IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

UTKARSH KUMAR GUPTA

01-03-2024

OUTLINE

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

EXECUTIVE SUMMARY

- In this capstone project, we will predict if the SpaceX Falcon 9 first stage will land successfully using several machine learning classification algorithms.
- The main steps in this project include:
  - Data collection, wrangling, and formatting
  - Exploratory data analysis
  - Interactive data visualization
  - Machine learning prediction
- It is also concluded that Logistic Regression may be the best machine learning algorithm to predict if the Falcon 9 first stage will land successfully or not.

INTRODUCTION

- In this capstone, we will 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; 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. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.
- The main question that we are trying to answer is, for a given set of features about a Falcon 9 rocket launch which include its payload mass, orbit type, launch site, and so on, will the first stage of the rocket land successfully?

Section 1

Methodology

METHODOLOGY

- The overall methodology includes:

1. Data collection, wrangling, and formatting, using:
   - SpaceX API
   - Web scraping
2. Exploratory data analysis (EDA), using:
   - Pandas and NumPy
   - SQL
3. Data visualization, using:
   - Matplotlib and Seaborn
   - Folium
   - Dash
4. Machine learning prediction, using
   - Logistic regression
   - Support vector machine (SVM)
   - Decision tree
   - K-nearest neighbors (KNN)

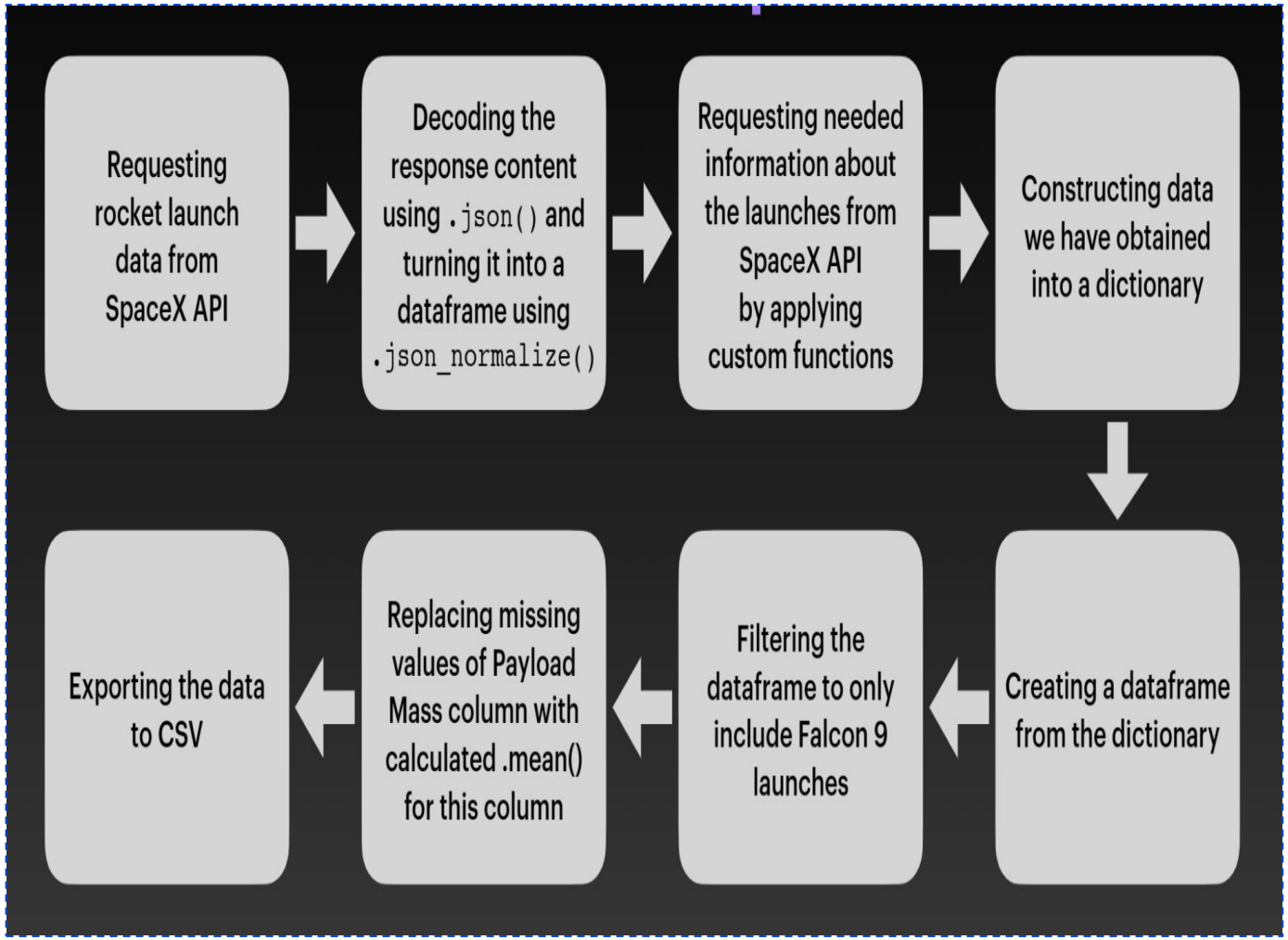
Data collection

- 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, Booster Version, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, LatitudeData
- 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

- The API used is https://api.spacexdata.com/v4/launches/past

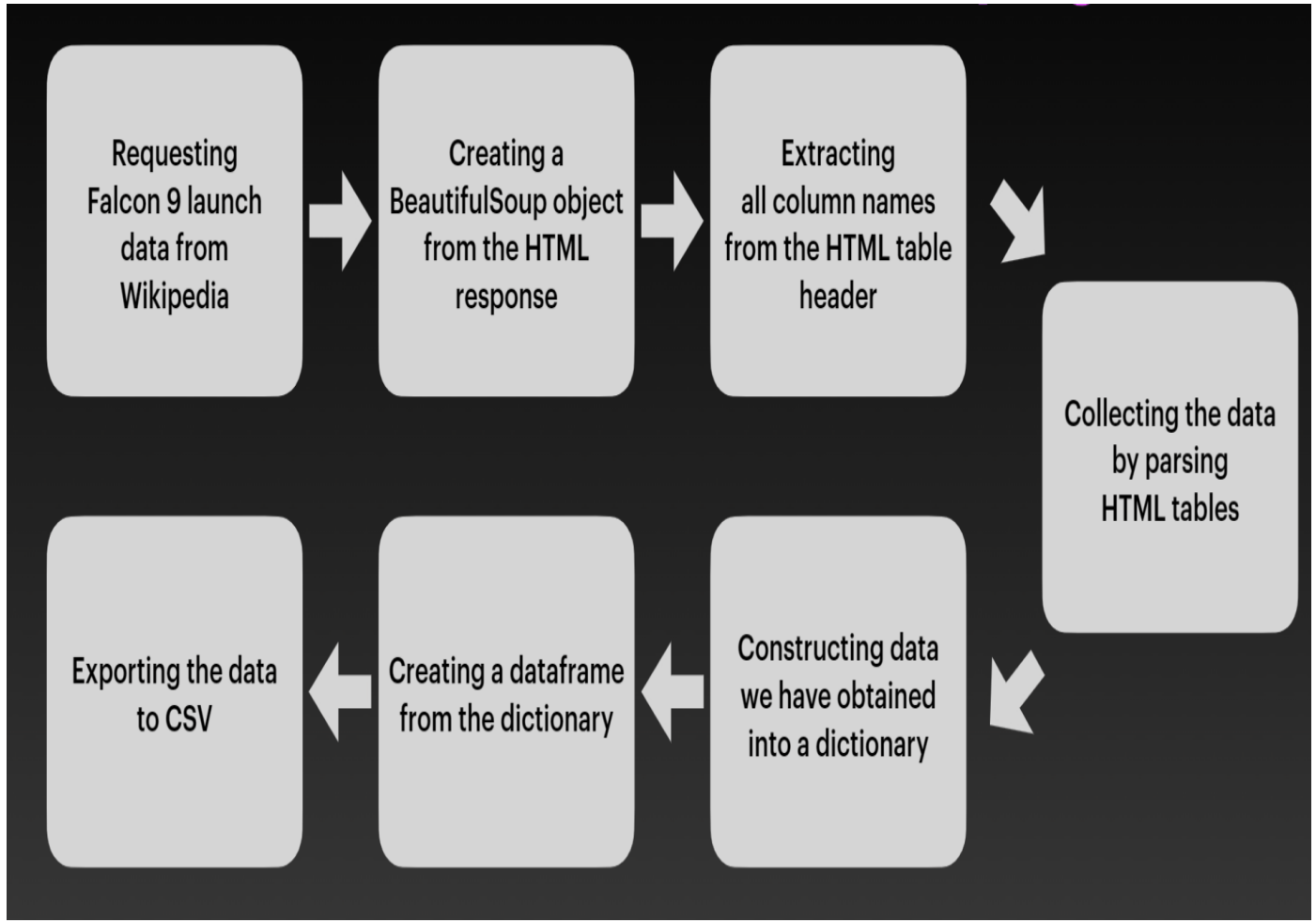
- GitHub URL of SpaceX API calls notebook



Data Collection - Scraping

- The data is scraped from https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

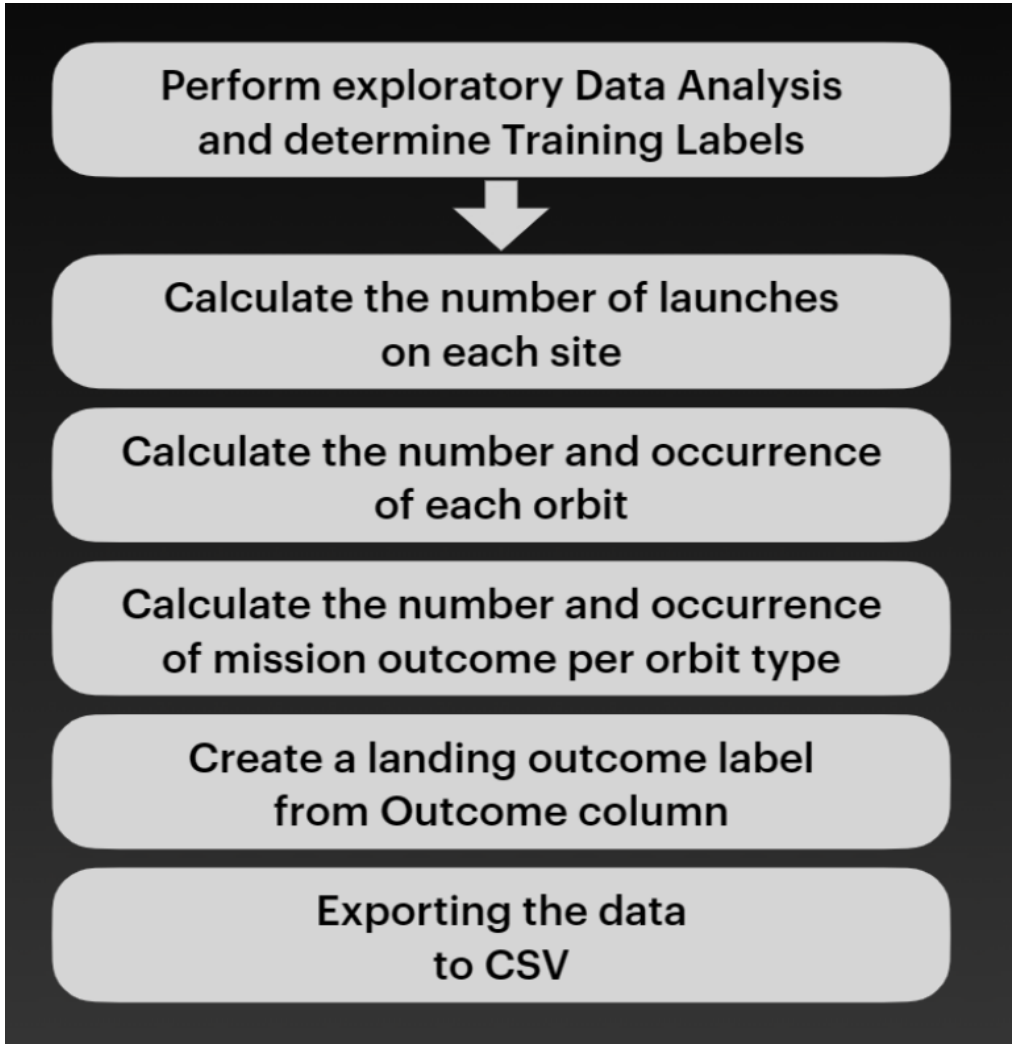
- GitHub URL of the web scraping notebook



Data Wrangling

- The data is later processed so that there are no missing entries and categorical features are encoded using one-hot encoding.
- An extra column called ‘Class’ is also added to the data frame. The column ‘Class’ contains 0 if a given launch is failed and 1 if it is successful.
- In the end, we end up with 90 rows or instances and 83 columns or features.

- GitHub URL of data wrangling notebook



EDA with Data Visualization

- Charts plotted:
  - 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 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 URL

EDA with SQL

Performed SQL queries:

- 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 v1.1Listing 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 payload mass greater than 4000 but less than 6000Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015Ranking 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 Url

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 its latitude and longitude coordinates as a start location.
- ❑Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts

Coloured Markers of the launch outcomes for each Launch Site:-

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

Distances between a Launch Site to its proximities:-

- ❑Added coloured 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 URL

Build a Dashboard with Plotly Dash

Launch Sites Dropdown List:-

- ❑ Added a dropdown list to enable Launch Site selection.

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:-

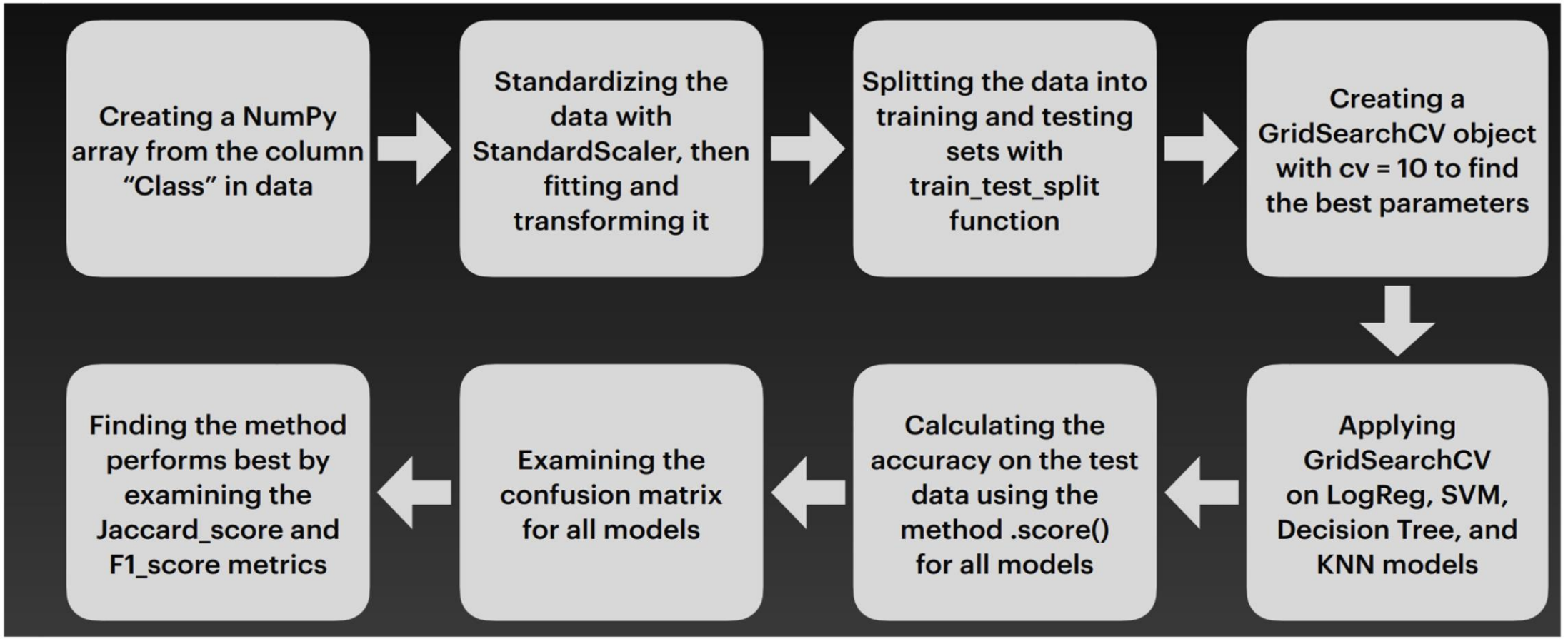
- ❑ Added a slider to select Payload range.

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.

- GitHub URL

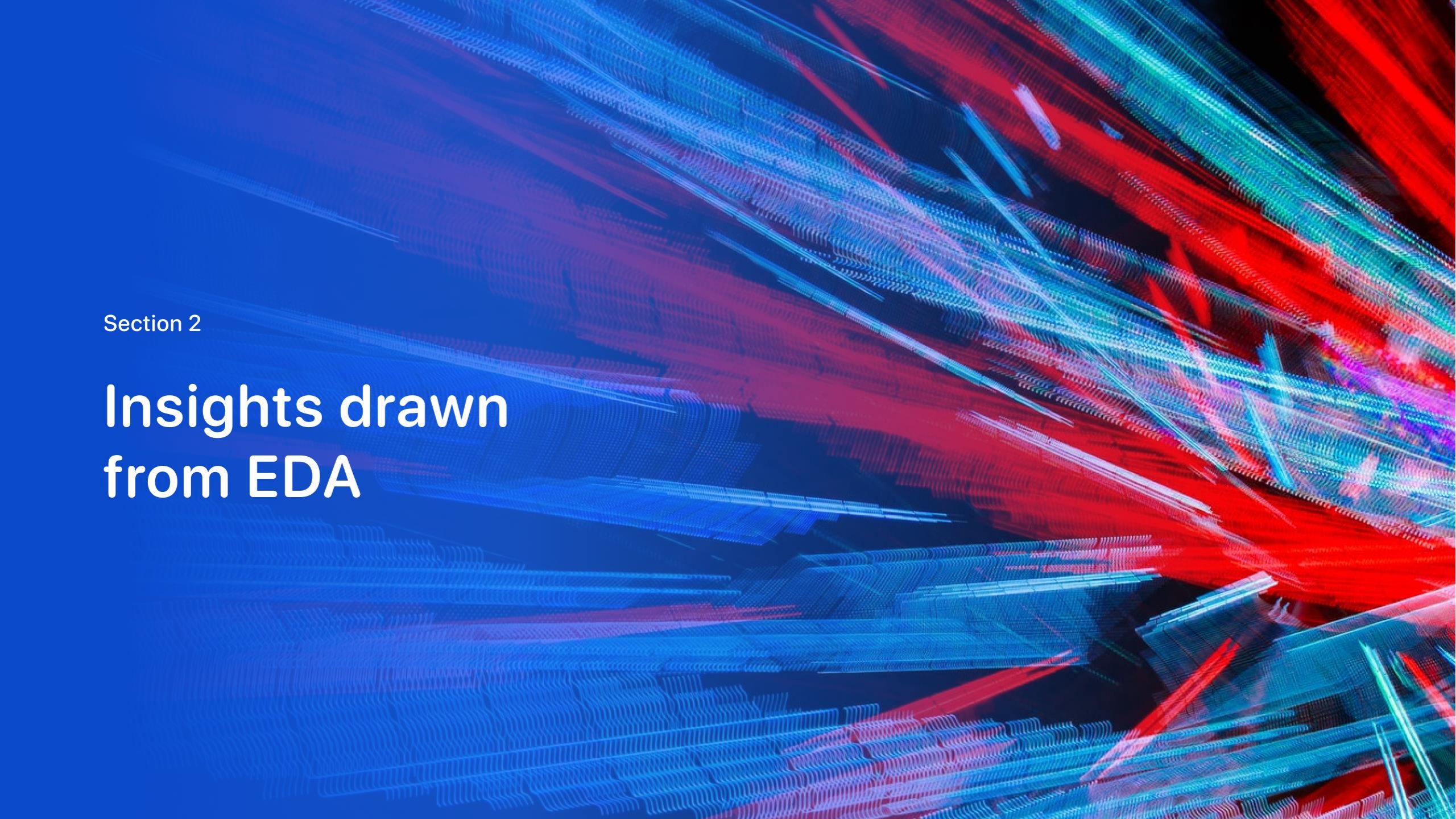
Predictive Analysis (Classification)



GitHub URL

Results

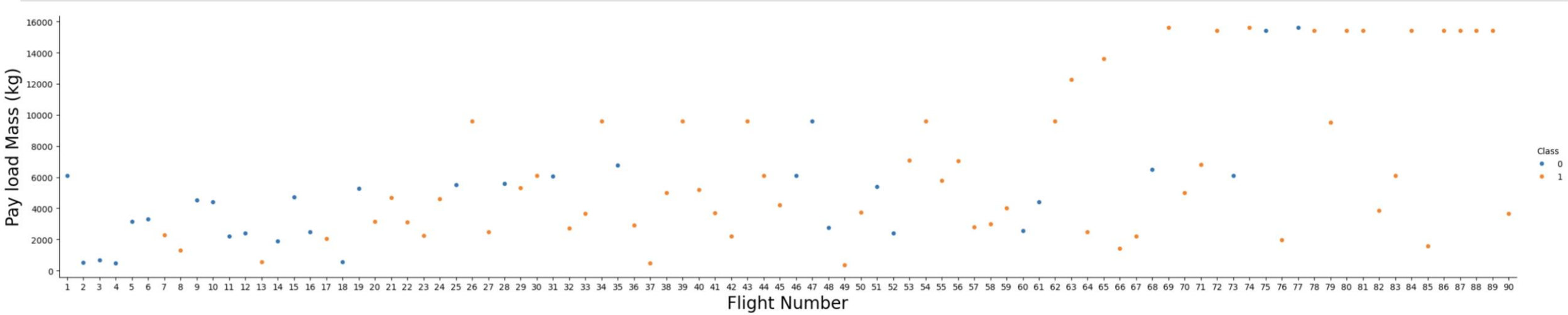
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



Section 2

Insights drawn from EDA

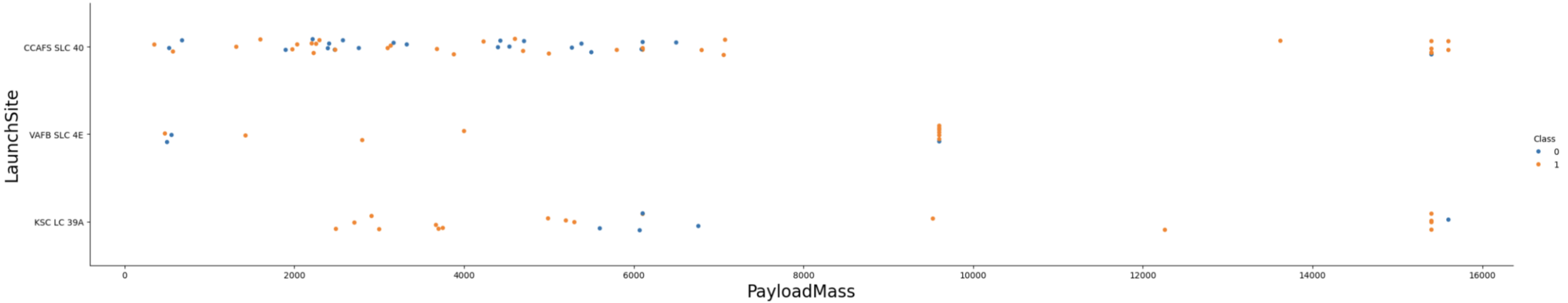
Flight Number vs. Launch Site



EXPLANATION:

- The earliest 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



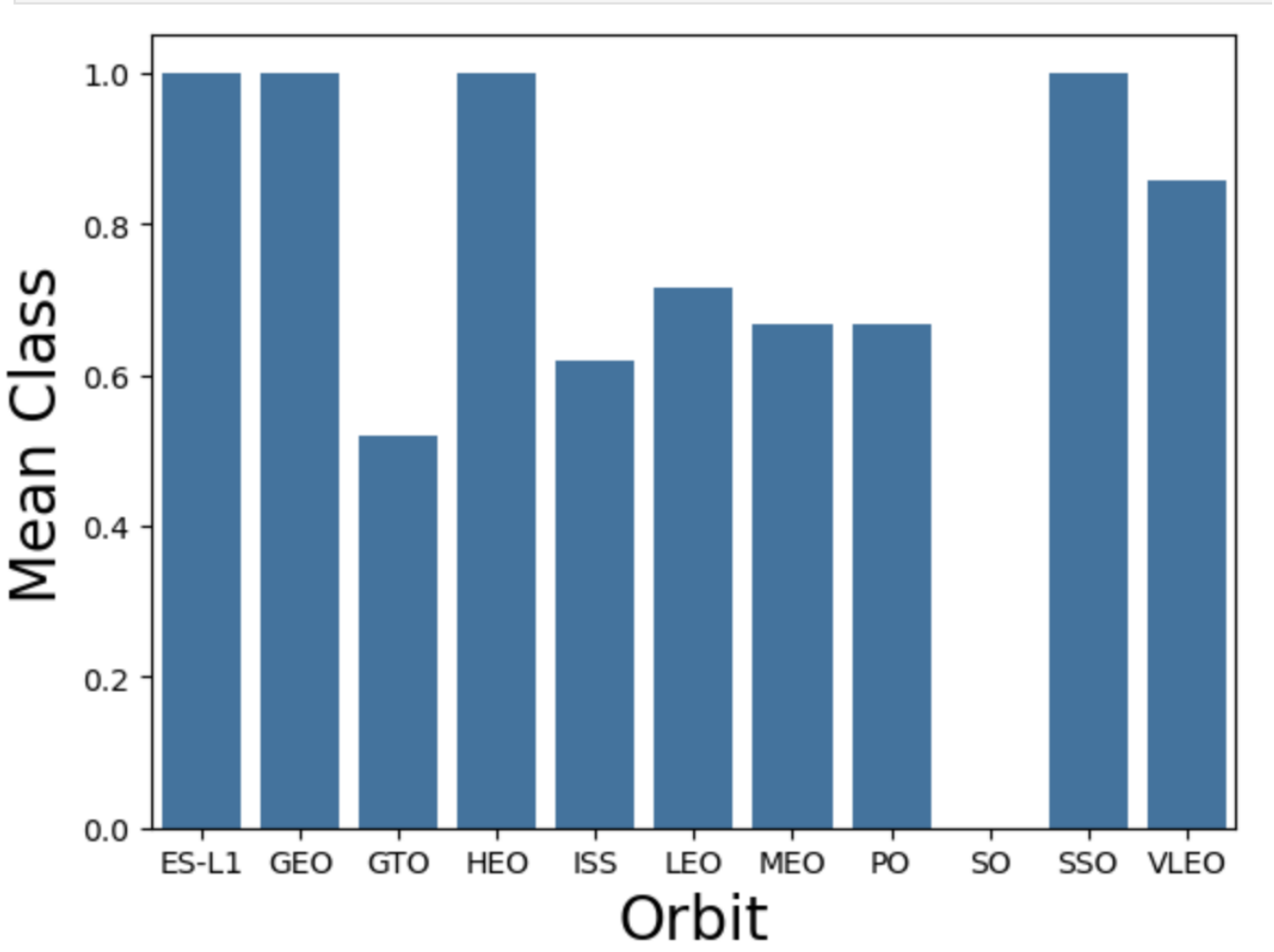
EXPLANATION:

- 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.

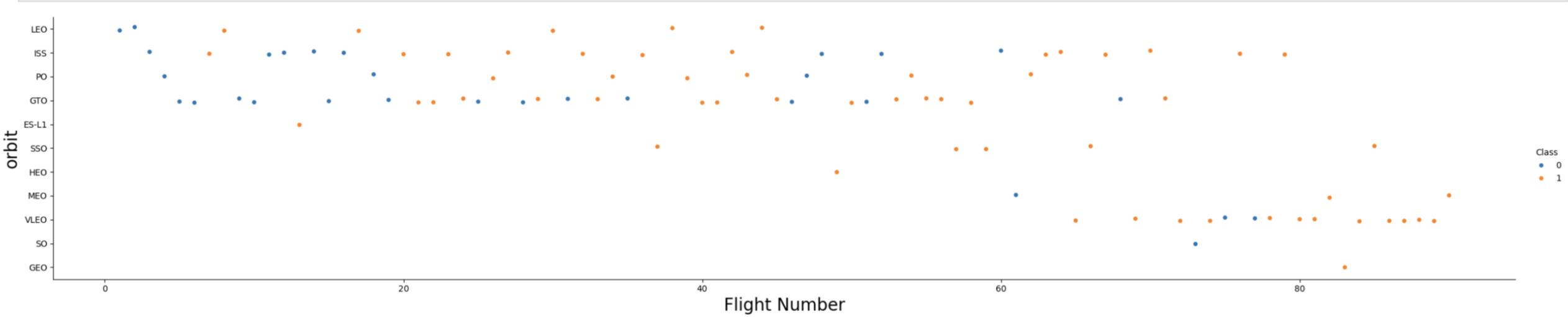
Success Rate vs. Orbit Type

EXPLANATION:

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



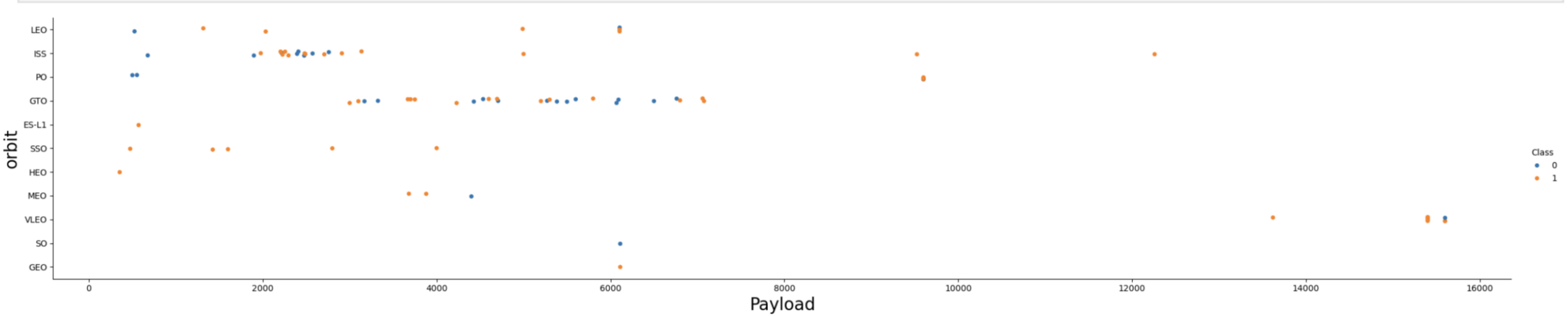
Flight Number vs. Orbit Type



EXPLANATION:

- 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



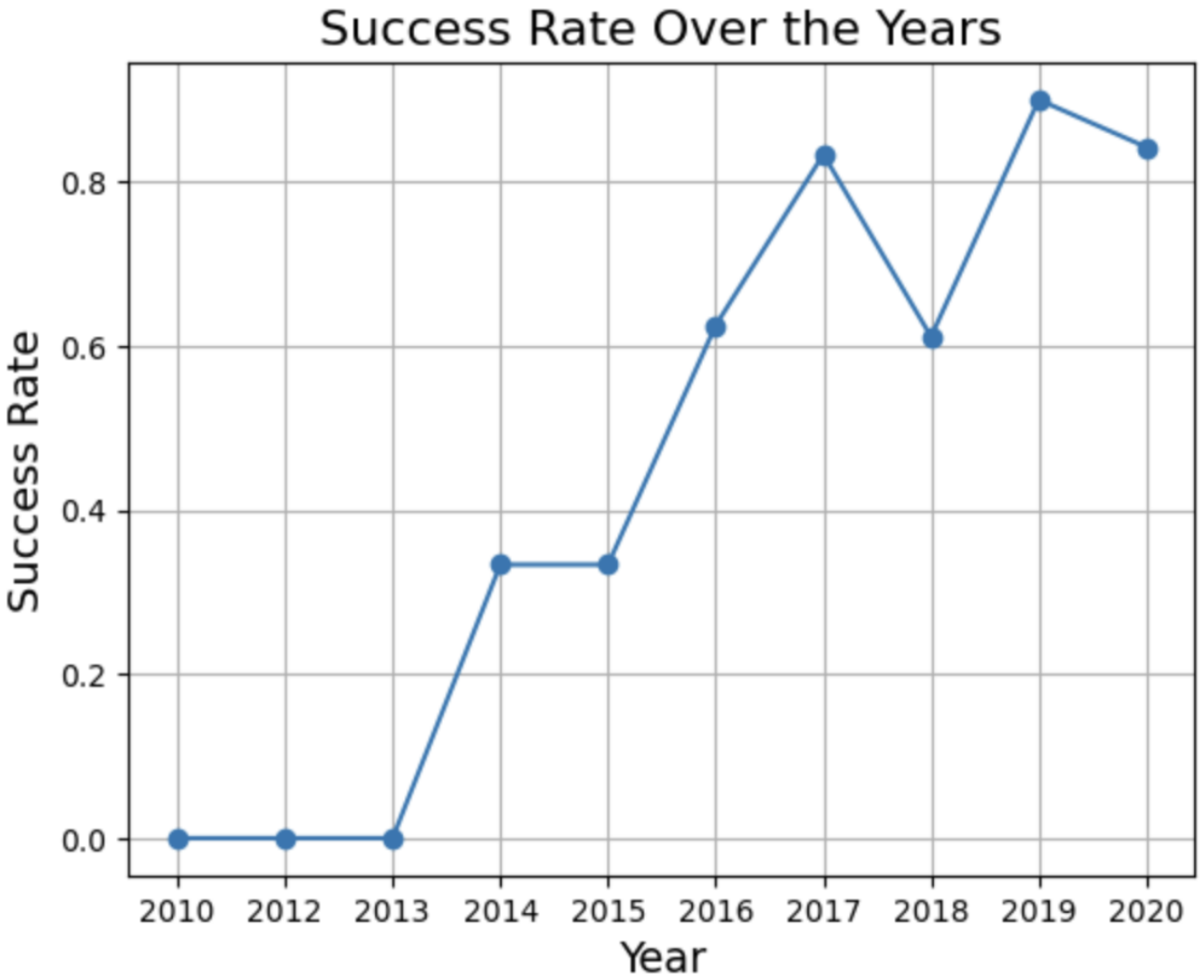
EXPLANATION:

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

Launch Success Yearly Trend

EXPLANATION:

- The success rate since 2013 kept increasing till 2020.



All Launch Site Names

- **Displaying 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

Launch Site Names Begin with 'CCA'

- **Displaying 5 records where launch sites begin with the string 'CCA'**

```
[10]: %sql select * from SPACEXTABLE 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_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

Total Payload Mass

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

```
]: %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

Average Payload Mass by F9 v1.1

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

```
%sql select AVG("PAYLOAD_MASS__KG_") from SPACEXTABLE where "Booster_Version" = "F9 v1.1"
```

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

AVG("PAYLOAD_MASS__KG_")
2928.4

First Successful Ground Landing Date

- **Listing the date when the first succesful landing outcome in ground pad was acheived.**

```
%sql select MIN("Date") AS "First Successful Landing Date"FROM SPACEXTABLE WHERE "Landing_Outcome"='Success (ground pad)'
```

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

First Successful Landing Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- **Listing 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" from SPACEXTABLE where "Landing_Outcome"="Success (drone ship)"
AND ("PAYLOAD_MASS__KG_" >=4000 AND "PAYLOAD_MASS__KG_"<=6000);
```

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

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

Total Number of Successful and Failure Mission Outcomes

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

```
%sql SELECT "Mission_Outcome", COUNT(*) AS outcome_count FROM SPACEXTABLE GROUP BY "Mission_Outcome";
```

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

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

Boosters Carried Maximum Payload

- **Listing 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

2015 Launch Records

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

```
%sql SELECT strftime('%m',"Date") AS Month_Number,"Landing_Outcome","Booster_Version","Launch_Site" FROM SPACEXTABLE
WHERE substr("Date", 0, 5) = '2015'
AND "Landing_Outcome" LIKE '%Failure%'  AND "Landing_Outcome" LIKE '%drone ship%';
```

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

Month_Number	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- **Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground**

```
%sql SELECT "Landing_Outcome", COUNT(*) AS outcome_count FROM SPACEXTABLE
WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY "Landing_Outcome" ORDER BY outcome_count DESC;
```

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

Landing_Outcome	outcome_count
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

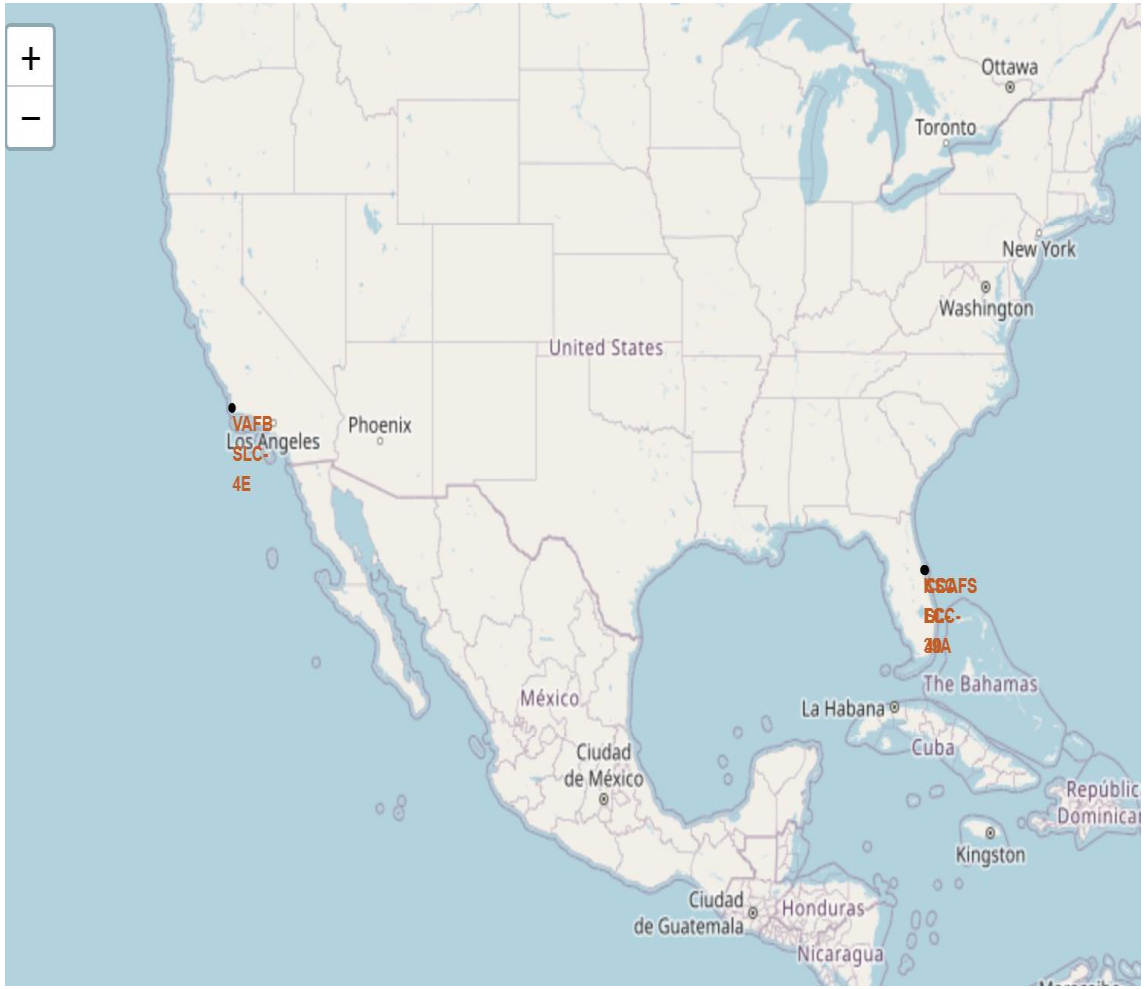
Section 3

Launch Sites Proximities Analysis

All Launch sites location markers on global map

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

- 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.

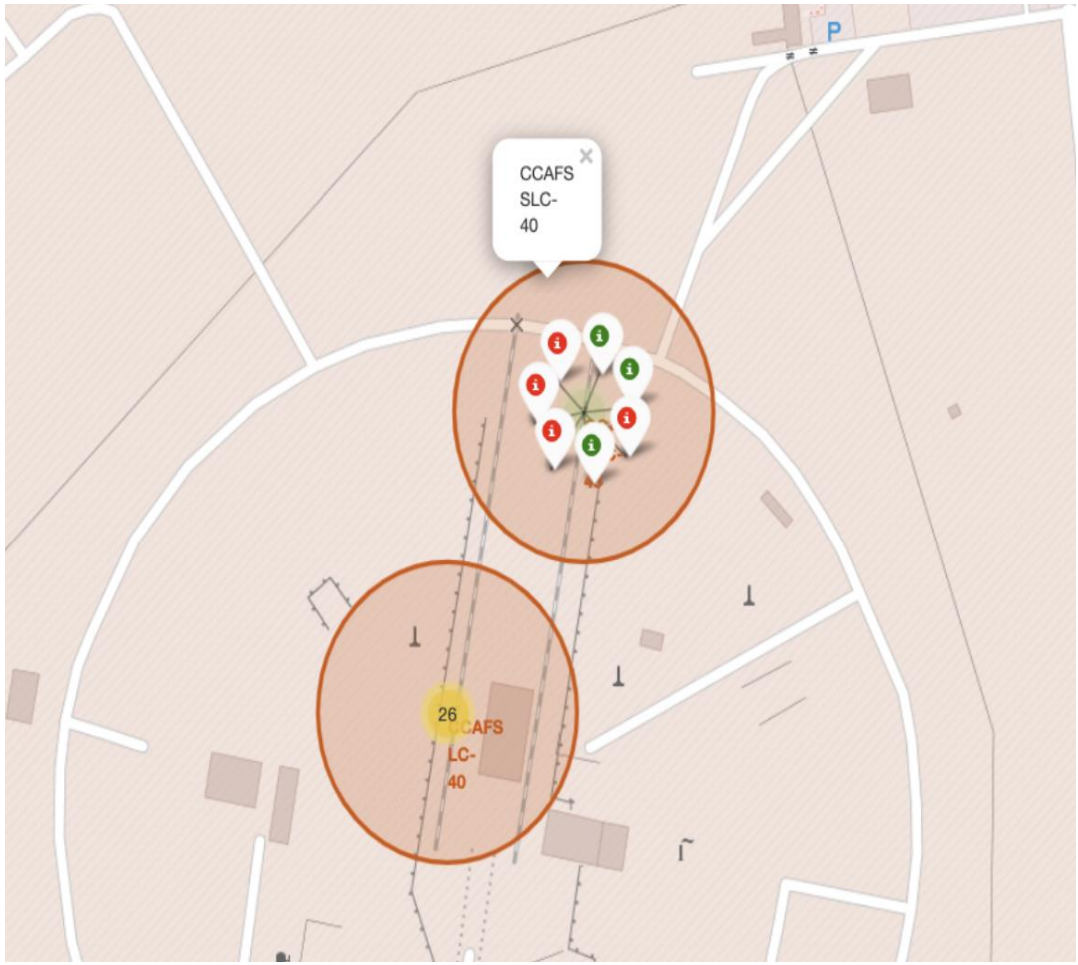


Coloured labelled launch records on map

• From the colour-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.

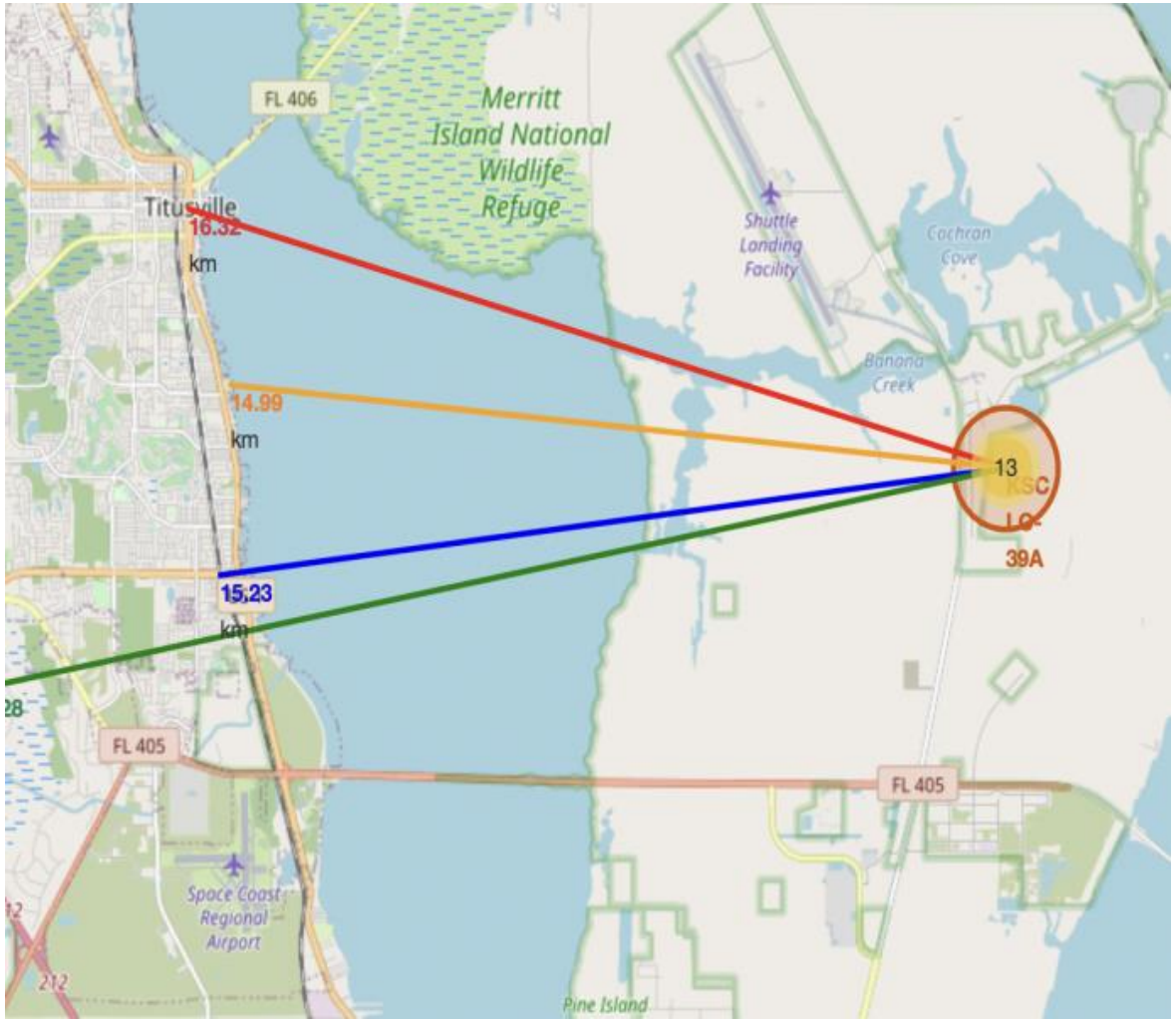


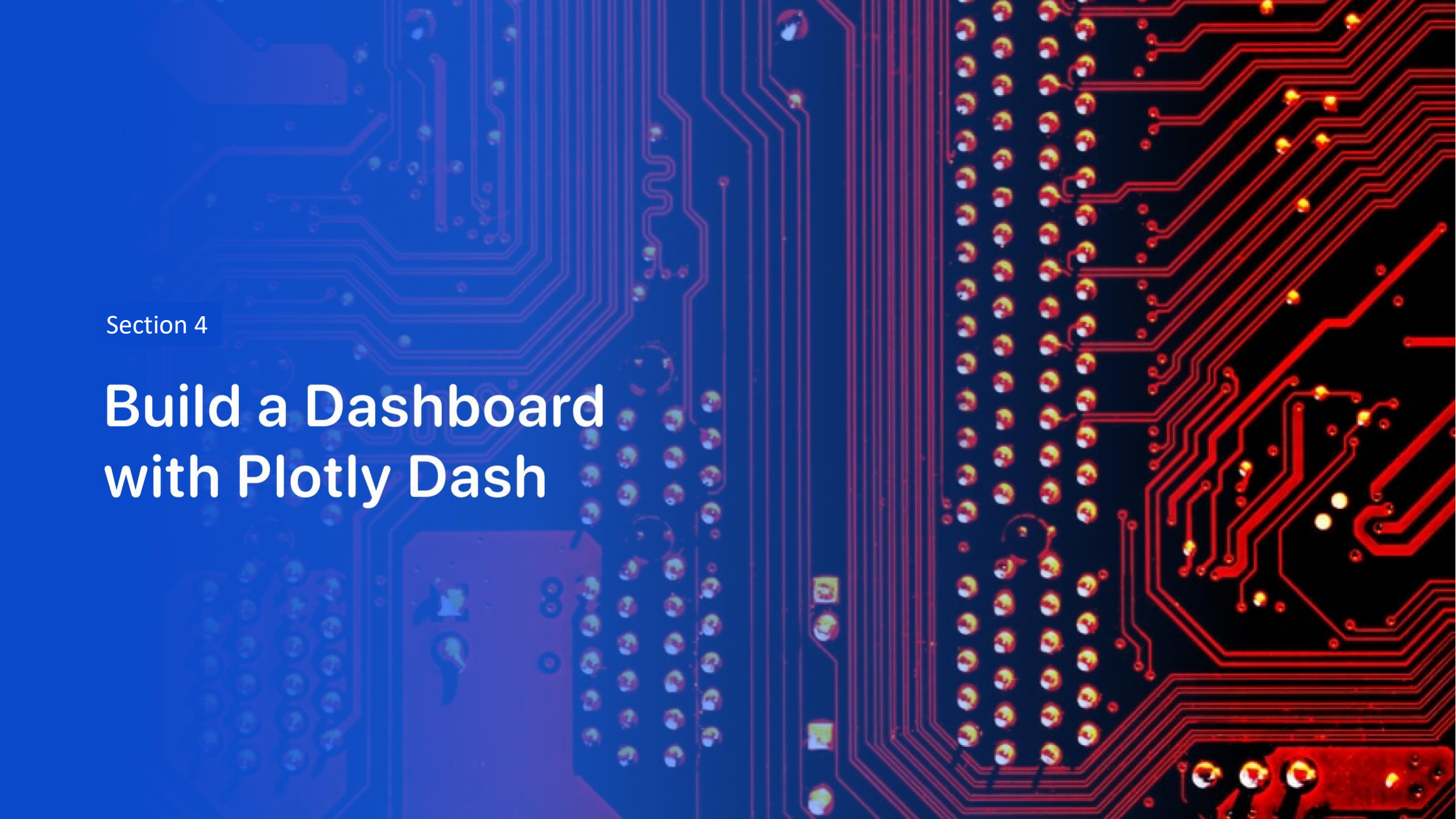
Distance from launch site to its proximities

- From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:
  - relative close to railway (15.23 km)
  - relative close to highway (20.28 km)
  - relative close to coastline (14.99 km)

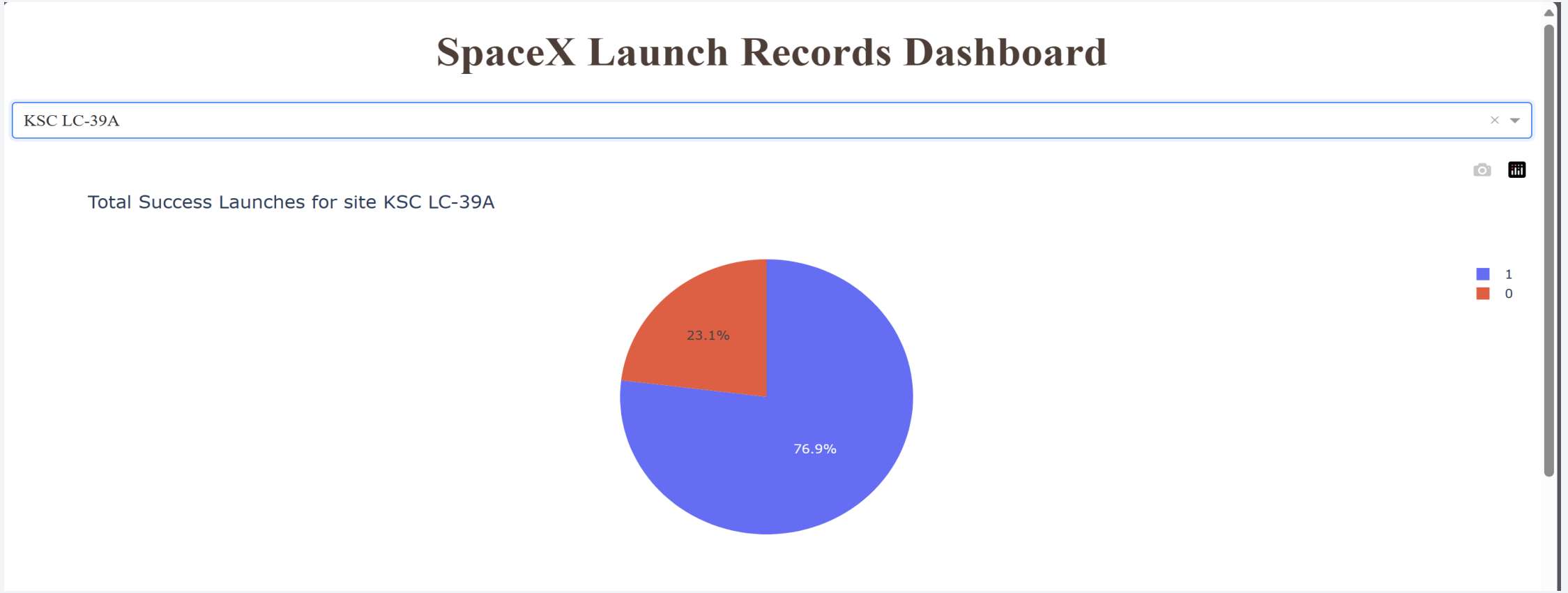
- Also the launch site KSC LC-39A is :
  - relative close to its closest city Titusville (16.32 km).

- Failed rocket with its high speed can cover distances like 15-20 km in few seconds. It could be potentially dangerous to populated areas.




Section 4
Build a Dashboard with Plotly Dash

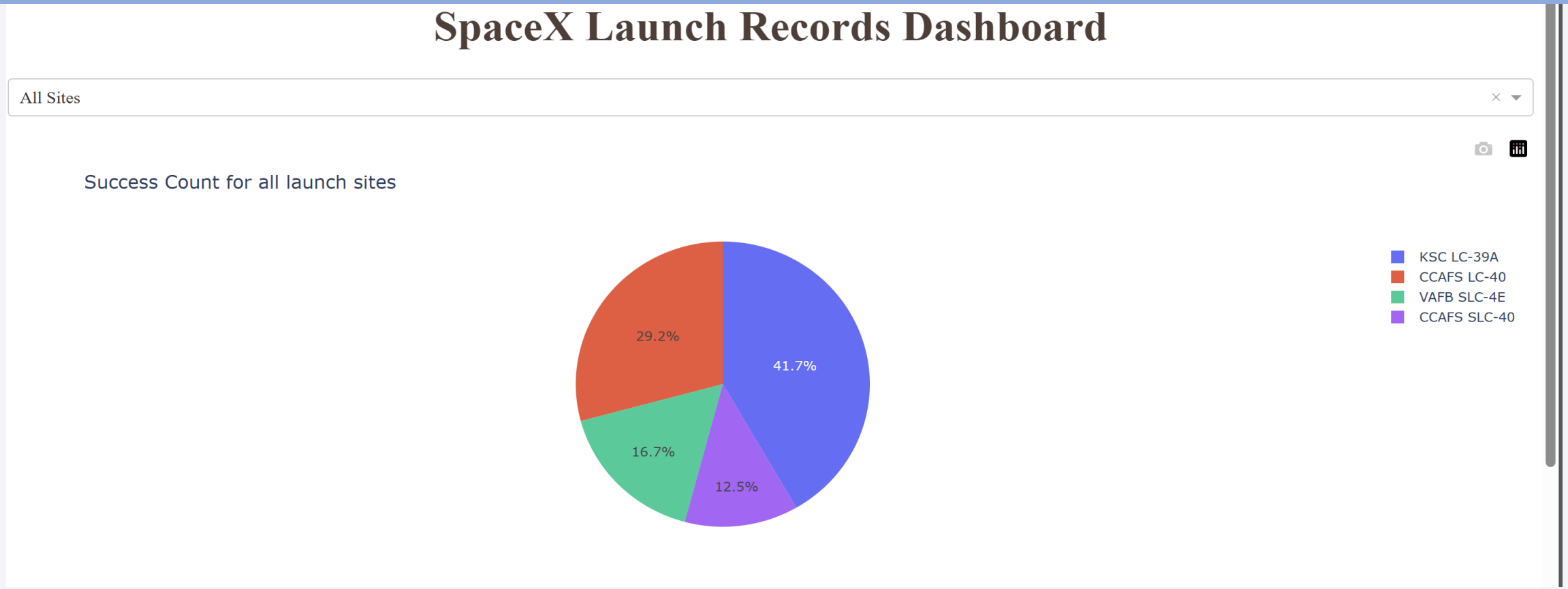
launch success count for all sites



Explanation:

- The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches.

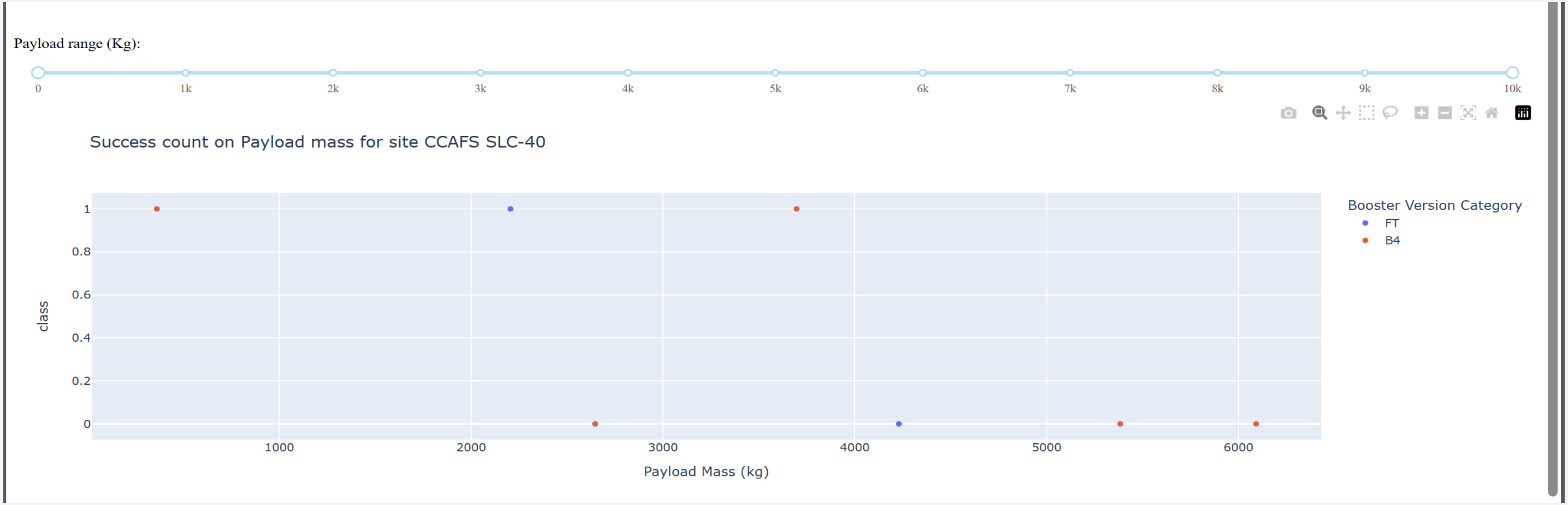
launch site with highest launch success ratio



Explanation:

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

launch site with highest launch success ratio



Explanation:

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

Section 5

Predictive Analysis (Classification)

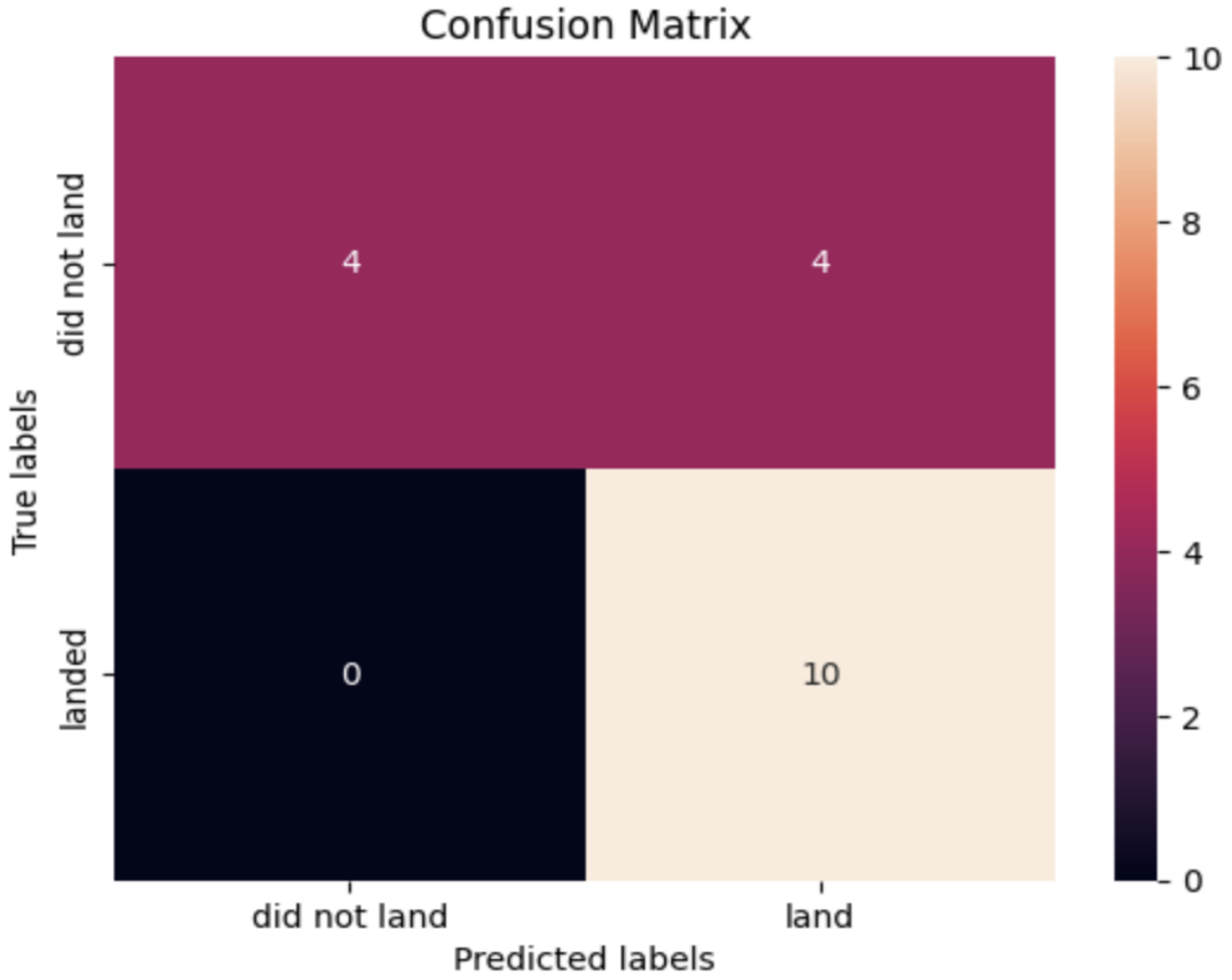
Classification Accuracy

- KNN has a highest cross-validation score (0.888) .A higher cross-validation score indicates better average performance across different folds during model evaluation.
- Although the Decision Tree has a slightly better test set score (0.7222 vs. 0.777), it's essential to prioritize cross-validation performance. The test set score may vary due to specific characteristics of the test data
- Therefore, KNN is the best model based on the scores.

	Logistic Reg.	SVM	Decision tree	KNN
Cross validation accuracy	0.8222	0.8222	0.8888	0.8444
Test data accuracy	0.7777	0.7777	0.7222	0.7777

Confusion Matrix

- The confusion matrix of the best performing model - KNN



Conclusions

- KNN Model is the best algorithm for this dataset.
- 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

- Special thank to:

IBM

Coursera

Instructors

Thank you!