# Winning Space Race with Data Science

Eric Vierzigmann
15.01.2024

# Outline

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

# Executive Summary

Methodologies

- Explorative Data Analysis
  - SQL Queries
  - Visualization
- Geographical Visualization
- Interactiv Visualization
- Training of different Machine Learning Models

Results

- Correlation between various features and success rate
- Predicting of successful landing possible
- Training and Test Data to small

# Introduction

Goal: We want to outperform SpaceX

Question: Can we predict the cost of a rocket launch from SpaceX?

- The cost of a launch highly dependes on the reusage of the first stage
- Can we predict whether the first stage of a rocket will land successfully?

Section 1

# Methodology

# Methodology

## Executive Summary

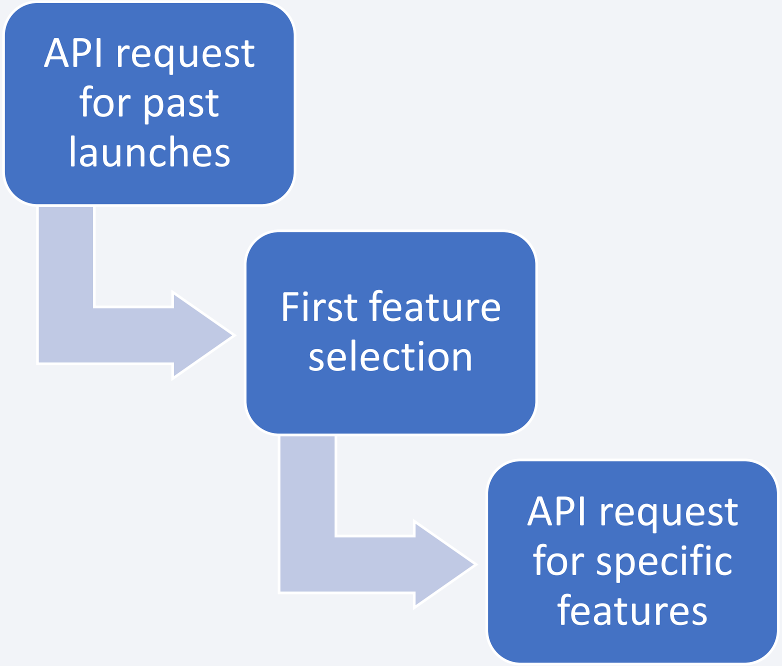
- Data collection methodology:
  - API requests & webscrapping
- Perform data wrangling
  - Create binary outcome class
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Linear Regression, SVM, Decision Tree, kNN
  - Gridsearch for best Hyperparameters

# Data Collection

- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts

# Data Collection – SpaceX API

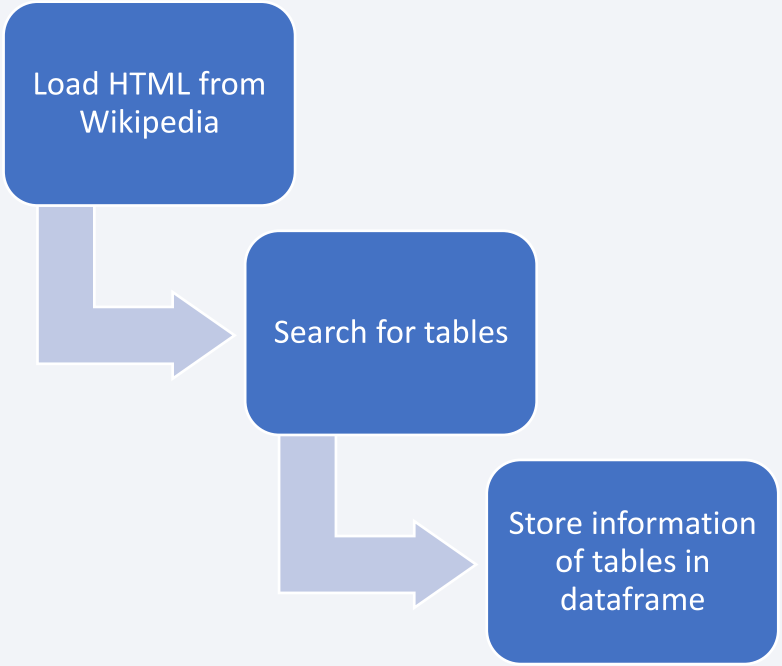
- Loading data for past launches
- Selecting relevant features and feature IDs
  - Rocket ID
  - Payload ID
  - Launchpad ID
  - Core ID
- Loading additional data by using the IDs
- Storing the data in a dataframe




GitHub: [CourseraFinalProject/01_jupyter-labs-spacex-data-collection-api.ipynb at main · dashm00n/CourseraFinalProject (github.com)]

# Data Collection - Scraping

- Loading HTML file from Wikipedia
- Creating BeautifulSoup object
- Searching for tables (name='table')
- Searching for columns (name='th')
    - Drop unnecessary columns
- Iterating over rows and storing information in a dataframe




Github: [CourseraFinalProject/02_jupyter-labs-webscraping.ipynb at main · dashm00n/CourseraFinalProject (github.com)]

# Data Wrangling

- Getting first insights by searching for null values and counting unique entries
- Generating an output variable to be predicted
  - Landing outcome successful (1) or not (0)

Collect unique landing outcomes → Map success or failure to each outcome → New column with a binary landing outome

Github: [CourseraFinalProject/03_labs-jupyter-spacex-Data wrangling.ipynb at main · dashm00n/CourseraFinalProject (github.com)]

# EDA with Data Visualization

To get a rough overview, some features and their correlation are shown in graphs

- Scatter plots of all launches with landing outcome as label
  - Compared features:
    - Launch Site
    - Payload Mass (kg)
    - Flight Number
    - Orbit
- Success Rate per Orbit in a bar chart

Github: [CourseraFinalProject/05_jupyter-labs-eda-dataviz.ipynb at main · dashm00n/CourseraFinalProject (github.com)]

# EDA with SQL

- Get unique launch sites
- Get first 5 entires of CCAFS launch sites
- Get total payload that was transported for NASA (CRS)
- Get average payload that was transported with Booster Version F9 v1.1
- Get unique landing outcomes
- Get date first successful landing on a ground pad
- Get Booster Versions that landed successfully on a drone shipe having a payload mass between 4000 and 6000kg
- Get the number of each unique mission outcome
- Get Booster Versions that carried the maximum payload mass

Github: [CourseraFinalProject/04_jupyter-labs-eda-sql-coursera_sqllite.ipynb at main · dashm00n/CourseraFinalProject (github.com)]

# Build an Interactive Map with Folium

- A folium map was created for the geographical representation of the launches and the launch sites, this contains the following objects:
  - Markers
  - Icons
  - Lines
  - Line-Text
- These objects can be used to determine insights such as the distances to certain land features, other launch sites and more

Github: CourseraFinalProject/06_lab_jupyter_launch_site_location.ipynb at main · dashm00n/CourseraFinalProject (github.com)

# Build a Dashboard with Plotly Dash

- A pie chart was created to show the distribution of starts per launch sites
- A pie chart with the success rate can also be displayed for each launch site
- A scatterplot was created for the correlation between landing outcome and payload mass
    - Display of all launch sites
    - Display of a specific launch site
    - Display of the respective booster version per launch
    - Choosing of the displayed payload range

Github: CourseraFinalProject/06_lab_jupyter_launch_site_location.ipynb at main · dashm00n/CourseraFinalProject (github.com)

# Predictive Analysis (Classification)

- For classification whether a landing would be successful, 4 types of models were analysed
  - Linear Regression
  - Support Vector Machine
  - Decision Tree
  - K Nearest Neighbors
- The hyperparameter for each model were tuned using Gridsearch


Github: [CourseraFinalProject/08_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb at main · dashm00n/CourseraFinalProject (github.com)]

# Results

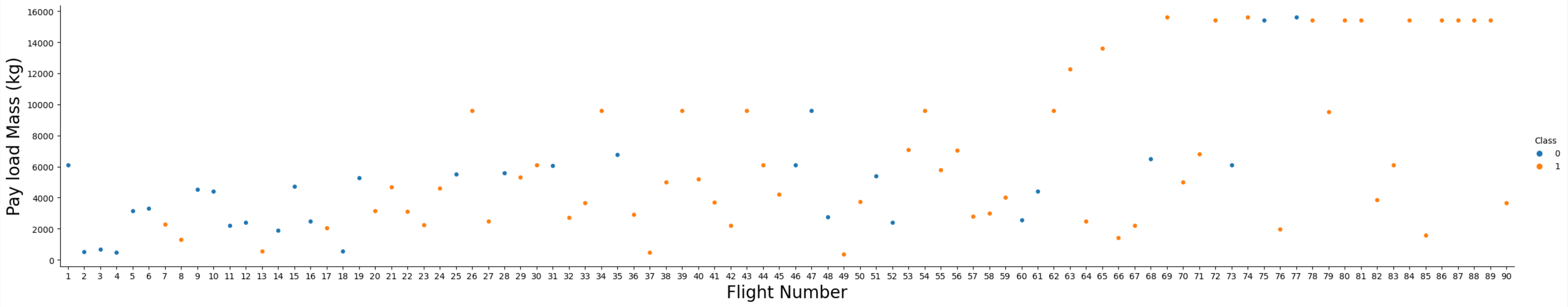
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

# Insights drawn from EDA

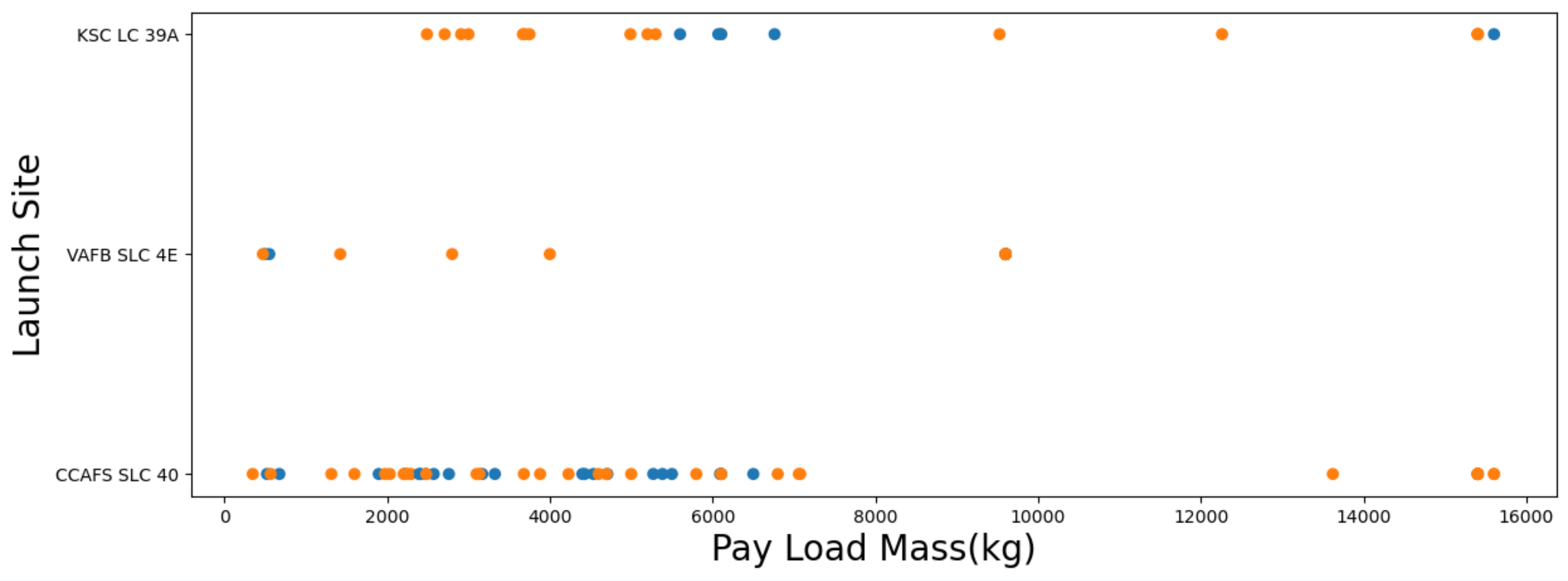
# Flight Number vs. Launch Site

- The payload mass increases with larger flight number
- The success rate of landing outcomes increases with larger flight number



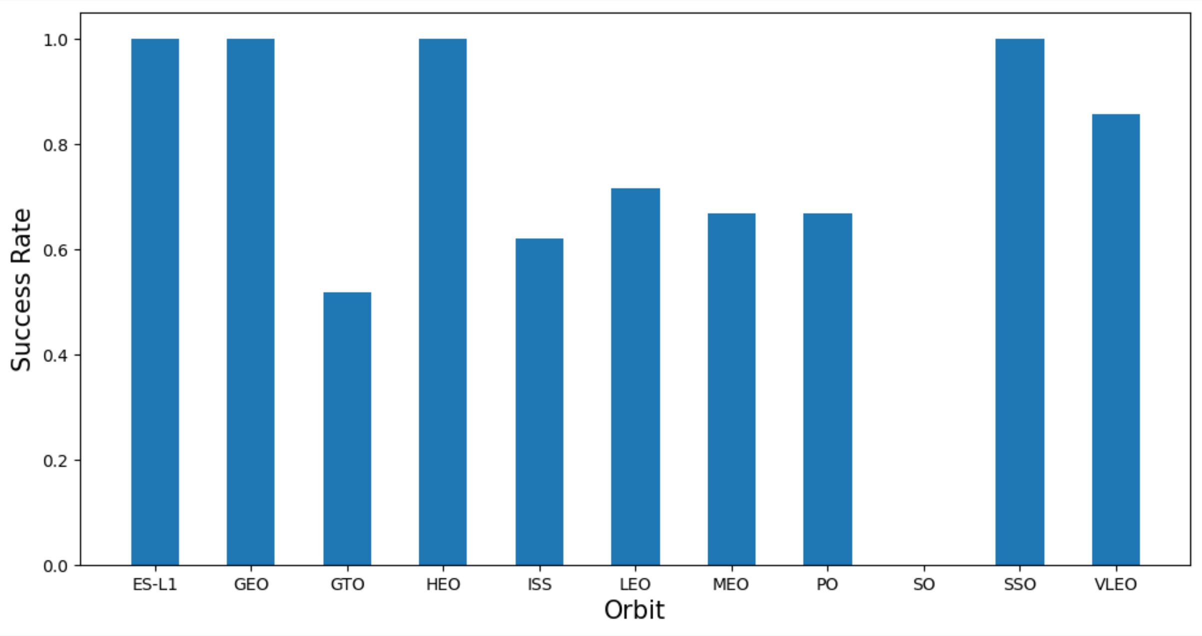
# Payload vs. Launch Site

- VAFB SLC 4E dosen't carry payloads heavier than 10000kg



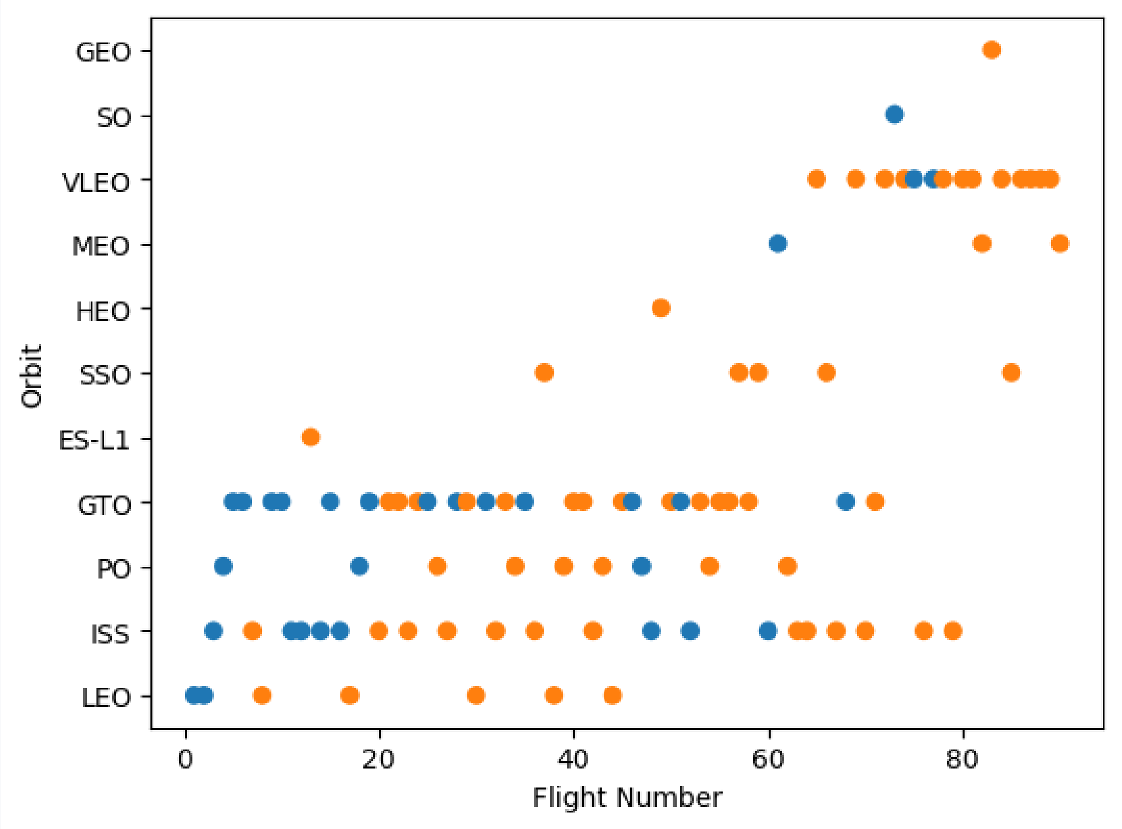
# Success Rate vs. Orbit Type

- ES-L1, GEO, HEO and SSO have a 100% success rate
- No rocket return successfully from SO
- The rockets from GTO, ISS, LEO, MEO and PO have a 50-70% chance of landing successfully



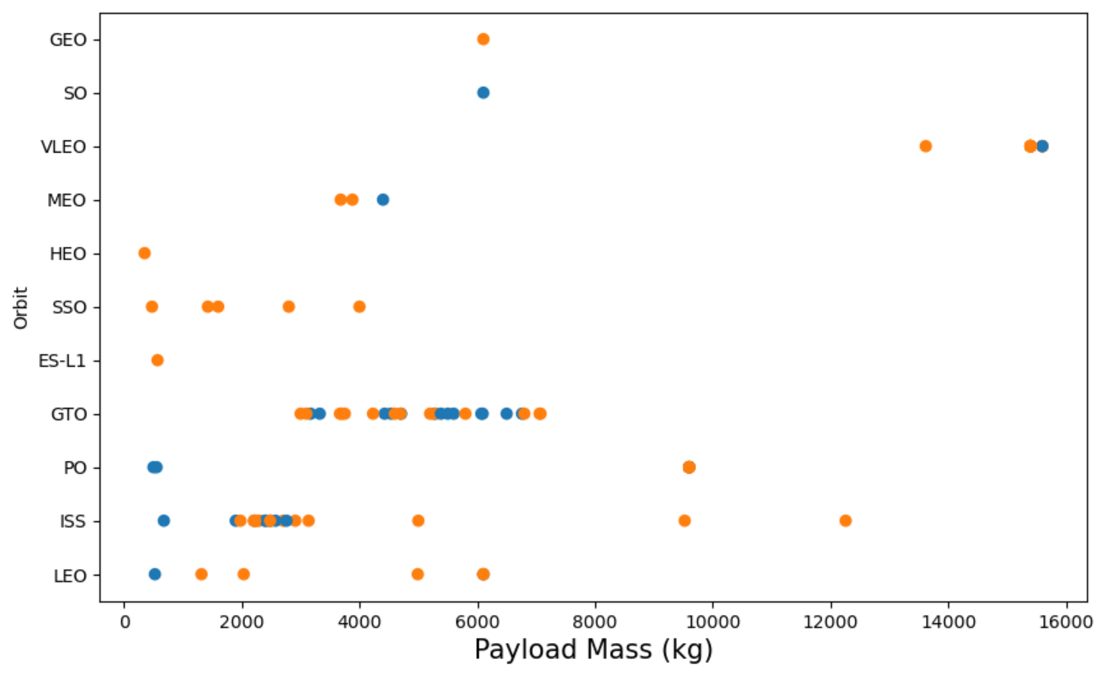
# Flight Number vs. Orbit Type

- In the beginning, rockets only flew to GTO, PO, ISS and LEO
- Launches to GEO only startet recently
- The success rate of LEO increases with the flight number
- There is nearly no correlation between the orbit and the flight number regarding the landing success



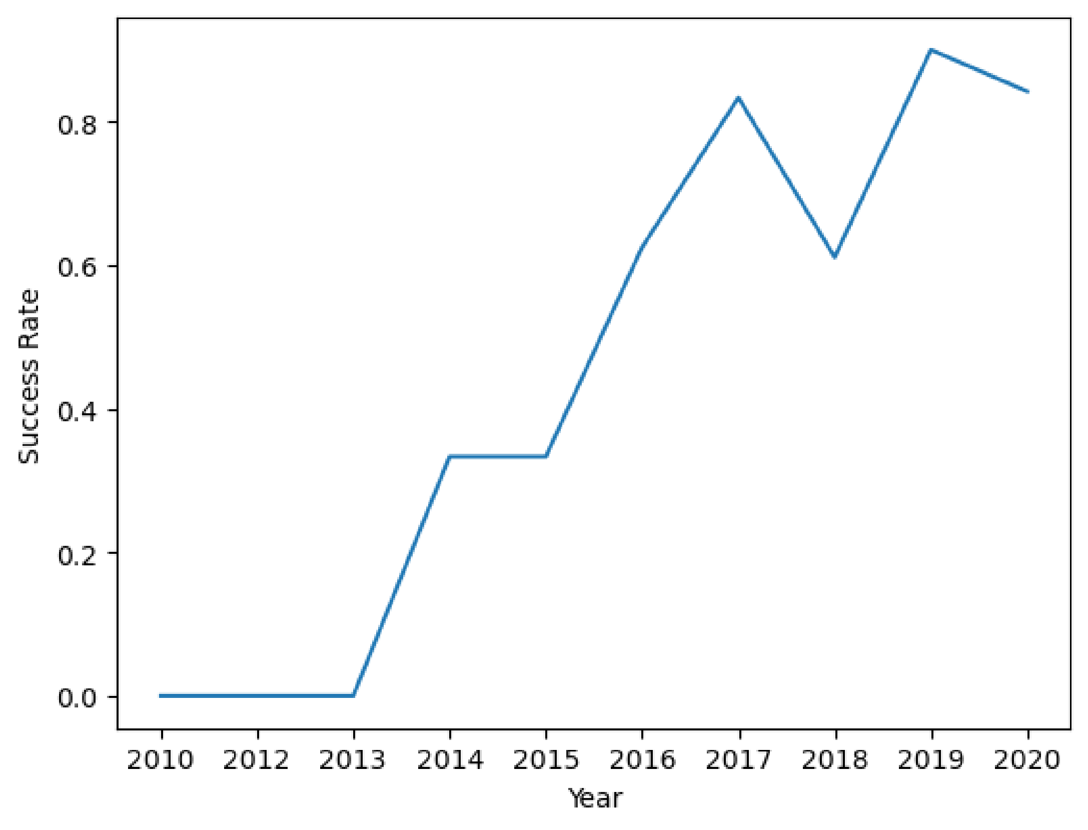
# Payload vs. Orbit Type

- With heavy payloads the positive landing rate are more for Polar, LEO and ISS
- For GTO, there is no correlation between payload and successful landing
- In VLEO, there are only heavy payloads over 14000kg



# Launch Success Yearly Trend

- In total, the success rate is increasing over the years
- It kept constant from 2014-2015
- It shows a drop in the years 2018 and 2021



# All Launch Site Names

- There are 4 different launch sites from SpaceX
  - CCAFS LC-40
  - CCAFS SLC-40
  - VAFB SLC-4E
  - KSC LC-39A

- The launch sites CCAFS LC-40 and SLC-40 are located next to each other

# Launch Site Names Begin with 'CCA'

- In the first 3 years, there were only very few launches from launch sites beginning 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 | 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

Query:

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer LIKE '%NASA (CRS)%’
```

Total amount of payload that NASA (CRS) transported with SpaceX rockets:

| SUM(PAYLOAD_MASS__KG_) |
|---|
| 48213 |

# Average Payload Mass by F9 v1.1

Query:

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version LIKE 'F9 v1.1%’
```

Average amount of payload that rockets with a F9 v1.1 Booster transported:

| AVG(PAYLOAD_MASS__KG_) |
| --- |
| 2534.6666666666665 |

# First Successful Ground Landing Date

Query:

```
%sql SELECT min(Date) FROM SPACEXTABLE WHERE Landing_Outcome='Success (ground pad)’
```

First successful landing on a ground pad:

| min(Date) |
| --- |
| 2015-12-22 |

# Successful Drone Ship Landing with Payload between 4000 and 6000

Query:

```
%sql SELECT Booster_Version, PAYLOAD_MASS__KG_ FROM SPACEXTABLE WHERE Landing_Outcome='Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
```

There are only 4 successful Drone Ship Landing with Payload between 4000 and 6000:

| Booster_Version | PAYLOAD_MASS__KG_ |
|---|---|
| F9 FT B1022 | 4696 |
| F9 FT B1026 | 4600 |
| F9 FT B1021.2 | 5300 |
| F9 FT B1031.2 | 5200 |

# Total Number of Successful and Failure Mission Outcomes

Query:

```
%sql SELECT Mission_Outcome, count(*) FROM SPACEXTABLE GROUP BY Mission_Outcome
```

Almost all missions were successful – there is only one record of a failure during flight:

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

# Boosters Carried Maximum Payload

Query:

```
%sql SELECT Booster_Version, PAYLOAD_MASS__KG_ FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_=(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

Booster Versions that carried the maximum payload of 15600kg:

| 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

Query:

```
%sql SELECT SUBSTR(Date, 6,2), Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE
WHERE SUBSTR(Date, 0, 5) = '2015' and Landing_Outcome LIKE 'failure (drone ship)’
```

Records that show failure of landing on drone ships in the year 2015:

| SUBSTR(Date, 6,2) | 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 |

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

## Query:

```
%sql SELECT Landing_Outcome, count(Landing_Outcome) AS count_outcome FROM SPACEXTABLE
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome ORDER BY count_outcome DESC
```

## Ranking of Landing Outcomes:

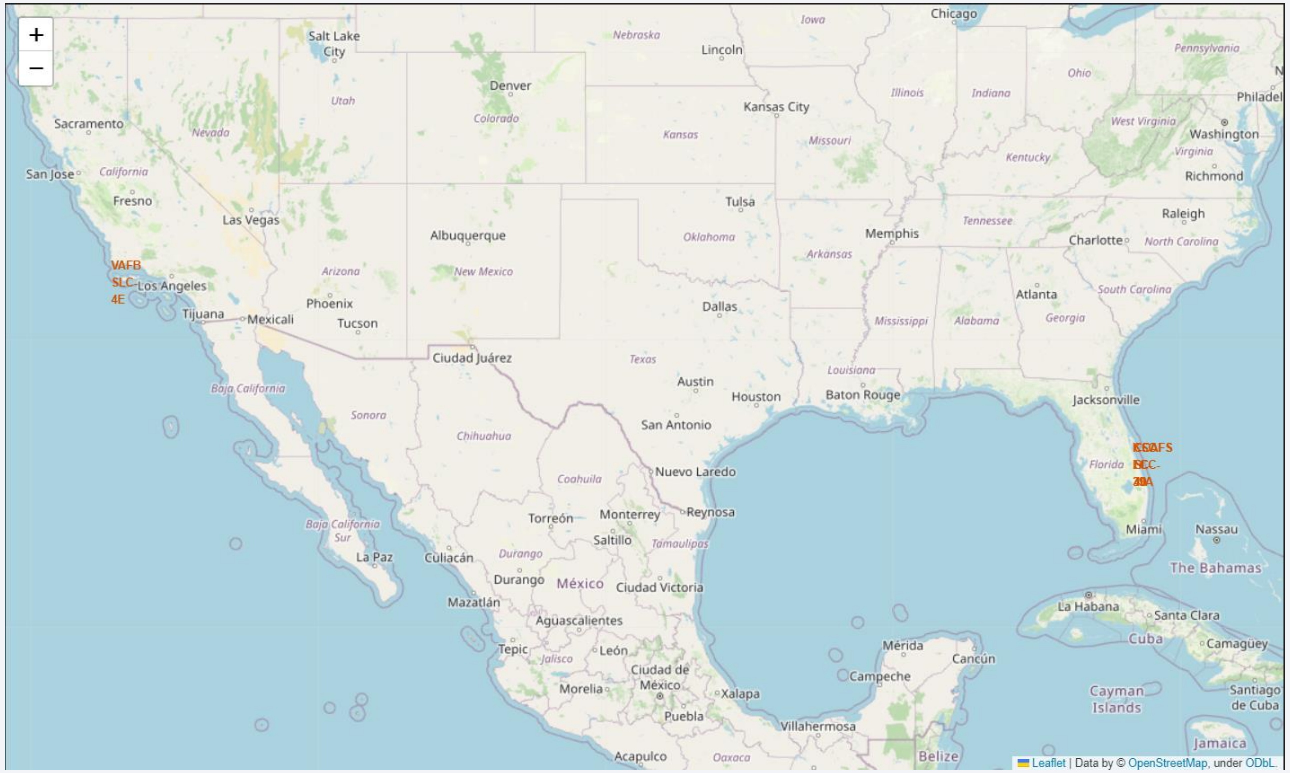
| Landing_Outcome | count_outcome |
|---|---|
| No attempt | 10 |
| Success (drone ship) | 5 |
| Failure (drone ship) | 5 |
| Success (ground pad) | 3 |
| Controlled (ocean) | 3 |
| Uncontrolled (ocean) | 2 |
| Failure (parachute) | 2 |
| Precluded (drone ship) | 1 |

Section 3

# Launch Sites Proximities Analysis

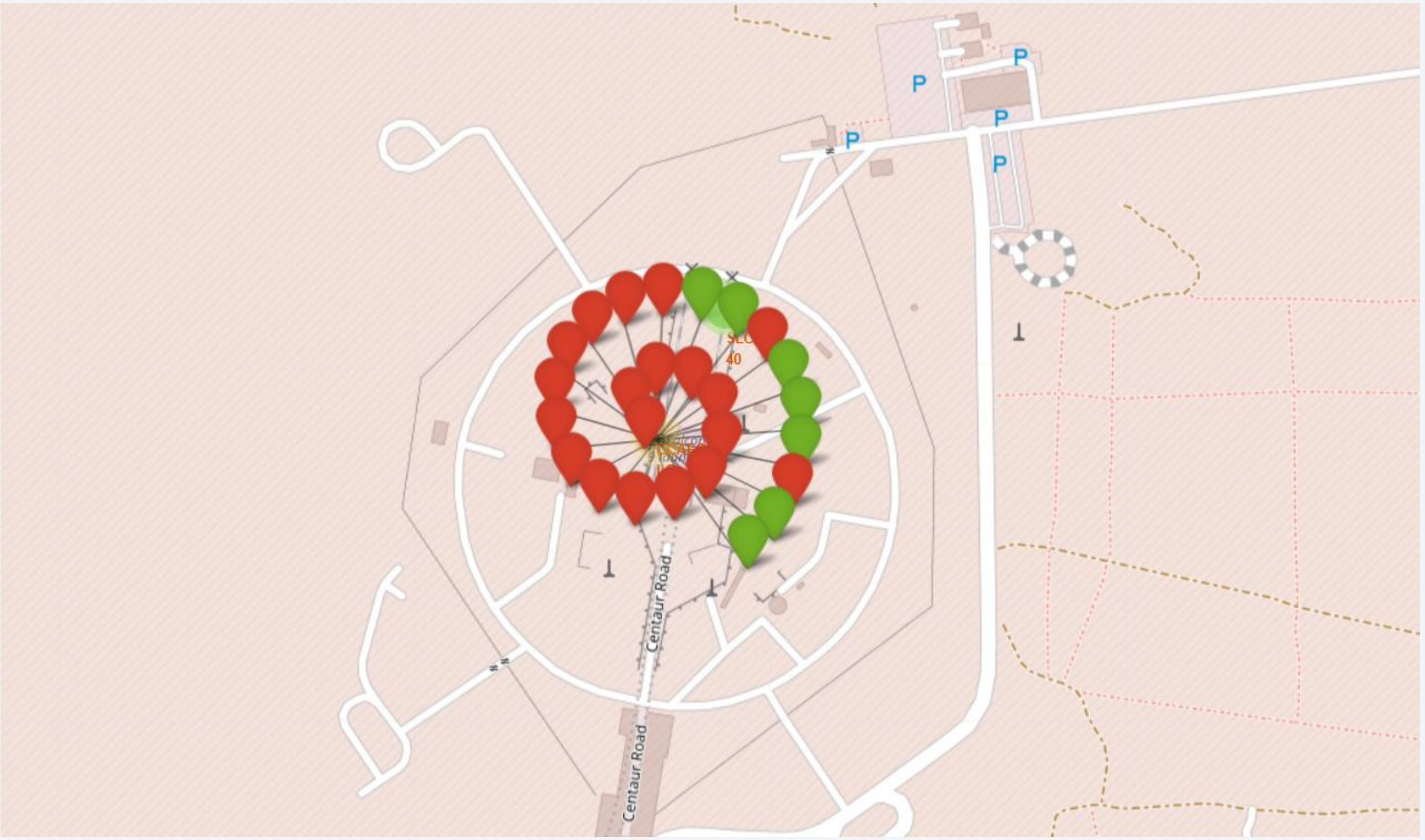
# Launch Site Locations of SpaceX

- All launch sites are at the coast
- 3 out of 4 are located next to each other
- All launch sites are located at a southernmost possible point of the USA
    - Closest to the equator as possible



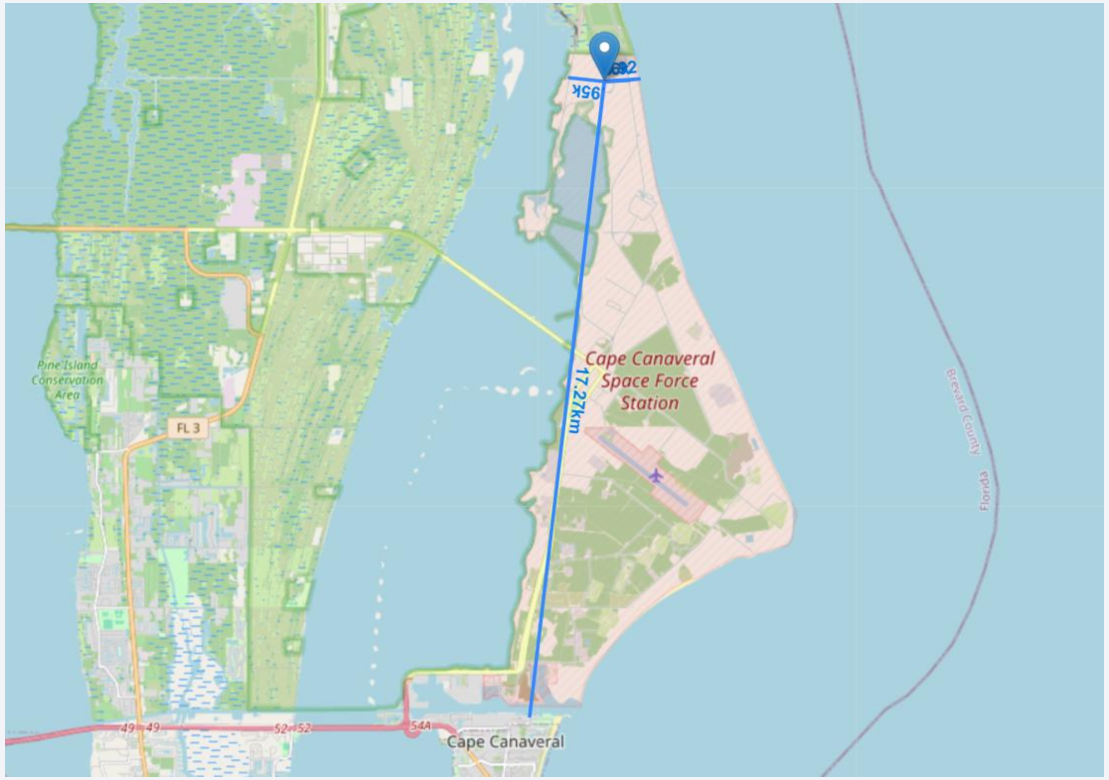
# Landing Outcomes at CCAFS LC-40

- Launches at CCAFS LC-40
- At the beginning, all landings failed
- In recent times, the success rate increased



# Landmarks near CCFAS LC-40

- The launch site CCFAS LC-40 is pretty close to the next road, railway and the coast line (less then 1km)
- It is far away from the next city Cape Canaveral (17.2km)

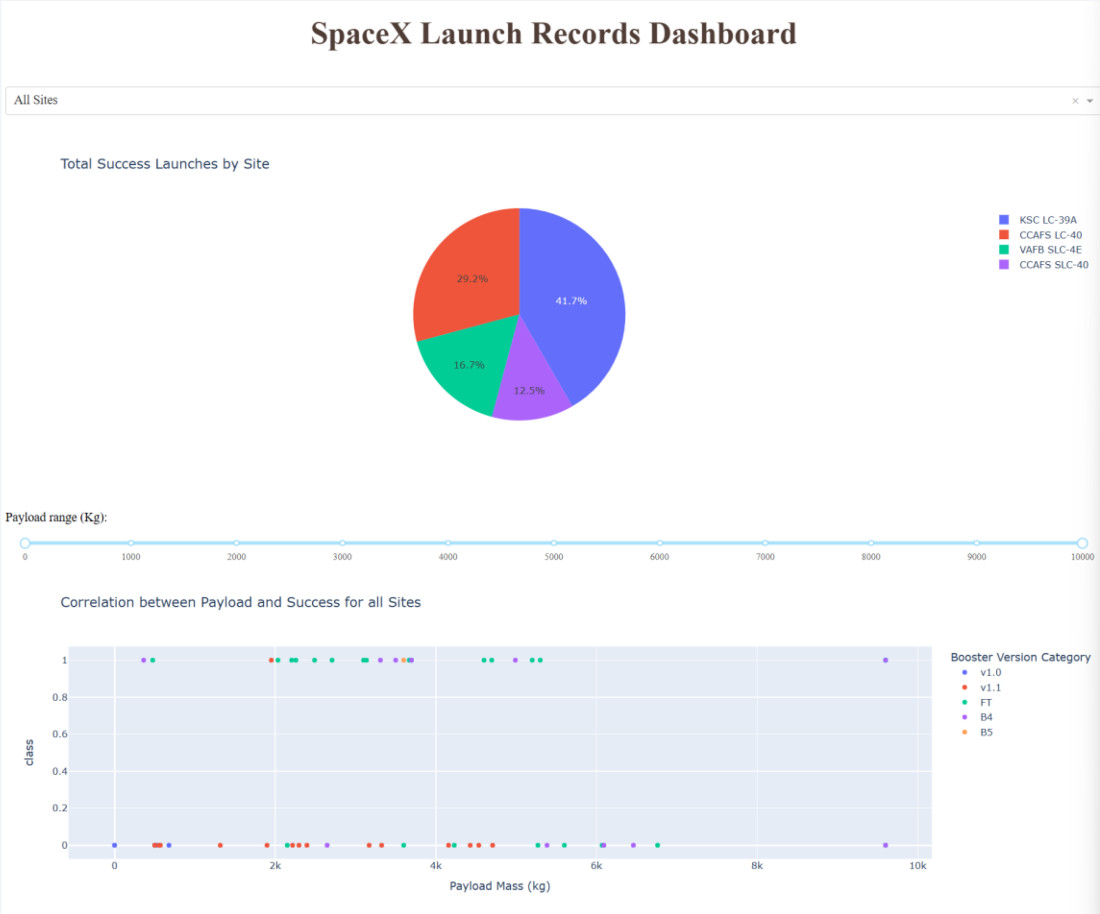


Section 4

# Build a Dashboard with Plotly Dash

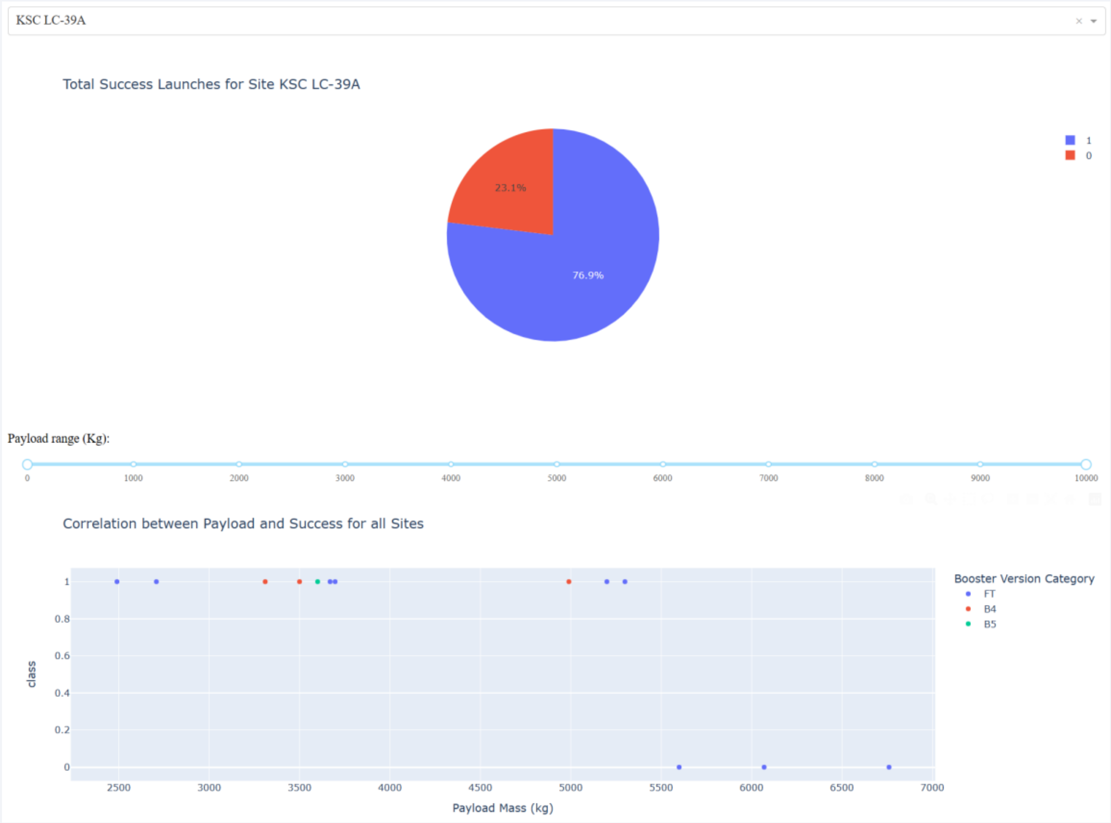
# SpaceX Launch Records

- KSC LC-39 has the most successful landings
- CCAFS SLC-40 has the least successful landings
- Rockets with Booster Version v1.1 are most likely not going to land successfully



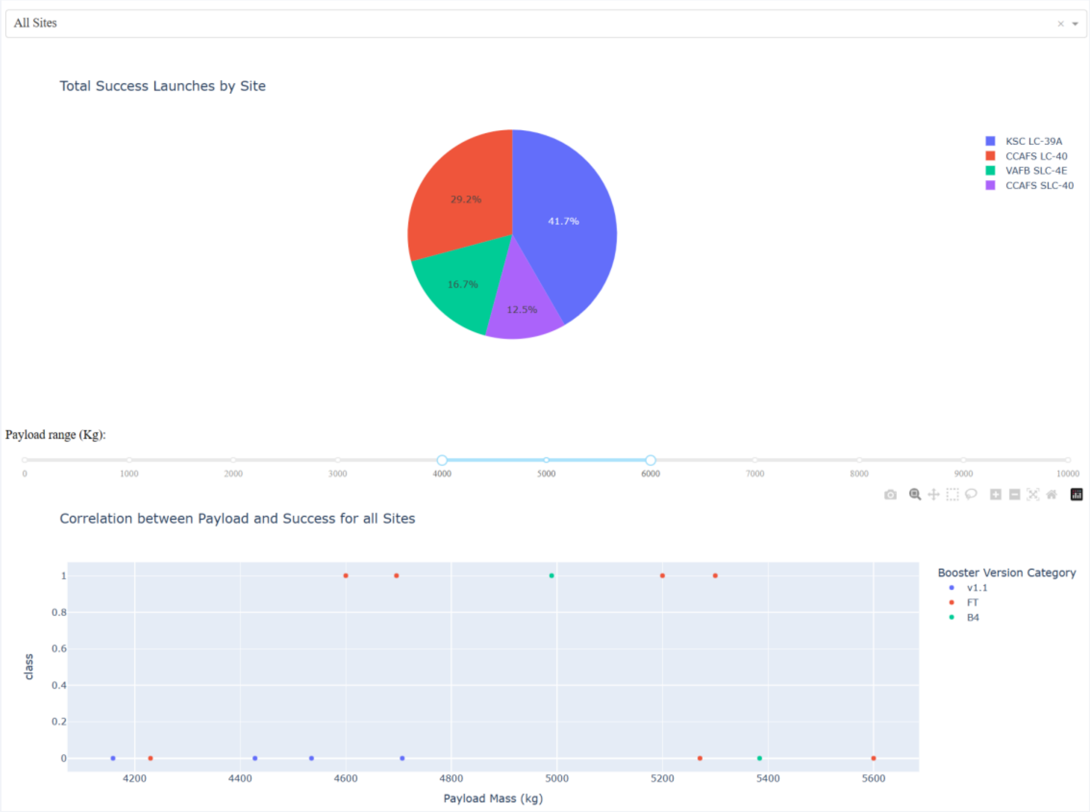
# Landing Outcomes of KSC LC-39

- KSC LC-39 with the most total successful landings also has the highest success rate
- It mostly uses the Booster Version FT
- The Rockets carried pay load masses between 2500 and 6750kg



# Payload vs Success in specific range

- Payload between 4000 and 6000kg
- In this range, all rockets with Booster Version v1.1 failed during landing

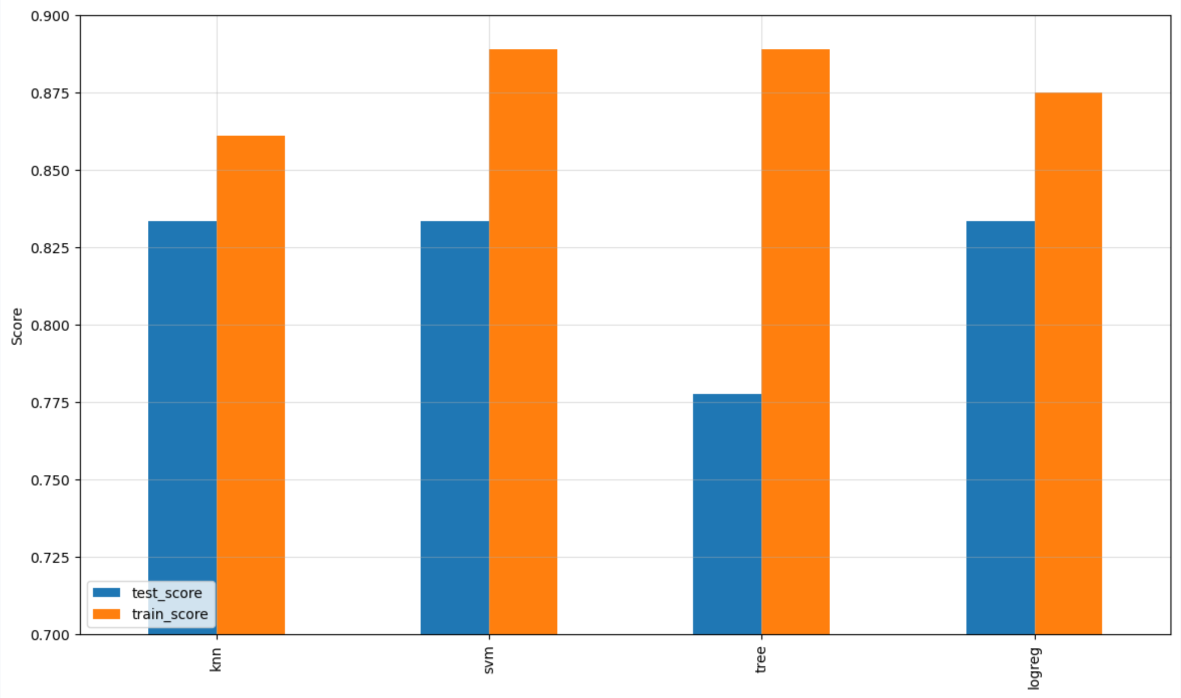


Section 5

# Predictive Analysis (Classification)

# Classification Accuracy

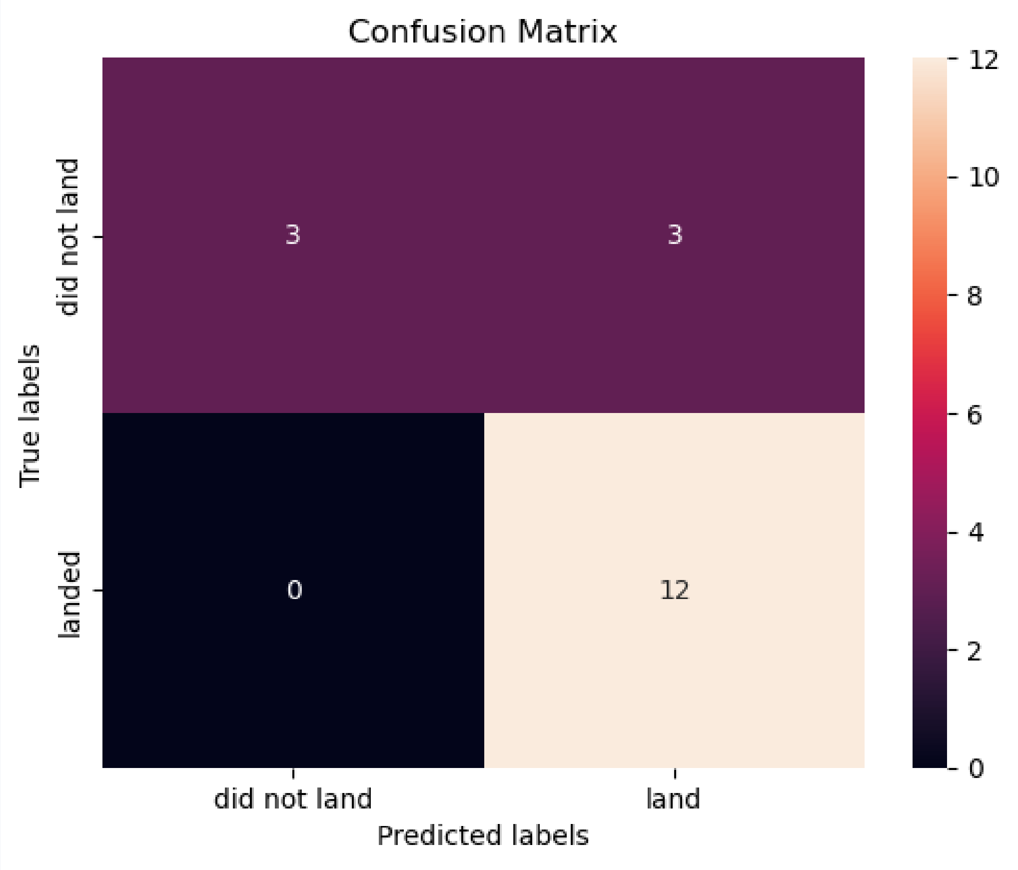
- The kNN model has the lowest accuracy
- The models SVM and Decision Tree have the highest test score
- The train score of SVM is better then the one of the Decision Tree




Therefore SVM is the best model for this usecas

# Confusion Matrix

- There are no False Negative predictions
- The main issue is the number of False positive predictions



# Conclusions

- We found some correlations regarding the successful landing, for example:
  - Payload mass
  - Orbit type
  - Whether the rocket is already reused
  - Booster Version
- We can predict with great certainty when a rocket will land safely again
- If our prediction is that the rocket will not land safely, there is still a chance that it will land correctly
- With a bigger dataset, we could train a better model with higher accuracy

# Conclusions

- We found some correlations regarding the successful landing, for example:
  - Payload mass
  - Orbit type
  - Whether the rocket is already reused
  - Booster Version
- We can predict with great certainty when a rocket will land safely again
- If our prediction is that the rocket will not land safely, there is still a chance that it will land correctly
- With a bigger dataset, we could train a better model with higher accuracy

# Appendix

Columns from the original data:

```
['static_fire_date_utc', 'static_fire_date_unix', 'tbd', 'net', 'window', 'rocket',
'success', 'details', 'crew', 'ships', 'capsules', 'payloads', 'launchpad', 'auto_update',
'failures', 'flight_number', 'name', 'date_utc', 'date_unix', 'date_local',
'date_precision', 'upcoming', 'cores', 'id', 'fairings.reused', 'fairings.recovery_attempt',
'fairings.recovered', 'fairings.ships', 'links.patch.small', 'links.patch.large',
'links.reddit.campaign', 'links.reddit.launch', 'links.reddit.media',
'links.reddit.recovery', 'links.flickr.small', 'links.flickr.original', 'links.presskit',
'links.webcast', 'links.youtube_id', 'links.article', 'links.wikipedia', 'fairings']
```

# Appendix

Head of dataframe after first cleansing

| | FlightNumber | Date | BoosterVersion | PayloadMass | Orbit | LaunchSite | Outcome | Flights | GridFins | Reused | Legs | LandingPad | Block | ReusedCount | Serial | Longitude | Latitude | Class |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 2010-06-04 | Falcon 9 | 6104.959412 | LEO | CCAFS SLC 40 | None None | 1 | False | False | False | NaN | 1.0 | 0 | B0003 | -80.577366 | 28.561857 | 0 |
| 1 | 2 | 2012-05-22 | Falcon 9 | 525.000000 | LEO | CCAFS SLC 40 | None None | 1 | False | False | False | NaN | 1.0 | 0 | B0005 | -80.577366 | 28.561857 | 0 |
| 2 | 3 | 2013-03-01 | Falcon 9 | 677.000000 | ISS | CCAFS SLC 40 | None None | 1 | False | False | False | NaN | 1.0 | 0 | B0007 | -80.577366 | 28.561857 | 0 |
| 3 | 4 | 2013-09-29 | Falcon 9 | 500.000000 | PO | VAFB SLC 4E | False Ocean | 1 | False | False | False | NaN | 1.0 | 0 | B1003 | -120.610829 | 34.632093 | 0 |
| 4 | 5 | 2013-12-03 | Falcon 9 | 3170.000000 | GTO | CCAFS SLC 40 | None None | 1 | False | False | False | NaN | 1.0 | 0 | B1004 | -80.577366 | 28.561857 | 0 |

# Thank you!