IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Baby Yaw Lwi
2023-08-12

Outline

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

Executive Summary

- Summary of all methodologies
  - Data collection
  - Data wrangling
  - Exploratory data analysis (EDA)
  - Interactive visual analytics
  - Predictive analytics
- Summary of all results
  - Exploratory data analysis results
  - Interactive analytics results
  - Predictive analysis results

Introduction

- The target prediction of this project is the success or failure of the Falcon9 first stage landing.
- If we could correctly predict the outcome of the first stage, we could determine the cost of a launch, which is vital information to bid against SpaceX.

Section 1

Methodology

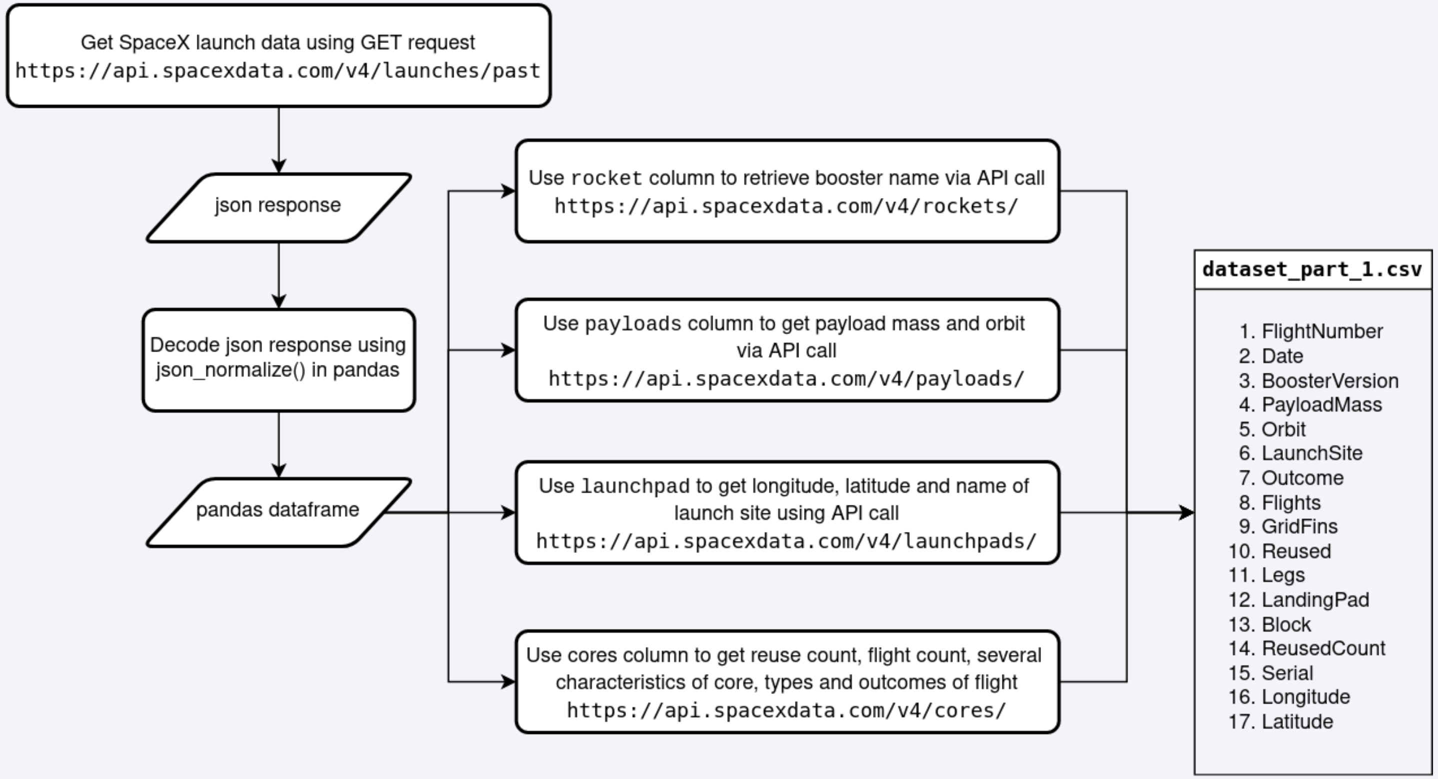
Methodology

Executive Summary

- Data collection methodology:
  - Launch data is collected from SpaceX API and web scraping
- Perform data wrangling
  - Class label data is calculated using outcome column of launch 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
  - Several classification models are tested and the best model is selected

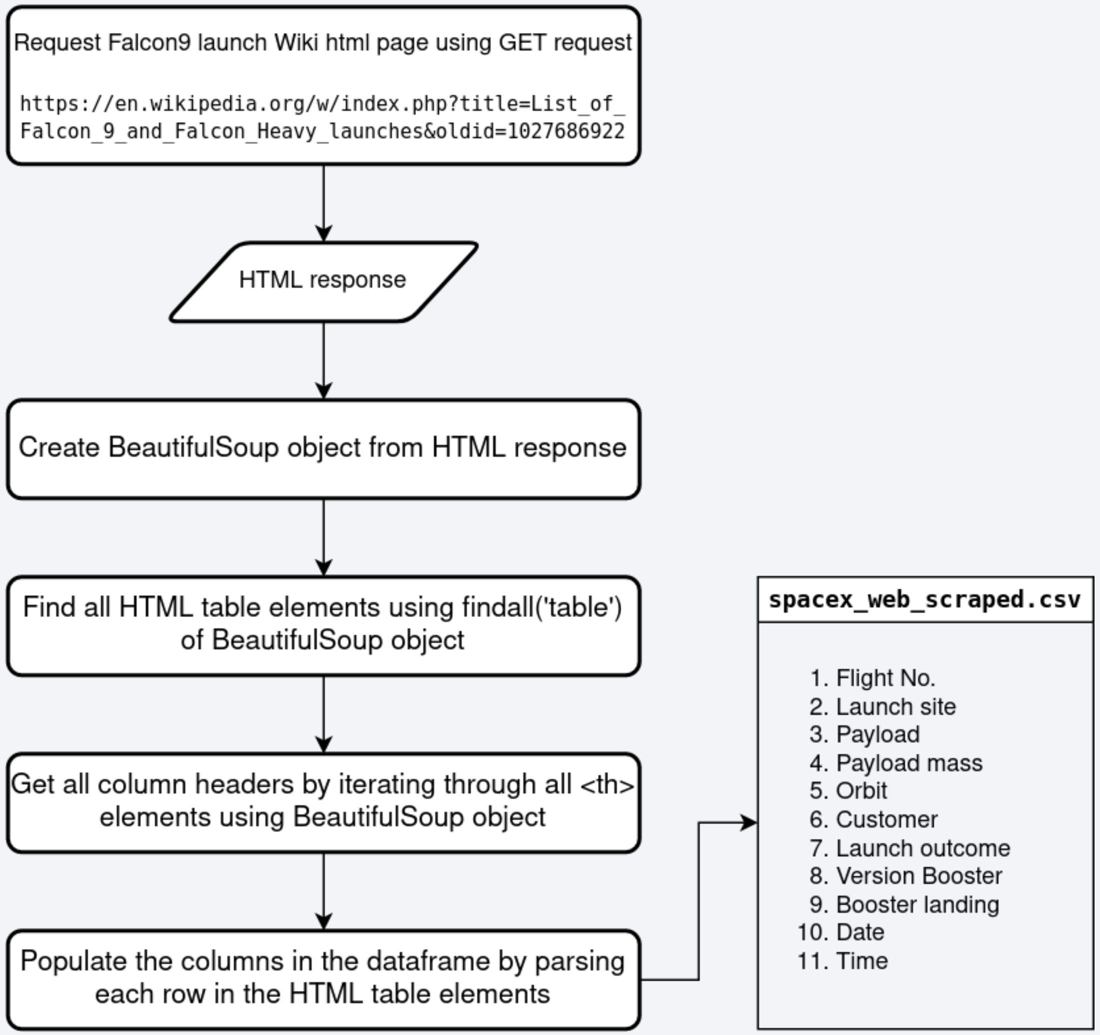
Data Collection – SpaceX API

- Collect historical launch data via SpaceX API
- Transform the json response to pandas dataframe using json_normalize() function
- Extract data about booster, payload mass, orbit, launch site data, core data, types and outcomes of the flight from corresponding additional APIs using collected data columns
- Combine all 17 data columns : FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- https://github.com/HpareBaby/Space-Race-Project-IBM/blob/master/capstone_W1_data_collection_api.ipynb



Data Collection - Scraping

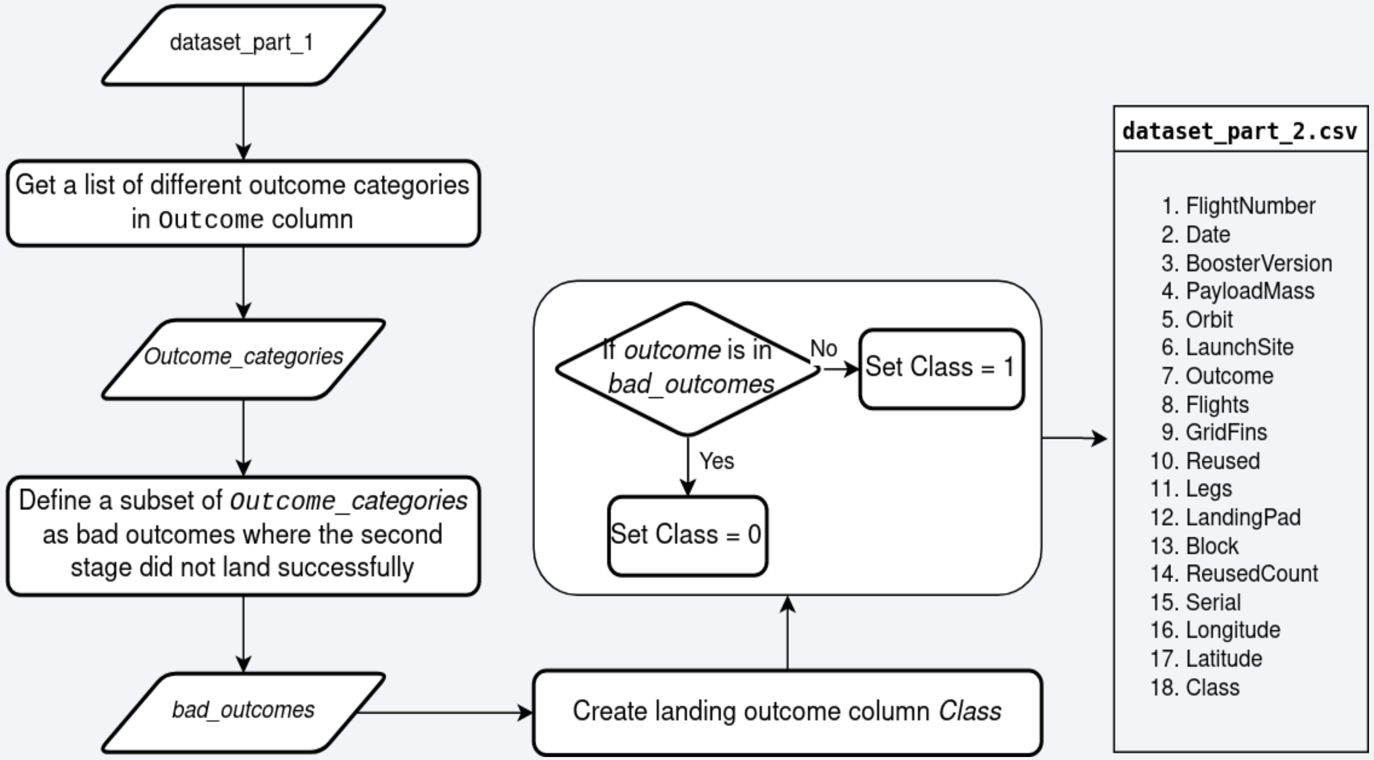
- Get the static Falcon9 launch Wiki page using GET request
- Find all HTML table elements using BeautifulSoup object
- Get all column headers by iterating through <th> elements using BeautifulSoup object
- Populate the dataframe by parsing each row and extracting data from the HTML table elements
- https://github.com/HpareBaby/Space-Race-Project-IBM/blob/master/capstone_W1_data_collection_web_scraping.ipynb



Data Wrangling

- Calculate the occurrence counts and number of different outcomes for each orbit
- Get a list of different outcome categories in Outcome column
- Define a subset of outcome categories as bad outcomes where the second stage did not land successfully, which include:
  - 'False ASDS', 'False Ocean', 'False RTLS', 'None ASDS', 'None None'
- Create a landing outcome label column Class.
- Set Class = 0, if outcome is in bad outcomes.

  Set Class = 1, otherwise.
- https://github.com/HpareBaby/Space-Race-Project-IBM/blob/master/capstone_W1_data_wrangling.ipynb



EDA with Data Visualization

The following charts are plotted for exploratory data analysis:

- Scatterplot of Flight Number vs. Launch Site to determine how the relationship between flight number and launch site would affect the launch outcome
- Scatterplot of Payload Mass vs. Launch Site to determine how the relationship between payload and launch site would affect the launch outcome
- Bar chart of average success rate of each orbit type to determine how orbit type affect the success rate
- Scatterplot of Flight Number vs. Orbit type to determine how the relationship between flight number and orbit type would affect the launch outcome
- Scatterplot of Payload Mass vs. Orbit type to determine how the relationship between payload and orbit type would affect the launch outcome
- Line chart to observe the yearly trend of average success rate from 2010 to 2020

https://github.com/HpareBaby/Space-Race-Project-IBM/blob/master/capstone_W2_eda_dataviz.ipynb

EDA with SQL

We used SQL queries to:

- List the unique launch sites
- List 5 records where launch sites begin with 'CCA'
- Calculate total payload mass carried by boosters launched by NASA (CRS)
- Calculate average payload mass carried by booster version F9 v1.1
- Get the date when the first successful landing outcome in ground pad was achieved
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Calculate the total number of successful and failure mission outcomes
- List the names of the booster versions which have carried the maximum payload mass
- List the month names, landing outcomes in drone ship, booster versions, launch site for the 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

https://github.com/HpareBaby/Space-Race-Project-IBM/blob/master/capstone_W2_eda_sql.ipynb

Build an Interactive Map with Folium

The following objects are added to the interactive map using Folium:

- Circle objects to mark the launch sites on the map
- Popup objects to display the launch site names when clicked
- MarkerCluster objects to display the successful and failed launches in each launch site with color-labeled markers
- MousePosition object to find out the coordinates of the location on mouse hover
- Polyline objects to draw a line between a launch site and selected landmarks in its proximity

https://github.com/HpareBaby/Space-Race-Project-IBM/blob/master/capstone_W3_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

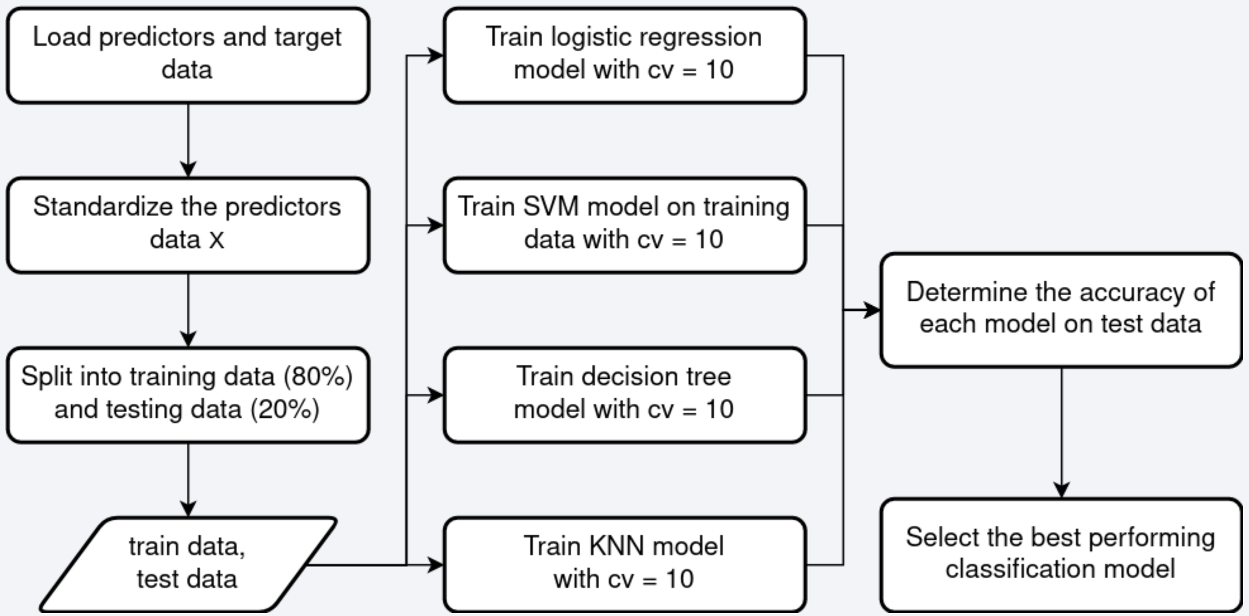
The following plots and interactions are added to the dashboard:

- Dropdown list to select launch sites
- Pie chart to show the number of successful and failed launches
- Range slider to specify the payload range
- Scatter plot to show the relationship between payload and success

https://github.com/HpareBaby/Space-Race-Project-IBM/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)

- Standardize the predictor variables
- Split the data into training (80%) and testing datasets (20%)
- Train the logistic regression, SVM, decision tree and KNN models and tune the corresponding hyperparameters of each model using GridSearchCV
- Determine the accuracy performance of each model on test data
- Select the best performing classification model



https://github.com/HpareBaby/Space-Race-Project-IBM/blob/master/capstone_W4_SpaceX_Machine_Learning_Prediction.ipynb

Results

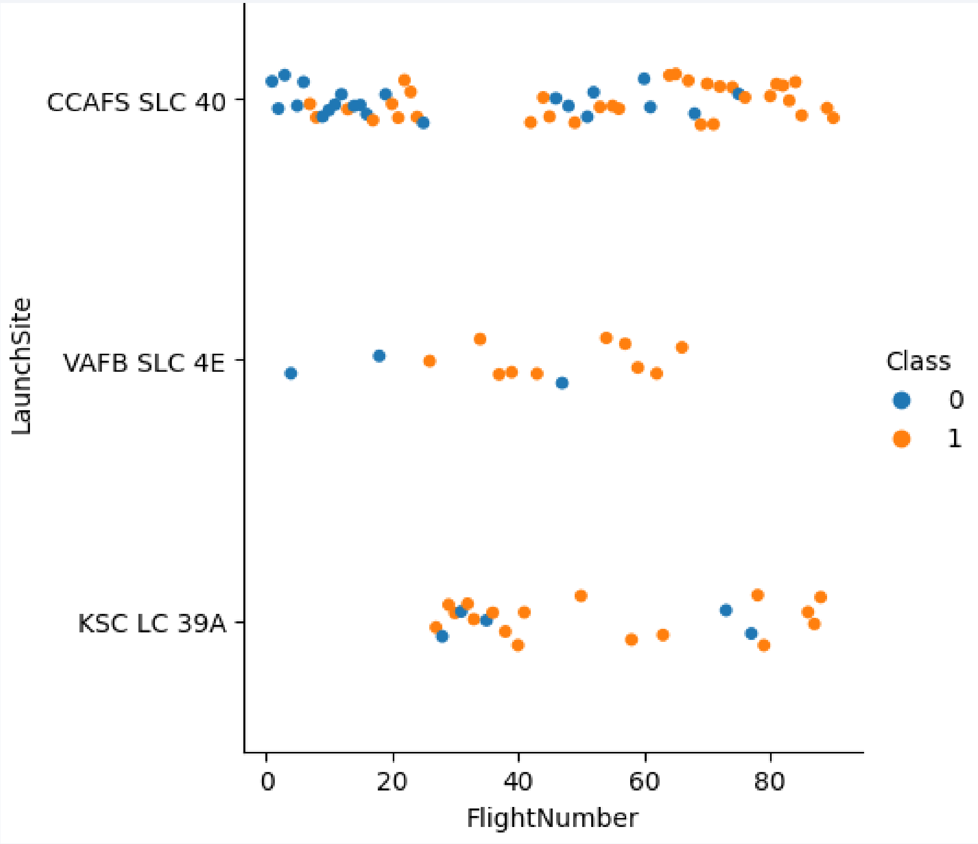
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

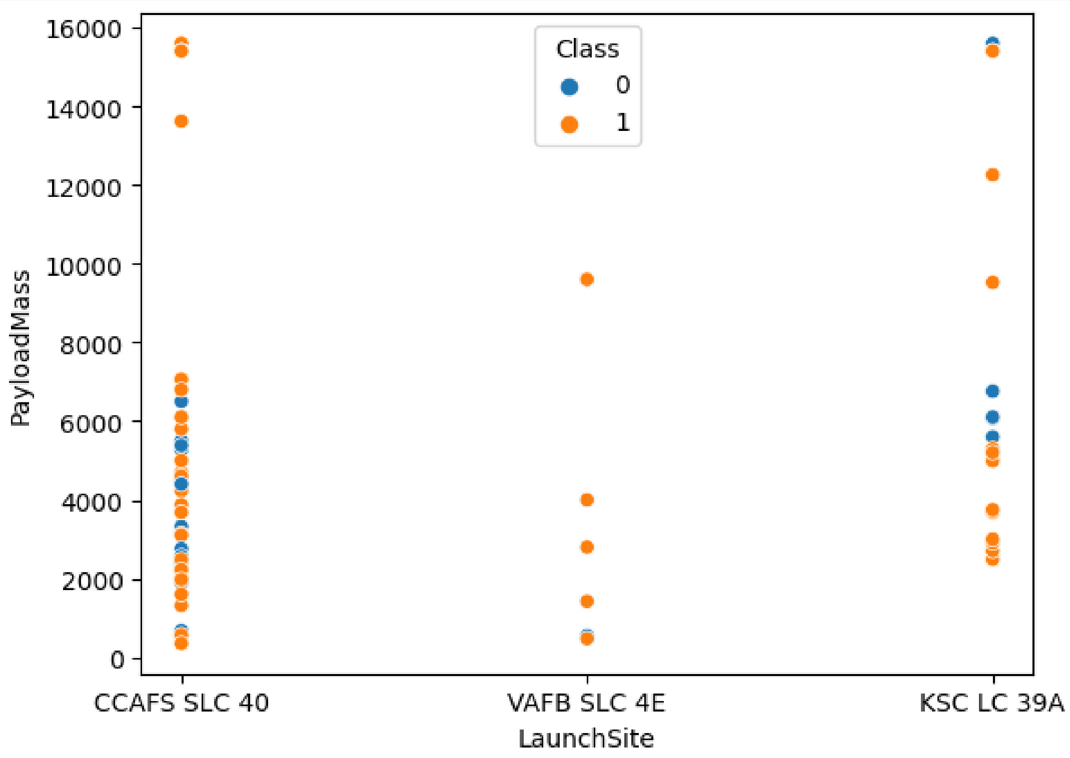
Flight Number vs. Launch Site

- Scatterplot of Flight Number vs. Launch Site to determine how the relationship between flight number and launch site would affect the launch outcome
- There is no relationship between success and flight number for CCAFS SLC 40 orbit.
- Generally, higher flight numbers have higher success rate for launch sites VAFB SLC 4E and KSC LC 39A.



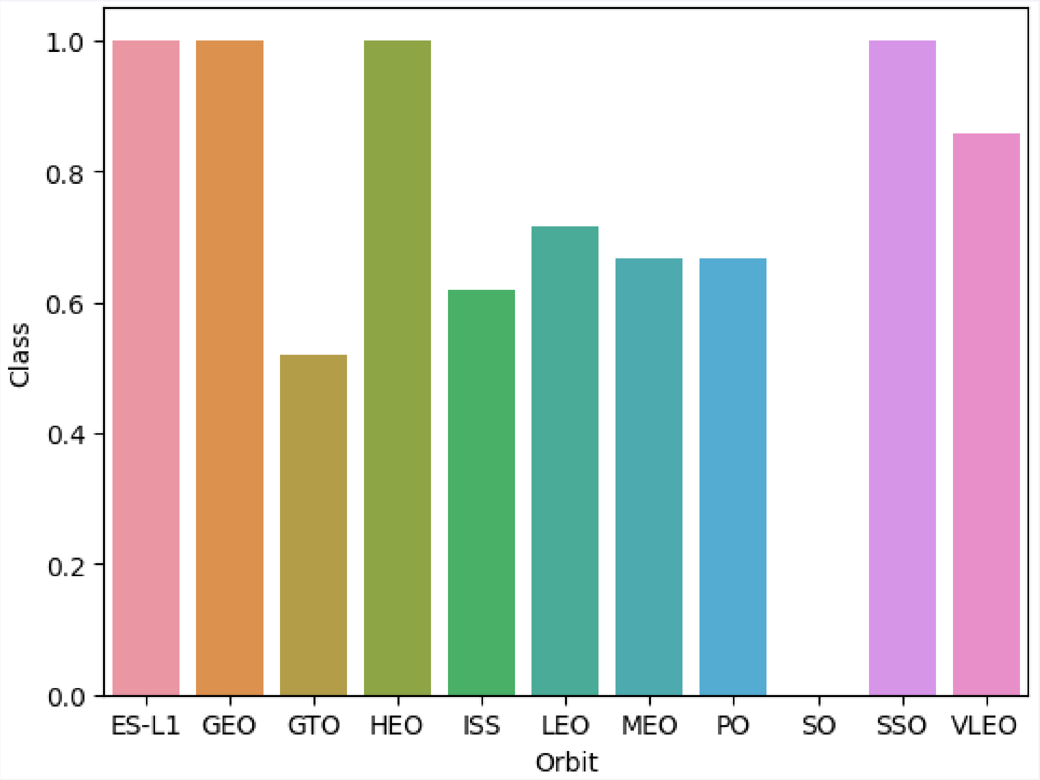
Payload vs. Launch Site

- Scatterplot of Payload Mass vs. Launch Site to determine how the relationship between payload and launch site would affect the launch outcome
- There are no rockets launched with heavy payload mass greater than 10,000 from the launch site VAFB SLC 4E.
- Heavier payloads, greater than 10,000, have higher success rate compared to lighter payloads for launch sites CCAFS SLC 40 and KSC LC 39A.



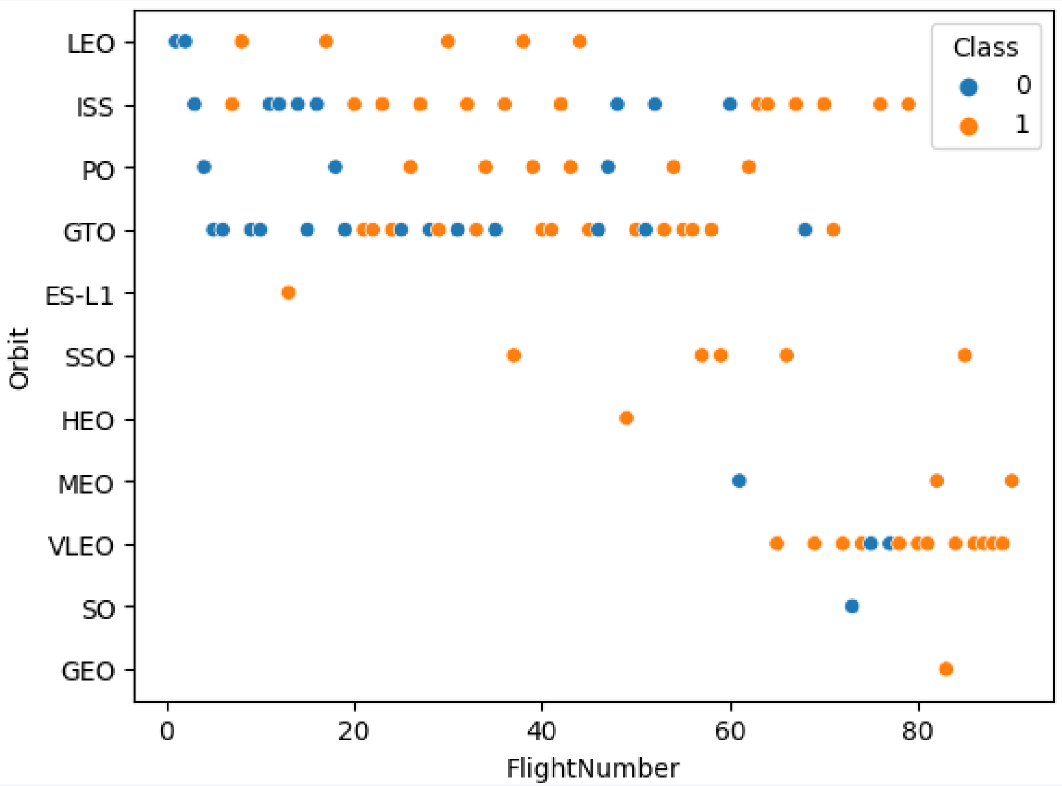
Success Rate vs. Orbit Type

- Bar chart of average success rate of each orbit type to determine how orbit type affect the success rate
- Orbit types ES-L1, GEO, HEO and SSO have the highest success rate of 1, whereas SO has the lowest success rate of 0.
- The orbit with the second highest success rate (0.857) is VLEO.
- Orbit types MEO and PO have the same success rate of 0.667.



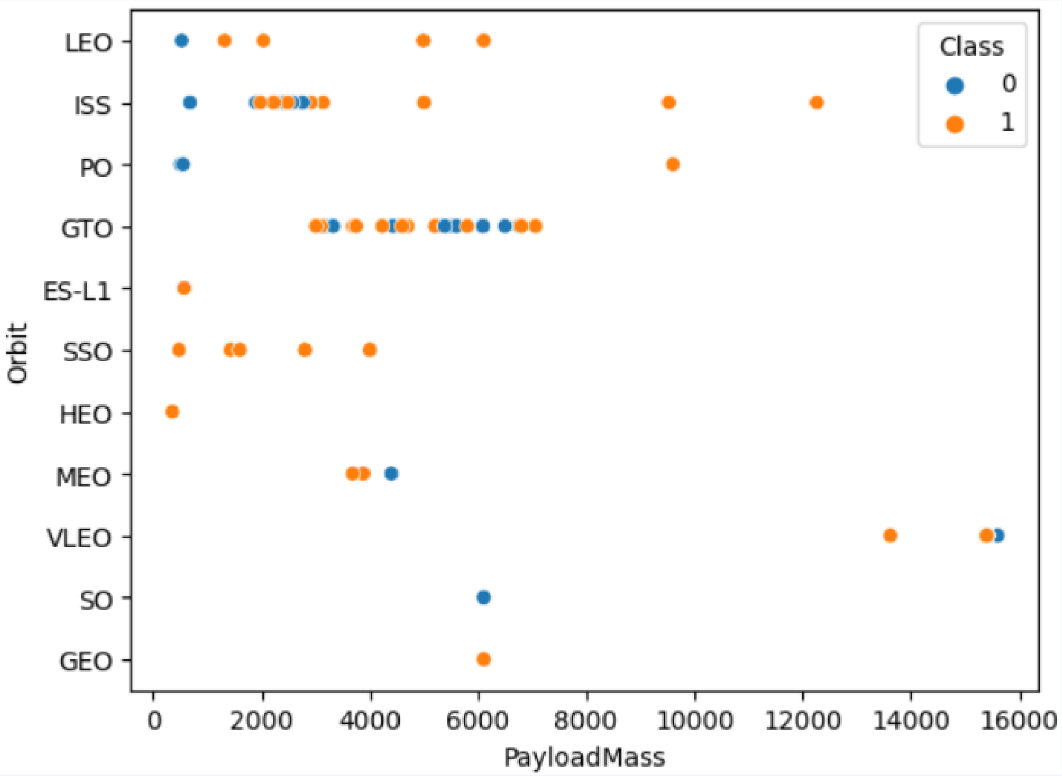
Flight Number vs. Orbit Type

- Scatterplot of Flight Number vs. Orbit type to determine how the relationship between flight number and orbit type would affect the launch outcome
- The number of successful landing increases with increasing flight number for orbit type LEO.
- There is no distinct relationship between success and flight number for GTO orbit.
- There are alternating non-linear relationship between positive outcomes and flight number for orbits of type ISS, PO and VLEO.



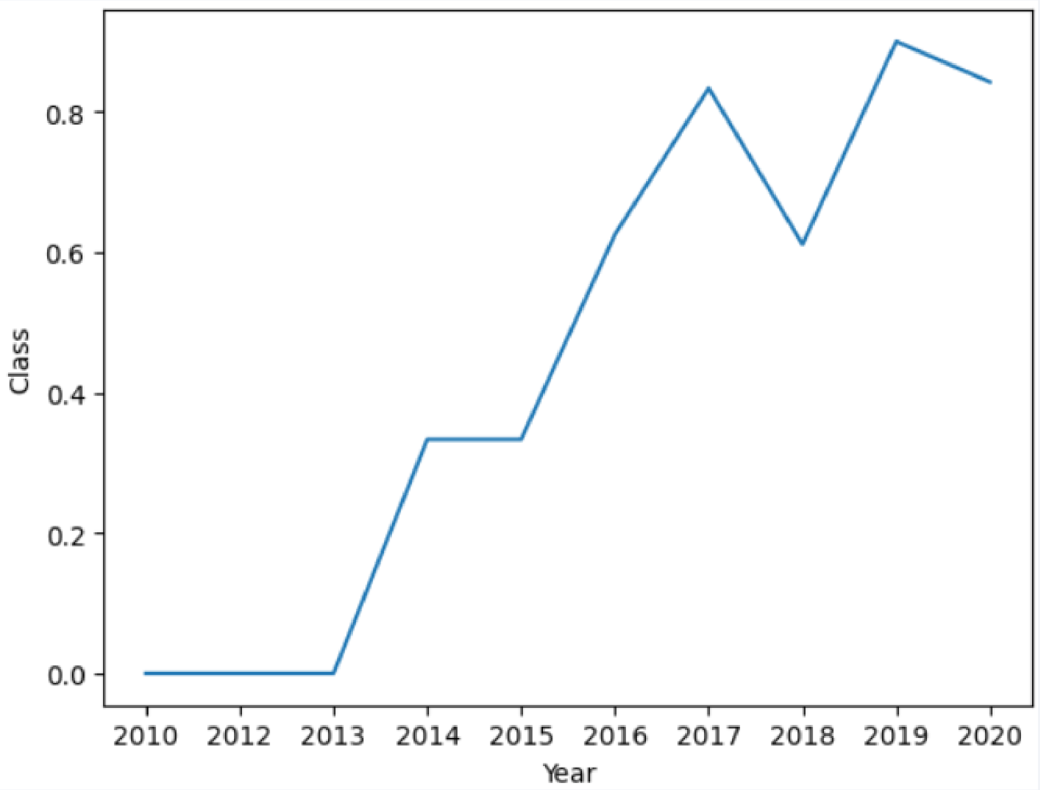
Payload vs. Orbit Type

- Scatterplot of Payload Mass vs. Orbit type to determine how the relationship between payload and orbit type would affect the launch outcome
- There are more successful landings with heavy payloads compared to lighter payload masses for orbit types PO, LEO and ISS.
- Orbit type GTO has almost equal distribution of positive and negative landings across different payload mass.



Launch Success Yearly Trend

- Line chart to observe the yearly trend of average success rate from 2010 to 2020
- The lowest success rate was 0 and it continued with no change from 2010 to 2013
- The success rate is in an increasing trend since 2013, with a peak of 0.9 in 2019



All Launch Site Names

- Find the names of the unique launch sites
- SELECT DISTINCT Launch_Site FROM 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
None

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5

```
%%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_
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer = 'NASA (CRS)'

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer = 'NASA (CRS)'
```

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

SUM(PAYLOAD_MASS__KG_)
45596.0

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE 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
- SELECT MIN(Date) FROM SPACEXTBL WHERE Landing_Outcome = 'Success (ground pad)'

```
%%sql
SELECT MIN(Date) FROM SPACEXTBL WHERE Landing_Outcome = 'Success (ground pad)'

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

MIN(Date)
01/08/2018

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
- SELECT DISTINCT Booster_Version FROM SPACEXTBL WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000

```
SELECT DISTINCT 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.
```

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
- SELECT Mission_Outcome, COUNT(*) AS COUNT FROM SPACEXTBL GROUP BY Mission_Outcome

```
SELECT Mission_Outcome, COUNT(*) AS COUNT FROM SPACEXTBL GROUP BY Mission_Outcome
```

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

Mission_Outcome	COUNT
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- SELECT DISTINCT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)

```
SELECT DISTINCT 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, booster versions, and launch site names for in year 2015
- SELECT SUBSTR(Date, 4, 2) AS MONTH, Landing_Outcome, Booster_Version, Launch_Site
- FROM SPACEXTBL WHERE Landing_Outcome LIKE 'Failure%' AND substr(Date, 7, 4)= '2015'

```
SELECT
SUBSTR(Date, 4, 2) AS MONTH,
Landing_Outcome,
Booster_Version,
Launch_Site
FROM SPACEXTBL --
WHERE Landing_Outcome LIKE 'Failure%' AND
substr(Date, 7, 4)= '2015'
```

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

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
- SELECT Landing_Outcome, COUNT(1) AS COUNT FROM SPACEXTBL WHERE Date >= '04/06/2010' AND Date <= '20/03/2017' GROUP BY Landing_Outcome ORDER BY COUNT DESC

```
SELECT Landing_Outcome, COUNT(1) AS COUNT
FROM SPACEXTBL
WHERE Date >= '04/06/2010'
AND Date <= '20/03/2017'
GROUP BY Landing_Outcome
ORDER BY COUNT DESC
```

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

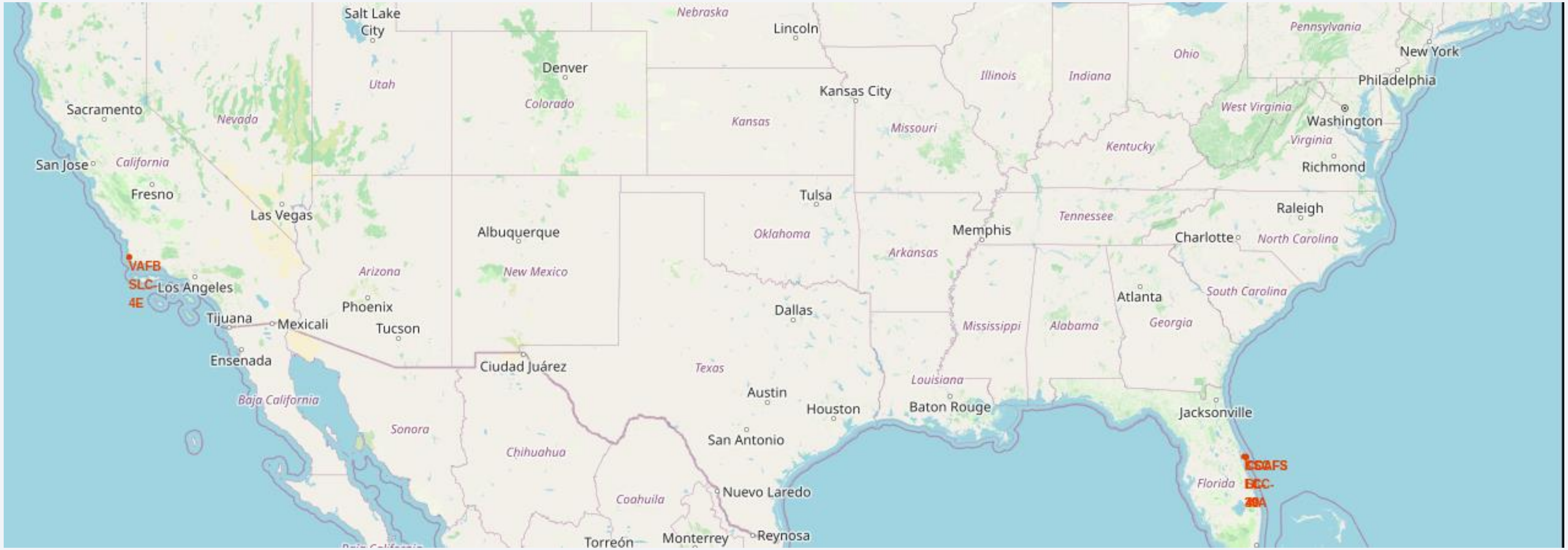
Landing_Outcome	COUNT
Success	20
No attempt	9
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

Section 3

Launch Sites Proximities Analysis

Launch site locations

- There are 4 launch sites, with one on the west coast and 3 on the east coast.

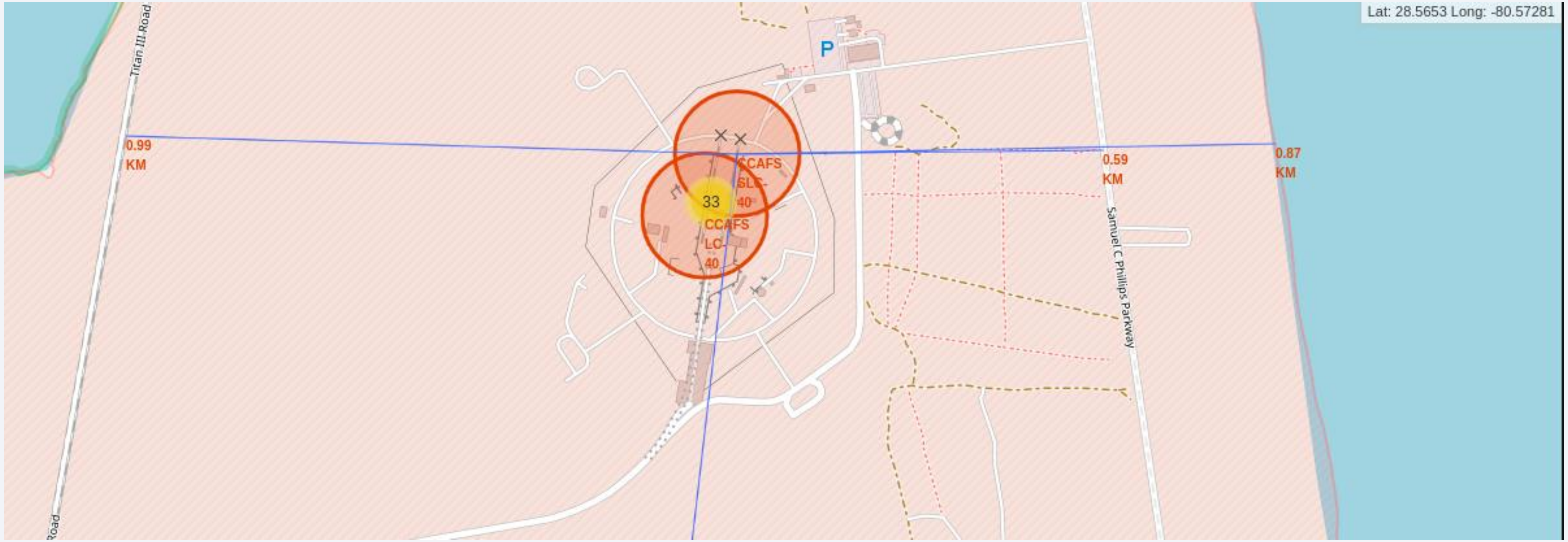


Successes and failures in each launch site

- Launch site CCAFS LC-40 has the largest number of launches, 33, with the lowest success rate of 26.9%.
- Launch site KSC LC-39A has the highest success rate of 76.9%.

Launch site proximities

- Launch site (CCAFS SLC-40) is closest to highway with only about 0.59 KM distance.
- The second closest in distance is coastline with 0.87 KM away.
- The distance from the launch site to highway is about 0.99 KM.
- The launch site is farthest from the city with a distance of 51.16 KM.

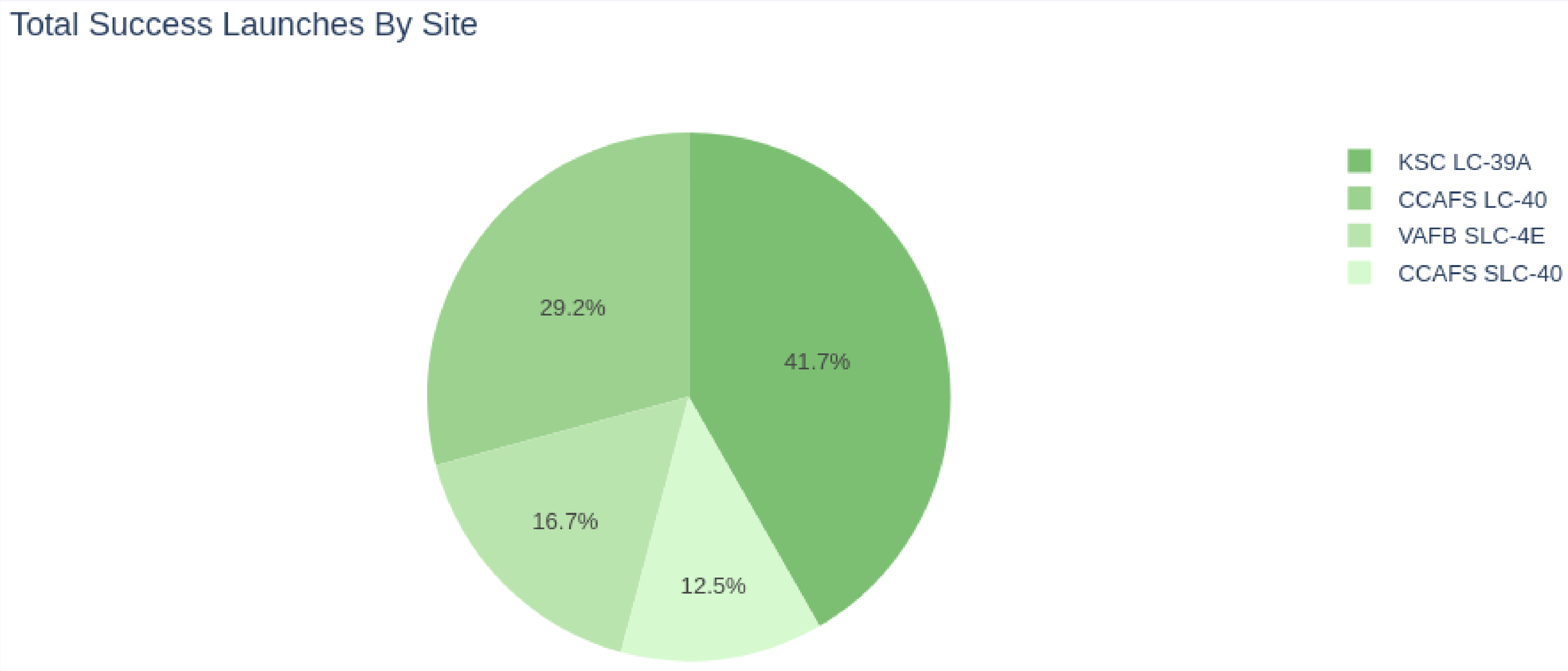


Section 4

Build a Dashboard with Plotly Dash

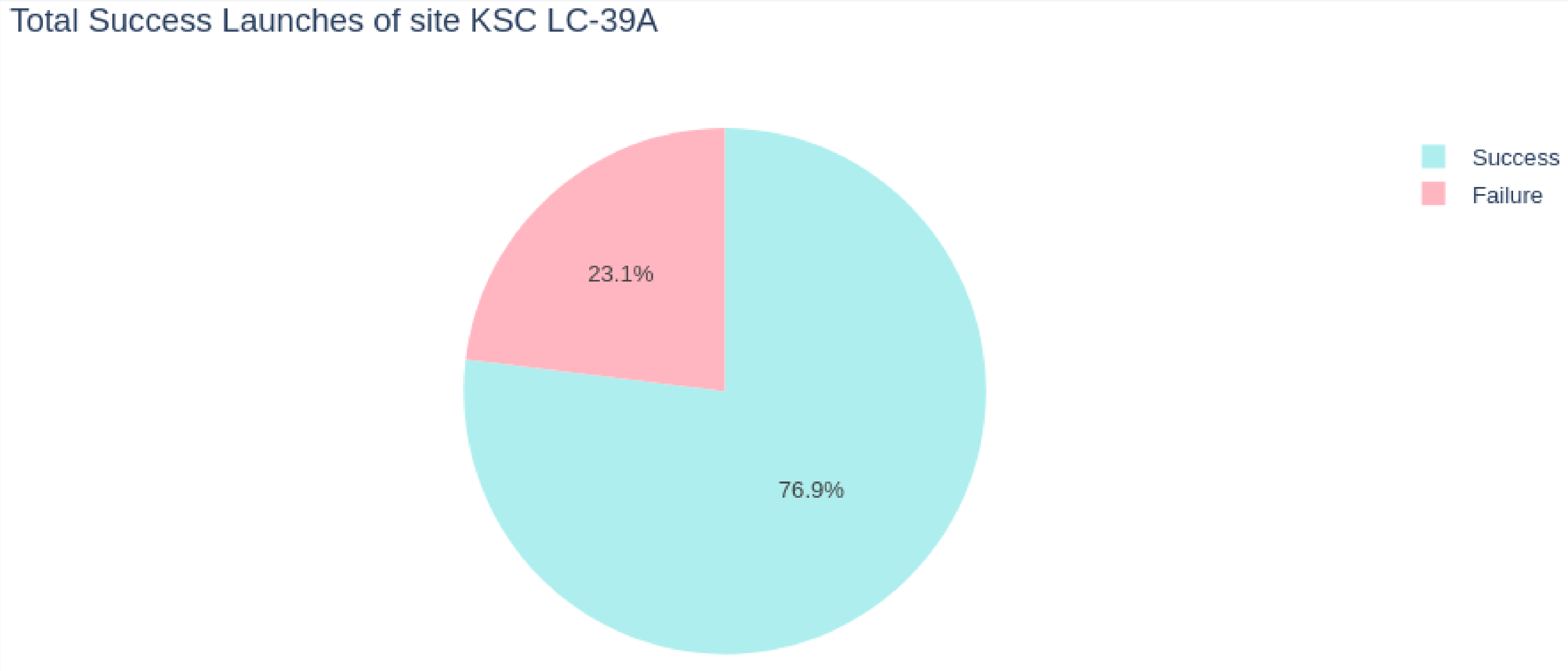
Total success launches by site

- Launch site KSC LC-39A has the highest attribution of 41.7% of all successful landings, whereas CCAFS SLC-40 has the lowest attribution of 12.5%.



Launch site with highest success rate

- Launch site KSC LC-39A has the highest success rate of 76.9%.



Correlation between payload and success

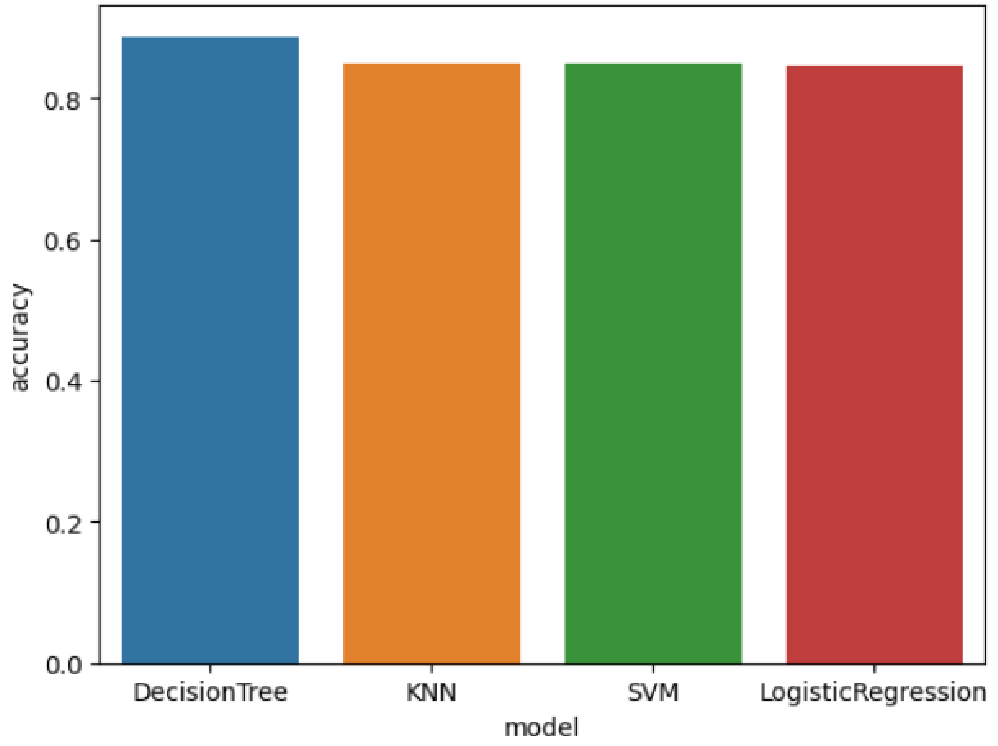
- Booster version category BT has 100% success rate with only 1 launch. FT has the second highest success rate with 66.67% of total 24 launches.

Section 5

Predictive Analysis (Classification)

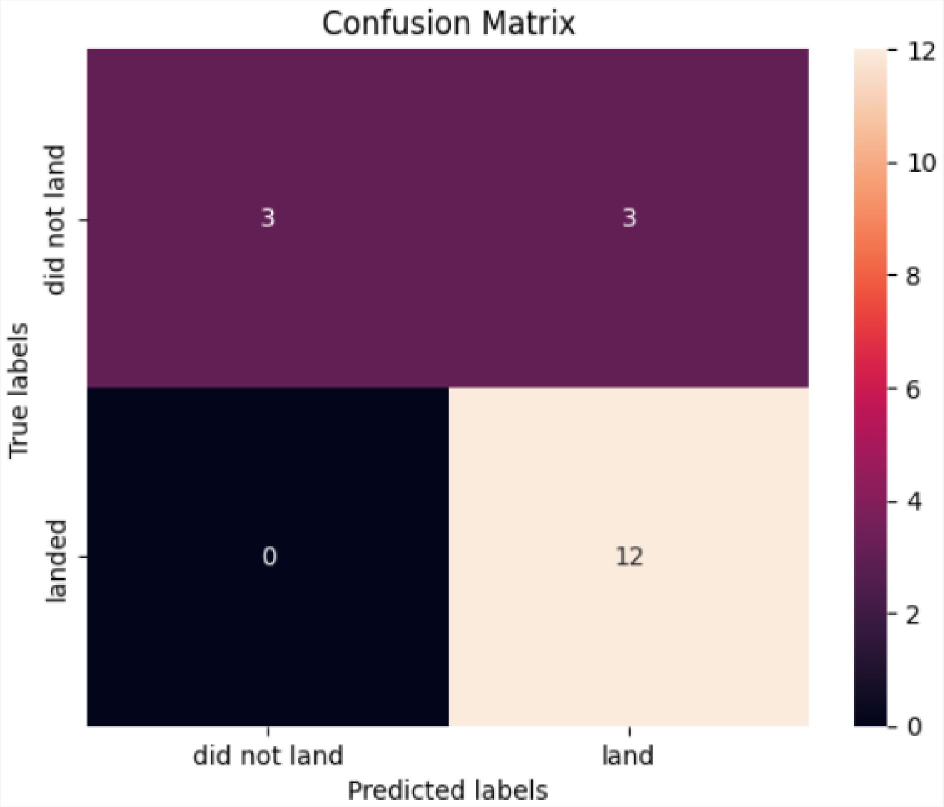
Classification Accuracy

- Decision tree model has the highest training accuracy with 88.75%.



Confusion Matrix

- Confusion matrix shows the performance of the decision tree model on test data
- The model could predict all the actual positive outcomes correctly, with 100% recall rate
- However, there are 3 false positives predictions on the test data



Conclusions

- Orbit types ES-L1, GEO, HEO and SSO have the highest success rate of 1, whereas SO has the lowest success rate of 0.
- The success rate is in an increasing trend since 2013, with a peak of 90% in 2019.
- Launch site CCAFS LC-40 has the largest number of launches, 33, with the lowest success rate of 26.9%.
- Launch site KSC LC-39A has the highest success rate of 76.9%.
- Decision tree model has the highest model accuracy with 100% recall rate.

Appendix

- Python code snippet for decision tree model training with GridSearchCV

```
parameters = {
     'criterion': ['gini', 'entropy'],
     'splitter': ['best', 'random'],
     'max_depth': [2*n for n in range(1,10)],
     'max_features': ['auto', 'sqrt'],
     'min_samples_leaf': [1, 2, 4],
     'min_samples_split': [2, 5, 10]
}
tree = DecisionTreeClassifier()
tree_cv = GridSearchCV(tree, parameters, cv=10)
tree_cv.fit(X_train, Y_train)
```

Thank you!