WINNING SPACE RACE WITH DATA SIENCE

Michał Pacocha

09.09.2024

Outline

Executive Summary

Summary of methodologies:

The scope of the project focuses on identifying key factors that contribute to the success of rocket landings. The following methodology was applied to achieve the project's objectives:

- **Data Collection:** Gathered data through the SpaceX API and web scraping techniques.
- **Data Wrangling:** Processed and structured the data to create categorical variables for success/failure of landings.
- **Data Exploration:** Utilized SQL and Python (Pandas, Matplotlib and Seaborn) for data exploration and visualization.
- **Data Analysis:** Performed geospatial analysis using Folium and created interactive dashboards with Plotly Dash.
- **Model Building:** Developed predictive models for landing outcomes using Logistic Regression, Support Vector Machine, Decision Trees, and K-Nearest Neighbors.

Basic insight:

- The first successful landing took place in December 2015.
- Out of 101 missions, only one ended in failure (overall mission outcomes, not landing outcomes).
- Orbits ES-L1, GEO, HEO, and SSO have a 100% landing success rate.
- Since 2013, the landing success rate has been steadily increasing.
- The best performing machine learning model is the Linear Regression model, although all models exhibit similar behavior.

Introduction

Background:

SpaceX is an American aerospace company founded by Elon Musk in 2002. Its mission is to reduce space transportation costs and enable the colonization of Mars. SpaceX developed the Falcon rockets and Dragon spacecraft, achieving milestones like the first privately funded spacecraft to reach orbit and the first reusable rocket.

Reason:

The project aims to answer the question of what factors are related to the successful landing of the first stage of SpaceX's Falcon 9 rockets. By gaining this knowledge, a competing company can gain valuable insights and potentially improve their approach, increasing efficiency and success rates in space missions.

METHODOLOGY

Data Collection - API

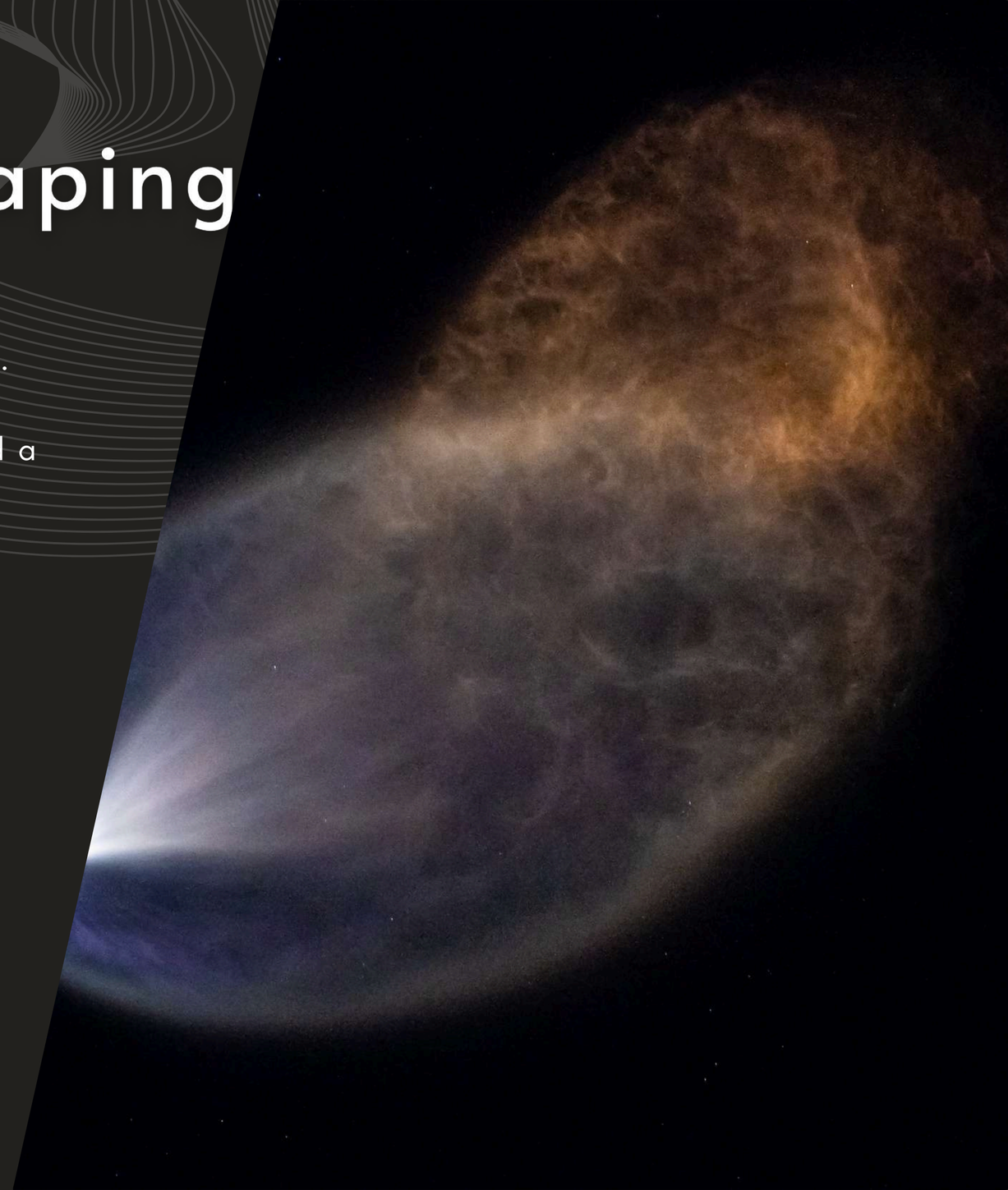
- **Requested rocket launch data** from the SpaceX API using requests.get().
- **Decoded the response content** as JSON using .json().
- **Created a DataFrame** from the decoded JSON using json_normalize().
- **Requested additional data** about the launches from the SpaceX API.
- **Created a dictionary** from the requested data.
- **Converted the dictionary** into a DataFrame using pd.DataFrame().
- **Filtered the data** to include only Falcon 9 launches.

Jupyter Notebook

Data Collection - scraping

- **Requested the Falcon 9 Launch Wiki page f**rom its URL using requests.get().
- **Created a BeautifulSoup object** from the HTML response.
- **Found all tables** on the wiki page using .find_all().
- **Extracted column names** one by one using .find_all() and a custom extract_column_from_header() function.
- **Parsed** the launch HTML tables.
- **Created a dictionary** from the data.
- **Converted the dictionary** into a DataFrame.

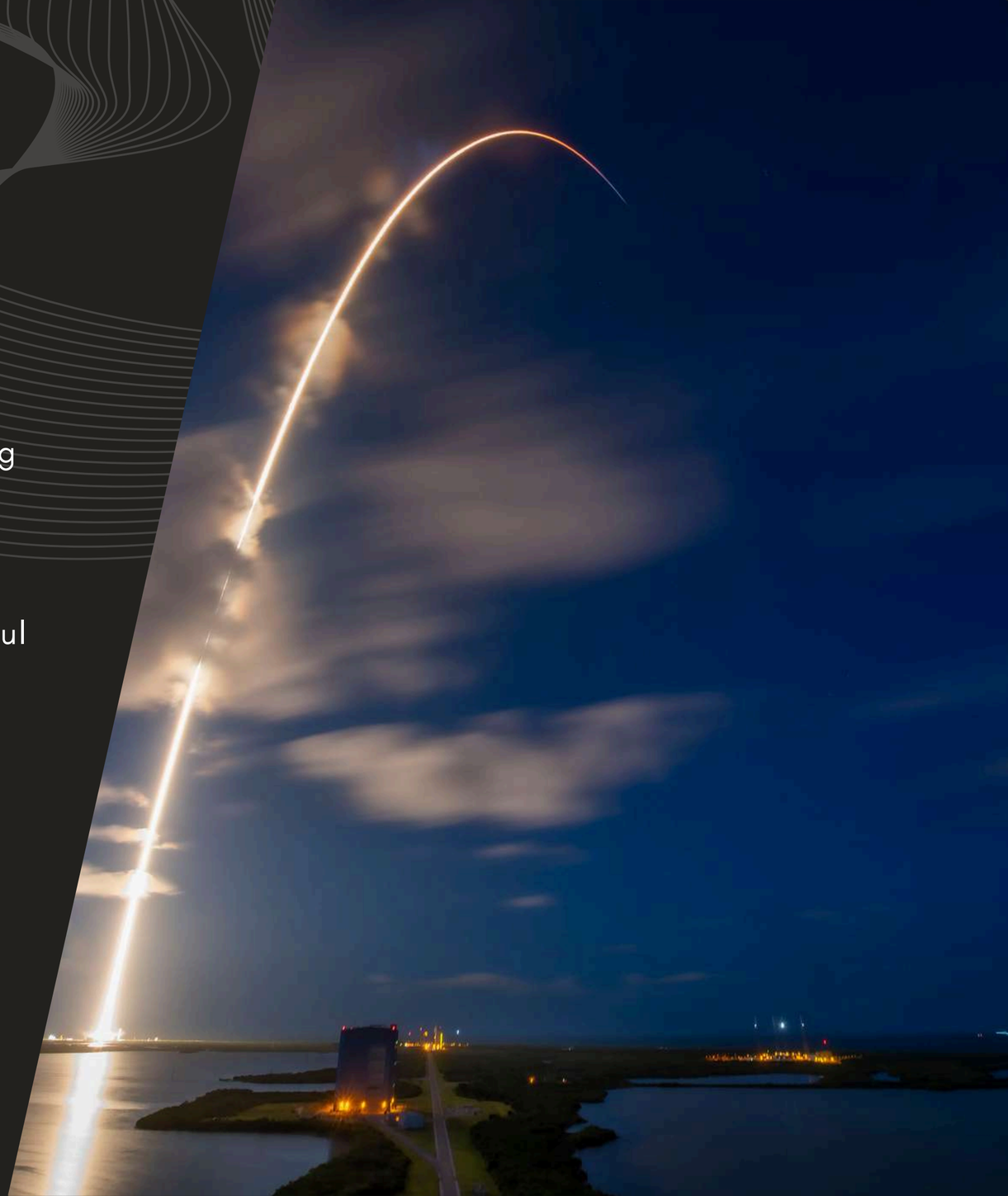
Jupyter Notebook



Data Wrangling

- **Dealt with missing values** for Payload Mass by replacing them with the mean.
- **Calculated** the number of launches at each site using value_counts().
- **Calculated** the number and occurrence of each orbit using value_counts().
- **Calculated** the number and occurrence of each mission outcome for the orbits using value_counts().
- **Created a set** of outcomes where landing was unsuccessful (e.g., None, None; False ASDS).
- **Created a binary** landing outcome column (0 = Failed, 1 = Successful).

Jupyter Notebook



EDA with Data Visualization

- Scatter plot to visualize the relationship between Flight Number, Launch Site and landing outcomes.
- Scatter plot to visualize the relationship between Payload Mass, Launch Site and landing outcomes.
- Bar plot to visualize the relationship between success rate of each orbit type.
- Scatter plot to visualize the relationship between FlightNumber, Orbit type and landing outcome.
- Scatter plot to visualize the relationship between Payload, Orbit type and landing outcomes.
- Line plot to visualize the landing success yearly trend.

Jupyter Notebook

EDA with SQL

- Query displaing unique launch sites in the space mission
- Query displaing 5 records where launch sites begin with the string 'CCA'
- Query displaing the total payload mass carried by boosters launched by NASA (CRS)
- Query displaing average payload mass carried by booster version F9 v1.1
- Query displaing the date when the first succesful landing outcome in ground pad was acheived
- Query displaing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Query displaing the total number of successful and failure mission outcomes
- Query displaing the names of the booster versions which have carried the maximum payload mass.
- Query displaing the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- Query displaing the ranking of landing outcomes in 2015

Jupyter Notebook

Data Analysis - Interactive Map with Folium

- Added a circle object to visualize each launch site.
- Added a marker object to label each launch site.
- Added a marker object in a marker cluster to visualize successful and failed landings for each launch site.
- Added circle objects to represent launch site proximities.
- Added marker objects to display the distance from each launch site to its proximities.
- Added line objects to visualize the distance from each launch site to its proximities.

Jupyter Notebook

Data Analysis - Interactive dashboard with Ploty Dash

- Added a dropdown menu to the select launch site.
- Added a pie chart to visualize propotion of successful and failed landing for all launch sites and each individual site.
- Added a slider to select the payload range.
- Added a strip plot to visualize successful and failed landing by launch site and payload range.

Code

Model building - Predictive Analysis (Classification)

- **Selected features** for the model.
- **Created dummy variables** for categorical columns using pd.get_dummies().
- **Cast all numeric columns** to float64.
- **Created a NumPy array** from the binary column for successful/failed landing outcomes.
- **Standardized features** using StandardScaler().
- **Split data into training and testing sets** using train_test_split().
- **Created** Logistic Regression, Support Vector Machine, Decision Tree, and K-Nearest Neighbors objects.
- **Found the best parameters** for each model using GridSearchCV().
- **Fitted** each model.
- **Created a confusion matrix** for each model.
- **Evaluated each model** using Accuracy Score, Jaccard Score, F1 Score, and Recall Score.
- **Further evaluated models** by calculating Mean Cross-Validation using cross_val_score().
- **Picked** the best-performing model.

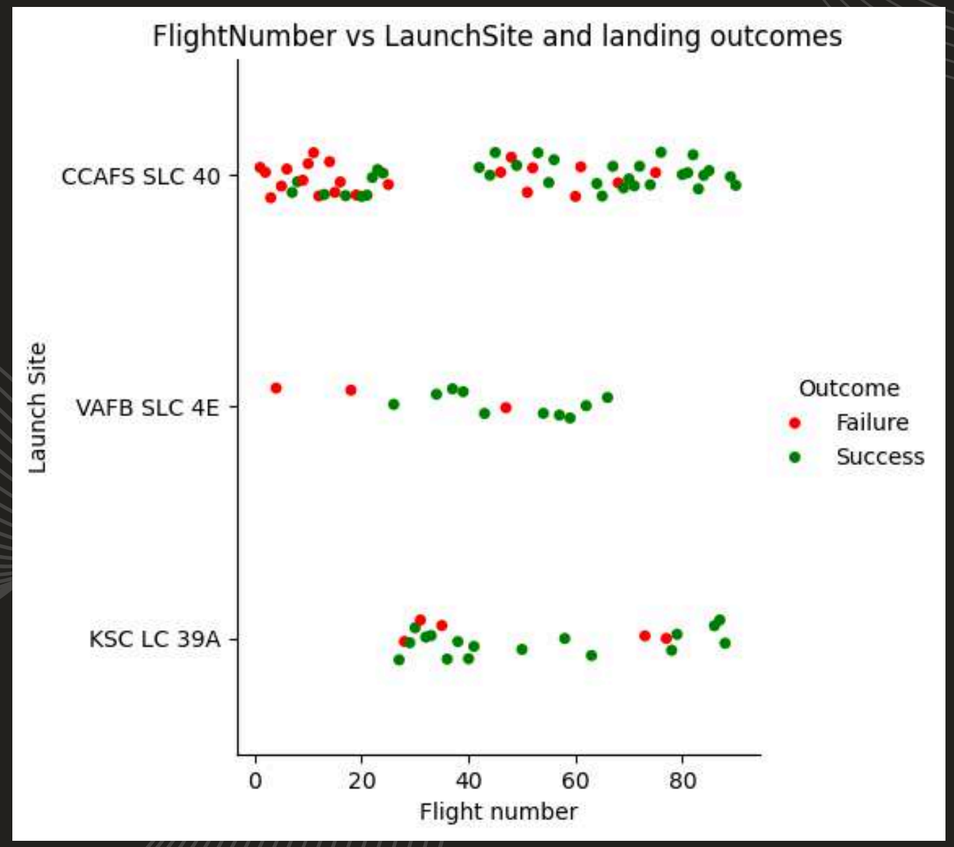
Jupyter Notebook


RESULTS

INSIGHTS DRAWN BY EDA

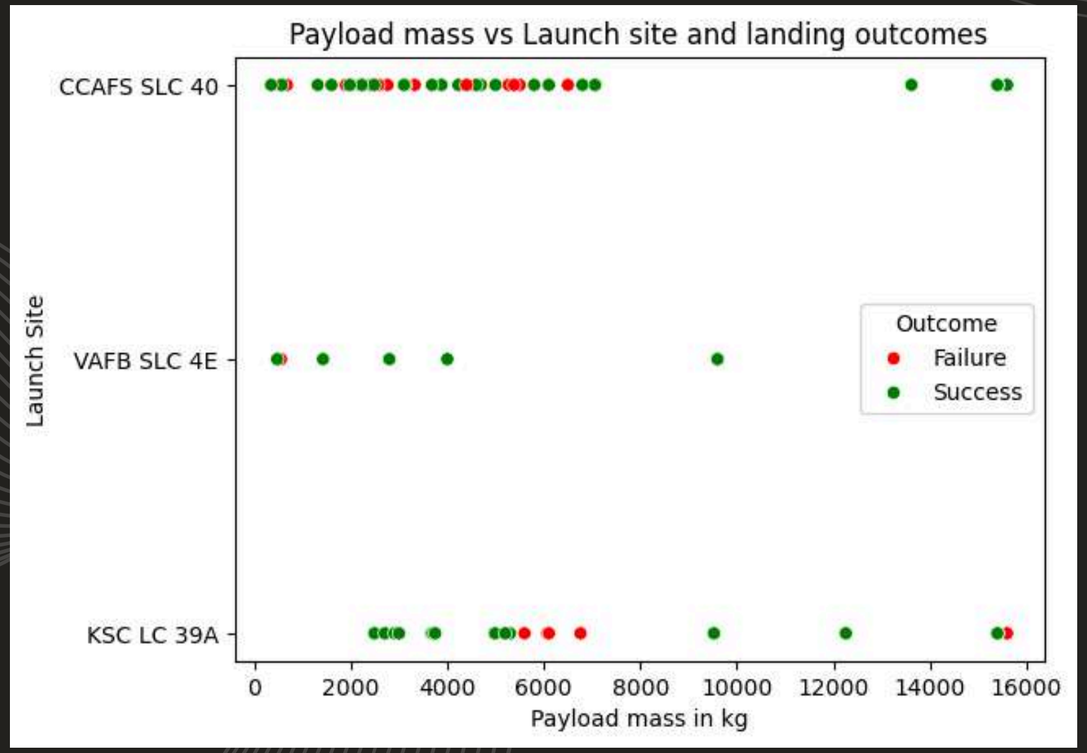
Flight Number vs Launch Site

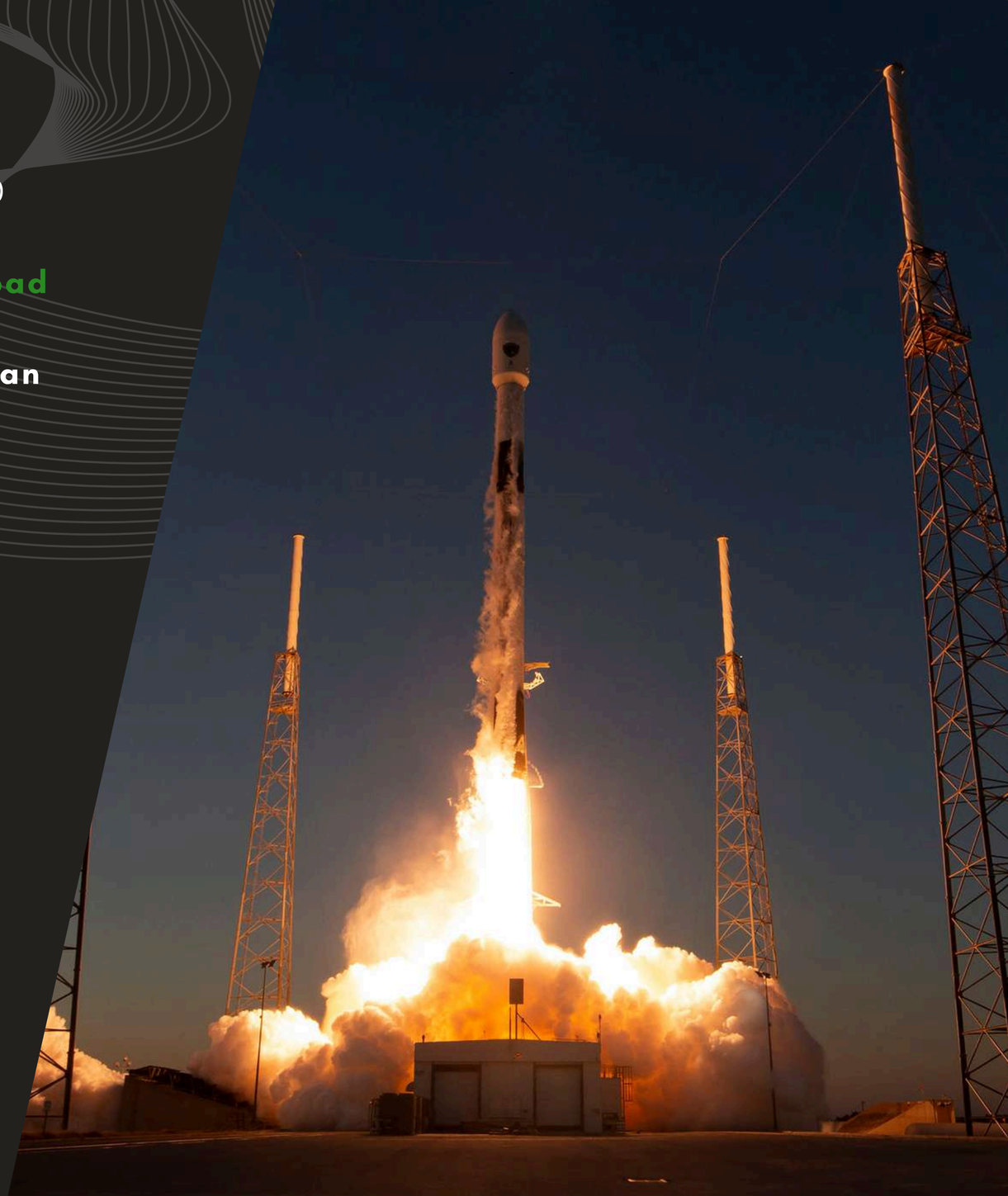
- **CCAFS SLC 40** was the **first used launch site.**
- **The first launches** from **CCAFS SLC 40** were mostly **unsuccessful.**
- **VSFB SLC 4E** appears to be **currently not in use.**
- **Most launches** have taken place from **CCAFS SLC 40.**
- **Recent launches** have a **significantly higher success rate.**



Payload vs Launch Site

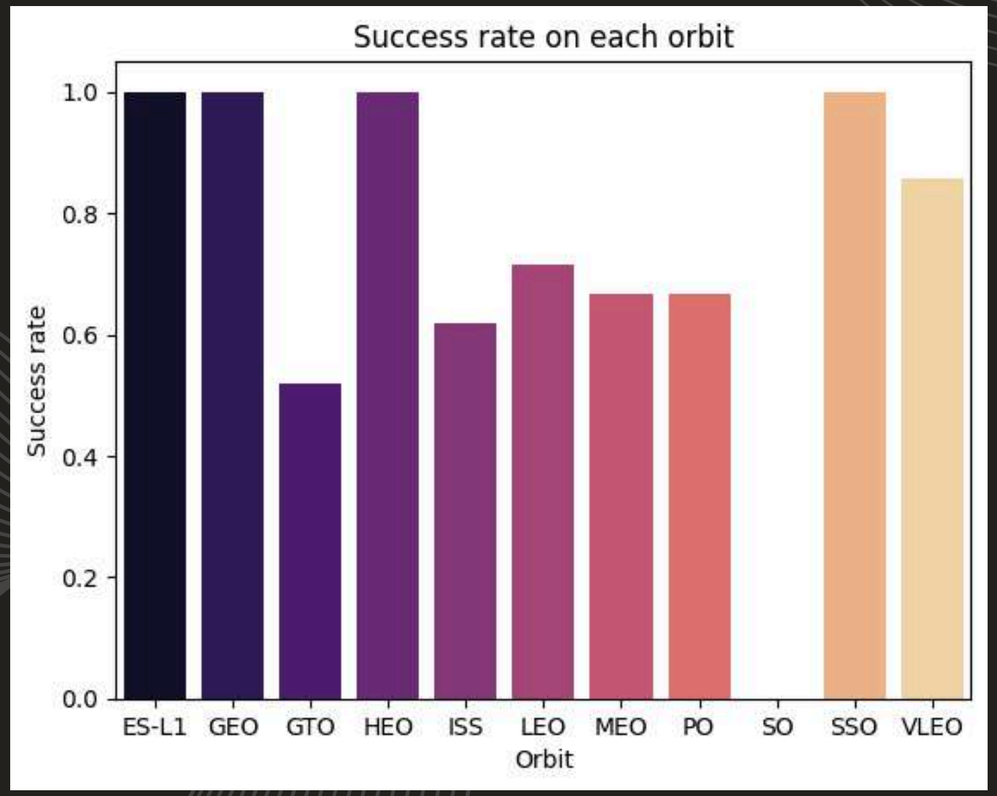
- There have been **no heavy payload launches** (over 10 000 kg) from **VAFB SLC 4E**.
- **CCAFS SLC 40** has a **100% success rate for heavy payload launches** (over 10 000 kg).
- **KSC LC 39A** only hosts launches with payloads **greater than 2 000 kg**.
- **Most payload mass range from 0 kg to 8 000 kg**.





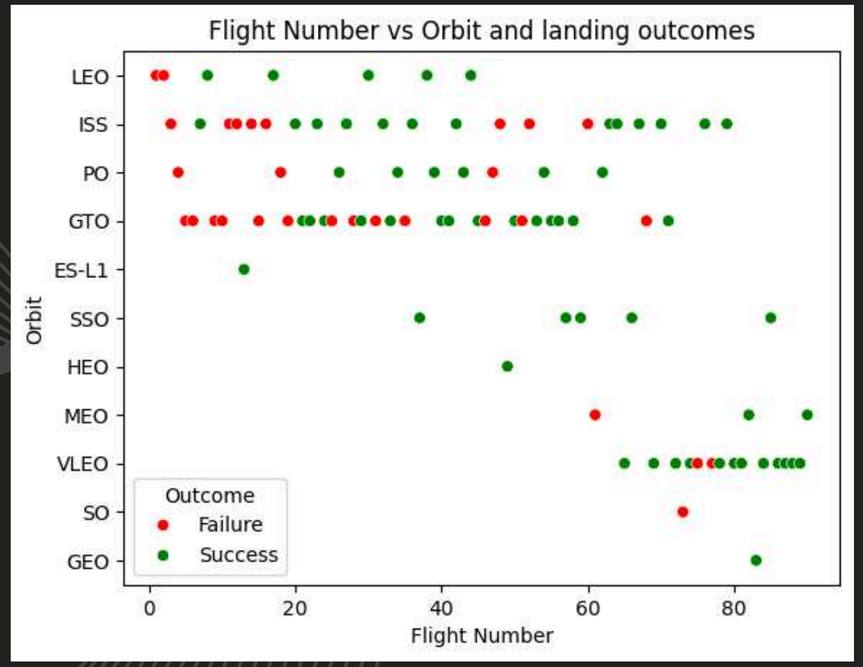
Success Rate vs Orbit Type

- **Orbits ES-L1, GEO, MEO, and SSO** have a **100% success rate**.
- **Orbits GTO, ISS, LEO, MEO, PO, and VLEO** have a **success rate ranging from 50% to 85%**.
- **Orbit SO** has a **0% success rate**.



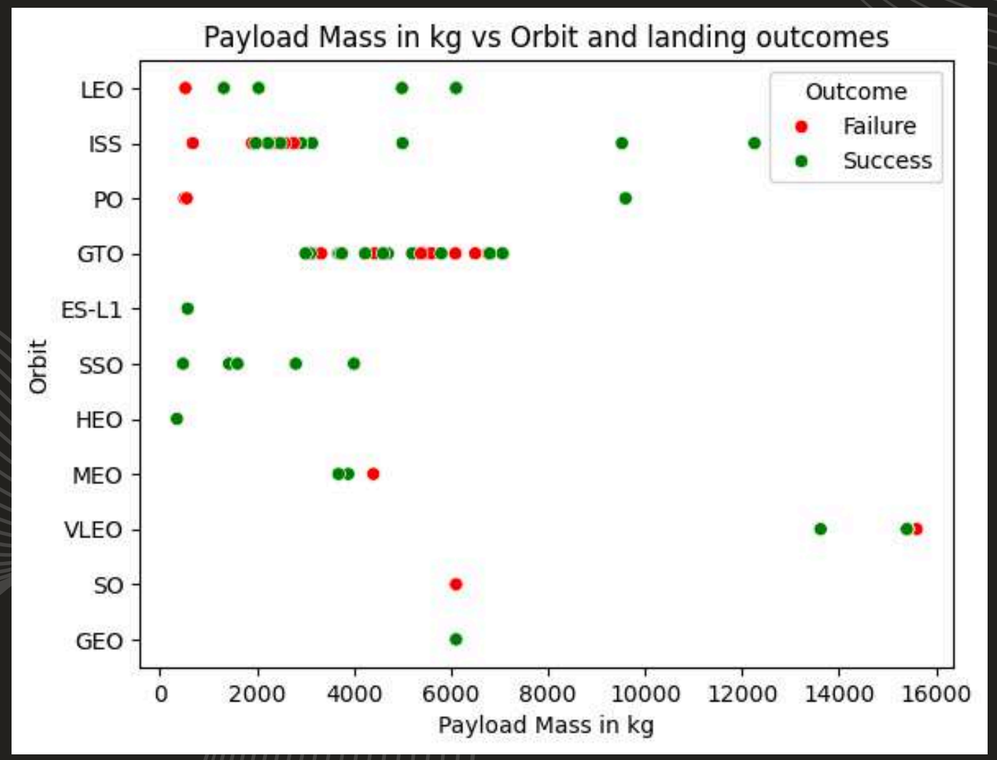
Flight Number vs Orbit Type

- **The first launches** were to **orbits LEO, ISS, PO, and GTO,** which are also among **the most frequently used.**
- **Orbits like ES-L1, HEO, GEO, and SO** have only had **one launch each.**
- **Recently, the most frequently used orbits are SSO, MEO, and VLEO.**
- Excluding orbits with only one launch, **GTO has the worst success rate.**
- Excluding orbits with only one launch, **SSO and VLEO have the highest success rates.**



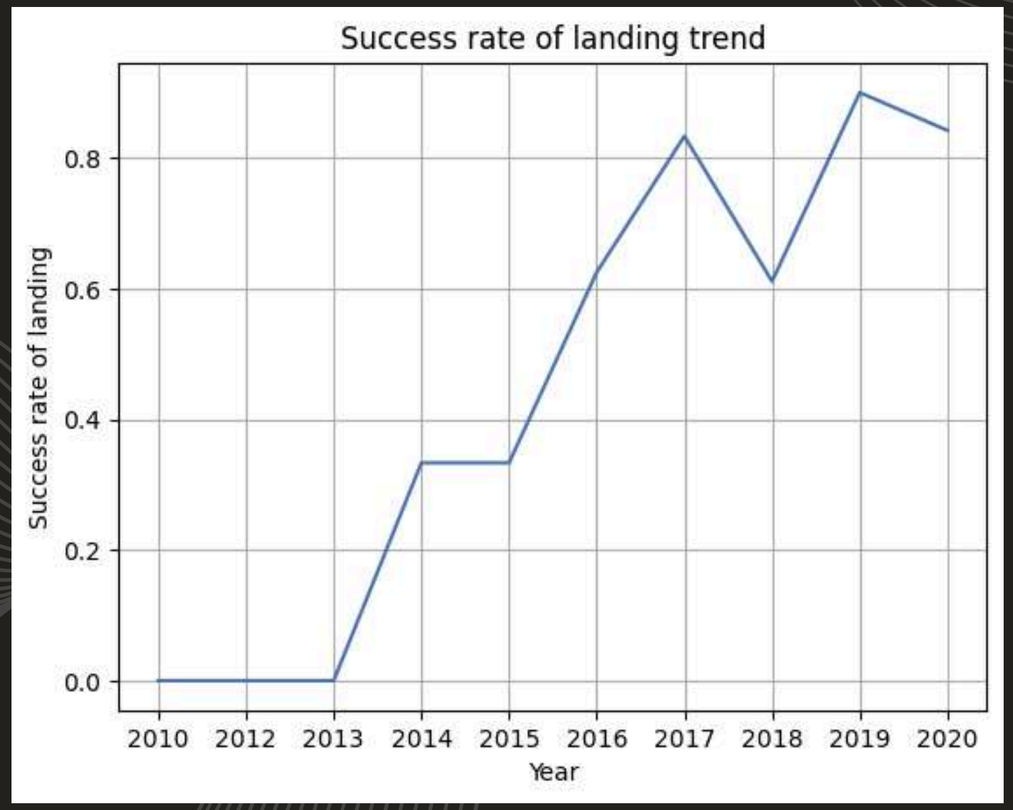
Payload vs Orbit Type

- **Heavy payloads** are used in **ISS, PO, and VLEO orbits**.
- **SO** is used for **lower payloads** and has a **100% success rate**.
- **GTO payloads** range from **3,000 kg to 8,000 kg** and **outcome is mixed.**



Launch Success Yearly Trend

- **Since 2013**, the success rate **has been steadily growing**.
- **In 2018**, the **success rate dropped** from around 80% in 2017 to around 60%.
- **In 2019**, the **success rate reached its peak**.



Launch Sites Info

All launch site names:

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

- **CCAFS LC-40** - located at Cape Canaveral, Florida.
- **VAFB SLC-4E** - located at Vandenberg Space Force Base in California.
- **KSC LC-39A** - located at Kennedy Space Center in Florida.
- **CCAFS SLC-40** - located at Cape Canaveral, Florida.

Launch Sites Names Begin 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

Payload Mass Info

The total payload mass carried by boosters launched by NASA (CRS):

- 45 596 kg

Total_payload_mass_kg for NASA (CRS)
45596

Average payload mass carried by booster version F9 v1.1:

- 2 928.4 kg

Avg mass carried by bosster ver. F9 v1.1
2928.4

Missions Info

Date when the first succesful landing outcome in ground pad was acheived:

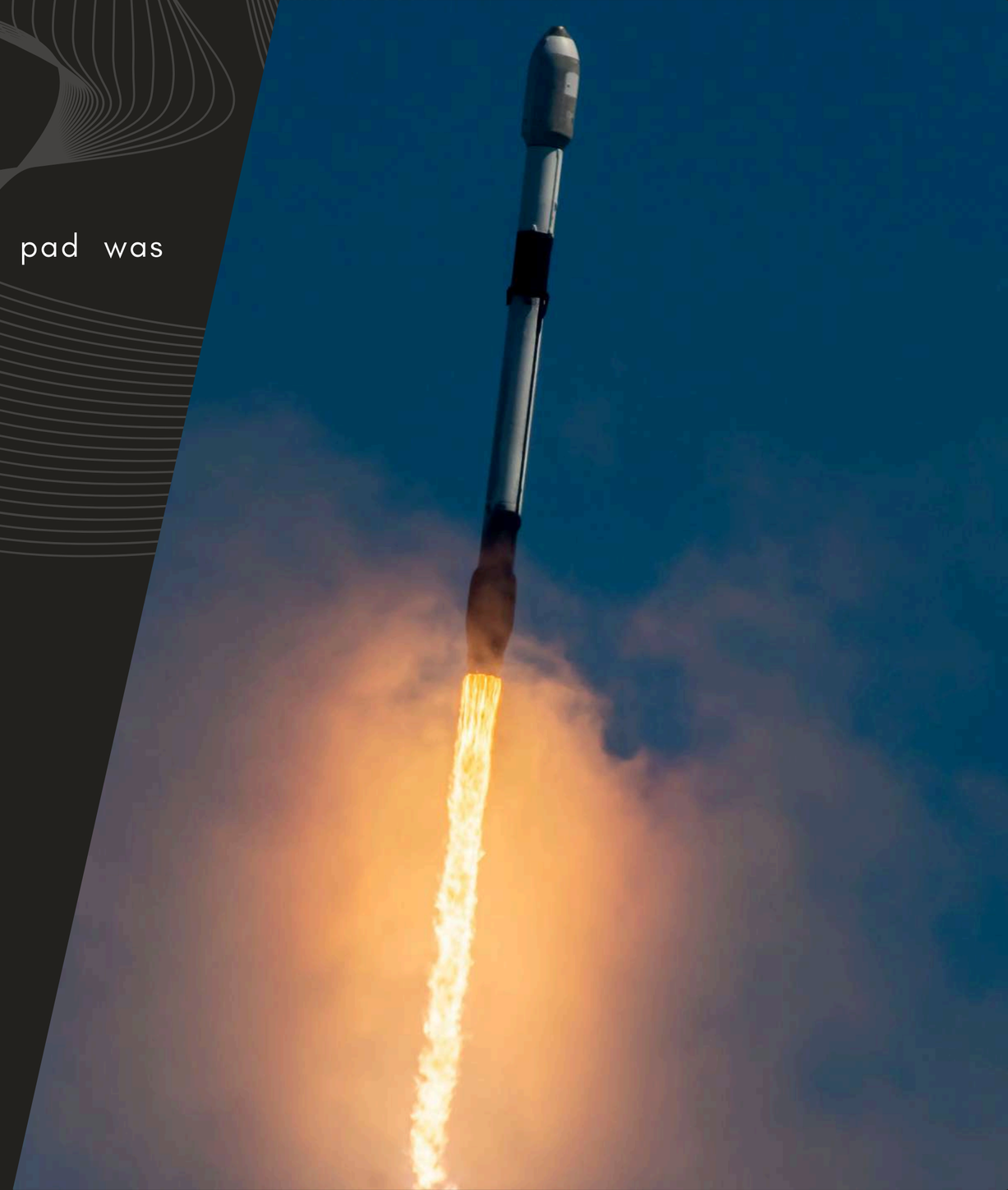
- December twenty-second, twenty fifteen

MIN(Date)
2015-12-22

The total number of successful and failure mission outcomes:

- **100 Successful misions** and **1 failed mission**

Successful missions outcomes	Failed missions
100	1



Boosters Info

Boosters which have success in drone ship and have payload mass greater than 4 000 but less than 6 000:

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

FT Boosters seem to have **the best** outcomes for nedium payload mass

Boosters which have carried the maximum payload mass:

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

Only **B5 Booster** carry the **maxiumum payload mass**

Landing Outcomes Info

Failures of landing attempts on the drone ship in 2015

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

In 2015 there was **2 fialed drone ship landing outcomes**. Both with similar bosster version **B1012** and **B1015**. Also both landings were performed at the same launch site **CCAFS LC-40**

Ranked landing outcomes between June 4, 2010, and March 20, 2017:

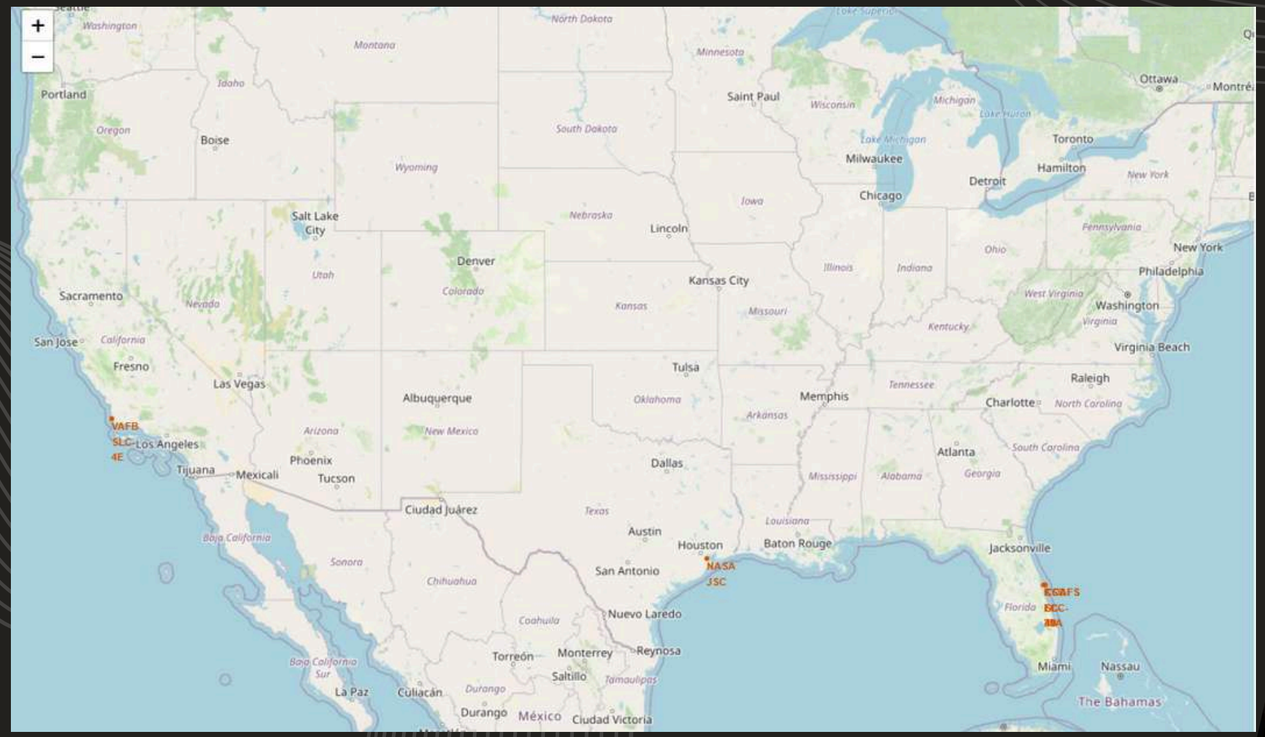
Number of landing outcomes between 2010-06-04 and 2017-03-20	Landing_Outcome
10	No attempt
5	Success (drone ship)
5	Failure (drone ship)
3	Success (ground pad)
3	Controlled (ocean)
2	Uncontrolled (ocean)
2	Failure (parachute)
1	Precluded (drone ship)

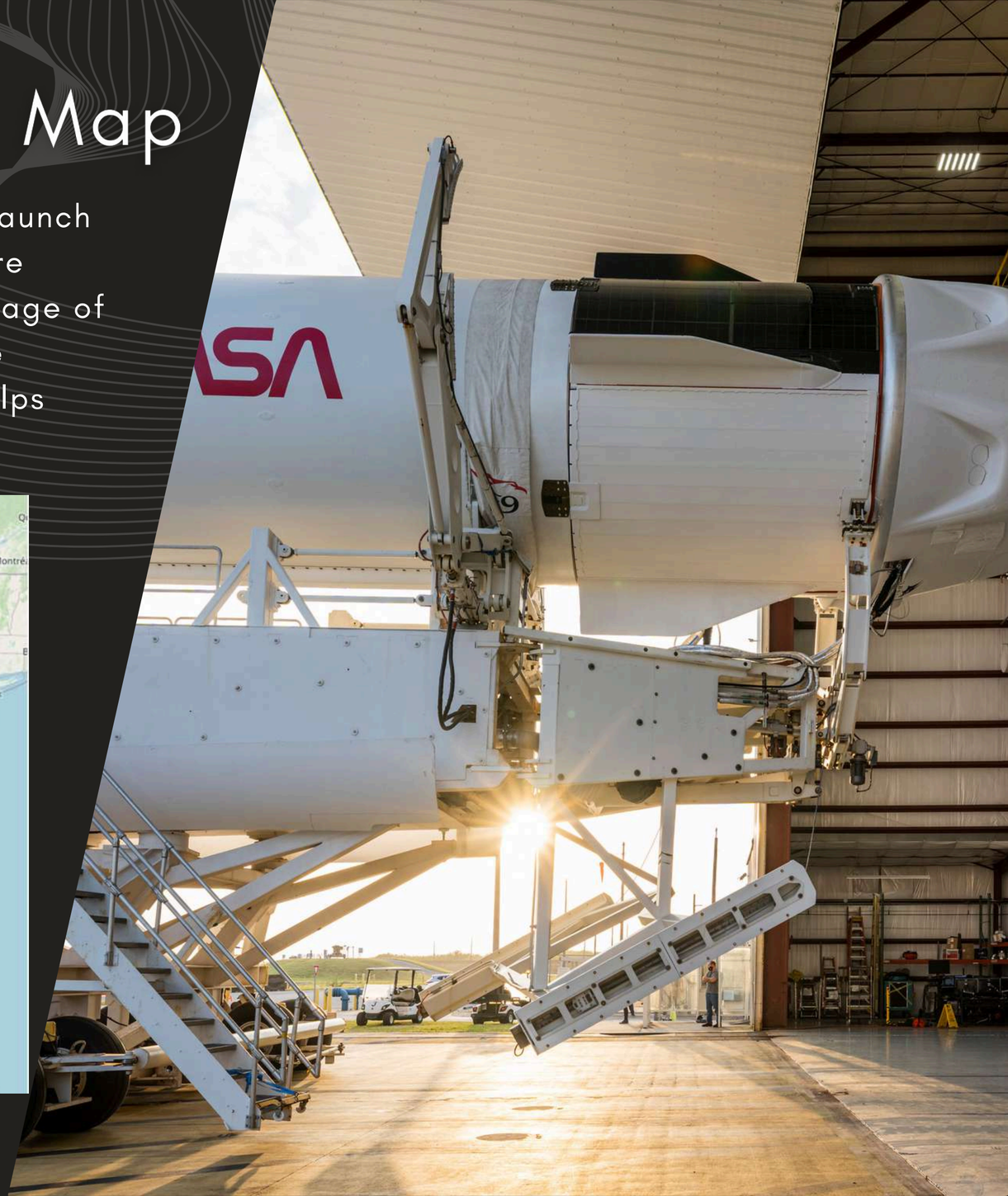
Out of a total of 31 missions, **11** ended with **success**, **10** with **failure**, and in **10** cases, there was **no landing attempt** which was most frequent

LAUNCH SITES PROXIMITIES ANALYSIS

Launch Sites Locations on Map

Map shows the **locations of launch sites**. As we can see, all launch sites are in the **southern states of the US**. Launch facilities are positioned as **close to the equator** as possible to take advantage of the **Earth's rotational velocity**, which provides a **boost to the rocket's launch speed and improves fuel efficiency**. This helps achieve the required velocity for reaching orbit.



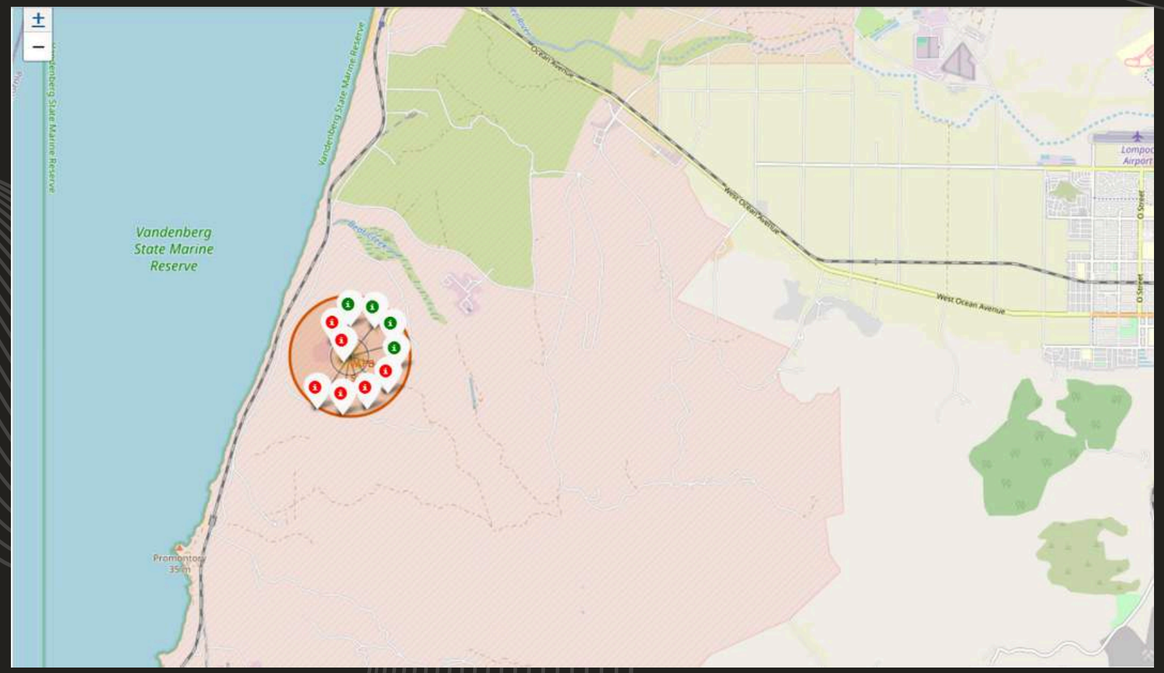


Launch Outcomes for **VAFB**

SLC 4E

Thanks to the interactive map we can easily see number of successful and failed launch otucomes for each launch site. Here we have **VAFB SLC 4E**:

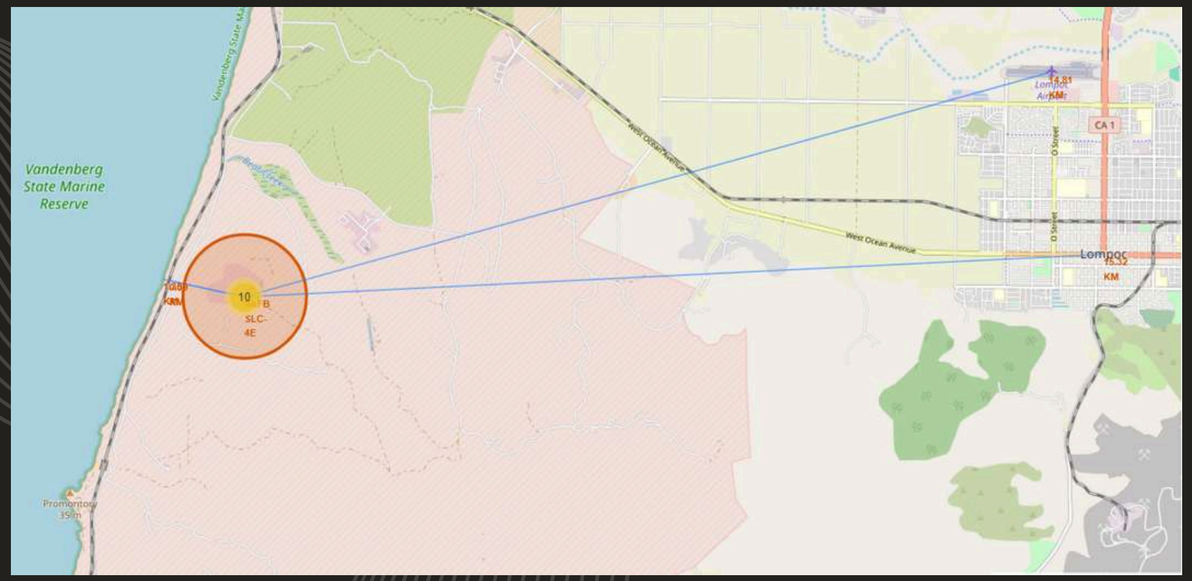
- **4 successful** Launches
- **6 failed** Launches
- Success rate 60%



VAFB SLC 4E Proximities

Launch sites are built **near cities and transportation hubs** such as railways, and airports, as well as **near coastlines**. This proximity to transportation infrastructure facilitates the **efficient movement of equipment and personnel**. Being near the coast allows rockets to be launched over open water, **minimizing the risk to populated areas** in case of an incident during launch.

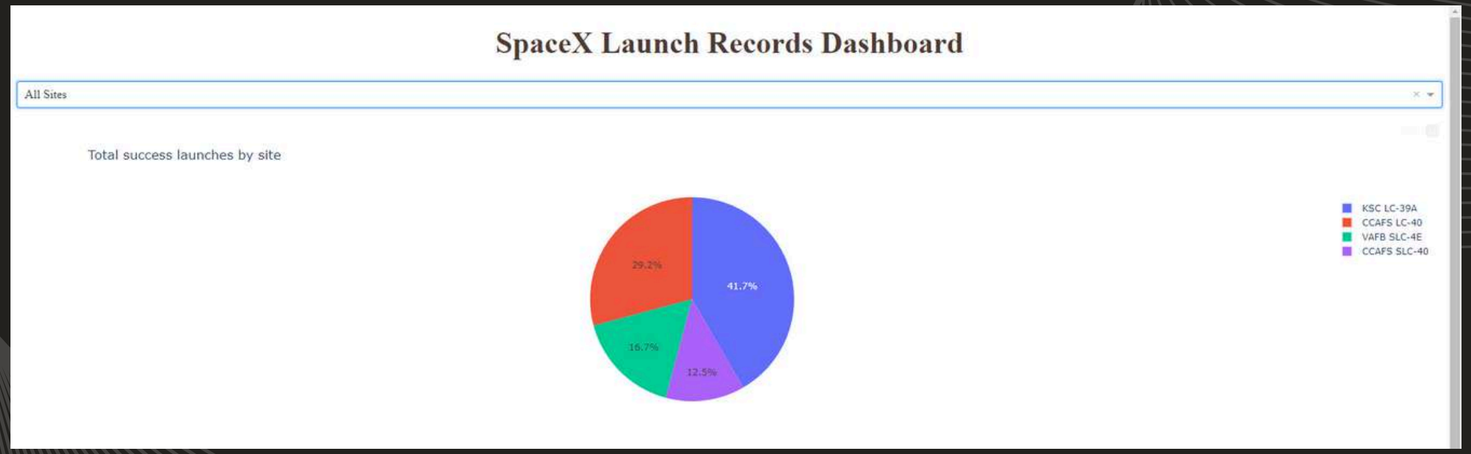
- **0.09 km** to **railway**
- **1.35 km** to **coastline**
- **14.81** km to **airport**
- **15.32** km to **city**




LAUNCH
OUTCOMES
ANALYSIS WITH
INTERACTIVE DASH

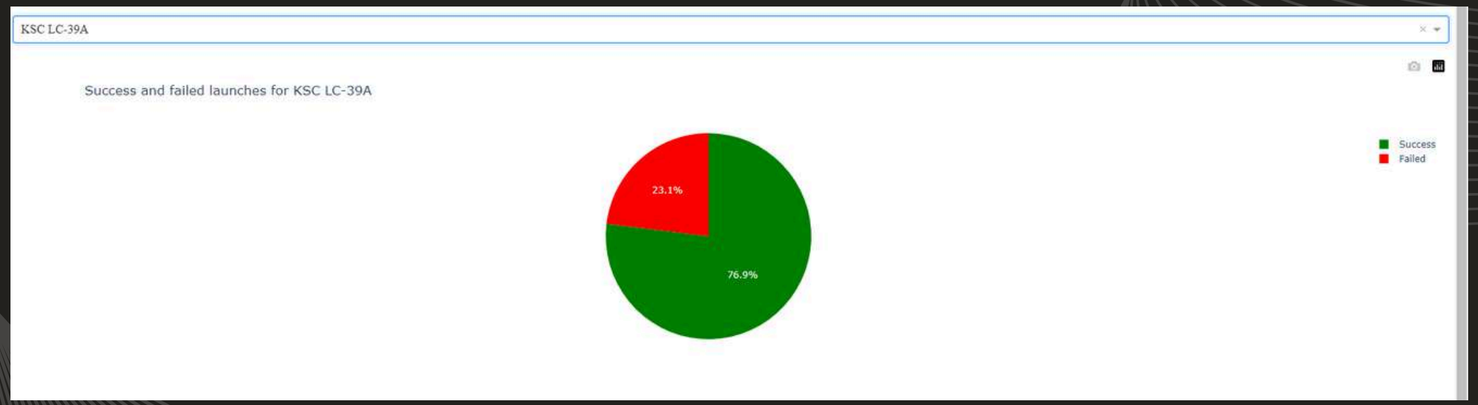
Count of launch succes for each site

- **Highest launch** success **KSC L-39A**
- **Lowest launch** success **CCAFS SLC-40**



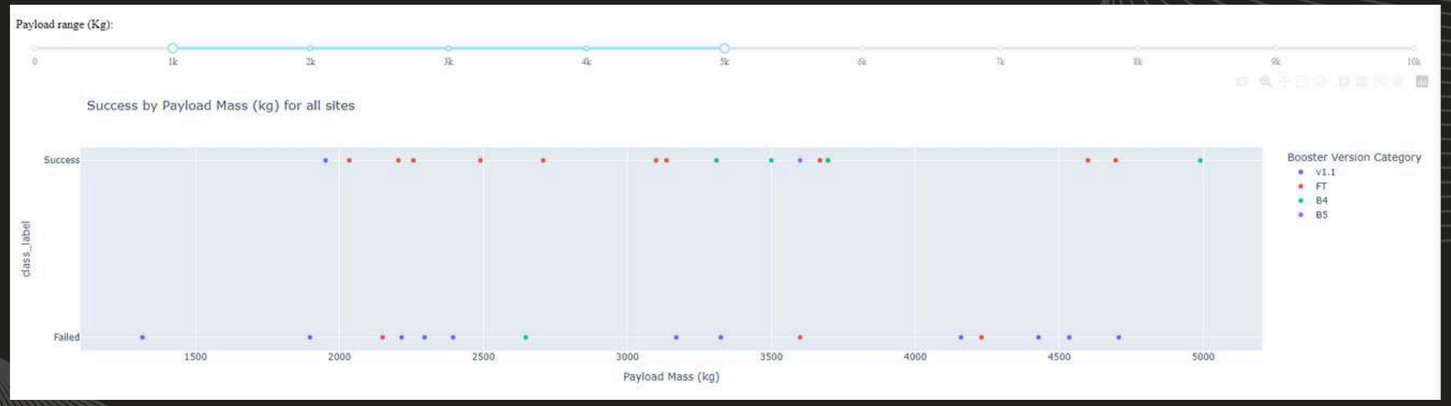
Launch success ratio for **KSC L-39A**

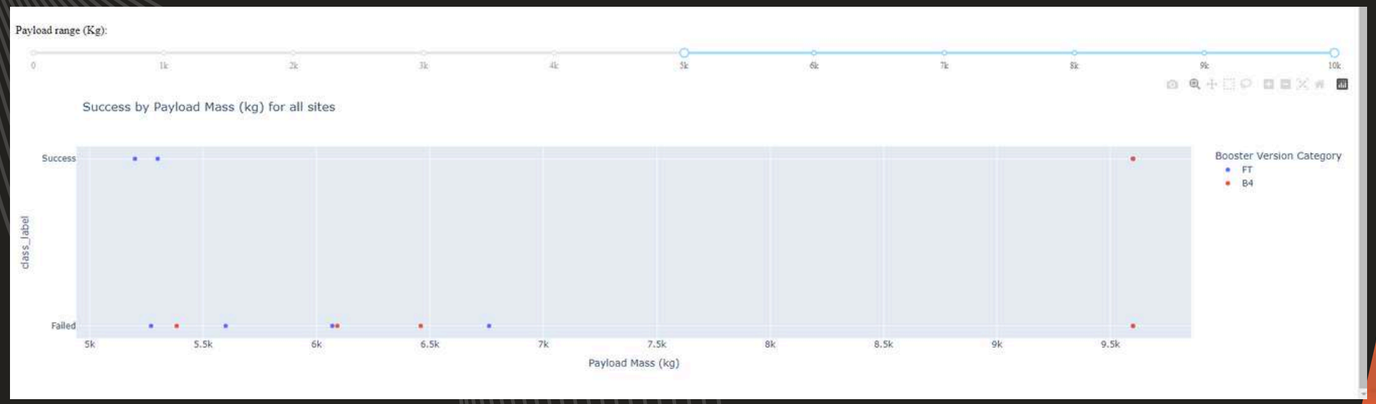
- **76,9% Success rate**
- **23,1% Failure rate**



Payload vs Launch Outcome for all sites

- **Lower payload** = **better success rate** of launch outcome: **46%** in range 0kg-5 000kg, **27%** **in** range 5 000kg-10 000kg
- There is much **more mission** with **lower payload** mass: **39** in range 0kg-5 000kg, **11** in range 5 000kg-10 000kg





PREDICTIVE ANALYSIS (CLASSIFICATION)

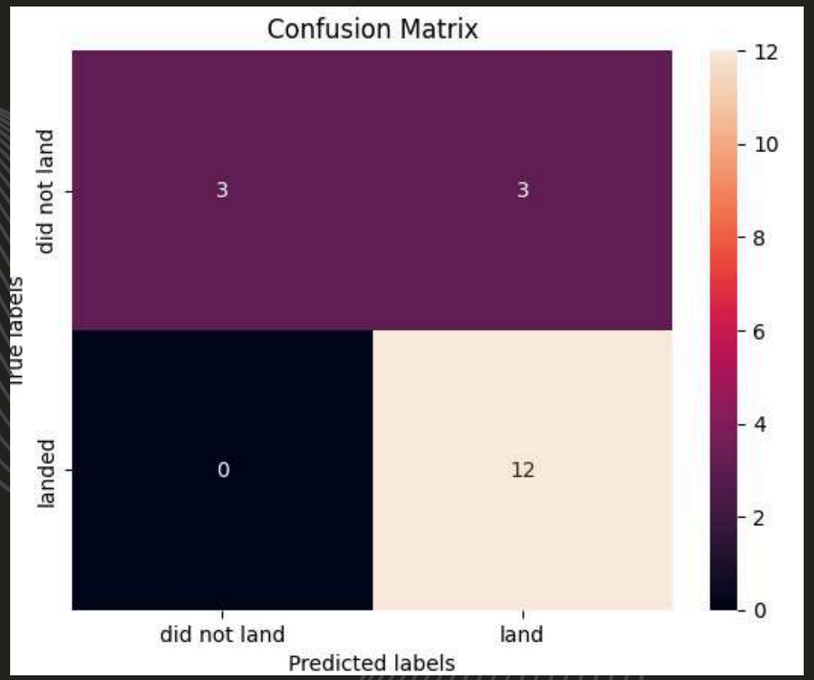
Models evaluation

- **Accuracy Score, Jaccard Score, F1 Score, and Recall Score** are at the **same level**, which is likely due to the **small dataset**.
- The **mean cross-validation score** is slightly **higher** for **Logistic Regression** compared to other models.

Method	Accuracy Score	Jaccard Score	F1 Score	Recall score	Mean corss-validation score
Logistic Regression	0.833333	0.8	0.888889	1.0	0.811111
Support Vector Machine	0.833333	0.8	0.888889	1.0	0.800000
Decision Tree	0.833333	0.8	0.888889	1.0	NaN
K Nearest Neighbors	0.833333	0.8	0.888889	1.0	0.800000

Confusion Matrix for Logistic Regression

- All confusion matrices were identical.
- **True positive: 3**
- **False negative: 3**
- **False positive: 0**
- **True negative: 12**
- A false negative count of 3 means that the model incorrectly predicted 3 instances as land when they should have been did not land.
- A false positive count of 0 means that the model correctly predicted all instances of successful landings.



CONCLUSIONS

Conclusions

1. **Best Model:**

- The best model is **Logistic Regression** with **parameters: 'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'**.

2. **Launch Sites:**

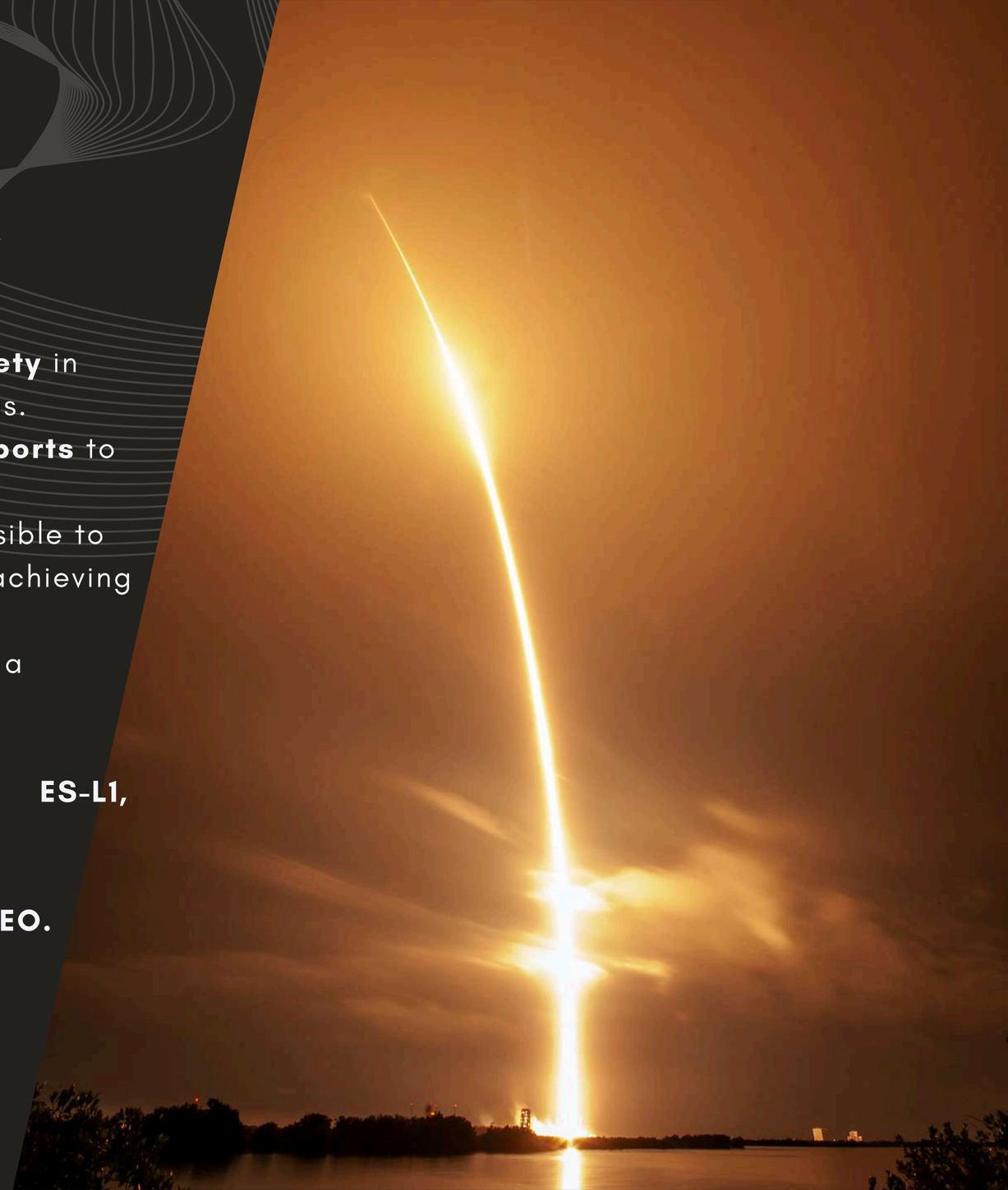
- **Launch sites** are often positioned **near coastlines** to **ensure safety** in case of problems during launch, minimizing risks to populated areas.
- **Launch sites** are typically located **near cities, railways, and airports** to **optimize accessibility and infrastructure support**.
- **Launch sites** are typically located as close to the **equator** as possible to **take advantage of the Earth's rotational speed**, which helps in achieving higher launch efficiency.
- **KSC LC-39A** has the **highest count of successful launches**, with a success rate of **76.9%**.

3. Orbits:

- Orbits **ES-L1, GEO, MEO, and SSO** have a **100% success rate but ES-L1, HEO, GEO, and SO** have only had **one launch each**.
- **GTO** is **the most commonly used orbit**.
- **Recently, the most frequently** used orbits ar**e SSO, MEO, and VLEO.**

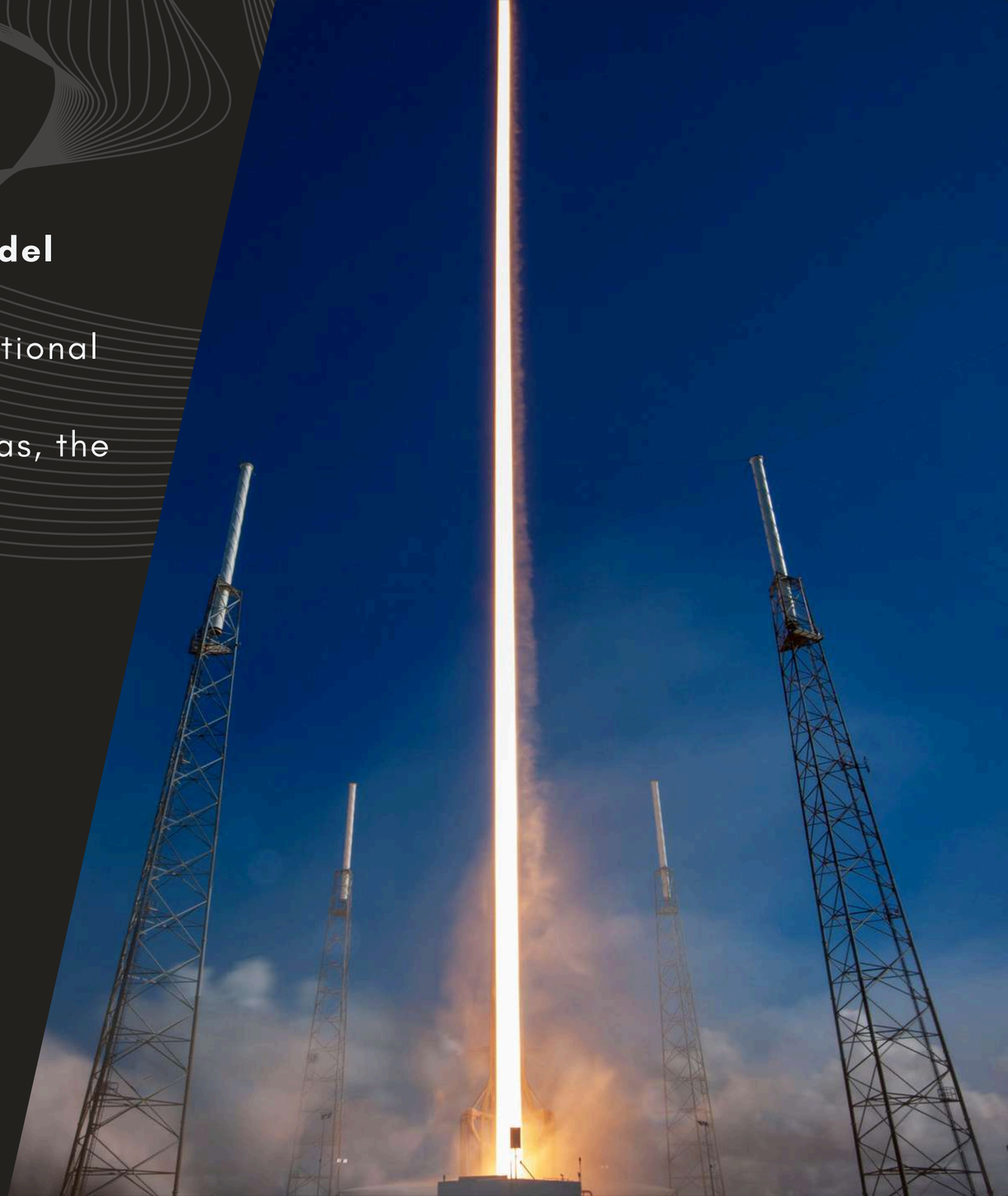
4. **Payload Mass:**

- Missions with **lower payload** masses are more likely to result in a **successful launch outcome**.



Things to consider

- **The dataset is small,** which **causes problems** during **model creation,** such as **overfitting**.
- Every year, new missions are launched, which provide additional data.
- Use of **different classification methods** compared such as, the **Naïve Bayes classifier**.



Credits:

- All photos used in this presentation were download from: flickr.com