IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Tom Brown
December 2, 2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - Exploratory Data Analysis with Data Visualization
  - Exploratory Data Analysis with SQL
  - Building an interactive map with Folium
  - Building a Dashboard with Plotly Dash
  - Predictive analysis (Classification)
- Summary of all results
  - Exploratory Data Analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Introduction

- Background
  - SpaceX is a successful company of the commercial space age, making space travel affordable. The company advertises rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.
- Problems you want to find answers
  - How do variables such as payload mass, launch site, number of flights, and orbits impact the success of the first stage landing?
  - Does the rate of successful landings increase over the years?
  - What is the best algorithm that can be used for binary classification in this case?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Utilizing SpaceX Rest API to gather Flight Number, Date, Booster Version, Payload Mass, Orbit, Launch Site, Outcome, Flights, Grid Fins, Reused, Legs, Landing Pad, Block, Reused Count, Serial, Longitude, Latitude
- Perform data wrangling
  - Utilizing Web Scrapping from Wikipedia to gather Flight No., Launch site, Payload, Payload Mass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Create array of data; standardize data; split data into training and testing data; perform GridSearch on four distinct classification models

Data Collection

- The data collection process involved a combination of API requests from SpaceX rest API and web scraping data from a table in SpaceX's Wikipedia entry.

- Both rest API and Wikipedia web scraping collection methods were utilized to allow for a more comprehensive view of the data and allowed for more detailed analysis.

Data Collection – SpaceX API

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

  Flight Number, Date, Booster Version, Payload Mass, Orbit, Launch Site, Outcome, Flights, Grid Fins, Reused, Legs, Landing Pad, Block, Reused Count, Serial, Longitude, Latitude

- [GitHub Jupyter Notebook]

Request rocket launch data from SpaceX API

Decode the response content using .json() and turn it into a dataframe using .json_normalize()

Request required data about the launches from SpaceX API by applying custom functions

Constructing dictionary from data obtained from API

Export the data to CSV

Replace missing values of Payload Mass column with calculated .mean()

Filtering the dataframe to only include Falcon 9 launches

Creating a dataframe from the dictionary

Data Collection - Scraping

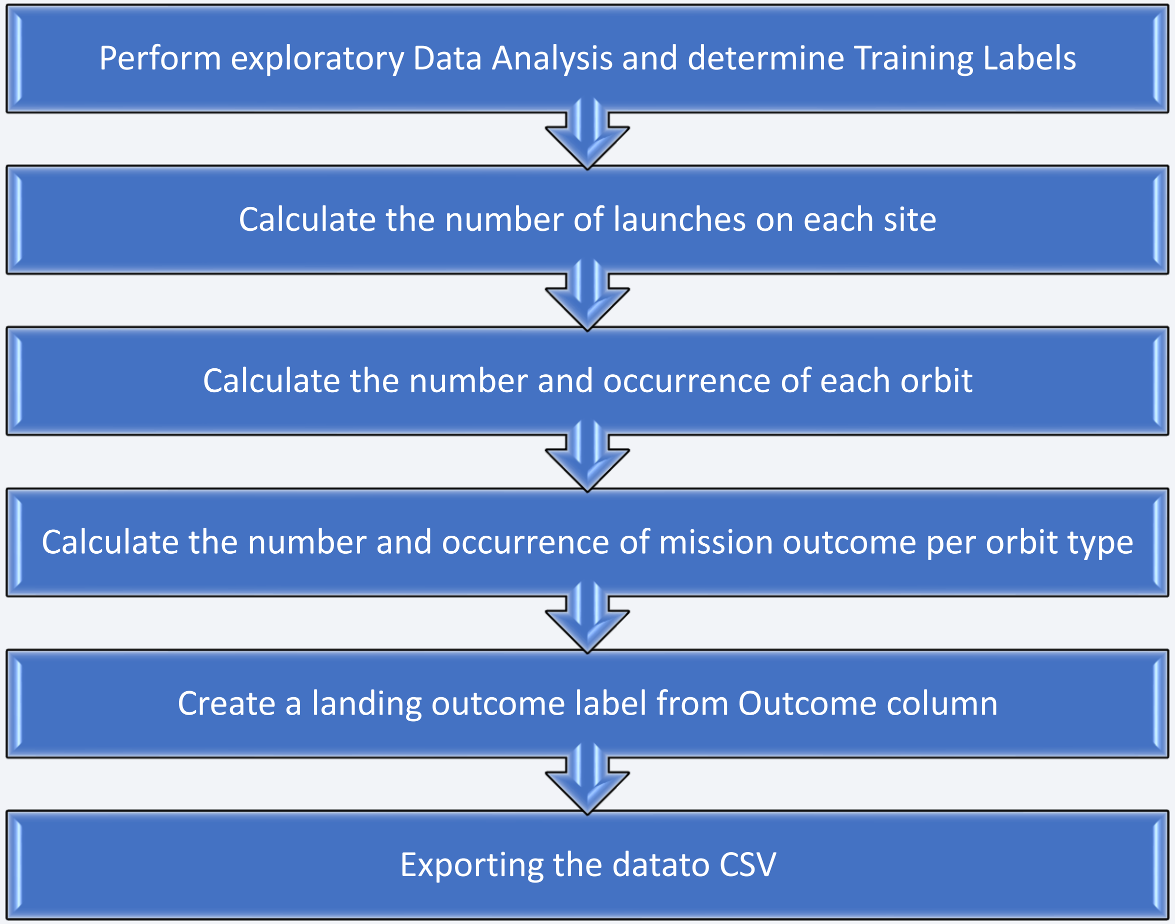
- **Columns obtained by using Wikipedia Web Scraping:**

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

- [GitHub Jupyter Notebook]

Data Wrangling

- There were several different cases where the booster did not land successfully. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully for the ground pad. 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. These conditions were coded to Training labels with “1” meaning successful landing, “0” meaning it was unsuccessful. This allows the values to be analyzed by the model.
- [GitHub jupyter Notebook]



EDA with Data Visualization

Generated Charts	
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	Various scatter plots showing relationship between variables.
Bar chart showing comparison between categories	Line chart showing trends over time.

- GitHub Jupyter Notebook

EDA with SQL

Generated SQL Queries to produce the following information:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string ‘CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing 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
- Ranking the count of landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order
- GitHub Jupyter Notebook

Build an Interactive Map with Folium

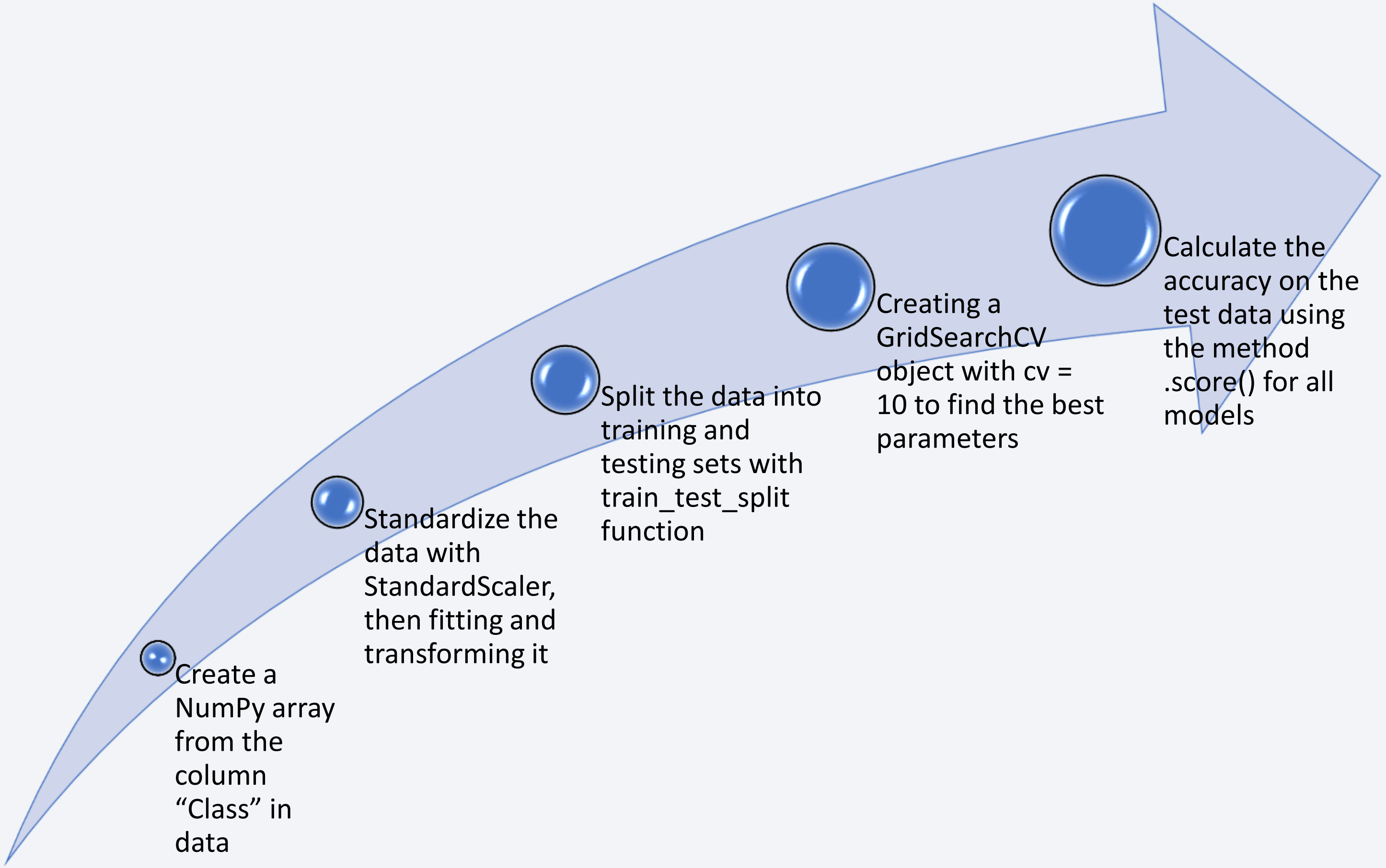
Generated Interactive Folium Maps for the following data:

- Markers of all Launch Sites
- Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.
- 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.
- Colored Markers of the launch outcomes for each Launch Site
- 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 significant Landmarks
- Include colored lines to show distances between the Launch Site CCAFS-SLC-40 and significant landmarks like Railway, Highway, Coastline and Closest City.
- GitHub Jupyter Notebook

Build a Dashboard with Plotly Dash

- Launch Sites Dropdown List
- Added a pie chart to show the total successful launches count for all sites or the selected launch site
    - Success vs. Failed counts for the site, if a specific Launch Site was selected.
- Added a slider to select Payload range.
- Added a scatter chart to show the correlation between Payload and Launch Success.
- Github Plotly Dashboard

Predictive Analysis (Classification)



- [Github Jupyter Notebook]()

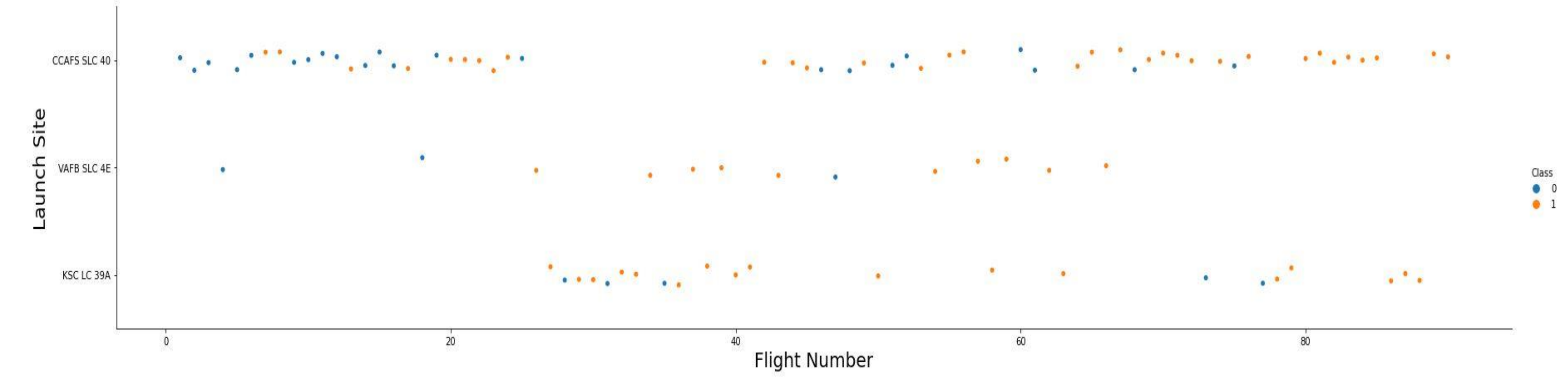
Results

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

Section 2

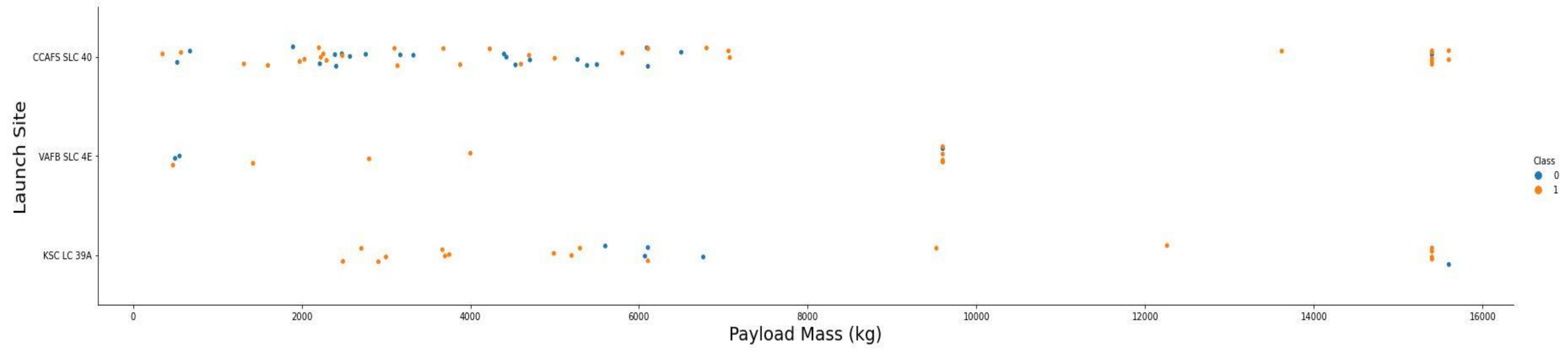
Insights drawn from EDA

Flight Number vs. Launch Site



- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- Higher success rates appear to increase with each flight.

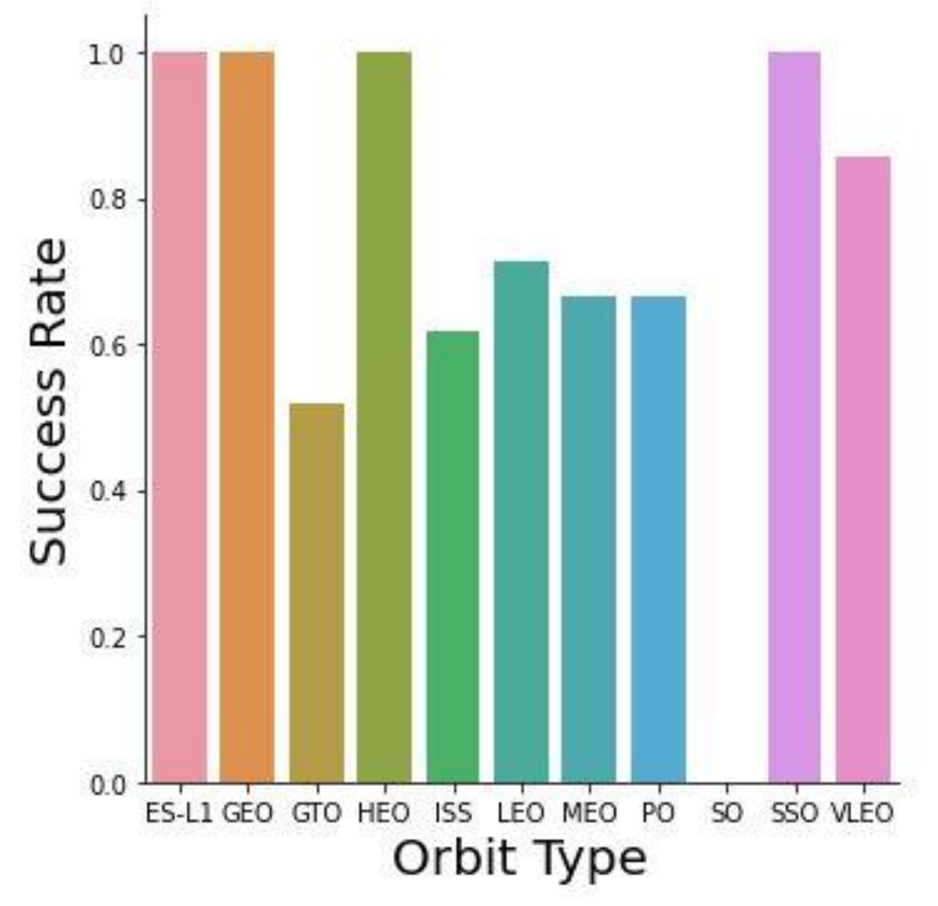
Payload vs. Launch Site



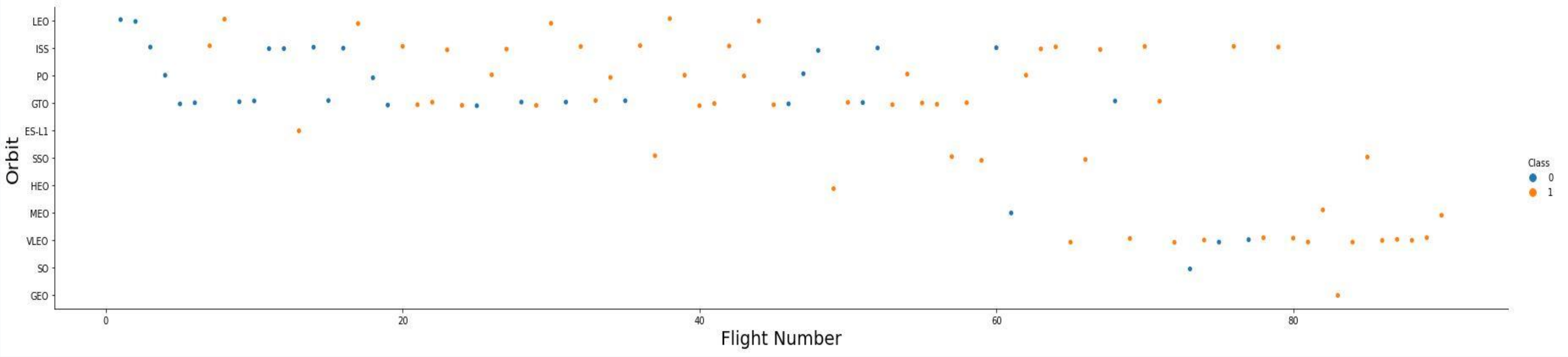
- For every launch site the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg.

Success Rate vs. Orbit Type

- Orbits ES-L1, GEO, HEO, SSO have a 100% success rate
- Orbit SO has a 0% success rate
- All other orbits have a success rate between 50% and 85%.

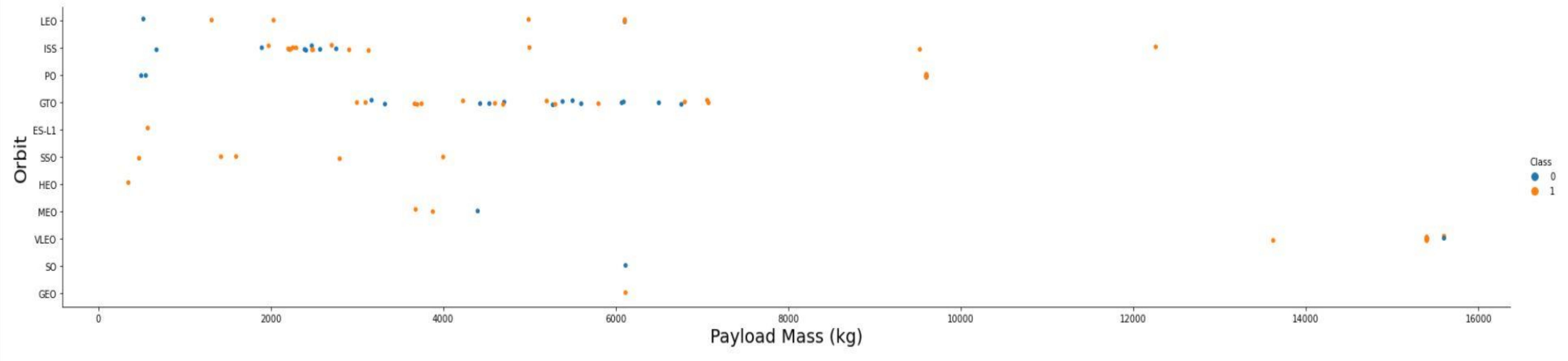


Flight Number vs. Orbit Type



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

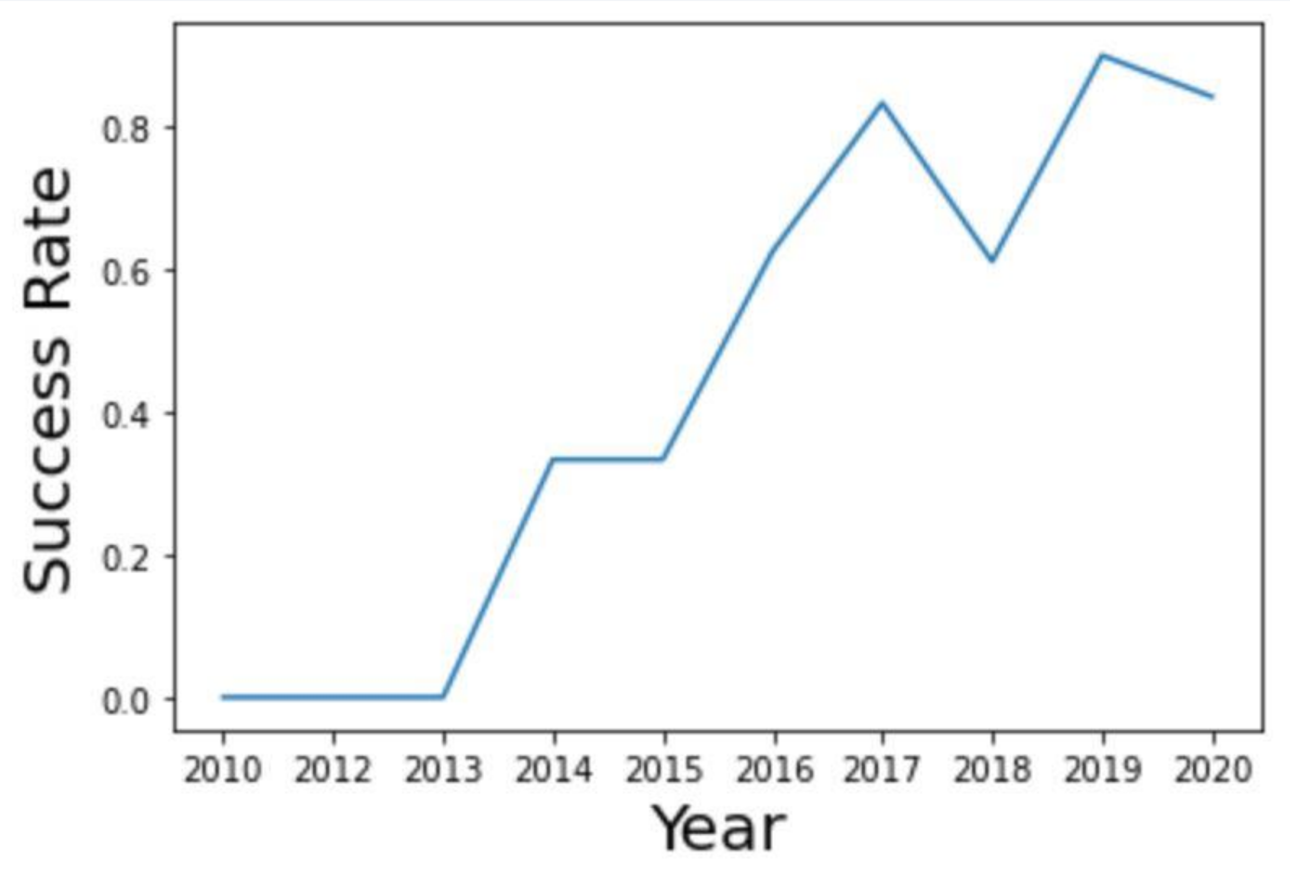
Payload Mass vs. Orbit Type



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

Launch Success Yearly Trend

- With the exception of 2018 and 2020, the success rate has steadily increased since the program began.



All Launch Site Names

```
In [8]: %sql select distinct launch_site from SPACEX;

 * ibm_db_sa://npx07993:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[8]:

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

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

Launch Site Names Begin with 'CCA'

```
In [9]: %sql select * from SPACEX where launch_site like 'CCA%' limit 5;

 * ibm_db_sa://npx07993:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[9]:

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

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

Total Payload Mass

```
In [10]: %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEX where customer = 'NASA (CRS)';

 * ibm_db_sa://npx07993:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.

Out[10]:
```

total_payload_mass
45596

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

Average Payload Mass by F9 v1.1

```
In [11]: %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEX where booster_version like '%F9 v1.1%';

 * ibm_db_sa://npx07993:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.

Out[11]:
```

average_payload_mass
2534

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

First Successful Ground Landing Date

```
In [12]: %sql select min(date) as first_successful_landing from SPACEX where landing__outcome = 'Success (ground pad)';

 * ibm_db_sa://npx07993:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.

Out[12]:
```

first_successful_landing
2015-12-22

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

Successful Drone Ship Landing with Payload between 4000 and 6000

```
In [13]: %sql select booster_version from SPACEX where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000;

 * ibm_db_sa://npx07993:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[13]:

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

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

Total Number of Successful and Failure Mission Outcomes

```
In [14]: %sql select mission_outcome, count(*) as total_number from SPACEX group by mission_outcome;

 * ibm_db_sa://npx07993:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[14]:

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

- Listing the total number of successful and failure mission outcomes.

Boosters Carried Maximum Payload

In [15]:
```
%sql select booster_version from SPACEX where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEX);
```

```
 * ibm_db_sa://npx07993:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[15]:

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

```
In [17]: %sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEX where landing__outcome = 'Fa:

 * ibm_db_sa://npx07993:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[17]:

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

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

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

```
In [20]: %sql select landing__outcome, count(*) as count_outcomes from SPACEX where date between '2010-06-04' and '2017-03-20' group by landi
```

```
 * ibm_db_sa://npx07993:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[20]:

landing__outcome	count_outcomes
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

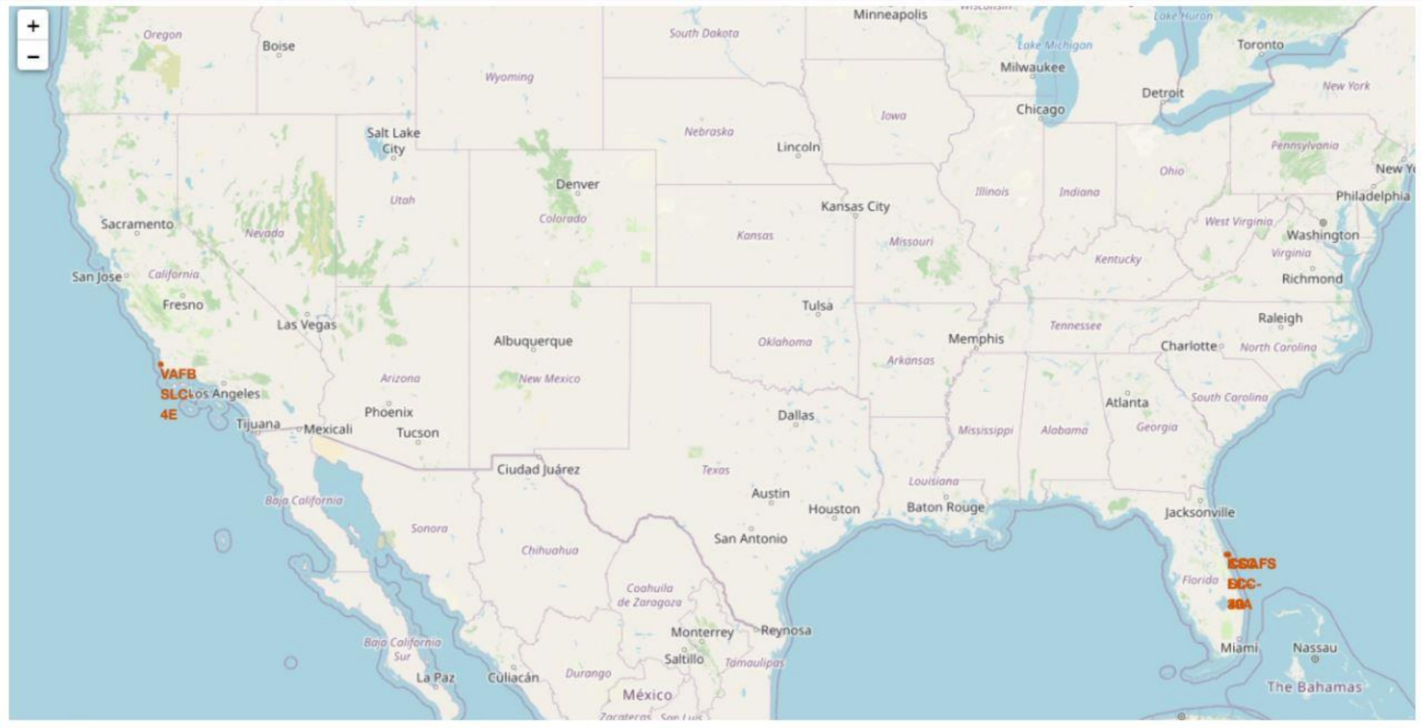
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

Section 3

Launch Sites Proximities Analysis

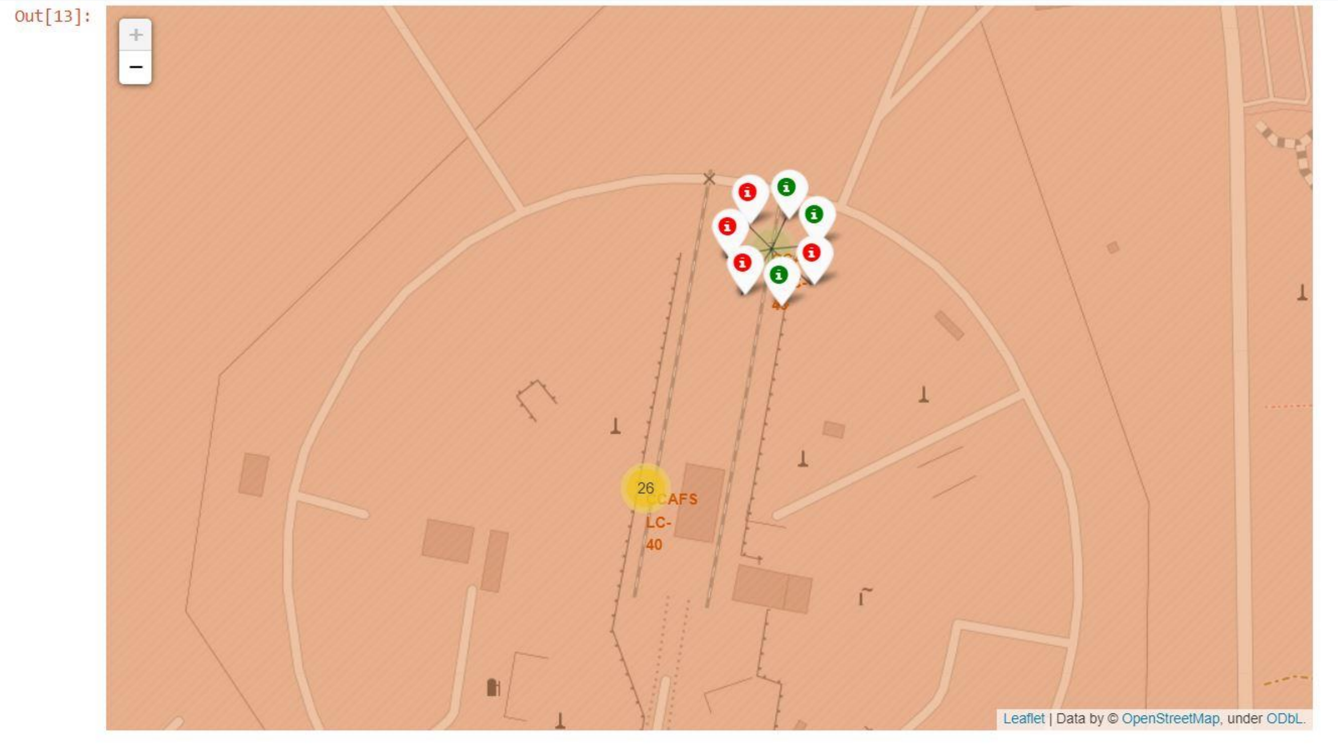
SpaceX Launch Sites

- Launch sites are in proximity to the Equator line.
- The land is moving faster at the equator than any other place on the surface of the Earth.
- Ships launched from the equator it go up into space, stay in space because of inertia and they are traveling at the same speed as they were launched.
- Launch sites are in close proximity to the coast to minimize the risk of having any debris dropping or exploding near people.



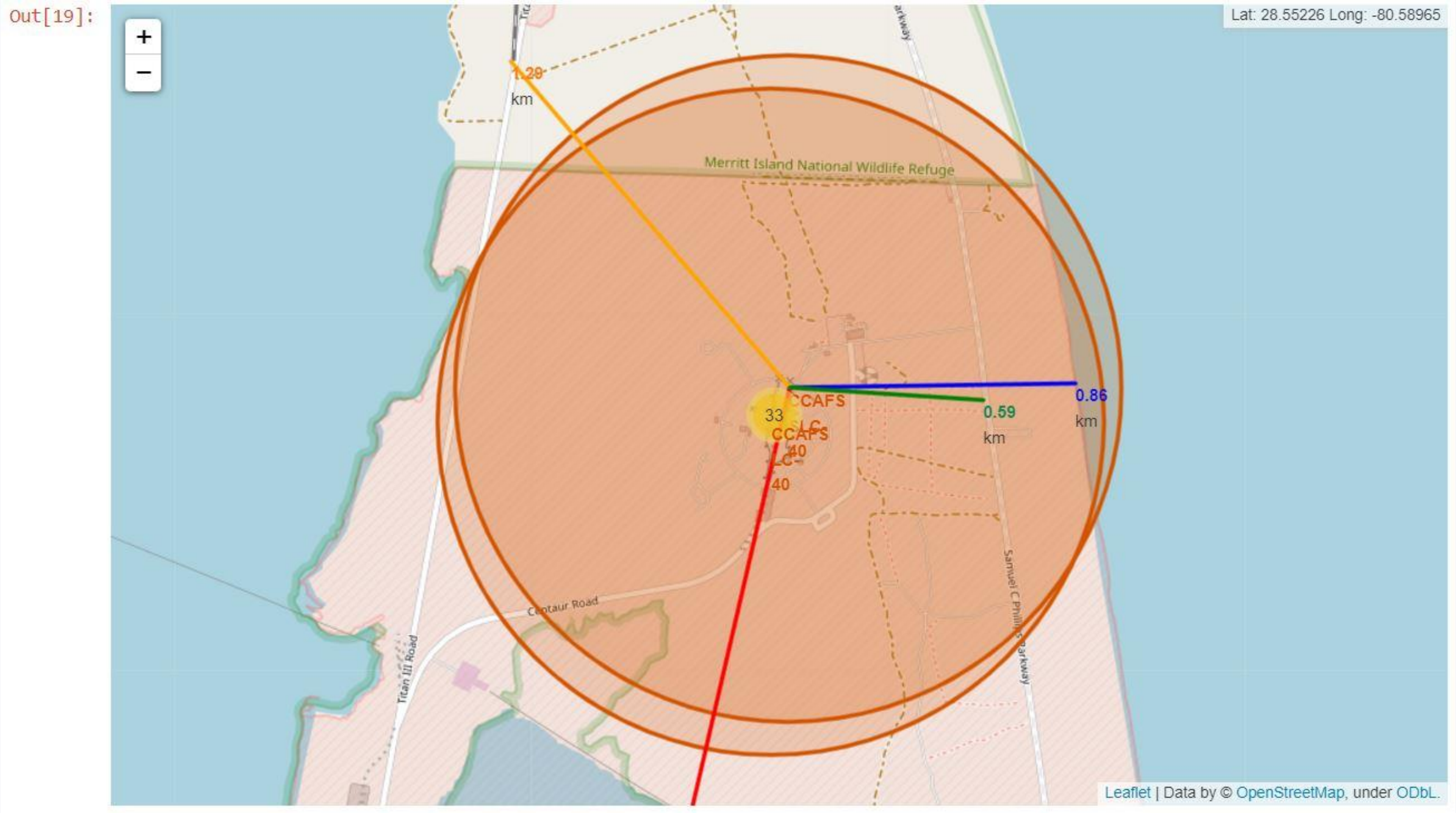
Marked Color Launch Locations by Coordinates

- Successful and failed launches are easily identified because of the colored markers.
- Green Marker = Successful Launch
- Red Marker = Failed Launch



Launch Location Relative to Landmarks

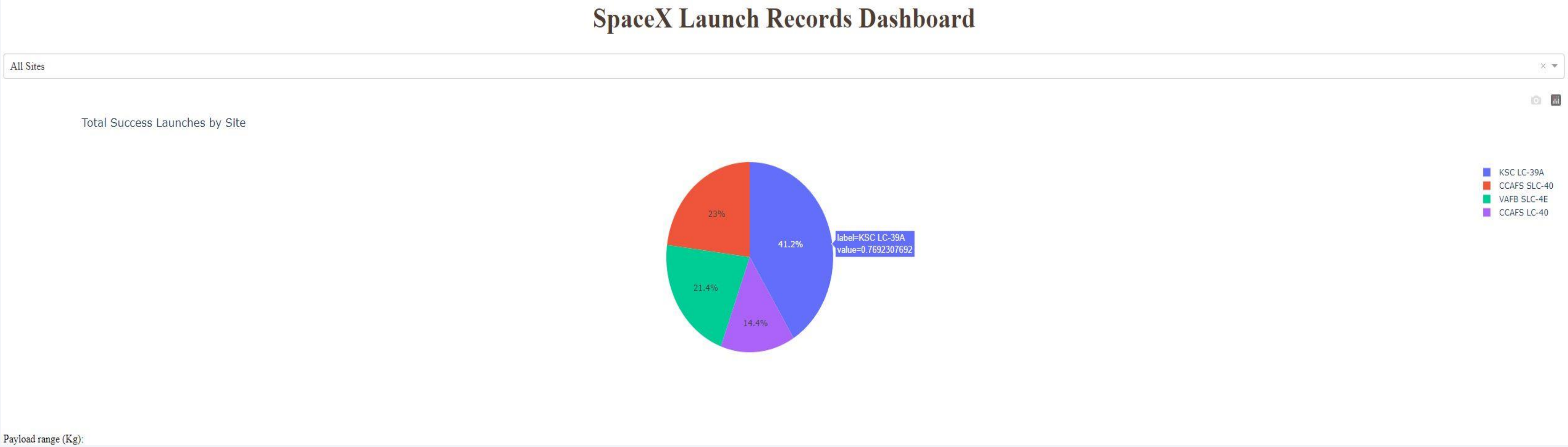
- From the visual analysis of the launch site CCAFS-SLC-40 we can clearly see that it is:
  - Proximity to City (17.96 km)
  - Proximity to Coastline (0.86 km)
  - Proximity to Highway (0.59 km)
  - Proximity to Railway (1.29 km)
- Failed rocket with its high speed can cover distances like 15-20 km in few seconds. It could be potentially dangerous to populated areas.



Section 4

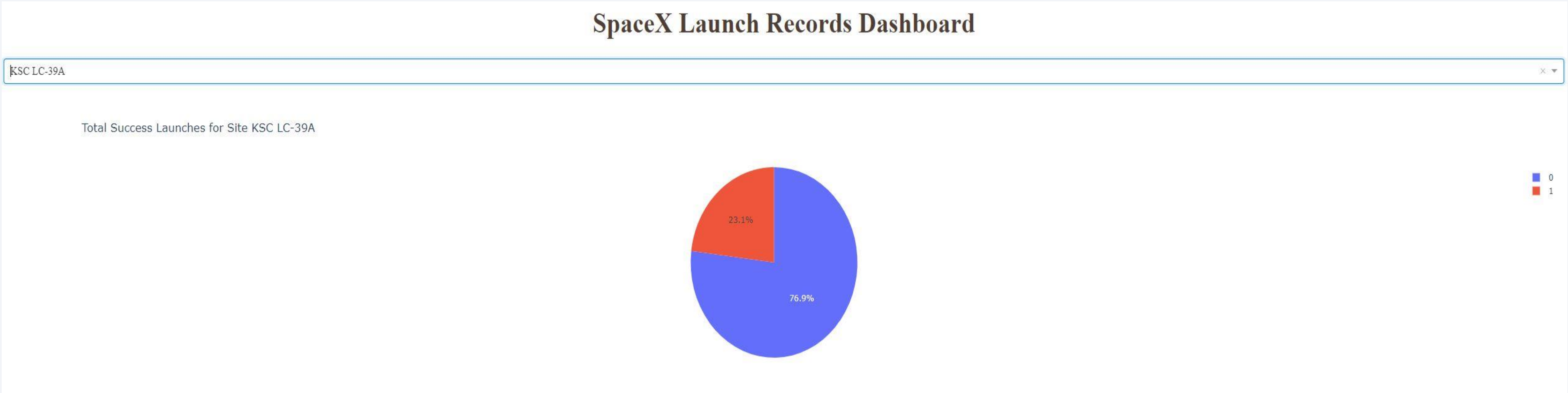
Build a Dashboard with Plotly Dash

Launch Percentages for All Launch Sites



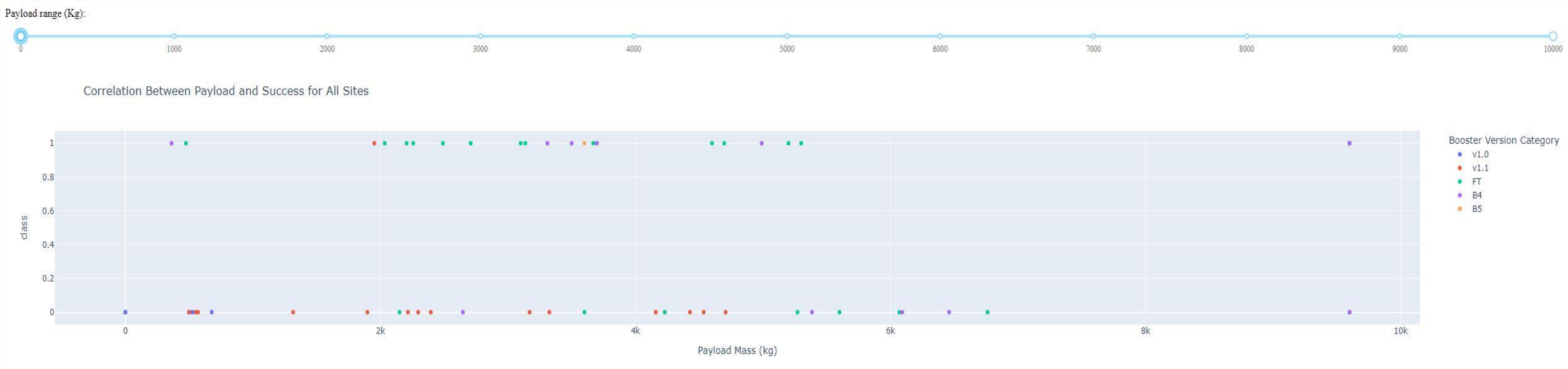
- The chart shows that from all the sites, KSC LC-39A has the most successful launches.

Launch Percentage for Most Successful Site



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

Payload Mass vs. Launch Outcome All Sites



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

Section 5

Predictive Analysis (Classification)

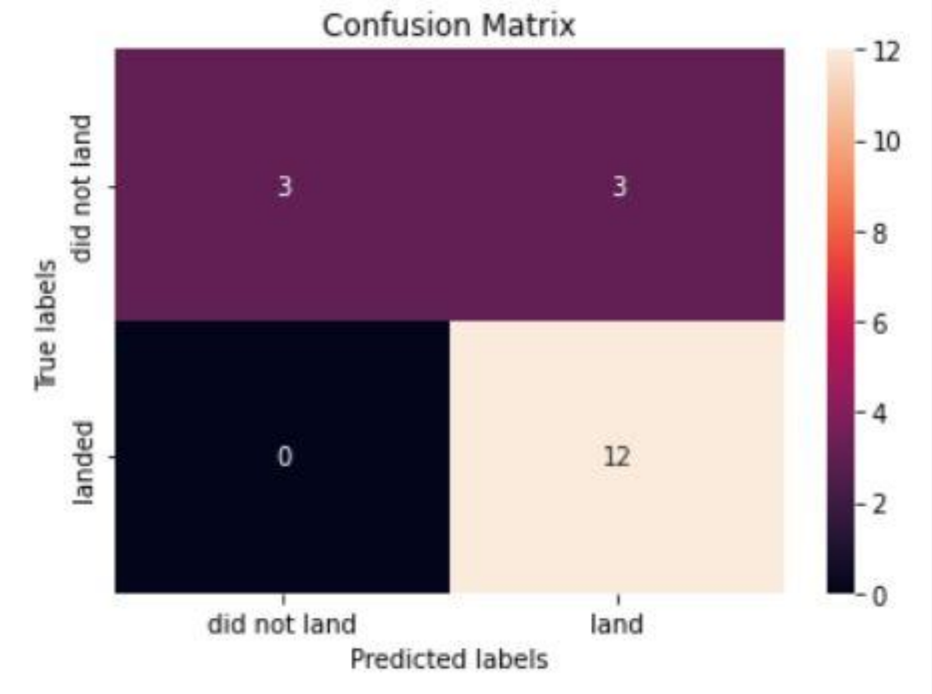
Classification Accuracy

- Based on the scores of the Test Set, we can not confirm which method performs best.
- Same Test Set scores may be due to the small test sample size (18 samples).

	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	0.888889	0.833333

Confusion Matrix

- Examining the confusion matrix, the logistic regression can distinguish between the different classes. The largest problem is false positives.



Conclusions

- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.
- Launches with a low payload mass show better results than launches with a larger payload mass.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- The success rate of launches increases over the years.

Thank you!