IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jonas Lee
30/9/24

Outline

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

Executive Summary

Summary of methodologies

- Data collection using API and web scraping
- Data wrangling
- Exploratory Data Analysis with Data Visualization and SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive modeling

Summary of all results

- Exploratory Data Analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Introduction

Project background and context

SpaceX is the leader in the commercial space industry, significantly reducing the cost of space travel. The company offers Falcon 9 rocket launches for $62 million, while competitors charge over $165 million, largely due to SpaceX's ability to reuse the first stage. By predicting whether the first stage will successfully land, we can estimate launch costs. This study will leverage public data and machine learning models to forecast the reusability of SpaceX's first stage.

Questions to answer

How do factors like payload mass, launch site, number of flights, and orbital trajectories influence the success of first-stage landings?

What are some useful insights we can draw from available data?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX Rest API
  - Web scrapping from Wikipedia
- Perform data wrangling
  - Assign training labels for supervised models by converting mission outcomes into a binary format.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Determining the best classification algorithm (Logistic regression, SVM, decision tree, & KNN)

Data Collection

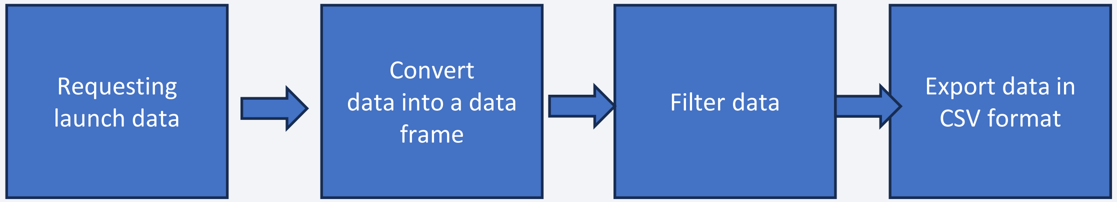
- Describe how data sets were collected.
- Data sets were collected from SpaceX REST API and web scrapped from Wikipedia
- These datasets are then filtered and exported in the CSV format where it will be further processed in our modelling.
- You need to present your data collection process use key phrases and flowcharts

Data Collection – SpaceX API

More precise operations can be found in referenced jupyter notebook.

Notebook link:

https://github.com/J0na3/IBM-DS-Capstone-/blob/main/IBM%20capstone/jupyter-labs-spacex-data-collection-api.ipynb

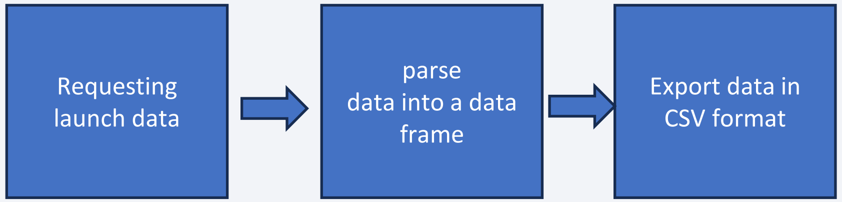


Data Collection - Scraping

- More precise operations can be found inside of the referenced notebook.

Notebook link:

https://github.com/J0na3/IBM-DS-Capstone-/blob/main/IBM%20capstone/jupyter-labs-webscraping.ipynb



Data Wrangling

Data set referenced many different outcomes. We converted these outcome to binaries, to be processed further in our analysis.

Summary of operations

- Determine the launch frequency per site
- Determine the count and frequency of each distinct orbit
- Determine the count and frequency of mission outcomes for each orbit type
- Create a landing outcome label

Notebook link:

https://github.com/J0na3/IBM-DS-Capstone-/blob/main/IBM%20capstone/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

Charts plotted:

1. Flight Number vs. Launch Site, Scatter chart
2. Payload Mass vs. Launch Site, scatter chart
3. Orbit Type vs. Success Rate, Bar chart
4. Flight Number vs. Orbit Type, scatter chart
5. Payload Mass vs Orbit Type, scatter chart
6. Success Rate Yearly Trend, Line chart

- Scatter plots illustrate the relationship between variables and can be utilized in machine learning models if a relationship is present.
- Bar charts compare discrete categories, highlighting the relationship between specific categories and a measured value.
- Line charts depict trends in data over time, making them ideal for time series analysis

https://github.com/J0na3/IBM-DS-Capstone-/blob/main/IBM%20capstone/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

SQL Queries:

https://github.com/J0na3/IBM-DS-Capstone-/blob/main/IBM%20capstone/jupyter-labs-eda-sql-coursera_sqllite.ipynb

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
- List the 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

Build an Interactive Map with Folium

Added Markers, circles on all launch sites.

-To identify and show proximity.

Used lines to measure distances between sites and other proximities.

https://github.com/J0na3/IBM-DS-Capstone-/blob/main/IBM%20capstone/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

Dashboard interactions:

- Added a Launch Site Drop-down Input component to the dashboard to filter Dashboard visual by launch site.
- Added a Pie Chart to the Dashboard to show total success launches when ‘All Sites’ is selected and show success and failed counts when a particular site is selected
- Added a Payload range slider to the Dashboard to select different payload ranges to identify visual patterns
- Scatter chart created to observe how payload may be correlated with mission outcomes for selected site(s). The colour-label Booster version on each scatter point provided missions outcomes with different boosters

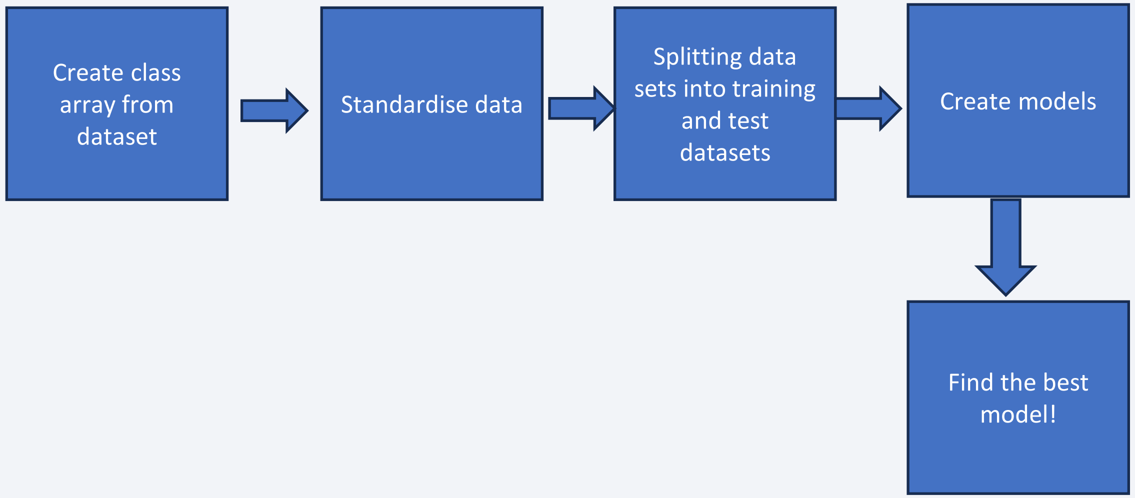
https://github.com/J0na3/IBM-DS-Capstone-/blob/main/IBM%20capstone/plotly%20dash%20app.py

Predictive Analysis (Classification)

1. Logistic regression
2. SVM
3. decision tree,
4. KNN

These where the algorithms used in the modelling process.

Metrics used for evaluation where Jaccard score and f1 score metrics.



- https://github.com/J0na3/IBM-DS-Capstone-/blob/main/IBM%20capstone/SpaceX-Machine-Learning-Prediction-Part-5-v1.ipynb

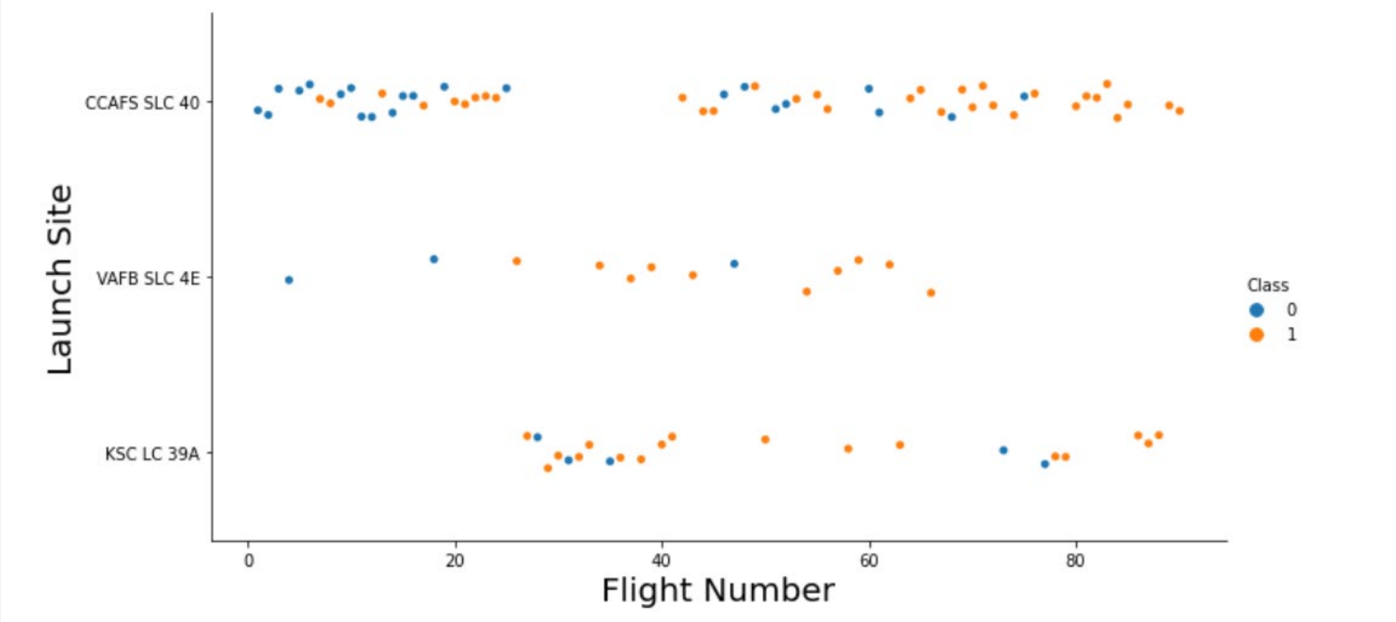
Results

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

Section 2

Insights drawn from EDA

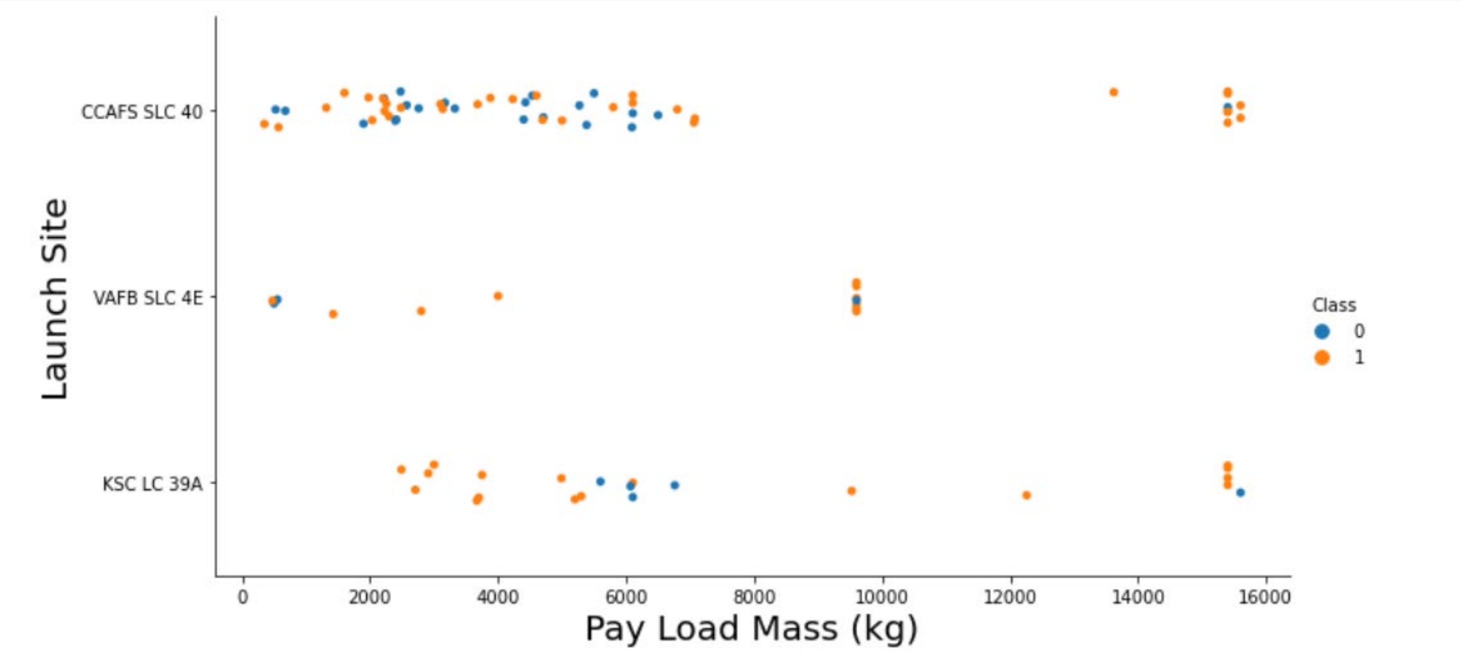
Flight Number vs. Launch Site



- Success rate and number of flights seems to be correlated. .: newer flights have a higher rate of success.
- Most launches are on CCAFS SLC 40.

Payload vs. Launch Site

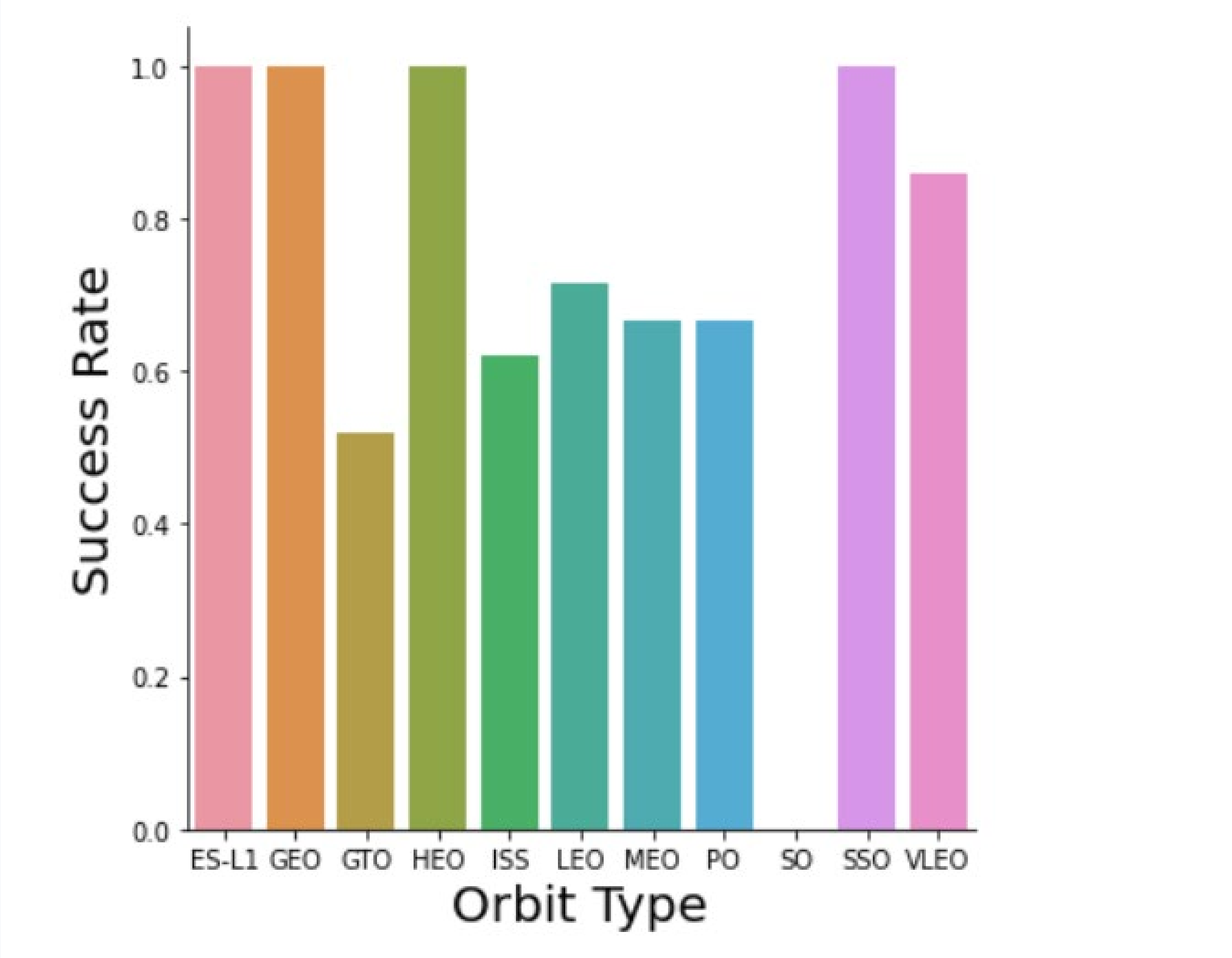
Although the relationship isn't very clear, it is evident that the higher the payload mass, the higher the success rate.



Success Rate vs. Orbit Type

Orbits ES-LI GEO HEO SSO have a 100% success rate. Suggesting they are the best orbits.

Orbit SO is the worst with a 0% success rate.

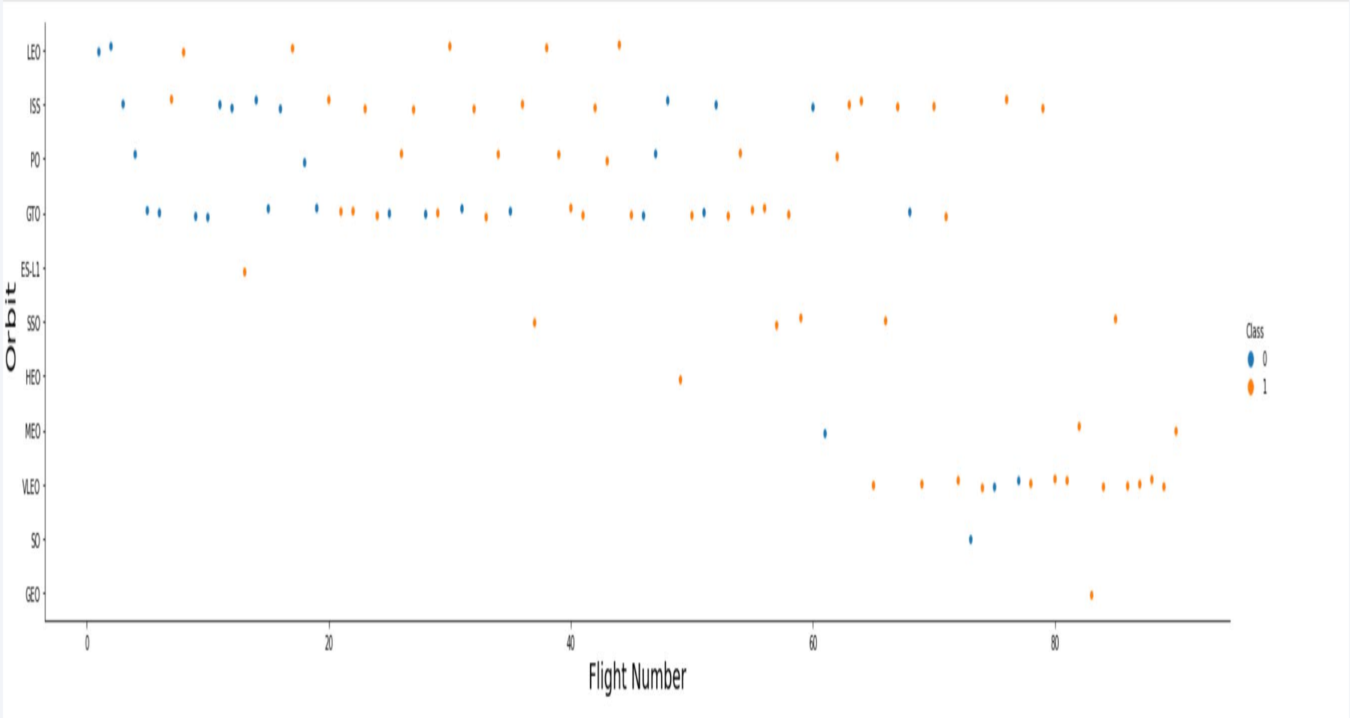


Flight Number vs. Orbit Type

Most flights orbit ISS, GTO and VLEO.

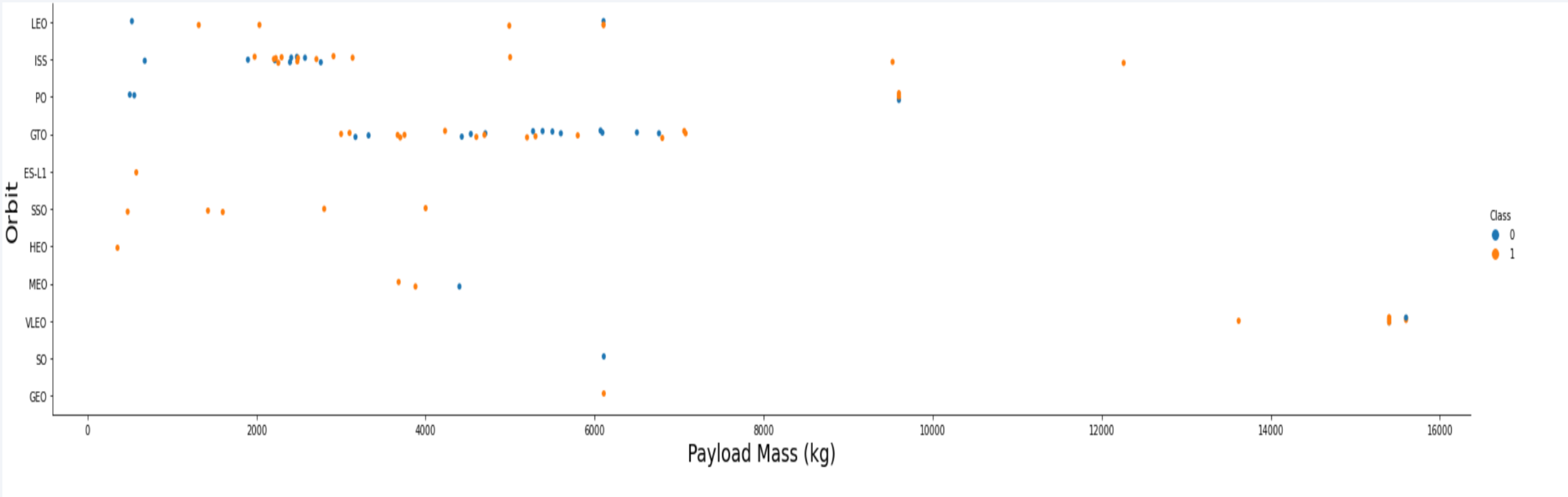
A disproportionate share of flights outside of these 3 with GEO, HEO, SEO and ES-LI with a single flight each.

This affects our success rates as the data set is too small and thus not representative.



Payload vs. Orbit Type

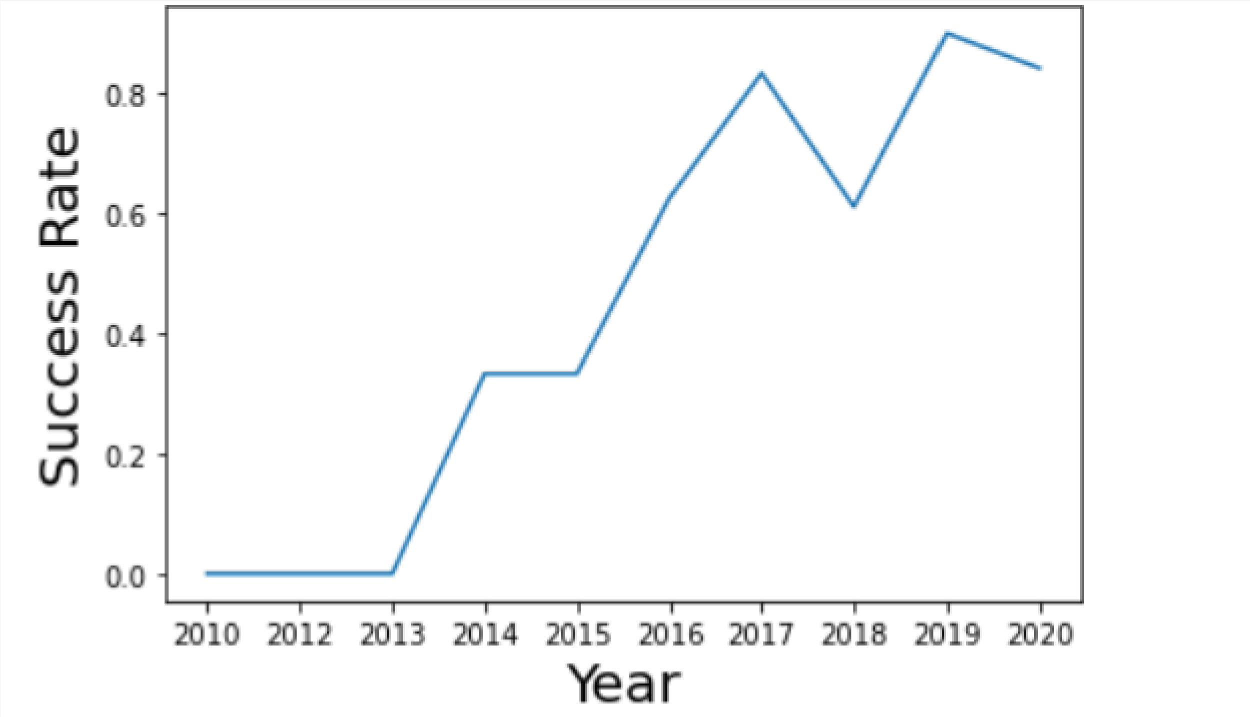
- Majority of payloads between 2500 and 7000 are orbiting GTO



Launch Success Yearly Trend

Overall clear improvement in year-on-year success rates.

However significant dip in 2018. May be worth looking into why that is.



All Launch Site Names

SQL query: “select distinct launch_site from spacextable”

Distinct only returns unique values.

Result:

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

Launch Site Names Begin with 'CCA'

Query: select * from spacextable where launch_site like ‘CCA%’ limit 5

This displays the 5 records where launch sites is CCA. It returns launch sites from CCAFS, as it starts 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	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:

Select sum(payload_mass_kg) from spacextable where customer = ‘NASA (CRS)’

This adds all the values in payload mass for the customer NASA crs

Results

sum(payload_mass__kg_)
45596

Average Payload Mass by F9 v1.1

Query:

select avg(payload_mass_kg_) from spacextable where booster_version like “
F9 v1.1”

Returns the avg payload mass from the dataset where booster version is similar to F9 v1.1

Result:

avg(payload_mass__kg_)
2534.6666666666665

First Successful Ground Landing Date

Query:

Select min(date) as first_successful_landing from spacextable where landing_outcome = ‘Success (ground pad)’

Returns the first date where landing outcome was successful and on ground pad.

Result:

min(date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Query: select booster_version from spacextable where (payload_mass_kg between 4000 and 6000) and landing outcome = ‘Success (drone ship)

Selects booster version of successful drone ship landings where payload is between 4000 and 6000.

Result:

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:

Select mission_outcome, count* as total_number from spacextable group by mission outcome

This groups the outcomes together and returns a count of each outcome.

Result:

Mission_Outcome	count(*)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

Query:

Select booster_version from spacextable where payload_mass_kg = (select max(payload_mass_kg) from spacextable)

Selects the booster versions carrying the max payload from the dataset.

Results:

Booster_Version
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:

select substr(date,6,2) as months ,date, booster_version, launch_site, landing_outcome from spacextable where landing_outcome = 'Failure (drone ship)' and date between "2015-01-01" and "2015-12-31”

Selects date, booster version, launch site and outcome of failed drone ships during 2015

months	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

Query: select landing_outcome, count(*) as count_outcomes from spacextable where date between '2010-06-04' and '2017-03-20' group by landing_outcome order by count_outcomes desc

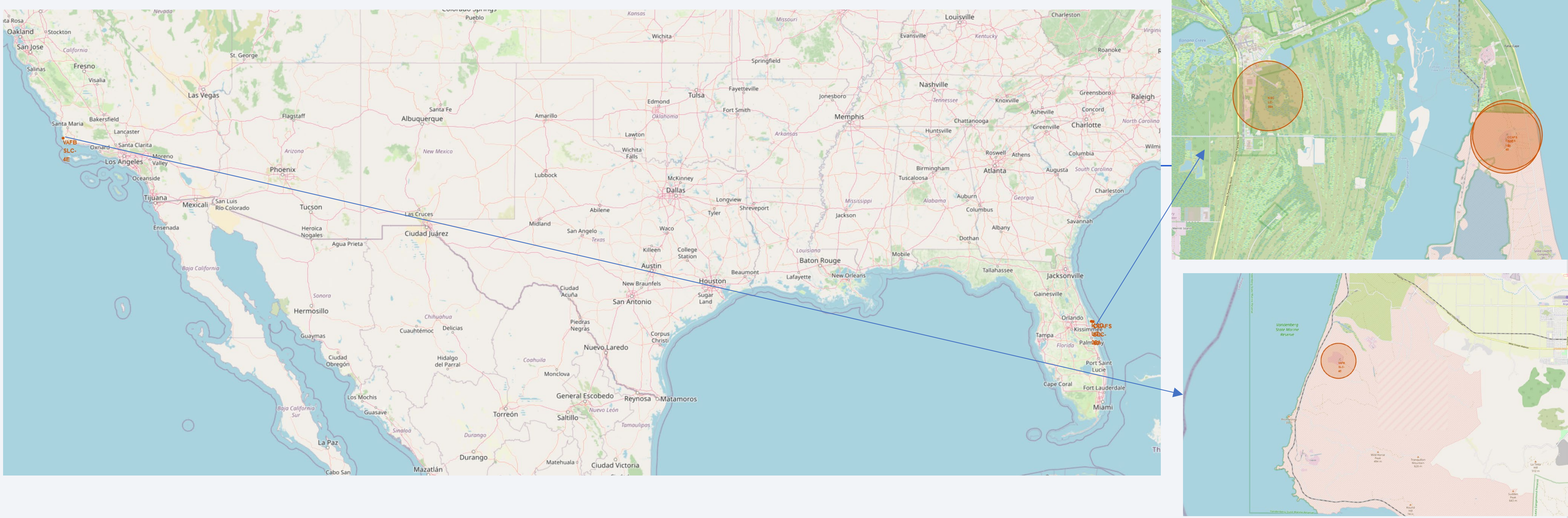
Selects and counts the landing outcome between the two dates and groups together similar outcomes. Finally ordering them in descending order of outcomes.

Results:

Landing_Outcome	count_outcomes
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

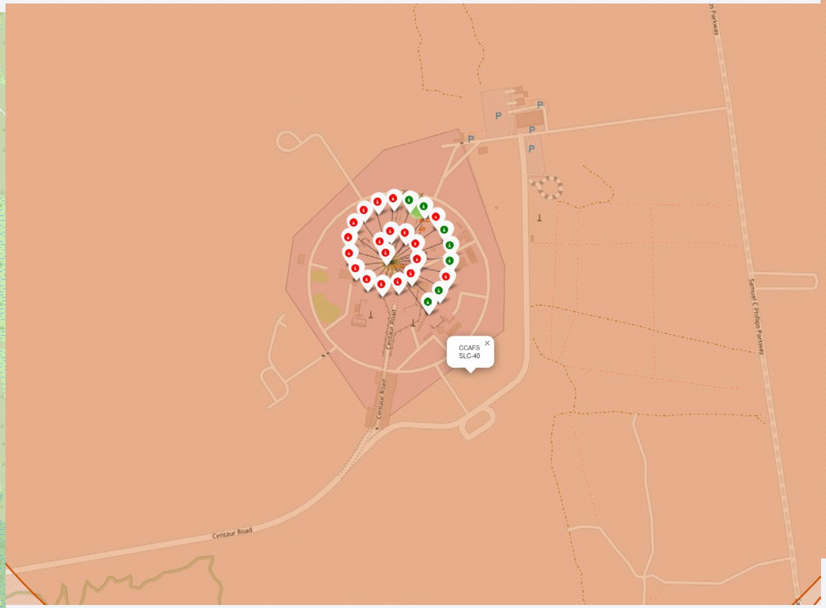
Launch Sites Proximities Analysis

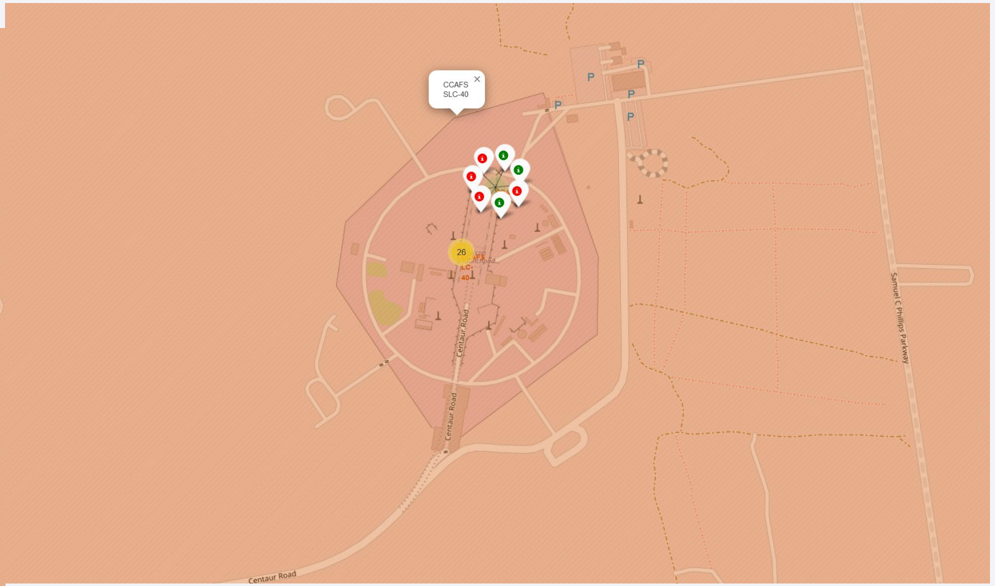


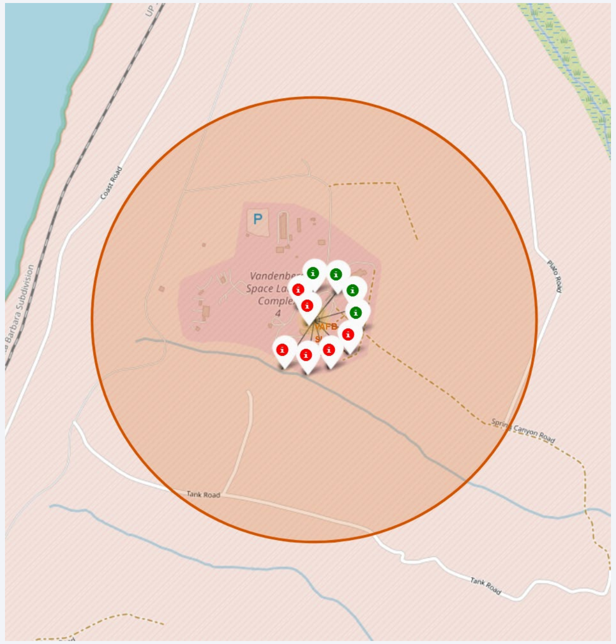
We can see from this graph that all the launch sites are located on the coast.
This is most likely due to the limited number of proximities which minimises risk to external civilians

3 of the sites are located in close proximity to one another. While VAFB SLC-4E is the only site in the west.

Success/Failures of launch sites





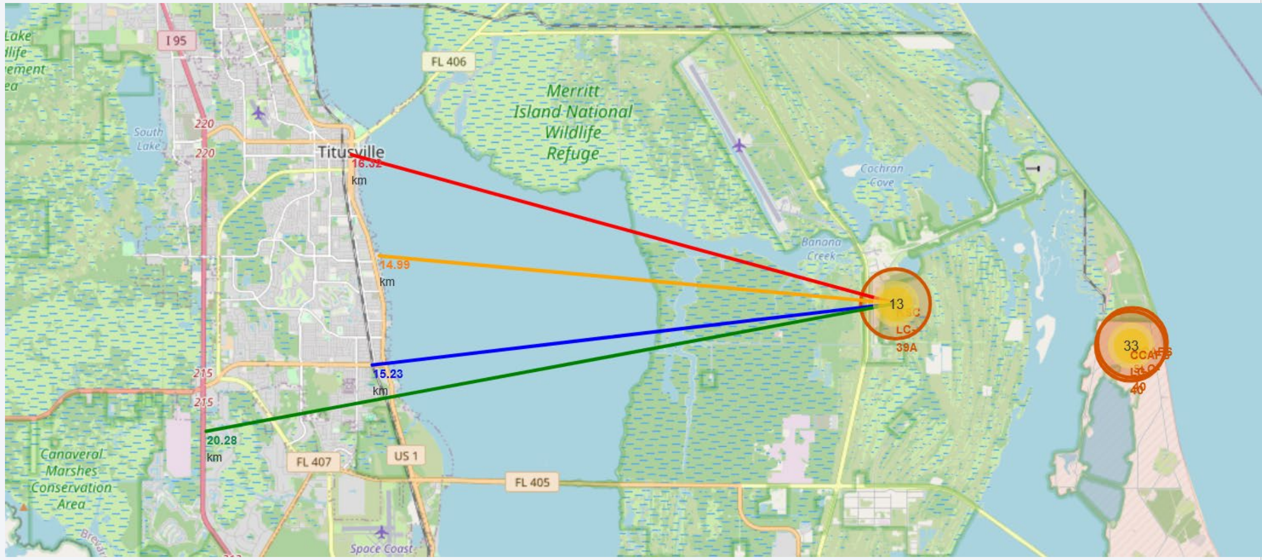


From this graph, we can see that CCAFS SLC40 is has the most launches. However, KSC LC-39A evidently has the highest success rate.

Green = Success
Red = Unsuccessful

Proximity distances

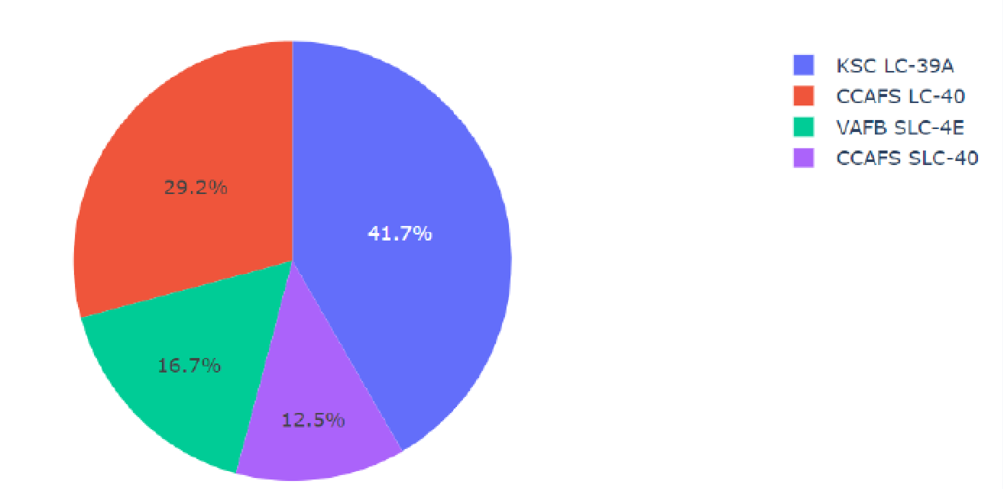
- From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:
- relative close to railway (15.23 km)
- relative close to highway (20.28 km)
- relative close to coastline (14.99 km)
- Also the launch site KSC LC-39A is relative close to its closest city Titusville (16.32 km)
- Failed rocket with its high speed can cover distances like 15-20 km in few seconds.It could be potentially dangerous to populated areas.



Section 4

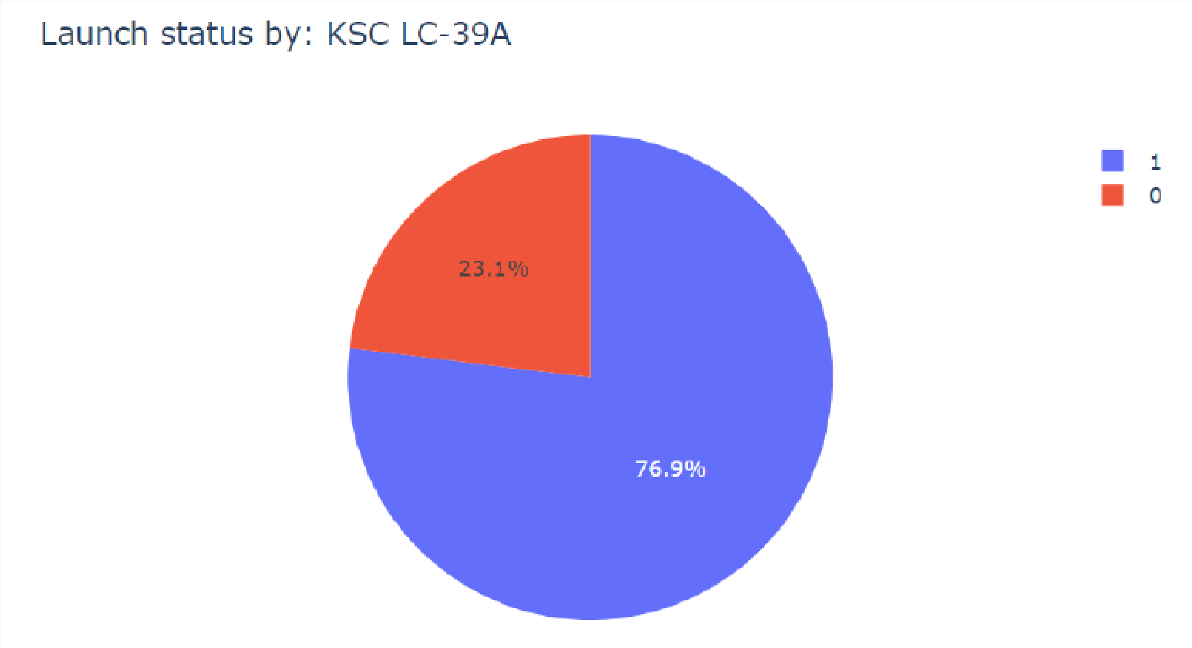
Build a Dashboard with Plotly Dash

Total successful launches for all sites



KSC LC-39A has the most successful launches.

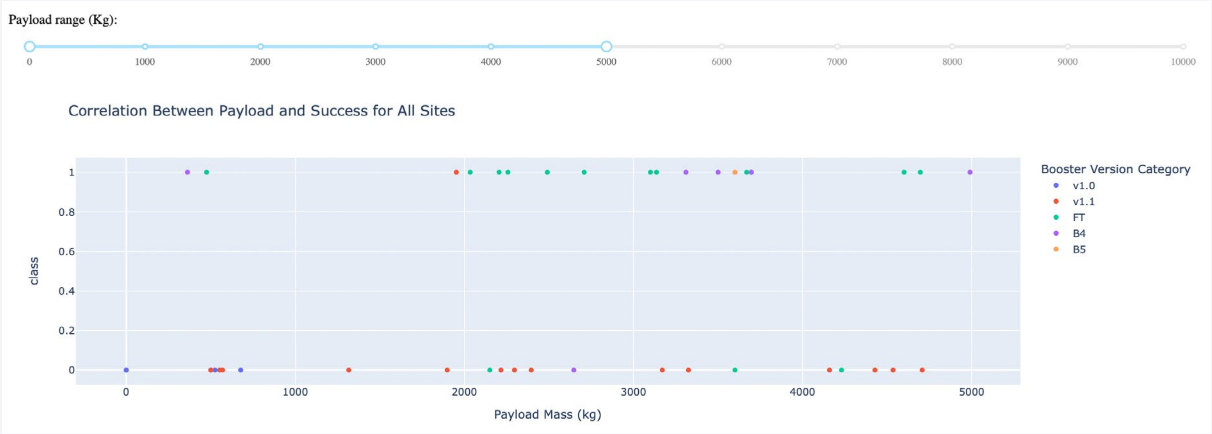
Highest Lauch Success Ratio

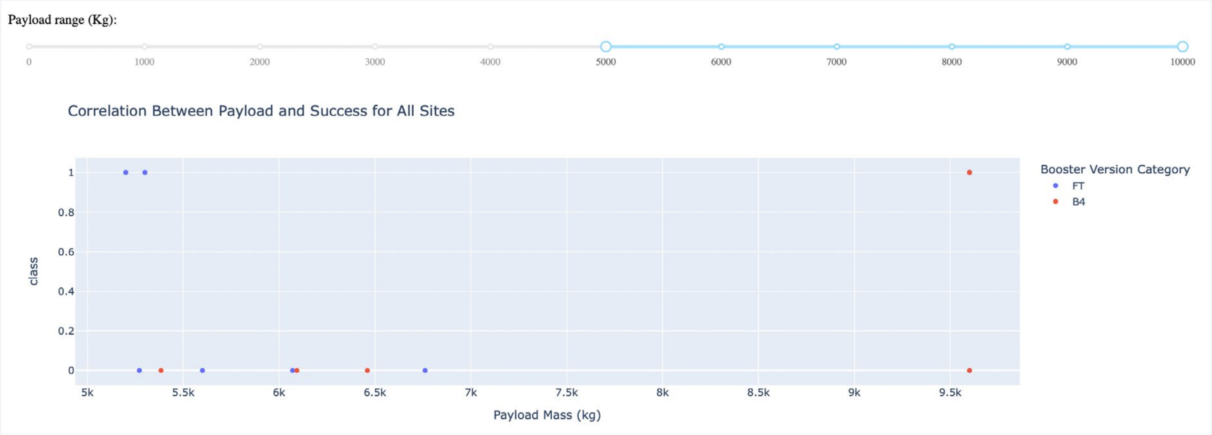


KSC LC-39A also has the highest success rate at 76.9%

Payload Mass and Launch Outcomes

Payloads between 2000 and 5500 kg have the highest success rate.





Section 5

Predictive Analysis (Classification)

Classification Accuracy

From the full set results, we can conclude that the Decision tree model is the best model.

It has the highest scores across the board compared to the others.

It should be noted that the test set results are poor. Due to a small sample size.

Test set

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

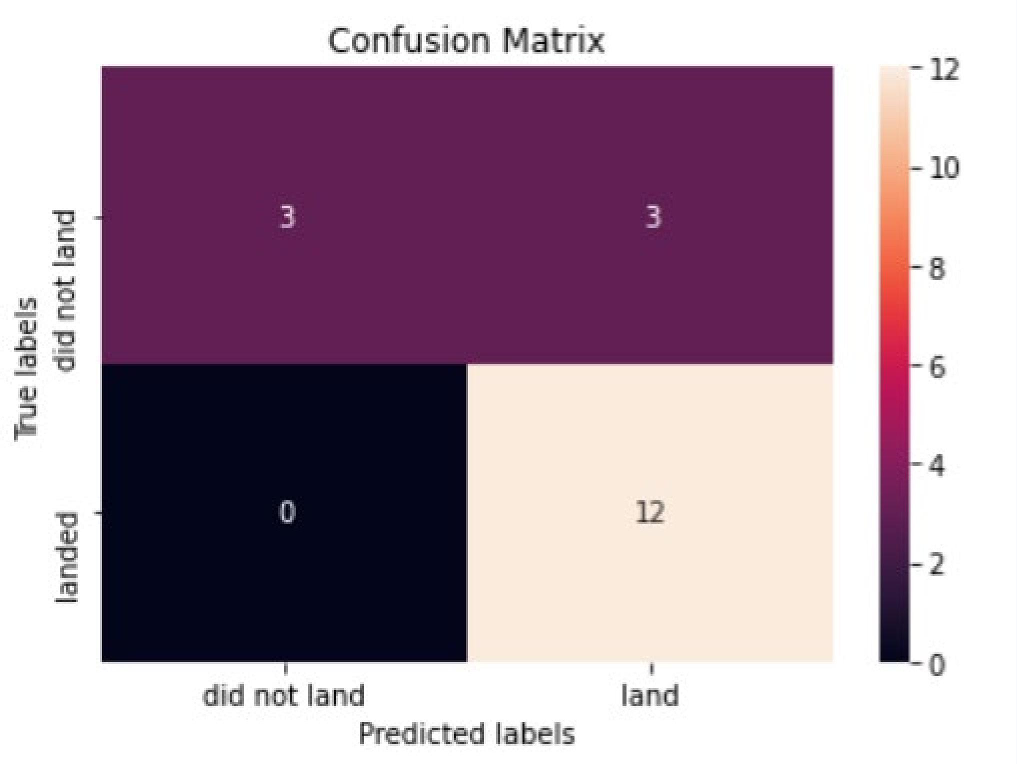
Full set

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

Confusion Matrix

The confusion matrix reveals that logical regression effectively differentiates between classes.

The issue is the high rate of false positives.



Conclusions

- The decision tree model is the best model
- Most Launches are coastal. With a fair distance from civilian infrastructure
- KSC LC-39A has the highest success rate of all the launches.
- Orbits GEO, HEO, SSO and ES-L1 have a 100% success rate.

Thus, as space Y , we should use the decision tree model in our predictive analysis. Moreover, we should launch near coastal areas away from civilian infrastructure. To increase our success chances we should use site KSC LC-39A and lean towards orbits with 100% success rates if possible.

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!