Applied Data Science final report

Ivan Perez Torres
23.02.2024

Outline

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

Executive Summary

Objective:

Collect, process, analyze and visualize data from the company SpaceX for identifying factors that may determine a successful rocket landing using python

Summary of methodologies:

1. Data collection from SpaceX REST API and web scraping

2. Data wrangling for the target objective: landing outcome

3. Data exploratory analysis with:

- Visualization tools such as Matplotlib, Pandas, Seaborn and Numpy.
- SQL
- Geographical visualizations and markers using Folium package

4.Predictive Analysis building Machine learning models using logistic regression, support vector machine, decision tree and K-nearest neighbor.

Summary of all results:

- Exploratory analysis: improvement of the outcomes, Orbit ES-L1, GEO, HEO and SSO have 100% success rate and SO a 0% success rate; and launch Site VAFBSLC 4E may not be available for higher payload of 10000.
- Predictive analysis: all tested models have a similar performance, being decision tree model slightly better.

Introduction

Project background and context:

SpaceX has gained worldwide attention for a series of historic milestones. It is the only private company ever to return a spacecraft from low-earth orbit, which it first accomplished in December 2010. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars whereas other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage. In this project, we used public data and machine learning models to predict whether SpaceX or a competing company can reuse the first stage.

Problems you want to find answers

1. Landing outcome factors: payload mass, launch site, number of flights, and orbits
2. Rate of successful landings over time
3. What predictive model would successfully predict the landing outcome

Section 1

Methodology

Methodology

Executive Summary

- **Data collection methodology:** Data was collected using SpaceX REST API and web scraping from Wikipedia
- **Perform data wrangling:** Data filtering, missing values handling and one-hot encoding for outcome variable for later analysis and modeling
- **Perform exploratory data analysis (EDA) using visualization and SQL:**
- **Perform interactive visual analytics using Folium and Plotly Dash**
- **Perform predictive analysis using classification models: Build and evaluate models to find the best model with the best parameters.**

Data Collection

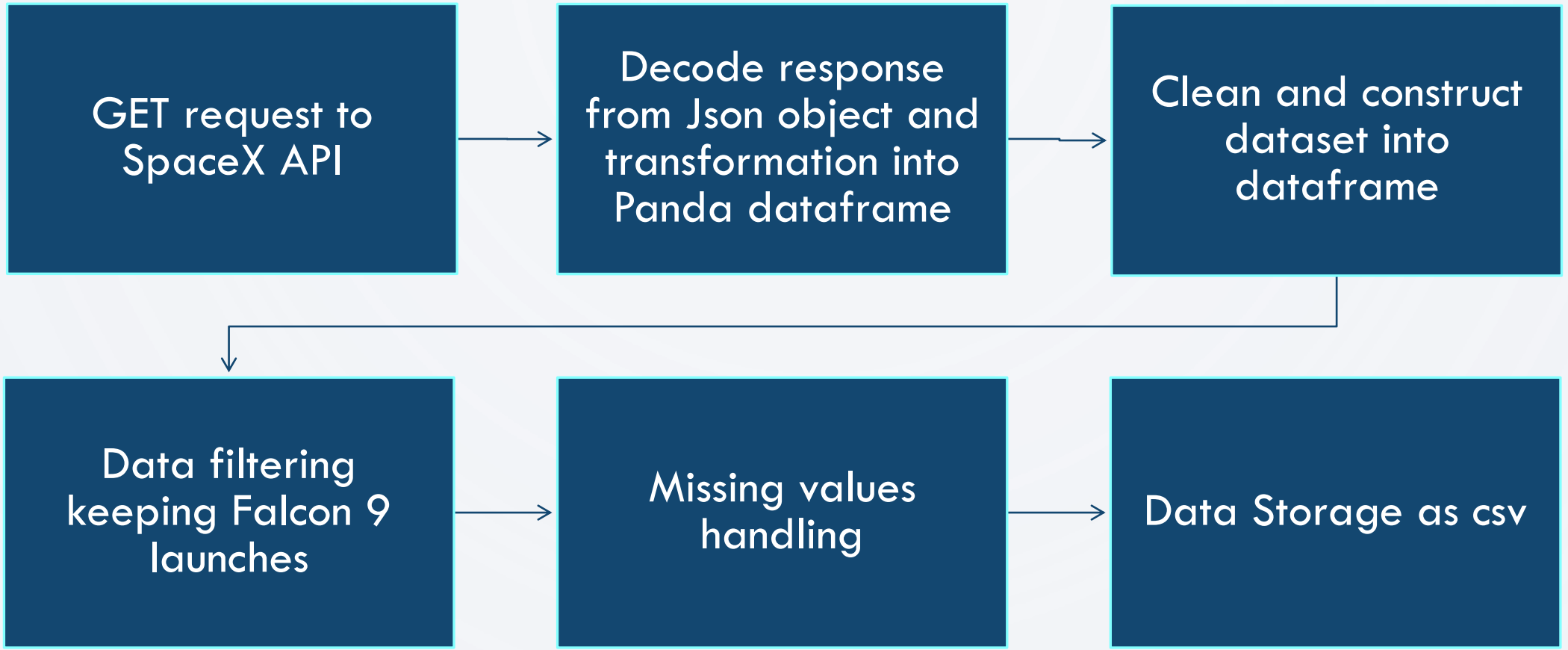
API data:

- Request data from SpaceX API using GET resquest.
- Decode response content Json using .json() and convert to a pandas dataframe with .json_normalize()
- Data cleaning, and filling of missing values

Web Scraping:

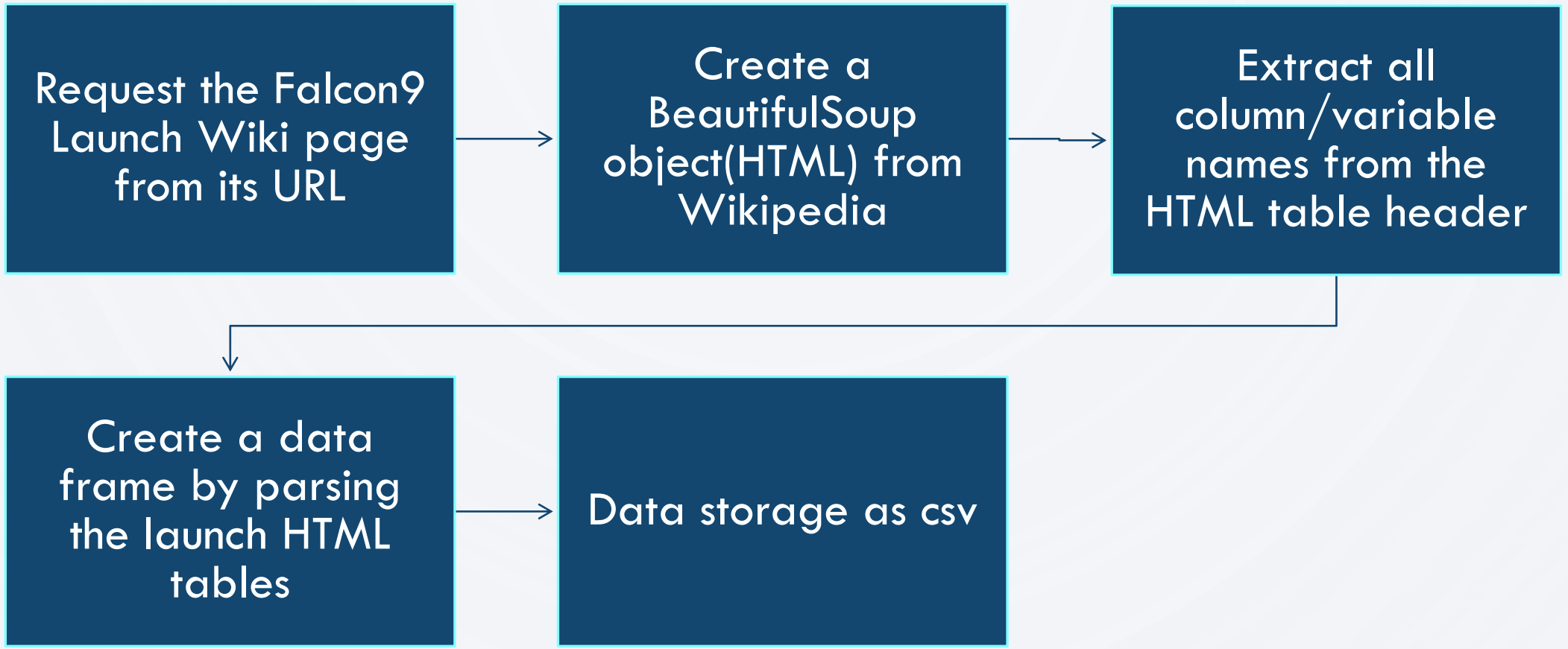
- Extract records as HTML from Wikipedia, parse the object and convert it to a pandas dataframe
- Filtering data for Falcon 9 launch records with BeautifulSoup
- Replacing of missing values with average for Payload_Mass variable
- Export data to csv file

Data Collection – SpaceX API



- Code in GitHub: https://github.com/Psivan5g/IBM-Final-Report/blob/2b69ba3bdbb3c4bde42e234567f8296edb1470c7/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping



- Code in GitHub: https://github.com/Psivan5g/IBM-Final-Report/blob/2b69ba3bdbb3c4bde42e234567f8296edb1470c7/jupyter-labs-webscraping.ipynb

Data Wrangling

Exploratory Data analysis for data labels with:

- Calculation of launches per site
- Number of occurrence of orbit
- Number of occurrence of mission outcome of the orbits

Creation of a landing outcome label with two values: 1 (Success) and 0 (Failure)

- Code in GitHub: https://github.com/Psivan5g/IBM-Final-Report/blob/92e376591d0a79bd4fc58912f0365a2309c2bec3/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

Variables compared and charts:

- Flight Number vs Launch Site → Scatter plot
- Flight Number vs Payload → Scatter plots
- Payload Mass vs Launch Site → Scatter plot
- Payload Mass vs Orbit type → Bar Charts
- Flight Number vs Orbit type → Scatter plot
- Payload Mass vs Outcome → Scatter plot
- Year vs Outcome → Line plot

- Code in GitHub: https://github.com/Psivan5g/IBM-Final-Report/blob/d400295c02028bf545e0273c72ee56c5827af06f/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

Queries:

- SpaceX dataset loading into PstgreSQL database with SQLite
- Name of unique launch sites in space mission
- Total payload mass carried by boosters launched by NASA(CRS)
- Average payload mass carried by booster version F9 V1.1
- Total number of successful and failure outcomes
- Boosted version and launch sites in failed landing in drone ship

- Code in GitHub: https://github.com/Psivan5g/IBM-Final-Report/blob/35fb7e709b61017ef58542a5f5a1b5923baddcac/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Load a world map with Folium
- Mark the launch sites, added to the map and circled then with information about the success and failure of the launches with a color code.
- Calculated the distance between the launch sites and its proximities.

- Code in GitHub: https://github.com/Psivan5g/IBM-Final-Report/blob/a9381ed42c774befff9ef80dd5e2dba75acef33e/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- Build an interactive dashboard with Plotly dash
- Plotted pie charts
- Plotted scatter graph showing the relationship between variable Outcome and Payload Mass for the different booster version

- Code in GitHub: https://github.com/Psivan5g/IBM-Final-Report/blob/d400295c02028bf545e0273c72ee56c5827af06f/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

Predictive Analysis (Classification)

- Loaded data with numpy and pandas, transformed and split into training and testing
- Build machine learning models: K nearest neighbour, decision tree, support vector machine and logistic regression
- Tuned hyperparameters using GridSearchCV
- Improving model using feature engineering and algorithm tuning.
- Evaluation of model using bar charts and confusion matrix

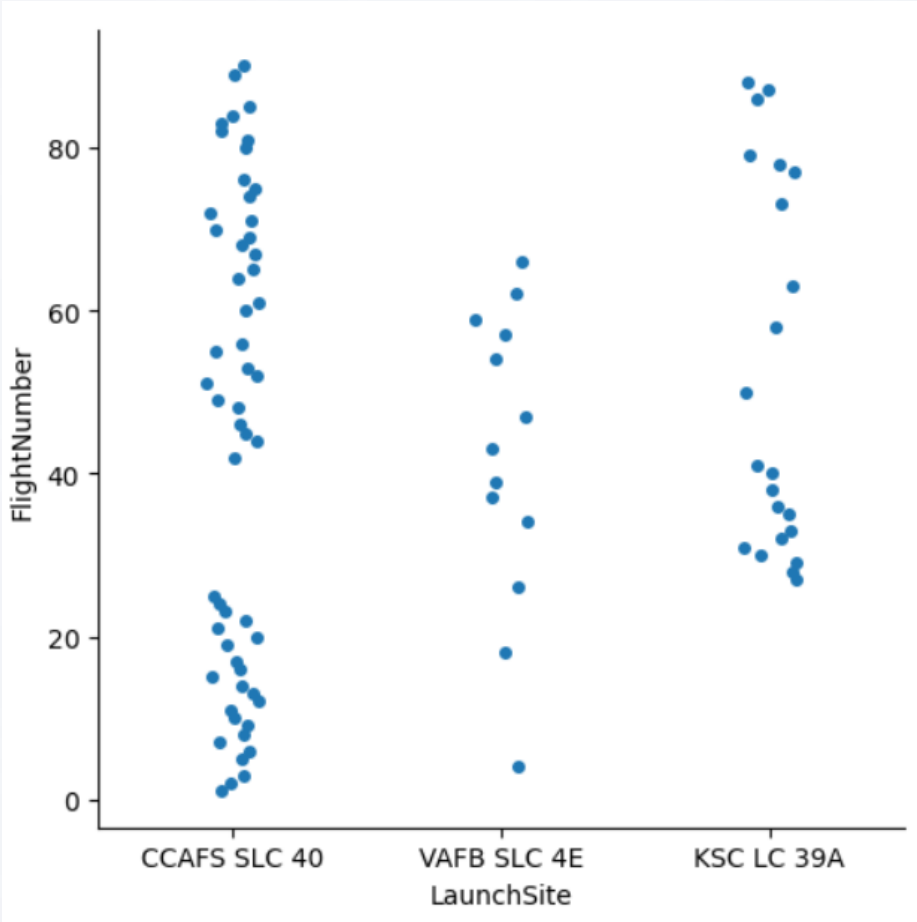
- Code in GitHub: https://github.com/Psivan5g/IBM-Final-Report/blob/6a895070dd4d1713959250ba4cab19179576fc45/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Section 2

Insights drawn from EDA

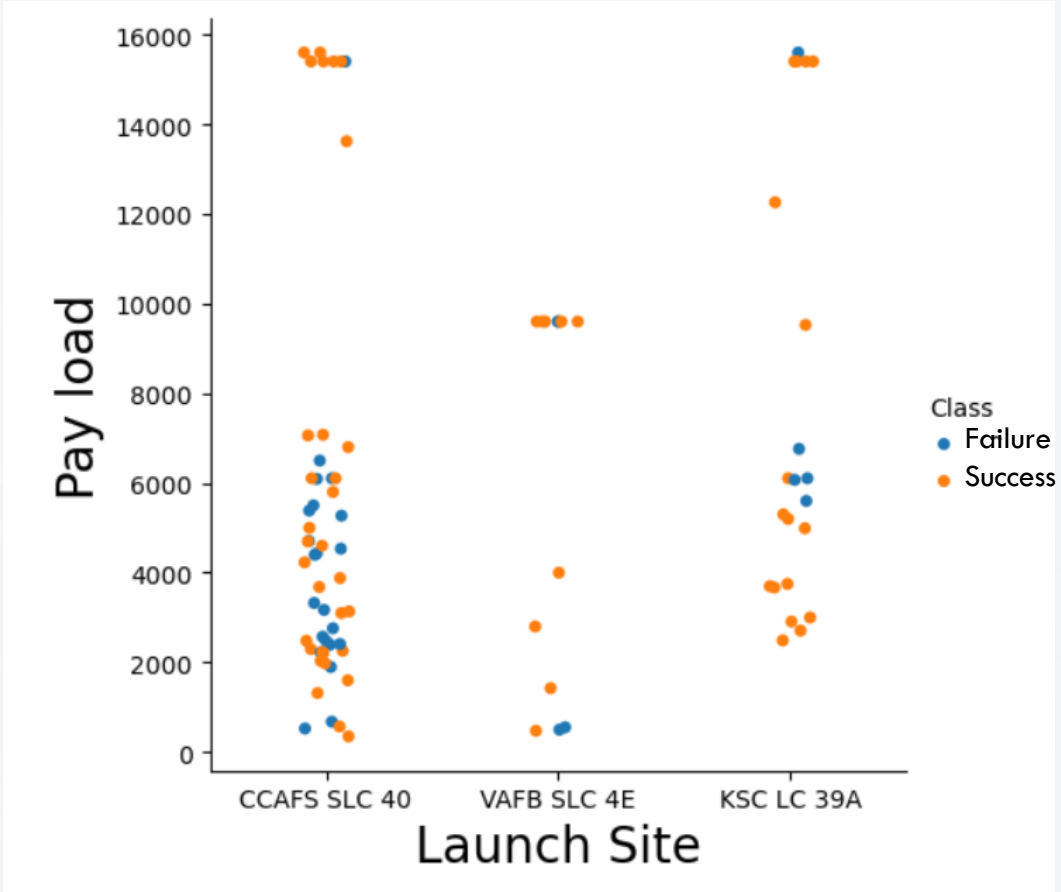
Flight Number vs. Launch Site

- CCAFS SLC 40 is the most popular place whereas VAF SLC 4E is the less place used for launching
- Further analysis could be done for determine how and why the CCAFS is more popular and what is wrong with VAFB



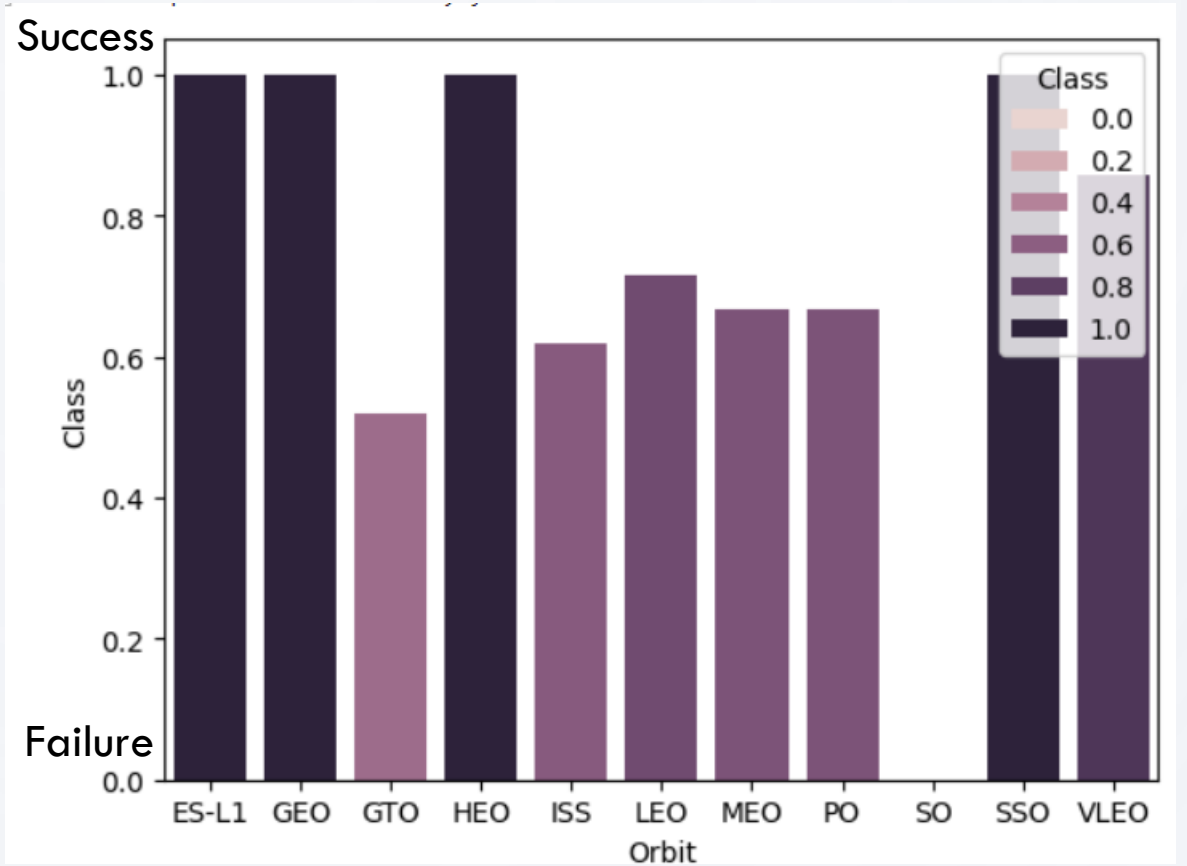
Payload vs. Launch Site

- VAFB SLC 4E has no launches with more tan 10000. Knowing why could give information about the suitability of that launching site.
- CCADS SLC40 has no launches from 8000 to 14000 of load
- Added outcome variable: seems that even though VAFB SLC 4E is not much used, the success rate is quite high. Additionally, it seems that higher Payload leads to more success outcome (from 10000 to 16000)



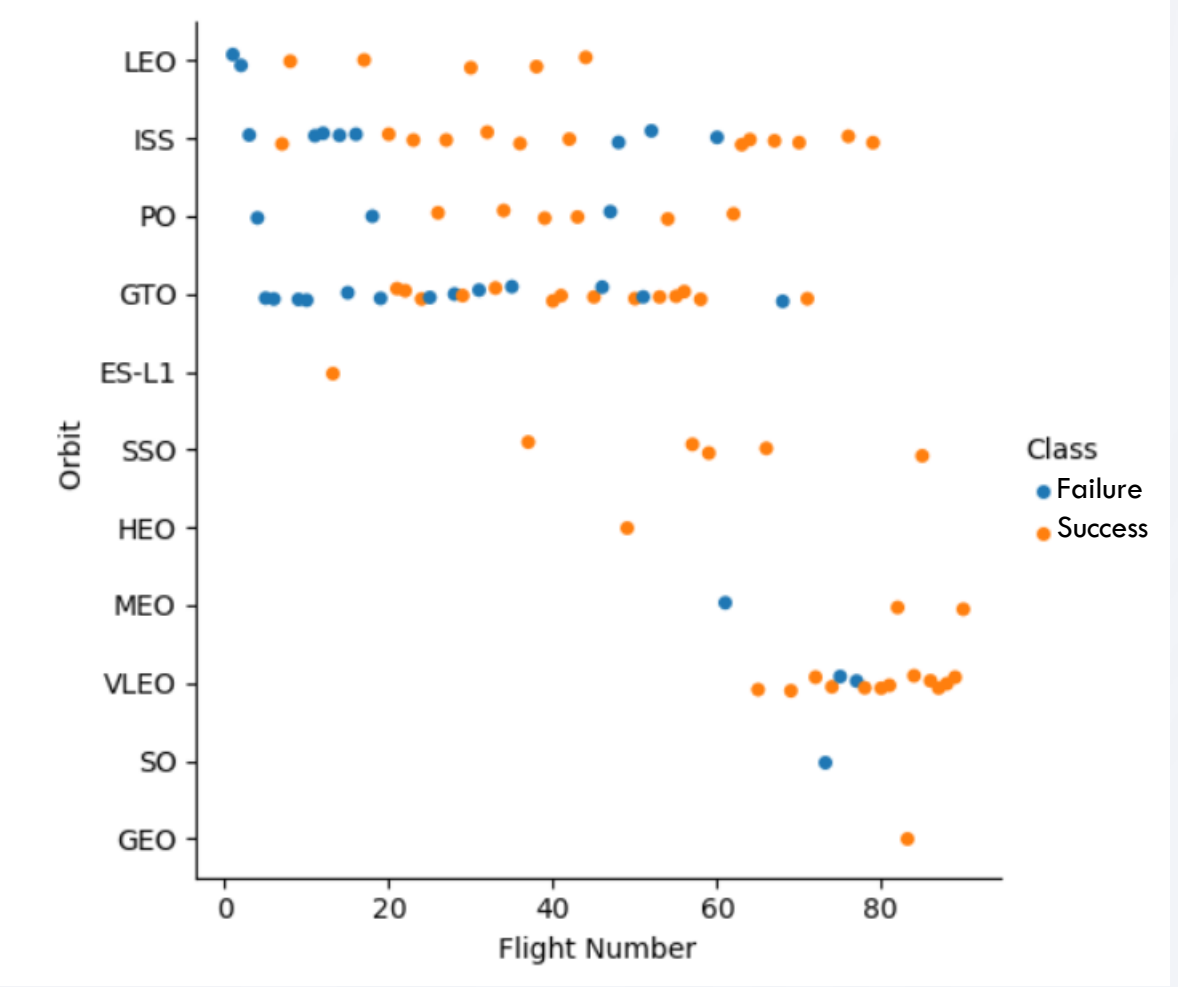
Success Rate vs. Orbit Type

- Launches to orbit ES-L1, GEO, HEO and SSO have high success rate whereas launches to orbit SO has the lowest.
- Mention to VLEO orbit, with 80% success rate
- Orbits with higher rate could lead with more secure launching, if we take into account the number of flights launched there



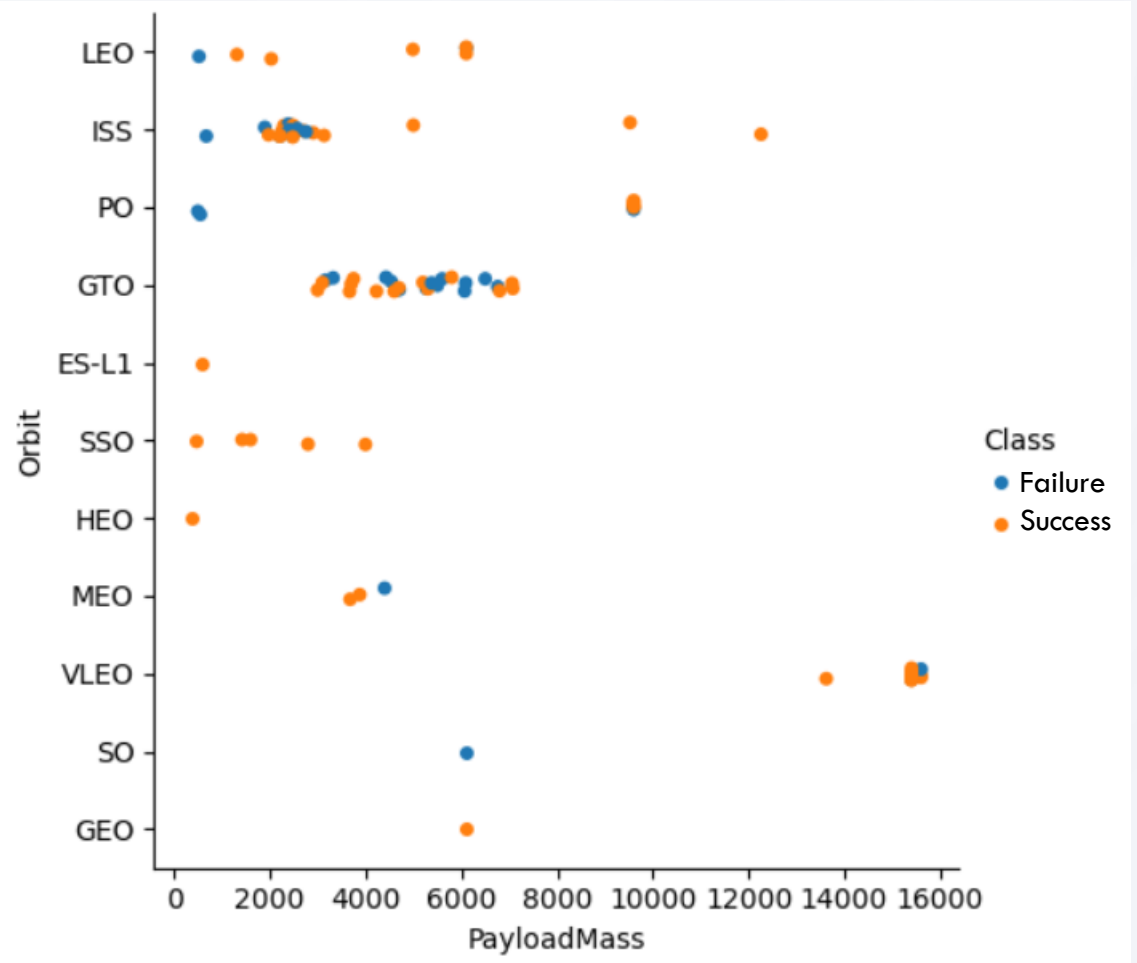
Flight Number vs. Orbit Type

- This plot gives explanation why the extreme rates in the last barplot. Orbits SO, HEO, ES-L1, and GEO have only one data point, meaning that it is difficult to generalize the results.
- However, VLEO (with 80% success rate) have more datapoints, being a promising orbit.



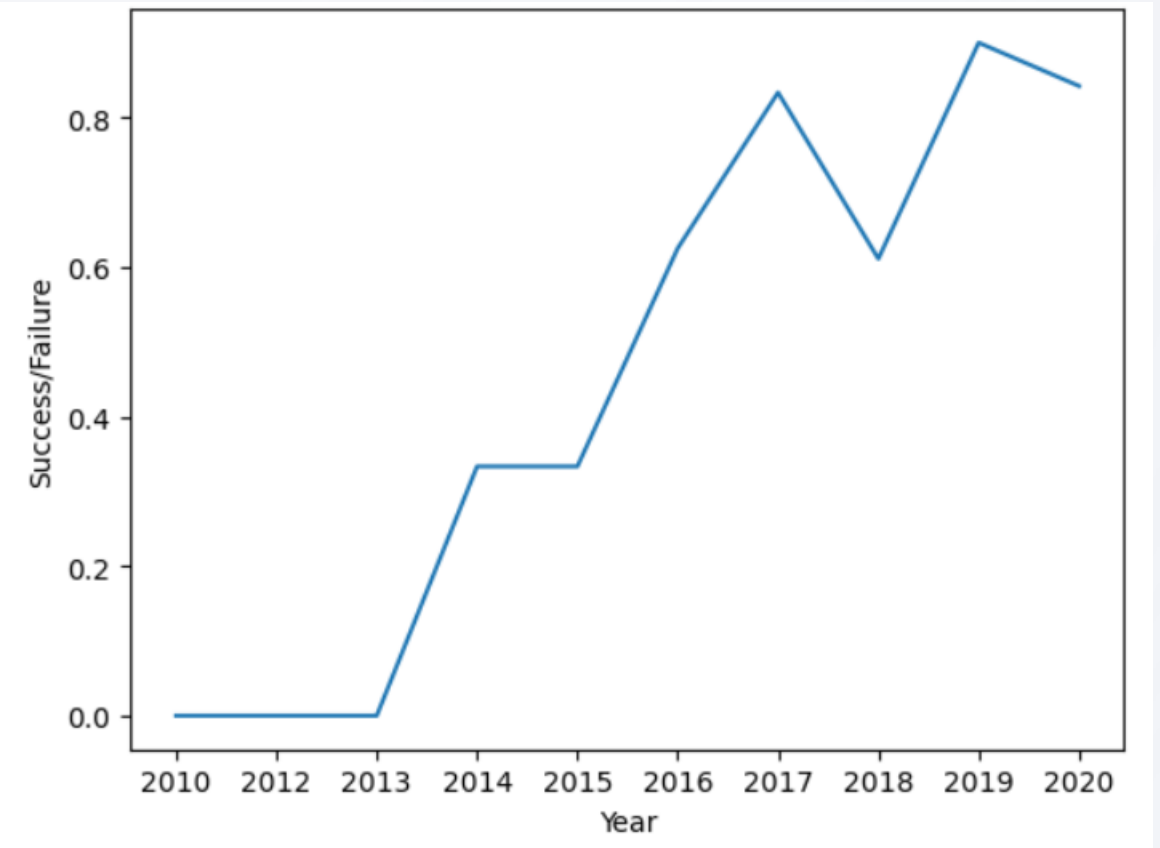
Payload vs. Orbit Type

- In VLEO, the Payload average Payload is high, around 16000 kg whereas in SSO, the Payload is low, below 4000 Kg



Launch Success Yearly Trend

- Clear improvement through the years from the starting to the ending
- Between 2017 and 2018 there was a decreasing in the performance, further research could be done for getting some insights from it



All Launch Site Names

- Query: %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'

- Query: %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
2010-06-04	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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

- Query: %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer like '%NASA%';

sum(PAYLOAD_MASS__KG_)
107010

Average Payload Mass by F9 v1.1

- Query: %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version like '%F9%';

avg(PAYLOAD_MASS__KG_)
6138.287128712871

First Successful Ground Landing Date

- Query: %sql select min(Date) from SPACEXTBL where Mission_Outcome like '%Success%';

min(Date)
2010-06-04

Successful Drone Ship Landing with Payload between 4000 and 6000

- Query: %sql select distinct(Booster_Version) from SPACEXTBL where Landing_Outcome like '%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

- Query: %sql select count(MISSION_OUTCOME) as missionoutcomes from SPACEXTBL GROUP BY MISSION_OUTCOME;

missionoutcomes
1
98
1
1

Boosters Carried Maximum Payload

- Query: %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

2015 Launch Records

- Query: %sql SELECT substr(Date, 6,2),MISSION_OUTCOME,BOOSTER_VERSION,LAUNCH_SITE FROM SPACEXTBL where substr(Date,0,5)='2015';

substr(Date, 6,2)	Mission_Outcome	Booster_Version	Launch_Site
01	Success	F9 v1.1 B1012	CCAFS LC-40
02	Success	F9 v1.1 B1013	CCAFS LC-40
03	Success	F9 v1.1 B1014	CCAFS LC-40
04	Success	F9 v1.1 B1015	CCAFS LC-40
04	Success	F9 v1.1 B1016	CCAFS LC-40
06	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

- Query: %sql SELECT LANDING_OUTCOME, COUNT(*) AS COUNT_LAUNCHES FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING_OUTCOME ORDER BY COUNT_LAUNCHES DESC;

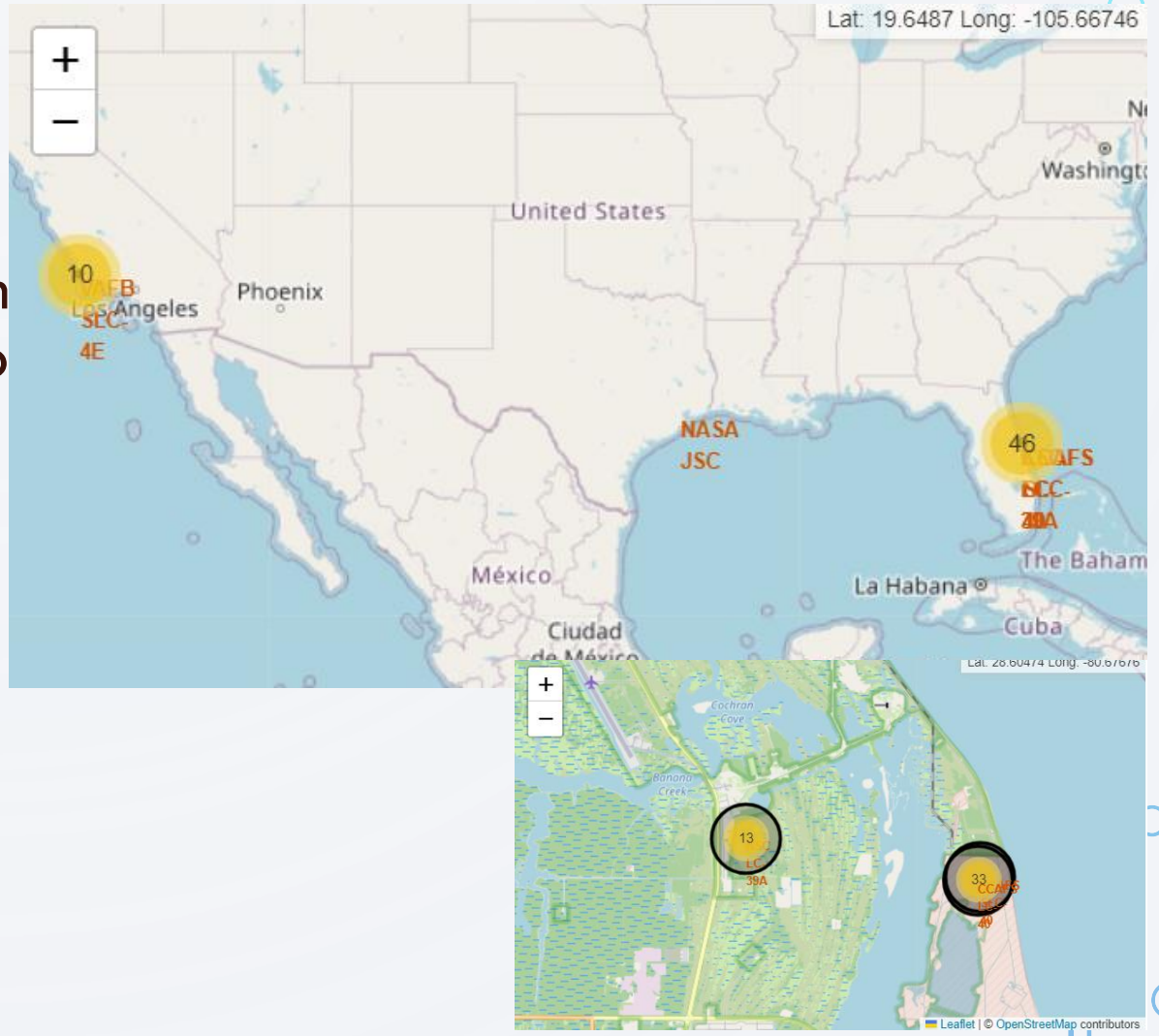
Landing_Outcome	COUNT_LAUNCHES
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

Folium map: Launch sites in USA

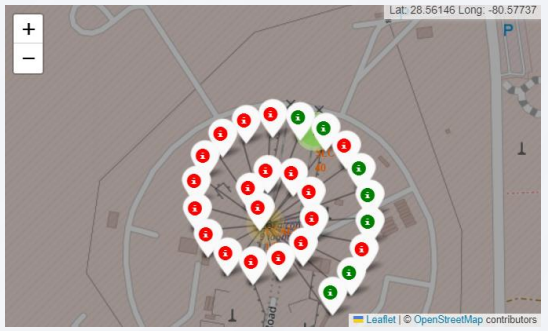
- There are mainly two launching areas: One on the east coast of USA and the other on the west coast.
- The East coast have three different areas in the same city, close to each other. And two of them are almost in the same place (CCAFS)
- Most of the launchings have been done from the east coast

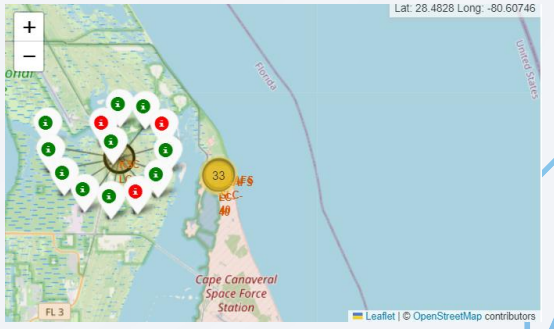


Folium Map: Outcome landings per sites

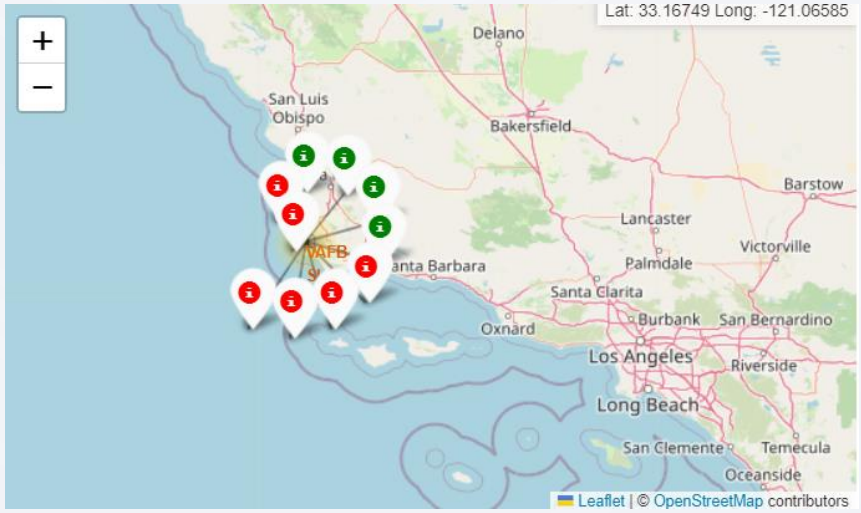
- KSC LC 39 A has high rate of successful landings.

East coast:



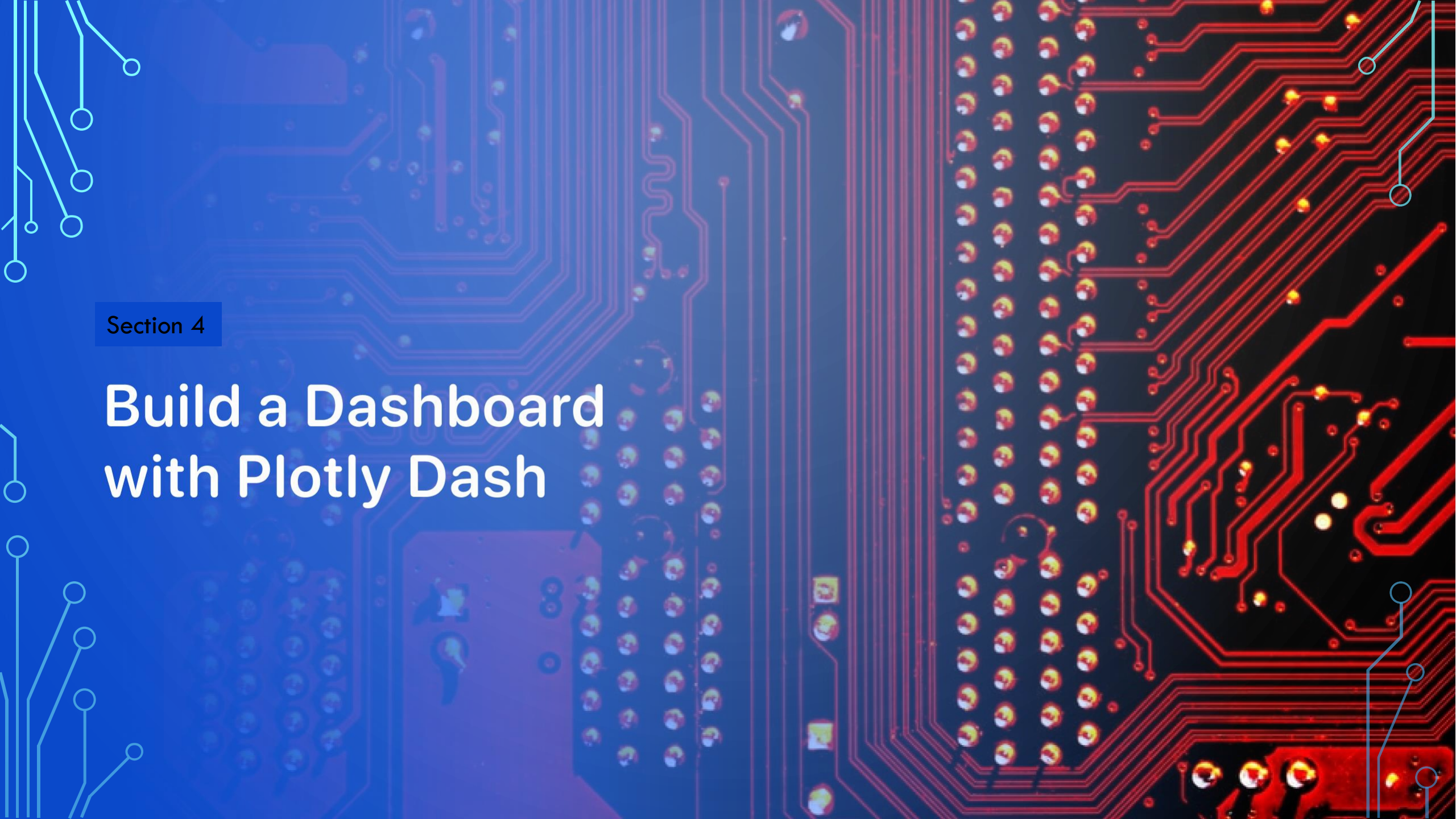


West coast:



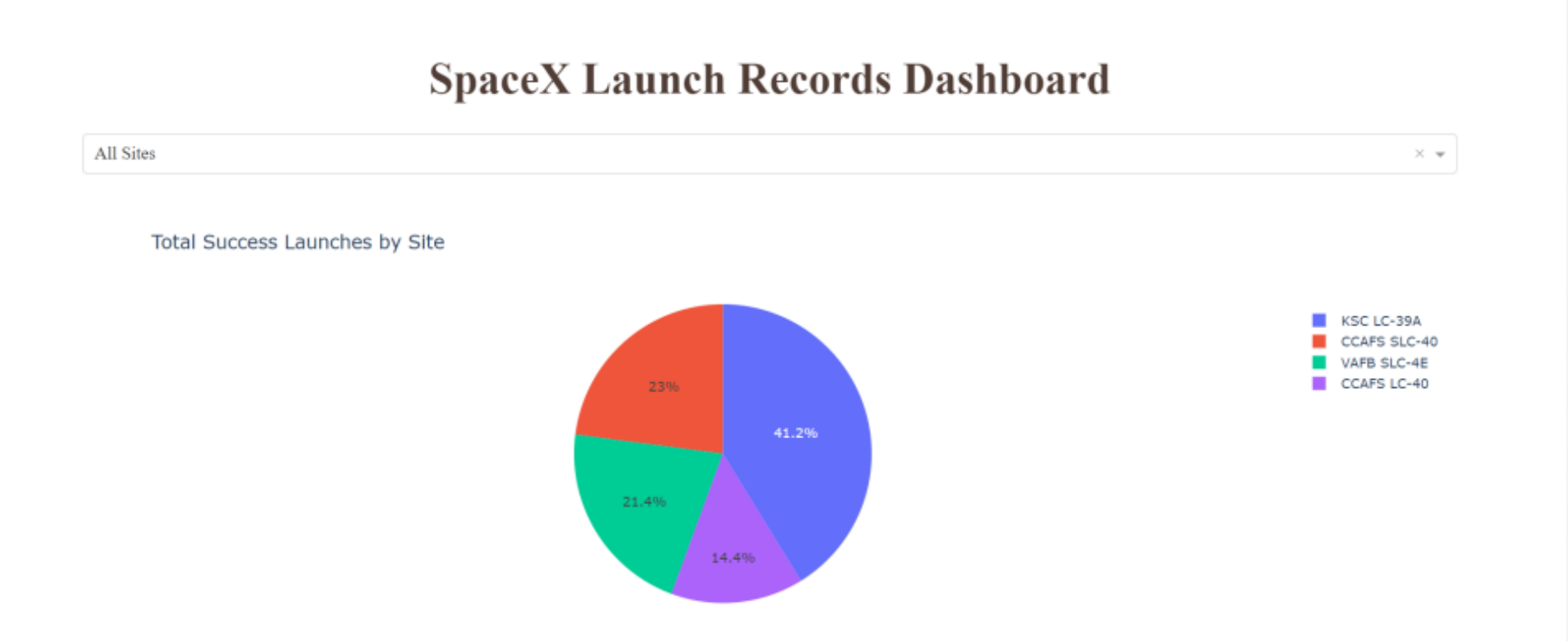
Folium Map: Distance of CCAFS stations

- The proximity to the coast could be one factor that affect the success.


Section 4
Build a Dashboard with Plotly Dash

SpaceX Launch Records – Pie plot

- Based on this records, KSC seems to be better option for launching with higher success whereas CCAFS LC-40 seems to be the slighter worst option.



Correlation Between Payload and Success per Site

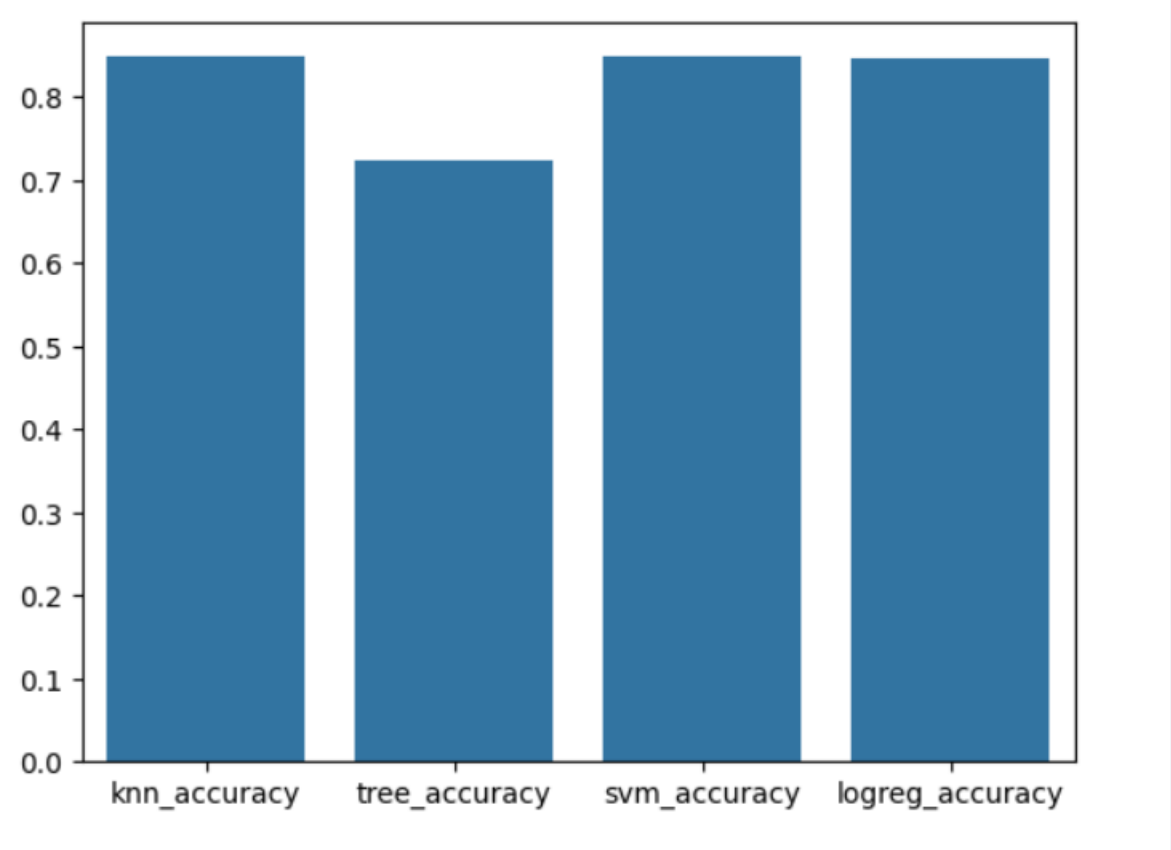
- No further insights can be done from this plot

Section 5

Predictive Analysis (Classification)

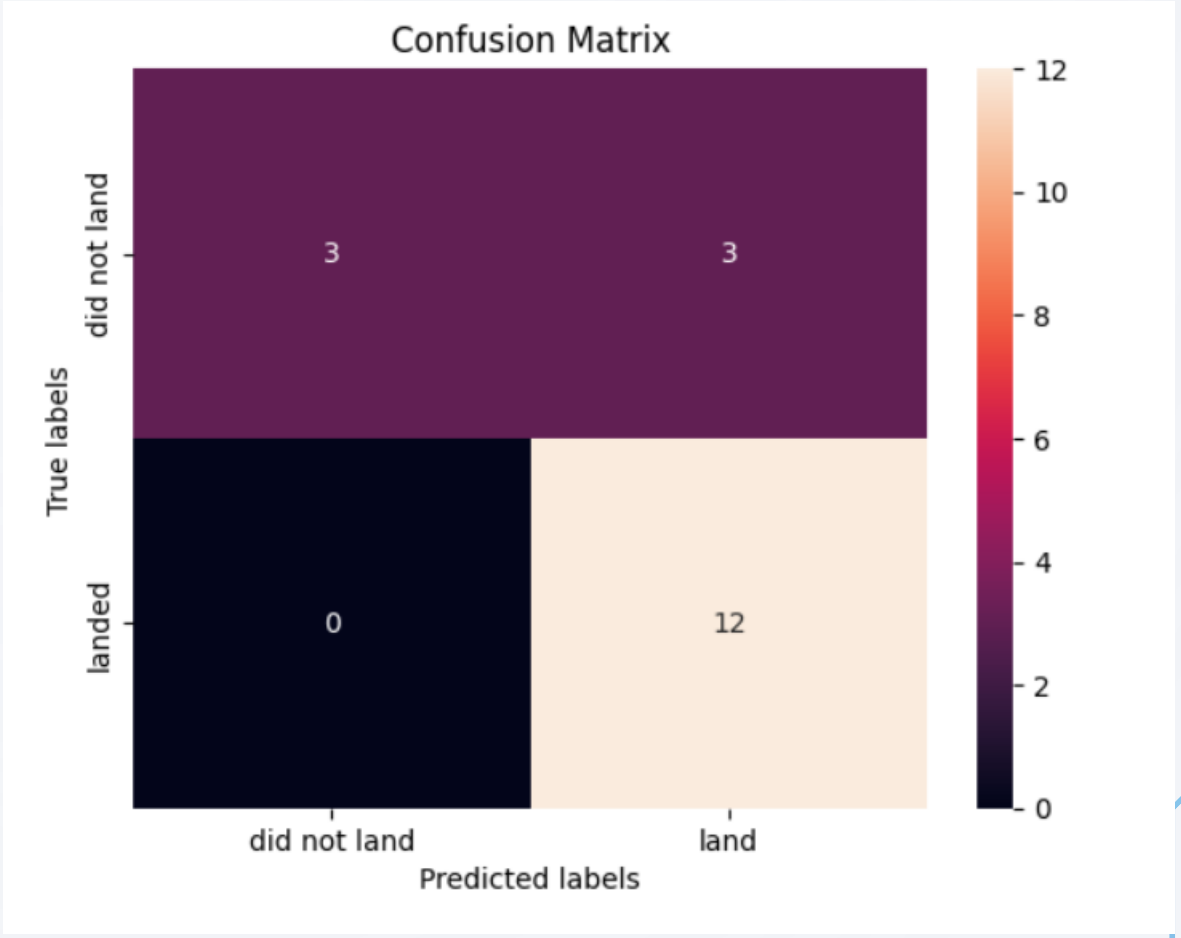
Classification Accuracy

- Base on the accuracy, all models performed similarly. However, decision tree was slightly less accurate compared with the other models



Confusion Matrix

- All models performance had similar results
- The true positive performance is accurate, however , the models cannot determine accurate the negative outcomes
- More data could help for improving the false positive results



Conclusions

- Localization: KSC LC-39 A has the higher success rate among launch sites. Especially with maximum success with launches below 60000 kg
- Orbits: VLEO orbit seems a promising starting orbit with high success rate and relatively enough information to support
- The higher the Payload, the most probable success
- Our models are relatively accurate for performance predictions, however decision tree was slightly less accurate than the others

Thank you!