Winning Space Race with Data Science

Abid Amir
12-11-2022

Outline

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

Executive Summary

- **Summary of methodologies**
  - **The Methodologies were used for Analyzing data:**
    - Data collection using Web Scraping and SPACEX API
    - EDA , including data Wrangling, Data visualization and interactive Visual Analytics
    - Machine learning Prediction
- **Summary of all results**
  - Possibility of Collection valuable Data from public resources
  - EDA allowed to :
    - summarize main characteristics of the data .
    - gain better understanding of the data set.
    - uncover relationships between different variables.
    - extract important variables for the best to predict success of launchings.
  - Machine Learning Prediction showed the best model to predict which characteristics are important to drive this opportunity by the best way, using all collected data.

Introduction

- **Project background and context**
  - A new Company SAPCE Y that would like to compete SPACE X founded by Billionaire industrialist Allon Musk. By gathering information about Space X and creating dashboards we will evaluate the viability of SPACEY to bid SPACEX.
- **Problems you want to find answers:**
  - The best way to estimate the total cost for launches, by predicting successful landings of the first stage of rockets.
  - What is the best Launch Site
  - The interaction amongst various features that determine the success rate of a successful landing.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data of SpaceX was collected from 2 sources:
    - SPACE X API :(https://api.spacexdata.com/v4/launches/past)
    - WebScraping from Wikipedia:(https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922)
- Perform data wrangling
  - Collected data was enriched by creating a landing outcome label from outcome column after summarizing and analyzing features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash

Methodology

Executive Summary

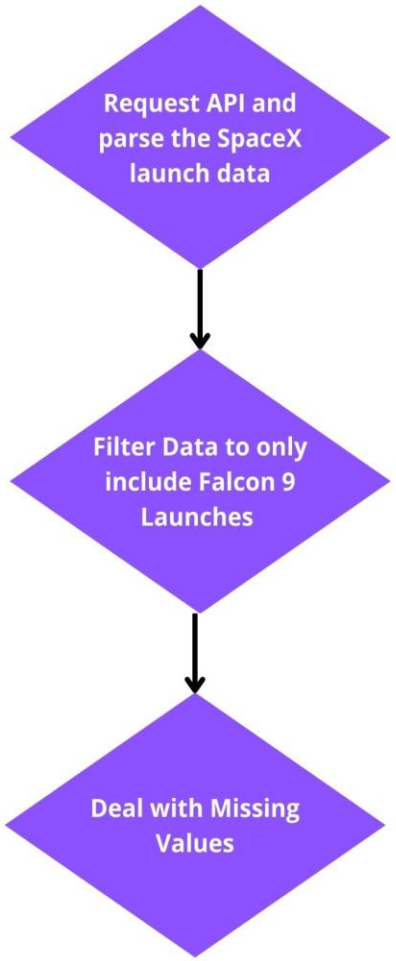
- Perform predictive analysis using classification models
  - Data that was collected until this step were normalized, divided in training and test data sets and evaluated by four different classification models, being the accuracy of each model evaluated using different combinations of parameters.

Data Collection

- How data sets were collected ?
  - Data was collected using The GET request from SPACE X API
  - Next, we decoded the response content as a Json using json() function call and turn it into a pandas dataframe using json_normalize.
  - We then filtred data, dealing with missing value by filling it where necessary.
  - In addition, we performed web scraping from Wikipedia for Falcon 9 launch records with BeautifulSoup
  - The objective was to extract the launch records as HTML table, parse the table and convert it to a pandas dataframe for future analysis

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
- Source Code:(https://github.com/EmirAbid/IBM-Data-Science-Capstone-/blob/main/Data%20Collection%20API.ipynb)

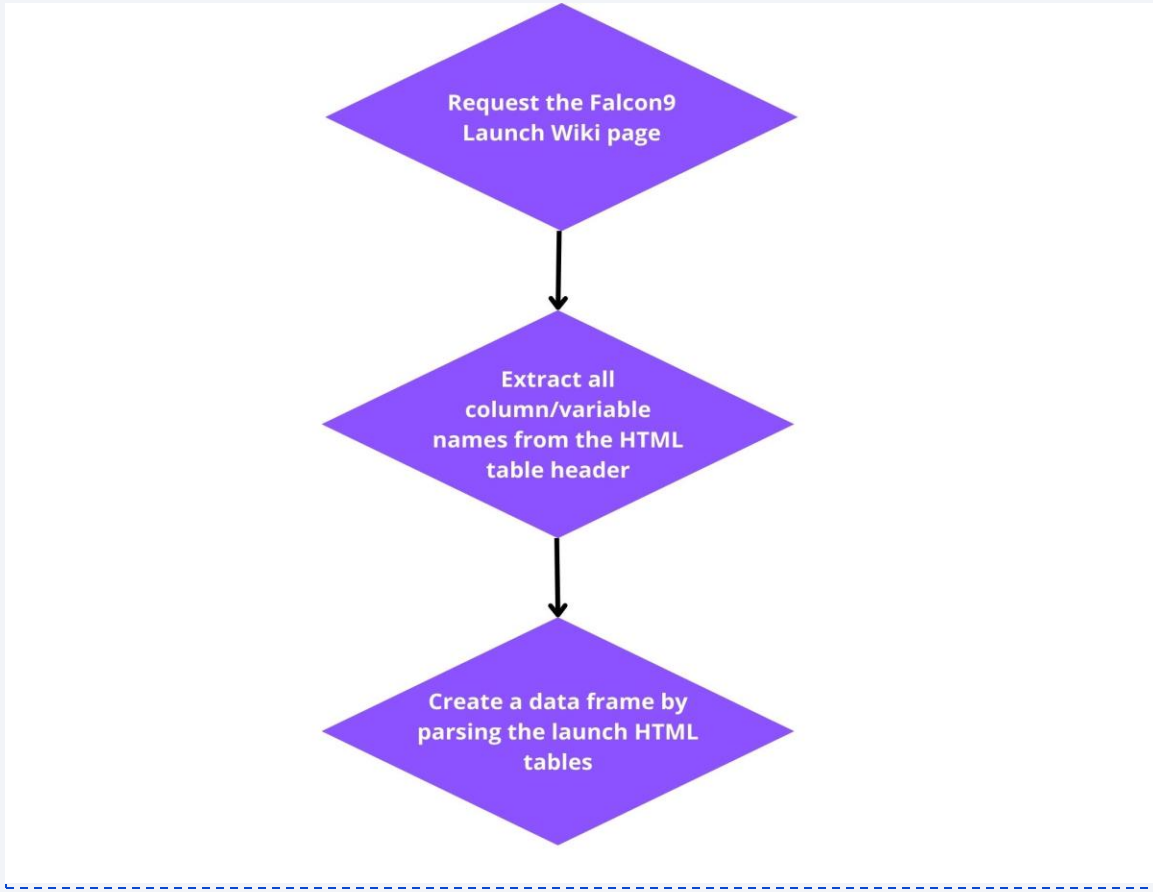


Data Collection - Scraping

- Data from SpaceX launches can also be obtained from Wikipedia

- Data are Downloaded from Wikipedia according to the flowchart and then persisted

- Code Source:(https://github.com/EmirAbid/IBM-Data-Science-Capstone-/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb)



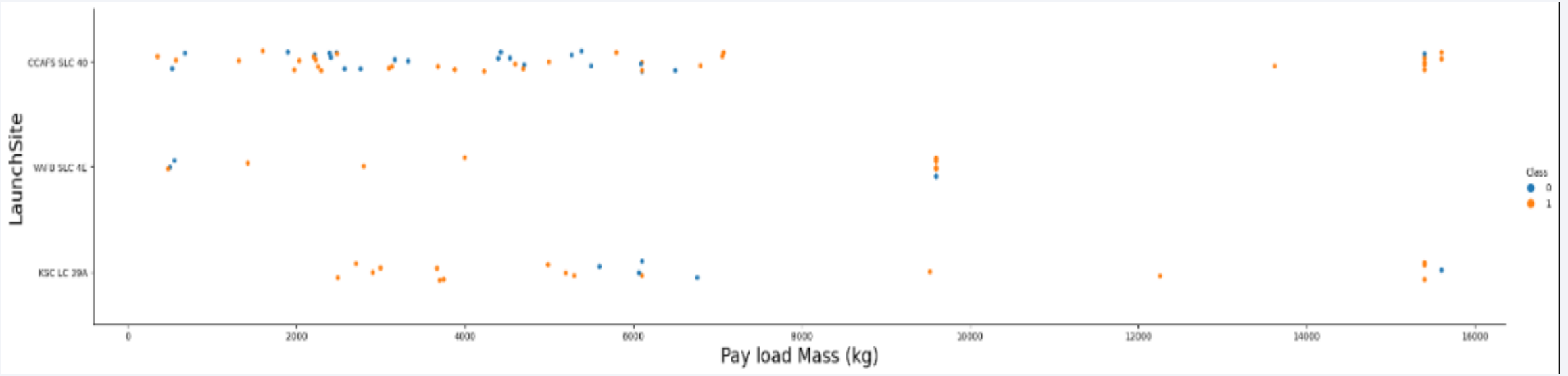
Data Wrangling

- Performing some initial exploratory data analysis on the data set
- Then Calculating the number of launches on each site.
- Calculating the number and occurrence of each orbit.
- Calculating the number and occurence of mission outcome per orbit type.
- Finally, create a landing outcome label from Outcome column.

- **Code Source:** (https://github.com/EmirAbid/IBM-Data-Science-Capstone-/blob/main/Data%20Wrangling.ipynb)

EDA with Data Visualization

- To explore data, scatterplots and barplots were used to visualize the relationship between pair of features:
  - Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, Payload and Orbit.



Code Source:(https://github.com/EmirAbid/IBM-Data-Science-Capstone-/blob/main/Data%20Visualization.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.

Code Source:(https://github.com/EmirAbid/IBM-Data-Science-Capstone-/blob/main/EDA%20Sqllite.ipynb)

Build an Interactive Map with Folium

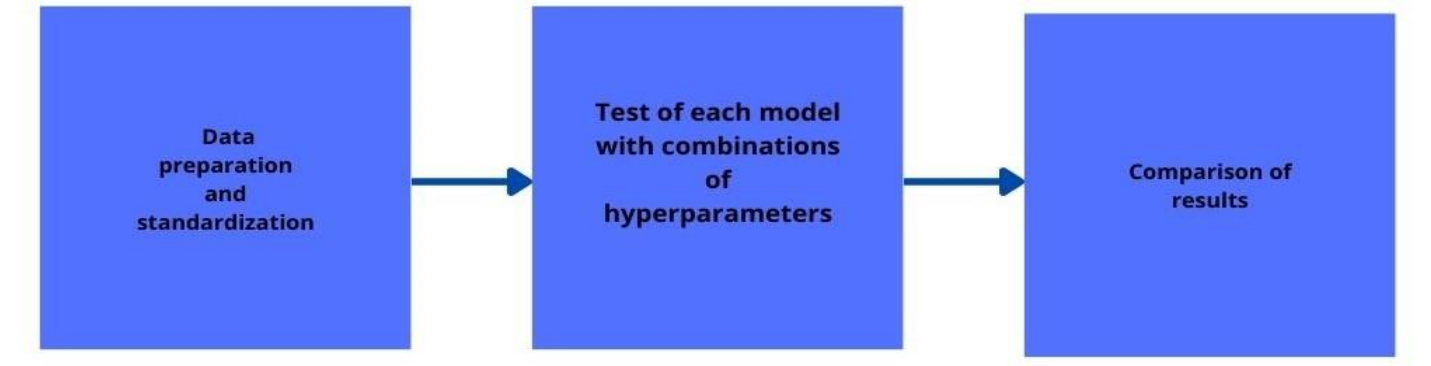
- Markers, circles, lines and marker clusters were used with Folium Maps
  - Markers indicate points like launch sites.
  - Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center.
  - Marker clusters indicates groups of events in each coordinate, like launches in a launch site
  - Lines are used to indicate distances between two coordinates.
- Code source:(https://github.com/EmirAbid/IBM-Data-Science-Capstone-/blob/main/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb)

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
- Explain why you added those plots and interactions
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

Predictive Analysis (Classification)

- Loading, transforming the data, split the data into training and testing.
- Building a four classification model (logistic regression, support vector machine, decision tree and k nearest neighbors) and tune different hyperparameters using GridSearchCV.
- Calculating the accuracy using score() function.
- Improving the model using feature engineering and algorithm tuning.
- Comparing the four model



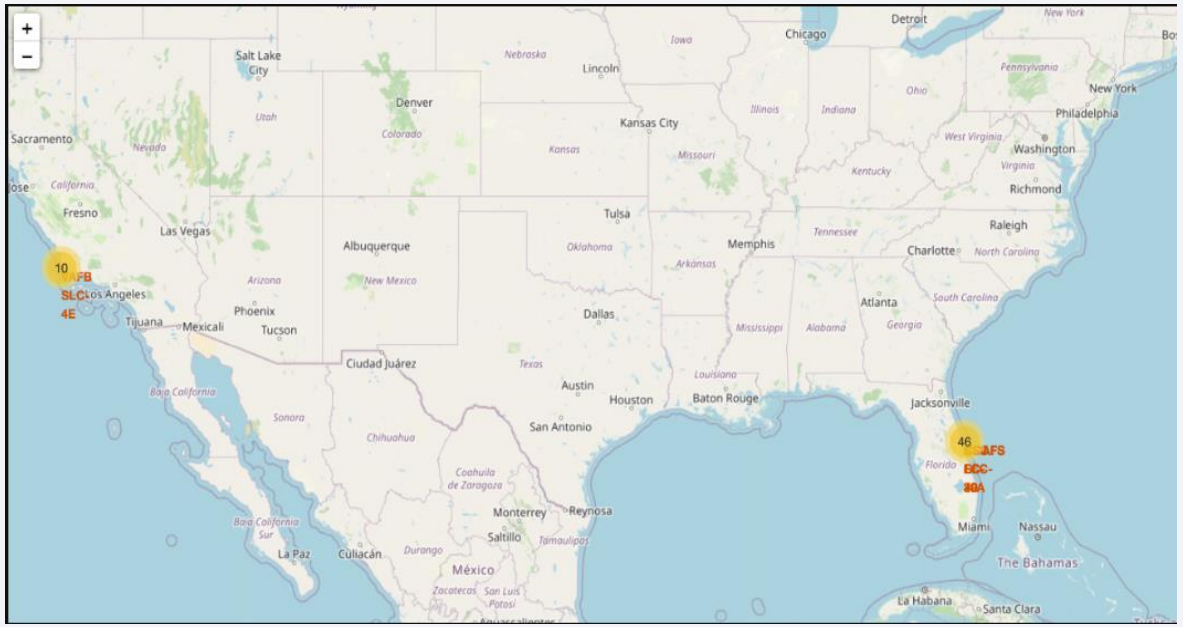
Code Source: (https://github.com/EmirAbid/IBM-Data-Science-Capstone-/blob/main/Machine%20Learning%20Prediction.ipynb)

Results

- Exploratory data analysis results:
  - SpaceX uses 4 launches Sites: CCAFS LC-40 ,CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E
  - Average payload mass carried by booster version F9 v1.1: 2928
  - The first success landing outcome happened in 2015 fiver year after the first launch;
  - Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015;
  - The number of landing outcomes became as better as years passed.

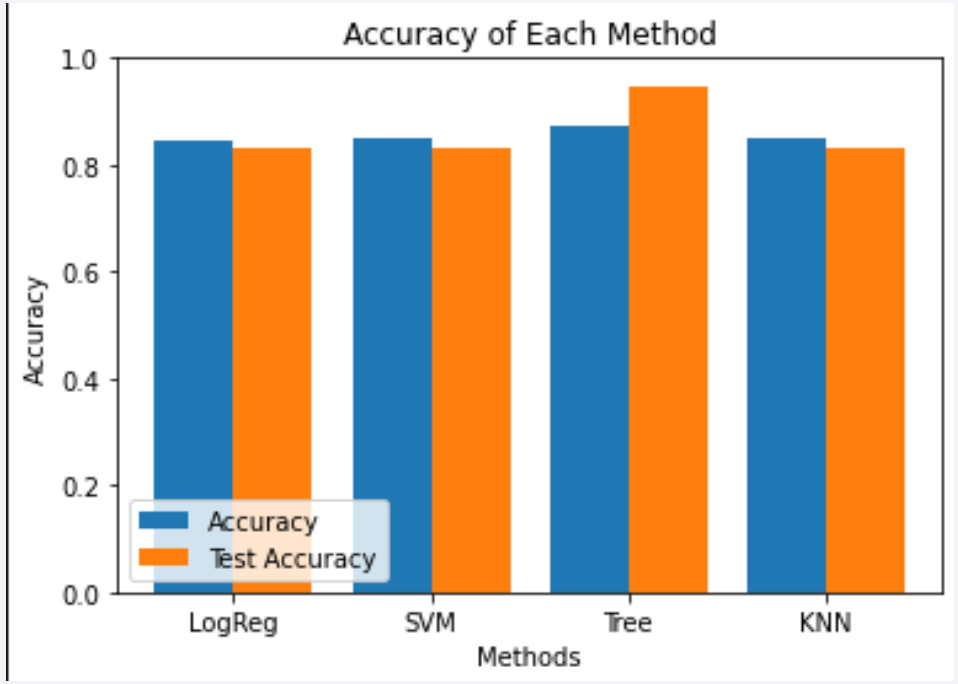
Results

- Using interactive analytics was possible to identify that launch sites use to be in safety places, near sea, for example and have a good logistic infrastructure around.



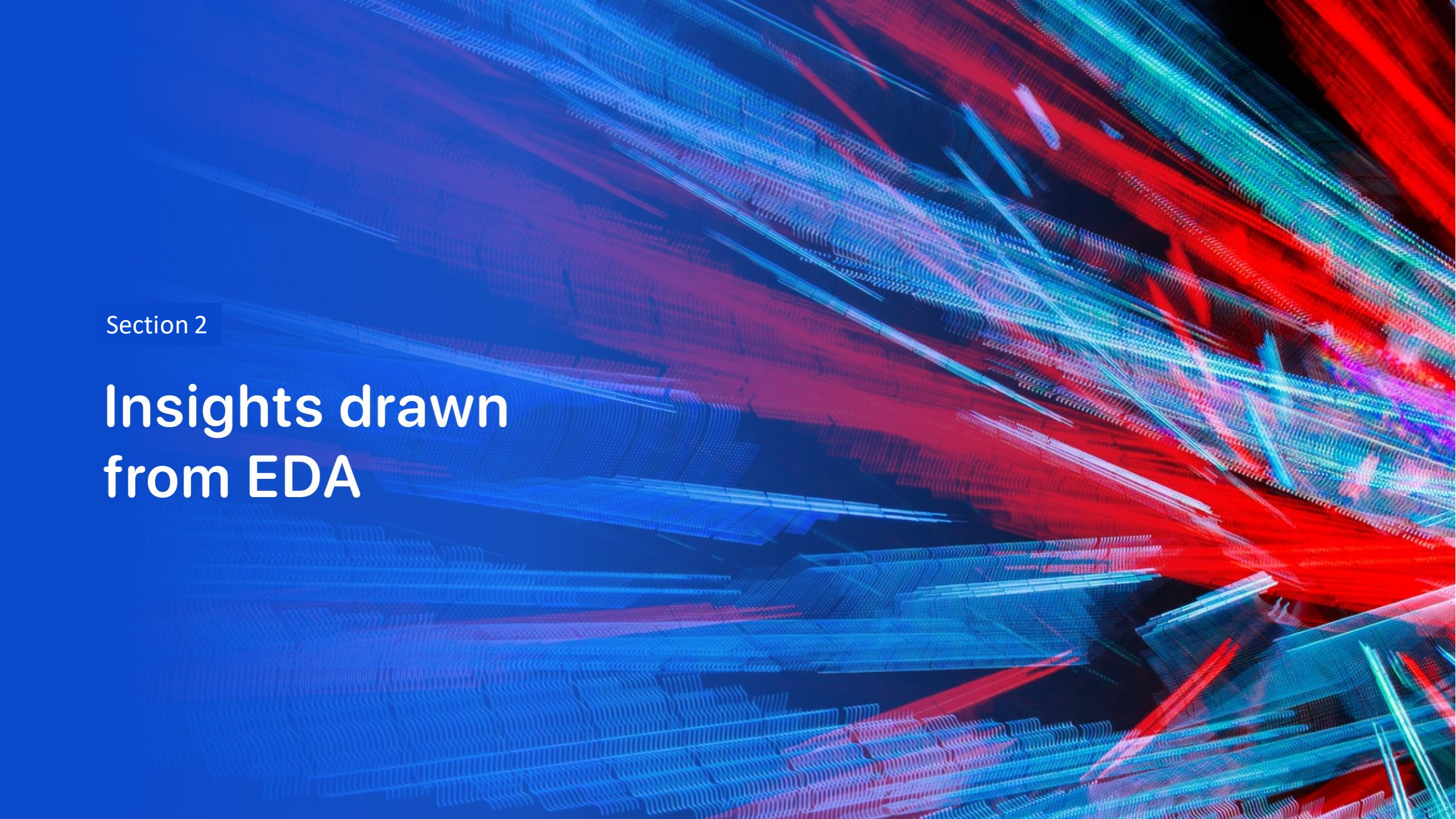
Results

Predictive analysis revealed that Decision Tree Classifier is the best model to predict successful landings which have accuracy approximately 87% and accuracy for test data approximately 94%

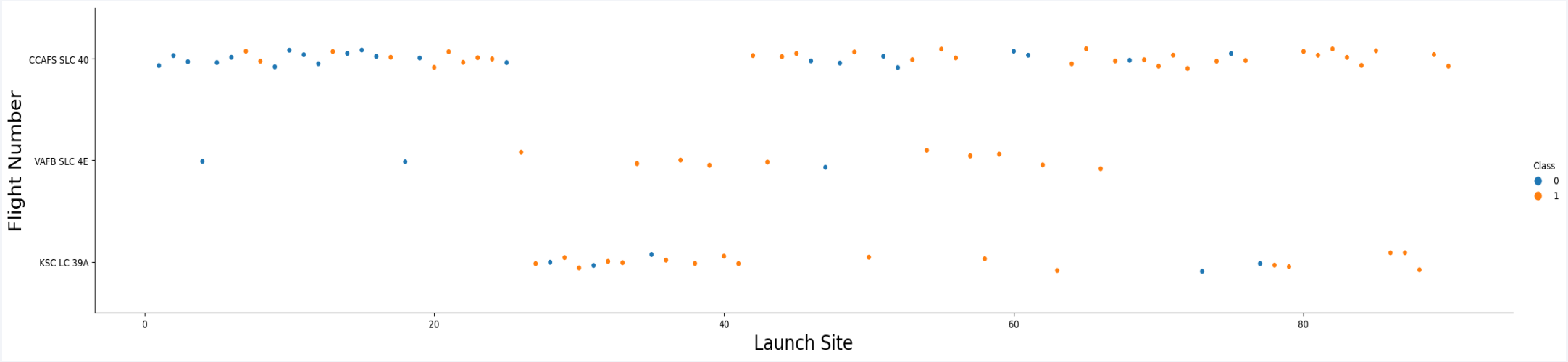


Section 2

Insights drawn from EDA

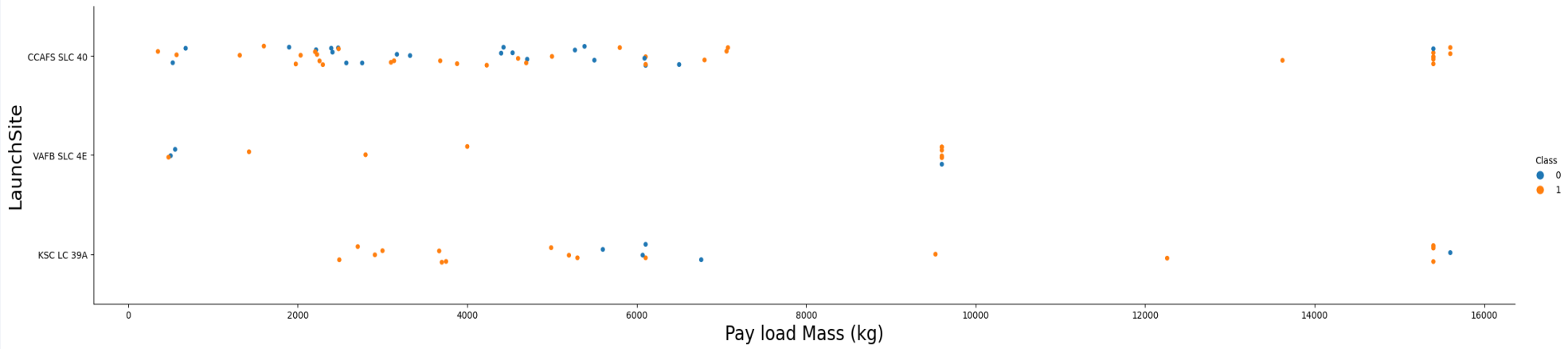


Flight Number vs. Launch Site



- According to the plot above, it revealed that the best launch site is CCAF5 SLC 40, the recent launches were successful ,which has the biggest number of flight
- In the second place we found VAFB SLC and in the third place KSC LC 39A

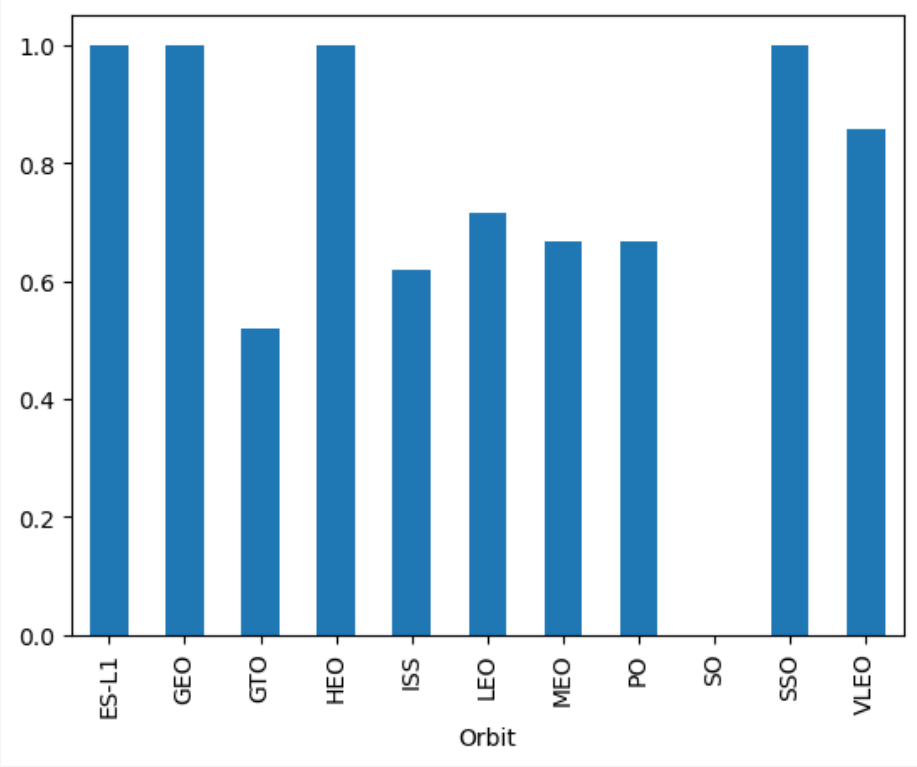
Payload vs. Launch Site



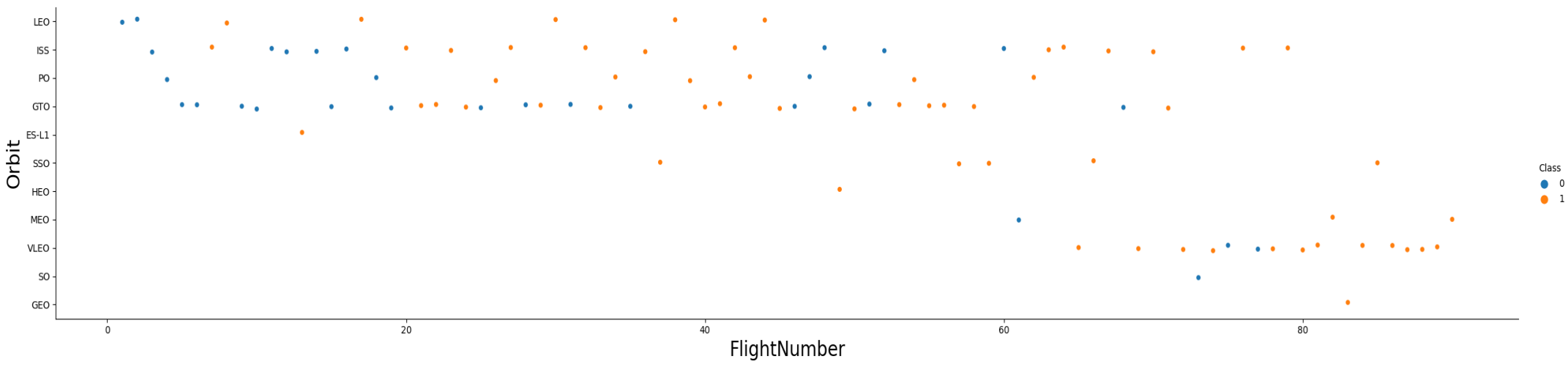
- Payloads over 9,000kg have excellent success rate.
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.

Success Rate vs. Orbit Type

- The biggest success rates happens to orbits:
  - ES L1
  - GEO
  - HEO
  - SSO.
- Followed by:
  - VLEO (above 80%)
  - LFO (above 70%).

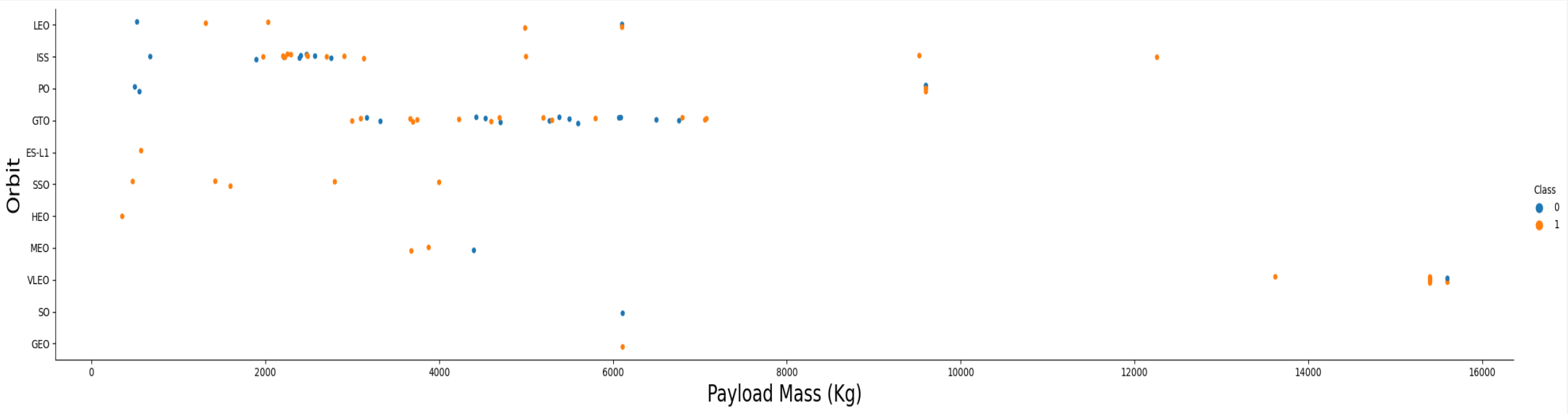


Flight Number vs. Orbit Type



- Apparently, more the FlightNumber increase the more success rate is improved
- VLEO orbit seems a new business opportunity, due to recent increase of its frequency.

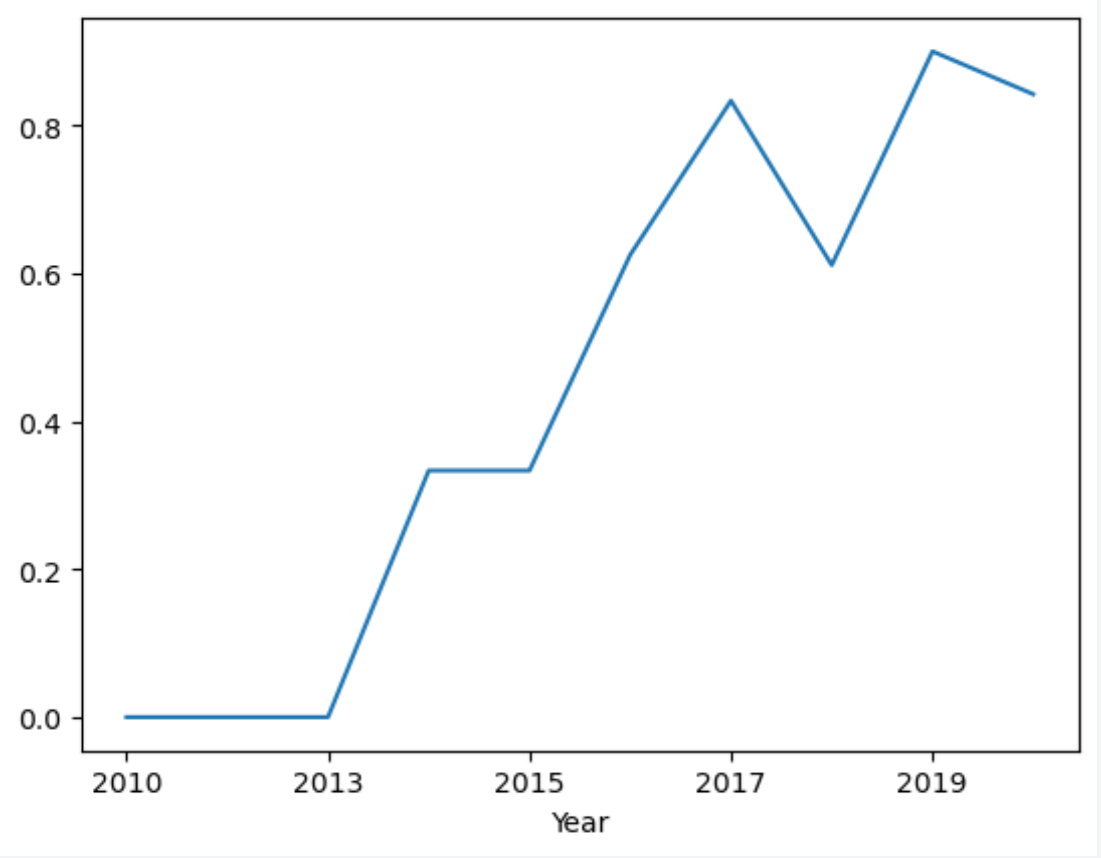
Payload vs. Orbit Type



- It is revealed above, that there is no relation between Payload and Orbit Type.
- ISS has a better range of Payload and a good rate of success.

Launch Success Yearly Trend

- Between 2010 and 2013 launch Success is constant and equal to zero
- Between 2013 and 2019 the curve is increased until achieve 89% of success



All Launch Site Names

There are four launch sites : CCAFS LC-40 ,CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E

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

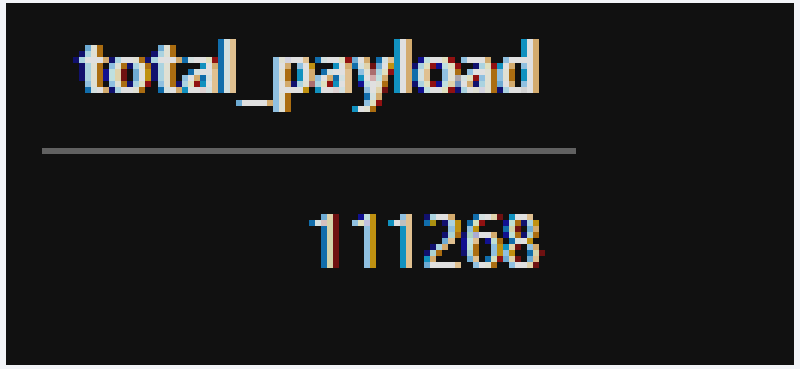
Launch Site Names Begin with 'CCA'

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

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

Total Payload Mass

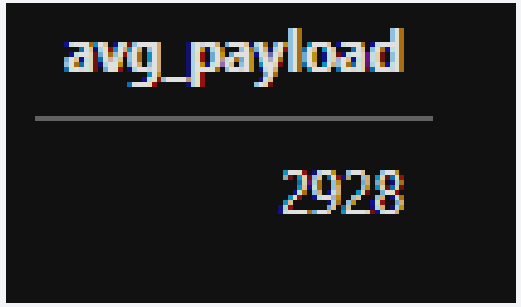
- the total payload mass carried by boosters launched by NASA (CRS)



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

Average Payload Mass by F9 v1.1

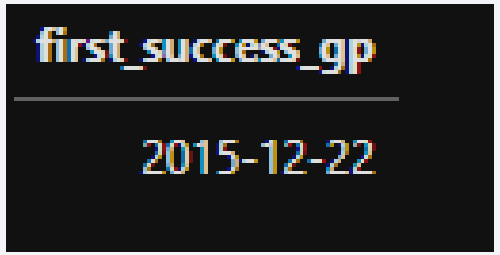
- Average payload mass carried by booster version F9 v1.1:



Filtering data by the booster version above and calculating the average payload mass we obtained the value of 2,928 kg.

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad



By filtering data by successful landing outcome on ground pad and getting the minimum value for date it's showed that the first occurrence happened on 12/22/2015.

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

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

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes :

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

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass

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

2015 Launch Records

- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

booster_version	launch_site	MONTH
F9 v1.1 B1012	CCAFS LC-40	1-10
F9 v1.1 B1015	CCAFS LC-40	4-14

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:

landing_outcome	quantity
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

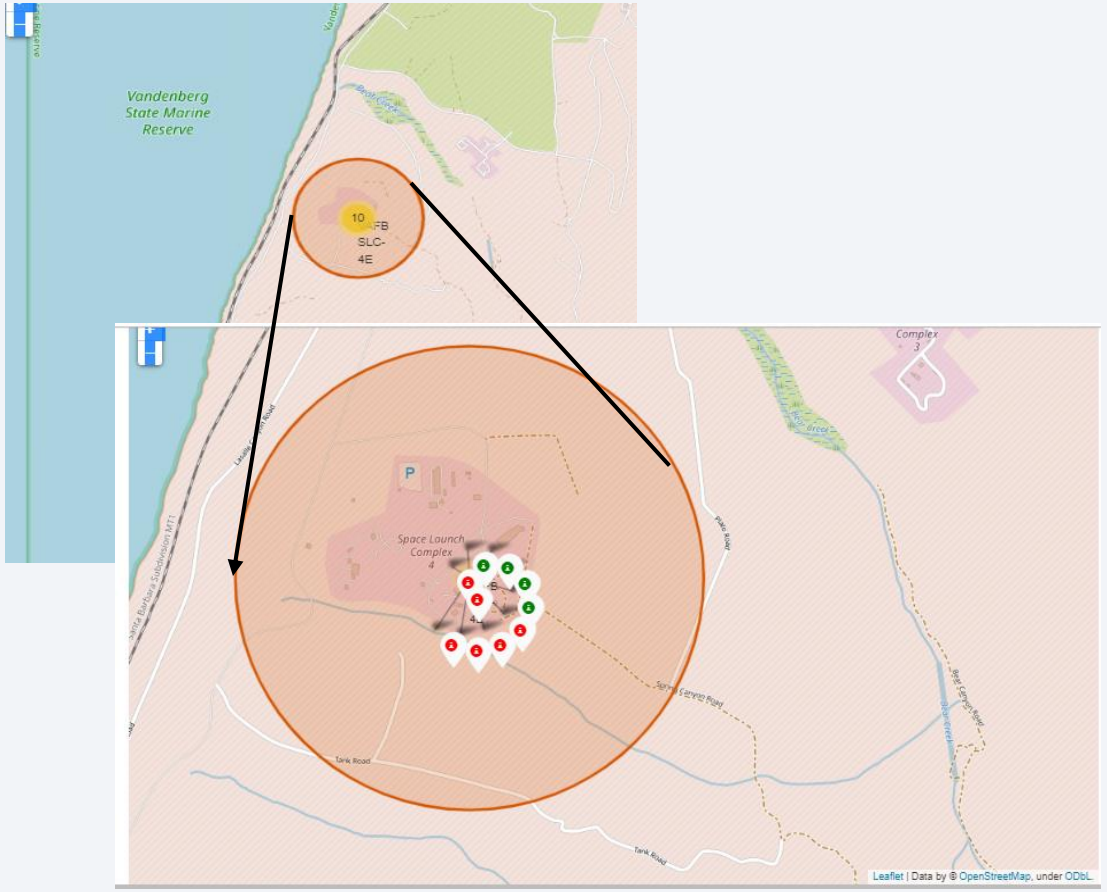

Section 3
Launch Sites
Proximities Analysis

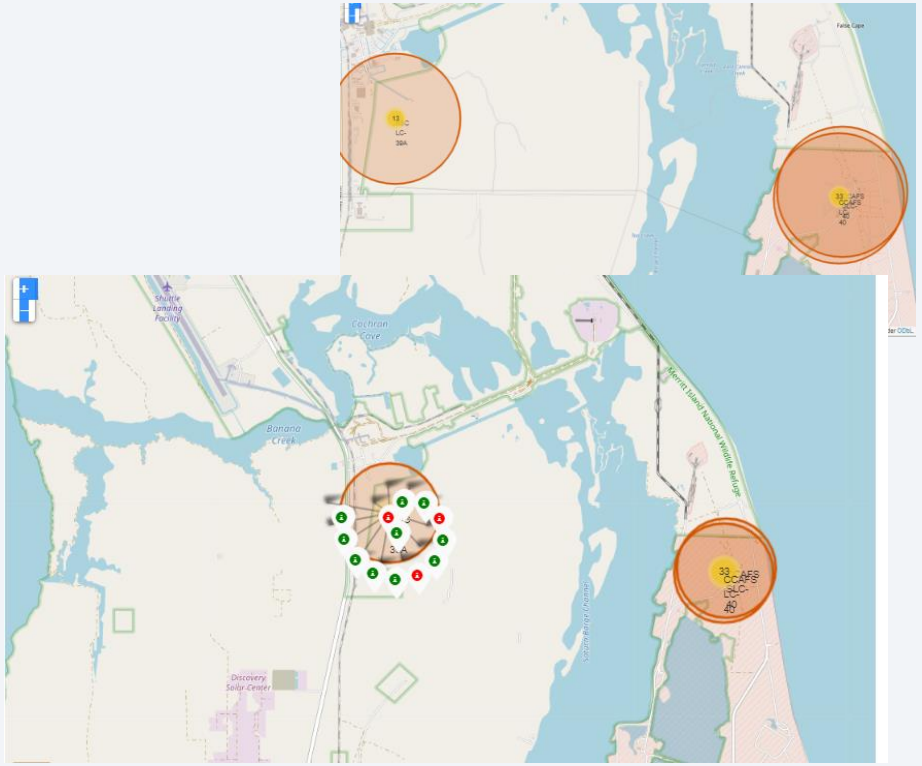
All Launches Sites

- The folium show that Spacex launches located in USA, near the sea.

Launch Outcomes by Site

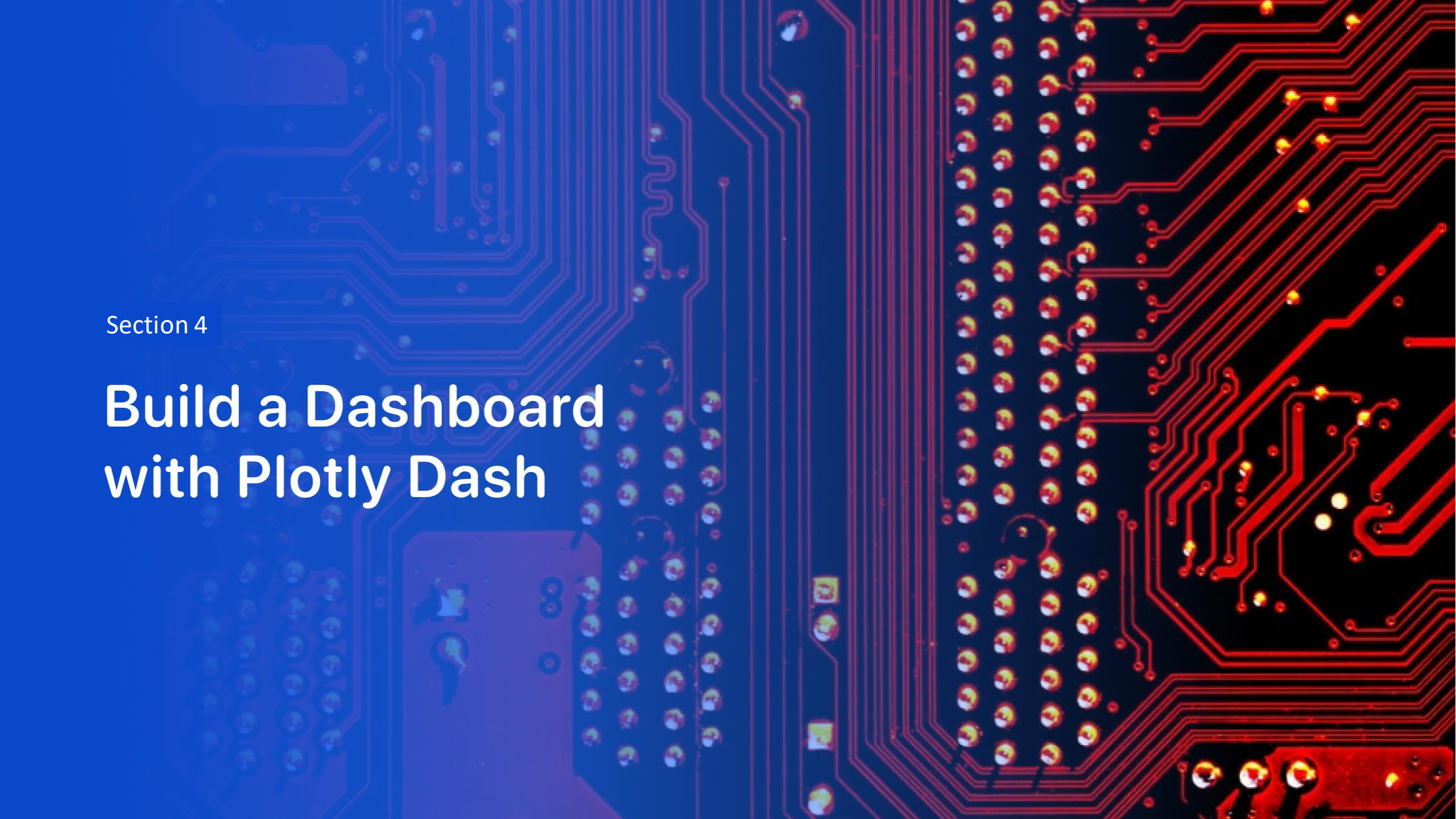
Example of VAFB SLC-4E/KSC LC 39A launch site launch outcomes



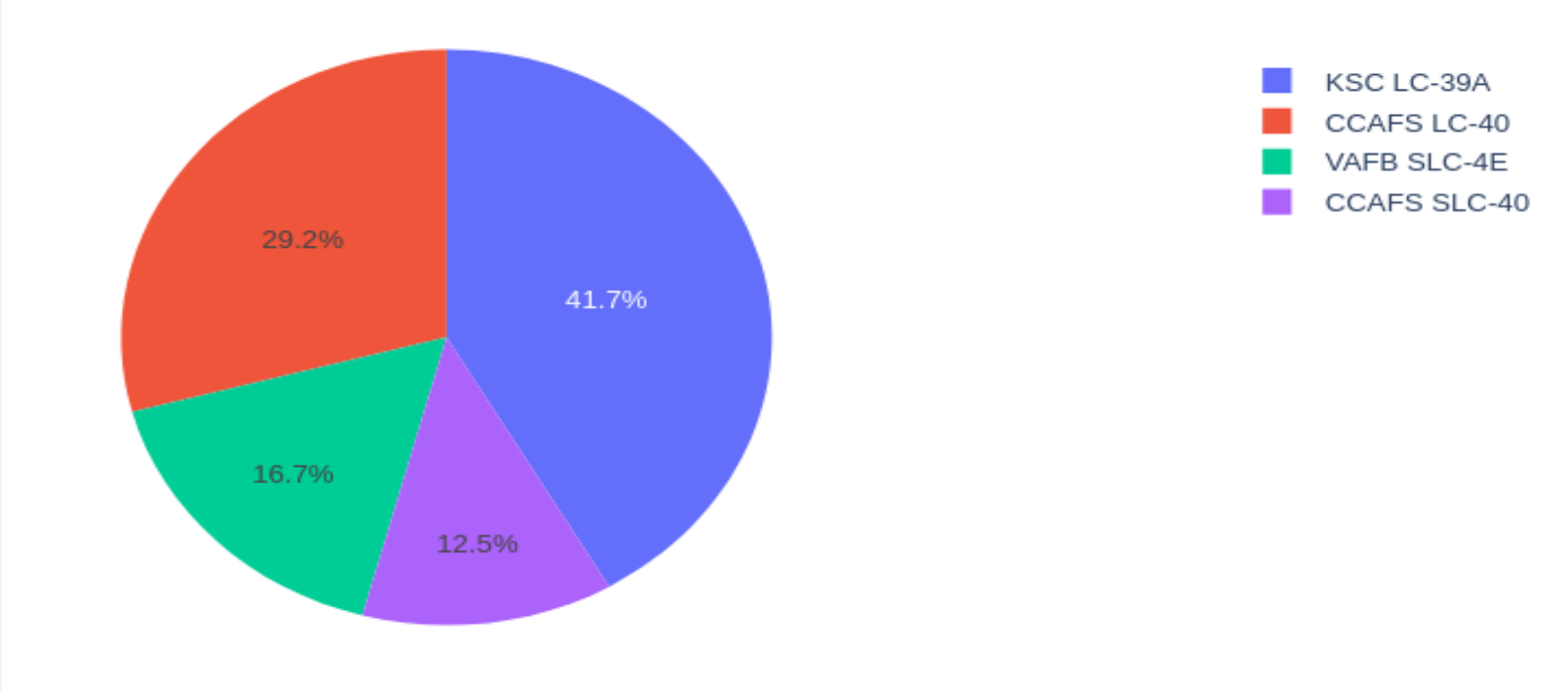


Logistics and Safety

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

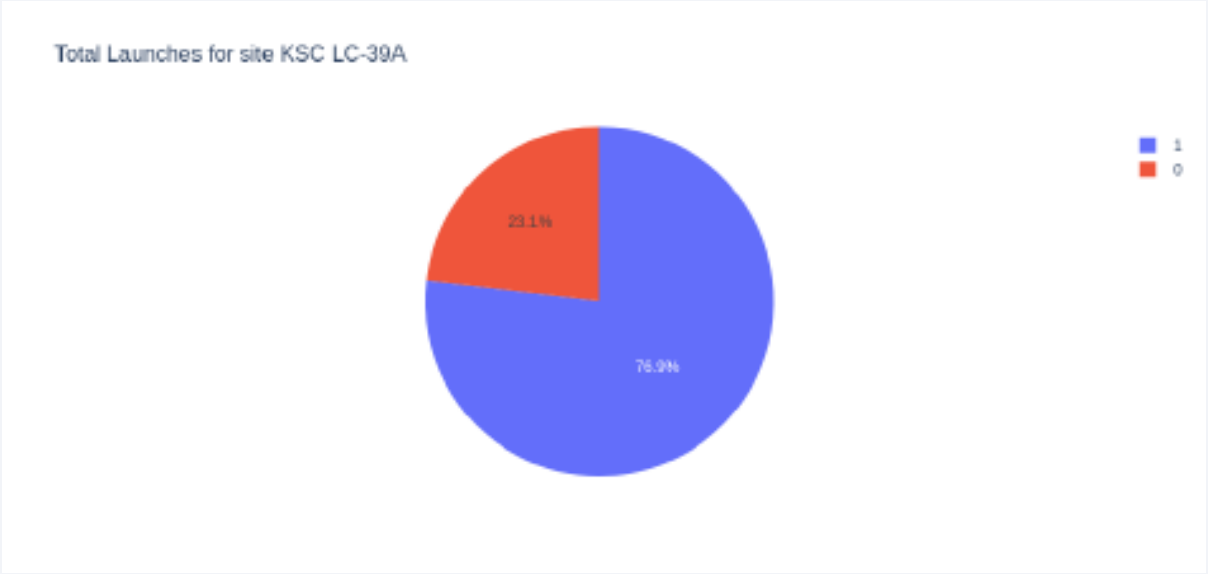

Section 4
Build a Dashboard with Plotly Dash

The Success Rate by each launch Site



- KSC LC-39A has the most successful rate , equal to 41.7% .

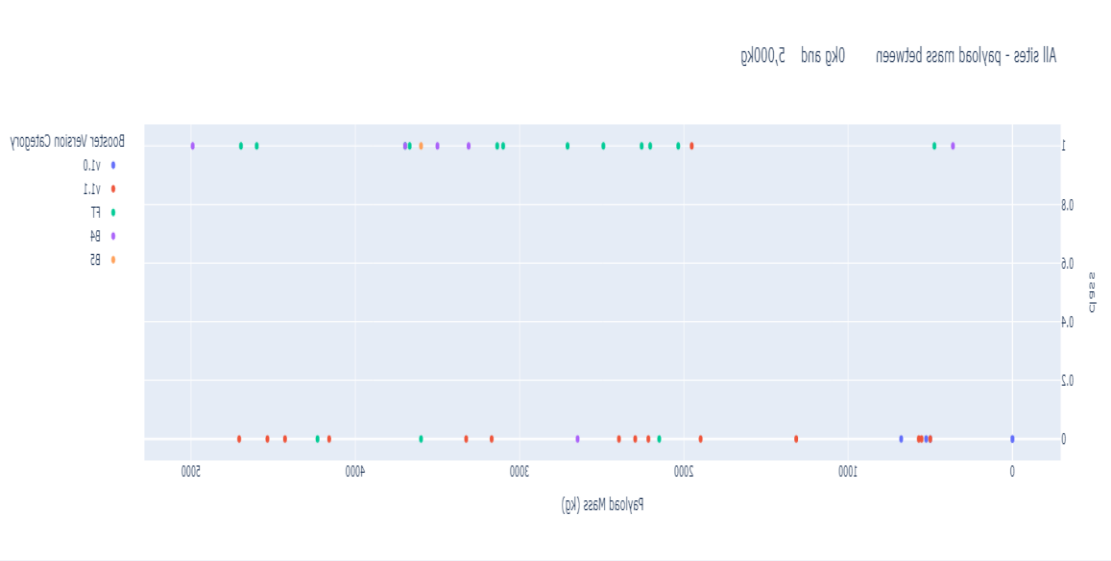
Launch Success Ratio for KSC LC-39A



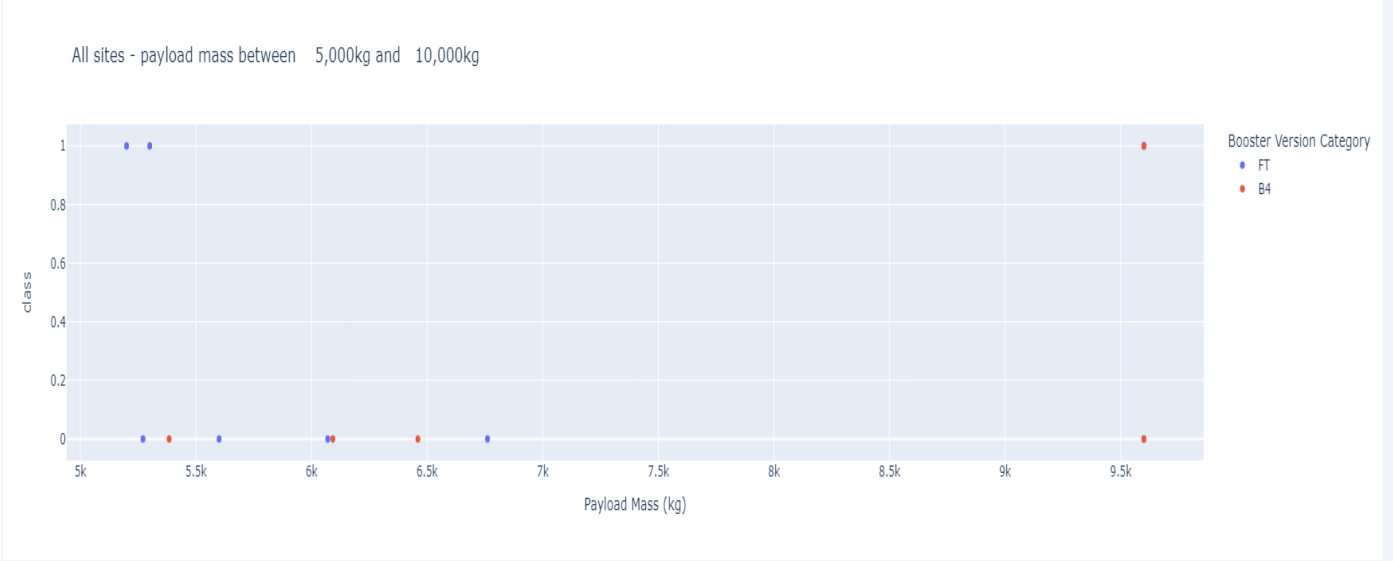
- KSC LC-39 has achieved 76.9% of success and 23.1% of failure

Payload vs. Launch Outcome

Low Weighted Payload 0kg-5000kg



Heavy Weighted Payload 5000kg-10000kg

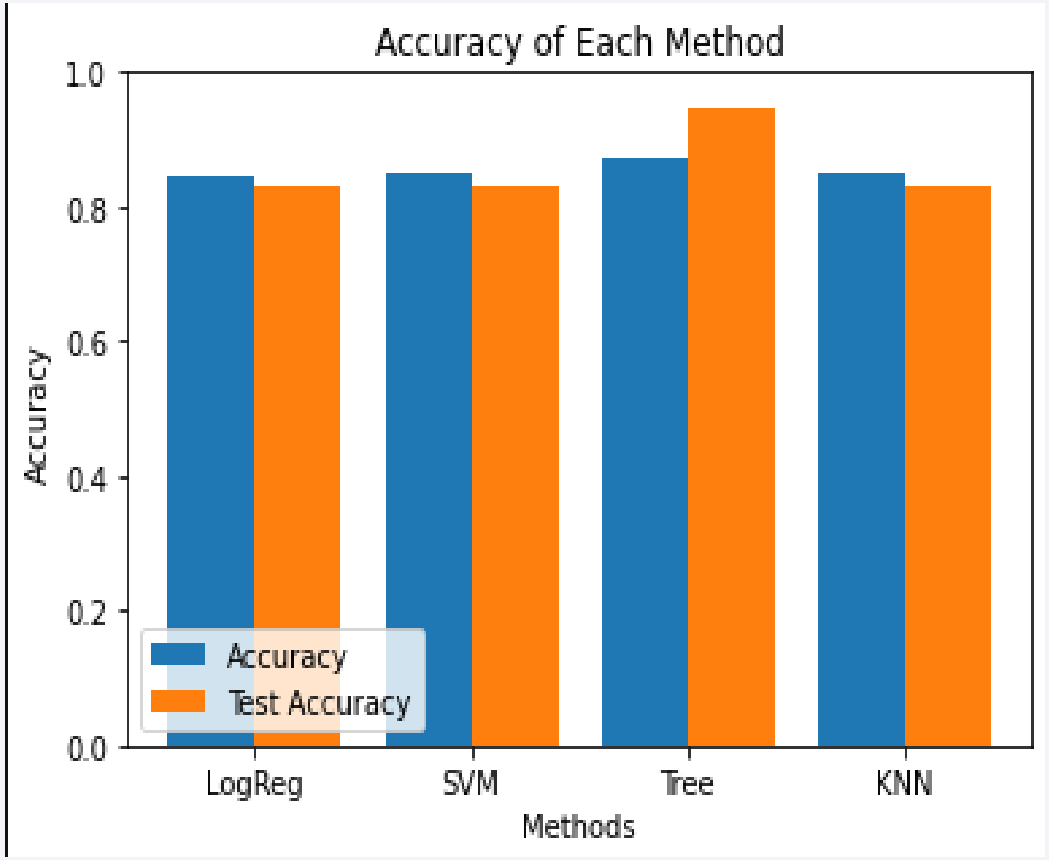


Section 5

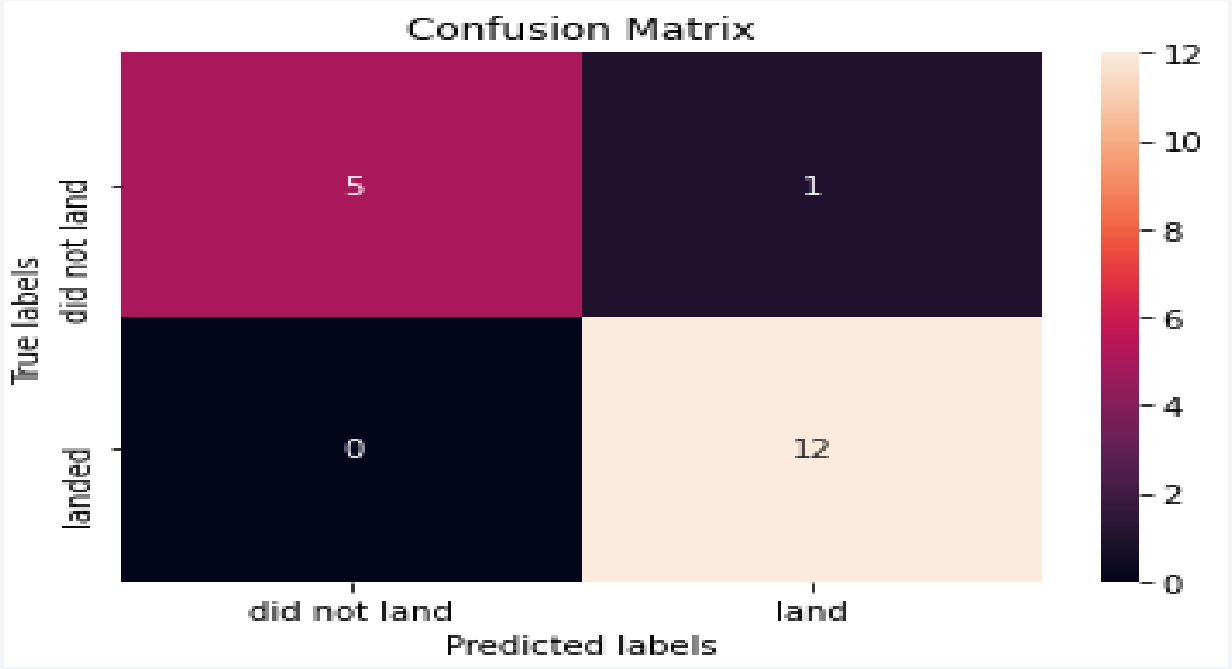
Predictive Analysis (Classification)

Classification Accuracy

- We compared the four classification Model and the result is shown beside
- The result show that Decision Tree Model has the highest accuracuy, equal to 87%(for trainig_set) and 96% (for the testing set)



Confusion Matrix



- The confusion Matrix of Decision Tree Classifier shows that the accuracy rating is high

Conclusions

- Launches above 7000kg are less risky.
- More the FlightNumber increase, more the accuracy of success increase
- 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 best model for this task

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!