IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Subhransu Chaudhuri

12-06-2025

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis for Data Visualization
  - Interactive Visual Analytics with Folium
  - Interactive Dashboard with Plotly Dash
  - Predicted Analysis
- Summary of all results
  - Exploratory Data Analysis Result
  - Screenshots of Interactive Visual Analytics
  - Predicted Analysis Result

Introduction

SpaceX is the most successful company of the commercial space age. One reason id that the rocket launches are relatively inexpensive. SpaceX advertises **Falcon 9** rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars, much of the savings is because SpaceX can reuse the first stage. Therefore, if we the people at SpaceY can determine if the first stage will land, we can determine the cost of the launch and thus compete with SpaceX.

To do this following questions need to be answered :-

- How do various variables like payload mass, launch site, orbits, etc. affect the success of the first stage landing ?
- Since we will use machine learning models we will need to know, which algorithm has the most accuracy ?
- How to know which future rocket launches will have a successful first stage landing ?

Section 1

Methodology

Methodology

- Data collection methodology:
  - Using SpaceX rest API
  - Using web scrapping from Wikipedia
- Perform data wrangling
  - Filtering the data
  - Dealing with missing data
  - 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

result.

Data Collection

Data collection process involved a combination of API requests from SpaceX Rest API and Web Scraping data from a table in SpaceX's Wikipedia entry.

We had to use both of these data collection methods in order to get complete information about the launches for a more detailed analysis.

Data Columns obtained by using SpaceX REST API:

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite,

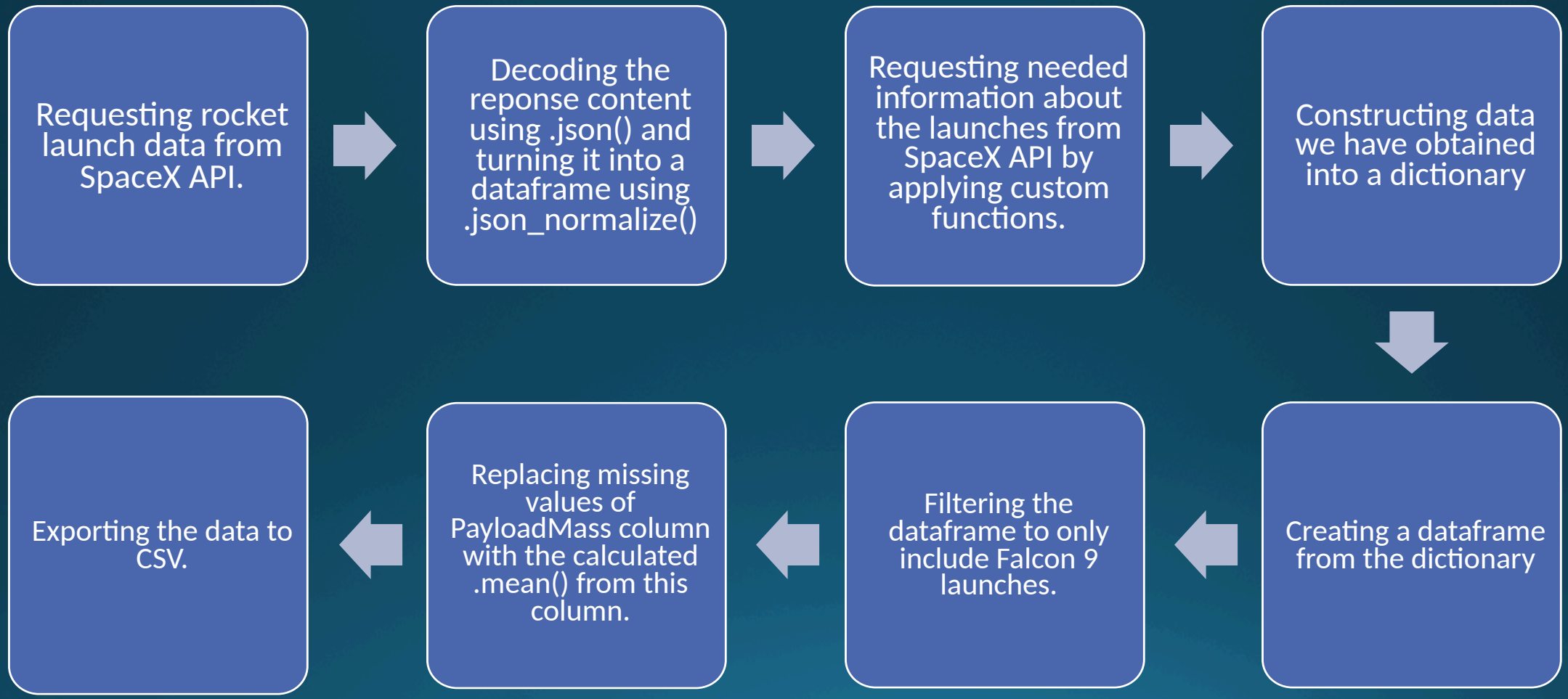
Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude.

Data Columns obtained by using Wikipedia Web Scrapping:

Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch

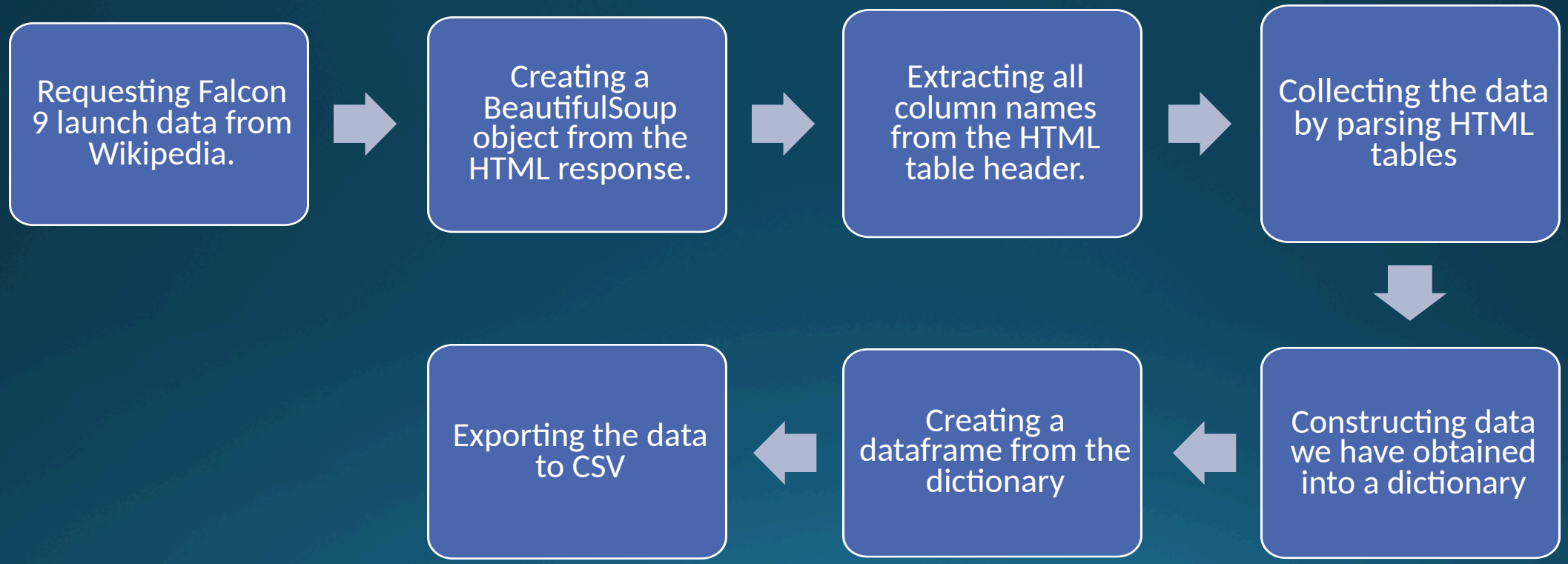
outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



Click here for more info

Data Collection - Scraping



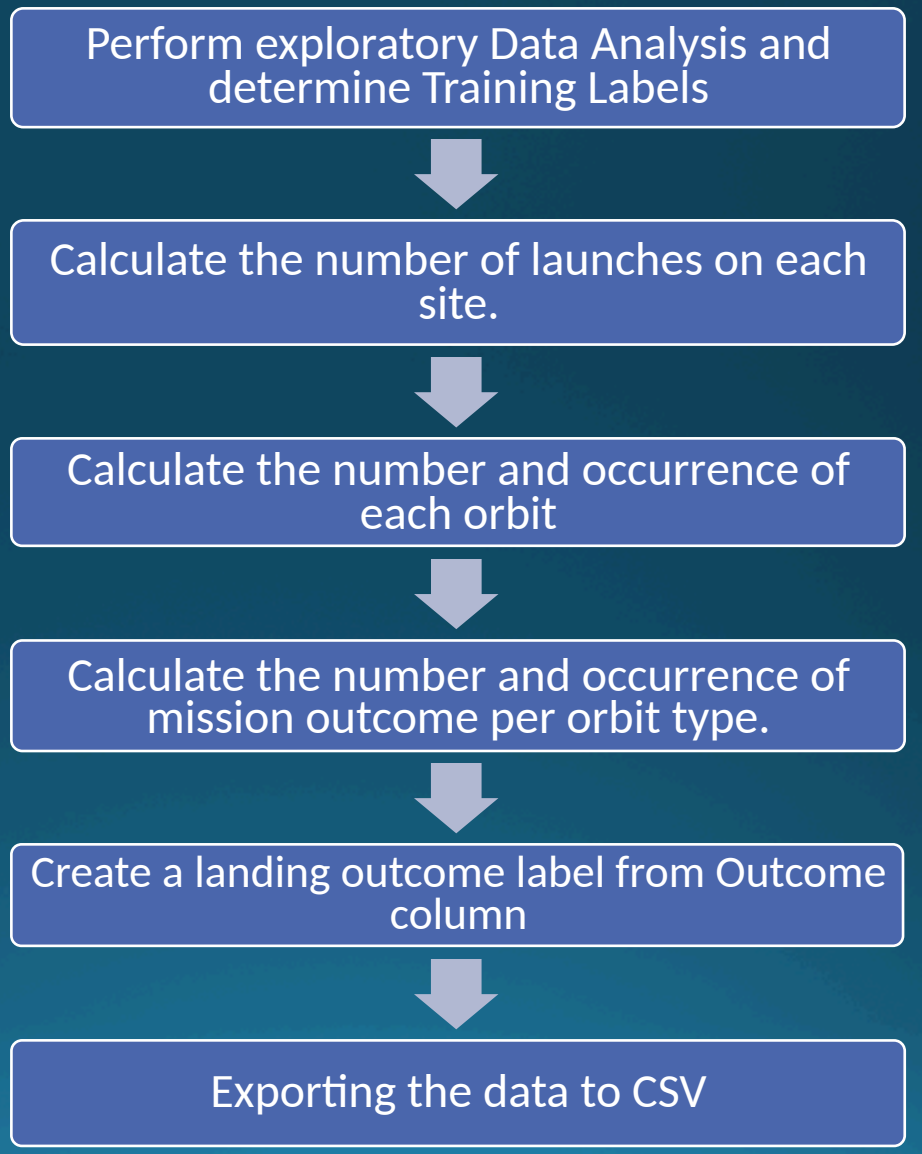
Click here for more info

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 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.

Click here for more info



EDA with Data Visualization

Various 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
- Success Rate Yearly Trend

Click here for more info

EDA with SQL

Performed the following SQL queries:

- Displaying the names of the unique launch sites in the space stations.
- 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 name of the boosters which have success in drone ships 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.

Click here for more info

Build an Interactive Map with Folium

Markers of all Launch Site:

- Added Markers with Circle, Popup Label and Text Label of NASA Johnson Space Centre 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.

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

Click here for more info

Build a Dashboard with Plotly Dash

Launch Sites Dropdown List:

- Added a dropdown list to enable Launch Site selection.

Pie Chart showing Success Launches(All Sites/ Certain Sites):

- 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.

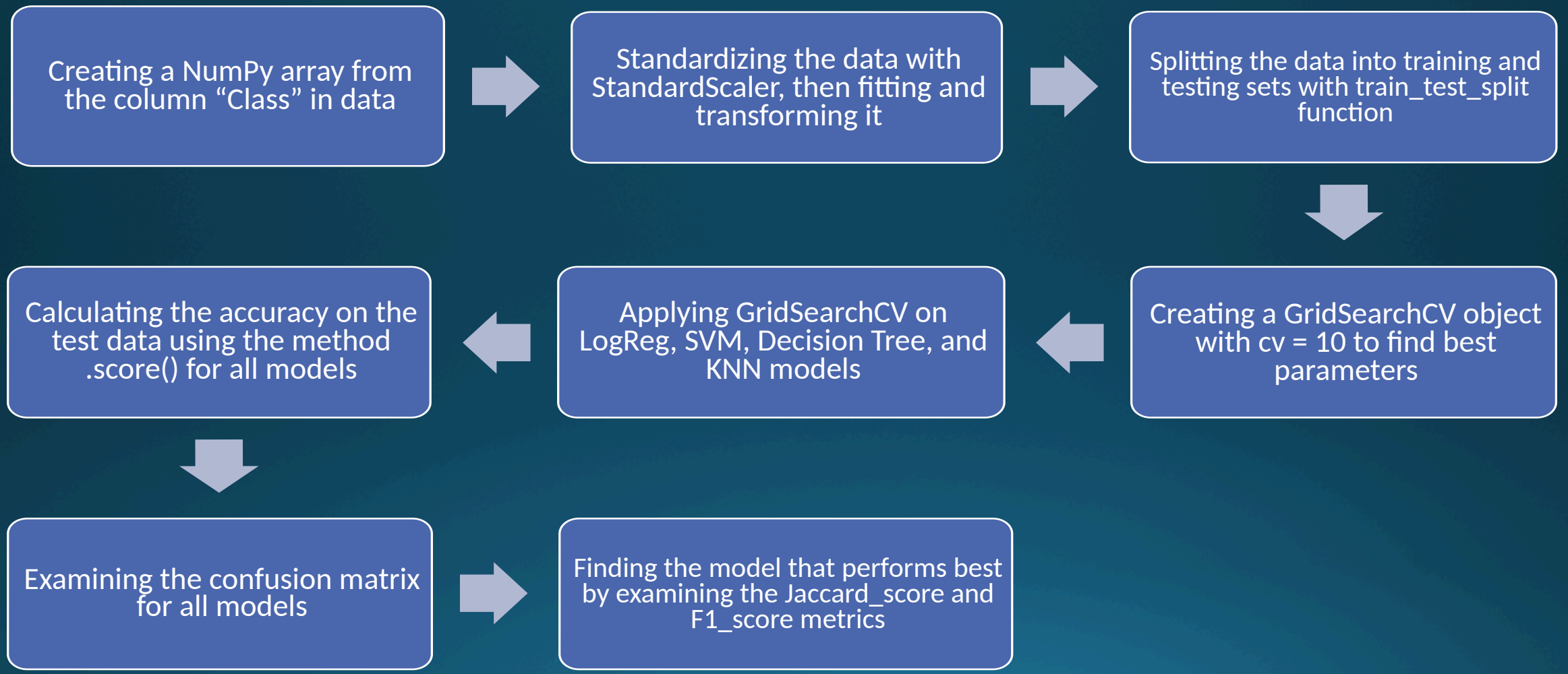
Slider of Payload Mass Range:

- 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

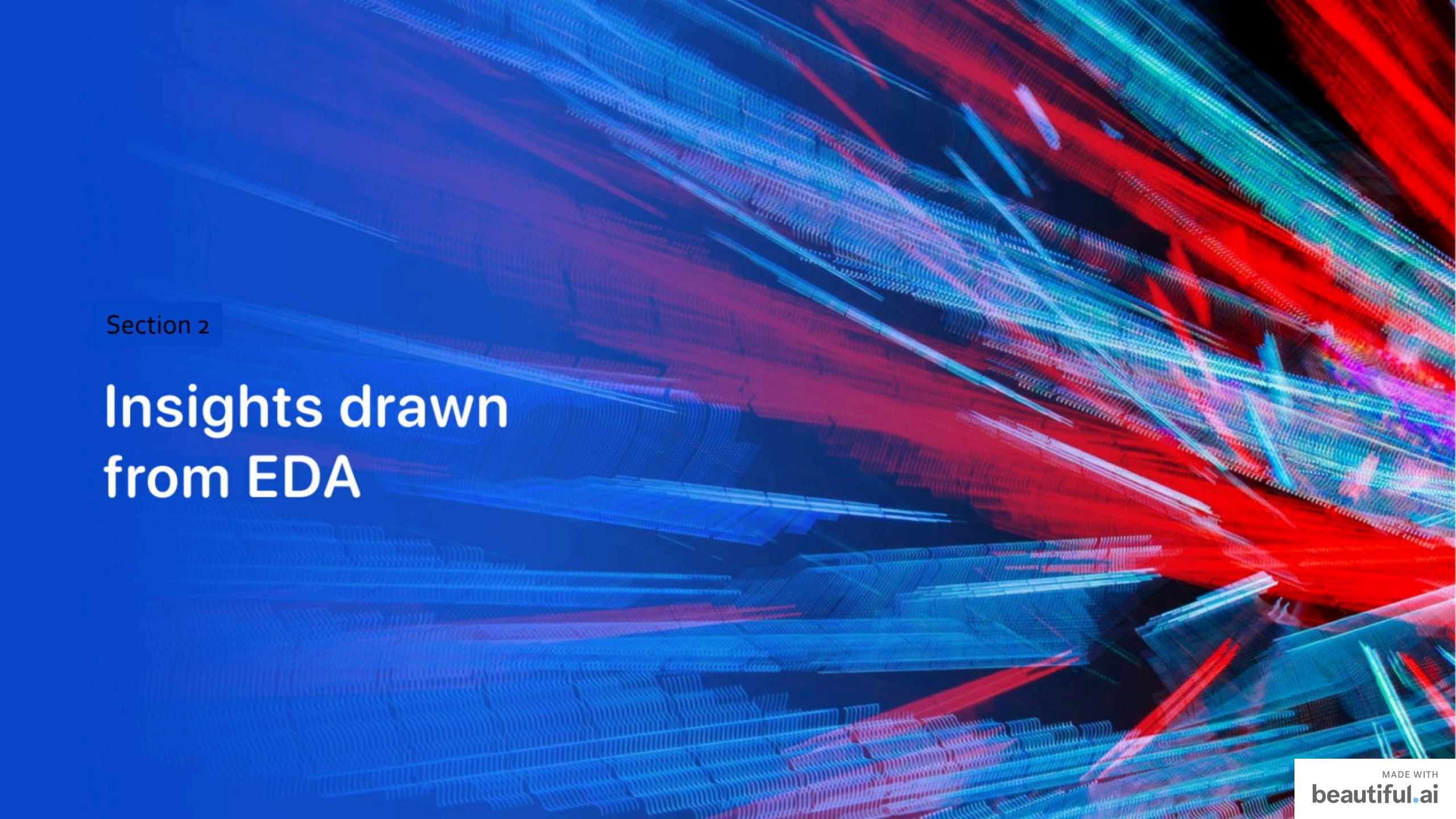
Predictive Analysis (Classification)



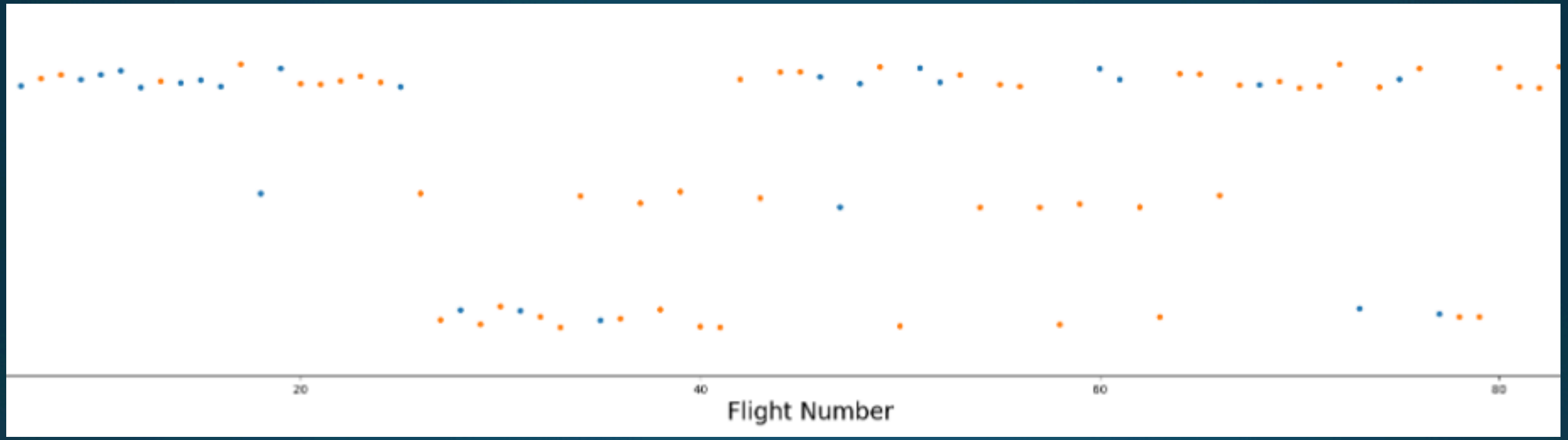
Click here for more info

Results

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

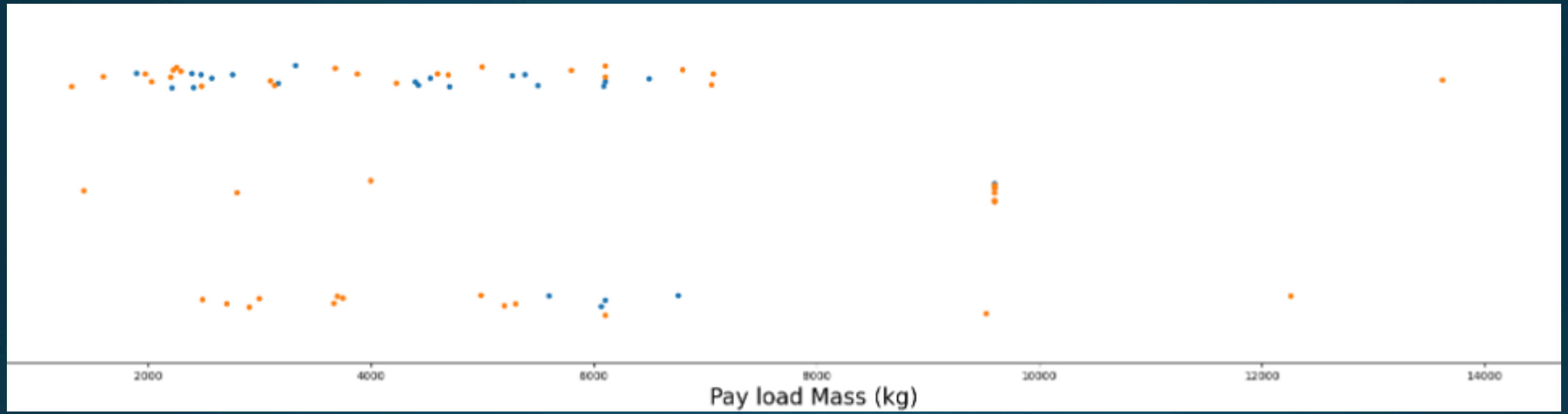


Flight Number vs. Launch Site



- The earliest flights all failed while the latest flights all succeeded
- The CCFAS SLC-40 launch site has about half all launches
- VAFB SLC 4E and KSC LC 39A have a higher success rate

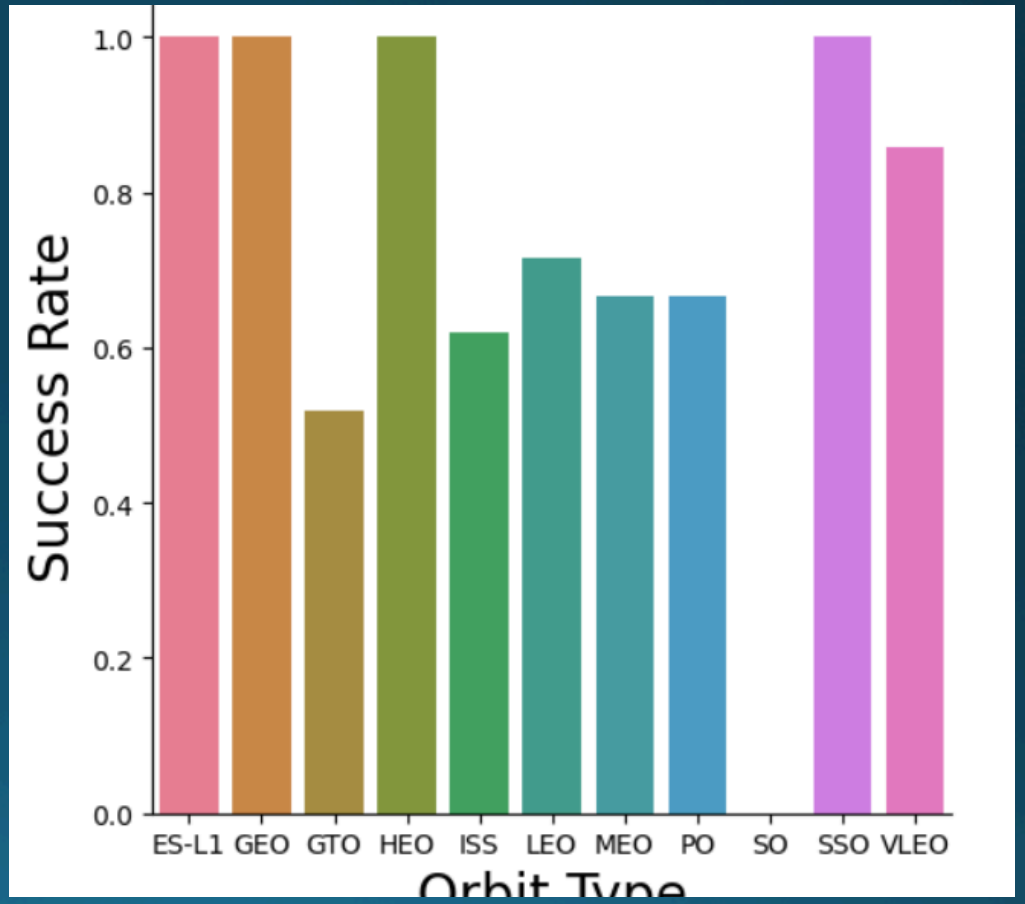
Payload vs. Launch Site



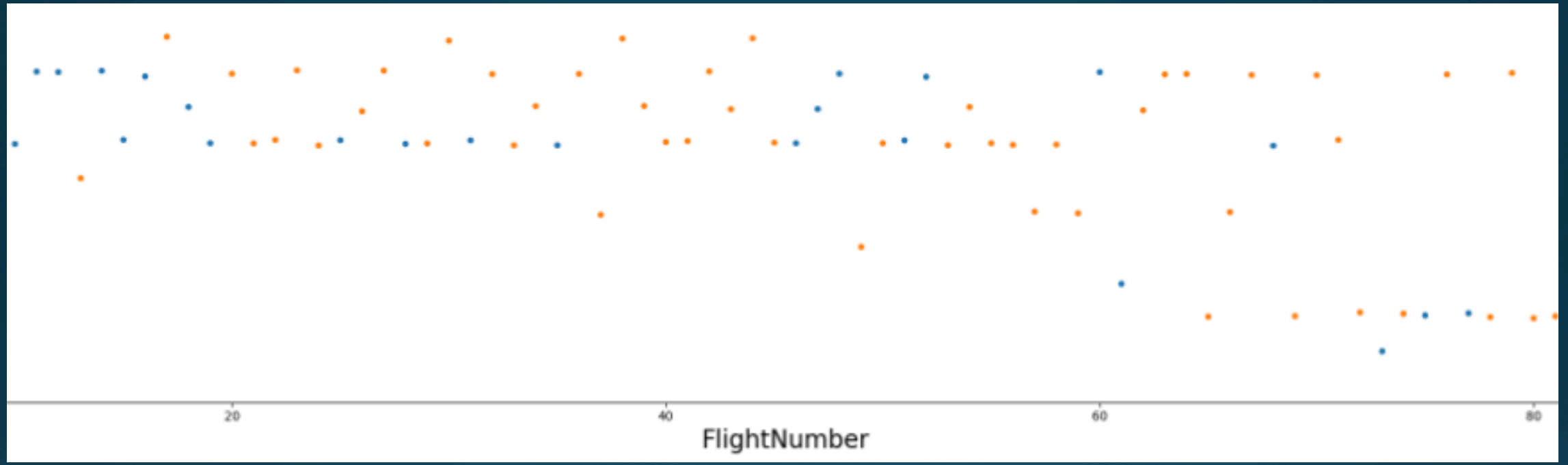
- For every launch site higher the payload mass, the higher the success rate
- Most of the launches with payload mass over 7000kg were successful
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

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 50% to 90%: GTO, ISS, LEO, MEO, PO, VLEO

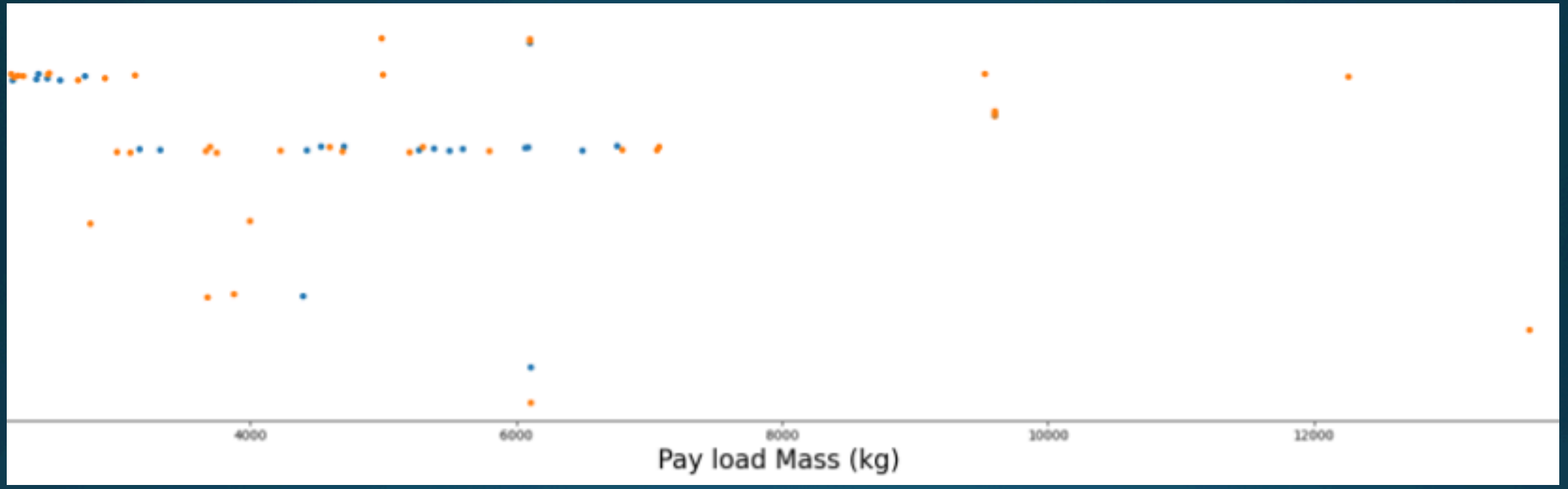


Flight Number vs. Orbit Type



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 and GTO orbit.

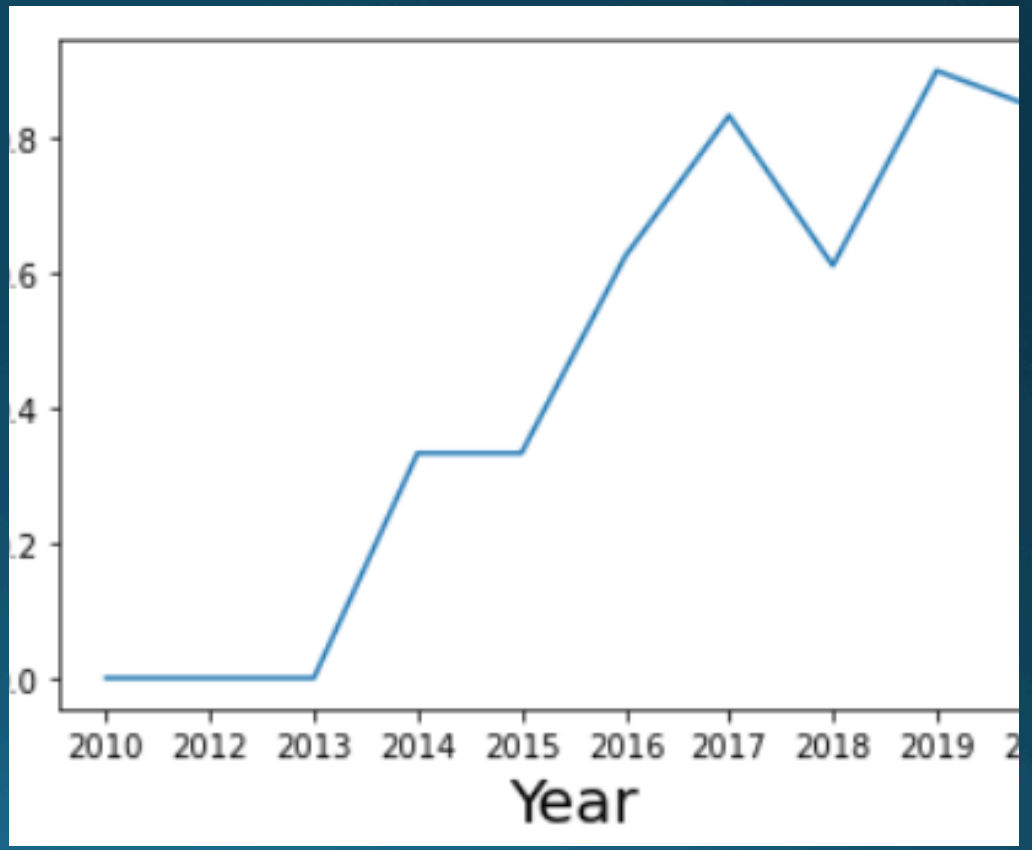
Payload vs. Orbit Type



Heavy payload 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 on increasing.



All Launch Site Names

```
%sql select distinct launch_site from SPACEXTBL
```

```
 * sqlite:///my_data1.db
Done.
```

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

Displaying all unique launch site names

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5
```

```
 * sqlite:///my_data1.db
Done.
```

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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 details of launch site starting with ‘CCA’

Total Payload Mass

```
ct sum(PAYLOAD_MASS__KG_) as total_payload_mass from SPACEXTBL where customer = 'NA
:///my_data1.db
```

oad_mass
45596

Displaying the total payload mass carried by booster launched by NASA

Average Payload Mass by F9 v1.1

```
t avg(PAYLOAD_MASS__KG_) as Average_Payload_Mass from SPACEXTBL where Booster_Version :
```

```
///my_data1.db
```

yload_Mass
2928.4

Displaying the average Payload Mass Carried by ‘Booster Version F9 v1.1’

First Successful Ground Landing Date

```
ql select min(Date) from SPACEXTBL where Landing_Outcome = 'Success (ground pad
 sqlite:///my_data1.db
ne.
in(Date)
15-12-22
```

Displaying the date of the first successful ground landing

Successful Drone Ship Landing with Payload between 4000 and 6000

```
select Booster_Version from SPACEXTBL where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and
```

```
qlite:///my_data1.db
.
```

ster_Version
F9 FT B1022
F9 FT B1026
9 FT B1021.2
9 FT B1031.2

Displaying the list of successful drone ship landing with payload mass between 4000 and 6000.

Total Number of Successful and Failure Mission Outcomes

```
%sql select Mission_Outcome, count(*) from SPACEXTBL group by Mission_Outcome
```

```
 * sqlite:///my_data1.db
Done.
```

Mission_Outcome	count(*)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Displaying the total number of Successful and Failed Mission Outcomes

Boosters Carried Maximum Payload

```
 * sqlite:///my_data1.db
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

Displaying the list of Booster Version carrying maximum payload mass.

2015 Launch Records

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Displaying Launch Records of the year 2015

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

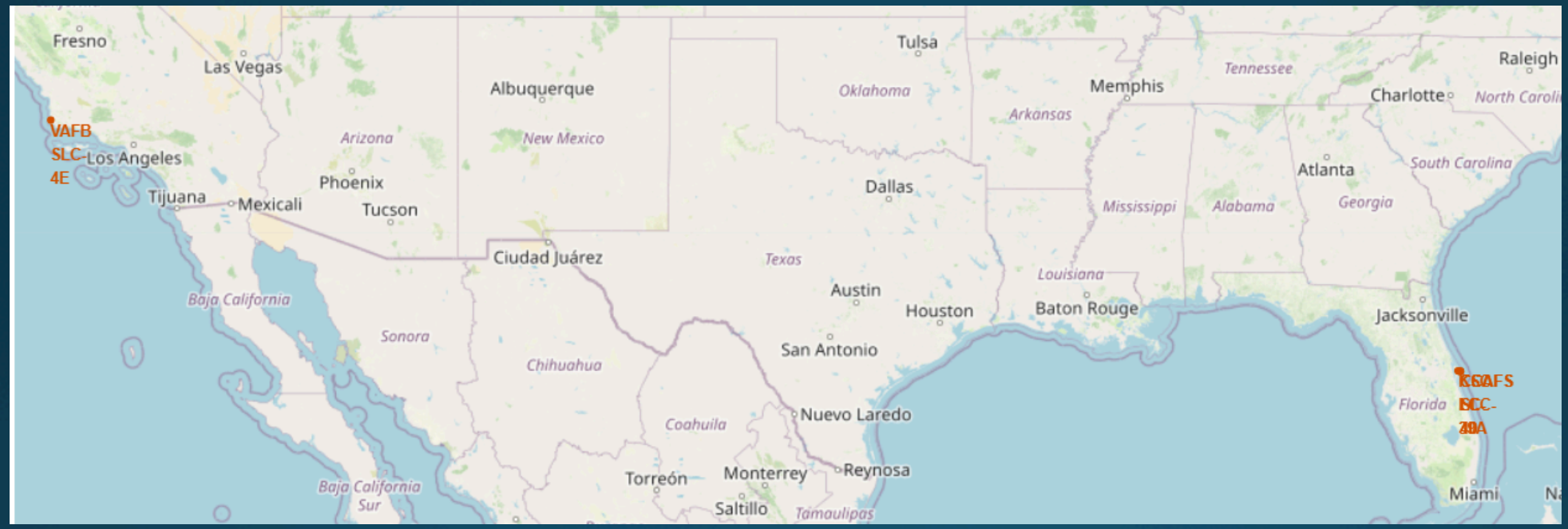
Landing_Outcome	count_outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Ranking Landing Outcomes between 2010-06-4 and 2017-03-20

Section 3

Launch Sites Proximities Analysis

Launch Sites



The launch sites are quite closer to the equator. The Earth's Equator rotates at approximately 1000 miles per hour. This rotational speed translates to a free velocity boost for rockets launched eastward, reducing the amount of fuel needed for orbital insertion.

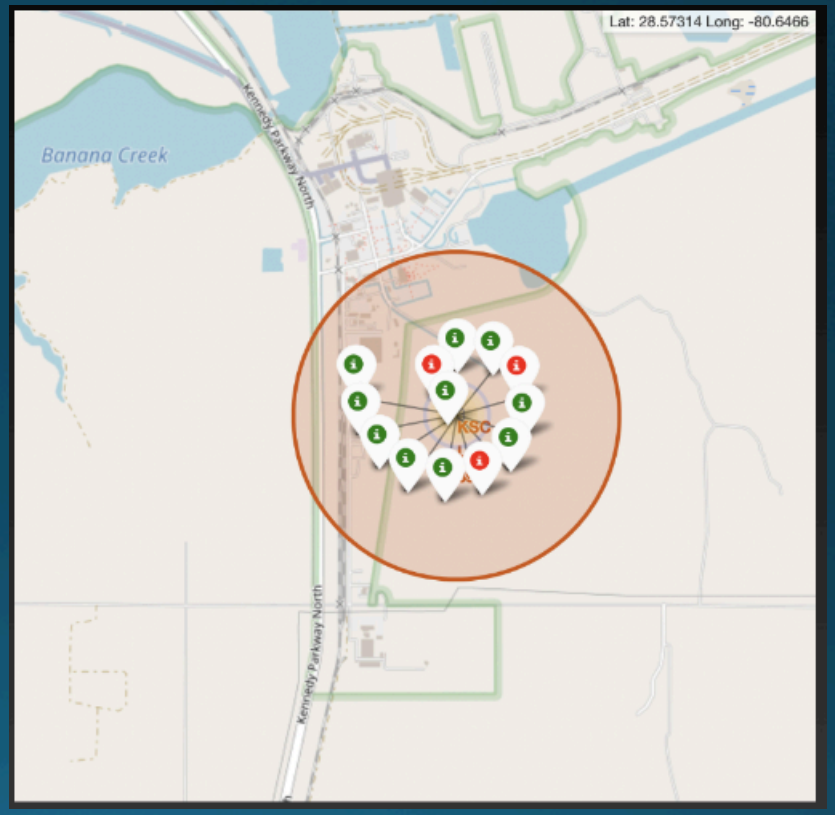
All launch sites are really close to the coast because while launching rockets towards the ocean it minimizes the risk of having any debris dropping or exploding near people.

Lauch Site KSC LC-39A

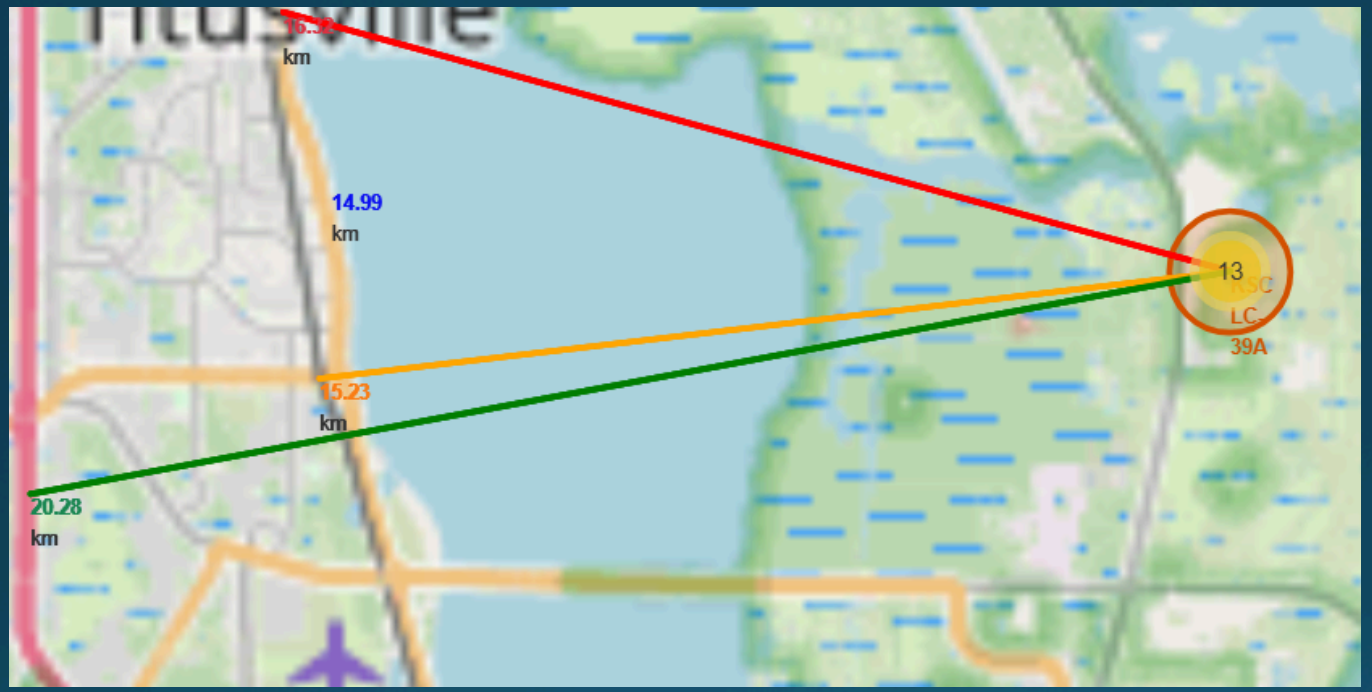
From the coloured markers we should be able to predict which launch site has high success rate.

The green markers indicate a successful landing, whereas the red markers indicate a failed landing.

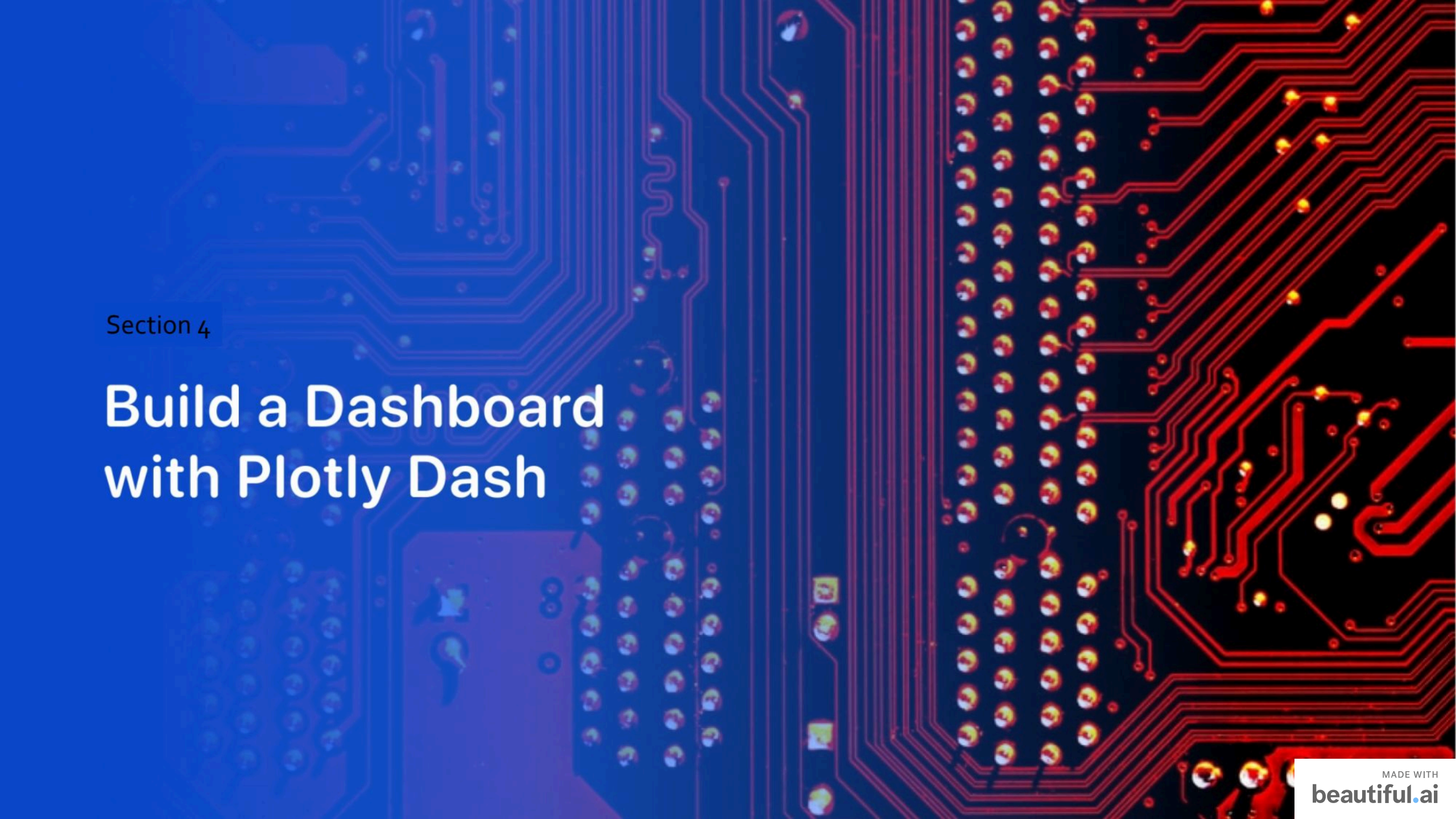
So, by observing we can see the launch site KSL LC-39A has a high success rate.



Distances between a Launch Site to its Proximities



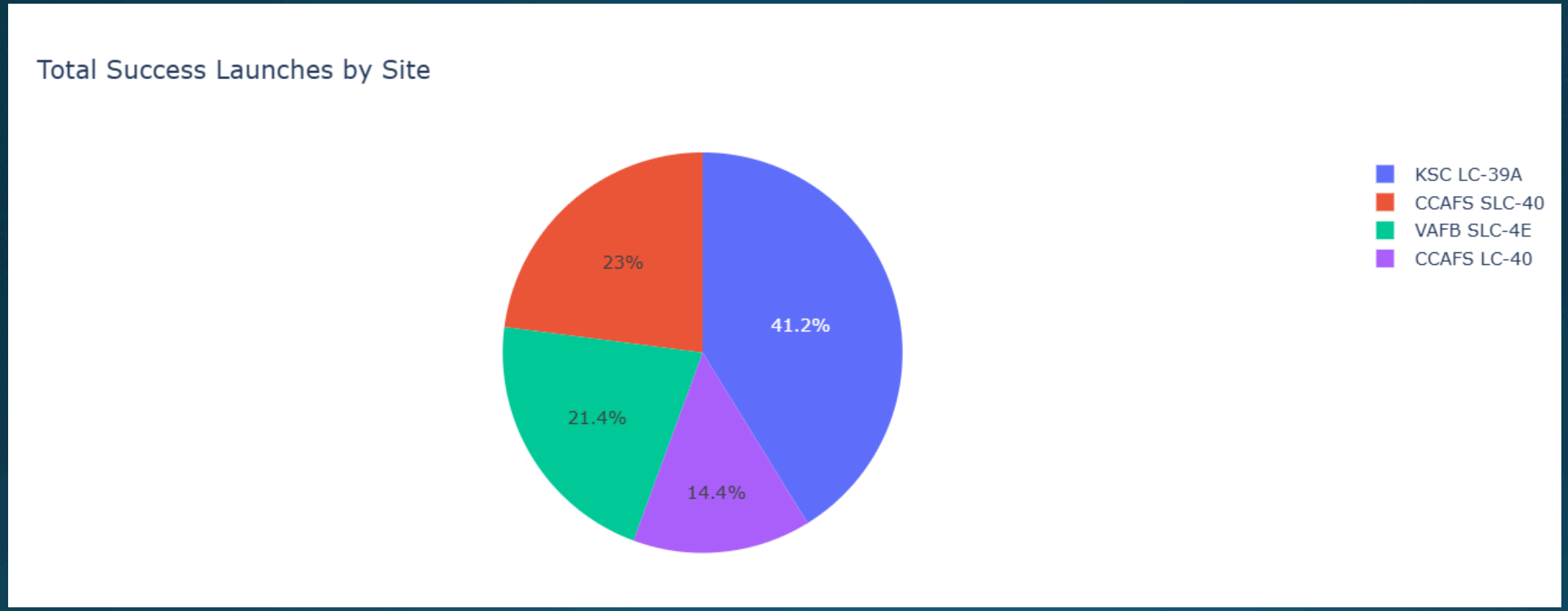
From observation we can see that the launch site KSC LC-39A is relatively close to coastline, highway, railway and the city. The closeness of the launch site to the city i.e., Titusville is potentially dangerous because failed rockets with its high speed can cover distances like 15-20 km in a few seconds.



Section 4

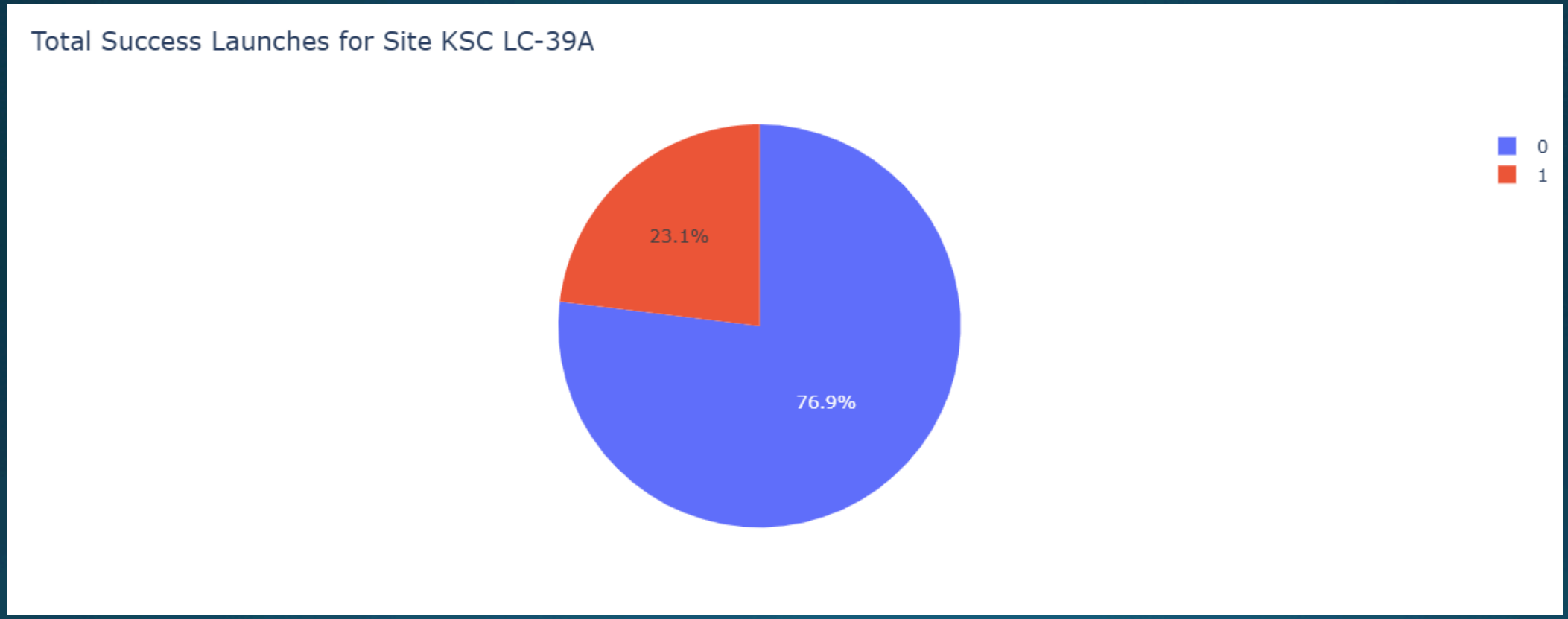
Build a Dashboard with Plotly Dash

Launch success for each Site



From this pie chart we can that KSC LC-39A has the most successful launches.

<Dashboard Screenshot 2>



From the chart we can observe that KSC LC-39A has the highest success rate of 76.9% with only 3 failed launches from 13 launches.

Correlation between Payload and Success

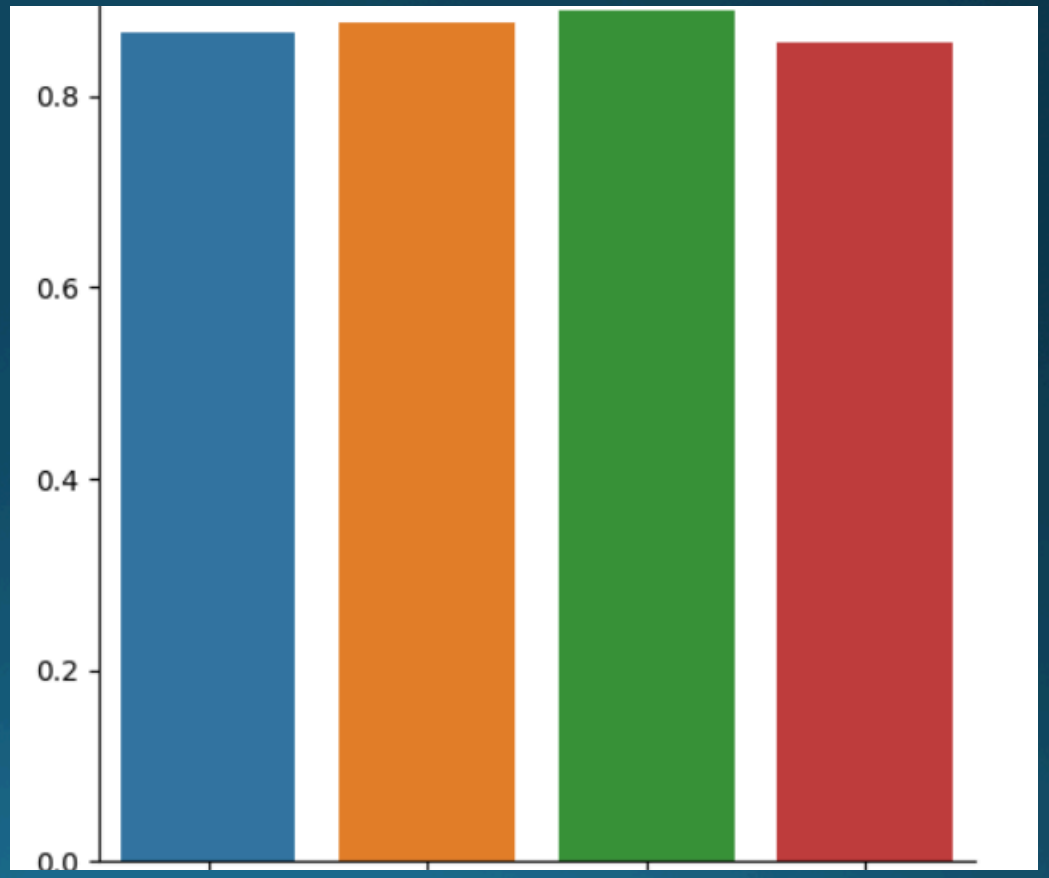
From these plots we can see that all the payload mass for site KSC LC-39A lie in the range 2400 to 6800.

Section 5

Predictive Analysis (Classification)

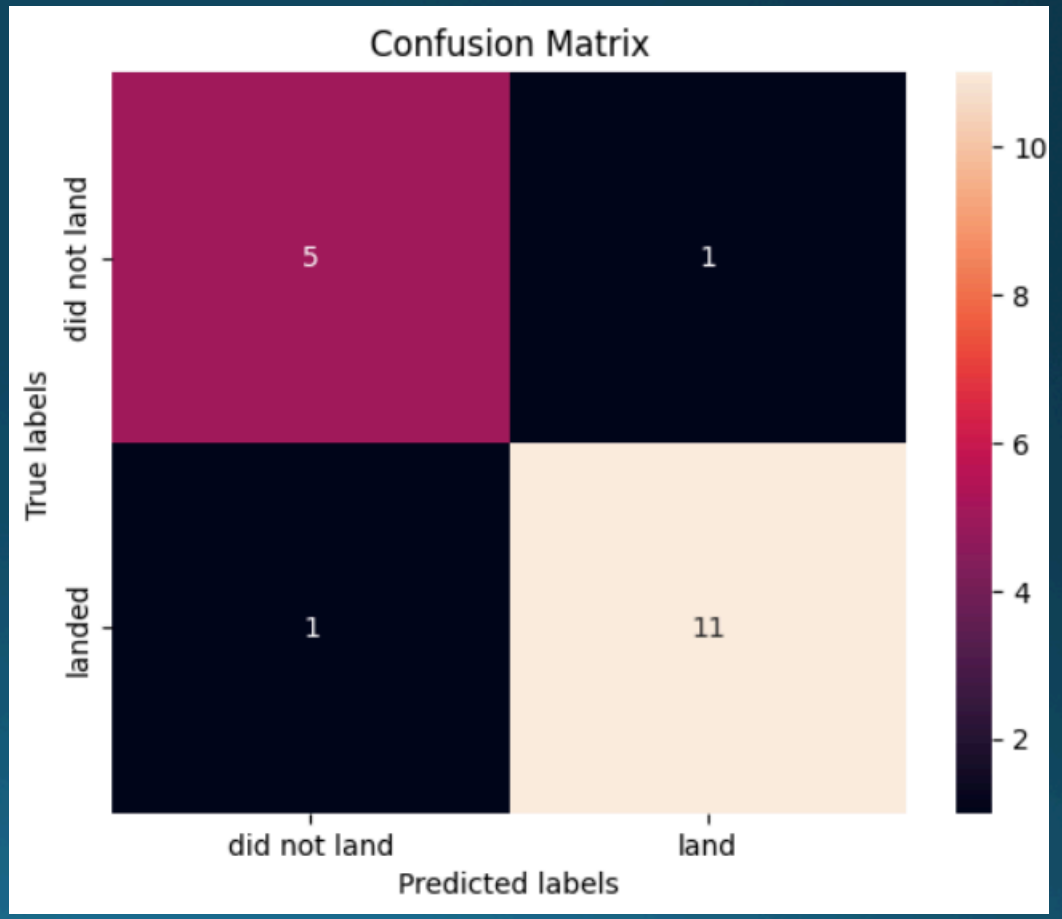
Classification Accuracy

This plot shows that Decision Tree Model has the highest accuracy.



Confusion Matrix

From this confusion matrix, we can observe that there is only one inaccuracy for both True Positive and False Negative



Conclusions

- The success rate of launches has increased over the years.
- Considering all factors we can conclude that the flights with lighter payload mass has higher success rate
- KSC LC-39A has the highest success rate of all the launch sites
- Most of the launch sites are situated close to Equator and much closer to coasts.
- For computing the results **Decision Tree Model** is the best algorithm

Appendix

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.843750	0.819444
F1_Score	0.909091	0.916031	0.915254	0.900763
Accuracy	0.866667	0.877778	0.888889	0.855556

This shows the scores for different models used in this project.

Thank you!