IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Fabien Laborde
21/10/2024

Outline

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

Executive Summary

- Predictive analysis approach with binary classification on Falcon 9 stage 1 landing
- Data Collection and Understanding
  - Data collected from SpaceX Public API and Wikipedia Falcon 9 page web-scraping with BeautifulSoup
  - EDA with SQL and visualizations with Seaborn, Folium and Interactive Dashboard
- Data Preparation
  - Selection of relevant data and imputation of missing values
  - Feature engineering of the target feature and One-hot-encoding
- Modeling and Evaluation
  - Logistic Regression, SVM, Decision Tree and KNN with cross-validation for hyperparameter tuning
  - Similar results with 83,33% accuracy and 50% of false positives on a small sample of 18 launches

Introduction

- Background and Context
  - SpaceX has a major advantage over competition thanks to its first stage landing
  - SpaceX launches are about 2.5 times cheaper than competitors (62 M$ vs up to 165M$)
  - The outcome of the first stage landing is the key factor in determining the launch cost
  - Using SpaceX experience and public data, help SpaceY compete by predicting landing outcomes

- Problem
  - Can we predict if the first stage will land successfully to minimize the cost of the launch ?

Section 1

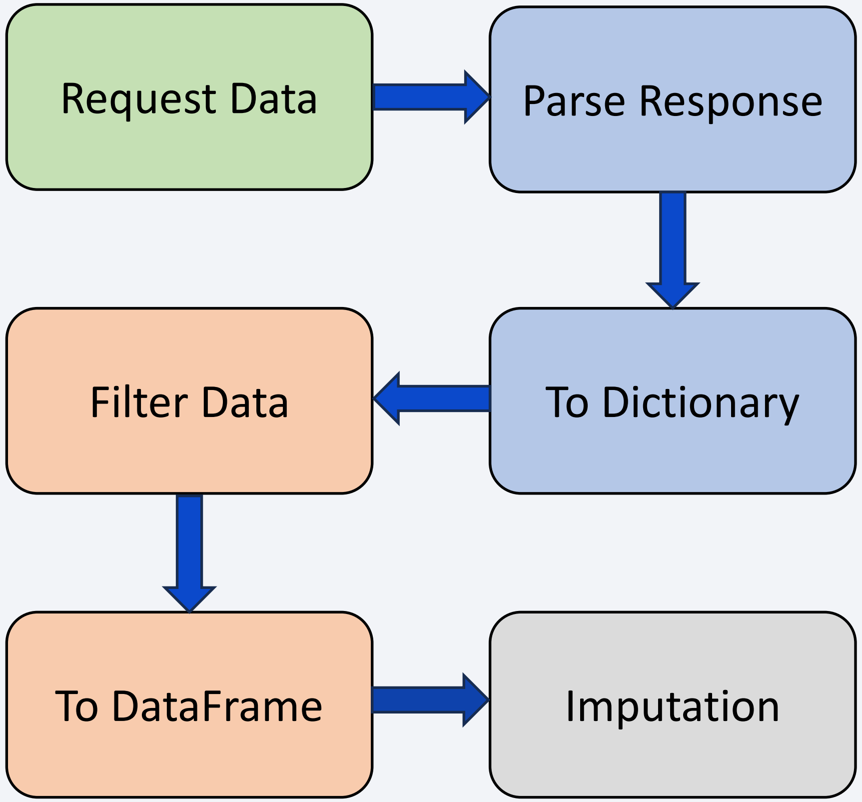
Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API REST calls and Wikipedia web-scraping with BeautifulSoup
- Perform data wrangling
  - Imputation of missing values, engineering of target feature and one-hot-encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Evaluation of Logistic Regression, SVM, Decision Tree and KNN models accuracy
  - 10-fold grid search cross-validation for hyperparameter tuning

Data Collection



Data sources

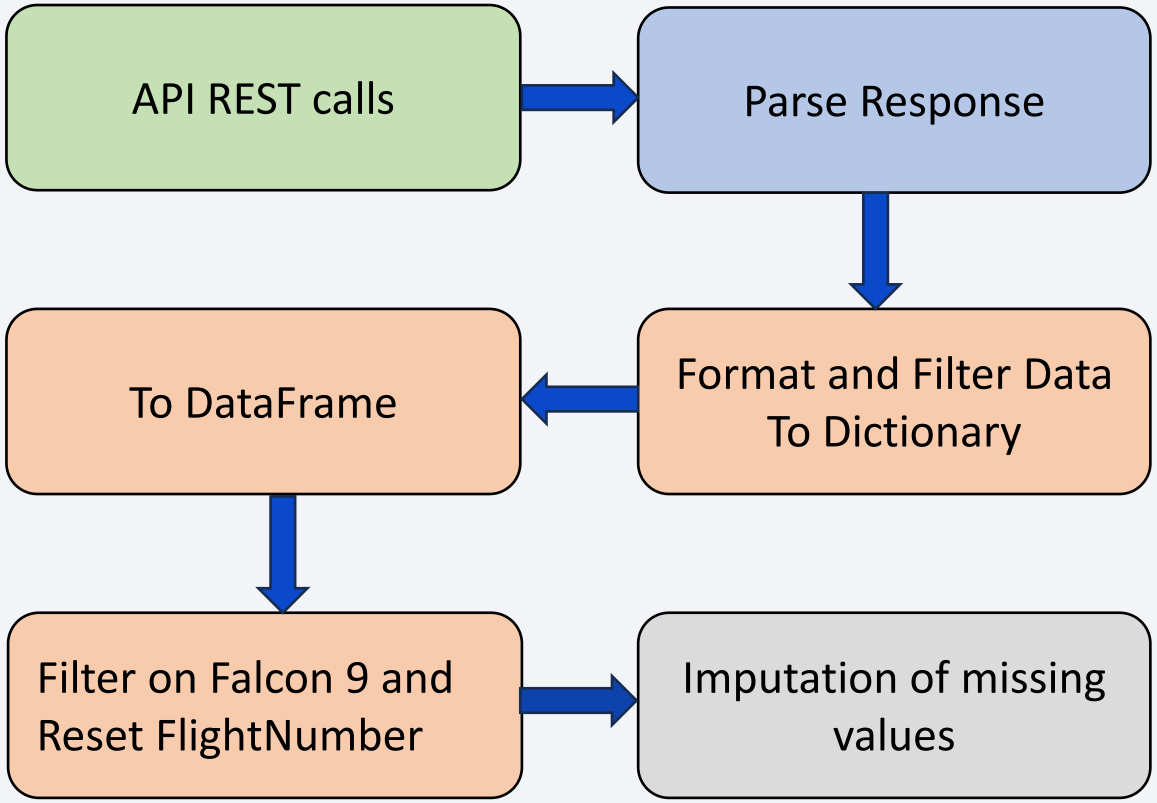
- SpaceX API REST calls
- Wikipedia web-scraping

Methodology

- Request the data
- Parse the response into a Dictionary
- Filter the data and create features into a DataFrame
- Handle the missing values

Data Collection – SpaceX API

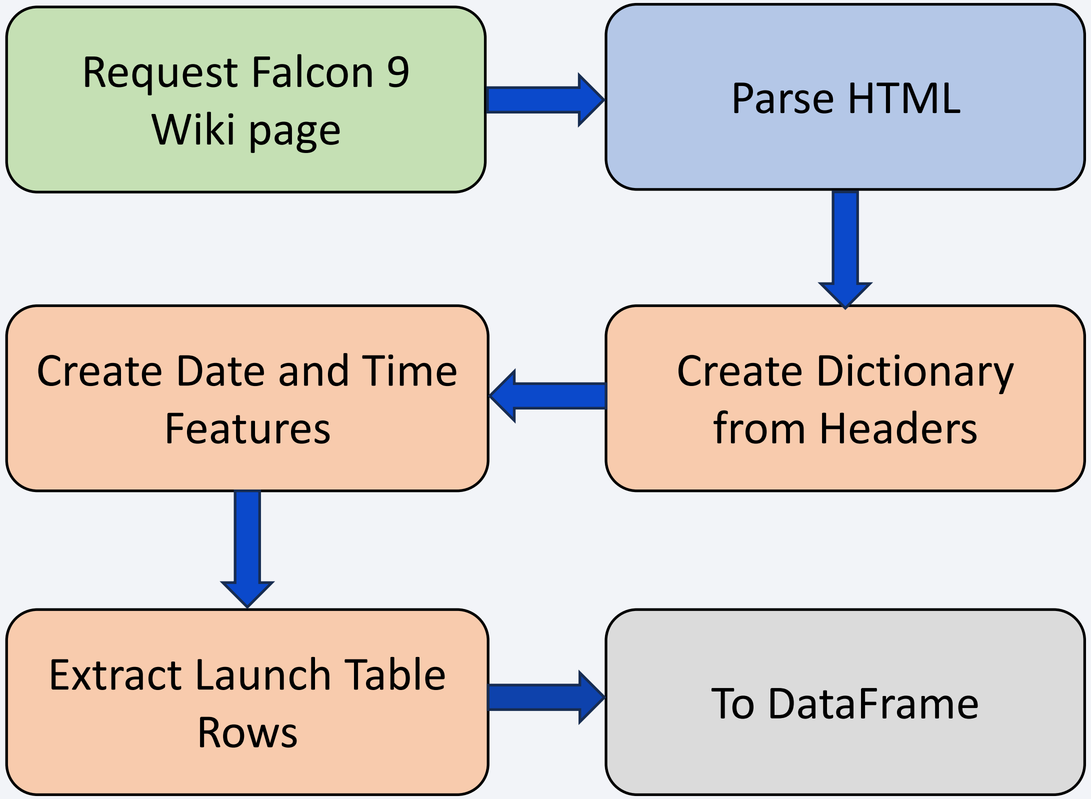
- API REST Calls on Rockets, Launchpads, Payloads, and Cores
- Parsing JSON Response with json_normalize()
- Selection of a subset and formatting into a dictionary
- Casting to DataFrame
- Selection of rows for Falcon 9 boosters
- FlightNumber reindexing
- Imputation of the 5 missing PayloadMass with the mean



URL: https://github.com/LabFSquared/IBM-Data-Science/blob/main/W1_01_jupyter-labs-spacex-data-collection-api-v2.ipynb

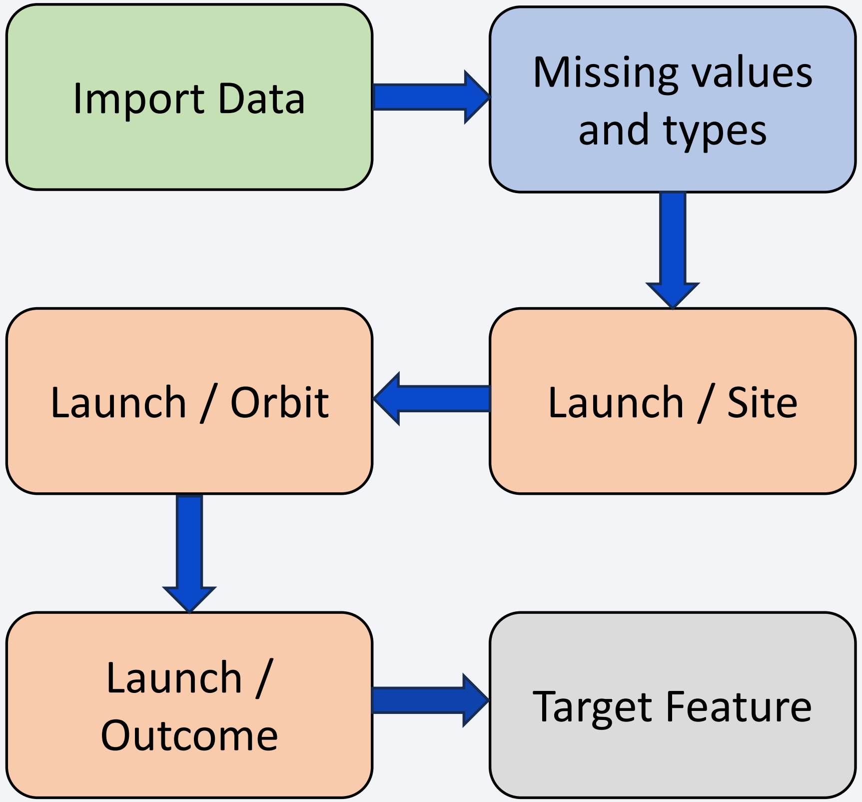
Data Collection - Scraping

- Request of Falcon 9 Wikipedia page
- Parsing HTML Response with BeautifulSoup
- Creation of a dictionary from the table headers
- Creation of Date and Time features
- Extraction of the Launch table rows into the dictionary iteratively
- Casting to DataFrame



URL: https://github.com/LabFSquared/IBM-Data-Science/blob/main/W1_02_jupyter-labs-webscraping.ipynb

Data Wrangling



Methodology

- Import the data from csv into a DataFrame
- Check the missing values and the types of the columns
- Calculate the number of launch per site
- Calculate the number of launch per orbit type
- Calculate the number of launch per outcome type
- Map the outcome success to 1 and failure to 0
    - Success: Outcome containing “True”
    - Failure: Outcome containing “False” or “None”

URL: https://github.com/LabFSquared/IBM-Data-Science/blob/main/W1_03_labs-jupyter-spacex-Data%20wrangling-v2.ipynb

EDA with Data Visualization

- **Scatter Plot:** correlations between variables and influence on landing outcome
  - Flight Number vs Launch Site
  - Flight Number vs Orbit Type
  - Payload Mass vs Launch Site
  - Payload Mass vs Orbit Type

- **Bar Chart:** comparison of number of flights and success rates between orbits
  - Success Rate by Orbit Type

- **Line Plot:** evolution of success rate
  - Launch Success Yearly Trend

URL: https://github.com/LabFSquared/IBM-Data-Science/blob/main/W2_02_jupyter-labs-eda-dataviz-v2.ipynb

EDA with SQL

- Unique names of launch sites
- Five records of the launch sites beginning with "CCA"
- Total payload mass launched by NASA
- Average payload mass carried by boosters version F9 v1.1
- Date of the first successful landing in ground pad
- Names of boosters succeeding in drone ship landing with a payload mass between 4000 and 6000
- Total number of success and failure by outcome type
- Name of boosters that have carried the maximum payload mass
- Records with a failure in drone ship landing in 2015
- Ranking of the outcome type counts between 2010-06-04 and 2017-03-20

URL: https://github.com/LabFSquared/IBM-Data-Science/blob/main/W2_01_jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Folium map with launch site locations, landing outcome markers and distance lines
  - Launch site locations are visualized using circles
  - Number of launches and their outcome are visualized using color coded markers
  - Distances to important locations such as highways, railways, cities and coastline are visualized using lines

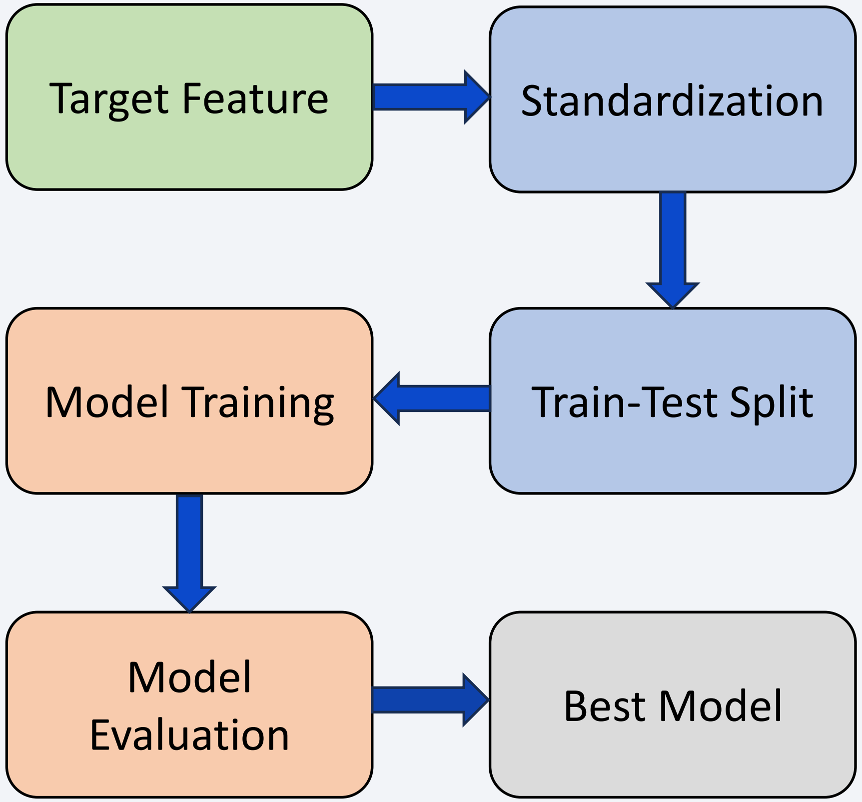
URL: https://github.com/LabFSquared/IBM-Data-Science/blob/main/W3_01_lab-jupyter-launch-site-location-v2.ipynb

Build a Dashboard with Plotly Dash

- Interactive Dashboard using Plotly Dash
    - Dropdown list to select a specific launch site or all sites and their corresponding indicators
    - Pie chart to show the share of success rate across all sites or the success-to-failure ratio for a specific launch site
    - Scatter plot of Payload Mass vs Outcome with color coded Booster Categories for the selected option to show the correlation between payload, booster category and their influence on landing outcomes
    - Slider to select a range of payload mass for the scatter plot

URL: https://github.com/LabFSquared/IBM-Data-Science/blob/main/W3_02_spacex_dash_app.py

Predictive Analysis (Classification)



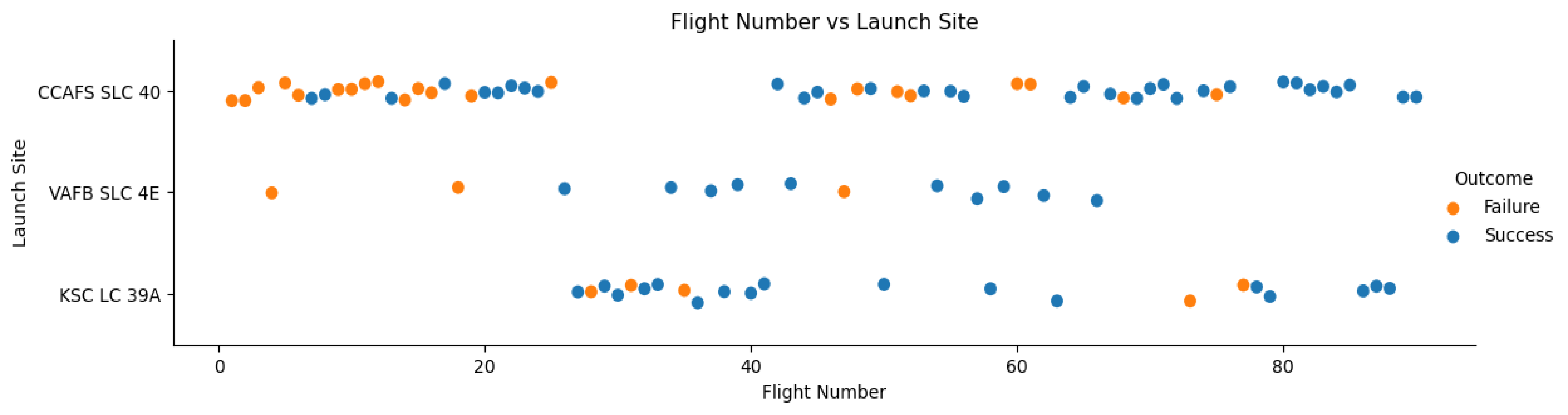
- Split target feature ‘class’ from DataFrame to numpy array
- Standardize independent features with StandardScaler()
- Train-test split with test size of 20%
- Training of Logistic Regression, SVM, Decision Tree and KNN models with a 10-fold GridSearchCV for hyperparameter tuning
- Evaluation of models accuracy on train and test sets
- Evaluation of precision and recall with confusion matrix
- Comparison of models accuracy

URL: https://github.com/LabFSquared/IBM-Data-Science/blob/main/W4_SpaceX-Machine-Learning-Prediction-Part-5-v1.ipynb

Section 2

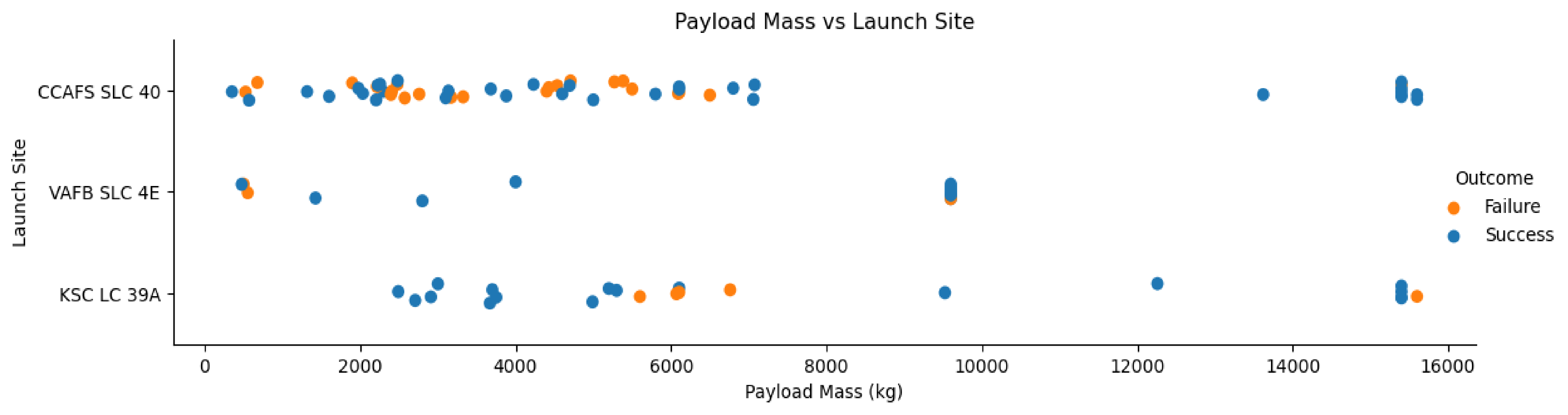
Insights drawn from EDA

Flight Number vs. Launch Site



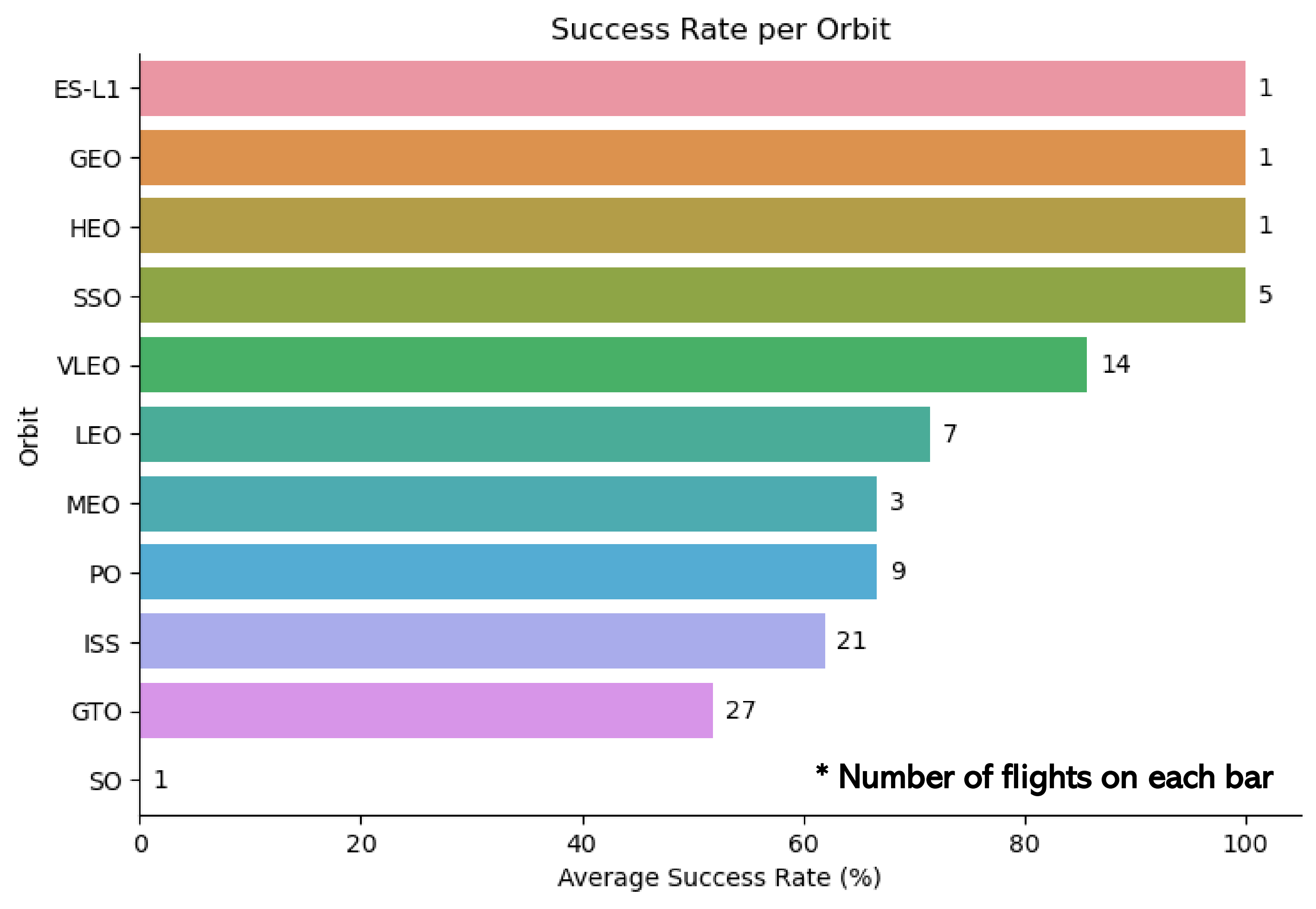
- Major breakthrough around flight 20 with a massive increase in success rate
- Most of launches from Florida, especially CCAFS, likely due to its proximity with the equator
- Between flight 20 and ~45, launches were conducted from KSC instead of CCAFS

Payload vs. Launch Site



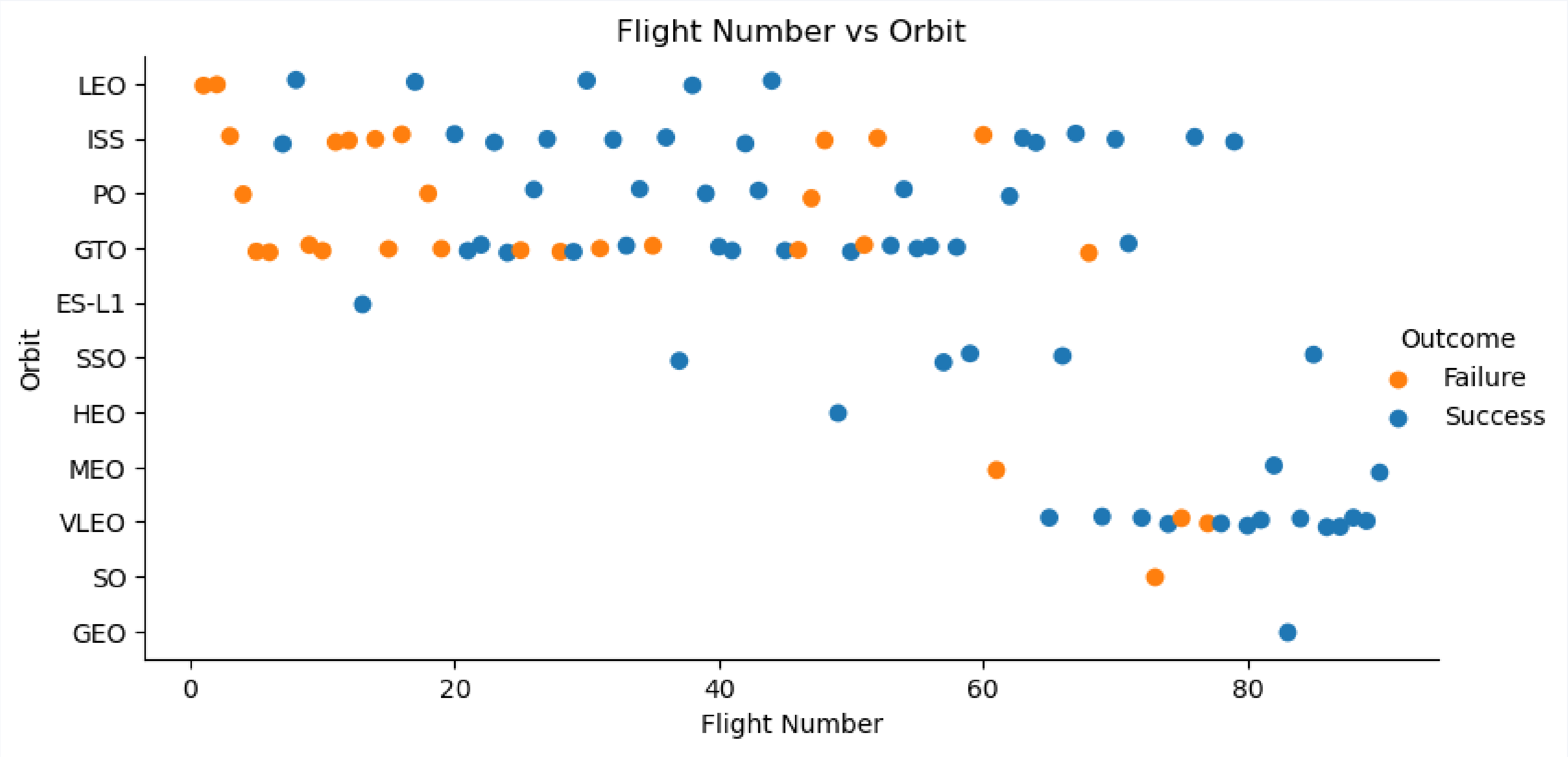
- Most of the payload mass under ~7000 kg
- High success rate for heavier payloads over 9000 kg, mainly launched from Florida (CCAFS and KSC)
- Low correlation between payload mass and success rate for CCAFS

Success Rate vs. Orbit Type



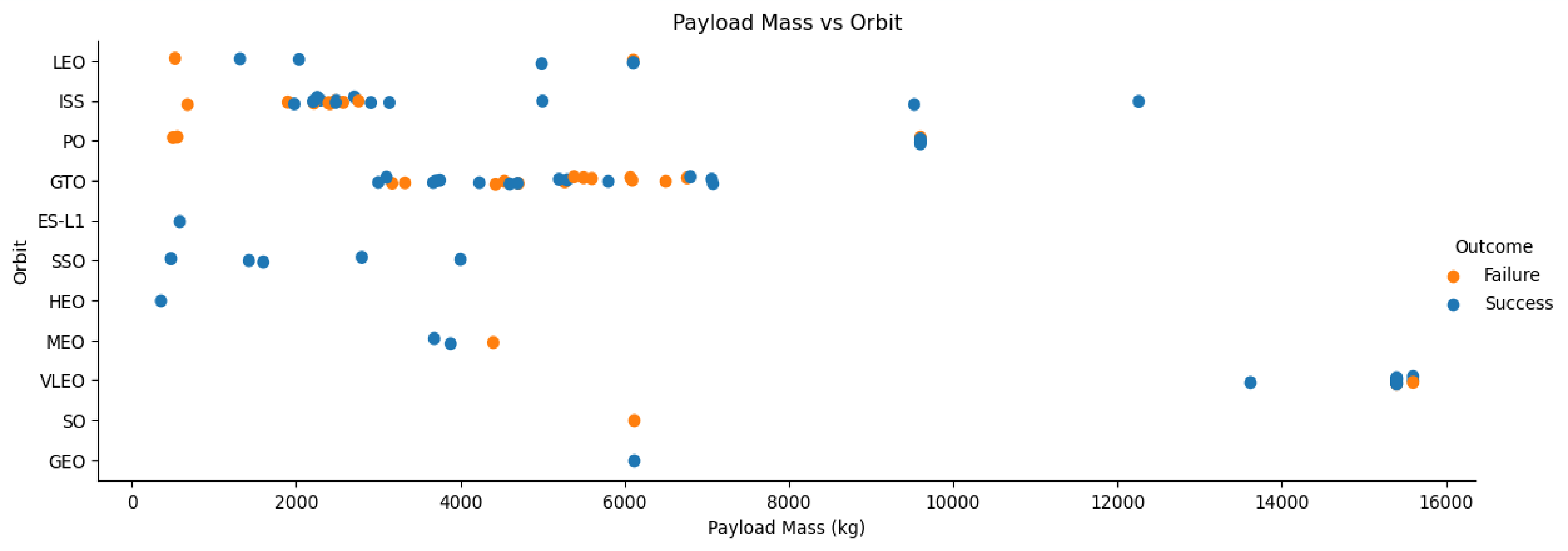
- GTO has the most flights for a relatively low ~50% success rate
- VLEO has a ~85% success rate in a relatively large number of attempts
- SSO succedeed in all five attempts
- ES-L1, GEO, HEO, and SO had one flight, with SO being the only failure

Flight Number vs. Orbit Type



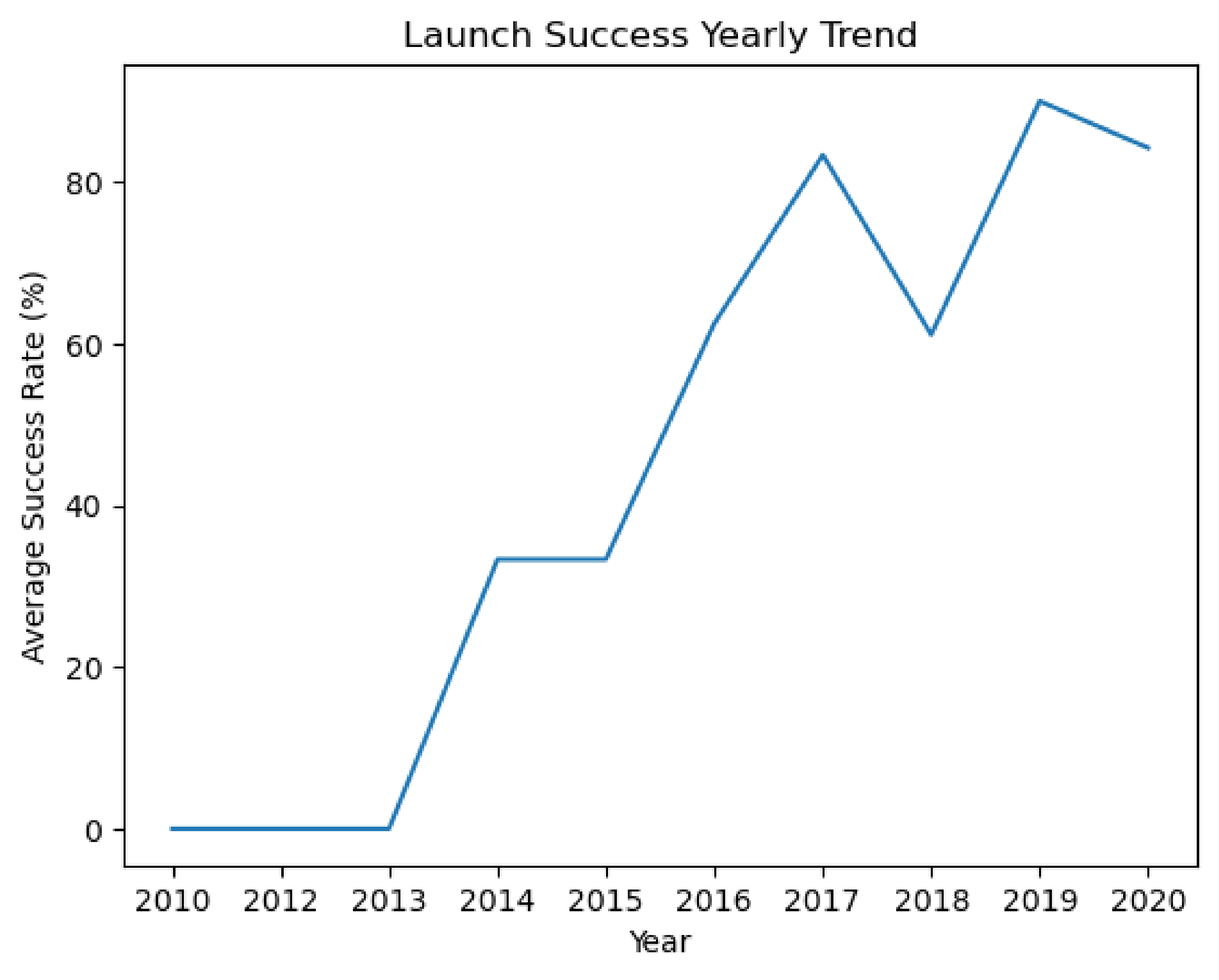
- Most launches are to low orbits (LEO, ISS, PO, SSO, GTO and VLEO)
- ~50% of launches after flight 65 are to VLEO, indicating a shift in strategy or mission type

Payload vs. Orbit Type



- VLEO allows heavier payloads due to lower energy requirements
- Low correlation between payload mass and success rate for GTO

Launch Success Yearly Trend



- Massive increase in success since 2013
- ~35% in 2014 to ~85% in 2017
- ~60% in 2018 likely due to increased complexity of missions and higher launch frequency
- Over 85% success since 2019

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
- CCAFS LC-40 and CCAFS SLC-40 refer to the same launch site at Cape Canaveral, differing only in naming conventions

Launch Site Names Begin with 'CCA'

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

Python

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

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

- 5 records where launch sites begin with `CCA`
- CCAFS stands for Cape Canaveral Air Force Station, LC-40 stands for Launch Complex 40

Total Payload Mass

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

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

SUM(PAYLOAD_MASS__KG_)
45596

- Total payload carried by boosters for NASA’s Commercial Resupply Services (CRS)
- CRS is NASA's program for contracting companies to deliver cargo to the International Space Station

Average Payload Mass by F9 v1.1

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

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

AVG(PAYLOAD_MASS__KG_)
2928.4

- Average payload mass carried by booster version F9 v1.1
- Low end of the payload mass range (0-15600)

First Successful Ground Landing Date

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

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

MIN(Date)
2015-12-22

- Date of the first successful landing outcome on ground pad

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_ BETWEEN 4000 AND 6000
```

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

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

- Boosters with successful landing on drone ship and payload mass between 4000 and 6000 kg
- Only F9 FT booster variants

Total Number of Successful and Failure Mission Outcomes

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

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

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

- Total number of successful and failure mission outcomes
- ~99% success indicates that mission success is independent of landing success (~66%)

Boosters Carried Maximum Payload

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

```
 * 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

- Booster which have carried the maximum payload mass
- Only F9 B5 booster variants likely designed to maximize payload mass (15,600 kg)

2015 Launch Records

```
%%sql
SELECT Landing_Outcome, Booster_Version, Launch_Site, substr(Date, 6, 2) as Month FROM SPACEXTABLE
WHERE Landing_Outcome = 'Failure (drone ship)' AND substr(Date, 0, 5) = '2015'
```

* sqlite:///my_data1.db
Done.

Landing_Outcome	Booster_Version	Launch_Site	Month
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	01
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	04

- Failed landing outcomes in drone ship, booster versions, launch sites and months in 2015
- Not consecutive launches according to the sequence B10xx, despite occurring in a short time span

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

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

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

Landing_Outcome	Count
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

- Ranking of landing outcome counts between 2010-06-04 and 2017-03-20 in descending order

Section 3

Launch Sites Proximities Analysis

Launch Site Locations

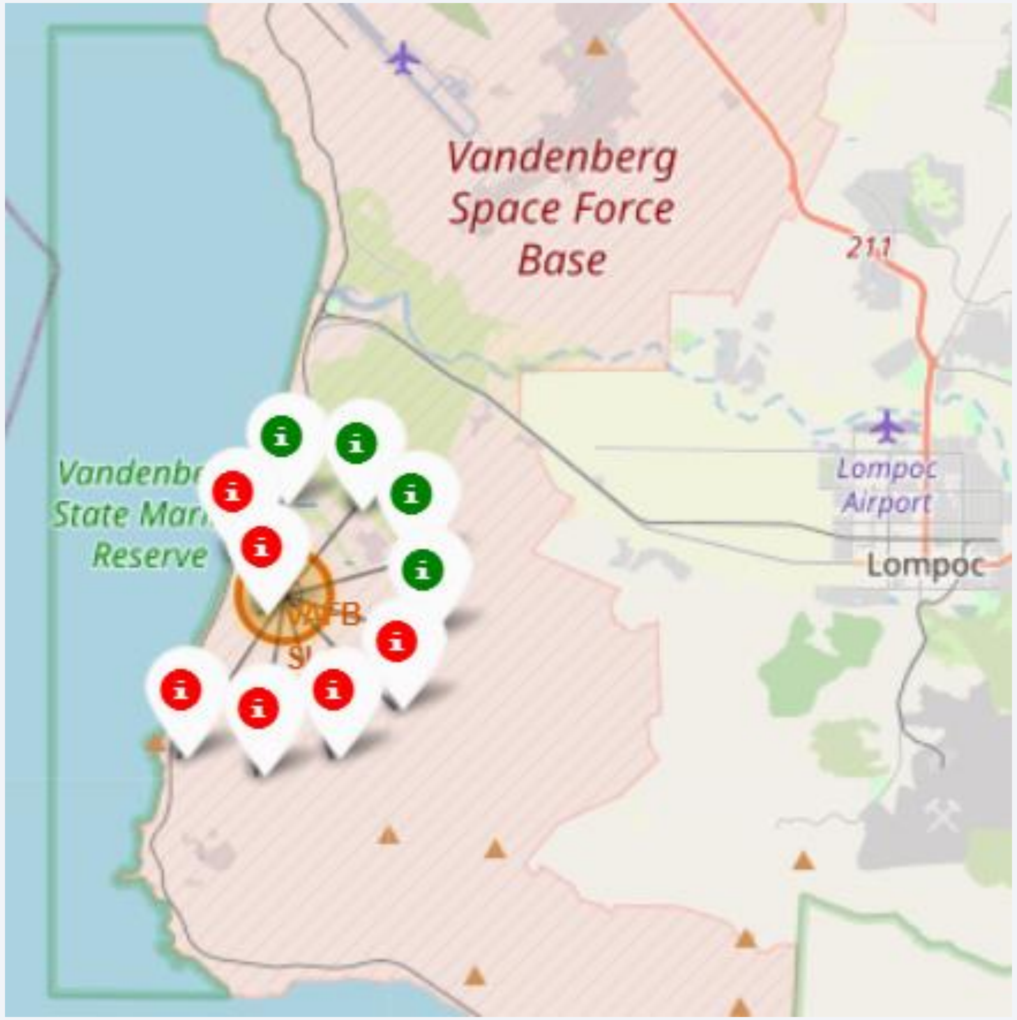
Global map with the 4 launch sites

Florida area with CCAFS LC-40, CCAFS SLC-40 and KSC LC-39A

- All launch sites are situated in relatively isolated areas along the coastline to maximize safety

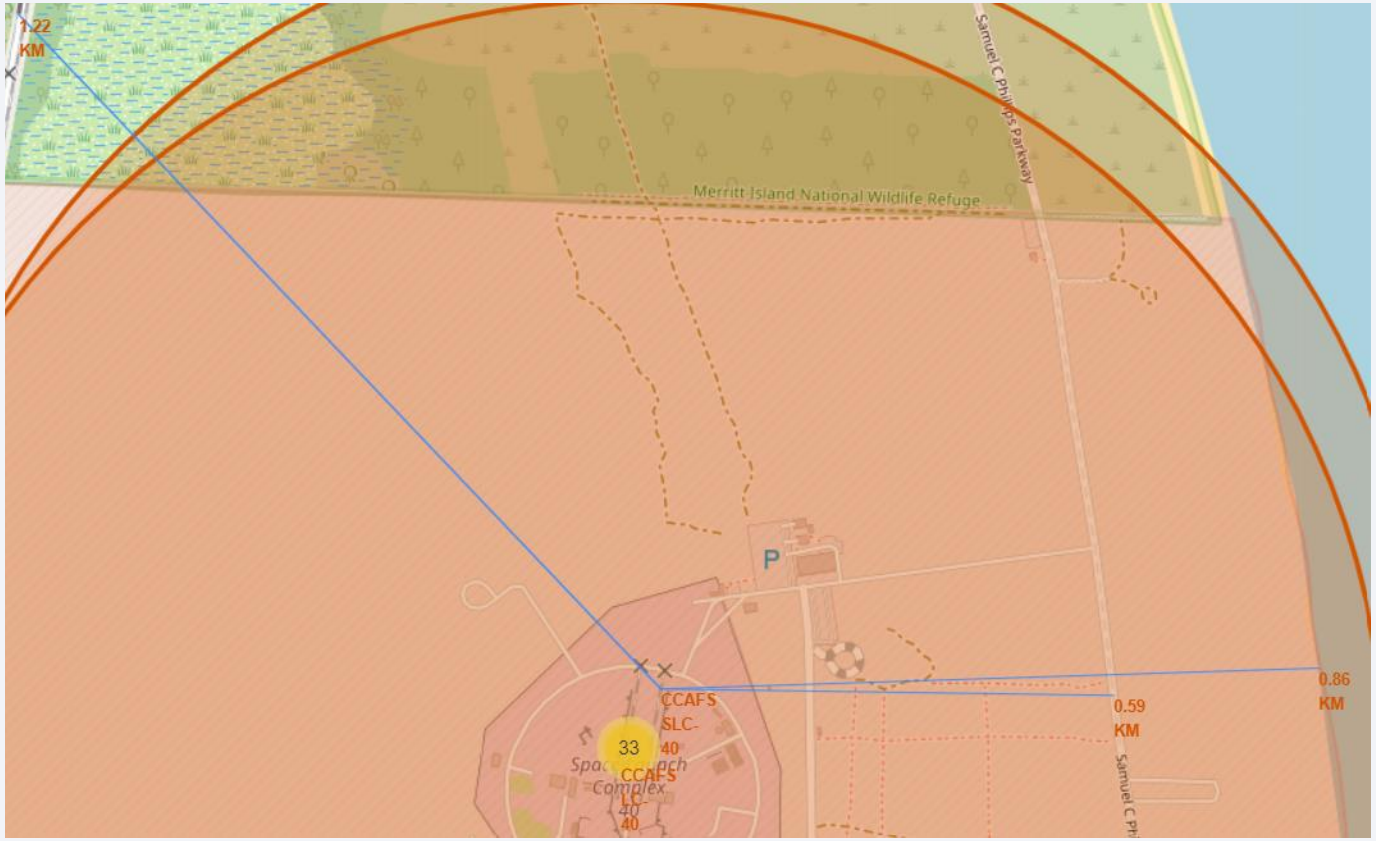
- CCAFS LC-40, CCAFS SLC-40 and KSC LC-39A in Florida
- VAFB SLC-4E in California

Launch Outcome Markers



- Launch outcomes in VAFB SLC-4E
- 4 successful landings in green
- 6 failed landings in red

Launch Site Proximities



On the left: distance between launch site and closest city, Cape Canaveral
On the right: distance between launch site and ocean, highway and railway

Distances
Ocean: 0,86 km
Highway: 0,59 km
Trainway: 1,22 km
City: 17,47 km

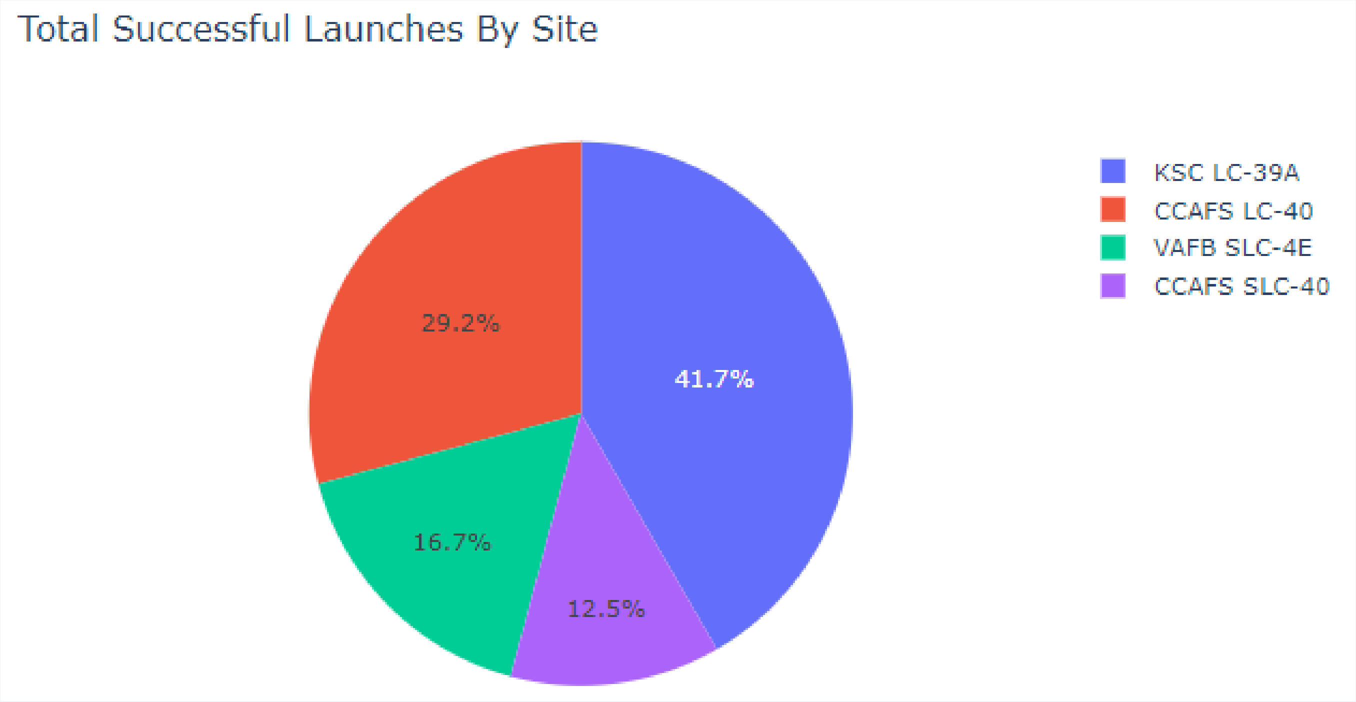
Transportation
Launch site is near highways and railways for efficient transport of personnel and equipment

Safety
Located close to the coast but distanced from cities to enhance safety

Section 4

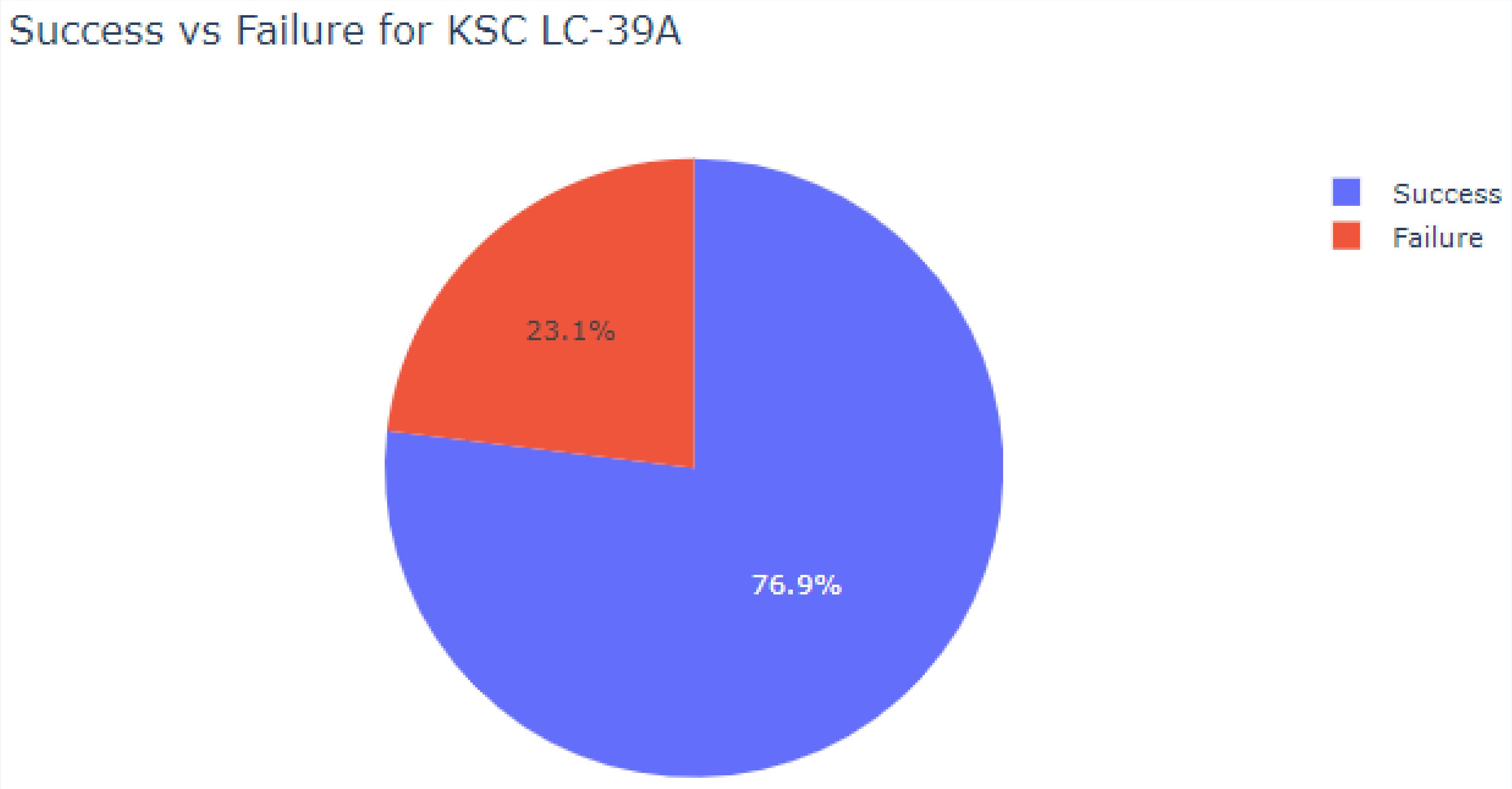
Build a Dashboard with Plotly Dash

Total Successful Launches by Site



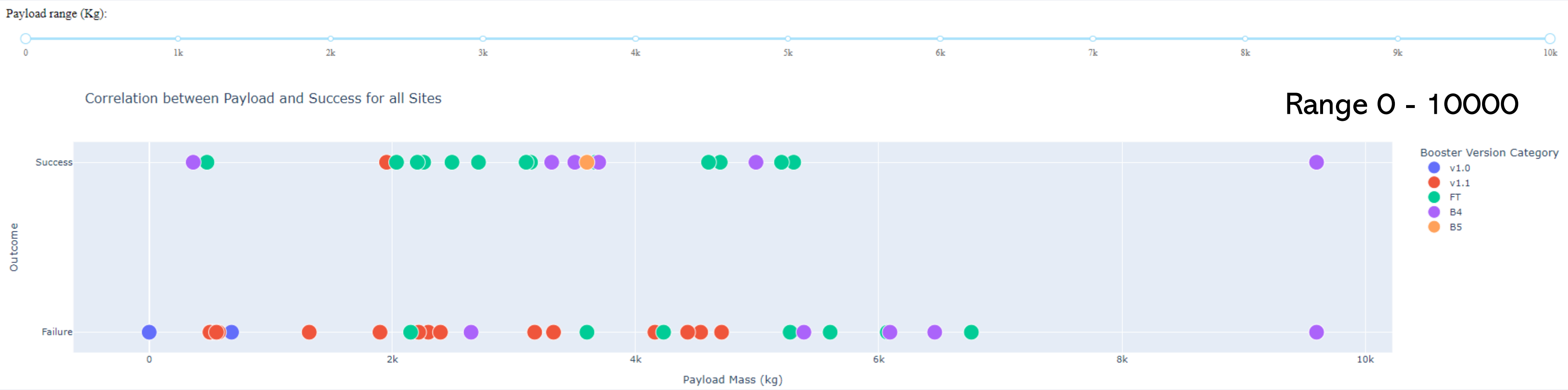
- CCAFS LC-40 and CCAFS SLC-40 refer to the same launch site at Cape Canaveral
- 83.3% of successful launches occur in Florida at CCAFS and KSC, the most active sites
- VAFB's lower share (16.7%) is due to fewer launches, given its focus on PO and SSO orbits

Highest Success Rate Launch Site



- KSC LC-39A boasts the highest success rate at 76.9%, with 10 successes and 3 failures

Payload Mass vs Outcome per Booster Category



Range 0-10000

Most Successful

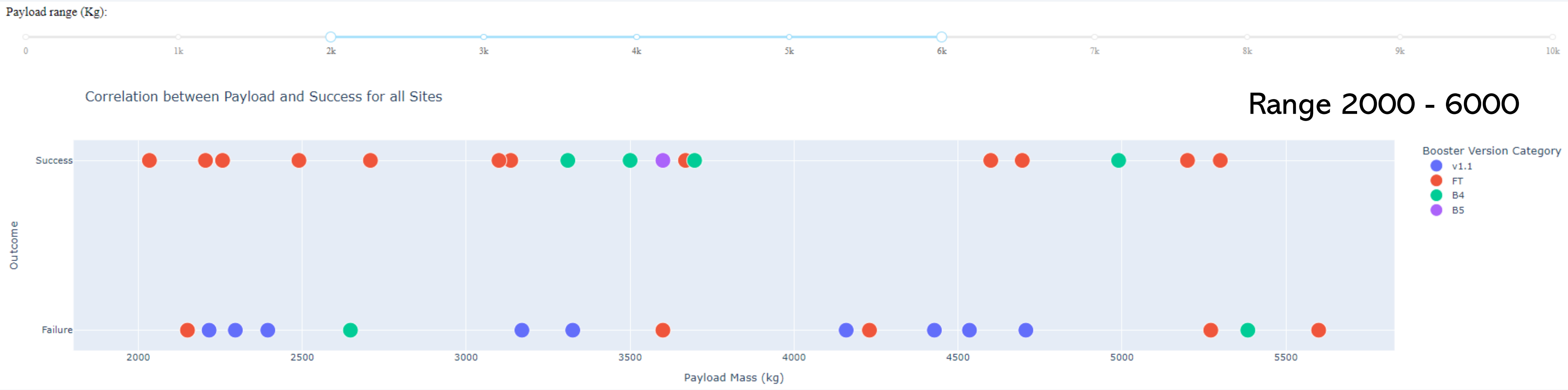
FT Boosters

Most Failures

v1.1 Boosters

Heaviest Payload

B4 Boosters



Range 2000-6000

Most Successful

FT Boosters

Most Failures

v1.1 Boosters

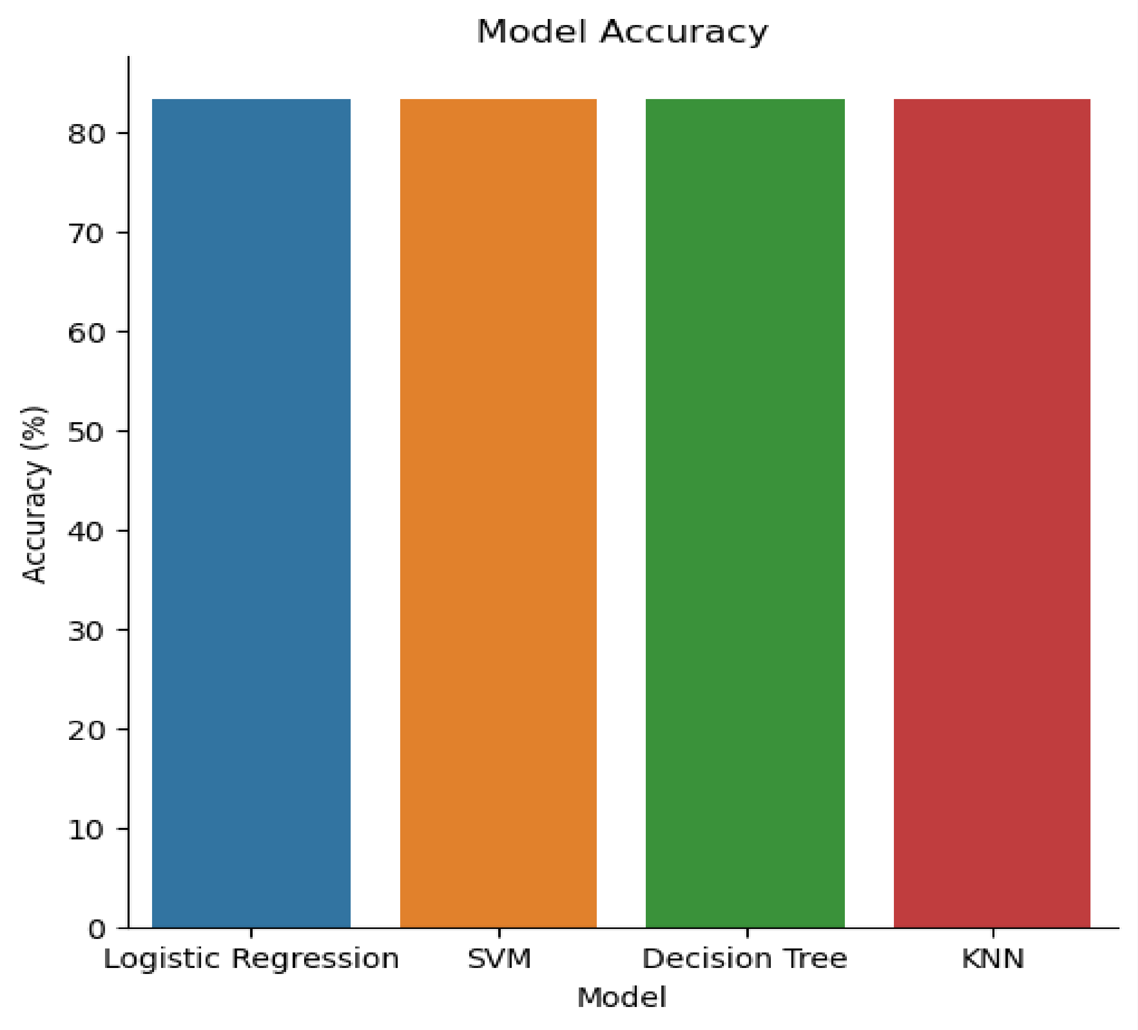
No Launch

v1.0 Boosters

Section 5

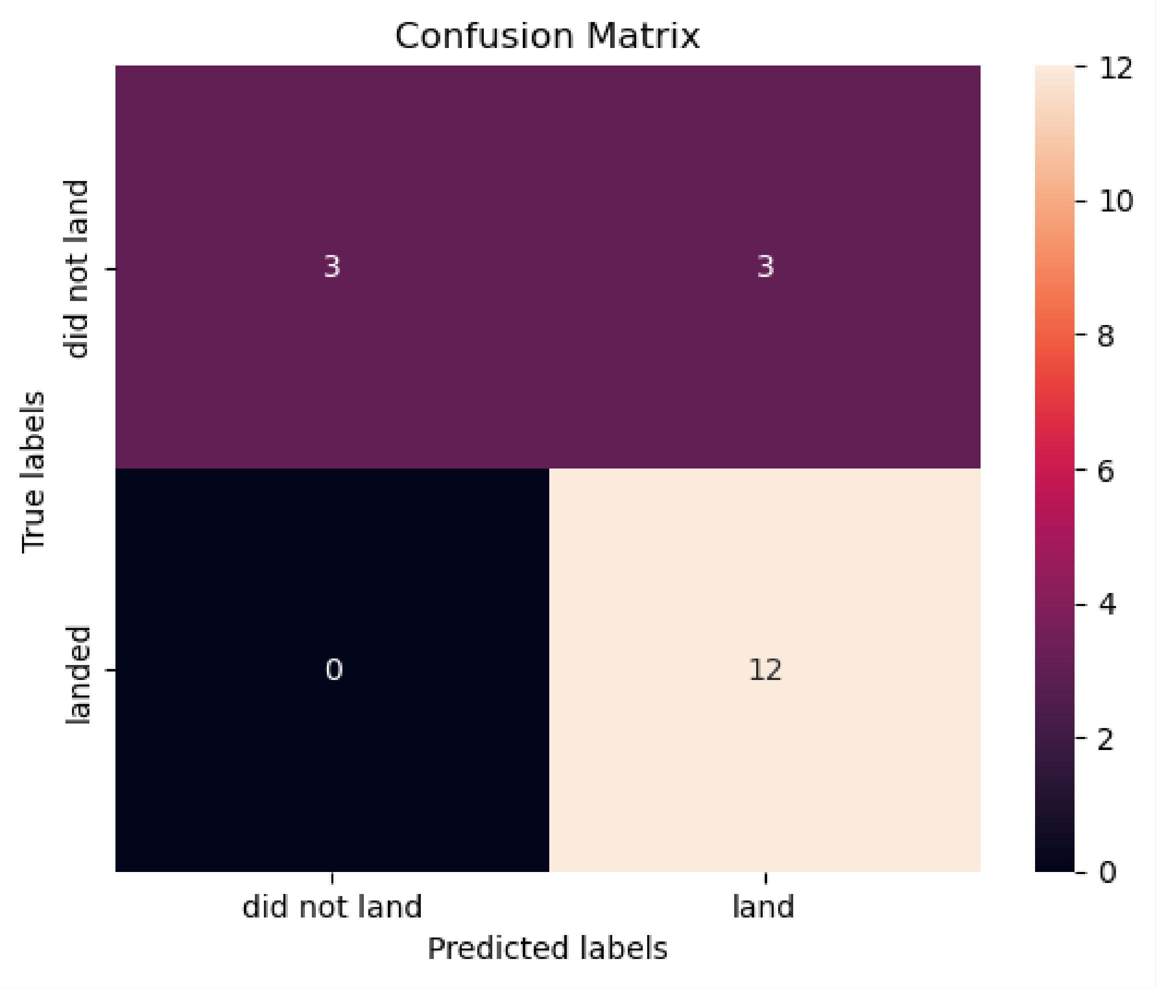
Predictive Analysis (Classification)

Classification Accuracy



- Small sample size of 18 launches
- 83,33% accuracy for all models on the test set
- Signs of overfitting on the Decision Tree with 87,67% accuracy on the train set
- More data is needed to ensure good generalization

Confusion Matrix



- Small sample size of 18 launches
- Same results for all models
- 83,33% correctly classified
    - 100% of successful landings correctly classified
    - 50% of failed landings misclassified (false positive)

Conclusions

- SpaceX leads the industry with launch costs at less than half those of competitors
- SpaceX's booster reusability is the key factor enabling significant cost reductions
- Predicting landing outcomes supports decision-making to optimize costs
- Data was collected from SpaceX API and Wikipedia page
- EDA with Pandas and SQL, using Seaborn, Folium and Plotly Dash for visualizations and dashboarding
- A predictive analysis with a binary classification approach was conducted
    - The models demonstrate good predictive capabilities with an accuracy of 83.33%
    - A critical concern is that 50% of actual failures were classified as false positives
    - In addition, the sample size is too small to train reliable models
    - Train the models on a larger dataset and set a higher threshold for positive outcomes

Thank you!