IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ross
17 September 2024

Outline

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

Executive Summary

Summary of methodologies

- Data Collection using the SpaceX API and cleaning the data

- Data Wrangling using Pandas to conduct Exploratory Data Analysis and determine the training/success variables.

- Explore the data further using Panadas and Matplolib to visualize multiple variable relationships and conduct some feature engineering

- Explore the data through SQL queries to gather a range of statistics regarding different variables

- Use Folium to gain geographic insight into the data and investigate the relationships between the launch sites to successes/failures.

- Build a Dash Application to visualize and compare data interactively to show insights and patterns.

- Build several Machine Learning Models and then test and compare the models to find the best one.

Summary of Results

- The more launches that have occurred, the greater the chance of success

- The Booster Version FT is likely to produce the greatest chance of success

- Launch Sites are near the coast, close to the Equator, infrastructure but away from cities

- Launch Site CCAFS SLC-40 has the highest chance of success

- Various classification models return similar results and have a problem returning False Positives.

- Orbits ES-L1, GEO, HEO, SSO have the best success rate.

URL: https://github.com/rsx8/IBM_DS_SpaceX_Capstone

Introduction

Project background:

SpaceX has managed to make Space Travel cheaper by selling it's Falcon 9 rocket launches for $62M, undercutting other providers that cost $165M. This is because the Falcon 9 rocket is reusable, provided the Falcon 9 can land safely.

We are a rival company that would like to bid against SpaceX for a rocket launch.

The Problem:

We need to be able to determine the cost of a launch for our rocket over the cost of the launch for a Falcon 9 rocket. To do this we need to determine how likely the Falcon 9 rocket is to succeed in landing.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
    - Data collecting through SpaceX API and Webscraping
- Perform data wrangling
    - Data loaded into Pandas Dataframe and convert data into Training Labels
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Build several models to predict successful landings, find the best parameters for each and compare their performance scores to determine the best model.

Data Collection – SpaceX API

Steps for SpaceX API Data Collection

1. Retrieve the data from the API using: `response = requests.get("https://api.spacexdata.com/v4/launches/past")`
2. Decode the content using `.json()` and turn into a DataFrame: `df = pd.json_normalize(json_raw)`
3. Filter out the unwanted columns from the DataFrame
4. Extract/Clean the data from the intial DataFrame into first a Dictionary and then into a new DataFrame.
5. Filter only for Falcon 9 Booster Versions
6. Replace missing values with the `.mean()` function over the rest of the column.
7. Export to `.to_csv` to keep a safe backup of the data.

URL: https://github.com/rsx8/IBM_DS_SpaceX_Capstone

Data Collection - Scraping

Steps for SpaceX Web Scapring

1. Retrieve the data from the API using: `response = requests.get("https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922")`
2. Decode the content using `BeautifulSoup: soup = BeautifulSoup(response.text, 'html.parser’)`
3. Extract the tables: `html_tables = soup.find_all('table')`
4. Extract the column/variable names from the desired table
5. Using the column names, create an empty dictionary and extract the values for the columns into it.
6. Convert the Dictionary into a Dataframe and proceed to Data Wrangling

URL: https://github.com/rsx8/IBM_DS_SpaceX_Capstone

Data Wrangling

Steps for Data Wrangling

1. Understand the data you have but looking for null values (`df.isnull().sum()`) and looking at which columns are numerical or categorical (`df.dtypes`)
2. Determine what are the Training Labels and transform/create them as needed. In the SpaceX case, we took 8 unique Categorical Landing Outcomes and transformed them into a Binary column.

URL: https://github.com/rsx8/IBM_DS_SpaceX_Capstone

EDA with Data Visualization

In order to identity patterns the following graphs were created:

1. Flight Number vs Launch Site
2. Payload Mass vs Launch Site
3. Success Rate vs Orbit Type
4. Flight Number vs Orbit Type
5. Payload Mass vs Orbit Type
6. Successes over the years

URL: https://github.com/rsx8/IBM_DS_SpaceX_Capstone

EDA with SQL

SQL Queries performed:

1. Unique Launch Sites
2. Search records where launch site begins with the string ‘CCA’
3. Total amount of payload launched by NASA
4. The average payload carried by Booster Version F9 V1.1
5. The data of the first successful ground landing
6. Boosters with success on a Drone Ship with a payload 4k-6K
7. The counts of success/failure regarding mission outcome
8. All the Booster version name
9. Counts of the landing outcomes

URL: https://github.com/rsx8/IBM_DS_SpaceX_Capstone

Build an Interactive Map with Folium

Folium Site Map Includes:

1. Location or all Launch Sites
2. Marker clusters of each Launch, colored green/red to easily show success/failure
3. Distance to infrastructure and points of interest such as urban centres and open areas (the coast).
4. The Equator relative to the Launch Site locations.

URL: https://github.com/rsx8/IBM_DS_SpaceX_Capstone

Build a Dashboard with Plotly Dash

Plotly Dash:

1. An interactive pie chart to show the success rate at each Launch Site or which Launch Site had the most successes across all Sites. This was to see if there was a pattern between Launch Site and Success.
2. A scatter plot that showed payload against success noting different booster versions. This was to see if there was a relationship between Success, Payload Mass and/or Booster Version:

URL: https://github.com/rsx8/IBM_DS_SpaceX_Capstone

Predictive Analysis (Classification)

Steps for predictive Analysis:

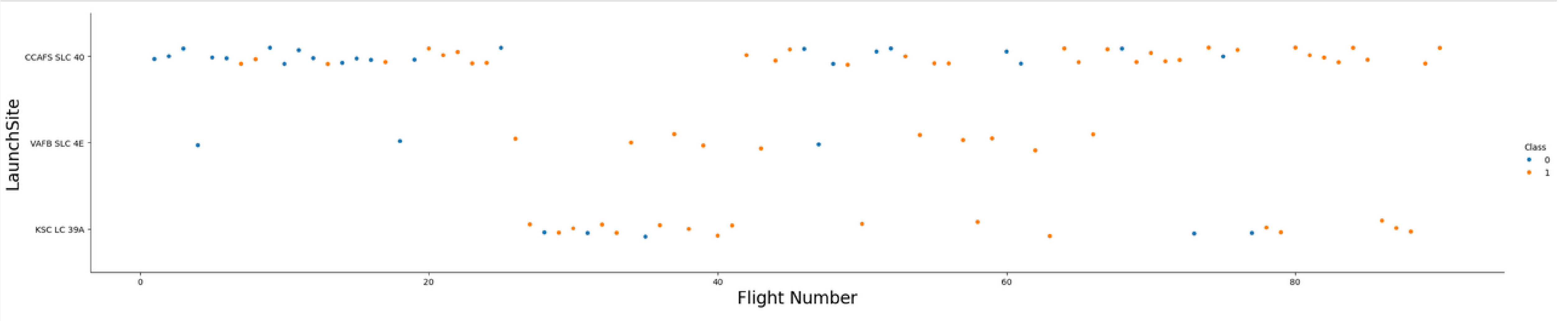
1. Transform the data so that it removes any biases from where that data where values are large/small.
2. Split the data into Training/Test sets, usually 80%/20% split.
3. Select the desired algorithm (Logistic Regression, Support Vector Machine, Decision Tree, K Nearest Neighbour) and create a dictionary of different hyperparameters.
4. Use GridSearchCV to find the best hyperparameters and use the best hyperparameters to get an accuracy score.
5. Compare across multiple different algorithms (repeating steps 3 and 4) until you've found the model with greatest accuracy.
6. Create and view the confusion matrix of the chosen model to quickly see how the model performs against, False Positives and False Negatives.

URL: https://github.com/rsx8/IBM_DS_SpaceX_Capstone

Section 2

Insights drawn from EDA

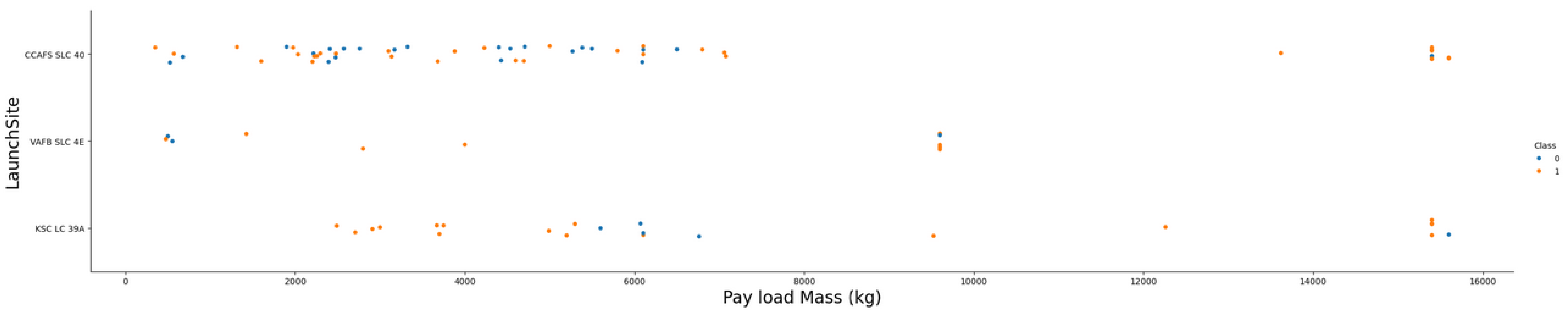
Flight Number vs. Launch Site



Key points:

1. Success rate (Class = 1/Orange Points) improves as Flight Number increases.
2. High Chance of landing failure before Flight Number reaches 20
3. CCAFS SLC 40 is the most common Launch Site

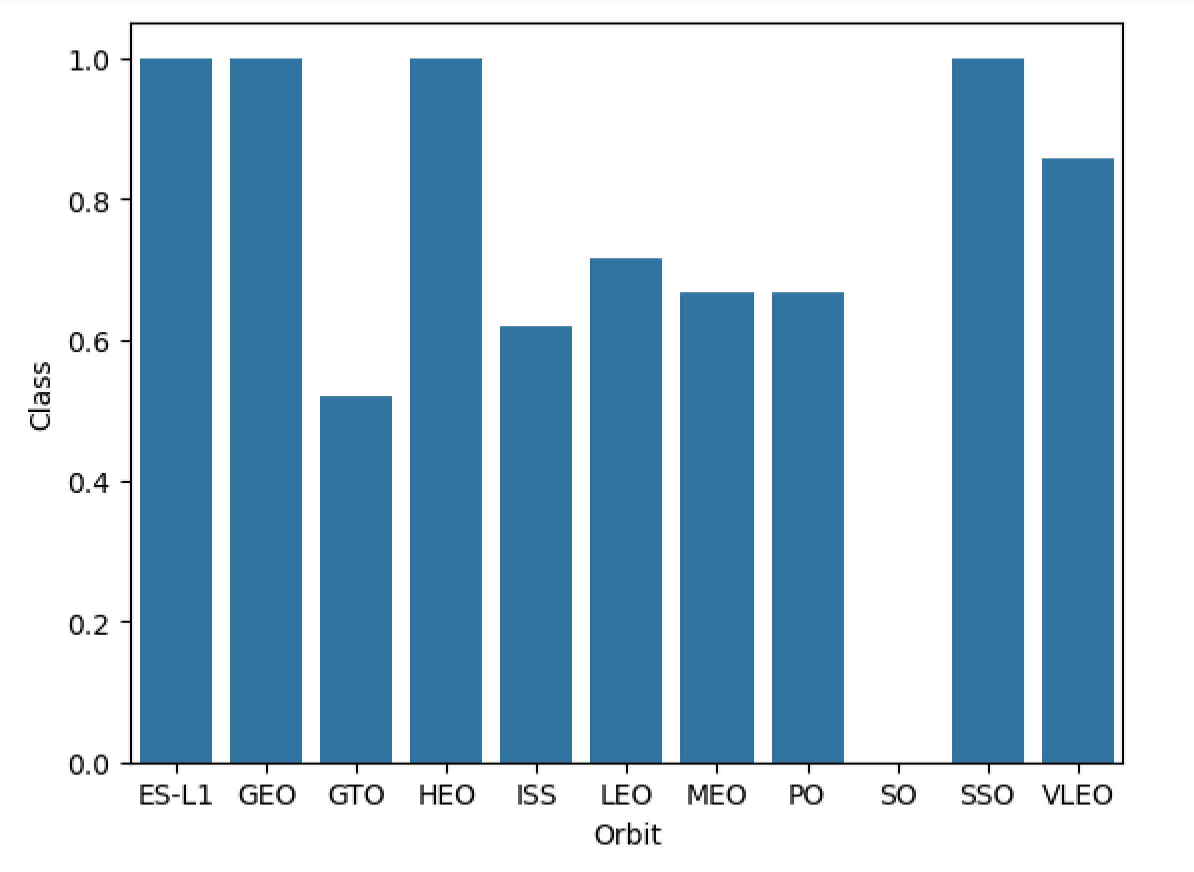
Payload vs. Launch Site



Key points:

1. VAFB SLC 4E Launch Site does not launch with a mass higher than 10,000 kg
2. Higher payloads greater than 8,000 kg tend to be more successful
3. KSC LC 39A has better success with < 5,000 kg and > 9,000 kg Payload Mass
4. Most launches occurred from site CCAFS SLC 40

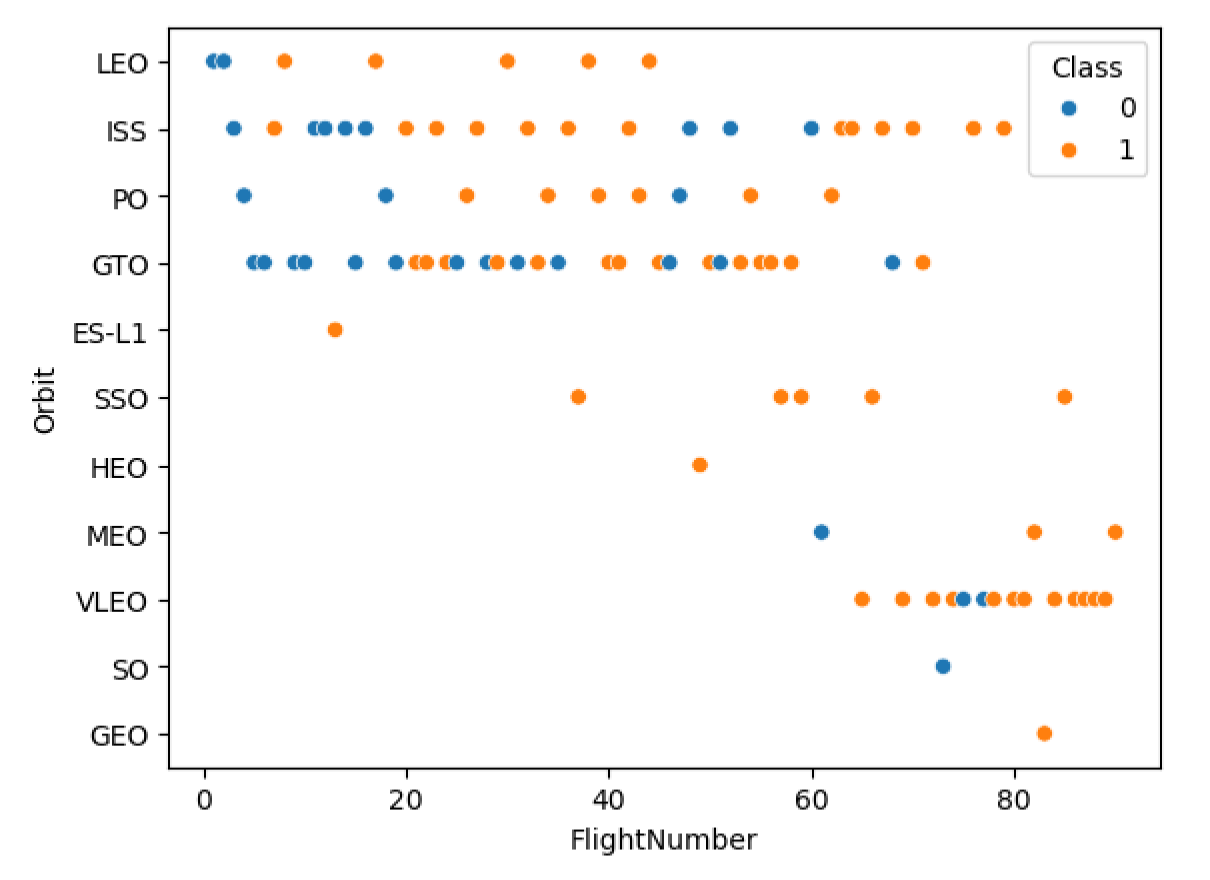
Success Rate vs. Orbit Type



Key points:

1. Success rate 100% for Orbits: ES-L1, GEO, HEO, SSO
2. Success rate 50%-85% for Orbits: GTO, ISS, LEO, MEO, PO and VLEO
3. Success rate 0% for SO Orbits

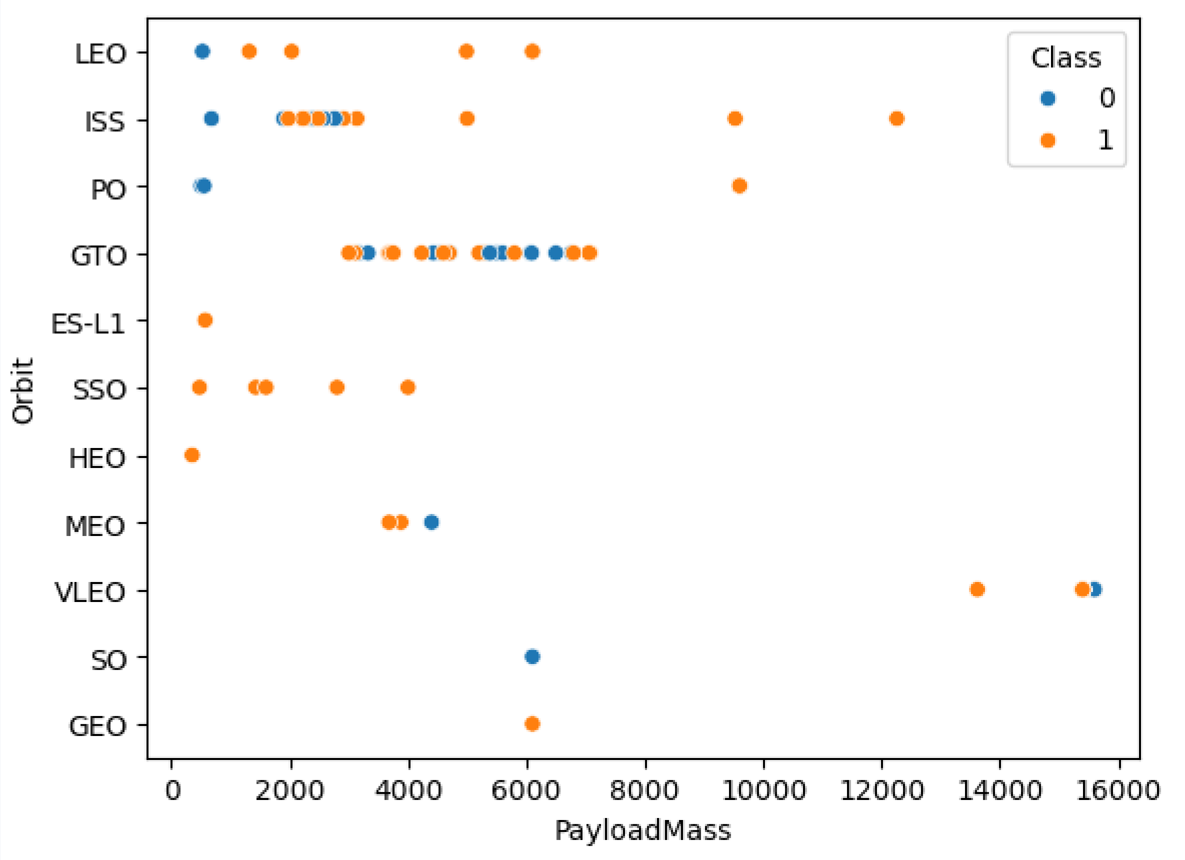
Flight Number vs. Orbit Type



Key points:

1. As Flight number increases success rate increases
2. GTO and ISS have a mixed success rate.
3. ISS Orbit remains a consistent target for a launch
4. LEO Launches decline as the flight number increases and VLEO increases when LEO stops
5. VLEO has the greatest frequency of recent launches
6. Most launches where either targeting GTO, ISS or VLEO orbits

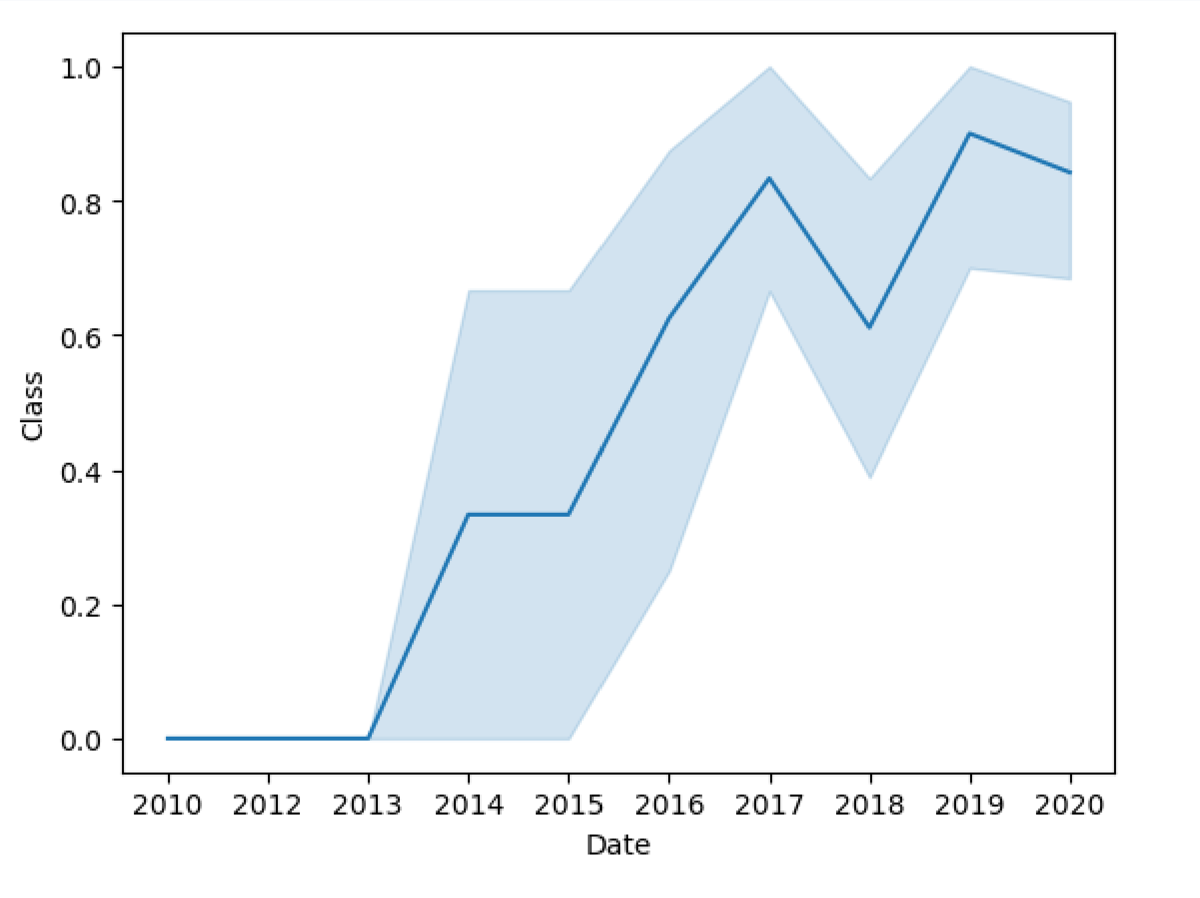
Payload vs. Orbit Type



Key points:

1. VELO has the heaviest payloads at > 13,000 kg
2. The ISS Orbits have some outlier payloads but otherwise are consistent at 2,000 - 4,000 kg
3. GTO has a spread from 3,000 - 8000 kg but with inconsistent success rate.
4. SSO, HEO and ES-L1 where all successful with Payloads under or equal to ~4,000 kg
5. As Payload increases, generally success rate increases, except for GTO.

Launch Success Yearly Trend



Key points:

1. As the years increase, generally the success rate has increased.
2. This peaked in 2019 at about 90%
3. Before 2013 the chance of success was 0%

All Launch Site Names

```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE
```

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

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

Key points:

1. There are 4 launch Sites listed in the SPACEXTBALE

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEXTABLE 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	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	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	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Key points:

1. 5 records where the Launch_Site begins with ‘CCA’
2. It Shows a range of Booster Versions
3. Mission Outcome is not directly related to Landing Outcome – they both don’t need to be successful.

Total Payload Mass

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

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Payload like '%CRS%'
```

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

SUM(PAYLOAD_MASS__KG_)
111268

Key points:

1. The total sum of the payloads that have been launched by Space X for NASA (CRS) is **111,269 kg**

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 SPACEXTABLE WHERE Booster_Version like 'F9 v1.1%'
```

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

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

Key points:

1. The average payload mass carried by booster version F9 v1.1 is: **2,534 kg**

First Successful Ground Landing Date

```
%sql SELECT * FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)' ORDER BY DATE ASC LIMIT 3
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2015-12-22	1:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)
2016-07-18	4:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)

Key points:

1. The first successful Ground Landing was on the **22nd of December 2015**
2. The second successful Ground Landing was in July of 2016.
3. The third successful Ground Landing was in February of 2017.

Successful Drone Ship Landing with Payload between 4000 and 6000

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 * FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ >= 4000 AND PAYLOAD_MASS__KG_ <= 6000
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2016-05-06	5:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-08-14	5:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-10-11	22:53:00	F9 FT B1031.2	KSC LC-39A	SES-11 / EchoStar 105	5200	GTO	SES EchoStar	Success	Success (drone ship)

Key points:

1. Only 4 launches (with payloads 4000-6000kg) have had a successful landing on a Drone Ship.

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
%sql SELECT Mission_Outcome, COUNT(*) AS Cnt FROM SPACEXTABLE GROUP BY Mission_Outcome
```

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

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

Key points:

1. Nearly all Mission Outcomes have been a success.
2. This is in contrast the landing outcomes (seen in earlier slides) that have not always been successful.

Boosters Carried Maximum Payload

```
%sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

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

Booster_Version
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

Key points:

1. Many F9 B5 boosters have carried the maximum payload from the SPACEXTABLE.

```
%sql SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE
```

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

MAX(PAYLOAD_MASS__KG_)
15600

The Maximum Payload so far has been 15,600 kg

2015 Launch Records

```
%sql SELECT substr(Date, 6,2) AS 'Month', Landing_Outcome, Booster_Version, Date FROM SPACEXTABLE WHERE Landing_Outcome ='Failure (drone ship)' and substr(Date,0,5)='2015'
```

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

Month	Landing_Outcome	Booster_Version	Date
01	Failure (drone ship)	F9 v1.1 B1012	2015-01-10
04	Failure (drone ship)	F9 v1.1 B1015	2015-04-14

Key points:

1. There have been 2 failed outcomes when landing on Drone Ship in 2015
2. They both used the same Booster Version

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

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

```
%sql SELECT Landing_Outcome, COUNT(*) AS CNT FROM SPACEXTABLE WHERE DATE > '2010-06-04' AND DATE < '2017-03-20' GROUP BY Landing_Outcome ORDER BY CNT DESC
```

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

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

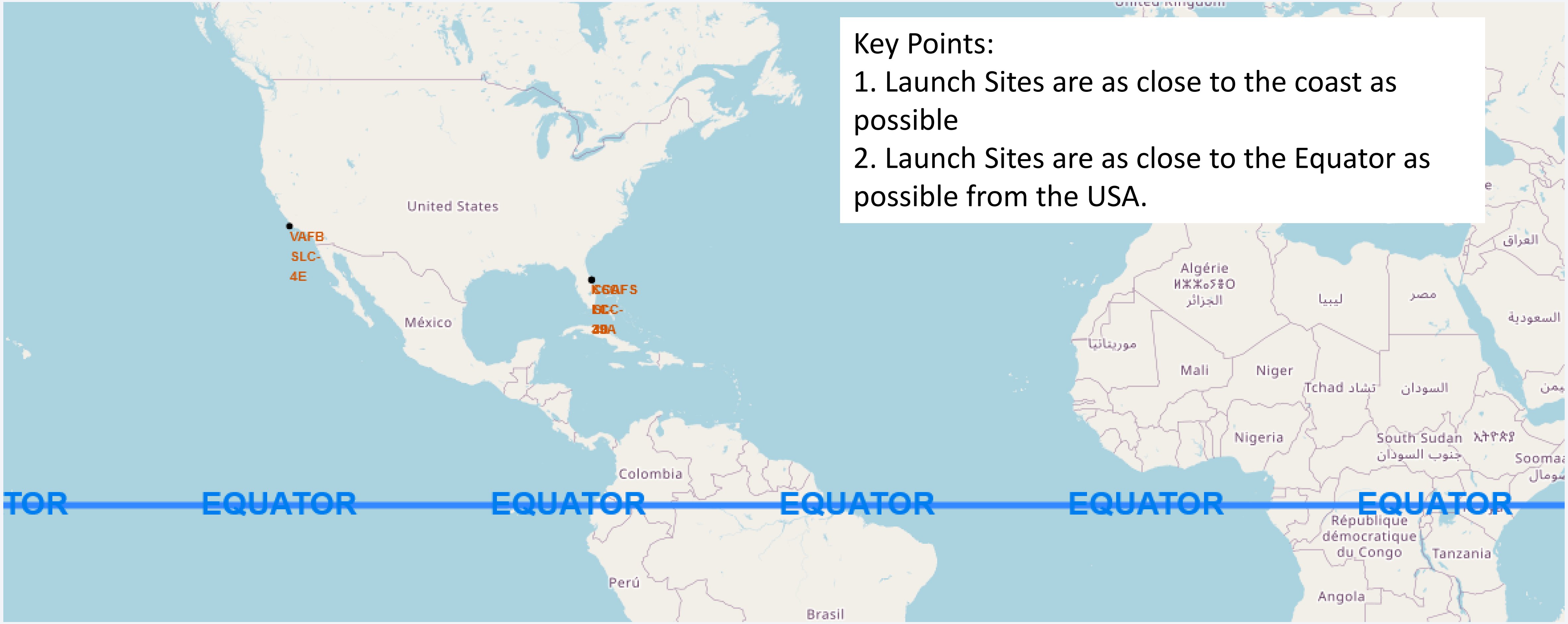
Key points:

1. In 10 launches, no attempt to land was made (this was the most common option between 2010 and 2017)
2. Success and Failure on Drone Ship landings were equal (at 5 counts each)
3. Controlled landing on a ground pad or controlled landing into the ocean were the next most common (and equal at 3 counts each).
4. Only 8 successful landings recorded against 22 unsuccessful landings.

Section 3

Launch Sites Proximities Analysis

Location of 4 Launch Sites used by SpaceX



Key Points:

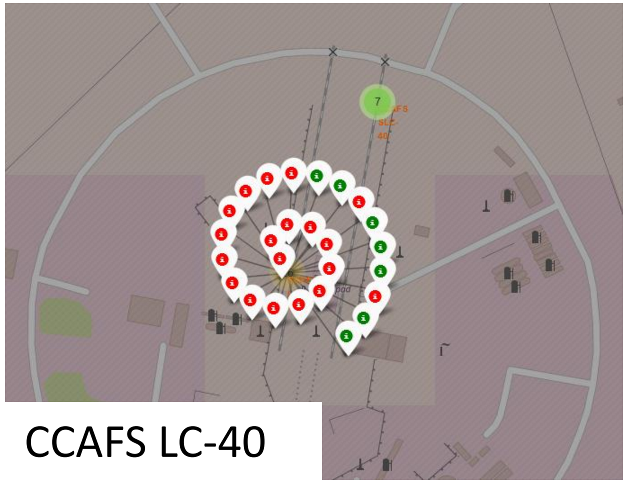
1. Launch Sites are as close to the coast as possible
2. Launch Sites are as close to the Equator as possible from the USA.

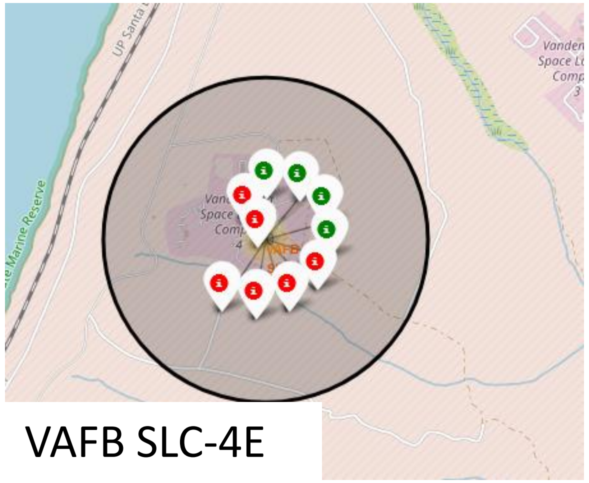
Successes and Failures from each Site (Zoomed out)

Successes and Failures from each Site (Zoomed In)

KSC LC39A


CCAFS SLC-40

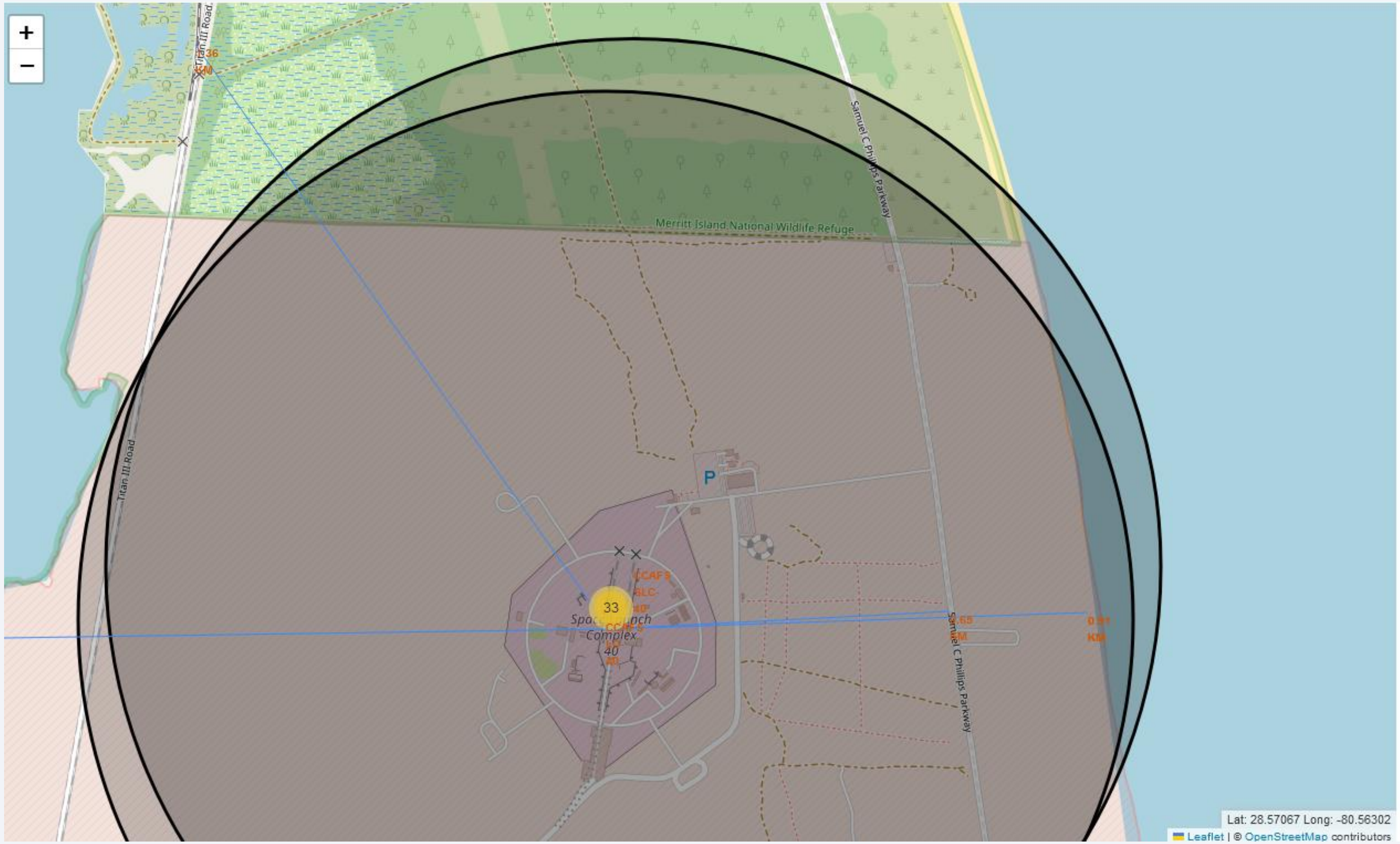

CCAFS LC-40


VAFB SLC-4E

Key Points:

1. More launches on the East Coast
2. More successes than failures at KSC LC39A

Distance From Landmarks



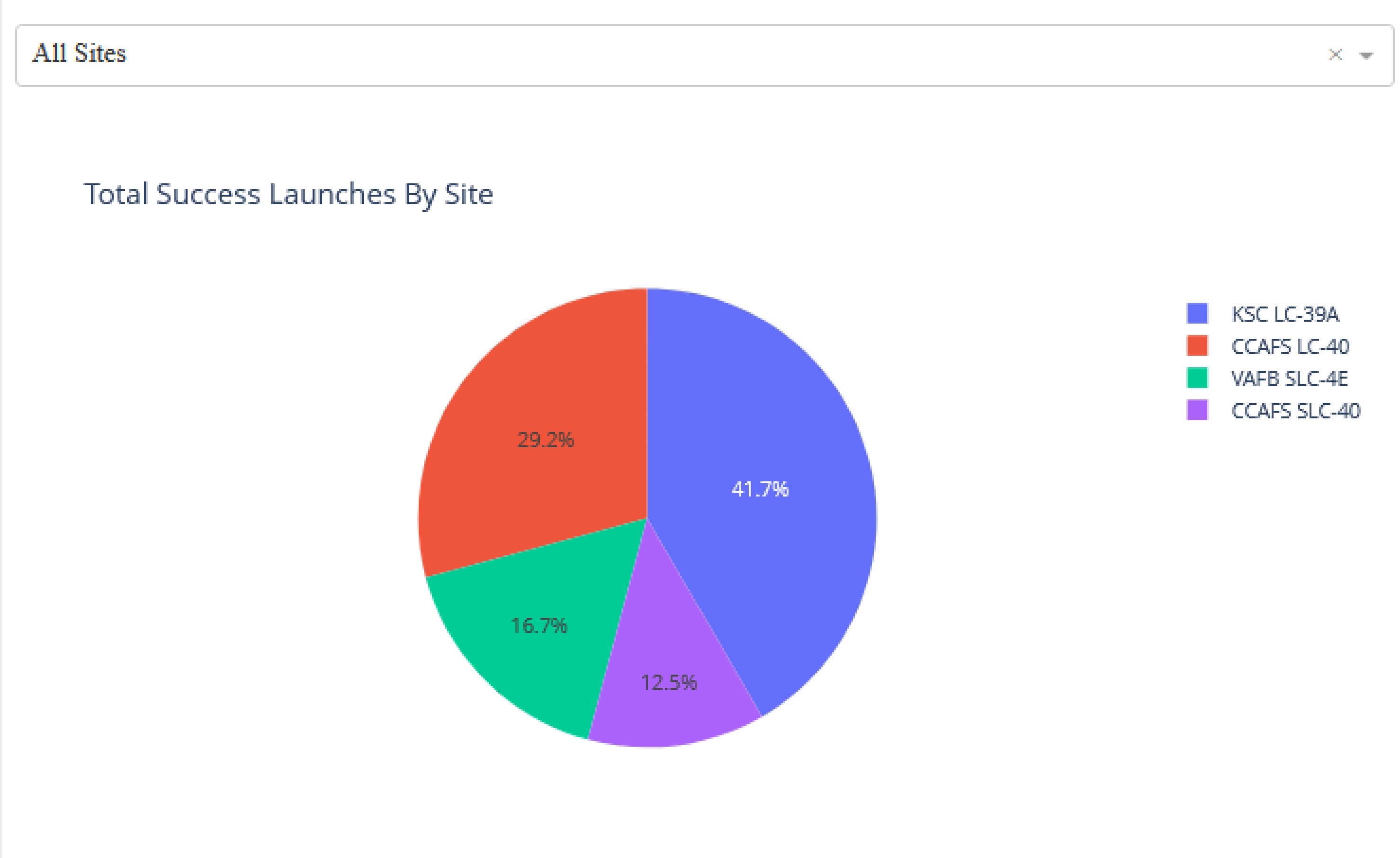
Key Points:

1. Launch Sites are usually near the coast to within a few kilometres.
2. The same is true for Railways, Roads.
3. City's (Orlando in the case of the image left) are a long distance away of upwards of 70km.

Section 4

Build a Dashboard with Plotly Dash

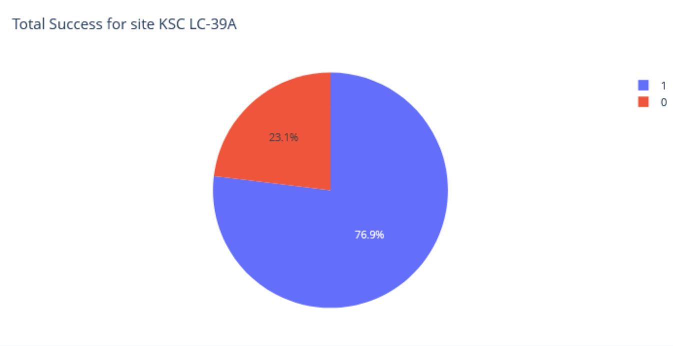
Launch Site Successes

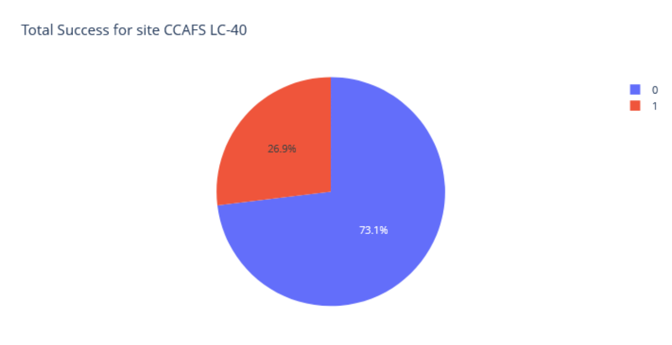


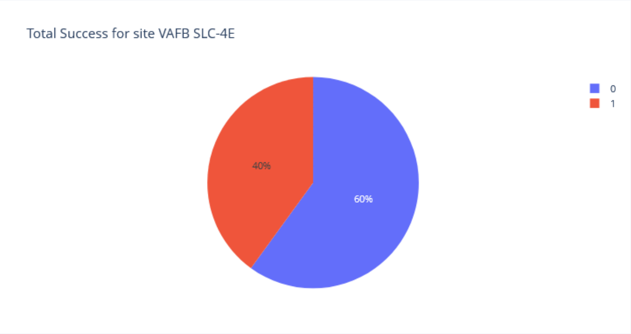
Key Points:

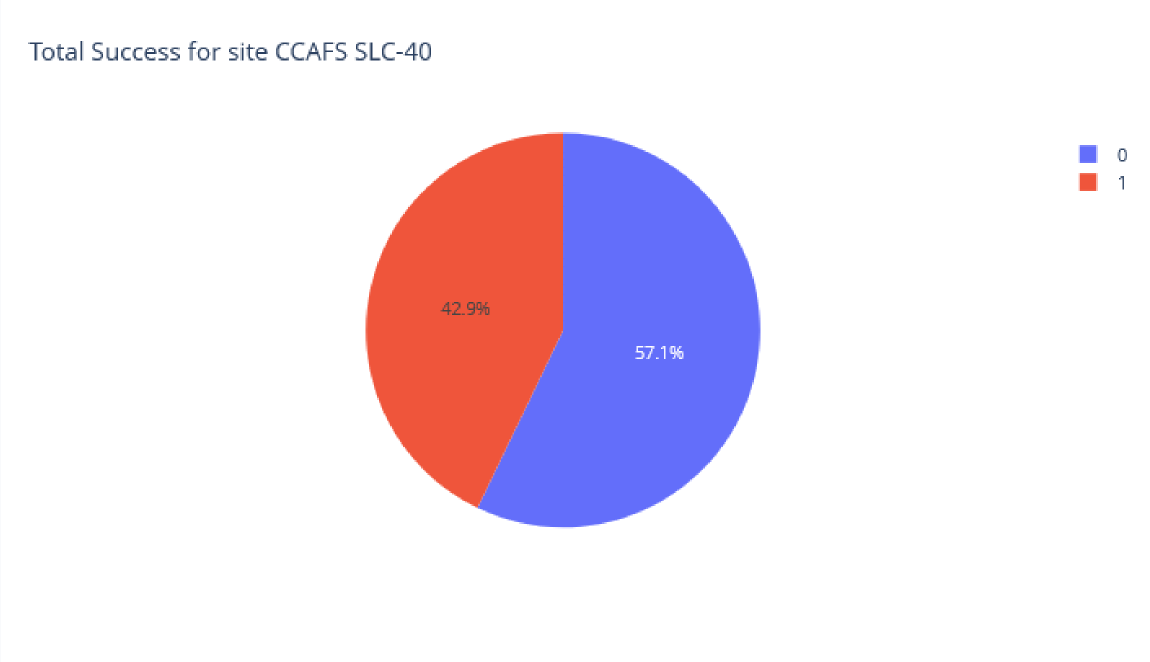
1. KSC LC-39A had the most successful landings out of the 4 launch sites.

Launch Site with highest success rate



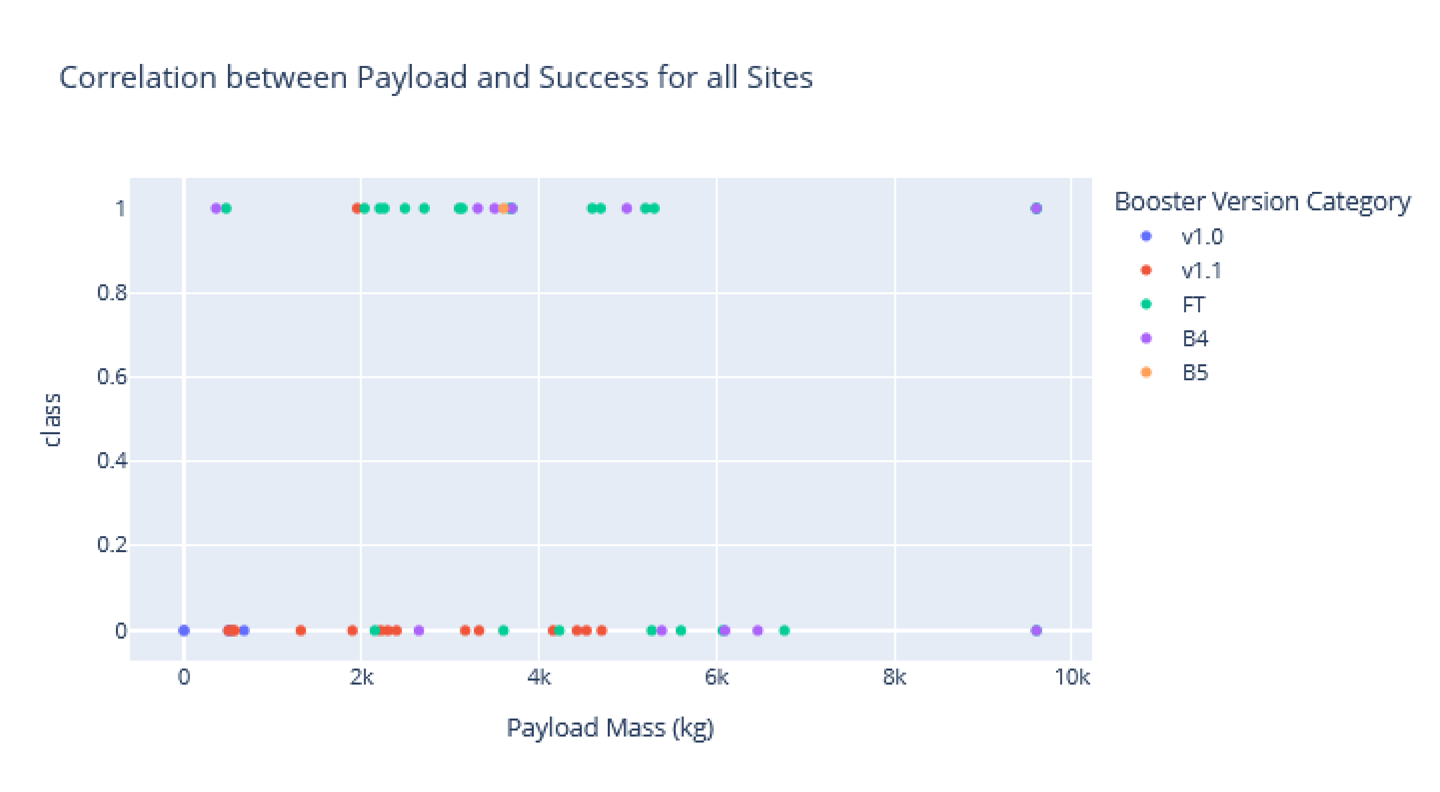






Key Points:
1. CCAFS SLC-40 had the highest success rate at 42.9%
2. KSC LC-39A (previous slide) may have had the most successes, but it also had the poorest success rate (23.1%) out of all the Launch Sites.

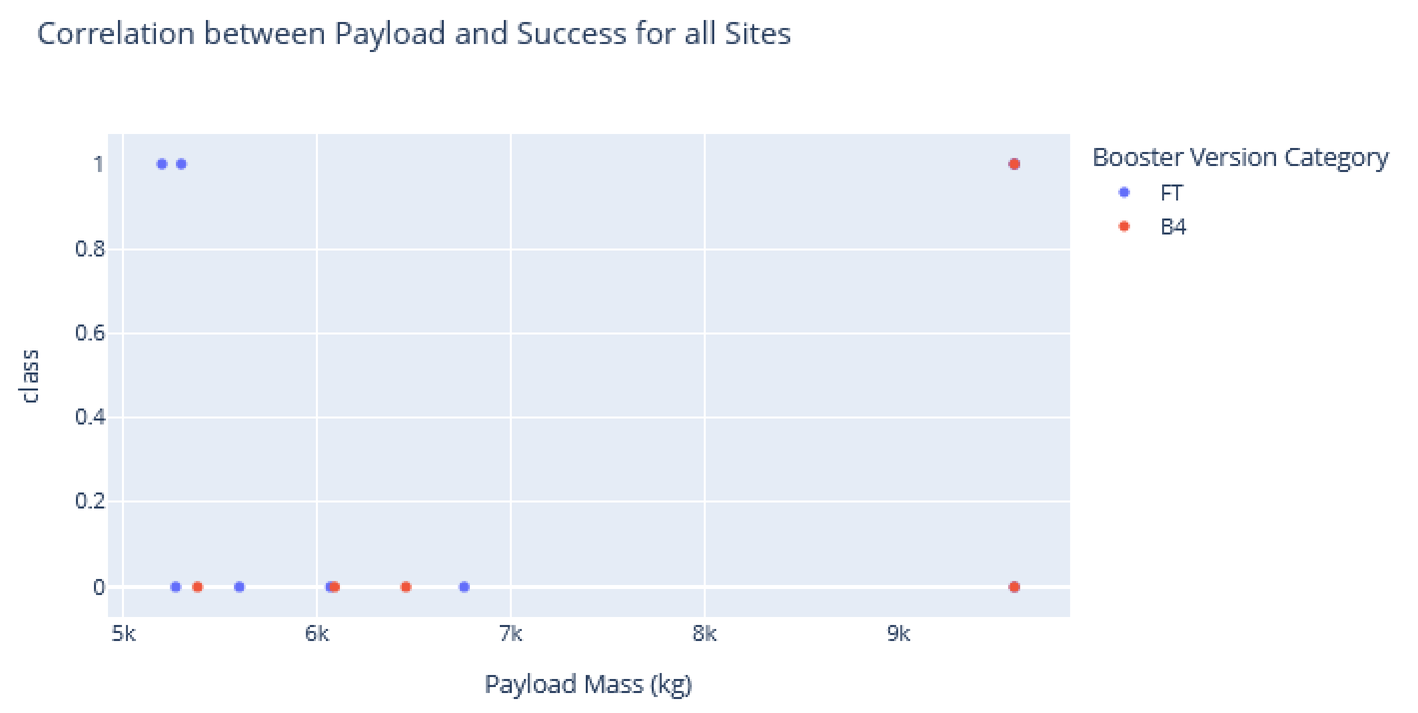
Correlations between Payload and Success



Key Points:

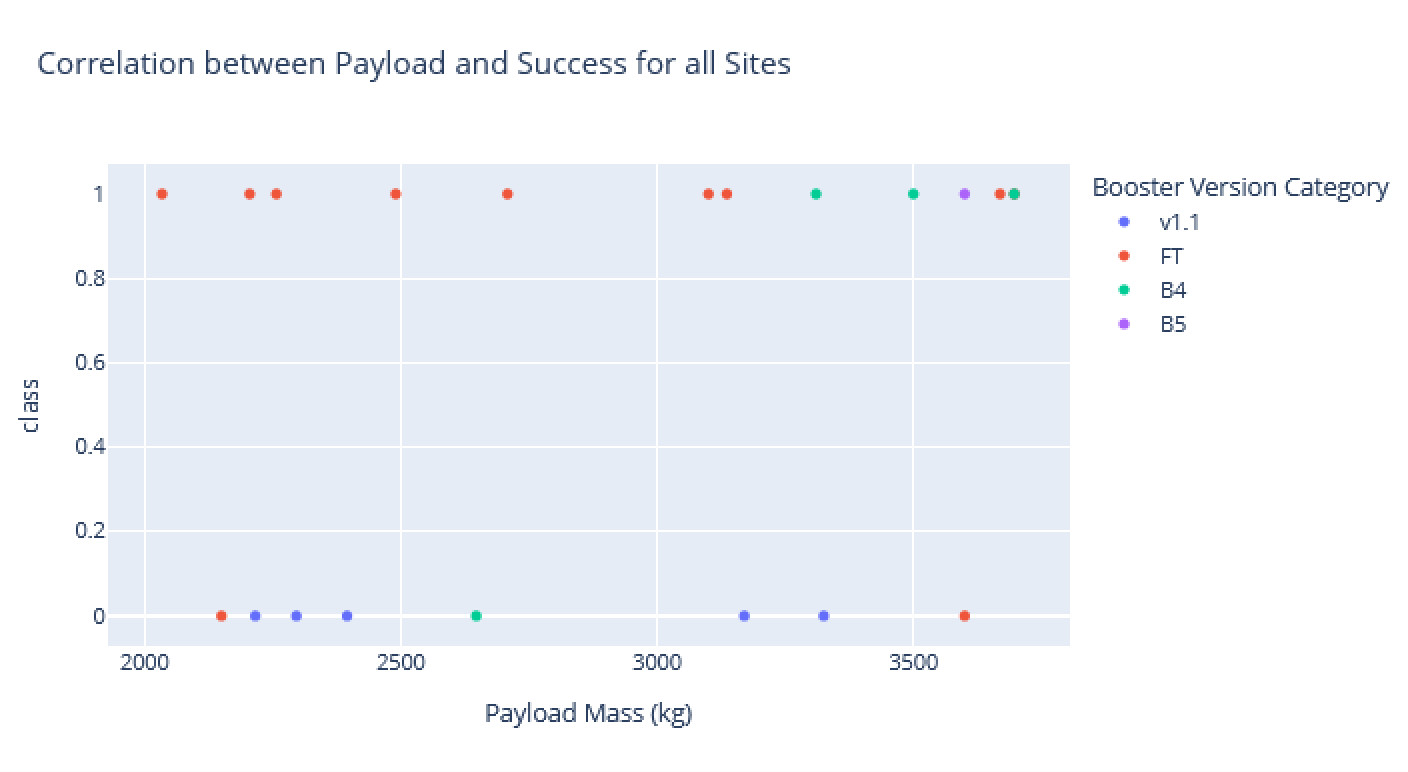
1. Across all Payloads, Booster Version FT has the most successes
2. Booster Version v1.1 has the most failures

Correlations between Payload and Success



Key Points:
1. Only 2 Booster Versions carry over 5K payloads and these usually end unsuccessfully.

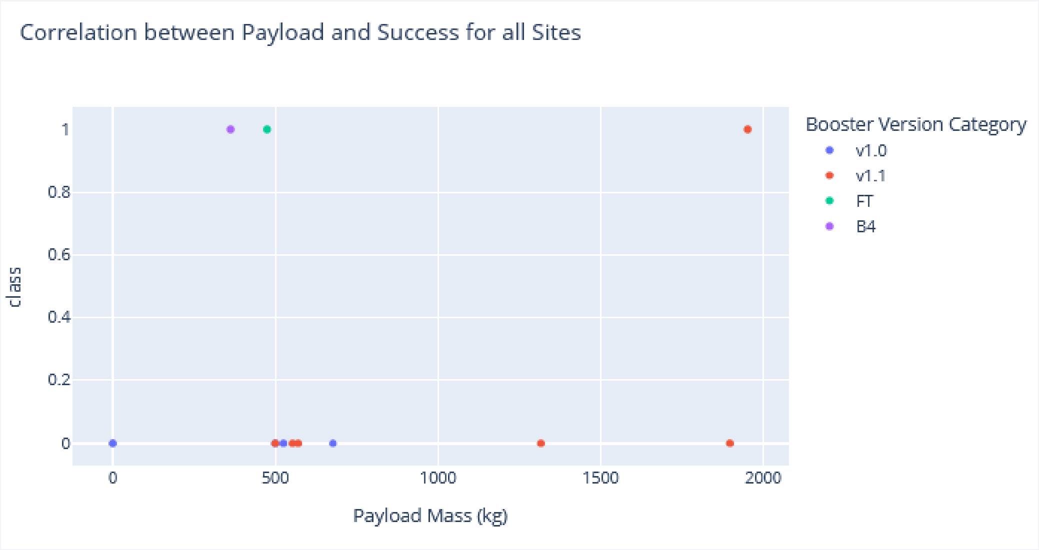
Correlations between Payload and Success

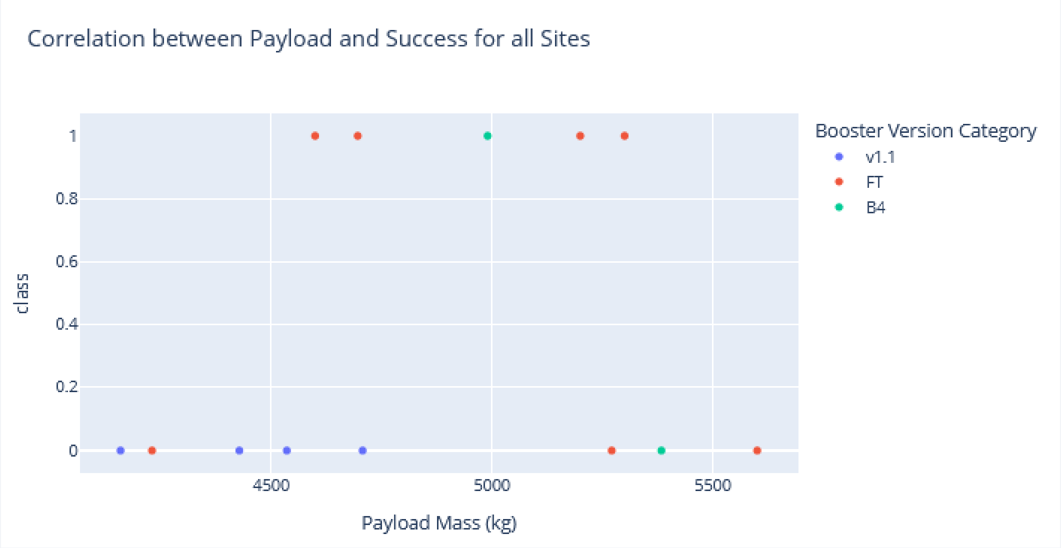


Key Points:

1. Payloads between 2K and 4K are the only interval where Success is more likely than failure.
2. Again, Success is most likely using the FT.

Correlations between Payload and Success





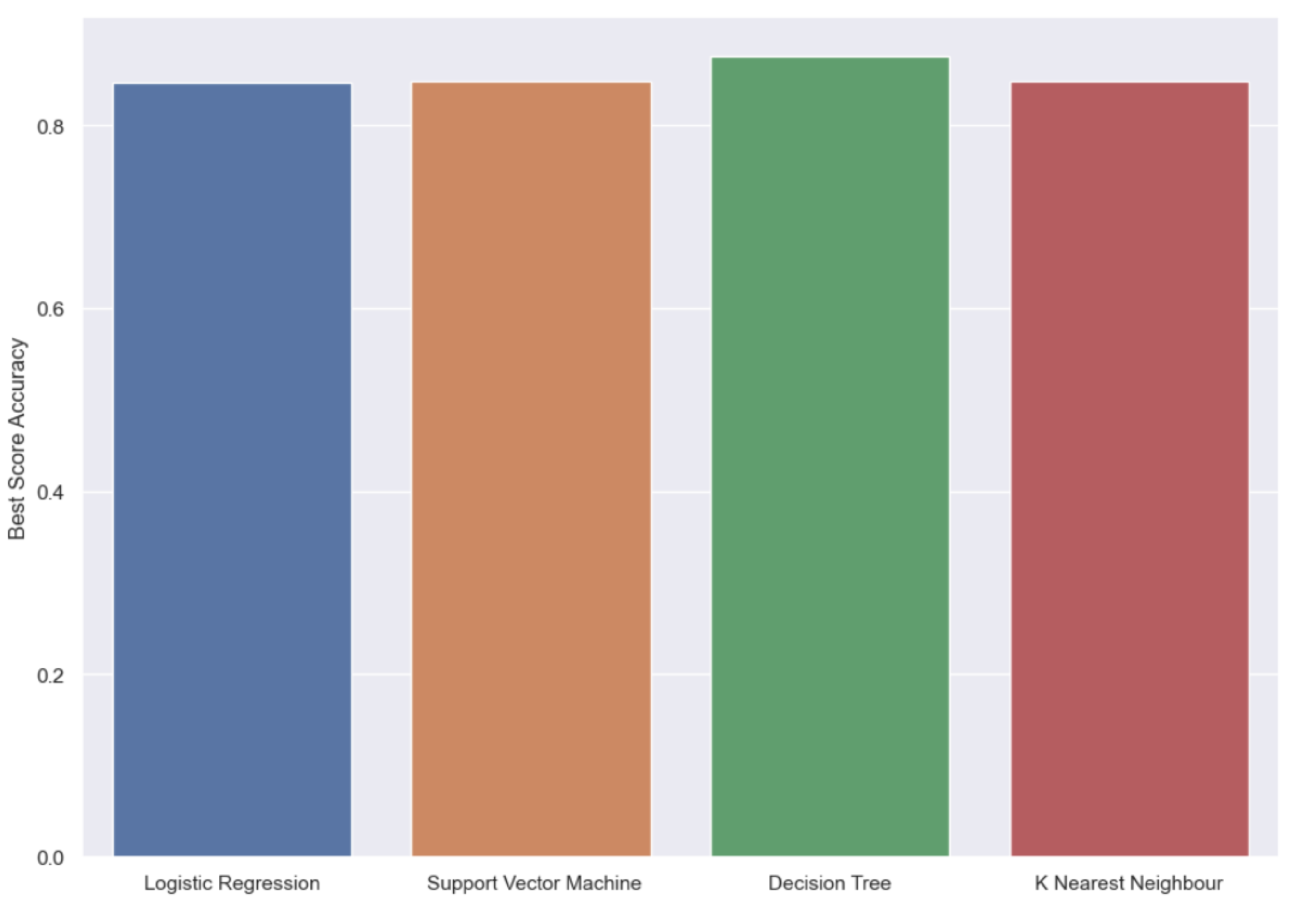
Key Points:

1. Payloads between 0 and 2K had the lowest success rate.
2. Closely followed by the range 4K – 6K.

Section 5

Predictive Analysis (Classification)

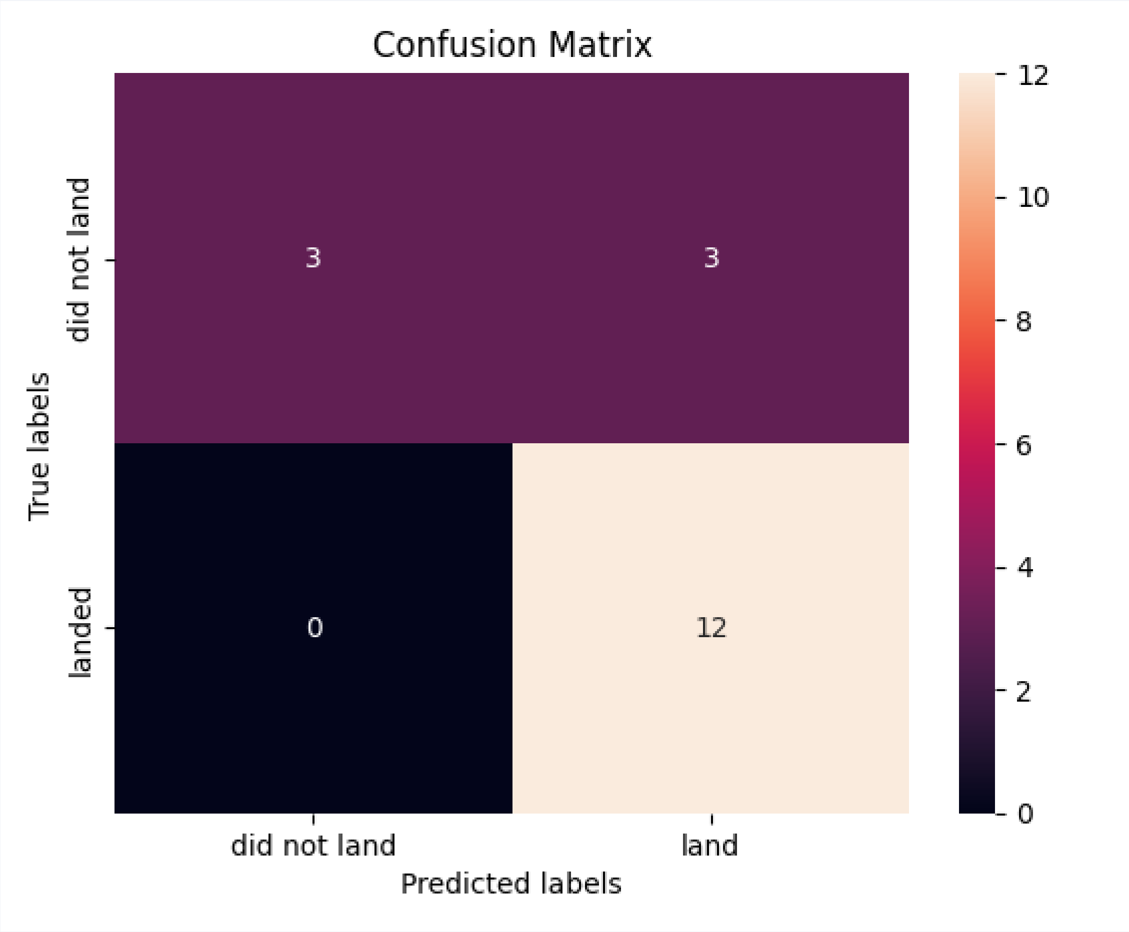
Classification Accuracy



Key Points:

1. Decision Tree has marginally the best accuracy.
2. Accuracy obtained from best_score_ from the GridSearchCV object created for Algorithm

Confusion Matrix



The Confusion Matrix shows the following:

1. For all that landed, the model predicted correctly that it would land.
2. The model found it harder to predict if a Falcon 9 Rocket would not land. It Predicted 3 False Positives and 3 True Negatives.
3. All Confusion Matrixes where the same as where the classification reports below:

```
Classification report for Decision Tree
              precision    recall  f1-score   support

           0       1.00      0.50      0.67         6
           1       0.80      1.00      0.89        12

    accuracy                           0.83        18
   macro avg       0.90      0.75      0.78        18
weighted avg       0.87      0.83      0.81        18
```

Conclusions

- The more launches that have occurred, the greater the chance of success
- The Booster Version FT is likely to produce the greatest chance of success
- Launch Sites are near the coast, close to the Equator, infrastructure but away from cities
- Launch Site CCAFS SLC-40 has the highest chance of success
- Various classification models return similar results and have a problem returning False Positives.
- Orbits ES-L1, GEO, HEO, SSO have the best success rate.

Appendix

Model Accuracies

Classification report for logistic regression

	precision	recall	f1-score	support
0	1.00	0.50	0.67	6
1	0.80	1.00	0.89	12
accuracy			0.83	18
macro avg	0.90	0.75	0.78	18
weighted avg	0.87	0.83	0.81	18

Classification report for Support Vector Machine

	precision	recall	f1-score	support
0	1.00	0.50	0.67	6
1	0.80	1.00	0.89	12
accuracy			0.83	18
macro avg	0.90	0.75	0.78	18
weighted avg	0.87	0.83	0.81	18

Classification report for Decision Tree

	precision	recall	f1-score	support
0	1.00	0.50	0.67	6
1	0.80	1.00	0.89	12
accuracy			0.83	18
macro avg	0.90	0.75	0.78	18
weighted avg	0.87	0.83	0.81	18

Classification report for K Nearest Neighbour

	precision	recall	f1-score	support
0	1.00	0.50	0.67	6
1	0.80	1.00	0.89	12
accuracy			0.83	18
macro avg	0.90	0.75	0.78	18
weighted avg	0.87	0.83	0.81	18

URL: https://github.com/rsx8/IBM_DS_SpaceX_Capstone

Thank you!