IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Daniel Lecheler
8/2/2025

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
  - Using Folium to build an interactive map
  - Using Plotly Dash to build a Dashboard
  - Predicitve Analysis (with Classification)
- Summary of all results
  - Results of Exploratory Data Analysis
  - Interactive Analytics Demo (screenshots)
  - Predictive Analysis results

Introduction

- Project background and context
  - SpaceX is the dominant company in space travel as the era of commercial space travel takes flight. With a mission of making space travel accessible and affordable, the company advertises its *Falcon 9* rocket launches on its websites which cost an average of $62 million which is almost one tenth of what their competitors are offering. This reduction in total cost is primarily due to the reusing of the first stage of the rocket and therefore the total cost of the launch can be deduced based on knowing whether the first stage landed without issue. Using public information and data combined with machine learning models, this project will predict the if SpaceX will reuse the first stage of their rocket
- Problems you want to find answers
  - How do the following variables effect the success of the first stage of the rocket landing:
    - Payload mass
    - Launch site location
    - Number of flights
    - Orbits of Earth
  - Does the rate of the first stage landing successfully increase over time?
  - What is the best algorithm that can be used for binary classification for this project?

Section 1

Methodology

Methodology

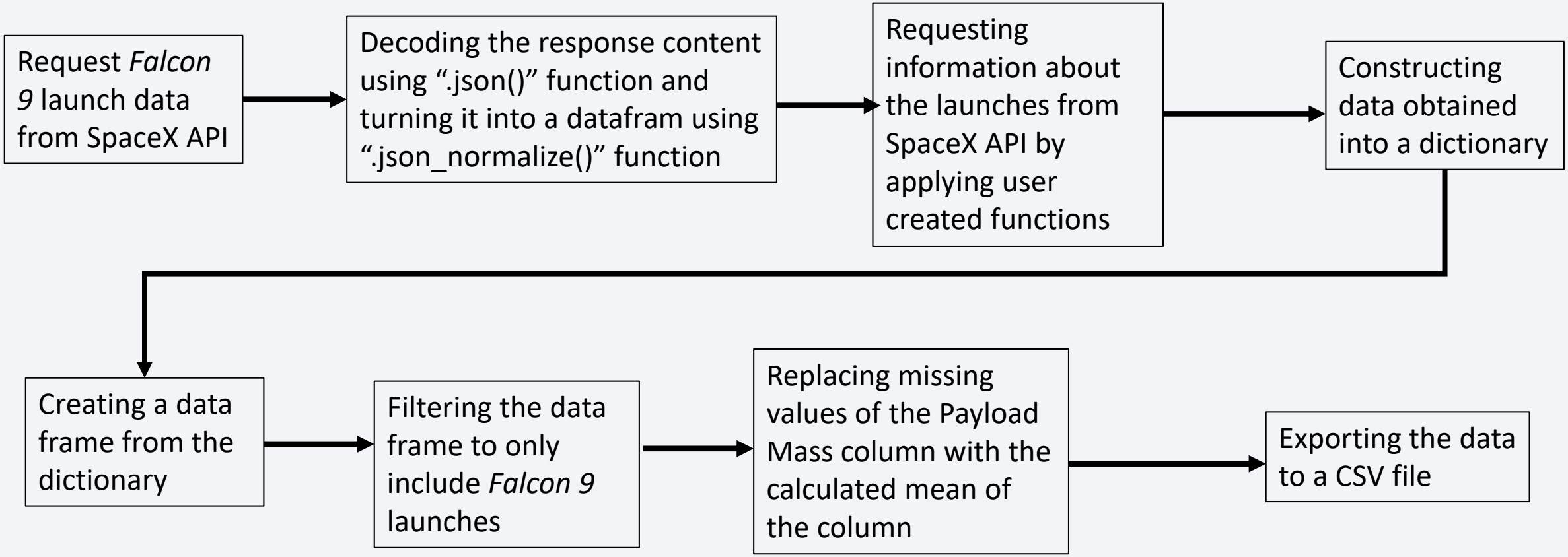
Executive Summary

- Data collection methodology:
  - Using SpaceX Rest API
  - Using Web Scrapping from Wikipedia
- Perform data wrangling
  - Filtering the data
  - Dealing with missing variable values
  - Using One Hot Encoding to prep the data for 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 the classification models to optimize the results

Data Collection

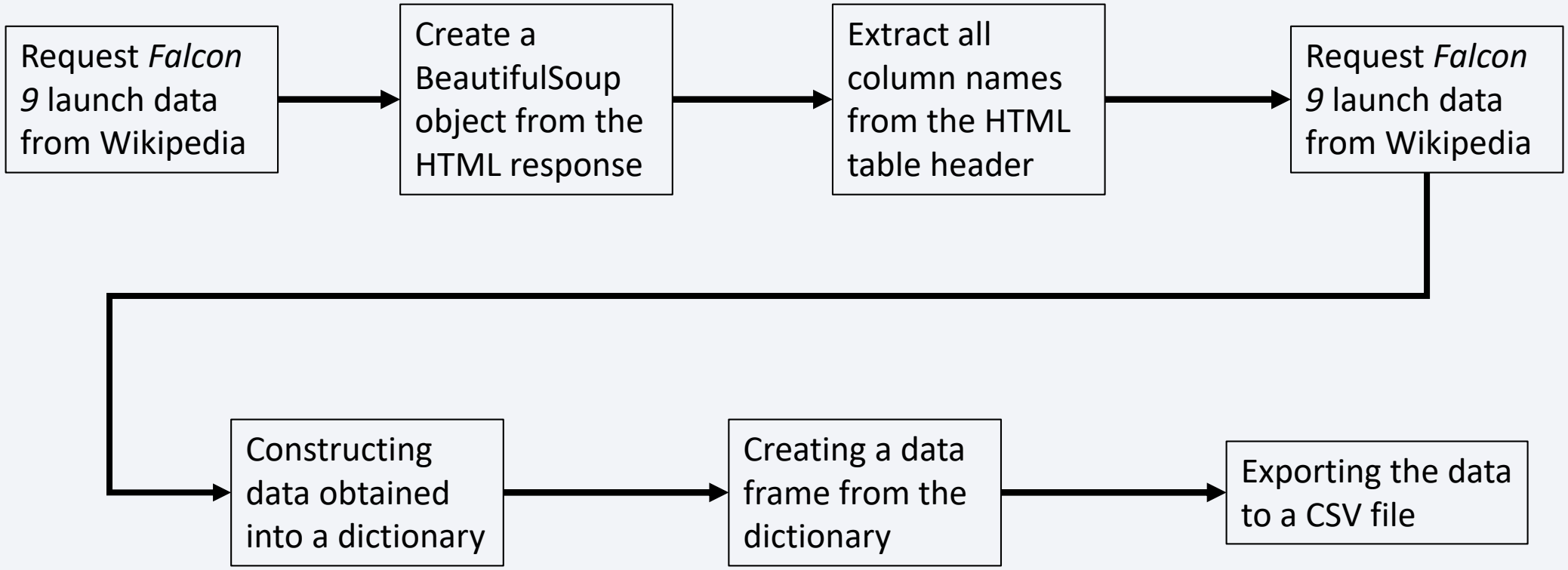
- The data collection process combined the use of API request from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia page
  - The following data columns were obtained by using SpaceX RestAPI:
    - FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, Launch Site, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
  - The following data columns were obtained using Web Scraping:
    - Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



Link to GitHub: Data Collection API

Data Collection - Scraping



Link to GitHub: Data Collection with Web Scraping

Data Wrangling

Link to GitHub: Data Wrangling

- In the data set, there are several different cases where the first stage booster did not land successfully. Some of these involved an attempted landing that failed due to an accident. The following are the different types of outcomes and what they each mean
  - True Ocean: Mission outcome was considered successful since the landing was in the specific section of ocean that was expected
  - False Ocean: Mission outcome was considered a failure since the landing was outside of the expected section of ocean
  - True RTLS: Mission outcome was successful since the landing was on the ground pad
  - False RTLS: Mission outcome was a failure since the rocket failed to land on the ground pad
  - True ASDS: Mission outcome was successful since the landing was on a drone ship
  - False RTLS: Mission outcome was a failure since the rocket failed to land on a drone ship
- The true or false first word prefix is converted to the Boolean 1 and 0 to indicate success or failure in the launches

EDA with Data Visualization

Link to GitHub: [EDA with Data Visualization]

- Charts that were 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 were used to show the relationship between variables and if that relationship existed, they were used for machine learning models
- Bar charts showed comparisons among discrete categories with a goal of showing the relationship between the specific categories and the measured value
- Line charts showed trends in data over time (AKA time series)

EDA with SQL

Link to GitHub: EDA with SQL

- SQL queries that were performed:
  - Displaying the names of the unique launch sites
  - Displaying 5 records where the launch site began with the string “CCA”
  - Displaying the total payload mass caried by boosters launched by NASA
  - Displaying the average payload mass caried by booster version F9 v1.1
  - Listing the date of the first successful landing on the ground pad
  - Listing the names of the boosters which have success in drone ship and have a payload mass between 4000 and 6000
  - Listing the total number of successful and failed mission outcomes
  - Listing the names of the booster versions which carried the maximum payload mass
  - Listing the failed landing outcomes in drone ship, the booster version used, and the launch site names in 2015
  - Raning the count of landing outcomes between the date of 6/4/2010 and 3/20/2017 from most recent to oldest

Build an Interactive Map with Folium

Link to GitHub: Interactive Visual Analytics with Folium

- Markers of all launch sites:
  - The NASA Johnson Space Center is marked with a circle, pop-up label based on it's exact latitudinal and longitudinal coordinates
  - All launch sites are marked with a circle, pop-up label based on their exact latitudinal and longitudinal coordinates
- Colored Markes of the launch outcomes for each site
  - Added colors to markers to indicate success (green) or failure (red) using Marker Cluster to identify which launch sites have the highest success rates
- Approximate instances between sites
  - Added colored lines to show distances between the launch sites and geographical indicators (coastline), transportation routes (railway, highway), and urban centers (the closest city)

Build a Dashboard with Plotly Dash

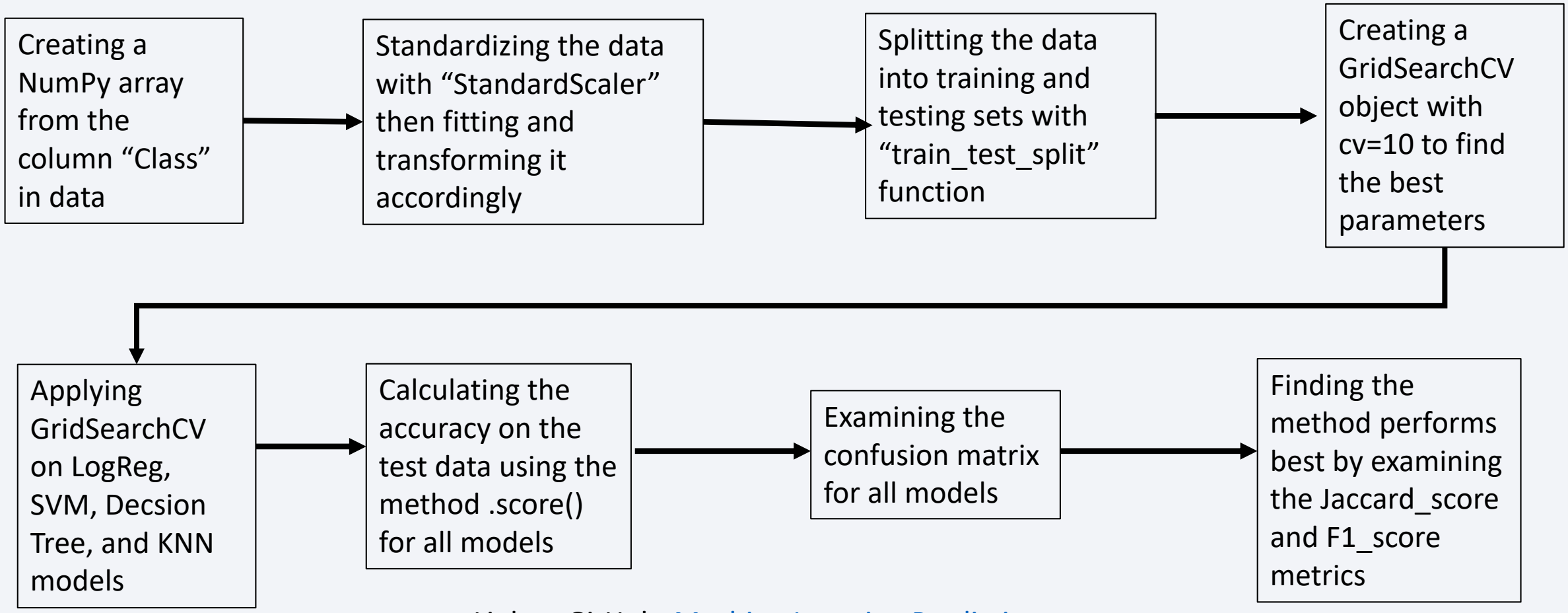
Link to GitHub: SpaceX Dash App

- Launch Sites Drop-down List:
  - Added a drop-down list to enable launch site selection
- Pie Chart of Successful Launches:
  - Added a pie chart to show the total successful launches count for all sites and the Success vs. Failure 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 different Booster Versions:
  - Added a scatter chart to show the correlation between Payload and Launch Success

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
- You need present your model development process using key phrases and flowchart
- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose

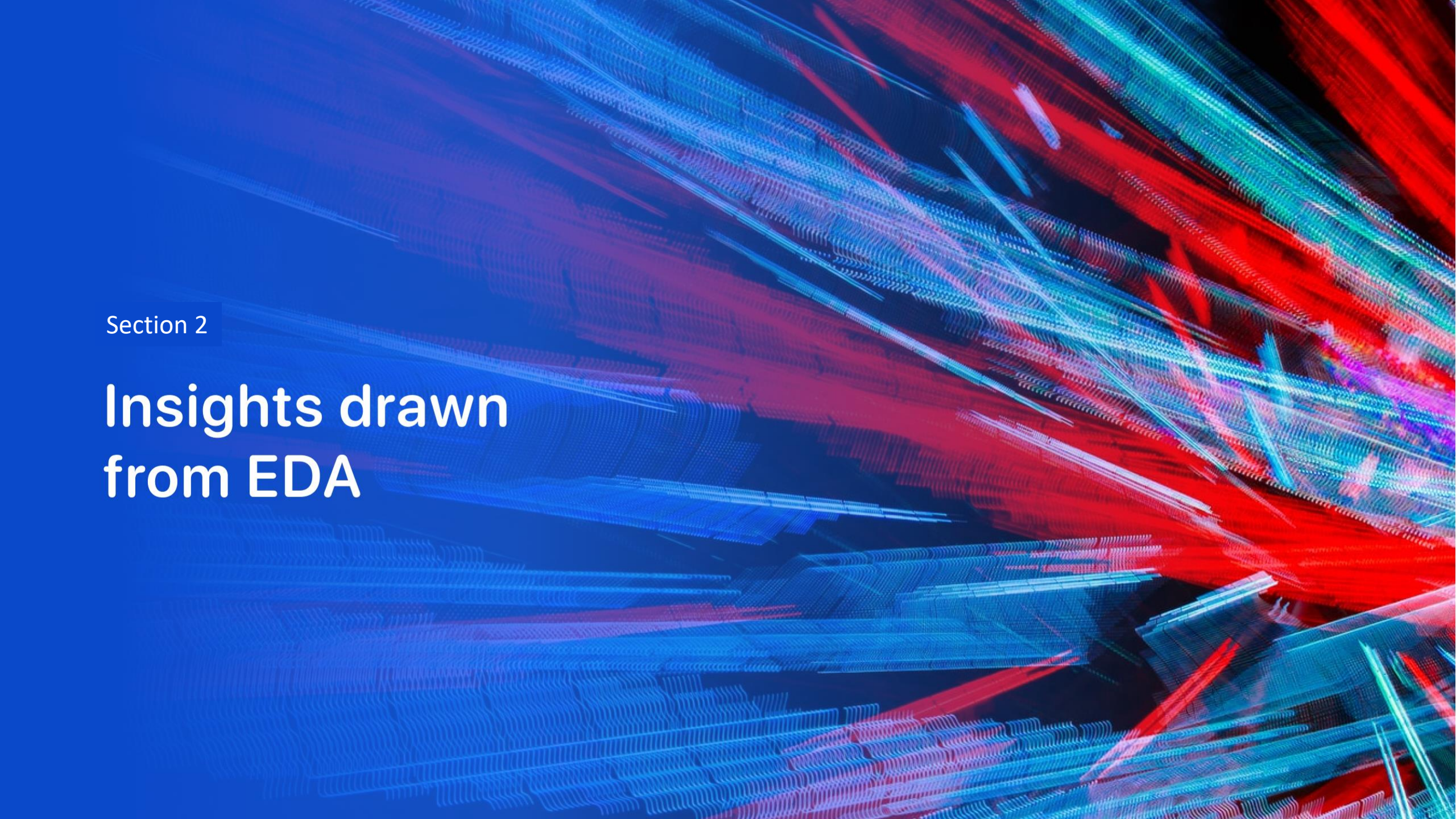
Predictive Analysis (Classification)



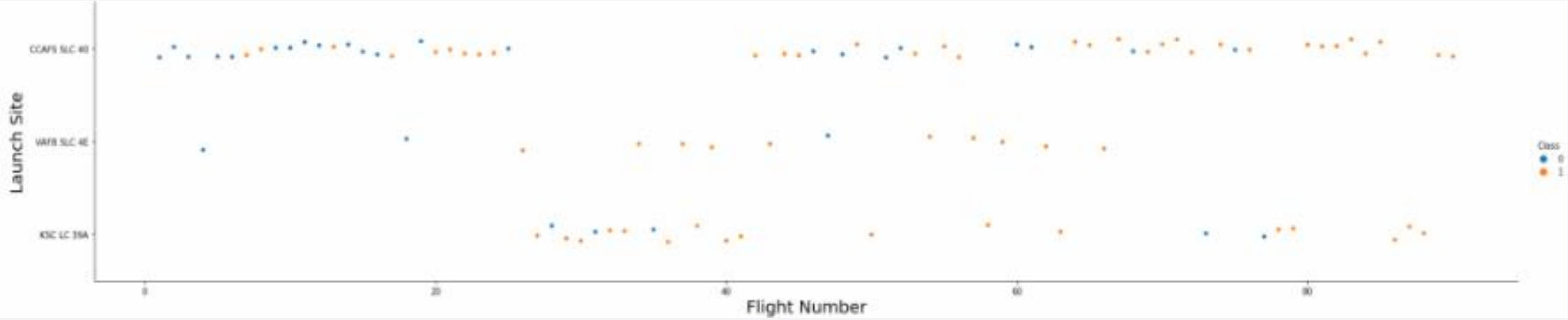
Link to GitHub: [Machine Learning Prediction]

Results

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

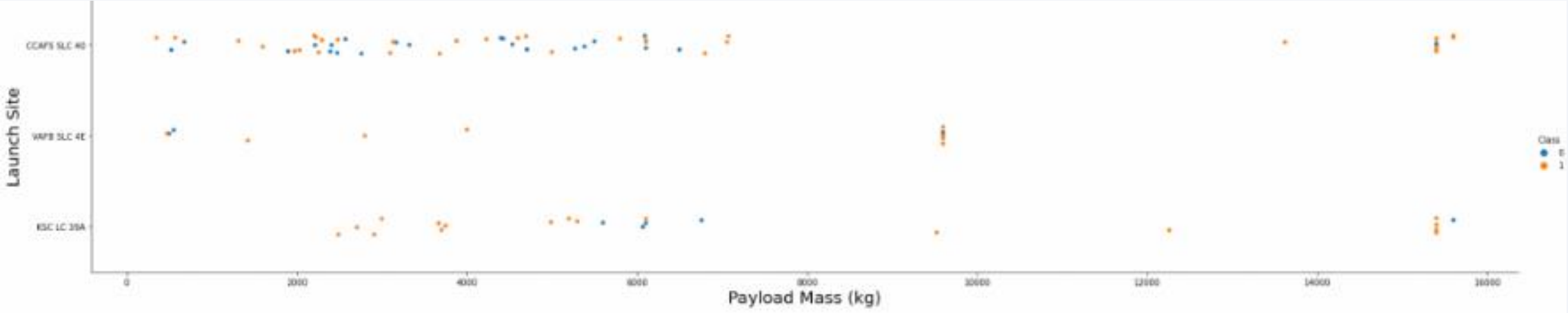

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- Explanation:
  - The earliest 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

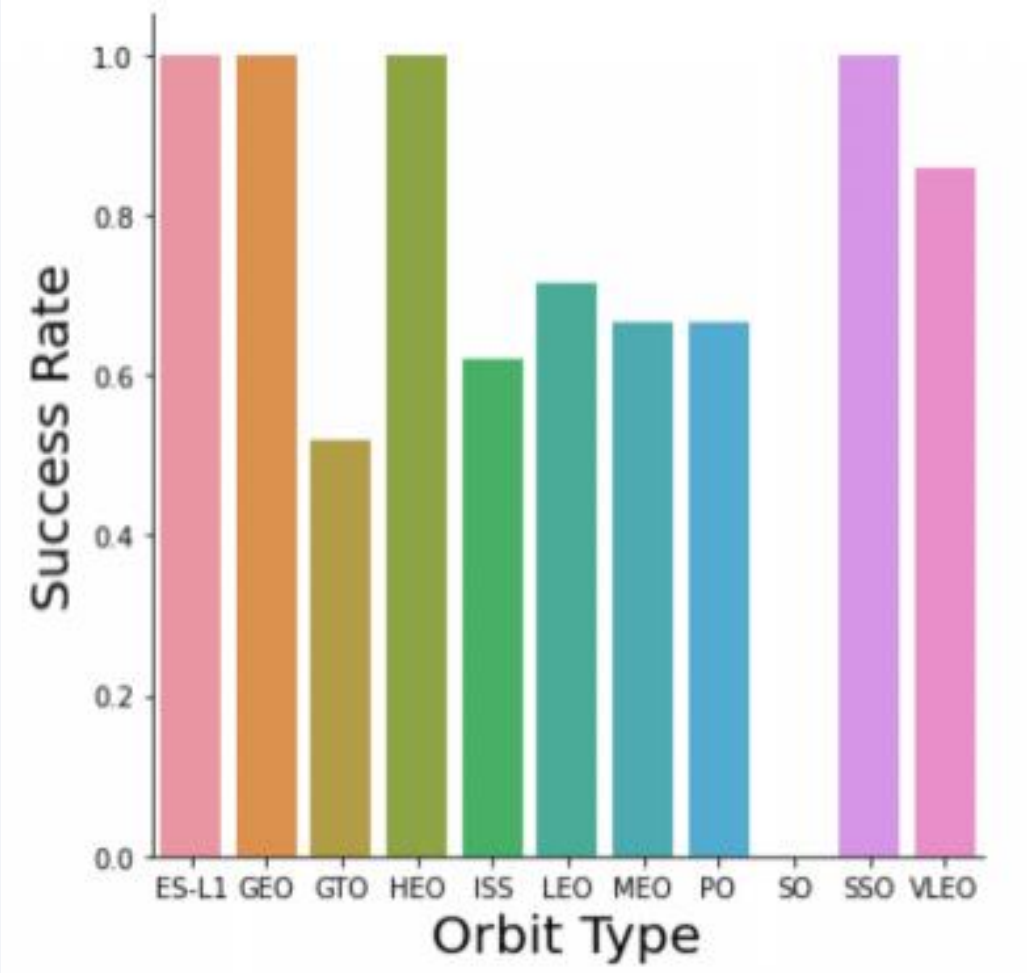
Payload vs. Launch Site



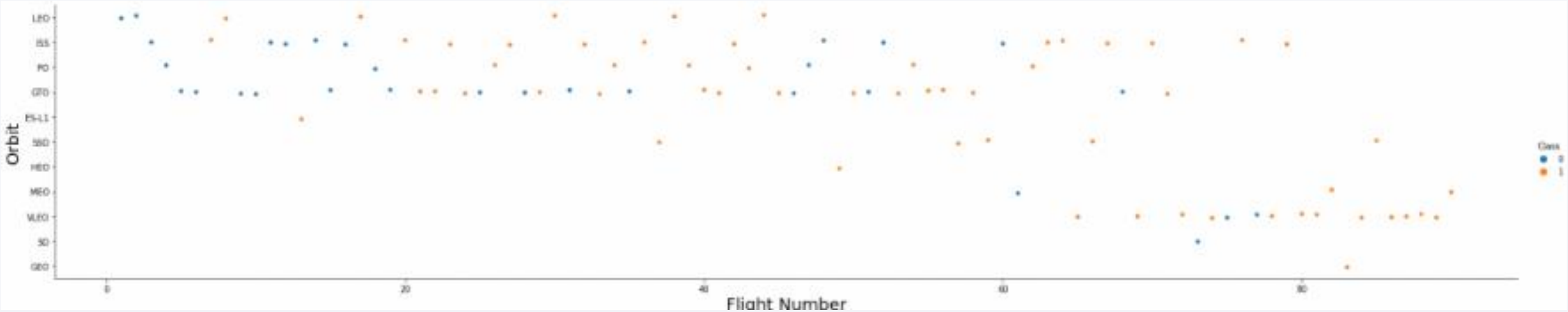
- Explanation
  - For every launch site the higher the payload mass, the higher the success rate.
  - Most of the launches with payload mass over 7000 kg were successful.
  - KSC LC 39A has a 100% success rate for payload mass under 5500 kg too

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

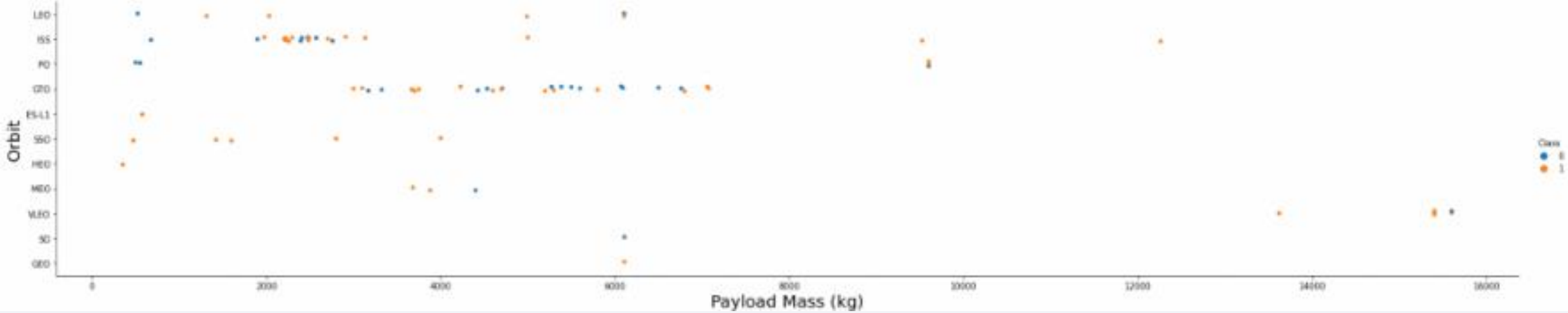


Flight Number vs. Orbit Type



- 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

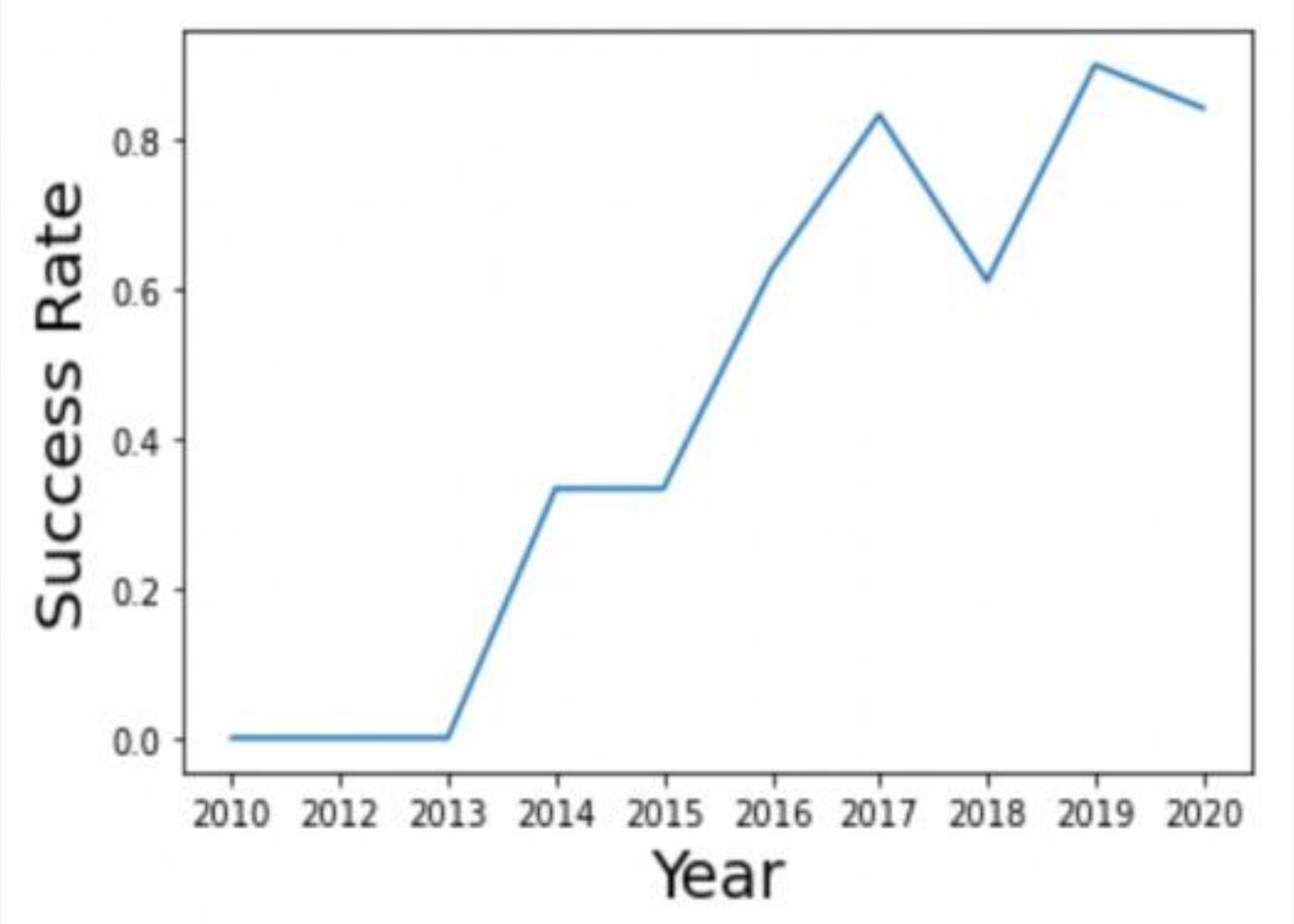
Payload vs. Orbit Type



- 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

```
In [4]: %sql select distinct launch_site from SPACEXDATASET;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[4]:

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

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

Launch Site Names Begin with 'CCA'

```
In [5]: %sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqblod8lcg.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

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

Total Payload Mass

```
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

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

Average Payload Mass by 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.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[7]:

average_payload_mass
2534

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

First Successful Ground Landing Date

```
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.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[8]:

first_successful_landing
2015-12-22

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

Successful Drone Ship Landing with Payload between 4000 and 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.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[9]:

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

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

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.bs2io90l08kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[10]:

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

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

Boosters Carried Maximum Payload

```
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.bs2io90108kqb1od8lcg.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

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

2015 Launch Records

```
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.bs2io90108kqblod8lcg.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)

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

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

```
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.bs2io90108kqblod8lcg.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

- 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

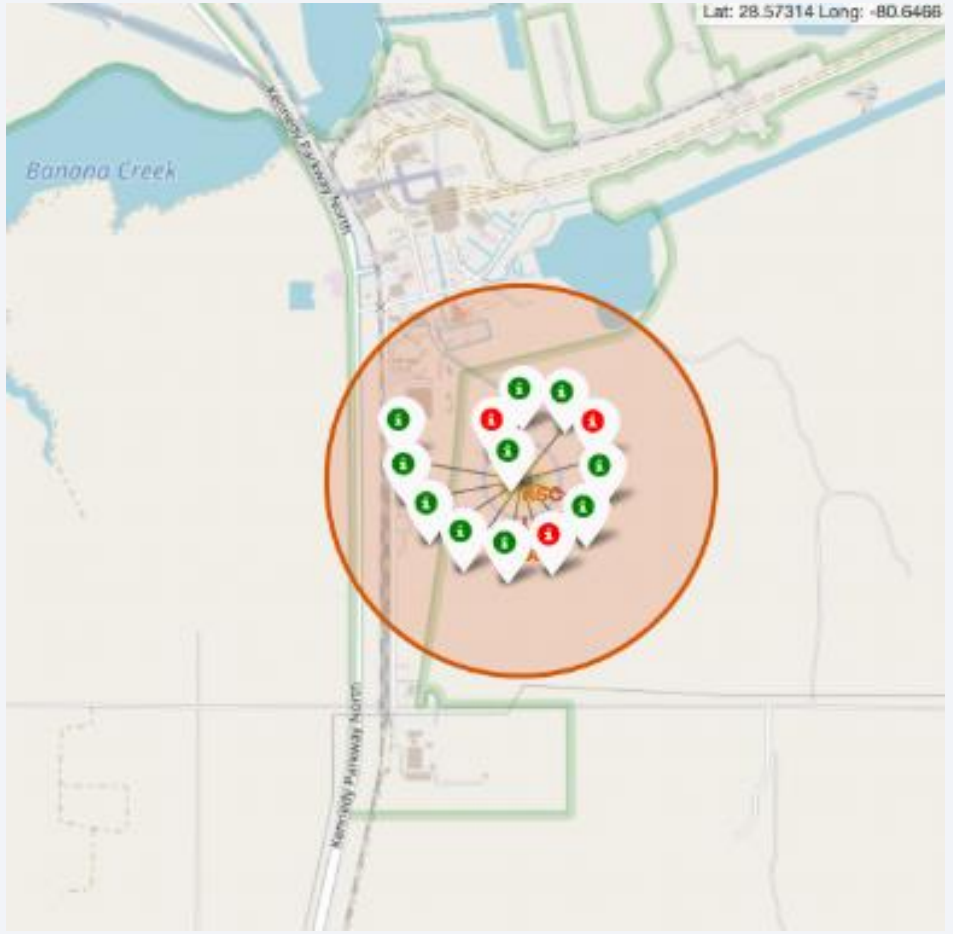

Section 3
Launch Sites
Proximities Analysis

All launch sites' location markers on a global map

- Explanation:
  - 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

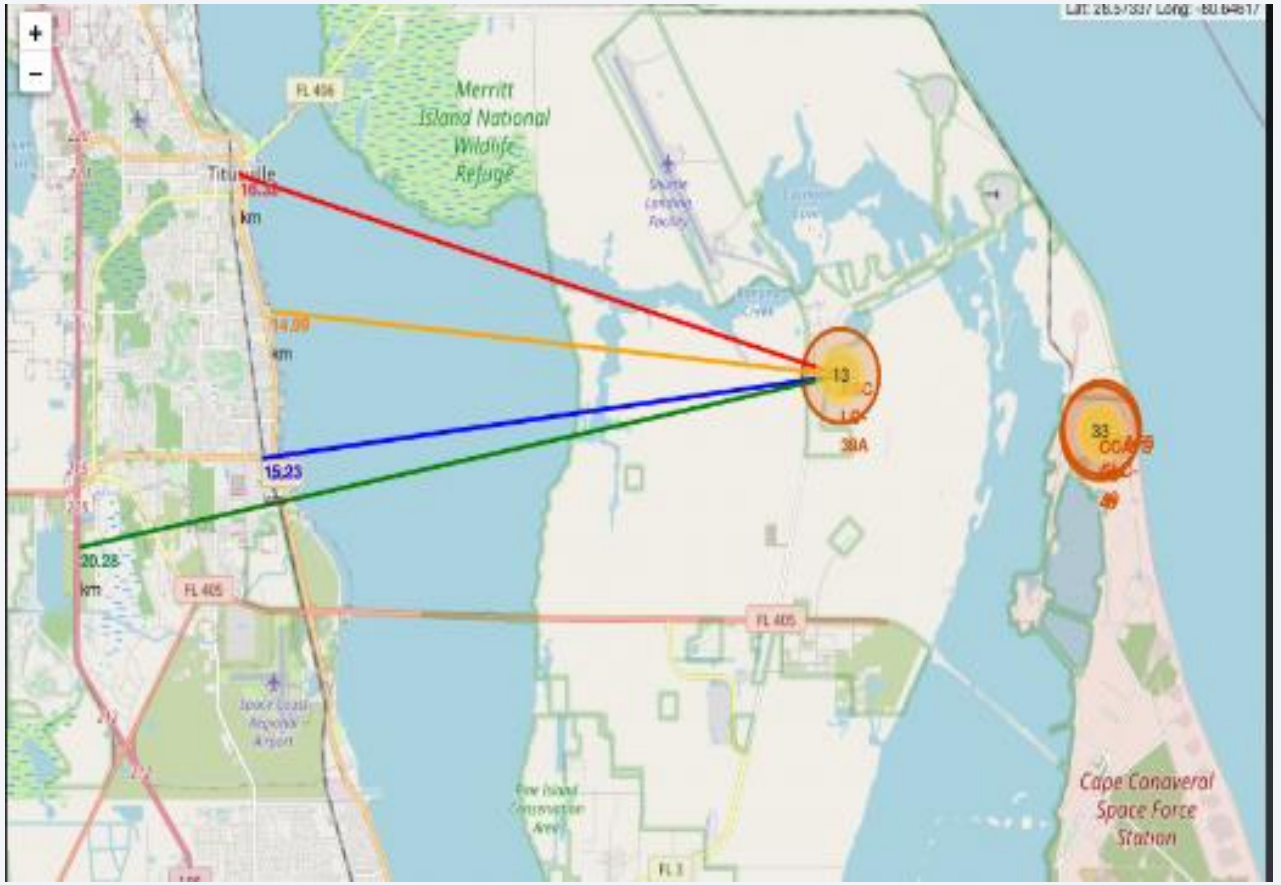
<Folium Map Screenshot 2>

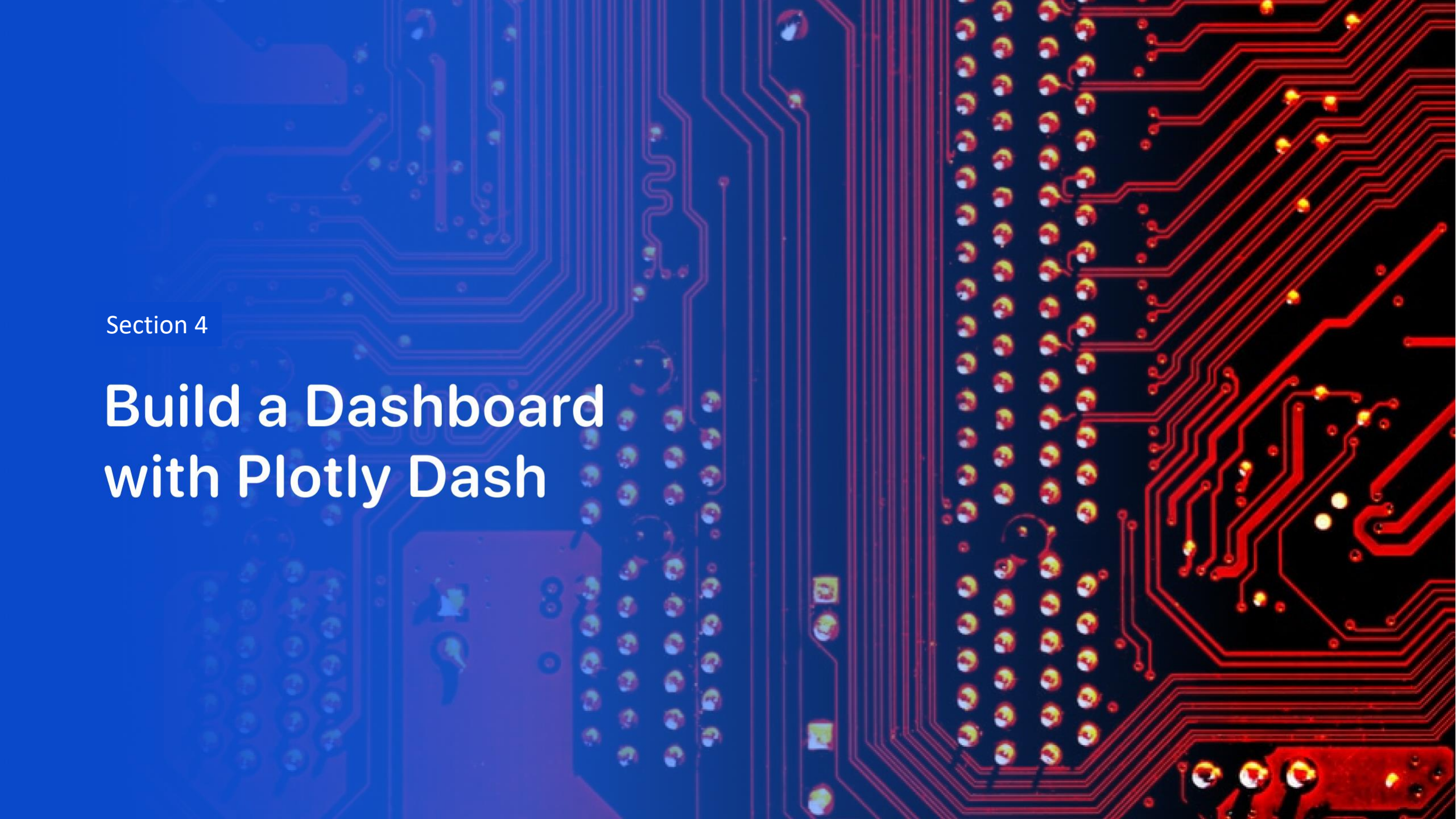
- Explanation:
  - From the color-labeled markers we should be able to easily identify which launch sites have relatively high success rates.
    - Green Marker = Successful Launch
    - Red Marker = Failed Launch
  - Launch Site KSC LC-39A has a very high Success Rate.



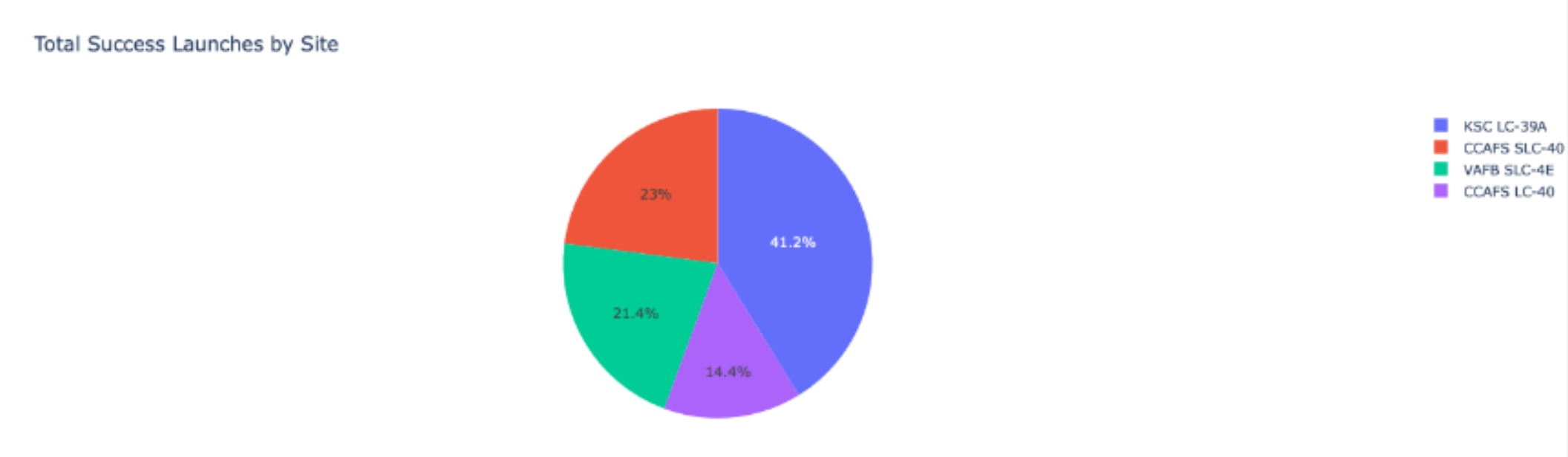
Distance from the launch site KSC LC-39A to its proximities

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



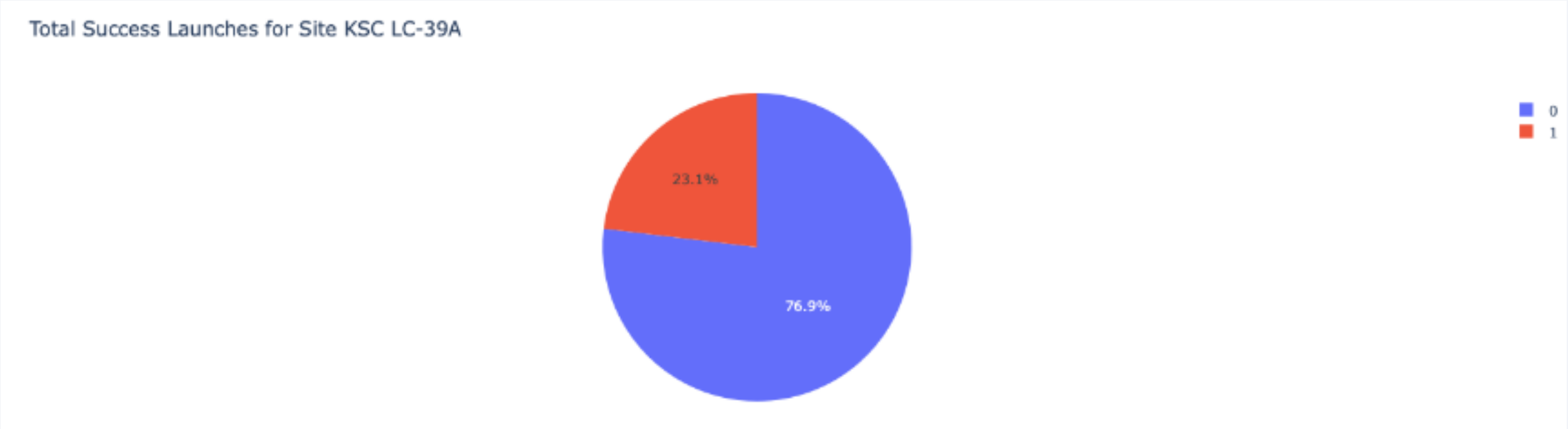

Section 4
Build a Dashboard with Plotly Dash

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:
  - 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

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

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Explanation:
  - 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 overall 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

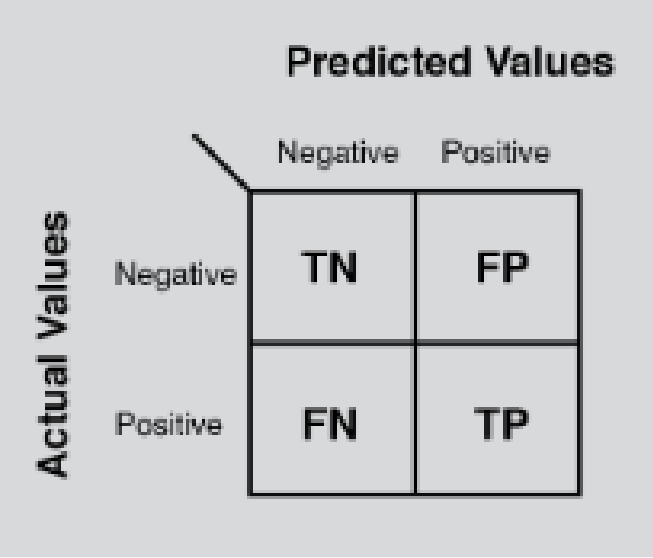
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

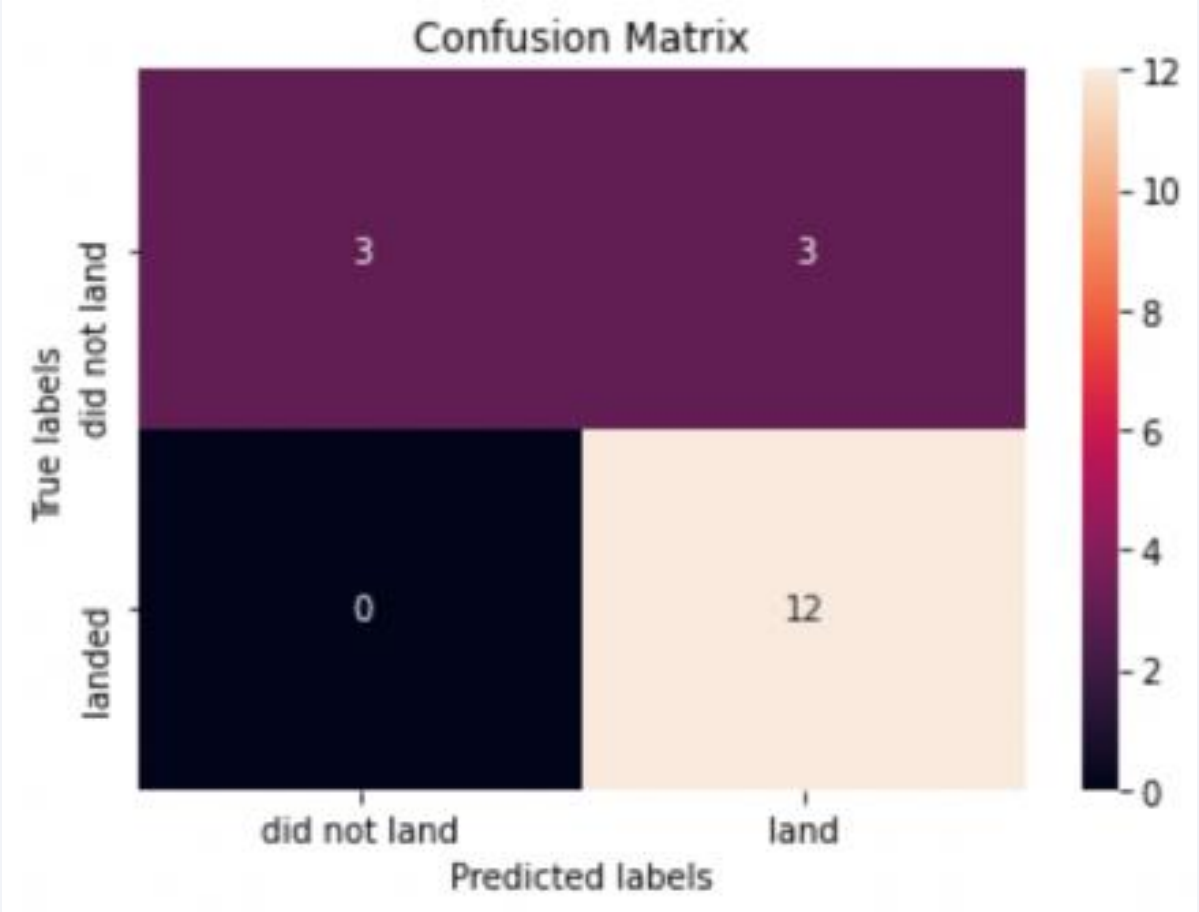
Scores and Accuracy of the Entire Data Set

	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

- Explanation:
  - 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

- Decision Tree Model is the best algorithm for this dataset.
- Launches with a smaller 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 a coast of a large body of water.
- The success rate of launches increases over the time.
- KSC LC-39A has the highest success rate of all the launch sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.

Thank you!