IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

William Lathlean
June 2024

Outline

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

Executive Summary

- SpaceX has developed a reusable first stage for their Falcon 9 rocket.
- Reuse of the first stage provides a huge savings, over $100 million each launch.
- Using a variety of methodologies, we will predict:
  - Success or failure for each launch
  - The optimum payload size
  - Best launch sites
- Summary of methodologies
  - Data Collection (using API, Web Scraping)
  - Data Wrangling
  - Exploratory Data Analysis (using SQL, Data Visualization)
  - Interactive Visual Analytics (using Folium)
  - Interactive Dashboard (using Plotly Dash)
  - Launch Outcome Prediction (using Machine Learning)
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

Introduction

- Background
  - The commercial space age has begun, making space travel more accessible through companies like Virgin Galactic, Rocket Lab, Blue Origin, and SpaceX.
  - SpaceX stands out with significant accomplishments, such as sending spacecraft to the International Space Station, launching the Starlink satellite internet constellation, and conducting manned missions, facilitated by cost-effective rocket launches due to the reusability of the Falcon 9's first stage.
- Problems to Solve
  - Determine the price of each rocket launch for a new competitor, Space Y, by analyzing SpaceX's costs.
  - Predict if SpaceX will reuse the first stage of their Falcon 9 rockets using a machine learning model and public information, without relying on traditional rocket science methods.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - From SpaceX using REST API
  - From Wikipedia using BeautifulSoup
- Perform data wrangling
  - Data cleaning, filtering, dealing with missing values
  - Class label creation
  - Performed one-hot encoding
- 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

Collect Data from SpaceX

Retrieve Spacex launch data via URL using request.get (REST API) → Decode JSON response using .json() method → Convert to Pandas data frame using .json_normalize() method

Collect Data from Wikipedia

Retrieve Spacex launch data via URL using request.get (HTML) → Create a BeautifulSoup object from the HTML response using html.parser → Locate the table containing desired data → Create a dictionary from the table data → Convert to Pandas data frame using pd.DataFrame on the values in the dictionary

Data Collection – SpaceX API

- Full code available on GitHub:
- [WL_jupyter-labs-spacex-data-collection-api.ipynb]

Get data through call to API:
response = requests.get(static_json_url)

↓

Check response code:
response.status_code
(200 is success)

↓

Decode JSON response:
data_json = response.json()

↓

Convert JSON data to DataFrame:
data = pd.json_normalize(data_json)

Data Collection - Scraping

- Full code available on GitHub:
- WL_jupyter-labs-webscraping-2.ipynb

Perform an HTTP GET method:
response = requests.get(static_url)

↓

Check response code:
response.status_code
(200 is success)

↓

Create a BeautifulSoup object from the HTML:
soup = BeautifulSoup(response.text, 'html.parser')

↓

Use the find_all function to locate tables
html_tables = soup.find_all('table')

↓

Parse the table into a dictionary and create a data frame

Data Wrangling

- Full code available on GitHub:
- WL_jupyter-labs-spacex-Data wrangling.ipynb

Determine the number of launches on each site:

```
print(df['LaunchSite'].value_counts())
```

```
CCAFS SLC 40    55
KSC LC 39A      22
VAFB SLC 4E     13
Name: LaunchSite, dtype: int64
```

↓

Calculate the number and occurrence of each orbit:

```
print(df['Orbit'].value_counts())
```

↓

Calculate the number and occurence of mission outcome of the orbits:

```
landing_outcomes = df['Outcome'].value_counts()
```

↓

Create a set of unsuccessful second stage landings:

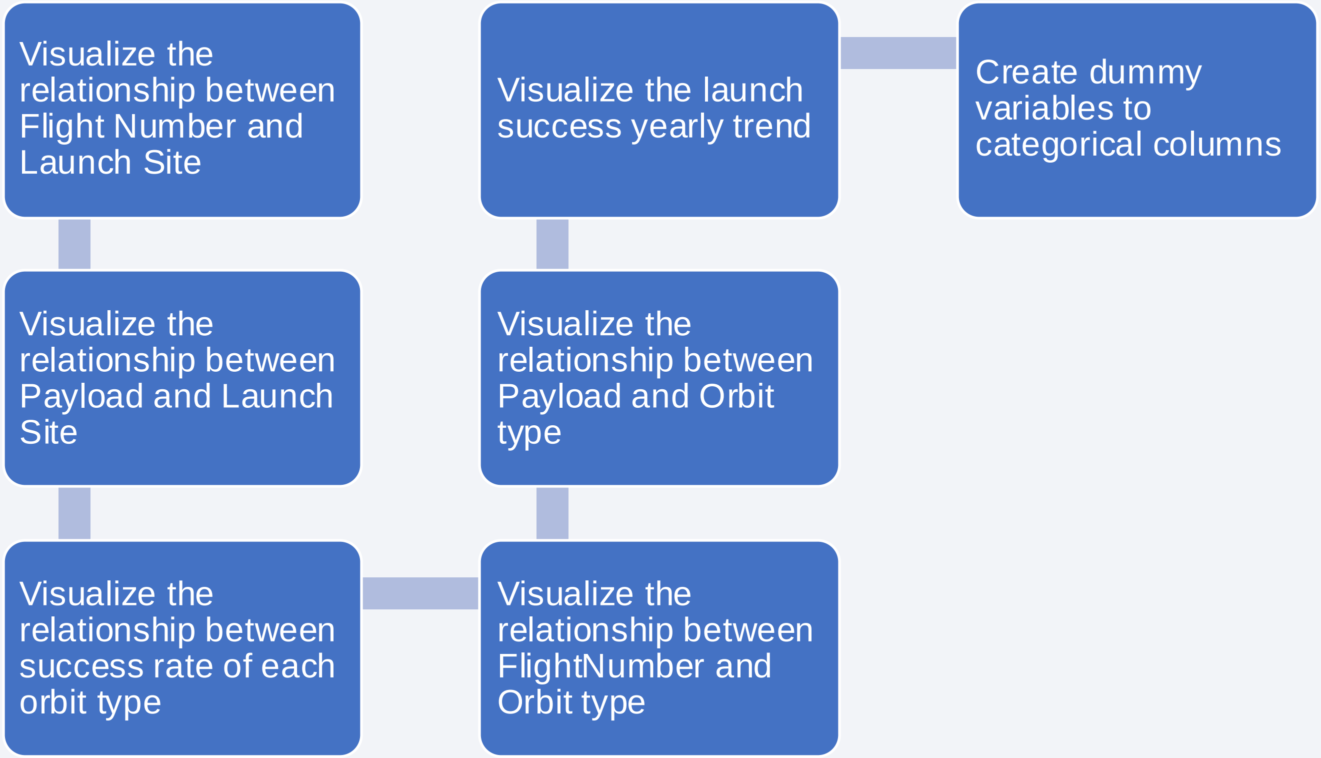
```
bad_outcomes=set(landing_outcomes.keys()[[1,3,5,6,7]])
```

↓

Create new column "Class" to label outcomes as successful (1) or failed (0).

EDA with Data Visualization

- Full code available on GitHub:
- [WL_jupyter-labs-eda-dataviz.ipynb]



EDA with SQL

- Full code available on GitHub:
- WL_jupyter-labs-eda-sql-coursera_sqllite.ipynb

Load the SQL extension and establish a connection with the database

Use sqlite and SQL commands to query the database

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

```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE;
```

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

```
%sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" like 'CCA%' LIMIT 5
```

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

```
%sql SELECT Customer, sum(payload_mass__kg_) as total_payload FROM SPACEXTABLE WHERE Customer LIKE 'NASA (CRS)%';
```

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

```
%sql select AVG(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version like 'F9 v1.1%';
```

List the date when the first succesful landing outcome in ground pad was achieved:

```
%sql select Landing_Outcome, min(Date) from SPACEXTABLE where "Landing_Outcome" = "Success (ground pad)";
```

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

```
%%sql
select distinct Booster_Version, Landing_Outcome,
payload_mass__kg_ as Payload
from spacextable
where Landing_Outcome = "Success (drone ship)" and
(payload_mass__kg_ < 6000 and payload_mass__kg_ > 4000)
order by payload_mass__kg_ asc;
```

List the total number of successful and failure mission outcomes:

```
%%sql
select Mission_Outcome, count(Mission_Outcome) as
"Total_Missions"
from spacextable
group by Mission_Outcome;
```

List the names of the booster_versions which have carried the maximum payload mass:

```
%%sql
select distinct Booster_Version, PAYLOAD_MASS__KG_
from spacextable
where PAYLOAD_MASS__KG_ = (select MAX(PAYLOAD_MASS__KG_)
from spacextable)
order by Booster_Version;
```

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.

Build an Interactive Map with Folium

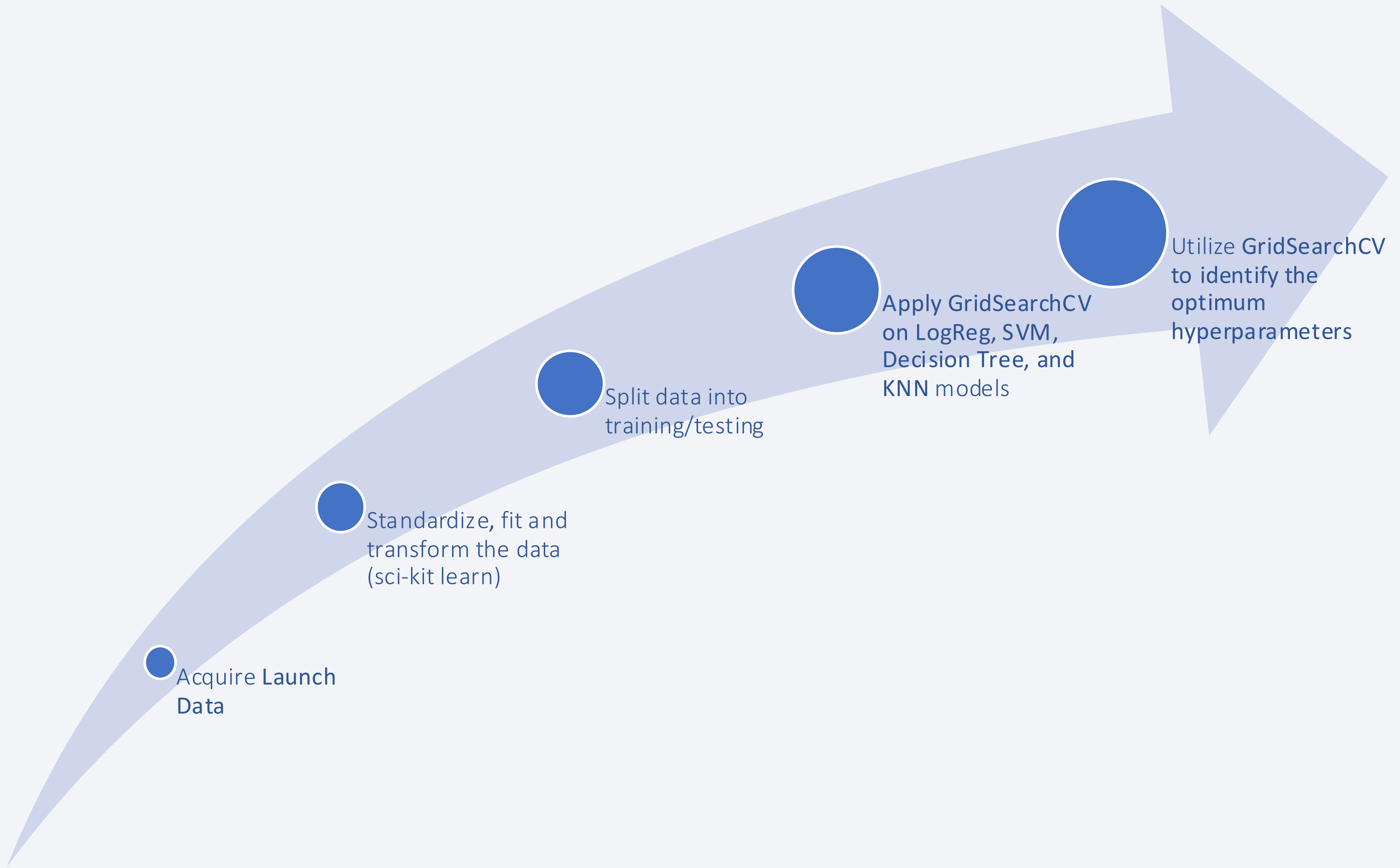
- Map objects were added, such as:
  - Circles (to indicate specific locations, like Johnson Space Center, and SpaceX launch sites)
  - Markers (showing individual launches and indicating success or failure)
  - Labels (launch site names)
  - Lines (showing relationships and distances between objects)
- These objects were added to help visually recognize geographical patterns about launch sites and their results.
- Jupyter lab code: WL_lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- An interactive dashboard was created using Plotly Dash displaying the following:
  - A drop down list to allow the user to select a specific launch site (or all sites) for the data to be displayed.
  - A pie chart displaying the total successful launches by selected site(s)
  - A scatter plot showing the correlation between Payload Mass (Kg) to Outome for all of the various booster versions launched from selected site(s).
  - A slider to select payload range to be displayed in the scatter plot

- GitHub URL: spacex_dash_app.ipynb

Predictive Analysis (Classification)



- GitHub URL: [SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb]()

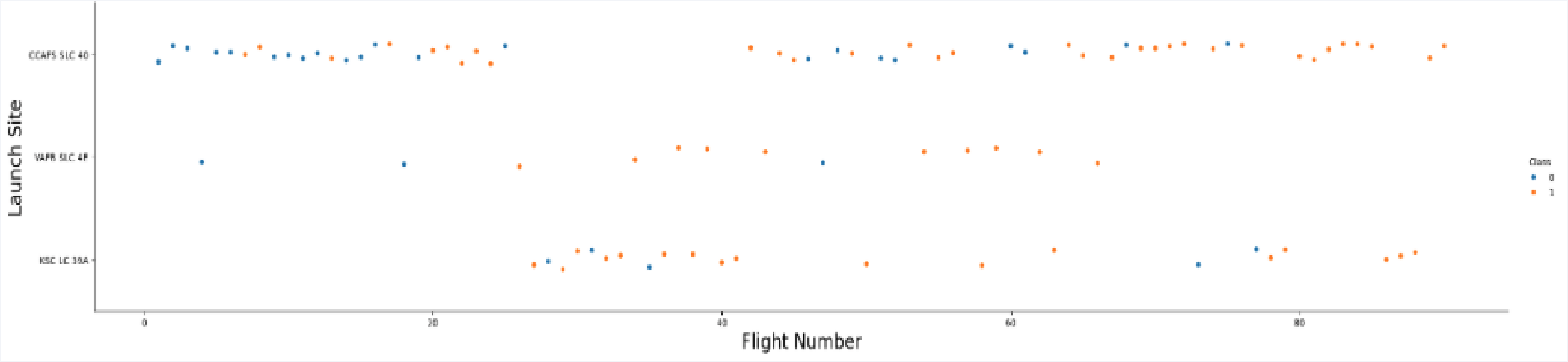
Results

- Exploratory Data Analysis
  - Most successful launch site = KSC LC-39A
  - Launch success 2013-2020 upward trend
  - Orbit types with highest success rate (100%)
    - SSO (5 launches)
    - ES-L1 (1 launch)
    - GEO (1 launch)
    - HEO (1 launch)
- Visual Analysis
  - Four launch sites, all are situated near coastlines
  - Launch sites are within relatively close proximity to railways, highways, airports and/or military bases.
  - More launches from East cost sites
- Predictive Analysis
  - Decision Tree classifier performed best at 94.44% accuracy (as opposed to other methods ranged 83%-88%)

Section 2

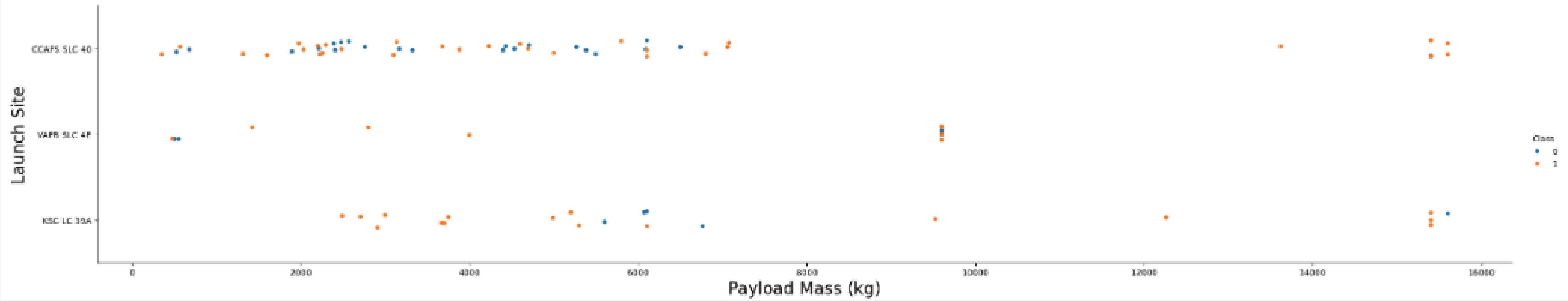
Insights drawn from EDA

Flight Number vs. Launch Site



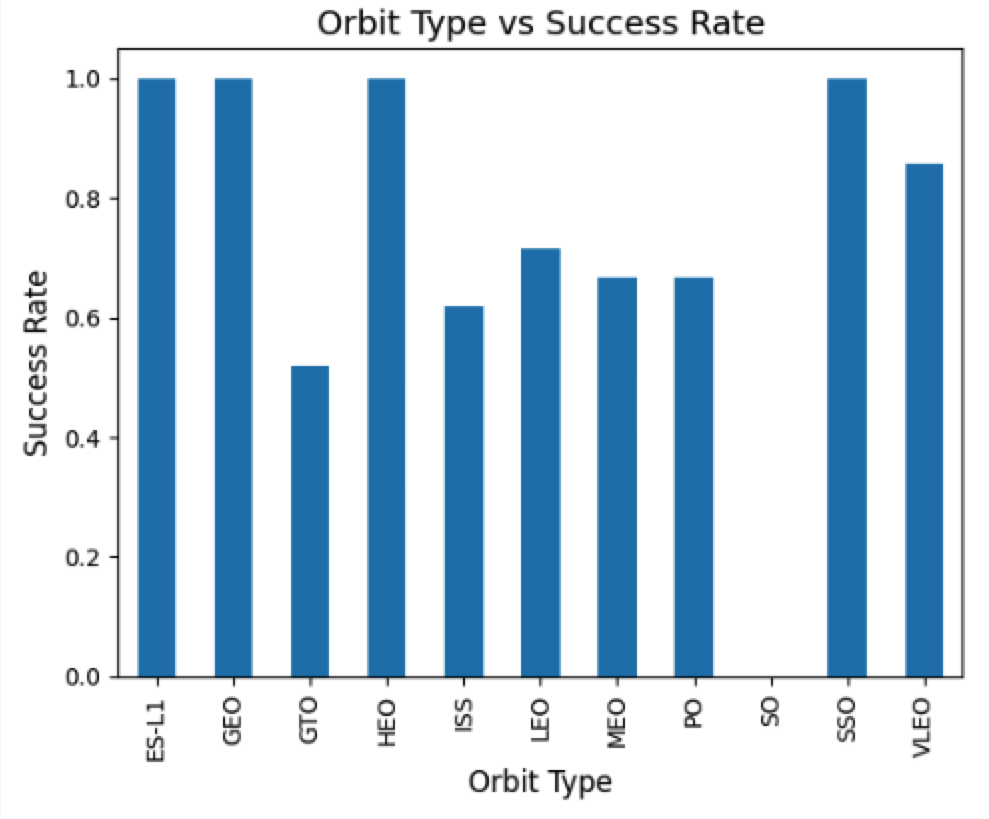
- Launch success (orange dots) increased with number of flights.
- CCAFS SLC 40 site had most launches overall

Payload vs. Launch Site



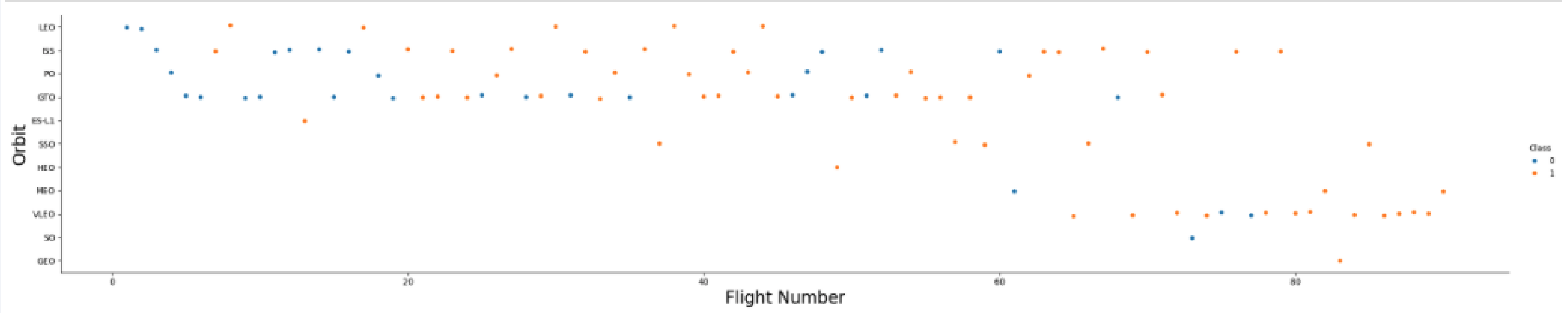
- No launches from VAFB-SLC site with payload mass greater than 10,000 kg
- It appears that launch success is greater for payloads over 10,000 kg

Success Rate vs. Orbit Type



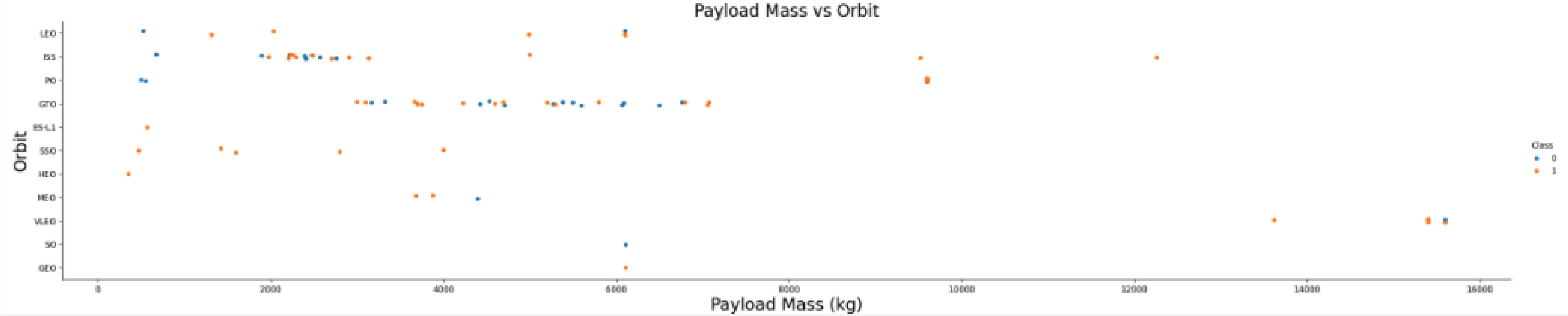
- SSO, ES-L1, GEO, and HEO type orbits were 100% successful.
- SO orbit type has no successful launches.

Flight Number vs. Orbit Type



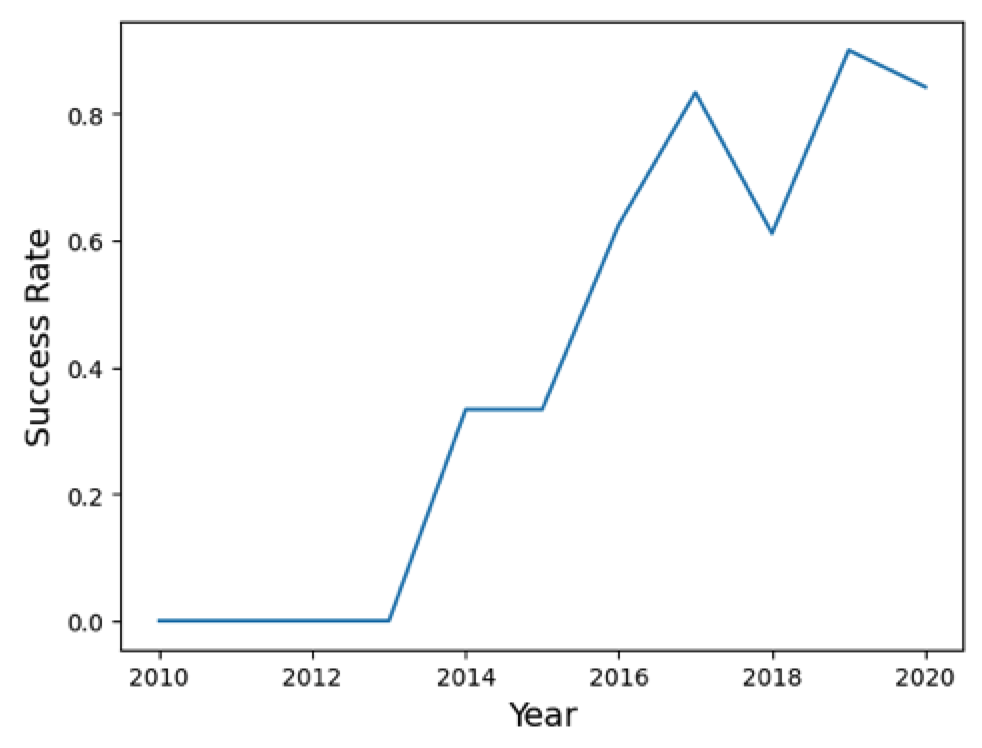
- LEO orbits appear to increase in success with the number of flights
- VLEO orbits have been more consistently successful
- GTO data is inconclusive, no obvious correlation

Payload vs. Orbit Type



- Polar, LEO and ISS have increased success/positive landing rates with higher payloads
- Again, GTO data is inconclusive, no apparent correlation.

Launch Success Yearly Trend



- The launch success rate has been increasing between 2013 - 2020

All Launch Site Names

- Using the keyword DISTINCT in our query, we can show unique launch site names.
- %sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE;

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

Launch Site Names Begin with 'CCA'

- To find 5 records where launch sites begin with `CCA`, we use:
  - LIKE and the % wildcard to search for Launch_Site values that match the pattern 'CCA%'
  - LIMIT 5 to limit the return results to five rows of data
- %sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" like 'CCA%' LIMIT 5

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

Total Payload Mass

- To calculate the total payload carried by boosters from NASA, we use:

  %sql SELECT Customer, sum(payload_mass__kg_) as total_payload FROM SPACEXTABLE WHERE Customer LIKE 'NASA (CRS)%';

Customer	total_payload
NASA (CRS)	48213

Average Payload Mass by F9 v1.1

- To calculate the average payload mass carried by booster version F9 v1.1

```
%sql select AVG(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version like 'F9 v1.1%';
```

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- To find the dates of the first successful landing outcome on ground pad

  %sql select Landing_Outcome, min(Date) from SPACEXTABLE where "Landing_Outcome" = "Success (ground pad)";

Landing_Outcome	min(Date)
Success (ground pad)	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- To list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
%%sql
select distinct Booster_Version, Landing_Outcome, payload_mass__kg_ as Payload
from spacextable
where Landing_Outcome = "Success (drone ship)"
    and (payload_mass__kg_ < 6000
    and payload_mass__kg_ > 4000)
order by payload_mass__kg_ asc;
```

Booster_Version	Landing_Outcome	Payload
F9 FT B1026	Success (drone ship)	4600
F9 FT B1022	Success (drone ship)	4696
F9 FT B1031.2	Success (drone ship)	5200
F9 FT B1021.2	Success (drone ship)	5300

Total Number of Successful and Failure Mission Outcomes

- To calculate the total number of successful and failure mission outcomes

```
%%sql
select
   case
      when Mission_Outcome like 'Success%' then 'Success'
      when Mission_Outcome like 'Failure%' then 'Failure'
   end as Outcome,
   count(*) as Total_Missions
from spacextable
group by Outcome;
```

Outcome	Total_Missions
Failure	1
Success	100

Boosters Carried Maximum Payload

- To list the names of the booster which have carried the maximum payload mass:

```
%%sql
select distinct Booster_Version, PAYLOAD_MASS__KG_
from spacextable
where PAYLOAD_MASS__KG_ = (select MAX(PAYLOAD_MASS__KG_) from spacextable)
order by Booster_Version;
```

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

2015 Launch Records

- To list the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%%sql
select
   case
      when substr(Date, 6, 2) = '01' then 'January'
      when substr(Date, 6, 2) = '02' then 'February'
      when substr(Date, 6, 2) = '03' then 'March'
      when substr(Date, 6, 2) = '04' then 'April'
      when substr(Date, 6, 2) = '05' then 'May'
      when substr(Date, 6, 2) = '06' then 'June'
      when substr(Date, 6, 2) = '07' then 'July'
      when substr(Date, 6, 2) = '08' then 'August'
      when substr(Date, 6, 2) = '09' then 'September'
      when substr(Date, 6, 2) = '10' then 'October'
      when substr(Date, 6, 2) = '11' then 'November'
      when substr(Date, 6, 2) = '12' then 'December'
   end as Month_Name,
   Landing_Outcome,
   Booster_Version,
   Launch_Site
from spacextable
where substr(Date,0,5)='2015' and Landing_Outcome like 'Failure (drone ship)'
;
```

Month_Name	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- To 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:

```
%%sql
select
   Landing_Outcome,
   count(*) as Outcome_Count,
   rank() over (order by count(*) desc) as Rank
from spacextable
where Date between '2010-06-04' and '2017-03-20'
group by Landing_Outcome
order by Outcome_Count desc
```

Landing_Outcome	Outcome_Count	Rank
No attempt	10	1
Success (drone ship)	5	2
Failure (drone ship)	5	2
Success (ground pad)	3	4
Controlled (ocean)	3	4
Uncontrolled (ocean)	2	6
Failure (parachute)	2	6
Precluded (drone ship)	1	8

Section 3

Launch Sites Proximities Analysis

SpaceX Launch Sites Studied

- Data pertaining to four launch sites was explored.

	Launch Site	Lat	Long
0	CCAFS LC-40	28.562302	-80.577356
1	CCAFS SLC-40	28.563197	-80.576820
2	KSC LC-39A	28.573255	-80.646895
3	VAFB SLC-4E	34.632834	-120.610745

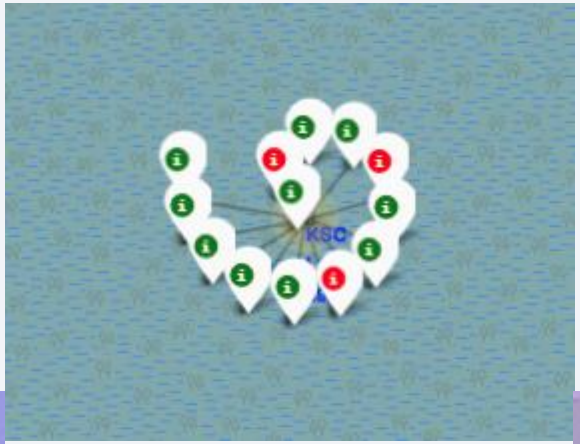
- Launch sites located near the coast
- One West coast site
- Three East coast sites
- Locations are in the south, relatively close proximity to the equator

Launch Outcome Cluster Markers

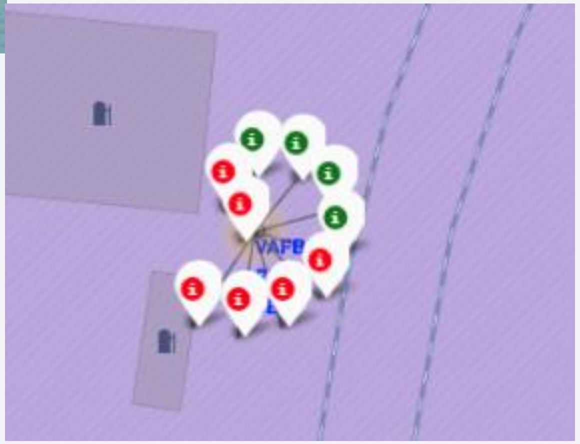
- By adding colored markers to the map for each launch outcome, we can visually identify which sites have the most launches and highest success rates

CCAFS LC-40

CCAFS SLC-40

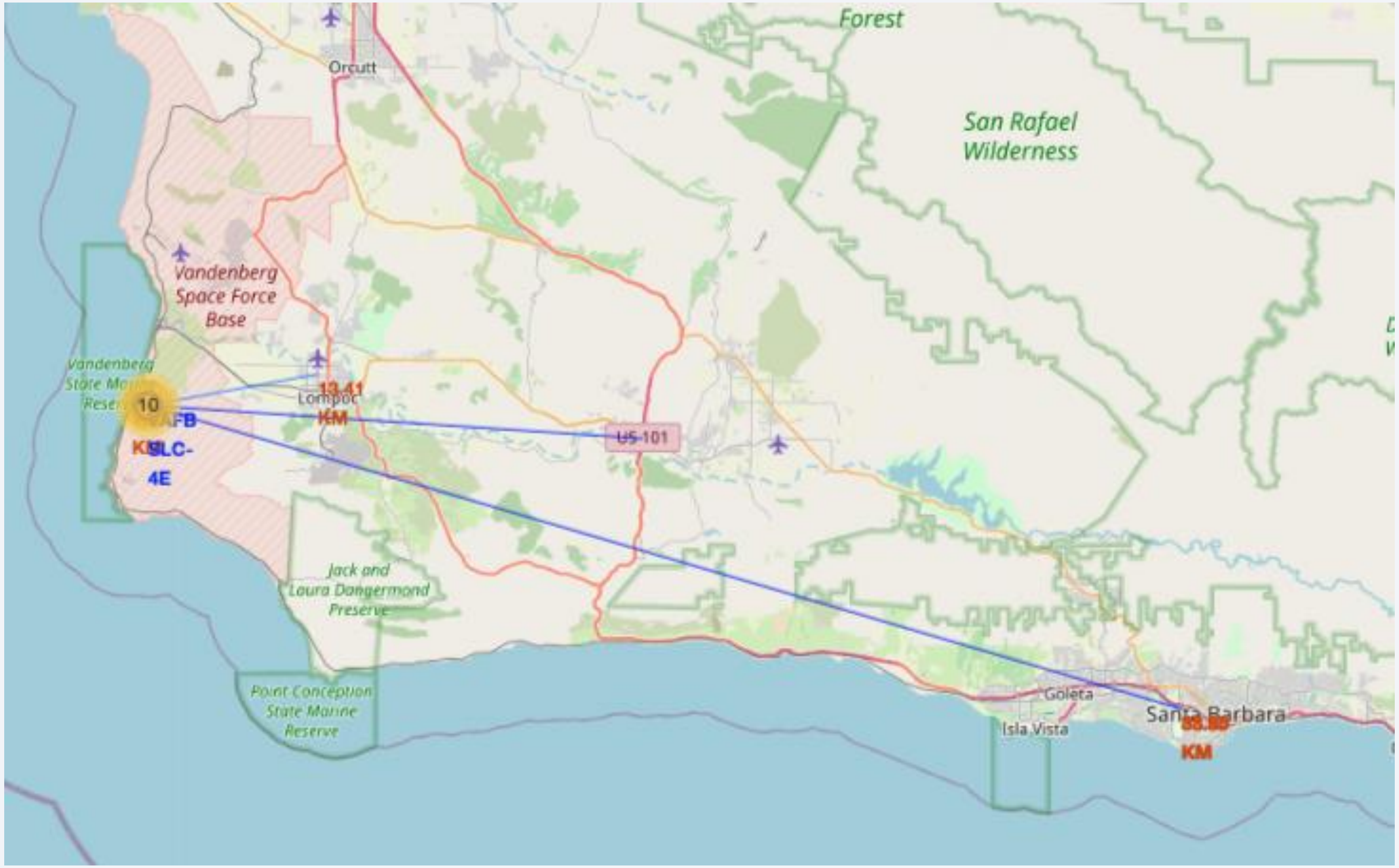


KSC LC-39A



VAFB SLC-4E

Proximity to Launch Site



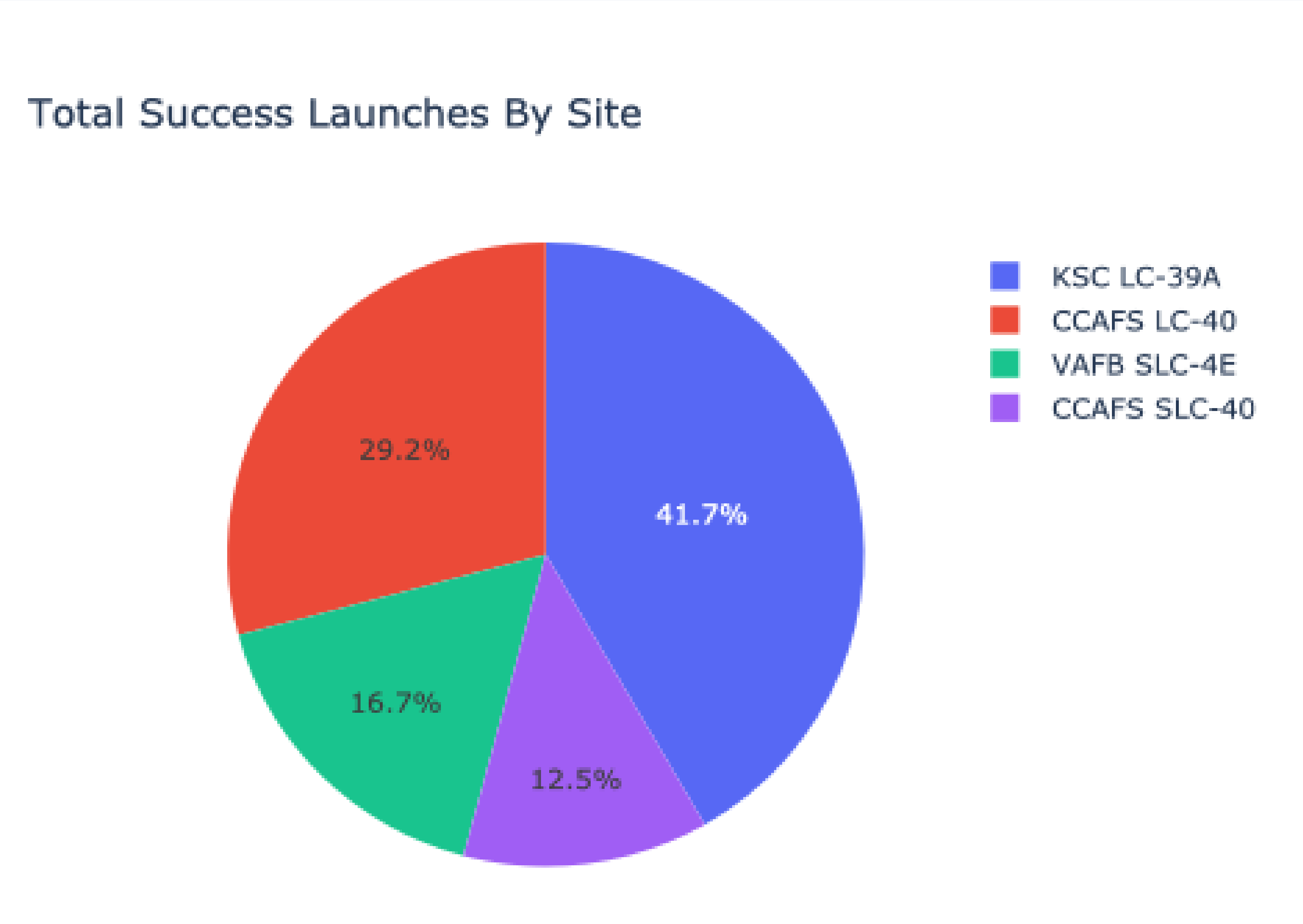
- Using the map we can visually locate landmarks, cities, resources, etc and the distance to each.
- For the site shown here, distances were determined to the nearest large city/town, to the nearest airport, and the nearest highway.

Section 4

Build a Dashboard with Plotly Dash

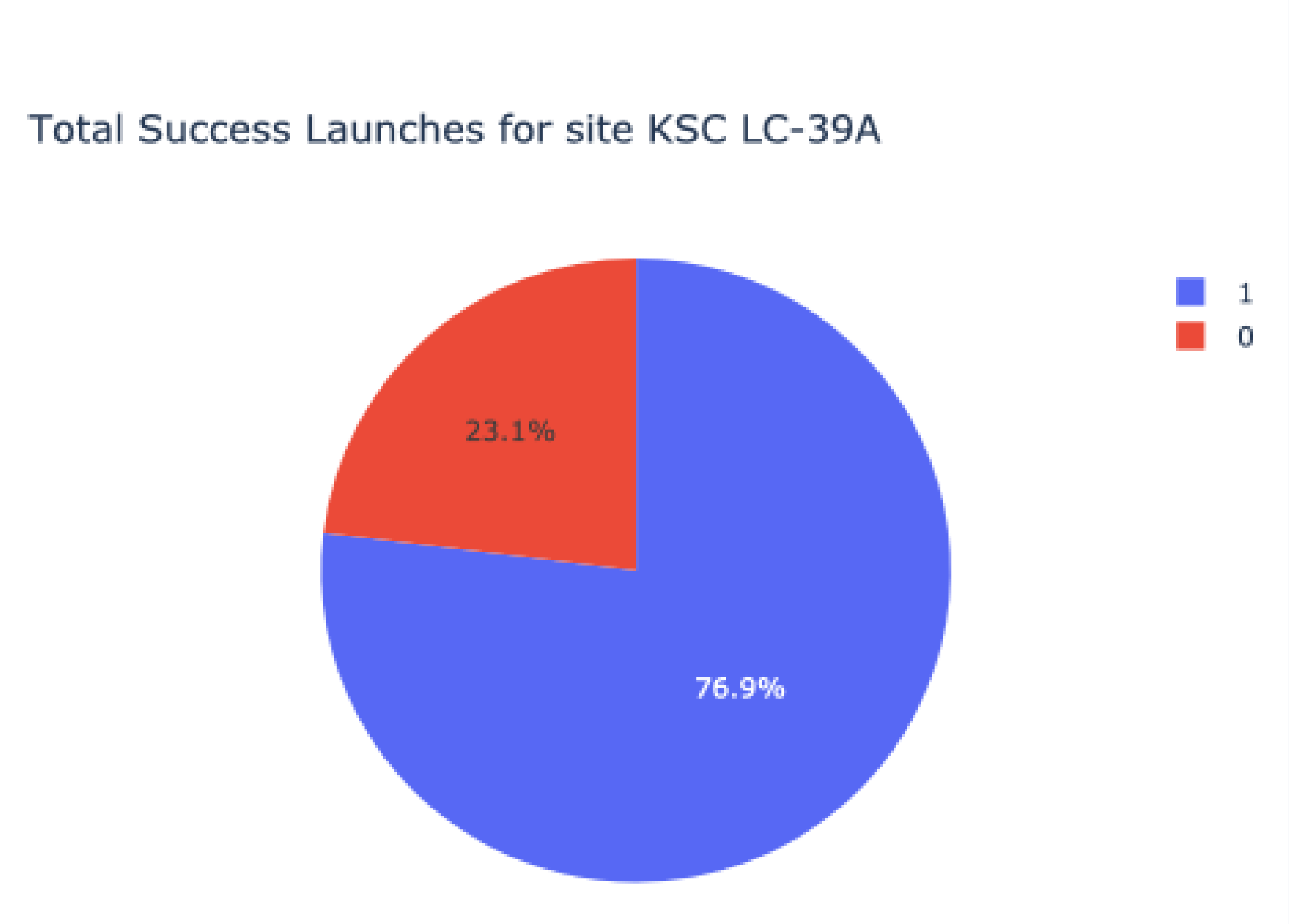
Total Launch Success Counts – All Sites

- KSC LC-39A had the highest count of successful launches.



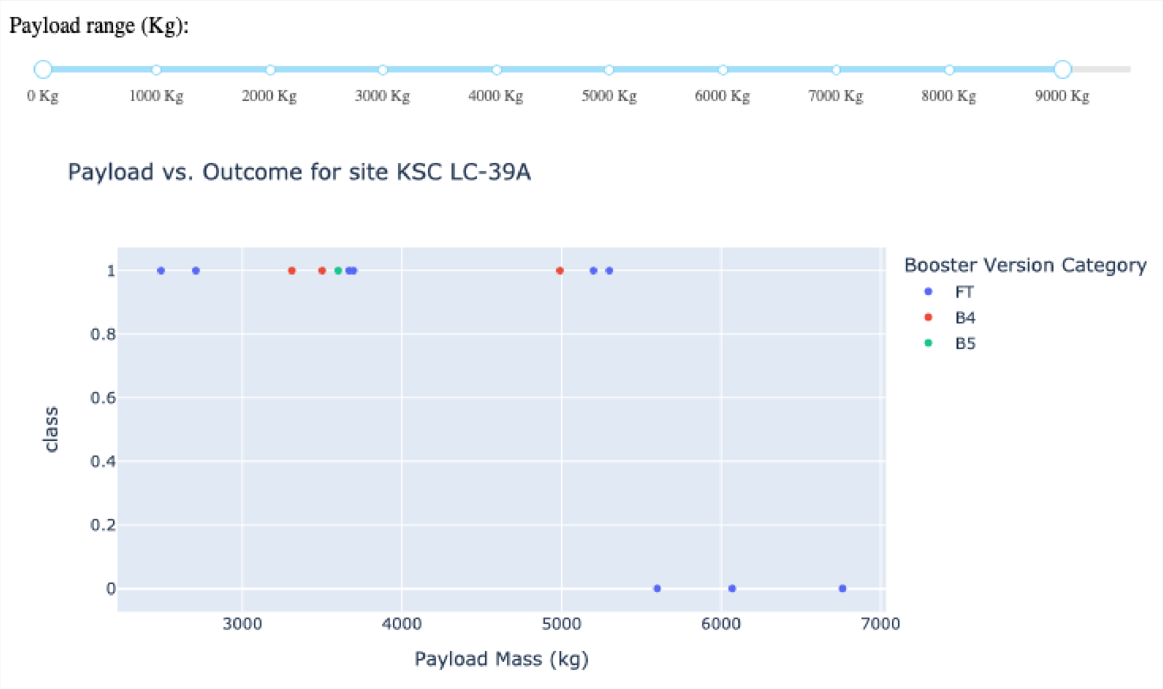
Highest Launch Success Ratio

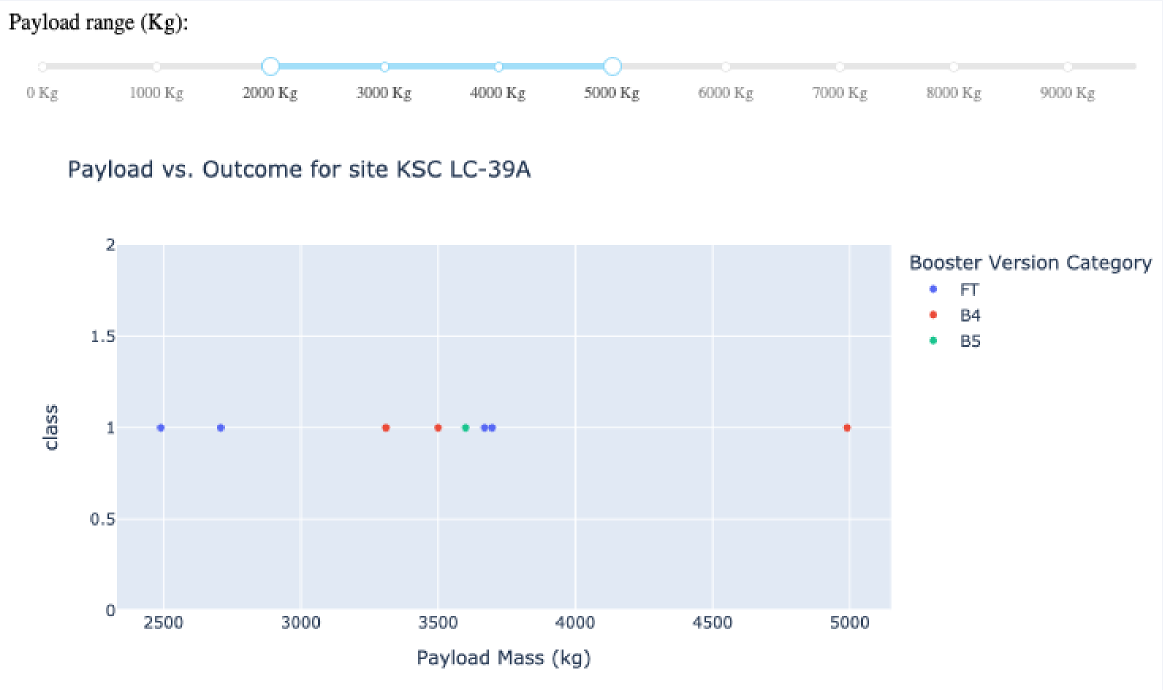
- KSC LS-39A achieved a launch success ratio of 76.9%



Payload to Success Correlation

- Using the slider to select various ranges of payloads, we can view the respective data for each booster version and the outcomes for the chosen range.



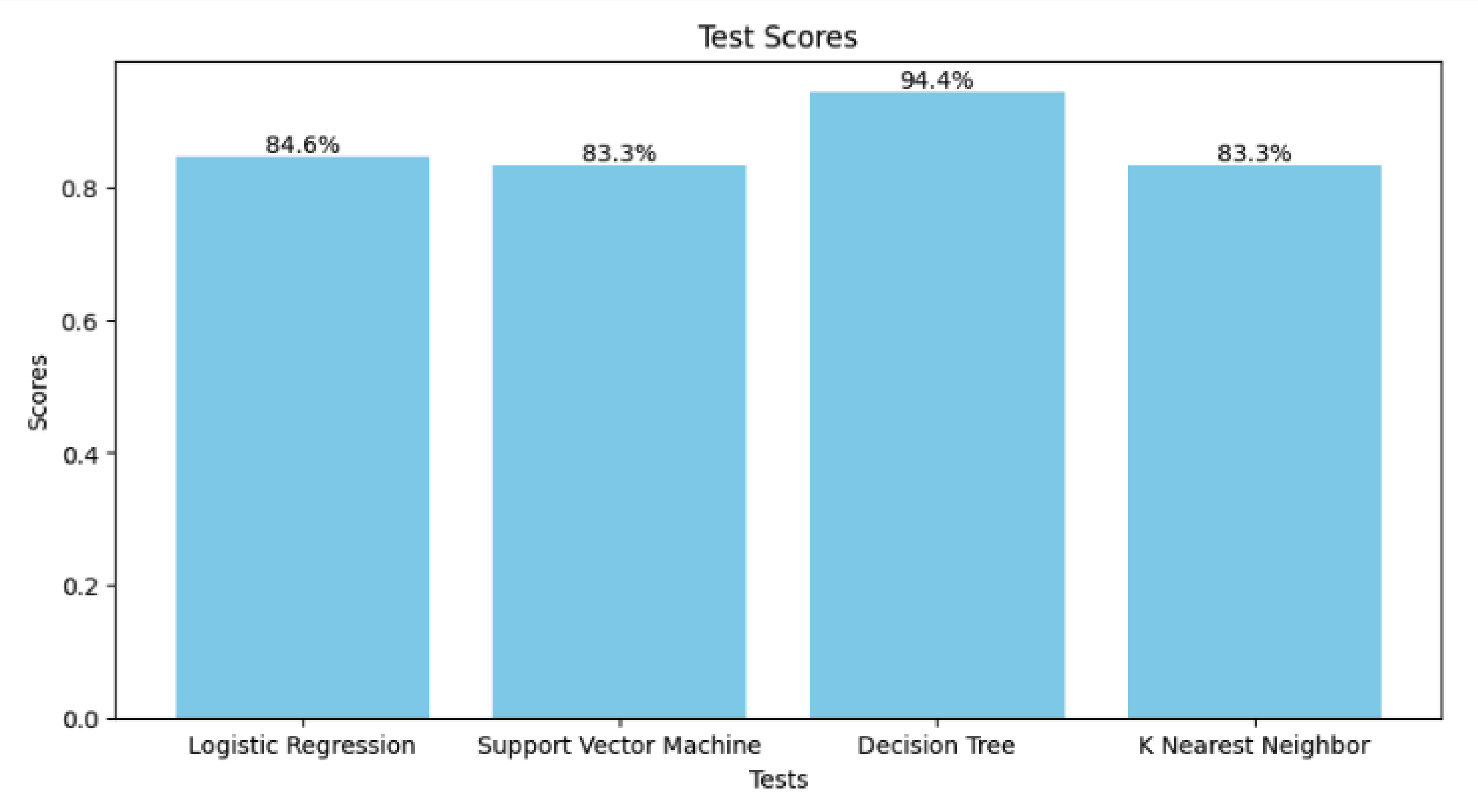


Section 5

Predictive Analysis (Classification)

Classification Accuracy

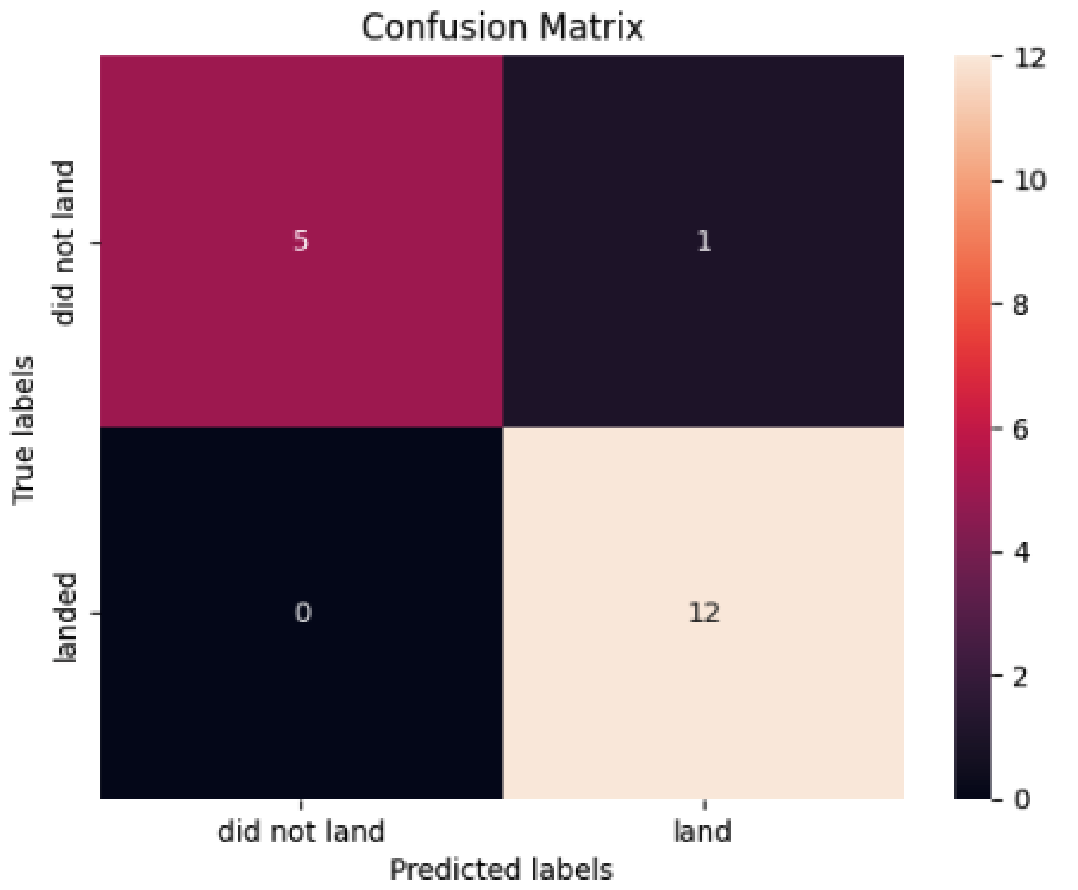
- Decision Tree model had the highest accuracy.



Confusion Matrix

- The confusion matrix compares the actual target values with those predicted by the model.
- The quadrants display True Negatives (TN = 5), False Positives (FP = 1), False Negatives (FN = 0), and True Positives (TP = 12)
- Accuracy:
  = (TP+TN)/(TP+TN+FP+FN)
  = (12+5)/(12+5+1+0)
  = 17/18
  ≈ 0.944 or 94.4%

```
yhat = tree_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



Conclusions

- Overall launch success has increased over time (2013-2020)
- As the number of launches increases, so does the success rate
- Most successful launch site is KSC LC-39A
- Launch success is greater for payloads over 10,000 kg
- Orbit types SSO (5 launches), ES-L1 (1 launch), GEO (1 launch), and HEO (1 launch) have the highest success rate (100%)
- Decision Tree classifier performed best at 94.44% accuracy (as opposed to other methods ranged 83%-88%)

Appendix

- Jupyter Notebooks and other material available on GitHub:

  [Space-Y-Project]

Thank you!