Winning Space Race with Data Science

Doron Fingold
August 2024

IBM Developer
SKILLS NETWORK

Outline

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

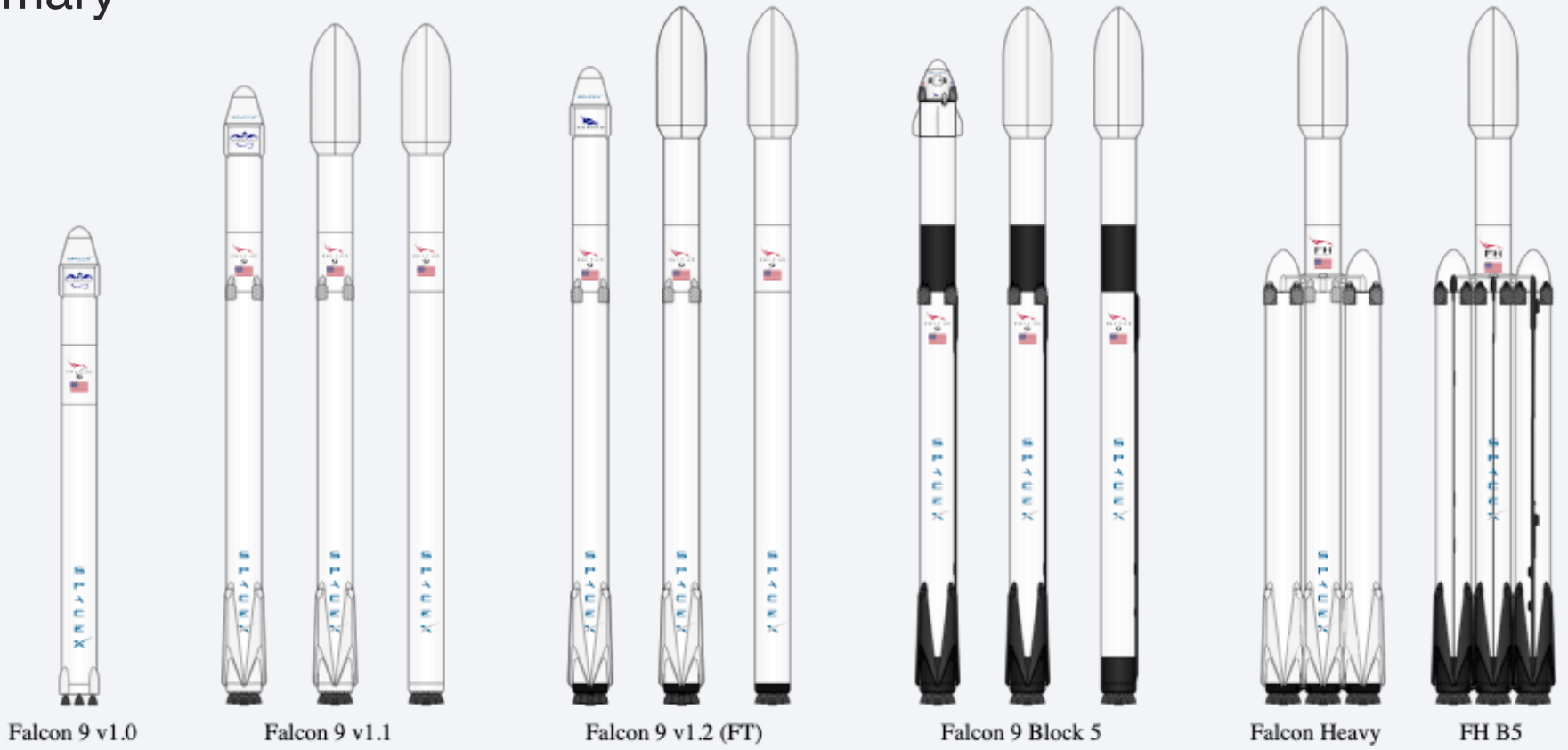


Image by Lucabon (based on work of Markus Säynevirta and Craigboy and Rressi)

Executive Summary

- SpaceX 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.
- We want to identify what factors contribute to successful landing.
- Launch records were obtained from SpaceX API as well as Wikipedia.
- It was transformed for Exploratory Data Analysis (EDA).
- EDA included visualization, mapping launch sites locations and interactive dashboards.
- Predictive analysis using machine leaning methods including; Logistic Regression, KNN, SVM and Classification Trees.
- Key findings include:
  - Rate of success increased dramatically in the first decade.
  - The Decision Tree Classifier was the highest scoring prediction model.

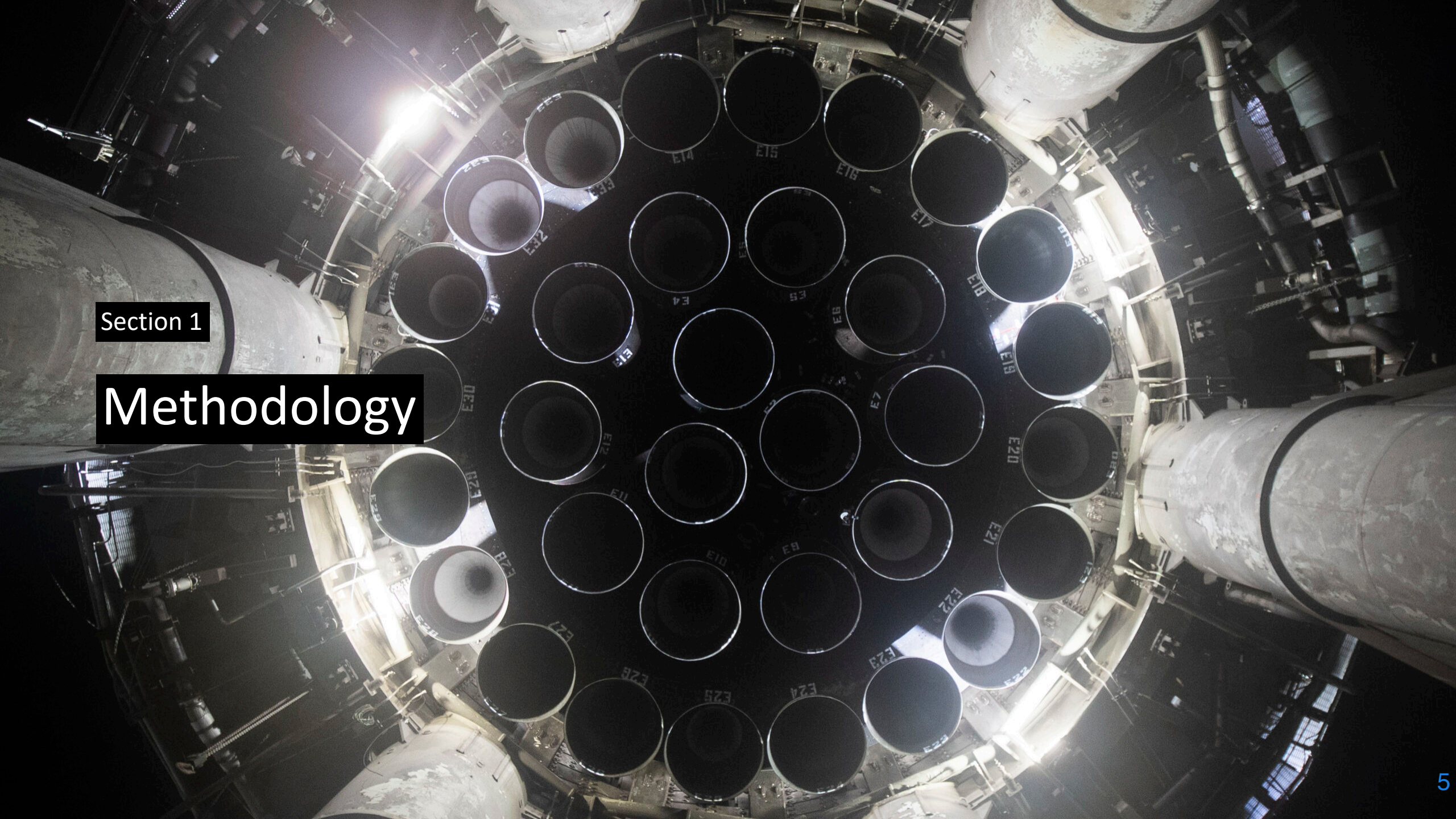
Introduction

SpaceY is entering the space industry and would like to compete against existing companies. Most notably, SpaceX.

SpaceX advertises their rocket launches at a starting rate of 62 million dollars, compared to other providers cost of upward of 165 million dollars each.

Much of the cost saving is due to SpaceX ability to reuse their first stage.

If we can determine the factors contributing to successful launch outcomes, we can help determine the potential cost of a launch, and help SpaceY win the space race.



Section 1

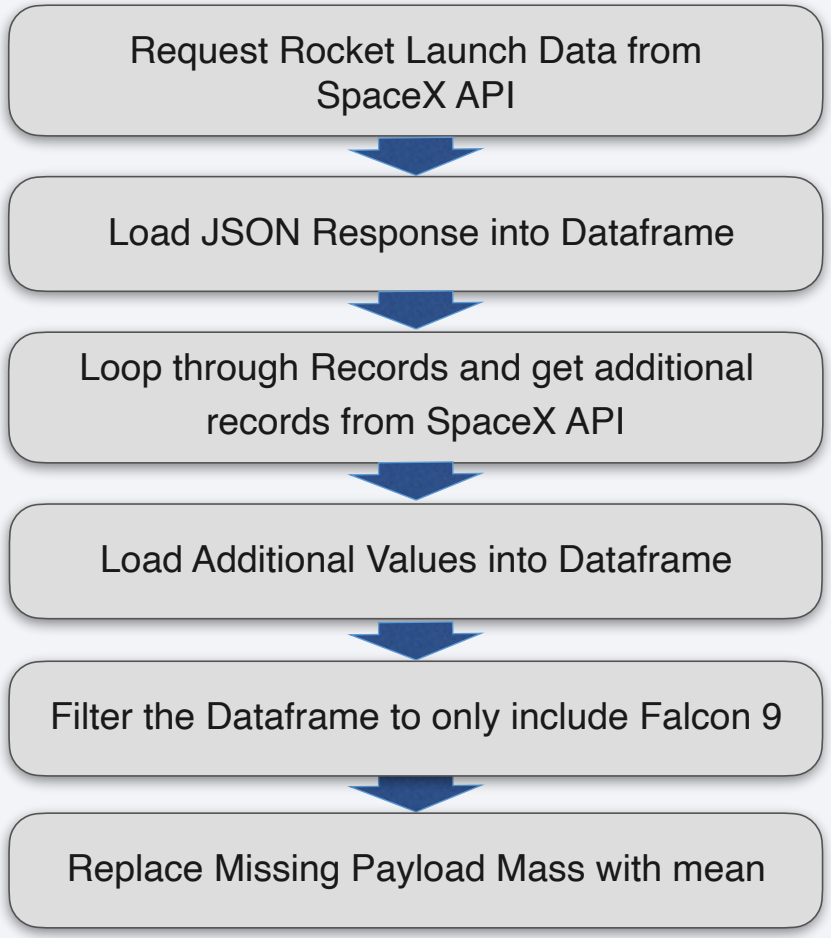
Methodology

Methodology

- Data collection:
  - Space X REST API
  - Web scrapping from a Wikipedia page titled "List of Falcon 9 and Falcon Heavy launches”
- Data wrangling:
  - Removed data related to boosters other then Flacon 9
  - Replaces missing payload mass with mean of known payloads
  - Narrowed down the booster landing outcome to bad outcome (0) and otherwise (1)
- Exploratory data analysis (EDA):
  - Visualization
  - Insight with SQL
  - Mapping Launch Sites
  - Interactive Dashboard
- Predictive analysis:
  - Evaluate the accuracy of machine learning classification models for predicting launch outcomes

Data Collection – SpaceX API

- Retrieved data from SpaceX REST API:
  - Past Launches - https://api.spacexdata.com/v4/launches/past
    - **Flight number, date, landing outcome, type of landing, number of flights, with grid-fins, core reused, legs used, landing pad**
- Then using the IDs from the launch records to retrieve additional records:
  - Boosters - https://api.spacexdata.com/v4/rockets/
    - **Booster name**
  - Launch Sites - https://api.spacexdata.com/v4/launchpads/
    - **Launch site, longitude, latitude**
  - Payloads - https://api.spacexdata.com/v4/payloads/
    - **payload mass and orbit**
  - Rocket Cores - https://api.spacexdata.com/v4/cores/
    - **Version of cores, number of times core reused, serial**
- See notebook on Github

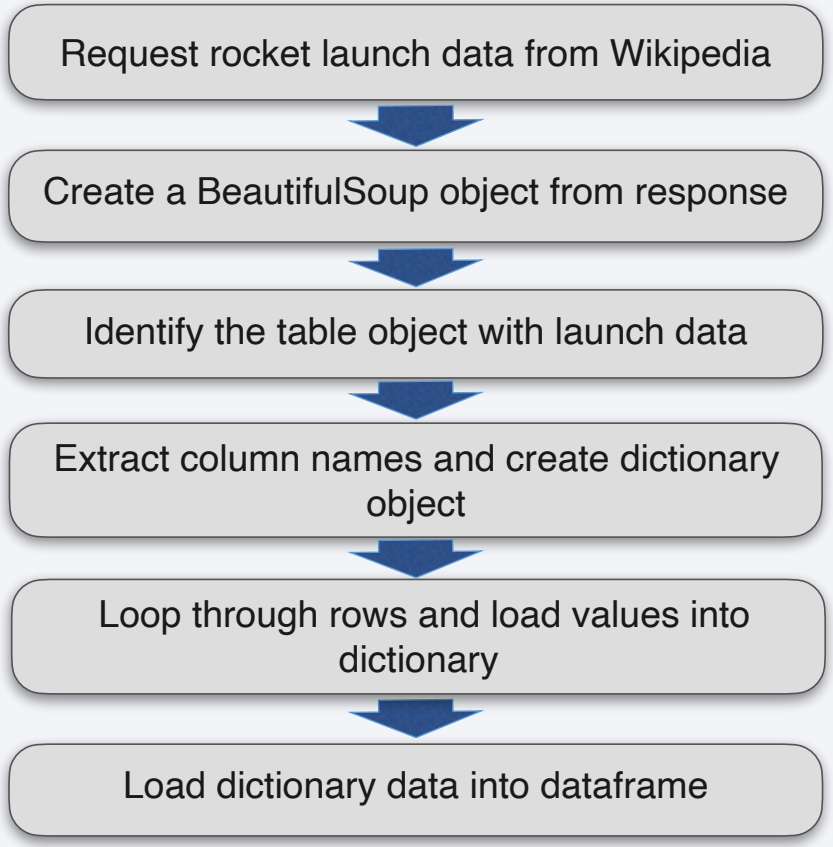


Data Collection - Scraping

- From wikipedia page titled “List of Falcon 9 and Falcon Heavy launches" we collected the following values:

  Flight No. ,
  Date and time ( ),
  Launch site,
  Payload,
  Payload mass,
  Orbit,
  Customer,
  Launch outcome

- See notebook on Github



Data Wrangling

- A total of 8 outcomes were recorded. These outcomes were classified into 2 possible outcomes. This would be the value we would like to predict:

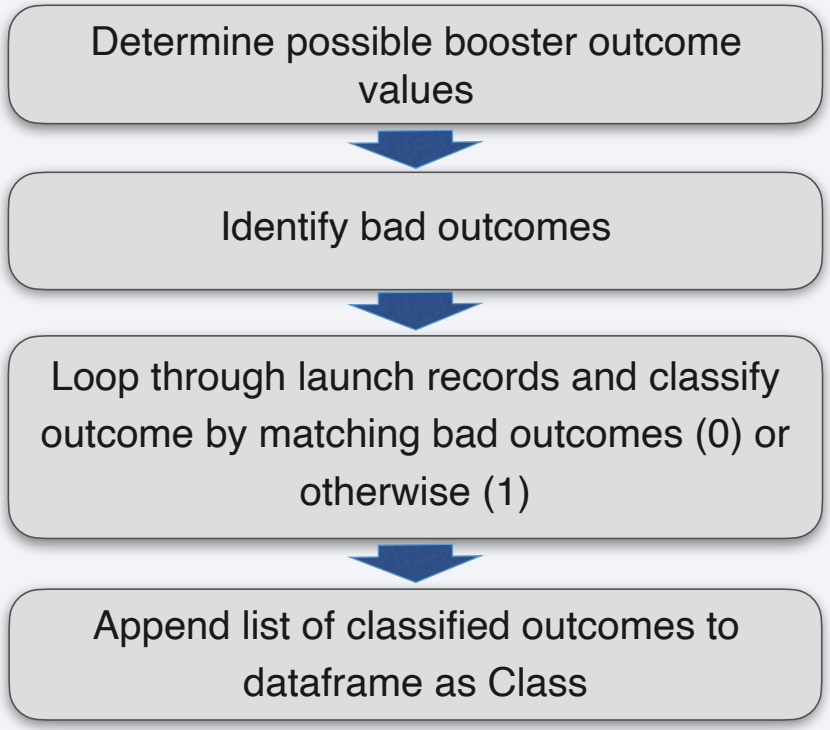
  (0) - Bad Outcome

  False ASDS, False Ocean, False RTLS, None ASDS', 'None None'

  (1) - Otherwise

  True ASDS, True RTLS, True Ocean

- See notebook on GitHub



EDA with Data Visualization

- Explored possible effect some variables have on others using Scatter Plot:
  - Payload mass vs. Number of flights
  - Flight number vs. Launch site
  - Payload mass vs. Launch site
  - Flight number vs. Orbit
  - Payload mass vs Orbit
- Explored using Bar Chart the Success Rate per Orbit.
- Explored using a Line Graph the success rate over the years.
- Applied OneHotEncoder to the columns Orbits, LaunchSite, LandingPad, and Serial.
- See notebook on Github

EDA with SQL

Exploratory data analysis with SQL included the following:

- Listed launch sites
- Explored launch sites beginning with ‘CCA'
- Found the total payload mass carried by boosters launched by NASA
- Found the average payload mass carried by booster version F9 v1.1
- Found the date when the first successful landing outcome in ground pad was achieved.
- Listed the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- Counted the total number of successful and failure mission outcomes.
- Listed the names of the booster_versions which have carried the maximum payload mass.
- Listed the month names, failure landing_outcomes in drone ship ,booster versions and launch_site for the months in year 2015.
- Counted the total of the different landing outcomes between the date 2010-06-04 and 2017-03-20.

- See notebook on Github

Build an Interactive Map with Folium

Exploratory data analysis with Folium Interactive Map:

- Looking for some of the factors to consider when choosing an optimal location for building a launch site by analyzing the existing launch site locations records
- Marked all launch sites on a map
- Marked the success/failed launches for each site on the map
- Calculated the distances between a launch site to its proximities

- See notebook on Github

Build a Dashboard with Plotly Dash

- As part of our exploratory data analysis (EDA) we used Plotly dashboard. We wanted to see if there is relationship between launch sites as well as payload mass and successful outcomes. The dashboard allows us to select launch sites and define the payload mass ranges which produce the following charts:
  - Interactive pie chart that helps visualizes the booster landing success rate by launch site.
  - Interactive scatter plot that illustrates the relationship between payload mass, launch site and landing success.

- See Plotly code on GitHub

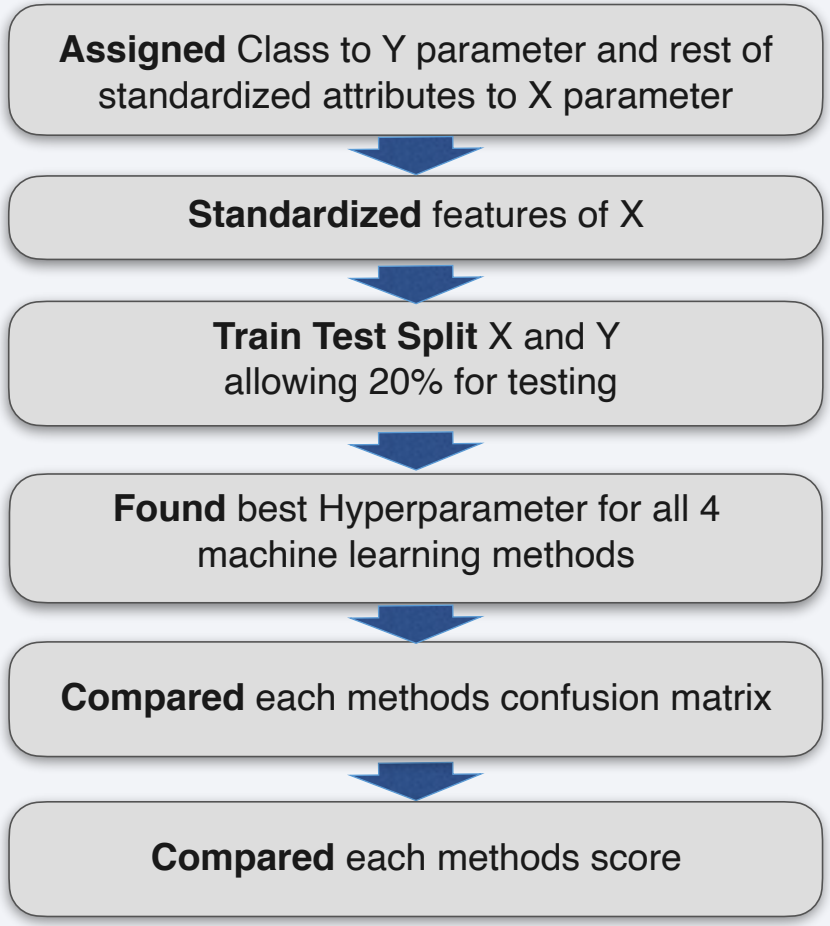
Predictive Analysis (Classification)

Exploratory Data Analysis with Machine Learning Prediction

Performed exploratory data analysis and determine training labels

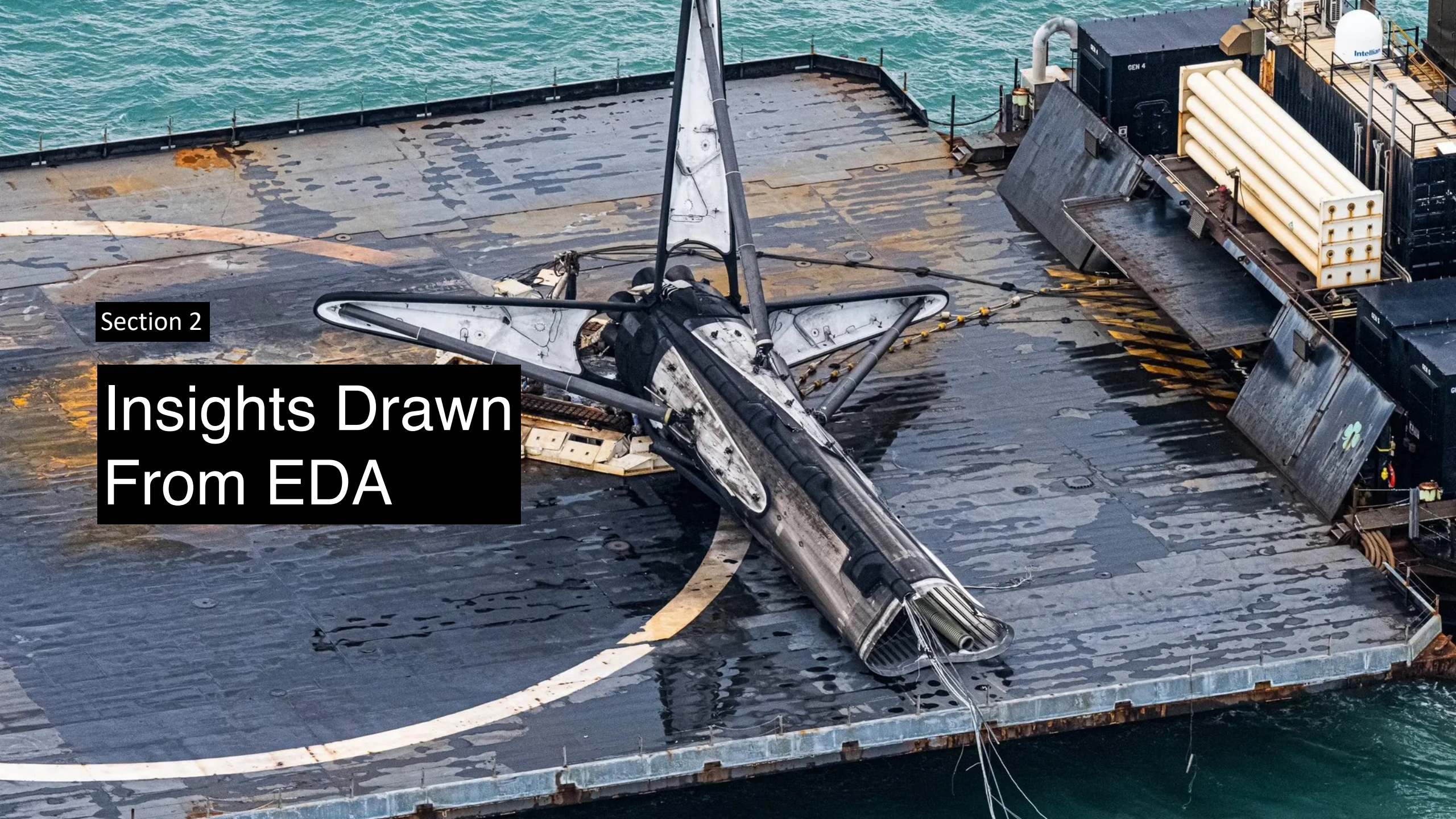
- **Created a column for the class**
- **Standardized the data**
- **Data was split into training data and test data**
- **Found best Hyperparameter for SVM, Classification Trees and Logistic Regression**
- **Identify the method performing best using test records**

- See Notebook on Github



Results

- Flight number is somewhat correlated with success. This might be due to quality of the initial build as we see a lot of failures in the low number of flights.
- From 2010 to 2020 the success rate has kept raising surpassing 50% by 2016 and reaching above 80% by 2019.
- Launch sites on east cost seem to have a slight higher success rate, 69% vs 60% in the west. However there were more launches from the east cost so it is not conclusive.
- Decision Tree Classifier as got the highest accuracy score out of the different machine learning prediction methods we evaluated.

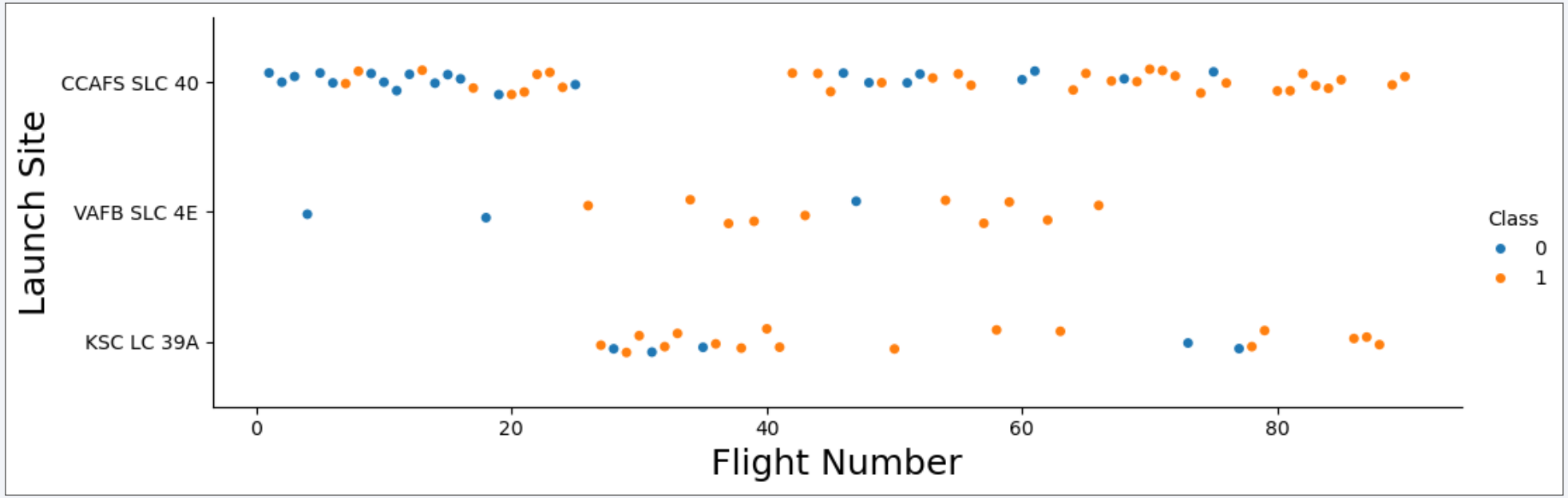


Section 2

Insights Drawn From EDA

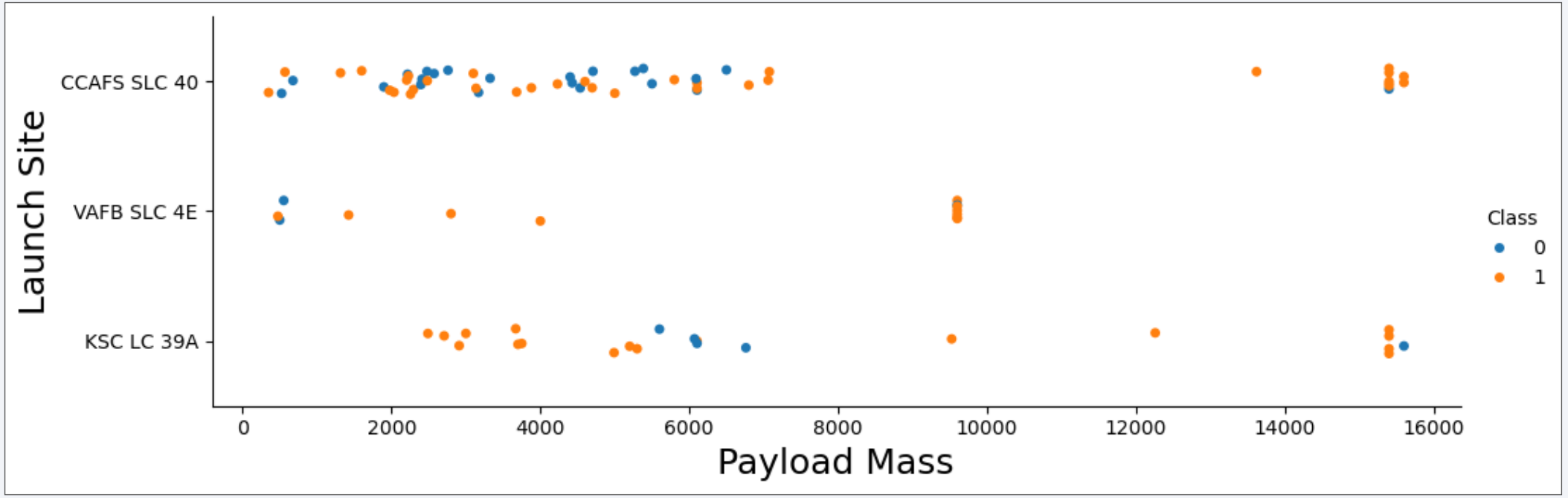
Flight Number vs. Launch Site

Looking for correlation between number of flights, launch site and successful landing. We notice that successful landing is more less even across launching sites. We also notice a larger concentration of bad outcomes at the lower flight number range.



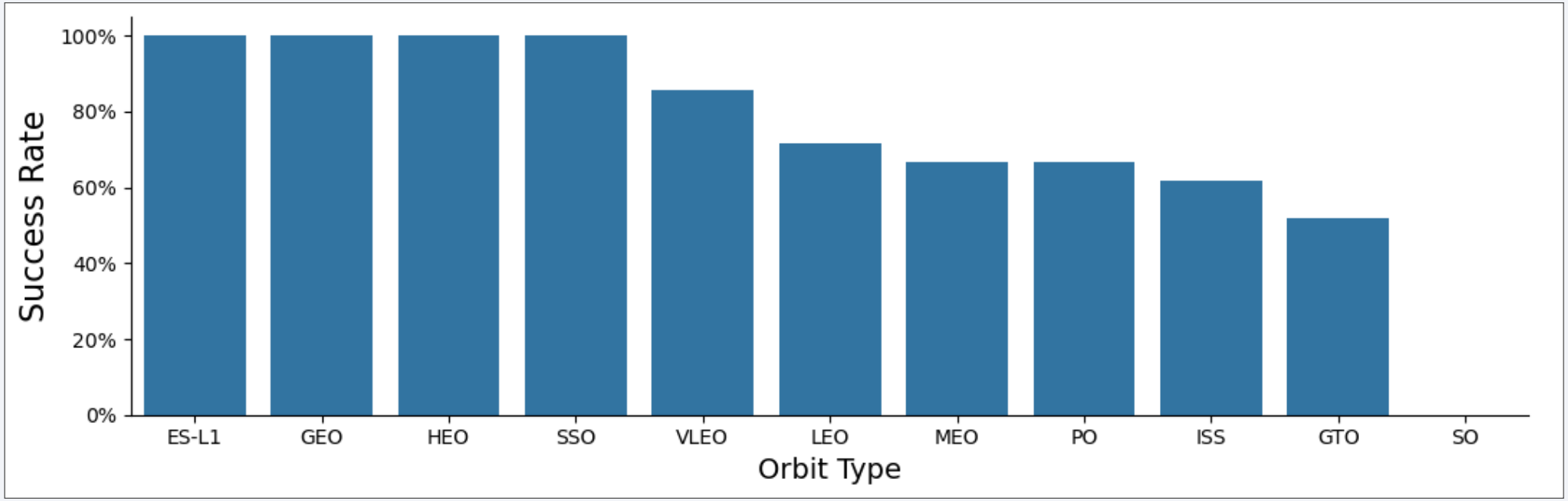
Payload vs. Launch Site

Looking at correlation between launch sites, payload mass and outcome we notice; most bad outcome happen at payload mass below 7000(kg), especially at CCAFS SLC 40.



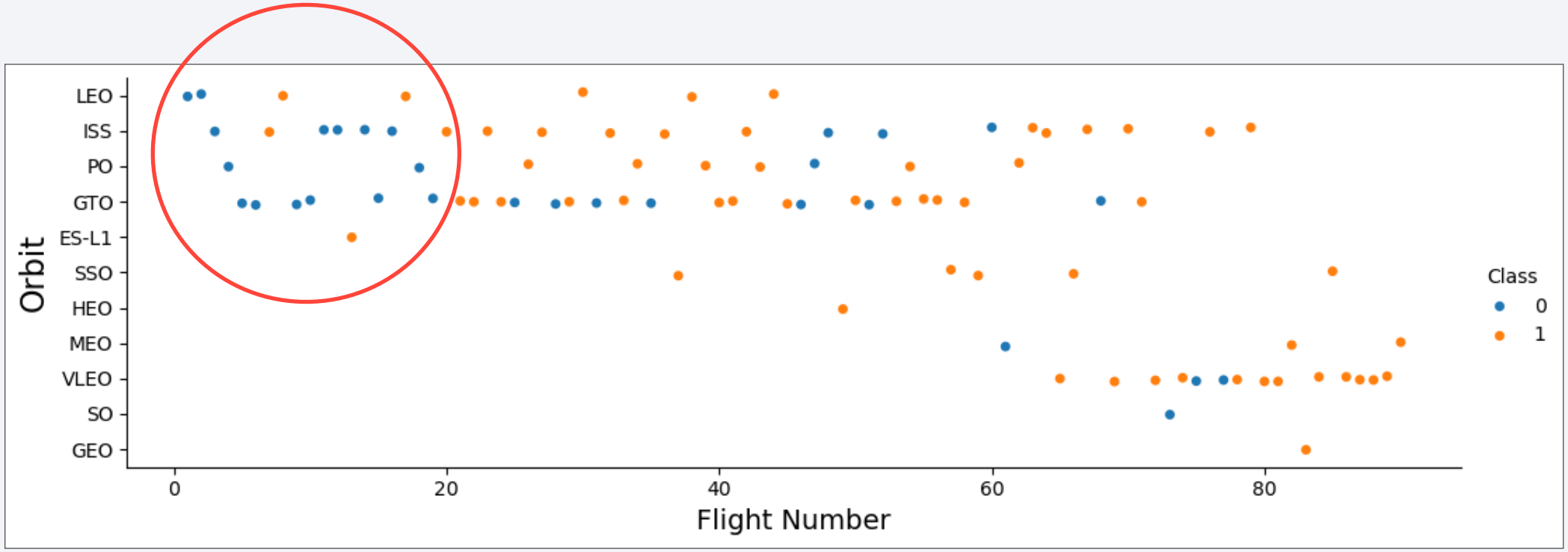
Success Rate vs. Orbit Type

Listing the orbit types by the outcome reveals that 4 orbits had no bad outcomes. 6 other Orbits had mostly good outcome with success rate between 50% and 80%. SO orbit has not gained a successful outcome yet however the next slide shows it only had one launch record.



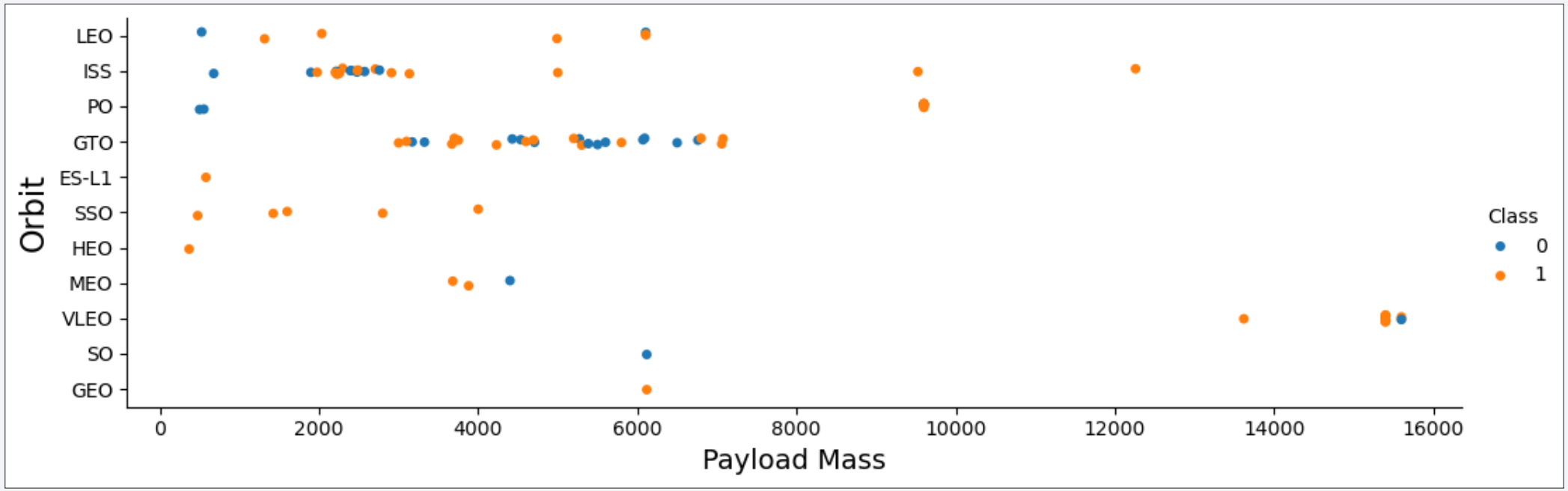
Flight Number vs. Orbit Type

When we review the correlation between number of flights, orbit type and outcome we notice again a higher concentration of bad outcome at the lower number of flights ranges.



Payload vs. Orbit Type

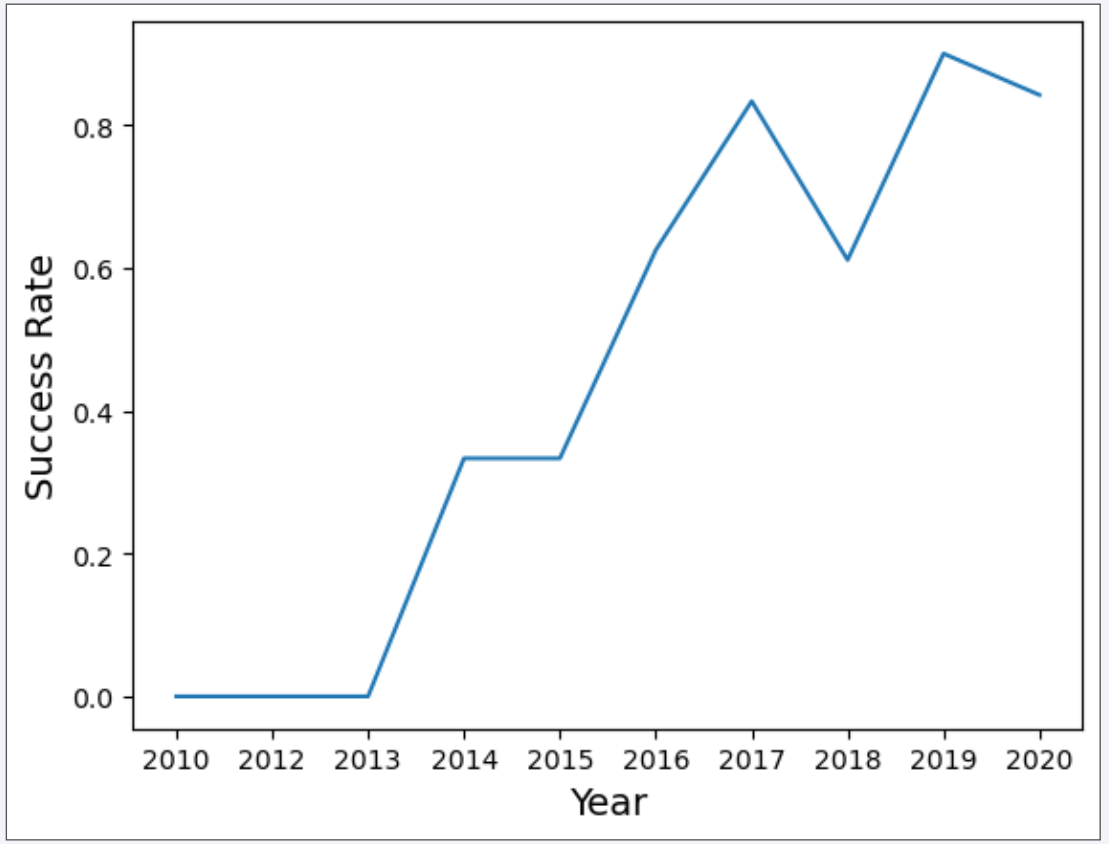
When we evaluate the relationship between Payload mass, orbit type and outcome, we can see that bad out comes are more common with lower payload mass, however not enough to draw any conclusive conclusions.



Launch Success Yearly Trend

Looking into the yearly success trend we can identify a clear improvement over time.

This might suggest that analysis done with years prior to 2016 or even 2017 might skew the result negatively.



All Launch Site Names

Using SQL we identified the launch sites names.

```
SELECT DISTINCT launch_site
FROM   spacextable
ORDER  BY launch_site;
```

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

Launch Site Names Begin with 'CCA'

Using SQL we found 5 records where launch sites begin with `CCA`

```
SELECT *
FROM   spacextable
WHERE  launch_site LIKE 'CCA%'
LIMIT  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	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

Using SQL we calculated the total payload carried by boosters from NASA (CRS)

```
SELECT Sum(payload_mass__kg_) AS 'Sum of payload carried for NASA(CRS)'
FROM   spacextable
WHERE  customer = 'NASA (CRS)';
```

Sum of payload carried for NASA(CRS)
45596

Average Payload Mass by F9 v1.1

Using SQL we calculated the average payload mass carried by booster version F9 v1.1

```
SELECT Round(Avg(payload_mass__kg_), 2) AS 'Average Payload for F9 v1.1'
FROM   spacextable
WHERE  booster_version LIKE 'F9 v1.1%';
```

Average Payload for F9 v1.1
2534.67

First Successful Ground Landing Date

Using SQL we listed the date when the first successful landing outcome in ground pad was achieved.

```
SELECT Min(date) AS 'First Ground pad landing'
FROM   spacextable
WHERE  landing_outcome = 'Success (ground pad)';
```

First Ground Pad Landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
SELECT booster_version
FROM   spacextable
WHERE  payload_mass__kg_ BETWEEN 4000 AND 6000
       AND landing_outcome = 'Success (drone ship)';
```

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

Total Number of Successful and Failure Mission Outcomes

Using SQL we listed the total number of successful and failure mission outcomes

```
SELECT  (SELECT Count(*)
         FROM   spacextable
         WHERE  landing_outcome LIKE '%Success%')     AS 'Success',
        (SELECT Count(*)
         FROM   spacextable
         WHERE  landing_outcome NOT LIKE '%Success%') AS 'Failure';
```

Success	Failure
61	40

Boosters Carried Maximum Payload

Using SQL we listed the names of the booster versions which have carried the maximum payload mass.

```
SELECT DISTINCT booster_version,
                payload_mass__kg_
FROM   spacextable
WHERE  payload_mass__kg_ = (SELECT Max(payload_mass__kg_)
                            FROM   spacextable);
```

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

2015 Launch Records

```
SELECT CASE Substr(date, 6, 2)
         WHEN '01' THEN 'January'
         WHEN '02' THEN 'February'
         WHEN '03' THEN 'March'
         WHEN '04' THEN 'April'
         WHEN '05' THEN 'May'
         WHEN '06' THEN 'June'
         WHEN '07' THEN 'July'
         WHEN '08' THEN 'August'
         WHEN '09' THEN 'September'
         WHEN '10' THEN 'October'
         WHEN '11' THEN 'November'
         WHEN '12' THEN 'December'
         ELSE ''
       END                  AS 'Month',
       Substr(date, 0, 5) AS 'Year',
       landing_outcome,
       booster_version,
       launch_site
FROM   spacextable
WHERE  Substr(date, 0, 5) = '2015'
       AND landing_outcome = 'Failure (drone ship)';
```

Using SQL we listed the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

Month	Year	Landing_Outcome	Booster_Version	Launch_Site
January	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

Using SQL we ranked the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20

```
SELECT landing_outcome,
       Count(landing_outcome) AS 'Count'
FROM   spacextable
GROUP  BY landing_outcome
HAVING Substr(date, 1, 4)
       || Substr(date, 6, 2)
       || Substr(date, 9, 2) BETWEEN '20100604' AND '20170320'
ORDER  BY count DESC;
```

Landing_Outcome	Count
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

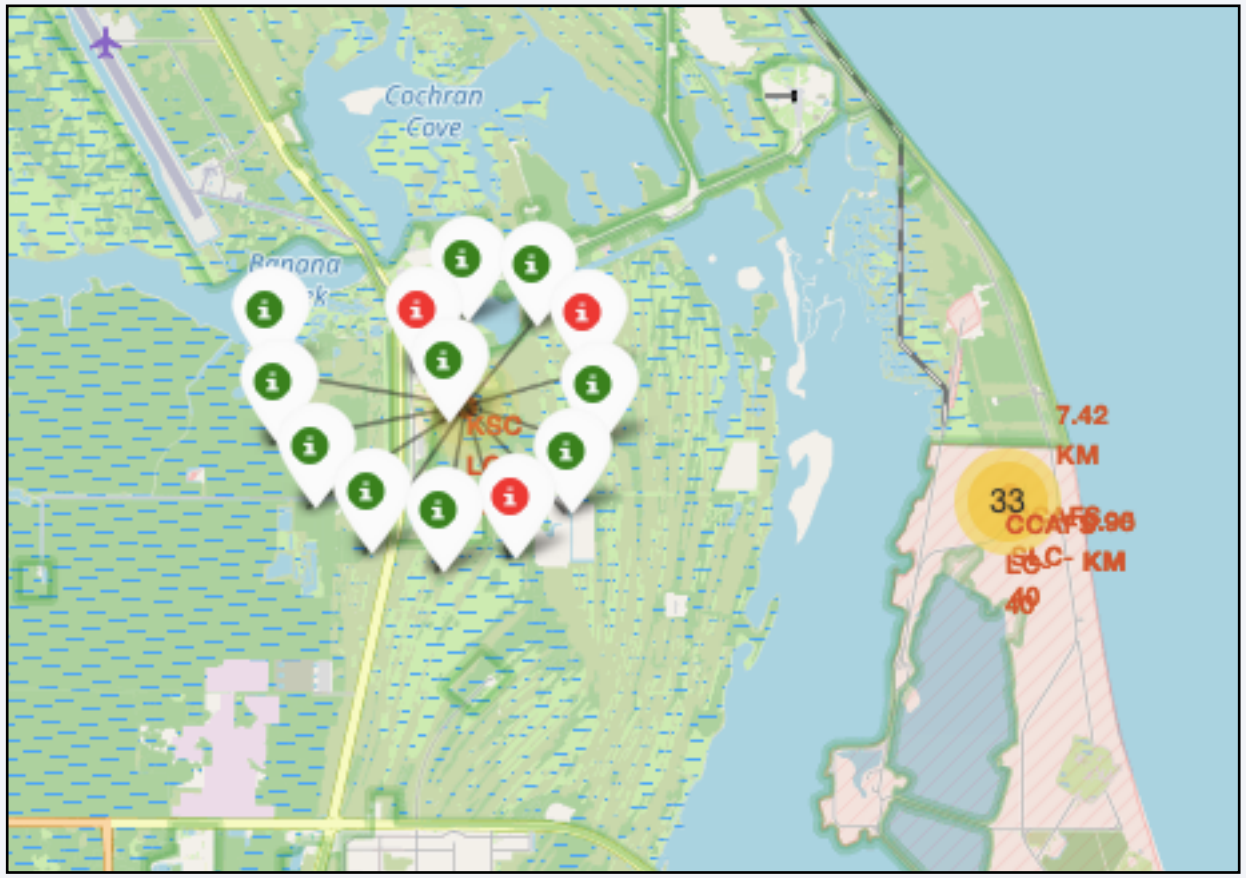
Launch Sites Proximities Analysis

Marking Launch site locations on a map

Using Folium we marked the launch sites on a map to explore the a link between location and possible outcome. We notice that all sites are by the coast line. 3 Sites on the east cost and another site on the west cost.

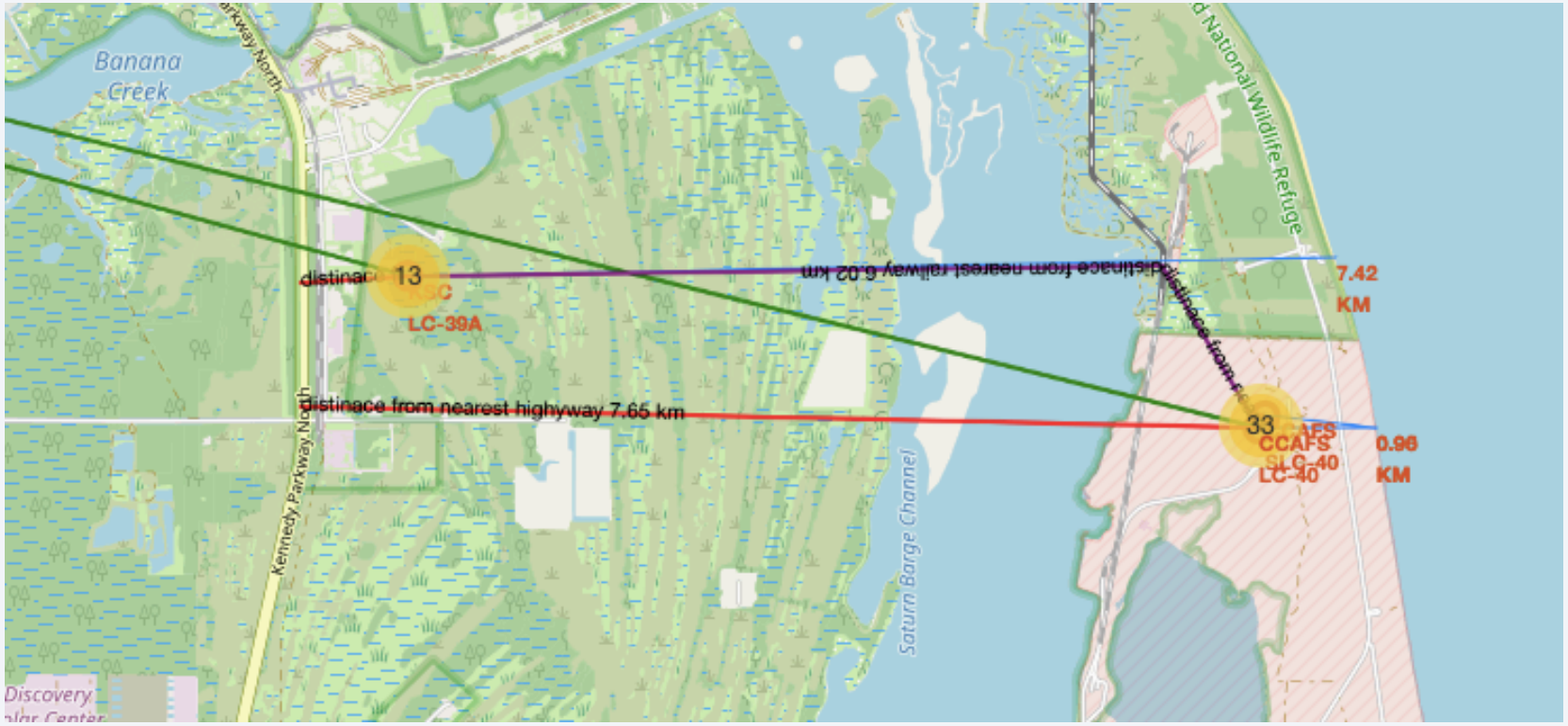
Marking Success/Failed Launches on the Map

As part of our exploratory data analysis we added marker clusters to the launch sites to identify the success/failed launches. The result is inconclusive. There isn't a strong link between the location of the launch site and success rate.



Marking nearest City, Railroad and Highway

We explored a possible connection between the vicinity of cities, highways and railroads and outcomes.

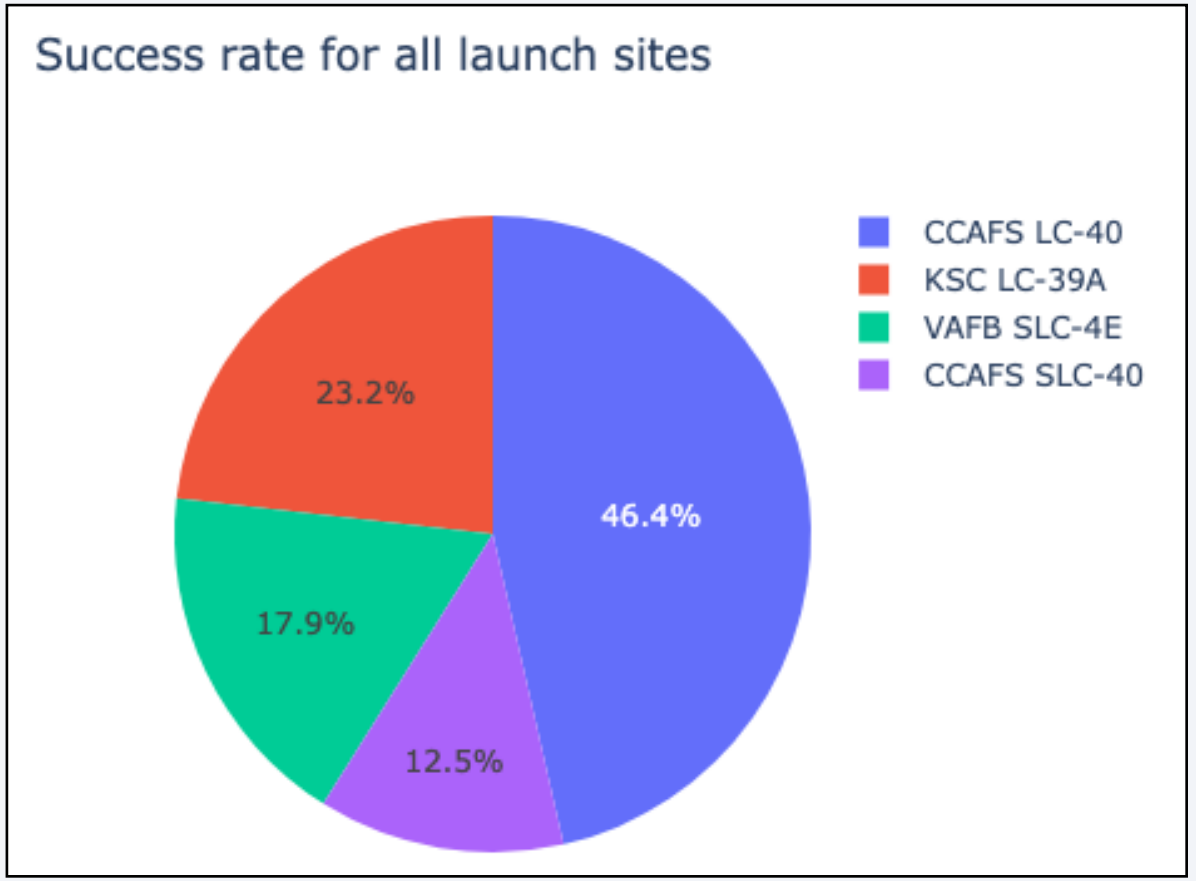


Section 4

Build a Dashboard With Plotly Dash

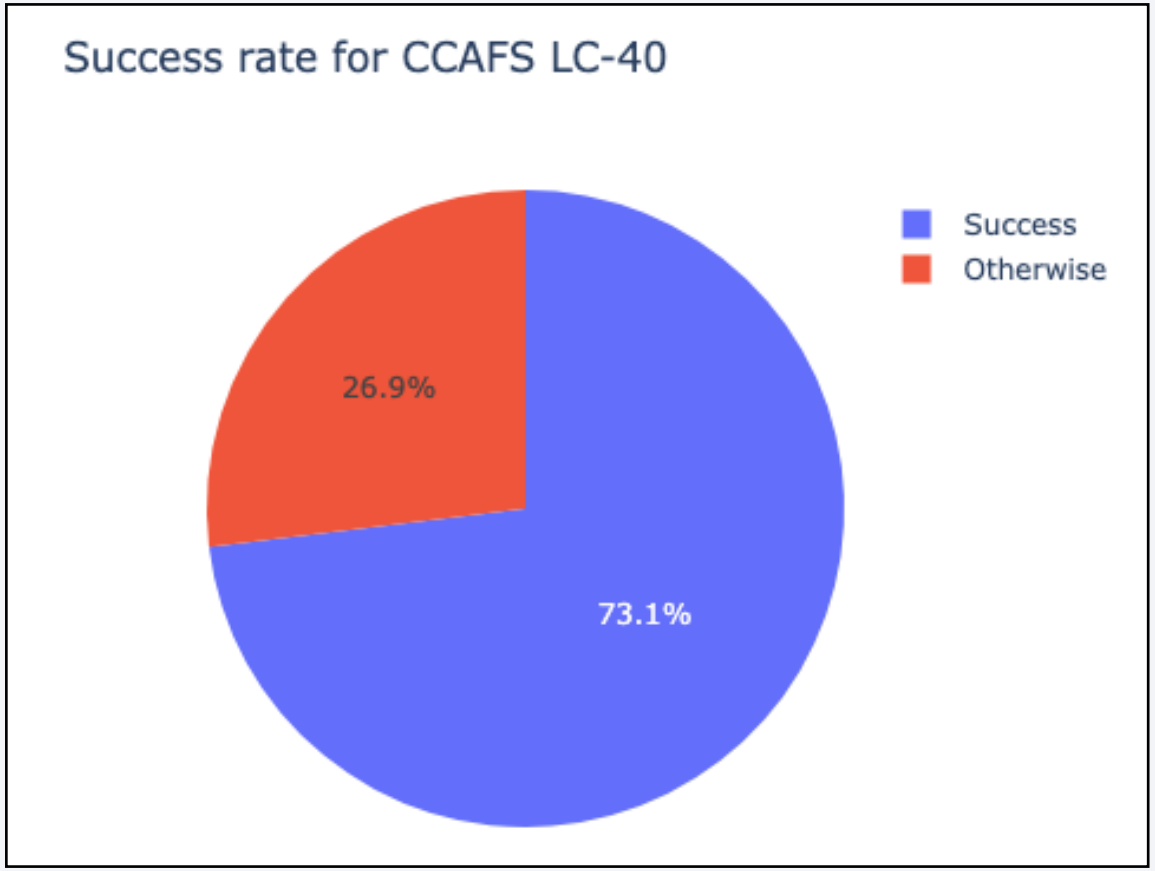
Success Rate for all Launch Sites with Plotly

Using Plotly we Implemented an interactive dashboard for exploratory data analysis. We compared the launch sites and found that CCAFS LC-40 had the highest share of stage one successful landings.



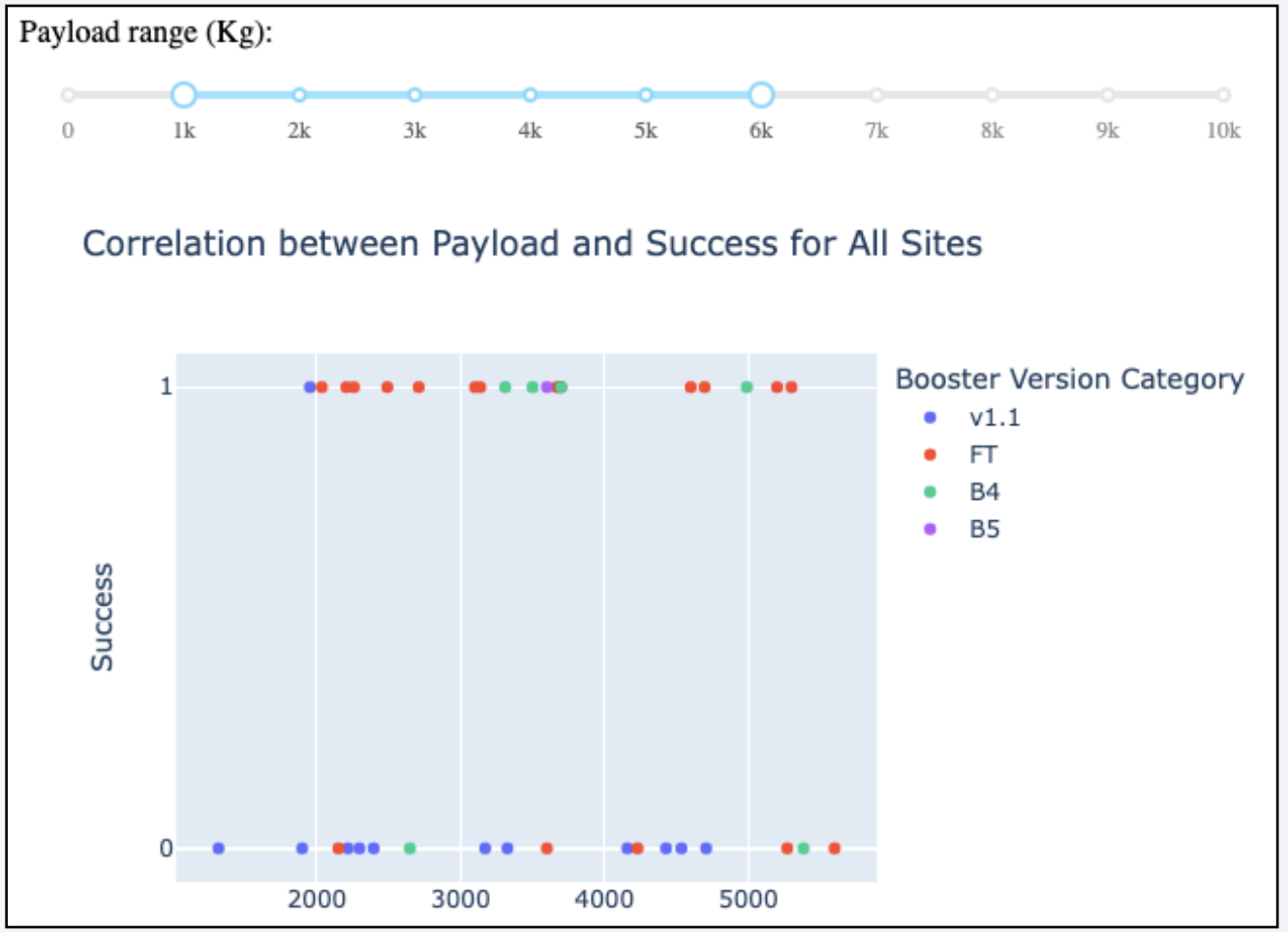
Plotly - Launch Site with Highest Success Rate

Exploring with Plotly dashboard we identified CCAFS LC-40 to be the launch site with the highest success rate of 73.1%



Plotly - Payload Mass and Successful Outcome

Using Plotly dashboard for exploratory data analysis we identified the payload mass range with highest concentration of successful outcomes to be between 2,000 kg and 5,500 kg. We also identified FT and B4 to be the best performing booster versions.



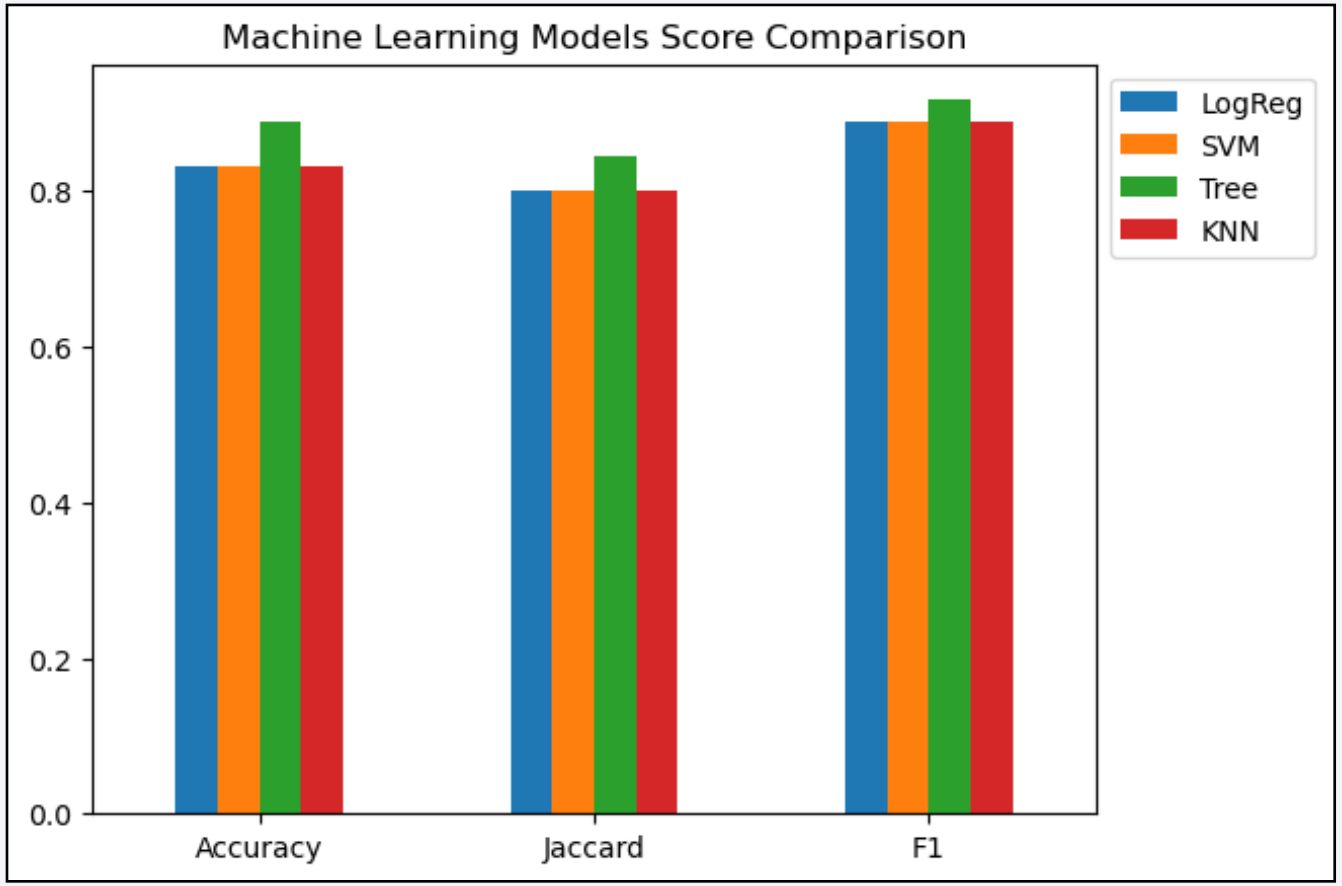
Section 5

Predictive Analysis (Classification)

Classification Accuracy

After training different classification machine learning models, with 80% of the dataset, we tested their accuracy score with the remaining 20%.

We found that the decision tree classifier got the highest accuracy results.

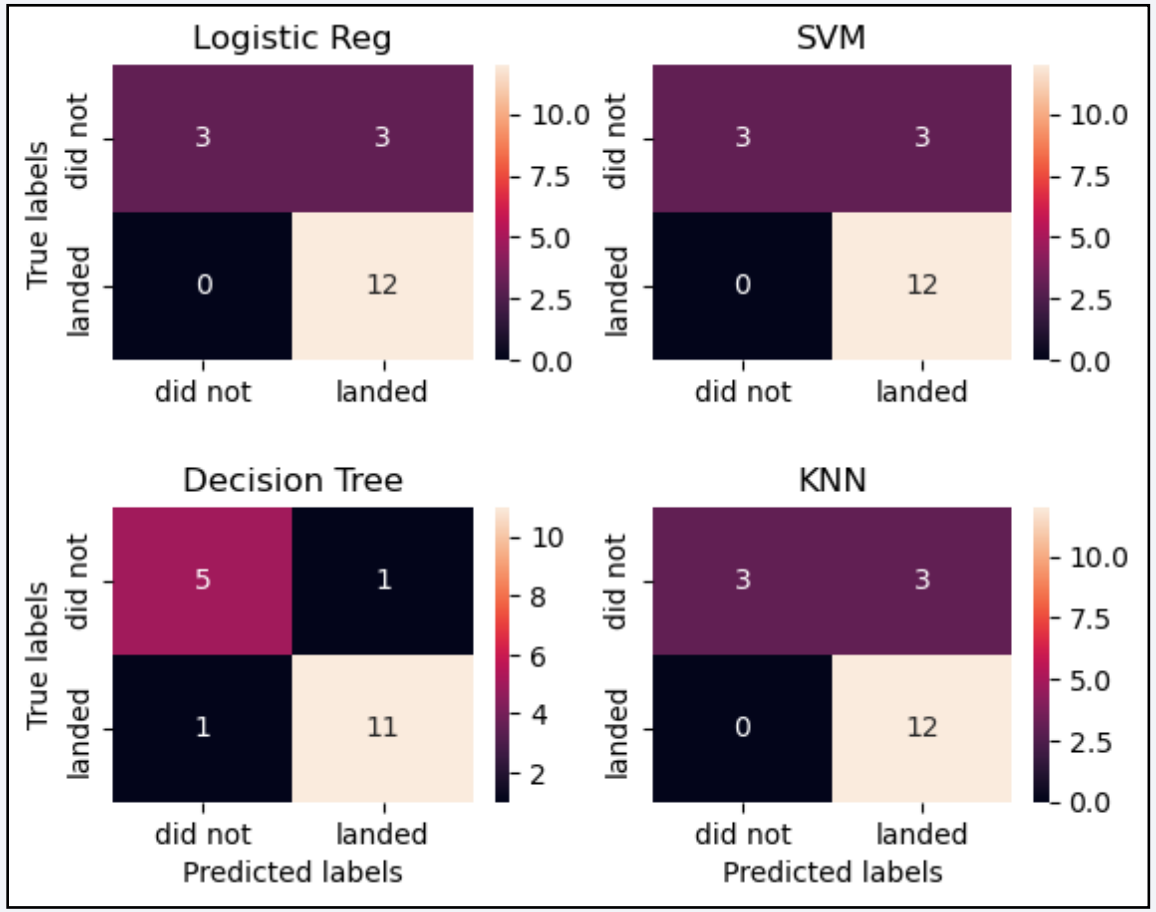


Confusion Matrix

Using the predicted test data and actual test data we plotted a confusion matrix for each machine learning model.

When comparing the Matrixes we noticed that while KNN, Logistical Regression and SVM predicted accurately the successful landings, they failed to predict 3 unsuccessful landings.

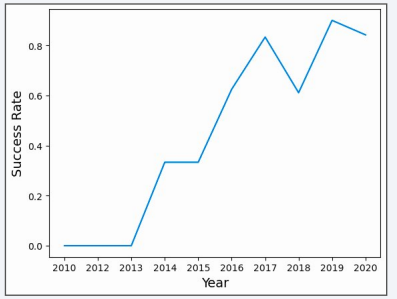
Decision Tree model was able to classify correctly all but 1 from each outcome class.

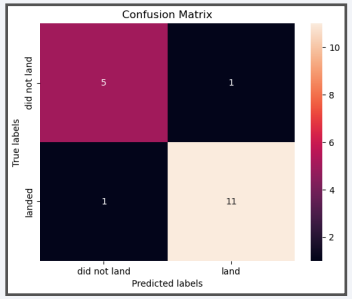


Conclusions

Main finding

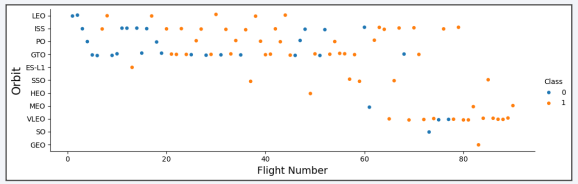
- Successful landing rate increased over the years.
- Decision Tree Classifier shown to be the most accurate in classifying landing outcomes, with accuracy score of 0.889

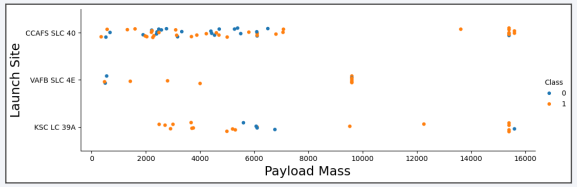


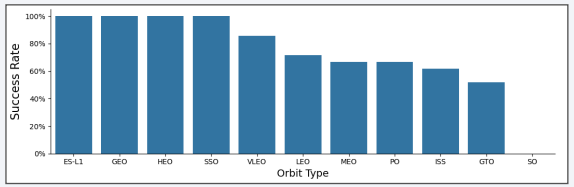


Other findings

- A larger concentration of bad outcomes at the lower flight number range.
- Most bad outcome happen at payload mass below 7000(kg), especially at CCAFS SLC 40. However that is also the range with the highest number of launches.
- Orbits ES-L1, GEO, HEO and SSO had no bad outcomes. However the number of attempts is far lower than other orbits.

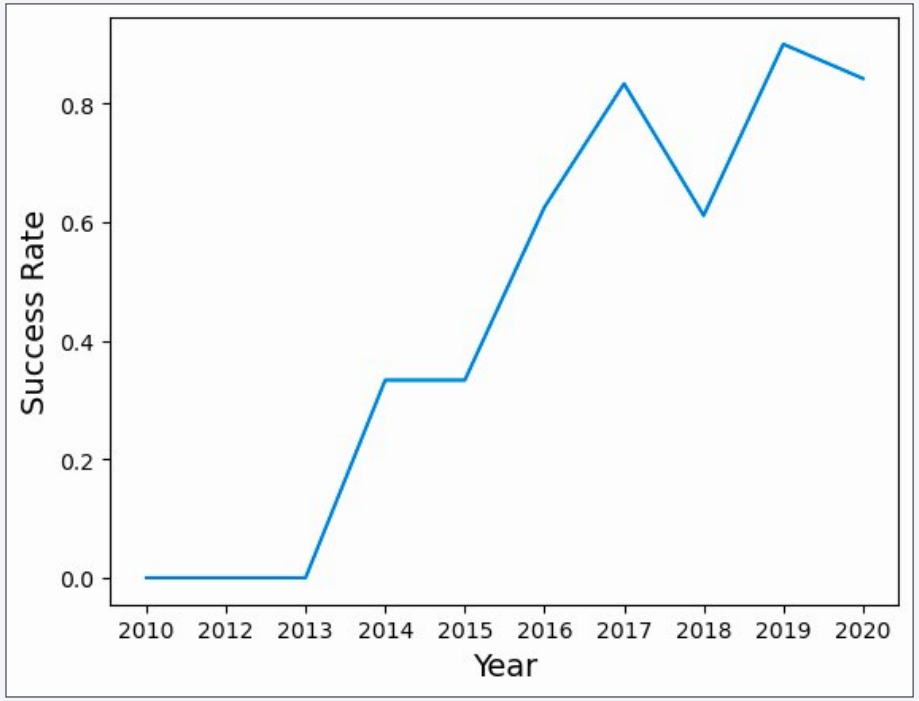






Next steps

- Given that the success rate increased over the years and considering the data we evaluated, ended in 2020, we ought to preform this analysis with more up to date dataset.



?


Thank you

Appendix

Orbit references

LEO: Low Earth orbit (LEO)is an Earth-centred orbit with an altitude of 2,000 km (1,200 mi) or less (approximately one-third of the radius of Earth),[1] or with at least 11.25 periods per day (an orbital period of 128 minutes or less) and an eccentricity less than 0.25.[2] Most of the manmade objects in outer space are in LEO [1].

VLEO: Very Low Earth Orbits (VLEO) can be defined as the orbits with a mean altitude below 450 km. Operating in these orbits can provide a number of benefits to Earth observation spacecraft as the spacecraft operates closer to the observation[2].

GTO A geosynchronous orbit is a high Earth orbit that allows satellites to match Earth's rotation. Located at 22,236 miles (35,786 kilometers) above Earth's equator, this position is a valuable spot for monitoring weather, communications and surveillance. Because the satellite orbits at the same speed that the Earth is turning, the satellite seems to stay in place over a single longitude, though it may drift north to south," NASA wrote on its Earth Observatory website [3] .

SSO (or **SO**): It is a Sun-synchronous orbit also called a heliosynchronous orbit is a nearly polar orbit around a planet, in which the satellite passes over any given point of the planet's surface at the same local mean solar time [4] .

ES-L1 :At the Lagrange points the gravitational forces of the two large bodies cancel out in such a way that a small object placed in orbit there is in equilibrium relative to the center of mass of the large bodies. L1 is one such point between the sun and the earth [5] .

HEO A highly elliptical orbit, is an elliptic orbit with high eccentricity, usually referring to one around Earth [6].

ISS A modular space station (habitable artificial satellite) in low Earth orbit. It is a multinational collaborative project between five participating space agencies: NASA (United States), Roscosmos (Russia), JAXA (Japan), ESA (Europe), and CSA (Canada) [7]

MEO Geocentric orbits ranging in altitude from 2,000 km (1,200 mi) to just below geosynchronous orbit at 35,786 kilometers (22,236 mi). Also known as an intermediate circular orbit. These are "most commonly at 20,200 kilometers (12,600 mi), or 20,650 kilometers (12,830 mi), with an orbital period of 12 hours [8]

HEO Geocentric orbits above the altitude of geosynchronous orbit (35,786 km or 22,236 mi) [9]

GEO It is a circular geosynchronous orbit 35,786 kilometres (22,236 miles) above Earth's equator and following the direction of Earth's rotation [10]

PO It is one type of satellites in which a satellite passes above or nearly above both poles of the body being orbited (usually a planet such as the Earth [11]

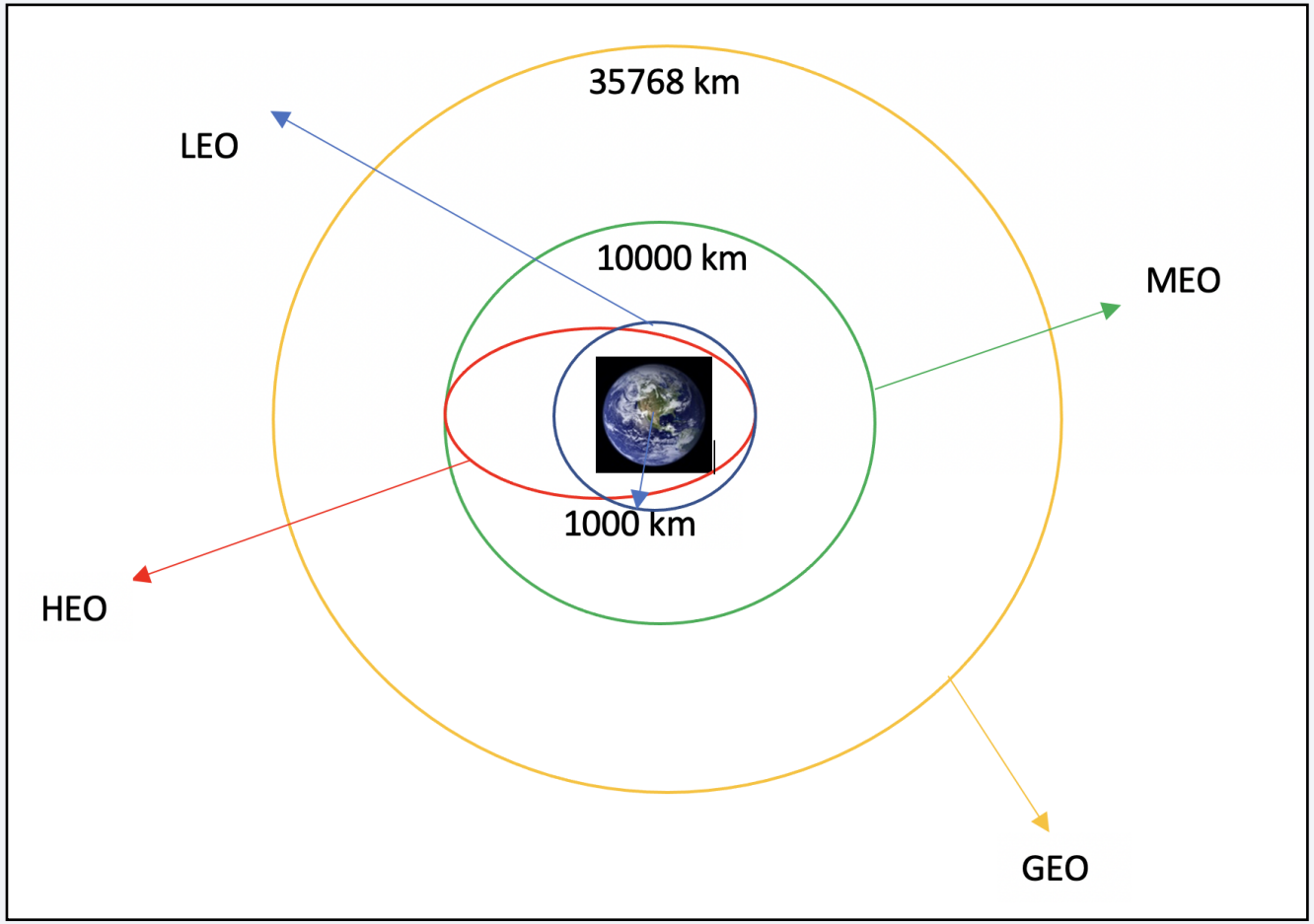
Appendix

Launch sites references

- **CCAFS SLC 40**, **CCAFS LC 40**: Cape Canaveral Space Launch Complex 40
- **VAFB SLC 4E** : Vandenberg Air Force Base Space Launch Complex 4E
- **KSC LC 39A**: Kennedy Space Center Launch Complex 39A

Appendix

Orbits illustration



Appendix

Defining Outcomes

- Outcome values we considered to be successful outcomes (classified as 1):
  - **True ASDS** - means the mission outcome was successfully landed to a drone ship.
  - **True RTLS** - means the mission outcome was successfully landed to a ground pad.
  - **True Ocean** - means the mission outcome was successfully landed to a specific region of the ocean.
- Outcomes values we consider to be bad outcomes (classified as 0):
  - **None None** , **None ASDS** - represent a failure to land.
  - **False ASDS -** means the mission outcome was unsuccessfully landed to a drone ship.
  - **False RTLS** - means the mission outcome was unsuccessfully landed to a ground pad.
  - **False Ocean** - means the mission outcome was unsuccessfully landed to a specific region of the ocean.

Appendix - Falcon 9 Flight Stages