IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Luca Iacovella
26 Feb. 2024

Outline

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

Executive Summary

- Used methodologies
  - Data collection
    - SpaceX API
    - Wikipedia webscraping
  - Data wrangling in Pandas
  - Exploratory Data Analysis (EDA)
    - with SQL
    - with visualization
  - Interactive Visual Analytics with Folium and Plotly Dash
  - Machine learning model building in Scikit-Learn
- Results
  - Different graphs indicate correlation between launch properties and success rate
  - Decision tree predicts success well

Introduction

- SpaceX wants to bring affordable space travel
- Launches from different sites
- Each launch has different properties
- These properties might influence the success of a mission
- Goal: predict mission success based on launch properties

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - By making API-calls to SpaceX’s API
  - By scraping the Wikipedia article on Falcon launches
- Perform data wrangling
  - Extracted landing success status
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Trained and evaluated 4 different machine learning models

Data Collection

- SpaceX API data:
  - Source: https://api.spacexdata.com/v4/
  - Collected using API-calls
  - Additional API-calls for more detailed information

- Wikipedia article Falcon launches:
  - Source: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
  - Collected using BeautifulSoup4 webscraping
  - Manual cleaning was then performed

Data Collection – SpaceX API

GET /launches/past

API-call to get historic overview of rocket launches

For each rocket
GET /rockets/rocket

API-call to get the rocket name

For each launchpad
GET /launchpads/launchpad

API-call to get the geolocation of the launchpad

For each payload
GET /payloads/payload

API-call to get the payload (in kg) and orbit

[GitHub URL]

Data Collection - Scraping

GET Falcon launches Wikipedia article

Get the Wikipedia article and parse it using BeautifulSoup

Extract and clean columns

Clean columns (by removing unnecessary HTML) and remove unused columns

Parse table records into DataFrame

Iterate over soup object, clean the values and insert into Pandas DataFrame

Save result as CSV

Store the resulting DataFrame as CSV for later use

GitHub URL

Data Wrangling

- Calculated amount of launches per site
- Calculated occurrence of each orbit
- Calculated amount of mission outcomes per type
- Calculated success rate across all missions

GitHub URL

EDA with Data Visualization

- The visaluation EDA mainly consisted finding correlation with success rates
    - Per launch site, influence of flight number on success rate
    - Per launch site, influence of payload mass on success rate
    - Influence of orbit type on success rate
    - Per orbit type, influence of flight number on success rate
    - Per orbit type, influence of payload mass on success rate
    - Influence of launch year on success rate

GitHub URL

EDA with SQL

- SQL was used to query the launch data with regards to
    - Launch sites
    - Payload mass
    - Booster versions
    - Mission outcomes

GitHub URL

Build an Interactive Map with Folium

- Folium was used to visualize the launch sites on a map
- The interactive map included the amount of launches per location
- The proximity of the following was studies:
    - Coasts
    - Railways
    - Highways
    - Cities

[GitHub URL]

Build a Dashboard with Plotly Dash

- Study success rate of
  - Launch sites
  - Payload mass
  - Booster version category
- Launch site success with Pie chart for
- Payload mass and booster version category with scatter plot to distinguish good and bad ranges

[GitHub URL]

Predictive Analysis (Classification)

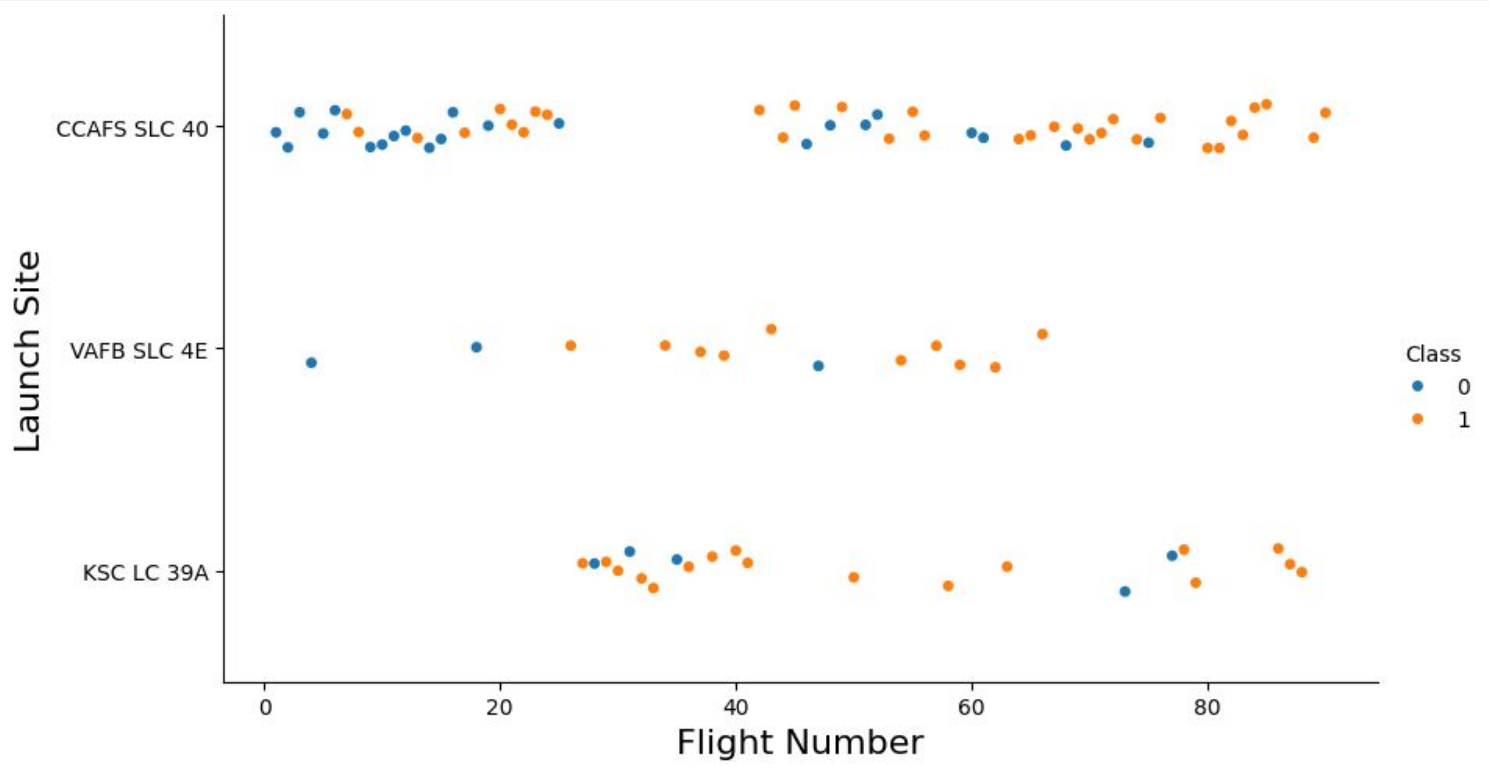
- One-hot encoded features
- Separate test set from training set
- Standardize data
- Fit training data to different models
- Evaluate models using test data

GitHub URL

Section 2

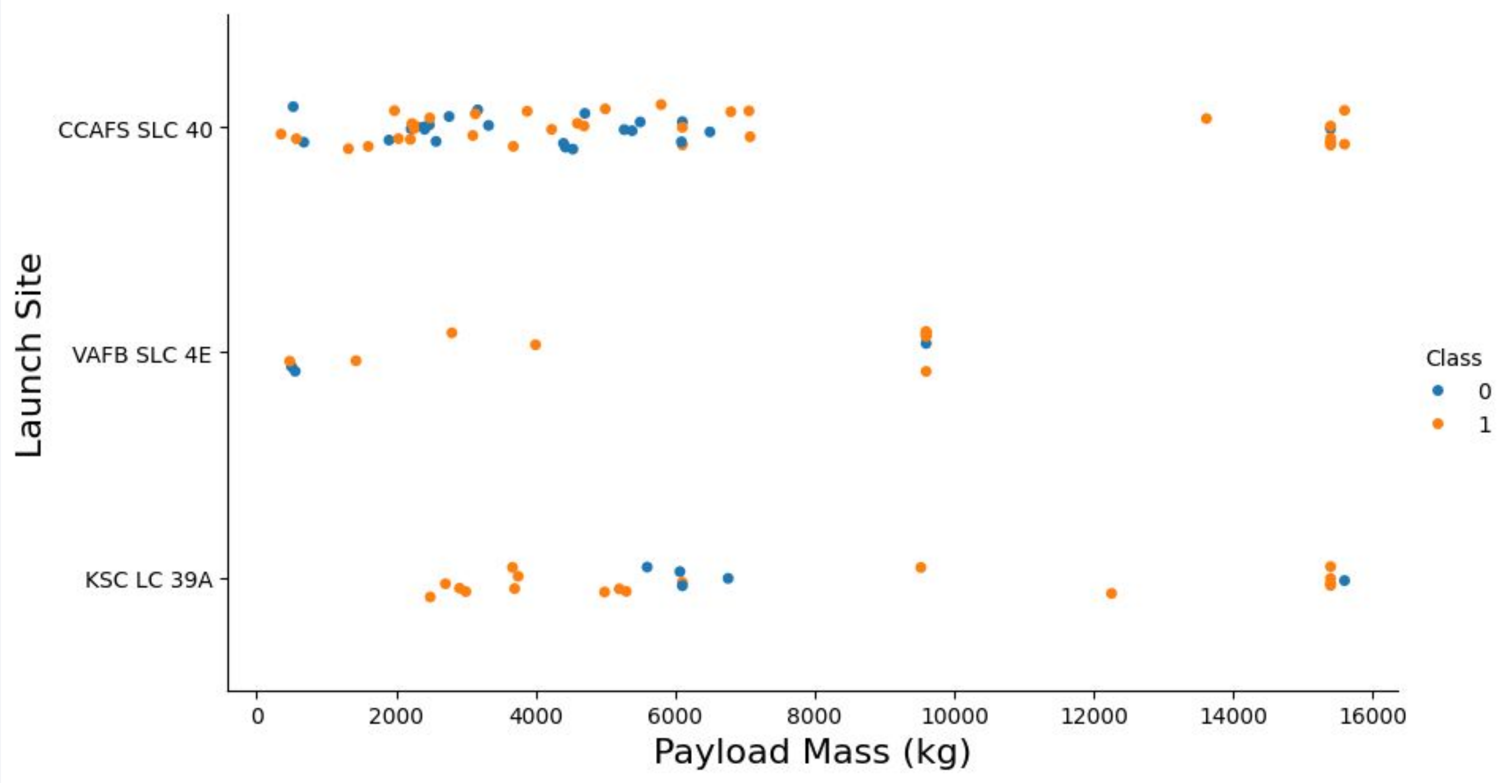
Insights drawn from EDA

Flight Number vs. Launch Site



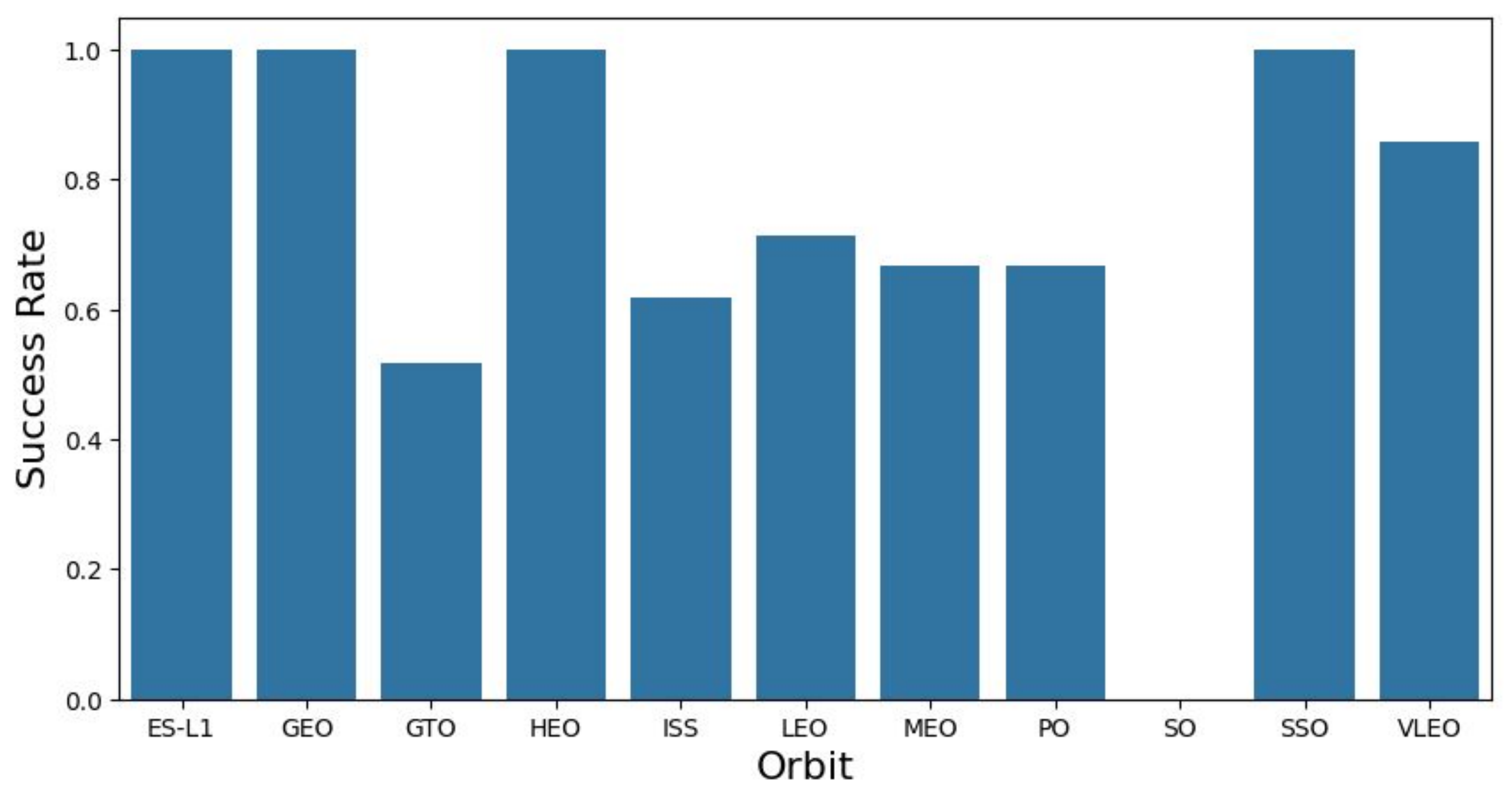
- Orange dots indicate successful launches
- A higher flight number correlates with success rate
- This trend is true for all launch sites

Payload vs. Launch Site



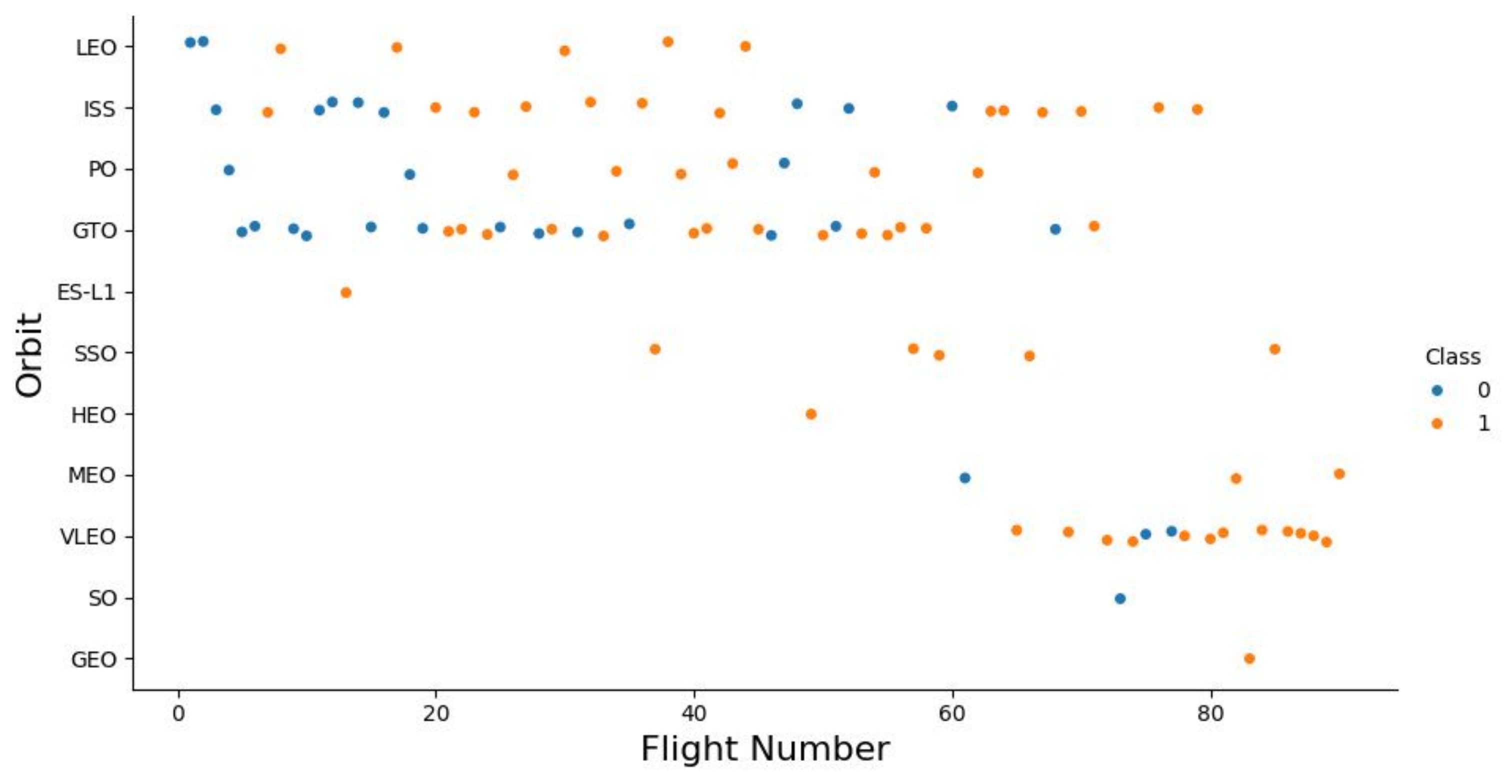
- VAFB SLC launch site seems to have a max payload of 10,000 kg
- Higher payload mass seems to correlate with higher success rate, for all launch sites
- This could be due to low payload mass implying early test launches

Success Rate vs. Orbit Type



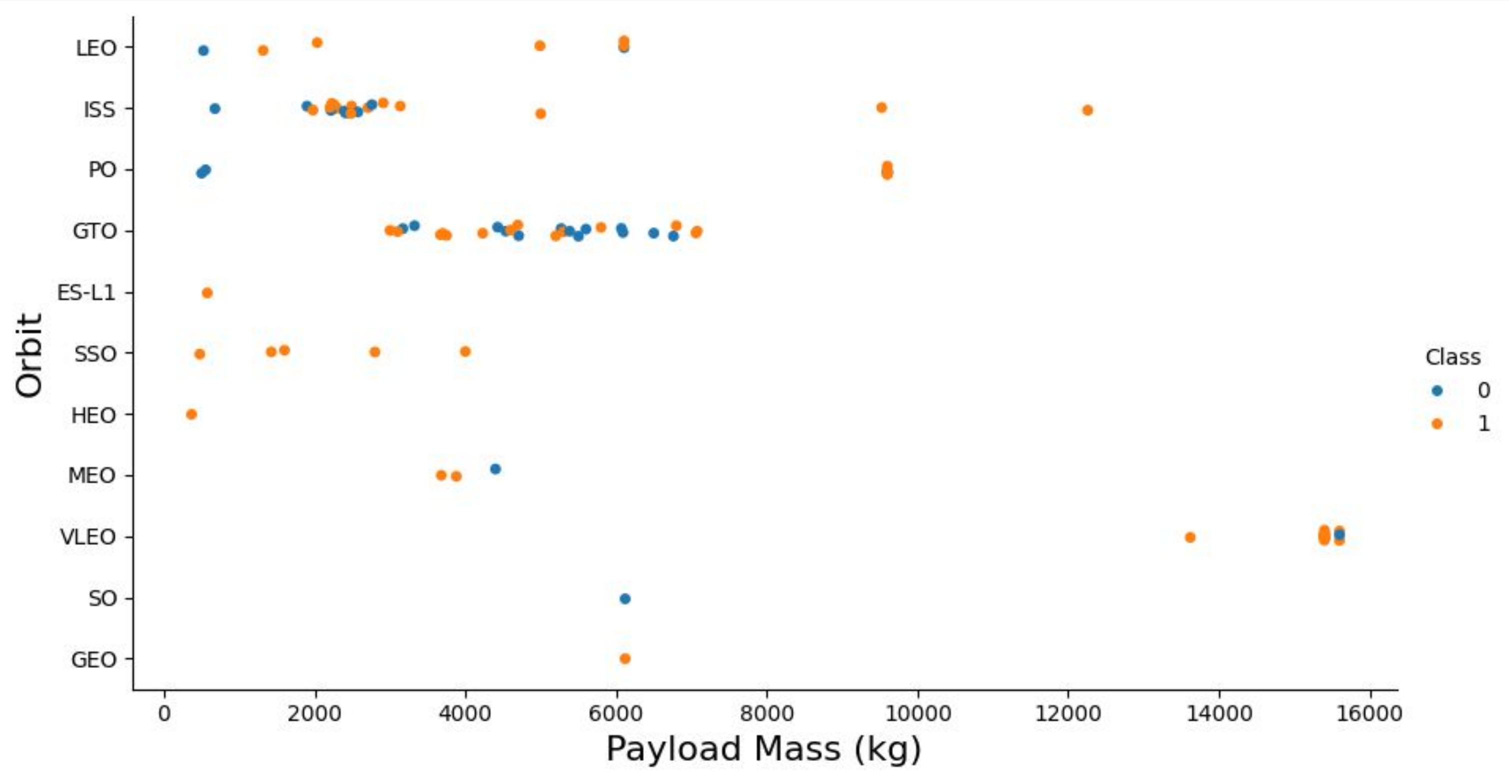
- SO orbit has not been achieved
- ES-L1, GEO, HEO, SSO always succeeded
- The other orbit success rates range from 50% to 90%

Flight Number vs. Orbit Type



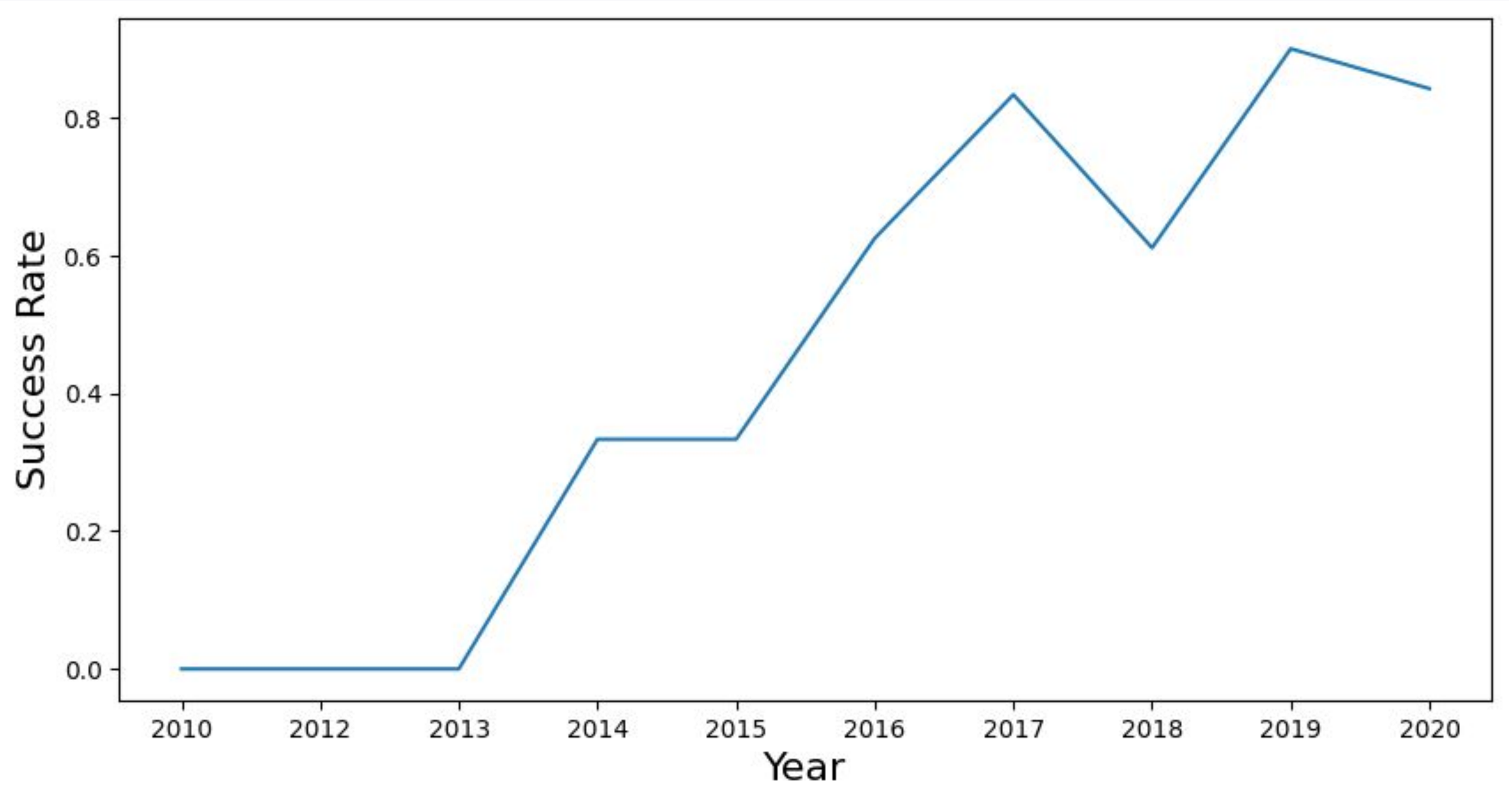
- Success rate and flight number relation depends on orbit, e.g.:
- LEO orbit success relates to higher flight number
- GTO orbit success rate and flight number seem unrelated

Payload vs. Orbit Type



- Success rate and payload relation dependent on orbit too
- No relation for GTO
- Relation for LEO, ISS, PO

Success rate yearly trend



- As the years pass, success rate increases
- General trend, not perfect

All Launch Site Names

```
SELECT DISTINCT Launch_Site
FROM SPACEXTBL
```

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

- Use the DISTINCT keyword to get all unique Launch_Sites

Launch Site Names Begin with 'CCA'

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

- Used the LIKE keyword to match pattern
- Used % operator to match multiple characters
- Used LIMIT keyword to keep 5 records

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

```
SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD_MASS
FROM SPACEXTBL
WHERE Customer = 'NASA (CRS)'
```

TOTAL_PAYLOAD_MASS
45596

- Used SUM aggregator function to add all payload mass values

Average Payload Mass by F9 v1.1

```
SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD_MASS
FROM SPACEXTBL
WHERE Booster_Version = 'F9 v1.1'
```

AVG_PAYLOAD_MASS
2928.4

- Used AVG aggregator function to get the average payload mass value

First Successful Ground Landing Date

```sql
SELECT MIN(Date)
FROM SPACEXTBL
WHERE Landing_Outcome = 'Success (ground pad)'
```

MIN(Date)
2015-12-22

- Used the MIN aggregator function to get the lowest date value

Successful Drone Ship Landing with Payload between 4000 and 6000

```
SELECT DISTINCT Booster_Version
FROM SPACEXTBL
WHERE Landing_Outcome = 'Success (drone ship)'
GROUP BY Booster_Version
HAVING SUM(PAYLOAD_MASS__KG_) BETWEEN 4000 AND 6000
```

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

- Grouped by Booster_Version
- Filtered each group using SUM aggregator
- Used BETWEEN keyword to define necessary range

Total Number of Successful and Failure Mission Outcomes

```
SELECT Mission_Outcome, COUNT(*) AS COUNT
FROM SPACEXTBL
GROUP BY Mission_Outcome
```

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

- Grouped by Mission_Outcome
- Aggregated using COUNT function to get the amount per outcome

Boosters Carried Maximum Payload

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

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

- Used sub-query to get maximum payload mass
- Filtered Booster_Versions that have this maximum payload mass

2015 Launch Records

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

Month	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

- Used substr function to get the Month and Year position from the Date string

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

```
SELECT Landing_Outcome, COUNT(*)
FROM SPACEXTBL
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY Landing_Outcome DESC
```

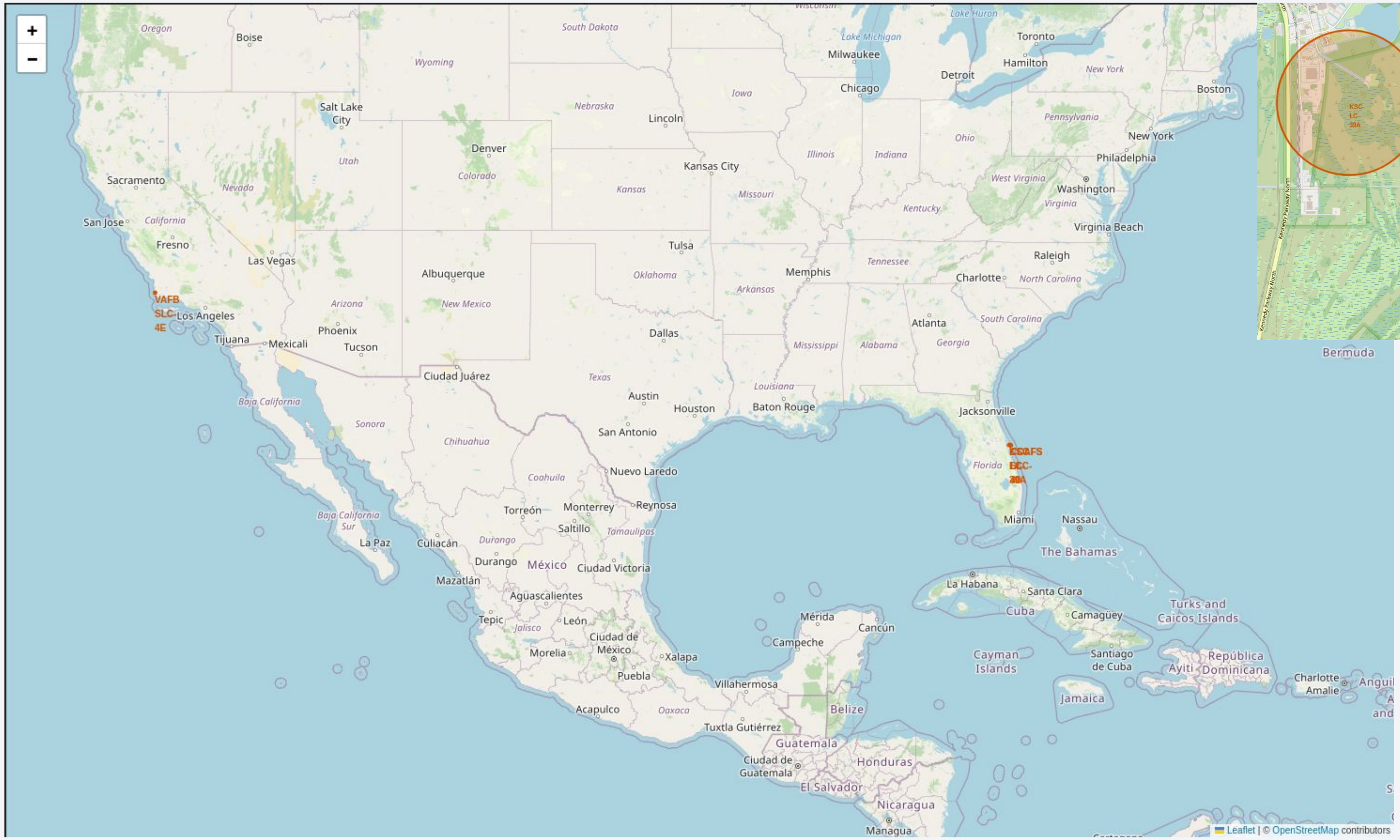
Landing_Outcome	COUNT(*)
Uncontrolled (ocean)	2
Success (ground pad)	3
Success (drone ship)	5
Precluded (drone ship)	1
No attempt	10
Failure (parachute)	2
Failure (drone ship)	5
Controlled (ocean)	3

- Used lexicographical property of Date format to filter based on Date range
- Ordered by descending Landing_Outcome label

Section 3

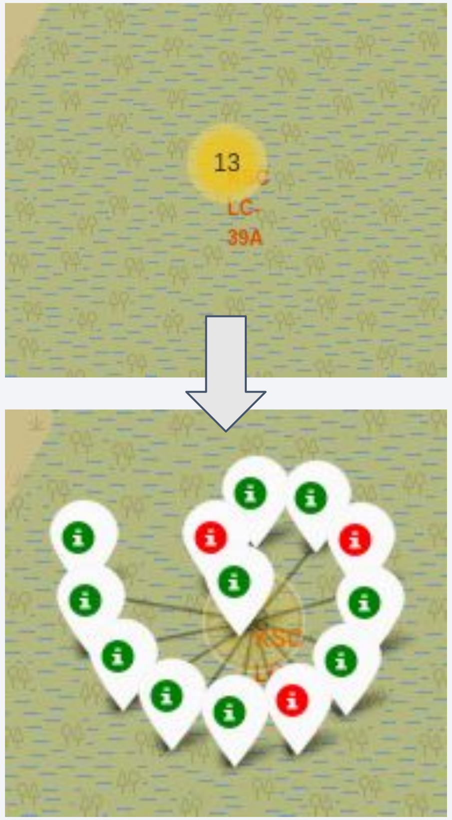
Launch Sites Proximities Analysis

Launch Sites on Map



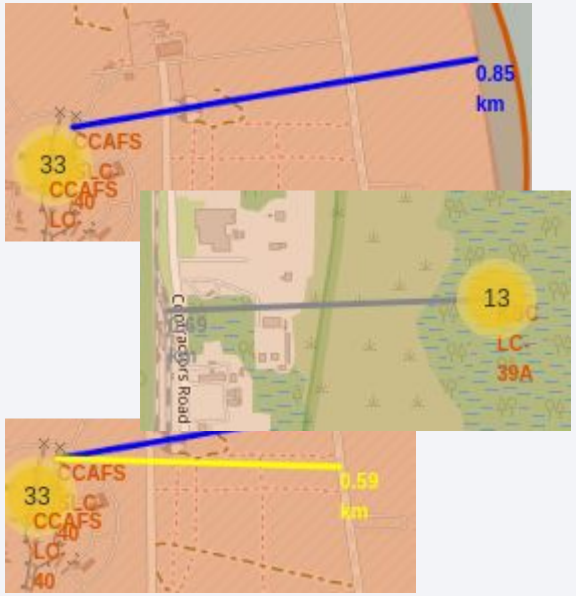
- Launch sites on map
  - Highlighted with circle
  - Labelled by name

Launch Outcomes per Launch Site

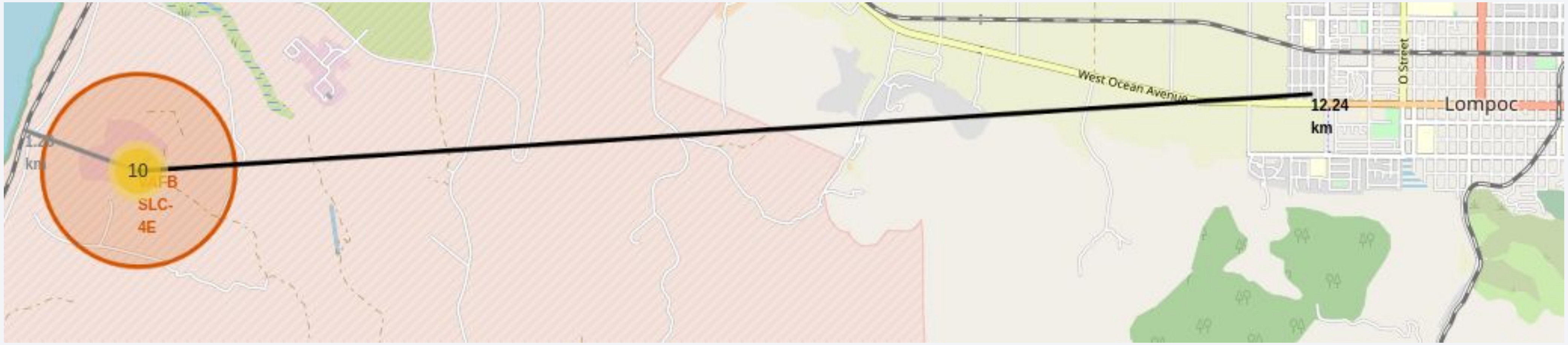


- Launch sites have a number in yellow circle for the total amount of launches
- When clicked, it reveals the successful (green) and failed (red) launches

Launch Site Proximity



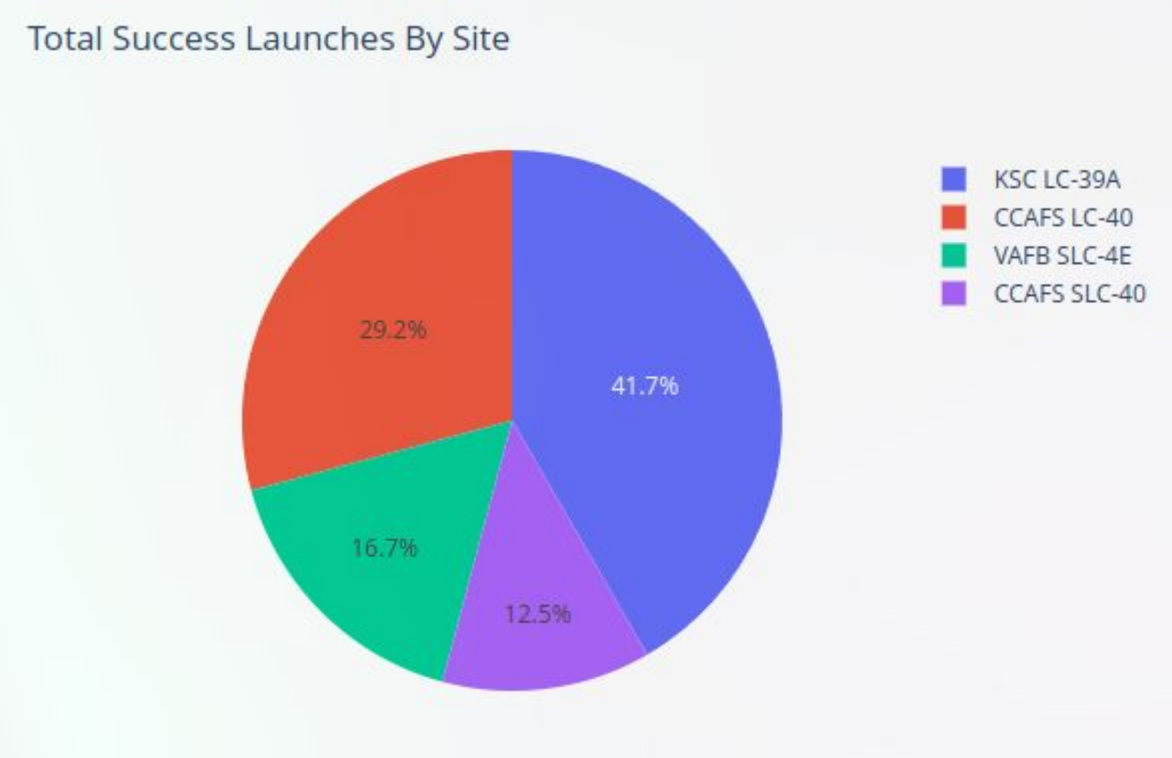
- Launch sites are close to coasts, railways and highways
  - Close to coast might be for safety (in case of failure) if the launches are directed towards the coast
  - Proximity might be for logistic reasons: ships, trains and trucks might be used to deliver rocket parts
- Launch sites seem to be at least 10 km from cities with residential areas, likely for safety



Section 4

Build a Dashboard with Plotly Dash

Launch Site Success Rate

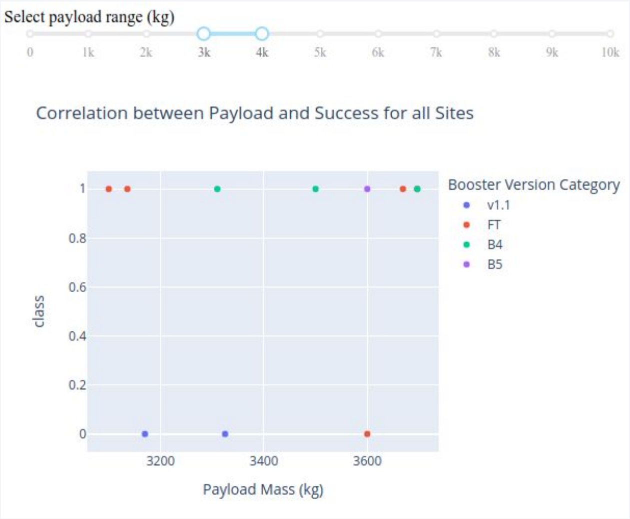


- Pie chart representing number of successful launches per launch site
- KSC LC-39A has the most successful launches, CCAFS SLC-40 the least

KSC LC-39A Success Rate

- Pie chart representing the successful launches and failed launches of KSC LC-39A
- This launch site has the highest success rate (also highest total successful launches)

Payload Mass and Booster Version Cat.



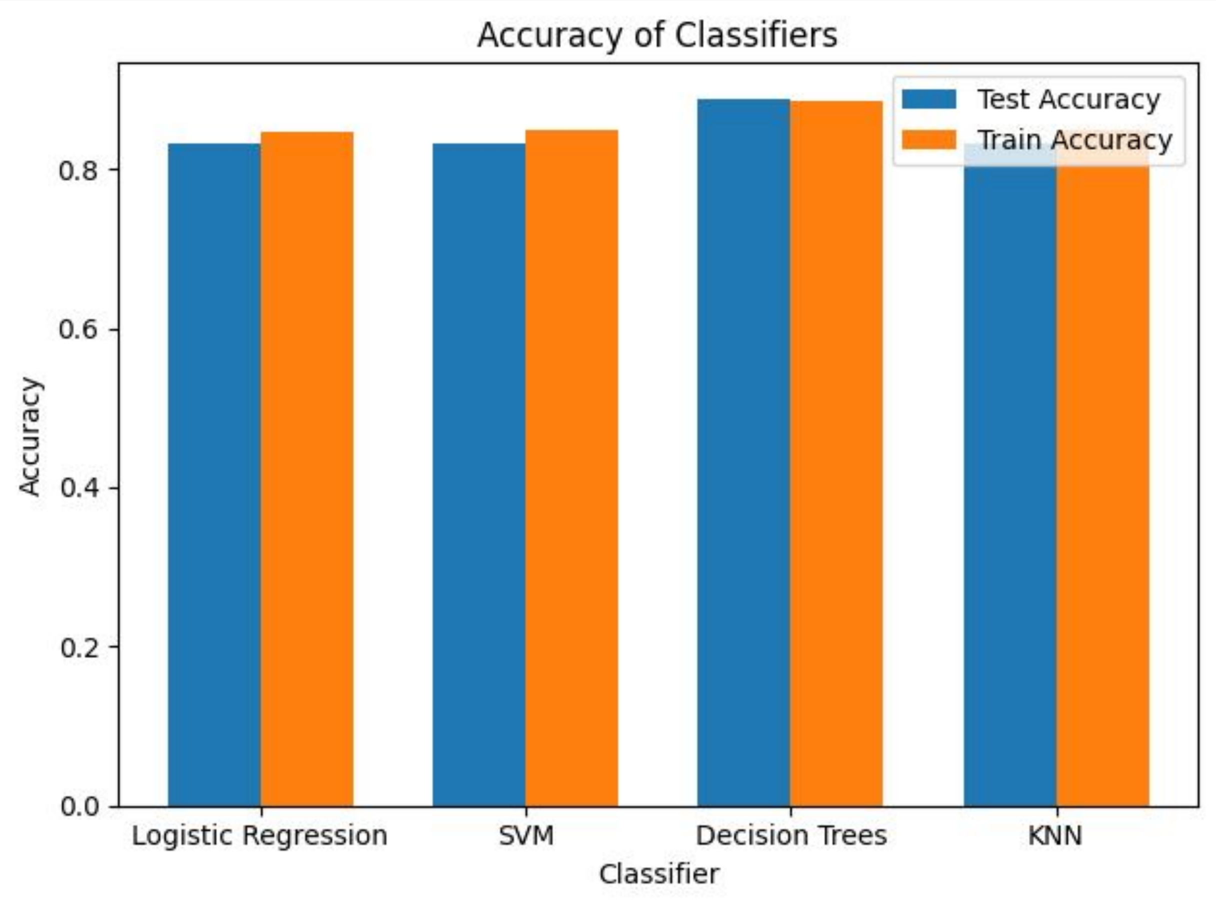
- Each dot represents a launch, colored by the booster version category
- Y-axis represents if the launch was a success
- X-axis represents the payload mass
- The F9 FT booster has the highest success rate
- The highest success rate is found between 3k-4k kg and the lowest between 6k-8k kg

Section 5

Predictive Analysis (Classification)

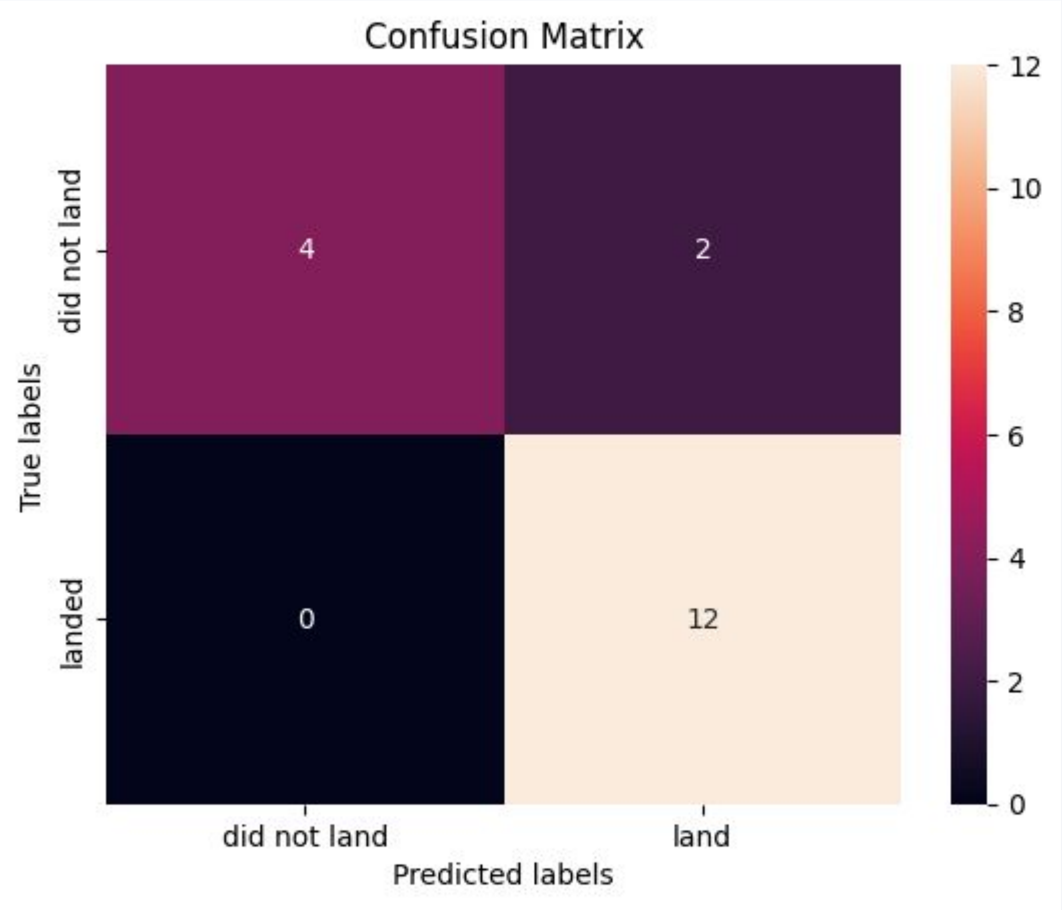
Classification Accuracy

- Classification of landing outcome
- One-hot encoding of features
- Split data in train and test set
- Trained and evaluated using 4 different types of models
  - Logistic regression
  - SVM
  - Decision tree
  - kNN



Confusion Matrix

- Best model: Decision tree
- Acceptable performance: 89%
- Good recall
- Precision needs improvement
  - Model does not predict well in case of non-landing
  - Too many false positives



Conclusions

- We used geodata to find out more properties of the launch sites
- Orbit type, flight number, launch year, payload mass seem to correlate with success rate
- Some launch sites had better success rates than others
- The decision tree model had the highest classification accuracy

Thank you!