IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Valeria Pamato
19th August 2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies
  - Data collection – API
  - Data collection – WebScraping
  - Data Wrangling
  - EDA with Data Visualization
  - EDA with SQL
  - Interactive Visual Analytics with Folium
  - Interactive Dashboard with Plotly Dash
  - Predictive Analysis with Machine Learning
- Summary of all results
  - EDA results
  - Interactive Visual Analytics results
  - Interactive Dashboard results
  - Predictive Analysis (ML) results

Introduction

- Project background and context

SpaceX is a highly successful rocket company. Due to a system they created that allows them to land the rocket after launch and reuse it, their rockets are substantially less expensive than those of the other businesses. The cost of the rocket can be determined if the first stage lands. Other businesses might make use of this information.

- Problems you want to find answers

What is the success rate of the launces?

How rocket characteristics interact with launch success?

Can we predict the success rate based on past launches?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API and Web Scraping
- Perform data wrangling
  - One-hot encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

For this project we collected data regarding Falcon 9 rocket and the characteristics of every launch for this type.

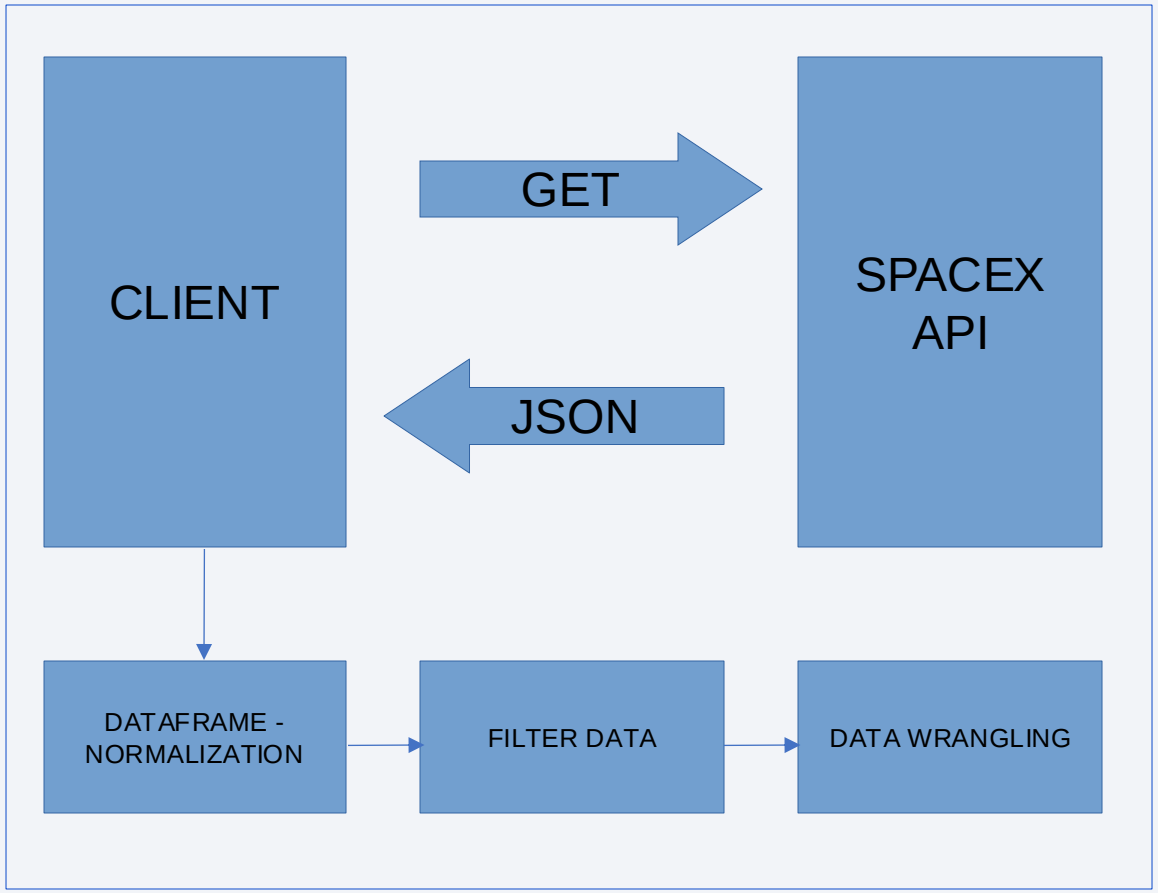
The data was gathered in two ways:

1) SPACEX API using the get request. The resulting json file was normalized and transformed into a DataFrame using pandas. Then we dealt with missing values using the mean value to replace them.
2) Web scraping Falcon 9 Wikipedia page using BeautifulSoup. The valuable data was parsed from the HTML table to a DataFrame.

Data Collection – SpaceX API

- The data was collected from the SpaceX API using get request and turned into a DataFrame.
- Some cleaning, data wrangling and formatting was done.

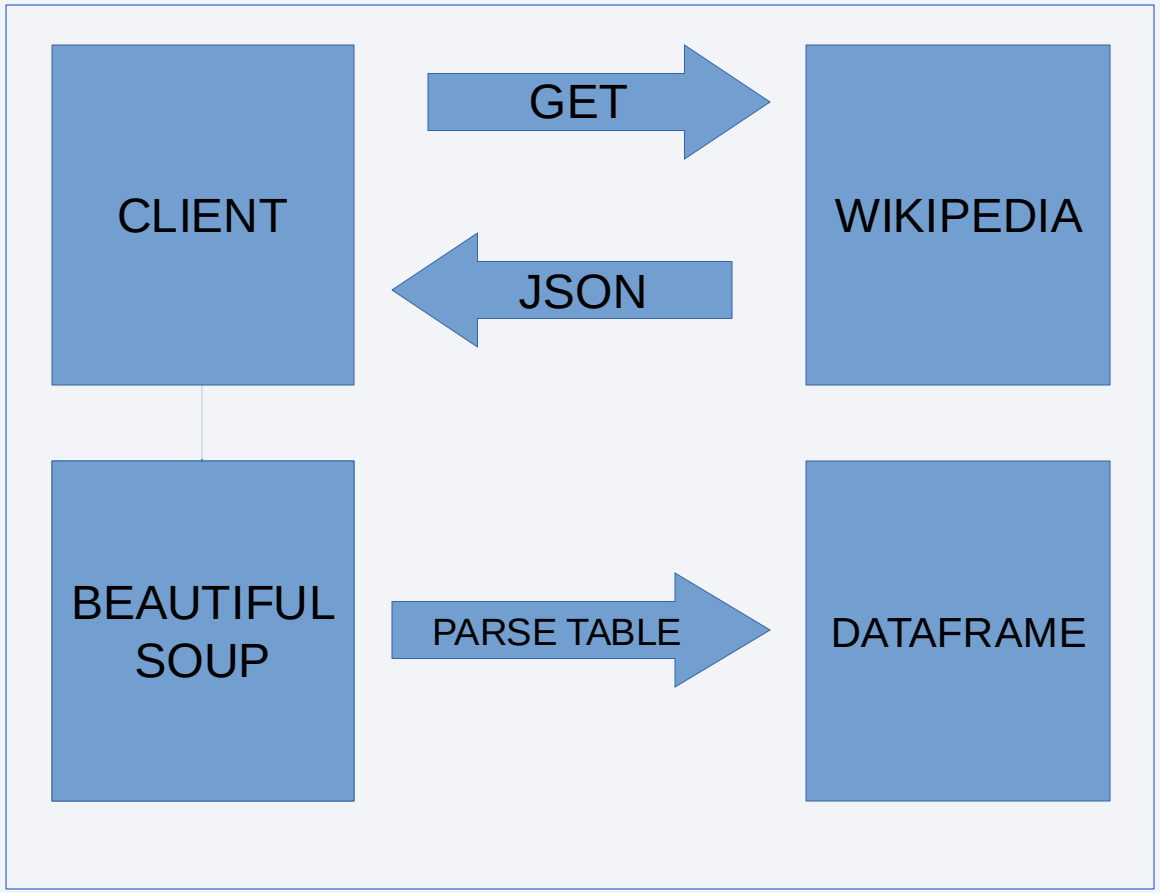
- https://github.com/valeriapamato/Winning-Space-Race-with-Data-Science---Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- We used Beautiful Soup to web scrape the Falcon 9 Wikipedia page
- The HTML table was parsed into a DataFrame

- https://github.com/valeriapamato/Winning-Space-Race-with-Data-Science---Capstone/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

N° OF ORBIT LAUNCHES	
SITE	**LAUNCHES**
CCAFS SLC 40	55
KSC LC 39A	22
VAFB SLC 4E	13

THE 4 MOST FREQUENT ORBITS	
ORBIT	**OCCURRENCES**
GTO	27
ISS	21
VLEO	14
PO	9

MISSION OUTCOMES	
True ASDS	False ASDS
True RTLS	False RTLS
True Ocean	False Ocean
None ASDS	None None

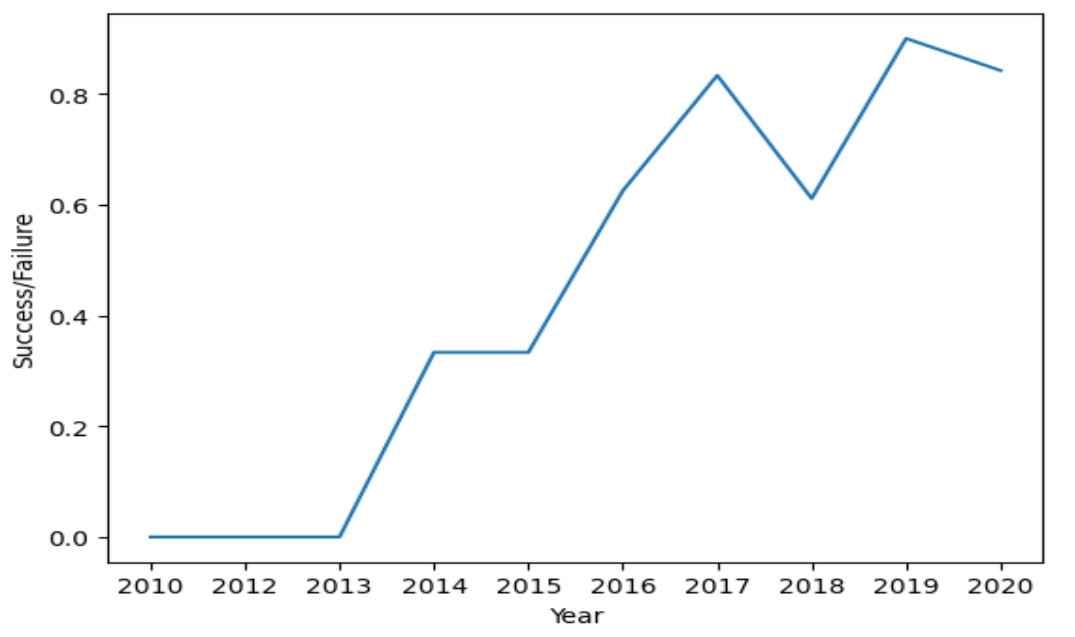
- We performed some Exploratory Data Analysis (EDA).
- We calculated the number of launches on each site, the number and occurrence of each orbit, the number and occurence of mission outcome of the orbits.
- We created a landing outcome label for training supervised models.

https://github.com/valeriapamato/Winning-Space-Race-with-Data-Science---Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

We performed exploratory Data Analysis :

- Visualized the relationship between Flight Number and Launch Site, Payload Mass and Launch Site, success rate of each orbit type, FlightNumber and Orbit type, Payload Mass and Orbit type.
- Visualized the launch success yearly trend.



We performed Feature Engineering creating dummy variables to categorical columns will be used in success prediction.

https://github.com/valeriapamato/Winning-Space-Race-with-Data-Science---Capstone/blob/main/edadataviz.ipynb

EDA with SQL

We loaded the SpaceX dataset into Db2 database.

We applied EDA using SQL:

- names of the unique launch sites in the space mission
- total payload mass carried by boosters launched by NASA (CRS)
- date when the first succesful landing outcome in ground pad was acheived.
- total number of successful and failure mission outcomes
- records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015

https://github.com/valeriapamato/Winning-Space-Race-with-Data-Science---Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

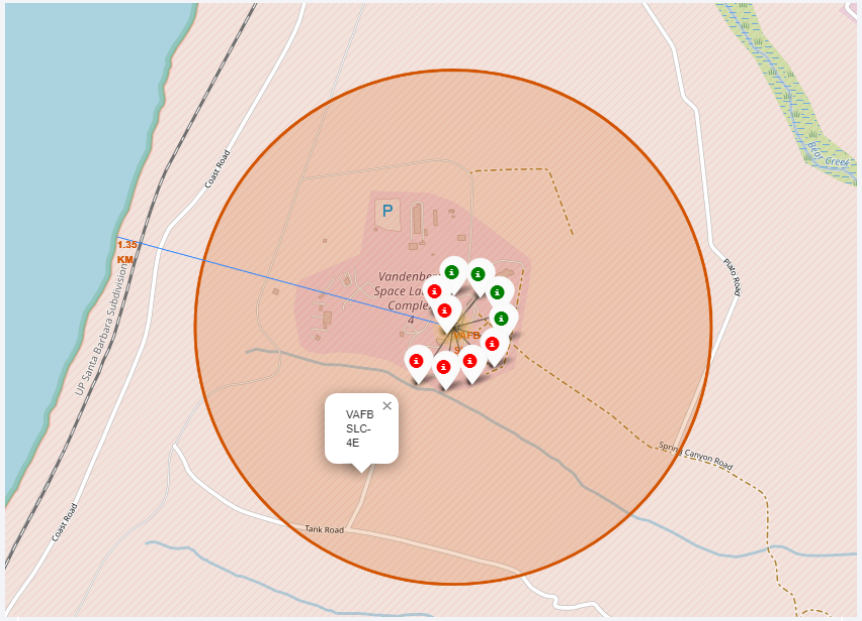
Build an Interactive Map with Folium

We marked all launch sites and the success/failed launches for each site on the map using markers, circles and lines. To mark the success/fail we assigned the outcomes to a class 0 and 1.

We calculate the distances between a launch site to its proximities and marked this distance with lines and markers.

With these maps we understood that the launch sites are all in the coastline and close to the equator. The success/fail was marked with different colors to easily understand and we analyzed the distance between the launch sites and railways, highways, cities.

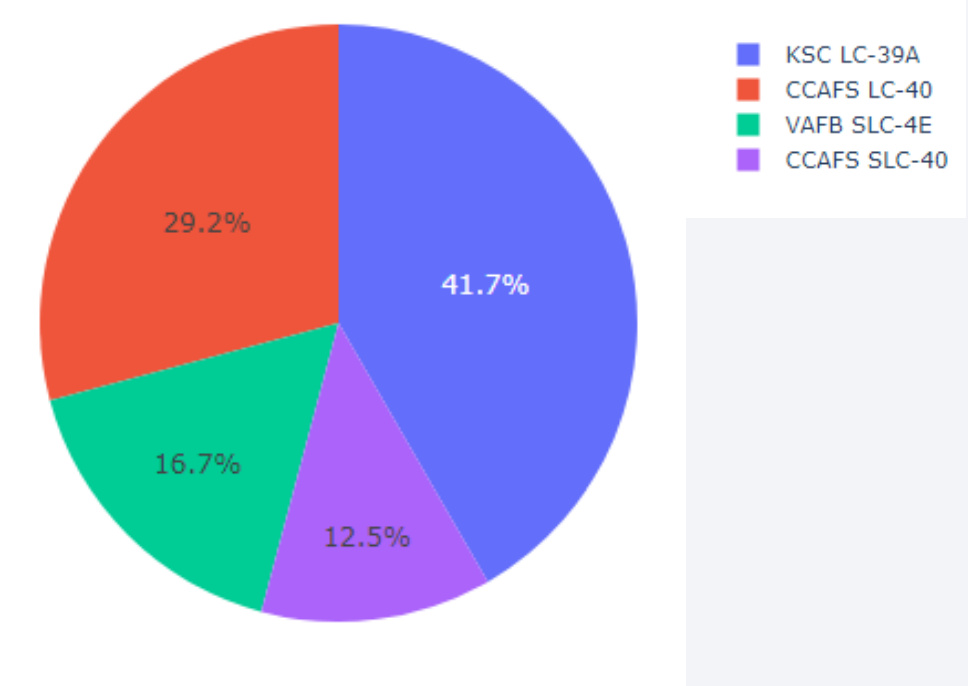
https://github.com/valeriapamato/Winning-Space-Race-with-Data-Science---Capstone/blob/main/lab_jupyter_launch_site_location.ipynb



VAFB SLC-4E site with success/failed launches and distance from coastline (1.35km)

Build a Dashboard with Plotly Dash

LAUNCH SUCCESS COUNTS



We created an interactive dashboard with Plotly Dash with a drop-down to select a launch site and range slider to select payload.

We plotted a pie chart with the number of successful launches for each site and a scatterplot to show the relationship between Mass Payload and Launch Outcome with different booster versions.

https://github.com/valeriapamato/Winning-Space-Race-with-Data-Science---Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

We performed exploratory Data Analysis, determined Training Labels and performed different classification alghorithms:

- We load the data using pandas and numpy, performed a train-test split.
- We performed Logistic Regression, a SVC, a Decision Tree and a KNN Classification
- For all the models we created a GridSearchCV to find the best hyperparameters and calculated the accuracy and confusion matrix to evaluate the models.
- At the end we found the best performing model.

https://github.com/valeriapamato/Winning-Space-Race-with-Data-Science---Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

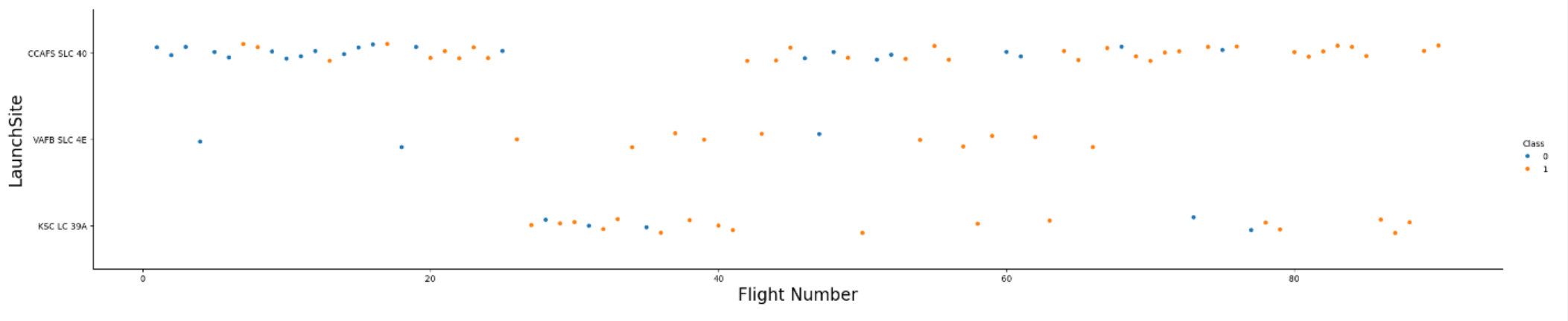
Results

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

Section 2

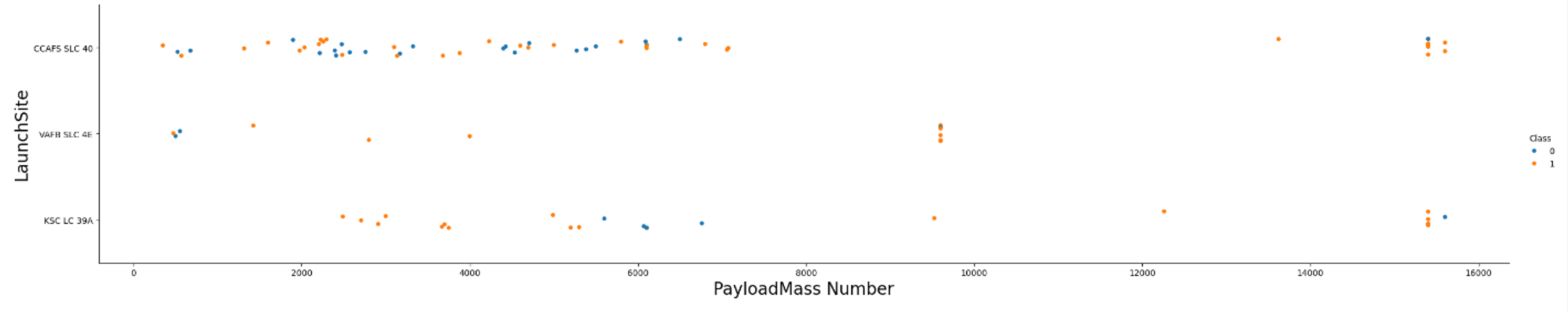
Insights drawn from EDA

Flight Number vs. Launch Site



From this plot we find that as the number of flights increases, the launch success rate increases.

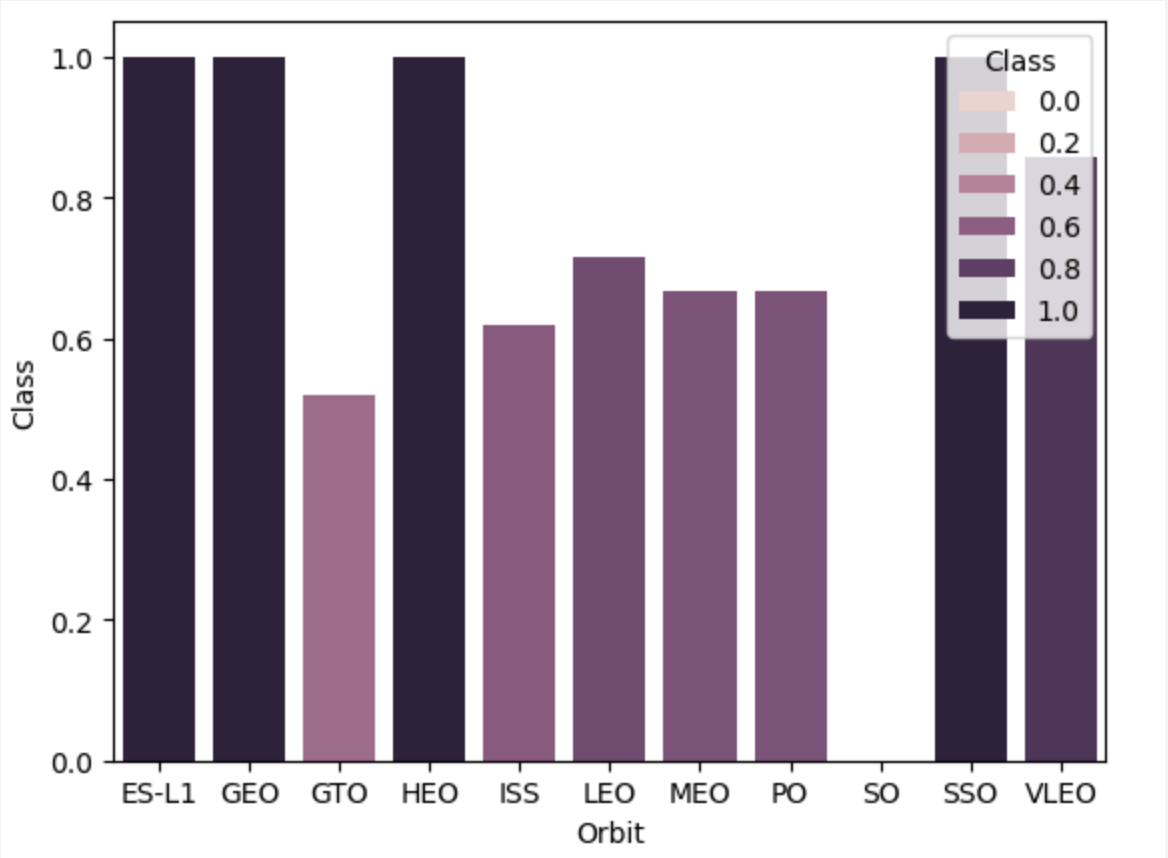
Payload vs. Launch Site



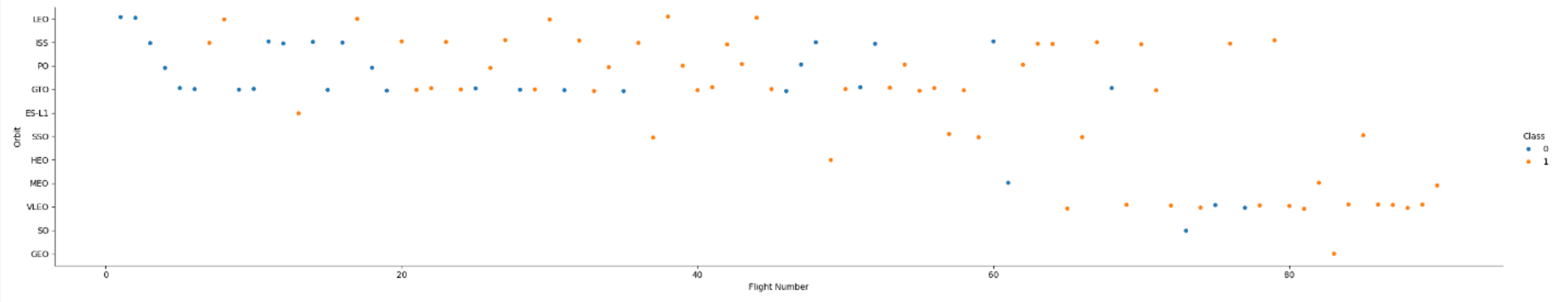
- From this plot we find that:
  - For the VAFB-SLC launch site there are no rockets launched for heavy payload mass
  - For most of the CCAFS SLC 40 site's launches with low mass there is no clear distinction between success and failure, while there are a few launches with very high mass that have been mostly successful

Success Rate vs. Orbit Type

- The most successful orbit types are:
- ES-L1
- GEO
- HEO
- SSO

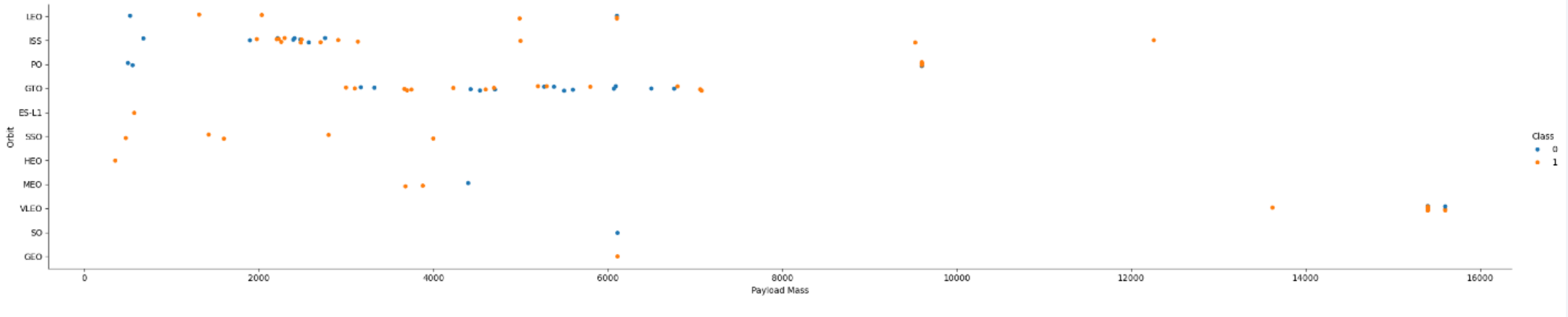


Flight Number vs. Orbit Type



- In LEO orbit, success seems to be related to the number of flights.
- In GTO orbit appears to be no relationship between flight number and success.

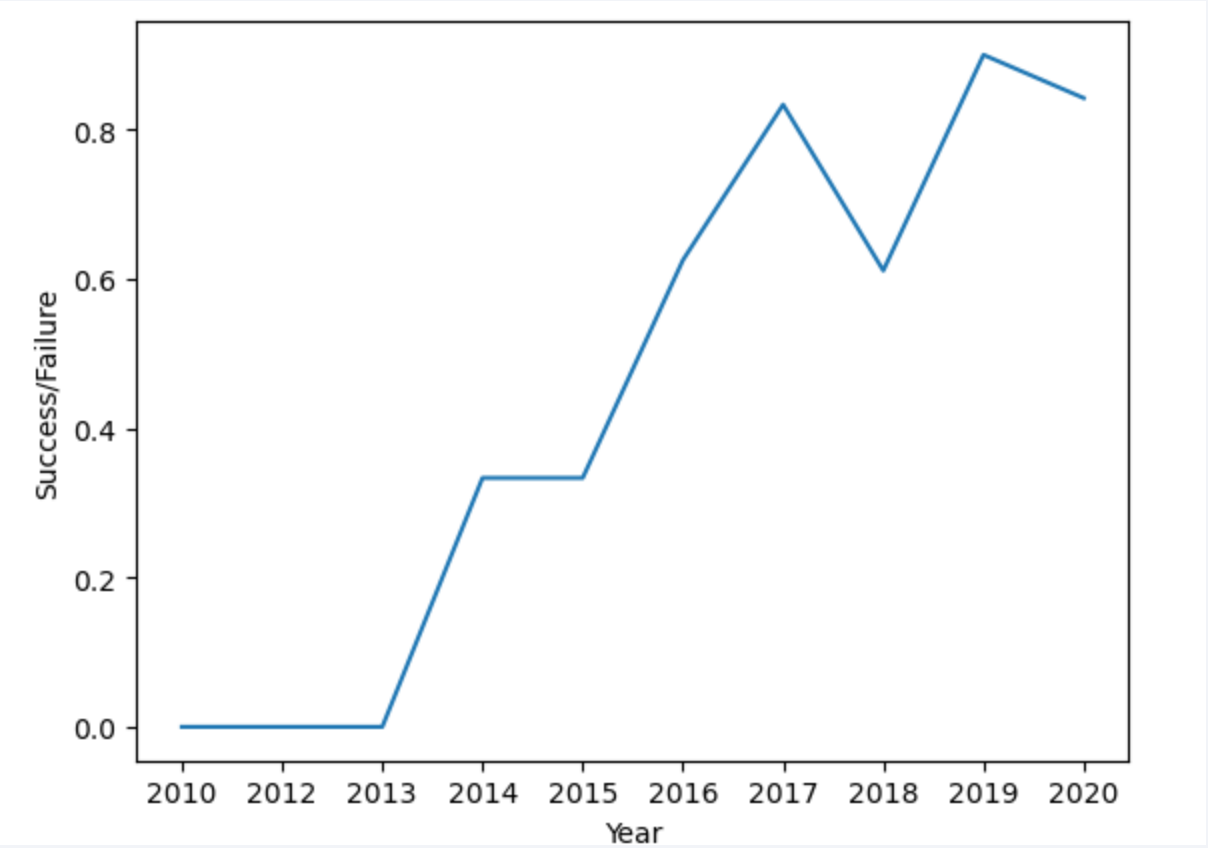
Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- For GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend

- We can observe that the success rate since 2013 kept increasing till 2020



All Launch Site Names

We use the DISTINCT key word to retrieve the unique launch site names

```
%sql select distinct("Launch_Site") from SPACEXTBL
```

```
 * sqlite:///my_data1.db
Done.
```

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

Launch Site Names Begin with 'CCA'

We used the query using the LIKE keyword to retrieve a specific string

Display 5 records where launch sites begin with the string 'CCA'

```
%sql select * from SPACEXTBL where "Launch_Site" like 'CCA%' limit 5
```

```
 * sqlite:///my_data1.db
Done.
```

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

We used the SUM keyword to calculate the total payload mass and the WHERE to filter the customer

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql select sum("PAYLOAD_MASS__KG_") from SPACEXTBL where Customer = 'NASA (CRS)'
```

```
 * sqlite:///my_data1.db
Done.
```

sum("PAYLOAD_MASS__KG_")
45596

Average Payload Mass by F9 v1.1

We used the AVG keyword to calculate the average payload mass and the WHERE to filter the booster version using LIKE

Display average payload mass carried by booster version F9 v1.1

```
%sql select avg("PAYLOAD_MASS__KG_") as "average payload mass" from SPACEXTBL where Booster_Version like "F9 v1.1%"
```

```
 * sqlite:///my_data1.db
Done.
```

average payload mass
2534.6666666666665

First Successful Ground Landing Date

We used the MIN keyword to find the first successful date and WHERE to filter the success in ground pad

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%sql select min(Date) from SPACEXTBL where Landing_Outcome = "Success (ground pad)"
```

```
 * sqlite:///my_data1.db
Done.
```

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

We used the WHERE to filter the landing outcome and the payload to find the ones 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

```
%sql select Booster_Version from SPACEXTBL where Landing_Outcome = "Success (drone ship)" and PAYLOAD_MASS__KG_> 4000 and PAYLOAD_MASS__KG_<6000
```

```
 * sqlite:///my_data1.db
Done.
```

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

We used the COUNT to retrieve the number of outcomes and the GROUP BY to get the total number based on the mission outcome

Task 7

List the total number of successful and failure mission outcomes

```
%sql select count(Mission_Outcome),Mission_Outcome  from SPACEXTBL group by Mission_Outcome
```

```
 * sqlite:///my_data1.db
Done.
```

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

Boosters Carried Maximum Payload

We used a subquery to find the booster versions with maximum payload mass

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%sql select Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

```
 * sqlite:///my_data1.db
Done.
```

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

We use the SUBSTR to find the Month, the WHERE to filter the year and the landing outcome

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

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%sql select substr(Date,6,2) as Month, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTBL where substr(Date,0,5)='2015' and Landing_Outcome = 'Failure (drone ship)'
```

```
 * sqlite:///my_data1.db
Done.
```

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

We used COUNT to retrieve the number of landing outcome, WHERE to filter the dates, the GROUP BY to group by landing outcome and ORDER BY the count

```
%sql select Landing_Outcome,count(Landing_Outcome) as Count_Outcomes  from SPACEXTBL where Date between '2010-06-04' and '2017-03-20'  group by Landing_Outcome ORDER BY Count_Outcomes DESC
```

```
 * sqlite:///my_data1.db
Done.
```

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

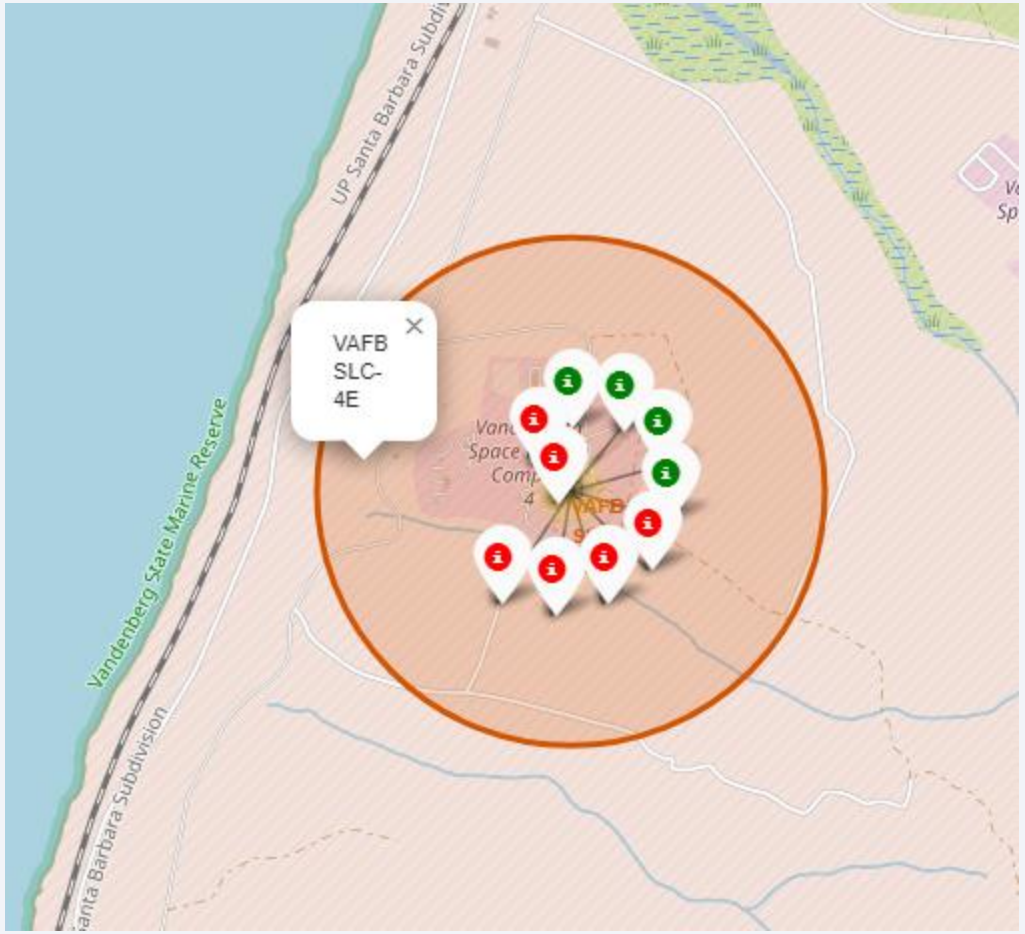

Section 3
Launch Sites
Proximities Analysis

SpaceX Launch Site locations

We can see that all the launch sites are located in the USA, near a coastline and on the equator

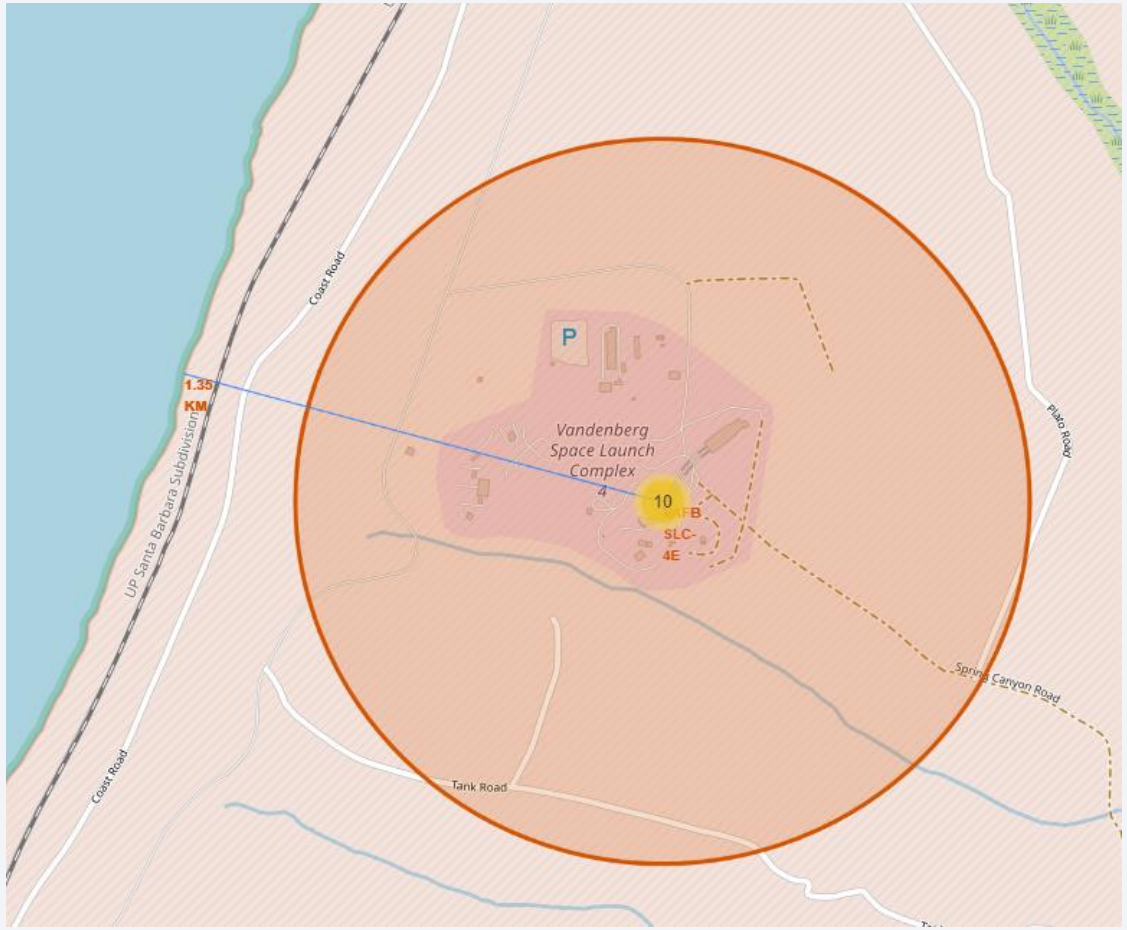
Success/Failed launces at VAFB SLC-4E

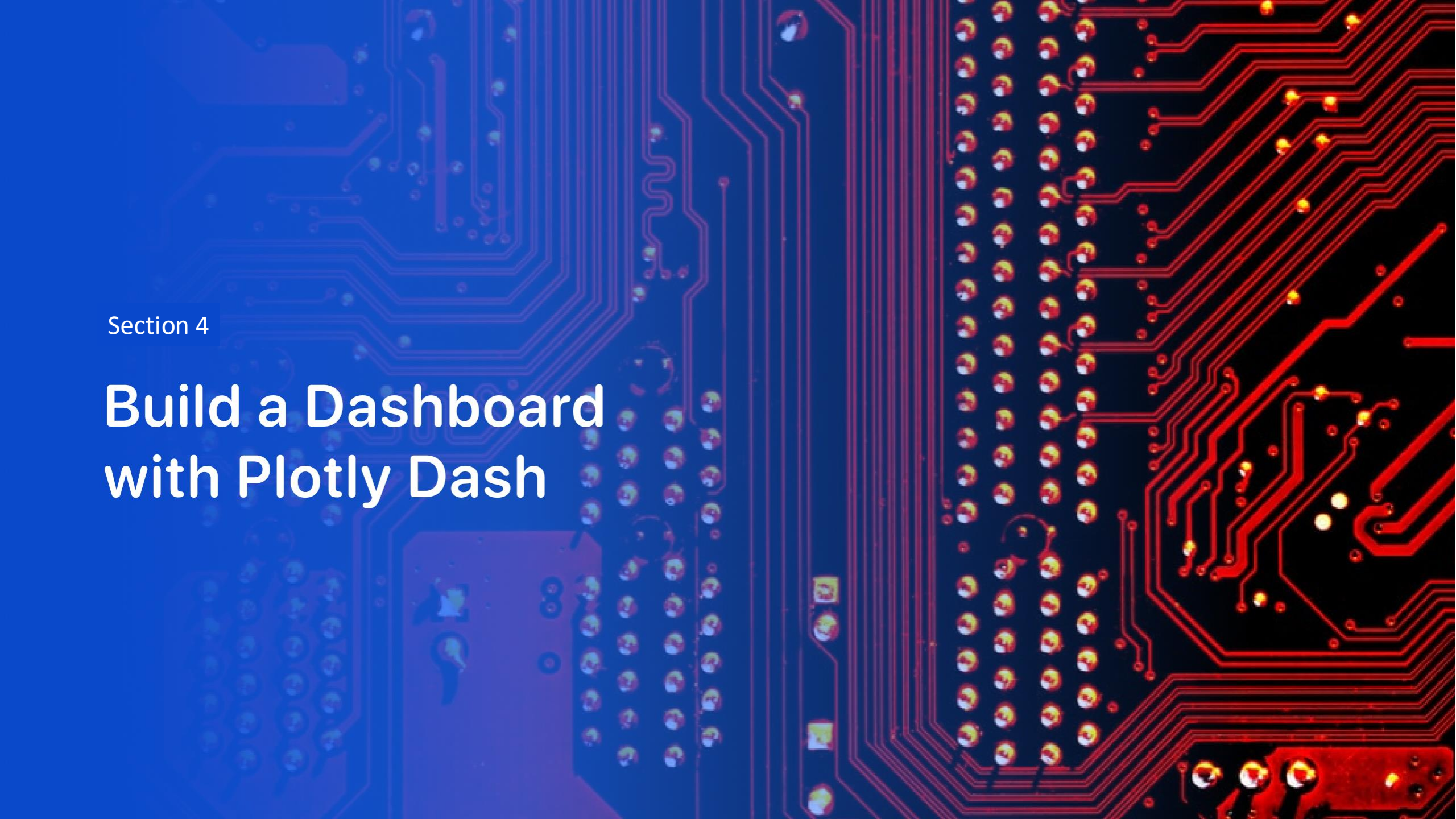
We can analyze the success/failed launches at the VAFB SLC-4E location. With the different colors is easy to see that the failed launches (red) were more than the success launches (green).



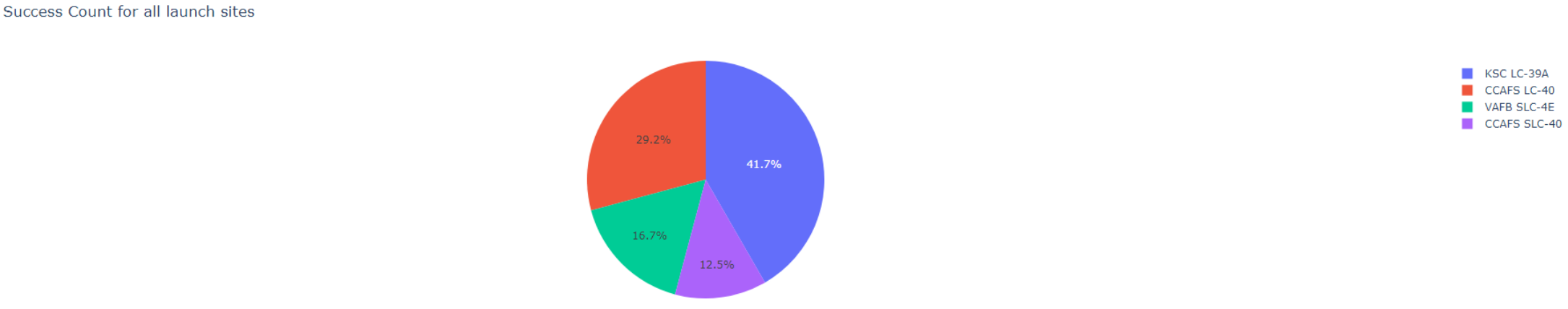
Distance between VAFB SLC-4E and coastline

We can see that the VAFB SLC-4E launch site is 1.35km from the coastline



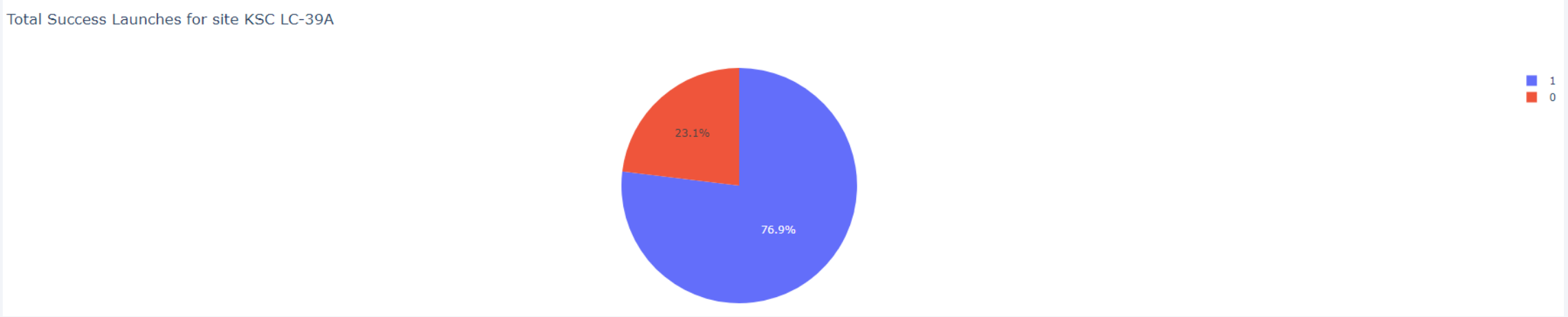

Section 4
Build a Dashboard with Plotly Dash

Launch success for all sites



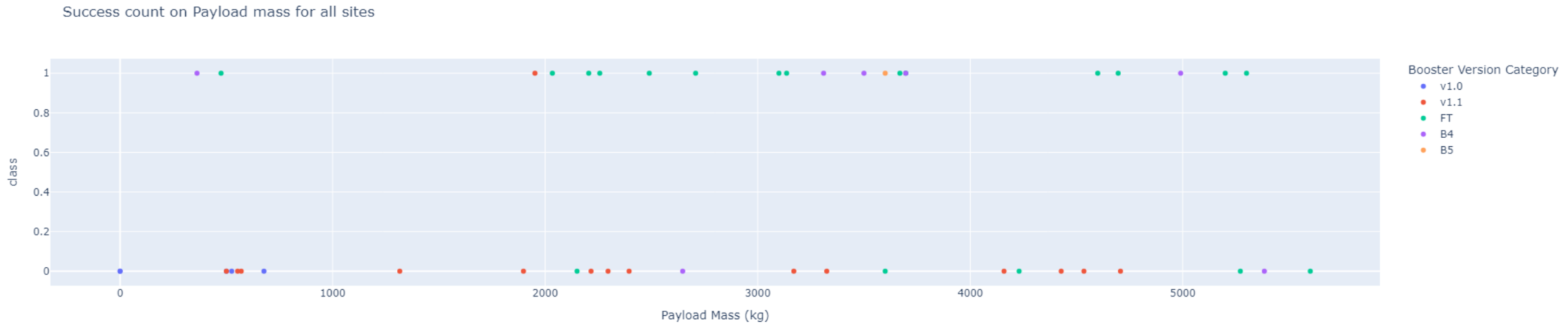
From this pie chart we can find that the majority of the successful launches were from the KSC LC-39A location

Success Ratio at KSC LC-39A



From this chart we can find that the KSC LC-39A location has a 76.9% success ratio

Relationship between Payload and Launch Outcome for all sites



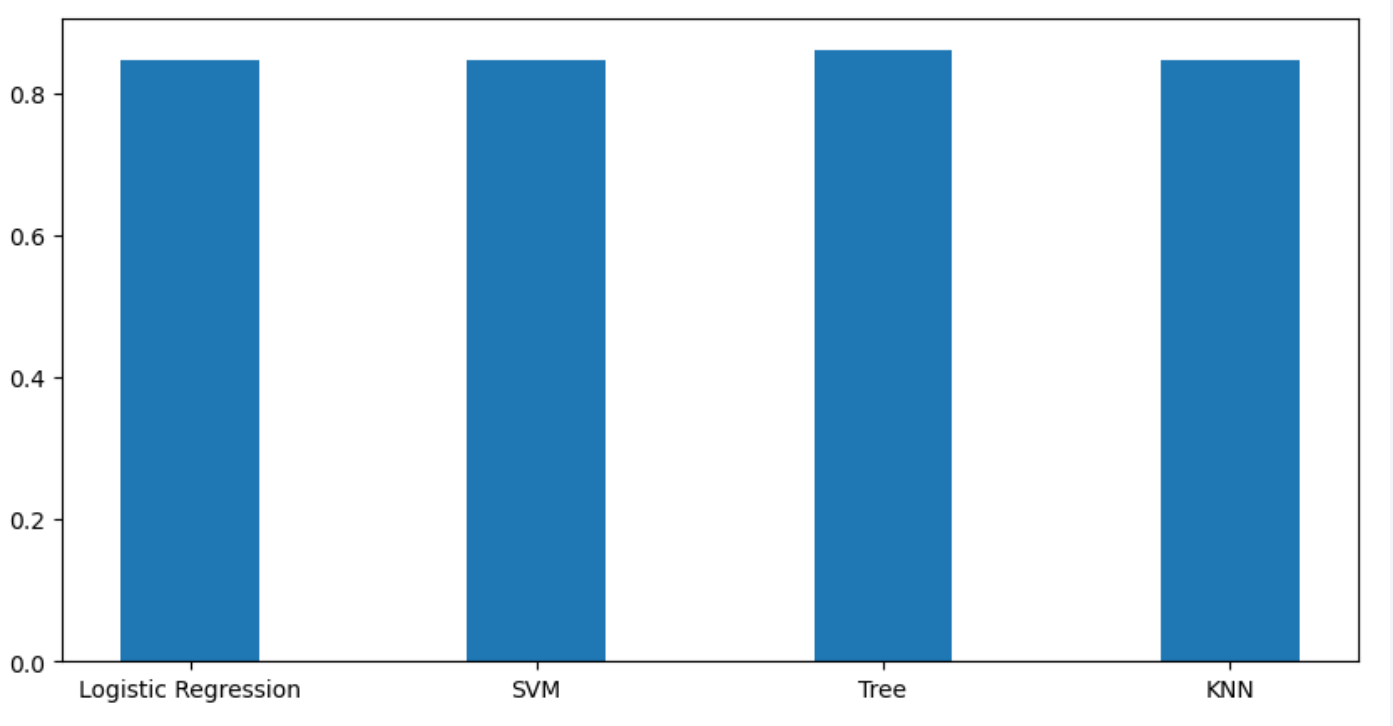
From this chart we can find that the FT booster has the largest success rate, while the v1.1 booster has the lowest success rate.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

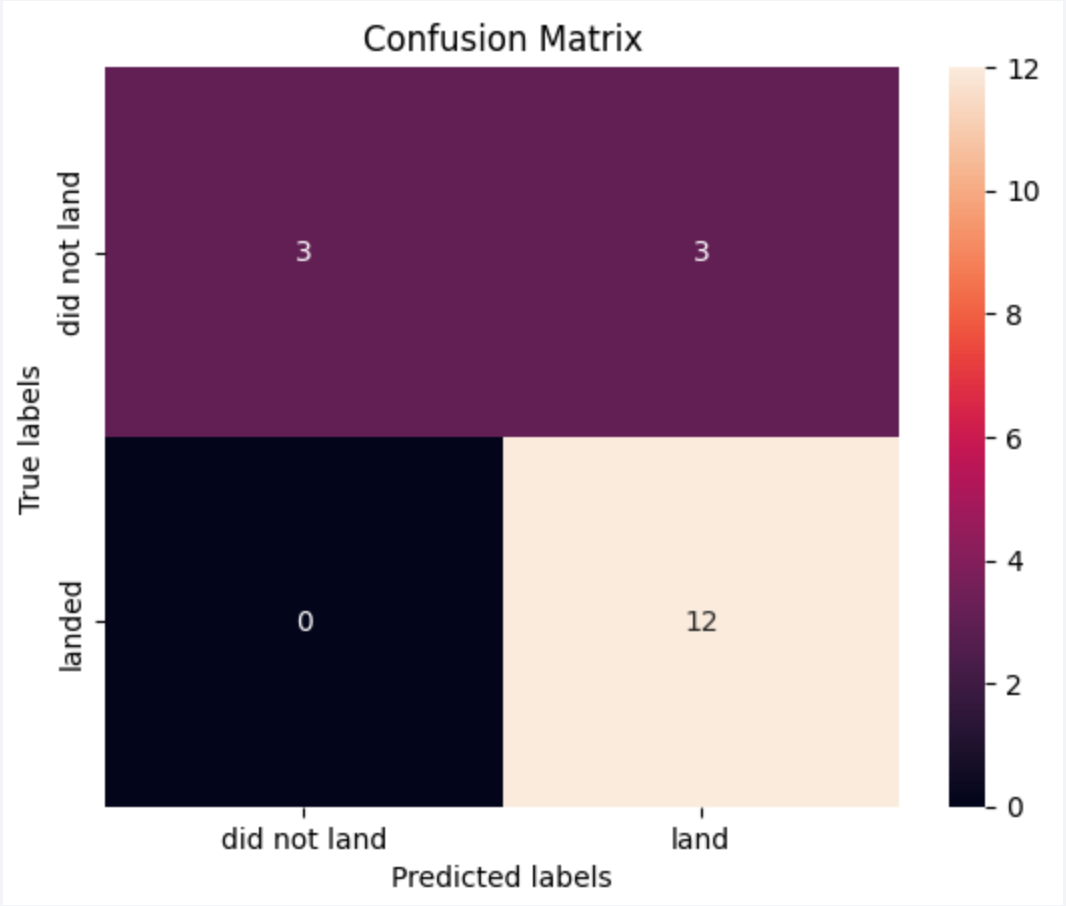
- The decision tree model has the best accuracy of all the models.



	Model	Accuracy
0	Logistic Regression	0.846429
1	SVM	0.848214
2	Tree	0.862500
3	KNN	0.848214

Confusion Matrix

- The confusion matrix of the decision tree shows us that the model is able to identify and distinguish between classes.
- There is problem of false positives, that is, it tends to distinguish landed launches as non-landed launches in 3 out of 15 cases, or 20% of the cases.



Conclusions

- There has been an increase in the success rate since 2013.
- The orbits with the highest success rate are: ES-L1, GEO, HEO, SSO
- The launch sites are all close to the coasts, probably for safety reasons
- The KSC LC-39A site is the one with the highest success rate the decision tree is the most accurate machine learning model

Thank you!