

Winning Space Race with Data Science

Dr. Raj Kumar
21-06-2022



Outline

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



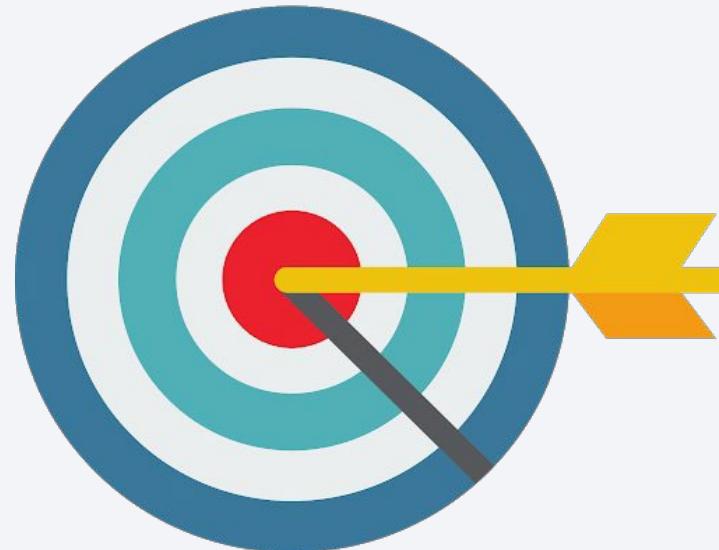
Executive Summary

Summary of methodologies

- ❖ Data collection
- ❖ Data wrangling
- ❖ EDA with SQL
- ❖ EDA with data visualization
- ❖ 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

SpaceX is the most successful company of the commercial space age, making space travel affordable. The company advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage

Problems you want to find answers

- How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
- Does the rate of successful landings increase over the years?
- What is the best algorithm that can be used for binary classification in this case?

Section 1

Methodology

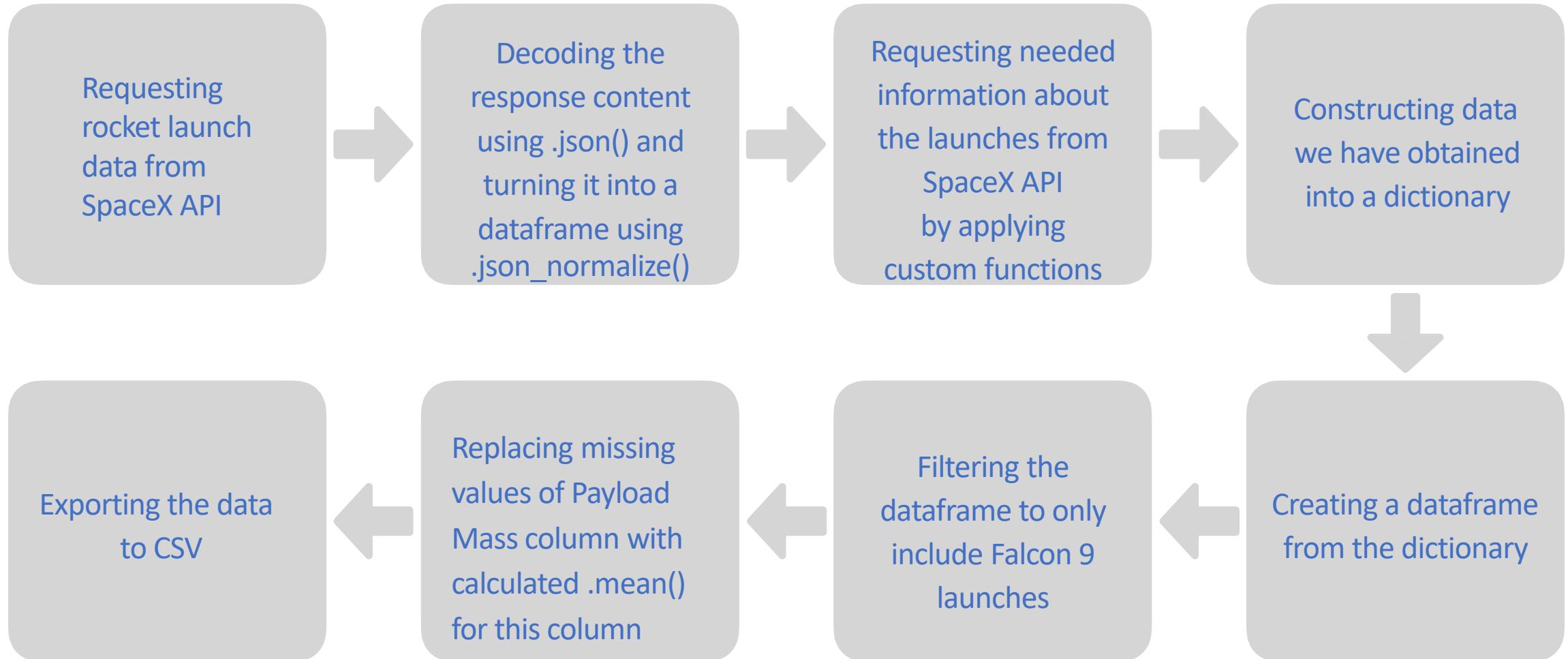
Methodology

Executive Summary

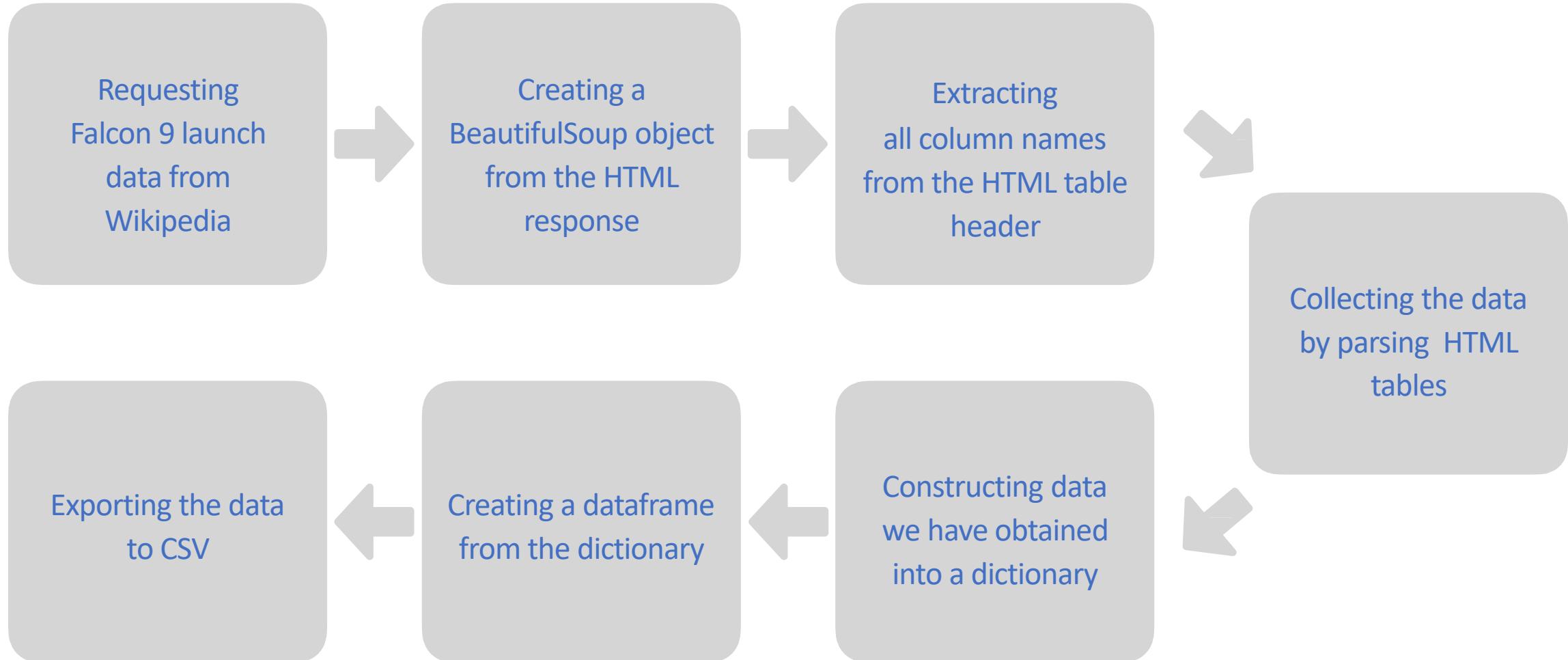
- ❖ Data collection methodology:
 - ✓ Using SpaceX Rest API
 - ✓ Using Web Scrapping from Wikipedia
- ❖ Performed data wrangling
 - ✓ Filtering the data
 - ✓ Dealing with missing values
 - ✓ Using One Hot Encoding to prepare the data to a binary classification
- ❖ Performed exploratory data analysis (EDA) using visualization and SQL
 - ✓ Plotting : Scatter Graphs, Bar Graphs to show relationships between variables to show patterns of data.
- ❖ Performed interactive visual analytics using Folium and Plotly Dash
- ❖ Performed predictive analysis using classification models
 - ✓ How to build, tune, evaluate classification models



Data collection – SpaceX API



Data collection – scraping



Data wrangling

Introduction

There are several different cases where the booster did not land successfully

- True Ocean means the mission outcome was successfully landed to a specific region of the ocean.
- False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean.

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

Process

Perform Exploratory Data Analysis EDA on dataset

Calculate the number of launches at each site

Calculate the number and occurrence of each orbit

Calculate the number and occurrence of mission outcome per orbit type

Export dataset as .CSV

Create a landing outcome label from Outcome column

Work out success rate for every landing in dataset

Each launch aims to an dedicated orbit, and here are some common orbit types:

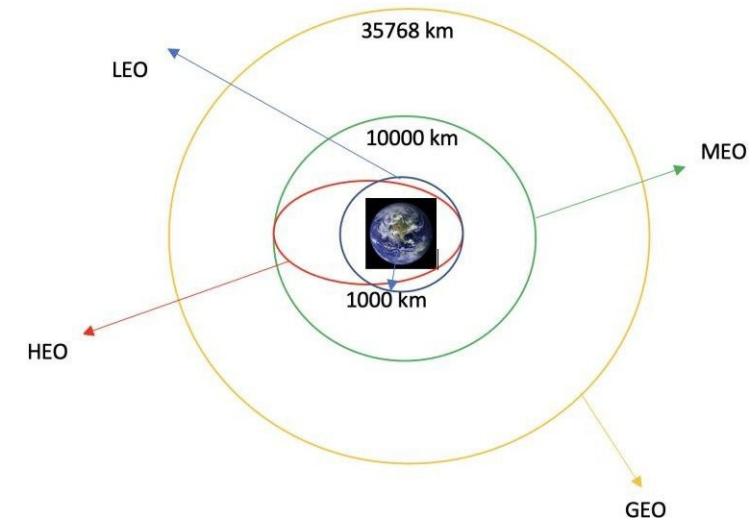


Diagram showing common orbit types SpaceX uses

EDA with Data Visualization

- ✓ 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

- Load dataset into DB2 database.
- Queried using sql python integration.
- Findout many questions given in hand on lab.
- Queried information about the launch site name, mission outcomes, various payload size of customer and booster versions and landing outcomes.

Build an Interactive Map with Folium

Markers of all launch sites:

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

Coloured Markers of the launch outcomes for each Launch Site:

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

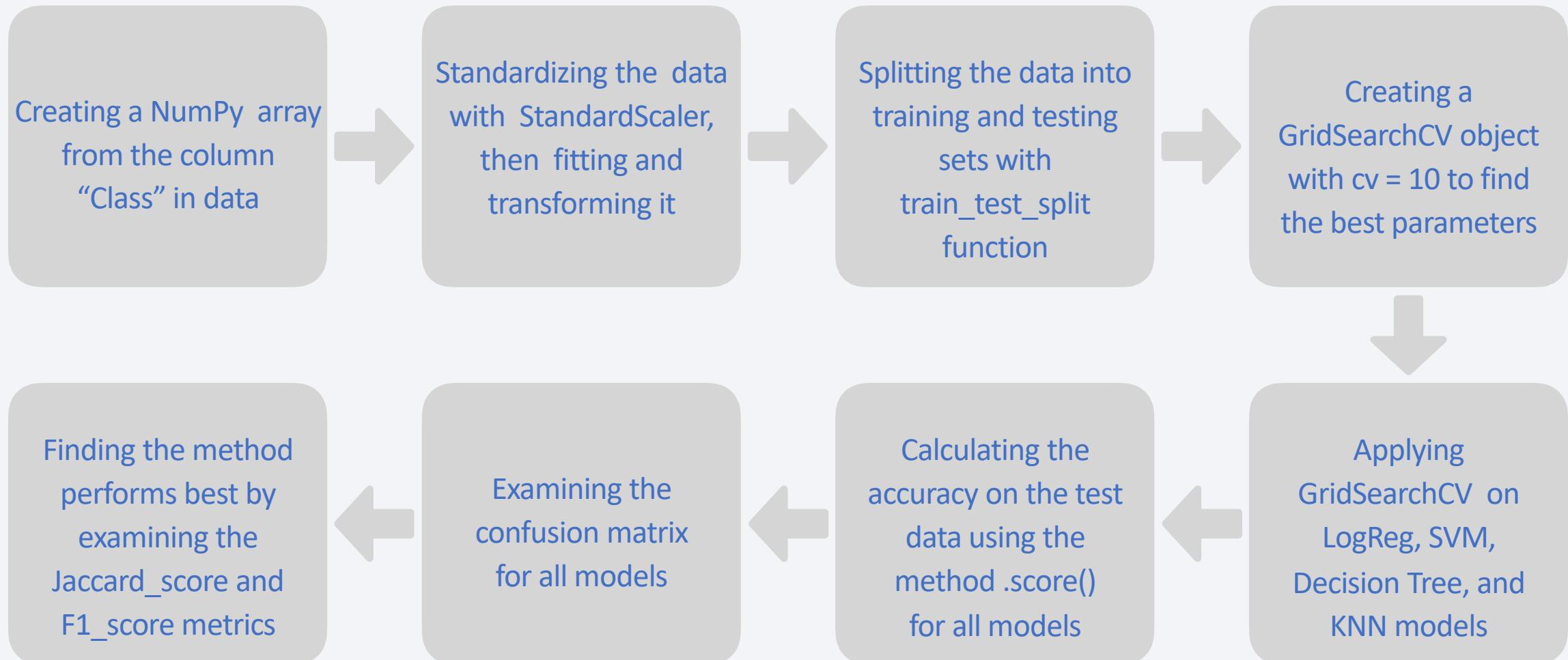
Distances between a Launch Site to its proximities:

- Added coloured Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City.

Build a Dashboard with Plotly Dash

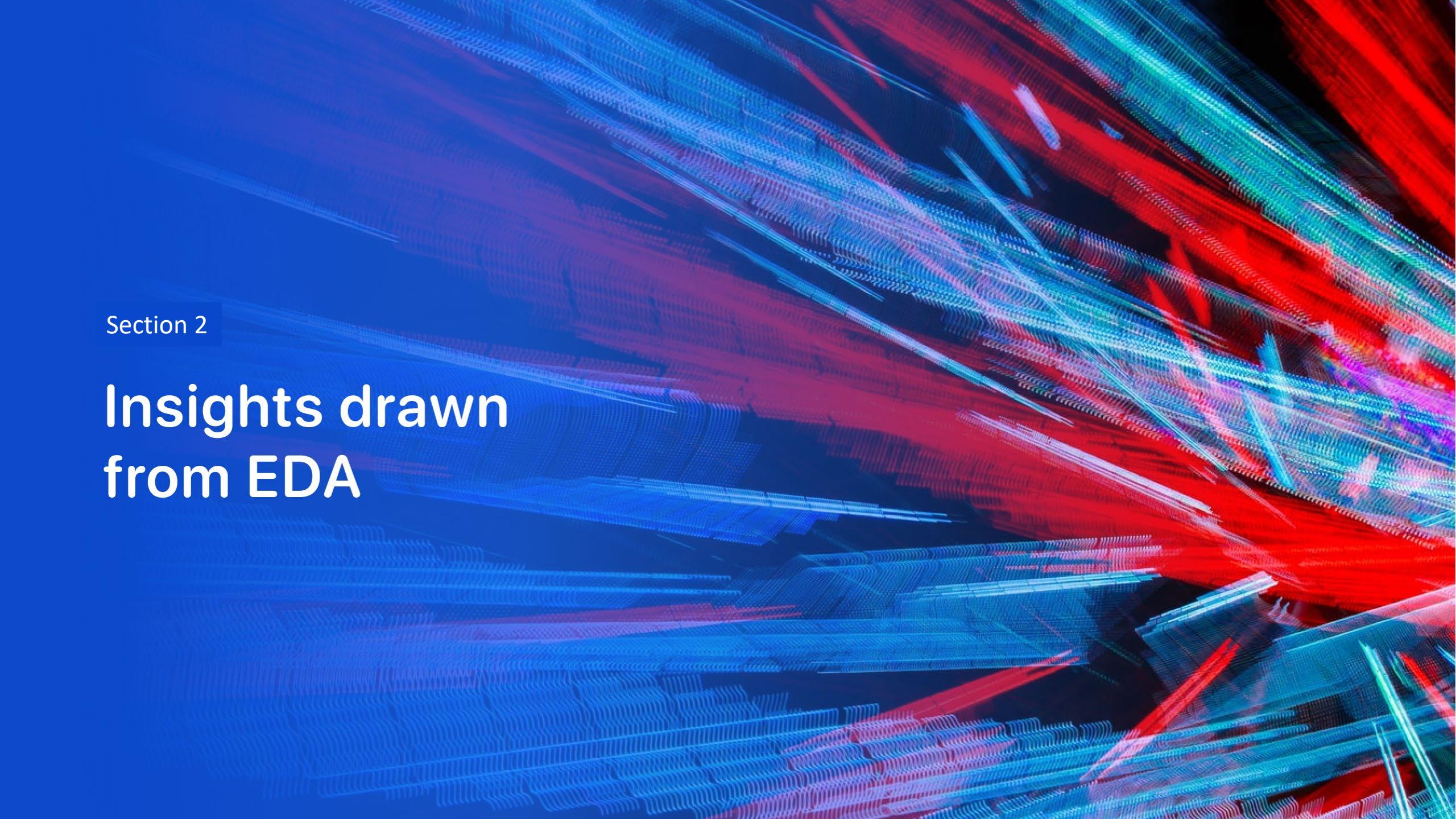
- **Launch sites dropdown lists:**
 - Added a dropdown list to enable Launch Site selection.
- **Pie Chart showing Success Launches (All Sites/Certain Site):**
 - Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific 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

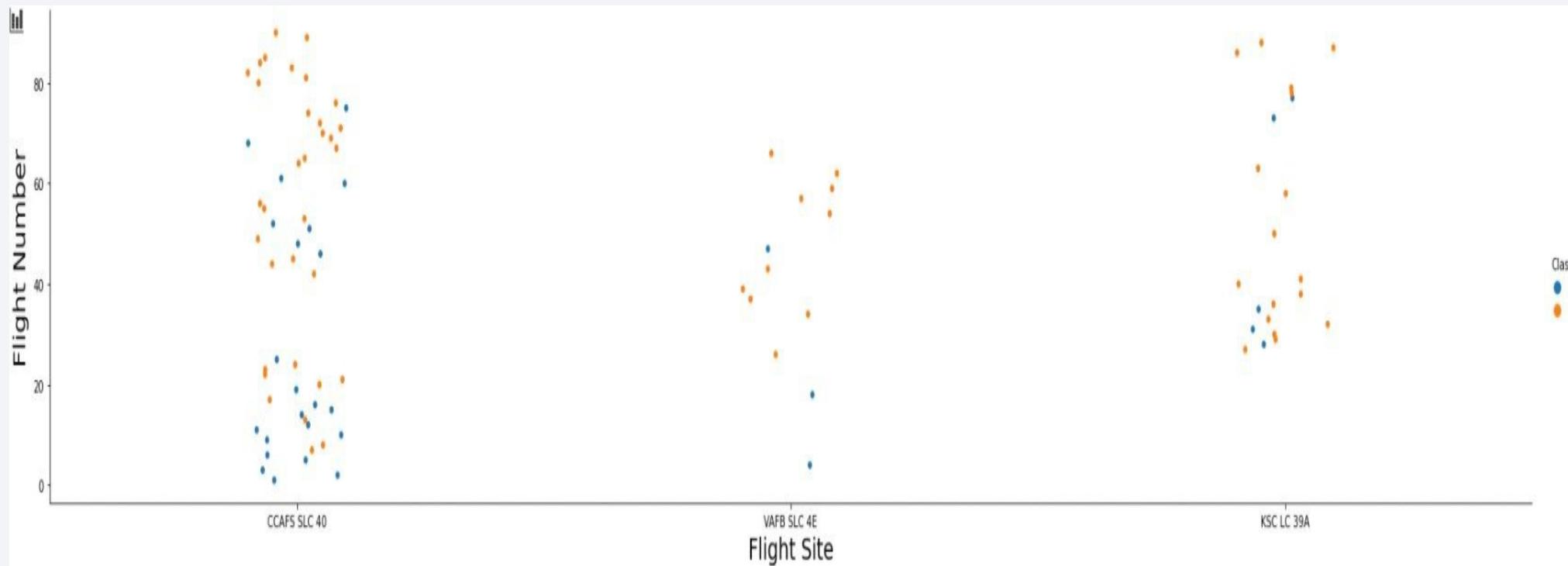
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

The background of the slide features a complex, abstract digital visualization. It consists of numerous thin, glowing lines that create a sense of depth and motion. The lines are primarily blue and red, with some green and purple highlights. They form a grid-like structure that curves and twists across the frame, resembling a 3D wireframe or a network of data points. The overall effect is futuristic and dynamic, suggesting concepts like data flow, digital communication, or complex systems.

Section 2

Insights drawn from EDA

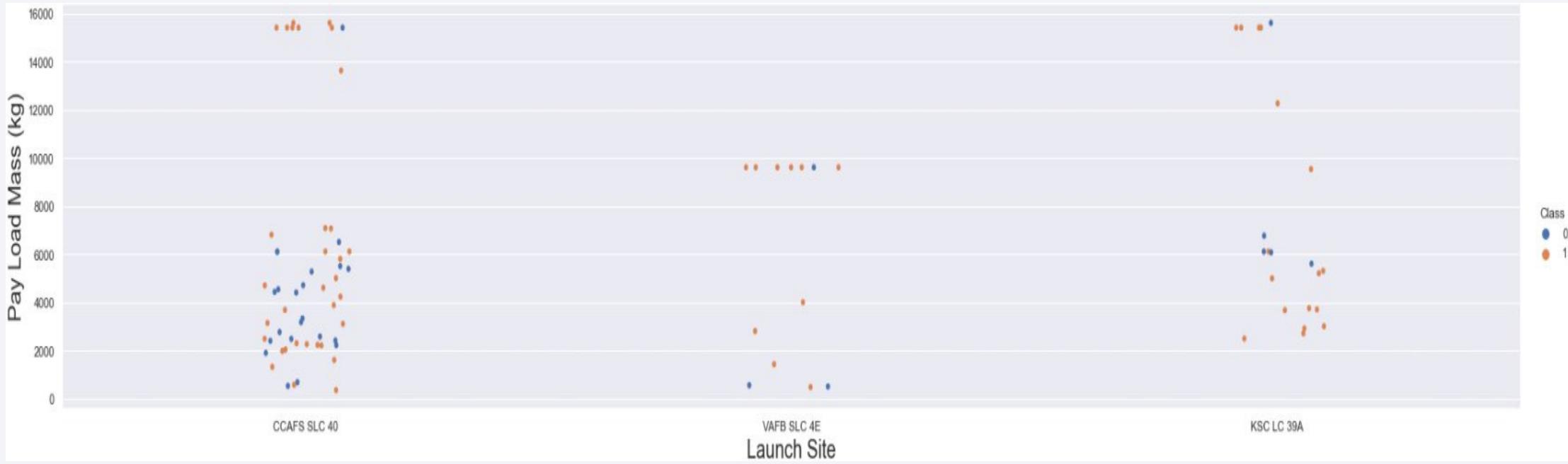
Flight Number vs. Launch Site



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

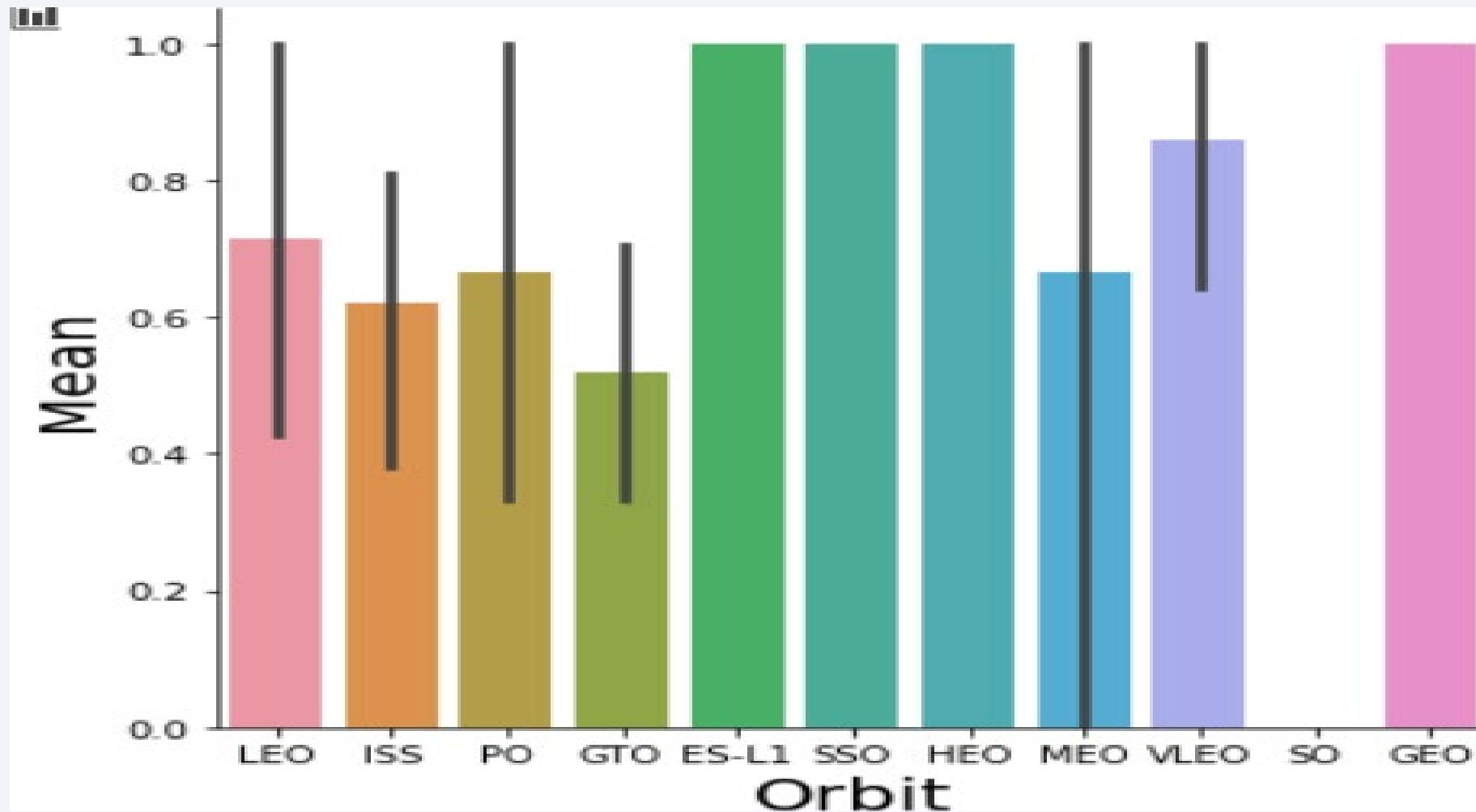
Payload vs. Launch Site



Explanation:

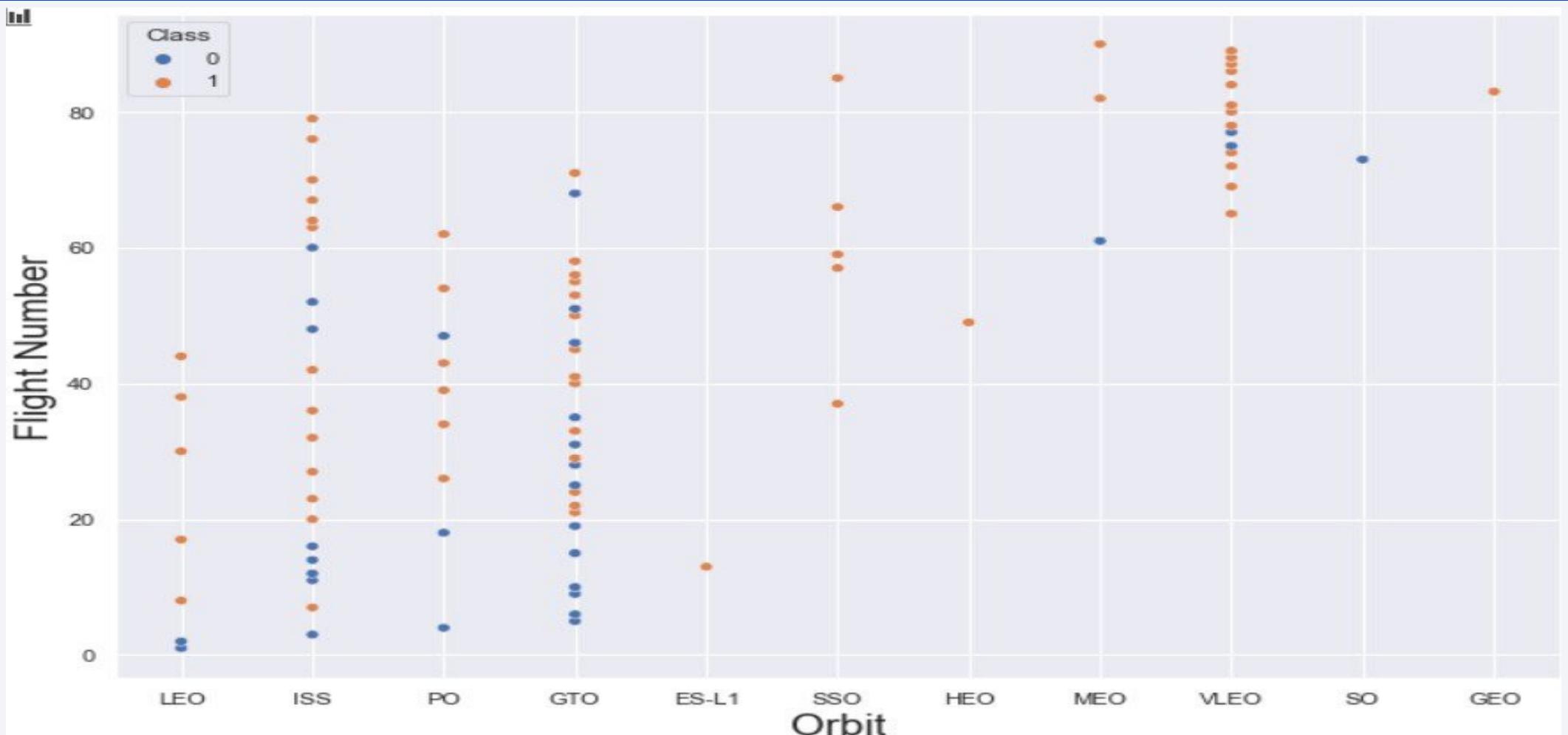
- For every launch site the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

Success Rate vs. Orbit Type



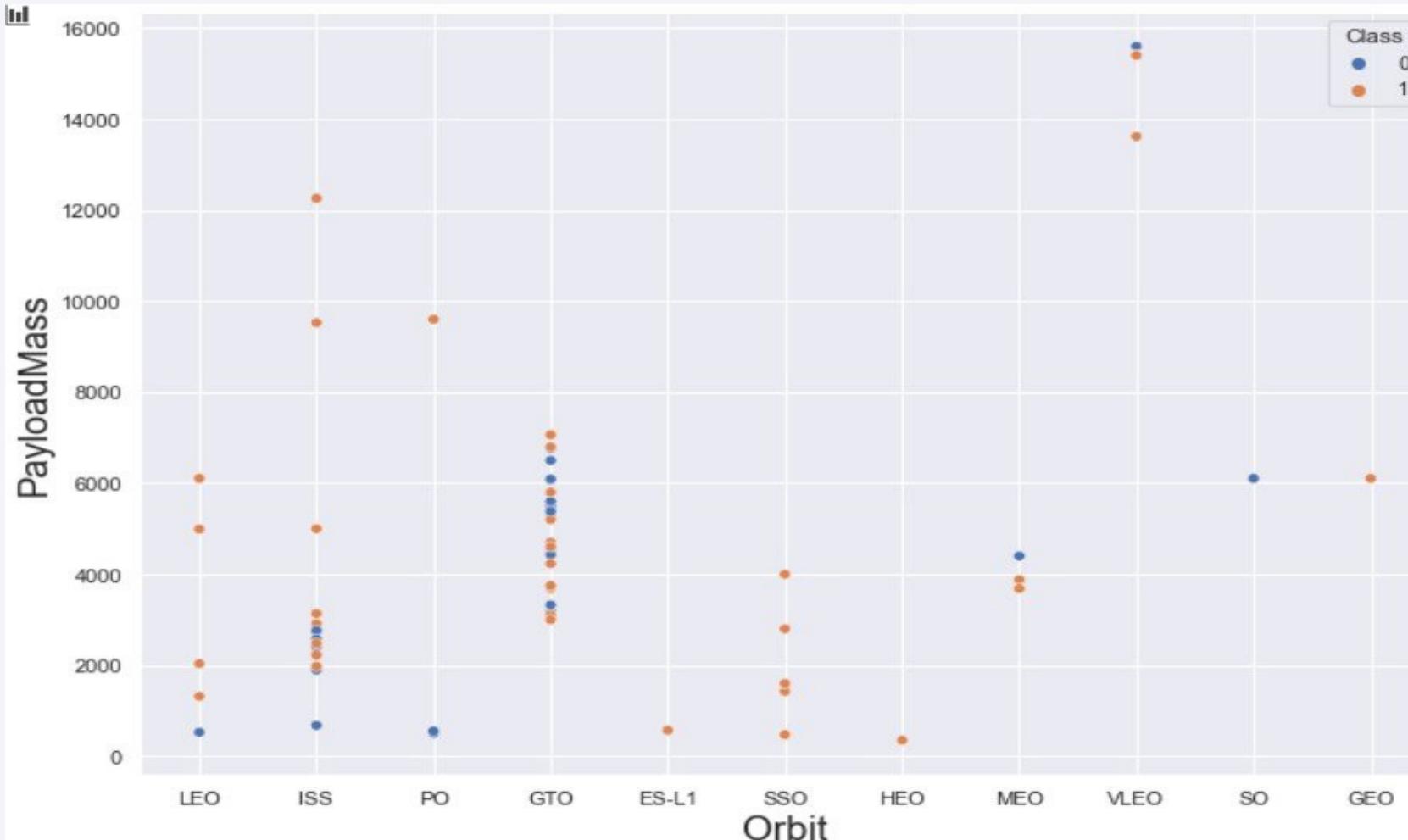
Orbit GEO, HEO, SSO, ES-L1 has the best Success Rate

Flight Number vs. Orbit Type



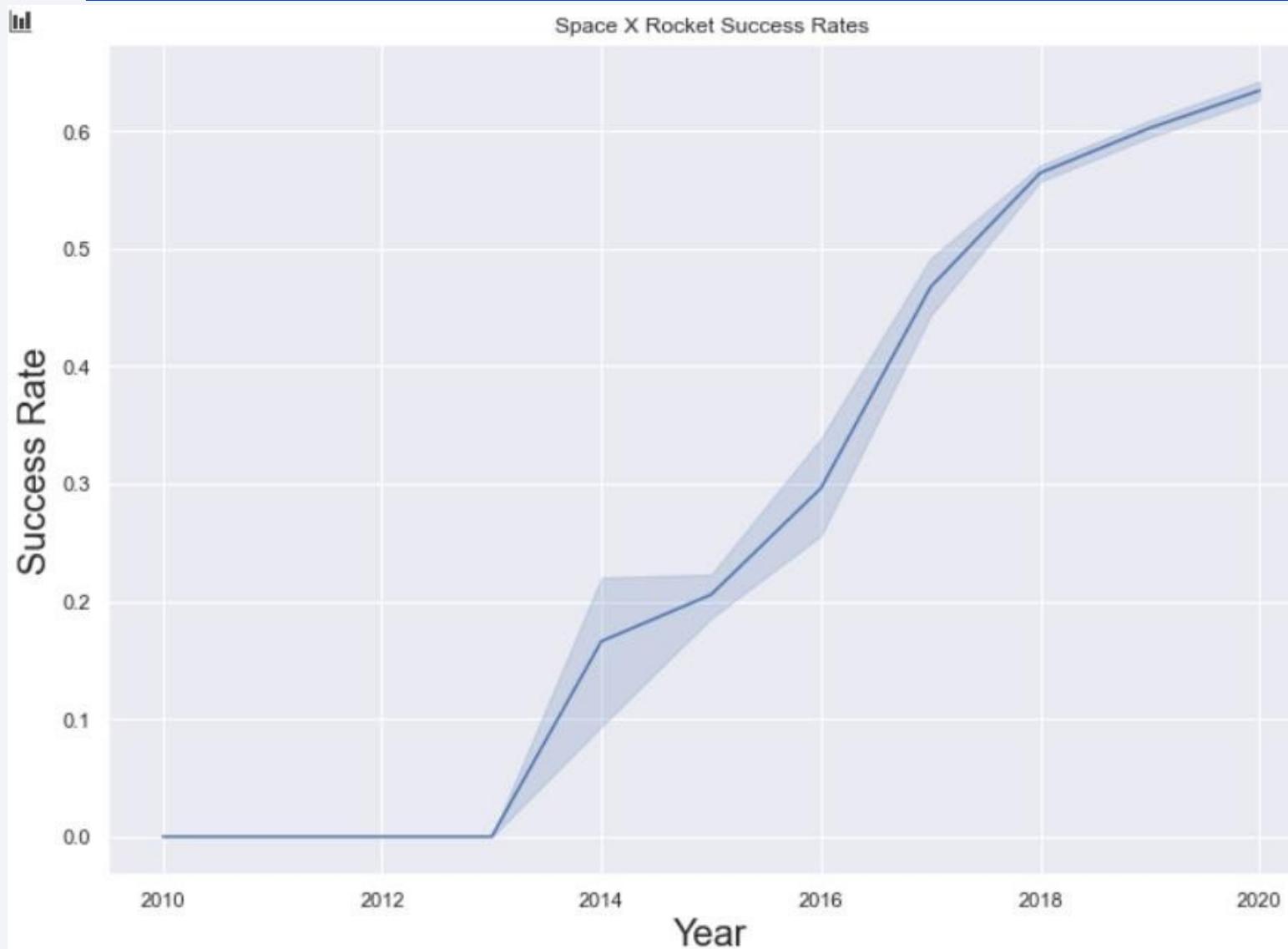
LEO orbit appears Success related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



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

```
In [4]: %sql select distinct launch_site from SPACEXDATASET;
```

```
* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31198/bludb  
Done.
```

Out[4]:

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

Displaying the names of the 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.bs2io90108kqb1od8lcg.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

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

Out[6]:

total_payload_mass
45596

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.bs2io90108kqb1od8lcg.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.bs2io90108kqb1od8lcg.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 4000 and 6000;
```

```
* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqblod8lcg.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.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31198/bludb  
Done.
```

booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

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.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)

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.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

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.

The background of the slide is a photograph taken from space at night. It shows the curvature of the Earth against the dark void of space. City lights are visible as numerous small white and yellow dots, primarily concentrated in the lower right quadrant where the United States and Mexico would be. In the upper left quadrant, the green and blue glow of the aurora borealis (Northern Lights) is visible in the upper atmosphere.

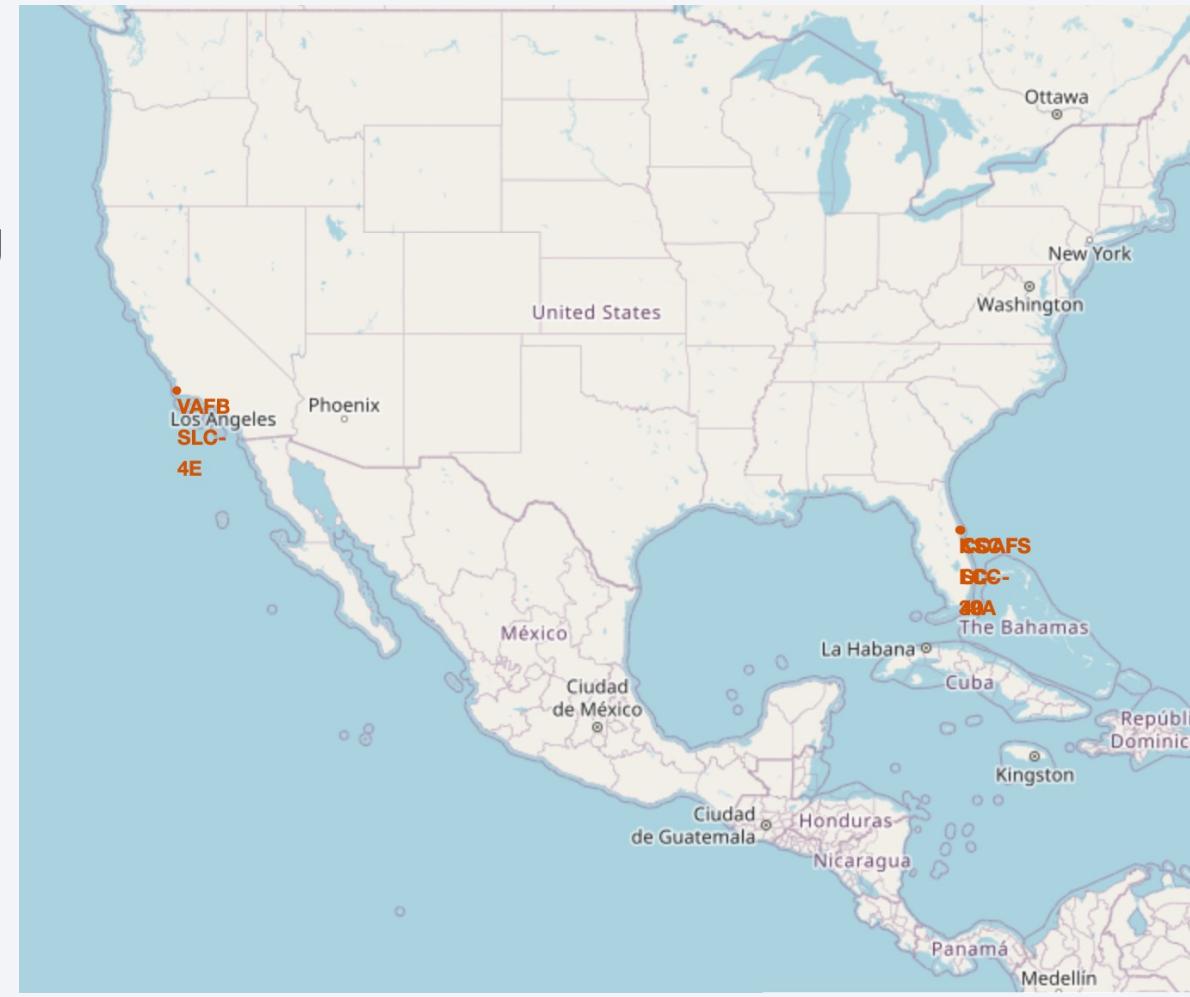
Section 3

Launch Sites Proximities Analysis

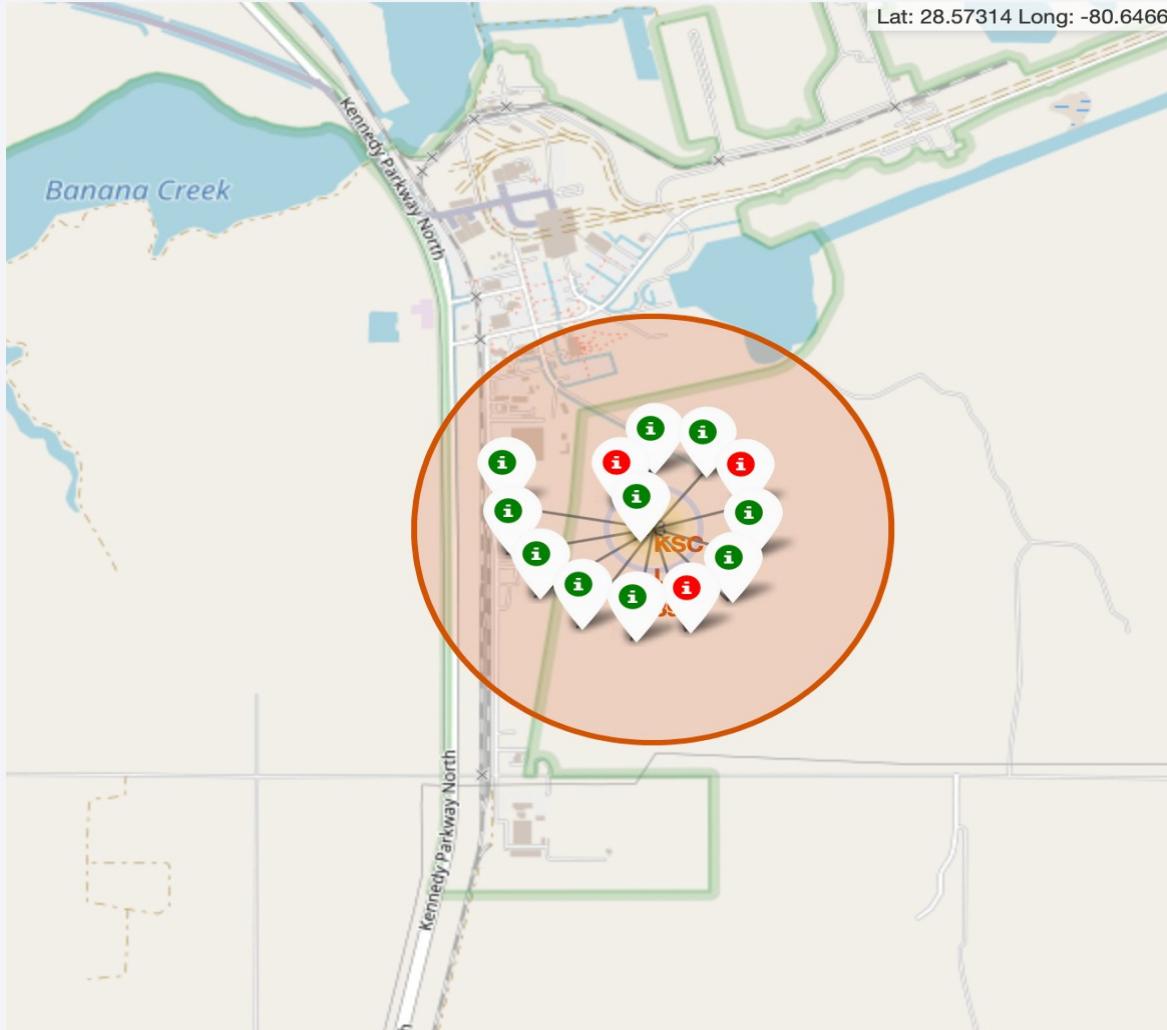
All launch sites' location markers on global map

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

All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimises the risk of having any debris dropping or exploding near people.

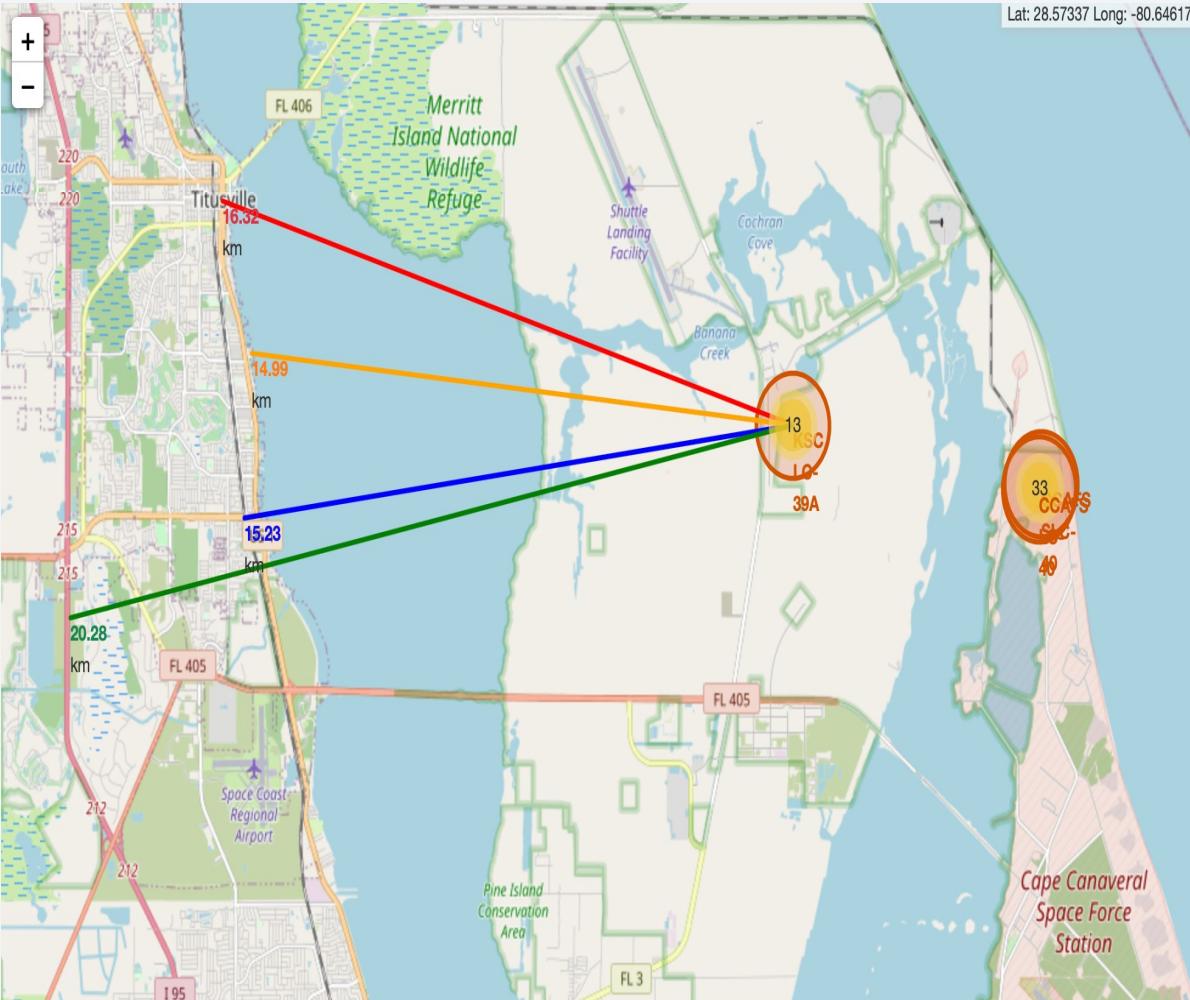


Color-labeled launch record on map



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

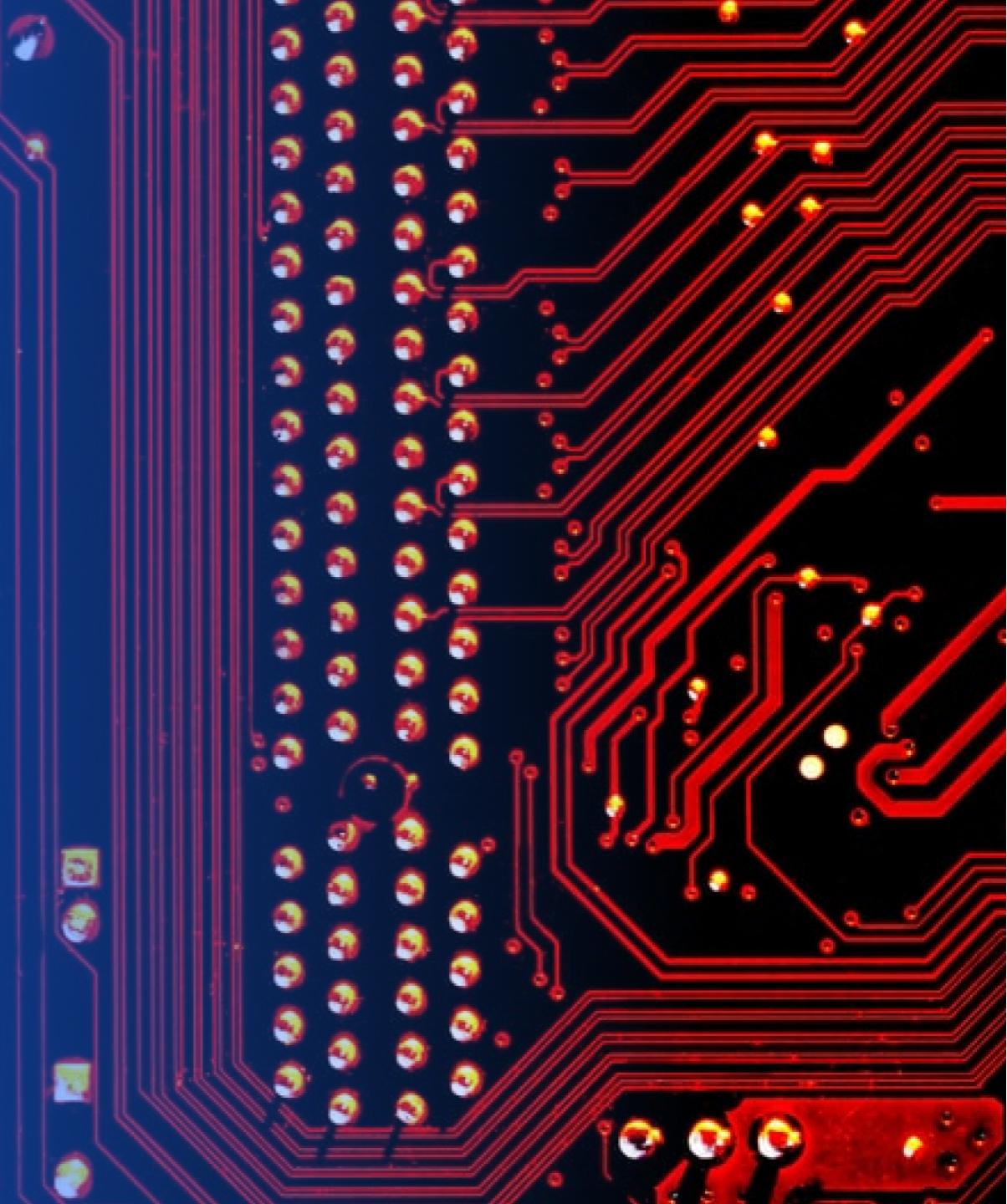
Distance from launch site KLCSC-39A to its proximities



- 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.

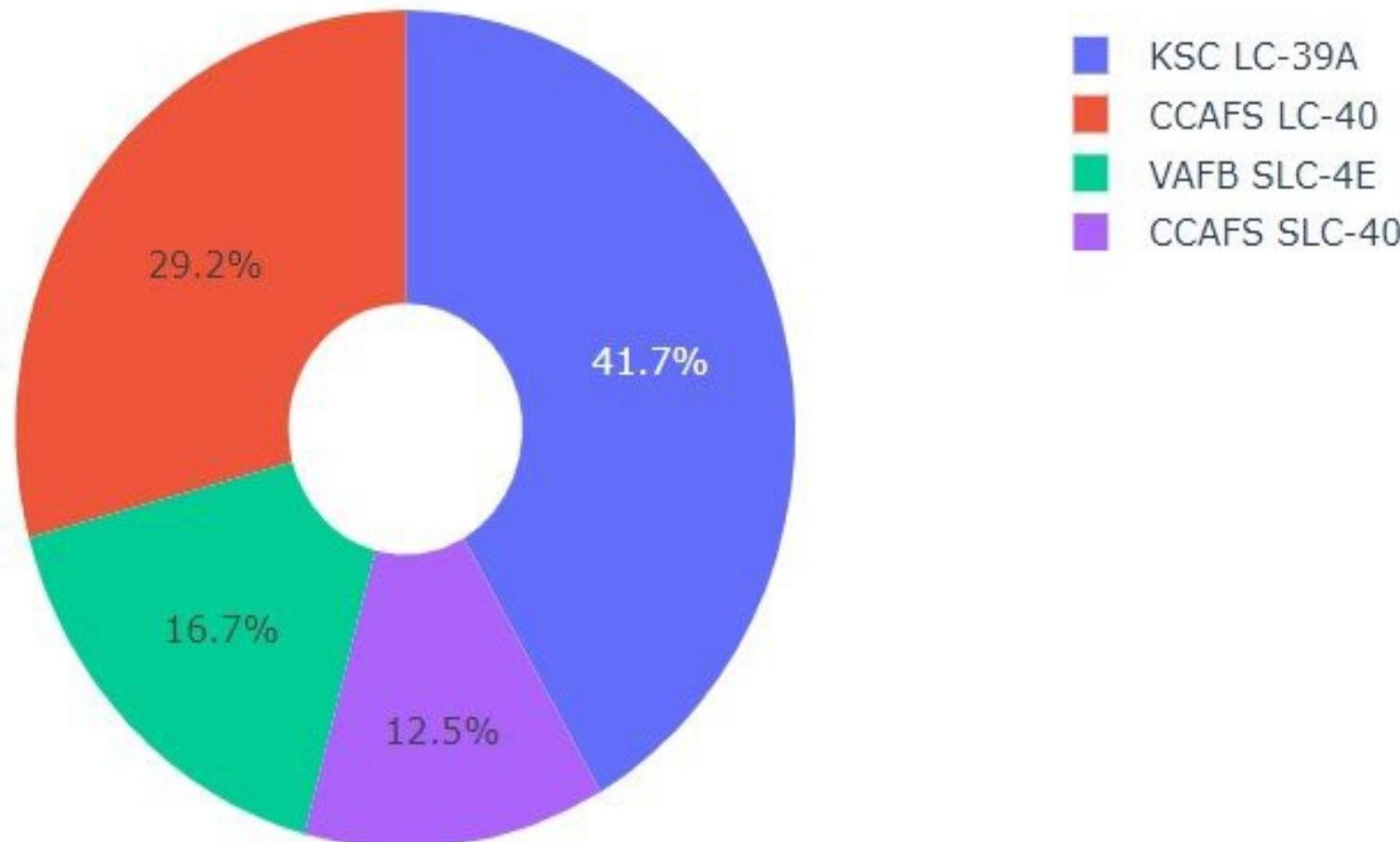
Section 4

Build a Dashboard with Plotly Dash



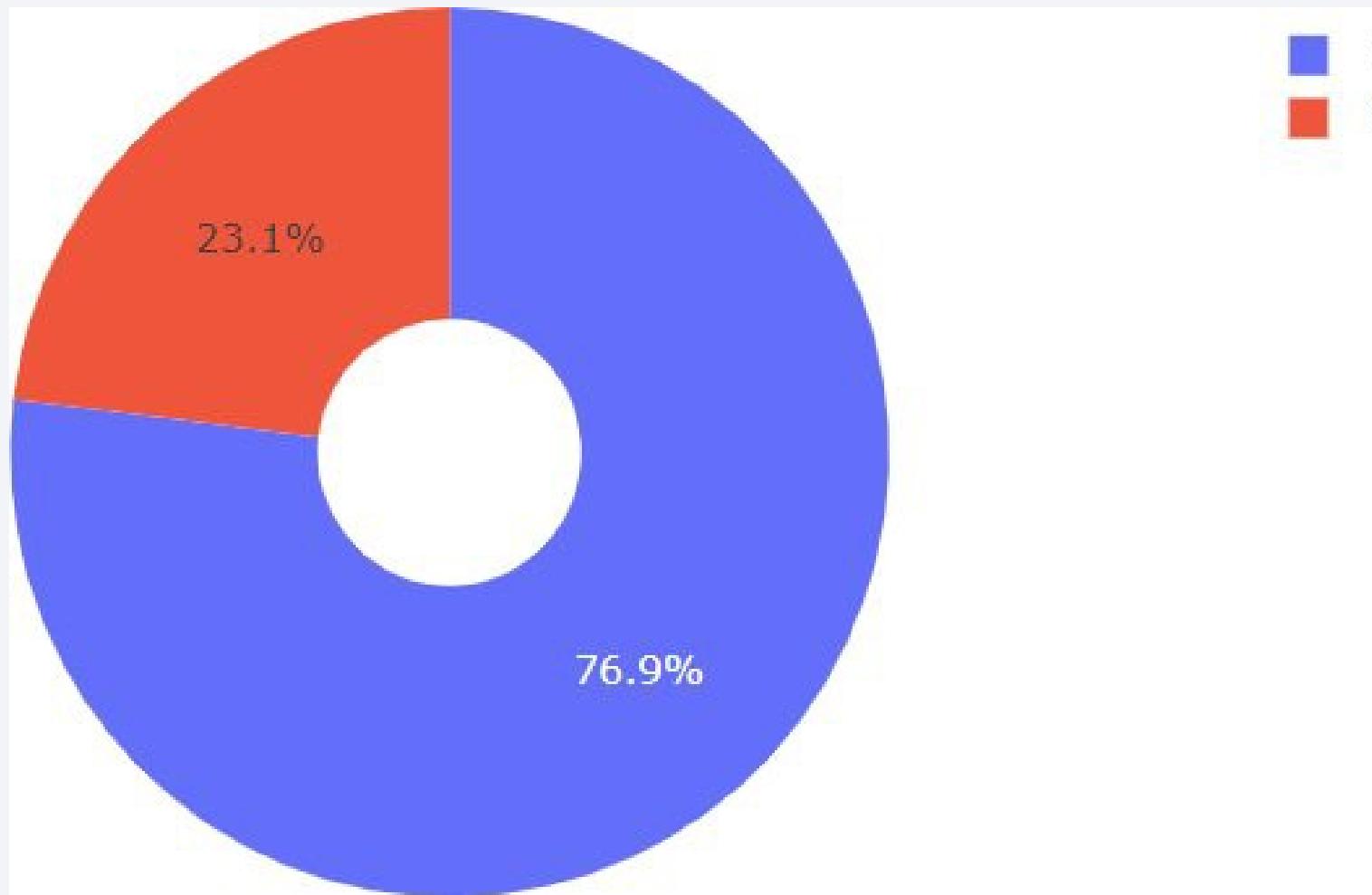
Launch success count for all site

Total Success Launches By all sites



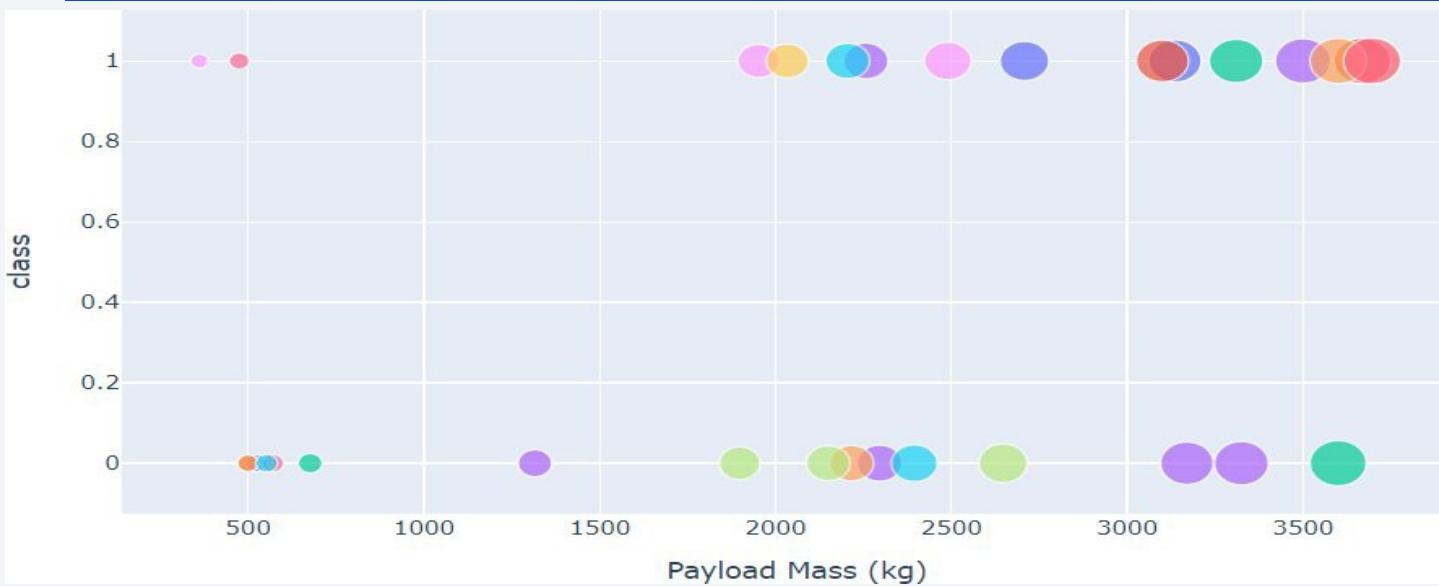
The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches.

Launch site with highest success launch ratio

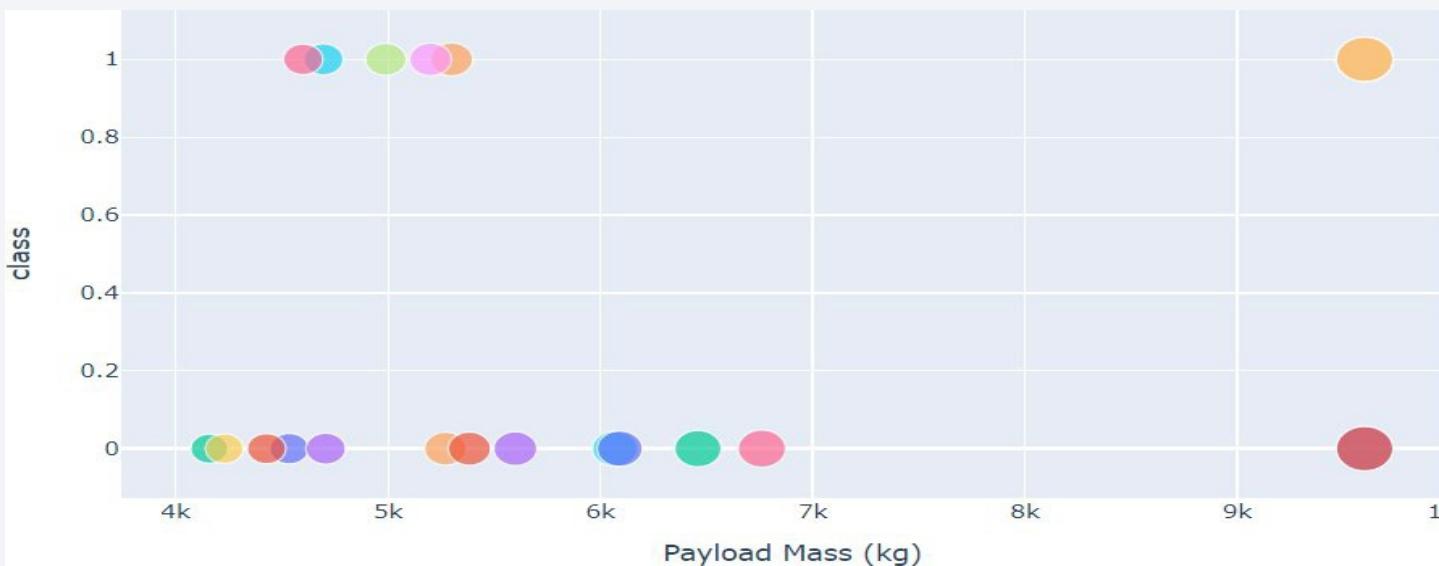


KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.

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

As you can see our accuracy is extremely close but we do have a winner its down to decimal places! using this function

```
bestalgorithm = max(algorithms, key=algorithms.get)
```

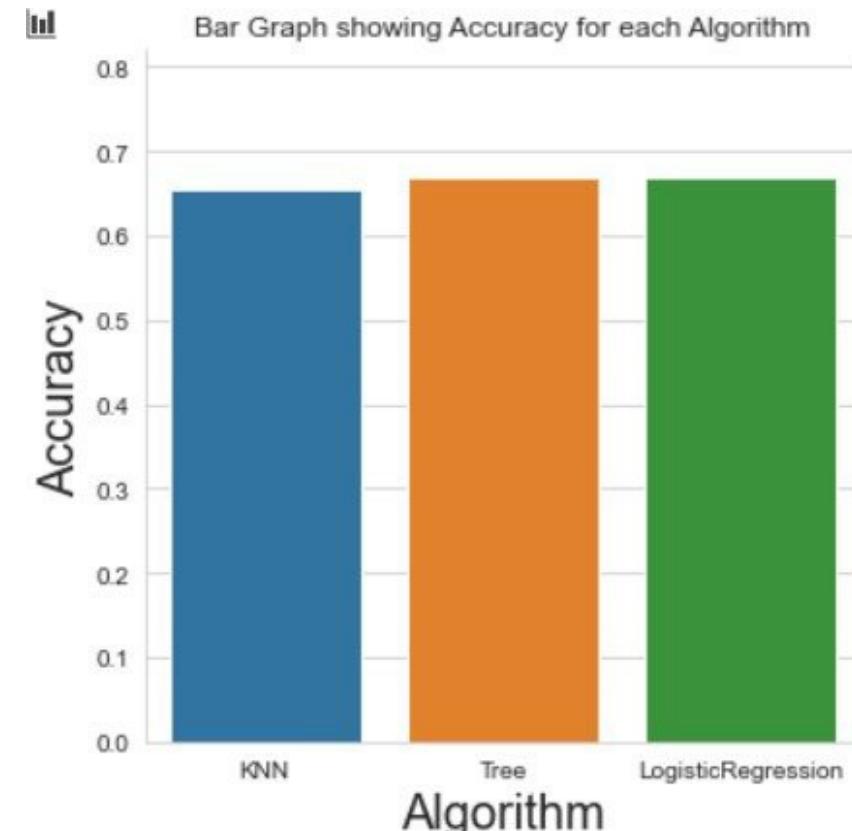
Accuracy	Algorithm
0 0.653571	KNN
1 0.667857	Tree
2 0.667857	LogisticRegression

The tree algorithm wins!!

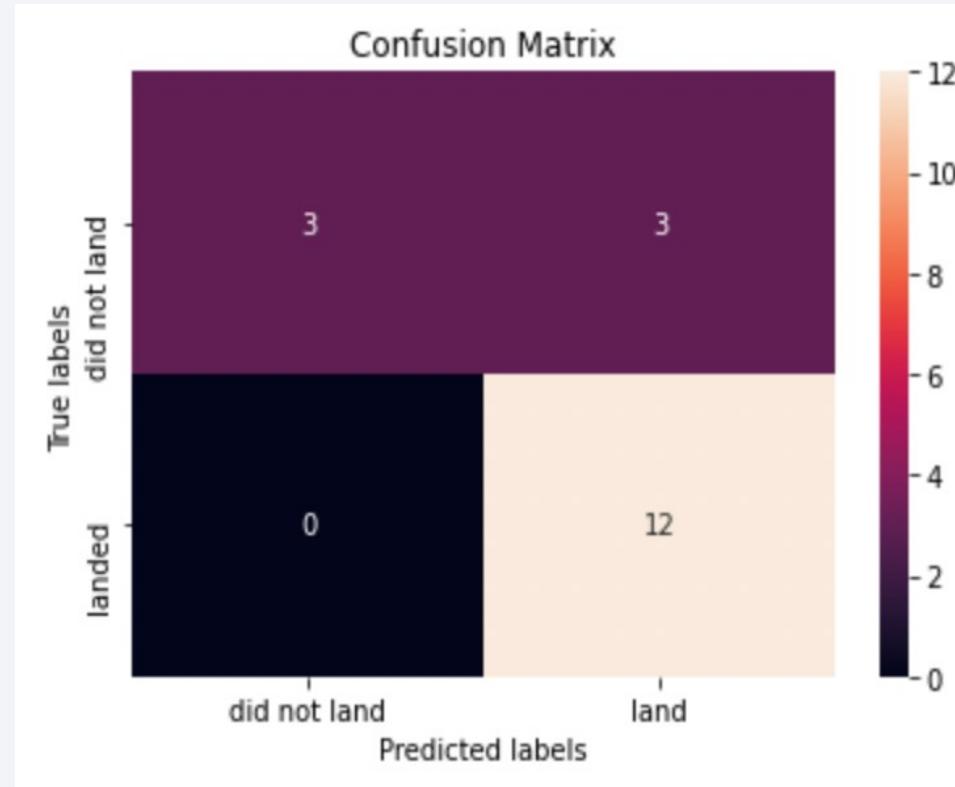
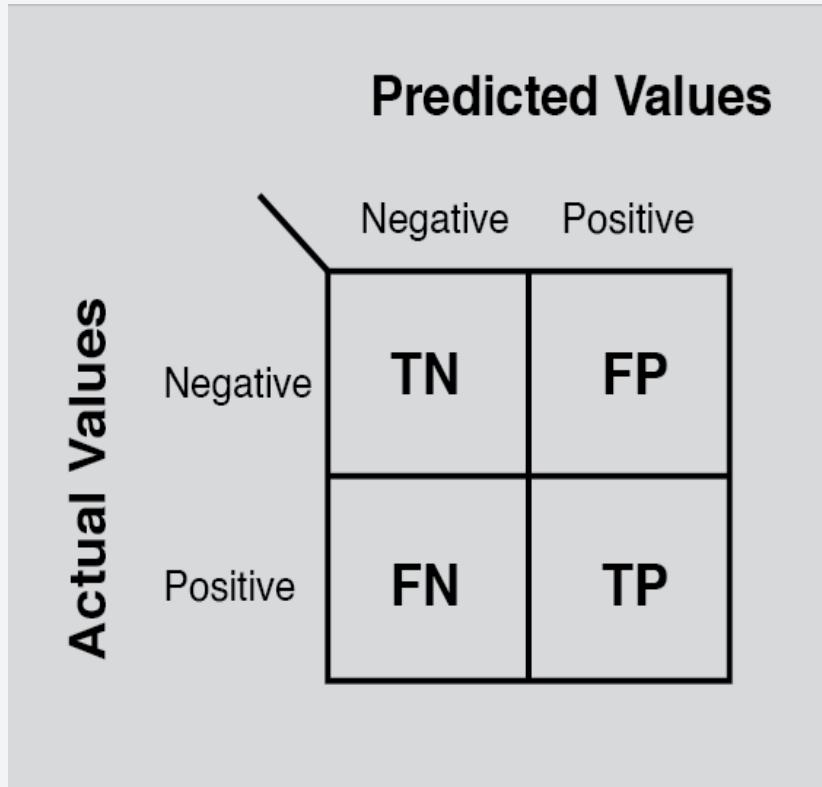
```
Best Algorithm is Tree with a score of 0.6678571428571429
```

```
Best Params is : {'criterion': 'gini', 'max_depth': 2, 'max_features': 'auto', 'min_samples_leaf': 1, 'min_samples_split': 2, 'splitter': 'best'}
```

After selecting the best hyperparameters for the decision tree classifier using the validation data, we achieved 83.33% accuracy on the test data.



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

1. Decision Tree Model is the best algorithm for this dataset.
2. Launches with a low payload mass show better results than launches with a larger payload mass.
3. Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
4. The success rate of launches increases over the years.
5. KSC LC-39A has the highest success rate of the launches from all the sites.
6. Orbits ES-L1, GEO, HEO and SSO have 100% success rate.

Special Thanks to:



Instructors

Coursera

IBM

Thank you!

