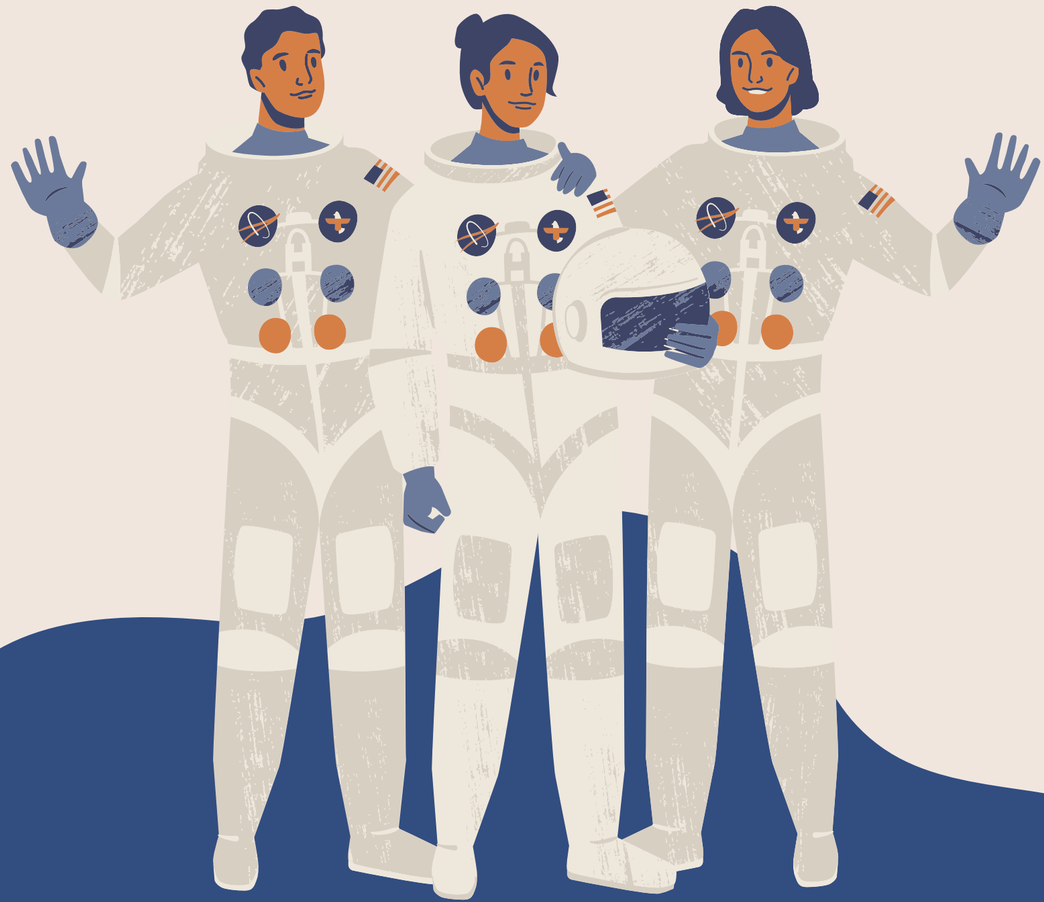
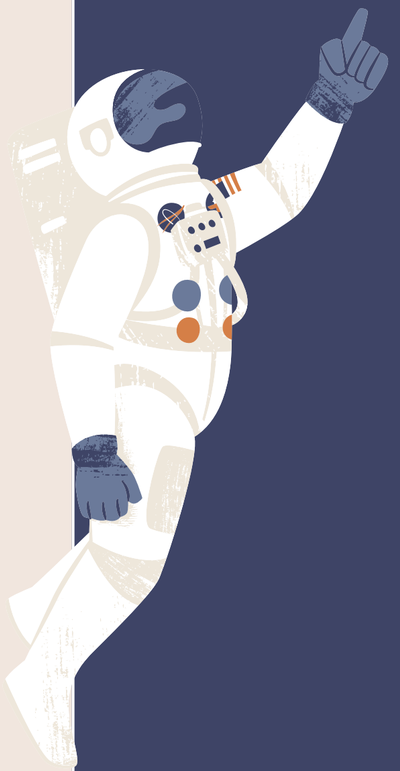
APPLIED DATA SCIENCE CAPSTONE

SPACE Y

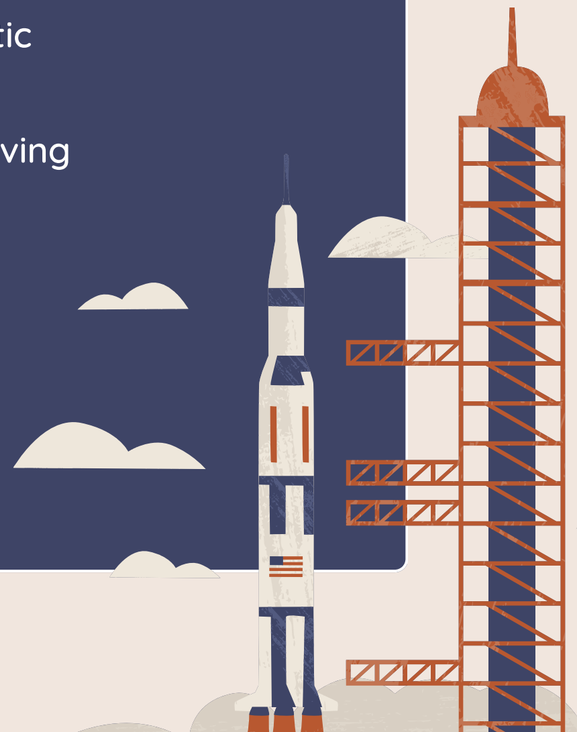


BY: JOHN EINER A. ZOLETA



EXECUTIVE SUMMARY

- The goal is to find out what characteristics are most important for a first-stage landing to be successful and how to make space missions more efficient.
- Methods: web scraping and API data collecting; exploratory data analysis (EDA); logistic regression, decision trees, and SVM predictive models.
- Important conclusions: Success rates have risen with time, with the KSC LC-39A site having the best success rate. Some orbits have a success rate of 100%.



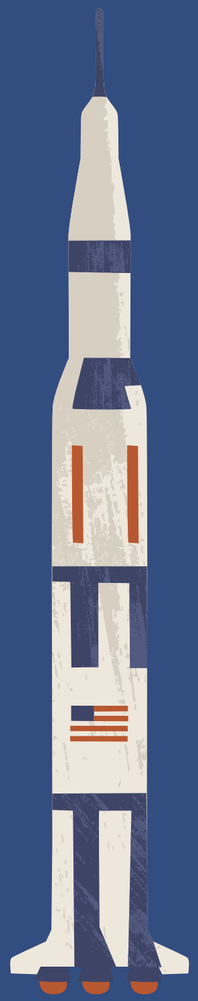
INTRODUCTION

Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch.

Which variables most affect landings that are successful?
What impact do different operating circumstances have on success rates?
How can we use past data to forecast the chance of a successful landing?

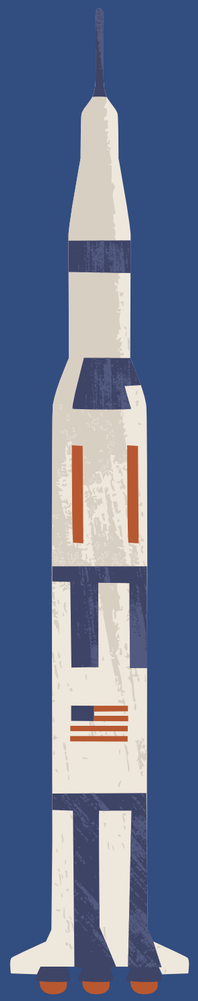
METHODOLOGY

- Methodology for data collection: Wikipedia site scraping and the SpaceX API were used to gather data.
- Perform data wrangling; use one-hot encoding for categorical features.
- Use visualization and SQLL for exploratory data analysis (EDA).
- Use Folium and Plotly Dash for interactive visual analytics.
- Use classification models for predictive analysis, and learn how to create, adjust, and assess classification models.



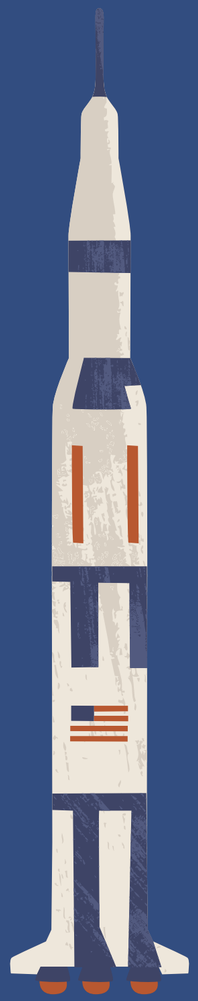
DATA COLLECTION: API

- Request via the Space X APIs.
- Lists (Launch Site, Booster Version, Payload Data) +.JSON file.
- Json_normalize to DataFrame conversion of JSON data.
- Dictionary-relevant information.
- To a DataFrame, cast a dictionary.
- Only include Falcon 9 launches in the data filter.
- Use the mean to impute missing PayloadMass values.



DATA COLLECTION: WEB SCRAPING

- Request Wikipedia HTML.
- Parse using BeautifulSoup.
- Aquire launch info HTML table.
- Assign dictionary to the Dataframe.
- Extract data from table cells iteratively and store it in a dictionary.
- Create dictionary.



DATA WRANGLING

Make a training label that designates landing outcomes, with 0 denoting a failure and 1 denoting a success. The landing site and the mission's outcome are the two main components of the outcome column. There will be a new label column called "Class"; if the mission's outcome was successful, the value will be 1; if not, it will be 0.

Mapping the Values:

- Assign 1 to the True Ocean, True RTLS, and True ASDS landings that were successful.
- Give the following landing attempts a score of 0: False Ocean, False RTLS, None None, False ASDS, and None ASDS.

EDA WITH VISUALIZATION

Important factors such as Flight Number, Payload Mass, Launch Site, Orbit, Class, and Year were subjected to exploratory data analysis (EDA).

Utilized Visualizations:

- To look into the interactions between these variables and find possible patterns to apply in machine learning training, scatter plots, line charts, and bar plots were used.
- Plots examined consist of:
  - Flight Number versus Mass of Payload
  - Launch Site vs. Flight No.
  - Launch Site vs. Payload Mass
  - Orbit vs. Success Rate
  - Aircraft Number versus Orbit
  - Payload Volume versus Orbit
  - Annual Trend of Success Rate

In order to guide the feature selection procedure for training the machine learning models, these visualizations were essential in spotting trends and connections between the variables.

EDA WITH SQL

The dataset was successfully imported into an IBM DB2 database, and to obtain more in-depth knowledge of the data, SQL queries were run via Python integration.

Questions centered on:

- Getting the names of launch sites.
- Examining the results of missions.
- examining payload dimensions for different clients.
- Examining booster iterations and landing performance.

These questions had a crucial role in developing a thorough comprehension of the dataset, which further influenced the process of analysis and model construction.

MAP WITH FOLIUM

Key data, such as launch locations, landings that were successful or unsuccessful, and proximity to significant infrastructure, such highways, railroads, cities, and coastlines, were plotted on folium maps.

This graphic aids in explaining the rationale behind the launch sites' strategic locations and offers insightful information about their deployment. It also emphasizes the connection between regional considerations and landing success.

DASHBOARD WITH PLOTLY DASH

An interactive scatter plot and pie chart are included in the dashboard to examine launch site performance and landing success.

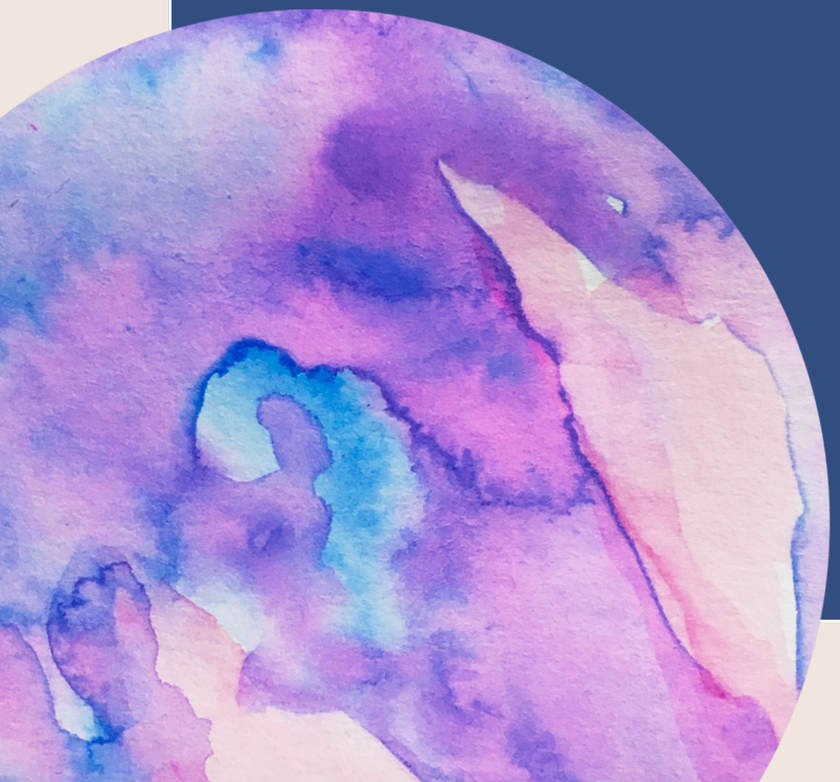
- Users can choose between seeing the distribution of successful landings across all launch locations and concentrating on the success rates of individual sites by navigating the pie chart.
- Users can select a specific launch site or all launch sites, and they can also use a slider to change the payload mass from 0 to 10,000 kg. These two inputs are accepted by the scatter plot.

While the scatter plot shows how success varies depending on launch site, payload mass, and booster version category, the pie chart clearly visualizes the success rates at various launch sites.

PREDICTIVE ANALYSIS

The features were scaled using Standard Scaler after the dataset's label column "Class" was removed.

- Next, train_test_split was used to divide the dataset into training and testing sets.
- For a variety of machine learning models, GridSearchCV with 10-fold cross-validation was used to determine the ideal parameters.
- The models put to the test are K-Nearest Neighbors (KNN), Decision Tree, Support Vector Machine (SVM), and Logistic Regression (LogReg).
- After each model was given a score on the test set, Confusion Matrices were created to assess how well it performed.
- The models' scores were graphically compared using a bar plot, which highlighted the models' efficacy according to the chosen evaluation metrics.



RESULTS AND DISCUSSION

Analyzing exploratory data:

- Over time, launch success rates have gotten better and better.
- Out of all launch locations, the KSC LC-39A facility has the best success rate.
- ES-L1, GEO, HEO, and SSO are among the orbits that have a 100% success rate.

Graphical Analysis:

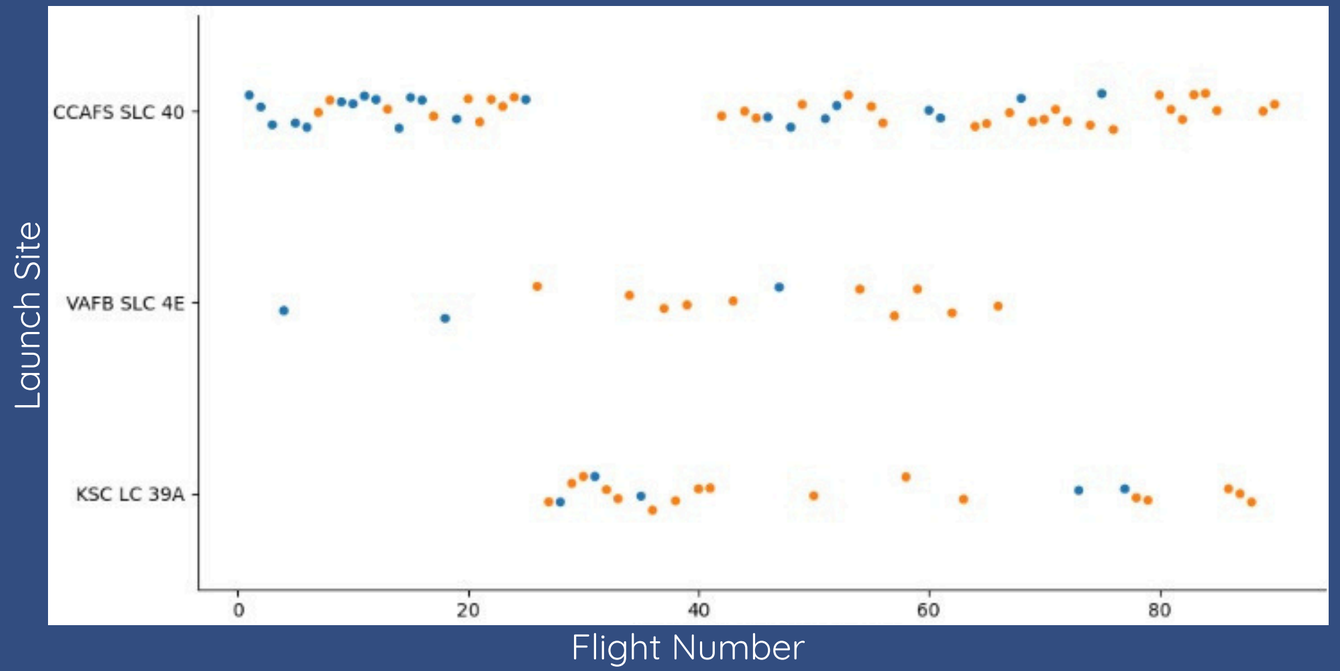
- To maximize launch conditions, the majority of launch locations are positioned strategically along the coast and close to the equator.
- In order to reduce risk in the case of a launch failure, these locations are positioned far enough from cities, highways, and railroads while still being reachable for assistance and logistics.

Analytics that predicts:

- For this dataset, the Decision Tree model proved to be the most successful predictive model, providing the best level of accuracy in predicting launch outcomes.

FLIGHT NUMBER VS LAUNCH SITE

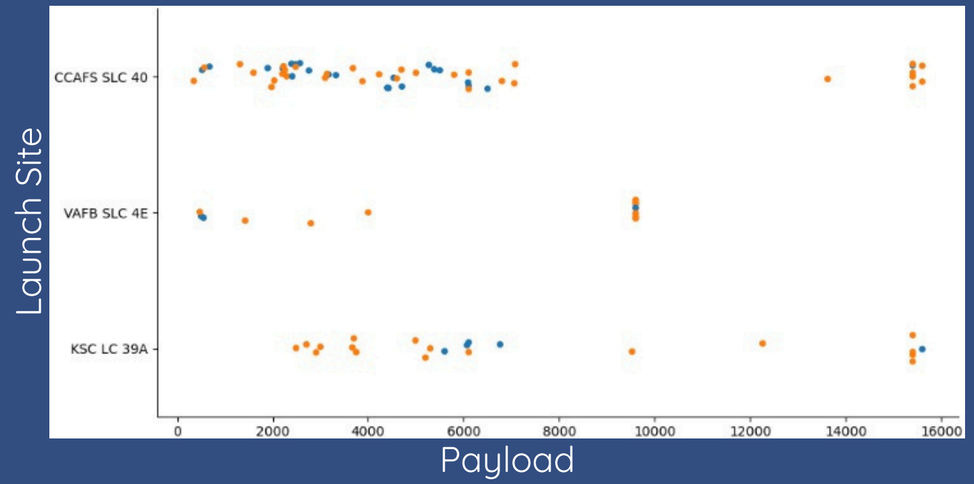
The figure makes a clear trend evident: the success rate rises with the number of flights at a particular launch point. This implies that additional launches lead to better results, probably because of more site improvements and expertise.



PAYLOAD VS LAUNCH SITE

Exploratory Data Analysis

- A higher success rate is typically correlated with a larger payload mass.
- Most of the launches that had payloads larger than 7,000 kg were successful.
- For payloads under 5,500 kg, the KSC LC-39A launch site has consistently achieved 100% success rate.
- There have been no missions launched from the VAFB SLC-4E location that carry payloads larger than about 10,000 kg.

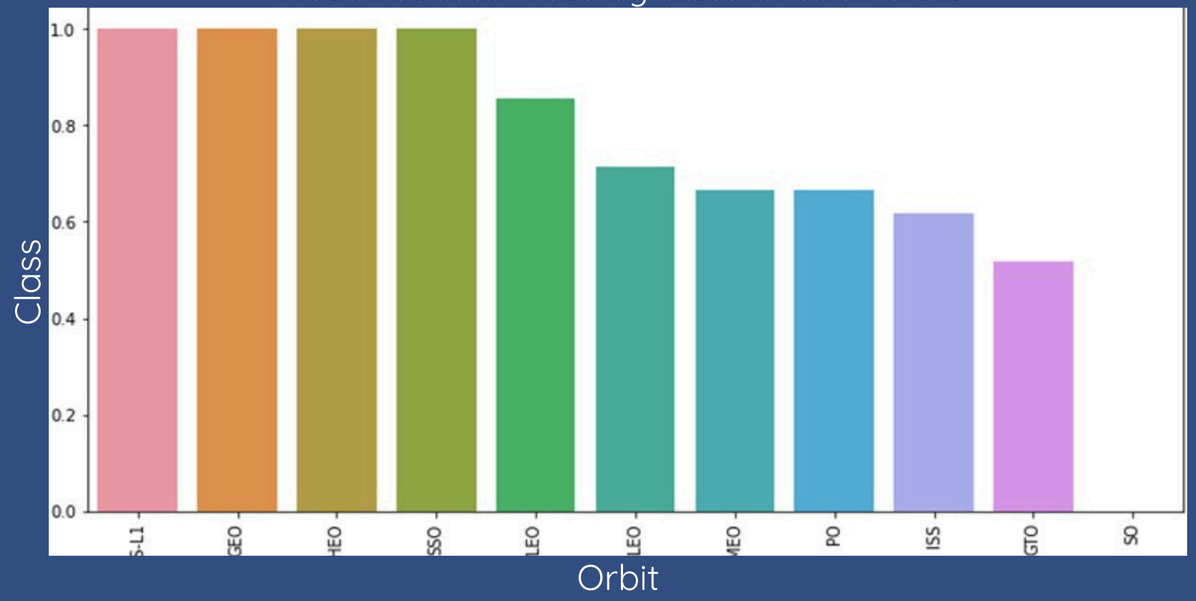


SUCCESS RATE VS ORBIT TYPE

Orbit's Success Rate:

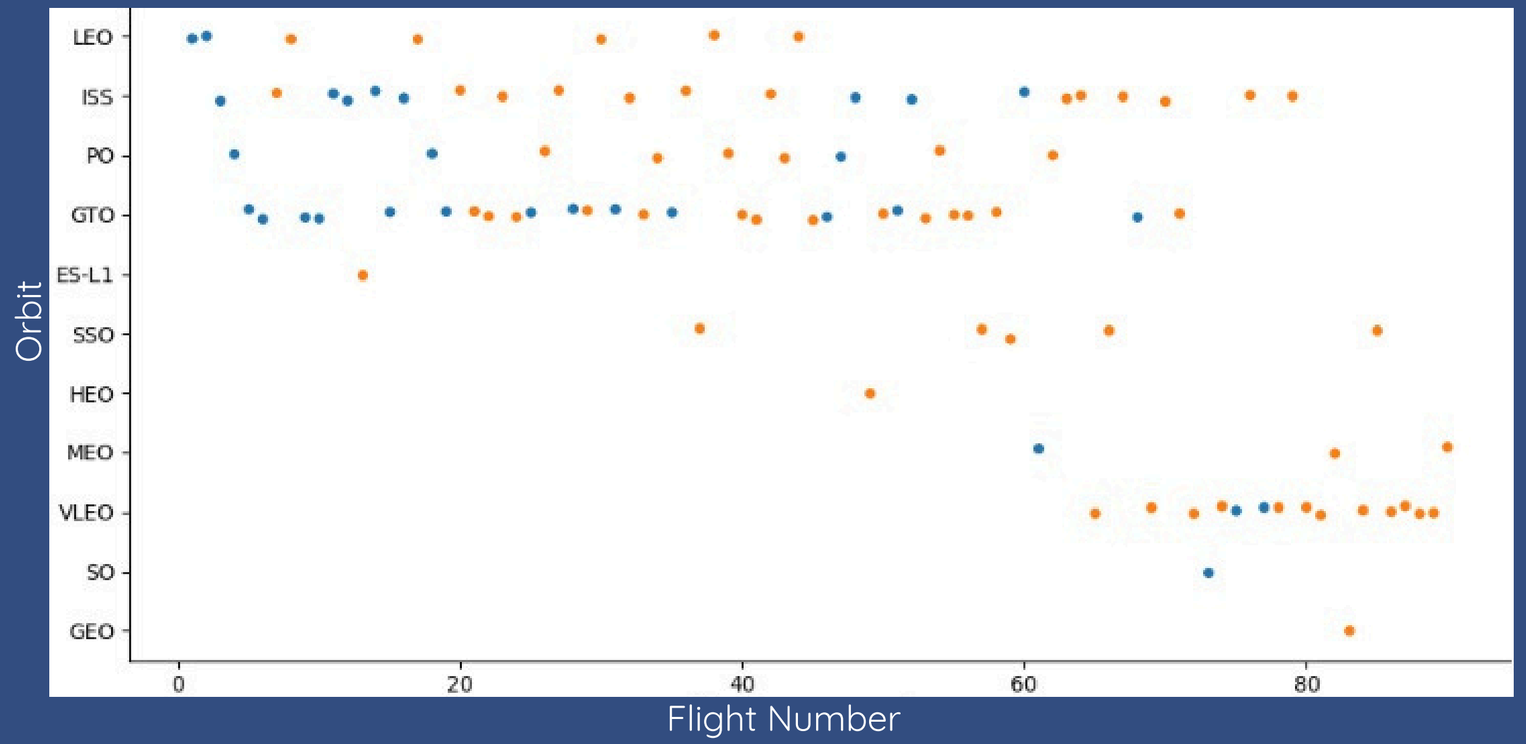
Orbits of ES-L1, GEO, HEO, and SSO have 100% success rate.
Success rate: 50% - 80% for GTO, ISS, LEO, MEO, and PO orbits.
0% of the time: orbit of SO.

Plot of success rate by class of each orbit



PAYLOAD VS ORBIT TYPE

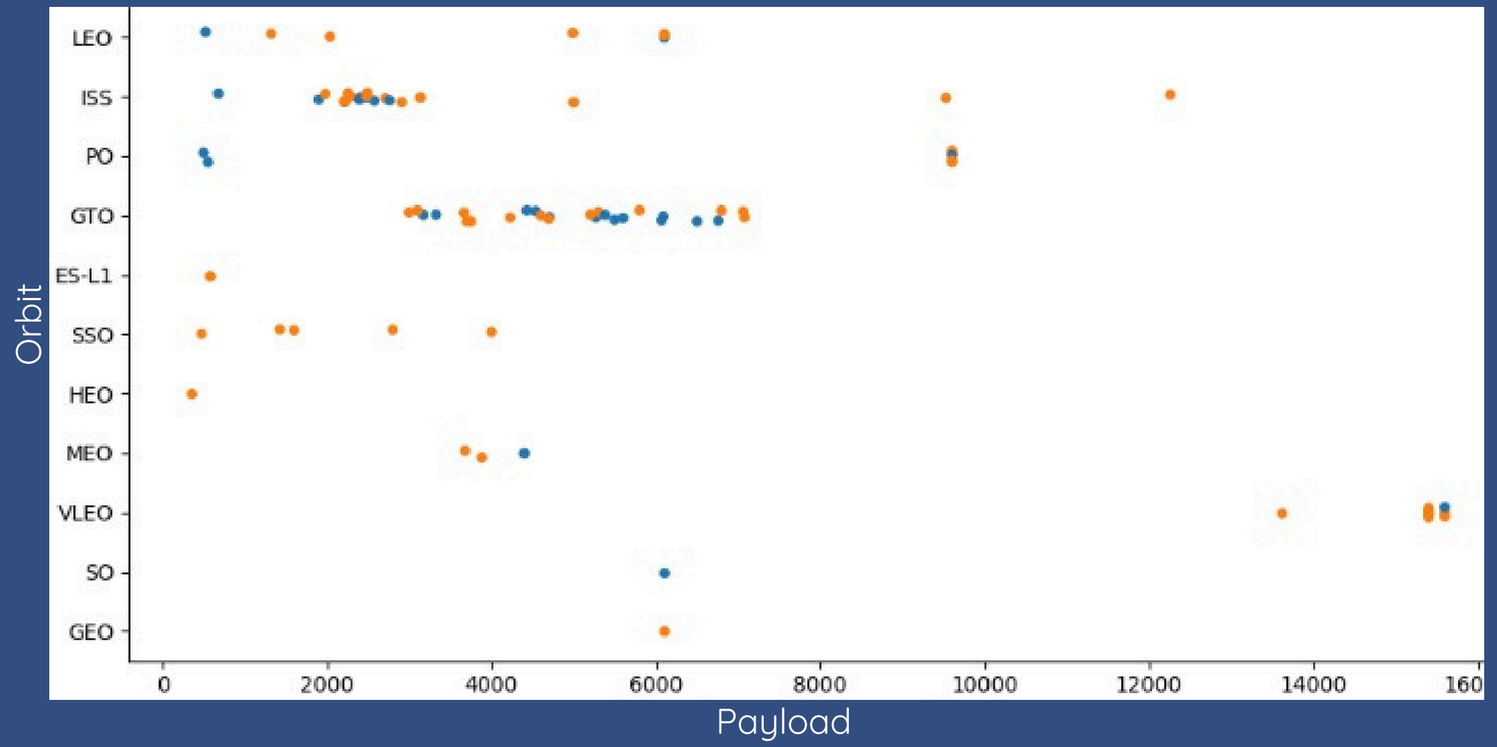
The link between Flight Number and Orbit Type is depicted in the plot below. A clear relationship exists between the number of flights and success rates in the low-Earth orbit (LEO), suggesting that higher flight frequencies lead to better results. On the other hand, there is no obvious correlation between flight number and success in the GTO orbit.



PAYLOAD VS ORBIT TYPE

Payload versus Orbit

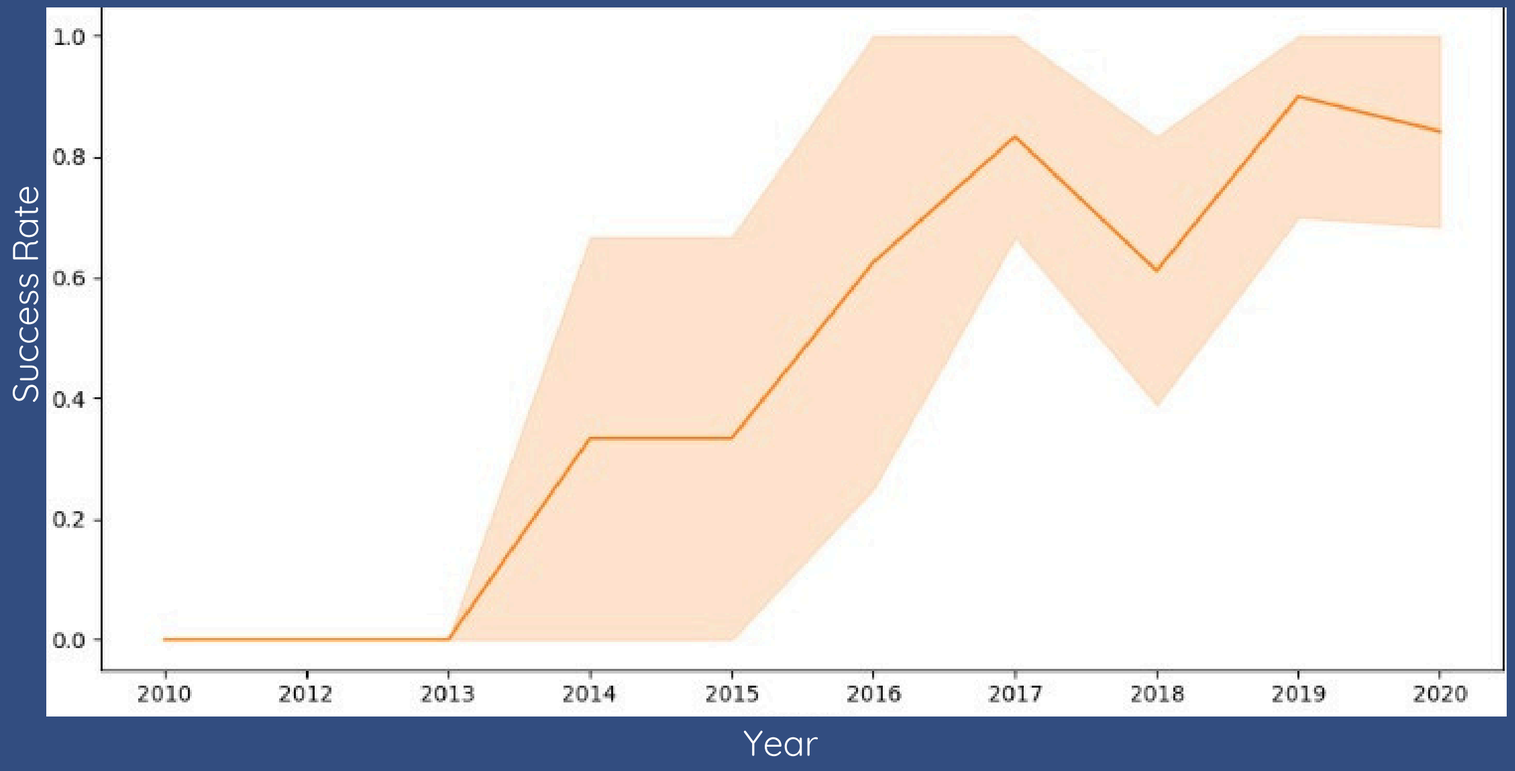
- Performance of heavy payloads is often greater in LEO, ISS, and PO orbits.
- When it comes to bigger payloads, the GTO orbit has inconsistent results, showing a range of success rates.



LAUNCH SUCCESS OVER TIME

Launch Performance over Time:

- Between 2013 and 2017, the success rate significantly improved, and between 2018 and 2019, it continued to climb.
- Nonetheless, success rates decreased from 2017 to 2018 and from 2019 to 2020 once more.
- All things considered, the pattern suggests that the success rate has increased steadily since 2013.



EDA WITH SQL

LAUNCH SITE INFORMATION

Query:

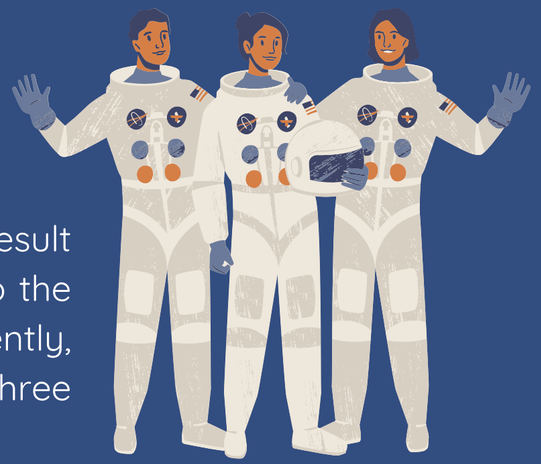
```
%%sql
SELECT UNIQUE LAUNCH_SITE
FROM SPACEXDATASET;
```

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

Launch Site Names

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

The distinct launch site names can be found by running a database query. Perhaps as a result of data input errors, it appears that CCAFS SLC-40 and CCAFS SLC-40 probably relate to the same launch location. Additionally, its previous designation was CCAFS LC-40. Consequently, we may say that CCAFS SLC-40, KSC LC-39A, and VAFB SLC-4E are probably the only three distinct launch site values.



LAUNCH SITE INFORMATION

Query:

```
%%sql
SELECT *
FROM SPACEXDATASET
WHERE LAUNCH_SITE LIKE ‘CCA%’
LIMIT 5;
```

The Launch Site name starts with "CCA"; retrieve the first five entries from the database. This will assist in pinpointing the precise launch locations linked to the CCA prefix and any relevant data entries.

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

Query:

```
%%sql SELECT SUM(PAYLOAD_MASS_KG_) \
        FROM  SPACEXTBL  \
        WHERE  CUSTOMER  =  'NASA  (CRS)';
```

We determined that the overall payload carried by boosters for NASA amounts to 45,596 kg

	total_payloadmass
0	45596

AVERAGE PAYLOAD MASS

Query:

```
%%sql SELECT SUM(PAYLOAD_MASS_KG_) \
        FROM  SPACEXTBL  \
        WHERE  BOOSTER_VERSION =  'F9 v1'
```

We observed that the average payload mass carried by the booster version F9 v1 is 2,928.4 kg.

	avg_payloadmass
0	2928.4

FIRST SUCCESSFUL GROUND LANDING

Query:

```
%%sql SELECT MIN(DATE) \
FROM  SPACEXTBL  \
WHERE  LANDING__OUTCOME = 'Success (ground pad)'
```

December 22, 2015 marked the first successful landing on a ground pad.

	firstsuccessfull_landing_date
0	2015-12-22

SUCCESSFUL LANDING WITH PAYLOAD BETWEEN 4000 AND 6000

Query:

```
%%sql SELECT MIN(DATE) \
FROM  SPACEXTBL  \
WHERE  LANDING__OUTCOME = 'Success (drone ship)'  \
PAYLOAD_MASS__KG   BETWEEN  4000 AND  6000;
```

With a payload mass of between 4,000 and 6,000 kg, this query finds the four booster types that successfully landed on a drone ship.

	boosterversion
0	F9 FT B1022
1	F9 FT B1026
2	F9 FT B1021.2
3	F9 FT B1031.2

TOTAL NUMBER OF SUCCESSFUL AND FAILED MISSION

Query:

```
%%sql SELECT MISSION OUTCOME,  COUNT(*)  as  total_number  \
FROM  SPACEXTBL  \
GROUP  BY  MISSION_OUTCOME;
```

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

Total Number of Successful and Failed Mission Outcomes:

- 1 Failure during flight
- 99 Successful Missions
- 1 Success (with unclear payload status)

BOOSTERS WITH MAXIMUM PAYLOAD

Using a subquery in the WHERE clause in conjunction with the MAX() function, we were able to determine which booster models had the maximum payload.

Query:

```
%%sql SELECT boosterversion  \
FROM  SPACEXTBL  \
WHERE  PAYLOAD_MASS_KG  =  (SELECT MAX(PAYLOAD_MASS_KG) FROM SPACEXBTL;
```

	boosterversion	payloadmasskg
0	F9 B5 B1048.4	15600
1	F9 B5 B1048.5	15600
2	F9 B5 B1049.4	15600
3	F9 B5 B1049.5	15600
4	F9 B5 B1049.7	15600
5	F9 B5 B1051.3	15600
6	F9 B5 B1051.4	15600
7	F9 B5 B1051.6	15600
8	F9 B5 B1056.4	15600
9	F9 B5 B1058.3	15600
10	F9 B5 B1060.2	15600
11	F9 B5 B1060.3	15600

FAILED LANDINGS

The month, landing outcome, booster version, payload mass (kg), launch site, and all launches in 2015 that resulted in a stage 1 drone ship landing failure are retrieved by this query. There were two instances of this kind.

Query:

```
%%sql SELECT SUBSR(Date,4,2) as month, Date, Booster_Version,  Launch_Site,  [Landing_Outcome]  \
FROM  SPACEXTBL  \
WHERE  [Landing_Outcome]  =  'Failure (drone ship')  and  substr(Date,7,4)='2015';
```

month	Date	Booster_Version	Launch_Site	Landing _Outcome
01	10-01-2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	14-04-2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

SUCCESSFUL LANDINGS

Query:

```
%%sql SELECT [Landing_Outcome], count(*)  as  count_outcome  \
FROM  SPACEXTBL  \
WHERE  DATE between '04-06-2010'  and  '20-03-2017'  group by [Landing_Outcome] order  by  count_outcomes  DESC;
```

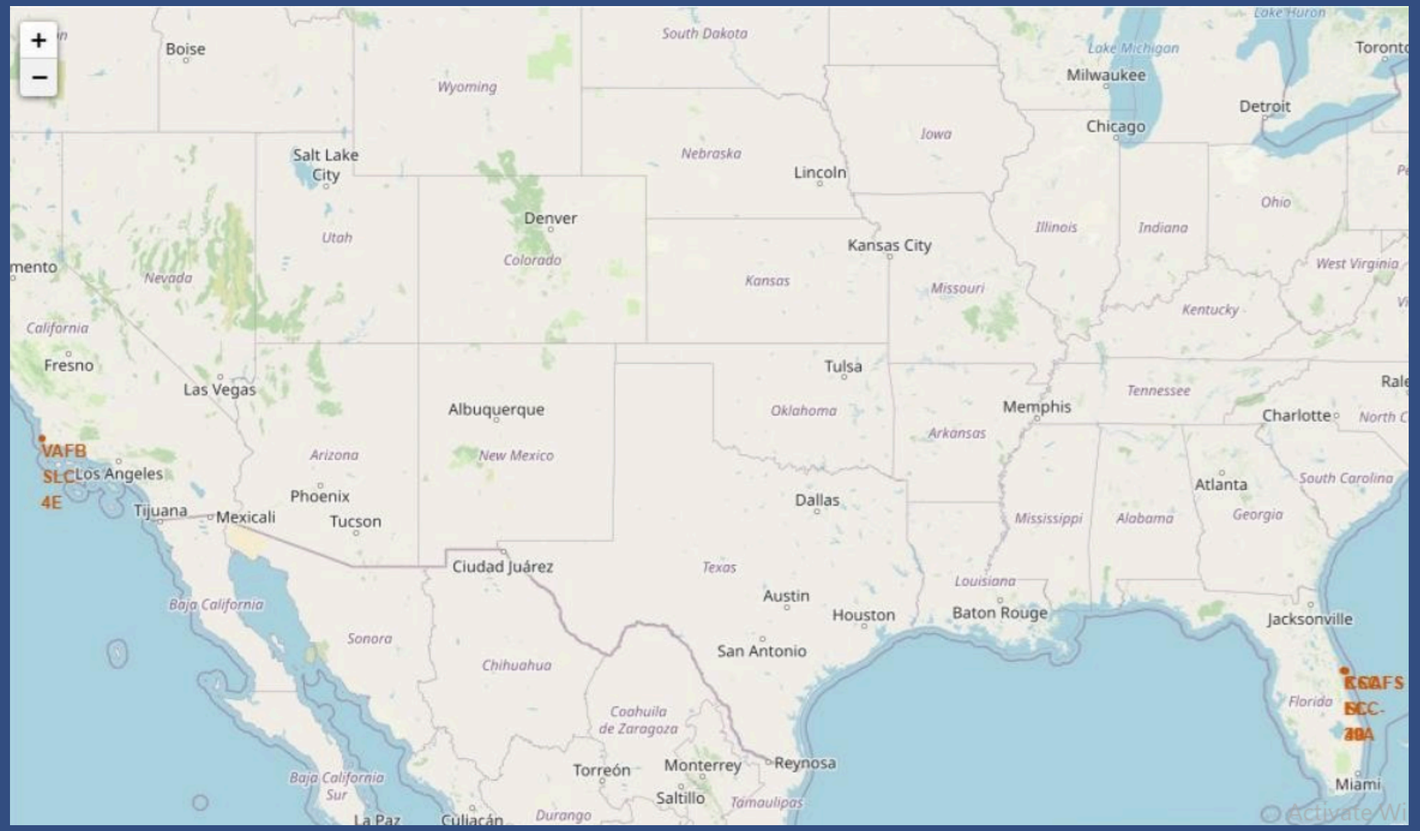
We used the WHERE clause to filter for results between 04-06-2010 and 20-03-2017. From the data, we extracted Landing results and the COUNT of landing outcomes.

To further organize the results:

- To group the landing results, we used the GROUP BY clause.
- The aggregated outcomes were sorted in descending order using the ORDER BY clause.

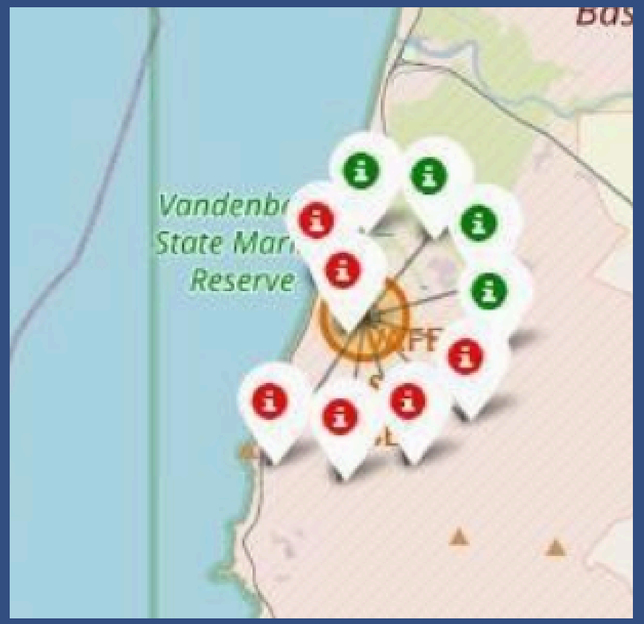
Landing _Outcome	count_outcomes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

LAUNCH SITES



COLOR-CODED LAUNCH MARKERS

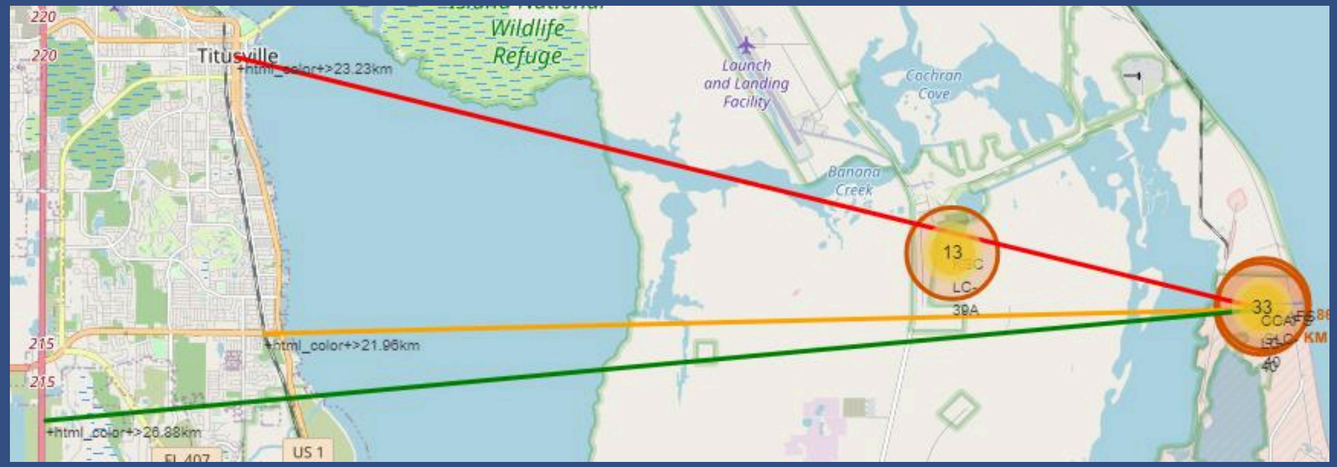
Clicking on clusters on the Folium map shows the result of each landing: successful landings are shown by green icons, and unsuccessful landings are indicated by red icons. For instance, the chart at VAFB SLC-4E displays four successful landings and six unsuccessful landings.



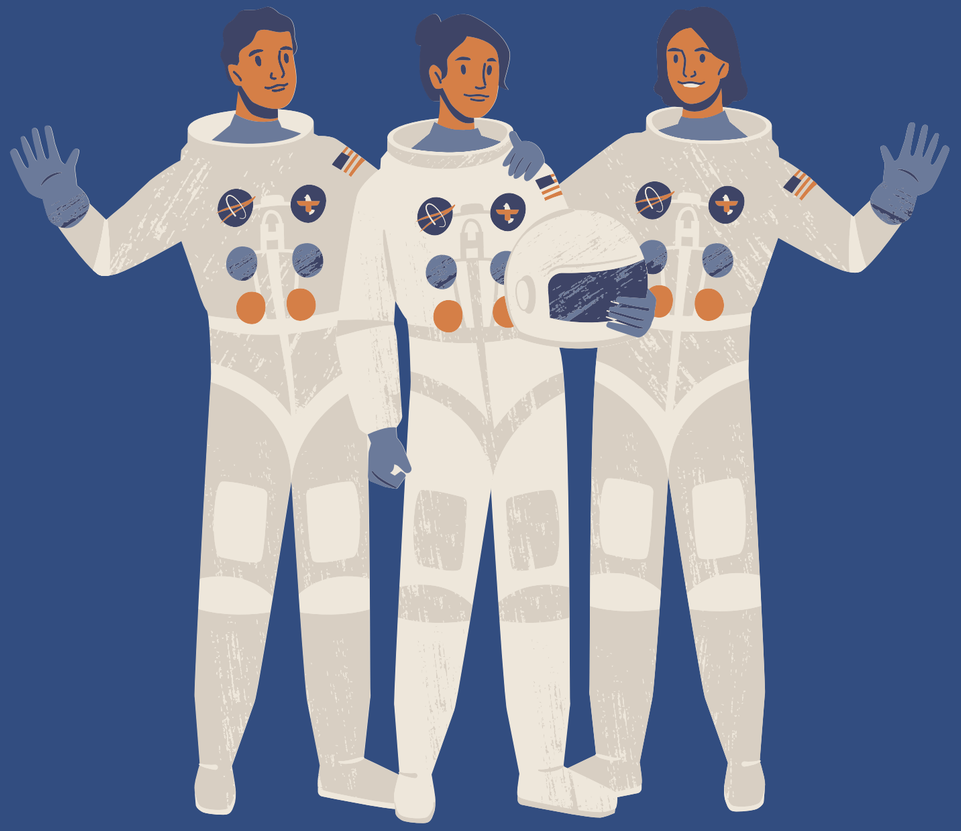
DISTANCE TO PROXIMITIES

Nearby CCAFS SLC-40:

- 0.86 kilometers from the closest coast.
- The closest train station is 21.96 kilometers away.
- Distance from closest city: 23.23 kilometers.
- 26.88 miles away from the closest motorway.



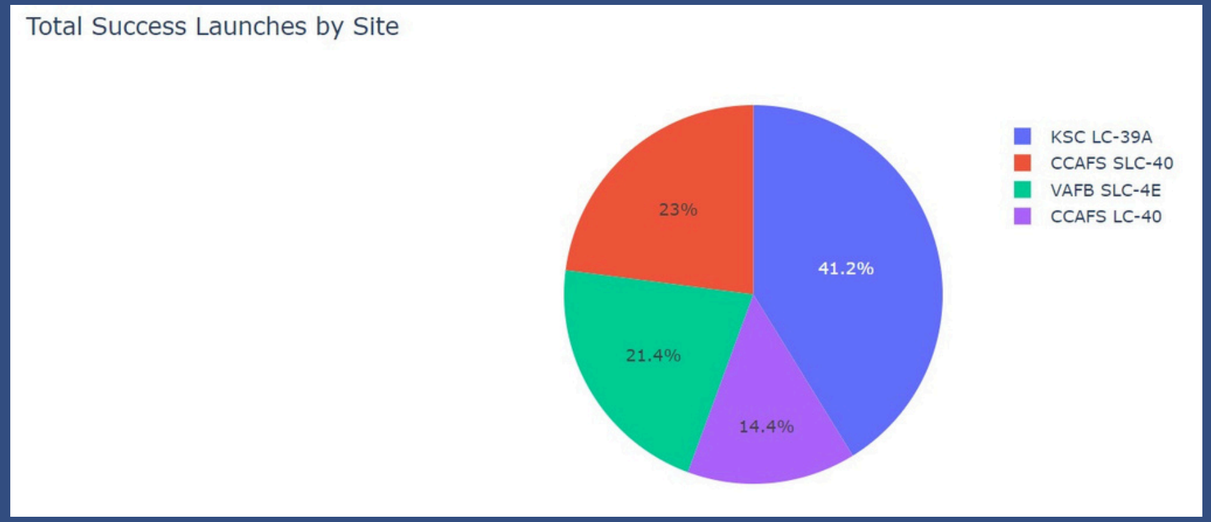
DASHBOARD VIA PLOTLY DASH



SUCCESSFUL LAUNCHES IN DIFFERENT SITES

Launch Performance by Location:

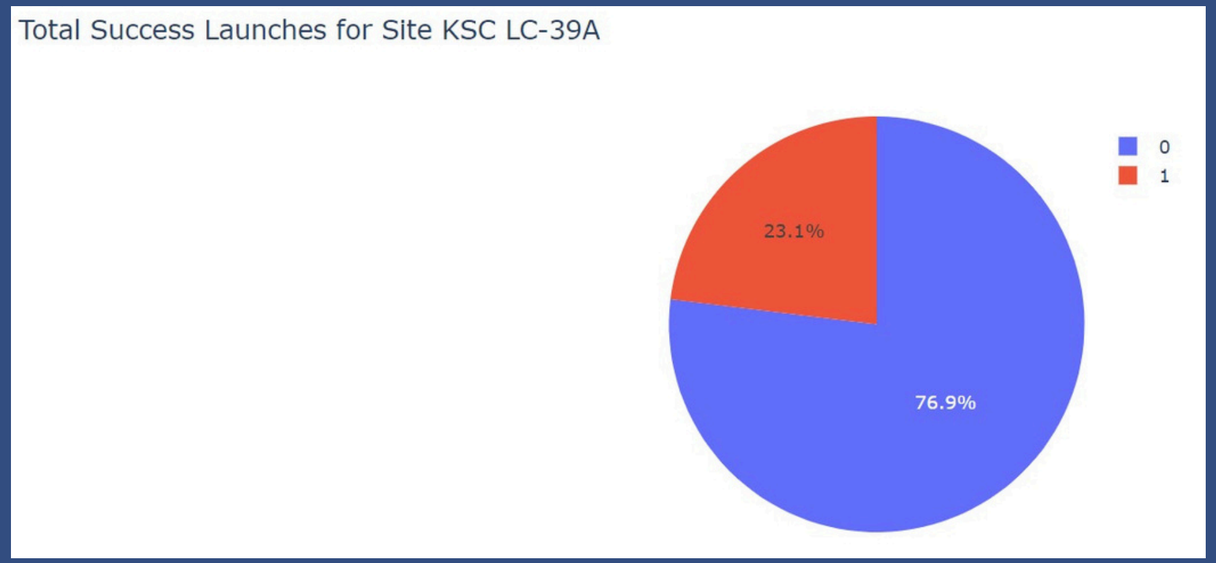
- Success as a Percentage of Total: Out of all launch sites, KSC LC-39A has had the most successful launches.



HIGHEST LAUNCH SUCCESS

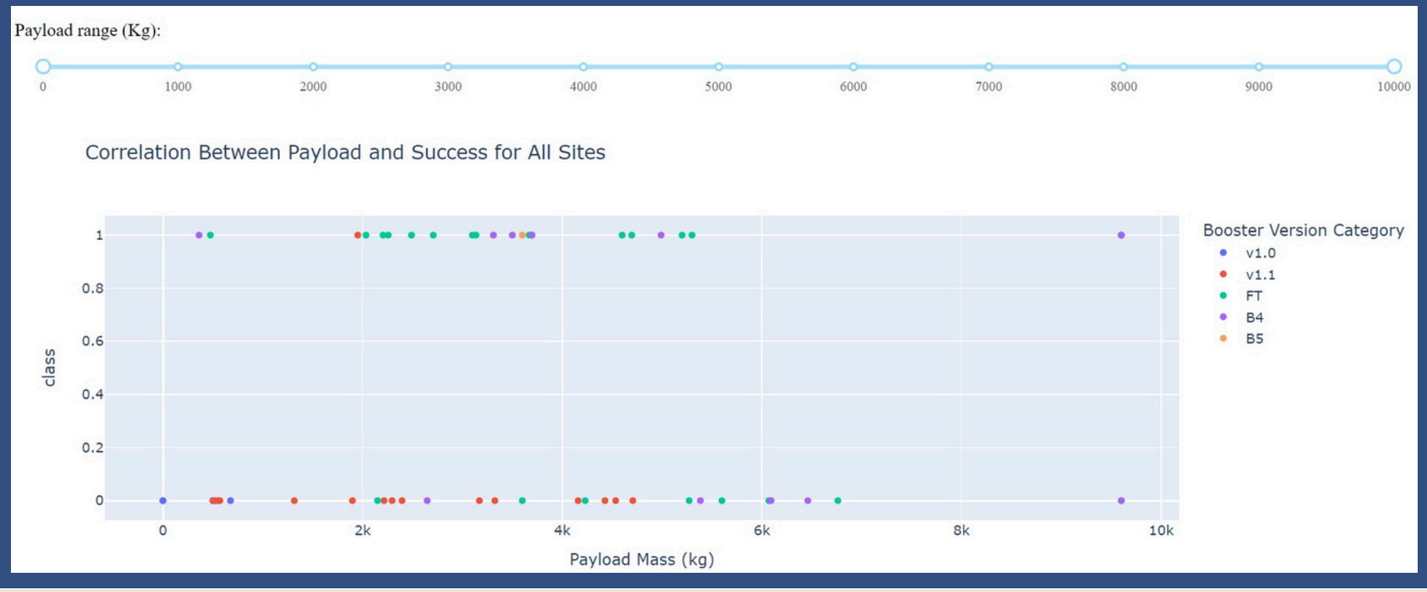
Launch Performance (KSC LC-29A)
Success as a Percentage of Total: Of all launch locations, KSC LC-39A has the highest success percentage (76.9%).

- Ten launches were successful, and three failed.

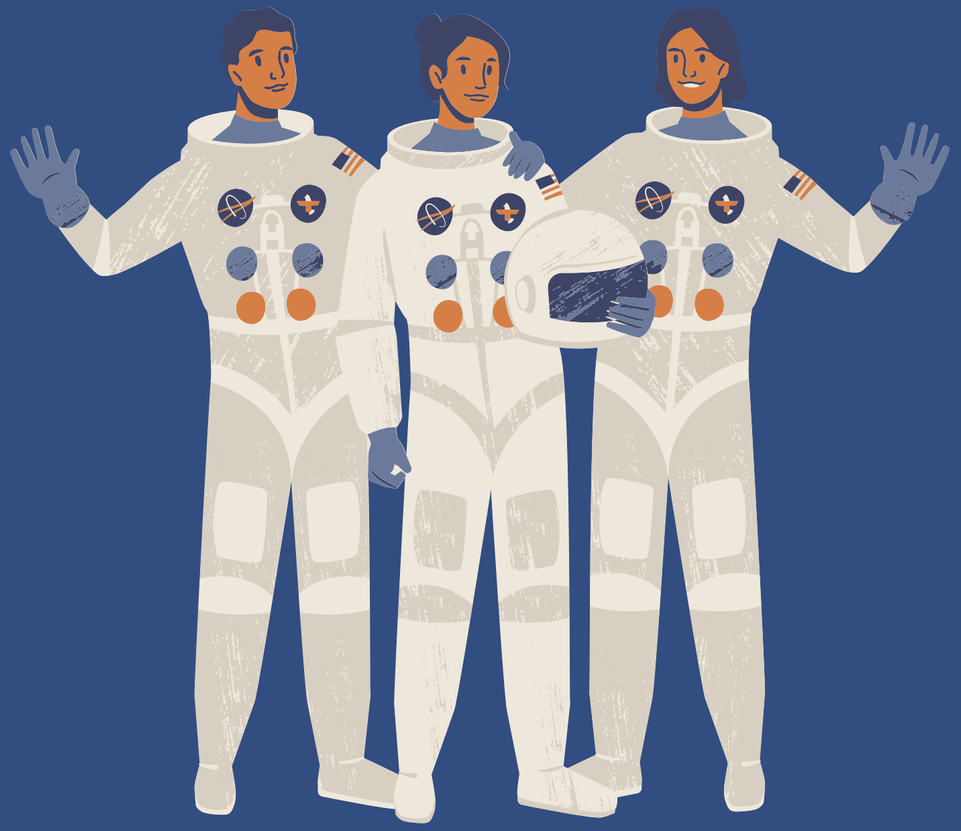


PAYLOAD VS SUCCESS - BOOSTER VERSION

- The maximum success rate is seen with payloads weighing between 2,000 and 5,000 kg; a result of 1 denotes a successful outcome, while a result of 0 denotes an unsuccessful one.

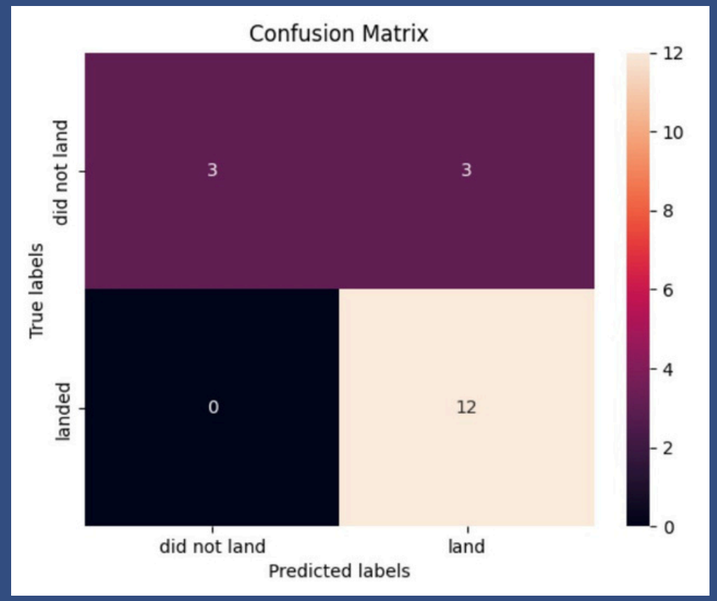


PREDICTIVE ANALYSIS

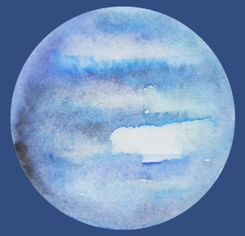
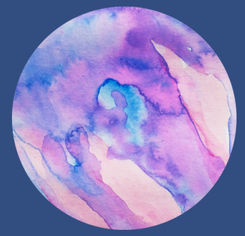
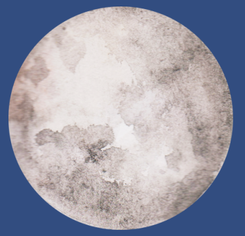


CONFUSION MATRICE

- Because every model produced the same outcomes for the test set, the confusion matrix is the same for every model. To be more precise, the models projected:
-
- When the genuine label was a successful landing, there were twelve successful landings.
- When the genuine label was a failure landing, there were three unsuccessful landings.
- Three landings that were successful despite the genuine label being failing (false positives).
- This suggests that our models often overestimate the likelihood of successful landings.



CONCLUSION

For Space Y, we created a machine learning model that can forecast the success of the first landing, possibly saving $100 million USD every launch. We generated a complete dataset stored in a DB2 SQL database and produced a visualization dashboard using information from a public SpaceX API and web scraping.

Model Accuracy: The decision tree model fared somewhat better than the others, with 83% accuracy.
Launch Success Trends: Since 2013, there has been a rise in success rates. KSC LC-39A, for example, had a 100% success rate for payloads weighing less than 5,500 kg.
Geographical Insights: All launch sites are close to the shore for safety, and the majority are close to the equator to maximize fuel efficiency.
Payload Mass: Higher success rates are correlated with payload masses that are larger.
Orbital Success Rates: As shown by the ES-L1, GEO, HEO, and SSO orbits a complete success rate.
In order to strengthen the model and increase its forecast accuracy and establish Space Y as a formidable rival in the space launch sector, we advise ongoing data collection.

THANK YOU VERY MUCH!