IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Rushikesh
Sonawane
February 5th, 2023

Outline

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

Executive Summary

- In this project, we tried to predict the landing outcomes of SpaceX Falcon 9 rocket booster to predict the cost of a launch. We did this by first collecting all the necessary data from SpaceX API and web-scraping, then we do some data wrangling and exploratory data analysis to determine what features best to predict our outcome. We also build a map and interactive dashboard to better understand our data. Finally, we built a machine learning model to predict the landing outcome.
- From our work, we found that Flight Number, Payload Mass, Orbit, Launch Site, Flights, GridFins, Reused, Legs, Landing Pad, Block, Reused Count, and Serial can be used to estimate first-stage landing outcome, and that all SpaceX launch sites are in close proximity to Equator line and coastline. We also decided that the best model to use to predict landing outcome based on our data is the Decision Tree Classifier model

Introduction

The commercial space age is here, companies are making space travel affordable for everyone. Perhaps the most successful company is SpaceX because the **cost of their rocket launches** is relatively inexpensive. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 milliondollars each, much of the savings is because SpaceX can reuse the first stage. Unlike other rocket providers, SpaceX's Falcon 9 Can recover the first stage. Sometimes the first stage does not land.
Sometimes it will crash as shown in this clip. Other times, Space X will sacrifice the first stage due to the mission parameters like payload, orbit, and customer.

Introduction

- For our company to be able to compete with SpaceX, our **goal** is to **determine the price of each launch**. If we can **determine if the first stage will land**, we can determine the cost of a launch.

- We will do this by gathering information about Space X and determine if SpaceX will reuse their first stages. To achieve this, we will **train a machine learning model and use public information to predict if SpaceX will reuse the first stage**.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX launch data that is gathered from SpaceX REST API, Falcon 9 launch records obtained from scraping Wiki webpage.
- Perform data wrangling.
  - Analyzing features that contributes to outcome, creating outcome classes (0 for failed landing, 1 for successful landing)
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building multiple machine learning algorithm with tuned hyperparameters and choosing the best model to implement.

Data Collection

- SpaceX launch data that is gathered from SpaceX REST API. The SpaceX REST API endpoints, or URL, starts with api.spacexdata.com/v4/. We have the different end points, for example: /capsules and /cores.

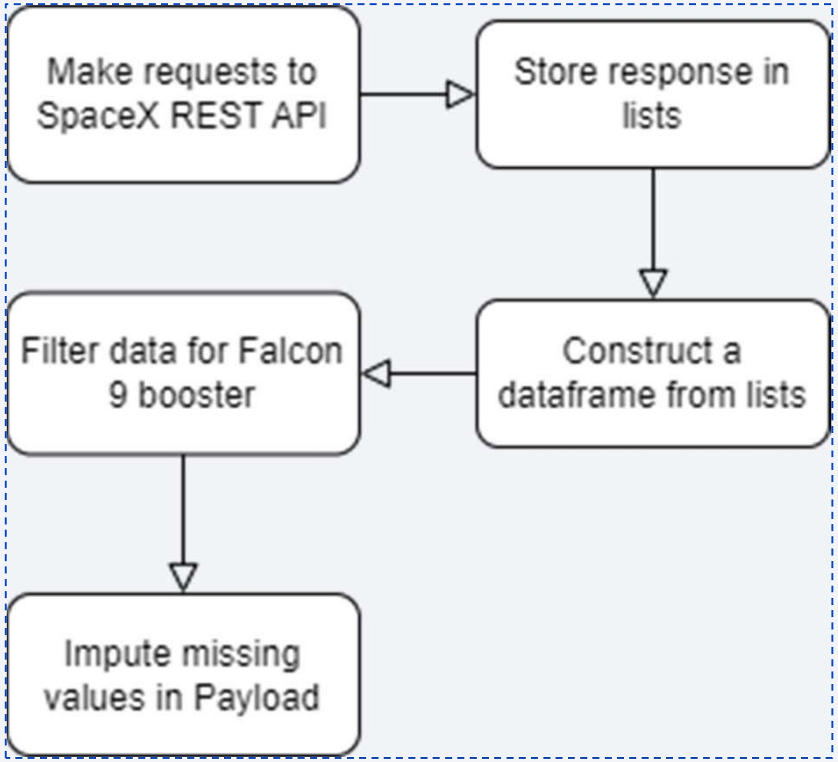
- Falcon 9 Launch data is web-scraped from related Wiki pages then parsed using Beautiful Soup package.

Data Collection - SpaceX API

- SpaceX API gives us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome. We specifically work with the endpoint:

api.spacexdata.com/v4/launches/past

- GitHub URL to SpaceX API calls https://github.com/SonawaneRushi/Data-Science-Castone

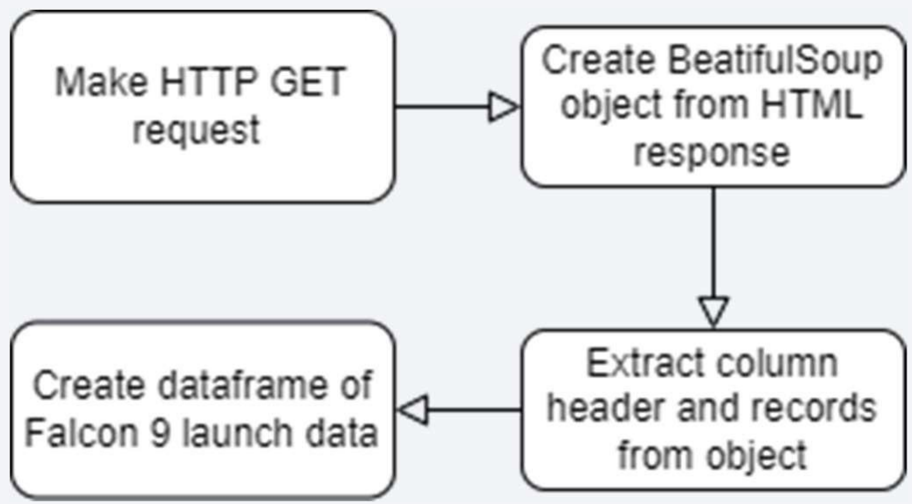


Data Collection - Scraping

- Falcon 9 booster launches data is obtained from:

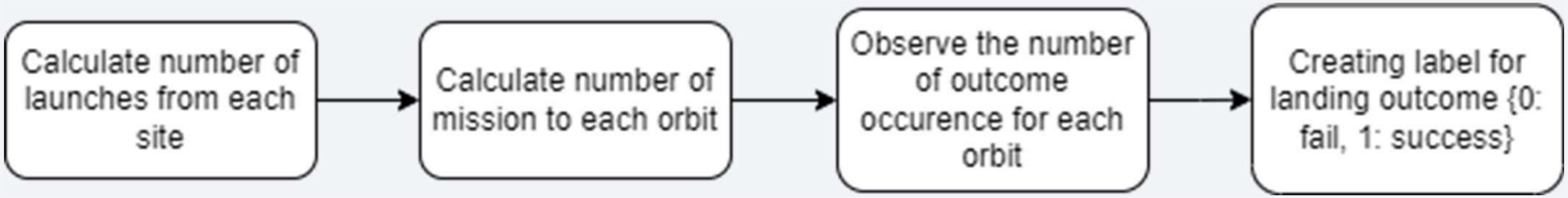
https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

GitHub URL to Falcon 9 launch data web-scraping
https://github.com/SonawaneRushi/Data-Science-Castone



Data Wrangling

- EDA was performed on the obtained data to determine training featured
- GitHub URL to data wrangling process https://github.com/SonawaneRushi/Data-Science-Castone.



EDA with Data Visualization

Plots created:

- Scatter plot of Flight Number (indicating the continuous launch attempts.) vs Payload and overlaying the outcome of the launch to see how flight number and payload mass affect landing outcome
- Relationship between Flight Number and Launch Site with scatter plot to see their effect on landing outcome
- Relationship between Payload and Launch Site to see the outcome pattern for payload mass and launch site pairs
- Relationship between success rate for each orbit type with bar chart
- Relationship between FlightNumber and Orbit type to their respective landing outcome
- Relationship between Payload and Orbit type and their outcome
- Line plot to see the launch success yearly trend

Engineering the features: creating dummy variables, ast all numeric columns to `float64`,

GitHub URL to EDA with data visualization notebook https://github.com/SonawaneRushi/Data-Science-Castone.

EDA with SQL

SQL queries performed:

- Displaying the names of the unique launch sites in the space mission
- Display 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.
- 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.
- Rank the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017

GitHub URL to EDA with data visualization notebook https://github.com/SonawaneRushi/Data-Science-Castone.

Build an Interactive Map with Folium

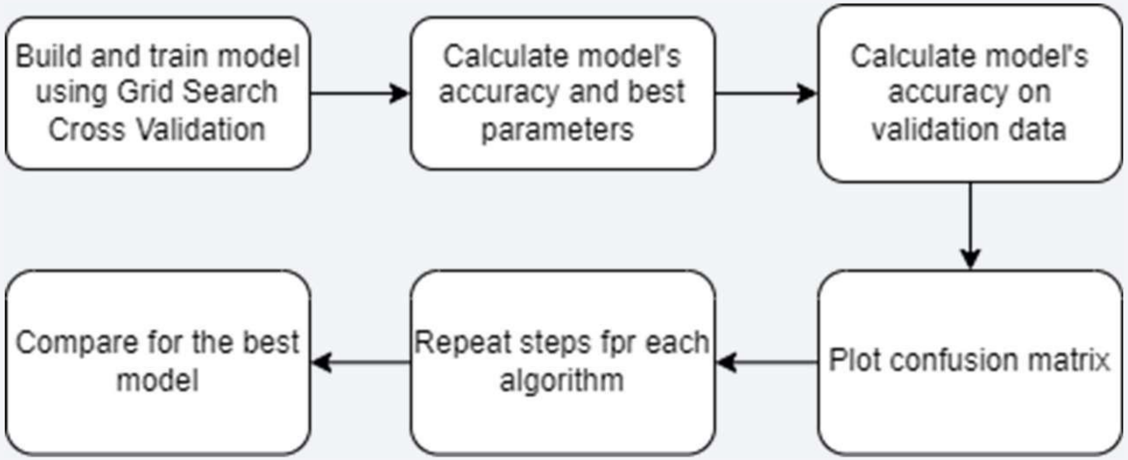
- Mark all launch sites on a map with circle and marker to see if launch sites in proximity to the Equator line and in very close proximity to the coast
- Mark the success/failed launches for each site on the map with MarkerCluster() to identify which launch sites have relatively high success rates.
- Calculate the distances between a launch site to its coastline proximities
- GitHub URL of interactive map with Folium https://github.com/SonawaneRushi/Data-Science-Castone. .

Build a Dashboard with Plotly Dash

- Add a Launch Site drop-down input component and a callback function to render success-pie-chart based on selected site dropdown
- Add a range slider to select payload and a callback function to render the success-payload-scatter-chart scatter plot
- GitHub URL to Plotly Dash lab https://github.com/SonawaneRushi/Data-Science-Castone. .

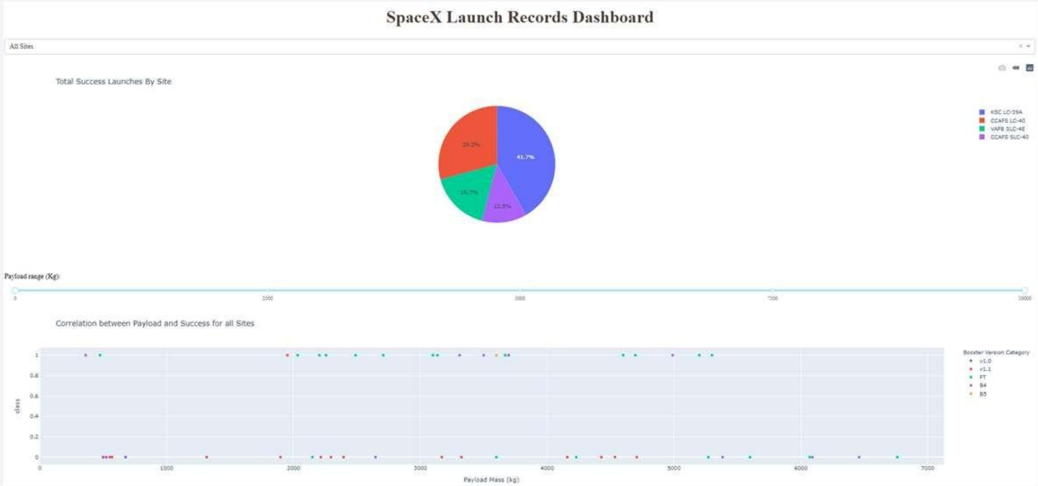
Predictive Analysis (Classification)

- Built multiple algorithm (Logistic Regression, Support Vector Machine, Decision Tree, K-Nearest Neighbor), evaluated, improved, and found the best performing classification model using Grid Search Cross Validation
- GitHub URL of predictive analysis lab https://github.com/SonawaneRushi/Data-Science-Castone.



Results

- From the exploratory data analysis, it is decided that the features to be used to predict outcome are: Flight Number, Payload Mass, Orbit, Launch Site, Flights, GridFins, Reused, Legs, Landing Pad, Block, Reused Count, and Serial



Interactive analytics dashboard demo

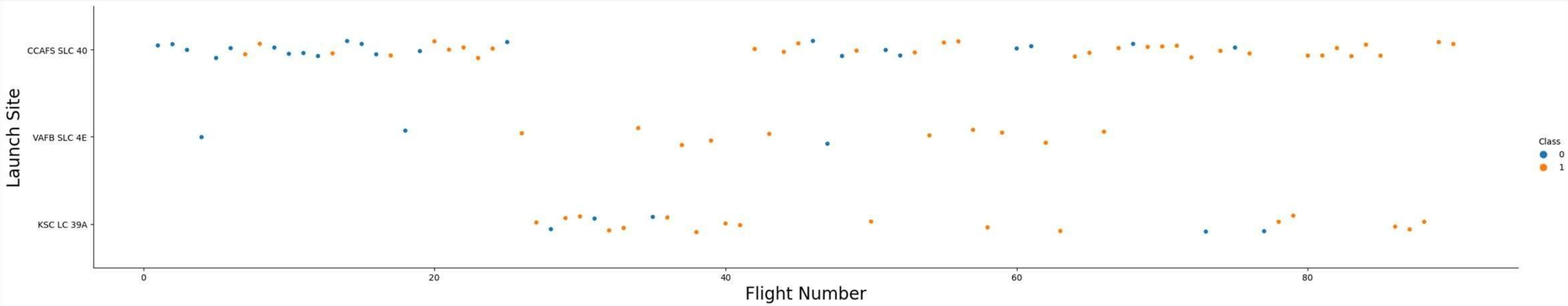
- From predictive analysis, it was found that the best model to use for our data is Decision Tree Classifier. With criterion equals to 'gini', 12 maximum depth, 'auto' as max features, 4 minimum sample leafs, 5 minimum sample split, and 'random' as splitter. The model yield 0.875 as the best accuracy score

Section 2

Insights drawn from EDA

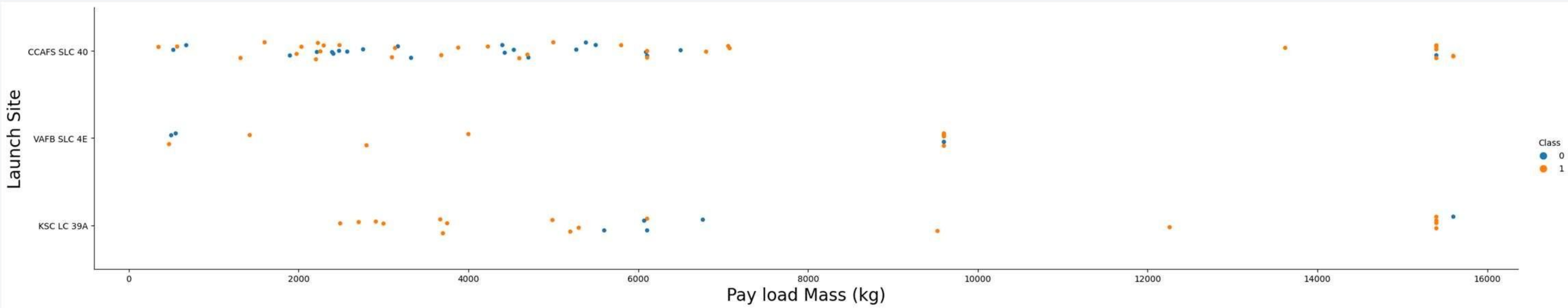
Flight Number vs. Launch Site

- CCAFS SLC-40 is used for many early attempt (lower flight number)
- CCAFS SLC-40 appear to have the lowest landing success rate compared to the other two sites
- As attempt or flight number increase, doesn't always result in successful landing of the first stage



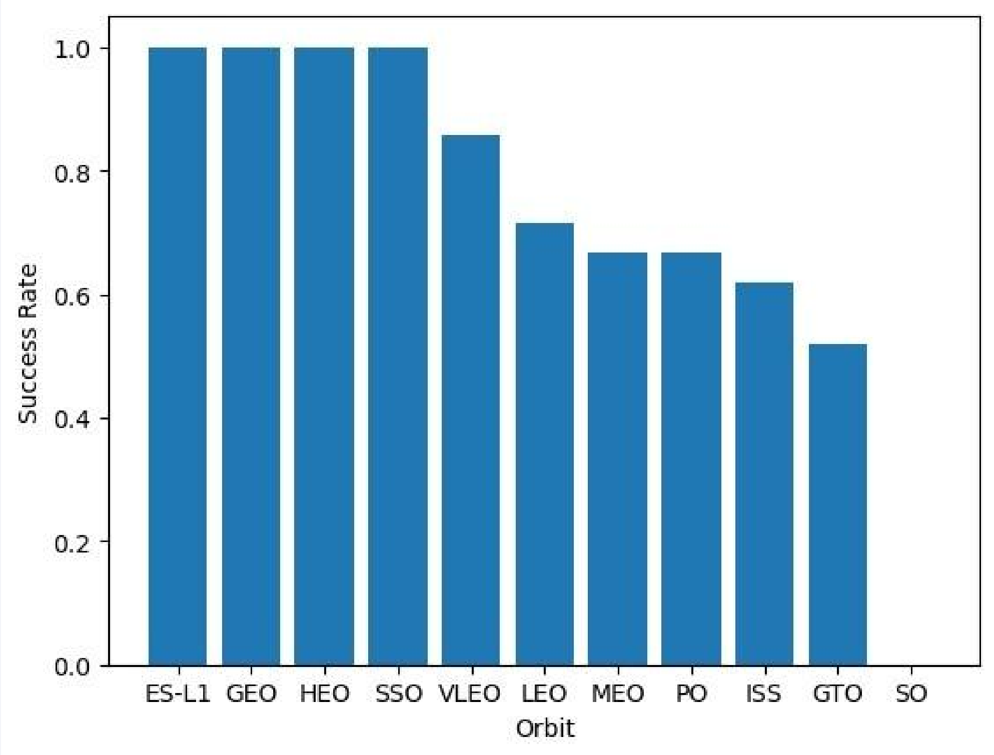
Payload vs. Launch Site

- CCAFS SLC 40 and KSC LC 39A have records of launches with heavy payload mass (>10000 kg), whereas VAFB SLC 4E doesn't (<10000 kg).



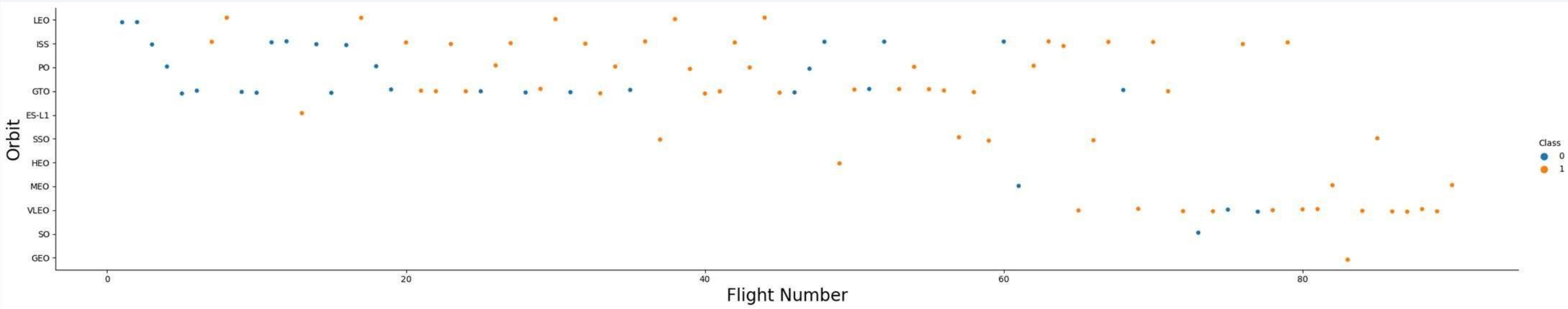
Success Rate vs. Orbit Type

- Some orbit have success rate of 100% (1 fro ES-L1, GEO, HEO, and SSO)
- SO has 0% success rate.



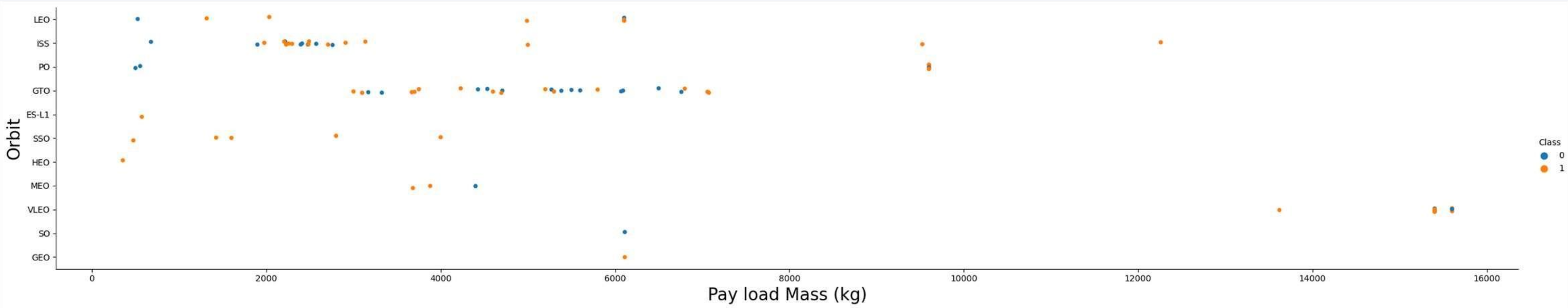
Flight Number vs. Orbit Type

- Some orbit has more rocket launches towards them than the other
- 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.



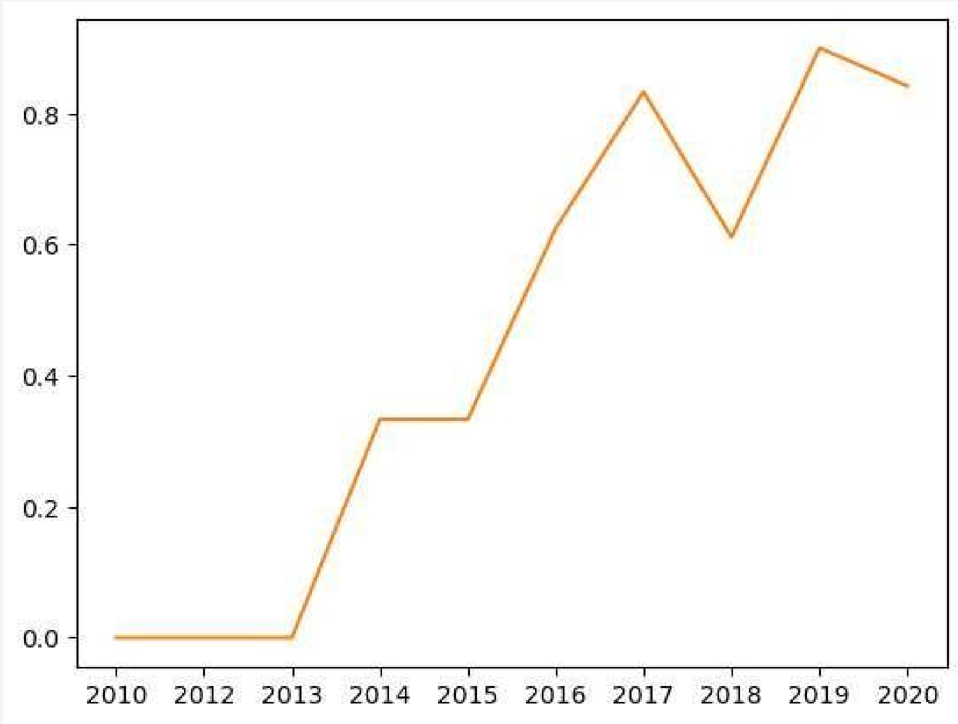
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However, for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.



Launch Success Yearly Trend

- The sucess rate since 2013 kept increasing till 2020



All Launch Site Names

- There are 4 distinct launch sites in our dataset corresponding to rocket launch with Falcon 9 booster.

```
%sql SELECT DISTINCT(launch_site) FROM spacex
```

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

Launch Site Names Begin with 'CCA'

- Below are the first 5 records of launces from site beginning with ‘CCA’
- The record shows launches from CCAFS LS-40, all with successful mission outcome and either failed or no attempt for first stage landing.

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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 carried by all boosters from NASA is 45,596 kg

1
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2,928 kg.

1
2928

First Successful Ground Landing Date

- December 22nd 2015 mark the first successful landing outcome on ground pad

2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- F9 FT B1021 2, F9 FT B1031 2, F9 FT B1022, F9 FT B1026 are the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

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

Total Number of Successful and Failure Mission Outcomes

- There was 1 launch with failed mission outcome, and 100 for successful mission outcomes (1 with unclear payload status)

mission_outcome	number_of_event
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Below are the list of boosters 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

- In year 2015, there are 2 failed landing outcomes outcomes in drone ship, listed below with their booster versions and launch site names.

booster_version	launch_site	landing__outcome	YEAR
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)	2015
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)	2015

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

- Between the date 2010-06-04 and 2017-03-20, No Attempt is the number one landing outcome with 10 counts of event, followed by failed in drone ship and success in drone ship with 5 events each

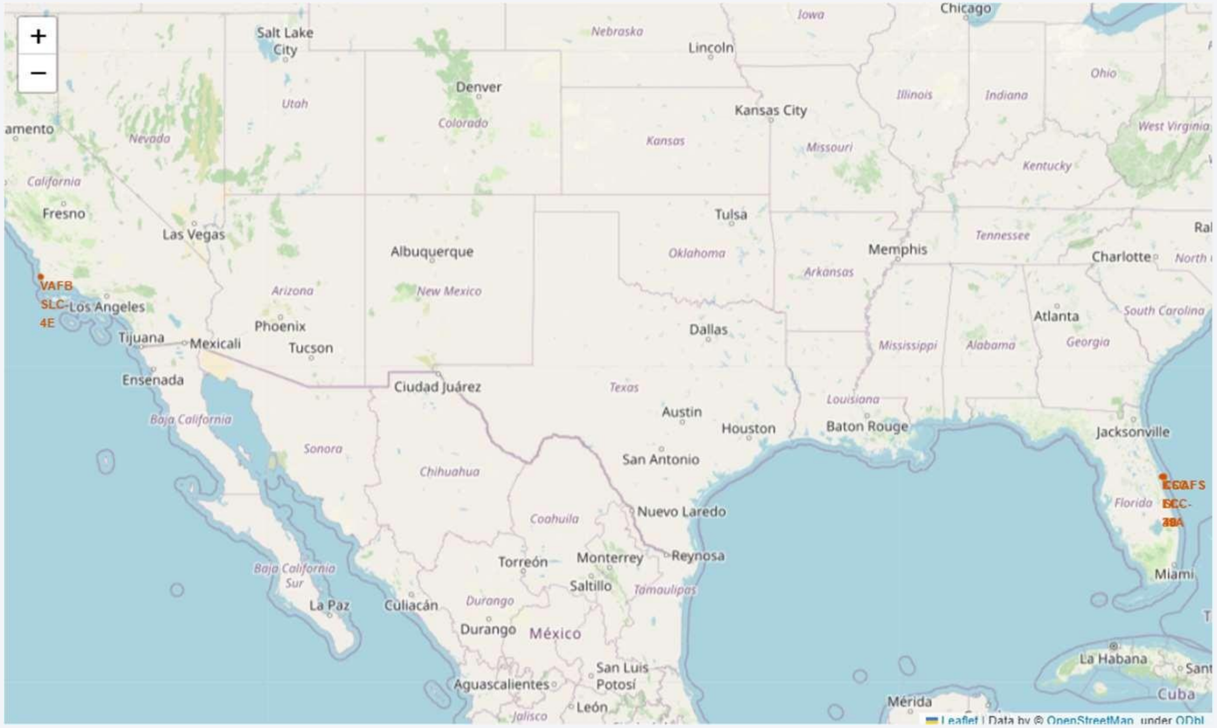
landing__outcome	count_of_event
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

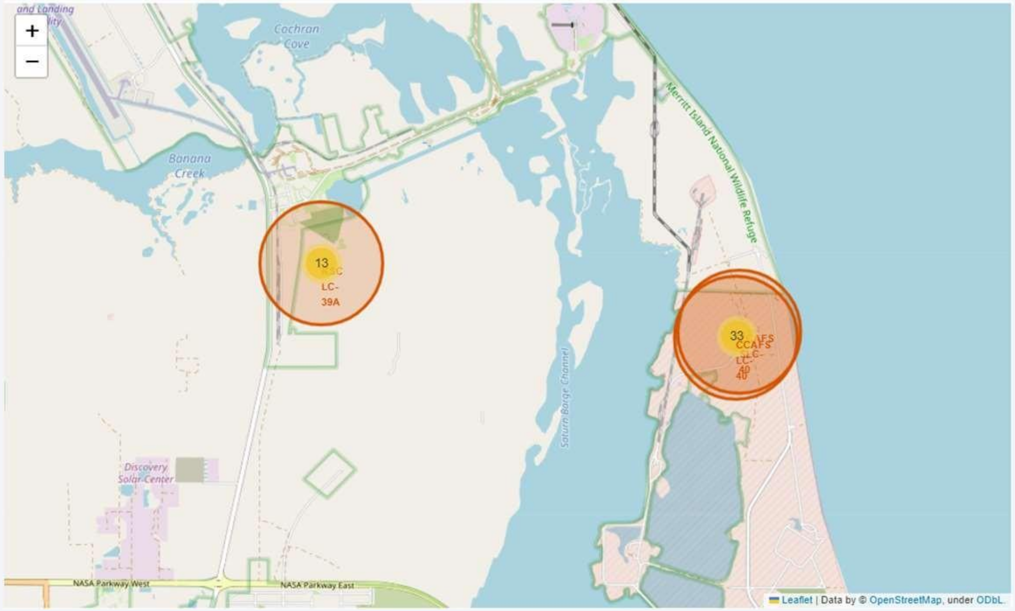
Map of Launch Sites

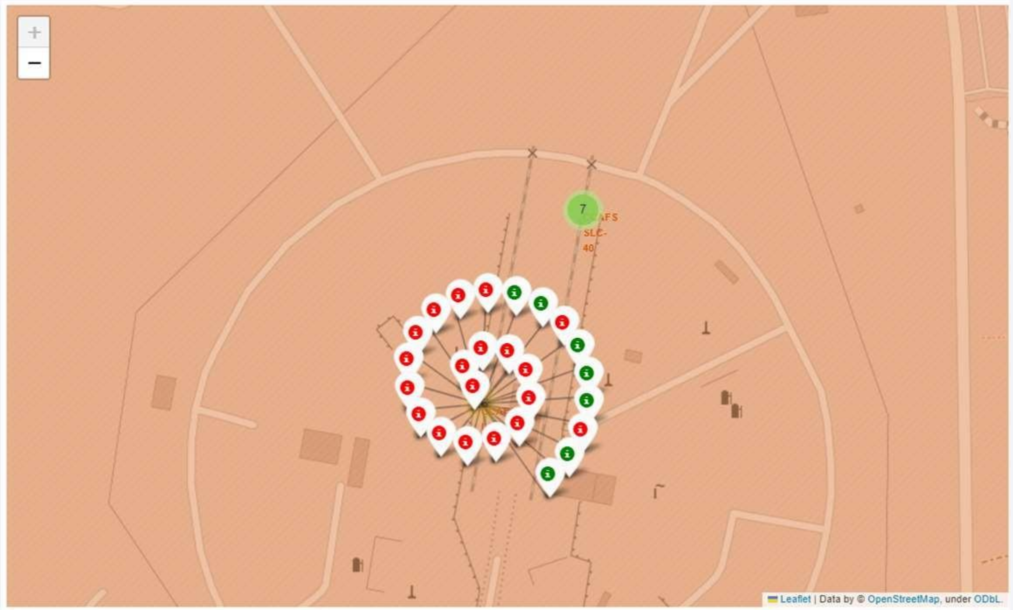
- From the map we can see that all launch sites are in proximity to the Equator line and are all near the coast



Adding marker for each launch in each site

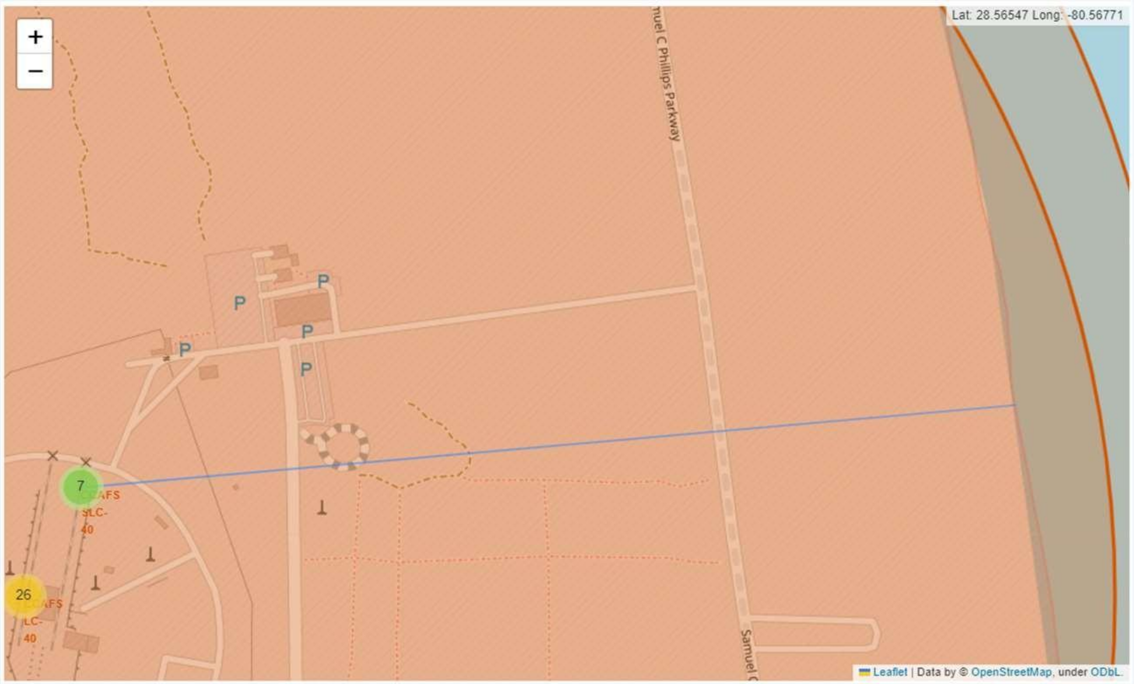
- The CCAFS LC 40 launched the most mission (26)





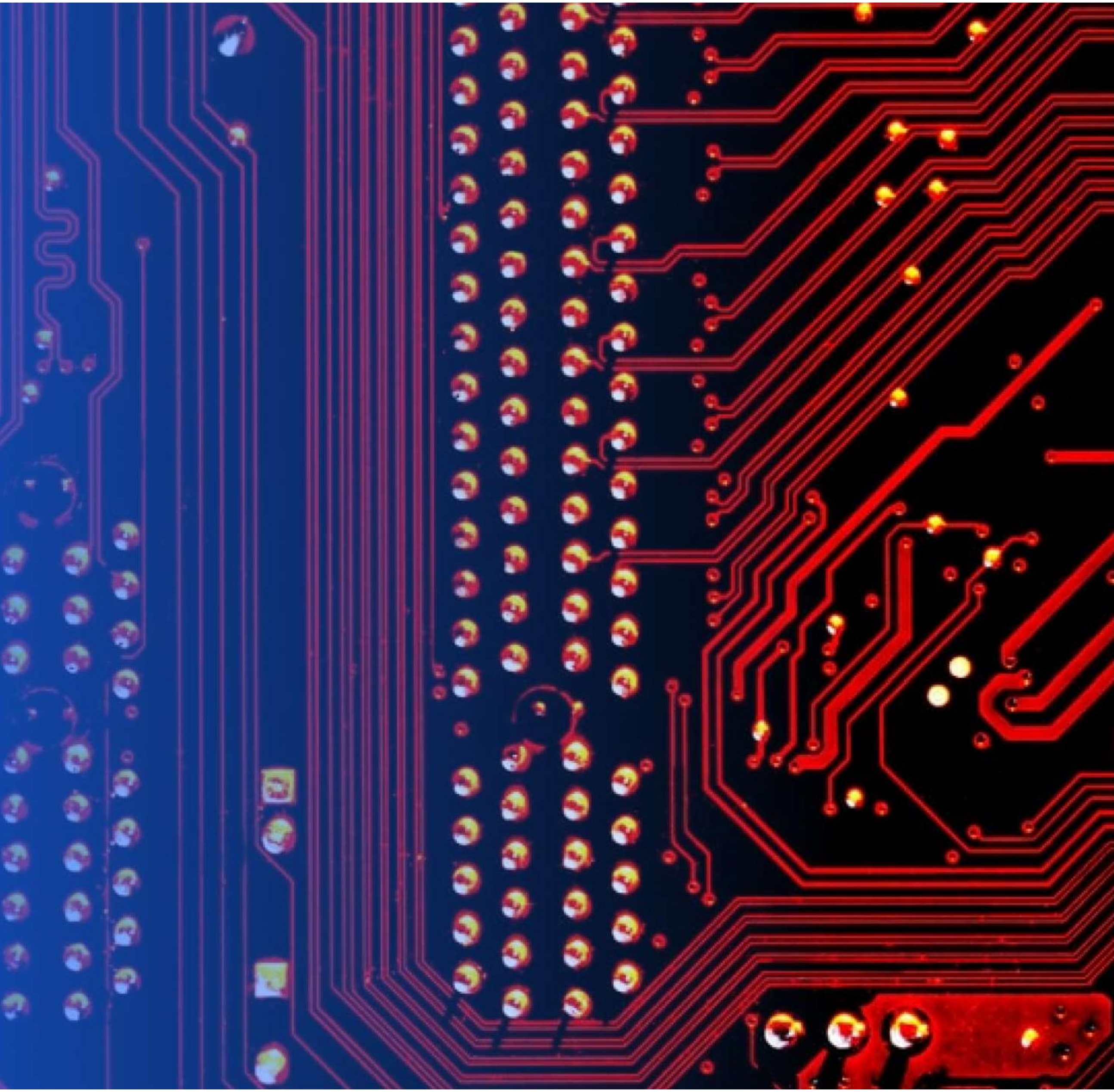
Distance between a launch site to its proximities

- Shown is the map of CCAFS SLC 40 to its coastline proximities



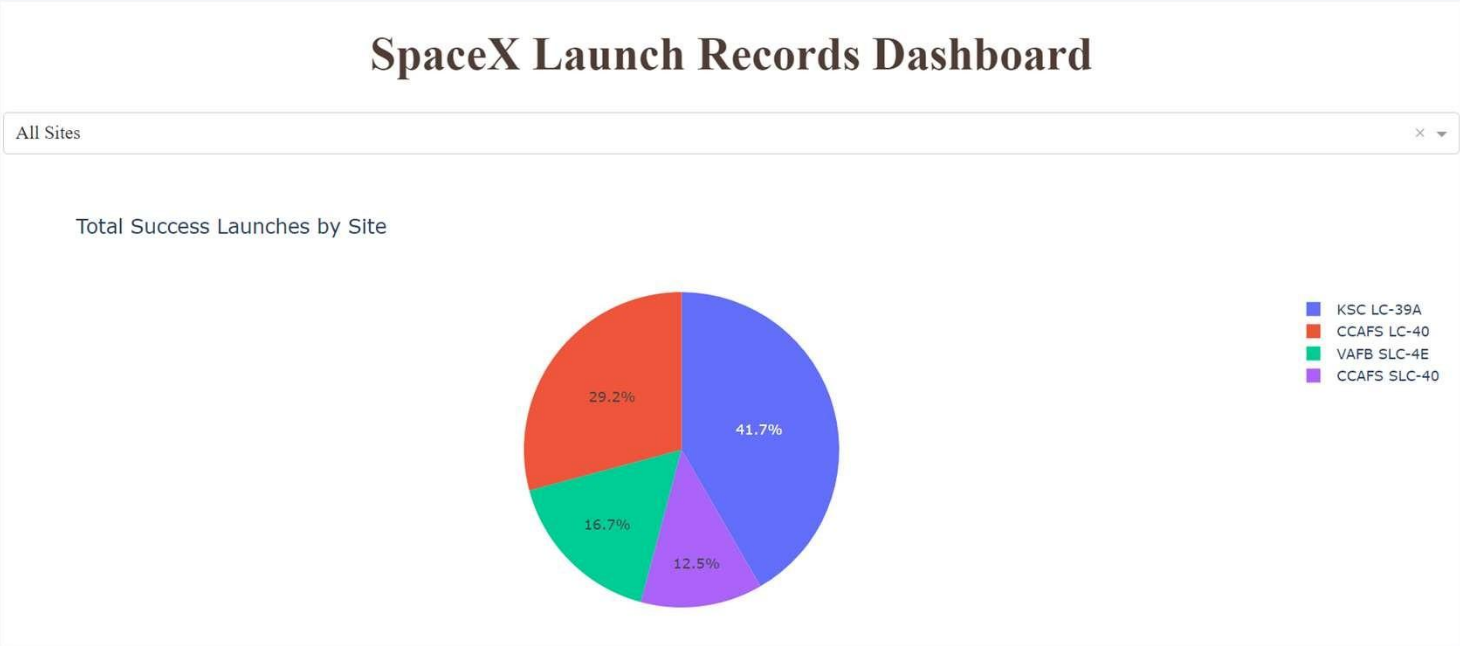
Section 4

Build a Dashboard with Plotly Dash



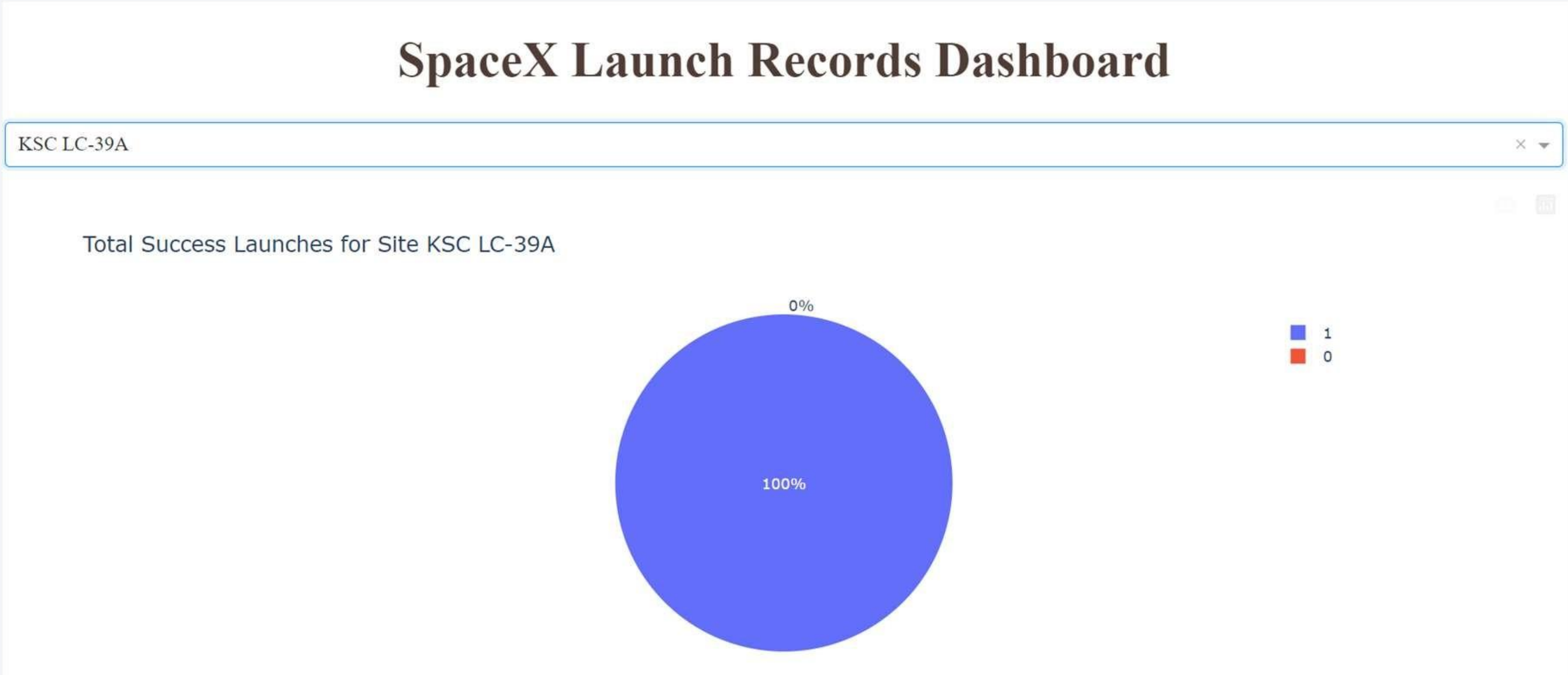
SpaceX Launch Records Dashboard

- The dashboard screenshot shown below showed that KSC LC-39A success launches makes up for 49 of the total successful launches for all sites



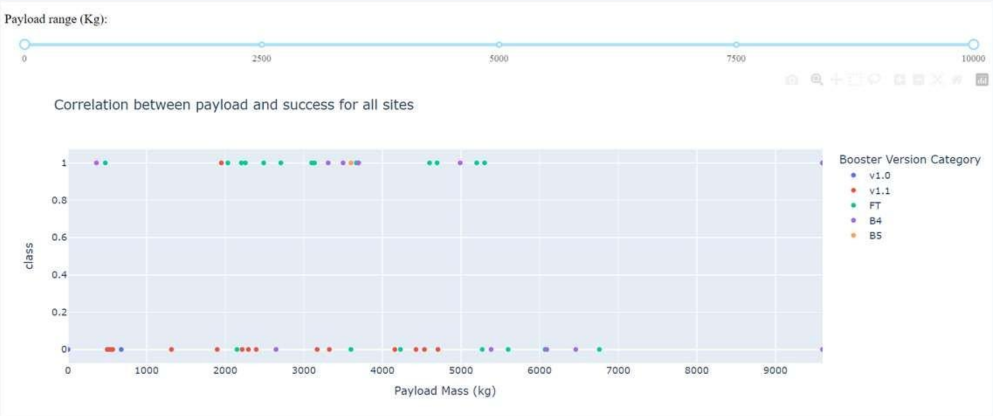
KSC LC-39A

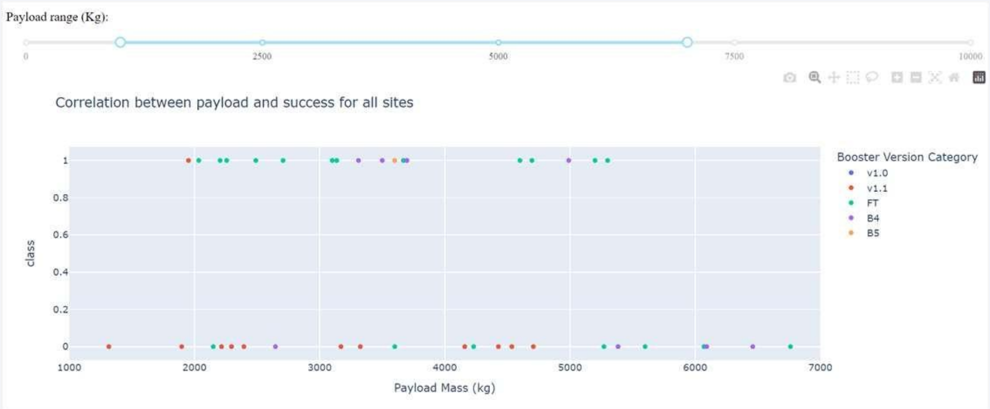
- KSC LC-39 A showed 100% result for class 1, that is successful landing outcome



Correlation Between Payload and Success

- The screenshot below showed the correlation between payload mass and landing outcome for all booster version in all launch sites, shown in different payload mass range.



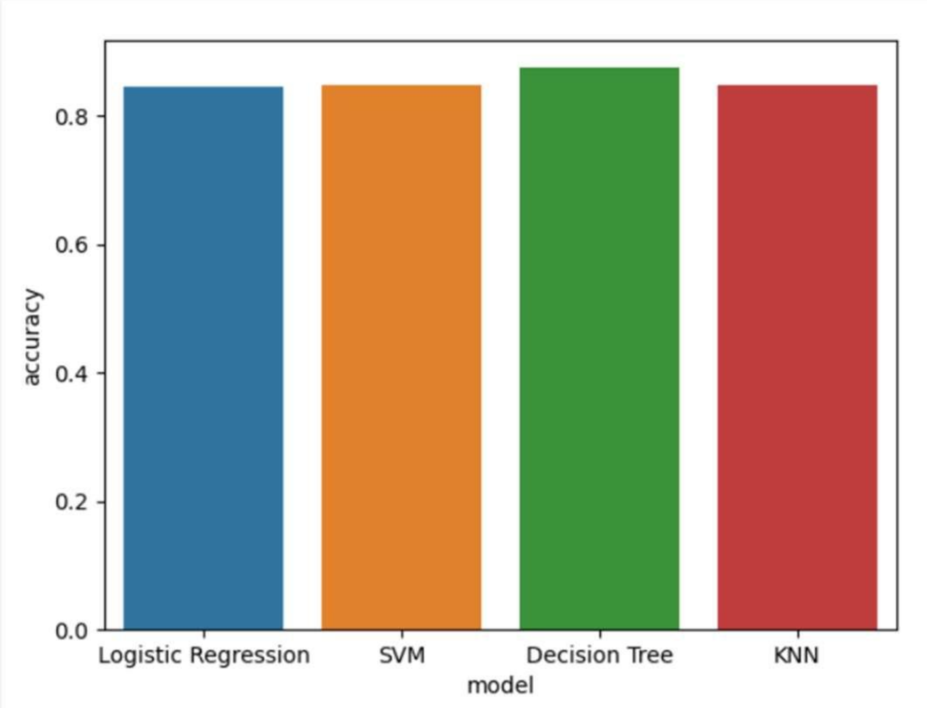


Section 5

Predictive Analysis (Classification)

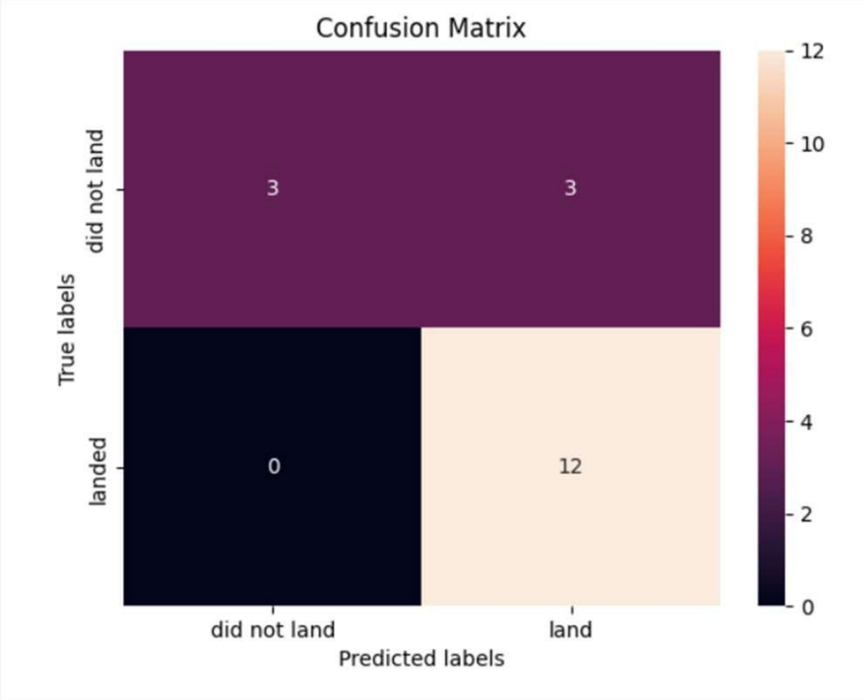
Classification Accuracy

- From our experiment, it was found that Decision Tree has the best accuracy among all models built



Confusion Matrix

- The Decision Tree model built successfully predicted all observation with the true values of landed. However, the model’s problem is its false positives, that is the observation that does lot land, but was predicted by the model to land



Conclusions

- Flight Number, Payload Mass, Orbit, Launch Site, Flights, GridFins, Reused, Legs, Landing Pad, Block, Reused Count, and Serial can be used to estimate first-stage landing outcome
- All SpaceX launch sites are in close proximity to Equator line and coastline
- The best model to use to predict landing outcome based on our data is the Decision Tree Classifier model

Appendix

```
%%sql
SELECT distinct(booster_version)
FROM spacex
WHERE (landing__outcome= 'Success (drone ship)') and (payload_mass__kg_ between 4000 and 6000)
```

```
%%sql
SELECT booster_version, launch_site, landing__outcome, EXTRACT(YEAR from date) as year
FROM spacex
WHERE (landing__outcome= 'Failure (drone ship)') and (EXTRACT(YEAR from date)=2015)
```

```
%%sql
SELECT mission_outcome, count(*) as number_of_event
FROM spacex
GROUP BY mission_outcome
```

```
%%sql
SELECT landing__outcome, count(*) as count_of_event
FROM spacex
WHERE date between '2010-06-04' and '2017-03-20'
GROUP BY landing__outcome
ORDER BY count(*) DESC
```

```
%%sql
SELECT sum(payload_mass__kg_)
FROM spacex
GROUP BY customer
HAVING customer='NASA (CRS)'
```

```
%%sql
SELECT AVG(payload_mass__kg_)
FROM spacex
GROUP BY booster_version
HAVING booster_version='F9 v1.1'
```

```
%%sql
SELECT min(date)
FROM spacex
WHERE landing__outcome like 'Success%'
```


Thank you!