



IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Salman Almalki
2023-03-06



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 a Dashboard with Plotly Dash
 - Machine Learning Prediction
- Summary of all results
 - Exploratory data analysis results
 - Interactive analytics demo in screenshots
 - Predictive analysis results

Introduction

- Project background and context

Space X 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 Space X 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 space X for a rocket launch. The goal of the project is to build a machine learning pipeline to predict whether the first stage will successfully land.

- Problems you want to find answers

- What influences if the rocket will land successfully?
- Where is the best place to make launches?
- What conditions does SpaceX have to achieve to get the best results and ensure the best rocket success landing rate?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

Data of Space X was collected from 2 sources:

- Space X API
- WebScraping from Wikipedia

- Perform data wrangling

- One Hot Encoding data fields for Machine Learning and dropping irrelevant columns

- Perform exploratory data analysis (EDA) using visualization and SQL

- Perform interactive visual analytics using Folium and Plotly Dash

- Perform predictive analysis using classification models

- How to build, tune, evaluate classification models

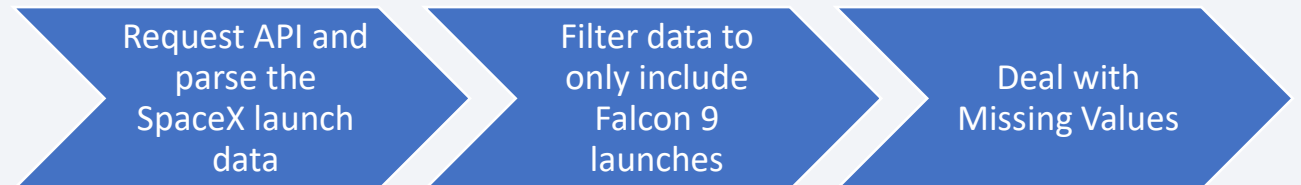
Data Collection

The following datasets was collected by

- worked with SpaceX launch data that is gathered from the SpaceX REST API.
- This API will give us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.
- Our goal is to use this data to predict whether SpaceX will attempt to land a rocket or not.
- The SpaceX REST API endpoints, or URL, starts with `api.spacexdata.com/v4/`.
- Another popular data source for obtaining Falcon 9 Launch data is web scraping Wikipedia using BeautifulSoup.

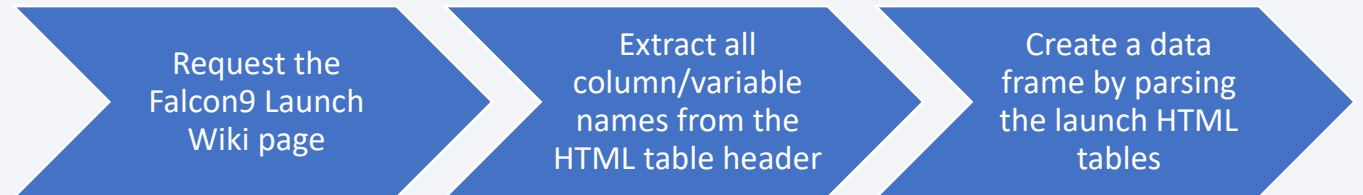
Data Collection – SpaceX API

- SpaceX offers a public API from where data can be obtained and then used;
- This API was used according to the flowchart beside and then data is persisted.
- GitHub URL of the completed SpaceX API calls notebook
<https://github.com/Salman-mlk77/Final-Project/blob/d1699560a01cc69d4ecdcb8e95b51f8d6c4e424d/Data%20Science%20Pro.ipynb>



Data Collection - Scraping

- Data from SpaceX launches can also be obtained from Wikipedia
- Data are downloaded from Wikipedia according to the flowchart and then persisted.
- GitHub URL of the completed web scraping notebook
<https://github.com/Salman-mlk77/Final-Project/blob/d1699560a01cc69d4ecdcb8e95b51f8d6c4e424d/Data%20Science%20Project2.ipynb>



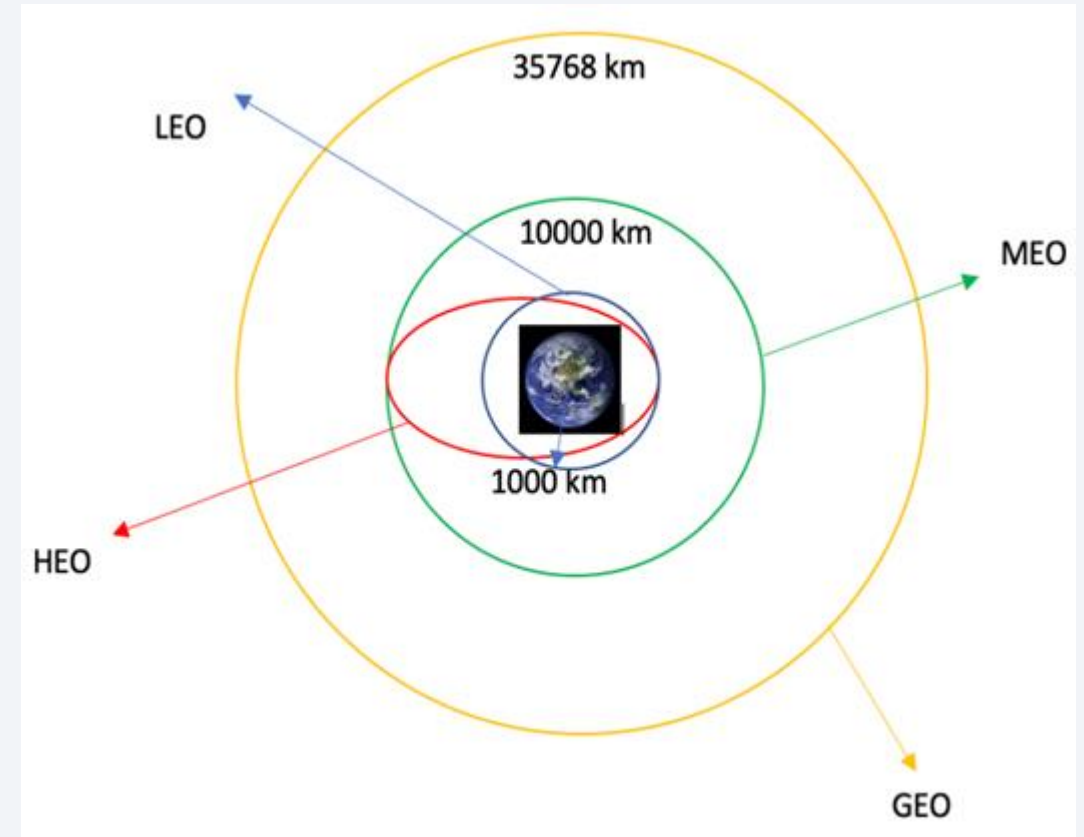
Data Wrangling

We performed exploratory data analysis and determined the training labels.

We calculated the number of launches at each site, and the number and occurrence of each orbits

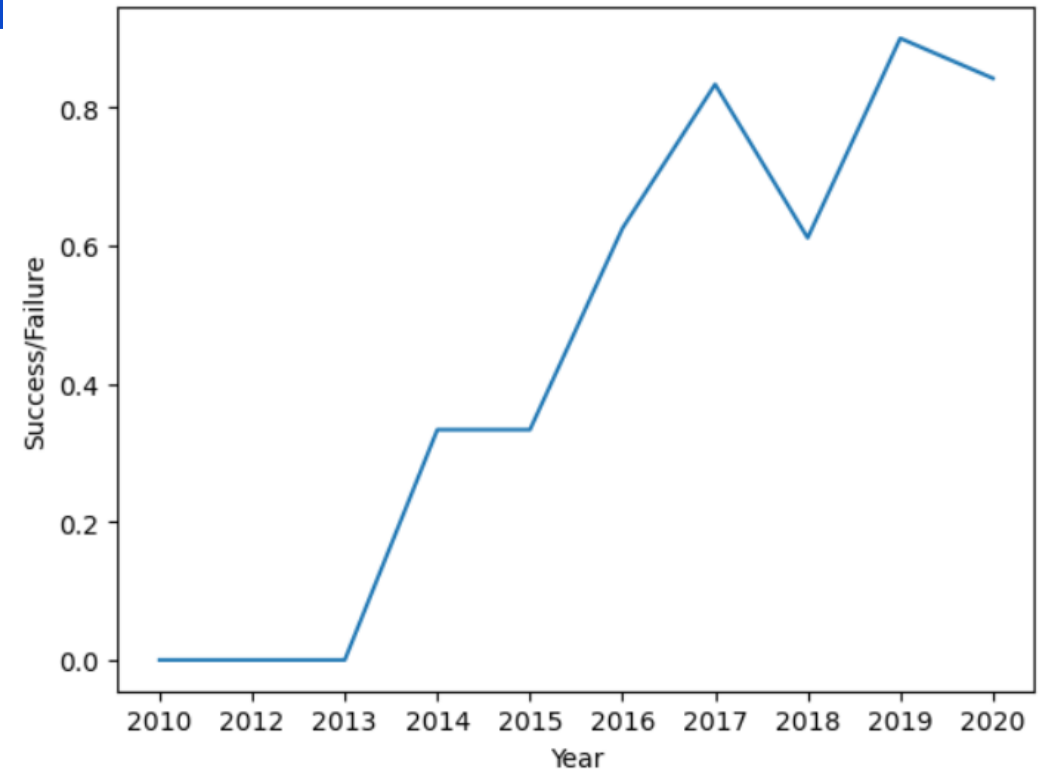
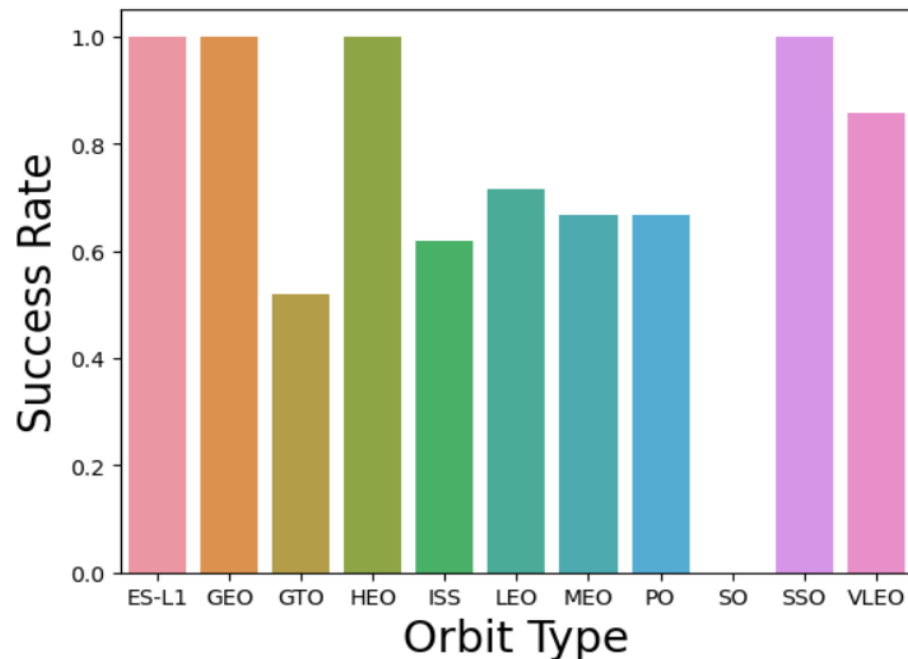
We created landing outcome label from outcome column and exported the results to csv.

GitHub URL of the completed data wrangling notebook
<https://github.com/Salman-mlk77/Final-Project/blob/d1699560a01cc69d4ecdcb8e95b51f8d6c4e424d/Data%20Science%20Pro3.ipynb>



EDA with Data Visualization

- We explored the data by visualizing the relationship between flight number and launch Site, payload and launch site, success rate of each orbit type, flight number and orbit type, the launch success yearly trend.



- GitHub URL of the completed EDA with Data Visualization notebook
<https://github.com/Salman-mlk77/Final-Project/blob/d1699560a01cc69d4ecdcb8e95b51f8d6c4e424d/Data%20Science%20Pro5.ipynb>

EDA with SQL

The following SQL queries were performed:

- Names of the unique launch sites in the space mission;
- Top 5 launch sites whose name begin with the string 'CCA';
- Total payload mass carried by boosters launched by NASA (CRS);
- Average payload mass carried by booster version F9 v1.1;
- Date when the first successful landing outcome in ground pad was achieved;
- Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg;
- Total number of successful and failure mission outcomes;
- Names of the booster versions which have carried the maximum payload mass;
- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015;
- Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20 GitHub URL of your completed EDA with SQL notebook

GitHub URL of the completed EDA with SQL notebook <https://github.com/Salman-mlk77/Final-Project/blob/d1699560a01cc69d4ecdcb8e95b51f8d6c4e424d/Data%20Science%20pro4.ipynb>

Build an Interactive Map with Folium

- We marked all launch sites, and added map objects such as markers, circles, lines to mark the success or failure of launches for each site on the folium map
- We assigned the dataframe launch_outcomes(failures, successes) to classes 0 and 1 with **Green** and **Red** markers on the map in a MarkerCluster()
- We calculated the distances between a launch site to its proximities
- GitHub URL of completed interactive map with Folium map
<https://github.com/Salman-mlk77/Final-Project/blob/d1699560a01cc69d4ecdcb8e95b51f8d6c4e424d/Data%20Science%20Pro6.ipynb>

Build a Dashboard with Plotly Dash

The following graphs and plots were used to visualize data

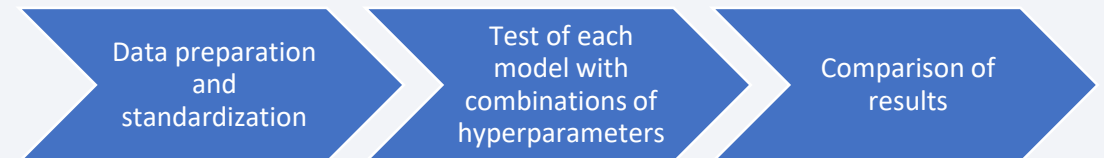
- Percentage of launches by site
- Payload range

This combination allowed to quickly analyze the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads.

GitHub URL of completed Plotly Dash lab <https://github.com/Salman-mlk77/Final-Project/blob/457e6febd39c609ac2de7d7736ed247978b204a0/Dash.py>

Predictive Analysis (Classification)

- We loaded the data using numpy and pandas, transformed the data, split our data into training and testing.
- We built different machine learning models and tune different hyperparameters using GridSearchCV.
- We used accuracy as the metric for our model, improved the model using feature engineering and algorithm tuning.
- We found the best performing classification model.



GitHub URL of completed predictive analysis lab

https://github.com/Salman-mlk77/Final-Project/blob/457e6febd39c609ac2de7d7736ed247978b204a0/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

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

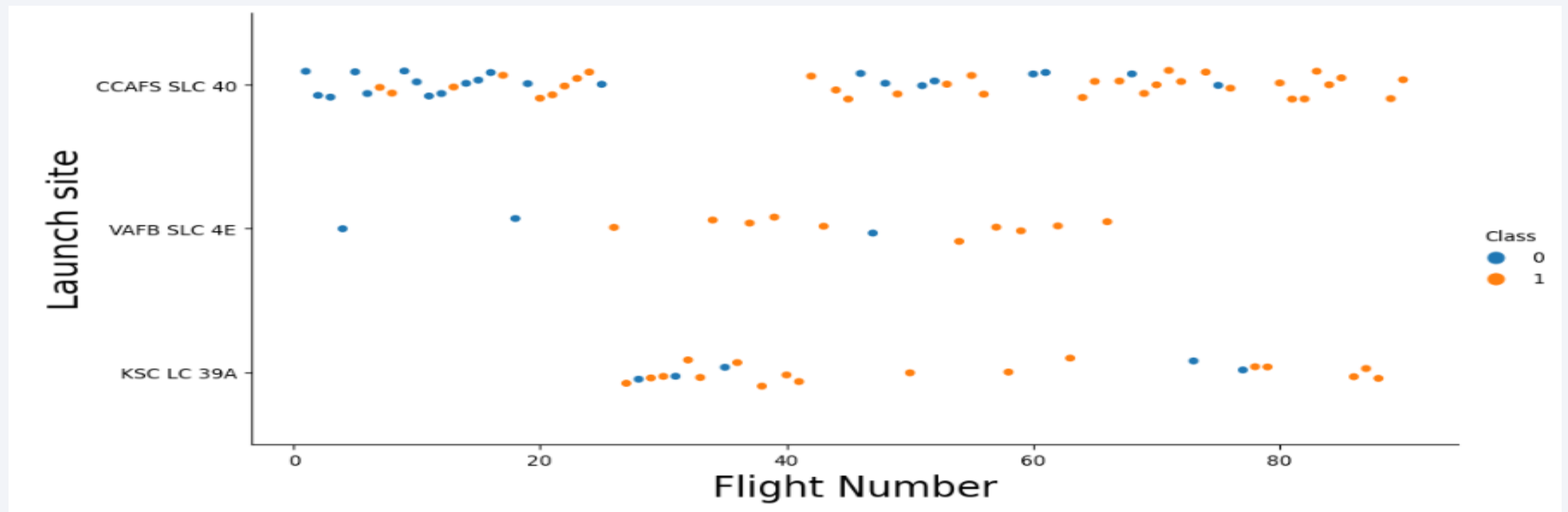
The background of the slide is an abstract composition. It features a dark blue base color. Overlaid on this are numerous diagonal streaks in shades of red and cyan. A faint, light blue grid pattern is also visible, particularly in the lower half of the image. The overall effect is dynamic and technological.

Section 2

Insights drawn from EDA

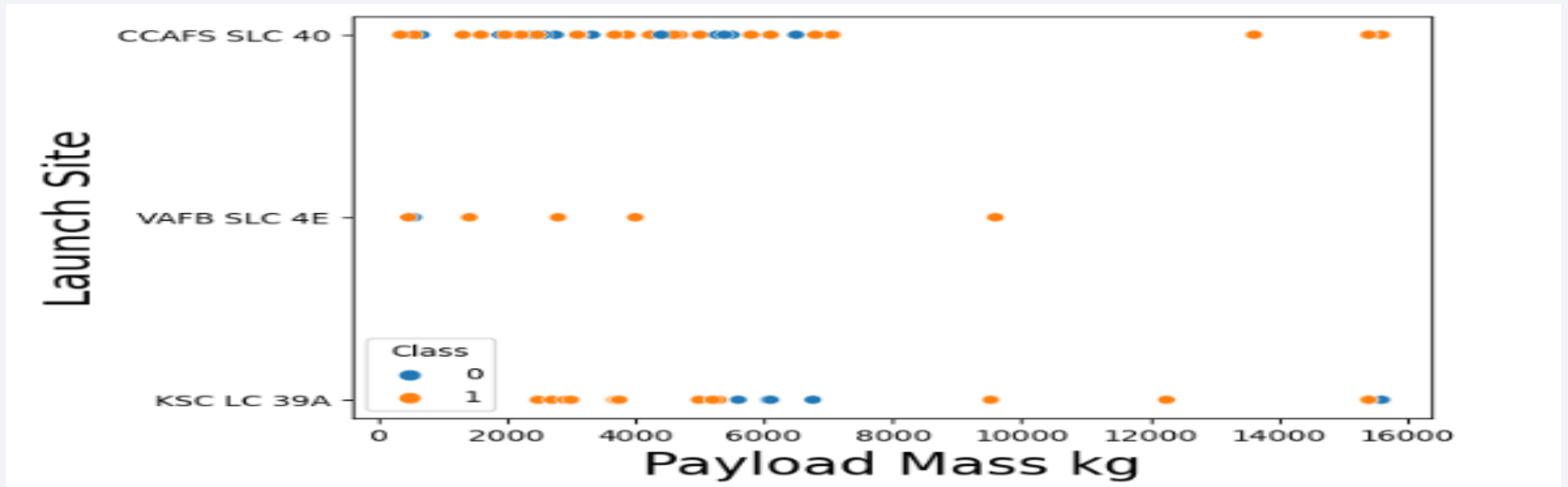
Flight Number vs. Launch Site

- From the plot, we found that the larger the flight amount at a launch site, the greater the success rate at a launch site.



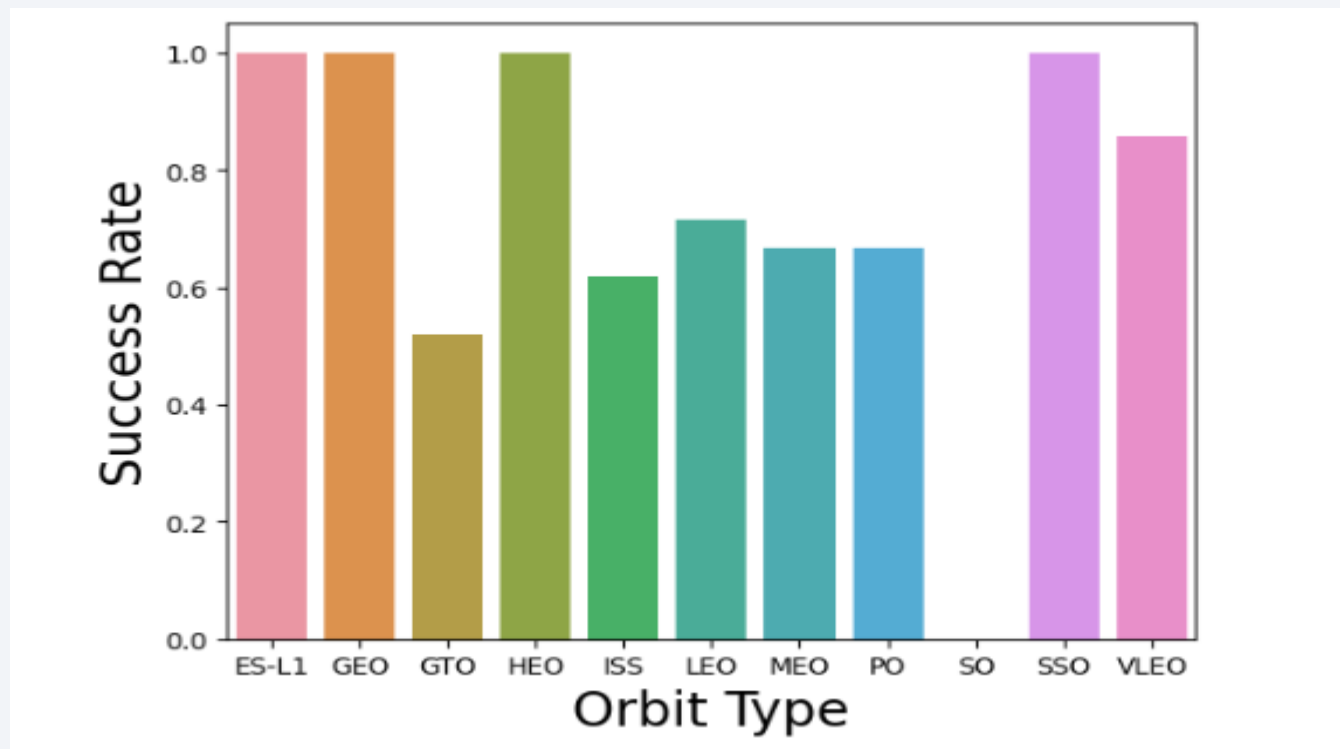
Payload vs. Launch Site

- Payloads over 9,000kg (about the weight of a school bus) have excellent success rate;
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.



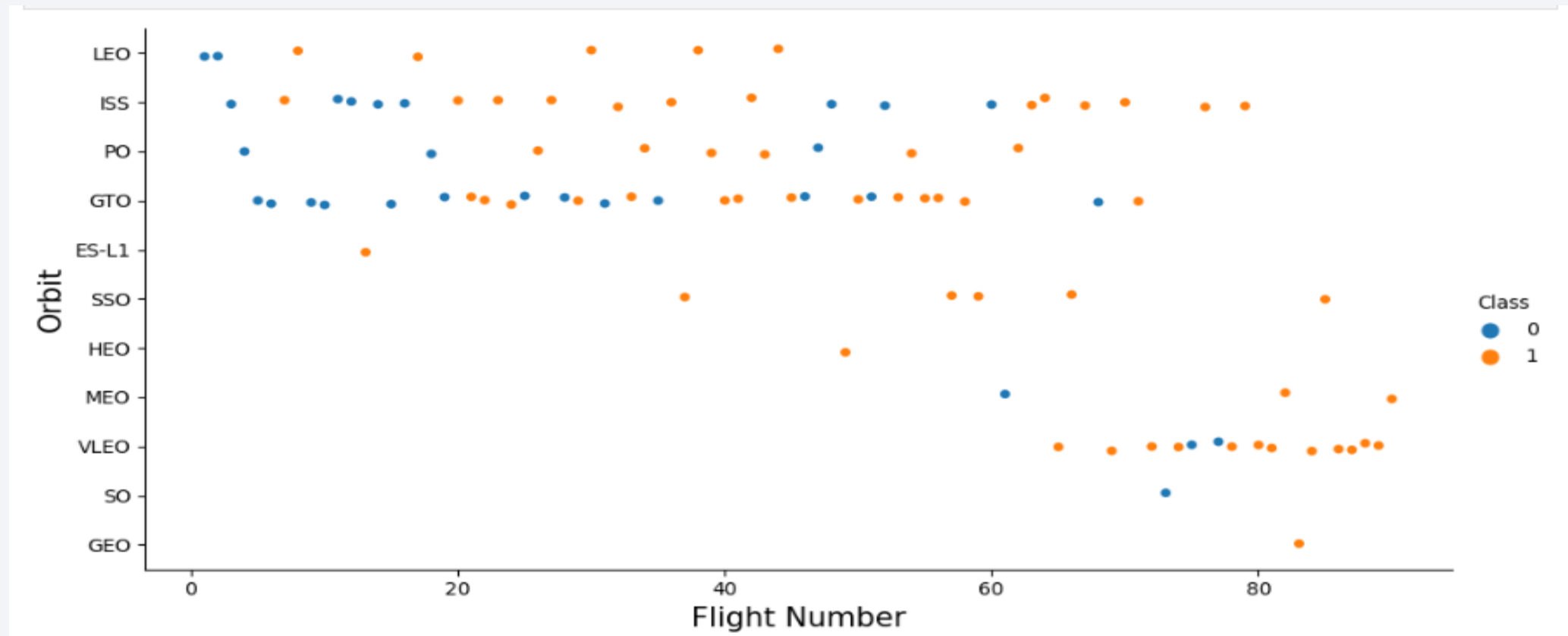
Success Rate vs. Orbit Type

- From the plot, we can see that ES-L1, GEO, HEO, SSO, VLEO had the most success rate.



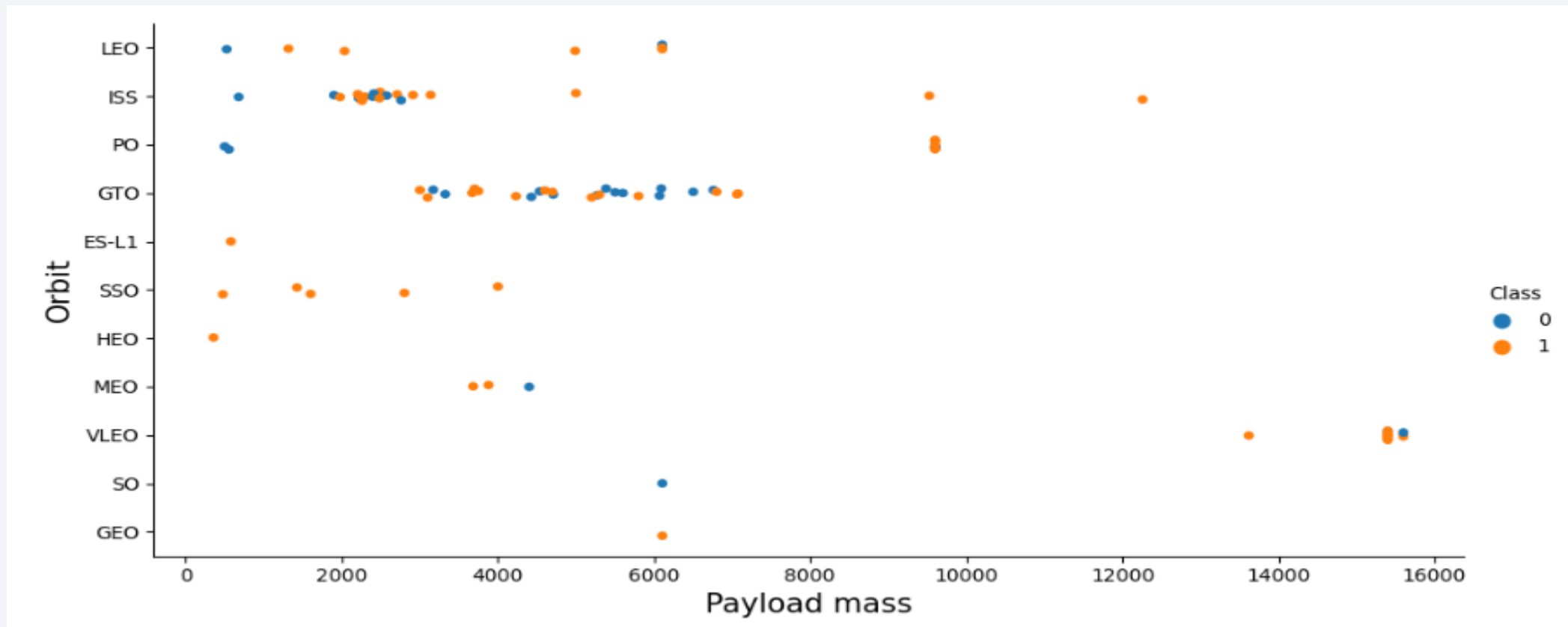
Flight Number vs. Orbit Type

- You should see that 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 when in GTO orbit.



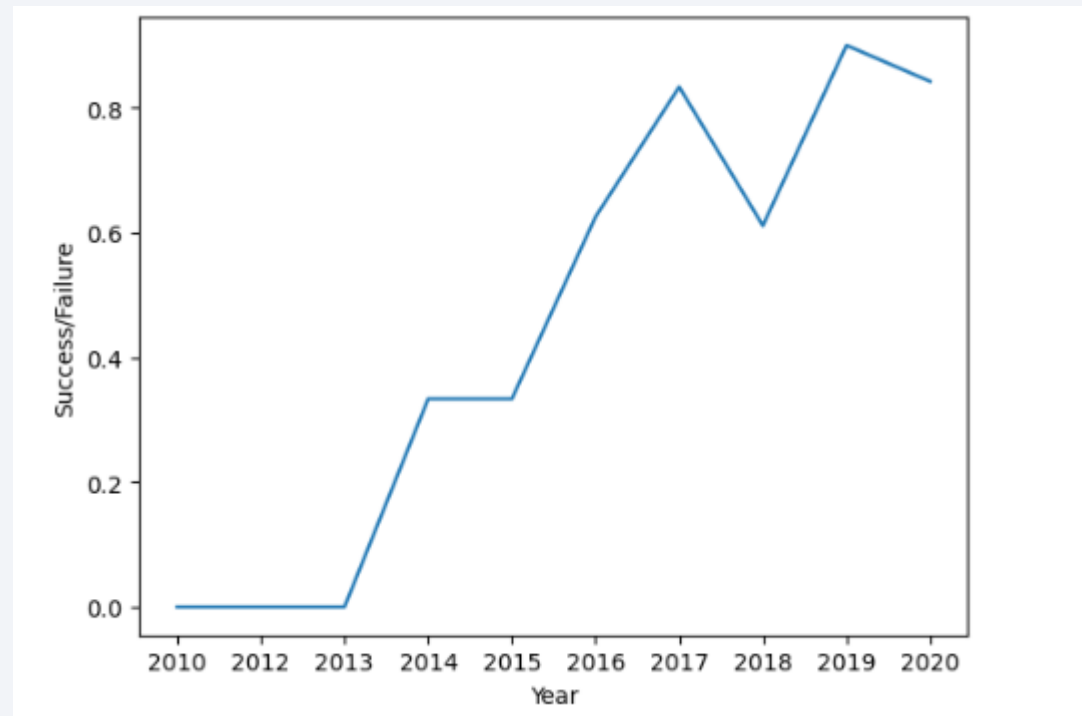
Payload vs. Orbit Type

- We can observe that with heavy payloads, the successful landing are more for PO, LEO and ISS orbits.



Launch Success Yearly Trend

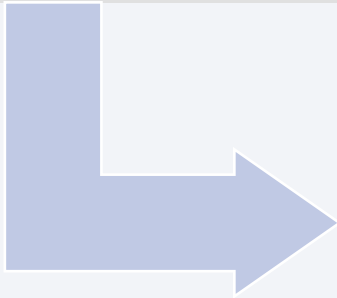
- You can observe that the success rate since 2013 kept increasing till 2020



All Launch Site Names

- According to data, there are four launch sites:
- They are obtained by selecting unique occurrences of “launch_site” values from the dataset.

```
%sql select Unique(LAUNCH_SITE) from SPACEXTBL;
```

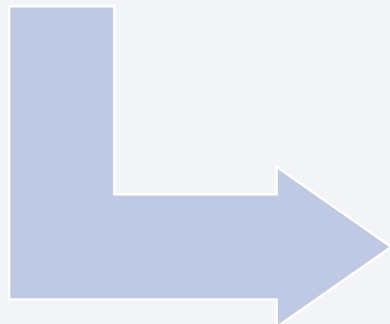


launch_site
CCAFS LC-40
CCAFS SLC-40
CCAFSSLC-40
KSC LC-39A
VAFB SLC-4E

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with 'CCA':

```
%sql SELECT LAUNCH_SITE from SPACEXTBL where (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5;
```

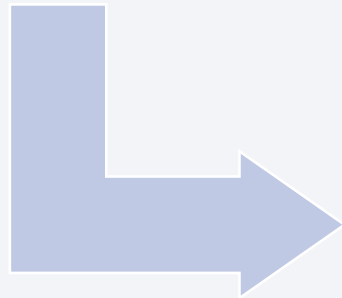


launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

- Total payload carried by boosters from NASA:

```
%sql select sum(PAYLOAD_MASS_KG_) as payloadmass from SPACEXTBL;
```



payloadmass

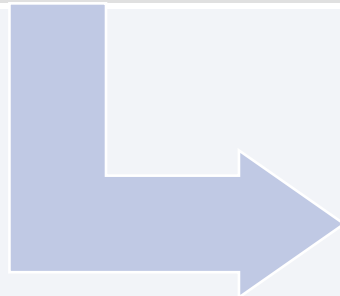
619967

- Total payload calculated above, by summing all payloads whose codes contain 'CRS', which corresponds to NASA.

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1:

```
%sql select avg(PAYLOAD_MASS_KG_) as payloadmass from SPACEXTBL;
```



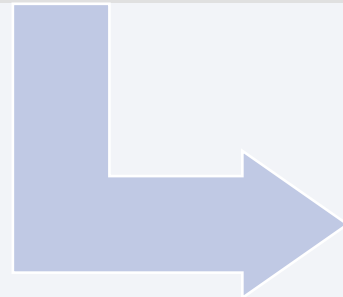
payloadmass
6138

- Filtering data by the booster version above and calculating the average payload mass we obtained the value of 6138.

First Successful Ground Landing Date

- First successful landing outcome on ground pad

```
%sql select min(DATE) from SPACEXTBL;
```



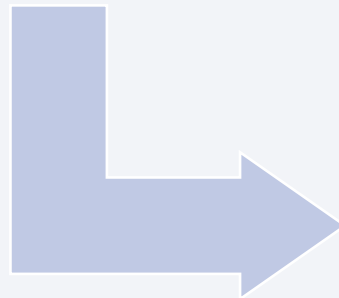
1
2010-06-04

- By filtering data by successful landing outcome on ground pad and getting the minimum value for date it's possible to identify the first occurrence, that happened on 2010-06-04

Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
%sql select BOOSTER_VERSION from SPACEXTBL where LANDING__OUTCOME='Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000;
```



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

Total Number of Successful and Failure Mission Outcomes

- Number of successful and failure mission outcomes

```
%sql select count(MISSION_OUTCOME) as missionoutcomes from SPACEXTBL GROUP BY MISSION_OUTCOME;
```

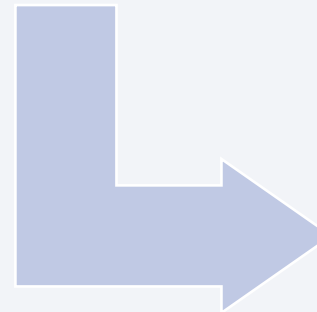


missionoutcomes
1
99
1

Boosters Carried Maximum Payload

- Boosters which have carried the maximum payload mass

```
%sql select BOOSTER_VERSION as boosterversion from SPACEXTBL where PAYLOAD_MASS_KG_=(select max(PAYLOAD_MASS_KG_) from SPACEXTBL);
```



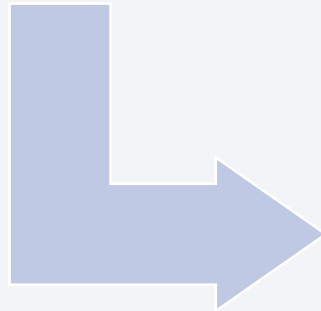
boosterversion
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

- These are the boosters which have carried the maximum payload mass registered in the dataset.

2015 Launch Records

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

```
%sql SELECT MONTH(DATE),MISSION_OUTCOME,BOOSTER_VERSION,LAUNCH_SITE FROM SPACEXTBL where EXTRACT(YEAR FROM DATE)='2015';
```

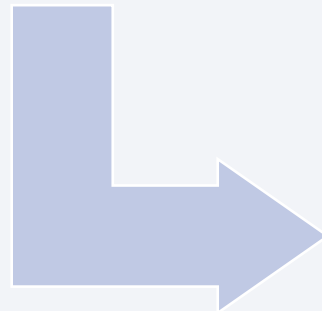


1	mission_outcome	booster_version	launch_site
1	Success	F9 v1.1 B1012	CCAFS LC-40
2	Success	F9 v1.1 B1013	CCAFS LC-40
3	Success	F9 v1.1 B1014	CCAFS LC-40
4	Success	F9 v1.1 B1015	CCAFS LC-40
4	Success	F9 v1.1 B1016	CCAFS LC-40
6	Failure (in flight)	F9 v1.1 B1018	CCAFS LC-40
12	Success	F9 FT B1019	CCAFS LC-40

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

- Ranking of all landing outcomes between the date 2010-06-04 and 2017-03-20:

```
%sql SELECT LANDING__OUTCOME FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' ORDER BY DATE DESC;
```



	landingoutcome	count
0	No attempt	10
1	Success (drone ship)	6
2	Failure (drone ship)	5
3	Success (ground pad)	5
4	Controlled (ocean)	3
5	Uncontrolled (ocean)	2
6	Precluded (drone ship)	1
7	Failure (parachute)	1

- This view of data alerts us that “No attempt” must be taken in account.

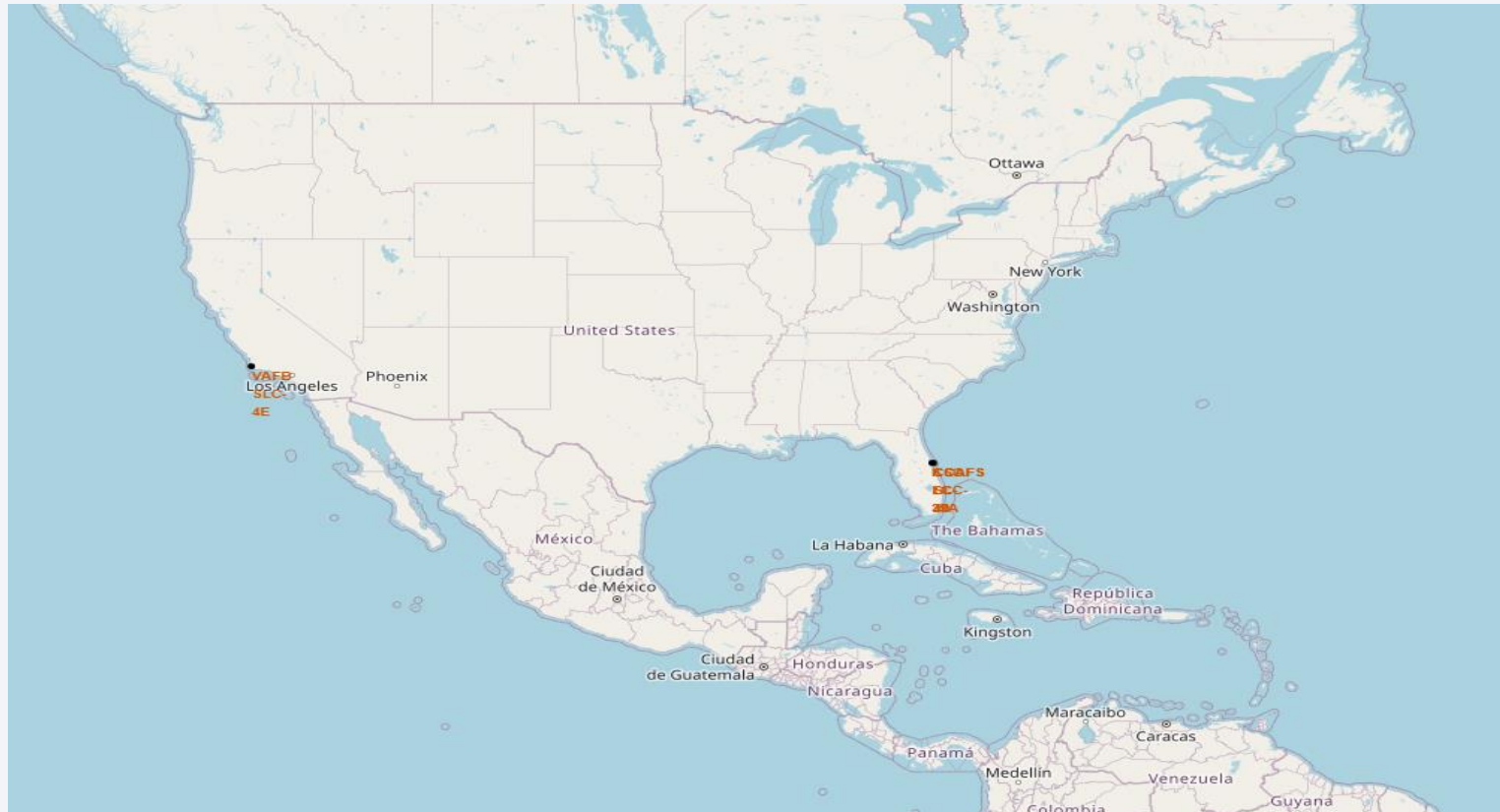
A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The background is a deep blue gradient.

Section 3

Launch Sites Proximities Analysis

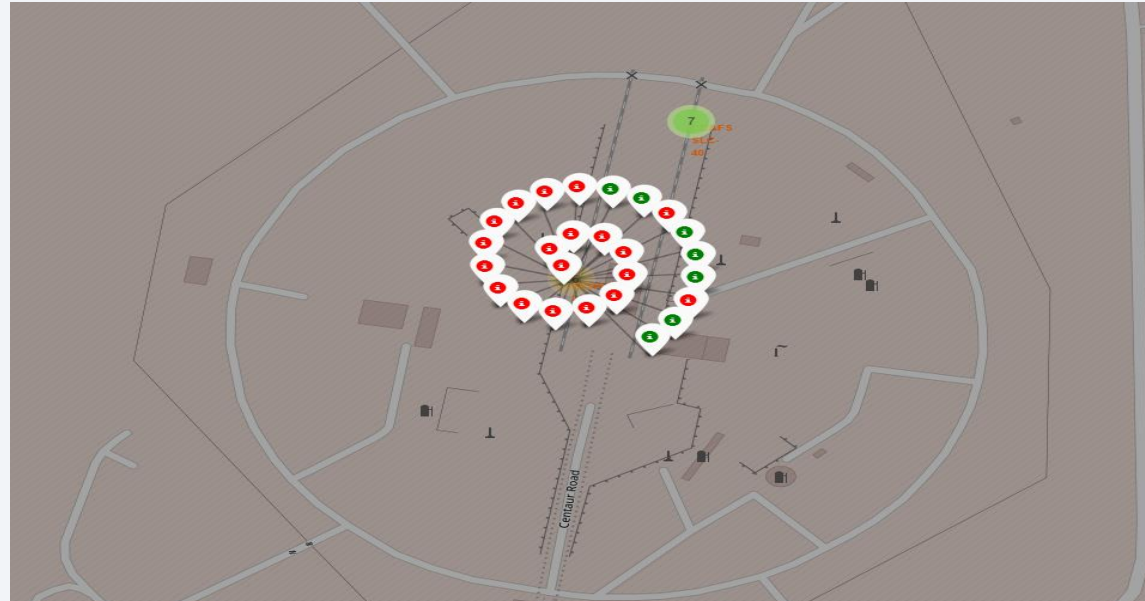
All launch sites

- Launch sites are near sea, probably by safety, but not too far from roads and railroads.



Launch Outcomes by Site

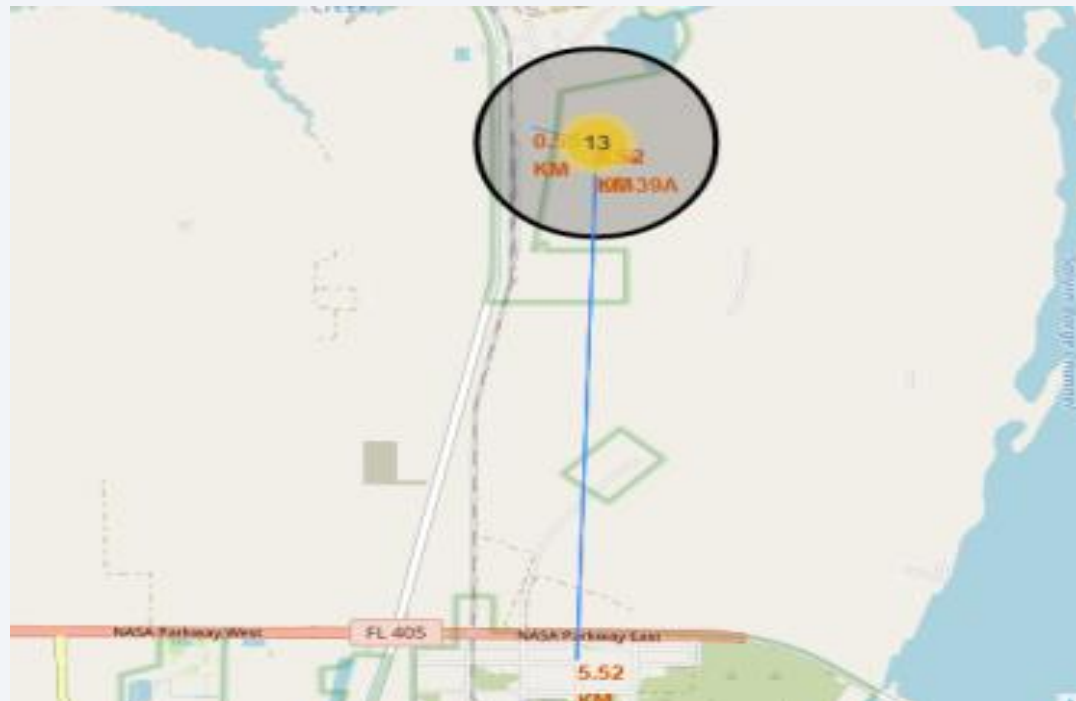
- Example of CCAFC LS-40 launch site launch outcomes



- Green markers indicate successful and red ones indicate failure.

Logistics and Safety

- Launch site KSC LC-39A has good logistics aspects, being near railroad and road and relatively far from inhabited areas.





Section 4

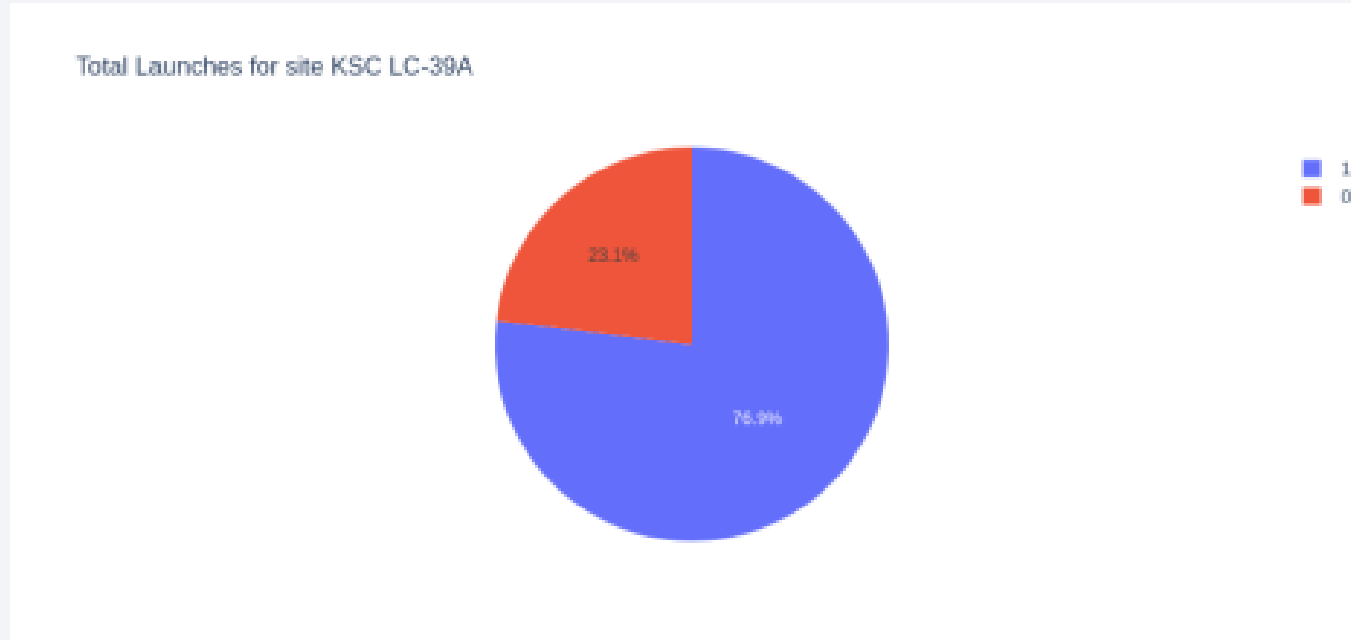
Build a Dashboard with Plotly Dash

Successful Launches by Site

- The place from where launches are done seems to be a very important factor of success of missions

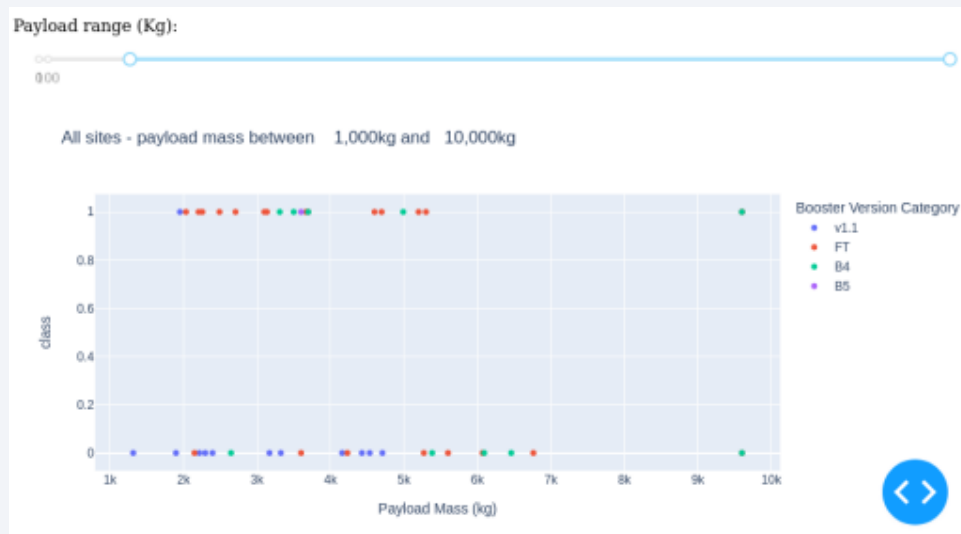


Launch Success Ratio for KSC LC-39A

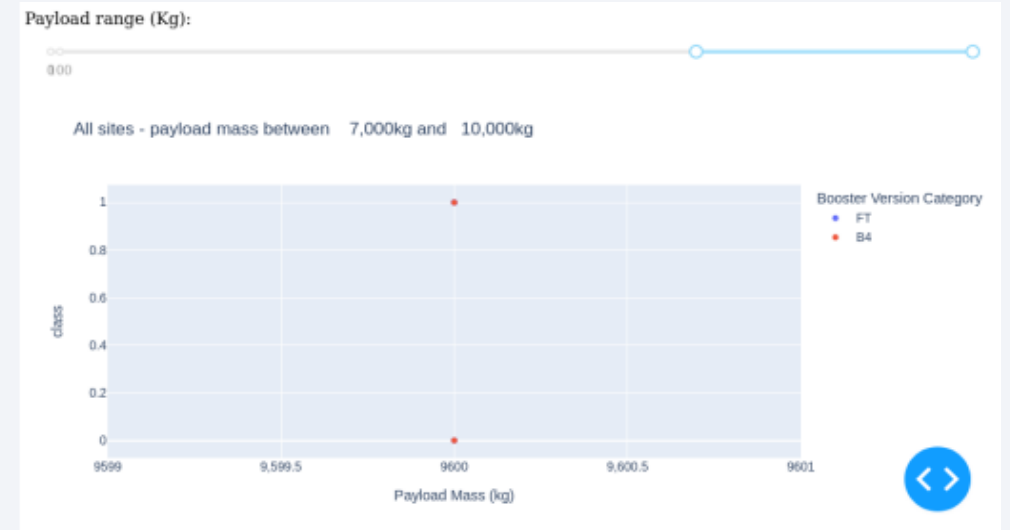


- Launch Success Ratio for KSC LC-39A

Payload vs. Launch Outcome



- Payloads under 6,000kg and FT boosters are the most successful combination.



- There's not enough data to estimate risk of launches over 7,000kg

Section 5

Predictive Analysis (Classification)

Classification Accuracy

```
parameters = {'criterion': ['gini', 'entropy'],
              'splitter': ['best', 'random'],
              'max_depth': [2*n for n in range(1,10)],
              'max_features': ['auto', 'sqrt'],
              'min_samples_leaf': [1, 2, 4],
              'min_samples_split': [2, 5, 10]}
```

```
tree = DecisionTreeClassifier()
```

```
grid_search = GridSearchCV(tree, parameters, cv=10)
tree_cv = grid_search.fit(X_train, Y_train)
```

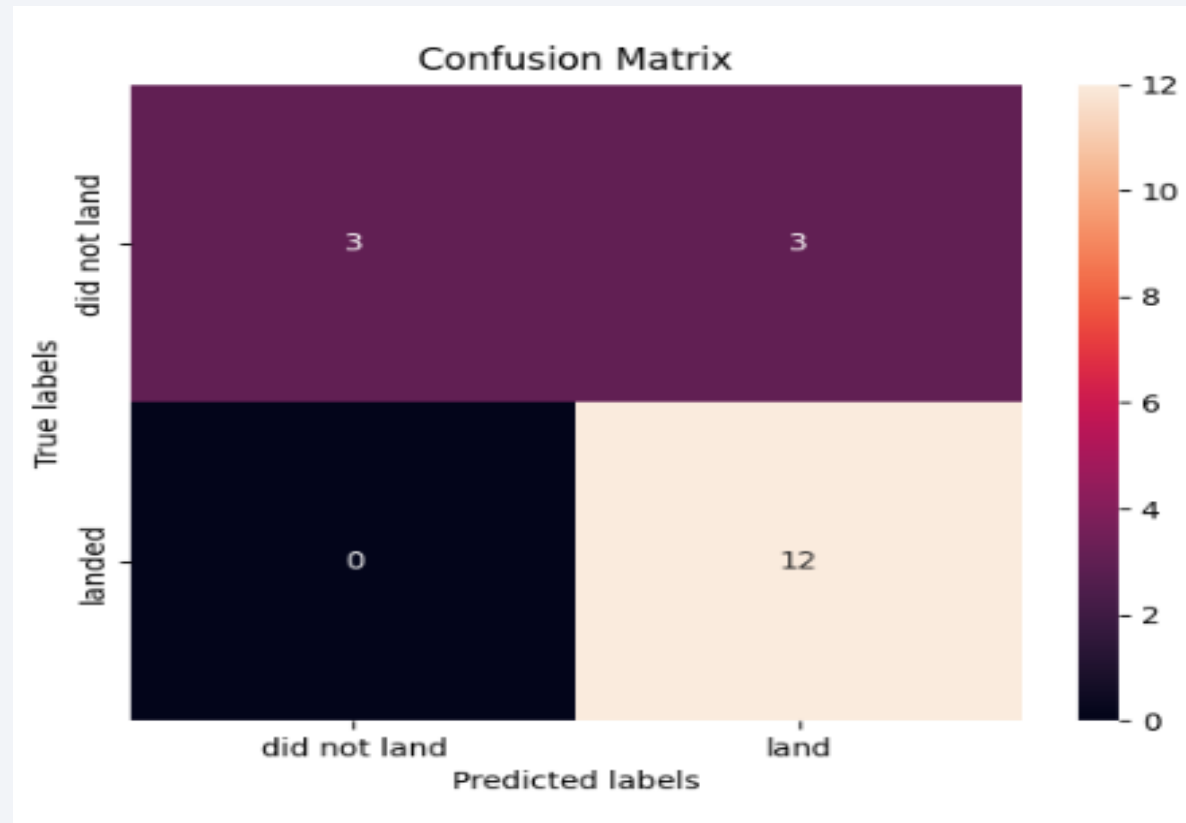
```
print("tuned hpyerparameters :(best parameters) ",tree_cv.best_params_)
print("accuracy :",tree_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters) {'criterion': 'entropy', 'max_depth': 4, 'max_features': 'auto', 'min_samples_leaf': 2, 'min_samples_split': 2, 'splitter': 'random'}
accuracy : 0.9017857142857144
```

- The decision tree classifier is the model with the highest classification accuracy

Confusion Matrix

- The confusion matrix for the decision tree classifier shows that the classifier can distinguish between the different classes. The major problem is the false positives .i.e., unsuccessful landing marked as successful landing by the classifier.



Conclusions

We can conclude that:

- The larger the flight amount at a launch site, the greater the success rate at a launch site.
- Launch success rate started to increase in 2013 till 2020.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- KSC LC-39A had the most successful launches of any sites.
- The Decision tree classifier is the best machine learning algorithm for this task.

Appendix

- Some of Cartality from Google Map

Thank you!

