IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Nana Antwi Frimpong

19/06/2023

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

- Project background and context
- In the commercial space age, SpaceX is the most successful company. They are making space travel affordable. They advertise one of their rockets Falcon 9 which is at the cost of 62 million dollars while other providers offer their rockets at an upward cost of 165 million dollars each. SpaceX can reuse the first stage of their rockets hence why the massive savings on their prices offered. If we can determine if the first stage will land, we will be able to determine the cost of a launch as well. Based on public information made available by this company and machine learning models, we will work on predicting if SpaceX will reuse the first stage.
- Questions to be answered

- How do variables such as launch site, payload mass, number of flights, and orbits affect the success of the first stage landing?

- Can or Does the rate of successful landings increase over the years?

- What is the best algorithm available that can be used for binary classification in the case we are studying?

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology:

- Using SpaceX Rest API
- Using Web Scrapping from Wikipedia

Perform data wrangling

- Filtering the data
- Dealing with missing values
- Using One Hot Encoding to prepare the data to a binary classification

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

In this project, a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry were used in the data collection process.

Both data collection methods were used to get complete information about the launches for a more detailed analysis.

- **Data Columns are obtained by using Wikipedia Web Scraping:**

Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

- **Data Columns obtained by using SpaceX REST API were:**

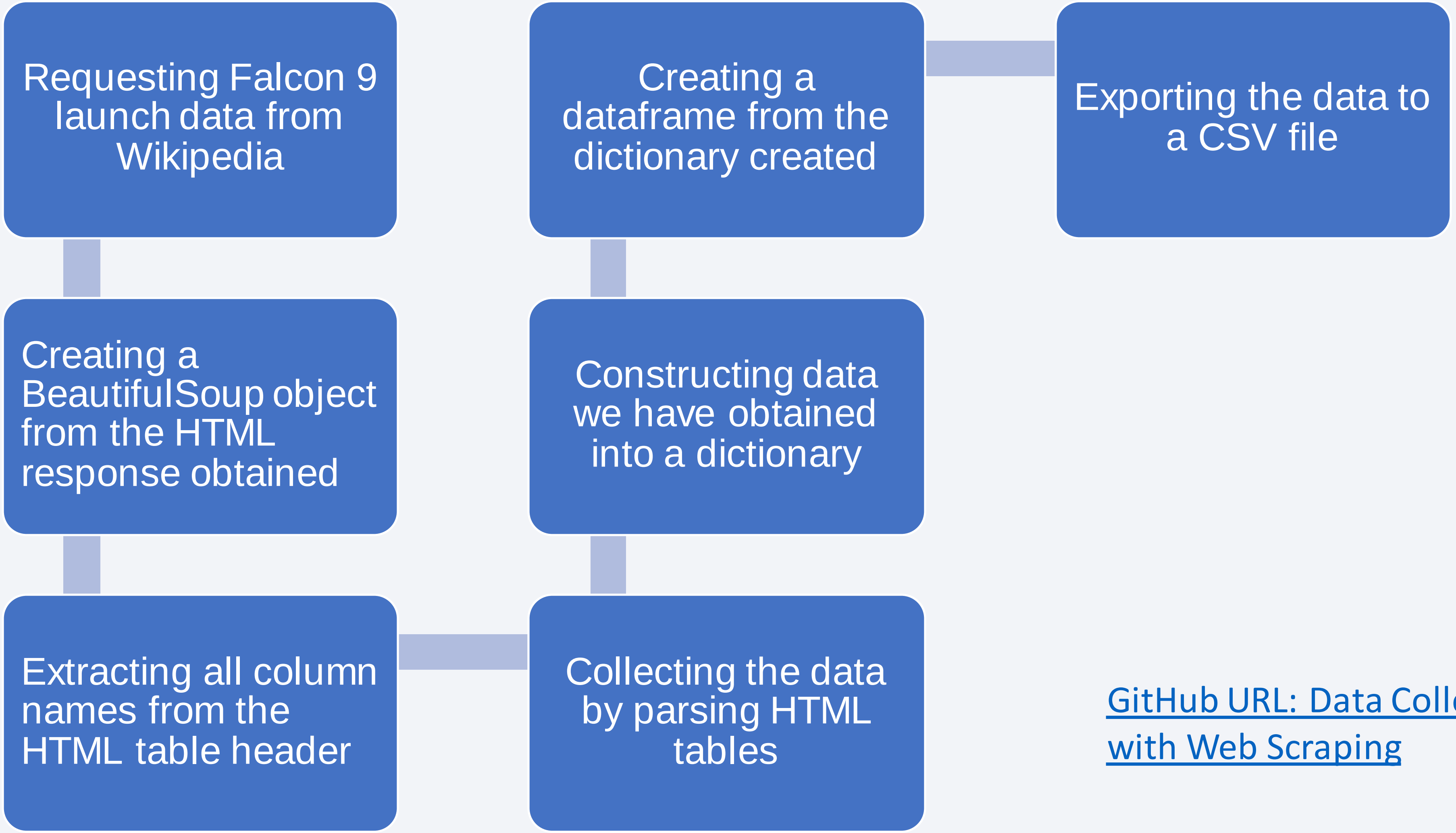
FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

Data Collection – SpaceX API

1. Requesting rocket launch data from the SpaceX API
2. Decoding the response content with .json() and transforming into a dataframe using .json_normalize()
3. Requesting needed information about the launches from SpaceX API using custom functions
4. Constructing data we have obtained into a dictionary
5. Creating a dataframe from the dictionary
6. Filtering the dataframe to only Falcon 9 launches
7. Replacing missing values of Payload Mass column by calculating .mean() for the column
8. Exporting the data to a CSV file.

GitHub URL: Data Collection API

Data Collection - Scraping



[GitHub URL: Data Collection with Web Scraping]

Data Wrangling

From the data set obtained, there are several different cases where the booster was not able to land successfully. At some given times, landing was attempted and was unsuccessful. From the examination, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. Also, True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad. While True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship.

For this project, a conversion of those outcomes into Training Labels with “1” which means the booster successfully landed and “0” means it was unsuccessful was made.

1. Perform exploratory Data Analysis and then determine what the Training Labels are
2. Calculate the total number of launches made on each site
3. Calculate the total number and occurrence of each orbit
4. Calculate the total number and occurrence of mission outcome per each orbit type
5. Create a landing outcome label from Outcome column
6. Exporting the data to a CSV file

[GitHub URL: Data Wrangling]

EDA with Data Visualization

Charts plotted:

- Flight Number vs. Payload Mass
- Flight Number vs. Launch Site
- Payload Mass vs. Launch Site
- Orbit Type vs. Success Rate
- Flight Number vs. Orbit Type
- Payload Mass vs Orbit Type
- Success Rate Yearly Trend.

A. Bar charts were used to show comparisons among discrete categories within the data sets.

B. This was to show the relationship between specific categories and measured values.

C. Scatter plots were also used show the relationship between variables. Here, If a relationship exists, they could be used in machine learning model.

D. Line charts were used to show trends in data over time (time series).

GitHub URL: EDA with Data Visualization

EDA with SQL

Performed SQL queries:

- To display the names of each launch site in the space mission
- To display 5 records where launch sites begin with the string ‘CCA' from the data set
- To list the total number of successful and failure mission outcomes
- To display the total payload mass carried by boosters launched by NASA (CRS)
- To display average payload mass carried by booster version F9 v1.1
- To list the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- To list the date when the first successful landing outcome in ground pad was achieved
- To list the total number of successful and failure mission outcomes
- To list the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- To rank and count the number 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 URL: EDA with SQL]

Build an Interactive Map with Folium

Add Markers to all Launch Sites:

- Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.

- Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.

Add Colored Markers as launch outcomes for each Launch Site:

- Added colored Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates. Distances between a Launch Site to its proximities:

- Added colored Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City.

[GitHub URL: Interactive Visual Analytics with Folium]()

Build a Dashboard with Plotly Dash

Launch Sites Dropdown List:

- A dropdown list is added to enable Launch Site selection.

Pie Chart showing Success Launches (All Sites/Certain Site):

- A pie chart is added to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.

Slider of Payload Mass Range:

- A slider is added to select Payload range.

Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:

- A scatter chart is added to show the correlation between Payload and Launch Success.

GitHub URL: SpaceX Dash App

Predictive Analysis (Classification)

Creating a NumPy array from the column "Class" in data

Standardizing the data with StandardScaler, then fitting and transforming it

Splitting the data into training and testing sets usine the train_test_split function

Creating a GridSearchCV object with cv = 10 to find the best parameters

Applying GridSearchCV on LogReg, SVM, Decision Tree, and KNN models

Calculating the accuracy on the test data with the method .score() for all models

Examining the confusion matrix for all models

Finding the method performs best by examining the Jaccard_score and F1_score metrics

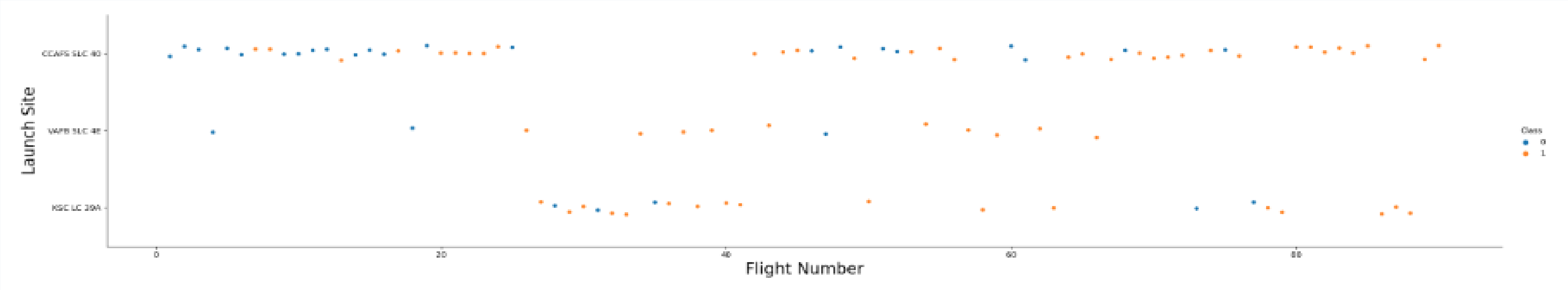
Results

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

Section 2

Insights drawn from EDA

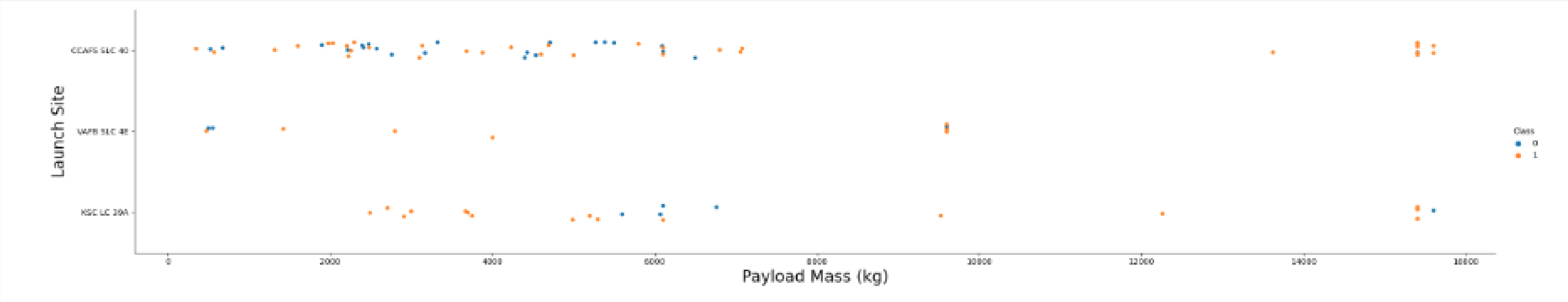
Flight Number vs. Launch Site



Here, we can see that:

- VAFB SLC 4E and KSC LC 39A have higher success rates
- The earliest flights all failed while the latest flights all succeeded.
- It can be assumed that each new launch has a higher rate of success.
- The CCAFS SLC 40 launch site has about a half of all launches.

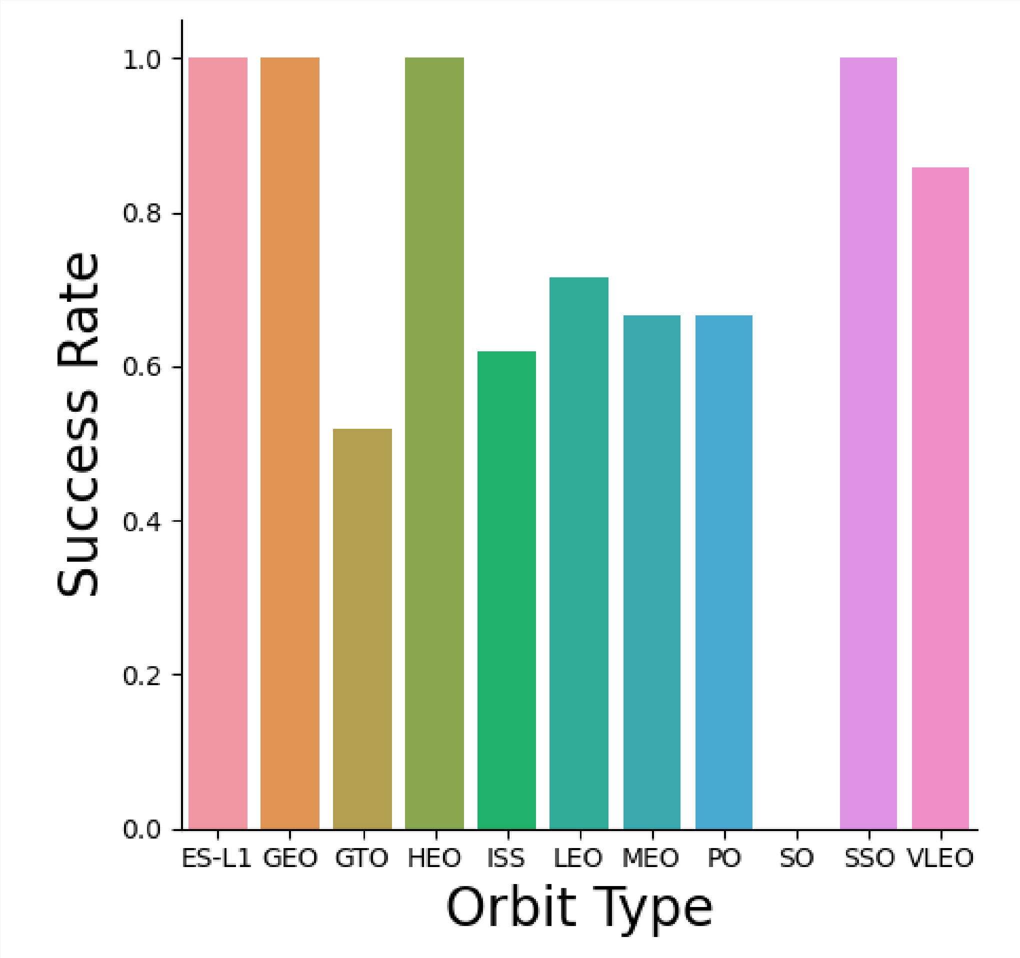
Payload vs. Launch Site



Here, we can see that:

- ➢ For every launch site the higher the payload mass, the higher the success
- ➢ Most of the launches with payload mass over 7000 kg were successful rate.
- ➢ KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

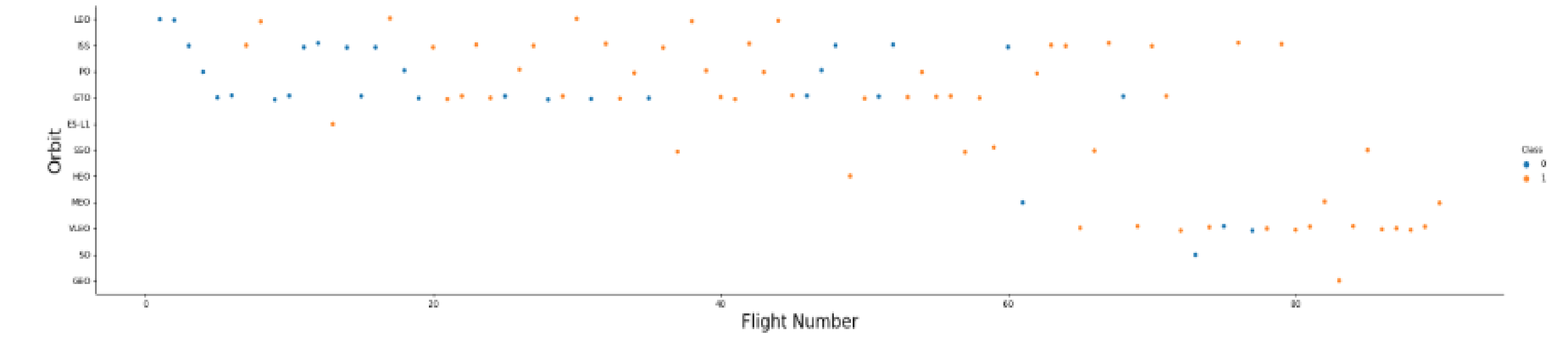
Success Rate vs. Orbit Type



Here, we can see that:

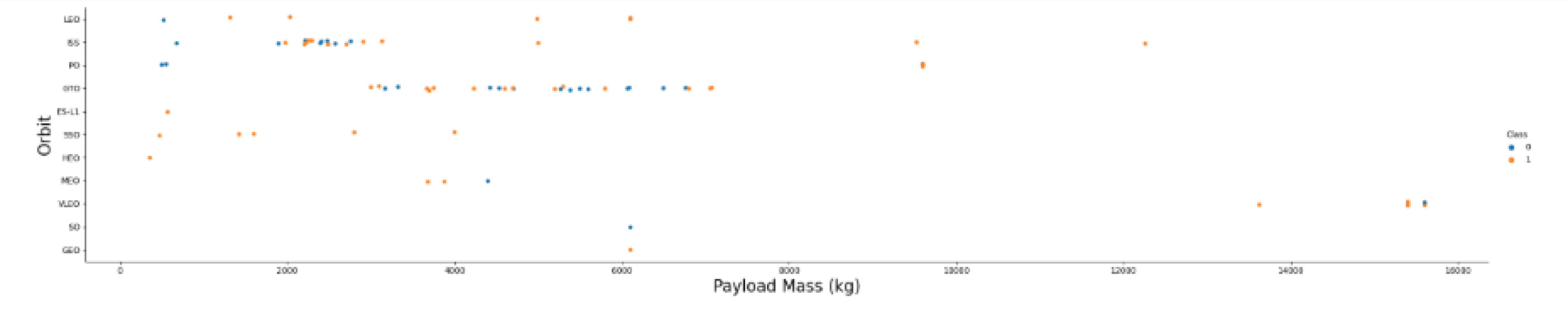
- Orbits with success rate between 50% and 85%:
  - GTO, ISS, LEO, MEO, PO
- Orbits with 100% success rate:
  - ES-L1, GEO, HEO, SSO
- Orbits with 0% success rate:
  - SO

Flight Number vs. Orbit Type



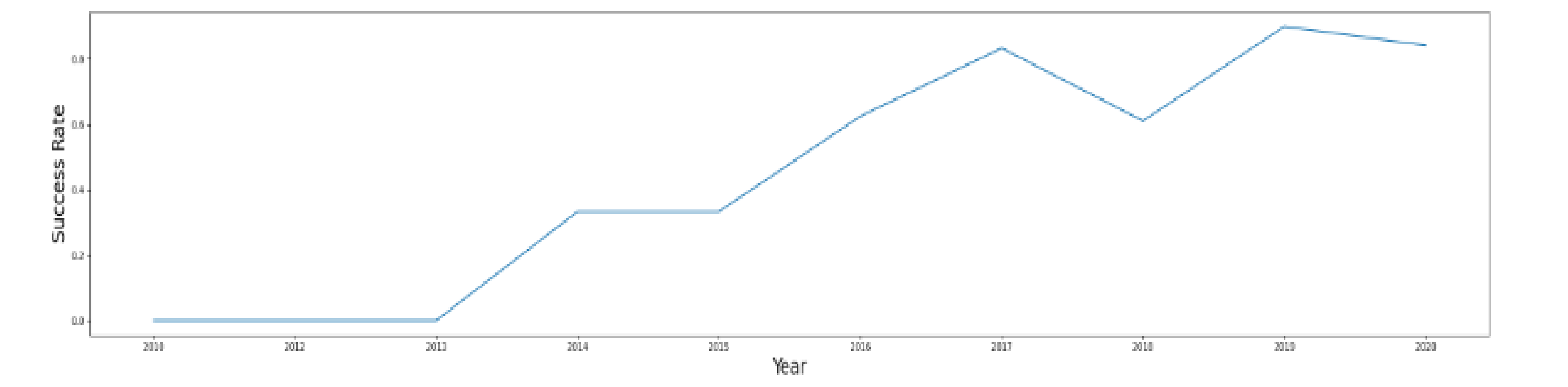
- ➢ In the LEO orbit, the Success is related to the number of flights.
- ➢ There seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



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

Launch Success Yearly Trend



➢ From 2013, the success rate kept increasing till 2020.

All Launch Site Names

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

```
In [54]: %sql select distinct launch_site from SPACEXTBL;

 * ibm_db_sa://pmg68780:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
   sqlite:///my_data1.db
Done.
```

Out[54]:

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

Launch Site Names Begin with 'CCA'

Displaying 5 records where launch sites begin with the string 'CCA'.

In [55]:
```
%sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5;
```

```
 * ibm_db_sa://pmg68780:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
   sqlite:///my_data1.db
Done.
```

Out[55]:

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing_outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	None	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	None	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	None	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	0:35:00	F9 v1.0 B0006	CCAFS LC-40	None	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	None	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

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

```
In [56]: %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXTBL where customer = 'NASA (CRS)';

 * ibm_db_sa://pmg68780:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
   sqlite:///my_data1.db
Done.
```

Out[56]:

total_payload_mass
45596

Average Payload Mass by F9 v1.1

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

```
In [57]: %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXTBL where booster_version like '%F9 v1.1%';

 * ibm_db_sa://pmg68780:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
   sqlite:///my_data1.db
Done.
```

Out[57]:

average_payload_mass
2534

First Successful Ground Landing Date

Listing the date when the first successful landing outcome in ground pad was achieved.

```
In [60]: %sql select min(date) as first_successful_landing from SPACEXTBL where landing_outcome = 'Success (ground pad)';

 * ibm_db_sa://pmg68780:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
   sqlite:///my_data1.db
Done.
```

Out[60]:

first_successful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.

```
In [61]: %sql select booster_version from SPACEXTBL where landing_outcome = 'Success (drone ship)' and payload_mass__kg_ between 400

 * ibm_db_sa://pmg68780:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
   sqlite:///my_data1.db
Done.
```

Out[61]:

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

Total Number of Successful and Failure Mission Outcomes

Listing the total number of successful and failure mission outcomes.

```
In [62]: %sql select mission_outcome, count(*) as total_number from SPACEXTBL group by mission_outcome;

 * ibm_db_sa://pmg68780:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
   sqlite:///my_data1.db
Done.
```

Out[62]:

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1
None	898

Boosters Carried Maximum Payload

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

In [65]:
```
%sql select booster_version from SPACEXTBL where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTBL);
```

```
 * ibm_db_sa://pmg68780:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
   sqlite:///my_data1.db
Done.
```

Out[65]:

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

Listing the names of the booster versions which
have carried the maximum payload mass.

2015 Launch Records

Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015.

In [66]:
```
%%sql select monthname(date) as month, date, booster_version, launch_site, landing_outcome from SPACEXTBL
      where landing_outcome = 'Failure (drone ship)' and year(date)=2015;
```

```
 * ibm_db_sa://pmg68780:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
   sqlite:///my_data1.db
Done.
```

Out[66]:

MONTH	DATE	booster_version	launch_site	landing_outcome
October	2015-10-01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

Ranking 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.

Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

In [67]:

```
%%sql select landing_outcome, count(*) as count_outcomes from SPACEXTBL
      where date between '2010-06-04' and '2017-03-20'
      group by landing_outcome
      order by count_outcomes desc;
```

```
 * ibm_db_sa://pmg68780:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
   sqlite:///my_data1.db
Done.
```

Out[67]:

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

Section 3

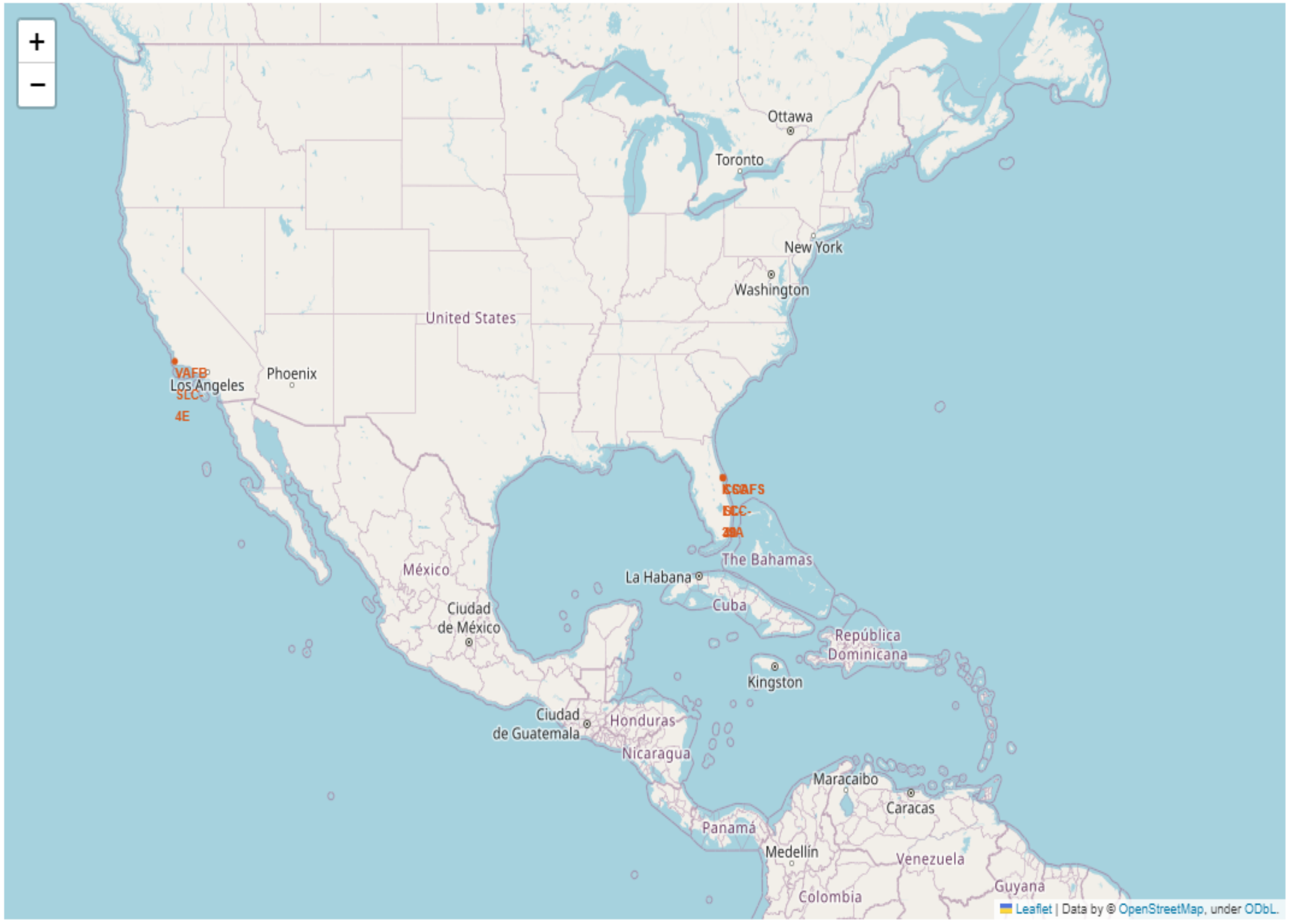
Launch Sites Proximities Analysis

All launch sites' location markers on a global map

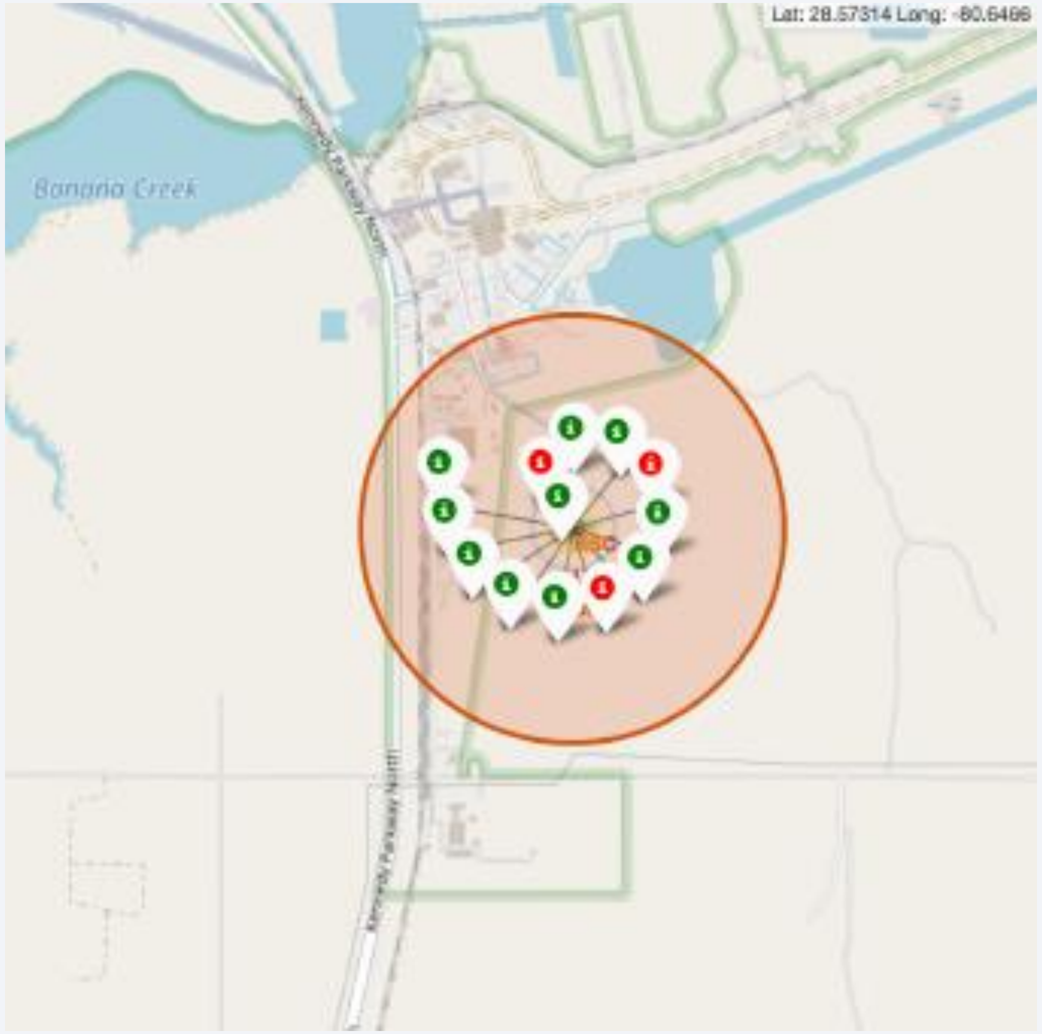
- Most of Launch sites are in proximity to the Equator line. As land is moving faster at the equator than any other place on the surface of the Earth, anything on the surface of the Earth at the equator is already moving at 1670 km/hour. Because of inertia, if a ship is launched from the equator , it goes up into space, and it is also moving around the Earth at the same speed it was moving before launching. This way, the spacecraft keep up a good enough speed to stay in orbit.

- All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimizes the risk of having any debris dropping or exploding near people.

Out[13]:



Color-labeled launch records on the map

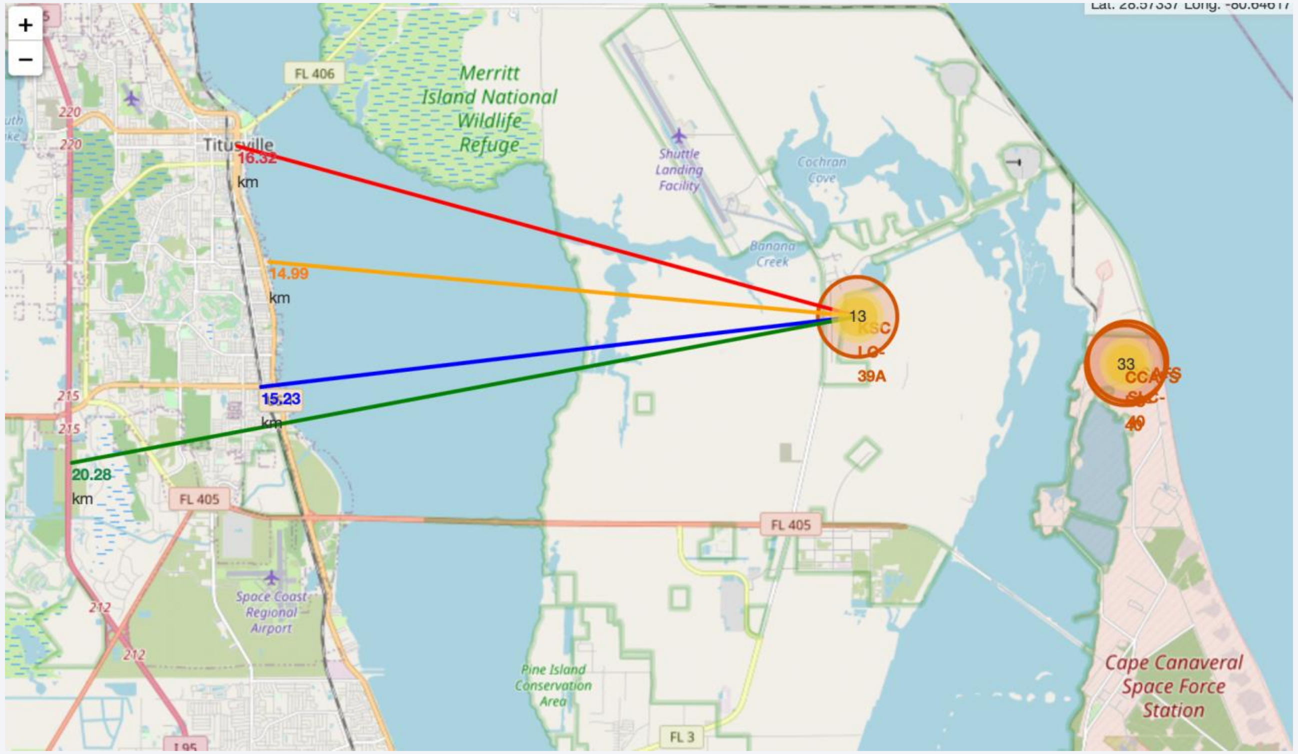


➢ From the color-labeled markers we should be able to easily identify which launch sites have relatively high success rates.

- Green Marker = Successful Launch
- Red Marker = Failed Launch

➢ Launch Site KSC LC-39A is the site with a very high Success Rate.

<Folium Map Screenshot 3>



- ➢ From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:

  - relatively close to railway (15.23 km)
  - relatively close to highway (20.28 km)
  - relatively close to coastline (14.99 km)

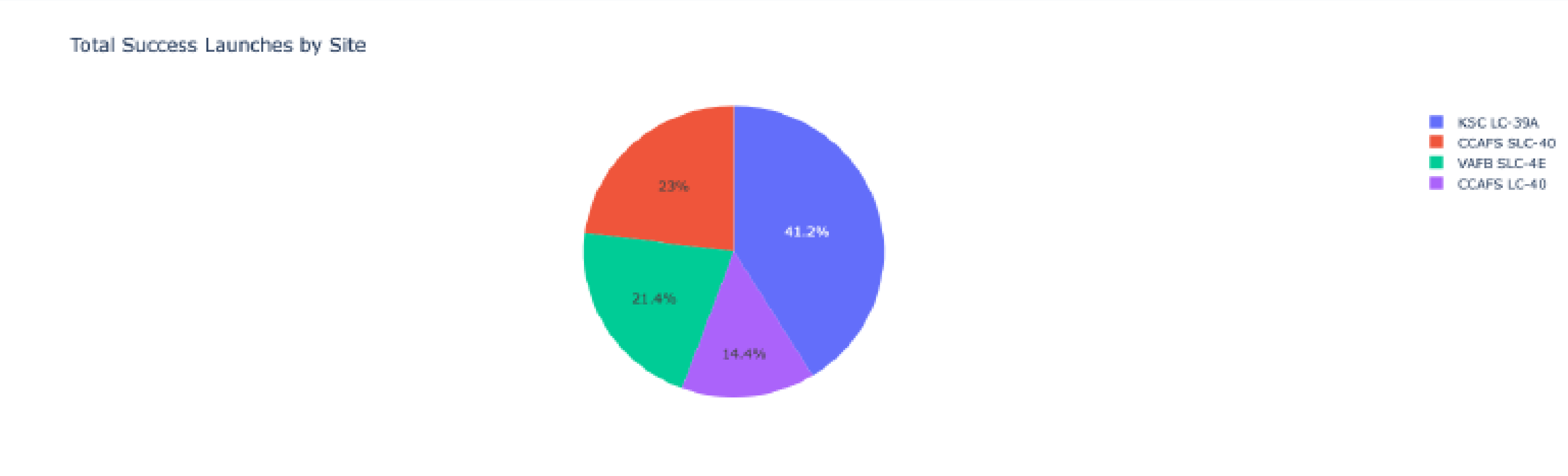
- ➢ The launch site KSC LC-39A is relatively close to its closest city Titusville (16.32 km).

- ➢ Failed rocket with its high speed can cover distances like 15-20 km in few seconds. It could be potentially dangerous to populated areas.

Section 4

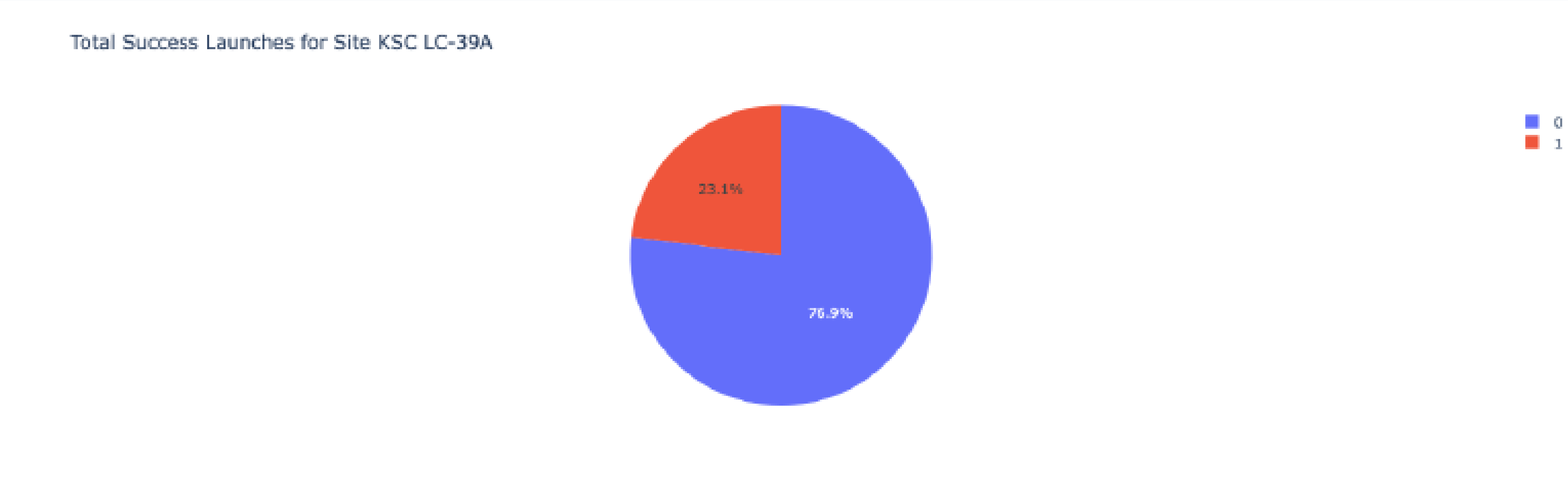
Build a Dashboard with Plotly Dash

Launch success count for all sites



The chart clearly shows that KSC LC-39A has the most successful launches compared to all other sites.

<Dashboard Screenshot 2>



With 10 successful and only 3 failed landings, KSC LC-39A has the highest launch success rate (76.9%).

Payload Mass vs. Launch Outcome for all sites

The charts on the left here show that payloads between 2000 and 5500 kg have the highest success rate.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Out[100…

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

Scores and Accuracy of the Test Set

Based on the scores of the Test Set, we cannot confirm which method performs best.

- Same Test Set scores may be due to the small test sample size which was 18 samples. Because of this, we tested all methods based on the whole dataset.

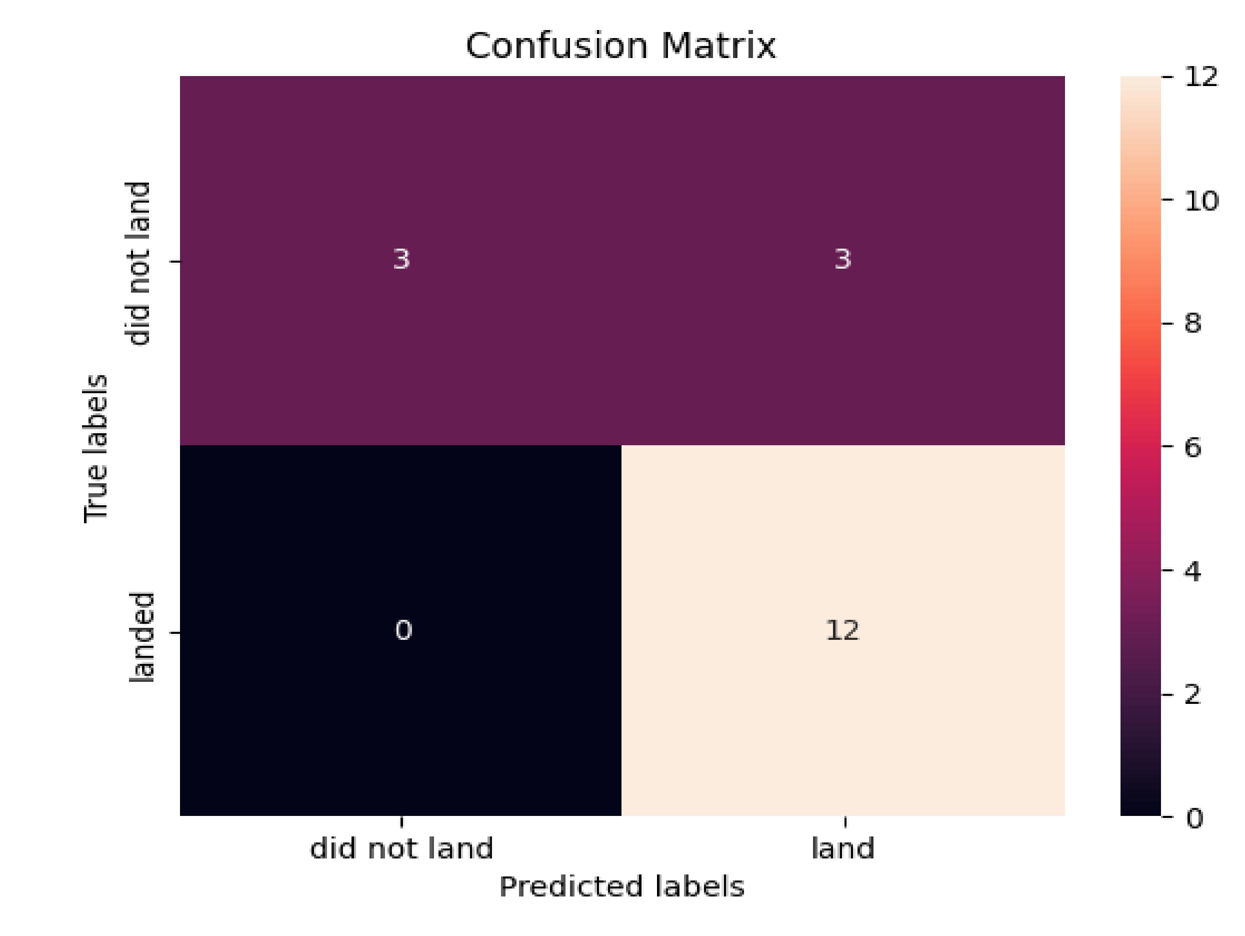
Out[92]:

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.830986	0.819444
F1_Score	0.909091	0.916031	0.907692	0.900763
Accuracy	0.866667	0.877778	0.866667	0.855556

Scores and Accuracy of the Entire Data Set

- The scores of the whole Dataset confirm that the best model is the Decision Tree Model. This model has not only higher scores, but also the highest accuracy.

Confusion Matrix



From the confusion matrix, we see that logistic regression can distinguish between the different classes.

The major problem is false positives.

Conclusions

- The best algorithm for this dataset is Decision Tree Model.
- Launches with a low payload masses show better results than launches with a larger payload masses.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast for safty reasons.
- The success rate of launches increases over the years.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate compared to other orbits

Appendix

[GITHUB URL FOR FULL PROJECT]

Special Thanks to:

COURSERA IBM INSTRUCTORS

SpaceX

Project Jupyter | Home

Python

WIKIPEDIA

GitHub

Thank you!