IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Suman C
09/09/2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - Exploratory Data Analysis with SQL and Visualization
  - Building Interactive Maps & Dashboards
  - Predictive Analysis
- Summary of all results
  - Exploratory Data Analysis results
  - Interactive Maps and Dashboards Screenshots
  - Predictive Analysis Results

Introduction

- Project background and context

  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.

- Problems you want to find answers
  - What are the best variables invloves in the success of the first stage landing?
  - What is the best algorithm that can be used for binary classification in this project

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data collected from SpaceX api and web scraping from wikipedia
- Perform data wrangling
  - Data was enhanced by creating a landing outcome label based on outcome data
- 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

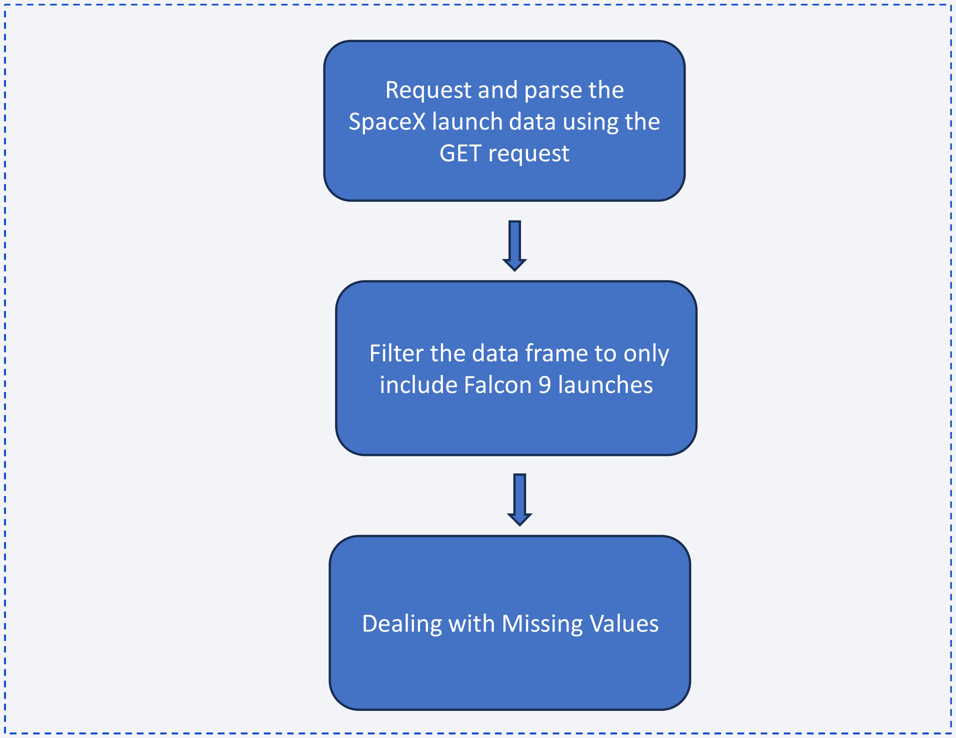
- Data collection was done using get request to the SpaceX API.
- Cleaned the data, checked for missing values and fill in missing values where necessary.
- Performed web scraping from Wikipedia for Falcon 9 launch records with BeautifulSoup

Data Collection – SpaceX API

- Data is obtained from spaceX public api.

- Notebook Link:

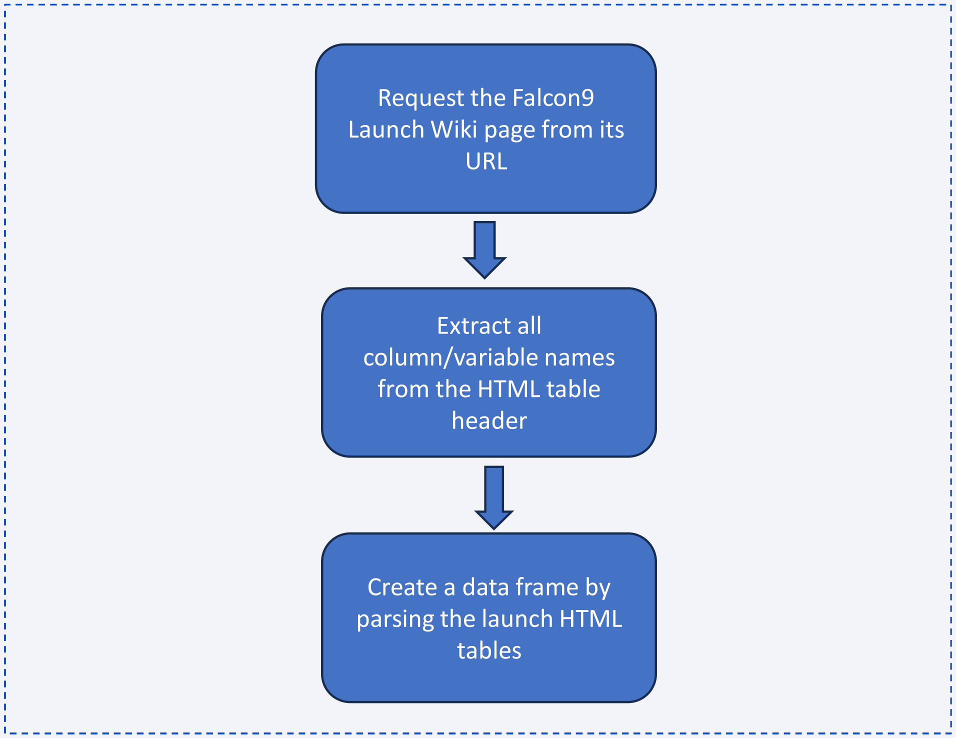
https://github.com/chintapatlasumanrao/ds-proj/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Extracted a Falcon 9 launch records HTML table from Wikipedia
- Parsed the table and converted it into a pandas dataframe
- GitHub notebook url:

  https://github.com/chintapatlasumanrao/ds-proj/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Performed exploratory Data Analysis and determined Training Labels
  - Calculated the number of launches on each site
  - Calculated the number and occurrence of each orbit
  - Calculated the number and occurrence of mission outcome per orbit type
  - Created a landing outcome label from Outcome column

- Github notebook url:
  https://github.com/chintapatlasumanrao/ds-proj/blob/main/labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

- Performed exploratory Data Analysis and Feature Engineering using Pandas and Matplotlib
  - To show the relationship between variables, used scatter plots
  - Bar charts are used for showing comparisons among discrete categories
  - Line charts are used to display trends in data over time
- GitHub URL of completed EDA with data visualization notebook:

  https://github.com/chintapatlasumanrao/ds-proj/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- Performed SQL queries
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first succesful landing outcome in ground pad was acheived.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
  - 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.
  - 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.
- GitHub URL of completed EDA with SQL notebook:

  https://github.com/chintapatlasumanrao/ds-proj/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Marked all launch sites, and added map objects such as markers, circles, lines to mark the success or failure of launches for each site on the folium map.
- Using the color-labeled marker clusters, identified which launch sites have relatively high success rate.
- Calculated the distances between a launch site to its proximities
- GitHub URL of completed interactive map with Folium map:

  https://github.com/chintapatlasumanrao/ds-proj/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- Built an interactive dashboard with Plotly dash
- Added a dropdown list to enable Launch Site selection.
- Plotted pie charts showing the total launches by a certain sites
- Added a slider to select Payload range.
- Added a scatter chart to show the correlation between Payload and Launch Success
- GitHub URL of completed Plotly Dash lab:

  https://github.com/chintapatlasumanrao/ds-proj/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Loaded using numpy and pandas, transformed the data, split our data into training and testing.
- Created different machine learning models and tune different hyperparameters using GridSearchCV.
- Used accuracy as the metric for our model, improved the model using feature engineering and algorithm tuning.
- Figured out best performing classification model.
- GitHub URL of your completed predictive analysis lab

  https://github.com/chintapatlasumanrao/ds-proj/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

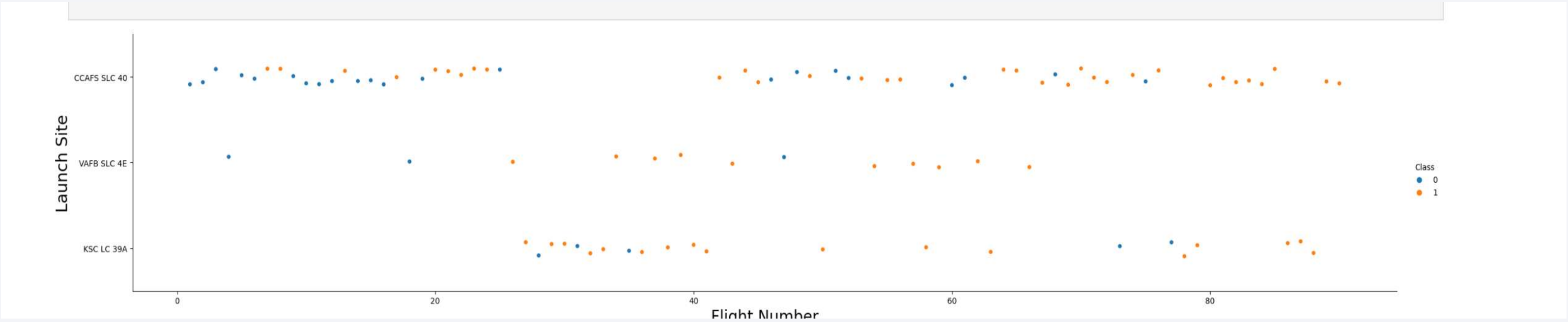
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

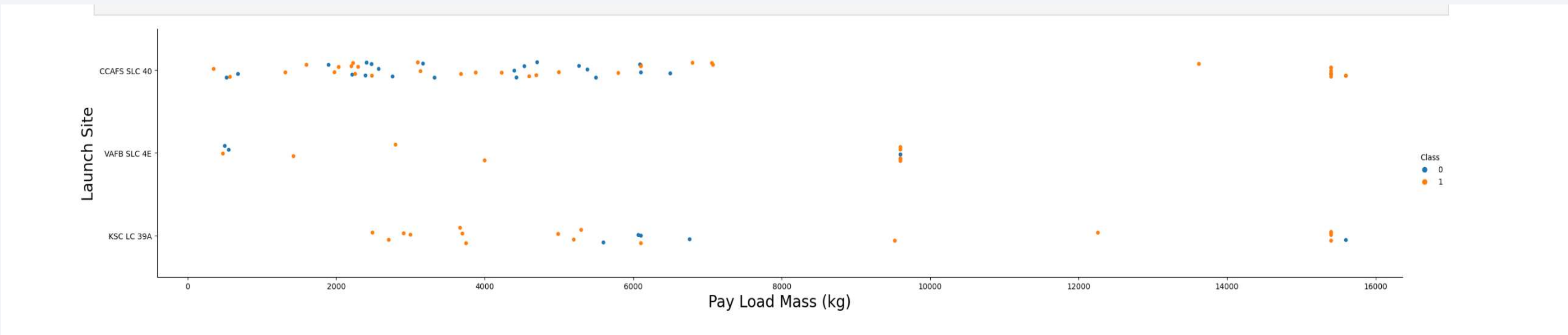
Flight Number vs. Launch Site

- Show a scatter plot of Flight Number vs. Launch Site



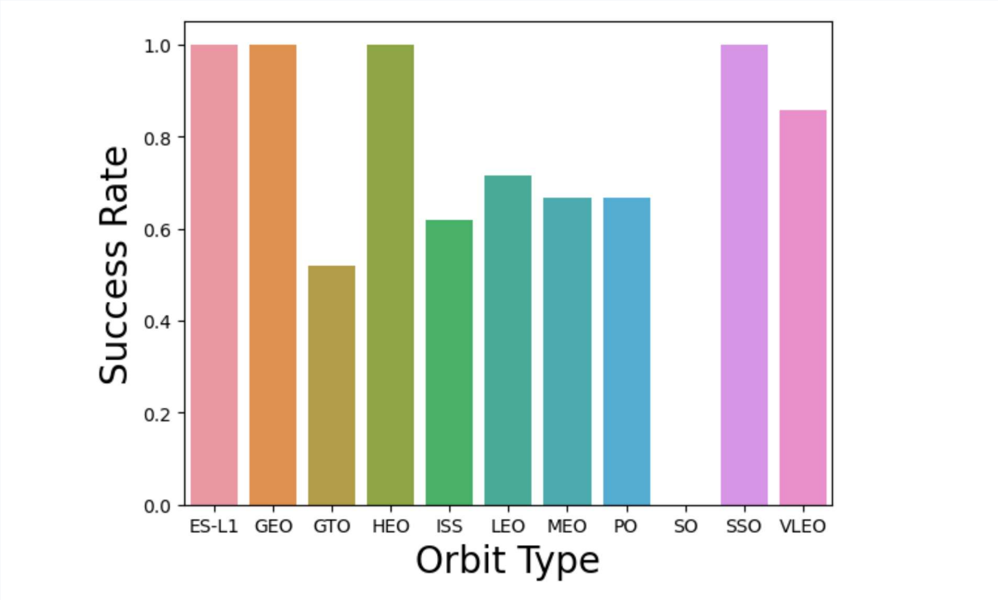
Payload vs. Launch Site

- Show a scatter plot of Payload vs. Launch Site



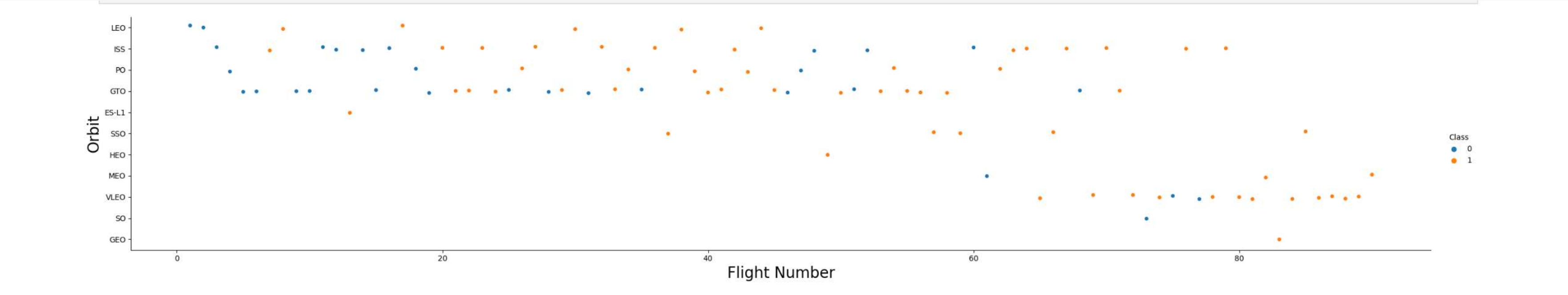
Success Rate vs. Orbit Type

- Show a bar chart for the success rate of each orbit type



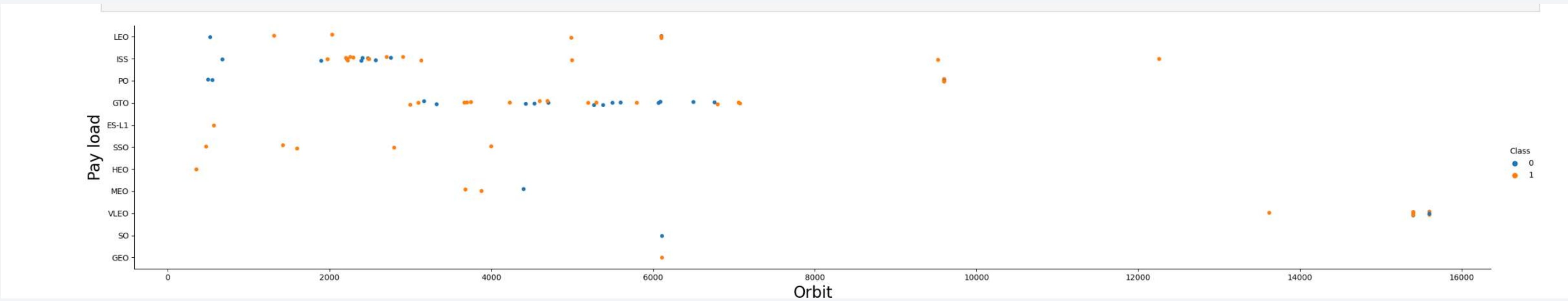
Flight Number vs. Orbit Type

- Show a scatter point of Flight number vs. Orbit type



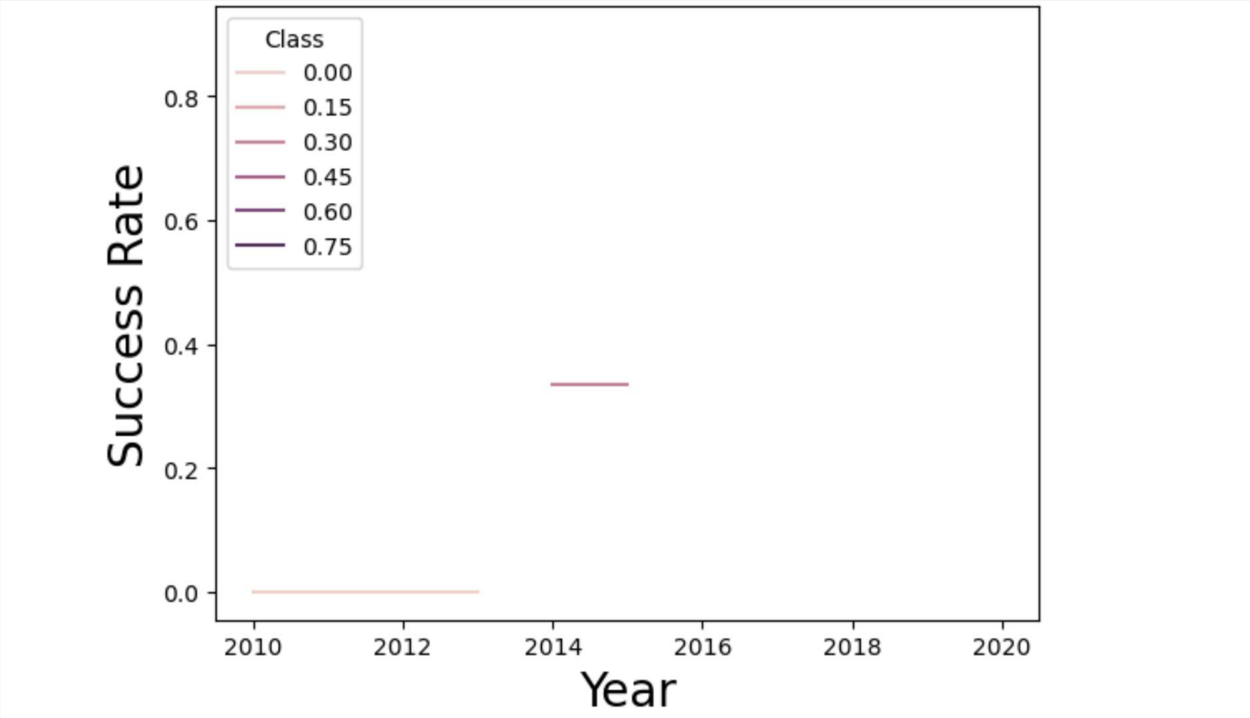
Payload vs. Orbit Type

- Show a scatter point of payload vs. orbit type



Launch Success Yearly Trend

- Show a line chart of yearly average success rate



All Launch Site Names

- Find the names of the unique launch sites
- Used distinct keyword to find the unique launch sites

```
In [14]: %sql select distinct  "Launch_Site" from SPACEXTABLE

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

Out[14]:

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Used LIKE operator and LIMIT for Find 5 records where launch sites begin with `CCA`

In [15]:
```
%sql select * from SPACEXTABLE where "Launch_Site" LIKE 'CCA%' LIMIT 5
```

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

Out[15]:

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	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- Used SUM function to get the total payload

```
In [19]: %sql select sum("PAYLOAD_MASS__KG_") as "Total Payload" from SPACEXTABLE where "Customer" = 'NASA (CRS)'

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

Out[19]:

Total Payload
45596

Task 4

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Used AVG function to get the total average payload mass

```
In [20]: %sql select avg("PAYLOAD_MASS__KG_") as "Total Payload" from SPACEXTABLE where "Booster_Version" = 'F9 v1.1'

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

Out[20]:

Total Payload
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Min function used to get the first successful landing outcome on ground pad

In [28]:
```
%sql select min("Date") from SPACEXTABLE where "Landing_Outcome" = 'Success (ground pad)'
```

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

Out[28]:

min("Date")
2015-12-22

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 600

In [41]:
```
%%sql select distinct "Booster_Version" from SPACEXTABLE where "Landing_Outcome" = 'Success (drone ship)'
AND "PAYLOAD_MASS__KG_" between 4000 and 6000
```

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

Out[41]:

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

In [50]:
```
%sql select substr("Mission_Outcome", 1, 7), count(*)  from SPACEXTABLE group by 1
```

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

Out[50]:

substr("Mission_Outcome", 1, 7)	count(*)
Failure	1
Success	100

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Used subquery to get the booster names list which carried the maximum payload mass

In [53]:
```
%%sql select distinct "Booster_Version" from SPACEXTABLE where "PAYLOAD_MASS__KG_" =
(select max("PAYLOAD_MASS__KG_") from SPACEXTABLE)
```

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

Out[53]:

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

In [66]:
```
%%sql select substr("Date", 6, 2) as month, "Landing_Outcome", "Booster_Version", "Launch_Site" from SPACEXTABLE where
"Landing_Outcome" = 'Failure (drone ship)' AND substr("DATE",1,4)='2015'
```

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

Out[66]:

month	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

In [71]:

```
%%sql select "Landing_Outcome", count(*)  from SPACEXTABLE where "Date" between '2010-06-04' and '2017-03-20'
group by "Landing_Outcome" order by 2 desc
```

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

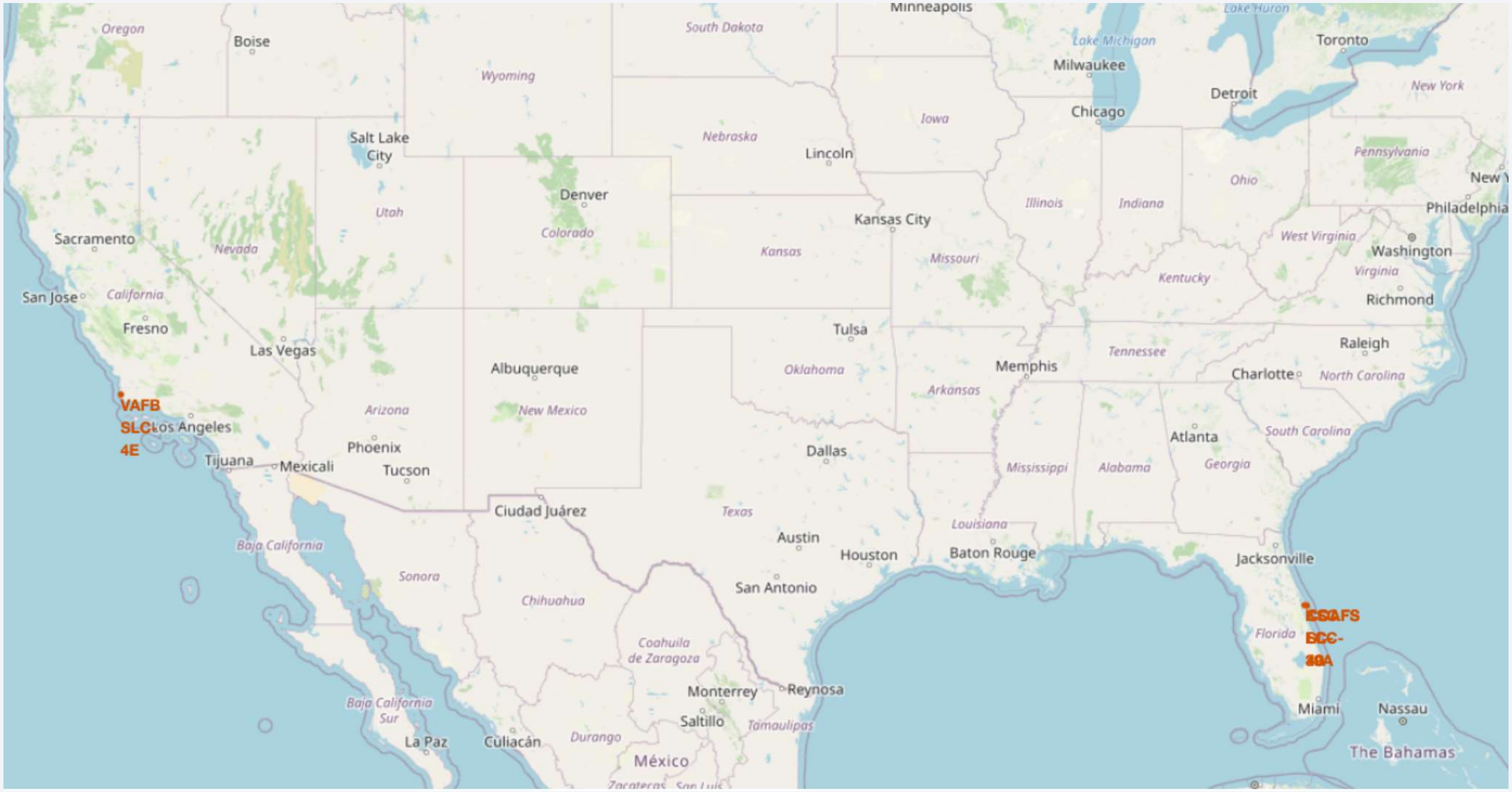
Out[71]:

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

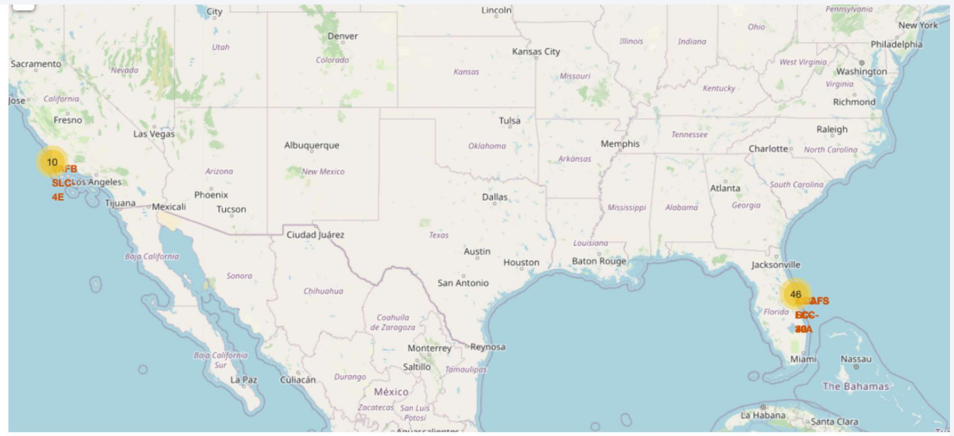
Section 3

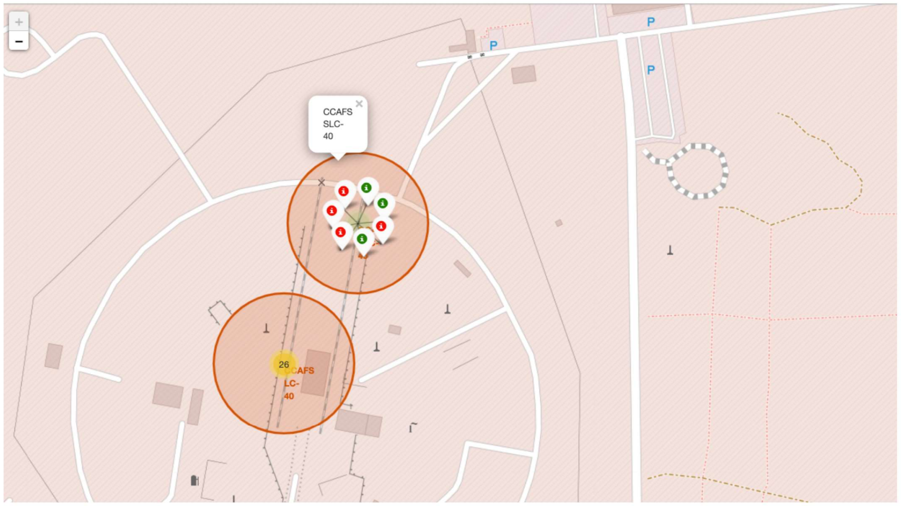
Launch Sites Proximities Analysis

All launch sites – Space x



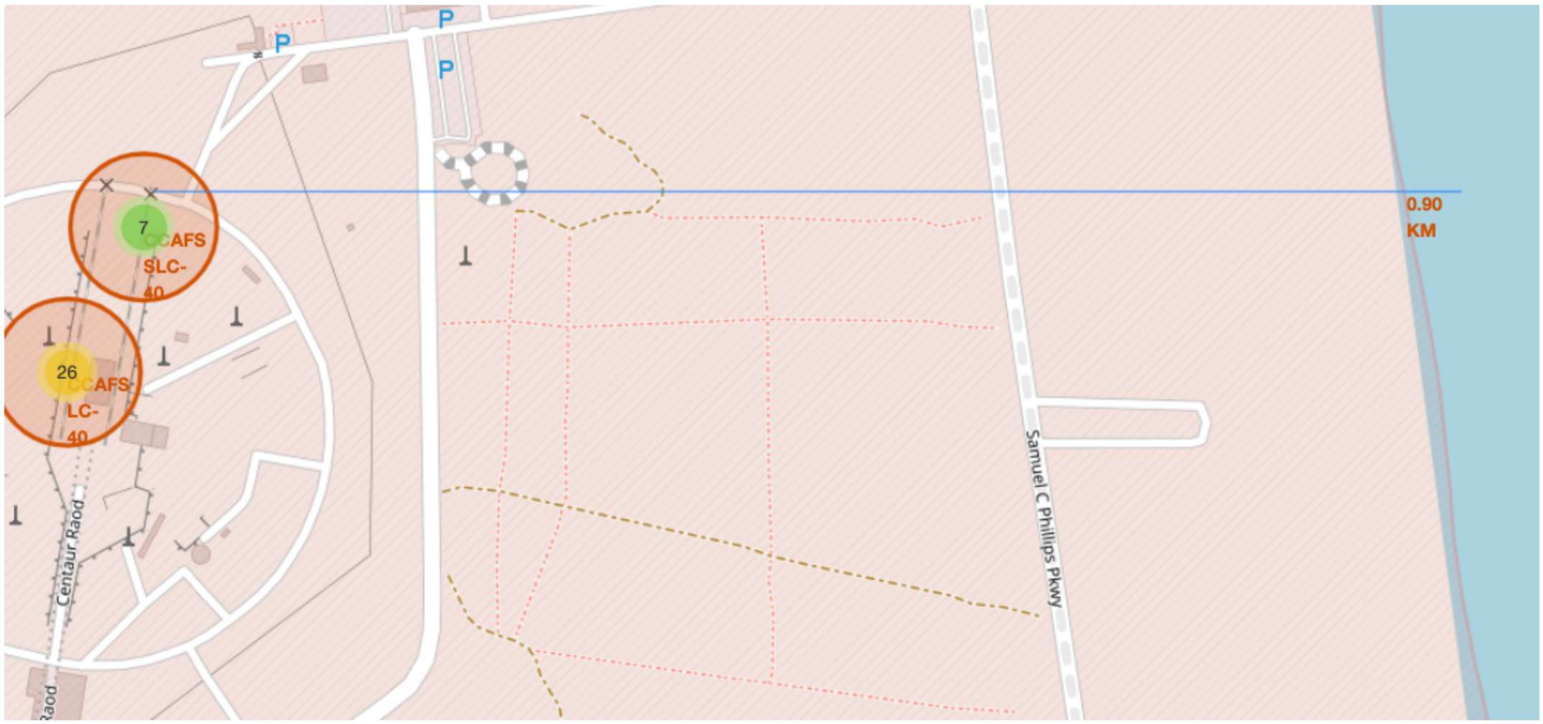
Launch sites with color labels





From the color-labeled markers in marker clusters, you should be able to easily identify which launch sites have relatively high success rates.

Launch Site distance to Proximities

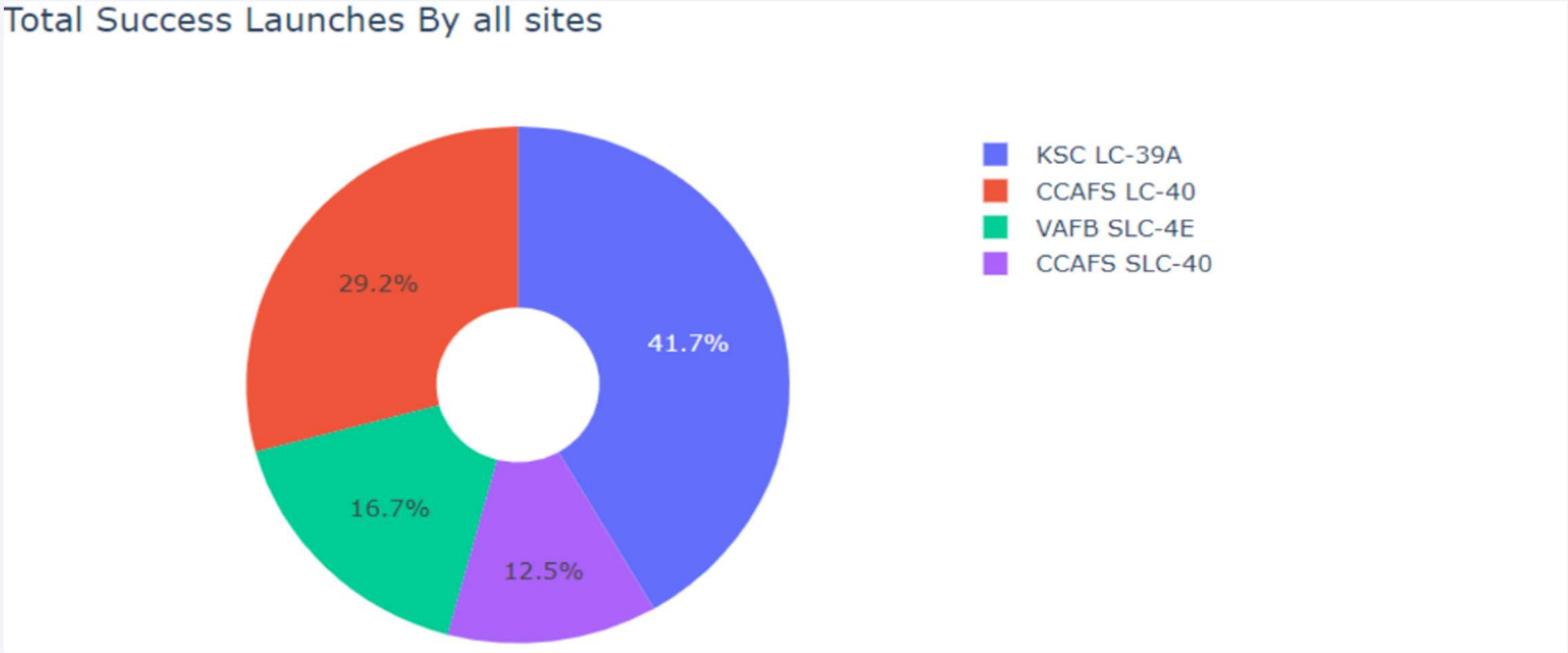


Section 4

Build a Dashboard with Plotly Dash

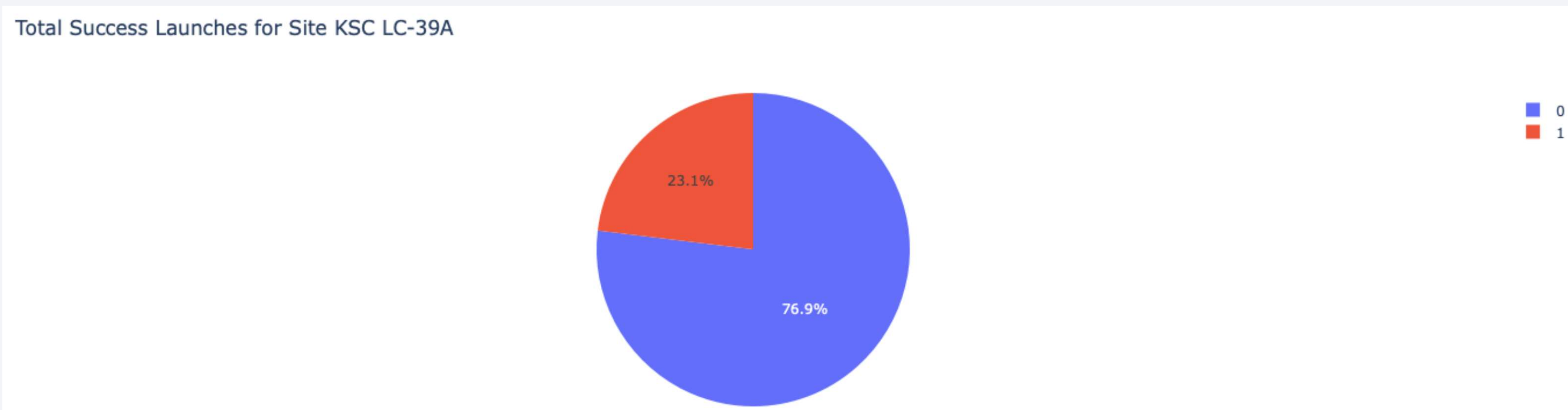
Success Launches by all sites

- Screenshot of launch success count for all sites, in a piechart

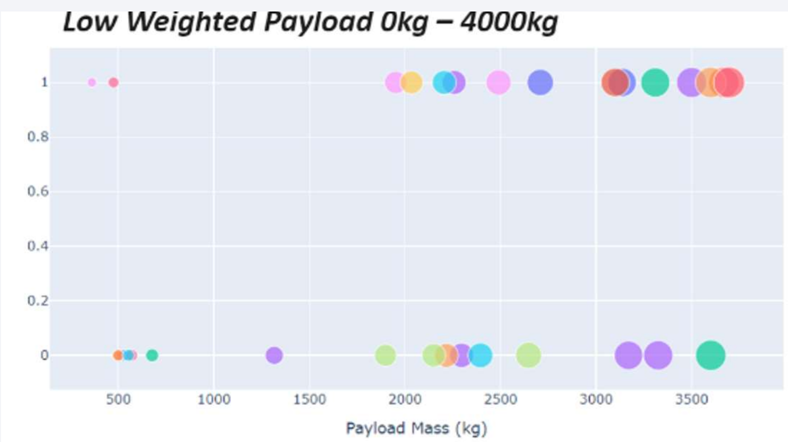


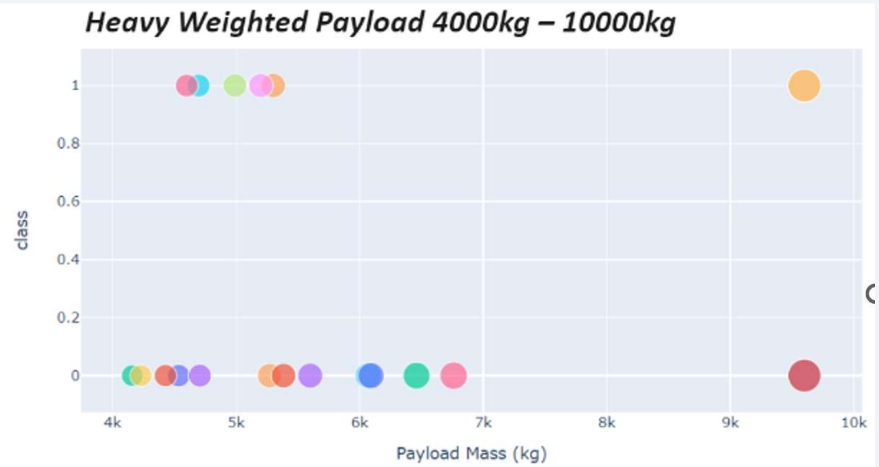
Success Launches for Site

- Screenshot of the piechart for the launch site with highest launch success ratio



Payload vs. Launch Outcome scatter plot for all sites





Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The decision tree classifier is the model with the highest classification accuracy

In [60]:
```
print("Model\t\tAccuracy\tTestAccuracy")#,logreg_cv.best_score_)
print("LogReg\t\t{}\t\t{}".format((logreg_cv.best_score_).round(5), logreg_cv.score(X_test, Y_test).round(5)))
print("SVM\t\t{}\t\t{}".format((svm_cv.best_score_).round(5), svm_cv.score(X_test, Y_test).round(5)))
print("Tree\t\t{}\t\t{}".format((tree_cv.best_score_).round(5), tree_cv.score(X_test, Y_test).round(5)))
print("KNN\t\t{}\t\t{}".format((knn_cv.best_score_).round(5), knn_cv.score(X_test, Y_test).round(5)))
```

```
Model           Accuracy        TestAccuracy
LogReg          0.84643         0.83333
SVM             0.84821         0.83333
Tree            0.875           0.83333
KNN             0.84821         0.83333
```

Confusion Matrix

- The confusion matrix for the decision tree classifier shows that the classifier can distinguish between the different classes. The major problem is the false positives .i.e., unsuccessful landing marked as successful landing by the classifier.

Conclusions

- The larger the flight amount at a launch site, the greater the success rate at a launch site.
- Launch success rate started to increase in 2013 till 2020.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- KSC LC-39A had the most successful launches of any sites.


Thank you!