IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

David Wertheimer
November 2023

Outline

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

Executive Summary

- Objectives
  - Analyze data related to commercial space flight.
  - Since the overall cost is related to the reusability of rocket first-stage boosters, we will focus our analysis in this area and specifically on the performance of the SpaceX Falcon 9 booster.
- Summary of methodologies
  - Various data science and machine learning methods will be used to perform data analysis, discovery, and visualization of results.
- Summary of all results
  - Using these approaches, we've learned much about SpaceX's ability to land booster stages successfully. In this presentation, we will explore our approach more fully and share findings.

Introduction

- We are interested in predicting if the Falcon 9 first stage will land successfully.
- SpaceX advertises Falcon 9 rocket launches on its website at 62 million dollars; other providers charge 165 million dollars per launch.
- SpaceX can realize savings because they can reuse the first-stage boosters.
- So, if the first stage lands successfully, subsequent launch costs can be reduced. This information is helpful for a competing company that wants to bid against SpaceX for a rocket launch.

Section 1

Methodology

Methodology:

Executive Summary

- Data collection methodology:
  - SpaceX launch and booster data was obtained using two methods: REST API requests to the SpaceX APIs and by scraping data from Wikipedia.
- After data collection, the following methods were used to prepare and analyze the data:
  - Perform Data wrangling
  - Perform Exploratory data analysis (EDA) using visualization and SQL
  - Perform Interactive visual analytics using Folium and Plotly Dash
  - Perform Predictive analysis using classification models

Data Collection

- For this project, data was collected from two data sources:

1. SpaceX REST API:
    - The SpaceX REST API version 4 was used to access past launch data and additional details.
2. Wikipedia Pages:
    - Data was scraped from the Wikipedia "List of Falcon 9 and Falcon Heavy launches" page.

Each source provided access to launch details such as flight number, date, payload mass, orbit type, launch outcome, booster version, landing outcome, and other data.

Data Collection – SpaceX API

- The https://api.spacexdata.com/ public API was accessed to pull SpaceX launch and booster data.
- The steps outlined in the flowchart were followed to collect, clean and persist data from the SpaceX API.
- The completed SpaceX API calls notebook that implements this process is available on GitHub.

TASK 1. Request and parse the SpaceX launch data using the GET request REST API https://api.spacexdata.com/v4/launches/past

↓

Extracted required columns and replaced IDs with details from the following API endpoints to construct a data frame.

https://api.spacexdata.com/v4/rockets
https://api.spacexdata.com/v4/launchpads
https://api.spacexdata.com/v4/payloads
https://api.spacexdata.com/v4/cores

↓

TASK 2. Filtered the data frame only to include Falcon 9 launches.

↓

TASK 3. Performed data wrangling to replace missing payload data with the mean payload mass.

↓

Export the Data Frame to a CSV file for future use.

Data Collection - Scraping

- Data for SpaceX launches is also available on a Wikipedia page.
- Data was scraped from the Wikipedia "[List of Falcon 9 and Falcon Heavy launches]" page and saved for future use.
- The completed web scraping notebook is available on [GitHub].

TASK 1. Request the Falcon9 Launch Wiki page from its URL
https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches

↓

TASK 2: Extract all column/variable names from the HTML table.

↓

TASK 3: Create a header into a dictionary structure by parsing the launch HTML tables

↓

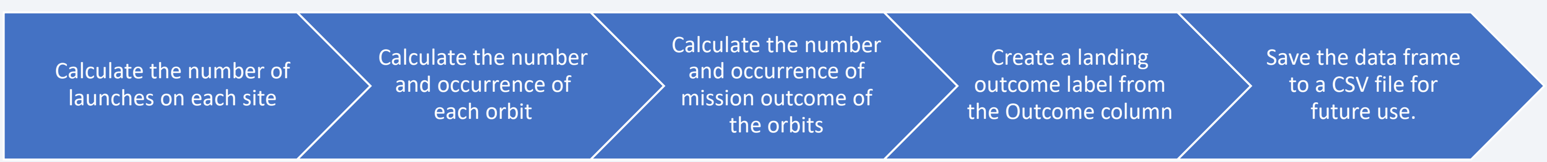
Convert the dictionary into a data frame

↓

Export the Data Frame to a CSV file for future use.

Data Wrangling

- Exploratory Data Analysis (EDA) was performed to find patterns in the data and determine the label for training supervised models.
- This exploration looked at launches from each site, orbits, mission outcome by orbit, and overall mission outcome.
- Since not all landing attempts were successful, it was helpful to convert landing outcome data into training labels where "1" indicates a successful landing and "0" indicates a failed landing.



- The completed data wrangling notebook is available for review on GitHub.

EDA with Data Visualization

Parameters	Plot Type	Why	Observation
• Flight Number • Payload Mass • Landing Outcome	**Scatter Plot** with color representing landing outcome	See the impact of launch attempts over time and payload mass would impact launch outcome.	As SpaceX gained experience and confidence, they increased payload mass and achieved greater landing success.
• Flight Number • Launch Site • Landing Outcome	**Scatter Plot** with color representing landing outcome	To visualize the relationship between flight number (attempts over time) and the launch site.	As the number of launches increased, the landing success rate also increased. It's unclear from the visualization if the launch site had much impact.
• Payload Mass • Launch Site • Landing Outcome	**Scatter Plot** with color representing landing outcome	Visualize the relationship between Payload and Launch Site. Does the mass of the payload combined with the launch site show us anything?	The most unambiguous indication from the visualization is that no launches over 10,000 kg were attempted from VAFB SLC 4E.
• Orbit • Landing Outcome	**Bar Plot** representing success rate as a percentage.	Visualize the relationship between the success rate of each orbit type.	The Orbits with the highest level of successful landing outcomes were ES-L1, GEO, HEO, and SSO.
• Flight Number • Orbit • Landing Outcome	**Scatter Plot** with color representing landing outcome	Visualize the relationship between Flight Number (attempts over time) and Orbit type.	The LEO orbit success rate is related to the number of flights (no failures after the first two), while there seems to be no relationship between the number of flights for the GTO orbit (ongoing failures).
• Payload Mass • Orbit • Landing Outcome	**Scatter Plot** with color representing landing outcome	Visualize the relationship between Payload Mass and type of Orbit.	With heavy payloads, successful landings are higher for Polar, LEO, and ISS orbits. However, GTO cannot be distinguished well, as successful and failed landings are both present in the visualization.
• Year • Success Rate	**Line Plot** showing landing success.	Visualize the launch success yearly trend.	The success rate steadily improved from 2013 to 2020, except for a notable dip in 2018.

- The completed EDA with data visualization notebook is available on GitHub.

EDA with SQL

- SQL queries performed to understand the data set were:
  - Query the unique launch site names
  - Query 5 records where the launch site name begins with “CCA”
  - Query the total payload mass for launches for NASA
  - Query the average payload mass for F9 v1.1 version boosters
  - Determine the date when the first successful ground pad landing was achieved
  - List the names of the boosters which have success in drone ships and have payload mass greater than 4,000 but less than 6,000 kg
  - List the total number of successful and failed mission outcomes
  - List the names of booster versions that carried the maximum payload mass.
  - Query the records and display the month, drone ship landing failures, booster versions, and launch site for 2015.
  - Rank the count of landing outcomes between 2010-06-04 and 2017-03-20 in descending order.

- The complete EDA with SQL notebook is available on [GitHub]

Build an Interactive Map with Folium

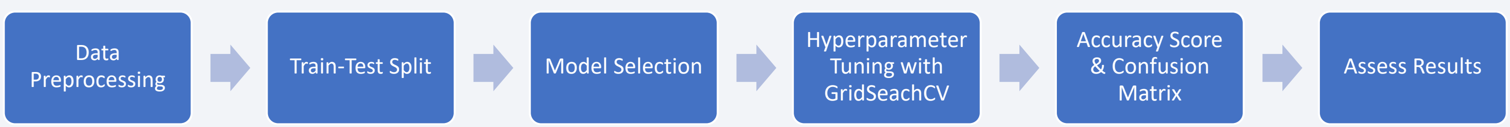
- Map objects (markers, circles, lines, etc.) were added to the Folium maps to add in visualization.
- Markers indicate launch pad locations, and Circles highlight the surrounding area on the Folium maps.
- Lines were added to indicate the distance from the launch pads to railroad tracks, other infrastructure, and shorelines.
- Marker clusters show the launches from each pad and color-coded to indicate successful (green) or failed (red) booster landings.
- These Markers, Circles, and Lines were added to help visualize launch sites, landing successes, and distances to essential features on the map.
- The complete Notebook with interactive Folium maps is available on GitHub

Build a Dashboard with Plotly Dash

- An interactive dashboard was built to assist with discovering insights from the SpaceX dataset more easily than with static graphs.
- The dashboard application contains the following elements:
  - A dropdown list to select individual or all launch pads
  - A Range slider to choose payload mass range in kgs.
  - A Pie Chart that shows the success statistics based on the site selection dropdown
  - A Scatter Plot that shows Landing success based on the selected payload mass range and by color-coded booster version.
- The dashboard was built to answer questions like “Which site has the highest success rate?” and “Which Booster version has the highest launch success rate?”
- The source code for the Plotly Dash interactive Dashboard is available on GitHub.

Predictive Analysis (Classification)

- Data prepared in earlier modules was used for predictive analysis. Initially loaded into a Pandas data frame and then transformed and standardized in a Numpy array.
- The data was then split into training and testing sets, with a 20% holdout set for testing the models after training.
- To determine which classification model would provide the best prediction accuracy, multiple models were tested using hyperparameter optimization with GridSearchCV to fine-tune their performance.
- Model performance was assessed using a combination of a Confusion Matrix from accuracy scoring.

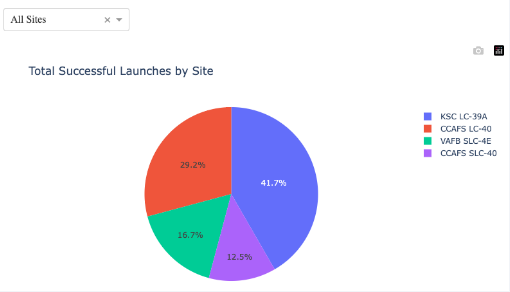


- A notebook that contains the full analysis is available on GitHub.

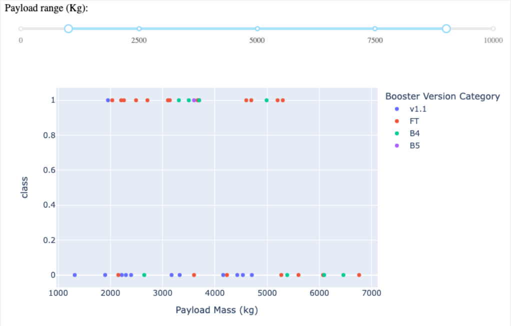
Results

- Exploratory data analysis results
  - SpaceX uses four launch pads
  - The total mass for NASA payloads was 45,596 kg
  - The average payload mass for Falcon 9 launches was 2,928.4 kg
  - The first successful booster landing occurred on December 22, 2015
  - Four boosters for launches with payloads of between 4,000 and 6,000 kgs successfully landed on the drone ship.
  - SpaceX has a mission success rate of 99 out of 101.
  - SpaceX has had mixed success with booster landings when a landing was attempted but has improved over time.
- Predictive analysis results
  - The predictive analysis showed that each model, when optimized, performed equally well as the other models. Each had an accuracy of 83.33%.

Interactive Dashboard Screenshots



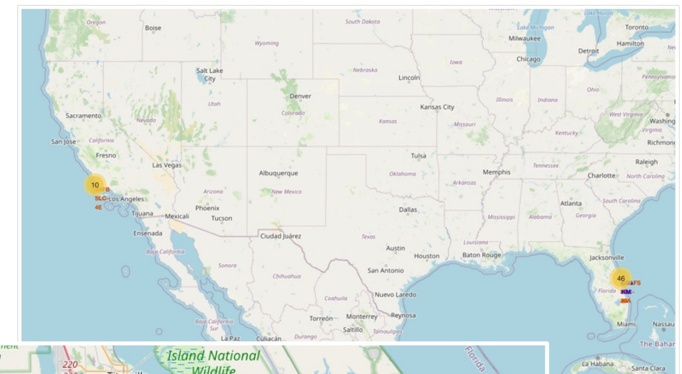
Pie Chart Showing Successful Launches By Launch Site

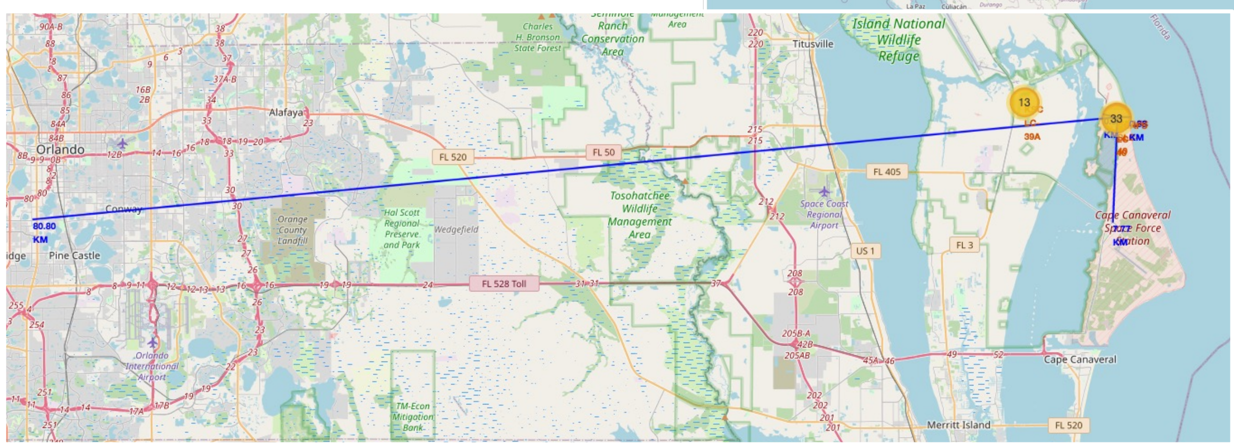


Scatter Plot that shows booster landing success versus payload mass and booster version.

Results

- Using geographic visualizations, it is possible to determine that launch sites are located to optimize safety, access to infrastructure, and as close to the equator as is reasonably achievable.
- For safety reasons, sites are located close to the shoreline and away from populated areas.
- Access to infrastructure includes being close to railroads and the shoreline for ocean transportation of large equipment and rocket components.
- Being close to the equator optimizes launch performance; thus, the sites are in the southern US.



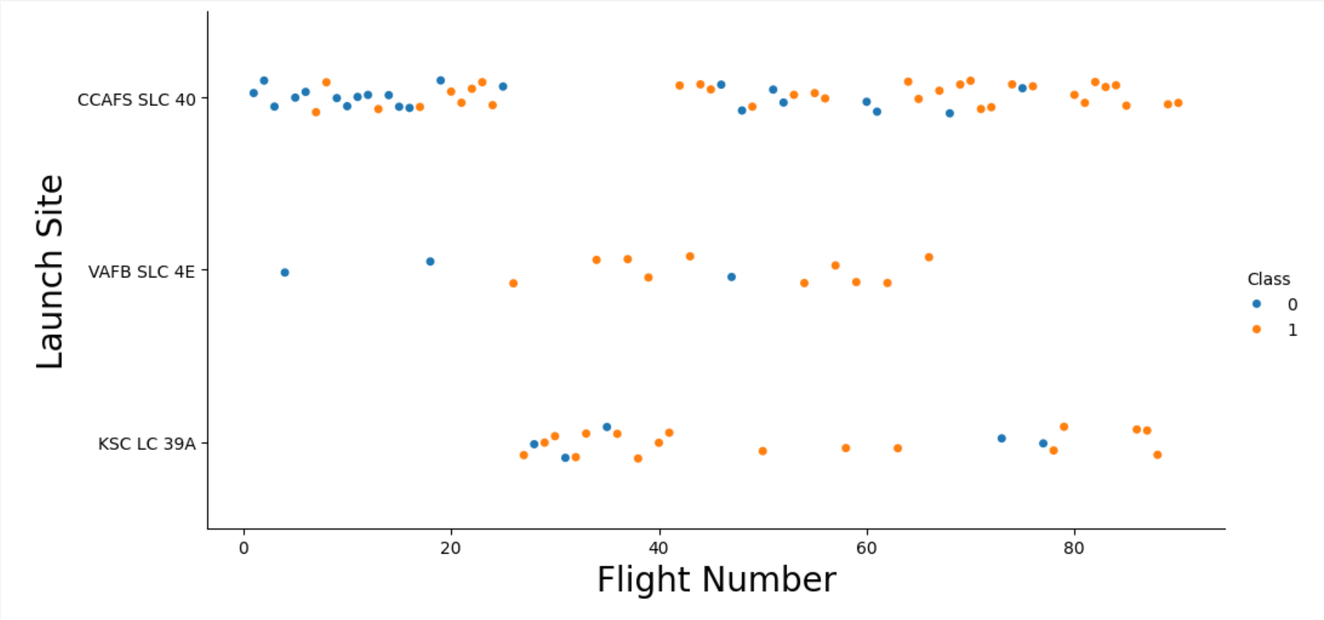


Section 2

Insights drawn from EDA

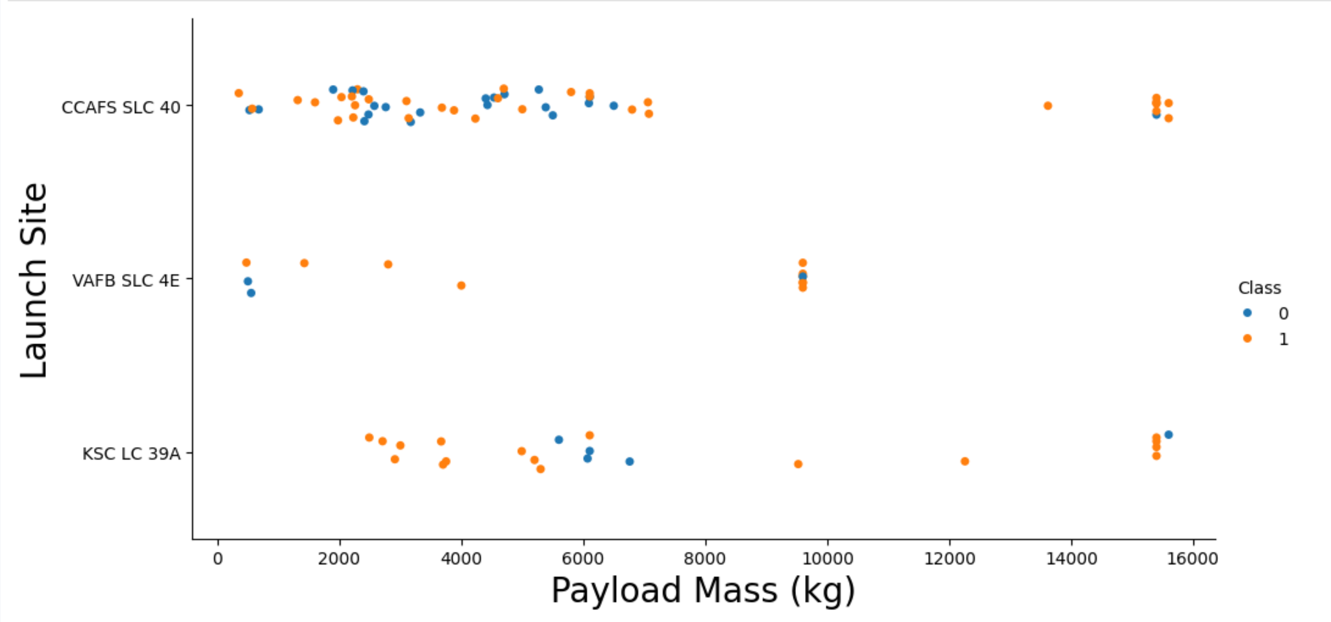
Flight Number vs. Launch Site

This plot shows that the success of booster landings increased as the number of flights increased.



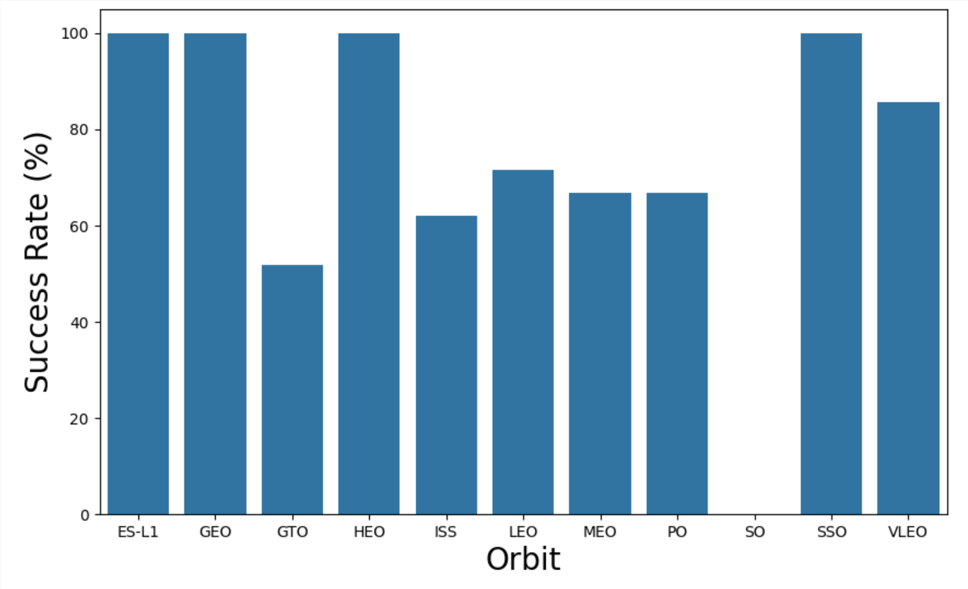
Payload vs. Launch Site

Two observations can be drawn from this plot. One, all launches from VAFB SLC 4E had a payload mass of less than 10,000 kg. Two, as the payload mass increases, so does the overall success of landing boosters (it's plausible that as SpaceX gained experience and confidence, they were able to increase payload mass along with increasing the success of booster recovery).



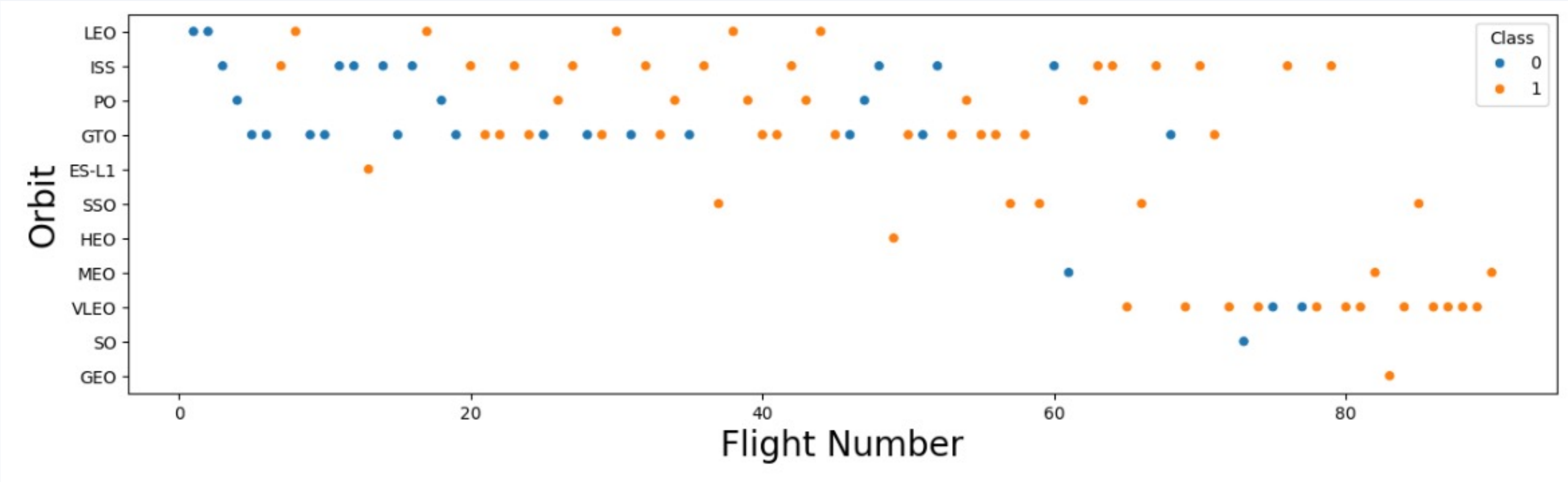
Success Rate vs. Orbit Type

From this plot we can see that ES-L1, GEO, HEO, and SSO had the highest success rates.



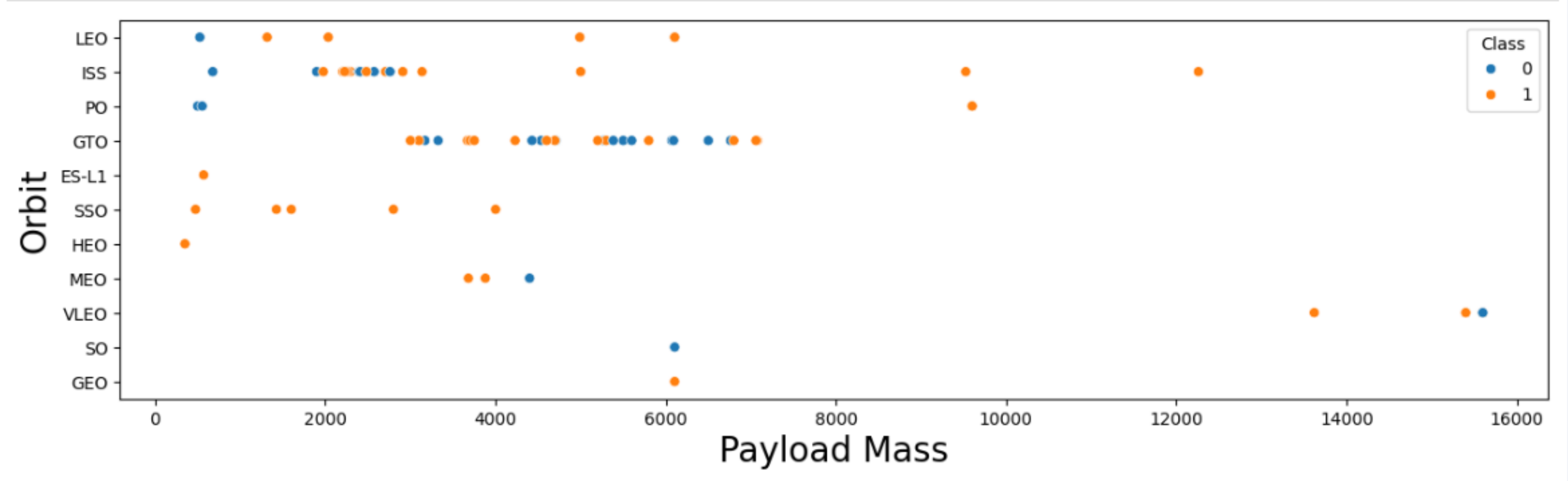
Flight Number vs. Orbit Type

This plot shows that over time, SpaceX transitioned from a program centered around LEO, ISS, PO, and GTO orbits to a significant increase in VLEO orbits. It also shows that the success rate has improved for LEO with the number of flights, whereas for GTO, it has not.



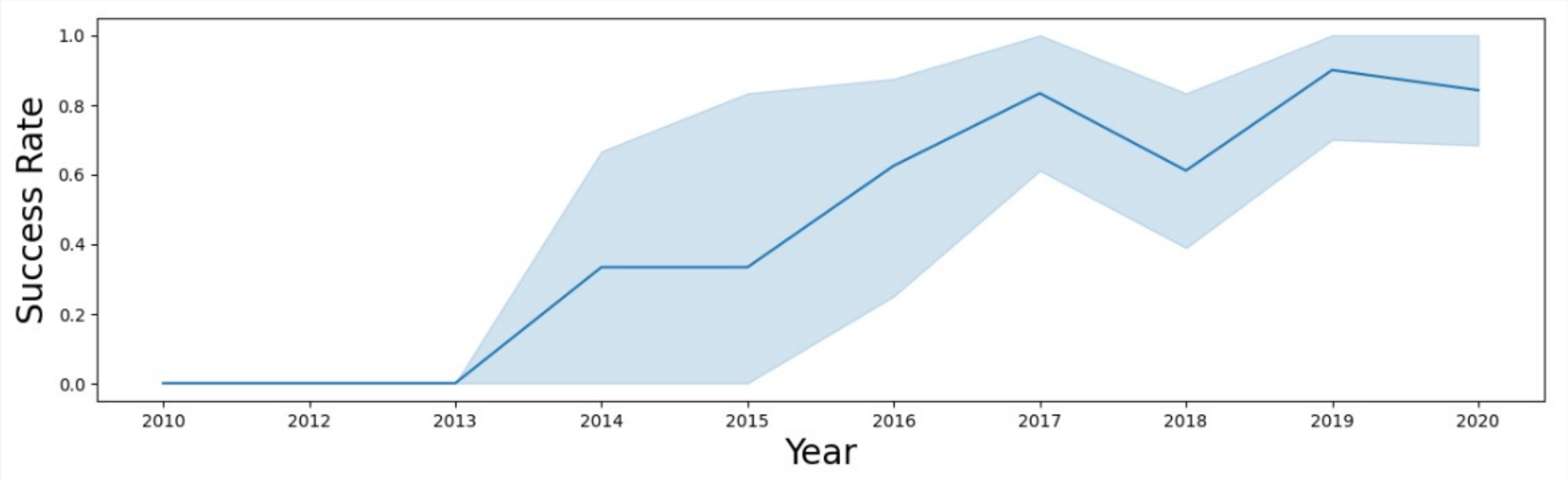
Payload vs. Orbit Type

With heavy payloads, landing success is greater for Polar, LEO, and ISS orbits. However, for GTO we cannot distinguish this as easily since both successful and failed landings are observed.



Launch Success Yearly Trend

This plot shows that since 2013, the success rate in landing boosters has steadily improved, with the notable exceptions of no improvement between 2014 and 2015 and a dip in the success rate in 2018.



All Launch Site Names

```
%sql SELECT DISTINCT("Launch_Site") FROM SPACEXTBL;

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

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

A SQL query was used to pull a set of unique ('distinct') launch site names. The query returned a list of 4 launch sites, as shown above.

Note: CCAFS LC-40 and CCAFS SLC-40 (aka "Slick 40") are the same pad. It was renamed at some point in the past.

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEXTBL WHERE "Launch_Site" like "CCA%" LIMIT 5;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04 00:00:00	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08 00:00:00	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 00:00:00	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 00:00:00	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01 00:00:00	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

A SQL query was written to list five examples of launches that occurred at Launch Sites whose name began with ‘CCA.’ The query and its results are shown above.

Total Payload Mass

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "Customer" = "NASA (CRS)";
```

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

SUM("PAYLOAD_MASS__KG_")
45596

A SQL query was written to calculate the total payload mass carried by launches for NASA, as a customer of SpaceX. The query and result of 45,596 kg are shown above.

Average Payload Mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1

```
%sql SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "Booster_Version" = "F9 v1.1";
```

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

AVG("PAYLOAD_MASS__KG_")
2928.4

A SQL query was built to calculate the average payload mass for all launches in the data set. The query and the result of the average payload mass of 2,928.4 kg are shown above.

First Successful Ground Landing Date

```
%%sql
SELECT MIN(date(Date)) AS "First Successful Ground Pad Landing"
FROM SPACEXTBL WHERE "Landing_Outcome" = "Success (ground pad)";
```

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

First Successful Ground Pad Landing
2015-12-22

A SQL query was used to find the launch date of the first successful booster landing, December 22, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

The names of boosters that have successfully landed on a drone ship with a payload mass greater than 4000 but less than 6000 kg were pulled using the SQL query shown to the right, along with the results.

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%%sql
SELECT DISTINCT("Booster_Version")
FROM SPACEXTBL
WHERE ("PAYLOAD_MASS__KG_" > 4000)
and ("PAYLOAD_MASS__KG_" < 6000)
and "Landing_Outcome" = "Success (drone ship)";
```

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

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

Total Number of Successful and Failure Mission Outcomes

- Here is the query and the results of a request to list mission successes and failures. As you can see, with a record of 1 in flight Failure for 101 attempts, SpaceX runs a highly effective program.

List the total number of successful and failure mission outcomes

```
%%sql
SELECT "Mission_Outcome", count(*)
FROM SPACEXTBL
GROUP BY "Mission_Outcome";
```

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

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

Boosters Carried Maximum Payload

Using a subquery, we determined the maximum payload mass (15,600 kg) and then pulled a list of the booster names for launches that carried that mass.

The list of the 12 boosters, their names, and the mass carried are listed to the right, along with the query that generated the results.

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%%sql
SELECT "Booster_Version", "PAYLOAD_MASS__KG_"
FROM SPACEXTBL
WHERE "PAYLOAD_MASS__KG_" = (
    SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTBL
);
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

Using a SQL query, we can quickly pull a list of records that display the month of failed landings on drone ships and the booster versions and launch site for launches in 2015.

```
%%sql
SELECT strftime("%Y", Date) as Year,
    strftime("%m", Date)as Month,
    "Landing_Outcome",
    "Booster_Version",
    "Launch_Site"
FROM SPACEXTBL
WHERE strftime("%Y", Date) = "2015"
    and "Landing_Outcome" = "Failure (drone ship)"
ORDER BY Date
```

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

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

Here, we're ranking the count of landing outcomes between 2010-06-04 and 2017-03-20 in descending order.

To complete this, we filtered the launches by the required date range, then grouped the "Landing Outcomes" and sorted them by the number of Outcomes in descending order.

```
%%sql
SELECT Landing_Outcome, count(Landing_Outcome) as Count
FROM SPACEXTBL
WHERE Date BETWEEN "2010-06-04" and "2017-03-20"
GROUP BY Landing_Outcome
ORDER BY count(Landing_Outcome) desc
```

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

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

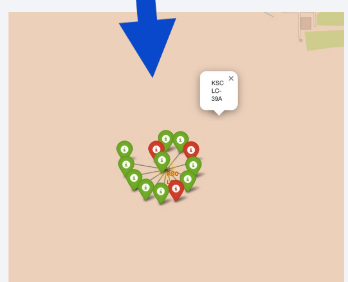
Section 3

Launch Sites Proximities Analysis

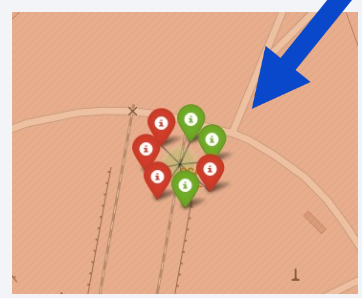
Map of SpaceX Launch Sites

SpaceX Launch Pads and Booster Landing Success Rates

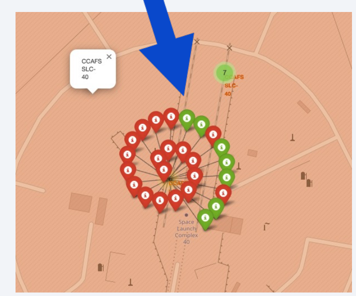
Here’s a visual representation of the four SpaceX launch pads and markers showing the success or failure to land booster stages. **Red** indicates failure, and **Green** indicates success.



KSC LC-39A



CCAFS LC-40



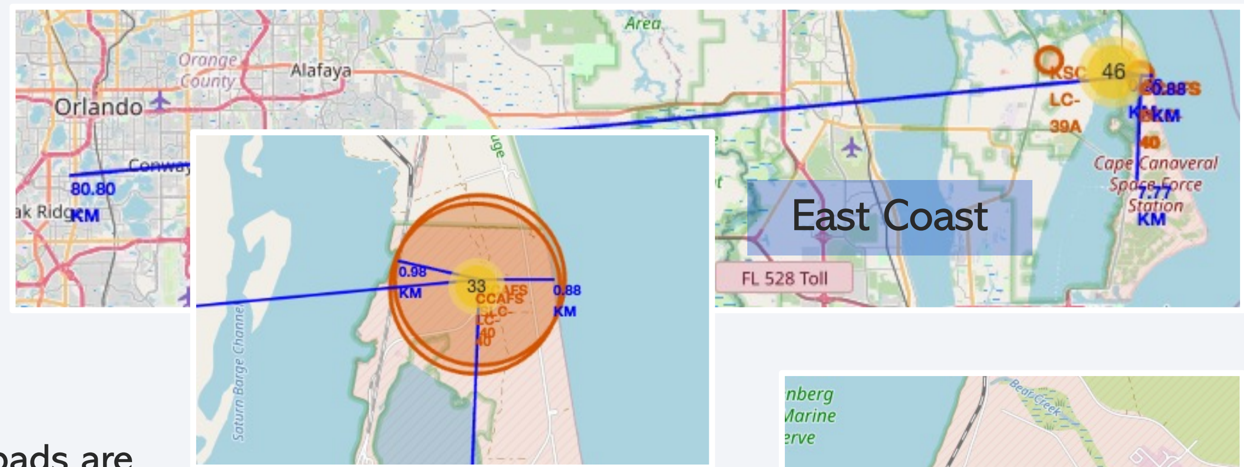
CCAFS SLC-40

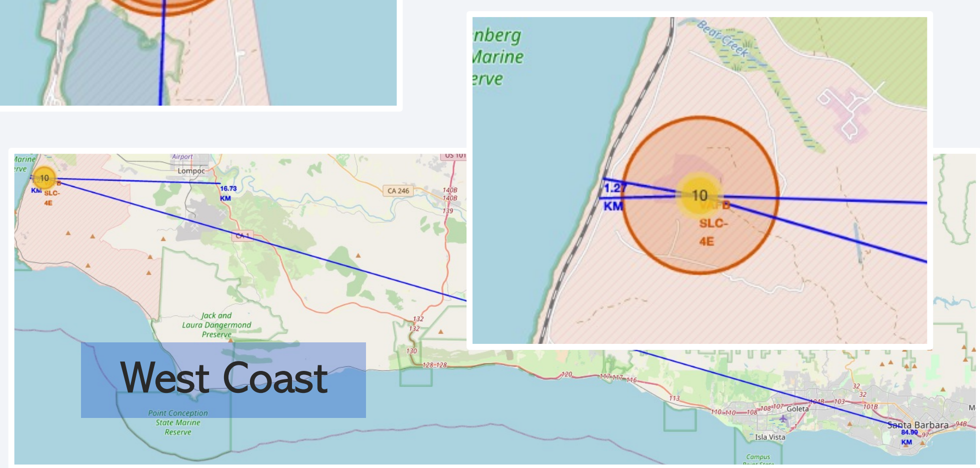
**WEST COAST LAUNCH PAD
CALIFORNIA**

VAFB SLC-4E

SpaceX Launch Site distances to key landmarks

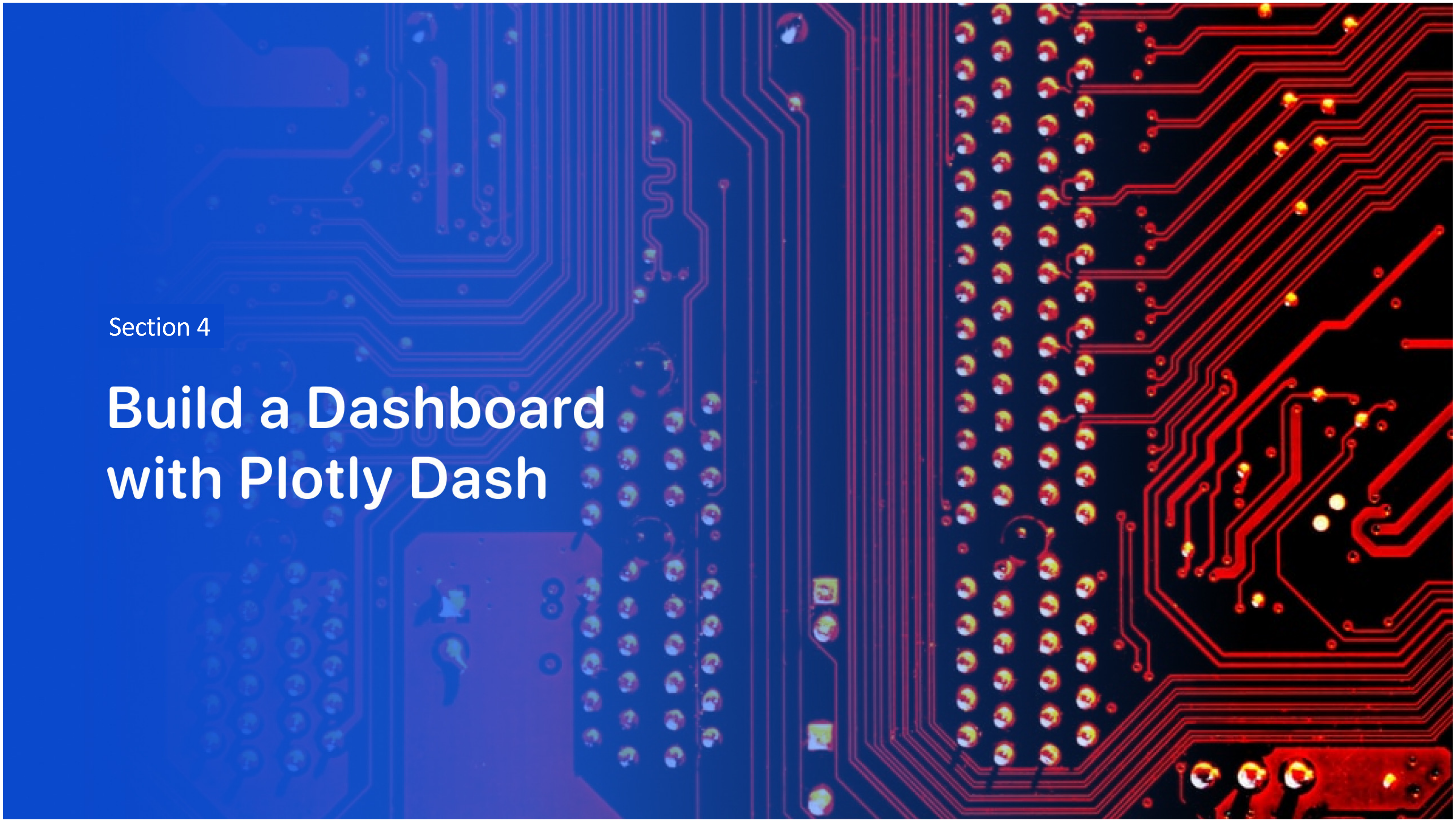
These maps show the location of one east coast and one west coast launch pad and their proximity to key landmarks.





As you can see, the launch pads are close to the shoreline and railroads within 1 to 2 km for safety, rail, and ocean transportation needs.

However, the pads are further away from highways and cities (10-100 km) due to safety concerns.

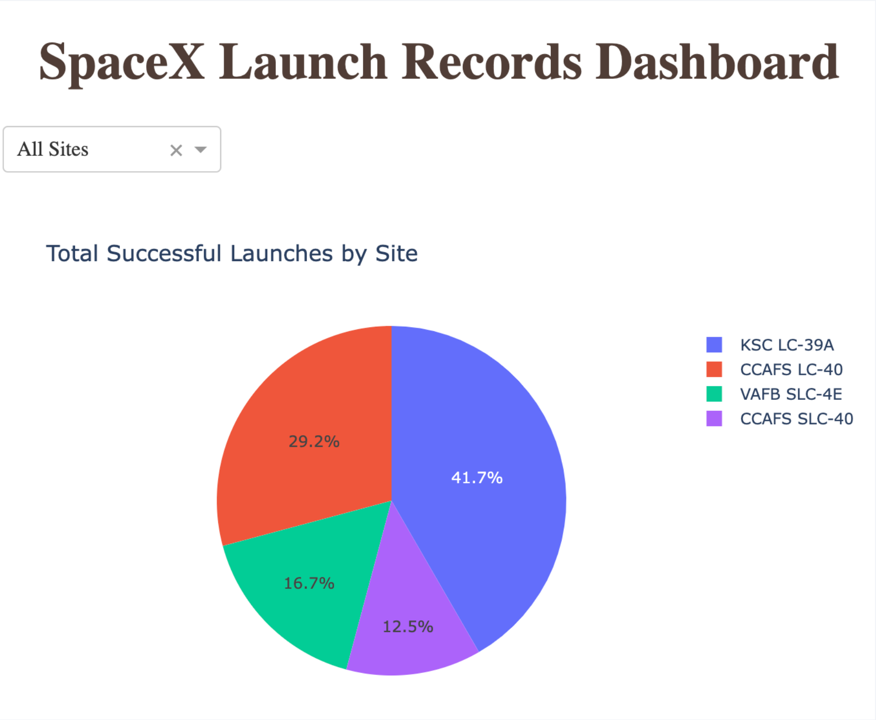

Section 4
Build a Dashboard
with Plotly Dash

SpaceX Launch Success Dashboard

This screenshot from the SpaceX dashboard shows the launch success for all launch sites.

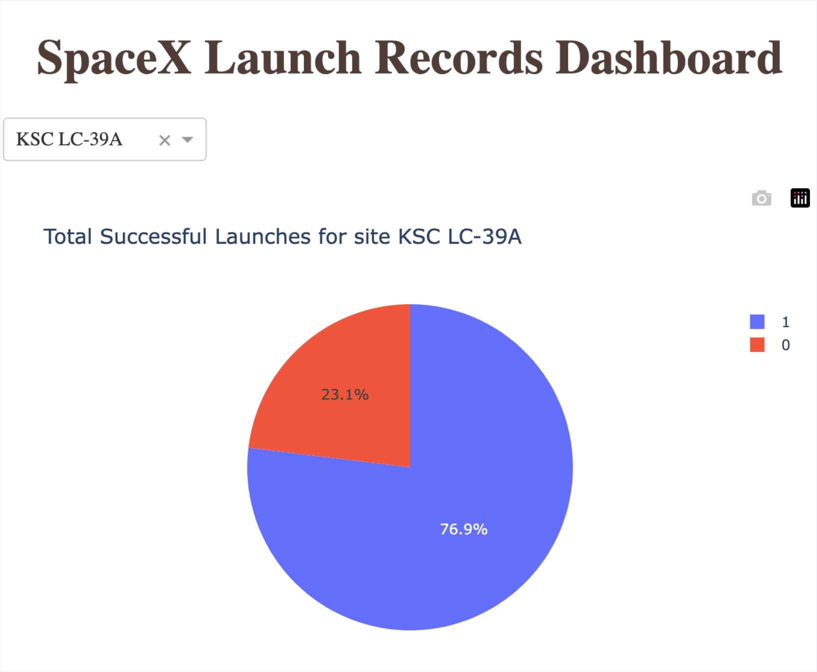
An interactive dashboard to explore this data is helpful since it's possible to drill down into the details and gain additional insights.

As you can see, KSL LC-39A has the highest success rate while CCAFS SLC-40 has the lowest.



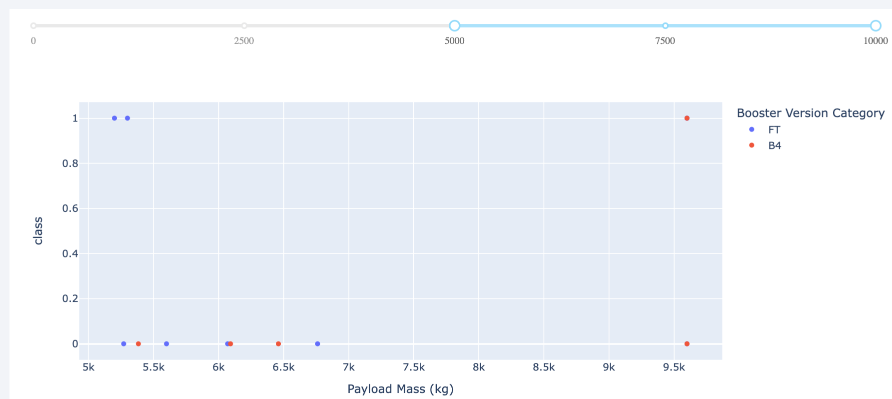
SpaceX Dashboard Drill Down

Continuing from the previous slide, we can drill down into the details for the most successful pad, KSC LC-39A, to explore additional information about this launch site's success and failure rates.



SpaceX Dashboard Success Rates for Different Payload Ranges

- The top scatter plot shows launch success and failure rates with payload masses less than 5,000 kg. The bottom plot shows the results for 5,000 kg or greater payload masses.
- The class indicates 1 for success and 0 for failure. The color code represents various booster versions.
- As these plots show, there are significantly greater successes for lower mass launches (only three successes for high mass launches).
- Although not demonstrated here, clicking on the booster to filter further the data displayed by the version number is possible.



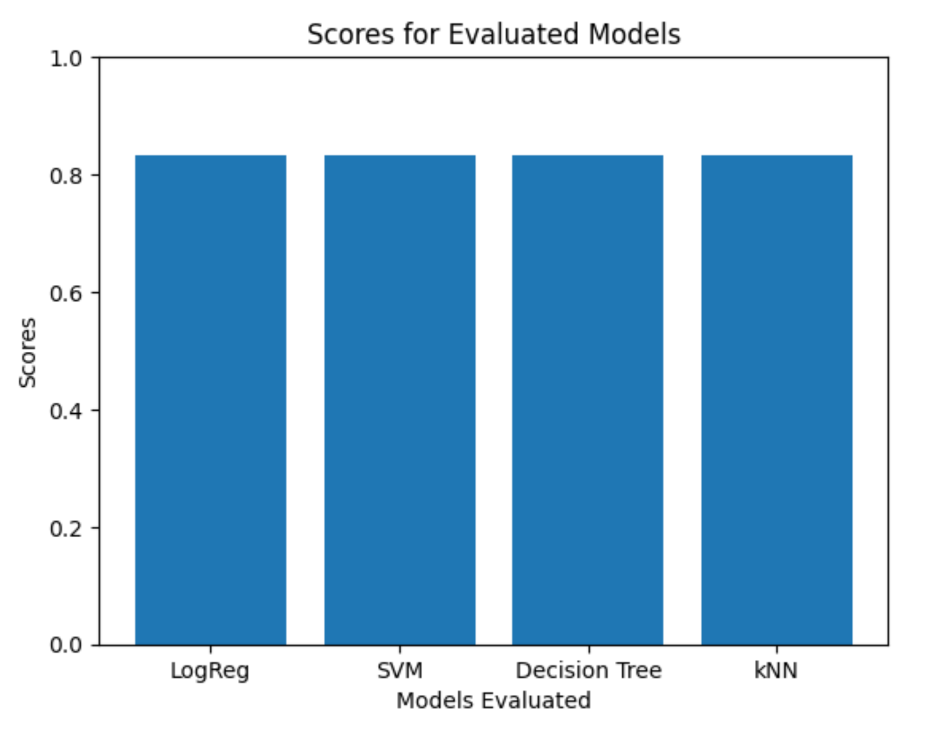
Section 5

Predictive Analysis (Classification)

Classification Accuracy

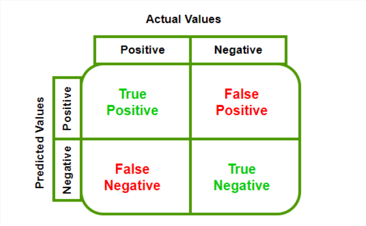
For the analysis, all models performed similarly. The bar plot shown to the right is the same for each model (Logistic Regression, SVM, Decision Tree, and KNN).

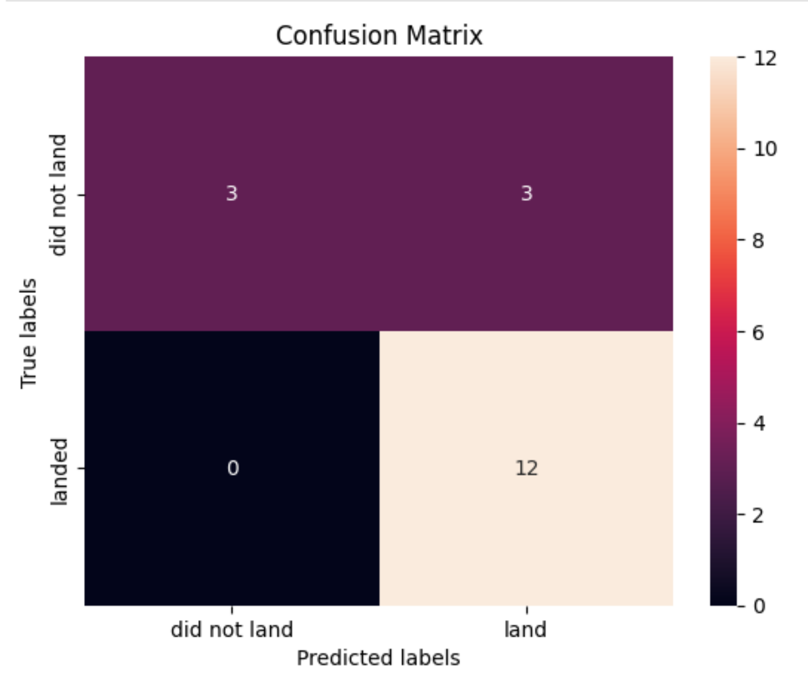
The optimized models all performed fairly well.



Confusion Matrix

- Interpreting this confusion matrix (see the key provided below), we can see that the matrix shows a high rate of “false positives” while there are no “false negatives.”
- The models predict successful landings more accurately than failed landings.





Conclusions

- SpaceX has built a strong program allowing them to land and reuse rocket boosters successfully. Because of this, they can keep costs low for customers.
- From our analysis using available data, we can predict with certainty if a booster landing will be successful.
- In exploring the data and visualizations, we also learned a lot about SpaceX's criteria for positioning launch sites.
- Launch success rates have been improving steadily over time since 2013.
- Booster landing success is most frequent with ES-L1, GEO, HEO, SSO, and VLEO orbits.
- All of the machine learning models, once optimized, performed similarly.


Thank You!