SpaceY Mission

Jaime Sanz

24/12/2023

Presentation Index

- Executive Summary (3)
- Introduction (4)
- Methodology (6)
- Results (16)
- Conclusion (46)
- Appendix (47)

Executive Summary

For this project the following steps were completed:

- **Gathered information** from the public SpaceX API and the SpaceX page on Wikipedia to define a 'class' column that identifies whether landings were successful.
- **Analyzed the data** using SQL queries, visual representations, folium-generated maps, and interactive dashboards.
- **Wrangled the data**. Selected pertinent columns as predictors and transformed categorical variables into binary format through one-hot encoding. Standardized the dataset and utilized GridSearchCV to optimize parameters for various machine learning algorithms. Evaluated and compared the performance of each model visually by their accuracy scores.
- **Implemented four distinct ML algorithms**: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K-Nearest Neighbors. All models yielded comparably good results with an average accuracy of approximately 83.33%, although there was a tendency to overestimate the number of successful landings. Additional data is required to enhance the precision and effectiveness of the models

Introduction

Context:

- We have entered the era of commercial space exploration.
- SpaceX offers the most competitive rates at $62 million compared to $165 million USD.
- This advantage stems mainly from their capability to retrieve the initial stage of their rockets.
- Competitor Space Y is looking to rival SpaceX's achievements.

Challenge:

- Space Y has commissioned us to develop a machine learning algorithm that can accurately forecast the successful retrieval of a rocket's first stage.

Data Collection and Wrangling methodology

Approach to Data Collection:

- Merged information from SpaceX's public API with content from its Wikipedia entry.

Data Preparation:

- Processed and organized data, categorizing actual landings as either successful or not.

Data Exploration:

- Conducted exploratory data analysis (EDA) through both graphical representation and SQL queries.

Visualization:

- Created interactive visual representations using Folium for mapping and Plotly Dash for dashboard analytics.

Predictive Modelling:

- Engaged in predictive analytics with various classification algorithms.
- Optimized these models by employing GridSearchCV for hyperparameter tuning.

EDA and interactive visual analytics Methodology

Data Collection

Data Acquisition Strategy:

- The process included fetching data through Space X's public API and extracting information from a table in Space X's Wikipedia page.

Upcoming Presentations:

- The subsequent slide will present a diagram outlining the data retrieval process from the API, followed by another detailing the data extraction via web scraping.

Data Points from Space X API:

- Variables such as Flight Number, Date, Booster Version, Payload Mass, Orbit, Launch Site, Outcome, Number of Flights, Grid Fins, Reusability status, Legs, Landing Pad, Block number, Reuse Count, Serial number, and geographical coordinates (Longitude, Latitude).

Data Extracted from Wikipedia:

- Collected attributes include Flight Number, Launch Site, Payload details, Payload Mass, Orbit, Customer, Outcome of the Launch, Booster Version, Landing status of the Booster, Date, and Time.

Data Wrangling

Establishing a Training Label for Landing Results:

- Develop a label for training purposes, marking successful landings as '1' and failures as '0’.

Components of the Outcome Variable:

- The 'Outcome' variable is composed of two parts: 'Mission Outcome' and 'Landing Location’.

Creation of the 'Class' Label:

- Introduce a new column labeled 'class' assigned with a '1' when the 'Mission Outcome' is affirmative, and a '0' in all other scenarios. The assignment is as follows:
- For outcomes labeled as 'True' for ASDS (Autonomous Spaceport Drone Ship), RTLS (Return to Launch Site), and Ocean landings, the class is assigned a '1'.
- For outcomes that are 'None' or 'False' for ASDS, RTLS, Ocean landings, or combinations thereof, the class is assigned a '0'.

EDA with visualization results

Conducting Exploratory Data Analysis (EDA):

- EDA was carried out focusing on the attributes Flight Number, Payload Mass, Launch Site, Orbit, Class, and Year.

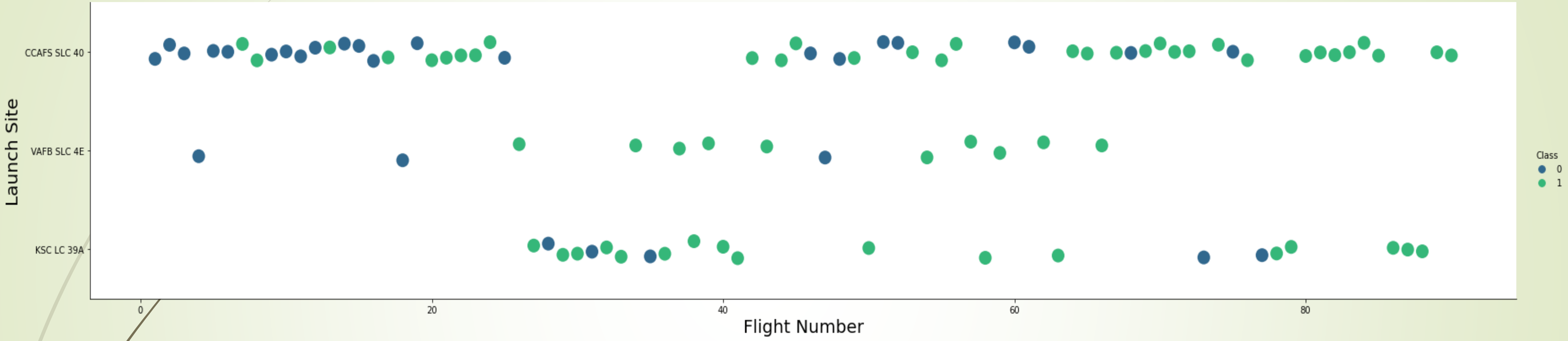
Visualization Techniques Employed:

- To discern patterns, the following comparisons were visualized: Flight Number against Payload Mass, Flight Number with respect to Launch Site, Payload Mass in relation to Launch Site, Orbit compared to Success Rate, Flight Number versus Orbit, Payload against Orbit, and the trend of success over the years.

Graphical Representations Used:

- A variety of graphs such as scatter plots for trend identification, line charts for temporal evolution, and bar plots for categorical comparison were utilized to investigate the associations among the variables. This analysis assists in determining the viability of these variables as predictors in the machine learning model.

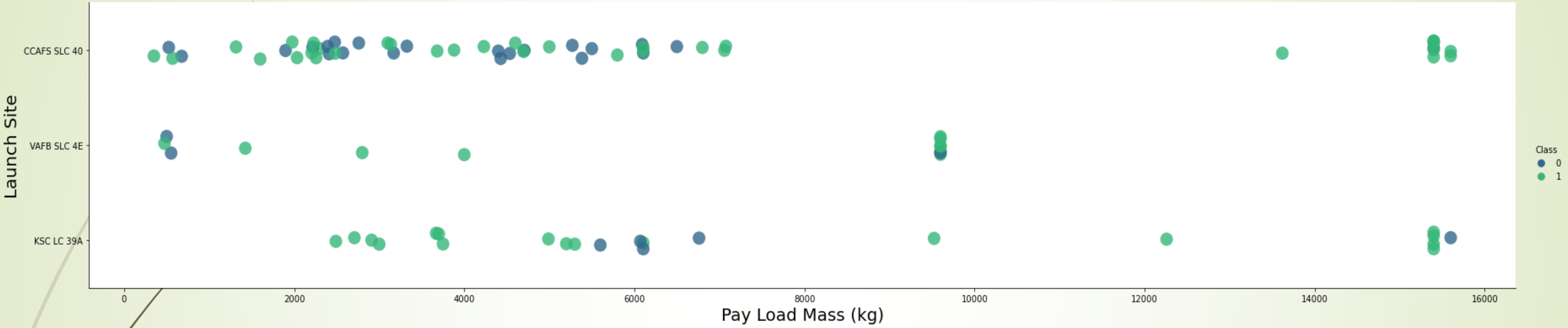
Flight Number vs. Launch Site



Green = successful launch
Purple = unsuccessful launch

Graphic suggests an increase in success rate over time (indicated in Flight Number). Likely a big breakthrough around flight 20 which significantly increased success rate. CCAFS appears to be the main launch site as it has the most volume.

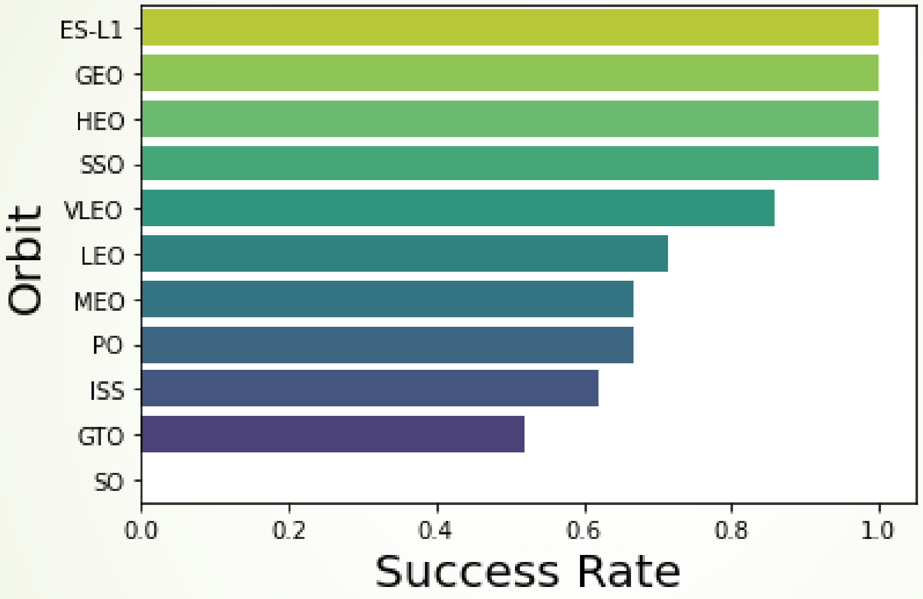
Payload vs. Launch Site



Green = successful launch
Purple = unsuccessful launch

Payload mass appears to fall mostly between 0-6000 kg. Different launch sites also seem to use different payload mass.

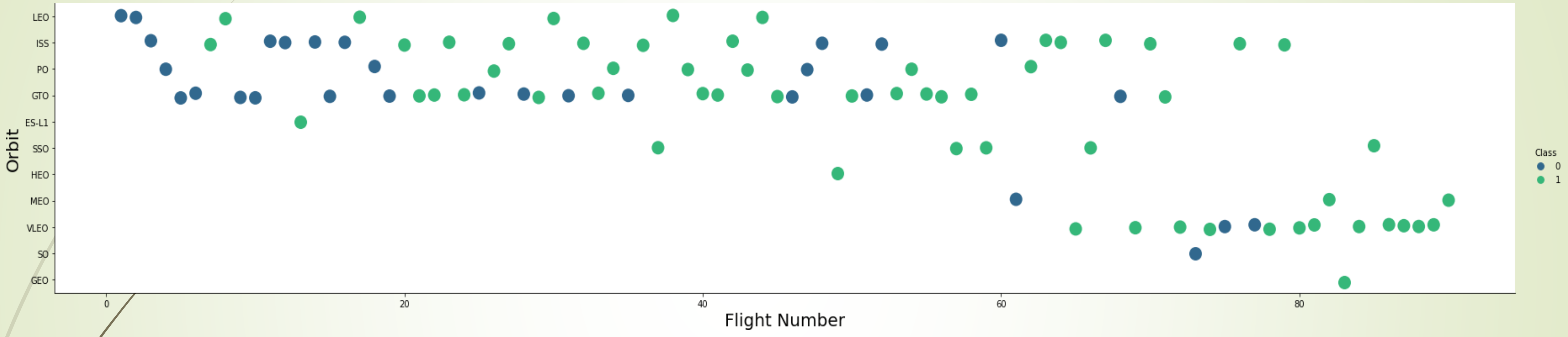
Success rate vs. Orbit type



Insights:

- ES-L1 (1), GEO (1), HEO (1) have 100% success rate (sample sizes in parenthesis)
- SSO (5) has 100% success rate
- VLEO (14) has a 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

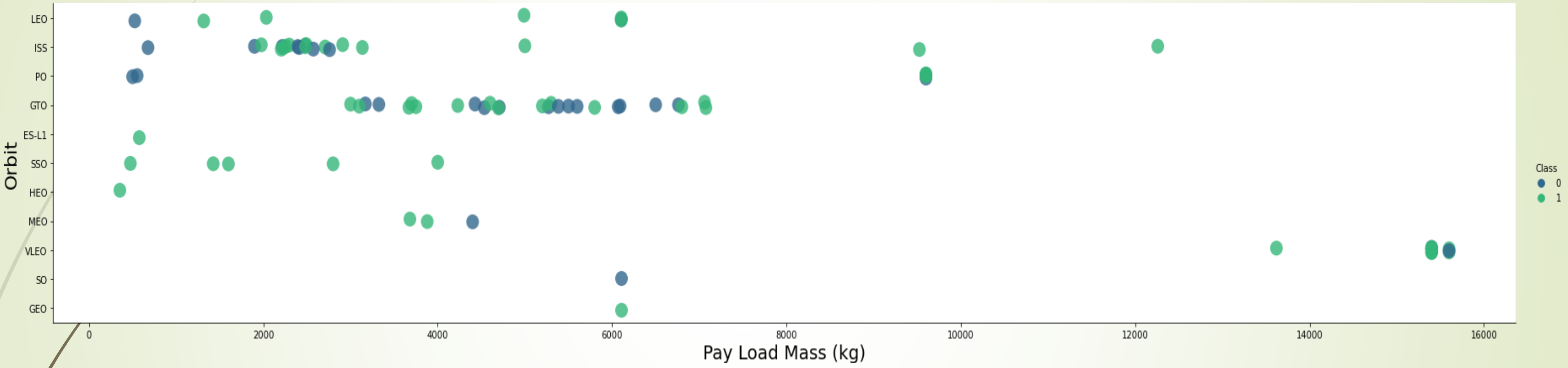


Green = successful launch
Purple = unsuccessful launch

Insights:

- 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



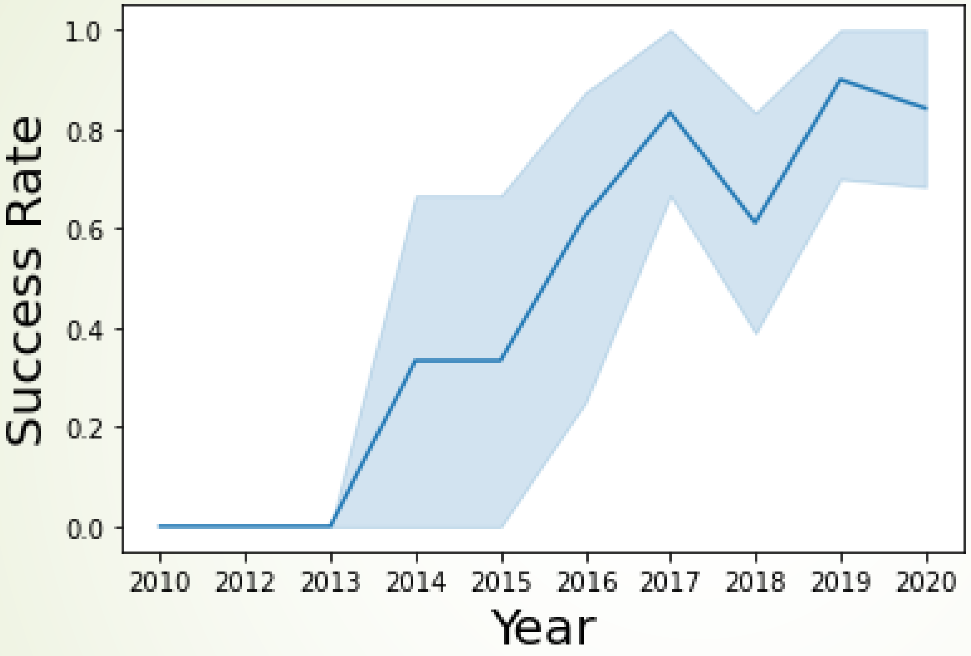
Green = successful launch

Purple = unsuccessful launch

Insights:

- 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



95% confidence interval
(light blue shading)

Insights:

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

EDA with SQL results

Data Integration:

- The dataset was imported into the IBM DB2 database system.

Data Querying Process:

- Utilized the SQL Python interface for executing queries.

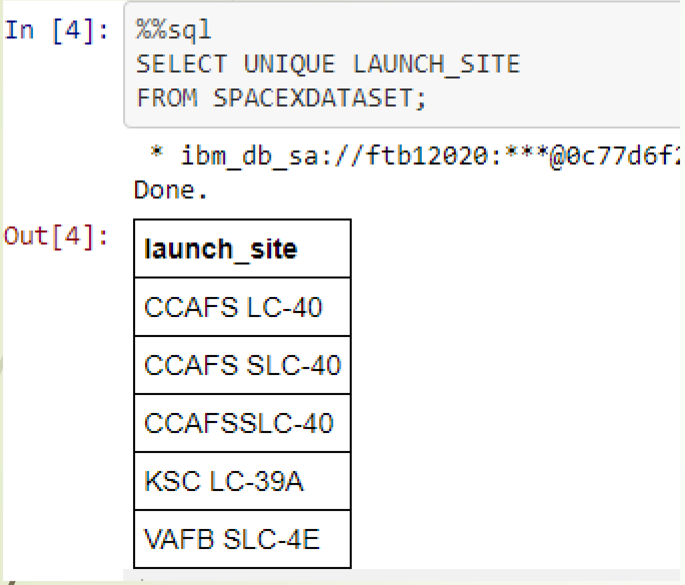
Objective of Queries:

- The intent behind the queries was to deepen the understanding of the data collected.

Scope of Data Extraction:

- Information was retrieved regarding the names of launch sites, the results of missions, the diverse payload capacities of customers, different versions of boosters, and the outcomes of landings.

All Launch Site Names



In [4]:

```
%%sql
SELECT UNIQUE LAUNCH_SITE
FROM SPACEXDATASET;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f
Done.
```

Out[4]:

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

Query unique launch site names from database.

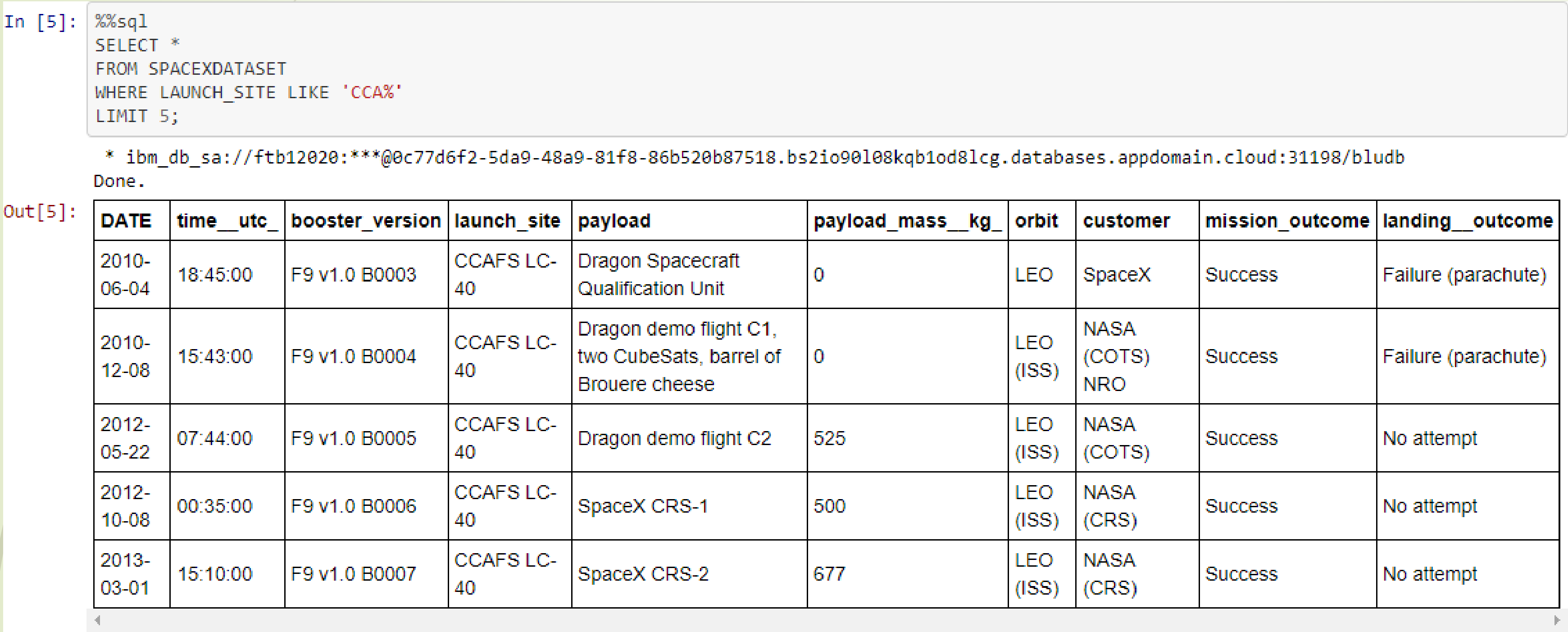
CCAFS SLC-40 and CCAFSSLC-40 likely all represent the same launch site with data entry errors.

CCAFS LC-40 was the previous name.

Likely only 3 unique launch_site values:

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

Launch Site Names Beginning with `CCA`



In [5]:

```
%%sql
SELECT *
FROM SPACEXDATASET
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```

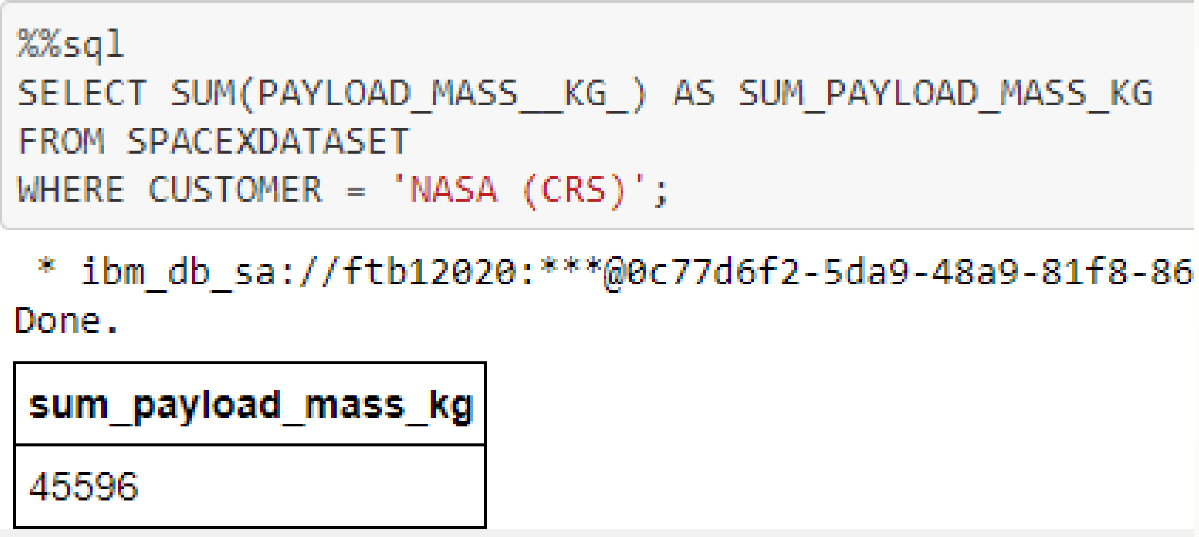
```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[5]:

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 from NASA



```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS SUM_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE CUSTOMER = 'NASA (CRS)';
```

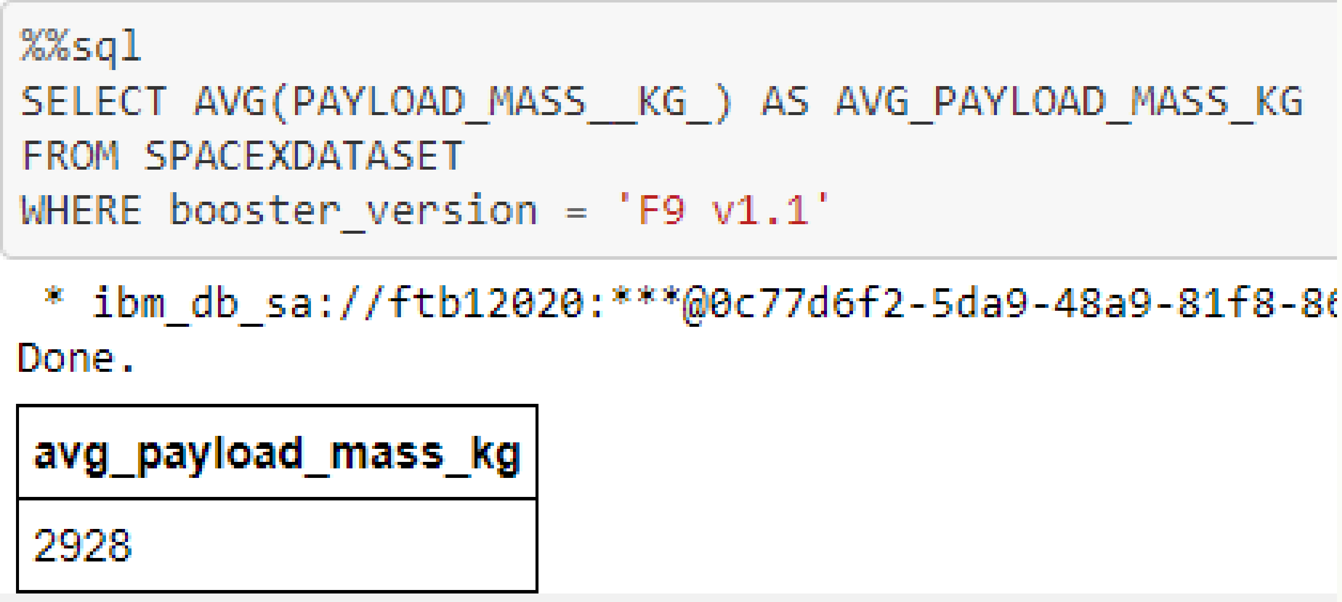
```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86
Done.
```

sum_payload_mass_kg
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 AVG_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE booster_version = 'F9 v1.1'
```

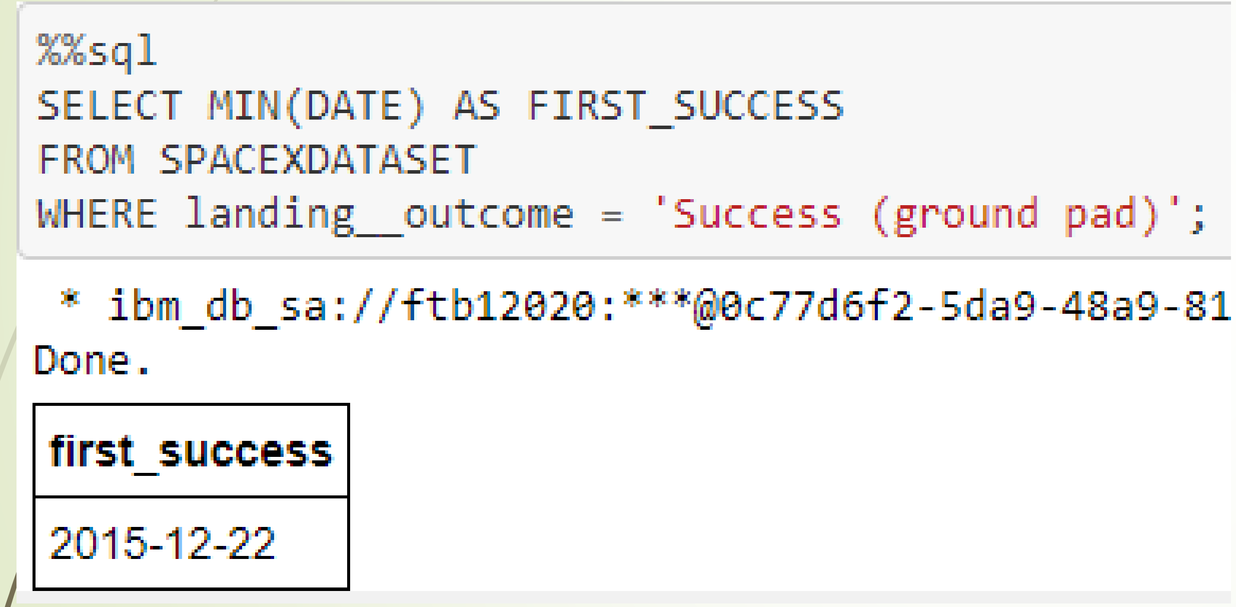
```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86
Done.
```

avg_payload_mass_kg
2928

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 Pad Landing Date



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

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81
Done.
```

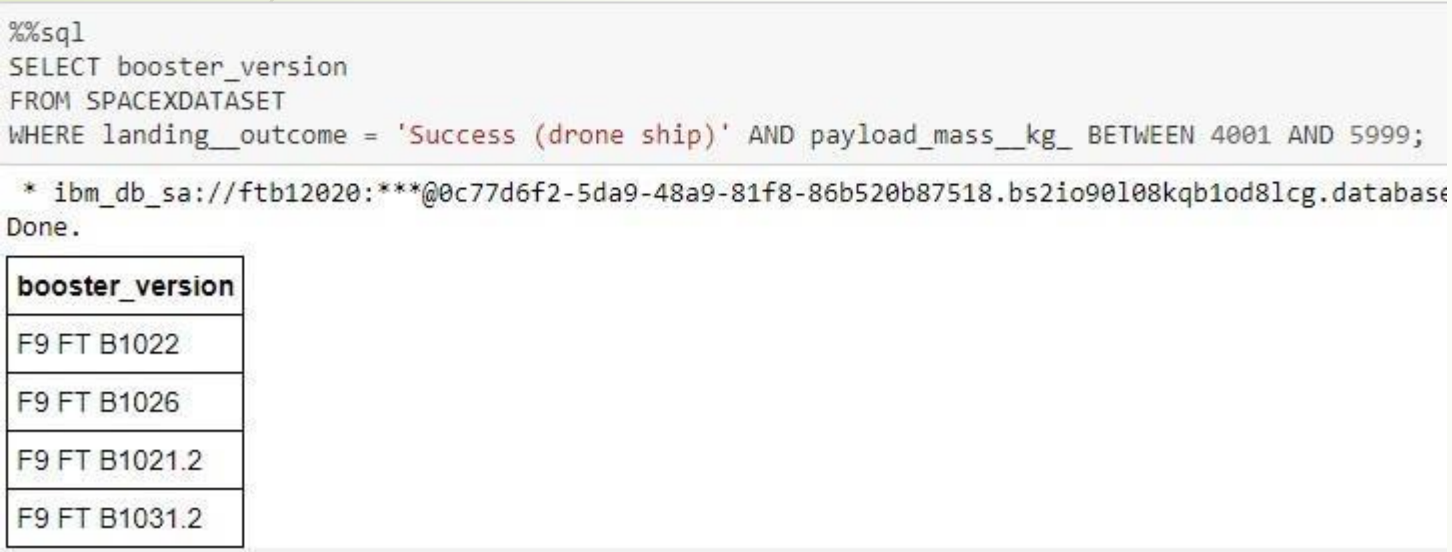
first_success
2015-12-22

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



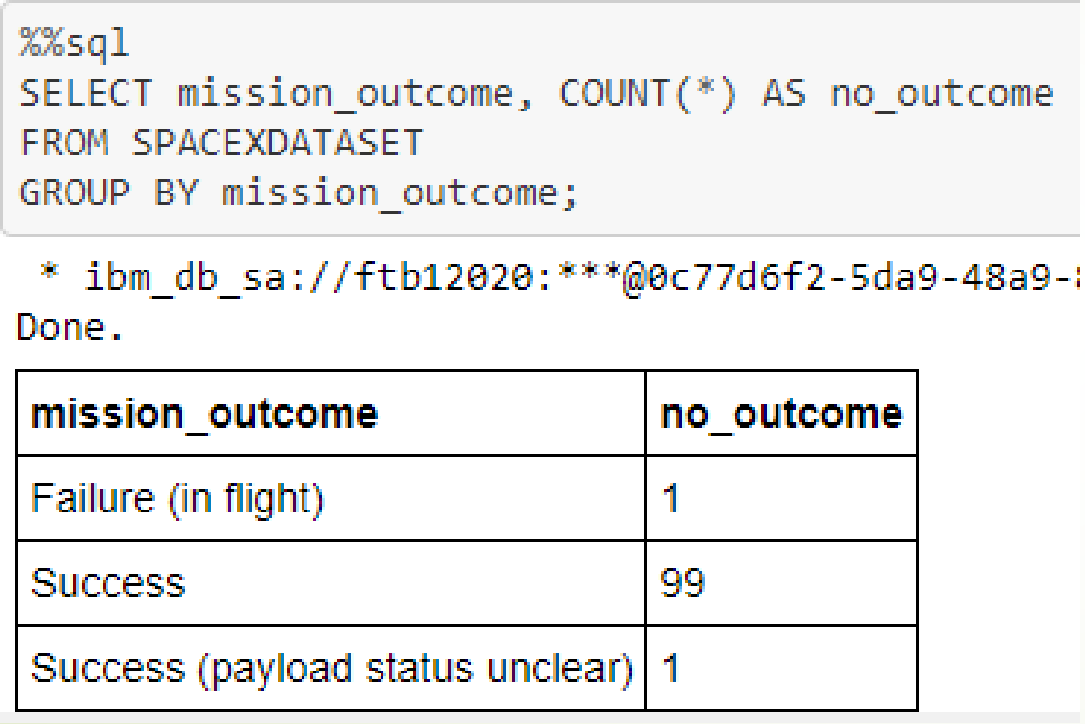
```
%%sql
SELECT booster_version
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4001 AND 5999;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.database
Done.
```

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

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

Total Number of Each Mission Outcome



```
%%sql
SELECT mission_outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
GROUP BY mission_outcome;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-
Done.
```

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

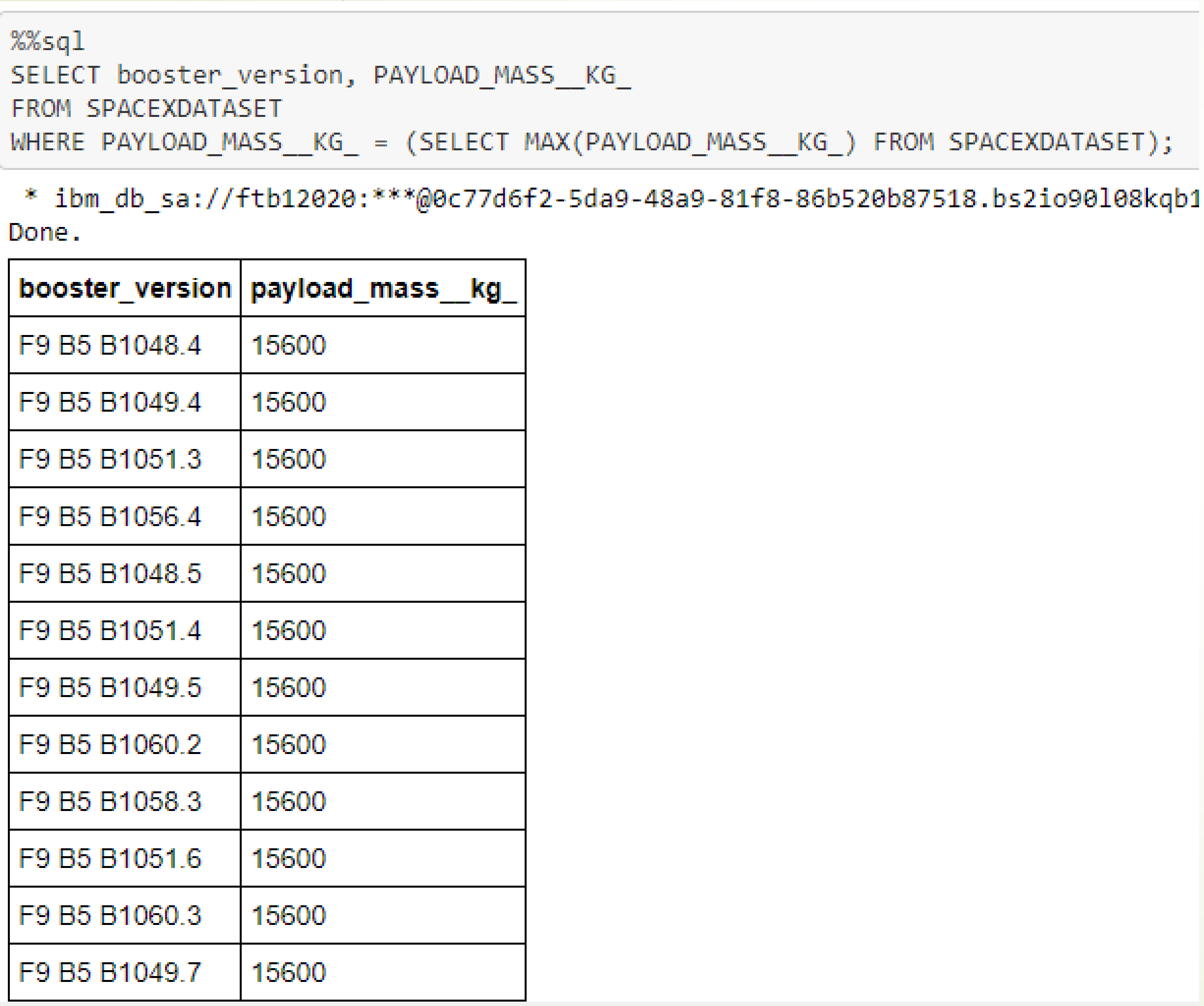
This query returns a count of each mission outcome.

SpaceX appears to achieve its mission outcome nearly 99% of the time.

This means that most of the landing failures are intended.

Interestingly, one launch has an unclear payload status and unfortunately one failed in flight.

Boosters that Carried Maximum Payload



```
%%sql
SELECT booster_version, PAYLOAD_MASS__KG_
FROM SPACEXDATASET
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXDATASET);
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1
Done.
```

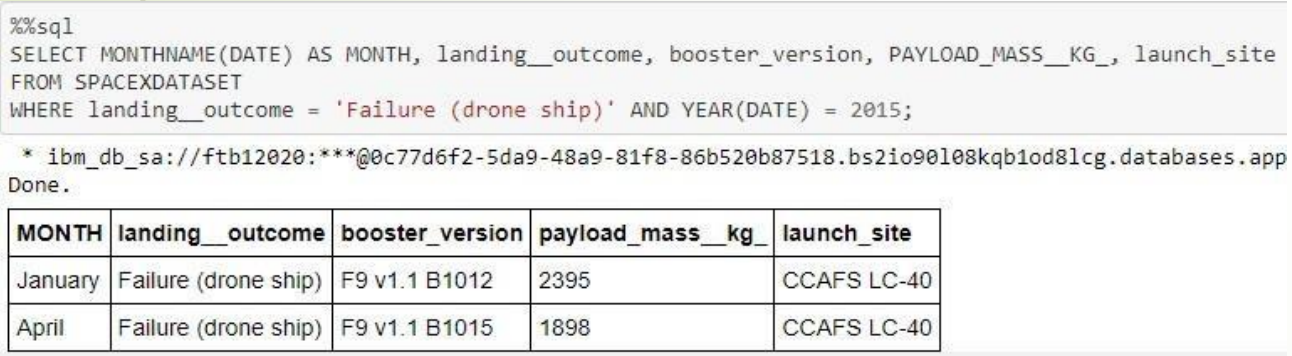
booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

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 Failed Drone Ship Landing 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;
```

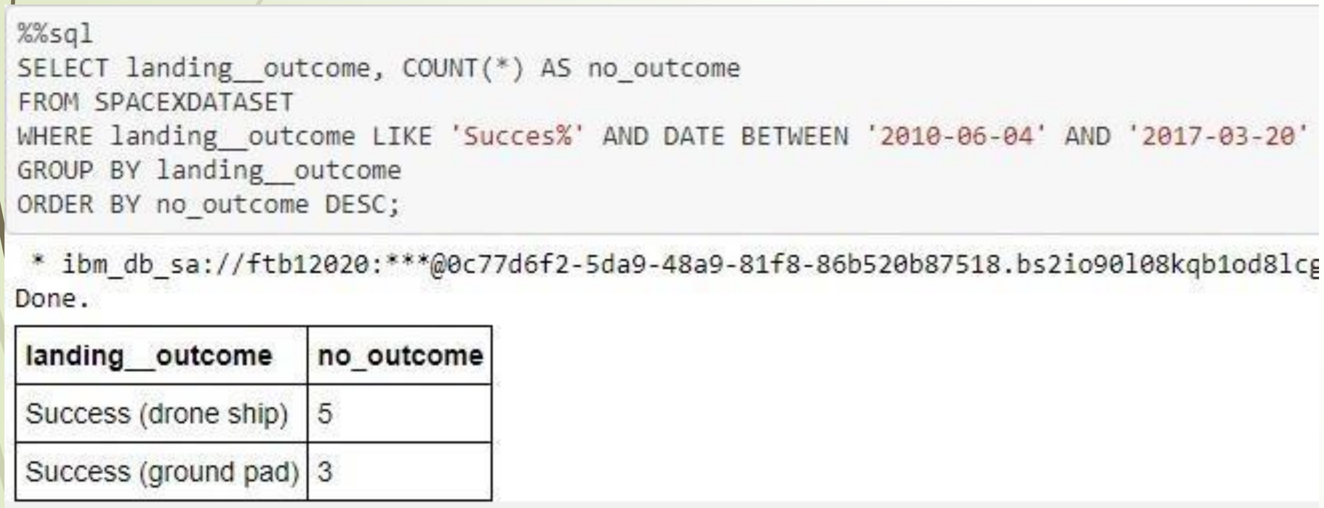
```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.app
Done.
```

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.

Ranking Counts of Successful Landings 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;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg
Done.
```

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.

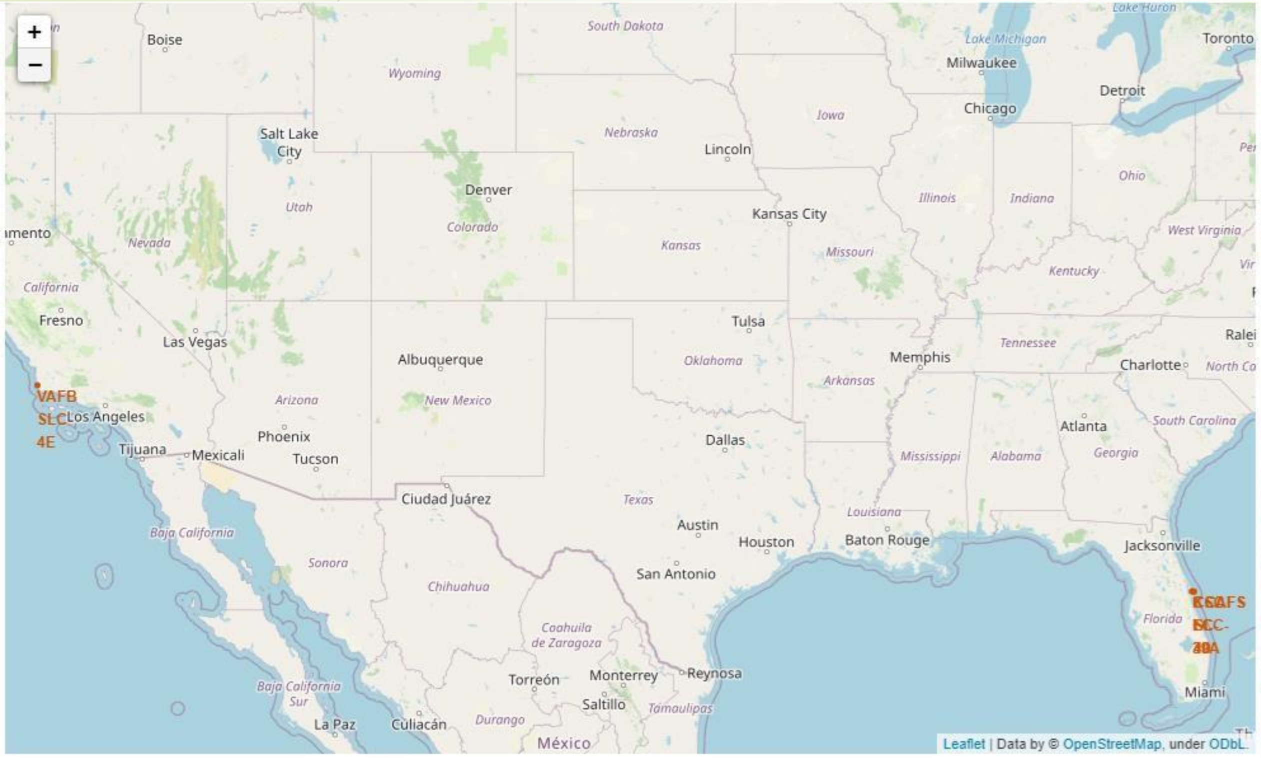
There are two types of successful landing outcomes: drone ship and ground pad landings.

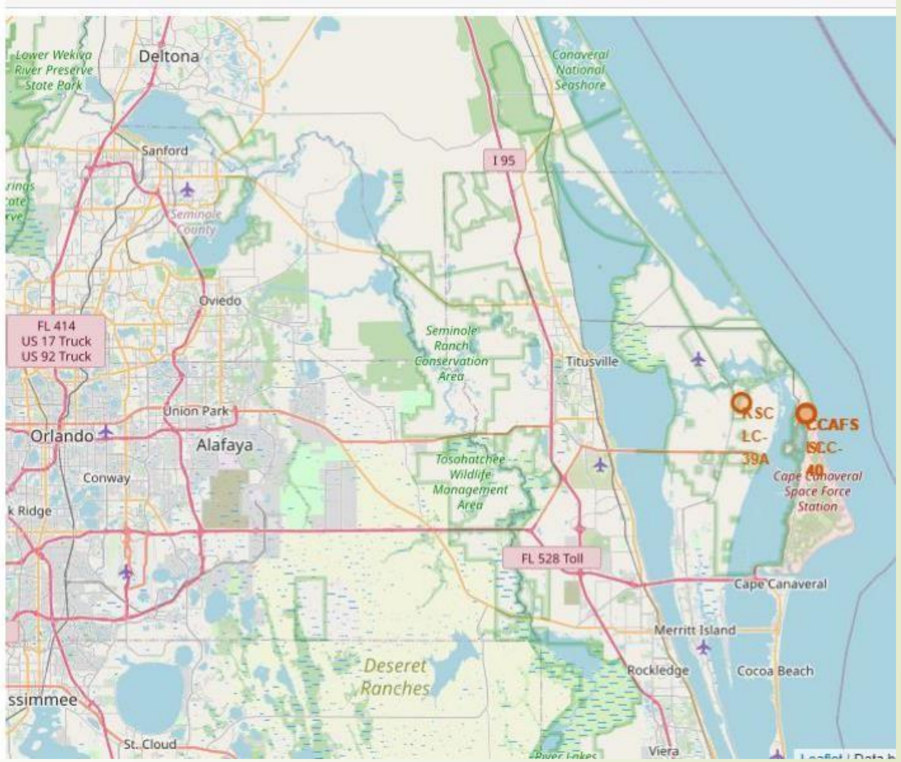
There were 8 successful landings in total during this time period

Interactive map with Folium results

- Folium maps are utilized to pinpoint rocket launch sites and denote the sites of both successful and unsuccessful landings, along with examples showcasing their proximity to essential infrastructure such as railways, highways, coasts, and cities. This visualization aids in comprehending the strategic placement of launch sites as well as the correlation between landing outcomes and their geographical locations

Launch Site Locations

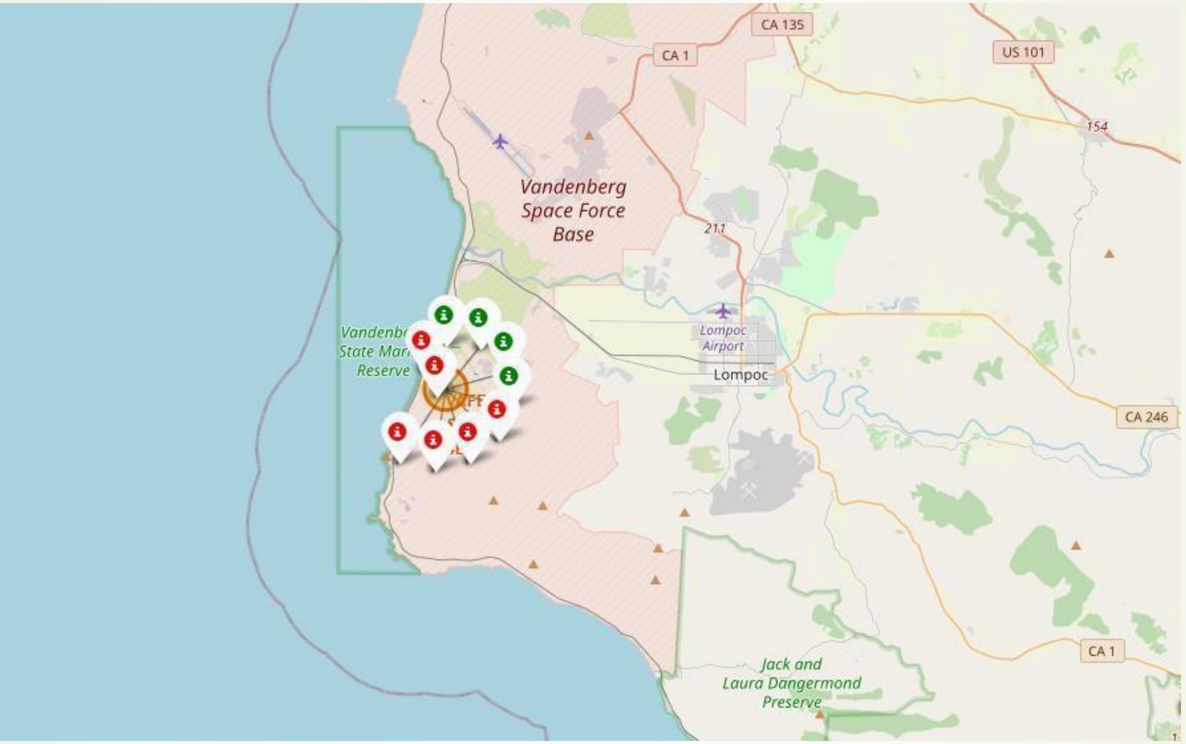




The left map shows all launch sites relative US map.

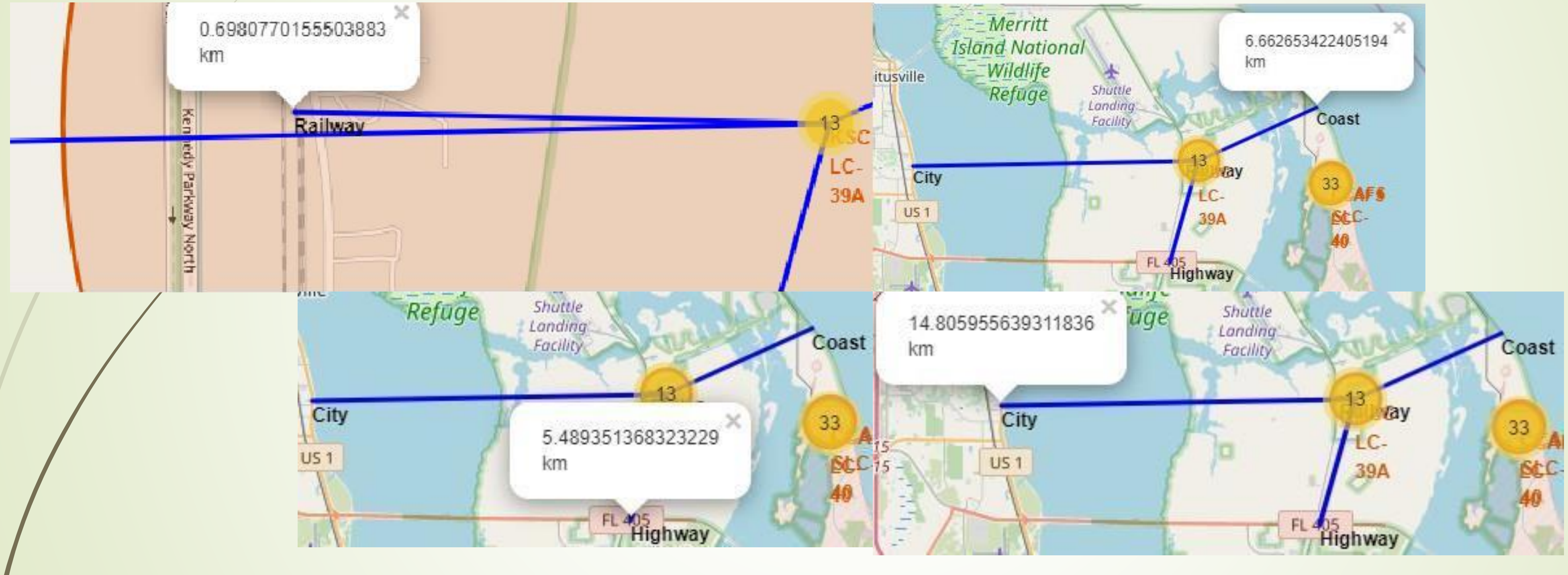
The right map shows the two Florida launch sites since they are very close to each other. All launch sites are near the ocean.

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



0.6980770155503883 km

6.662653422405194 km

5.489351368323229 km

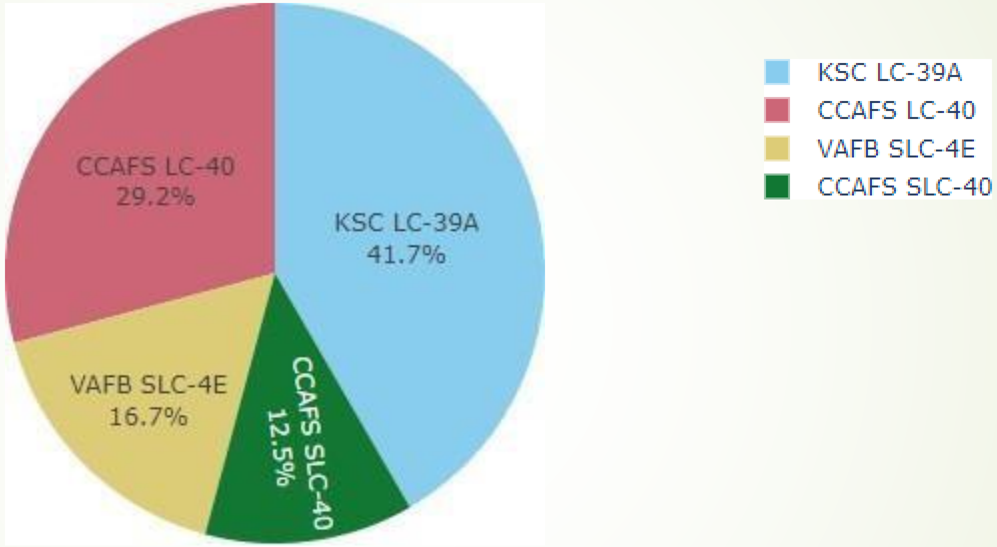
14.805955639311836 km

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.

Plotly Dash dashboard results

- Dashboard includes a pie chart and a scatter plot.
- The **pie chart** is interactive, allowing users to view the overall distribution of successful landings at all launch sites or to filter by each individual site’s success rate.
- The **scatter plot** is designed to accept two types of input: a choice between all sites or a specific site, and a range of payload masses from 0 to 10,000 kg, adjustable via a slider.
- This pie chart serves to illustrate the success rates associated with various launch sites. Meanwhile, the scatter plot offers insights into the variation of success rates based on different launch sites, the mass of the payload, and categories of booster versions

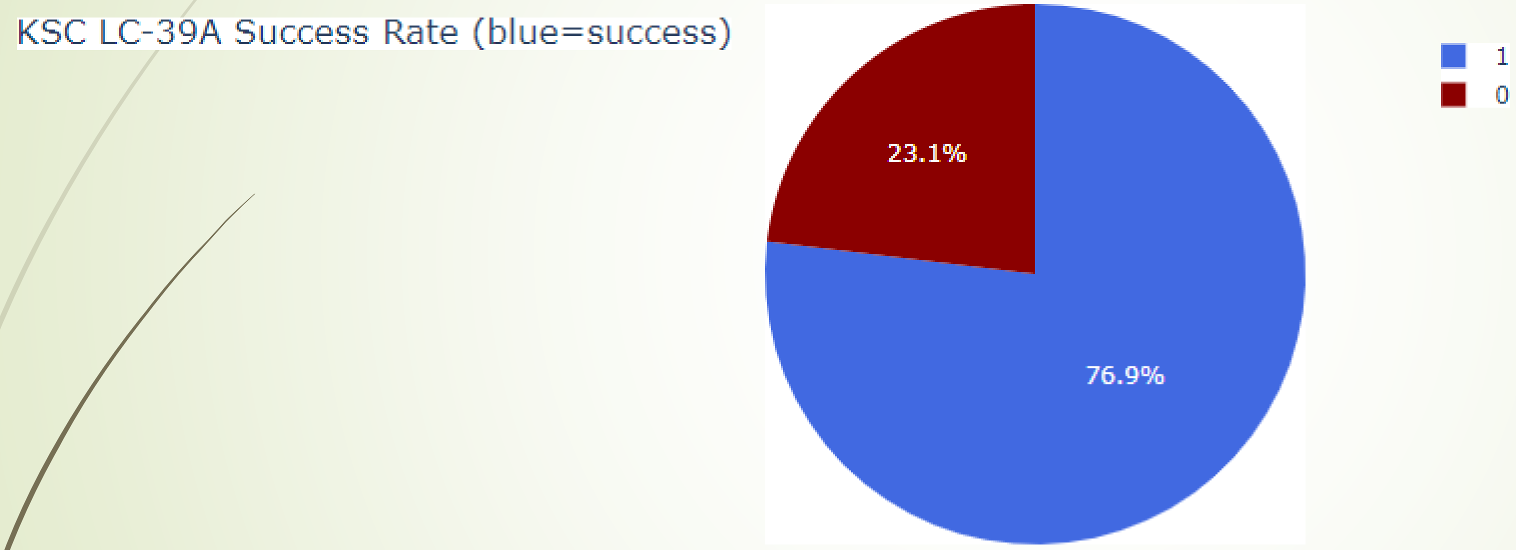
Successful Launches Across Launch Sites



Insights:

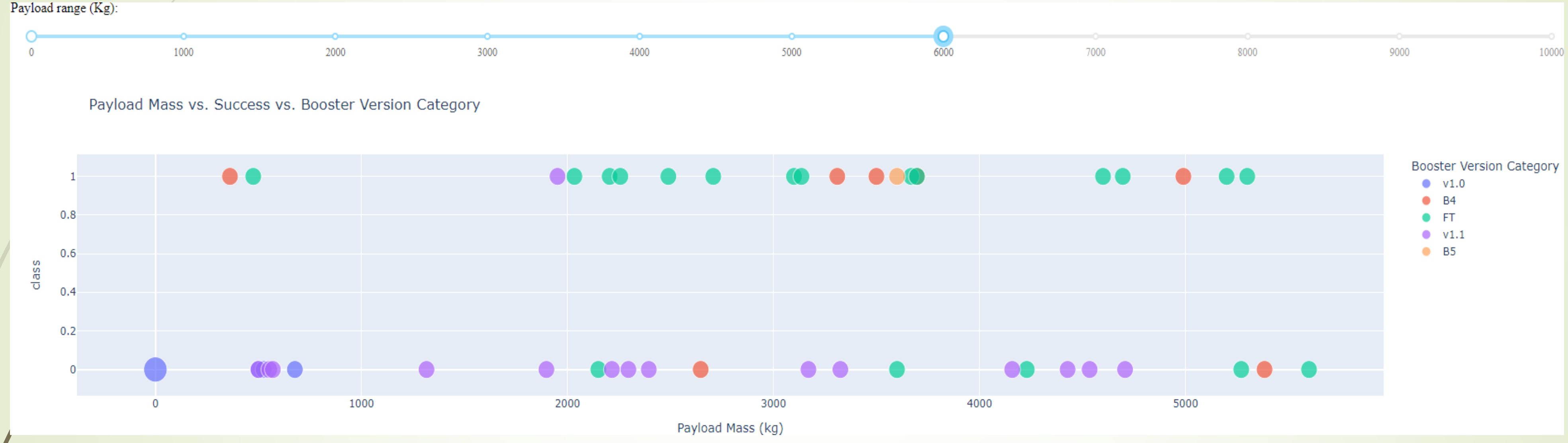
- This is the distribution of successful landings across all launch sites.
- CCAFS LC-40 is the old name of CCAFS SLC-40 so CCAFS and KSC have the same amount of successful landings, but a majority of the successful landings where performed before the name change.
- VAFB has the smallest share of successful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

Highest Success Rate Launch Site



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

Payload Mass vs. Success vs. Booster Version Category



Insights:

- 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 0 kg.

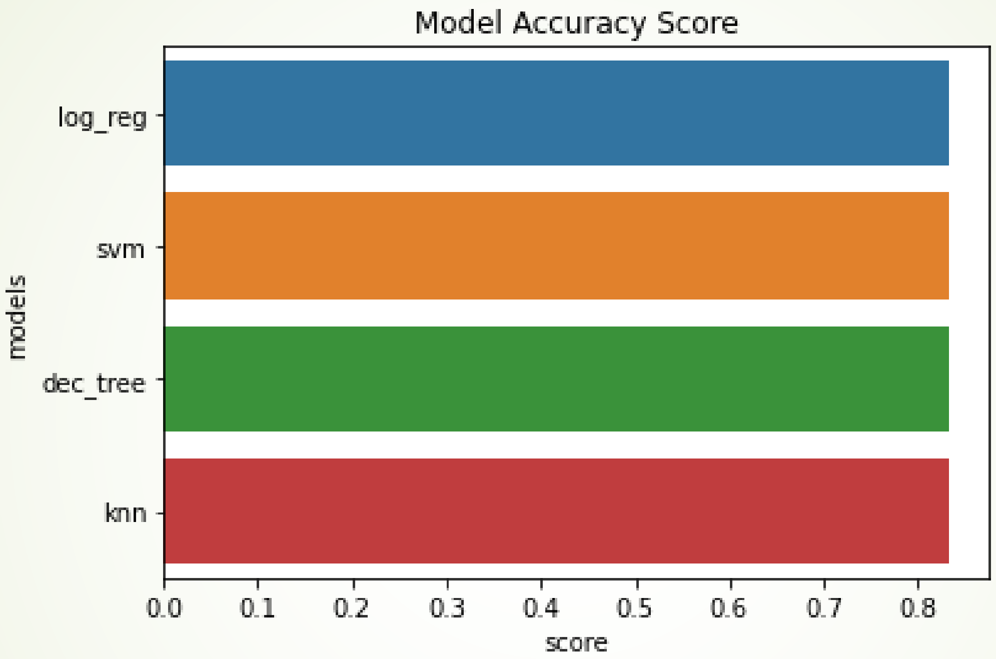
Predictive analysis methodology

- Load the Dataset
- Standardize the data
- Split the data into training and testing data
- Create a model with a GridSearchCV object
- Calculation of accuracy on test data using method score
- Calculation confusion matrix
- Find the best accuracy with a model

Predictive analysis (classification) results

GRIDSEARCHCV(CV=10) ON LOGISTIC REGRESSION, SVM, DECISION TREE, AND KNN

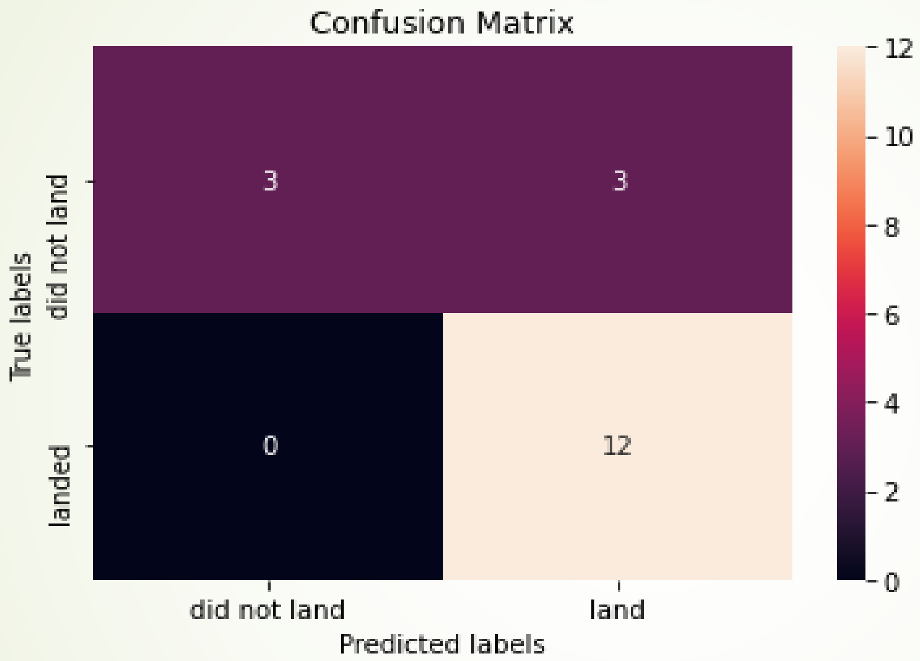
Classification Accuracy



Insights:

- All models had virtually the same accuracy on the test set at **83.33% accuracy**. It should be noted that test size is small at only sample size of18.
- 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



Correct predictions are on a diagonal from top left to bottom right.

Insights:

- Since all models performed the same for the test set, the confusion matrix is the same across all models. 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.

Conclusion

- Our mission was to construct an algorithm for Space Y, aimed at challenging SpaceX's market position.
- The model's objective is to forecast the successful descent of Stage 1, potentially conserving approximately $100 million USD.
- We harnessed data from SpaceX's public API and the SpaceX Wikipedia entry through web scraping techniques.
- Data categorization was conducted, and the resulting dataset was stored in a DB2 SQL database.
- A visualization dashboard was developed.
- The machine learning model we formulated achieved an 83% prediction accuracy.
- Allon Mask from SpaceY can utilize this model to assess the likelihood of a successful Stage 1 landing prior to launch, aiding in launch decision-making.
- To refine the model's predictive capabilities and enhance its accuracy, additional data should be amassed and analyzed.

Thank you!