

Applied Data Science Capstone Project

ABHISHEK KUMAR



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

Questions to be answered

Methodology

Data collection methodology:

- Using SpaceX Rest API
- Using Web Scrapping from Wikipedia

Performed data wrangling

- Filtering the data
- Dealing with missing values
- Using One Hot Encoding to prepare the data to a binary classification

Performed exploratory data analysis (EDA) using visualization and SQL

Performed interactive visual analytics using Folium and Plotly Dash

Performed predictive analysis using classification models

- Building, tuning and evaluation of classification models to ensure the best results

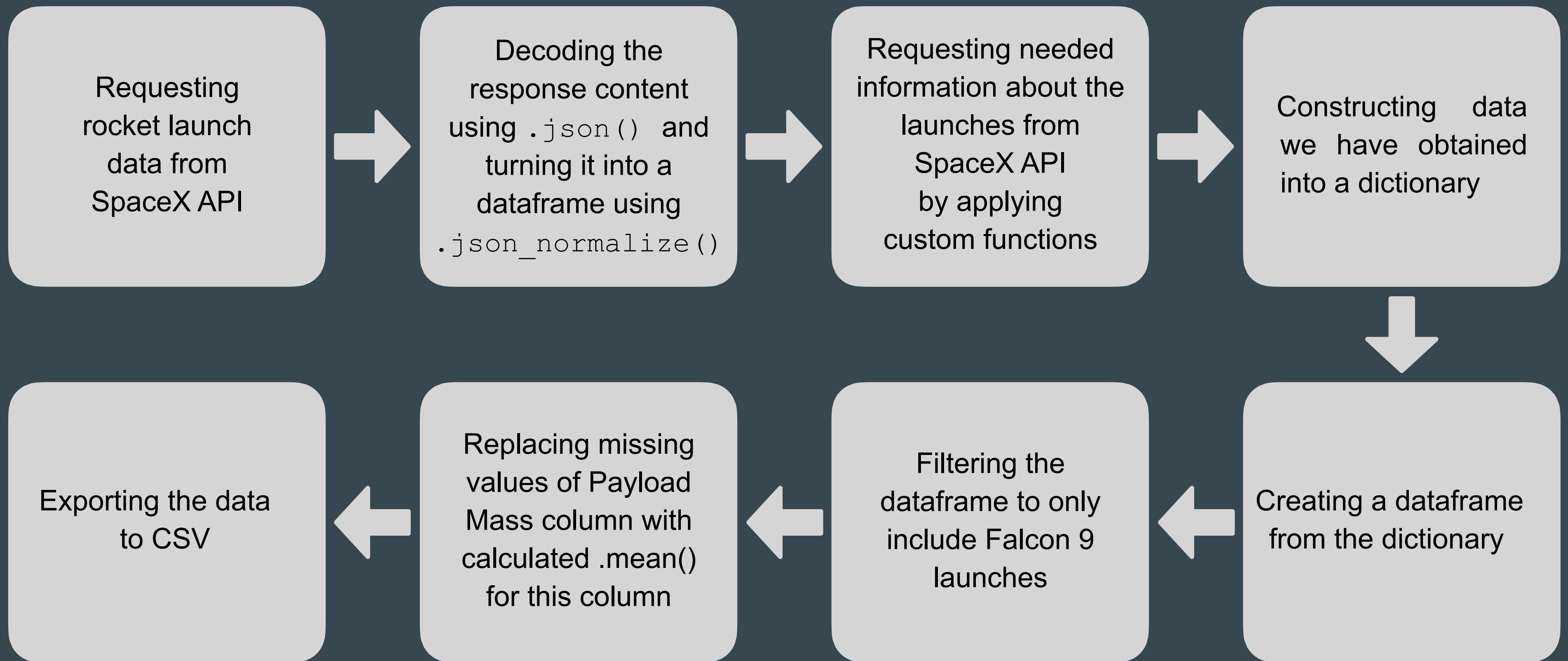
Data Collection

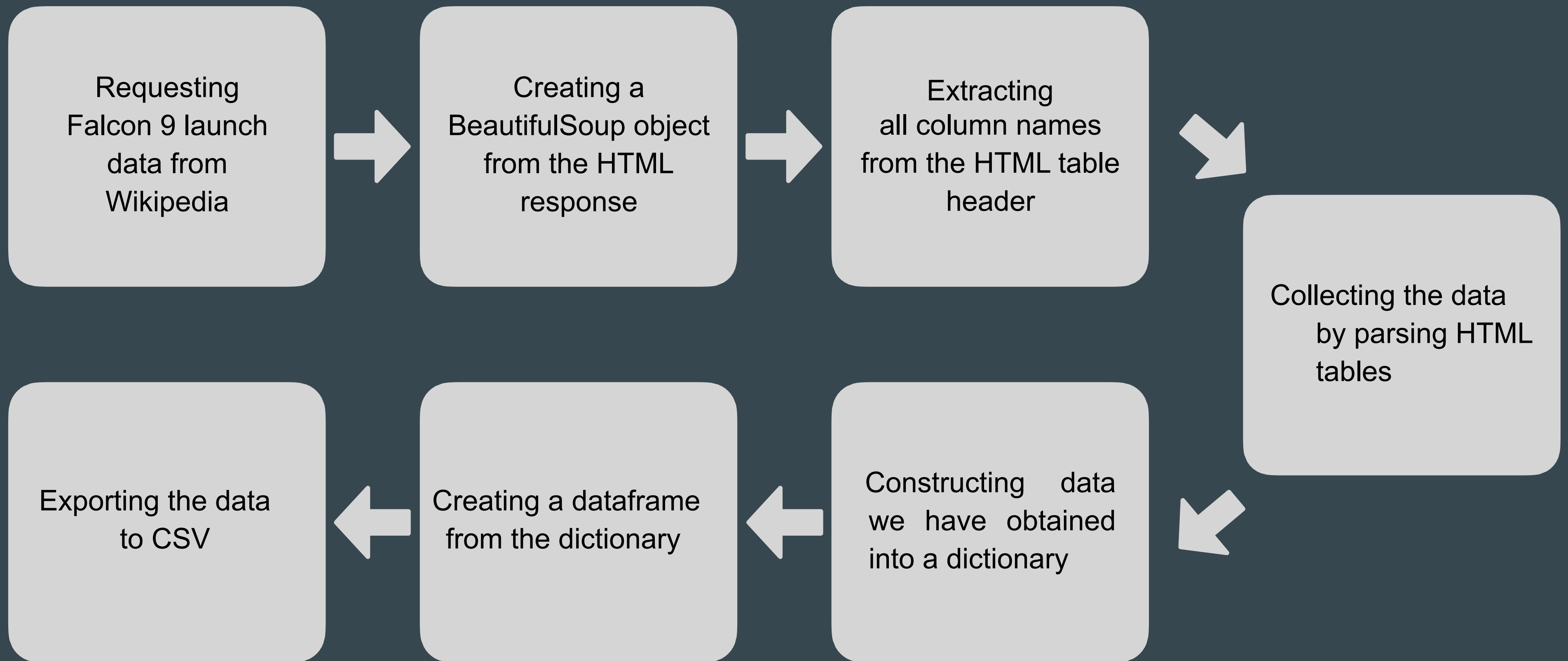
Data Columns are obtained by using SpaceX REST API:

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

Data Columns are obtained by using Wikipedia Web Scraping:

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





Data Wrangling

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

We mainly convert those outcomes into Training Labels with “1” means the booster successfully landed, “0” means it was unsuccessful.

Perform exploratory Data Analysis
and determine Training Labels



Calculate the number of launches
on each site

Calculate the number and occurrence of
each orbit

Calculate the number and occurrence of
mission outcome per orbit type

Create a landing outcome label
from Outcome column

Exporting the data
to CSV

EDA WITH DATA VISUALIZATION

Charts 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 and 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).

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 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have 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

FOLIUM MAP

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.

DASHBOARD - PLOTTY 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.

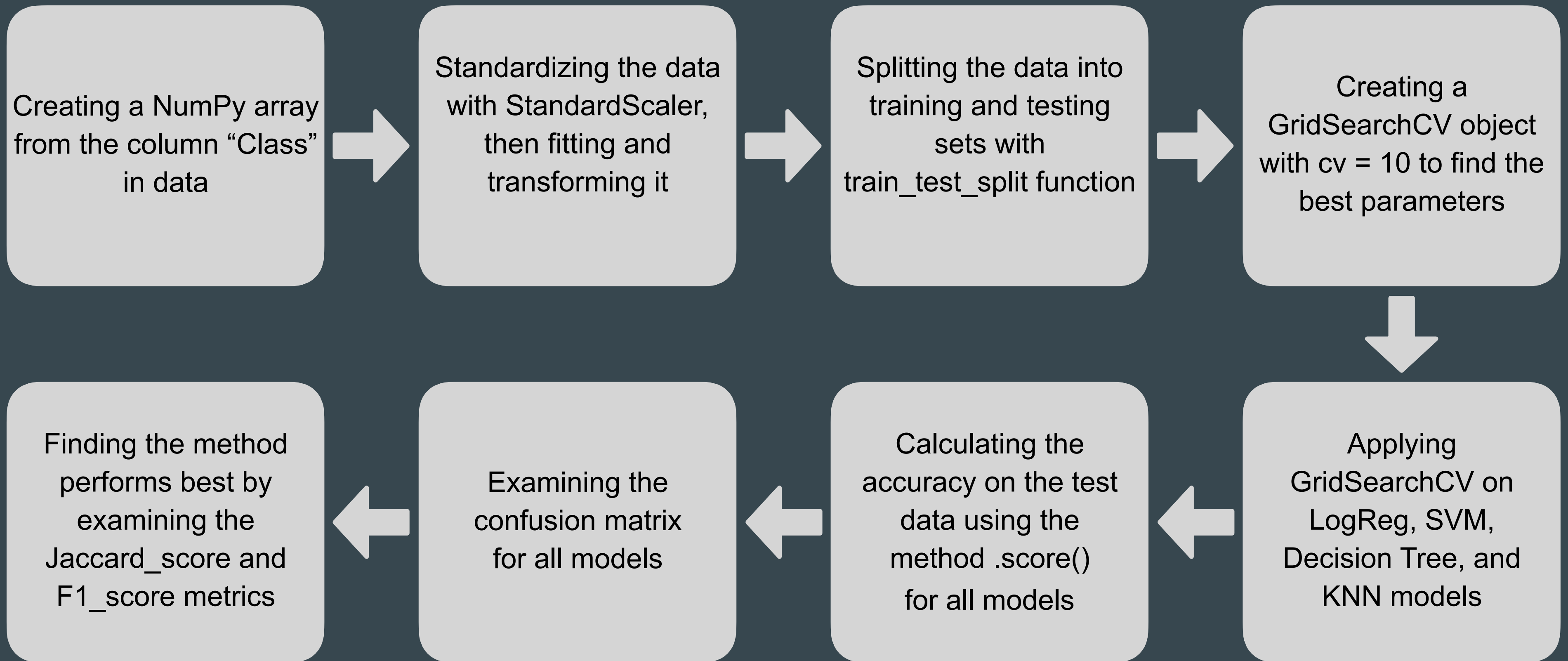
Slider of Payload Mass Range:

- 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

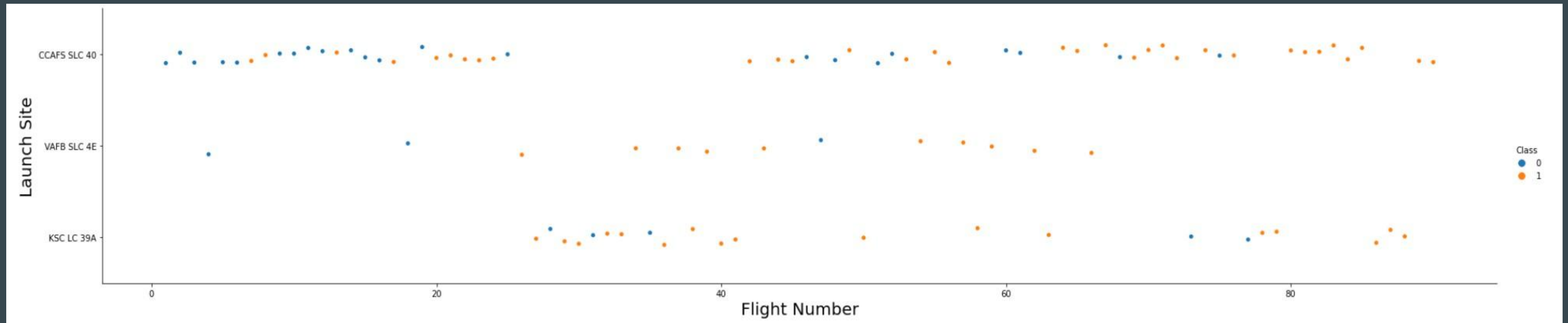


[GitHub URL: Machine Learning Prediction](#)

Results

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

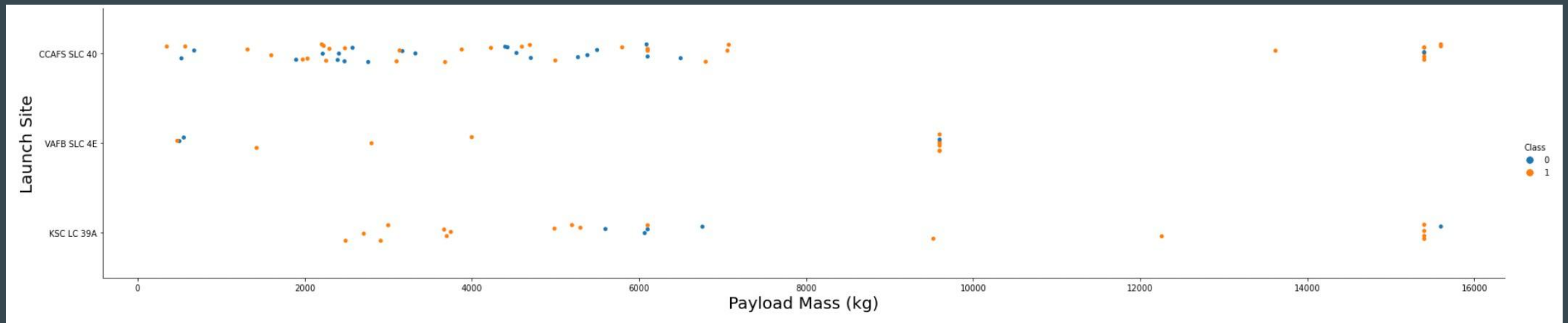
Flight No. VS Launch Sites



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 Sites



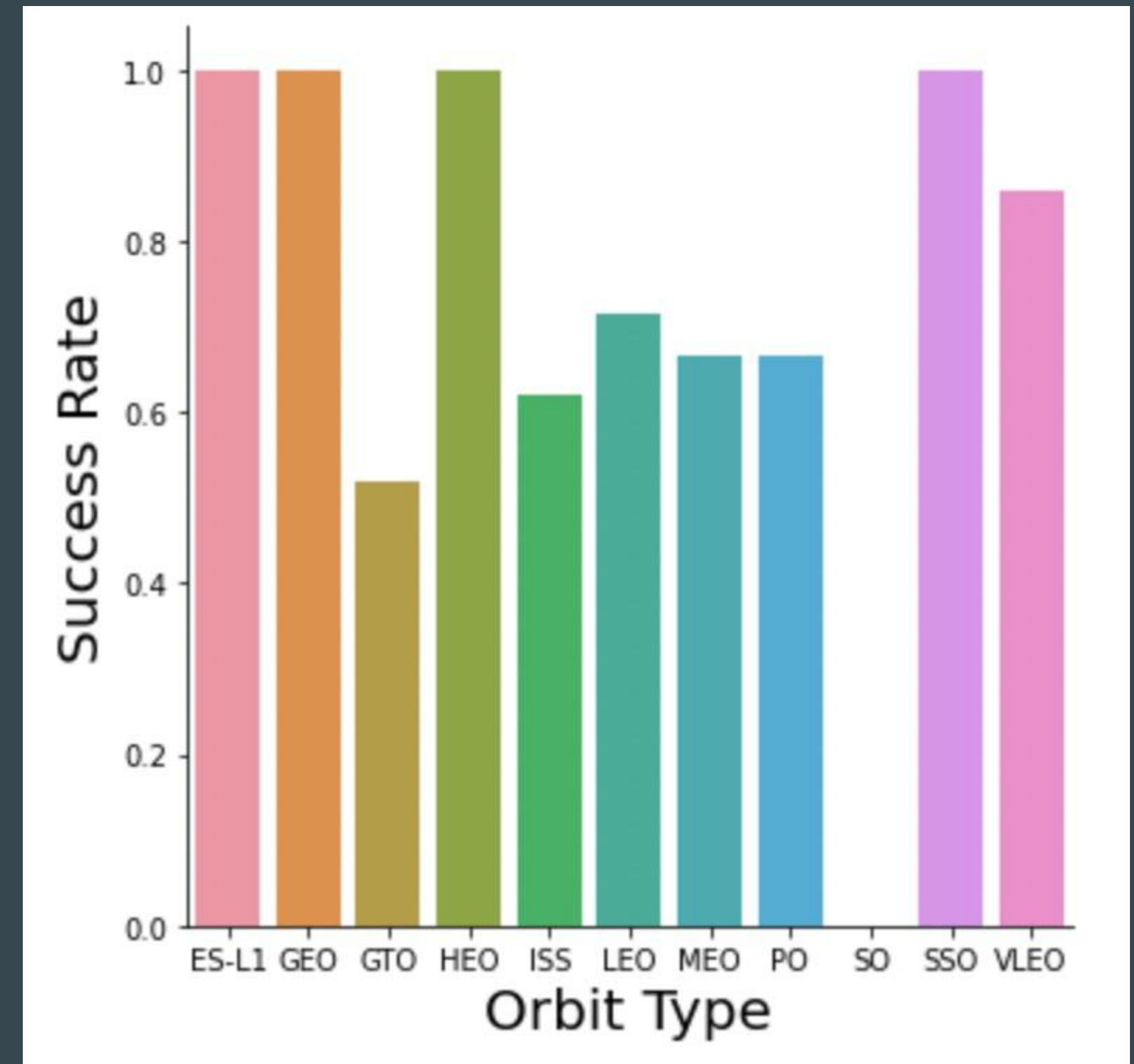
Explanation:

- 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.

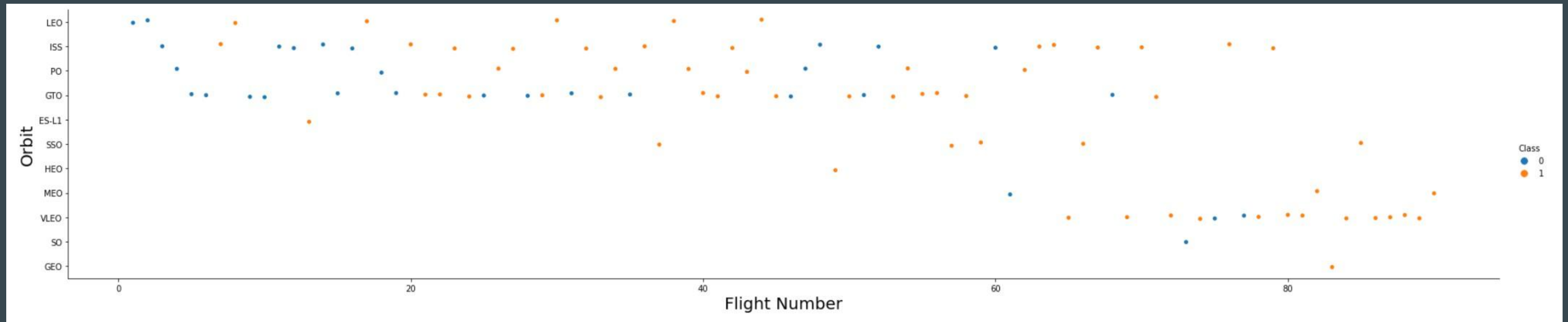
Success Rate VS Orbit

Explanation:

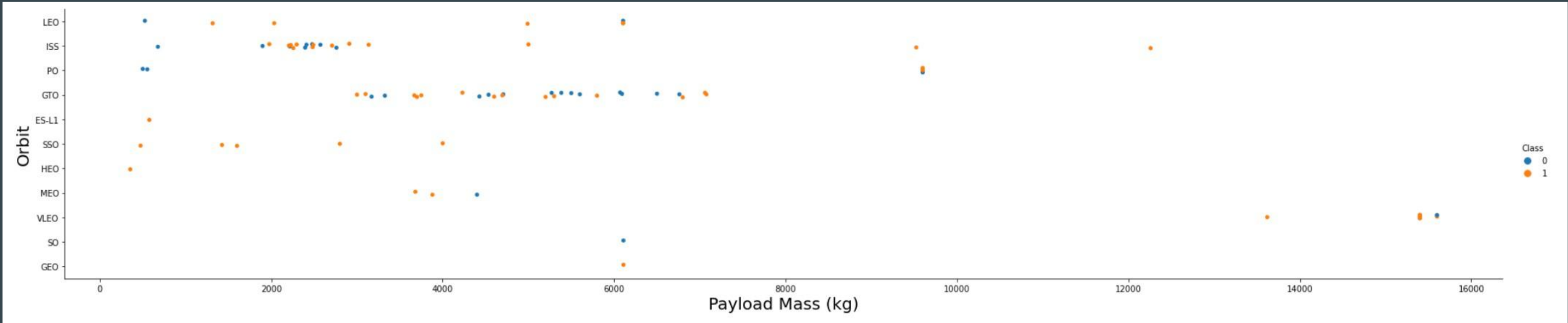
- Orbits with 100% success rate:
 - ES-L1, GEO, HEO, SSO
- Orbits with 0% success rate:
 - SO
- Orbits with success rate between 50% and 85%:
 - GTO, ISS, LEO, MEO, PO, VLEO



Flight No. VS Orbit



Payload Mass VS Orbit



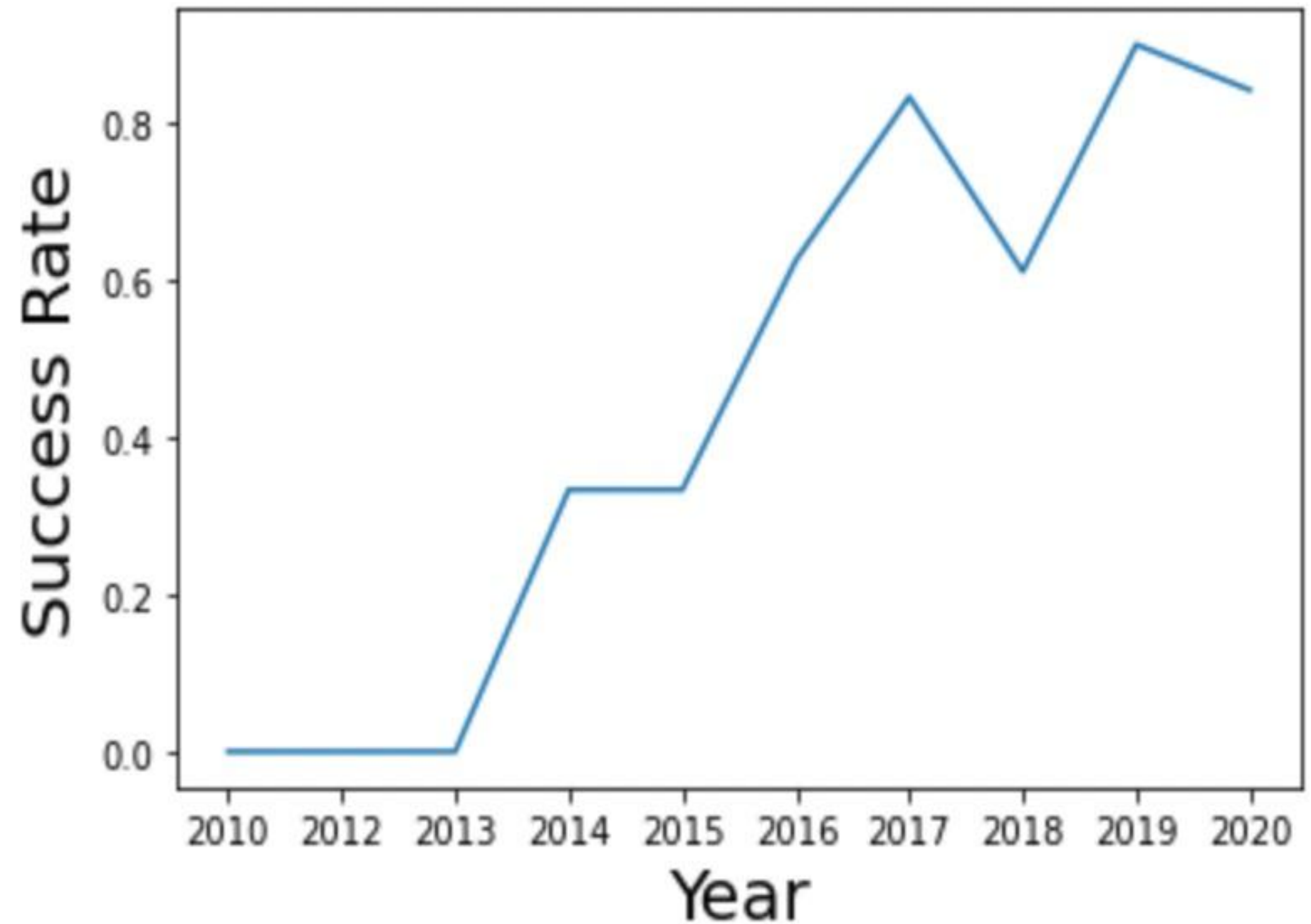
Explanation:

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

Launch Success Trends

Explanation:

- The success rate since 2013 kept increasing till 2020.



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 the unique launch sites in the space mission.

Launch sites with names starting with 'CCA'

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

```
* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqblod8l1cg.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.bs2io90108kqblod81cg.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).

AVG. PAyload MASS (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.bs2io90l08kqblod8lcg.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 Landing

```
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.bs2io90108kqblod8lcg.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 (Payload ->[4000-6000])

```
In [9]: %sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000;
```

```
* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod81cg.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.

Mission Outcome

```
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.bs2io90108kqb1od8lcg.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.

Max Payload - Booster

```
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.bs2io90108kqb1od81cg.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.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) |

Explanation:

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

Success Count Comparison

```
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.bs2io90108kqb1od8lcg.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 |

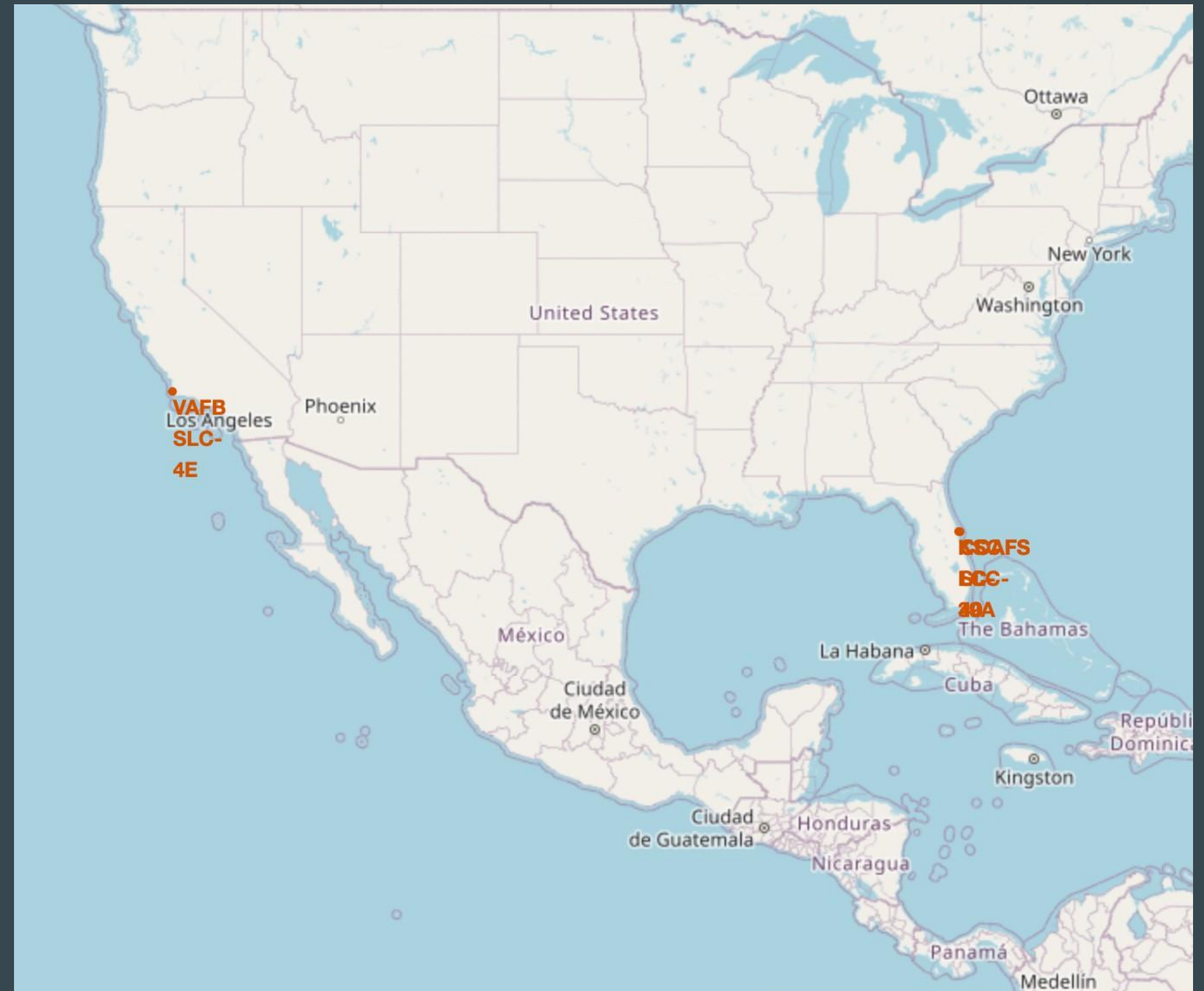
Explanation:

- 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.

Launch Sites on Map

Explanation:

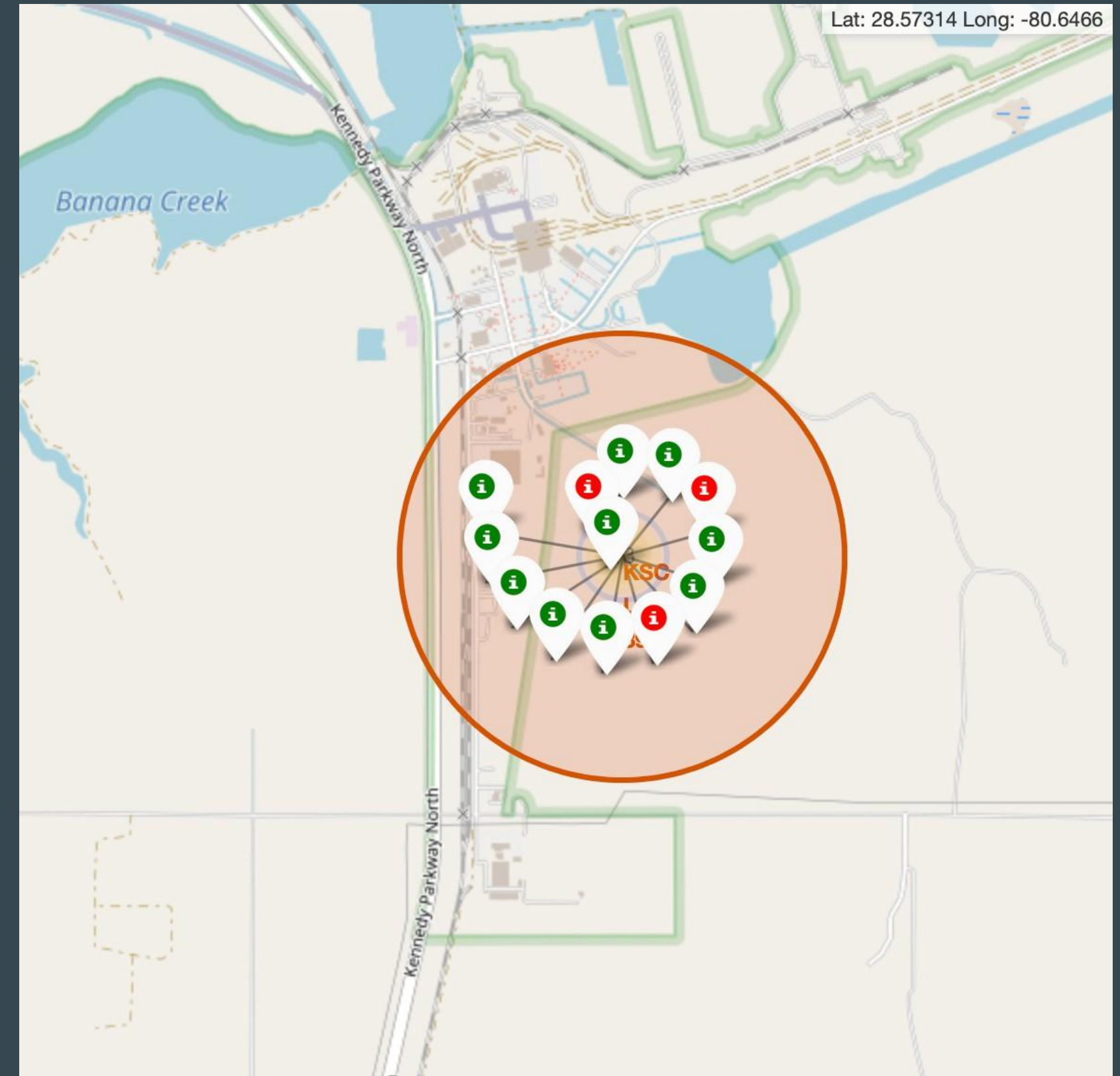
- Most of Launch sites are in proximity to the Equator line. The land is moving faster at the equator than any other place on the surface of the Earth. Anything on the surface of the Earth at the equator is already moving at 1670 km/hour. If a ship is launched from the equator it goes up into space, 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.



Labelled records on Map

Explanation:

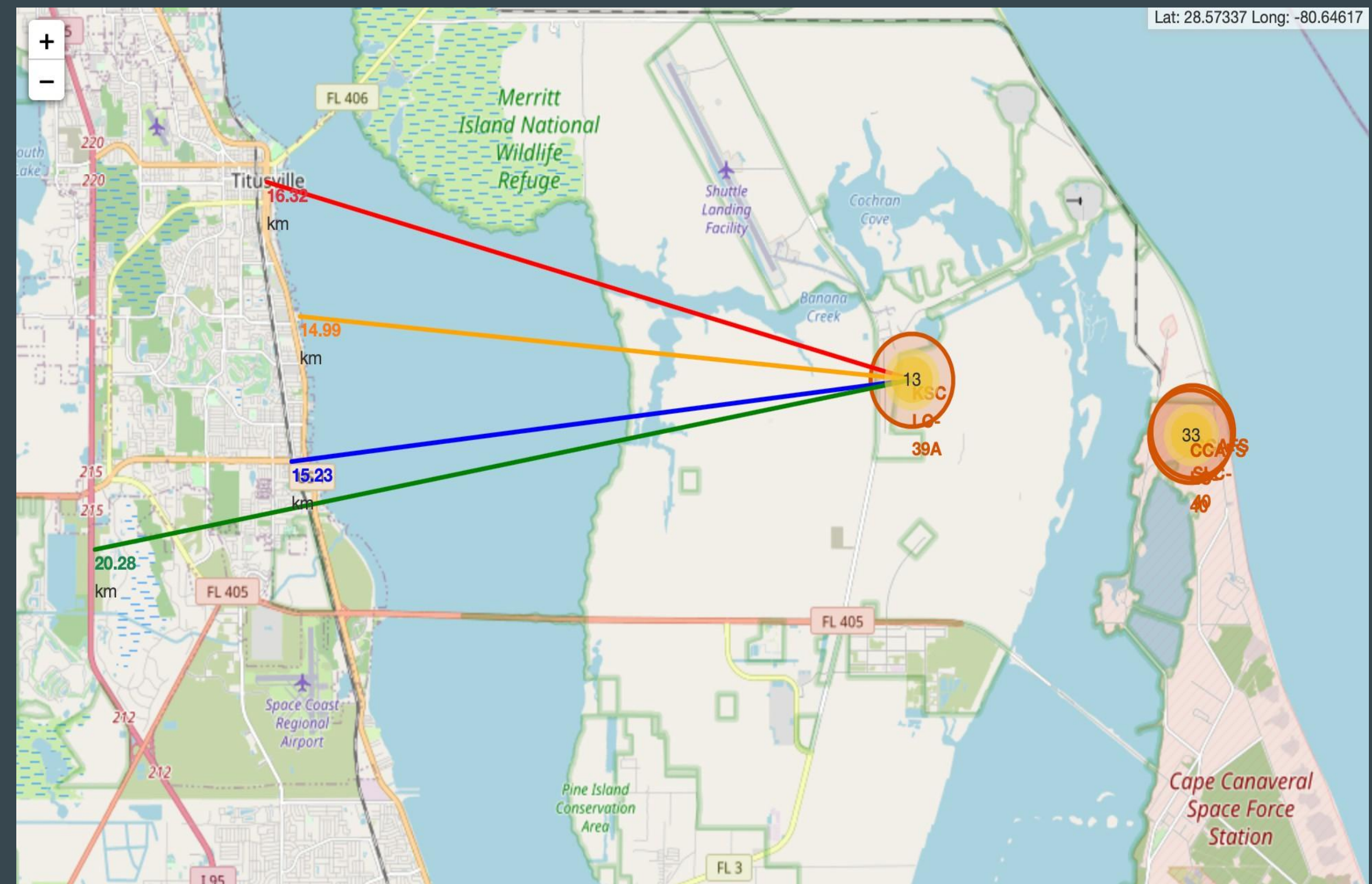
- From the colour-labeled markers we should be able to easily identify which launch sites have relatively high success rates.
 - **Green Marker** = Successful Launch
 - **Red Marker** = Failed Launch
- Launch Site KSC LC-39A has a very high Success Rate.



Distance from KSC LC-39A

Explanation:

- 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.



Launch Success Count

Total Success Launches by Site



Explanation:

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

Launch Site Success Ratio

Total Success Launches for Site KSC LC-39A



Explanation:

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

Payload VS Luanch Outcome

Explanation:

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



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 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

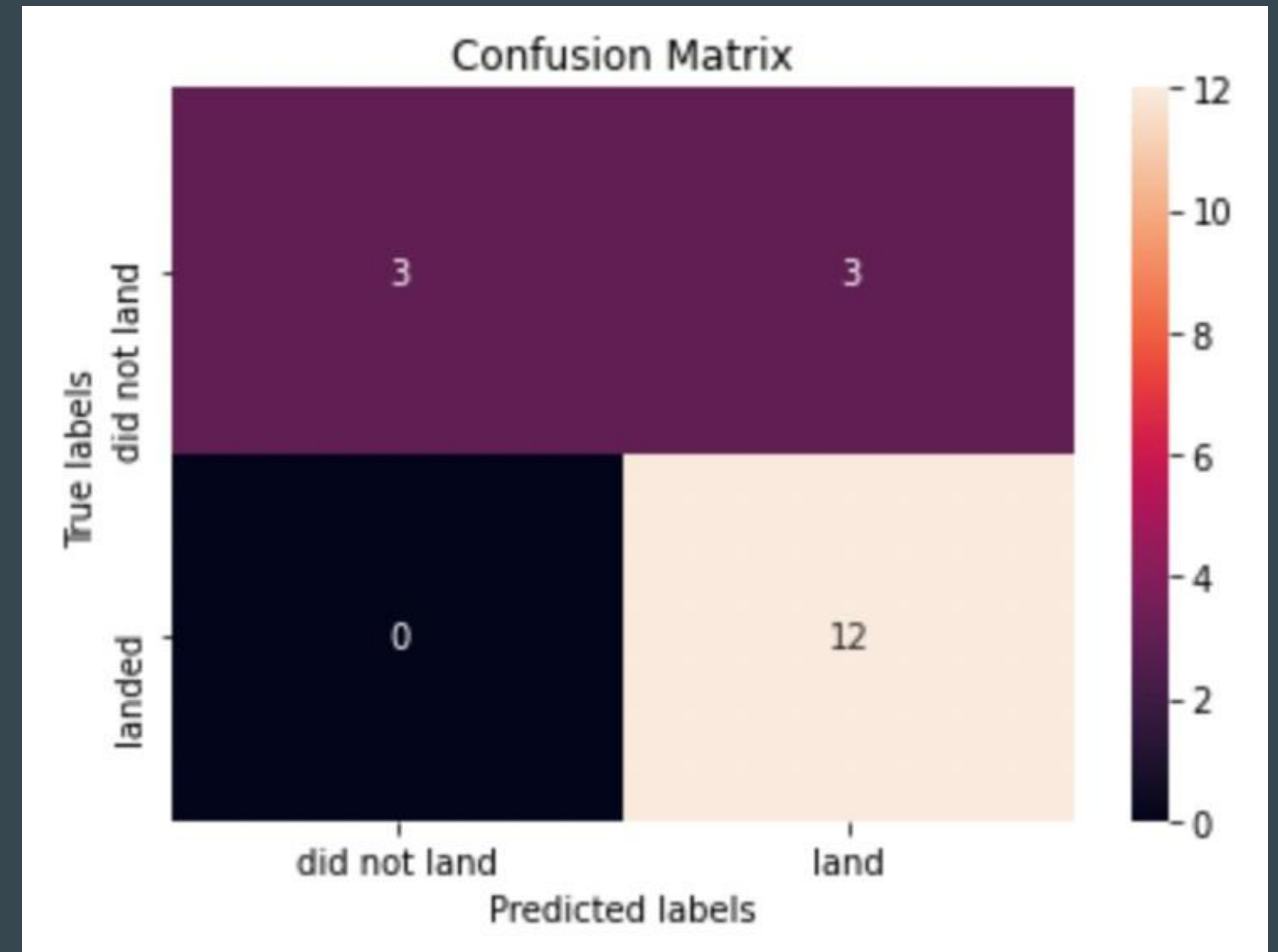
| | LogReg | SVM | Tree | KNN |
|----------------------|----------|----------|----------|----------|
| Jaccard_Score | 0.833333 | 0.845070 | 0.882353 | 0.819444 |
| F1_Score | 0.909091 | 0.916031 | 0.937500 | 0.900763 |
| Accuracy | 0.866667 | 0.877778 | 0.911111 | 0.855556 |

Confusion Matrix

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.

| | | Predicted Values | |
|---------------|----------|------------------|----------|
| | | Negative | Positive |
| Actual Values | Negative | TN | FP |
| | Positive | FN | TP |



Conclusion

- 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.