IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Vincent Rammelsberg
October 6, 2025

Outline

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

Executive Summary

- **Summary of methodologies**
  - Data Collection through SpaceX API
  - Data Collection via Webscraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Visualizations
  - Interactive Visual Analytics using Folium
  - Interactive Dashboard with Plotly Dash
  - Prediction of first stage landing outcome via Machine Learning models

- **Summary of all results**
  - Average success rate of rocket launches increased between 2013 and 2020
  - Launch site KSC LC-39A has the highest success rate
  - Orbit types ES-L1, GEO, HEO and SSO have a 100% success rate
  - All launch sites are in proximity to a coast
  - Machine Learning models were able to accurately predict first stage landing outcomes (accuracy > 80%)

Introduction

Background

- SpaceX advertises Falcon 9 rocket launches on ist 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.
- Therefore if we can determine if the first stage will land, we can determine the cost of a launch.
- Our company SpaceY wants to use this information to bid against SpaceX for a rocket launch.

Questions

- Which factors impact the success rate of first stage landings?
- Can we build a model that accurately predicts the first stage landing outcome?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data collection through SpaceX API
  - Webscraping from a Wikipedia page
- Perform data wrangling
  - Filter relevant data, handle missing values, apply OneHotEncoding to categorical features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Train and test multiple machine learning models
  - Tune hyperparameters of each model via GridSearch

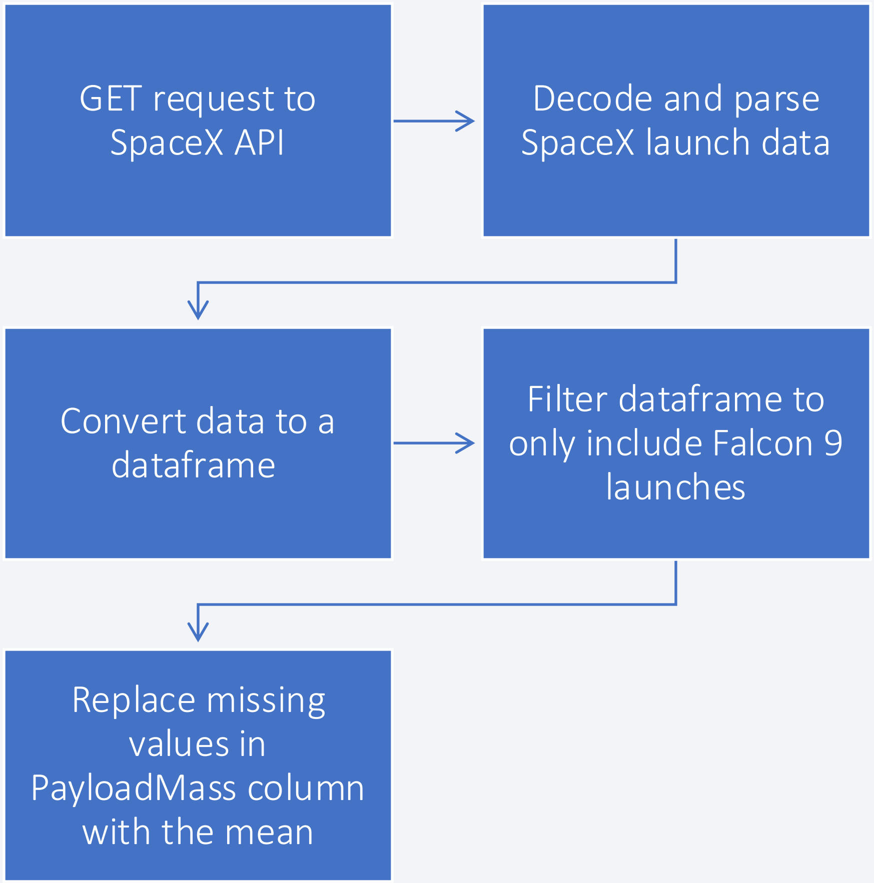
Data Collection

- Collect data through SpaceX API
  - Make a GET request to the SpaceX API
  - Decode response content using .json() & convert result into a dataframe using .json_normalize()
- Webscraping to collect Falcon 9 historical launch records from a Wikipedia page
  - Perform an HTTP GET method to request the Falcon9 Launch HTML page
  - Use BeautifulSoup to parse HTML table & convert it into a Pandas dataframe

Data Collection – SpaceX API

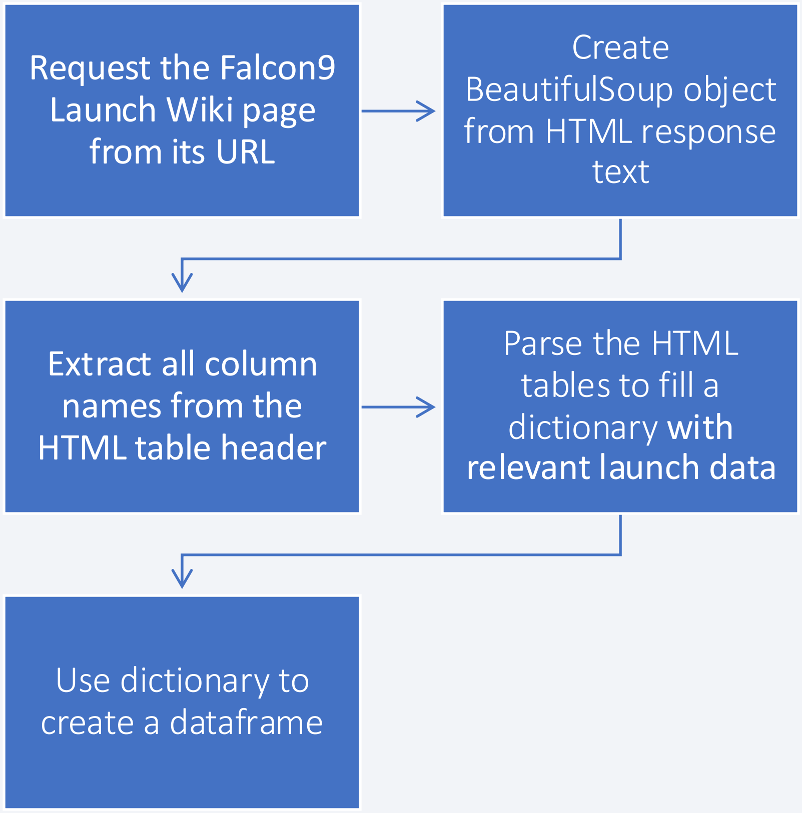
- The SpaceX API provides information about rocket launches done by SpaceX
- Collect relevant data through the SpaceX API
- We used GET-requests to several endpoints of the SpaceX API

- https://github.com/vincentR2/IBM_DataScience_Capstone/blob/main/01_DataCollection_API.ipynb



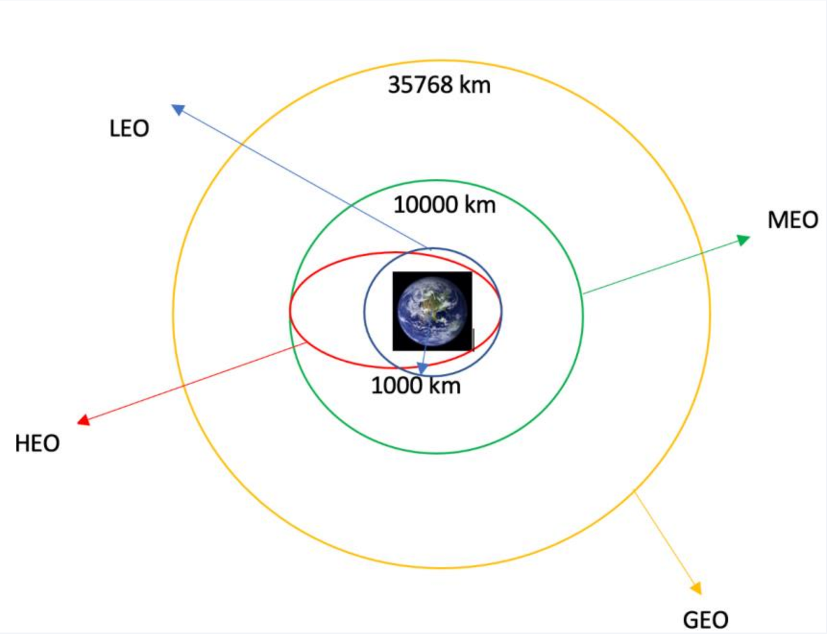
Data Collection - Scraping

- Perform web scraping to collect historical launch records from a Wikipedia page
- https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- https://github.com/vincentR2/IBM_DataScience_Capstone/blob/main/02_DataCollection_Webscraping.ipynb



Data Wrangling

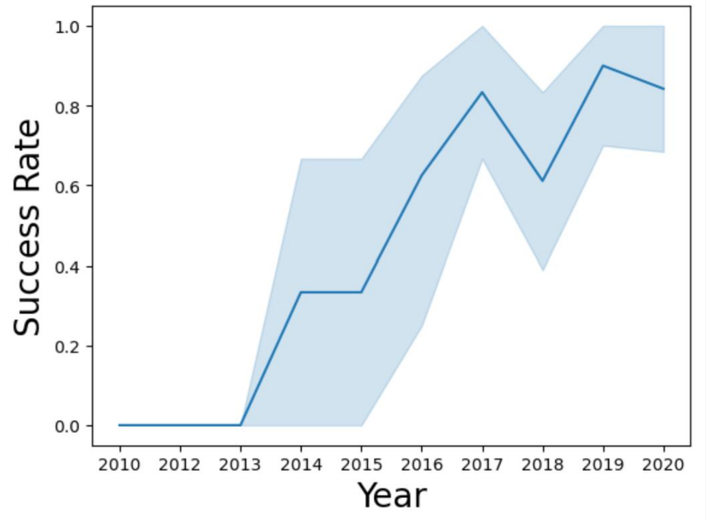
- Perform some Exploratory Data Analysis (EDA) to identify patterns in the data
  - Determine number of launches per launch site
  - Determine occurrences of each orbit type
- Data Wrangling
  - Convert outcomes of booster landings to a binary label
    - 0: booster landing was unsuccessfull
    - 1: booster landed successfully
  - Create dummy variables to categorical columns
- https://github.com/vincentR2/IBM_DataScience_Capstone/blob/main/03_DataWrangling.ipynb

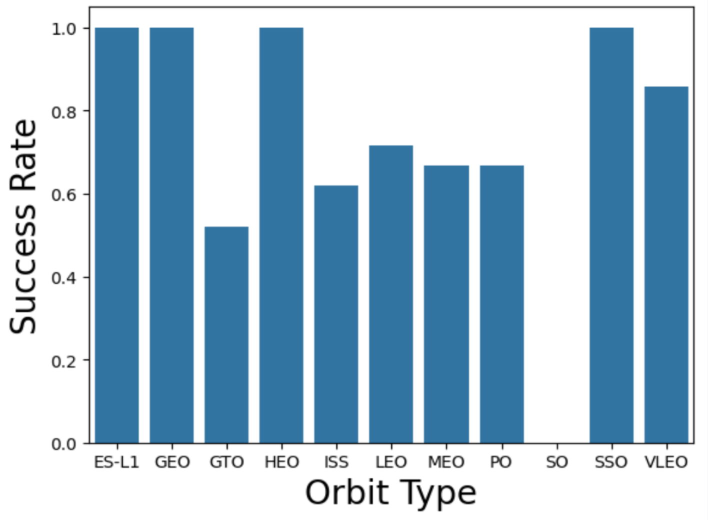


EDA with Data Visualization

- Perform EDA with data visualizations to identify relationships between the features
    - Relationship between Flight Number & Launch Site
    - Relationship between Payload Mass & Launch Site
    - Launch Success Rate over the Years (->Lineplot)
    - Success Rate of each Orbit Type (->Barplot)

- https://github.com/vincentR2/IBM_DataScience_Capstone/blob/main/05_EDA_DataVisualizations.ipynb





EDA with SQL

- Load the SpaceX dataset into the corresponding table in a Db2 database
- Execute SQL queries to gain insights about SpaceX launch data
  - Names of unique launch sites
  - Total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version "F9 v1.1"
  - Date of first successful landing outcome on ground pad
  - Total number of successful and failure mission outcomes
  - All booster versions that have carried the maximum payload mass
  - Count of landing outcomes between 2010-06-04 and 2017-03-20
- https://github.com/vincentR2/IBM_DataScience_Capstone/blob/main/04_EDA_SQL.ipynb

Build an Interactive Map with Folium

- Analyze the existing launch site locations with Folium
- Add a marker for each launch site on the folium map
  - Identify locations of existing launch sites
- Create colored-labeled markers for all launch records based on the launch outcome to identify launch sites with high success rates
  - Successful: green marker
  - Unsuccessful: red marker
- Calculate the distances between a launch site and its proximities (e.g. railway, highway, coastline, closest city)
  - Draw lines between launch site and its proximities
- https://github.com/vincentR2/IBM_DataScience_Capstone/blob/main/06_InteractiveVisualAnalytics_Folium.ipynb

Build a Dashboard with Plotly Dash

- We build an interactive SpaceX launch records dashboard with Plotly Dash
  - Users can select the launch site and a range of payload mass to gain specific insights about SpaceX launch records
- Pie Chart
  - Show the total successful number of successful launches for each all sites
  - Show the number of successful and failed launches for a specific site
- Scatter Plot
  - Show the correlation between payload mass (x-axis) and launch success (y-axis)
- https://github.com/vincentR2/IBM_DataScience_Capstone/blob/main/spacex-dash-app.py

Predictive Analysis (Classification)

- We create a machine learning pipeline to predict if the first stage will land successfully or not
    1. Standardize our dataset
    2. Split our dataset into a training set (80%) and a test set (20%)
    3. Train and finetune different machine learning models
        i. Create machine learning model object
        ii. Create GridSearchCV object with different hyperparameter settings
        iii. Find model with the best hyperparameter combinations and determine the accuracy on the validation data
    4. Test models on unseen test data
        i. Calculate accuracy on test data & plot confusion matrix
- https://github.com/vincentR2/IBM_DataScience_Capstone/blob/main/07_MachineLearning_Prediction.ipynb

Results

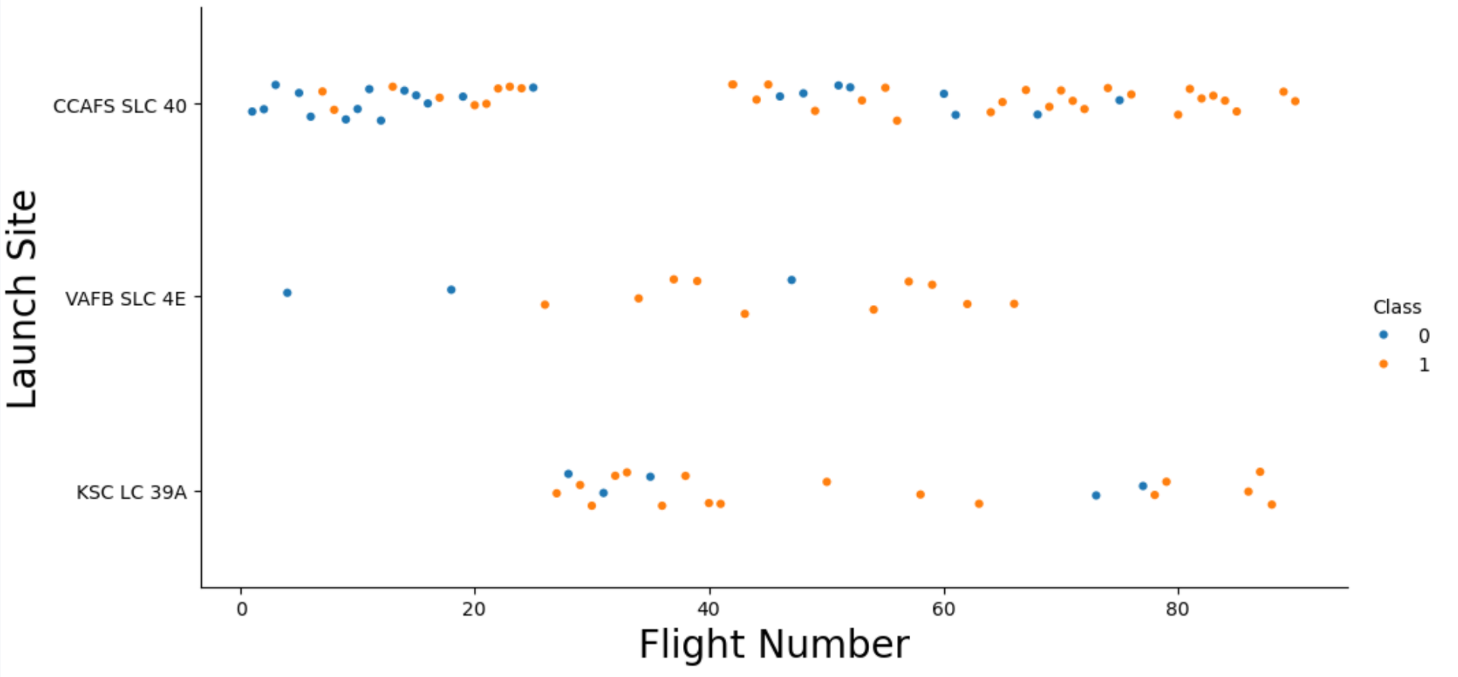
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

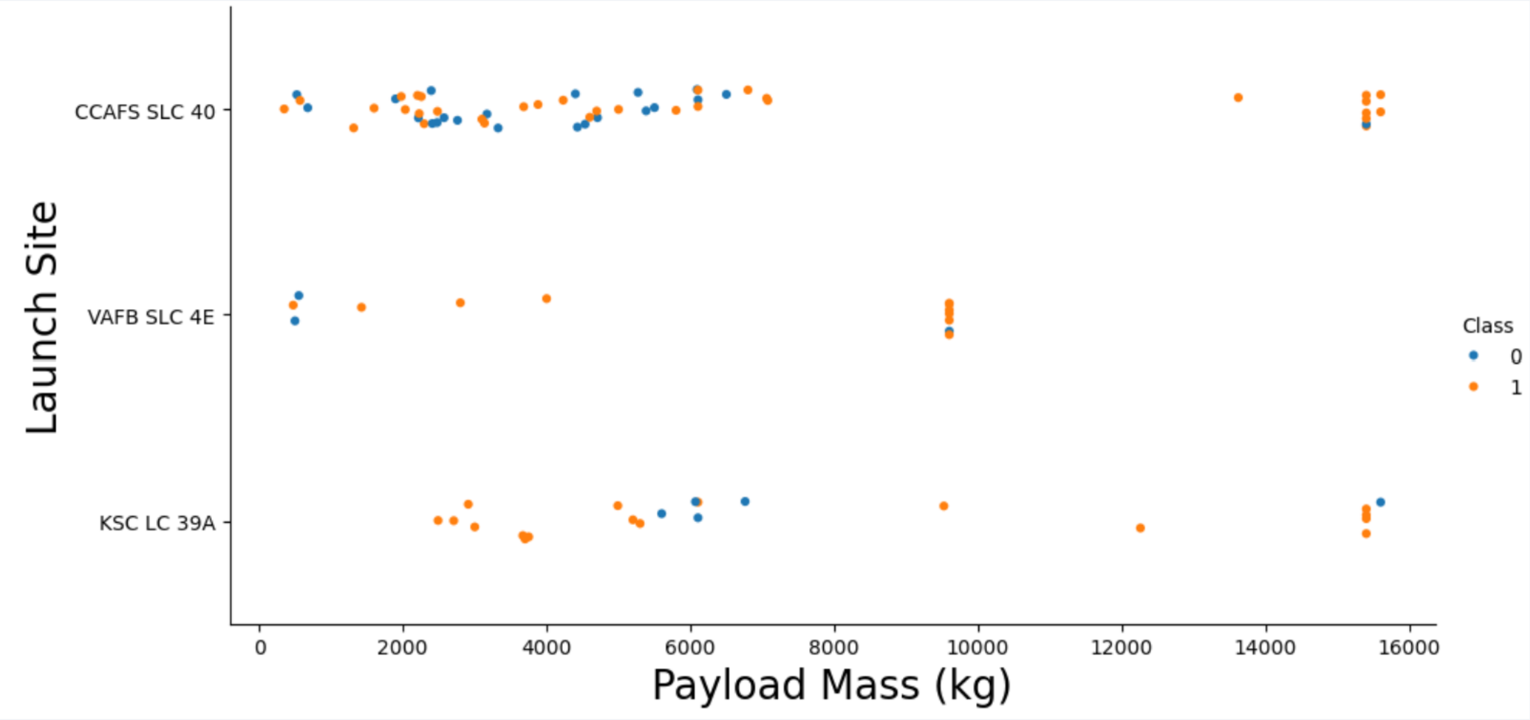
Flight Number vs. Launch Site

- The scatter plot visualizes the relationship between flight number (x-axis) and launch site (y-axis)
- Earlier flights had a lower success rate (more blue points)
- It also visualizes the success rate and amount of launches of each launch site



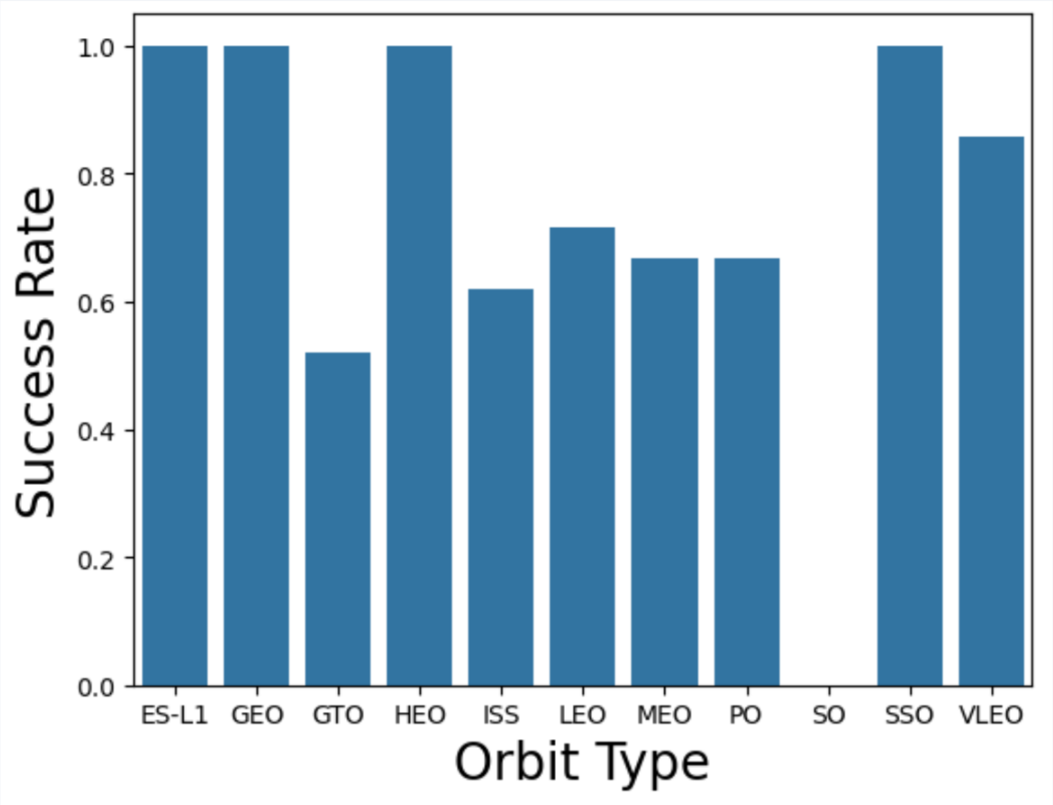
Payload vs. Launch Site

- As payload mass increases, the first stage is more likely to land successfully
- For launch site "VAFB SLC 4E" there are no rockets launched for heavy payloads (>10.000kg)

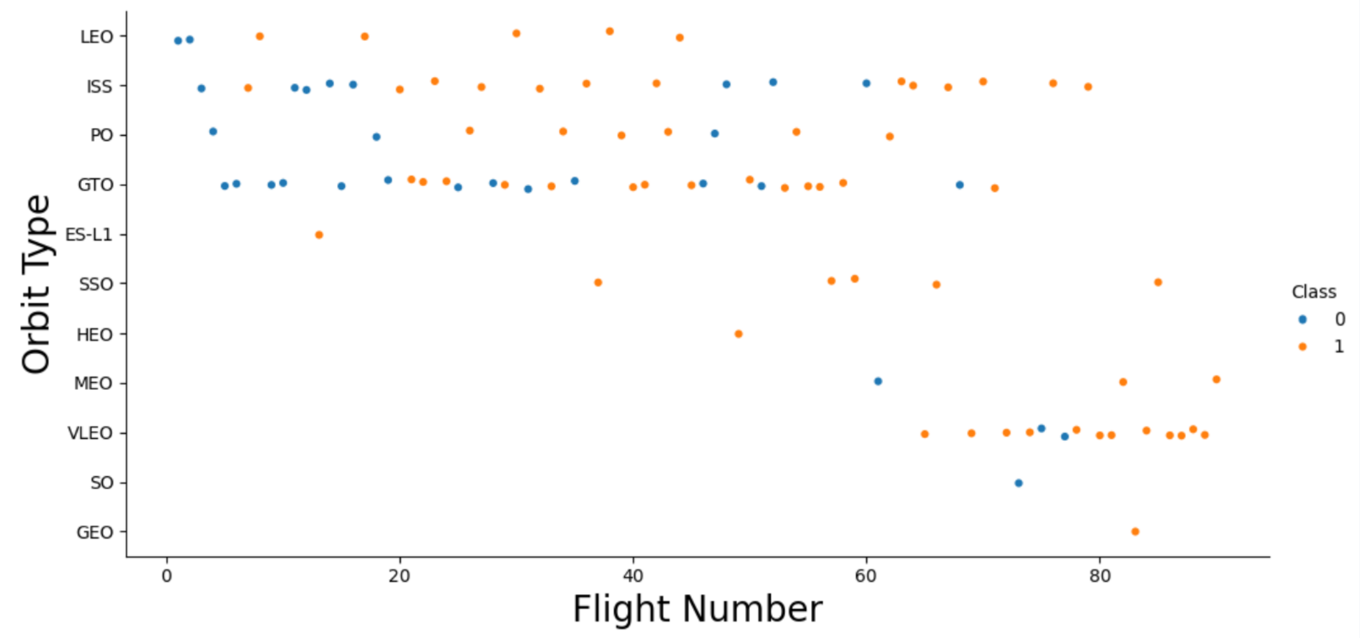


Success Rate vs. Orbit Type

- Orbit Types ES-L1, GEO, HEO and SSO have a perfect success rate (100%)
- SO has no successful rocket launches
- Except for SO, every orbit type has a success rate higher than 50%



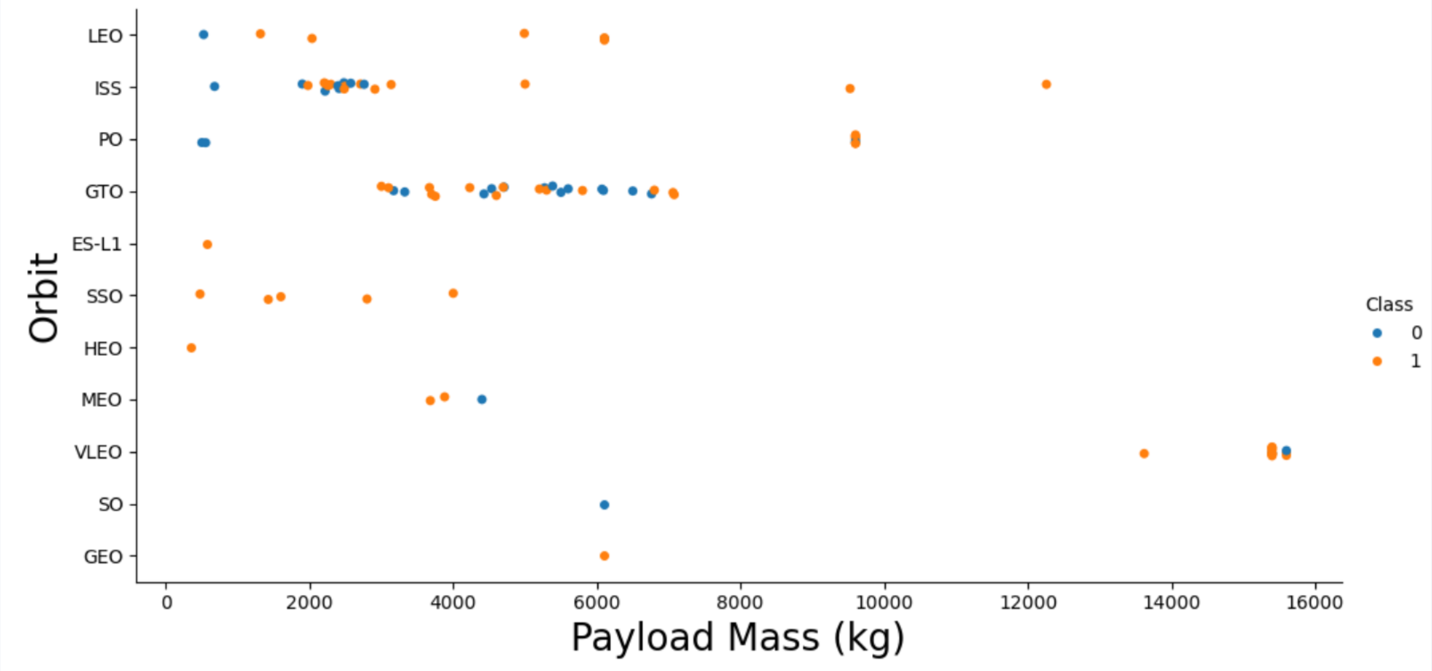
Flight Number vs. Orbit Type



- In the LEO orbit, success seems to be related to the number of flights
- Many orbit types are only represented in higher flight numbers

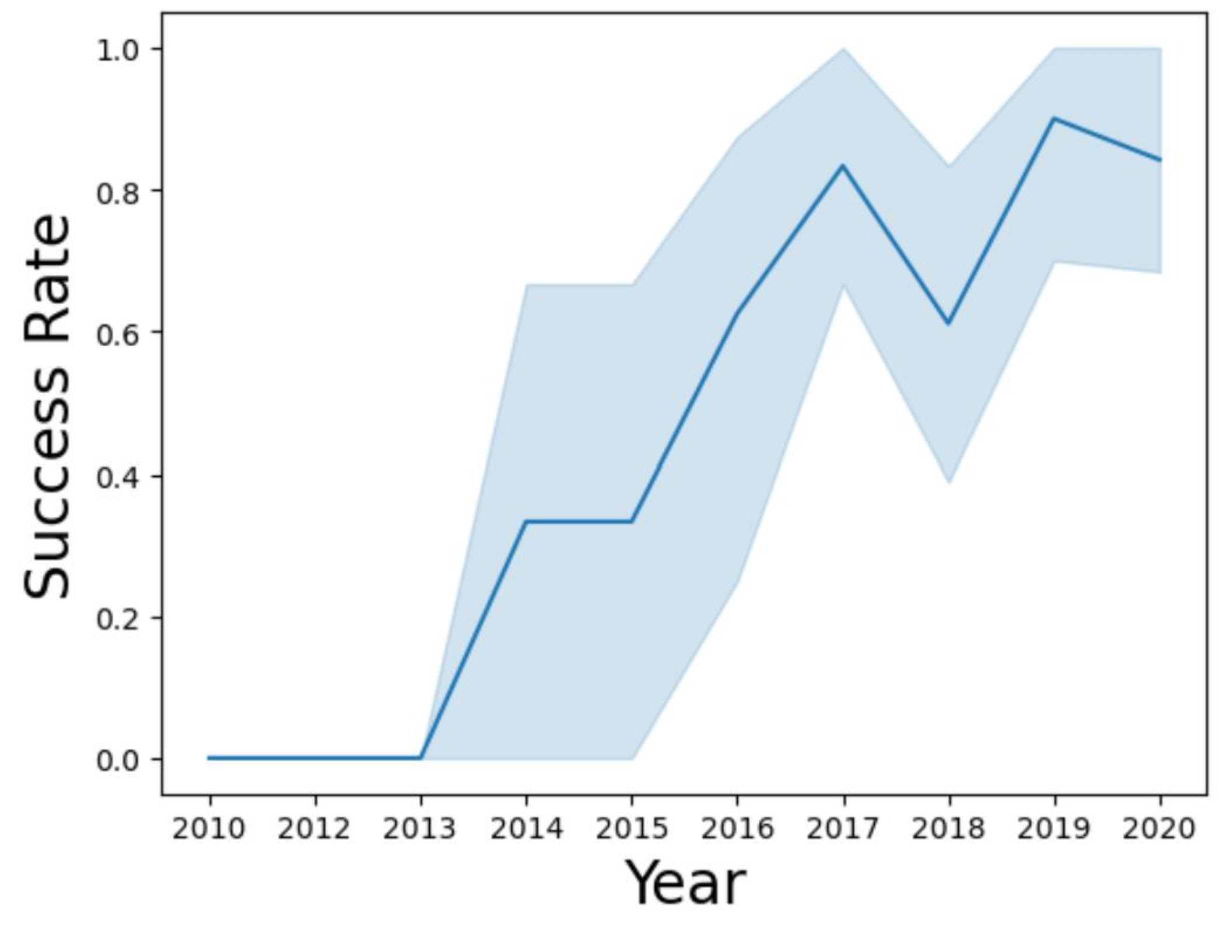
Payload vs. Orbit Type

- For Polar, LEO and ISS, the success rates increases with heavier payloads
- However, this trend does not apply to all orbits



Launch Success Yearly Trend

- The success rate increased drastically between 2013 and 2017
- In 2018, the success rate dropped by around 20%
- In 2020, the success rate was higher than 80%



All Launch Site Names

- There are four unique launch sites in the space mission
    - CCAFS LC-40
    - VAFB SLC-4E
    - KSC LC-39A
    - CCAFS SLC-40

```
%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'

```
%sql select * from SPACEXTBL 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_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

- The total payload carried by boosters from NASA is **45.596 kg**

```
%sql select Customer, sum(PAYLOAD_MASS__KG_) AS total_payload_mass from SPACEXTBL where Customer = 'NASA (CRS)'
```

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

Customer	total_payload_mass
NASA (CRS)	45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is **2.928,4 kg**

```
%sql select Booster_Version, avg(PAYLOAD_MASS__KG_) from SPACEXTBL \
where Booster_Version = "F9 v1.1";
```

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

Booster_Version	avg(PAYLOAD_MASS__KG_)
F9 v1.1	2928.4

First Successful Ground Landing Date

- **2015-12-22**: first successful landing outcome on ground pad

```
%sql select min(Date) as first_successfull_landing_groundpad from SPACEXTBL \
where Landing_Outcome LIKE 'Success (ground pad)'
```

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

first_successfull_landing_groundpad
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
%sql select Booster_Version, Landing_Outcome, PAYLOAD_MASS__KG_ from SPACEXTBL \
where Landing_Outcome like "Success (drone ship)" and PAYLOAD_MASS__KG_ between 4000 and 6000;
```

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

Booster_Version	Landing_Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes
  - Success: 99
  - Success (payload status unclear): 1
  - Failure: 1

```
%sql select Mission_Outcome, count(*) from SPACEXTBL group by Mission_Outcome;
```

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

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

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass (15.600 kg)

```
%sql select Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTBL \
    where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

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

- List of failed landing outcomes in drone ship in year 2015

```
%sql select substr(Date,6,2) as month, Date, Booster_Version, Launch_Site, Landing_Outcome \
from SPACEXTBL \
where Landing_Outcome = 'Failure (drone ship)' and substr(Date,0,5)='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)

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

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

```
%sql select Landing_Outcome, count(*) \
    from SPACEXTBL where Date between '2010-06-04' and '2017-03-20' \
    group by Landing_Outcome order by count(*) desc;
```

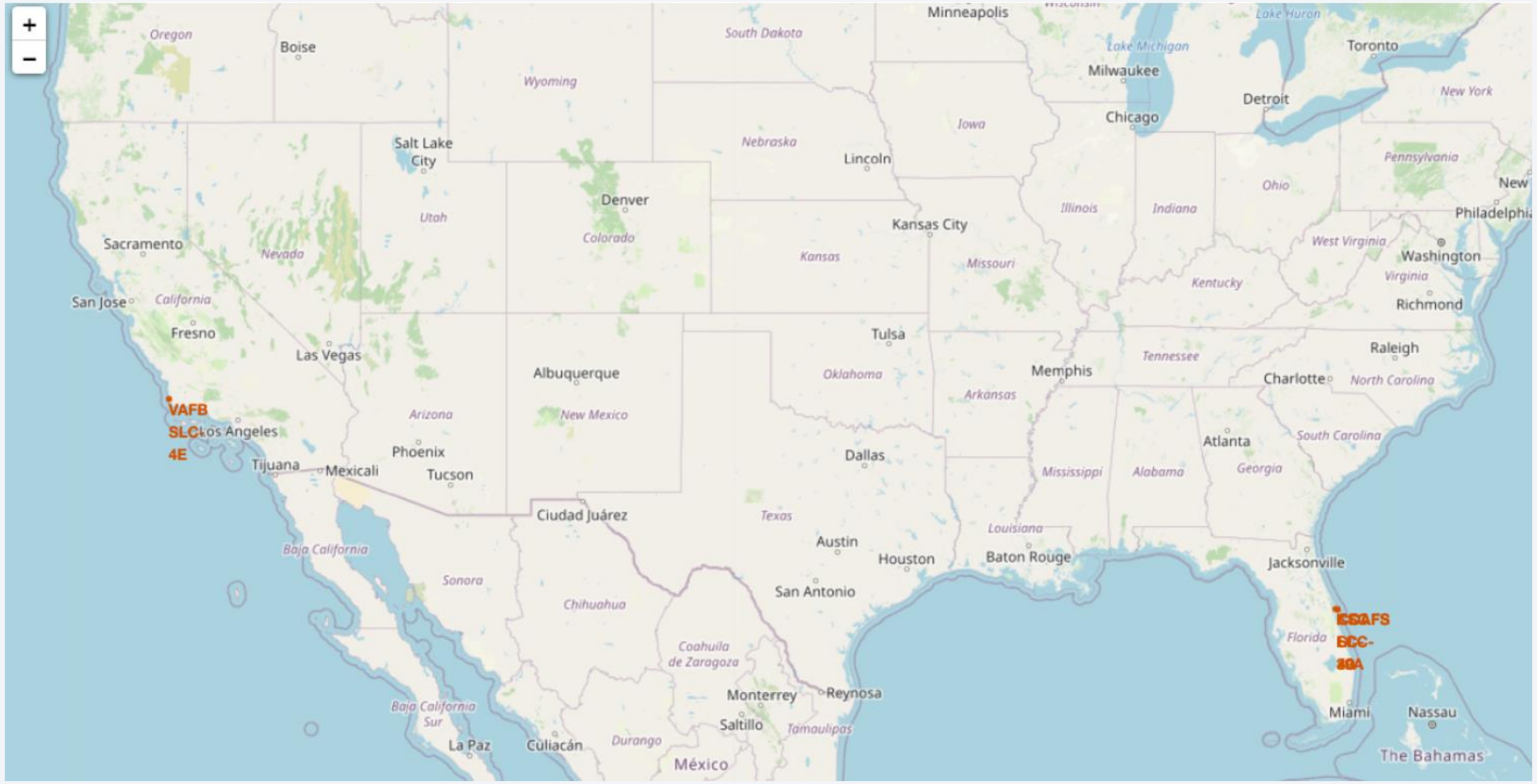
Landing_Outcome	count(*)
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

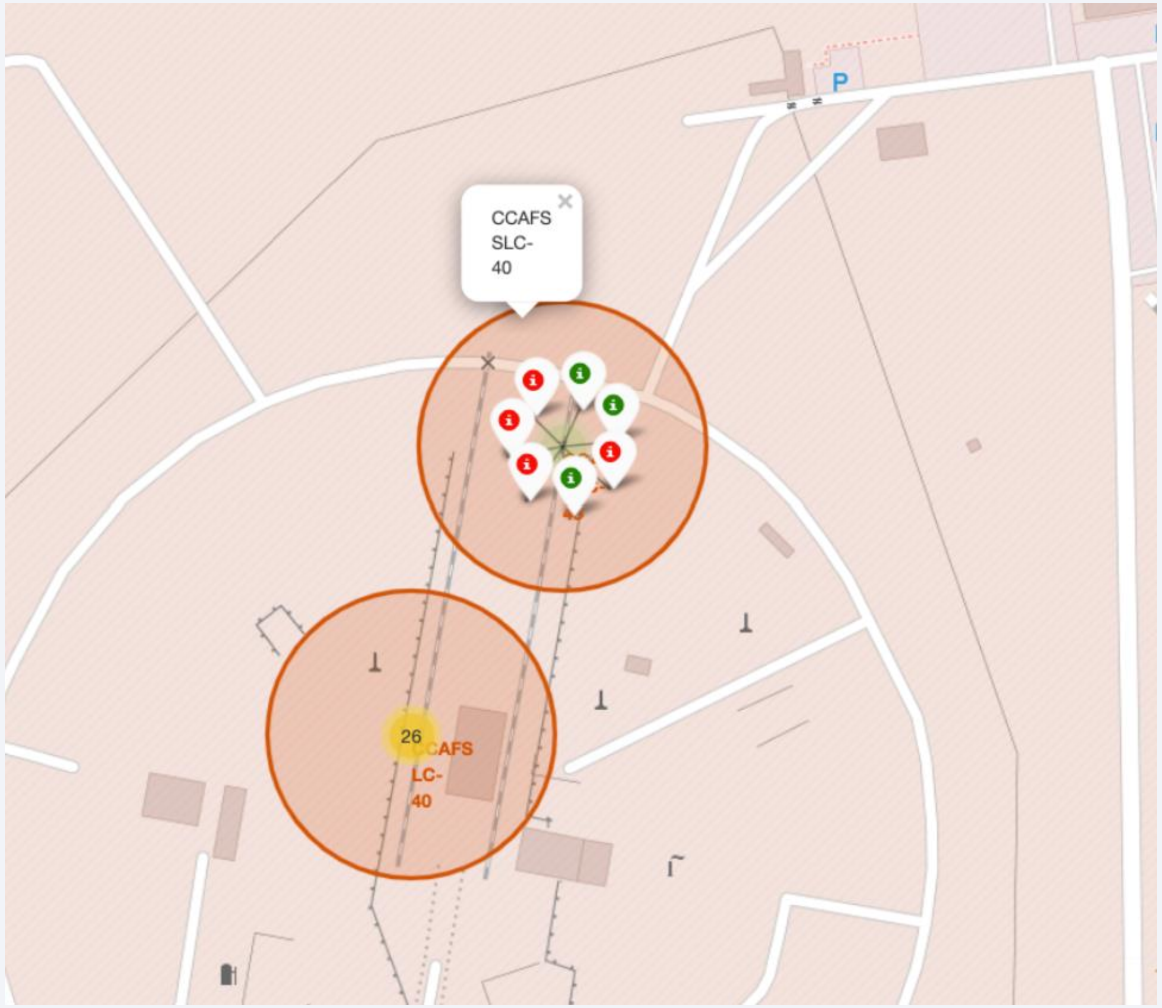
SpaceX Launch Site Locations

- Folium map shows the locations of the launch sites
- All launch sites are close to a coast in California and Florida
    - Reduce the risk to populated areas in case of launch failure



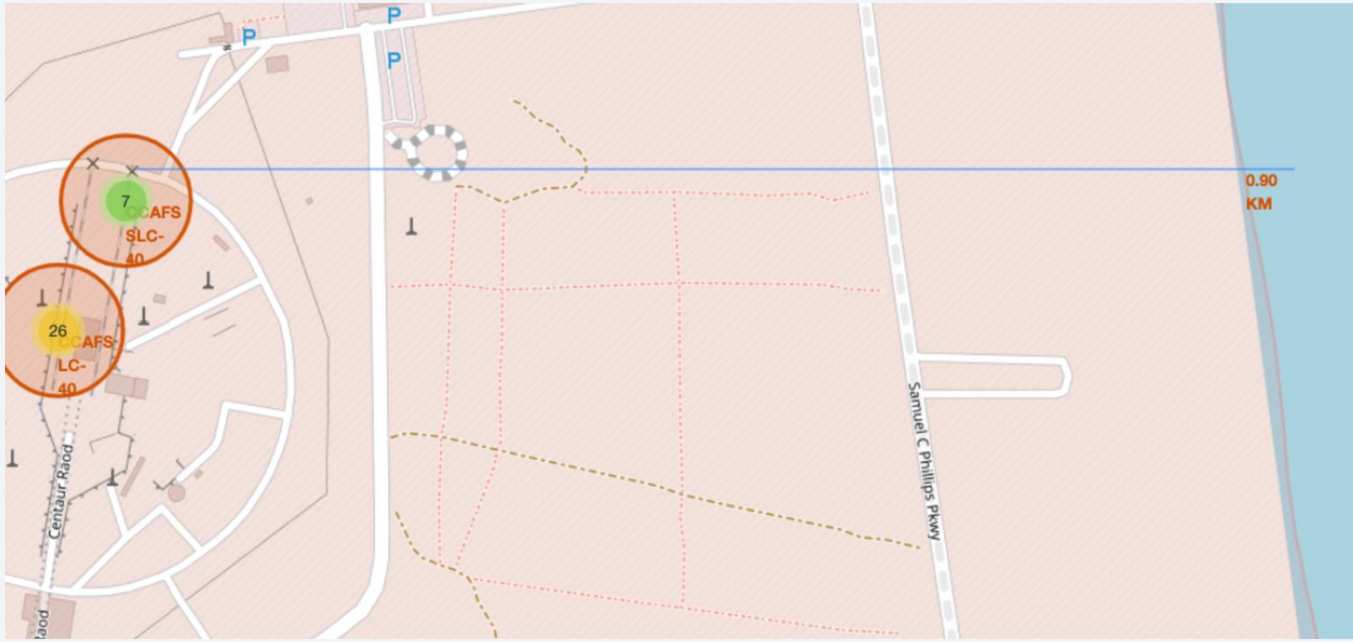
Launch outcomes of CCAFS SLC-40

- Folium map shows color-labeled launch outcomes of launch site **CCAFS SLC-40** in Florida
  - Green: successful launches
  - Red: unsuccessful launches



Explore Proximity of Launch Site

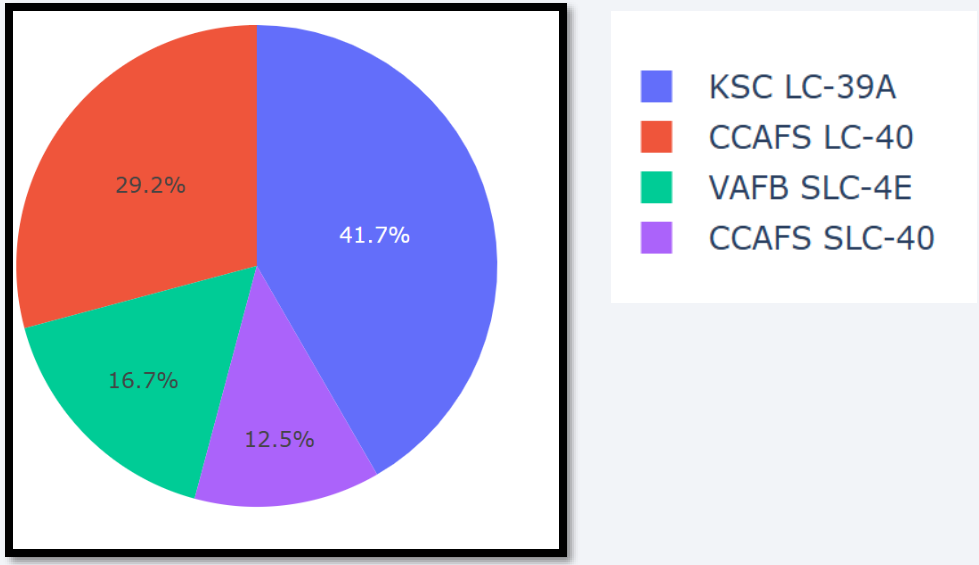
- Explore the proximities of launch site **AFS SLC-40** in Florida
- AFS SLC-40 is in close proximity to a coastline (0.90 km)



Section 4

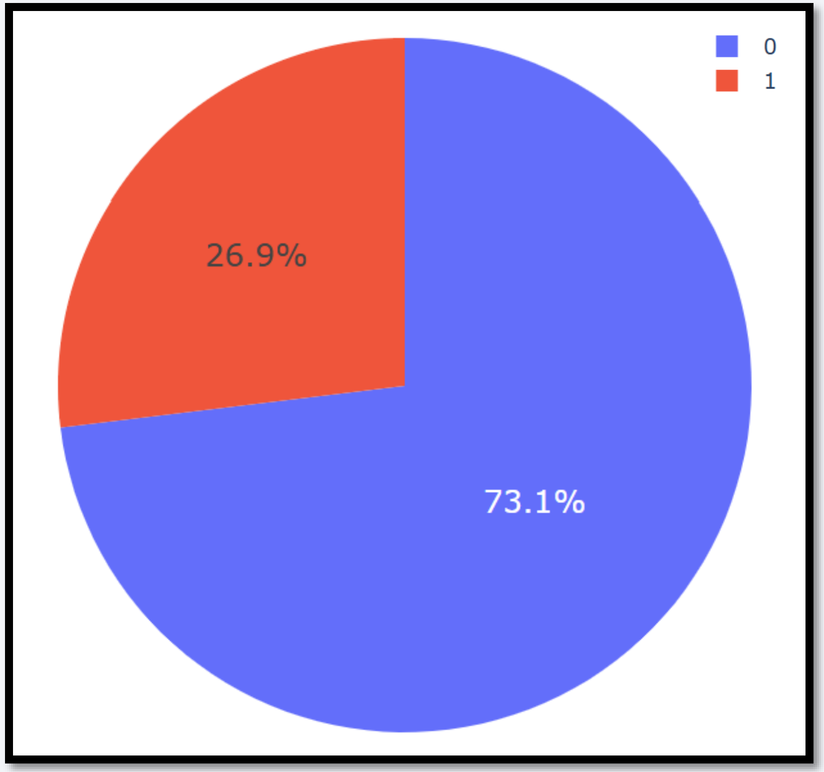
Build a Dashboard with Plotly Dash

Interactive Dashboard: Successful Landings



- Pie chart visualizes launch success count for all sites
- Launch site with highest success count: **KSC LC-39A**
- Launch site withl lowest success count: **CCAFS SLC-40**

Pie Chart: Launch Success Rate of CCAFS LC-40



- Pie Chart visualizes the success ratio of launch site CCAFS LC-40
    - Unsuccessful launches: 73.1%
    - Successful launches: 26.9%
- CCAFS LC-40 has the lowest success rate among all launch sites (26.9%)
- KSC LC-39A has the highest success rate (76.9%)

Scatter Plot: Payload vs. Launch Outcome

- Scatter plot visualizes the correlation between **Payload Mass** and **Launch Outcome** for all sites
  - The data points are color-labeled based on the booster version category

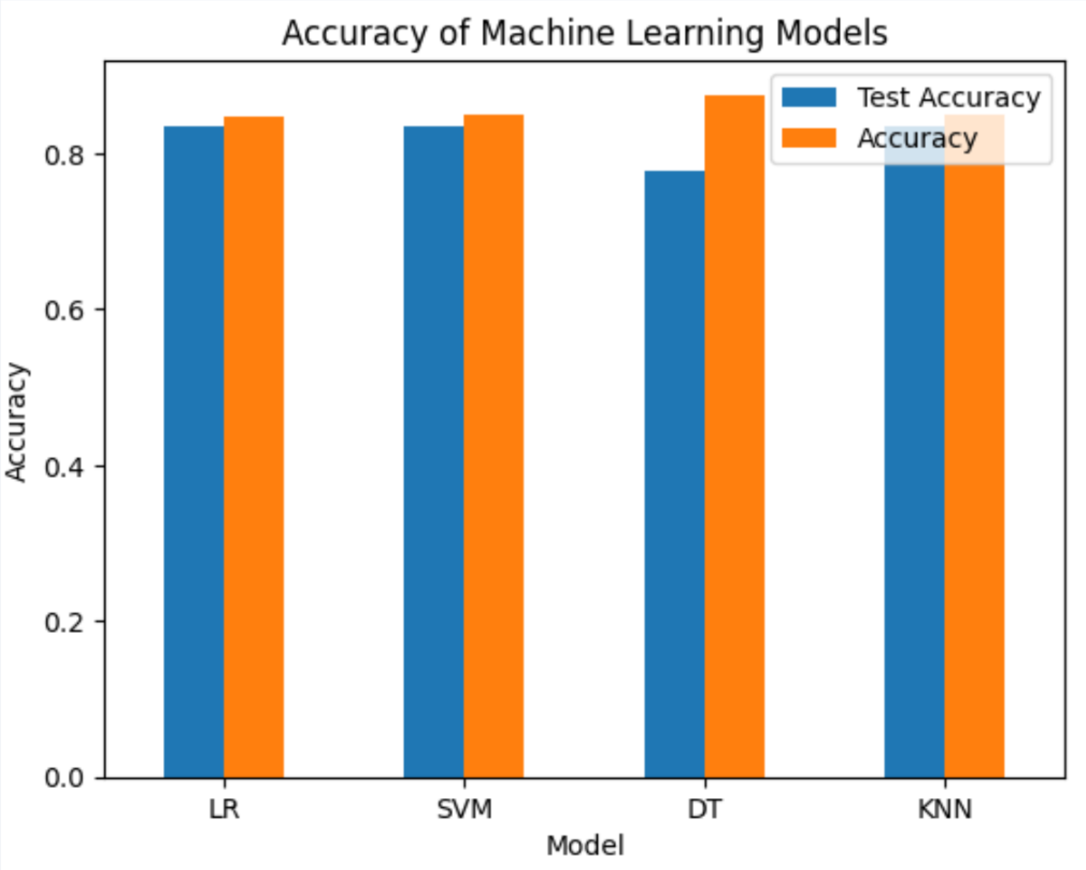
- Booster versions **B4** and **B5** have the highest success rates
- **FT** boosters have a high success rate for small payloads (< 6000 kg)

Section 5

Predictive Analysis (Classification)

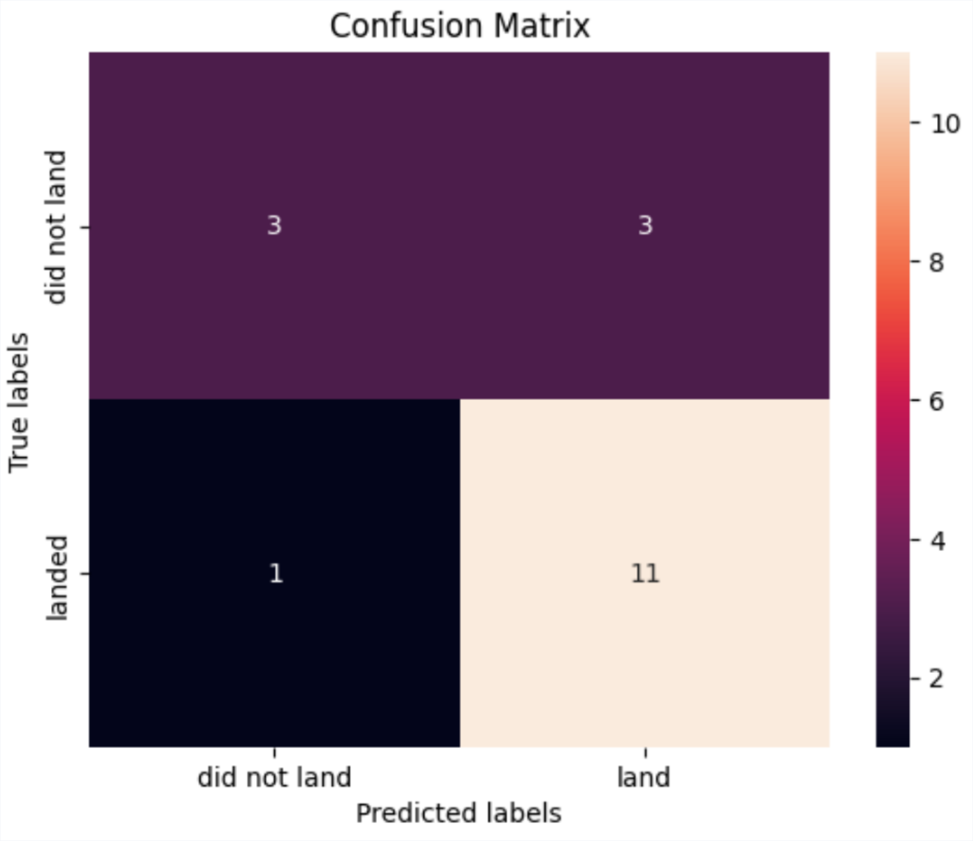
Classification Accuracy

- Bar Plot visualizes the **test accuracy** and the **accuracy on the validation data** of different machine learning models
- The Decision Tree (DT) has the highest accuracy on the validation data (87.5%)
- However, the Decision Tree also has the lowest test accuracy (77,8%) of all models
- Overall, all models have a similar performance



Confusion Matrix

- Confusion matrix of DecisionTreeClassifier provides summary of the performance on the test data
  - True Positives: 11
  - True Negatives: 3
  - False Positives: 3
  - False Negatives: 1



Conclusions

- The success rate of rocket launches increased over the last years (2013-2020)
    - 2020: more than 80% of rocket launches were successful
- The success rate is influenced by several factors (e.g. orbit, launch site, payload mass)
- Launch sites are typically close to the coast near the equator line
- We are able to accurately predict the first stage landing outcome with multiple machine learning models

Appendix

- Data Sources
  - SpaceX API (https://api.spacexdata.com/v4/rockets)
  - Wikipedia page titled List of Falcon 9 and Falcon Heavy launches (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
- Link to GitHub repository: GitHub-Repository

Thank you!