IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Gargi Chaudhuri

4/28/2023

Outline

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

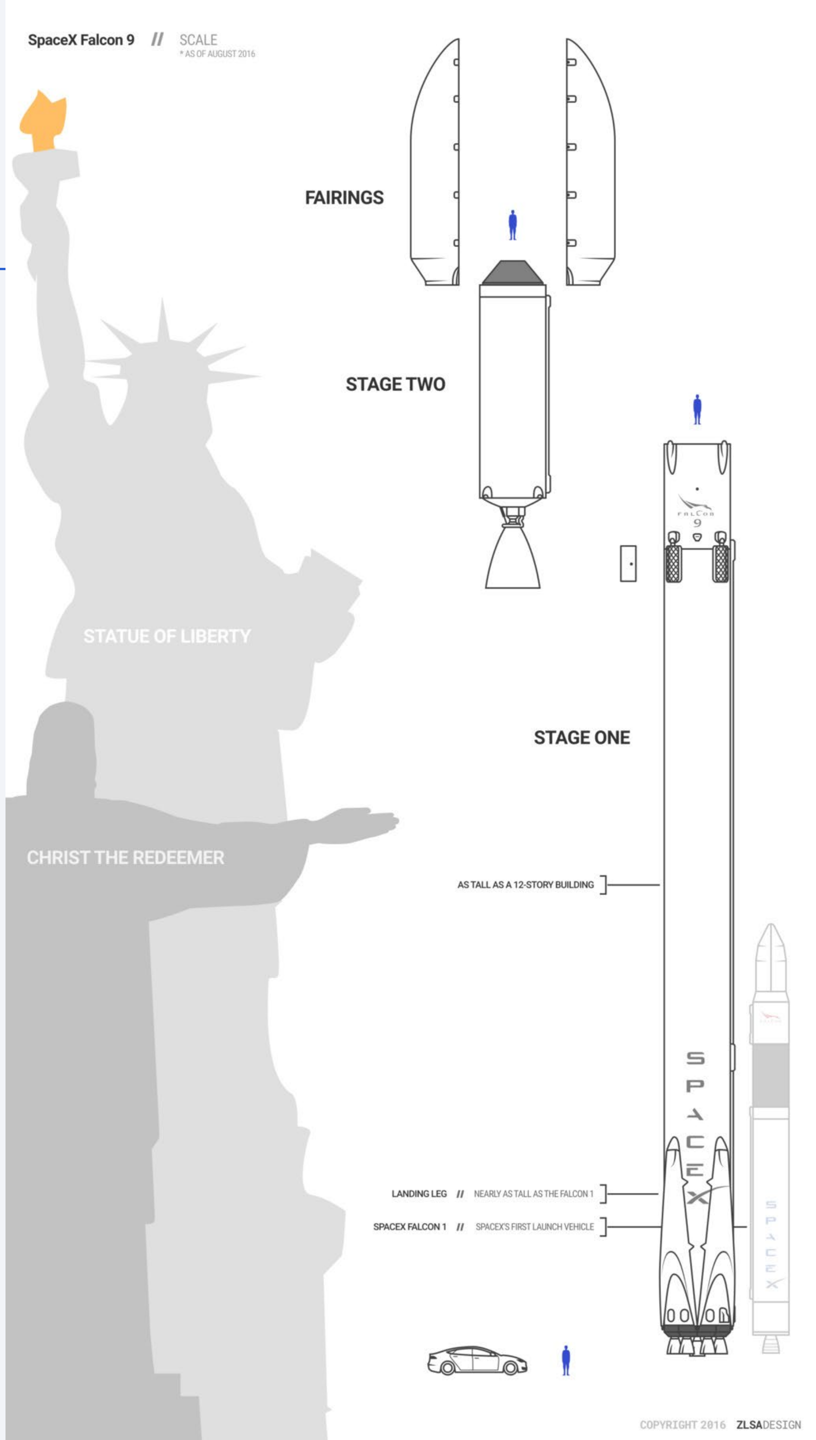
Executive Summary

- Summary of methodologies
  - Data Collection with API
  - Data Collection with Web Scraping lab
  - Data Wrangling
  - Exploratory Data Analysis (EDA) with Data Visualization
  - EDA with SQL
  - Interactive Visual Analytics with Folium
  - Interactive Dashboard with Plotly Dash
  - Machine Learning Prediction

- Summary of all results
  - As the flight number increases, the first stage is more likely to land successfully.
  - The more massive the payload, the less likely the first stage will return.
  - KSC LC-39A and VAFB SLC 4E has relatively higher success rate than CCAFS LC-40
  - The VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000).
  - With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
  - The orbits HEO, SSO, ES-L1, and GEO have the higher success rates and PO has none.
  - The success rate since 2013 kept increasing till 2020.
  - Decision tree classifier has the best classification accuracy (90%)

Introduction

- Project background and context
  - The commercial space age is here, companies are making space travel affordable for everyone.
  - Virgin Galactic is providing suborbital spaceflights. Rocket Lab is a small satellite provider. Blue Origin manufactures sub-orbital and orbital reusable rockets.
  - 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.
  - 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. To help us understand the scale of the Falcon 9, we are used these diagrams from Forest Katsch, at zlsadesign.com (on the right).
    - The payload is enclosed in the fairings. Stage two, or the second stage, helps bring the payload to orbit, but most of the work is done by the first stage. The first stage is shown here. This stage does most of the work and is much larger than the second stage. Here we see the first stage next to a person and several other landmarks. This stage is quite large and expensive. Unlike other rocket providers, SpaceX's Falcon 9 Can recover the first stage. Sometimes the first stage does not land. Sometimes it will crash as shown in this clip. Other times, Space X will sacrifice the first stage due to the mission parameters like payload, orbit, and customer.
- Problems you want to find answers
  - Space Y that would like to compete with SpaceX. Therefore, the goals of the capstone project are to:
    - To determine the price of each launch.
    - To predict if the Falcon 9 first stage will land successfully and will reuse the first stage
  - This information can be used if an alternate company wants to bid against SpaceX for a rocket launch (Coursera project website)



Section 1

Methodology

Methodology: Executive Summary

- Data collection methodology:
  - We collected data with SpaceX REST API and by webscraping from Wikipedia with BeautifulSoup. The data preparation process involved: first parsing the data into dataframe format under the collect columns; secondly, filtering the dataframe just to show Falcon 9 information; and finally, weeding out missing data and saving it in the CSV format.

- Perform data wrangling
  - In this section Exploratory Data Analysis (EDA) was performed to find some patterns in the data and determine what would be the label for training supervised models. In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident. Therefore, in this lab those outcomes were converted into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful to run predictive analysis in the next section

Methodology: Executive Summary …continues

- Perform exploratory data analysis (EDA) using visualization and SQL
    - In this assignment some exploratory data analysis was performed before running the machine learning based predictive model. Using scatterplot, correlations between Flight Number and PayLoadMass, Flight Number and Launch Site, Payload and Launch Site, and FlightNumber and Orbit type, all color coded with outcome of the launch were evaluated. Using a bar graph, the success rate of each orbit were compared, and line graph was used to show the yearly trend of launch success.
    - Using SQL, we queried the names of the unique launch sites, 5 records where launch sites begin with the string 'CCA', the total payload mass carried by boosters launched by NASA (CRS), average payload mass carried by booster version F9 v1.1, the date when the first successful landing outcome in ground pad was achieved, the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000, the total number of successful and failure mission outcomes, the names of the booster versions which have carried the maximum payload mass, the records which will display the month names, failure landing outcomes in drone ship, booster versions, launch site for the months in year 2015, and finally, the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 ranked in descending order.

Methodology: Executive Summary …continues

- Perform interactive visual analytics using Folium and Plotly Dash
  - Using folium, first, all the launch site circle markers were created to show the locations of the launch site. Secondly, the circle markers were colored green and red based on their success and failure, respectively. Finally, the distance from nearest coastline, highway, railway line and city was calculated and visualized using line and labelled texts.
  - In the interactive dashboard for Plotly Dash, we have used piegraph which shows the proportion of success and failures of all launch sites. The scatterplot shows the correlation between success/failure and payload mass (kg). The dropdown menu allows us to choose among each of the four launch sites and all launch sites. The range slider can filter the amount of payload mass in the scatterplot. The graphical view and interactivity helps to communicate the detailed information quickly.
- Perform predictive analysis using classification models
  - A machine learning pipeline was created to predict if the first stage will land given the data. Four machine learning models were tested: Logistic Regression (LogReg), Support Vector Machine (SVM), Decision Tree classifier (Tree_CV), and K-Nearest Neighbor (KNN) to train and fit the model. For each of the models, an optimal set of hyperparameters were tested and used for prediction. The best classification model was found by comparing the training and testing accuracy of all the models.

Data Collection

- We collected data with SpaceX REST API and by webscraping from Wikipedia with BeautifulSoup, about Booster name, launch site (with lat/long), mass of payload and the orbit that its going to, outcome of the landing, the type of the landing, number of flights with that core, whether grid fins were used, whether the core is reused, whether legs were used, the landing pad used, the block of the core which is a number used to separate version of cores, the number of times this specific core has been reused, and the serial of the core.

- The data preparation process involved: first parsing the data into dataframe format under the collect columns; secondly, filtering the dataframe just to show Falcon 9 information; and finally, weeding out missing data and saving it in the CSV format.

Get request to SpaceX REST API

Collected Data on

- Flight Number, Date of Launch, Rocket type (Booster Version), Payload mass, Orbit, Launch site, outcome, number of lights, grid fins used?, core is reused?, legs were used?, the landing pad used?, the block of the core which is a number used to separate version of cores, the number of times this specific core has been reused, and the serial of the core.

Data formatted and filtered

Data saved as CSV

Webscraping from Wikipedia using BoutifulSoup

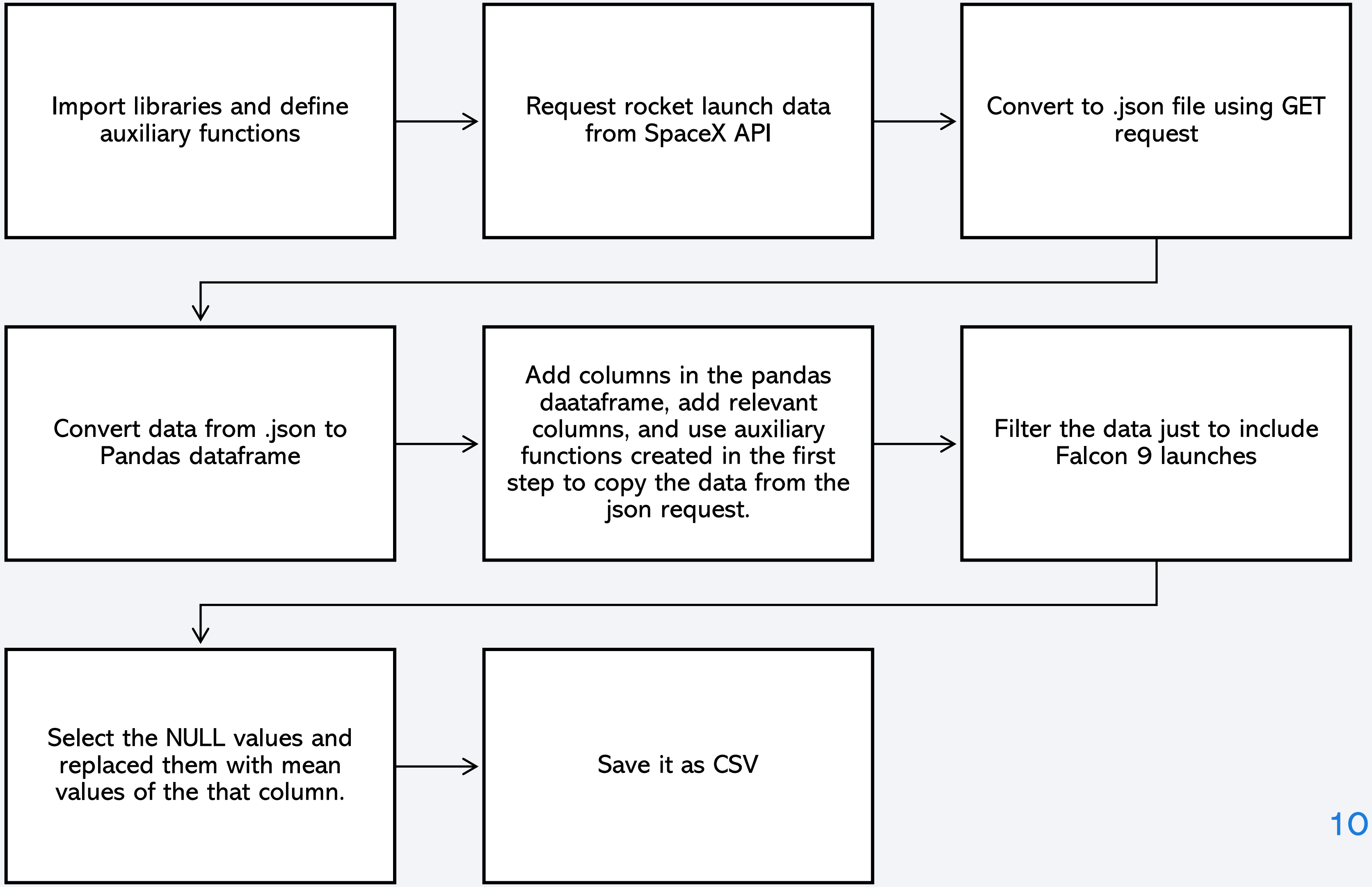
Collected Data on Falcon 9 historical launch records from a Wikipedia page titled List of Falcon 9 and Falcon Heavy launches

Data formatted and filtered

Data saved as CSV

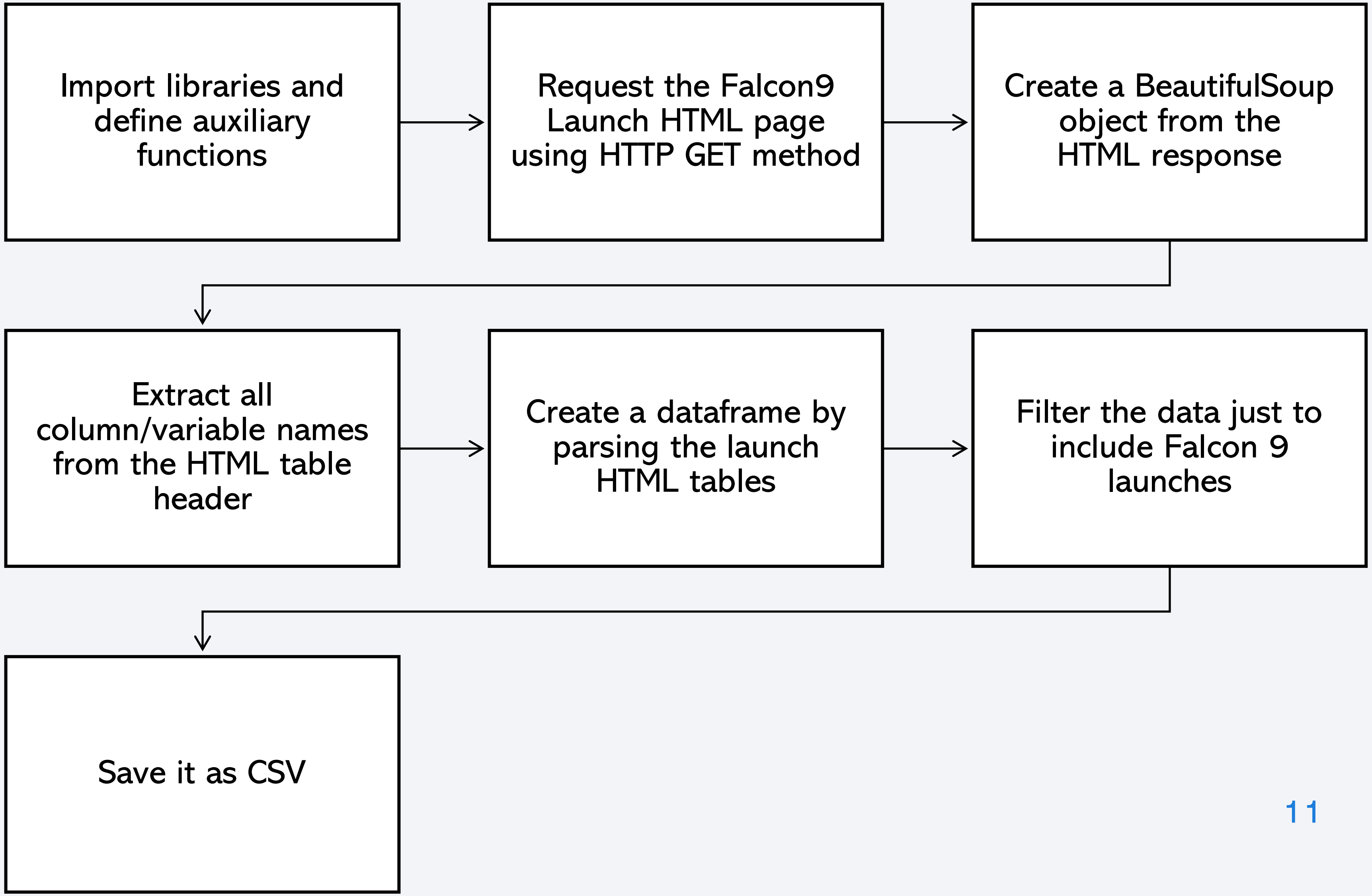
Data Collection – SpaceX API

- Github Link



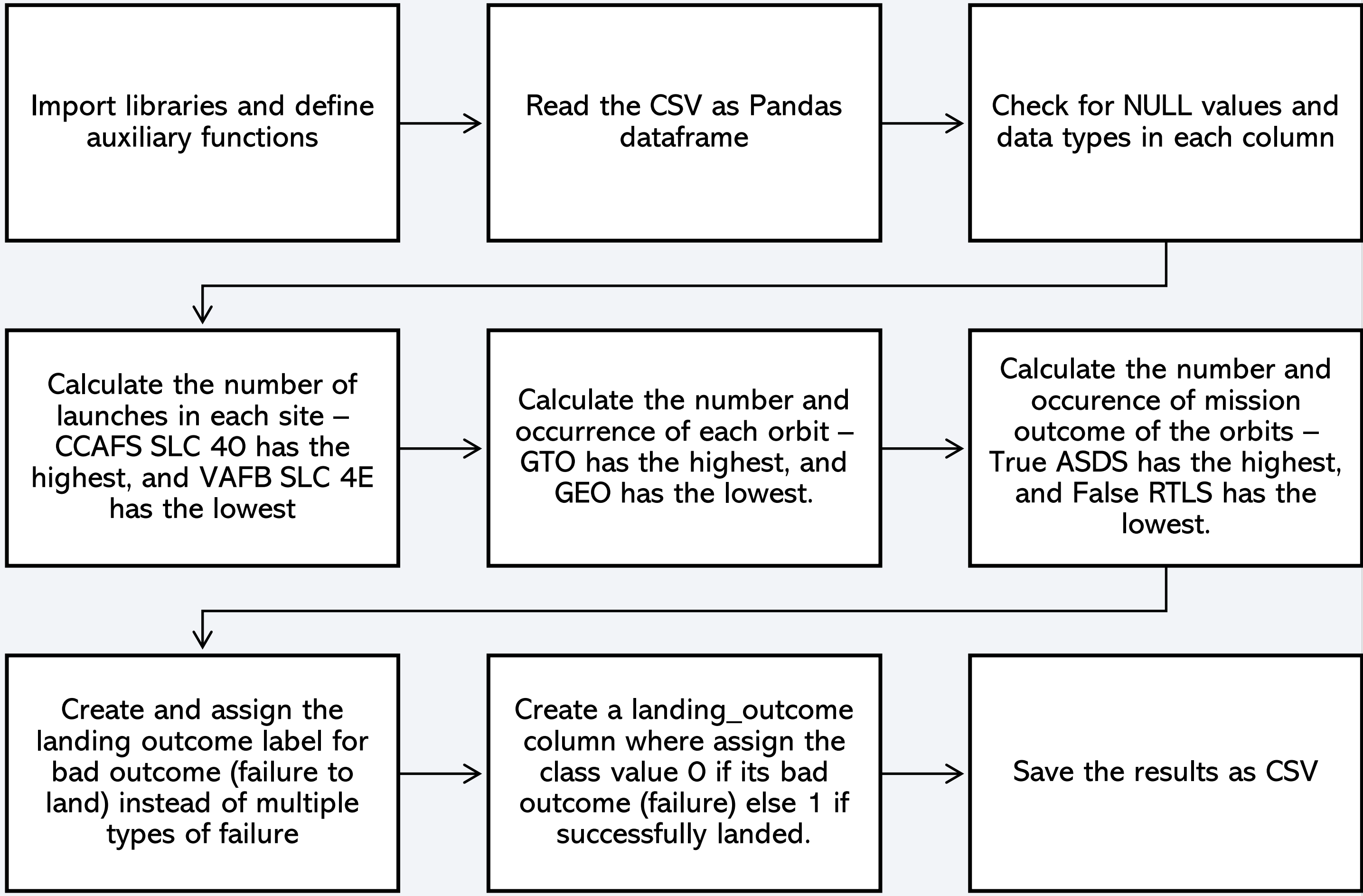
Data Collection - Scraping

- GitHub URL



Data Wrangling

- In this section Exploratory Data Analysis (EDA) was performed to find some patterns in the data and determine what would be the label for training supervised models.
- In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example:
  - 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.
  - True RTLS means the mission outcome was successfully landed to a ground pa. False RTLS means the mission outcome was unsuccessfully landed to a 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.
- Therefore, in this lab those outcomes were converted into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful. The flowchart describes the data processing steps
- [GitHub URL]



EDA with Data Visualization

- In this assignment some exploratory data analysis was performed before running the machine learning based predictive model
- Using scatterplot, the following correlations were evaluated:
  - Flight Number and PayLoadMass color coded with outcome of the launch. We see that as the flight number increases, the first stage is more likely to land successfully. The payload mass is also important; it seems the more massive the payload, the less likely the first stage will return.
  - Flight Number and Launch Site color coded with outcome of the launch. We see that KSC LC-39A and VAFB SLC 4E has relatively higher success rate than CCAFS LC-40
  - Payload and Launch Site color coded with outcome of the launch. We see that the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).
  - FlightNumber and Orbit type color coded with outcome of the launch. We see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.
- Using a bar graph, the success rate of each orbit were compared. HEO, SSO, ES-L1, and GEO have the higher success rates and PO has none.
- Using a line graph, the yearly trend of launch success was evaluated. We see that that the success rate since 2013 kept increasing till 2020.
- GitHub URL

EDA with SQL

- SQL queries performed
  1. Display the names of the unique launch sites in the space mission
  2. Display 5 records where launch sites begin with the string 'CCA’
  3. Display the total payload mass carried by boosters launched by NASA (CRS)
  4. Display average payload mass carried by booster version F9 v1.1
  5. List the date when the first successful landing outcome in ground pad was achieved.
  6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  7. List the total number of successful and failure mission outcomes
  8. List the names of the booster_versions which have carried the maximum payload mass.
  9. 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.
  10. Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

- GitHub URL

Build an Interactive Map with Folium

- 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.
- First, all the launch site circle markers were created to show the locations of the launch site. Secondly, the circle markers were colored green and red based on their success and failure, respectively. Finally, the distance from nearest coastline, highway, railway line and city was calculated and visualized using line and labelled texts.
- GitHub URL

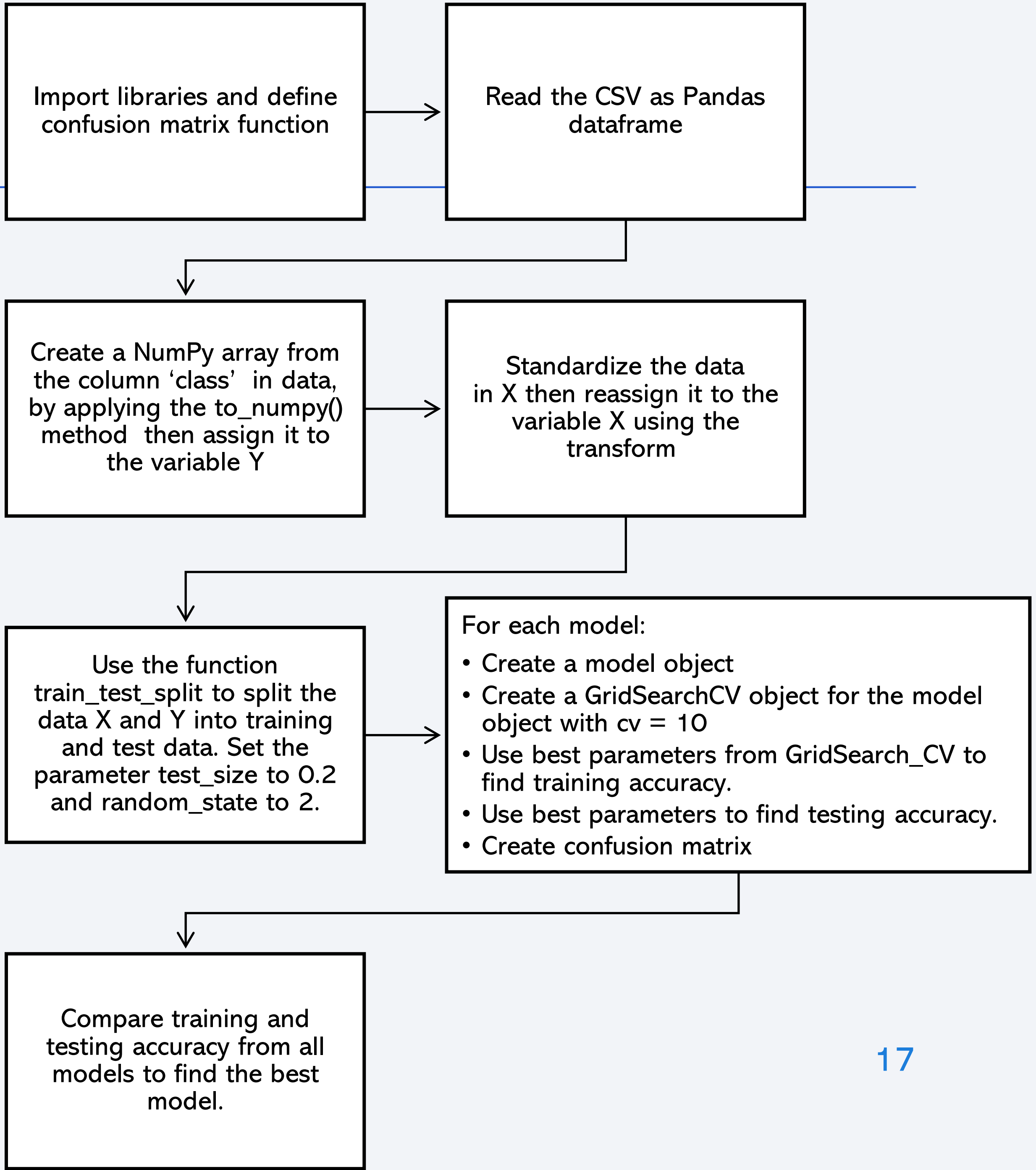
Build a Dashboard with Plotly Dash

- In the interactive dashboard for Plotly Dash, we have used piegraph which shows the proportion of success and failures of all launch sites. The scatterplot shows the correlation between success/failure and payload mass (kg).
- The dropdown menu allows us to choose among each of the four launch sites and all launch sites. The range slider can filter the amount of payload mass in the scatterplot. The graphical view and interactivity helps to communicate the detailed information quickly.
- GitHub URL

Predictive Analysis (Classification)

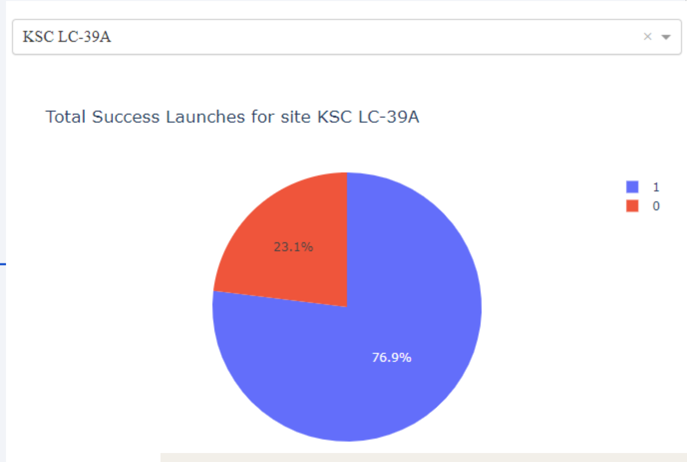
- In this lab, a machine learning pipeline was created to predict if the first stage will land given the data. First, the data was downloaded and formatted as a pandas dataframe. The ‘class’ column was used as dependent variable and rest of the factors were independent variables. Four machine learning models were tested: Logistic Regression (LogReg), Support Vector Machine (SVM), Decision Tree classifier (Tree_CV), and K-Nearest Neighbor (KNN) to train and fit the model. For each of the models, an optimal set of hyperparameters were tested and used for prediction. The best classification model was found by comparing the training and testing accuracy of all the models in a bargraph.

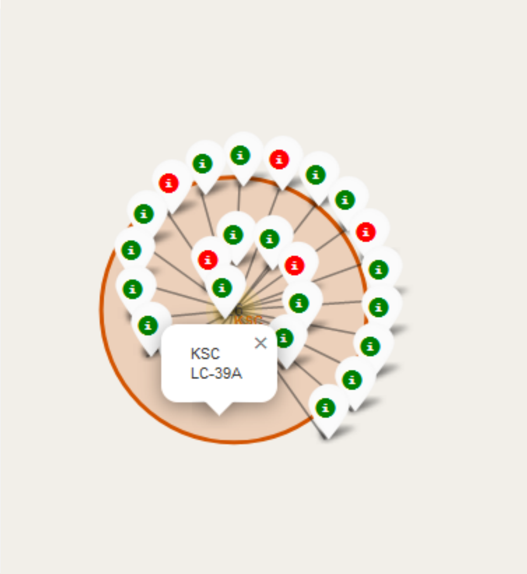
- GitHub URL

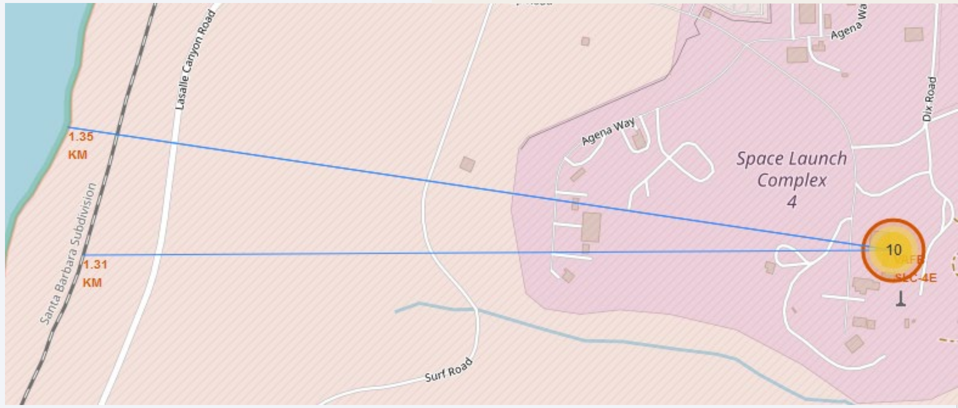


Results

- Exploratory data analysis results
  - As the flight number increases, the first stage is more likely to land successfully.
  - The payload mass is also important; it seems the more massive the payload, the less likely the first stage will return.
  - KSC LC-39A and VAFB SLC 4E has relatively higher success rate than CCAFS LC-40
  - The VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000).
  - With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
  - The orbits HEO, SSO, ES-L1, and GEO have the higher success rates and PO has none.
  - The success rate since 2013 kept increasing till 2020.
- Interactive analytics helped to explore the data more in depth.
- Predictive analysis results
  - Decision tree classifier has the best classification accuracy (90%)



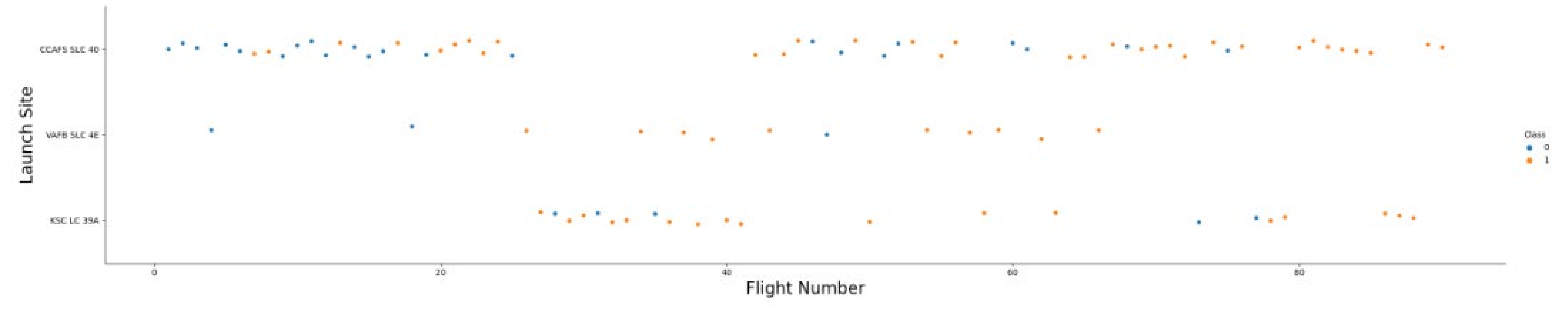




Section 2

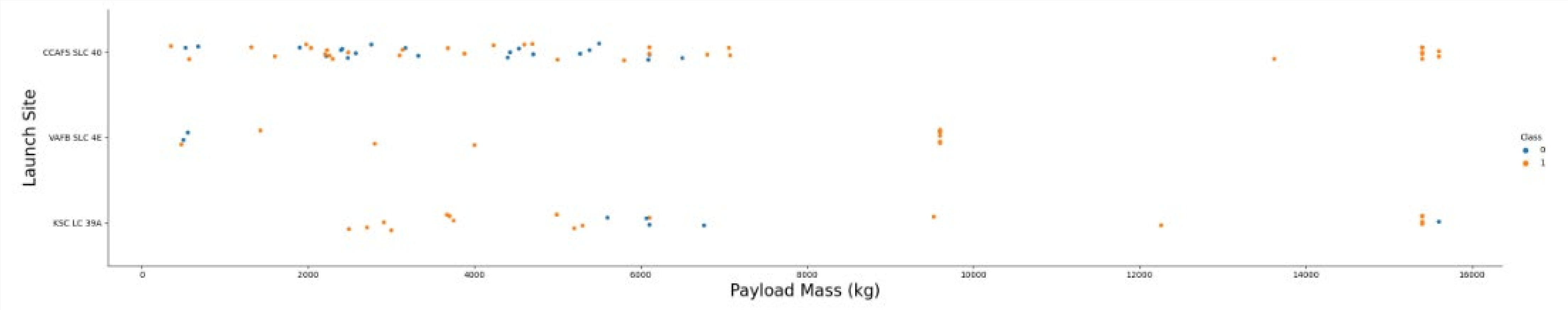
Insights drawn from EDA

Flight Number vs. Launch Site



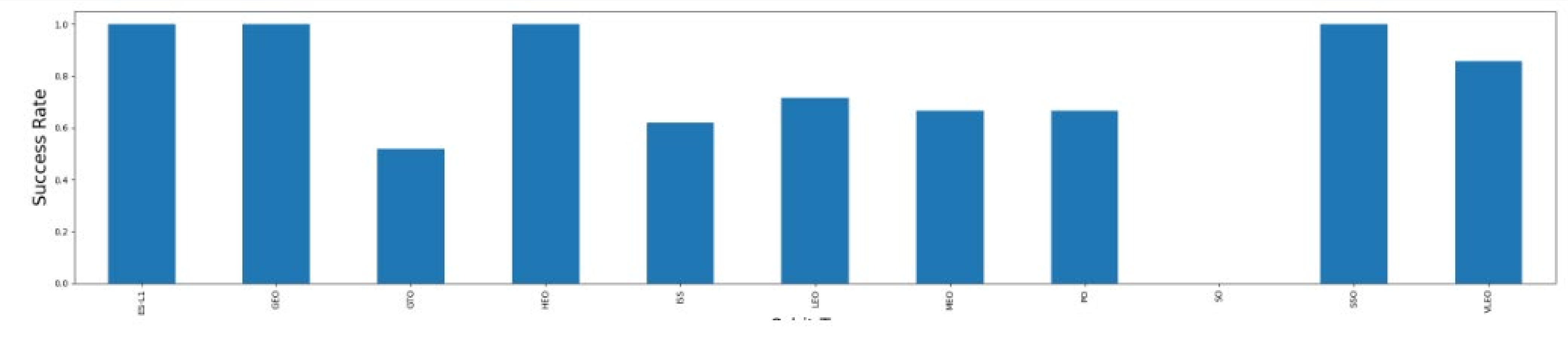
Flight Number and Launch Site color coded with outcome of the launch. We see that KSC LC-39A and VAFB SLC 4E has relatively higher success rate than CCAFS LC-40.

Payload vs. Launch Site



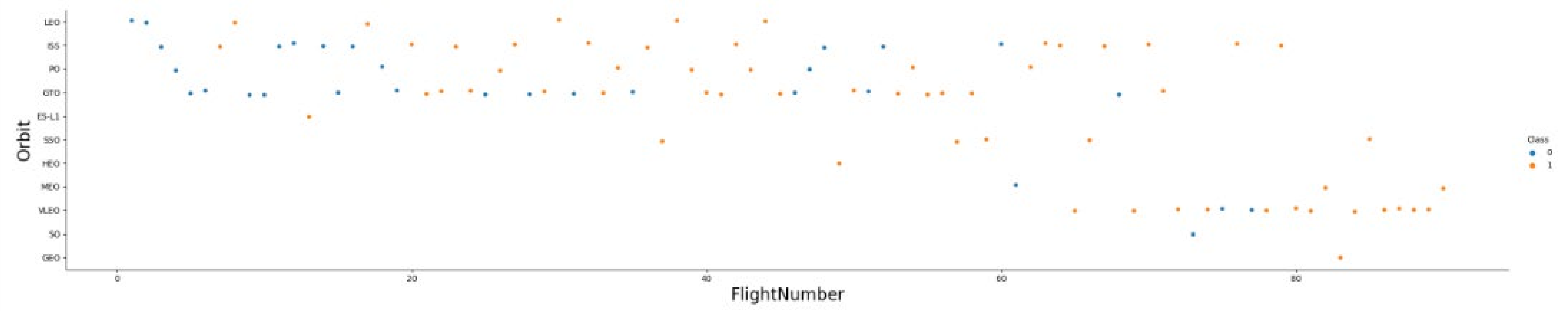
Payload and Launch Site color coded with outcome of the launch. We see that the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).

Success Rate vs. Orbit Type



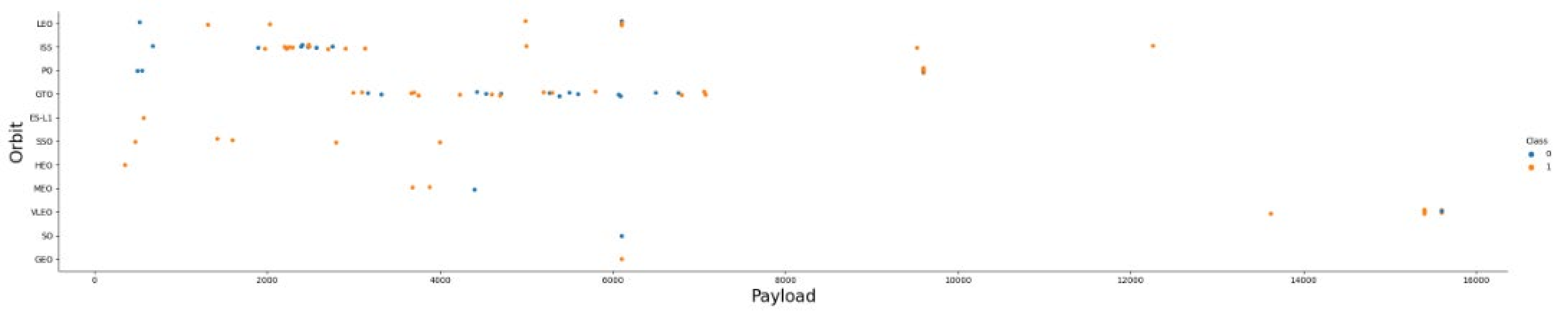
Bar graph showed the success rate of each orbit. HEO, SSO, ES-L1, and GEO have the higher success rates and PO has none.

Flight Number vs. Orbit Type



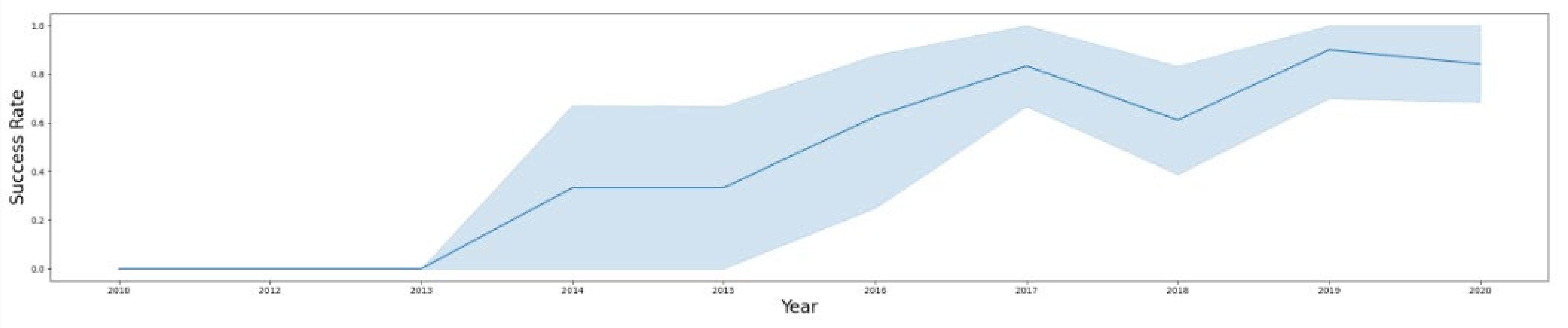
We see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



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

Launch Success Yearly Trend



We see that the overall success rate since 2013 kept increasing till 2020.

All Launch Site Names

- Unique launch site names listed

Task 1

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

In [10]:
```
%sql SELECT DISTINCT "LAUNCH_SITE" FROM SPACEXTBL
```

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

Out[10]:

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

Launch Site Names Begin with 'CCA'

Task 2

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

In [12]:
```
%sql SELECT * FROM SPACEXTBL WHERE "LAUNCH_SITE" LIKE '%CCA%' LIMIT 5
```

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

Out[12]:

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

5 records where launch sites begin with `CCA` listed here

Total Payload Mass

The total payload carried by boosters from NASA is listed here.

Task 3

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

```
In [13]: %sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "CUSTOMER" = 'NASA (CRS)'

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

Out[13]:

SUM("PAYLOAD_MASS__KG_")
45596

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1 is listed here.

Task 4

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

In [14]:
```
%sql SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "BOOSTER_VERSION" LIKE '%F9 v1.1%'
```

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

Out[14]:

AVG("PAYLOAD_MASS__KG_")
2534.6666666666665

First Successful Ground Landing Date

The date of the first successful landing outcome on ground pad is listed here.

Task 5

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

Hint:Use min function

[15]:
```
%sql SELECT MIN("DATE") FROM SPACEXTBL WHERE "Landing _Outcome" LIKE '%Success%'
```

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

[15]:

MIN("DATE")
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 is listed here.

Task 6

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

n [16]:

```
%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
Done.
```

ut[16]:

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

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

17]:

```
%sql SELECT (SELECT COUNT("MISSION_OUTCOME") FROM SPACEXTBL WHERE "MISSION_OUTCOME" LIKE '%Success%') AS SUCCESS, \
(SELECT COUNT("MISSION_OUTCOME") FROM SPACEXTBL WHERE "MISSION_OUTCOME" LIKE '%Failure%') AS FAILURE
```

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

17]:

SUCCESS	FAILURE
100	1

The total number of successful and failure mission outcomes is listed here.

Boosters Carried Maximum Payload

The names of the booster which have carried the maximum payload mass is listed here.

Task 8

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

In [18]:

```
%sql SELECT DISTINCT "BOOSTER_VERSION" FROM SPACEXTBL \
WHERE "PAYLOAD_MASS__KG_" = (SELECT max("PAYLOAD_MASS__KG_") FROM SPACEXTBL)
```

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

Out[18]:

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

Task 9

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

```
%sql SELECT substr("DATE", 4, 2) AS MONTH, "BOOSTER_VERSION", "LAUNCH_SITE" FROM SPACEXTBL\
WHERE "LANDING _OUTCOME" = 'Failure (drone ship)' and substr("DATE",7,4) = '2015'
```

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

MONTH	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

The failed landing outcomes in drone ship, their booster versions, launch site name, and month in the year 2015 is listed here.

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

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

Task 10

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

```
[20]:
%sql SELECT "LANDING _OUTCOME", COUNT("LANDING _OUTCOME") FROM SPACEXTBL\
WHERE "DATE" >= '04-06-2010' and "DATE" <= '20-03-2017' and "LANDING _OUTCOME" LIKE '%Success%'\
GROUP BY "LANDING _OUTCOME" \
ORDER BY COUNT("LANDING _OUTCOME") DESC ;
```

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

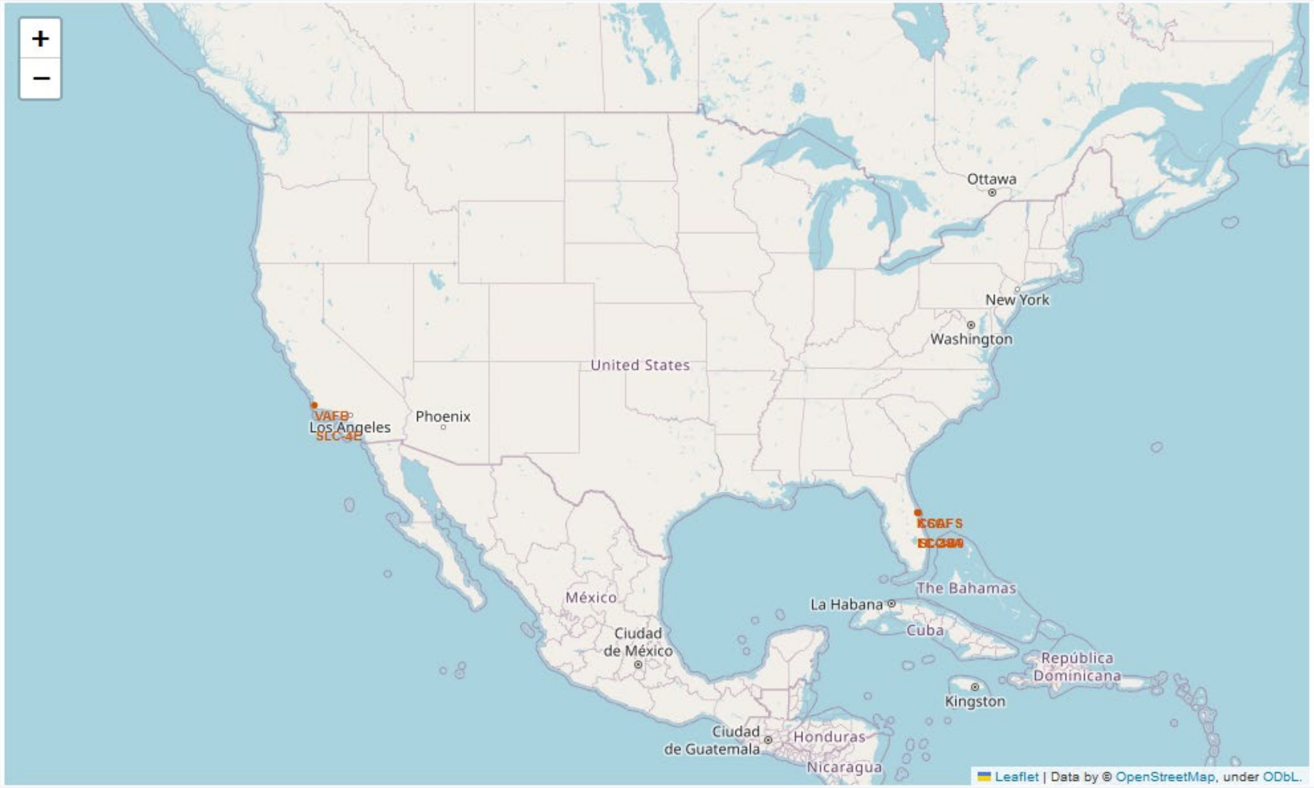
[20]:

Landing _Outcome	COUNT("LANDING _OUTCOME")
Success	20
Success (drone ship)	8
Success (ground pad)	6

Section 3

Launch Sites Proximities Analysis

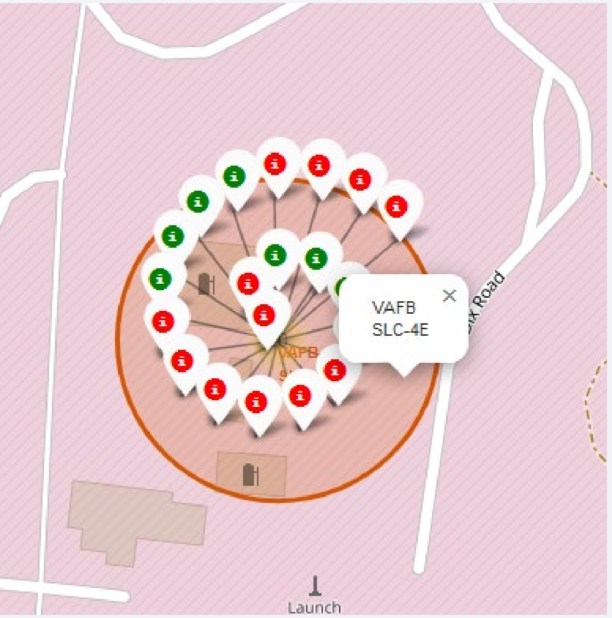
All launch sites on a map

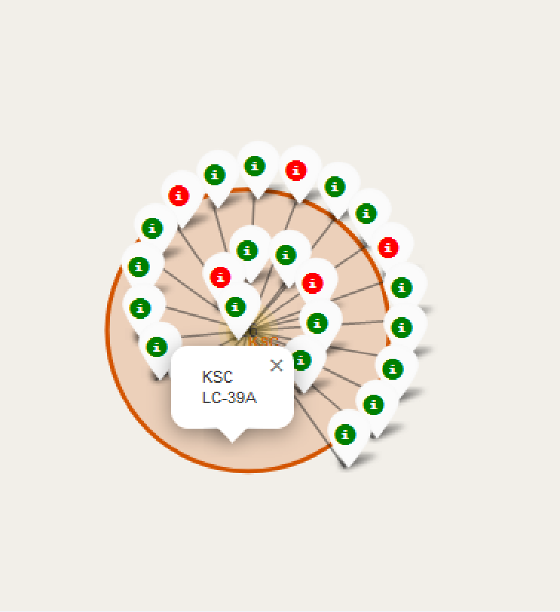


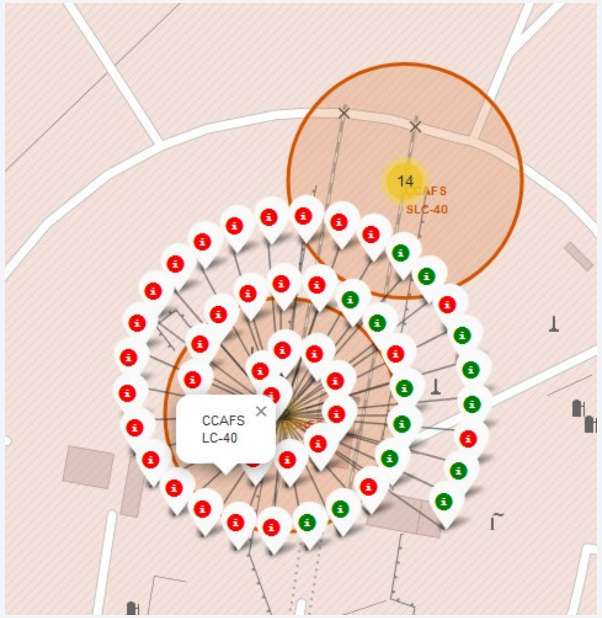
All the launch sites are along the coast, so that if there is a failed landing it can happen in the ocean for safety. The launch sites are far away from the equator.

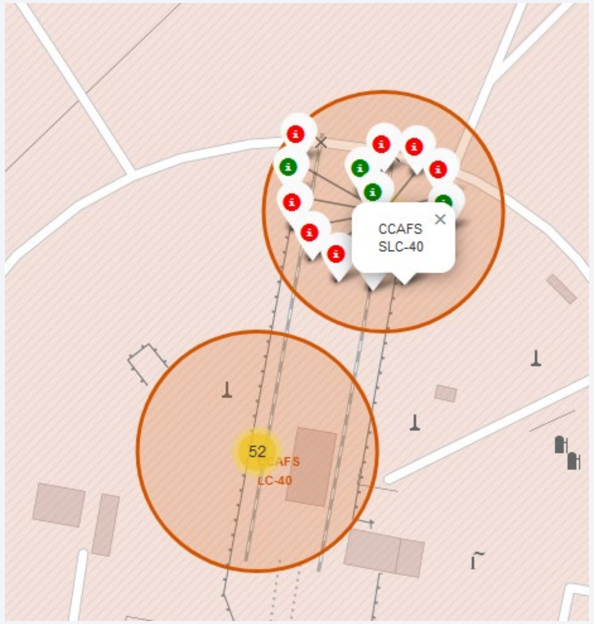
The successful and failed launches for each site on the map

To enhance the map the launch outcomes in each site has been labelled based on their success (green) and failure (red).



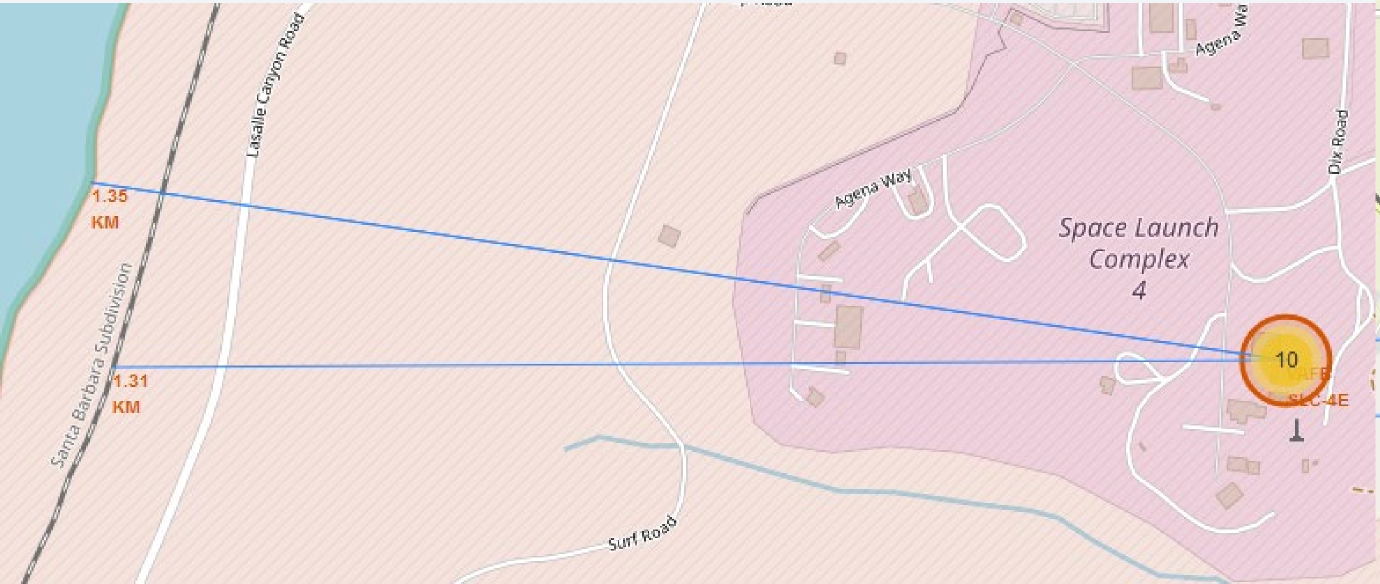


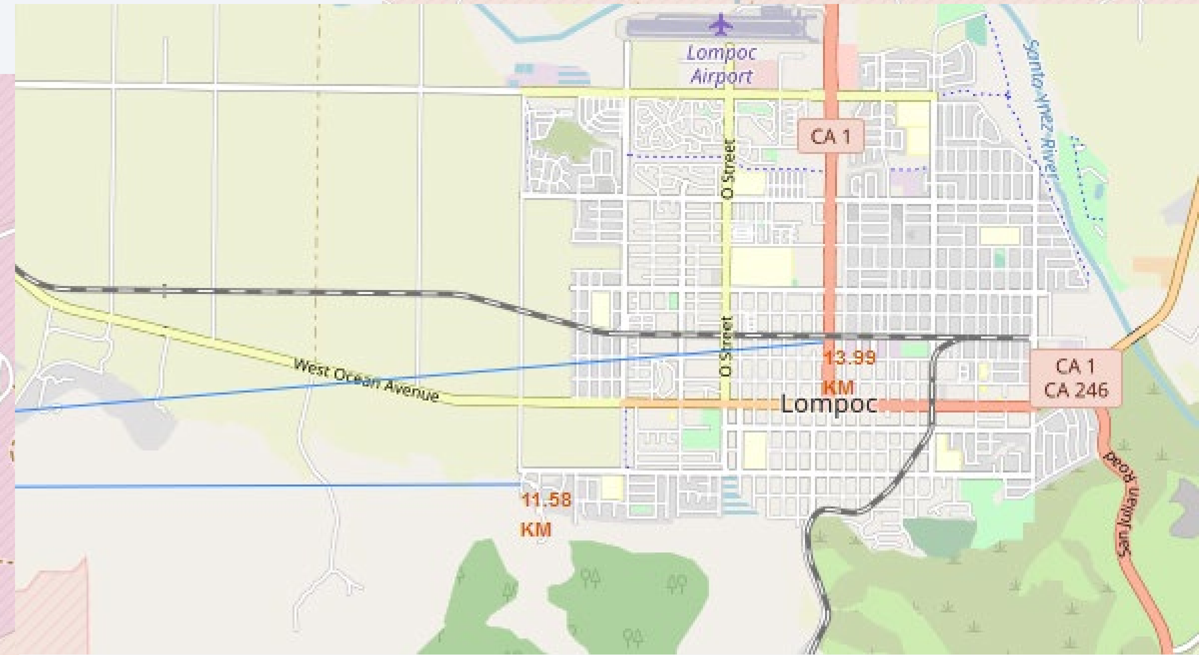




The distances between a launch site to its proximities

- The VAFB SLC-4E is 1.35KM from the coastline, 1.31 KM from the railway line. 5.59 KM away from major road, 11.58 KM from the edge of the Lompoc city, and 13.99KM from the city center.



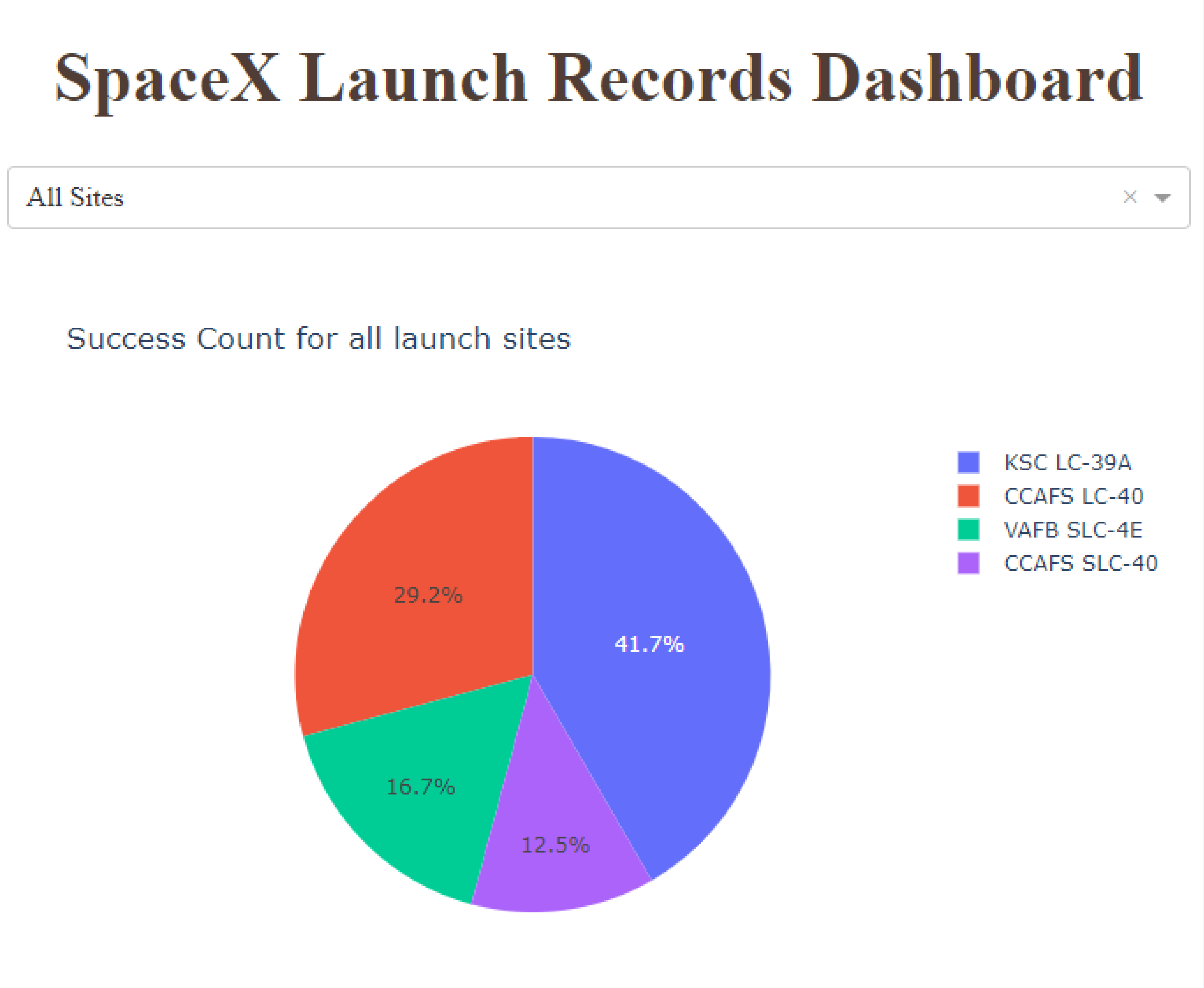


Section 4

Build a Dashboard with Plotly Dash

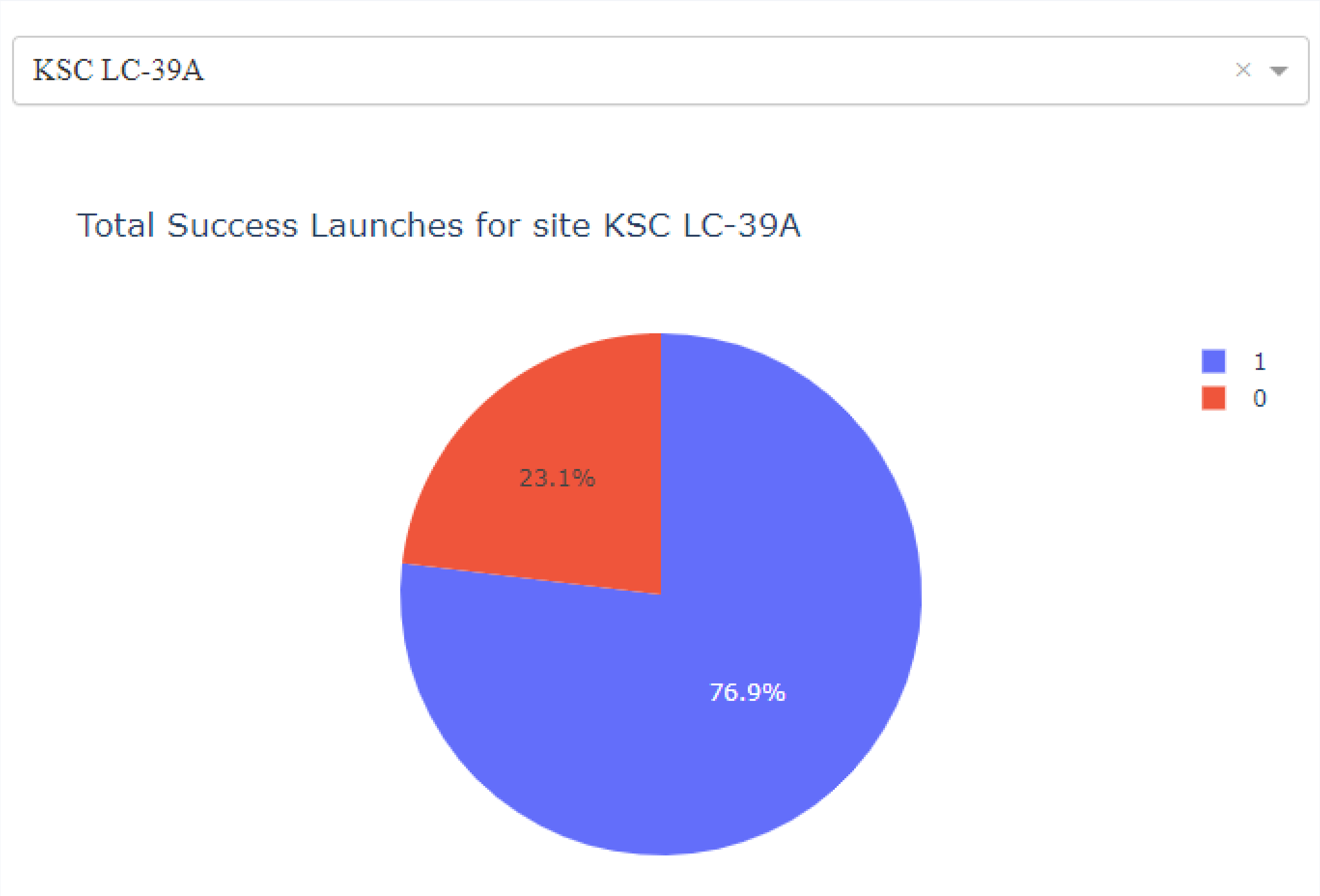
Piechart showing proportion of success for all launch sites

KSC LC-39A has the highest success rate whereas CCAFS SLC-40 has the lowest.

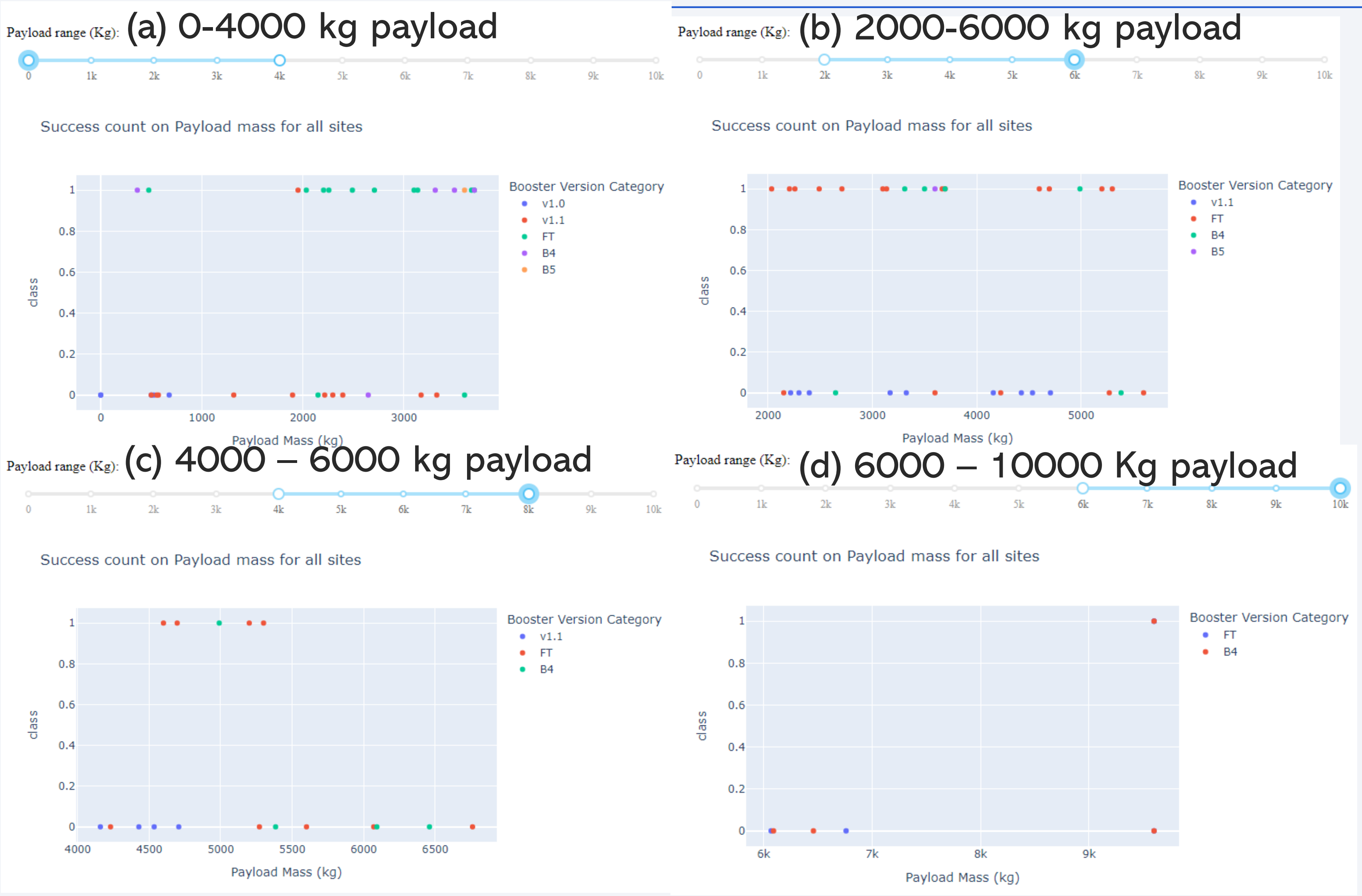


Piechart showing proportion of success and failure for the KSC LC-39A launch site

Among all the launch site KSC LC-39A has the highest success rate which is around 76.9%.



Scatterplot showing the relationship between payload and launch outcome



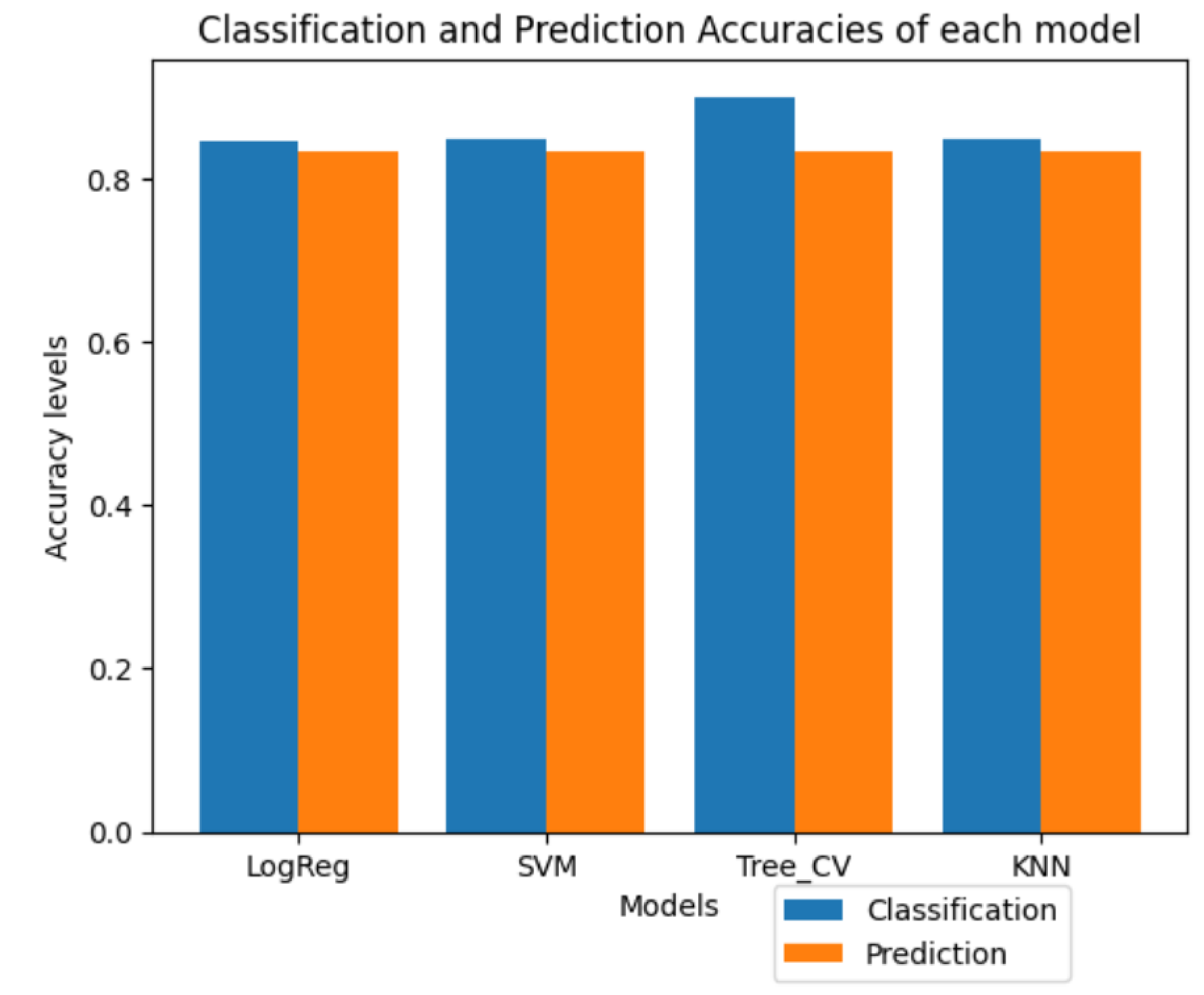
Scatterplot showing the relationship between payload and launch outcome. The payloads are filtered to show the relationship of success with payload between 0-4000 kg (a), 2000-4000 kg (b), 4000 – 6000 kg (c), and 6000 – 10000 kg (d). Lower payloads have higher success rates

Section 5

Predictive Analysis (Classification)

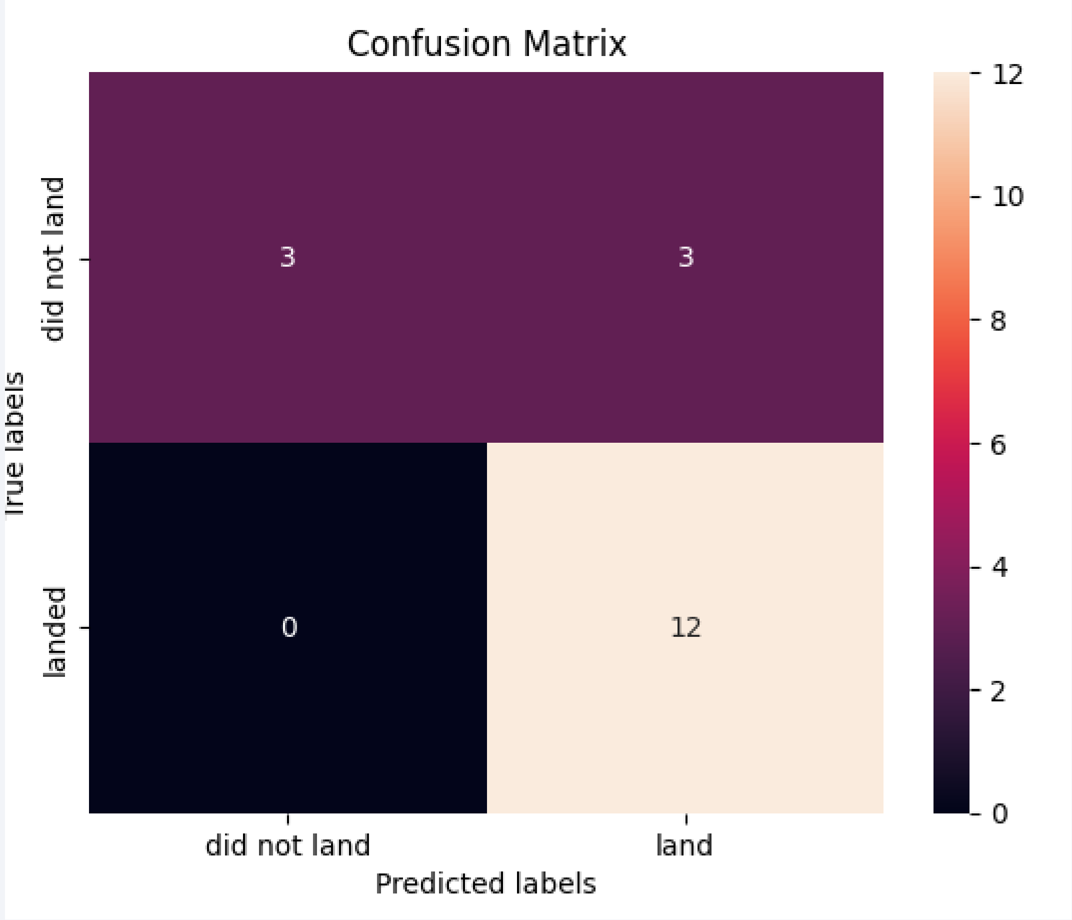
Classification Accuracy

Decision tree classifier has the highest classification accuracy among all the four models. The prediction accuracy remains same for all of them.



Confusion Matrix

The confusion matrix shows that mission success were correctly predicted for 12 data points (True Positive) and failure for 3 (True Negative) data points. The model only failed for 3 data points (False Positive) where it predicted as landed but in reality, it failed.



Conclusions

- KSC LC-39A and VAFB SLC 4E launch sites have relatively higher success rate than CCAFS LC-40 launch site.
- Heavier payloads leads to more failure. If heavy payloads succeed in landing, its mostly in the Polar, LEO and ISS orbit.
- Overall, the orbits HEO, SSO, ES-L1, and GEO have the higher success rates and PO has none.
- More flights in a launch site results in better success rates. The success rate since 2013 kept increasing till 2020.
- The Decision tree classifier has the highest classification accuracy. The prediction accuracy is same for all machine learning models tested in this project.

Appendix

- Bargraph to compare the different model training and testing accuracy
- Different code fixing due to either package version change or incorrect given code.

Thank you!