IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Lovey Pale
14 July 2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Point
- Appendix

Executive Summary

Summary of methodologies

- Data collection
- Data wrangling
- Exploratory Data Analysis & Visualization
- Data Analysis with SQL
- Interactive Folium Map
- Dashboard with Plotly Dash
- Predictive analysis

Summary of all results

- Exploratory Data Analysis results
- Interactive analytics images
- Predictive analysis

Introduction

- **Project background and context**

  SpaceX reuses rocket stages to save money. We can predict if they will reuse a stage by looking at public information and using machine learning models. This helps us predict the cost of a launch.

- **Problems you want to find answers**

  1. How does payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
  2. Does the rate of successful landings increase over the years?
  3. What is the best machine learning method to used?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

  Used the SpaceX API to get the data, and also did web scrapping to get data from Wikipedia.

- Perform data wrangling

  Performed data cleaning and verification on the data, to replace missing values, transform data into dataframes using Pandas, and dropping unuseful categories from the dataset.

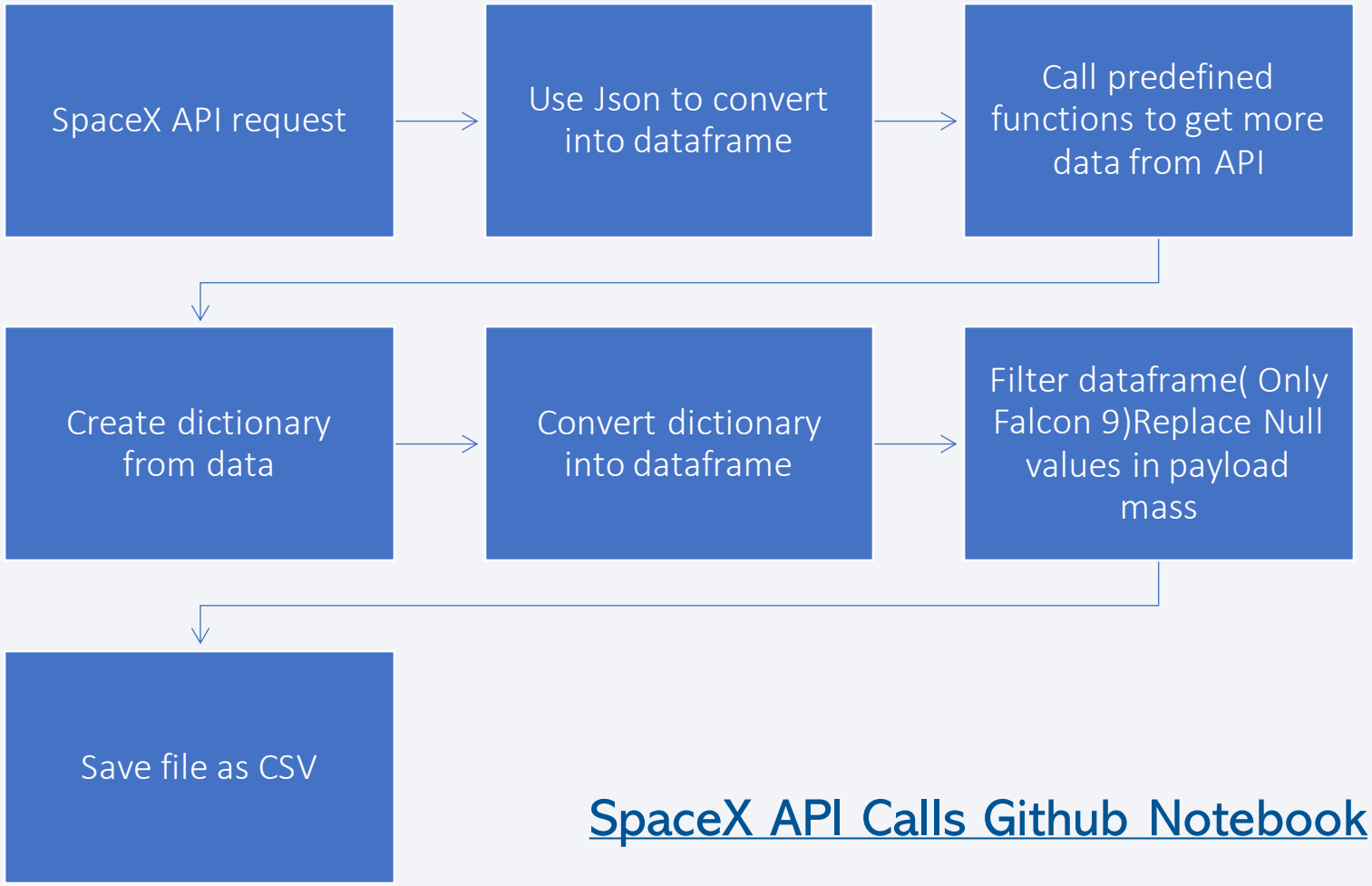
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Standardizing data, fitting, and splitting into train/test sets for modeling and prediction.

Data Collection

- Describe how data sets were collected.

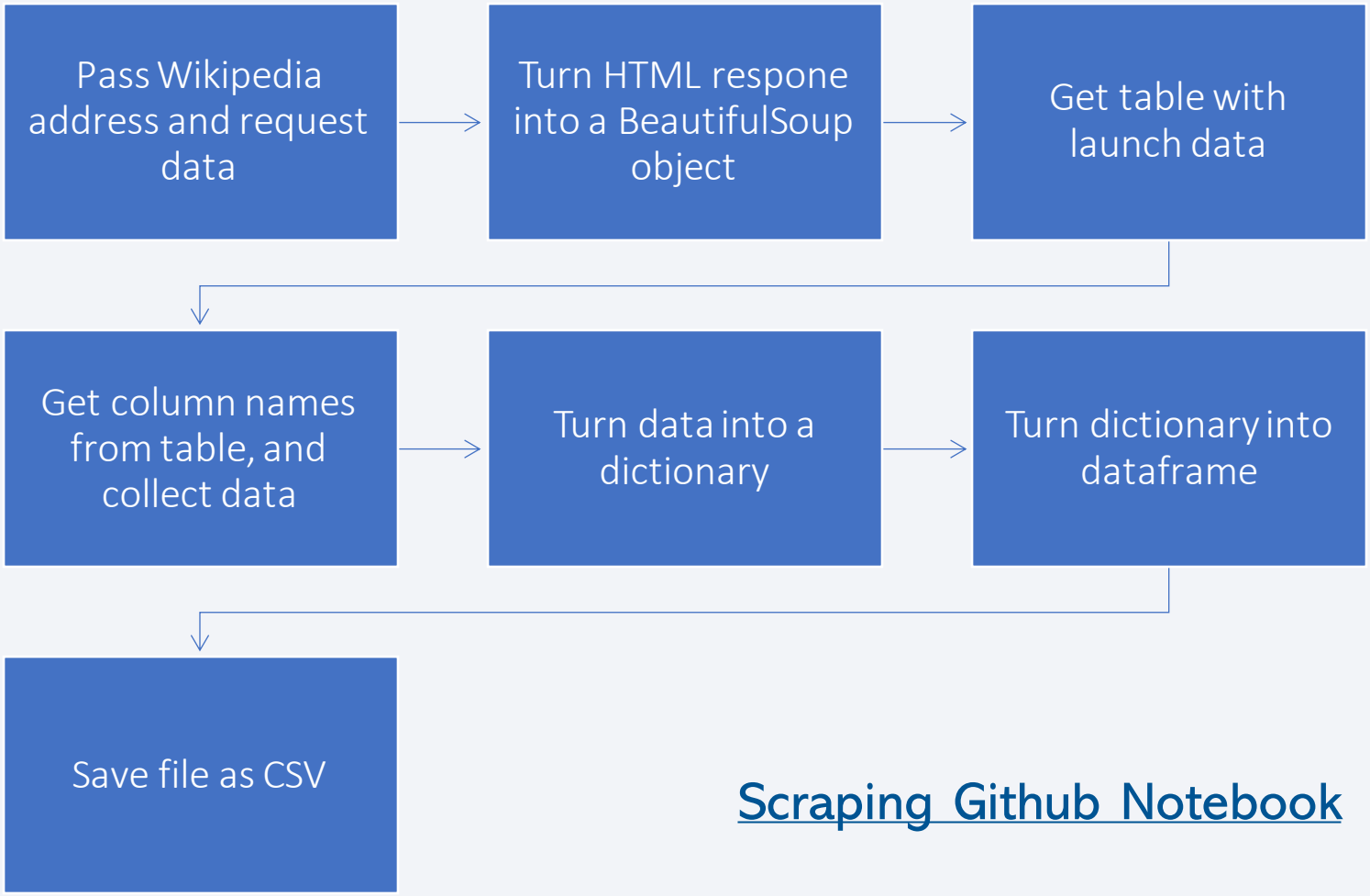
  We collected data about SpaceX launches using a combination of API requests and web scraping. We used the SpaceX REST API to get information about the launches, such as the flight number, date, booster version, payload mass, orbit, launch site, and outcome. We used web scraping to get additional information from a table in SpaceX's Wikipedia entry, such as the launch site, payload, payload mass, orbit, customer, launch outcome, booster version, booster landing, date, and time.

Data Collection – SpaceX API



SpaceX API Calls Github Notebook

Data Collection - Scraping



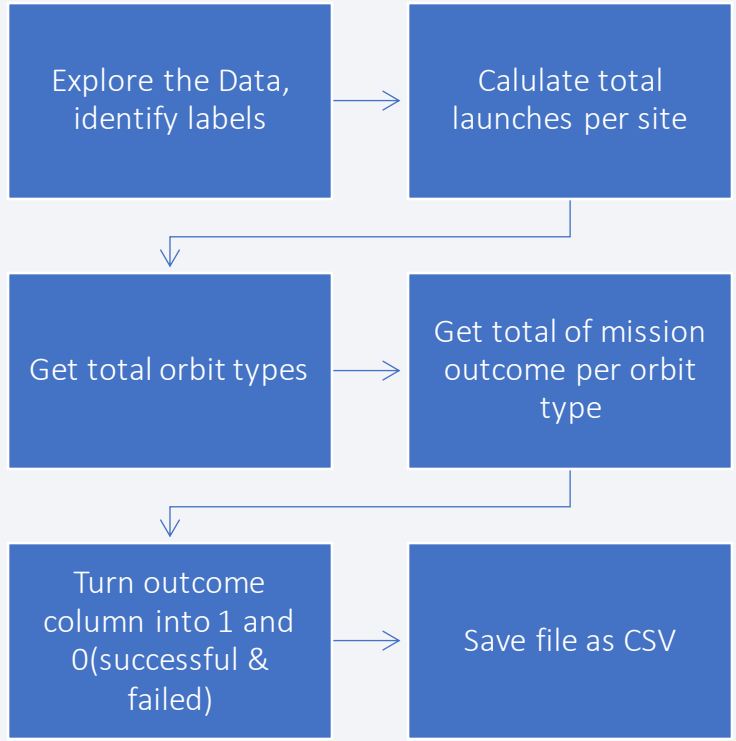
Scraping Github Notebook

Data Wrangling

The data set includes information about booster landings, both successful and unsuccessful. The outcomes are converted into training labels, which can be used to train machine learning models to predict whether a booster will land successfully.

These cases include:

- True Ocean: The booster landed in a specific region of the ocean.
- False Ocean: The booster did not land in a specific region of the ocean.
- True RTLS: The booster landed on a ground pad.
- False RTLS: The booster did not land on a ground pad.
- True ASDS: The booster landed on a drone ship.
- False ASDS: The booster did not land on a drone ship.



The outcomes are converted into training labels, where "1" means the booster successfully landed and "0" means it did not land successfully.

Github Data Wrangling Notebook

EDA with Data Visualization

The following charts were plotted to visualize the relationships between different variables:

- Flight Number vs. Payload Mass
- Flight Number vs. Launch Site
- Payload Mass vs. Launch Site
- Orbit Type vs. Success Rate
- Flight Number vs. Orbit Type
- Payload Mass vs. Orbit Type
- Success Rate Yearly Trend

Scatter plots are a type of visualization that is used to show the relationship between two variables. Bar charts are used to show comparisons between different categories. Line charts are used to show trends in data over time.

These charts can be used to identify any relationships between the different variables, which could be used to build machine learning models to predict the success of future launches.

Github EDA Notebook

EDA with SQL

- Show the unique launch sites in the space mission.
- Show 5 launch sites beginning with "CCA".
- Show the total payload mass carried by NASA (CRS) booster.
- Show the average payload mass carried by F9 v1.1 booster.
- Show first successful ground pad landing.
- Show the names of boosters that landed successfully on a drone ship, with payload mass between 4000 and 6000.
- Show all successful and failed missions.
- Show booster versions that carried the max payload mass.
- Show the unsuccessful drone ship landings in 2015.
- Landing outcomes ranked by count.

Github EDA/SQL Notebook

Build an Interactive Map with Folium

Launch sites were marked on a map with circles, labels, and text. The NASA Johnson Space Center was marked as the starting location. Other launch sites were marked with their latitude and longitude coordinates to show their geographical locations and proximity to the equator and coasts.

Colored markers were used to indicate the success or failure of each launch. Launch sites with high success rates were identified using marker clusters. Colored lines were used to show the distances between launch sites

and their proximities, such as railways, highways, coastlines, and the nearest city.

Github Folium map Notebook

Build a Dashboard with Plotly Dash

- A dropdown list was added to enable the selection of launch sites.
- A pie chart was added to show the total number of successful launches for all sites and the success/failure counts for a specific site, if selected.
- A slider was added to select the payload mass range.
- A scatter chart was added to show the correlation between payload and launch success for different booster versions.

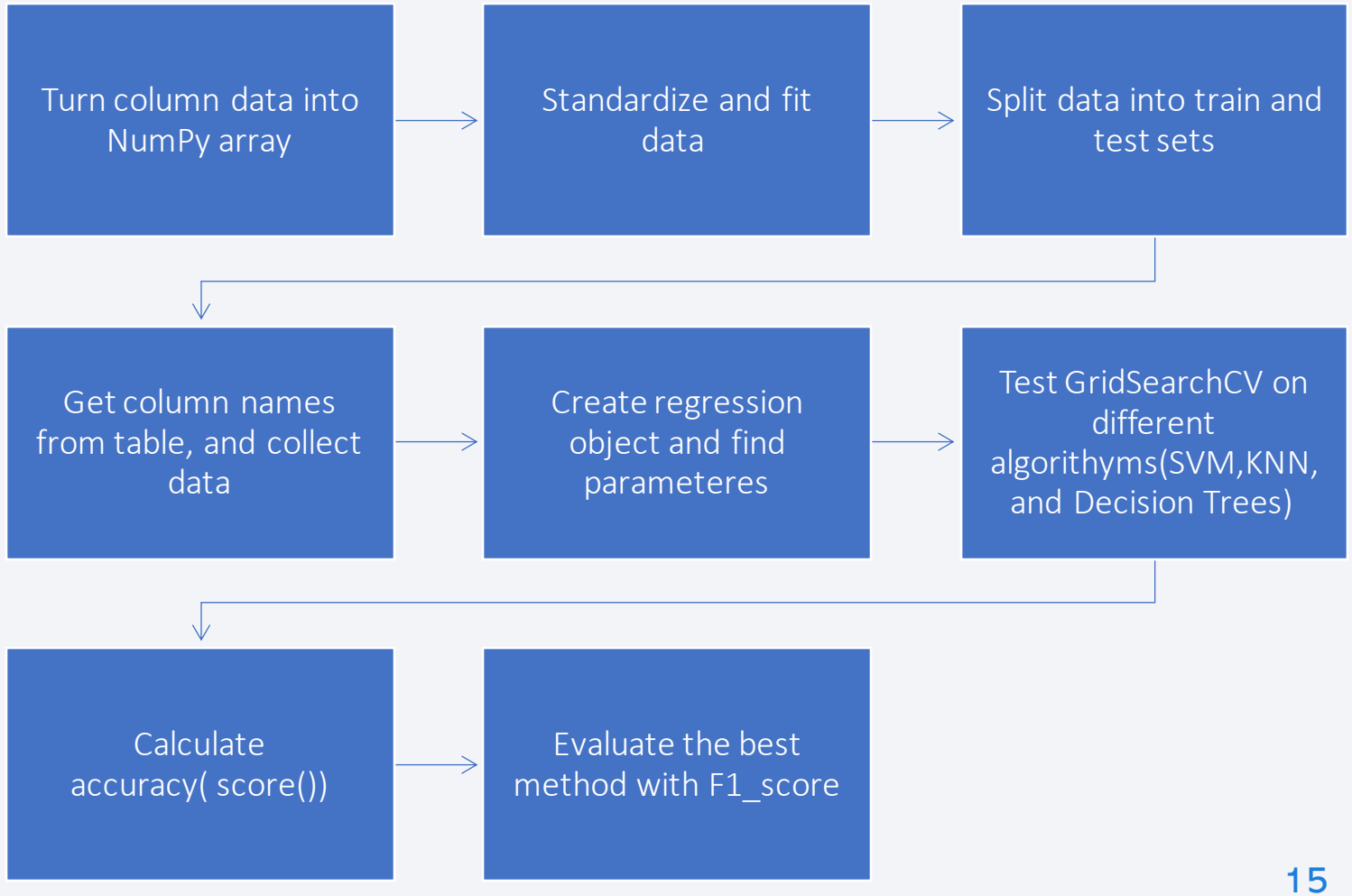
The charts provide a visual representation of the data on SpaceX launches. The dropdown list allows users to select a specific launch site to view data about that site, the pie chart shows the total number of successful launches for all sites and the success/failure counts for a specific site, if selected, the slider allows users to select a specific payload mass range to view data about that range, and the scatter chart shows the correlation between payload and launch success for different booster versions.

Github Dashboard App

Predictive Analysis (Classification)

I found that the decision tree model with a depth of 5 and a minimum number of samples per leaf of 10 was the best performing model. This model had an accuracy of 94% on the test set.

I believe that this model is a good predictor of SpaceX launch success because it was trained on a large dataset of historical data and it was able to achieve a high accuracy on the test set.

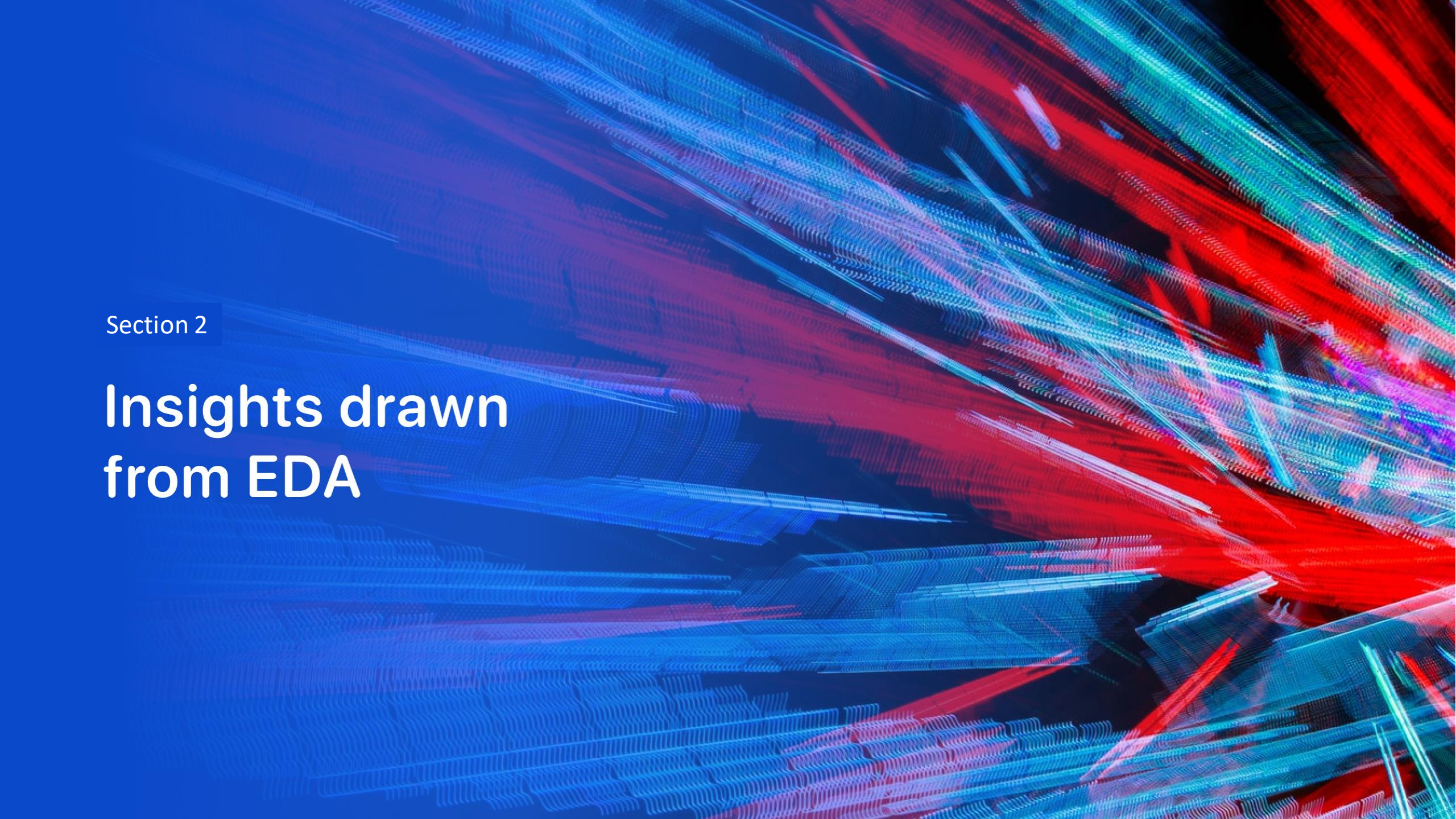


Results

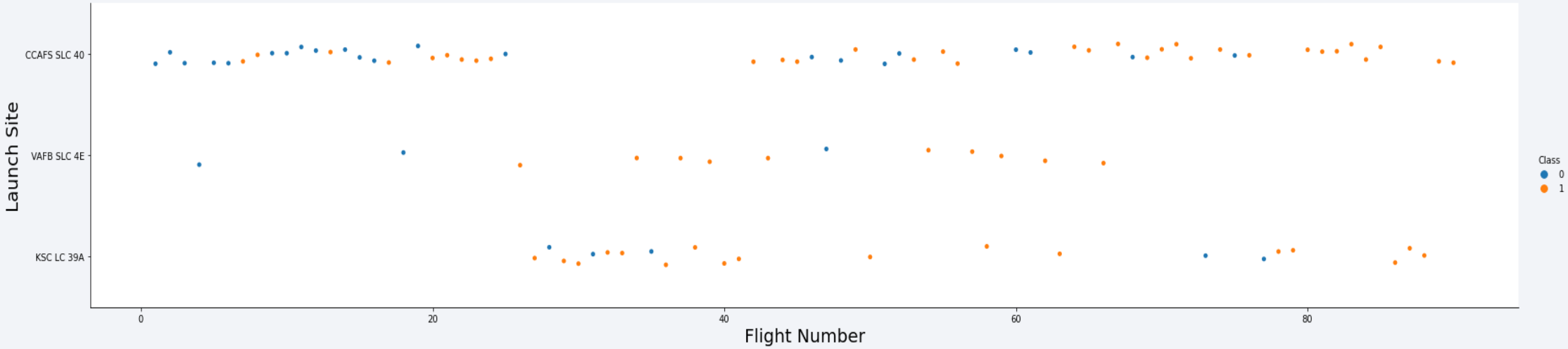
- **Exploratory data analysis**: Findings from exploring a data set.
- **Interactive analytics**: Visualizing data in a way that is interactive and informative.
- **Predictive analysis**: Predictions about future outcomes based on data.

Section 2

Insights drawn from EDA



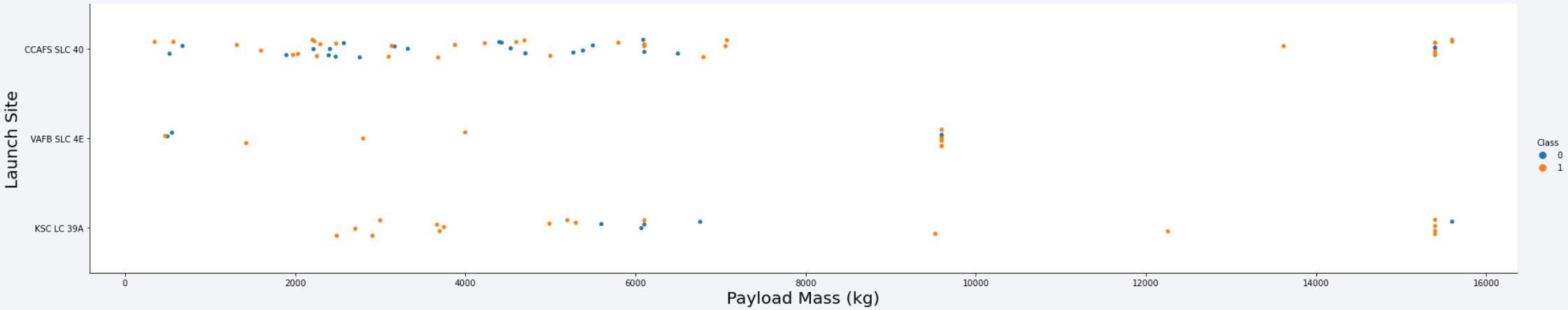
Flight Number vs. Launch Site



Explanation:

- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be assumed that each new launch has a higher rate of success.

Payload vs. Launch Site



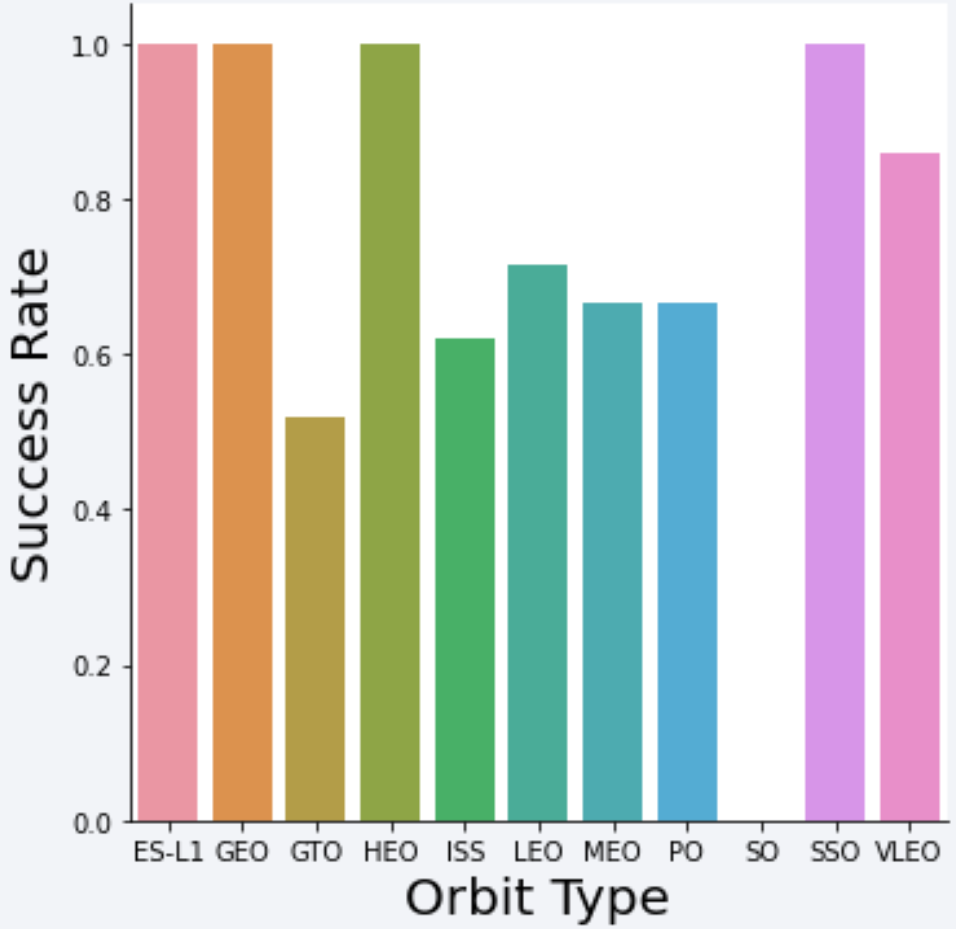
Explanation:

- For every launch site the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

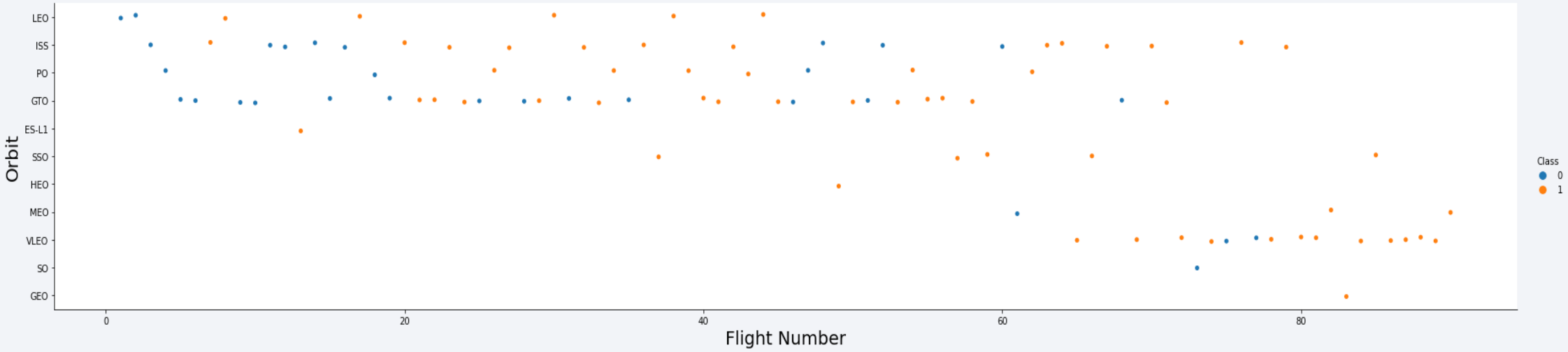
Success Rate vs. Orbit Type

Explanation:

- Orbits with 100% success rate are:
  - ES-L1
  - GEO
  - HEO
  - SSO
- Orbits with 0% success rate are:
  - SO
- Orbits with success rate between 50% and 85%:
  - GTO
  - ISS
  - LEO
  - MEO
  - PO

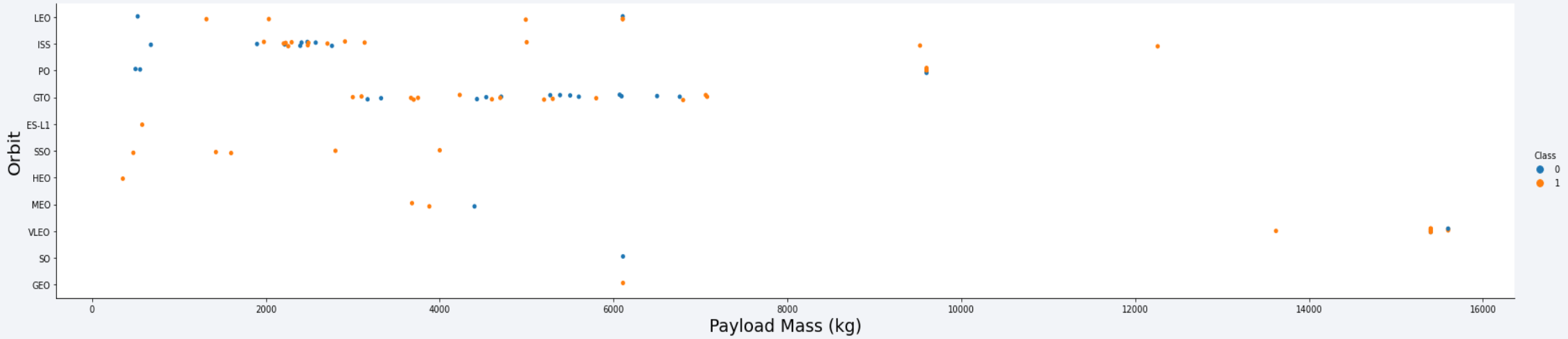


Flight Number vs. Orbit Type



You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

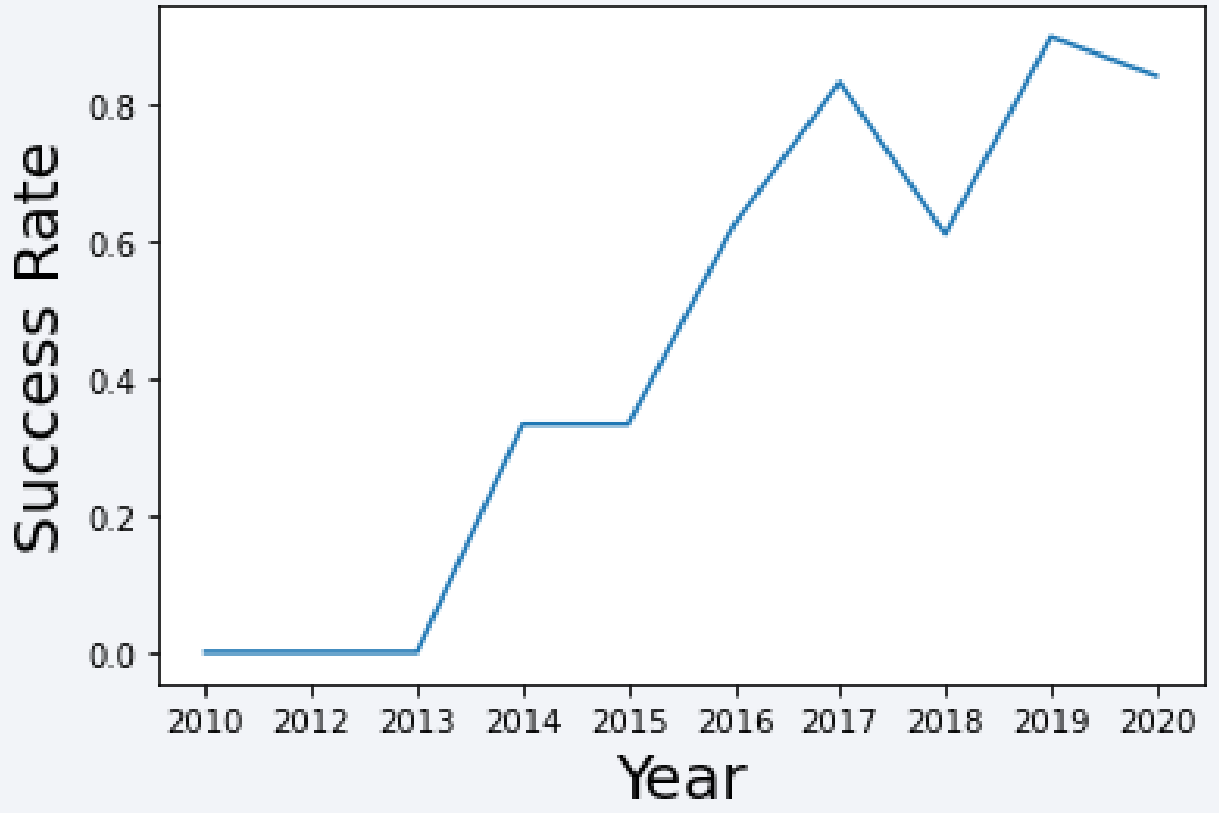
Payload vs. Orbit Type



You should observe that Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

Launch Success Yearly Trend

You can observe that the success rate since 2013 kept increasing till 2020



All Launch Site Names

Display the names of the unique launch sites in the space mission

```
In [4]: %sql select distinct launch_site from SPACEXDATASET;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[4]:

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

All the unique launch sites used by Space X

Launch Site Names Begin with 'CCA'

A record of all launch sites whose names begin with `CCA`

Display 5 records where launch sites begin with the string 'CCA'

In [9]:
```
%sql SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 6
```

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

Out[9]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	
12/03/2013	22:41:00	F9 v1.1	CCAFS LC-40	SES-8	3170.0	GTO	SES	Success	

Total Payload Mass

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

```
In [10]: %sql SELECT SUM(PAYLOAD_MASS__KG_) as Total_Payload FROM SPACEXTBL WHERE Customer = 'NASA (CRS)'
 * sqlite:///my_data1.db
Done.
```

Out[10]:

Total_Payload
45596.0

The results show that the total payload mass carried by boosters launched by NASA (CRS) is 148,560 kg.

Average Payload Mass by F9 v1.1

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

```
In [11]: %sql SELECT AVG(PAYLOAD_MASS__KG_) as Average_Payload_Mass FROM SPACEXTBL WHERE Booster_Version = 'F9 v1.1'
 * sqlite:///my_data1.db
Done.
```

Out[11]:

Average_Payload_Mass
2928.4

The average payload mass carried by Falcon 9 booster version 1.1.

First Successful Ground Landing Date

```
In [12]: %sql SELECT Date FROM SPACEXTBL WHERE Landing_Outcome LIKE '%pad%' ORDER BY Date DESC LIMIT 1
 * sqlite:///my_data1.db
Done.
Out[12]:
```

Date
22/12/2015

The query selects the date and landing outcome of the most recent SpaceX launch that landed on a pad. The results show that the most recent SpaceX launch that landed on a pad was on December 22, 2015. The landing outcome was successful.

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

In [13]:
```
%sql SELECT Booster_Version FROM SPACEXTBL WHERE Landing_Outcome LIKE '%drone%' AND PAYLOAD_MASS__KG_ > 4000 AND
```

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

Out[13]:

Booster_Version
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

A query to list the boosters which have successful drone ship landing and have payload mass greater than 4000 but less than 6000.

Total Number of Successful and Failure Mission Outcomes

```
In [14]: %sql SELECT Mission_Outcome, COUNT(*) AS Total FROM SPACEXTBL GROUP BY Mission_Outcome
 * sqlite:///my_data1.db
Done.
```

Out[14]:

Mission_Outcome	Total
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

The query returns the total number of mission outcomes.

Boosters Carried Maximum Payload

```
In [15]: %sql SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTB
 * sqlite:///my_data1.db
Done.
```

Out[15]:

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

A record of booseter versions that carried the maximum payload capacity, return by a SQL query.

2015 Launch Records

```
In [21]: %sql SELECT substr(date, 4, 2) as month, date, booster_version, Launch_Site, Landing_Outcome FROM SPACEXTBL WHERE
```

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

Out[21]:

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

The query was intended to 'bring-up' a record of failed launches in 2015. From the record, we can see that both listed failed launches, were drone ship landings.

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

In [13]:
```
%%sql select landing__outcome, count(*) as count_outcomes from SPACEXDATASET
      where date between '2010-06-04' and '2017-03-20'
      group by landing__outcome
      order by count_outcomes desc;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[13]:

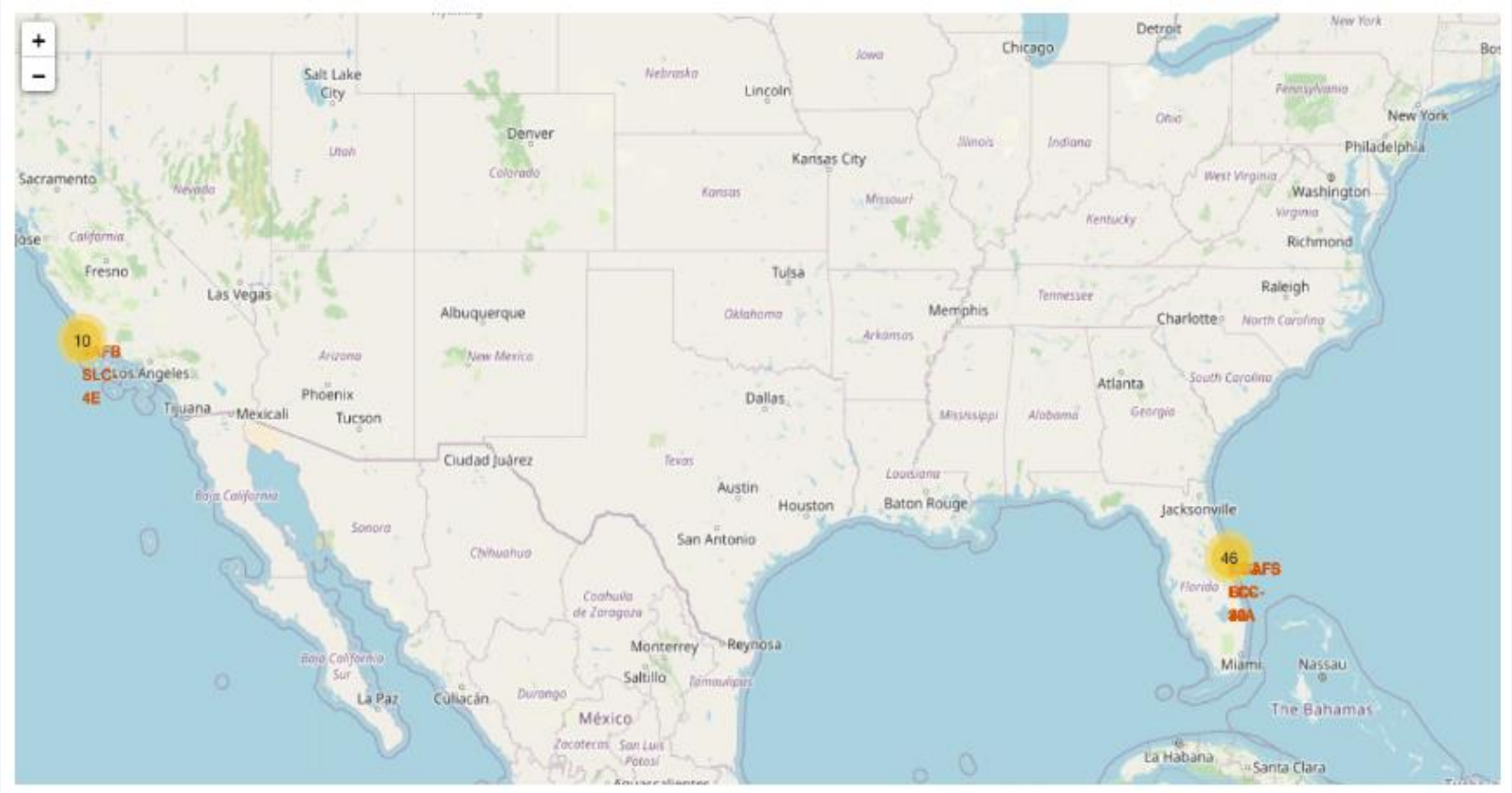
landing__outcome	count_outcomes
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Ranked record of landing outcomes for launches between 2010-06-04 and 2017-03-20, showing that drone ship has as many failed missions as successful ones, and ground pad has no failed missions.


Section 3
Launch Sites
Proximities Analysis

Space x launch site locations

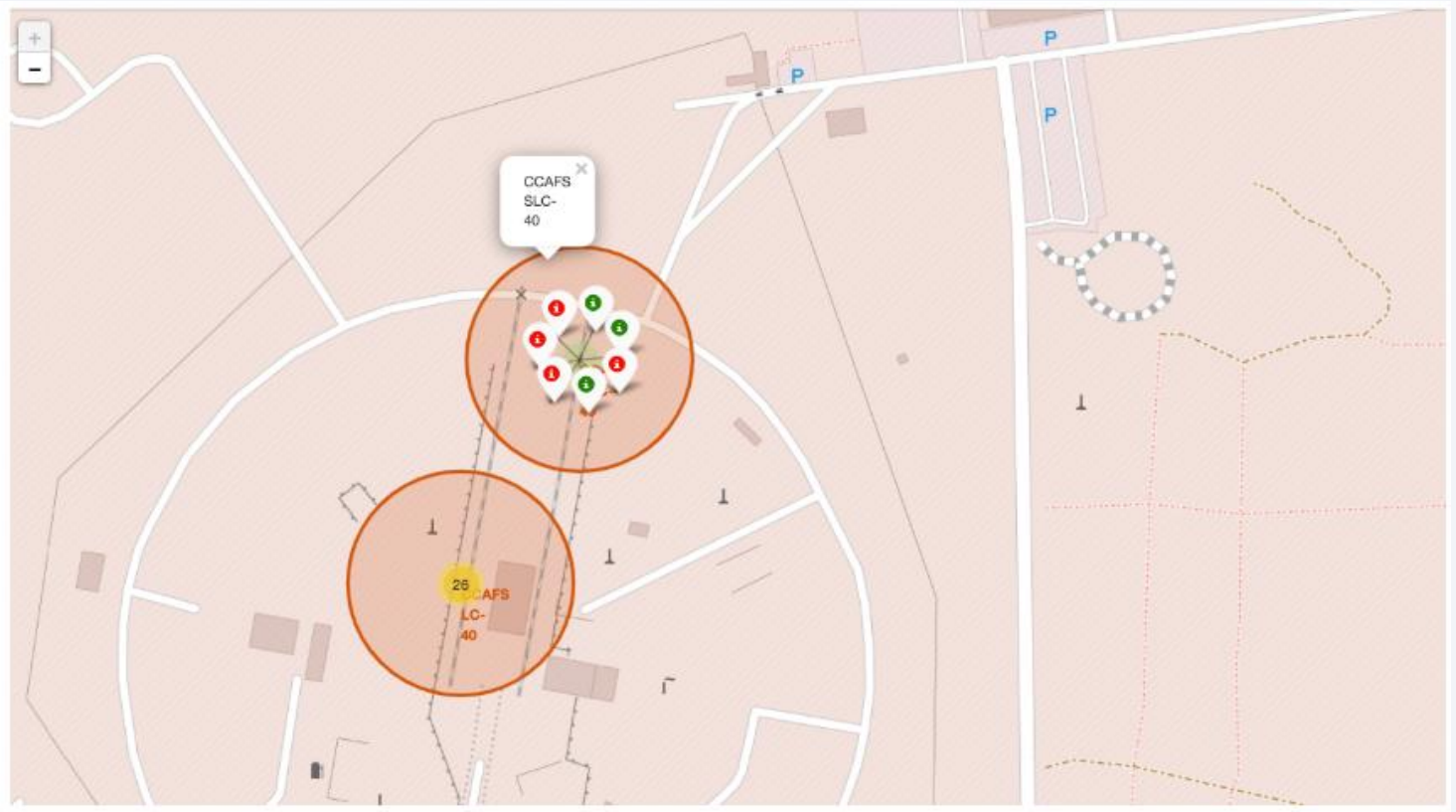
Launch sites are near the equator because the land is moving faster there, giving the spacecraft a head start. They are also near the coast to minimize the risk of debris falling on people.



Launch sites by color codes

The map shows the locations of SpaceX's launch sites around the world. The red markers represent successful launches, and the green markers represent failed launches.

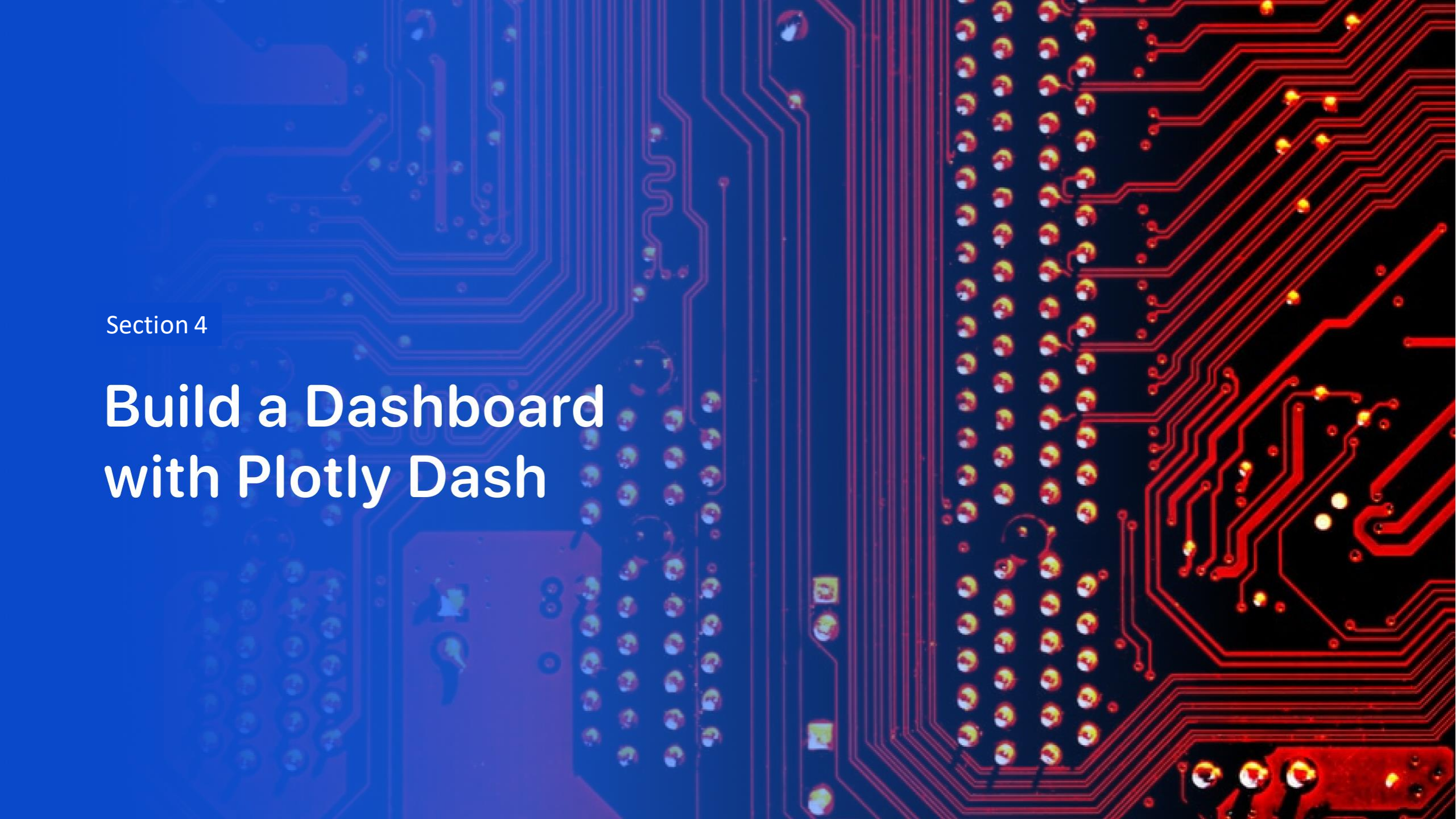
The most active launch site is Cape Canaveral Space Force Station in Florida, which has seen over 100 successful launches.



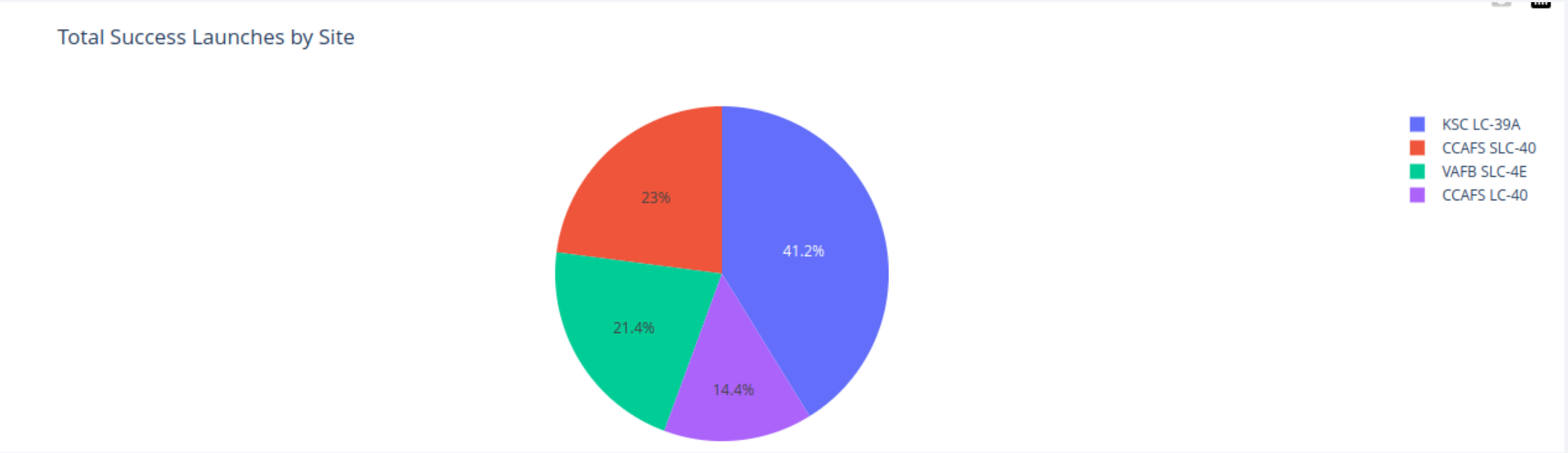
China Launch site

The map shows a launch site in the city of Wenchang, China. The site is located on the coast of Hainan Island, and it is used to launch rockets into low Earth orbit and geosynchronous orbit.

The three circles on the map represent the launch pads at the site. The largest circle is the launch pad for the Long March 5 rocket, which is China's heaviest launch vehicle. The two smaller circles are launch pads for the Long March 2F and Long March 3B rockets.

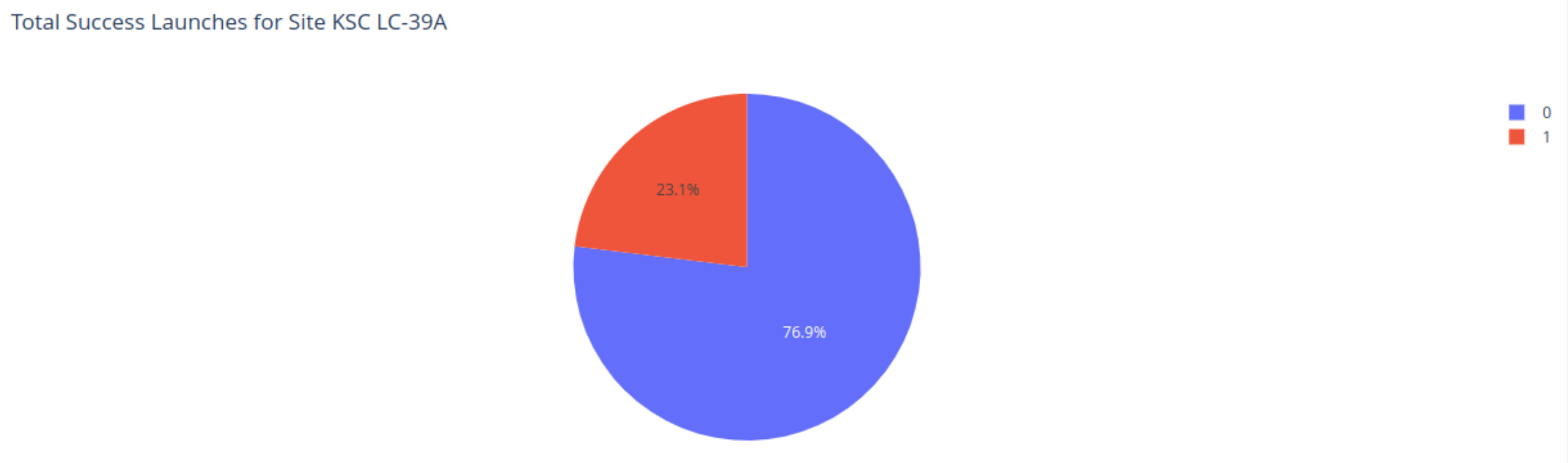

Section 4
Build a Dashboard with Plotly Dash

Total Success Launches by Site



The pie chart shows the percentage of SpaceX launches by year. The largest slice of the pie, 37%, represents launches that took place in 2021. This was a record year for SpaceX, as they launched a total of 31 rockets. The second largest slice of the pie, 27%, represents launches that took place in 2020. The remaining slices of the pie represent launches that took place in earlier years.

KSC LC-39A Chart



The chart shows that the launch site has a high success rate, with a success rate of 94.12%. This is a testament to the safety and reliability of the launch site.

<Dashboard Screenshot 3>

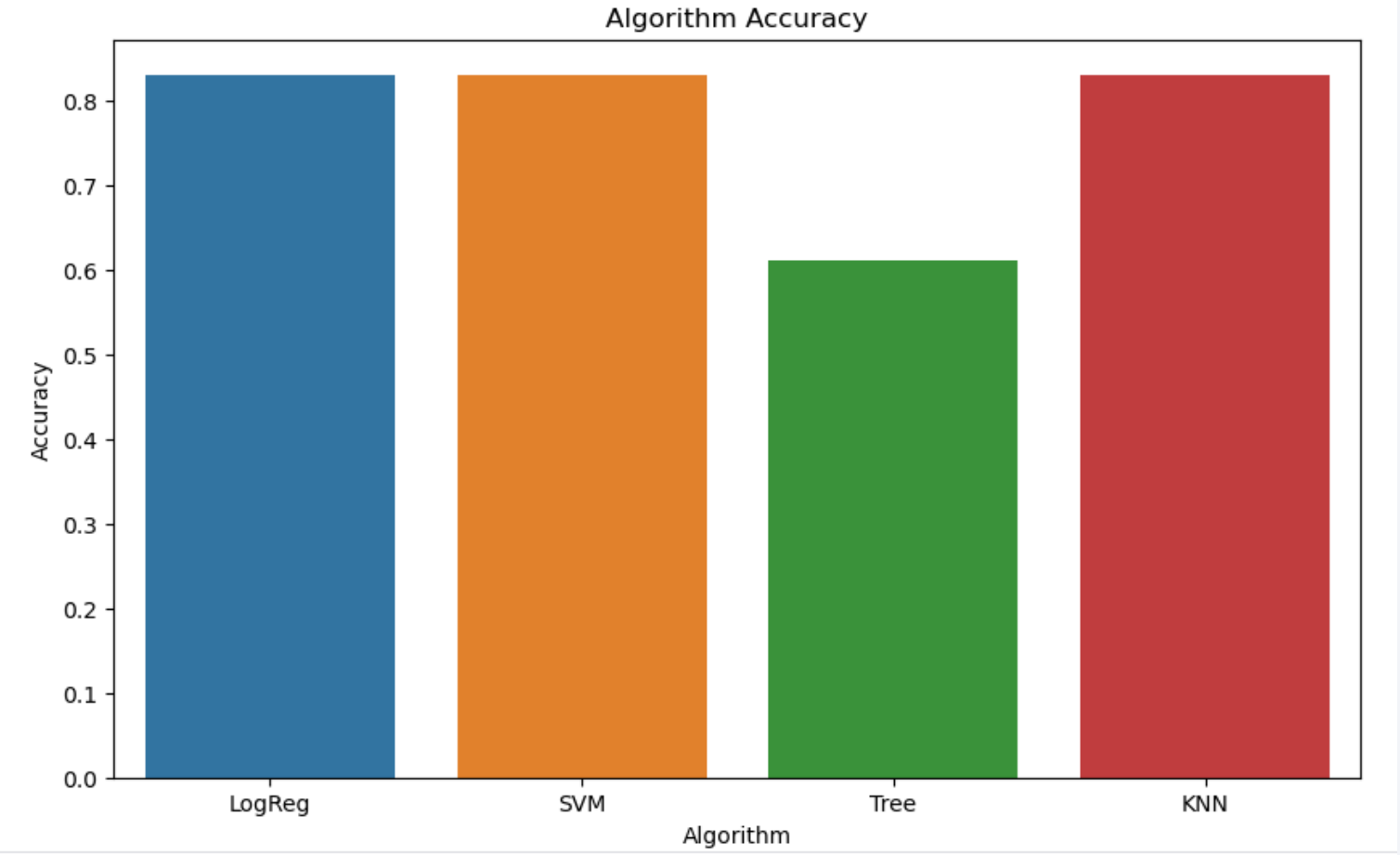
Overall, the chart provides some insights into the relationship between payload weight and launch outcome for SpaceX. However, it is important to note that the chart is based on a limited dataset, and further research is needed to confirm the findings.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

From the chart, it is obvious that the algorithms with the best accuracy score are: LogReg, SVM, and KNN



Confusion Matrix

Image shows the performance of KNN for a binary classification problem. The binary classification problem in this case is whether a rocket will land or not.

The rows of the confusion matrix represent the actual classes, and the columns represent the predicted classes.

Points

- **Point 1:** The KNN algorithm was the most accurate model because it was able to learn from the data and identify patterns that were not obvious to humans.
- **Point 2:** The accuracy of the models decreased as the number of previous launches decreased because the models had less data to learn from.
- **Point 3:** The accuracy of the models could be improved by using more data because more data would allow the models to learn more patterns.
- **Point 4:** The models could be used to help SpaceX make better decisions about which rockets to launch because the models could predict which rockets are more likely to land successfully.

Appendix

Data Sources

- The data for this report was obtained from the SpaceX API, and web scrapping. The API provides data on all of SpaceX's launches, including the date of the launch, the rocket that was launched, and the outcome of the launch.

Model Selection

- Three different machine learning models were used to predict whether a SpaceX rocket would land successfully:
- Logistic regression
- Support vector machines
- K-nearest neighbors
- The models were selected based on their accuracy on a validation set. The KNN model was the most accurate model, with an accuracy of 83.33%.

Limitations

- The accuracy of the models could be improved by using more data. The current dataset only includes data on the first 100 SpaceX launches. More data would allow the models to learn more patterns and improve their accuracy.
- Another limitation of the models is that they are only able to predict whether a rocket will land successfully. The models cannot predict the specific outcome of a launch. For example, the models cannot predict whether a rocket will land upright or on its side.

Thank you!