Data Science Capstone Project

Arya Kadakia

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion and Insights

Executive Summary

Methodologies

- Data Collection using web scraping and SpaceX API
- Exploratory Data Analysis (EDA), data wrangling, data visualization and interactive visual analytics
- Predictive analysis using machine learning

Summary

- It was possible to collected valuable data from public sources
- EDA allowed to identify which features are the best to predict success of launchings
- Machine learning helped us determine the best model to predict successful launches.

Introduction

SpaceX is the most successful company of the commercial space age, making space travel affordable. The company advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.

The objective is to evaluate the viability of a new company SpaceY to compete with SpaceX. They want to know the variables that are involved in a successful launch.

Research Questions:

- How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
- Does the rate of successful landings increase over the years?
- What is the best algorithm that can be used to predict successful launches?

Methodology

Methodology

1. Data collection
2. Data wrangling
3. Exploratory data analysis (EDA) using visualization and SQL
4. Interactive visual analytics using Folium and Plotly Dash
5. Predictive analysis using classification models

Data Collection: SpaceX API

SpaceX offers a public API from where data can be obtained for analysis (https://api.spacexdata.com/v4/rockets/)

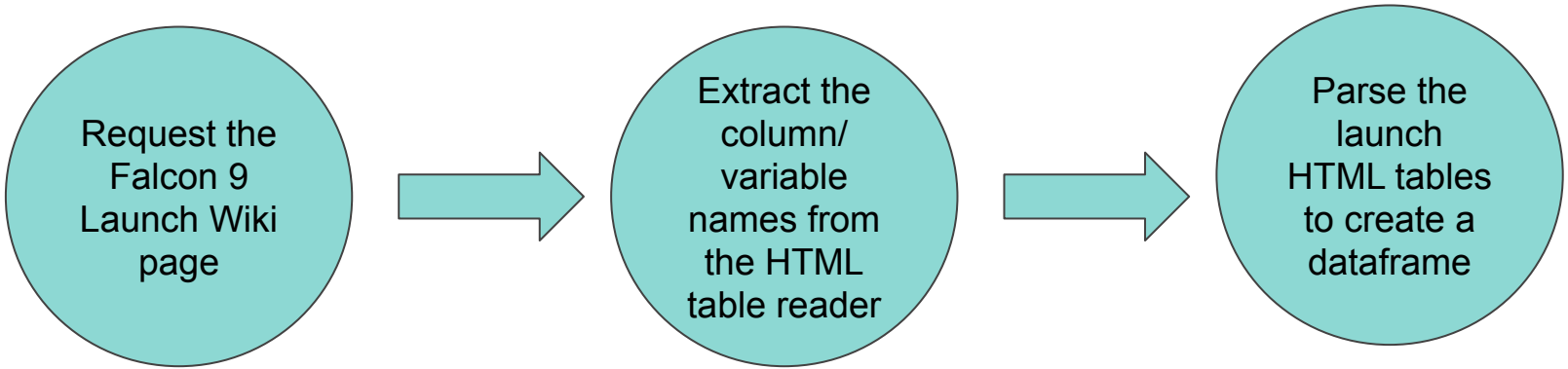
Data: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

Code: https://github.com/aryakadakia/datasciencecourse/blob/capstone/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection: Web Scraping

Data from SpaceX launches can also be obtained from Wikipedia.
(https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)

Data: Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time



Code: https://github.com/aryakadakia/datasciencecourse/blob/capstone/jupyter-labs-webscraping.ipynb

Data Wrangling

1. Exploratory Data Analysis (EDA) was performed to find patterns in the dataset.
2. Calculated number of launches per site, occurrences of each orbit, and occurrences of mission outcome per orbit type.
3. Landing Outcome label was created from Outcome column (1 = successfully landed, 0 = unsuccessful).

Code: https://github.com/aryakadakia/datasciencecourse/blob/capstone/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

Scatter, bar, and line plots were used to visualize the relationship between -

1. Launch Site X Flight Number
2. Launch Site X Payload Mass
3. Orbit type X Flight Number
4. Payload Mass X Orbit type
5. Success rate of each orbit
6. Launch success yearly trend

Code: https://github.com/aryakadakia/datasciencecourse/blob/capstone/edadataviz.ipynb

EDA with SQL

SQL Queries:

- Names of the unique launch sites in the space mission
- Top 5 launch sites whose name begin with the string 'CCA’
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date when the first successful landing outcome in ground pad was achieved
- Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg
- Total number of successful and failure mission outcomes
- Names of the booster versions which have carried the maximum payload mass
- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20

Code: https://github.com/aryakadakia/datasciencecourse/blob/capstone/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Folium Interactive Map

Features

1. Markers of each launch site
    a. Circles: highlighted areas around specific coordinates (e.g., NASA Johnson Space Center)
2. Colored markers of launch outcomes
    a. Marker clusters: success (Green) and failed (Red) launches were identified to determine launch sites with relatively high success rates.
3. Lines: distances between two coordinates
    a. e.g., between Launch Site KSC LC-39A and Railway, Highway, Coastline and Closest City

Code: https://github.com/aryakadakia/datasciencecourse/blob/capstone/lab_jupyter_launch_site_location.ipynb

Plotly Dashboard

Features

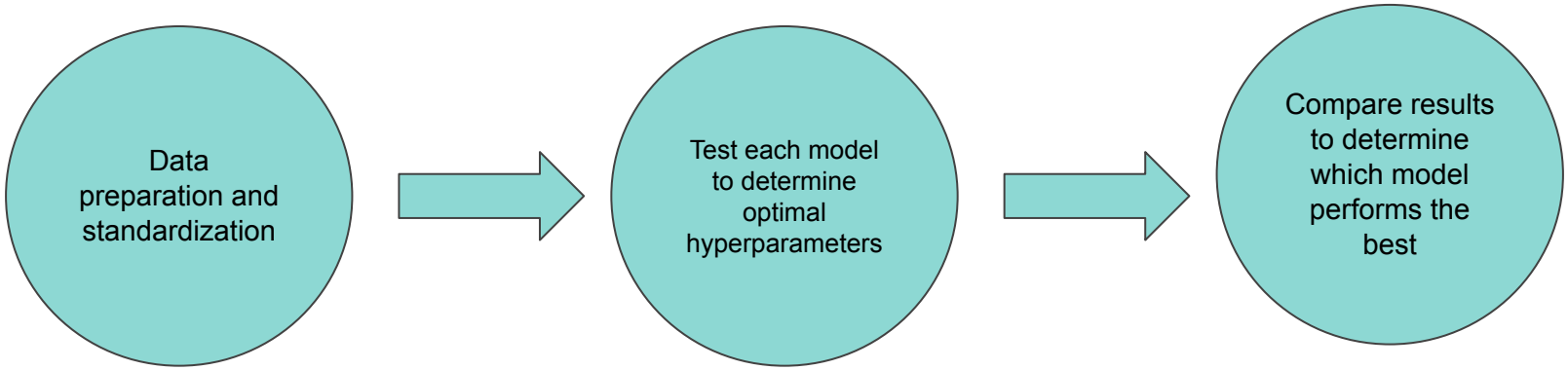
1. Dropdown list to enable Launch Site selection.
2. Pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the selected site.
3. Slider to select Payload range.
4. Scatter chart helped determine correlation between payload and launch success.

Code: https://github.com/aryakadakia/datasciencecourse/blob/capstone/spacex_dash_app.py

Predictive Analysis (Classification)

The following classification models were compared and evaluated -

1. Logistic regression
2. Support vector machine (SVM)
3. Decision tree
4. k-nearest neighbors (KNN)

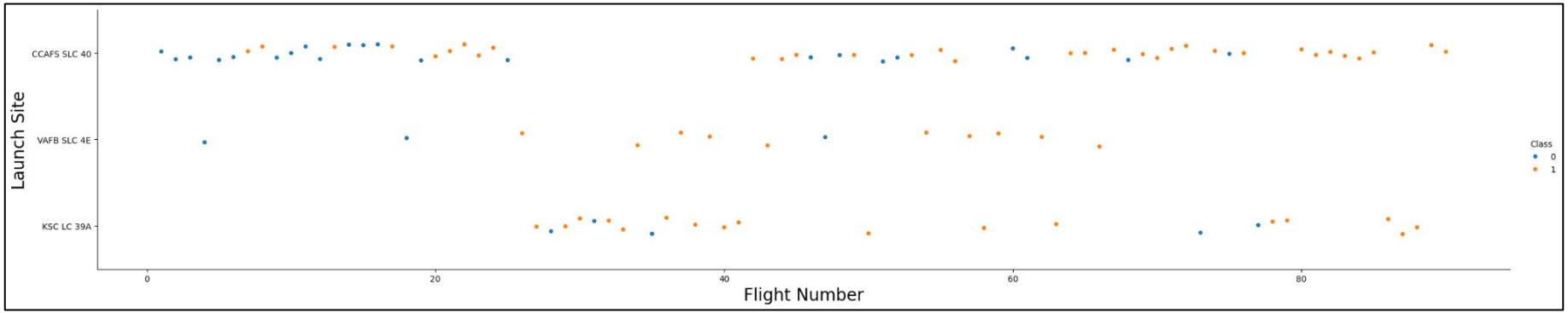


Code: https://github.com/aryakadakia/datasciencecourse/blob/capstone/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

EDA with Visualization

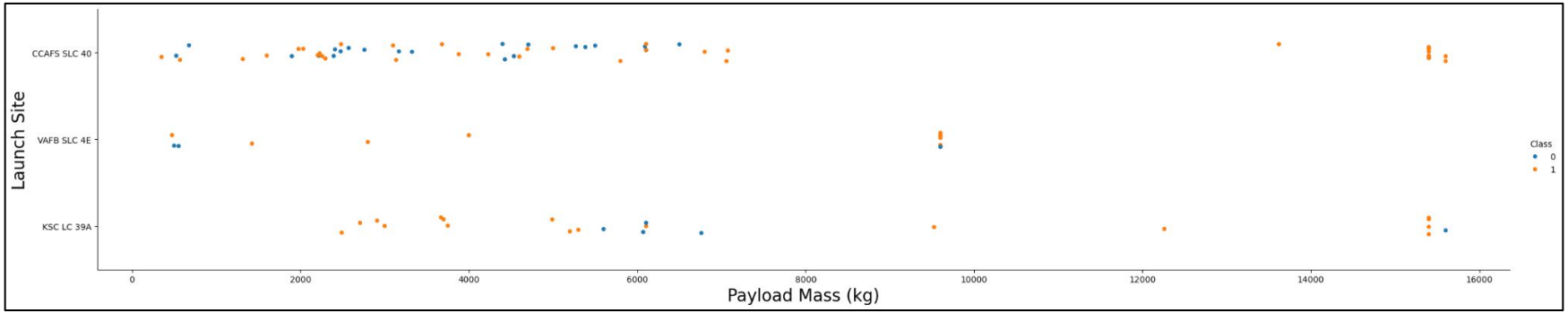
Flight Number vs. Launch Site



Success rate improves over time.

The best launch site seems to be CCAF5 SLC 40, where most of recent launches took place and were successful.

Payload Mass vs. Launch Site



Payloads over 9,000kg have an excellent success rate.

Payloads over 12,000 kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.

KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

Success rate vs. Orbit type

Orbits with 100% success rate:

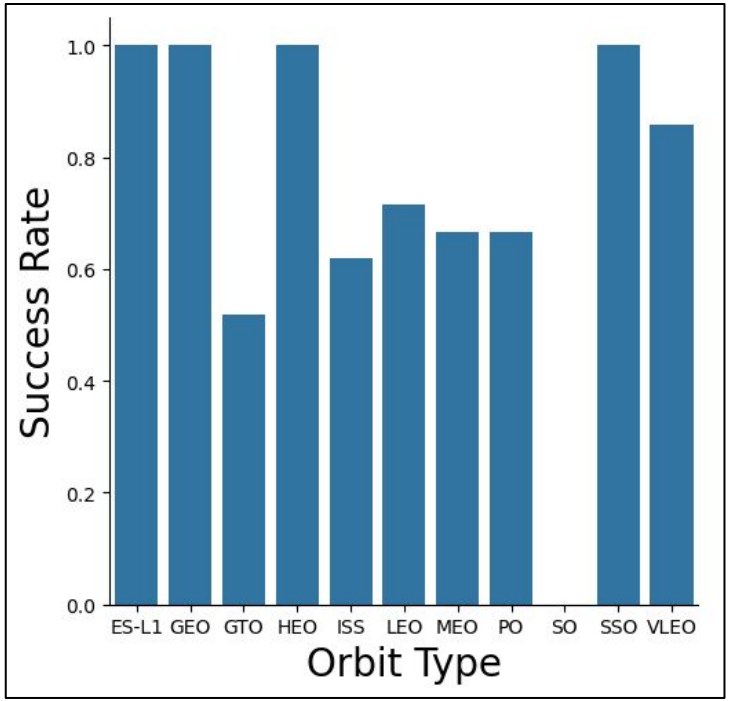
- ES-L1, GEO, HEO, SSO

Orbits with success rate between 50% and 85%:

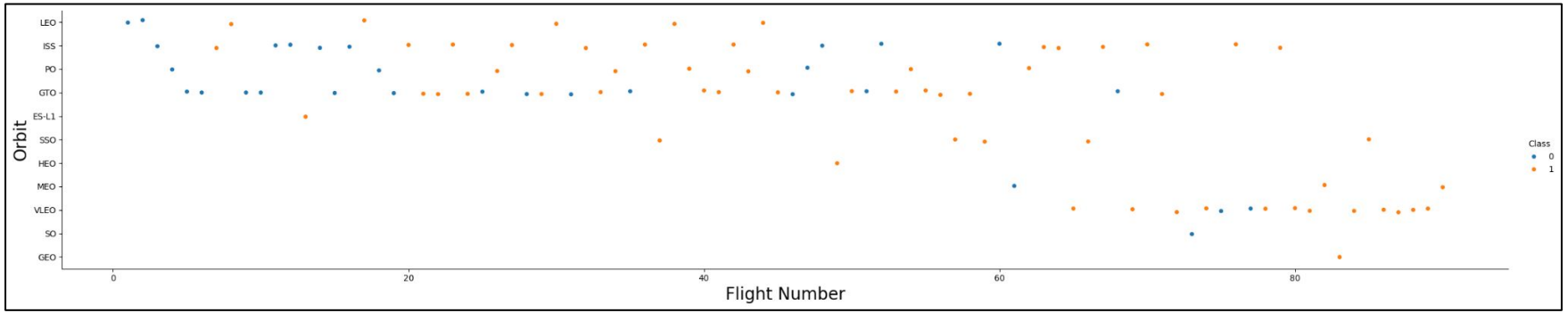
- GTO, ISS, LEO, MEO, PO, VLEO

Orbits with 0% success rate:

- SO



Flight Number vs. Orbit type

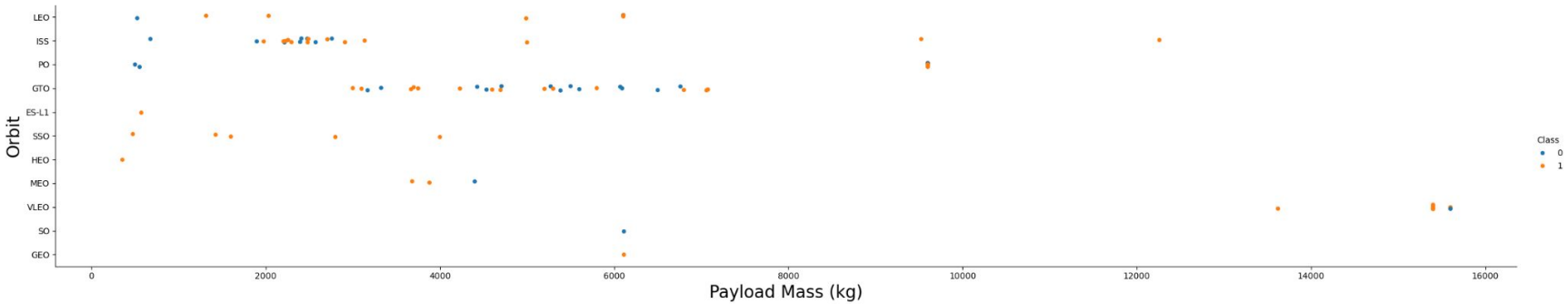


Success rate improved over time to all orbits.

VLEO orbit seems to be relatively new due to increase frequency towards the end.

LEO orbit has been consistently successful after its initial few launches.

Payload Mass vs. Orbit type



There seems to be no relation between payload and success rate for orbit GTO.

ISS orbit has the widest range of payloads.

SO and GEO orbits have limited launches

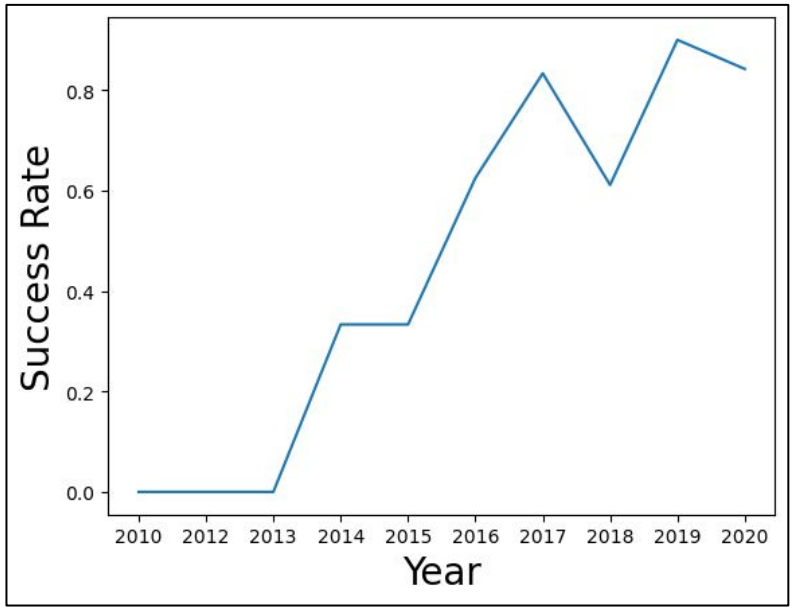
VLEO is the only orbit with payload mass > 13,000 kg.

Launch success yearly trend

Success rate started increasing in 2013 and kept until 2020;

It’s possible that the first three years involved a period of adjustments and technology advancements that hindered successful launches at that time.

There was a sharp drop in 2017 by almost 20%. It’s possible there were some shortages or financial restrictions that interfered with the launches.



EDA with SQL

All launch site names

```
%sql select distinct launch_site from SPACEXTABLE;
```

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

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

Names of the unique launch sites in the space mission

Launch site names begin with `CCA`

Display 5 records where launch sites begin with the string 'CCA'

```
%sql select * from SPACEXTABLE where launch_site like 'CCA%' limit 5;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (p
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 (p
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	N
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	N
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	N

5 records where launch sites begin with the string 'CCA'

Total payload mass

```
%sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXTABLE where customer = 'NASA (CRS)';
```

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

total_payload_mass
45596

Total payload mass carried by boosters launched by NASA (CRS)

Average payload mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1

```
%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXTABLE where booster_version like '%F9 v1.1%
```

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

average_payload_mass
2534.6666666666665

Average payload mass carried by booster version F9 v1.1

First successful ground landing date

```
%sql select min(date) as first_successful_landing from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)'
```

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

first_successful_landing
2015-12-22

Date when the first successful landing outcome in ground pad was achieved

Successful drone ship landing with payload between 4000 and 6000

```
%sql select booster_version from SPACEXTABLE where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_
```

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

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

Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

Total number of successful and failure mission outcomes

List the total number of successful and failure mission outcomes

```
%sql select mission_outcome, count(*) as total_number from SPACEXTABLE group by mission_outcome;
```

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

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

Total number of successful and failure mission outcomes

Boosters carried maximum payload

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

```
%sql select booster_version from SPACEXTABLE where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACE
```

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

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

Names of the booster versions which have carried the maximum payload mass

2015 launch records

```
%%sql SELECT
    SUBSTR(date, 6, 2) AS month,
    date,
    booster_version,
    launch_site,
    Landing_Outcome
FROM SPACEXTABLE
WHERE
    Landing_Outcome = 'Failure (drone ship)'
    AND SUBSTR(date, 1, 4) = '2015';
```

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

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015

Rank success count between 2010-06-04 and 2017-03-20

```
%%sql select Landing_Outcome, count(*) as count_outcomes from SPACEXTABLE
      where date between '2010-06-04' and '2017-03-20'
      group by Landing_Outcome
      order by count_outcomes desc;
```

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

Landing_Outcome	count_outcomes
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

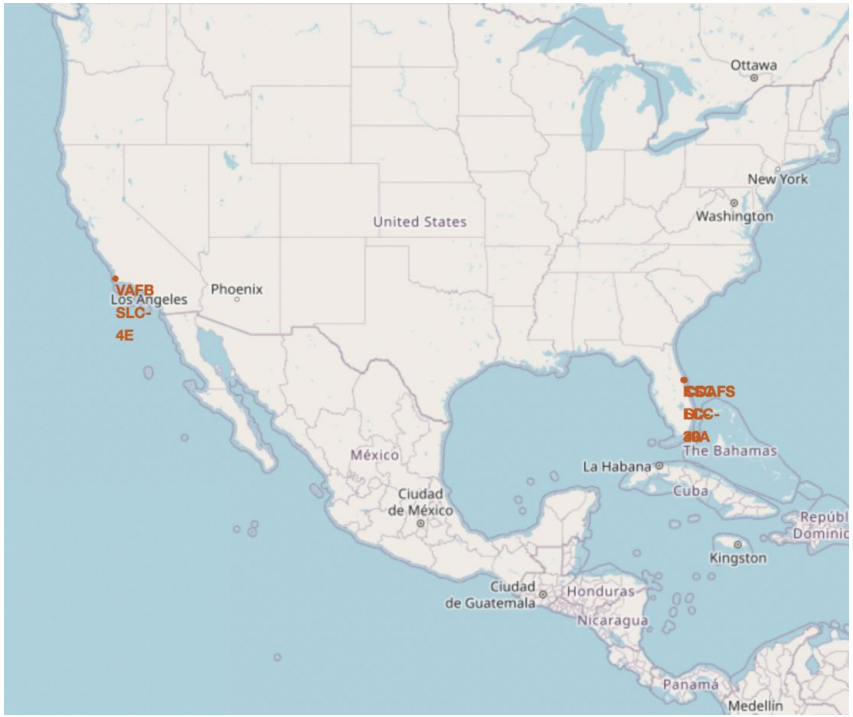
Count of landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order

Folium Interactive Map

All launch sites

All launch sites are in very close proximity to the coast, minimizing the risk of falling debris injuring civilians or destroying property.

Most of Launch sites are in proximity to the Equator. This helps with launches as there is less gravitational attraction at the Equator. The surface of the earth is also traveling faster there, allowing spacecrafts to make use of inertia during launch.

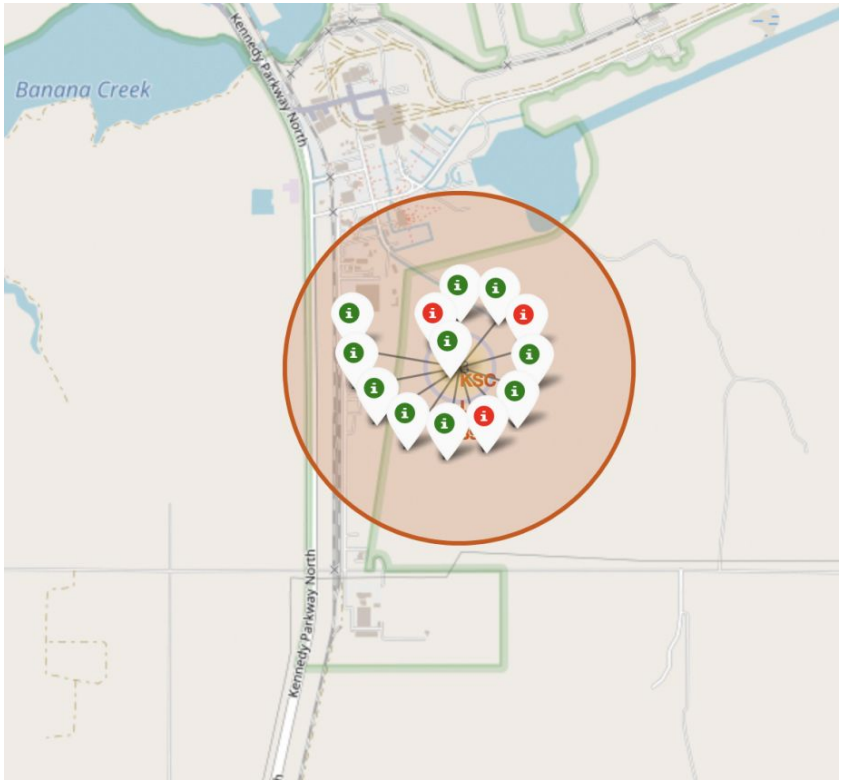


Site-level launch records

Green = Successful Launch

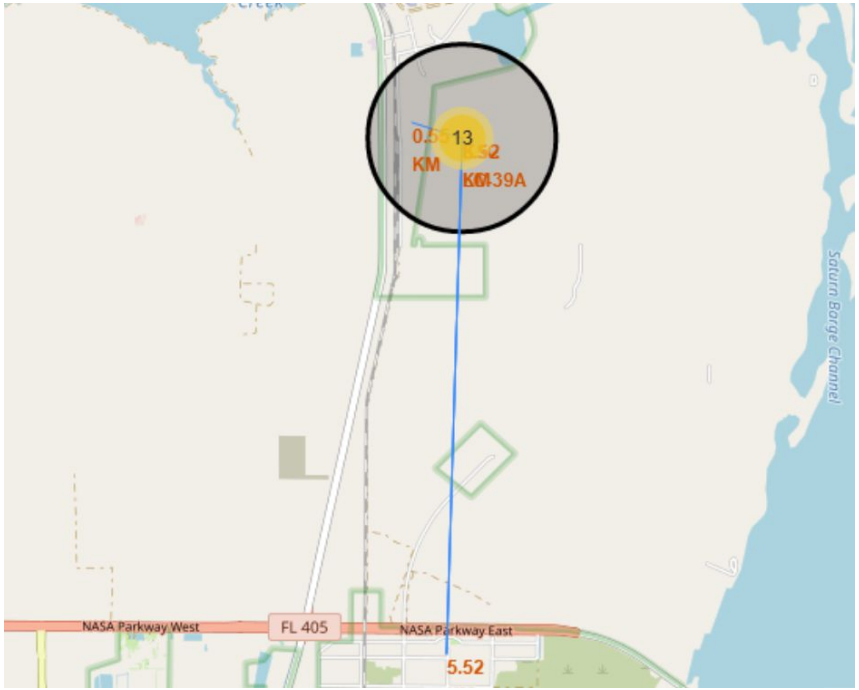
Red = Failed Launch

Launch Site KSC LC-39A has a very high success rate.



Distance from the launch site KSC LC-39A to its proximities

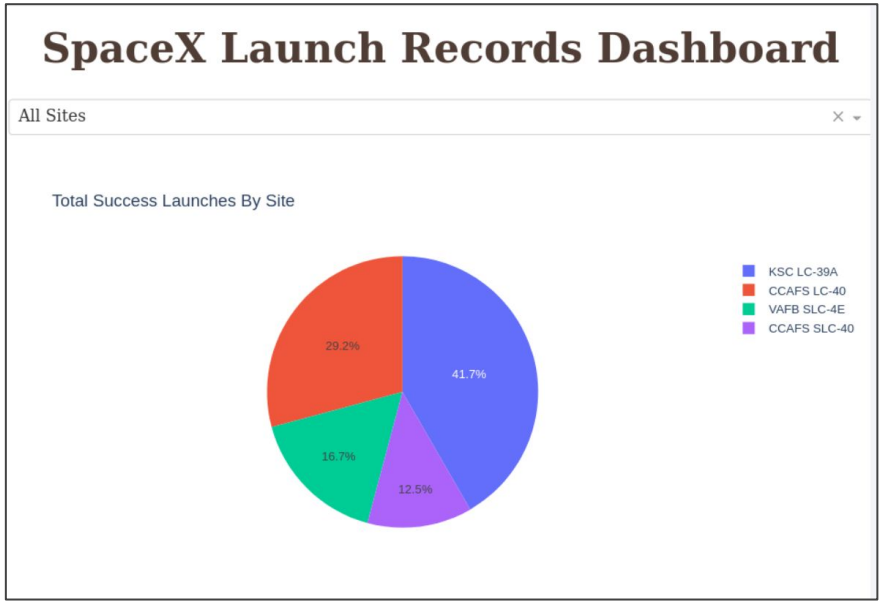
KSC LC-39A is relatively close to railway, which makes it more efficient to move equipment and supplies in and out.



Plotly Dashboard

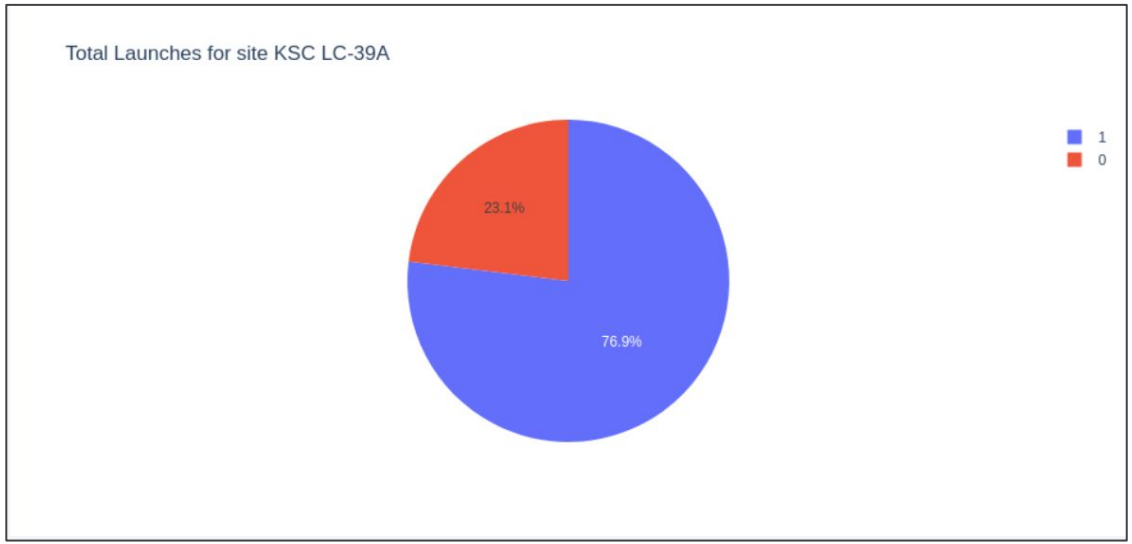
Launch success count for all sites

Launch site seems to hugely impact the success of missions.



Launch site with highest launch success ratio

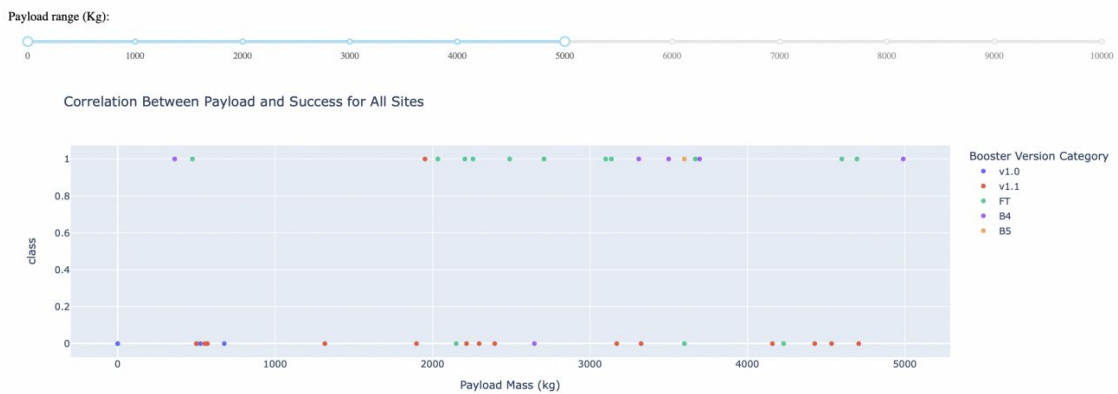
76.9% of launches are successful in KSC LC-39A (10 successful, 3 failed landings)

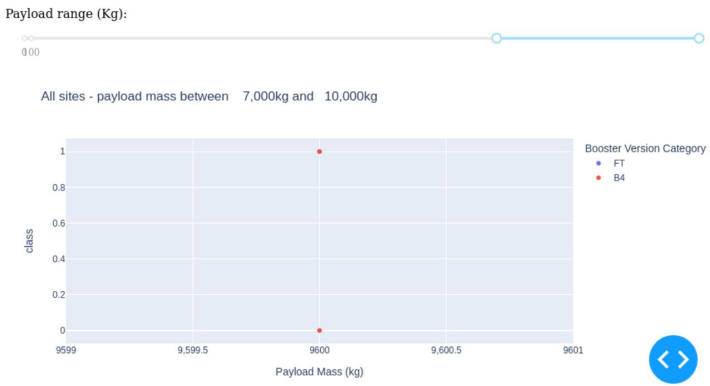


Payload Mass vs. Launch Outcome

Payloads under 5,500 kg and FT boosters are the most successful combination.

There’s not enough data to estimate risk of launches over 7,000 kg.





Predictive Analysis (Classification)

Classification Accuracy

Four classification models were tested. The model with the highest classification accuracy was the Decision Tree classifier, which had an accuracy of 87%.

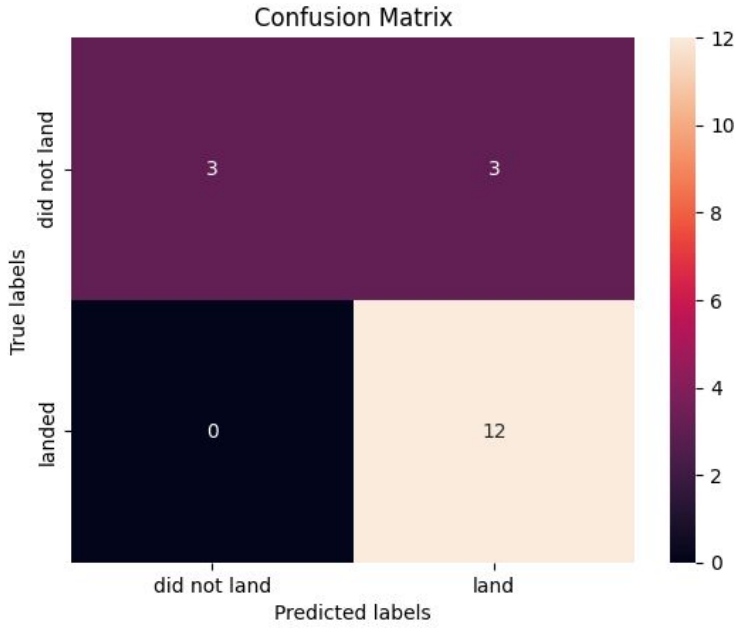
Similar Test Accuracy may be due to the small sample size (18 samples).

Model	Accuracy	TestAccuracy
LogReg	0.84643	0.83333
SVM	0.84821	0.83333
Tree	0.87679	0.77778
KNN	0.84821	0.83333

Confusion Matrix

Examining the confusion matrix for each model, we can see there is a persistent issue with false positives.

The models seemed to incorrectly classify some failed landings as successful ones.



Conclusion

Conclusion

- The Decision Tree classifier is the best algorithm to determine successful launches from this dataset.
- Launches with a low payload mass show better results than launches with a larger payload mass.
- Most of launch sites are in proximity to the Equator and are in close proximity to the coast.
- While most mission outcomes are successful, the success rate of launches seems to be increasing over time. This could be due to advancements in rocket technology, software, and safety protocols.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.
- Using this information, SpaceY can now determine the optimal strategy for a successful launch and first stage landing.