IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Sohail bounnite>
<10/21/2022>

Outline

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

Executive Summary

- **Summary of methodologies**
  - Data Collection (web scraping - spaceX API)
  - Data wrangling
  - EDA with Data Visualization
  - EDA With SQL
  - Building an interactive map with Folium
  - Building Dashboard with Plotly Dash
  - Predictive Analysis
- **Summary of all results**
  - EDA results
  - Interactive Analytics
  - Predictive Analytics

Introduction

- **Project background and context**

we will predict if the Falcon 9 first stage will land successfully. SpaceX 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.

- **Problems you want to find answers**

  - Determine if the first stage will land.
  - Determine the cost of a launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX API
  - Using web scraping from Wikipedia
- Perform data wrangling
  - Filtering the Data
  - Dealing with missing values
- 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 models and evaluating the different classification algorithms to ensure best results

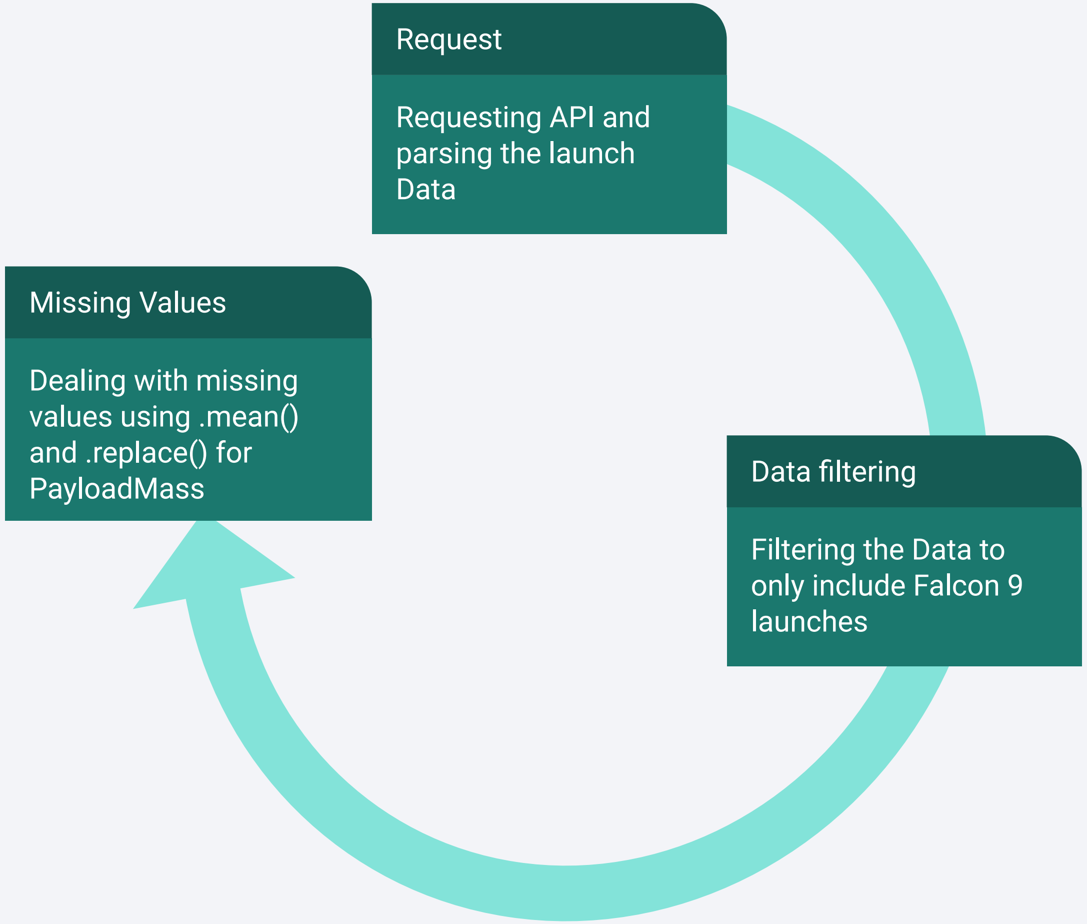
Data Collection

- Data sets that were used were collected from two main sources:
  - SpaceX API
  - Wikipedia using web scraping

Data Collection – SpaceX API

- Data columns By SpaceX API Are: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

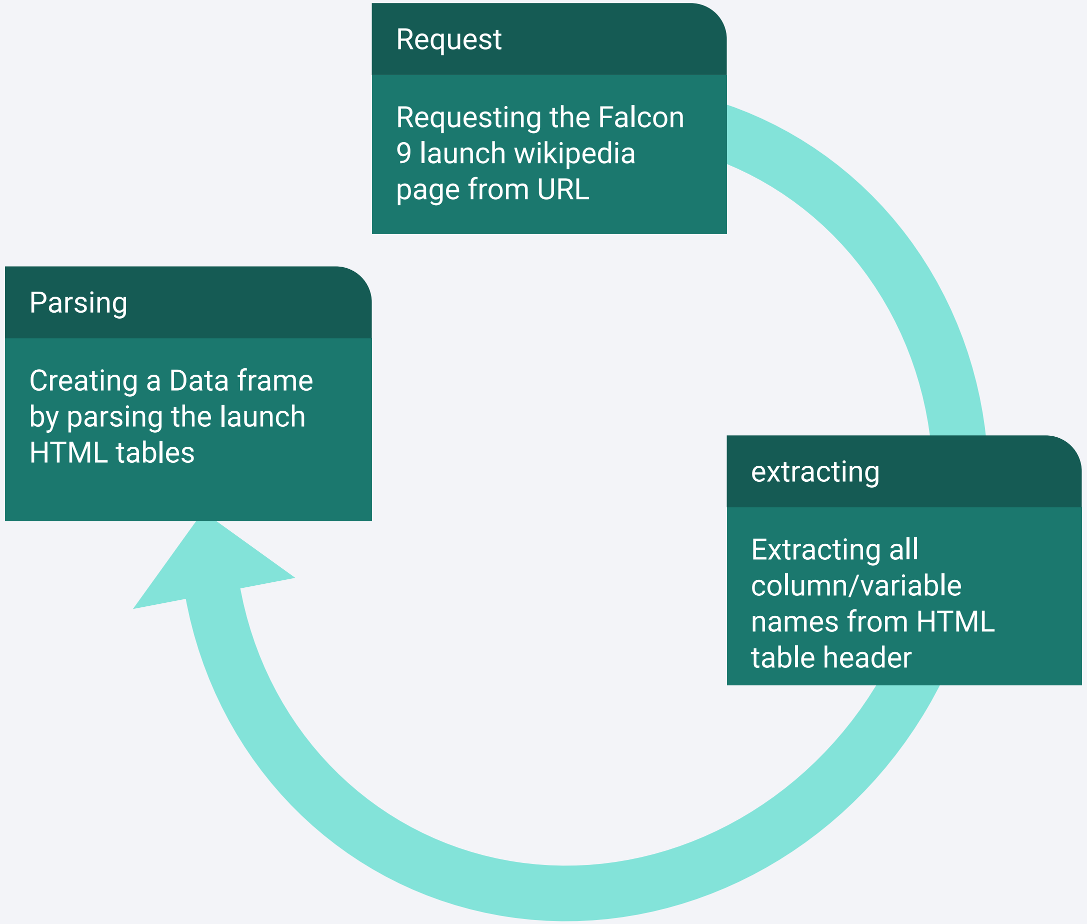
- Source code: https://github.com/derfal/IBM-Capstone-/blob/main/jupyter-labs-spacex-data-collection-api%20.ipynb



Data Collection - Scraping

- Data columns By Web Scraping Wikipedia Are: Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

- Source code: https://github.com/derfal/IBM-Capstone-/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- In this stage, we performed Some Exploratory Data Analysis (EDA) to find some patterns and determine what would be the label for training supervised models.

- 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;
    - 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.
    - 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.

source code: https://github.com/derfal/IBM-Capstone-/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- The chart that we plotted are the following:

  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

  Visualizing Data helps to tell story to make the relationships between the data that is plotted show up more clearly highlighting trends like in line charts as well as outliers in the case of scatterplots

  source code: https://github.com/derfal/IBM-Capstone-/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- the SQL queries that we performed are:
  - 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

source code: https://github.com/derfal/IBM-Capstone-/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters were used with Folium Maps to :
  - Markers indicate points like launch sites
  - Circles indicate highlighted areas around specific coordinates
  - Marker clusters indicates groups of events in each coordinate
  - Lines are used to indicate distances between two coordinates

source code: https://github.com/derfal/IBM-Capstone-/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- The plots/graphs and interactions that we added to the dashboard are:
    - Launch sites Dropdown list
    - Pie Chart showing Success Launches (All Sites/Certain Site)
    - Slider of Payload Mass Range
    - Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions

- source code: https://github.com/derfal/IBM-Capstone-/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

Data preparation and standardisation → Split the Data X and Y into training and test Data → Applying GridSearchCV on LogReg, SVM, Decision Tree, and KNN models → Finding the method performs best by examining the Jaccard_score and F1_score metrics

Source code : https://github.com/derfal/IBM-Capstone-/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

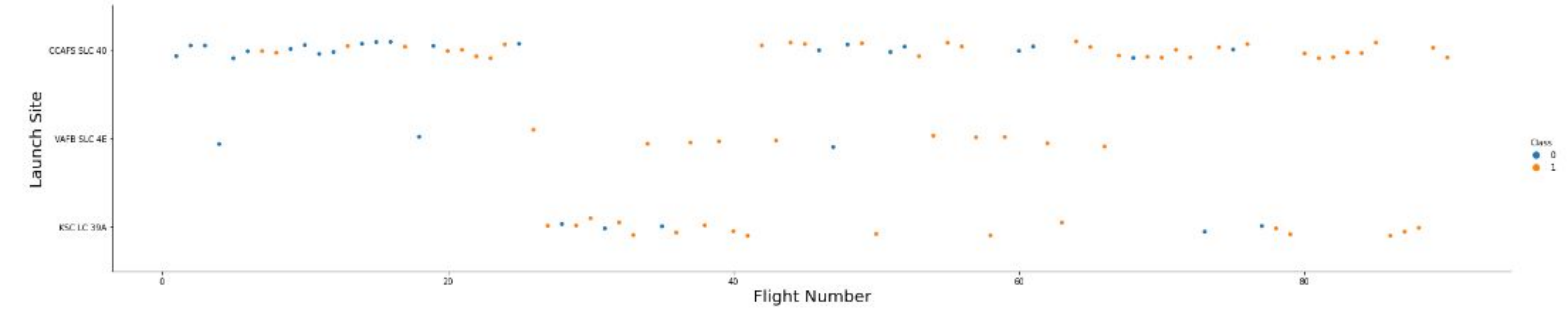
Results

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

Section 2

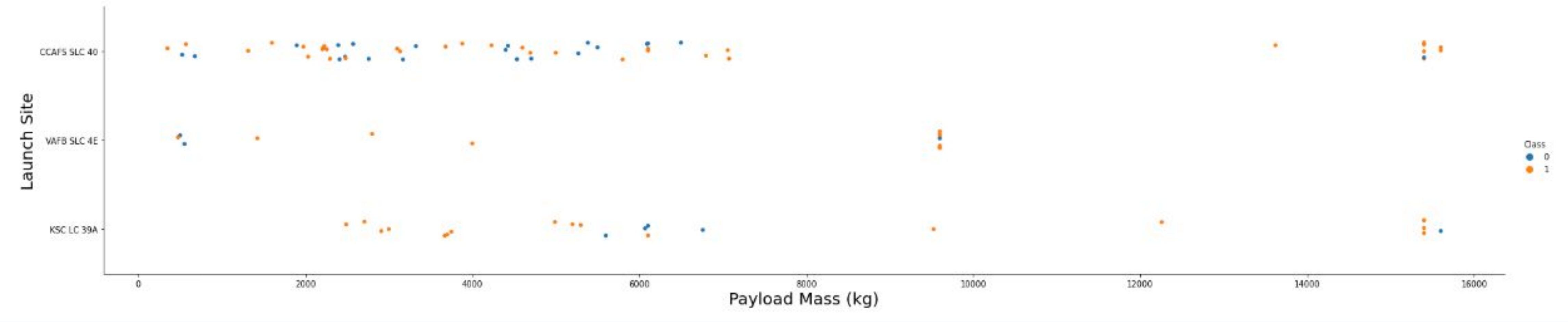
Insights drawn from EDA

Flight Number vs. Launch Site



- Launch Site CCAFS SLC 40 has the most launches
- Earliest launches were not successful but the latest were more successful
- VAFB SLC 4E has the highest success rate

Payload vs. Launch Site



- Overall the highest the payload the more successful the launch is
- KSC LC 39A is the exception for lower Payload Mass since it has a high success rate (100%) for Payload mass under 5500 Kg
- The launches above 7000Kg were more successful

Success Rate vs. Orbit Type

- Orbits with 100% success rate are:

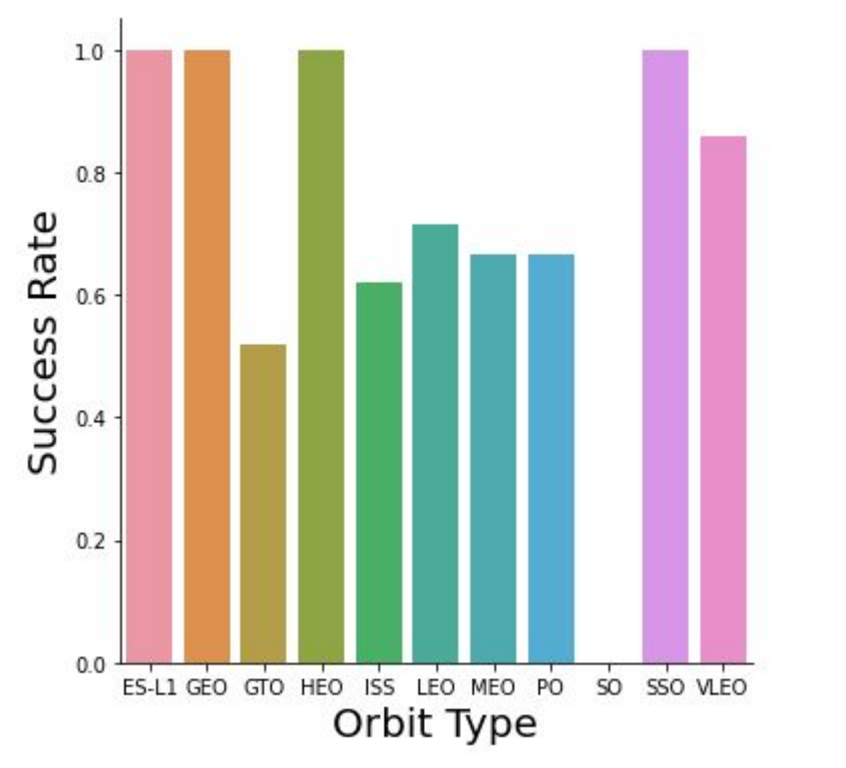
ES-L1 / GEO / HEO / SSO

- Orbits with 0% success rate are :

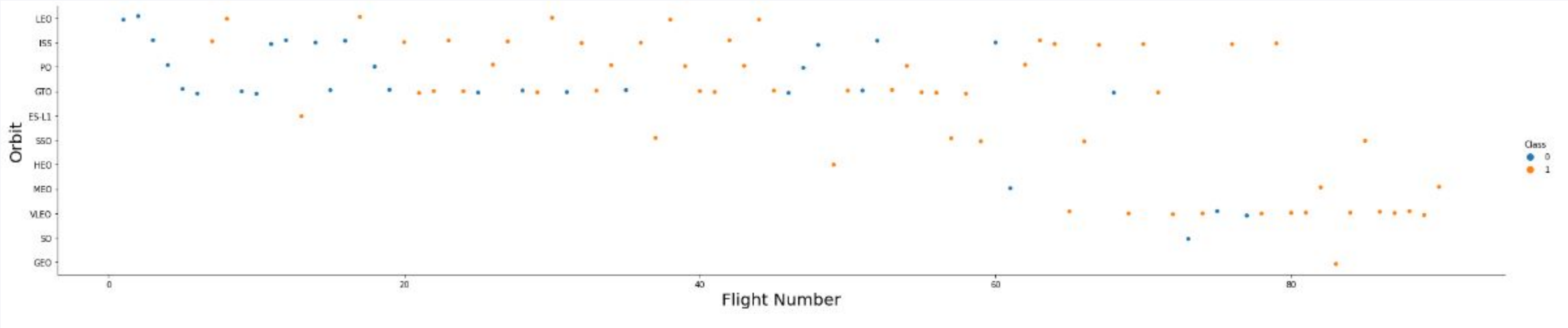
SO

- Orbits with success rate between 50% and 85% are:

GTO / ISS / LEO / MEO / PO / VLEO



Flight Number vs. Orbit Type



- Flight number is an important factor in some orbits like (LEO / ISS / PO) because the more flights we do the higher the success, while in other orbits like (GTO) it doesn’t make a difference.

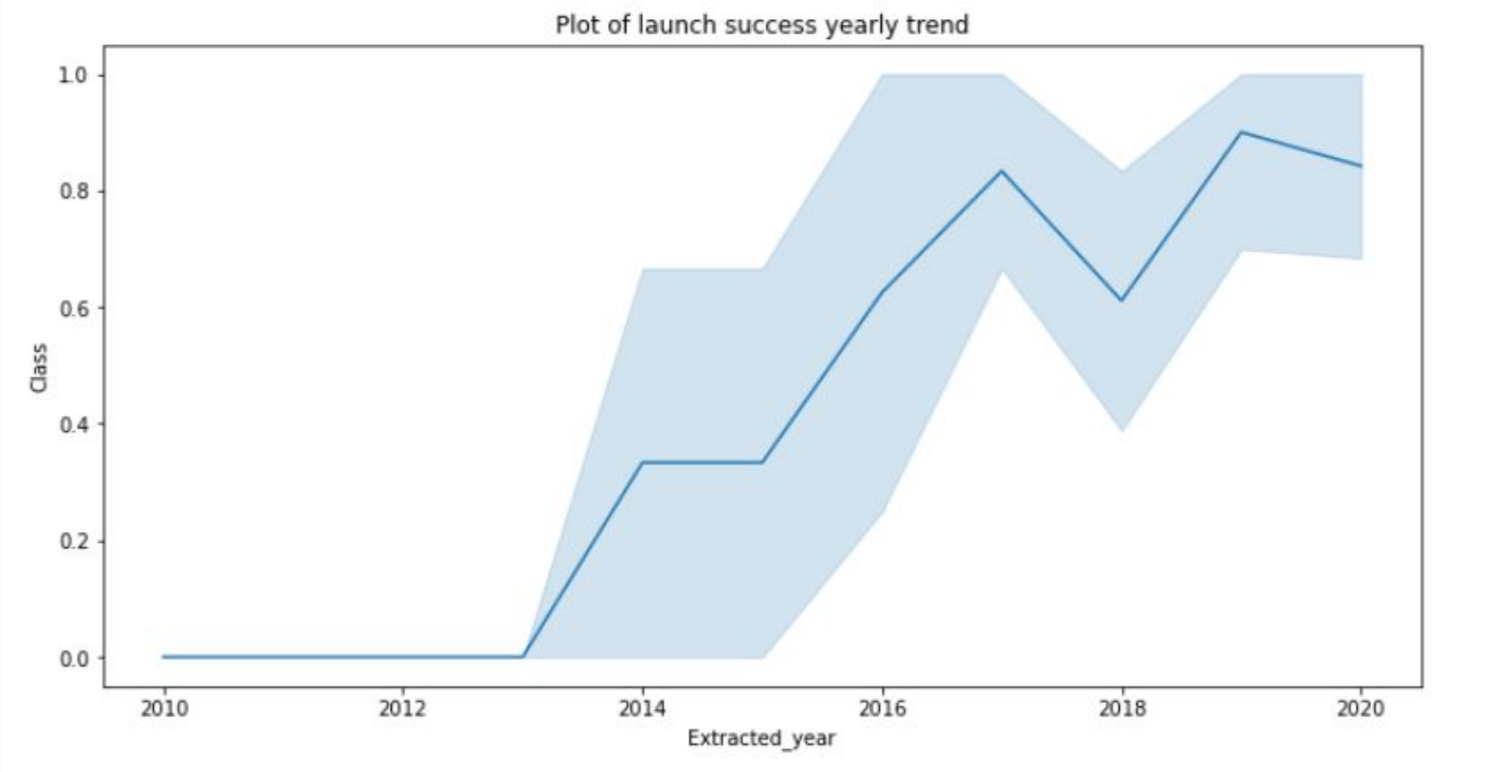
Payload vs. Orbit Type

- GTO and ISS have mixed results over relatively same payload mass as we can find success and fails in close payload mass ranges

- Other orbits tend to be more successful the more Payload Mass they have.

Launch Success Yearly Trend

- Between 2010 and 2013 the trend is flat at 0% success rate
- From 2013 to 2017 we see constant improvement in the success rate to reach 80% success rate
- In 2018 we see a sudden drop in success rate 60% before going up again until 2020



All Launch Site Names

```
%sql select distinct Launch_Site from SPACEXTBL

 * ibm_db_sa://zqj61447:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
   sqlite:///my_data1.db
Done.
```

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

- We used the keyword DISTINCT to find the names of launch sites without repetition
- The results are : CCAFS LC-40 / CCAFS SLC-40 / KSC LC-39A / VAFB SLC-4E

Launch Site Names Begin with 'CCA'

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

 * ibm_db_sa://zqj61447:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
   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	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

- We re displaying 5 records where launch site begin with ‘CCA’ using (like ‘CCA%’)

Total Payload Mass

```
%sql select sum(payload_mass__kg_) from SPACEXTBL WHERE customer = 'NASA (CRS)'
```

```
 * ibm_db_sa://zqj61447:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
   sqlite:///my_data1.db
Done.
```

1
45596

- We use the function sum to calculate the total Payload Mass by summing all payload whose codes contain CRS which corresponds to NASA

Average Payload Mass by F9 v1.1

```
%sql select avg(payload_mass__kg_) from SPACEXTBL WHERE booster_version = 'F9 v1.1'
```

```
 * ibm_db_sa://zqj61447:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
   sqlite:///my_data1.db
Done.
```

1
2928

- By using the function avg we can calculate the average. we used it to calculate the average Payload Mass by the booster version F9 v1.1 which is 2928 Kg.

First Successful Ground Landing Date

```
%sql SELECT MIN(Date) AS FirstSuccessfull_landing_date FROM SPACEXTBL WHERE Landing__outcome LIKE 'Success (ground pad)'
```

```
 * ibm_db_sa://zqj61447:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
   sqlite:///my_data1.db
Done.
```

firstsuccessfull_landing_date
2015-12-22

- By filtering data by successful landing outcome on ground pad and getting the minimum value for date we identified the first successful landing that occurred on 12/22/2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND LANDING__OUTCOME = 'Success (drone ship)';
```

```
 * ibm_db_sa://zqj61447:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
   sqlite:///my_data1.db
Done.
```

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

- We have identified the different successful Drone Ship landings with Payload Between 4000 Kg and 6000 Kg

Total Number of Successful and Failure Mission Outcomes

```
%sql select mission_outcome, count(mission_outcome) from SPACEXTBL GROUP BY mission_outcome
```

```
 * ibm_db_sa://zqj61447:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
   sqlite:///my_data1.db
Done.
```

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- From the query we got the total number of successful and failure mission outcomes which is shown in the table above.

Boosters Carried Maximum Payload

```
%sql select booster_version, payload_mass__kg_ from SPACEXTBL where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTBL)
```

```
 * ibm_db_sa://zqj61447:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
   sqlite:///my_data1.db
Done.
```

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

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

```
%sql select booster_version, launch_site from SPACEXTBL where landing__outcome = 'Failure (drone ship)' and year(DATE) = 2015
```

```
 * ibm_db_sa://zqj61447:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
   sqlite:///my_data1.db
Done.
```

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- Above we listed the failed landing outcomes in drone ship, their booster versions, and launch site names for the year 2015

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

```
%sql SELECT LANDING__OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING__OUTCOME ORDER BY QTY D
```

```
 * ibm_db_sa://zqj61447:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
   sqlite:///my_data1.db
Done.
```

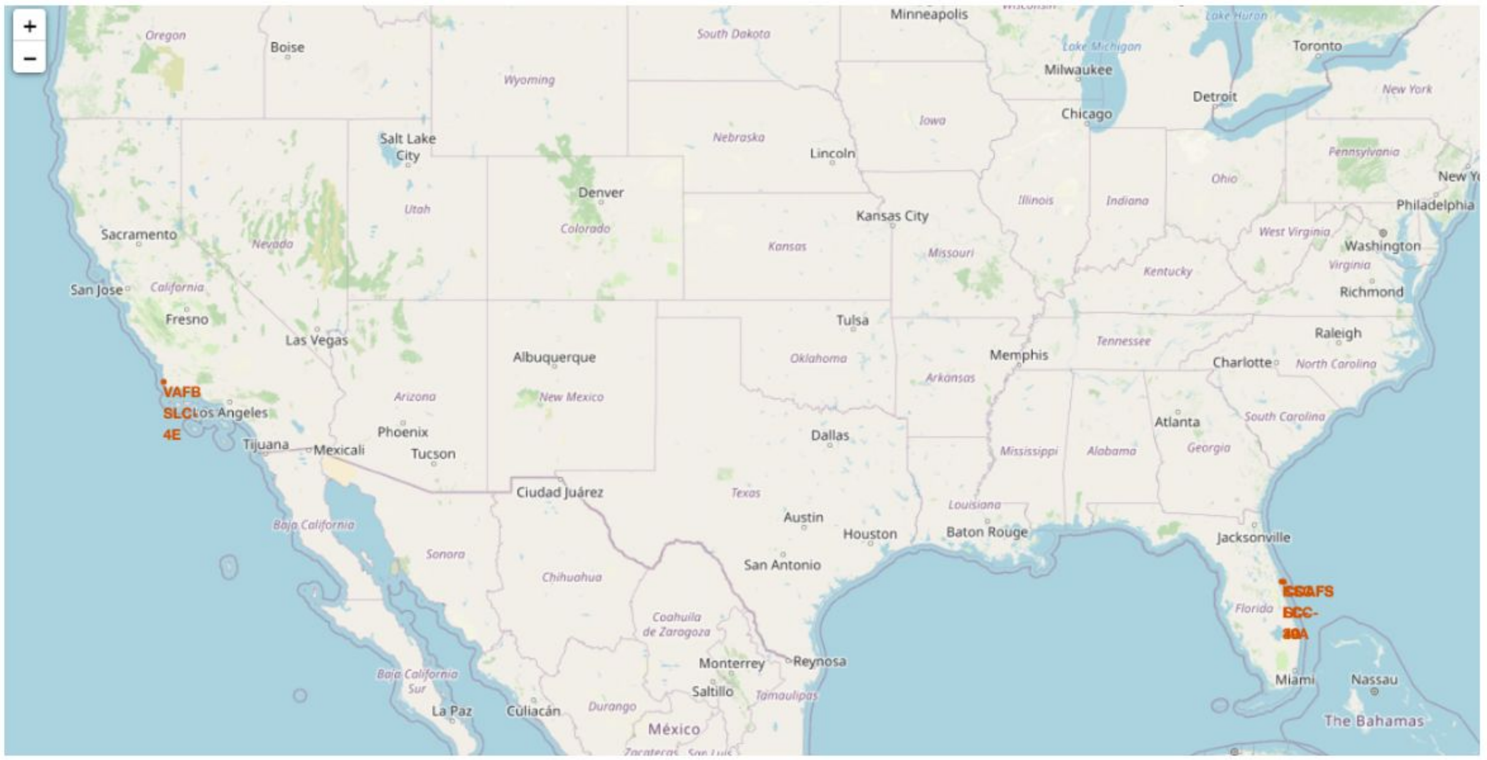
landing__outcome	qty
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

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

Section 3

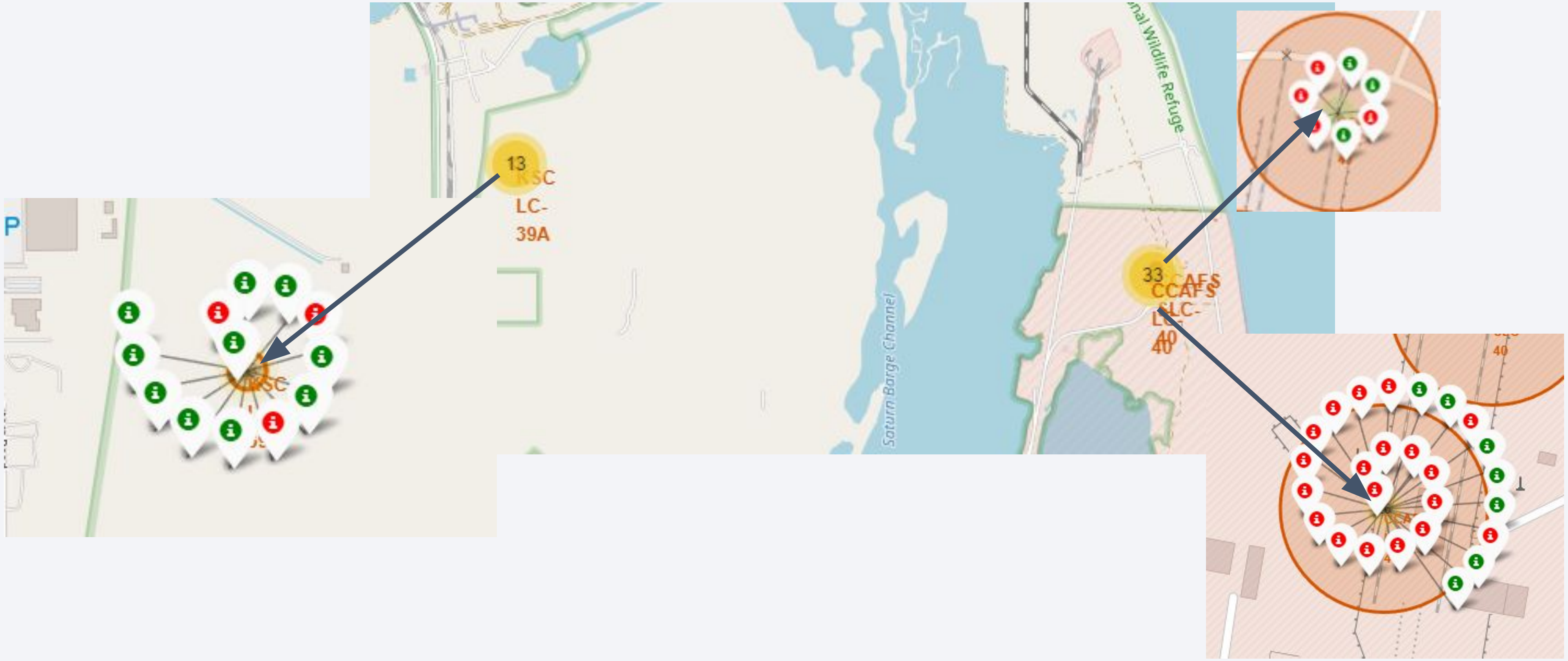
Launch Sites Proximities Analysis

All site launches



- All the launch sites are in locations near the sea far from residential and crowded areas. this choice is to avoid any accident or debris from the launches.

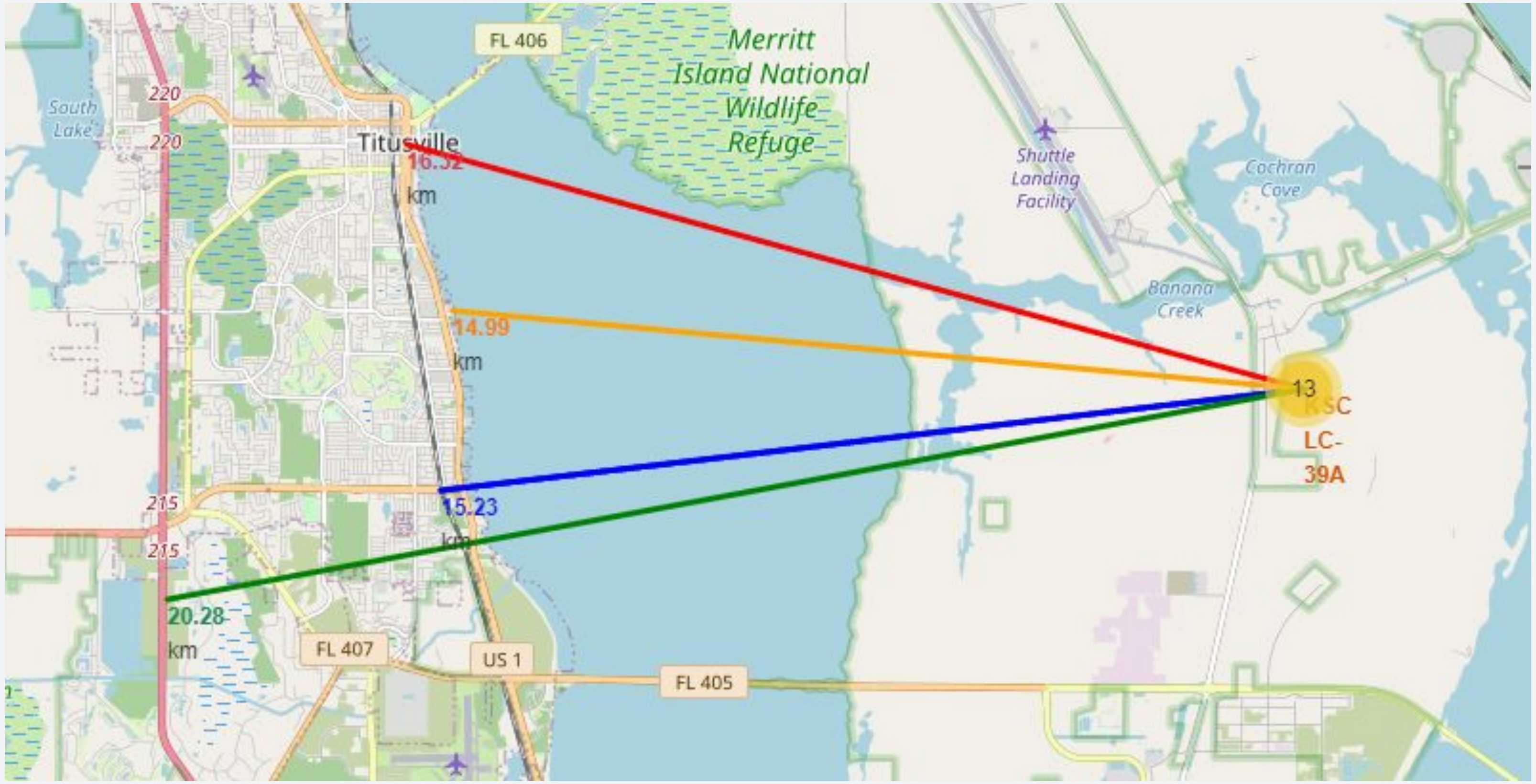
The success/failed launches on the map



- The Map shows different launch site with the markers (Green for successful and Red for failed launches)
- The site KSC LC-39A as shown on the map above has the highest successful launches

Distance between launch site to its proximities

- The map shows the distance between the launch site and:
  - City 16.32Km
  - Highway 15.23Km
  - Railway 20.28Km
  - Coastline 14.99Km

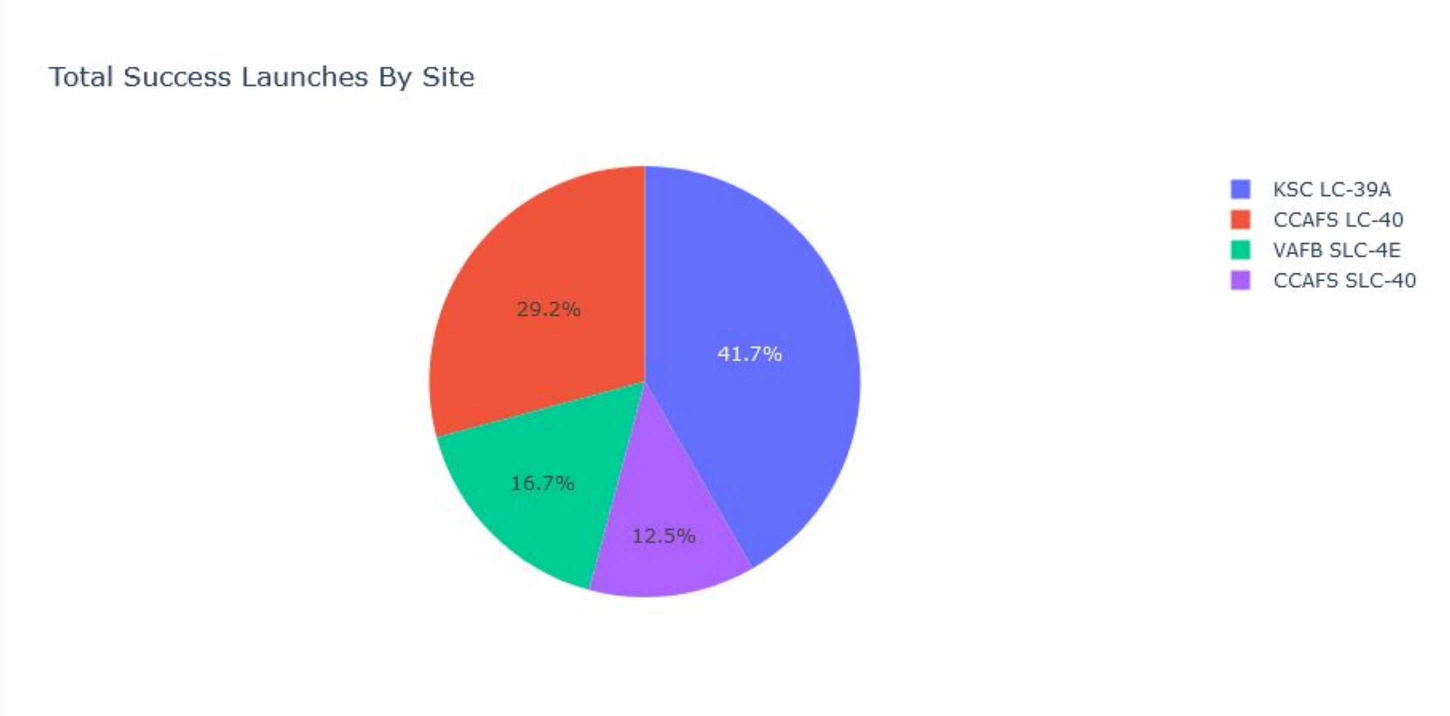


Section 4

Build a Dashboard with Plotly Dash

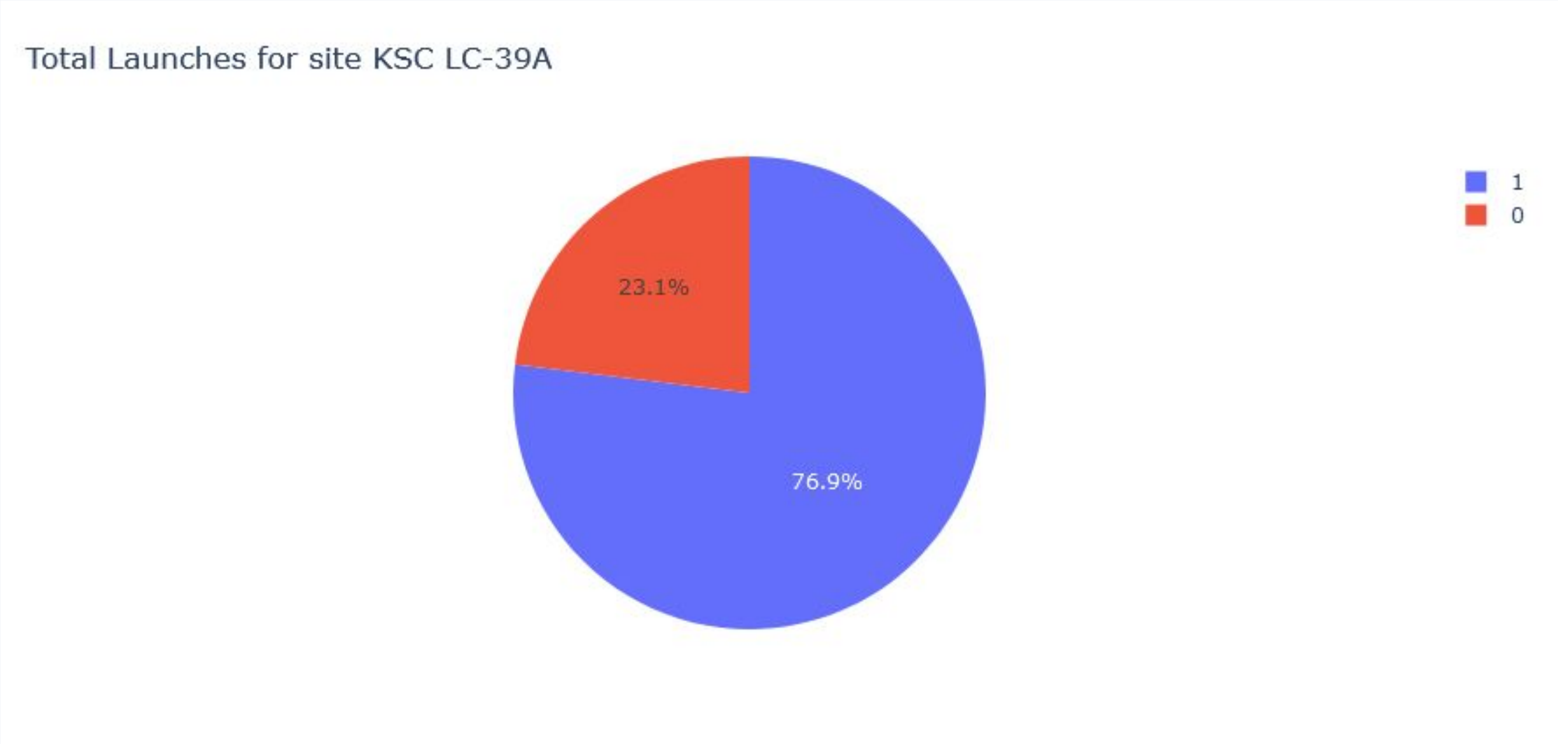
Total success launches by Site

- The Pie chart shows that KSC LC-39A has the highest success rate by 41.7% followed by CCAFS LC-40 with 29.2%

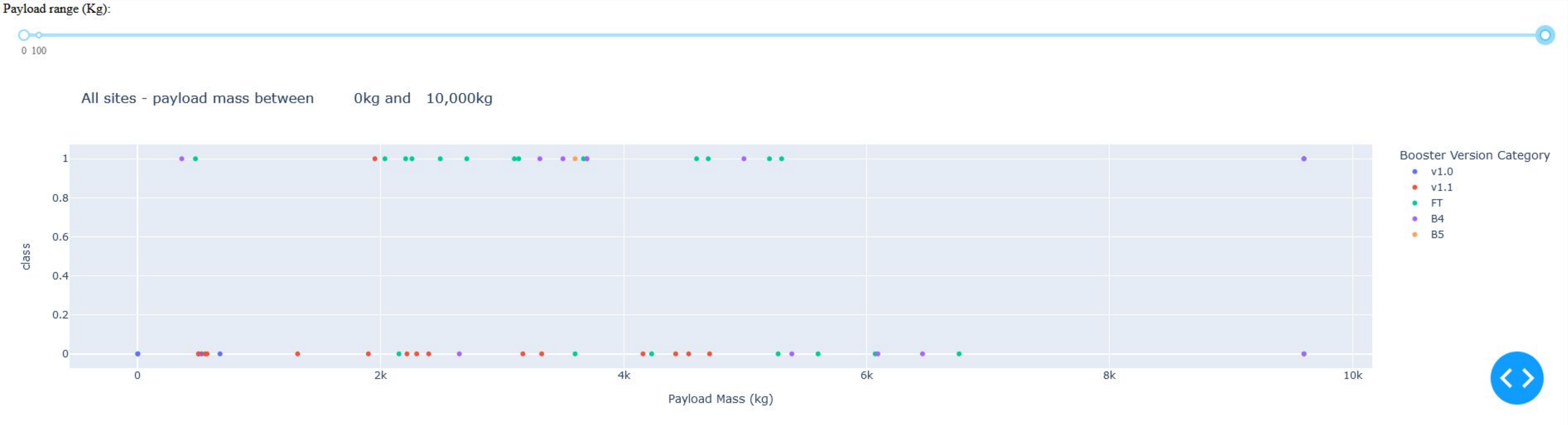


Total launches for site KSC LC-39A

- By Analyzing the Site KSC LC-39A we can get deeper and see that the site has a success rate of launches of 76.9% and only a rate of 23.1% of unsuccessful launches.

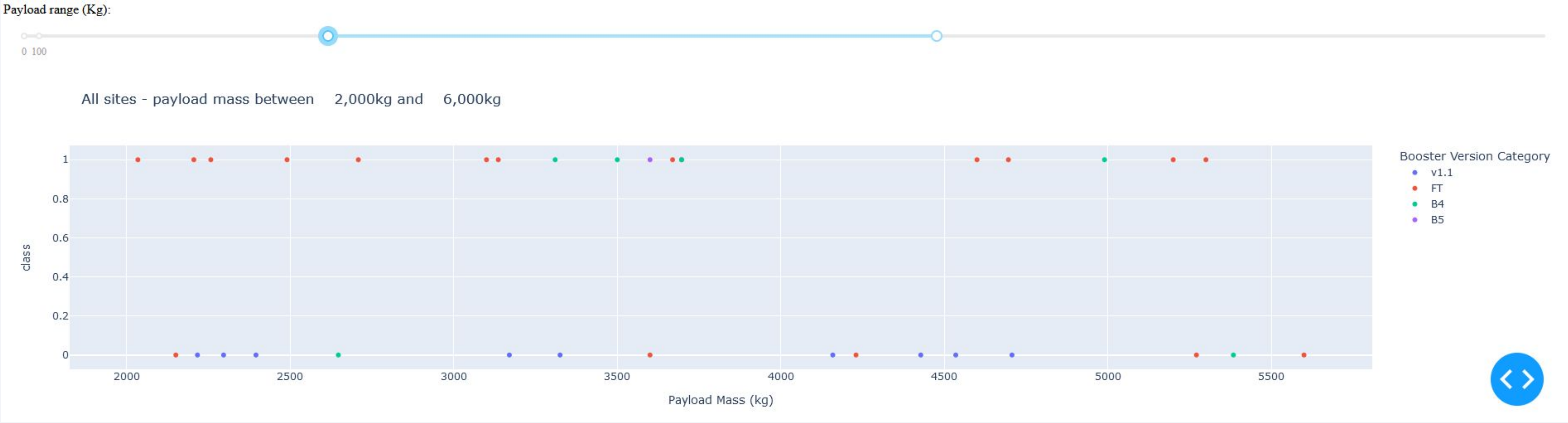


Success rate across all payload range



- The chart shows the success rate for all the booster version categories across all payload range.
- we can see that by narrowing the range we can have more success rates.

Success rate between 2000 Kg and 6000 Kg



- we can see that by narrowing the payload range between 2000 Kg and 5500 Kg we get to see more successful rates.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- 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	0.722222	0.833333

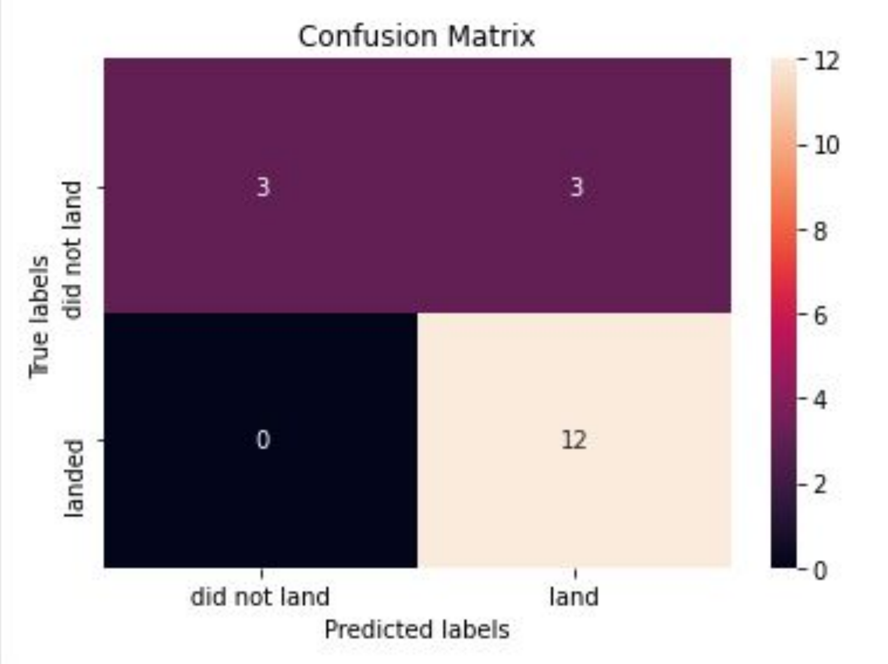
Scores and Accuracy of the Entire Data Set

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.895522	0.819444
F1_Score	0.909091	0.916031	0.944882	0.900763
Accuracy	0.866667	0.877778	0.922222	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.

	Negative	Positive
Negative	**TN** True Negative	**FP** False positive
Positive	**FN** False Negative	**TP** True Positive



Conclusions

- Decision tree model gave the best results for this particular Dataset
- Launches with Payload between 2000 Kg and 5500 Kg have the highest success rate
- Success rate increases over the years
- KSC LC-39A has the highest success rate of launches across sites
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate and VLEO is worth mentioning since it has a high success rate for number of launches
- All launch sites are close to the coast for safety purposes

Thank you!