IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Laurence McDonald
2021-09-19

Outline

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

Executive Summary

Proposal

SpaceX advertises Falcon 9 launches with a **cost** of **$62 mil**, **other providers** offer **$165 mil.** SpaceX offer less as they can **reuse** the **1st stage.** By **predicting** if the **1st stage** will **land,** we can determine the **cost** of a **launch** which will **help** other companies **bid against SpaceX** for a rocket Launch.

Outcome

Using features including (Flight Number, Payload Mass, Orbit, Launch Site, Flights, Grid Fins, Reused, Legs, Landing Pad, B Lock, Reused Count, Serial) from SpaceX API and Wiki, we can predict the outcome of a successful launch with an **accuracy** of **0.875** and an **AUC ROC** score of **0.75.**

We can predict all successful landing outcomes and thus can provide competitors with information to assist them in producing a competitive cost of launch.

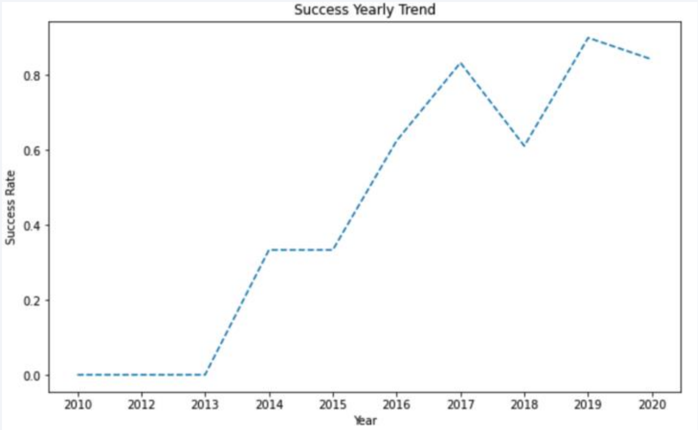


Figure 1: Falcon 9 1st Stage successful land rate.

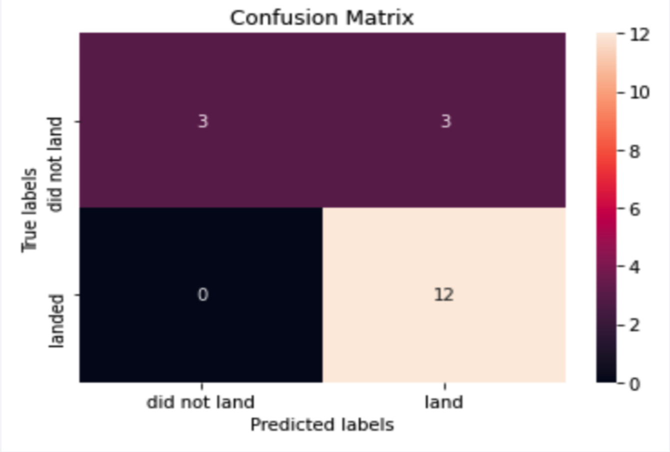


Figure 2: Confusion Matrix on test dataset prediction.

Introduction

<u>Problem Statement</u>

SpaceX advertises Falcon 9 launches with a **cost** of **$62 mil**, **other providers** offer **$165 mil.** SpaceX offer less as they can **reuse** the **1st stage.**

<u>Approach</u>

We will use launch data extracted from the SpaceX API and Wikipedia and explore correlations between the fields obtained and stage 1 launch and landing success.

By **predicting** if the **1st stage** will **land,** we can determine the **cost** of a **launch** which will **help** other companies **bid against SpaceX** for a rocket Launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected via two approaches:
    - SpaceX API data in the form of a json file that was translated into a pandas dataframe.
    - Web scraping Wikipedia HTML table via beautifulsoup.
- Perform data wrangling
  - The SpaceX API data had missing values in landing pads which were taken as nulls and pay load mass which was replaced with the mean value.
  - Features were used via get dummies and a boolean target generated 1, if 1st staged landed and 0, if otherwise.
- Perform exploratory data analysis (EDA) using visualization and SQL
  - The data was uploaded to a database and queried for profiling purposes and relationships visualized.
- Perform interactive visual analytics using Folium and Plotly Dash
  - The data explored in terms of launch site via geographical visualization and a dashboard describing key attributes of launch success by launch site.
- Perform predictive analysis using classification models
  - A predictive model was built based on the features extracted from the data collection process.

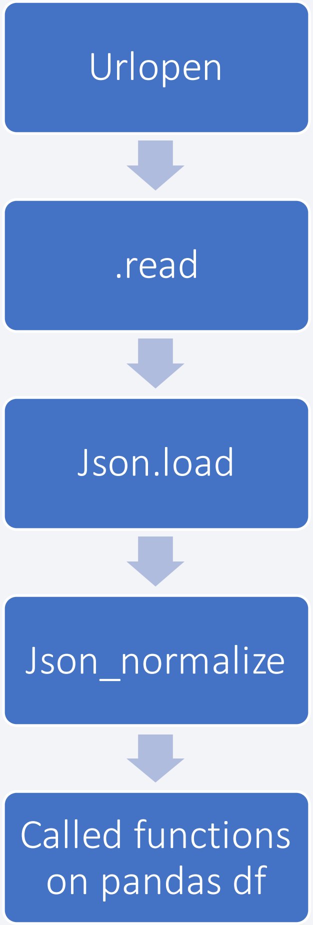
Data Collection

Data was collected via two approaches:

- SpaceX API data in the form of a json file that was translated into a pandas dataframe.
- Web scraping Wikipedia HTML table via beautifulsoup.

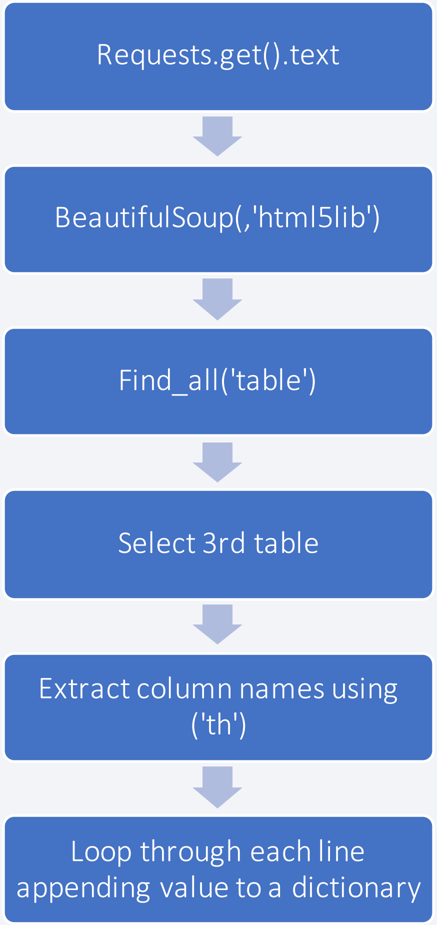
Data Collection – SpaceX API

- SpaceX API data was extracted using 4 provided functions, getBoosterVersion, getLaunchSIte, getPayloadData, and getCoreData.
- We used a static response object for the project after initially reading the api data via a request.get().
- [Github Link - Data Collection SpaceX API]



Data Collection - Scraping

- Web scraping was conducted on Wiki Link using beautiful soup.
- Data was extracted by creating a beautiful soup object, selecting our desired table and enumerating through rows adding values to a dictionary containing the column heading.
- Github Link - Data Collection Wiki



Data Wrangling

- Handling of Nulls, initially landing pads was kept null meaning no landing and payload mass was filled with the average. The Data wrangling results in a datasets where 67% of launches have a 1st stage successful landing.
- [Github - Data Wrangling]()

Handling of Nulls → Explore nulls and column types → Explore number of lanches on each site → Explore the frequency on Orbits → Explore the landing column to identify the landing outcome → Generate target based on True landings 1, otherwise 0.

EDA with Data Visualization

- Relationships within the dataset were explored to identify variables which correlate with a landing outcome. Identifying some key features off the bat will help us understand if the model will be able to predict the landing outcome.
- In our case we were able to see some correlation to landing with, payload mass, launch site, orbit. We were also able to understanding further information about the data itself including the increased success rate through time.
- Github - Data Visualization

EDA with SQL

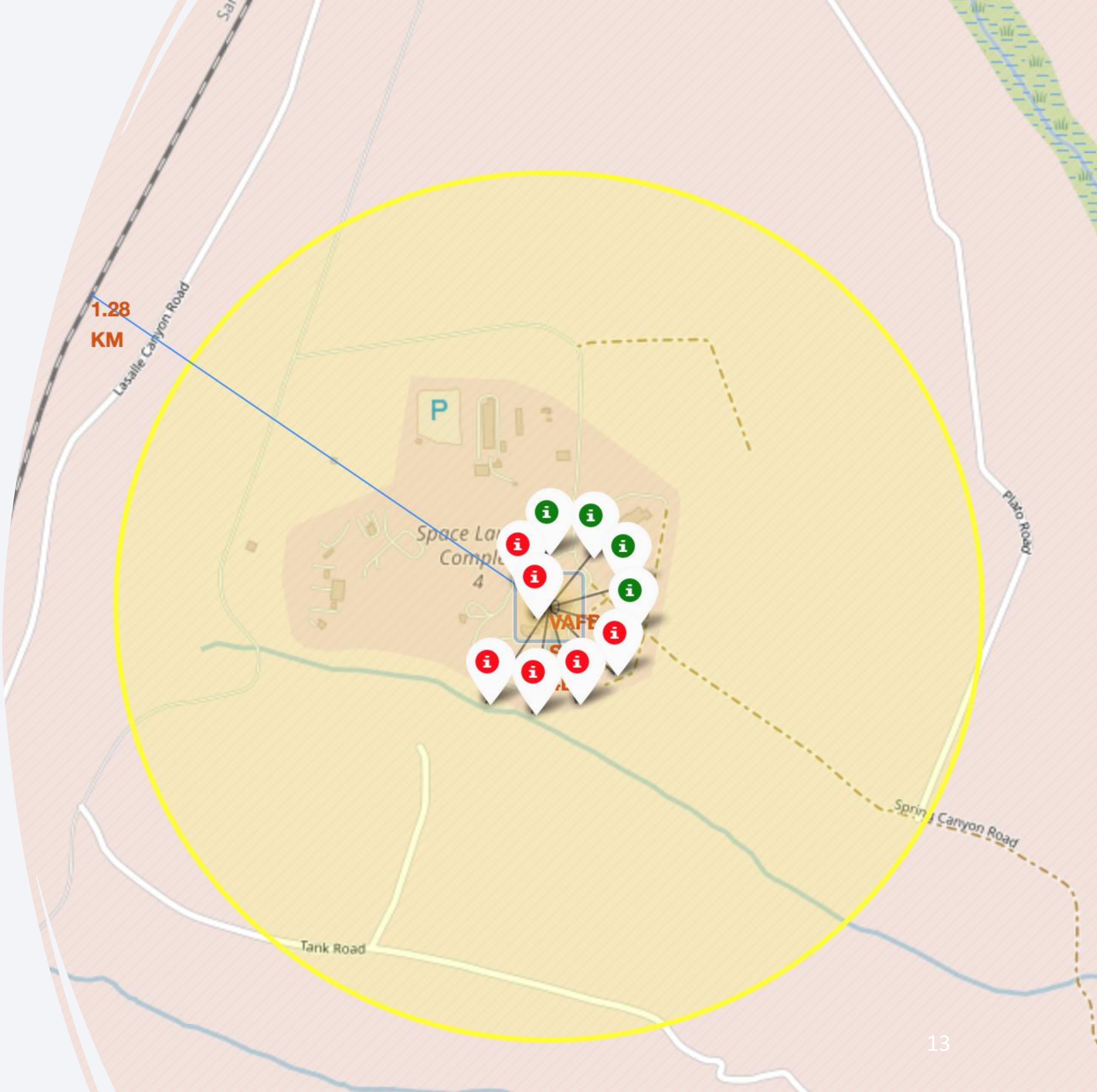
Below are a list of queries performed to profile the data.

1. ***Display the names of the unique launch sites in the space mission***
2. ***Display 5 records where launch sites begin with the string 'CCA'***
3. ***Display the total payload mass carried by boosters launched by NASA (CRS)***
4. ***Display average payload mass carried by booster version F9 v1.1***
5. ***List the date when the first successful landing outcome in ground pad was achieved***
6. ***List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000***
7. ***List the total number of successful and failure mission outcomes***
8. ***List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015***
9. ***Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order***

- Github Exploratory Analysis

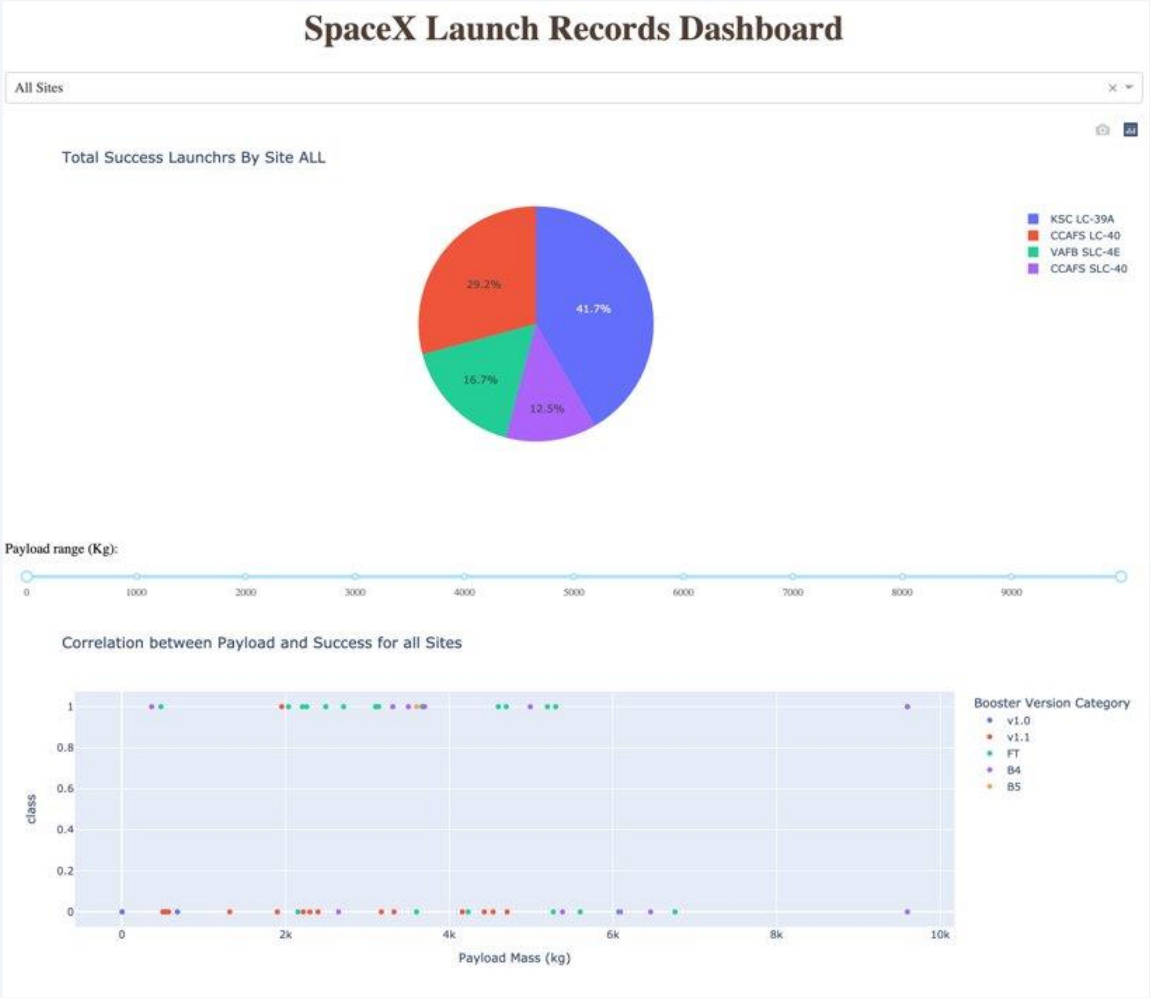
Build an Interactive Map with Folium

- Visualizing features withiin the dataset, we also developed a geographical view of the Falcon 9 launch sites used Folium. Creating Circle objects to mark lanch sites and markers to label them. We also created marker clusters to show the landing outcomes within the launch site, and lines to show proximity to landmarkes.
- [Github - Interactive Map]()

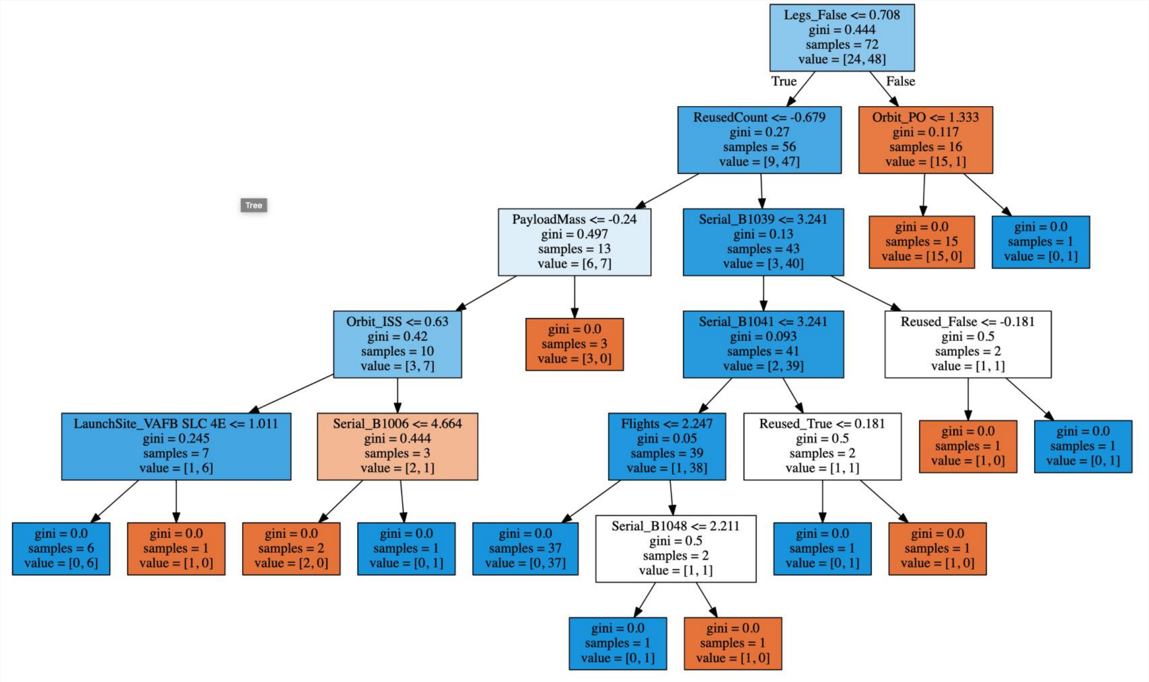


Build a Dashboard with Plotly Dash

- The Plotly Dasboard developed shows the relationship between launch site, payload mass, and landing outcome. This was key to provide information on launch sites.
- The pie chart shows the 1st stage landing success of the falcon 9 shuttle.
- The scatter plot below shows the success and failures by payload mass which can be toggled via the slider.
- [Github - Dashboard Code]()



Predictive Analysis (Classification)



Methodology

- Using features including (Flight Number, Payload Mass, Orbit, Launch Site, Flights, Grid Fins, Reused, Legs, Landing Pad, B Lock, Reused Count, Serial) we developed several models with the outcome below:
  - LR AUC: 0.75
  - TREE AUC: 0.75
  - SVM AUC: 0.75
  - KNN AUC: 0.75
  - LR accuracy : 0.8464285714285713
  - TREE accuracy : 0.875
  - SVM accuracy : 0.8482142857142856
  - KNN accuracy : 0.8482142857142858

Outcome

- We can predict the outcome of a successful launch using a decision tree classifier with an **accuracy** of **0.875** and an **AUC ROC** score of **0.75.**
- We can predict all successful landing outcomes and thus can provide competitors with information to assist them in producing a competitive cost of launch.

- Github - Predicting Landing Outcome

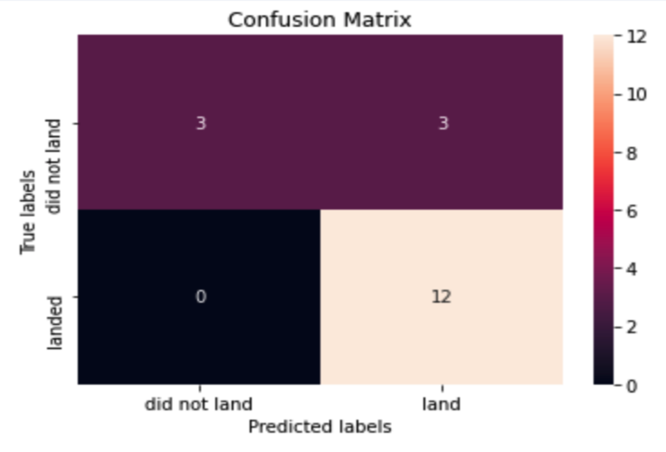


Figure 2: Confusion Matrix on test dataset prediction.

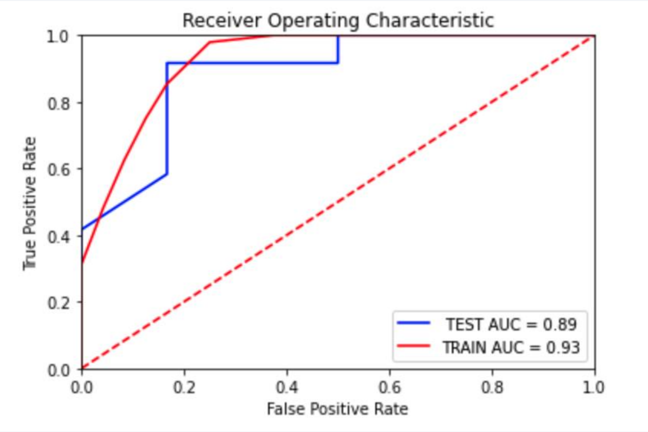


Figure 1: ROC Curve on Decision Tree Classifier

Results

<u>Proposal</u>

SpaceX advertises Falcon 9 launches with a **cost** of **$62 mil**, **other providers** offer **$165 mil.** SpaceX offer less as they can **reuse** the **1st stage.** By **predicting** if the **1st stage** will **land,** we can determine the **cost** of a **launch** which will **help** other companies **bid against SpaceX** for a rocket Launch.

<u>Outcome</u>

We have conducted data analysis to build a dashboard informing stakeholders of key launchsite data.

We have generated a geographical representation of this data.

Using features including (Flight Number, Payload Mass, Orbit, Launch Site, Flights, Grid Fins, Reused, Legs, Landing Pad, B Lock, Reused Count, Serial) from SpaceX API and Wiki, we can predict the outcome of a successful launch with an **accuracy** of **0.875** and an **AUC ROC** score of **0.75.**

We can predict all successful landing outcomes and thus can provide competitors with information to assist them in producing a competitive cost of launch.

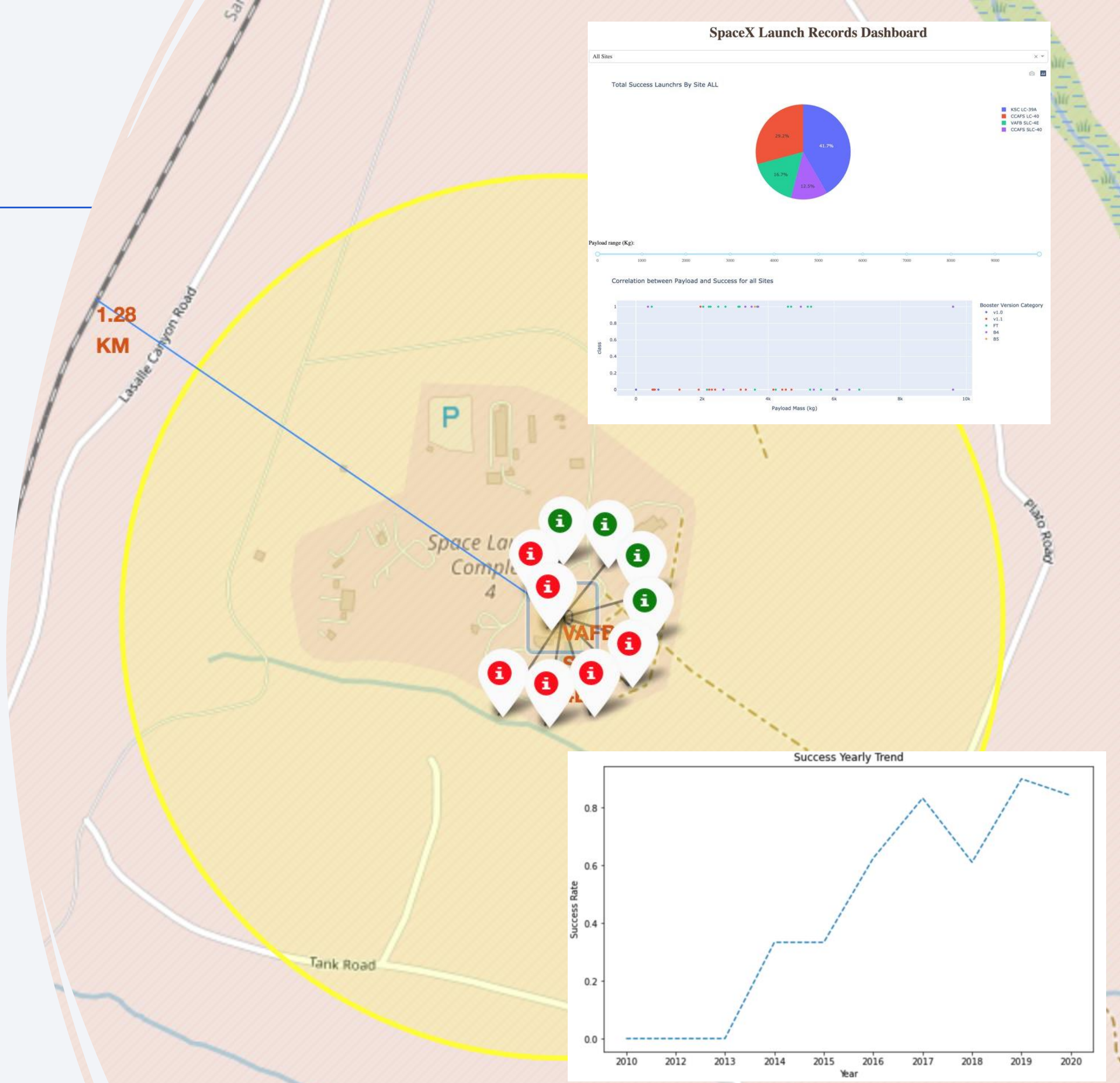
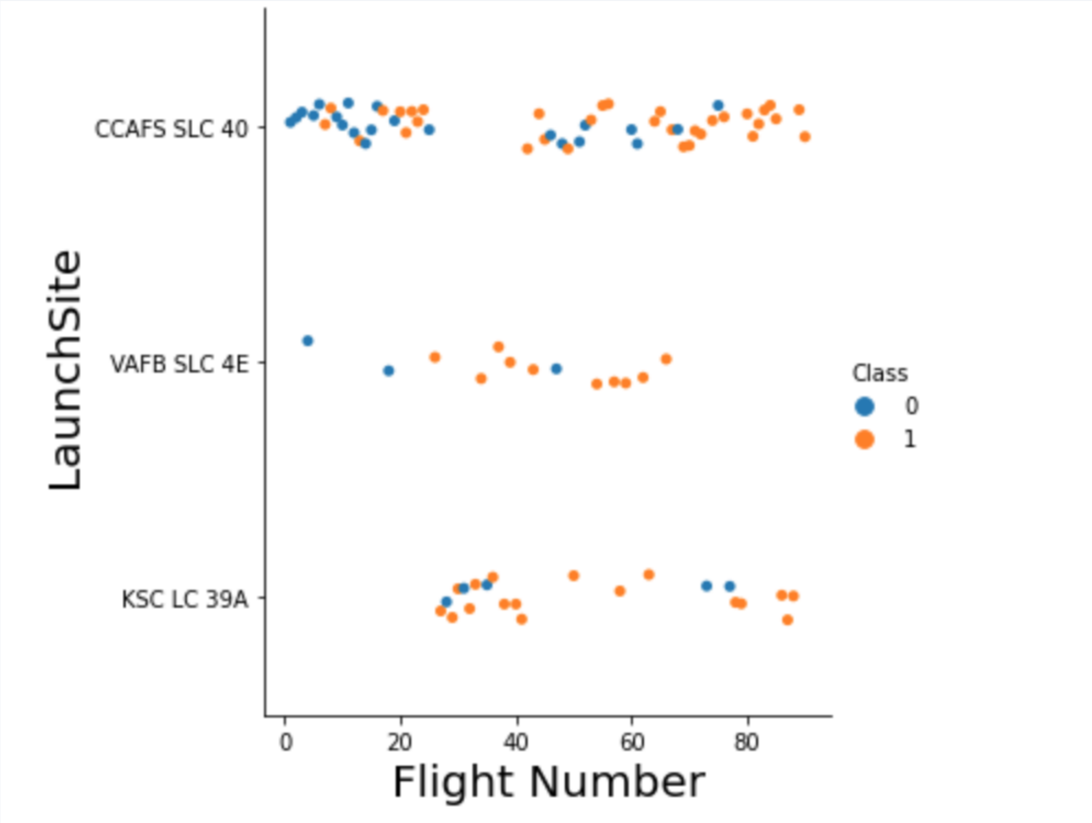


Figure 1: Falcon 9 1st Stage successful land rate.

Section 2

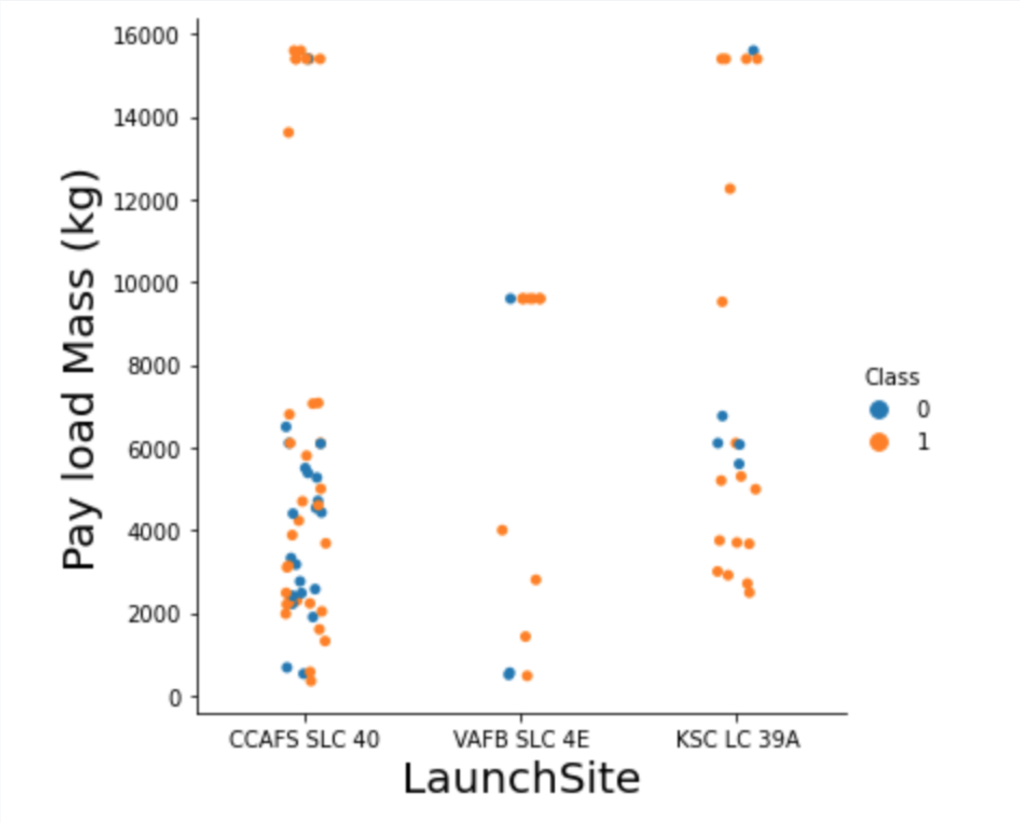
Insights drawn from EDA

Flight Number vs. Launch Site



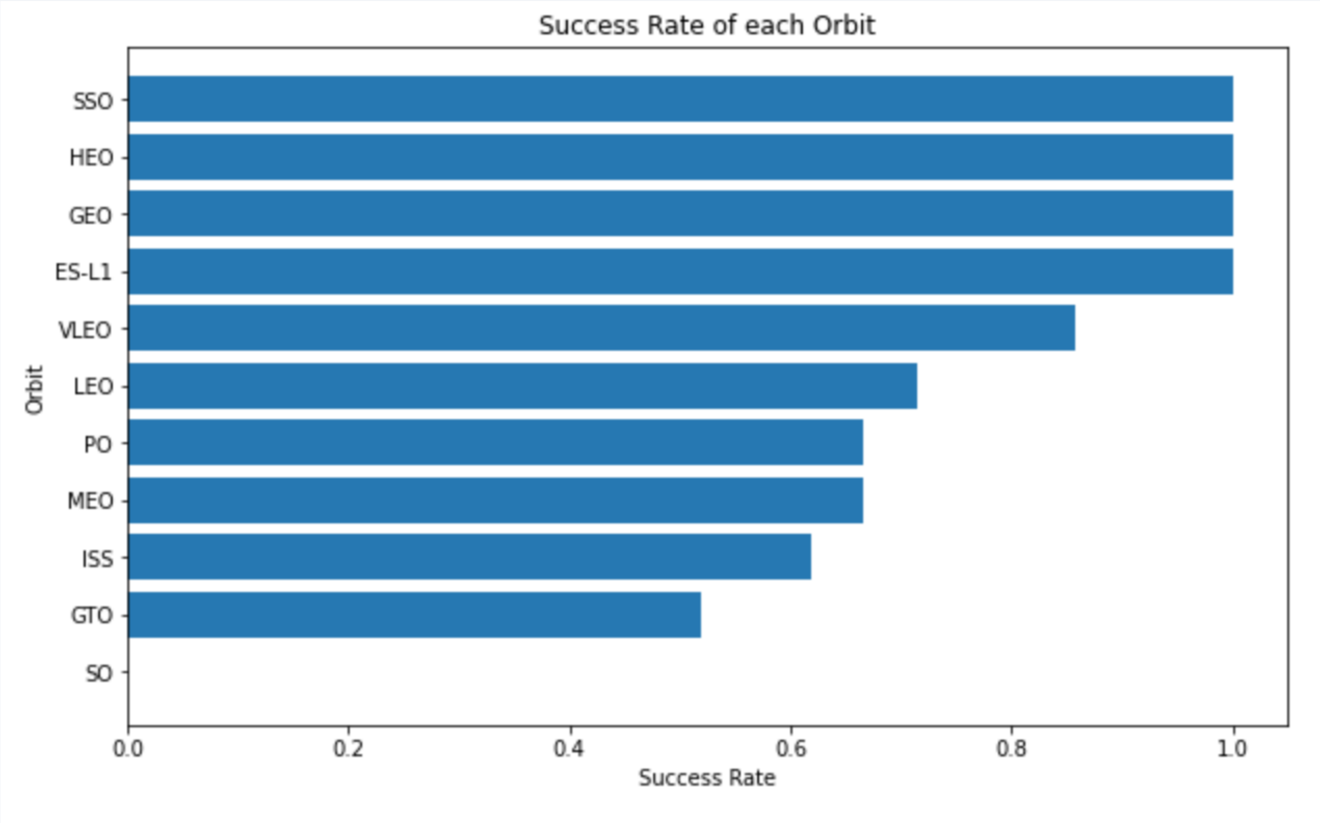
- We see that different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
- Also note VAFB SLC 4E has a much better success rate after the first 2 launces.

Payload vs. Launch Site



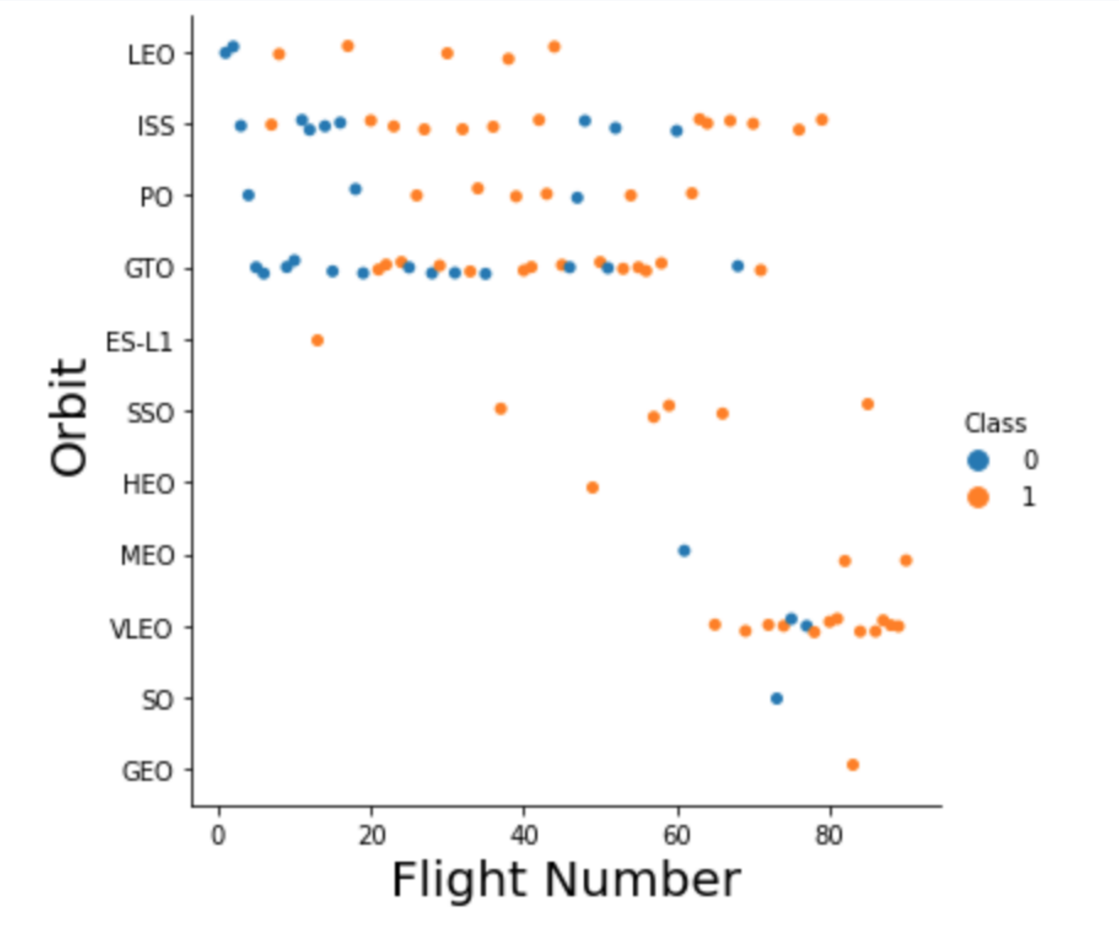
- The higher the payload mass the higher the success rate. The most successful site id VAFB and KSC has a nice seggregation between failed outcomes and positive outcomes.

Success Rate vs. Orbit Type



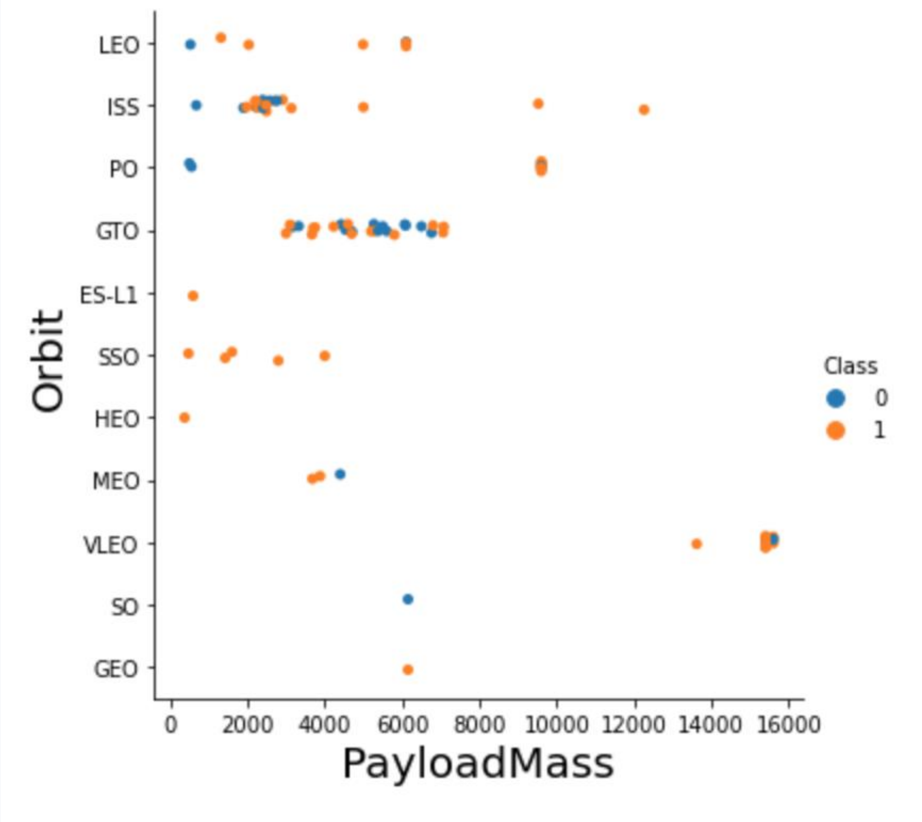
- The SSO, HEO, GEO, and ES-L1 all have a success rate of 100%. However, we do see below that they do not have many flights.

Flight Number vs. Orbit Type



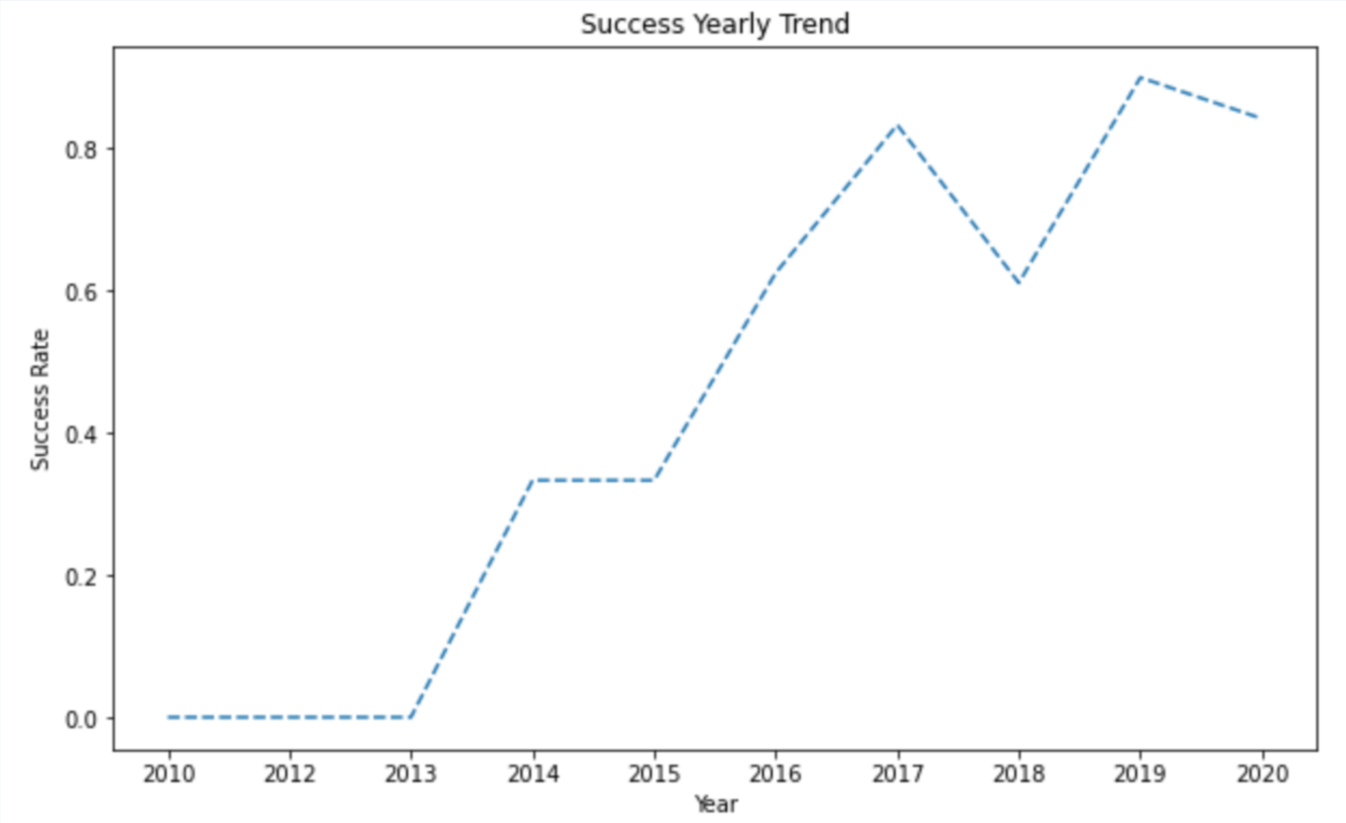
- In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



- Observe that Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

Launch Success Yearly Trend



- Observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

Find the names of the unique launch sites

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

There are 4 unique Launch Sites

Launch Site Names Begin with 'CCA'

Showing the first 5 records with Launcg site beginning with 'CCA'

	DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
0	2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
3	2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	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 carried by boosters from NASA is 45,596 kg.

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2534.66.

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad was 22-12-2015

Successful Drone Ship Landing with Payload between 4000 and 6000

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

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes

Success 99

Success (payload status unclear) 1

Failure (in flight) 1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- 74 F9 B5 B1048.4
- 77 F9 B5 B1049.4
- 79 F9 B5 B1051.3
- 80 F9 B5 B1056.4
- 82 F9 B5 B1048.5
- 83 F9 B5 B1051.4
- 85 F9 B5 B1049.5
- 92 F9 B5 B1060.2
- 93 F9 B5 B1058.3
- 94 F9 B5 B1051.6
- 95 F9 B5 B1060.3
- 99 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

	booster_version	launch_site
13	F9 v1.1 B1012	CCAFS LC-40
16	F9 v1.1 B1015	CCAFS LC-40

```
2010
2010
2012
2012
2013
2013
2013
2014
2014
2014
```

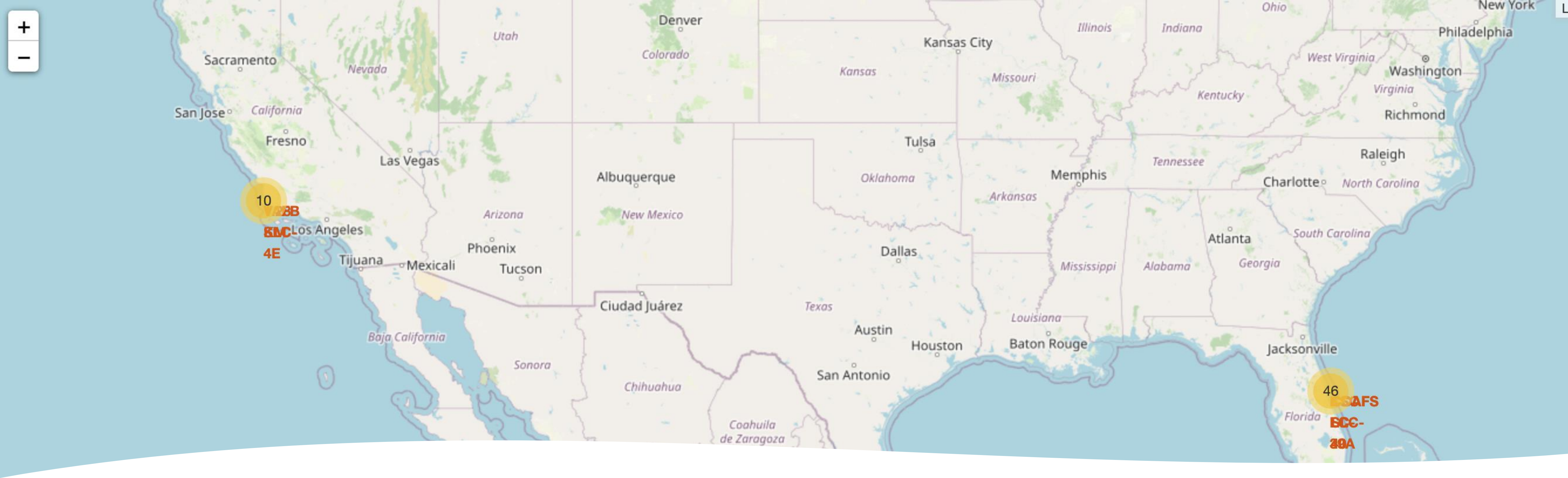
```
array(['Failure (parachute)', 'No attempt', 'Uncontrolled (ocean)',
       'Controlled (ocean)', 'Failure (drone ship)',
       'Precluded (drone ship)', 'Success (ground pad)',
       'Success (drone ship)', 'Success', 'Failure'], dtype=object)
```

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

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order
- No attempt 10
- Success (drone ship) 5
- Failure (drone ship) 5
- Success (ground pad) 3
- Controlled (ocean) 3
- Failure (parachute) 2
- Uncontrolled (ocean) 2
- Precluded (drone ship) 1

Section 4

Launch Sites Proximities Analysis



Launch Site Location

- Important to note there is 1 launch site on the west coast and 3 on the east coast where most launches have happened.

Launch Site Location and Success

- The Green markers represent a successful stage 1 landing and red indicate a failed landing.

Launch Site Location and Proximity

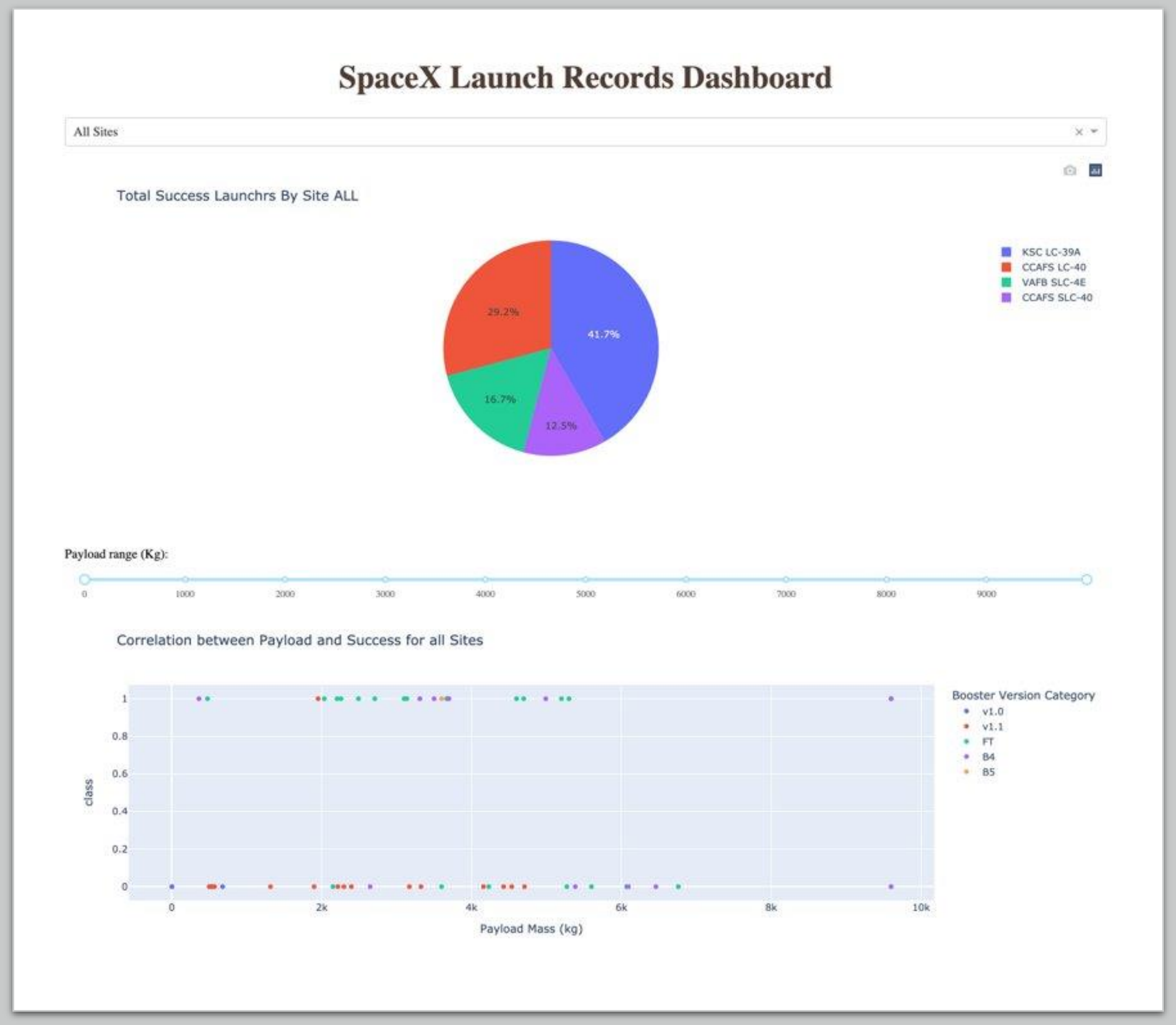
- The line and distance show the launch sites location to the Santa Barbara Subdivision.

Section 5

Build a Dashboard with Plotly Dash

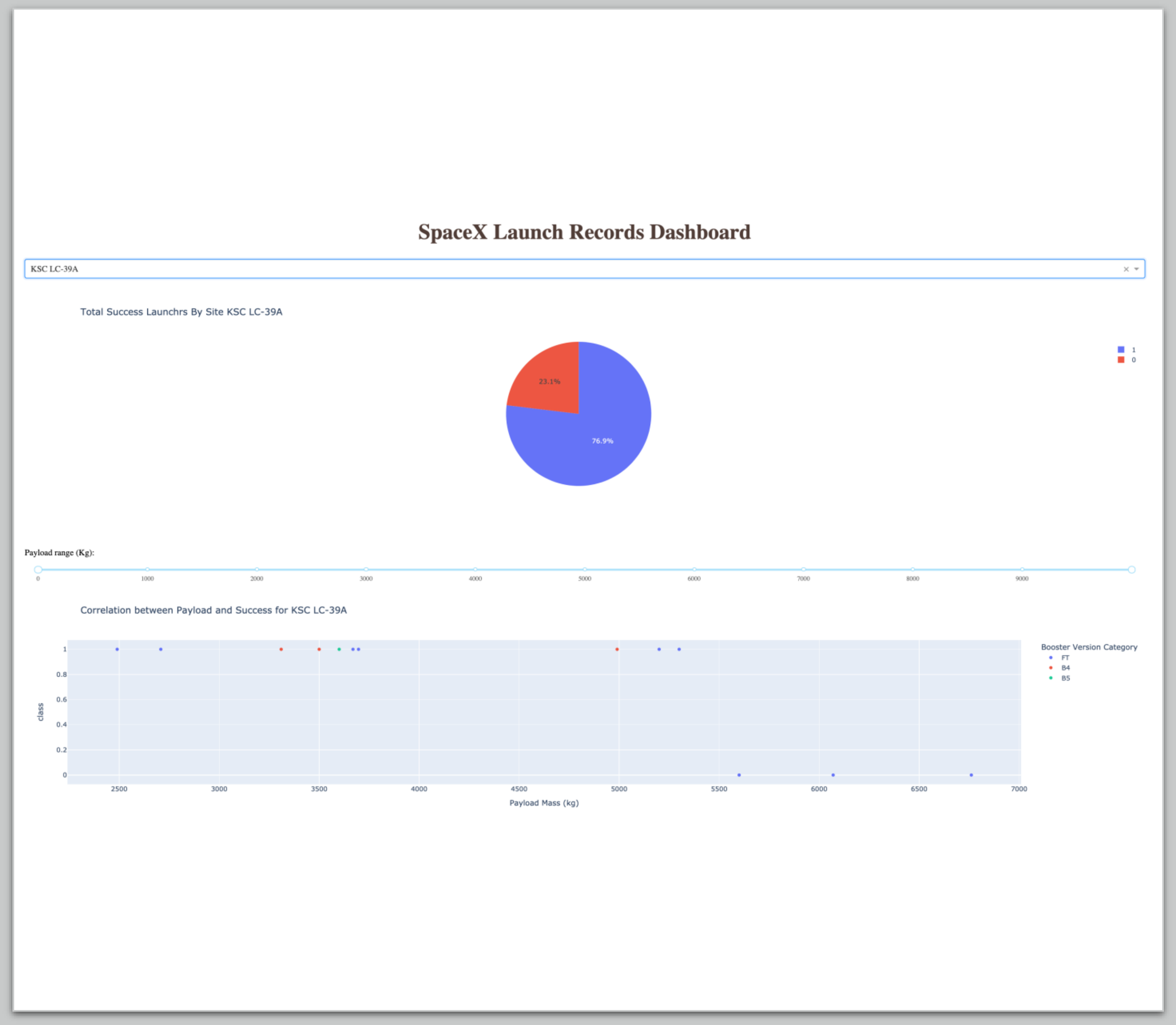
Dashboard Main View

- The vis to the right shows the main dashboard.
- The pie chart shows the 1st stage landing success of the falcon 9 shuttle.
- The scatter plot below shows the success and failures by payload mass which can be toggled via the slider.



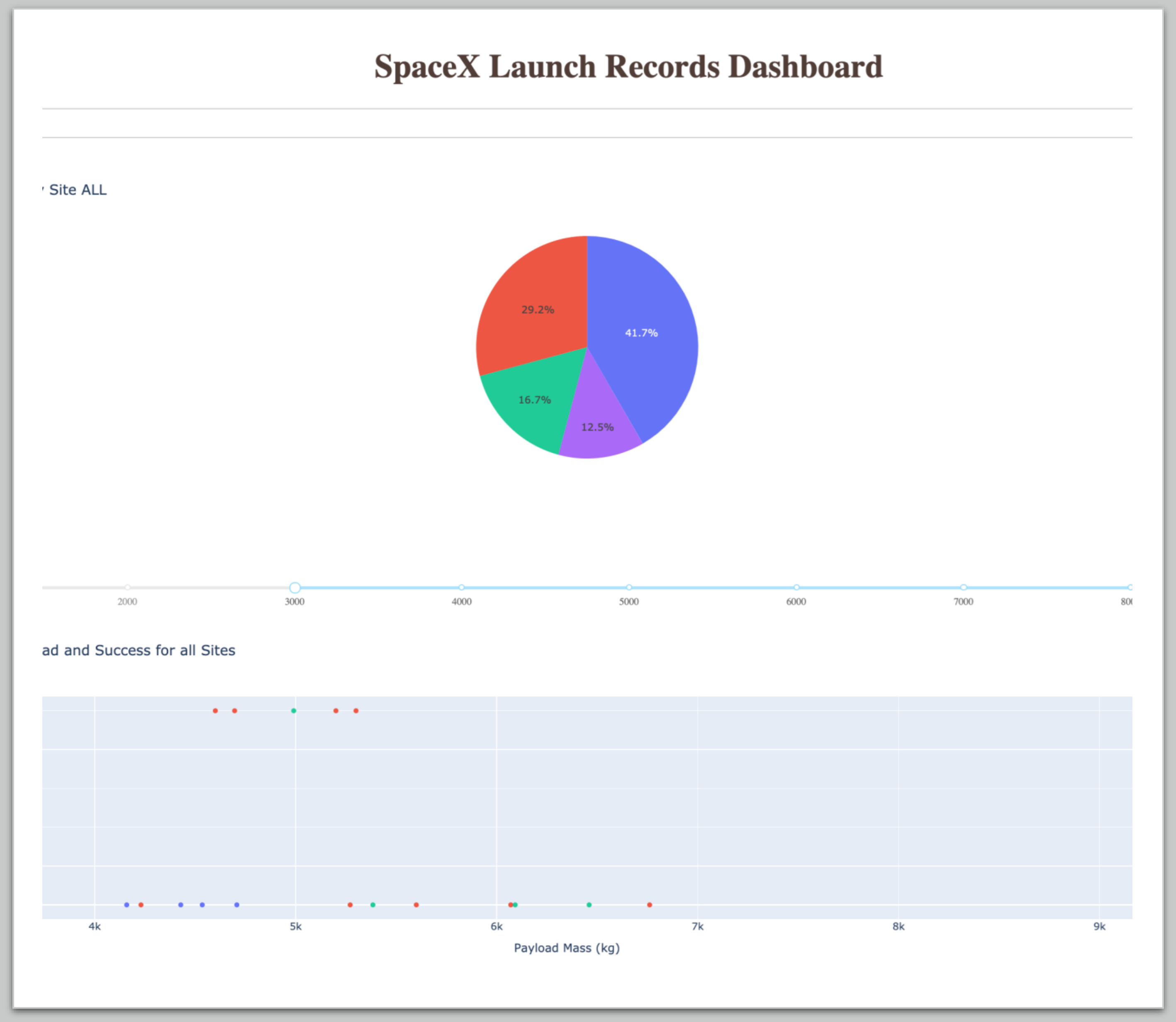
Dashboard Launch Site View

- The vis to the right shows a selected launch site dashboard.
- The pie chart shows the 1st stage landing success of the falcon 9 shuttle for a specific launch site.
- The scatter plot below shows the success and failures by payload mass which can be toggled via the slider.



Dashboard Payload View

- This snapshot shows the functionality of the payload slider which has been toggled to show launches >=3000kg.

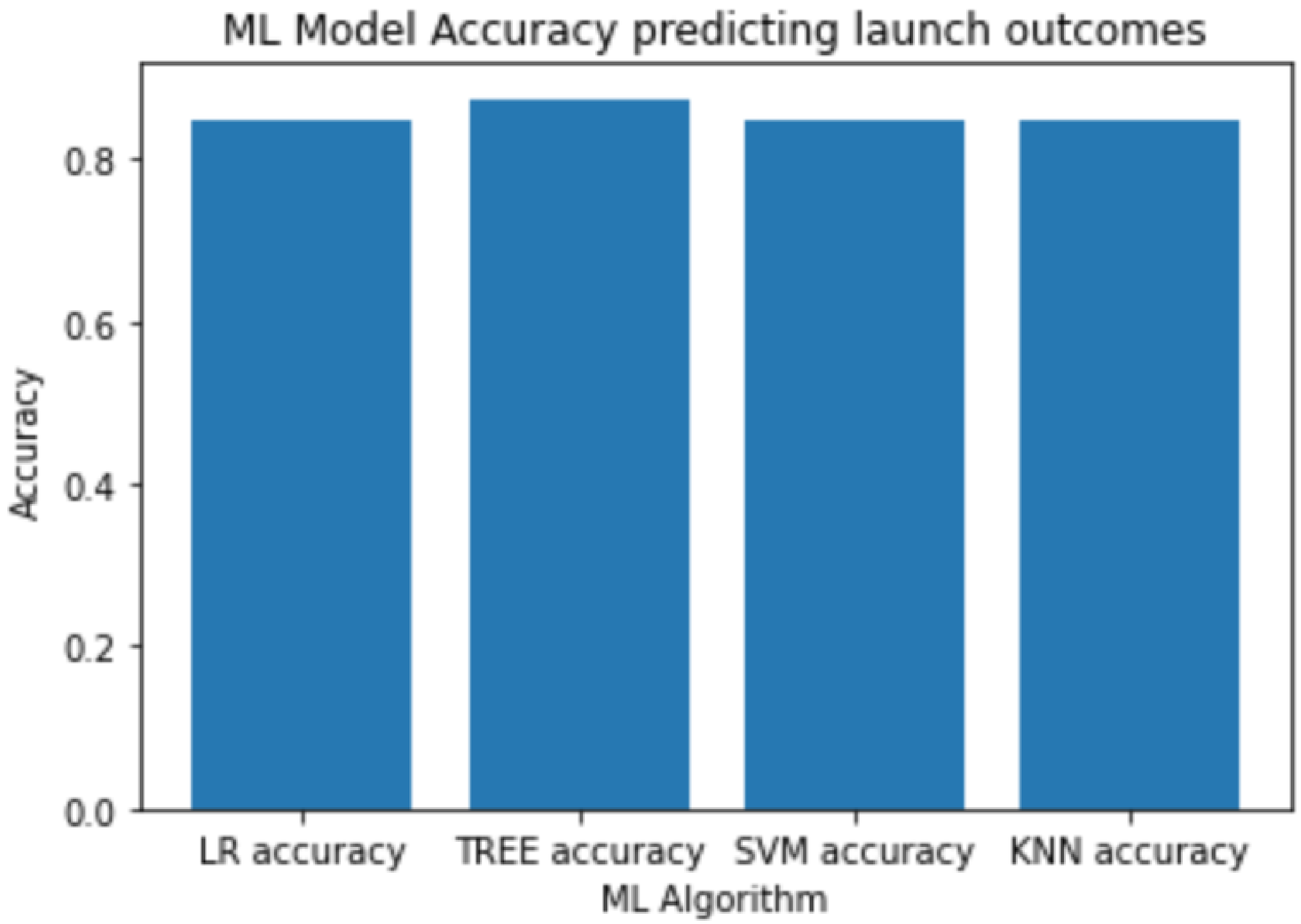


Section 6

Predictive Analysis (Classification)

Classification Accuracy

- The Decision Tree Classifier has the highest model accuracy and performs the best.



Confusion Matrix

- The figure on the right shows the confusion matrix for the Decision Tree classifier on the Test dataset.
- We can see we are 100% at classifying actual landings but do have 3 false positives where the shuttle hasn't landed.

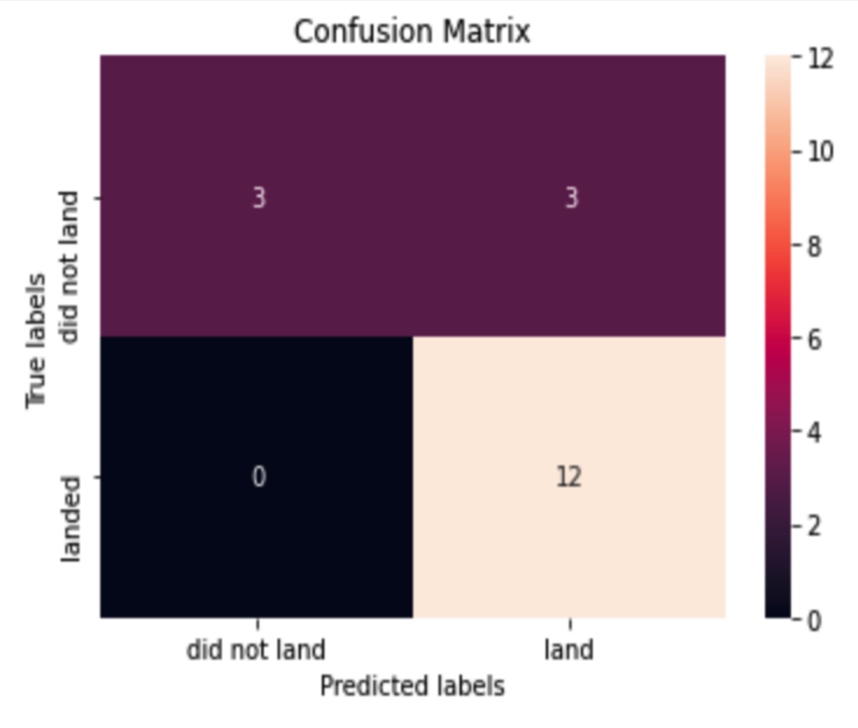


Figure 2: Confusion Matrix on test dataset prediction.

Conclusions

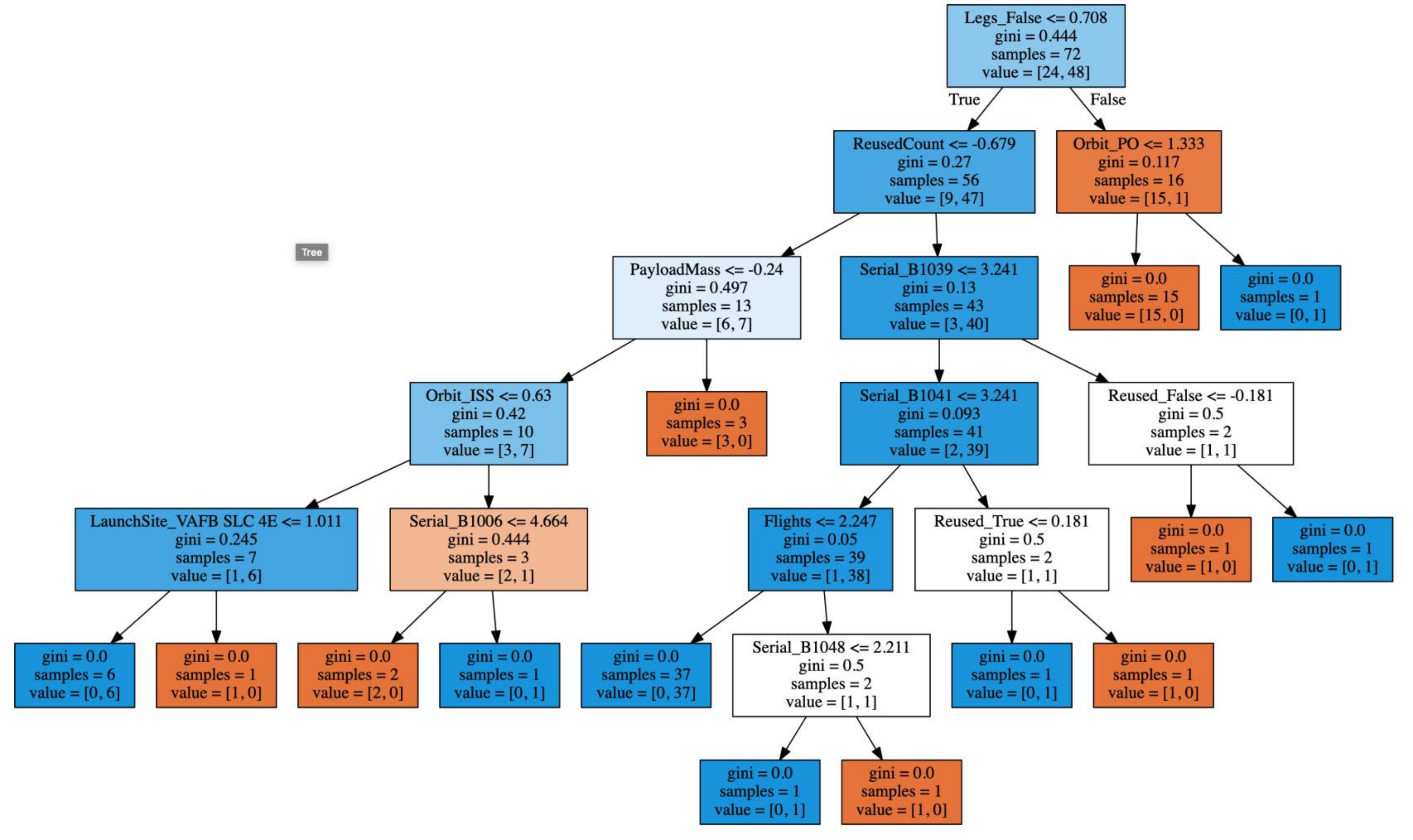
- We can confidently predict the landing outcome for a falcon 9 launch.

Further Considerations

- A further consideration maybe if we were to utilize launch site data, we could also include the weather at the time of launch. Possibly bringing in environmental factors.
- Analysis should be done to check the model isn't overfitting and it would be good to have out of time data however that is not yet available.
- Ther point would be to isolate the most important features although we have shown on the next slide how the decision tree classifier the determining the outcome.
- Further hyperparameter tuning would also be a good idea.

Appendix

- I have included the tree visual showing how the Decision Tree classifier has determined the tree.



Thank you!