IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Szilard Fazakas
03.11.2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection
  - Data wrangling
  - EDA with data visualization
  - EDA with SQL
  - Building an interactive map with Folium
  - Building Dashboard with Plotly Dash
  - Machine Learning (predictive analysis)
- Summary of all results
  - EDA results
  - Interactive analysis
  - Predictive analysis (classification)

Introduction

- Project background and context

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. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

- Problems you want to find answers

In this project we need to find the answer to the question, if the first stage will land successfully.

- What factors influence the successful landing?
- Relation between various features that determine the success

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using Space X API and web scraping from Wikipedia.
- Perform data wrangling
  - One-hot encoding was performed on various features for Machine Learning and cleaning data of null values and not useful columns.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification model
  - Perform exploratory Data Analysis and determine Training Label
  - KNN, Tree, Logistic Regression and SVM models have been built and evaluated to find the best classifier.

Data Collection

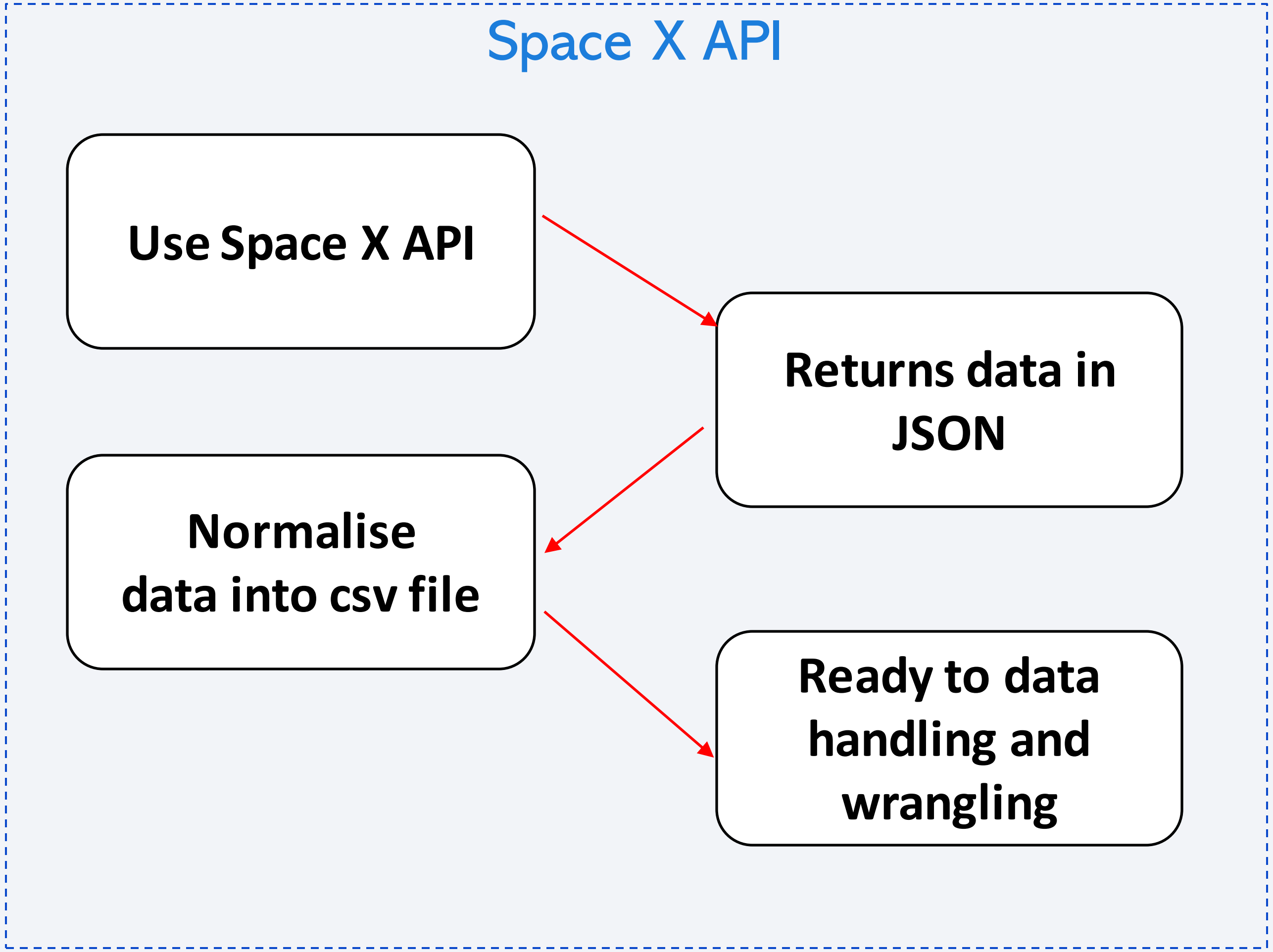
Data was collected using different methods

- First stage collection was done by using Space X API.
- Next the collected data is in Json format, we had to perform a conversion to Dataframe to make it more understandable and easier to work with.
- After the conversion, we need to clean the data, checking for missing information and filling in where it is necessary.
- We performed web scraping in addition collected data from Wikipedia.
- The reason to collect data from multiple sources was to get more accurate results.
- Web scraping extract the launch record in HTML format, where we need to find the necessary information and convert it to Dataframe for further analysis.

Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

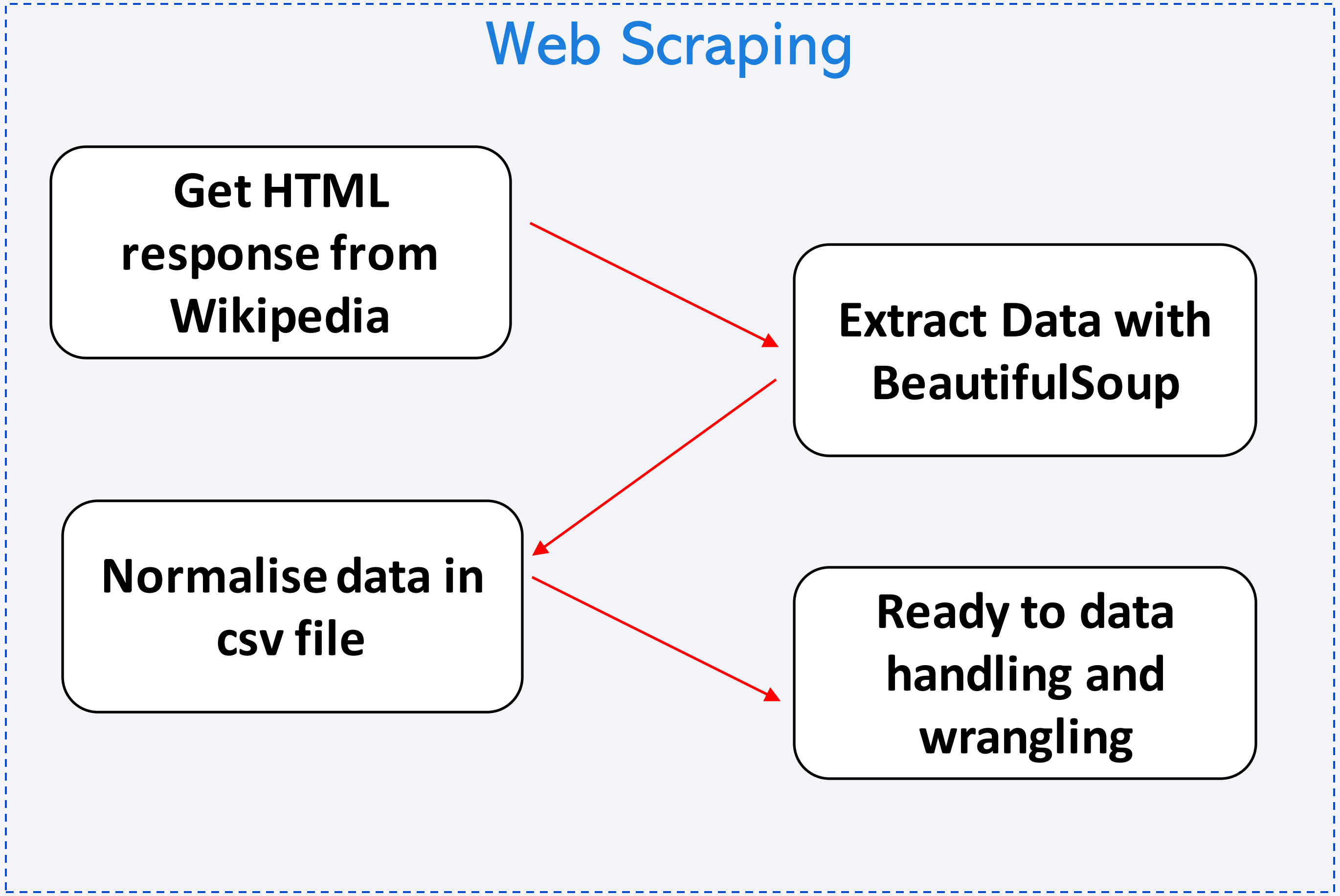
- ***Link to the Notebook***



Data Collection - Scraping

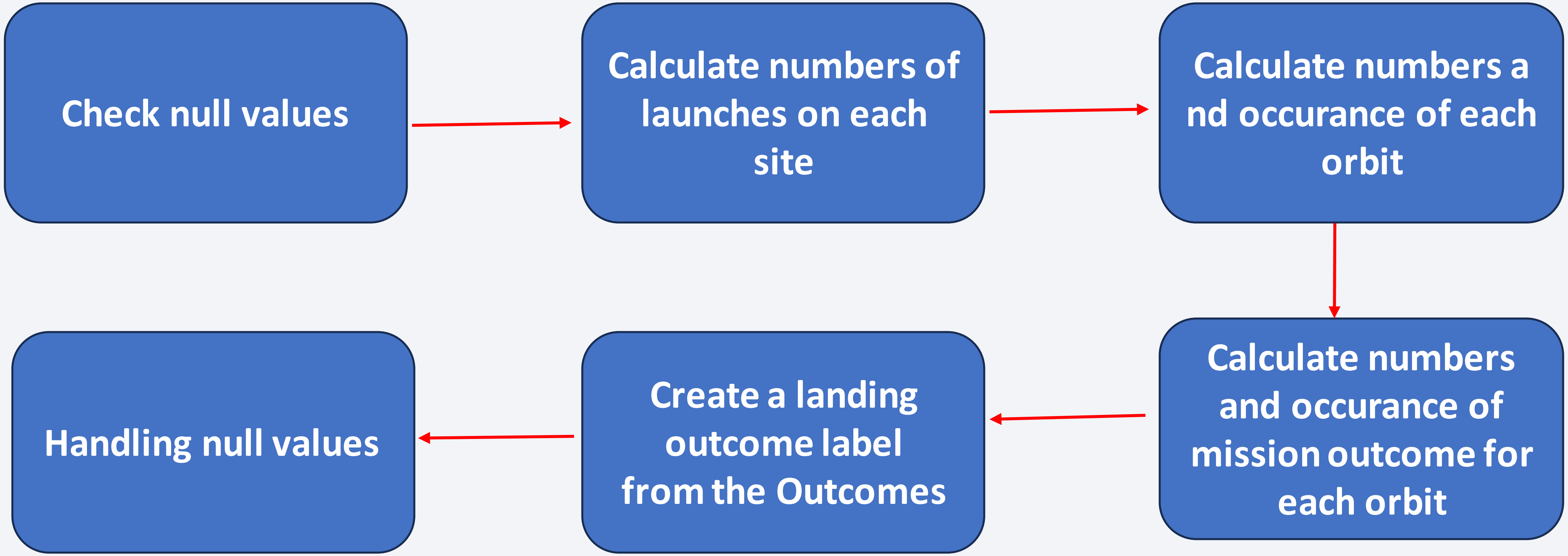
- Presenting web scraping, process using key phrases and flowcharts

- [Link to the web scraping Notebook]()

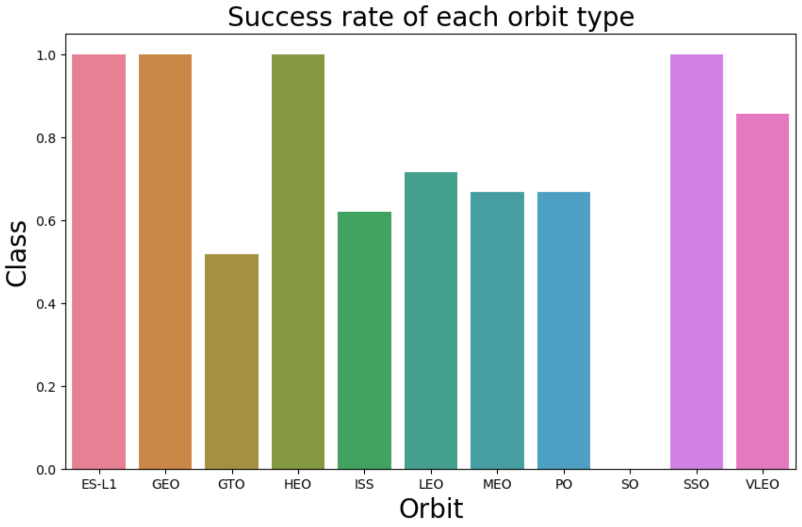


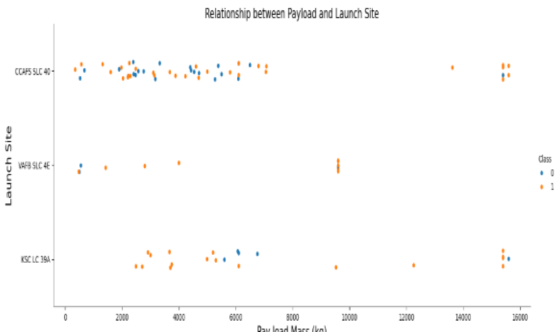
Data Wrangling

EDA Analysis

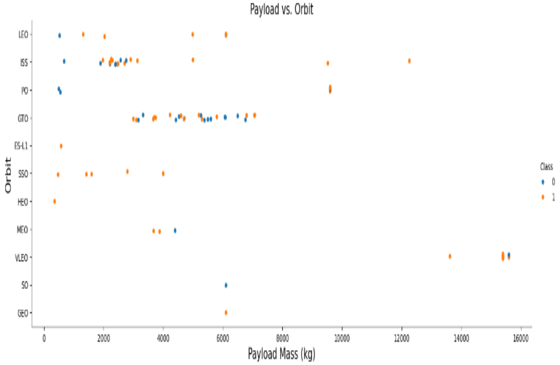


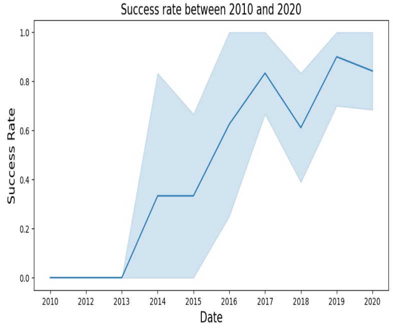
[LInk to Data Wrangling]

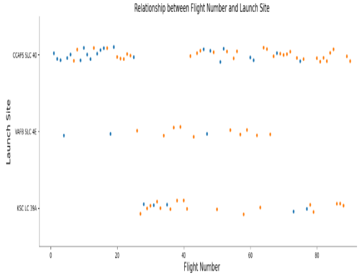


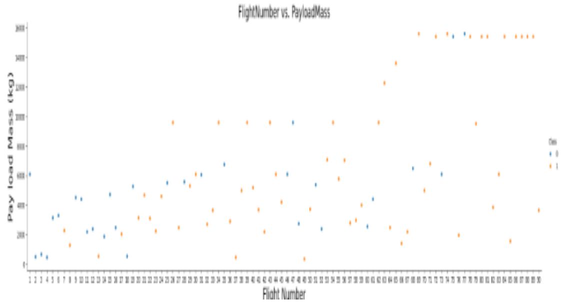


EDA with Data Visualization







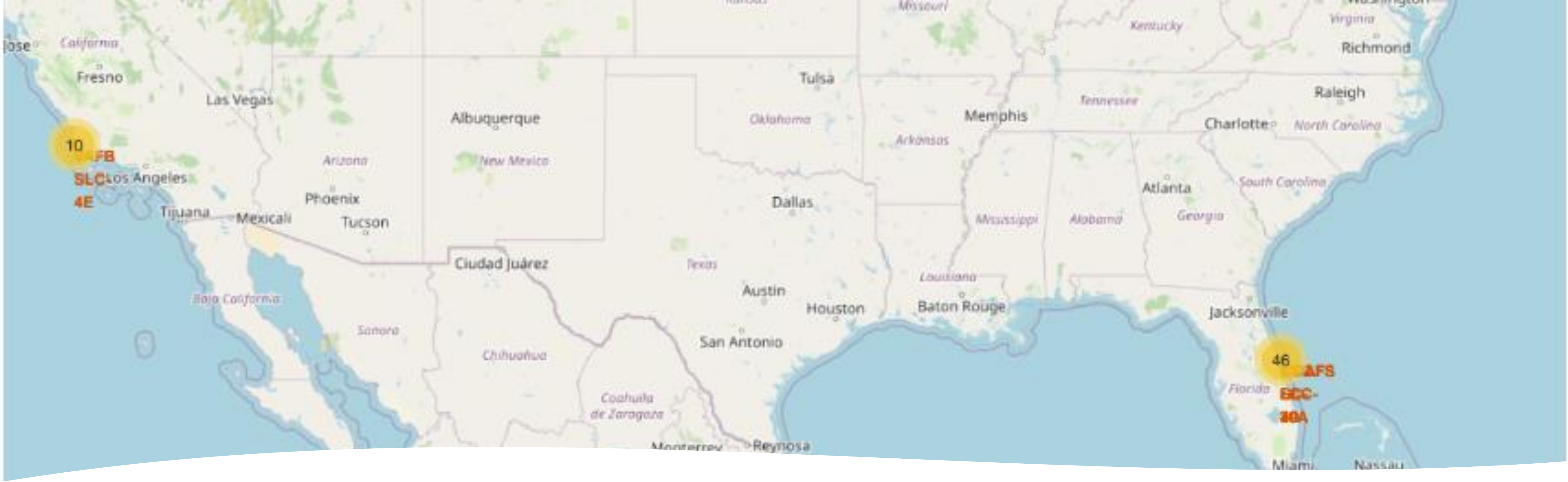


[Link to the visualization Notebook]()

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first succesful landing outcome in ground pad was acheived.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
  - List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
  - Rank 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.

[Link to the SQL Notebook]



Build an Interactive Map with Folium

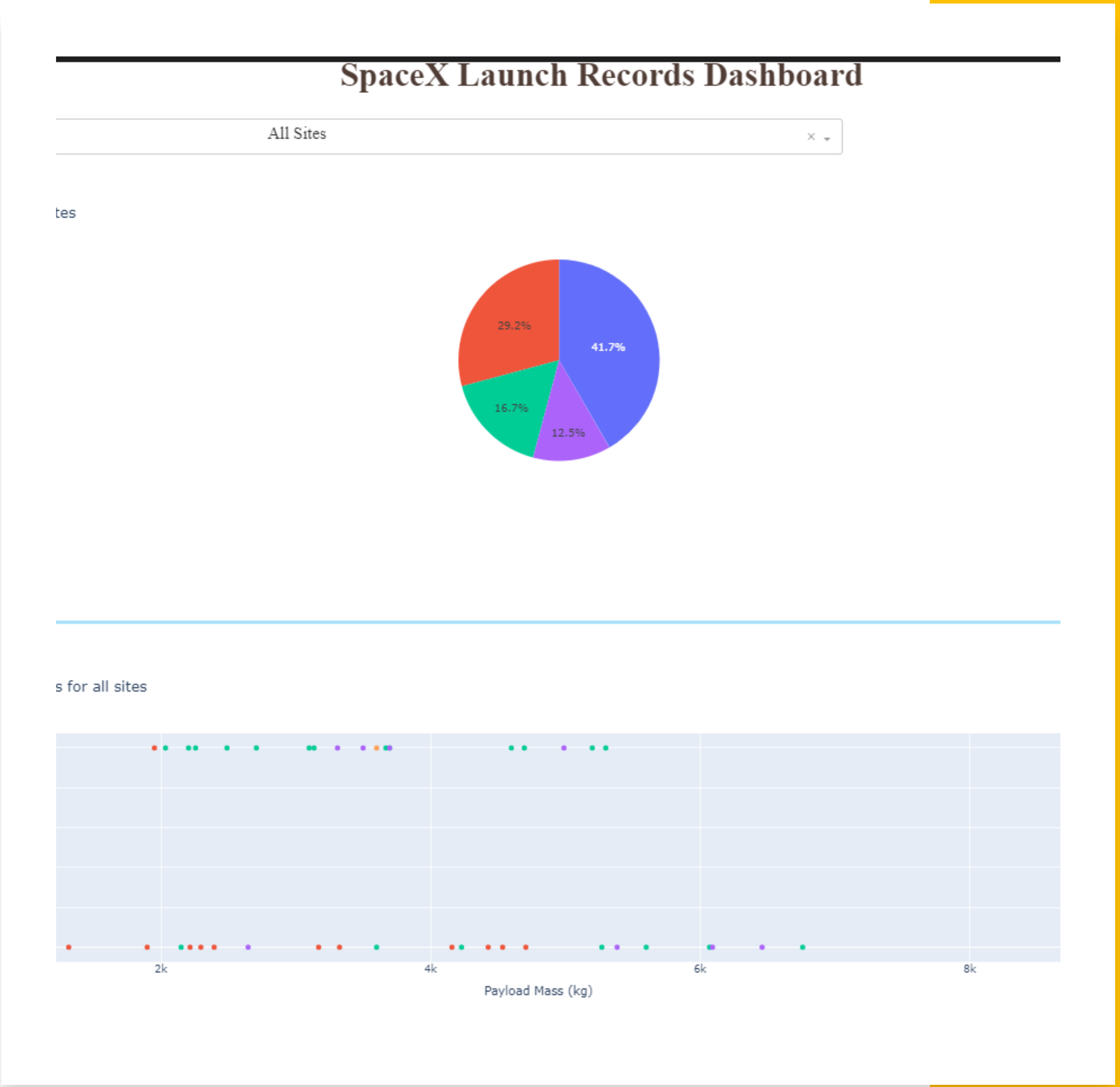
- Map markers, distance line have been added to the map to find an optimal launch location.

- [Link to the Interactive Map]()

Build a Dashboard with Plotly Dash

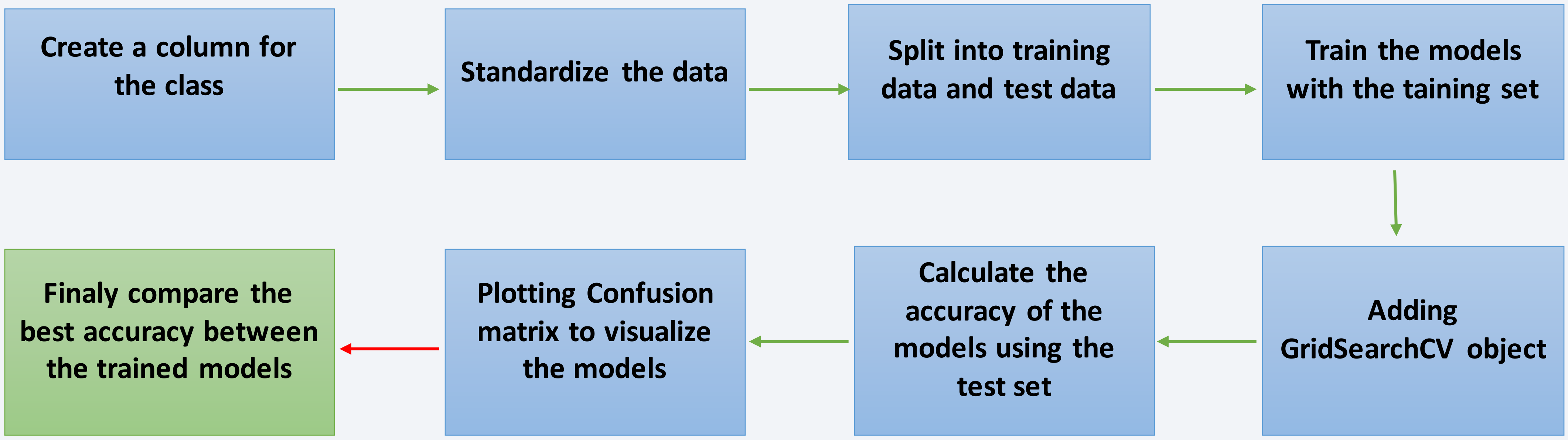
- A dashboard have been built for an interactive visualization.
  - Pie chart to show all successful launches
    (added dropdown menu to choose site)
  - Scater plot to show payload mass (added setable payload mass / kg scale)

[Link to the dashboard](#)



Predictive Analysis (Classification)

To find best Hyperparameter for SVM, Classification Trees and Logistic Regression performed a couple a steps:



[Link to the Predictive Analysis Notebook]

Results

- In this project there were four models trained SVM, Logistic Regression, KNN, and Regression Tree. All of them perform with an accuracy above 80%.
- I would like to highlight the Regression Tree achieved the best score of 89%
- The most successful launch site is KSC LC 39A
- Orbits with the highest successful rate are GEO, HEO, SSO, ES L1

```
print("tuned hpyerparameters :(best parameters) ",knn_cv.best_params_)
print("accuracy :",knn_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}
accuracy : 0.8482142857142858
```

```
print("tuned hpyerparameters :(best parameters) ",svm_cv.best_params_)
print("accuracy :",svm_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}
accuracy : 0.8482142857142856
```

```
print("tuned hpyerparameters :(best parameters) ",tree_cv.best_params_)
print("accuracy :",tree_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'criterion': 'entropy', 'max_depth': 4, 'max_features': 'auto', 'min_samples_lea
f': 2, 'min_samples_split': 5, 'splitter': 'random'}
accuracy : 0.8892857142857142
```

```
print("tuned hpyerparameters :(best parameters) ",logreg_cv.best_params_)
print("accuracy :",logreg_cv.best_score_)
```

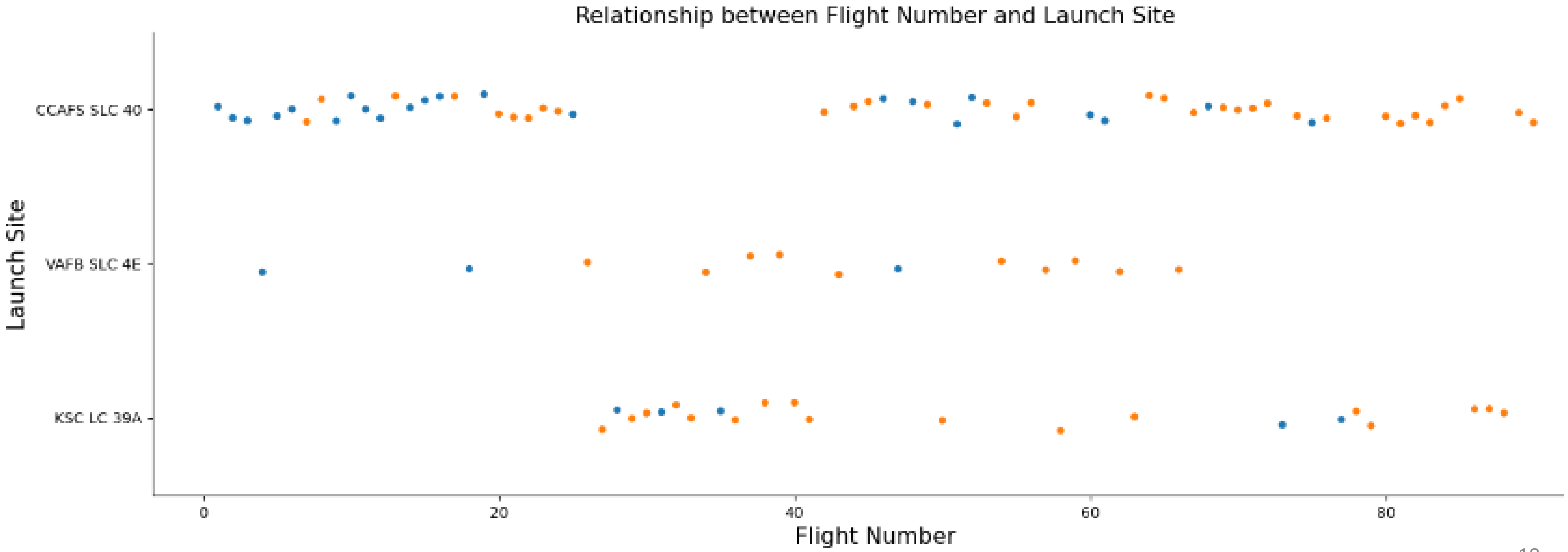
```
uned hpyerparameters :(best parameters)  {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}
ccuracy : 0.8464285714285713
```

Section 2

Insights drawn from EDA

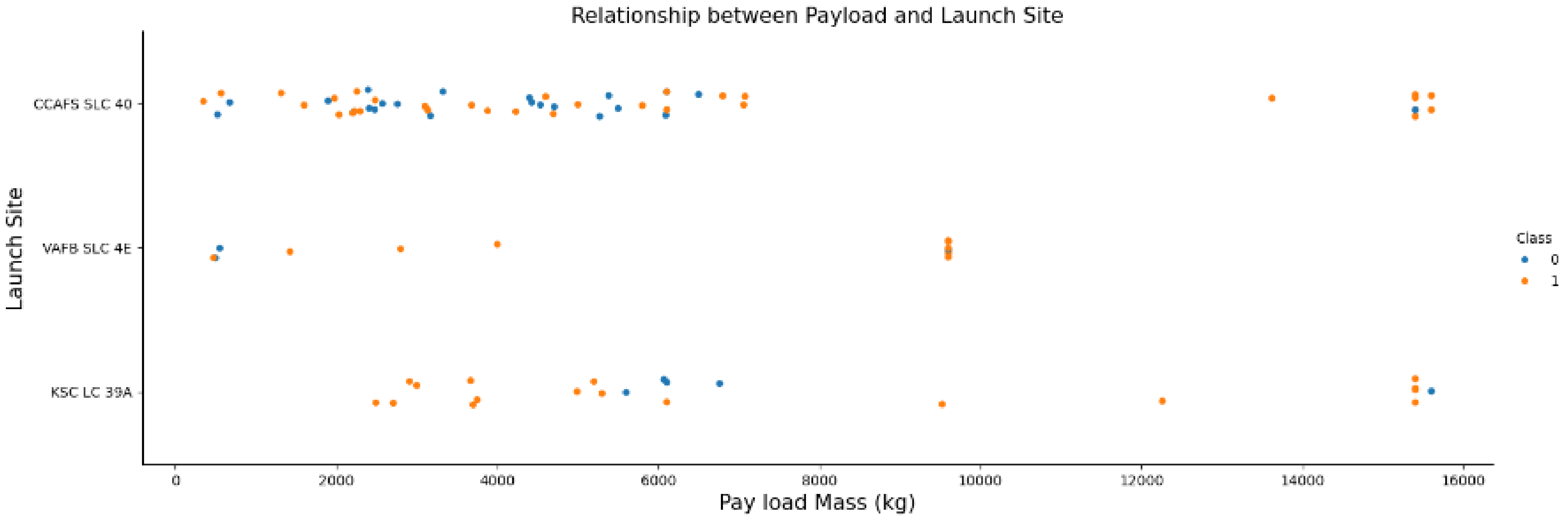
Flight Number vs. Launch Site

- This scatter plot shows a comparation between launch sites, clearly shows that CCAFS SLC 40 has the highest launch number



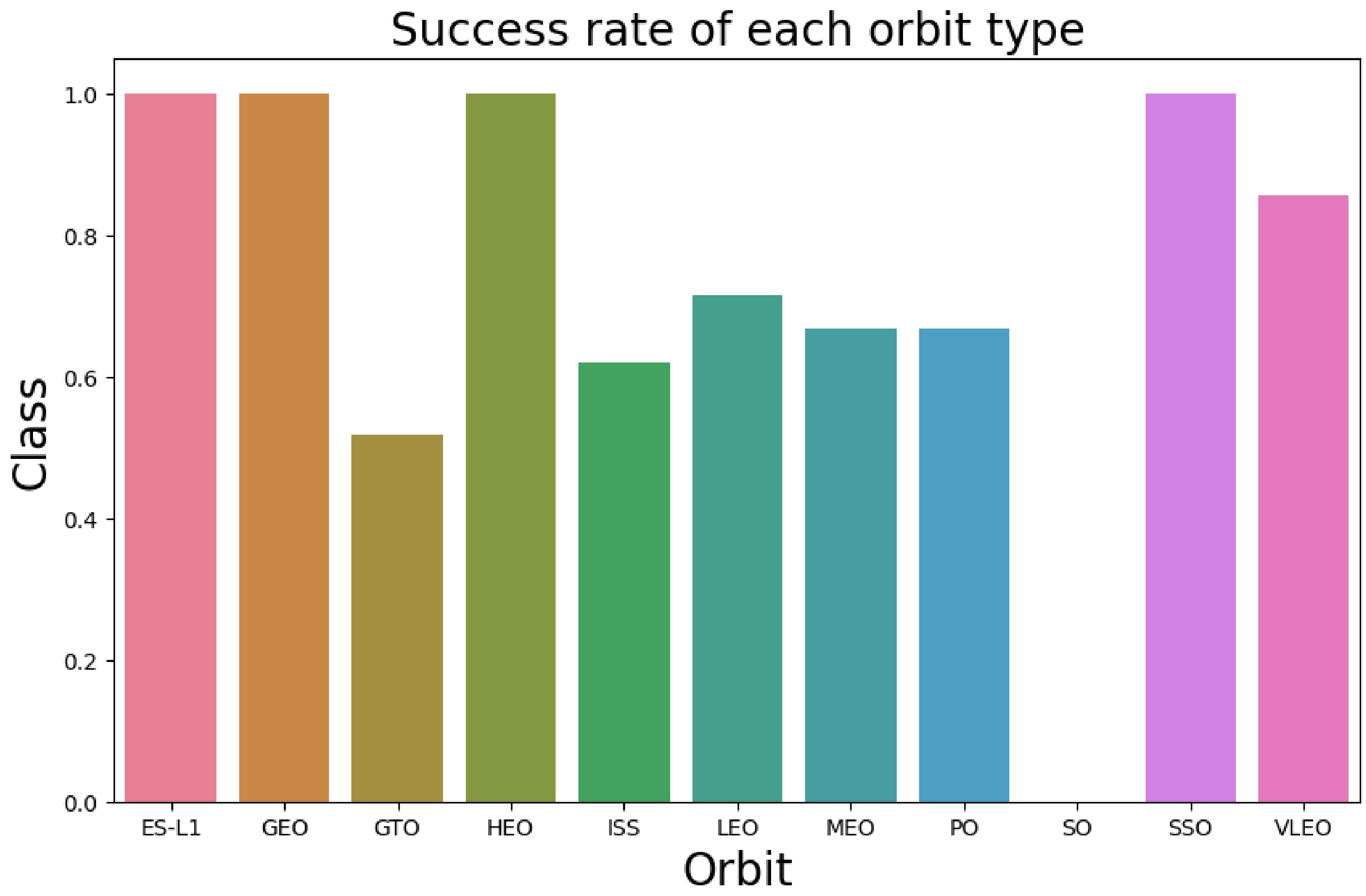
Payload vs. Launch Site

The most of the launches with less than 8000 payload mass (kg) have been launched from CCAFS SLC 40



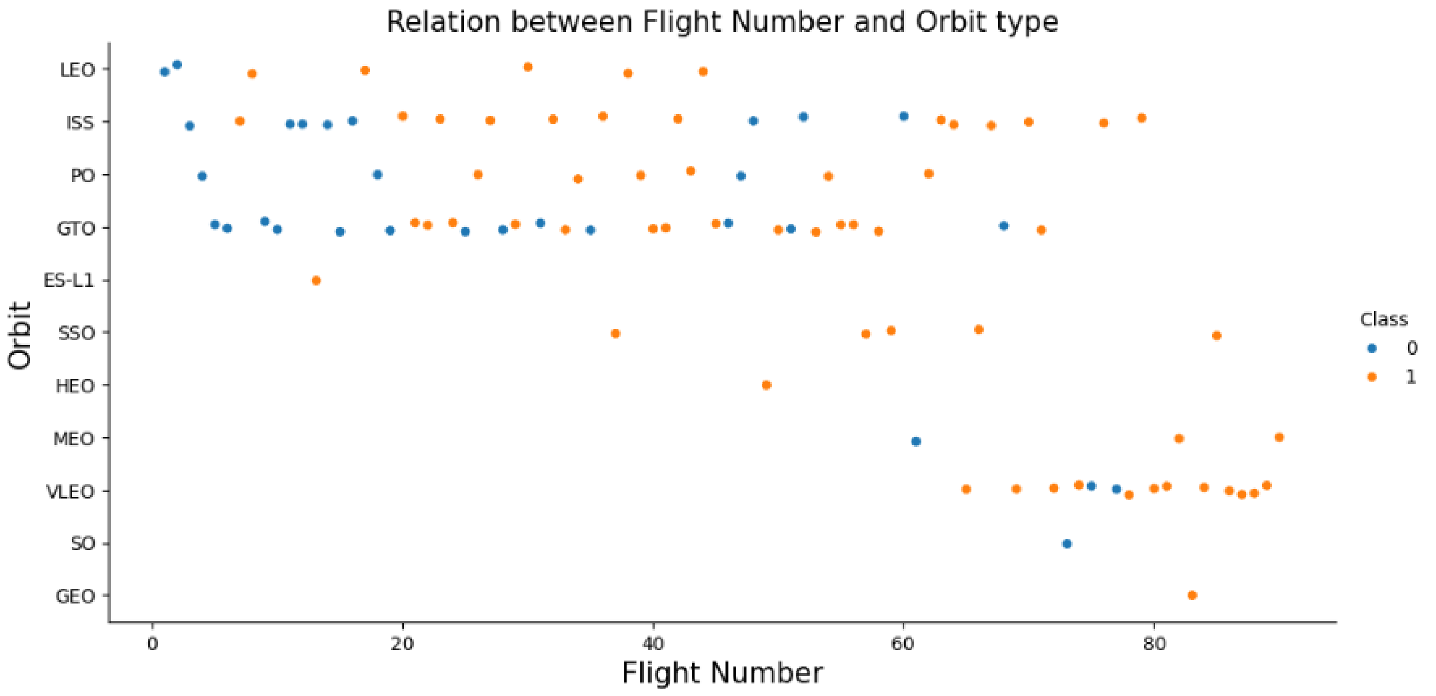
Success Rate vs. Orbit Type

- There are four Orbits with outstanding success rate
  - ES-L1
  - GEO
  - HEO
  - SSO



Flight Number vs. Orbit Type

- This plot shows that in recent years the number of lunches outstanding for the Orbit VLEO and shows a great success rate.

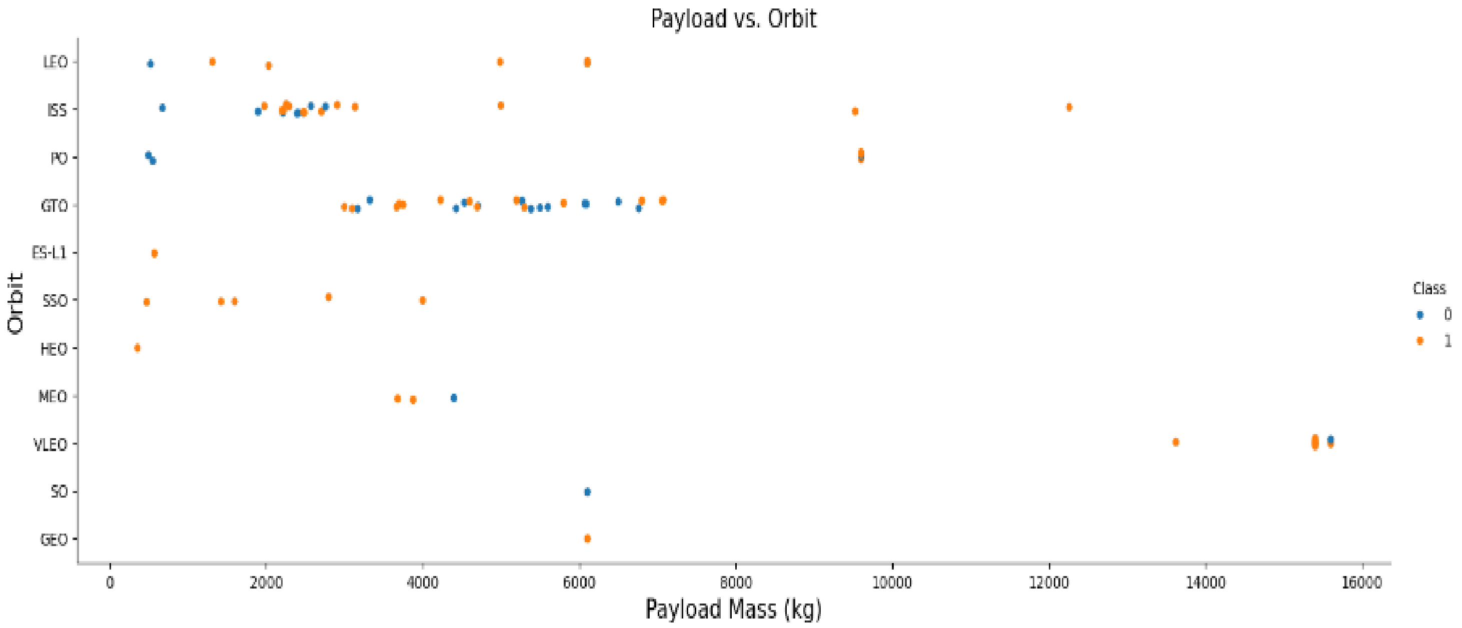


Payload vs. Orbit Type

This plot highlight great correlation at two orbits and payload.

- ISS from 2000 to 4000 payload mass

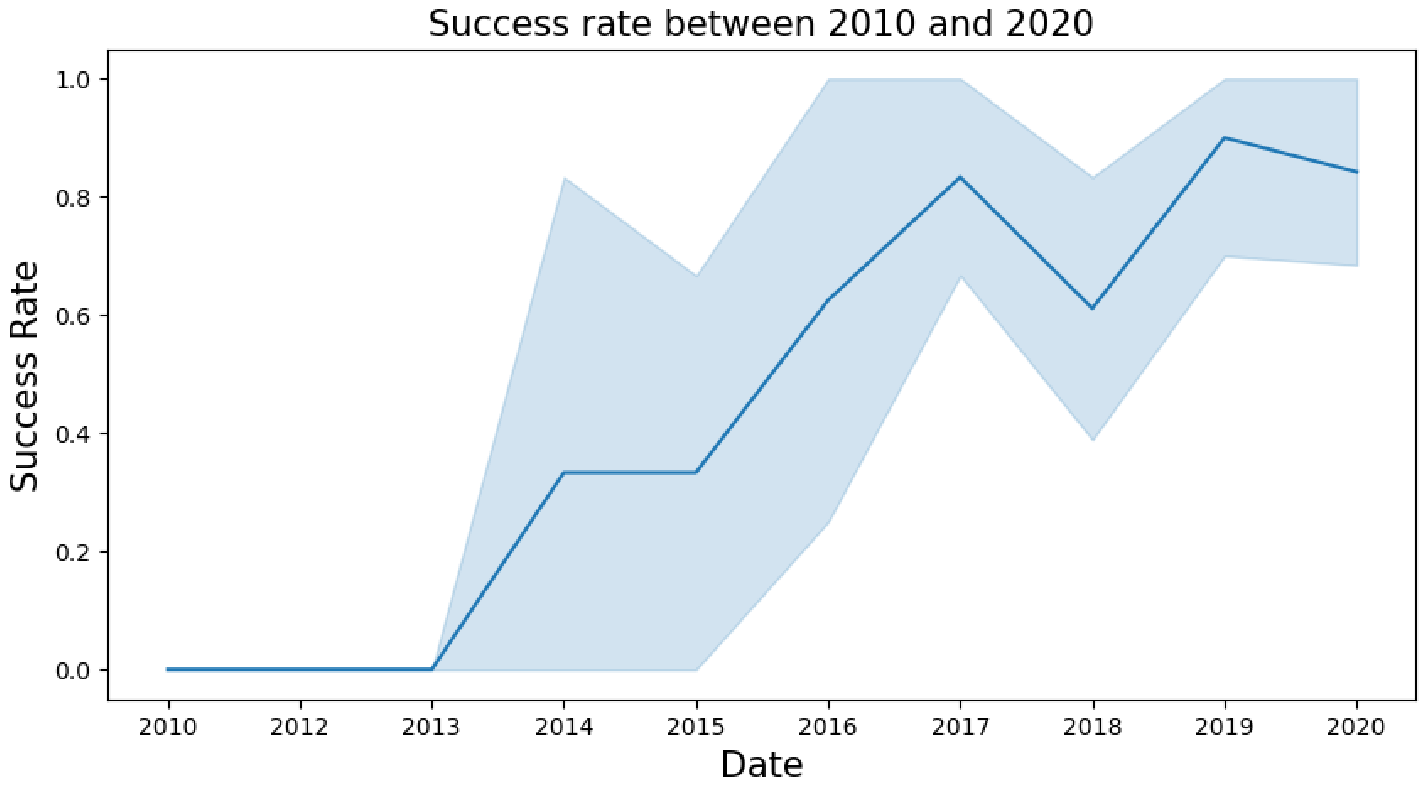
- GTO from 3500 to 7500 payload mass



Launch Success Yearly Trend

Launch success rate showing a great increase since 2013.

We can see a fallback in 2018 then from 2019 increased and stabilized, that could happen because the advanced technology and the gained experience from the previous years.



All Launch Site Names

- To find unique launch site names a SQL code have been used. (`%sql select distinct(LAUNCH_SITE) from SPACEXTABLE;`)

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

Launch Site Names Begin with 'CCA'

- Here we performed a SQL code to find data about launch site where the site name starting with CCA

- `%sql select * from SPACEXTABLE where (LAUNCH_SITE) LIKE '%CCA%' LIMIT 5;`

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- **%sql** select Customer, sum(PAYLOAD_MASS__KG_) as Total_Payload_Mass from SPACEXTABLE where Customer = 'NASA (CRS)'

Customer	Total_Payload_Mass
NASA (CRS)	45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- `%sql select Booster_Version, avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version = 'F9 v1.1';`

Booster_Version	avg(PAYLOAD_MASS__KG_)
F9 v1.1	2928.4

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad
- `%sql select min(Date) as First_Succesful_Landing, Landing_Outcome, Launch_Site from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)`

First_Succesful_Landing	Landing_Outcome	Launch_Site
2015-12-22	Success (ground pad)	CCAFS LC-40

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- **%sql** select Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTABLE where Landing_Outcome='Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000;

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes
- **%sql** select Mission_Outcome,count(Mission_Outcome) as Total_Mission_Outcome from SPACEXTABLE group by Mission_Outcome;

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

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass
- **%sql** select Booster_Version,PAYLOAD_MASS__KG_ from SPACEXTABLE where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTABLE);

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

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

- ```
  %sql select substr(Date, 6,2) as Months, Mission_Outcome, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE where Landing_Outcome like 'Failure%' and subs
  ```

Months	Mission_Outcome	Landing_Outcome	Booster_Version	Launch_Site
04	Success	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
10	Success	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
  ```

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

- Rank 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
- **%sql** select Landing_Outcome, count(Landing_Outcome) as Total_Landings from SPACEXTABLE where Date between '2010-06-04' AND '2017-03-20' group by Landing_Outcome order by Count(Landing_Outcome) DESC;

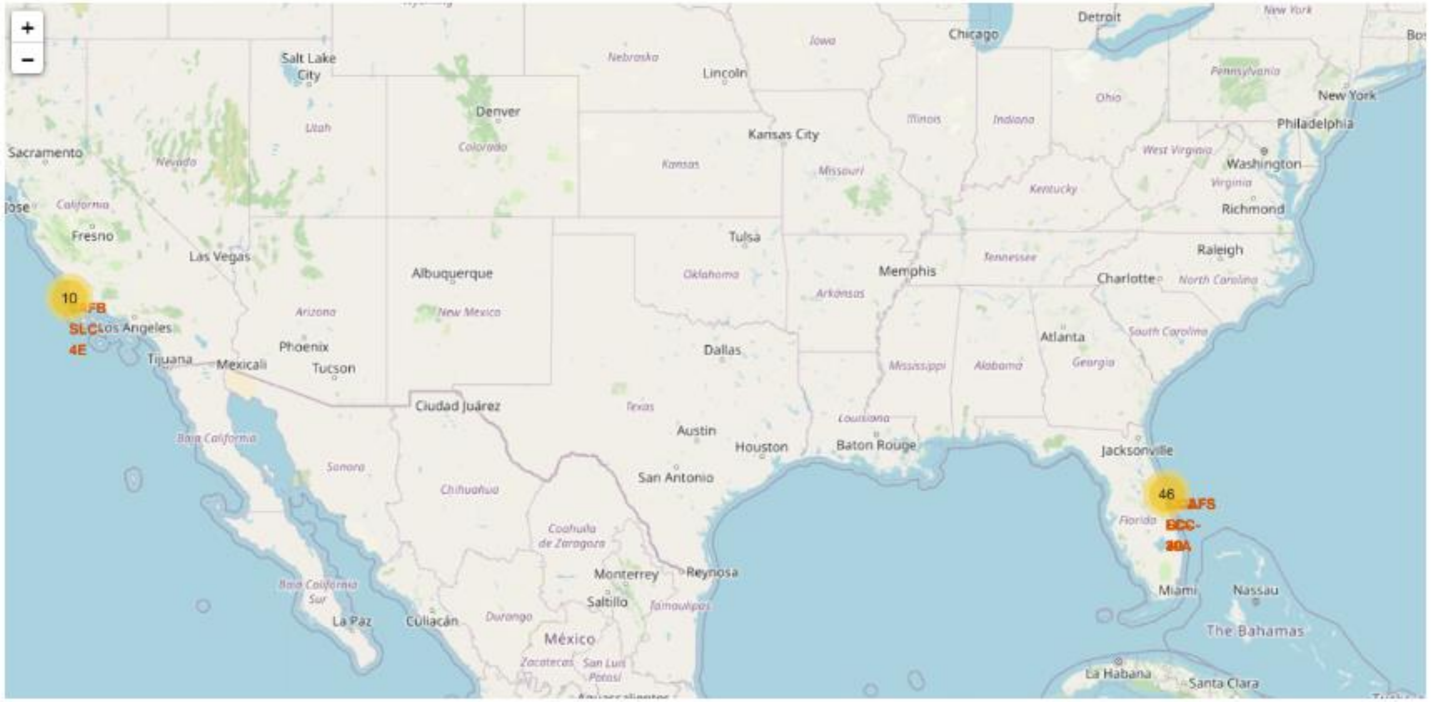
Landing_Outcome	Total_Landings
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

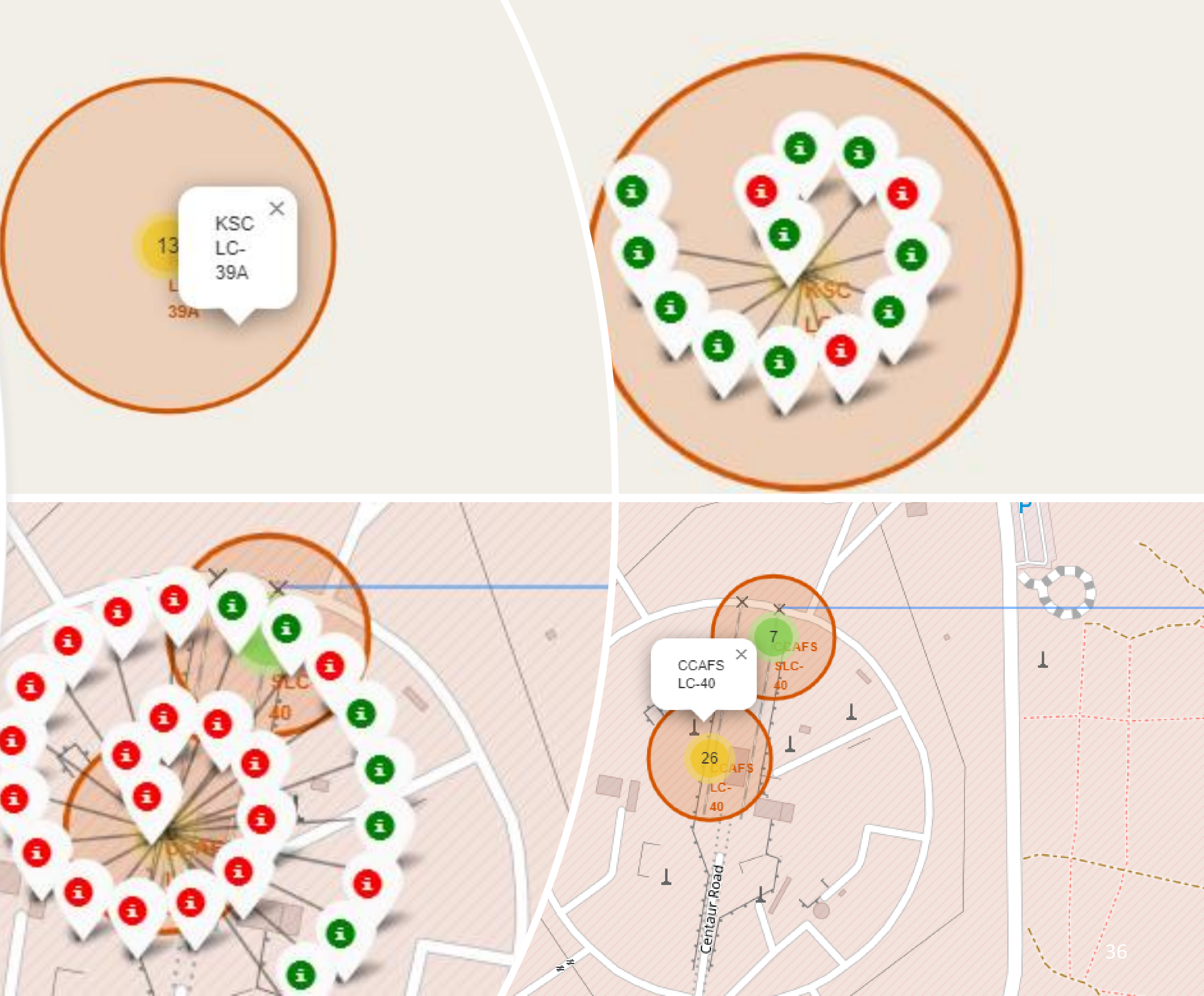
Launch Sites

The interactive map shows clearly, that all SpaceX launch sites are in United States we can find them in California and Florida.



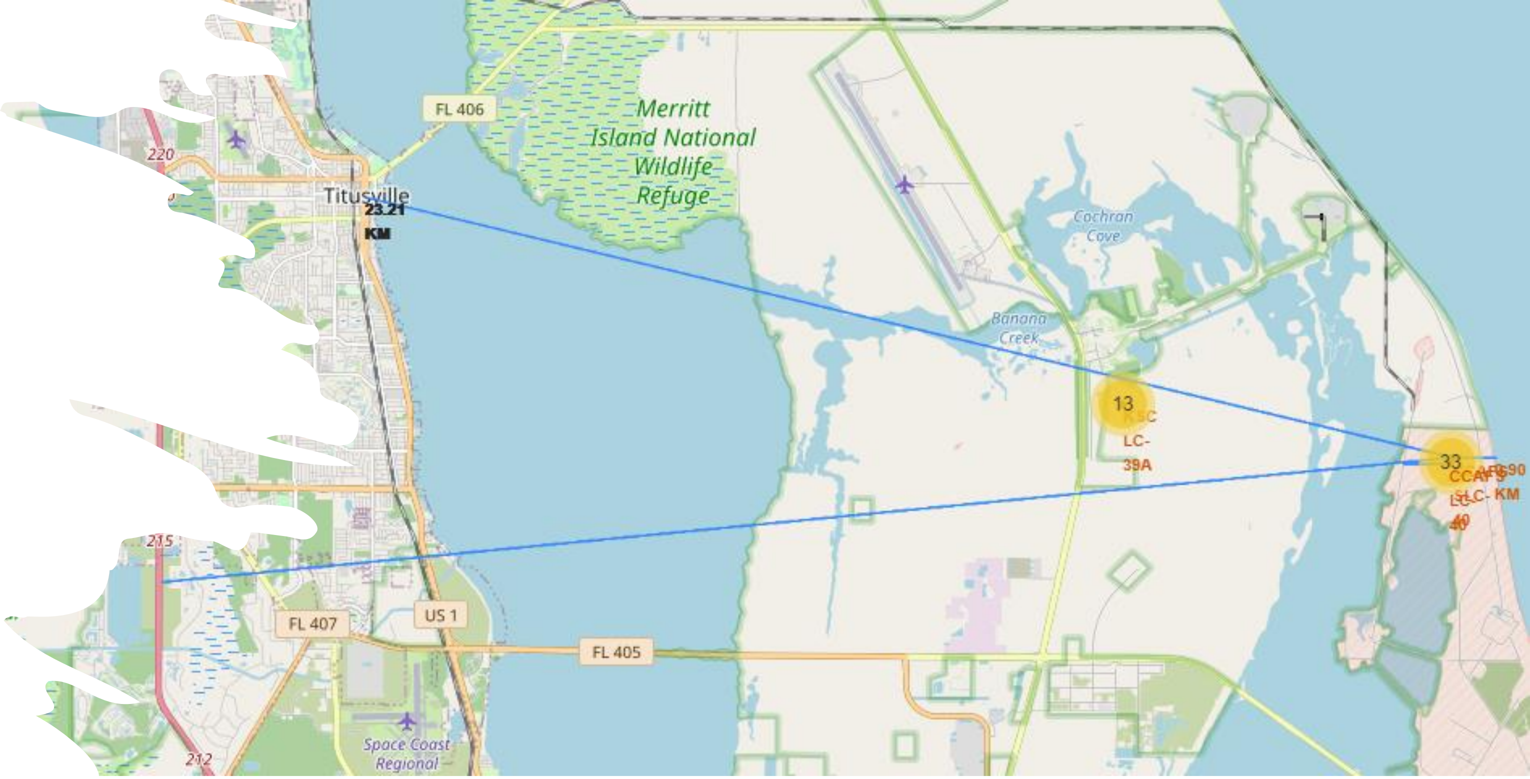
Florida Launch Sites

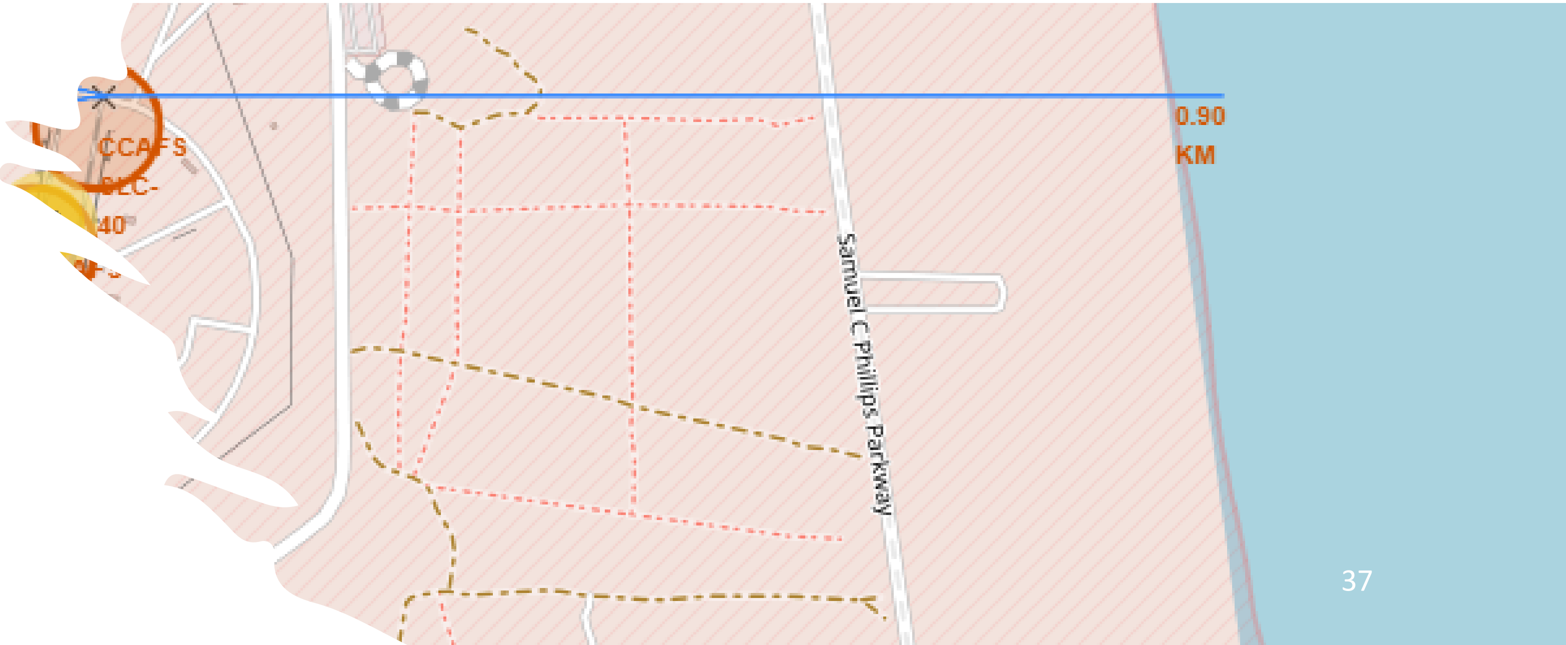
- In Florida there are two launch sites CCAFS LC-40 and KSC LC-39A
- The **Green** icons shows the successful launches
- The **Red** icons shows the failed launches



Florida Launch Site and Distances

- A couple of line added to the map to show some significant distance from launch site and closest city , highway, railway.
- It is clearly shows they are not close there is a safe distance from cities
- The only close point is the coastline



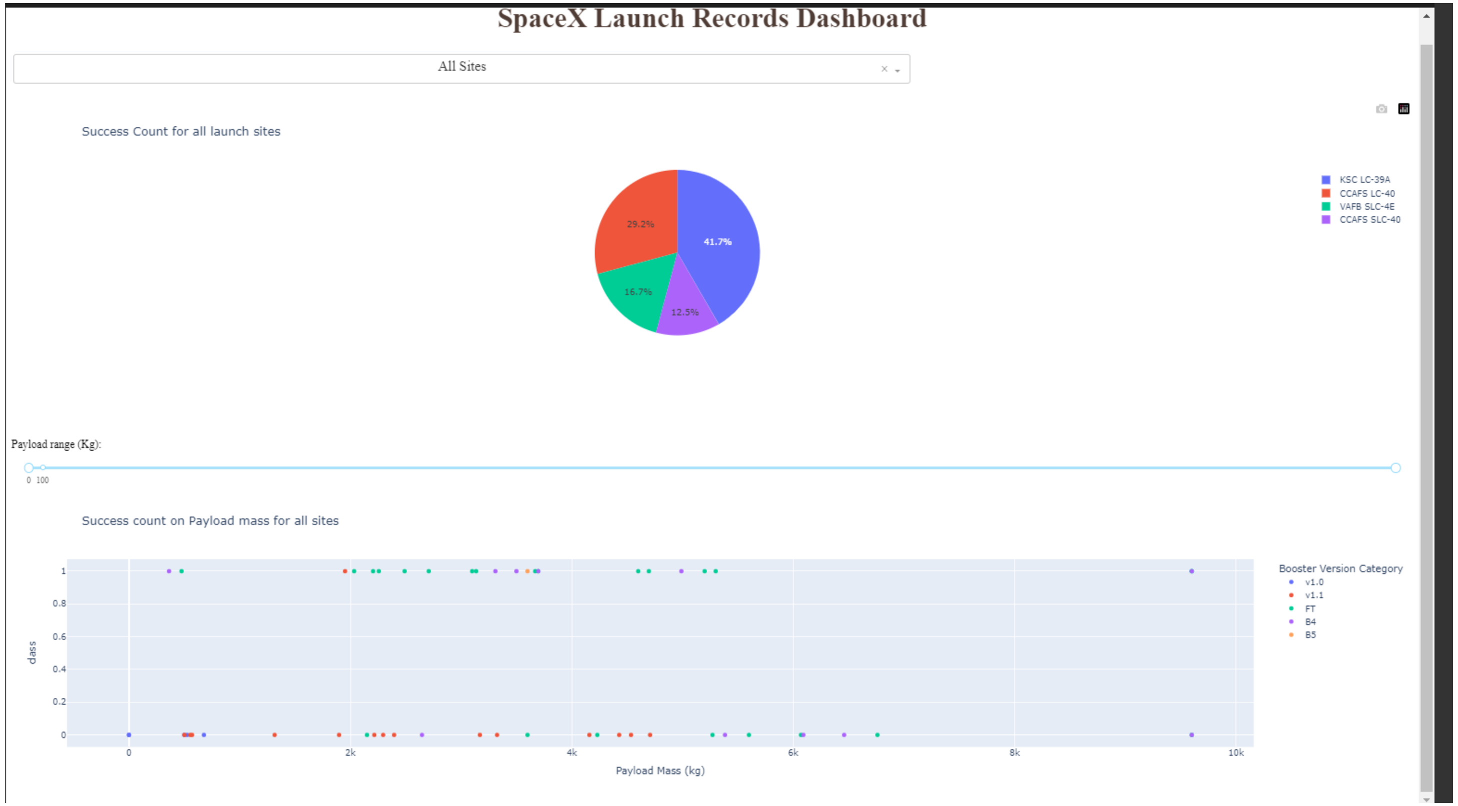


Section 4

Build a Dashboard with Plotly Dash

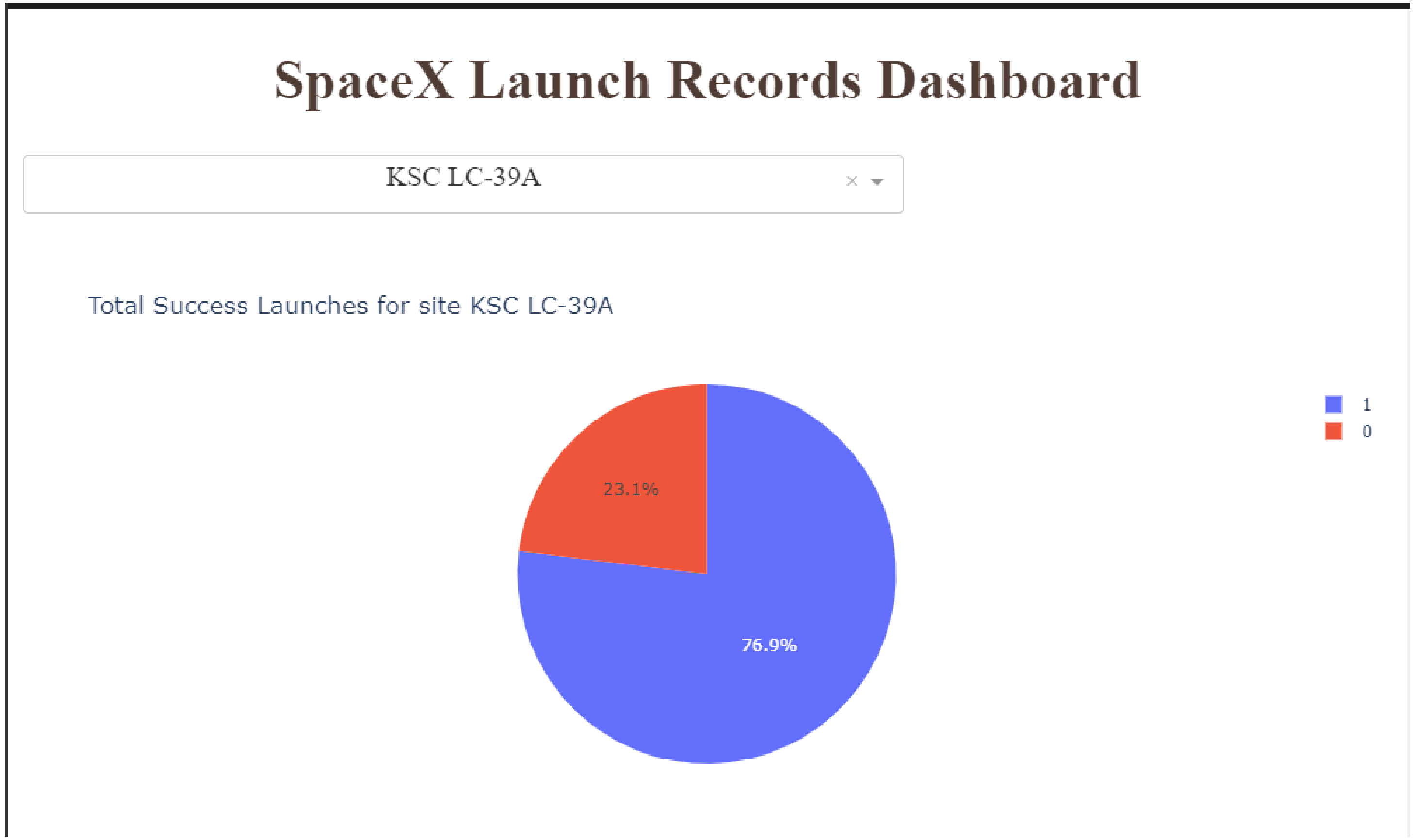
Total successful launches by all the sites

- This interactive Dashboard shows all launches and their success rate
- The bottom chart shows how the payload correlate with the successful launches



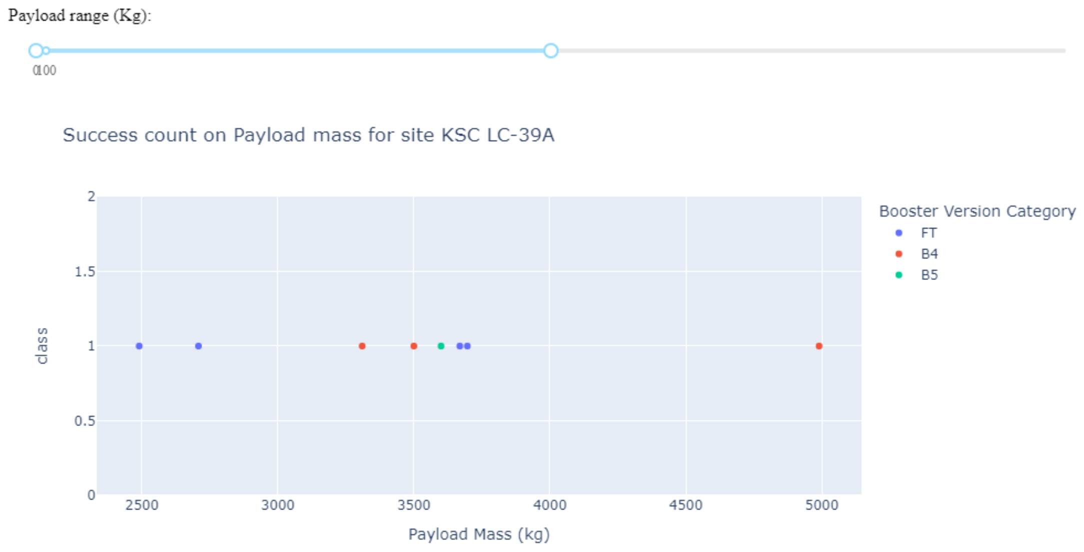
Success rate by site

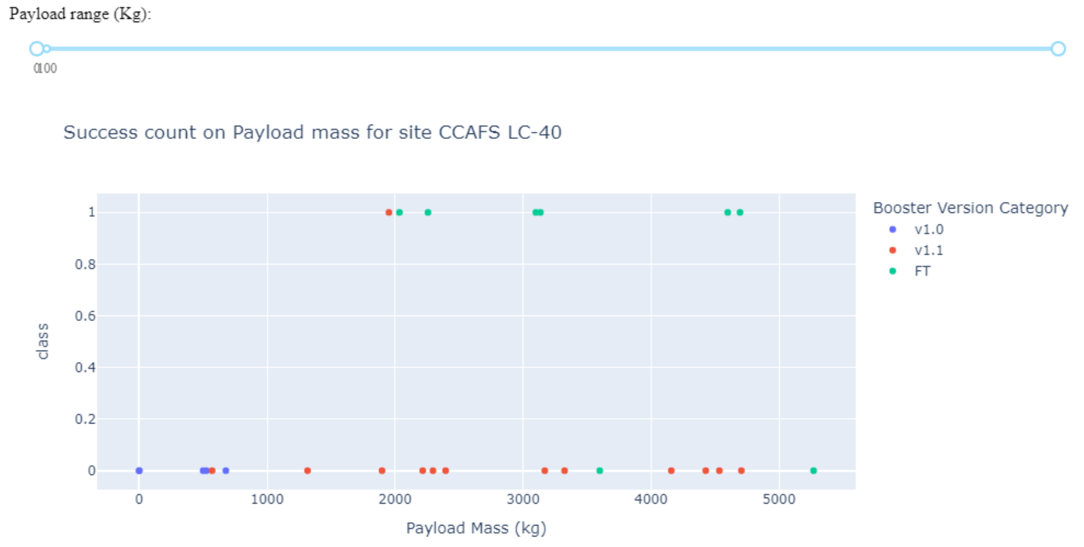
- At the dropdown menu we can choose different sites
- On this chart we can see launch site KSC LC-39A and the success rate



Scaling and visualizing the the payload correlation with success rate

- Here we can see two different launch site and the booster versions
- Clearly shows how the payload mass is correlate with the success rate
- On this interactive dashboard we can choose different scale of payload and it will highlight the successful launch and their booster version



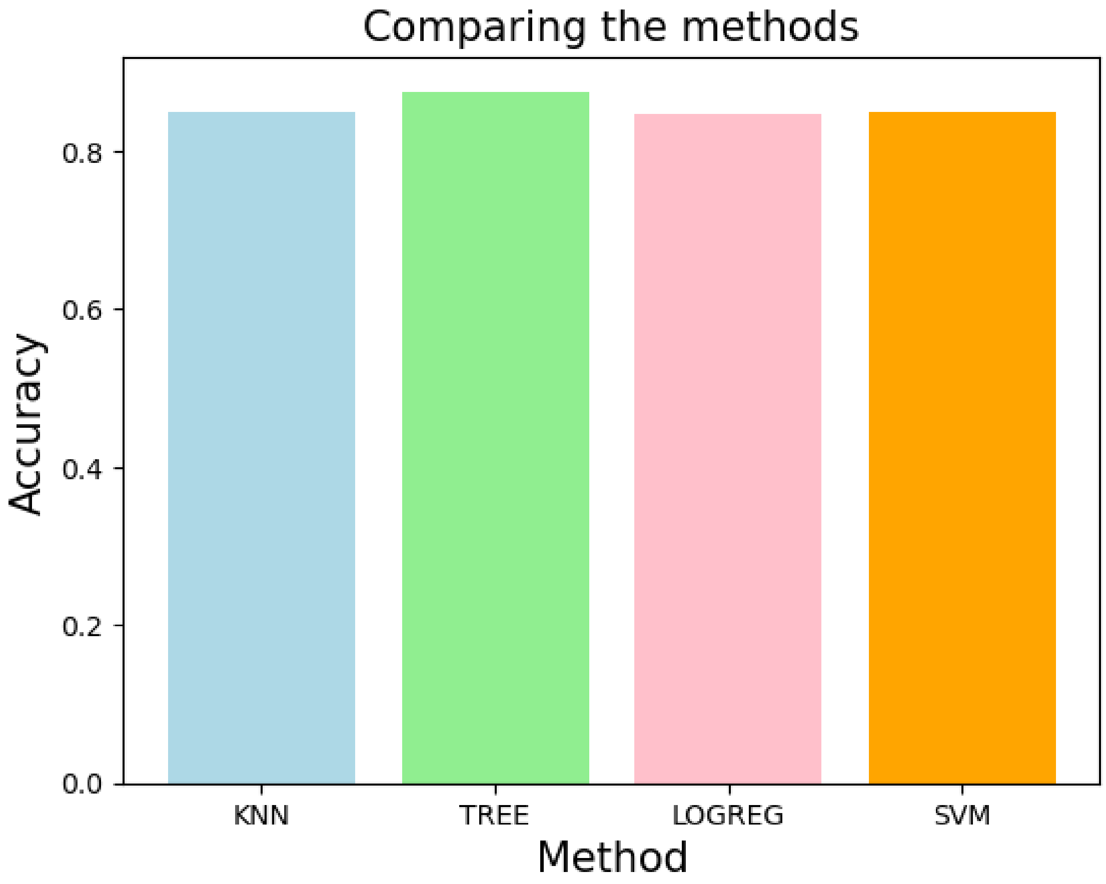


Section 5

Predictive Analysis (Classification)

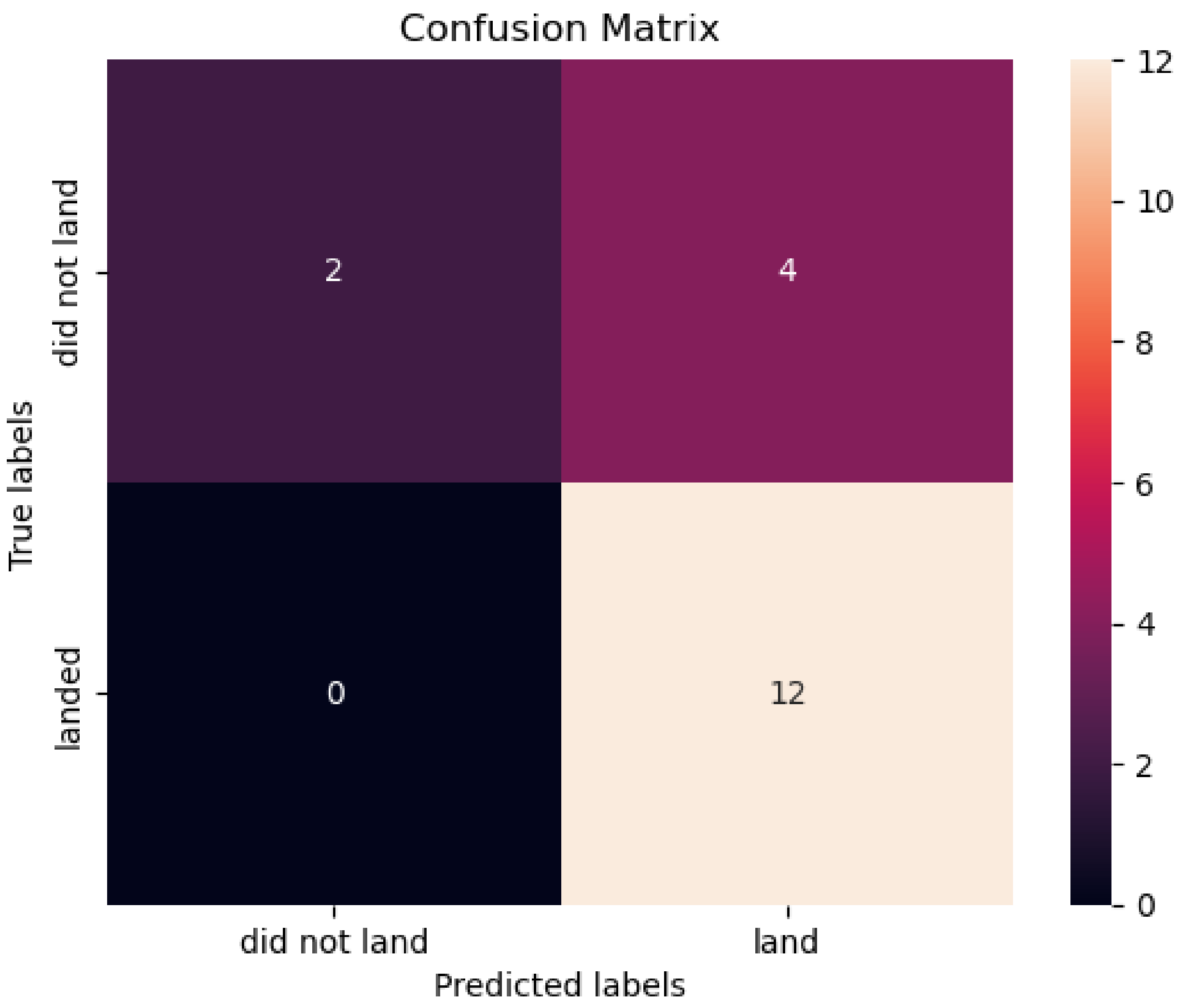
Classification Accuracy

- Comparing models to find the best accuracy
- On this bar chart we can see the Decision Tree has the highest score in accuracy



Confusion Matrix

- The Confusion Matrix for the Decision Tree
- The only problem that there are some errors in the Matrix where some occasions shows as landed but actually did not land



Conclusions

- The larger number of flight at a launch site, greater the success rate.
- Launch success rate started to increase in 2013 and stabilised in 2020.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the highest success rate.
- KSC LC-39A is the most successful launch site among the other sites.
- The Decision tree classifier is the best machine learning method for this task.

Thank you!