IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Andre Correa
10/07/22

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies for Data Analysis:

  - Data Collection using SpaceX API and web scrapping;

  - EDA (Exploratory Data Analysis) , including Data Visualization, Data Wrangling & Interactive Visual Analytics;

  - Machine Learning Prediction.

- Summary of all results:

  - Valuable and informative Data was collected from public sources;

  - EDA was fundamental to discover the best features to predict a launch's success;

  - 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

- Space X advertises Falcon 9 rocket launches with a cost of 62 million dollars, while other providers cost upwards 165 million dollars each. Much of the saving is due to Space X's ability to reuse the first stage. Therefore, if we can predict if the first stage will land successfully, we can determine the cost of the launch. The objective of this work is to find out the viability of a new company 'Space Y' to compete with 'Space X' for rocket launching services.

- Desirable conclusions:

  -The best way to predict the success of landing of the first stage of the rockets;

  -Determine whether the location of the launch is relevant for its success;

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Space X's data was collected from two sources:
    - Space X API [https://api.spacex.com/v4/rockets/]
    - Webscraping
    [https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches]
- Perform data wrangling
  - Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features.
- Perform exploratory data analysis (EDA) using visualization and SQL

Methodology

Executive Summary

- Perform interactive visual analytics using Folium and Plotly Dash.
- Perform predictive analysis using classification models
  - The data that was collected until this step was normalized, divided into training and testing data subsets and evaluated by four different classification models, being the accuracy of each model tested using different combinations of parameters.

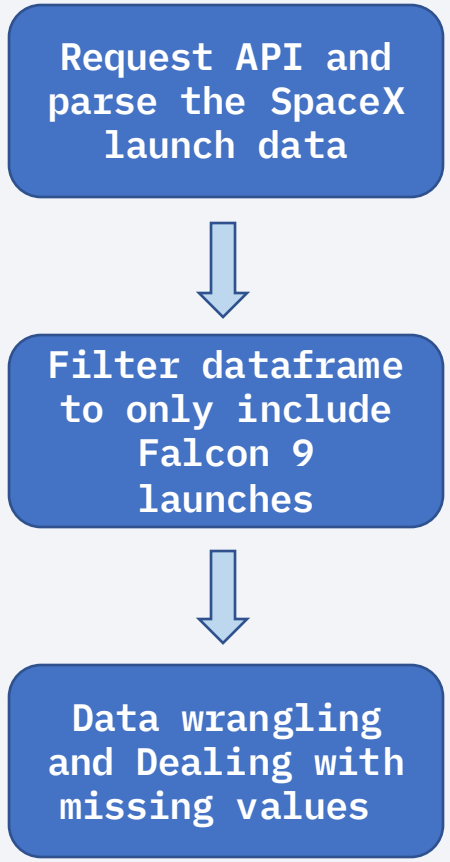
Data Collection

- The data set were collected from:

- Space X API [https://api.spacex.com/v4/rockets/] and from
- Wikipedia using webscraping techniques [https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches]

Data Collection – SpaceX API

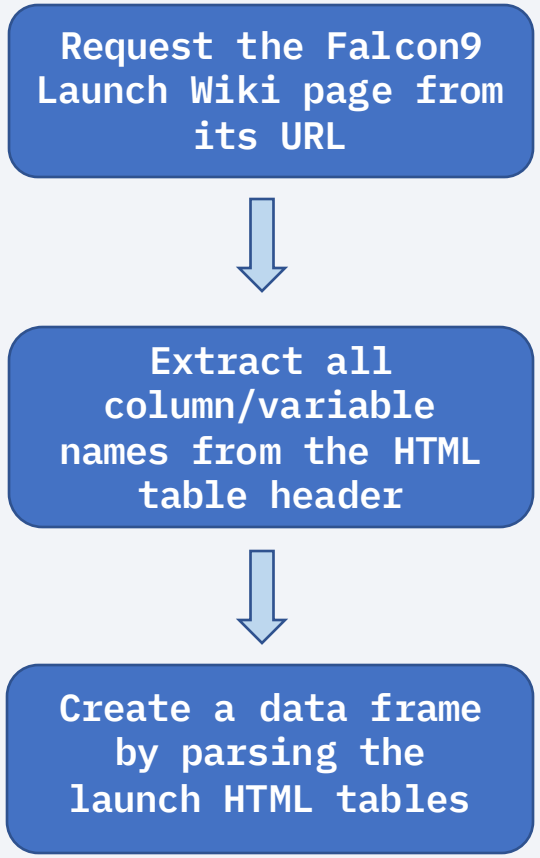
- Space X provides public access to API, where we obtained and used the data;
- After SpaceX API calls were made we used it according to the flowchart on the right.
- Source code: https://github.com/Andre-Larc/Rocket-Launch/blob/master/Data%20Collection%20API%20-%20From%20SpaceX.ipynb



Data Collection - Scraping

- Performed web scraping to collect Falcon 9 historical launch records from a Wikipedia page.
- The Data was obtained and processed according to the flowchart on the right
- Source Code:

https://github.com/Andre-Larc/Rocket-Launch/blob/master/Data%20Collection%20with%20Web%20Scraping%20-%20Rocket%20Launch.ipynb



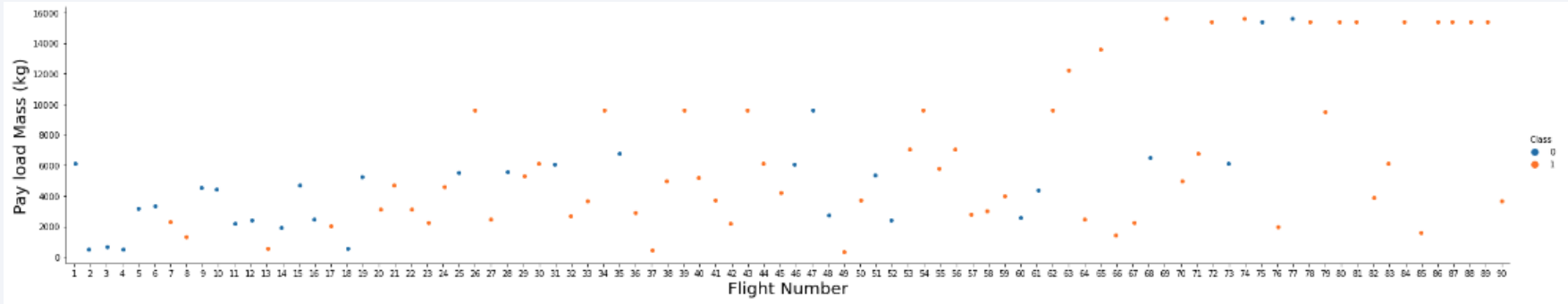
Data Wrangling

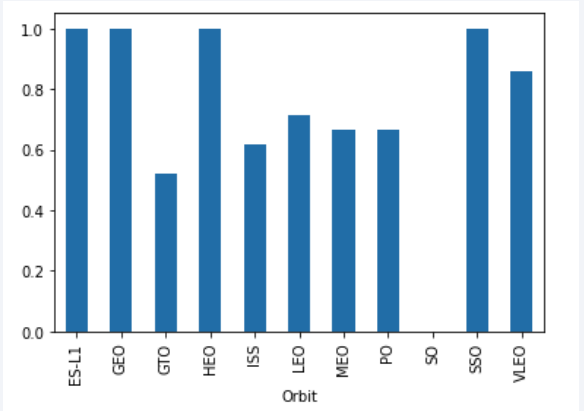
- Initially some Exploratory Data Analysis (EDA) was performed on the dataset.
- Then the summaries launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated.
- Finally, the landing outcome label was created from Outcome column.

- Source code: https://github.com/Andre-Larc/Rocket-Launch/blob/master/Data%20Wrangling%20-%20SpaceX.ipynb

EDA with Data Visualization

- Scatterplots and barplots were used to visualize the relationship between certain features:
  - Payload mass x Flight Number;
  - Launch site x Flight Number;
  - Launch site x Payload mass;
  - Orbit type x Flight Number;
  - Payload x Orbit type





- Source code: https://github.com/Andre-Larc/Rocket-Launch/blob/master/EDA%20with%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; and
    - 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.

- Source code: https://github.com/Andre-Larc/Rocket-Launch/blob/master/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters were used with Folium Maps.
  - Used 'Circles' to highlight an area around specific coordinates where the space stations, like NASA Johnson Space Station, are located;
  - Used 'Markers' to indicate the location of launch sites on the map;
  - Used 'Marker Clusters' indicating events in the surroundings of each coordinate like the position where each launch happened at a launch site;
  - Used 'Lines' to visually show the distance between two coordinates, like the distance from a launch site to the nearest coastline.

- Source code: https://github.com/Andre-Larc/Rocket-Launch/blob/master/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

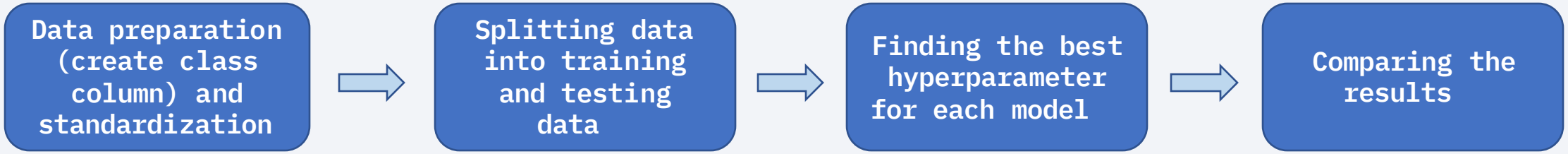
Build a Dashboard with Plotly Dash

- The following graphs and plots were used to visualize data:
  - Percentage of launches by site
  - Payload range
- This interaction was essential to analyze the relation between payloads and launch sites, helping to conclude the best location to launch from based on payload.

- Source code: https://github.com/Andre-Larc/Rocket-Launch/blob/master/SpaceX_Dash_App_2.py

Predictive Analysis (Classification)

- Compared 4 classification models: Logistic regression, support vector machine, decision tree and k nearest neighbors.



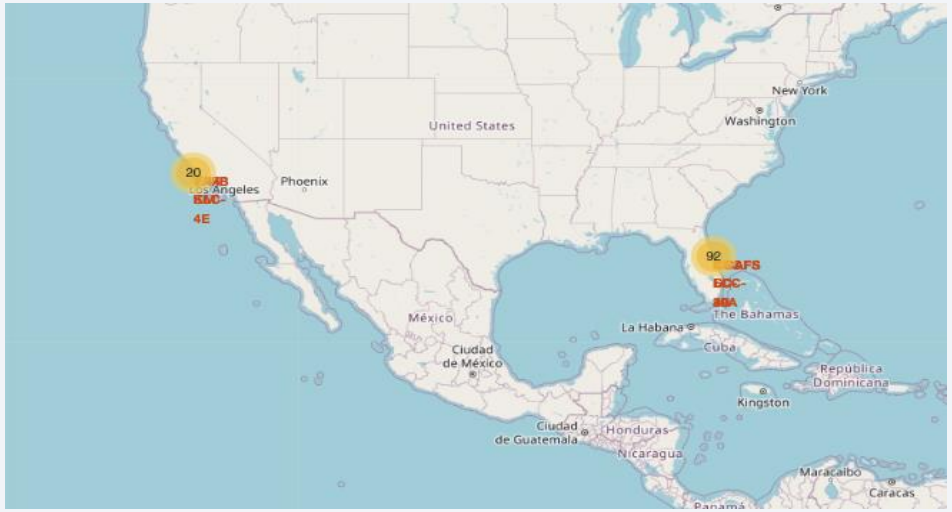
- Source code: https://github.com/Andre-Larc/Rocket-Launch/blob/master/Machine%20Learning%20Prediction_v2.ipynb

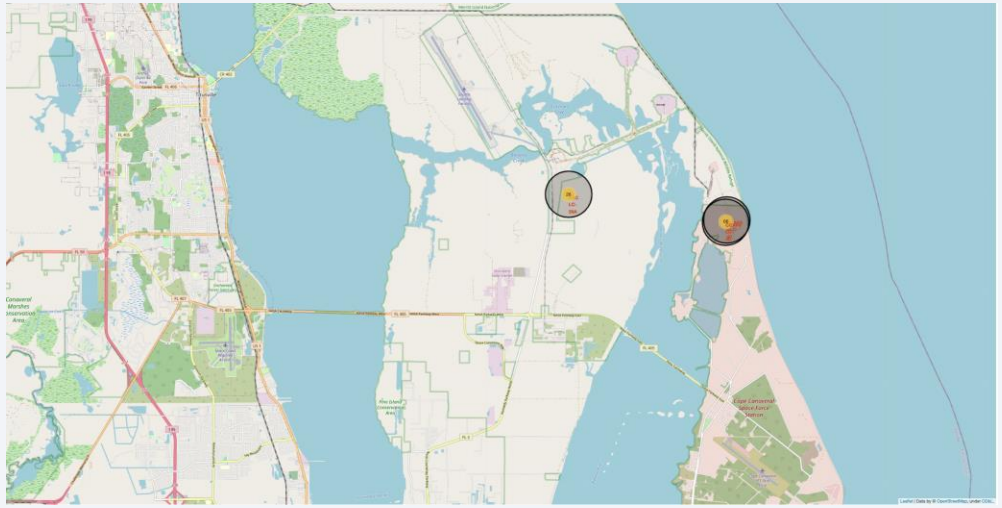
Results

- Exploratory data analysis results:
  - Space X uses 4 different launch sites;
  - The first launches were done for themselves and for NASA;
  - The average payload of F9 v1.1 booster is 2,928 kg;
  - The first successful landing only happened 5 years after SpaceX's first launch;
  - Many Falcon 9 booster versions were successful at landing on drone ships having payload above the average;
  - Two booster versions failed at landing on drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015;
  - The successful outcome rates at landing kept on increasing each year.

Results

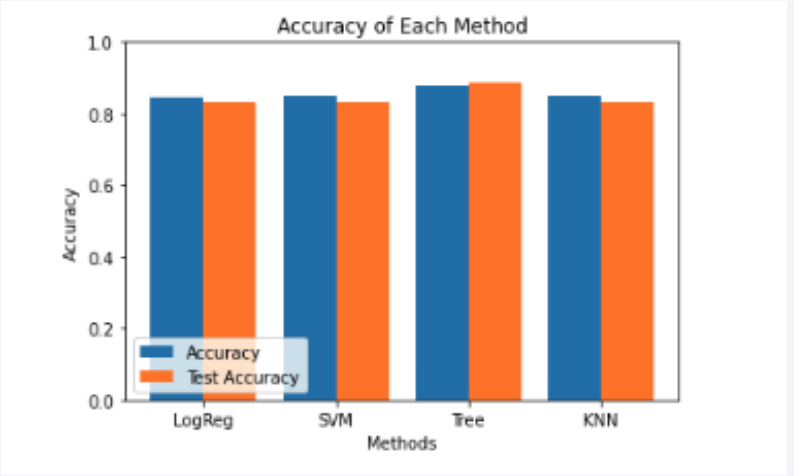
- With the support of interactive analysis on the map, it was concluded that all launch sites are geographically located in safe areas, near the coastline, and far enough from populated zones.
- The incidence of launches is greater on the East coast's launch sites.





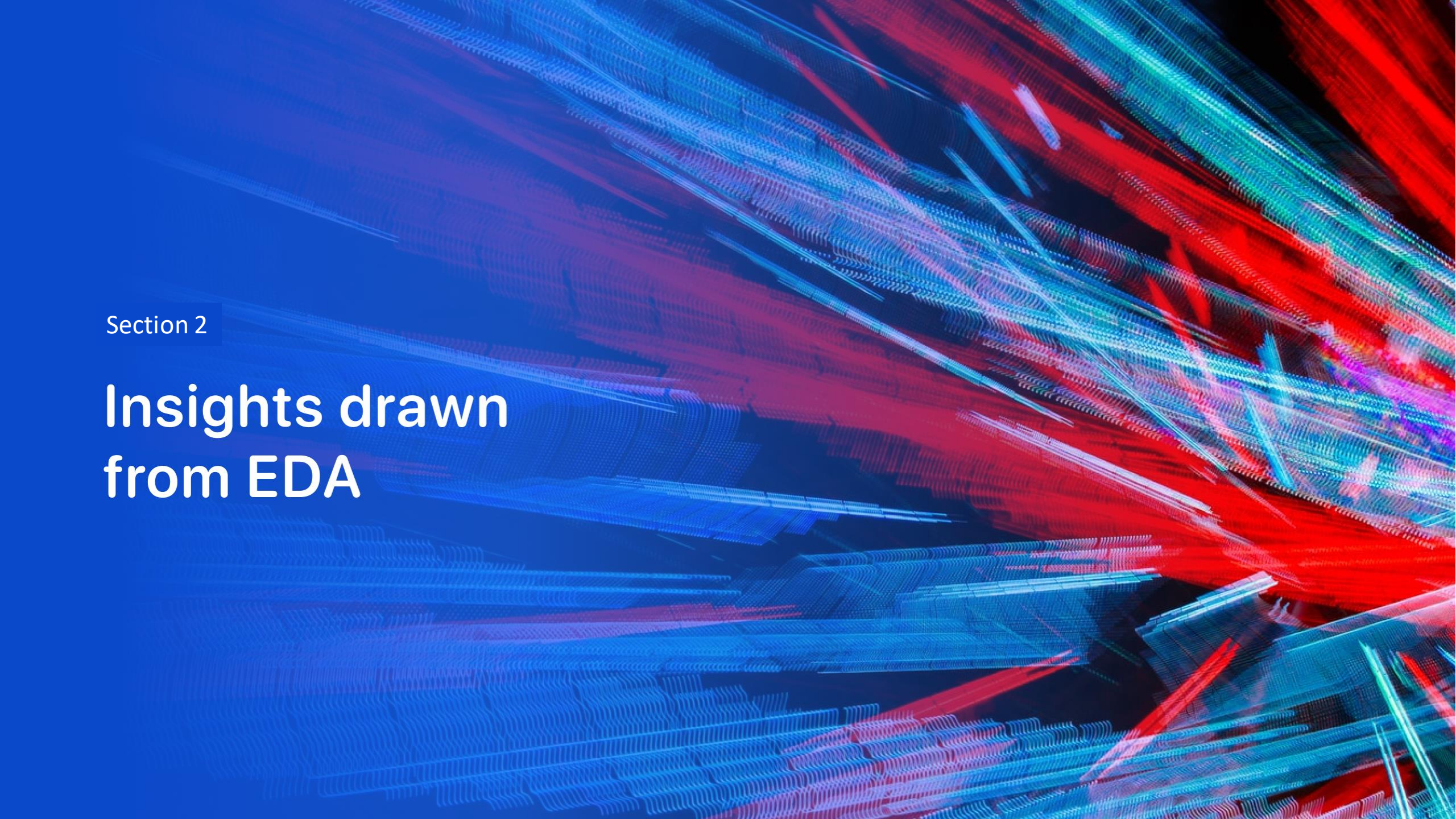
Results

- Predictive analysis shows that Decision Tree Classifier is the best model to predict successful landings, having an accuracy of 87,7% and accuracy for test data of 89%.

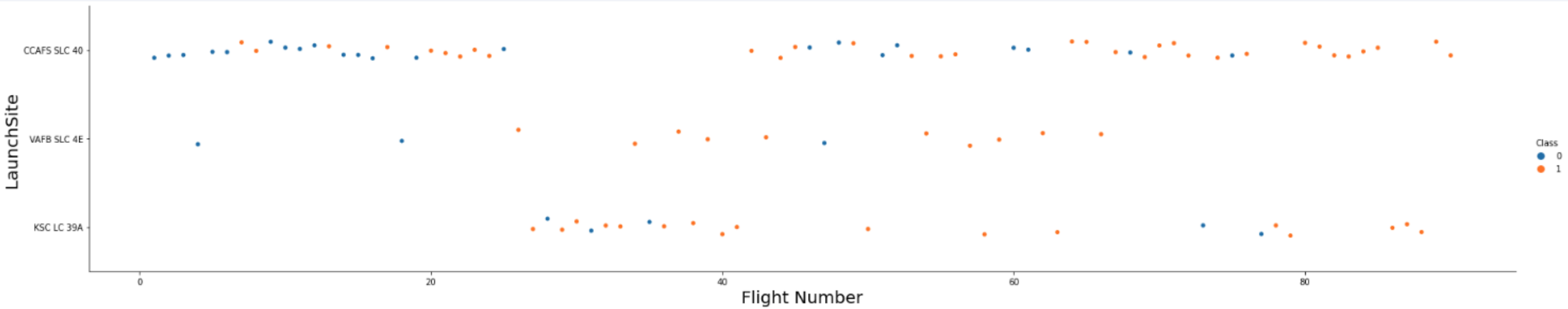


Section 2

Insights drawn from EDA

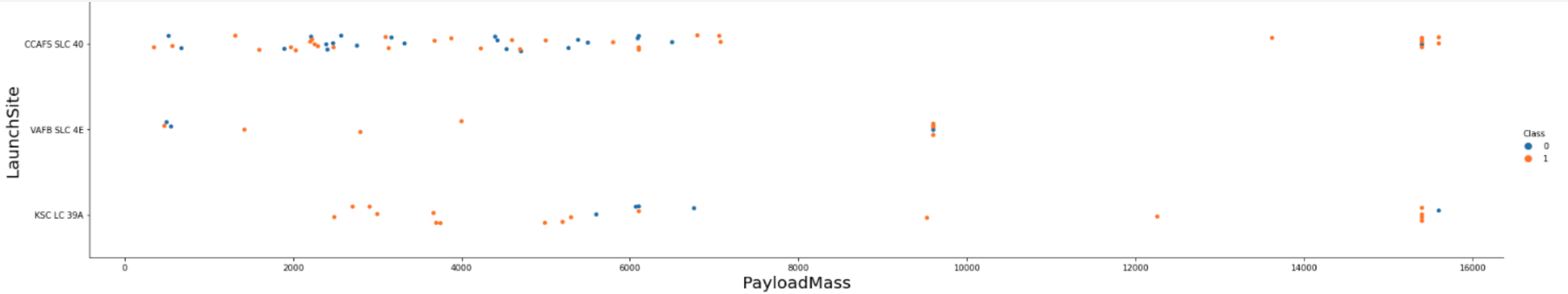


Flight Number vs. Launch Site



- Analyzing the scatter plot, we can observe that the best launch site nowadays is CCAFS SLC 40, which has a high success rate on most of recent launches;
- We can also observe that the success rate increased in general over time.

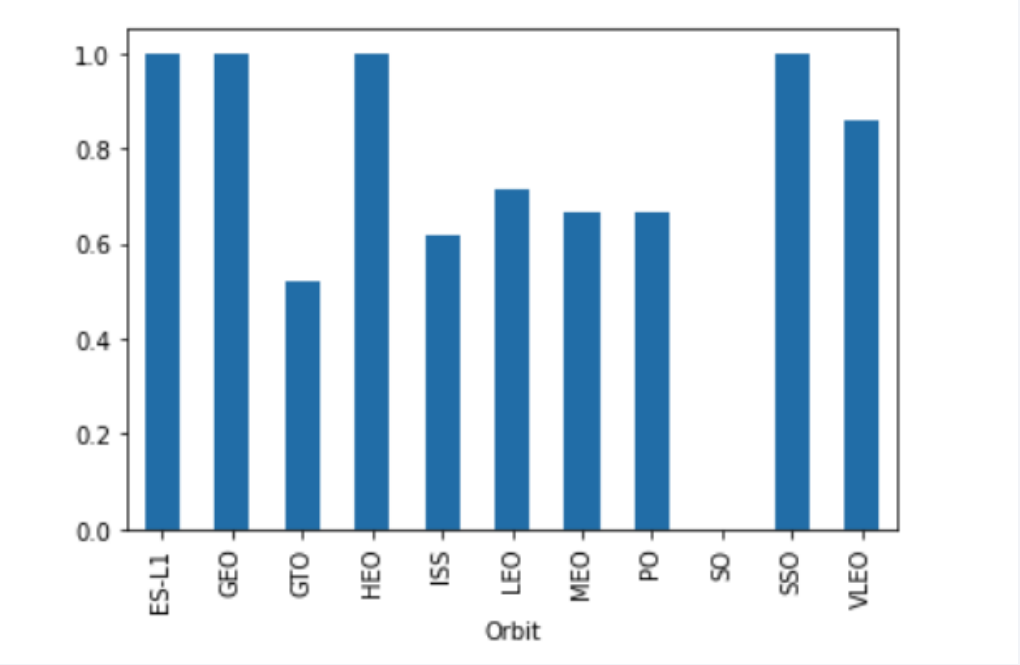
Payload vs. Launch Site



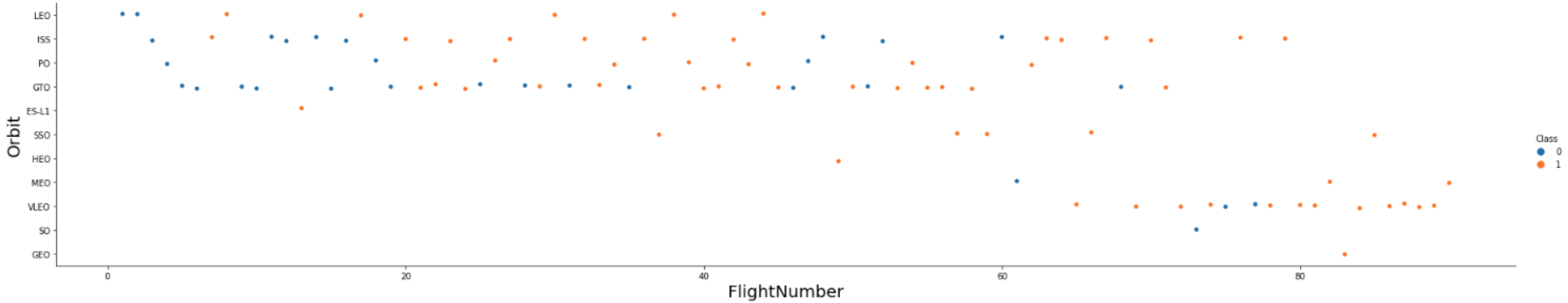
- Analyzing the scatterplot above we can observe that the high payload mass launches from both KSC LC 39A and CCAFS SLC 40 have a high success rate;
- Also, launches from KSC LC 39A had a good success rate with payload under around 6,000 Kg;
- No launches with Payload above 10000 Kg seem to be observed from VAFB SLC 4E.

Success Rate vs. Orbit Type

- Highest success rate:
  - ES-L1
  - GEO
  - HEO
  - SSO
  - VLEO
- Lowest success rate:
  - GTO
  - ISS

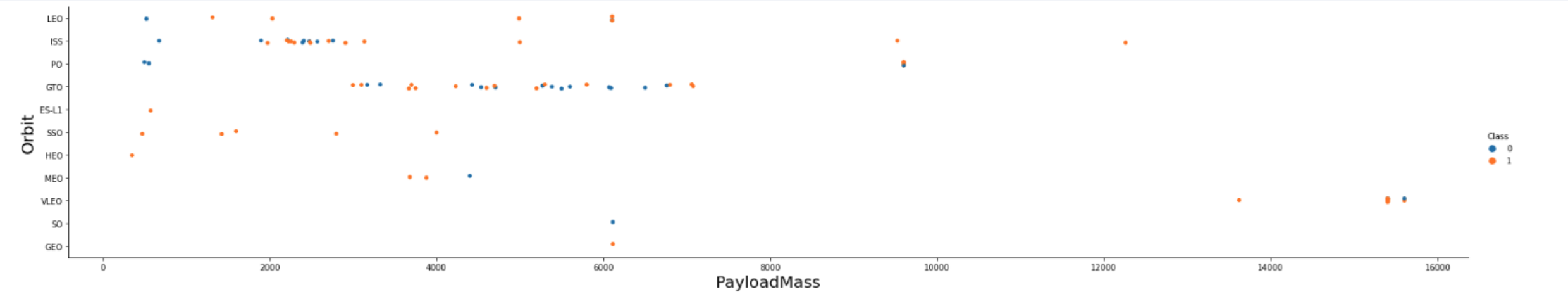


Flight Number vs. Orbit Type



- Most recent launches went for VLEO having a high success rate;
- GTO had the lowest success rate, starting to increase after flight N.50;
- Generally, the success rate increased over time.

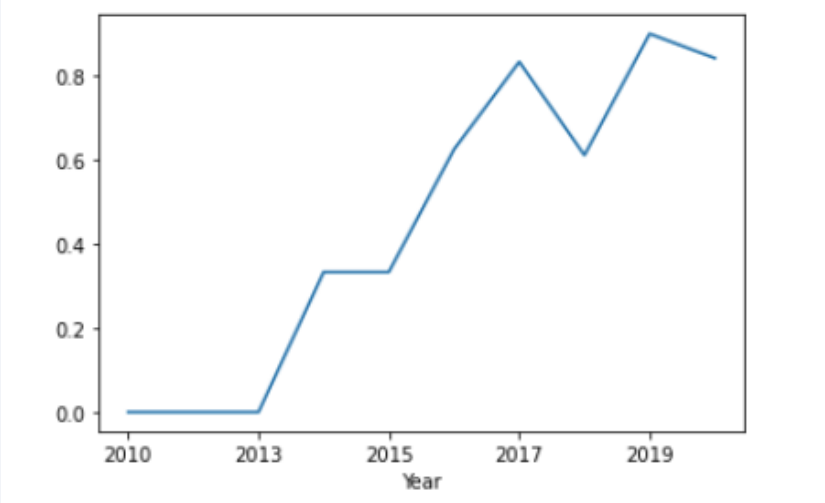
Payload vs. Orbit Type



- Hard to draw any conclusion from orbit GTO in relation to Payload mass;
- App, all launches to orbit SSO had a good success rate for Payload lower than 4000 Kg;
- All launches to ISS and LEO with Payload above 5000 Kg were successful.

Launch Success Yearly Trend

- After 2013, the success rate kept on rising with time;
- There was a drop of about 20% on the success rate during the year of 2017. Then starting to rise again during the year of 2018.



All Launch Site Names

- There are four launch sites Space X uses, and the following table shows their unique names:

- We query the data using SQL to find the unique values on the column 'launch_site'.

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

Launch Site Names Begin with 'CCA'

- We queried the data in order to find 5 records where launch sites begin with `CCA`, representing launches from Cape Canaveral:

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

- Total payload carried by boosters from NASA was of 111268 Kg:

total_payload
111268

- By filtering the Payload mass column by the code 'CRS', which represents boosters from NASA, we can get the sum of all their Payload mass.

Average Payload Mass by F9 v1.1

- The Average Payload Mass carried by booster version F9 v1.1 was of 2928 Kg:

avg_payload
2928

- We also filtered the data, selecting only the booster version F9 v1.1 in order to calculate the Average Payload Mass.

First Successful Ground Landing Date

- The first successful landing outcome on ground pad happened on December 22nd of 2015:

first_landing
2015-12-22

- We selected only the successful landing outcomes by filtering the data, and by getting the minimum value in the DATE column we found the earliest successful landing outcome.

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:

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

- By filtering the data we can select the unique booster versions with Payload Mass between 4000 Kg and 6000 Kg which also had a Successful Landing Outcome on Drone ship.

Total Number of Successful and Failure Mission Outcomes

- Total number of "Success" and "Failure" mission outcomes:

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

- By grouping by Mission Outcome and counting the occurrences in each group we found the numbers shown above.

Boosters Carried Maximum Payload

- Boosters which have carried

  the maximum Payload Mass:

- By filtering the data we acquired the unique booster versions with the maximum values in 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

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

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

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

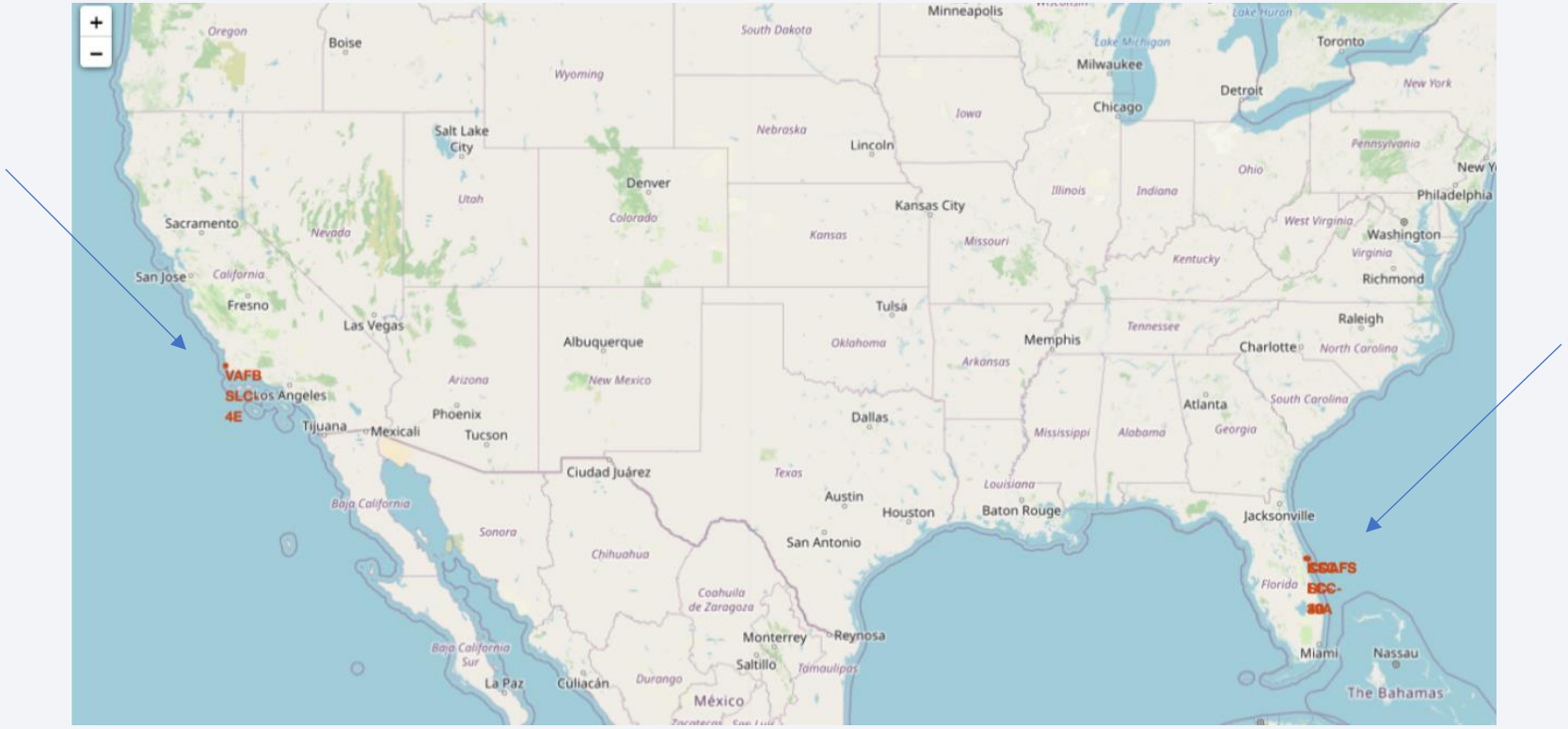
- Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20:

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

- By counting each quantity and grouping by Landing outcome we can analyzing that a big portion of launches during the selected period didn't even attempt to land.


Section 3
Launch Sites
Proximities Analysis

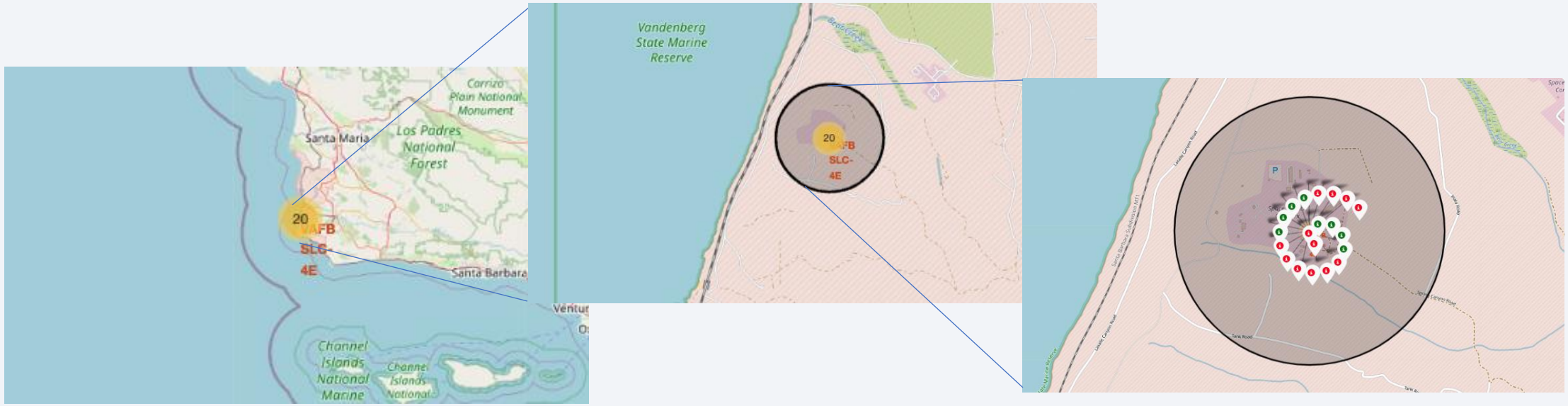
All Launch sites shown on the map



- All Launch sites are located near the coast on both East and West sides of the country. They also tend to be located towards the south being as close to the Equator line as possible inside the US territory.

Launch outcomes on the map

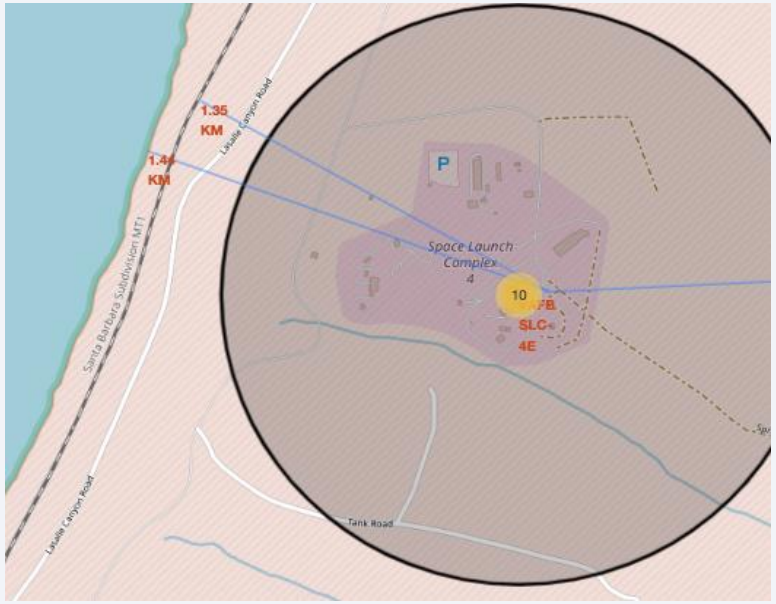
- Analyzing launch site 'VAFB SLC-4E' for example:

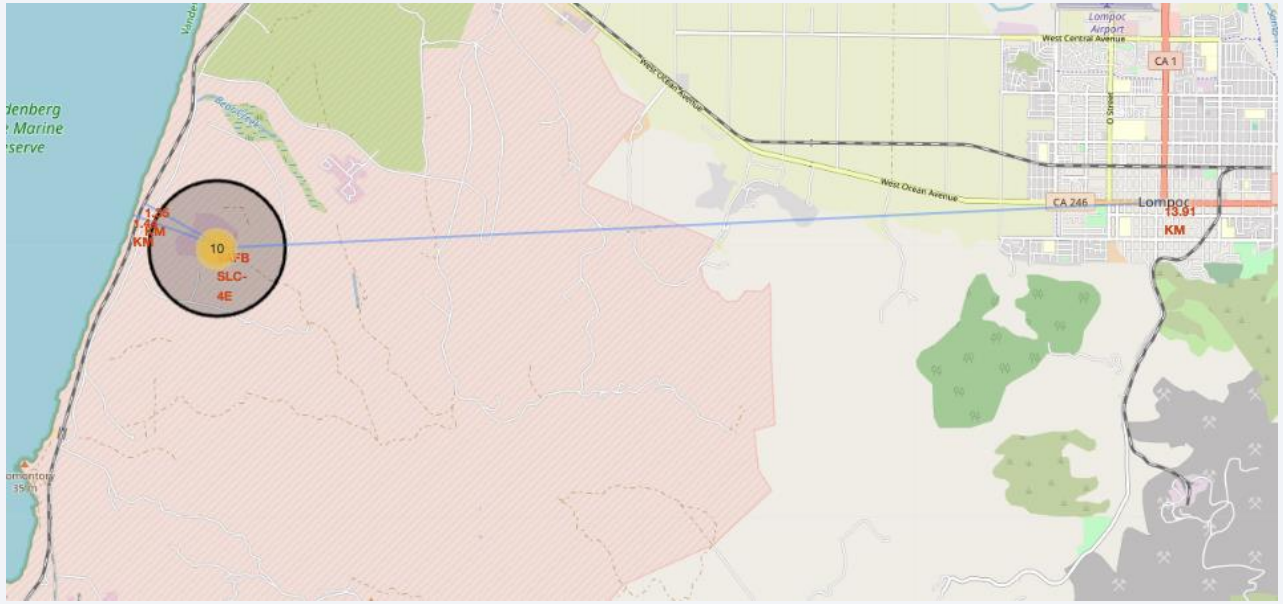


- We added markers with color codes to visually represent the launch outcome of each launch per location. You can see the green and red markers representing successful and failure outcomes respectively.

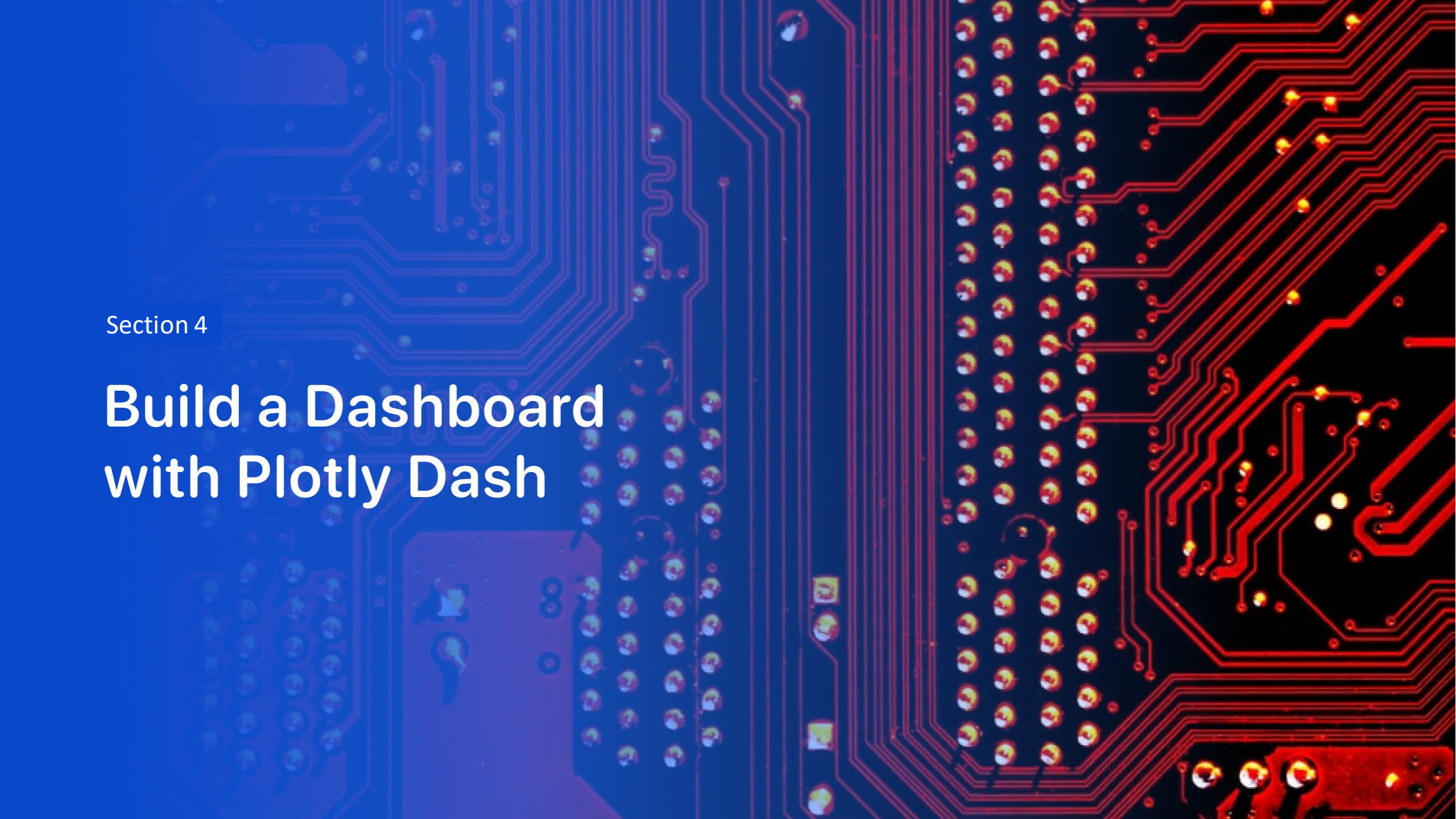
Safety and Logistics

- Still analyzing launch site 'VAFB SLC-4E' and its surroundings :



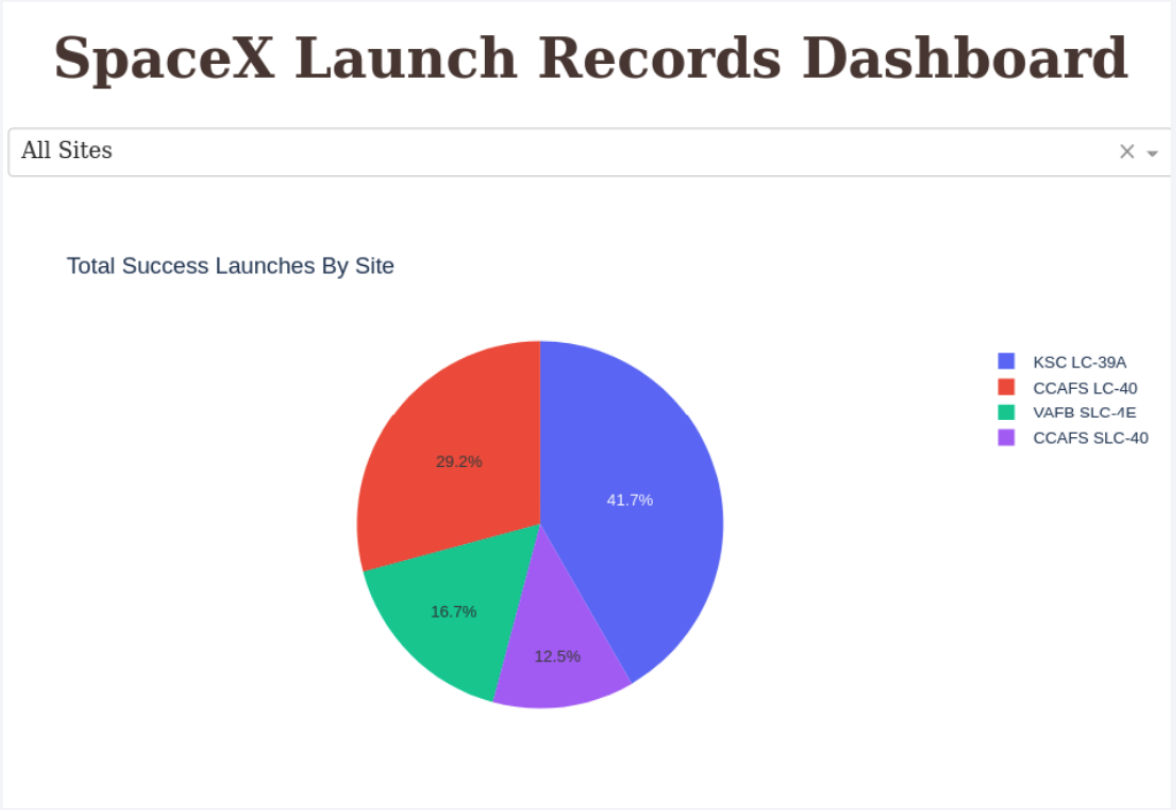


- We found and marked on the map the distances from the launch site to the nearest coastline (1.44 Km) and to the nearest town (13.91 Km). Meaning that the launch site is safely located near the coastline and far enough from any populated area. And close enough to some roads and railroads (Santa Barbara Subdivision MT1) for better logistics.

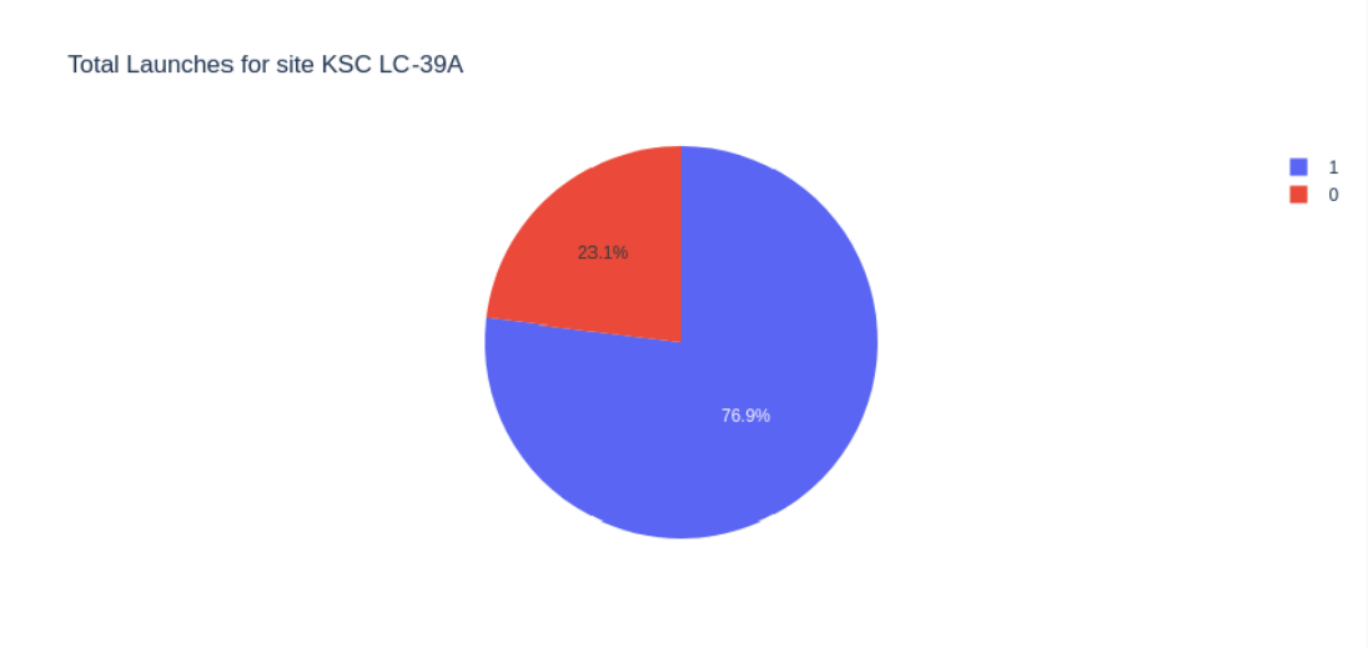

Section 4
Build a Dashboard with Plotly Dash

Successful Launches by Launch site

- We created an interactive dashboard, and analyzing the generated piechart we can see that the Launch site choice seems to be considerably relevant for the mission's outcome.

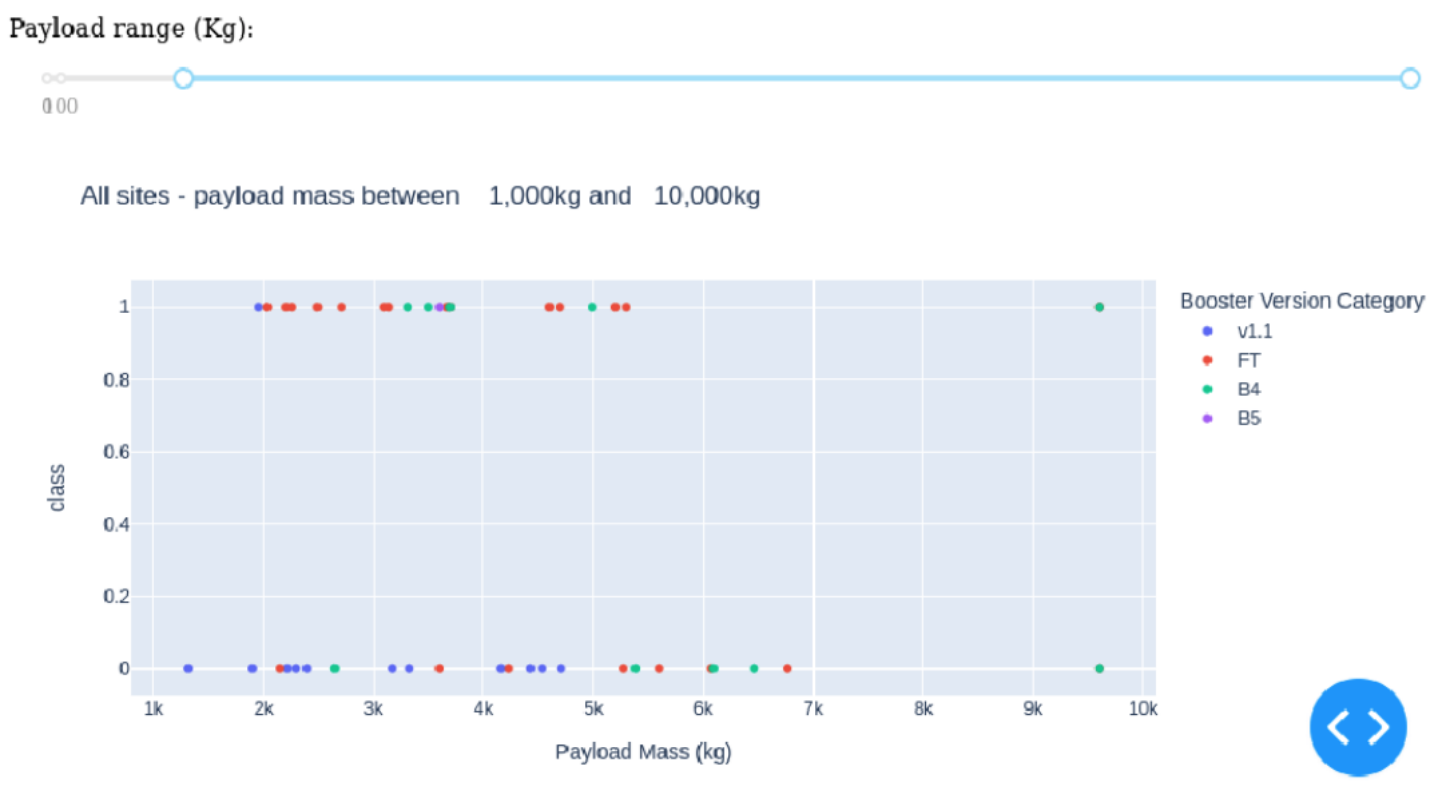


Launch Success Ratio - Piechart



- Interacting with our dashboard for the launch success ratio for each launch site, we found that the one with the highest success ratio is 'KSC LC-39A' with 76.9% successful launches.

Payload vs. Launch Outcome



- Analyzing our generated scatterplot, we can see that payloads under 6,000 Kg combined with the Booster version FT are the most successful.

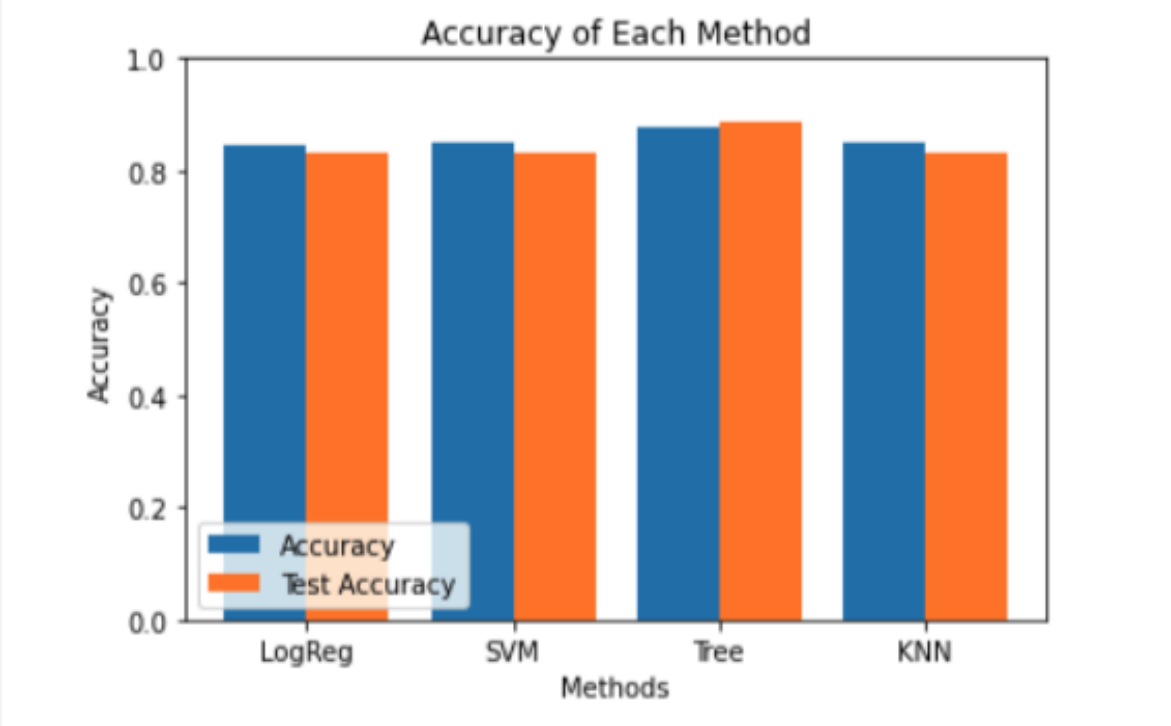
Section 5

Predictive Analysis (Classification)

Classification Accuracy

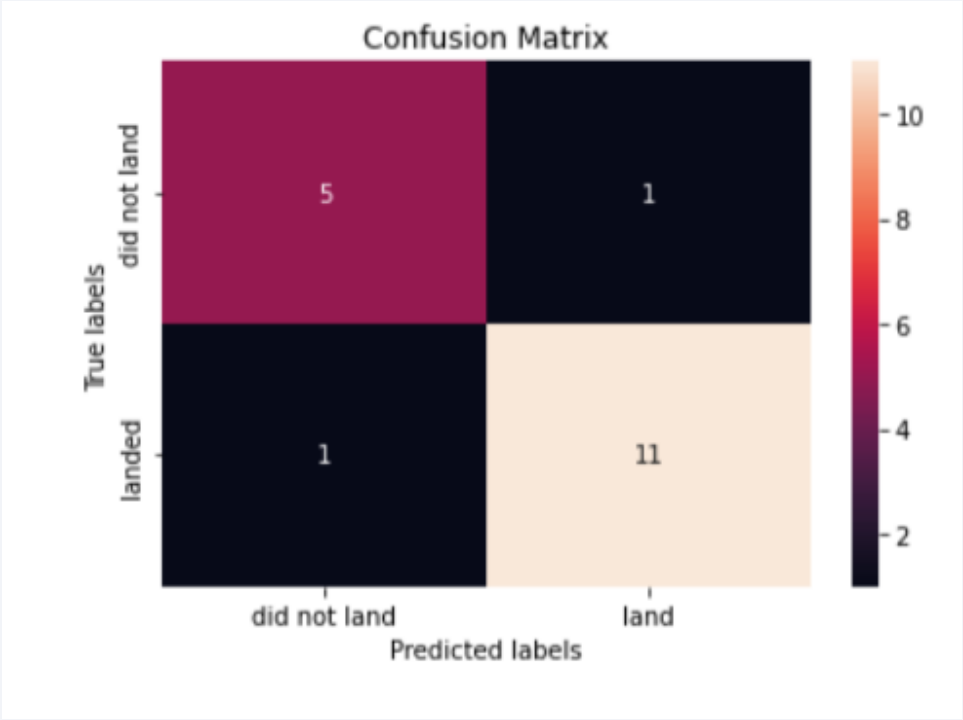
- We tested the accuracy of 4 classification models (LogReg, SVM, DecisionTree and KNN) which is visually represented on the right.

- Comparing all 4 models, we found that The Decision Tree model is the one with the highest accuracy over 87%.



Confusion Matrix

- To confirm that, we plot a Confusion Matrix and as we can see by the numbers in the True Positive and True Negative blocks are high, compared to the numbers in the False Positive and False Negative blocks which are very low.



Conclusions

- The best predictive algorithm or this dataset is the Decision Tree Model;
- Launch sites tend to be located near the coast and not far from the Equator line;
- Launches with low Payload Mass show better results than launches with larger Payload Mass;
- KSC LC-39A has the highest launch success rate compared to all other Launch sites;
- The success rate at all Launch sites increased over the years;
- Launches to the orbits ES-L1, SSO, GEO and HEO have the best success rate.

Thank you!