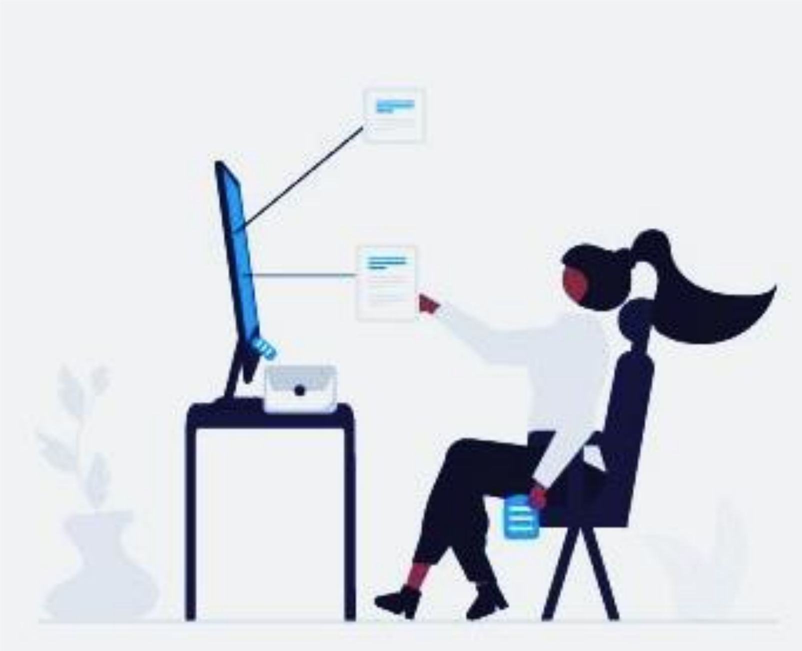
IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Kai-Wei Chang
2023/02/16

Outline

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



Executive Summary

Methodologies

- SpaceX launch data is collected via SpaceX REST API and web scrapping.
- Exploratory data analysis, visual analytics to identify features affecting landing success
- Multiple classification model built via machine learning to evaluate predictability of success landing based on the relevant features

Results

- As of 2017, the success landing rate has improved to 80%
- Missions to SSO orbit with light payload mass increase success landing rate
- Booster type FT should be used for the light payload mass missions
- Launch Site KSC LC-39A is likely give better launch success rate

Introduction

Background

Space X has advantages in space missions at lower cost compare with its competitors. A noticeable advantage is that SpaceX can recover and reuse its launch rockets at first stage. Successful recovery on first stage greatly reduce the launch cost for space mission. We, Space Y, is a competitor of using similar technology. We seek for business strategies that serve for our advantages over Space X.

Objective

Determine features that impact the success recovery of first stage from Space X. Therefore device our business approaches in optimizing the features and bring our cost per mission competitive with Space X.

Section 1

Methodology

Methodology

Executive Summary

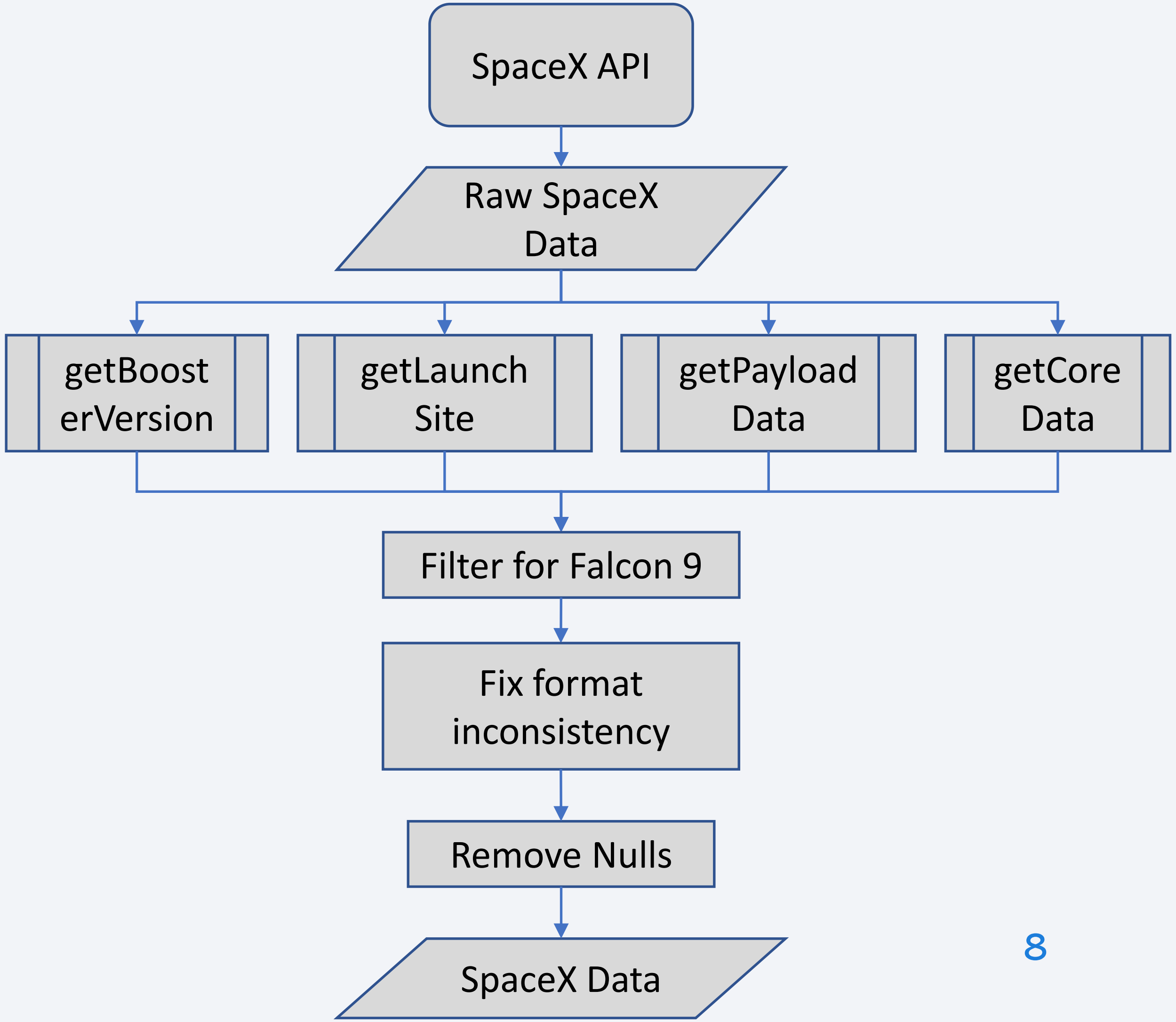
- Data collection methodology:
  - SpaceX launch data was gathered from SpaceX REST API.
  - Web scraping Falcon 9 and Falcon Heavy Launches Records from Wikipedia
- Perform data wrangling
  - Apply API to call and gather data, filter for Falcon 9 launches, and remove null data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Classification models including Logistic Regression, SVM, Decision Tree, and KNN are tested and compared. Models are optimized with parameter cross validation strategy.

Data Collection

- Data Collection – SpaceX API
- Data Collection – Scraping
- Data Wrangling

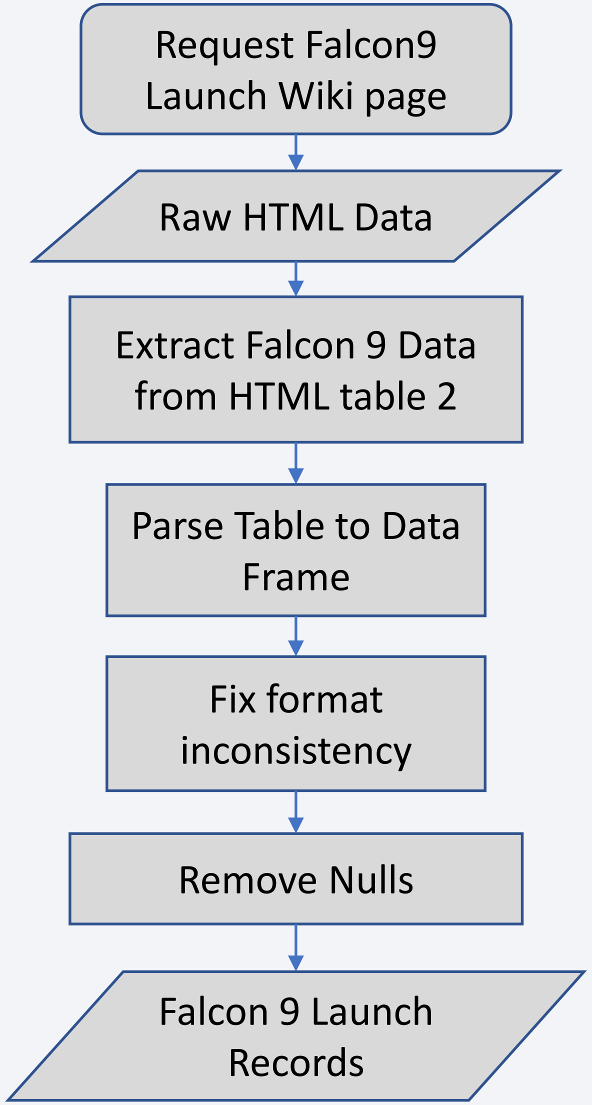
Data Collection – SpaceX API

- Raw data obtained using SpaceX API
- Key information including Booster version, Launch Site, Payload, and Core data are collected via data ID from the raw data
- Data is sampled for Falcon 9 rocket
- Data with missing formation is removed
- GitHub URL of the completed SpaceX API calls: https://github.com/kwchang101/IBM-Course-Projects/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Raw wiki page of Falcon 9 launch record is collected
- Parse html table of launch record to data frame
- Fix format inconsistency and remove nulls
- GitHub URL of the completed webs craping: https://github.com/kwchang101/IBM-Course-Projects/blob/main/jupyter-labs-webscraping.ipynb



Request Falcon9 Launch Wiki page

Raw HTML Data

Extract Falcon 9 Data from HTML table 2

Parse Table to Data Frame

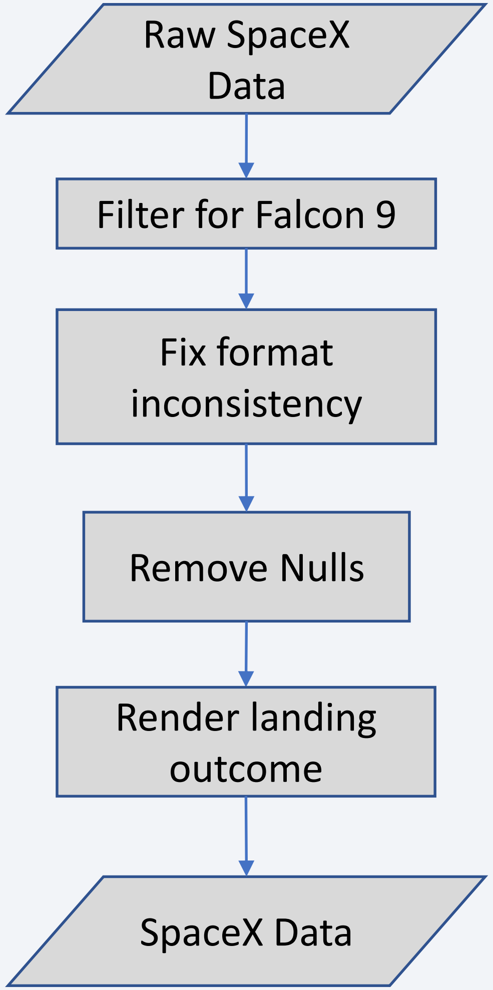
Fix format inconsistency

Remove Nulls

Falcon 9 Launch Records

Data Wrangling

- Space X data for Falcon 9 is sampled
- Fix format inconsistency
- Remove null records
- Preliminary investigate space X data for key features
- Render landing outcomes to Success (1) vs. Fail (0)
- Data wrangling process: https://github.com/kwchang101/IBM-Course-Projects/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- Rocket recovery may be affected by operation orbit (payload altitude)
- Space mission by orbit is investigated for potential correlation
- Payload mass in association with orbit of operation is investigated to narrow down the groups of orbital missions.
- Identify launch sites that have success recovery rate in according to payload mass
- Launch site in correlation with payload mass is also investigated
- EDA with data visualization: https://github.com/kwchang101/IBM-Course-Projects/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- From the Data base, in looked into primarily the booster details in association with payload mass.
- List of Booster capacities are listed
- Boosters with success landing at light weight payload mass (4000-6000 KG) are investigated
- Investigate past landing outcomes are investigated for data reliability to present scenario
- EDA with SQL: https://github.com/kwchang101/IBM-Course-Projects/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

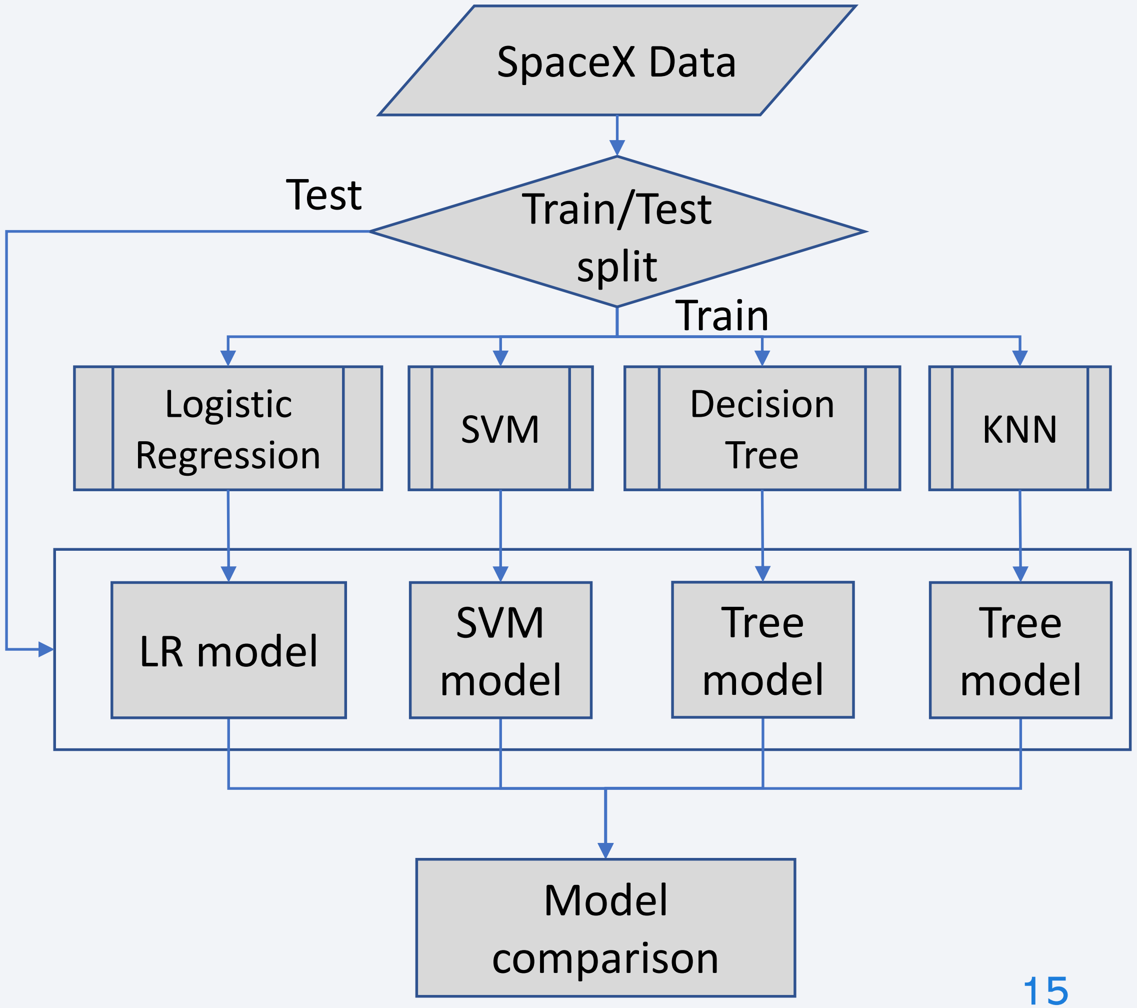
- The geological map marked with launch sites (circles). Success and fail recovery attempts are included. Blue lines indicate the distance (red) to nearest land marks (coastline, railway, highway, and city)
- Geographical locations of launch sites are associated with factors such as area availability, impacts to nearby city, etc. are not easily presentable by tables and figures. Interactive map is included for visual investigation on these features
- Interactive map with Folium map: https://nbviewer.org/github/kwchang101/IBM-Course-Projects/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Success counts for each launch site is addressed in pie chart
- Independent success rate is calculated for each selected launch site
- Scatter plot is implemented for visualization of success rate by each booster version, for each launch site
- A slider bar is implemented to track success rate by a range of payload mass
- This plot attempt is added for easy track the relationships success rate by range of payload mass, in association with launch sites
- Plotly Dash lab: https://github.com/kwchang101/IBM-Course-Projects/blob/main/spacex_dash_app.py

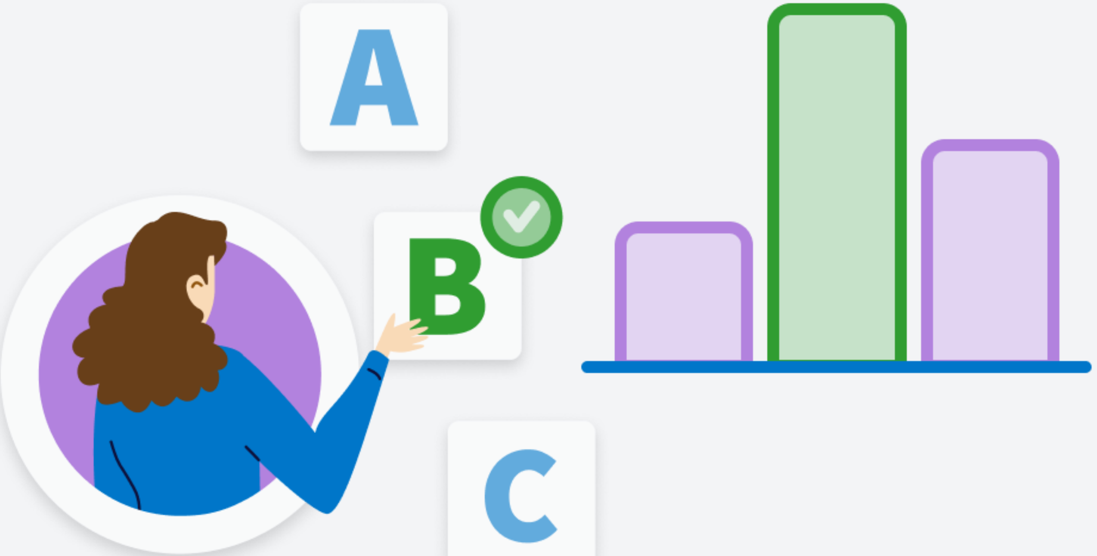
Predictive Analysis (Classification)

- Classification model is built based on the features of interest, based on the previous analysis
- The predictivity of success outcomes indicate the features, parameters, and the models best fit to our analysis
- Classification models including Logistic Regression, SVM, Decision Tree, and KNN are tested and compared.
- Models are optimized with parameter cross validation strategy.
- Predictive analysis: https://github.com/kwchang101/IBM-Course-Projects/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



Results

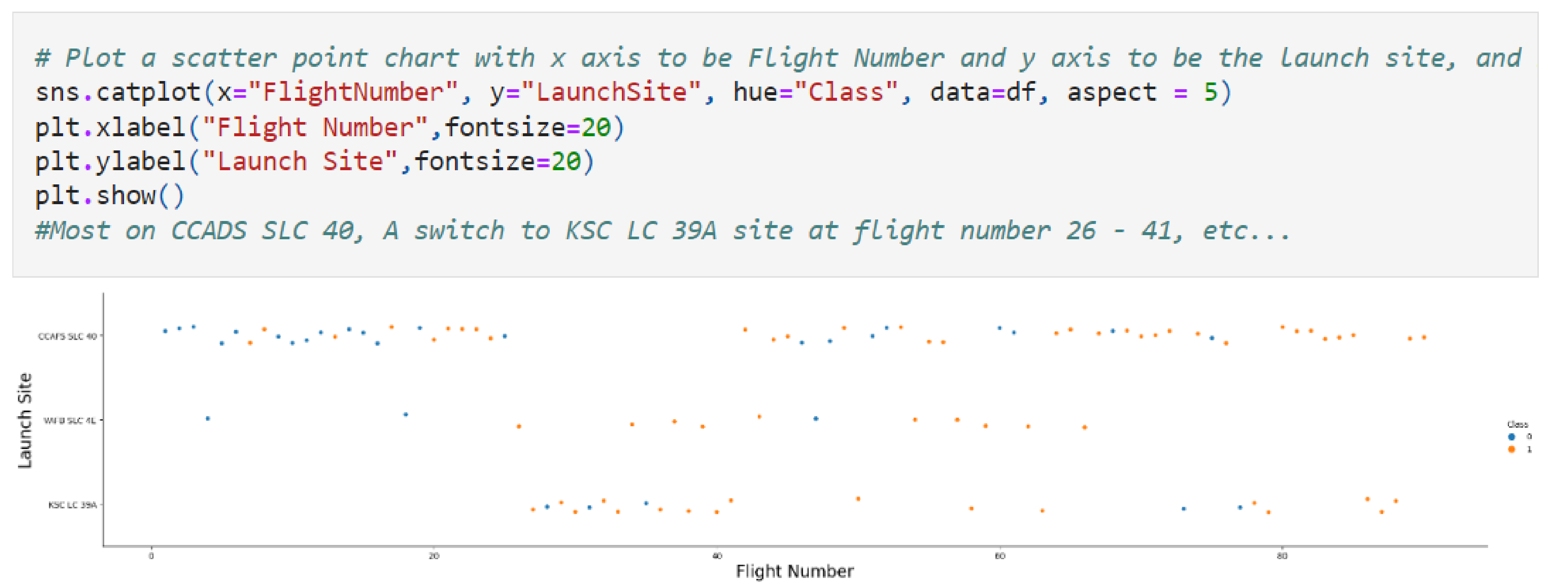
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



Section 2

Insights drawn from EDA

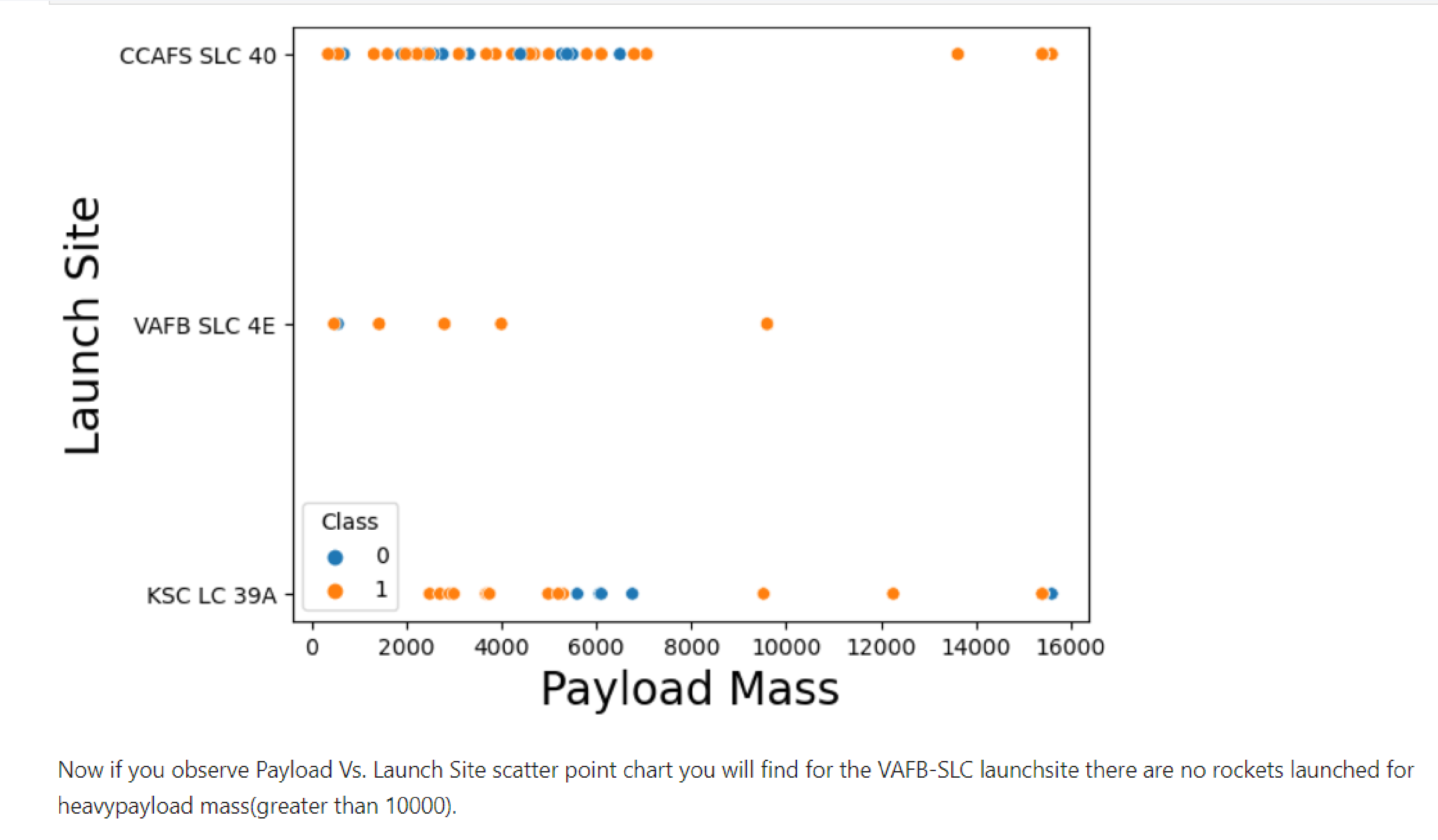
Flight Number vs. Launch Site



```
# Plot a scatter point chart with x axis to be Flight Number and y axis to be the launch site, and
sns.catplot(x="FlightNumber", y="LaunchSite", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
#Most on CCADS SLC 40, A switch to KSC LC 39A site at flight number 26 - 41, etc...
```

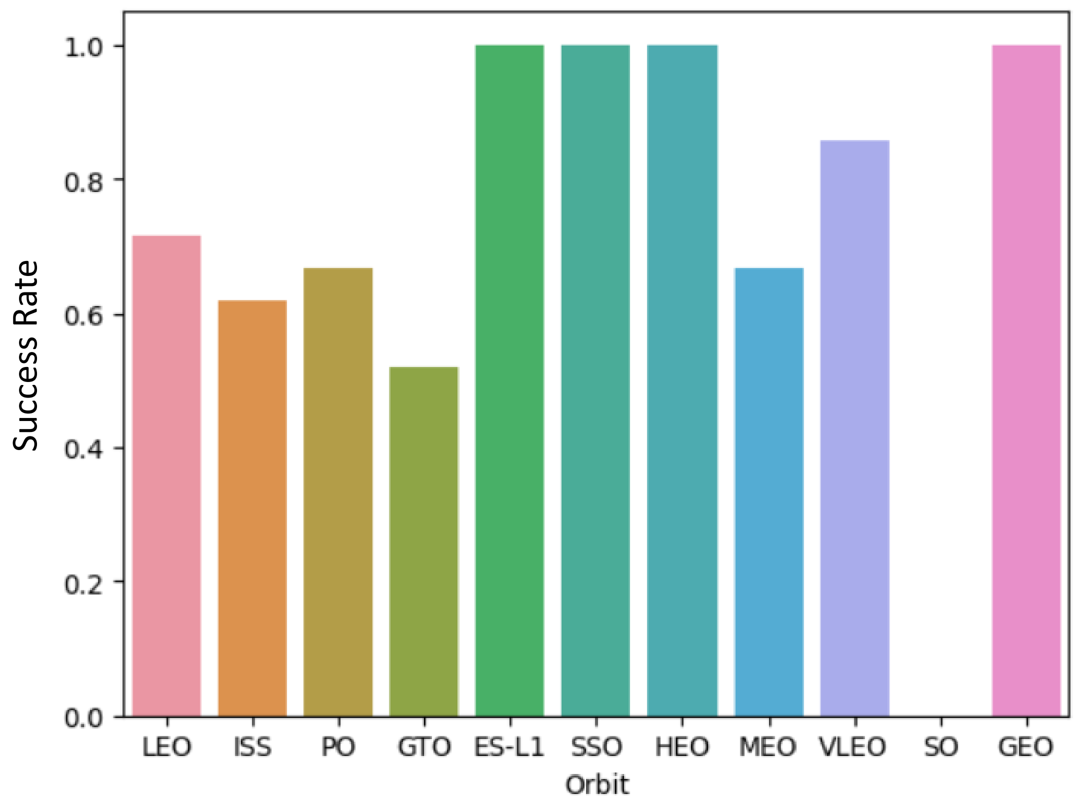
Most of launch sites occur on CCADS SLC 40, where later flight number have inproved success rate. Also, an increase access using KSC LC 39A launch site is found at flight number 26-41

Payload vs. Launch Site



Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

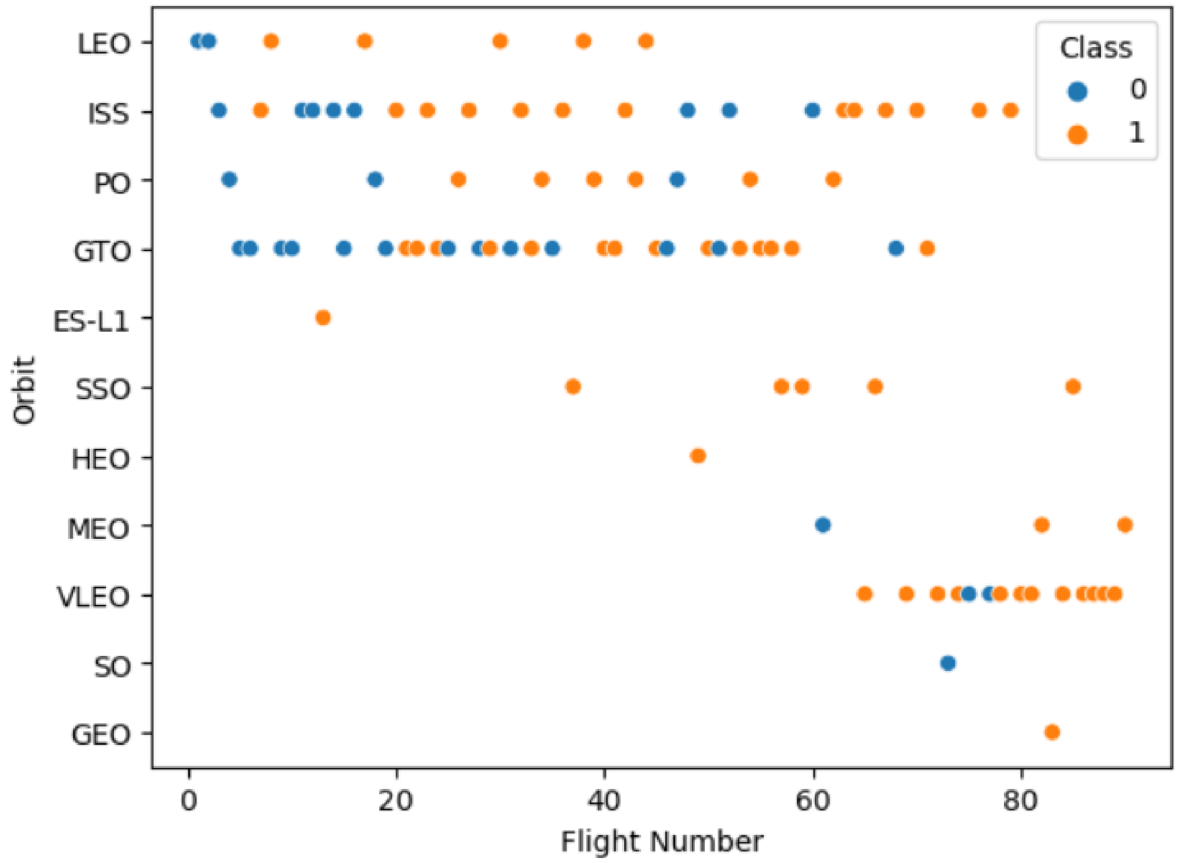
Success Rate vs. Orbit Type



Analyze the ploted bar chart try to find which orbits have high sucess rate.

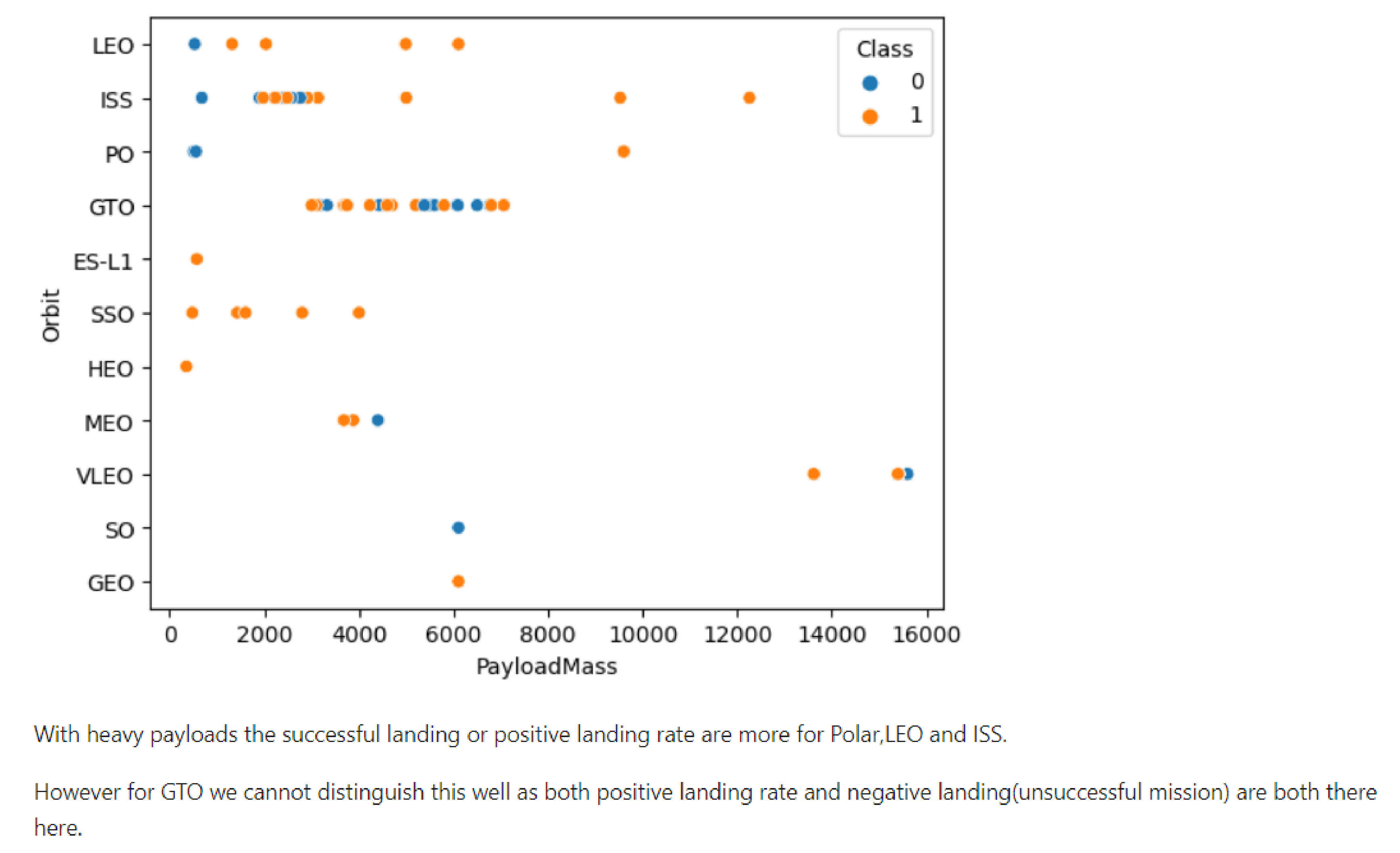
ES-L1, SSO, HEO, and GEO orbits have high success rate. SSO has multiple successes.

Flight Number vs. Orbit Type



You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

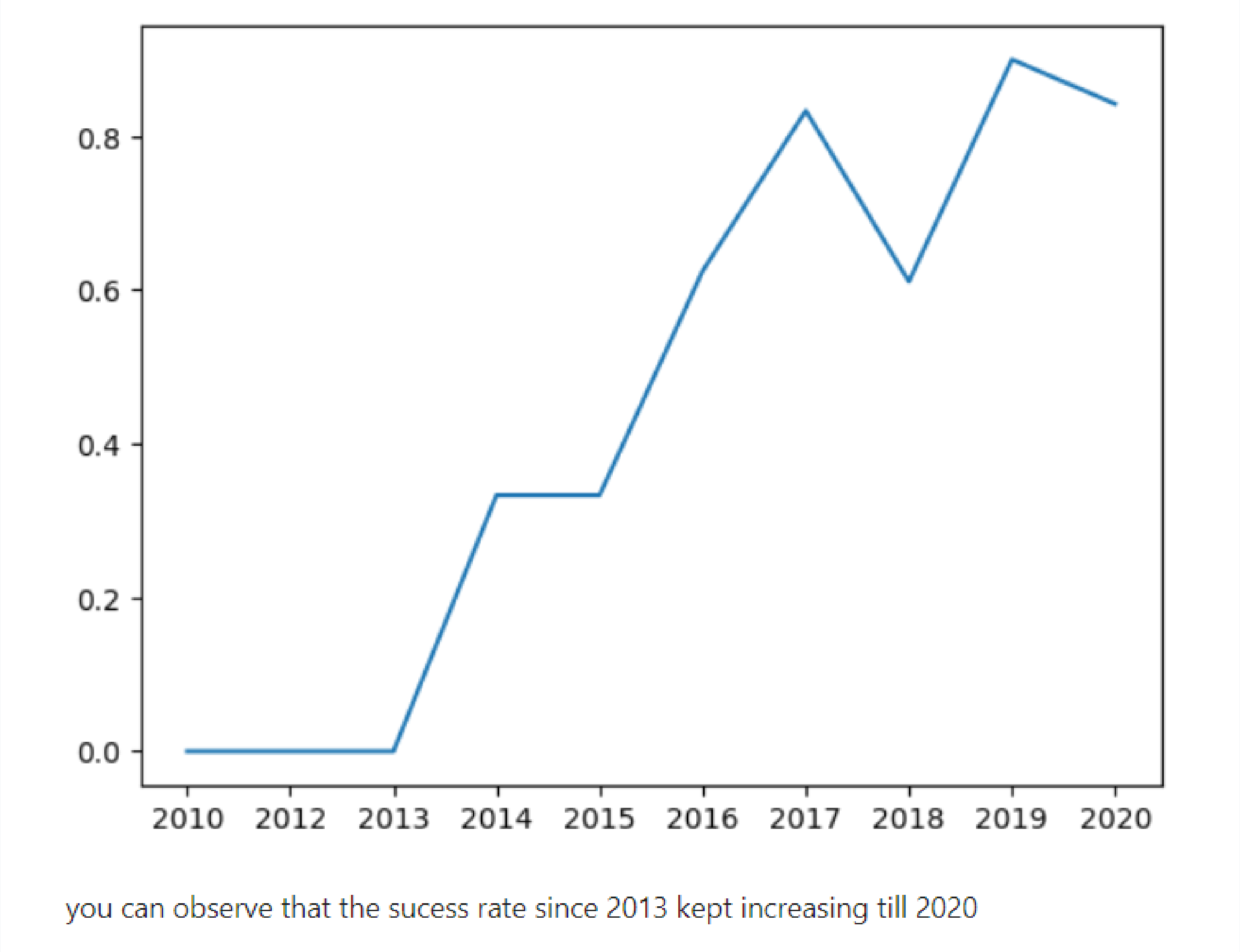
Payload vs. Orbit Type



With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

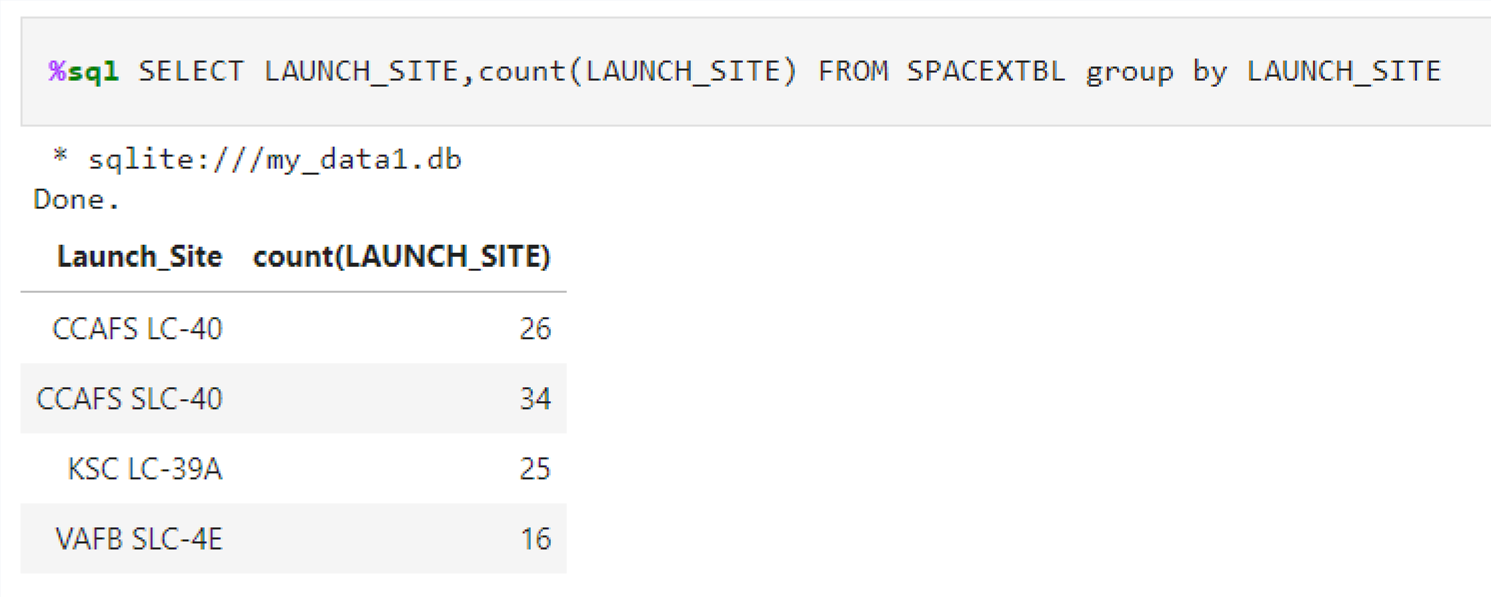
However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend



you can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names



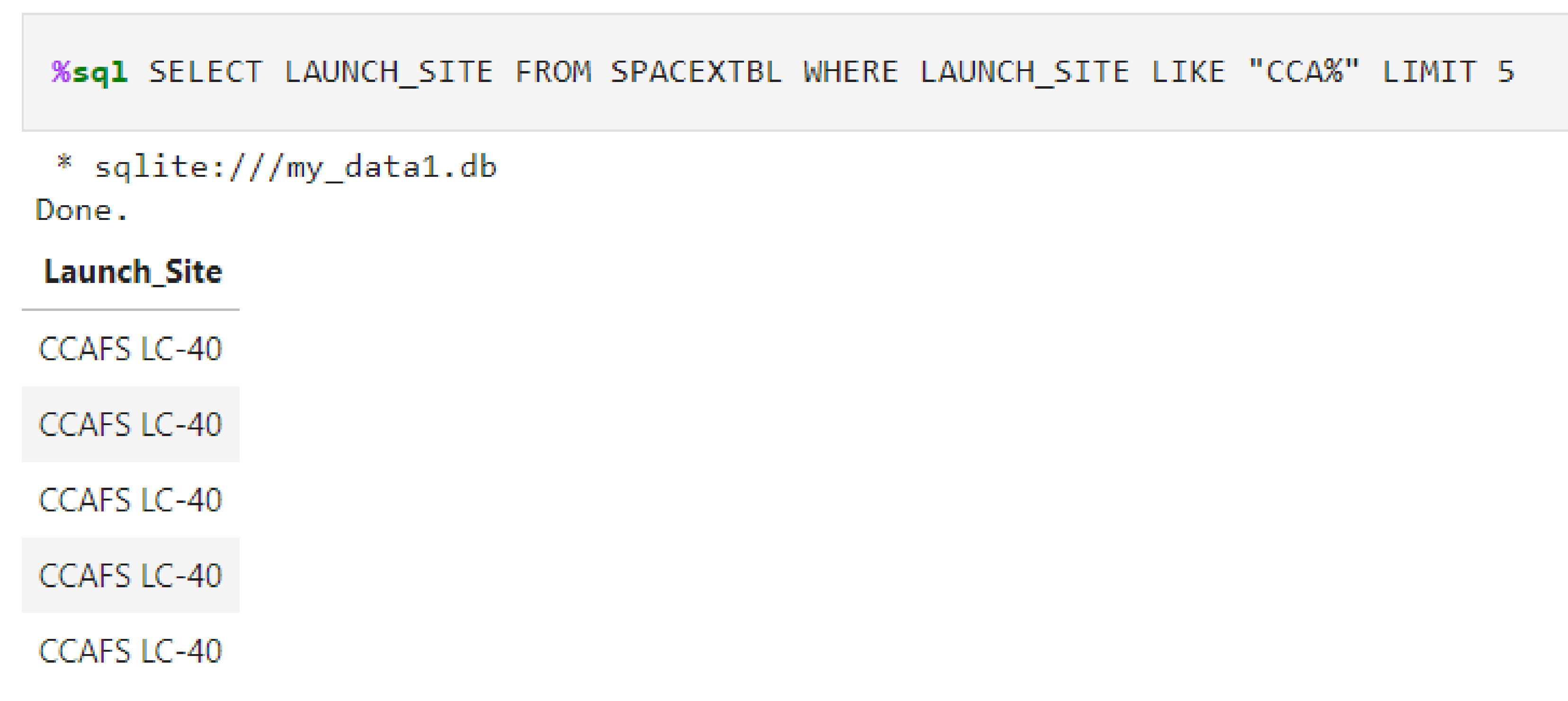
```
%sql SELECT LAUNCH_SITE,count(LAUNCH_SITE) FROM SPACEXTBL group by LAUNCH_SITE
```

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

Launch_Site	count(LAUNCH_SITE)
CCAFS LC-40	26
CCAFS SLC-40	34
KSC LC-39A	25
VAFB SLC-4E	16

- Four unique launch sites are presented in the data provided
- Number of record associate with each launch site is also reported, as an addition

Launch Site Names Begin with 'CCA'



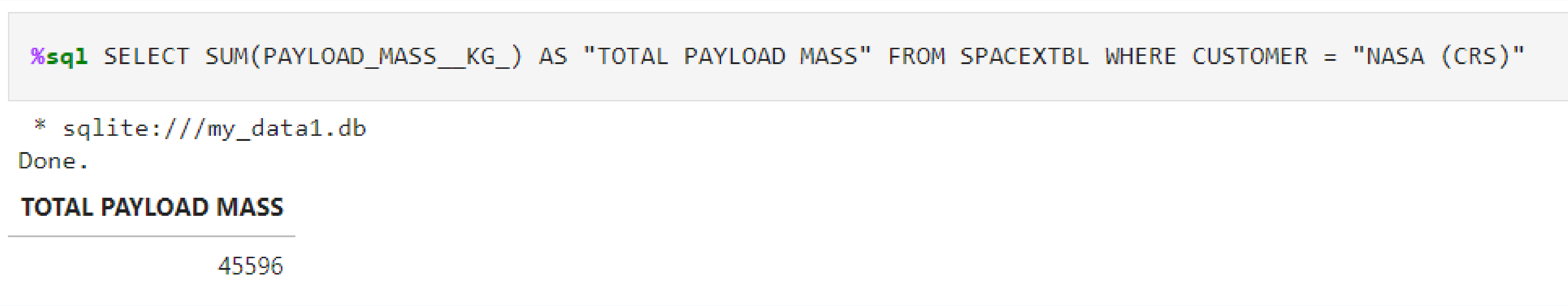
```
%sql SELECT LAUNCH_SITE FROM SPACEXTBL WHERE LAUNCH_SITE LIKE "CCA%" LIMIT 5
```

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

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

- Given CCAFS LC-40 and CCAFS SLC-40 both begin with “CCA”, the top 5 of the list all address to CCAFS LC-40 imply this launch site is first used in the operation

Total Payload Mass



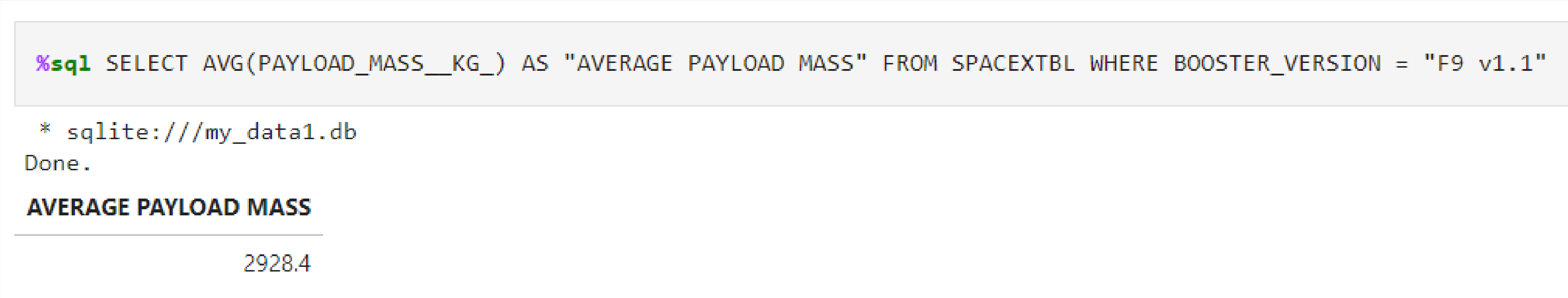
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS "TOTAL PAYLOAD MASS" FROM SPACEXTBL WHERE CUSTOMER = "NASA (CRS)"
```

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

TOTAL PAYLOAD MASS
45596

- Total payload mass by NASA (CRS) may be indicative to the “value” of NASA as a customer, however, number of missions, cargo size, mission numbers are yet to be addressed, and might not be relevant with the success rate.

Average Payload Mass by F9 v1.1



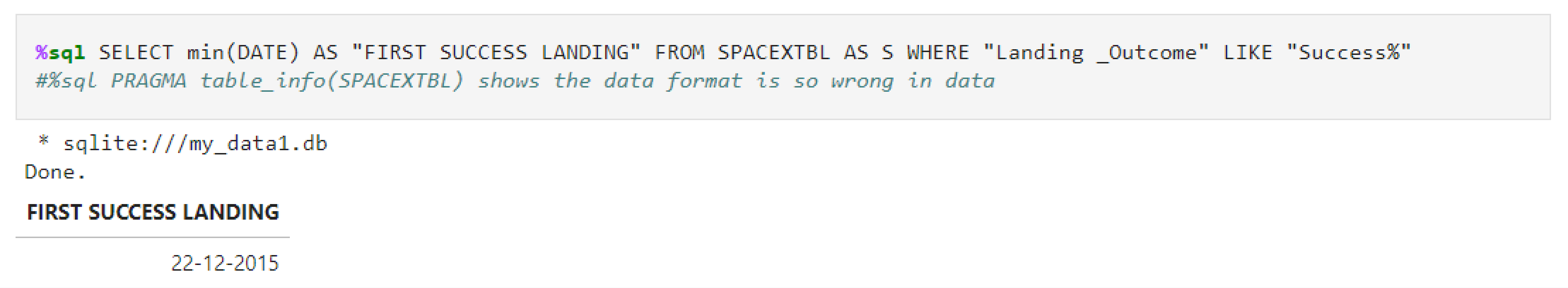
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS "AVERAGE PAYLOAD MASS" FROM SPACEXTBL WHERE BOOSTER_VERSION = "F9 v1.1"
```

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

AVERAGE PAYLOAD MASS
2928.4

- Booster version F9 v1.1 in average has a payload mass of 2928.4 KG, implying this booster is more used for light payload mass missions.

First Successful Ground Landing Date



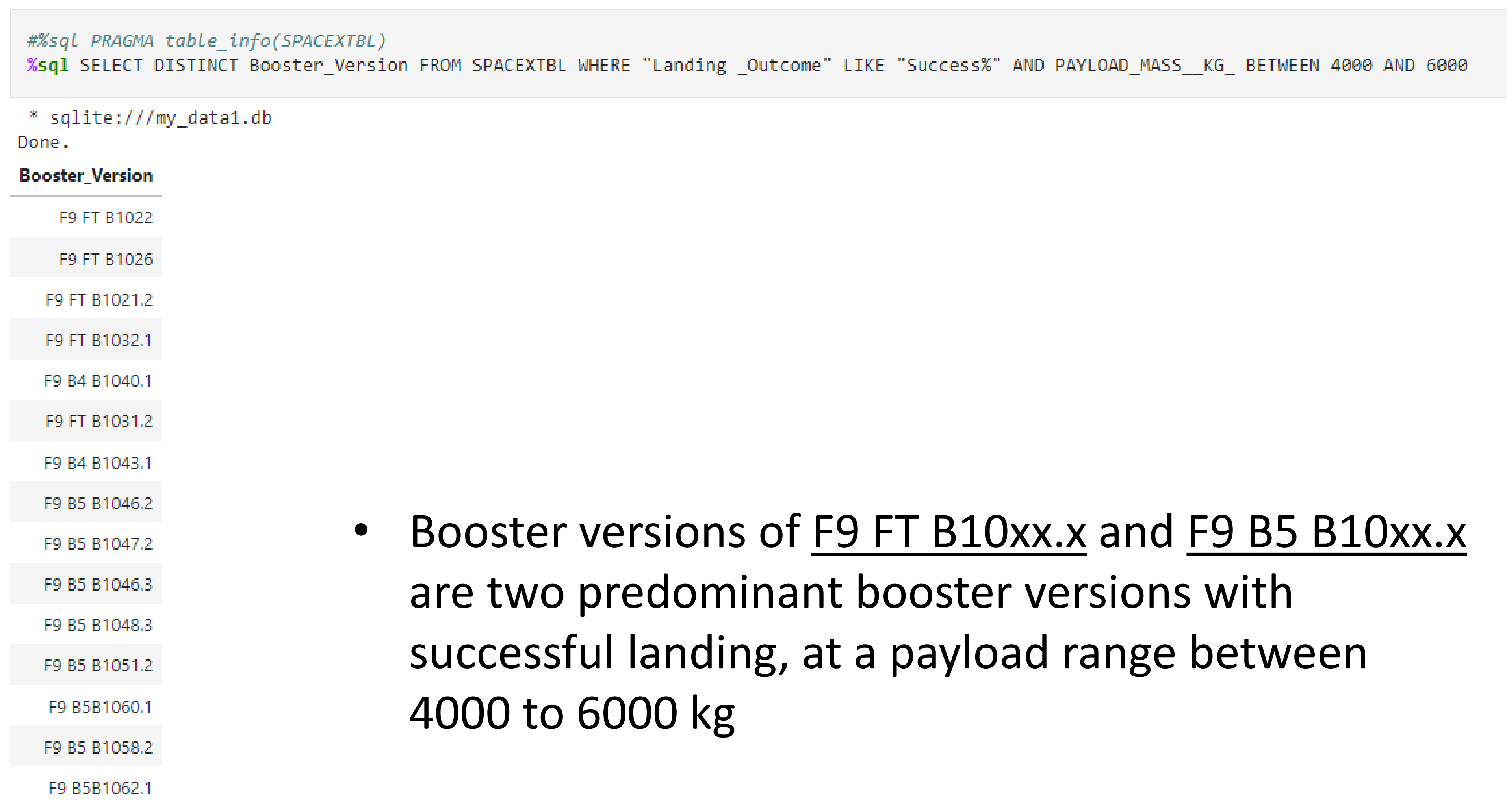
```
%sql SELECT min(DATE) AS "FIRST SUCCESS LANDING" FROM SPACEXTBL AS S WHERE "Landing _Outcome" LIKE "Success%"
#%sql PRAGMA table_info(SPACEXTBL) shows the data format is so wrong in data
```

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

FIRST SUCCESS LANDING
22-12-2015

- Successful ground landing by 22-12-2015 would mark a milestone of the maturation of ground landing technology

Successful Drone Ship Landing with Payload between 4000 and 6000



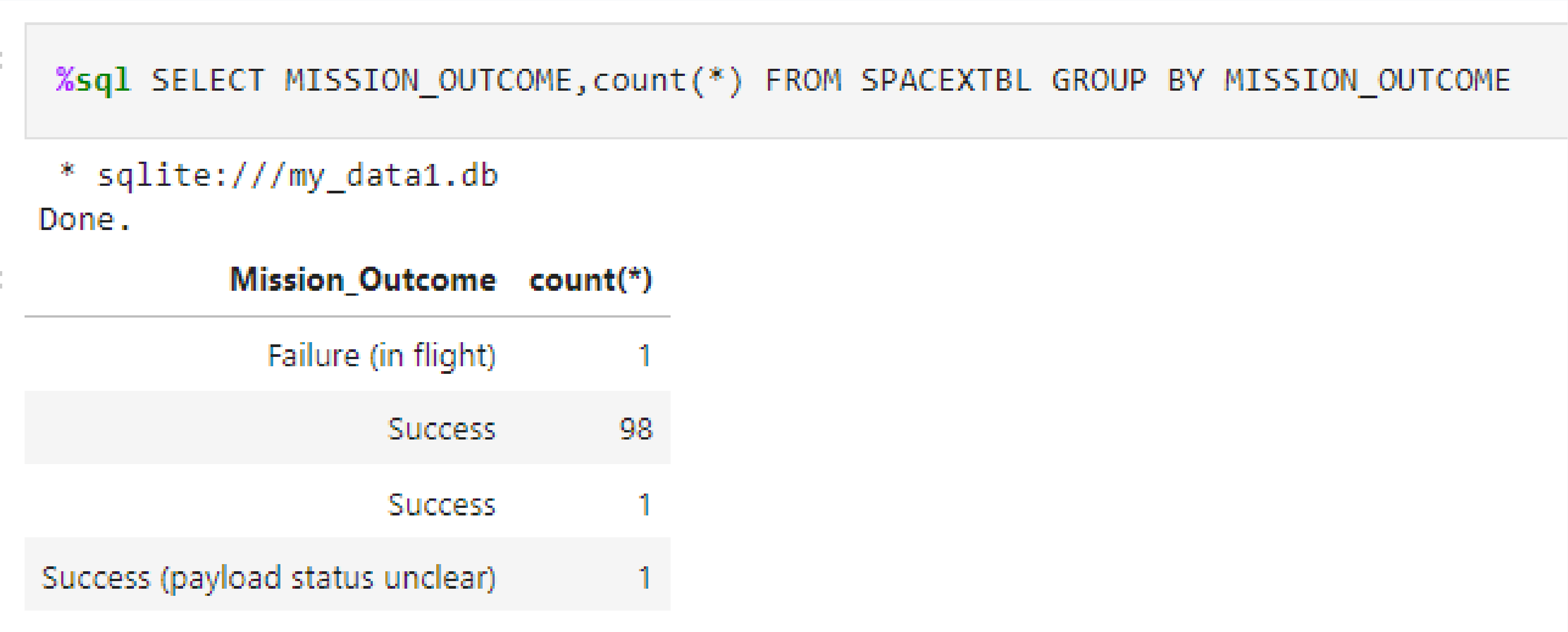
```
#%sql PRAGMA table_info(SPACEXTBL)
%sql SELECT DISTINCT Booster_Version FROM SPACEXTBL WHERE "Landing _Outcome" LIKE "Success%" AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
```

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

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 B4 B1043.1
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1046.3
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

- Booster versions of F9 FT B10xx.x and F9 B5 B10xx.x are two predominant booster versions with successful landing, at a payload range between 4000 to 6000 kg

Total Number of Successful and Failure Mission Outcomes



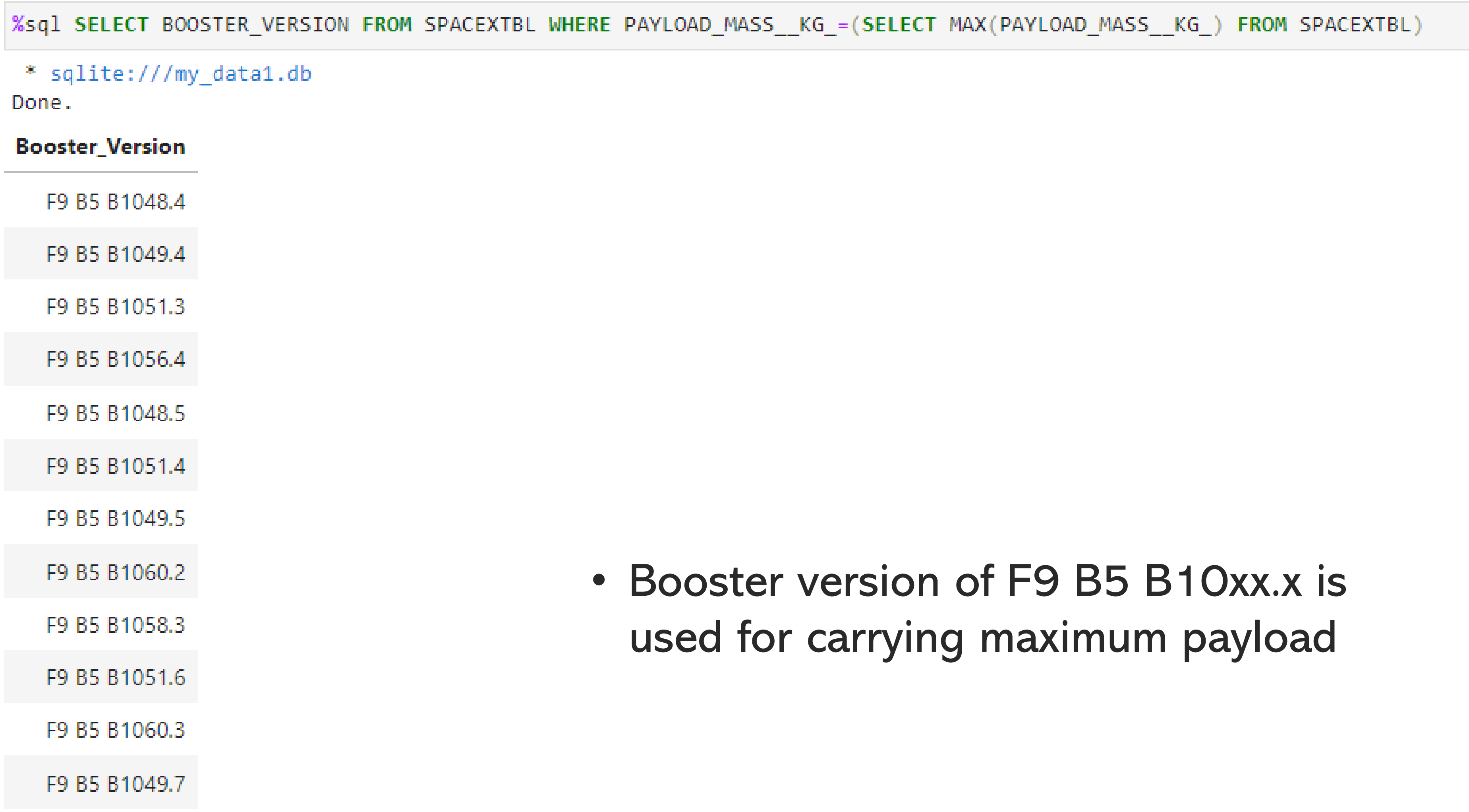
```
%sql SELECT MISSION_OUTCOME,count(*) FROM SPACEXTBL GROUP BY MISSION_OUTCOME
```

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

Mission_Outcome	count(*)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- Most missions are success disregard to the successfulness of drone ship landing.
- Some records may have format errors that lead to two “Success” categories

Boosters Carried Maximum Payload



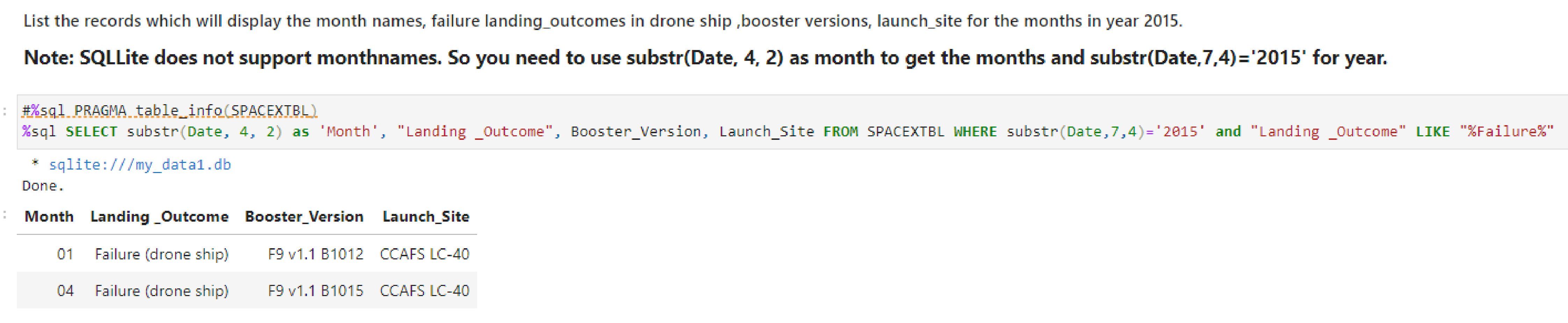
```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_=(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

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

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

- Booster version of F9 B5 B10xx.x is used for carrying maximum payload

2015 Launch Records



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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

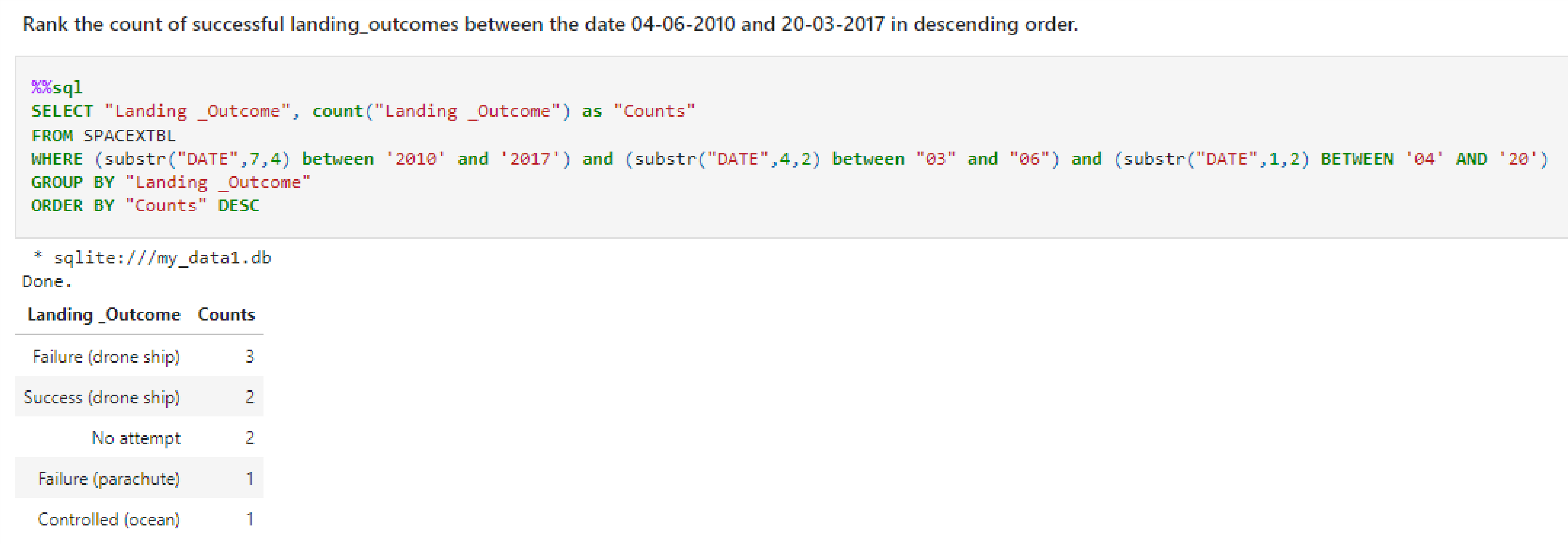
```
#%sql PRAGMA table_info(SPACEXTBL)
%sql SELECT substr(Date, 4, 2) as 'Month', "Landing _Outcome", Booster_Version, Launch_Site FROM SPACEXTBL WHERE substr(Date,7,4)='2015' and "Landing _Outcome" LIKE "%Failure%"
```

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

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

- In year 2015, two failure landing occurred on January and April. All occur at launch site CCADS LC-40, using booster version F9 v1.1 B101x

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



Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

```
%%sql
SELECT "Landing _Outcome", count("Landing _Outcome") as "Counts"
FROM SPACEXTBL
WHERE (substr("DATE",7,4) between '2010' and '2017') and (substr("DATE",4,2) between "03" and "06") and (substr("DATE",1,2) BETWEEN '04' AND '20')
GROUP BY "Landing _Outcome"
ORDER BY "Counts" DESC
```

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

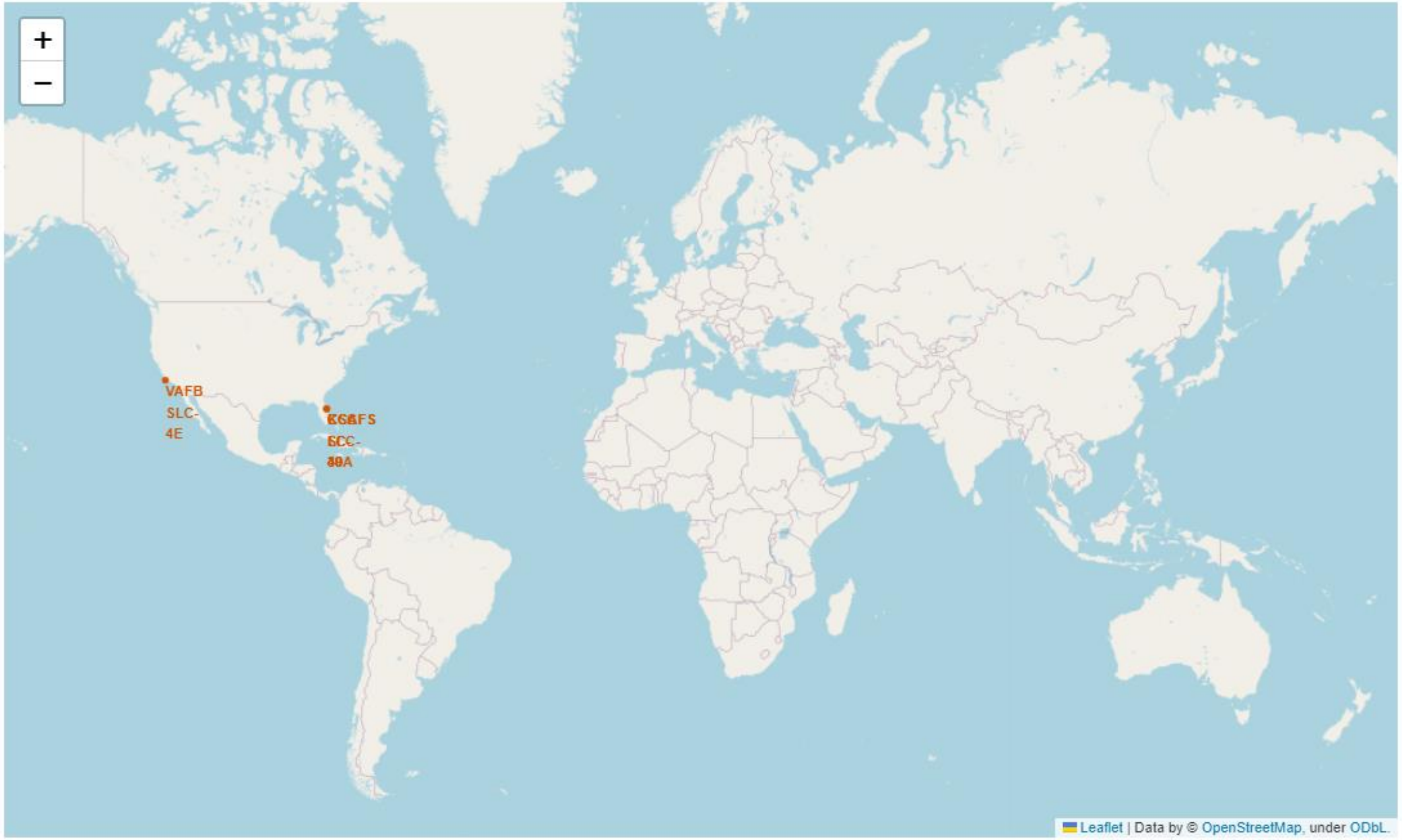
Landing _Outcome	Counts
Failure (drone ship)	3
Success (drone ship)	2
No attempt	2
Failure (parachute)	1
Controlled (ocean)	1

- Launches between 2010-06-04 and 2017-03-02 have relatively high failure counts. Two success incidences occurred.

Section 3

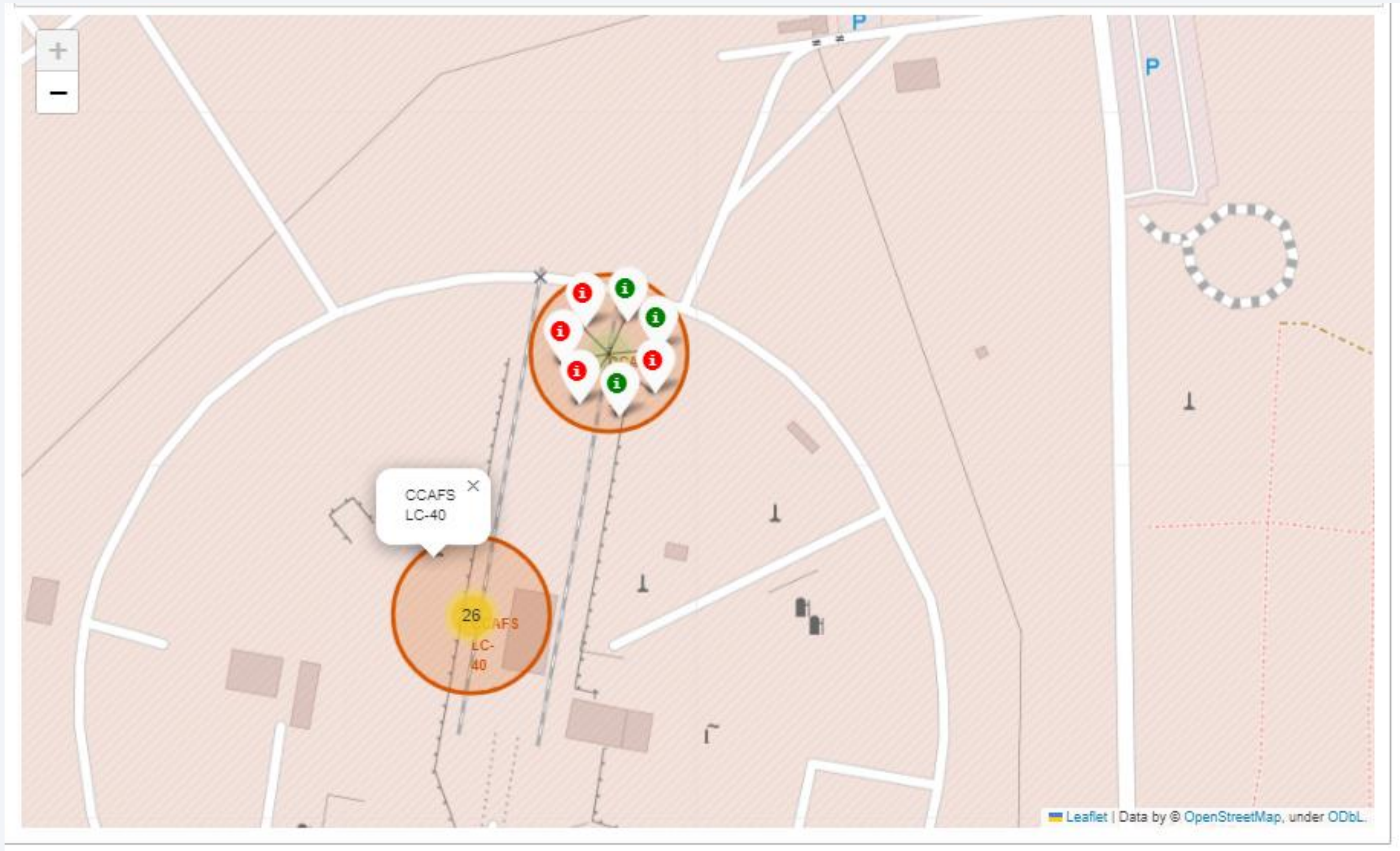
Launch Sites Proximities Analysis

Launch Sites that operate SpaceX missions



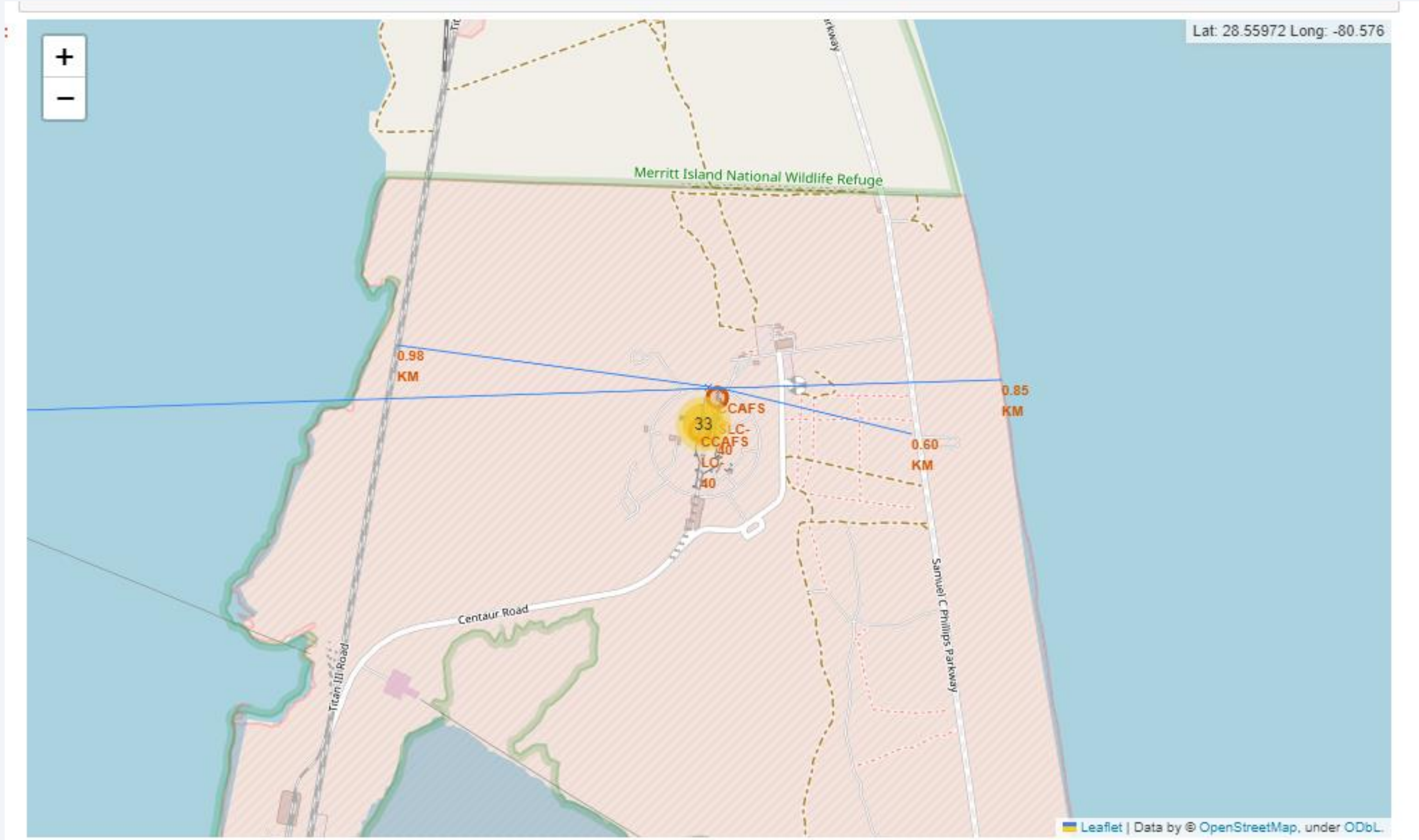
- SpaceX missions takes place at US, where 3 launch sites at coast of Atlantic Ocean, and 1 launch sites at coast of Pacific Ocean

Geographical Landing Outcomes on Launch Sites

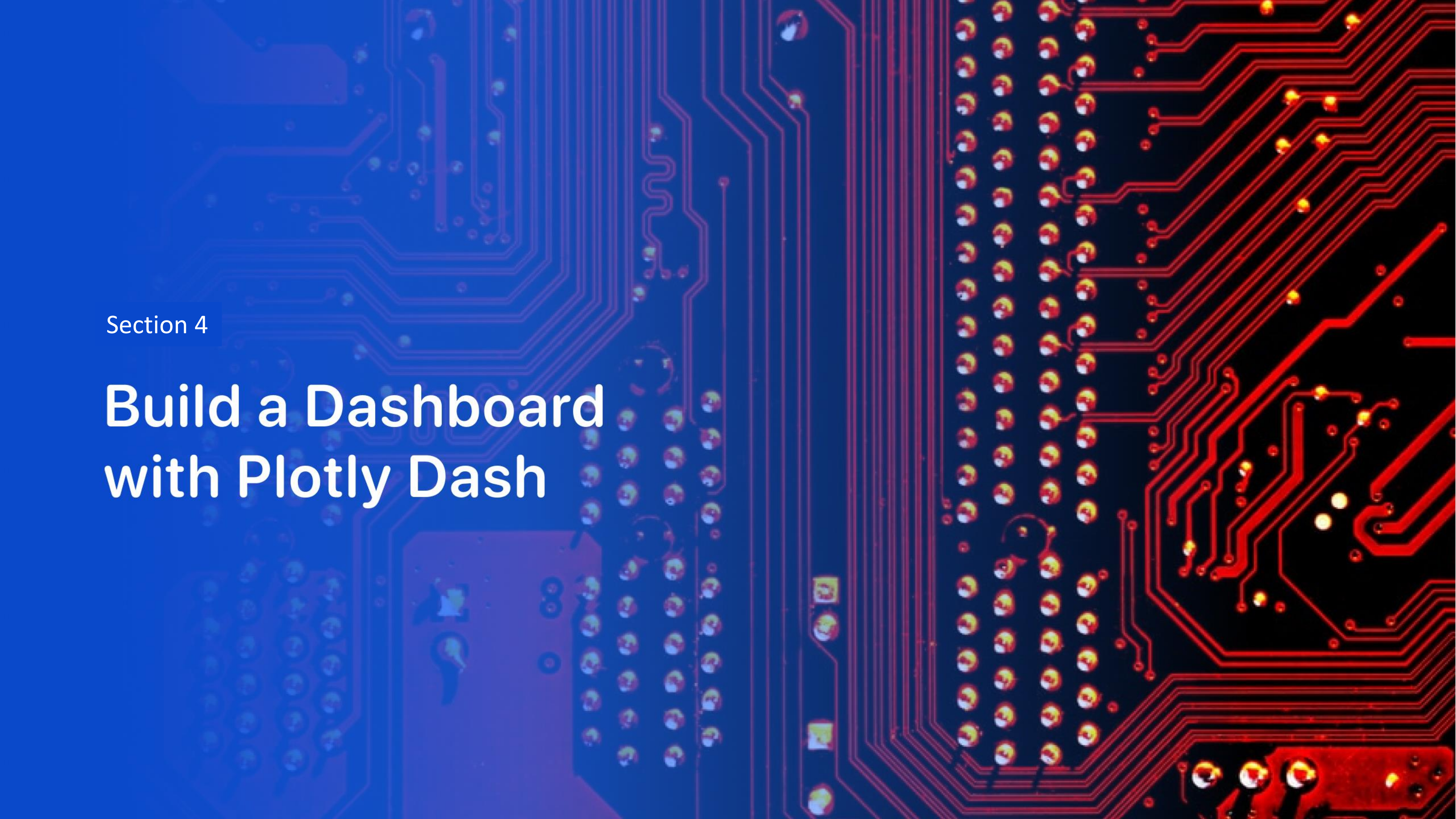


- Landing outcomes from each Space X missions, on each launch sites
- CCAFS LC-40 launch site is used 26 times according to the database

Geographical positioning of launch sites


+
−
Lat: 28.55972 Long: -80.576
Merritt Island National Wildlife Refuge
0.98 KM
0.85 KM
0.60 KM
33
CCAFS SLC-40
CCAFS LC-40
Centaur Road
Titan III Road
Samuel C Phillips Parkway
Leaflet | Data by © OpenStreetMap, under ODbL.

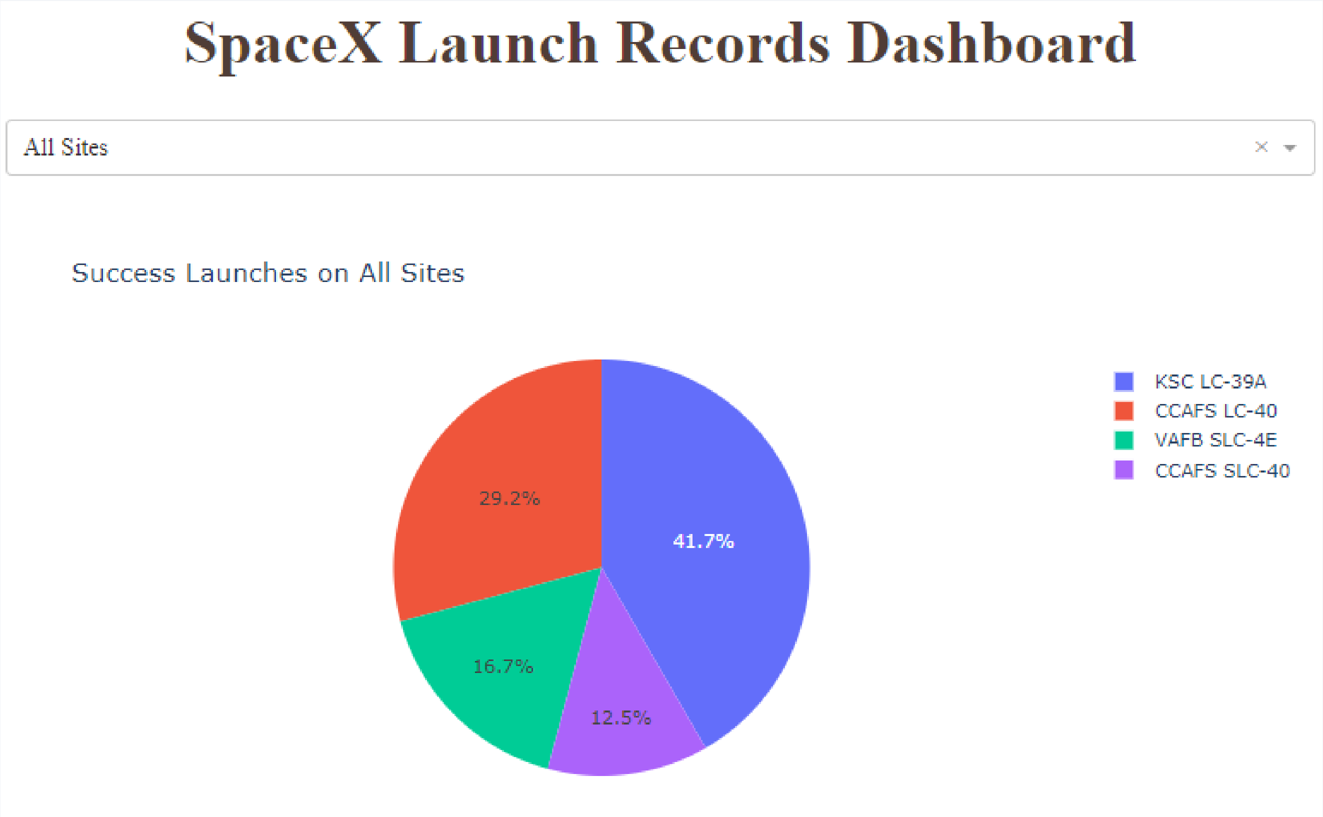
- The launch site is relatively close to coast, with access to land traffics including railway and highway.
- The launch site is far away from cities and crowded places.



Section 4

Build a Dashboard with Plotly Dash

Success Launch Incidences among Launch Sites

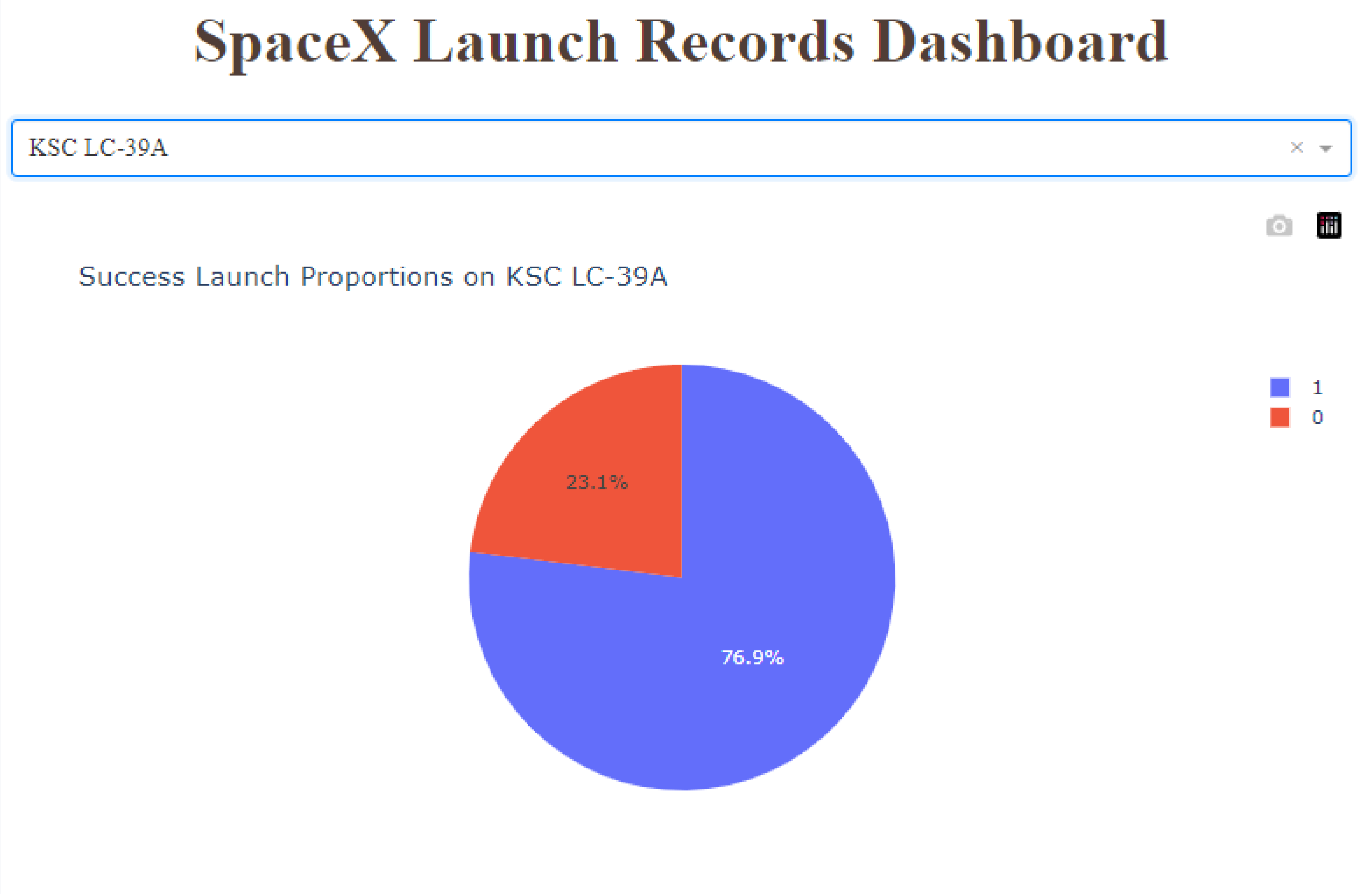


SpaceX Launch Records Dashboard

All Sites × ▾

- KSC LC-39A is composed of most success launch outcomes, follow by CCADS LC-40

Success Launch Rate at KSC LC-39A Launch Site



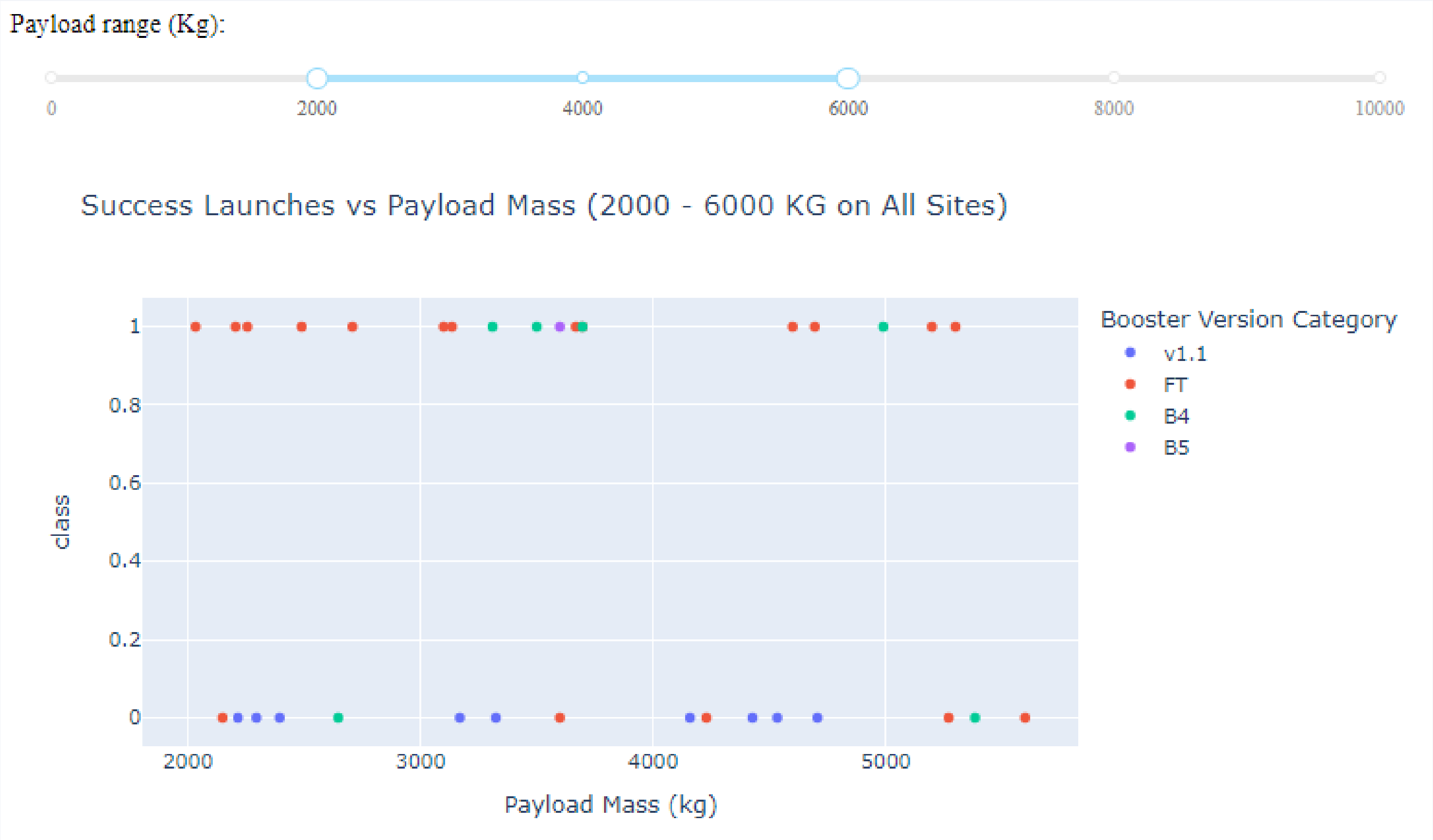
SpaceX Launch Records Dashboard

KSC LC-39A

Success Launch Proportions on KSC LC-39A

- KSC LC-39A has highest successful launch ratio of 76.9%

Booster Type Performance at Payload Range of 2000-6000 KG

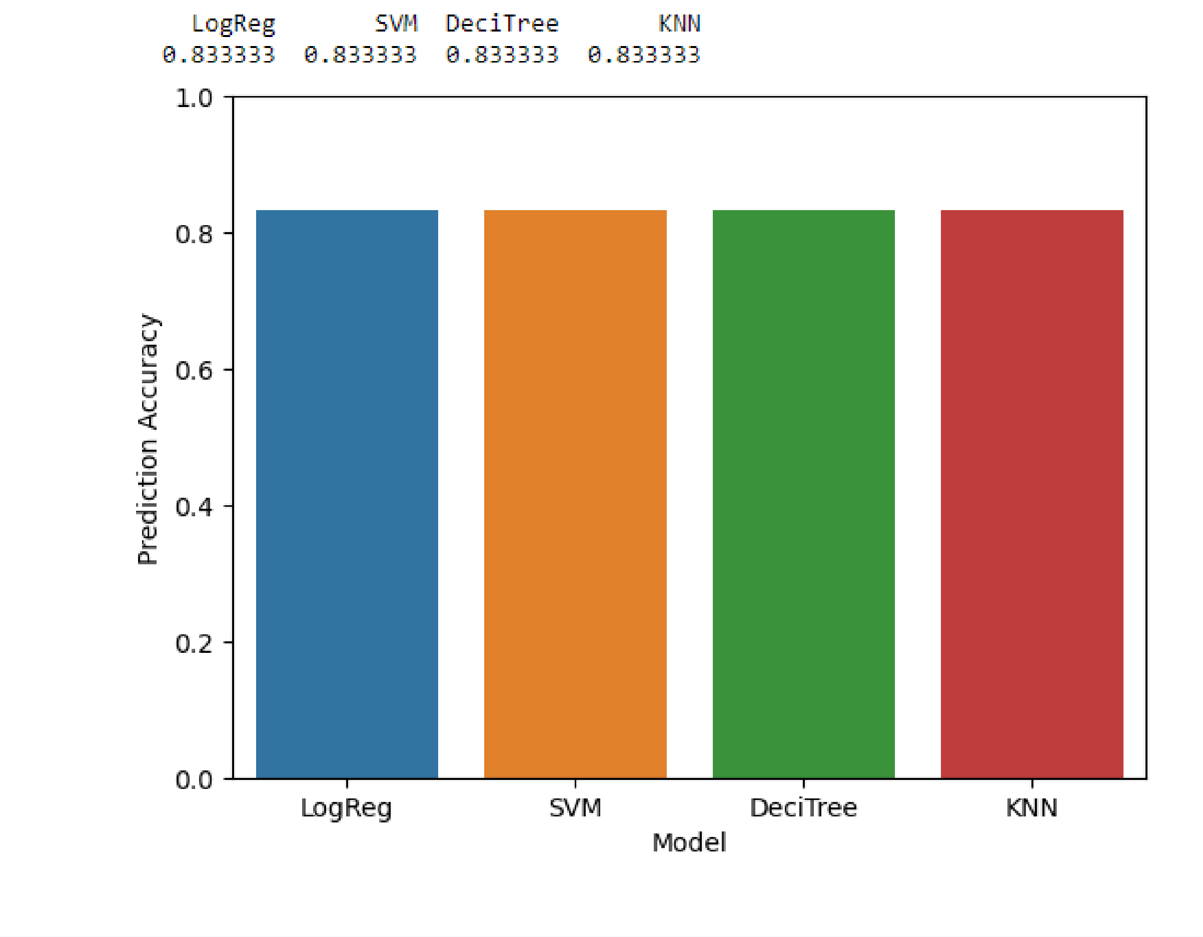


- Booster type FT poses high success launch rate at payload of 2000-6000 KG
- Booster type v1.1 has lowest success launch rate at payload of 2000-6000 KG

Section 5

Predictive Analysis (Classification)

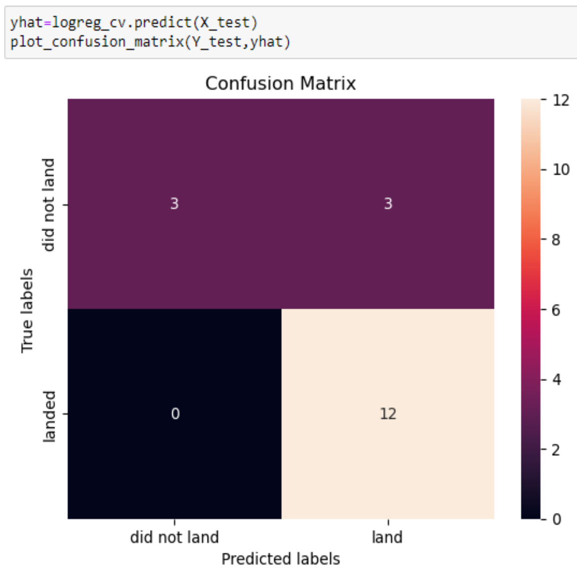
Classification Accuracy



```
  LogReg       SVM  DeciTree       KNN
0.833333  0.833333  0.833333  0.833333
```

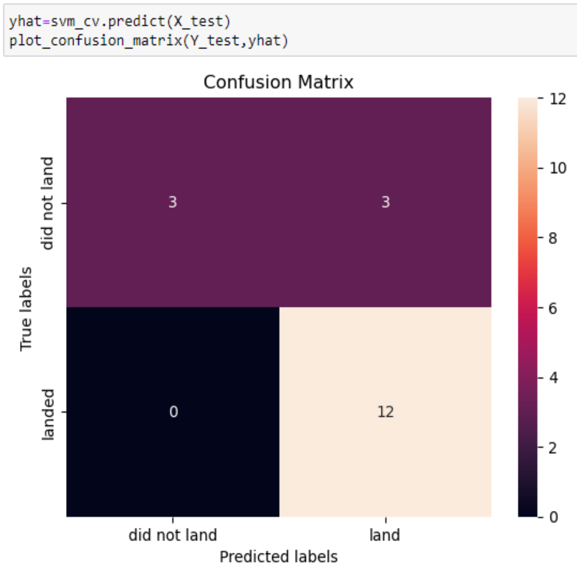
- All 4 tested models: Logistic Regression, SVM, Decision Tree, and KNN gave the same prediction accuracy of 83%

Confusion Matrix



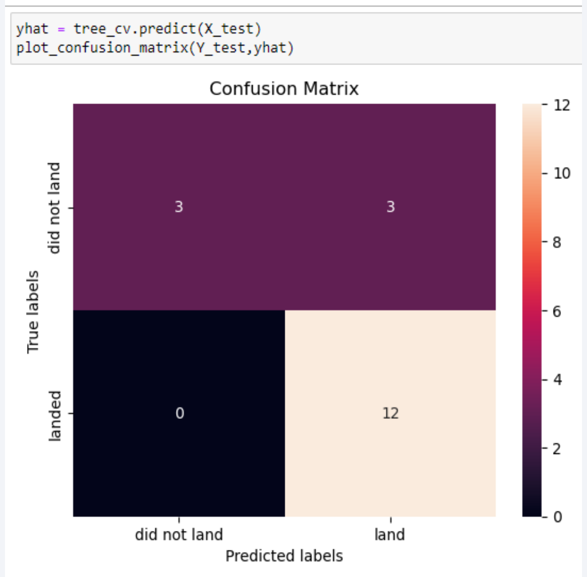
```
yhat=logreg_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```

Logistic Regression



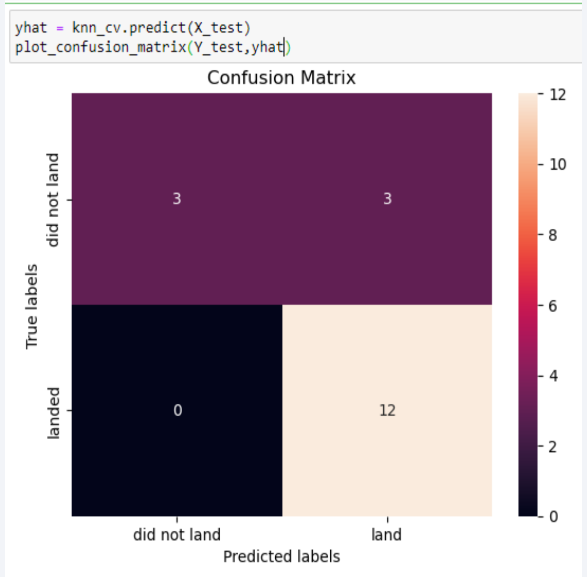
```
yhat=svm_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```

SVM



```
yhat = tree_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```

Decision Tree



```
yhat = knn_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```

KNN

- All 4 classification prediction models gave the same confusion matrices

Conclusions

- As of 2017, the success landing rate has improved to 80%
- Missions to SSO orbit has 100% success rate, at payload below 4000kg
- Booster type FT us suitable for low payload mass (<6000kg) missions
- Booster type B5 is suitable for various payload mass
- Launch Site KSC LC-39A propose highest success launch ratio
- All features mentioned above for a space mission are to be considered to ensure a high first stage recovery rate, and therefore can make our Space Y competitive with Space X

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!