IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Applied Data Science Capstone Project-
Predict Falcon 9 Rocket Launch

Xi Geng
Nov 27, 2021

Outline

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

Executive Summary

- Summary of methodologies
- Data collection
- Data wrangling
- Exploratory data analysis (EDA) using SQL
- EDA using data visualization
- Build an Interactive Map with Folium
- Build an interactive dashboard with Plotly Dash
- Predictive analysis using classification algorithm

- Summary of all results
- Insights drawn from EDA
- Understand dataset using SQL
- Launch sites proximities analysis results
- Interactive analytics of launches record on dashboard
- Predictive analysis results

Introduction

- Project background and context

The objective is to predict if the Falcon 9 first stage will land successfully. SpaceX 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 can be used if an alternate company wants to bid against SpaceX for a rocket launch.

- Problems you want to find answers

❑ How to collect information of detailed launch records?

❑ What kind of methodologies will be used to conduct data cleaning and waggling?

❑ What are the key features to influence if the rocket will land successfully?

❑ How to perform visualized predictive analysis?

❑ Which classification model perform best to predict the launch outcomes?

❑ What conditions does SpaceX have to achieve to get the best results and ensure the best rocket success landing rate.

Methodology- Executive Summary

- Data collection methodology:
  - SpaceX API
  - Web Scrapping from Wikipedia using BeautifulSoup
- Perform data wrangling
  - Replace the missing data with mean value
  - One hot encoding the categorical data
  - Standardization of numerical 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
  - Split train and test dataset
  - Use train data to train classification models (Logistic regression, SVM, Tree, KNN)
  - Use GridSearchCV to tune parameter
  - Evaluate the model based on the accuracy score

Data Collection

- SpaceX launch data is gathered from the SpaceX REST API.

❑ This API contains data about detailed rocket launch information including date, payload, launch and landing specifications, orbit, customers, outcomes, etc .

❑ Requests are sent to receive relevant launch information in the format of Jason.

❑ The Jason data is decoded and processed to a structured dataframe and exported to a csv file for further feature engineering and data wrangling and data analysis.

- Falcon 9 Launch data is also obtained through web scraping of Wikipedia using BeautifulSoup.

❑ Wikipedia contains up-to-date information with respect of Falcon 9 launches.

❑ Requests are sent to receive the data from html

❑ Beautiful soup can extract and parse the obtained data

❑ The extracted data can be exported to a csv file for further data wrangling.

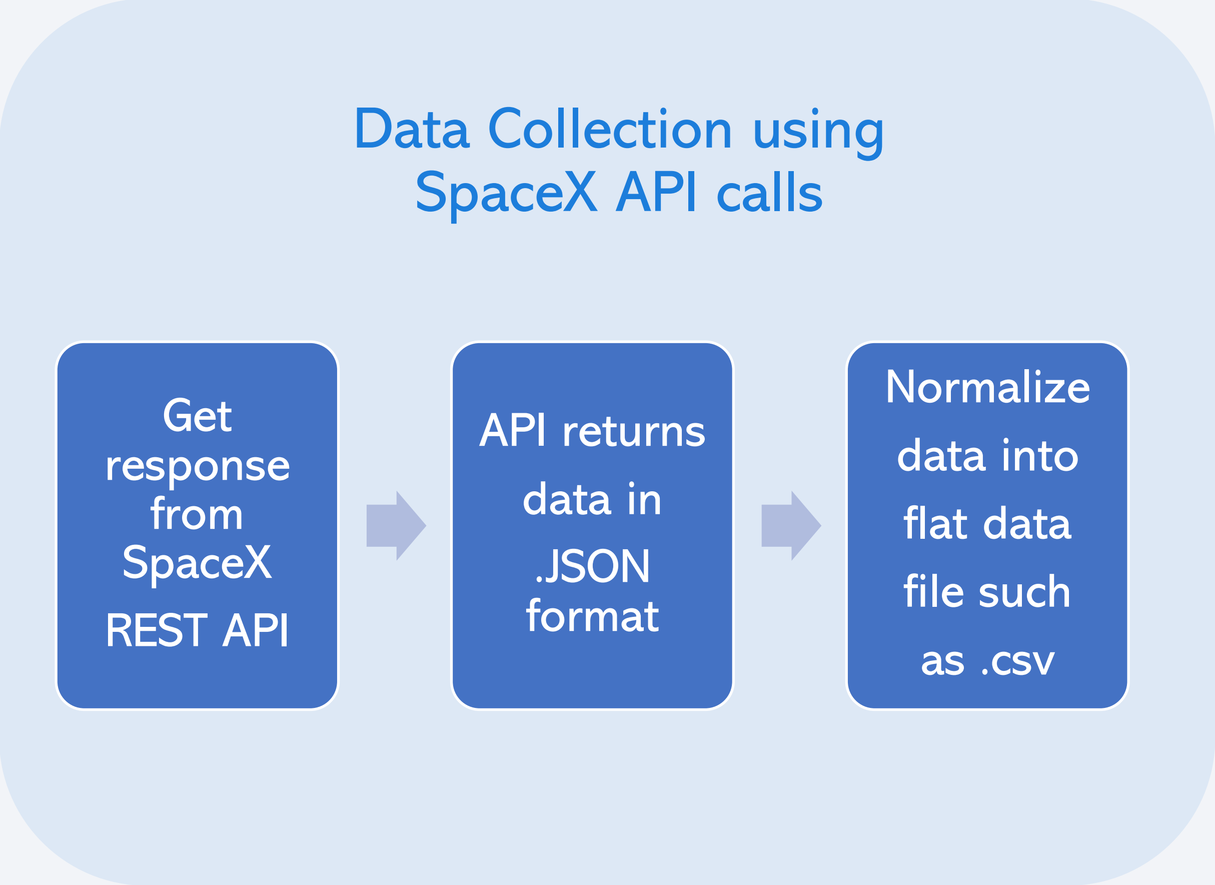
Data Collection – SpaceX API

- SpaceX launch data is gathered from the SpaceX REST API.

❑ This API contains data about detailed rocket launch information including date, payload, launch and landing specifications, orbit, customers, outcomes, etc .

❑ Requests are sent to receive relevant launch information in the format of Jason.

❑ The Jason data is decoded and processed to a structured dataframe and exported to a csv file for further feature engineering and data wrangling and data analysis.



Data Collection using SpaceX API calls

Get response from SpaceX REST API

API returns data in .JSON format

Normalize data into flat data file such as .csv

GitHub Link

Data Collection - Scraping

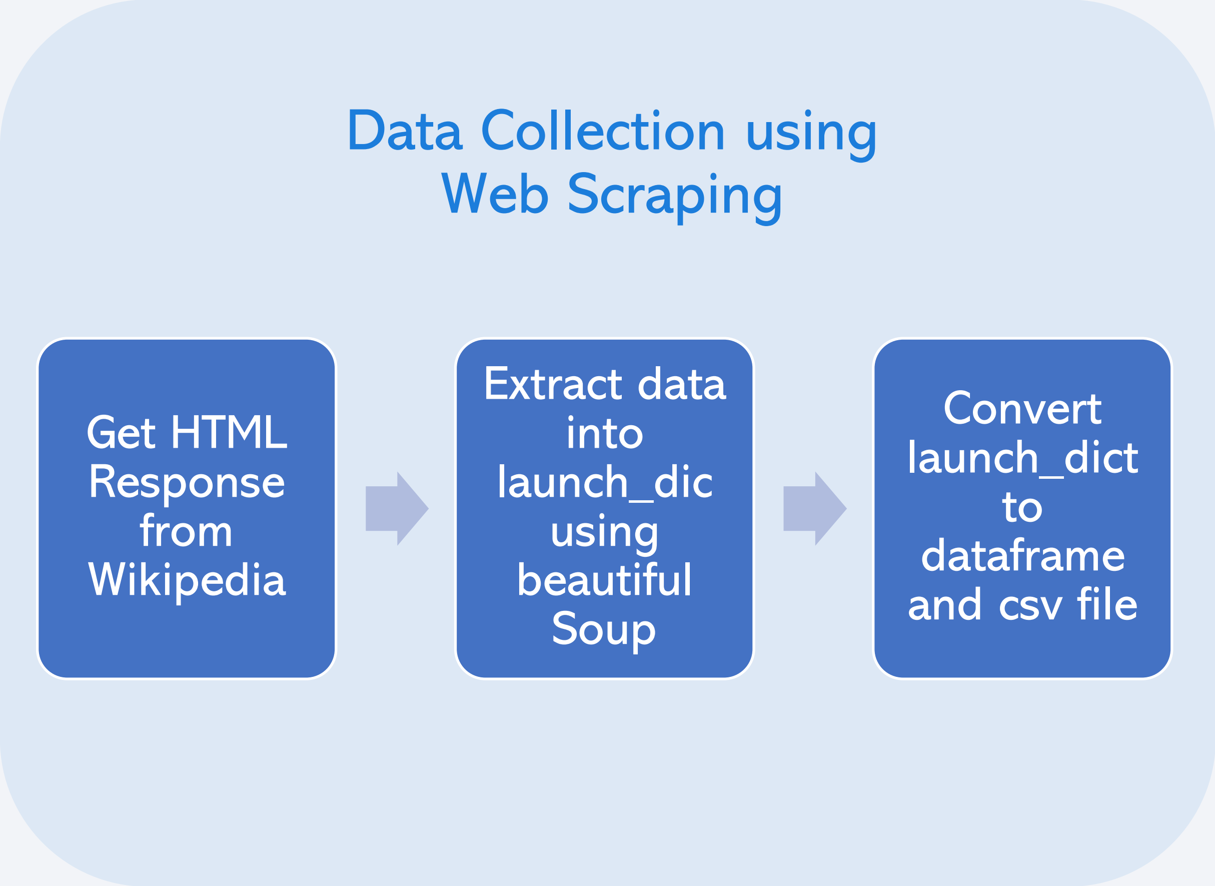
- Falcon 9 Launch data is also obtained through web scraping of Wikipedia using BeautifulSoup.

❑ Wikipedia contains up-to-date information with respect of Falcon 9 launches.

❑ Requests are sent to receive the data from html

❑ Beautiful soup can extract and parse the obtained table and place in a launch_dic

❑ The extracted data in the launch_dic can be converted to dataframe followed by exporting to a csv file for further data wrangling



Data Collection using Web Scraping

Get HTML Response from Wikipedia

Extract data into launch_dic using beautiful Soup

Convert launch_dict to dataframe and csv file

[GitHub Link]

Data Wrangling

- The obtained raw data have been processed as follows:

❑ Import libraries numpy and pandas

❑ Load the csv file to dataframe

❑ Examine data size, column name, data type, etc

❑ Find the missing data suing df.isnull().sum() and replace them with mean value

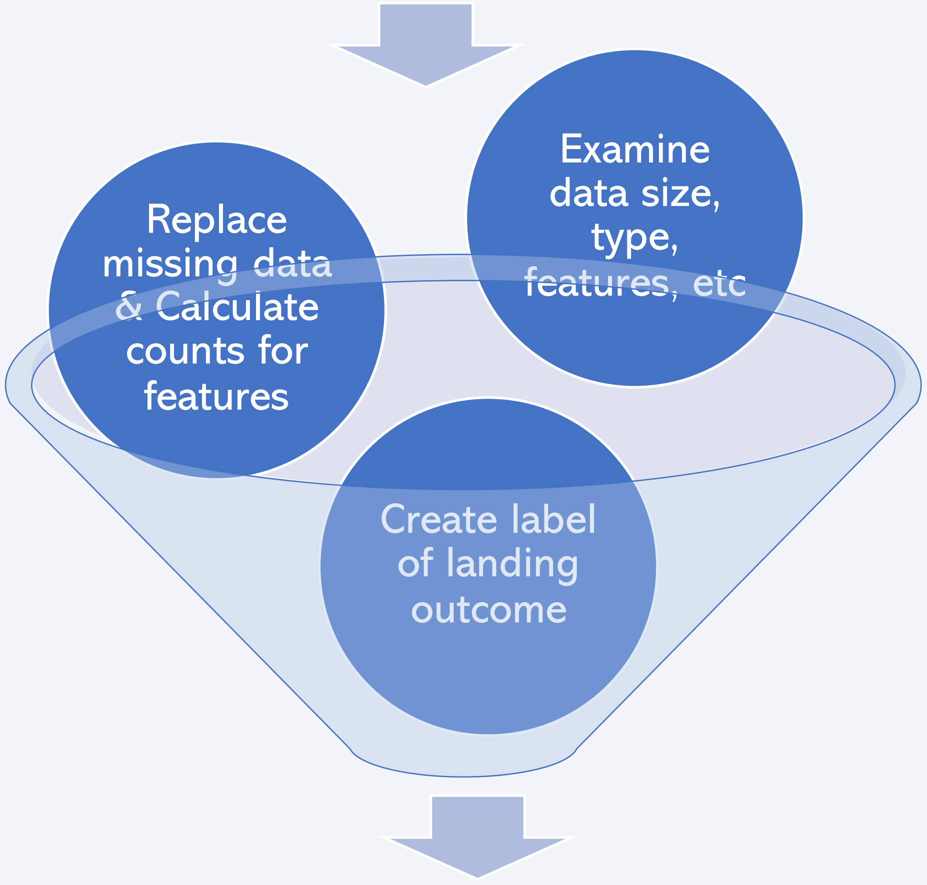
❑ Calculate (i) number of launches on each site (ii) the number and occurrence of each orbit (iii) the number and occurrence of mission outcome per orbit type using value_counts()

❑ Create a landing outcome label from Outcome column and calculate its success rate

❑ Export the dataframe to a csv file for further data analysis

GitHub Link

Load csv to dataframe


Replace missing data & Calculate counts for features
Examine data size, type, features, etc
Create label of landing outcome

Export to csv

EDA with Data Visualization

- Scatter plot: show relationships between two numeric variables and the key features such as flight number, launch sites, payloads and Orbit type are plotted.
  - ❑ Flight Number vs. Launch Site
  - ❑ Payload vs. Launch Site
  - ❑ Flight Number vs. Orbit Type
  - ❑ Payload vs. Orbit Type
- Bar Chart: compare performance between different groups and the different orbit types are compared in terms of success rate
  - ❑ Success Rate vs. Orbit Type
- Line plot: shows frequency of data along a number line, here is used to show the change of launch success over time
  - ❑ Launch Success Yearly Trend

GitHub Link

EDA with SQL

- Perform SQL queries to gather information and understand the dataset:
  - ❑ Show all launch site names
  - ❑ Display 5 records of launch site names begin with the string ‘CCA’
  - ❑ Calculate total payload mass launched by NASA (CRS)
  - ❑ Calculate average payload mass by F9 v1.1
  - ❑ Show the date of the first successful ground landing date
  - ❑ List the names of the boosters with payload mass between 4000 and 6000
  - ❑ Count the total number of successful and failure mission outcomes
  - ❑ List the booster_versions carrying maximum payload mass.
  - ❑ Display the records of the failed landing_outcomes in drone ship in 2005
  - ❑ Rank The Count of Landing Outcomes Between 2010-06-04 and 2017-03-20 in Descending Order

GitHub Link

Build an Interactive Map with Folium

- Map objects including markers, circles, MarkerCluster, MousePosition and lines are created and added to the folium maps.
  - ❑ The markers and circles are used to mark the location of launch sites.
  - ❑ The MarkClusters are used to summarize the total number of the launch missions.
  - ❑ The green and red labelled markers are used to distinguish successful and failure missions.
  - ❑ The MousePosition is used to indicate the coordinate of the map like latitude and longtitude.
  - ❑ The lines are used to display the distance between the launch sites to other places including the highways, railways, coastlines, cities, etc.

- The interactive maps provide visualized information regarding the launch sites with launch outcomes and their proximities.

GitHub Link

Build a Dashboard with Plotly Dash

- Dashboard title, dropdown buttons, rangesliders, pie charts and scatter figures have been added to the interactive dashboard with Plotly Dash.
    - ❑ The title shows the theme of the dashboard
    - ❑ Dropdown provides the users with multiple options to show the launch records of either all the launch sies or specific launch sites.
    - ❑ The rangeslides allow user to select certain range of payloads for visualization of the launch_outcomes.
    - ❑ The pie charts displays the successful launches for each launch sites as well as the success rate for individual launch sites.
    - ❑ The bubble scatter chart can indicate the correlation between the mission_outcomes at certain payload range.

GitHub Link

Predictive Analysis (Classification)

- Build Models

❑ Load dataset into NumPy and convert to dataframe using Pandas
❑ Transform numerical Data via scalar
❑ Encode the categorical data via one hot encoding
❑ Split data into train and test data sets
❑ Select 4 types of machine learning algorithms for prediction
❑ Find suitable parameters for each algorithms using GridSearchCV
❑ Train the models(GridSearchCVobjects) suing train dataset.

- Evaluate Models

❑ Check accuracy for each model using train dataset using tuned parameter
❑ Check accuracy for each model using test dataset
❑ Plot Confusion Matrix

GitHub Link

Predictive Analysis (Classification)

- Improve the Models

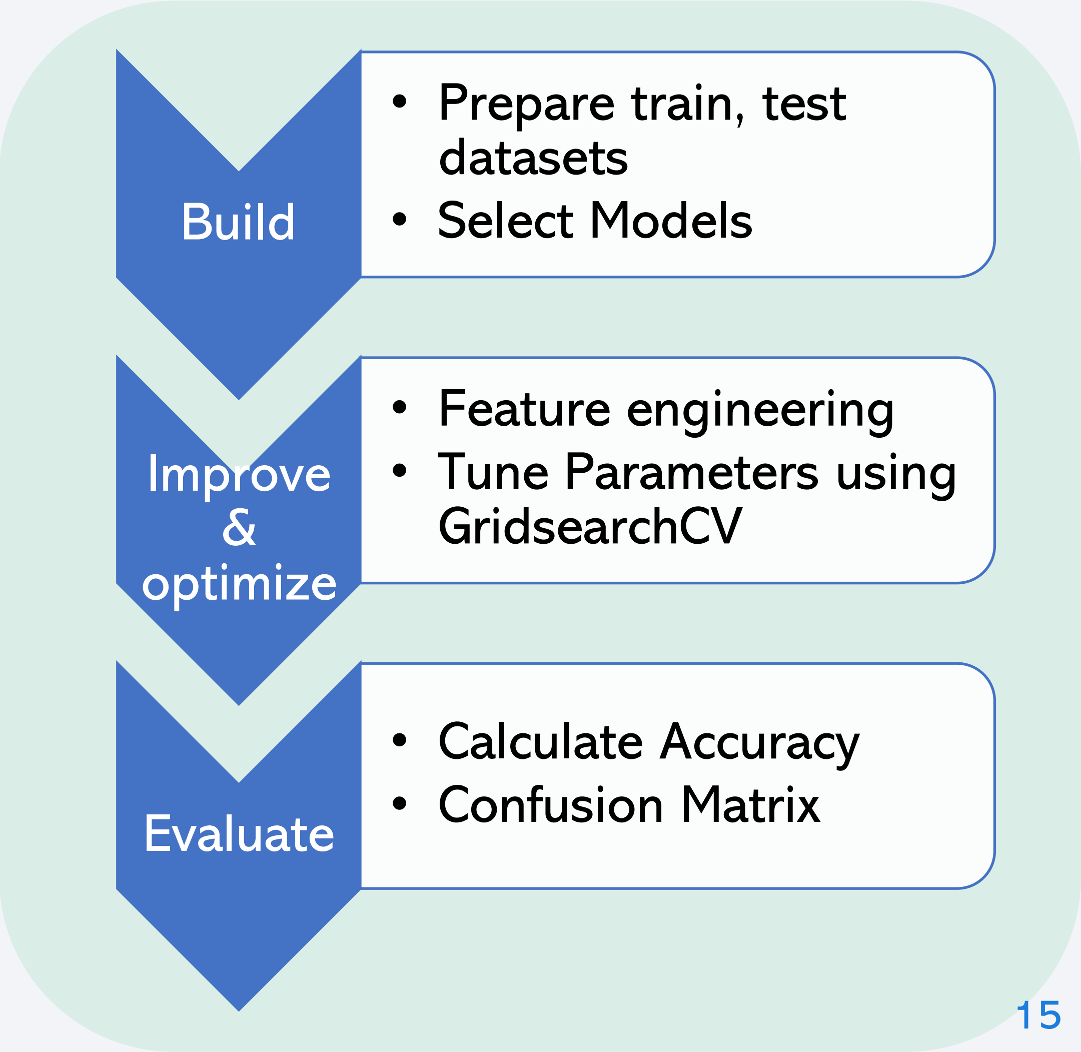
❑ Feature Engineering

❑ Algorithm Tuning

- Find the Best Performing Classification Model

❑ The model with the best accuracy score as best performing model

❑ The confusion matrix is also used to assess the performance of models.



Build
- Prepare train, test datasets
- Select Models

Improve & optimize
- Feature engineering
- Tune Parameters using GridsearchCV

Evaluate
- Calculate Accuracy
- Confusion Matrix

GitHub Link

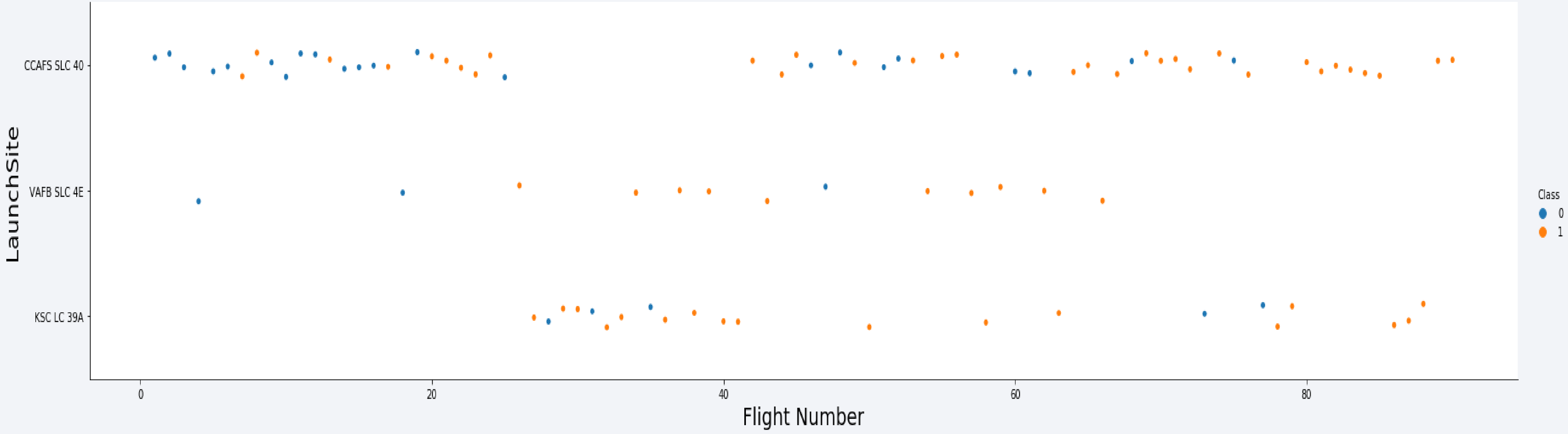
Results

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

Section 2

Insights drawn from EDA

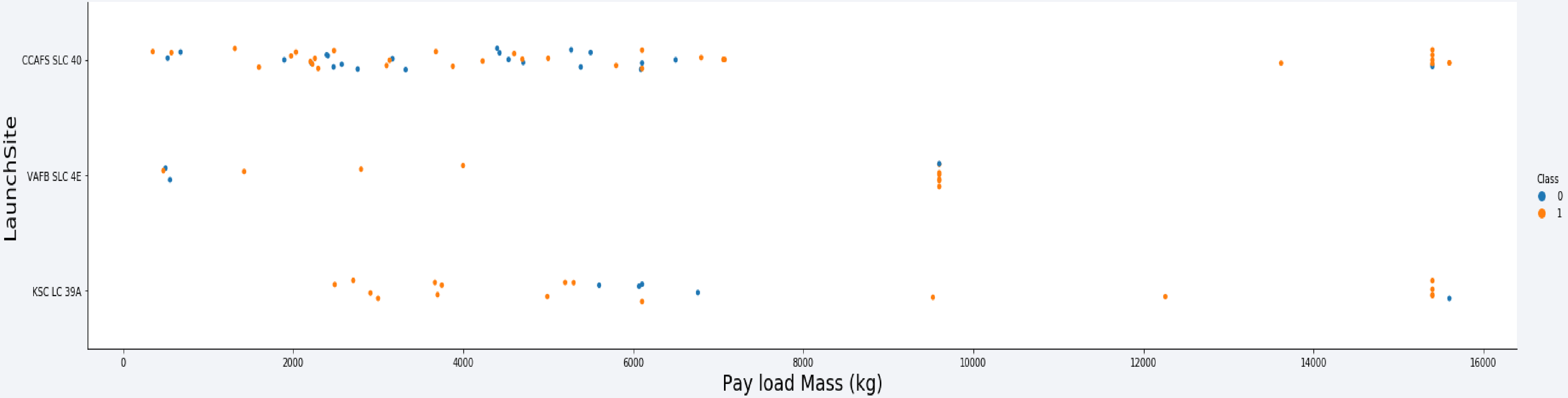
Flight Number vs. Launch Site



- With the increase in the flight number, the success rate increases at all 3 launch sites.

GitHub Link

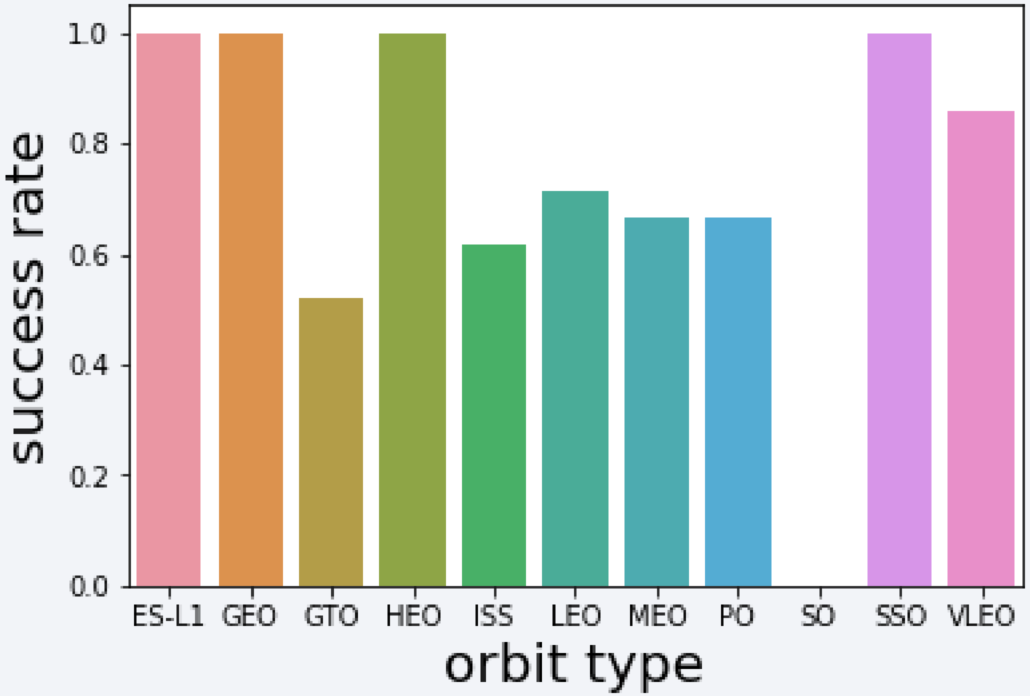
Payload vs. Launch Site



- For the CCAFS SLC40 site, it is suitable for launches with light and medium payloads (0-6500 kg).
- For the VAFB-SLC site, there are no rockets launched for heavy payload mass (greater than 10000kg).
- For the KSC LC39A site, it has relatively high successful launches for medium payload (e.g 5000kg-7000kg)

GitHub Link

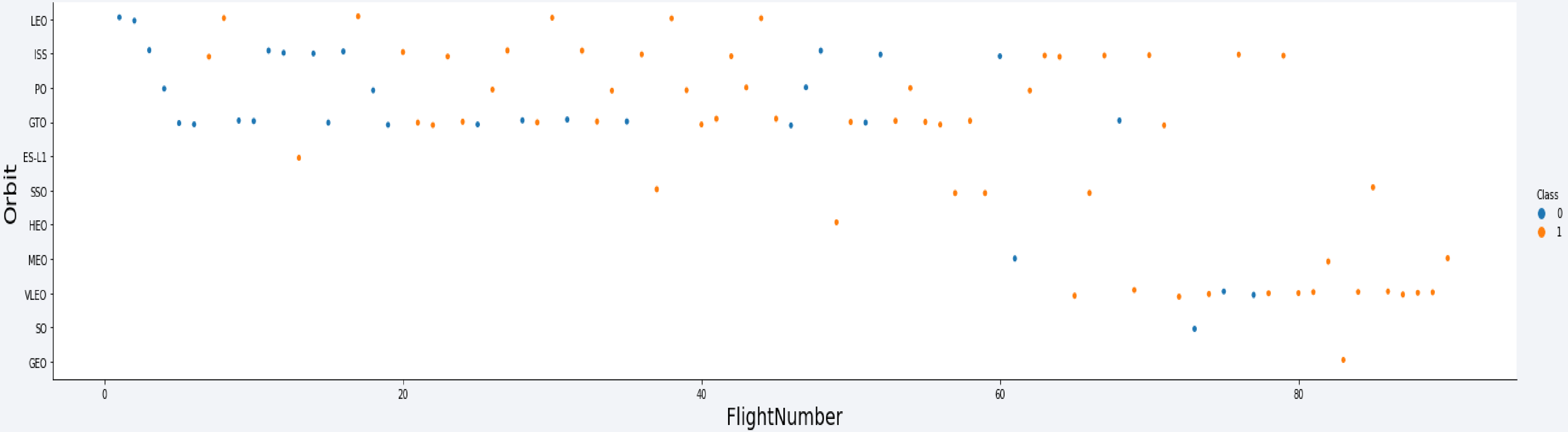
Success Rate vs. Orbit Type



- The bar chart shows that the orbits of ES-L1, GEO, HEO, SSO have the highest success rate.

GitHub Link

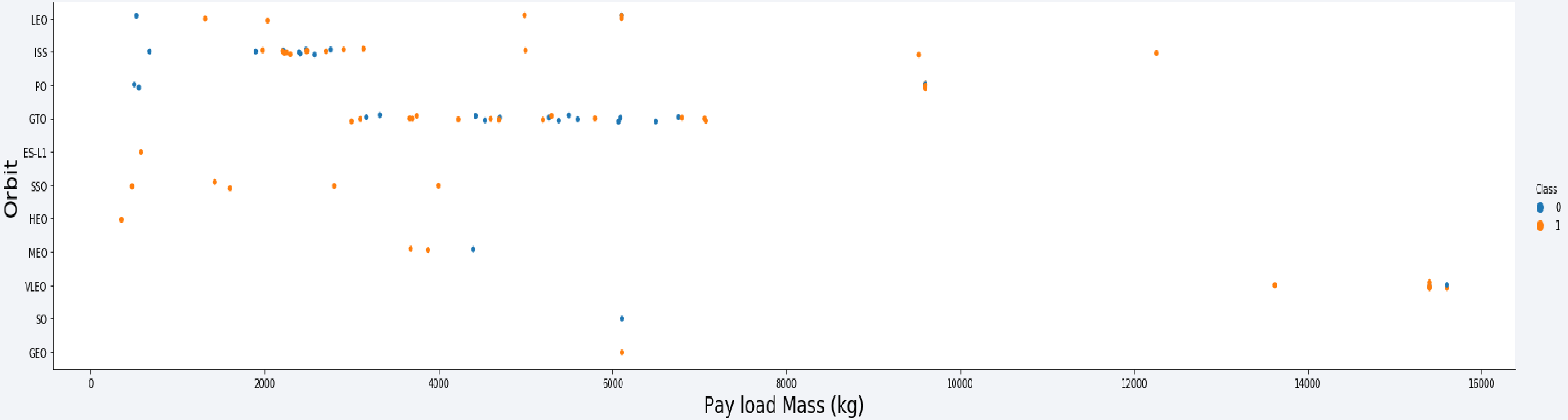
Flight Number vs. Orbit Type



- For LEO orbit, the success launch rate increases with the number of flights.
- There is no obvious relation to flight number when in other orbits such as GTO.
- Rocket launches at SSO orbit all succeed.

GitHub Link

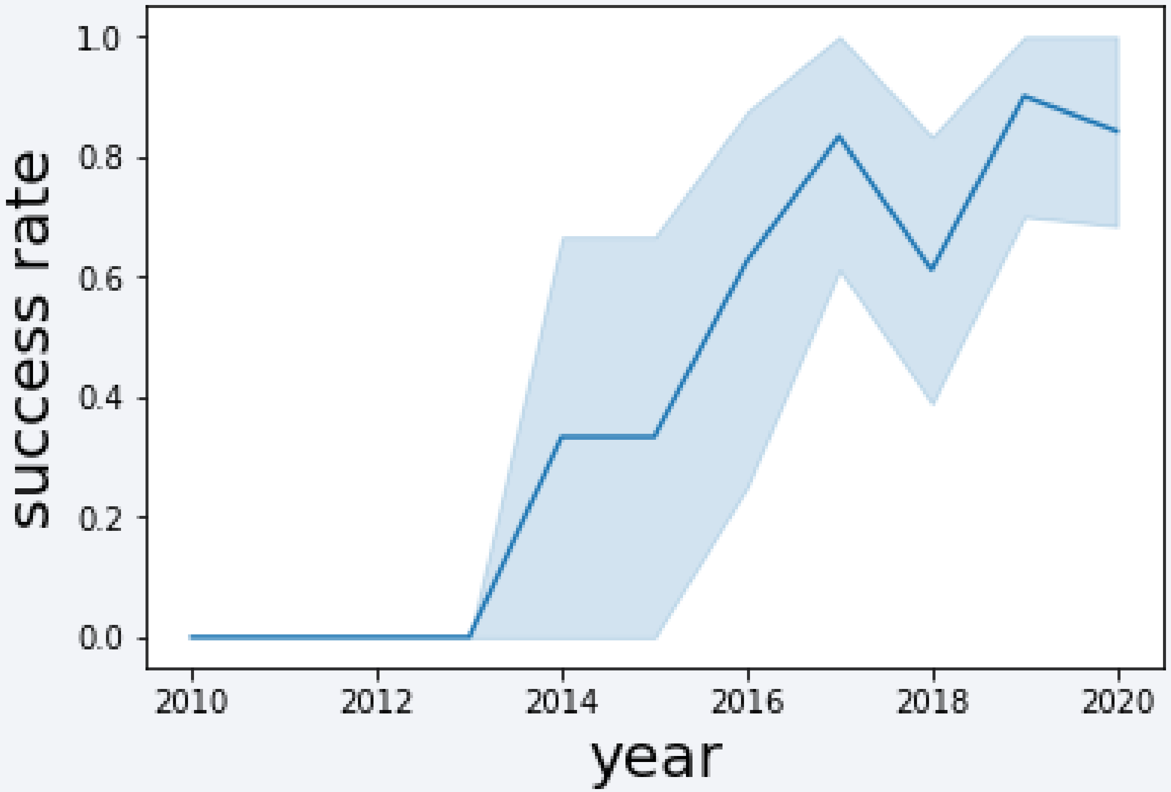
Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate increases for PO, LEO and ISS.
- For GTO we don’t distinguish the above trend since both failure and successful missions are observed.

GitHub Link

Launch Success Yearly Trend



- The line chart shows an uptrend of success rate, which kept increasing till 2020.

GitHub Link

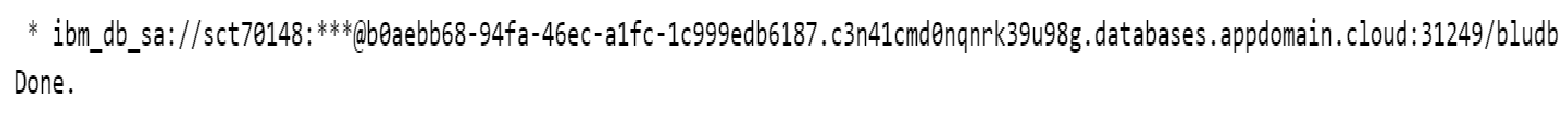
Section 3

Understand Dataset with SQL

All Launch Site Names

- Connect the IBM db2 database using service credential

  %sql ibm_db_sa://my-username:my-password@my-hostname:my-port/my-db-name?security=SSL

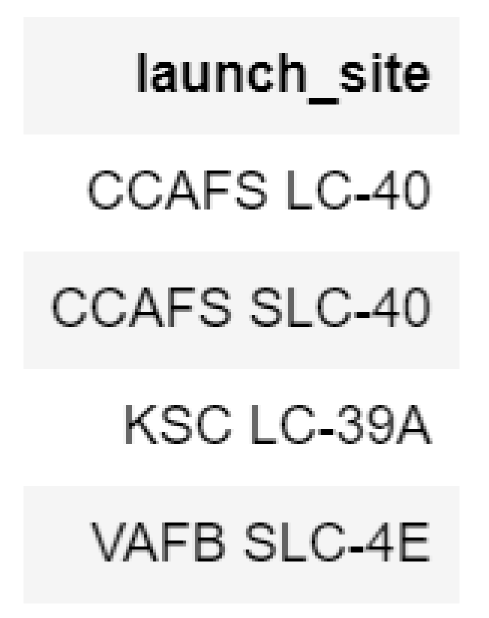


```
 * ibm_db_sa://sct70148:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:31249/bludb
Done.
```

- Execute the below SQL queries to find the results:

  %sql select distinct LAUNCH_SITE from SPACEXTBL

- Note that “Distinct” was used to query unique result.



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

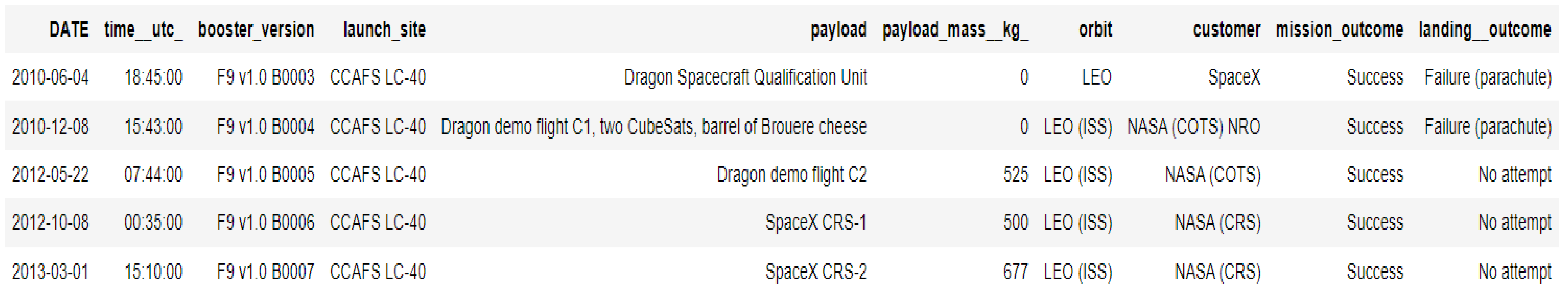
GitHub Link

5 Records of Launch Site Names Begin with 'CCA'

- Execute the below SQL query to find the results:

```
%sql SELECT * from SPACEXTBL where LAUNCH_SITE LIKE 'CCA%' LIMIT 5
```

- Note that “LIKE” was used to query site names begin with “CCA” while limit 5 to shown only 5 records.



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

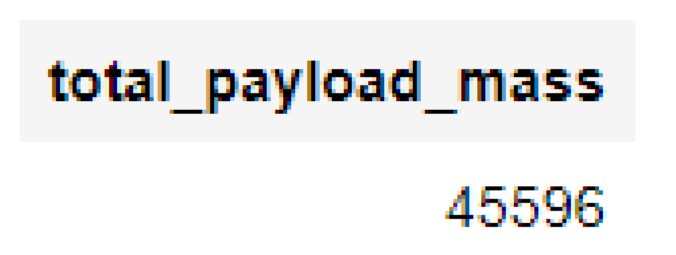
GitHub Link

Total Payload Mass Launched by NASA (CRS)

- Execute the below SQL query to find the total payload mass of 45596.

```
%sql select sum(payload_mass__kg_) as
Total_Payload_Mass from SPACEXTBL
where customer='NASA (CRS)’
```

- Sum() to calculate the total value, while where clause as a conational query to filter the results



```
total_payload_mass
             45596
```

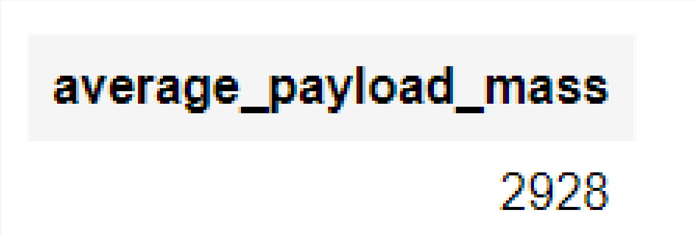
GitHub Link

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 was 2928 by SQL query:

```
%sql select avg(PAYLOAD_MASS__KG_) from
SPACEXTBL where BOOSTER_VERSION='F9 v1.1’
```

- Avg() to calculate the average value, while where clause as a conational query to filter the results



average_payload_mass
2928

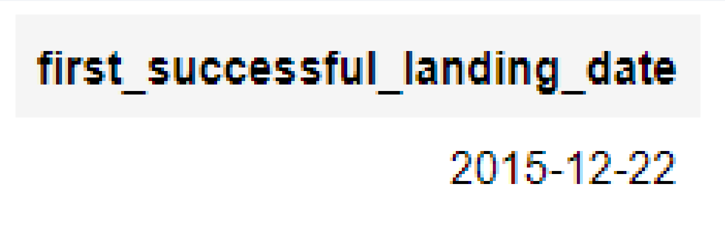
GitHub Link

First Successful Ground Landing Date

- The first successful landing date with outcome on ground pad was 2015/12/22 by SQL query:

```
%sql select min(DATE) as
first_successful_landing_date from SPACEXTBL where
LANDING__OUTCOME='Success (ground pad)’
```

- Min(Date) to find the first date, while where clause as a conational query to filter the results



first_successful_landing_date
2015-12-22

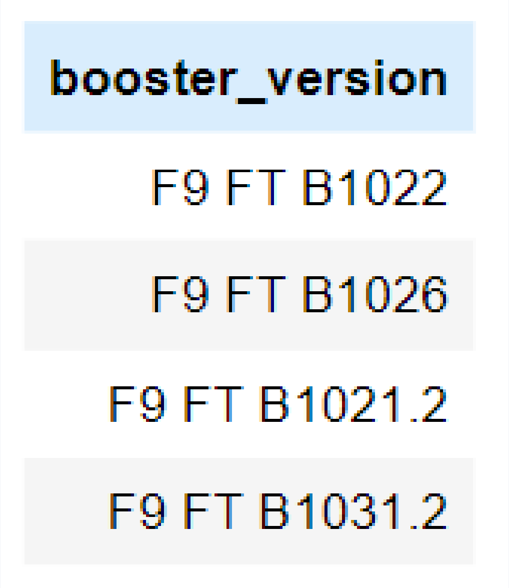
GitHub Link

Successful Drone Ship Landing with Payload between 4000 and 6000

- The list of boosters having successfully landed on droe ship with certain payload is shown in the right table by SQL query:

  %sql select BOOSTER_VERSION from SPACEXTBL where LANDING__OUTCOME='Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000

- WHERE clause as a conational query to filter the results with certain payloads and landing_outcome while ‘AND’ to apply multiple filters.

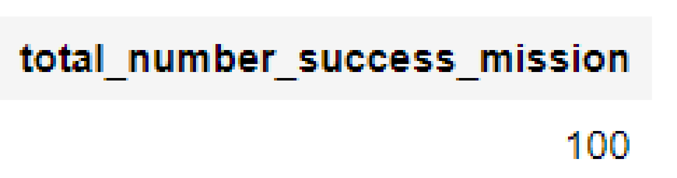


booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

GitHub Link

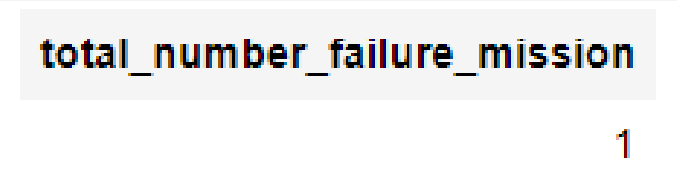
Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes is 100 and 1, respectively. The SQL queries are shown below:
- %sql select count(MISSION_OUTCOME) as total_number_success_mission from SPACEXTBL where MISSION_OUTCOME like 'Success%’



total_number_success_mission
100

- %sql select count(MISSION_OUTCOME) as total_number_failure_mission from SPACEXTBL where MISSION_OUTCOME like 'Failure%’



total_number_failure_mission
1

- Count() to count the total number, while the where clause to filter the results containing either ‘Success’ or ‘Failure’ in the mission outcomes.

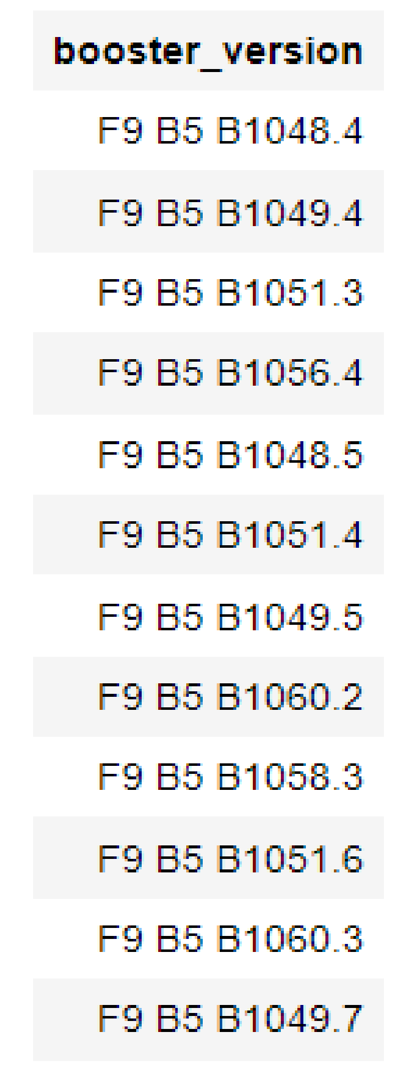
GitHub Link

Boosters Carried Maximum Payload

- The list of boosters with max payload is shown in the right table by SQL subquery:

```
%sql select BOOSTER_VERSION from SPACEXTBL
where PAYLOAD_MASS__KG_=(select
max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

- A subquery is used to select the max payload using max() function, then booster_versions are selected through a WHERE clause to filter the results with maximum payload.



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

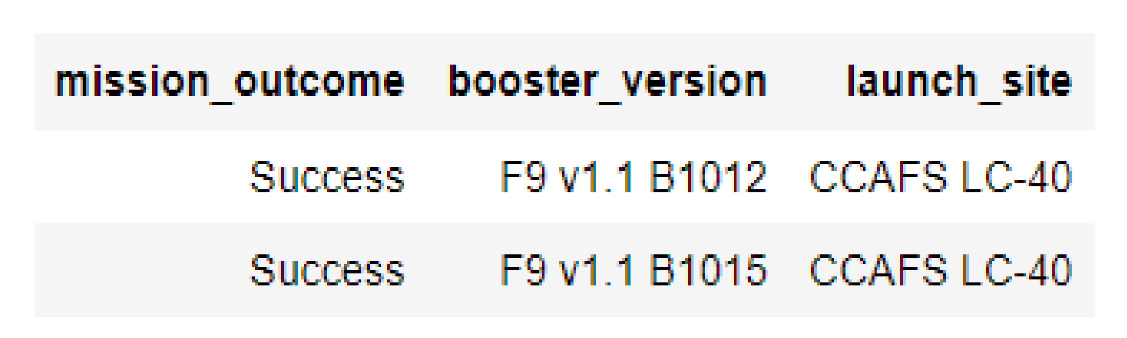
GitHub Link

2015 Launch Records

- The list the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 are obtained by SQL query:

```
%sql SELECT MISSION_OUTCOME,
BOOSTER_VERSION, LAUNCH_SITE FROM
SPACEXTBL where EXTRACT(YEAR FROM
DATE)='2015’ and
LANDING__OUTCOME='Failure (drone ship)’
```

- An extract() function is used to select the year 2015 from date, while where clause with two filters connected by ‘AND’ to filter the results.



mission_outcome	booster_version	launch_site
Success	F9 v1.1 B1012	CCAFS LC-40
Success	F9 v1.1 B1015	CCAFS LC-40

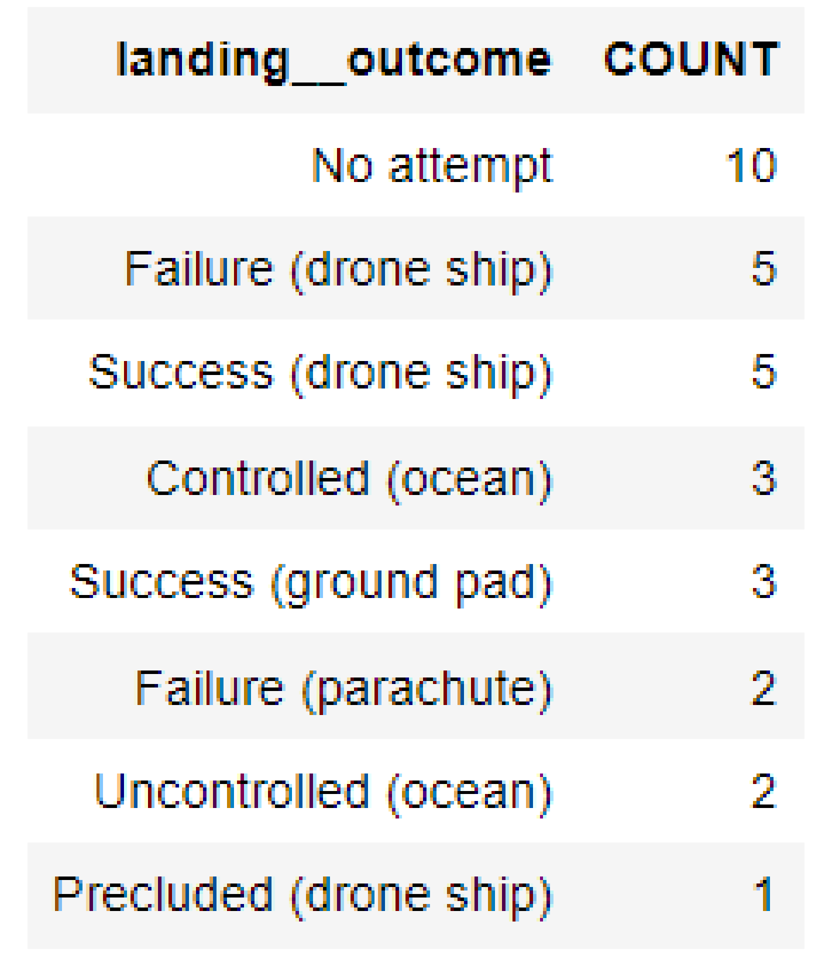
GitHub Link

Rank The Count of Landing Outcomes Between 2010-06-04 and 2017-03-20 in Descending Order

- The landing coutcomes are ranked as shown in the right table using the SQL query:

  %sql SELECT LANDING__OUTCOME, count(LANDING__OUTCOME) as count FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20’ GROUP BY LANDING__OUTCOME ORDER BY count DESC

- Count() and GROUP By are used to obtain counts for each outcome while WHERE clause to filter the DATE range. The obtained results are ranked using ORDER BY method in the descending order via DESC.



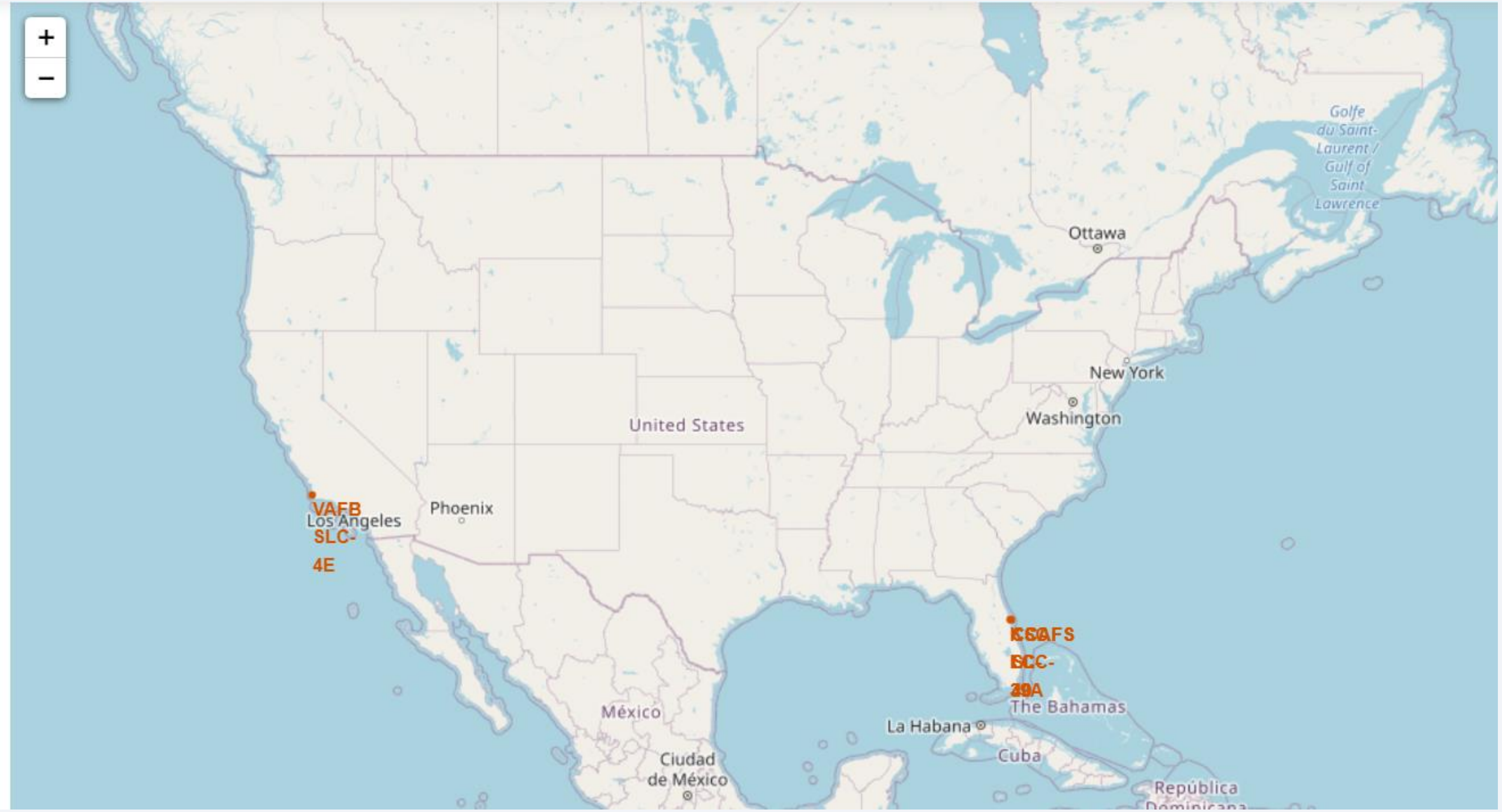
landing__outcome	COUNT
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

GitHub Link

Section 4

Launch Sites Proximities Analysis

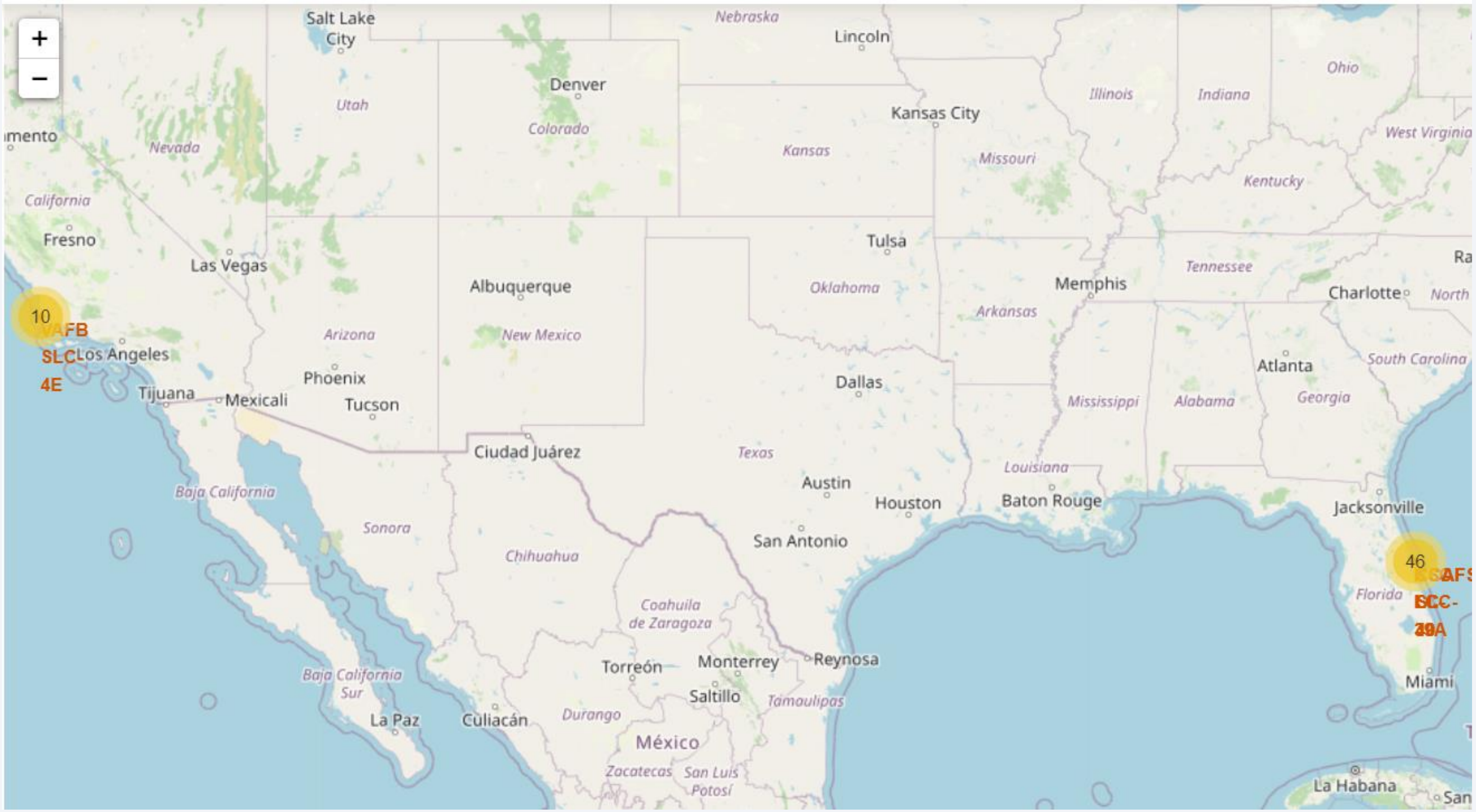
All Launch Sites Marked on the Map



GitHub Link

- 3 launch sites in FL and 1 launch site in CA.

Success/Failed Launches Color-Labeled on the Map

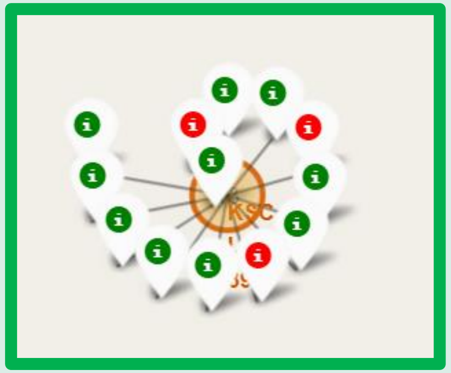


GitHub Link

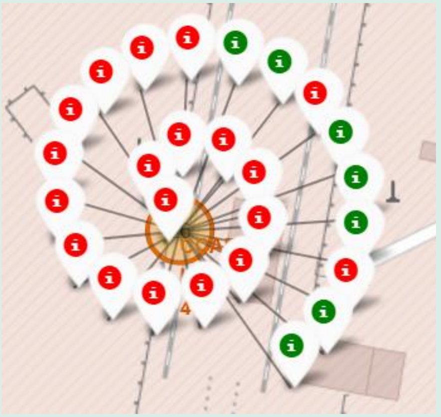
- All the launches are marked on the map as mark-clusters.

Success/Failed Launches Color-Labeled on the Map

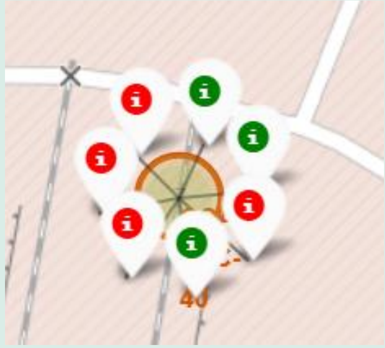
KSC LC-39A
13 launches



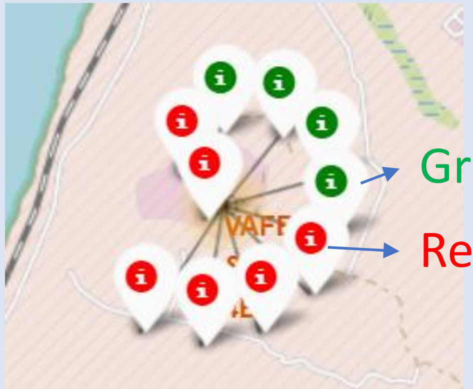
CCAFS LC-40
26 launches



CCAFS SLC-40
7 launches

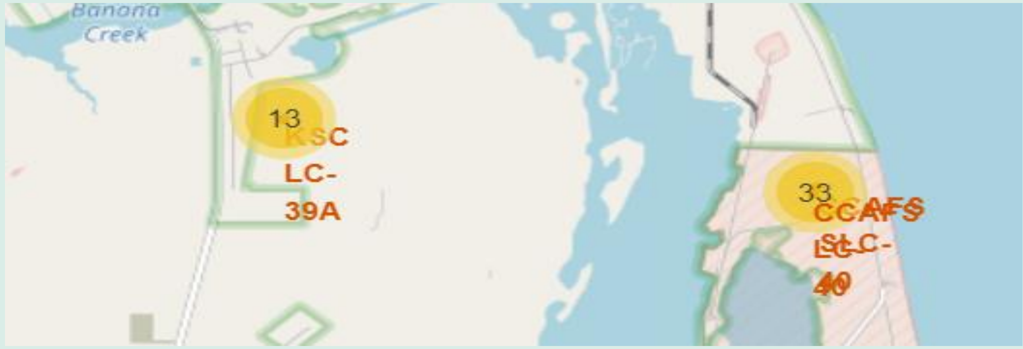


VAFB SLC-4E
10 launches



Green: success

Red: failure



FL launch sites

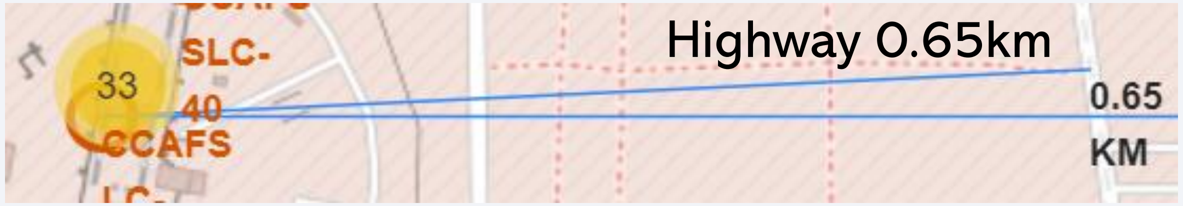
Base
Vandenberg State Marine Reserve
10
VAFB SLC-4E

CA launch sites

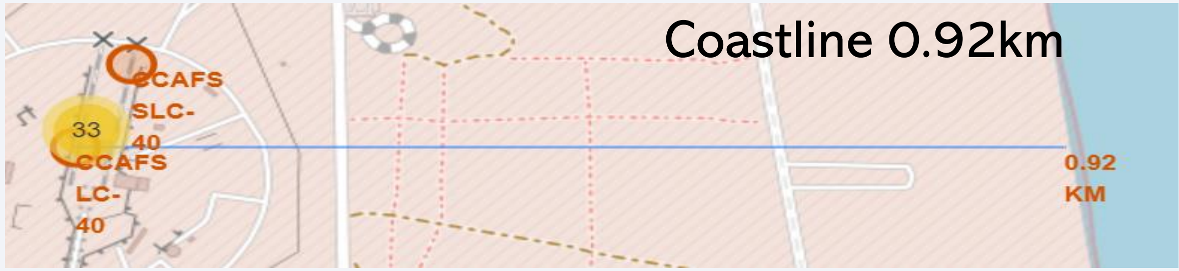
GitHub Link

- KSC LC-39A launch site in Florida has the highest success rate.

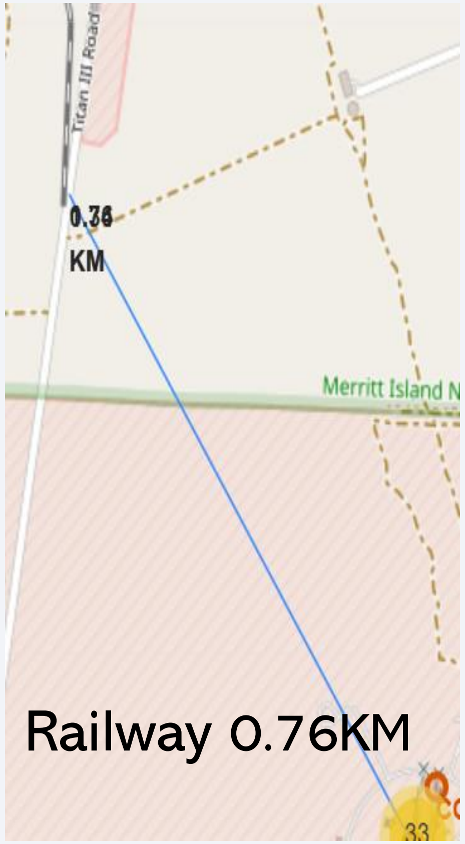
Distance of Launch Site (KSC LC-39A) to its Proximities



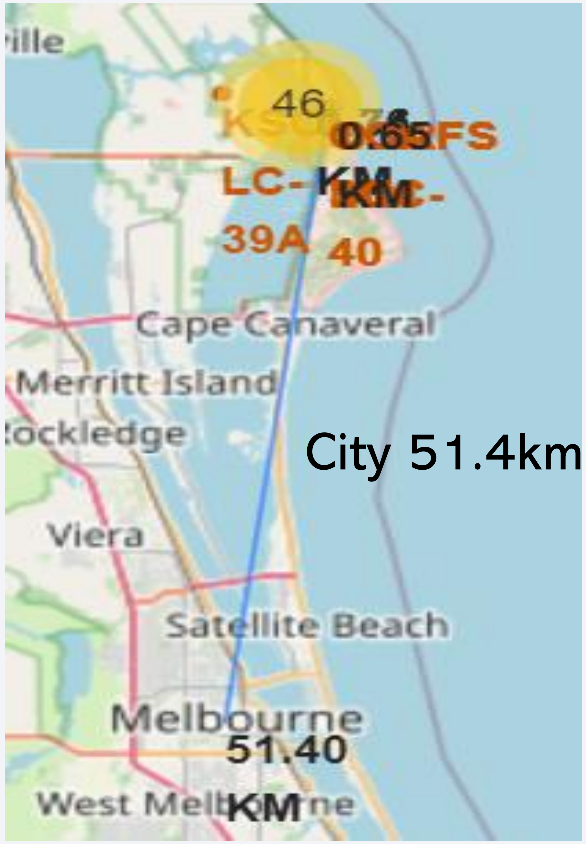
Highway 0.65km



Coastline 0.92km



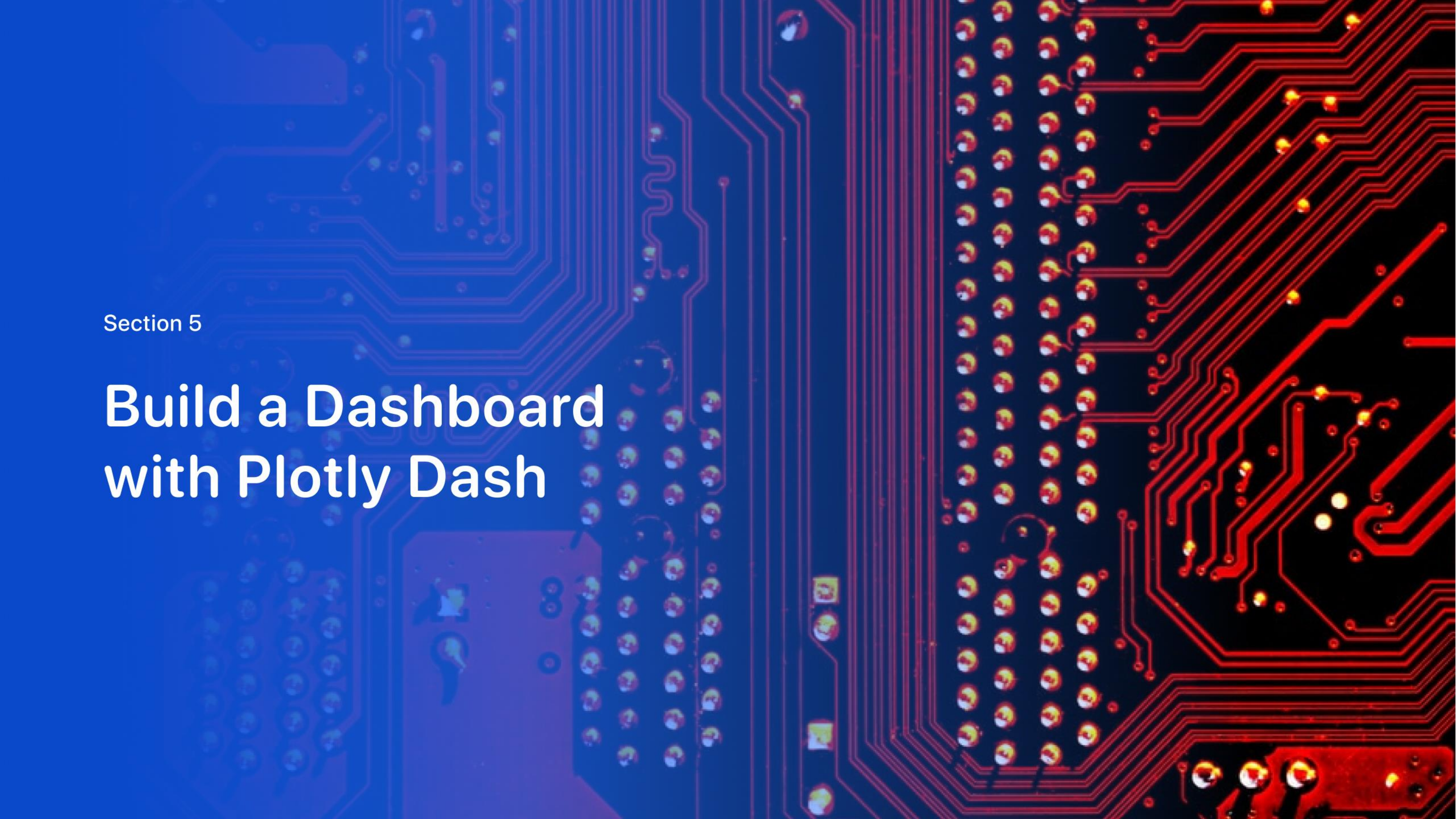
Railway 0.76KM



City 51.4km

- KSC LC-39A is in close proximity to highway, railway and coastline.
- KSC LC-39A keep certain distance away from cities.

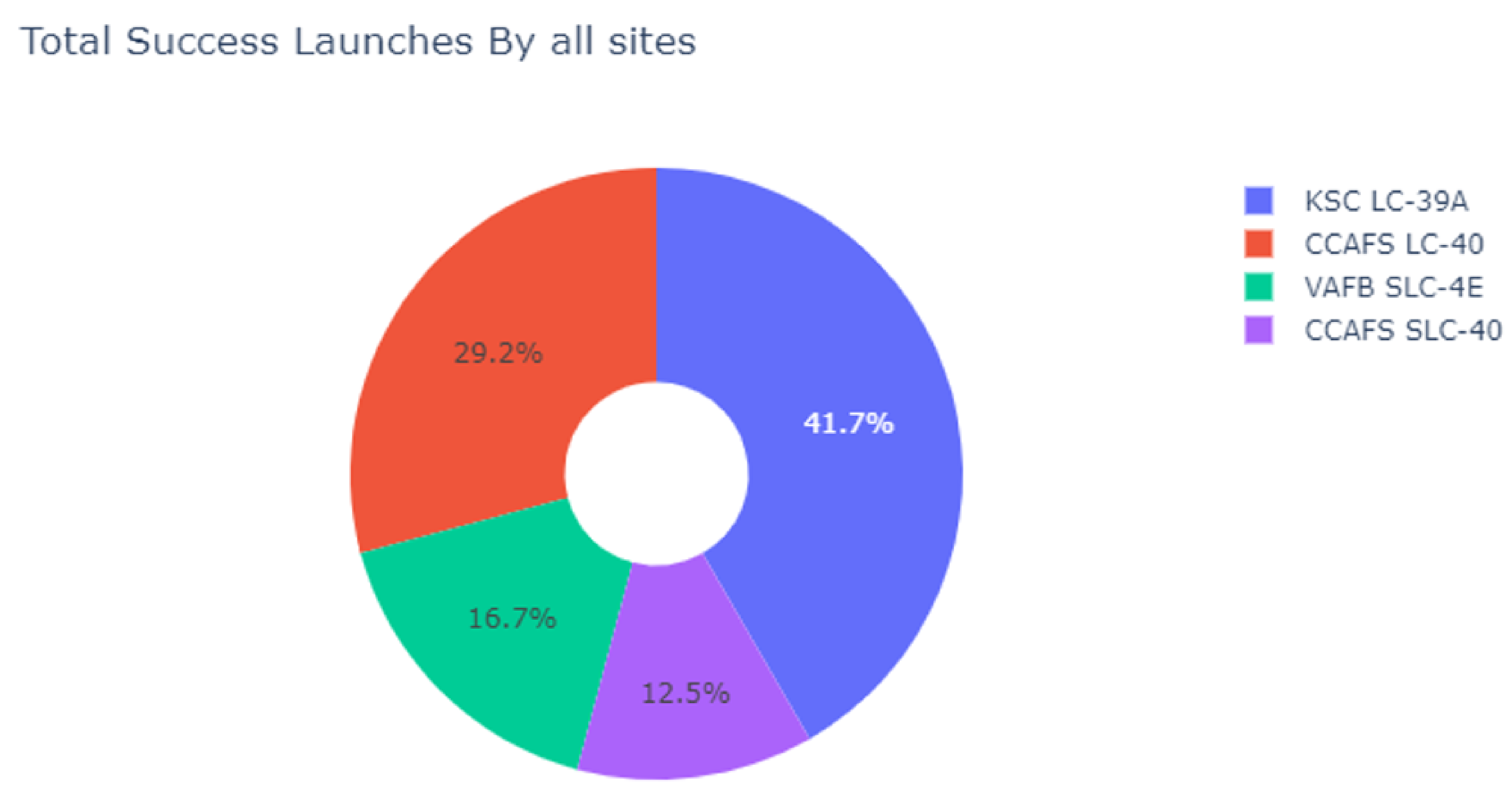
GitHub Link



Section 5

Build a Dashboard with Plotly Dash

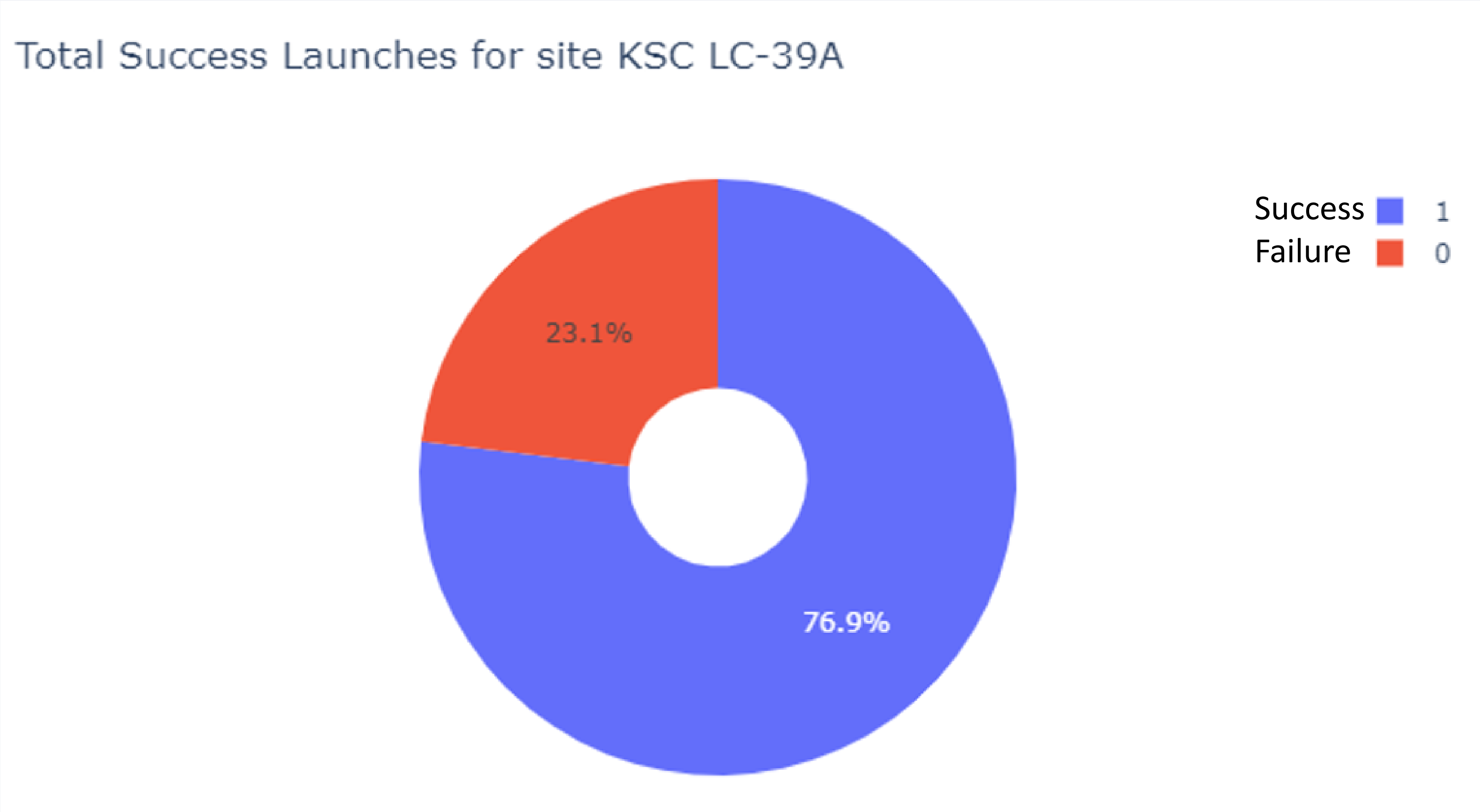
SpaceX Launch Records Dashboard- Launch Success Count



GitHub Link

- KSC LC-39A has most successful launches, followed by CCAFS LC-40, VAFB SLC-4E and CCAFS SLC-40.

SpaceX Launch Records Dashboard- Launch Site with Highest Launch Success Ratio



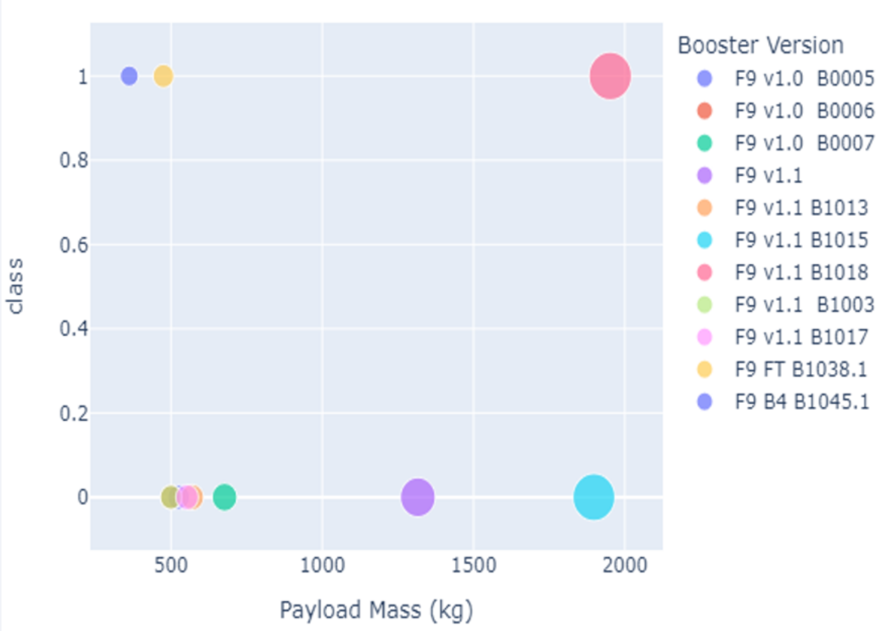
Total Success Launches for site KSC LC-39A

GitHub Link

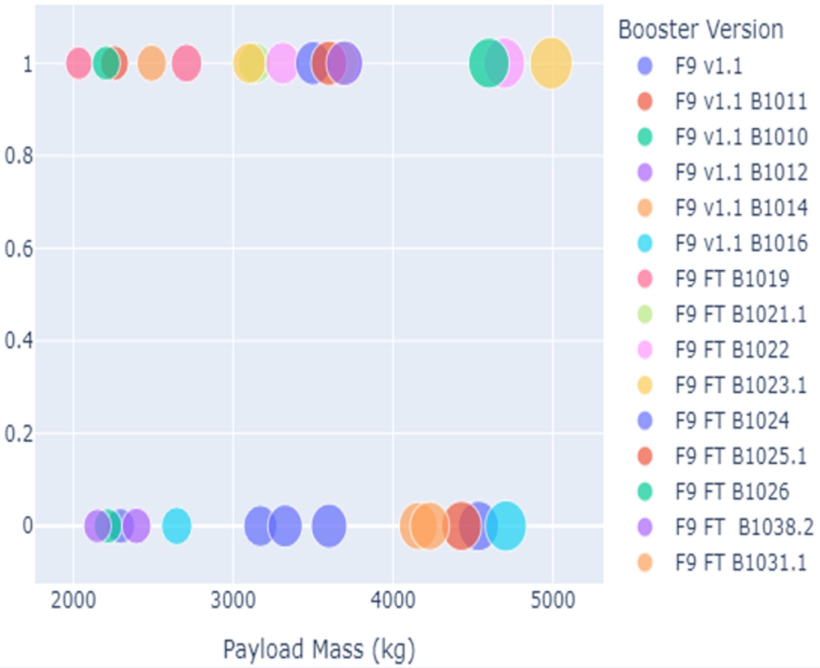
- KSC LC-39A has highest launch success rate of 76.9%.

SpaceX Launch Records Dashboard- Payload vs. Launch Outcome for All Sites

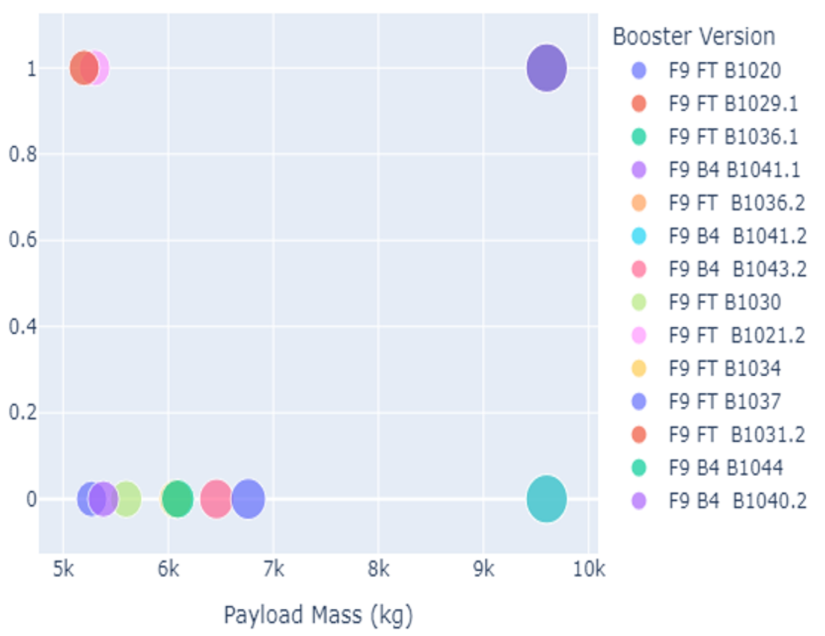
Payload: 0-2000kg



Payload: 2000-5000kg



Payload: 5000-10000kg



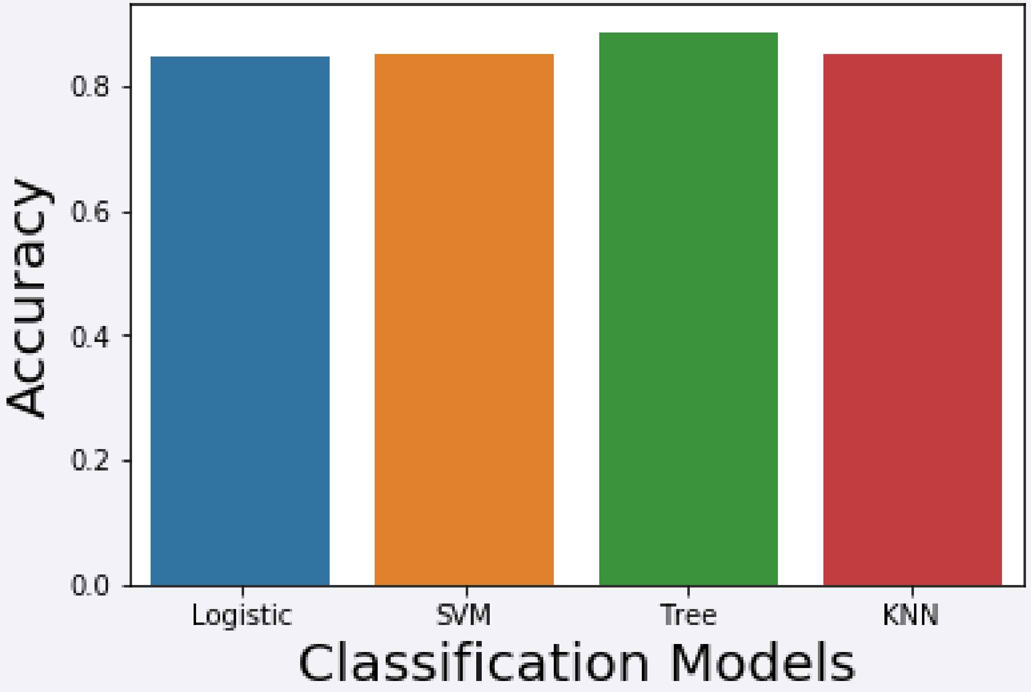
- For light payload (0-2000kg), the launch success rate is high.
- For medium payload (2000-5000kg), the success and failure launches counts is comparable.
- For heavy payload (>5000kg), the launch success rate is low.

GitHub Link

Section 6

Predictive Analysis (Classification)

Classification Accuracy of Train Data

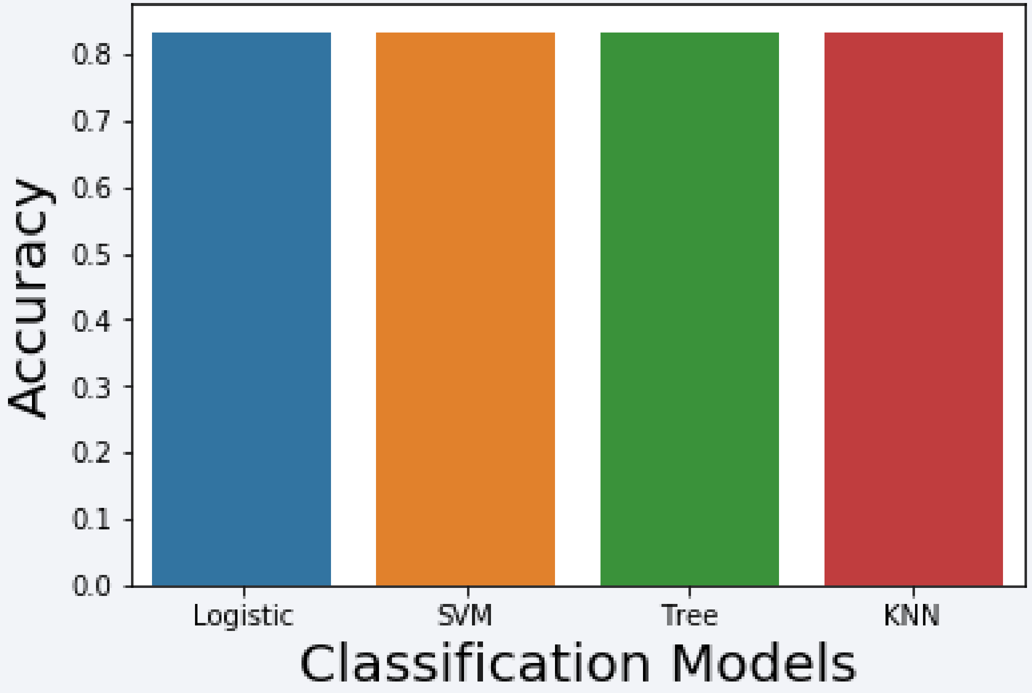


Classification models	Accuracy
Logistic regression	0.8464
SVM	0.8482
Decision tree	0.8857
KNN	0.8482

- The bar chart shows the accuracy for varied classification models based on the train data.
- Decision tree models gives rise to the highest accuracy of 0.8857 as the best performing model.

GitHub Link

Classification Accuracy of Test Data

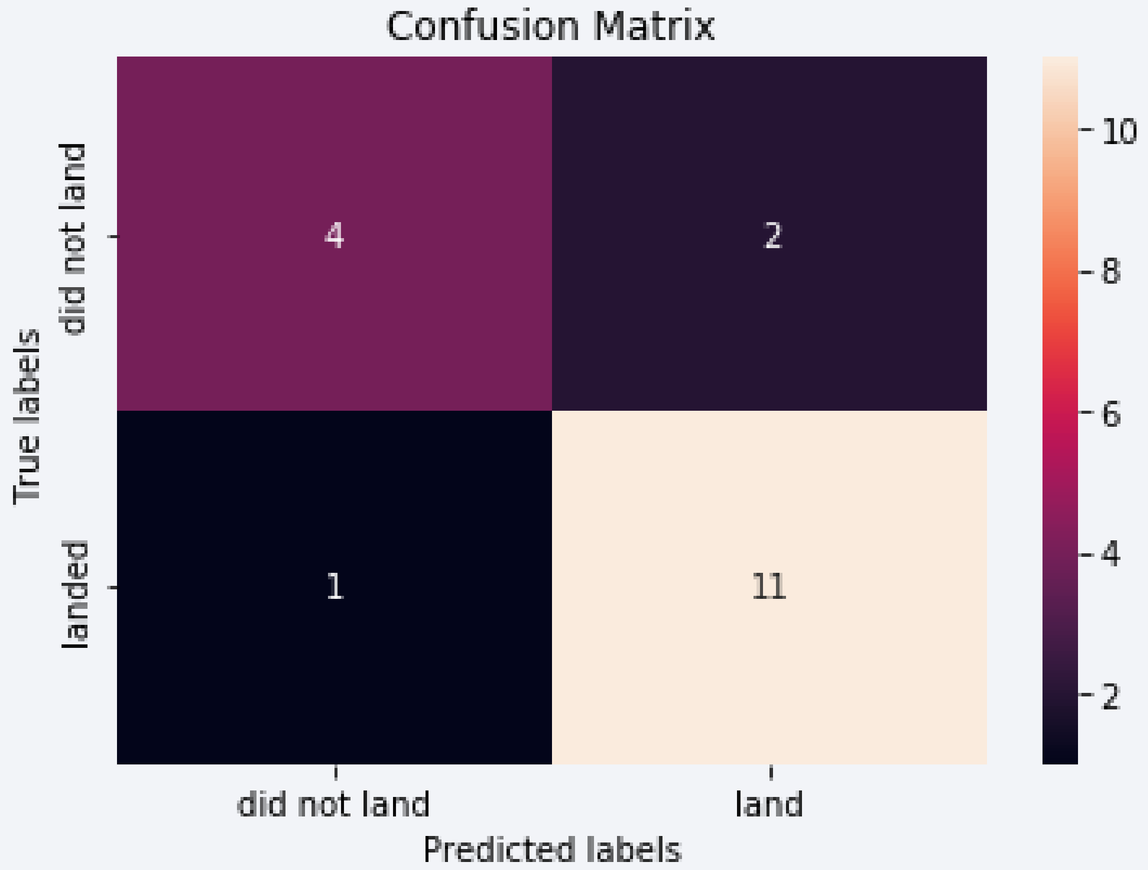


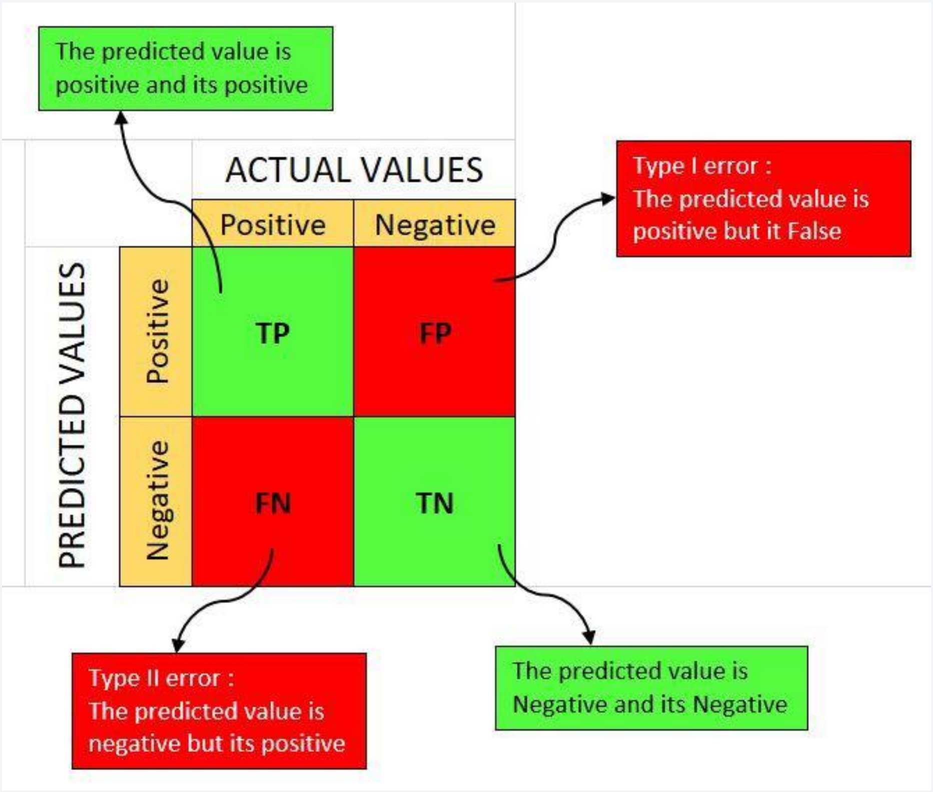
Classification models	Accuracy
Logistic regression	0.8333
SVM	0.8333
Decision tree	0.8333
KNN	0.8333

- The bar chart shows the accuracy for varied classification models based on the test data.
- All models renders the same accuracy of 0.8333.

GitHub Link

Confusion Matrix of Decision Tree





The predicted value is positive and its positive

		ACTUAL VALUES	
		Positive	Negative
PREDICTED VALUES	Positive	TP	FP
	Negative	FN	TN

Type I error : The predicted value is positive but it False

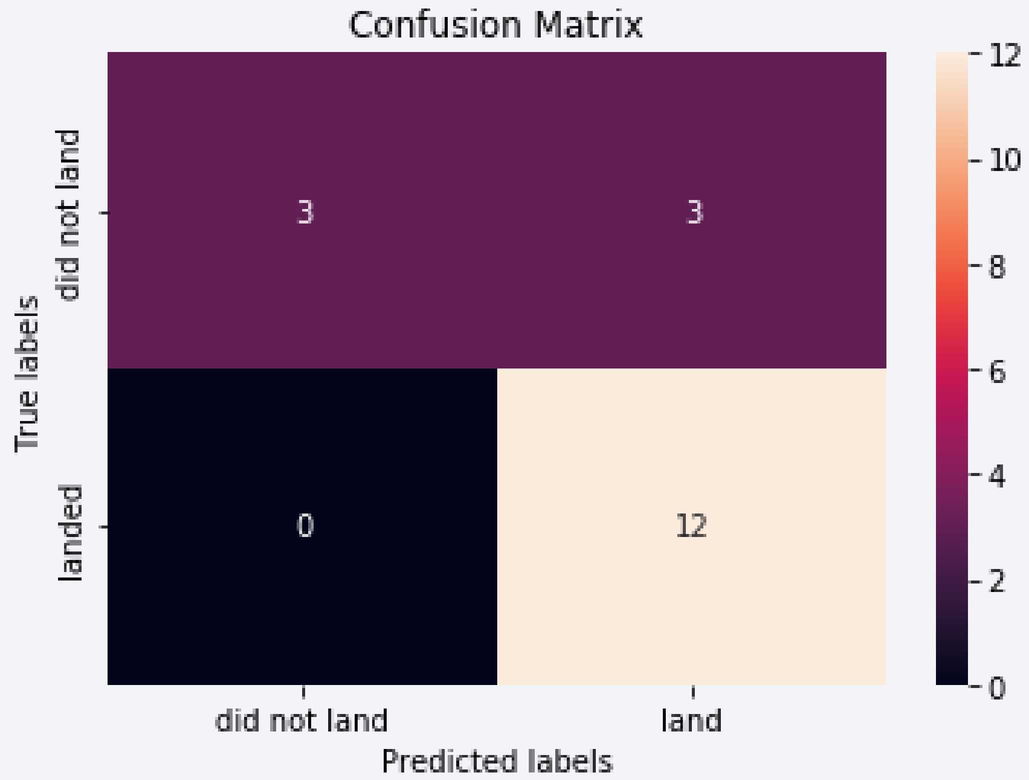
Type II error : The predicted value is negative but its positive

The predicted value is Negative and its Negative

A good model is one which has high true positive (TP) and true negative (TN) rates, while low FP and FN rates. Here TP%=0.222 and TN%=0.611, total accuracy (TP%+TN%)= 83.3%

GitHub Link

Confusion Matrix of Logistic, SVM and KNN



GitHub Link

Identical confusion matrix obtained for these 3 models and TP%=0.167 and TN%=0.667, total accuracy (TP%+TN%)= 83.3%

Conclusions (I)

- With the increase in the flight number, the success rate increases at all 3 launch sites.
- CCAFS SLC40 is suitable for light and medium payloads (0-6500 kg). And KSC LC39A has relatively high successful launches for medium payload (e.g 5000kg-7000kg).
- The orbits of ES-L1, GEO, HEO, SSO have the highest success rate.
- For LEO orbit, the success launch rate increases with the number of flights.
- With heavy payloads the successful landing or positive landing rate increases for PO, LEO and ISS.
- Success rate keeps increasing from 2013 to 2020.

Conclusions (II)

- There are 3 launch sites in FL and 1 launch site in CA, among which KSC LC-39A launch site in Florida has the highest success rate of 76.9%.
- KSC LC-39A is in close proximity to highway, railway and coastline, whereas it keeps certain distance away from cities.
- For light payload, the launch success rate is high, whereas for heavy payload (>5000kg), the launch success rate is low.
- Decision tree models gives rise to the highest accuracy of 0.8857 using train data as the best performing model.
- Confusion matrix shows the total accuracy (TP%+TN%)= 83.3% for decision tree models.

Appendix

- All the Jupiter notebook files have been uploaded to GitHub https://github.com/xgeng2021/XiGeng2021.
- No extra supplemental information available.

Thank you!