IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Batın Ö.
13.09.2023

Outline

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

Executive Summary

- "Analyzing SpaceX Data for Successful Landings
- Collected data from SpaceX API and Wikipedia.
- Created 'class' labels for successful landings.
- Explored data with SQL, visualizations, maps, and dashboards.
- Engineered features and applied one-hot encoding.
- Standardized data and optimized models using GridSearchCV.
- Visualized model accuracy.
- Results:
- Four ML models: Logistic Regression, Support Vector Machine, Decision Tree, K Neighbors.
- All achieved ~83.33% accuracy.
- Models tended to overpredict success.
- More data needed for improved accuracy.

Introduction

- **Background:**
  - The Commercial Space Age is Thriving
  - SpaceX Leads with Competitive Pricing ($62 million vs. $165 million USD)
  - SpaceX's Success Attributed to Rocket Recovery (Stage 1)
  - Space Y Aims to Challenge SpaceX
- **Problem:**
  - Space Y Entrusts Us with Training a Machine Learning Model for Predicting Successful Stage 1 Rocket Recovery

Section 1

Methodology

Methodology

- Data Collection & Analysis Process:

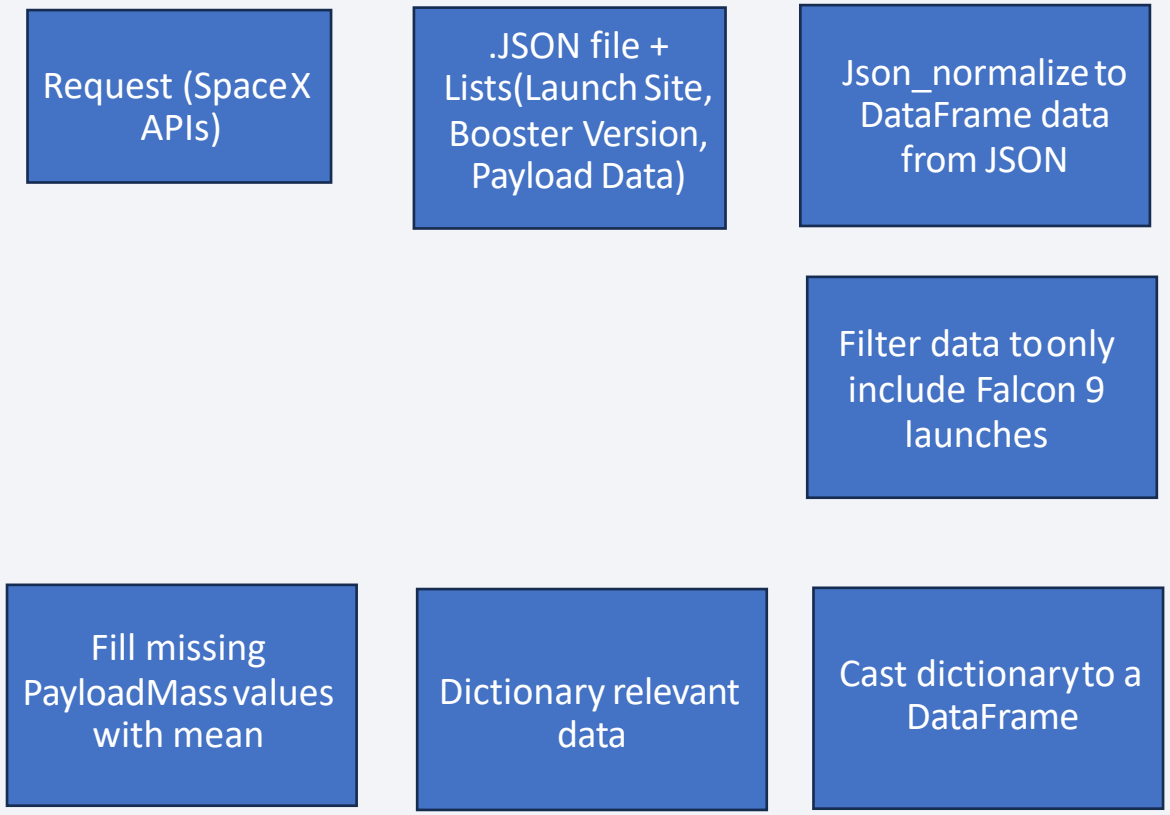
1. Gathered data from SpaceX API and Wikipedia
2. Cleaned and organized the data
3. Identified successful and unsuccessful landings
4. Explored data with visuals and SQL
5. Created interactive visuals with Folium & Plotly Dash
6. Predictive analysis using classification models
7. Fine-tuned models with GridSearchCV

Data Collection

- *Data collection involved two methods: API Requests and Web Scraping.*
- **API Requests:** Gathered data from SpaceX's public API.
- **Web Scraping:** Extracted data from a table in SpaceX's Wikipedia entry.
- **SpaceX API Data Columns:** FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- **Wikipedia Web Scraping Data Columns:** Flight No., Launch Site, Payload, Payload Mass, Orbit, Customer, Launch Outcome, Version Booster, Booster Landing, Date, Time

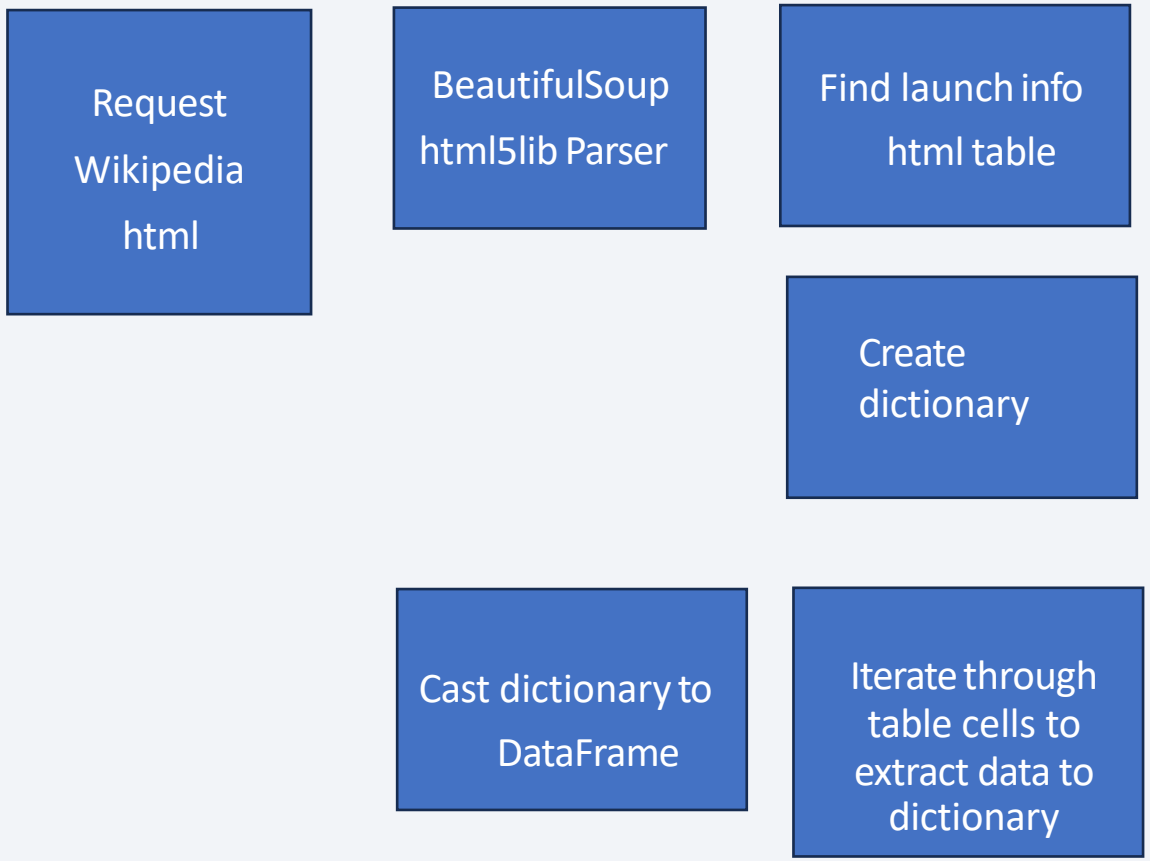
Data Collection – SpaceX API

- Github:
- https://github.com/batinium/applied-data-science-capstone/blob/main/ADSC_1_jupyter_labs_spacex_data_collection_api.ipynb



Data Collection - Scraping

- Github
- https://github.com/batinium/applied-data-science-capstone/blob/main/ADSC-2-jupyter_labs_webscraping.ipynb



Data Wrangling

- **Objective:** Create a training label for landing outcomes, where success is represented as '1' and failure as '0'.
- **Outcome Components:**
  - 'Mission Outcome'
  - 'Landing Location'
- **New Training Label Column:** 'class'
- **Value Mapping:**
  - If 'Mission Outcome' is True: 'Landing Location' True ASDS, True RTLS, & True Ocean → Set 'class' to 1
  - If 'Mission Outcome' is False: 'Landing Location' None None, False ASDS, None ASDS, False Ocean, False RTLS → Set 'class' to 0

Github: https://github.com/batinium/applied-data-science capstone/blob/main/ADSC_2_1data_wrangling_.ipynb

EDA with Data Visualization

- Exploratory Data Analysis (EDA) focused on key variables: Flight Number, Payload Mass, Launch Site, Orbit, Class, and Year. Various plots, including scatter plots, line charts, and bar plots, were employed to assess relationships among these variables. This analysis aimed to identify any meaningful associations that could be leveraged for machine learning model training.

- Github: https://github.com/batinium/applied-data-science-capstone/blob/main/ADSC_4_EDA_with_Visualization.ipynb

EDA with SQL

- Loaded data into IBM DB2 Database, queried with SQL Python integration to understand launch site names, mission outcomes, customer payload sizes, booster versions, and landing outcomes.

- Github: https://github.com/batinium/applied-data-science-capstone/blob/main/ADSC_3_sql_coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Using Folium maps, we visually display Launch Sites, both successful and unsuccessful landings, and their proximity to important landmarks such as Railway, Highway, Coast, and City. This helps us comprehend the rationale behind the choice of launch site locations and provides a clear visualization of successful landings in relation to these key points.

- Github: https://github.com/batinium/applied-data-science-capstone/blob/main/ADSC_5_Interactive_Visual_Analytics_with_Folium.ipynb

Build a Dashboard with Plotly Dash

- Dashboard Overview:
- **Key Visuals:** Pie Chart and Scatter Plot.
- **Pie Chart:** Choose between overall launch site success distribution or individual site success rates.
- **Scatter Plot:** Options to display data for all launch sites or specific sites, with payload mass adjustable from 0 to 10,000 kg.
- **Usage:**
- **Pie Chart:** Visualizes launch site success rates.
- **Scatter Plot:** Helps analyze success variations across launch sites, payload mass, and booster version categories.

- Github: https://github.com/batinium/applied-data-science-capstone/blob/main/ADSC_6_Machine_Learning_Prediction.ipynb

Predictive Analysis (Classification)

- Github: https://github.com/batinium/applied-data-science-capstone/blob/main/ADSC_6_Machine_Learning_Prediction.ipynb

Split label column 'Class' from dataset

Fit and Transform Features using Standard Scaler

Score models on split test set

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

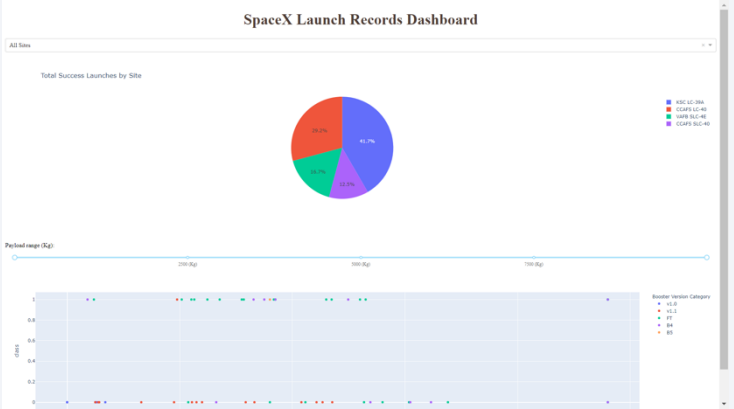
Barplot to compare scores of models

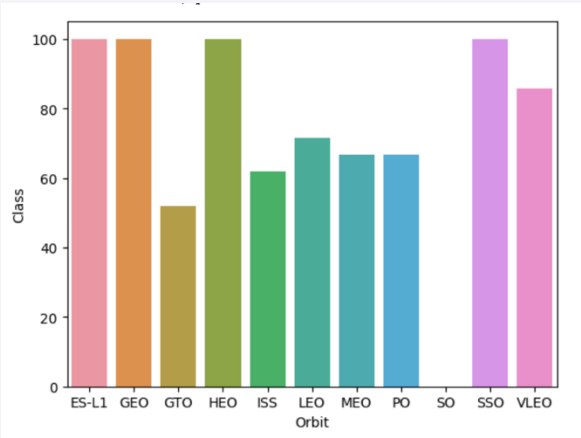
Train_test_split data

GridSearchCV (cv=10) to find optimal parameters

Confusion Matrix for all models

Results



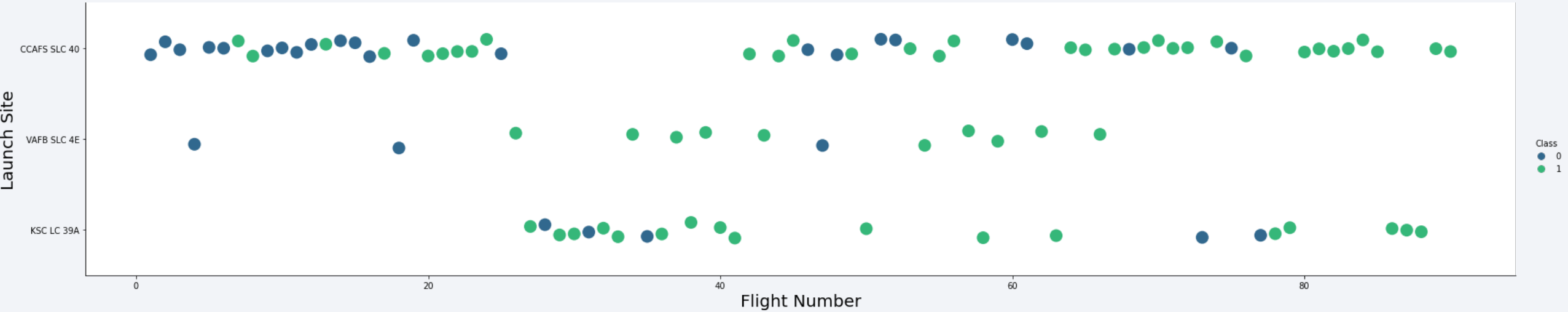


- This serves as an introductory glimpse into the forthcoming Plotly dashboard, offering a sneak peek at its various components. Within this dashboard, you can expect to find informative sections dedicated to Exploratory Data Analysis (EDA) presented through dynamic visualizations, EDA enriched with SQL insights, an Interactive Map powered by Folium, and culminating with the unveiling of our model's performance metrics, boasting an impressive accuracy rate of approximately 83%.

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

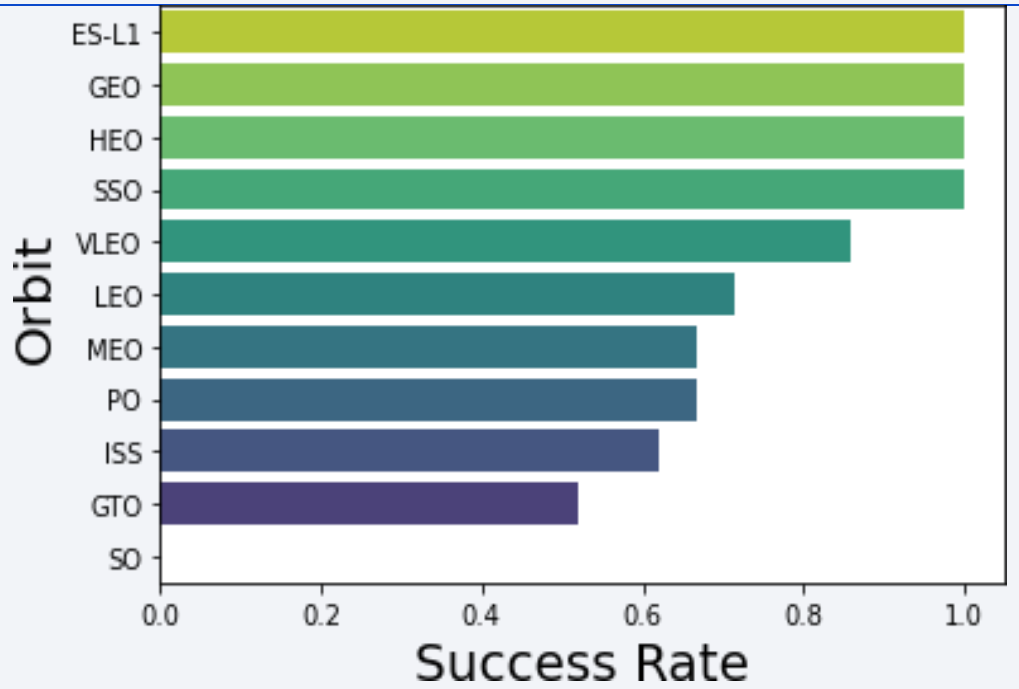


- The colors on the graphic convey information about rocket launches. Green means a launch was successful, while purple means it was not.
- The graph also shows that the success rate of launches has been improving as the flight number increases. There seems to be a notable improvement in success around flight number 20, which was likely a significant breakthrough.
- The launch site at CCAFS appears to be the busiest, with the most launches happening there.

Payload vs. Launch Site

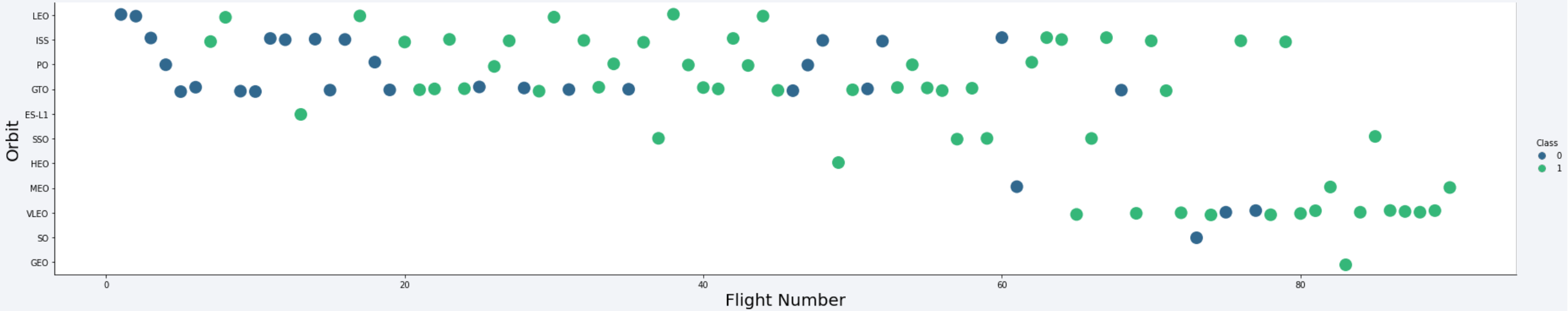
- Green means the launch was successful, while purple means the launch was unsuccessful.
- The weight of the payload typically ranges from 0 to 6000 kilograms. Additionally, various launch sites have different payload weight requirements.

Success Rate vs. Orbit Type



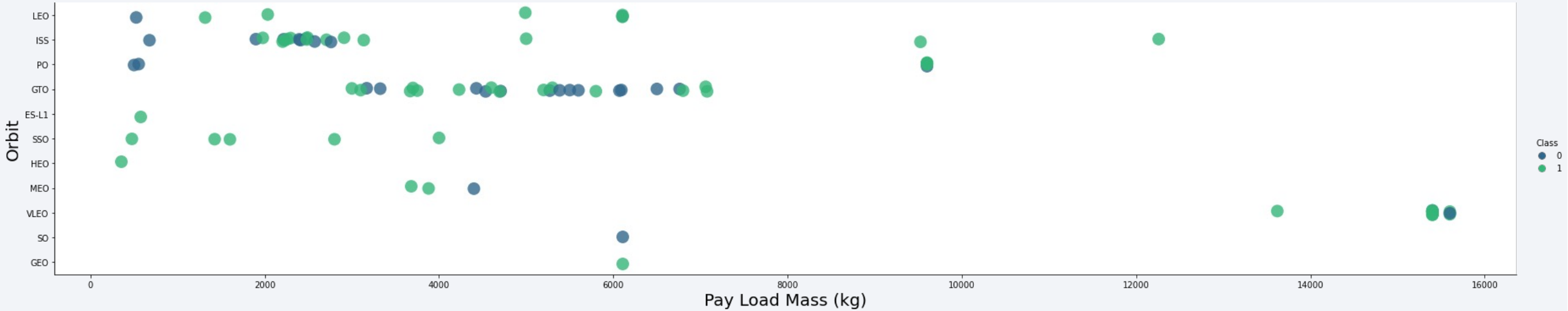
1. ES-L1 (1), GEO (1), and HEO (1) missions have a 100% success rate, with each having only one mission attempt (sample sizes in parentheses).
2. SSO (5) missions also have a 100% success rate, but there have been a total of five mission attempts in this category.
3. VLEO (14) missions have a decent success rate and have been attempted 14 times.
4. SO (1) missions have a 0% success rate, with one mission attempt resulting in failure.
5. GTO (27) missions have an approximately 50% success rate, making it the category with the largest number of mission attempts, totaling 27.

Flight Number vs. Orbit Type



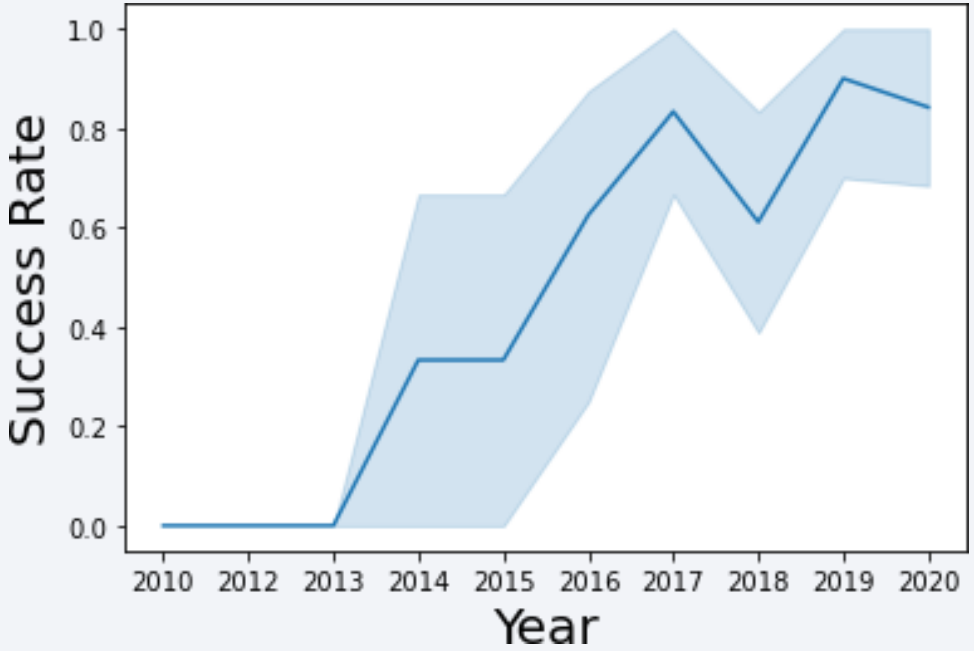
- Green means a successful launch, while purple signifies an unsuccessful one.
- SpaceX's choice of launch orbit has shifted as they've conducted more flights. It seems that the success of their launches is linked to this choice.
- Initially, SpaceX used Low Earth Orbit (LEO) for their launches and had moderate success. However, in their recent launches, they've returned to Very Low Earth Orbit (VLEO). It appears that SpaceX tends to perform better in lower orbits or orbits that are synchronized with the Sun's position.

Payload vs. Orbit Type



- When we look at the color codes, green means a successful launch, while purple indicates an unsuccessful one.
- It appears that the weight of the payload is connected to the type of orbit achieved. For orbits like Low Earth Orbit (LEO) and Sun-Synchronous Orbit (SSO), we often see relatively light payloads. However, when it comes to Very Low Earth Orbit (VLEO), the most successful orbits tend to have heavier payload masses, typically on the higher side of the range.

Launch Success Yearly Trend



- Success has generally been on the rise since 2013, with a small decrease in 2018. In recent years, the success rate has remained consistently around 80%.

All Launch Site Names

- I want to retrieve the names of unique launch sites from the database. It appears that 'CCAFS SLC-40' and 'CCAFSSLC-40' likely refer to the same launch site due to data entry errors. The previous name for this site was 'CCAFS LC-40.' So, there are likely only three unique launch site values: 'CCAFS SLC-40,' 'KSC LC-39A,' and 'VAFB SLC-4E.

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

Launch Site Names Begin with 'CCA'

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 calculates the combined weight of payloads in kilograms that were transported to the International Space Station (ISS) when NASA was the customer. The term 'CRS' stands for Commercial Resupply Services, which signifies that these payloads were delivered to the 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:***@0c77d6f2-5da9-48a9-81f8-8
Done.
```

avg_payload_mass_kg
2928

- This query is used to find the average weight of payloads carried by launches that utilized the booster version known as "F9 v1.1." The average payload mass for the F9 1.1 booster falls towards the lower end of our payload weight 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 provides information about the date of the first successful ground pad landing. The first successful ground pad landing did not occur until the end of 2015. In general, successful landings began to occur in 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 will provide a list of four booster versions that both successfully landed on a drone ship and had a payload mass within the range of 4000 to 6000, excluding the exact values of 4000 and 6000.

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 statement discusses the results of a query related to SpaceX missions. It reveals that SpaceX has a very high success rate, achieving its mission objectives in nearly 99% of cases. This suggests that most of the times when a mission doesn't go as planned, it's actually intentional, possibly for testing or other reasons. Interestingly, there is one mission with an uncertain payload status, and unfortunately, there was one mission that failed during its 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

- This query provides information about booster versions and their relationship to payload mass. It tells us that the query is looking for booster versions that were used to carry the highest payload mass, which was 15600 kg. These booster versions all belong to a category known as "F9 B5 B10xx.x." The fact that these booster versions are very similar suggests that there is a correlation between the payload mass and the specific booster version used. In other words, it seems that the choice of booster version is influenced by the weight of the payload it needs to carry.

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

- This query fetches information about SpaceX launches that took place in 2015 and meets the following criteria: the first stage of the rocket failed to land on a drone ship. The information retrieved includes the month of the launch, the outcome of the landing attempt, the version of the booster used, the mass of the payload in kilograms, and the launch site. There were two instances in 2015 where this specific situation occurred.

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

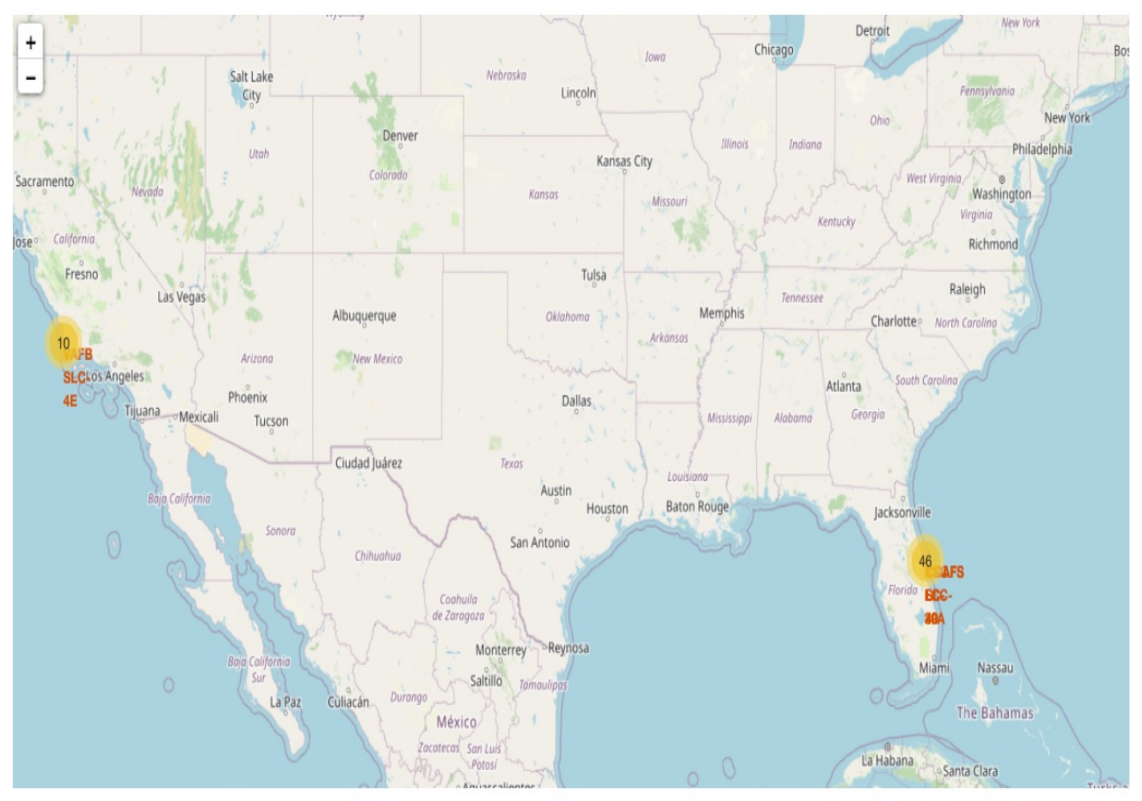
- This query provides a list of successful rocket landings that occurred between June 4, 2010, and March 20, 2017. Successful landings are categorized into two types: drone ship landings and ground pad landings. In total, there were 8 successful landings during this specified time frame.

Section 3

Launch Sites Proximities Analysis

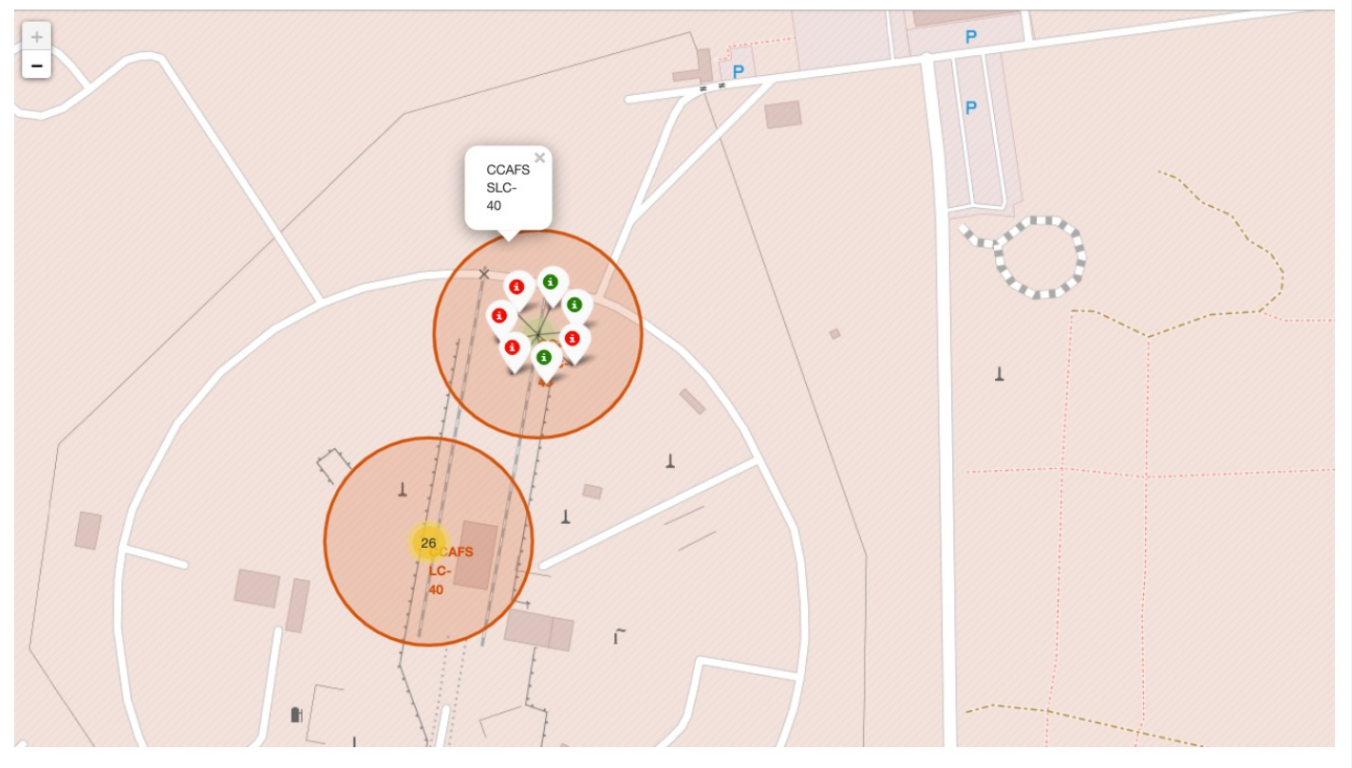
<Launch Site>

- The map displays all the launch sites in relation to the entire United States. We also have a closer view of the two launch sites in Florida, which are situated very close to each other. It's important to note that all of these launch sites are located near the ocean.



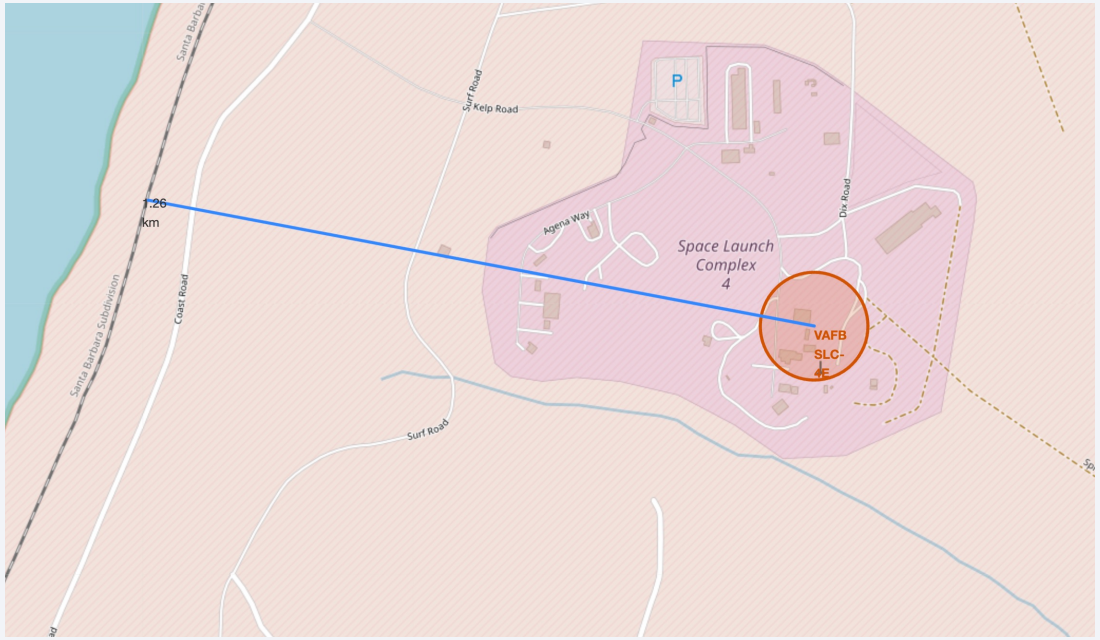
Landings

- You can click on the clusters on the Folium map to see information about each landing. Successful landings are represented by green icons, while failed landings are represented by red icons. For instance, in the case of VAFB SLC-4E on this map, there are 4 successful landings and 6 failed landings.



Launch Site Approximation

- Let's take KSC LC-39A as an example. Launch sites, like this one, are strategically located near railways for easy transportation of large equipment and supplies. They are also situated close to highways for the convenience of transporting both people and supplies.
- Additionally, launch sites are positioned near coastal areas and away from densely populated cities. This arrangement is designed to ensure that if there's a problem during a rocket launch, and the rocket needs to be aborted, it can safely fall into the sea instead of landing in a heavily populated area.

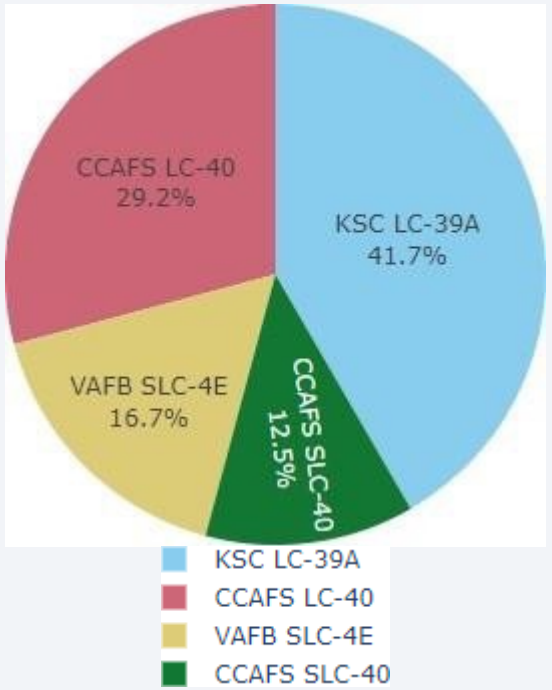


Section 4

Build a Dashboard with Plotly Dash

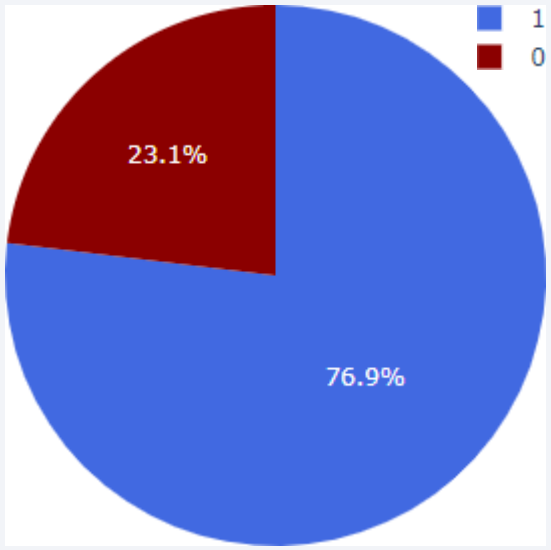
Successful Launches

- This chart shows the number of successful rocket landings at different launch sites. CCAFS LC-40 used to be called CCAFS SLC-40, so CCAFS and KSC have the same number of successful landings, but most of these happened before the name change. VAFB has the fewest successful landings, possibly because it has a smaller number of launches and launching rockets on the west coast is more challenging.



Success Rate of Launch Site

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



Payload Mass, Success and Booster Version

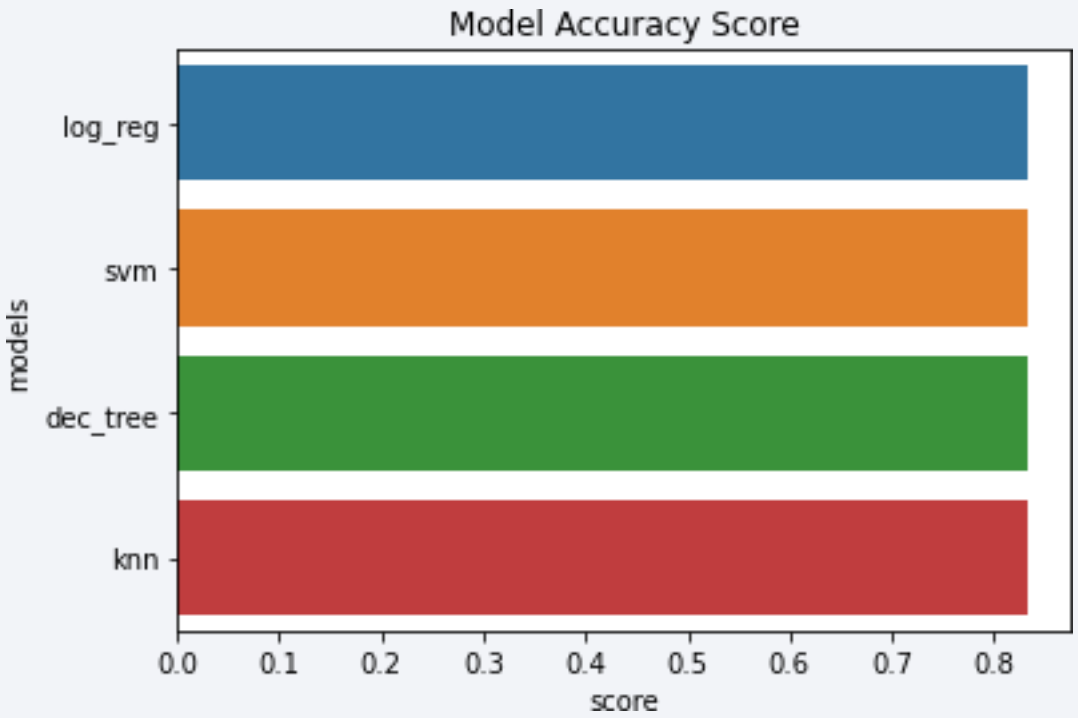
- The Plotly dashboard has a feature called the "Payload range selector," which allows you to choose a range of payload values. However, the default range is set from 0 to 10,000, even though the maximum payload capacity is 15,600.
- In the context of the data, there's a classification system where "1" indicates a successful landing, and "0" represents a failure. The scatter plot on the dashboard also takes into account the booster version category by using different colors, and it adjusts the size of the data points based on the number of launches.
- What's interesting is that within the range of 0 to 6,000 on the dashboard, there are two instances where there were failed landings, and the payloads for these failed landings were zero kilograms.

Section 5

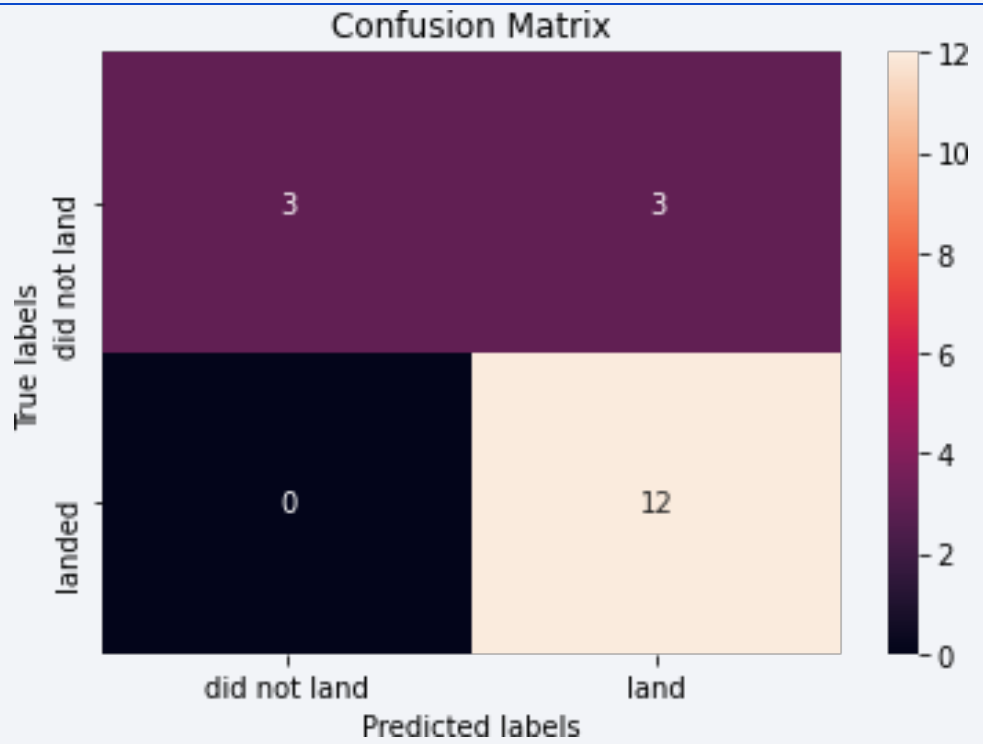
Predictive Analysis (Classification)

Classification Accuracy

- All the models achieved nearly identical accuracy scores of 83.33% on the test dataset. It's important to mention that the test dataset is quite small, consisting of only 18 samples. This limited sample size can lead to significant fluctuations in accuracy results, especially in models like the Decision Tree Classifier when running multiple times. To make a more confident determination about the best model, it is likely that we need a larger amount of data.



Confusion Matrix



- In the context of predictions, correct outcomes are aligned along a diagonal line from the top-left corner to the bottom-right corner.
- In our case, all the models we tested performed equally on the test dataset, so their confusion matrices (a way to assess prediction performance) are identical. Specifically:
- All the models correctly predicted 12 successful landings when the actual label indicated a successful landing.
- They also correctly predicted 3 unsuccessful landings when the true label was indeed an unsuccessful landing.
- However, the models made a mistake in predicting 3 successful landings when, in reality, the landings were unsuccessful (these are called false positives). This suggests that our models tend to overestimate the number of successful landings.

Conclusions

- Our task was to build a machine learning model for Space Y, a company competing with SpaceX. The purpose of this model is to predict when the first stage of a rocket will successfully land, which can save Space Y approximately $100 million USD.
- To develop this model, we gathered data from a public SpaceX API and also collected information by scraping SpaceX's Wikipedia page. We then organized this data, created labels for it, and stored it in a DB2 SQL database. To make the insights more accessible, we designed a dashboard for visualization.
- The machine learning model we constructed achieved an accuracy rate of 83%. This means that Allon Mask, the CEO of SpaceY, can use this model to make relatively accurate predictions about whether a launch will result in a successful Stage 1 landing before the actual launch takes place. This information can help SpaceY decide whether to proceed with the launch or not.
- To further enhance the model's performance, it is advisable to gather more data. Additional data can aid in identifying the most effective machine learning model and ultimately improving the accuracy of the predictions.

Appendix

- Github: https://github.com/batinium/applied-data-science-capstone/tree/main
- This project was made possible with https://www.coursera.org/search?query=IBM%20applied%20data%20capstone&

Thank you!