IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sara Ržen
20th May, 2025

Executive Summary

Summary of methodologies:

1. Data Collection
2. Data Wrangling
3. Exploring data using SQL
4. Visualising Data
5. Analysing Data using interactive visual analytics = Folium, Dash
6. Predictive Analysis = Machine Learning

Conclusion

Introduction

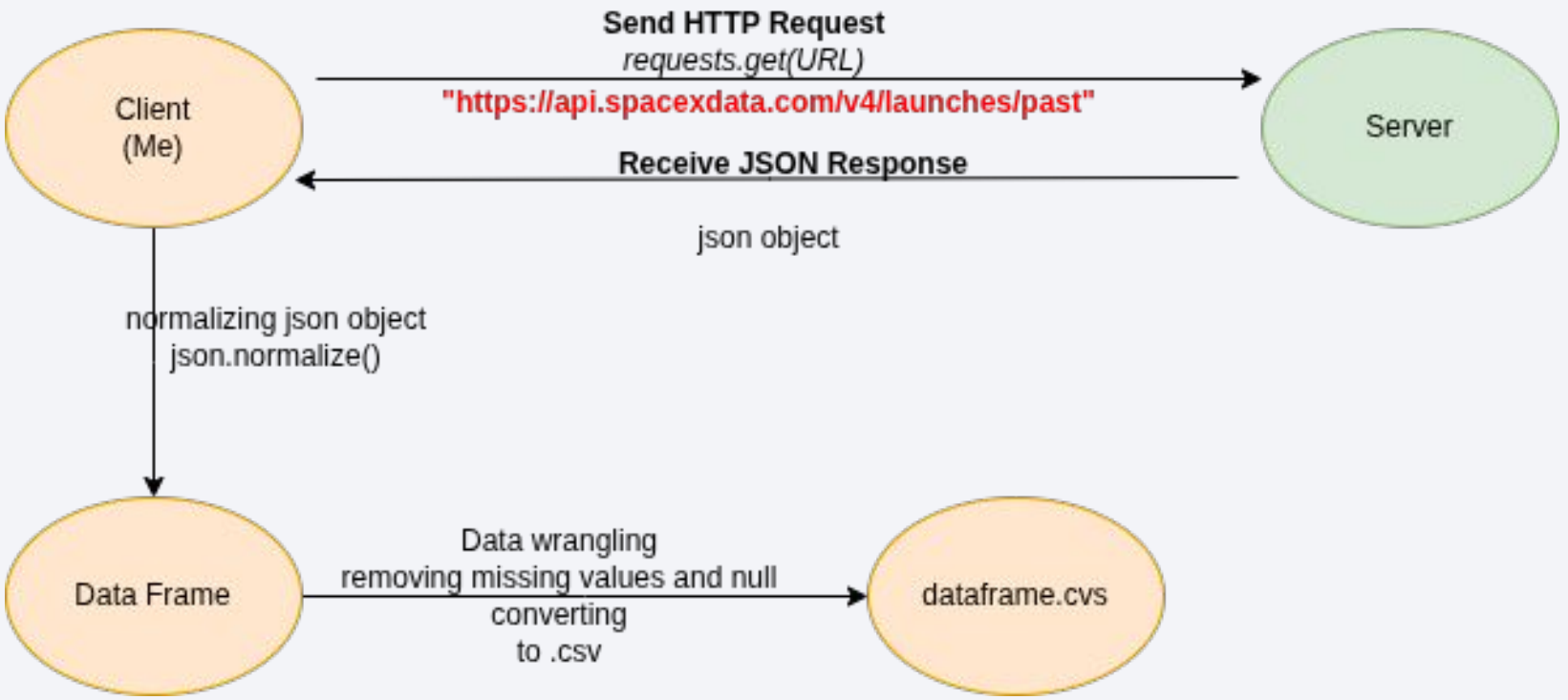
- In this capstone project, we analyse SpaceX data trying to predict if the Falcon 9 first stage (Booster Versions) will land successfully.
- Based on their cost saving methods and provided data, we determine best payload mass, best booster version, launch site & orbit destination.
- We have role-played as a competing alternative company SpaceY, who is interested in launching their own rockets.

Section 1

Methodology

Data Collection - SpaceX API

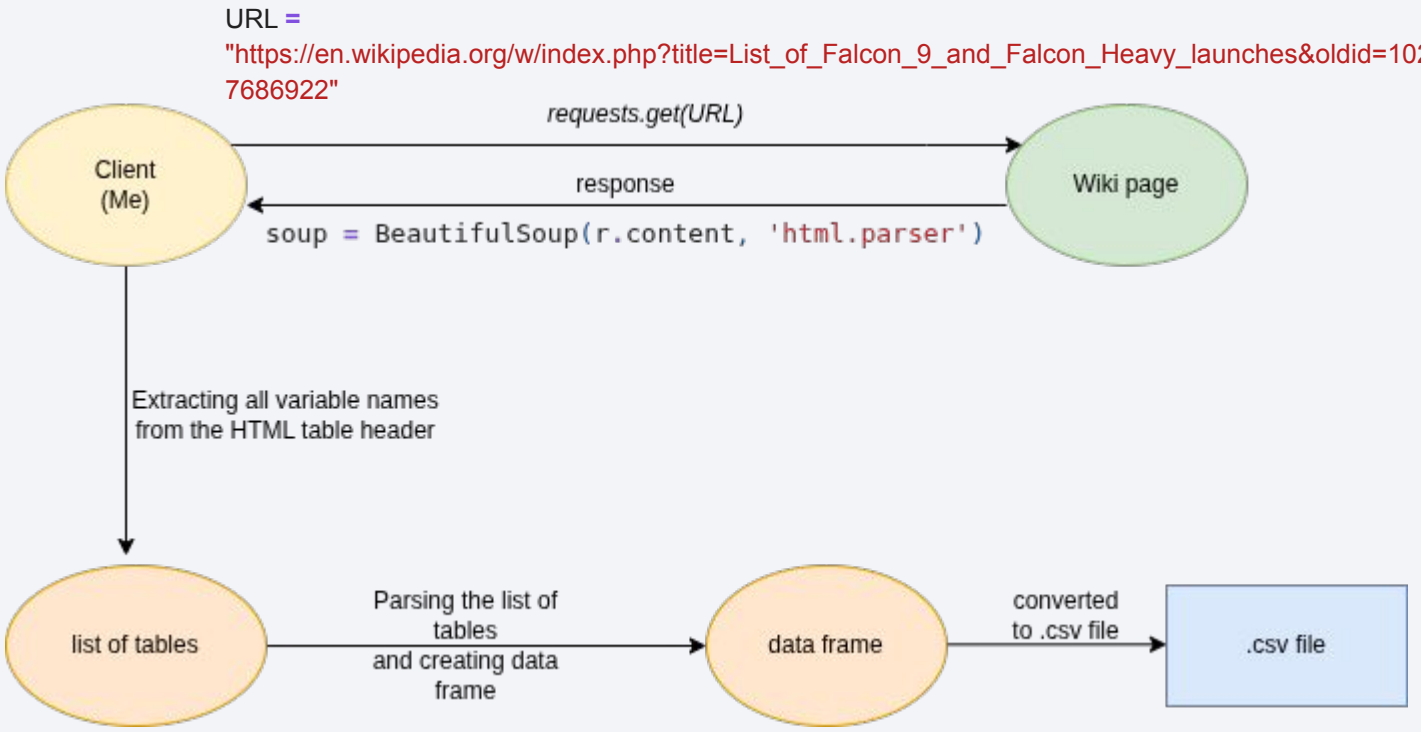
- Data was collected using REST API, scraping Wiki page on Falcon 9 launches
- Result was view by calling .json() method.
- JSON object was then converted into a dataframe and then normalized by json_normalize function.



GITHUB

Data Collection - Scraping

- Next task was to scrape Wiki pages and obtaining Falcon 9 Launch data. By using BeautifulSoup package I have successfully scraped HTML tables found on Wiki page.
- After scraping, next task was to extract all variable names from header of HTML tables.
- After scraping, I have created a data frame by parsing the launch HTML tables.
- At this step, I have also replaced and removed al N/A or NULL values for next step.
- Last step was to convert dataframe to .csv file.



GITHUB

Data Wrangling

- At this step we get to look at the data and get to know and identify key values for later steps.
- With this step we ensure, that our data has no null or missing values. If so, we replace or remove those values, since they could influence our data set.
- We have identified, that variable ‘LaunchSite’, was one of the key variables of our analysis.
- We have created a landing outcome label from Outcome column

GITHUB

EDA with Data Visualization

- Exploratory Data Analysis was done by using Pandas, Matplotlib, NumPy and seaborn libraries.
- With visualization we try to see if there is any correlations between different variables.
- We have plotted seven charts:
  - five categorical scatter plots, to determine correlation and relationships between the variables:
    - Payload Mass & Flight Number, hue='Class'
    - Launch Site & Flight Number, hue='Class'
    - Launch Site & Payload Mass, hue='Class'
    - Launch Site & Flight Number, hue='Class'
    - Flight Number & Orbit, hue='Class'
    - Payload Mass & Orbit, hue='Class'
  - one bar chart, which showed success rate of each orbit

GITHUB

EDA with SQL

Summary of the SQL queries:

- Display :
  - unique names
  - Where clause
  - “like” string
  - Distinct clause
  - Order by Clause
  - Limit
  - Group by
  - Order by
  - substring

GITHUB

Build an Interactive Map with Folium

- Summarize map objects, that were added to a map:
  - First circle was added to a location of NASA Johnson Space Center, on top of it circle, was added a marker with a title
  - For each launch location there was added a circle and a markup to add the name of the launch location
  - For each launch location there were added MarkerClusters, which showed how many successful and failed launches each launch location had.
  - Lastly, we added lines which connected launch site with proximities, like nearest city, coast and railway. On top of the lines, we added a markup with calculated distance between the two objects.

GITHUB

Build a Dashboard with Plotly Dash

- Key feature with this task was to show two graphs, pie chart showing launch sites and success rate and failure rate. Second graph is a scatter plot showing relationship between payload mass and outcome.
- Dropdown list allows us to change between launch sites and payload-slider changes corresponding to sliding button left and right, so we can see the whole scope of payload mass and success attached to it.

GITHUB

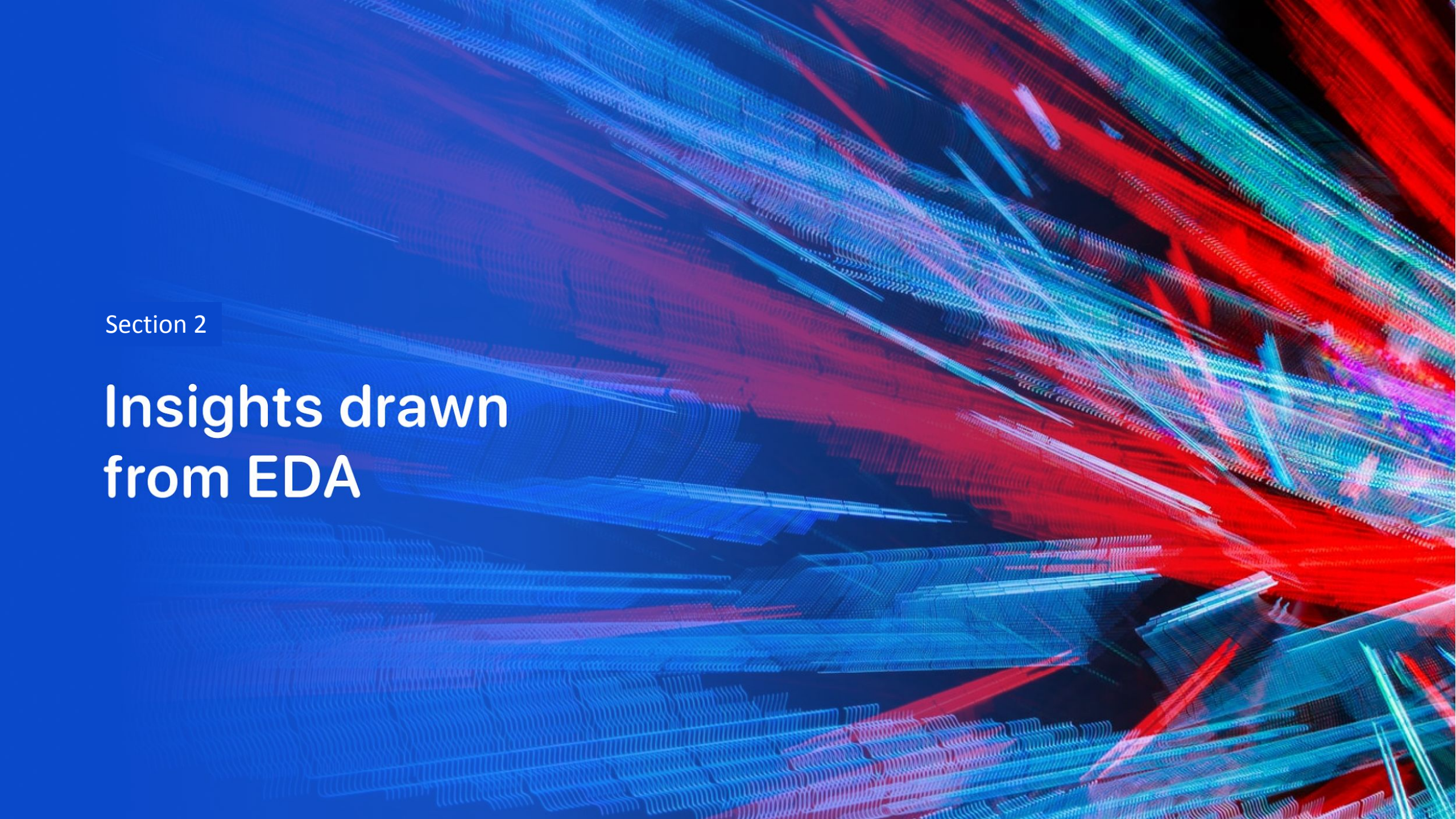
Predictive Analysis (Classification)

1. step was to create a NumPy array from the column Class in data, since our goal was to check if the success rate will be high or low, assigning variable Y.
2. step was to standardize data set and assign it to a variable X.
3. step was to split our data into a training and test sets. Our test set was a 20% of our data, stratifying Y.
4. Then we did four different classification models using GridSearchCV technique.
    - logistic regression model
    - support vector machine object
    - decision tree classifier object
    - k-nearest neighbours
5. For each model we also did confusion matrix and we tested the accuracy of test data using score method.
6. Lastly, based on test set accuracy, Logistic Regression was best.

GITHUB

Section 2

Insights drawn from EDA

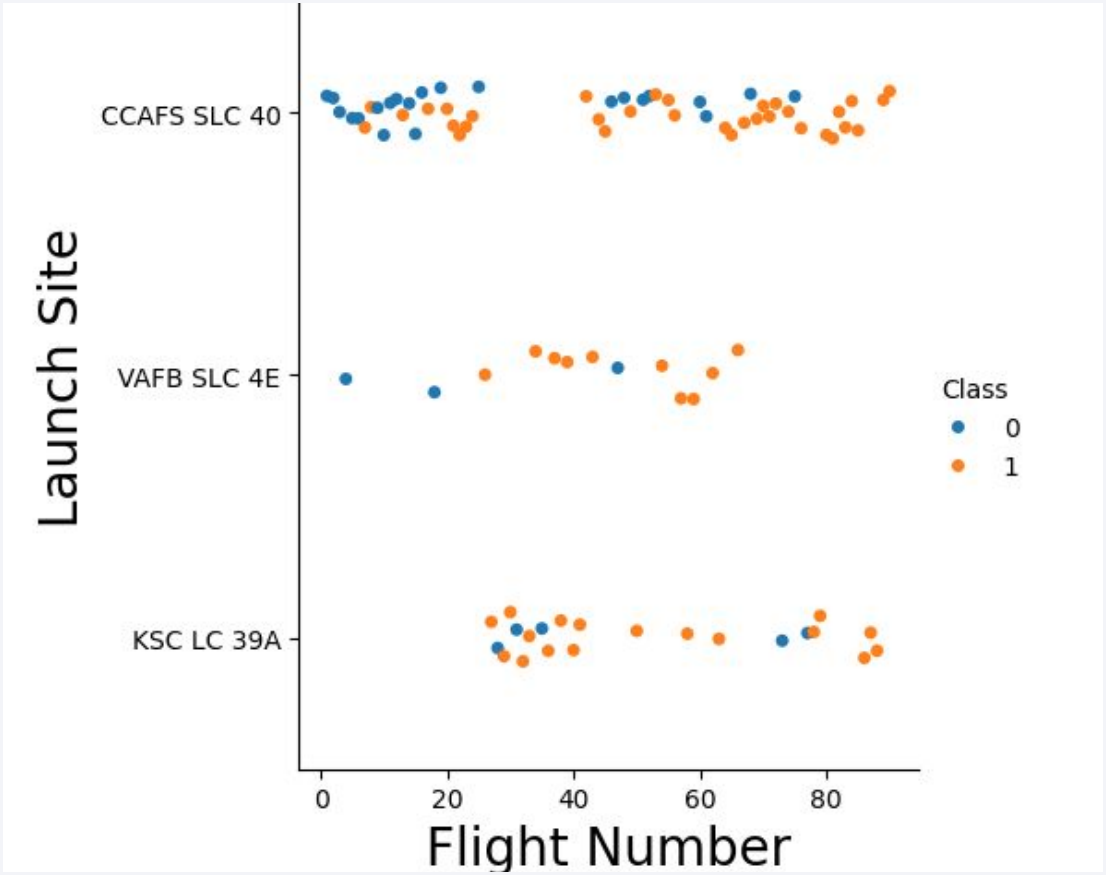


Flight Number vs. Launch Site

We can see three different Launch Sites and the relationship regarding Flight Number variable and success rate. As we can see, as the time progressed and more and more launches were made, we can see the density in the upper and lower part of chart, showing us that early failures occurred at all pads, but the steady improvements came with experience. CCAFS is the most successful pad.

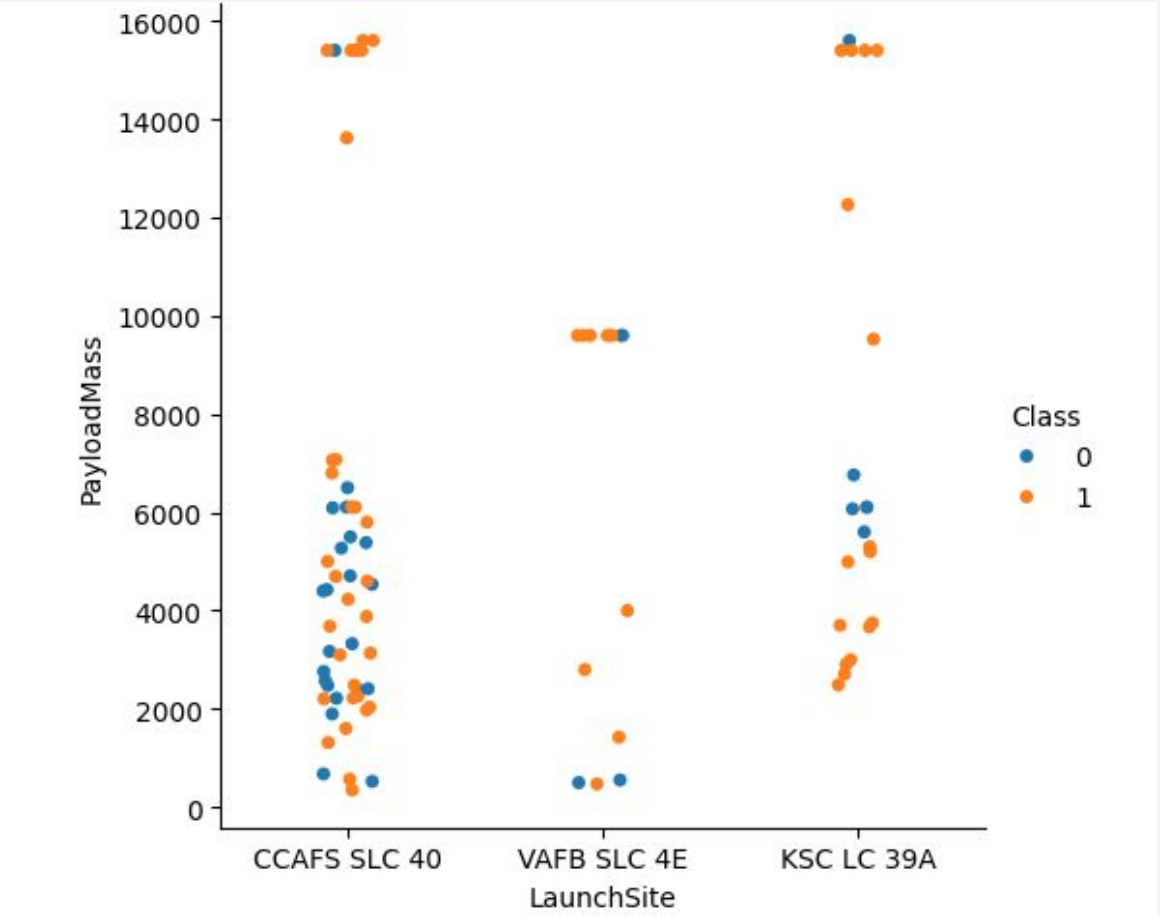
Bottom line:

From the 30th flight onward, there was almost a guaranteed success no matter witch pad was used, as the rocket design probably got more “broken in”.



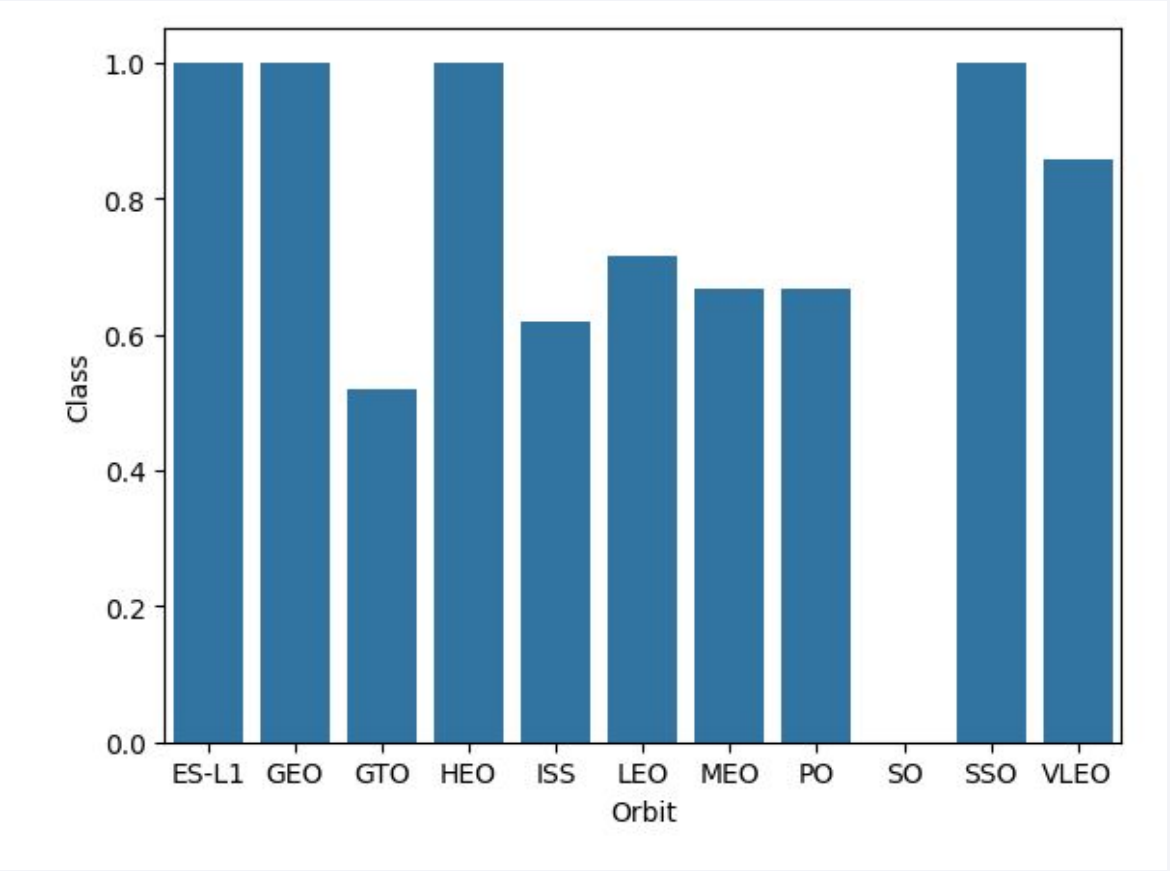
Payload vs. Launch Site

We can see that for VAFB SLC 4E there was no heavy payload mass (greater than 10000).



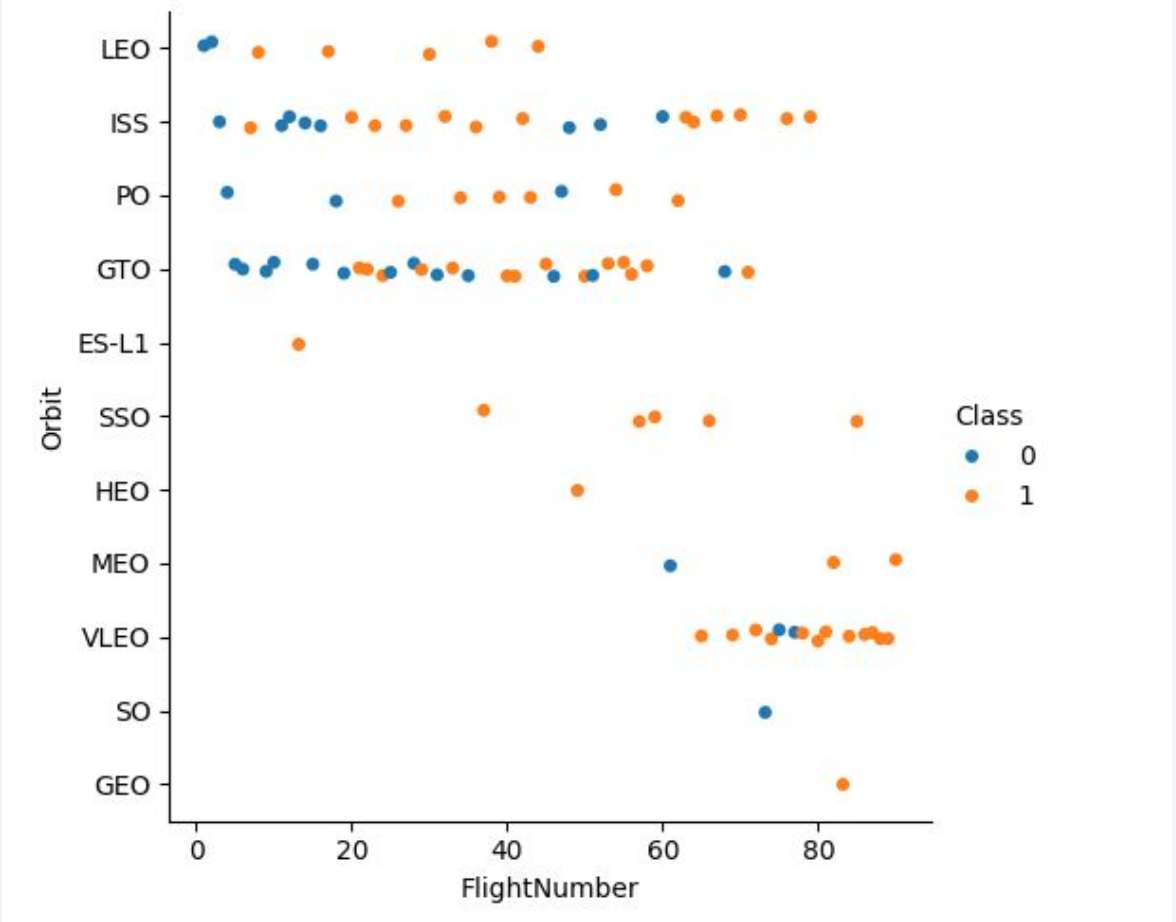
Success Rate vs. Orbit Type

There is a relationship between success rate and orbit type. There is a very strong correlation between success and ES-L1, GEO, HEO & SSO orbit.

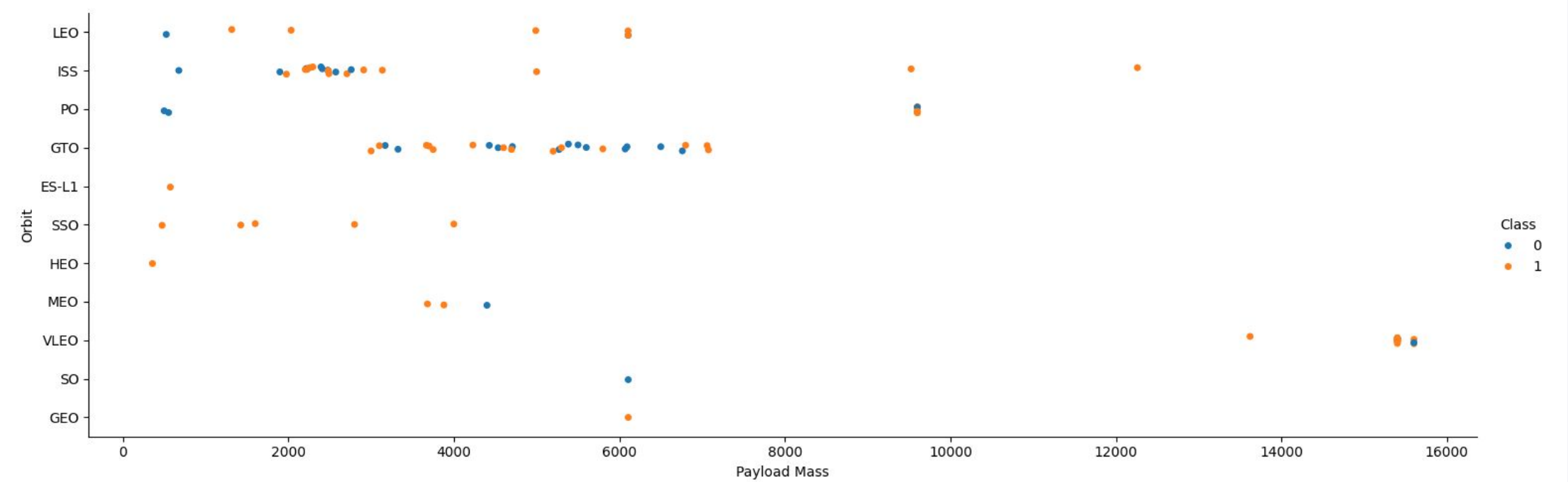


Flight Number vs. Orbit Type

We could see that in the VLEO orbit the Success apperats related to the number of flights; on the other hand there seems to be no relationship between flight number when in GTO orbit.



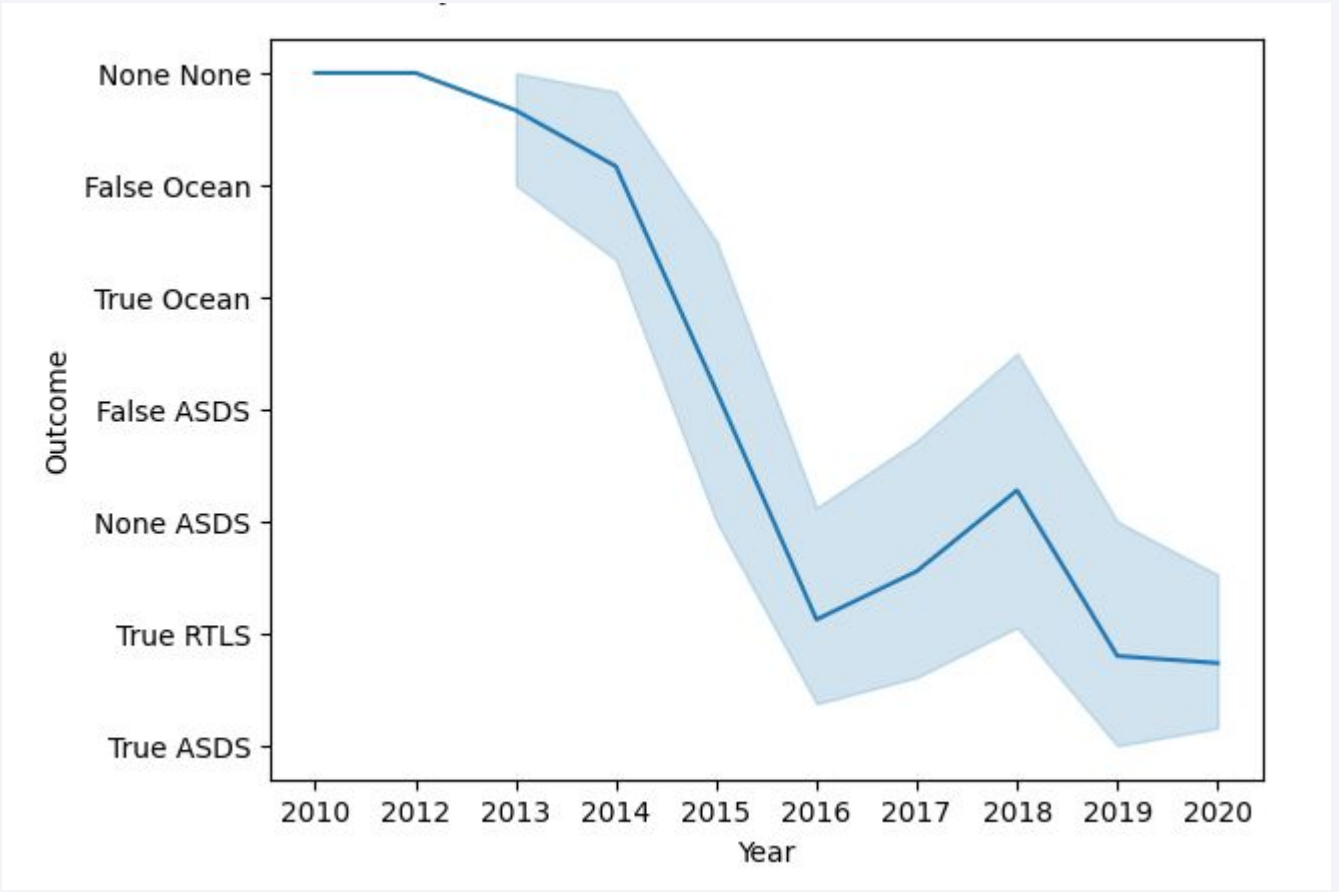
Payload vs. Orbit Type



As the payload mass increases there is also less and less successful landings. With heavy payload the successful landing or positive landing rate are more for Polar, LEO & ISS.

Launch Success Yearly Trend

We can observe that the successful landings started in 2013 with stabilization in 2014 and kept increasing until 2017.



All Launch Site Names

In data visualization step, we were working with data of four different launch Sites: CCAFS SLC-40, CCAFS LC-40, VAFB SLC-4E, KSC LC 39A.

```
array(['CCAFS LC-40', 'VAFB SLC-4E', 'KSC LC-39A', 'CCAFS SLC-40'],
      dtype=object)
```

Launch Site Names Begin with 'CCA'

First 5 records of launch sites that begin with ‘CCA’ were CCAFS LC-40. 4 out of 5 were made by NASA. First two made in 2010, next two in 2012 and the last one in 2013. Launches were all successful, but none of the landings.

Total Payload Mass

Total payload mass carried by booster launched by NASA (CRS) was 99980 KG.

Average Payload Mass by F9 v1.1

Average payload mass carried by booster version F9 v1.1 is 2928.4 KG.

First Successful Ground Landing Date

The first date with successful landing outcome in ground pad was 2015-12-22.

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 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

List of the total number of successful and failure mission outcomes:

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

Boosters Carried Maximum Payload

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

List of all the booster_versions that have carried the maximum payload mass:

2015 Launch Records

List of the records which displays the month names, failure landing_outcomes in drone ship, booster versions, launch_site for the months in year 2015.

month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

List of ranked landing outcomes between 2010-06-04 and 2017-03-20:

Landing_Outcome	total_count	outcome_rank
No attempt	10	1
Success (drone ship)	5	2
Failure (drone ship)	5	2
Success (ground pad)	3	4
Controlled (ocean)	3	4
Uncontrolled (ocean)	2	6
Failure (parachute)	2	6
Precluded (drone ship)	1	8

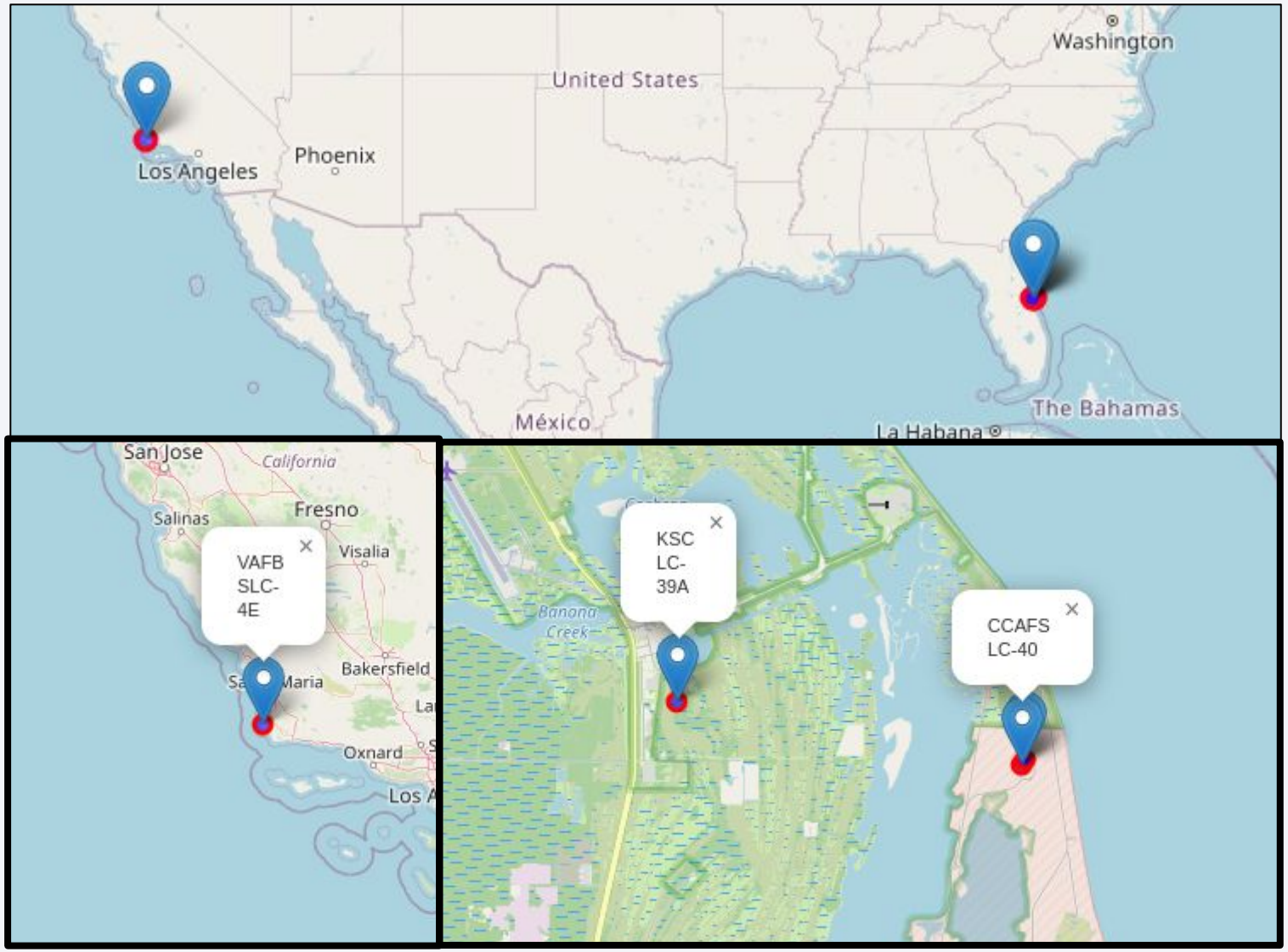

Section 3
Launch Sites
Proximities Analysis

Launch sites

Upper picture: launch site location.

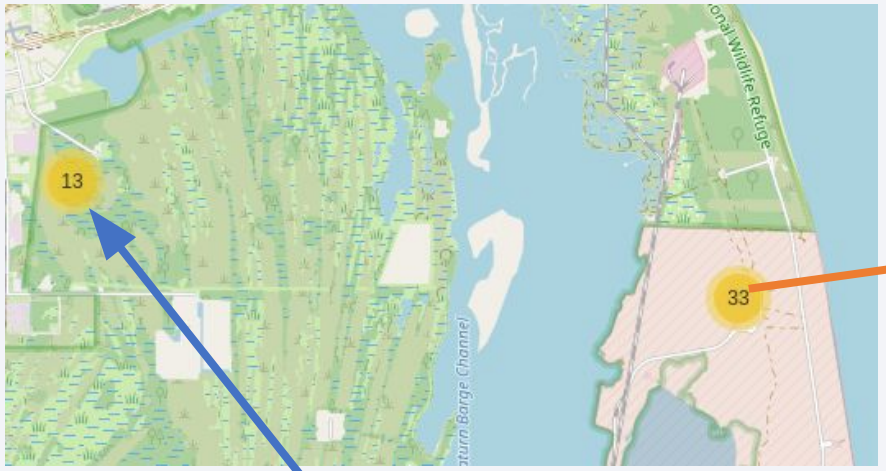
Bottom left: Launch site on west coast.

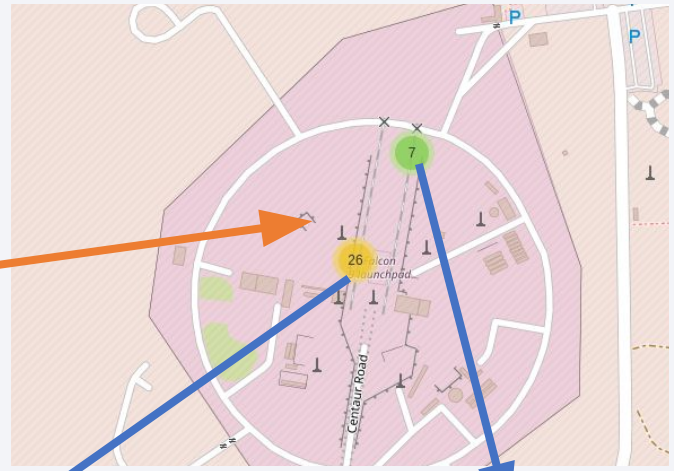
Bottom right: Three launch sites on east coast (CCAFS SLC-40 hidden).



Successful and failed launches for each launch site

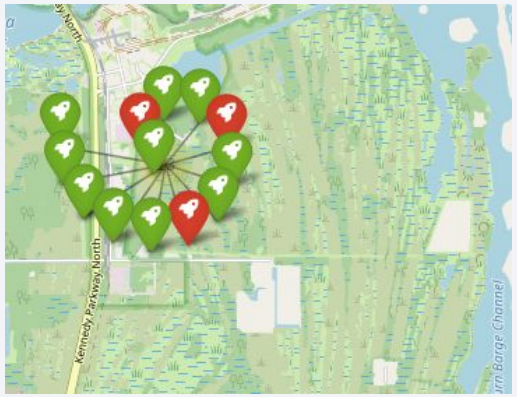
VAFB SLC-4E west coast:





10 launches, 40% success rate.

KSC LC-39A east coast:



13 launches, 77% success rate.

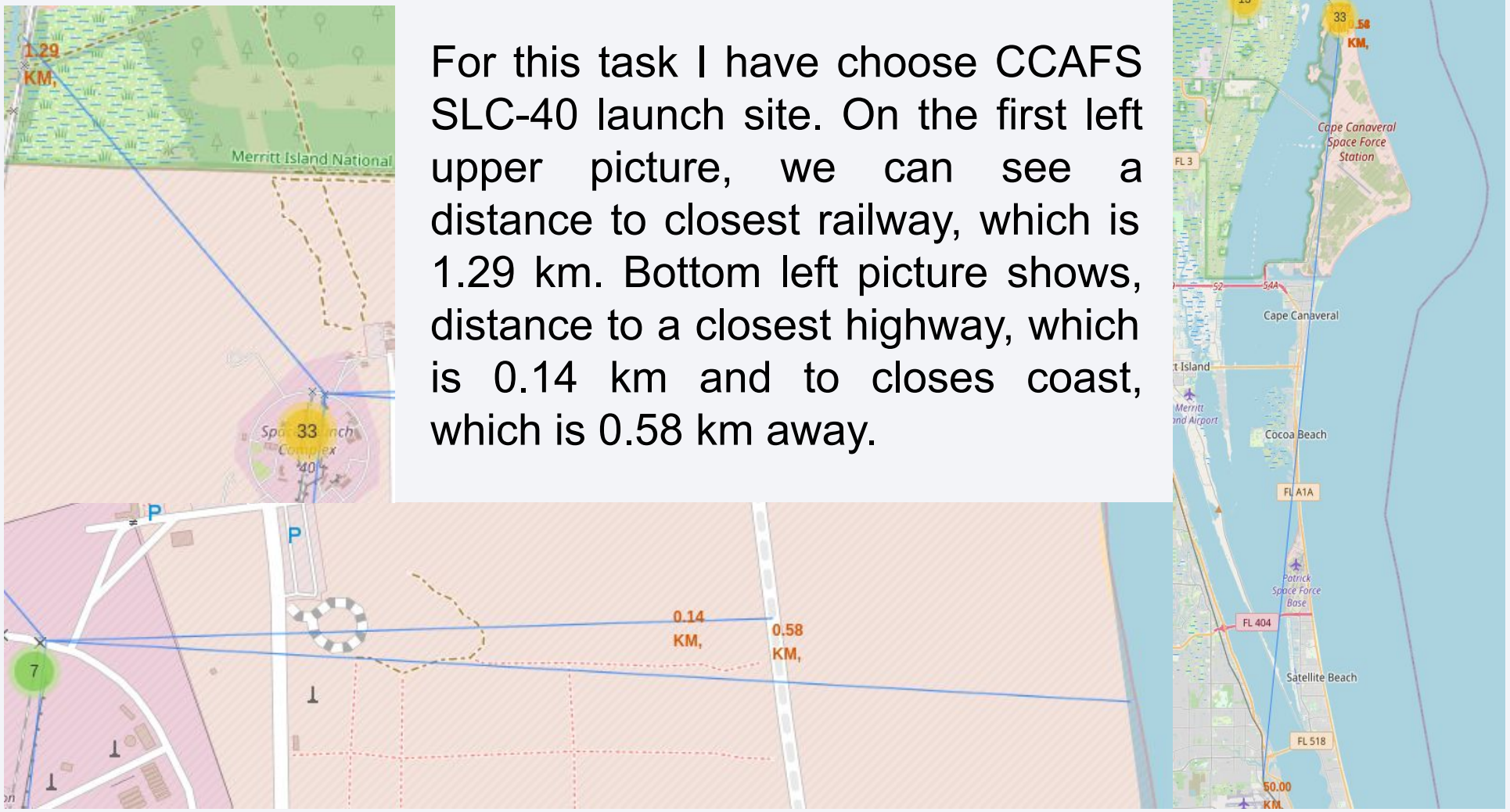
CCAFS LC-40:

26 launches, 27% success rate.

CCAFS SLC-40:

7 launches, 42% success rate.

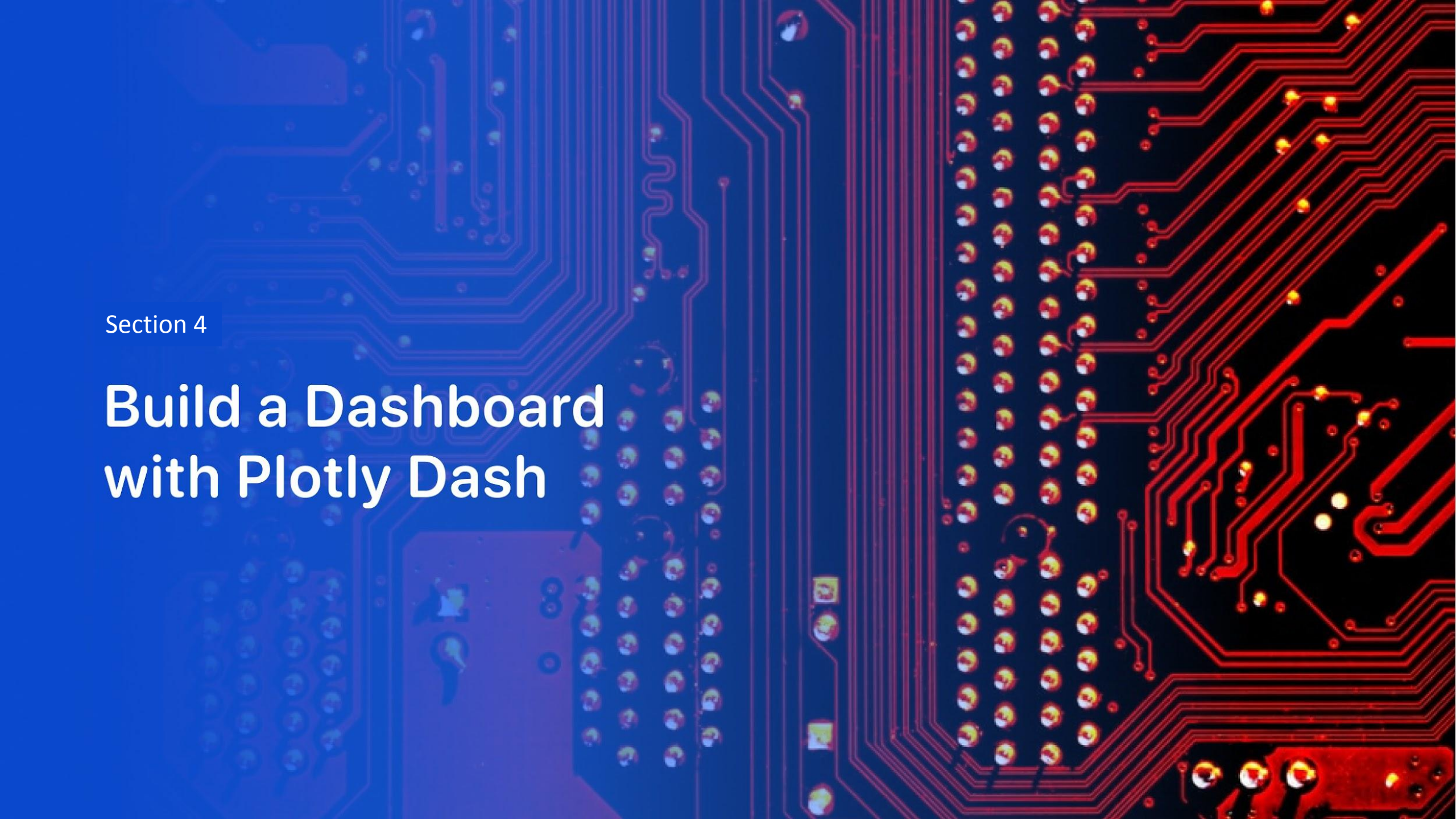
Distance between a launch site to its proximity



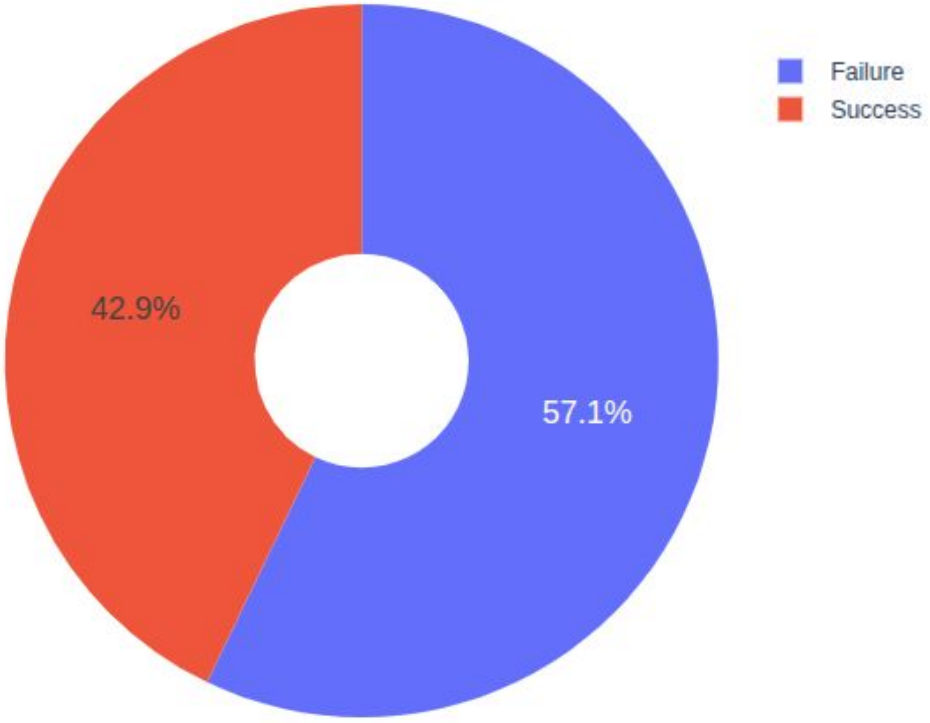
For this task I have choose CCAFS SLC-40 launch site. On the first left upper picture, we can see a distance to closest railway, which is 1.29 km. Bottom left picture shows, distance to a closest highway, which is 0.14 km and to closes coast, which is 0.58 km away.

This picture shows distance to a closest city, which is Melbourne, FL. 50 km.

We can see that launch sites are far away from city, probably due to safety.

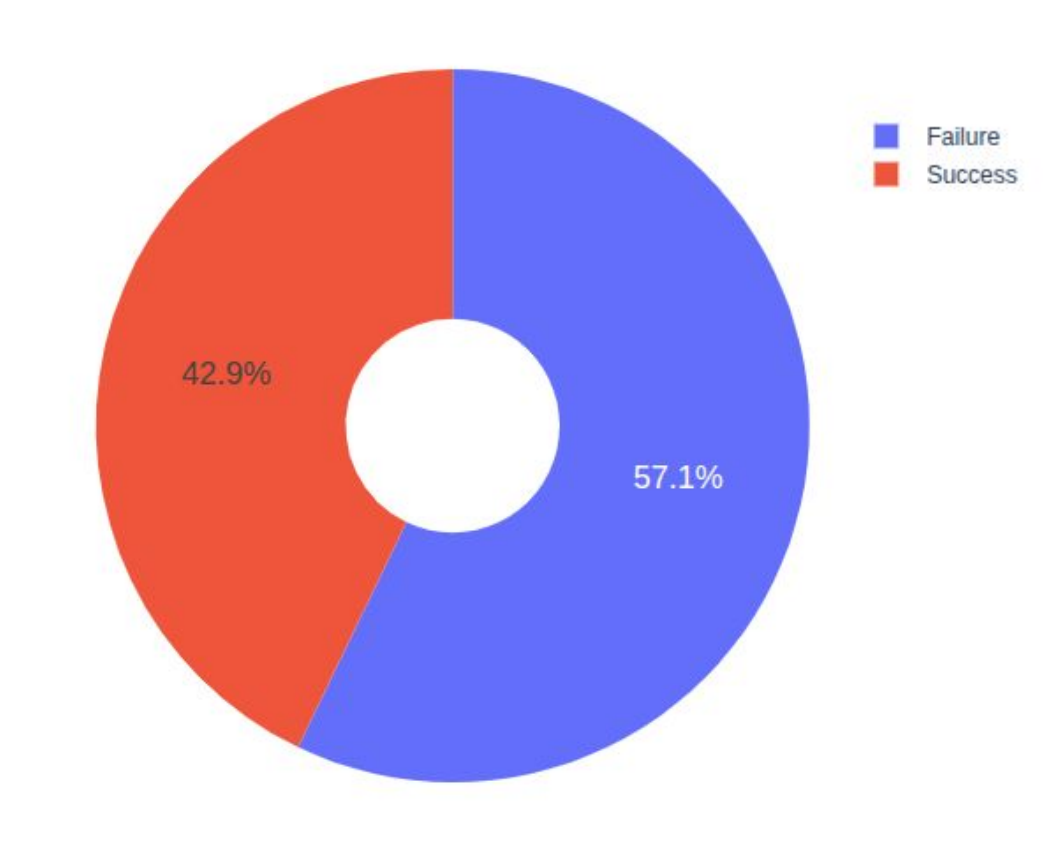

Section 4
Build a Dashboard with Plotly Dash

Launch Records



Pie chart shows a total success and failure rate of all launch sites combined. Success rate is 42.9% out of 56 noted launches.

Highest Success Rate Launch Site



The highest success rate was recorded at the launch site CCAFS SLC-40 and it was 42.9% out of 7 launches.

Payload vs. Launch Outcome

Payload range (Kg):

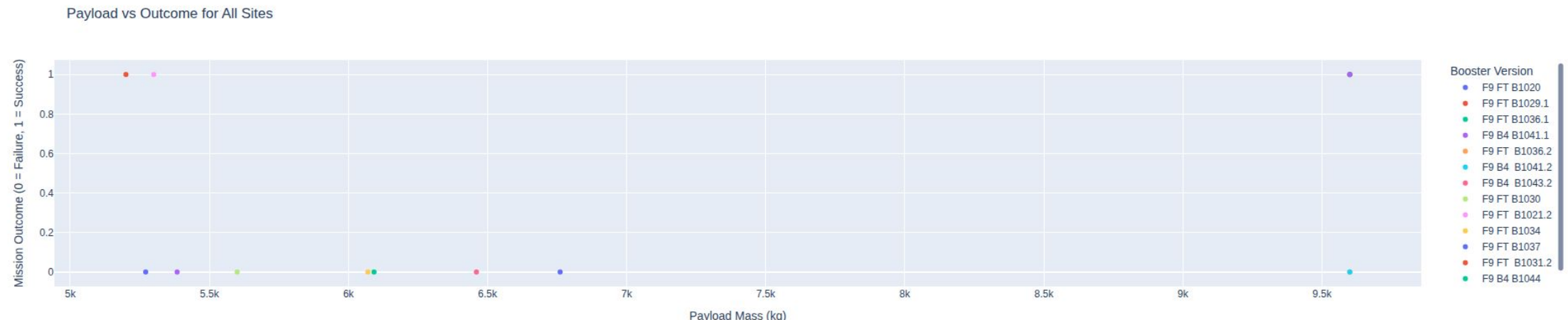
Payload range (Kg):

Payload range (Kg):

As payload mass increases success rate decreases. Range between 2000 kg and 5,500 kg yields best result.

Payload range (Kg):

0 100

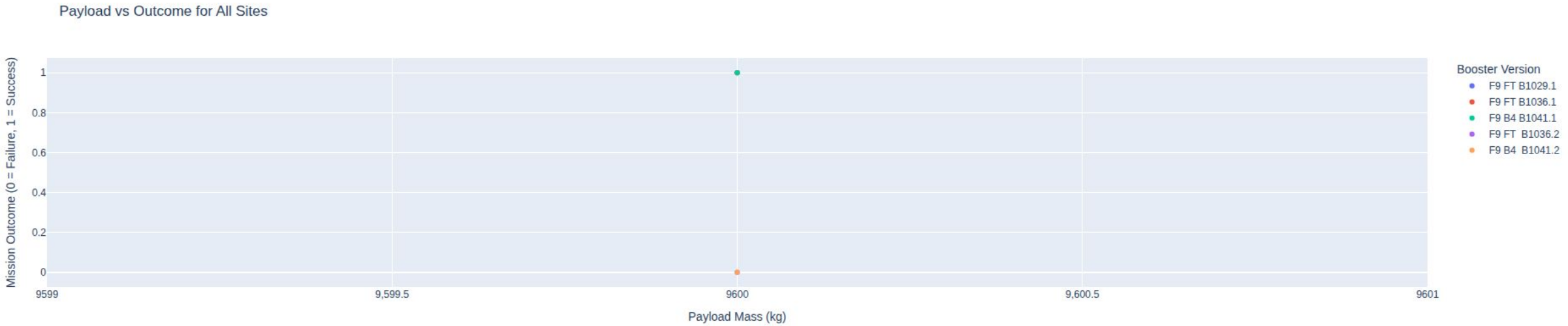


Payload range (Kg):

0 100

Payload range (Kg):

0 100



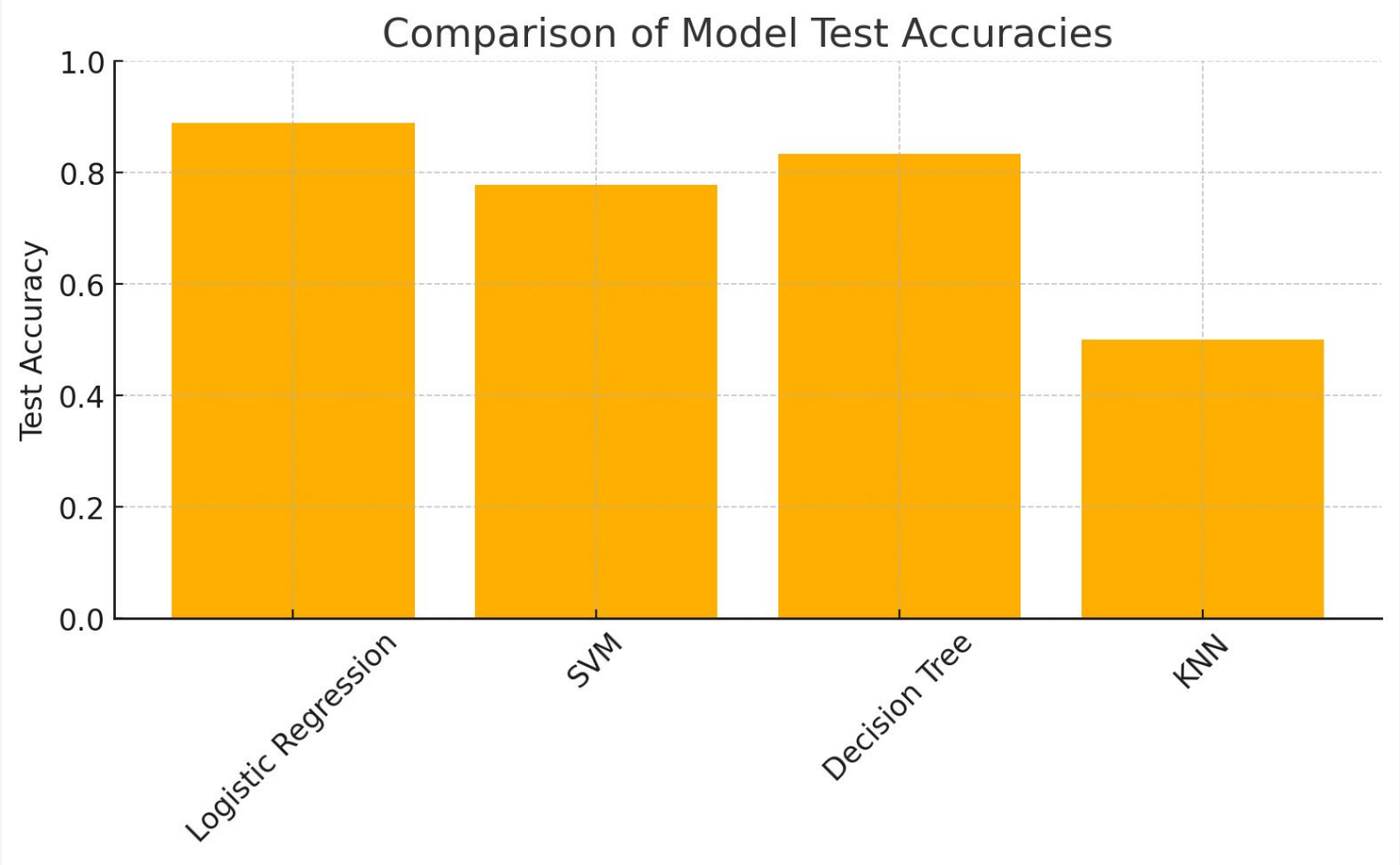
Successful max payload mass carried was by booster F9 B4 B1041.1 at 9600 kg. There was no noted launches after this point.

Section 5

Predictive Analysis (Classification)

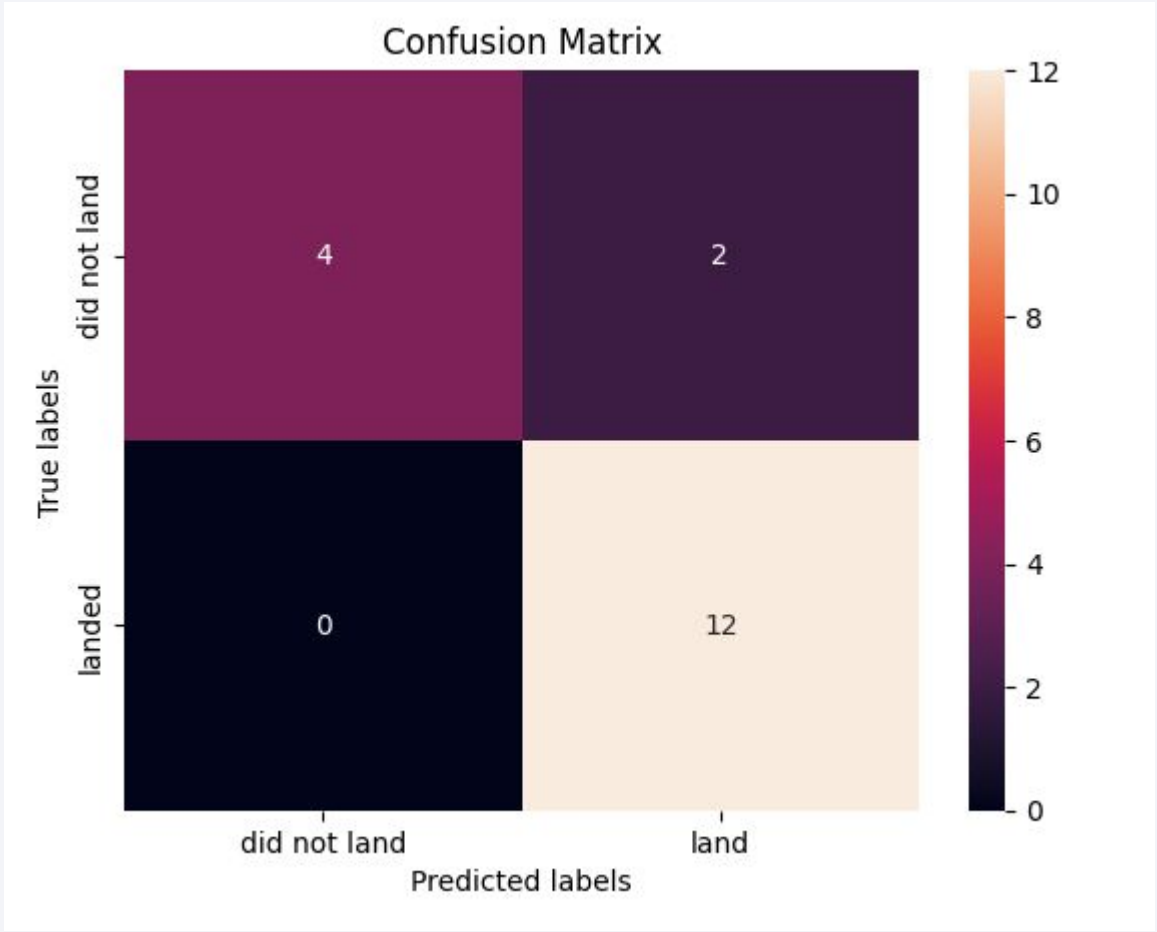
Classification Accuracy

Best model was Logistic Regression with a 88.9% accuracy.



Confusion Matrix

This model is excellent at finding actual landings (100% recall), but is sometimes over-predicts a landing when there wasn't one (2 false positives).
Overall accuracy is about 89%, with perfect sensitivity for landed flights but moderate specificity (≈67%) for "did not land".
True Positives = 12, 12 flights that truly landed were correctly predicted as landings. False Negatives = 0: No actual landings were missed - every landed flight was correctly identified.



Conclusions

- Successful landings started in 2013 and kept increasing until 2017
- Average payload mass carried by booster version F9 v1.1 is 2928.4 kg
- As we could see CCAFS was the most successful landing pad, best for payload mass from 2000 kg to 5500 kg
- Most successful destination orbits were ES-L1, GEO, HEO & SSO orbit
- From 4 launching sites, most successful is on east coast KSC LC-39A, with 77% success rate.
- All the launch sites are away from civilization

Appendix

GITHUB PROJECT REPOSITORY

Thank you!