IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Cassandra Torenn
06/23/2023

Outline

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

Executive Summary

Summary of methodologies

Data was collected using SpaceX REST API and webscraping, where it was then formatted and cleaned. Exploratory data analysis using visualization and SQL queries reveled that payload mass, orbit type, launch location, and rocket booster version all play a role in a successful landing.

Summary of all results

A dashboard reveled that the Cape Canaveral Space Launch Center is the most likely to have successful landings and that lower payloads have a greater chance of success. Data visualization reveled that rocket booster version FT has the greatest chance of success over most payloads. Using the interactive Folium map, its clear that good launch locations are far from large cities, but close to coastlines, railways, and highways. Finally, the best machine learning models to predict successful landings are SVM, KNN, and logistic regression.

Introduction

- The world has always looked up at the wild skies with wonder. Now we're heading in the direction of lay people exiting Earth's atmosphere and touching the stars with the advent of commercial space travel.
- To make this dream a reality, the cost of such ventures needs to fall to more obtainable figures. SpaceX hopes to discover how to bring their costs down by reusing the first stage of a rocket launch, netting a savings of over $100 million dollars per launch.
- Many factors go into whether a first stage can be reused and this study aims to address each in turn.
  - Launch site location
  - Payload weight
  - Rocket booster version
  - Type of orbit the rocket shoots for
  - Times the first stage has been used

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using the SpaceX API and by web scraping launch records from Wikipedia
- Perform data wrangling
  - Data standardization and normalization
  - Replacing missing values concerning payload mass with the mean payload mass
  - filtering the data and converting some attributes like Outcome to classes
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- SpaceX launch data was collected using
  - [SpaceX REST API]for launch information
    - Rocket or booster version and outcomes
    - Size of payloads
    - Landing & launching specifications

Data requested from SpaceX REST API → Data parsed and returned in .JSON format → Data normalized, filtered and reformatted to .csv

  - SpaceX Wikipedia page
    - Falcon 9 launch records extracted from HTML tables

From website request Falcon 9 launch data → Parse data using BeautifulSoup to extract Falcon 9 launch records → Convert data into dataframe using Pandas

Data Collection – SpaceX API

Request SpaceX launch data	• Use GET request to pull data from API • Convert and normalize data into a dataframe • Filter data for key information on rockets, payloads, launch pads, and cores
Store data in lists and then convert to dataframe	• Create global variables to hold lists of key information • Use functions to fill data into lists • Create dataset by combining columns into dictionary and covert to dataframe
Filter dataframe for specific information on Falcon 9 Launches	• Remove all data not concerning Falcon 9 launches • Apply data wrangling to clean data

- GitHub URL of the completed SpaceX API calls notebook

Data Collection - Scraping

Requesting Falcon 9 Launch information from Wikipedia	• Request data using HTTP GET method • Create BeautifulSoup object from the returned HTML
Extract data from HTML tables	• Determine where the tables reside on the website • Iterate through to collect all column names
Create dataframe by parsing the returned tables	• Create dictionary where the keys are the extracted column names • Extract data and assign to each key in dictionary

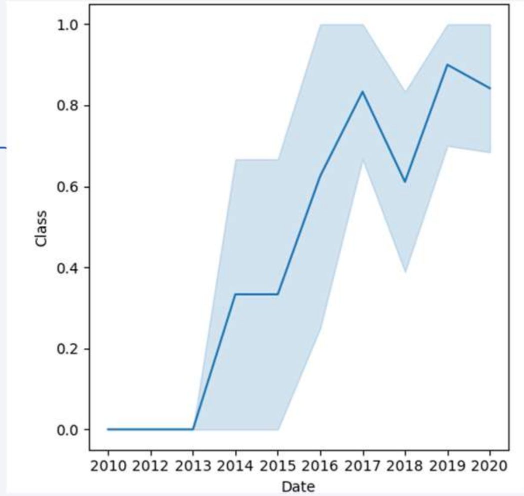
- GitHub URL of the completed web scraping notebook

Data Wrangling

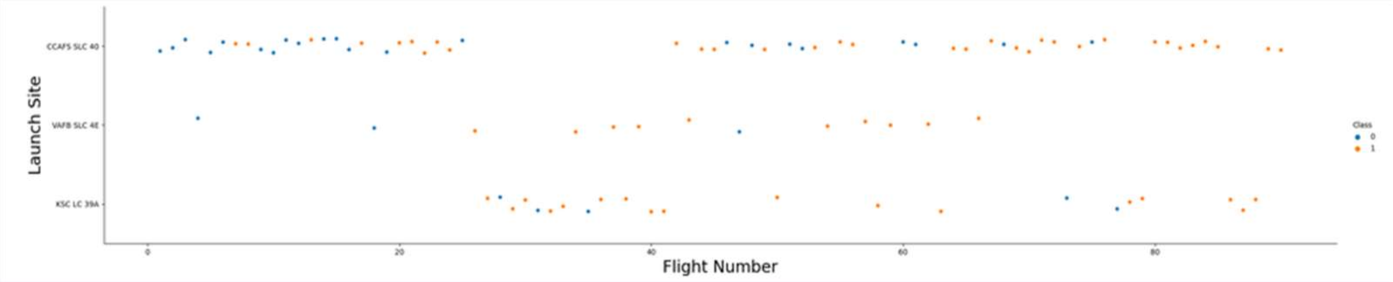
- Filter data
- Explore Data Types
- Determine missing values
  - Replacing missing values like payload mass with the mean of those values
- Calculate salient statistics
  - Launches per site
    - CCAFS SLC 40 =55
    - KSC LC 39A = 22
    - VAFB SLC 4E =13
  - Orbit types
    - LEO, VLEO, GTO, SSO, ES-L1, HEO, ISS, MEO, HEO, GEO, PO
  - Mission outcomes
    - Orbit type
    - Launch site
- Creating new variables
  - Determining which types of outcomes are successful
    - Store information as a class

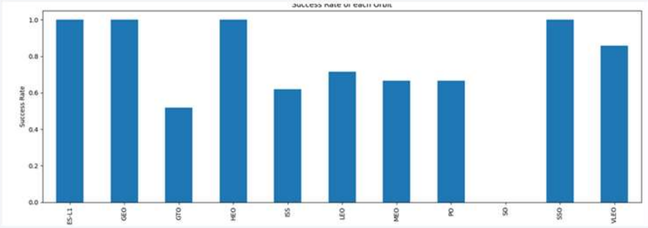
- GitHub URL of completed data wrangling

EDA with Data Visualization



- Bar charts – relationships between success rate and orbit type
- Line plots – yearly trend of launch successes
- Scatterplots – different attributes affect launch outcomes
  - flight number and payload mass
  - launch site and flight number
  - Launch site and payload mass
  - Flight number and orbit type
  - Payload and orbit type





GitHub URL of completed EDA with data visualization notebook

EDA with SQL

- Determined the unique launch sites and total number of successful mission outcomes
- Determined date of first successful landing on a ground pad
- Observed the payload mass carried by boosters that NASA launched
  - Total payload mass and which boosters have carried it
  - Average payload mass carried by booster version F9 v1.1
  - Which booster version has success in drone ship where the payload mass is between 4000-6000kg
- GitHub EDA
- GitHub EDA using SQL

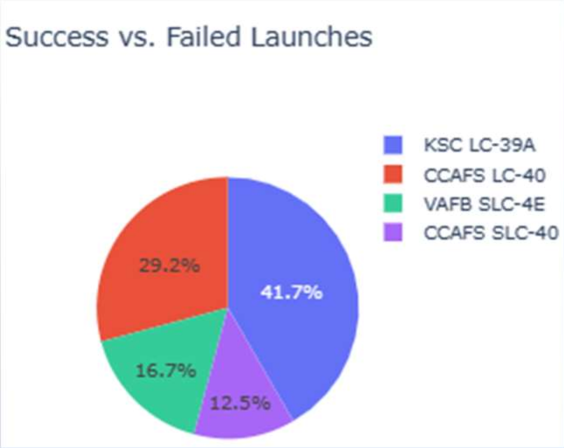
Build an Interactive Map with Folium

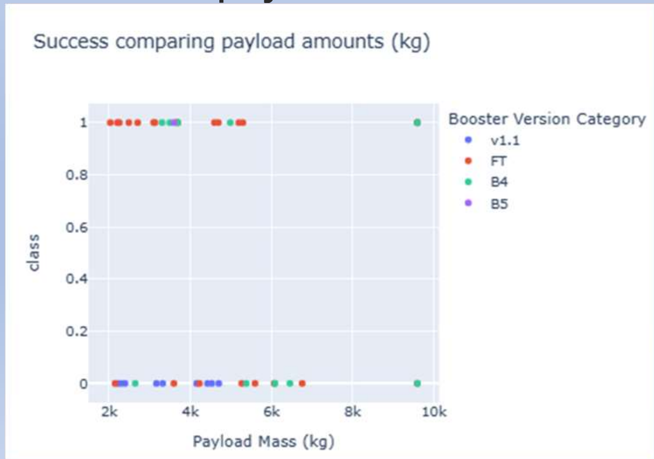
- Mark all launch site locations
  - Circle and marker at each site
- Mark the success or failed launches at each site
  - marker clusters with color coded markers indicate launch outcomes
- Mark distances between launch site and proximal objects such as highways, coastlines, railways, and major cities
  - Used mouse position to find coordinates of points of interest
  - Marked locations of POIs
  - Calculated distances between sites and POIs
    - Created marker labeling these distances for quick observation
  - Drew lines between using Polyline

GitHub URL of completed interactive map with Folium map

Build a Dashboard with Plotly Dash

- The goal of the dashboard is easy comparison of what factors influence a successful landing – specifically payload, launch site, and booster version.
  - First you'll find a dropdown menu listing the launch sites to quickly determine the successful and unsuccessful landings of the first stage in a pie chart.
  - Followed by a slider of payload masses and a scatter plot comparing the landing of the first stage with different payload masses, in terms of booster version.





[GitHub URL](#) of SpaceY Dashboard:

[GitHub URL](#) of completed predictive analysis

Predictive Analysis (Classification)

1. Load dataframe with data and name it X
2. Separate the outcome or 'Class' data from dataframe and output it as a series called Y
3. Standardize the original dataframe X using preprocessing.StandardScaler and fit transform the df
4. Split the data frames X and Y into testing and training data. 80% is training data and 20% is testing data
5. Create logistic regression object and use GridSearchCV to find the best parameters in the given dictionary
6. Train and fit the model, using the best parameters
7. Determine the accuracy of the model using data attribute best_score_ and visualize the accuracy using a confusion matrix
8. Repeat steps 5 to current with remaining classification models: KNN, SVM and Decision Tree
9. Compare the results of each confusion matrix and best_score_ method, choosing the model with the highest prediction success rate

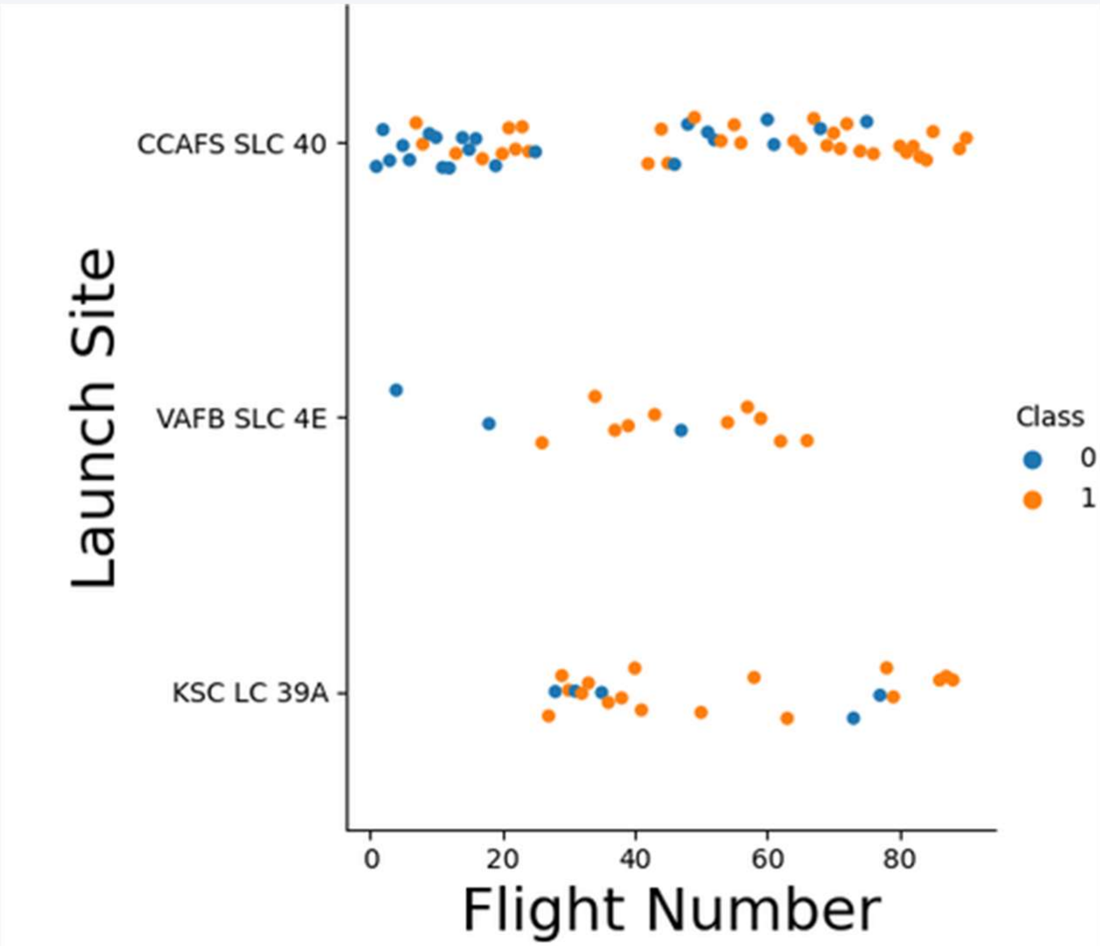
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

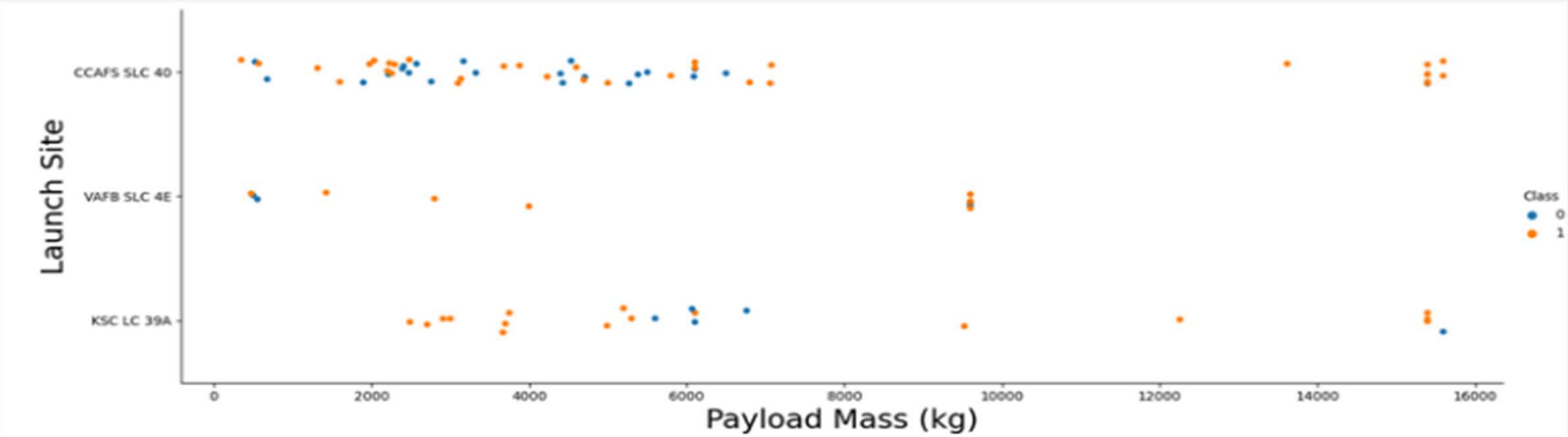
Insights drawn from EDA

Flight Number vs. Launch Site



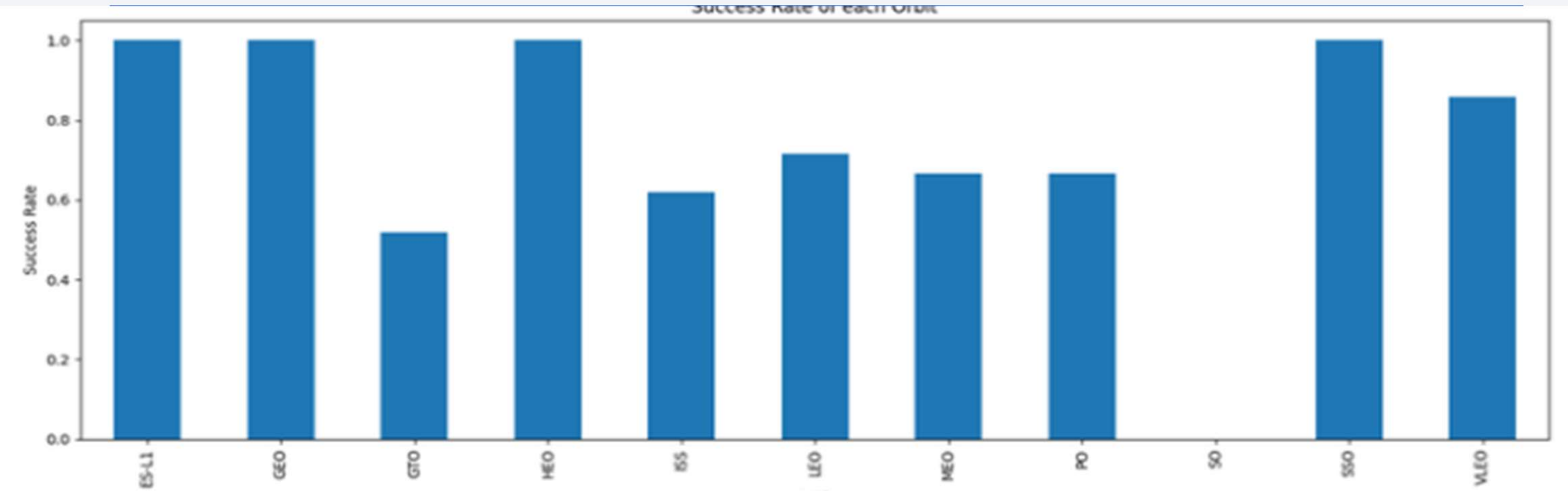
- The Class key represents if the landing was successful
  - blue is unsuccessful
  - orange is successful
- Observe that most flights were launched from Cape Canaveral AFS Space Launch, and the later the flight, the more likely the outcome was successful.
- The other two sites had fewer launches, but had more overall success.

Payload vs. Launch Site



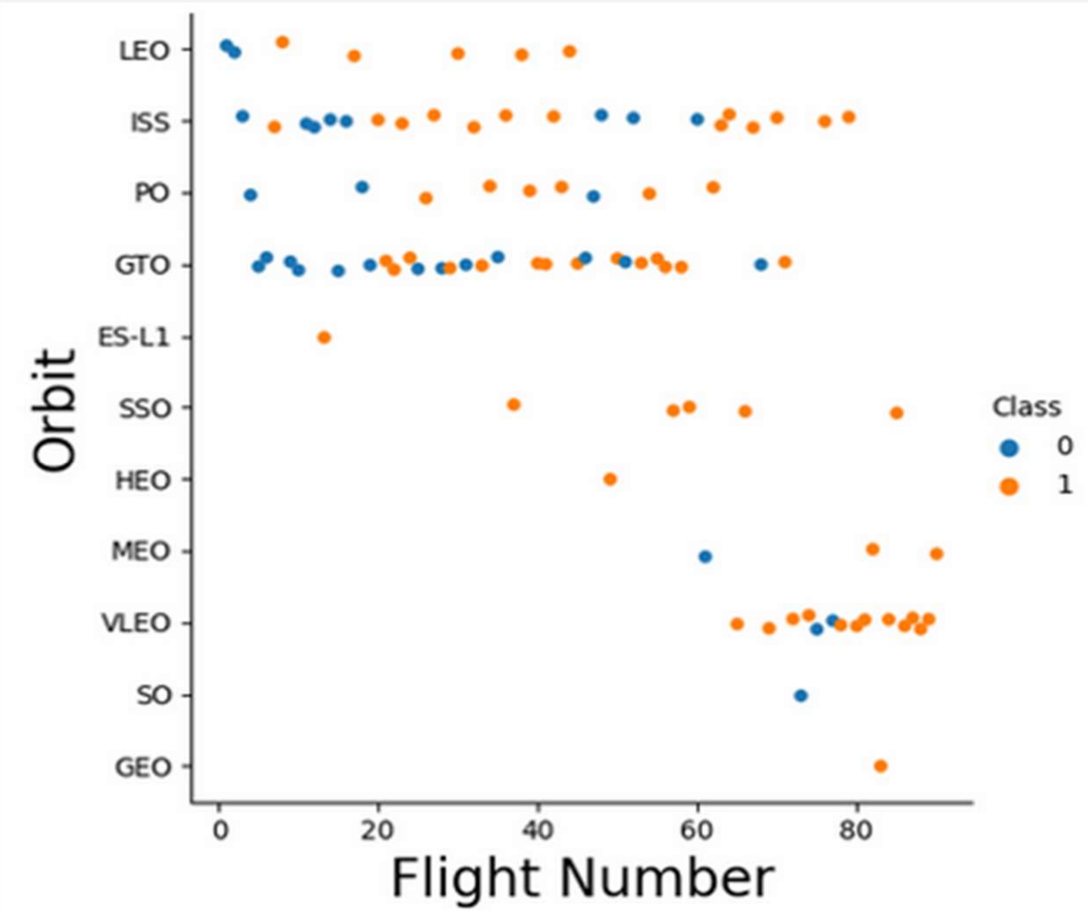
- CCAFS has launches across the greatest range of payload masses and is very successful with payloads larger than 6,000kg. However, no launches have a payload between 8,000 and 13,000kg at this site.
- VAFB has been overall successful from 0 to 10,000kg, but has not been used to launch any payloads that are heavier than 10,000kg.
- KSC does well with payloads between 2,000-16,000kg except between 5,000 to 7,000kg, where 4 of 5 launches were not successfully landed

Success Rate vs. Orbit Type



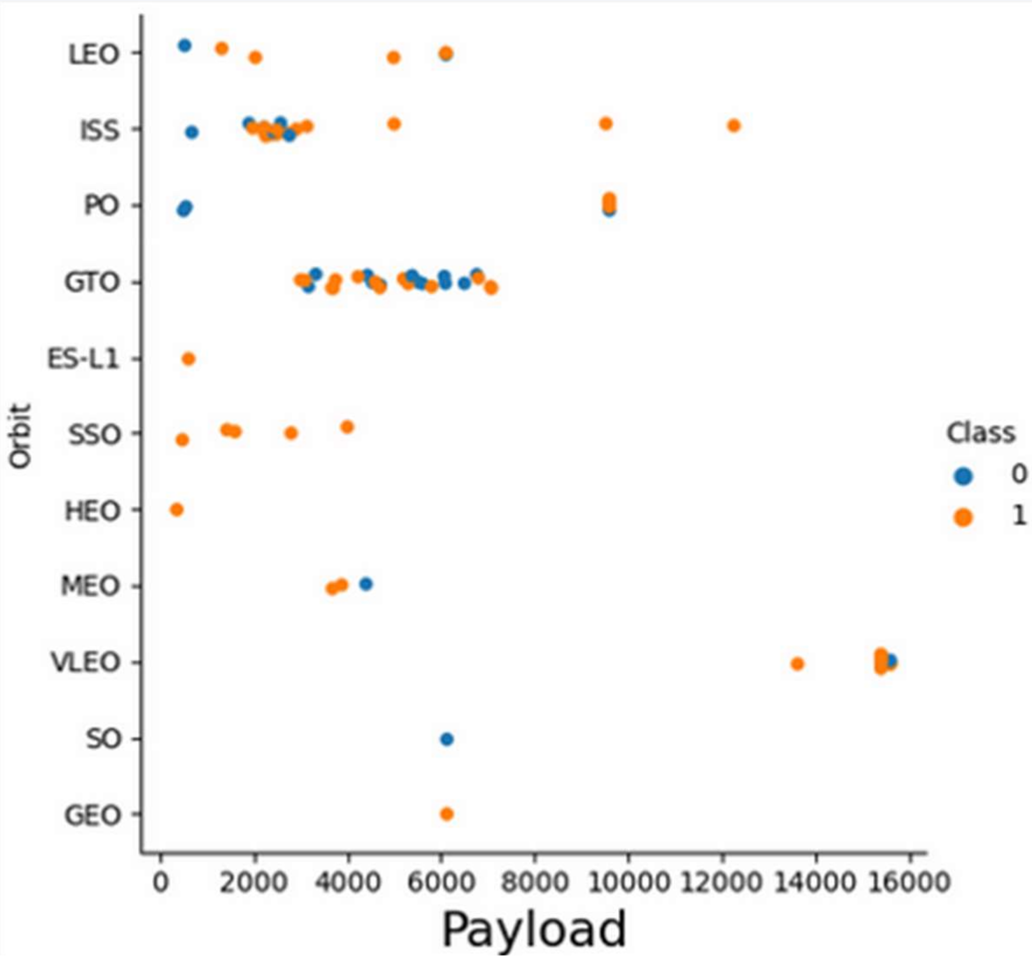
- For orbits ES-L1, GEO, HEO, and SSO the success rate was 100%, with MEO at over 80%.
- There have been no successful launches to the SO orbit type.
- All other have success rates between 50 and 70%.

Flight Number vs. Orbit Type



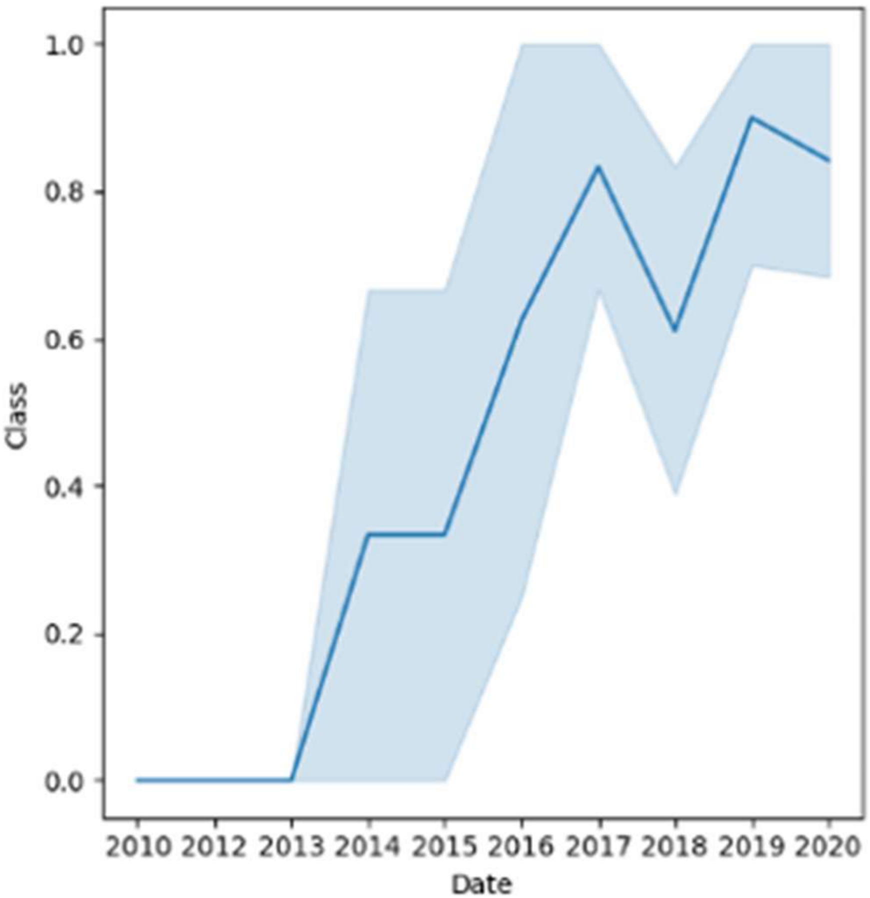
- Most of the beginning flights were to orbits of LEO, ISS, ES-L1, PO, and GTO, with moderate success. LEO orbits had the most success of these early flights.
- Other orbits were not attempted until the 40th flight and generally were quite successful. This is especially true for SSO, HEO, GEO, and VLEO

Payload vs. Orbit Type



- Launches to GTO shows no relationship between payload mass and success, although only loads from 3,000 to 8,000kg have been attempted.
- Launches to orbits ES-L1, SSO, HEO, and GEO have been successful at all payload masses, although all have been 6,000kg or less.
- Launches to LEO and ISS have more success with payload greater than 2,000kg and 4,000kg respectively.
- Launches to MEO have only carried around 4,000kg with 2 out of 3 successful landings, whereas VLEO has been mostly successful with masses greater than 13,000kg,

Launch Success Yearly Trend



- Over time, the successfulness of landings has improved, with no successes in the years 2010-2013.
- The success rate grew as time went on, improving to about 80% in 2017.
- A small dip occurred from 2017-2018, dropping the success rate to 60%, but it rebounded in 2019 and has stayed around 85% through 2020.

All Launch Site Names

```
%sql select distinct "Launch_site" from SPACEXTBL
```

The names of the unique launch sites were found using the above query. It uses sql magic to select a portion of the information stored in tables.

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

Launch Site Names Begin with 'CCA'

```
%%sql
select "Launch_site" from SPACEXTBL where "Launch_site" like '%CCA%' limit 5
```

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

There are 5 records where launch sites begin with `CCA` and they can be found using the query above.

Total Payload Mass

```
%%sql
SELECT SUM("PAYLOAD_MASS__KG_") as Total_Payload_Mass FROM SPACEXTBL
    WHERE "Customer" like '%NASA (CRS)%';
```

The above query calculates the total payload carried by boosters from NASA. It results in a total payload mass of 48213kg.

Total_Payload_Mass
48213.0

Average Payload Mass by F9 v1.1

```
%%sql
select AVG("PAYLOAD_MASS__KG_") from SPACEXTBL
    WHERE "Booster_Version" like '%F9 v1.1%'
```

Using sql magic, the average payload mass carried by booster version F9 v1.1 was calculated resulting in an average mass of 2534.667kg.

AVG("PAYLOAD_MASS__KG_")
2534.6666666666665

First Successful Ground Landing Date

```
%%sql
select Date, "Landing_Outcome" from SPACEXTBL where "Landing_Outcome" like '%Success (ground pad)%' limit 1
```

The date of the first successful landing outcome on a ground pad was found using the above query. It occurred on December 22nd, 2015.

Date	Landing_Outcome
22/12/2015	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
select "Booster_Version", "Landing_Outcome", "PAYLOAD_MASS__KG_" from SPACEXTBL
    WHERE "Landing_Outcome" like '%Success (drone%' AND
    "PAYLOAD_MASS__KG_" BETWEEN 4000 AND 6000
```

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are listed below.

Booster_Version	Landing_Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696.0
F9 FT B1026	Success (drone ship)	4600.0
F9 FT B1021.2	Success (drone ship)	5300.0
F9 FT B1031.2	Success (drone ship)	5200.0

Total Number of Successful and Failure Mission Outcomes

```
%%sql
SELECT "Mission_Outcome", COUNT("Mission_Outcome") FROM SPACEXTBL Group by "Mission_Outcome"
```

The total number of successful and failure mission outcomes was calculated to be 100 successful and 1 failure in flight.

Mission_Outcome	COUNT("Mission_Outcome")
None	0
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

```
%%sql
select "PAYLOAD_MASS__KG_", "Booster_Version" from SPACEXTBL where
    "PAYLOAD_MASS__KG_" = (select MAX("PAYLOAD_MASS__KG_") from SPACEXTBL)
```

PAYLOAD_MASS__KG_	Booster_Version
15600.0	F9 B5 B1048.4
15600.0	F9 B5 B1049.4
15600.0	F9 B5 B1051.3
15600.0	F9 B5 B1056.4
15600.0	F9 B5 B1048.5
15600.0	F9 B5 B1051.4
15600.0	F9 B5 B1049.5
15600.0	F9 B5 B1060.2
15600.0	F9 B5 B1058.3
15600.0	F9 B5 B1051.6
15600.0	F9 B5 B1060.3
15600.0	F9 B5 B1049.7

The names of the booster which have carried the maximum payload mass are listed to the right. This query needed a subquery to find the maximum payload mass as seen in the parentheses above.

2015 Launch Records

```
%%sql
select substr(Date, 4, 2) as Month, "Landing_Outcome", "Booster_Version", "Launch_Site"
    from SPACEXTBL WHERE substr(Date, 7, 4)='2015' and "Landing_Outcome" like '%Fail%'
```

The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015 are listed below.

Month	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

```
%%sql
select "Landing_Outcome", Count("Landing_Outcome") as Count from SPACEXTBL
    where Date between '04-06-2010' and '20-03-2017'
    group by "Landing_Outcome" order by count('Landing_outcome') desc
```

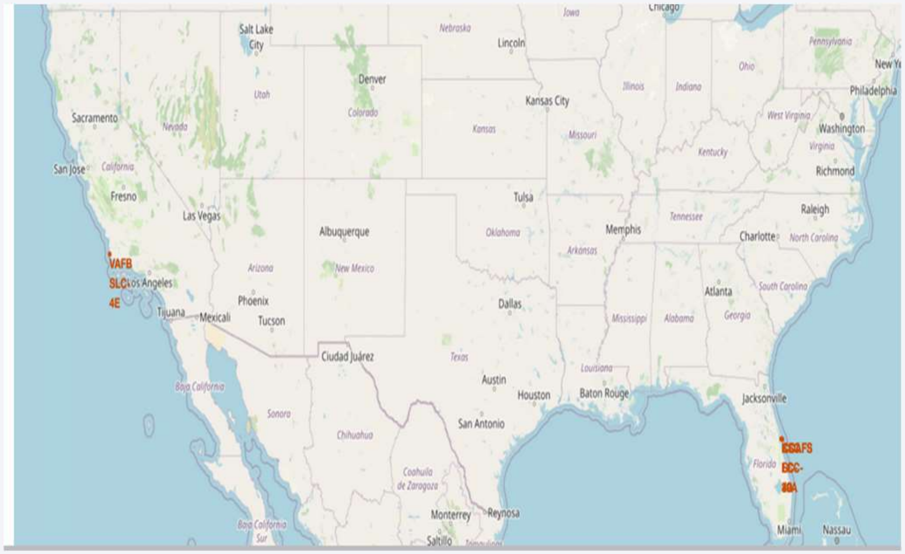
Landing_Outcome	Count
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

Ordered 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

Section 3

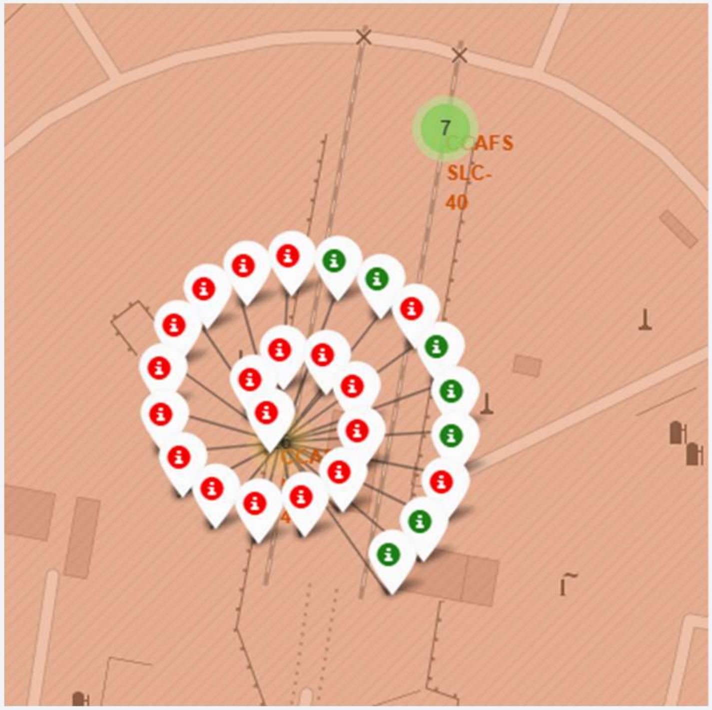
Launch Sites Proximities Analysis

Launch Site Locations in US



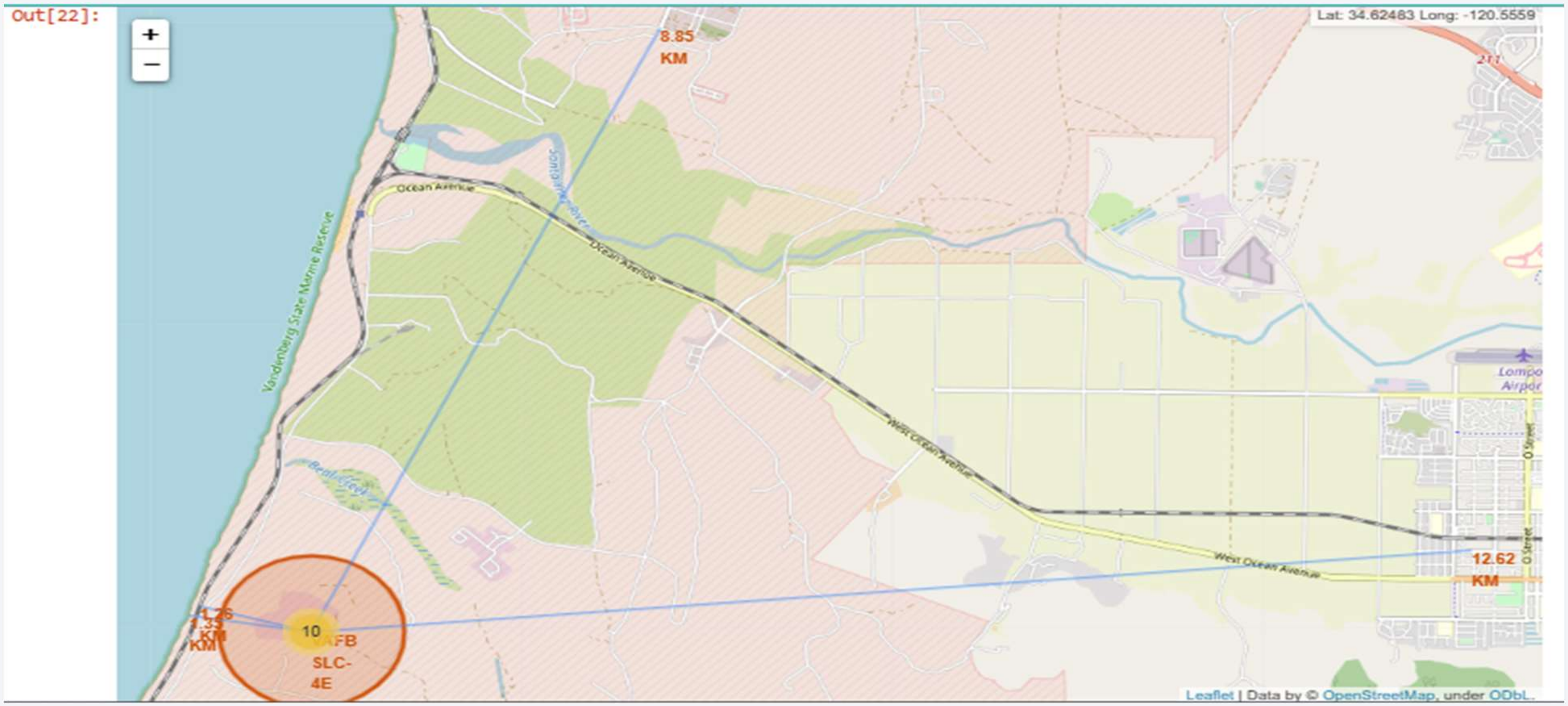
- The places for launches and landings are all very close to coastlines, far from major cities, but along highways. This is likely due to the possibilities of launch and landing failures that can result in large explosions, which the public needs to be protected from. Additionally, some landings are attempted in the ocean, so it makes sense to launch from very nearby. It is also ideal for highways and railroads to run close by since heavy equipment is needed to send a rocket to space.

Landing Outcomes for CCAFS LC-40

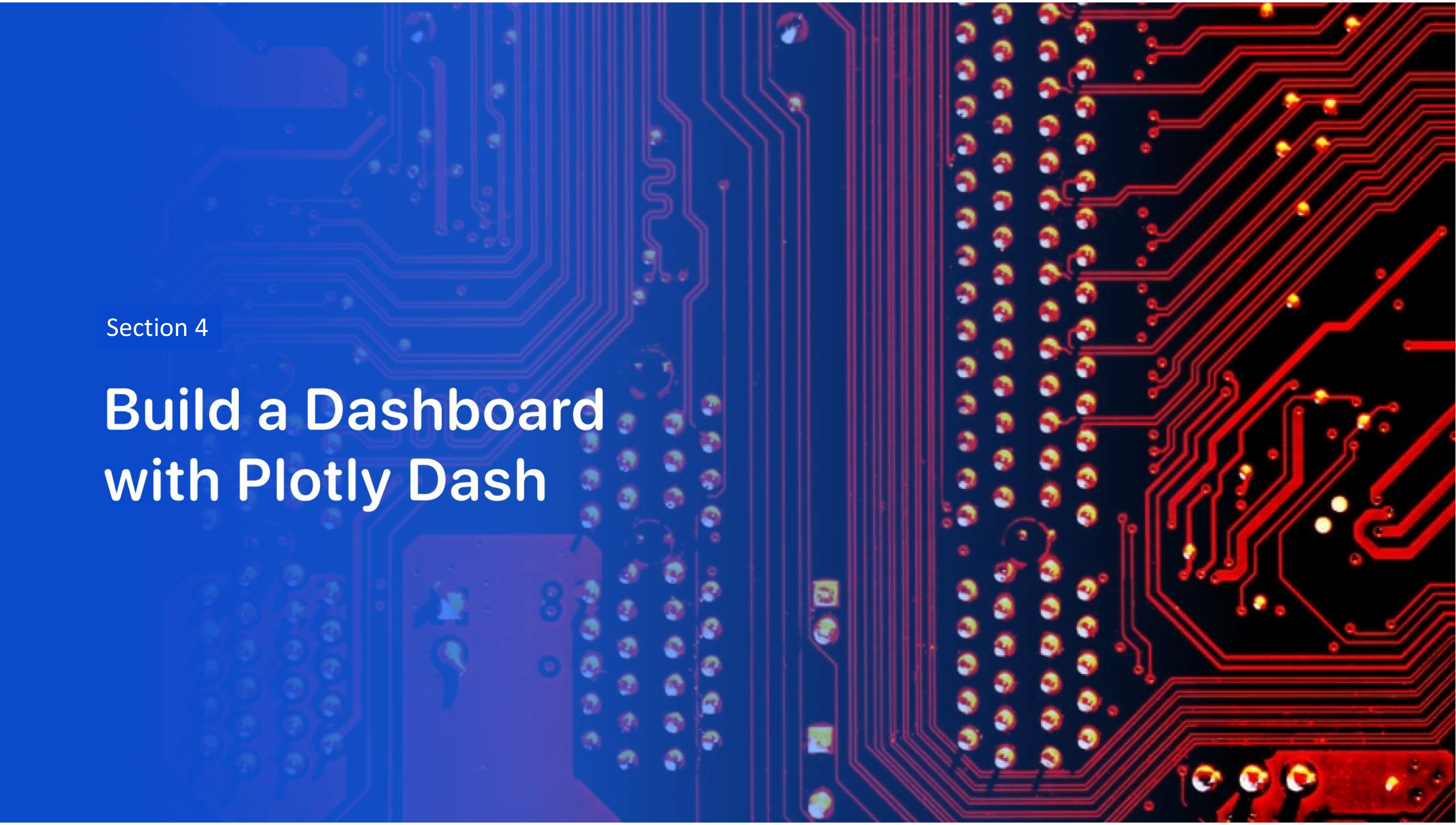


The green markers represent successful landing outcomes of the Stage 1 rocket from SpaceX launched at Cape Canaveral AFS Launch Center 40. The red markers indicate a failed landing.

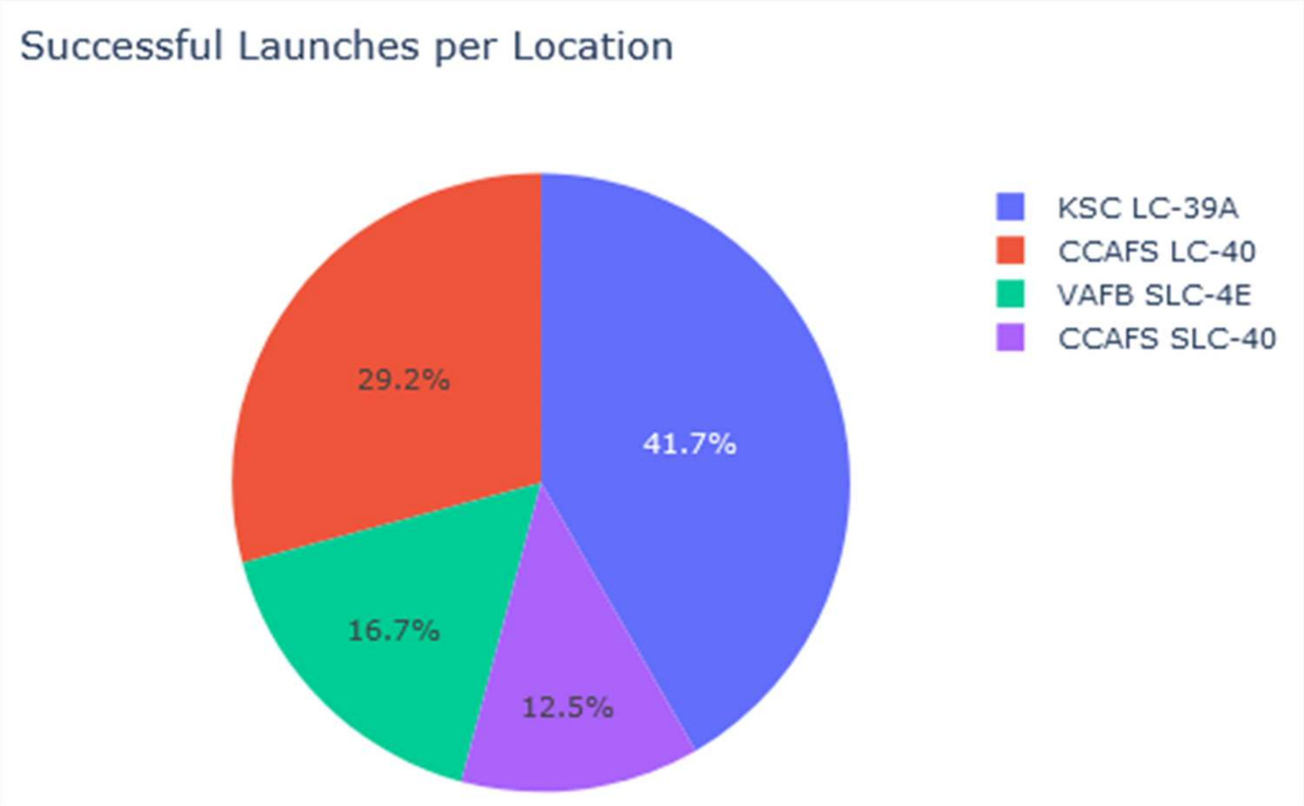
Proximities to Vandenberg Launch Site



Launch site is near to a coastline and far from large cities, likely to mitigate harm if a launch or landing goes awry. There is nearby access to rail lines and highways to facilitate moving heavy equipment in and out.

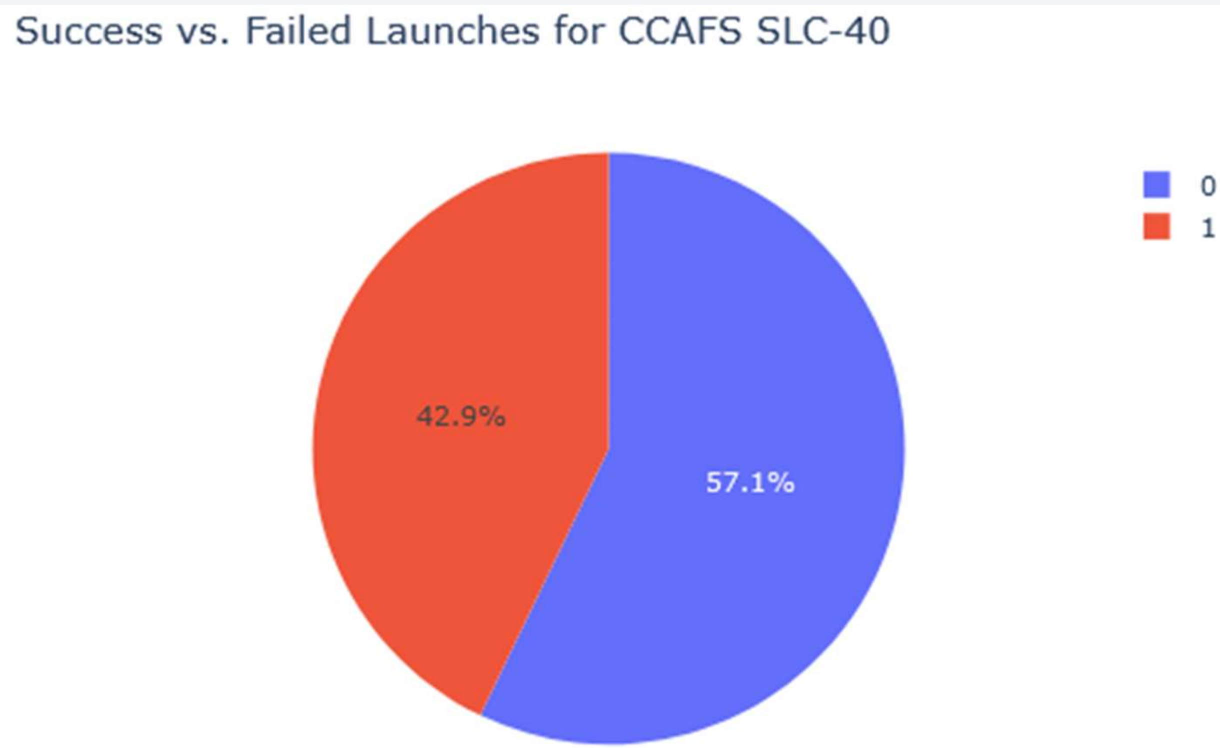

Section 4
Build a Dashboard
with Plotly Dash

Successful Landings per Location



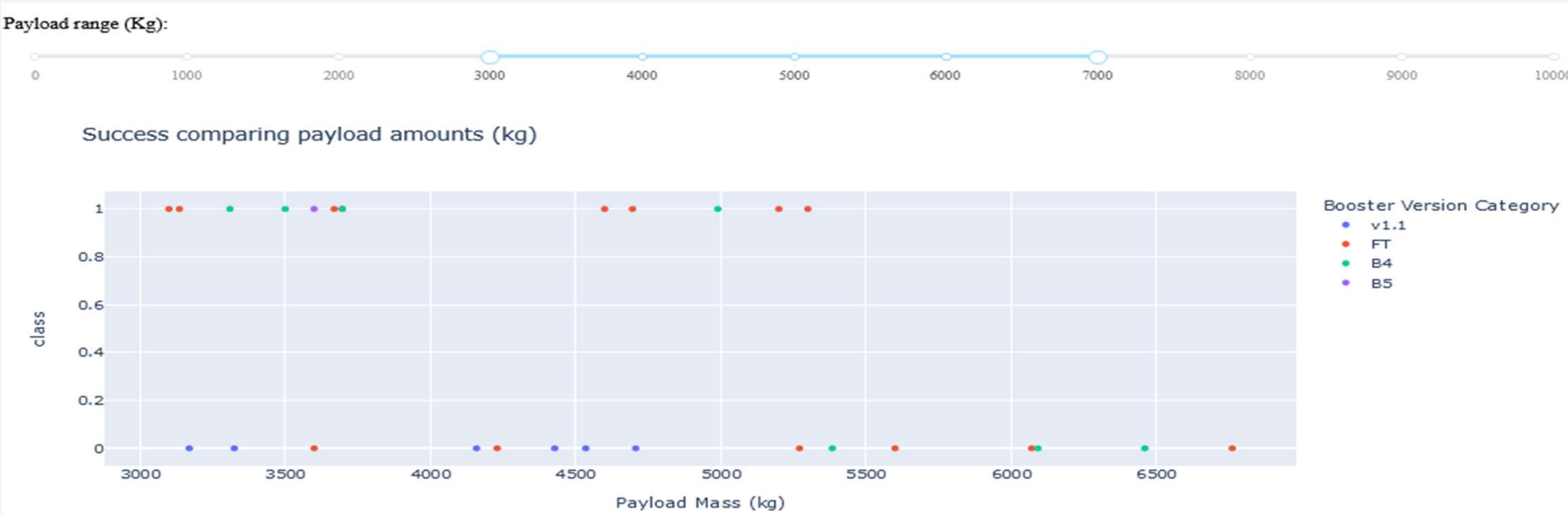
- Most of the successful landings occurred at the Kennedy Space Center (41.7%)
- Nearly 30% of successful landings happened at Cape Canaveral AFS Launch Center, but it's nearest neighbor CCAFS Space Launch Complex 40 is only responsible for 12.5%.
- Finally, VAFB SLC in California accounts for about 17% of successful landings.

Launch Site with the Highest Landing Success Ratio



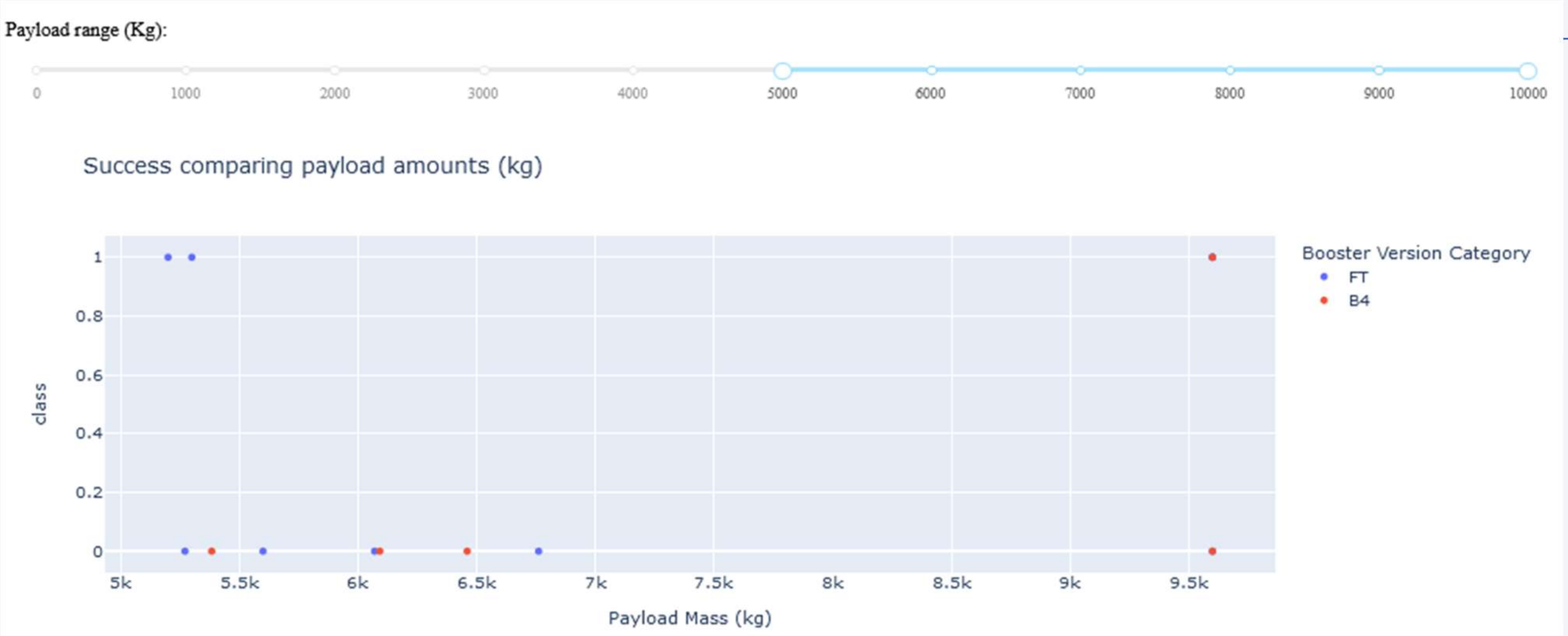
- Cape Canaveral AFS Space Launch Center has the highest success rate of any of the four sites observed. However, a 42.9% success rate at this site represents only 7 launches where 3 landings were successful

Comparison of Booster Versions, Landing Outcomes, and Payload Mass (3000-7000kg) at all sites



- All successful landings occurred using Booster Versions FT, B4, and B5. They all had payloads less than 5500kg.
- B4 was successful only for payloads less than 5000kg.
- There were no successful landings using V1.1, regardless of payload.
- The majority of successful landings occurred with payload masses (2000-3800kg) and (4500-5400)

Comparison of Booster Versions, Landing Outcomes, and Payload Mass (5000-10000kg) at all sites

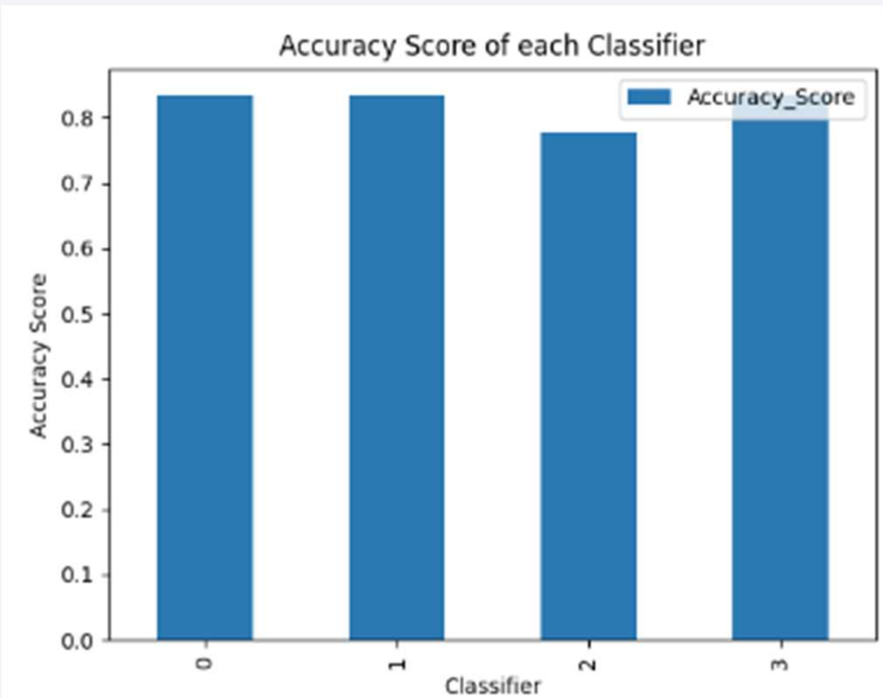


- Only Booster Versions FT and B4 have been employed when the payload mass is greater than 5000kg, and they are more likely to not land successfully at these masses.
- Only Booster Version B4 has been used with a payload of 9600kg, and so far has only been successful 50% of the time.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

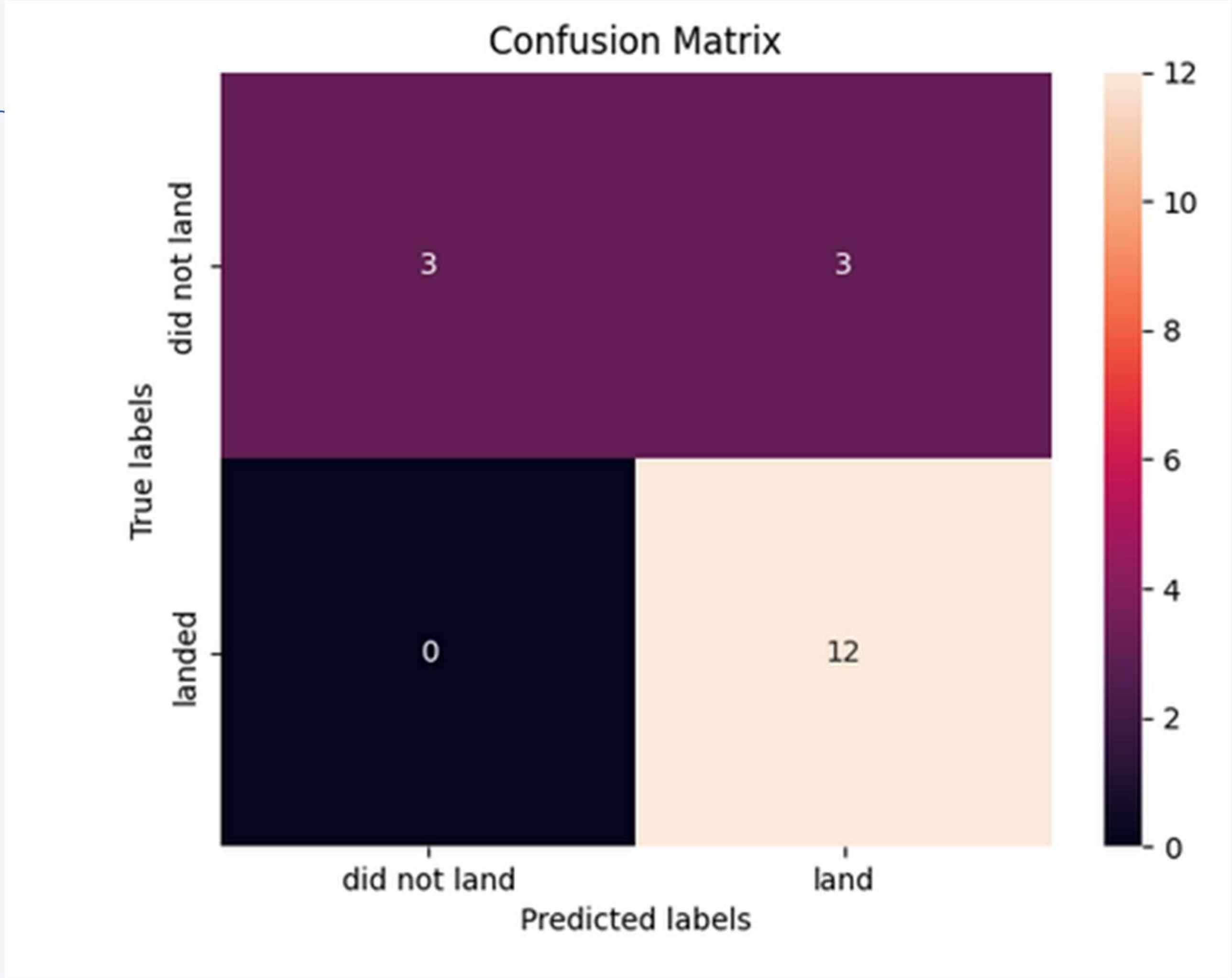


	Classifier_Type	Accuracy_Score
0	Logistic Regression	0.833333
1	SVM	0.833333
2	Decision Tree	0.777778
3	K Nearest Neighbor	0.833333

- Three models, Logistic Regression, Support Vector Machine, and K Nearest Neighbor, all share the same accuracy score of 83.333%. Only the Decision Tree model appear less accurate at 77.778%.

Confusion Matrix

- Three models, Logistic Regression, Support Vector Machine, and K Nearest Neighbor were the best performing models. It was able to predict correctly which landings would be successful. However, the models also has many false positives – in this case predicting that rocket will land when it actually does not. (Upper Right Quadrant)



Conclusions

- A stage 1 rocket's successful landing depends on a wider variety of factors including payload mass, booster version, launch site, and orbit type.
    - Most success happens with rocket booster versions FT, B4, and B5, with FT being the most successful along the spectrum of payload masses
    - Success is unlikely with payload masses between 5,500 to 9,500kg, and most likely under 5000kg.
    - Success is most likely at the Cape Canaveral Space Launch Complex
- The best models for predictions of landing success are KNN, SVM, and Logistic Regression
    - But beware of false positives

Appendix

Find all relevant code, SQL queries, charts, notebook outputs, and data sets in this [GitHub repository](#).

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude	Class
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.561857	0
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.561857	0
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.561857	0
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610829	34.632093	0
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28.561857	0

Thank you!