IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Richard M. Delia
August 8, 2023

Outline

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

Executive Summary

Summary of Methodology

The methodology used to analyze the SpaceX rocket landing data is systematic, beginning with data collection, then data wrangling where we look for patterns in the data, followed by exploratory data analysis where we extract and present meaningful information via graphs and charts and SQL queries. Lastly, we performed machine learning to determine a model and factors that lead to successful landings.

Summary/Key Take-Away of All Results

By analyzing the SpaceX launch data we were able to determine the factors that lead to successful landings which we can use to build rockets that should have higher landing success rates. Some example results of the analysis include the following:

- Landings are progressively more successful with subsequent launches.
- Missions with certain orbit destinations have higher rates of success.
- The DecisionTreeClassifier is the recommended model with the highest successful predictive landing accuracy.
- Rockets with large payloads have higher rates of success.
- The FT booster has a very high success rate when compared to other boosters.

Introduction

Project Objective

Our objective is to identify the key factors that lead to successful rocket landings using data from the rocket launch history of SpaceX. This information will be used by our organization to build rockets that have a high probability of successful landing at a price that is competitive with SpaceX.

Problems to Answer

- Where and how do we collect the data we need?
- What variables have an impact on successful landings, and to what degree?
- How do we create a model to interpret and process the data into meaningful information in order to accurately predict successful landings?

Section 1

Methodology

Methodology

Executive Summary

- Data Collection Methodology

  The raw data was collected via the SpaceX Launch history API. Data was also scraped from the SpaceX website.

- Perform Data Wrangling

  This is where analysis was done to detect patters to determine the best variables that leads to successful landings.

- Perform Exploratory Data Analysis (EDA) using Visualization and SQL

  Plots and graphs were created visually to help understand relationships among variables that impact landing success. SQL was also used to calculate and find information to support the analysis.

- Perform Interactive Visual Analytics using Folium and Plotly Dash

  An interactive dashboard was created to allow the user to select launch sites. Updated success / failure rates and the boosters used are shown. A payload filter can be used as well.

- Perform Predictive Analysis using Classification Models

  Machine learning was used to train the system to identify models to predict successful landings with a high level of accuracy.

Data Collection

Datasets were collected via the SpaceX REST APIs. This information was consolidated for analysis using Jupyter notebooks and the Python programming language.

Flowchart Details:

1. Identified SpaceX API as the Dataset source.
2. Load the Launch History dataset.
3. Normalize, clean, and convert Launch History to a dataframe.
4. Load and clean the Rocket Booster, Launch Site, Payload, and Core datasets.
5. Consolidate all datasets and select data to a new Launch dataframe.
6. Filter the Launch dataframe to only include Falcon 9 rockets.
7. Remove and average missing values in the finalized Launch dataframe.

Data Collection - SpaceX API

Data Type	Data Gathered	Dataset Source
SpaceX Rocket Info	Booster Names	https://api.spacexdata.com/v4/rockets/
SpaceX Launchpads	Launchpad Names & Locations	https://api.spacexdata.com/v4/launchpads/
SpaceX Payloads	Mass of the Payload & Orbit Type	https://api.spacexdata.com/v4/payloads/
SpaceX Cores	Rocket Core including physical attributes, landing type, if successful, # of flights & landing pads used	https://api.spacexdata.com/v4/cores/
SpaceX Launch History	Detailed data of previous launches	https://api.spacexdata.com/v4/launches/past

GitHub URL of Notebook: https://github.com/rdeliaus/testrepo/blob/main/Lab%201%20-%20Collecting%20the%20Data.ipynb

Data Collection - Scraping

The Space X Falcon 9 and Falcon Heavy launch records were scraped from their website. This was done by reading the HTML launch data using Python and converting it to a dataframe.

1. Read Webpage → 2. Parse Results → 3. Extract Launch Table → 4. Parse Launch Table Data → 5. Create the Dataframe

Flowchart Details:

1. Read the SpaceX webpage via HTTP and store the response.
2. Parse the result using the BeautifulSoup library so the data can be readily searched.
3. Locate the 3rd table in the HTTP response which contains launch data and extract the column headers.
4. Parse the launch data into a data dictionary.
5. Create a dataframe from the extracted launch table.

GitHub URL of Notebook:
https://github.com/rdeliaus/testrepo/blob/main/Web%20Scraping%20Falcon%209%20and%20Falcon%20Heavy%20Launches.ipynb

Data Wrangling

Exploratory data analysis was performed to detect patterns and to determine labels to use for training a model. New data values were created to represent landing success and failures to facilitate additional data analysis.

Flowchart Details:

1. Load the dataset from our data scraping step into a dataframe.
2. Calculate the number of launches for each site.
3. Calculate the number and types of orbits.
4. Calculate the number of successful and unsuccessful landings by landing site type.
5. Associate the landing success/failure class/status to the dataframe.
6. Calculate the average landing success rate overall.

GitHub URL of Notebook:
https://github.com/rdeliaus/testrepo/blob/main/Landing%20Prediction%20Lab%202%20Data%20Wrangling.ipynb

EDA with Data Visualization

Charts were generated to visually facilitate the understanding of the relationships among data labels and how they impact landing success.

Chart Type	Purpose	Example Insights
Scatter Plot Charts	These were used to visually show relationships among data labels including weather landings were successful or not.	• Payload mass is larger with more flights. • Success rate is larger with more flights. • Most launches have payloads under 7000KG.
Bar & Line Charts	Both Bar and Line charts were used to show the relationship between the landing success rate and the year and orbit.	• Orbits with highest success rate are ES-L1, GEO, HEO, and SSO. • The number of successful landings typically increases with time, with a noticeable decline in 2018.

GitHub URL of Notebook: https://github.com/rdeliaus/testrepo/blob/main/EDA%20with%20Visualization.ipynb

EDA with SQL

The information extracted via SQL queries from the launch data included the following:

- A list of each launch site used
- A list of Cape Canaveral launch sites used
- The total payload (in kg) launched for NASA (CRS)
- The average payload (in kg) launched using the F9 v1.1 booster rocket
- The first successful landing date
- The list of all boosters that successfully landed on a drone ship with a payload between 4000-6000 kg
- The total number of success and failed missions
- The list of boosters that carried the maximum payload in kg
- The list of failures by booster versions, launch site, and by month in 2015
- A ranked list of landing outcomes between 6/4/2010 and 3/20/2017

GitHub URL of Notebook: https://github.com/rdeliaus/testrepo/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

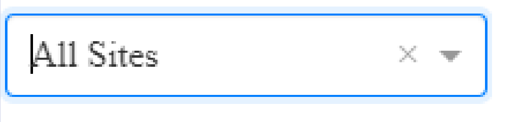
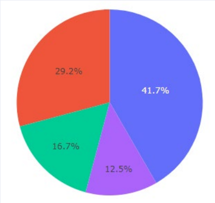
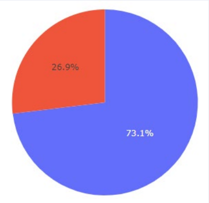
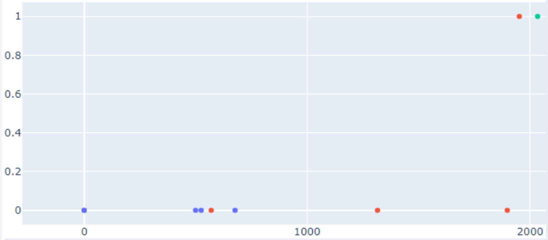
Folium maps are used to visually display the locations of launch sites, success and failures, as well as proximities to landmarks.

Map Objects	Purpose	Example Insights
Launch Site Map	This was used to create a map displaying all launch sites across the US.	• Visually see many launch sites in Florida. • Visually see most launch sites are near a coast.
Launch Site Markers	This is used to mark launch site coordinates, this can also include colors, and text labels.	• Visually see names associated with launch sites.
Circle Objects	Launch Sites are highlighted using circles so they can be readily seen.	• Visually assists in finding launch sites, and proximities to other landmarks.
Marker Clusters	This groups launch sites in close proximity to summarize launch success rates	• Visually see the number of launches by site. • Visually see the landing success rate by color.
Lines with Distance	Shows distance from a given launch site to other landmarks including oceans, cities, railways, etc.	• Most sites are within a few km of an ocean and a railway. • Most sites are at least 10 km from any city.

GitHub URL of Notebook: https://github.com/rdeliaus/testrepo/blob/main/Analysis%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

The following objects were used to create the interactive dashboard:

	Object	Purpose
	Launch Site Dropdown	Allows user to select which sites are in scope for all of the dashboard reporting objects.
 	Pie Chart	If all sites are selected, successful landings by each site is shown. If a specific site is selected the successful/unsuccessful values are shown.
	Payload Slider	Allows user to select the range of payload to filter the results of the Payload Mass and Launch Success of the scatter plot.
	Scatter Plot	Shows the Payload Mass (KG) by success rate, where 1=success & 0=failure. This also shows the booster type by color. This is filtered by the Launch Site Dropdown and the Payload slider.

GitHub URL of Notebook: https://github.com/rdeliaus/testrepo/blob/main/7-buildadashboard_spacex_dash_app.py

Predictive Analysis (Classification)

In order to predict successful landings we perform machine learning tasks including variations of regression modeling to select a model with the highest level of predictive accuracy.

Flowchart Details:

1. Load the datasets from our previous steps to create a new dataframe.
2. Standardize the data.
3. Split the data into a training and test subsets for modeling.
4. Perform various regression modeling techniques on the data.
5. Score the regression results.
6. Review Confusion Matrices to distinguish between the different classes and predicted vs. actual values.
7. Use the best score to select the optimal regression technique to use. (Decision Tree Classifier at 94.4%).

GitHub URL of Notebook: https://github.com/rdeliaus/testrepo/blob/main/Machine_Learning_Prediction.ipynb

Results

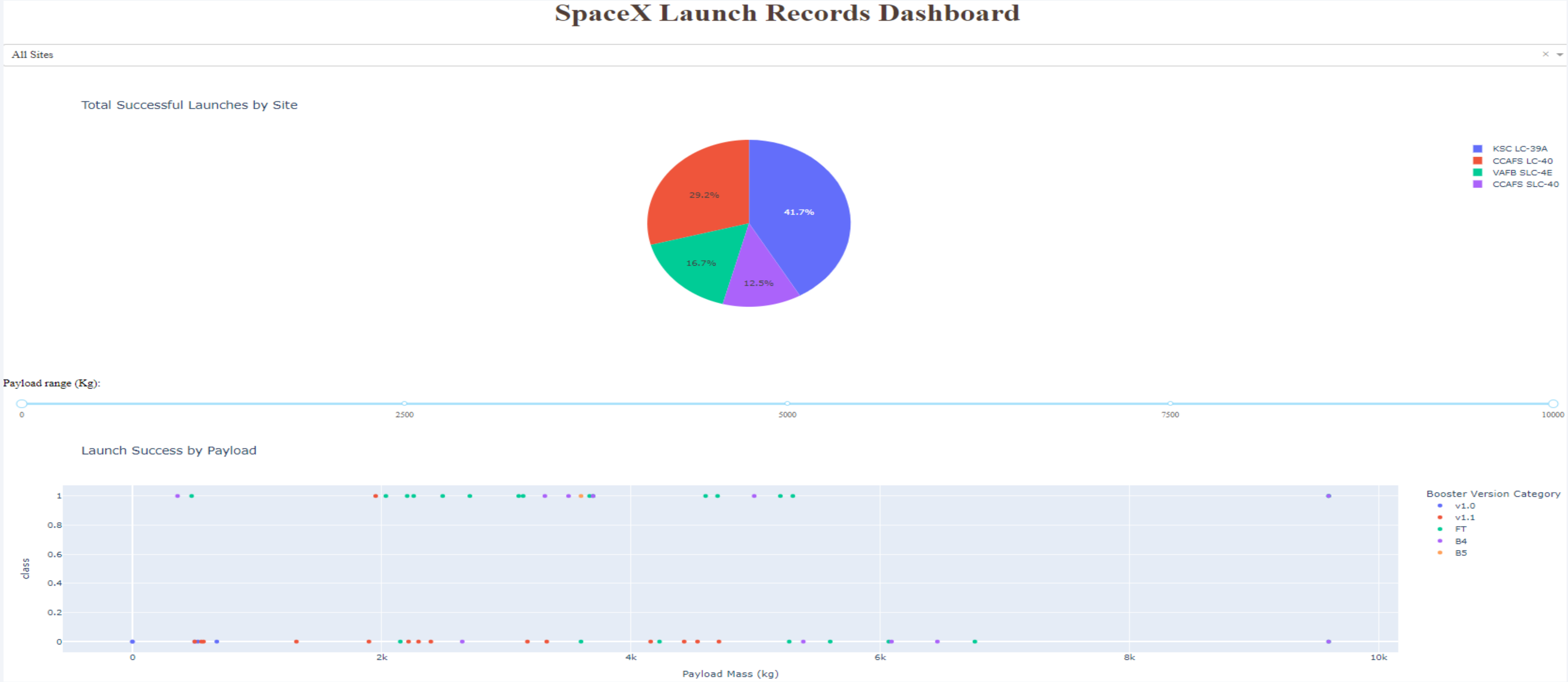
- Exploratory Data Analysis Results

  With Data Visualization we are able to see relationships among the data including the following:
  - Payload mass and success rate are greater with more flights
  - Most payloads are under 7000Kg.
  - Orbits with high success rates are ES-L1, GEO, HEO, AND SSO.
  - The rate of successful landings improves with additional launches.

  With SQL we are able to extract a wide variety of important information, some of which include the following:
  - Launch sites used.
  - Total payload by site.
  - Average payload for given rocket booster type.
  - Total successful and failed missions.
  - etc.

- Predictive Analysis Results
  - Using machine learning we identified the DecisionTreeClassifier Model to have the best predictive accuracy at 94.4%.

Results - Interactive Analytics Demo Screenshot

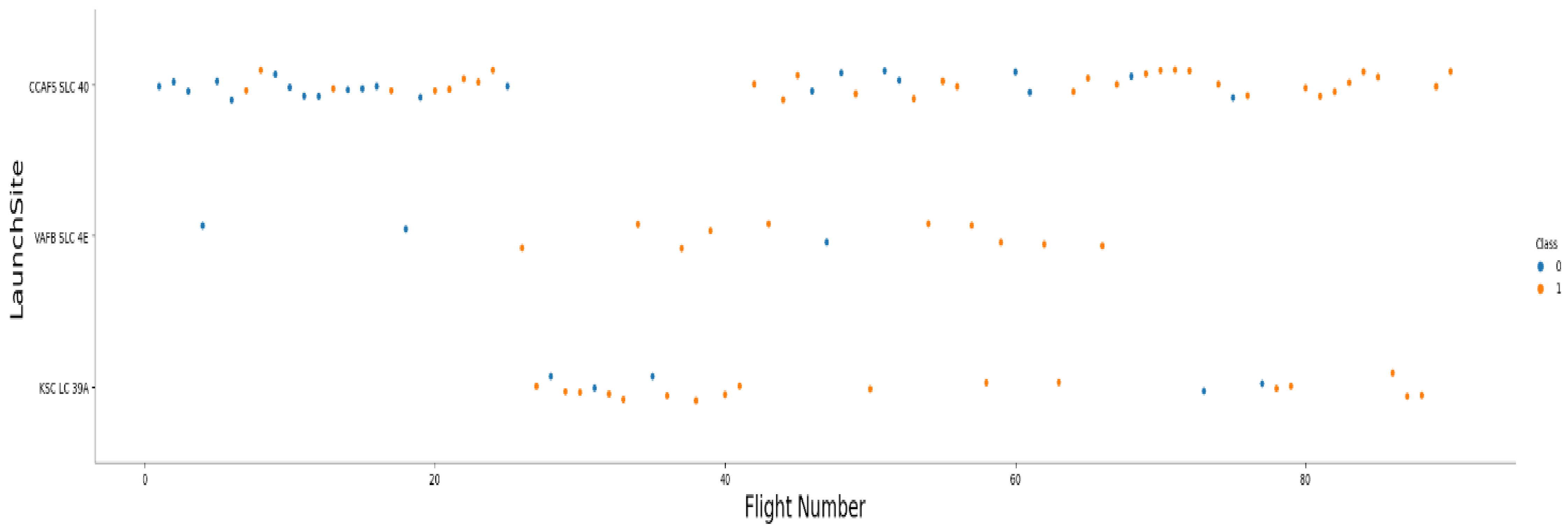


* See "Build a Dashboard with Plotly" section for more details and examples.

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

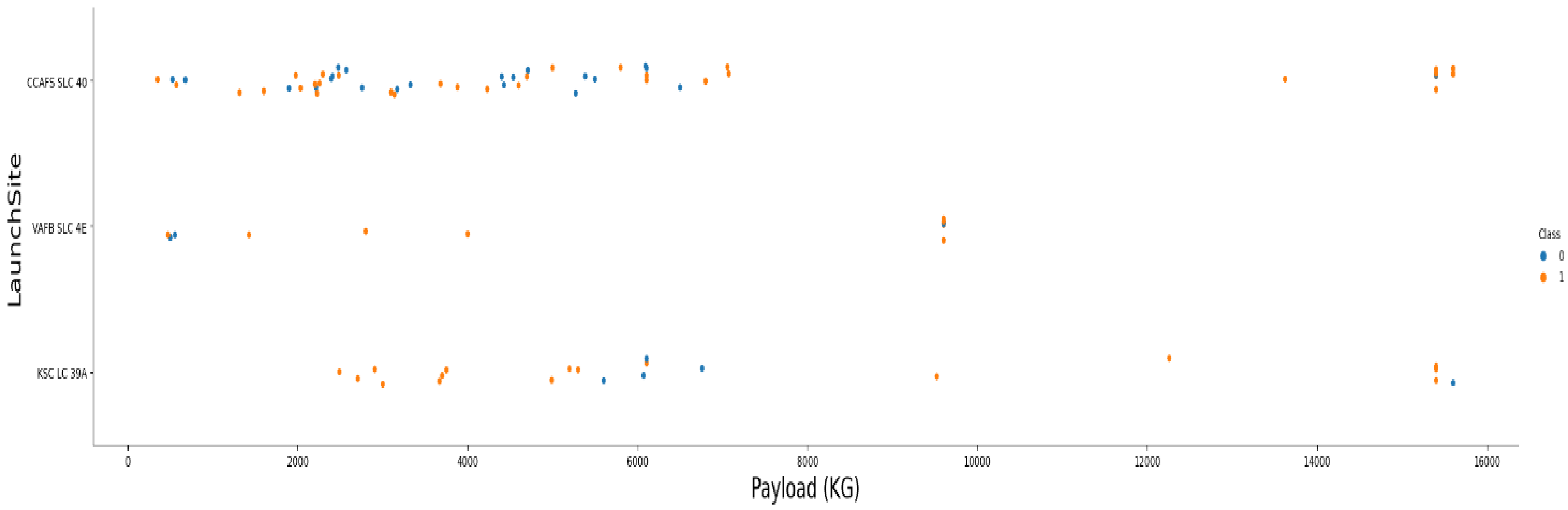


Class 0 (blue) = Failed
Class 1 (orange) = Success

From the graph we can see the following:

- CCAFS SLC 40 hosts the majority of launches
- As more rockets are launched, the success rate improves

Payload vs. Launch Site

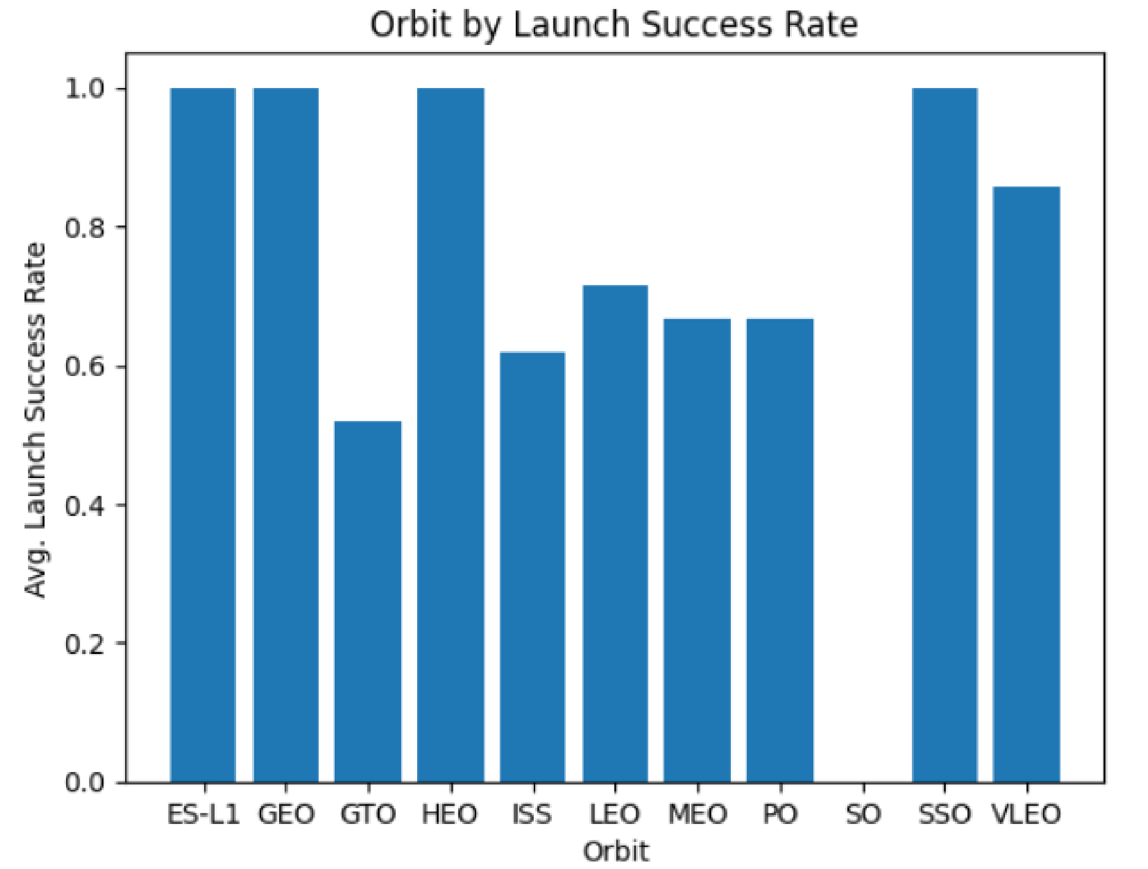


Class 0 (blue) = Failed
Class 1 (orange) = Success

From the graph we can see the following:

- Most launches have payloads under ~7,000 KG.
- VAFB-SLC has no payloads > 10,000KG.
- KSC LC 39A has a high failure rate around ~6,000KG.
- All rockets have a high success rate above ~9,000KG.

Success Rate vs. Orbit Type

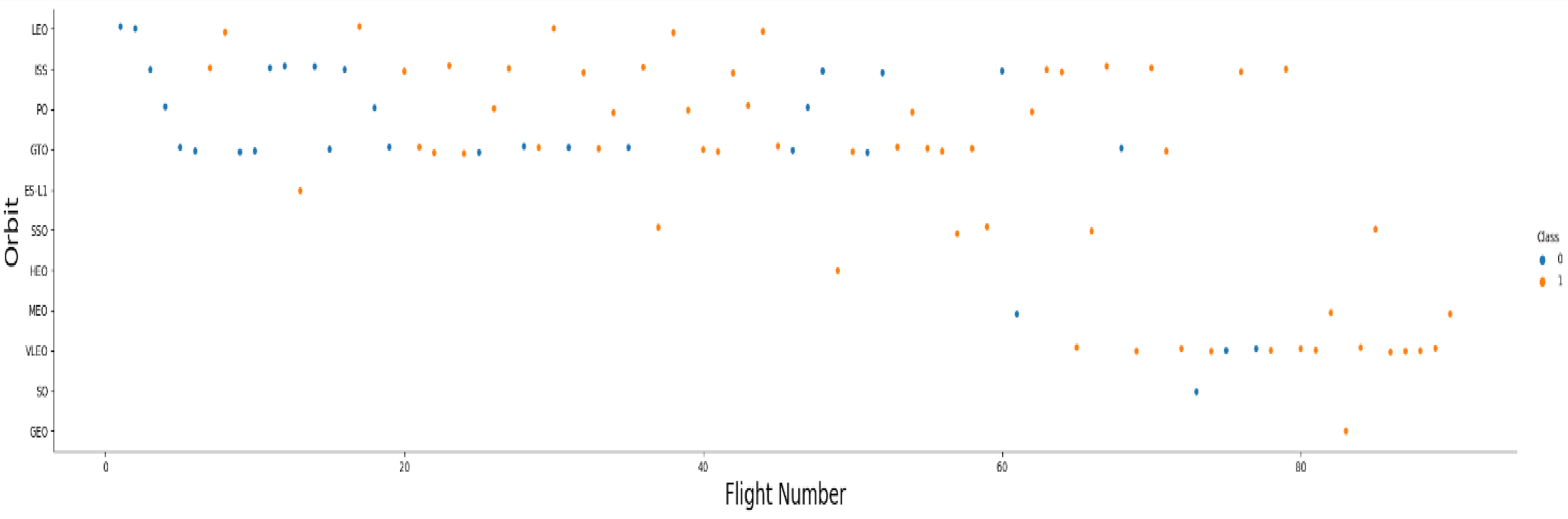


Class 0 (blue) = Failed
Class 1 (orange) = Success

From the graph we can see the following:

- ES-L1, GEO, HEO, and SSO have a 100% success rate (but these also had the lowest # of launches).
- GTO has the worst success rate at ~ 50%.
- No rockets were destined for an SO orbit.

Flight Number vs. Orbit Type

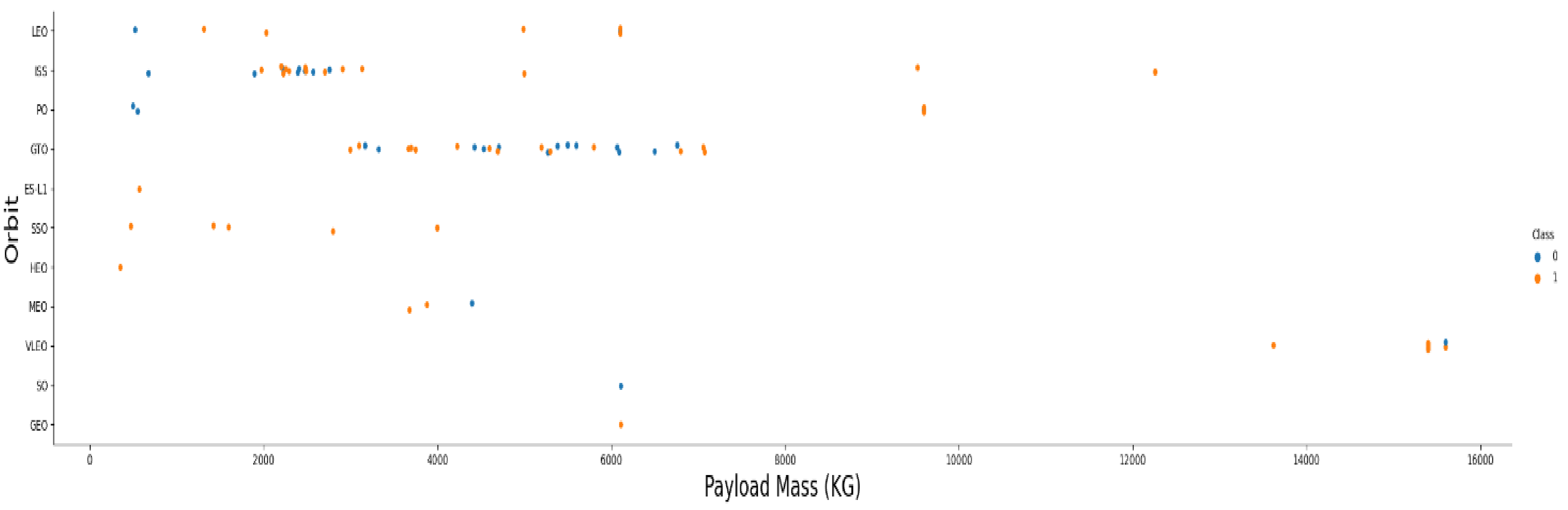


Class 0 (blue) = Failed
Class 1 (orange) = Success

From the graph we can see the following:

- The majority of the first 60 flights were for LEO, ISS, PO, AND GTO orbits.
- A high number of launches after 60 were for VLEO orbit rockets with a success rate of ~86%.
- In general, the more launches, the higher the success rate by orbit.

Payload vs. Orbit Type

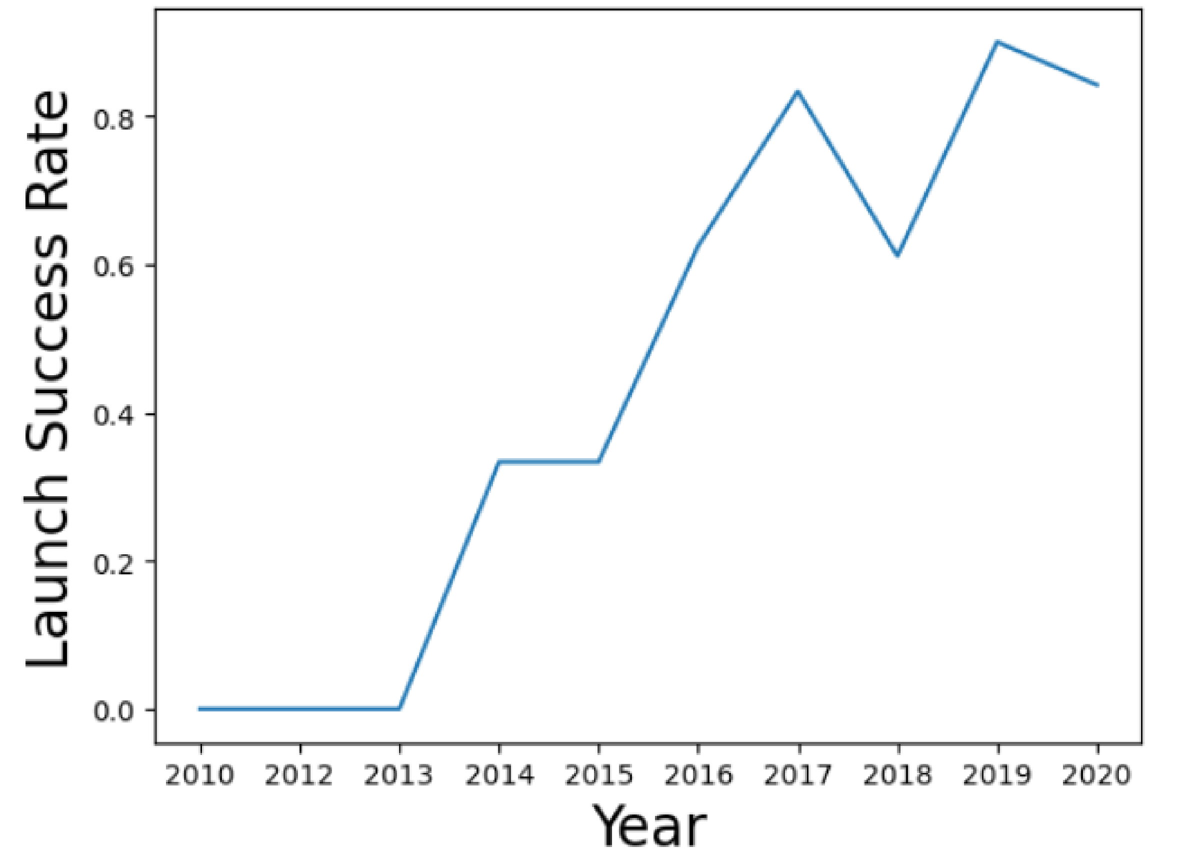


Class 0 (blue) = Failed
Class 1 (orange) = Success

From the graph we can see the following:

- Most Payloads are less than ~7,000KG.
- Most Payloads above ~7,000KG are successful.
- ES-L1, SSO, HEO, and MEO orbits all have payloads < 4,000KG with 100% success rates.

Launch Success Yearly Trend



From the graph we can see the following:

- Success rate for the first 3 years was zero.
- Overall, launch success rate increases over time.
- There is a noticeable rise in failures from 2017-2018.
- There is a downward success trend starting in 2019.

Class 0 (blue) = Failed
Class 1 (orange) = Success

All Launch Site Names

Launch Site Names	SQL Query Used	Query Description
Launch_Site CCAFS LC-40 VAFB SLC-4E KSC LC-39A CCAFS SLC-40	sql select distinct launch_site from spacextbl	This query selects all unique launch sites from the spacextbl to present a complete list of launch sites.

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

SQL Query Used	Query Description
sql select * from spacextbl where launch_site LIKE "CCA%" limit 5	This query selects all columns from the spacextbl where the launch site starts with letters CCA.

Total Payload Mass

Total Payload Mass	SQL Query Used	Query Description
Total Payload Mass (KG) 45596	sql select sum(PAYLOAD_MASS__KG_) as "Total Payload Mass (KG)" from spacextbl where Customer="NASA (CRS)"	This query totals the payload mass from the spacextble for the customer NASA CRS.

Average Payload Mass by F9 v1.1

Avg. Payload Mass	SQL Query Used	Query Description
Average Payload Mass (KG) 2928.4	sql select avg(PAYLOAD_MASS__KG_) as "Average Payload Mass (KG)" from spacextbl where Booster_Version = "F9 v1.1"	This query averages the payload mass from the spacextble for the F(v1.1 Booster rocket.

First Successful Ground Landing Date

Landing Date	SQL Query Used	Query Description
Date Landing_Outcome 2018-08-01 Success (ground pad)	sql select Date, Landing_Outcome from spacextbl where Landing_Outcome="Success (ground pad)" order by Date desc limit 1	This query selects the date and successful landing outcome from the spacextble sorted by ascending date and the result restricted to just 1 thus showing the first successful landing date.

Successful Drone Ship Landing with Payload between 4000 and 6000

<table>
<tr><th>Booster & Payloads</th><th>SQL Query Used</th><th>Query Description</th></tr>
<tr>
<td>
<table>
<tr><th>Booster_Version</th><th>PAYLOAD_MASS__KG_</th></tr>
<tr><td>F9 FT B1022</td><td>4696</td></tr>
<tr><td>F9 FT B1026</td><td>4600</td></tr>
<tr><td>F9 FT B1021.2</td><td>5300</td></tr>
<tr><td>F9 FT B1031.2</td><td>5200</td></tr>
</table>
</td>
<td>sql select Booster_Version,PAYLOAD_MASS__KG_ from spacextbl where Landing_Outcome="Success (drone ship)" and PAYLOAD_MASS__KG_>4000 and PAYLOAD_MASS__KG_<6000</td>
<td>This query selects the booster version and payload from the spacextble for successful landings on the drone ship and where the payload is between 4000-6000KG.</td>
</tr>
</table>

Total Number of Successful and Failure Mission Outcomes

Success & Failures	SQL Query Used	Query Description
Successes Failures 100 1	sql select sum(Mission_Outcome like "%Success%") as Successes, sum(Mission_Outcome like "%Failure%") as Failures from spacextbl	This query shows the total mission outcomes that are Successful and that are Failures from the spacextbl.

Boosters Carried Maximum Payload

<table>
<tr><th>Boosters & Payload</th><th>SQL Query Used</th><th>Query Description</th></tr>
<tr><td>
<table>
<tr><th>Booster_Version</th><th>PAYLOAD_MASS__KG_</th></tr>
<tr><td>F9 B5 B1048.4</td><td>15600.0</td></tr>
<tr><td>F9 B5 B1049.4</td><td>15600.0</td></tr>
<tr><td>F9 B5 B1051.3</td><td>15600.0</td></tr>
<tr><td>F9 B5 B1056.4</td><td>15600.0</td></tr>
<tr><td>F9 B5 B1048.5</td><td>15600.0</td></tr>
<tr><td>F9 B5 B1051.4</td><td>15600.0</td></tr>
<tr><td>F9 B5 B1049.5</td><td>15600.0</td></tr>
<tr><td>F9 B5 B1060.2</td><td>15600.0</td></tr>
<tr><td>F9 B5 B1058.3</td><td>15600.0</td></tr>
<tr><td>F9 B5 B1051.6</td><td>15600.0</td></tr>
<tr><td>F9 B5 B1060.3</td><td>15600.0</td></tr>
<tr><td>F9 B5 B1049.7</td><td>15600.0</td></tr>
</table>
</td><td>sql select Booster_Version, PAYLOAD_MASS__KG_ from spacextbl where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from spacextbl)</td><td>This query shows the booster version and payload from the spacextble where the payload is the maximum payload found in the spacextbl.</td></tr>
</table>

2015 Launch Records

Date	Year	Month	Landing_Outcome	Booster_Version	Launch_Site
01/10/2015	2015	10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
14/04/2015	2015	04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

SQL Query Used	Query Description
sql select date, substr(Date,7,4) as Year, substr(Date,4,2) as Month, landing_Outcome, Booster_Version, Launch_Site from spacextbl where landing_Outcome="Failure (drone ship)" and Year="2015"	This query parses the date field to obtain the year and month and the landing outcome, booster version, and launch site from the spacextble where failures occurred for the drone ship in the year 2015.

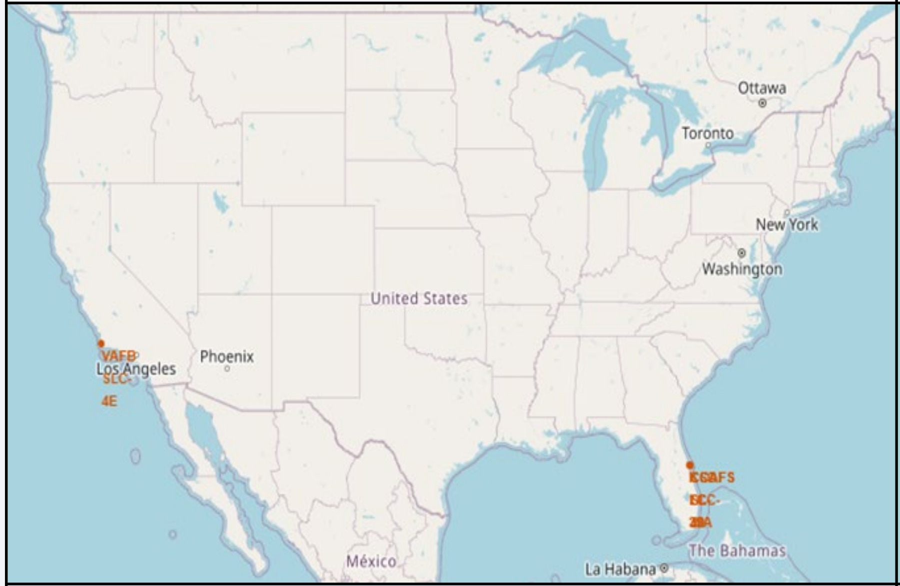
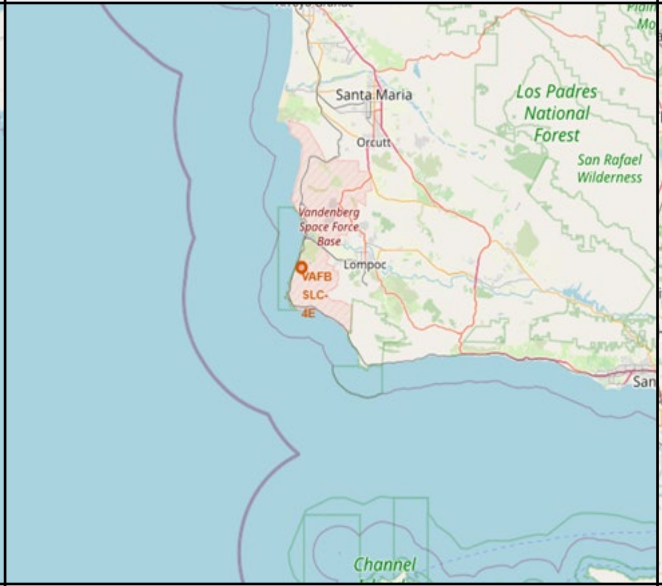
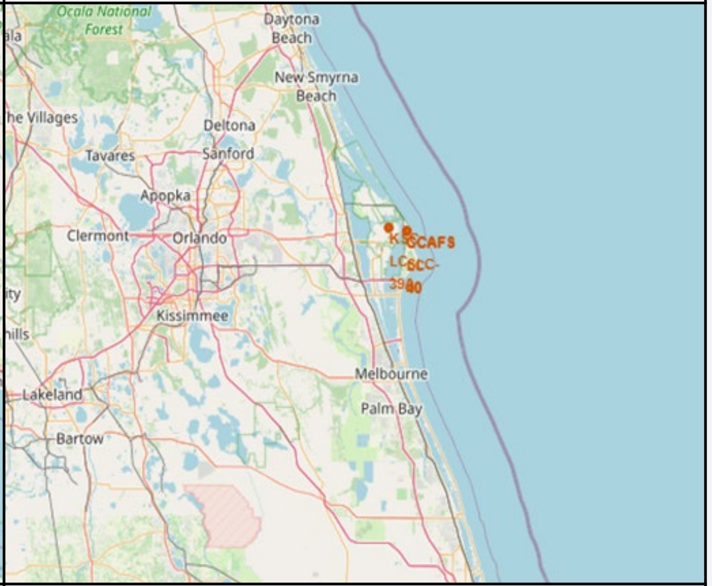
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

<table>
<tr><th>Landing Outcomes</th><th>SQL Query Used</th><th>Query Description</th></tr>
<tr>
<td>
<table>
<tr><th>Landing_Outcome</th><th>Count</th></tr>
<tr><td>Success</td><td>20</td></tr>
<tr><td>No attempt</td><td>10</td></tr>
<tr><td>Success (drone ship)</td><td>8</td></tr>
<tr><td>Success (ground pad)</td><td>7</td></tr>
<tr><td>Failure (drone ship)</td><td>3</td></tr>
<tr><td>Failure</td><td>3</td></tr>
<tr><td>Failure (parachute)</td><td>2</td></tr>
<tr><td>Controlled (ocean)</td><td>2</td></tr>
<tr><td>No attempt</td><td>1</td></tr>
</table>
</td>
<td>sql select Landing_Outcome, count(Landing_Outcome) as Count from spacextbl where Date>"04-06-2010" and Date<"20-03-2017" group by Landing_Outcome order by Count desc</td>
<td>This query shows the landing outcome and count from the spacextble where the dates are between 4/6/2010 and 3/20/2017. The outcomes are grouped to show the totals and a total count sorted by the count in descending order.</td>
</tr>
</table>

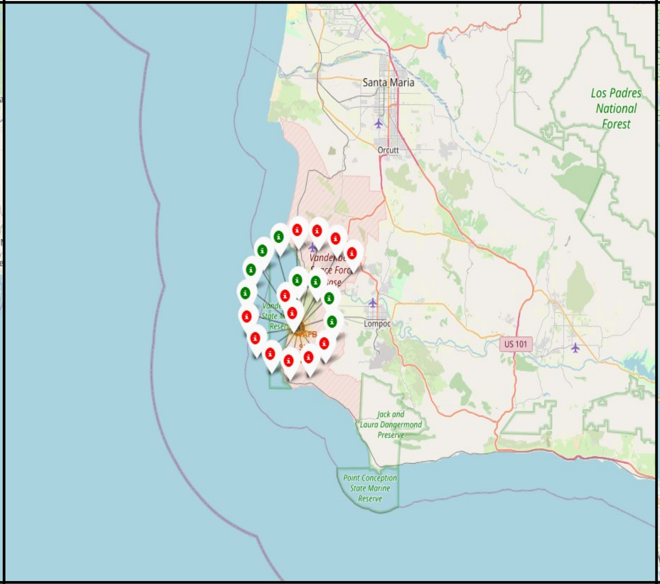
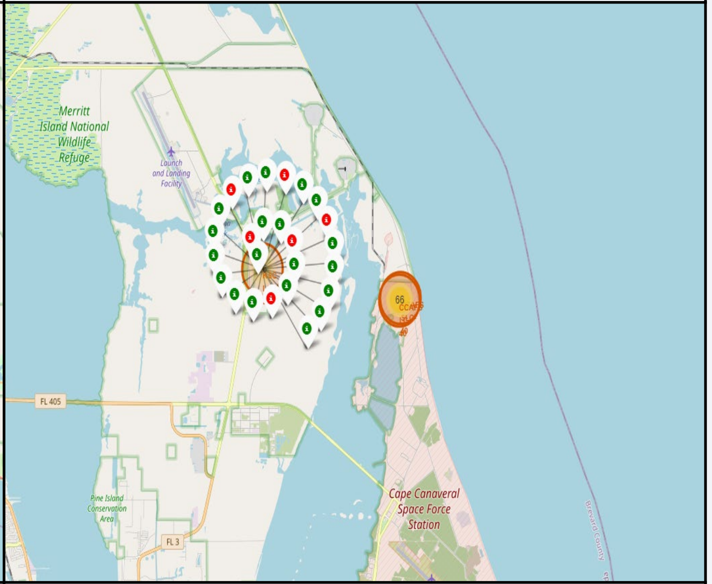
Section 3

Launch Sites Proximities Analysis

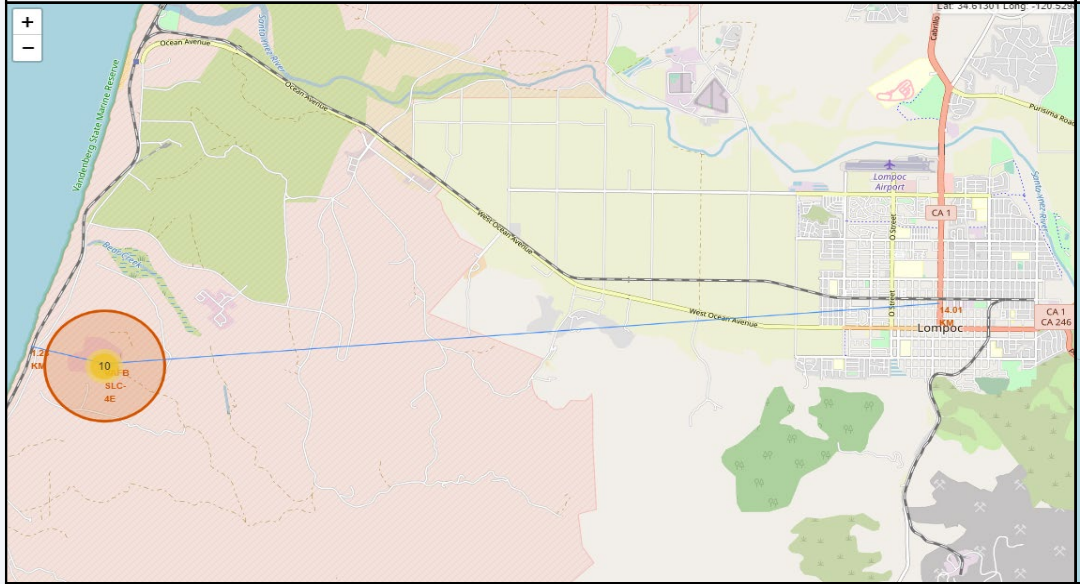
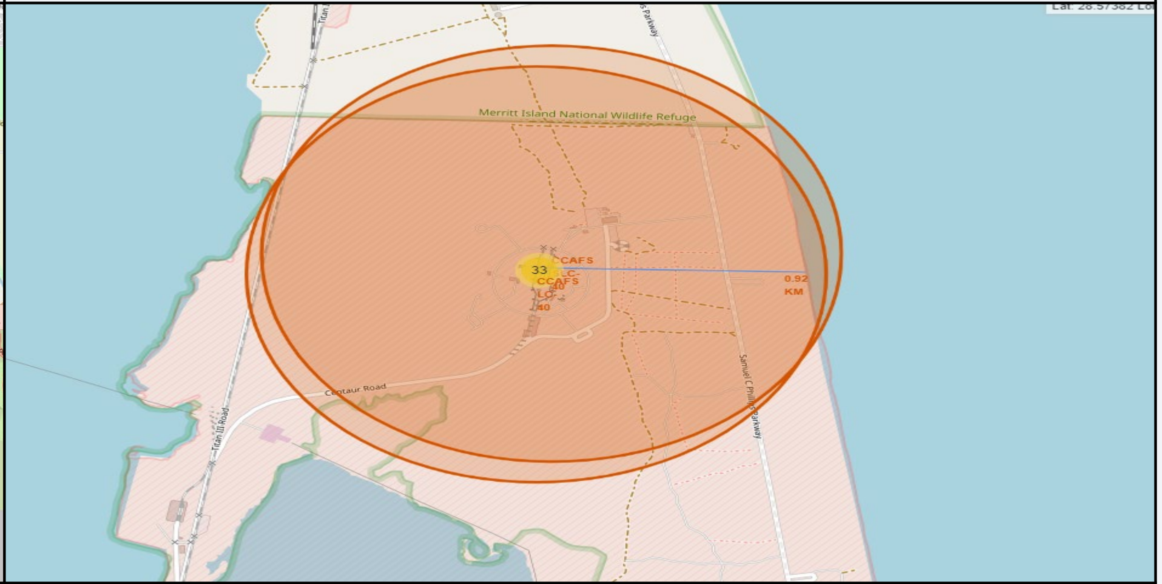
Global Launch Site Locations

Launch sites across the United States	Launch sites in California	Launch Sites in Florida
		
We can see that launch sites are concentrated on the southern east and west coastlines of the U.S.	Vandenberg Space Force Base hosts a single launch site in southern California.	Florida hosts three of the launch sites at Cape Canaveral, all in very close proximity to each other.

Launch Site Success/Failures

<table>
<tr><th>Success / Failed Launches across the United States</th><th>Success / Failed Launches in California</th><th>Success / Failed Launches in Florida</th></tr>
<tr><td>
</td>
<td>
</td>
<td>
</td></tr>
<tr><td>This map includes the total number of launches across the U.S. where we implement clustering to group launch sites together to summarize counts.</td><td>The number of both successful and unsuccessful landings is show by color in California (8 success & 12 failures)</td><td>Successful and unsuccessful landings in Florida are shown with KSC LC-39A expanded to show launch success by color.)</td></tr>
</table>

Launch Site Proximity to Landmarks

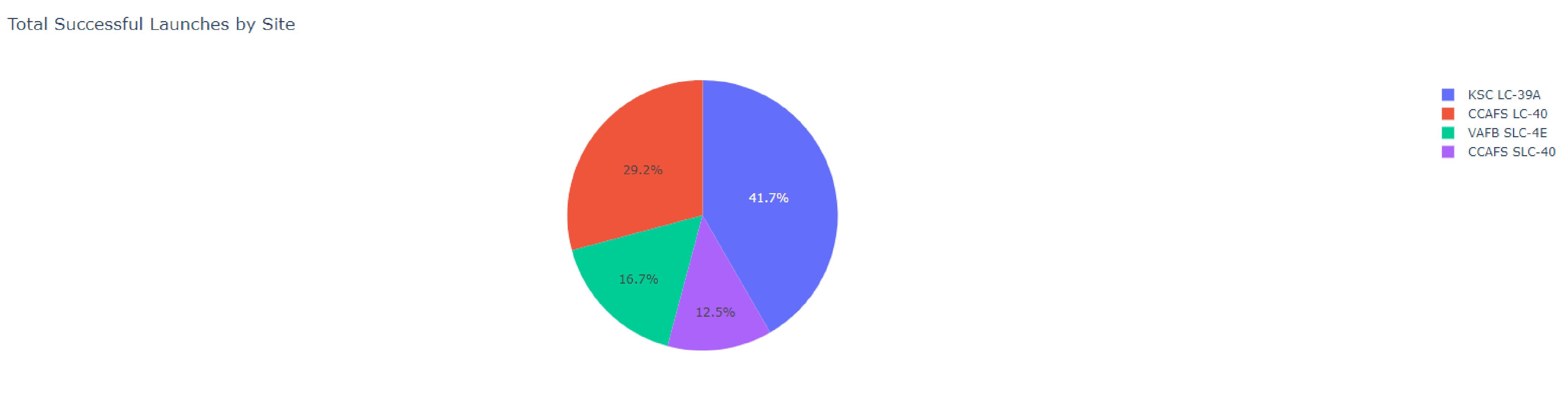
California Launch Site Landmark Proximity	Florida Launch Sites Proximity
	
We can see that Space Launch Complex 4 is approximately 1.28KM from the coast, and 14km from the city of Lompoc.	In this example we can see that both CCAFS SLC-40 and CCAFS LC-40 are less than 1km from a coastline.

In general we see that launch sites are generally within a few km of a coastline, and are all relatively close to railways, but are typically some distance (10km+) from any sort of major landmark, city, etc.

Section 4

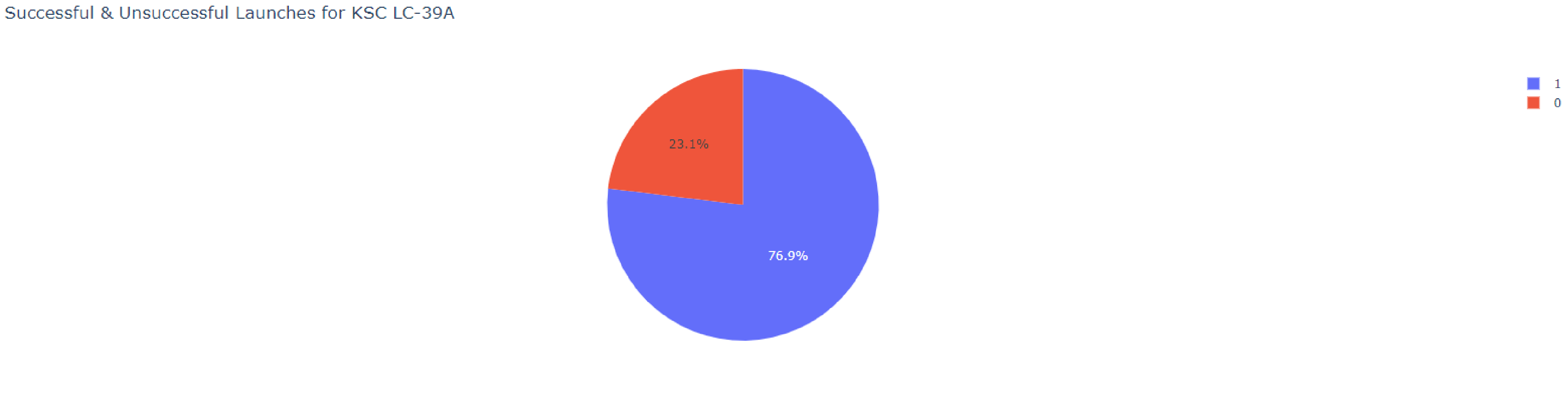
Build a Dashboard with Plotly Dash

SpaceX Launch Records Dashboard



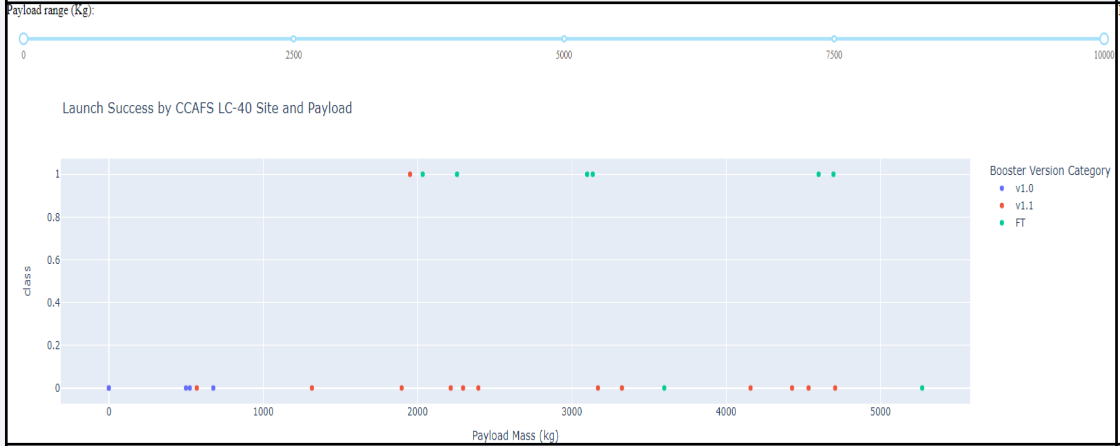
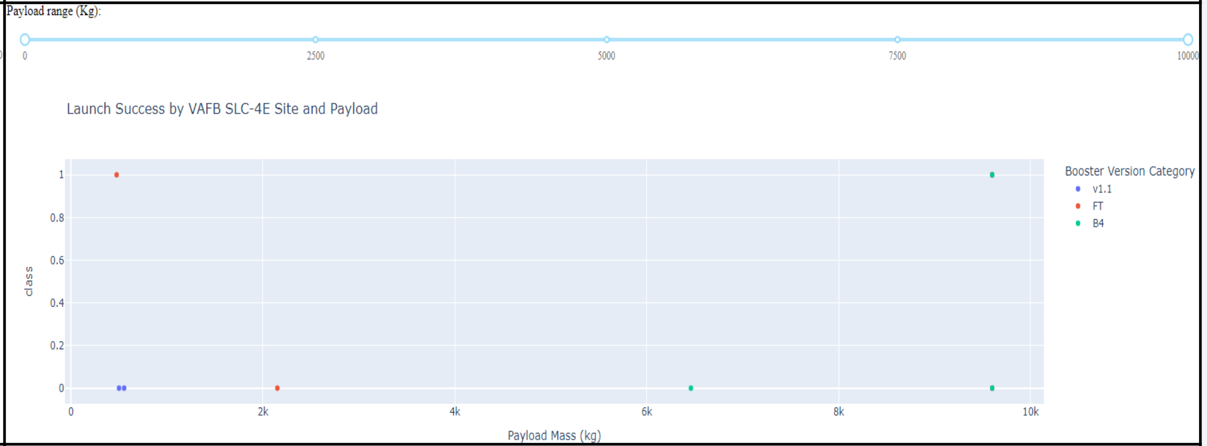
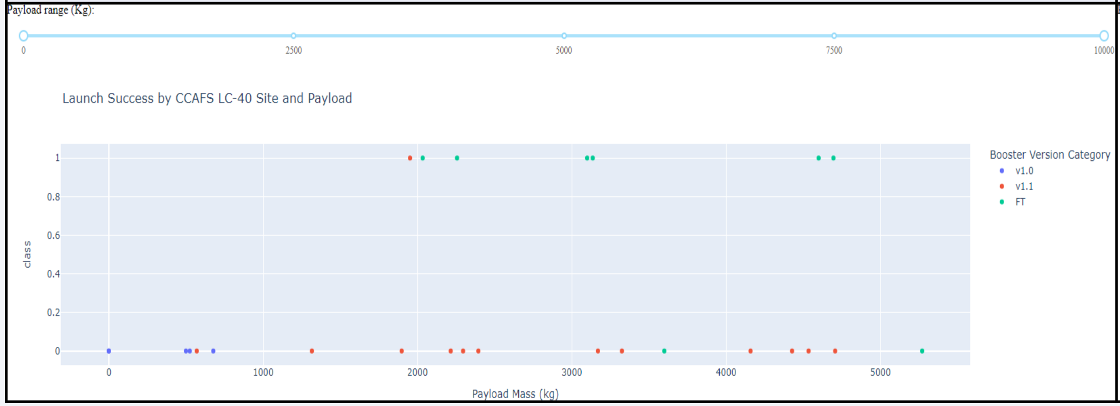
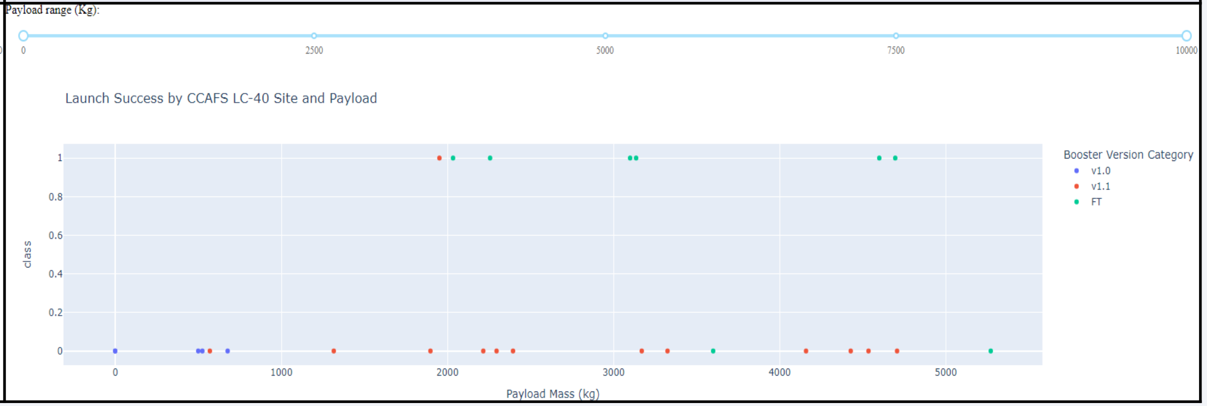
The pie chart shows the percentage of successful landing sites by site. As we can see KSC LC-39A has the highest successful rate at 41.7%, whereas CCAFS SLC-40 has the lowest at 12.5%. It is also important to also take into account the *number* of landings when evaluating the success rate.

Landing Site with Highest Success Rate



This pie chart shows that the landings site with the highest success rate is KSC LC-39A at 76.9%. It is important to note that the lower the number of landings, the more variability in the success rate, as the number of landings increases, the variability of the success rate diminishes. Basically, the more landings, the more predictable the success rate.

Launch Outcomes by Payload and Booster

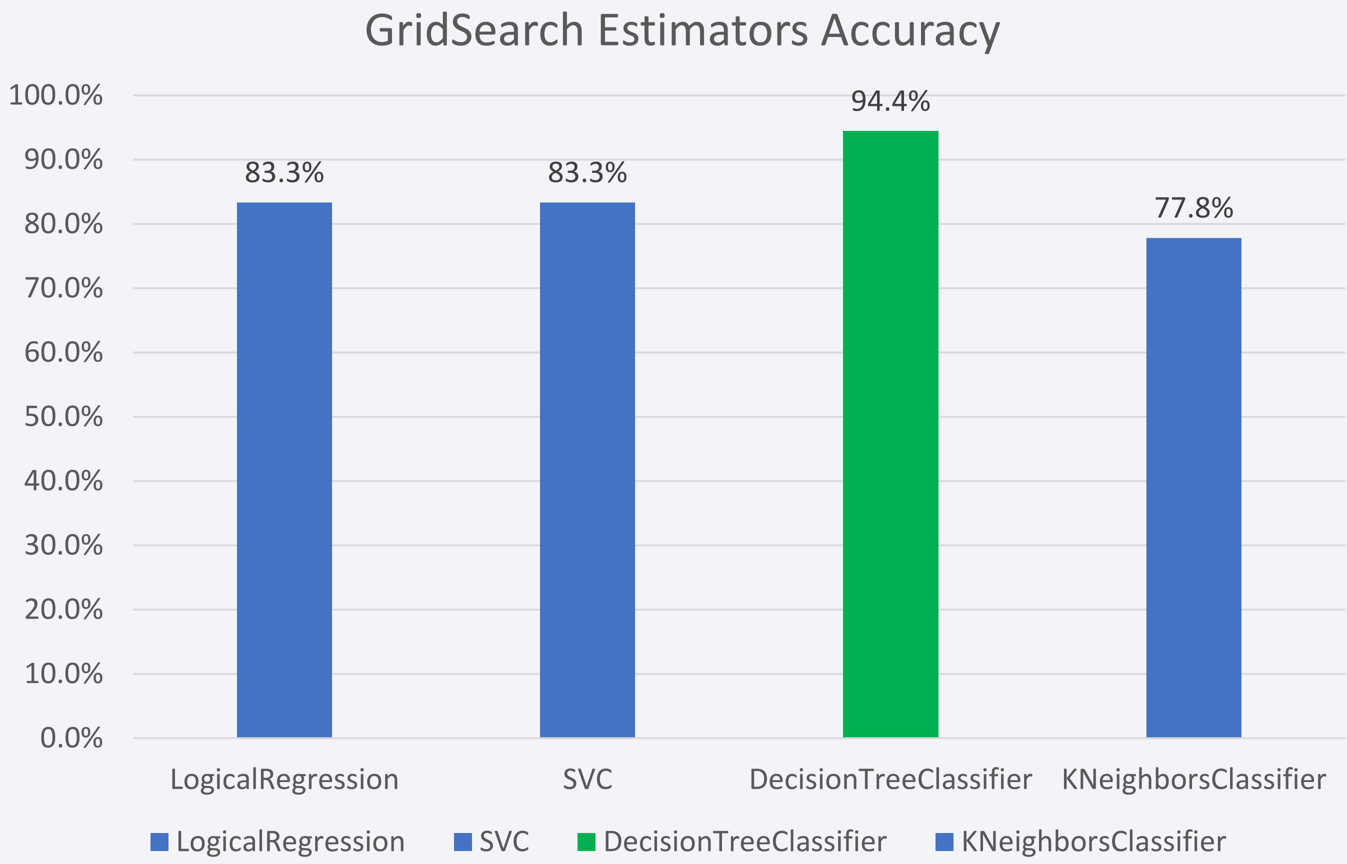
CCAFS LC-40 Site	VAFB SLC-4E Site
	
KSC LC-39A Site (Limited Payload < 5000kg)	**CCAFS SLC-40 Site (Payload > 5000 kg)**
	

The charts show the success rate by payload and booster category. The last two sites are filtered by payload for demonstration. A significant take away to note from the charts is the high success rate of the FT booster.

Section 5

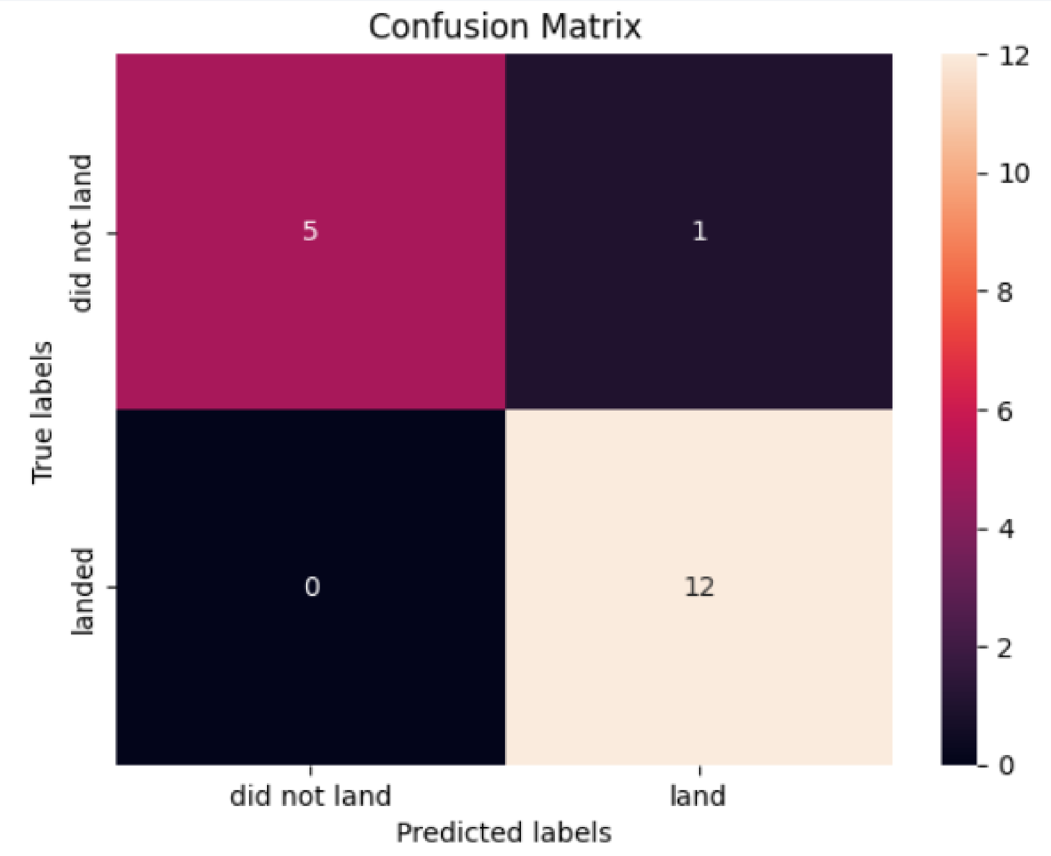
Predictive Analysis (Classification)

Model Classification Accuracy



The best model to use is the DecisionTreeClassifier which has an accuracy rate of 94.4%.

DecisionTreeClassifier Confusion Matrix



The best performing model is the DecisionTreeClassifier. In the Confusion Matrix shown we note the following:

- 12 landings were successfully predicted. (True Positive)
- 1 predicted landing did not land. (False Positive)
- 5 samples were correctly predicted not to land. (True Negative)
- There were no launches predicted not to land that actually landed. (False Negative)

Conclusions

- The more launches we perform, the higher we can expect our landing success rate, so we will need a certain minimum number of launches to gain proficiency.
- ES-L1, GEO, HEO, AND SSO orbits are preferred as they have a high landing success rates.
- We should use the DecisionTreeClassifier as our machine learning model as it has the highest predictive accuracy based on the SpaceX data.
- We should consider launching rockets with large payloads >9000kg.
- We should consider using the FT booster as the rocket type of choice.
- Our launch site(s) should be within 1-2km of an ocean and railway, and at least 10km from any significant landmark or city.

Thank you!