IBM DATA SCIENCE CAPSTONE PROJECT

WINNING THE SPACE RACE WITH DATA SCIENCE

OUTLINE

1. Executive Summary
2. Introduction
3. Methodology
4. Results
5. Conclusion
6. Appendix

SECTION 1

EXECUTIVE SUMMARY

EXECUTIVE SUMMARY

SUMMARY OF METHODOLOGIES:

This project follows these steps:

- Data Collection
- Data Wrangling
- Exploratory Data Analysis
- Interactive Visual Analytics
- Predictive Analysis (Classification)

SUMMARY OF RESULTS:

This project produced the following outputs and visualizations:

- Exploratory Data Analysis (EDA) results
- Geospatial analytics
- Interactive dashboard
- Predictive analysis of classification models

SECTION 2

INTRODUCTION

INTRODUCTION

- SpaceX has gained worldwide attention for a series of historic milestones.
- It is the only private company ever to return a spacecraft from low-earth orbit, which it first accomplished in December 2010. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars wheras 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 SpaceX for a rocket launch.
- Based of SpaceX launch data and analysis, this project resulted in a model to predict (with probability) if a Falcon 9 launch will result in a successfully landed 1st stage.

SECTION 3

METHODOLOGY

EXECUTIVE SUMMARY

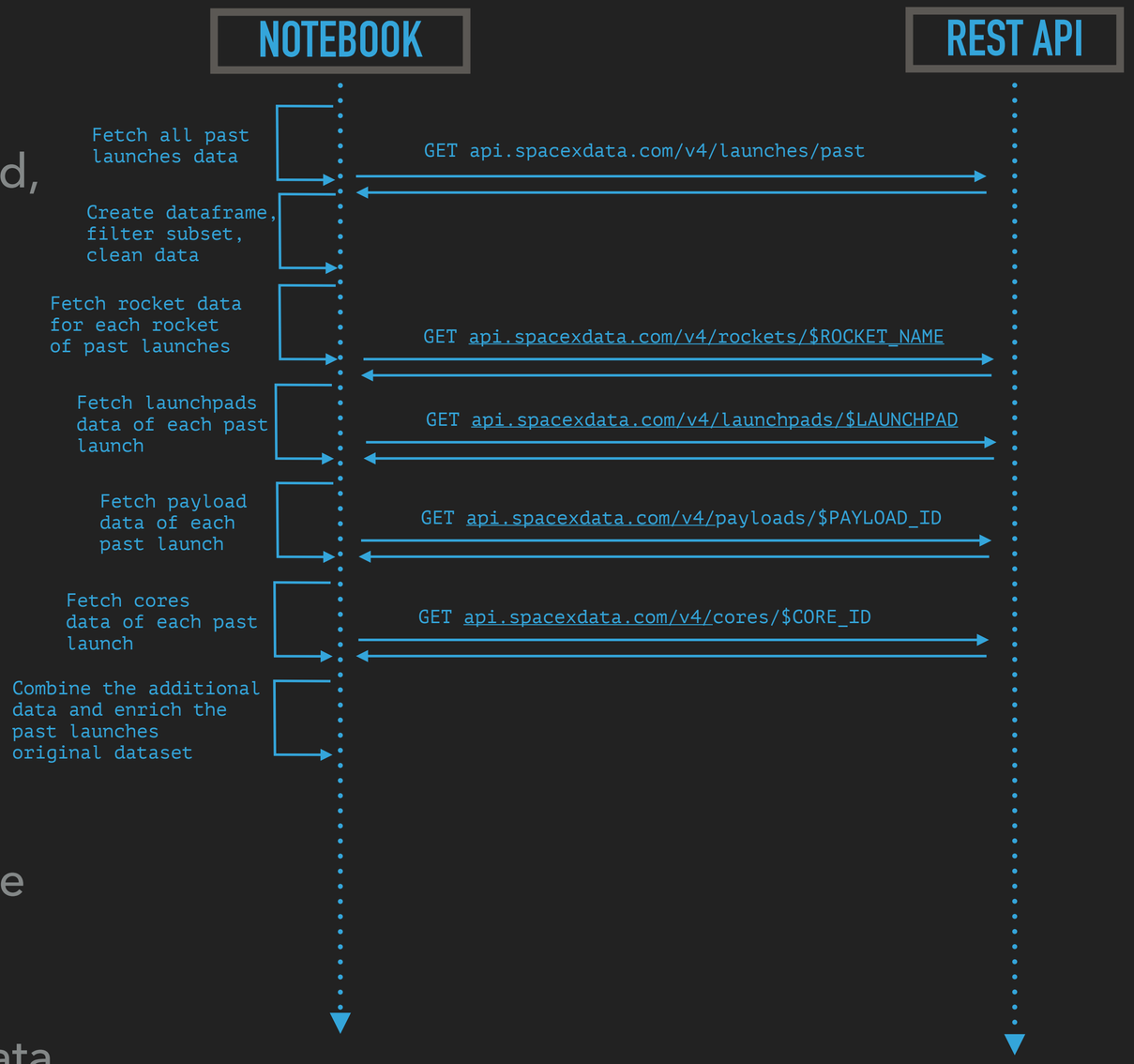
1. Data Collection
    1. Issuing requests to the SpaceX API using the Python requests library
    2. Web Scraping Wikipedia pages on Falcon 9 launches
2. Data Wrangling
    1. Remove NaN and empty values from dataset
    2. Determined:
        1. Number of launches on each site
        2. Number and occurrence of each orbit
        3. Number and occurrence of mission outcome per orbit type
    3. Added a landing outcome label with values:
        1. 1 - successful booster landing
        2. 0 - unsuccessful booster landing
3. Exploratory Data Analysis (EDA)
    1. Manipulate and evaluate the SpaceX dataset using SQL
    2. Visualize relationships between features and find patterns using Pandas, Matplotlib and Seaborn
4. Interactive Visual Analytics
    1. Geospatial analytics using Folium
    2. Creating an interactive dashboard using Plotly Dash
5. Data Modelling and Evaluation
    1. Using Scikit-Learn to:
        1. Pre-process (standardize) the data
        2. Split the data into training and testing data using train_test_split
        3. Train different classification models
        4. Find hyperparameters using GridSearchCV
    2. Plot confusion matrices for each classification model
    3. Assessing the accuracy of each classification model

See notebook on Github

DATA COLLECTION – SPACEX API

Retrieved data about launches, vehicles used, payload, launch specification and outcomes via the SpaceX REST API:

1. Fetch past SpaceX launches and parse JSON response
2. Fit data set into a Pandas data frame
3. Clean dataset: filter data to a subset, clean data, parse dates
4. Fetch booster versions, rocket payload, rocket core data, for every observation in the cleaned dataset
5. Enrich the cleaned dataset with the newly fetch data



See notebook on Github

DATA COLLECTION – WEB SCRAPING

Retrieved data about past Falcon 9 and Falcon Heavy launches from the respective Wikipedia page:

1. Fetch HTML-formatted data from Wikipedia page
2. Extract all columns and data from launches table
3. Parse launches HTML table and build Python dictionary with data
4. Build Pandas DataFrame from dictionary data

See notebook on Github

DATA MANIPULATION/WRANGLING

The SpaceX launches data is dataset with 100 launches of the Falcon 9 rocket. For each launch, the data set contains one observation (i.e. one row in a table), with different features (columns).

The dataset is of mixed datatypes (see table).

All features do not have empty (NaN) values, with the exception of the **LandingPad** feature whose 40% of the values are empty.

Feature Name	Type
FlightNumber	int64
Date	object
BoosterVersion	object
PayloadMass	float64
Orbit	object
LaunchSite	object
Outcome	object
Flights	int64
GridFins	bool
Reused	bool
Legs	bool
LandingPad	object
Block	float64
ReusedCount	int64
Serial	object
Longitude	float64
Latitude	float64

See notebook on Github

DATA MANIPULATION/WRANGLING

Initial exploratory data analysis:

The number of launches are split between three launch sites:

Launch site	Launches #
CCAFS SLC 40	55
KSC LC 39A	22
VAFB SLC 4E	13

Each launch aims to an dedicated orbit, there are 11 most common orbits in use. The launches in the dataset are distributed as such:

Orbit	Launches #
GTO	27
ISS	21
VLEO	14
PO	9
LEO	7
SSO	5
MEO	3
ES-L1	1
HEO	1
SO	1
GEO	1

See notebook on Github

DATA MANIPULATION/WRANGLING

Landings are distributed across Drone Ship landing (ASDS), controlled Ocean landing, ground landing pad (RTLS) and a landing failure (None):

Target	Success?	Landings #
ASDS	Yes	41
None	Failure	19
RTLS	Yes	14
ASDS	No	6
Ocean	Yes	5
Ocean	No	2
ASDS	Failure	2
RTLS	No	1

To determine if a booster will land, we also created a target label (column) which represents the landing outcome for the given launch, called "Outcome". The possible values are a bad outcome (0) or a good outcome (1).

Steps we took to create the label:

1. Define a set of unsuccessful (bad) outcomes
2. Create a list, `landing_class`, where the element is 0 if the corresponding row in `Outcome` is in the set `bad_outcome`, otherwise, it's 1
3. Create a `Class` column that contains the values from the list `landing_class`
4. Export the DataFrame as a .csv file.

See notebook on Github

EDA WITH DATA VISUALIZATION

The following types of plots were used to visualise the data set:

1. SCATTER PLOTS

Scatter plots were produced to visualize the relationships between:

- Flight Number and Launch Site
- Payload and Launch Site
- Orbit Type and Flight Number
- Payload and Orbit Type

Scatter plots can be helpful in examining the connections or associations that exist between two numerical variables.

2. BAR PLOTS

Bar plots were produced to visualize the relationships between:

- Success Rate and Orbit Type

Bar plots are employed to contrast a categorical variable against a numerical value. The orientation of the bar plot, whether horizontal or vertical, depends on the amount of data being presented.

3. LINE PLOTS

Line plots were produced to visualize the relationships between:

- Success Rate and Year (i.e. the launch success yearly trend)

Line plots typically display numerical data on both the x-axis and y-axis and are commonly employed to illustrate the fluctuations of a variable over a period of time.

See notebook on Github

EDA WITH SQL

To draw insights from the dataset SQL queries were used:

- Present the names of unique launch sites involved in the space mission.
- Exhibit 5 records where launch sites initiate with the characters 'CCA.'
- Display the total payload weight transported by NASA (CRS) launched boosters.
- Show the average payload weight carried by booster version F9 v1.1.
- Enumerate the date when the first successful landing was accomplished on a ground pad.
- Enlist the names of boosters that succeeded on a drone ship and carried a payload mass ranging between 4000 and 6000 kg.
- Provide a list of the total number of successful and unsuccessful mission outcomes.
- Display the names of booster versions that have carried the maximum payload mass.
- Enumerate the landing outcomes that failed on drone ships, their booster versions, and the launch site names for 2015.
- Arrange the count of landing outcomes such as Failure (drone ship) or Success (ground pad) between the dates 2010-06-04 and 2017-03-20 in descending order.

See notebook on Github

BUILD AN INTERACTIVE MAP WITH FOLIUM

To visualize the data on an interactive map:

1. Mark all launch sites on a map
 • Use folium.Map object to initialise a map
 • Use folium.Circle and folium.Marker objects to mark launch sites on the map

2. Put markers for the successful and failed launches on the map
 • Marker colour of successful launch (class = 1) is set to green, and failed launch (class = 0) to red.
 • Create an icon as a text label, assigning the icon_color as the marker_colour determined previously.

3. Calculate the distances between a launch site nearby points of interest
 • Each point of interest's distance from a launch site is caluculated using the point's latitude and longitude
 • Each point's location was marked on the map using a folium.Marker object
 • A folium.PolyLine was drawn on the map to display the distance line between the launch site and point of interest

See code on Github

BUILD A DASHBOARD WITH PLOTLY DASH

The following plots were added to a Plotly Dash dashboard to have an interactive visualisation of the data:

1. Pie chart (px.pie()) showing the total successful launches per site
 • This makes it clear to see which sites are most successful
 • The chart could also be filtered (using a dcc.Dropdown() object) to see the success/failure ratio for an individual site

2. Scatter graph (px.scatter()) to show the correlation between outcome (success or not) and payload mass (kg)
 • This could be filtered (using a RangeSlider() object) by ranges of payload masses
 • It could also be filtered by booster version

See notebook on Github

PREDICTIVE ANALYSIS (CLASSIFICATION)

The following steps were taking to develop, evaluate, and find the best performing classification model:

1. Model development
- Load dataset
- Perform necessary data transformations (standardise and pre-process)
- Split data into training and test data sets, using train_test_split()
- Decide which type of machine learning algorithms are most appropriate

2. Model evaluation
For each chosen algorithm, check the tuned hyperparameters and accuracy (score and best_score_) using the output GridSearchCV object

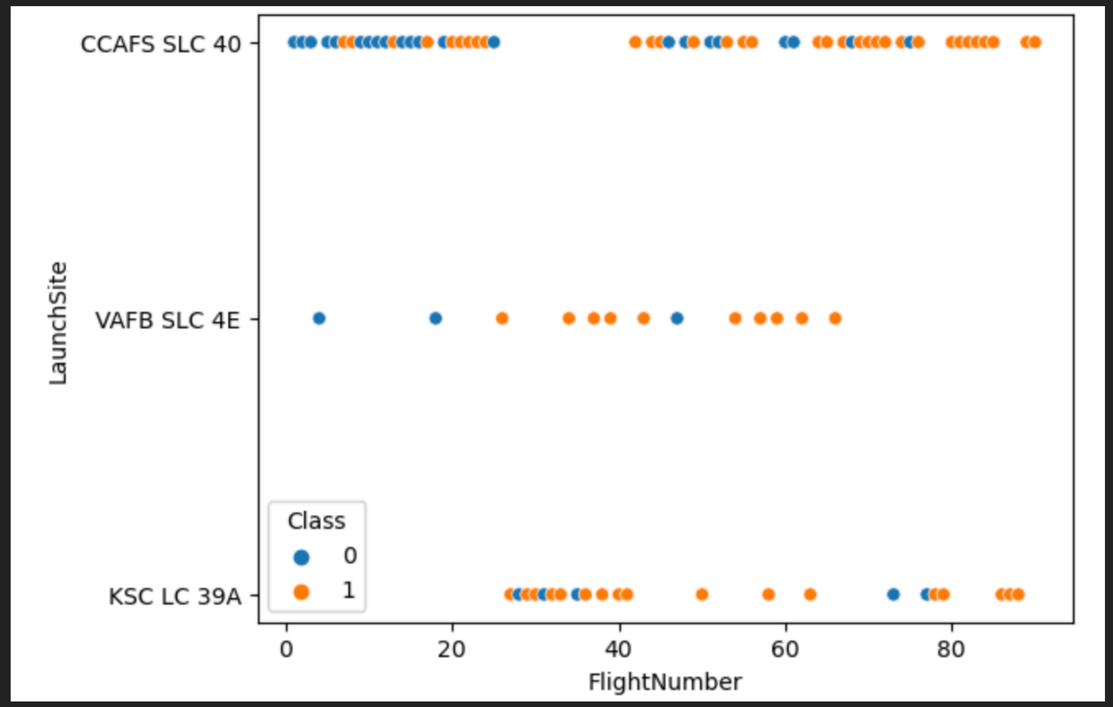
3. Finding the best classification model
- Review the accuracy scores for all chosen algorithms
- The model with the highest accuracy score is determined as the best performing model

SECTION 4

INSIGHTS DRAWN FROM EDA

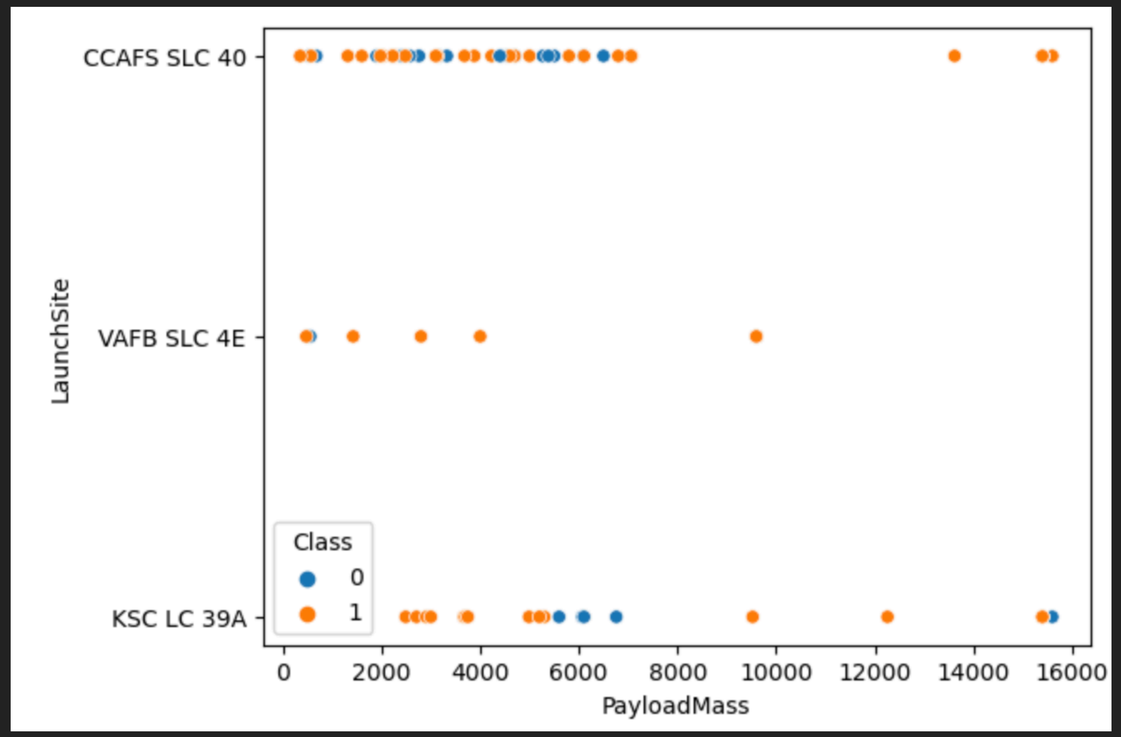
LAUNCH SITE V.S. FLIGHT NUMBER

- The scatter plot of Launch Site vs. Flight Number shows that:
  - The rate of success at a launch site is positively correlated to the number of flights from said launch site
  - CCAFS SLC 40 and VAFB SLC 4E is where most early (and unsuccessful) flights have been launched from
  - No early flights were launched from KSC LC 39A, so the launches from this site are more successful.
  - There's a significant uptick of successful landings (Class = 1) after more than 30 launches



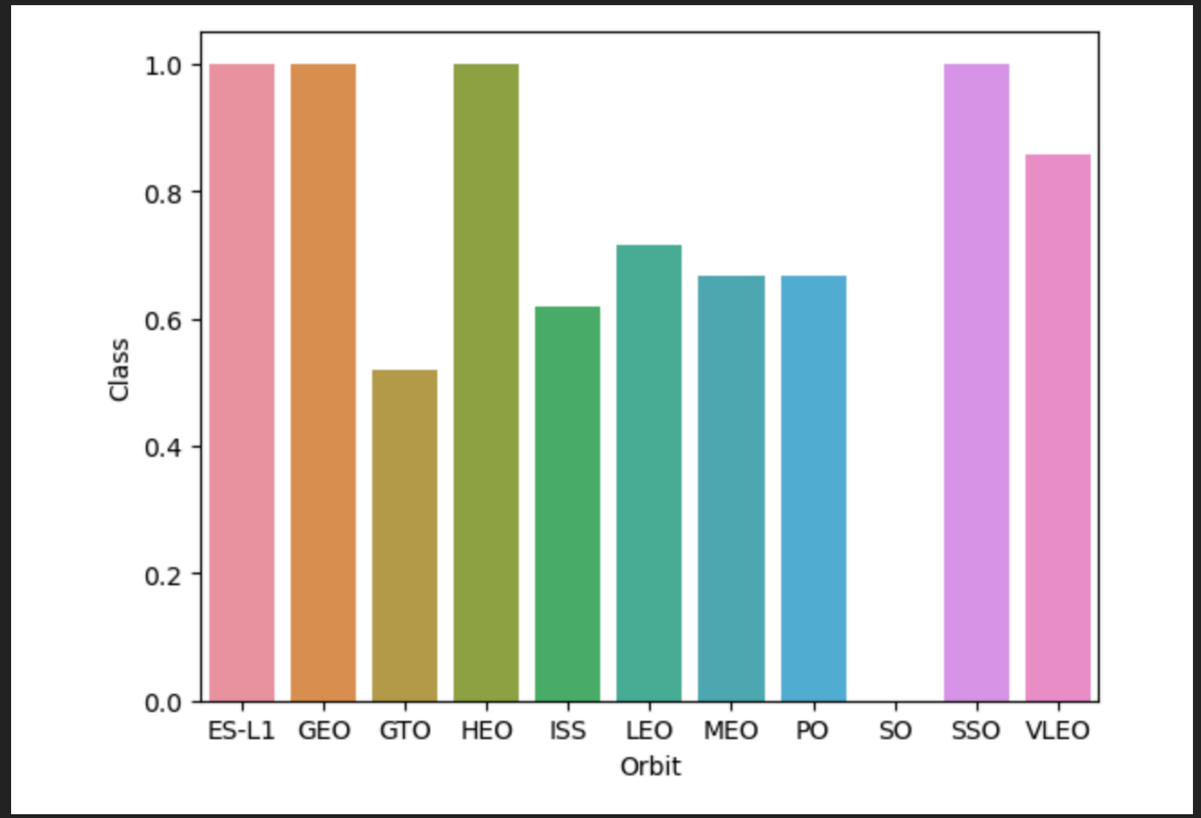
LAUNCH SITE V.S. PAYLOAD MASS

- The scatter plot of Launch Site vs. Payload mass shows that:
  - Successful landings are more common for payloads exceeding 7000 kg, however, there is a limited amount of information regarding these heavier launches.
  - There is no distinct relationship between the payload mass and success rate for launches from a specific site.
  - The majority of launches from CCAFS SLC 40 involved lighter payloads, although there were some instances with heavier payloads.



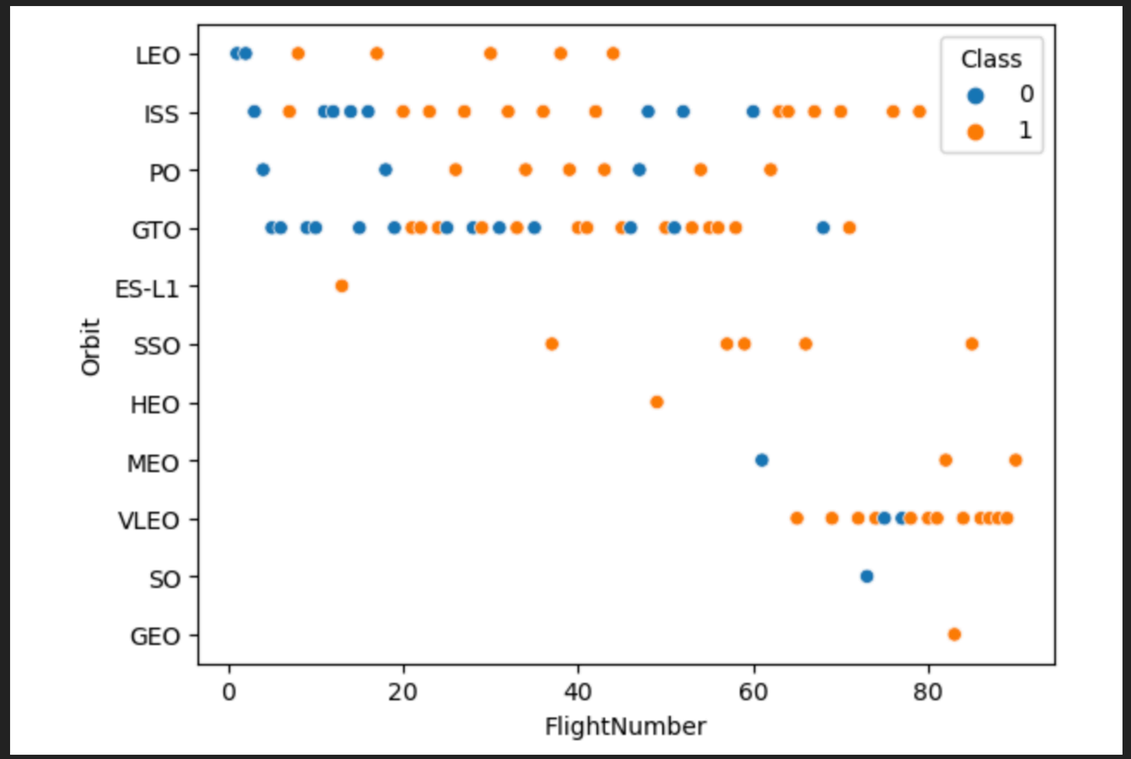
SUCCESS RATE V.S. ORBIT TYPE

- The bar plot of Success Rate v.s. Orbit Type shows that the following orbits have the highest (100%) success rate:
  - ES-L1 (Earth-Sun First Lagrangian Point)
  - GEO (Geostationary Orbit)
  - HEO (High Earth Orbit)
  - SSO (Sun-synchronous Orbit)
- The orbit with the lowest (0%) success rate is:
  - SO (Heliocentric Orbit)



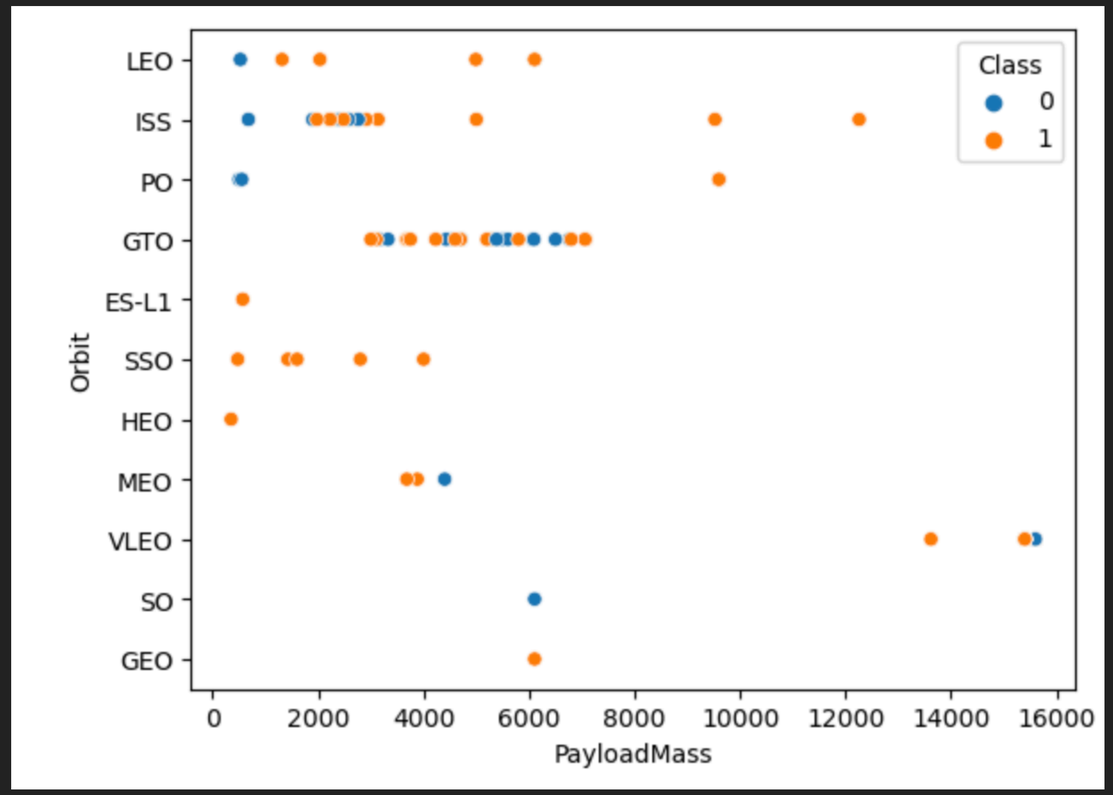
FLIGHT NUMBER V.S. ORBIT TYPE

- This scatter plot depicts the relationship between Orbit Type and Flight number and provides new insights that were not evident in previous plots. For instance:
  - The plot indicates that the 100% success rates for GEO, HEO, and ES-L1 orbits can be attributed to the fact that there was only one flight into each respective orbit.
  - Furthermore, the 100% success rate for SSO is even more remarkable, given that there were five successful flights.
  - The data also shows that there is little correlation between Flight Number and Success Rate for GTO.
  - Generally, as Flight Number increases, the success rate tends to increase as well. This trend is particularly evident for LEO, where unsuccessful landings occurred only during the early launches (low flight numbers).



PAYLOAD V.S. ORBIT TYPE

- The Orbit Type vs. Payload Mass scatter plot shows:
  - PO, ISS and LEO orbit types have more success with heavy payloads.
  - There’s an unclear relationship between payload mass and success rate in the GTO orbit
  - VLEO (Very Low Earth Orbit) launches are associated with heavier payloads, intuitively this checks out



ALL LAUNCH SITE NAMES

- The query used to find all site names:
  - select distinct(launch_site) from landings
- It selects the launch_site column from all rows in the landings table, but it filters them to return only distinct (unique) values

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

LAUNCH SITE NAMES BEGIN WITH 'CCA'

- The query used to find all site names:
  - select * from landings where launch_site LIKE 'CCA%' limit 5
- It selects all columns from all rows in the landings table, where the launch_site column begins with the ‘CCA’ letters.
- It returns only five rows, as required in the instructions, by using the LIMIT statement.

Out[9]:

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

- The query used to find the total payload mass:
  - select sum(payload_mass__kg_) as total_payload_mass from landings where customer = 'NASA (CRS)'
- The SUM keyword is used to calculate the total of the LAUNCH column, and the SUM keyword (and the associated condition) filters the results to only boosters from NASA (CRS)

total_payload_mass
45596

AVERAGE PAYLOAD MASS BY F9 V1.1

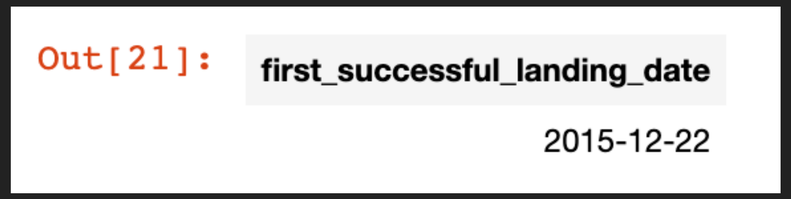
- The query used to find the average payload mass by Falcon9 v1.1:
  - select avg(payload_mass__kg_) as average_payload_mass from landings where booster_version LIKE 'F9 v1.1%'
- The AVG keyword is used to calculate the average of the PAYLOAD_MASS__KG_ column, and the WHERE keyword (and the associated condition) filters the results to only the F9 v1.1 booster version

Out[15]:

average_payload_mass
2534

FIRST SUCCESSFUL GROUND LANDING DATE

- To find the the dates of the first successful landing outcome on ground pad, we use the MIN keyword to calculate the minimum date, in combination with the WHERE clause to filter the result with only the successful ground pad landing:
  - SELECT MIN(date) AS first_successful_landing_date
    FROM landings
    WHERE landing__outcome = 'Success (ground pad)'



SUCCESSFUL DRONE SHIP LANDING WITH PAYLOAD BETWEEN 4000 AND 6000

- To list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000, we use the WHERE clause to filter the results to include only those that satisfy both conditions in the brackets (as the AND keyword is also used). The BETWEEN keyword allows for 4000 < x < 6000 values to be selected.

  - SELECT DISTINCT(booster_version) AS booster_name
    FROM landings
    WHERE landing__outcome = 'Success (drone ship)' AND
    payload_mass__kg_ BETWEEN 4000 AND 6000

TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSION OUTCOMES

- To calculate the total number of successful and failure mission outcome, we used the COUNT keyword to calculate the total number of mission outcomes, and the GROUPBY keyword to group these results by the type of mission outcome:
  - SELECT COUNT(*) AS count, mission_outcome
    FROM landings
    GROUP BY mission_outcome

Out[26]:

COUNT	mission_outcome
1	Failure (in flight)
99	Success
1	Success (payload status unclear)

BOOSTERS CARRIED MAXIMUM PAYLOAD

- To list the names of the booster which have carried the maximum payload mass, we use a subquery in combination with the DISTINCT, WHERE and MAX keywords. The MAX keyword is used in the subquery which returns the maximum payload mass, which is then used in the WHERE clause to filter the rows and find the relevant booster version:

  - SELECT DISTINCT(booster_version)
    FROM landings
    WHERE payload_mass__kg_ = (
      SELECT MAX(payload_mass__kg_) FROM landings
)

Out[29]:

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 LAUNCH RECORDS

- To list the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015, we use a combination of WHERE keyword and the YEAR function which extracts the year from a DATE column:
  - SELECT booster_version, launch_site, landing__outcome
    FROM landings
    WHERE landing__outcome = 'Failure (drone ship)' AND YEAR(date) = 2015

Out[38]:

booster_version	launch_site	landing__outcome
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

RANK LANDING OUTCOMES BETWEEN 2010-06-04 AND 2017-03-20

- To rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order, we use a combination of WHERE keyword to filter rows, BETWEEN keyword to select rows where a column is between between two dates, and then ORDER BY to order the result set and a GROUP BY to group the result rows by the landing type:

  - SELECT COUNT(*) AS count, landing__outcome
    FROM landings
    WHERE date BETWEEN '2010-06-04' AND '2017-03-20'
    GROUP BY landing__outcome
    ORDER BY count desc

Out[40]:

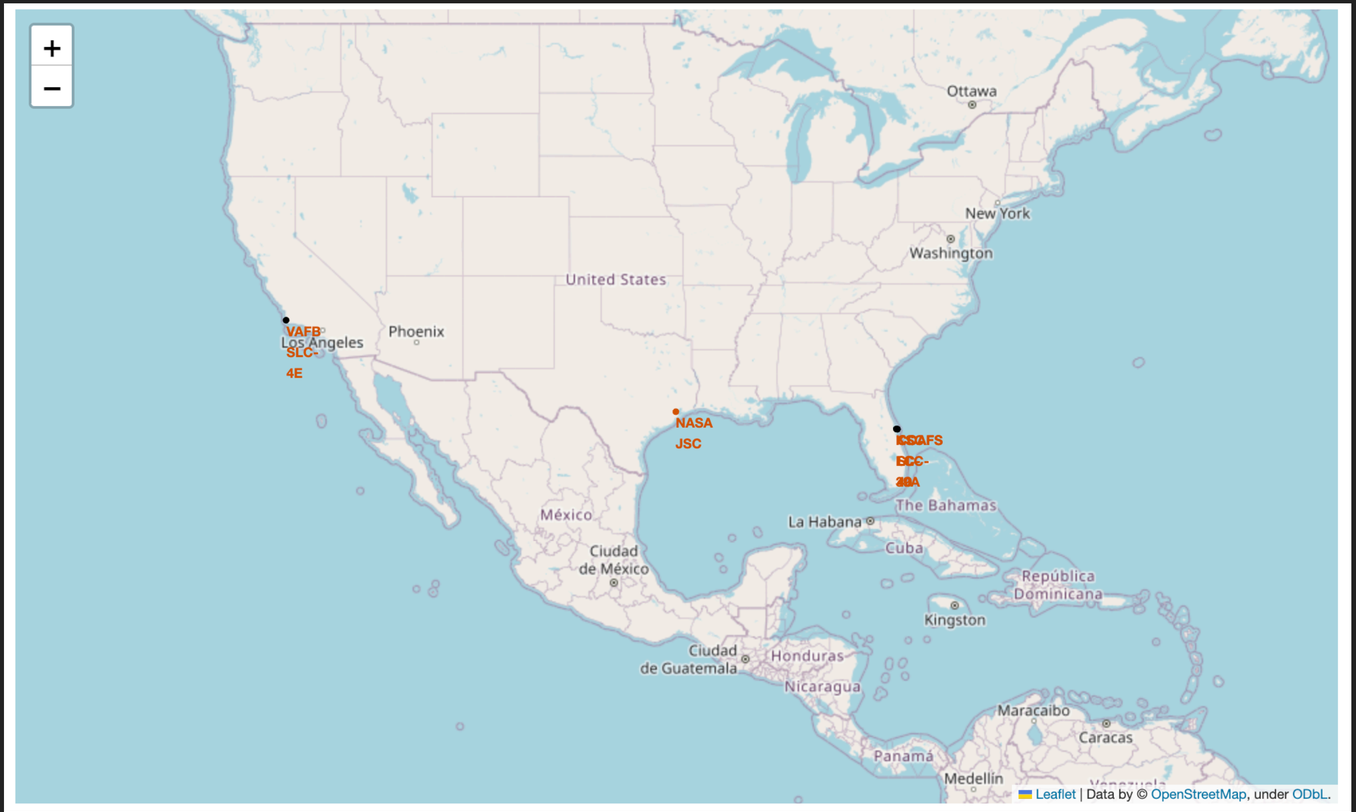
COUNT	landing__outcome
10	No attempt
5	Failure (drone ship)
5	Success (drone ship)
3	Controlled (ocean)
3	Success (ground pad)
2	Failure (parachute)
2	Uncontrolled (ocean)
1	Precluded (drone ship)

SECTION 5

LAUNCH SITES PROXIMITIES ANALYSIS

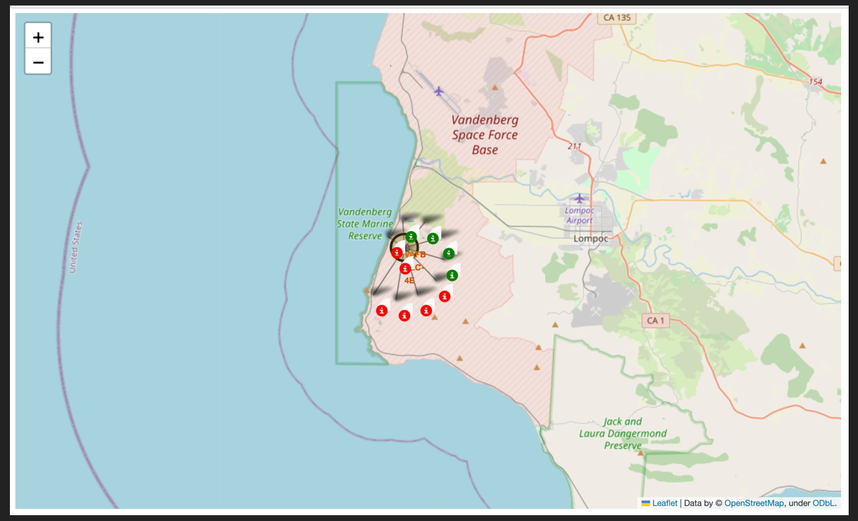
Photo by SpaceX on Unsplash

ALL LAUNCH SITES

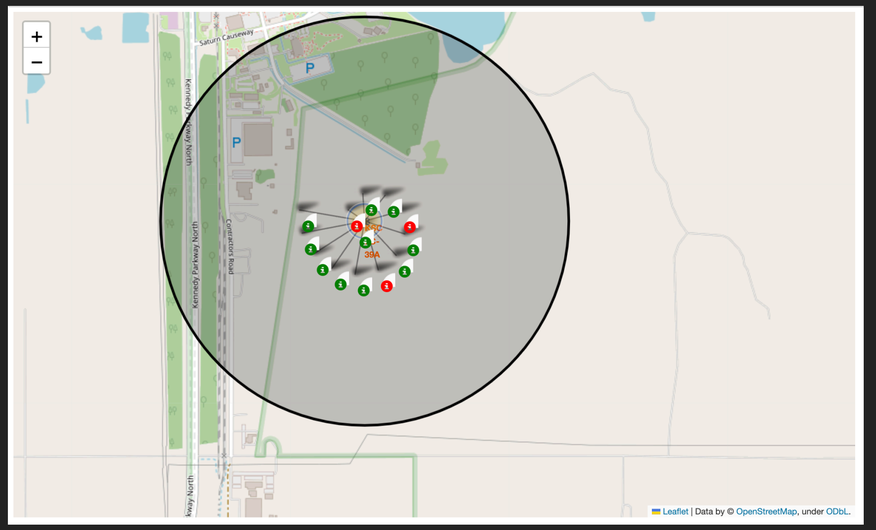


- All SpaceX launch sites are on coasts of the United States of America, specifically Florida and California.

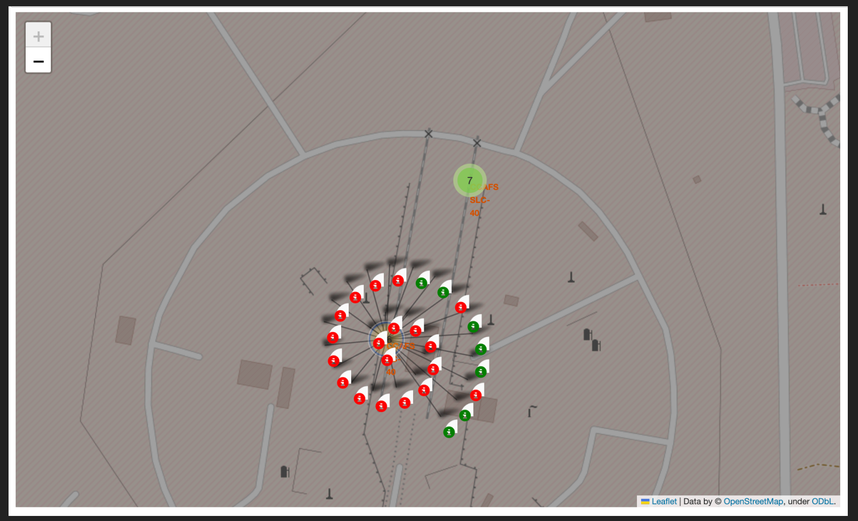
SUCCESS/FAILED LAUNCHES FOR EACH SITE



KSC LC-39A

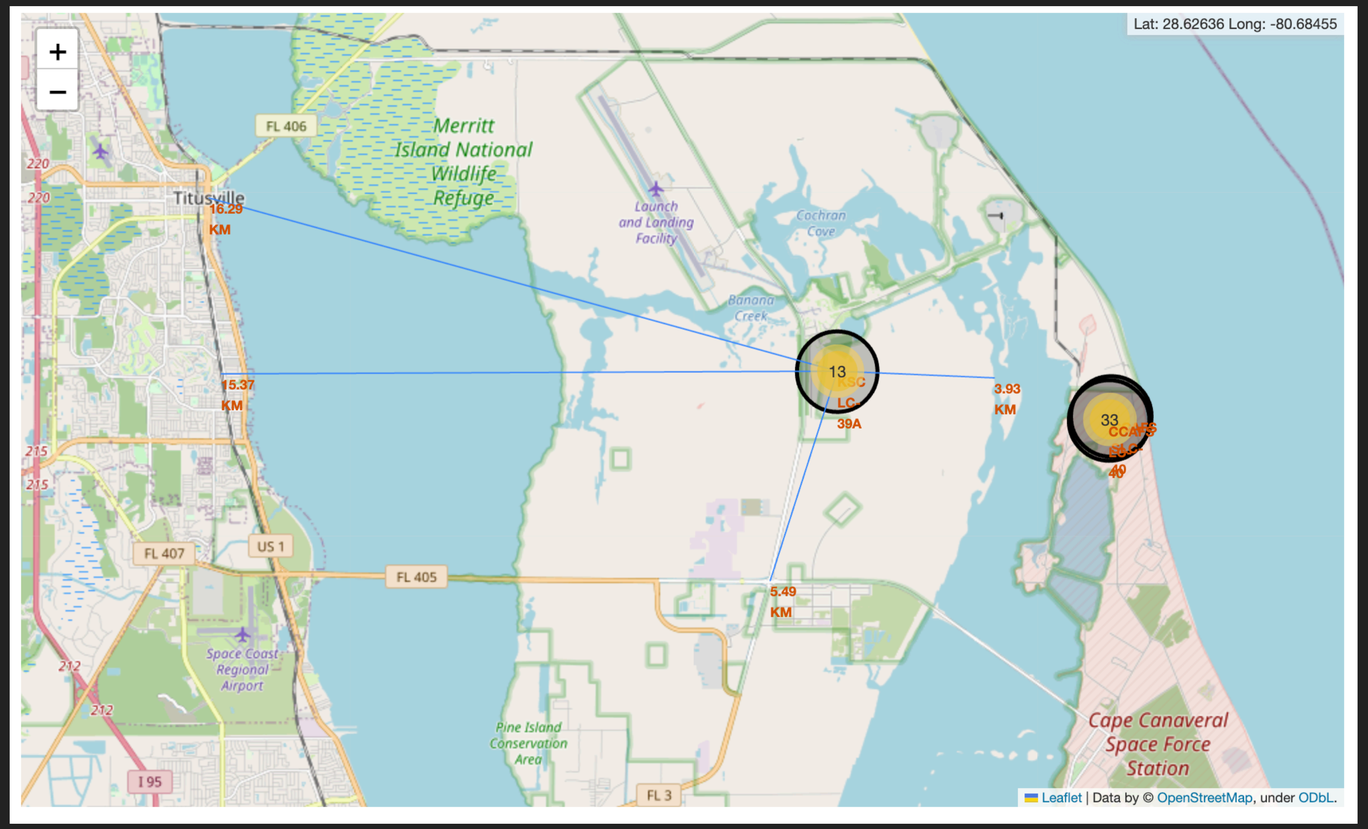


VAFB SLC-4E



CCAFS SLC-40

DISTANCES BETWEEN A LAUNCH SITE TO ITS PROXIMITIES



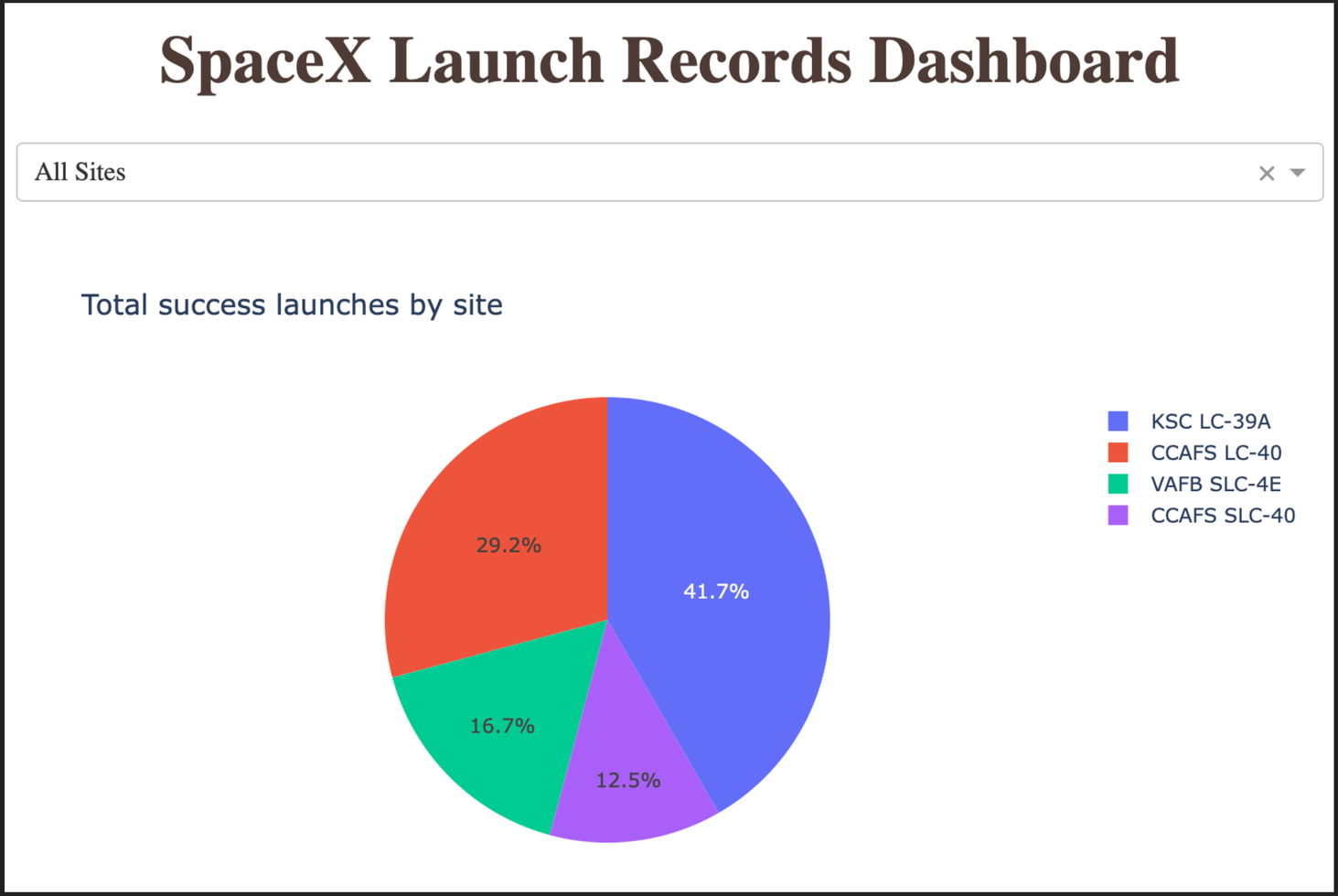
- Site name: KSC LC-39A
- Distance to motorway: 5.49KM
- Distance to railway: 15.37KM
- Distance to city: 16.29KM
- Distance to coast: 3.93KM

SECTION 6

BUILD A DASHBOARD WITH PLOTLY DASH

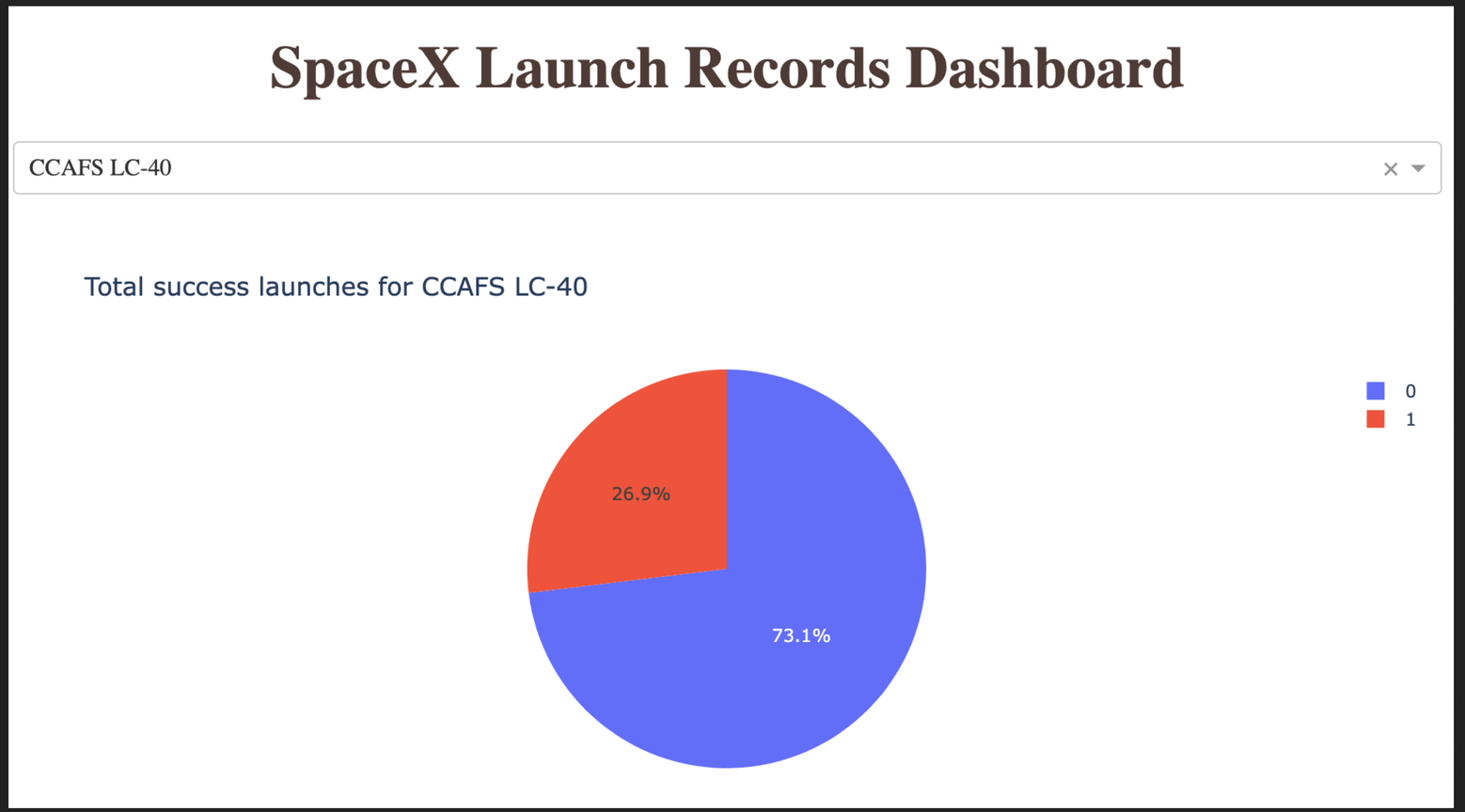
Photo by Leonel Fernandez on Unsplash

LAUNCH SUCCESS COUNT FOR ALL SITES



- The launch site with the most successful launches is **KSC LC-39** - a staggering 41.7% of the total launches were successful.

LAUNCH SUCCESS RATIOS FOR LAUNCH SITE WITH LOWEST LAUNCH SUCCESS



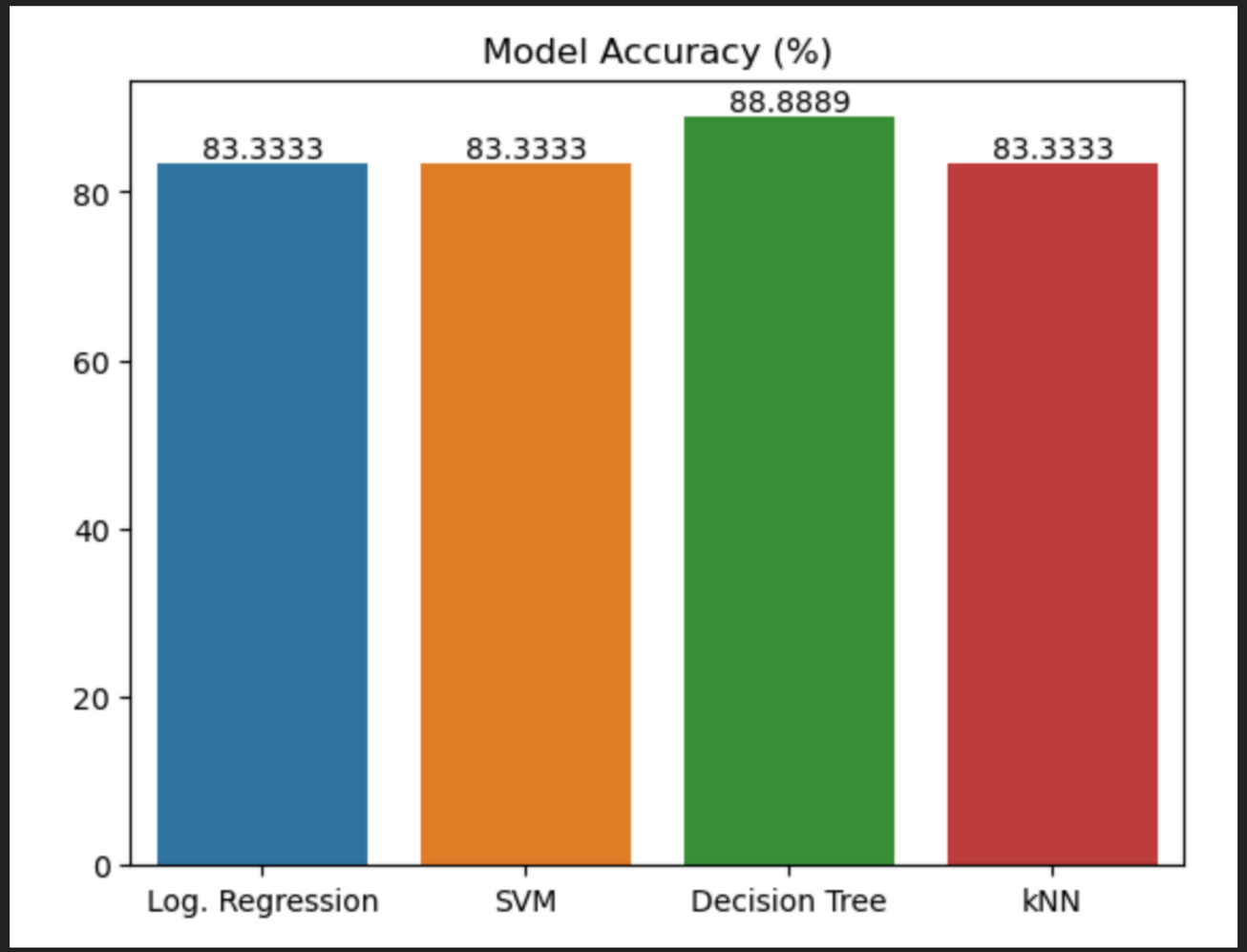
- **CCAFS LC-40** had the lowest success rate at 26.9% success.

SECTION 7

PREDICTIVE ANALYSIS (CLASSIFICATION)

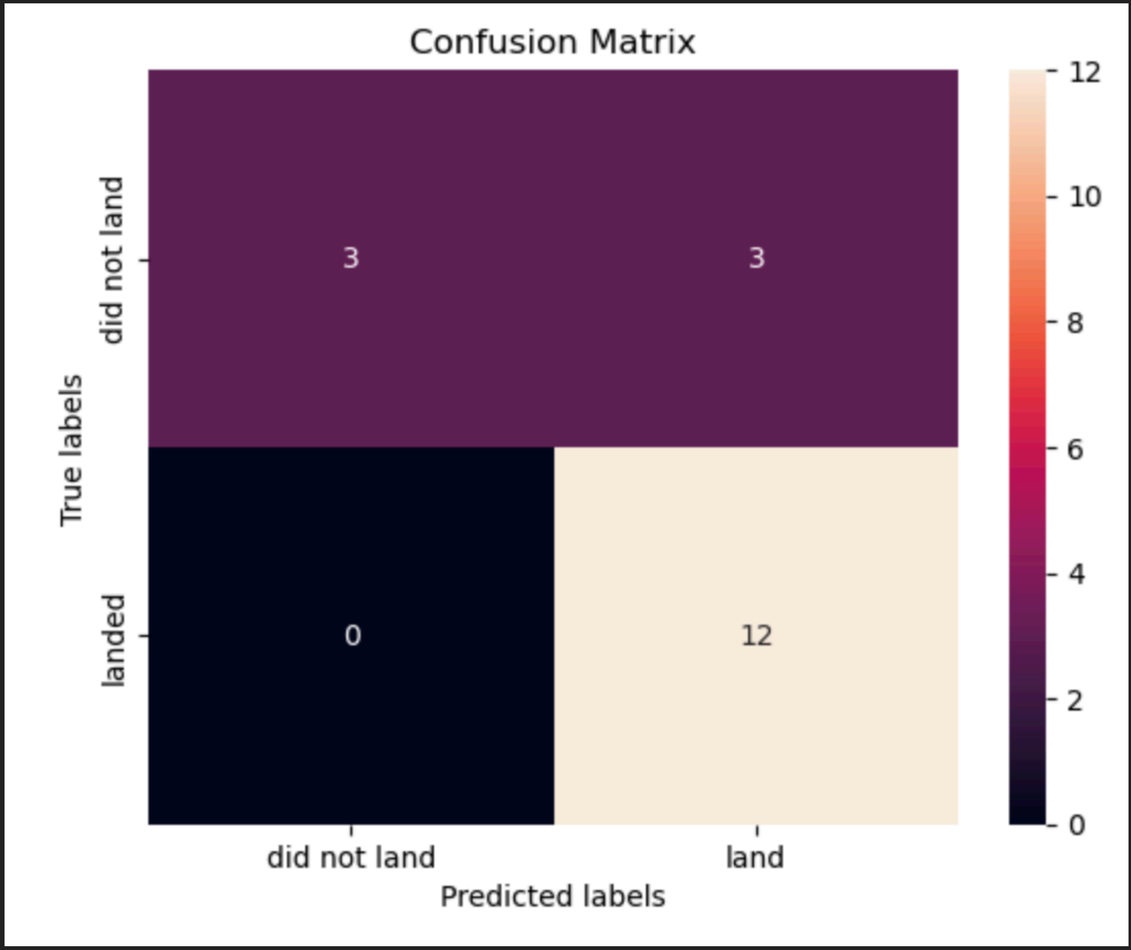
Photo by Markus Winkler on Unsplash

CLASSIFICATION ACCURACY



The most accurate model is **Decision Tree,** with accuracy of **~88.9%**.

CONFUSION MATRIX



- The best performing classification model is the Decision Tree model, with an accuracy of 88.89%.
- This is explained by the confusion matrix, which shows only 3 out of 18 total results were classified incorrectly (a false positive, see the top-right corner).
- The other 15 results are correctly classified (3 did not land, 12 did land).

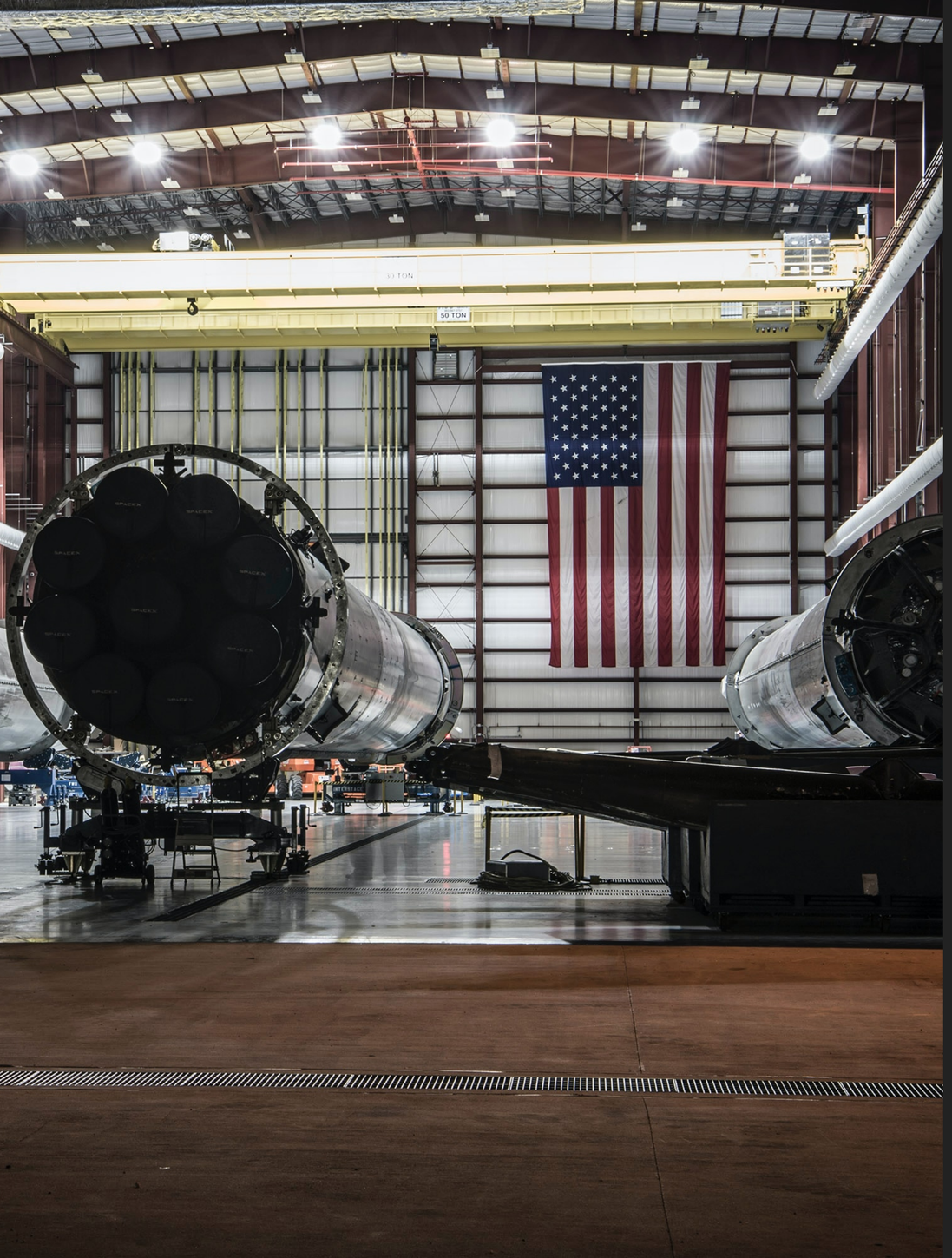
SECTION 8

CONCLUSIONS

Photo by SpaceX on Unsplash

CONCLUSIONS

- The data indicates that as a launch site accumulates more flight experience, the success rate improves, as evident from the increasing success rate from 2013, despite minor setbacks in 2018 and 2020.
- Between 2010 and 2013, all landings failed, but after 2016, there was consistently over a 50% chance of success.
- Orbit types ES-L1, GEO, HEO, and SSO had a 100% success rate; however, GEO, HEO, and ES-L1 each only had one flight. SSO stands out with 5 successful flights. The orbit types PO, ISS, and LEO had more success with heavier payloads. VLEO launches, which are associated with heavier payloads, intuitively make sense.
- The launch site KSC LC-39 A had both the highest number (41.7%) and rate (76.9%) of successful launches. Success with payloads over 4000kg was lower compared to lighter payloads.
- The Decision Tree model had the highest accuracy at 94.44%.



SECTION 9

APPENDIX

Photo by SpaceX on Unsplash

ADDITIONAL LINKS & SCRIPTS

- The whole capstone project, including this presentation, is available on Github.