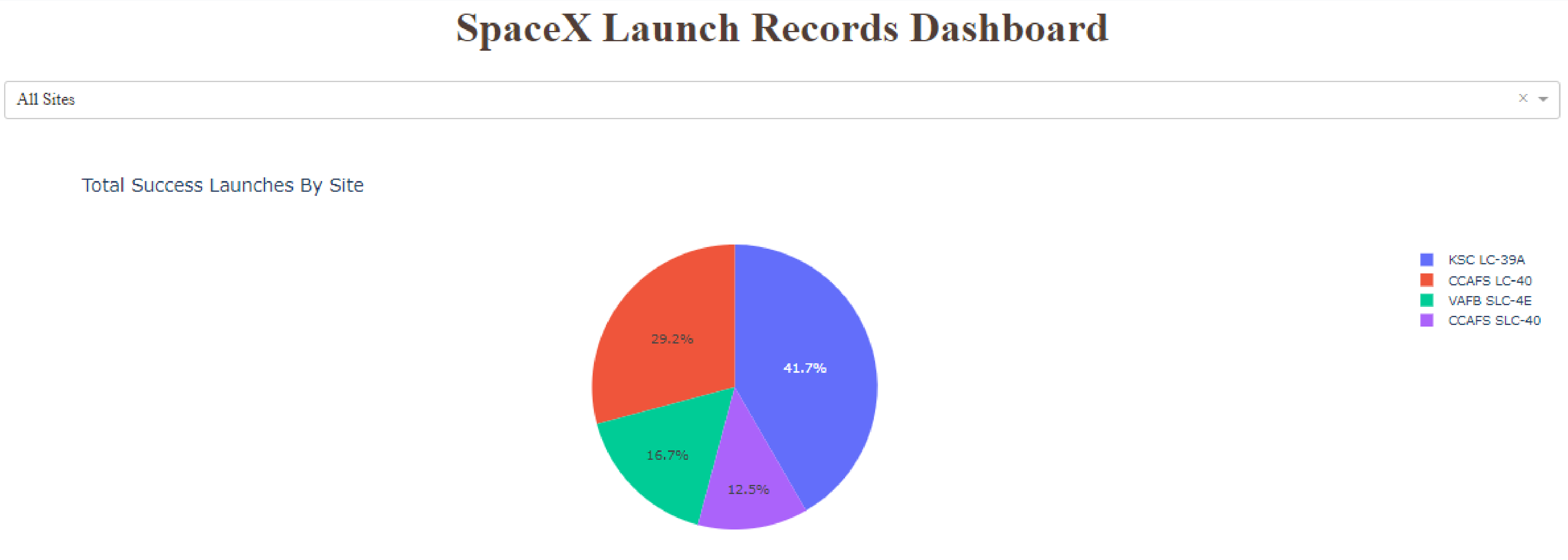

FL 406
Merritt Island National Wildlife Refuge
Titusville
16.31 KM
Launch and Landing Facility
Cochran Cove
Banana Creek
13
KSC LC-39A

Section 4

Build a Dashboard with Plotly Dash

Dashboard – Total Successes by Site

- Among all four launch sites, KSC LC-39A had the most successful launches (41.7% of all successful launches), followed by CCAFS LC-40 (29.2%)

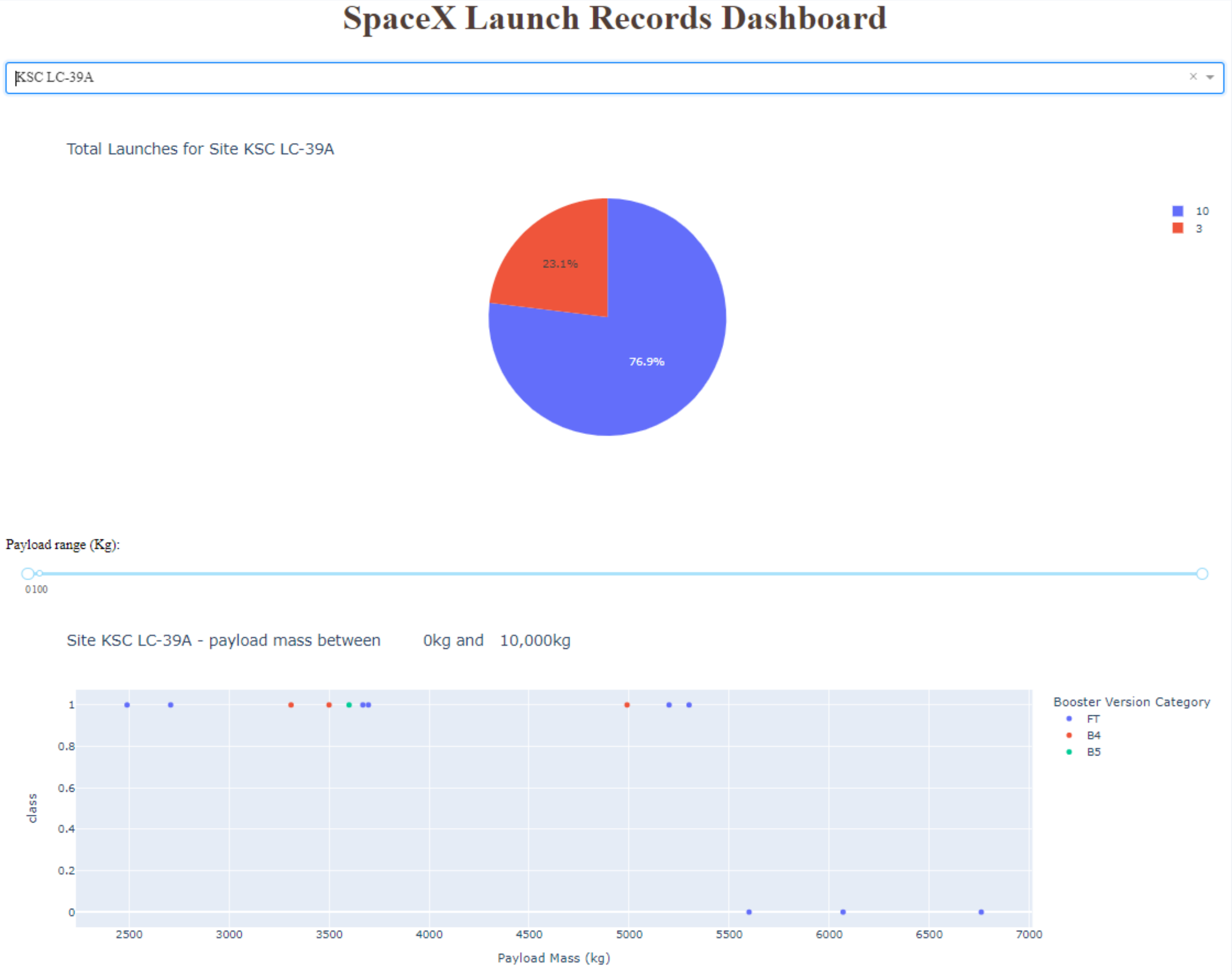


SpaceX Launch Records Dashboard

All Sites

Dashboard – Launches at KSC LC-39A

- The success rate was 76.9% at KSC LC-39A
- The payload mass was less than 7,000 kg for all launches at this site
- The payload mass was between 2,000 kg and 5,500 kg for all successful launches
- The Falcon 9 booster versions launched were FT (9 launches), B4 (3 launches), and B5 (1 launch)



SpaceX Launch Records Dashboard

KSC LC-39A

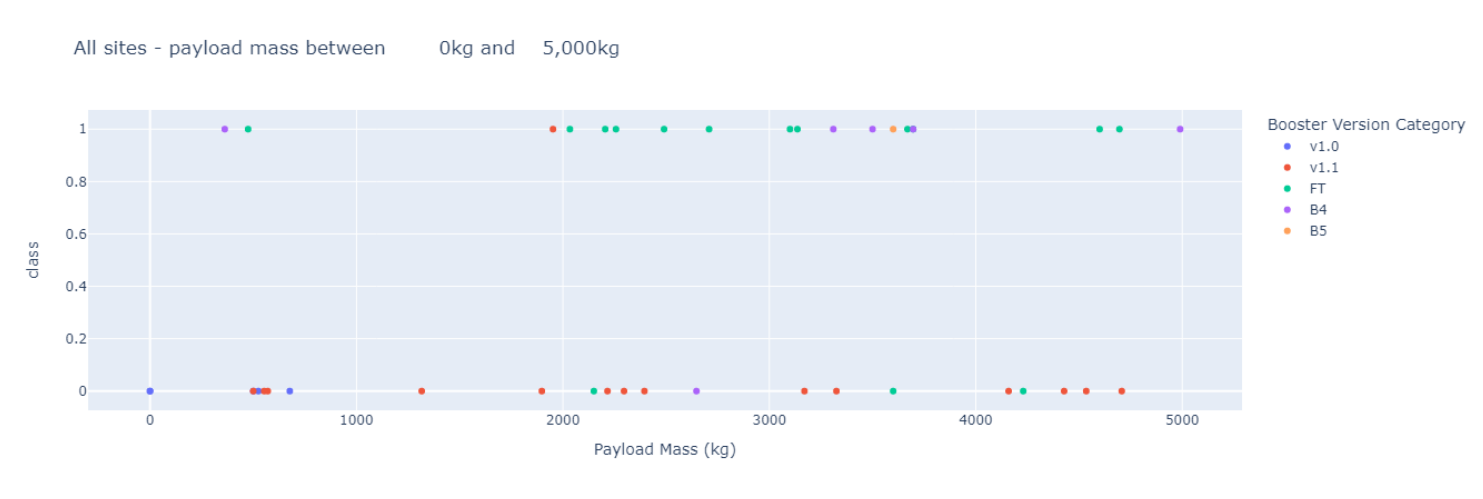
Total Launches for Site KSC LC-39A

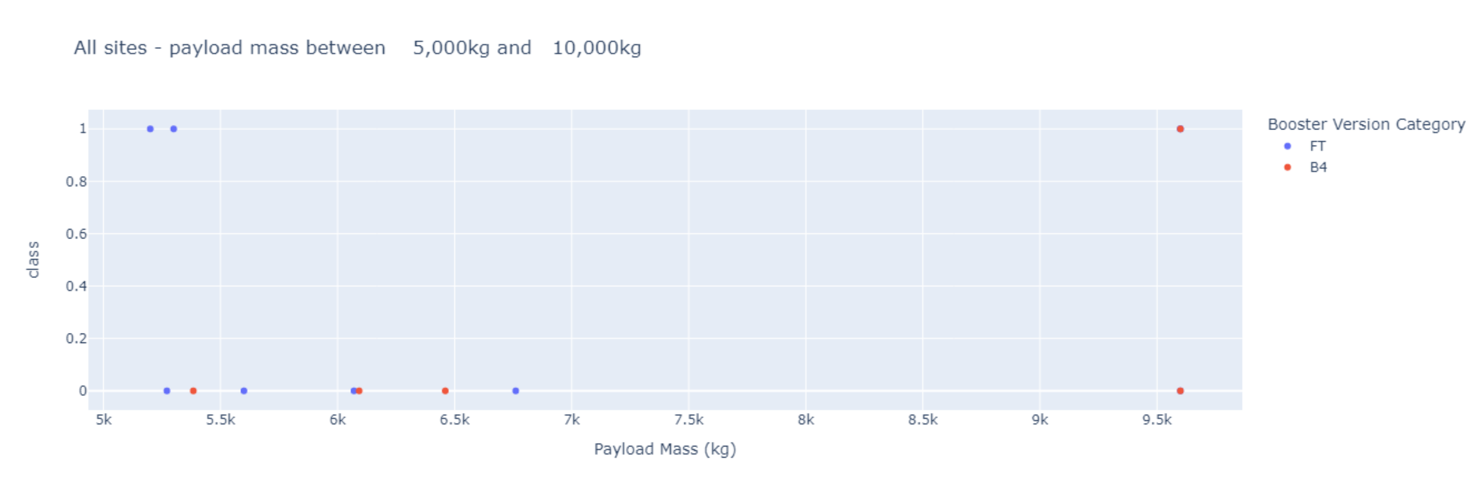
Payload range (Kg):

Site KSC LC-39A - payload mass between 0kg and 10,000kg

Dashboard – Payload vs Launch Outcome for All Sites

- Payload between 3,000 kg to 4,000 kg had the highest success rate (7 out of 10)
- Heavier payload greater than 5,500 kg had low success rate
- Falcon 9 booster version FT had the greatest number of successful launches, and success rate was 67%



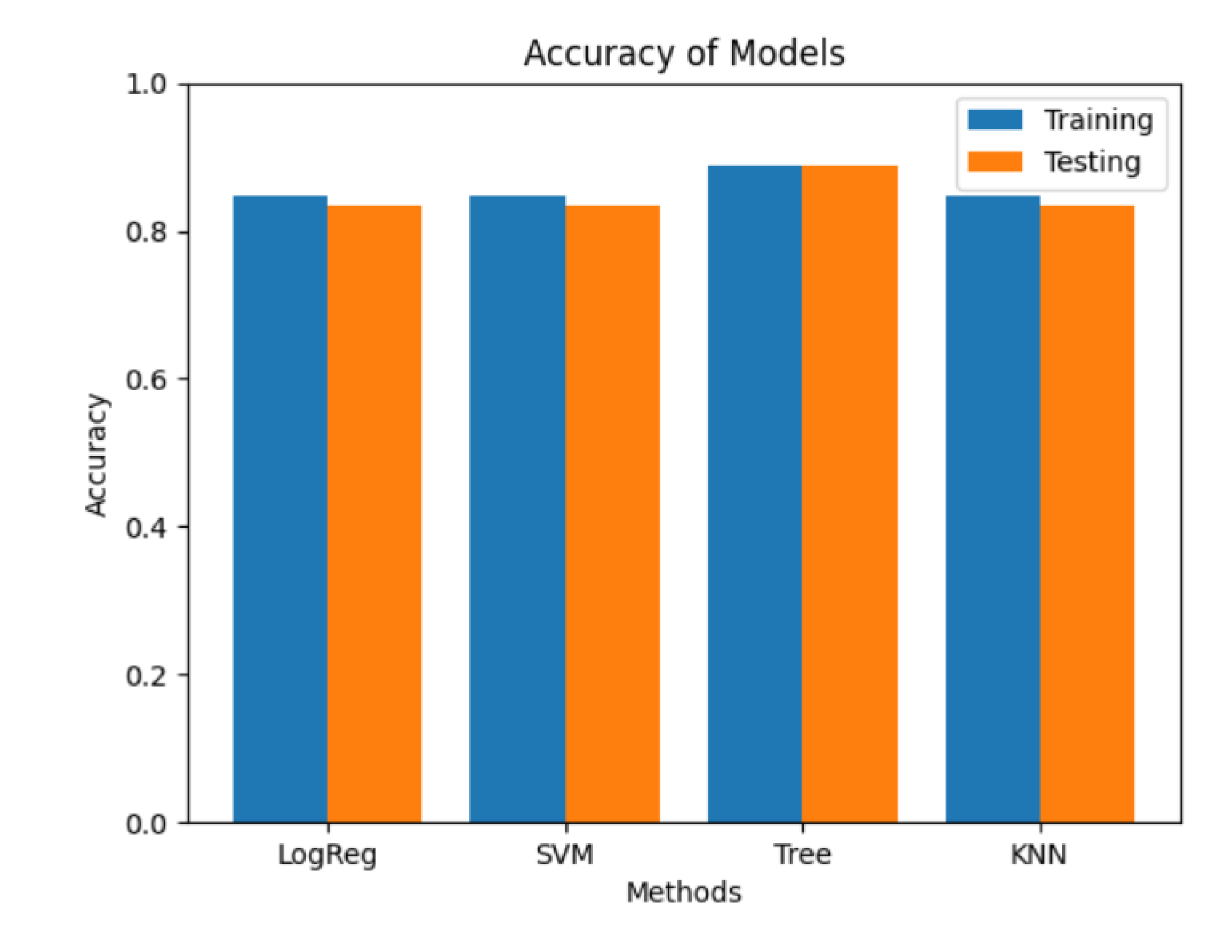


Section 5

Predictive Analysis (Classification)

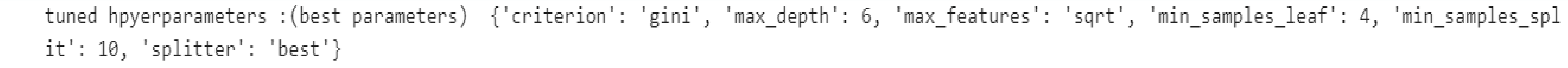
Classification Accuracy

- SVM, Classification Trees and Logistic Regression, Classification Trees, Support Vector Machine(SVM) and K-nearest neighbors (KNN) models were built and compared
- Decision Trees model had slighter higher accuracy than other models
- Logistic Regression, SVM, and KNN models had the same accuracy (0.83) on the testing data



Classification Accuracy

- Best hyperparameters of Decision Trees model



```
tuned hpyerparameters :(best parameters)  {'criterion': 'gini', 'max_depth': 6, 'max_features': 'sqrt', 'min_samples_leaf': 4, 'min_samples_spl
it': 10, 'splitter': 'best'}
```

- Decision Trees model performed the best based on accuracy, precision, and ROC-AUC score, but the differences compared to other models were not very significant

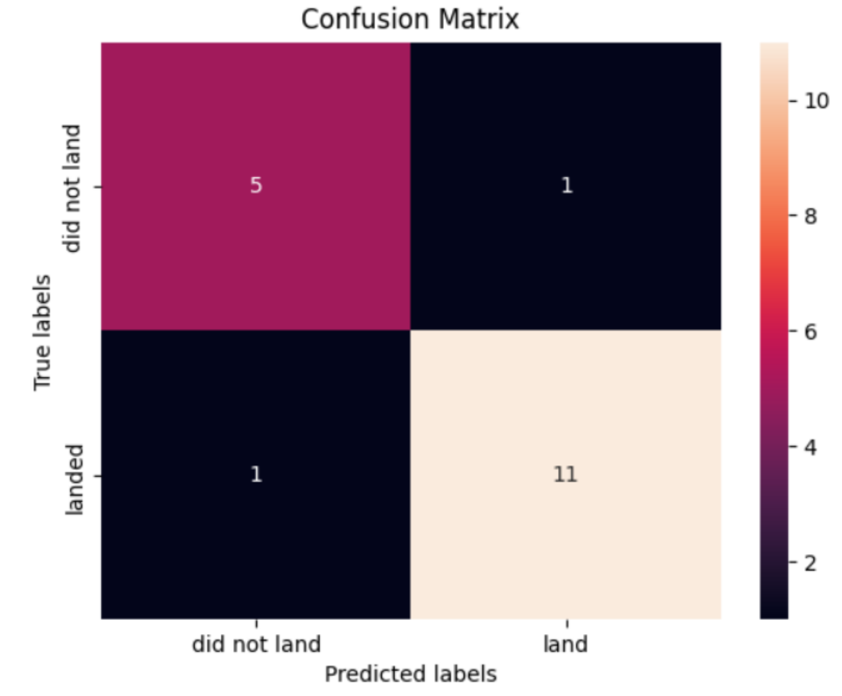
- Decision Trees model is preferred though the recall is slighter lower; in this study, precision is a better performance metric than recall because avoiding false positive is more important for the purpose of the business application

Model	Accuracy	Precision	F1-score	Recall	ROC-AUC
LogReg	0.833	0.800	0.889	1.000	0.750
SVM	0.833	0.800	0.889	1.000	0.750
Tree	0.889	0.917	0.917	0.917	0.875
KNN	0.833	0.800	0.889	1.000	0.750

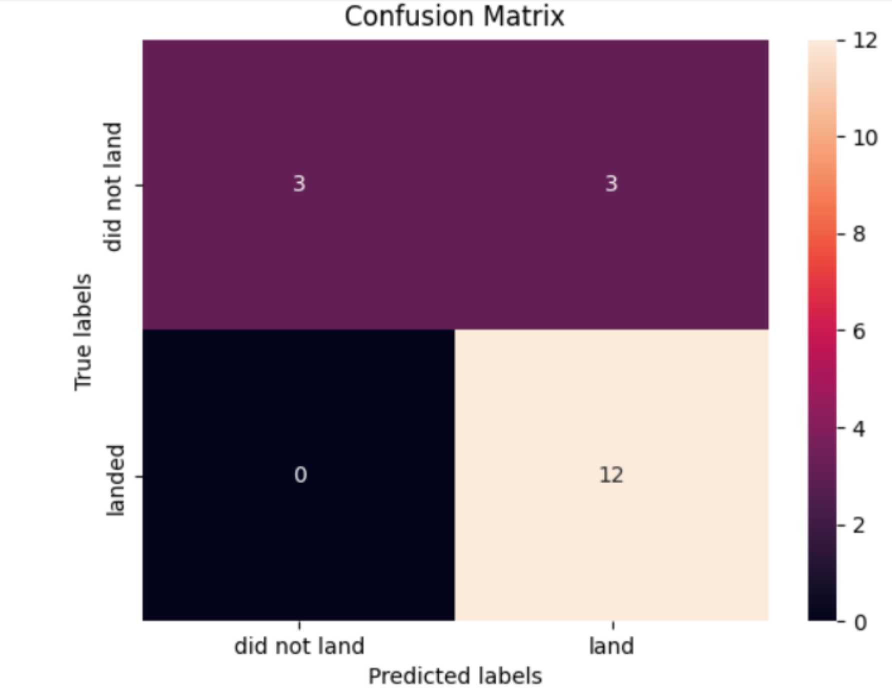
Confusion Matrix

- False positive was an issues for all models, but Decision Trees model had less false positives
- The matrices of Logistic Regression, SVM, and KNN were identical

Decision Trees



Logistic Regression, SVM, and KNN



Conclusions

- Several factors, such as launch site, payload, orbit, and especially the number of previous launches, may affect the launch outcomes
- The success rate was 100% for orbits GEO, HEO, SSO, and ES-L1, but the number of total launches at these orbits were relatively less GTO and LSS
- Most of the heavy payload (>10,000kg) launches were successful. Depending on the orbit, the payload mass can affect the success rate as some orbits better fit for a light or heavy payload mass
- For all Falcon 9 launches with payload mass below10,000 kg, the launches with payload between 2,000 and 5,500 kg were more likely to succeed, with payload between 3,000 and 4,000 kg had the highest success rate
- The success rate increased over the years since 2013, with a slight drop in 2018; the success rate was about 80% in recent years
- All launch sites are in very close proximity to coastline. KSC LC-39A had the highest success rate of 76.9%. With this data set, it is difficult to explain why this site had more success launches than other sites, and more information (e.g., meteorological data) is needed for further study
- For this data set, Decision Trees model performed best at predicting the launching outcome. Future study with more relevant variables could be conducted to improve accuracy and reduce false positives

Appendix

- Data sets , Python code snippets, and Notebook outputs are available on GitHub:

  https://github.com/hikarikk/Final-Presentation.git

Thank you!