IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

João Tomazetti
04/08/2022

Outline

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

Executive Summary

- Methodologies:
  - Webscraping;
  - Data Analysis;
  - Machine Learning;

- Summary of the results:
  - Data analysis was fundamental for the identification of the strong features;
  - Machine Learning presented the best model to deal with the situation;

Introduction

- The project background was to analyze the space-X data for a new company, giving the necessary data for a new launch;
- Problems you want to find answers:
  - Where to launch;
  - The cost;
  - Wich stage/orbit gives the best success rate;

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - *Data was collected from the following sources:*
    - *Space X API;*
    - *Webscrap from Wikipedia;*
- Perform data wrangling
  - *Organized and improves data with pandas;*
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash

Data Collection

- Perform predictive analysis using classification models
  - Data was normalized and divided between test and train sets, and then teste with predictive models;

Data Collection – SpaceX API

- SpaceX public API was used, and cleansed so we could have only the necessary data;

- GitHub URL of the completed SpaceX API calls notebook: https://github.com/Alpha-jor/IBM-Capstone-Project/blob/main/Collecting%20Data.ipynb

Data Collection - Scraping

- More data is obtained through Wikipedia, and then webscraped with python and normalized with pandas;

- GitHub URL of the completed web scraping notebook: https://github.com/Alpha-jor/IBM-Capstone-Project/blob/main/Web%20Scraping.ipynb

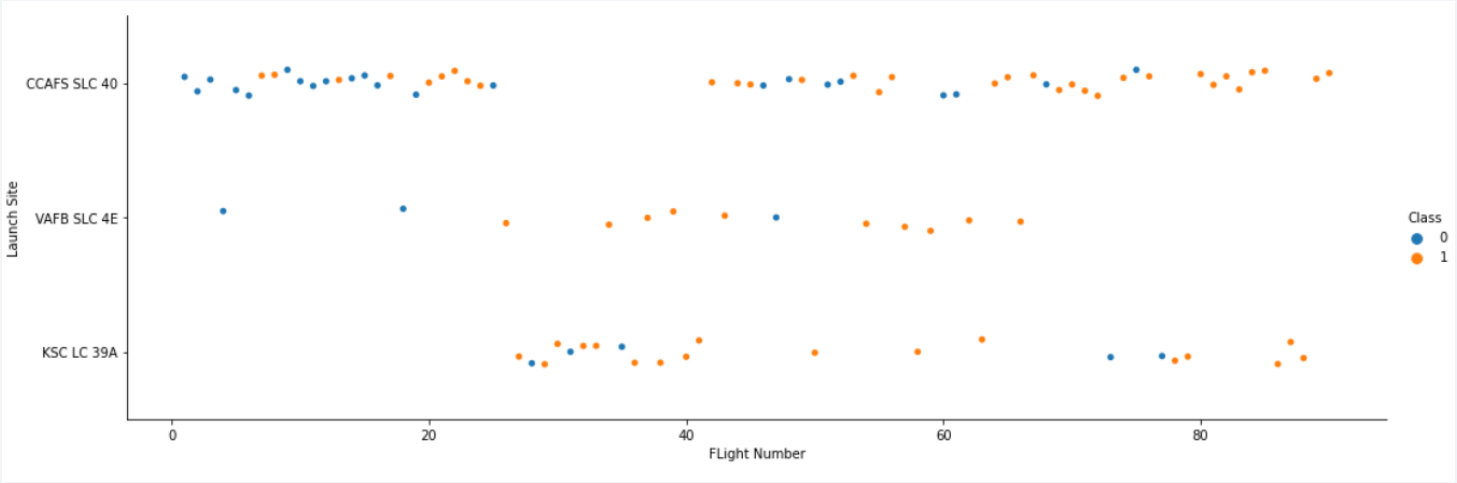
Data Wrangling

- Data that was of the most interest were further analyzed, like orbit frequency and the result of the mission (success/failure)

- GitHub URL of your completed data wrangling related notebooks: https://github.com/Alpha-jor/IBM-Capstone-Project/blob/main/Wrangling%20EDA.ipynb

EDA with Data Visualization

- Scatter and bar plots were created with the following features:
  - Payload Mass vs Flight Number;
  - Launch Site vs Flight Number;
  - Launch Site vs Flight Number;
  - Launch Site vs Payload Mass;
  - Orbit vs Flight Number;
  - Orbit vs Payload Mass;



- https://github.com/Alpha-jor/IBM-Capstone-Project/blob/main/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- SQL queries performed:
  - Unique launch sites in the space mission;
  - 5 records where the launch site beginned with CCA;
  - Total payload mass carried by boosters launched by NASA;
  - 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 greater than 4000 but less than 6000;
  - 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 the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order;

https://github.com/Alpha-jor/IBM-Capstone-Project/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Map objects used:
  - Markers;
  - Circles;
  - Lines;
  - Marker clusters;
- Those were used to indicate points in the map, highlight areas, show the distance between two locations and group events;

https://github.com/Alpha-jor/IBM-Capstone-Project/blob/main/Interactive%20Visuals%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- Plots/graphs and interactions added to the dashboard:
    - Range of the Payload;
    - Percentage of launches in each site
- They allowed the visualization of which site has a better performance;

https://github.com/Alpha-jor/IBM-Capstone-Project/blob/main/dash_app.py

Predictive Analysis (Classification)

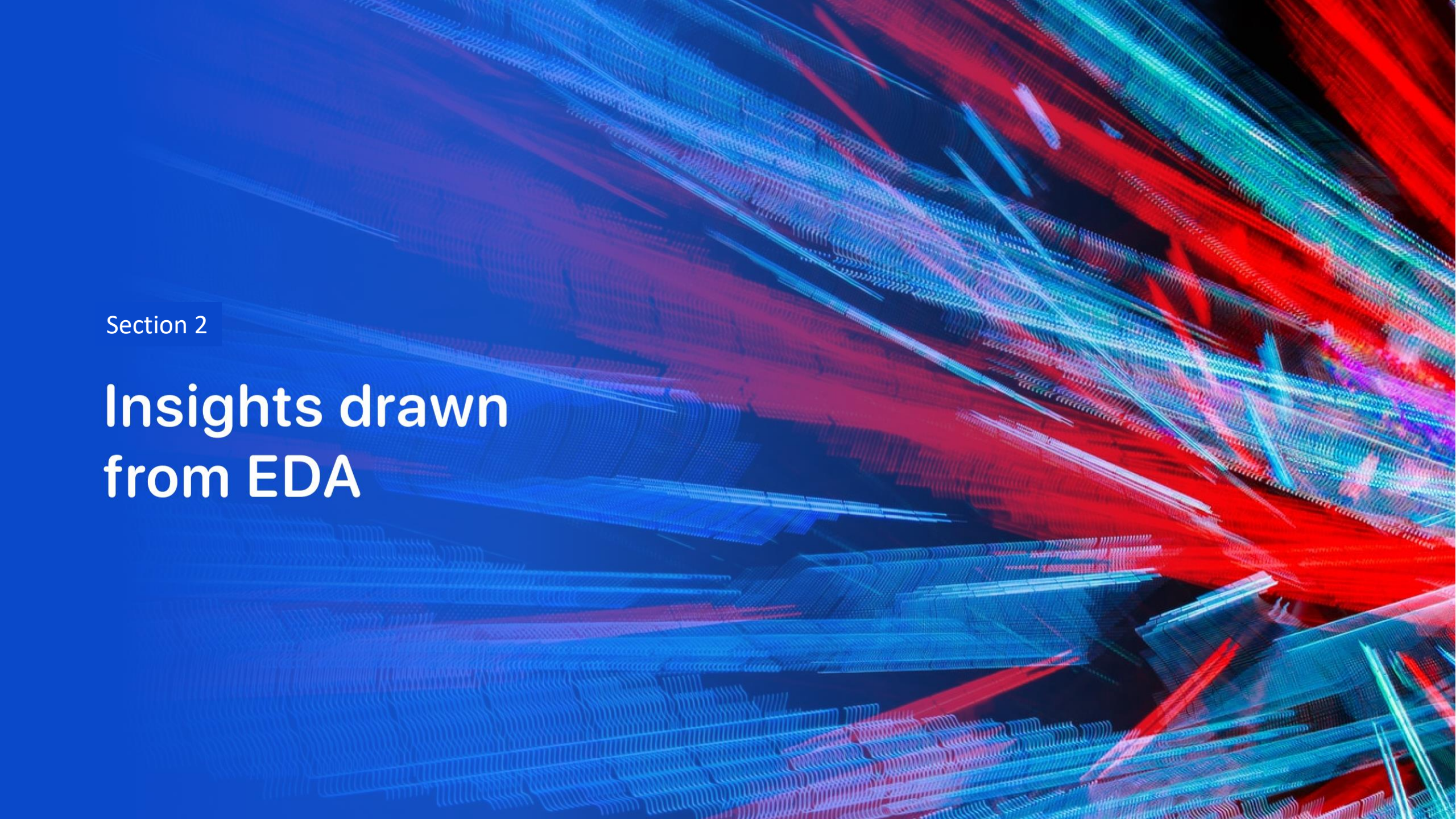
- Models were built through data standardization;
- 4 models tested for accuracy (logistic regression, KNN, decision tree and SVM);
- Results were compared;

https://github.com/Alpha-jor/IBM-Capstone-Project/blob/main/Machine%20Learning.ipynb

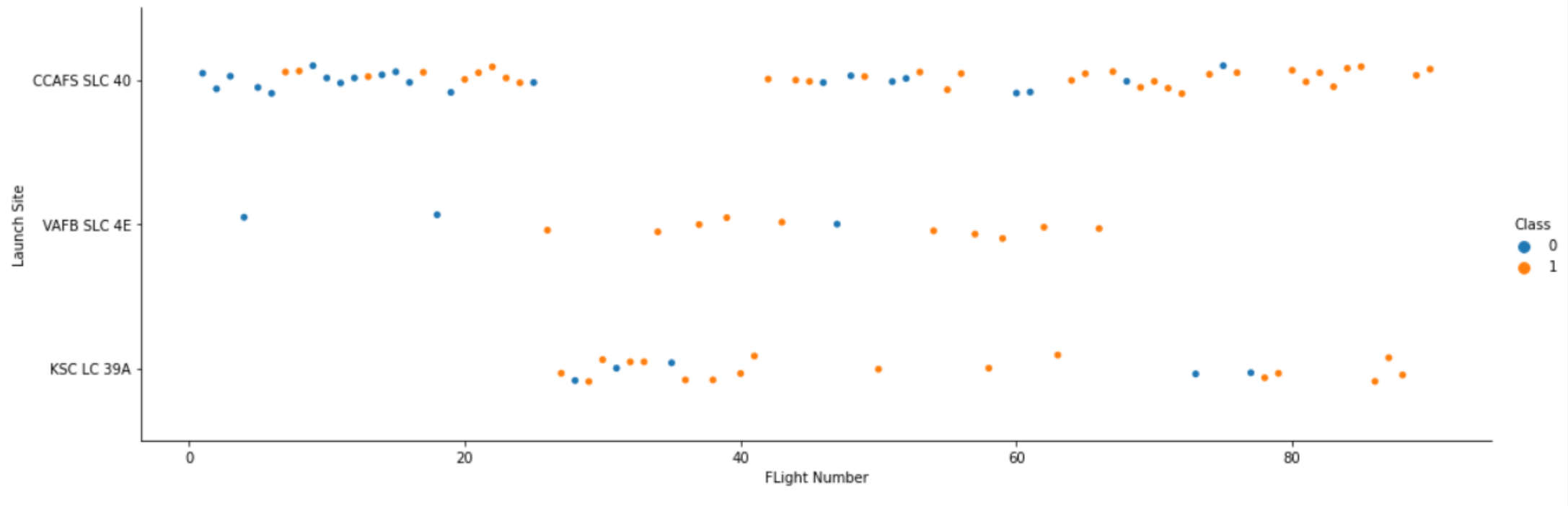
Results

- Exploratory data analysis results
  - Nearly 100% of the missions were successful;
  - The outcomes improved as the years passed;
- Predictive analysis results

Model	Accuracy	TestAccuracy
LogReg	0.84643	0.83333
SVM	0.84821	0.83333
Tree	0.875	0.83333
KNN	0.84821	0.83333

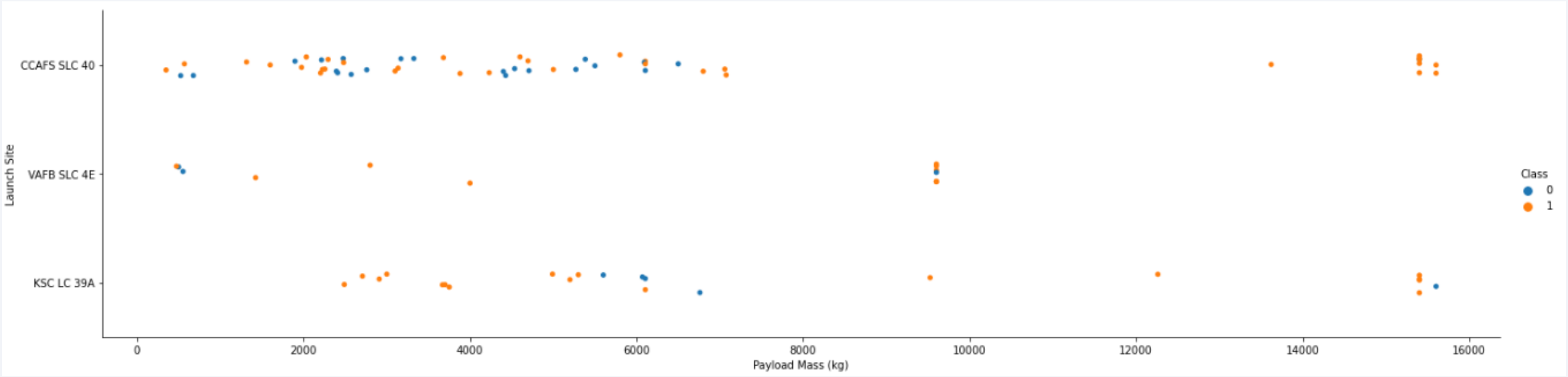

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



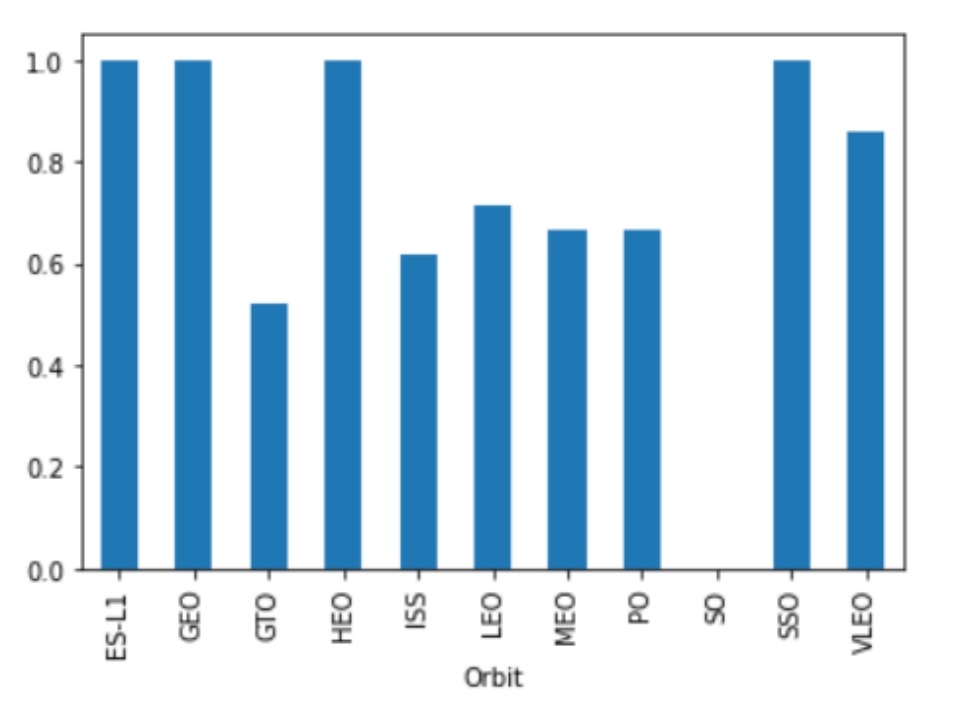
In a general way, the success rate increased with the flight number;

Payload vs. Launch Site



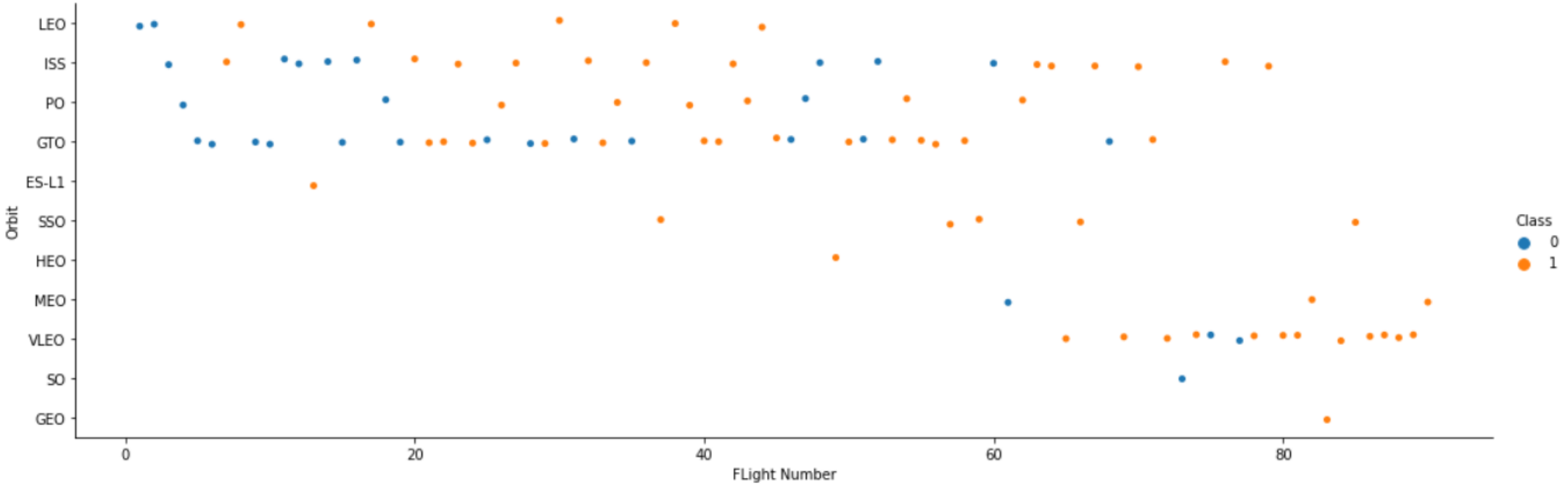
- There were no heavy payload mass for the VAFB SLC 4E;

Success Rate vs. Orbit Type



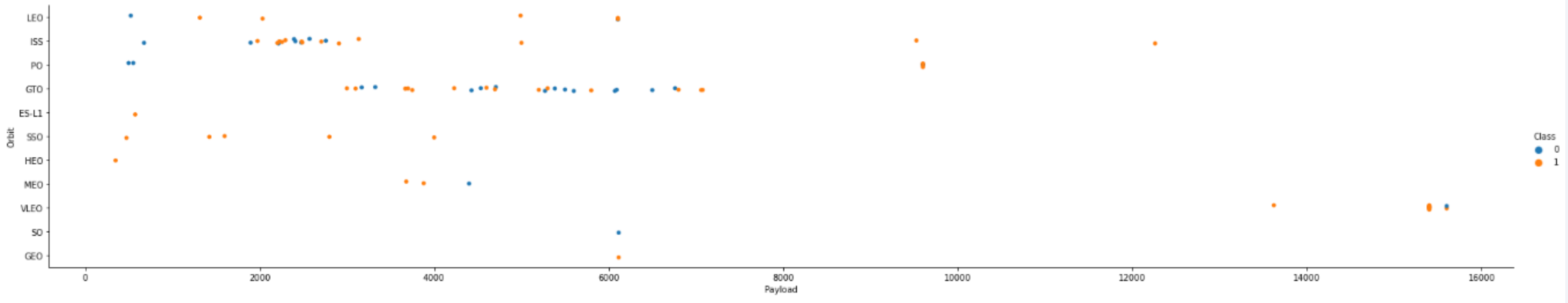
- The higher success rates were from ES-L1, GEO, HEO and SS0;
- The worst success rates were from GTO, ISS and SO;

Flight Number vs. Orbit Type



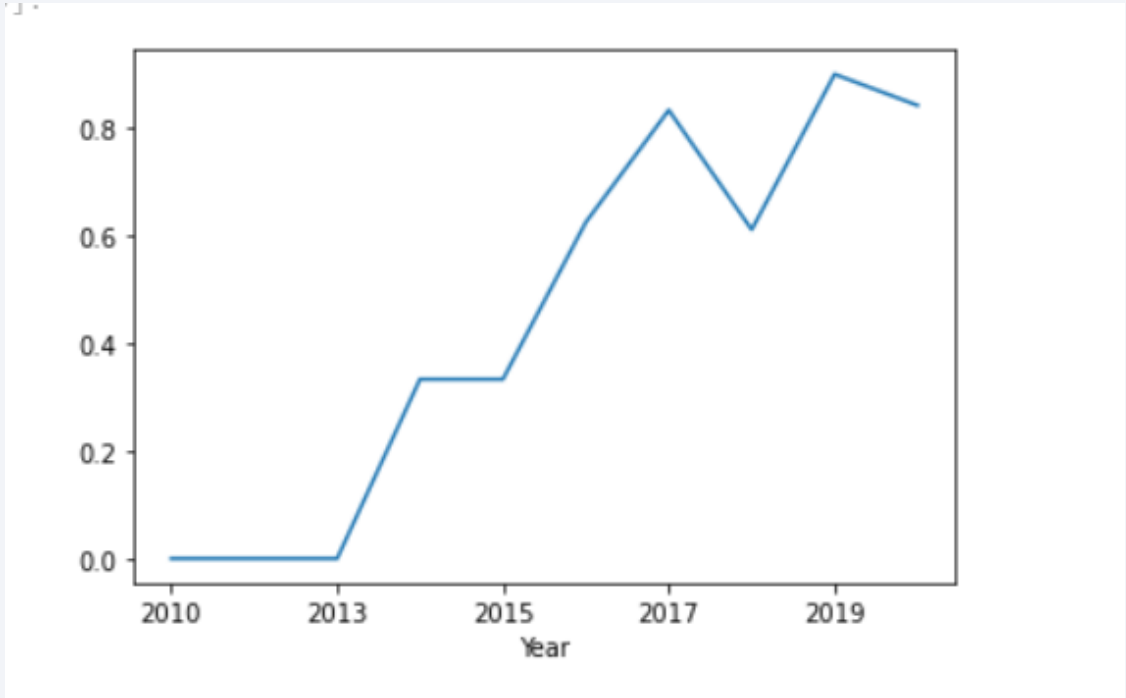
- For the most part the success rate increased with the flight numbers, execpt for GTO, with a weak correlation;

Payload vs. Orbit Type



- We can see certain increase in success rate over flight numbers on LEO and ISS orbits

Launch Success Yearly Trend



- The success rate kept increasing from 2013 until 2019

All Launch Site Names

- These are all the four launch sites:

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

- They were obtained with the following query:

```
sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
```

Launch Site Names Begin with 'CCA'

- These are the first 5 records which launch site names begin with ‘CAA’:

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

- They were obtained with the following query:

```
sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

Total Payload Mass

- These is the total payload mass by boosters launched by NASA:

total
111268

- It was obtained with the following query:

```
sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL FROM SPACEXTBL WHERE PAYLOAD LIKE '%CRS%';
```

Average Payload Mass by F9 v1.1

- These is the average payload mass by F9 v1.1 boosters:

average
2928

- It was obtained with the following query:

```
sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVERAGE FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1'
```

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad:

first_success
2015-12-22

- It was obtained with the following query:

```
sql SELECT MIN(DATE) AS FIRST_SUCCESS FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Success (ground pad)';
```

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

- They were obtained with the following query:

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

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

- It was obtained with the following query:

```
sql SELECT MISSION_OUTCOME, COUNT(*) AS QUANTITY FROM SPACEXTBL GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;
```

Boosters Carried Maximum Payload

- The names of the boosters which have carried the maximum payload mass:
- They were obtained with the following query:

```
sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL) ORDER BY BOOSTER_VERSION;
```

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

- They were obtained with the following query:

```
sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND DATE_PART('YEAR', DATE) = 2015;
```

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

- 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:

landing__outcome	quantidade
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

- It was obtained with the following query:

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

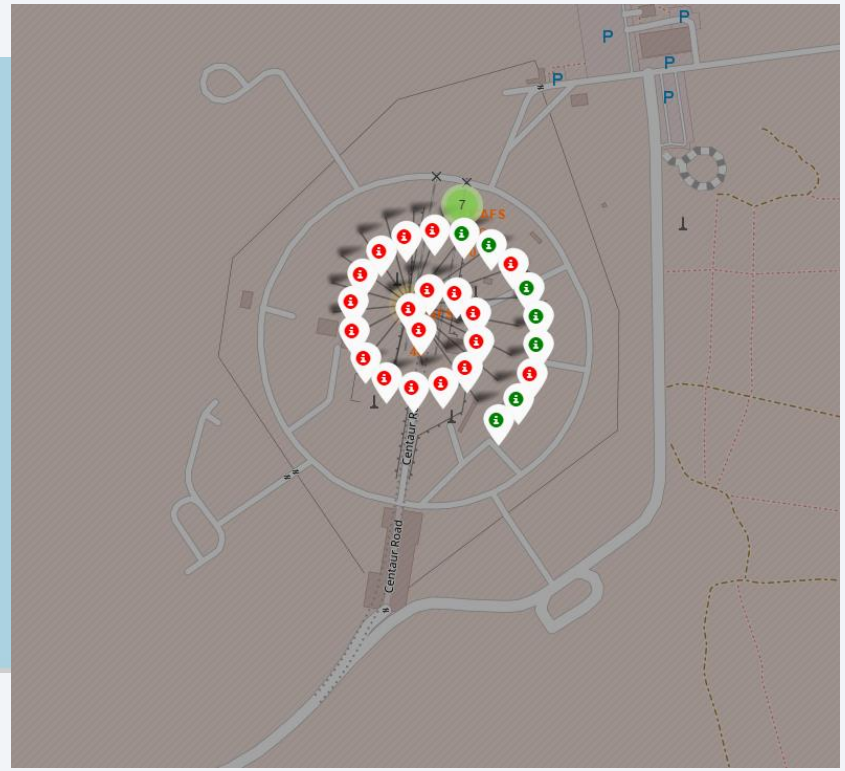
Section 3

Launch Sites Proximities Analysis

Launch Sites Locations

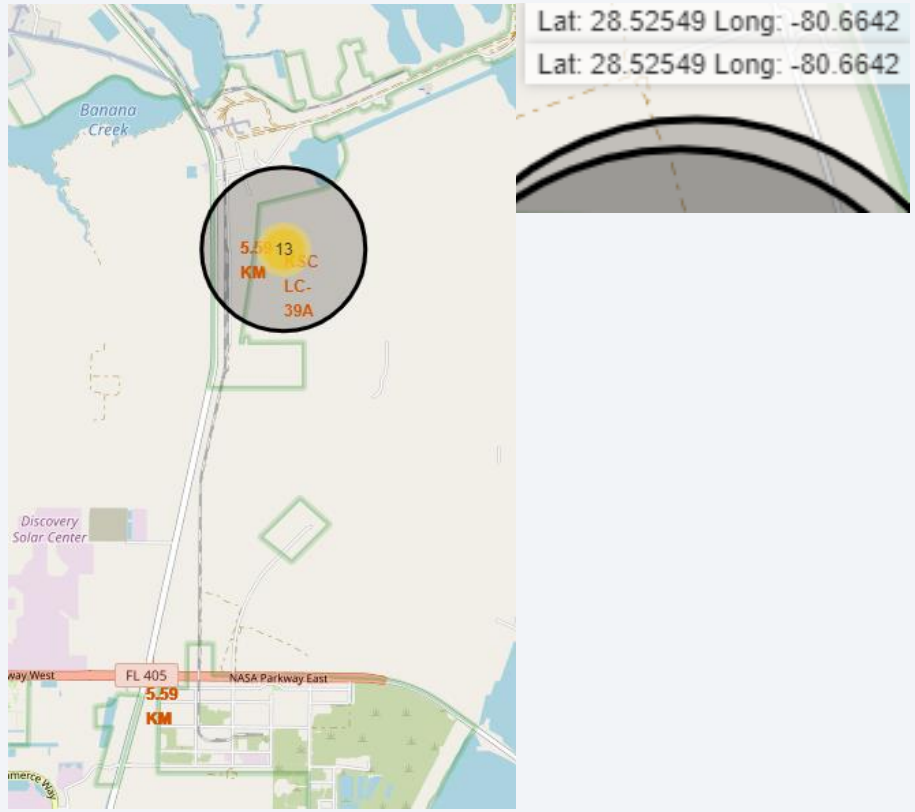
All launch sites are located on the coast, so it doesnt affect people on the cities.

Outcomes by Launch Sites (Color Markers)

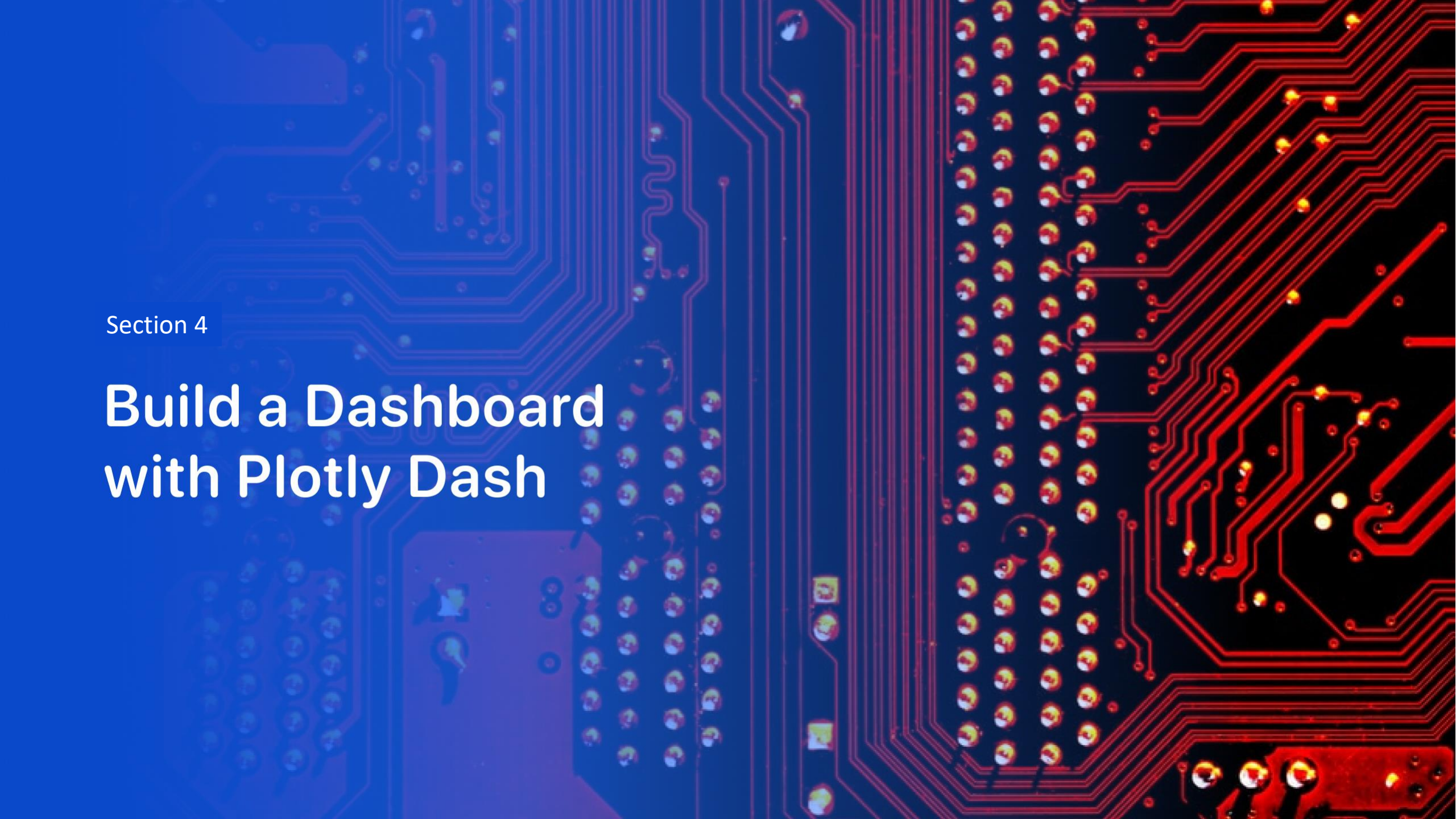


Green markers show that the outcome was succesfull

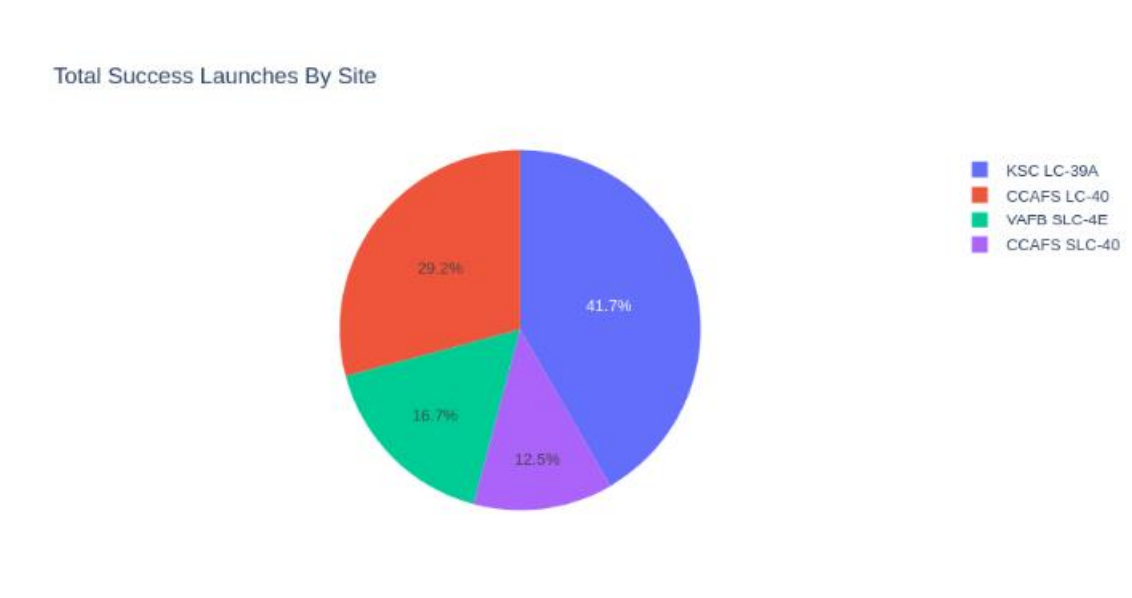
<Folium Map Screenshot 3>



The KSC LC-39A is 5.59 kilometers close to a railroad, while being far from buildings.

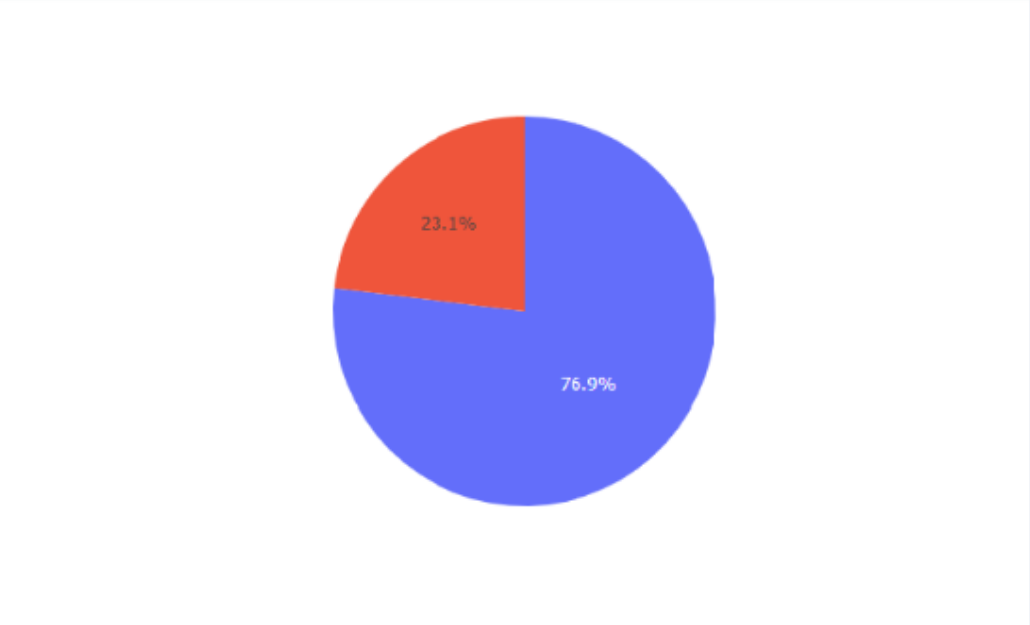

Section 4
Build a Dashboard with Plotly Dash

Success Percentage by Launch Site



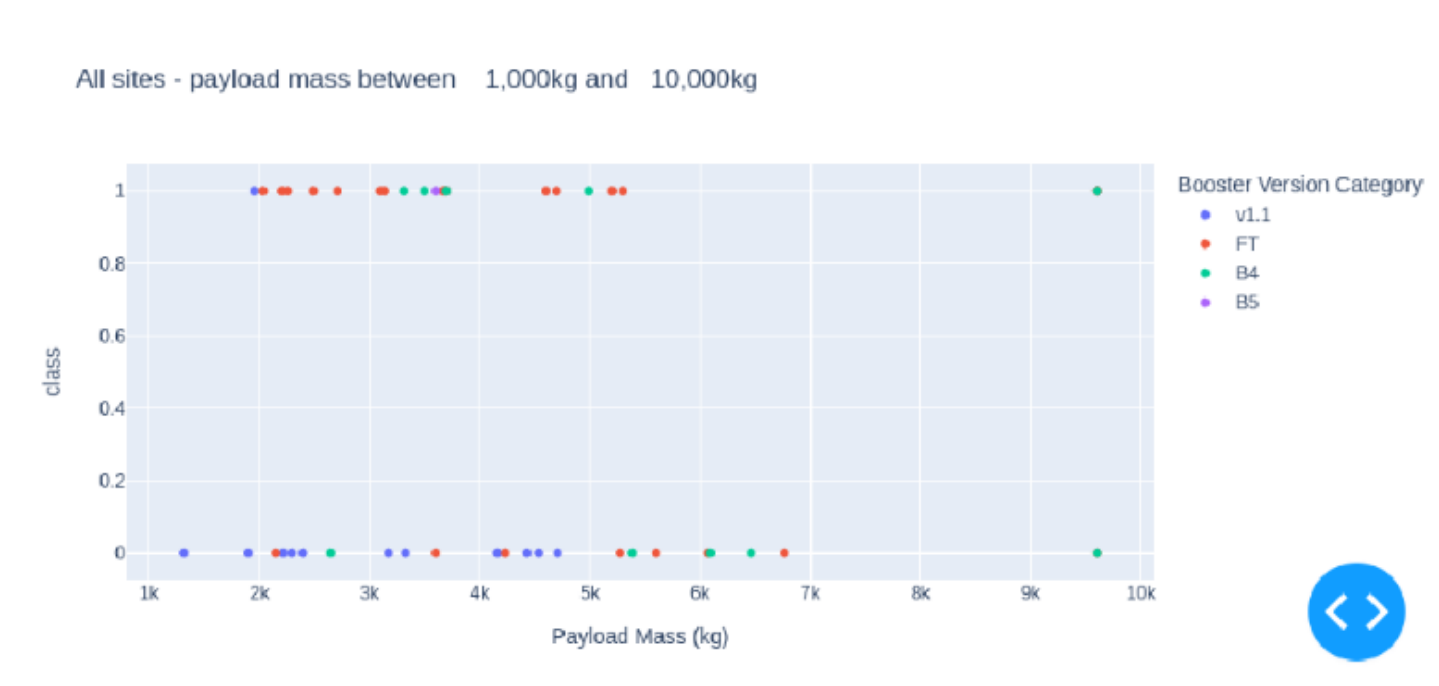
The highest success rate belongs to the KSC LC-39A site

Success Rate for the KSC LC-39A Site



This shows the success rate for the KSC LC-39A site, blue being the successful outcomes accounting for 76.9%

<Dashboard Screenshot 3>



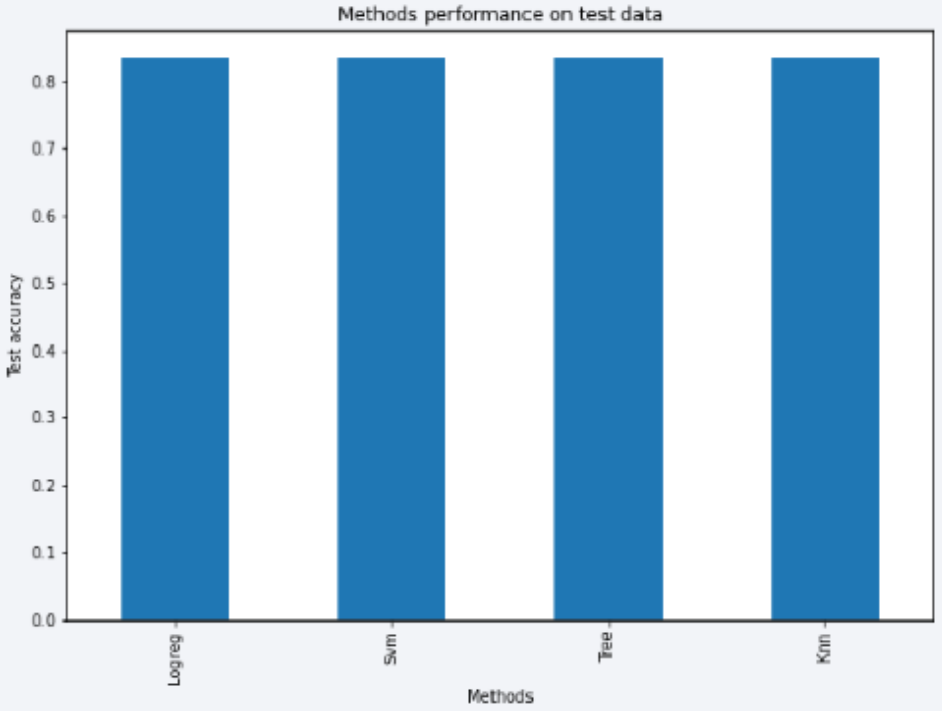
Smaller payload masses have a higher success rate

Section 5

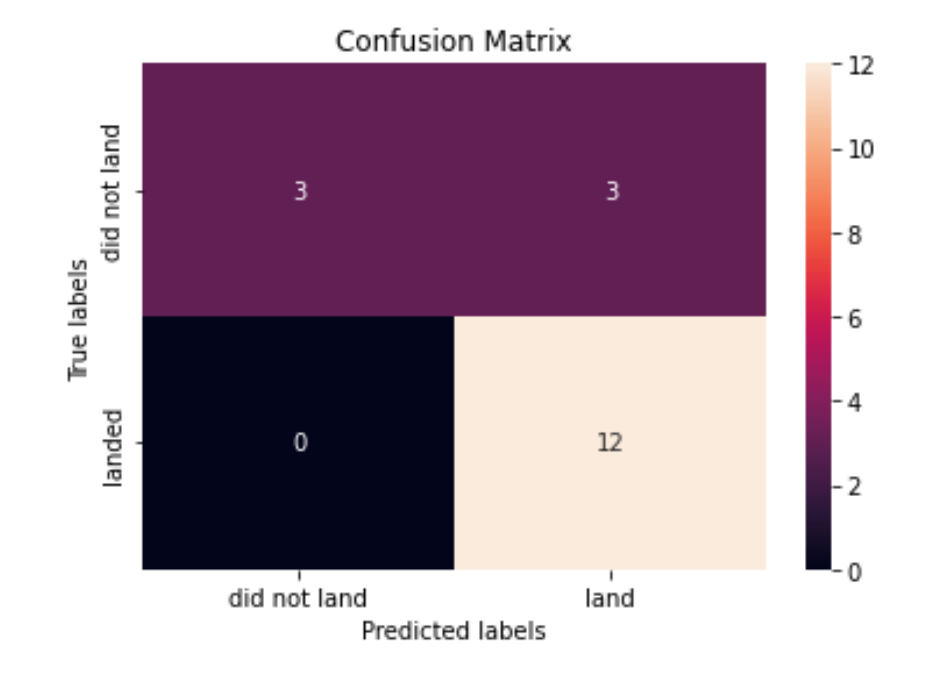
Predictive Analysis (Classification)

Classification Accuracy

- All methods had the same test score accuracy, being 83,33%;
- However, the decision tree had the highest train accuracy;



Confusion Matrix



This is the decision tree confusion matrix, with the only problem being the false negatives (3).

Conclusions

- A mission success can benefit from many factors, the most relevant ones being payload mass, launch sites and how many launches there have been before;
- The best launch site is KSC LC-39A;
- The orbits impacted the result but not as much as expected;
- The best predictive model was the Decision Tree, by a small train accuracy margin;

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!