IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Gökhan Sancak
15-07-2022

Outline

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

Executive Summary

Summary of Methodologies

- Data Collection (API + Web scraping)
- Data Wrangling
- EDA with Data Visualization
- EDA with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive Analysis (Classification)

Summary of results

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

Introduction

- Project background and context

  In this project, we will predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 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. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

- Problems you want to find answers
  - What factors determine the successful or failing of a rocket landing?
  - How much of an influence do these factors have on such a landing?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API
  - Web scraping from Wikipedia
- Perform data wrangling
  - OneHot Encoding data and removal of irrelevant fields
- 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
  - Standardize data and split into training and test data
  - Find best Hyperparameter for SVM, Classification Trees and Logistic Regression

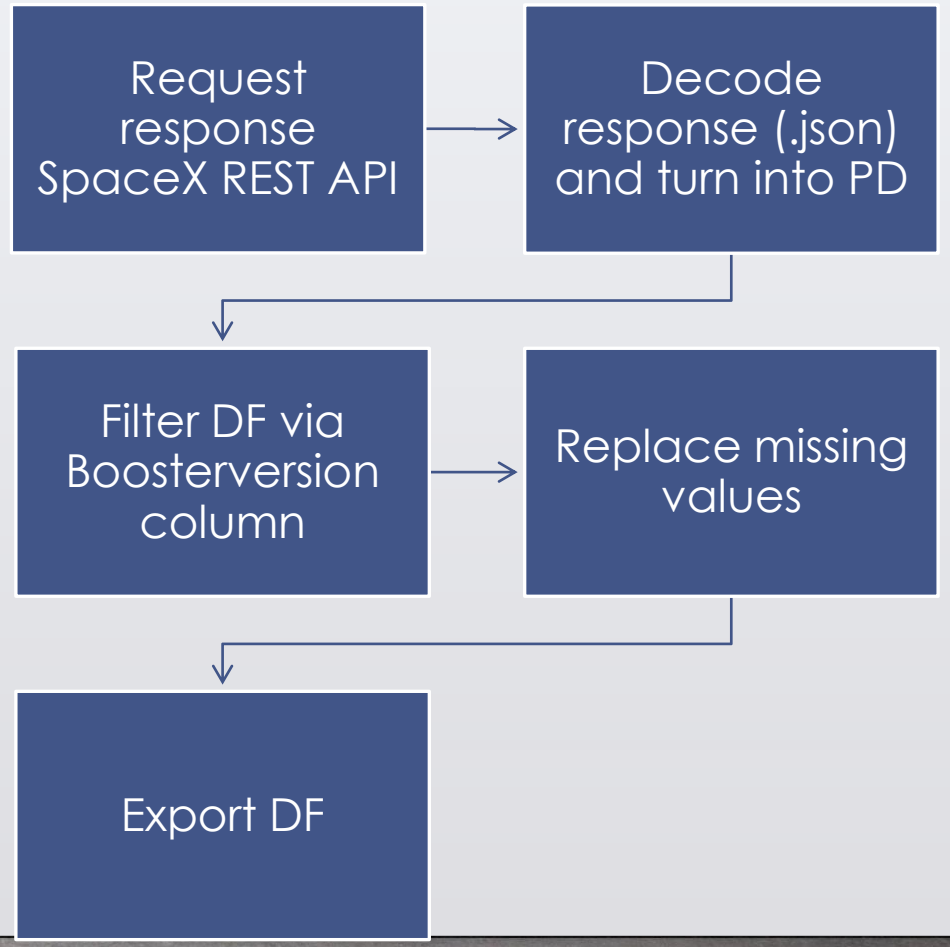
Data Collection

- Collected datasets:
  - SpaceX API for SpaceX Launch data: prediction of landing success
    - Launch data includes: rocket name, payload, launch site, landing outcome
  - Falcon 9 historical launch records (via web scraping (BeatifulSoup) from Wikipedia): examples of successful and unsuccessful landings

Data Collection – SpaceX API

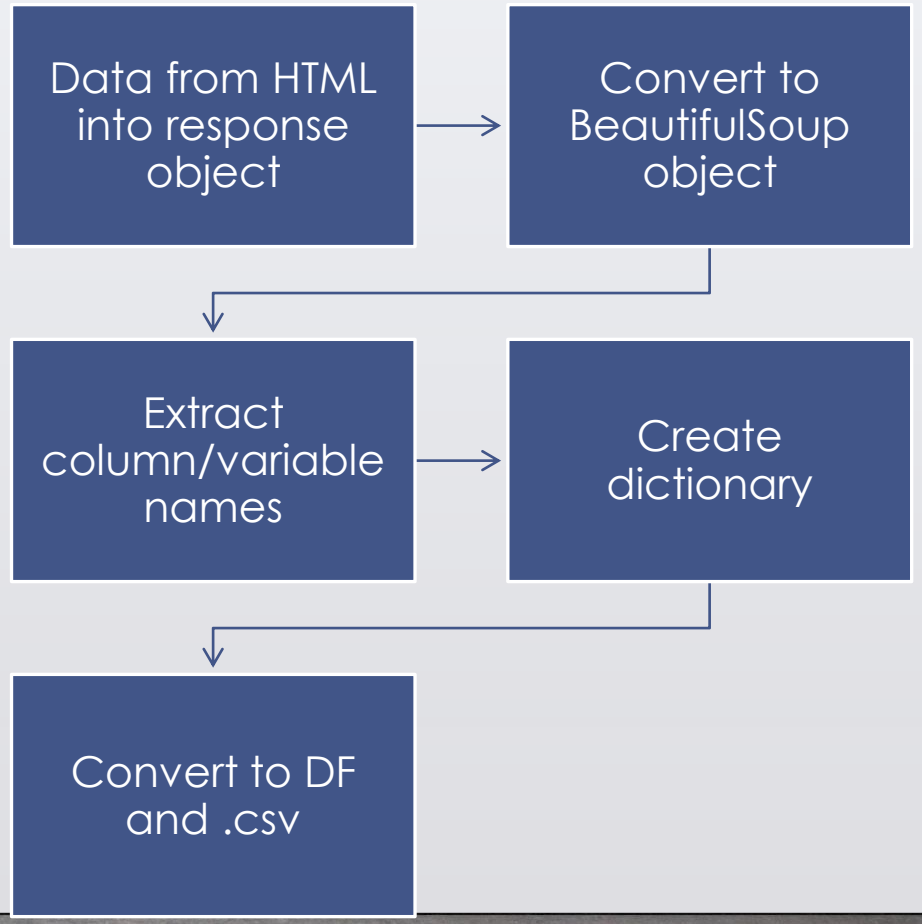
- Request SpaxeX REST API via url and store in response
- Decode response content using .json and turn into Pandas dataframe
- Filter dataframe to only include Falcon 9 launches using BoosterVersion column
- Replacing missing values with mean for PayloadMass column
- Export dataframe to CSV file

- Github SpaceX API



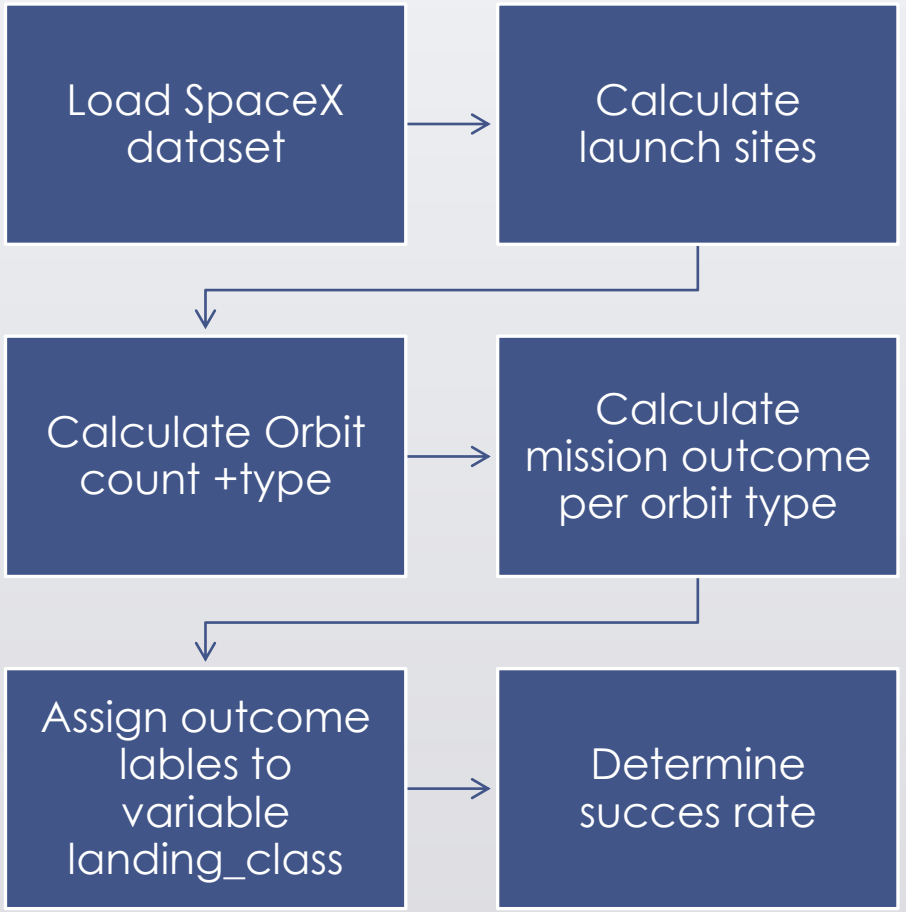
Data Collection - Scraping

- Scrape data from Wikipedia URL on 9th June 2021 in HTML response object
- Create BeautifulSoup object from HTML response
- Extract column/variable names
- Create dictionary with keys from column names
- Convert dictionary to dataframe and dataframe to .csv
- Github Data Collection - Scraping



Data Wrangling

- Load the SpaceX dataset
- Calculating number of launch sites via value_counts in LaunchSite column
- Every launch aims for an orbit which we also count by the column 'Orbit'
- Calculate the number and occurrence of mission outcome per orbit type (for. Example: True Ocean=successful landing, False Ocean=unsuccessful ocean landing)
- Create variable landing_class so we can assign outcome labels (1=successful landing, 0=unsuccessful)
- Determine success rate of landing class and export to .csv
- Github Data Wrangling



EDA with Data Visualization

- Scatter plot
  - Flight Number vs. PayloadMass, Flight Number vs. LaunchSite, PayloadMass vs. LaunchSite, Orbit vs. Flight Number, PayloadMass vs. Orbit
  - Scatter plots are particularly helpful graphs when we want to see if there is a linear relationship among data points. They indicate both the direction of the relationship between the x variables and the y variables, and the strength of the relationship.
- Bar Chart
  - Orbit vs. Mean
  - A bar diagram makes it easy to compare sets of data between different groups at a glance. The graph represents categories on one axis and a discrete value in the other. The goal is to show the relationship between the two axes. Bar charts can also show big changes in data over time
- Line Chart
  - Year vs. Success Rate
  - A line graph is a graphical display of information that changes continuously over time. Within a line graph, there are various data points connected by a straight line that reveals a continuous change in the values represented by the data points
- Github EDA with Data Visualization

EDA with SQL

- SQL queries:

- Display the names of the unique launch sites in the space mission (select distinct)
- Display 5 records where launch sites begin with the string 'CCA' (select *, where, limit 5)
- Display the total payload mass carried by boosters launched by NASA (CRS) (select sum)
- Display average payload mass carried by booster version F9 v1.1 (select avg)
- List the date when the first succesful landing outcome in ground pad was acheived. (select min(DATE), where)
- List the date when the first succesful landing outcome in ground pad was acheived. (select, where, =)
- List the total number of successful and failure mission outcomes (select count)
- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery (select, where, =select)
- 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. (select extract)
- Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order. (select, order by…desc)
- Github EDA with SQL

Build an Interactive Map with Folium

- Marked all launch sites via a Circle Marker with a label of it's name. Used the latitude and longitude coordinates for the launch site locations. This way we can easily identify launch sites in the map.
- Marked the successful (Green, class=1) and unsuccessful (Red, class=0) launches for each launch site in a MarkerCluster(). This way we can easily identify which launch sites have relatively high success rates.
- Using MousePosition we can select landmarks as highways or coast lines to measure the distance to a launch site (with drawn lines and a marker). We can use this info to gain insights into safety and access to launch sites.
- Github Interactive Map with Folium

Build a Dashboard with Plotly Dash

- Dropdown menu/list
  - To select SpaceX Launch records
- Pie chart
  - To show the total successful launches count for all sites
  - To select a specific launch and show the Success vs. Failed counts for the site
- Slider
  - To select the payload range of the rocket as to identify insights into success for different payloads
- Scatter chart
  - To show the correlation between payload and launch success
- Github Dashboard with Ploty Dash

Predictive Analysis (Classification)

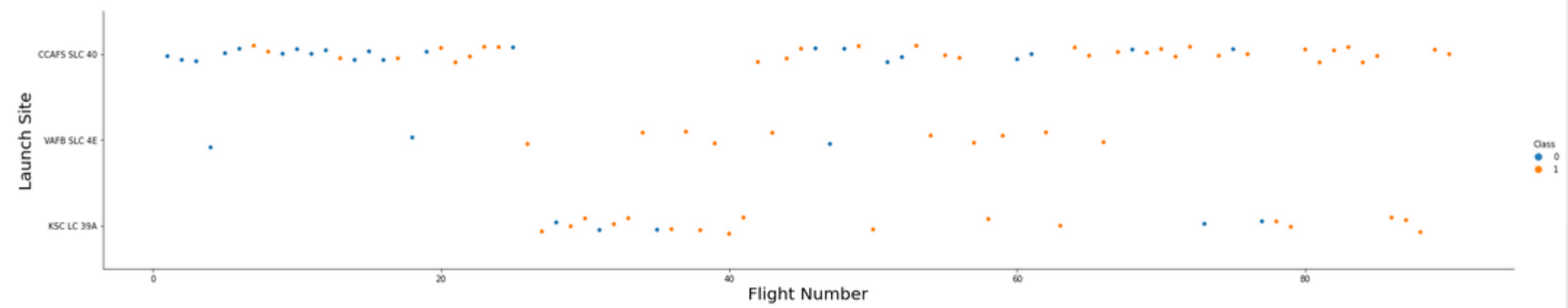
- Building the model
  - Load dataset and create NumPy array and assign output to Pandas series
  - Standardize and transform data
  - Split data into training and test data
  - Use different machine learning algorithms and create GridSearchCV objects
  - Fit the objects from the dictionary parameters
- Evaluating the model
  - Calculate the accuracy on the test data for each model
  - Find the best parameters
  - Plot confusion matrices
- Improving the model
  - Change algorithm
- Finding the best performing model
  - Check for best accuracy per model
  - Model with highest accuracy (closest to 1) is the best performing model
- Github Predictive Analysis (Classification)

Results

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


Section 2
Insights drawn from EDA

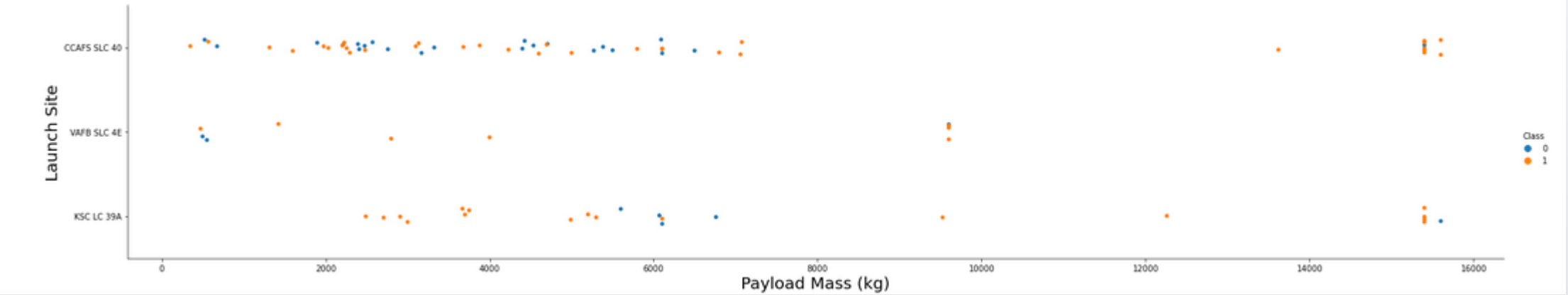
Flight Number vs. Launch Site



Scatter plot of Flight Number vs. Launch Site

- CCAFS SLC 40 Launch Site we see that as flight number increases it becomes more likely for a successful landing
- Generally successful landings increase as flight numbers rise for all launch sites

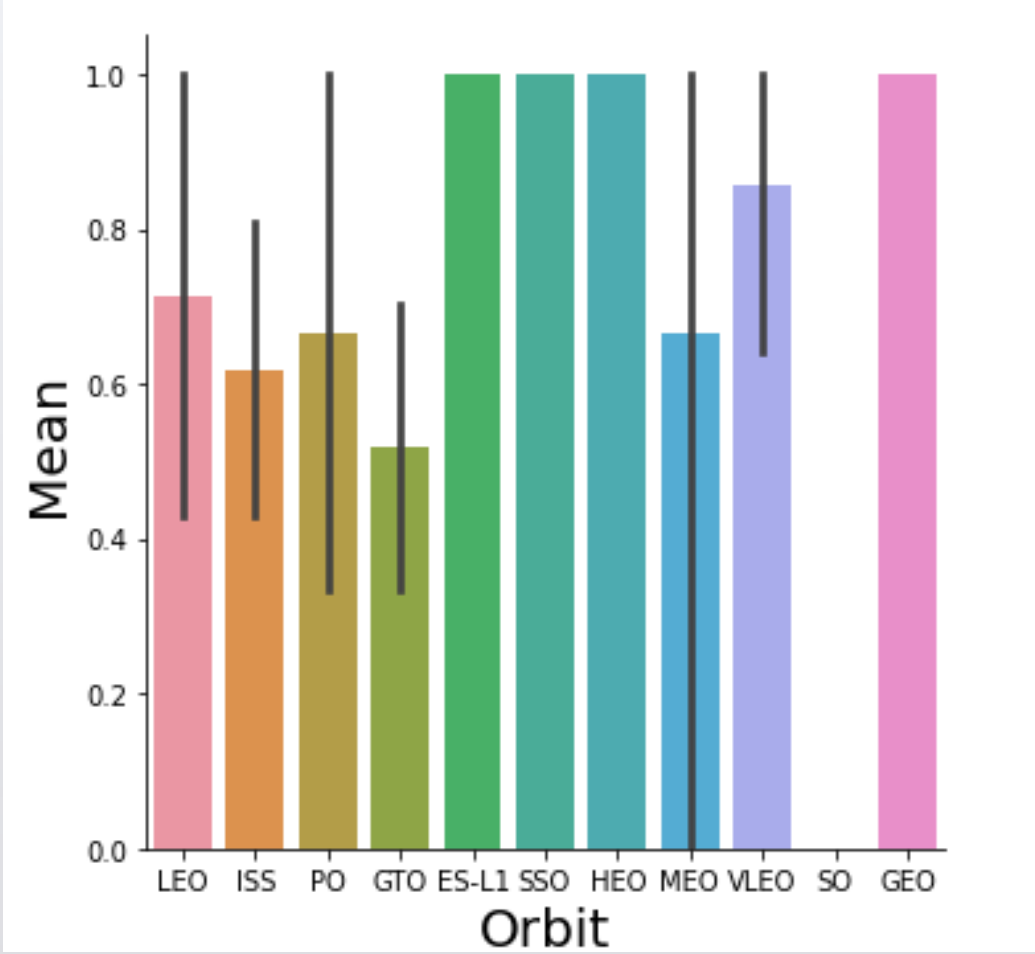
Payload vs. Launch Site



- Scatter plot of Payload vs. Launch Site
- VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).
- CCAFS SLC40 is more successful with a payload heavier than approx. 7000 kg
- KSC LC 39A is mostly unsuccessful around 5000 – 7000 kg

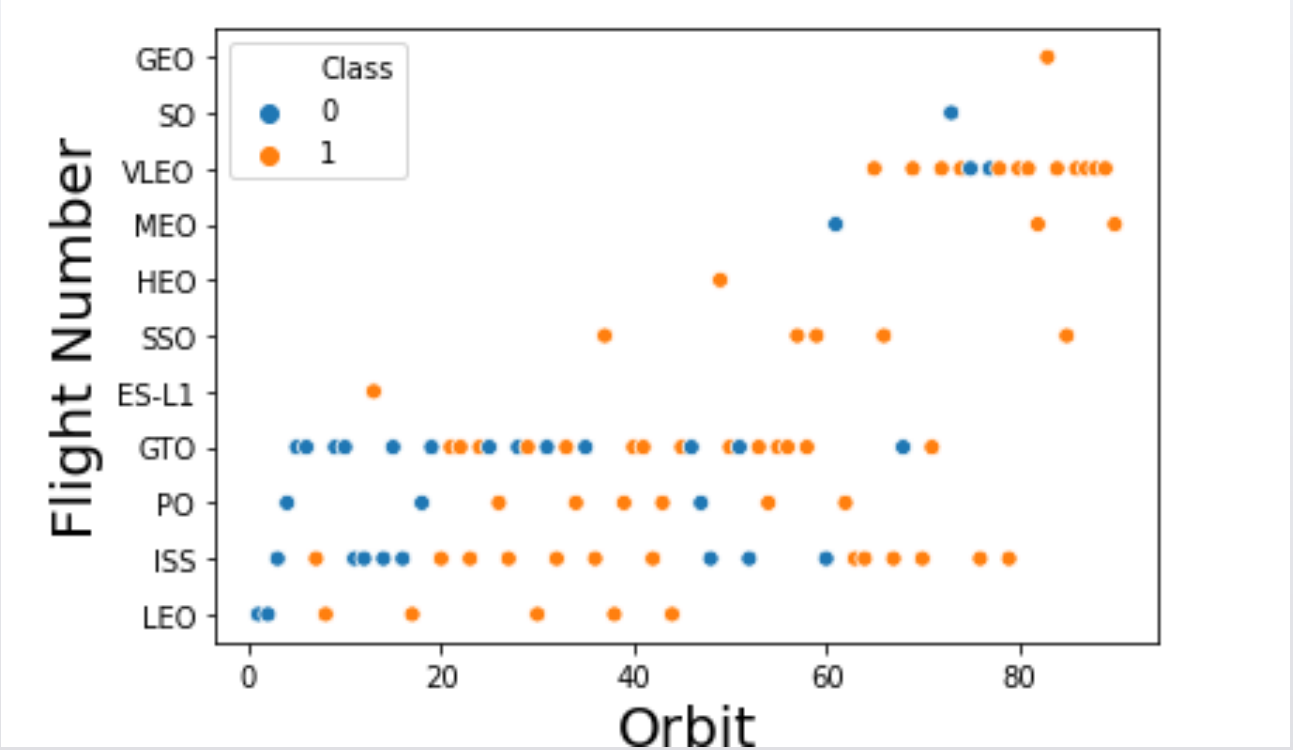
Success Rate vs. Orbit Type

- Bar chart for the success rate of each orbit type
- The Orbits ES-L1, SSO, HEO and GEO have the high success rates



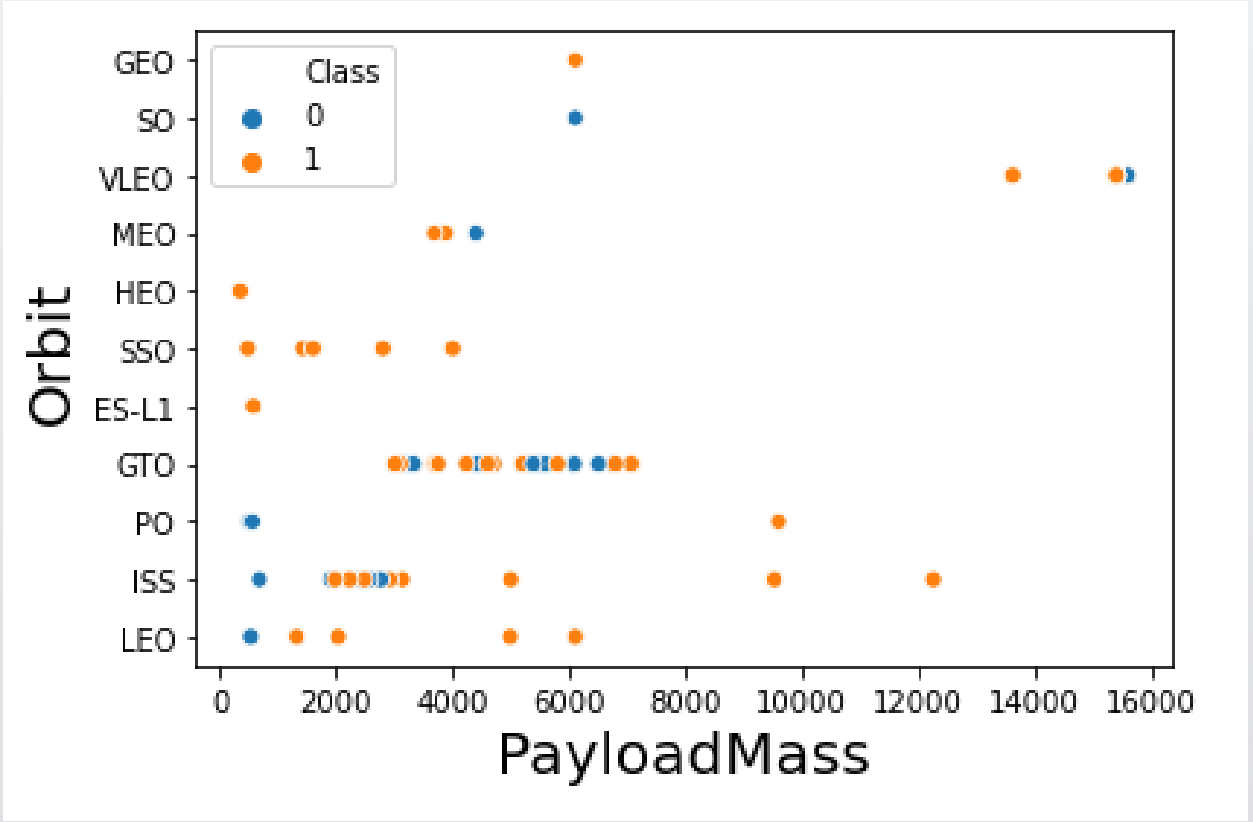
Flight Number vs. Orbit Type

- Scatter point of 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.



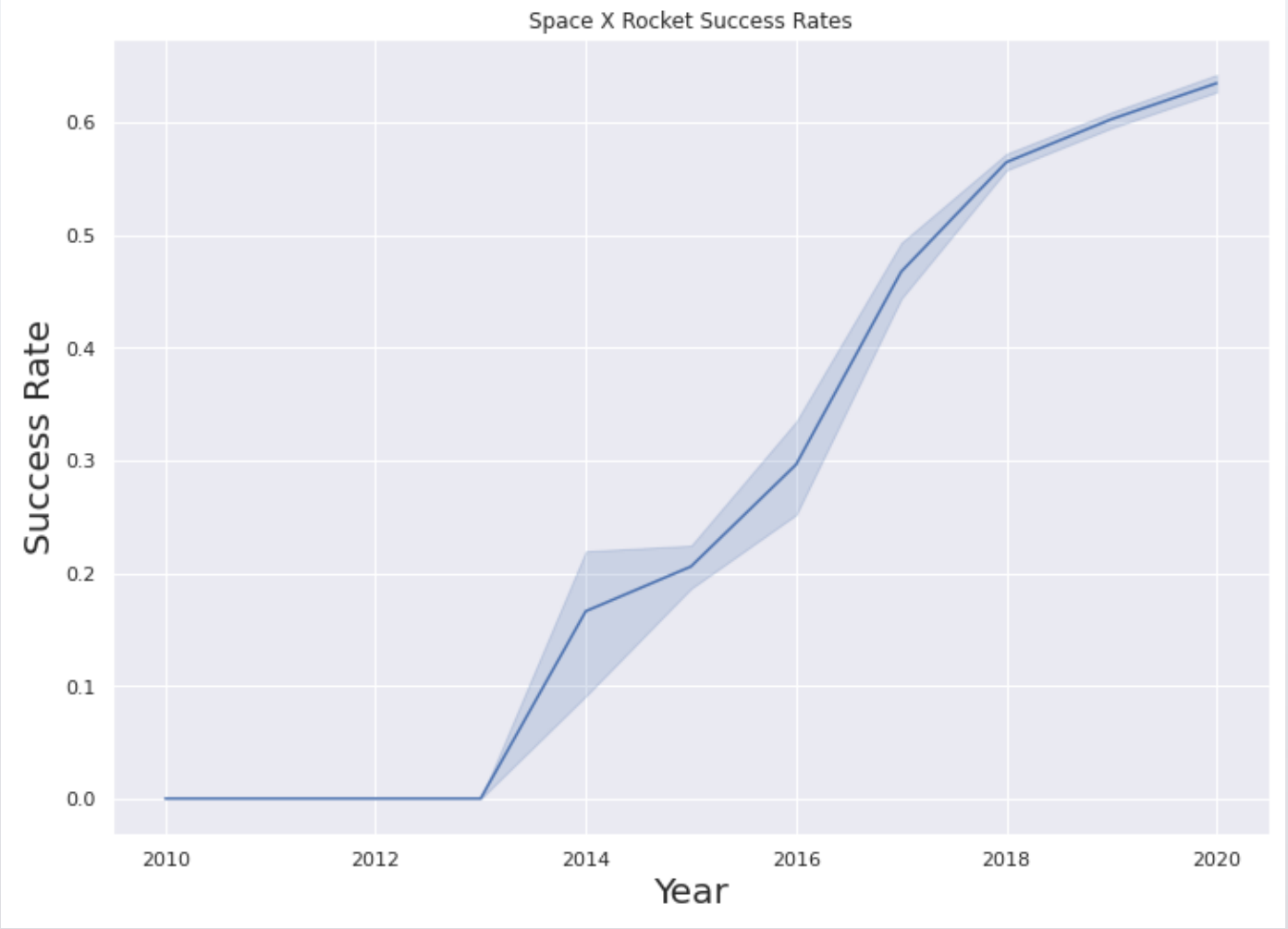
Payload vs. Orbit Type

- Scatter point of payload vs. orbit type
- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS
- 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

- Line chart of yearly average success rate
- The success rate keeps increasing between 2013 and 2020



All Launch Site Names

- Names of the unique launch sites: CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40
- In the query 'Distinct' is used which will only show us unique values in the Launch_site column for the SpaceXtbl

```
%sql select distinct(LAUNCH_SITE) from SPACEXTBL
```

```
 * sqlite:///my_data1.db
Done.
```

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA` (highlighted in picture)
- ‘Limit 5’ in the query means it will only show 5 records from SPACEXTBL and ‘like’ keyword shows us Launch Sites with ‘CCA%’. The percentage indicates Launch Site name must start with CCA

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

```
 * sqlite:///my_data1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Total payload carried by boosters from NASA: 45596 kg
- ‘Sum’ function calculates total in column Payload_Mass_KG_
- ‘Where’ clause filters dataset to only take Customer NASA (CRS)

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

```
 * sqlite:///my_data1.db
Done.
```

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1: 2928.4 kg
- ‘AVG’ function calculates average of column Payload_Mass_KG_
- ‘Where’ clause filters dataset to only use Booster_version F9 v1.1

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

```
 * sqlite:///my_data1.db
Done.
```

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- ‘Min’ function takes minimum date in Column ‘Date
- ‘Where’ clause filters dataset to only use ‘Landing_Outcome=Success(ground pad)’

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

```
 * sqlite:///my_data1.db
(sqlite3.OperationalError) no such column: Landing_Outcome
[SQL: select min(DATE) from SPACEXTBL where Landing_Outcome = 'Success (ground pad)']
(Background on this error at: http://sqlalche.me/e/e3q8)
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- ‘Booster_Version’ selected from SPACEXTBL
- ‘Where’ clause filters dataset to ‘Landing_Outcome=Success (drone ship)’
- ‘And’ clause filters that Payload_Mass_KG_ has 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
```

```
 * sqlite:///my_data1.db
(sqlite3.OperationalError) no such column: Landing_Outcome
[SQL: select BOOSTER_VERSION from SPACEXTBL where Landing_Outcome  = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000]
(Background on this error at: http://sqlalche.me/e/e3q8)
```

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes: 99
- ‘Count’ clause calculates the amount Mission_Outcome
- ‘Where’ clause filters dataset Mission_Outcome in success and failure

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

```
 * sqlite:///my_data1.db
Done.
```

count(MISSION_OUTCOME)
99

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass
- Selecting Booster_version from SpaceXTBL with 'Where' clause to only take Payload_Mass_LKG_ column. Subquery selects maximum value of this column and checks which booster_version it belongs to

```
%sql select BOOSTER_VERSION from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
 * sqlite:///my_data1.db
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

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Select 2015 year from SPACEXTBL with ‘where’ clause to filter for ‘Landing_Outcome=Failure in flight’

```
%sql SELECT EXTRACT(select YEAR(Date, 2015)) from SPACEXTBL where Landing__Outcome = 'Failure (in flight)')
```

```
 * sqlite:///my_data1.db
(sqlite3.OperationalError) near "select": syntax error
[SQL: SELECT EXTRACT(select YEAR(Date, 2015)) from SPACEXTBL where Landing__Outcome = 'Failure (in flight)')]
(Background on this error at: http://sqlalche.me/e/e3q8)
```

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
- Selecting all data from SPACEXTBL by * and filtering it with ‘where’ clause for Landing_Outcome with ‘like’ so we get the result ‘Success’ with dates between 2010-06-04 and 2017-03-20
- ‘Desc’ orders the selected dates in descending order

```
%sql select * from SPACEXTBL where Landing__Outcome like 'Success%' and (DATE between '2010-06-04' and '2017-03-20') order by date desc
```

```
 * sqlite:///my_data1.db
(sqlite3.OperationalError) no such column: Landing__Outcome
[SQL: select * from SPACEXTBL where Landing__Outcome like 'Success%' and (DATE between '2010-06-04' and '2017-03-20') order by date desc]
(Background on this error at: http://sqlalche.me/e/e3q8)
```

Section 3

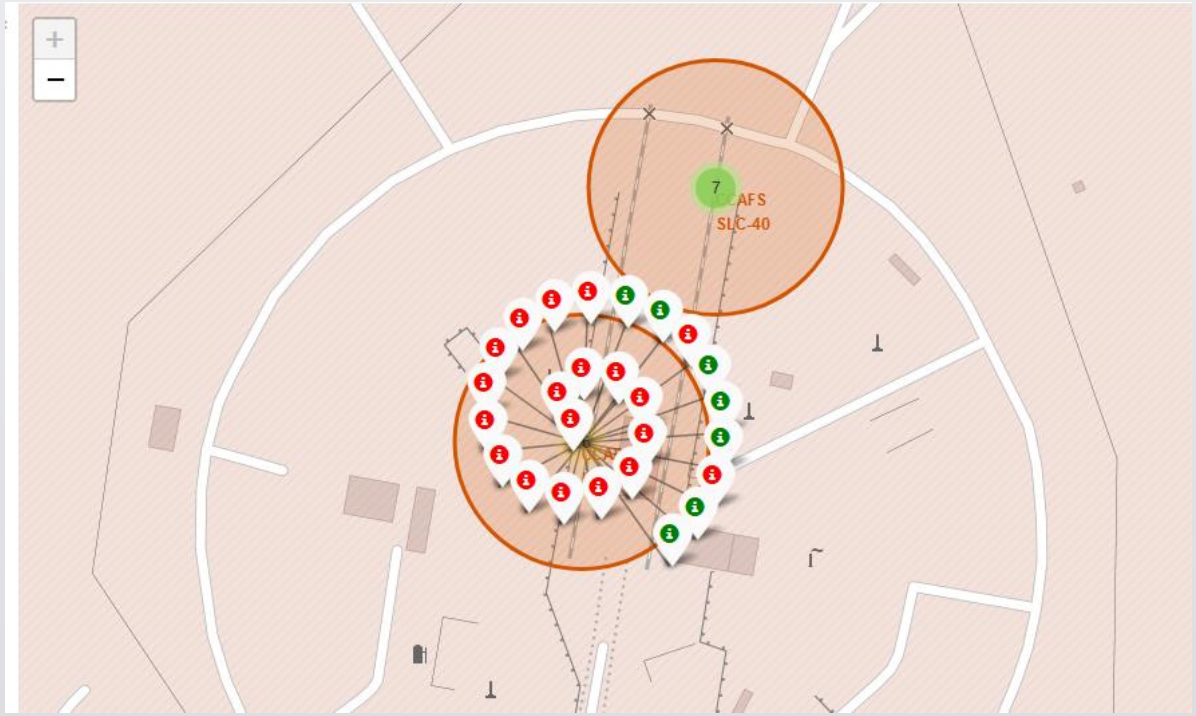
Launch Sites Proximities Analysis

Launch Site Locations Analysis with Folium

- All launch sites' location markers on a global map
- The launch sites of SpaceX are located in the USA near the east and west coast.

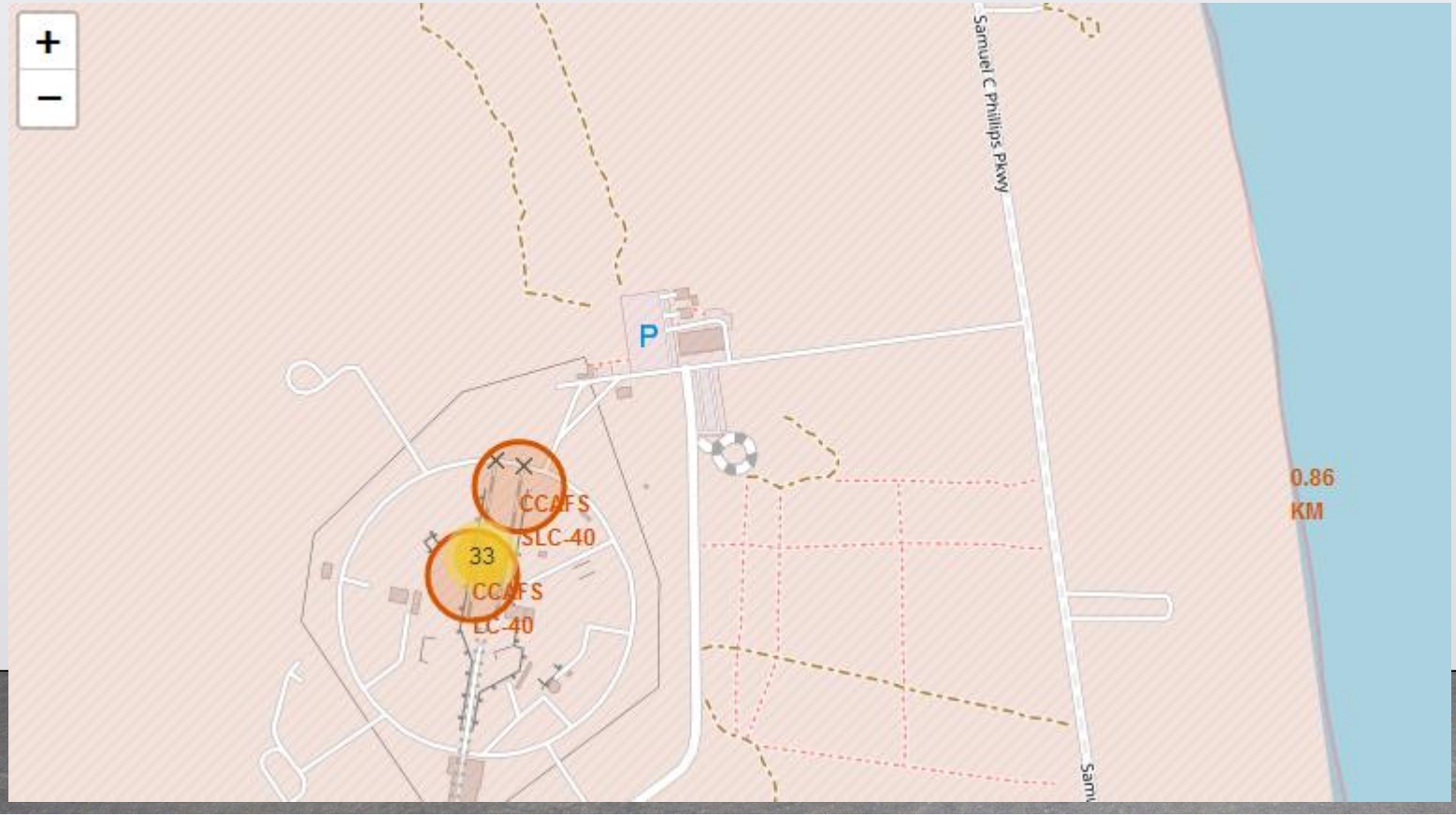
Launch Sites Folium map with marker cluster

- CCAFS LC-40 Launch Sites with green markers as successful launches and red markers as failed launches



Folium map of coastline distance launch site

- Distance of launch Site CCAFS SLC-40 to the east coastline (0.86 km)

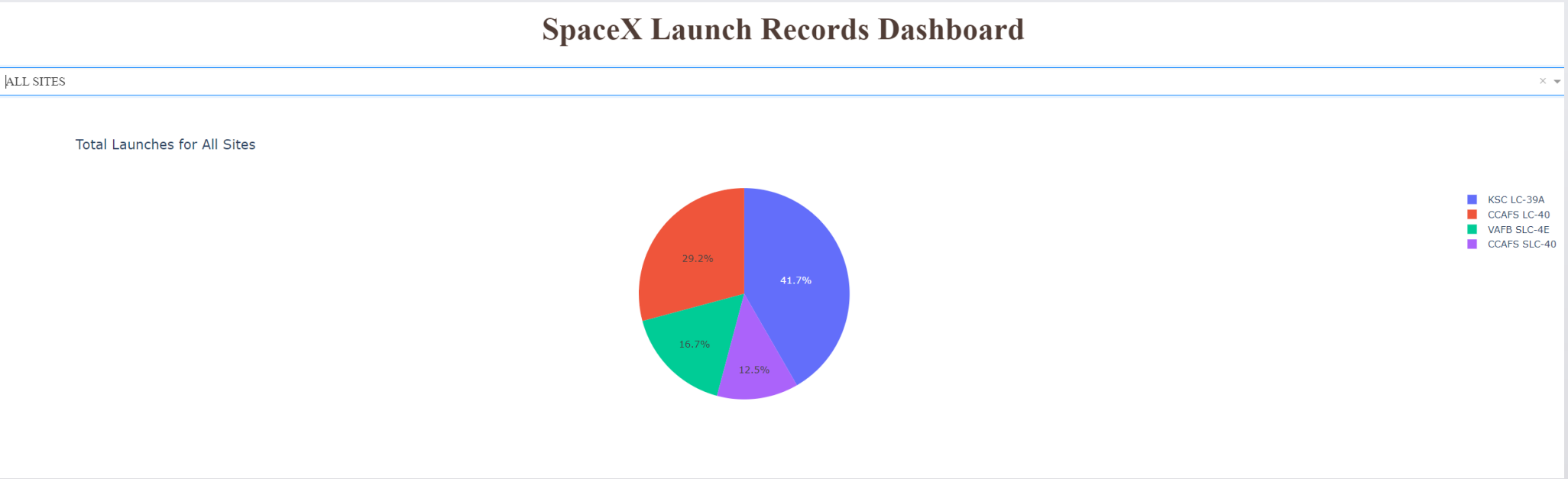



Section 4
Build a Dashboard with Plotly Dash

Launch success count for all sites, in a piechart

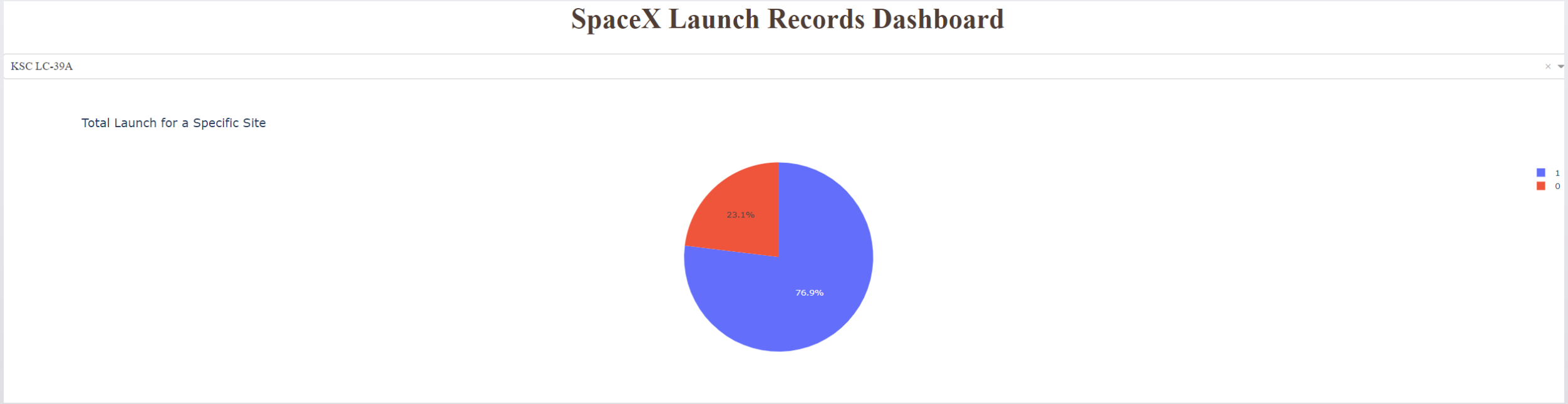
Highest launch success: KSC LC-39A (41.7%)

Lowest launch success: CCAFS SLC-40 (12.5%)

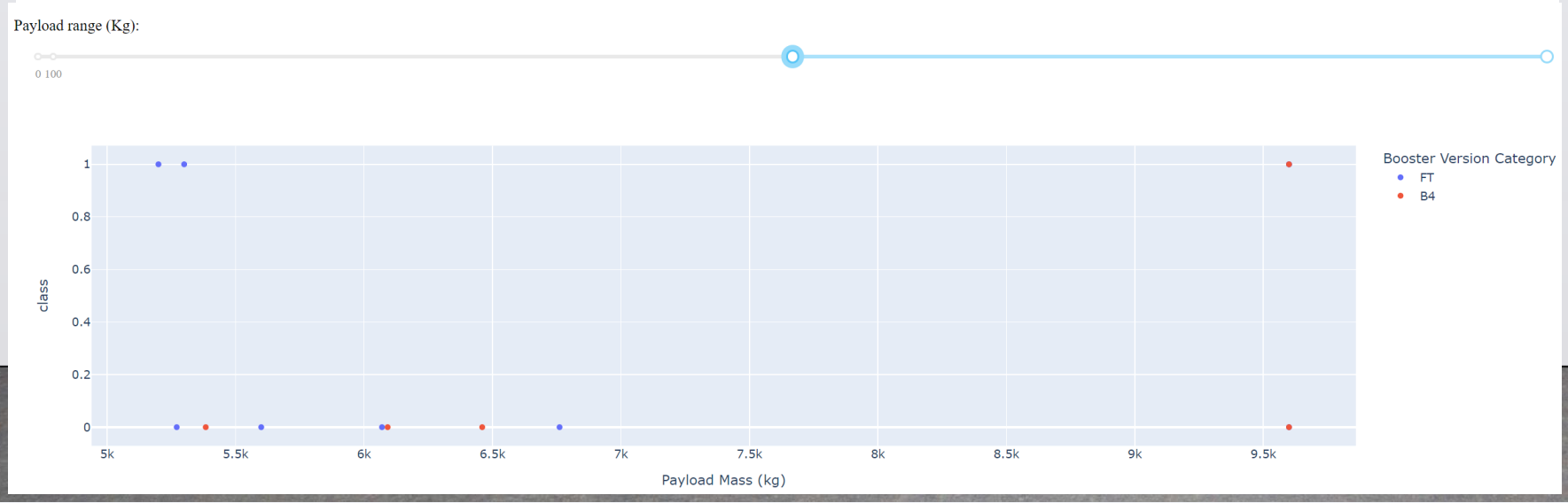


Piechart for the launch site with highest launch success ratio

Highest launch success ratio for KSC-LC-39A (76.9 %)



Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider

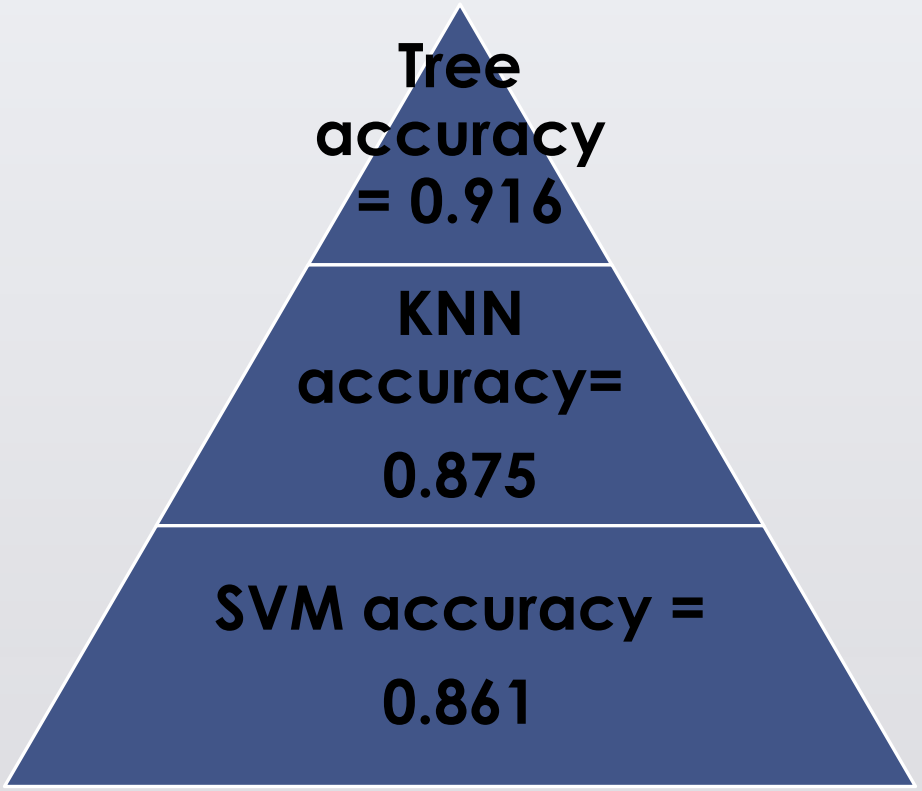


- Payload with lower weight between 2000 and 4000 have higher success rates then payload higher than 5000 kg

Section 5

Predictive Analysis (Classification)

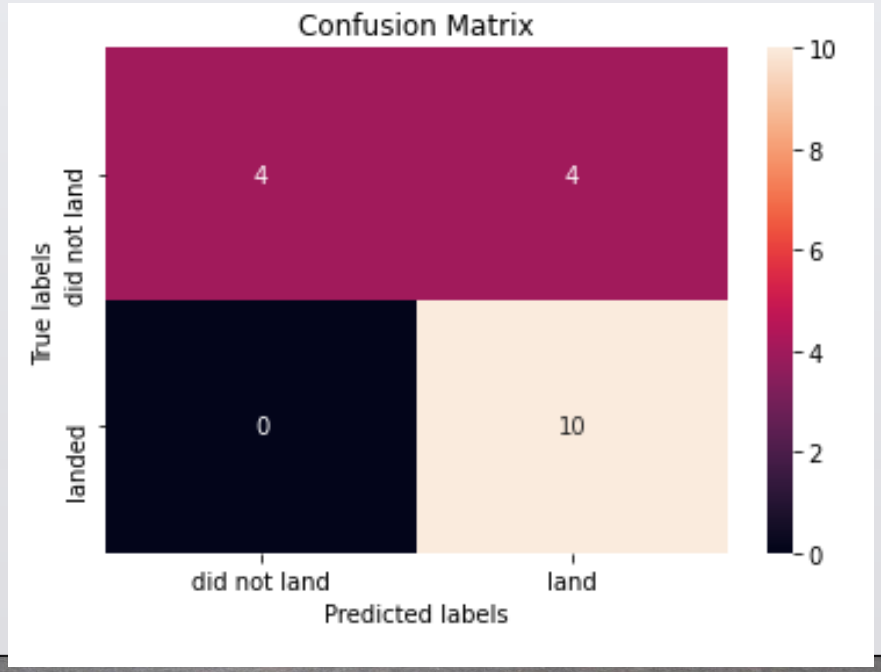
Classification Accuracy



The Tree alghorithm has the higher accuracy.

Confusion Matrix

- Confusion matrix for Tree.
- Tree can distinguish correctly between classes (beige area) but there are also false positive (red, upper right)



Conclusions

- The Falcon 9 first stage will land more successfully
  - By an increase of launches/flights (for example CCAFS SLC 4)
  - Lower weight payloads
- Launches from KSC LC-39A are the most succesful
- The Orbits ES-L1, SSO, HEO and GEO have the highest success rates
- Since 2013 (until) 2020 there has been an increase in the success rate
- It’s best to use the Tree Algorithm for classification since it has the highest accuracy

Appendix

- All files on Github: https://github.com/Gohan61/capstone10DSP/tree/master

Thank you!