IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Allen Yonemoto>
<6/12/2024>

Outline

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

Executive Summary

- Summary of methodologies

1. EDA with Data Visualization
2. Interactive map with Folium
3. Dashboard with Plotly Dash
4. Predictive Analysis

- Summary of all results

1. EDA results
2. Interactive analytics demo
3. Predictive Analysis Results

Introduction

- Project background and context

SpaceX has been the most successful company of the commercial space age, to make space travel affordable. The company advertises Falcon 9 rocket launches on the website, with a cost worth 62 million dollars. Other providers cost upward of 165 million dollars each, with much of the savings since SpaceX can reuse the first stage. If we are able to determine if the first stage will land, we can determine the cost of a launch. Based on our public information and machine learning models, we will predict if SpaceX will reuse the first stage.

- Problems you want to find answers

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 throughout the years?

What best algorithm can we use for binary classification?

Section 1

Methodology

Methodology

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 method to prepare the data to 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 evaluating classification models to display 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 wiki entry.

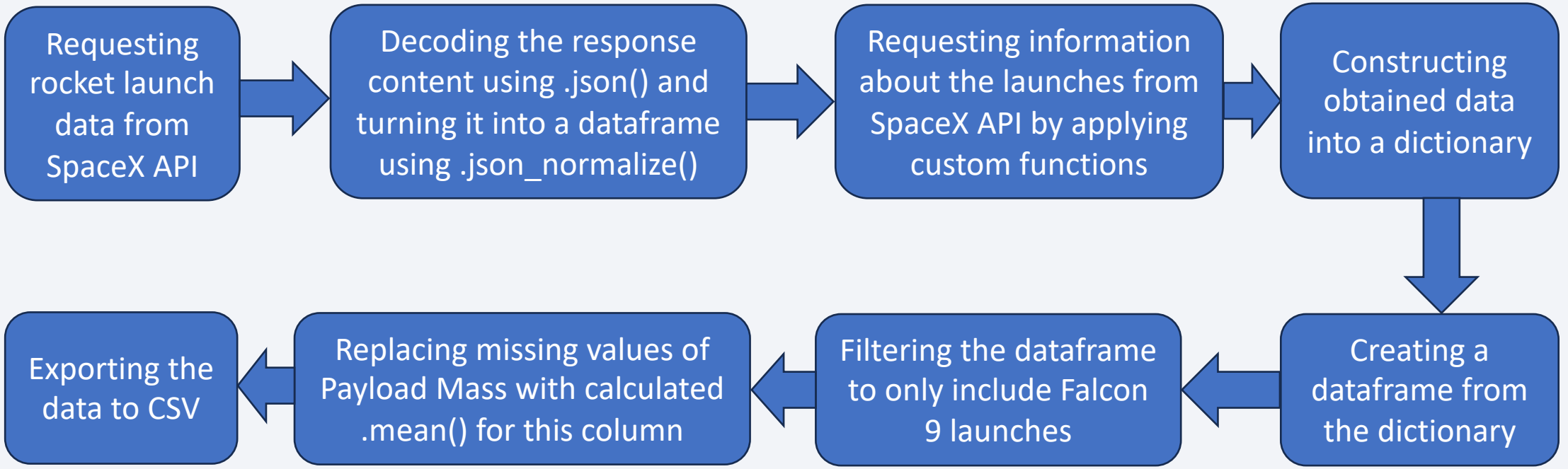
Data Columns are obtained by using the SpaceX REST API:

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

Data Columns are obtained by using the Wikipedia Web Scraping:

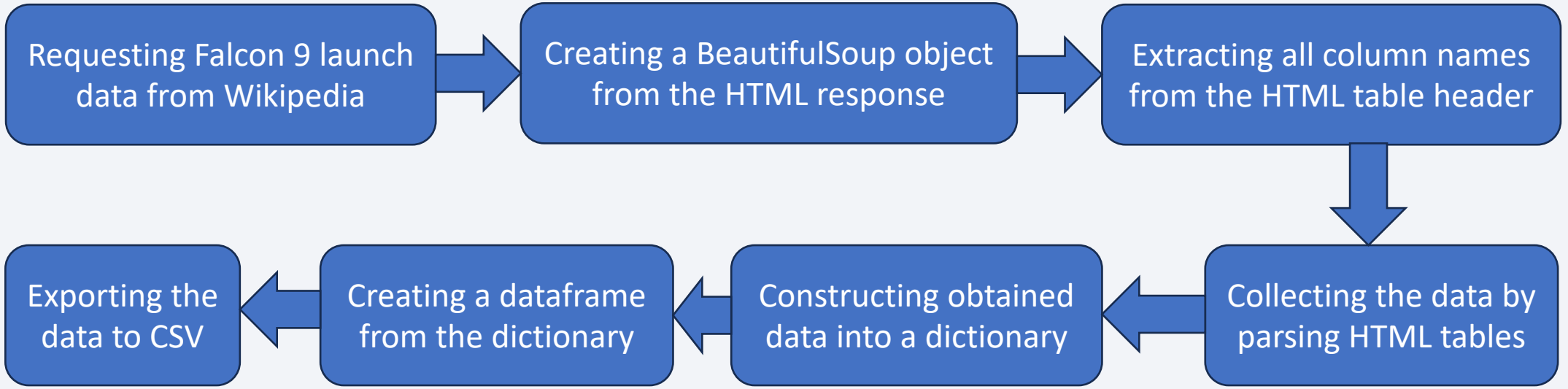
- Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



GitHub URL: Data Collection API

Data Collection - Scraping

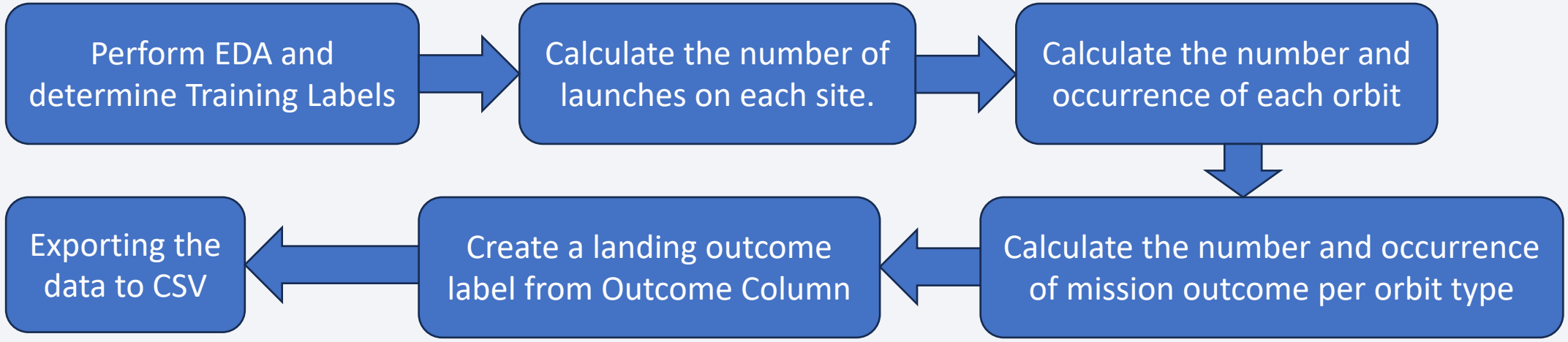


GitHub URL: Data Collection with Web Scraping

Data Wrangling

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

We mainly convert the outcomes into Training Labels with “1” meaning the booster successfully landed and “0” meaning unsuccessful.



GitHub URL: Data Wrangling

EDA with Data Visualization

Charts plotted:

- Flight Number v. 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 show the relationship between variables. Any existing relationship could be used in a machine learning model.

Bar plots show comparison among discrete sections. The idea is to show the relationship between selected categories that are compared and a measured value.

Line charts display trends in data over time.

EDA with SQL

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string ‘CCA’
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have a payload mass in between 4000 and 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in 2015
- Ranking the count of landing outcomes or success between 4/6/2010 and 3/20/2017 in descending order.

Build an Interactive Map with Folium

Markers of all Launch Sites:

- Added marker with Circle, Popup Label, and Text Label of NASA Johnson Space Center using the latitude and longitude coordinates for starting location.
- Added the markers with Circle, Popup Label, and Text Label of all launch sites using the latitude and longitude coordinates to show geographical locations and the proximity Equator and coasts.

Colored Markers of all launch outcomes for each launch site

- Used the green marker for successful launches and red marker for failed launches by using the marker cluster to identify which launch sites the highly successful rates appeared in.

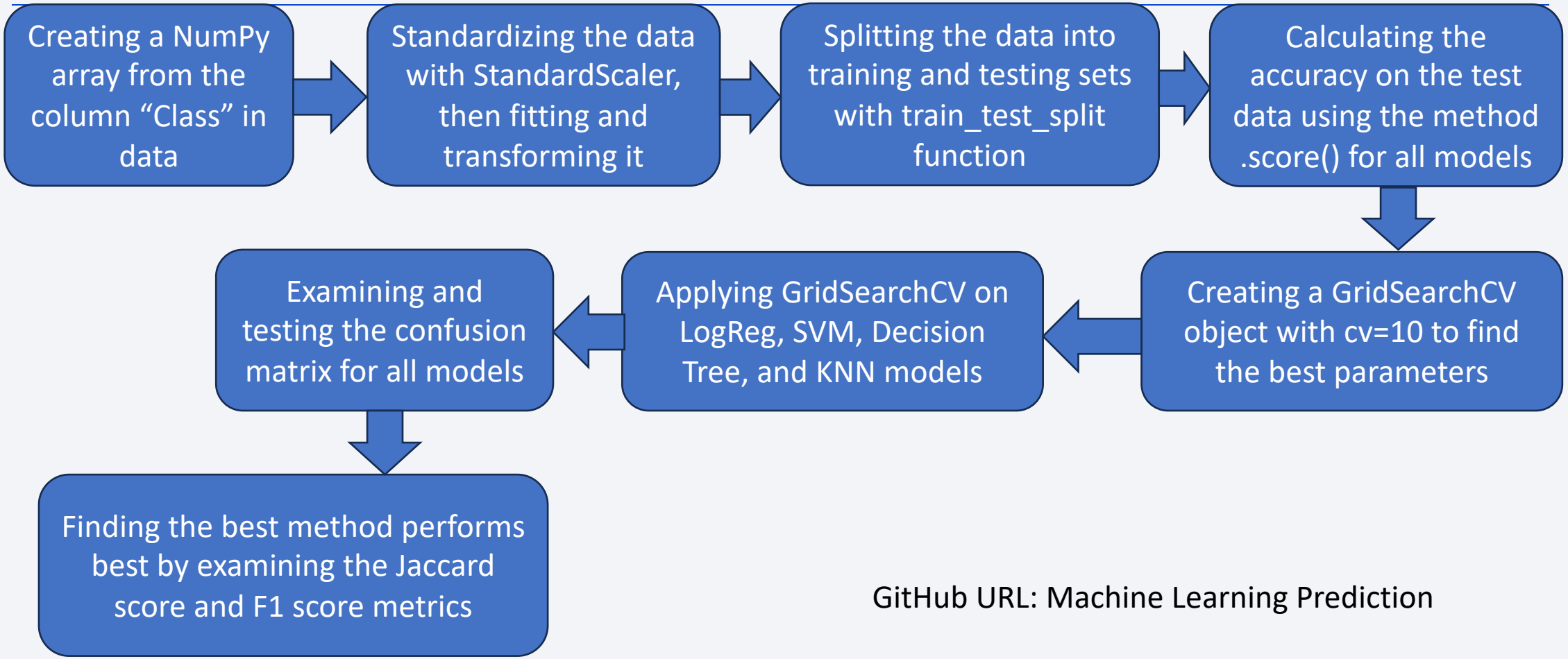
Distances between a Launch Site to its proximities:

- Used colored lines to show the distances between a Launch Site and its proximities such as Railway, Highway, Coastline, and Closest City.

Build a Dashboard with Plotly Dash

- Added a dropdown list to enable Launch Site selection.
- Added a pie chart to show the total number of successful launches of all sites and Success vs. Failed counts for the site, whenever a Launch Site was selected.
- Added a slider to select Payload range.
- Added a scatter chart to show the correlation between Payload and Launch Success.

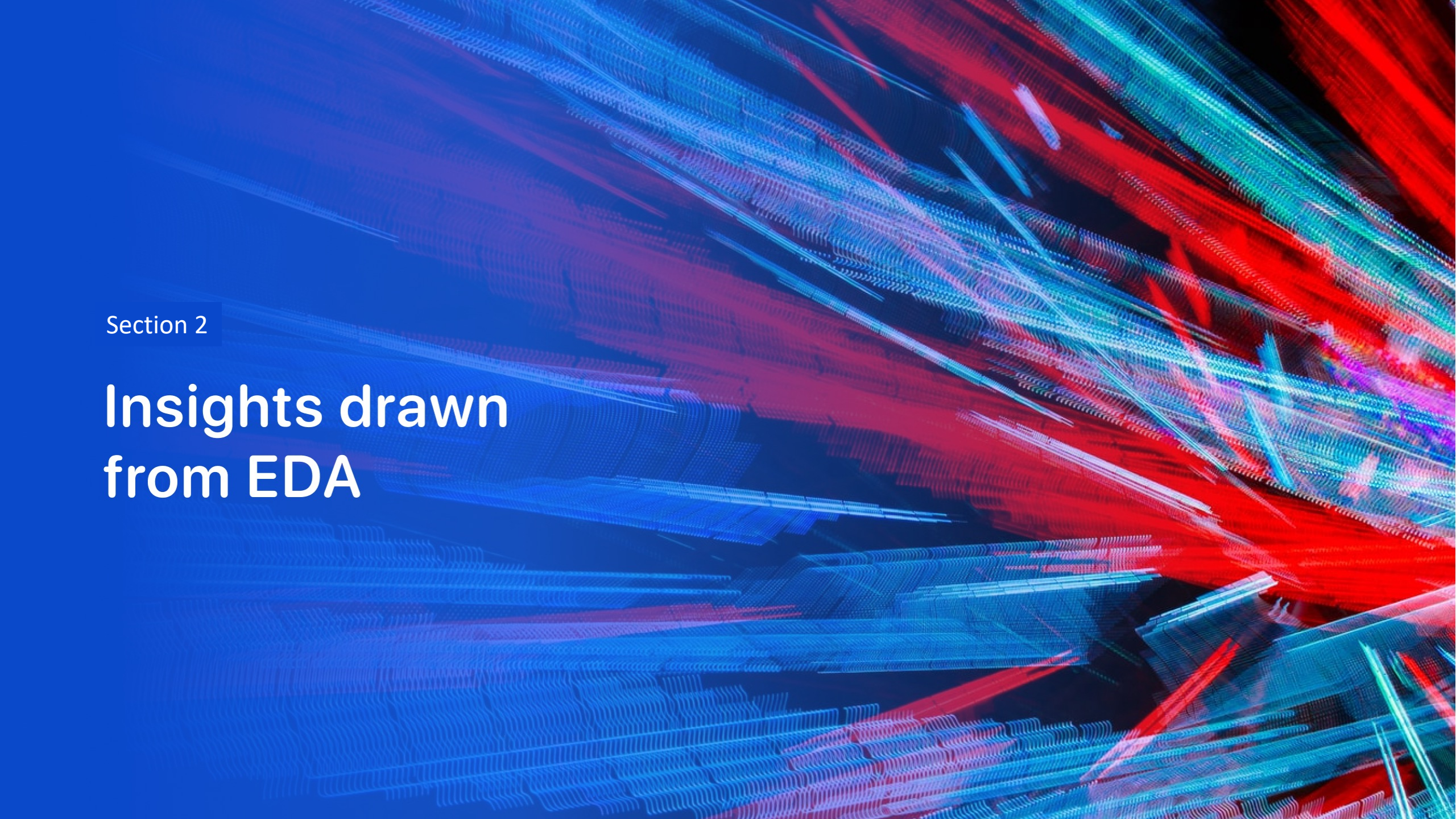
Predictive Analysis (Classification)



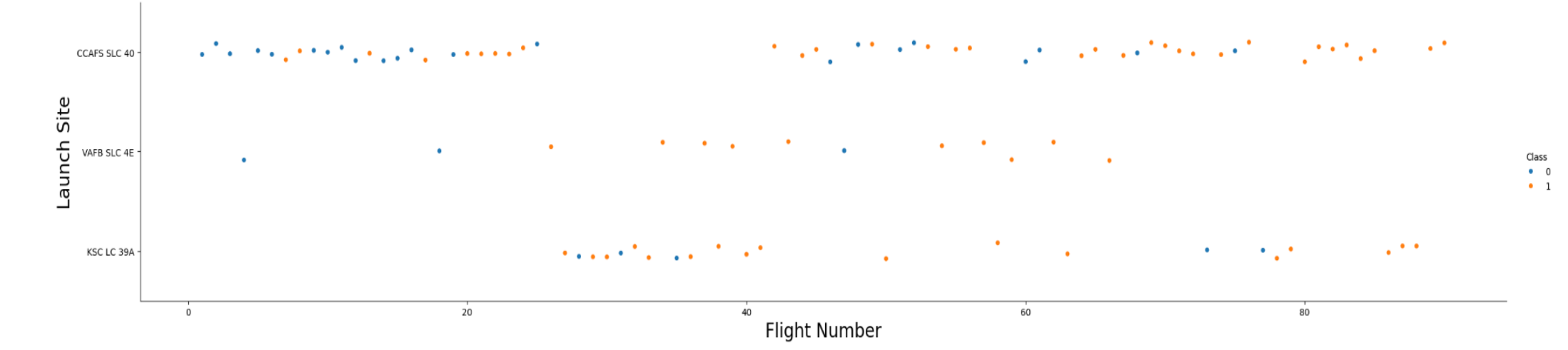
GitHub URL: Machine Learning Prediction

Results

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


Section 2
Insights drawn from EDA

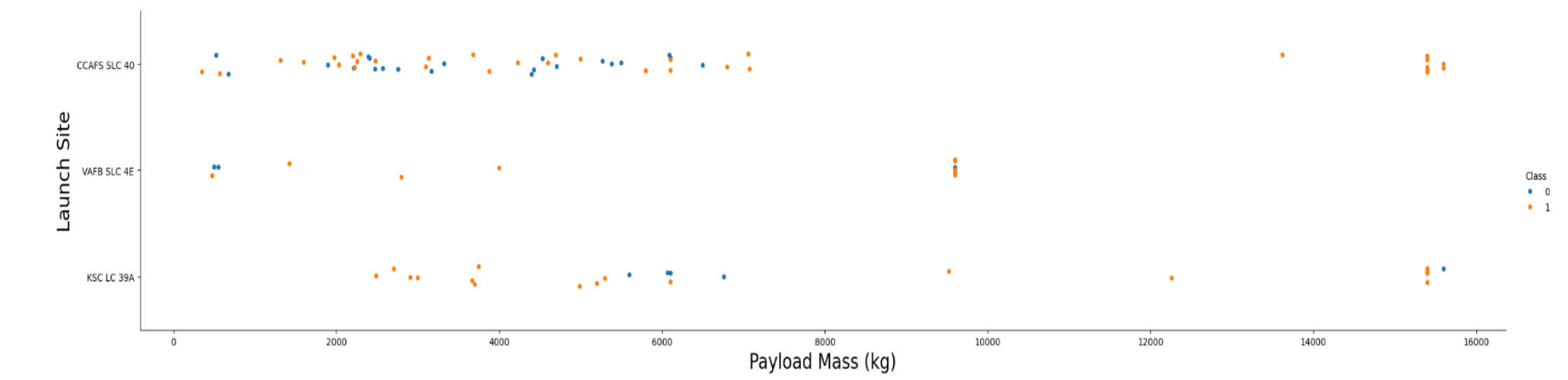
Flight Number vs. Launch Site



Details:

- The earliest flights failed while the latest and most recent flights succeeded.
- The CCAFS SL40 launch site has about half of all launches.
- VAFB SCL 4E and KSC LC 39A have the most highest success rates measured.
- There could be an assumption that a newer launch will have a higher rate of success.

Payload vs. Launch Site



Details:

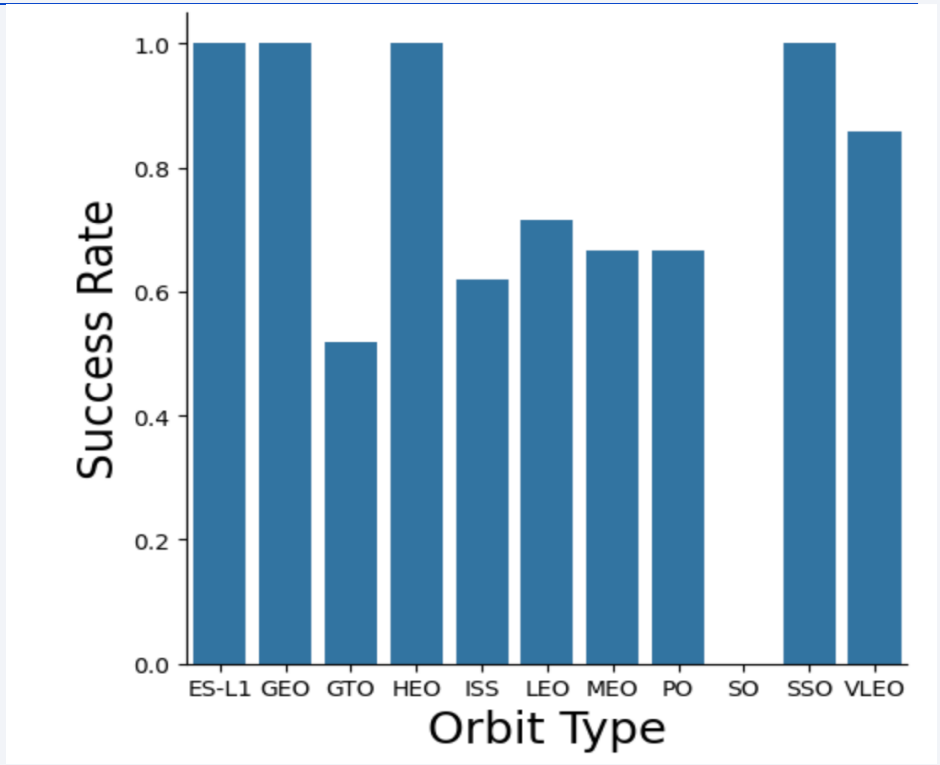
- For every launch site, the higher the payload mass, the higher the success rate is.
- Most of the launches with payload mass over 7000kg were successful.
- KSC LC 39A has 100% success rate for payload mass under 5500kg too.

Success Rate vs. Orbit Type

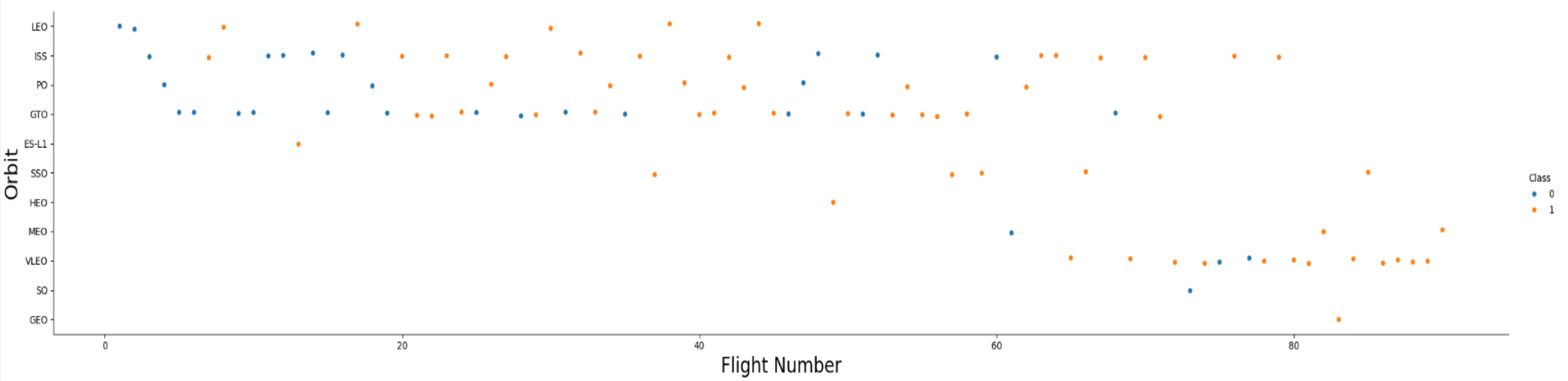
Details:

- Orbits with 100% success rate:
  - ES-L1, GEO, HEO, SSO
- Orbits with no success rate: SO
- Orbits with success rate between 50% and 85%:

  - GTO, ISS, LEO, MEO, PO



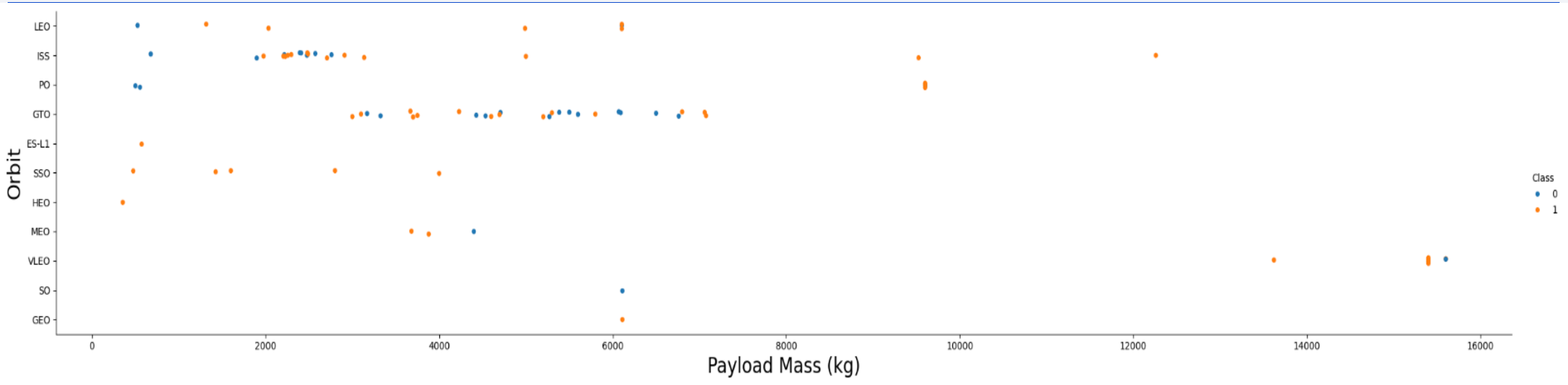
Flight Number vs. Orbit Type



Details:

- In the LEO orbit, the Success happens to be related to the number of flight. There appears to be no relationship between flight number when it is in the GTO orbit.

Payload vs. Orbit Type

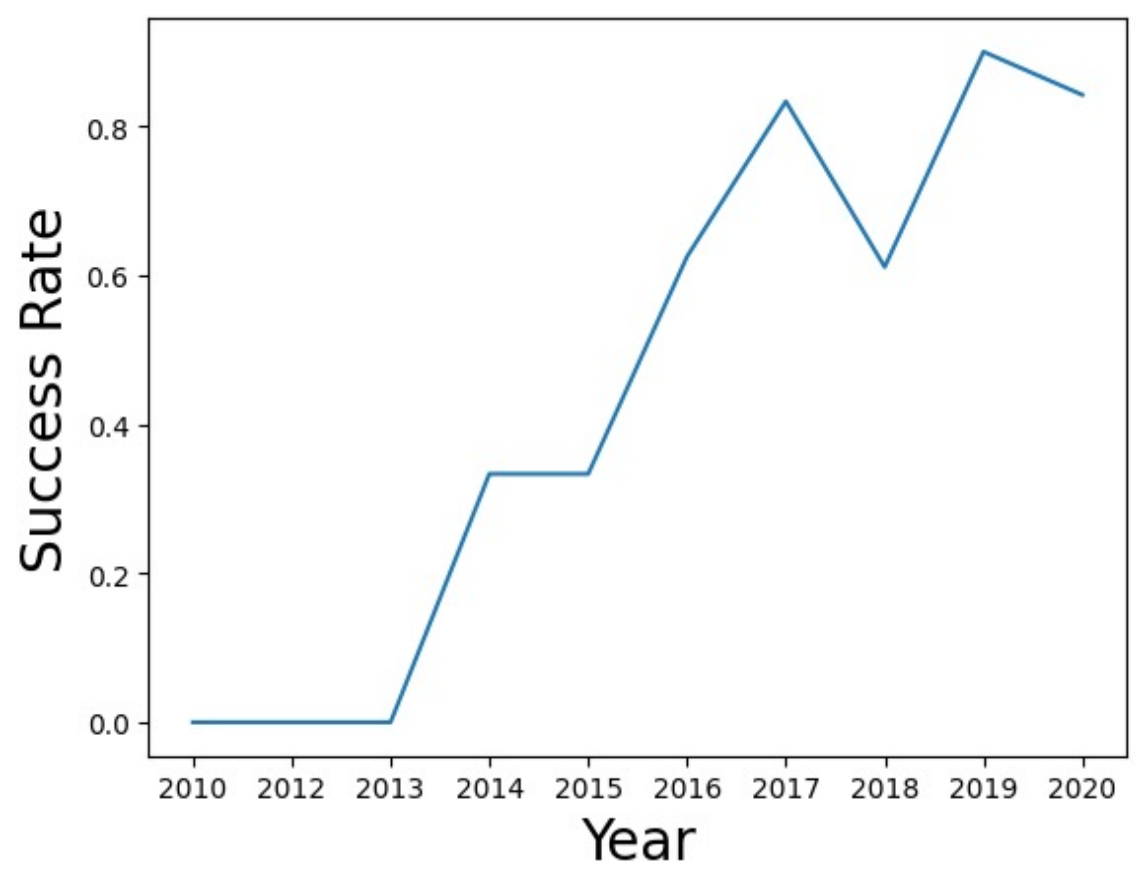


Details:

- Heavy payloads have a negative influence on GTO orbits and positive influence on GTO and ISS orbits.

Launch Success Yearly Trend

Details: The launch success rate increased from 2013 to 2020.



All Launch Site Names

Task 1

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

```
[9]: sql select distinct launch_site from SPACEXTBL;
 * sqlite:///my_data1.db
Done.
```

[9]:

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

Details:

- Displaying the names of the unique launch sites in the space mission.

Launch Site Names Begin with 'CCA'

Task 2

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

[10]:
```
%sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5;
```

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

[10]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcom
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachut
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 (parachut
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attem
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attem
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attem

Details:

- Displaying 5 records where launch sites begin with ‘CCA’.

Total Payload Mass

Task 3

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

```
[11]: %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXTBL where customer = 'NASA (CRS)';
```

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

[11]:

total_payload_mass
45596

Details:

- Displaying the total payload mass carried by boosters launched by NASA (CRS).

Average Payload Mass by F9 v1.1

Task 4

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

```
[12]: %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXTBL where booster_version like '%F9 v1.1%';
 * sqlite:///my_data1.db
Done.
```

[12]:

average_payload_mass
2534.6666666666665

Details:

- Displaying average payload mass carried by booster version F9 v1.1.

First Successful Ground Landing Date

Task 5

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

Hint:Use min function

```
[13]: %sql select min(date) as first_successful_landing from SPACEXTBL where landing_outcome = 'Success (ground pad)';
 * sqlite:///my_data1.db
Done.
```

[13]:

first_successful_landing
2015-12-22

Details:

- List the date when the first successful landing outcome in ground pad was achieved.

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
[14]: %sql select booster_version from SPACEXTBL where landing_outcome = 'Success (drone ship)' and payload_mass__kg_ between
```

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

[14]:

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

Details:

- Listing the names of the boosters that have success in drone ship and have payload mass greater than 4000 and less than 6000

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

```
[15]: %sql select mission_outcome, count(*) as total_number from SPACEXTBL group by mission_outcome;
```

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

[15]:

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

Details:

- Listing the total number of successful and failure mission outcomes.

Boosters Carried Maximum Payload

Task 8

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

```
[16]: %sql select booster_version from SPACEXTBL where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTBL);
```

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

[16]:

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

Details:

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

2015 Launch Records

Task 9

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, date, booster_version, launch_site, landing_outcome from SPACEXTBL where landing_
```

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

Details:

- Listing the failed landing outcomes in drone ship, their booster versions, and launch site names for the months in 2015

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

Task 10

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 count_outcomes from SPACEXTBL where date between '2010-06-04' and '2017-03-20'
```

```
 * 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

New Tab
chrome://newtab

Details:

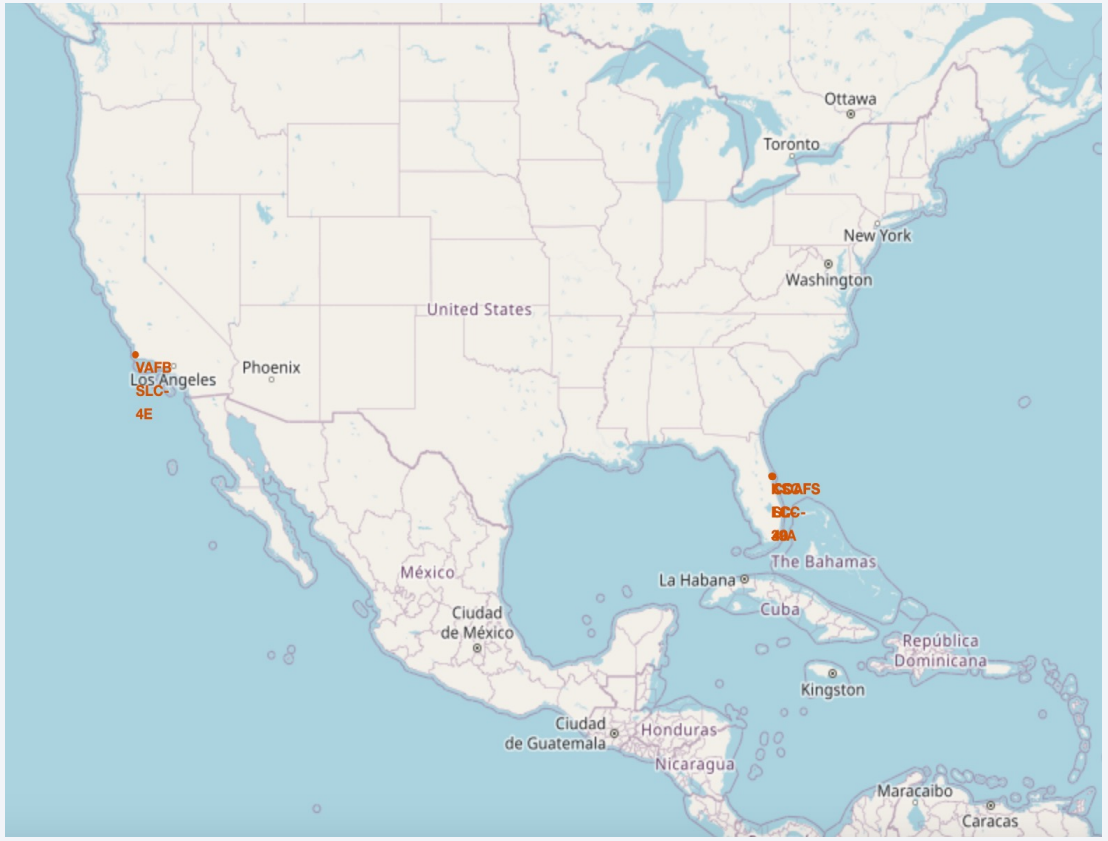
- Ranking the count of landing outcomes or Success between 2010-06-04 and 2017-03-20 in descending order.


Section 3
Launch Sites
Proximities Analysis

All launch sites' location markers on a global map

Details:

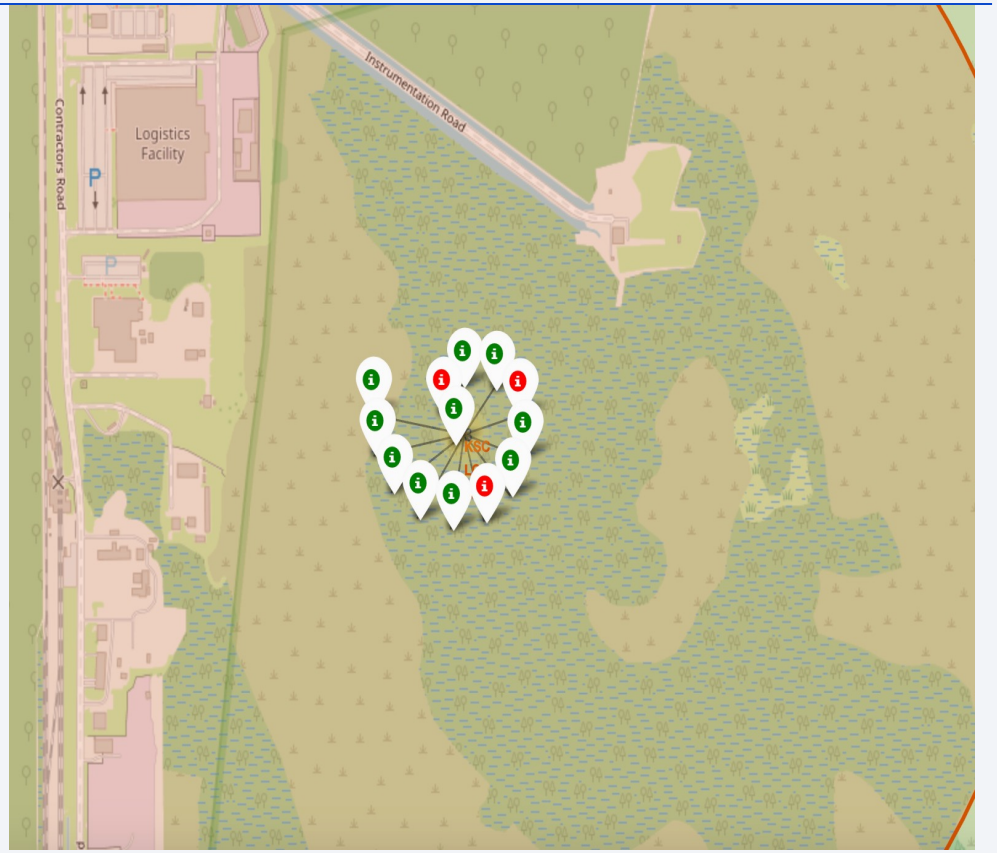
- Most Launch sites are in proximity to the Equator line. The land moves faster at the equator than any other place in the surface of Earth. Anything on the surface of Earth at the equator is already moving at 1670km/hour. If a ship is launched from the equator, it goes up to space and is also moving around the Earth at the same speed it was moving before launching due to inertia. The speed will help the spacecraft keep up a good enough speed to stay in orbit.
- All launch sites are in 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

Details:

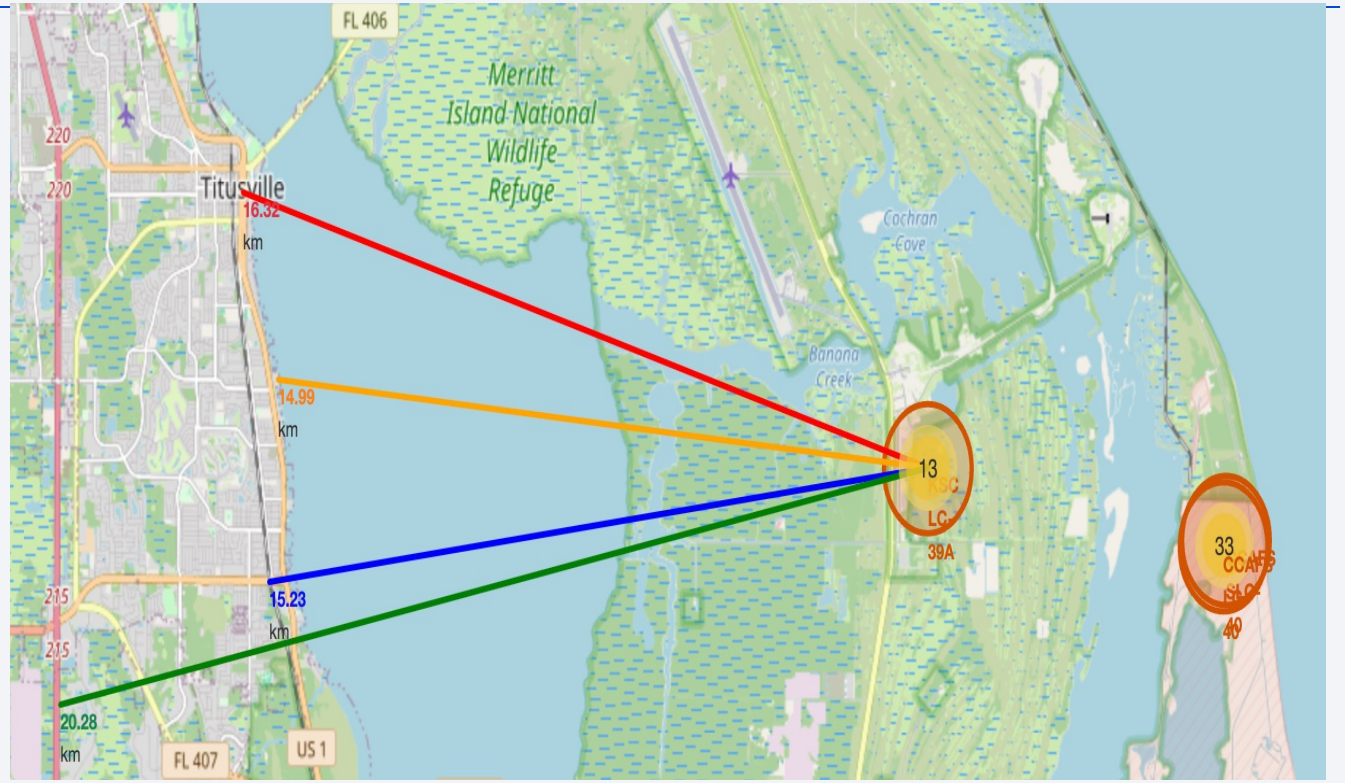
- By using the color-labeled markers, we will easily identify which launch sites have relatively high success rates.
- Green Marker = Successful Launch
- Red Marker = Failed Launch
- Launch Site KSC LC-39A has a high Success Rate.

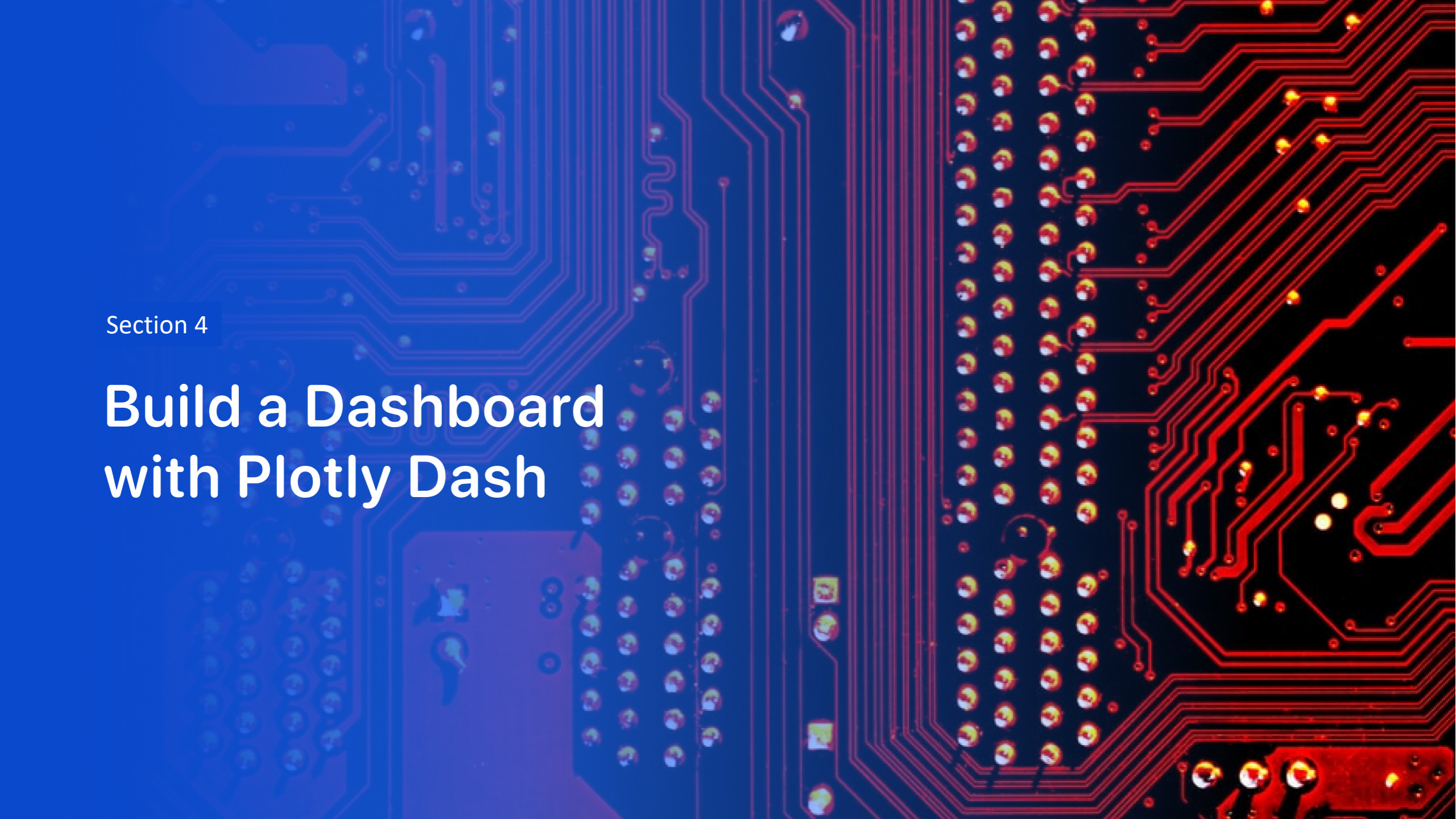


Distance from the launch site KSC LC-39A to its proximities

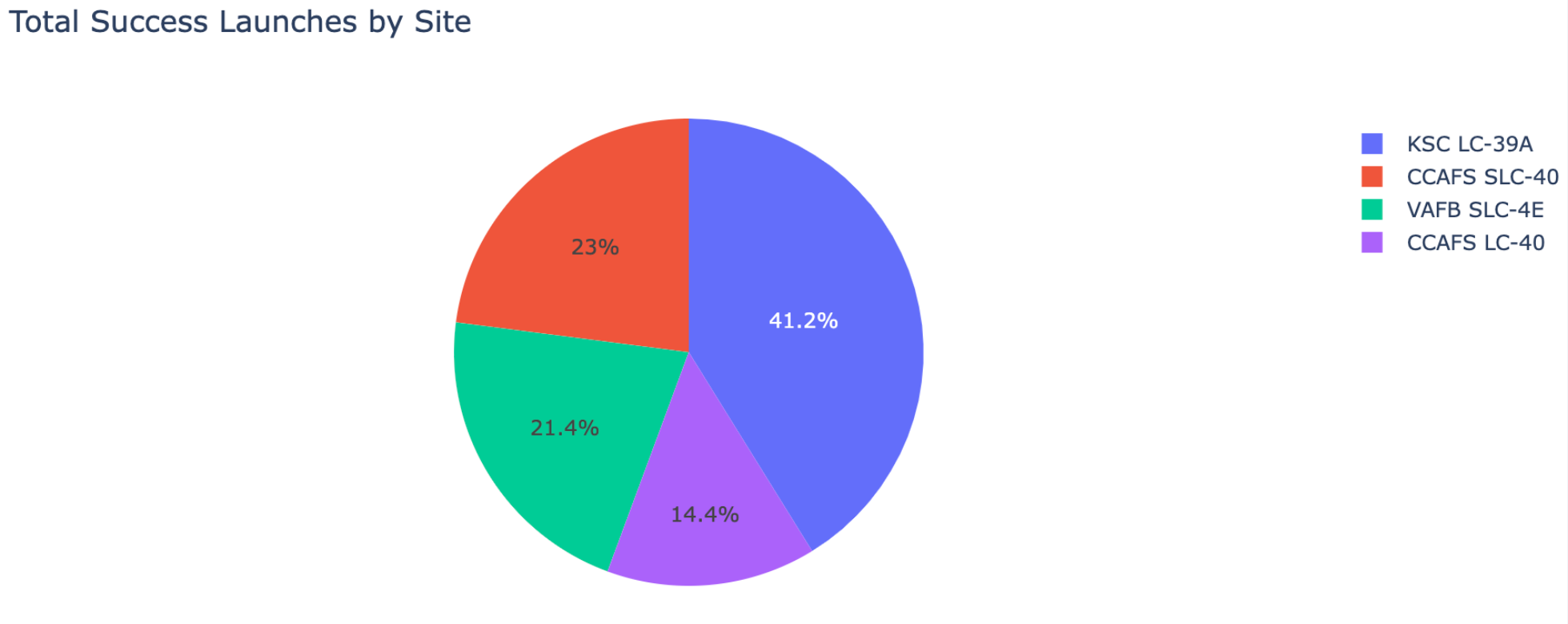
Details:

- From the visual analysis of the launch site KSC LC-39A we can see that it is:
  - relative close to railway (15.23km)
  - relative close to highway (20.28km)
  - relative close to coastline (14.99km)
- The launch site KSC LC-39A is relative close to its closest city, Titusville by 16.32km.
- The failed rocket with its high speed can cover distances like 15-20km in few seconds, which can be potentially dangerous to populated areas.




Section 4
Build a Dashboard with Plotly Dash

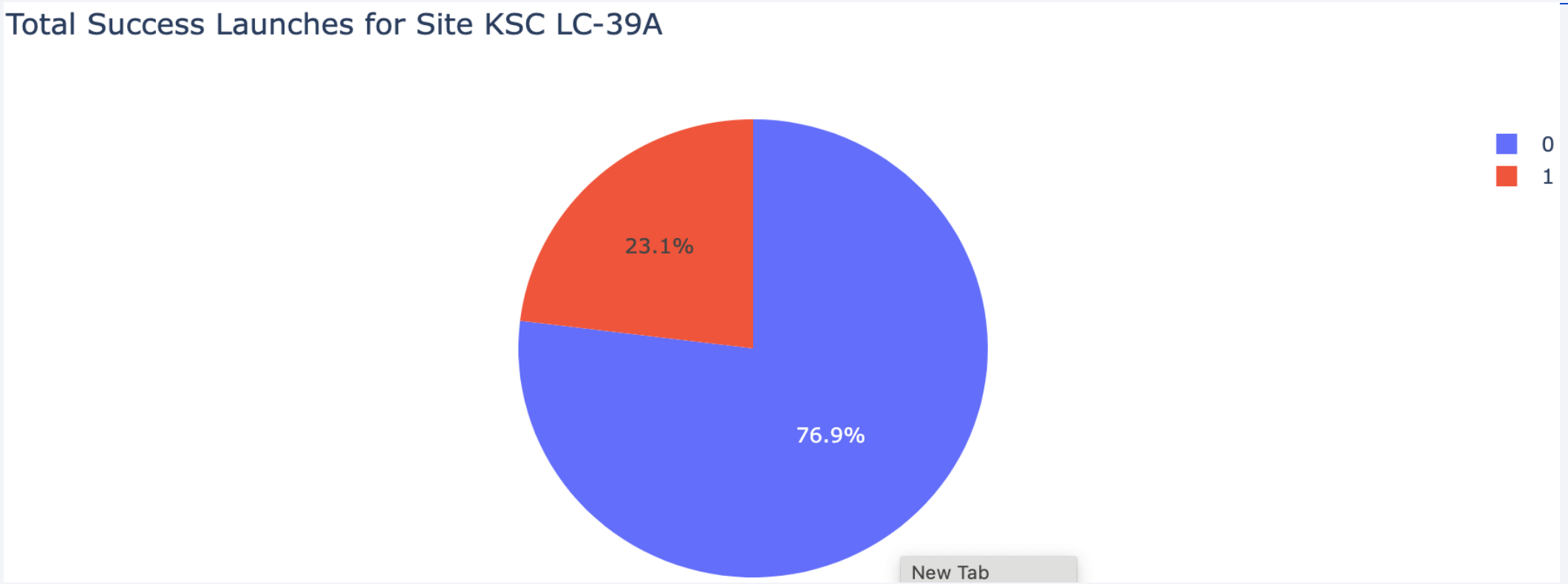
Launch Success Count for all Sites



Details:

- The chart shows the percentage count from all sites with KSC LC-39A having the most successful launches.

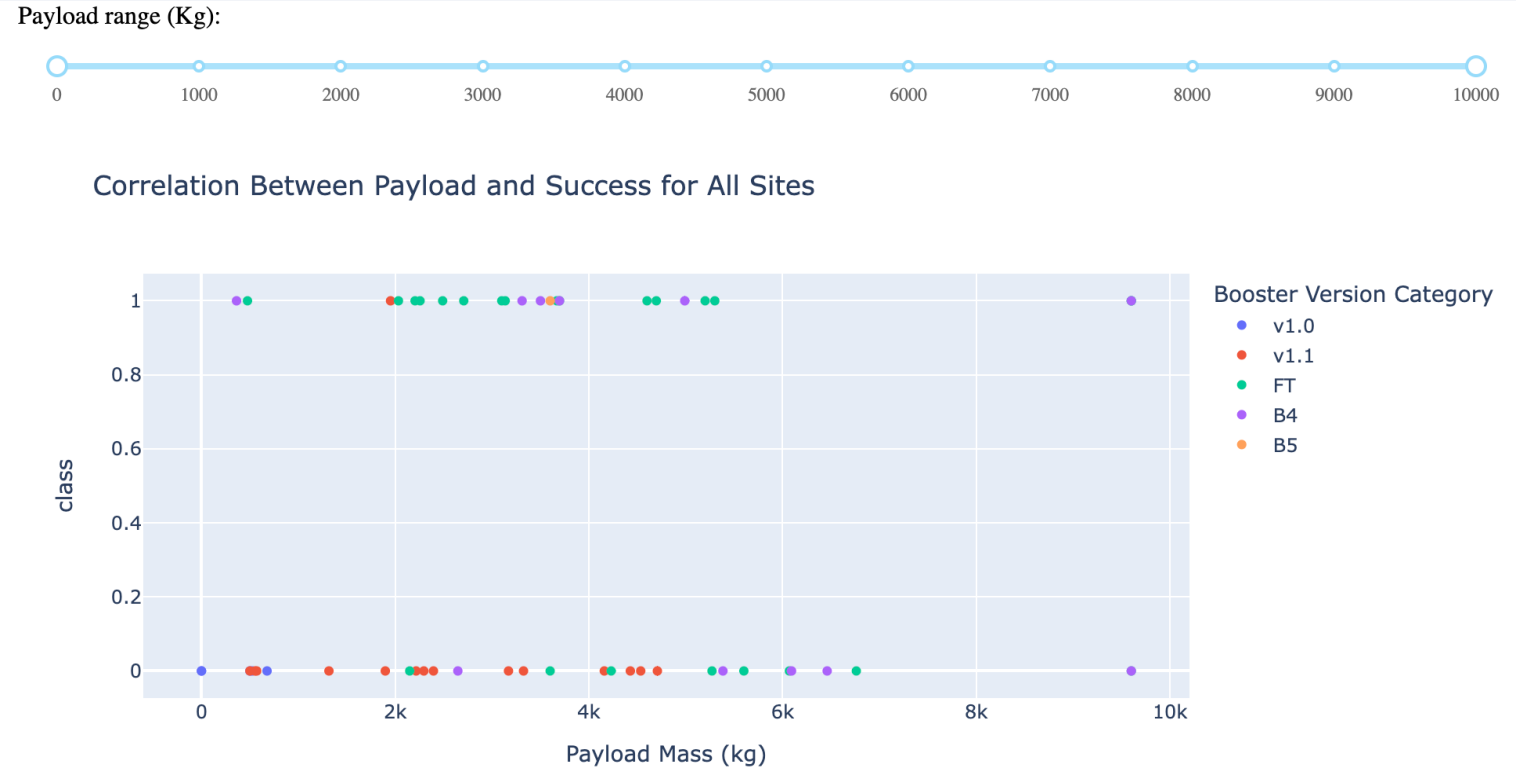
Launch site with highest Launch Success Ratio



Details:

- KSC LC-39A has the highest launch success rate at 76.9% with 10 successful and only 3 failed landings.

Payload Mass vs. Launch Outcome for all sites



Details:

- The chart shows that payloads between 2000kg and 5500kg have the highest success rate.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Scores and Accuracy of 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 Whole Data Set

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.666667	0.819444
F1_Score	0.909091	0.916031	0.800000	0.900763
Accuracy	0.866667	0.877778	0.755556	0.855556

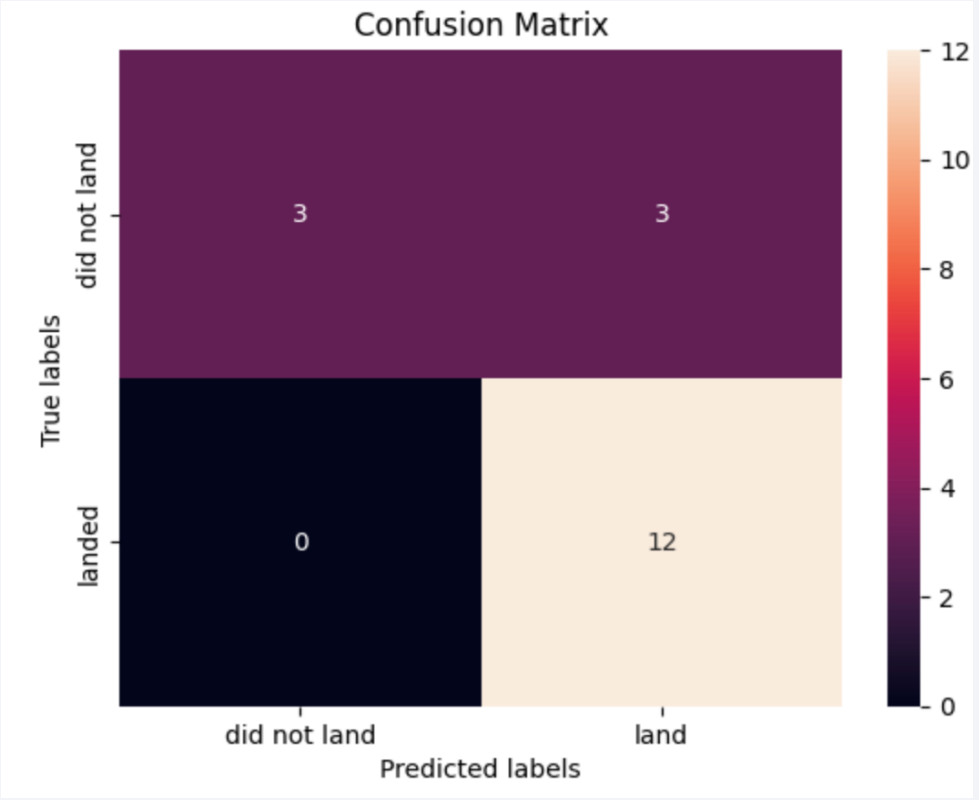
Details:

- Due to the scores of the Test Set being the same for all categories, we are not able to decide which method works best.
- Same Test Set scores are likely to be due to the small test sample size. So, we tested all methods based on the whole dataset.
- The scores of the whole dataset show that the Decision Tree Model works best. Not only does the model have higher scores, but also has the highest accuracy.

Confusion Matrix

Details:

- Examining the confusion matrix, the logistic regression can distinguish between the different classes. The major problem is false positives.



Conclusions

- The Decision Tree Model is the best algorithm to use for the dataset.
- Launches with low payload mass show better results than launches with larger payload mass.
- Most of the launch sites are in proximity to the Equator line, and all sites are in 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 sites.
- Orbits ES-L1, GEO, HEO, and SSO have a perfect 100% success rate.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!