IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Omar Muñoz
January 2024

Outline

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

Executive Summary

- Web scraping, data visualization, and predictive modeling can help us get insights into SpaceX's Falcon 9 landings and the reasons for their success or failure.
- Ever since 2013, Falcon 9 landings have become more successful. Nowadays its success rate is above 80%.
- Data science team has been able to **prototype** a ML model that predicts whether the landing will be successful or not with **88% accuracy**. This model can help with SpaceY's pricing decisions.

Introduction

- Falcon 9 launches cost approx. 60% less than its competitors because SpaceX proposes to reuse the first stage.
- Predicting SpaceX's first-stage reuse can support SpaceY pricing decisions.
- Can we determine which factors contribute to SpaceX's successful first stage landings?
- Can we accurately predict SpaceX's landing outcome?

Falcon 9 reusable first stage

Source: https://www.space.com/39172-spacex-to-skip-first-stage-landing-for-upcoming-iridium-launch.html

Section 1

Methodology

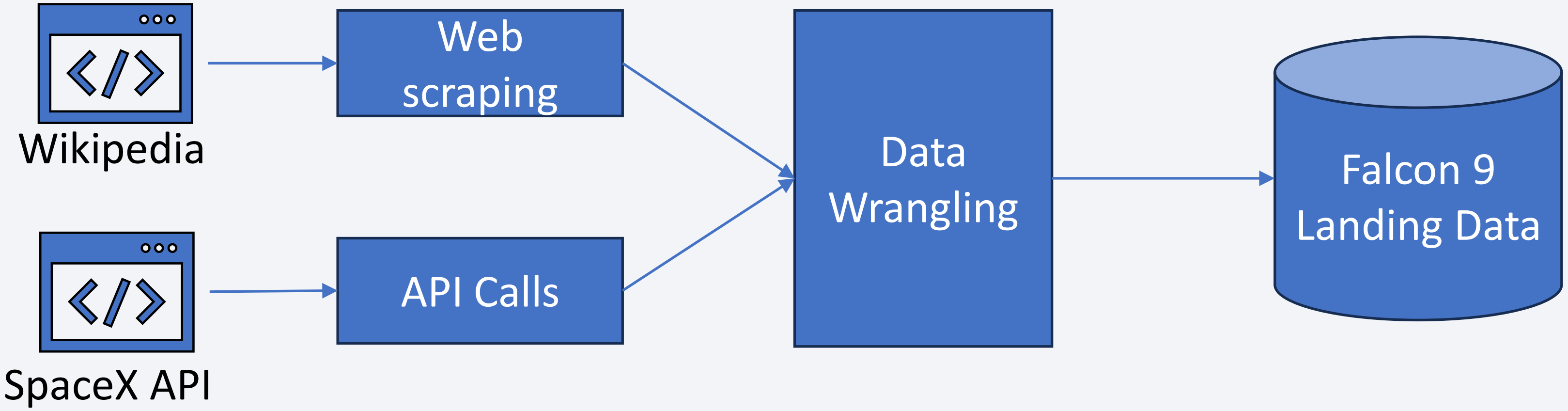
Methodology

Executive Summary

- Data collection methodology:
  - SPACEX API & Web Scraping
- Perform data wrangling
  - Filter Falcon 9 data, extract relevant attributes, fix missing values, one hot encoding, label encoding, among others.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Four different ML models (Logistic regression, SVM, Decision Tree, KNN) were tuned, and compared using the accuracy metric.

Data Collection

- Two sources of information for collecting data:
  - Wikipedia historical records: [List of Falcon 9 and Falcon Heavy launches.]
  - Non-official SPACEX API: https://docs.spacexdata.com/
- Data collection pipeline:

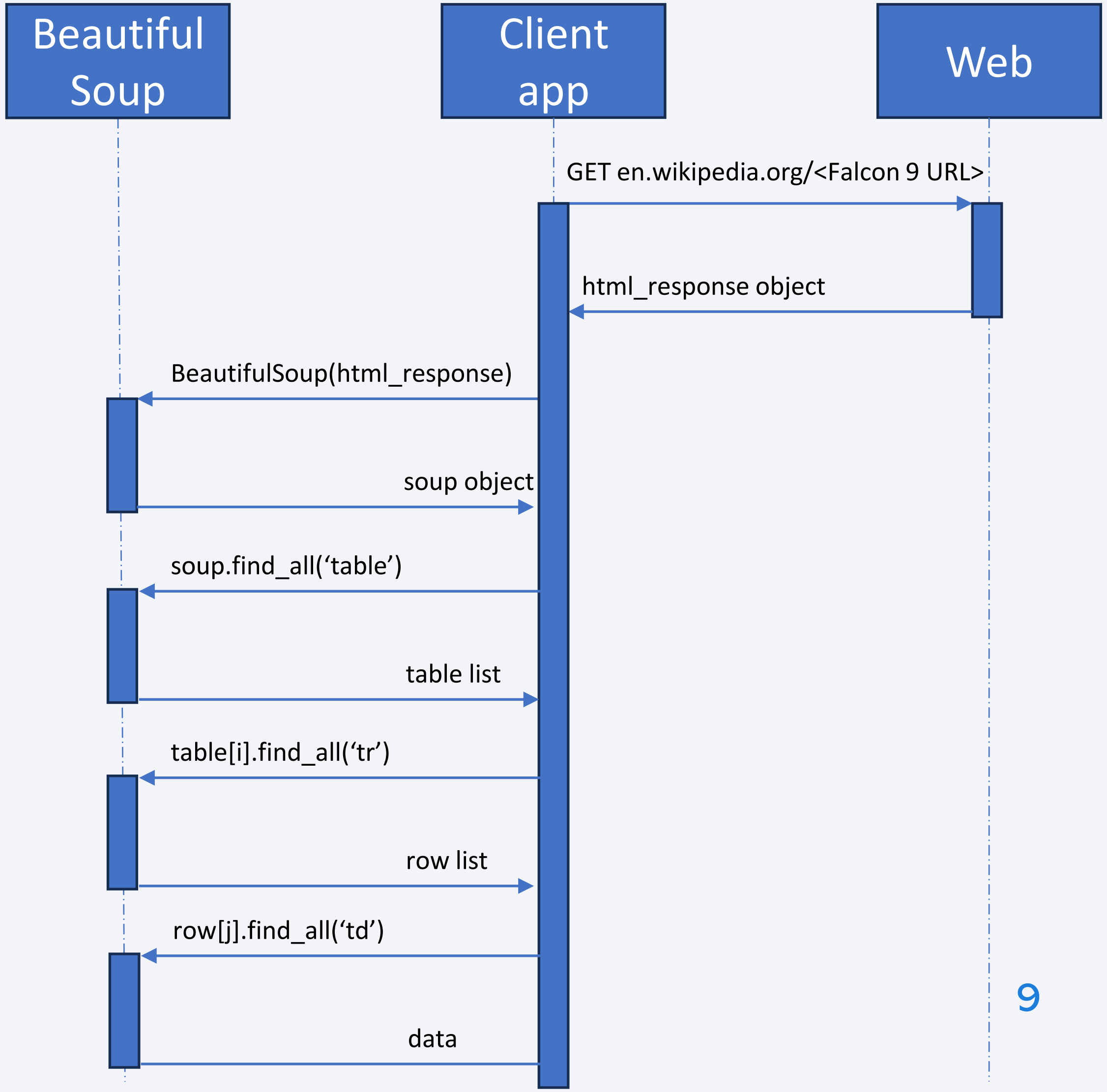


Data Collection – SpaceX API

- SpaceX REST API
  - “Open-Source API for launch, rocket, core, capsule, starlink, launchpad, and landing pad data.”
- Using GET calls we obtain past launch data and further GET calls help us map IDs to names and values.
- The JSON files are parsed and a DataFrame is created.
- [Jupiyter Notebook]

Data Collection - Scraping

- Using the **requests** and **BeautifulSoup** libraries we can scrap and parse the Falcon 9 launch records from Wikipedia.
- BeautifulSoup extracts the raw data and further preprocessing is required for constructing a DataFrame from the extracted info.
- [Jupyter Notebook]()



Data Wrangling

- Missing values:
  - 5/90 Records without Payload Mass → Replaced with average value.
  - 26/90 Records without Landing Pad → Remained the same as it represents no landing pad was used.
- One-hot encoding for categorical variables:
  - Orbit, LaunchSite, LandingPad, Serial, GridFins, Reused, Legs.
  - In the end we use 83 features.
- Mapping Landing Outcomes to target variable:
  - 'False ASDS', 'False Ocean', 'False RTLS', 'None ASDS', 'None None' → 0 (Fail)
  - 'True ASDS', 'True RTLS', 'True Ocean' → 1 (Success)
- Related notebooks:
  - Web scraping
  - SpaceX API
  - Data Wrangling
  - Machine Learning Prediction

EDA with Data Visualization

- Analyze how the Launch Site relates to relevant factors:
  - Flight Number vs. Launch Site
  - Payload vs. Launch Site
- Analyze how the target orbit for the mission relates to relevant factors:
  - Success Rate vs. Orbit Type
  - Flight Number vs. Orbit Type
  - Payload vs. Orbit Type
- Analyze SpaceX maturity over time
  - Landing Success Yearly Trend
- Related notebook: EDA with Data Visualization

EDA with SQL

Related notebook: [EDA with SQL]

SQL queries to:

- Determine the unique Launch Sites
- Show the data for the first five launches from any Cape Canaveral facility
- Compute the total payload mass carried by boosters launched by NASA
- Compute average payload mass carried by booster version F9 v1.1
- Find the date of the first successful landing outcome in ground pad
- List the names of boosters which have successfully landed on drone ship and had payload mass between 4000 and 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster which have carried the maximum payload mass
- List the landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20

Build an Interactive Map with Folium

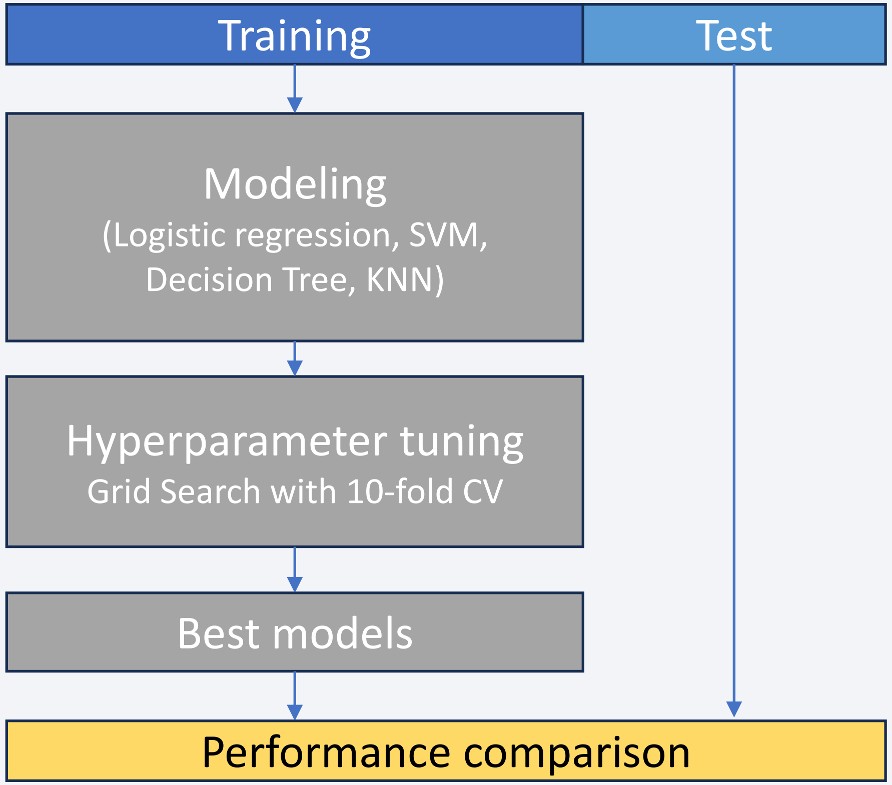
- Map visualization enables us to understand location data better:
  - Adding circles and markers helps us to visually locate the different SpaceX launch sites.
  - Marker colors can help us visualize the landing outcomes (success vs. failure) right on top of the launch site locations.
  - We can use lines to visually assess the launch site proximity to relevant landmarks such as railways, highways, coastlines, and cities.
- Related notebook: [Interactive Map with Folium]()

Build a Dashboard with Plotly Dash

- A dashboard can help us to interactively analyze data of interest.
  - Pie charts were added to explore the most successful Launch Sites.
  - Scatter plots were added to explore the correlation between the Payload and the Success for different Launch Sites.
- Related python script: Dash App

Predictive Analysis (Classification)

- Falcon 9 data is divided in 80% training and 20% testing.
- Four different ML models were fitted on the training split:
  - Hyperparameter tuning was done using 10-fold CV and Grid Search on the training set.
- Accuracy on the testing set is used to determine the best model.
- Related notebook: Predictive Analysis



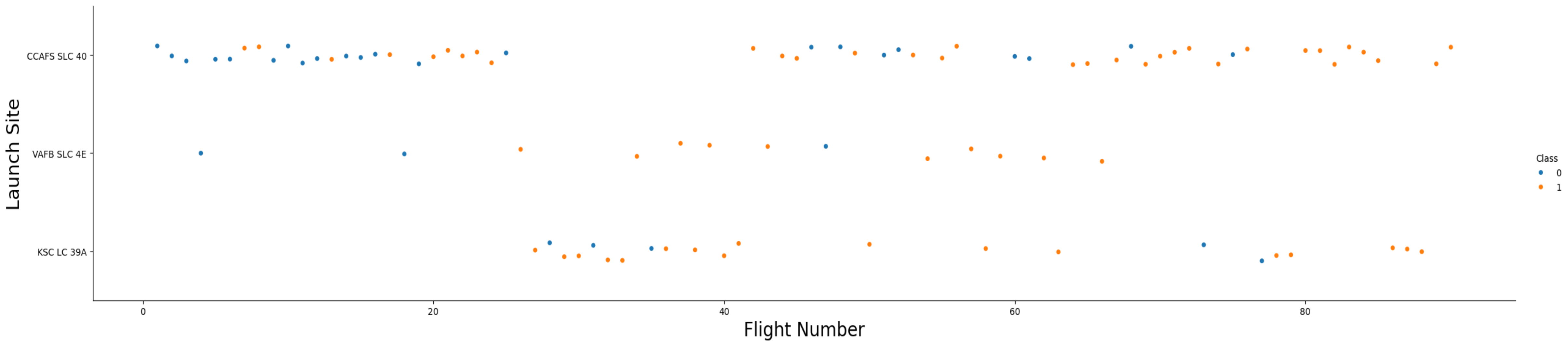
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

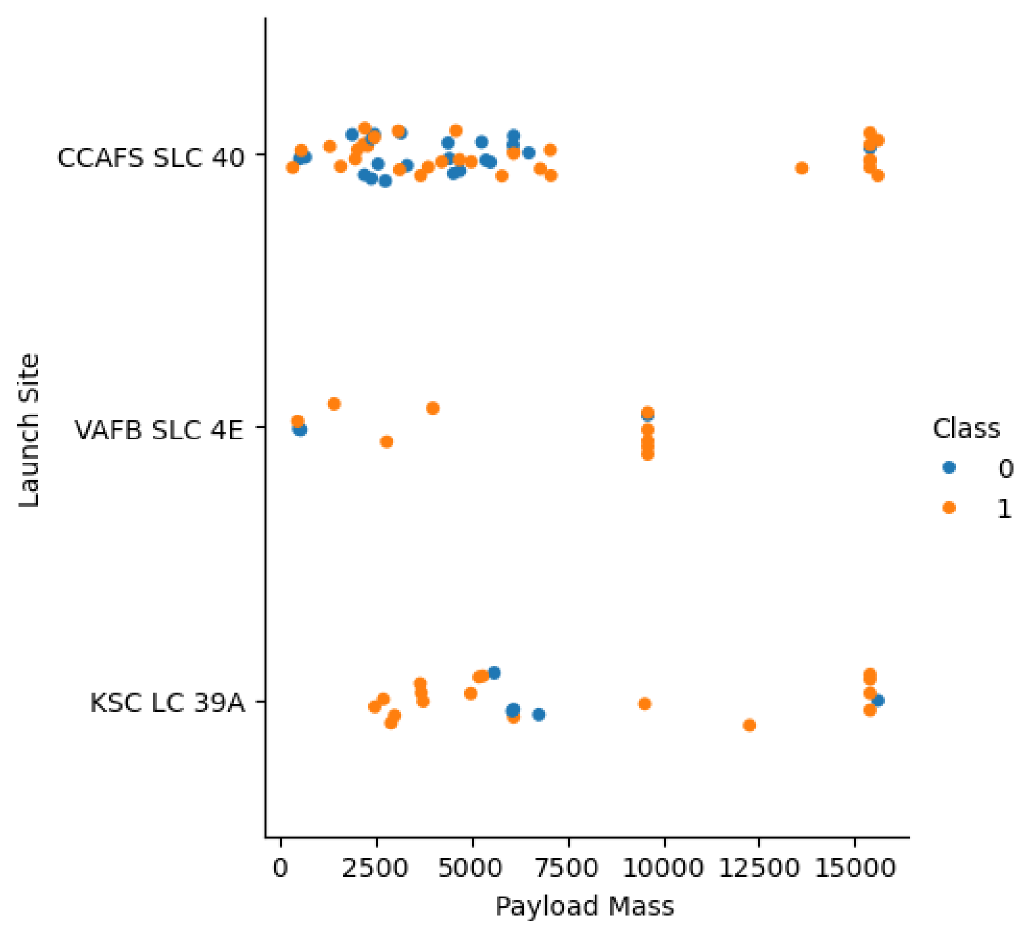
Flight Number vs. Launch Site



- We observe different launch sites have different success rates.
  - CCAFS SCL 40 is successful 60% of the time
  - KSC LC-39A and VAFB SCL 4E are successful approx. 77% of the time.
- CCAFS SCL 40 has more failures on the first flights and more successes on the last flights.

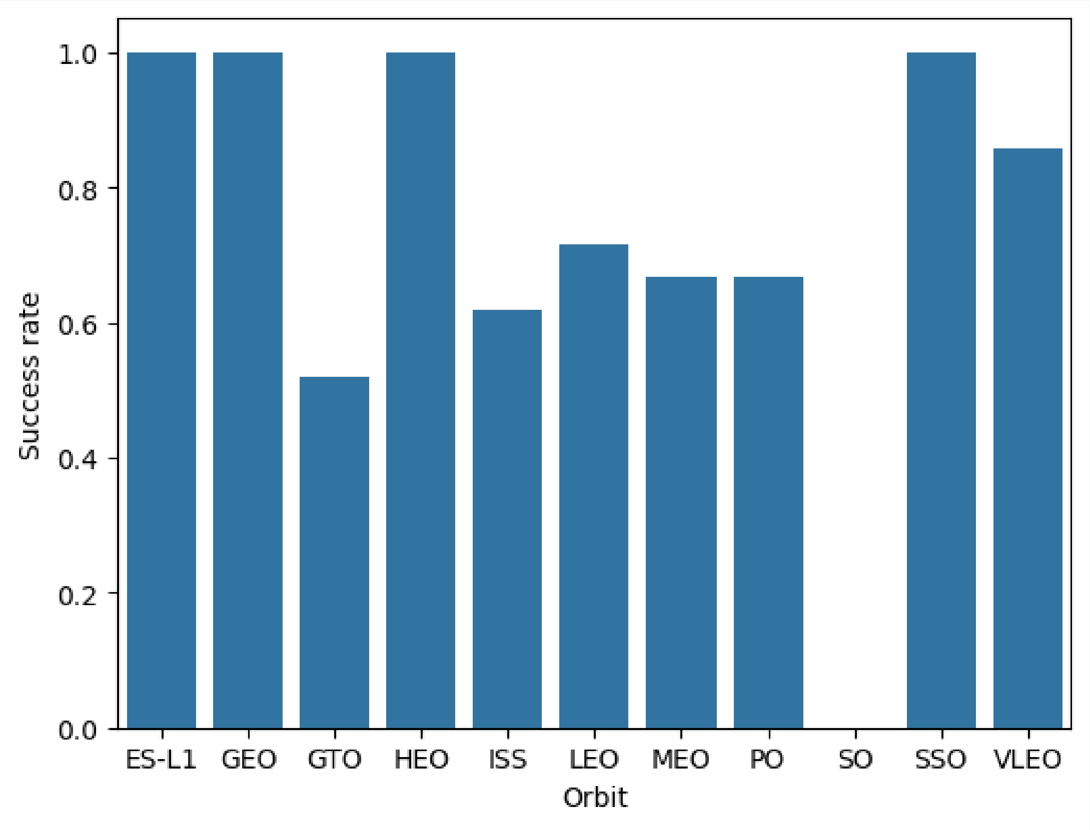
Payload vs. Launch Site

- We observe that there are no rockets launched for heavy payload mass (greater than 1,000 kg) for the VAFB launch site.
- For the CCAFS launch site, a large portion of successful landings come from heavy rockets (15,000 kg)



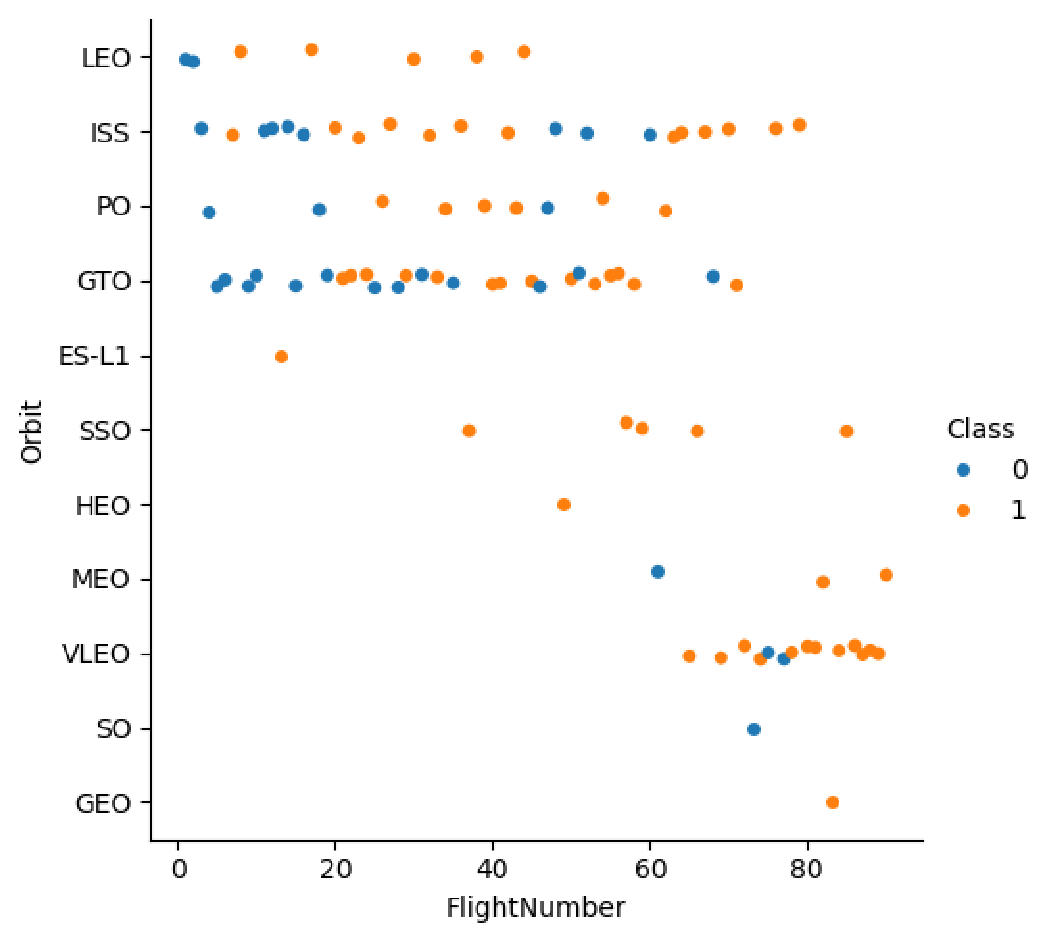
Success Rate vs. Orbit Type

- The target orbits with the highest success rates are: ES-L1, GEO, HEO, and SSO.
- The lowest success rates correspond to: SO, GTO, and ISS.



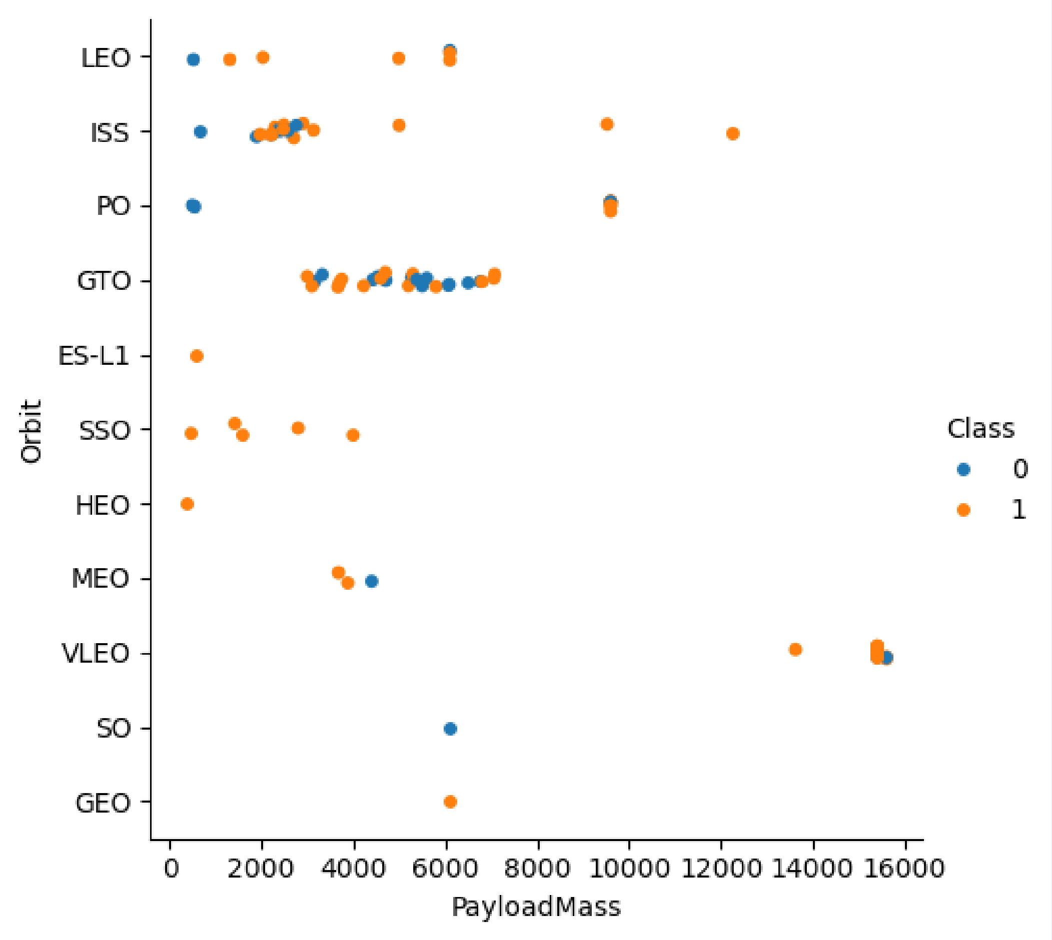
Flight Number vs. Orbit Type

- The success of LEO orbit missions appears to be related to the number of flights.
- There is no clear relationship between the number of flights and the success for GTO orbit missions.
- We have little data for GEO, SO, MEO, HEO, and ES-L1.



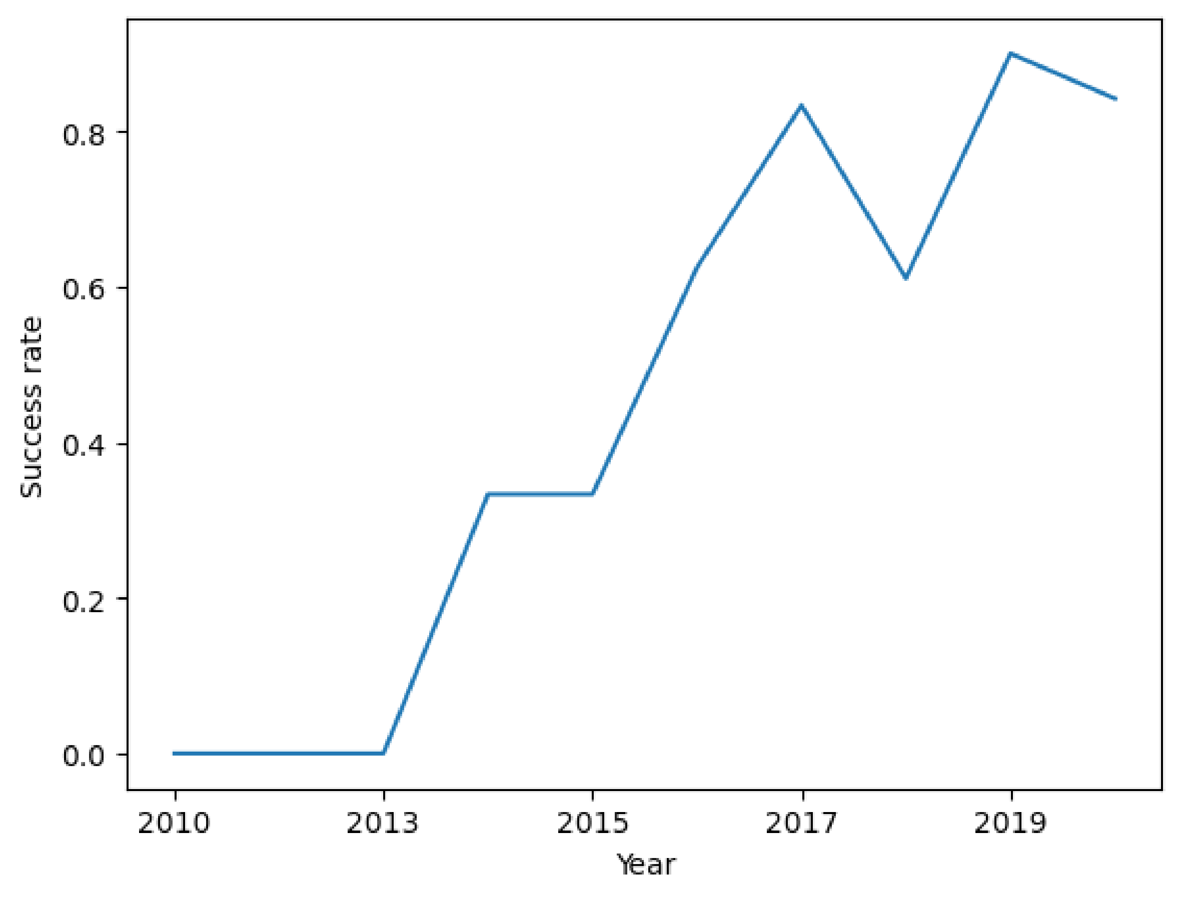
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for PO, LEO and ISS.
- For GTO payload does not seem to be related to the outcome.



Launch Success Yearly Trend

- Starting from 2013, landings for SpaceX launches have been increasingly successful.
- Nowadays, the success rate is above 80%.



All Launch Site Names

- DISTINCT command can be used on the SQL database for finding all Launch Site names.
- In total we have four different launch sites.

In [9]:
```
%sql SELECT DISTINCT("Launch_Site") FROM SPACEXTABLE
```

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

Out[9]:

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

Launch Site Names Begin with 'CCA'

- Launch Sites beginning with CCA correspond to Cape Canaveral. These can be retrieved using the LIKE command within the WHERE statement.
- Using the LIMIT command, we display only the first five.

In [13]:

```
%sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5
```

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

Out[13]:

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

Total Payload Mass

- We can use the built-in SUM function to determine the total amount of payload for a given customer.
- The total payload carried by boosters from NASA is 107,010 kg.

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

In [16]:
```
%sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "Customer" LIKE '%NASA%'
```

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

Out[16]:

SUM("PAYLOAD_MASS__KG_")
107010

Average Payload Mass by F9 v1.1

- We can use the built-in AVG function to compute the mean payload for a given Booster Version.
- The average payload mass carried by booster version F9 v1.1 is 2,928.4 kg.

```
In [18]: %sql SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "Booster_Version"='F9 v1.1'

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

Out[18]:

AVG("PAYLOAD_MASS__KG_")
2928.4

First Successful Ground Landing Date

- Ordering the entries where there was a successful ground landing by date, allows us to determine the date of the first successful landing: 2015-12-22

```
In [24]: %%sql

SELECT "Date" FROM SPACEXTABLE WHERE "Landing_Outcome"='Success (ground pad)' ORDER BY "Date" LIMIT 1;

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

Out[24]:

Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Four Booster Versions with payload mass between 4,000 and 6,000 have been used in successful drone ship landings.
- This result can be found based on the SQL query on the right.

```sql
SELECT DISTINCT("Booster_Version") FROM
SPACEXTABLE WHERE
"Landing_Outcome"='Success (drone ship)'
AND PAYLOAD_MASS__KG_ BETWEEN 4001 AND
5999;
```

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

Total Number of Successful and Failure Mission Outcomes

- By grouping the entries by mission outcome and counting the number of entries in each group, we observe that the table contains 100 successful outcomes vs. 1 failed outcome.

In [33]:

```
%%sql
SELECT "Mission_Outcome", COUNT("Mission_Outcome") FROM SPACEXTABLE GROUP BY "Mission_Outcome"
```

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

Out[33]:

Mission_Outcome	COUNT("Mission_Outcome")
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Twelve different booster versions have been used with the maximum payload mass.
- The SQL query on the right allows us to list all of them.
- The boosters used with maximum payload mass are variants of the B5 booster.

```
SELECT DISTINCT("Booster_Version") FROM
SPACEXTABLE WHERE "PAYLOAD_MASS__KG_"=
(SELECT MAX("PAYLOAD_MASS__KG_") FROM
SPACEXTABLE)
```

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

2015 Launch Records

- We observed seven launches during 2015, all of which were launched from CCAFS LC-40.
- From these, two landings failed in drone ship.

```
SELECT
        "Landing_Outcome",
        "Booster_Version",
        "Launch_Site",
        SUBSTR("Date",6,2) AS MONTH,
        SUBSTR("Date",0,5) AS YEAR
FROM SPACEXTABLE WHERE YEAR='2015';
```

Landing_Outcome	Booster_Version	Launch_Site	MONTH	YEAR
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	01	2015
Controlled (ocean)	F9 v1.1 B1013	CCAFS LC-40	02	2015
No attempt	F9 v1.1 B1014	CCAFS LC-40	03	2015
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	04	2015
No attempt	F9 v1.1 B1016	CCAFS LC-40	04	2015
Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40	06	2015
Success (ground pad)	F9 FT B1019	CCAFS LC-40	12	2015

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

- Most launches between 2010-06-04 and 2017-03-20, did not attempt to land.
- The second most common outcomes were successful and failed drone ship landings, each with an occurrence of five.

```sql
SELECT
        "Landing_Outcome",
        COUNT(*) AS CNT
FROM SPACEXTABLE
WHERE DATE("Date") BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY "Landing_Outcome"
ORDER BY CNT DESC;
```

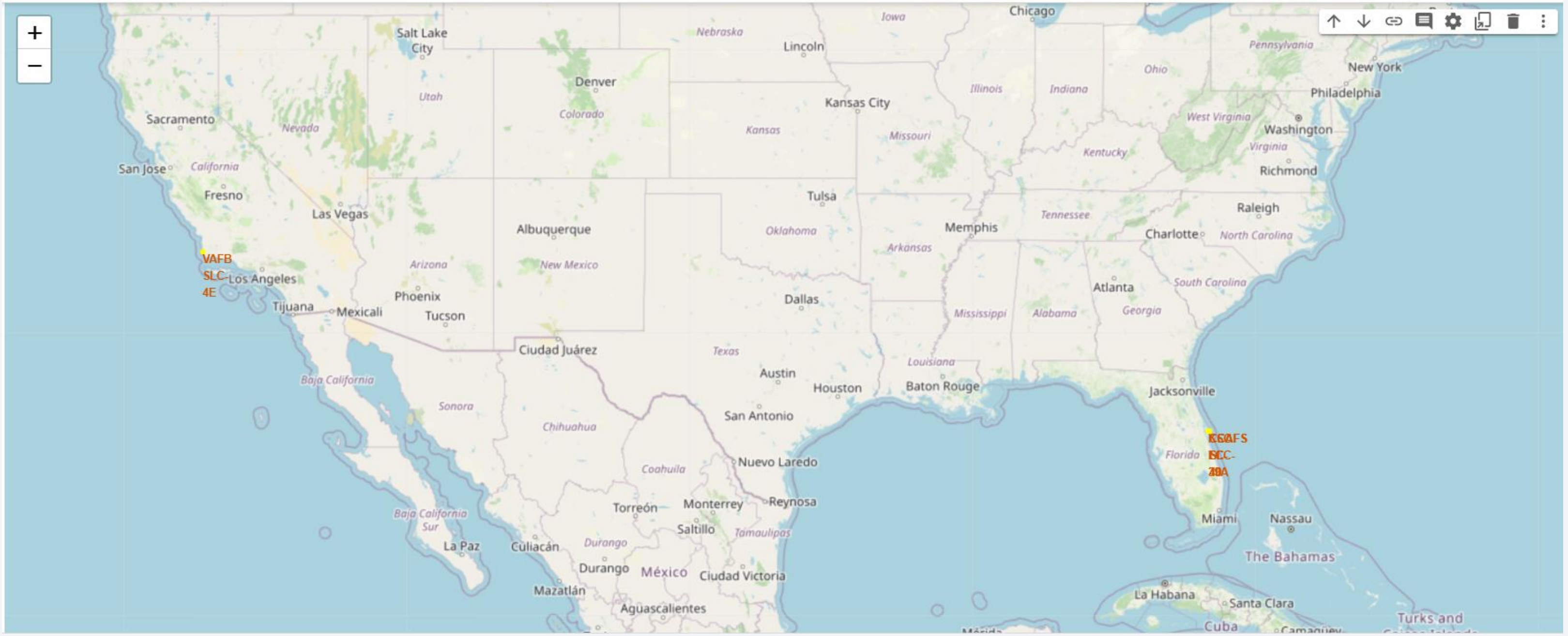
Landing_Outcome	CNT
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

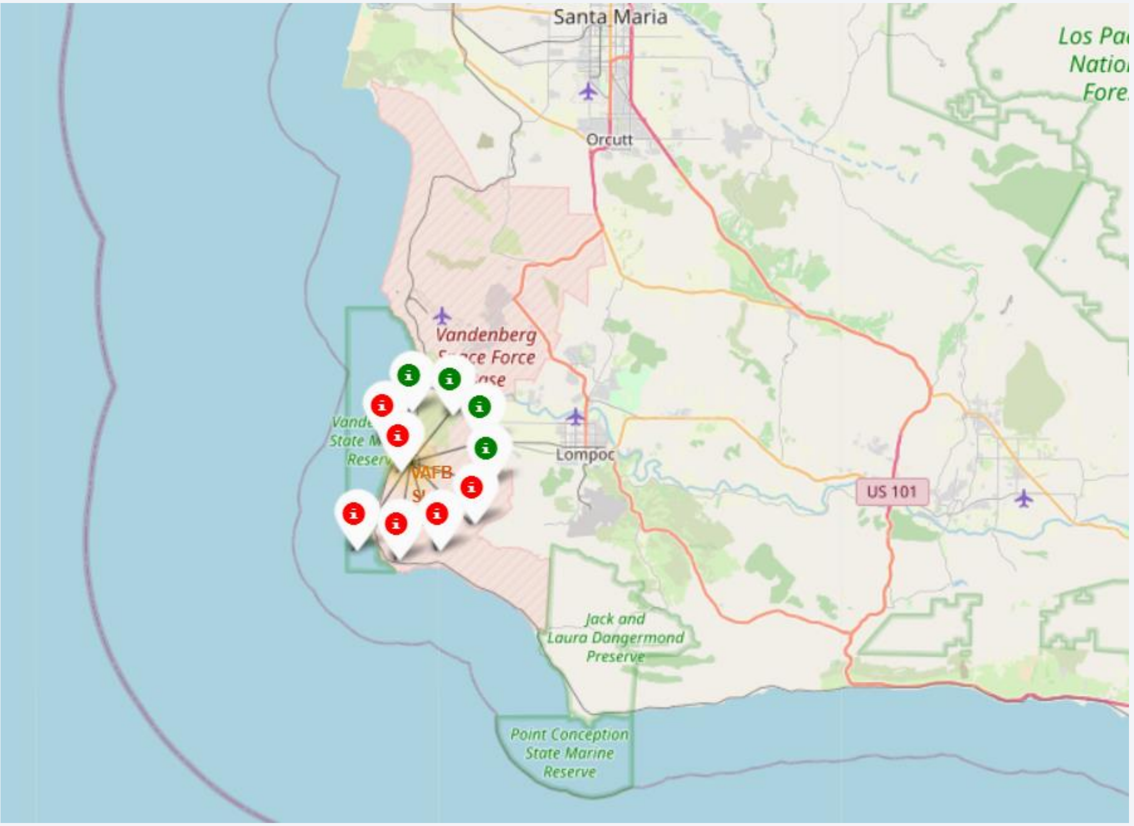
SpaceX Launch Site Locations

- SpaceX has launch sites both in the East and West coasts of the United States.
- All the launch sites are close to the coastline as shown on the right.



VAFB SCL-4E Launch Outcomes

- Folium allows us to add color markers to visualize the launch outcomes of a particular site within their location.
- For example, the picture shows the launch outcomes for the VAFB SCL-4E site. We observe only 4/10 launches were successful.



CCAFS SLC-40 proximities

- CCAFS SCL-40 is approx.
  - 0.59 km away from the Samuel Phillips Parkway.
  - 0.87 km away from the coastline.
  - 1.30 km away from the NASA Railroad.
- The information can be conveniently displayed within the map using markers.

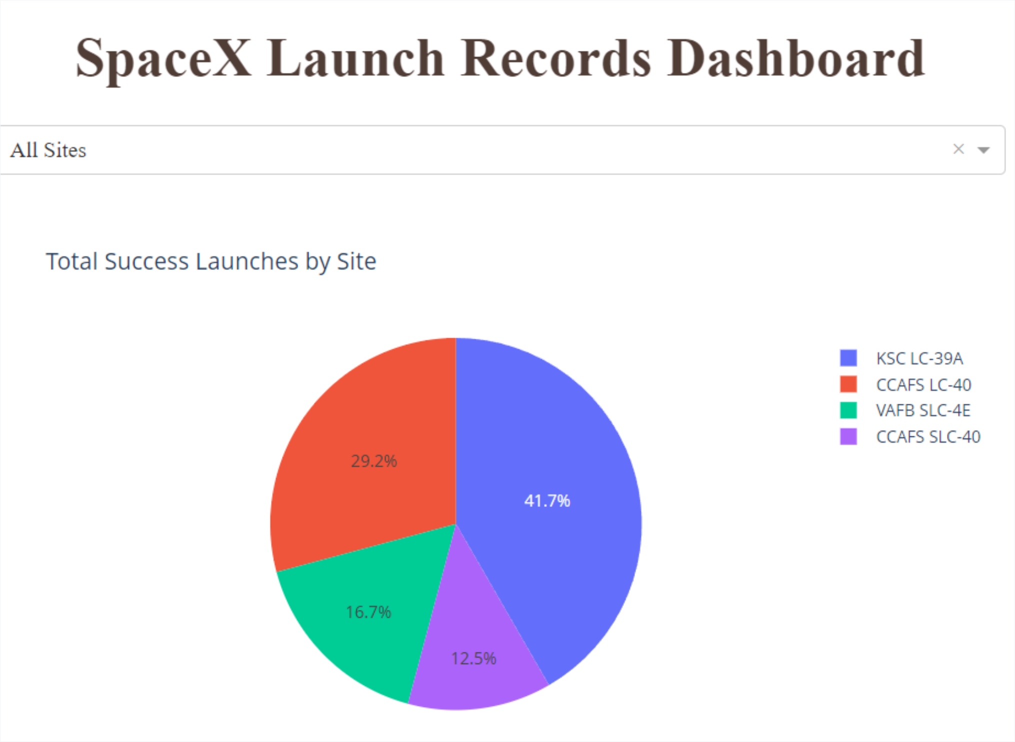
Section 4

Build a Dashboard with Plotly Dash

Success Launches by Site

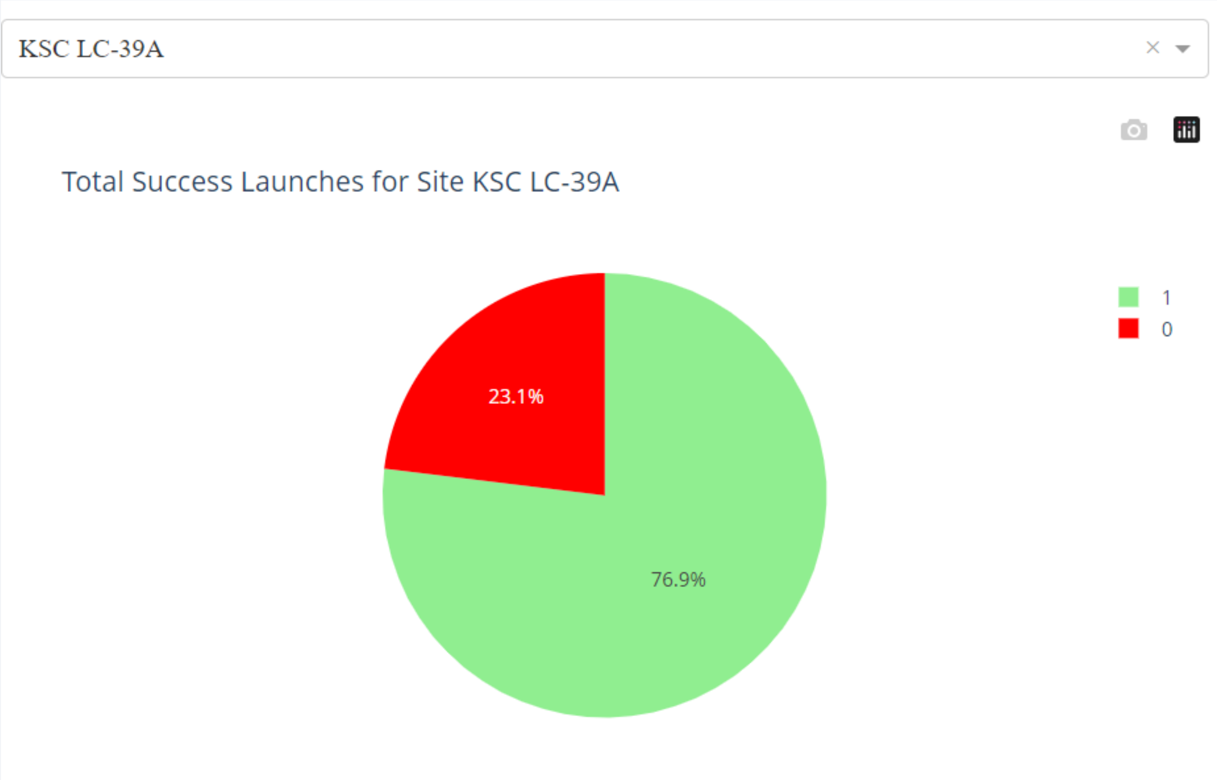
- We can see the proportion of successful launches by Launch Site using a Pie Chart like the one used in the interactive dashboard on the right.

- Most successful launches come from KSC LC-39A



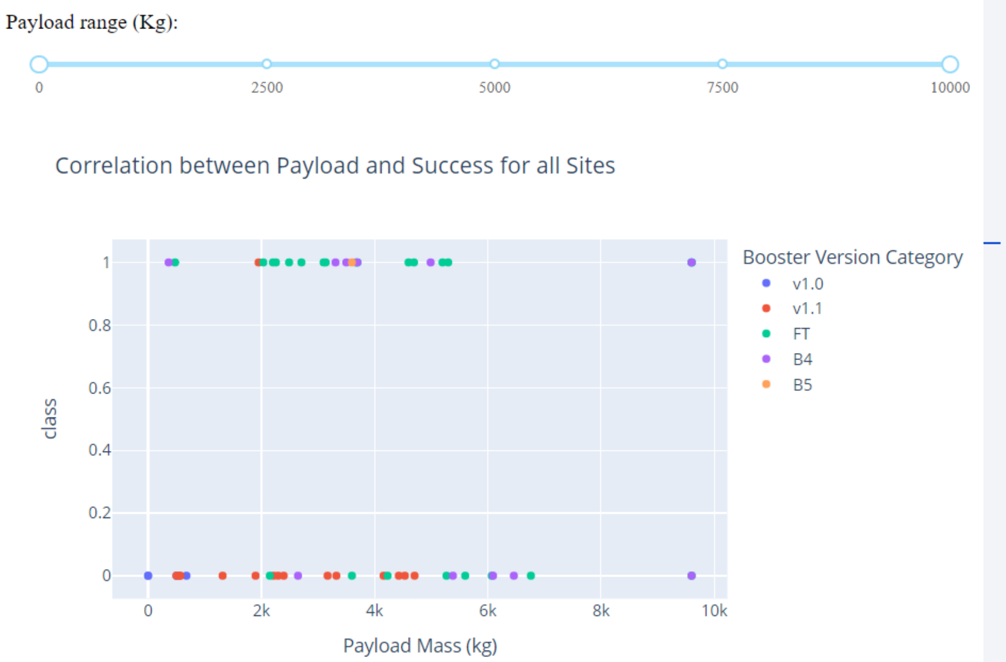
Launch Site with Highest Success Ratio

- Interactive exploration of the success rate of different launch sites allowed us to find that KSC LC-39A is also the launch site with the highest success ratio.
- The success ratio for KSC LC-39A is approx. 77%.



Payload and Success Rate

- The top image shows the launch outcome for different payloads and booster versions across all the payload range.
- The bottom image shows focuses on payloads between 3,100 and 3,700 kg where we observe the largest success rate (70%).
    - Within this range we observe that B4 boosters are the most common for successful outcomes.

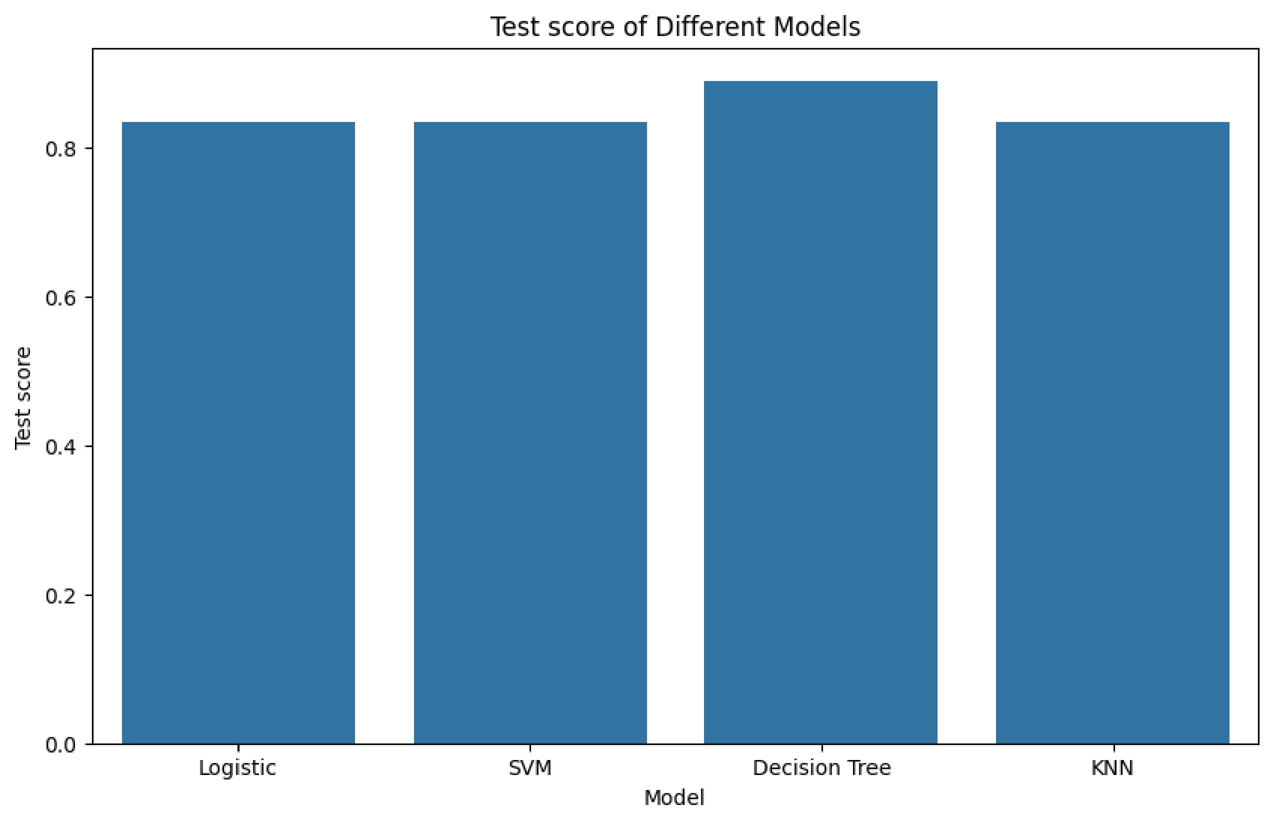


Section 5

Predictive Analysis (Classification)

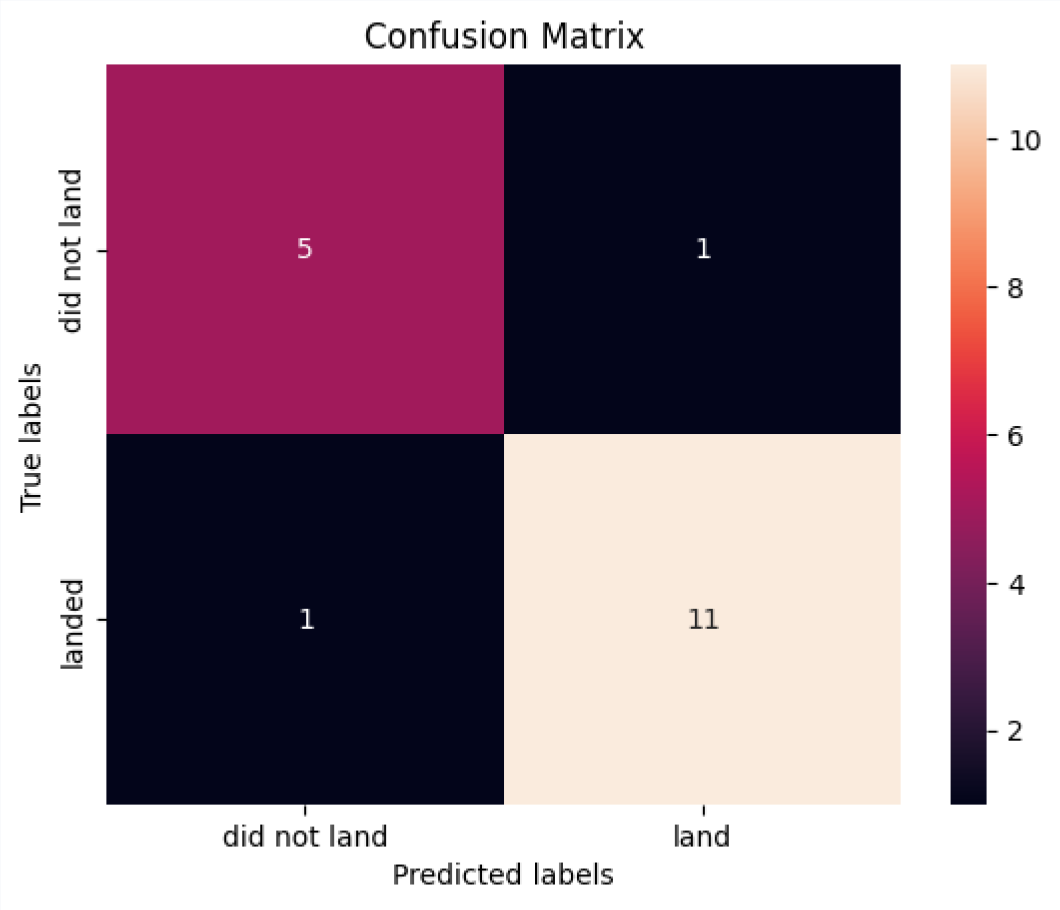
Classification Accuracy

- Among the four different classifiers, the Decision Tree classifier showed the best performance on the testing set with an accuracy of 88%.



Confusion Matrix

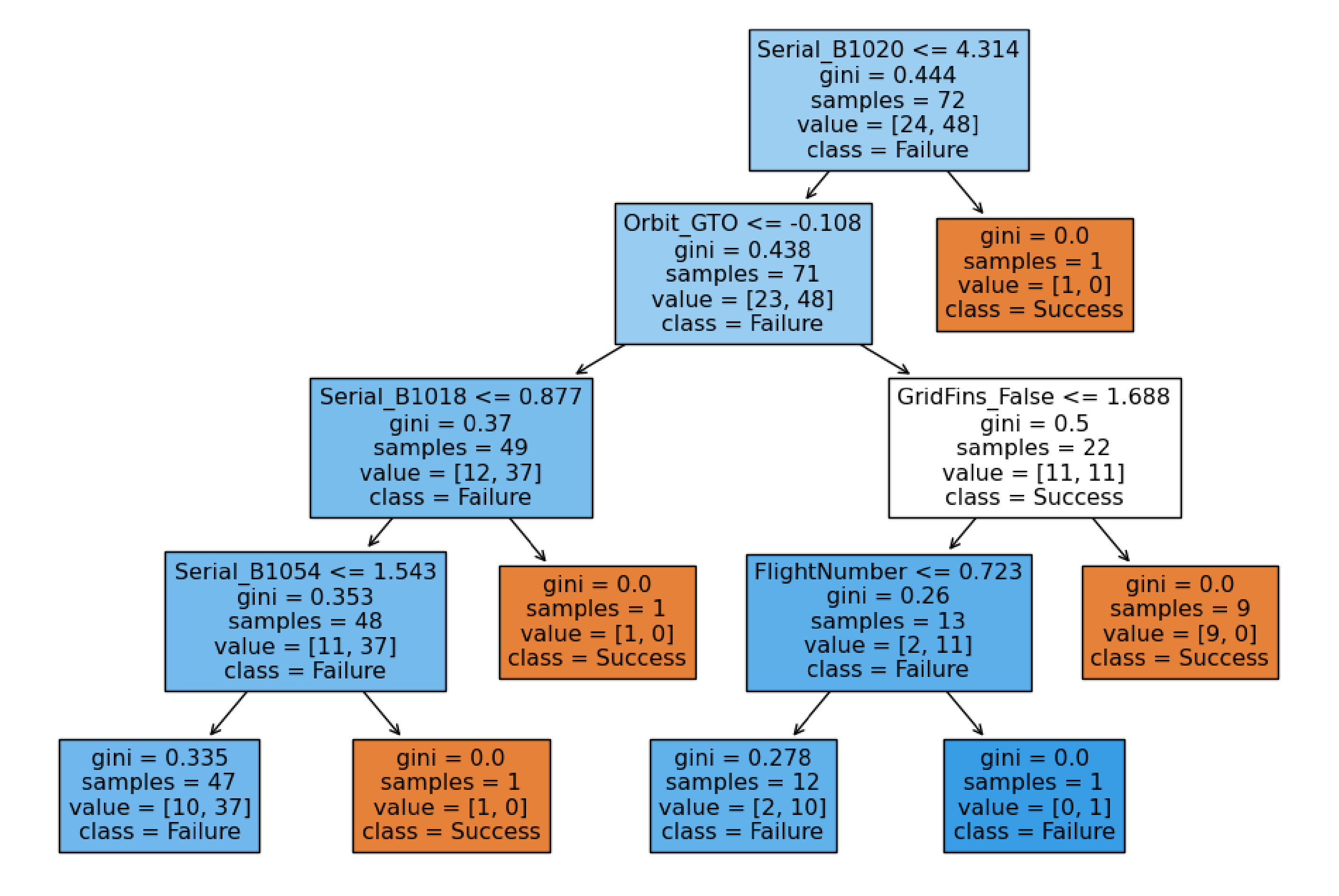
- The confusion matrix for the Decision Tree shows that the model makes only two mistakes for the 18 samples in the testing data.
- One is a False Positive and one is a False Negative.



Conclusions

- EDA revealed that factors such as Flight Number, Orbit Type, Payload, and Launch Site may be good predictors for the landing outcome.
- A Decision Tree was developed to predict the landing outcome achieving an 80%+ accuracy on the testing set.
  - Booster Type, Orbit Type, Flight Number and Grid Fins are the main variables used by the model.
- Determining if a launch will land can help us determine the cost of a launch. This is helpful to bid against SpaceX and to help set SpaceY dynamic prices.

Appendix: Final Decision Tree Model



Thank you!