Data Science and Machine Learning Capstone Project

Kavyansh Sharma

https://github.com/kavyansh091/capstone_project

Outline

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

Executive Summary

- In this capstone project, we embarked on a data-driven journey to predict the successful landing of Falcon 9 first stages, leveraging data collected from public SpaceX APIs and the SpaceX Wikipedia page.
- The primary objective was to determine if the first stage of the Falcon 9 rocket would land successfully, with the ultimate goal of estimating the cost of a launch—a crucial factor for competitors bidding against SpaceX.
- We developed four distinct machine learning models, including Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors, each revealing remarkably similar outcomes.

Introduction

- Companies such as Virgin Galactic, Rocket Lab, Blue Origin, and SpaceX are growing in the commercial space business.
- SpaceX has accomplished extraordinary feats including as ISS missions, Starlink internet connectivity, and cost-effective Falcon 9 launches.
- Our objective (Space Y) is to use machine learning to anticipate Falcon 9 first stage reusability and launch costs.

Methodology

1. Data Collection and Preprocessing

   - Data Sources: We collected data on Falcon 9 first-stage landings from various sources, including a RESTful API and web scraping.
   - Data Cleaning: We meticulously cleaned and prepared the data to ensure its accuracy and reliability for analysis.

2. Exploratory Data Analysis (EDA)

   - Data Visualization: We employed data visualization techniques to gain insights and identify patterns within the dataset.
   - Data Insights: Through EDA, we gathered valuable information that guided our modeling process.

3. Interactive Visual Analytics and Dashboard

    - Proximity Analysis: We built an interactive map using Folium to analyze the launch site proximity.
    - Launch Records Analysis: A dynamic dashboard was created using Plotly Dash to interactively explore launch records.

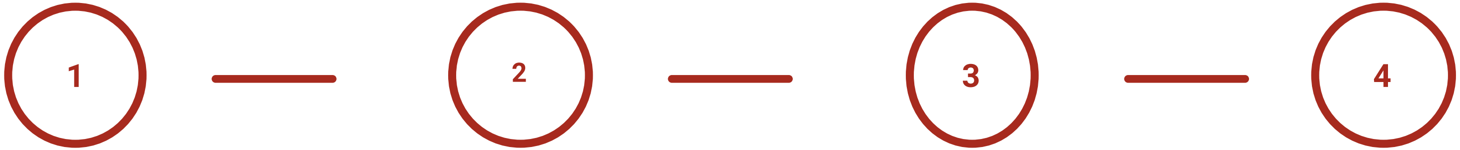
4. Predictive Analysis (Classification)

    - Model Selection: We considered various classification models, including Support Vector Machine (SVM), Classification Trees, and Logistic Regression.
    - Data Splitting: The dataset was split into training and testing data to assess model performance.
    - Hyperparameter Tuning: We conducted a hyperparameter grid search to optimize model parameters.
    - Model Evaluation: The models were evaluated on test data to select the best-performing method.

Data Collection Overview

- Data collection is a crucial step in our project to predict the successful landing of Falcon 9 first stages.
- We gathered data from reliable sources, including SpaceX's API and Wikipedia pages, to ensure data accuracy and completeness.
- 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



Access SpaceX API

accessed SpaceX's API to retrieve real-time and historical data

Data Retrieval

we obtained information such as launch dates, success status, and launch site coordinates.

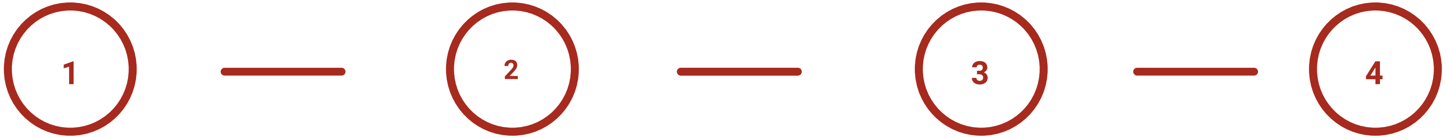
Data Integration

retrieved data was integrated into our dataset for analysis.

Filtering Data

Filter data to only include Falcon 9 launches

Data Collection - Wikipedia Data Scraping



Web Scraping

performed web scraping on relevant Wikipedia pages to extract additional data.

Information Extraction

Extracted information included mission details, payload types, and launch site descriptions.

Data Cleaning

scraped data underwent thorough cleaning to eliminate inconsistencies.

Creating Dataframe

Creating dictionary to dataframe

Data Wrangling

- Within our SpaceX project, meticulous data cleaning was conducted. This involved addressing missing values, outliers, and formatting inconsistencies specific to Falcon 9 launch and landing data.
- Step 1: Examine the 'Mission Outcome' and 'Landing Location' columns in your dataset.
- Step 2: Create a new column 'class' to represent the training labels.
- Step 3: Use the following value mapping for 'Mission Outcome' to determine the 'class' values:
- If 'Mission Outcome' is 'True ASDS', 'True RTLS', or 'True Ocean', set 'class' to 1.
- If 'Mission Outcome' is 'None None', 'False ASDS', 'None ASDS', 'False Ocean', or 'False RTLS', set 'class' to 0.
- Step 4: The 'class' column will now contain 1 for successful landings and 0 for failed landings based on the 'Mission Outcome' values

EDA with Data Visualization

- In our SpaceX project, data visualization played a pivotal role in exploring Falcon 9 launch and landing data. Visualizations were created to intuitively understand success patterns and mission characteristics.
- **Variables Analyzed:**
  - Flight Number: We examined the impact of flight numbers on various aspects of Falcon 9 missions.
  - Payload Mass: The relationship between payload mass and mission outcomes was explored.
  - Launch Site: We assessed the launch site's influence on mission success.
  - Orbit: The choice of orbit and its correlation with success rates was examined.
  - Class: We considered the newly created 'class' variable (0 for failure, 1 for success).
  - Year: We analyzed yearly trends in mission success.

- **Plots Utilized:**
  - Flight Number vs. Payload Mass: Scatter plots were used to visualize how payload mass varied with flight number, providing insights into mission payload capacity.
  - Flight Number vs. Launch Site: A bar plot was employed to showcase the distribution of launches across different sites over time.
  - Payload Mass vs. Launch Site: We used scatter plots to explore the relationship between payload mass and launch site choices.
  - Orbit vs. Success Rate: Bar charts illustrated the success rates for different orbits, aiding in the understanding of mission outcomes.
  - Flight Number vs. Orbit: Line charts allowed us to observe trends in orbit choices across Falcon 9 missions.
  - Payload vs. Orbit: Scatter plots visually depicted the relationship between payload mass and chosen orbits.
  - Success Yearly Trend: Line charts revealed trends in mission success rates over the years, helping us identify long-term patterns.

Data Integration and SQL Queries

- We loaded our dataset into an IBM DB2 Database and utilized Python's SQL integration to query for critical insights, including:

- Launch Site Names: To understand mission origins.
- Mission Outcomes: For success and failure patterns.
- Payload Sizes: To assess payload capacity.
- Booster Versions: To track technology advancements.
- Landing Outcomes: To study Falcon 9 first-stage landing success.

Building an Interactive Map with Folium

- In our SpaceX project, we harnessed Folium to create interactive maps that visualized the proximity of launch sites. This was integral in assessing the impact of launch site location on mission success.

- Folium allowed us to display real-time launch site information, making it easy for users to explore the locations and factors that influenced Falcon 9 first-stage landings.

Building a Dashboard with Plotly Dash

- Plotly Dash was a crucial component of our SpaceX project's dashboard, providing users with an intuitive interface to access and analyze Falcon 9 launch records.

- The dashboard featured interactive visualizations created with Plotly, allowing users to explore mission data interactively, facilitating decision-making for alternate rocket launches.

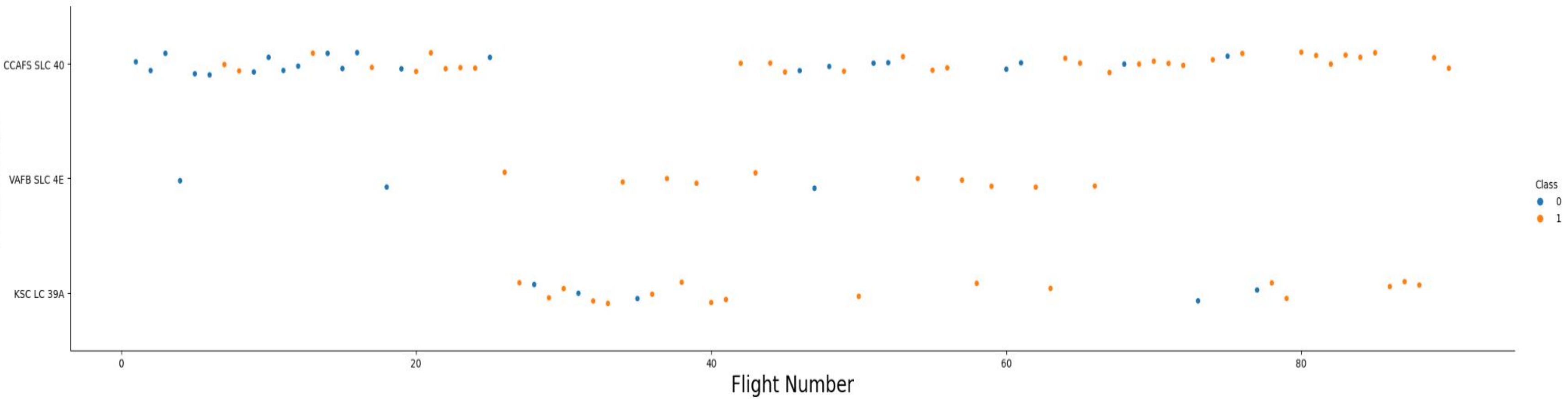
- The flexibility of Plotly Dash enabled us to customize the dashboard's features to cater to the specific needs of companies bidding against SpaceX for rocket launches, offering them actionable insights

EDA with Visualization

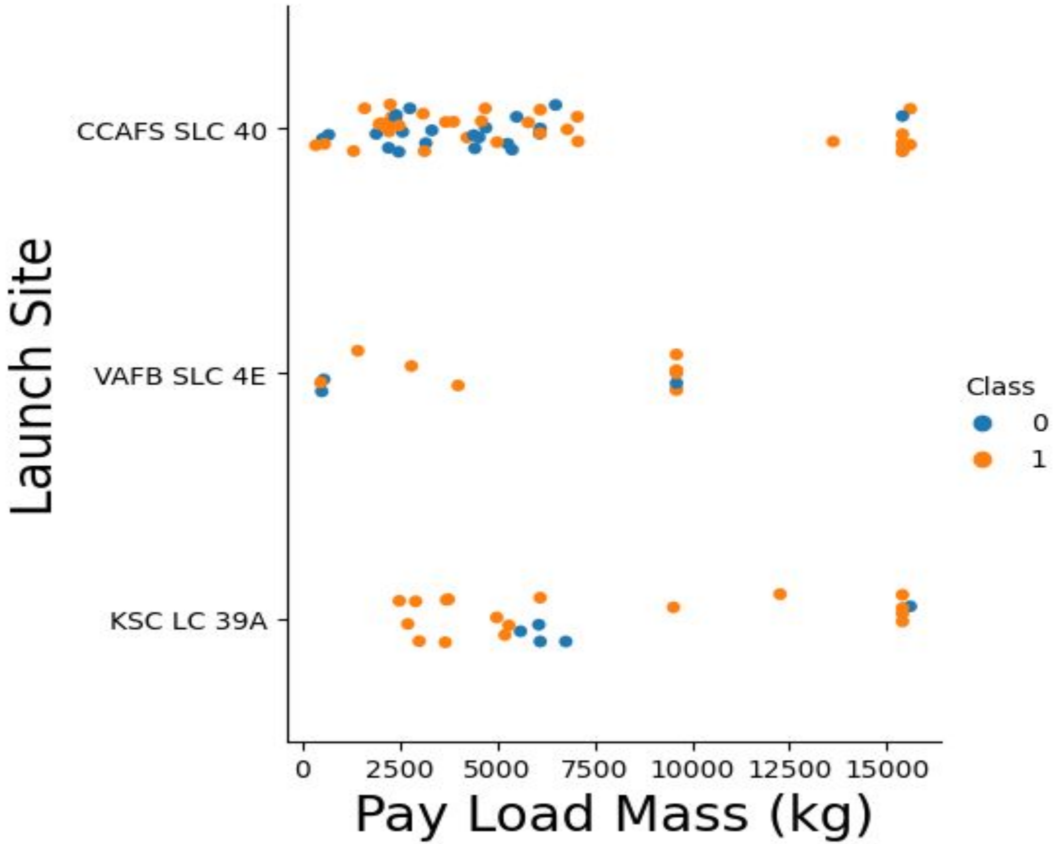
Exploratory data analysis with visualization

Flight Number vs. Launch Site

C
L
s
v



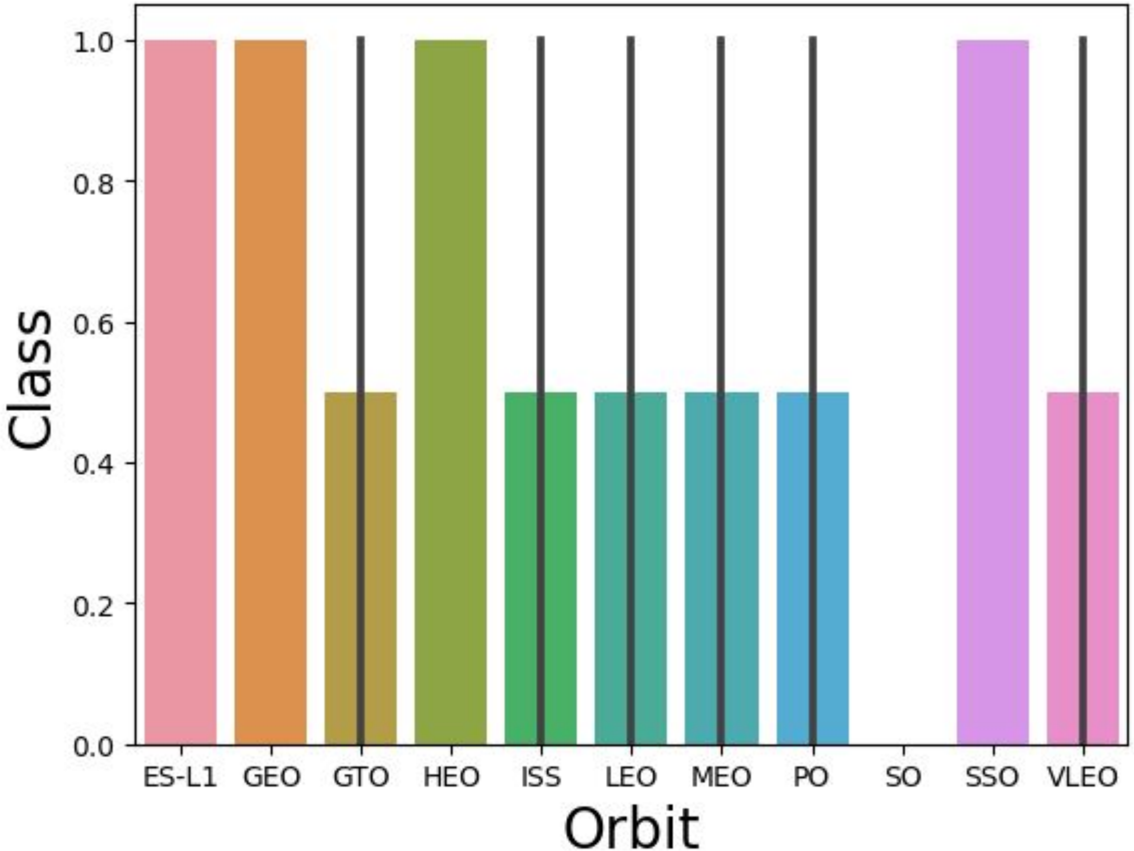
Payload vs. Launch Site



Payload mass appears to fall mostly between 0-6000 kg.

Different launch sites also seem to use different payload mass.

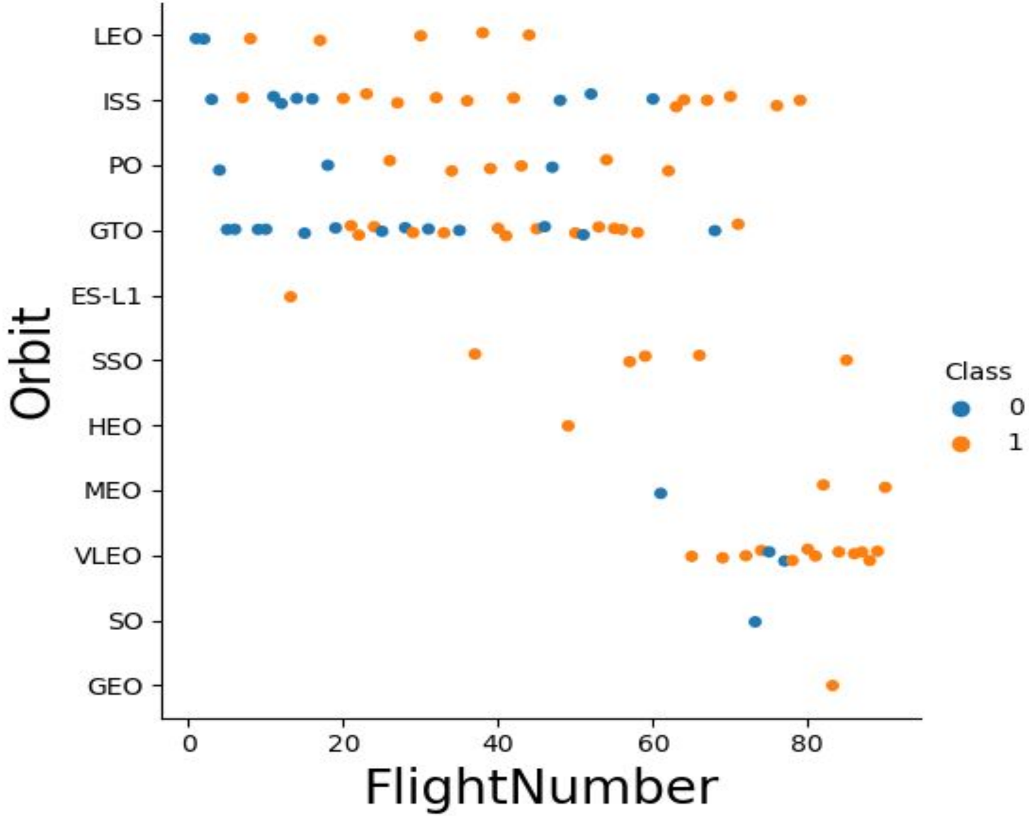
Relationship between success rate of each orbit type



Success rate scale

- With 0 as 0%
- 0.6 as 60%
- 1 as 100%

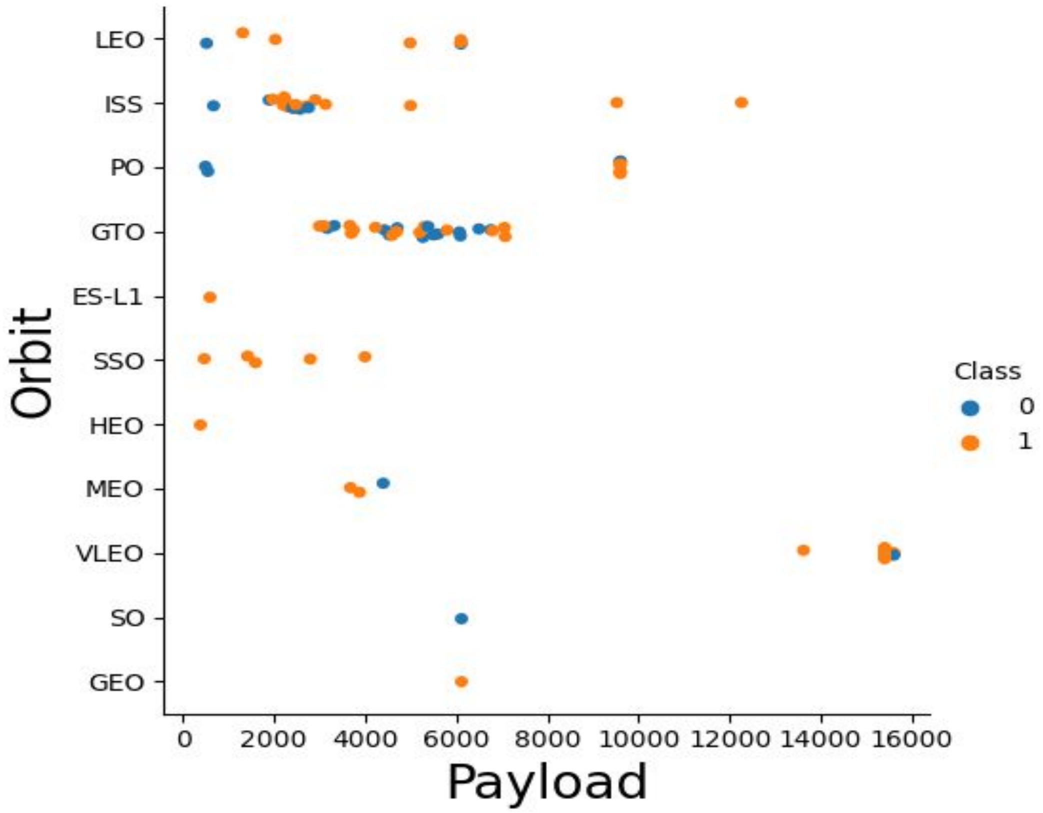
Flight Number vs. Orbit type



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

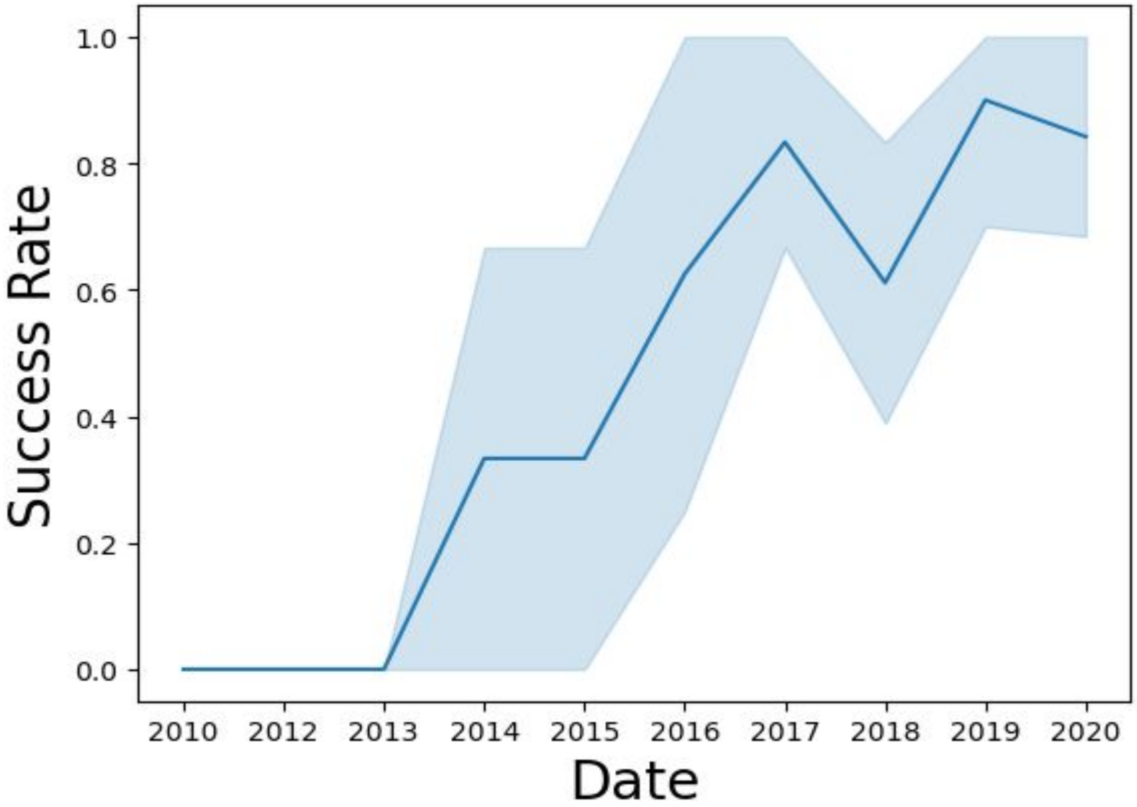
Payload vs. Orbittype



Payload mass seems to correlate with orbit

LEO and SSO seem to have relatively low payload mass

Launch Success Yearly Trend

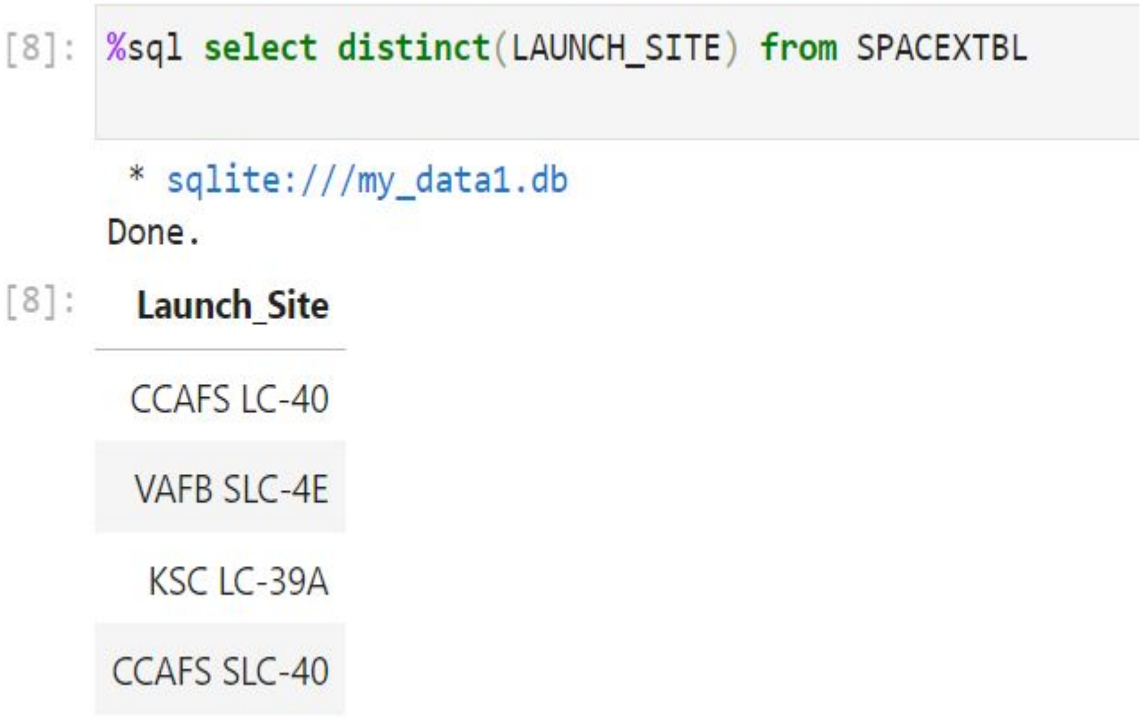


95% confidence interval (light blue shading)

EDA with SQL

Exploratory data analysis with sql

List all Launch site names



```
[8]: %sql select distinct(LAUNCH_SITE) from SPACEXTBL

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

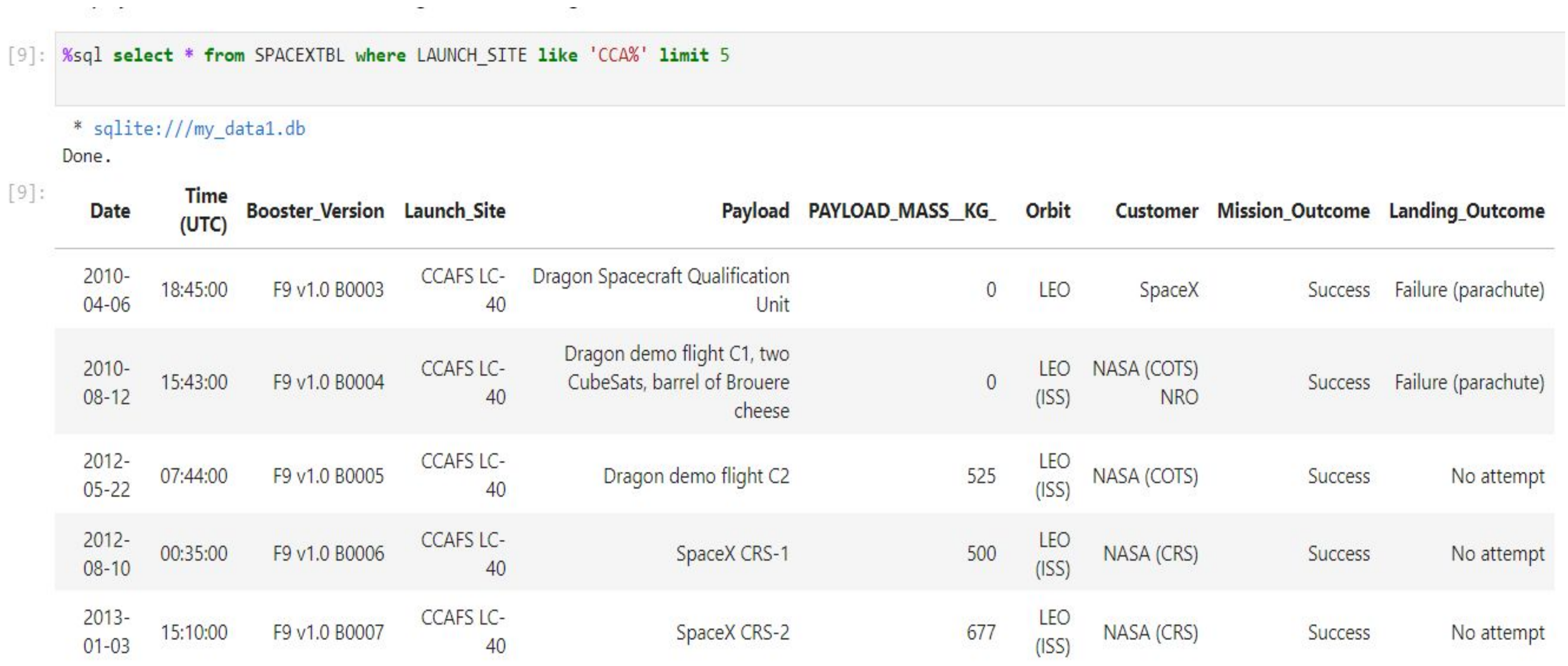
[8]:

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

Unique launch site names from the database:

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

List all Launch site beginning with “CCA”



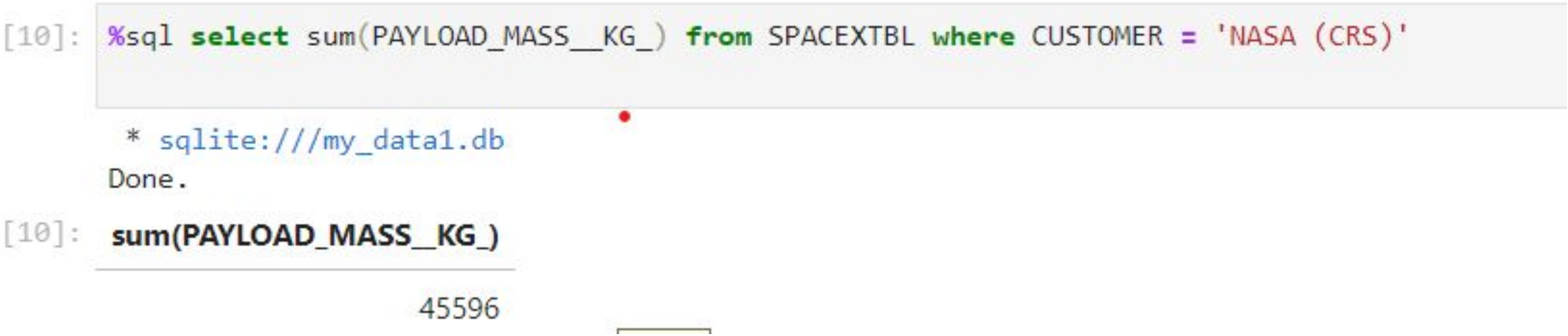
```
[9]: %sql select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' limit 5
```

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

[9]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass from NASA

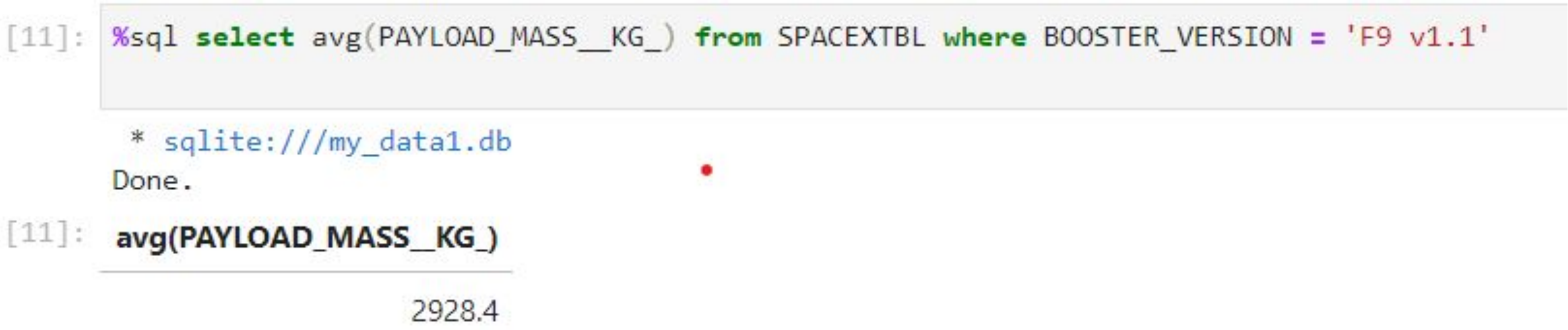


```
[10]: %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where CUSTOMER = 'NASA (CRS)'

       * sqlite:///my_data1.db
      Done.

[10]: 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).

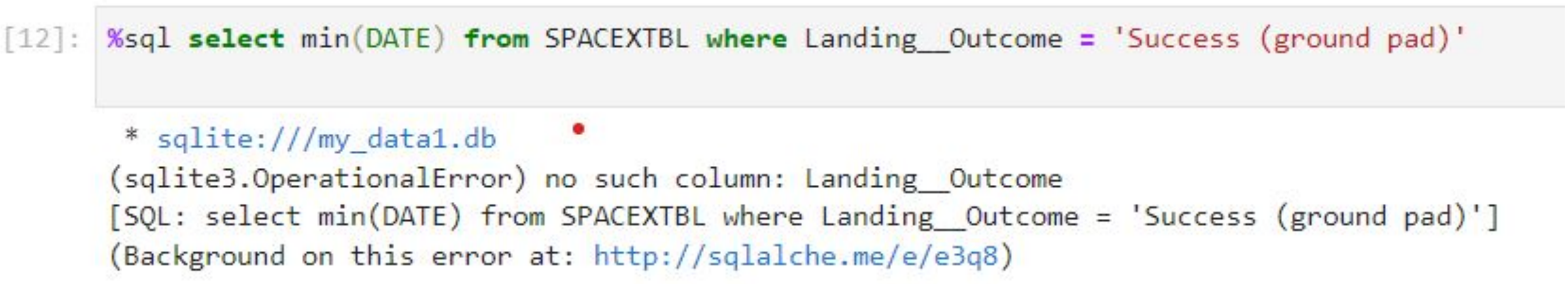
Average Payload Mass from NASA



```
[11]: %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where BOOSTER_VERSION = 'F9 v1.1'

 * sqlite:///my_data1.db
Done.

[11]: avg(PAYLOAD_MASS__KG_)
                    2928.4
```

This query calculates the average payload mass or launches which used booster version F9 v1.1

Successful Ground pad Landing date



```
[12]: %sql select min(DATE) from SPACEXTBL where Landing__Outcome = 'Success (ground pad)'
```

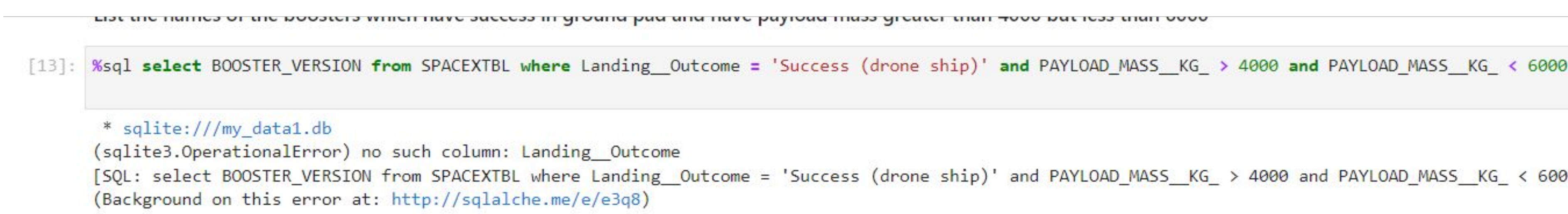
```
 * sqlite:///my_data1.db
(sqlite3.OperationalError) no such column: Landing__Outcome
[SQL: select min(DATE) from SPACEXTBL where Landing__Outcome = 'Success (ground pad)']
(Background on this error at: http://sqlalche.me/e/e3q8)
```

This query returns the first successful ground pad landing date.

First ground pad landing wasn’t

until the end of 2015.

Successful Drone Ship Landing with Payload Between 4000 and 6000



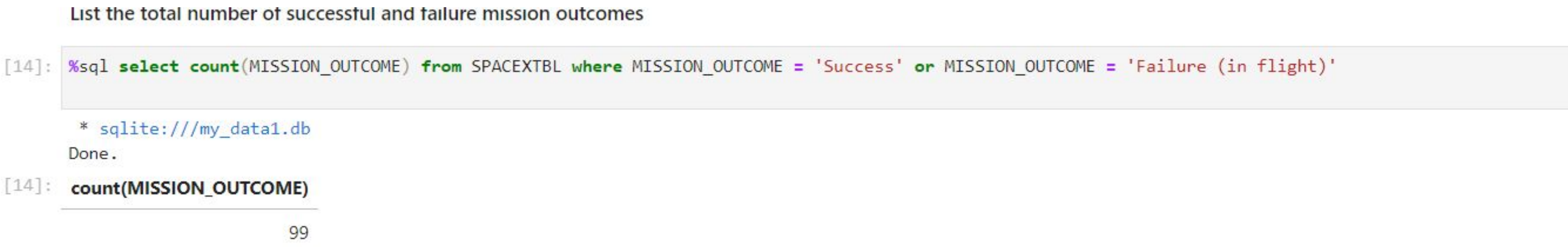
List the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000

```
[13]: %sql select BOOSTER_VERSION from SPACEXTBL where Landing__Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000
```

```
 * sqlite:///my_data1.db
(sqlite3.OperationalError) no such column: Landing__Outcome
[SQL: select BOOSTER_VERSION from SPACEXTBL where Landing__Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 600
(Background on this error at: http://sqlalche.me/e/e3q8)
```

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 Each Mission Outcome



List the total number of successful and failure mission outcomes

```
[14]: %sql select count(MISSION_OUTCOME) from SPACEXTBL where MISSION_OUTCOME = 'Success' or MISSION_OUTCOME = 'Failure (in flight)'
```

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

[14]:

count(MISSION_OUTCOME)
99

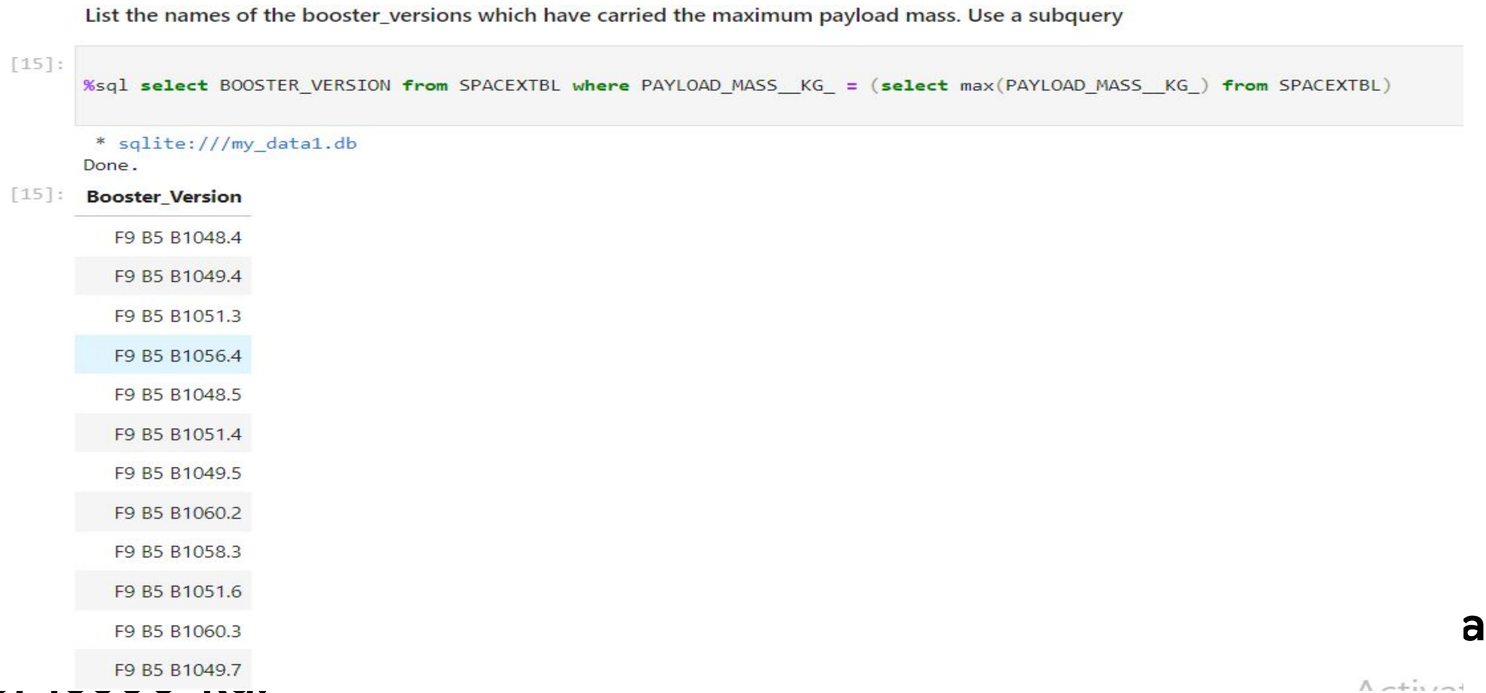
This query returns a count of each mission outcome.

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

This means that most of the landing

failures are intended.

Boosters that Carried Maximum Payload



List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
[15]:  %sql select BOOSTER_VERSION from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)

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

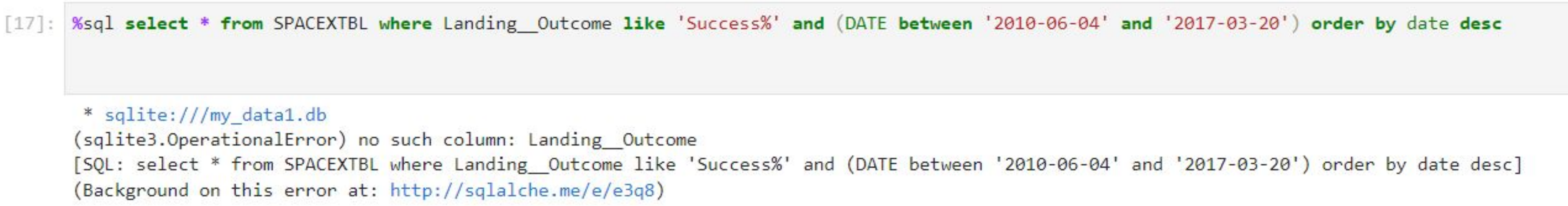
[15]:

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

ayload mass

These booster versions are very similar and all are of the F9 B5 B10xx.x variety.

Ranking Counts of Successful Landings Between 2010-06-04 and 2017-03-20



```
[17]: %sql select * from SPACEXTBL where Landing__Outcome like 'Success%' and (DATE between '2010-06-04' and '2017-03-20') order by date desc
```

```
 * sqlite:///my_data1.db
(sqlite3.OperationalError) no such column: Landing__Outcome
[SQL: select * from SPACEXTBL where Landing__Outcome like 'Success%' and (DATE between '2010-06-04' and '2017-03-20') order by date desc]
(Background on this error at: http://sqlalche.me/e/e3q8)
```

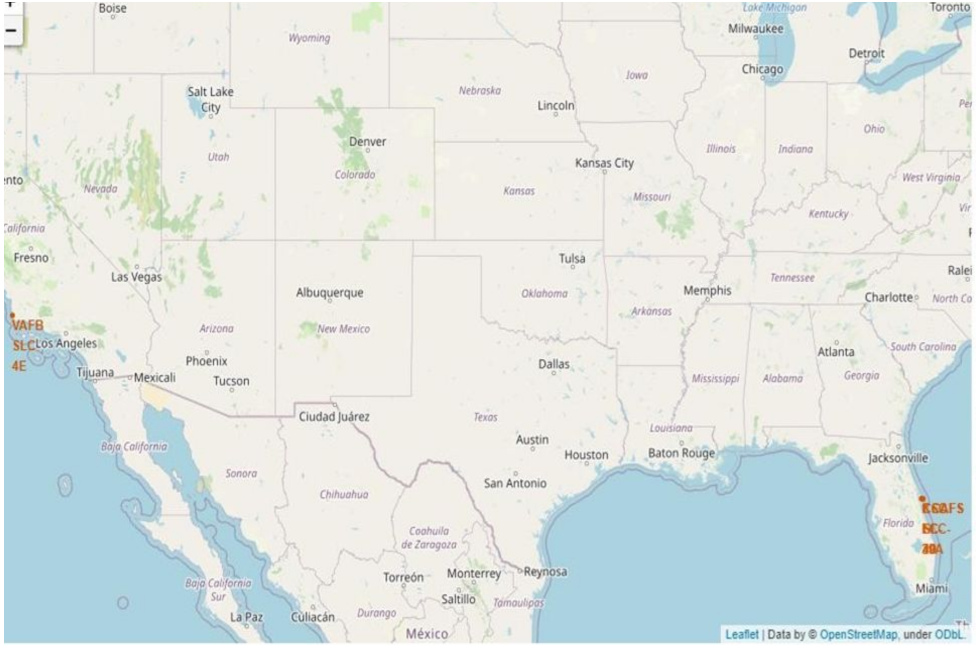
This query returns a list of successful landings and between 2010-06-04 and 2017-03-20 inclusively.

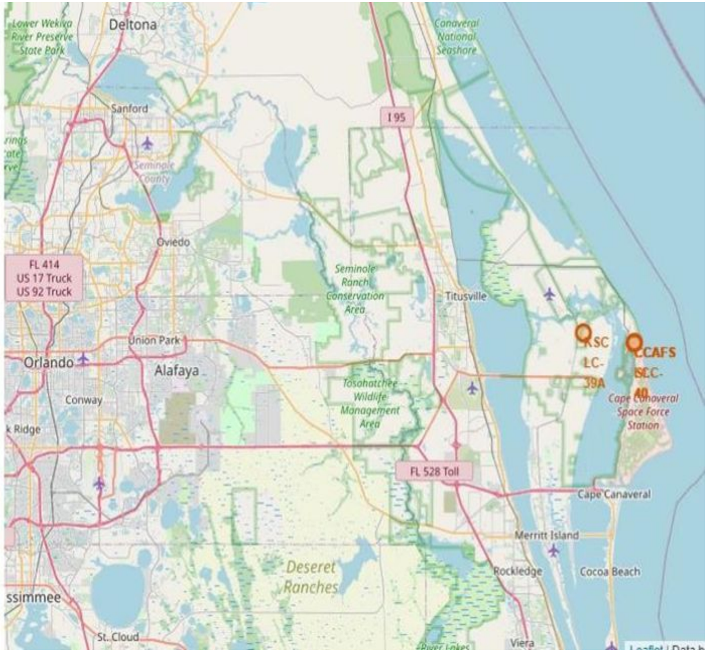
There are two types of successful landing outcomes: drone ship and ground pad landings.

Interactive Map with Folium

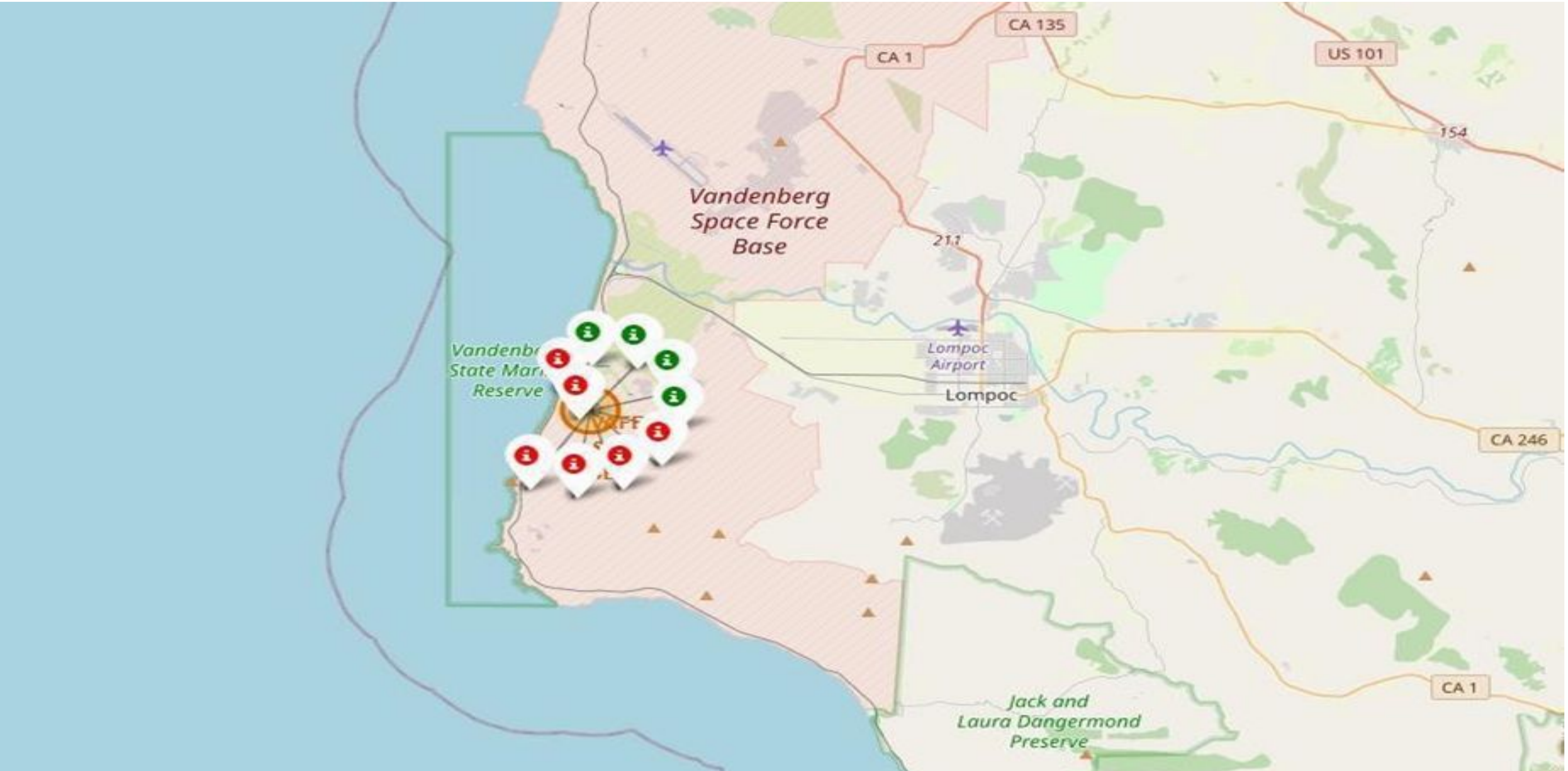
Visualizing with maps using folium

Launch Site Location





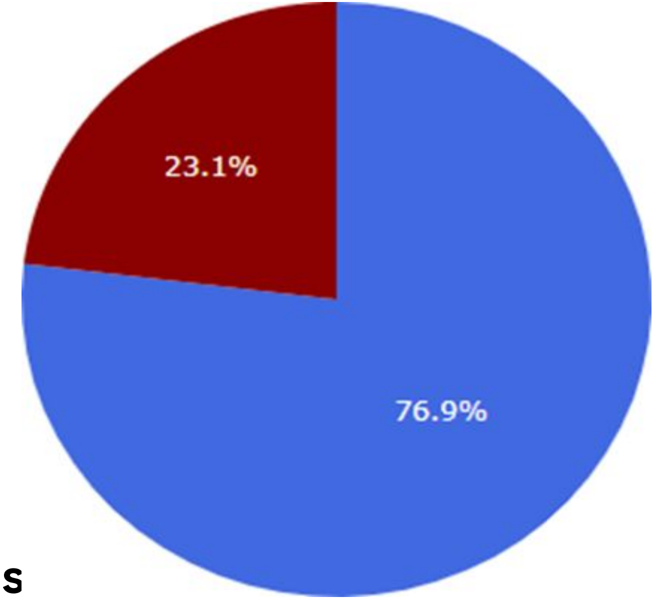
Color-Coded Launch Markers



Dashboard with Plotly Dashboard

Building a dashboard with plotly

Highest Success Rate Launch Site

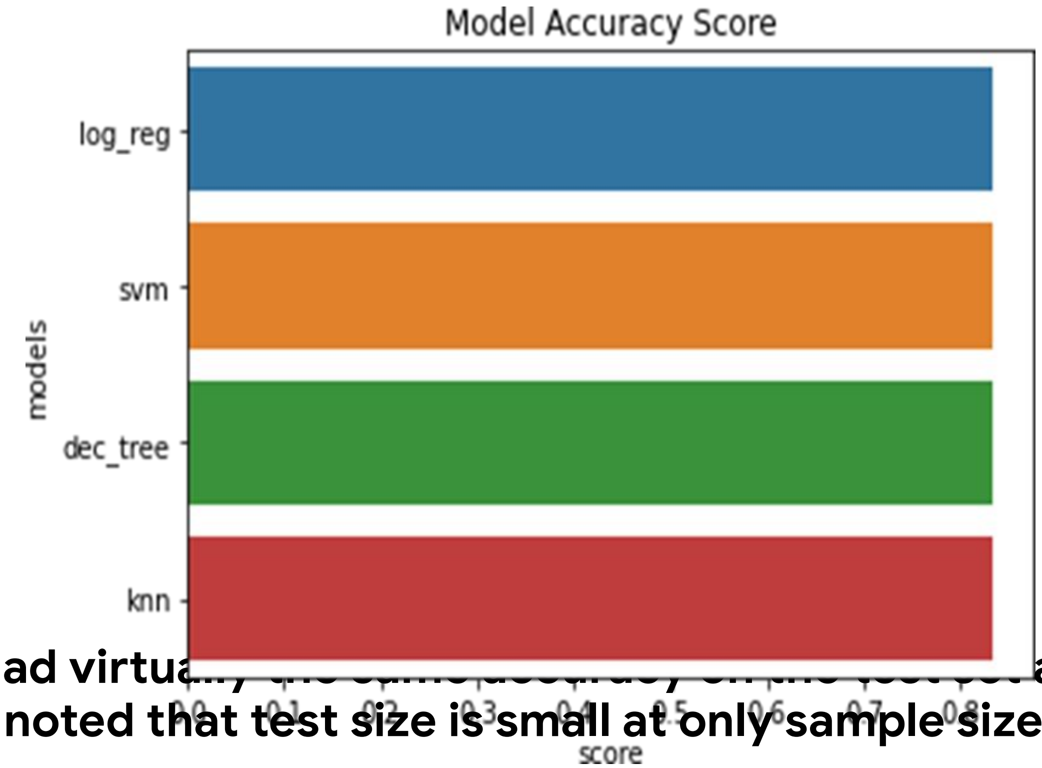


KSC LC-39A has the highes ssful landings and 3 failed landings.

Predictive Classification

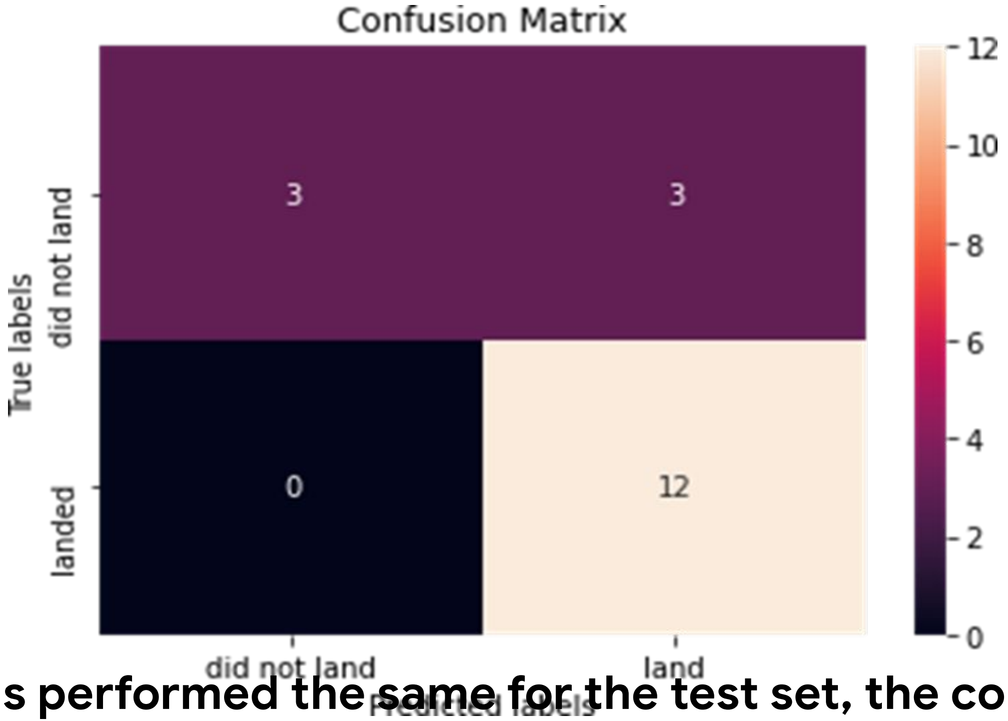
Perform a analysis with multiple algorithms

Model and classification accuracy



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 of 18.

Visualizing confusion matrix



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.

Conclusion

- In summary, we have successfully developed a machine learning model for Space Y aimed at competing with SpaceX.
- The primary objective of this model is to forecast the likelihood of a successful Stage 1 landing, potentially saving approximately $100 million USD per launch.
- Developed an 83% accurate ML model for Space Y to predict Stage 1 landing success.
- Gathered data from SpaceX API and Wikipedia, stored it in a DB2 SQL database.
- Created a visualization dashboard for easy decision-making.
- Enables Allon Mask and Space Y to make informed launch decisions.
- Ongoing data collection efforts recommended for improving accuracy.

Appendix

- Github Repository URL : https://github.com/kavyansh091/capstone_project
- Special Thanks to all the instructors