IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Juan José Saameño Pérez
October 30th, 2021

Outline

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

Executive Summary

- Data was collected from public SpaceX API and SpaceX Wikipedia page. The data was then explored with SQL, visualization, folium maps and dashboards. Afterwards, this data was sstandardized and it was used GridSearchCV to find the best parameters to create Machine Learning models.
- Four machine learnin models were used: Logistic regression, support vector machine, decision tree classifier and K-Nearest Neighbours. All of them produced similar results with an accuracy of 83.33%. All of them predicted successful landings, although more data would improve the accuracy and prediction models.

Introduction

Background

- Commercial Space is getting more and more popular
- SpaceX has the best prices for rocket launchings.
  - They can get back the stage 1 of the rocket
- SpaceY wants to compete with SpaceX

The mission

- SpaceY needs to train a machine learning model to predict successful recovery of stage 1

Section 1

Methodology

Methodology

- Executive Summary
- Data collection methodology:
  - Data from SpaceX public API and SpaceX Wikipedia page combined
- Perform data wrangling
  - Classifying true landings as successful and unsuccessful otherwise
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Models built and tuned using GridSearchCV

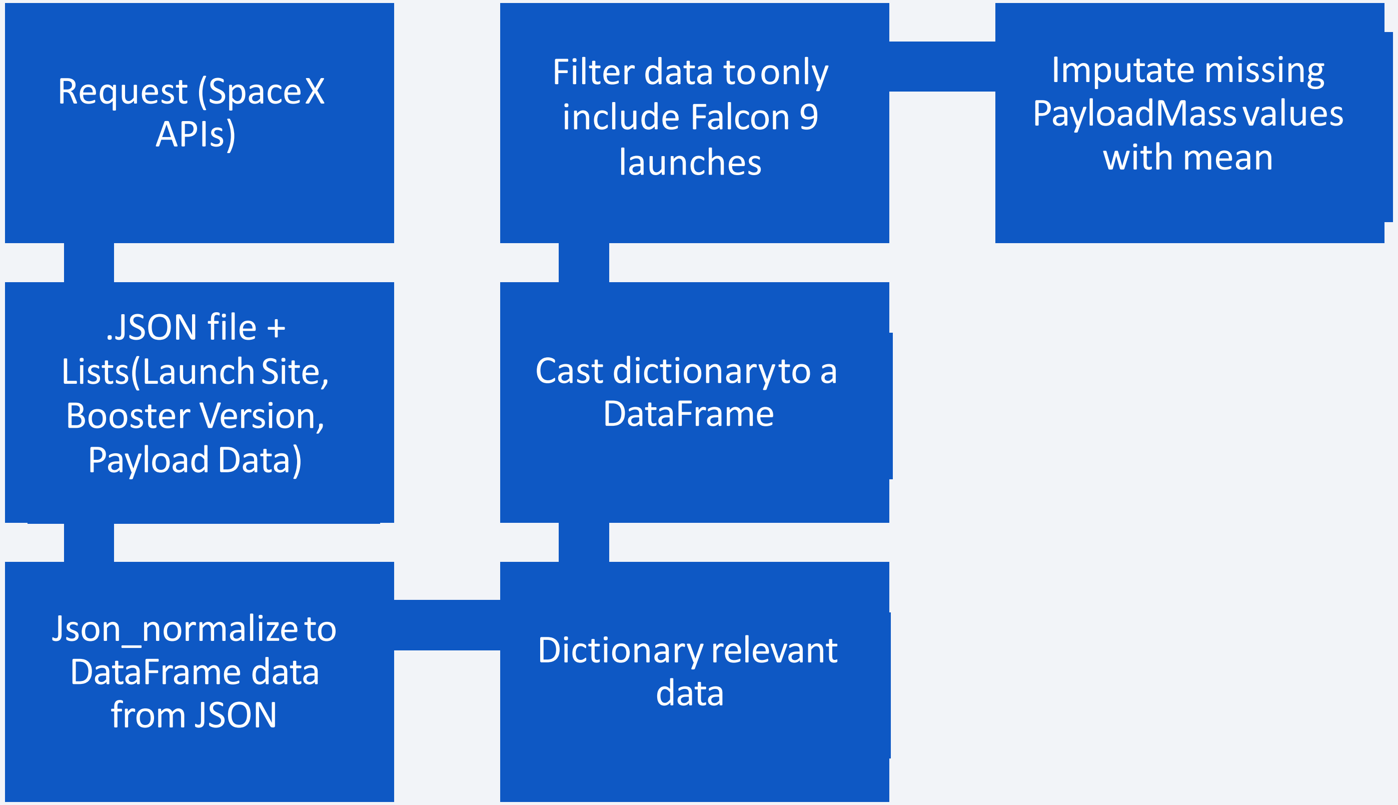
Data Collection

- Data collection process involved a combination of API requests from Space X public API and web scraping data from a table in Space X's Wikipedia entry.
- The next slide will show the flowchart of data collection from API and the one after will show the flowchart of data collection from webscraping.
- Space X API Data Columns:
  - FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins,
  - Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- Wikipedia Webscrape Data Columns:
  - Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API

- GitHub link:

https://github.com/juanjinho/Applied-Data-Science-Capstone/blob/ad9fe86dcb4a961c4e4fabcf2a557678cd82fe1f/Data%20Collection%20API.ipynb



Request (SpaceX APIs)

.JSON file + Lists(Launch Site, Booster Version, Payload Data)

Json_normalize to DataFrame data from JSON

Dictionary relevant data

Cast dictionary to a DataFrame

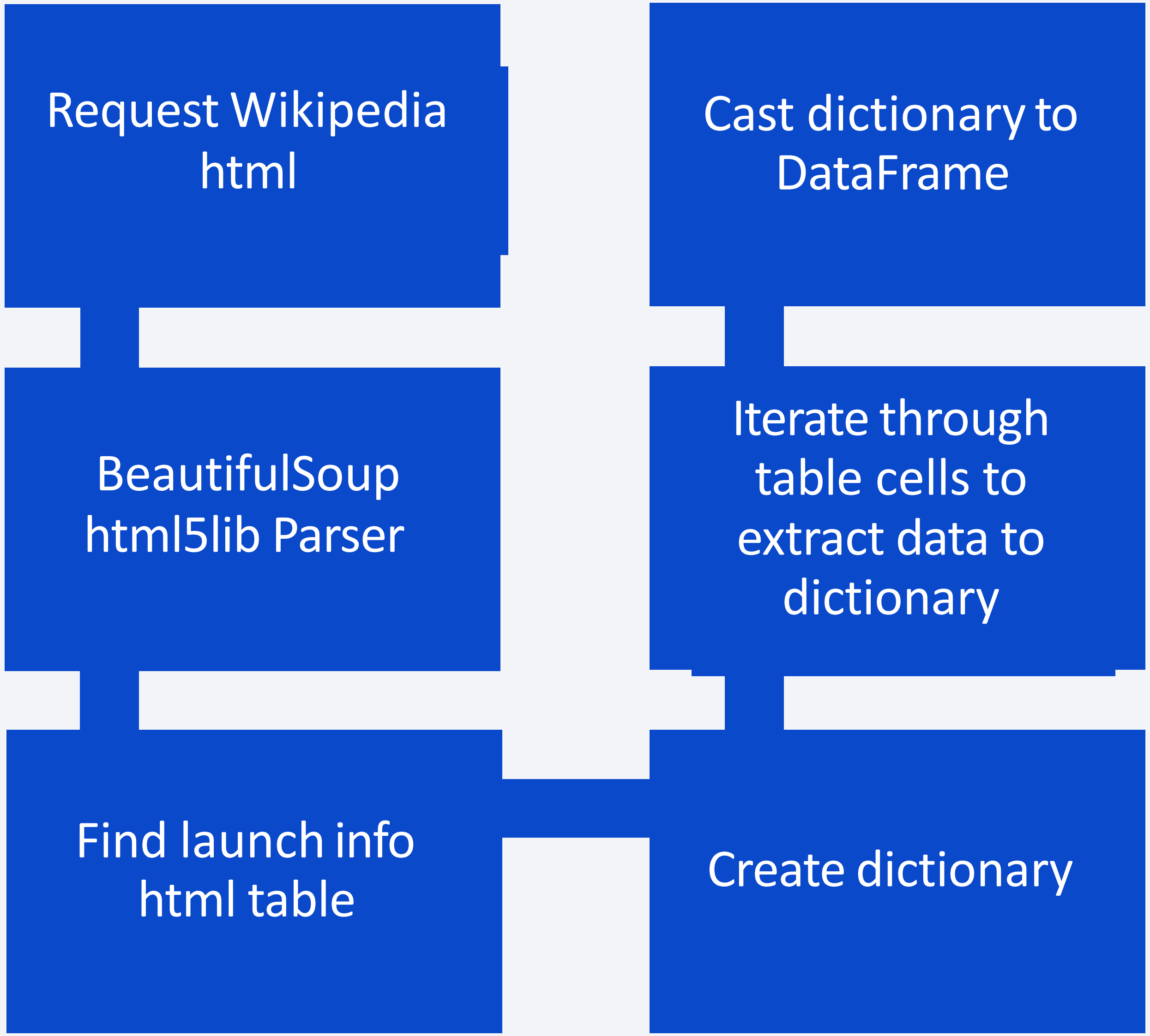
Filter data to only include Falcon 9 launches

Imputate missing PayloadMass values with mean

Data Collection - Scraping

- GitHub link:

https://github.com/juanjinho/Applied-Data-Science-Capstone/blob/ad9fe86dcb4a961c4e4fabcf2a557678cd82fe1f/Data%20Collection%20with%20Web%20Scraping.ipynb



Request Wikipedia html

BeautifulSoup html5lib Parser

Find launch info html table

Create dictionary

Iterate through table cells to extract data to dictionary

Cast dictionary to DataFrame

Data Wrangling

- Create a training label with landing outcomes where successful = 1 & failure = 0.

- Outcome column has two components: 'Mission Outcome' 'Landing Location'
- New training label column 'class' with a value of 1 if 'Mission Outcome' is True and 0 otherwise.
- Value Mapping:
  - True ASDS, True RTLS, & True Ocean – set to -> 1
  - None None, False ASDS, None ASDS, False Ocean, False RTLS – set to -> 0

- GitHub url:

https://github.com/juanjinho/Applied-Data-Science-Capstone/blob/ad9fe86dcb4a961c4e4fabcf2a557678cd82fe1f/Lab%202:%20Data%20wrangling.ipynb

EDA with Data Visualization

- Exploratory Data Analysis performed on variables Flight Number, Payload Mass, Launch Site, Orbit, Class and Year.
- Plots Used:
  - Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit vs. Success Rate, Flight Number vs. Orbit, Payload vs Orbit, and Success Yearly Trend
  - Scatter plots, line charts, and bar plots were used to compare relationships between variables to decide if a relationship exists so that they could be used in training the machine learning model
- GitHub url:
- https://github.com/juanjinho/Applied-Data-Science-Capstone/blob/ad9fe86dcb4a961c4e4fabcf2a557678cd82fe1f/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- Loaded data set into IBM DB2 Database.
- Queried using SQL Python integration.
- Queries were made to get a better understanding of the dataset.
- Queried information about launch site names, mission outcomes, various pay load sizes of customers and booster versions, and landing outcomes

- GitHub url:

https://github.com/juanjinho/Applied-Data-Science-Capstone/blob/ad9fe86dcb4a961c4e4fabcf2a557678cd82fe1f/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Folium maps mark Launch Sites, successful and unsuccessful landings, and a proximity example to key locations: Railway, Highway, Coast, and City.
- This allows us to understand why launch sites may be located where they are. Also visualizes successful landings relative to location.
- GitHub url:

https://github.com/juanjinho/Applied-Data-Science-Capstone/blob/ad9fe86dcb4a961c4e4fabcf2a557678cd82fe1f/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

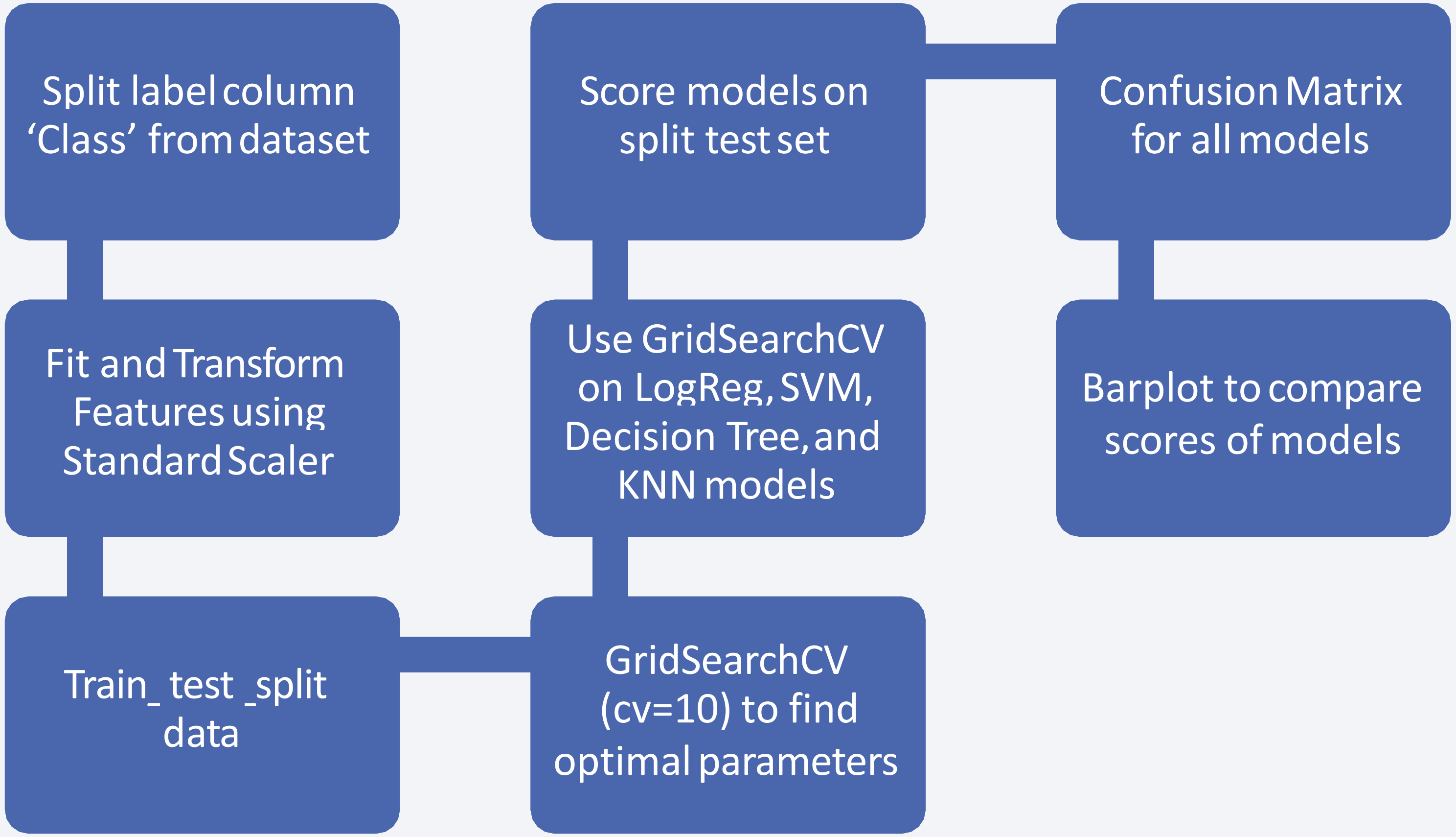
- Dashboard includes a pie chart and a scatter plot.
  - Pie chart can be selected to show distribution of successful landings across all launch sites and can be selected to show individual launch site success rates.
  - Scatter plot takes two inputs: All sites or individual site and payload mass on a slider between 0 and 10000 kg.
  - The pie chart is used to visualize launch site success rate.
  - The scatter plot can help us see how success varies across launch sites, payload mass, and booster version category.
- GitHub url:

https://github.com/juanjinho/Applied-Data-Science-Capstone/blob/ad9fe86dcb4a961c4e4fabcf2a557678cd82fe1f/Ploty_Dash.py

Predictive Analysis (Classification)

- GitHub link:

https://github.com/juanjinho/Applied-Data-Science-Capstone/blob/ad9fe86dcb4a961c4e4fabcf2a557678cd82fe1f/Machine%20Learning%20Prediction.ipynb



Split label column ‘Class’ from dataset

Fit and Transform Features using Standard Scaler

Train_ test _split data

GridSearchCV (cv=10) to find optimal parameters

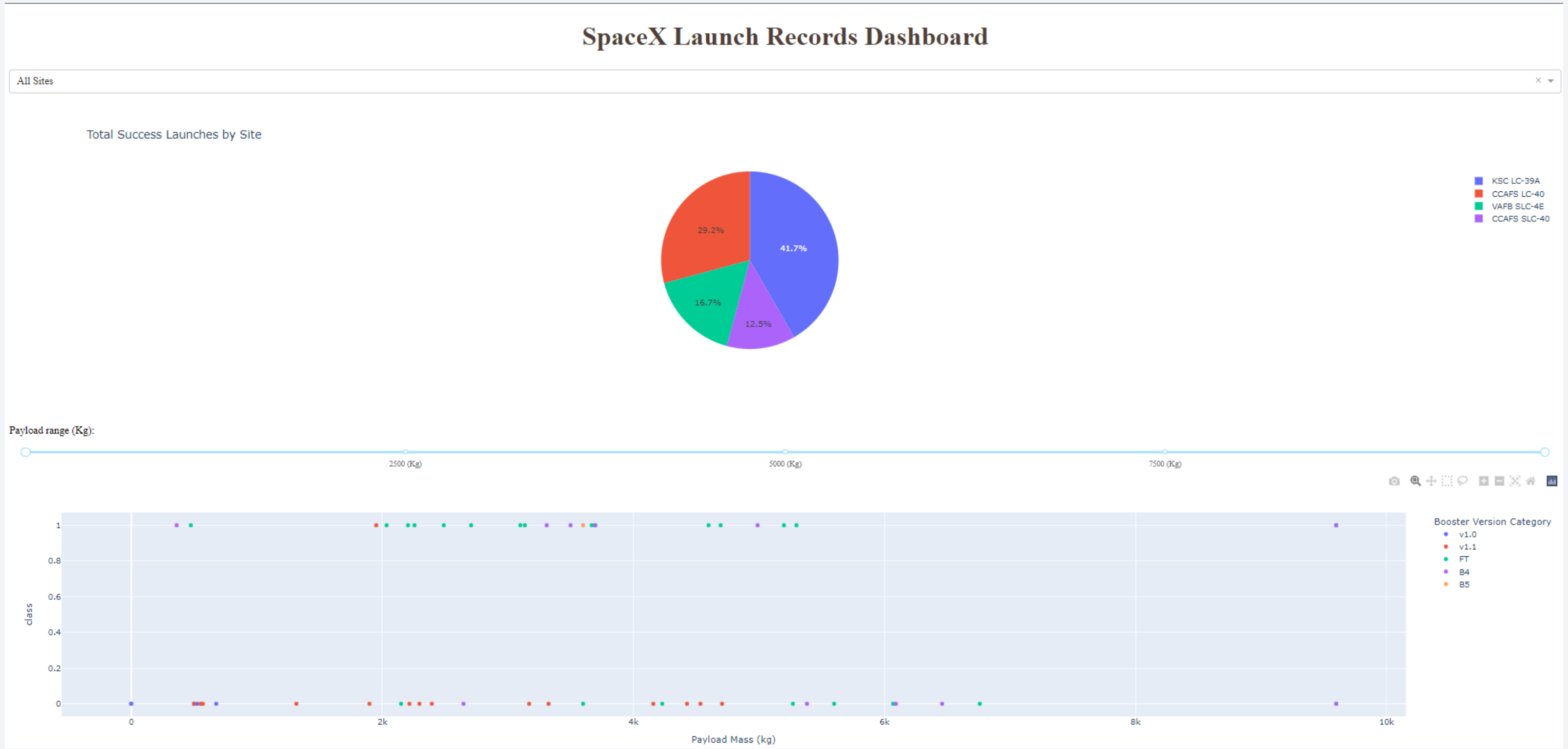
Use GridSearchCV on LogReg, SVM, Decision Tree, and KNN models

Score models on split test set

Confusion Matrix for all models

Barplot to compare scores of models

Results

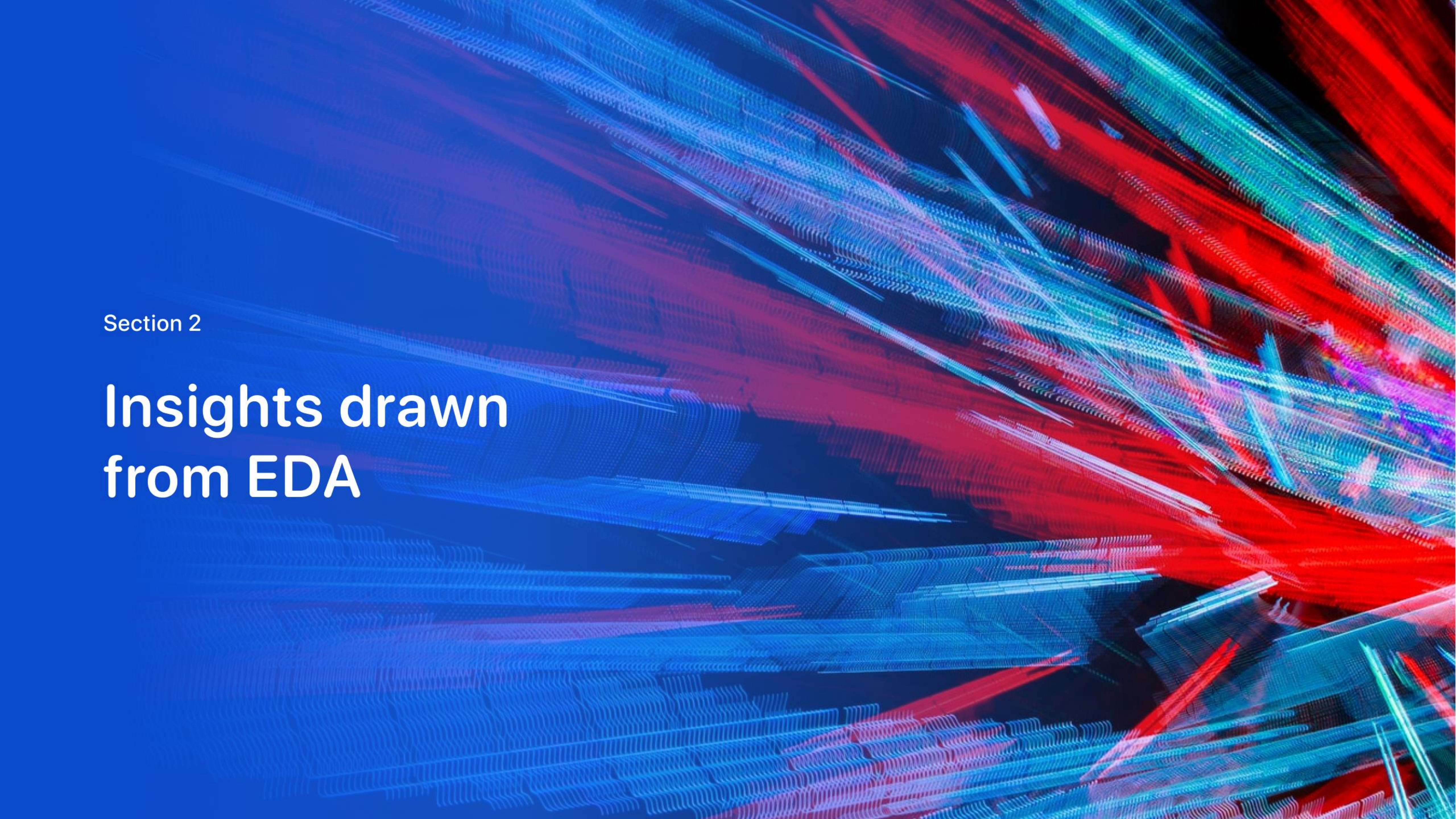


SpaceX Launch Records Dashboard

All Sites

Payload range (Kg):

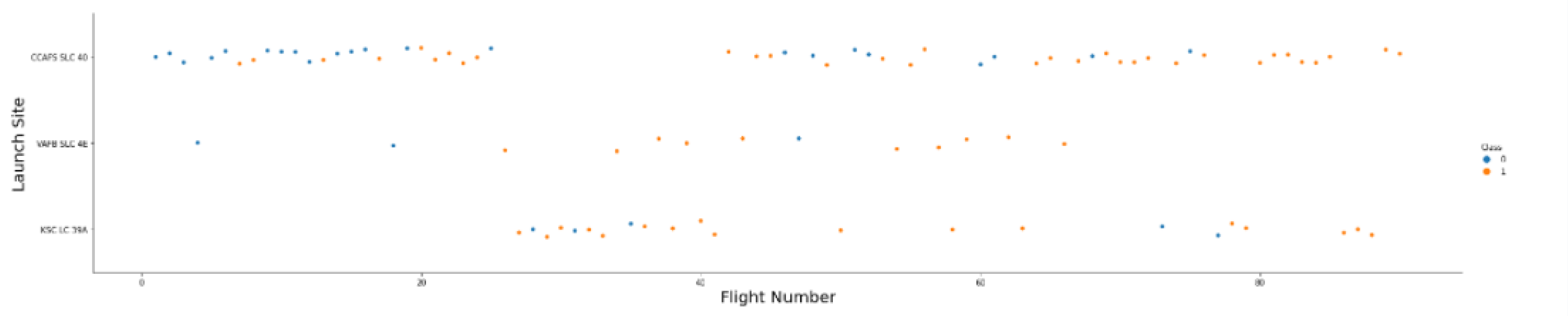
This is a preview of the Plotly dashboard. The following sides will show the results of EDA with visualization, EDA with SQL, Interactive Map with Folium, and finally the results of our model with about 83% accuracy.



Section 2

Insights drawn from EDA

Flight Number vs. Launch Site



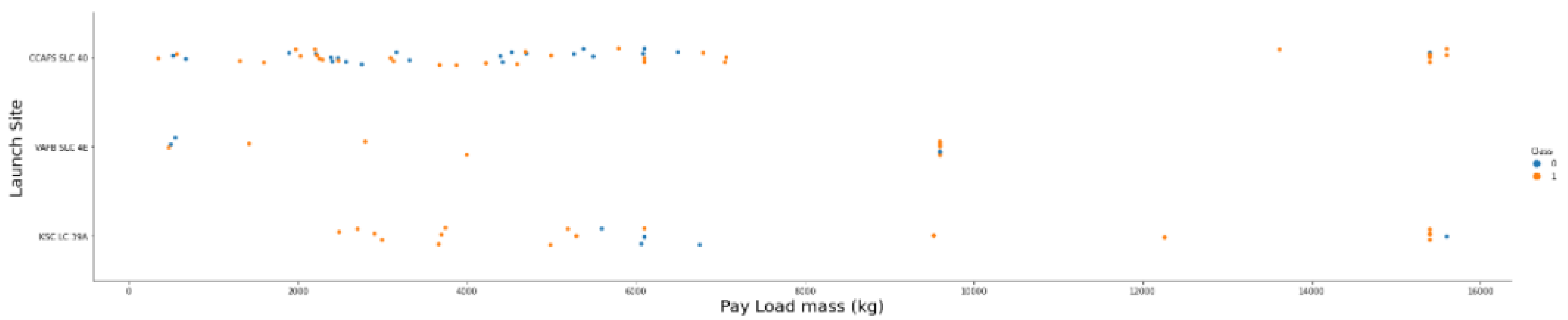
Orange = Successful launch; Blue = Unsuccessful launch

Success rate increases over time

Success increases after flight 20 approximately

CCAFS appears to be the main launch site

Payload vs. Launch Site



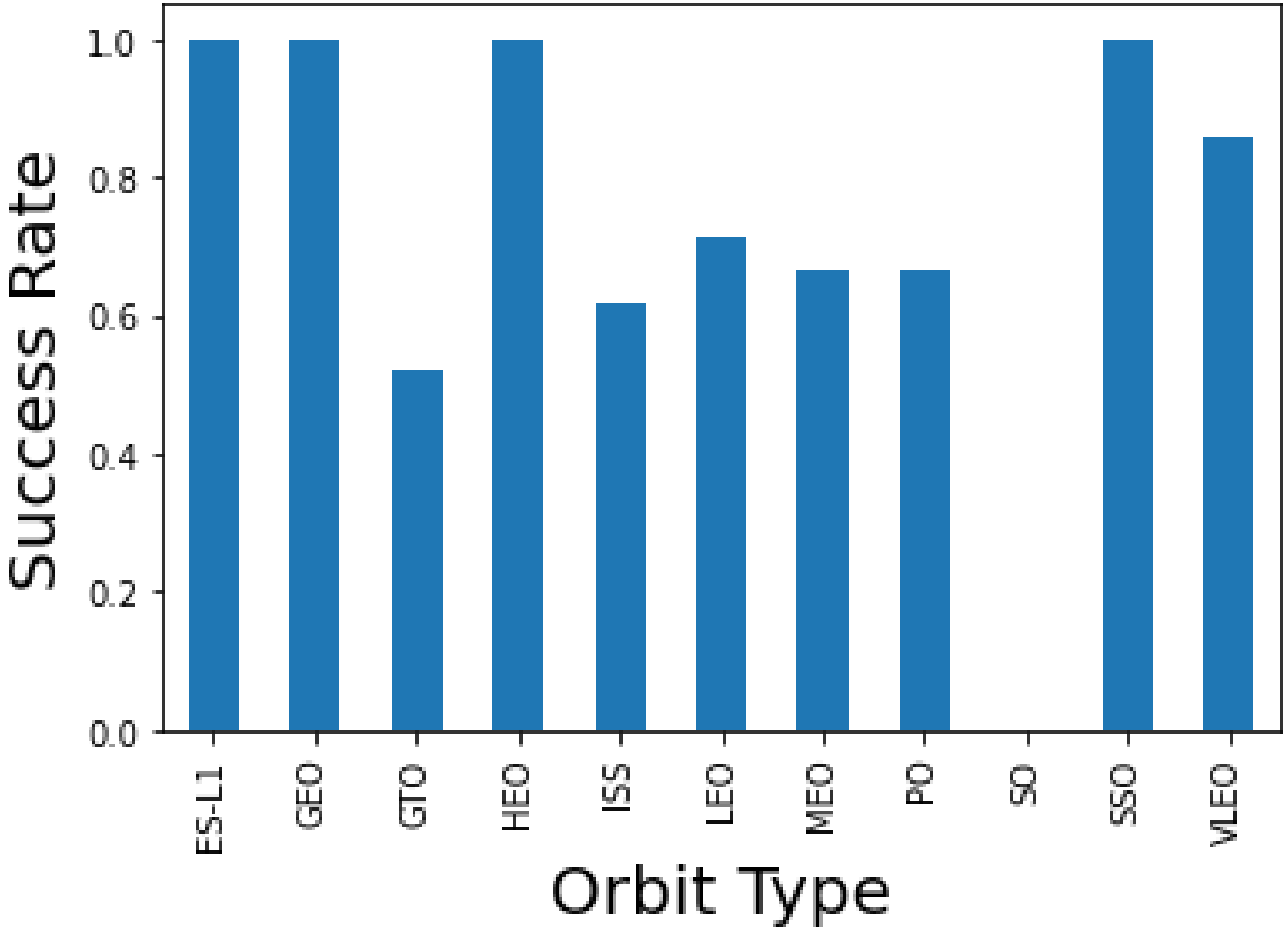
Orange = Successful launch; Blue = Unsuccessful launch

Payload mass appears to fall between 0-6000 kg

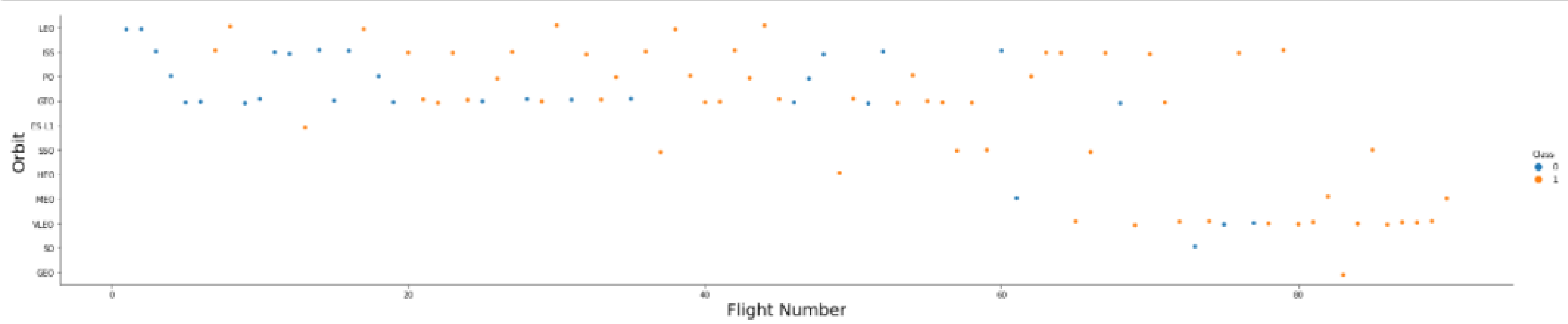
The payload seems to influence the launch site.

Success Rate vs. Orbit Type

- ES-L1 (1), GEO (1), HEO (1) have 100% success rate (sample sizes in parenthesis) SSO (5) has 100% success rate
- VLEO (14) has decent success rate and attempts
- SO (1) has 0% success rate
- GTO (27) has the around 50% success rate but largest sample



Flight Number vs. Orbit Type



Orange = Successful launch; Blue = Unsuccessful launch

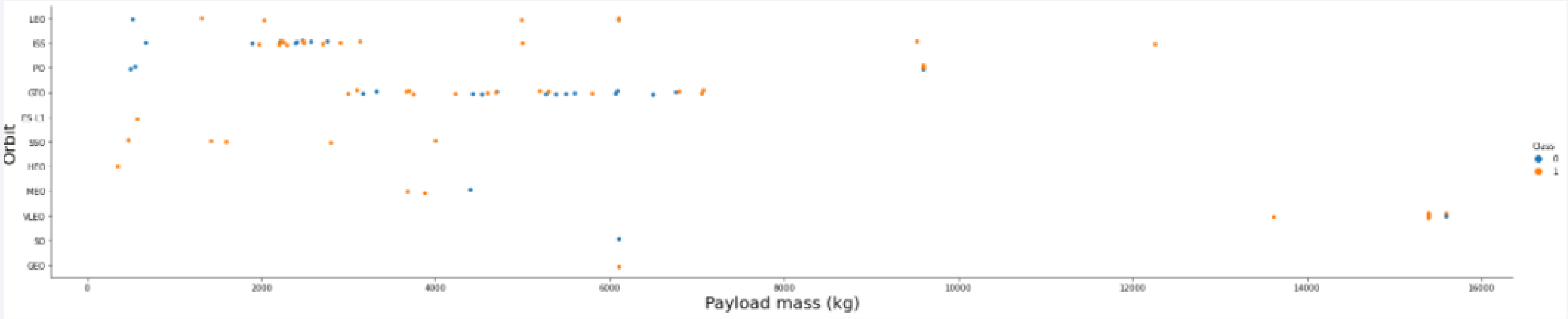
Launch Orbit preferences changed over Flight Number.

Launch Outcome seems to correlate with this preference.

SpaceX started with LEO orbits which saw moderate success LEO and returned to VLEO in recent launches.

SpaceX appears to perform better in lower orbits or Sun-synchronous orbits

Payload vs. Orbit Type

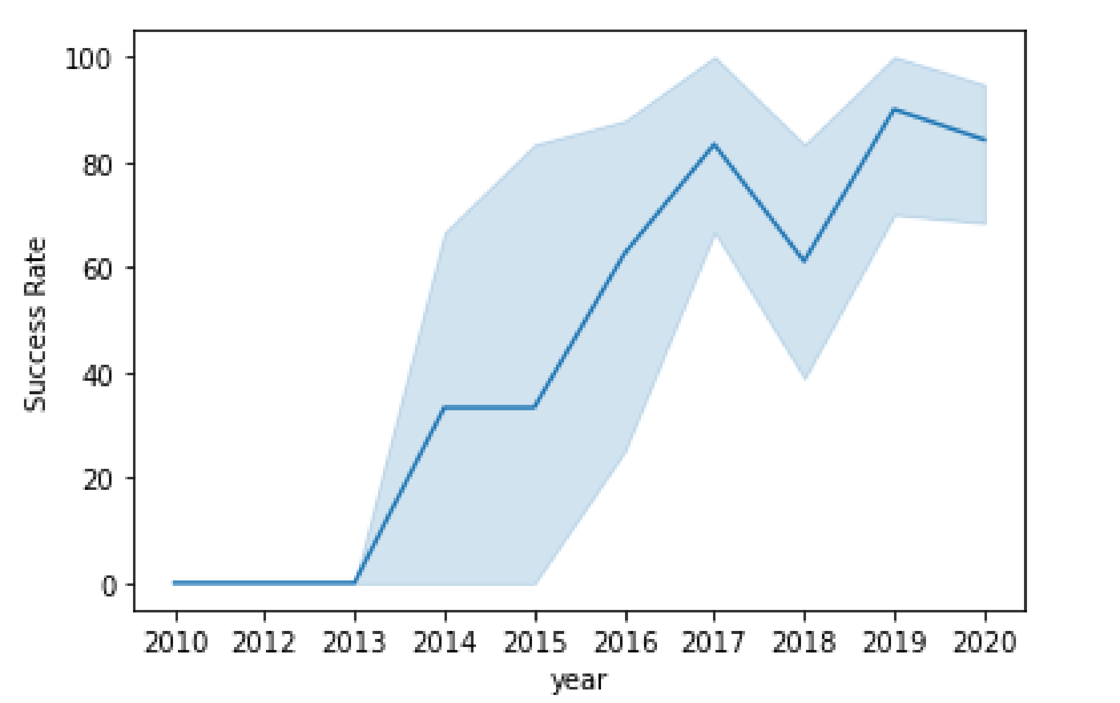


Orange = Successful launch; Blue = Unsuccessful launch

- Payload mass seems to correlate with orbit
- LEO and SSO seem to have relatively low payload mass
- The other most successful orbit VLEO only has payload mass values in the higher end of the range

Launch Success Yearly Trend

- Success generally increases over time since 2013 with a slight dip in 2018
- Success in recent years at around 80%



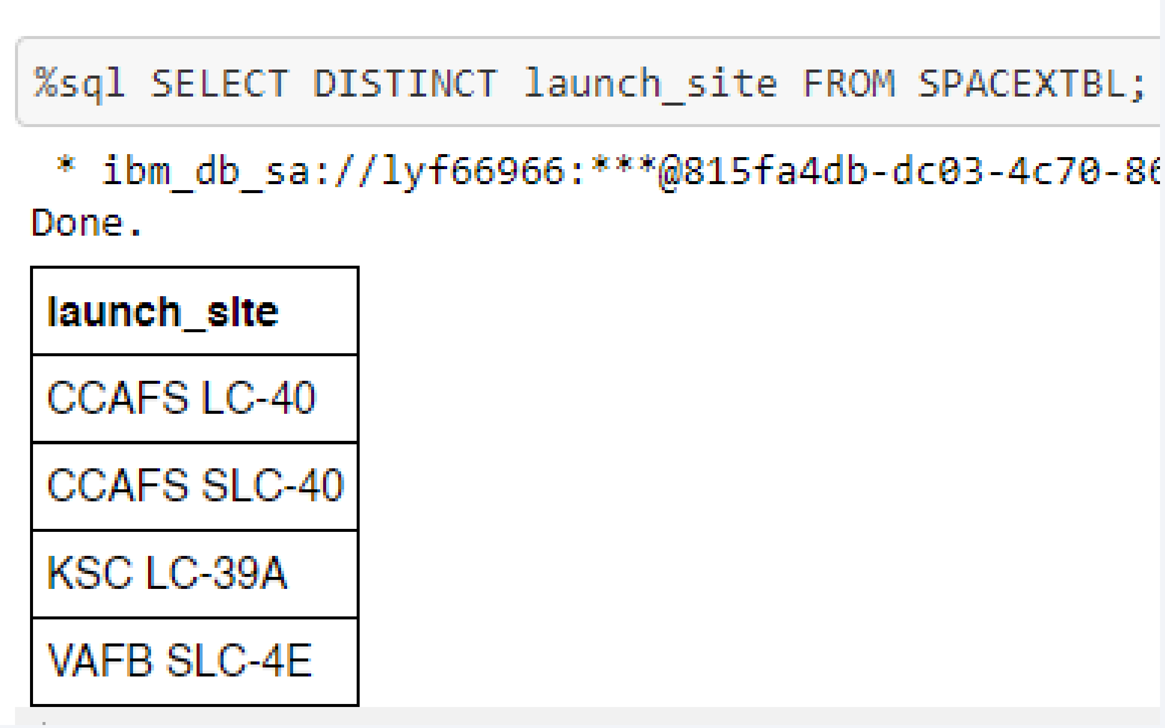
All Launch Site Names

Query unique launch site names from database.

CCAFS SLC-40 and CCAFSSLC-40 likely all represent the same

launch site with data entry errors.

CCAFS LC-40 was the previous name. Likely only 3 unique launch_site values: CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E

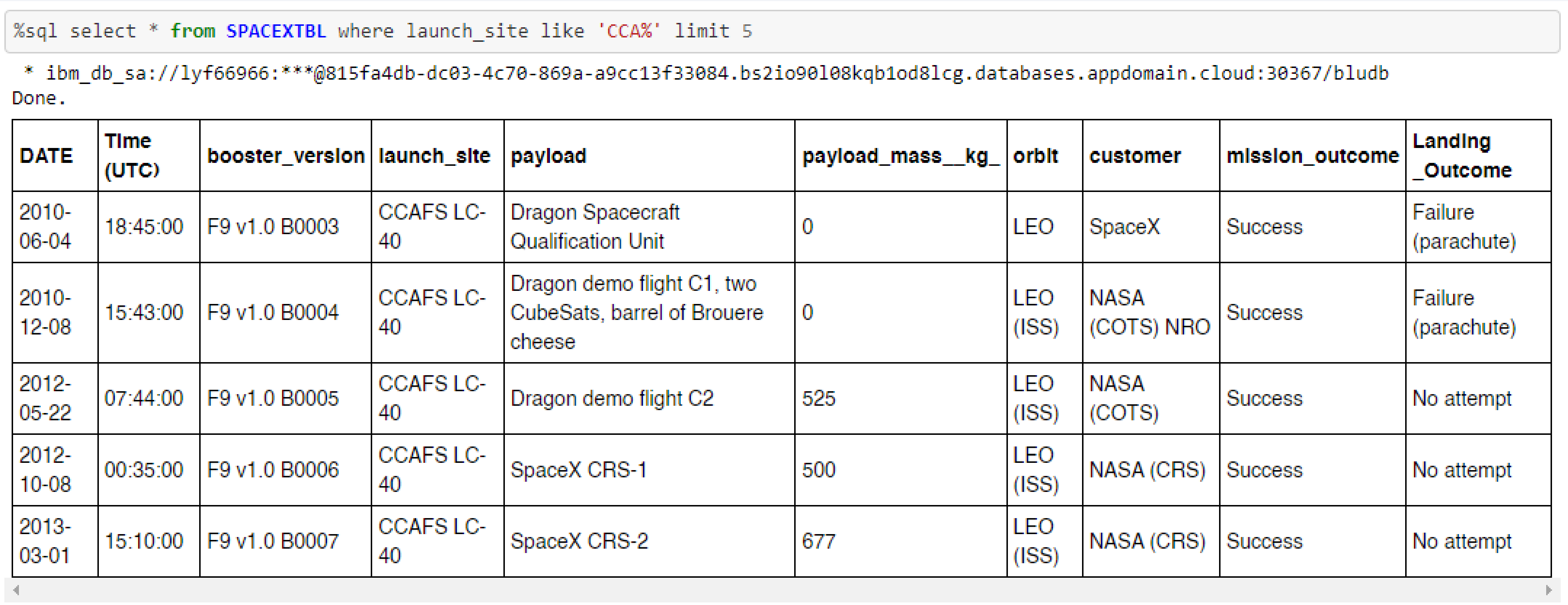


```
%sql SELECT DISTINCT launch_site FROM SPACEXTBL;
```

```
 * ibm_db_sa://lyf66966:***@815fa4db-dc03-4c70-86
Done.
```

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

Launch Site Names Begin with 'CCA'



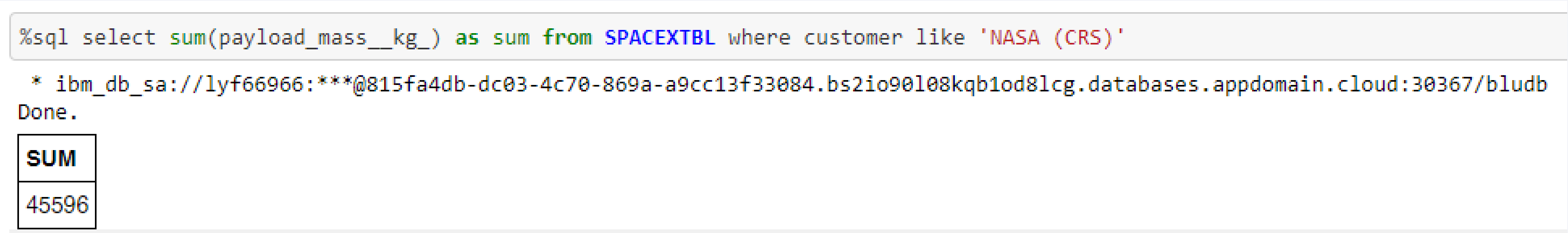
```
%sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5
```

```
 * ibm_db_sa://lyf66966:***@815fa4db-dc03-4c70-869a-a9cc13f33084.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30367/bludb
Done.
```

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

- First five entries in database with Launch Site name beginning with CCA.

Total Payload Mass



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

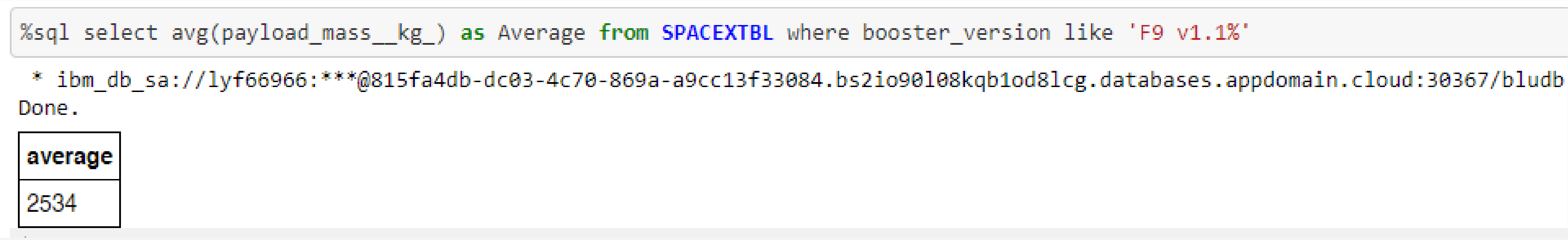
```
 * ibm_db_sa://lyf66966:***@815fa4db-dc03-4c70-869a-a9cc13f33084.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30367/bludb
Done.
```

SUM
45596

This query sums the total payload mass in kg where NASA was the customer.

CRS stands for Commercial Resupply Services which indicates that these payloads were sent to the International Space Station (ISS).

Average Payload Mass by F9 v1.1



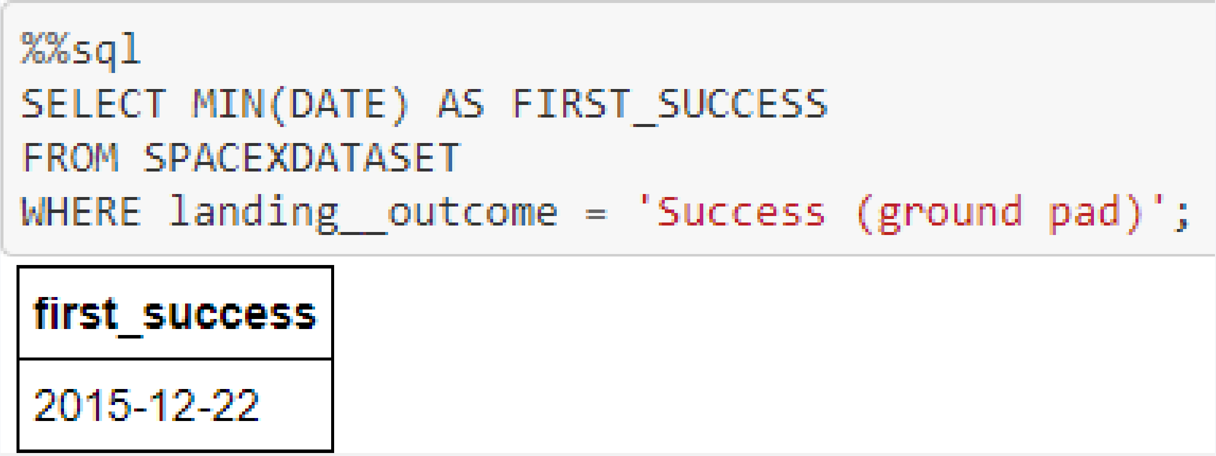
```
%sql select avg(payload_mass__kg_) as Average from SPACEXTBL where booster_version like 'F9 v1.1%'
```

```
 * ibm_db_sa://lyf66966:***@815fa4db-dc03-4c70-869a-a9cc13f33084.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30367/bludb
Done.
```

average
2534

- This query calculates the average payload mass or launches which used booster version F9 v1.1
- Average payload mass of F9 1.1 is on the low end of our payload mass range

First Successful Ground Landing Date



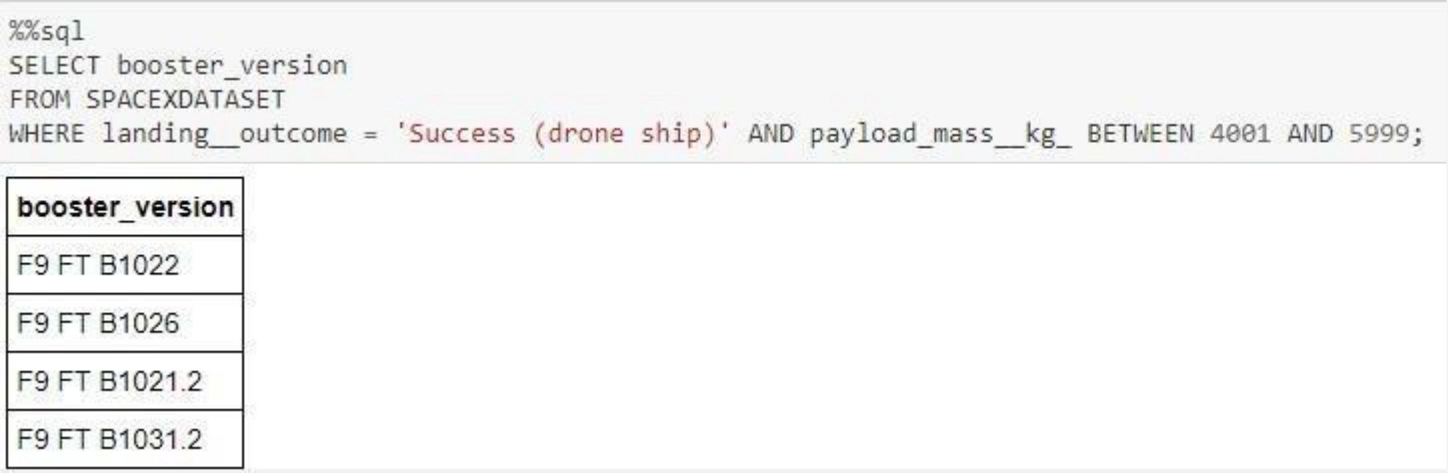
```
%%sql
SELECT MIN(DATE) AS FIRST_SUCCESS
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (ground pad)';
```

first_success
2015-12-22

- This query returns the first successful ground pad landing date.
- First ground pad landing wasn’t until the end of 2015.
- Successful landings in general appear starting 2014.

Successful Drone Ship Landing with Payload between 4000 and 6000

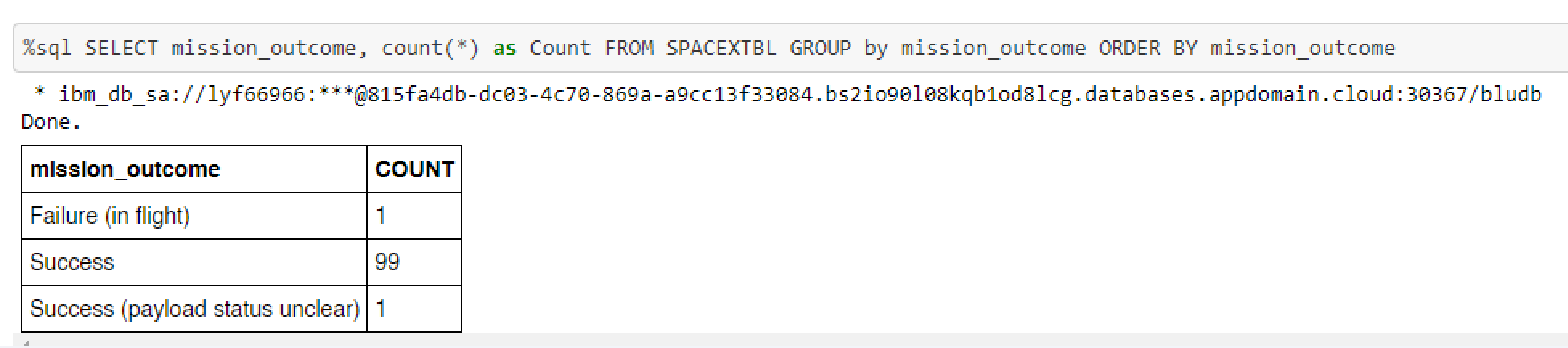
- This query returns the four booster versions that had successful drone ship landings and a payload mass between 4000 and 6000 noninclusively.



```
%%sql
SELECT booster_version
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4001 AND 5999;
```

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

Total Number of Successful and Failure Mission Outcomes



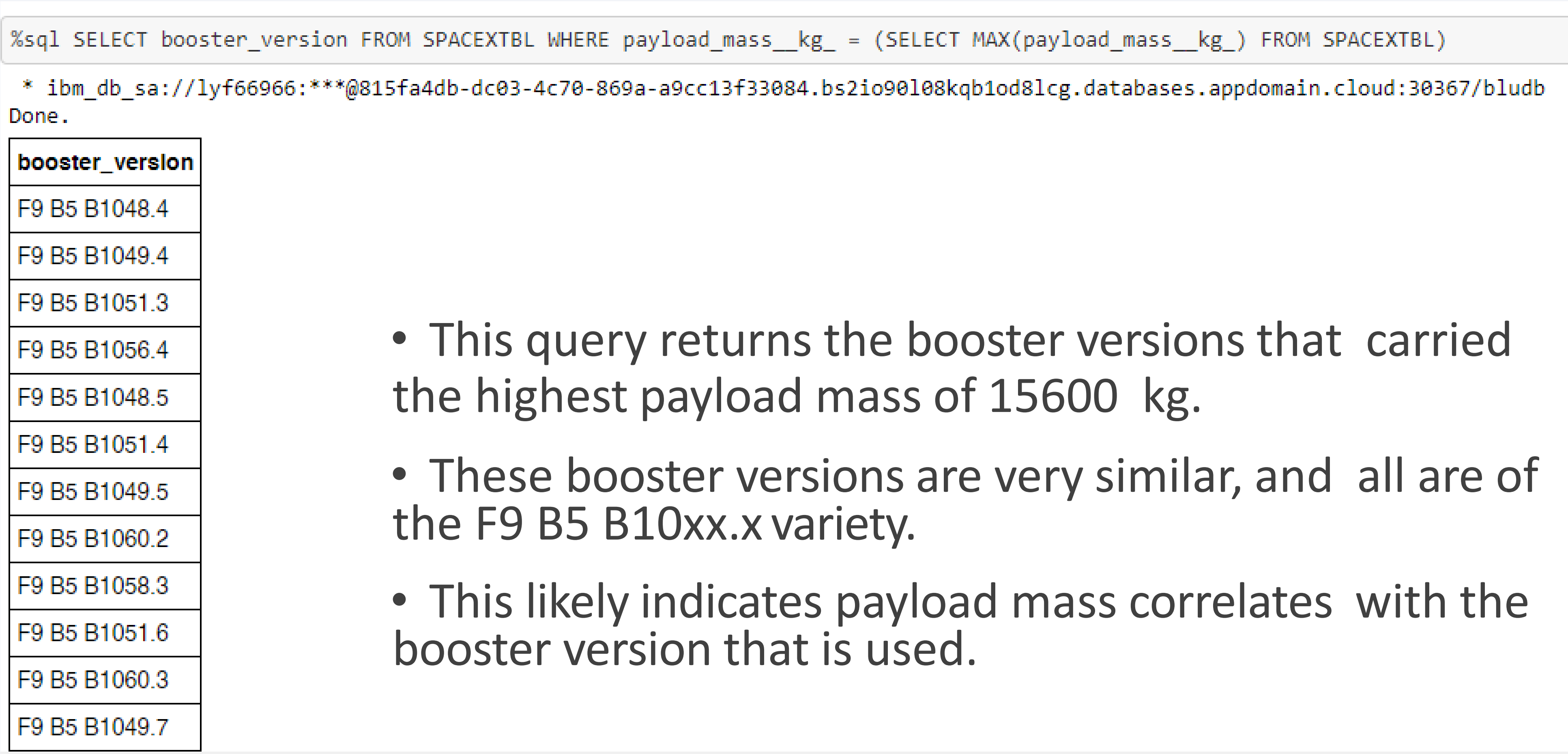
```
%sql SELECT mission_outcome, count(*) as Count FROM SPACEXTBL GROUP by mission_outcome ORDER BY mission_outcome
```

```
 * ibm_db_sa://lyf66966:***@815fa4db-dc03-4c70-869a-a9cc13f33084.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30367/bludb
Done.
```

mission_outcome	COUNT
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- This query returns a count of each mission outcome.
- SpaceX appears to achieve its mission outcome nearly 99% of the time.
- This means that most of the landing failures are intended.
- One launch has an unclear payload status and unfortunately one failed in flight.

Boosters Carried Maximum Payload



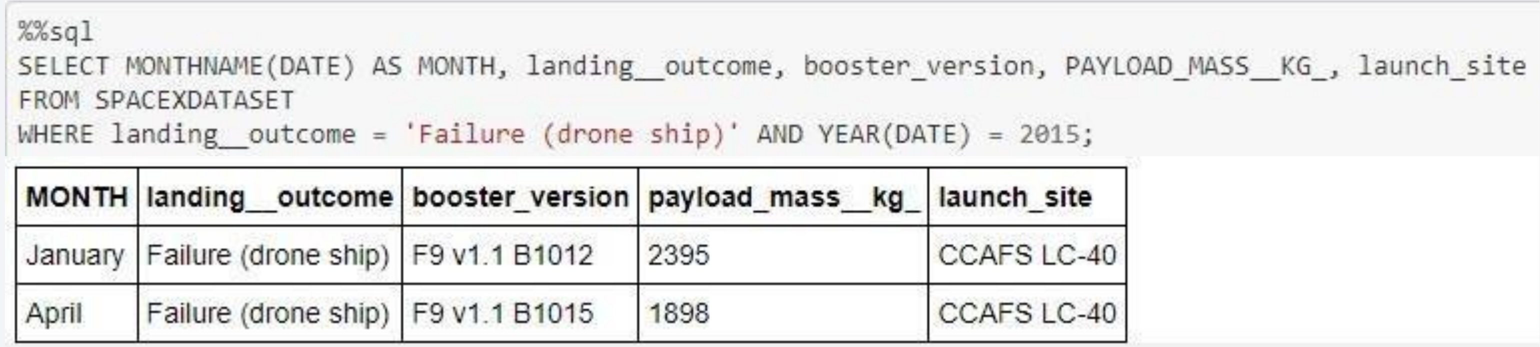
```
%sql SELECT booster_version FROM SPACEXTBL WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM SPACEXTBL)
```

```
 * ibm_db_sa://lyf66966:***@815fa4db-dc03-4c70-869a-a9cc13f33084.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30367/bludb
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

- This query returns the booster versions that carried the highest payload mass of 15600 kg.
- These booster versions are very similar, and all are of the F9 B5 B10xx.x variety.
- This likely indicates payload mass correlates with the booster version that is used.

2015 Launch Records

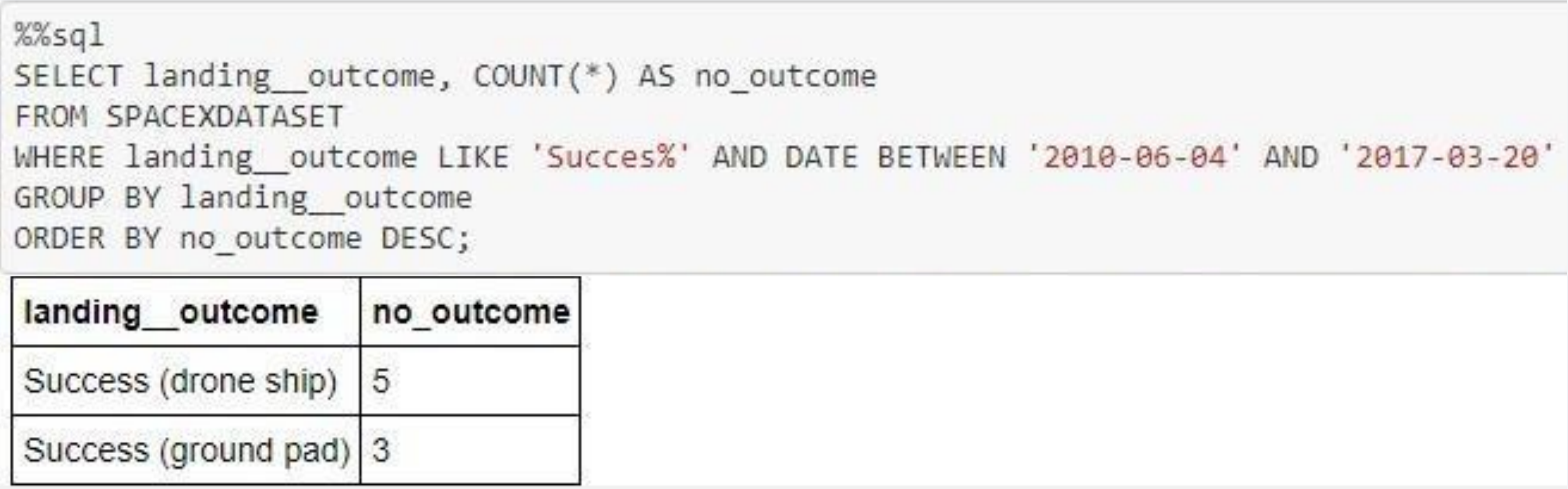


```
%%sql
SELECT MONTHNAME(DATE) AS MONTH, landing__outcome, booster_version, PAYLOAD_MASS__KG_, launch_site
FROM SPACEXDATASET
WHERE landing__outcome = 'Failure (drone ship)' AND YEAR(DATE) = 2015;
```

MONTH	landing__outcome	booster_version	payload_mass__kg_	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	2395	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	1898	CCAFS LC-40

- This query returns the Month, Landing Outcome, Booster Version, Payload Mass (kg), and Launch site of 2015 launches where stage 1 failed to land on a drone ship.
- There were two such occurrences.

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



```
%%sql
SELECT landing__outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
WHERE landing__outcome LIKE 'Succes%' AND DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY landing__outcome
ORDER BY no_outcome DESC;
```

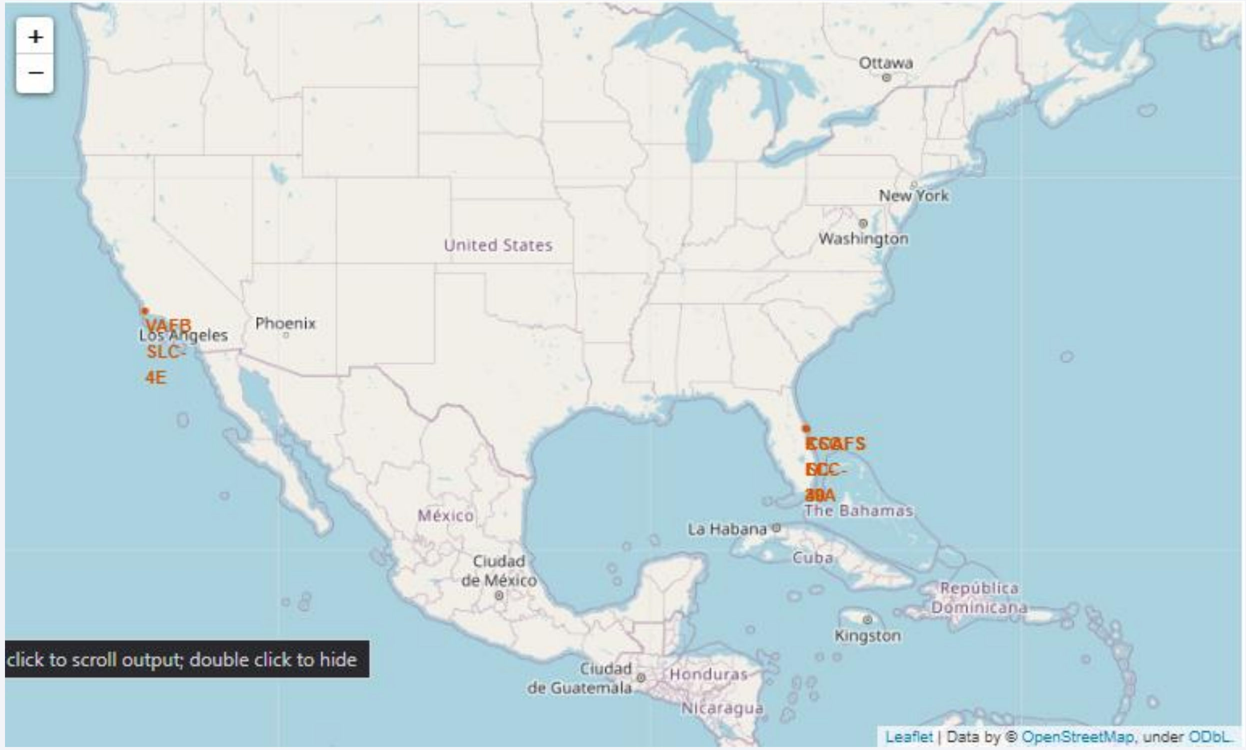
landing__outcome	no_outcome
Success (drone ship)	5
Success (ground pad)	3

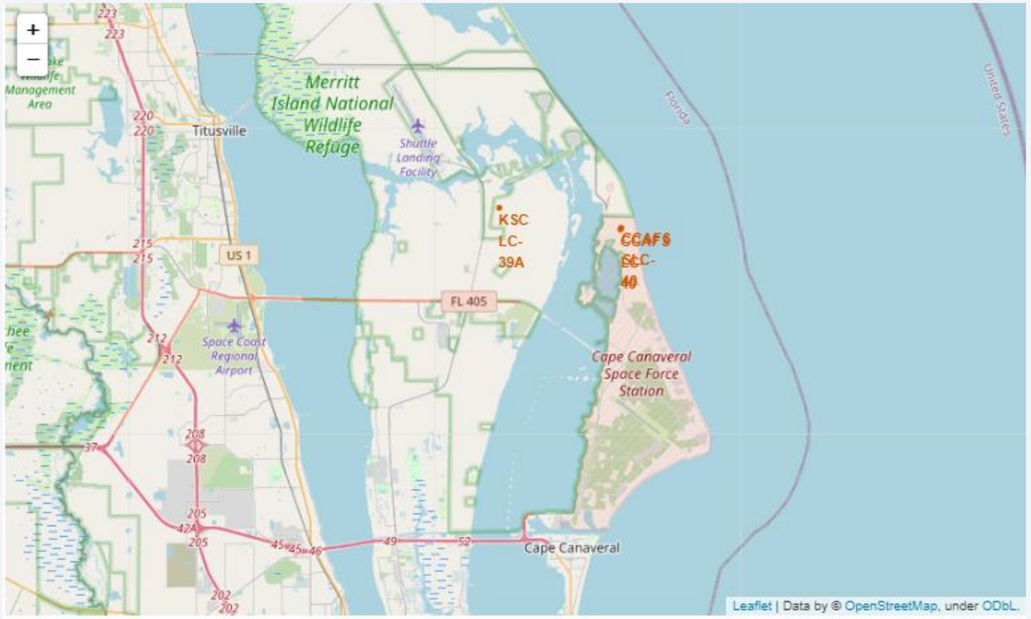
- This query returns a list of successful landings and between 2010-06-04 and 2017-03-20 inclusively.
- There are two types of successful landing outcomes: drone ship and ground pad landings.
- There were 8 successful landings in total during this time period

Section 4

Launch Sites Proximities Analysis

Launch site locations

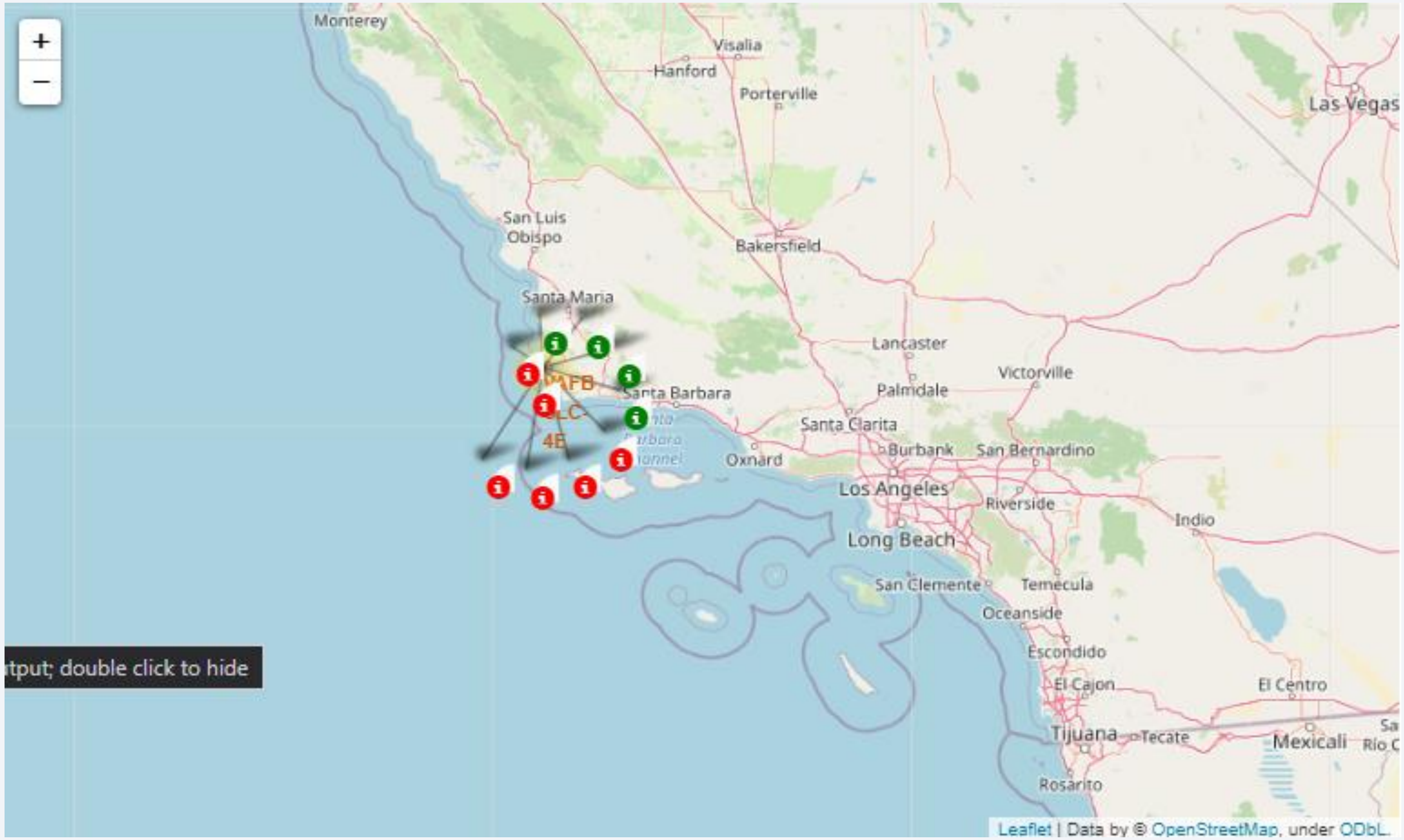




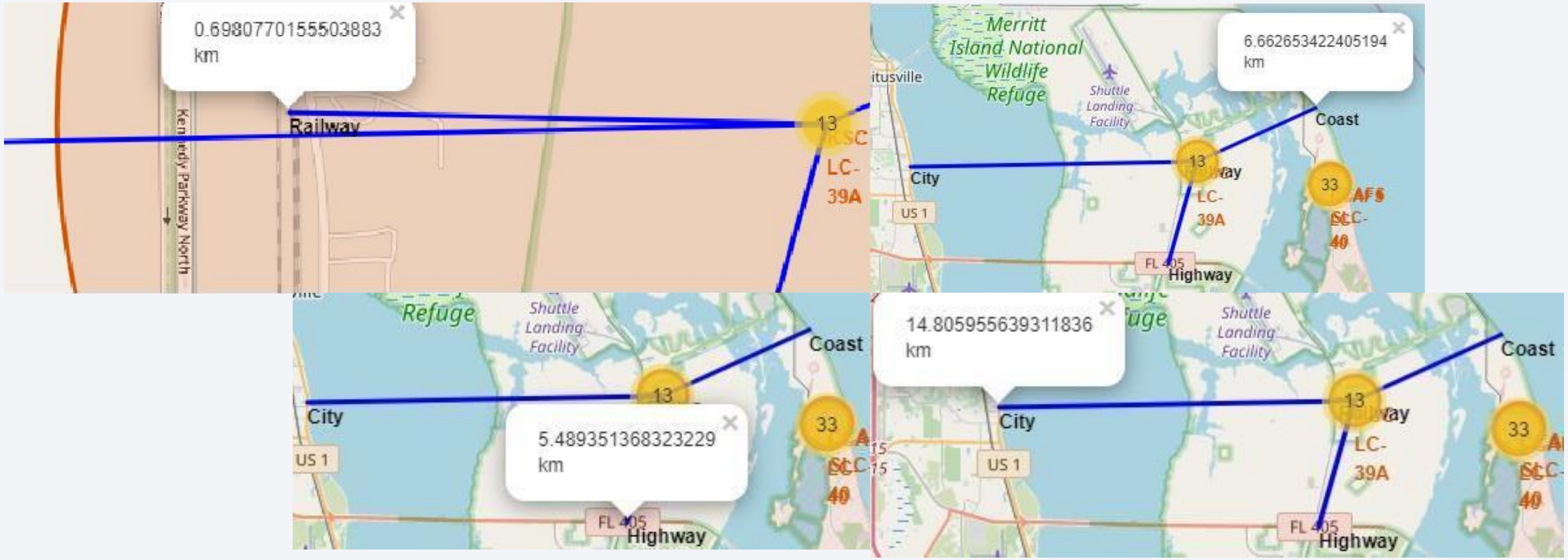
- The left map shows all launch sites relative US map.
- The right map shows the two Florida launch sites since they are very close to each other. All launch sites are near the ocean.

Color coded launch markers

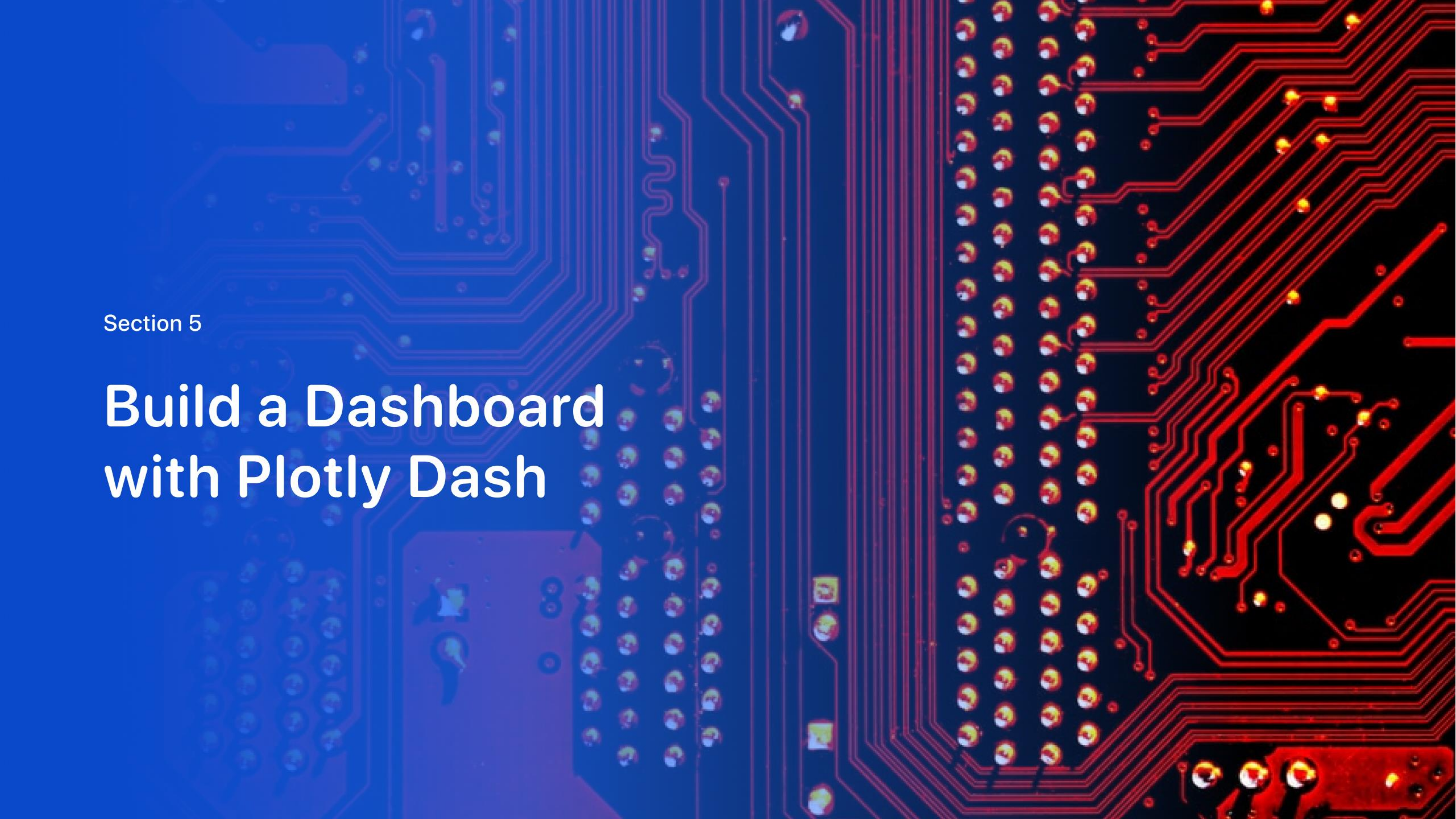
- Clusters on Folium map can be clicked on to display each successful landing (green icon) and failed landing (red icon). In this example VAFB SLC-4E shows 4 successful landings and 6 failed landings.



Key location proximities



Using KSC LC-39A as an example, launch sites are very close to railways for large part and supply transportation. Launch sites are close to highways for human and supply transport. Launch sites are also close to coasts and relatively far from cities so that launch failures can land in the sea to avoid rockets falling on densely populated areas.

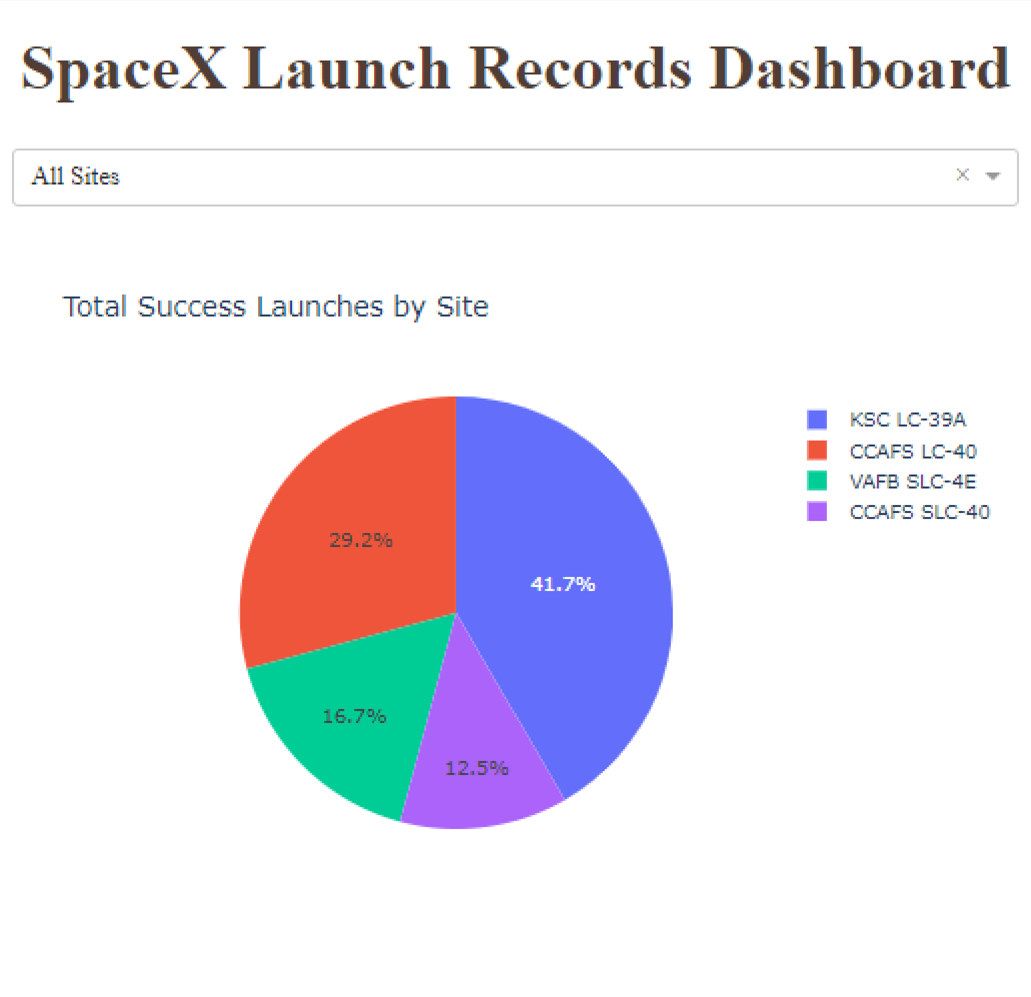


Section 5

Build a Dashboard with Plotly Dash

Successful launches by launch site

This is the distribution of successful landings across all launch sites. CCAFS LC-40 is the old name of CCAFS SLC-40 so CCAFS and KSC have the same number of successful landings, but a majority of the successful landings were performed before the name change. VAFB has the smallest share of successful landings. This may be due to smaller sample and increase in difficulty of launching in the west coast.

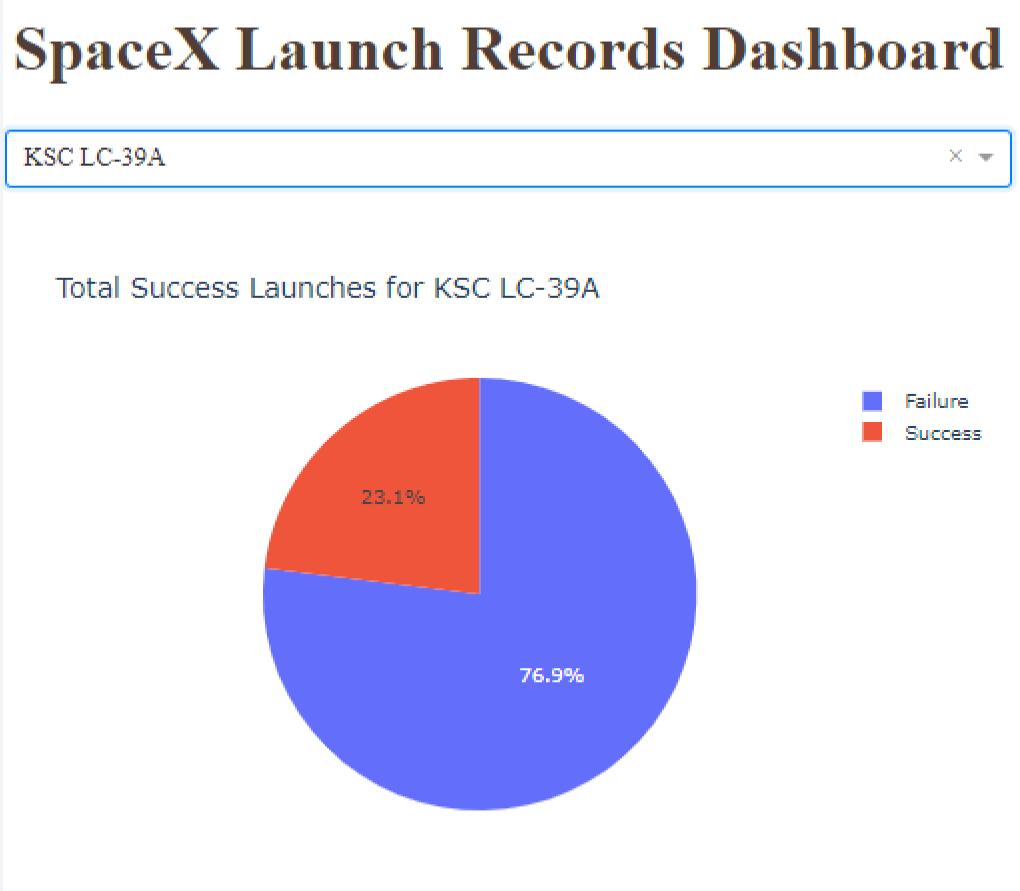


SpaceX Launch Records Dashboard

All Sites

Highest success rate

KSC LC-39A has the highest success rate with 10 successful landings and 3 failed landings.

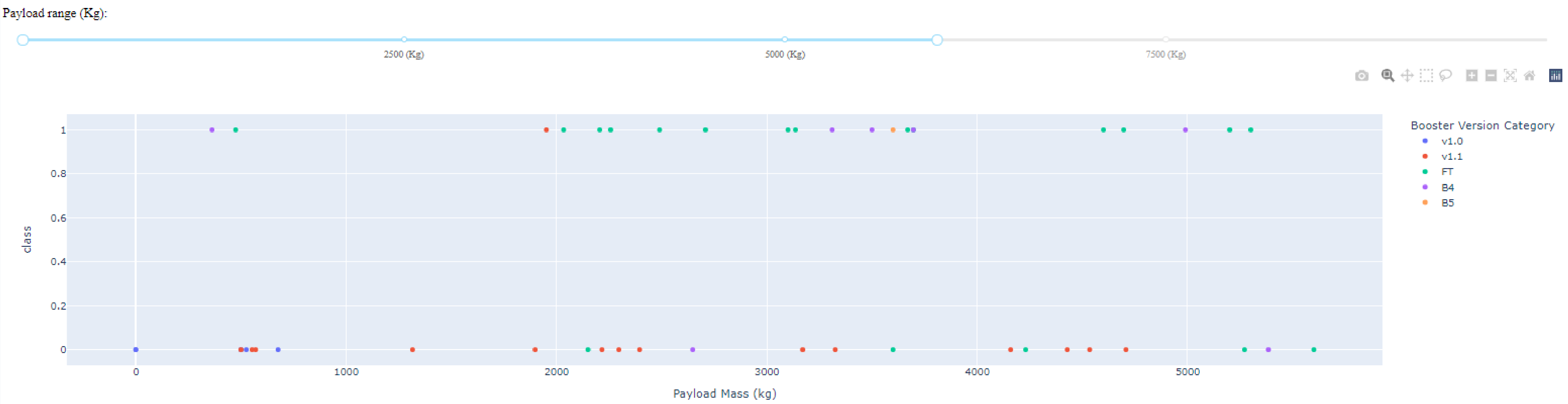


SpaceX Launch Records Dashboard

KSC LC-39A

Total Success Launches for KSC LC-39A

Payload mass – Success - Booster

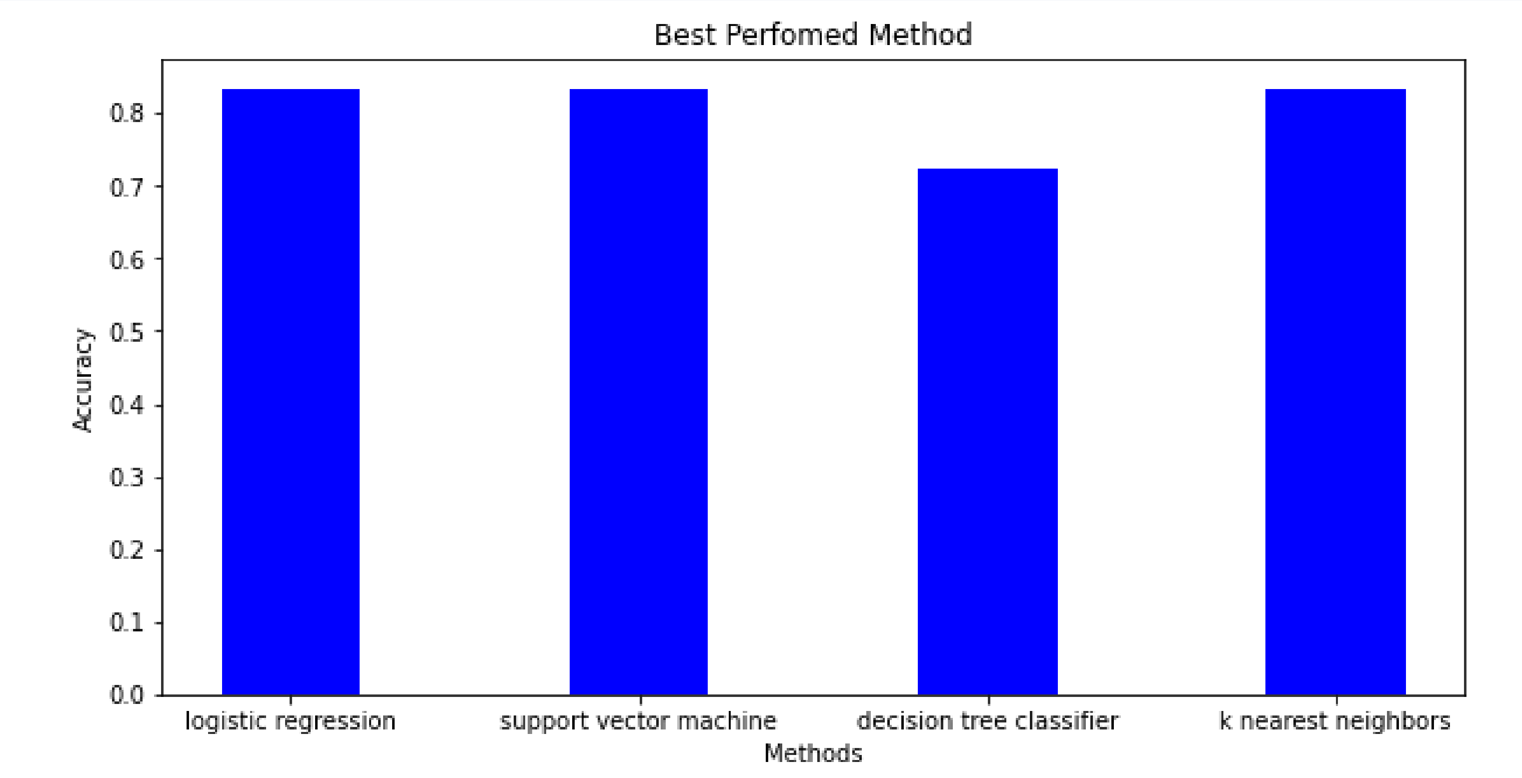


- Plotly dashboard has a Payload range selector. However, this is set from 0-10000 instead of the max Payload of 15600. Class indicates 1 for successful landing and 0 for failure. Scatter plot also accounts for booster version category in color and number of launches in point size. In this particular range of 0-6000, interestingly there are two failed landings with payloads of zero kg.

Section 6

Predictive Analysis (Classification)

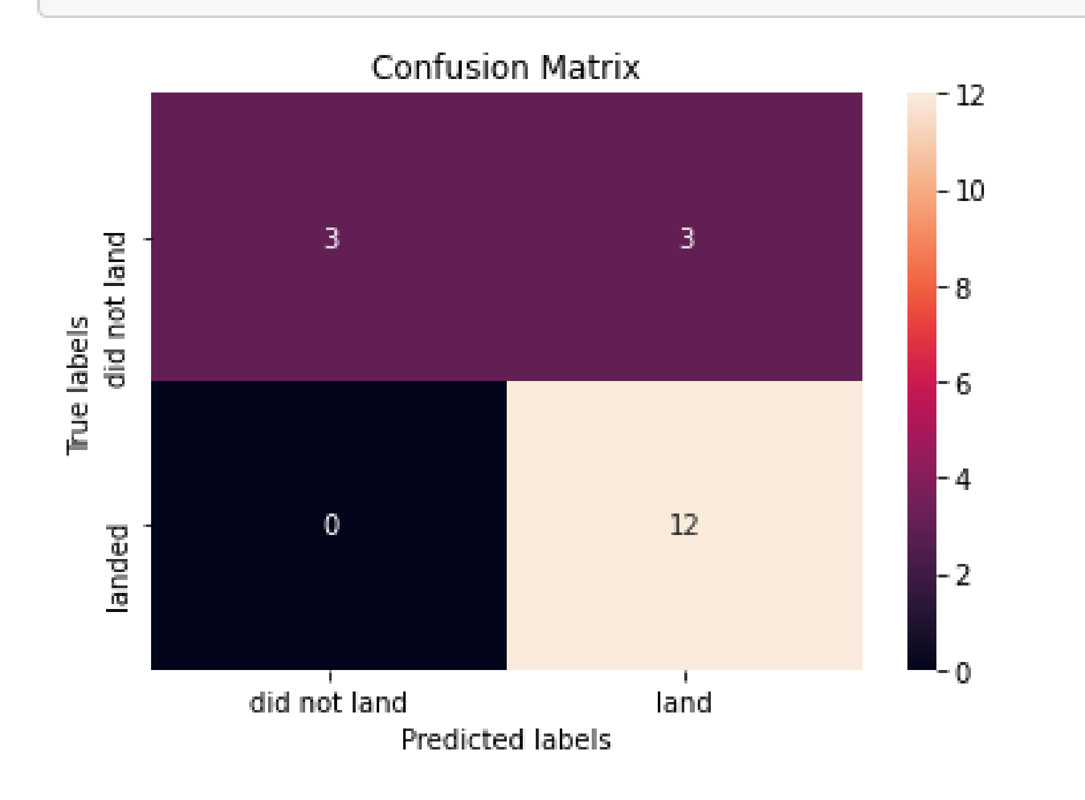
Classification Accuracy



- All models had virtually the same accuracy on the test set at 83.33% accuracy except for the decision tree classifier with 0.72.
- It should be noted that test size is small at only sample size of 18.
- This can cause large variance in accuracy results, such as those in Decision Tree Classifier model in repeated runs.
- We likely need more data to determine the best model.

Confusion Matrix

- The confusion matrix is the same for the logistic regression, support vector machine and k-nearest neighbours. The models predicted 12 successful landings when the true label was successful landing.
- The models predicted 3 unsuccessful landings when the true label was unsuccessful landing.
- The models predicted 3 successful landings when the true label was unsuccessful landings (false positives). Our models over predict successful landings.



Conclusions

- Our task: to develop a machine learning model for Space Y who wants to bid against SpaceX
- The goal of model is to predict when Stage 1 will successfully land to save ~$100 million USD
- Used data from a public SpaceX API and web scraping SpaceX Wikipedia page
- Created data labels and stored data into a DB2 SQL database
- Created a dashboard for visualization
- We created a machine learning model with an accuracy of 83%
- Allon Mask of SpaceY can use this model to predict with relatively high accuracy whether a launch will have a successful Stage 1 landing before launch to determine whether the launch should be made or not
- If possible, more data should be collected to better determine the best machine learning model and improve accuracy

Appendix

- GitHub repository url:

https://github.com/juanjinho/Applied-Data-Science-Capstone/tree/master

- Instructors:

Rav Ahuja, Alex Aklson, Aije Egwaikhide, Svetlana Levitan, Romeo Kienzler, Polong Lin, Joseph Santarcangelo, Azim Hirjani, Hima Vasudevan, Saishruthi Swaminathan, Saeed Aghabozorgi, Yan Luo

- Special Thanks to you, reader who is going to grade me.

Thank you!