IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Abtaal Aatif
19th Dec, 2024

Outline

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

Executive Summary

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

Introduction

- Project background and context

  Making space travel accessible, SpaceX is the most prosperous company of the commercial space age. On its website, the business promotes the 62 million USD Falcon 9 rocket launches. Because SpaceX can reuse the first stage, they can save a significant amount of money compared to other carriers who charge up to 165 million dollars apiece. The cost of a launch can therefore be ascertained if we can predict whether the first stage will land. Using machine learning models and publicly available data, we will forecast if SpaceX will reuse the first stage.

- Problems you want to find answers

  - What effects do factors like payload mass, launch location, number of lights, and orbits have on the first stage landing's success?
  - Over time, does the number of successful landings rise?
  - Which algorithm works best in this situation for binary classification?

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 values
  - Using One Hot Encoding to prepare the data to a binary classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building, tuning and evaluation of classification models to ensure the best results

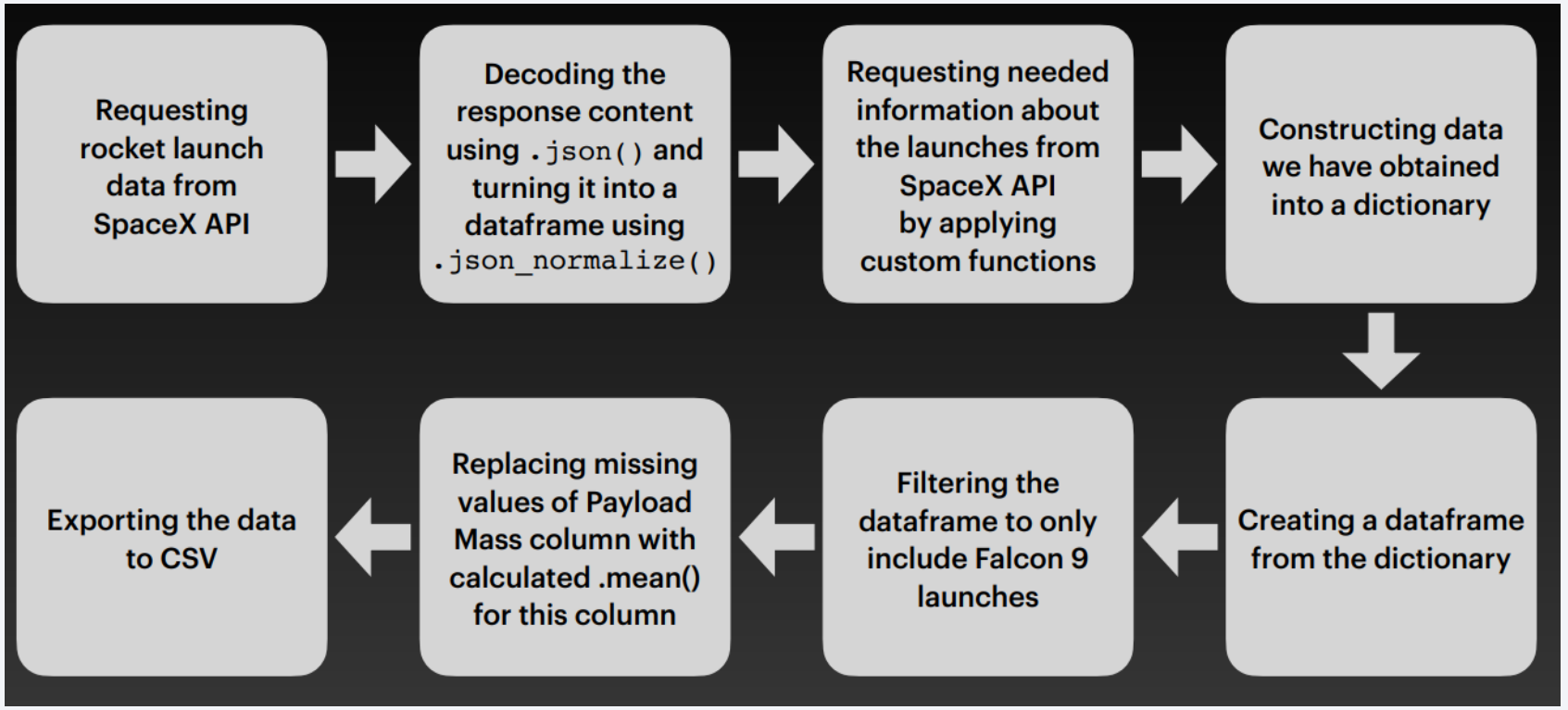
Data Collection

- Describe how data sets were collected
  - A combination of web scraping information from a table in SpaceX's Wikipedia entry and API queries from the SpaceX REST API were used in the data collection procedure.
- You need to present your data collection process use key phrases and flowcharts
  - To obtain comprehensive information about the launches for a more thorough analysis, we had to employ both of these data collection techniques.

Data Collection – SpaceX API

Below, you find the link for accessing the notebook that uses the API:

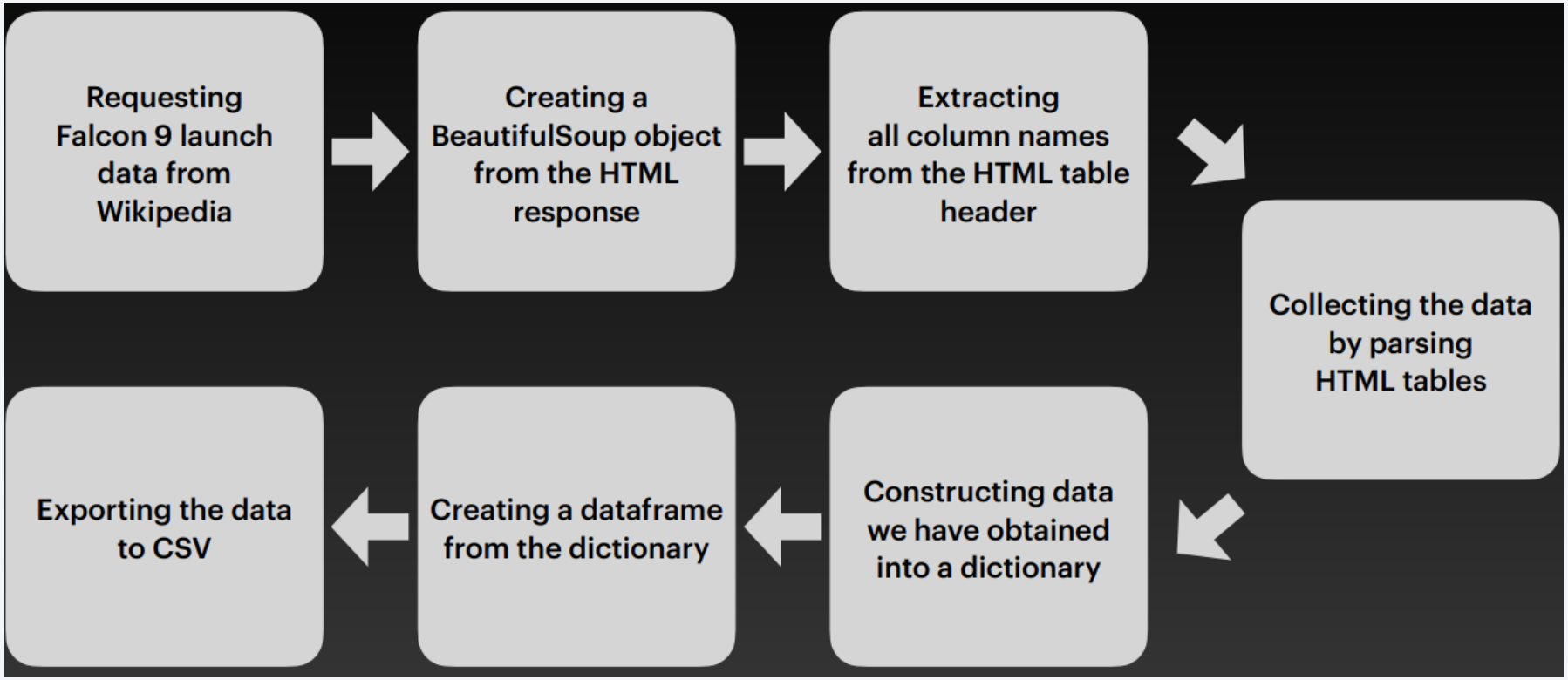
- https://github.com/Null9703/IBM-Data-Science-Capstone-Project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

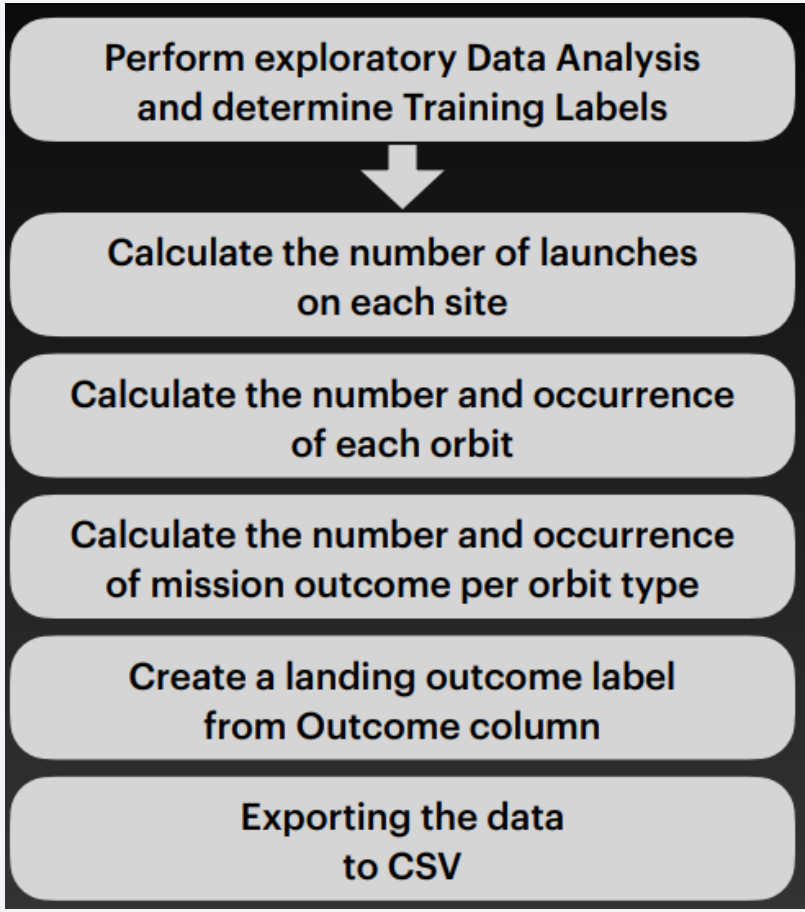
Below, you find the link for accessing the webscraping notebook:

- https://github.com/Null9703/IBM-Data-Science-Capstone-Project/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

There are multiple instances in the data set where the booster failed to land. Occasionally, an accident causes a landing attempt to fail; for instance, True Ocean indicates that the mission outcome was successfully landed to a specific area of the ocean, whereas False Ocean indicates that the mission outcome was unsuccessfully landed to a specific area of the ocean. When the mission outcome was successfully landed on a ground pad, it is known as true RTLS. An unsuccessful mission outcome landing on a ground pad is indicated by a false RTLS.A successful mission outcome landing on a drone ship is referred to as true ASDS. An unsuccessful mission outcome on a drone ship is indicated by a false ASDS. We mainly convert those outcomes into Training Labels with “1” means the booster successfully landed, “0” means it was unsuccessful.



- Data Wrangling Notebook: https://github.com/Null9703/IBM-Data-Science-Capstone-Project/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- **The charts we plotted were:** Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type and Success Rate Yearly Trend.
- Scatter plots illustrate how variables relate to one another. They could be incorporated into a machine learning model if a relationship is present.
- Comparisons between distinct categories are displayed in bar charts. The objective is to demonstrate the connection between a measured value and the specific categories under comparison.
- Data patterns throughout time are displayed in line charts (time series).
- **The notebook for eda with data visualization:** https://github.com/Null9703/IBM-Data-Science-Capstone-Project/blob/main/edadataviz.ipynb

EDA with SQL

- SQL queries were written for:
  - 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 irst 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 6000
  - Listing 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 2015
  - 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
- EDA with SQL notebook: https://github.com/Null9703/IBM-Data-Science-Capstone-Project/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

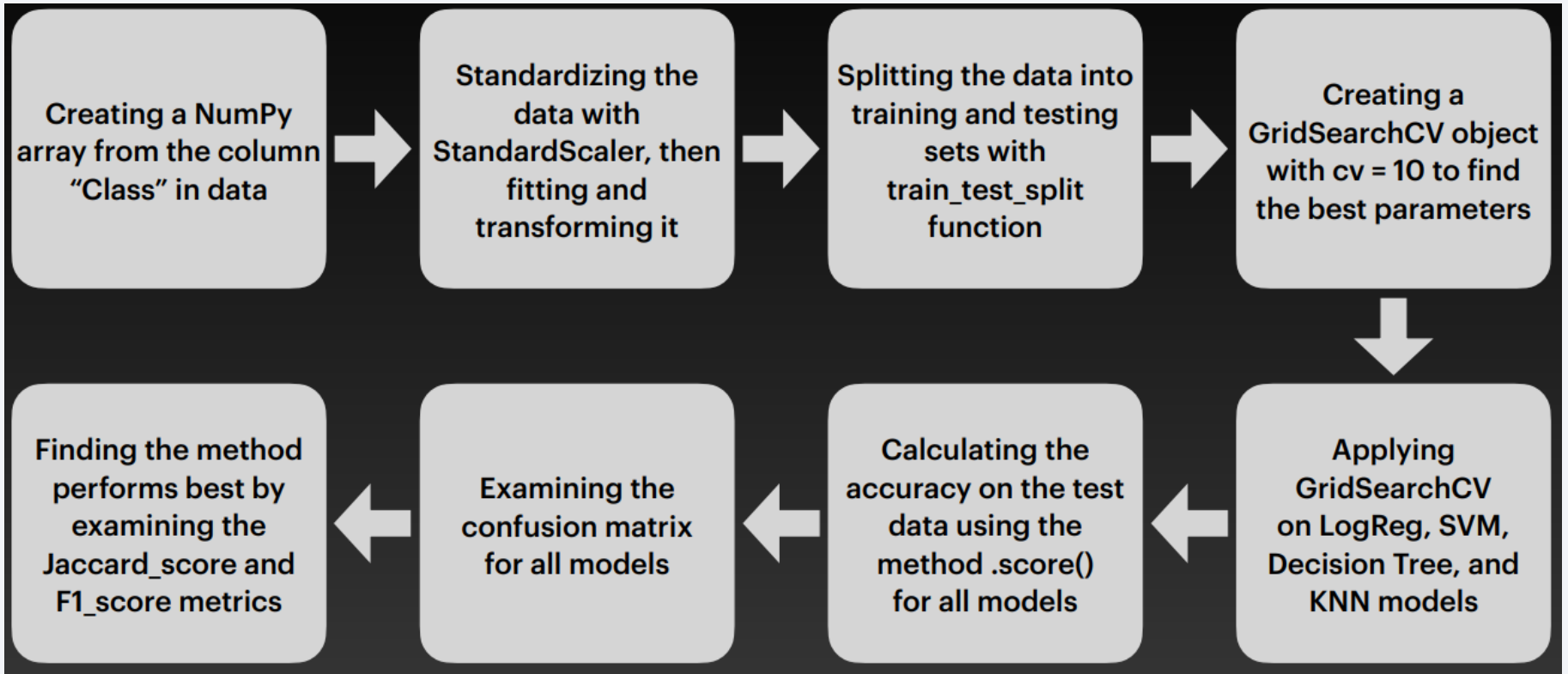
- After creating a map, we added markers:
  - Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.
  - Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.
- Added colored Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.
- Added colored Lines to show distances between the Launch Site and its proximities like Railway, Highway, Coastline and Closest City.
- Notebook for Folium: https://github.com/Null9703/IBM-Data-Science-Capstone-Project/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Added a dropdown list to enable Launch Site selection.
- 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.
- Added a slider to select Payload range.
- Added a scatter chart to show the correlation between Payload and Launch Success.
- Plotly Dash app: https://github.com/Null9703/IBM-Data-Science-Capstone-Project/blob/main/spacex_dash_app.py

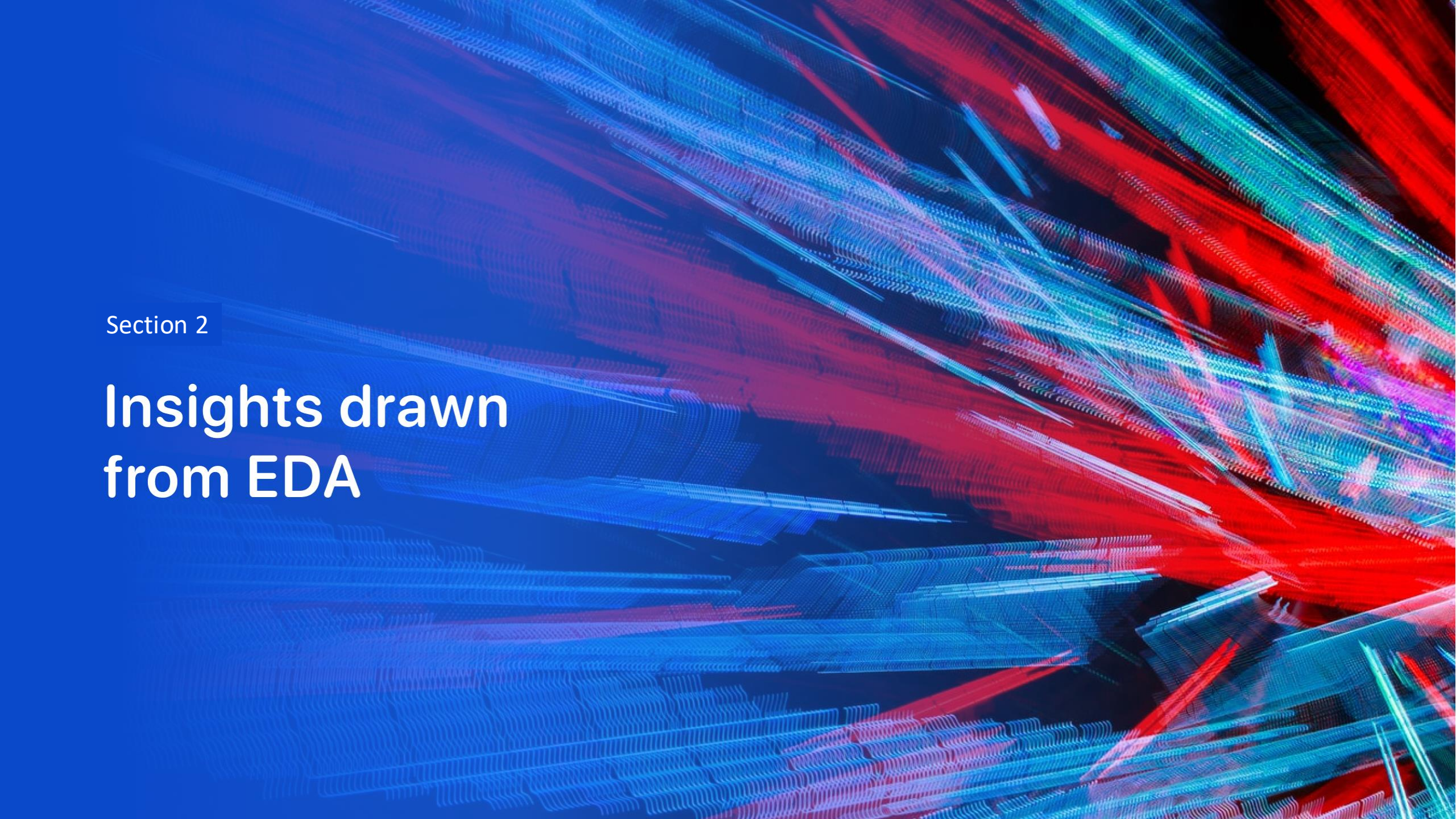
Predictive Analysis (Classification)

- Notebook for predictive analysis: https://github.com/Null9703/IBM-Data-Science-Capstone-Project/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

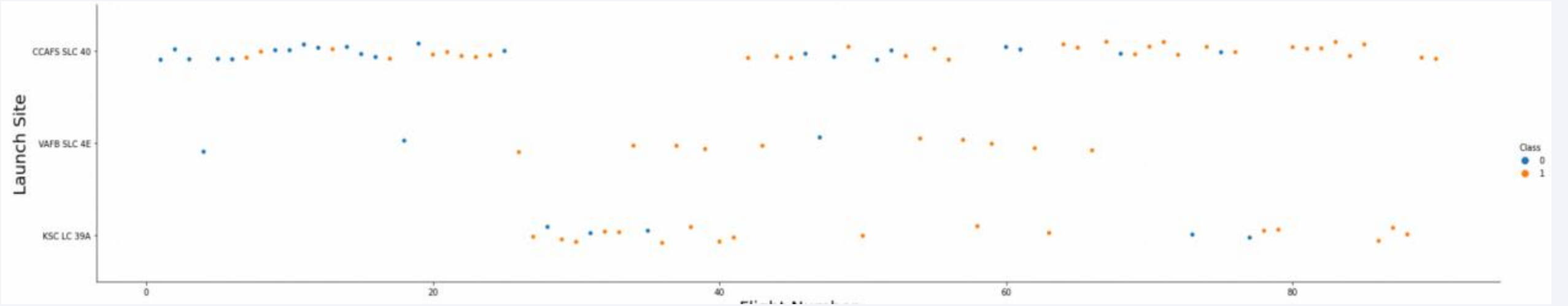


Results

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

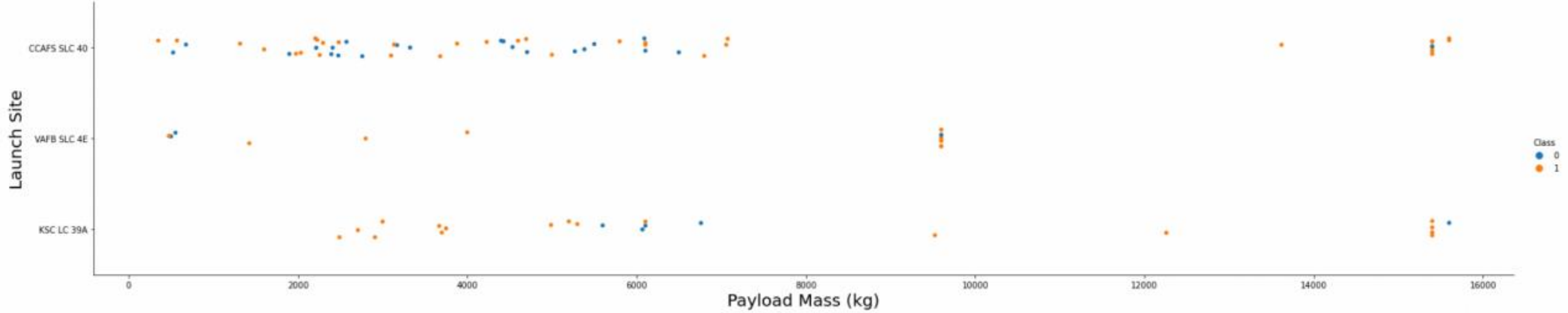

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



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

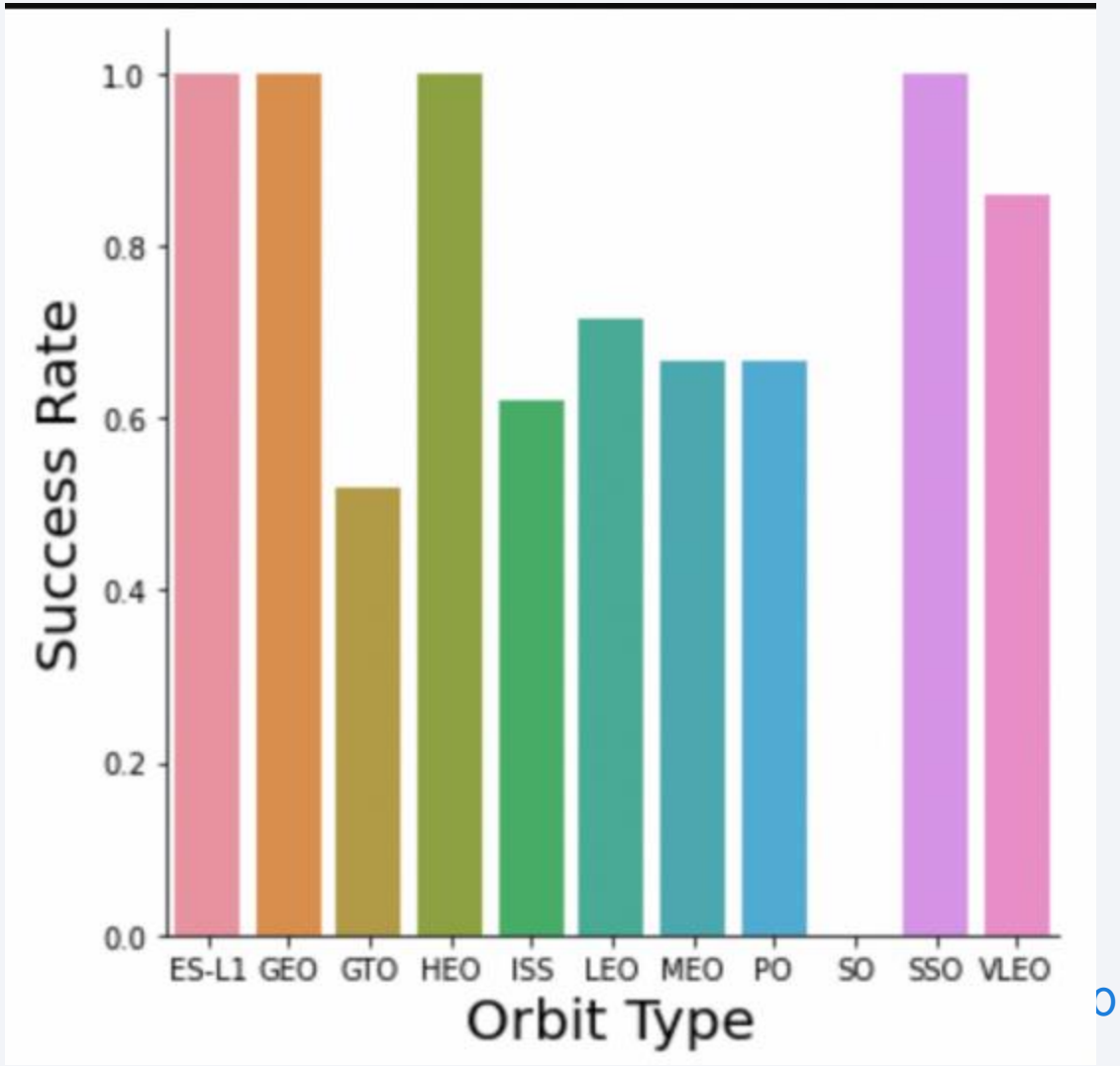
Payload vs. Launch Site



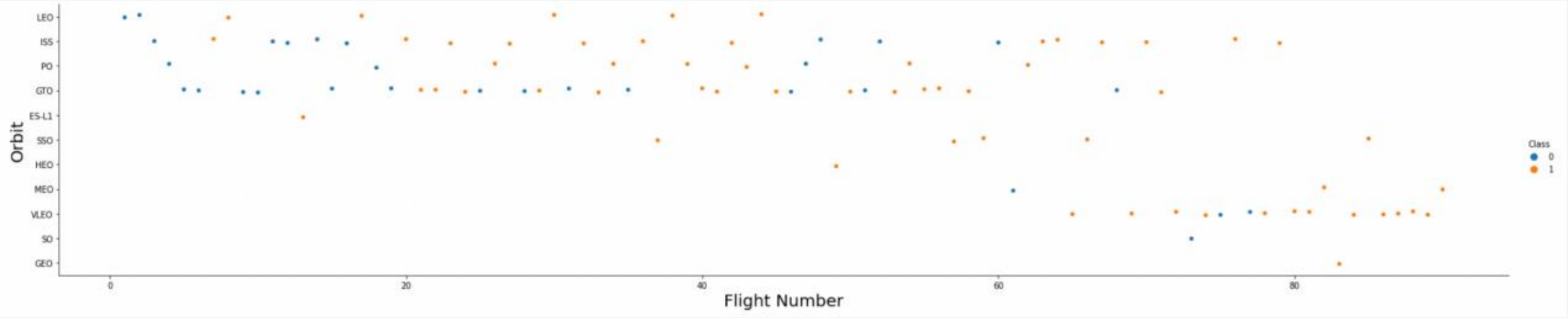
- For every launch site the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

Success Rate vs. Orbit Type

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

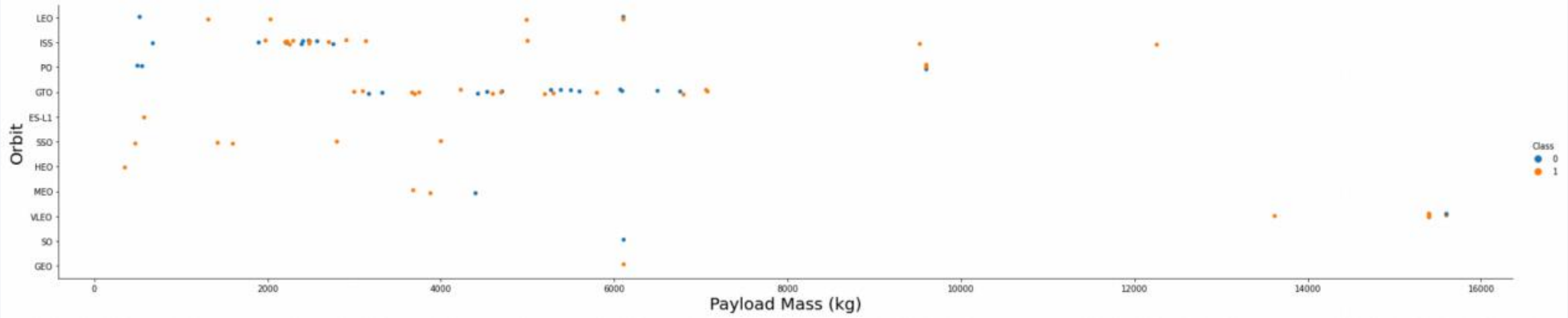


Flight Number vs. Orbit Type



- In the LEO orbit the Success appears related to the number of lights;
- There seems to be no relationship between light number when in GTO orbit.

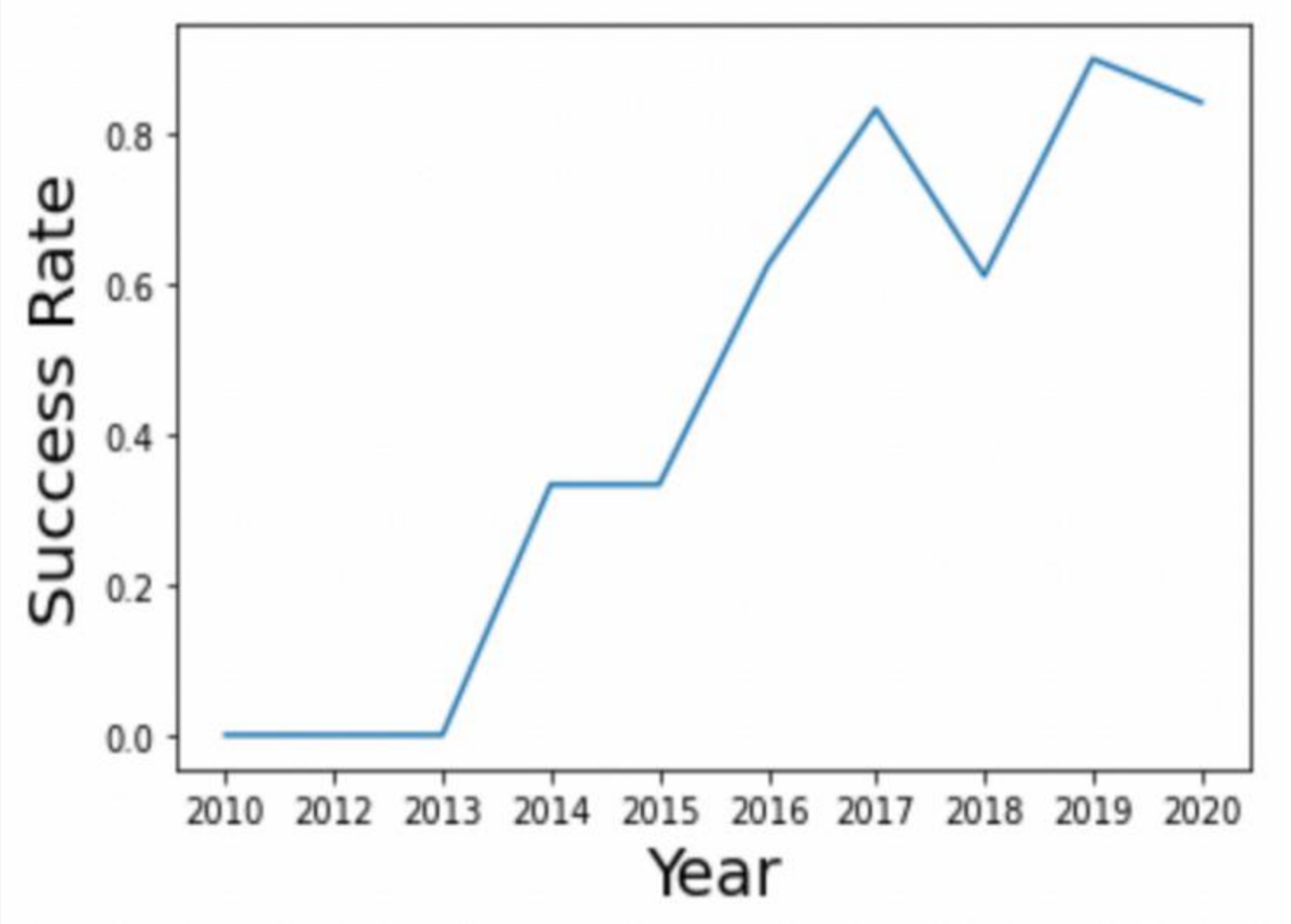
Payload vs. Orbit Type



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

Launch Success Yearly Trend

- The success rate since 2013 kept increasing till 2020.



All Launch Site Names

- 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'

```
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

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

Total Payload Mass

- 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

```
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

- 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.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[8]:

first_successful_landing
2015-12-22

- 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.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

- 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.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[10]:

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

- 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.bs2io90l08kqblod8lcg.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

- 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.bs2io90108kqb1od8lcg.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)

- 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.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

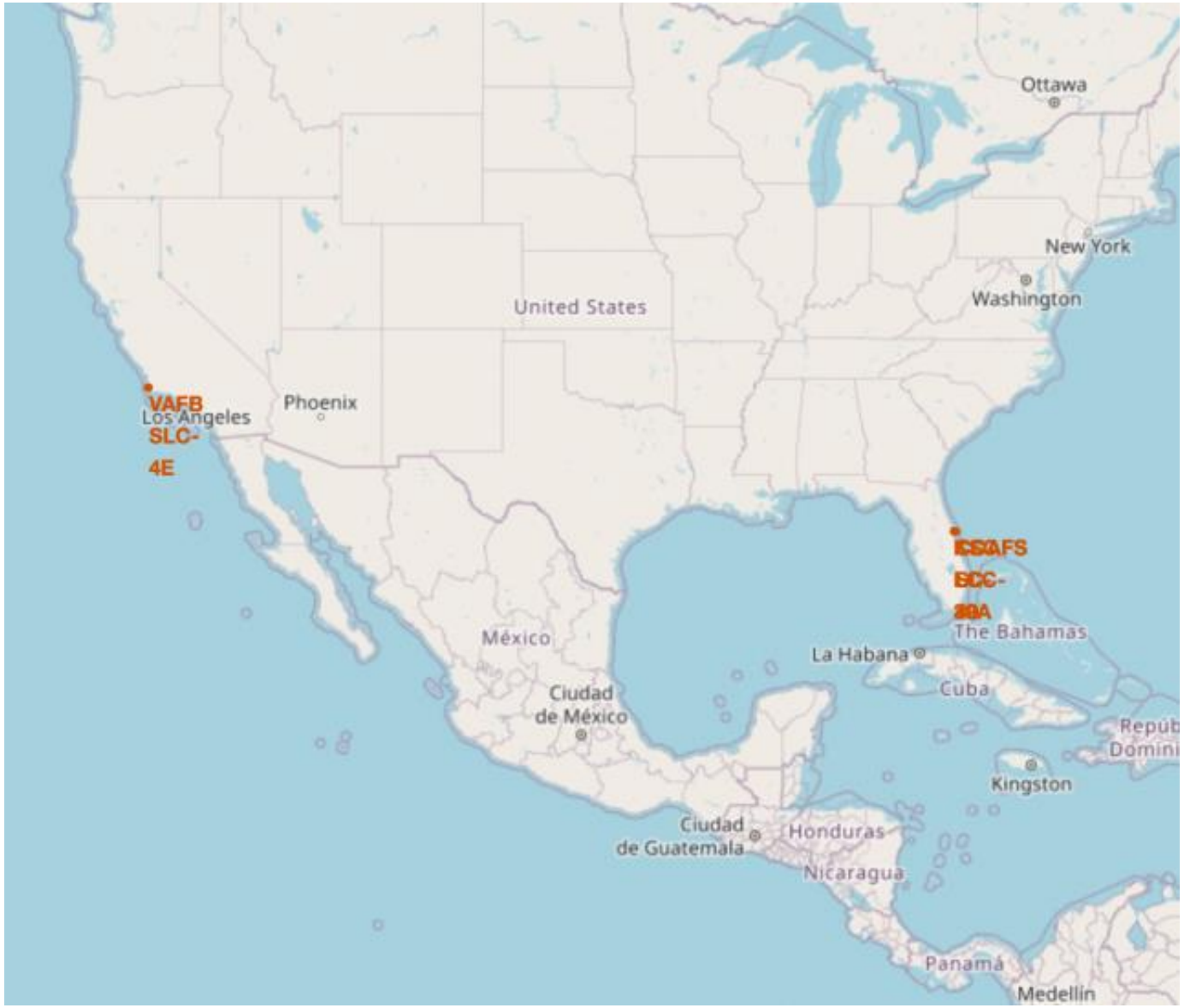
- Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20 in descending order.

Section 3

Launch Sites Proximities Analysis

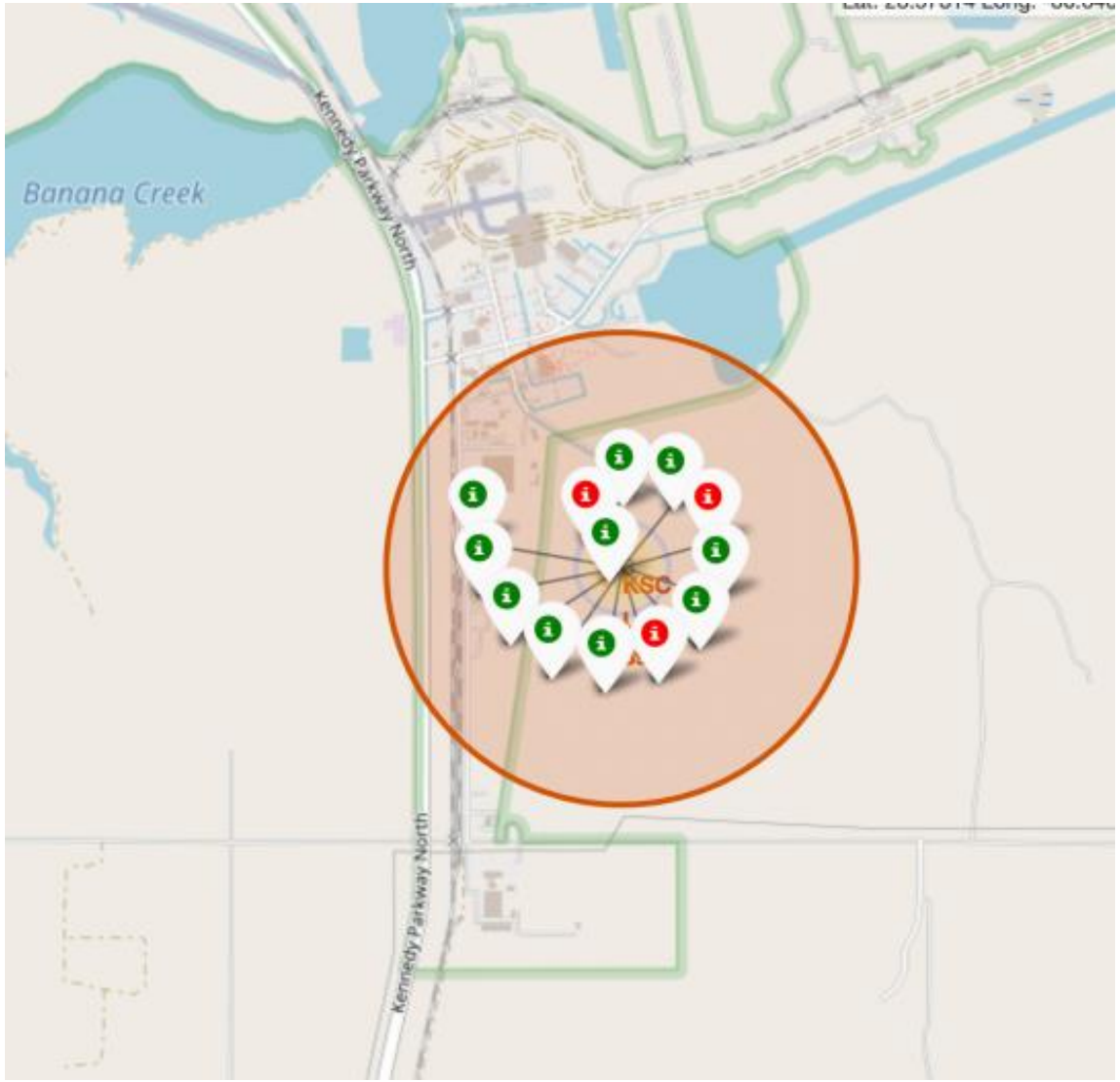
Map with all Launch Sites

- The majority of launch locations are close to the equator. The equator is the location on Earth's surface where the land moves the fastest. At the equator, everything on Earth's surface is already traveling at 1670 km/h. When a ship is launched from the equator, it travels through space at the same speed as before the launch while also orbiting the planet. Inertia is the cause of this.
- This velocity will assist the spacecraft in maintaining a sufficient speed to remain in orbit.
- Since all launch locations are so close to the coast, there is less chance of any debris falling or exploding close to humans when rockets are launched towards the ocean.



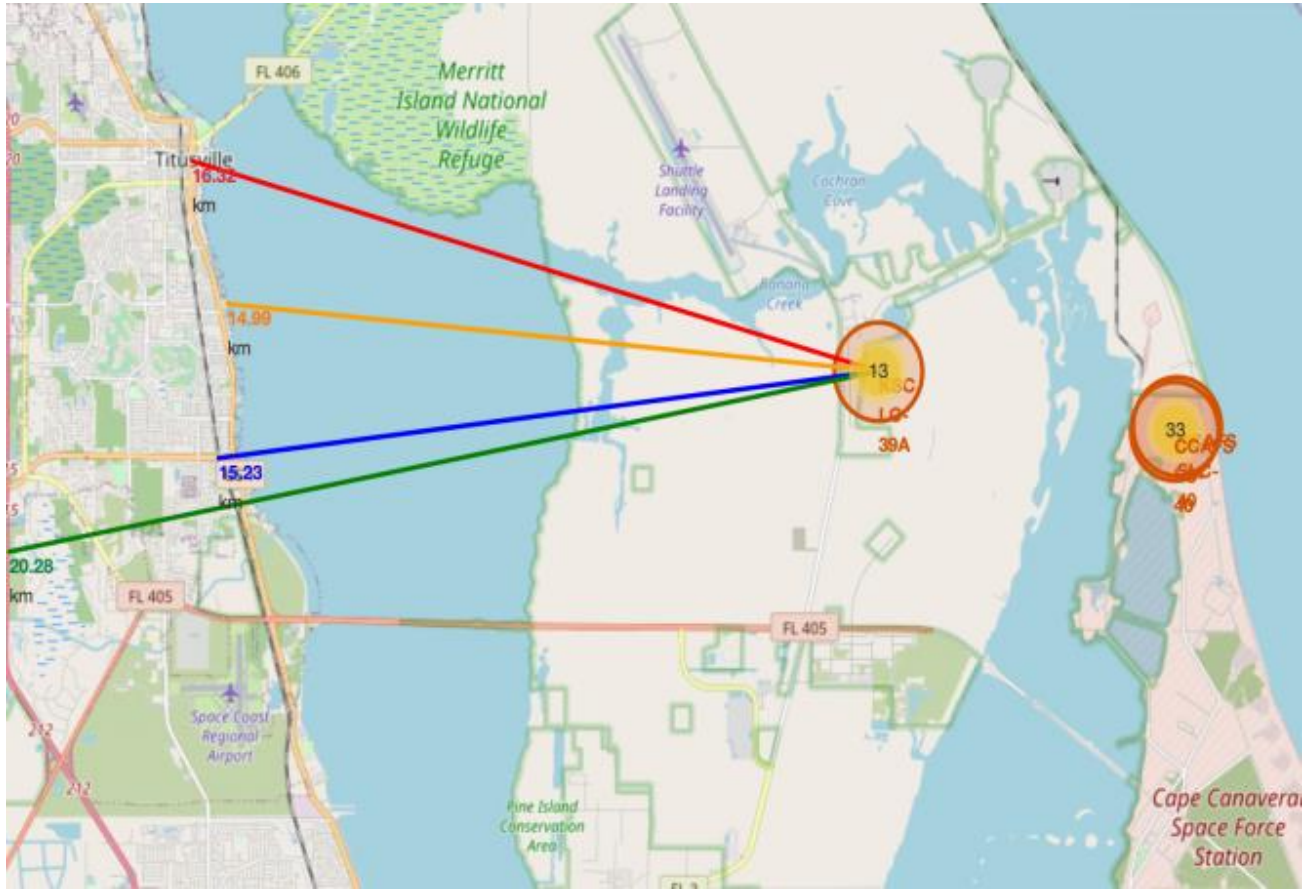
Map of Launch Records

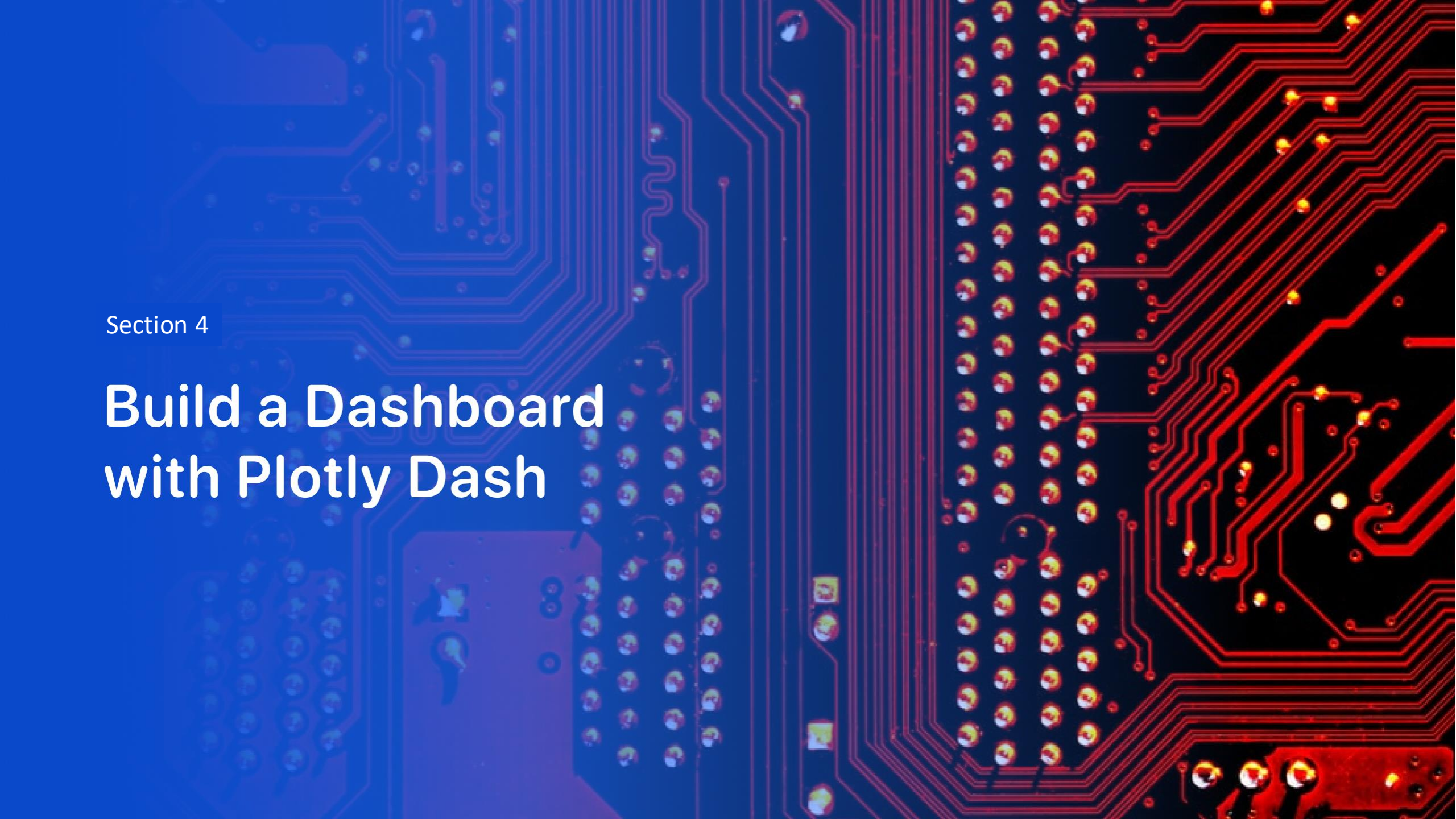
- From the color-labeled markers we should be able to easily identify which launch sites have relatively high success rates.
- Green represents success while red represents failure.
- Launch Site KSC LC-39A has a very high Success Rate.



Distance from the launch site KSC LC-39A to neighboring areas.

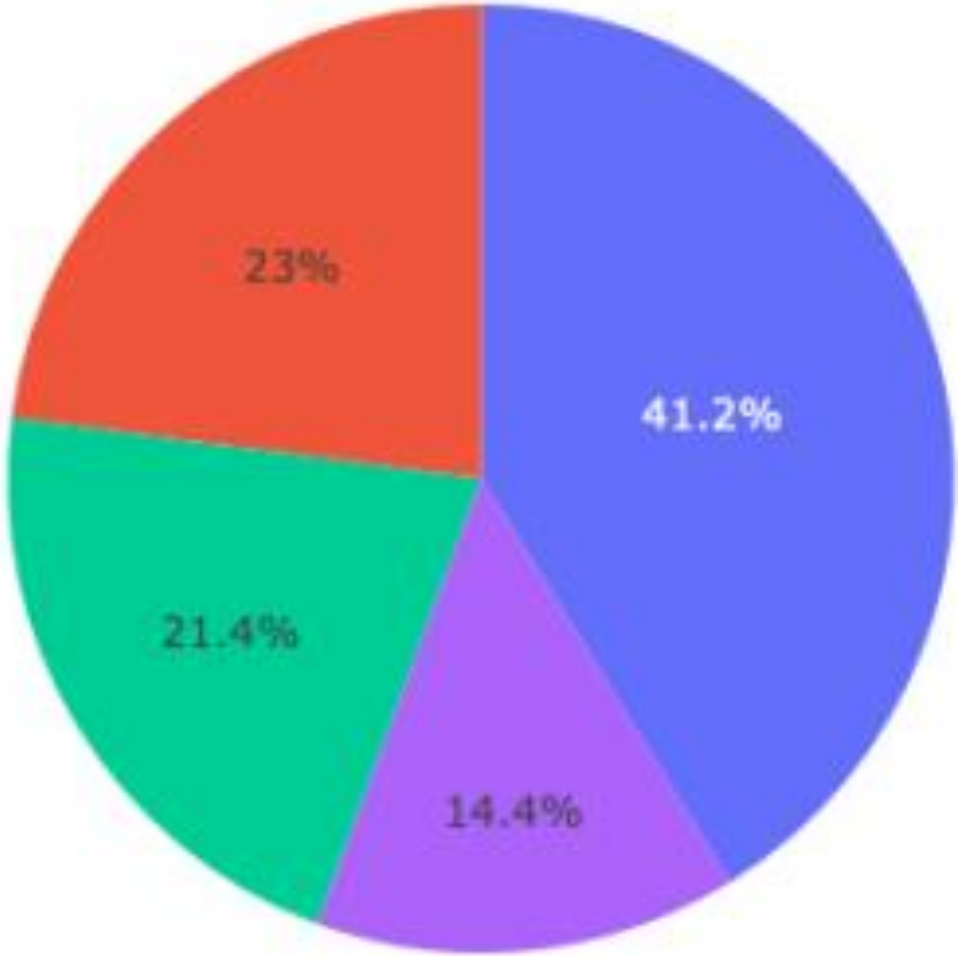
- From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:
  - relative close to railway (15.23 km)
  - relative close to highway (20.28 km)
  - relative close to coastline (14.99 km)
- Also the launch site KSC LC-39A is relative close to its closest city Titusville (16.32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



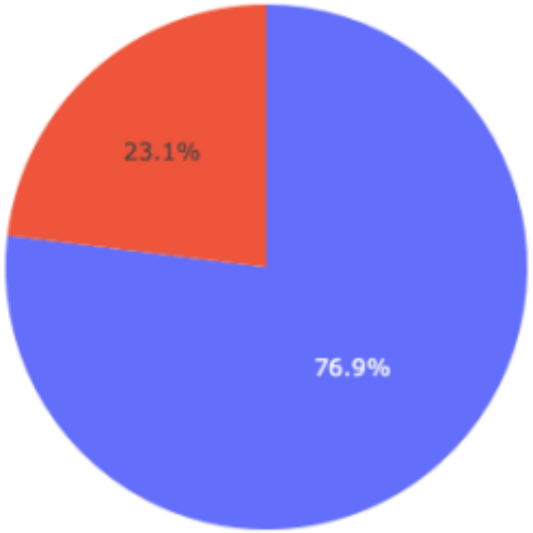

Section 4
Build a Dashboard with Plotly Dash

Success rates of sites

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



Total Success Launches for Site KSC LC-39A

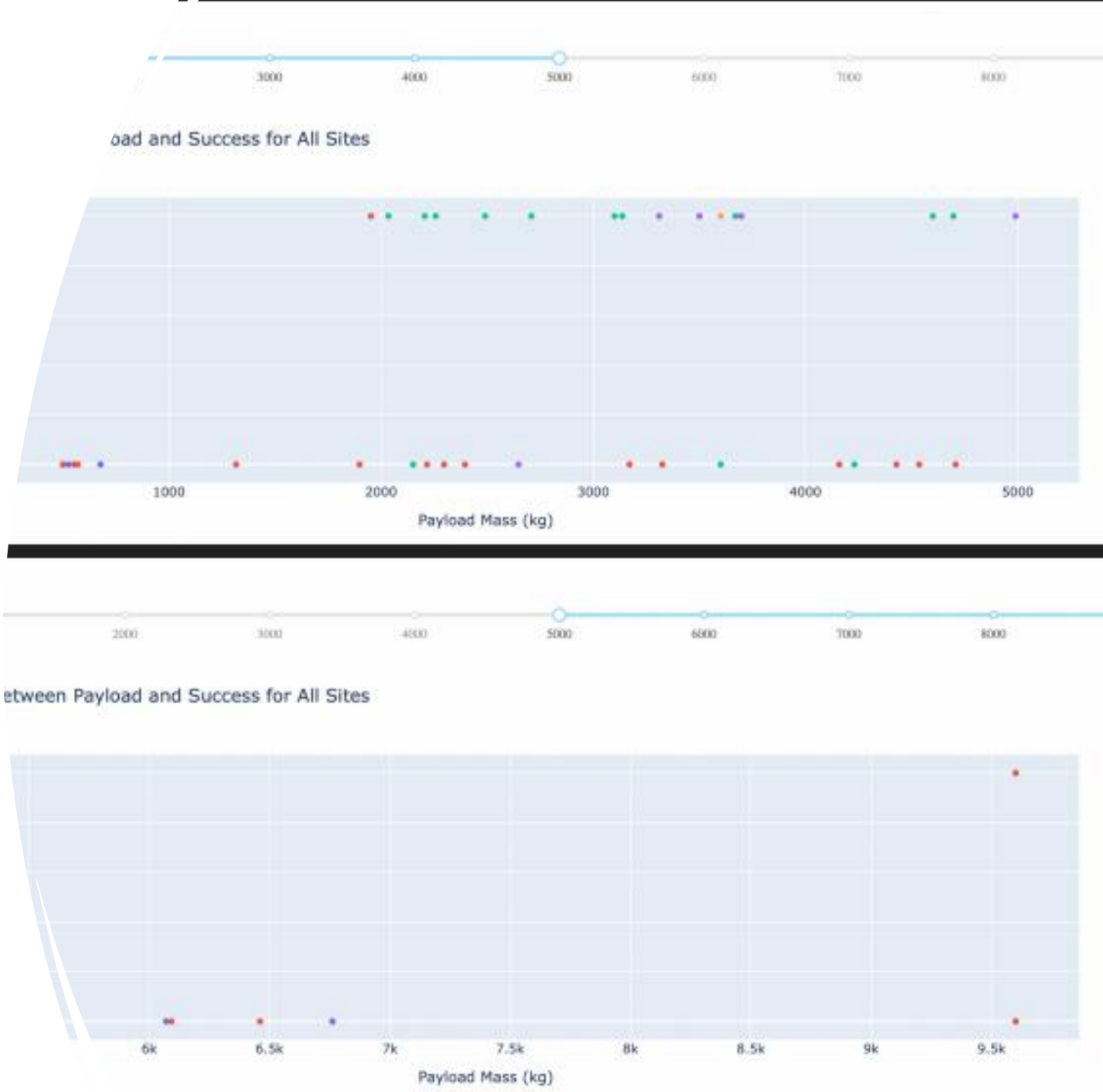


KSC LC-39A performance

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

Payload mass vs launch outcome

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



Section 5

Predictive Analysis (Classification)

Classification Accuracy

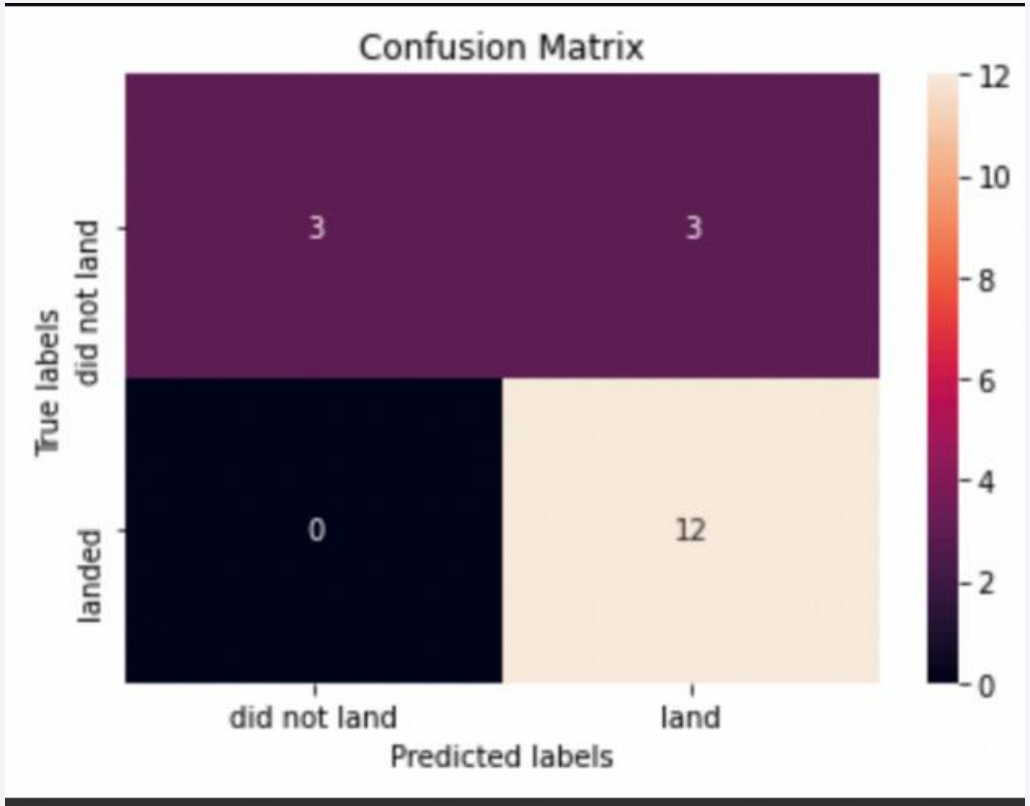
- Based on the scores of the Test Set, we cannot confirm which method performs best.
- Same Test Set scores may be due to the small test sample size (18 samples). Therefore, we tested all methods based on the whole Dataset.
- The scores of the whole Dataset confirm that the best model is the Decision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

	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

	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

- Decision Tree 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 Thanks to:
  - IBM
  - Coursera
  - Instructors

Thank you!