IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Denis Kolenko
30.07.2024

Outline

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

Executive Summary

Summary of methodologies

- Data Collection
- Data Wrangling
- Web Scraping
- Exploratory analysis using Pandas, Seaborn, Matplotlib
- Exploratory with SQL
- Building a dashboard with Plotly Dash
- Building map with Folium
- Predictive analysis (Classifications, Machine learning)

Summary of all results

In-Depth Data Insights

- Uncovered key trends and patterns in the data.

Interactive Dashboards and Maps

- Developed dashboards and maps for easy analysis and visualization.

Machine Learning Model Performance

- Models achieved excellent results and high metrics.

Successful Web Scraping

- Extracted data from web sources, enriching our analysis.

Key Findings from SQL Analysis

- SQL queries highlighted key dependencies and anomalies.

Introduction

- SpaceX advertises Falcon 9 rocket launches at a cost of $62 million, whereas it costs upwards of $165 million for each launch with other providers.
- SpaceX is the only private company to ever return a spacecraft from low-Earth orbit, which they first accomplished in December 2010.
- SpaceX has revolutionized the space industry by developing reusable rockets. Their Falcon 9 rockets are designed to be reused multiple times, significantly reducing the cost of space travel. The first successful landing of a Falcon 9 rocket's first stage occurred in December 2015.
- SpaceX is actively developing the Starship spacecraft, designed for missions to Mars and beyond. Starship is intended to be a fully reusable spacecraft capable of carrying large numbers of passengers and cargo, potentially enabling human colonization of other planets.

The project task is to predict if the first stage of SpaceX Falcon 9 will land successfully.

Section 1

Methodology

Methodology

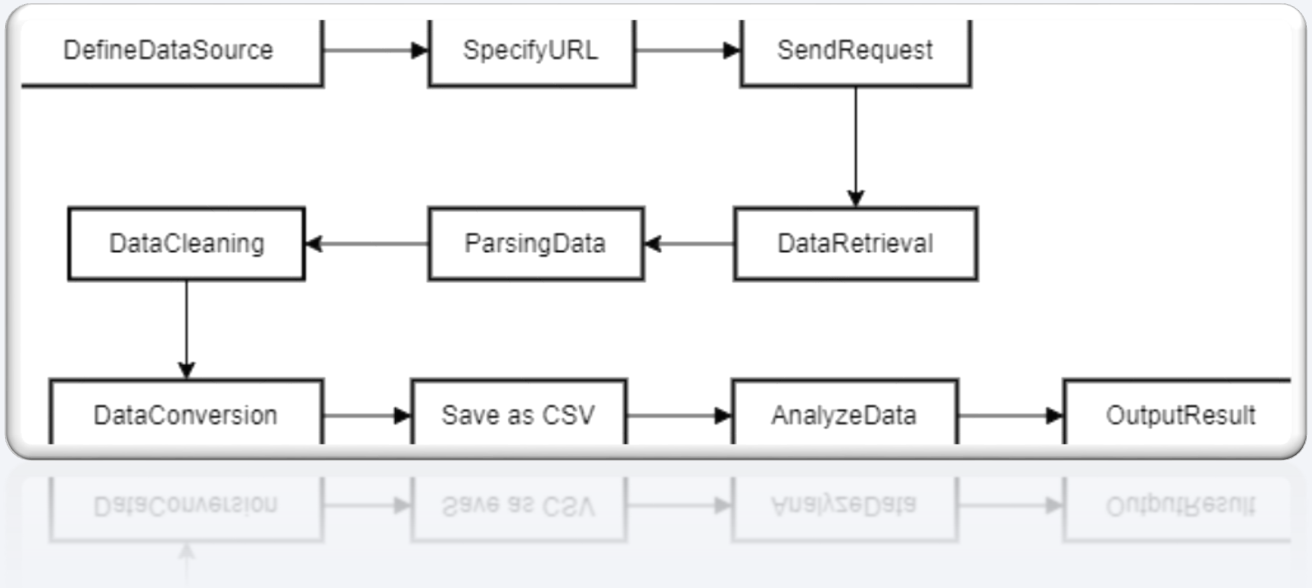
Executive Summary

- Data collection methodology:
  - Web scraping to collect Falcon 9 launch records from [Wikipedia]
  - Get request to SpaceX API.
- Perform data wrangling:
  - A classification variable "landing_class" *(0 = unsuccessful, 1 = successful)* was created based on the variable "Outcome". Data were cleaned of omissions and irrelevant columns.
- Perform exploratory data analysis (EDA) using visualization and SQL.
- Perform interactive visual analytics using Folium and Plotly Dash.
- Perform predictive analysis using classification models:
  - The classification models *(LR, KNN, SVM, DT)* were analyzed, and their accuracy was evaluated on test data to determine the best option.

Data Collection

Two main sources can be used to collect SpaceX launch data:

- The SpaceX REST API (*URL*) provides information about rockets, payloads delivered, launch characteristics, and landing results.
- Falcon 9 launch data can also be extracted using Wikipedia web scraping with the BeautifulSoup library.



Module 1

Data Collection – SpaceX API

[GitHub]

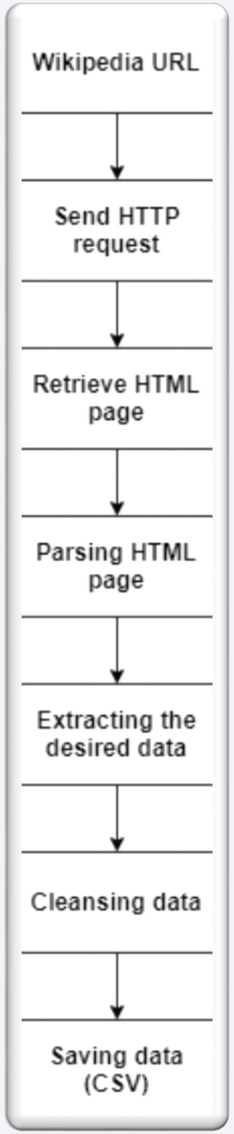
- Get response from API. *response = requests.get (URL)*
- Converting response to JSON. data = response.json(); data=pd.json_normalize(data)
- Filter data using custom functions. getLaunchSite(data); getPayloadData(data); getCoreData(data)
- Missing values for 'PayloadMass' set replaced with the mean of 'PayloadMass'. PayLoadMean = data_falcon9[‘PayloadMass’].mean()
- Final step exploit data in CSV files. data_falcon9.to_csv(‘dataset_part_1.csv’, index=False’)

Read more on the github

Data Collection - Scraping

Falcon 9 historical launch records were collected from a Wikipedia page called "List of Falcon 9 and Falcon Heavy launches". (URL)

	Flight No.	Date	Time	Version Booster	Launch Site	Payload	Payload mass	Orbit	Customer	Launch outcome	Booster landing
0	1	4 June 2010	18:45	F9 v1.0B0003.1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	Failure
1	2	8 December 2010	15:43	F9 v1.0B0004.1	CCAFS	Dragon	0	LEO	NASA	Success	Failure
2	3	22 May 2012	07:44	F9 v1.0B0005.1	CCAFS	Dragon	525 kg	LEO	NASA	Success	No attempt\n
3	4	8 October 2012	00:35	F9 v1.0B0006.1	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	No attempt
4	5	1 March 2013	15:10	F9 v1.0B0007.1	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	No attempt\n



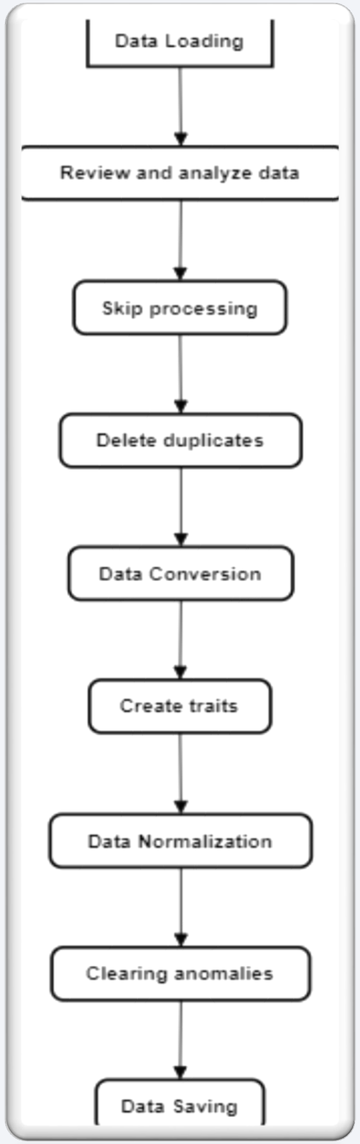
GitHub

Data Wrangling

Exploratory data analysis (EDA) was conducted to identify patterns and define learning labels. A classification variable was created to represent the outcome of the launch, with "0" indicating a failed first stage launch and "1" indicating a successful launch. These values were added to the "Class" column of the data frame.

```
0 True ASDS
1 None None
2 True RTLS
3 False ASDS
4 True Ocean
5 False Ocean
6 None ASDS
7 False RTLS
```

	Class
0	0
1	0
2	0
3	0
4	0
5	0
6	1
7	1



[GitHub]

EDA with Data Visualization

- **Line Chart**
  - To visualize the annual trend of launch success.
- **Bar plot**
  - To visualize the relationship between the success rate of each type of orbit.
- **Catplot**
  - To visualize the relationship between payload and launch site, flight number and orbit type, and payload and orbit type.

GitHub

EDA with SQL

The following SQL queries were executed during data analysis:

1. Identify unique launch sites for space missions
2. Search for launch sites starting with "CCA"
3. Calculate the total payload mass delivered by NASA rockets (CRS)
4. Calculate the average payload mass for F9 v1.1
5. Determine the date of the first successful landing at a landing site
6. Identify rockets that successfully landed on a drone spike and had a payload within a given mass range
7. Count the total number of successful and unsuccessful missions
8. Identify the versions of rockets with the highest payload
9. Filter records for 2015 that failed to land
10. Count different landing outcome

GitHub

Build an Interactive Map with Folium

The following elements have been added to the Folium map for interactive visual analysis:

1. Launch sites, displayed as circles.
2. Successful and unsuccessful launches at each site, represented as clusters of markers.
3. Lines showing distances between launch sites and nearby objects, such as railroads, highways, coastlines, and cities.

GitHub

Build a Dashboard with Plotly Dash

Developed an application on Plotly Dash to interactively analyze SpaceX launch data in real-time.

- The control panel includes input elements, such as a dropdown list for selecting a launch site and a payload range slider, which interacts with a pie chart and a scatter plot.
- The primary goal of the interface is to provide user-friendly tools for selecting different payload ranges and identifying visual patterns.
- The payload range slider allows users to visually track the relationship between payload and mission results for selected sites

GitHub

Predictive Analysis (Classification)

In the process of machine learning development, the following steps were taken:

1. Target column addition:
   - A column with class labels was added to the original dataset in order to further train the model.
2. Data normalization:
   - All data was normalized to a common standard in order to ensure that algorithms work correctly.
3. Data partitioning:
   - The data was split into training and testing samples using the `train_test_split` function.
4. Training sample division:
   - The training sample was then split into a validation subsample for model optimization and hyperparameter selection using the `GridSearchCV` function.
5. Hyperparameter optimization:
   - Hyperparameters for models like SVM, decision trees, K-nearest neighbors, and logistic regression were tuned to achieve better results.
6. Model accuracy assessment:
   - Models like SVM and KNN were validated on test data to evaluate their accuracy.

[GitHub]

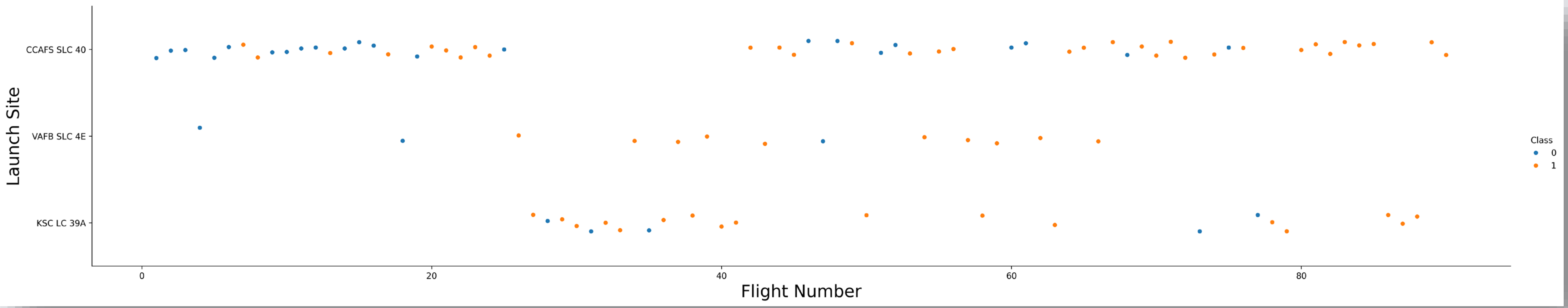
Results

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

Section 2

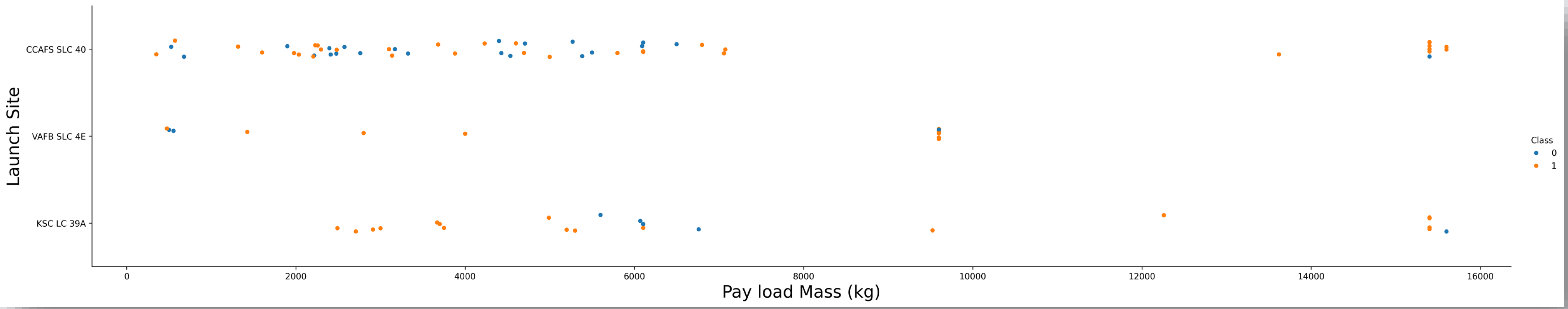
Insights drawn from EDA

Flight Number vs. Launch Site



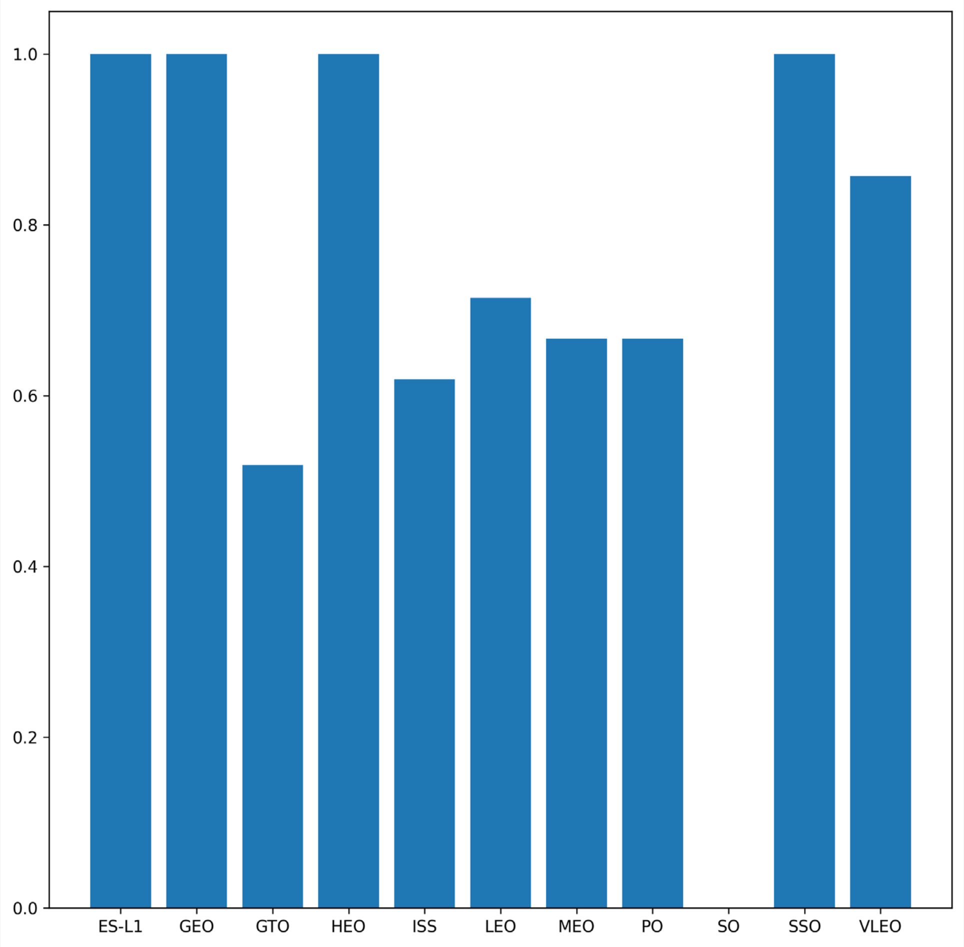
Our analysis shows that launch sites differ in the number of continuous successful launches. In particular, the VAFB SLC 4E and KSC LC 39A sites have a significantly higher percentage of successful launches than other sites.

Payload vs. Launch Site



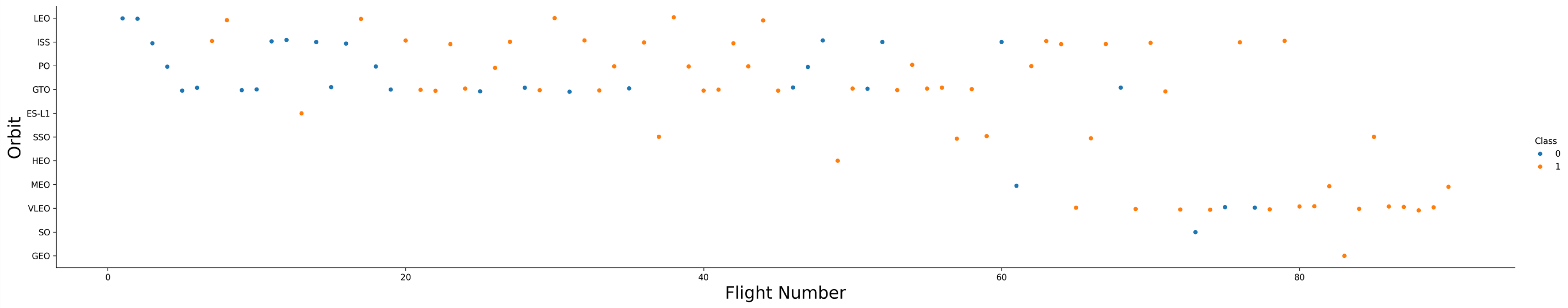
VAFB SLC does not have rockets capable of launching heavy payloads into orbit. In addition, CARS SLC 40 and SC LC 39A launch pads mainly launch rockets with payloads that weigh less than 8,000 kilograms in mass.

Success Rate vs. Orbit Type



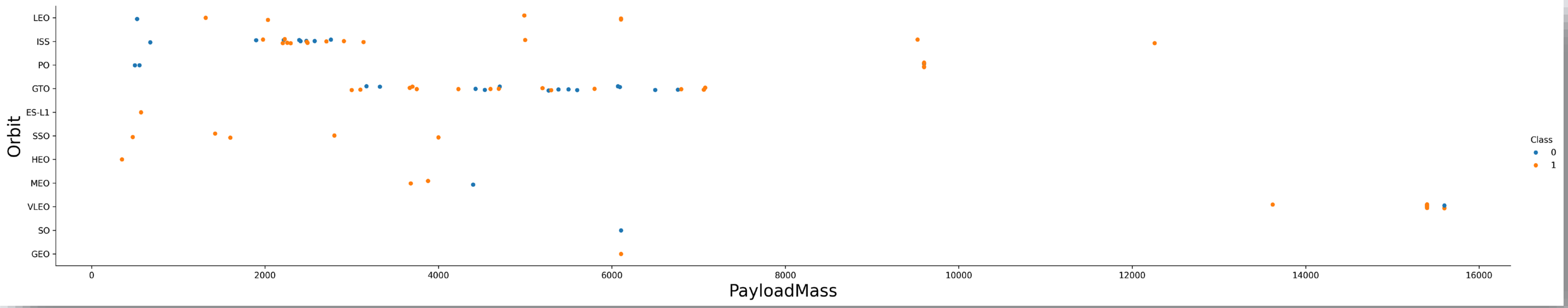
ES-L1, GEO, HEO and SSO orbits have the highest success rate of 100%, while for SO orbit, this indicator is zero.

Flight Number vs. Orbit Type



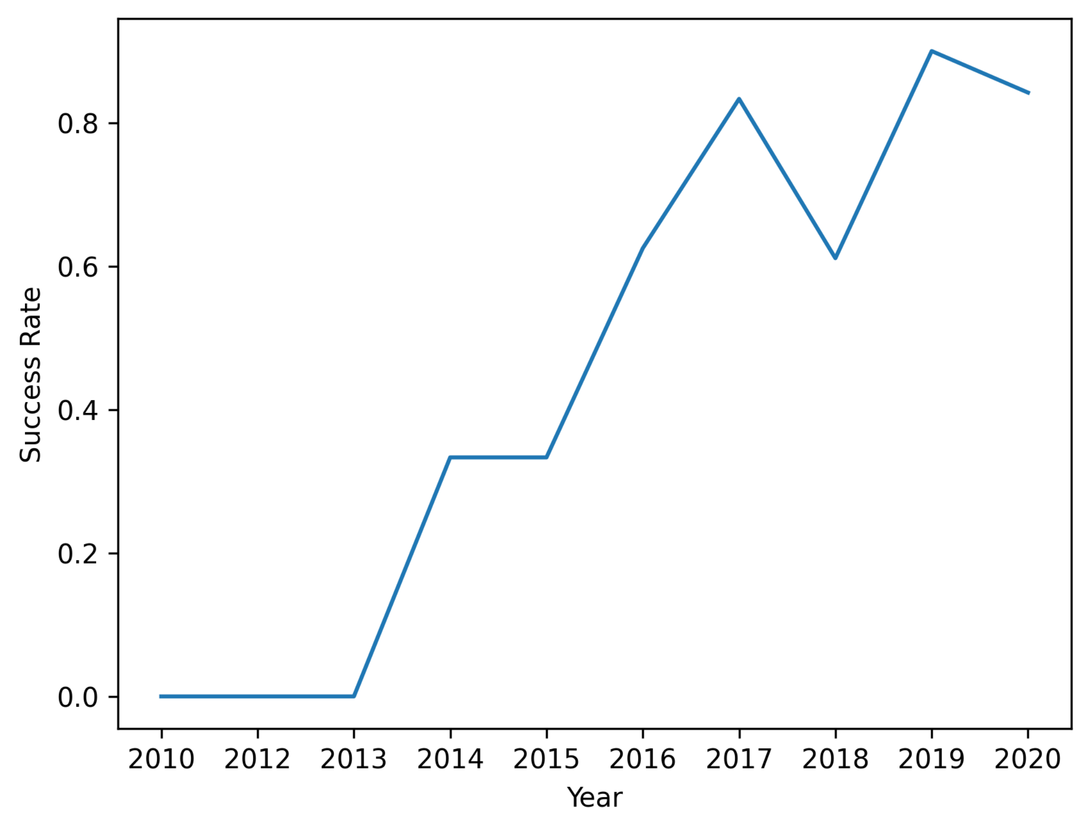
In LEO (Low Earth Orbit), the success rate appears to be directly correlated with the number of flights. However, in GTO (Geosynchronous Transfer Orbit) no such correlation is observed, and the number of flights does not seem to have a significant impact on the success rate.

Payload vs. Orbit Type



For launches with a heavy payload, the success rate of landing is significantly higher for polar, LEO, and IS orbits. However, in the case of the TO orbit, it is not possible to distinguish this dependence as both successful and unsuccessful landings occur here

Launch Success Yearly Trend



The success rate has been increasing continuously since 2013, reaching its highest level in 2020, with the exception of a slight decrease in 2018.

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

The SQL results show four different launch sites:

- CCAFS LC-40,
- VAFB SLC-4E,
- KSC LC-39A,
- CCAFS SLC-40.

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEXTABLE 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

According to the SQL results, there were five launches from launch sites starting with 'CCA'. Of these, four were carried out for NASA and one for SpaceX.

Total Payload Mass

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

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

Total_Payload_Mass
45596

According to SQL, the total mass of payload delivered by NASA's CRS rockets was 45,596 kg.

Average Payload Mass by F9 v1.1

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

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

Average_Payload_Mass
2928.4

According to the SQL results, the average payload mass carried by the F9 v1.1 rocket is 2928.4 kg.

First Successful Ground Landing Date

```
%sql SELECT MIN(Date) AS First_Successful_Landing_Date FROM SPACEXTABLE WHERE Landing_Outcome LIKE '%ground%';
```

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

First_Successful_Landing_Date
2015-12-22

According to SQL, the first successful ground landing took place on December 22, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome LIKE '%drone ship%' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
```

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

Booster_Version
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

SQL results show that four launch vehicles successfully landed on the drone ship with payload masses ranging from 4,000 to 6,00 kg.

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

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT Mission_Outcome, COUNT(*) AS Count FROM SPACEXTABLE 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

According to SQL, there are 100 successful missions and 1 failed mission.

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

The SQL results show that the following launch vehicles had the maximum payload mass

2015 Launch Records

```
%%sql
SELECT substr(Date, 6, 2) AS Month, Landing_Outcome, Booster_Version, Launch_Site
FROM SPACEXTABLE WHERE substr(Date, 1, 4) = '2015' AND Landing_Outcome
LIKE 'Failure (drone ship)' ORDER BY Month;
```

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

Month	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

SQL query results for 2015 show that the following failed drone ship landings occurred:

- January: Rocket with booster version F9 v1.1 B1012, CCAFS launch site LC-40
- April: Rocket with booster version F9 v1.1 B1015, CCAFS LC-40 launch pad

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

```
%%sql
SELECT Landing_Outcome, COUNT(*) 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.
```

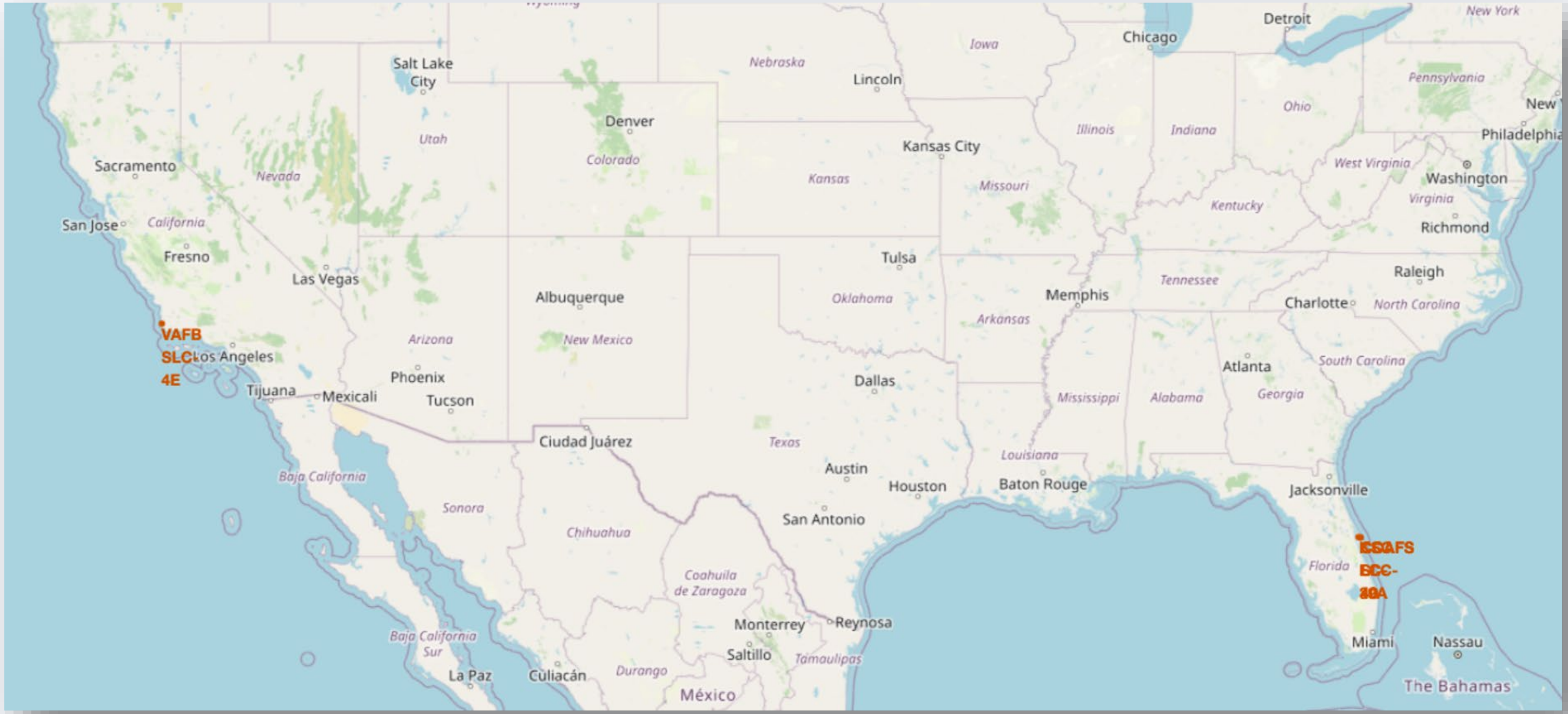
Landing_Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

The results of an SQL query showing the ranking of the number of successful and unsuccessful landings in the period from June 4, 2010 to March 20, 2017

Section 3

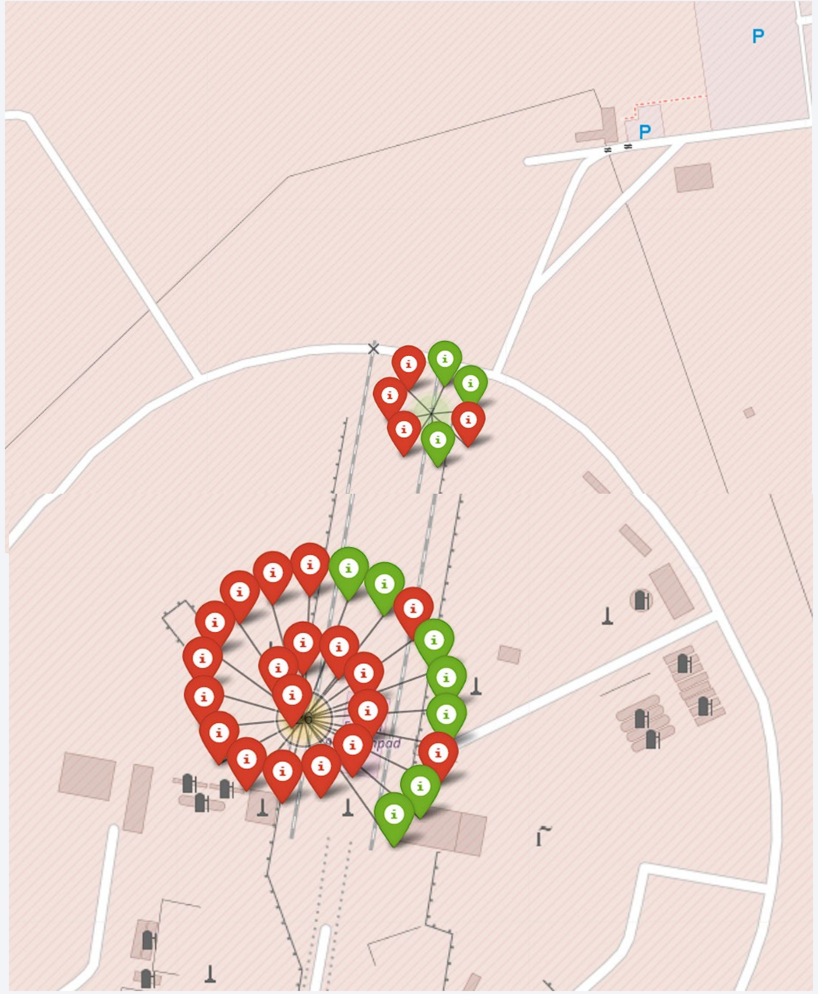
Launch Sites Proximities Analysis

Launch Sites

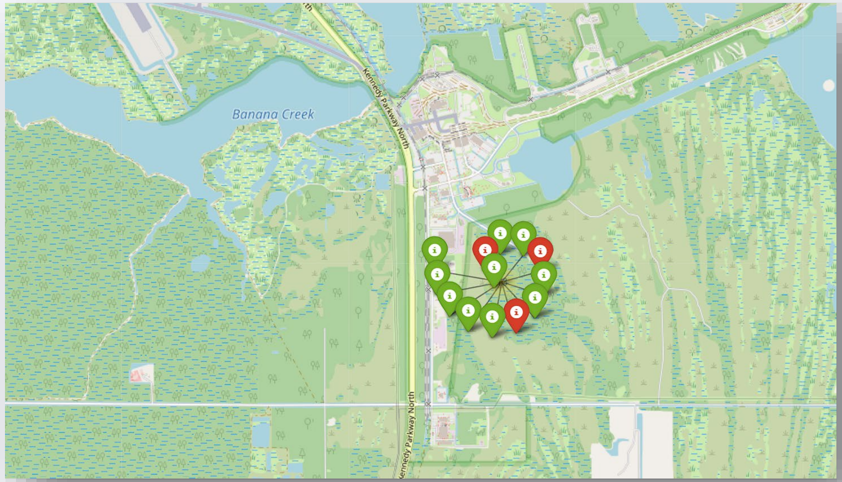


All launch sites are marked on the map.

Successful/Failed Launches



All successful and failed launches are shown on the map.



The distance to the nearest tags from CCAFS LC-40.

The nearest airport
is approximately 51.54 kilomet
ers from the launch site.

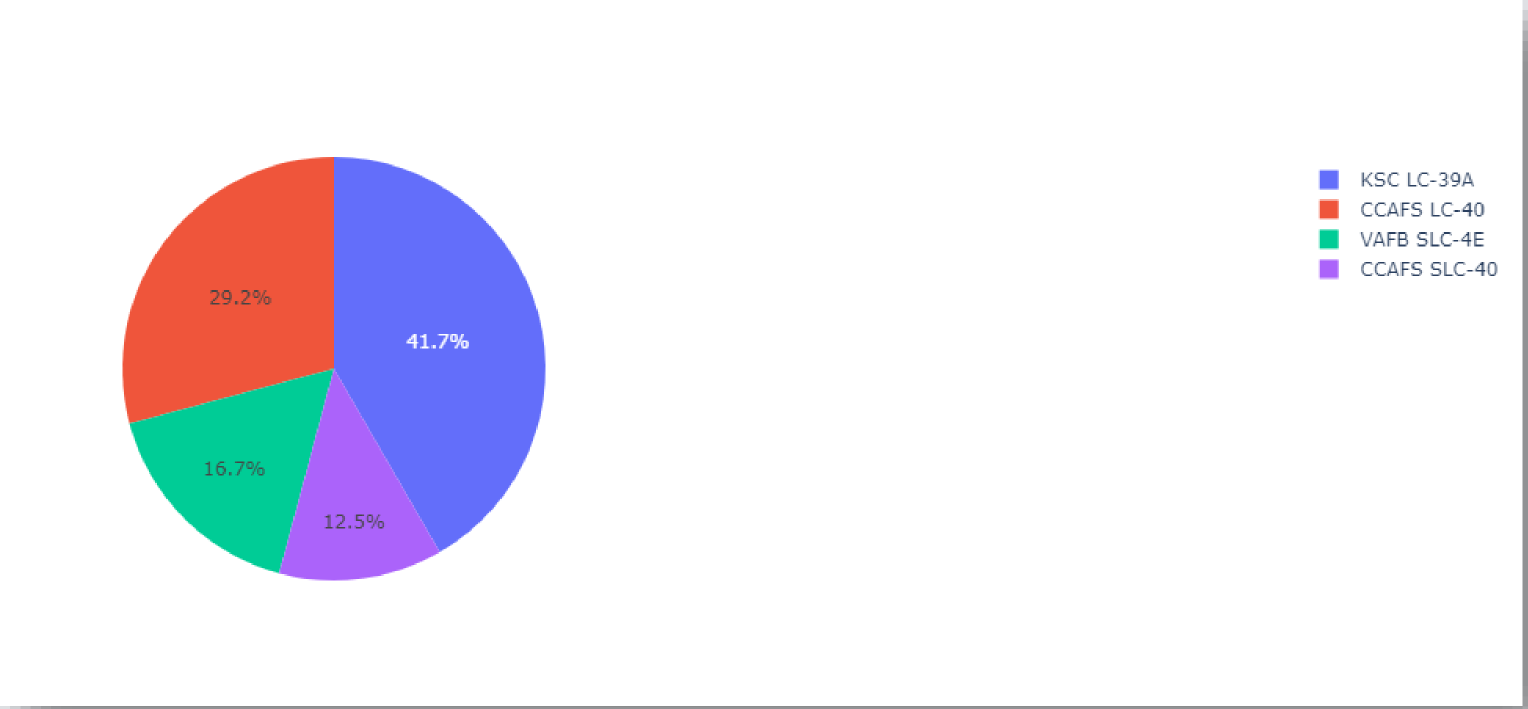
Nearest points:

- Railroad - 1.27 km from the launch point
- Road - 0.65 km from the launch point
- Shoreline - 0.91 km from the launch point

Section 4

Build a Dashboard with Plotly Dash

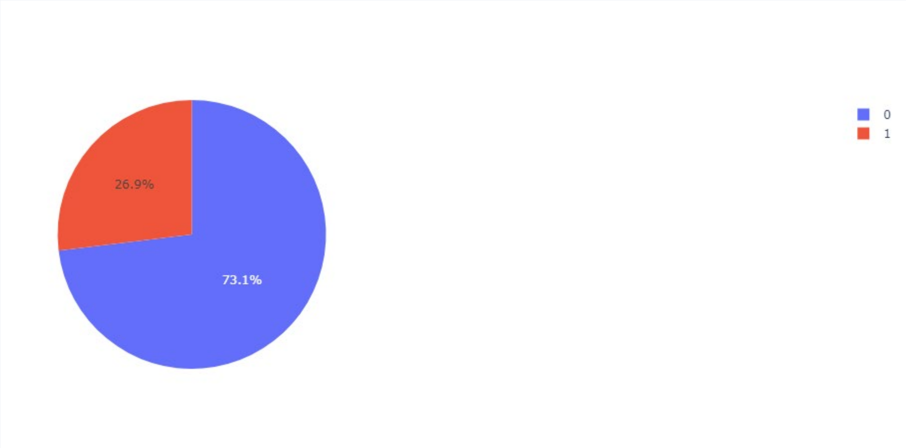
All launch successes at the sites



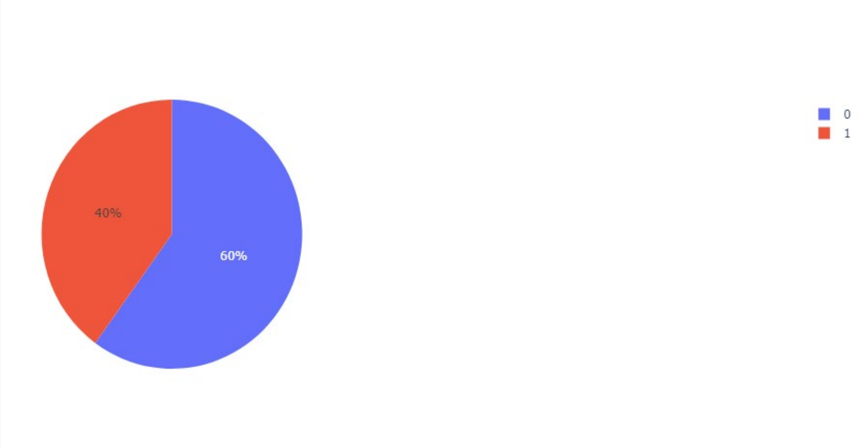
The percentage of successful launches on different sites. The highest percentage is 41.7%, registered at KSC LC-39A

Percentages of successful launches by site

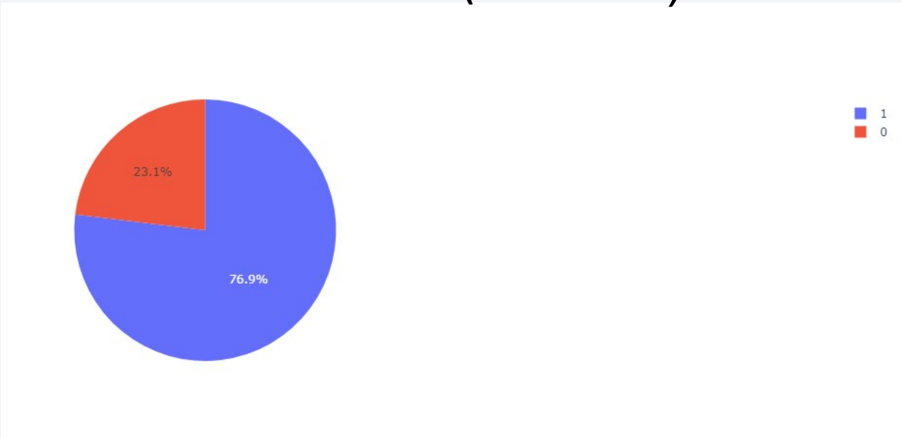
CCAFS LC-40



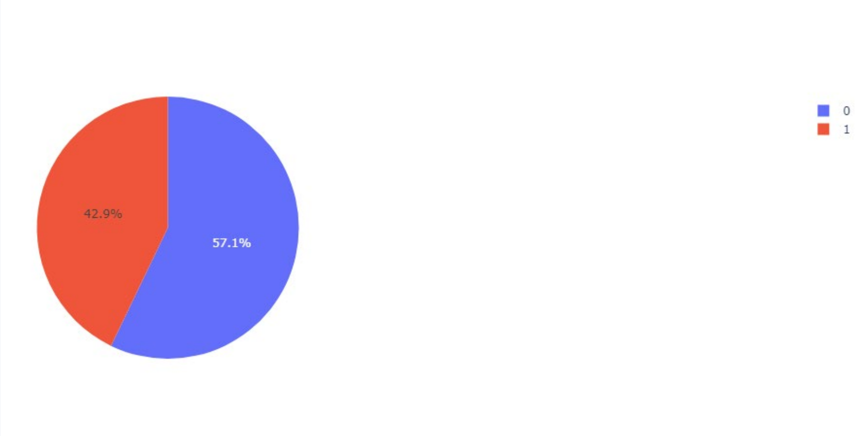
VAFB SLC-4E



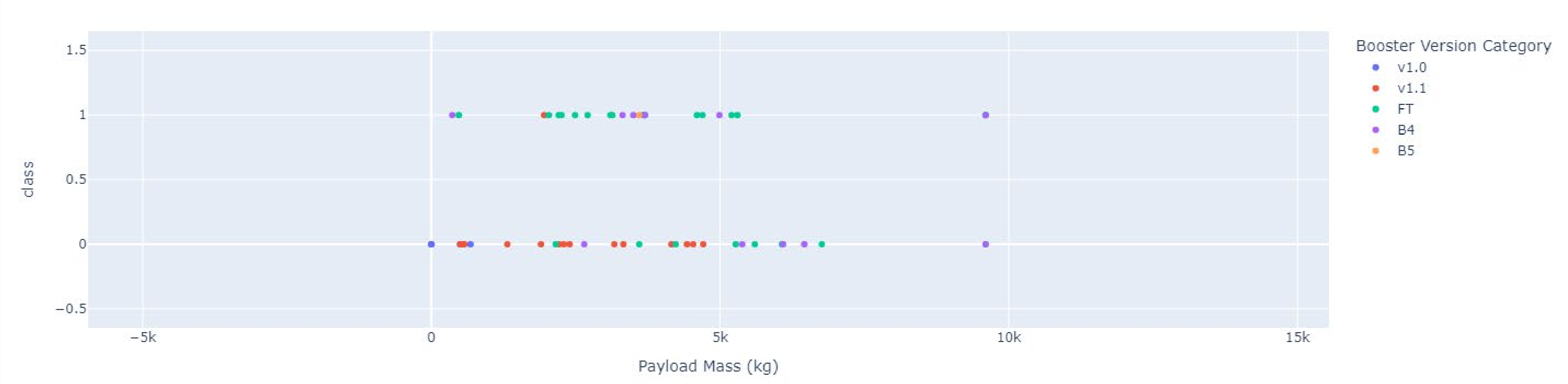
KSC LC-39A (the best)



CCAFS SLC-40 (the bad)



PayloadMass success count for all launch sites.

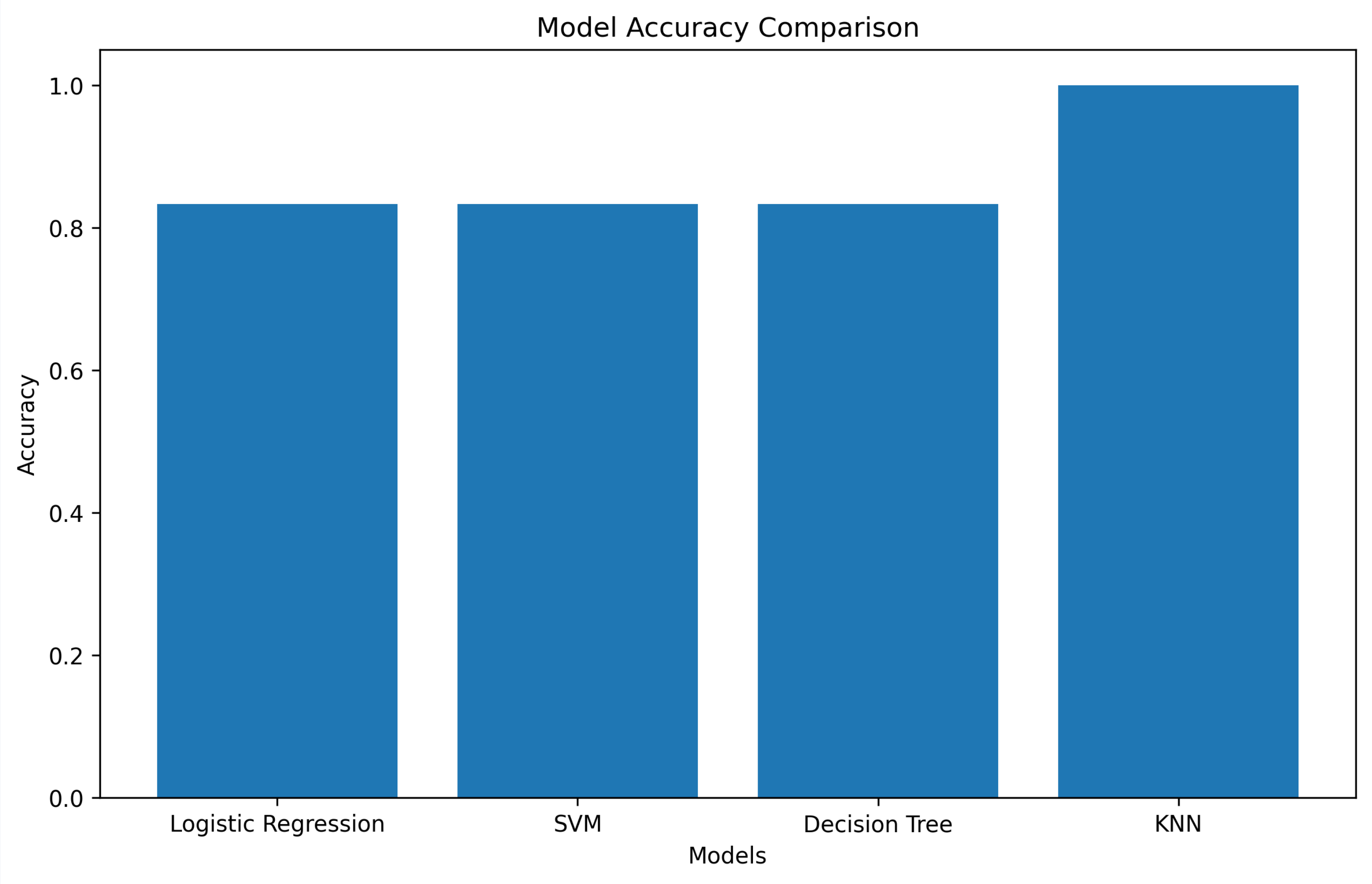


The graph shows the success rate of rocket launches (0 - failure, 1 - success) for different carrier versions (v1.0, v1.1, FT, B4, B5), depending on the payload mass (from 0 to 15,000kg). Most successful launches occur in the 0-5,000 kg range, with the FT version having the most successes.

Section 5

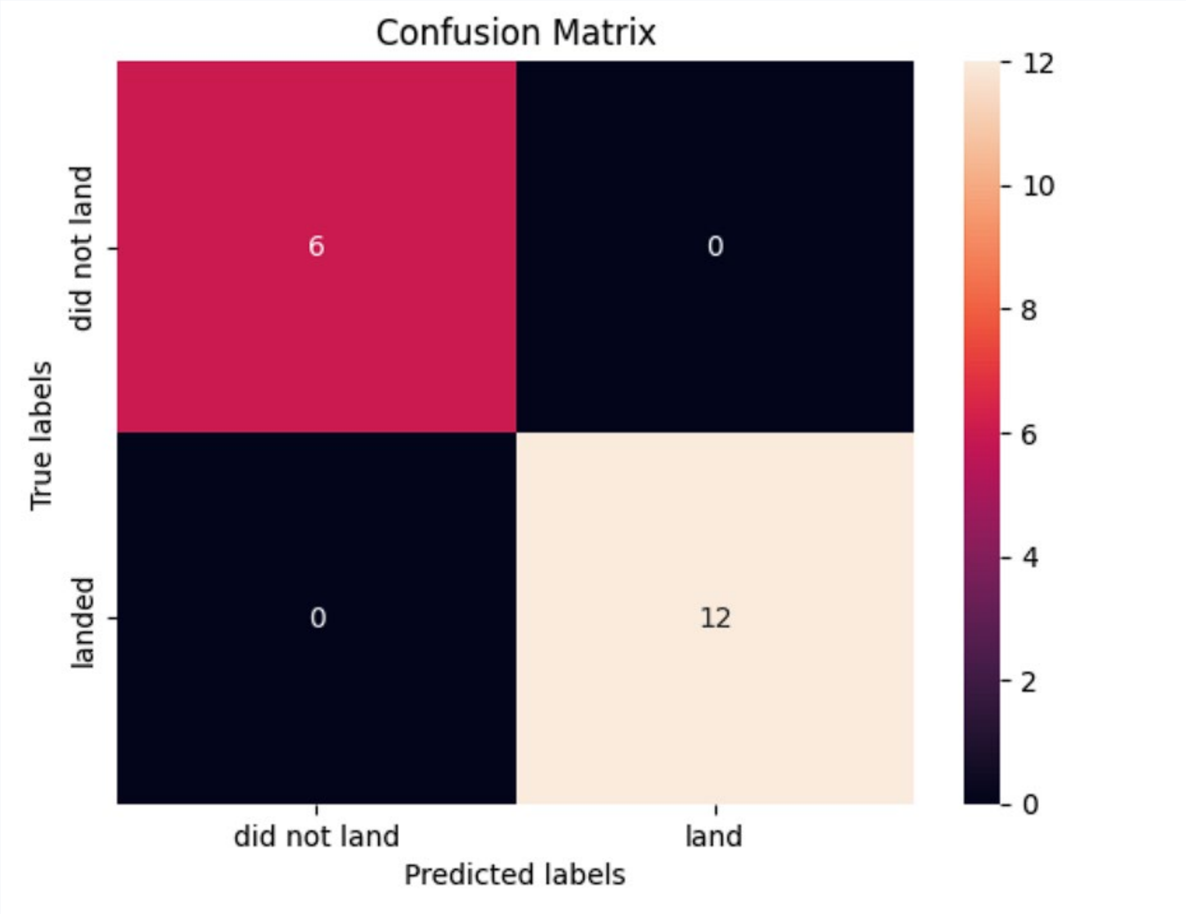
Predictive Analysis (Classification)

Classification Accuracy



It can be seen from the graph that KNN has the highest accuracy, while decision tree, SVM, and logistic regression have about the same accuracy, ranked next to KNN.

Confusion Matrix



The k-NN model is ideal for the following reasons:

- 100% accuracy:
  - The model does not allow for classification errors.
- Correct determination of outcomes:
  - It correctly classifies both successful and unsuccessful landings.
- Reliability:
  - Suitable for decision making in critical situations such as space launches.
- Best Performance:
  - Outperforms other models on all key metrics.

Conclusions

- **Based on the comprehensive analysis of SpaceX Falcon 9 launch data, the following key conclusions can be drawn:**

1. **Launch site and success rate:**
   KSC LC-39A demonstrates a significantly higher success rate compared to other launch sites, indicating potential advantages in its infrastructure and location.
2. **Payload mass and orbit type:**
   Successful landing outcomes are positively correlated with heavier payloads and specific orbit types, such as LEO and Polar, suggesting limitations for certain mission profiles.
3. **Flight number and success rate:**
   Although a correlation between flight number and success rate is observed for LEO, it is inconsistent across other orbit types, implying the influence of factors beyond simple experience.
4. **Predictive model accuracy:**
   The K-Nearest Neighbors (KNN) model achieves 100% accuracy in predicting Falcon 9 first stage landing success, demonstrating its reliability for critical decision-making in space missions.

Appendix

- **All PNG files:** *click here*
- **Module 1:** *click here*
- **Module 2:** *click here*
- **Module 3:** *click here*
- **Module 4:** *click here*
- **Full project:** *click here*
- **My GitHub:** *click here*

Thank you!