IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

SINDHE PANDURANG

08-08-2023

Outline

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

Executive Summary

- - Utilized data from public SpaceX API and SpaceX Wikipedia page.
- - Introduced a 'class' column for categorizing successful landings.
- - Analyzed data using SQL, data visualization, Folium maps, and interactive dashboards.
- - Engineered features, one-hot encoded categorical variables, and standardized data.
- - Employed GridSearchCV to optimize machine learning algorithms, achieving around 83.33% accuracy.
- - Developed four machine learning models: Logistic Regression, SVM, Decision Tree, KNN.
- - Attained consistent 83.33% accuracy across models.
- - Models tended to overestimate successful landings.
- - Dataset enrichment needed for improved model performance and precision.

Introduction

Background:

- Welcome to the era of Commercial Space Exploration.
- SpaceX leads with competitive pricing ($62M vs. $165M USD).
- Key factor: Rocket part recovery, mainly Stage 1.

- ## The Challenge:

- Space Y aims to rival SpaceX's success.
- Assigned task: Develop a machine learning model.
- Objective: Predict the likelihood of a successful Stage 1 recovery.

Section 1

Methodology

Methodology

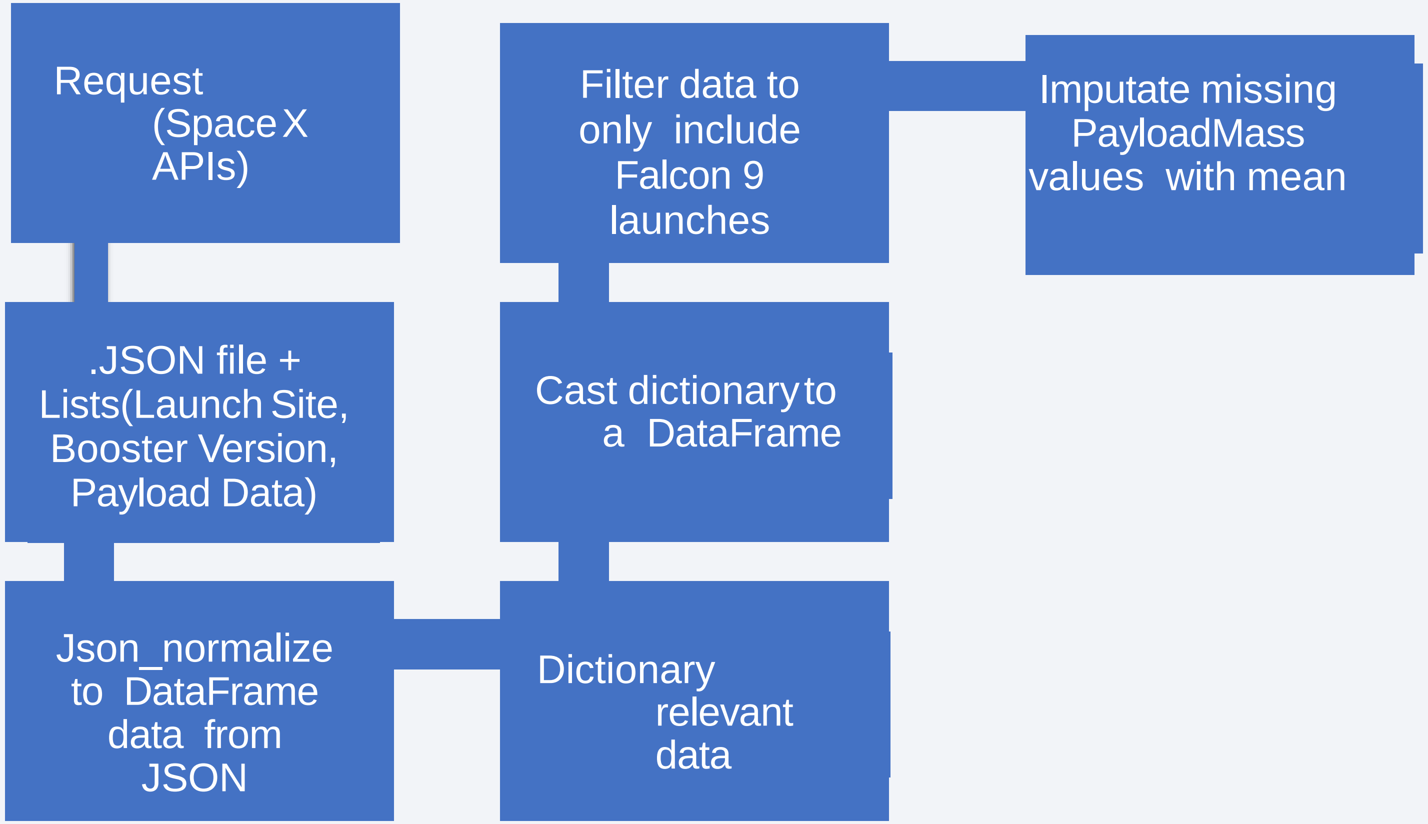
- 1. Data Gathering Approach:
- - Aggregated information from SpaceX's official API and Wikipedia page
- 2. Outcome Classification:
- - Distinguished between successful and unsuccessful landings
- 3. Exploratory Data Analysis:
- - Employed visualization tools and SQL queries
- 4. Interactive Visual Analytics:
- - Utilized Folium for geographical visualizations
- - Employed Plotly Dash for interactive data presentation
- 5. Predictive Analysis:
- - Implemented classification models for predictive insights
- - Fine-tuned models using GridSearchCV for optimization

Data Collection

- Data aggregation involved a blend of Space X public API queries and extracting data from a table in Space X’s Wikipedia page.
- Slide 2 illustrates the API-based data collection flow, followed by Slide 3 showcasing the web scraping data collection process.
- 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: https://github.com/SindhePandurangBITS/IBM_DS_Professional/blob/a809dddf6357089f018a1a5aa67517f1f76e20ee/jupyter-labs-spacex-data-collection-api.ipynb



Request (Space X APIs)

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

Json_normalize to DataFrame data from JSON

Dictionary relevant data

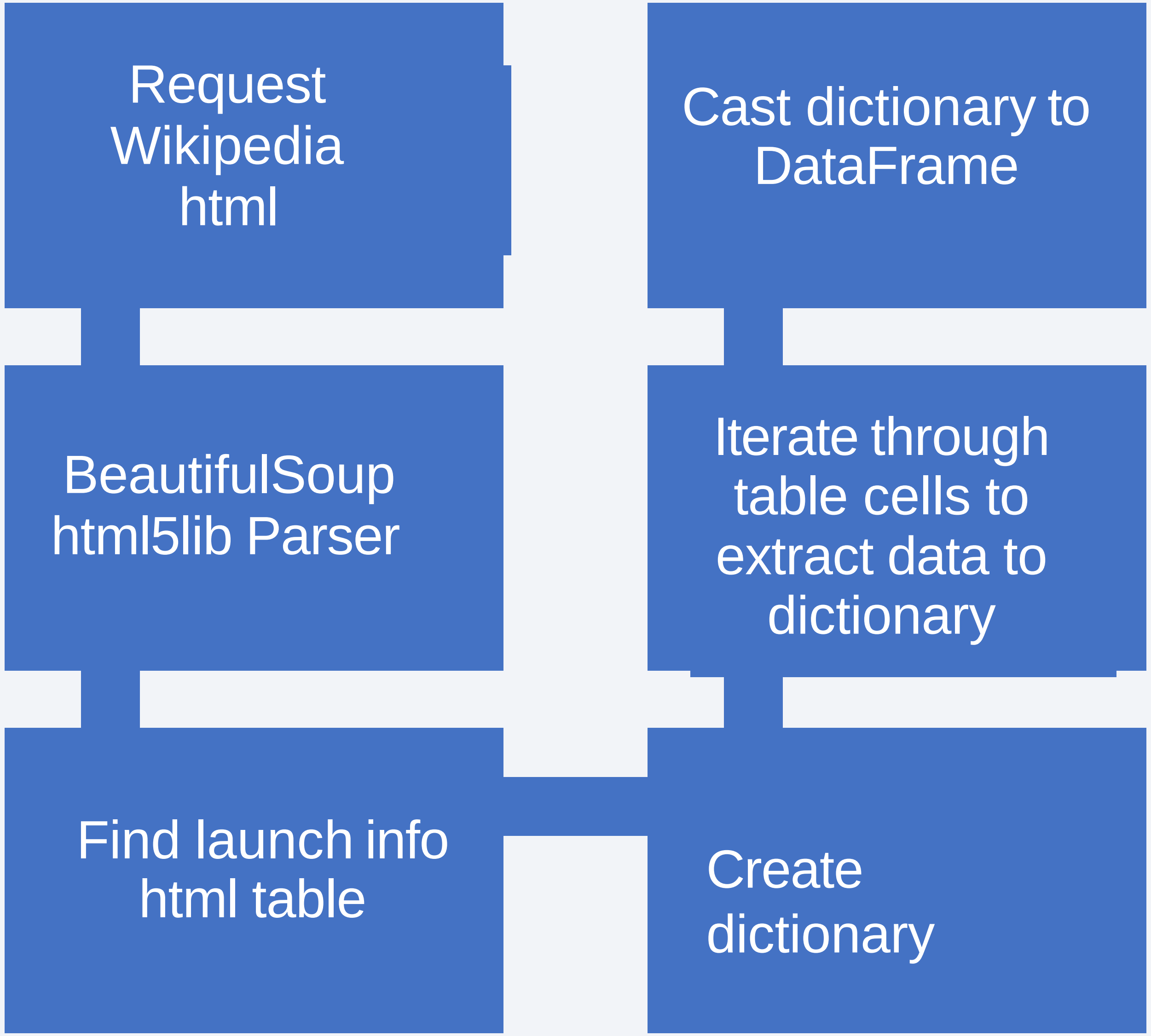
Cast dictionary to a DataFrame

Filter data to only include Falcon 9 launches

Imputate missing PayloadMass values with mean

Data Collection - Scraping

- GitHub:https://github.com/SindhePandurangBITS/IBM_DS_Professional/blob/a809dddf6357089f018a1a5aa67517f1f76e20ee/jupyter-labs-webscraping.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

- Introduce a binary training label: 1 for success, 0 for failure.
- The label derives from two aspects: 'Mission Outcome' and 'Landing Location'.
- A new column, 'class', is added to indicate success (1) or failure (0).
- Success conditions: True ASDS, True RTLS, True Ocean.
- Failure conditions: None, False ASDS, None ASDS, False Ocean, False RTLS.
- GitHub url:

https://github.com/SindhePandurangBITS/IBM_DS_Professional/blob/a809dddf6357089f018a1a5aa67517f1f76e20ee/IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.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/SindhePandurangBITS/IBM_DS_Professional/blob/a809dddf6357089f018a1a5aa67517f1f76e20ee/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- Uploaded dataset into IBM DB2 Database.
- Employed SQL-Python integration for querying.
- Executed queries to comprehend the dataset.
- Extracted details on launch site names, mission outcomes, customer payload sizes, booster versions, and landing results.
- GitHub url:

https://github.com/SindhePandurangBITS/IBM_DS_Professional/blob/a809dddf6357089f018a1a5aa67517f1f76e20ee/jupyter-labs-eda-sql-coursera_sqllite.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/SindhePandurangBITS/IBM_DS_Professional/blob/a809dddf6357089f018a1a5aa67517f1f76e20ee/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

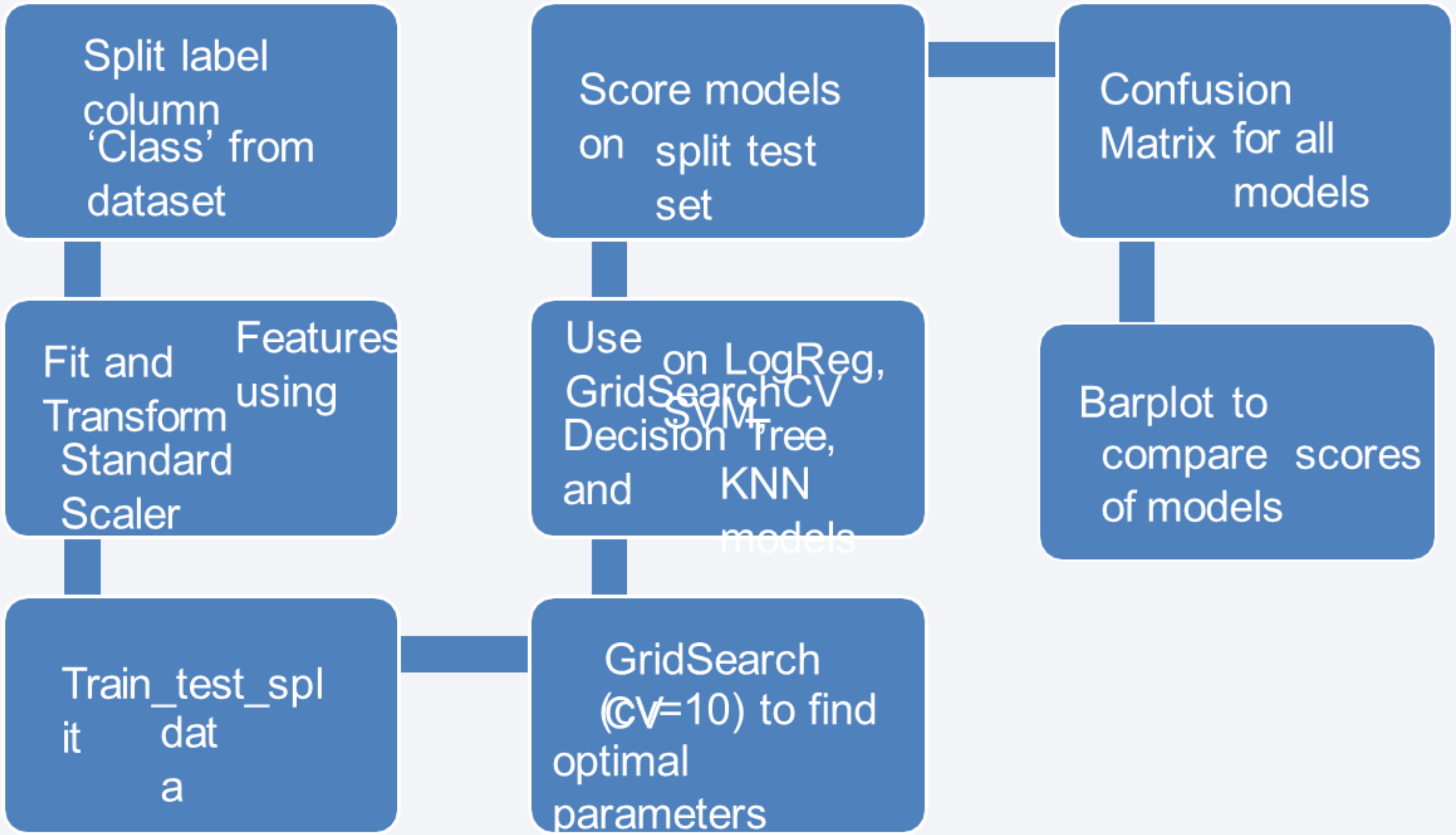
- The dashboard features a pie chart displaying overall successful landing distribution and site-specific success rates.
- It also includes a scatter plot for exploring success across sites, payload mass, and booster versions.
- The scatter plot's inputs are launch sites and payload mass (0-10000 kg).
- GitHub url:

https://github.com/SindhePandurangBITS/IBM_DS_Professional/blob/a809dddf6357089f018a1a5aa67517f1f76e20ee/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

Predictive Analysis (Classification)

- GitHub Url:

https://github.com/SindhePandurangBITS/IBM_DS_Professional/blob/a809dddf6357089f018a1a5aa67517f1f76e20ee/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb



Split label column 'Class' from dataset

Fit and Transform Features using Standard Scaler

Train_test_split data

GridSearch (CV=10) to find optimal parameters

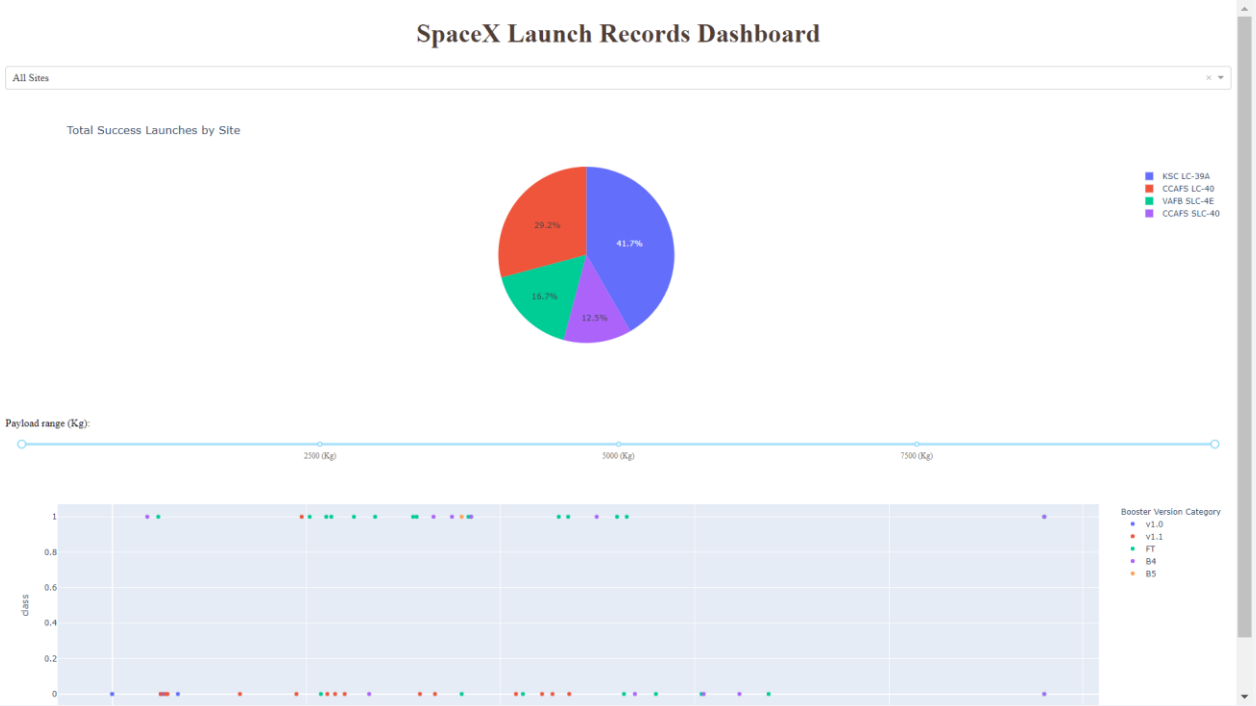
Use GridSearchCV on LogReg, SVM, Decision Tree, and KNN models

Score models on split test set

Confusion Matrix for all models

Barplot to compare scores of models

Results



SpaceX Launch Records Dashboard

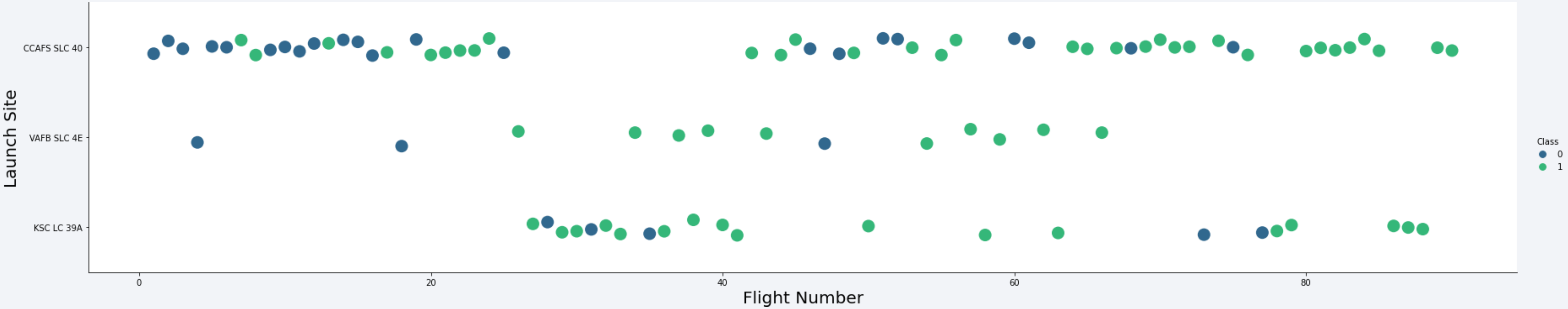
All Sites

- 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

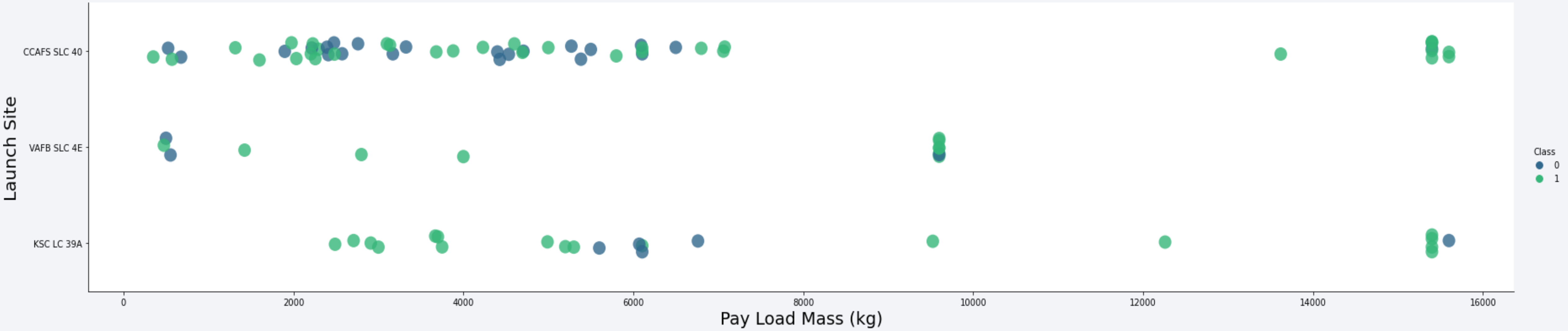
Insights drawn from EDA

Flight Number vs. Launch Site



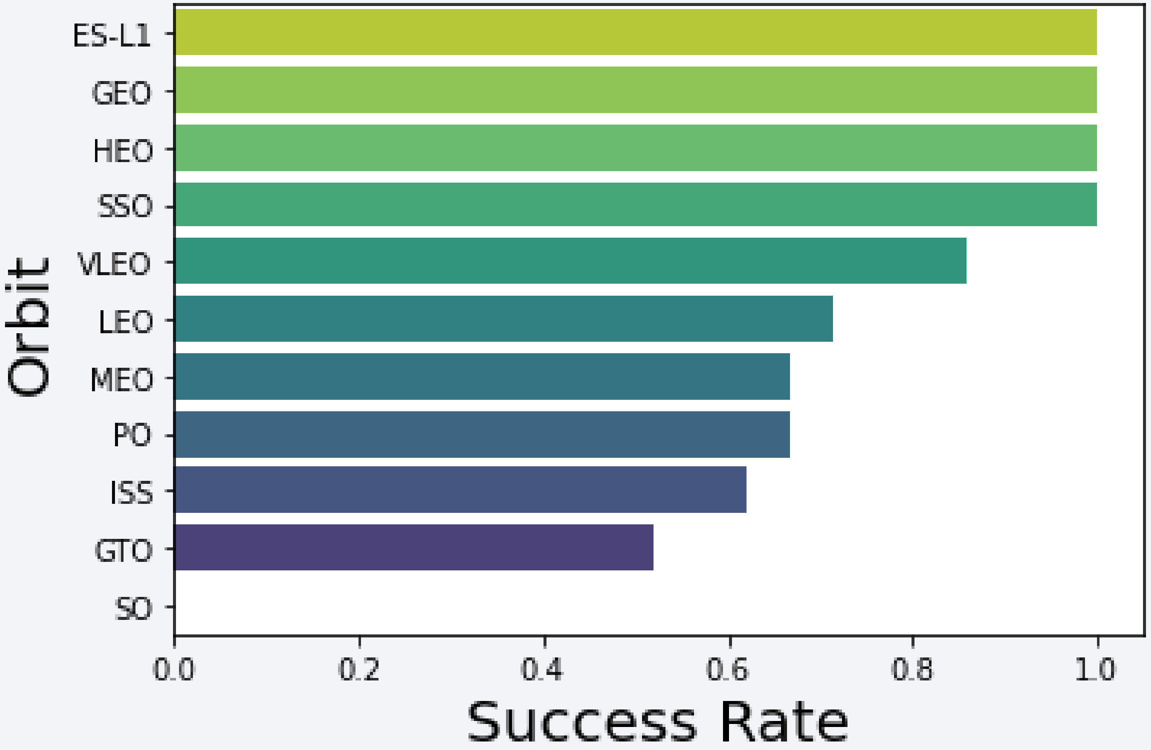
- Graphic suggests an increase in success rate over time (indicated in Flight Number). Likely a big breakthrough around flight 20 which significantly increased success rate. CCAFS appears to be the main launch site as it has the most volume.

Payload vs. Launch Site



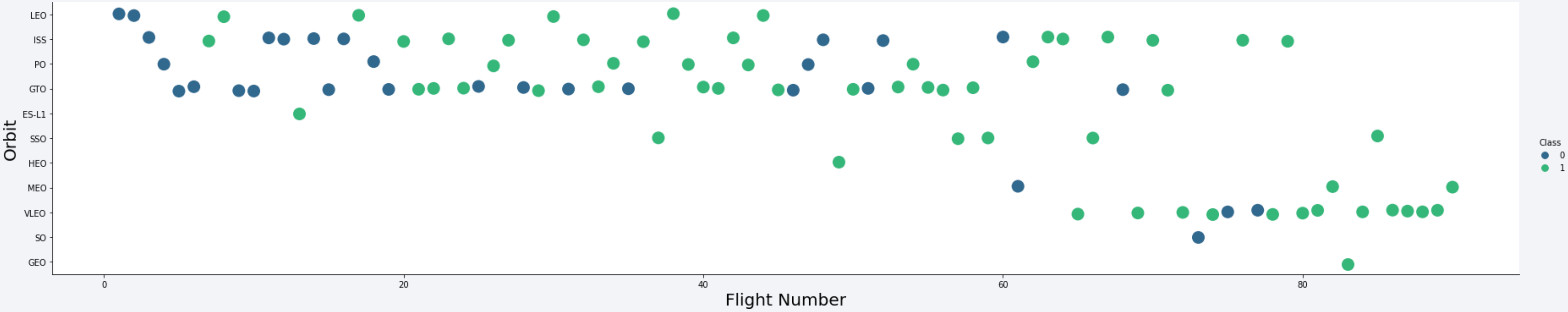
- Payload mass appears to fall mostly between 0-6000 kg.
- Different launch sites also seem to use different payload mass.

Success Rate vs. Orbit Type



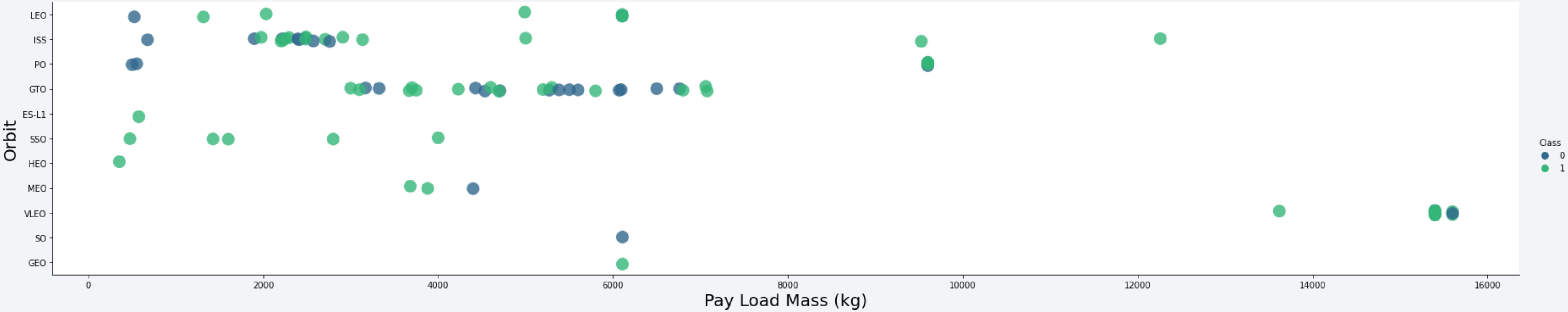
- ES-L1 (1), GEO (1), HEO (1) have 100% success rate (sample sizes in parenthesis) SSO (5) has 100% success rate
- VLEO (14) has decent success rate and attempts
- SO (1) has 0% success rate
- GTO (27) has the around 50% success rate but largest sample

Flight Number vs. Orbit Type



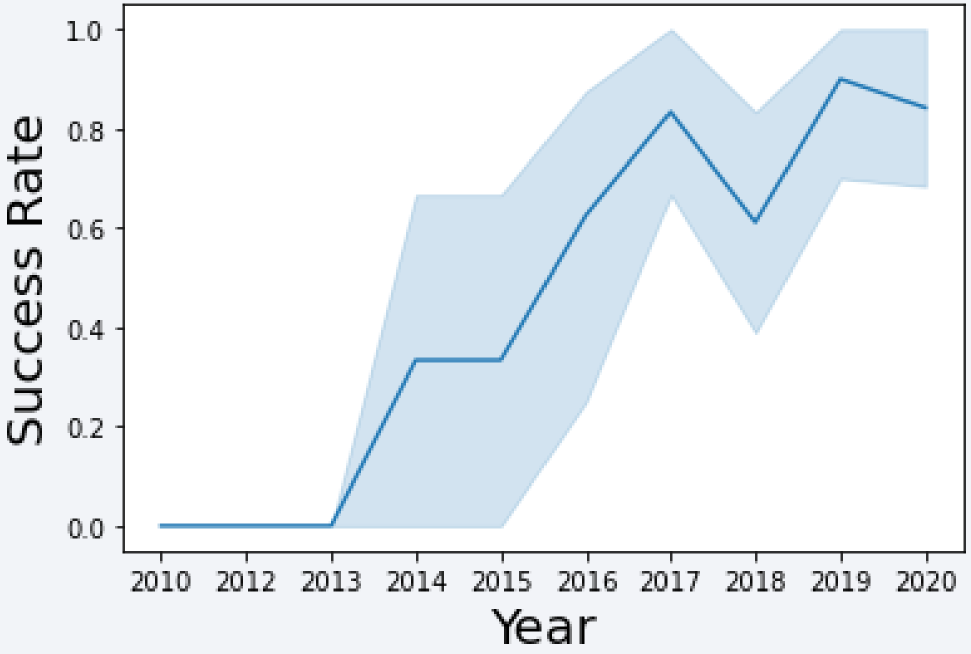
- Launch Orbit preferences changed over Flight Number. Launch Outcome seems to correlate with this preference.
- SpaceX started with LEO orbits which saw moderate success LEO and returned to VLEO in recent launches SpaceX appears to perform better in lower orbits or Sun-synchronous orbits

Payload vs. Orbit Type



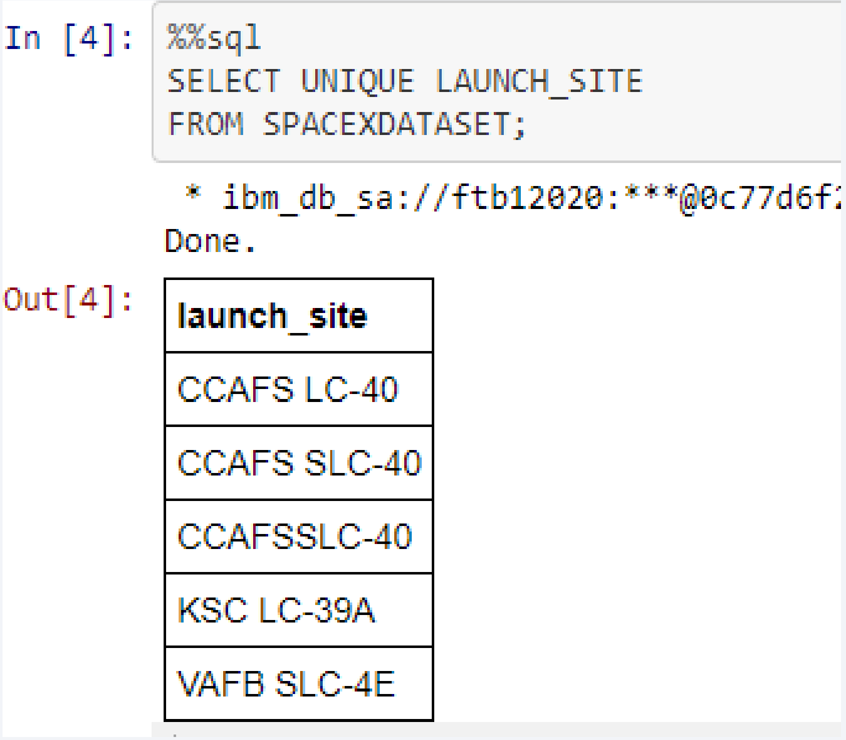
- Payload mass seems to correlate with orbit
- LEO and SSO seem to have relatively low payload mass
- The other most successful orbit VLEO only has payload mass values in the higher end of the range

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



In [4]:

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

```
 * ibm_db_sa://ftb12020:***@0c77d6f2
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

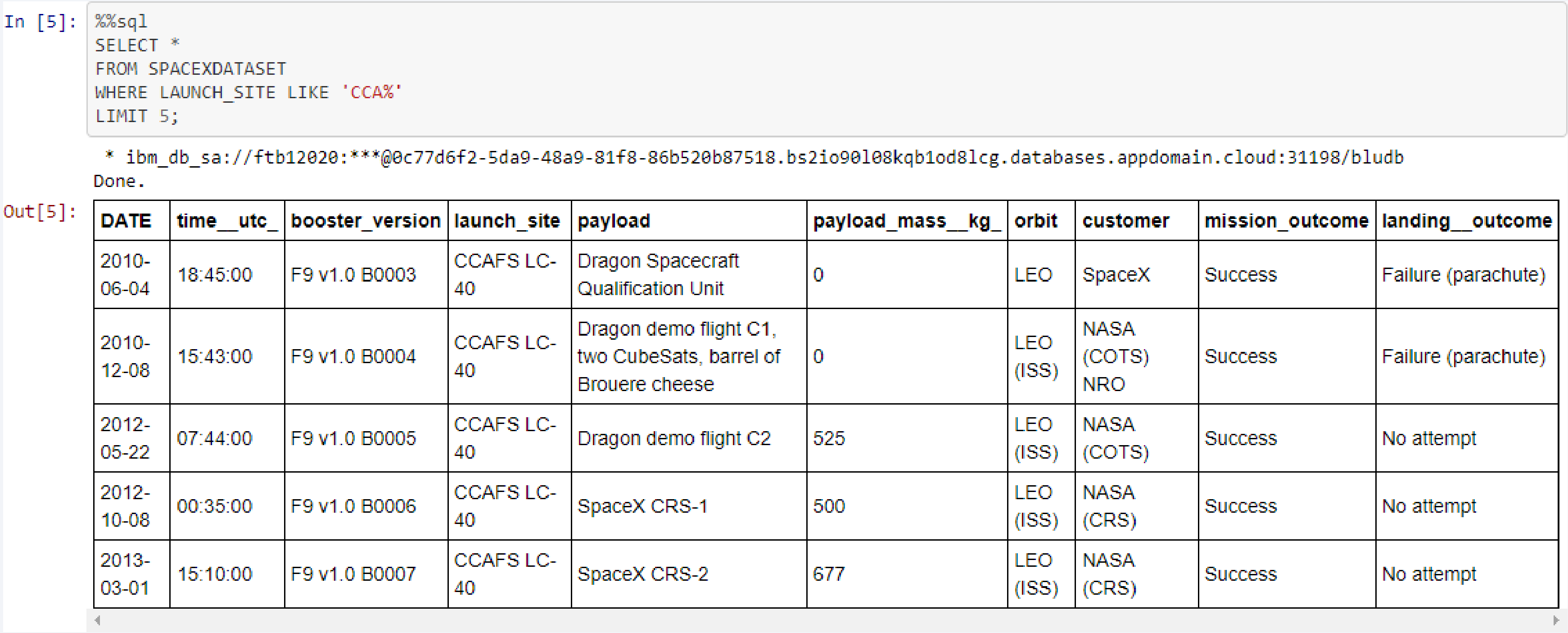
3 unique launch_site

values: CCAFS SLC-40,

KSC LC-39A, VAFB SLC-

4E

Launch Site Names Begin with 'CCA'



In [5]:

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

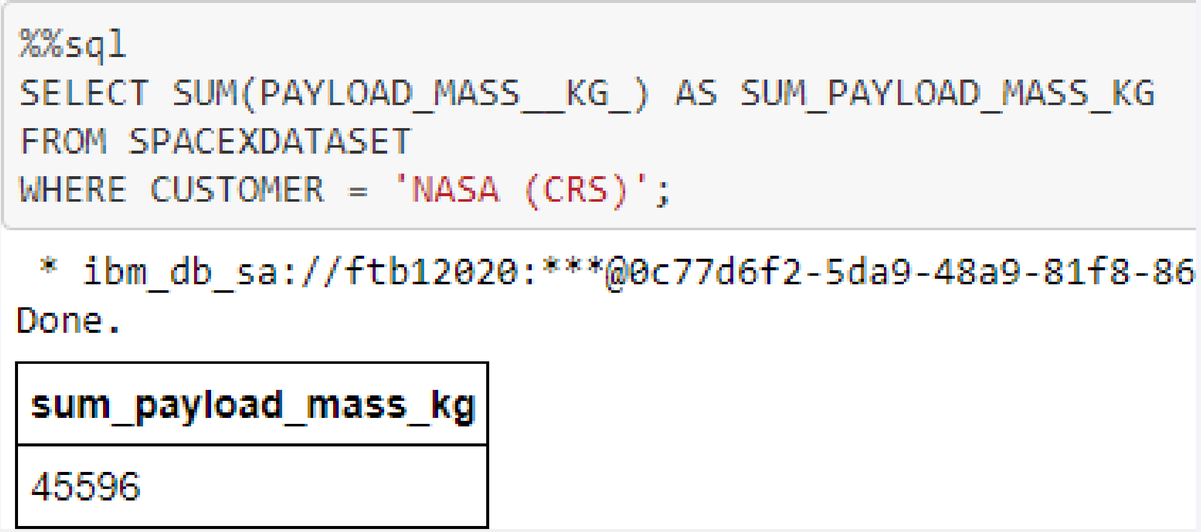
```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

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



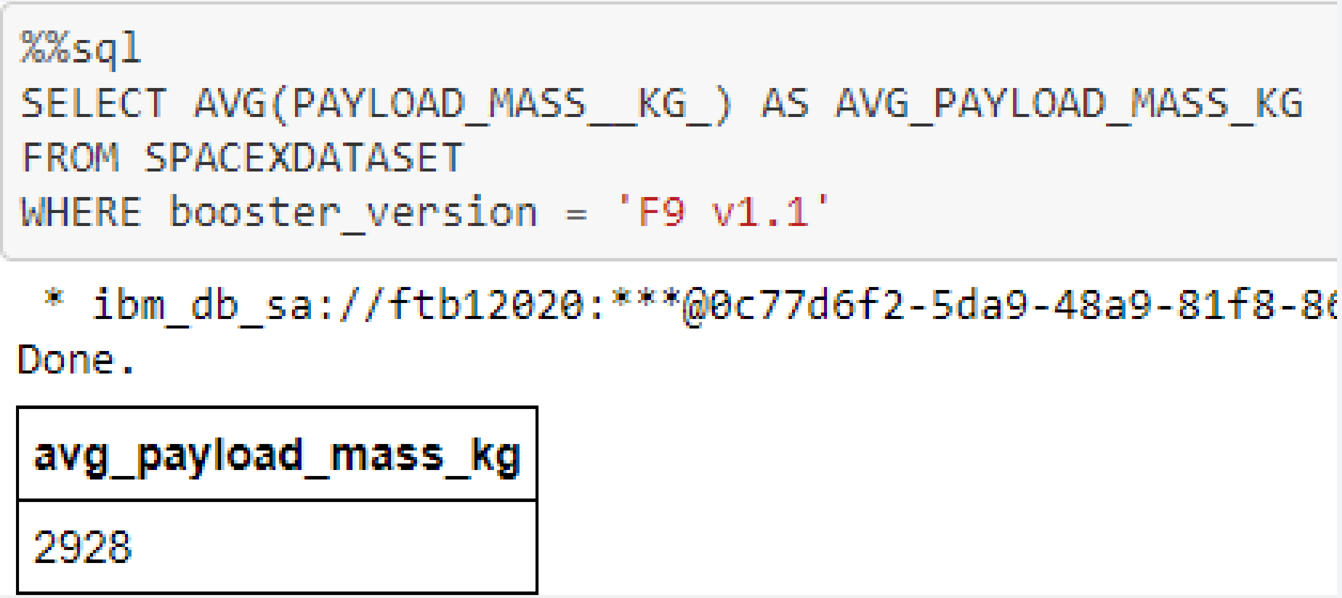
```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS SUM_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE CUSTOMER = 'NASA (CRS)';
```

```
 * 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



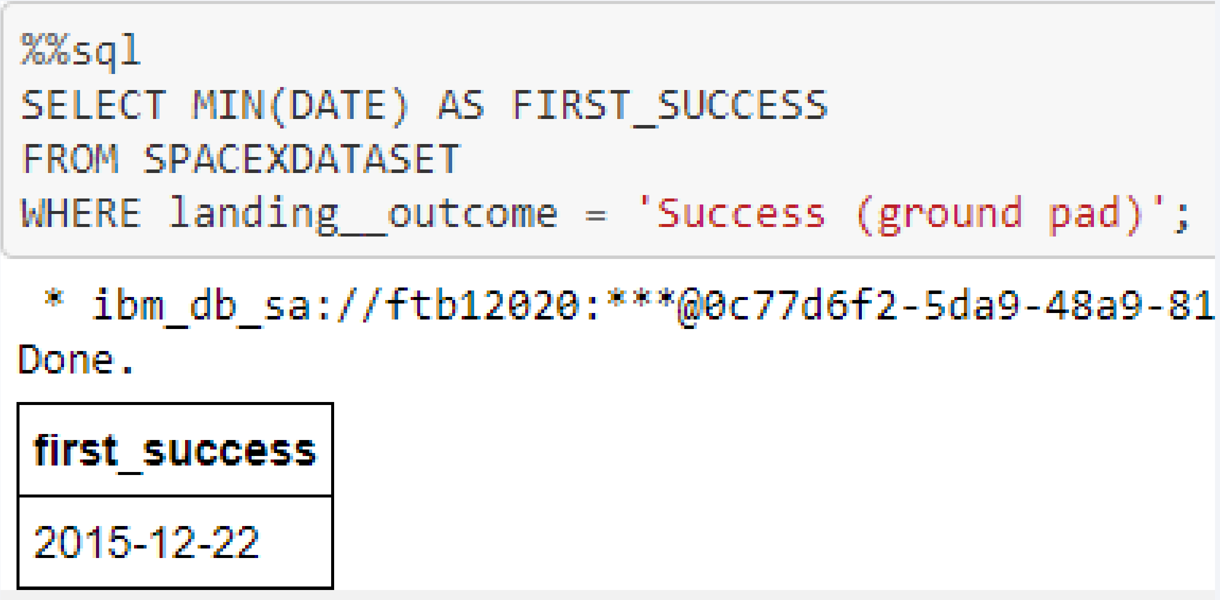
```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE booster_version = 'F9 v1.1'
```

```
 * 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 Landing Date



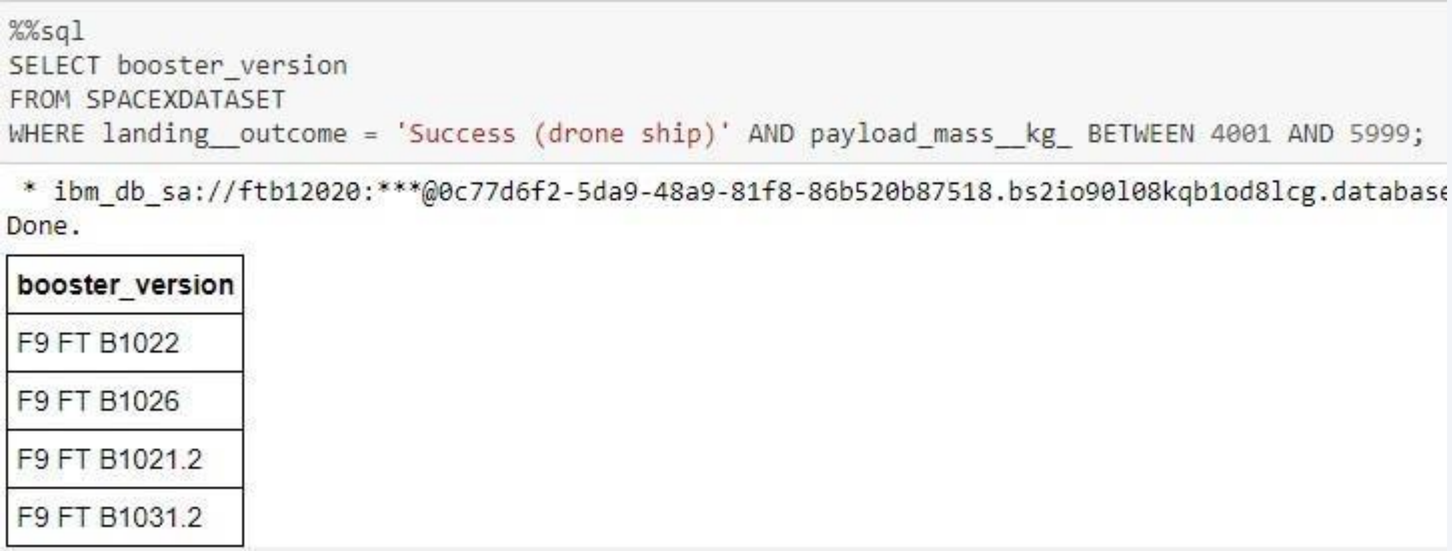
```
%%sql
SELECT MIN(DATE) AS FIRST_SUCCESS
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (ground pad)';
```

```
 * 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



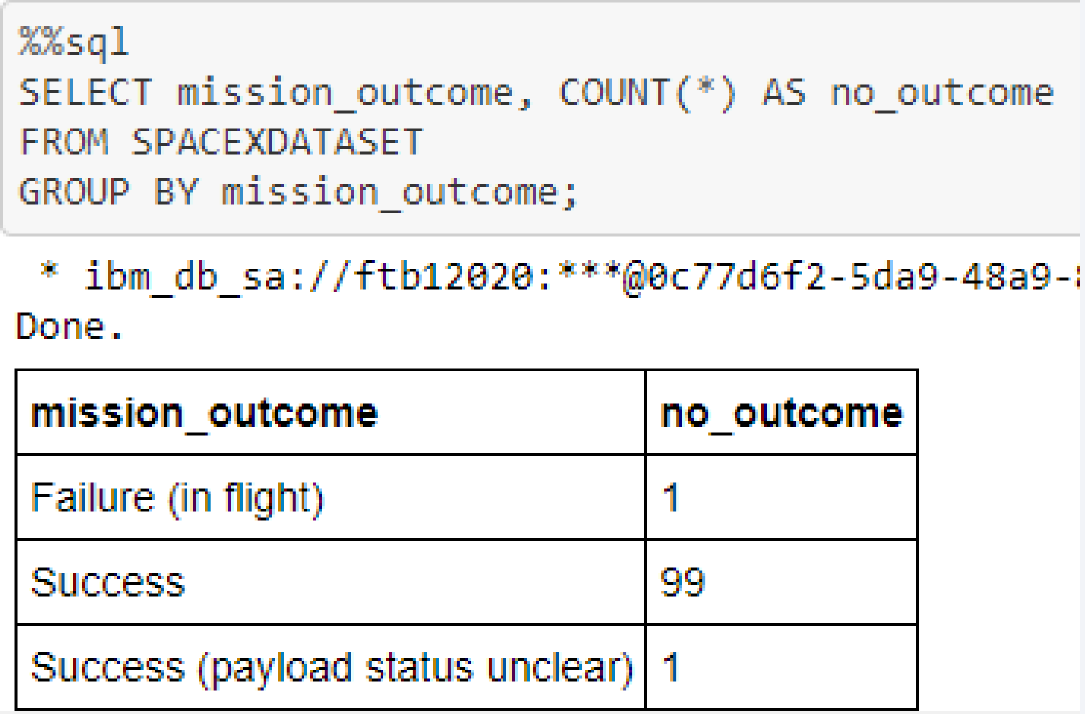
```
%%sql
SELECT booster_version
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4001 AND 5999;
```

```
 * 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 non inclusively.

Total Number of Successful and Failure Mission Outcomes



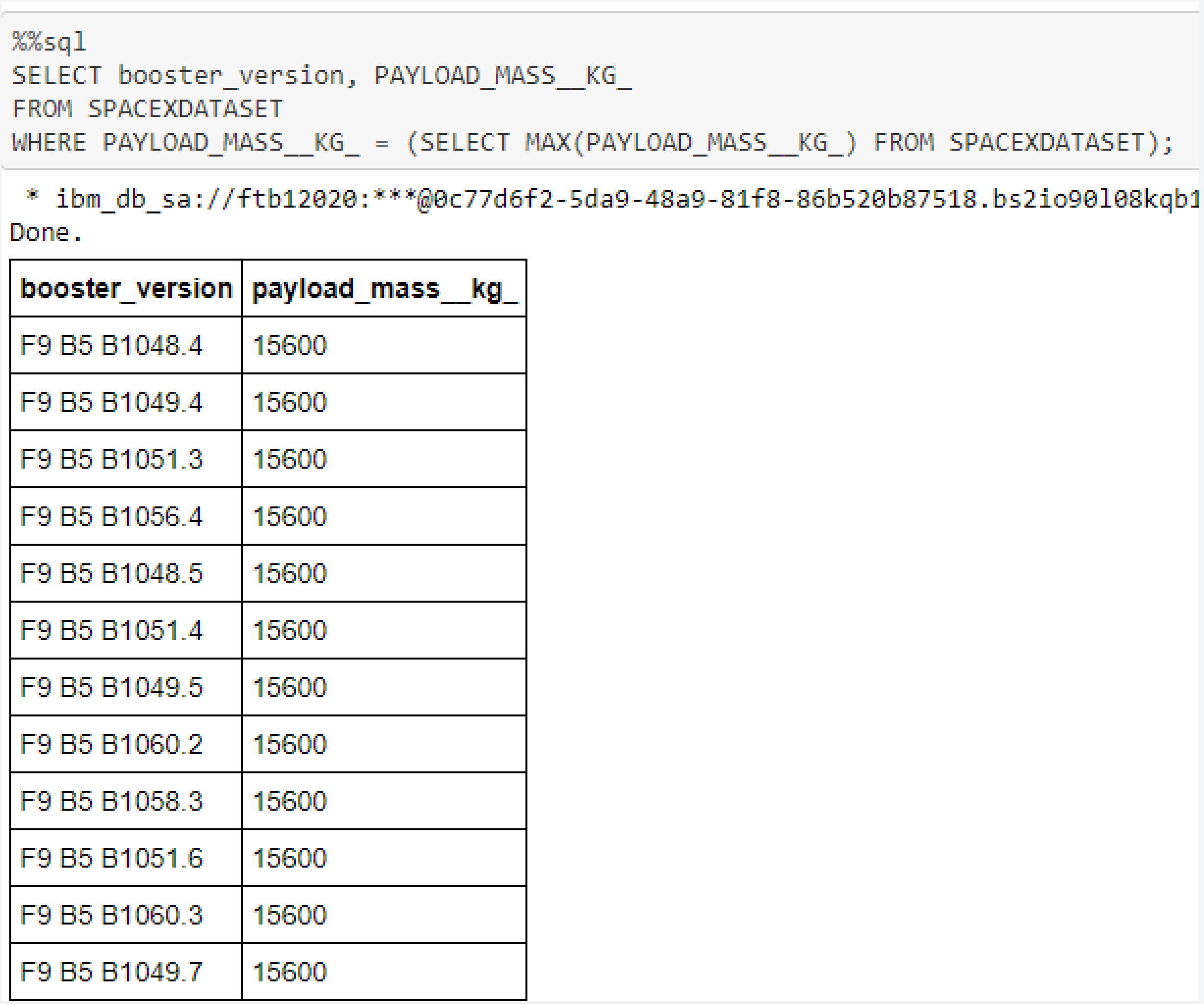
```
%%sql
SELECT mission_outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
GROUP BY mission_outcome;
```

```
 * 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 Carried Maximum Payload



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

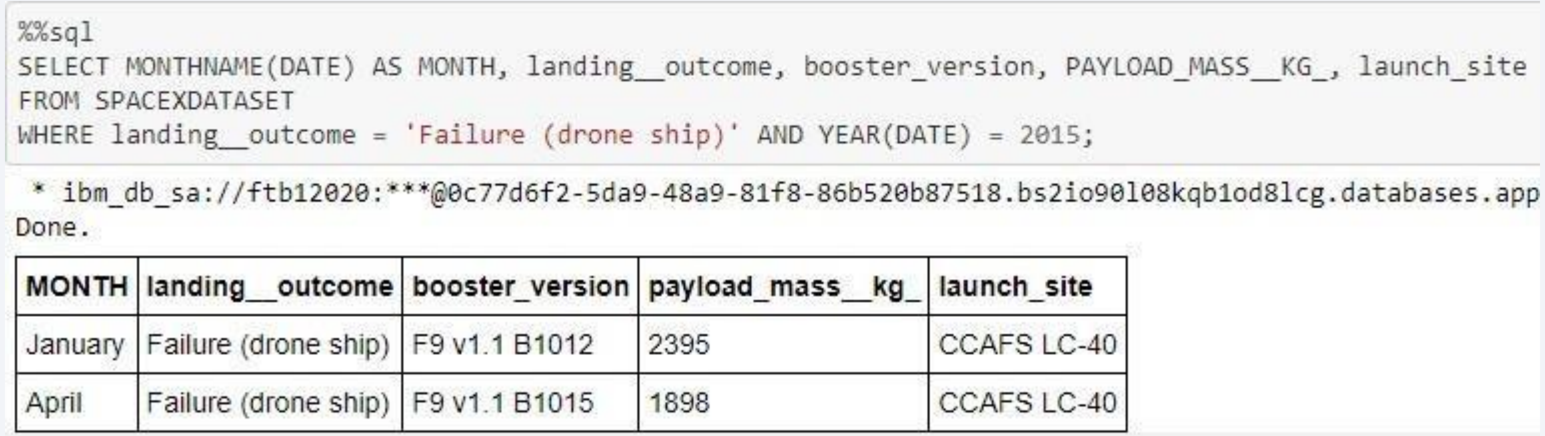
```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1
Done.
```

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

- The query identifies F9 B5 B10xx.x booster versions that transported the maximum 15600 kg payload mass, suggesting a strong correlation.

- These similar booster versions strongly imply that the choice of booster version relates to the payload mass requirement.

2015 Launch Records



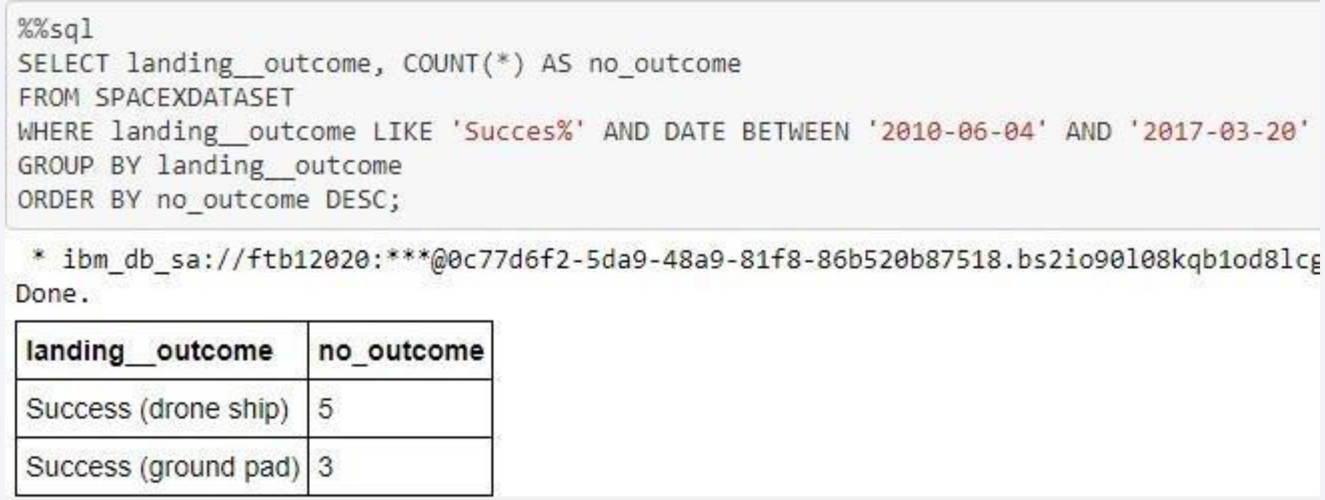
```
%%sql
SELECT MONTHNAME(DATE) AS MONTH, landing__outcome, booster_version, PAYLOAD_MASS__KG_, launch_site
FROM SPACEXDATASET
WHERE landing__outcome = 'Failure (drone ship)' AND YEAR(DATE) = 2015;
```

```
 * 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

- Query retrieves details of 2015 launches: Month, Landing Outcome, Booster Version, Payload Mass (kg), and Launch site. Specifically focuses on stage 1 failures to land on drone ship.
- Identifies two instances of such failures in 2015.

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;
```

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

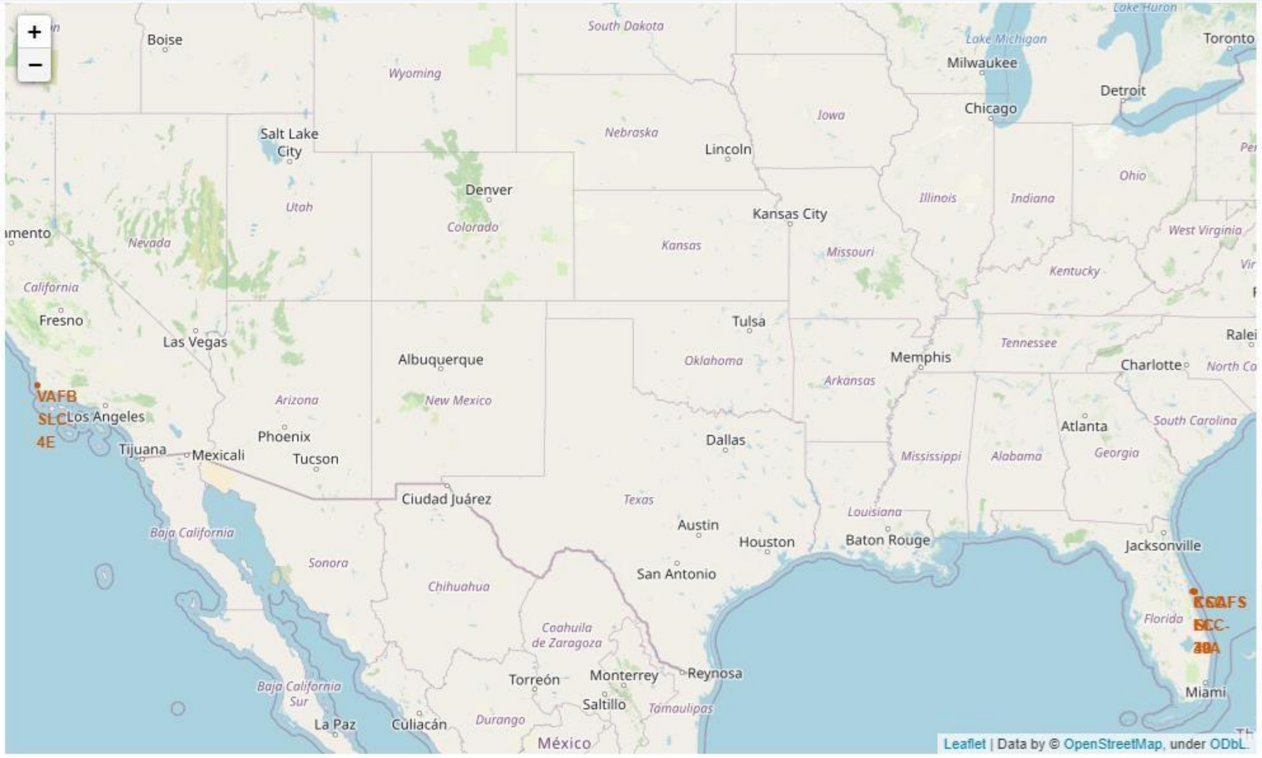
landing__outcome	no_outcome
Success (drone ship)	5
Success (ground pad)	3

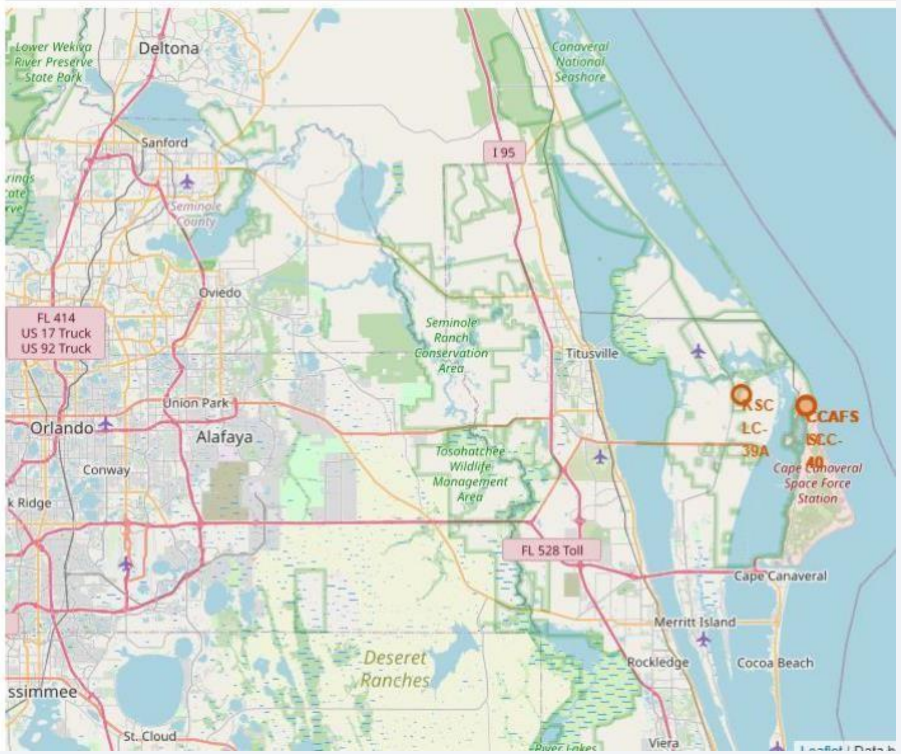
- The query provides a list of 8 successful landings between June 4, 2010, and March 20, 2017, including both drone ship and ground pad outcomes.
- Launch sites are depicted on maps: the left map displays all sites on the US map, while the right map focuses on the two closely situated launch sites in Florida, all of which are near the ocean.

Section 3

Launch Sites Proximities Analysis

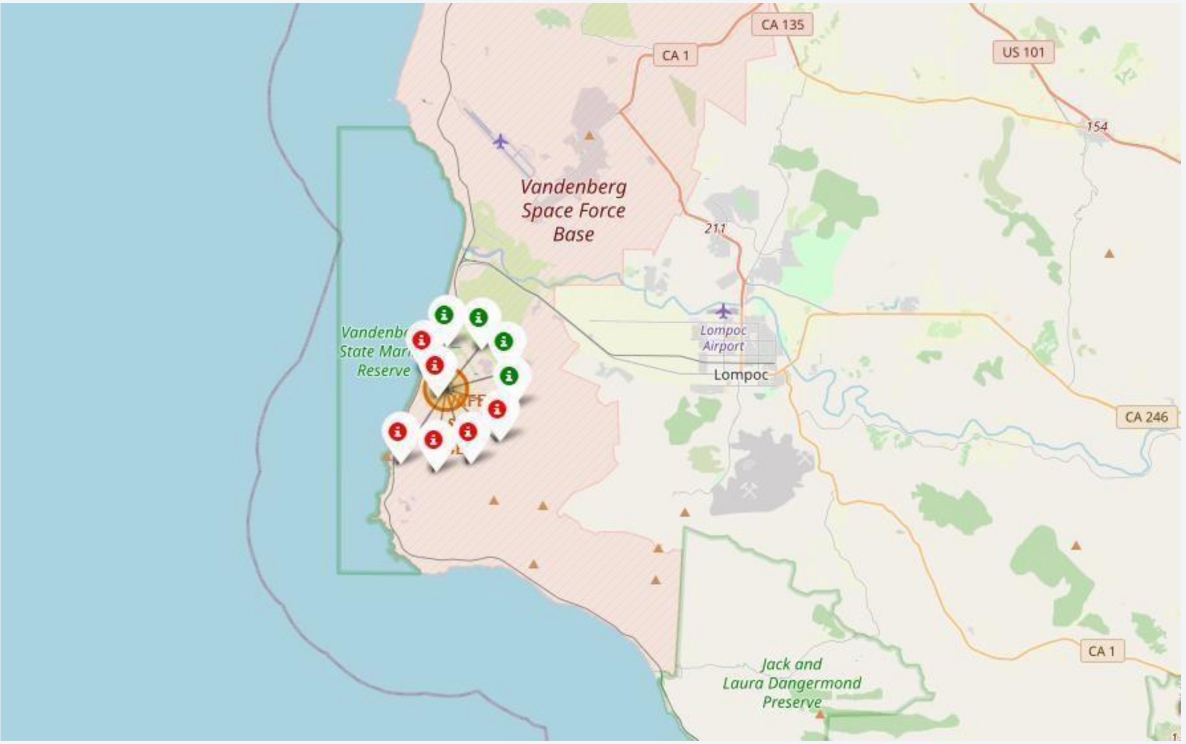
Launch Site Locations





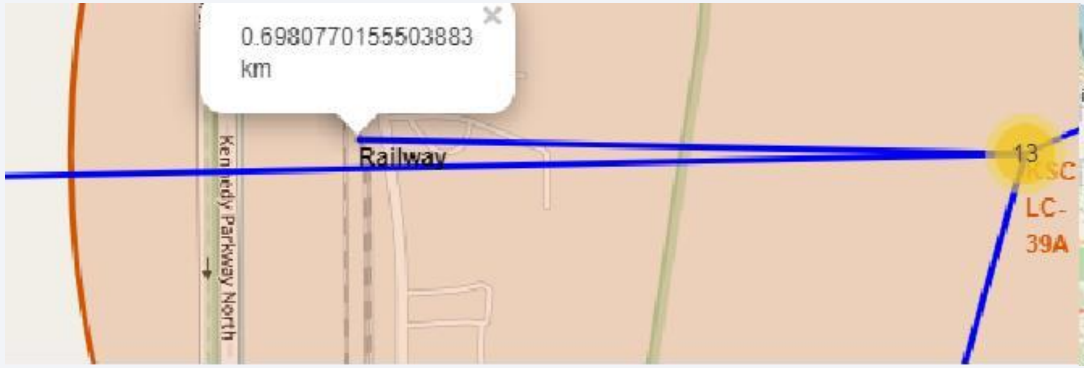
- The left map illustrates global launch sites in relation to the US map, while the right map specifically focuses on the close proximity of the two Florida launch sites.
- All launch sites are strategically located near ocean areas.

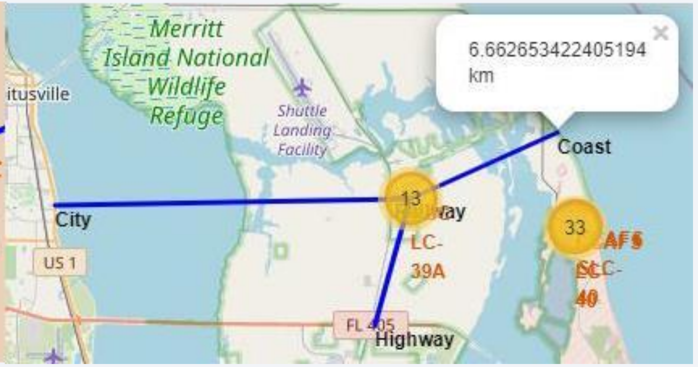
Color-Coded Launch Markers

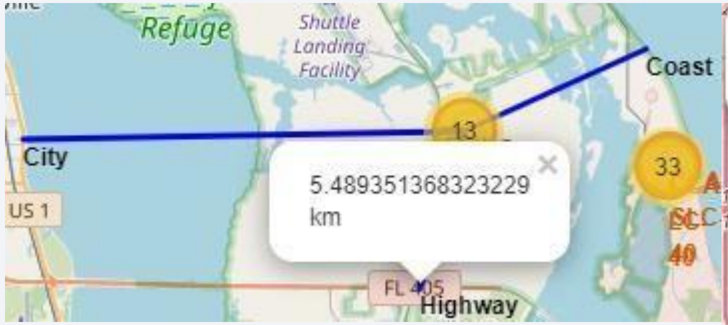

CA 135
CA 1
US 101
154
Vandenberg Space Force Base
211
Lompoc Airport
Lompoc
Vandenberg State Marine Reserve
CA 246
CA 1
Jack and Laura Dangermond Preserve

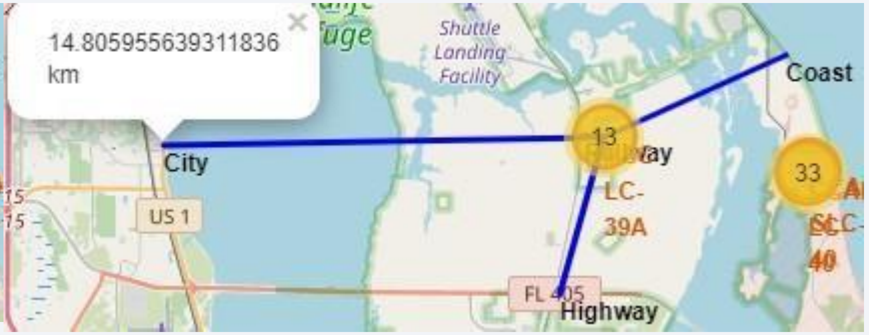
- Clickable clusters on Folium map reveal successful (green) and failed (red) landings.
- VAFB SLC-4E illustrates 4 successful and 6 failed landings in this instance.

Key Location Proximities

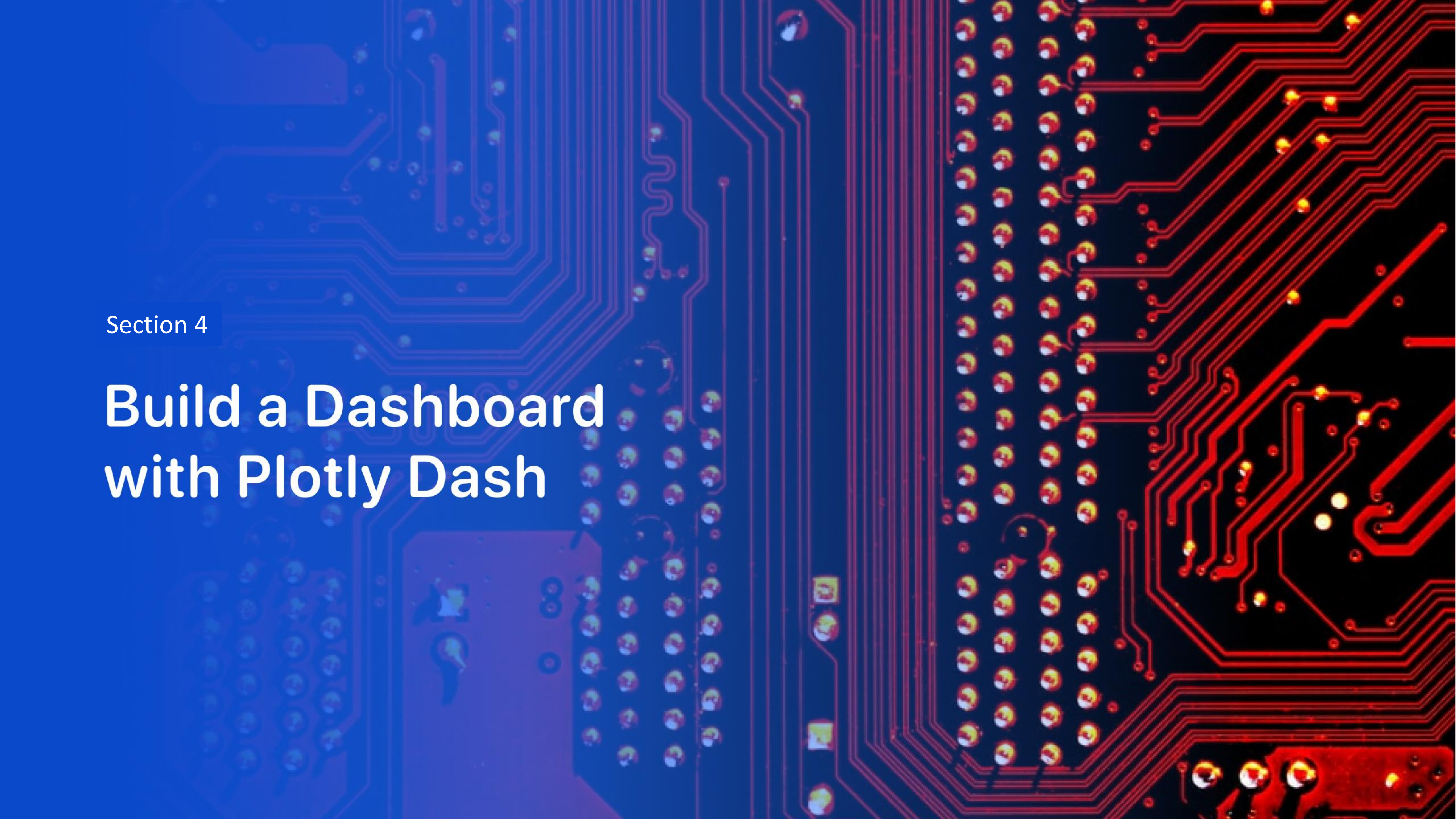








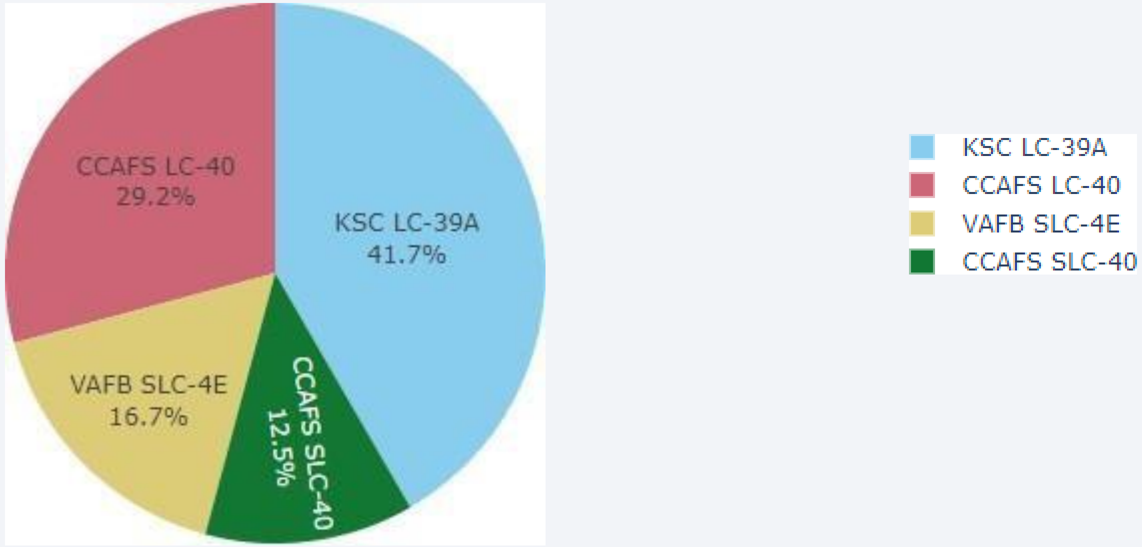
- Launch sites like KSC LC-39A are strategically located near railways and highways, facilitating efficient large-scale supply transportation as well as human transit.
- These launch sites are positioned near coasts, ensuring safe sea landings for launch failures and preventing rocket debris from endangering densely populated urban areas.



Section 4

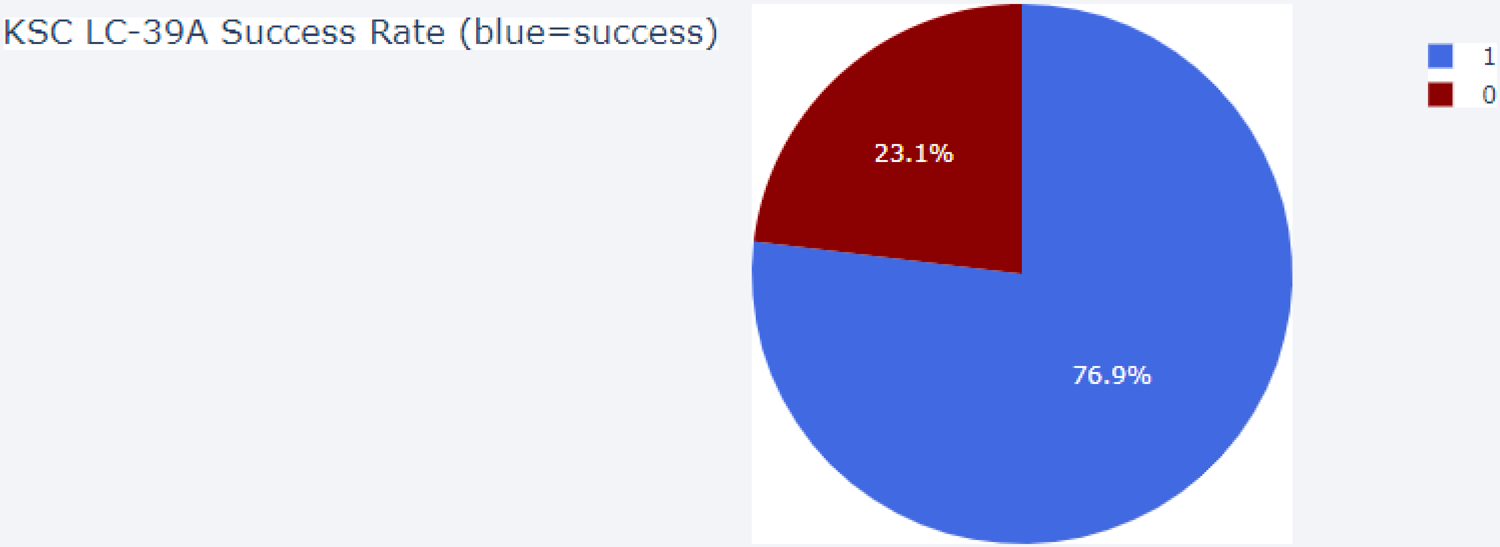
Build a Dashboard with Plotly Dash

Successful Launches Across Launch Sites



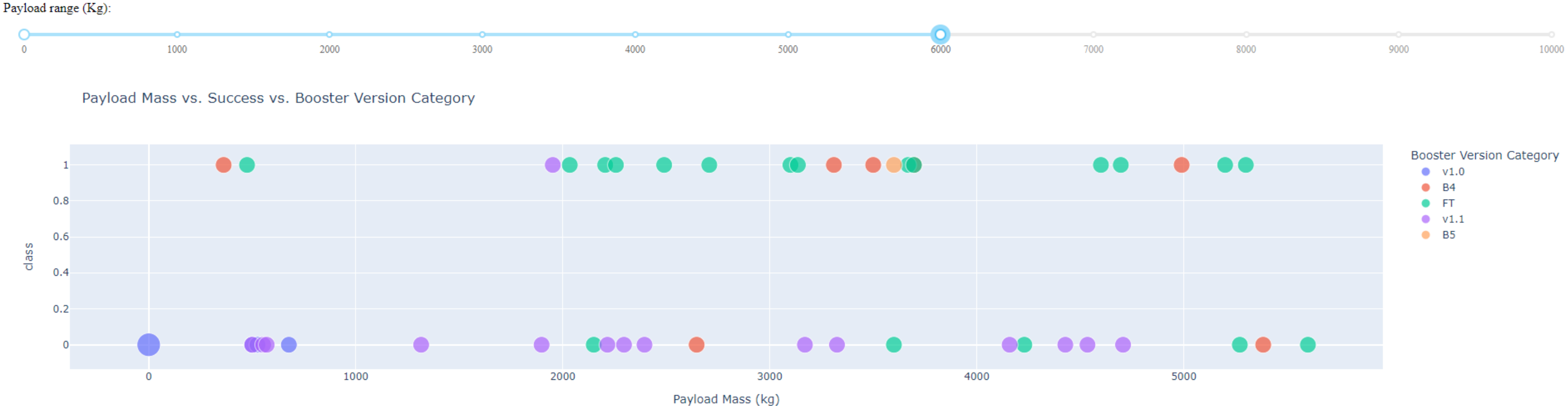
- Successful landings distributed across launch sites: CCAFS SLC-40 (formerly CCAFS LC-40) and KSC have equal successes due to name change, mainly prior to it.
- VAFB shows the lowest success share, possibly due to its smaller sample and higher west coast launch challenges

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



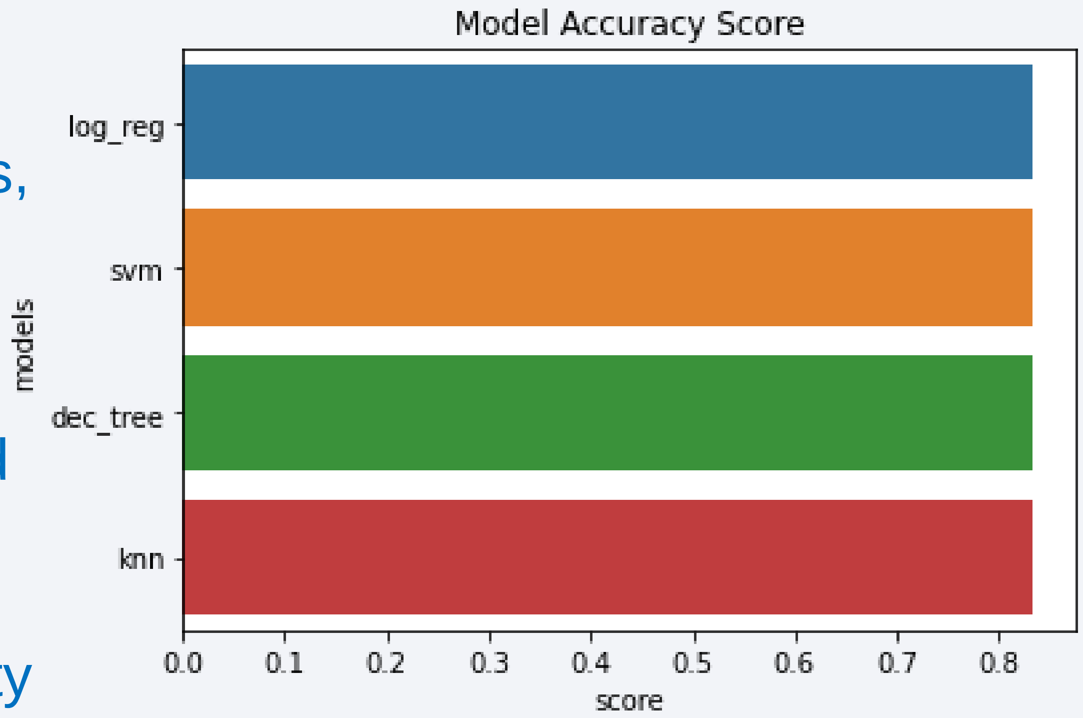
- Plotly dashboard's Payload range selector limited to 0-10000, not reflecting the max Payload of 15600.
- Class 1 indicates successful, 0 for failed landing. Scatter plot uses color for booster version and point size for launches.
- Notably, within the range 0-6000, two failed landings occurred with payloads of zero kg.

Section 5

Predictive Analysis (Classification)

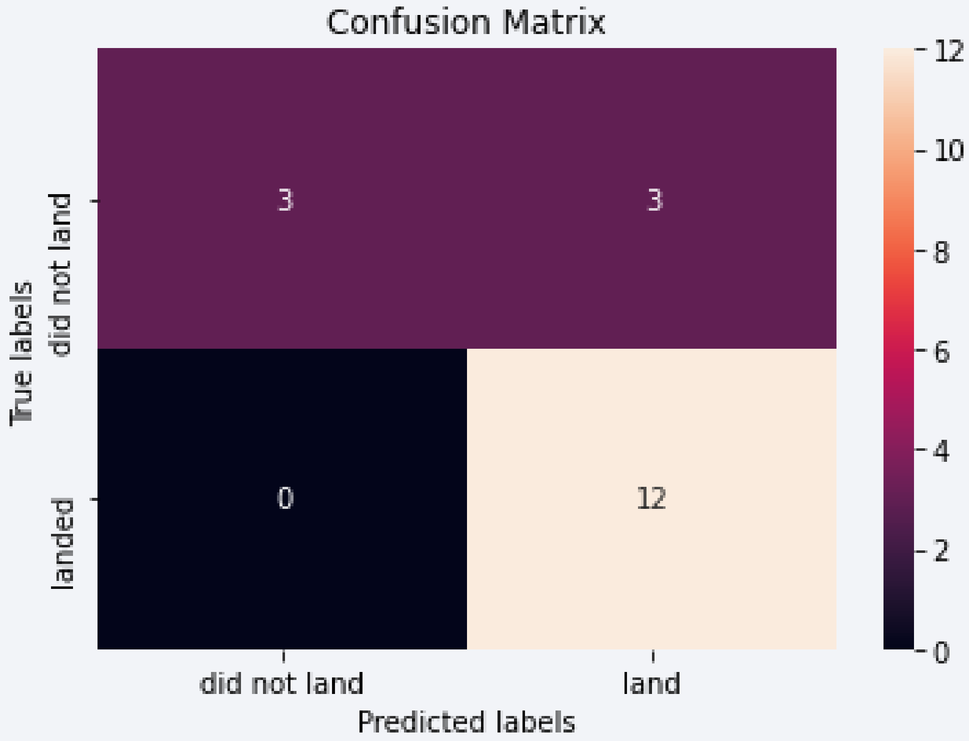
Classification Accuracy

- All models demonstrated similar performance with an 83.33% accuracy on the test set. Worth mentioning, the test set consisted of only 18 samples, which is relatively small.
- The limited test size of 18 samples could lead to significant fluctuations in accuracy scores. This was evident in the variability of accuracy results observed during repeated runs of the Decision Tree Classifier model.
- A larger dataset is required to enhance the reliability of model selection and performance evaluation.



Confusion Matrix

- Since consistent results were observed among all models on the test set, the confusion matrix remained consistent as well. The model forecasts aligned with true outcomes in 12 instances of successful landings.
- In cases of unsuccessful landings, the models accurately predicted 3 occurrences, but also made 3 incorrect successful landing predictions (false positives). This reveals a tendency for our models to overestimate successful landings.



Conclusions

- Objective: Develop a machine learning model for Space Y to compete with SpaceX.
- Aim: Predict successful Stage 1 landing to save around $100 million USD.
- Data Sources: Utilized SpaceX API and web scraping of SpaceX Wikipedia.
- Data Handling: Labels generated, data stored in DB2 SQL database.
- Visualization: Created a dashboard for clear data representation.
- Achievements: Built ML model with 83% accuracy.
- Practicality: Enables SpaceY to predict successful Stage 1 landing, aiding launch decisions.
- Future Steps: Collect more data to enhance ML model for improved accuracy.

Appendix

- **GitHub repository url:**

https://github.com/SindhePandurangBITS/IBM_DS_Professional.git

- **Instructors:**

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 All Instructors:**

https://www.coursera.org/professional-certificates/ibm-data-science?#instructors

Thank you!