IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Miguel Angel Trejo
January 2024

Outline

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

Executive Summary

Summary of methodologies.

- Data collection.
- Data wrangling.
- Exploratory Data Analysis with Data Visualization.
- Exploratory Data Analysis with SQL.
- Building an interactive map with Folium.
- Building a Dashboard with Plotly Dash.
- Predictive analysis (Classification).

Summary of all results.

- Exploratory Data Analysis results.
- Interactive analytics demo in screenshots.
- Predictive analysis results.

Introduction

Project background and context.

SpaceX is the most successful company of the commercial space age, making space travel affordable. The company advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.

Questions to be answered.

- How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
- Does the rate of successful landings increase over the years?
- What is the best algorithm that can be used for binary classification in this case?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX REST API.
  - Using web scraping from Wikipedia
- Perform data wrangling
  - Filtering the data.
  - Dealing with missing values.
  - Using One Hot Encoding to prepare the data to a binary classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models.
  - Building, tuning and evaluation of classification models to ensure the best results.

Data Collection

Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry.

We had to use both of these data collection methods in order to get complete information about the launches for a more detailed analysis.

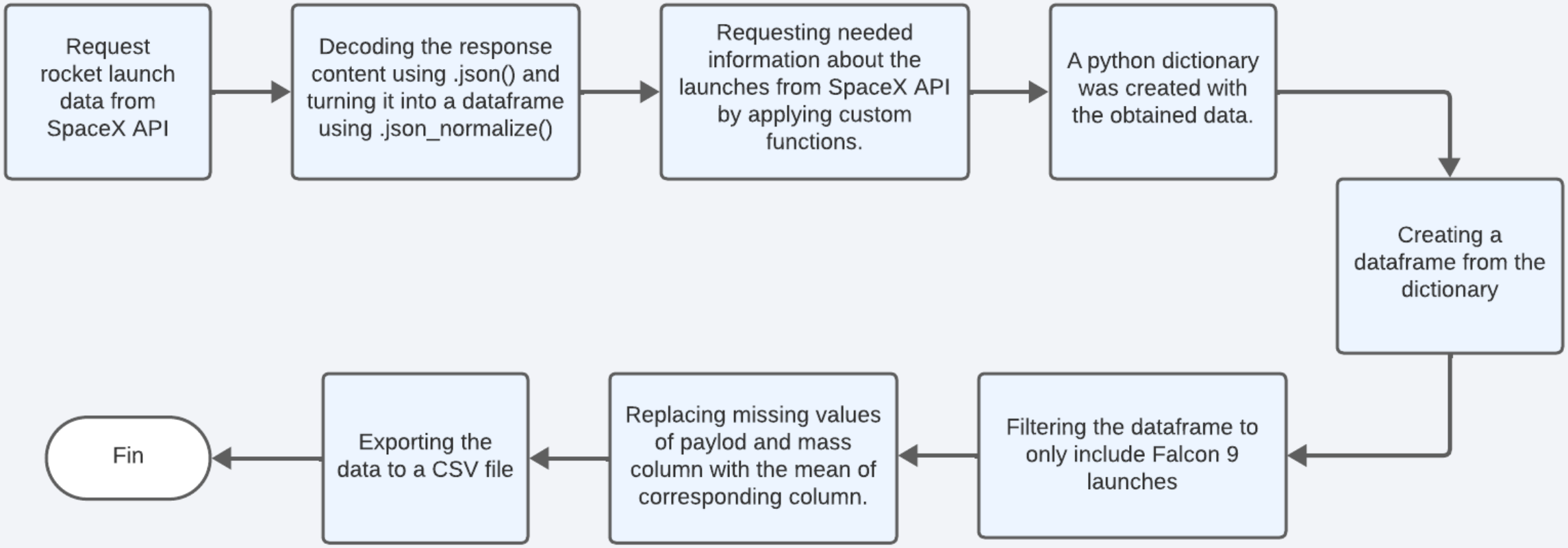
Data Columns are obtained by using SpaceX REST API:

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude.

Data Columns are obtained by using Wikipedia Web Scraping:

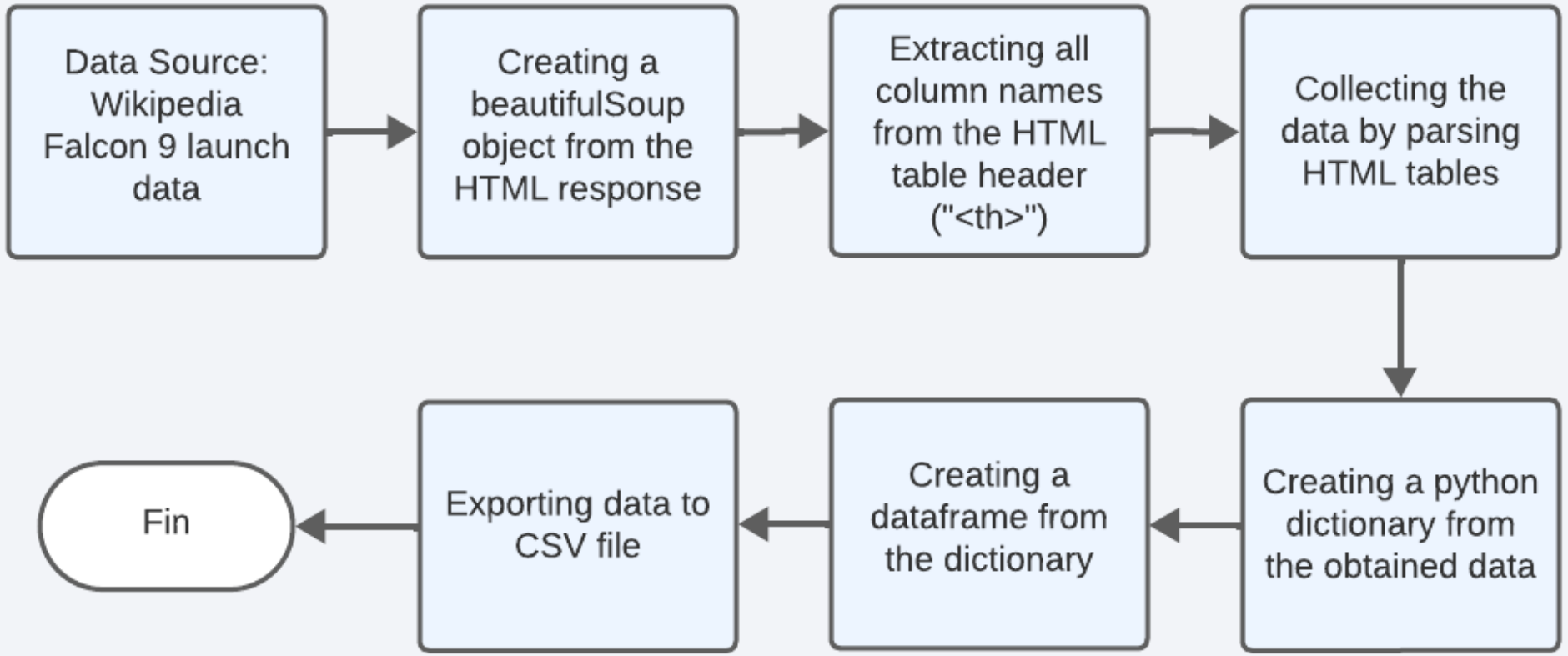
Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time.

Data Collection – SpaceX API



GitHub Link

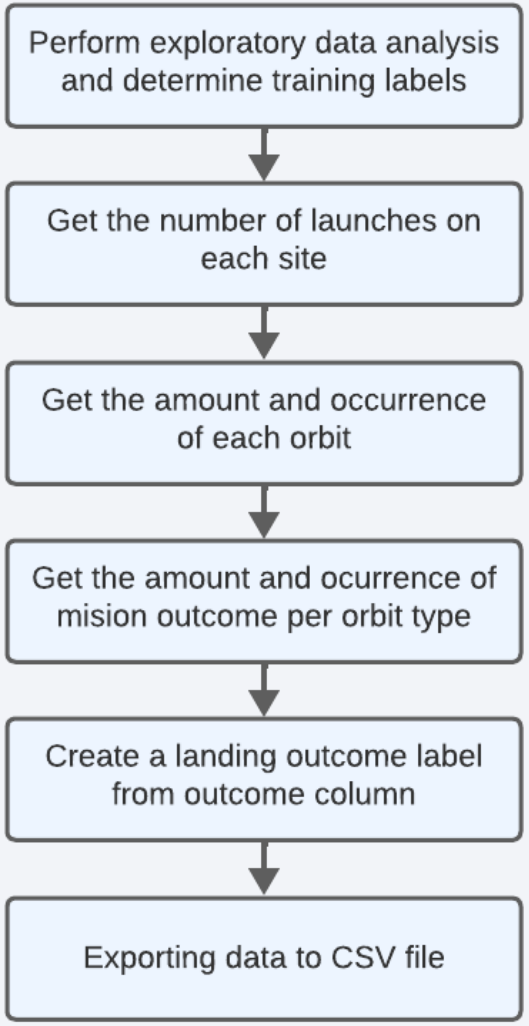
Data Collection - Scraping



GitHub link

Data Wrangling

In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad. True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship.



GitHub link

EDA with Data Visualization

- 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 Success Rate Yearly Trend.

Scatter plots show the relationship between variables. If a relationship exists, they could be used in machine learning model. Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value. Line charts show trends in data over time (time series).

GitHub link.

EDA with SQL

SQL queries performed.

- Display the names of the unique launch sites in the space mission.
- Display 5 records where launch sites begin with the string 'CCA'.
- Display the total payload mass carried by boosters launched by NASA (CRS).
- Display average payload mass carried by booster version F9 v1.1.
- List the date when the first successful landing outcome in ground pad was achieved.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- List the total number of successful and failure mission outcomes.
- List the names of the booster versions which have carried the maximum payload mass. Use a subquery.
- List the records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015.
- 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.

GitHub link

Build an Interactive Map with Folium

- Marker with circle, pop up label and text label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.
- Marker with circle, pop up label and text label of all launch sites using their latitude and longitude coordinates to show their geographical location and proximity to Equator and coasts.

Colored Markers of the launch outcomes for each Launch Site:

- Added colored Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.

Added colored Markers of success (Green) and failed (Red) launches using Marker Cluster to

Identify which launch sites have relatively high success rates.

- Added colored Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City.

GitHub Link.

Build a Dashboard with Plotly Dash

Launch Sites Dropdown List:

- Added a dropdown list to enable Launch Site selection using dcc.Dropdown from plotly library.

Pie Chart showing Success Launches (All Sites/Certain Site):

- Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.

Slider of Payload Mass Range:

- Added a slider to select Payload range using dcc.Slider from plotly library.

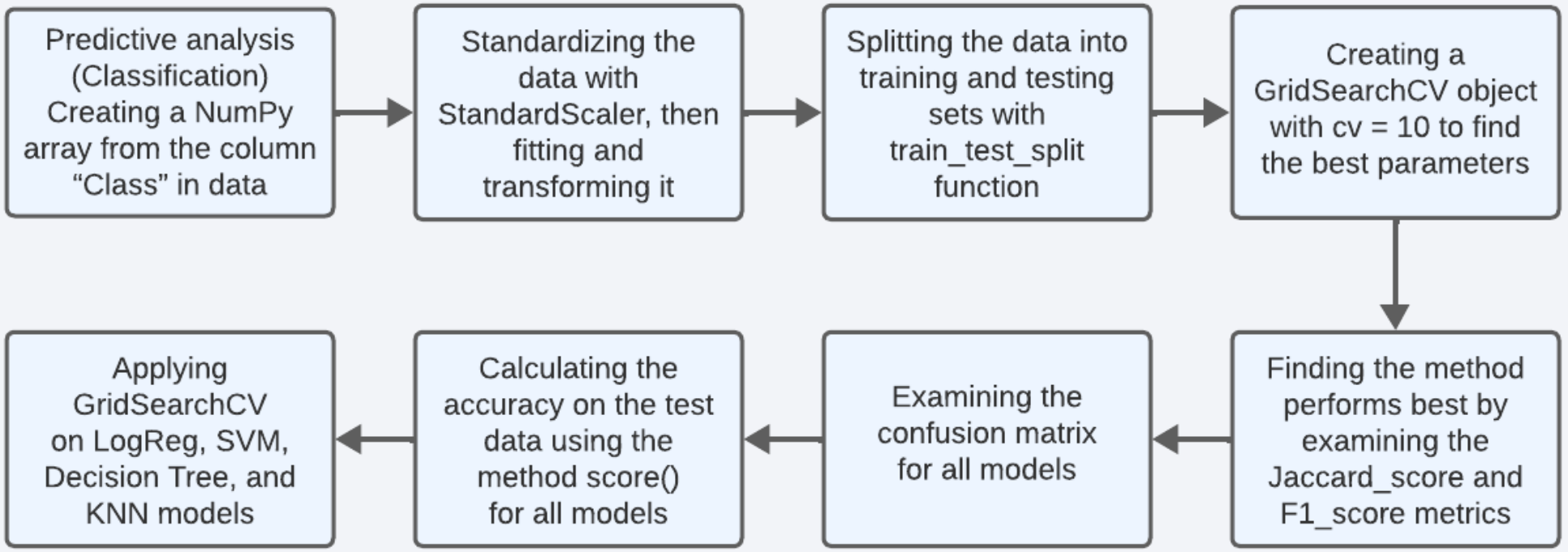
Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:

- Added a scatter chart to show the correlation between Payload and Launch Success.

Graphs was created using plotly.express library.

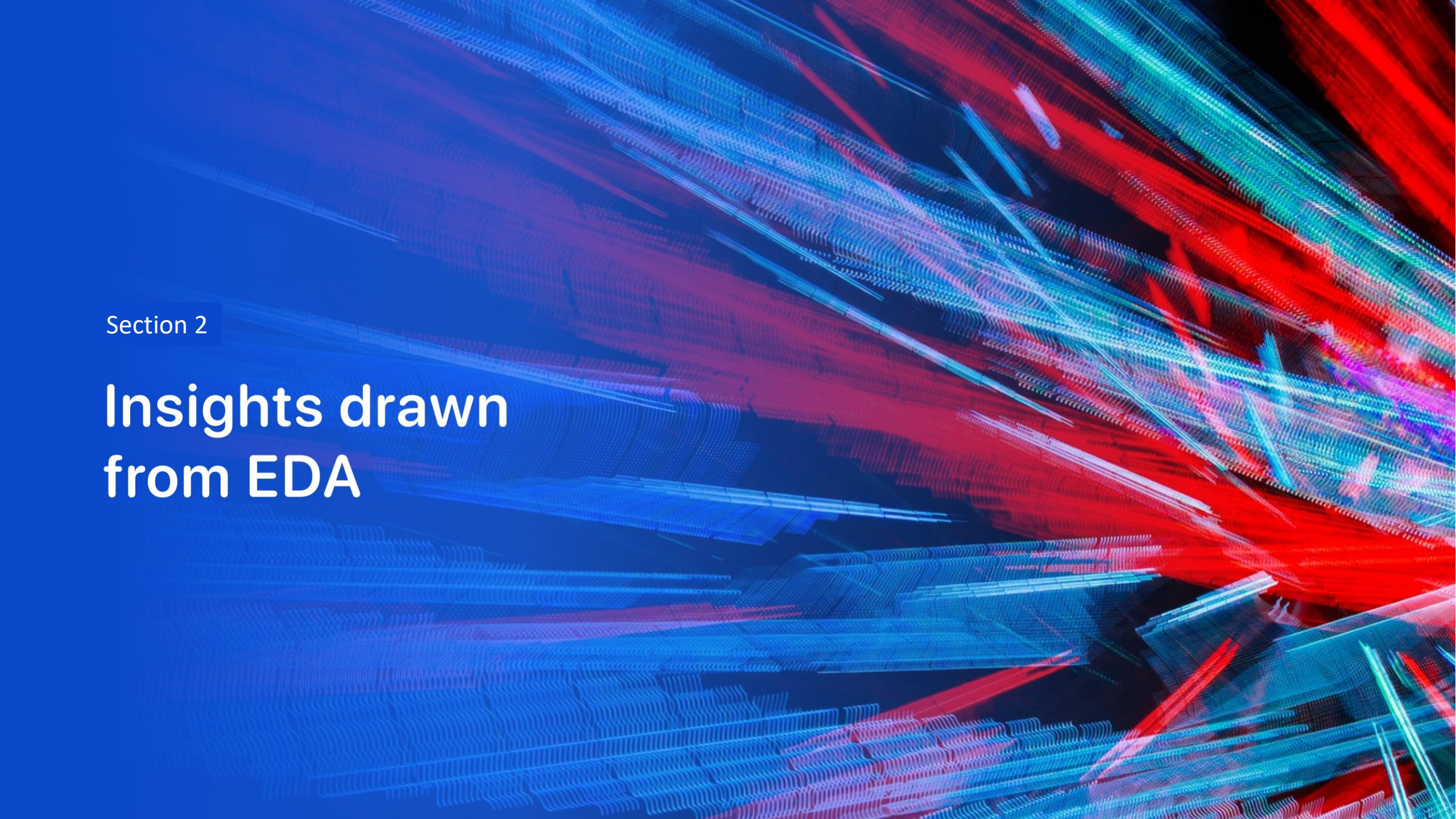
GitHub Link.

Predictive Analysis (Classification)

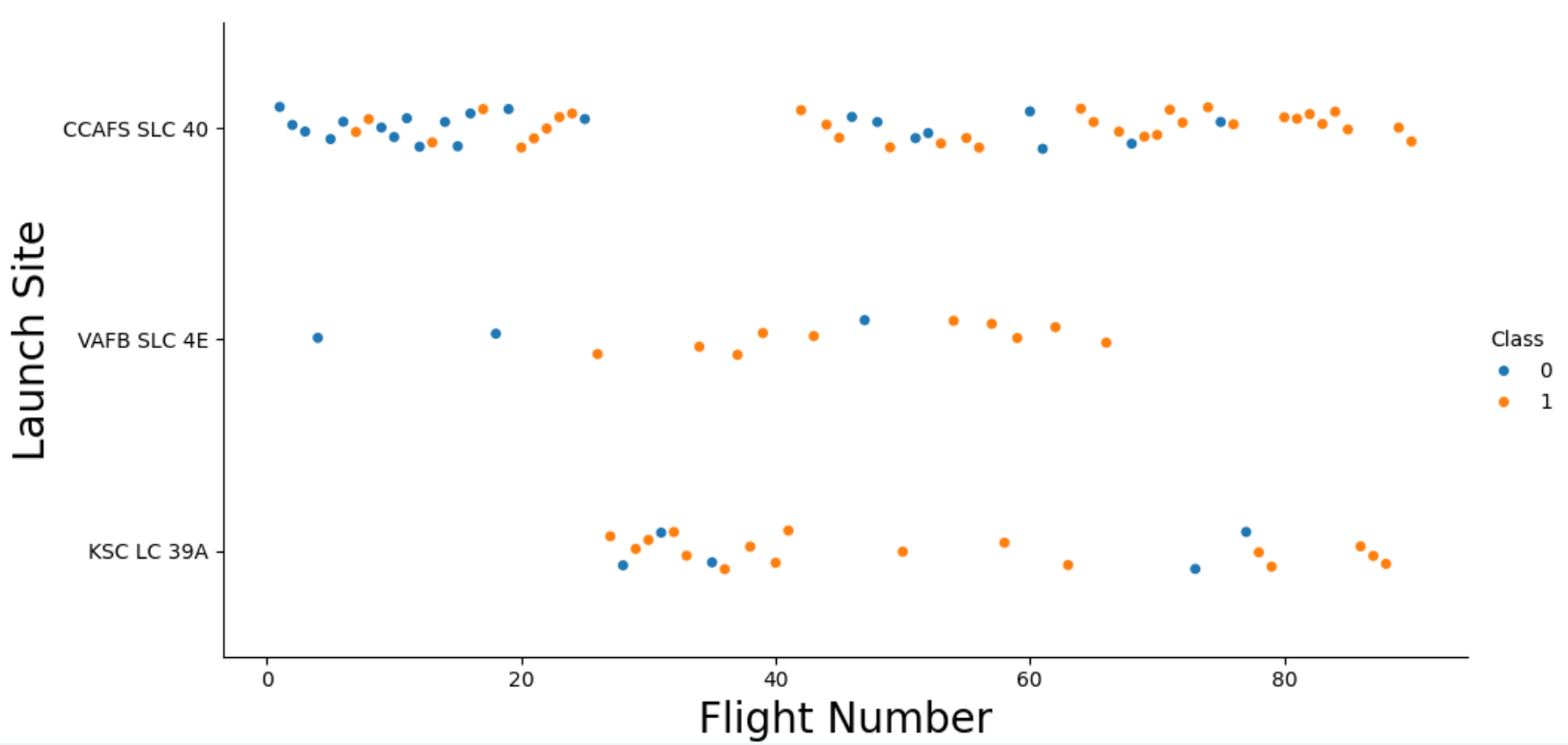


Results

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

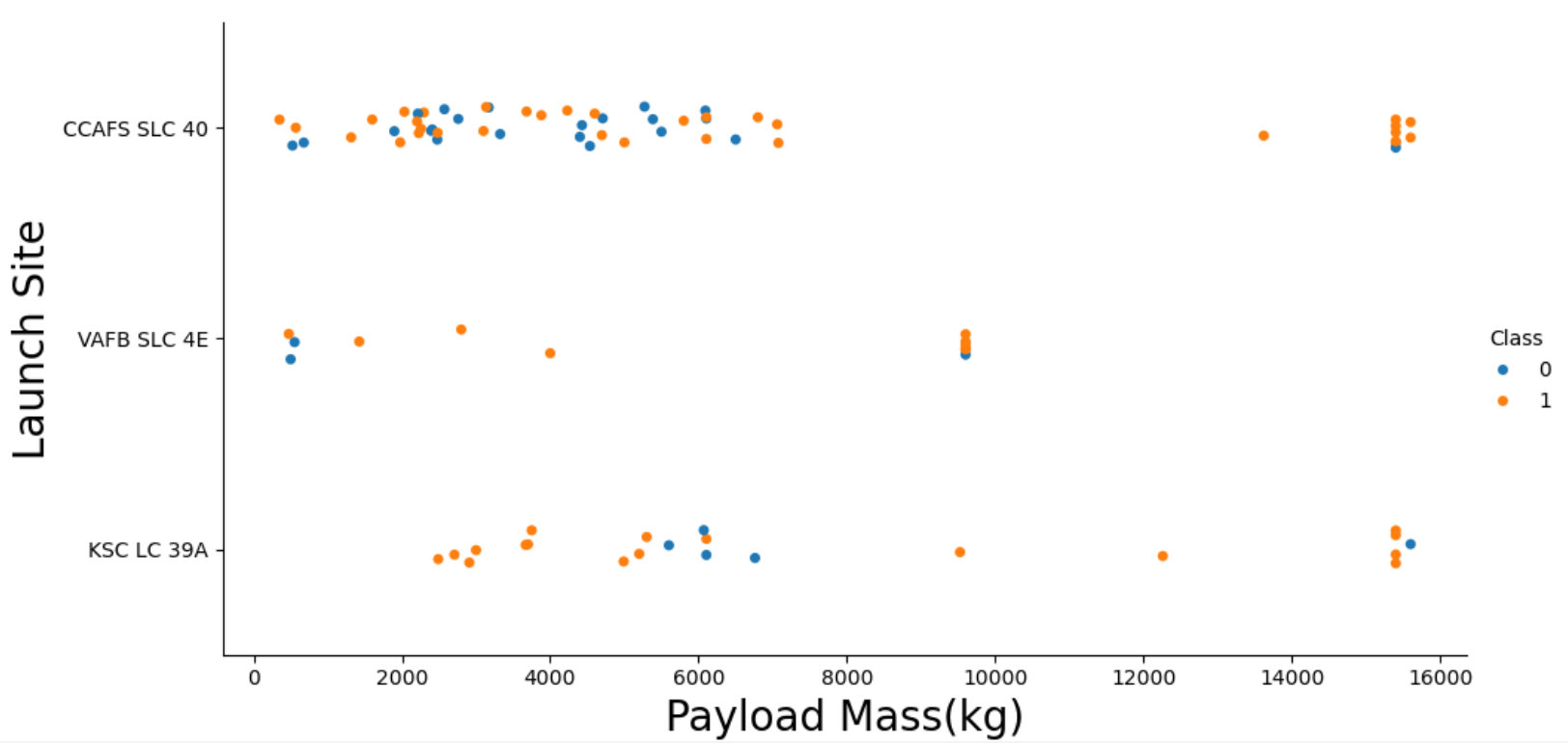

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



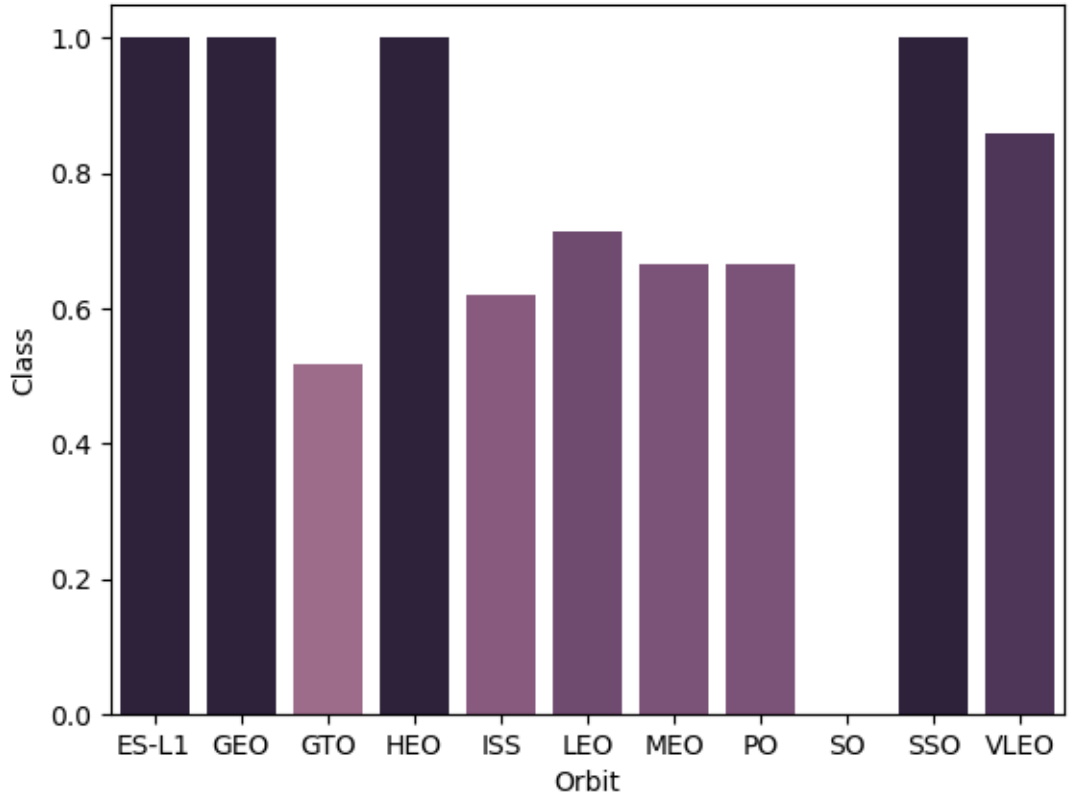
- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be assumed that each new launch has a higher rate of success.

Payload vs. Launch Site



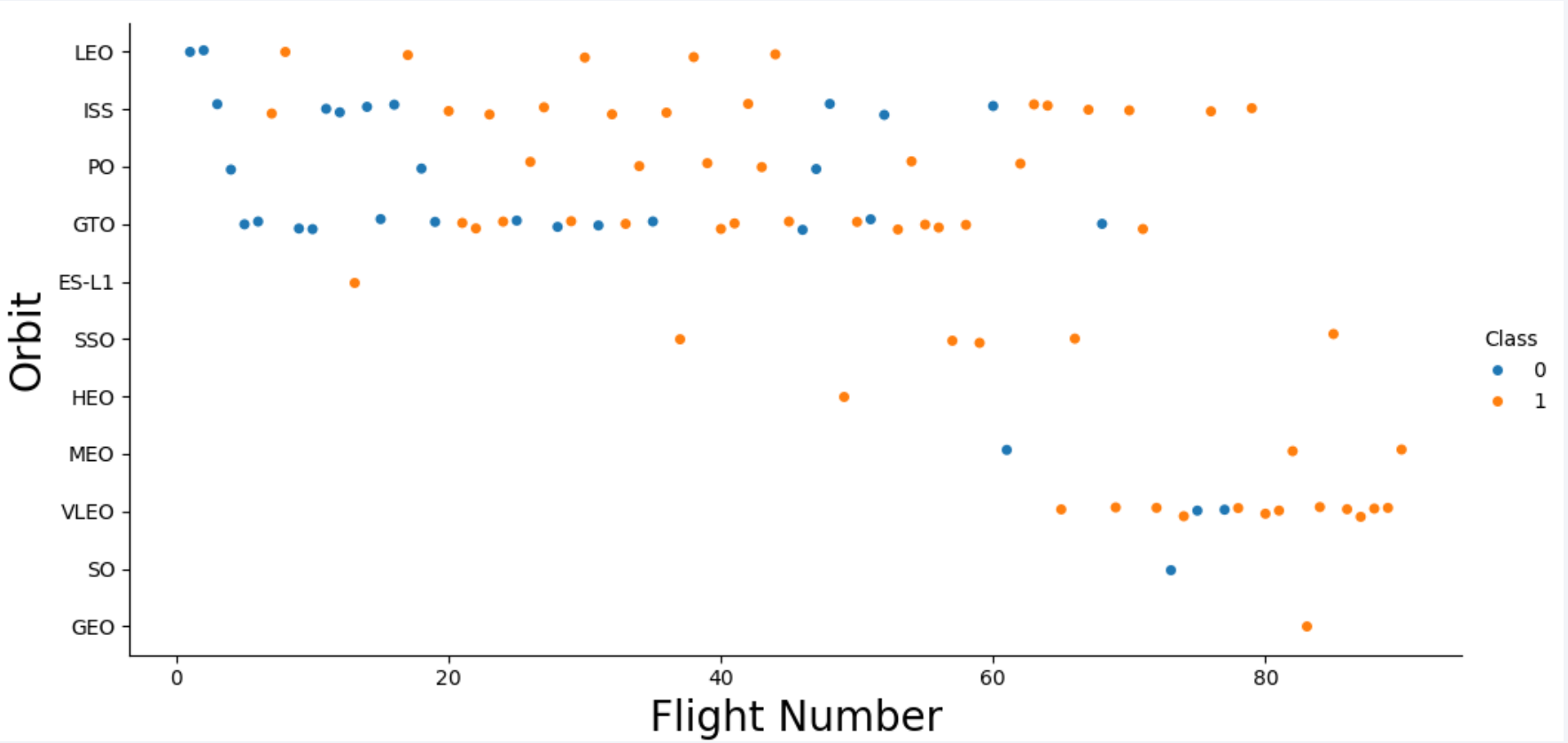
- For every launch site the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

Success Rate vs. Orbit Type



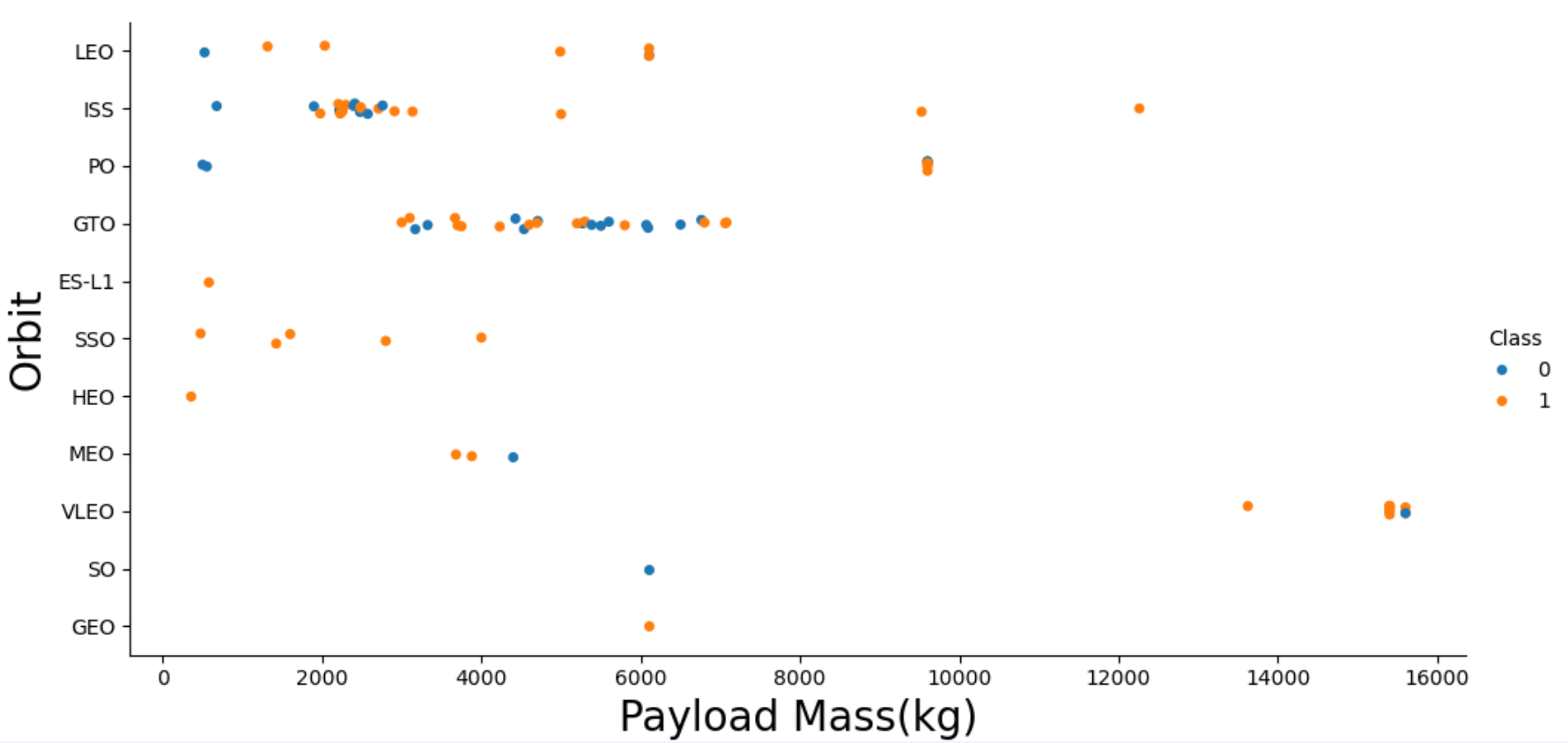
- Orbits with 100% success rate:
  - ES-L1, GEO, HEO, SSO
- Orbits with 0% success rate:
  - SO
- Orbits with success rate: between 50% and 85%:
  - GTO, ISS, LEO, MEO, PO

Flight Number vs. Orbit Type



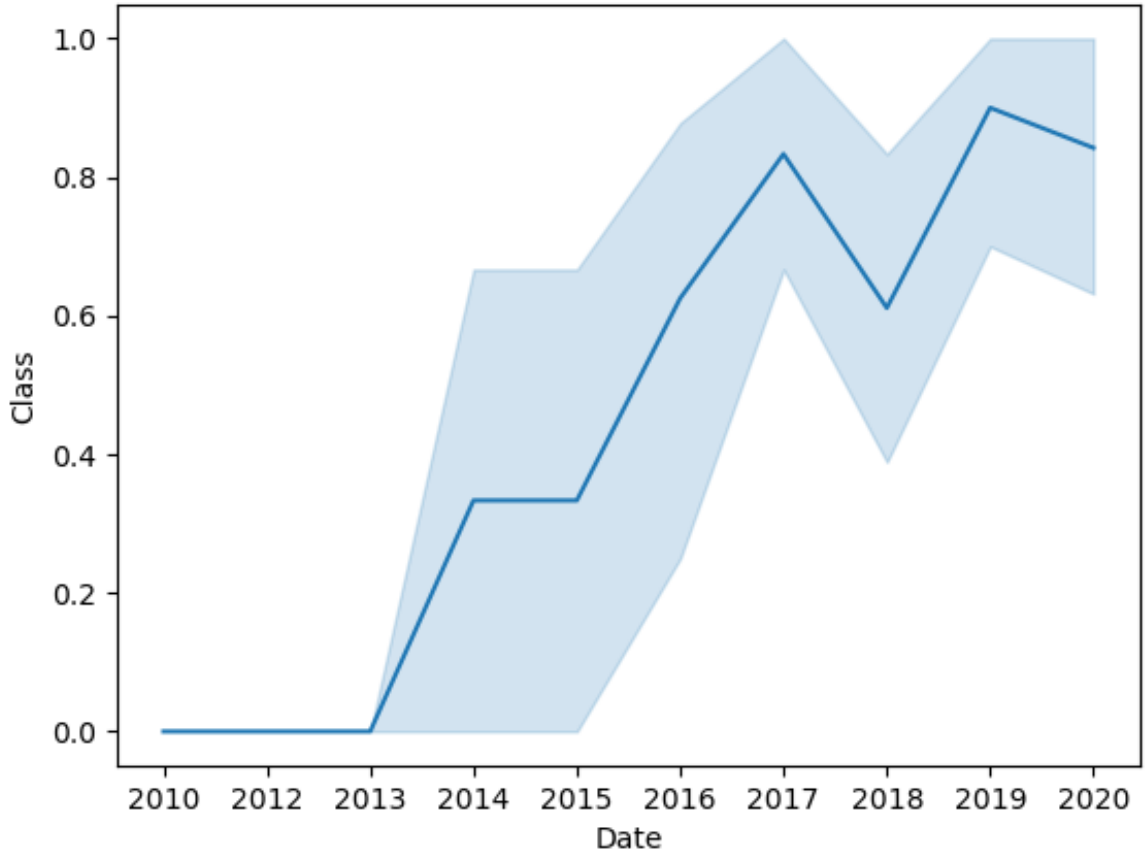
In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



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

Launch Success Yearly Trend



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

```
%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

To obtain the launch sites names we use SELECT DISTINCT statement

Launch Site Names Begin with 'CCA'

Display 5 records where launch sites begin with the string '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

The like operator was used to find the sites names begin with ‘CCA’ pattern.

Total Payload Mass

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

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = "NASA (CRS)"
```

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

sum(PAYLOAD_MASS__KG_)
45596

SUM function was used to obtain the total payload mass for specified customer.

Average Payload Mass by F9 v1.1

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

```
%%sql select round(avg(PAYLOAD_MASS__KG_), 4) as avg_payload
from SPACEXTABLE where Booster_Version like "%F9 v1.1%"
```

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

avg_payload
2534.6667

We filter the results using the like operator and the average function.

First Successful Ground Landing Date

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%%sql select min(Date), Launch_Site, Landing_outcome
from SPACEXTABLE
where Landing_Outcome = "Success (ground pad)"
```

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

min(Date)	Launch_Site	Landing_Outcome
2015-12-22	CCAFS LC-40	Success (ground pad)

Function min was used in the select statement to obtain the first successful landing outcome in ground pad

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 Booster_Version, Landing_outcome, PAYLOAD_MASS__KG_
from SPACEXTABLE
where Landing_Outcome = "Success (drone ship)" and PAYLOAD_MASS__KG_ between 4000 and 6000
```

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

Booster_Version	Landing_Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

Between operator was used in the select statement to obtain the results

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes.

```
%%sql select Mission_Outcome, count(*) as Total
from SPACEXTABLE group by Mission_Outcome
```

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

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

Group by clause and count function was used to obtain the results to the select statement.

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%%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

To obtain the results was necessary to execute a subquery where the max function was used.

2015 Launch Records

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%%sql select substr(Date, 6, 2) as Month, substr(Date, 0, 5) as Year, Booster_Version, Launch_Site
from SPACEXTABLE where Landing_Outcome = "Failure (drone ship)" and substr(Date, 0, 5) = '2015'
```

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

Month	Year	Booster_Version	Launch_Site
01	2015	F9 v1.1 B1012	CCAFS LC-40
04	2015	F9 v1.1 B1015	CCAFS LC-40

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 Date, Landing_Outcome, count(Landing_Outcome) as Total
from SPACEXTABLE
where Date between '2010-06-04' and '2017-03-20'
group by Landing_Outcome order by count(Landing_Outcome) desc
```

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

Date	Landing_Outcome	Total
2012-05-22	No attempt	10
2016-04-08	Success (drone ship)	5
2015-01-10	Failure (drone ship)	5
2015-12-22	Success (ground pad)	3
2014-04-18	Controlled (ocean)	3
2013-09-29	Uncontrolled (ocean)	2
2010-06-04	Failure (parachute)	2
2015-06-28	Precluded (drone ship)	1

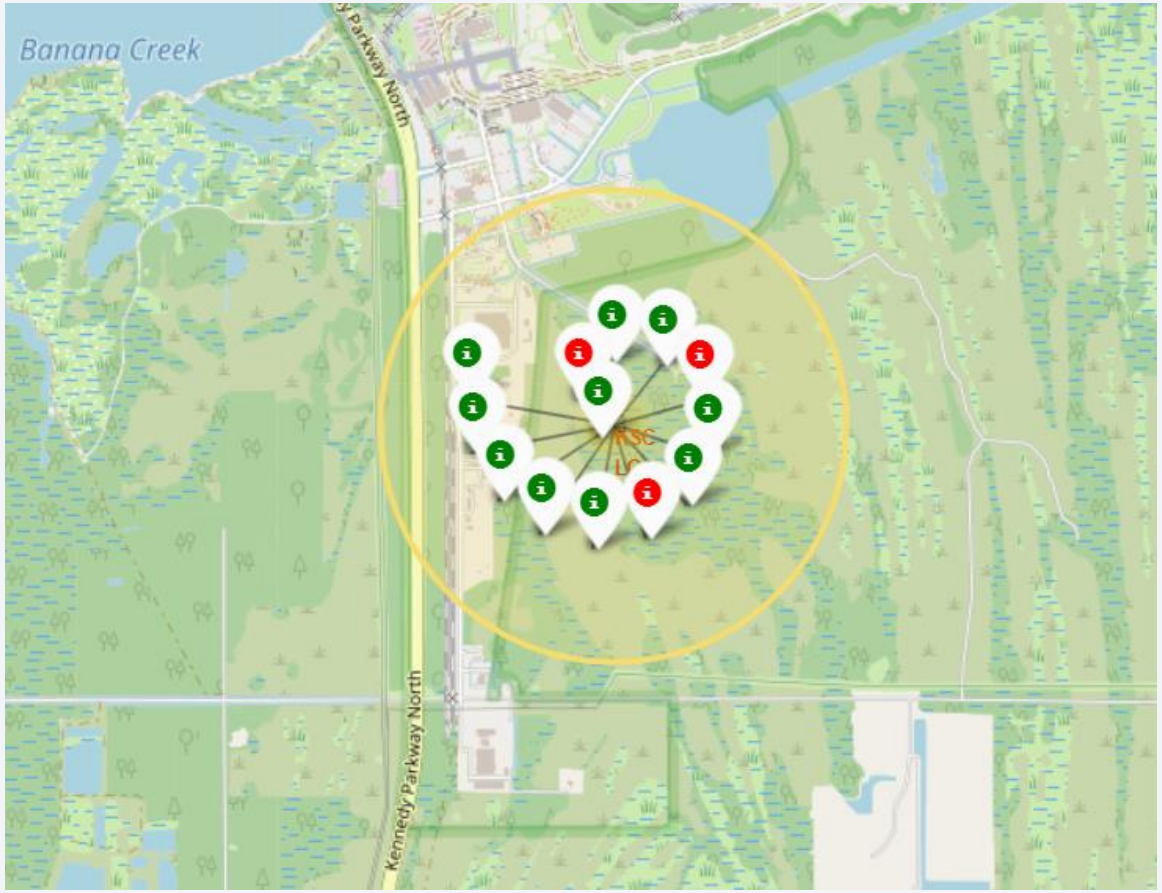
Section 3

Launch Sites Proximities Analysis

Markers of launch sites

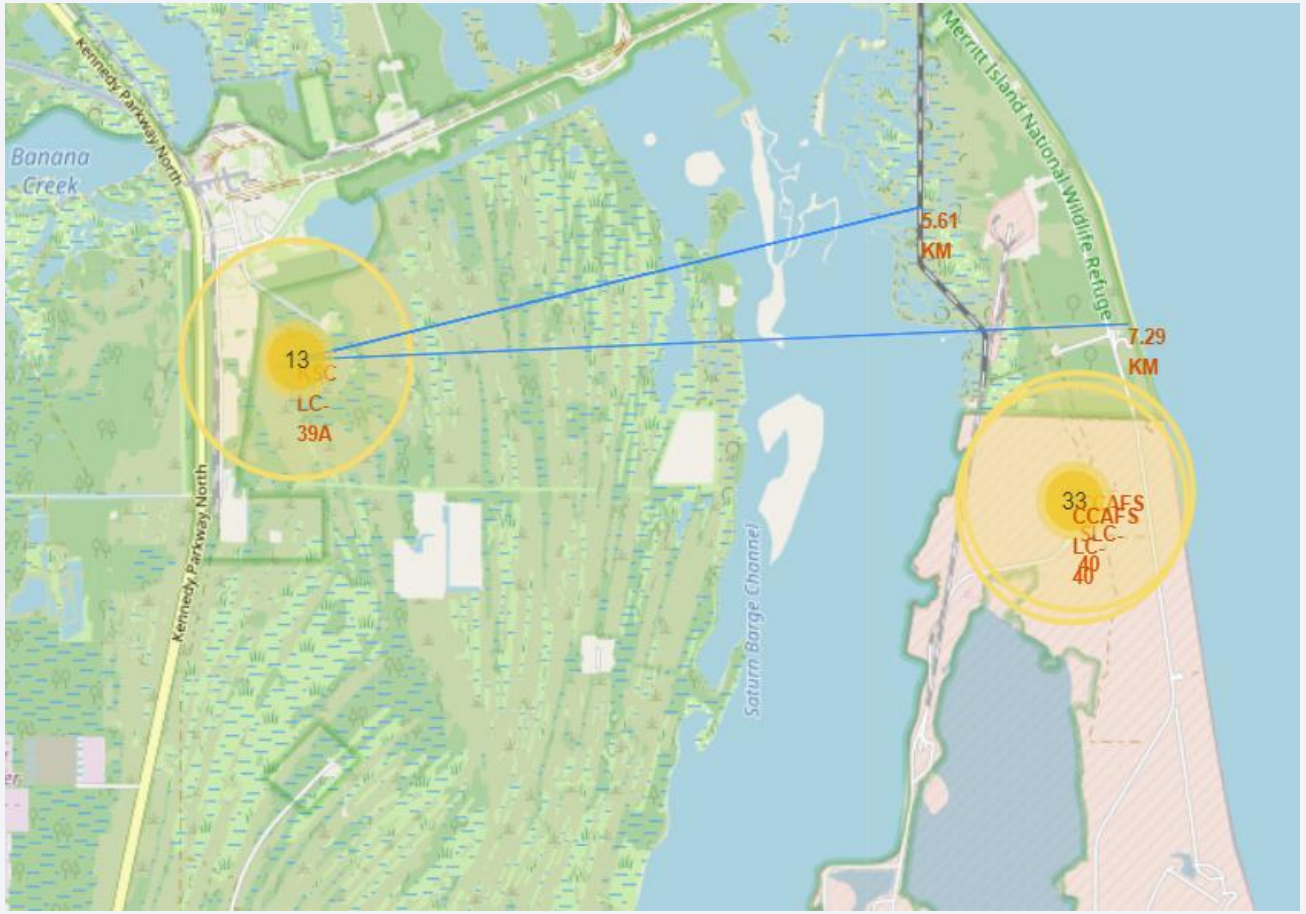
- Most of Launch sites are in proximity to the Equator line. If a spacecraft is launched from a site near Earth's equator, it can take optimum advantage of the Earth's substantial rotational speed. Sitting on the launch pad near the equator, it is already moving at a speed of over 1650 km per hour relative to Earth's center. This can be applied to the speed required to orbit the Earth (approximately 28,000 km per hour).
- All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimizes the risk of having any debris dropping or exploding near people.

Color-labeled launch records on the map

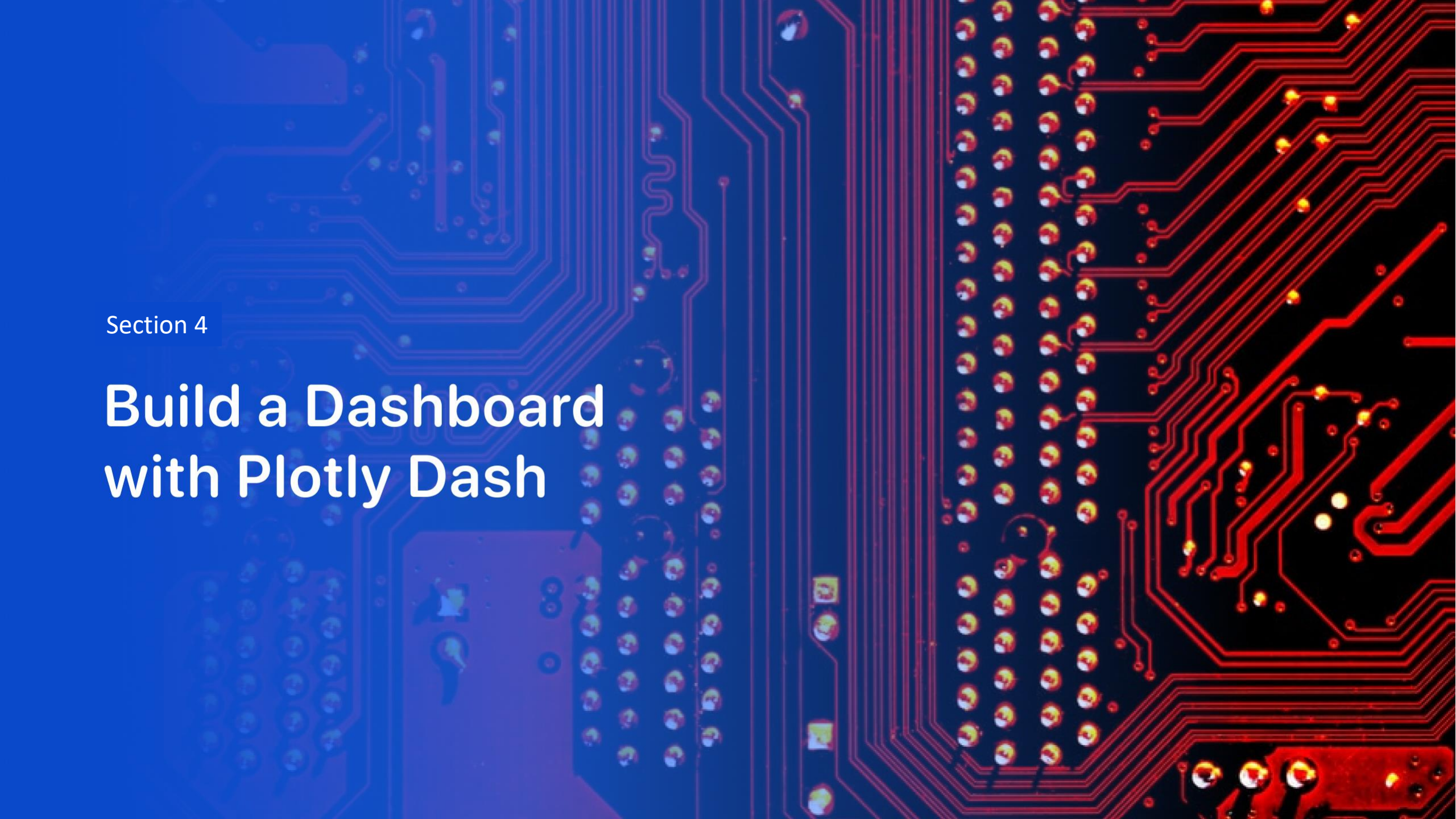


- From the color-labeled markers we should be able to easily. identify which launch sites have relatively high success rates.
  - Green Marker = Successful Launch
  - Red Marker = Failed Launch
- Launch Site KSC LC-39A has a very high Success Rate.

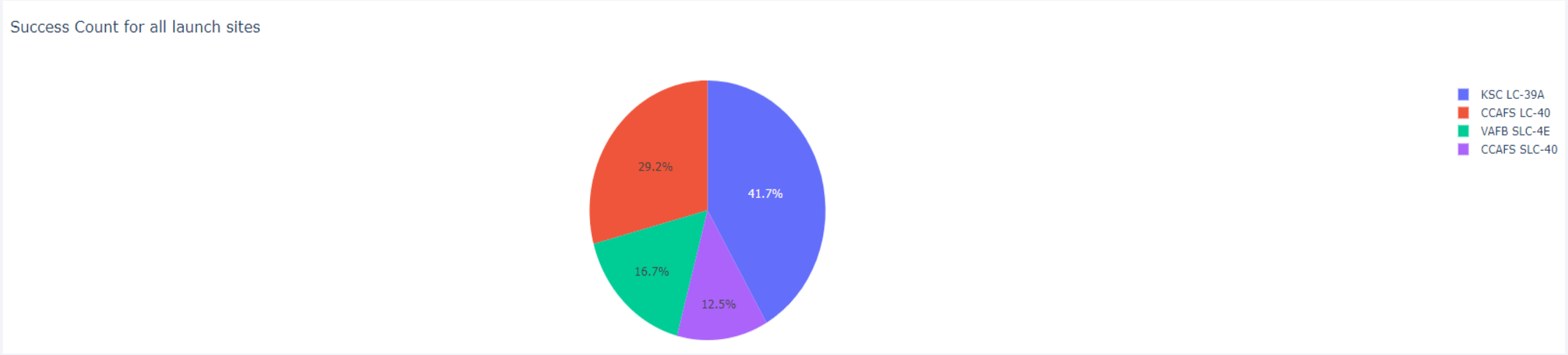
Distance from the launch site KSC LC-39A to its proximities



- Launch sites are in close proximity to railways, highways and coast line.
- Launch site KSC LC-39A is relative close to its closest city Titusville (16.32 km).

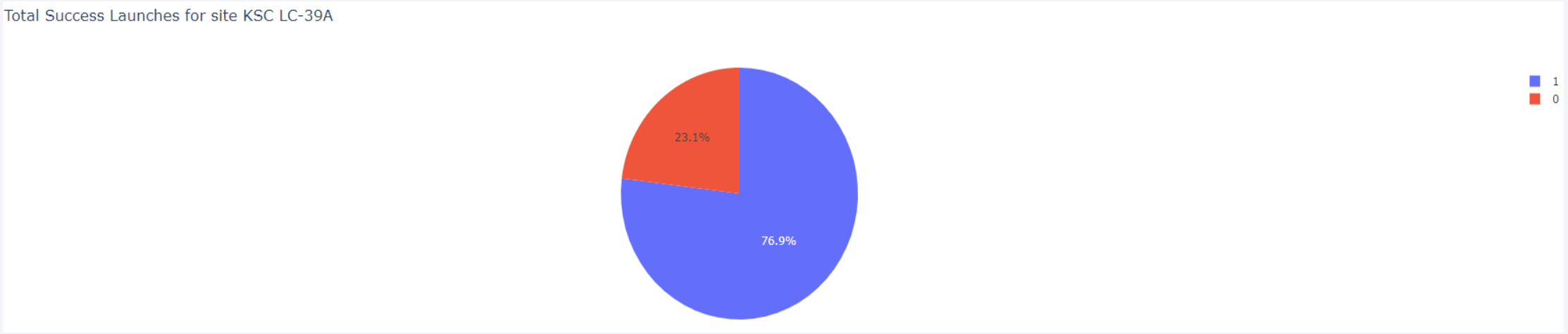

Section 4
Build a Dashboard with Plotly Dash

Launch success count for all sites



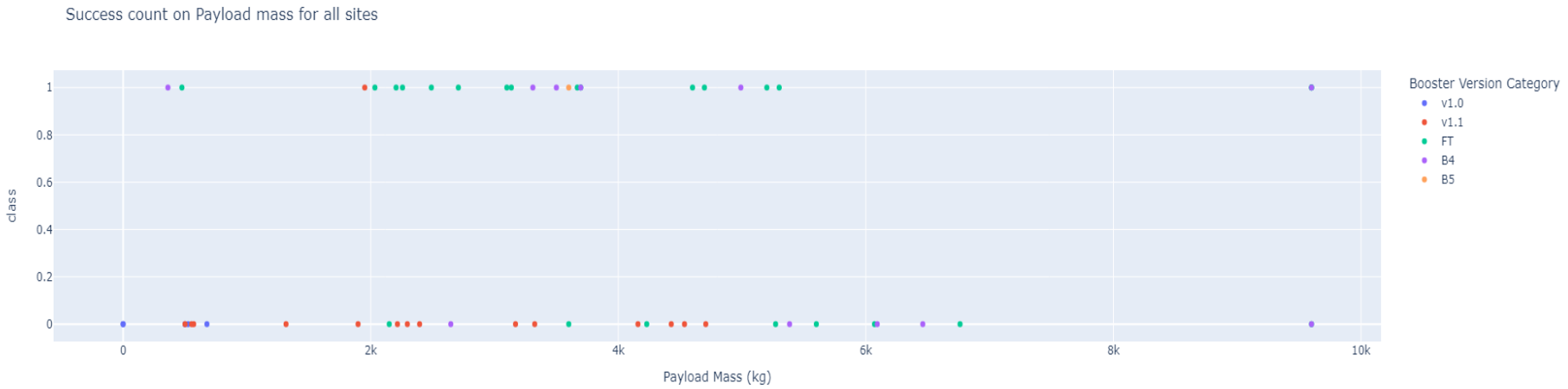
The chart shows that KSC LC-39A has the most successful launches.

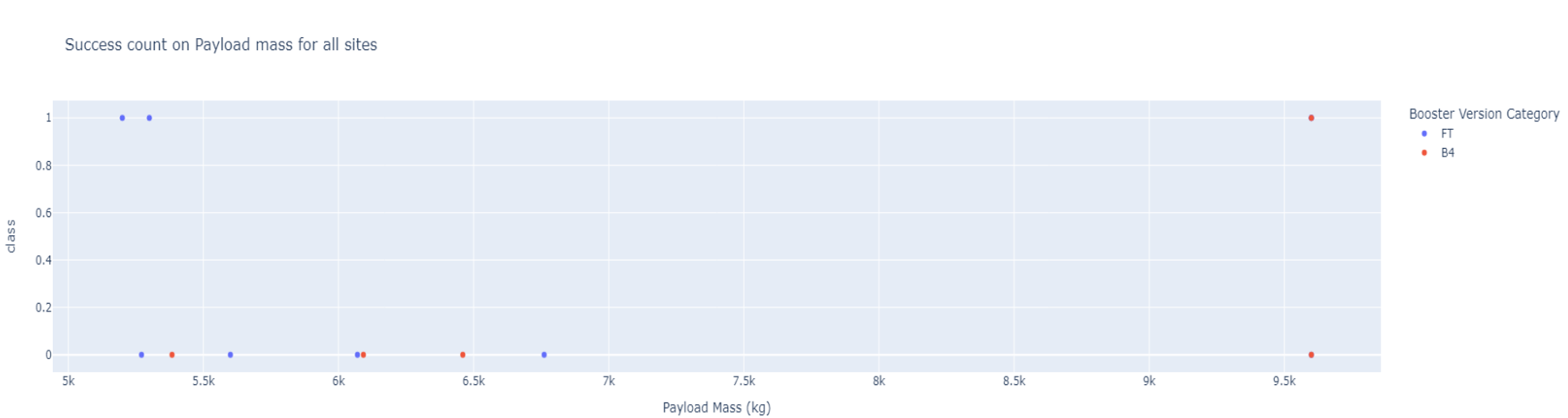
Launch site with highest launch success ratio.



For KSC LC-39A there are success rate of 76.9% with 10 successful and only 3 failed landings.

Payload Mass vs. Launch Outcome for all sites



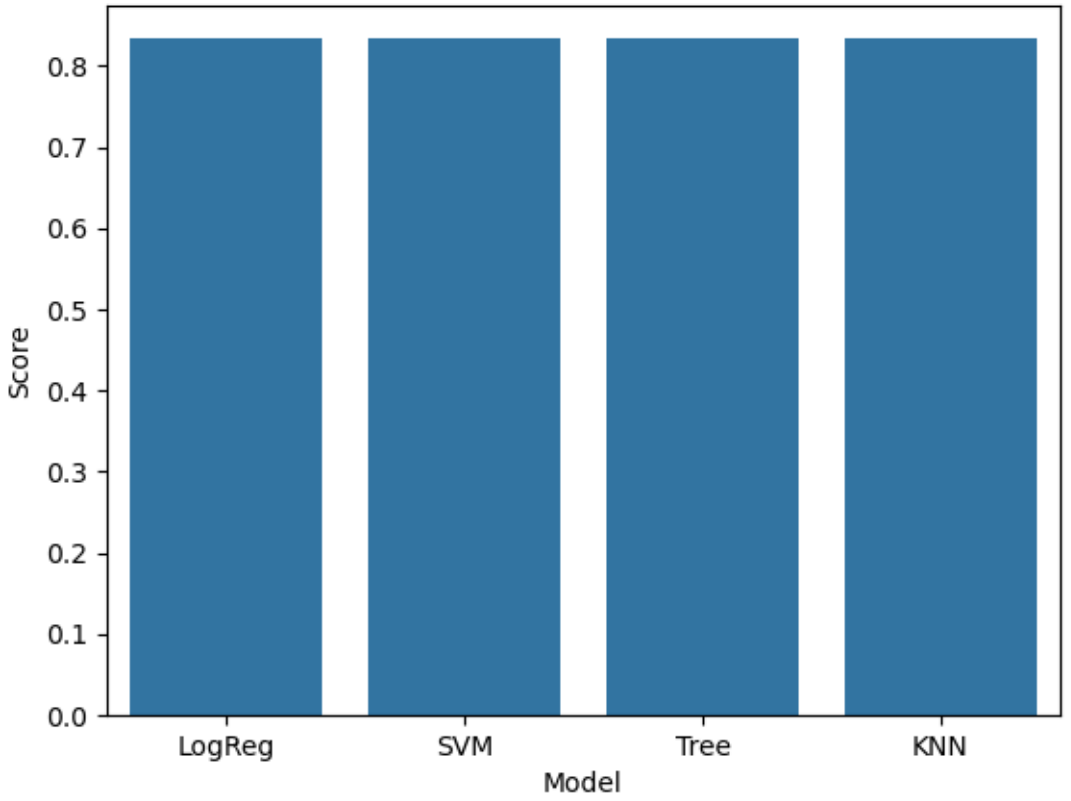


Payloads between 2000 and 5500 kg have the highest success rate.

Section 5

Predictive Analysis (Classification)

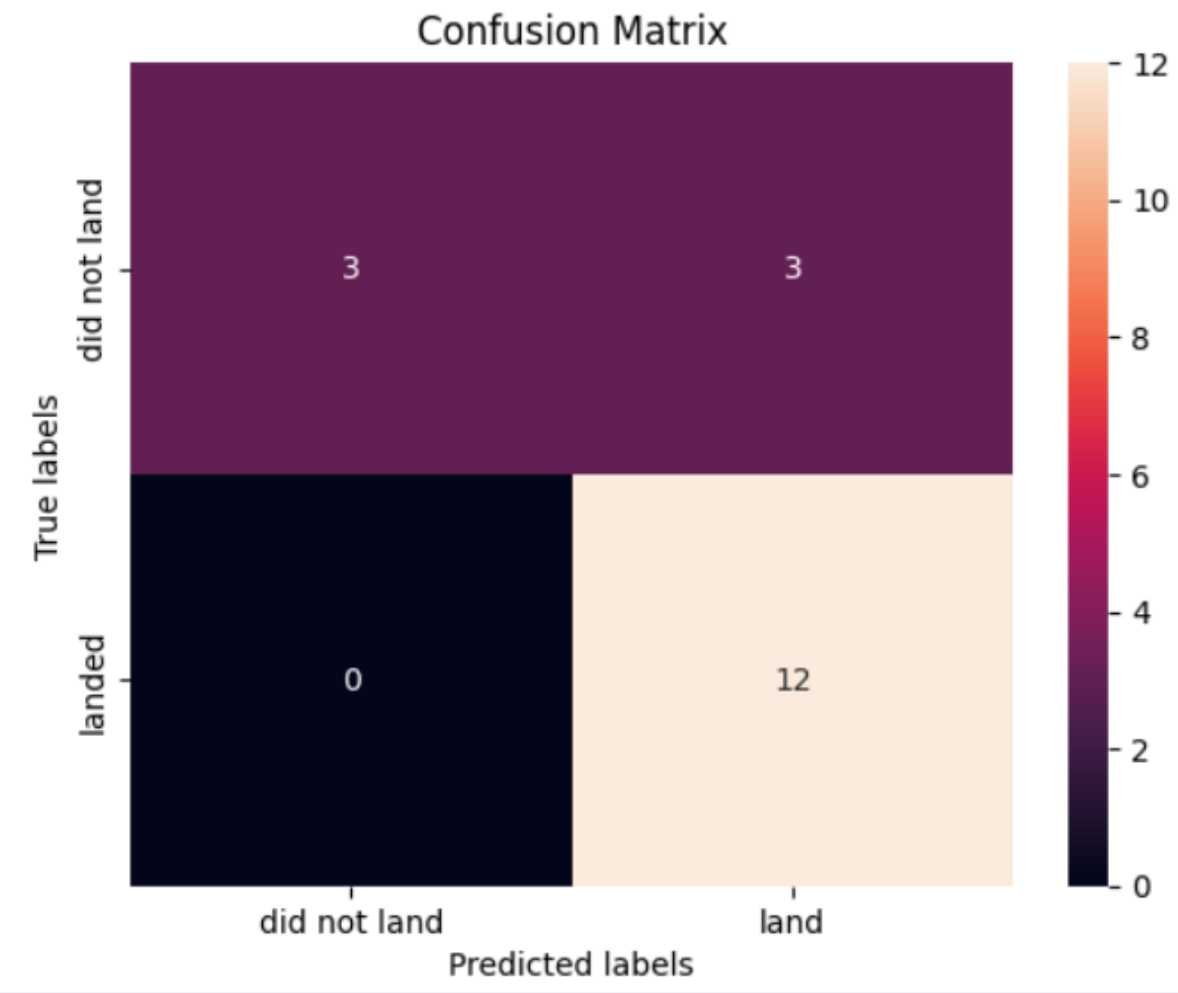
Classification Accuracy



Practically all algorithms give the same result.

- Accuracy for Logistics Regression method: 0.8333333333333334
- Accuracy for Support Vector Machine method: 0.8333333333333334
- Accuracy for Decision tree method: 0.8333333333333334
- Accuracy for K nearest neighbors method: 0.8333333333333334

Confusion Matrix



All models can distinguish between different classes.

The major problem is false positives.

Conclusions

- With the obtained results we can select any model to predict landing.
- Launches with a low payload mass show better results than launches with a larger payload mass.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- The success rate of launches increases over the years.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.

Appendix

- Code to create bar chart for the different classifiers.

```
labels = ['LogReg', 'SVM', 'Tree', 'KNN']
score = []
score.append(logreg_cv.score(X_test, Y_test))
score.append(svm_cv.score(X_test, Y_test))
score.append(tree_cv.score(X_test, Y_test))
score.append(knn_cv.score(X_test, Y_test))

list_of_tuples = list(zip(labels, score))
models = pd.DataFrame(list_of_tuples, columns=['Model', 'Score'])
ax = sns.barplot(models, x='Model', y='Score', hue='Model')
ax.bar_label(ax.containers[0], fontsize=10)
fig = bar.get_figure()
fig.savefig("bar.png")
```

Thank you!