IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Daniel Malfavón
06/28/2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Discussion
- Conclusion

Executive Summary

In this project established by IBM, the forecast will be developed if the SpaceX Falcon 9 first stage will be able to land correctly and satisfactorily based on the development of different algorithms for marching learning classification.

To do this, the project will delve into:

- Data collection, formatting and filtering.
- Exploring data analysis
- An interactive visualization of the data
- The development of Machine Learning for data prediction

Executive Summary

Based on what was analyzed, it will be defined whether there are correlations between the different variables present in the acquired data to find different patterns.

Finally, with these evaluations, machine learning methods such as decision trees or KNN will be developed to predict whether the SpaceX Falcon 9 will be able to land correctly.

Introduction

SpaceX, unlike many other companies in charge of launching rockets into space, they manage to reduce their launch costs by reusing the first propulsion stage of their constructions, so for them it is important to ensure that this first propulsion stage manages to land. in a satisfactory manner so as not to generate much expense.

Therefore, the problem is to determine if for the Falcon 9, based on many of the variables involved such as mass, type of orbit, or even the launch zone, it is possible to predict whether the first propulsion stage of the rocket will be able to land. satisfactorily.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Starting from the use of the SpaceX API and Web scraping
- Perform data wrangling
  - Using Jupyter notebooks
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Pandas
  - NumPy
  - SQL

Methodology

- Data visualization
  - Matplotlib
  - Seaborn
  - Folium
  - Dash
- Perform predictive analysis using classification models
  - Logistic Regression
  - Support Vector Machine (SVM)
  - Decision Tree
  - K-nearest neighbors (KNN)

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Data Collection

SpaceX API

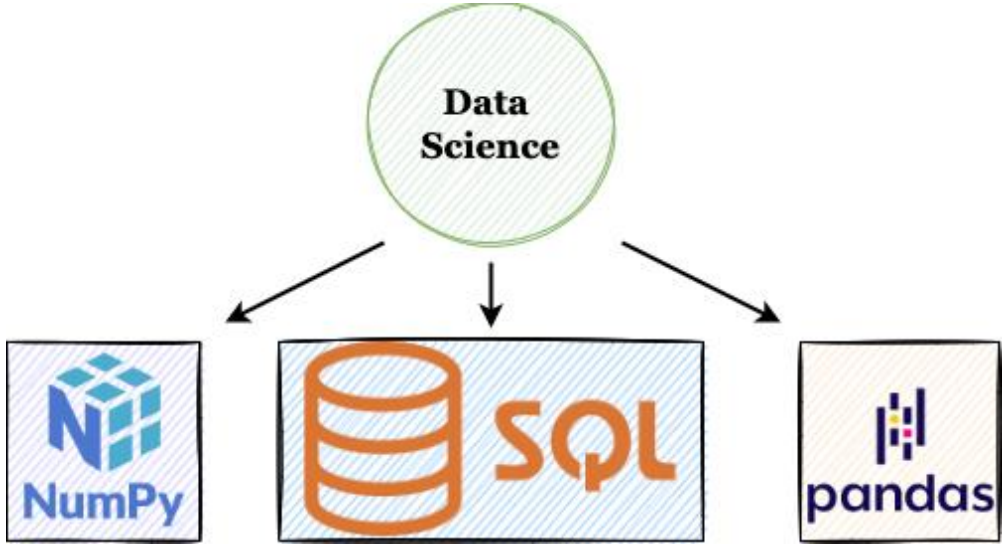
- The API used is: https://api.spacexdata.com/v4/launches/past
- Among the information that we can observe is the launch ID, date, rocket version, mass, type of orbit, among others. At the end of the data formatting, we are left with 90 rows and 17 columns of information.

Data Wrangling

- Subsequently, some adjustments were made to the data, first the empty data were eliminated depending on the case, if they were categorical, they were changed with the data with the highest frequency, and in the numerical continuous values they were replaced by the average of the rest of the data.
- Additionally, an extra column was added, which called 'Class' describes whether the landing was successful (showing a 1), or not, (0).
- Thus, our data represents 90 different launches with 80 columns, being information on 80 different characteristics.

EDA with Data Visualization

- In this part we are going to describe the tools used:
- Pandas and NumPy
  - These libraries are used to obtain information derived from our raw data.
- SQL
  - With it, it will allow us to isolate and extract data according to the limits or specifications that are predicted, such as the names of the models of the launch sites or the average mass that a certain propellant model weighs.



Build an Interactive Visualization

For this, they were used:

- Matplotlib and Seaborn
  - They are used to visualize data through scatter, bar and line graphs. All this to observe how the data is distributed and how it gives us the first look at how to predict its behavior
- Folium
  - This tool is used to visualize geographic data on a completely interactive map, allowing you to identify launch positions, references, and identify possible failure factors in launches, among other implications.

Build a Dashboard with Plotly Dash

- With Dash, we can generate a completely interactive site where we can select certain options as inputs in a drop-down menu to produce changes in the outputs, usually visual graphs such as pie charts, scatter charts, which show us the distribution of the data as well. as a correlation thereof.

Predictive Analysis (Classification)

Finally, once all the data has been obtained, we move on to the development of our machine learning models with the help of Scikit-learn. To do this, by standardizing and dividing the data into testing and training, it is possible to apply them in simulation models.

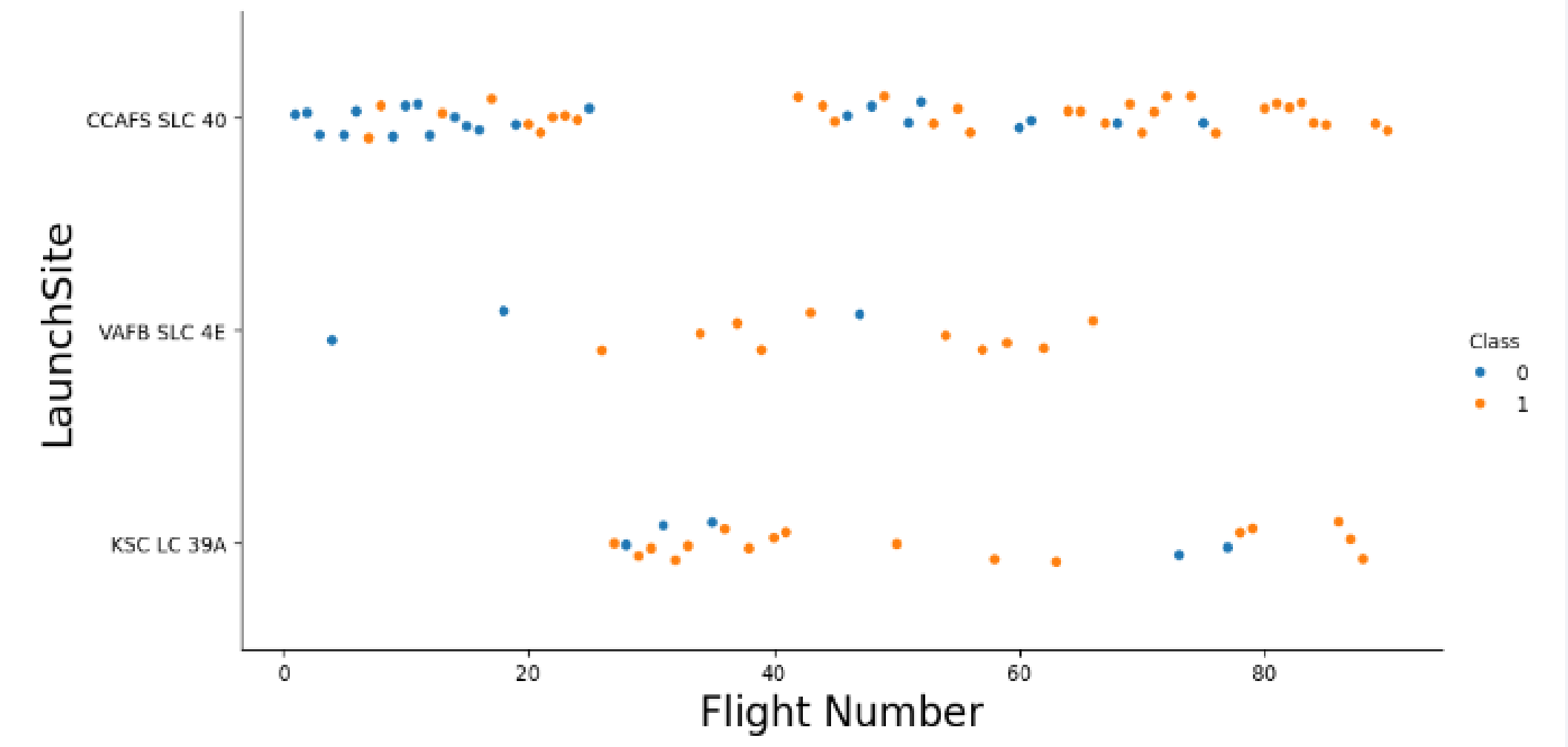
- Linear regression
- Support Vector Machine (SVM)
- Decision trees
- and K-nearest neighbors (KNN)

Thus, they were subsequently trained and tested to observe their capabilities and the accuracy score for each one.


Section 2
RESULTS

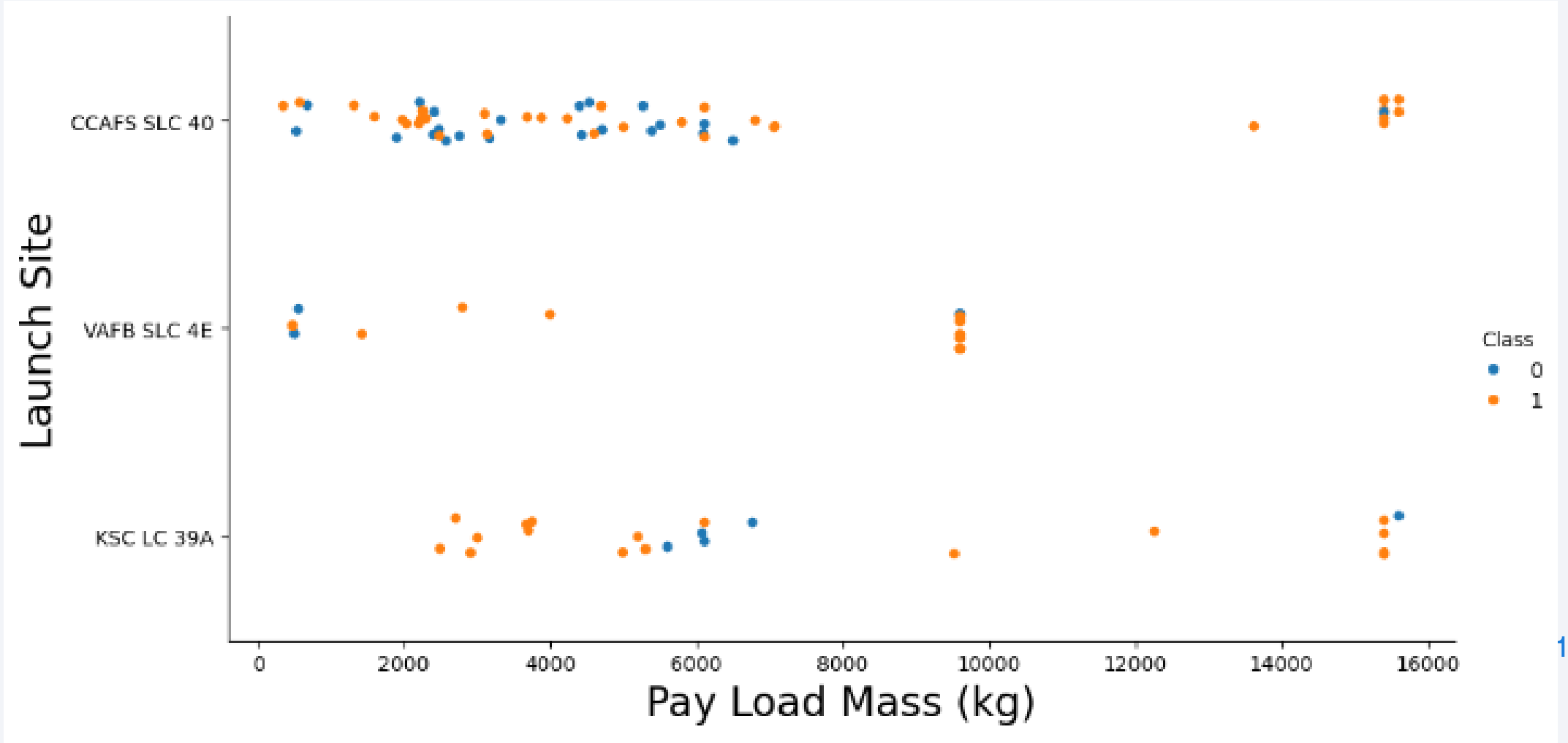
Flight Number vs. Launch Site

- The relationship between flight number and launch site



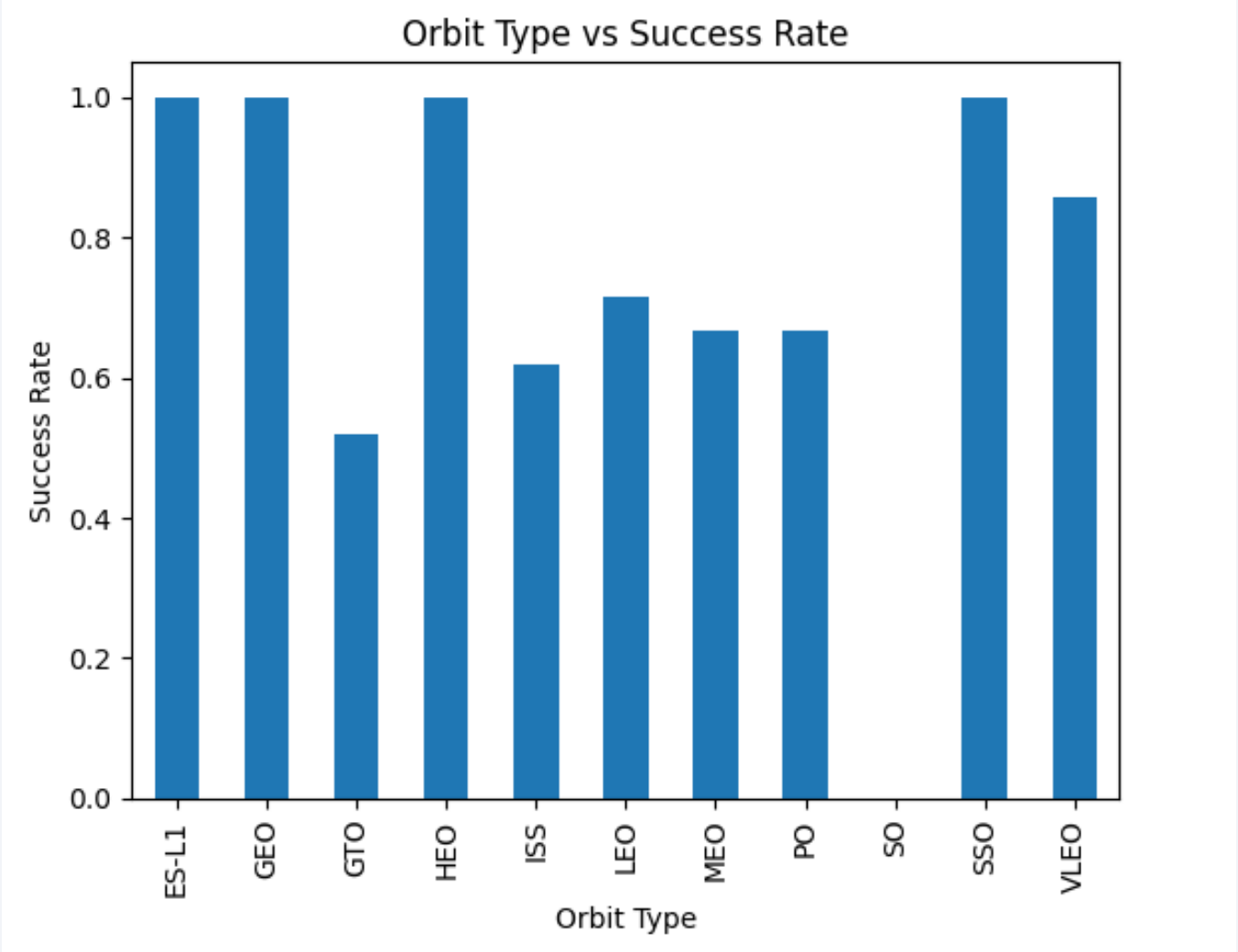
Payload vs. Launch Site

- The relationship between pay load mass and launch site



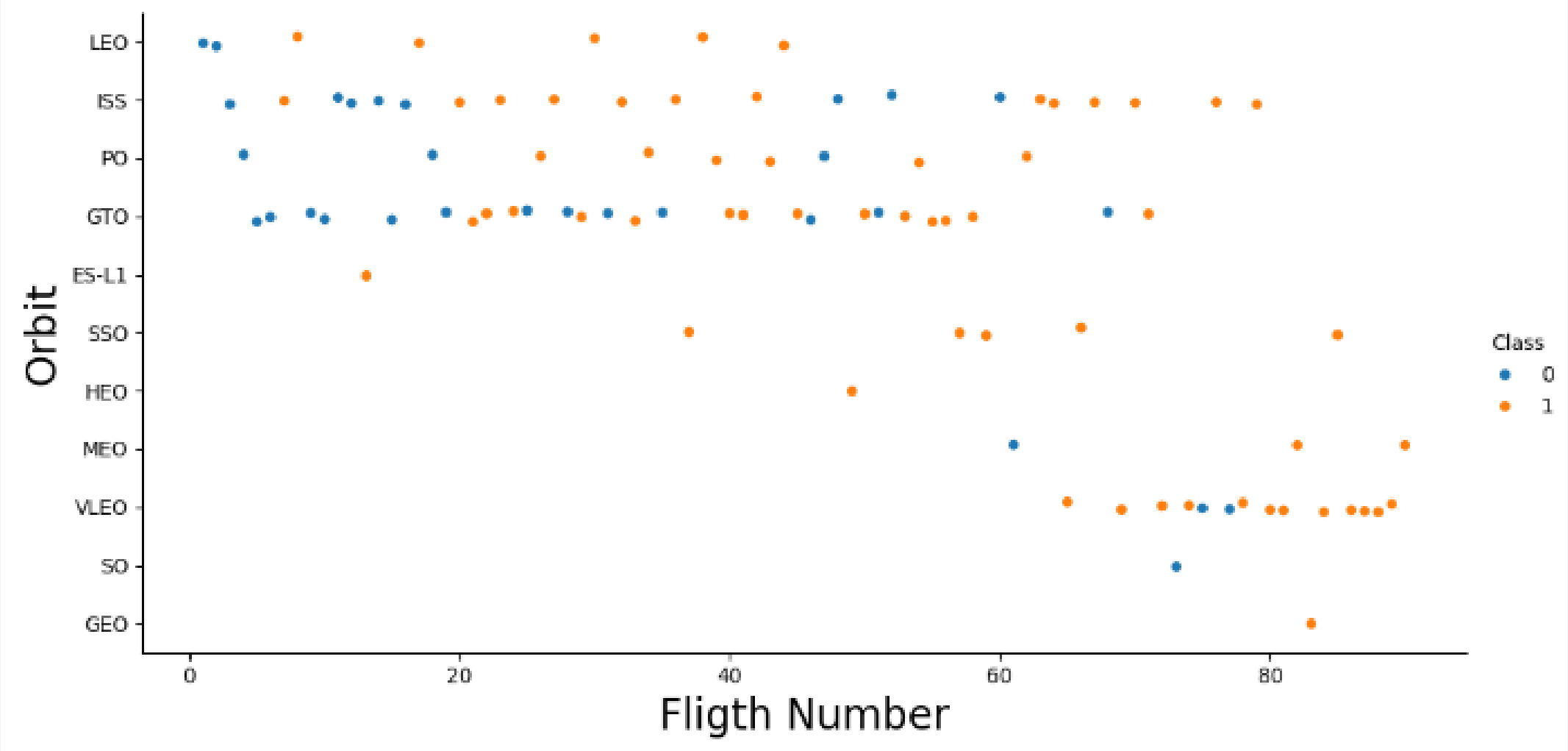
Success Rate vs. Orbit Type

- The relationship between success rate and orbit type



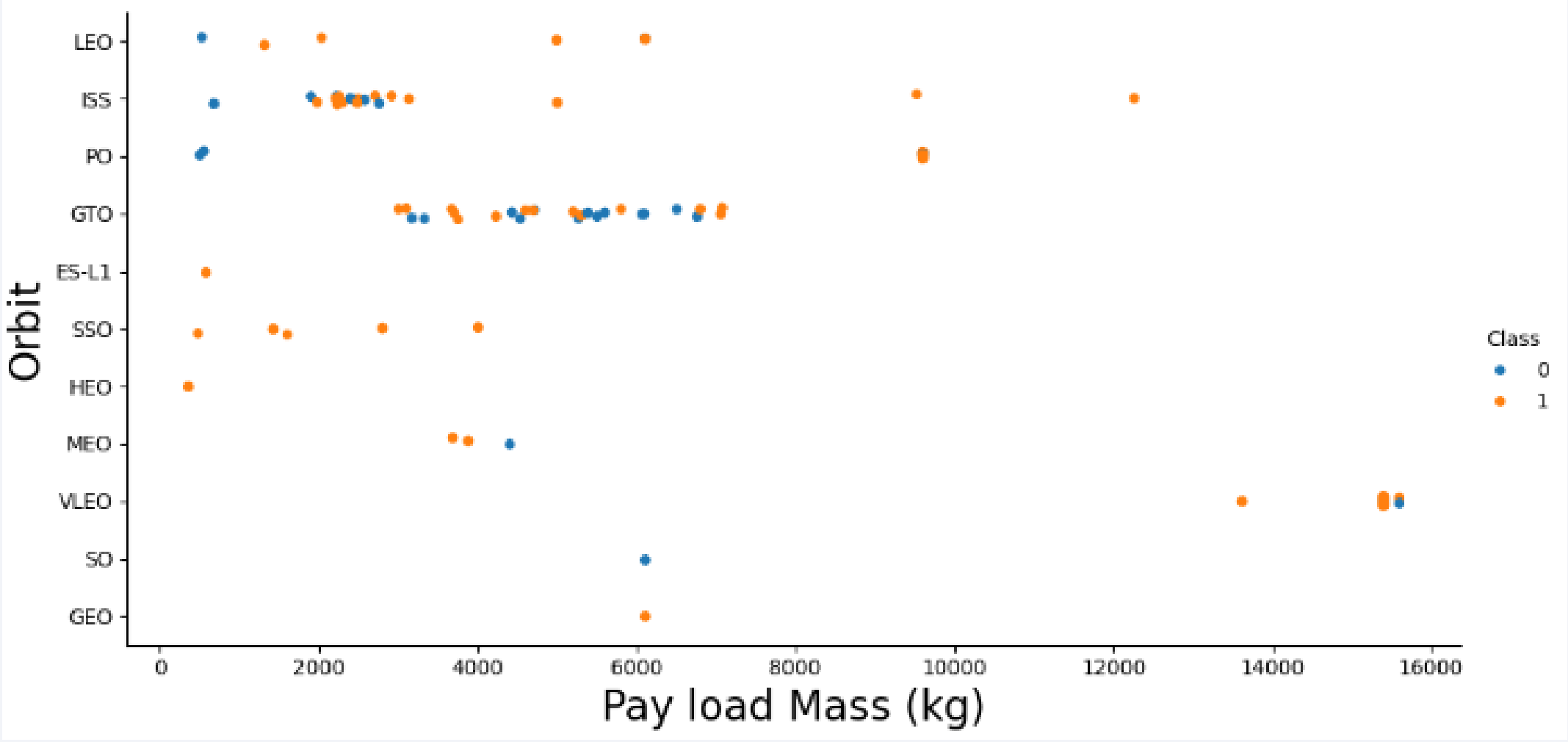
Flight Number vs. Orbit Type

- The relationship between flight number and orbit type



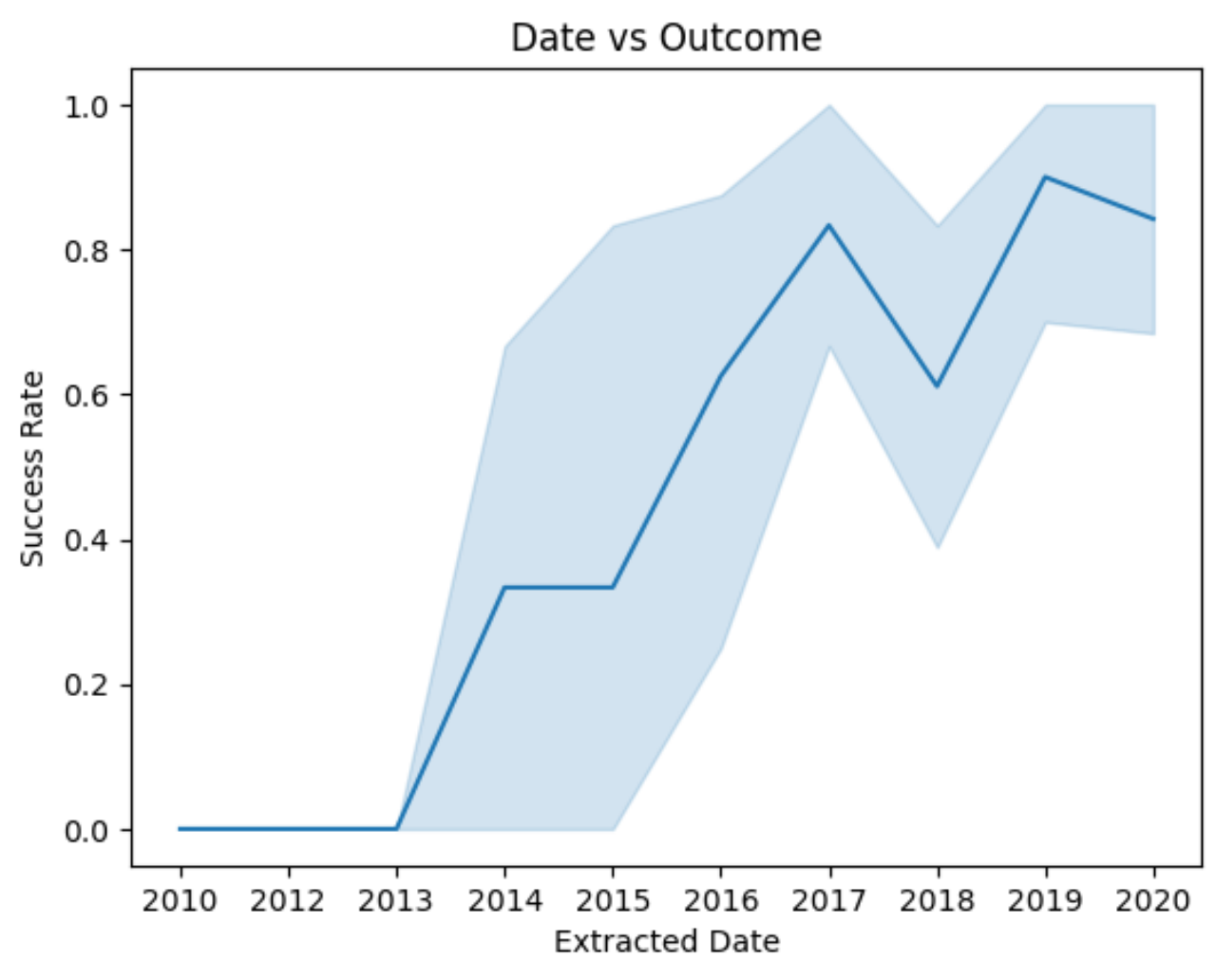
Payload vs. Orbit Type

- The relationship between pay load mass and orbit type



Launch Success Yearly Trend

- The launch success yearly trend



All Launch Site Names

The names of the unique launch sites in the space mission

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin 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

- The total payload mass carried by boosters launched by NASA (CRS)

Total Payload Mass
45596

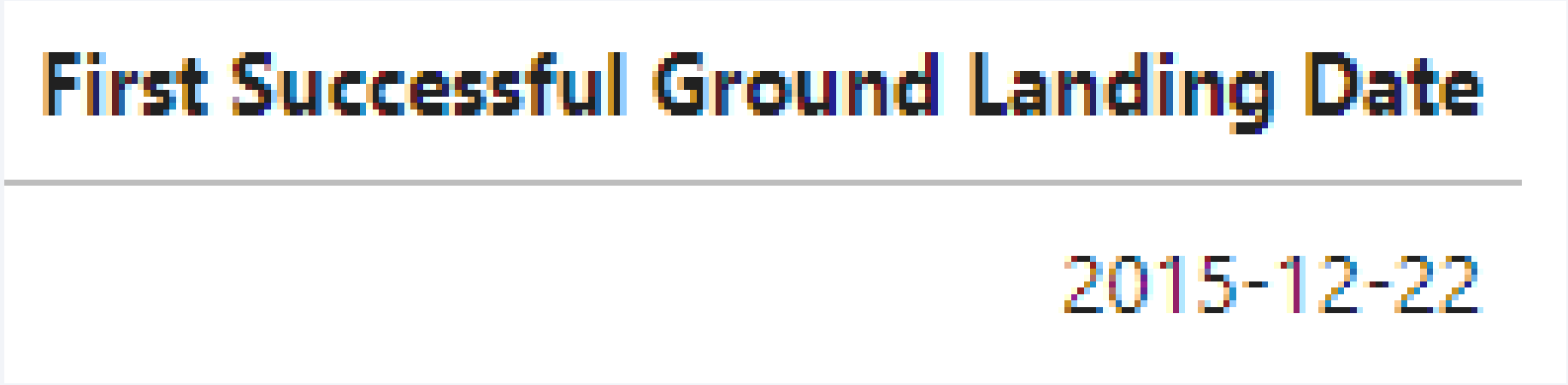
Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1

Average Payload Mass by F9 v1.1
2928.4

First Successful Ground Landing Date

- The date when the first successful landing outcome in ground pad was achieved is:



Successful Drone Ship Landing with Payload between 4000 and 6000

- 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 B1020
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes

Success	Failure
100	1

Boosters Carried Maximum Payload

- The names of the booster versions which have carried the maximum payload mass

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

- The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

Date	Booster_Version	Launch_Site
2015-01-10	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	F9 v1.1 B1015	CCAFS LC-40

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

- The count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

Landing_Outcome	COUNT
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

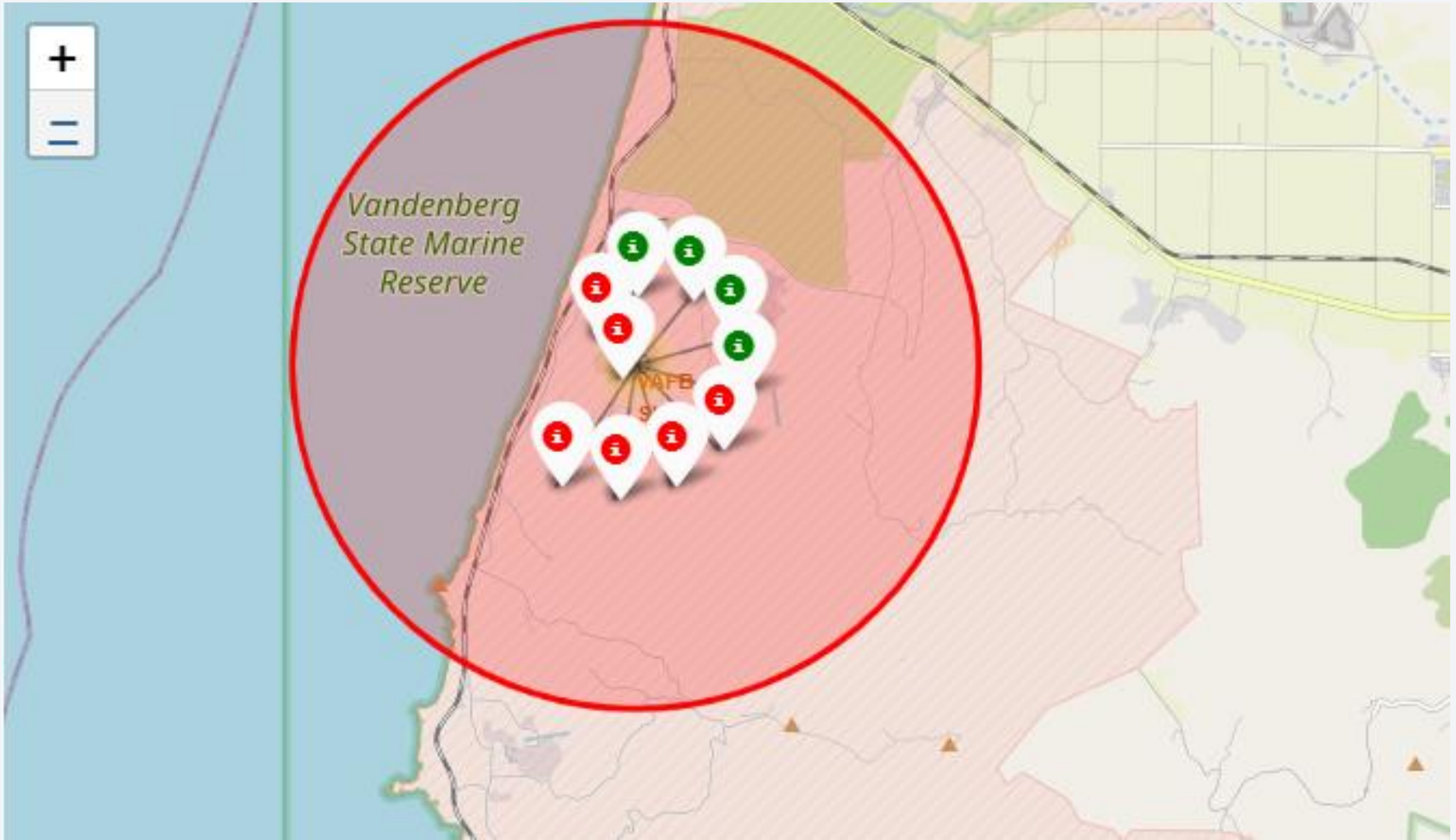
Launch Sites Proximities Analysis

All launch sites on a Folium map (1)

- All launch locations are shown on Map 1

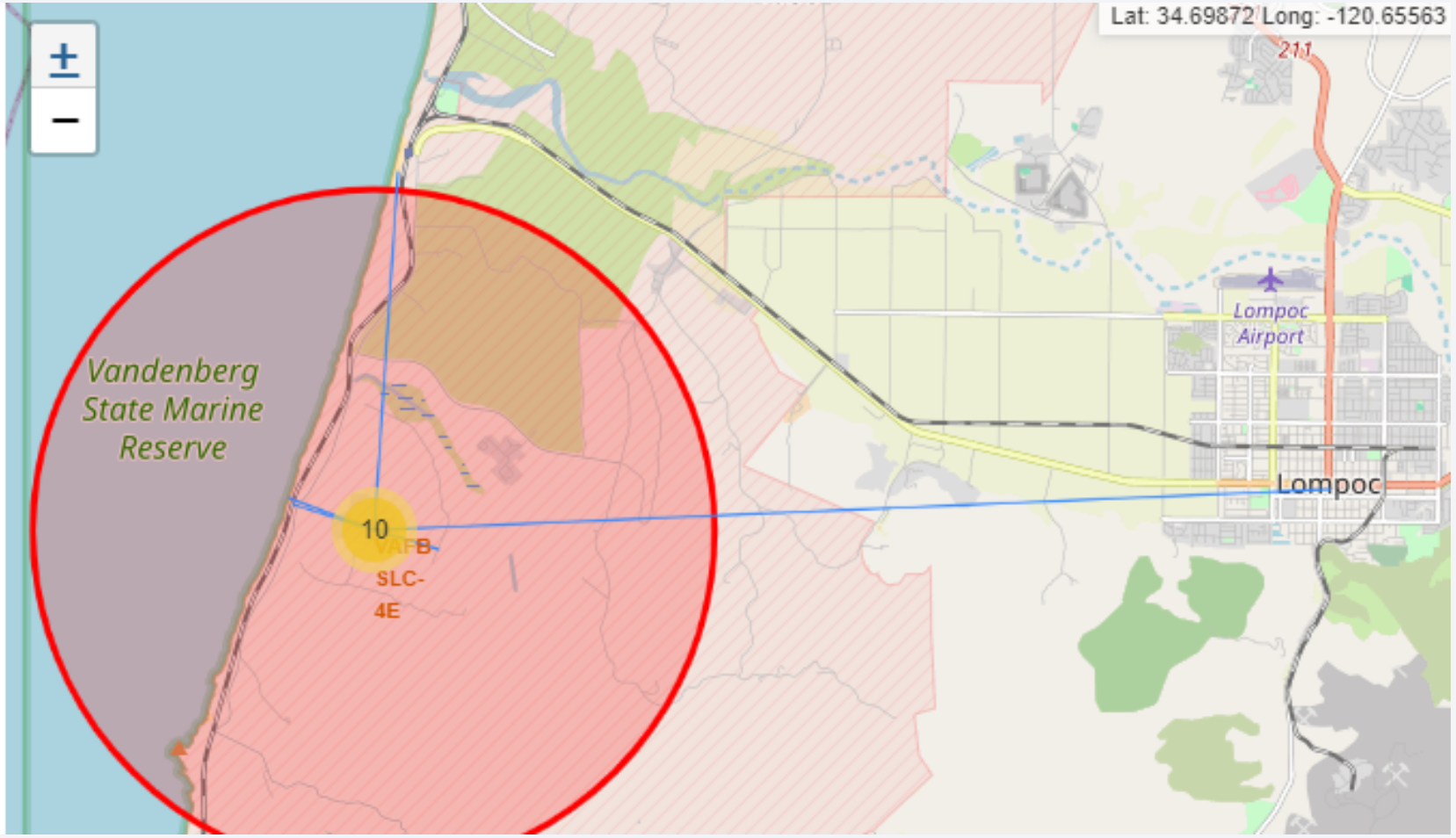
Succeeded and failed launches on a Folium map (2)

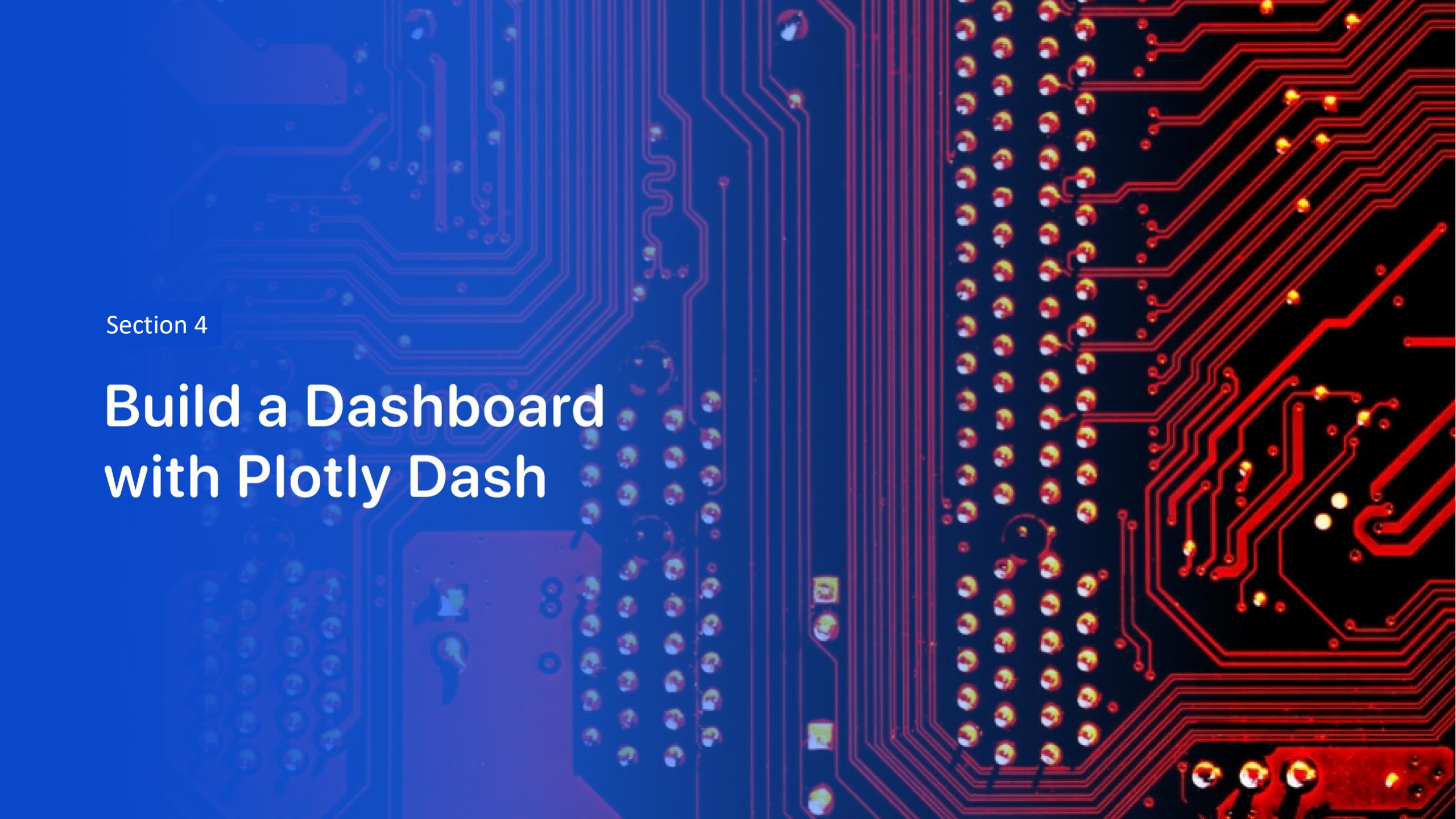
All Succeeded and failed launches for each site are shown on Map 2



Distances from a launch site on a Folium map (3)

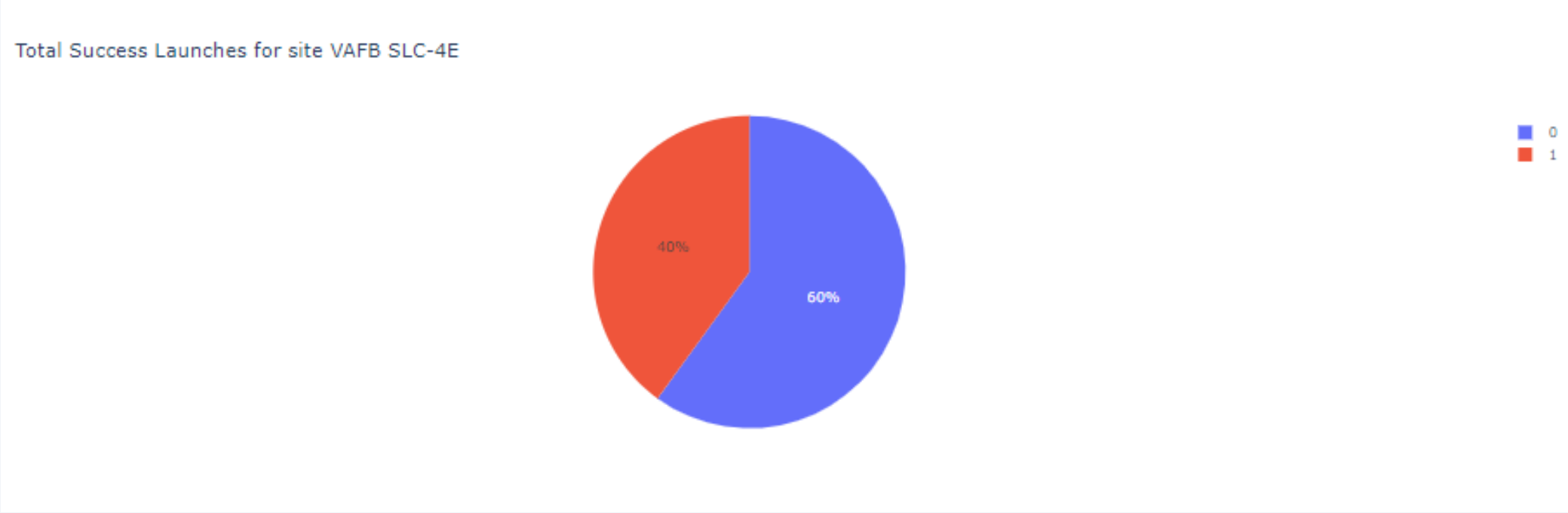
The distances from a launch site (VAFB SLC-4E) launch site to the nearest city, railway and highway are show on Map 3




Section 4
Build a Dashboard with Plotly Dash

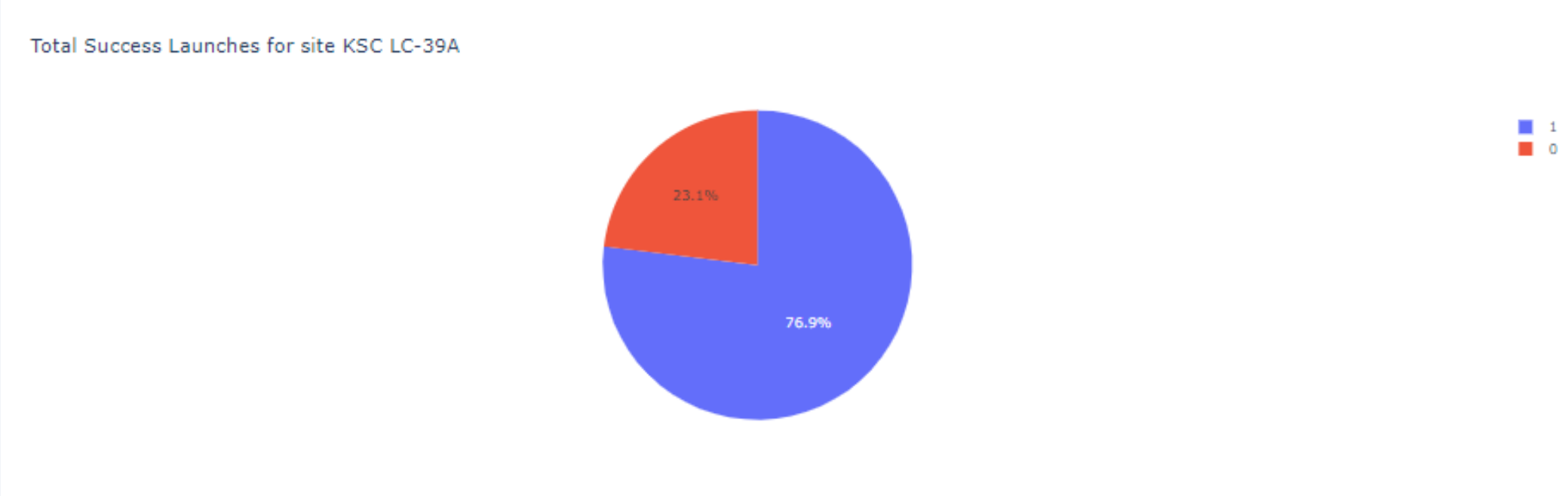
Pie Chart on a Dash site

- Here is a pie chart when we select the launch site VAFB SLC-4E. 0 is a failure land and 1 the opposite. We can see that 60% of the launches from this launch site were succeeded.



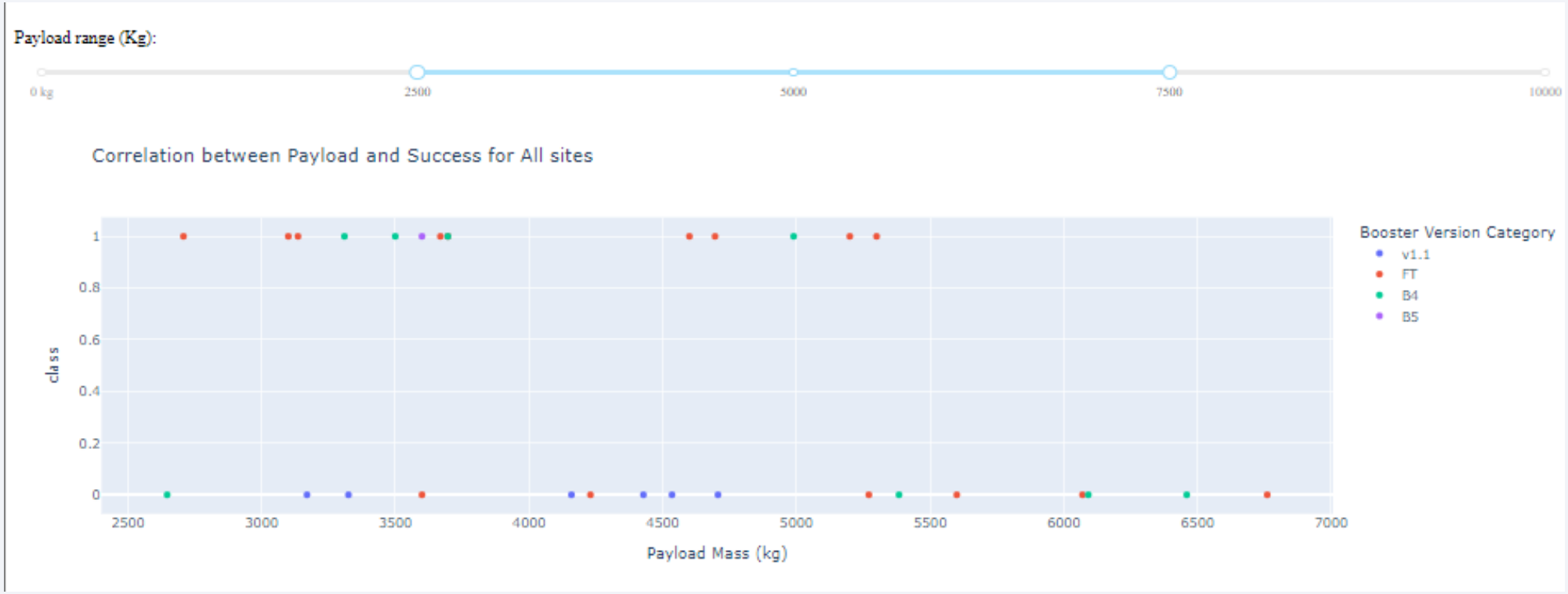
Pie Chart of the most succeeded launch site on a Dash site

- Here is a pie chart when we select the most succeeded launch site, KSC LC-39A. We can see that 76.9% of the launches from this launch site were succeeded.



Payload vs. Launch Outcome scatter plot on Dash site

- The picture shows a scatterplot when the payload mass range is set to be from 2500 to 7500 kg. A 0 represent a failed launches and 1 the opposite



Section 5

Predictive Analysis (Classification)

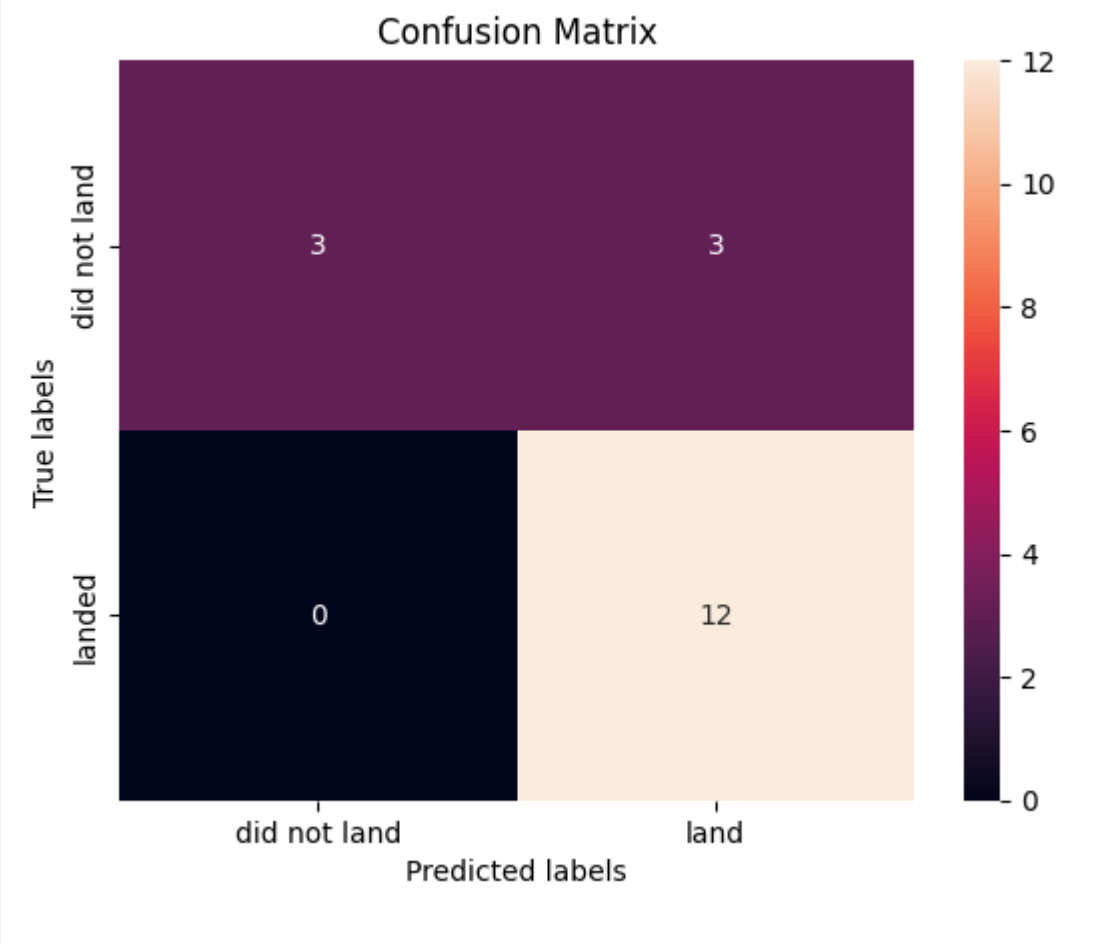
Classification Accuracy

- Here are the four results form each model where we can see that all of them have the same accuracy score (83.33%), but decision tree has the best GridSearchCV score of all of them (87.50%).

Model	Accuracy Score	GridSearchCV Best Score
Linear Regression	83.33%	84.64%
SVM	83.33%	84.82%
Decision Tree	83.33%	87.50%
KNN	83.33%	84.82%

Confusion matrix

- Here we can see the confusion matrix where all the four models output the same graphic.



Discussion

- From the data visualization, it is observed that certain characteristics can influence the outcome of the mission in various ways. For example, missions with heavy payloads are most successful in Polar, LEO and ISS orbits. However, in GTO orbits, it is difficult to distinguish between success and failure.
- Each characteristic can affect the outcome of the mission in a unique way. Although it is difficult to understand exactly how each influences, we can use machine learning algorithms to discover patterns in past data and predict the success of future missions based on these characteristics.

Conclusions

- In this project, we attempt to predict whether the first stage of a given Falcon 9 launch will land to determine the cost of the launch. Each characteristic of a Falcon 9 launch can affect the outcome of the mission in a particular way.
- Various machine learning algorithms are used to learn past data patterns from Falcon 9 launches and generate predictive models that can predict the outcome of a Falcon 9 launch.
- The predictive model generated by the decision tree algorithm was the one that showed the best performance among the 4 machine learning algorithms used, although this was only due to the GridSearchCV Score, since, if we compare the average score, the four models performed equally well. .

Thank you!