IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Elizabeth Gels
April 23, 2024

Outline

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

Executive Summary

- SpaceX has achieved historic milestones with Falcon 9. By reusing its first-stage boosters they've been able to greatly reduce the cost of a launch. Our startup has been tasked with predicting the cost of a rocket launch so other companies can make more informed bids against SpaceX.
- The SpaceX launch data was collected from The SpaceX REST API and by web scraping.
- During data wrangling we dealt with missing values and performed preliminary exploration to gain a better understanding of the data.
- In Exploratory data analysis (EDA) we built graphs and used SQL to gain valuable insights. We identified the dependent and independent variables for our model.
- We developed Interactive maps and dashboards using Folium and Plotly Dash.
- We perform predictive analysis using the following classification models: Logistic Regression, Support Vector Machine, Decision Trees and K-Nearest Neighbors.
- All four models had the same classification accuracy at 0.83; and the same confusion matrix. Although the model can be considered good at predicting, there might be room for improvement.

Introduction

- SpaceX is a private aerospace manufacturer and space transportation company founded in 2002. Its goal is to revolutionize space technology and enable human colonization of other planets. SpaceX has achieved several milestones, including launching and recovering spacecraft, sending astronauts to the International Space Station, and developing liquid-propellant rockets that reach orbit. (SpaceX – Wikipedia)
- SpaceX It is the only private company ever to return a spacecraft from low-earth orbit, which it was first accomplished in December 2010. SpaceX advertises Falcon 9 rocket launches on its website with a cost of $62 million whereas other providers cost upward of $165 million each. SpaceX can save millions of dollars by reusing the first stage of its rockets, if they are able to bring them back and have them land successfully.
- If we develop a model that can predict if the first stage of a rocket will land successfully then we could determine the cost of future launches. Then, competitors could make more informed decisions when bidding against SpaceX.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The SpaceX launch data was collected from The SpaceX REST API and by scraping the Wikipedia page: “List of Falcon 9 and Falcon Heavy launches”.
- Perform data wrangling:
  - Data wrangling included data cleansing (for missing values) and preliminary exploration to gain a better understanding of the data.
- Perform exploratory data analysis (EDA) using visualization and SQL.
- Perform interactive visual analytics using Folium and Plotly Dash.
- Perform predictive analysis using classification models.
  - Four classification models were built to predict if the first stage of the Falcon 9 would land successfully. Trained the data using Grid Search. Identified the model with the best classification accuracy.

Data Collection

- The SpaceX launch data was gathered from two sources:
  - The SpaceX REST API.
    - We used the Python requests library to perform a get request. This request obtained the data from the API in the form of a JSON. We converted the JSON data to a Pandas dataframe to see it as a flat table.
  - The Wikipedia page: “List of Falcon 9 and Falcon Heavy launches”.
    - We used the Python BeautifulSoup package to web scrape HTML tables containing Falcon 9 launch records. Then we parsed the data and convert it into a Pandas dataframe.

Data Collection - SpaceX API

SpaceX REST API

Endpoints (URLs):

https://api.spacexdata.com/v4/launches/past

https://api.spacexdata.com/v4/rockets

https://api.spacexdata.com/v4/launchpads

https://api.spacexdata.com/v4/payloads

https://api.spacexdata.com/v4/cores

Obtain data with a get request:
response=request.get(URL)

Successfully receive a .json response:
response.status_code = 200

Turn the content into a Pandas dataframe:
data=pd.json_normalize(response.json())

Filter dataframe to only include Falcon 9 launches.

Notebook GitHub URL:
https://github.com/elgels/IBM-Data-Science-Capstone/blob/b88c1a44dde3cdb7d0605a6bbff428eecd8d4f82/Collecting%20the%20Data%20API-api.ipynb

Data Collection - Scraping

Wikipedia page:

“List of Falcon 9 and Falcon Heavy launches”

Static URL

https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

Obtain data with a get request:
response = requests.get(Static_URL).text

Use the Python BeautifulSoup package to create an object from the HTML response:
soup = BeautifulSoup(response, 'html.parser’)

Find the HTML table with Falcon 9 records:
html_tables = soup.find_all(‘table’)

Iterate through the <th> elements of the table and parse the raw data to convert it into a Pandas dataframe.

Notebook GitHub URL:
https://github.com/elgels/IBM-Data-Science-Capstone/blob/17f3530a5a715101dbb1f5309a8ada631a619b5e/Webscraping%20Records%20from%20Wikipedia.ipynb

Data Wrangling

- In data wrangling, we followed this process:

Identify Missing Values. → Count observations (rows and columns). → Identify variables type (String, Numeric, Boolean). → Perform value counts to get familiar with the data. → Create new column “Class” to represent success or failure.

  - The Landing Pad column had 26 nulls (29% of total). It retained the None values to represent when landing pads were not used.
  - The dataframe consisted of 90 rows and 17 columns.
  - Date ranged from June 2010 through November 2020.
  - 7 variables were String type, 7 were Numeric and 3 were Boolean.
  - Value counts of variables Landing Site, Orbit and Outcome provided more insights into the data.
  - Created a new column “Class” to represent landing success (value of 1) or failure (value of 0).

- Notebook GitHub URL
  https://github.com/elgels/IBM-Data-Science-Capstone/blob/f06c6c75f64a4f498d8010f90722a225de894f10/Data%20Wrangling.ipynb

Data Wrangling - Insights

- Number of launches by Launch Site:

CCAFS SLC (Cape Canaveral)	KSC LC 39A (Kennedy Space Center)	VAFB SLC 4E (Vandenberg)
55	22	13

- Number and occurrence of each Orbit:

GTO	ISS	VLEO	PO	LEO	Other
27	21	14	9	7	12

- Outcome to determine the number of landing outcomes:

True ASDS (Successfully landed on a drone ship).	None (Failure to Land)	True RTLS (Successfully landed to a ground pad).	False ASDS (Unsuccessfully landed on a drone ship).	True Ocean (Successfully landed in the ocean).	Other
41	19	14	6	5	5

- The new column Class used column Outcome as a base to determine success or failure.

Data Wrangling - Variables in Dataframe

String	Description
Date	YYYY-MM-DD
Booster Version	Versions of the Falcon 9 boosters such as F9 v1.1 and F9 v1.1 B1012.
Orbit	Such as LEO (low orbit at an altitude of 2,000 km from Earth), VLEO (very low orbit at below 450 km), GTO (high Earth orbit at about 36,000 km).
Launch Site	Cape Canaveral (CCAFS SLC), Kennedy Space Center (KSC LC 39A) and Vandenberg (VAFB SLC 4E).
Outcome	Codes for successful and failed landings (in ocean, ground and drone ship).
Landing Pad	5 Ids of Landing Pads.
Serial	Serial Numbers such as B1049, B1051 and B1060.

Numeric	Description
Flight Number	Numbers from 1 through 90.
Payload Mass	Weight (Kg.) of cargo or equipment that the rocket is designed to carry into space.
Flights	Values from 1 through 6.
Block	Incremental improvements made to previous versions. From 1.0 through 5.0.
Reused Count	Values from 0 through 5.
Longitude	Longitude of the launch site.
Latitude	Latitude of the launch site.
Class	0 (if launch failed), 1 (if launch succeeded).

Boolean	Description
Grid Fins	True or False.
Reused	True or False.
Legs	True or False.

EDA with Data Visualization

- The following charts were plotted:

Scatter Plot
Flight Number vs Launch Site
Payload vs. Launch Site
Flight Number vs. Orbit Type

Bar Chart
Success Rate vs. Orbit Type

Line Chart
Yearly Average Success Rate

- We considered if the data can be used to predict a successful landing. We observed the launch sites had different success rates, even when combined with different levels of payload mass. Also, different orbits had different success rates, even when combined with different numbers of payload mass. We observed that the success rate has improved since 2013.

- Therefore, attributes such as launch sites, orbits, payloads and flight number can be used to help determine if the first stage will land successfully.

- We applied one hot encoding to categorical variables to prepare the data for the machine learning model.
- Notebook GitHub URL:
  https://github.com/elgels/IBM-Data-Science-Capstone/blob/a9fcb331e96890f06aeddcc399458acf4ec424d7/Data%20Visualization.ipynb

EDA with SQL

- We used SQL queries to find the following:
  - Display the names of the unique launch sites in the space mission.
  - Find 5 records where launch sites begin with ‘CCA’.
  - Calculate the total payload carried by boosters from NASA.
  - Display average payload mass carried by booster version F9 v1.1.
  - List the date when the first successful landing outcome in ground pad was achieved.
  - List the names of the boosters which had success in drone ship landing and had payload mass greater than 4,000 but less than 6,000.
  - Calculate the total number of successful and failure mission outcomes.
  - List the names of the boosters which have carried the maximum payload mass.
  - List the launch records which had a failed landing outcome (in drone ship) in 2015. Include the corresponding booster versions and launch sites.
  - Rank the count of landing outcomes, including Failure (drone ship) or Success (ground pad) between the dates 2010-06-04 and 2017-03-20, in descending order.

- Notebook GitHub URL:

  https://github.com/elgels/IBM-Data-Science-Capstone/blob/a63f676adc6ae45e4b0f431027a946582718523d/EDA%20using%20SQL.ipynb

Build an Interactive Map with Folium

- Instead of presenting static graphs, we used Folium (a Python map plotting library) to create interactive maps with features such as zoom-in, zoom-out, and the ability to easily find any coordinates on the map.
- We added circles, color-labeled markers, marker clusters, lines and icons to make it easy to identify the launch sites with the highest success rates.
- We analyze launch site geographical locations and their proximities to railways, highways, coastline and cities.
- Notebook GitHub URL:

  https://github.com/elgels/IBM-Data-Science-Capstone/blob/a9fcb331e96890f06aeddcc399458acf4ec424d7/Launch%20Sites%20Locations%20Analysis%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- We built a dashboard application with the Python Plotly Dash package. This package allowed us to add interactive features such as a drop-down list and a range slider.
- Interactive dashboards are more effective at finding insights from the data than with static graphs.
- The dashboard included a pie chart showing the share of total successful launches for all sites. Using the drop-down we could select a specific launch site to observe its detailed information.
- We also included a scatter plot of payload vs. outcome for all sites. Using the payload range slider we could select different payload ranges to observe the success rates. An added variable was booster versions which were represented in different colors.
- Notebook GitHub URL:

  https://github.com/elgels/IBM-Data-Science-Capstone/blob/a9fcb331e96890f06aeddcc399458acf4ec424d7/Build%20a%20Dashboard%20Application%20with%20Plotly%20Dash.py

Predictive Analysis (Classification)

- We aimed to predict if the first stage of the Falcon 9 would land successfully. We had the variable "Class" (1 if launch was successful, 0 if launch failed) as the dependent variable. The other features were the independent variables, except for Date, Outcome, Booster Version, Longitude and Latitude, which were excluded from the dataset.

- Before developing the models, we applied hot encoding to categorical variables: Orbit, Launch Site, Landing Pad, Serial, Grid Fins, Reused, and Legs. Then to find the best performing classification model, we followed this process:

Standardize the data. → Split the data into training and test data. → Initialize the model, perform Grid Search to find the best hyperparameters. → With the best hyperparameter values, determine the model with the best accuracy using the training data. → Find the model that performs best using the test data. Output the confusion matrix.

- Notebook GitHub URL:

  https://github.com/elgels/IBM-Data-Science-Capstone/blob/a9fcb331e96890f06aeddcc399458acf4ec424d7/Machine%20Learning%20Prediction.ipynb

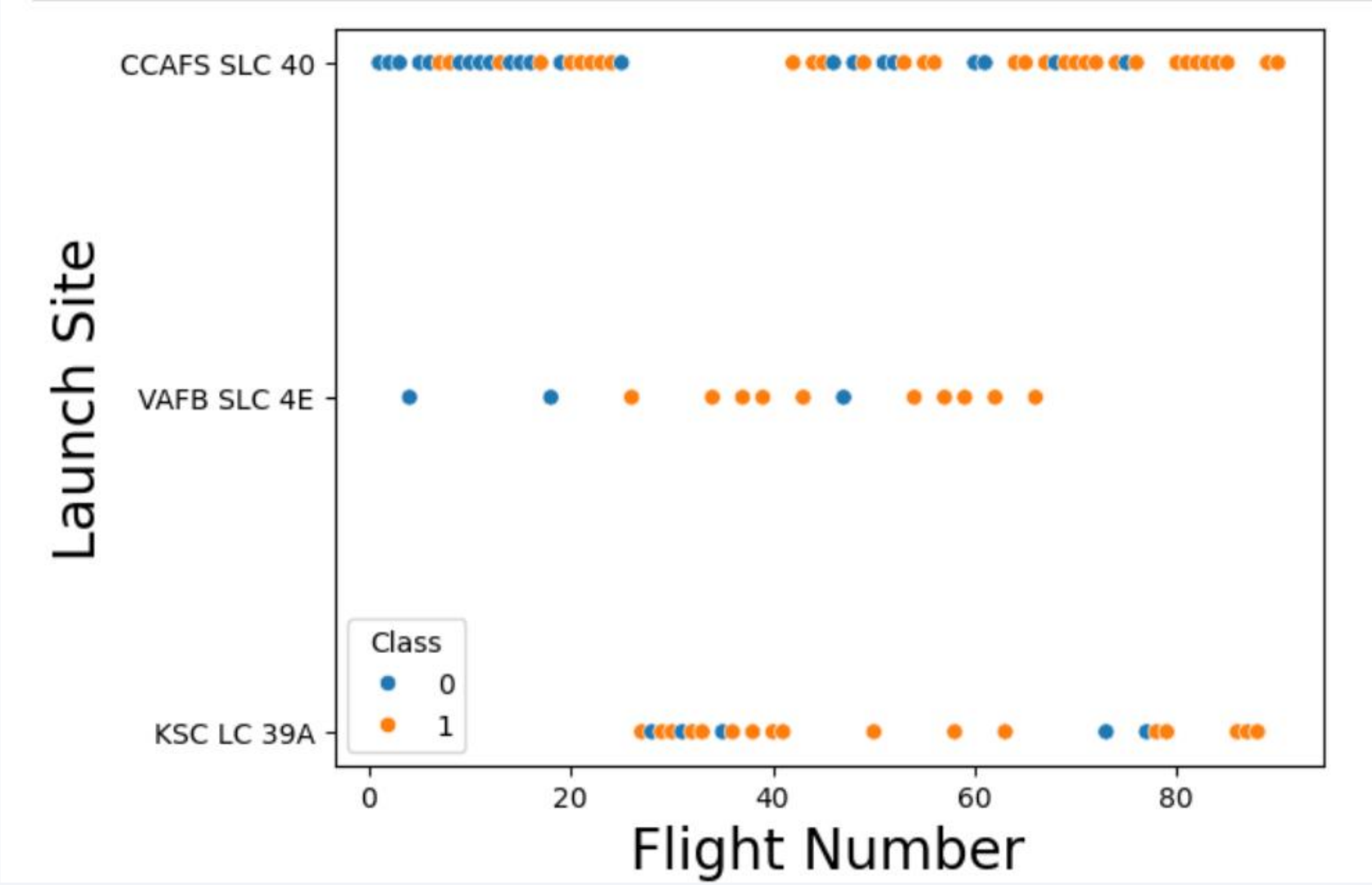
Results

- During exploratory data analysis we determined that the data we collected could be used to predict if the first stage of the rocket would land successfully. We identified important variables such as launch sites, orbits, payloads, flight number and booster versions.
- We developed maps (using Folium) which clearly showed that the Kennedy Space Center was the launch site with the highest success rate. Then we built an interactive dashboard (using Plotty Dash) that confirmed that the Kennedy Space Center had by far the largest success rates among all sites.
- We learned that launch sites are always close to railways and highways (for efficient transportation); and coastlines (for safety). Also, they keep a good distance from cities to prevent any debris falling into a populated area.
- When combining all sites, the payload range with the highest success rate was between 2,000 and 6,000 kg. At that range, the booster version category with the largest success rate was the FT.
- We developed and tested four predictive models: Logistic Regression, Support Vector Machine, Decision Trees and K-Nearest Neighbors. All four models had the same classification accuracy at 0.83.

Section 2

Insights drawn from EDA

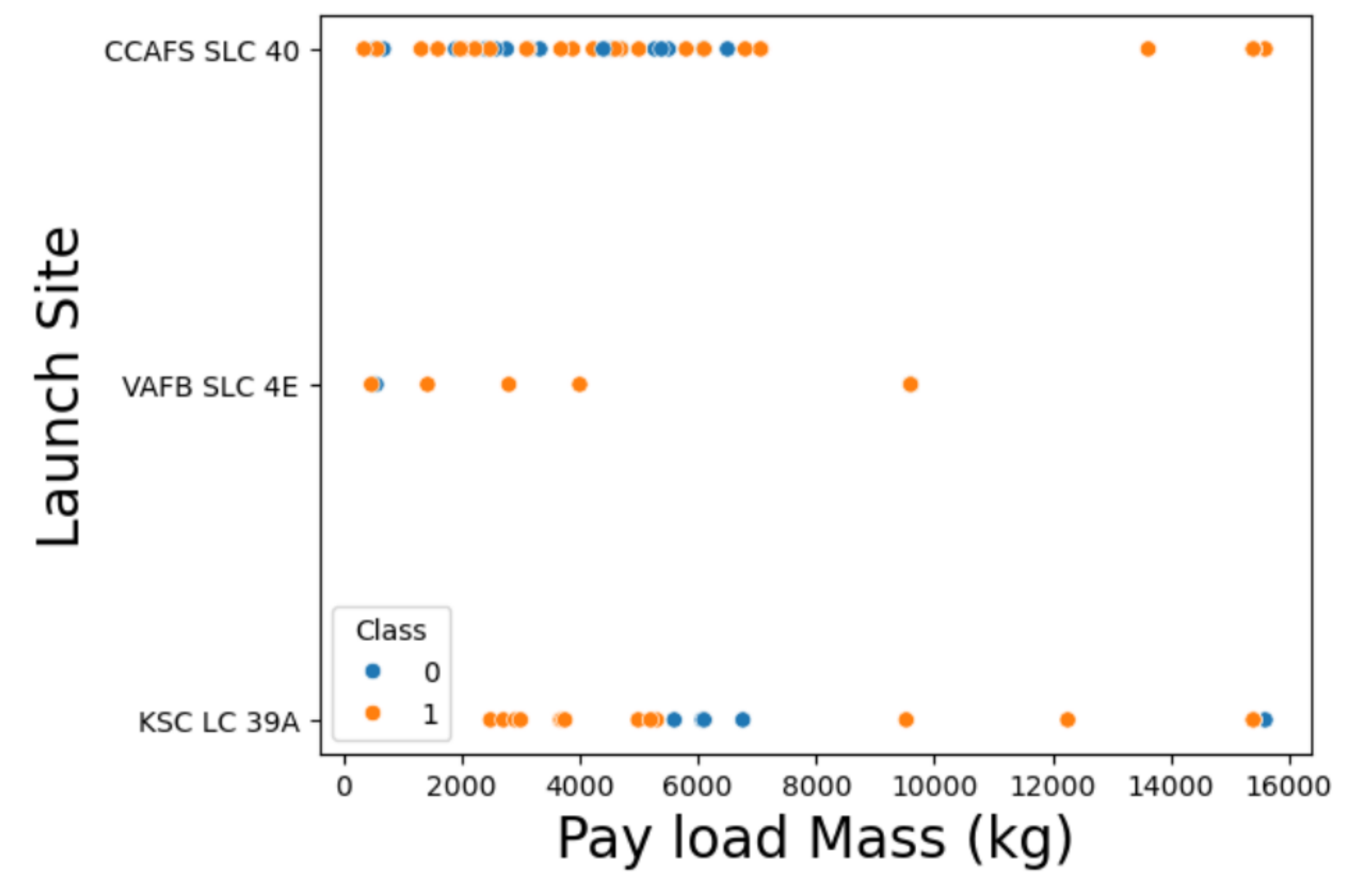
Flight Number vs. Launch Site



- As the flight number increases there seem to be more successes for all launch sites, however the successes and failures seem mixed specially until the 80th flight number mark.
- The launch site VAFB SLC did not have more flights beyond the 70th flight mark.

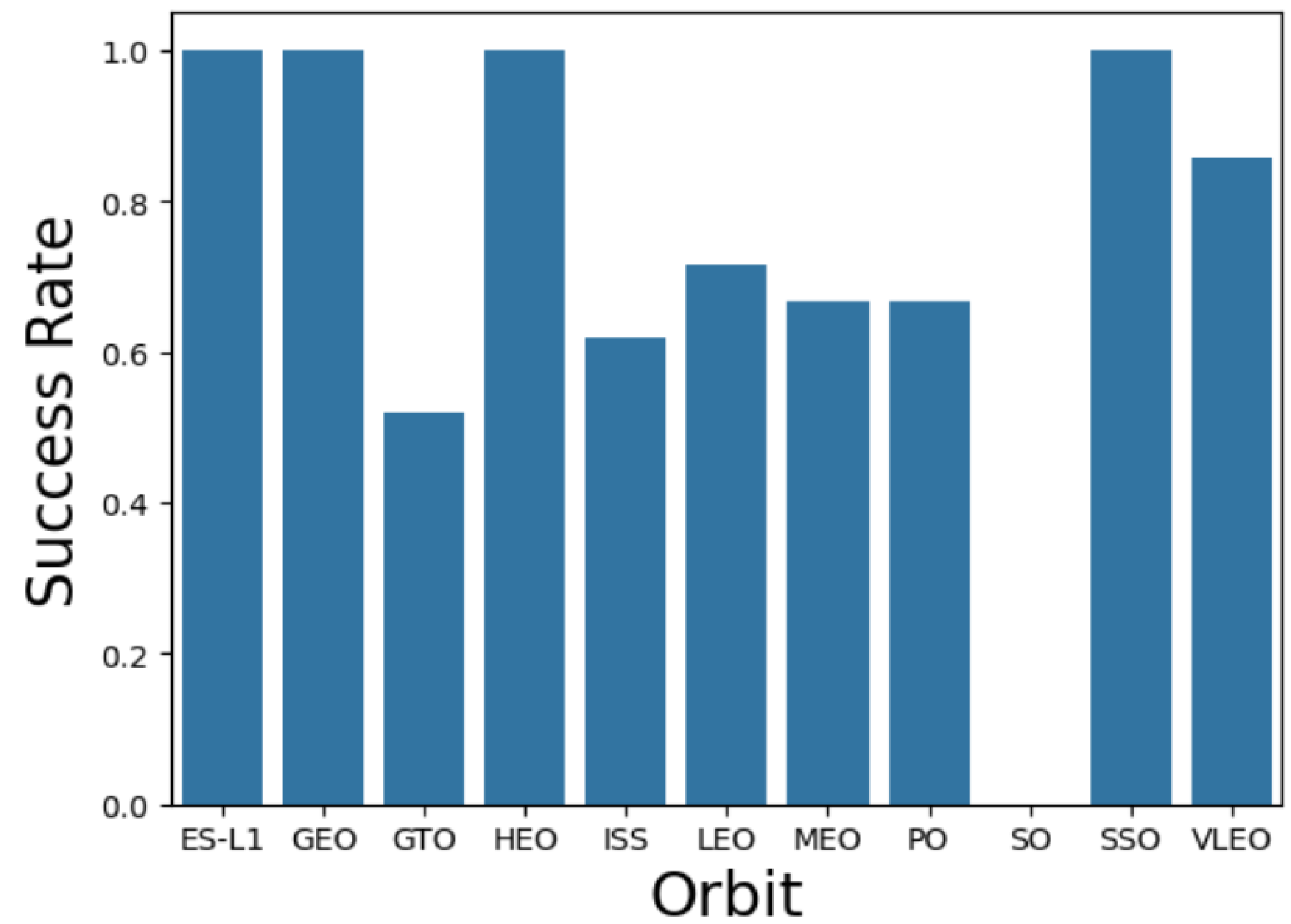
Payload vs. Launch Site

- For the CCAFS SLC 40 launch site (Canaveral Center), most rockets carried up to 8,000 kg; but with mixed success and failure rates. This site had a success rate of 100% when the payload mass was above 10,000 kg.
- For the VAFB-SLC (Vandenberg) launch site there were no rockets launched for heavy payload mass (greater than 10,000 kg).



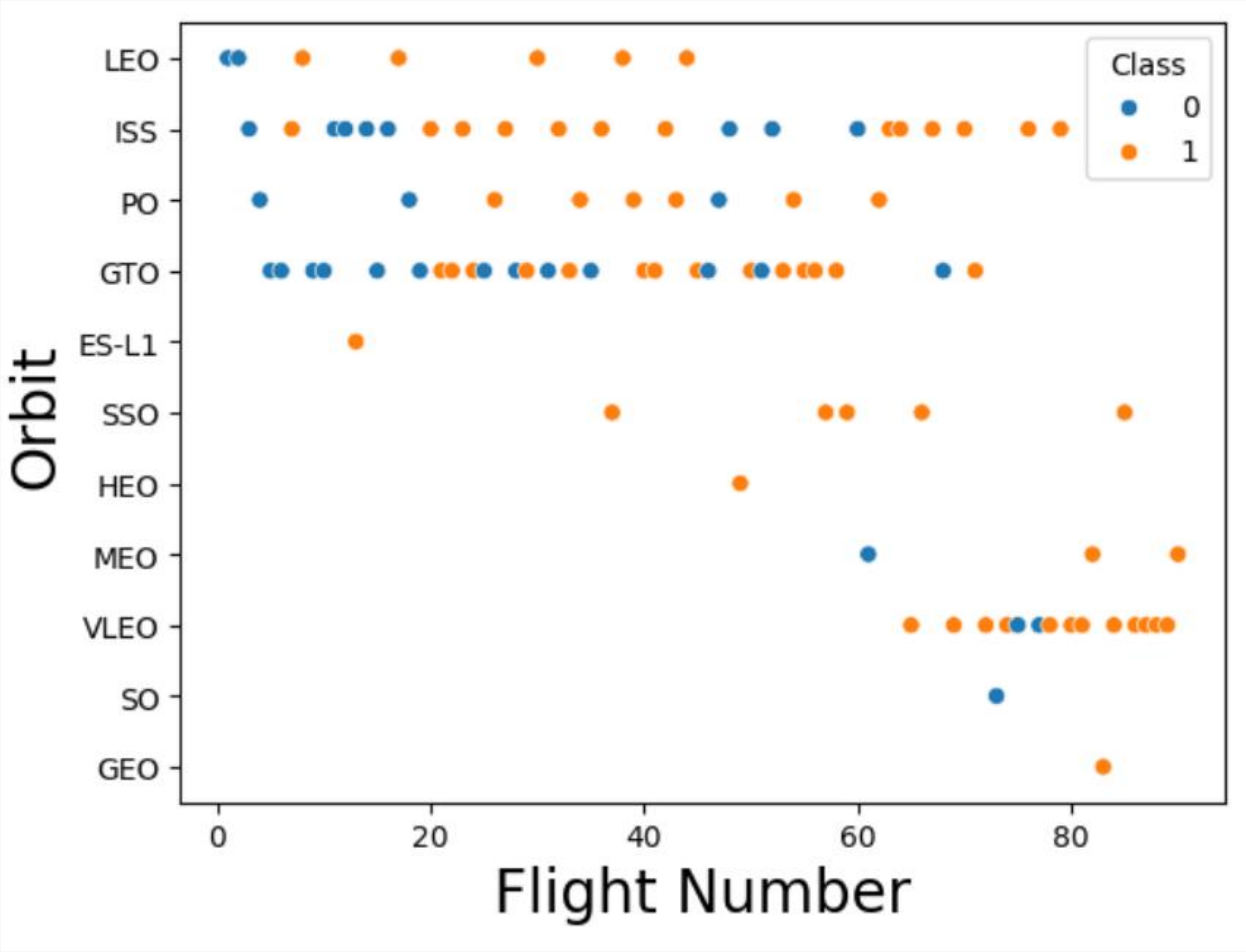
Success Rate vs. Orbit Type

- Orbits ES-L1, GEO, HEO and SSO have a high success rate of 1.
- ES-L1: The ES-Lis a gravitational point of equilibrium between Earth and Sun.
- GEO and GTO are geostationary orbits of 36,000 km from the Earth's equator, they are best known for the many satellites used for various forms of telecommunication.
- An HEO orbit is oblong, with one end nearer the Earth and other more distant.
- SSO is a particular kind of polar orbit at an altitude of about 800 km.
- It seems that the Falcon 9 had high success rates at low and high Earth orbits.



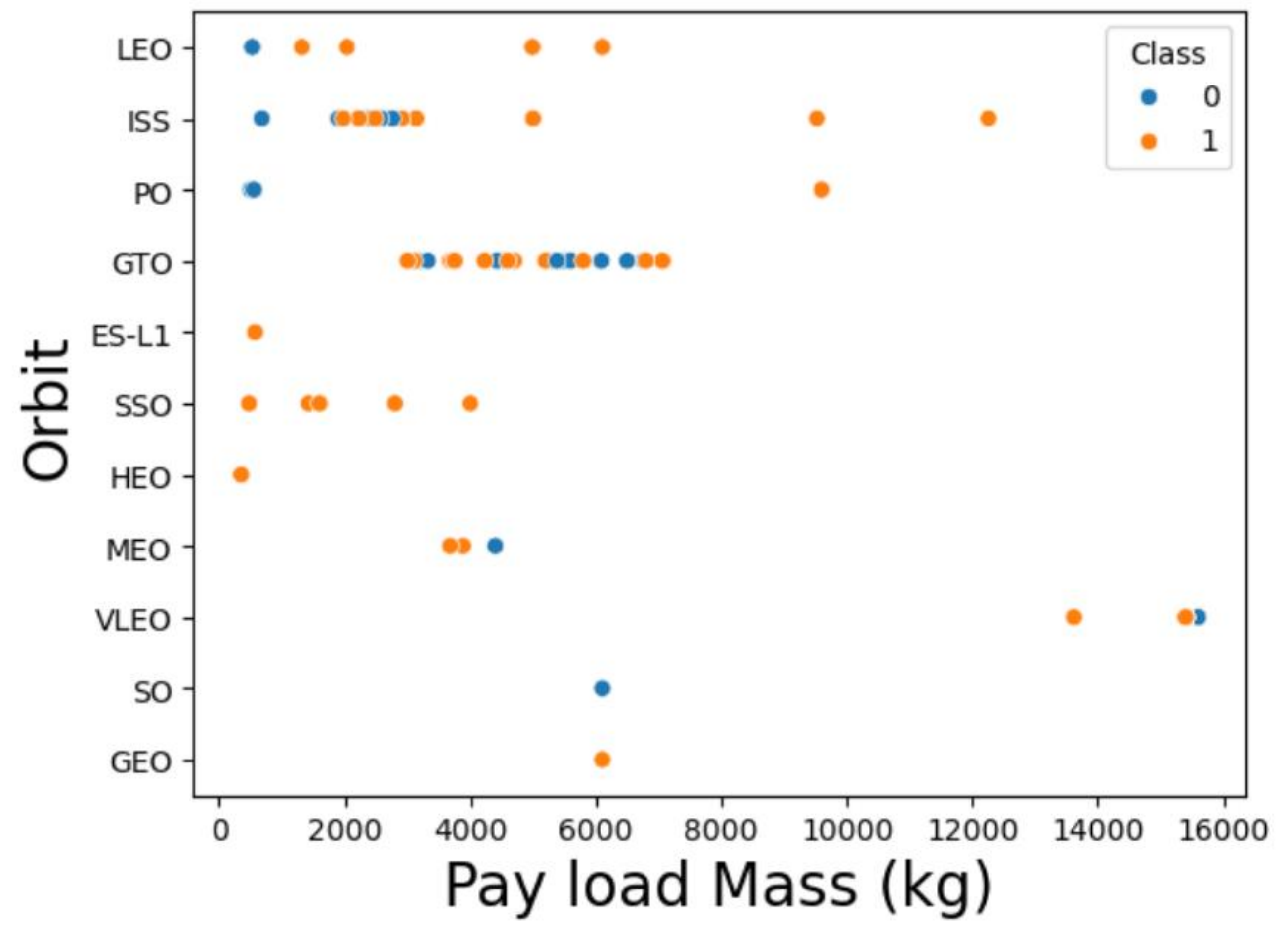
Flight Number vs. Orbit Type

- In the LEO orbit the success rate appears related to the number of flights; on the other hand, there seems to be no relationship between flight number and orbit type when in GTO.
- Flight numbers of 80 and over had several successes in VLEO orbit.



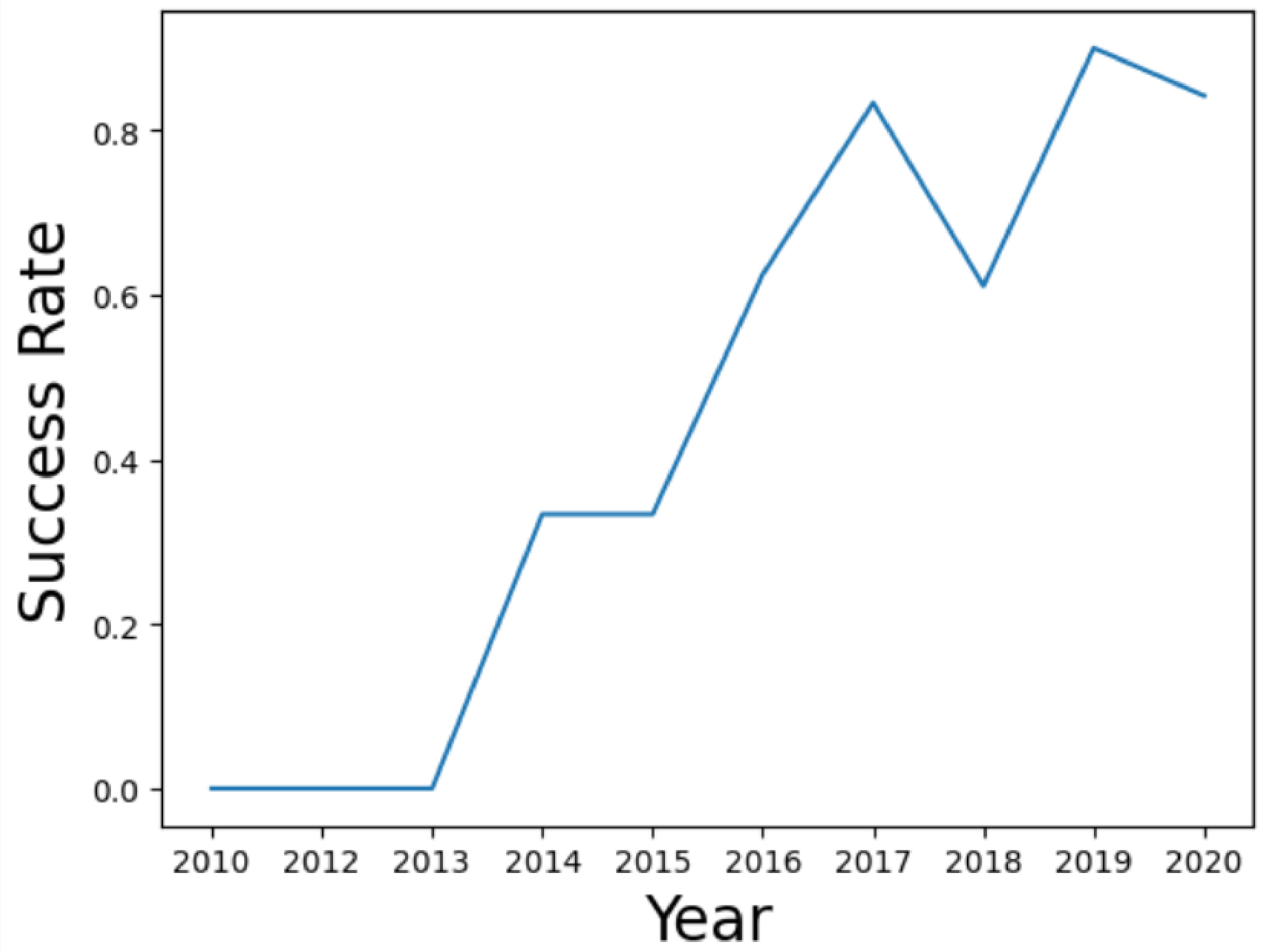
Payload vs. Orbit Type

- The few successful landings with heavy payloads (over 10,000 kg) were in PO, VLEO and ISS orbits.
- For GTO we cannot distinguish well between a successful landing and an unsuccessful mission in relation to heavier payloads.
- SSO had several successful launches at under 4,000 kg of payload mass.



Launch Success Yearly Trend

- The success rate has been increasing since 2013 until 2020.
- There was a sudden decrease in success rates in 2018. That year there were two failures to land: One in February (drone ship) and one in December (ground pad).



All Launch Site Names

SQL Query:	%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE;
Result:	Launch Sites: CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E

- Cape Canaveral Space Launch Complex (CCAFS SLC-40), and previously Launch Complex 40 (CCAFS LC-40) is a launch pad for rockets located at the north end of Cape Canaveral Space Force Station, Florida.
- Kennedy Space Center (KSC LC 39A) is located in Merritt Island, Florida.
- Vandenberg Space Force Base (VAFB SLC-4E) is located in Santa Barbara County, California.

Launch Site Names Begin with 'CCA'

SQL Query:	%sql SELECT "Launch_Site" FROM SPACEXTABLE WHERE "Launch_Site" LIKE "CCA%" LIMIT 5;
Result:	CCAFS LC-40, CCAFS LC-40, CCAFS LC-40, CCAFS LC-40, CCAFS LC-40

- In this query we retrieved the first five Launch Sites which name began with "CCA" and included any characters after that. The results show the same five Launch Sites (CCAFS LC-40), because they all met the requirements in the query.

Total Payload Mass

SQL Query:	%sql SELECT customer, SUM(PAYLOAD_MASS__KG_) AS total_payload FROM SPACEXTABLE WHERE customer="NASA (CRS)" GROUP BY customer;
Result:	Customer: NASA (CRS) Sum of Payload: 45,596

- The total payload carried by boosters belonging to customer NASA (CRS) was 45,596 kg.

Average Payload Mass by F9 v1.1

SQL Query:	%sql SELECT "Booster_Version", AVG(PAYLOAD_MASS__KG_) AS avg_payload FROM SPACEXTABLE WHERE "Booster_Version"="F9 v1.1" GROUP BY "Booster_Version"
Result:	Booster Version: F9 v1.1 Average Payload: 2,928.4 kg

- The average payload mass carried by booster version F9 v1.1 was 2,928.4 kg.

First Successful Ground Landing Date

SQL Query:	%sql SELECT "Landing Outcome", MIN("Date") AS min_date FROM SPACEXTABLE WHERE "Landing Outcome"="Success (ground pad)" GROUP BY "Landing Outcome";
Result:	Landing Outcome: Success (ground pad) Minimum Date: 2015-12-22

- The first successful landing outcome on ground pad was on December 22nd, 2015.

Successful Drone Ship Landing with Payload between 4,000 and 6,000

SQL Query:	%sql SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "Landing Outcome"="Success (drone ship)" AND (PAYLOAD_MASS__KG_>4000 AND PAYLOAD_MASS__KG_<6000) ;
Result:	Booster Version: F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2

- The names of the booster versions that had success in drone ship landing and a payload mass greater than 4,000 but less than 6,000, are:
    - F9 FT B1022, F9 FT B1026, F9 FT B1021.2 and F9 FT B1031.2.

Total Number of Successful and Failure Mission Outcomes

SQL Query:	%sql SELECT "Mission_Outcome", COUNT(*) AS count_misssion_outcome FROM SPACEXTABLE GROUP BY TRIM("Mission_Outcome");
Result:	Success: 99 Failure (in flight): 1 Success (payload status unclear): 1

- The total number of successful and failure mission outcomes are:
  - Success: 99
  - Failure (in flight): 1
  - Success (payload status unclear): 1

Boosters Carried Maximum Payload

SQL Query:	%sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) AS max_payload FROM SPACEXTABLE);
Result:	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 booster versions which have carried the maximum payload mass were:
    - 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 and F9 B5 B1049.7.

2015 Launch Records

SQL Query:	%sql SELECT substr("Date",1,4) AS year, substr("Date",6,2) AS month, "Landing Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTABLE WHERE "Landing Outcome"= "Failure (drone ship)" AND substr("Date",0,5)='2015'
Result:	2015-01, Failure (drone ship), F9 v1.1 B1012, CCAFS LC-40 2015-04, Failure (drone ship), F9 v1.1 B1015, CCAFS LC-40

- The following table shows the 2015 launch records which had a failed landing outcome in drone ship. The corresponding booster versions and launch sites were:

Year-Month	Landing Outcome	Booster Version	Launch Site
2015-01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015-04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

SQL Query:	%sql SELECT "Landing_Outcome", Count("Landing_Outcome") AS count_landing_outcomes FROM SPACEXTABLE WHERE ("Landing_Outcome" = "Failure (drone ship)" OR "Landing_Outcome" = "Success (ground pad)") AND ("Date" BETWEEN "2010-06-04" AND "2017-03-20") GROUP BY "Landing_Outcome" ORDER BY "count_landing_outcomes" DESC;
Result:	Count of Landing Outcomes: 5, 3

- The rank of landing outcomes with Failure (drone ship) or Success (ground pad); between the date 2010-06-04 and 2017-03-20 was:

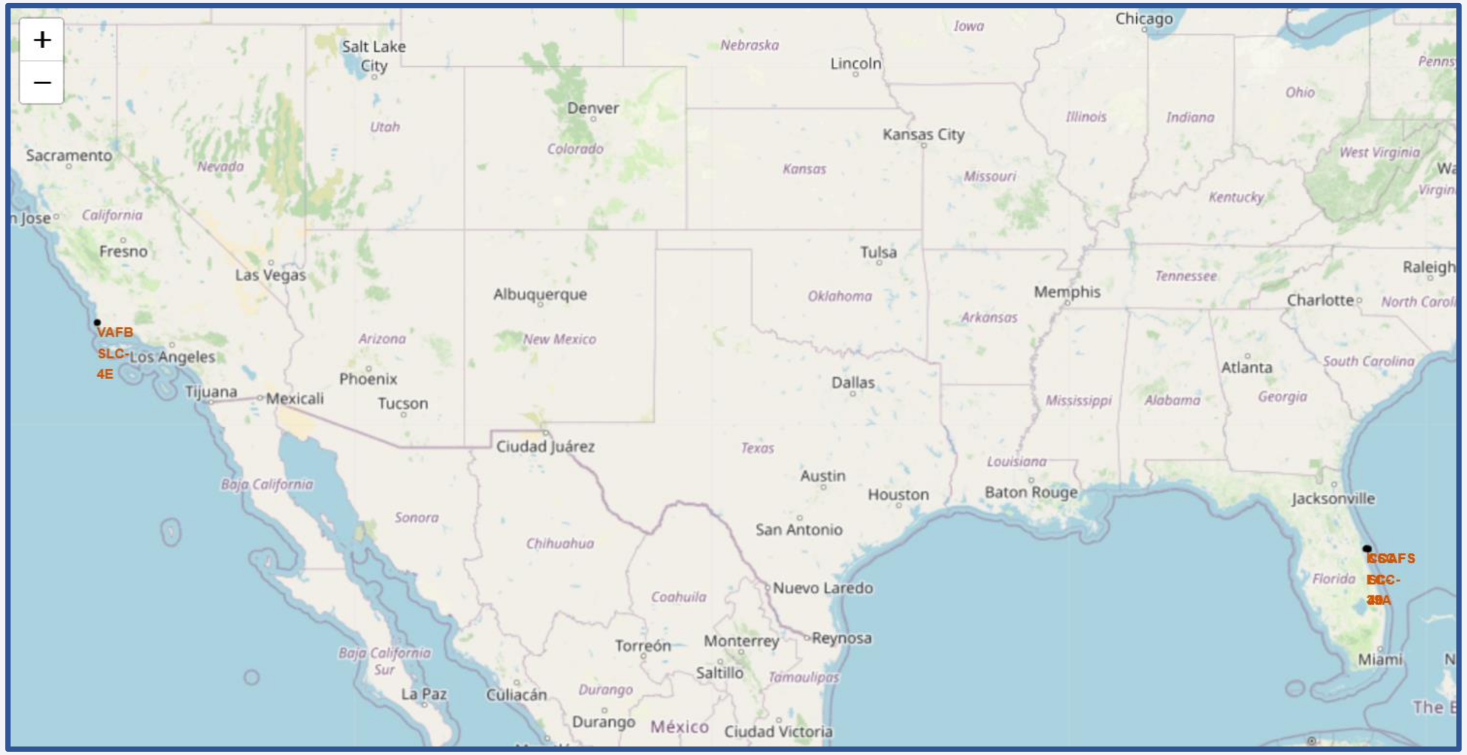
Landing Outcome	Count
Failure (drone ship)	5
Success (ground pad)	3

Section 3

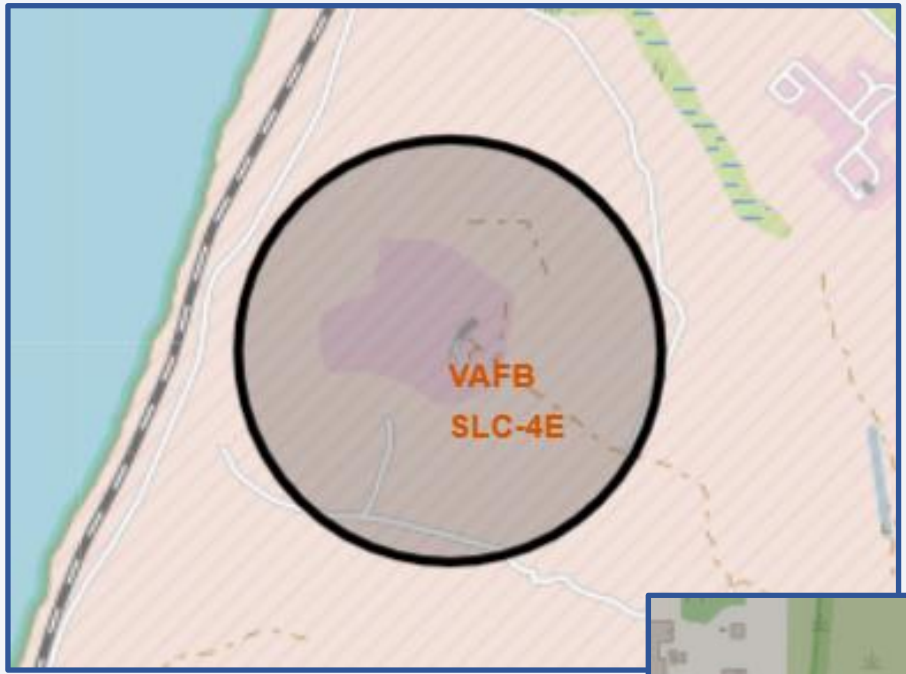
Launch Sites Proximities Analysis

Map of Launch Sites

- The launch sites in Florida (CCAFS SLC-40, KSC LC 39A) are closer to the Equator line that the launch site in California (VAFB SLC-4E).
- All launch sites are close to the Ocean. In Florida, the launch sites are close to the Atlantic Ocean and in Los Angeles, the site is close to the Pacific Ocean.



Map of Launch Sites - Close-Up



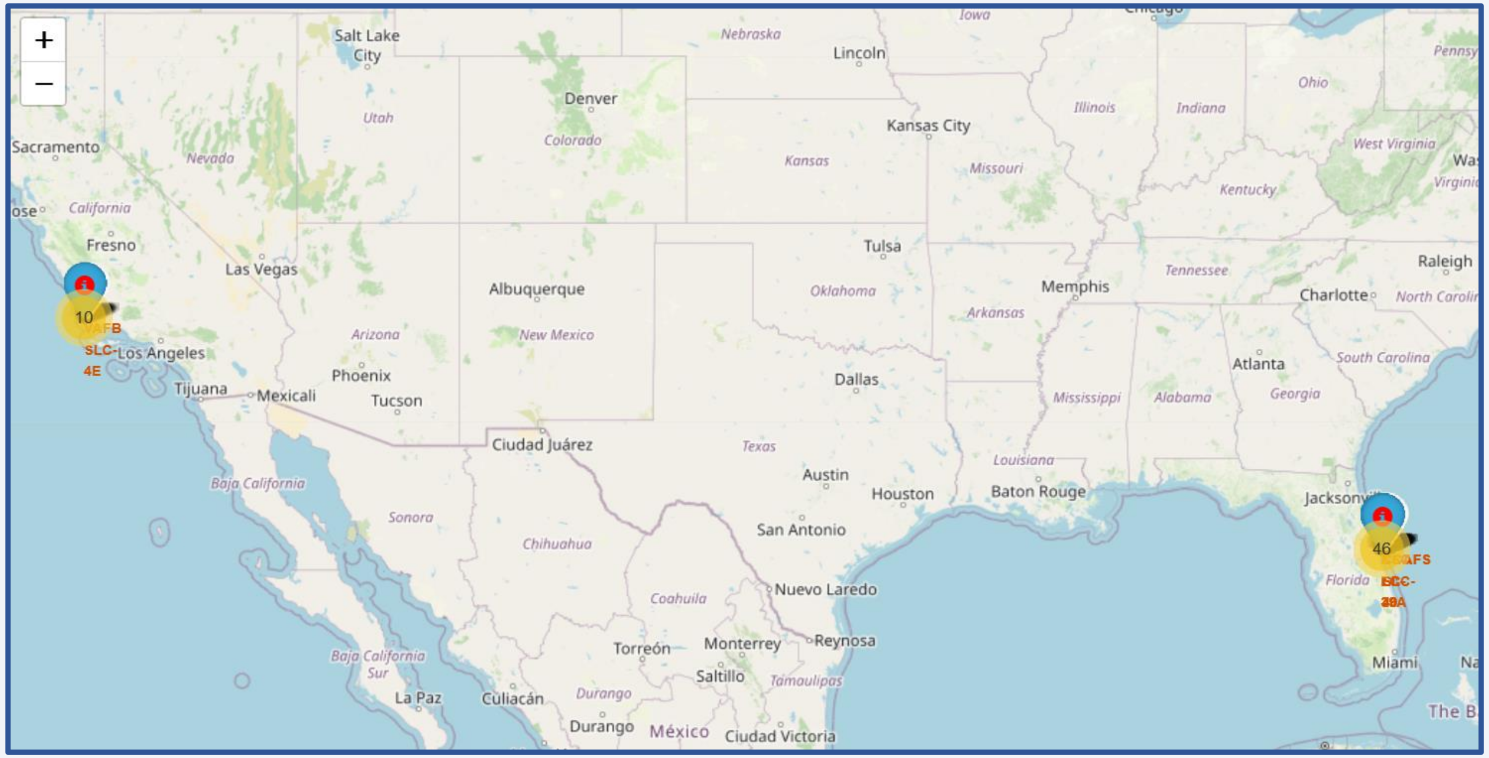
Vandenberg Space
Force Base, California.

Cape Canaveral Space
Force Station, Florida.

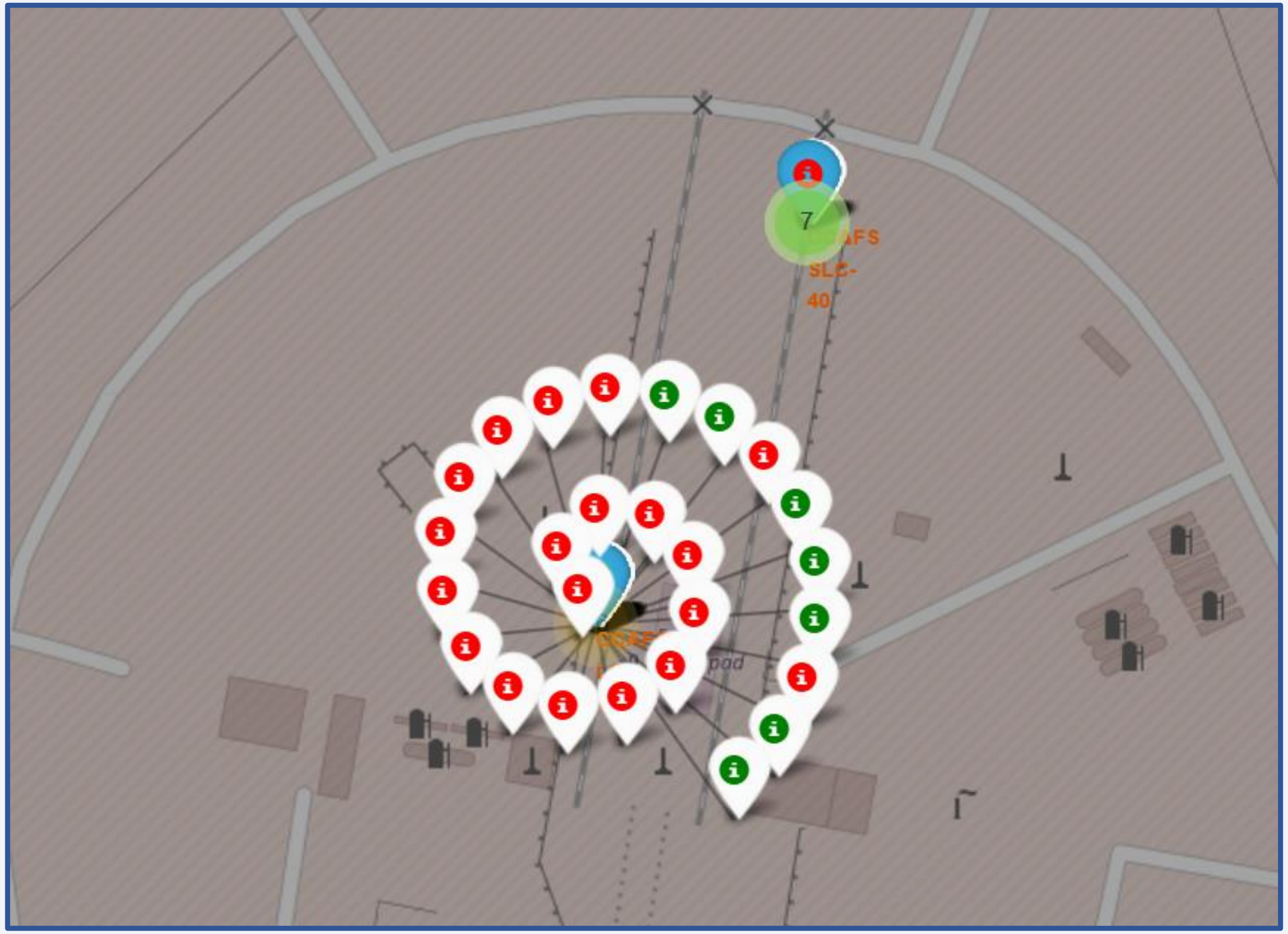
Kennedy Space
Center, Florida.

Map of Launch Sites and Outcomes

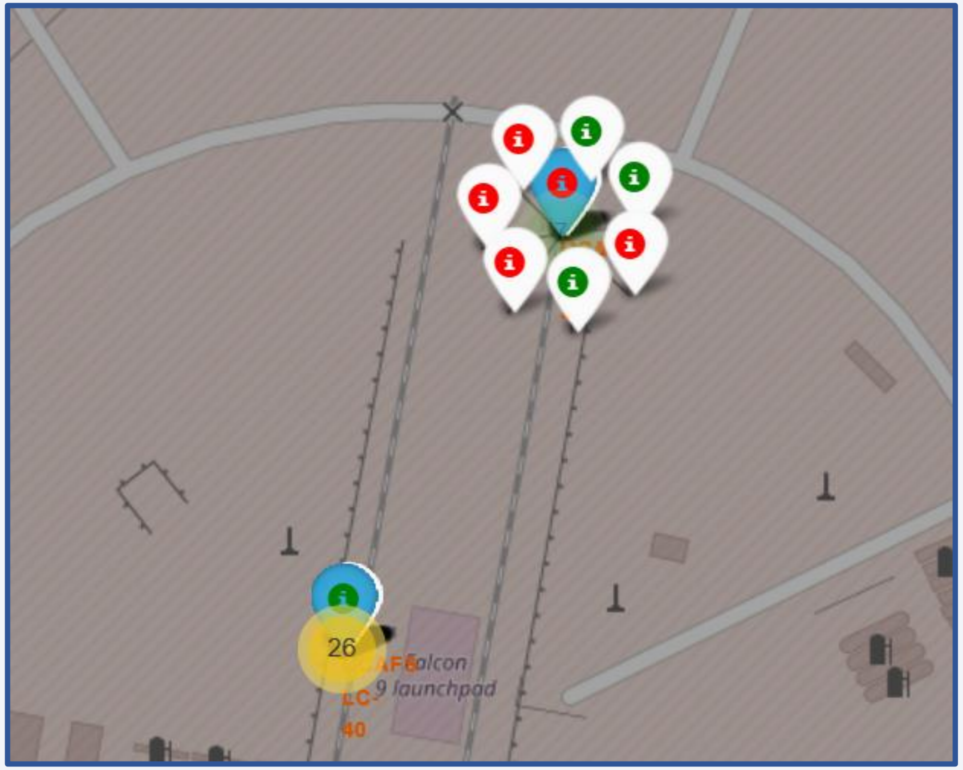
- This map shows the launch outcomes for each site. If a launch was successful, then we used a green marker and if a launch failed, we used a red marker.
- We observed that the Kennedy Space Center (KSC LC-39A) in Florida had the highest success rate.



Map of Launch Sites and Outcomes - Close-Up 1



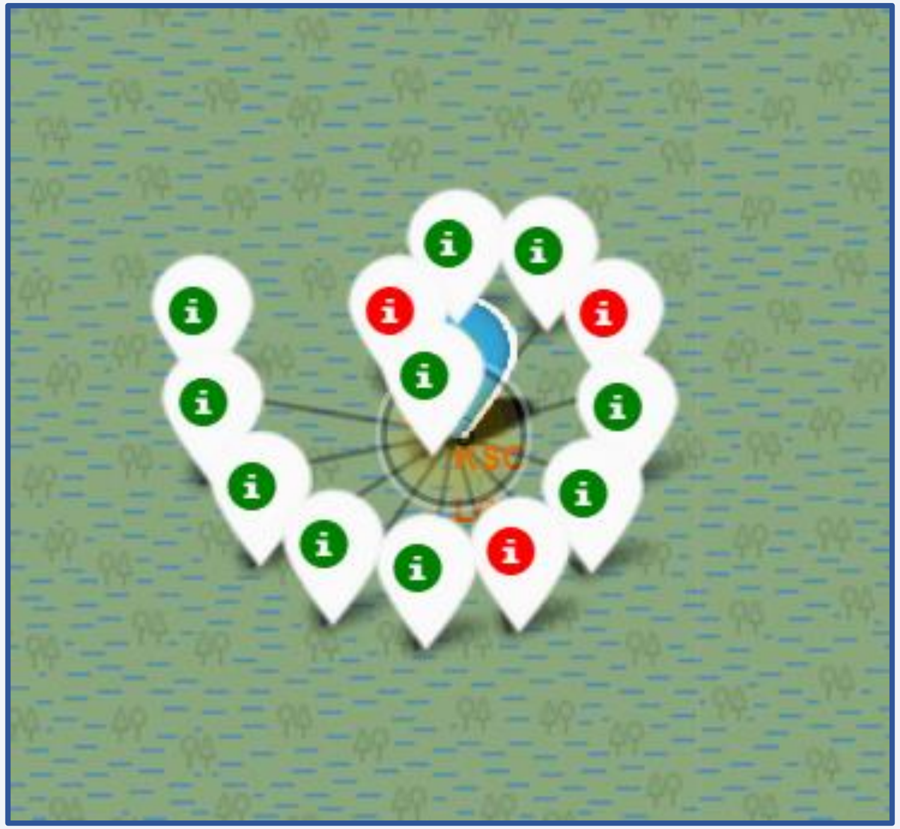
CCAFS LC -40



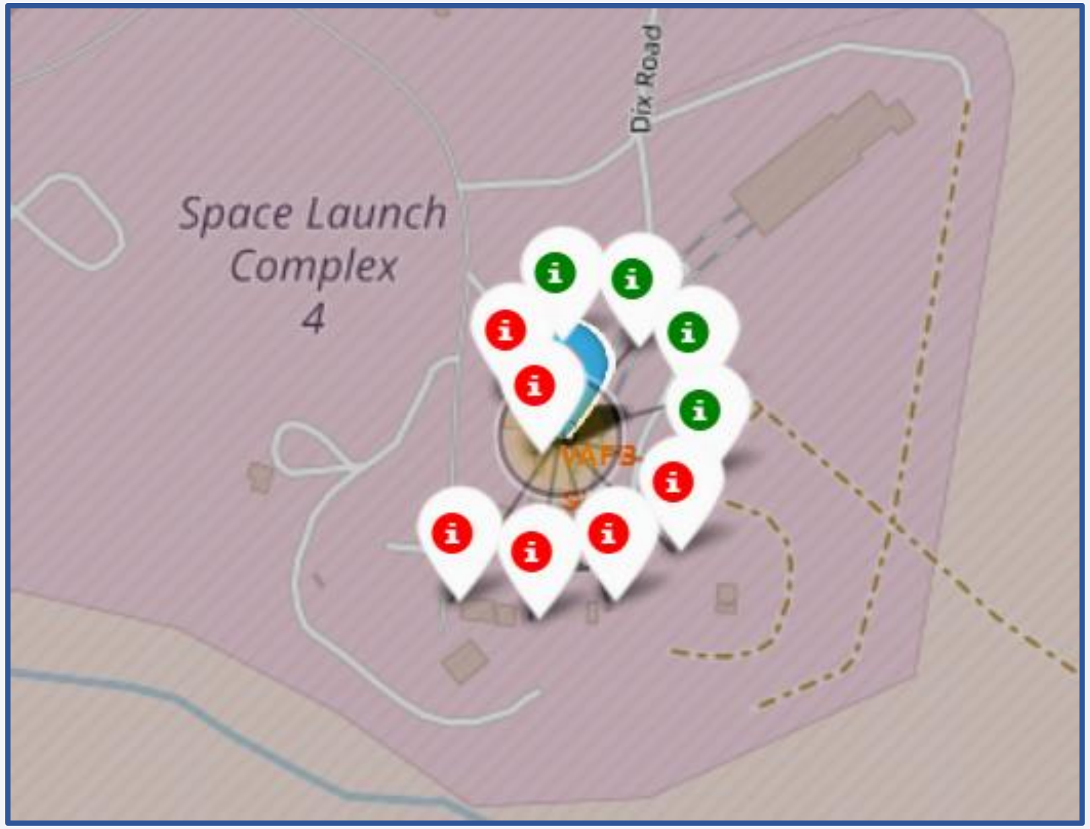
CCAFS SLC -40

Cape Canaveral Space Force Station, Florida.

Map of Launch Sites and Outcomes - Close-Up 2

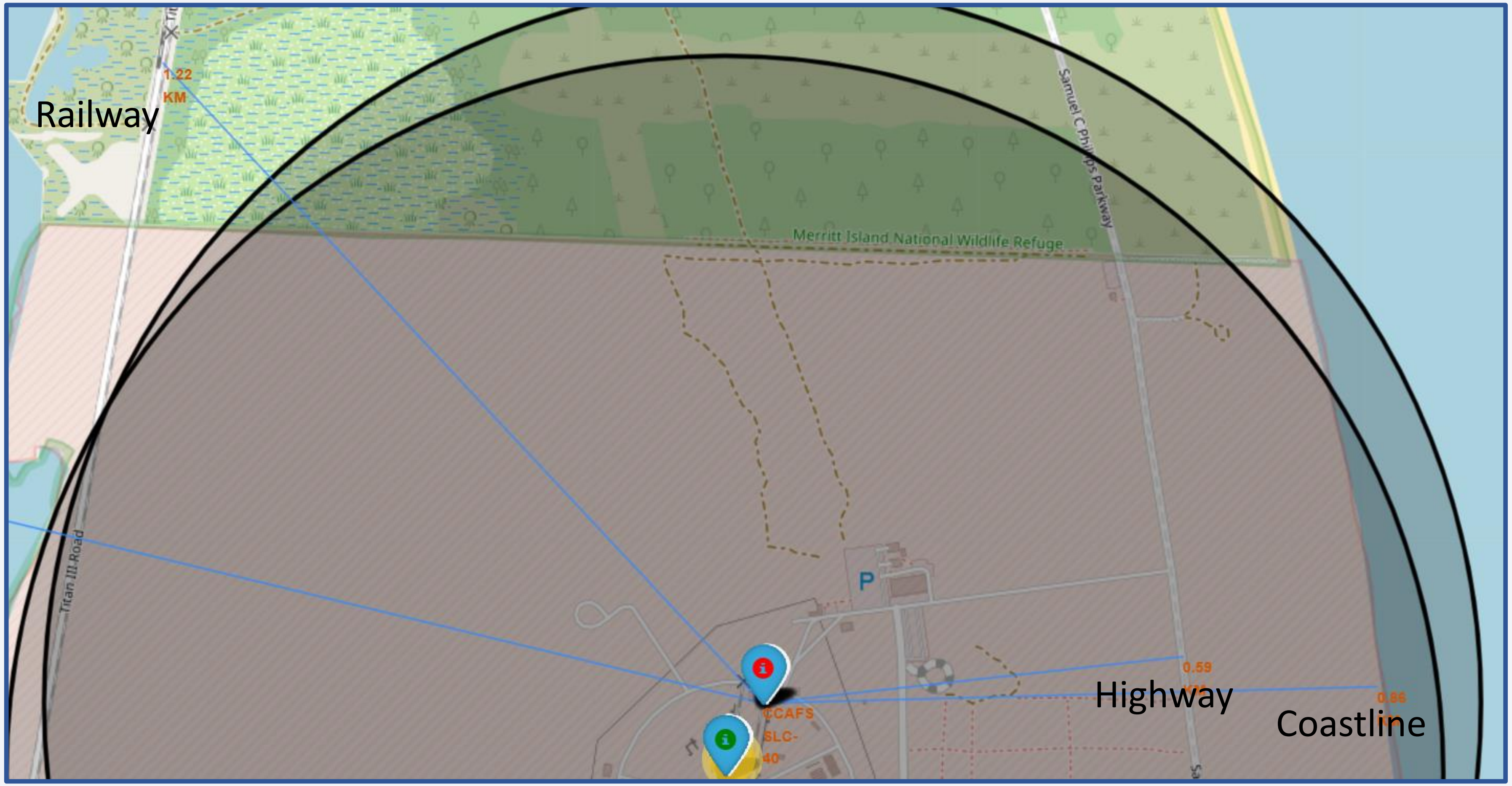


Kennedy Space Center
(KSC LC-39A), Florida.



Vandenberg Space Force Base
VAFB SLC-4E , California.

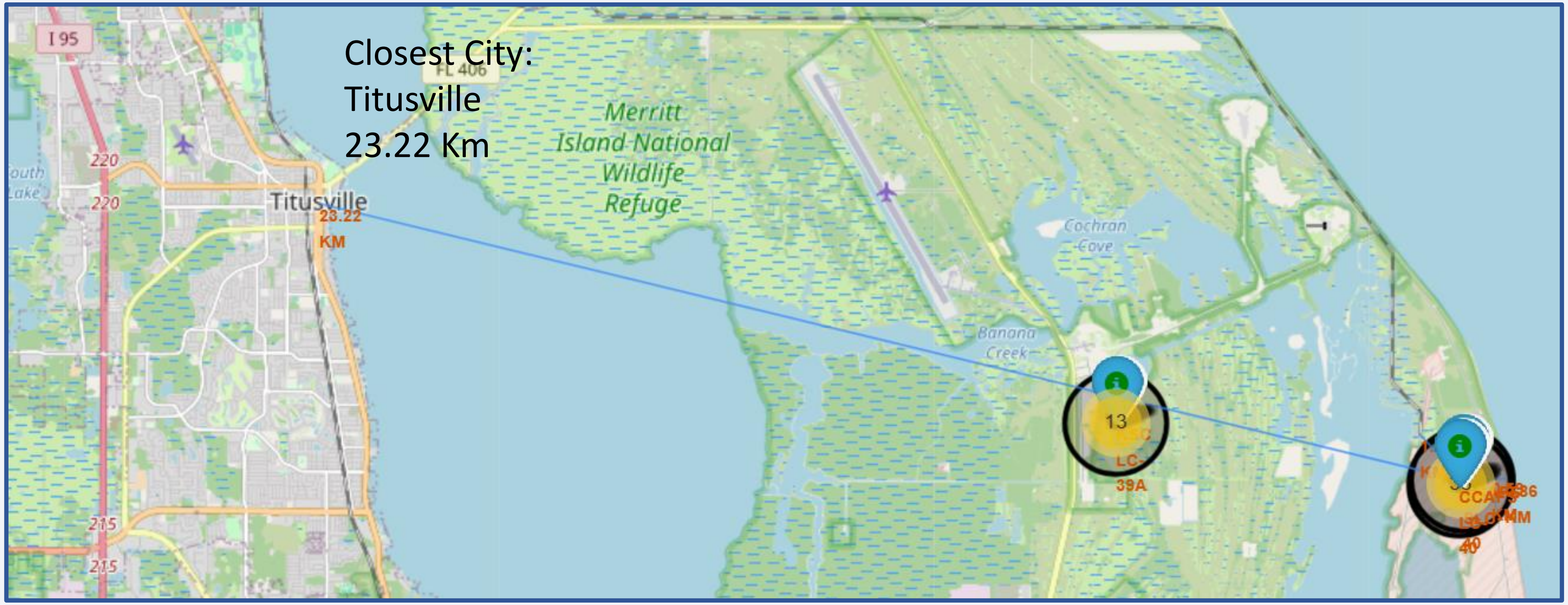
Canaveral Site and its Closest Railway, Highway and Coastline



Cape Canaveral Space Force Station (CCAFS SLC -40) in Florida and its proximities to:

- Railway: 1.22 Km
- Highway: 0.59 Km
- Coastline: 0.86 Km

Canaveral Site and its Closest City



Cape Canaveral Space Force Station (CCAFS SLC -40) in Florida is 23.22 km away from the closest city of Titusville.

Proximity to Railways, Highways, Coastline and Cities

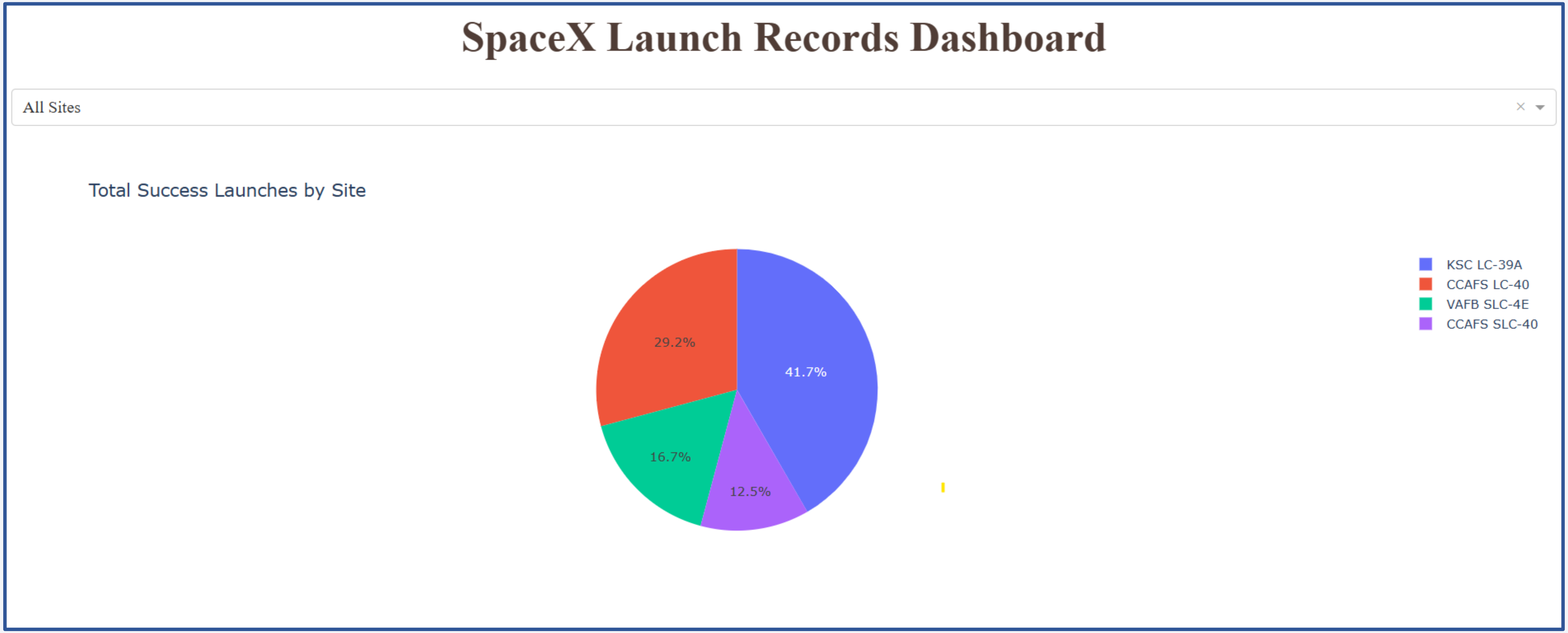
- Launch sites are in close proximity to railways and highways because it makes transportation of heavy equipment and materials more manageable. They also help maintain fuel efficiency.
- All launch sites are in close proximity to coastlines beacuse in case of failure of the launch, the rocket does not fall inland.
- All launch sites keep a good distance from cities because if something goes wrong during the ascent, the debris doesn't fall into a densely populated area.

Section 4

Build a Dashboard with Plotly Dash

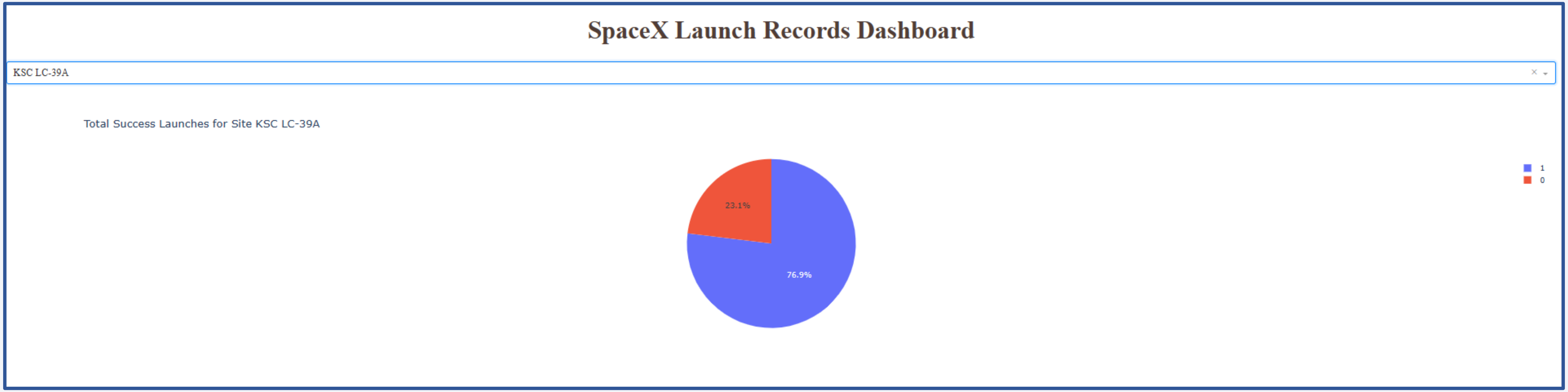
Success Rates Dashboard (for All Launch Sites)

- The Kennedy Space Center had the most successful launch rate at 41.7%. This means that out of the total launches for all sites, 41.7% were successful launches from the Kennedy Space Center. The Canaveral Launch Site followed with 29.2%.



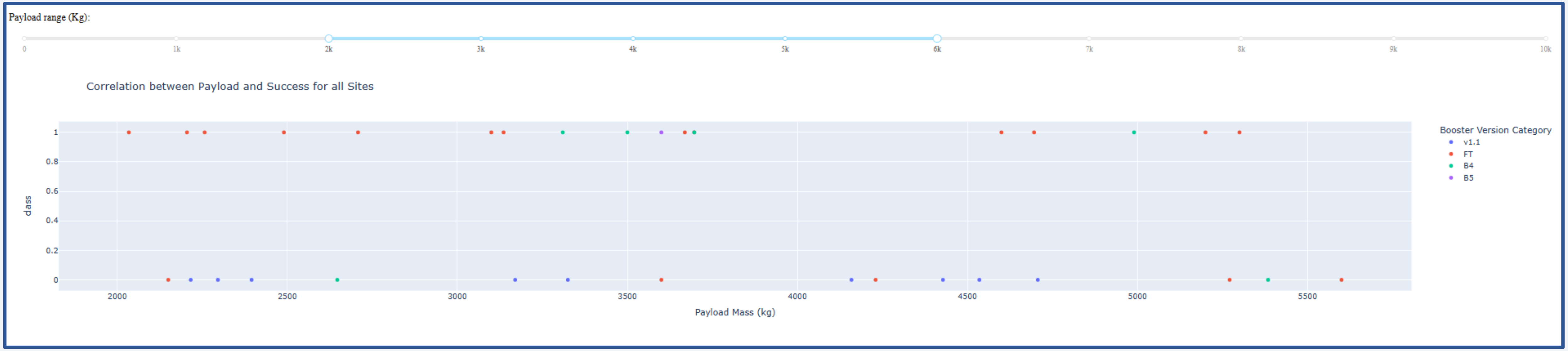
Success Rates Dashboard (for the Kennedy Space Center)

- For the Kennedy Space Center (the launch site with the highest success rate), 76.9% of their total launches were successful while 23.1% failed.



Payload vs. Launch Outcome Scatter Plot

- This scatter plot shows the successful launches (“Class” value of 1) and unsuccessful launches (“Class” value of 0) for all sites, by different payload ranges and booster version categories.
- The payload range can be selected in the range slider at the top of the graph; and the booster categories are represented in different colors for v1.1, FT, B4 and B5.
- The payload range with the highest success rate was between 2,000 and 6,000 kg. At that range, the booster version category with the largest success rate was FT while v1.1 only had failures.

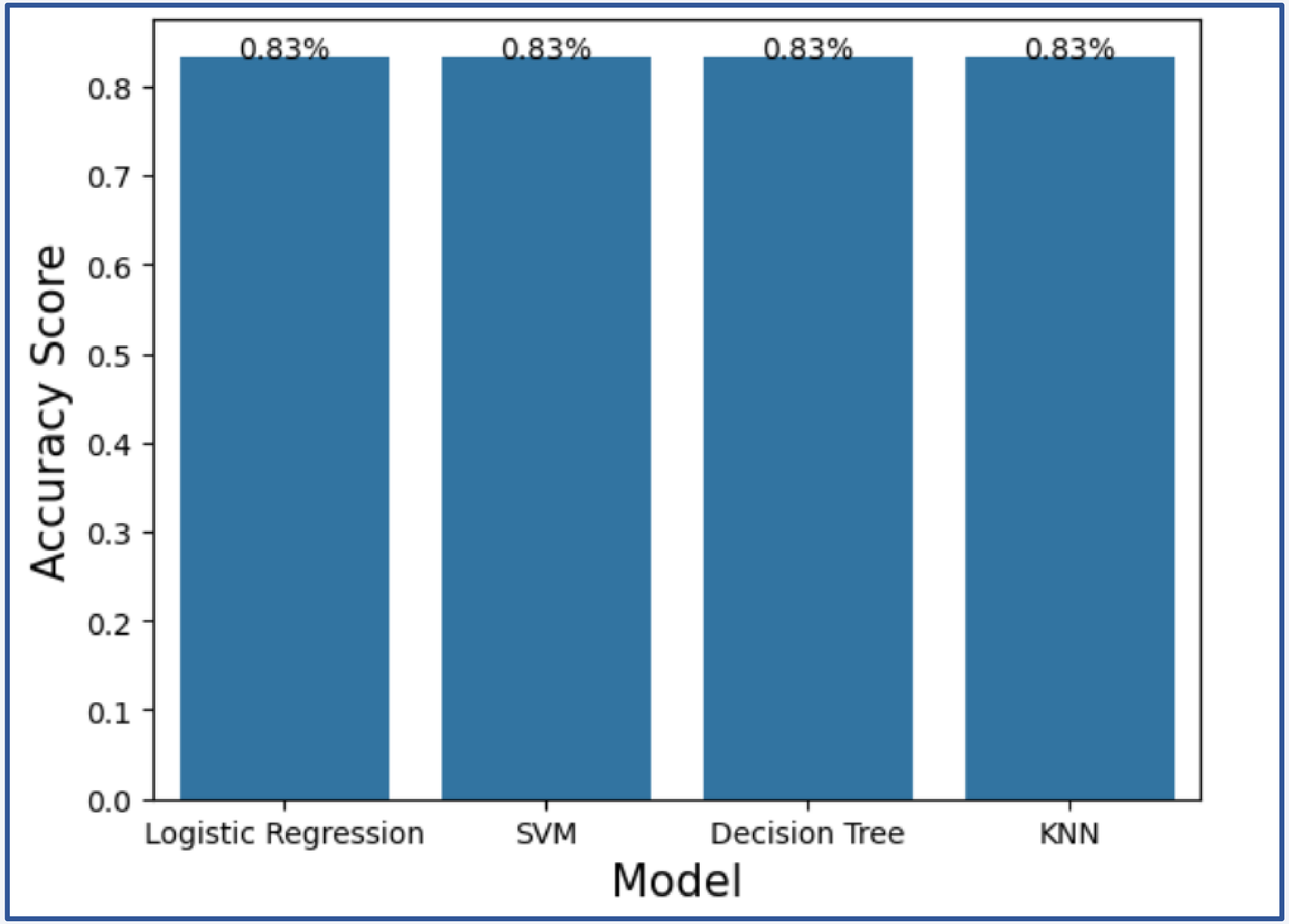


Section 5

Predictive Analysis (Classification)

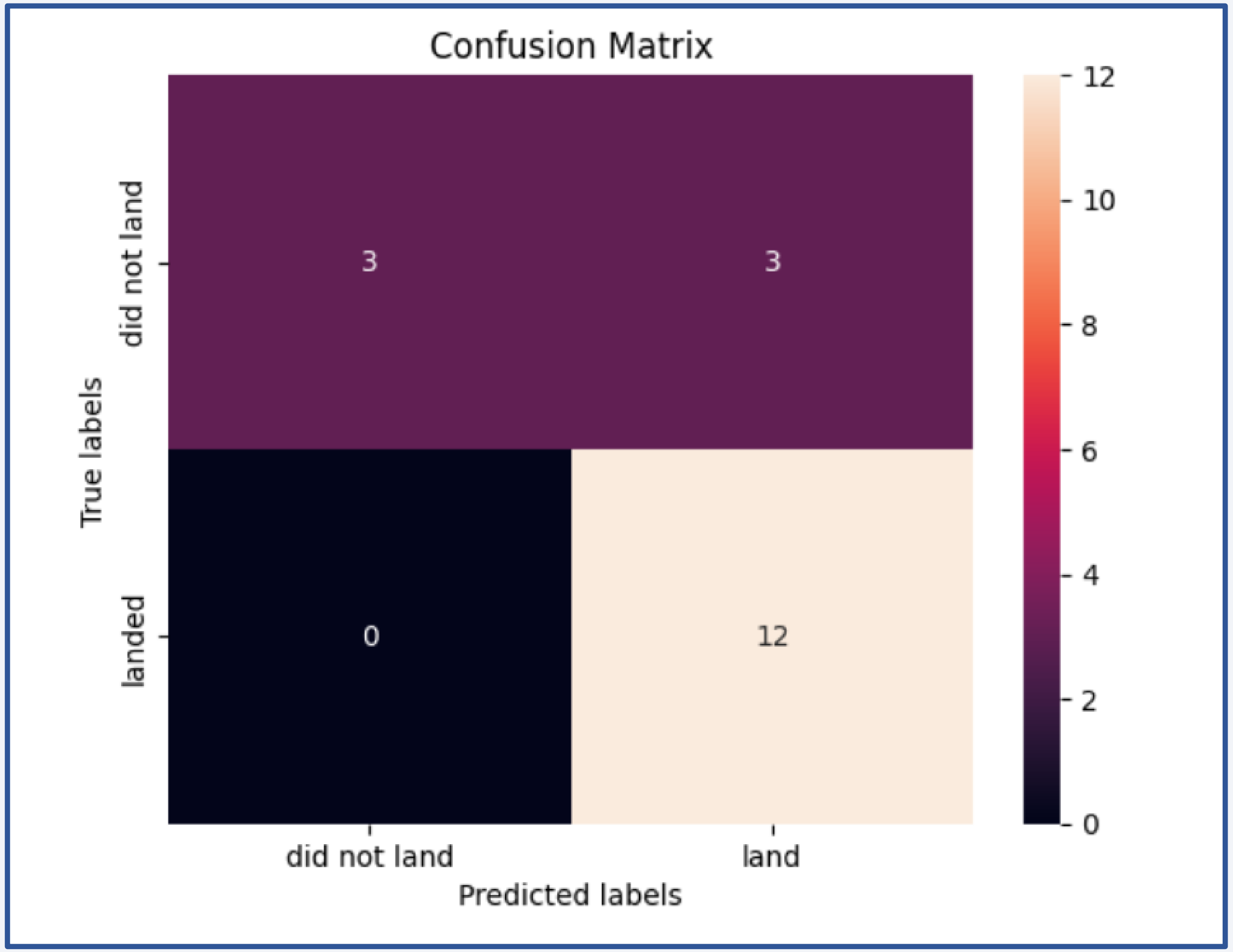
Classification Accuracy

- We developed and tested four predictive models:
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree
  - K-Nearest Neighbors
- All four models had the same classification accuracy at 0.83.



Confusion Matrix

- Since all four models had the same classification accuracy, they also had the same confusion matrix.
- The true positives are high (12) meaning the model is good at predicting the launches that will be successful.
- The problem is the false positives (3) because in these instances the model predicted a successful outcome (land) when in reality the launch failed (did not land).

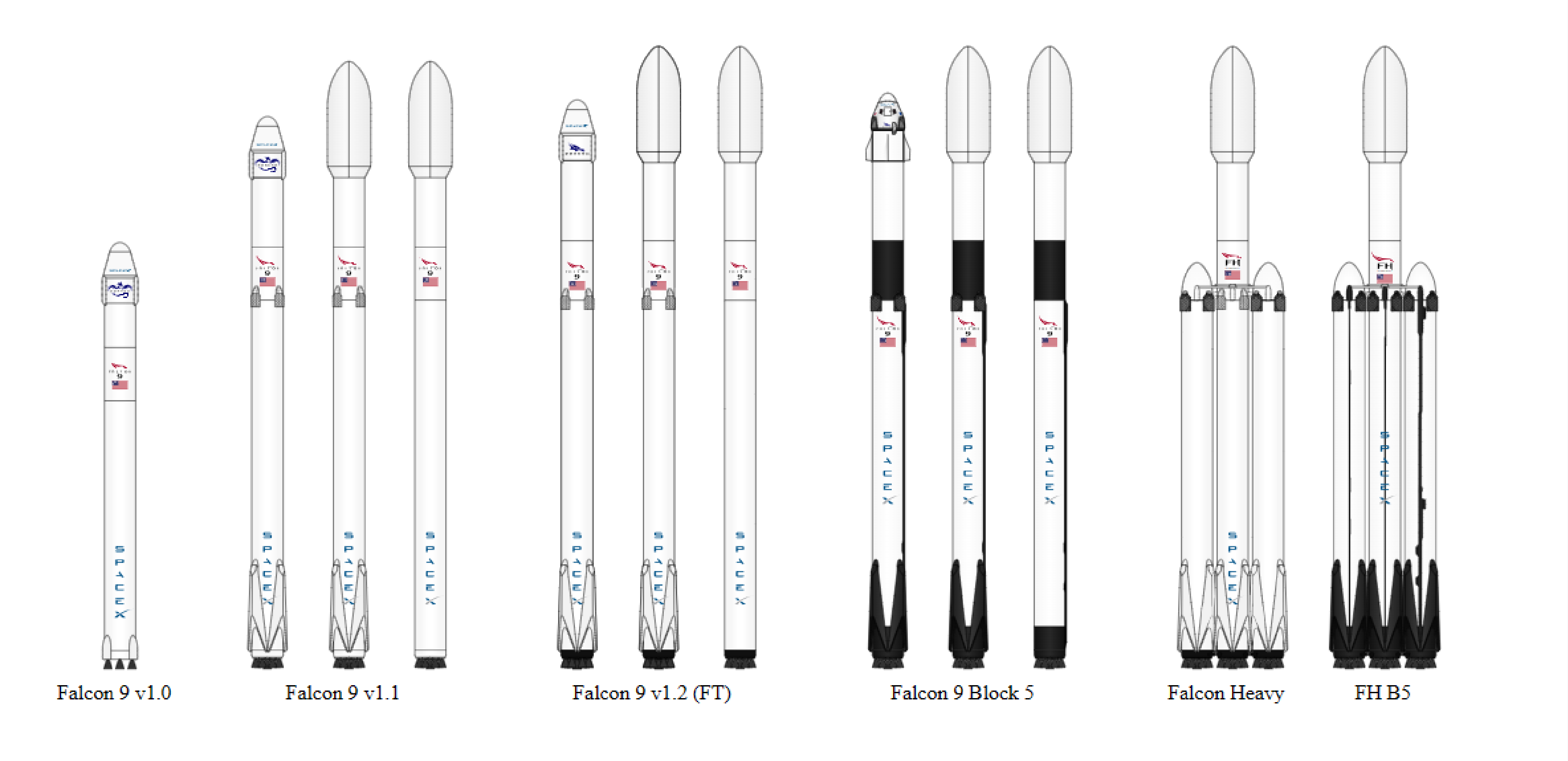


Conclusions

- SpaceX has achieved historic milestones with Falcon 9. They designed a rocket that features reusable first-stage boosters. With this capability, they were able to reduced the price of launching spacecraft to more than half the price of its competitors.
- We had the goal to develop a model that could predict if the first stage of a Falcon 9 would land successfully allowing it to be reused. If so, then we could determine the cost of a launch helping other companies to make more informed bids against SpaceX.
- We collected the data by API and web scraping. We cleaned and wrangled the raw data to convert it into a usable form. During exploratory data analysis we used graphs to observe the relationships between variables. We identified the dependent and independent variables for our model.
- We developed interactive maps and dashboards that clearly showed that the launch site with the highest success rate was the Kennedy Space Center.
- After preprocessing and data standardization, we developed and tested four predictive models: Logistic Regression, Support Vector Machine, Decision Trees and K-Nearest Neighbors.
- All four models had the same classification accuracy at 0.83; and the same confusion matrix. Although the model can be considered good at predicting, there might be room for improvement. For example, we could add other important variables such as rocket trajectory, weather conditions, technical anomalies, vehicle’s age and space traffic. These new independent variables could help predicting the success of a landing, increasing the accuracy of our model.

Appendix

Falcon 9 first-stage boosters

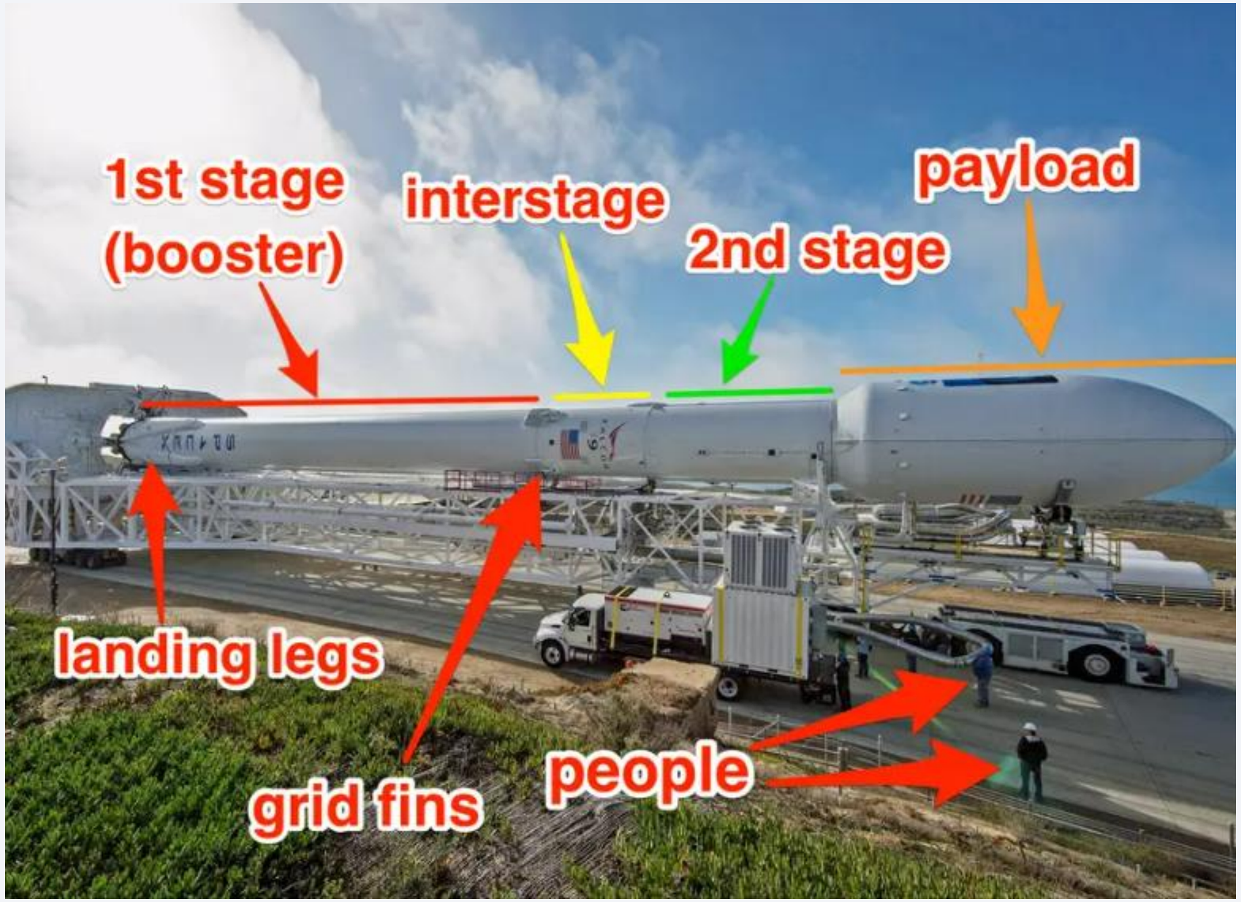


[Wikipedia Falcon 9]

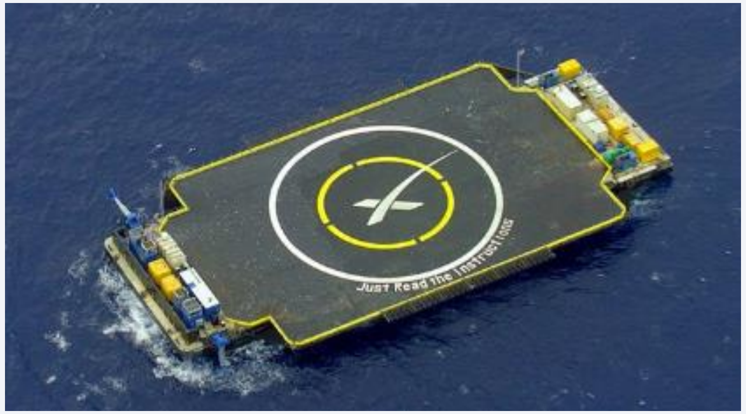
Appendix

- Falcon 9 is a rocket design that features reusable first-stage boosters. There are three booster types: Falcon 9 (F9), Falcon Heavy core (FH core) boosters, and Falcon Heavy side (FH side) boosters. Falcon 9 would be able to boost more than 9 tones to LEO orbit or more than 3 to geosynchronous transfer orbit (GTO) for $27 million. Falcon 9 Boosters are numbered with a B followed by a four-digit number.
- The “FT” in Falcon 9 stands for Full Thrust, which refers to the upgraded version of the Falcon 9 rocket developed by SpaceX. The Full Thrust version, also known as Falcon 9 v1.2, introduced several improvements over its predecessors, including increased thrust, enhanced performance, and the capability for both stages of the rocket to be reusable.
- Our data ranged from June 2010 to November 2020. Now we know that in 2023 SpaceX was very successful: They performed 91 launches of Falcon 9 with only 4 utilizing new boosters and successfully recovered the booster on all flights.

Appendix



Note: The second stage, powered by a single Merlin Vacuum Engine, delivers Falcon 9's payload to the desired orbit.



Landing Pad in Ocean.

A New Trick May Help SpaceX Launch Bigger Payloads, Save More Rockets (businessinsider.com)

Appendix

- View of database before applying One Hot Encoder to categorical variables.

```
]: features = df[['FlightNumber', 'PayloadMass', 'Orbit', 'LaunchSite', 'Flights', 'GridFins', 'Reused', 'Legs', 'LandingPad', 'Block', 'ReusedCoun
   features.head()
```

	FlightNumber	PayloadMass	Orbit	LaunchSite	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial
0	1	6104.959412	LEO	CCAFS SLC 40	1	False	False	False	NaN	1.0	0	B0003
1	2	525.000000	LEO	CCAFS SLC 40	1	False	False	False	NaN	1.0	0	B0005
2	3	677.000000	ISS	CCAFS SLC 40	1	False	False	False	NaN	1.0	0	B0007
3	4	500.000000	PO	VAFB SLC 4E	1	False	False	False	NaN	1.0	0	B1003
4	5	3170.000000	GTO	CCAFS SLC 40	1	False	False	False	NaN	1.0	0	B1004

Appendix

- Variables in database after Hot Encoding, ready for the machine learning model.

'FlightNumber’, 'PayloadMass’, 'Flights’, 'Block’, 'ReusedCount’,
‘Orbit_ES-L1', 'Orbit_GEO', 'Orbit_GTO', 'Orbit_HEO', 'Orbit_ISS’, 'Orbit_LEO’,
'Orbit_MEO', 'Orbit_PO', 'Orbit_SO', 'Orbit_SSO’, 'Orbit_VLEO’,
'LaunchSite_CCAFS SLC 40', 'LaunchSite_KSC LC 39A’, 'LaunchSite_VAFB SLC 4E’,
'LandingPad_5e9e3032383ecb267a34e7c7’, 'LandingPad_5e9e3032383ecb554034e7c9’,
'LandingPad_5e9e3032383ecb6bb234e7ca’, ’LandingPad_5e9e3032383ecb761634e7cb',
'LandingPad_5e9e3033383ecbb9e534e7cc’, 'Serial_B0003', 'Serial_B0005’, 'Serial_B0007', 'Serial_B1003’,
'Serial_B1004', 'Serial_B1005’, 'Serial_B1006', 'Serial_B1007’, 'Serial_B1008', 'Serial_B1010’, 'Serial_B1011',
'Serial_B1012’, 'Serial_B1013', 'Serial_B1015’, 'Serial_B1016', 'Serial_B1017’, 'Serial_B1018', 'Serial_B1019’,
'Serial_B1020', 'Serial_B1021’, 'Serial_B1022', 'Serial_B1023’, 'Serial_B1025', 'Serial_B1026’, 'Serial_B1028',
'Serial_B1029’, 'Serial_B1030', 'Serial_B1031’, 'Serial_B1032', 'Serial_B1034’, 'Serial_B1035', 'Serial_B1036’,
'Serial_B1037', 'Serial_B1038’, 'Serial_B1039', 'Serial_B1040’, 'Serial_B1041', 'Serial_B1042’, 'Serial_B1043',
'Serial_B1044’, 'Serial_B1045', 'Serial_B1046’, 'Serial_B1047', 'Serial_B1048’, 'Serial_B1049', 'Serial_B1050’,
'Serial_B1051', 'Serial_B1054’, 'Serial_B1056', 'Serial_B1058’, 'Serial_B1059', 'Serial_B1060’, 'Serial_B1062’,
'Reused_False’, 'Reused_True’,
'GridFins_False', 'GridFins_True’,
'Legs_False', 'Legs_True'

Thank you!