IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Harshil Bhatnagar
17 Nov 2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection
  - Data wrangling
  - Exploratory Data Analysis (EDA) with Data Visualization
  - Exploratory Data Analysis (EDA) with SQL
  - Building an interactive map with Folium
  - Building a Dashboard with Plotly Dash
  - Predictive analysis (Classification)
- Summary of all results
  - Exploratory Data Analysis (EDA) results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Introduction

Project Background and Context

- SpaceX has emerged as a leader in the commercial space industry, revolutionizing space travel by significantly reducing costs. The company offers Falcon 9 rocket launches at a competitive price of 62 million dollars, compared to over 165 million dollars charged by other providers. A key factor in these savings is the reusability of the Falcon 9's first stage. Accurately predicting whether the first stage will successfully land is crucial for estimating launch costs and enhancing reusability strategies. By utilizing public data and machine learning models, this project aims to forecast the likelihood of successful landings of SpaceX's Falcon 9 first stage.

Key Questions to Address

- How do factors like payload mass, launch site, flight count, and orbital parameters influence the success rate of first stage landings?
- Has there been an improvement in the success rate of landings over time?
- Which machine learning algorithm performs best for predicting the binary outcome of landing success in this context?

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

Methodology

Executive Summary

- 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

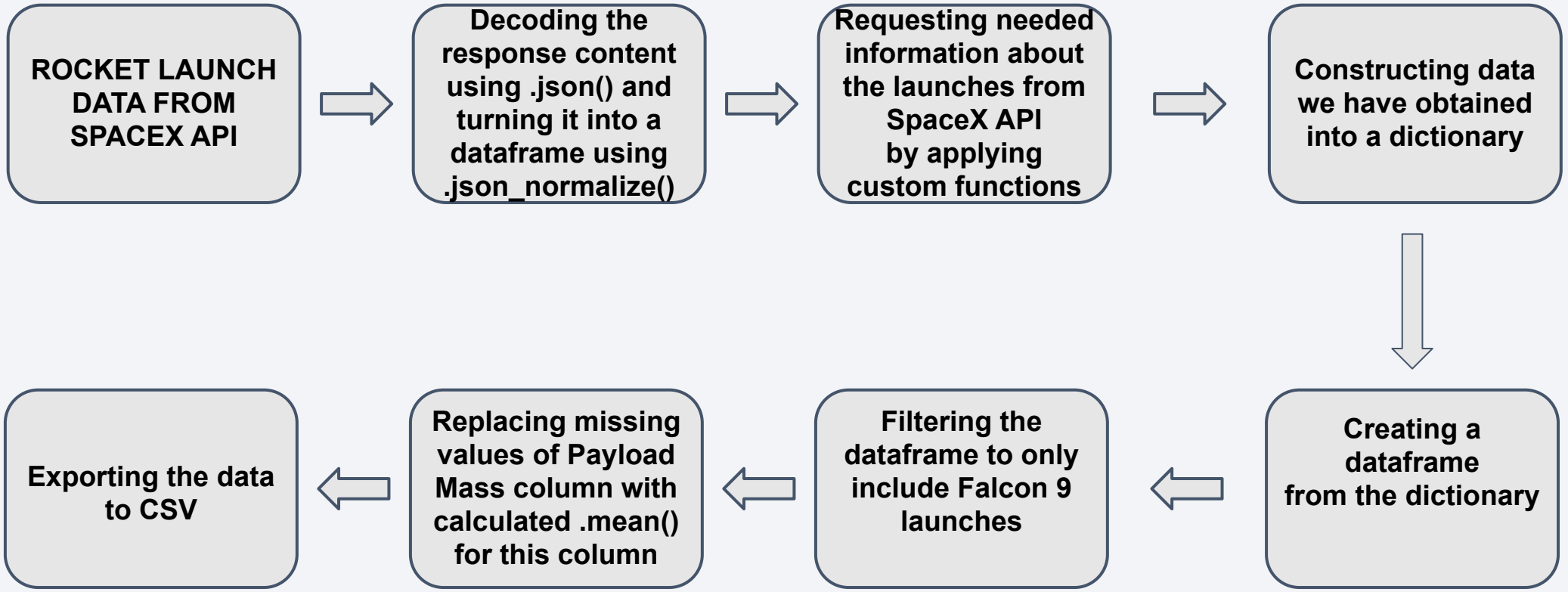
Data Collection

- The data collection process utilized a combination of API requests from the SpaceX REST API and web scraping from a Wikipedia table listing SpaceX launches. Both methods were necessary to gather comprehensive and detailed information about the launches, ensuring no critical data was missed for subsequent analysis.
- Data Columns obtained via SpaceX REST API:

  FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- Data Columns obtained via Web Scraping from Wikipedia:

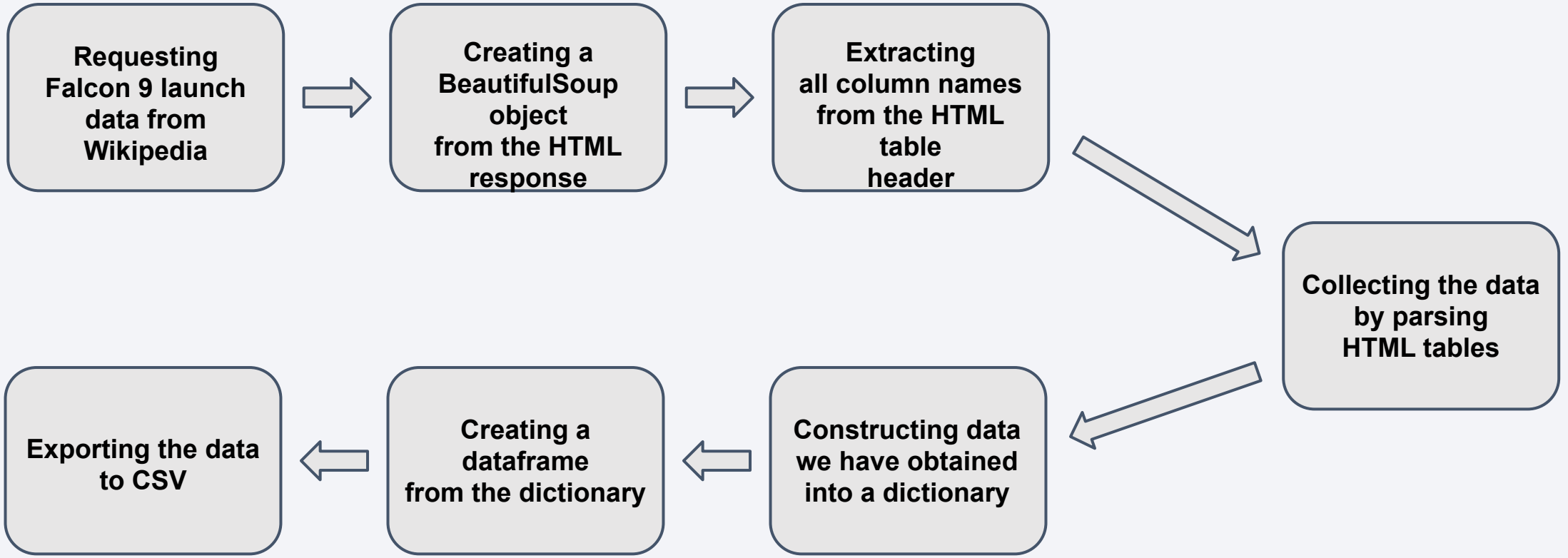
  Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time
- This dual approach provided a complete dataset for in-depth analysis of SpaceX’s launch history and performance.

Data Collection – SpaceX API



GITHUB LINK:
https://github.com/HarshilBhatnagar/IBM-DATA-SCIENCE-CAPSTONE

Data Collection – Scraping



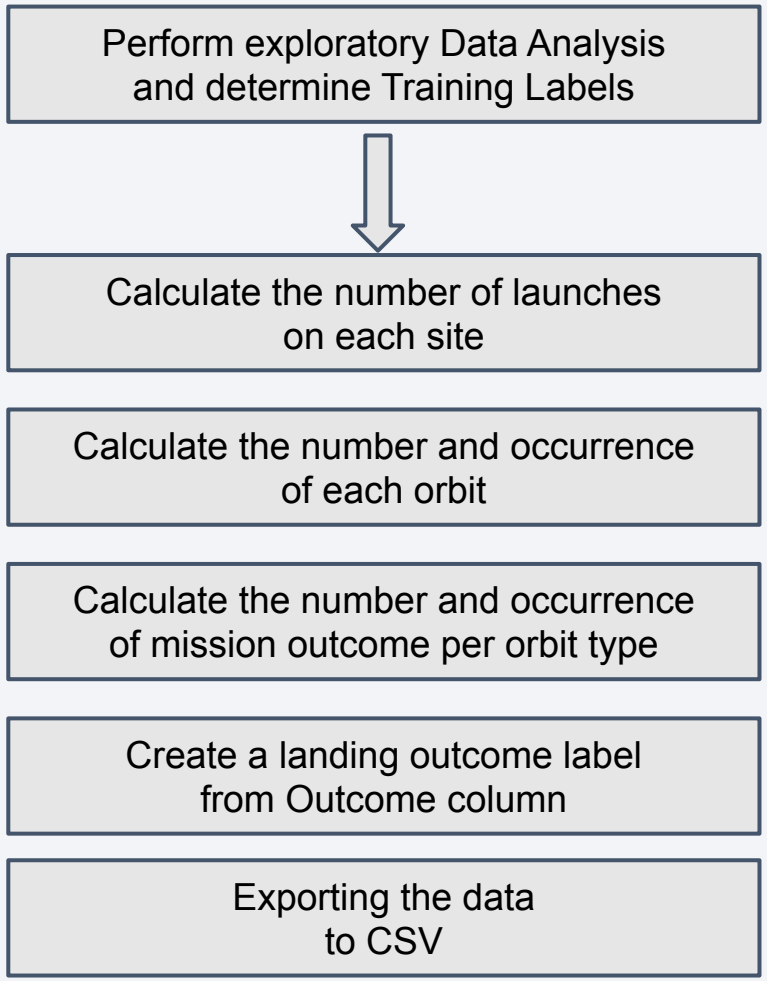
GITHUB LINK:
https://github.com/HarshilBhatnagar/IBM-DATA-SCIENCE-CAPSTONE

Data Wrangling

In the dataset, there are several scenarios where the Falcon 9 booster did not land successfully. Sometimes the landing attempt failed due to accidents or unfavorable conditions. For instance:

- True Ocean: The booster successfully landed in a designated ocean area.
- False Ocean: The landing attempt in the ocean was unsuccessful.
- True RTLS: The booster successfully returned and landed on a ground pad (Return to Launch Site).
- False RTLS: The landing attempt on a ground pad was unsuccessful.
- True ASDS: The booster successfully landed on a drone ship (Autonomous Spaceport Drone Ship).
- False ASDS: The landing attempt on a drone ship was unsuccessful.

These different outcomes were consolidated into binary labels for model training: 1 indicates a successful landing, while 0 indicates a failure.



GITHUB LINK:
https://github.com/HarshilBhatnagar/IBM-DATA-SCIENCE-CAPSTONE

EDA with Data Visualization

The following charts were created to explore the relationships in the data:

- Scatter Plots: Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Flight Number vs. Orbit Type, and Payload Mass vs. Orbit Type.
  - Scatter plots help identify potential relationships between variables, which can be useful for feature selection in machine learning models.
- Bar Charts: Orbit Type vs. Success Rate.
  - Bar charts provide a comparison across discrete categories, highlighting the differences in success rates for various orbit types.
- Line Charts: Yearly Trend of Success Rate.
  - Line charts display changes in success rates over time, revealing any trends or improvements in launch performance.

GITHUB LINK:
https://github.com/HarshilBhatnagar/IBM-DATA-SCIENCE-CAPSTONE

EDA with SQL

SQL Queries Performed

- Retrieved the unique launch site names from the dataset.
- Displayed 5 records where the launch site names begin with 'CCA'.
- Calculated the total payload mass carried by boosters launched by NASA (CRS missions).
- Computed the average payload mass for the booster version F9 v1.1.
- Identified the date of the first successful ground pad landing.
- Listed booster names that successfully landed on a drone ship with payload mass between 4000 and 6000 kg.
- Counted the total number of successful and failed mission outcomes.
- Retrieved the booster versions that carried the maximum payload mass.
- Listed the failed drone ship landings, along with booster versions and launch sites, for the year 2015.
- Ranked the count of landing outcomes (e.g., Failure on drone ship, Success on ground pad) between 2010-06-04 and 2017-03-20 in descending order.

Build an Interactive Map with Folium

Markers of All Launch Sites:

- Added a marker with a circle, popup label, and text label for NASA Johnson Space Center using its latitude and longitude as the starting location.
- Placed markers with circles, popup labels, and text labels for all launch sites, indicating their geographical locations and proximity to the equator and coastlines.

Colored Markers for Launch Outcomes:

- Added colored markers to indicate launch outcomes: green for successful launches and red for failed launches.
- Used a marker cluster to easily identify launch sites with higher success rates.

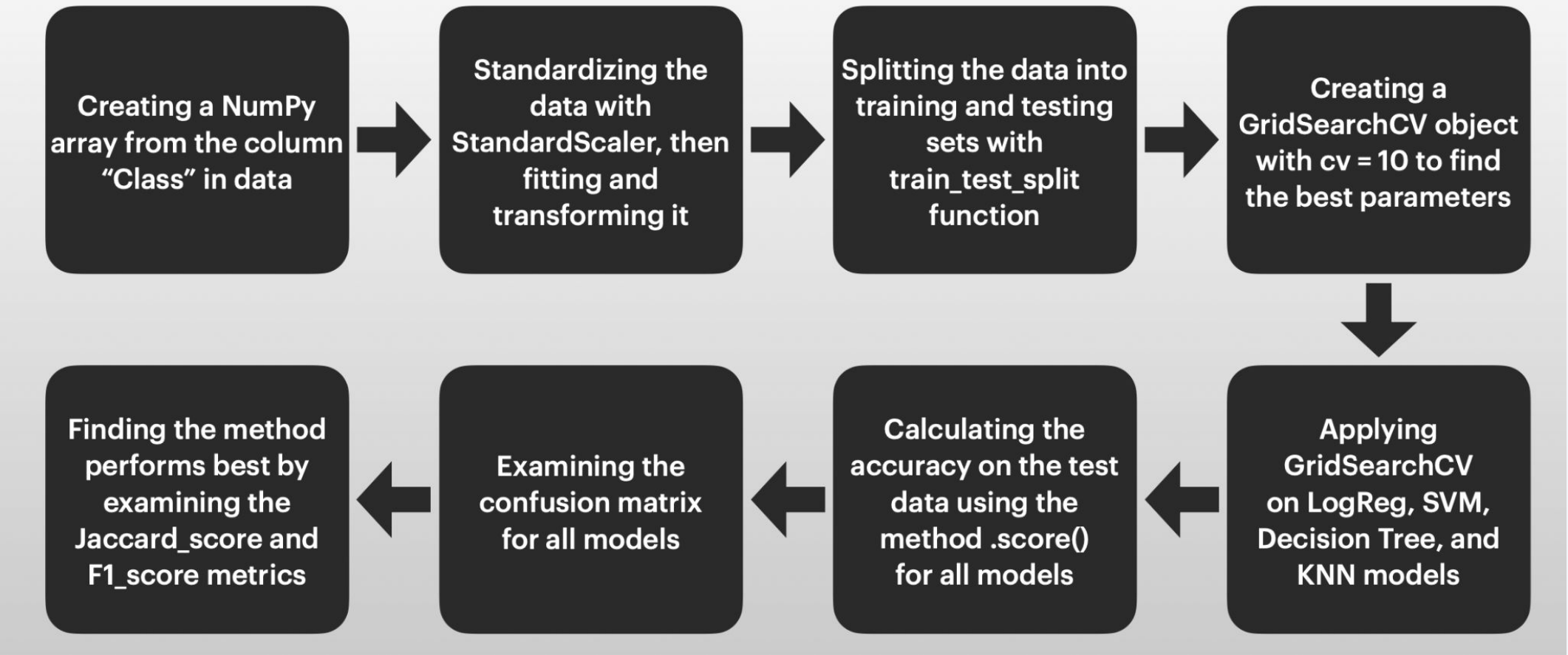
Distances from Launch Sites to Nearby Features:

- Added colored lines to display the distances between the KSC LC-39A launch site and nearby features, such as the railway, highway, coastline, and closest city.

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.

Predictive Analysis (Classification)

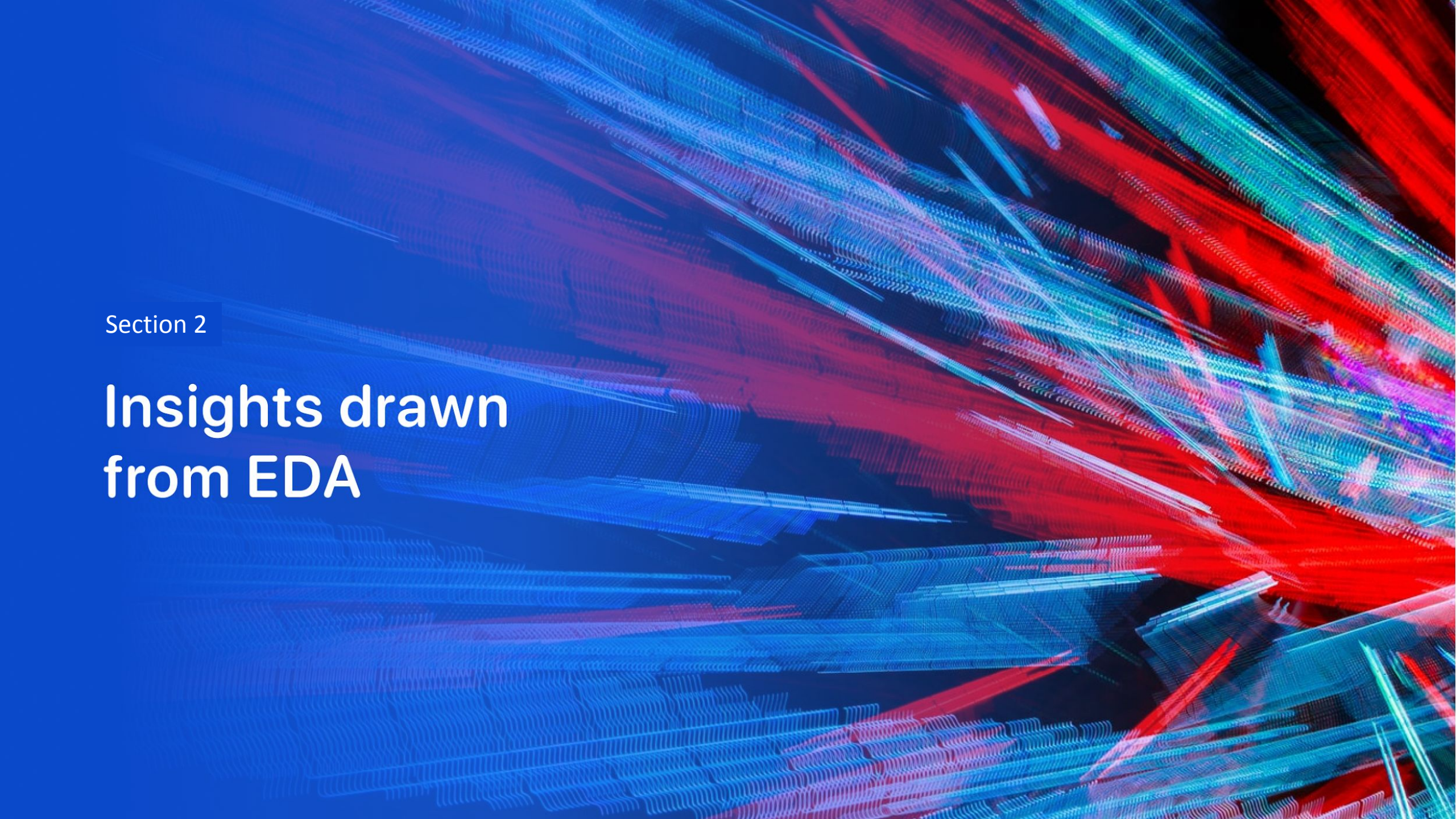


Results

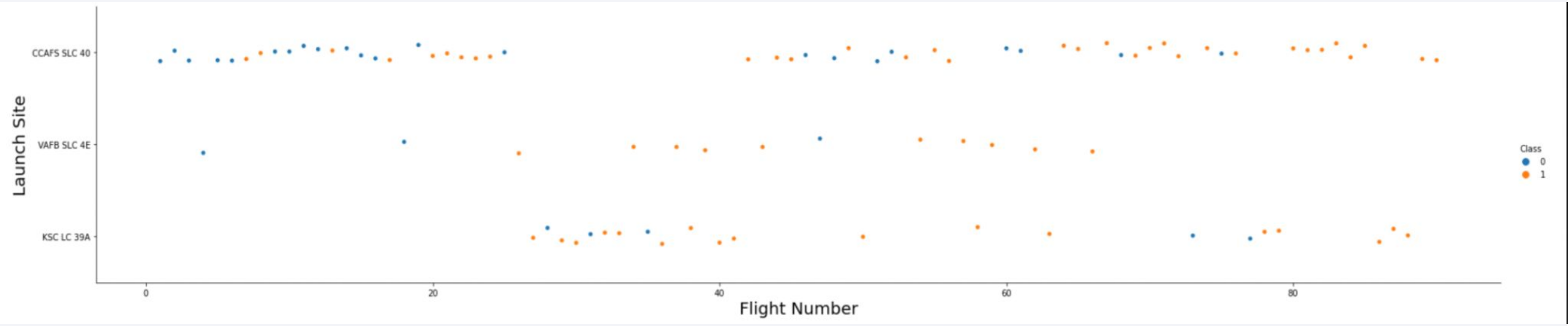
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA



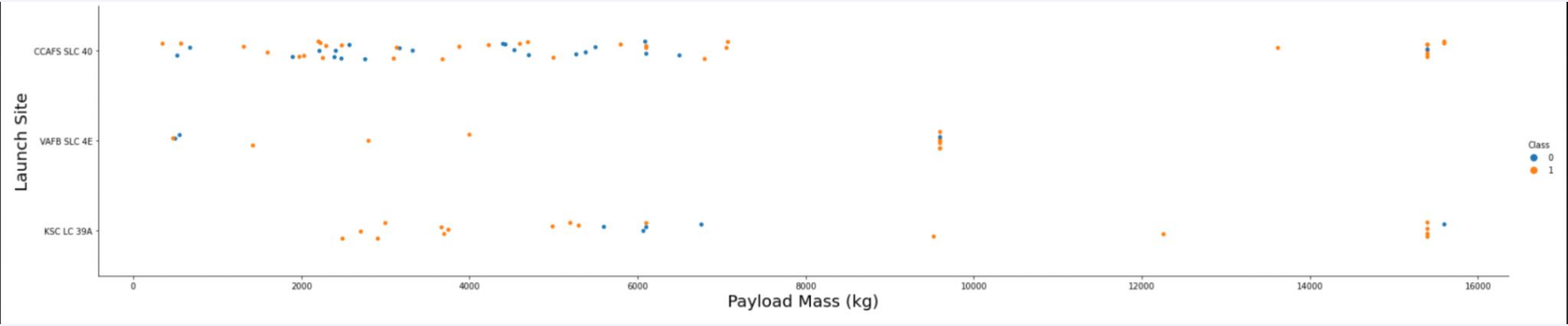
Flight Number vs. Launch Site



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



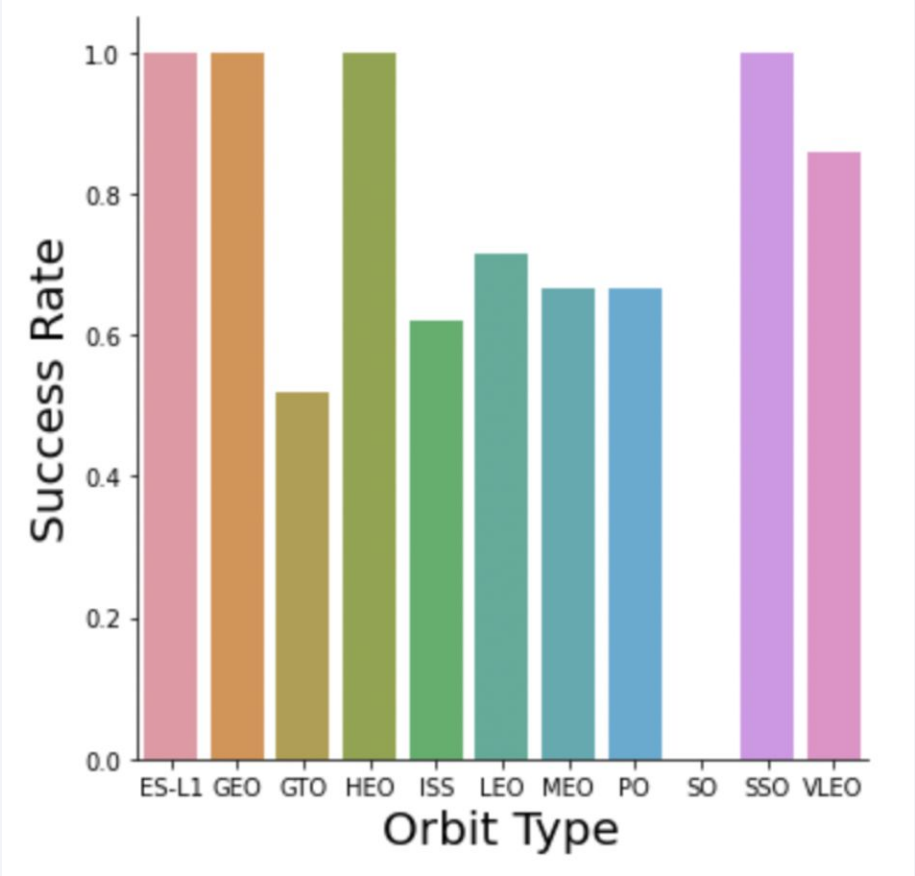
Explanation:

- 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.

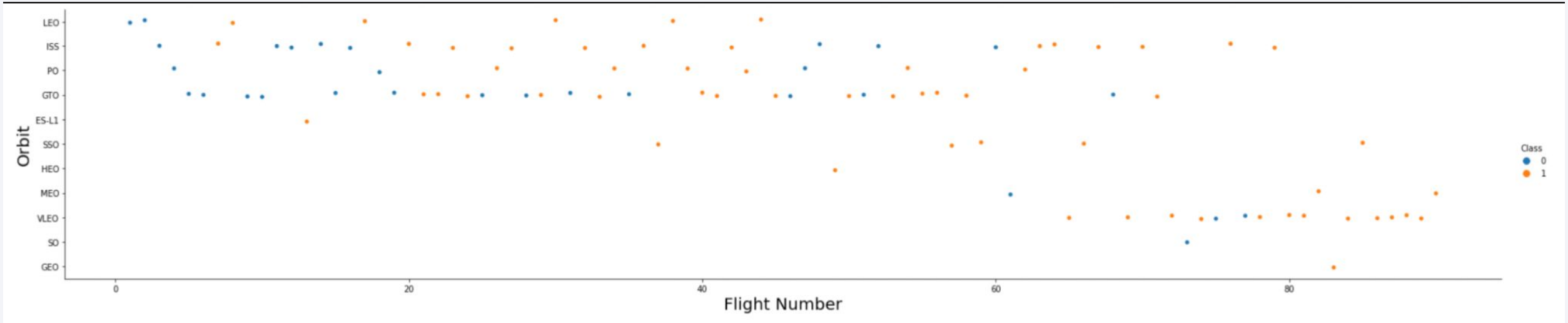
Success Rate vs. Orbit Type

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



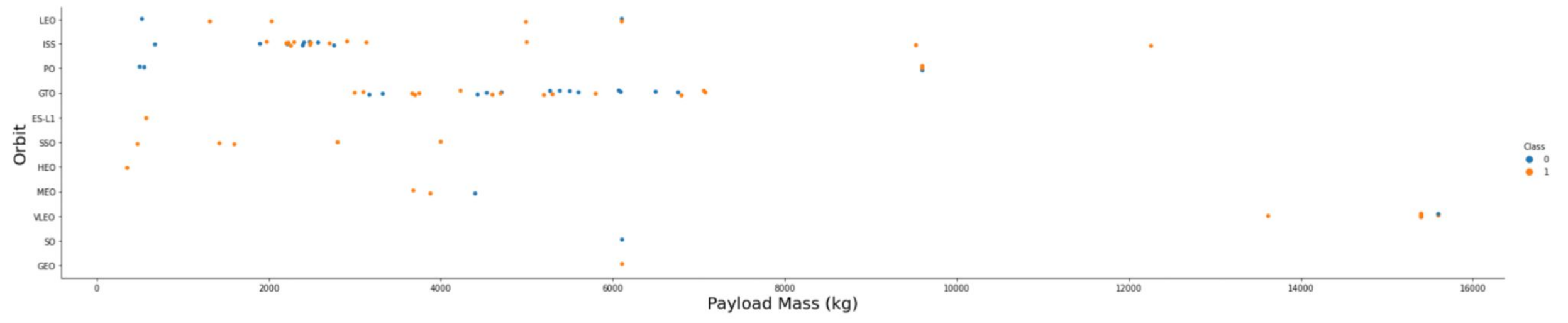
Flight Number vs. Orbit Type



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



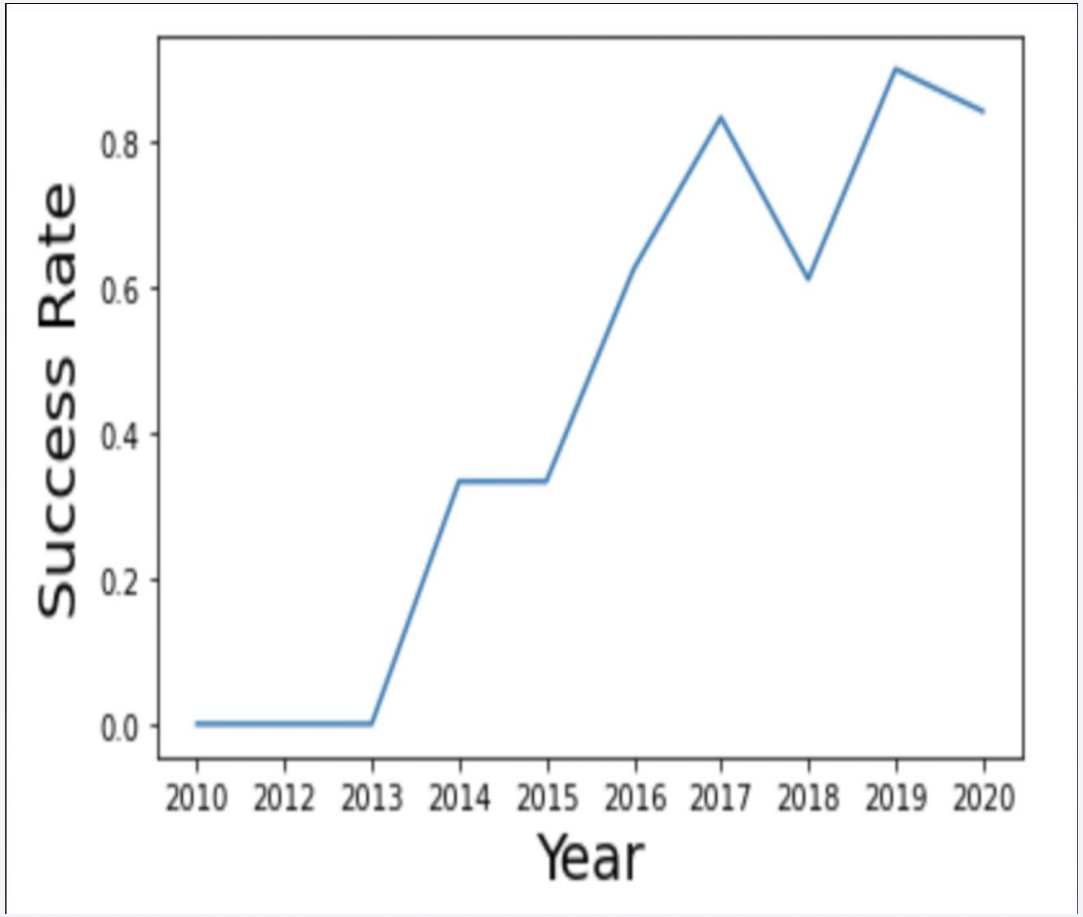
Explanation:

- 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

Explanation:

- Displaying 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

Launch Site Names Begin with 'CCA'

Explanation:

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

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

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

Out[5]:

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

Explanation:

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

```
In [6]: %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[6]:

total_payload_mass
45596

Average Payload Mass by F9 v1.1

Explanation:

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

```
In [7]: %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%';
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[7]:

average_payload_mass
2534

First Successful Ground Landing Date

Explanation:

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

```
In [8]: %sql select min(date) as first_successful_landing from SPACEXDATASET where landing__outcome = 'Success (ground pad)';
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[8]:

first_successful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Explanation:

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

In [9]:
```
%sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4
000 and 6000;
```

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

Out[9]:

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

Total Number of Successful and Failure Mission Outcomes

Explanation:

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

```
In [10]: %sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[10]:

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

Explanation:

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

```
In [11]: %sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET);
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[11]:
```

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

Explanation:

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

```
In [12]: %%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET
             where landing__outcome = 'Failure (drone ship)' and year(date)=2015;
```

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

Out[12]:

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

Explanation:

- Ranking 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.

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.
```

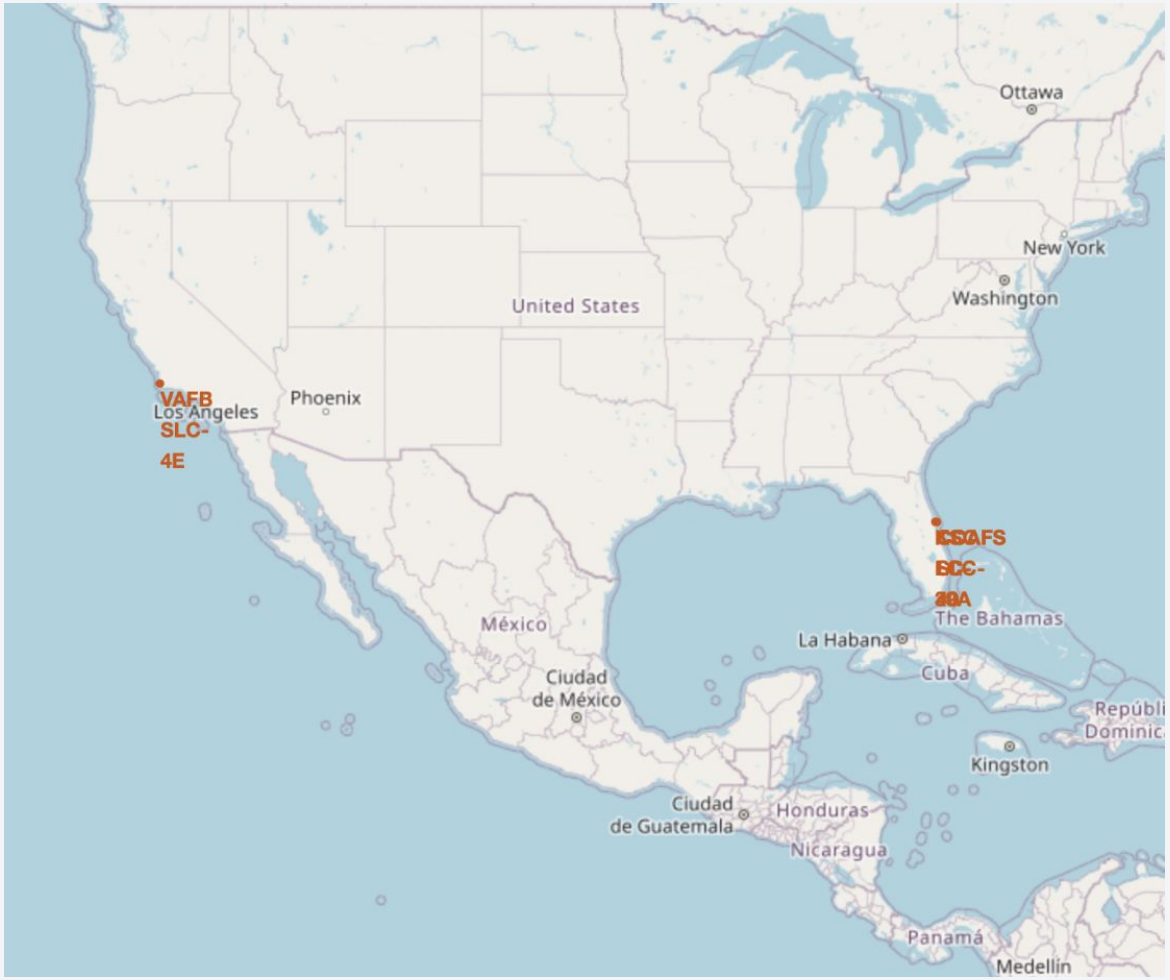
Out[13]:

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


Section 3
Launch Sites
Proximities Analysis

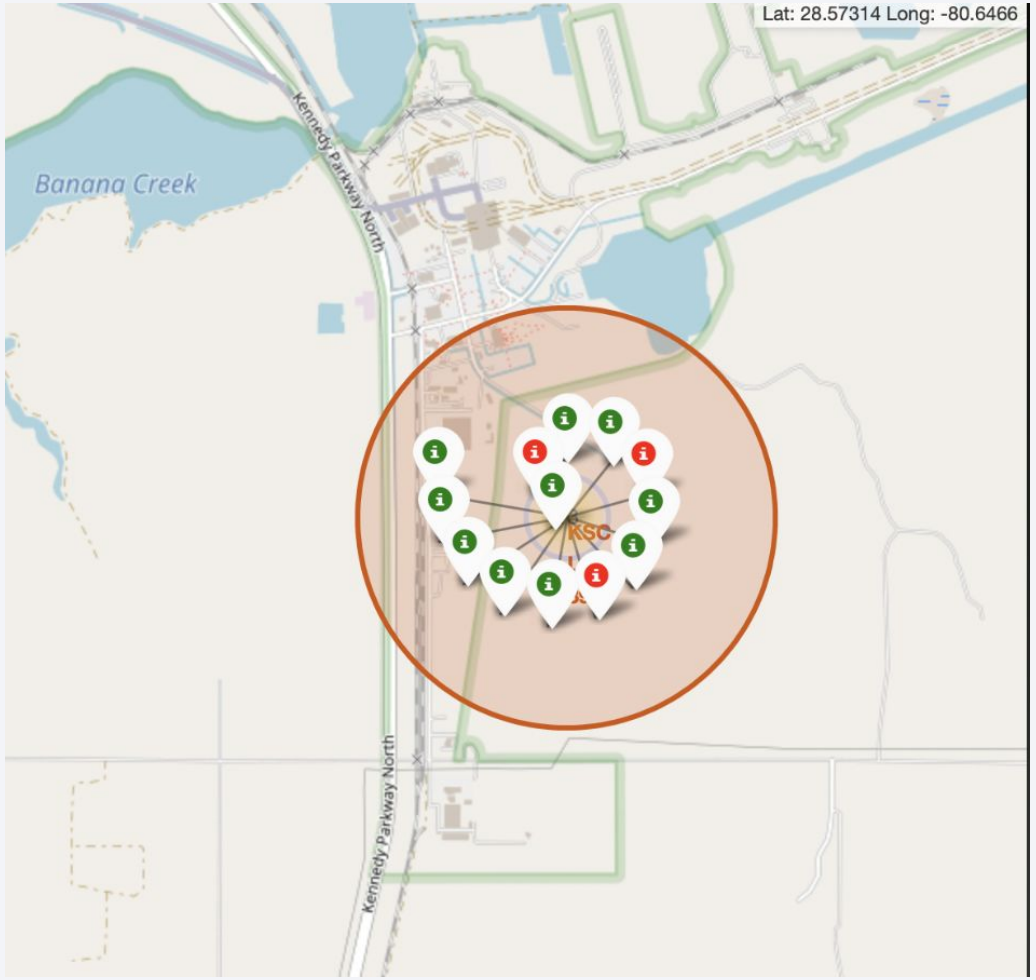
All launch sites' location markers on a global map

- 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 minimises the risk of having any debris dropping or exploding near people.



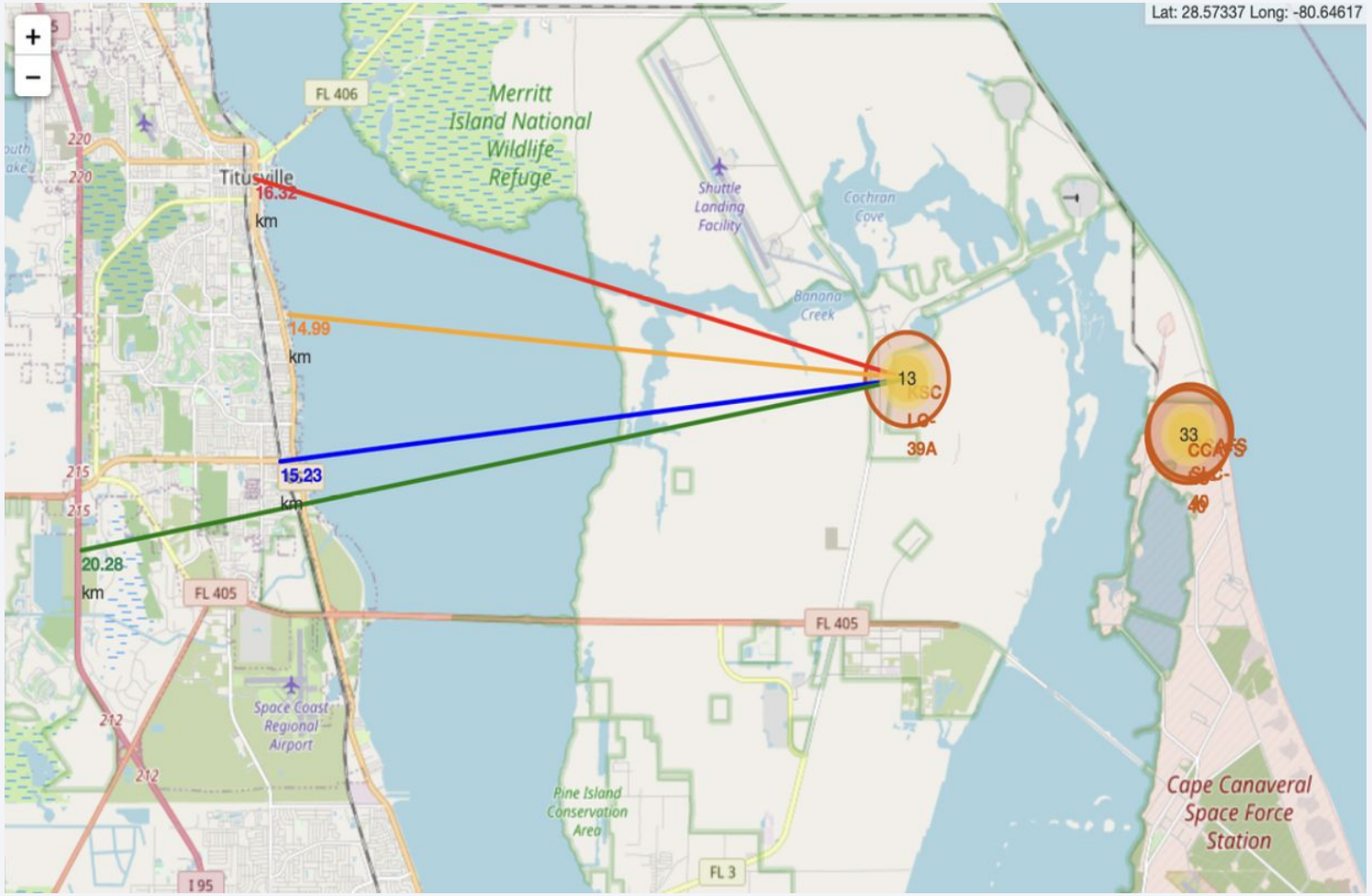
Colour-labeled launch records on the map

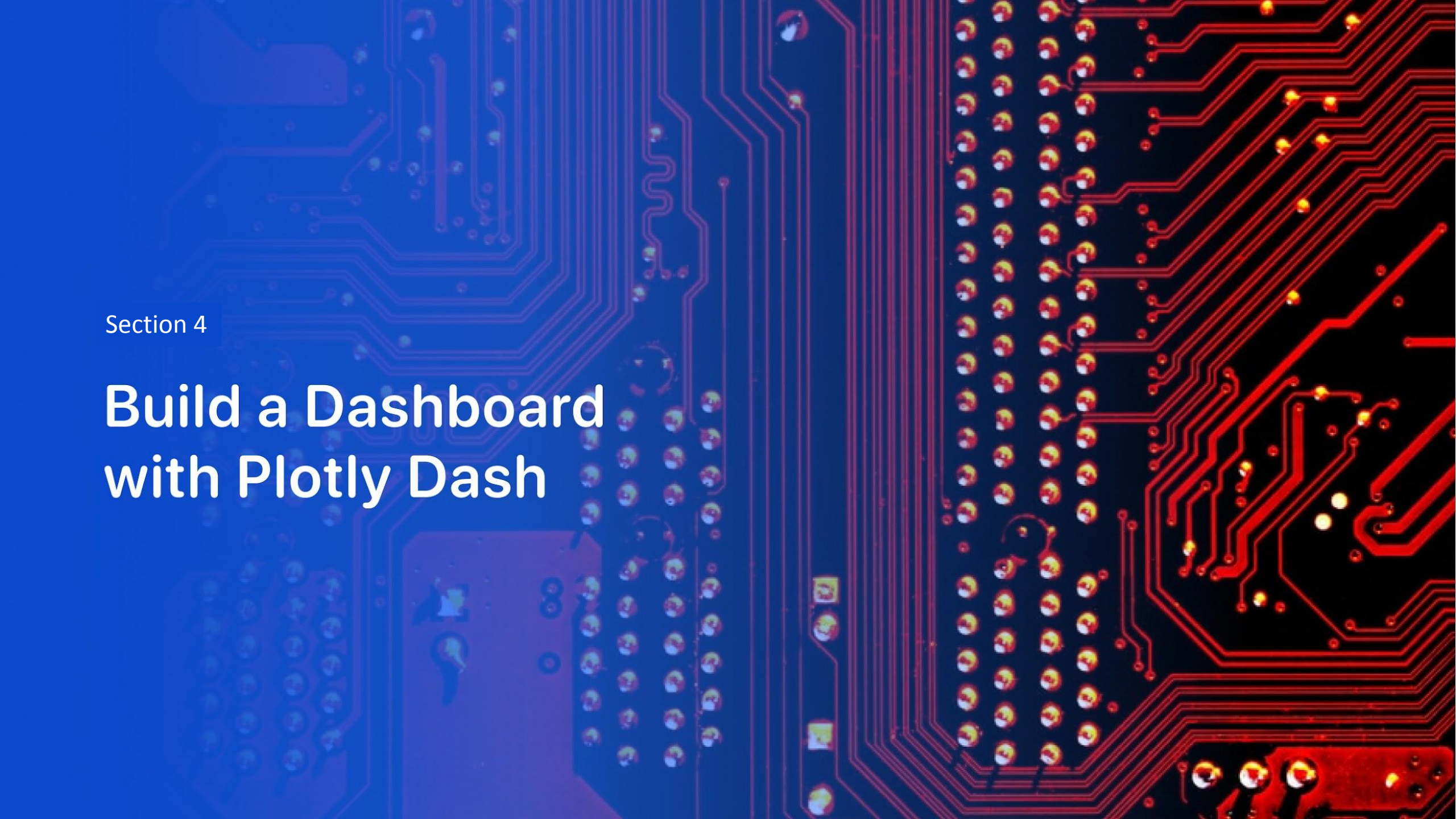
- From the colour-labeled markers we should be able to easily identify which launch sites have relatively high success rates.
  - Green Marker = Successful Launch
  - Red Marker = Failed Launch
- Launch Site KSC LC-39A has a very high Success Rate.



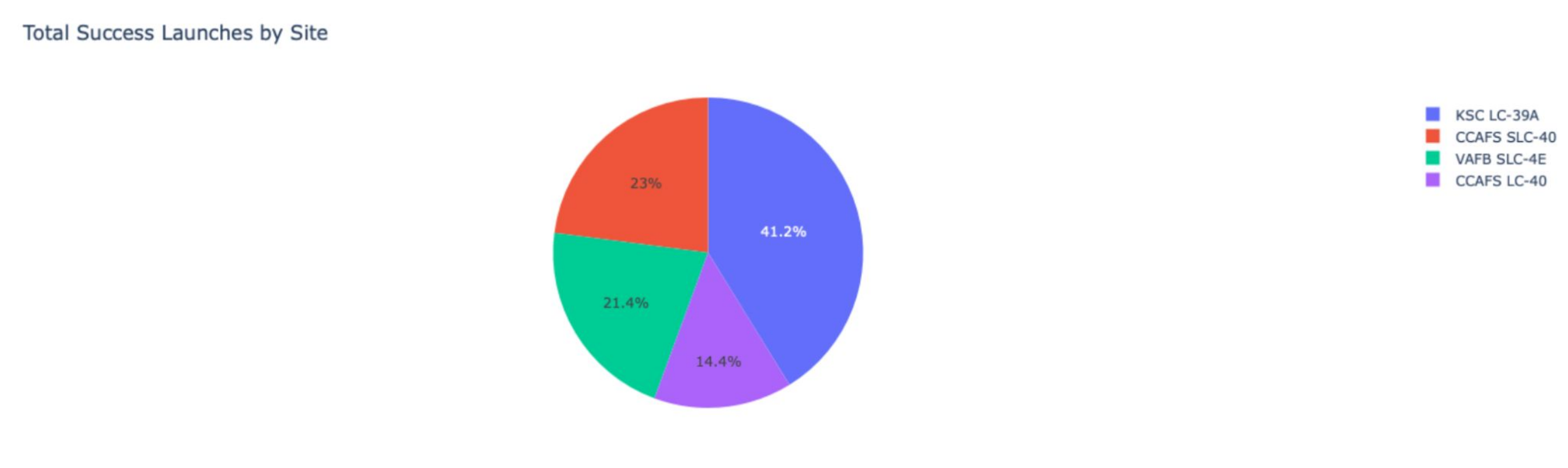
Distance from the launch site KSC LC-39A to its proximities

- 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.



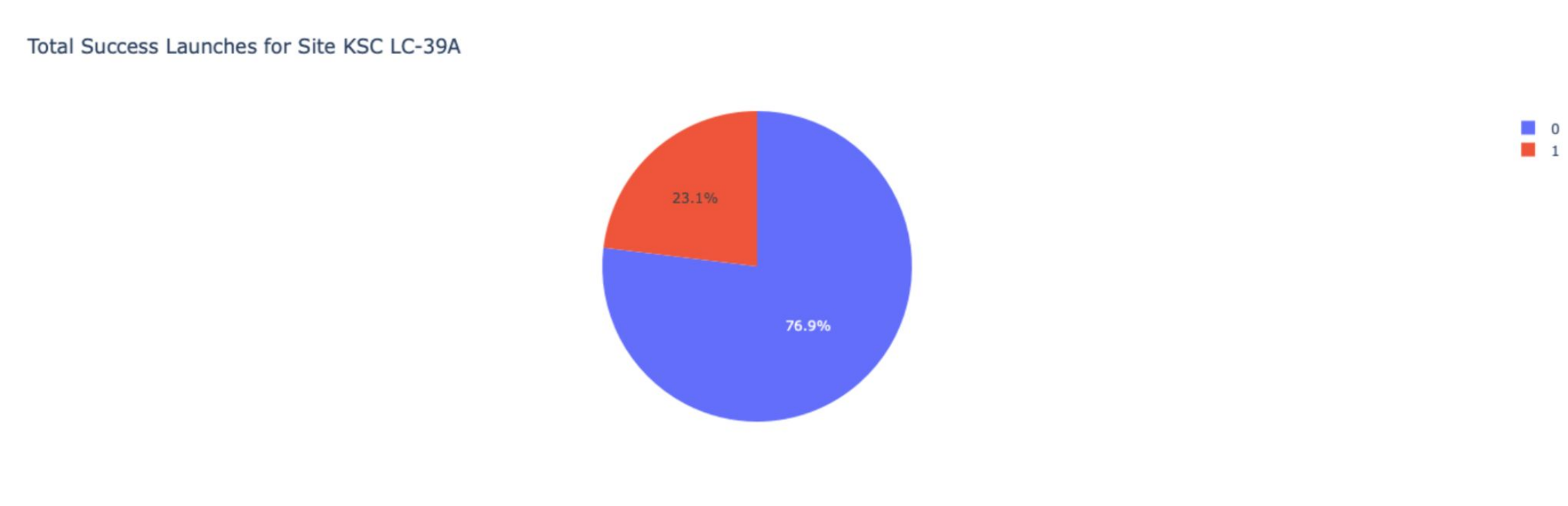

Section 4
Build a Dashboard with Plotly Dash

Launch success count for all sites



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

Launch site with highest launch success ratio



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

Payload Mass vs. Launch Outcome for all sites

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

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Based on the scores of the Test Set, we can not confirm which method performs best.
- Same Test Set scores may be due to the small test sample size (18 samples). Therefore, we tested all methods based on the whole Dataset.
- The scores of the whole Dataset confirm that the best model is the Decision Tree Model. This model has not only higher scores, but also the highest accuracy.

Scores and Accuracy of the Test set

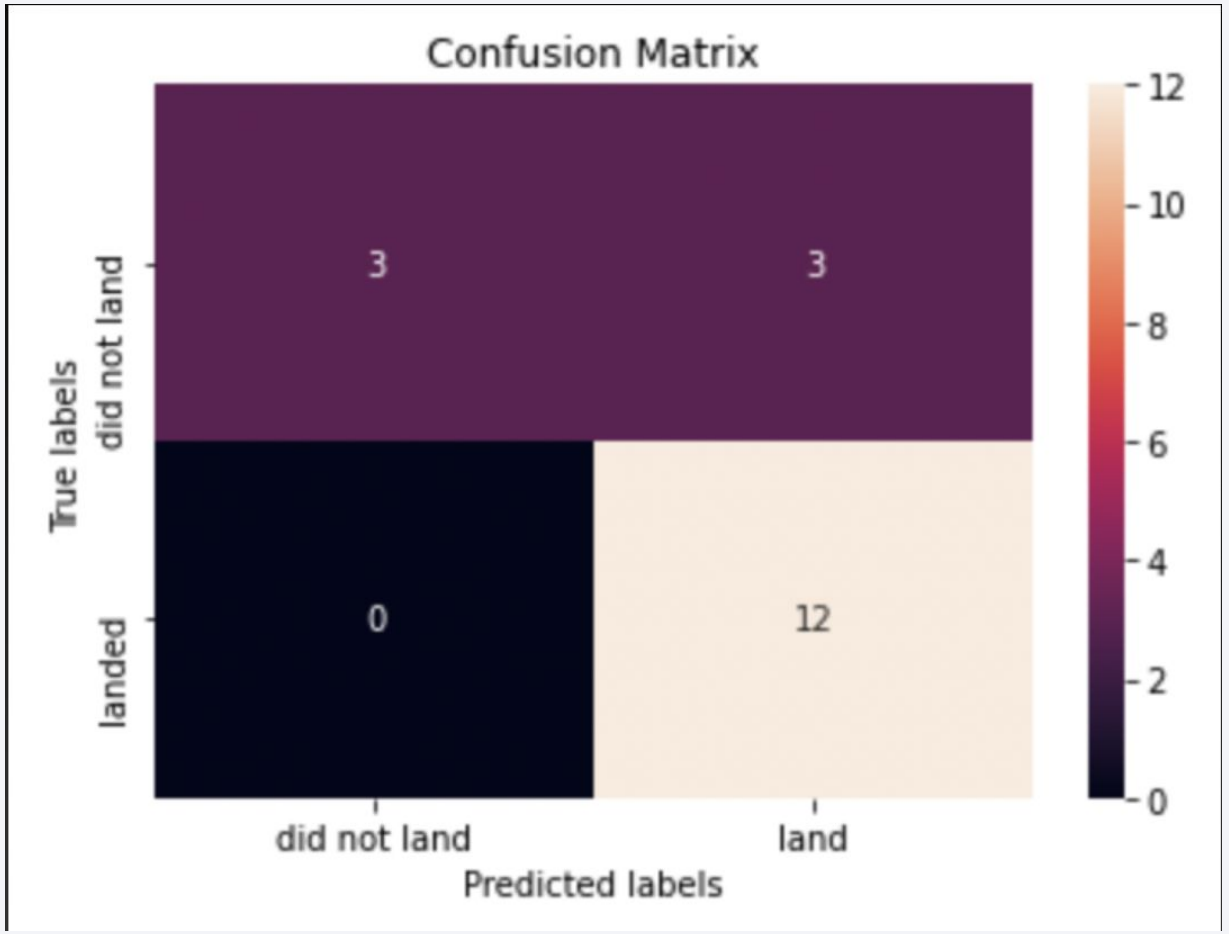
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

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

Confusion Matrix

- Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.



Conclusions

- The Decision Tree Model proved to be the most effective algorithm for predicting landing success with this dataset.
- Launches with lower payload mass tend to have higher success rates compared to those with heavier payloads.
- The majority of launch sites are strategically located near the equator and in close proximity to coastlines, optimizing launch efficiency.
- The success rate of launches has improved consistently over time, indicating advancements in technology and processes.
- Among all the launch sites, KSC LC-39A demonstrated the highest success rate, making it the most reliable site.
- Certain orbits, such as ES-L1, GEO, HEO, and SSO, achieved a 100% success rate, suggesting their suitability for specific mission profiles.

Appendix

SPECIAL THANKS TO :
COURSERA
IBM
INSTRUCTORS

Thank you!