IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

SAJULAL S L
30 September 2022

Outline

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

Section 1

Methodology

Executive Summary

- **Summary of methodologies**
  - Data collection through SpaceX API and web scraping from Wikipedia
  - Data wrangling and cleaning
  - Exploratory Data Analysis (EDA) using SQL, Data Visualization, GIS (FOLIUM) and Building an interactive dashboard
  - Machine Learning Predictive Analytics using different classification models – Classification Tree, Logarithmic Regression, SVM, KNN
- **Summary of all results**
  - EDA allowed to identify which features are the best to predict success of launchings;
  - Machine Learning Prediction showed the best model to predict which characteristics are important to drive this opportunity by the best way, using all collected data
  - Decision Tree Classifier can be used to predict successful landings and increase profits
  - The Decision tree classifier is the best machine learning algorithm for this task with an Accuracy of 90%.

Introduction

Project background and context

The commercial space age enables companies to make space travel affordable for everyone. There are several players like Virgin Galactic, Rocket Lab, Blue Origin and perhaps the most successful is SpaceX.
SpaceX's accomplishments include sending spacecraft to the International Space Station, Starlink, a satellite internet constellation providing satellite Internet access, Sending manned missions to Space, etc.
One reason SpaceX can do this is the rocket launches are relatively inexpensive. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.
The first stage is quite large, does most of the work and expensive. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch.

Problems you want to find answers

As a data scientist for Space Y, that would like to compete with Space X the following problems need to be answered.

- Determine the price of each launch by gathering information about Space X and creating dashboards.
- Predict whether Space X will reuse the first stage by training a machine learning model and using public information.

Methodology

Executive Summary

- Data collection methodology:
  - Collect rocket launch data via REST-API from SpaceX, get data from web scraping on Wikipedia
- Perform data wrangling
  - Modify and add Columns, remove outliers, transform, discuss missing data, One-Hot Encoding etc.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Split the data into Test/Train and then train several classification models (Classification Tree, Logarithmic Regression, SVM, KNN) to predict success or failure of recovery.
  - For each model the best hyper parameters are found using cross-validation
  - We finally compare the accuracy of these models using test data

Data Collection

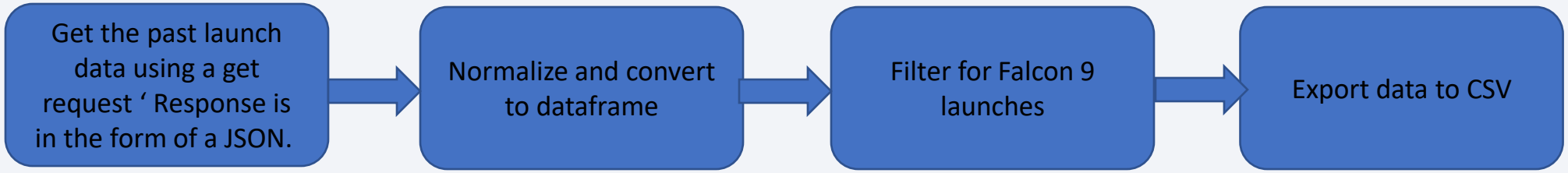
- **Data Collection “SpaceX API”**
  SpaceX provides a REST-API (https://api.spacexdata.com/v4/) where several JSON-files (data about launches, rocket used, payload delivered, launch specifications, landing specifications, and landing outcome) could be downloaded. These where combined, filtered for the Falcon9 data and then exported in csv format.

- **Data Collection “Web Scraping - Wikipedia”**
  Python libraries were used to retrieve the HTML of the webpage (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches) and the launch related data was extracted using the beautifulSoup package. The output was parsed and converted to a Pandas dataframe, filtered for Falcon 9 launches and exported to a csv file.

Data Collection – SpaceX API

Source Code

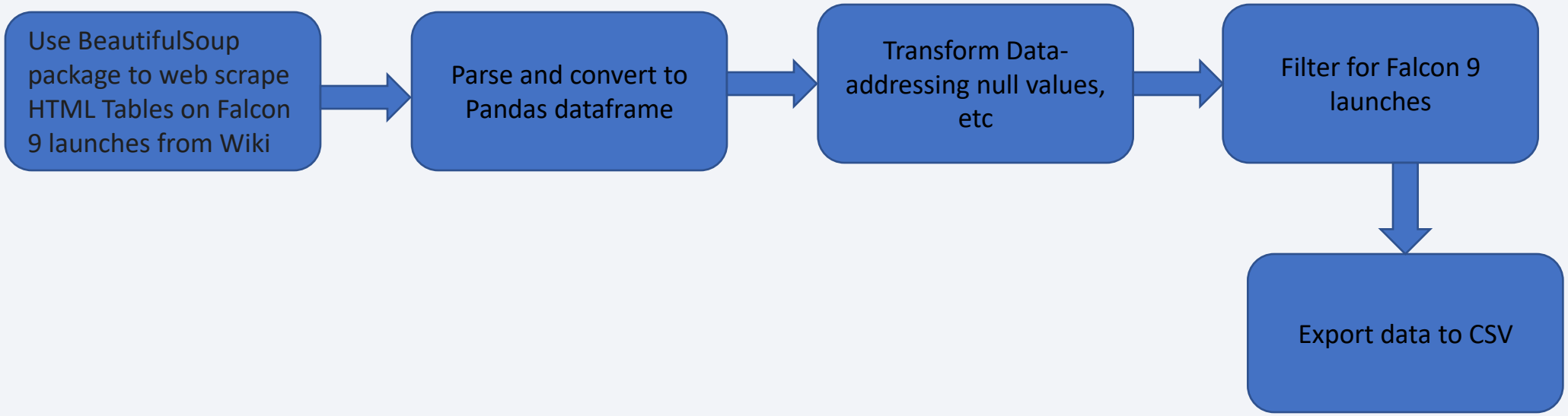
https://github.com/SAJULALSL/IBM-Capstone-Project/blob/master/Data%20Collection%20API.ipynb



Data Collection - Scraping

Source Code

https://github.com/SAJULALSL/IBM-Capstone-Project/blob/master/Data%20Collection%20with%20Web%20Scraping.ipynb



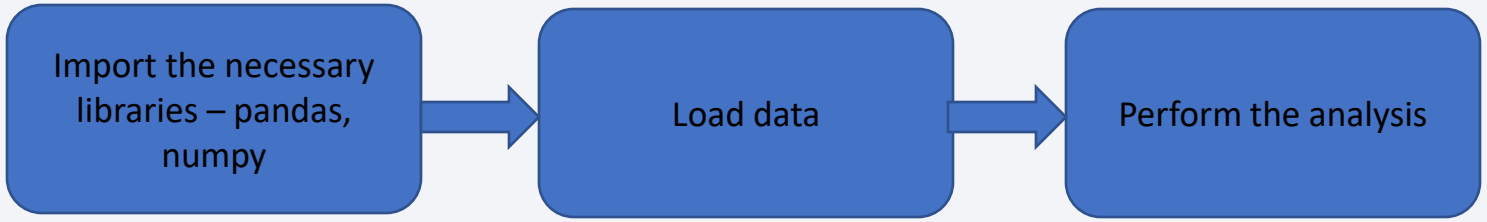
Data Wrangling

Source Code

https://github.com/SAJULALSL/IBM-Capstone-Project/blob/master/Data%20Wrangling.ipynb

Some Exploratory Data Analysis (EDA) was performed as mentioned below to find some patterns in the data and determine what would be the label for training supervised models

1. Calculate the number of launches on each site
2. Calculate the number and occurrence of each orbit
3. Calculate the number and occurrence of mission outcome per orbit type
4. Create a landing outcome label from Outcome column



EDA with Data Visualization

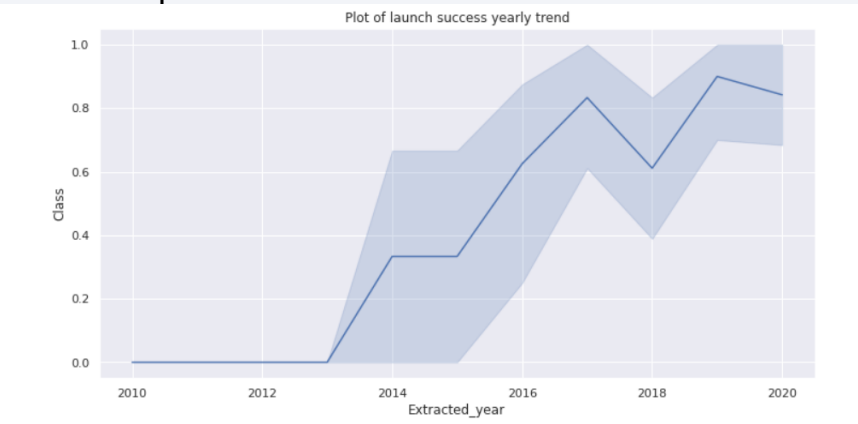
Source Code

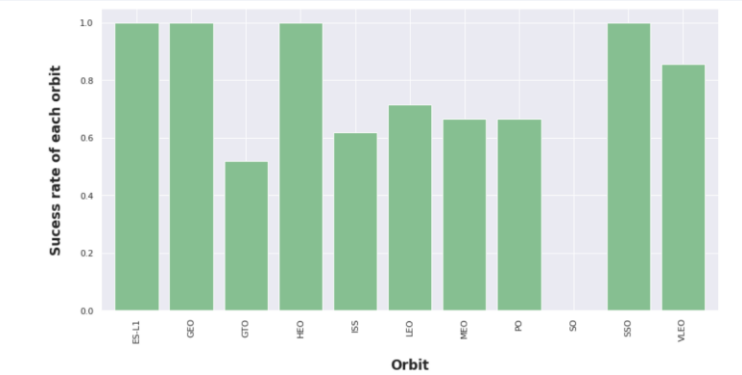
https://github.com/SAJULALSL/IBM-Capstone-Project/blob/master/EDA%20with%20Data%20Visualization.ipynb

As part of EDA, various data visualization tools are being used to generate insights from the launch data. Further, Feature Engineering being done using Pandas and Matplotlib to obtain some preliminary insights about how each important variable would affect the success rate.

1. Visualize the relationship between Flight Number and Launch Site
2. Visualize the relationship between Payload and Launch Site
3. Visualize the relationship between success rate of each orbit type
4. Visualize the relationship between Flight Number and Orbit type
5. Visualize the relationship between Payload and Orbit type
6. Visualize the launch success yearly trend

A few examples are as follows





EDA with SQL

Source Code

https://github.com/SAJULALSL/IBM-Capstone-Project/blob/master/EDA%20with%20SQL.ipynb

As part of EDA, load the SpaceX dataset into the corresponding table in a Db2 database and execute SQL queries understand the data set in detail.

1. Names of the unique launch sites in the space mission
2. 5 records where launch sites begin with the string 'CCA'
3. The total payload mass carried by boosters launched by NASA (CRS)
4. Average payload mass carried by booster version F9 v1.1
5. Date when the first successful landing outcome in ground pad was achieved.
6. Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
7. Total number of successful and failure mission outcomes
8. Names of the booster versions which have carried the maximum payload mass using a sub query
9. Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
10. 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

Build an Interactive Map with Folium

Source Code

https://github.com/SAJULALSL/IBM-Capstone-Project/blob/master/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

- The launch success rate may depend on many factors such as payload mass, orbit type, and so on. It may also depend on the location and proximities of a launch site, i.e., the initial position of rocket trajectories. Finding an optimal location for building a launch site certainly involves many factors and hopefully we could discover some of the factors by analyzing the existing launch site locations.
- We used Folium to add Markers, Circles and Lines to an interactive map showing the launch sites of the Falcon9 rocket
- Markers indicate points (launch sites) and the frequency of launches at each site (green=successful recovery of Stage One for this launch, red=failure of recovery)
- Circles were used to highlight areas around specific coordinates, like NASA Johnson Space Center
- Marker clusters indicates groups of events in each coordinate, like launches in a launch site
- Lines were used to indicate distances (closest proximity to coastline, city, railway, etc)
    1. Mark all launch sites on a map
    2. Mark the success/failed launches for each site on the map
    3. Calculate the distances between a launch site to its proximities

Build a Dashboard with Plotly Dash

Source Code

https://github.com/SAJULALSL/IBM-Capstone-Project/blob/master/Interactive%20Dashboard%20with%20Ploty%20Dash.ipynb

An interactive dashboard was built using Plotly dash that contains two types of graphs:

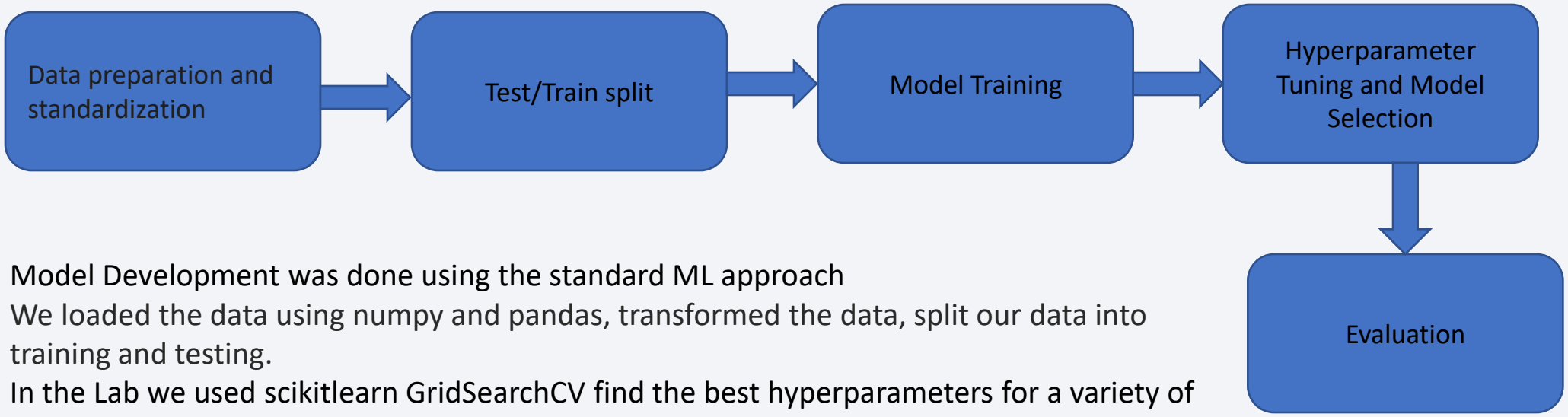
A **pie chart** showing the total number of launches by a site (or all of them combined). For a selected launch site, the pie chart showed a breakdown of launches (successful/failures)

A **scatter plot** of the outcome (successful recovery of Stage One Booster or not) versus the Payload Mass (Kg). The bound of the mass can be interactively changed by the user.

Predictive Analysis (Classification)

Source Code

https://github.com/SAJULALSL/IBM-Capstone-Project/blob/master/Machine%20Learning%20Prediction.ipynb



- Model Development was done using the standard ML approach
- We loaded the data using numpy and pandas, transformed the data, split our data into training and testing.
- In the Lab we used scikitlearn GridSearchCV find the best hyperparameters for a variety of models (Logistic Regression, SVM, Decision Tree, KNN). For every model the confusion matrix was created, and we compared the models using a simple score (accuracy).
- We used accuracy as the metric for our model, improved the model using feature engineering and algorithm tuning.
- Identified the best performing classification model.

Results

Exploratory data analysis results:

- Space X uses 4 different launch sites - CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E
- The first launches were done to Space X itself and NASA;
- The average payload of F9 v1.1 booster is 2,928 kg;
- The first success landing outcome happened in 2015 fiver year after the first launch;
- Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average;
- Almost 100% of mission outcomes were successful;
- Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015;
- The number of landing outcomes became as better as years passed. Obviously SpaceX learned how to do it better in the course of time.

Interactive analytics demo in screenshots:

- KSC LC-39A has the highest success rate for all sites
- KSC LC-39A has the highest percentage of successful recoveries of Stage One Booster
- Stage One Boosters are being recovered for a total of 76.9% of the cases
- Success rate for low to mid range payloads is higher compared to the high ones

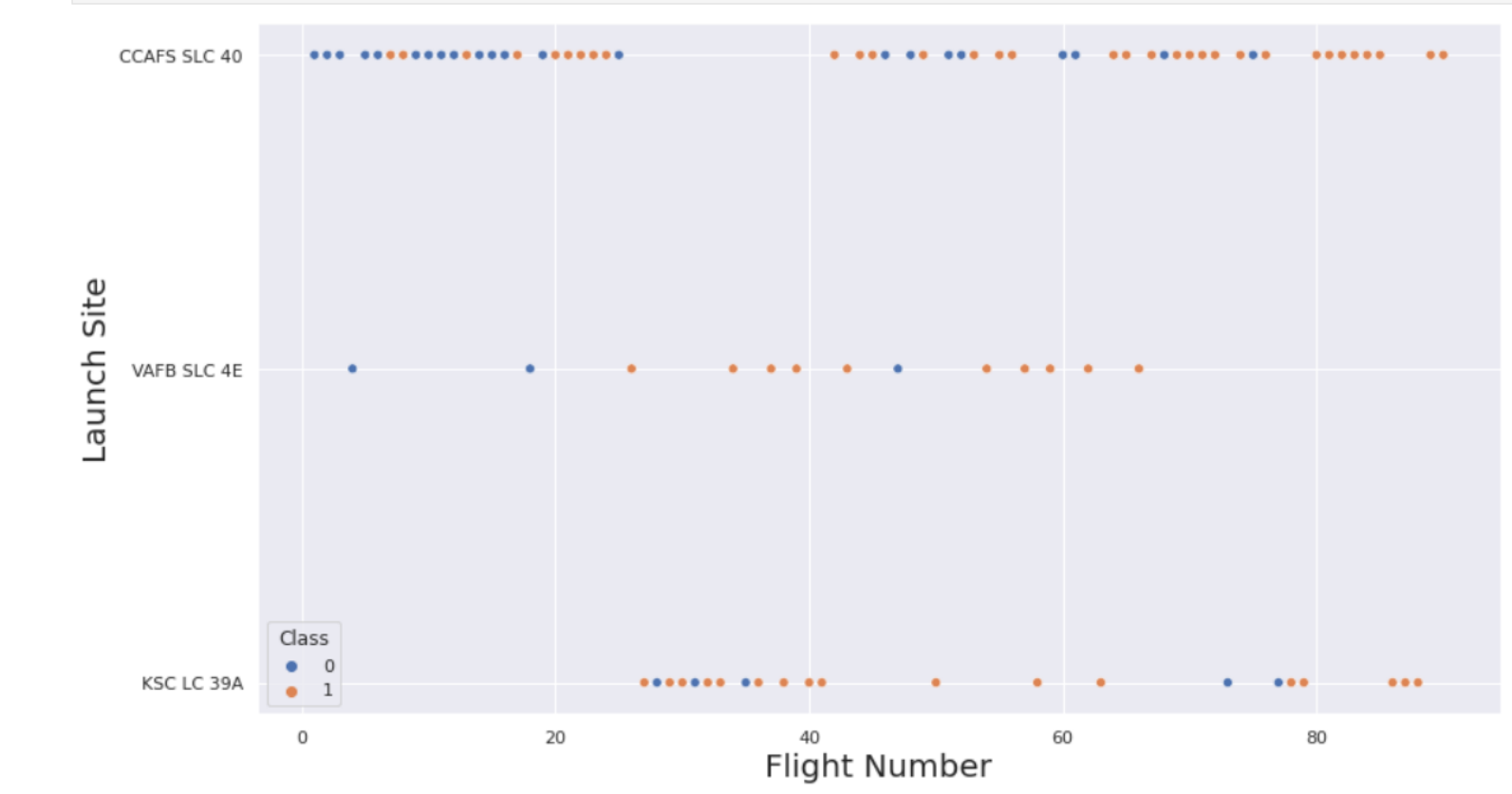
Predictive analysis results:

- Decision Tree Classifier can be used to predict successful landings and increase profits
- The Decision tree classifier is the best machine learning algorithm for this task with an Accuracy of 90%.

Section 2

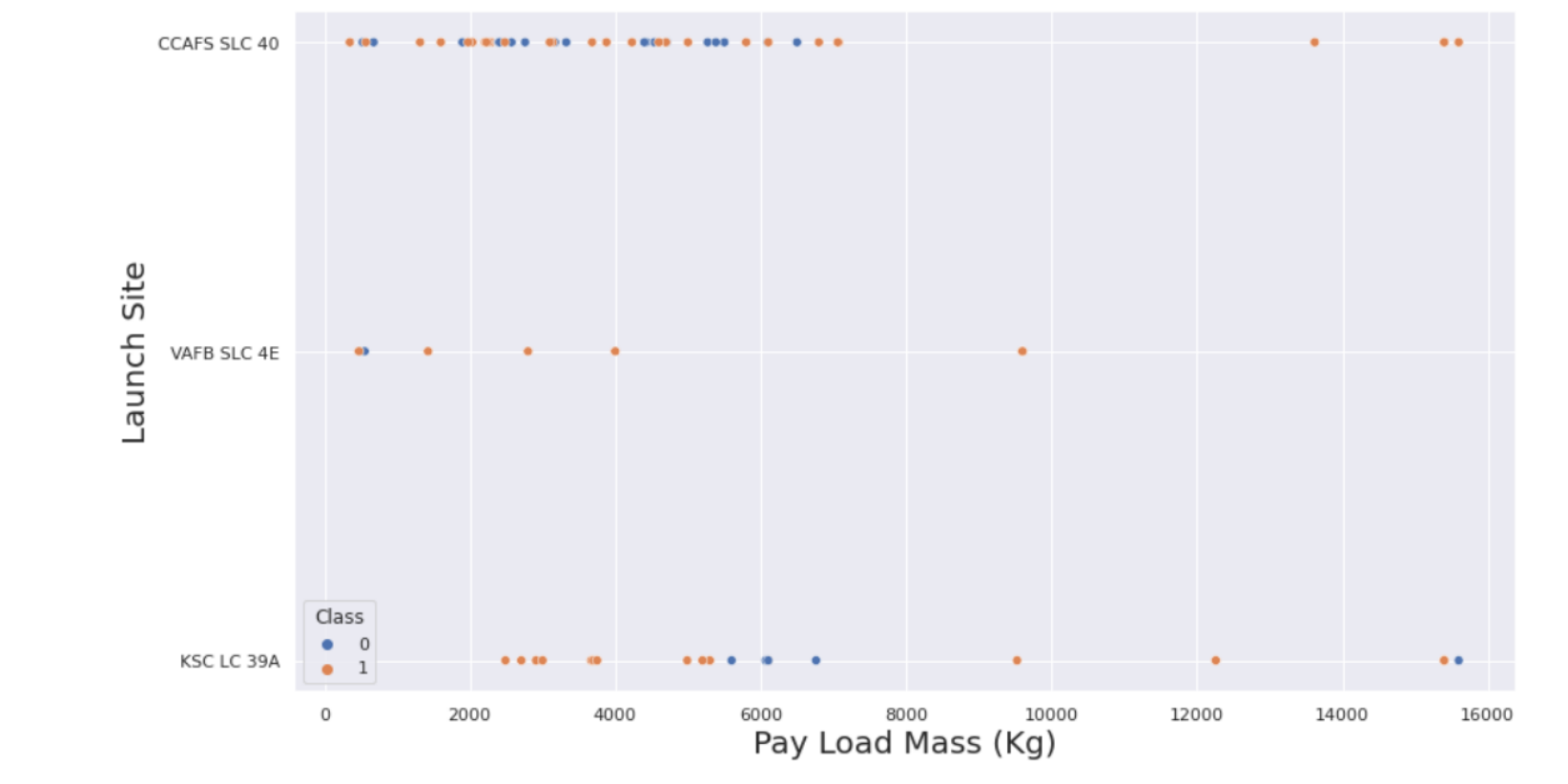
Insights drawn from EDA

Flight Number vs. Launch Site



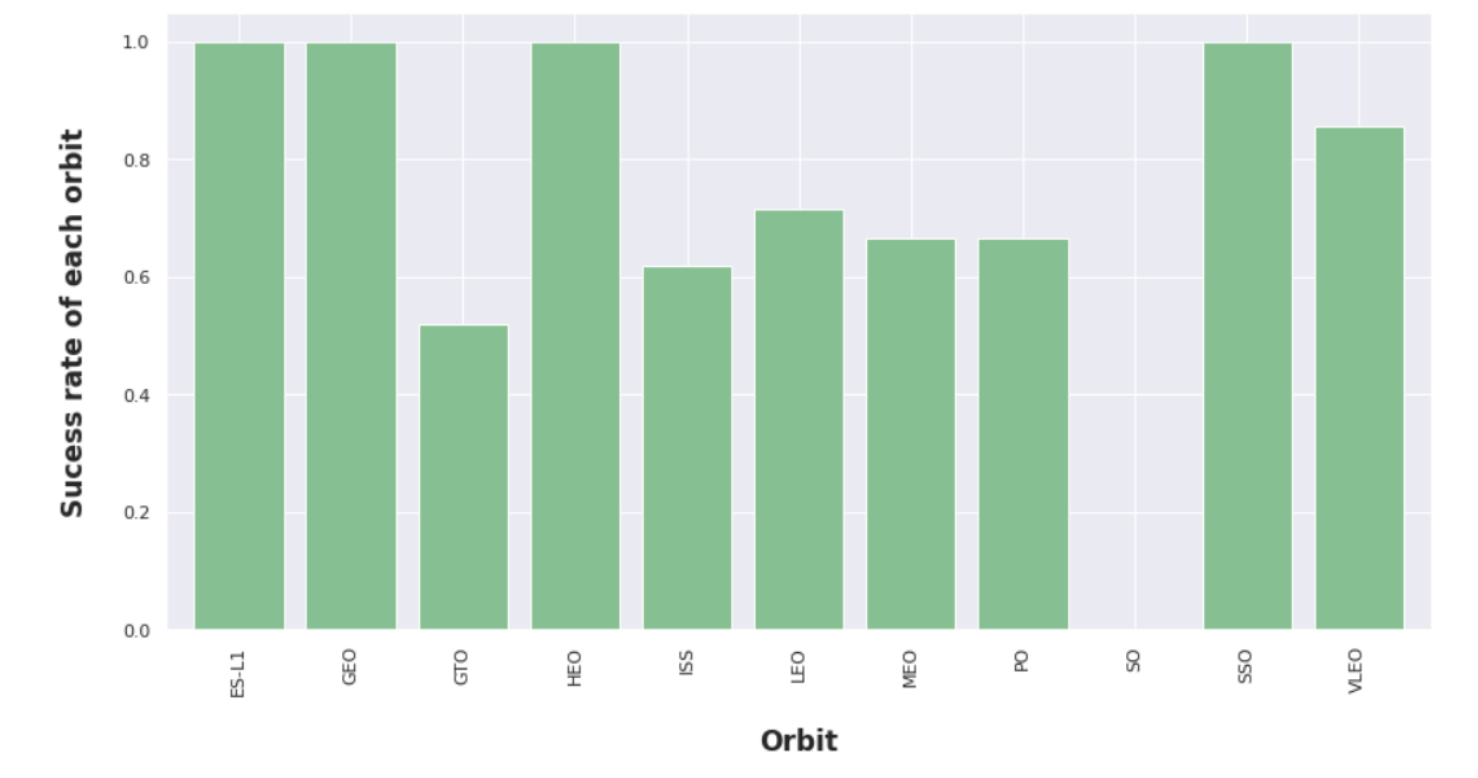
- The launch site CCAFS SLC 40 has the highest success. The launch sites VAFB SLC 4E and KSC LC 39A comes in the second and third places respectively in terms of success.
- More and more successful recoveries (orange dots) in the course of time (higher flight number) indicating that that at every site the situation gets better, chances of success are increasing

Payload vs. Launch Site



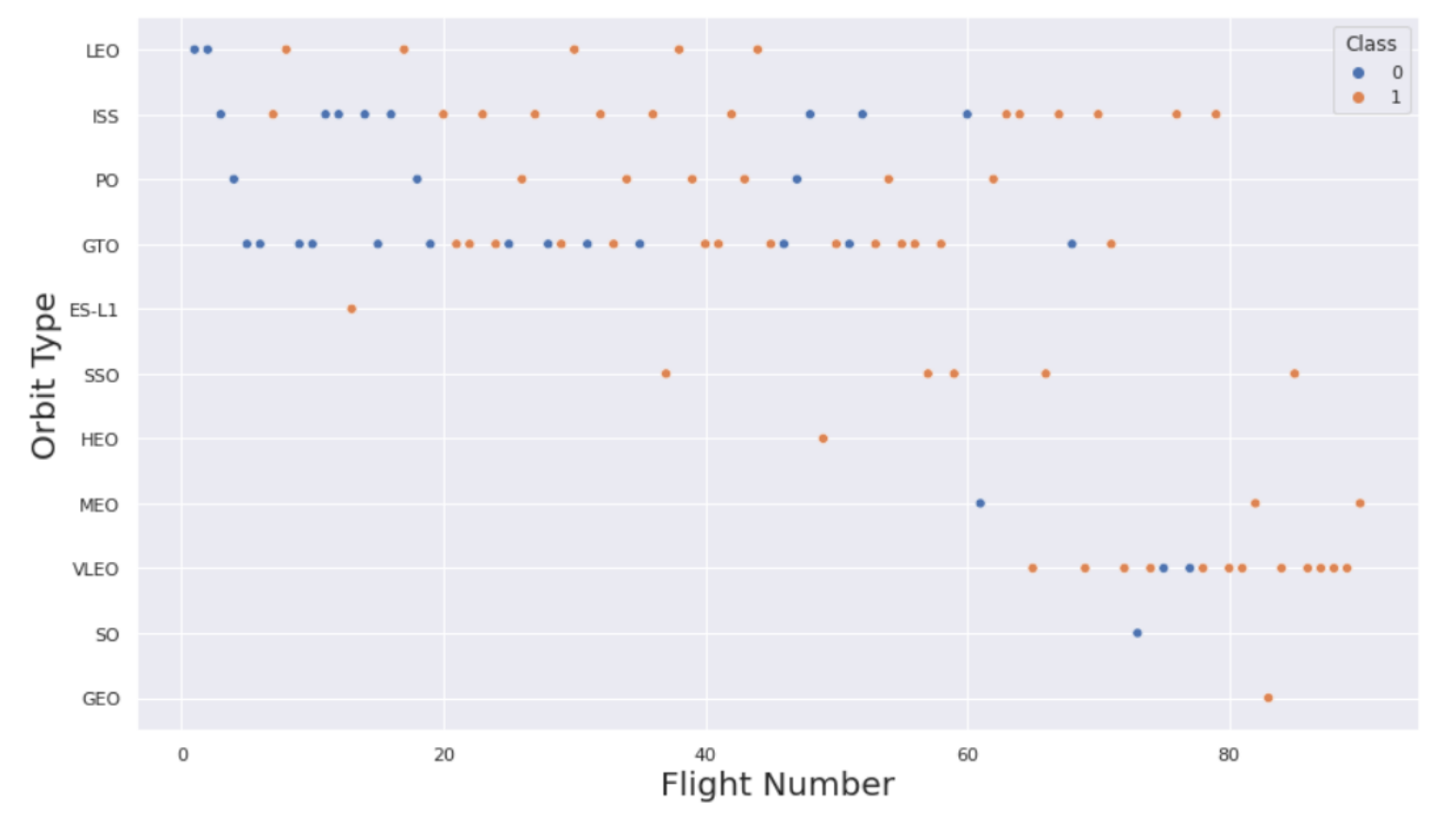
- The launch site CCAFS SLC 40 has the highest success across payload especially the low to med range
- The launch site CCAFS SLC 40 and KSC LC 39A has successes for high payloads (Payloads over 12,000kg) as well whereas VAFB SLC 4E doesn't have any success for the higher payloads.

Success Rate vs. Orbit Type



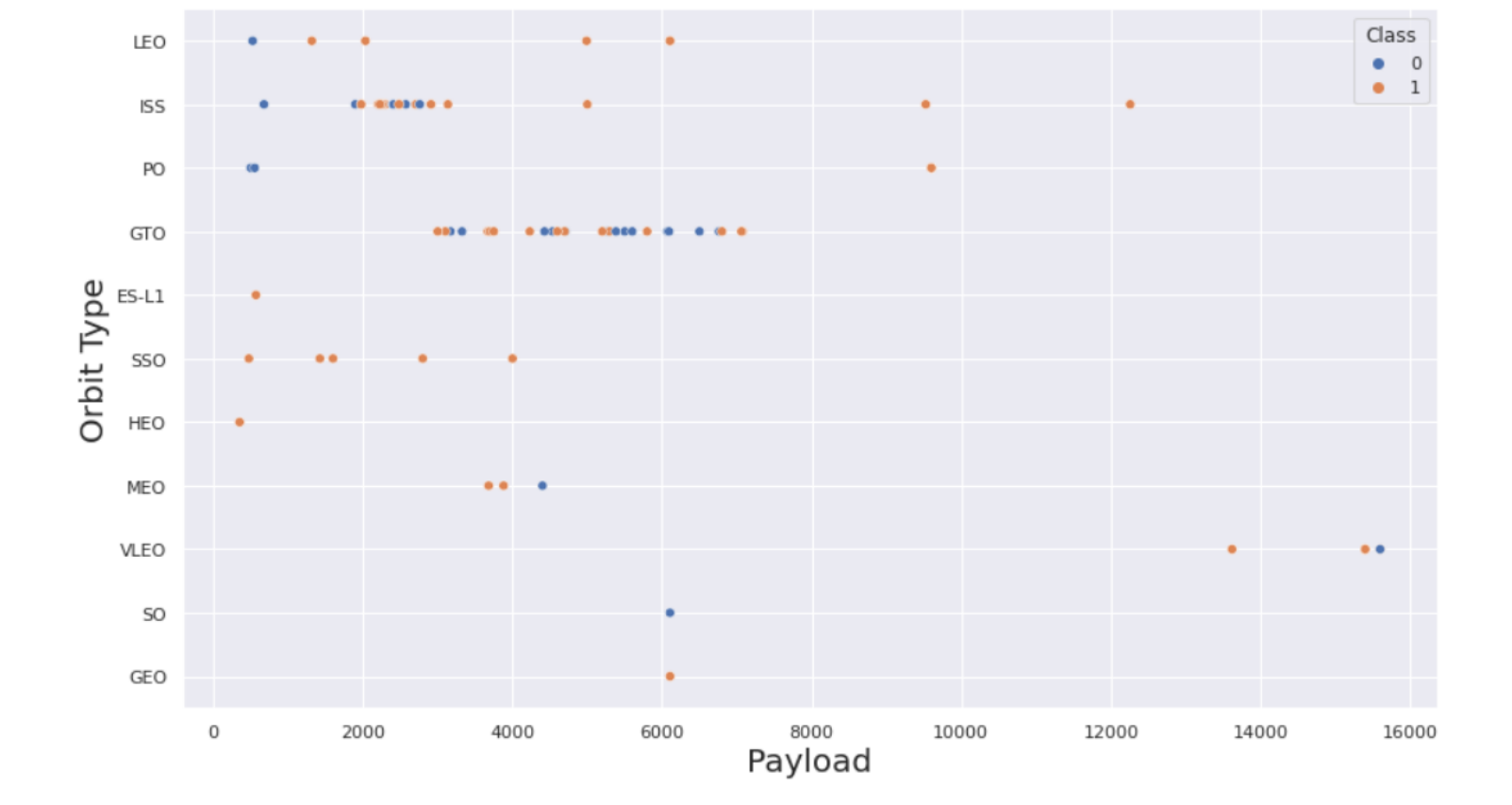
- From the plot, we can see that the orbit types ES-L1, GEO, HEO, SSO, had almost perfect success rate.
- SO, doesn't have any success

Flight Number vs. Orbit Type



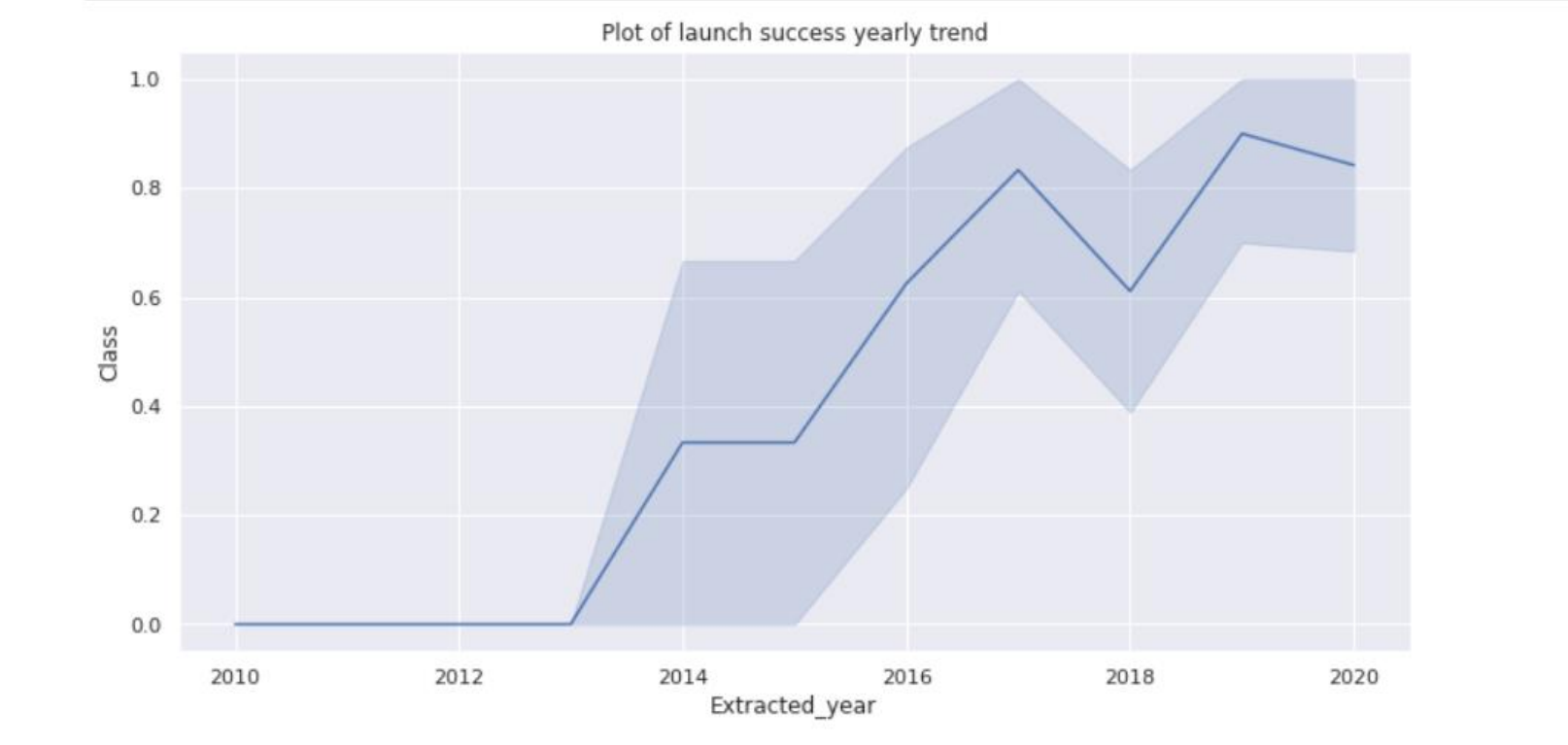
- Success rate improved over time to all orbits;
- In the VLEO orbit, seems a new business opportunity, due to recent increase of its frequency
- In the LEO orbit, success is related to the number of flights whereas in the GTO orbit, there is no relationship between flight number and the orbit.
- Not every orbit was in the program in the past.
- The GEO orbit is possbly an outlier.

Payload vs. Orbit Type



- There is no correlation between the Payload and the Orbit Type from a success rate perspective
- GTO has a narrow, continuous range of payloads, ISS Orbit has a a wide range of payloads. There are few launches to the orbits SO and GEO
- With heavy payloads, the successful landing are more for PO, LEO and ISS orbits.

Launch Success Yearly Trend



- It seems that the first three years were a period of adjustments and improvement of technology.
- The success rate is increasing over the years from 2013 till 2020
- SpaceX has achieved almost 80% success rate in 2020.

All Launch Site Names

Display the names of the unique launch sites in the space mission

```
%sql SELECT DISTINCT LAUNCH_SITE as "Launch_Sites" FROM SPACEXTBL;
```

```
 * ibm_db_sa://jmj86736:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32304/bludb
Done.
```

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

- There are 4 unique launch sites. ‘DISTINCT’ removes duplicates from the dataset.

Launch Site Names Begin with 'CCA'

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

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

```
 * ibm_db_sa://jmj86736:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32304/bludb
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	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

- This query displays 5 records where launch sites begin with `CCA`

Total Payload Mass

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

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

```
 * ibm_db_sa://jmj86736:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32304/bludb
Done.
```

1
45596

- Total payload mass carried by boosters launched by NASA (45596) is calculated by summing all payloads whose codes contain ‘CRS’, which corresponds to NASA.

Average Payload Mass by F9 v1.1

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

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where BOOSTER_VERSION = 'F9 v1.1'
```

```
 * ibm_db_sa://jmj86736:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32304/bludb
Done.
```

1
2928

- The average payload mass carried by booster version F9 v1.1 as 2928 was calculated by filtering data by the given booster version and calculating the average payload mass.

First Successful Ground Landing Date

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

Hint:Use min function

```
%sql select min(DATE) from SPACEXTBL where Landing__Outcome = 'Success (ground pad)'
```

```
 * ibm_db_sa://jmj86736:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32304/bludb
Done.
```

1
2015-12-22

- The date of the first successful landing outcome on ground pad was 22nd December 2015. It was calculated with the 'min' function.

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

```
%sql select BOOSTER_VERSION from SPACEXTBL where Landing__Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000
```

```
 * ibm_db_sa://jmj86736:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32304/bludb
Done.
```

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

- 'WHERE' clause was used to filter for boosters which have successfully landed on drone ship and applied the 'AND' condition to determine successful landing with payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
%sql select count(MISSION_OUTCOME) from SPACEXTBL where MISSION_OUTCOME = 'Success' or MISSION_OUTCOME = 'Failure (in flight)'
```

```
 * ibm_db_sa://jmj86736:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32304/bludb
Done.
```

1
100

- The wildcard like '%' was used to filter for **WHERE** MISSION_OUTCOME was a success or a failure.

Boosters Carried Maximum Payload

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)
```

```
 * ibm_db_sa://jmj86736:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32304/bludb
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

- The booster that have carried the maximum payload was arrived at using a subquery in the **WHERE** clause and the **MAX()** function.

2015 Launch Records

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE DATE LIKE '2015-%' AND \
LANDING__OUTCOME = 'Failure (drone ship)';
```

```
 * ibm_db_sa://jmj86736:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32304/bludb
Done.
```

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

- Combinations of the **WHERE** clause, **LIKE**, **AND**, and **BETWEEN** conditions were used to filter for failed landing outcomes in drone ship, their booster versions, and launch site names for year 2015
- There are 2 such occurrences

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

```
%sql select count(LANDING__OUTCOME),LANDING__OUTCOME from SPACEXTBL where DATE between '2010-06-04' and '2017-03-20' group by LANDING__OUTCOME order b
```

```
 * ibm_db_sa://jmj86736:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32304/bludb
Done.
```

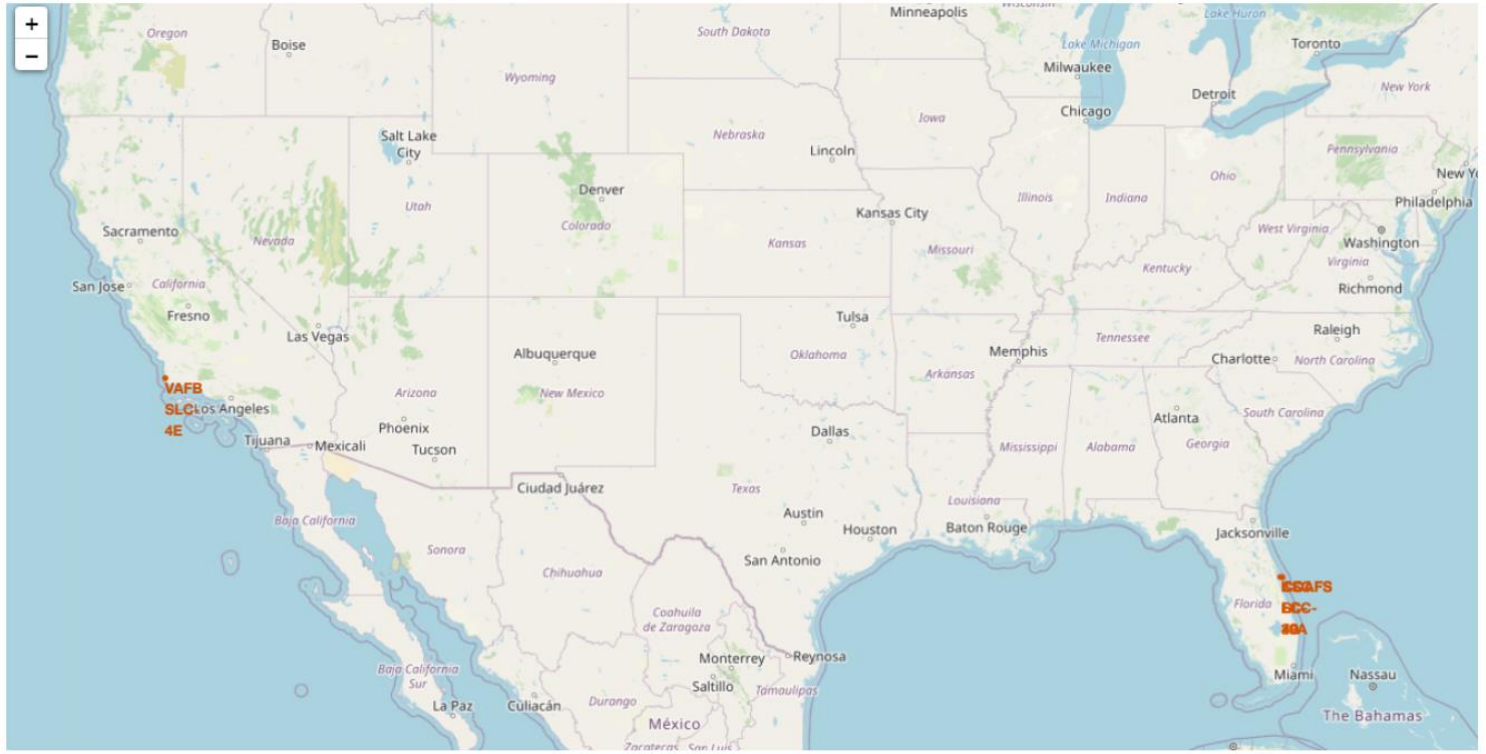
1	landing__outcome
10	No attempt
5	Failure (drone ship)
5	Success (drone ship)
3	Controlled (ocean)
3	Success (ground pad)
2	Failure (parachute)
2	Uncontrolled (ocean)
1	Precluded (drone ship)

- Landing outcomes and the **COUNT** of landing outcomes were selected from the data and used the **WHERE** clause to filter for landing outcomes **BETWEEN** 2010-06-04 to 2010-03-20.
- The **GROUP BY** clause was used to group the landing outcomes and the **ORDER BY** clause to order the grouped landing outcome in descending order.

Section 3

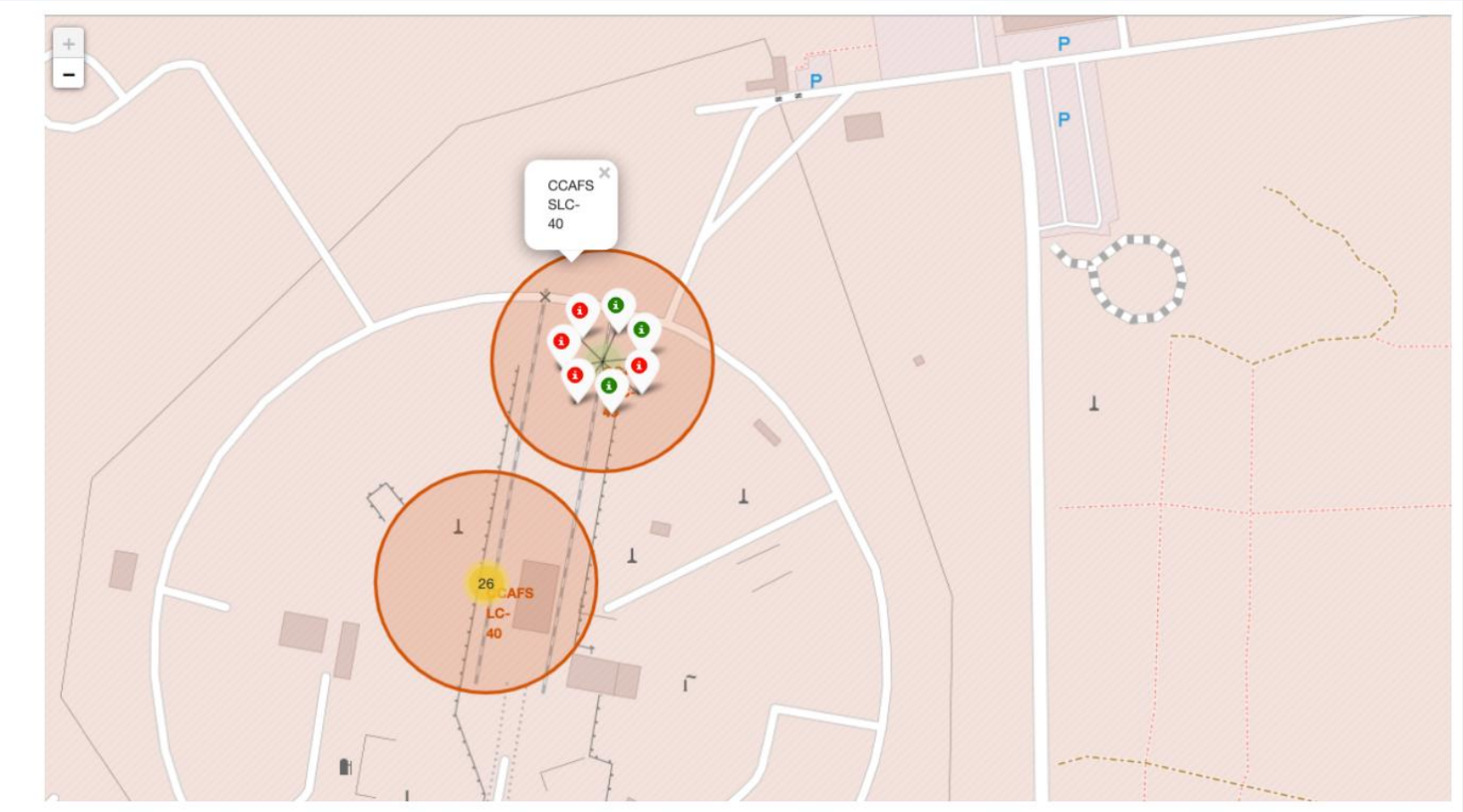
Launch Sites Proximities Analysis

Launch sites of SpaceX on the map of USA



- Launch sites are located across the eastern and western coasts of USA. This may be done from a safety perspective

Markers showing launch outcomes at launch sites

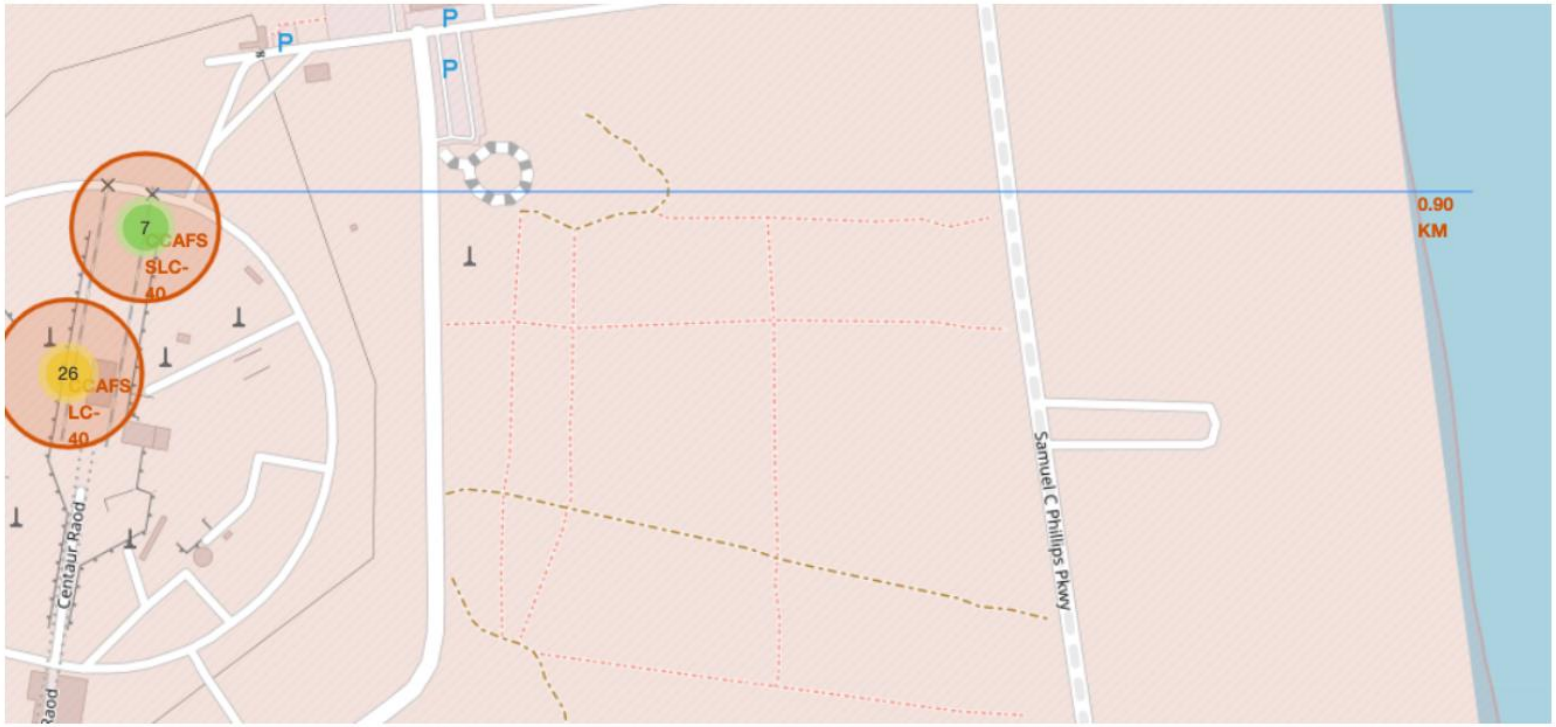


Each marker shows a launch.

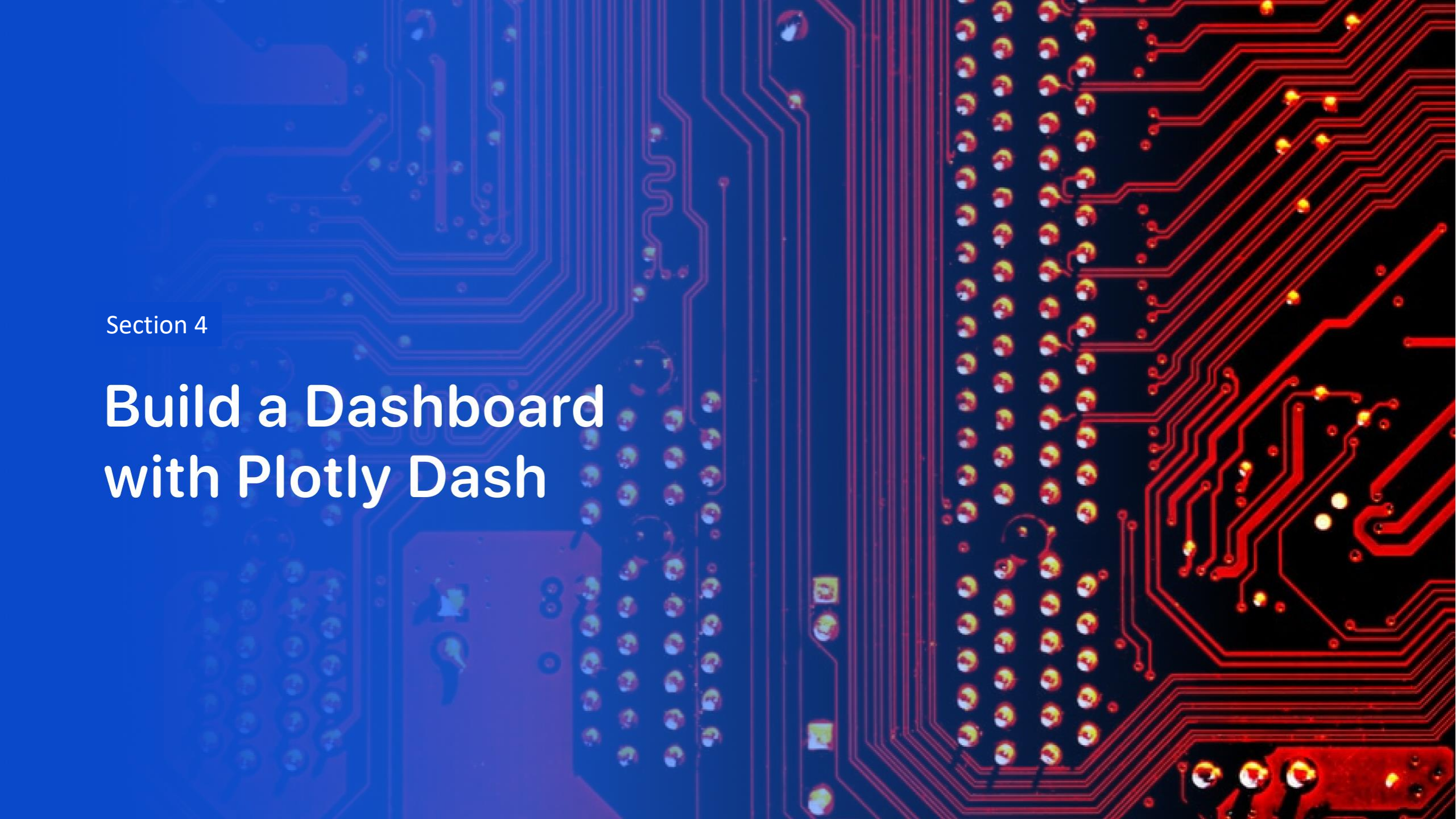
The color coding for the markers is -
Green = Success, Booster recovered,
Red = Failure, Booster not recovered

- At, CCAFS SLC - 40. there are 3 boosters got recovered where as 4 were not recovered successfully

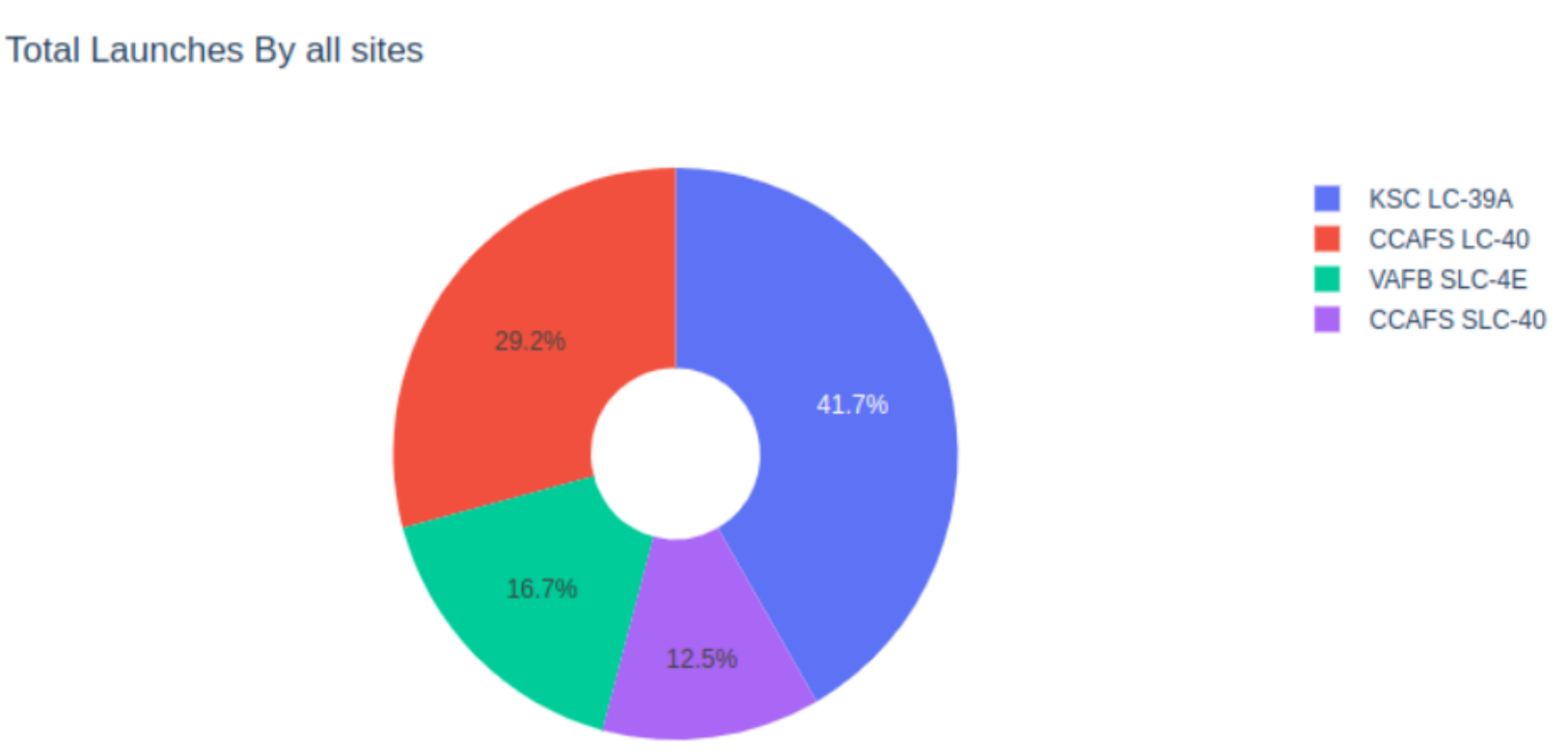
Proximity of the launch sites to landmarks



- This map shows the proximity of the launch sites to the various land marks
- CCAFS SLC – 40 is 0.9 KM from the coast line

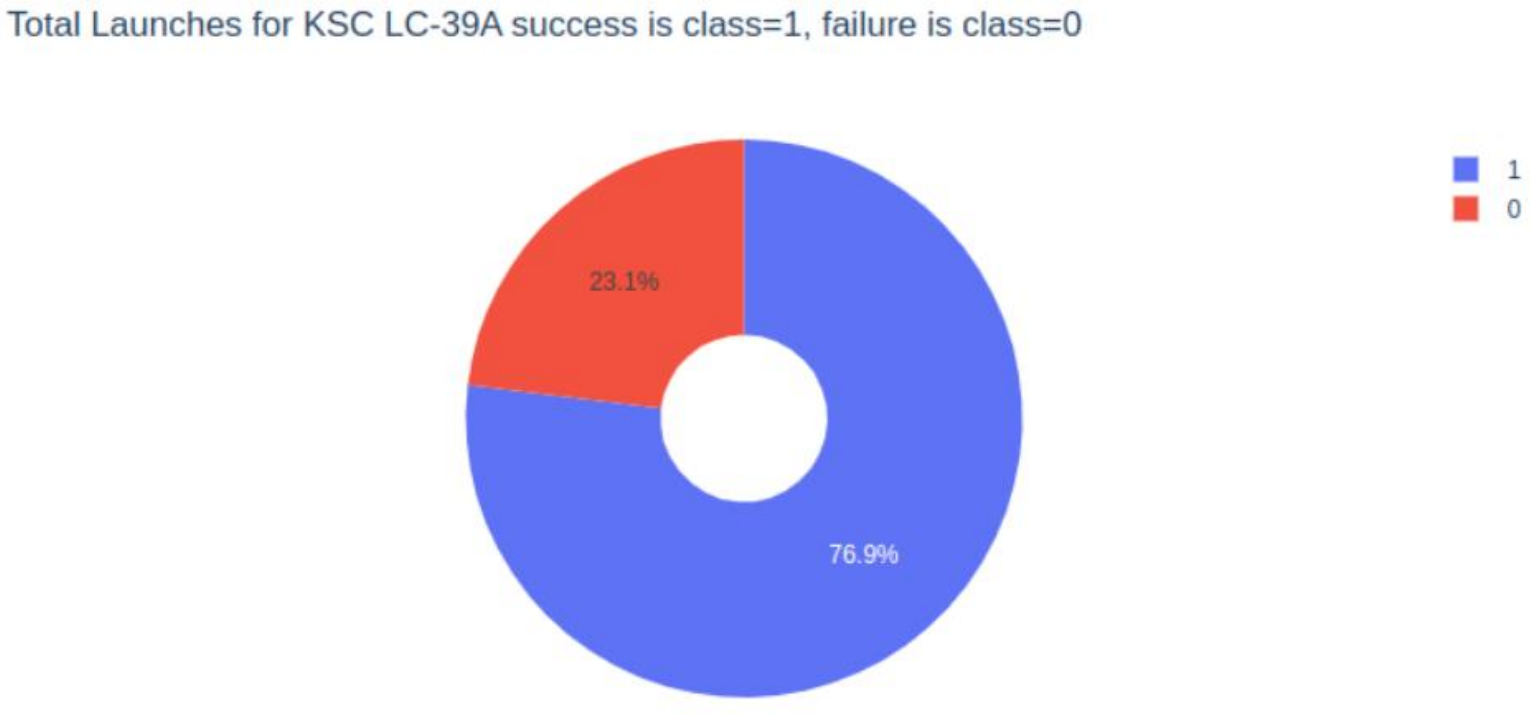

Section 4
Build a Dashboard with Plotly Dash

Pie Chart - Success percentage for each launch site



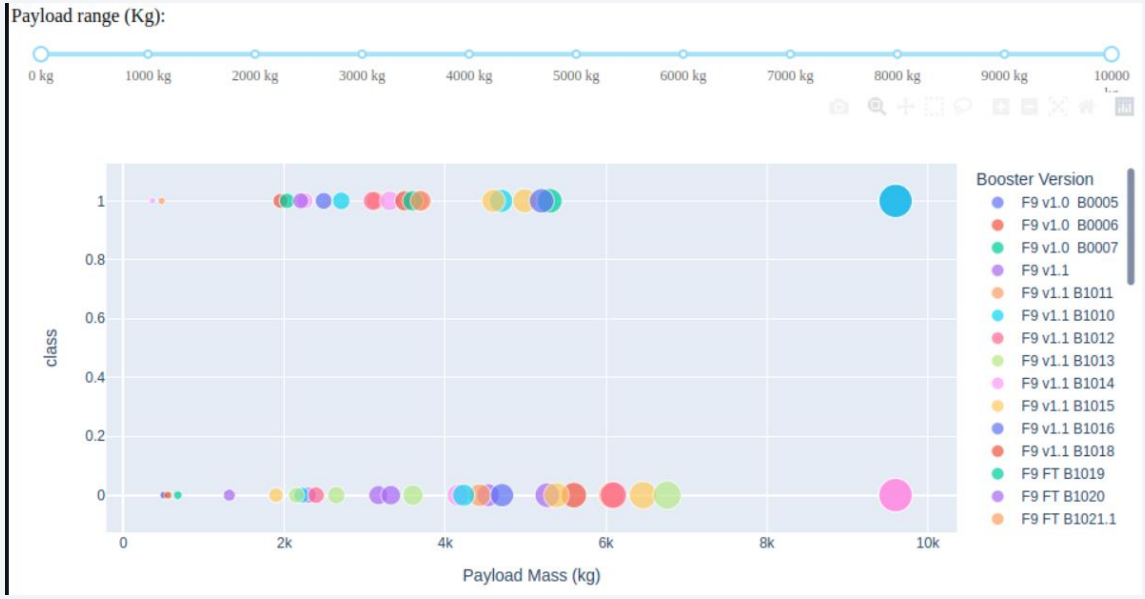
- KSC LC-39A has the highest success rate for all sites

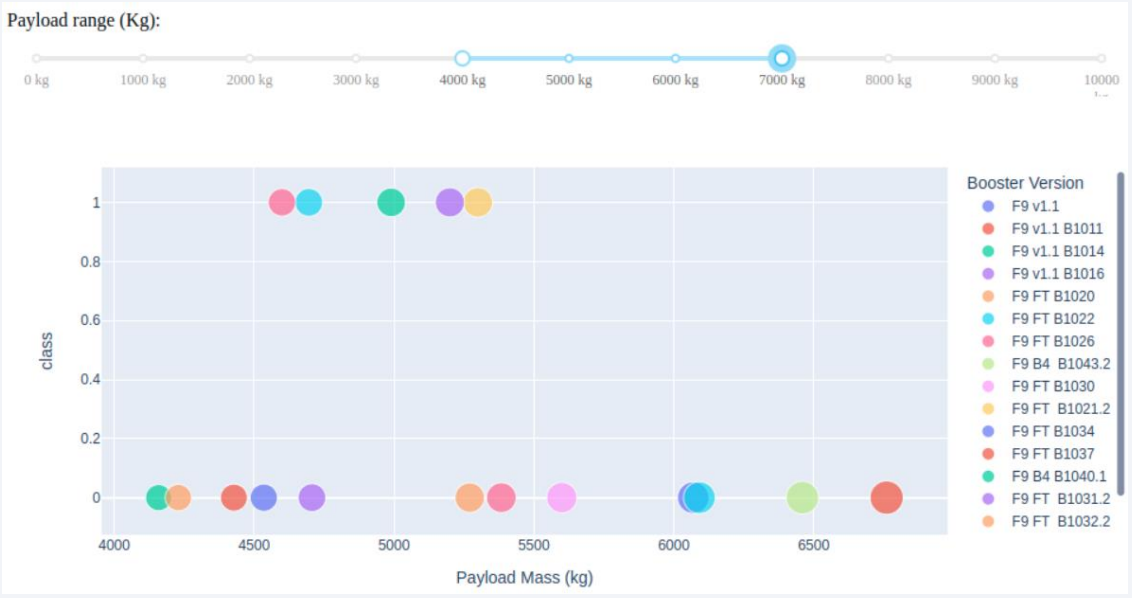
Pie Chart – Total launches at KSC LC-39A



- KSC LC-39A has the highest percentage of successful recoveries of Stage One Booster
- Stage One Boosters are being recovered for a total of 76.9% of the cases

Scatter plot - Payload vs Launch Outcome for all sites



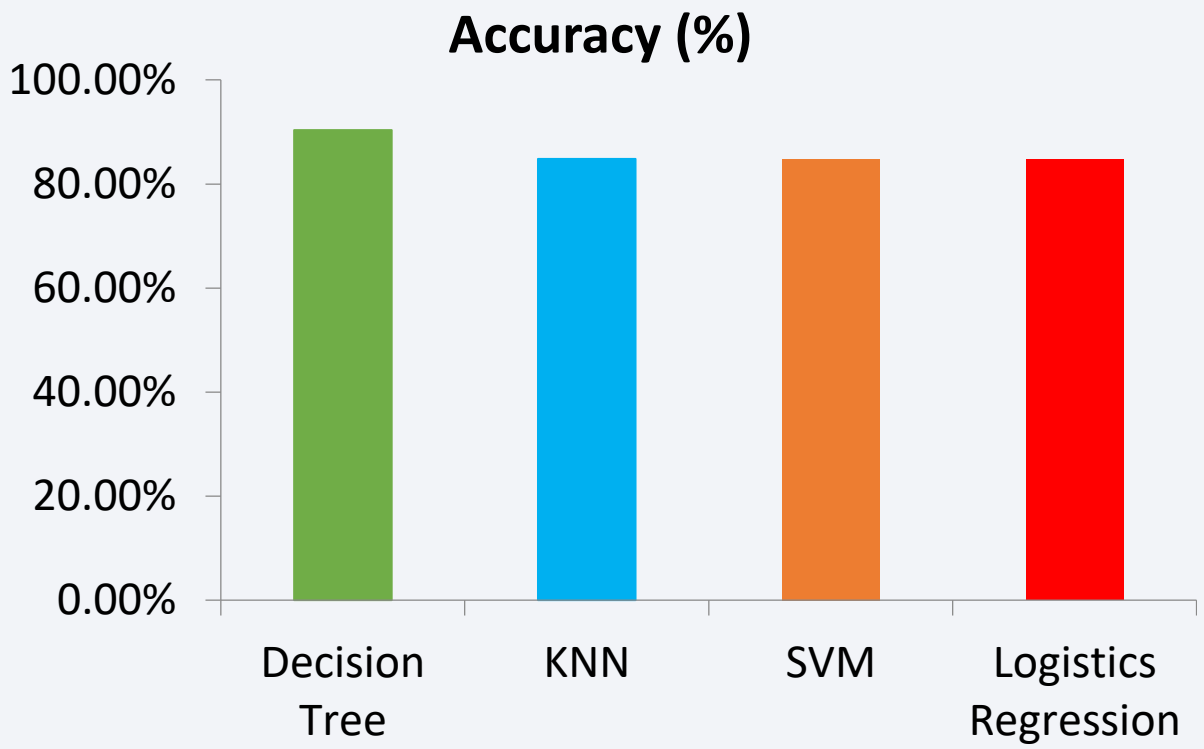


- Success rate for low to mid range payloads is higher compared to the high ones

Section 5

Predictive Analysis (Classification)

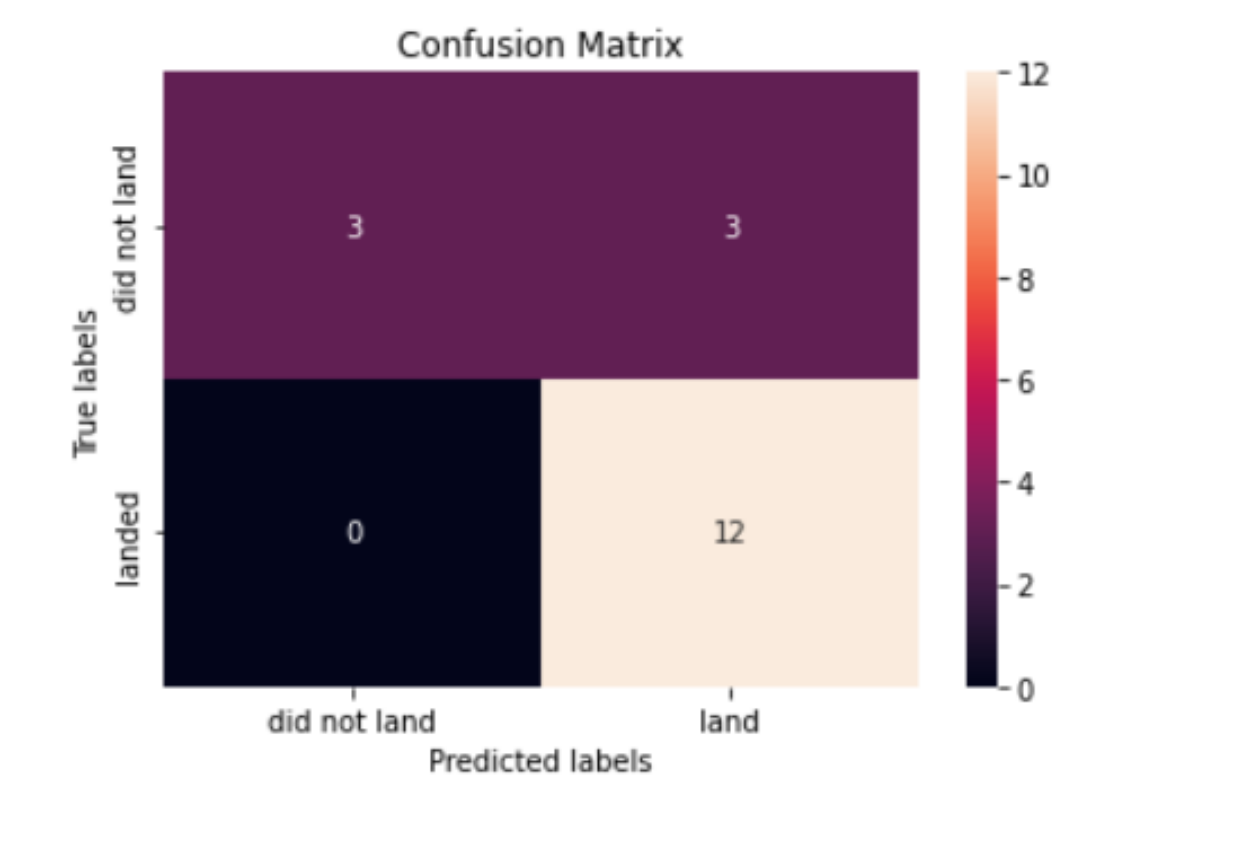
Classification Accuracy



	Accuracy
KNN	0.848214
Decision Tree	0.903571
Logistic Regression	0.846429
SVM	0.848214

- The Decision tree classifier is the best machine learning algorithm for this task with an Accuracy of 90%.

Confusion Matrix



- Decision Tree was the best performing model
- The confusion matrix for the decision tree classifier shows that the classifier can distinguish between the different classes. High numbers of true positive and true negative compared to the false ones

Conclusions

- SpaceX had already achieved a success rate for Booster recovery of over 80% in 2017. our ML models and predictions (83% accuracy) are not impressive
- The higher the number of flights the greater is the success rate may be because of gaining experience
- KSC LC-39A is the launch site with highest success rate.
- Launch success rate started to increase in 2013 till 2020. Although most of mission outcomes are successful, successful landing outcomes seem to improve over time, according the evolution of processes and rockets;
- Launches above 7,000kg are less risky;
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- Decision Tree Classifier can be used to predict successful landings and increase profits
- The Decision tree classifier is the best machine learning algorithm for this task with an Accuracy of 90%.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!