IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Kyle King
04/16/2024

Outline

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

Executive Summary

- In this capstone project we predict if the SpaceX Falcon 9 first stage will have a successful landing using several machine learning classification algorithms.

- The main steps of this project include:
  - Data collection, data wrangling, and data formatting
  - Exploratory data analysis
  - Interactive data visualization
  - Machine learning prediction

- The graphs shown that certain features of rocket launches have a correlation with the outcome of the launches.

- It can also be concluded that a decision tree might be the best machine learning algorithm to predict if the first stage will land successfully.

Introduction

- This capstone project was to predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars whereas other providers rockets cost upward of 165 million dollars each.
- Therefore if we can determine if a landing was successful then we can determine the cost of the launch.
- When it comes to SpaceX landings most of the unsuccessful landing were planned.
- The main question that we are trying to answer is, for a given set of features about a Falcon 9 rocket launch will the first stage of the rocket land successfully?

Section 1

Methodology

Methodology

- Data collection methodology:
  - SpaceX Rest API
  - Web Scraping from Wikipedia
- Performed data wrangling:
  - Filtering the data
  - Cleaning up missing data
  - Using One Hot Encoding to prepare the data to a binary classification
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models:
  - Building, tuning and evaluation of classification models to ensure the best results

Data Collection

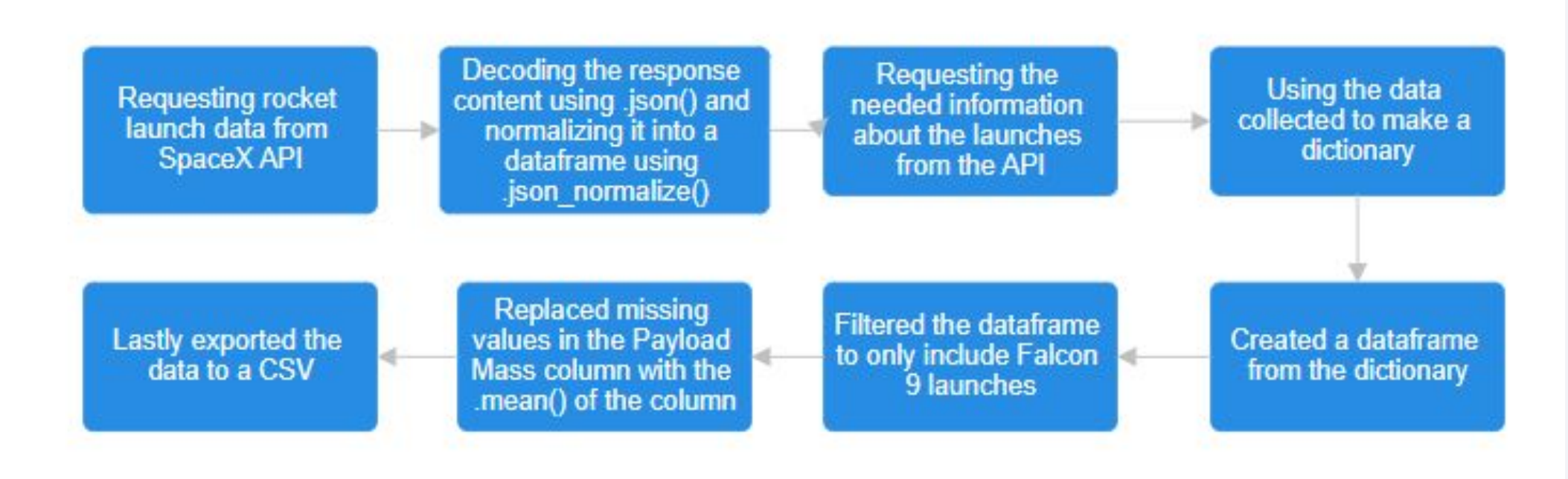
- The data collection process involved a combination of API requests from the SpaceX REST API and Web scraping data from the table on SpaceX Wikipedia.

- I used both of these data collection methods in order to get all the required information about the launches in order to make a more detailed analysis.

- Data obtained using Wikipedia Web Scraping:
  - Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

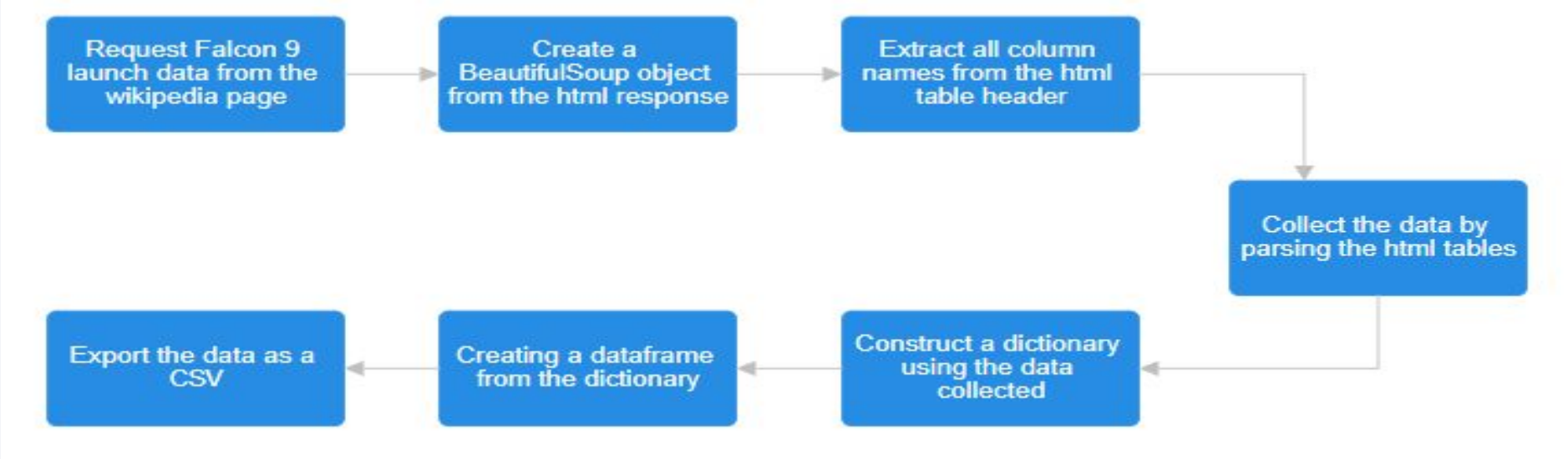
- Data obtained using SpaceX REST API:
  - FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

Data Collection – SpaceX API



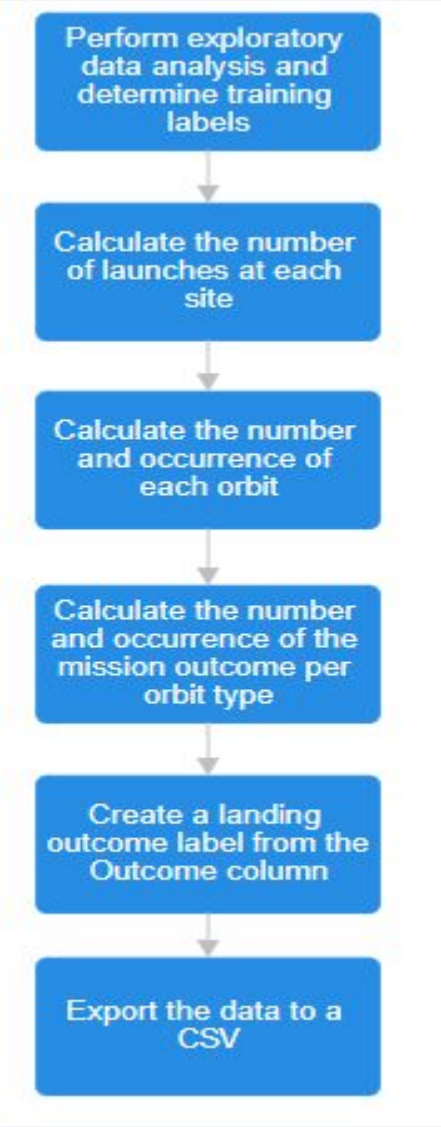
GitHub URL: Data Collection API

Data Collection - Scraping



GitHub URL: Data Collection with Web Scraping

Data Wrangling



In the data set there were several different cases where the booster did not land successfully. True ocean means the mission outcome was successful on landing to a specific region of the ocean while false ocean means the mission outcome was unsuccessful on landing to a specific region of the ocean.

True RTLS means the mission outcome was successful on landing to a ground pad. False RTLS means the mission outcome was unsuccessful on landing to a ground pad.

True ASDS means the mission outcome was successful on landing on a drone ship. False ASDS means the mission outcome was unsuccessful on landing on a drone ship.

We mainly converted those outcomes into training labels with ‘1’ meaning the booster landed successfully and ‘0’ meaning it was unsuccessful.

GitHub URL: Data Wrangling

EDA with Data Visualization

- The charts that were plotted were:
  - 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, and the Yearly Success Rate Trend

- I used scatter plots to show the relation between different variables. If there was a relationship then they could be used in a machine learning model.

- I used bar charts to show comparisons between discrete categories. The goal was to show relations between specific categories and their measured value.

- I used line charts to show trends in the data over a given time period.

GitHub URL: EDA using Data Visualization

EDA with SQL

- The SQL queries performed:
  - Display the names of the unique launch sites
  - Display 5 records where the launch site begins with ‘CCA’
  - Display the total payload mass carried by boosters launch by NASA CRS
  - Display the average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome was achieved
  - List the names of the boosters which had success in landing on the drone ship and has a payload mass greater than 4000 but less than 6000 kg
  - List the total number of successful and failure mission outcomes
  - List the names of the booster versions which have carried the maximum payload mass
  - List the failed landing outcomes on the drone ship, with their booster version and launch site name for the year 2015
  - Rank the count of landing outcomes between the dates 06-04-2010 and 03-20-2017 in descending order

GitHub URL: EDA using SQL

Build an Interactive Map with Folium

- Marker of all the launch sites:
  - I added Marker with Circle, Popup Label, and Text Label of NASA Space Center using its latitude and longitude coordinates as a start location.
  - I added Markers with Circle, Popup Label, and Text Label of all launch sites using their latitude and longitude coordinates to show the location and proximity to the Equator and the coastline.

- Colored Markers of the launch outcome for each launch site:
  - I added colored Markers of success and failure launches, which were green and red respectively, using Marker Cluster to identify which launch sites have a high success rate.

- Distance between a launch site and its proximities:
  - I added colored lines to show distances between the launch sites and its proximities like Highways, Railroads, coastline, and the nearest city.

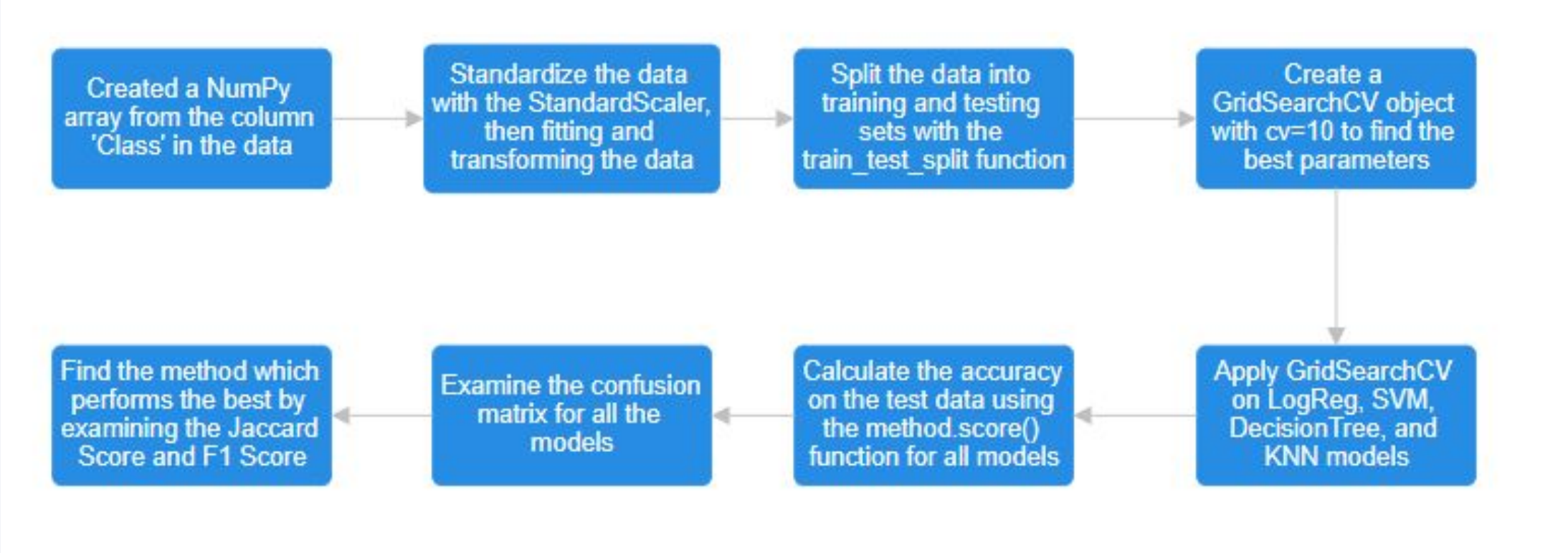
GitHub URL: Interactive Visuals with Folium

Build a Dashboard with Plotly Dash

- Launch Sites Dropdown List:
  - I added a dropdown list to enable launch site selection
- Pie Chart Showing Successful Launches:
  - I added a pie chart to show the total number of successful launches for all sites and the Success vs. Failed counts of a launch site
- Slider of Payload Mass Range:
  - I added a slider to select Payload Range
- Scatter Chart of Payload Mass vs. Success Rate for Different Booster Versions:
  - I added a scatter chart to show the relation between Payload Mass and Launch Success

GitHub URL: SpaceX Dash App

Predictive Analysis (Classification)



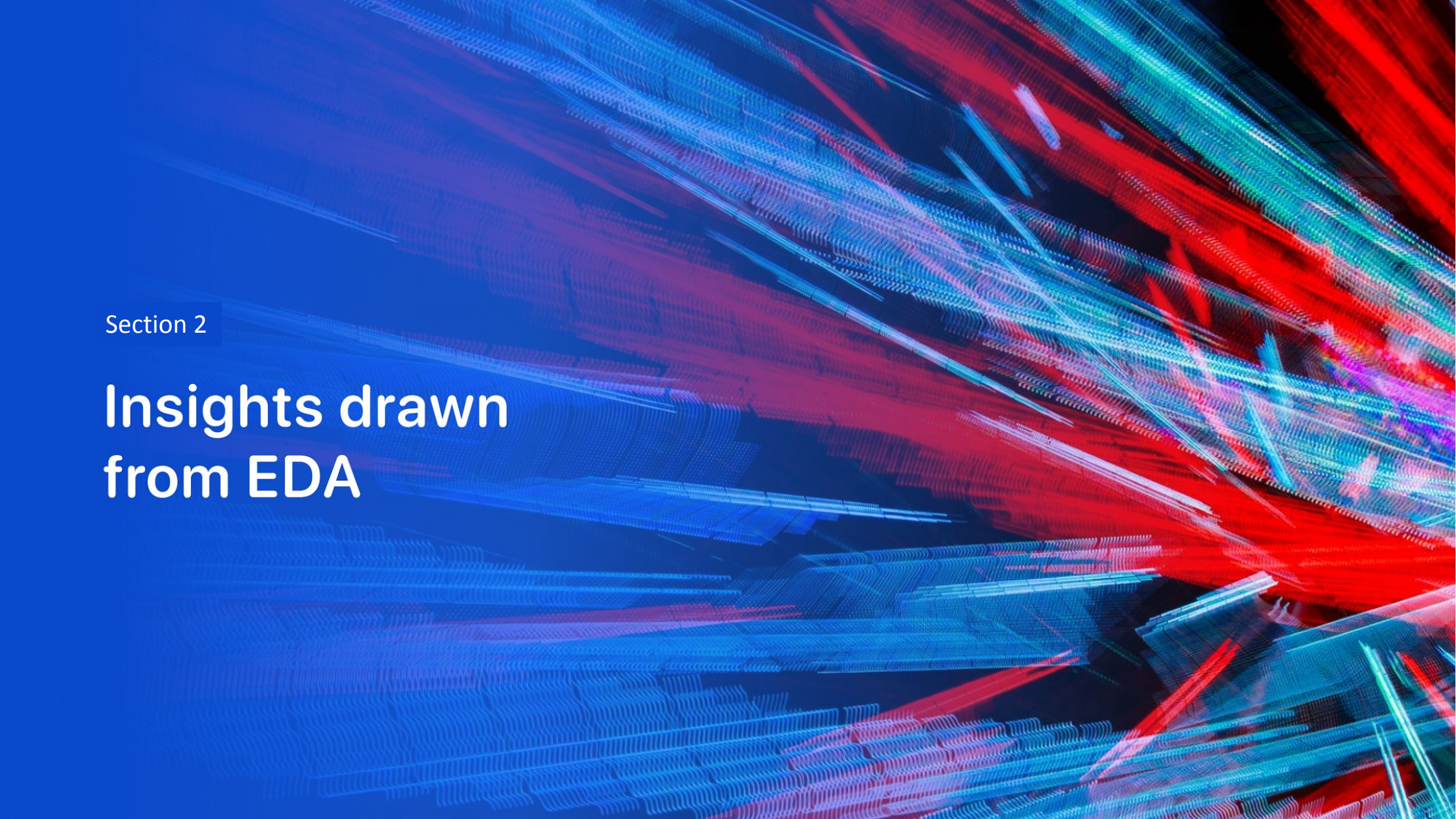
GitHub URL: Machine Learning Prediction

Results

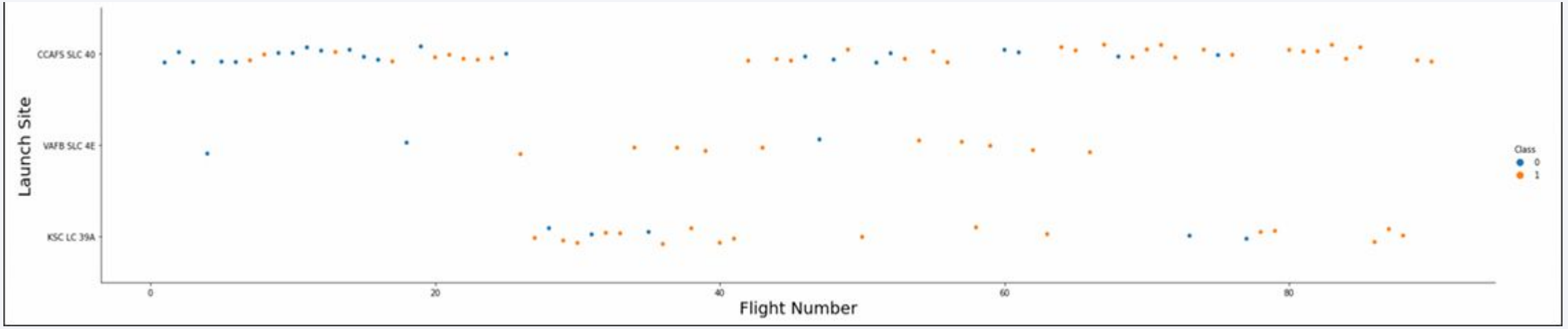
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA



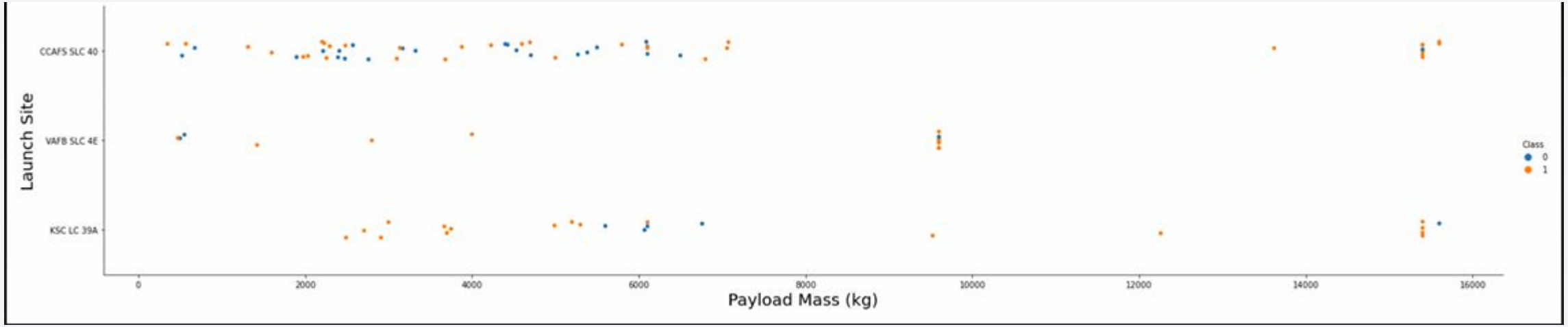
Flight Number vs. Launch Site



Explanation:

- The earliest fights all failed whereas the later flights all succeeded
- The CCAFS SLC 40 launch site has about half of all the launches
- VAFB SLC 4E and KSC LC 39A have a higher success rate then other sites
- One can assume that each of the newer launch sites have a high success rate.

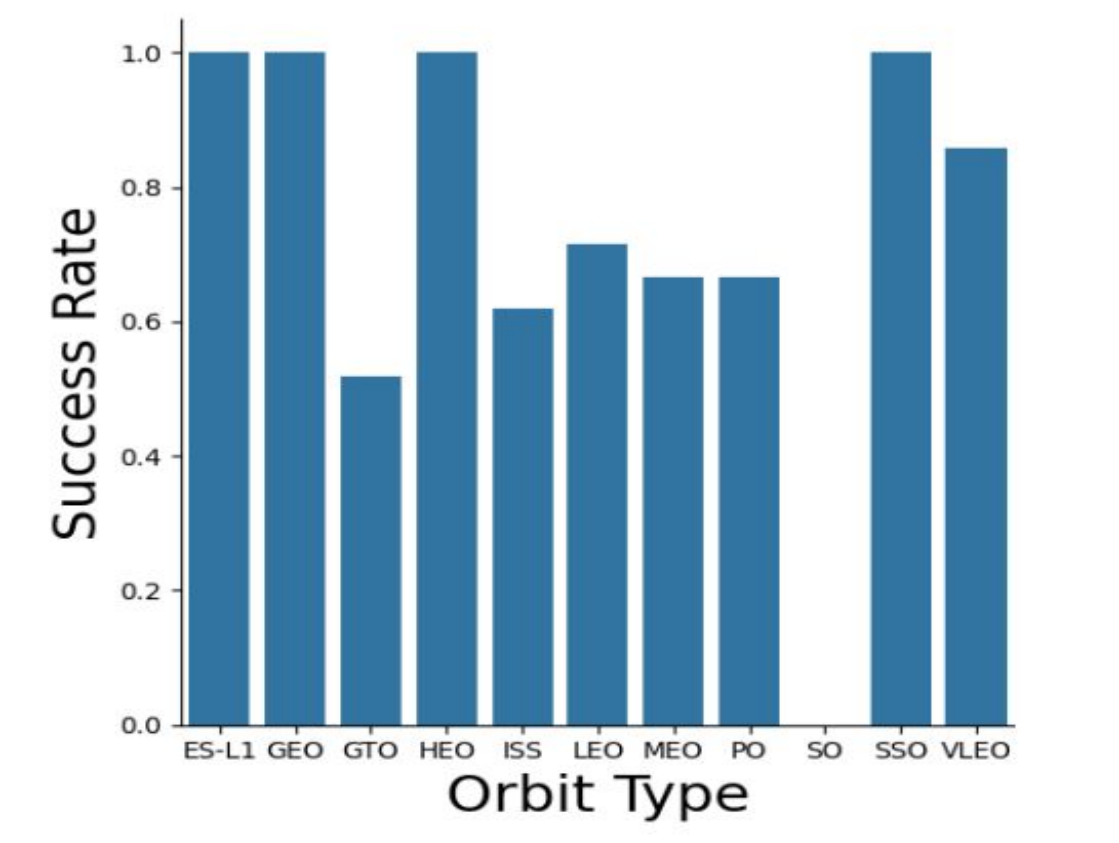
Payload vs. Launch Site



Explanation:

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

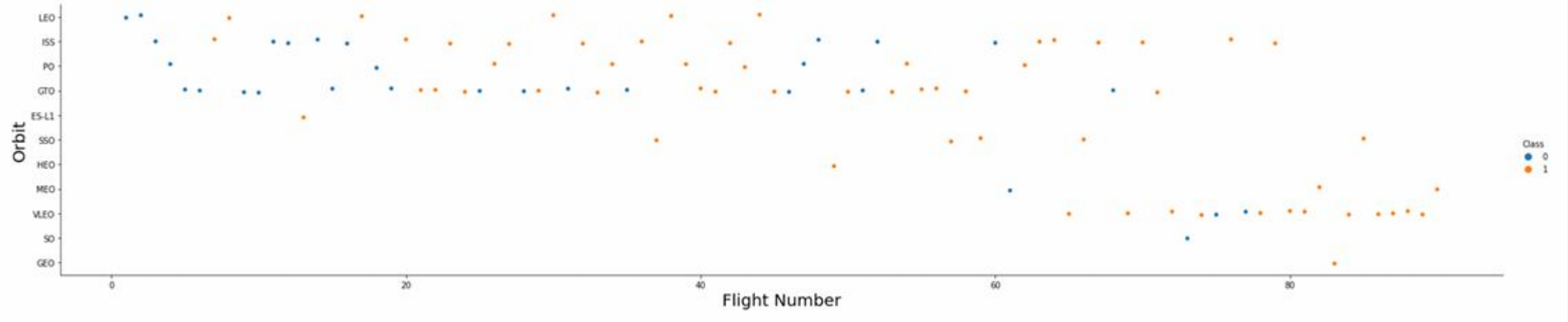
Success Rate vs. Orbit Type



Explanation:

- Orbits that have 100% success rate:
  - ES-L1, GEO, HEO, SSO
- Orbits with success rate between 50% and 80%:
  - GTO, ISS, LEO, MEO, PO
- Orbits that have 0% success rate:
  - SO

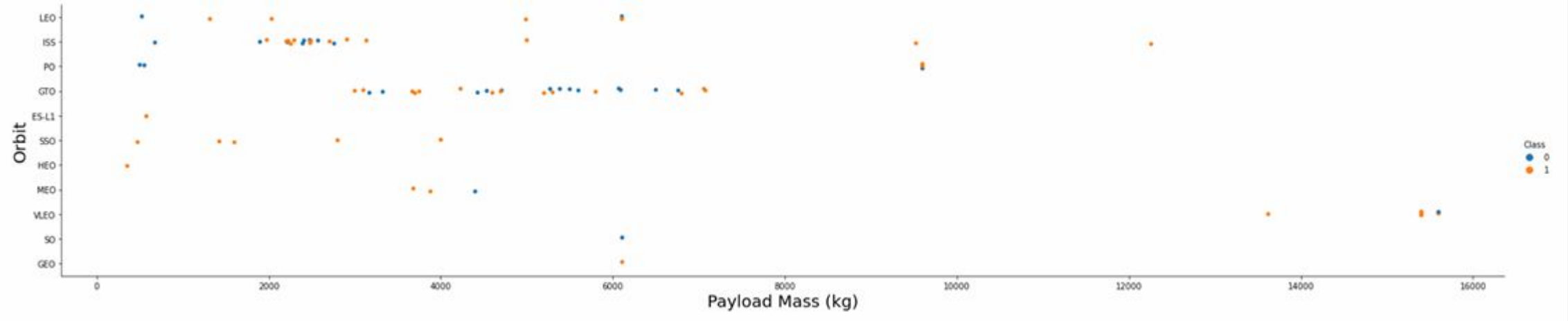
Flight Number vs. Orbit Type



Explanation:

- In the LEO orbit the success rate appears to be related to the number of flights
- There also seems to be no relation between the flight number and GTO orbit

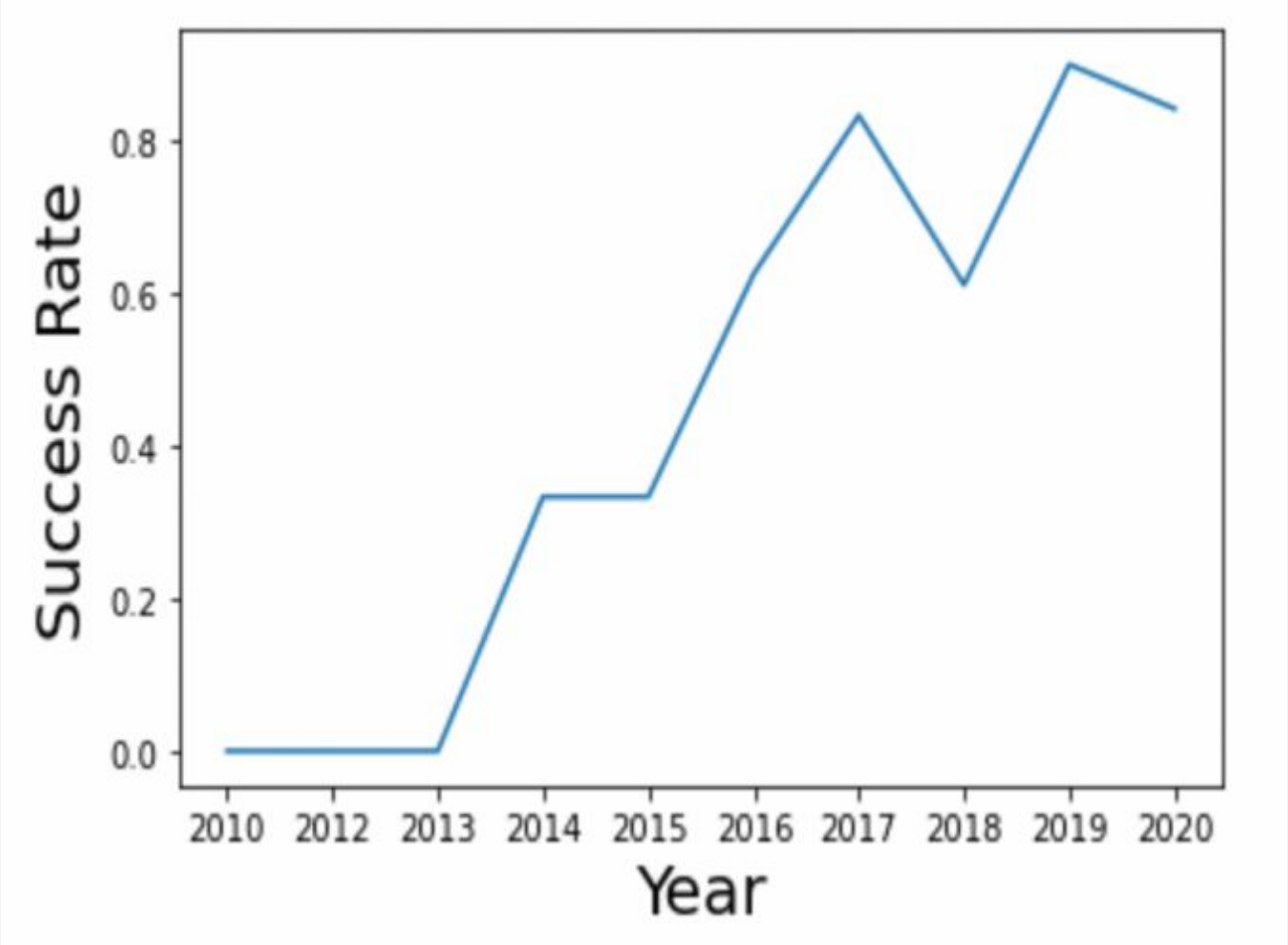
Payload vs. Orbit Type



Explanation:

- It seems that heavy payloads have a negative influence on GTO orbits and a positive influence on GTO and ISS orbits.

Launch Success Yearly Trend



Explanation:

- The success rate was steadily increasing since 2013 till a drop occurred in 2017. Followed by another rise in 2018.

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

Explanation:

- Display the names of each unique launch site

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

Explanation:

Display launch site beginning with ‘CCA’ and limiting it to only 5 results

Total Payload Mass

```
%sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXTABLE where customer = 'NASA (CRS)'
```

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

total_payload_mass
45596

Explanation:

Display the total payload mass carried by the booster launched by NASA CRS

Average Payload Mass by F9 v1.1

```
%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXTABLE where booster_version like '%F9 v1.1'
```

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

average_payload_mass
2928.4

Explanation:

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

First Successful Ground Landing Date

```
%sql select min(date) as first_successful_landing from SPACEXTABLE where landing_outcome = 'Success (ground pad)'
```

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

first_successful_landing
2015-12-22

Explanation:

Display the date when the first successful landing on the ground pad was achieved

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select booster_version from SPACEXTABLE where landing_outcome = 'Success (drone ship)' and payload_mass__kg_ between 4(
```

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

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

Explanation:

Display the names of the boosters which have has success landing and have a payload mass greater than 4000 but less than 6000 kg

Total Number of Successful and Failure Mission Outcomes

```
%sql select mission_outcome, count(*) as total_number from SPACEXTABLE group by mission_outcome
```

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

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

Explanation:

Display the total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

```
%sql select 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

Explanation:

Display the names of the booster versions which have carried the maximum payload mass.

2015 Launch Records

```
%sql select date, booster_version, launch_site from SPACEXTABLE where landing_outcome = 'Failure (drone ship)' and substr(Da
```

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

Date	Booster_Version	Launch_Site
2015-01-10	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	F9 v1.1 B1015	CCAFS LC-40

Explanation:

Display the failed landing outcomes along with their booster version and the date for the year 2015

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

```
%sql select landing_outcome, count(*) as count_outcomes from SPACEXTABLE where date between '2010-06-04' and '2017-03-20' gr
```

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

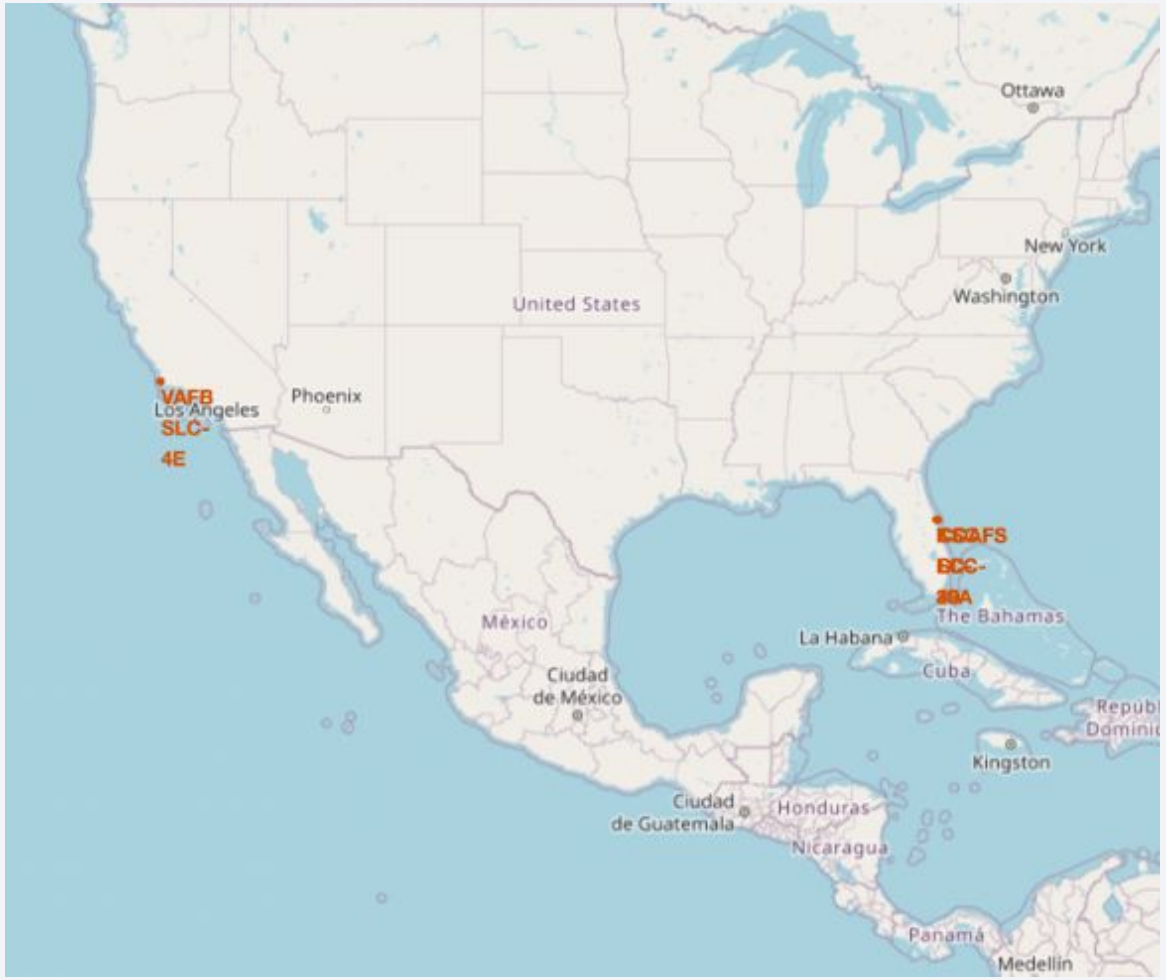
Landing_Outcome	count_outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Explanation:

Display the ranking of landing outcomes between the dates 06-04-2010 and 03-20-2017 in descending order.


Section 3
Launch Sites
Proximities Analysis

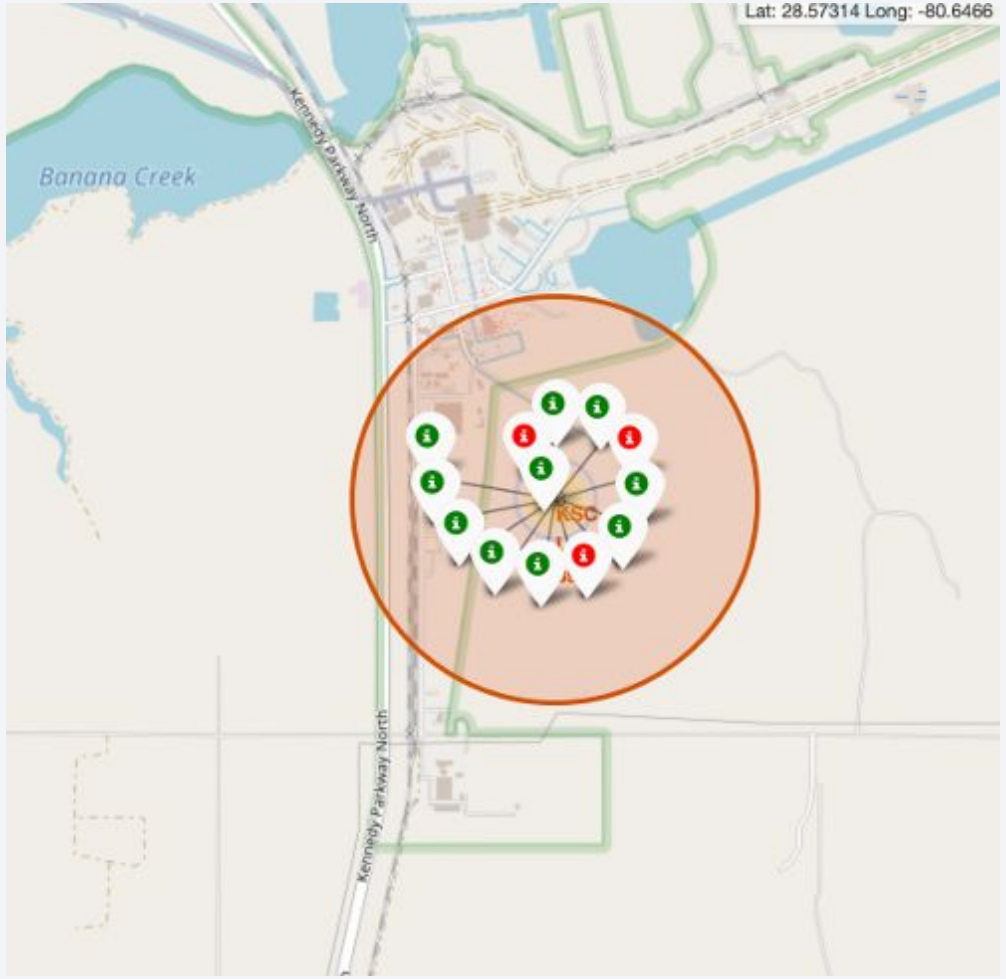
All launch sites location on a global map



Explanation:

- We can see that most of the launch sites are close in proximity to the Equator. Areas closer to the equator are moving faster than any other place at 1670 km per hour.
- Because of inertia when a ship is launched it is also moving around the earth at that speed which helps the spacecraft stay in orbit.
- We also see that launch sites are close to the coast to help minimize debris land in the ocean instead of in populated areas.

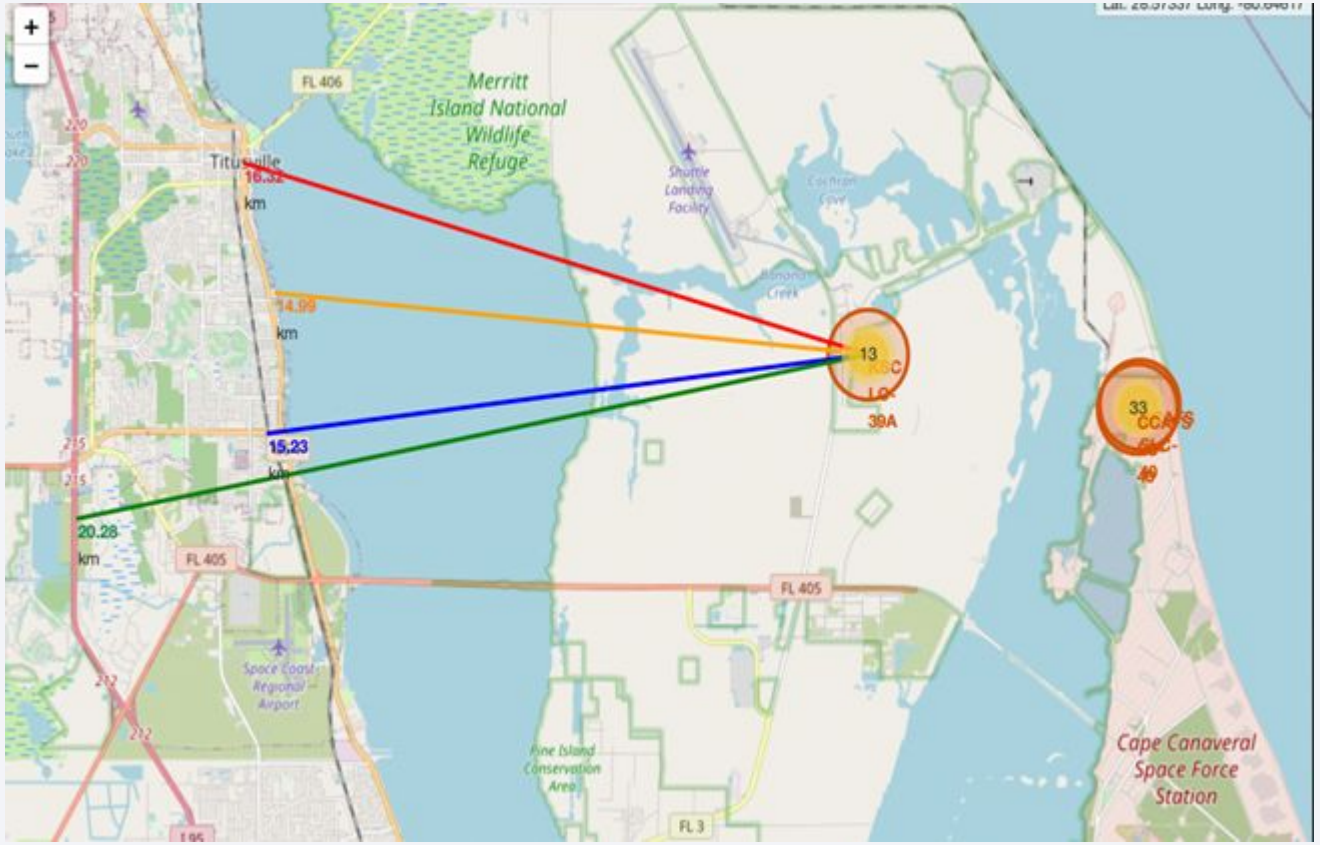
Colored launch records on a map



Explanation:

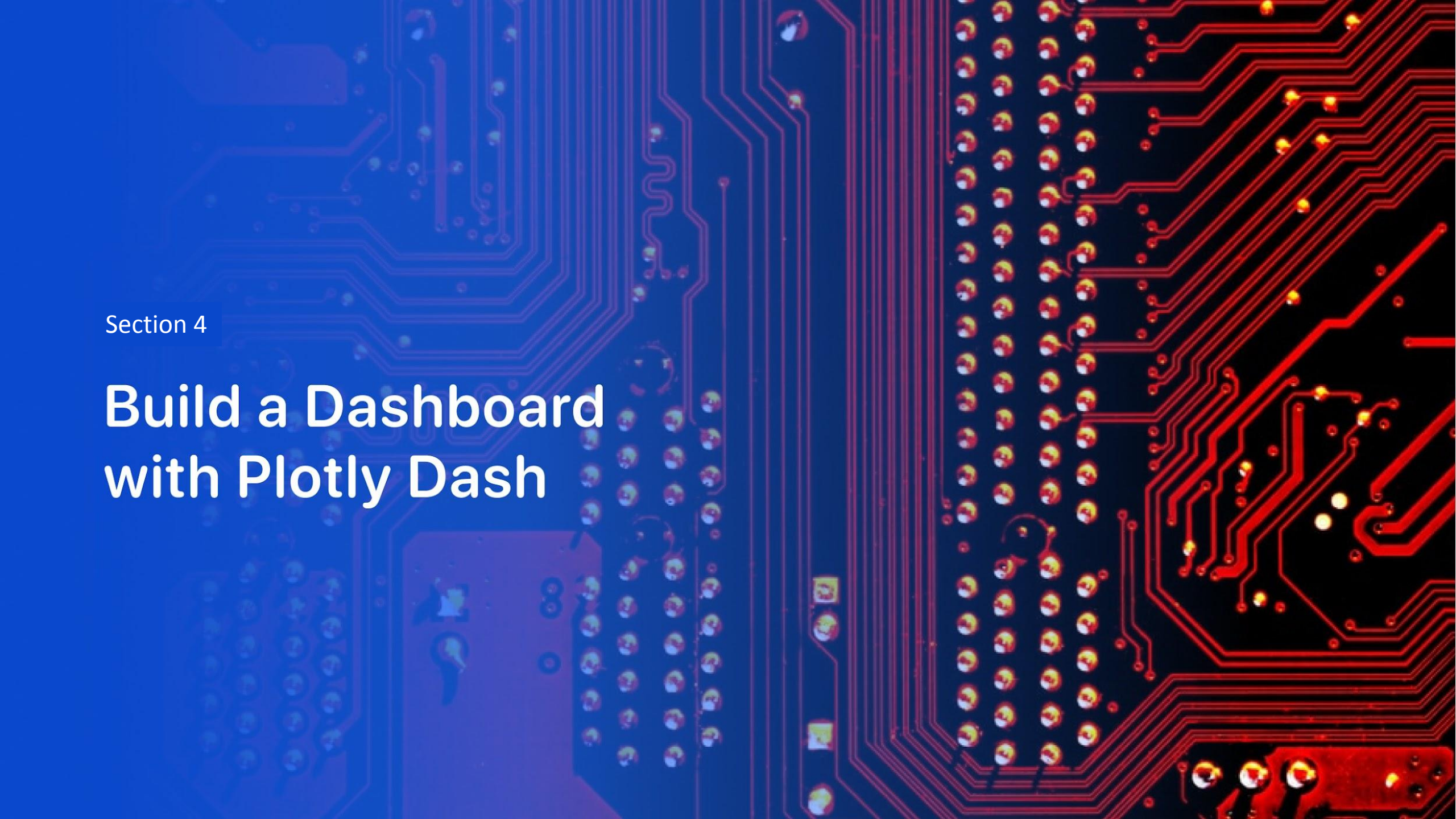
- One can use the colored markers to easily identify which launch sites have a high success rate.
- I used Green to display a successful launch and Red to display a failed launch.
- I also determined that launch site KSC LC-39A has a very high success rate with only have 3 failed out of the 13 total.

Distance from KSC LC-39A to its proximities

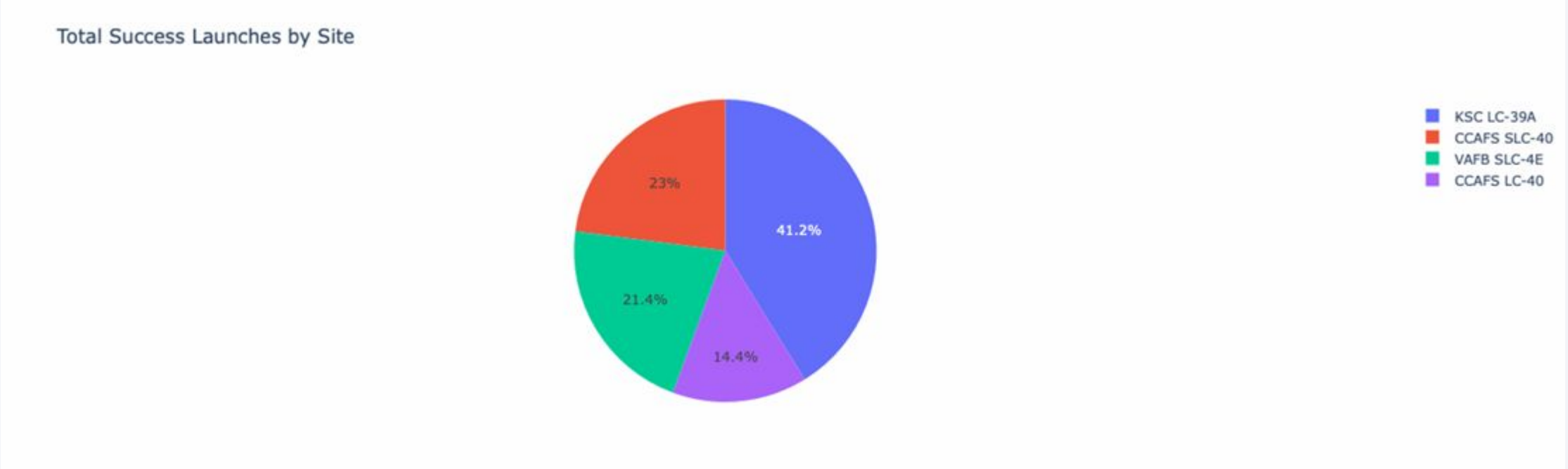


Explanation:

- We can see that launch site KSC LC-39A is relatively close to a railway, highway, and a coastline with distances of 15.23, 20.28, and 14.99 km respectively.
- The nearest city to the launch site is Titusville which is approximately 16.32 km away.
- It is important to be a safe distance from populated areas because the rocket ship is traveling so fast that it can cover that distance in just a mere seconds.


Section 4
Build a Dashboard with Plotly Dash

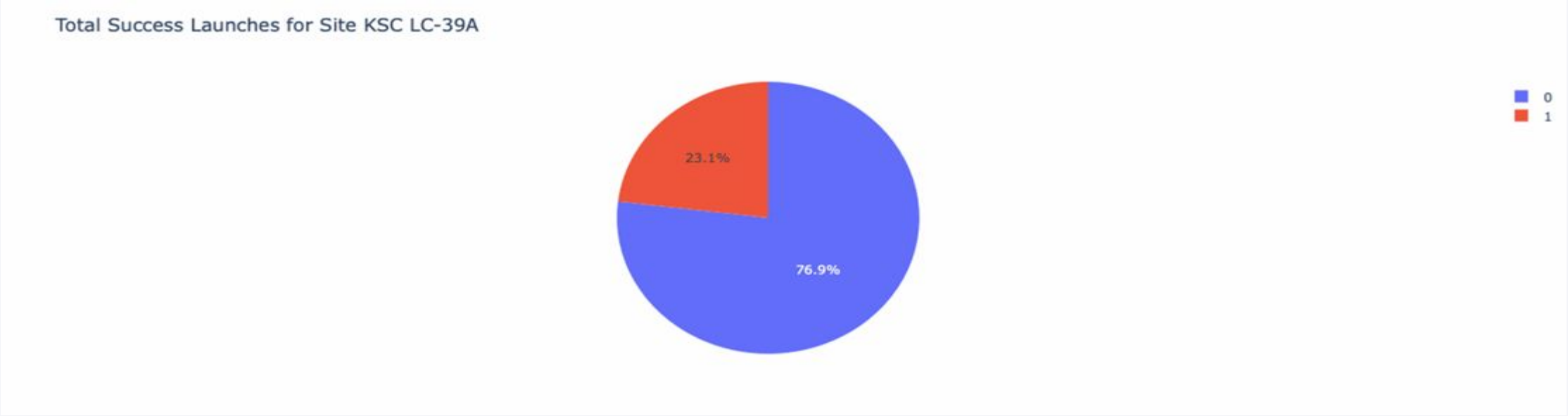
Launch success rate for all sites



Explanation:

- We can see from the pie chart that launch site KSC LC-39A has the highest success rate and launch site CCAFS LC-40 has the lowest success rate.

Launch site with the highest success rate



Explanation:

- From the pie chart we can see that launch site KSC LC-39A has the highest success rate at 76.9% with 10 successful landings out of the 13 total launches there.

Payload mass vs. Launch outcome for all sites

Explanation:

- We can see from the charts that payloads between 2000 and 5500 kg have the highest rate of success.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Scores and Accuracy of the Test Set

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

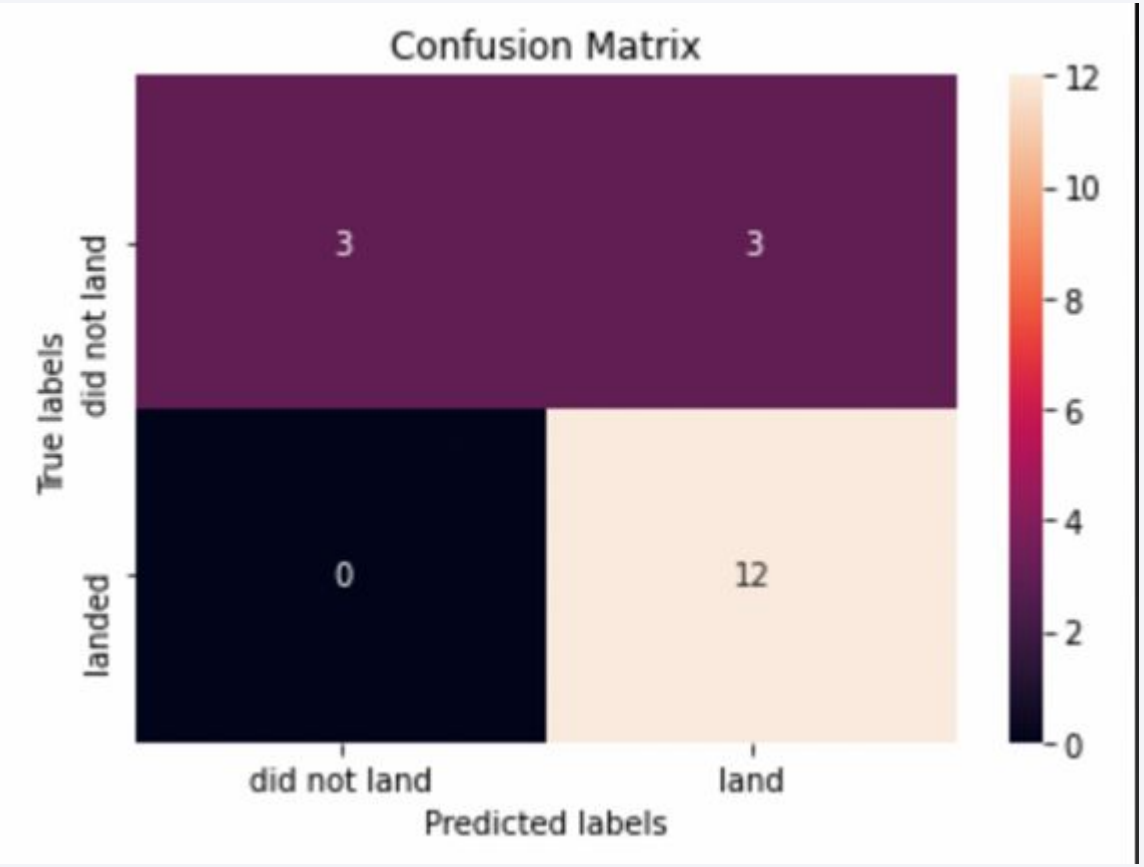
Scores and Accuracy of the Entire Data Set

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

Explanation:

- With the scores of the test set is is difficult to confirm which method performed the best.
- Same test set scores may have been due to the small sample size which lead to testing methods on the whole dataset.
- The scores of the entire dataset confirms that the best model to use is the the Decision Tree. Not only does it have the highest score it also has the highest accuracy.

Confusion Matrix



Explanation:

- From examining the confusion matrix one can see that logistic regression can easily distinguish between the different classes.

- One can also notice that a major problem that occurred with the data set is the false positives, which is the top right of the matrix.

Conclusions

- One can conclude that the Decision tree model was the best model for this dataset.
- Launches that have a lower payload mass show better success rates than launches with larger payload masses.
- Almost all of the launch sites are in close proximity to the Equator and the coastline.
- The success rate have gradually increased throughout the years.
- Orbits GEO, HEO, ES-L1, and SSO have a 100% success rate
- The launch site KSC LC-39A has the highest success rate out of all launch sites.

Appendix

- Special thanks to IBM, Coursera, and the instructors who created this certificate program which gives me the knowledge and confidence to pursue a career as a data scientist.

Thank you!