IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Christian Gmeiner
19.02.2023

# Outline

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

# Executive Summary

- Summary of methodologies
  - Data collection using webscraping and the provided SpaceX API
  - Data cleaning and wrangling, exploratory data analysis and visualisazion including a dashboard
  - Machine learning predictions and evaluation of different models
- Summary of all results
  - EDA showed a clear trend of increasing success rate over the past years in starts a s well as landings of reusable booster
  - Analysis of the launch sites established KSC LC 39A as most successful location
  - ML and evaluation directed us to a decision tree classifier as the best model for predictions

# Introduction

The objective is to evaluate the viability of the new company Space Y to compete with Space X.

Expected answers:

- Predicting launch cost based on successful landings and therefor reuse of first stage bosster rockets
- Launchfacility with the highest success rate

Section 1

# Methodology

# Methodology

## Executive Summary

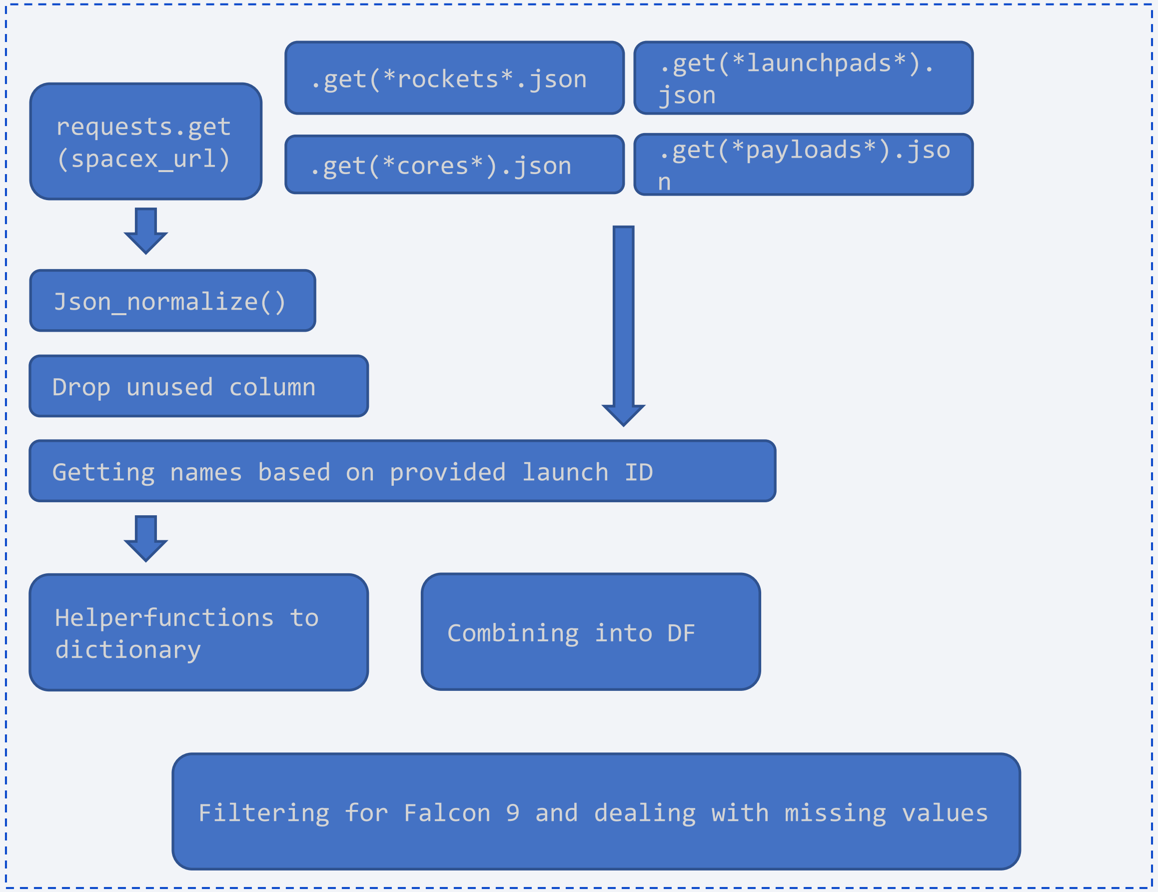
- Data collection methodology:
  - Two sources were used - Space X API and launch data from Wikipedia
- Perform data wrangling
  - Creating of a landing class columns represented a binary for all possible outcomes
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data was normalized and divides into training and test data
  - Each model was iteratively trained to find the best settings

# Data Collection

- Two different methods were used to gather the required date
  - Calls to the Space X API
  - Webscraping with BeautifulSoup
- Objects were parsed accordingly into dataframes and populated with additional data
- Dealt with missing values from API call

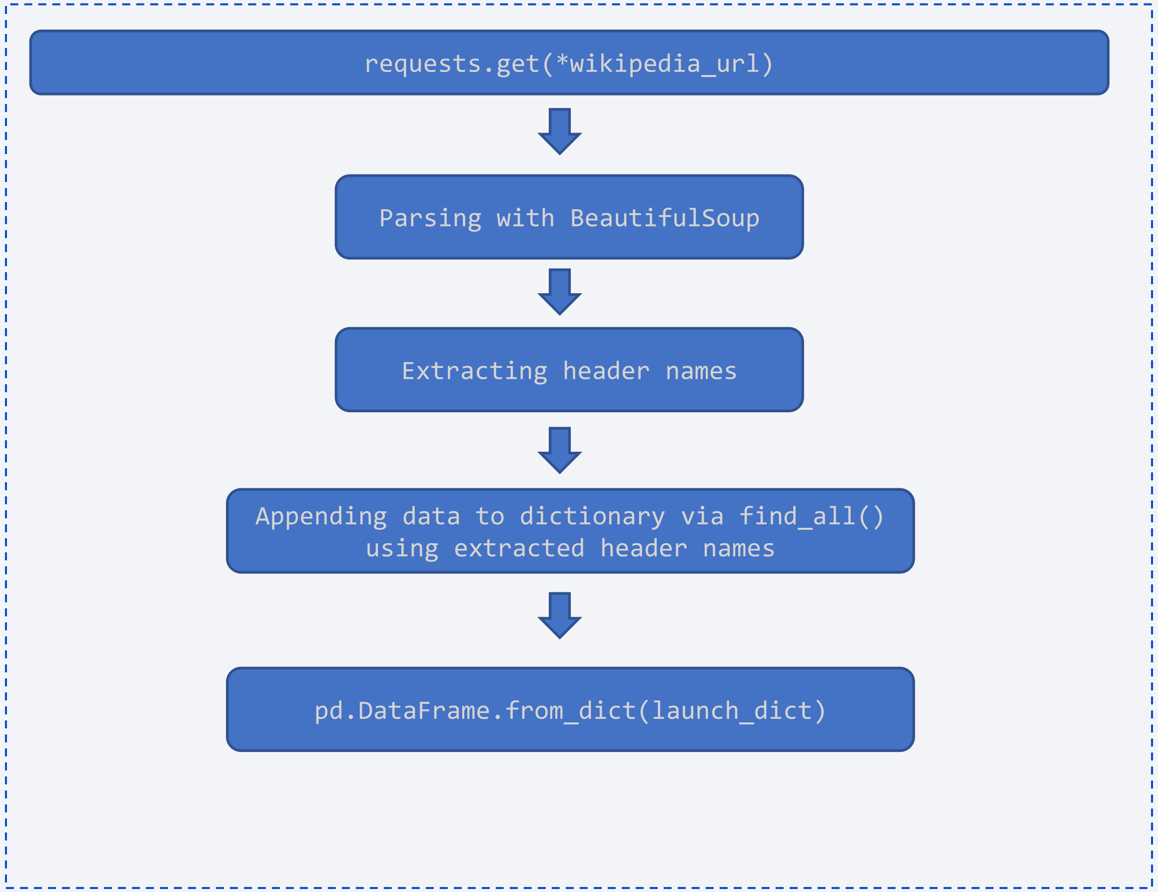
# Data Collection – SpaceX API

- Getting data from API as response
- https://api.spacexdata.com/v4/launches/past
- Normalizing json data into dataframe
- Additional API calls with helper functions to get more descriptive data as dictionary
- Combining everything into DF
- Replacing missing values in Payload mass with mean()
- link



# Data Collection - Scraping

- Getting response from
  `https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922`
- Parsing with beautifulSoup
- Extracting header information
- Extracting data into dictionary
- Combining into dataframe
- [link]



# Data Wrangling

- There were eight possible outcomes in Outcome column
- This needed to be cast into a binary form of success and failure for further steps
- The new class column was used for preliminary success rate calculation using mean()
- Overall success rate = 0,66%

- link

# EDA with Data Visualization

- Multiple scatter plots were used to show the relationship between different variables
  - To gather insights and see correlation
- A line graph to show the trend of increasing success rate over time
- A bar chart to show numeric relationship between Orbit and success rate

- link

# EDA with SQL

Multiple queries were performed to gather insights and additional metrics

- Names of the unique launch sites in the space mission.
- 5 records where launch sites 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 of the first successful landing outcome on ground pad.
- 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.
- Names of the booster_versions which have carried the maximum payload mass.
- Records which will display the month names, failure landing_ouutcomes in drone ship, booster versions, launch_site for 2015.
- Count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 descending.
- link

# Build an Interactive Map with Folium

- Marked the launch sites on map overview
- Created markers for each site showing the outcome as red and green markers
- Measured distances for each site for
    - Coastlines
    - Railways
    - Highways
    - Cities
- link

# Build a Dashboard with Plotly Dash

- Added a dropdown menu to select the four different launch sites to filter the charts
- Pie chart for visualizing the success rate
- Slider for payload mass to filter scatter plot and view which booster versions are successful depended on the mass selectes
- Scatter plot with payload mass, success rate and colorcoded booster versions

- link

# Predictive Analysis (Classification)

- Data normalization and split into test and train set
- Training models via GridSearchCV to finetune parameters
- Evaluation of models
- Displaying and comparing scores
- Bar chart to visualize the performance of each model type

- link

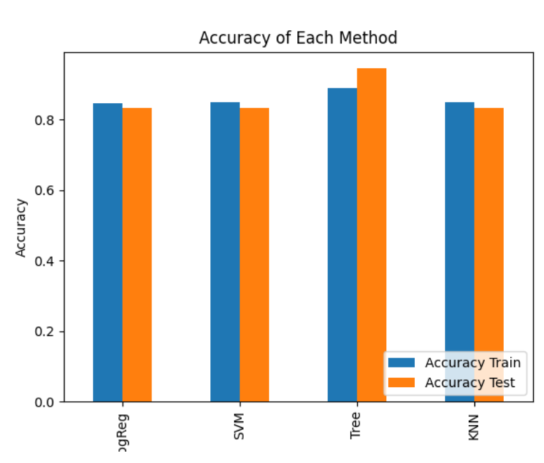
# Results

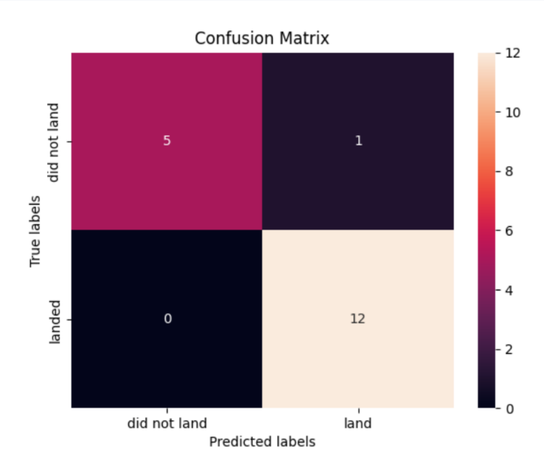
- While overall mission outcome is very high, 100 out of 101 successful, the landing outcome lies behind with 61 successful out of 101
  - This is due to 22 landings not even attempted
  - Both success rates increased significantly over time
  - First successful recovery was in 2015 – five years after the first rocket launch
- 3 out of 4 launch sites are located on the east coast
  - Including the highly successful KSC LC-39A

# Results

- The decision tree classifier was

Chosen as the best model for this task



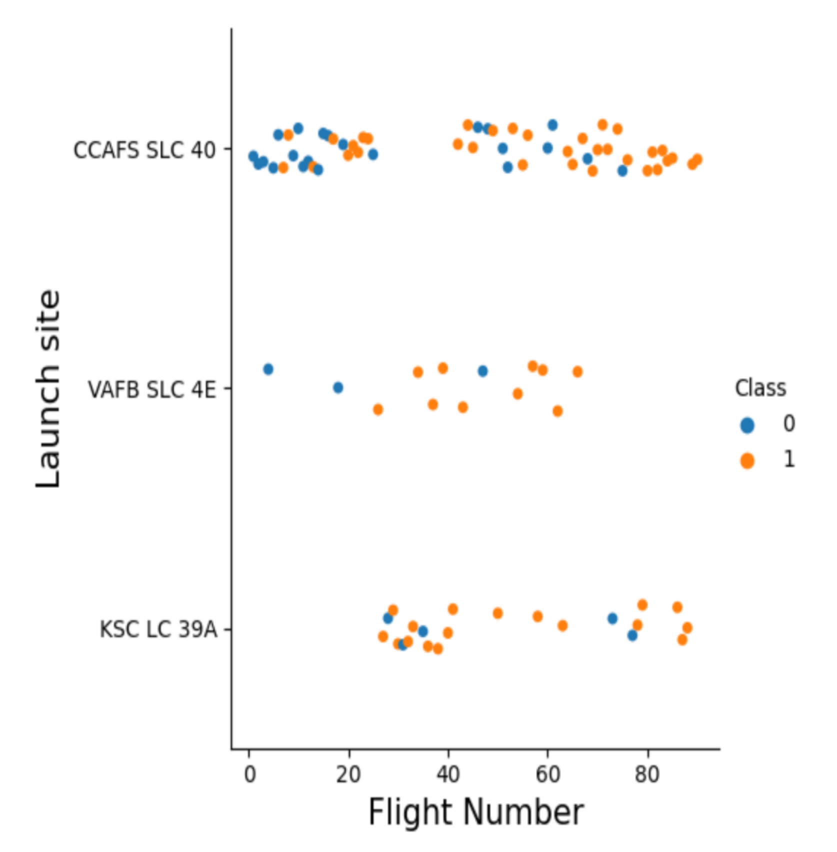



Section 2

# Insights drawn from EDA

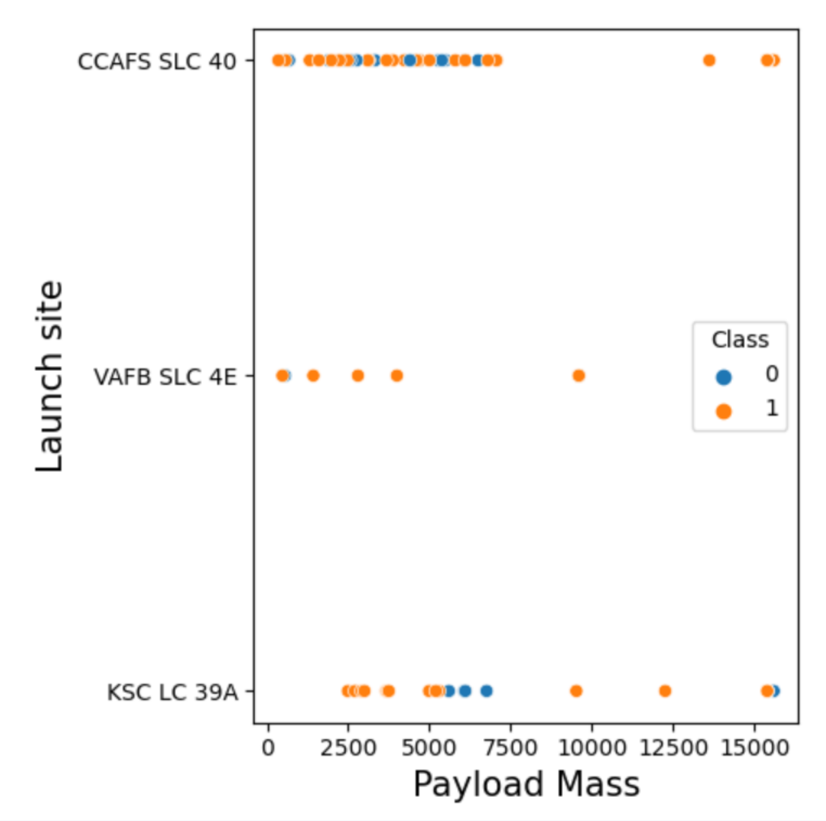
# Flight Number vs. Launch Site

- Most launches happen at CCAFS SLC 40
- Significat increase of success rate over time
- KSC LC 39A with the highest success rate



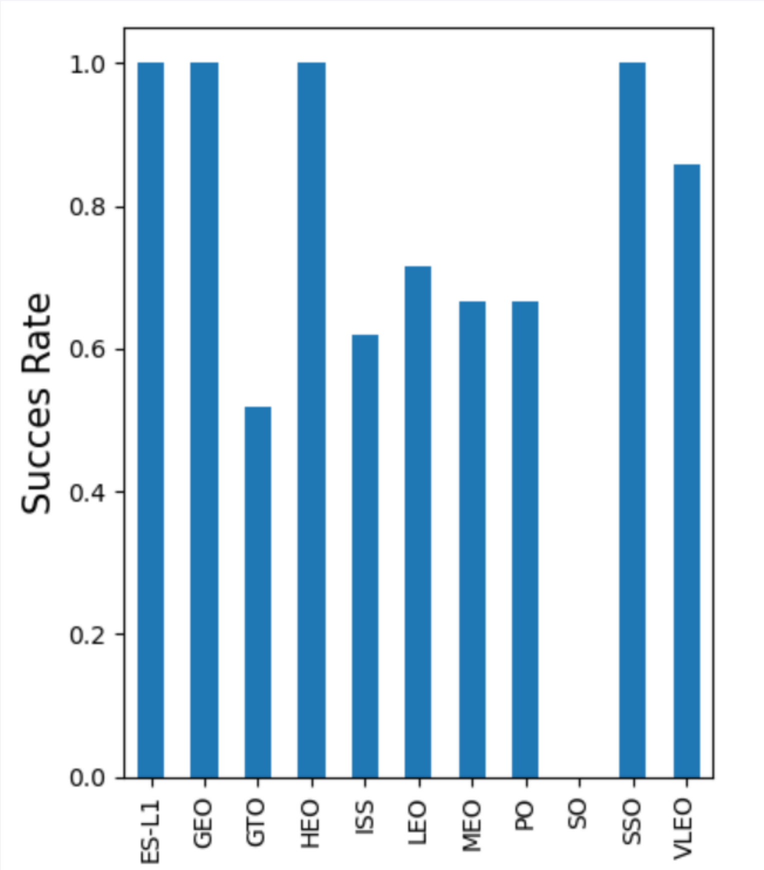
# Payload vs. Launch Site

- Maximum payload only achievable at KSC LC 39A and CCAFS SLC 40
- Higher payloads seem to have a higher chance of successful outcome



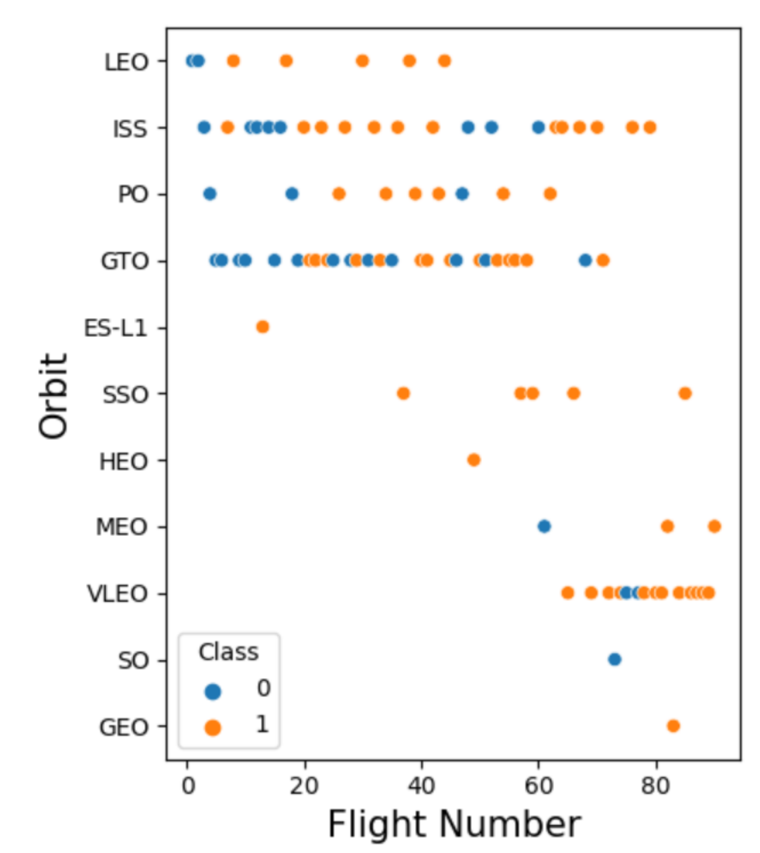
# Success Rate vs. Orbit Type

- Orbits with highest success rate:
  - ES-L1
  - GEO
  - HEO
  - SSO
- Lowest
  - GTO at about 50%



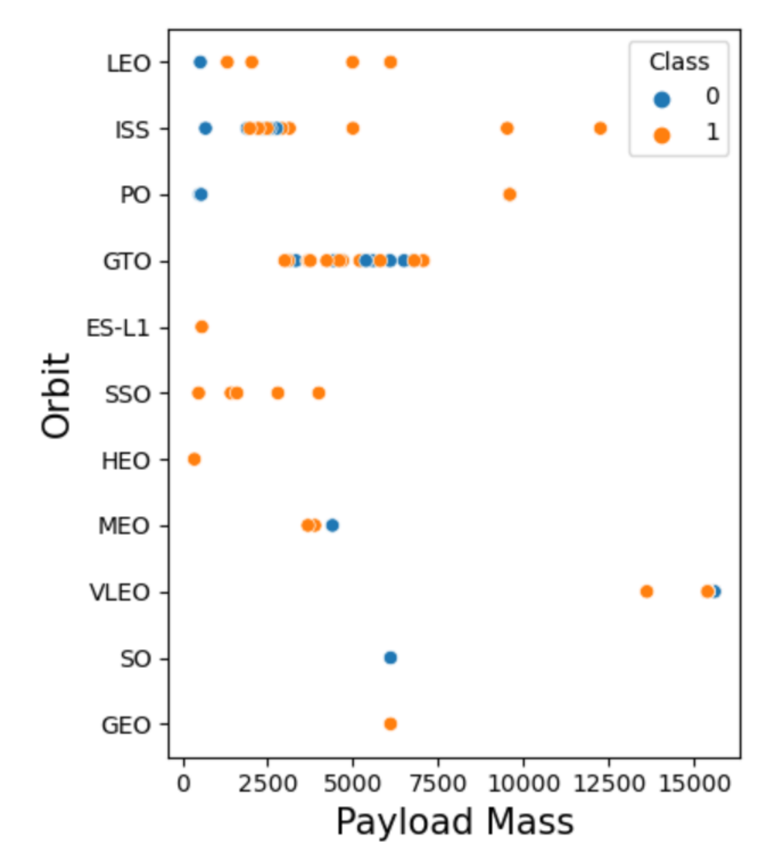
# Flight Number vs. Orbit Type

- Increase of success rate to every orbit type over the years
- Some even with 100%



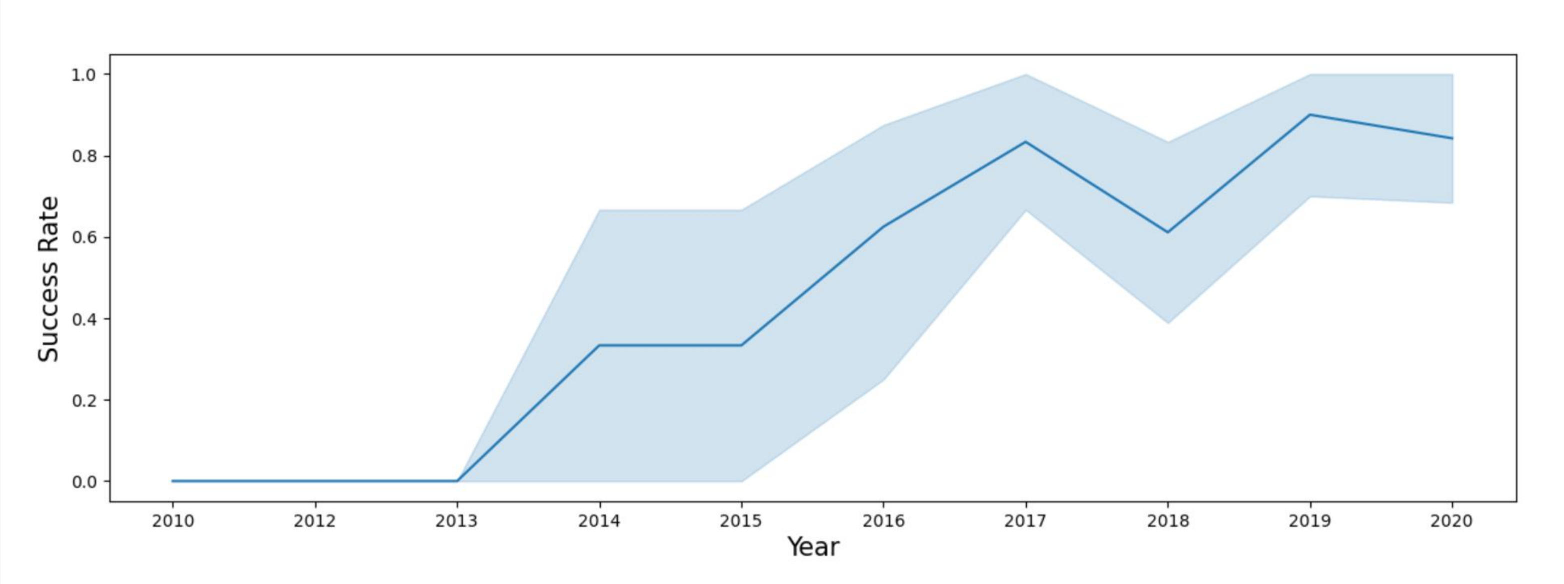
# Payload vs. Orbit Type

- Higher payload mass seems to to have a better chance reaching desired orbit
- With the exeption of GTO



# Launch Success Yearly Trend

Increasing success rate from 2013 to 2020



# All Launch Site Names

- There are four distivt Launch sites in the dataset

```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTBL;
```

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

# Launch Site Names Begin with 'CCA'

- Five entrees of CCAFS LC-40

%%sql SELECT *

FROM SPACEXTBL

WHERE Launch_Site like 'CCA%'

LIMIT 5;

| Date | Time (UTC) | Booster_Version | Launch_Site | Payload | PAYLOAD_MASS__KG_ | Orbit | Customer | Mission_Outcome | Landing _Outcome |
|---|---|---|---|---|---|---|---|---|---|
| 04-06-2010 | 18:45:00 | F9 v1.0 B0003 | CCAFS LC-40 | Dragon Spacecraft Qualification Unit | 0 | LEO | SpaceX | Success | Failure (parachute) |
| 08-12-2010 | 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) |
| 22-05-2012 | 07:44:00 | F9 v1.0 B0005 | CCAFS LC-40 | Dragon demo flight C2 | 525 | LEO (ISS) | NASA (COTS) | Success | No attempt |
| 08-10-2012 | 00:35:00 | F9 v1.0 B0006 | CCAFS LC-40 | SpaceX CRS-1 | 500 | LEO (ISS) | NASA (CRS) | Success | No attempt |
| 01-03-2013 | 15:10:00 | F9 v1.0 B0007 | CCAFS LC-40 | SpaceX CRS-2 | 677 | LEO (ISS) | NASA (CRS) | Success | No attempt |

# Total Payload Mass

- Sum of payload mass delivered for NASA

```
%%sql SELECT SUM(PAYLOAD_MASS__KG_)

FROM SPACEXTBL

WHERE Customer = 'NASA (CRS)’;
```

| SUM(PAYLOAD_MASS__KG_) |
|---:|
| 45596 |

# Average Payload Mass by F9 v1.1

- Average payload mass delivered by F9 v1.1 booster rocket

%%sql SELECT AVG(PAYLOAD_MASS__KG_)

FROM SPACEXTBL

WHERE Booster_Version = 'F9 v1.1';

| AVG(PAYLOAD_MASS__KG_) |
|---|
| 2928.4 |

# First Successful Ground Landing Date

- First successful landing on ground pad

```
%%sql SELECT MIN(Date)

FROM SPACEXTBL

WHERE "Landing _Outcome" = 'Success (ground pad)'
```

| MIN(Date) |
| --- |
| 01-05-2017 |

# Successful Drone Ship Landing with Payload between 4000 and 6000

- Booster versions with successful landings on drone ship with payload between 4000 and 6000 kg

```
%%sql SELECT Booster_Version

FROM SPACEXTBL

WHERE "Landing _Outcome" = 'Success (drone ship)'

AND PAYLOAD_MASS__KG_ between 4000 and 6000;
```

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

# Total Number of Successful and Failure Mission Outcomes

- Total number of mission outcomes

```
%%sql SELECT Mission_Outcome, count(*)

FROM SPACEXTBL

GROUP BY Mission_Outcome;
```

| Mission_Outcome | count(*) |
|---:|---:|
| Failure (in flight) | 1 |
| Success | 98 |
| Success | 1 |
| Success (payload status unclear) | 1 |

# Boosters Carried Maximum Payload

- Boosters capable of delivering maximum payload

```
%%sql SELECT Booster_Version, PAYLOAD_MASS__KG_

FROM SPACEXTBL

WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)

GROUP BY Booster_Version;
```

| Booster_Version | PAYLOAD_MASS__KG_ |
|---|---|
| F9 B5 B1048.4 | 15600 |
| F9 B5 B1048.5 | 15600 |
| F9 B5 B1049.4 | 15600 |
| F9 B5 B1049.5 | 15600 |
| F9 B5 B1049.7 | 15600 |
| F9 B5 B1051.3 | 15600 |
| F9 B5 B1051.4 | 15600 |
| F9 B5 B1051.6 | 15600 |
| F9 B5 B1056.4 | 15600 |
| F9 B5 B1058.3 | 15600 |
| F9 B5 B1060.2 | 15600 |
| F9 B5 B1060.3 | 15600 |

# 2015 Launch Records

- Landing_outcomes on drone ship with booster version and launch site name for 2015

```
%%sql SELECT substr(Date, 4, 2) AS “Month of year”, "Landing _Outcome", Booster_Version, Launch_Site

FROM SPACEXTBL

WHERE "Landing _Outcome" = 'Failure (drone ship)'

AND substr(Date,7,4) = '2015’;
```

| Month of year | Landing _Outcome | Booster_Version | Launch_Site |
|---|---|---|---|
| 01 | Failure (drone ship) | F9 v1.1 B1012 | CCAFS LC-40 |
| 04 | Failure (drone ship) | F9 v1.1 B1015 | CCAFS LC-40 |

Workaround due to use of sqlite

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

- Ranking of successful outcomes

```
%%sql SELECT "Landing _Outcome", count(*)

FROM SPACEXTBL

WHERE Date BETWEEN '2010-06-04' and '2017-03-20"

AND "Landing _Outcome" LIKE 'Success%'

GROUP BY "Landing _Outcome"

ORDER BY count(*) DESC;
```

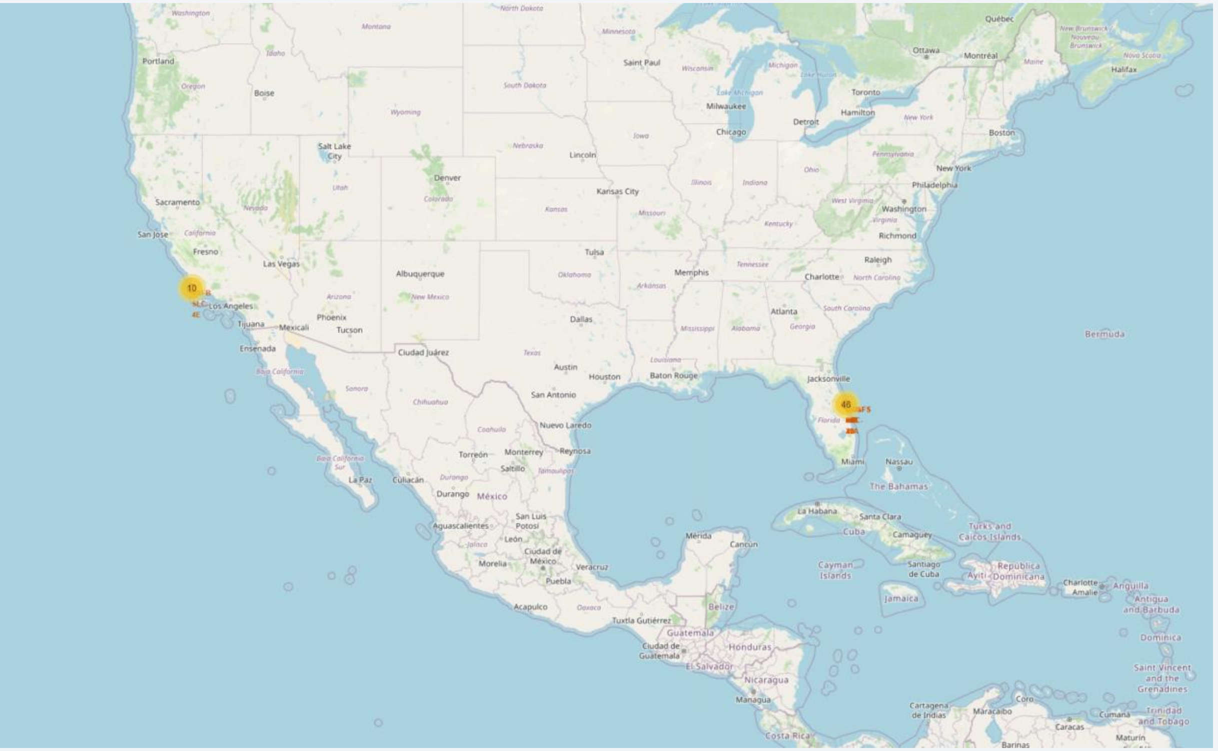
| Landing _Outcome | count(*) |
|---|---|
| Success (ground pad) | 5 |
| Success (drone ship) | 5 |

Needed to convert the date to datetime - df['Date'] = pd.to_datetime(df['Date']) before df.to_sql for the BETWEEN operation to work properly

Section 3

# Launch Sites Proximities Analysis

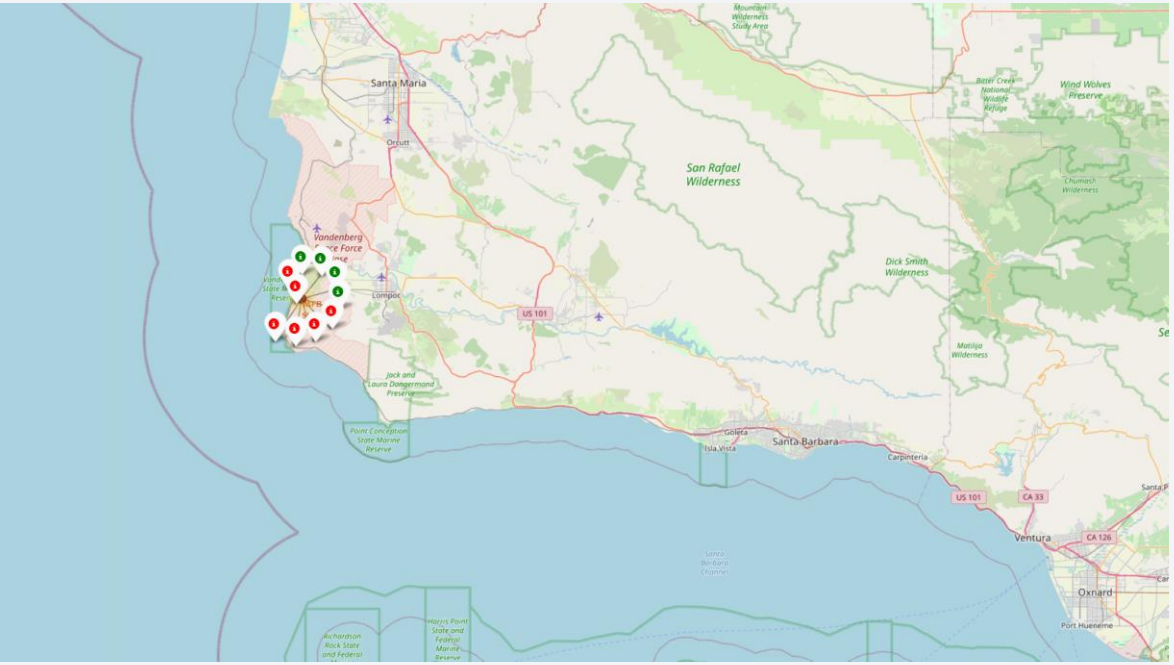
# Overview location of launch sites

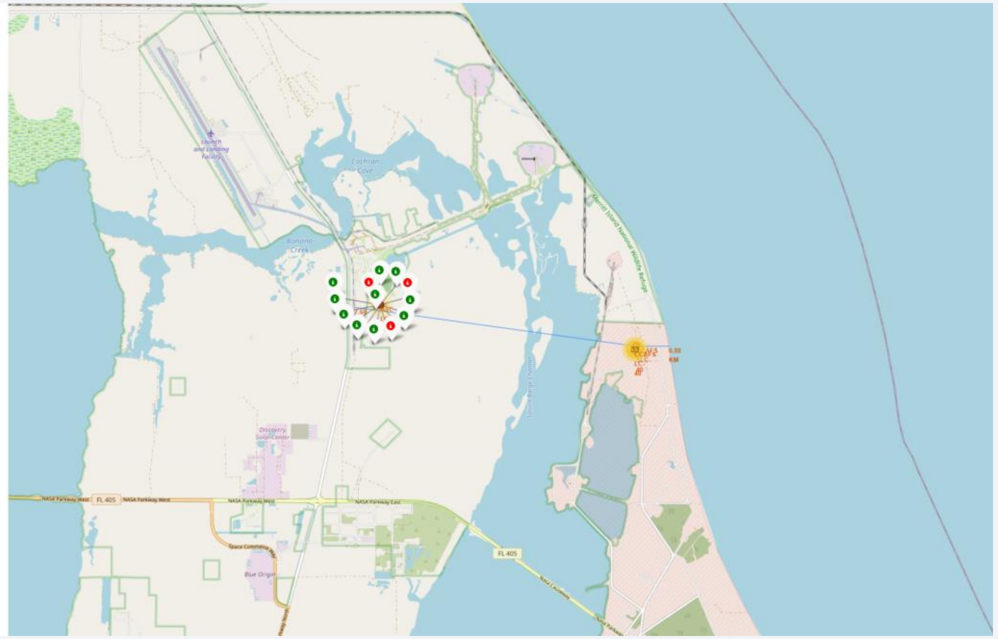



Launch site being on the coasts for safety reasons

# Launch outcomes by sites

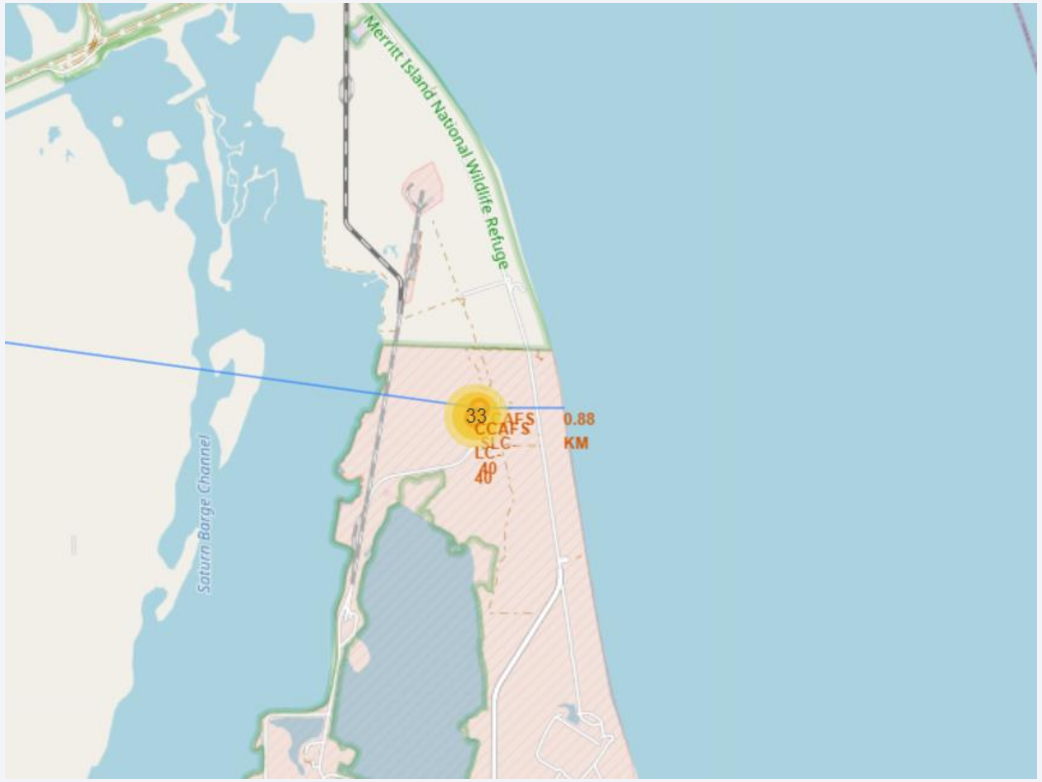
On the right we have KSC LC-39A with higher success rate indicated by the green markers



# Launch site safety and logistics

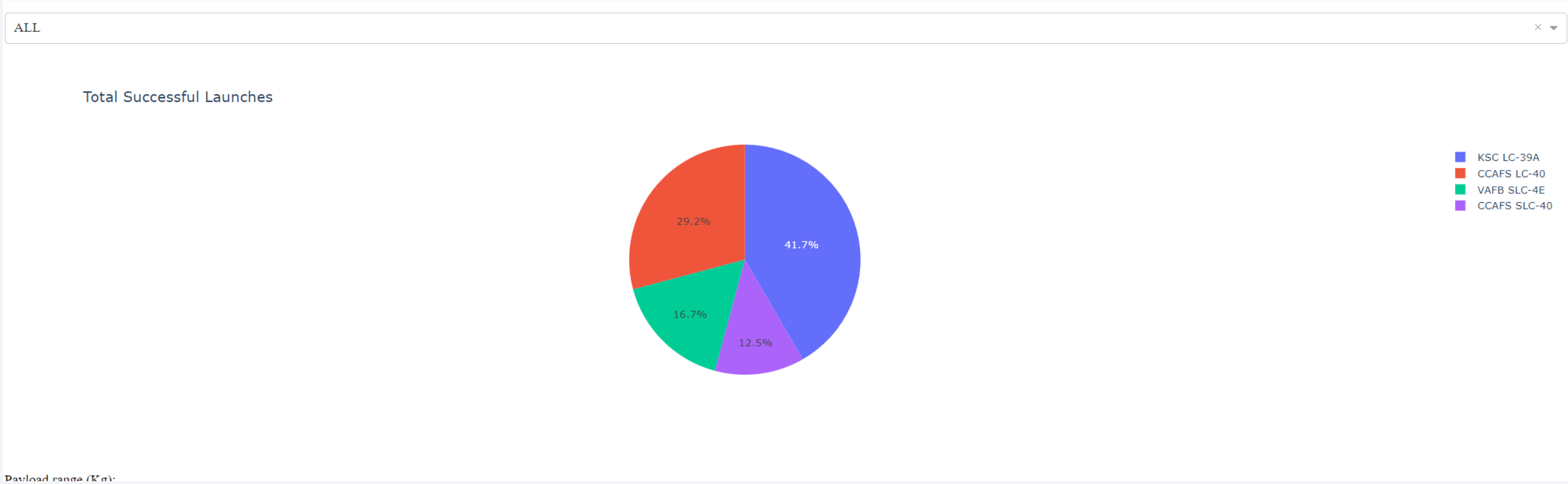
- All launch sites are near coastlines – (1-8km)
- Highways and Railways are near as well for logistics reasons – (1-6km)
- For safety launch sites tend to be 14-18km away from cities



Section 4

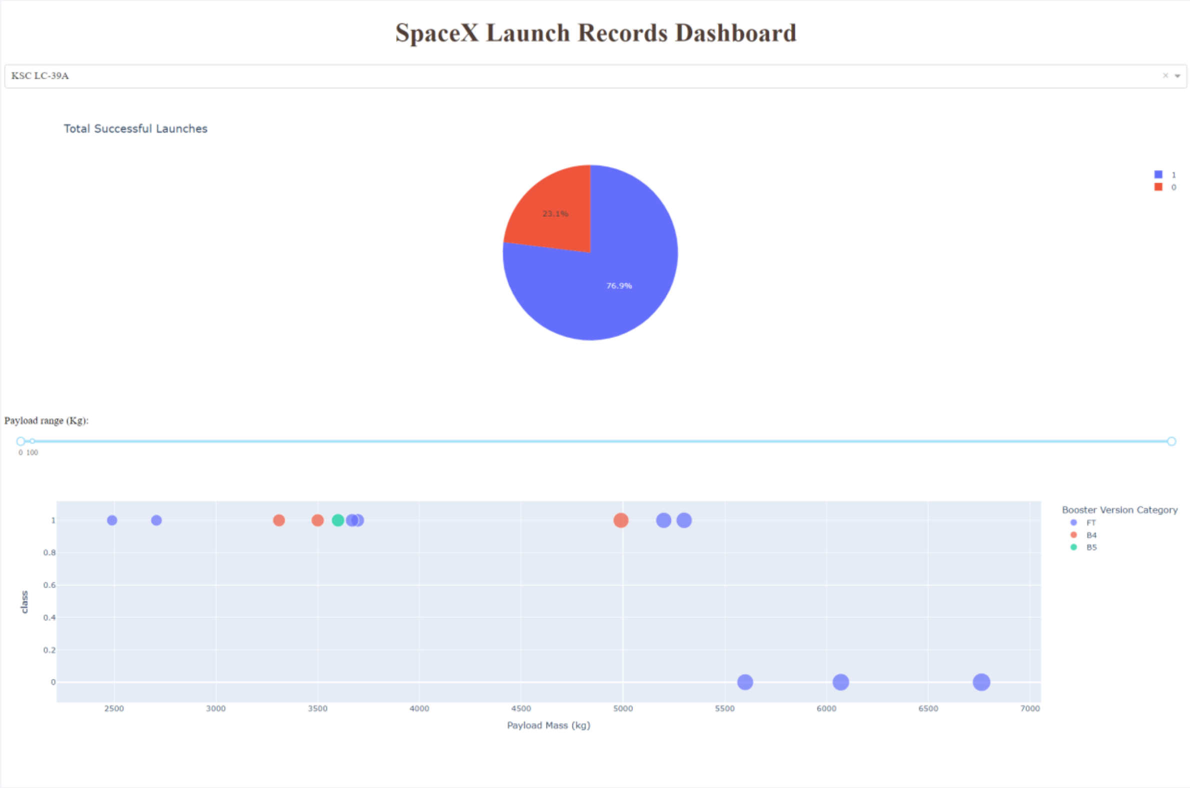
# Build a Dashboard with Plotly Dash

# Successful Launches by Site



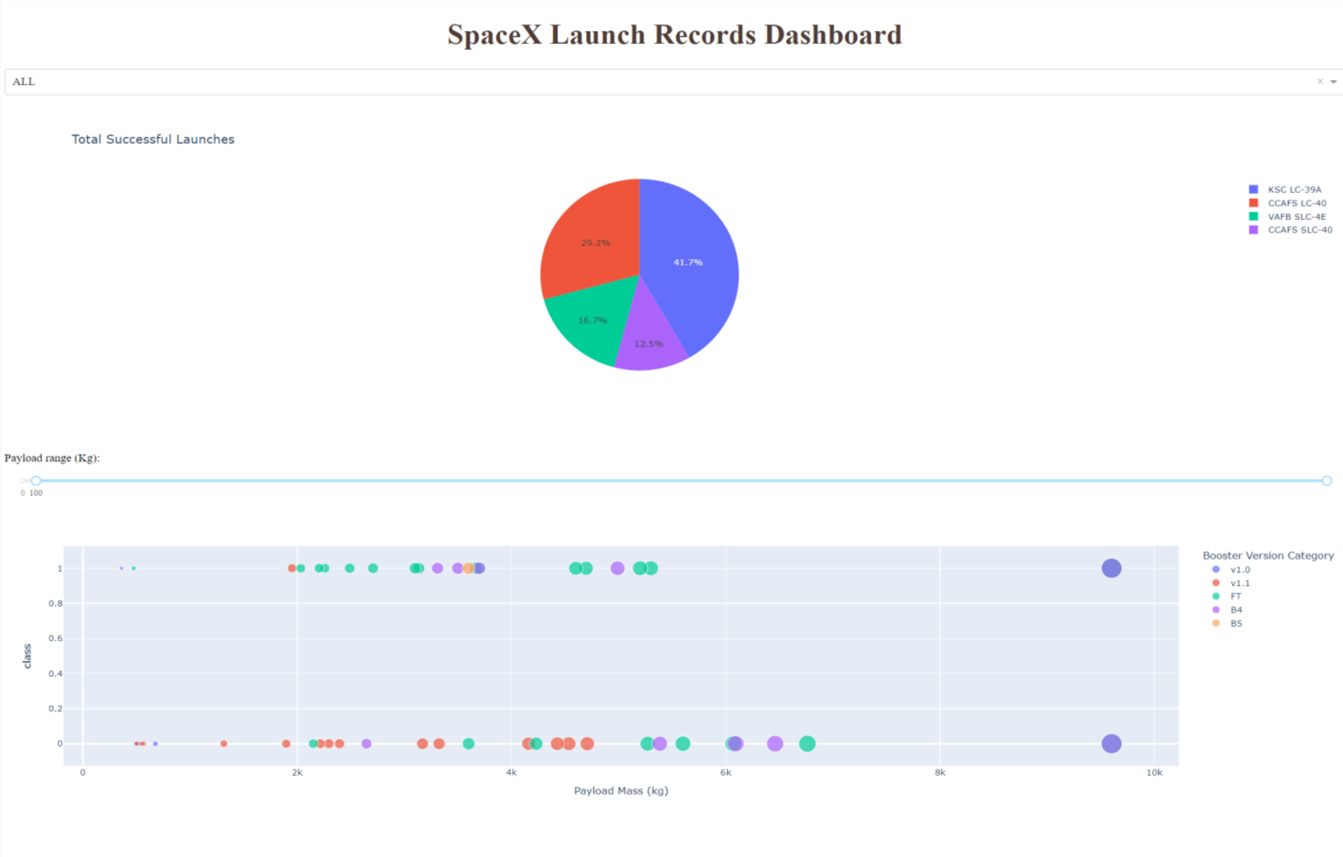

Successrate for all site showing KSC LC 39A is the best option for a successful start

# Launch Success Ratio for KSC LC 39A




Further exploration into the success rate of KSC LC 39A showing 76,9 % successful launches

# Payload vs. Launch Outcome




FT4 booster with low to moderate payload seems to be a very good setting

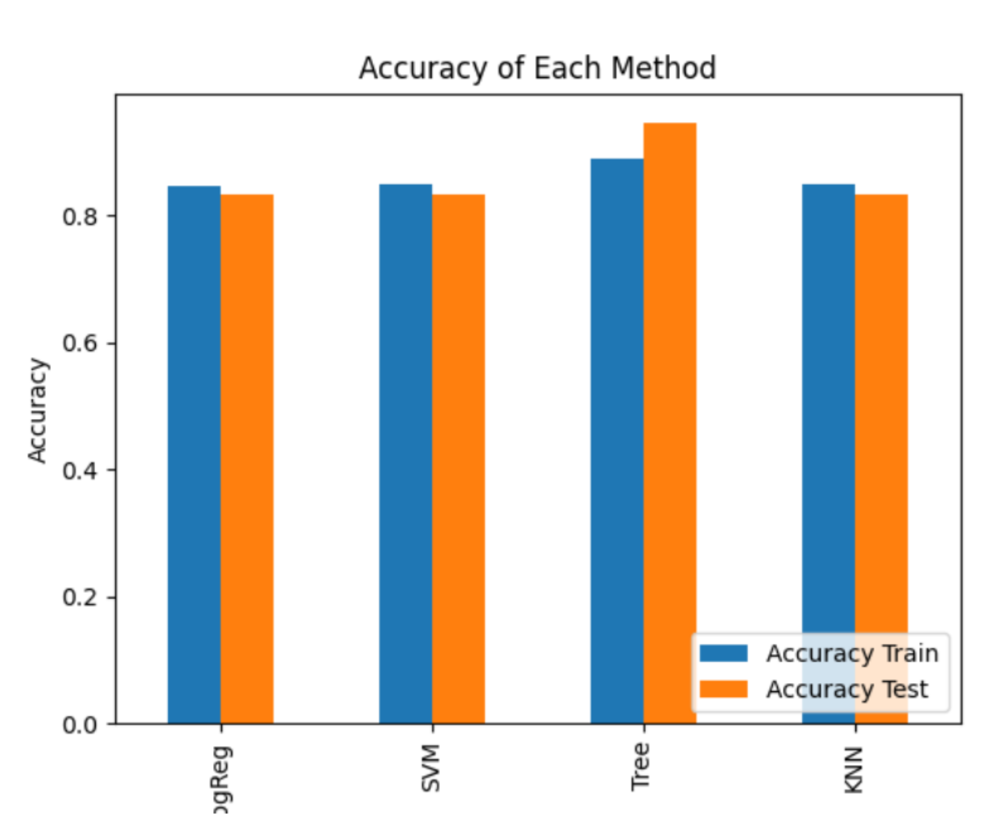
Section 5

# Predictive Analysis (Classification)

# Classification Accuracy

- Evaluation of accuracy with training and testing date clearly shows the higher performance of the decision tree classifier

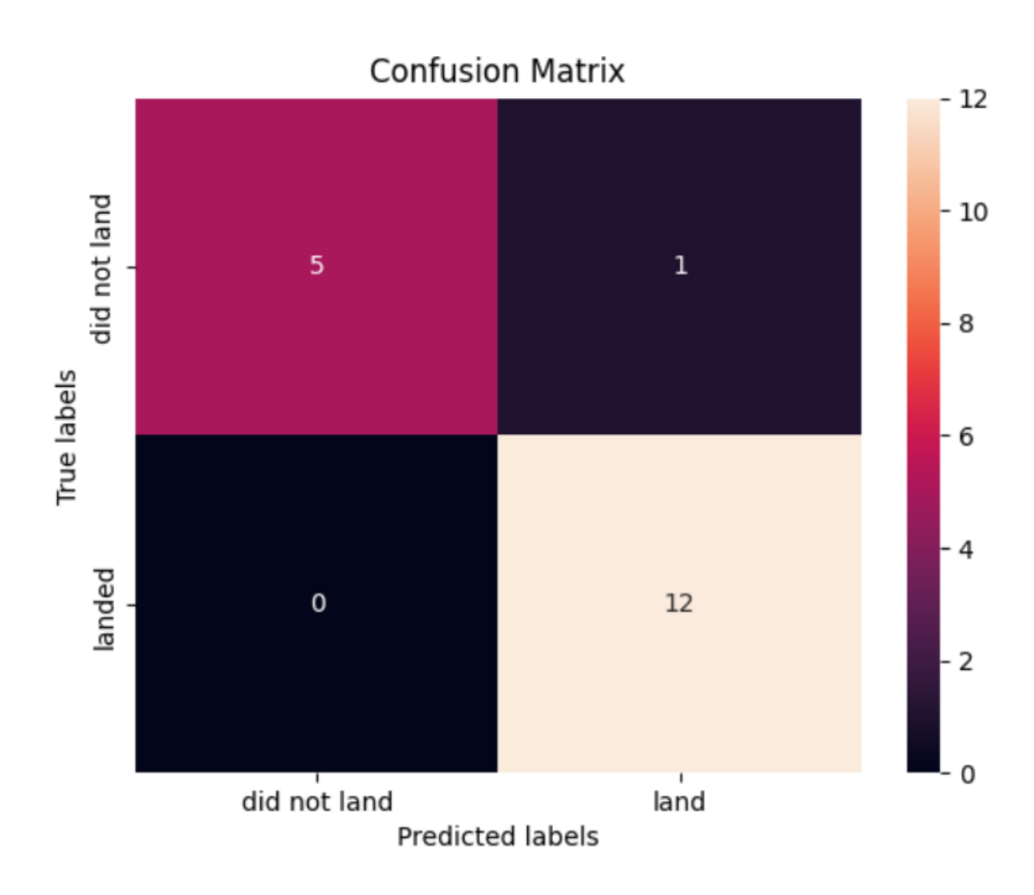
All other models have the same performance



# Confusion Matrix

Confusion matrix showing the prediction and actual labels of the data using the decision tree classifier

With one out of 18 misscalculated



# Conclusions

- Significant increases in recovery rate of booster since 2015
- To further cut costs of launches, every booster should be tried to recover
  - No attempt should be no option
- Most successful launch site is KSC LC-39A
  - Further invesigation is required for why
  - Focus on this site for future launches
- More focus on successful orbits
  - GEO, HEO, SSO, ES-L1 should be favored

# Thank you!