Winning Space Race with Data Science

Shaun Cruz
March 14 2022

Outline

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

Executive Summary

- Summary of methodologies
  - All methods performed using Python scripts
    - Data Collection through API and Web Scraping
    - Exploratory Data Analysis (EDA) via Data Wrangling, SQL, and Data Visualization
    - Visual Analytics interactively via Folium and Plotly Dash Dashboard
    - Machine Learning Prediction using scikit-learn (Support Vector Machines, Classification Trees, and Logistic Regression)

- Summary of all results
  - KSC LC-39A had the best success rate
  - Orbits ES-L1, GEO, HEO, SSO, VLEO had best success rate
  - Lighter payloads have higher success rate
  - All models predict the same, but the decision tree classifier worked best on the validation data

Introduction

- Project background and context
  - SpaceX Falcon 9 rocket launches cost 62 million dollars/launch vs Other providers at 165 million dollars/launch
  - Savings are due to reuse of the first stage.
  - If you can predict if SpaceX Falcon first stage will land, you can determine the total cost of SpaceX Falcon launch → Can other provider bid against SpaceX?
- Problems you want to find answers
  - Ultimate goal: Will the Falcon 9 first stage land successfully?
  - What factors (launch sites, payloads, F9 booster versions, etc.) affect success rate of launch?
  - Where are Launch sites located (in proximity to)?
  - What prediction model performs best: Support Vector Machines, Classification Trees, or Logistic Regression?

Section 1

Methodology

Methodology

Executive Summary

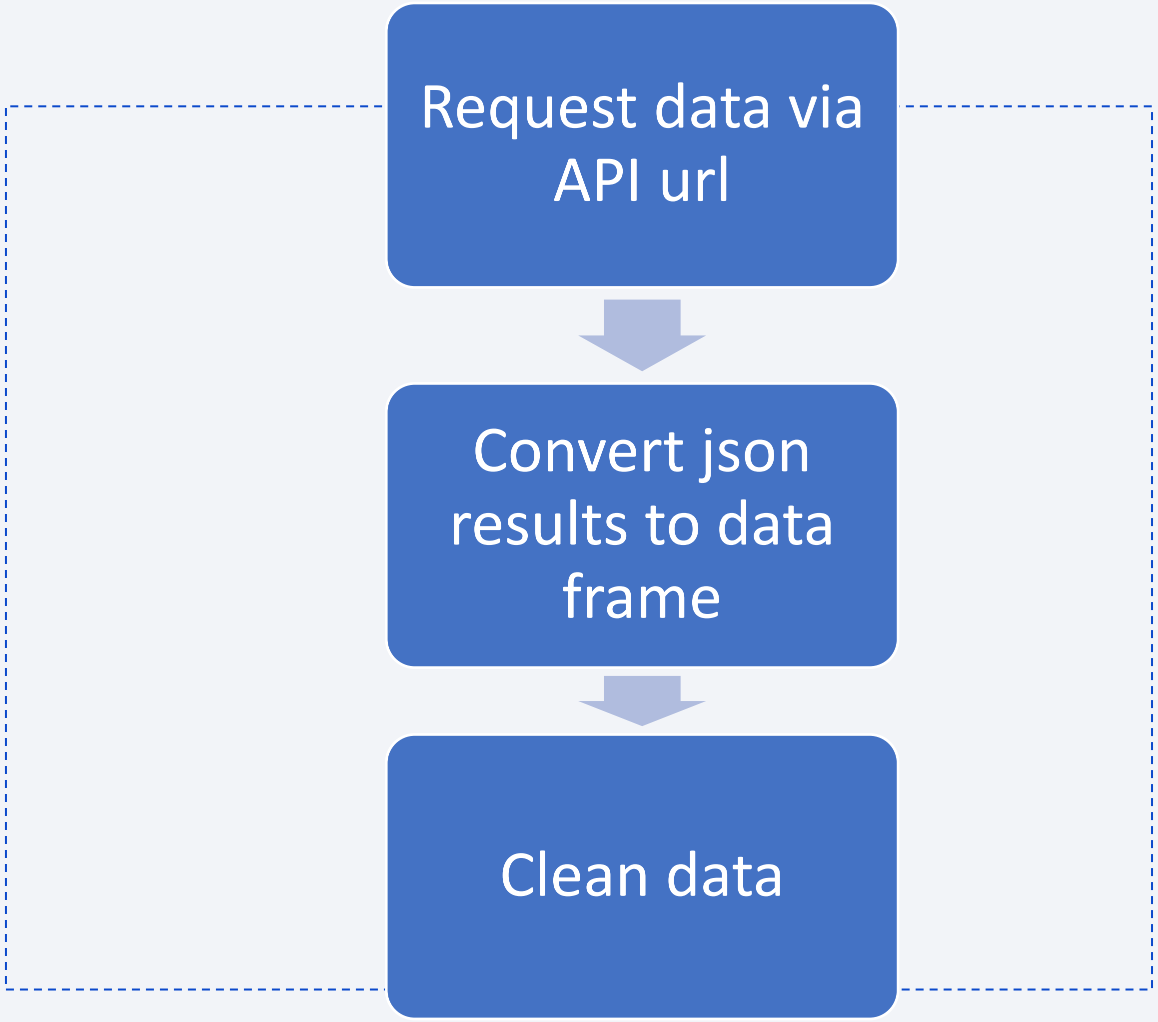
- Data collection methodology:
  - SpaceX API and Wikipedia (web scraping)
- Perform data wrangling
  - Data was cleaned and one-hot encoding was used on particular features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- SpaceX API
  - Requested from "https://api.spacexdata.com/v4/launches/past"
  - Requested and parsed JSON results using GET request
  - Decoded and turned into Pandas dataframe
  - Partitioned data into lists (BoosterVersion, PayloadMass, Orbit, etc.)
  - Combined all data into a dictionary, and filtered for only Falcon 9 launches
  - Cleaned data for any accordingly.
- Webscraping
  - Grabbed data from https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
    - Specifically from 9th June 2021 (https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922)
  - Used BeautifulSoup to extract and format data into dataframe, and converted into CSV

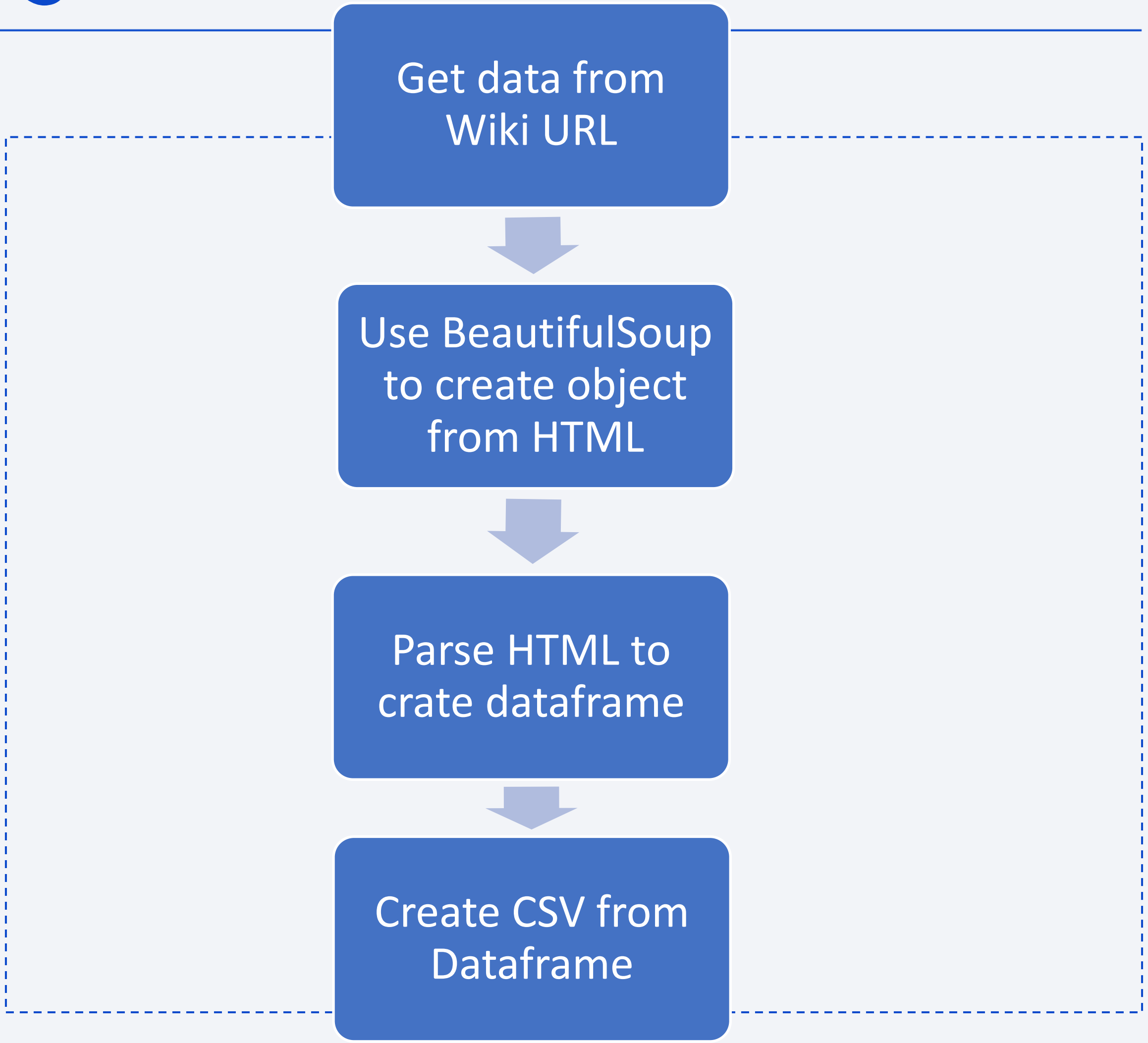
Data Collection – SpaceX API

- GitHub URL: https://github.com/shauncruz312/IBM-Data-Science-Capstone/blob/eb357d03a03f8baf950af4ca9168a611a3ad3e08/Data%20Collection%20API.ipynb



Data Collection - Scraping

- GitHub URL: https://github.com/shauncruz312/IBM-Data-Science-Capstone/blob/eb357d03a03f8baf950af4ca9168a611a3ad3e08/Data%20Collection%20with%20Web%20Scraping.ipynb



Data Wrangling

- GitHub URL: https://github.com/shauncruz312/IBM-Data-Science-Capstone/blob/eb357d03a03f8baf950af4ca9168a611a3ad3e08/EDA.ipynb

Calculate % of missing values for attributes

↓

Identify column types

↓

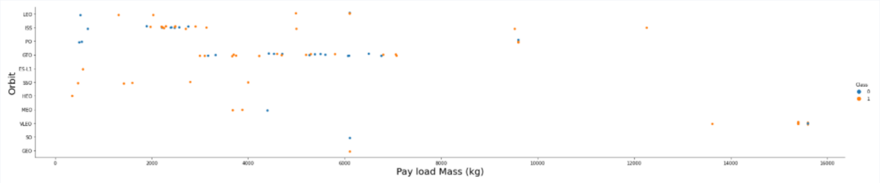
Calculate:

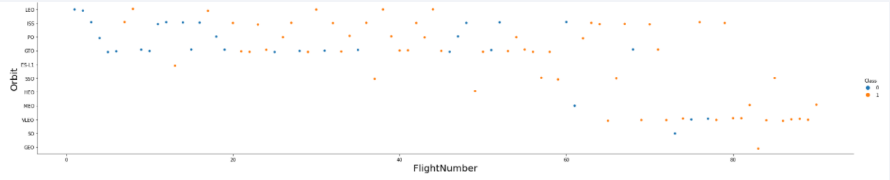
- Launches on each site
- Number and Occurrence of each orbit
- Number and Occurrence of mission outcome per orbit type

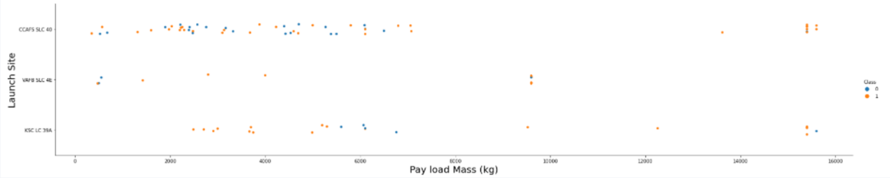
↓

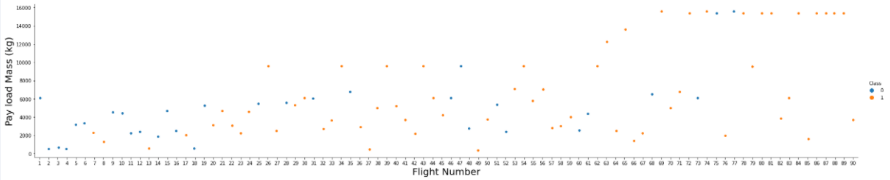
Create a landing outcome label from Outcome column

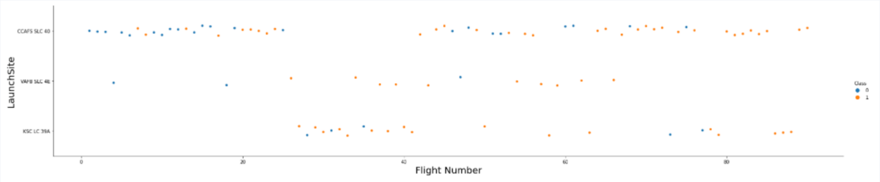
EDA with Data Visualization

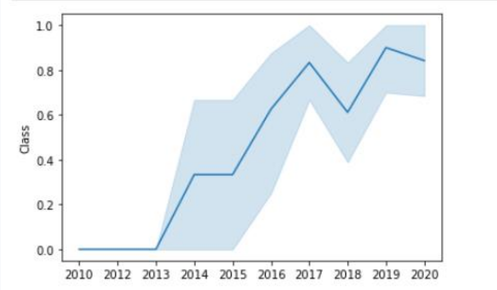




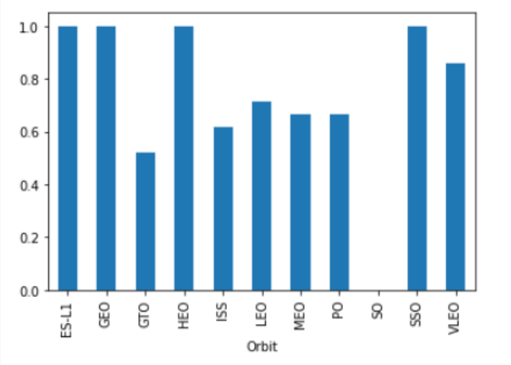








Launch Success Rate



Orbit Success Rate

- Various Plots made:
  - FlightNumber vs Payload Mass
  - FlightNumber vs LaunchSite
  - Payload vs LaunchSite
  - FlightNumber vs Orbit Type
  - Payload vs Orbit Type
  - Launch and Orbit Success Rate
- All plots to identify what features led to positive orbit/launch rate
- GitHub URL: https://github.com/shauncruz312/IBM-Data-Science-Capstone/blob/eb357d03a03f8baf950af4ca9168a611a3ad3e08/EDA%20with%20Visualization.ipynb

EDA with SQL

- SQL queries performed
  - SELECT * FROM HTB11006.SPACEXTBL LIMIT 5
  - SELECT DISTINCT(launch_site) FROM HTB11006.SPACEXTBL
  - SELECT DISTINCT(launch_site) FROM HTB11006.SPACEXTBL WHERE launch_site LIKE 'CCA%'
  - SELECT SUM(payload_mass__kg_) as total_payload_mass FROM HTB11006.SPACEXTBL WHERE CUSTOMER='NASA (CRS)'
  - SELECT AVG(payload_mass__kg_) as avg_payload_mass FROM HTB11006.SPACEXTBL WHERE BOOSTER_VERSION='F9 v1.1'
  - SELECT min(DATE) as min_date FROM HTB11006.SPACEXTBL WHERE LANDING__OUTCOME = 'Success (ground pad)'
  - SELECT BOOSTER_VERSION FROM HTB11006.SPACEXTBL WHERE LANDING__OUTCOME = 'Success (drone ship)' AND payload_mass__kg_ > 4000 AND payload_mass__kg_<6000
  - SELECT MISSION_OUTCOME, COUNT(*) as tot_num_missions FROM HTB11006.SPACEXTBL GROUP BY MISSION_OUTCOME
  - SELECT BOOSTER_VERSION, PAYLOAD_MASS__KG_ FROM HTB11006.SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from HTB11006.SPACEXTBL)
  - SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM HTB11006.SPACEXTBL WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND DATE LIKE '2015%'
  - SELECT LANDING__OUTCOME, count(*) as num_outcomes FROM (SELECT * FROM HTB11006.SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20') GROUP BY LANDING__OUTCOME ORDER BY num_outcomes DESC

- GitHub URL: https://github.com/shauncruz312/IBM-Data-Science-Capstone/blob/eb357d03a03f8baf950af4ca9168a611a3ad3e08/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Marked launch sites with success/failure indications
- Assigned launch outcomes (failure=0, 1=success)
- Identified what sites have high success rate for launches
- Calculated distances between launch site to coastline, city, railroad, and highway
  - Launch sites are always near railroads, highways, and coastlines
  - Launch sites are away from cities.
- GitHub URL: https://github.com/shauncruz312/IBM-Data-Science-Capstone/blob/eb357d03a03f8baf950af4ca9168a611a3ad3e08/Interactive%20Visual%20Analytics%20with%20Folium.ipynb
- IBM URL (since Folium doesn't work in GIT) https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/467fb4d0-9937-4d9f-8cbe-7294c408d3b9/view?access_token=fb58eddd8633997dc703a5bc1eb2c692eda0ba12a0e847a7e9e8cbf565129509

Build a Dashboard with Plotly Dash

- Pie chart showing total launch by site
- Scatter plot of Outcome vs Payload Mass for different Booster Versions
- This gives us an understanding of successes and locations
- GitHub URLS:
  - https://github.com/shauncruz312/IBM-Data-Science-Capstone/blob/eb357d03a03f8baf950af4ca9168a611a3ad3e08/spacex_dash_app.py
  - https://github.com/shauncruz312/IBM-Data-Science-Capstone/blob/eb357d03a03f8baf950af4ca9168a611a3ad3e08/SpaceXLaunchREcordsDashboard.PNG

Predictive Analysis (Classification)

- GitHub URL: https://github.com/shauncruz312/IBM-Data-Science-Capstone/blob/eb357d03a03f8baf950af4ca9168a611a3ad3e08/Machine%20Learning%20Prediction.ipynb

Load Data and standardize

↓

Split data into training/test data

↓

Train/Test and calculate accuracy (and Confusion Plot) for Logistic Regression, SVM, and Classification Tree, and KNN

↓

Compare Results

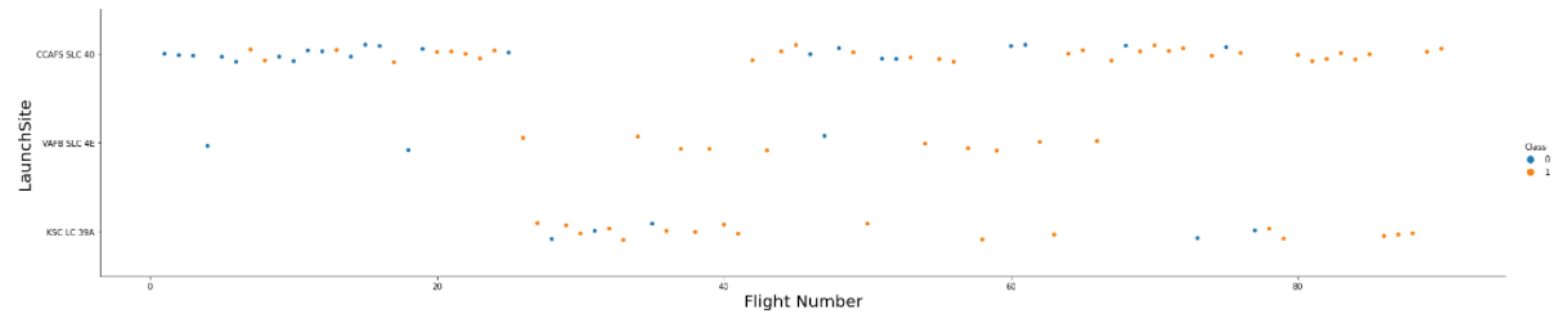
Results

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

Section 2

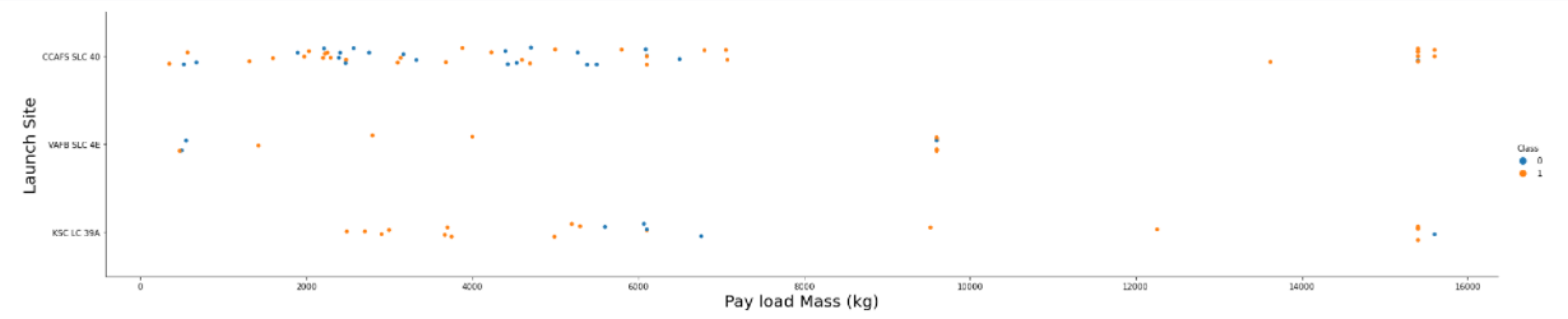
Insights drawn from EDA

Flight Number vs. Launch Site



- The higher the flight numbers at CCAFS SLC 40, the more successful. KSC LC 39A and VAFB SLC 4E generally had good success rate, but smaller sample size

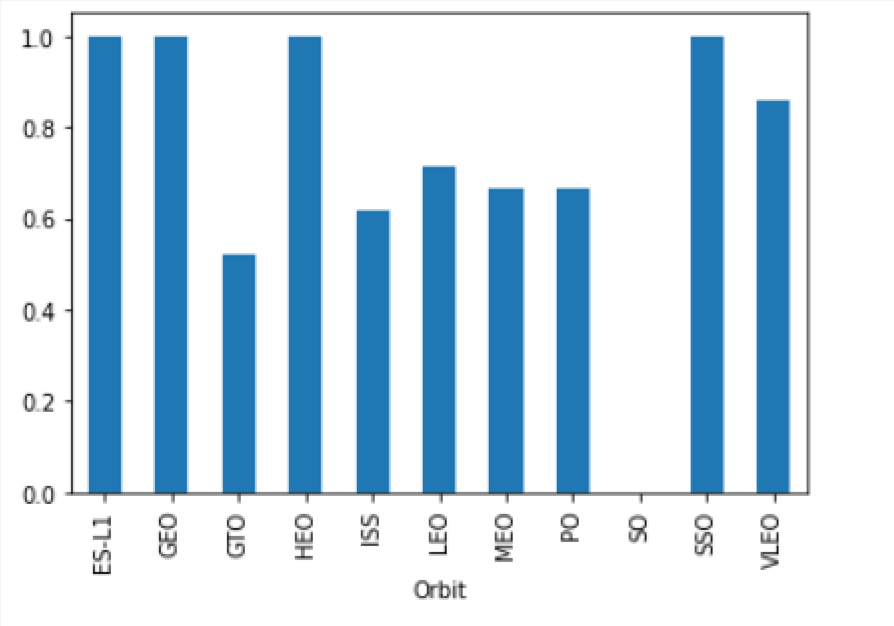
Payload vs. Launch Site



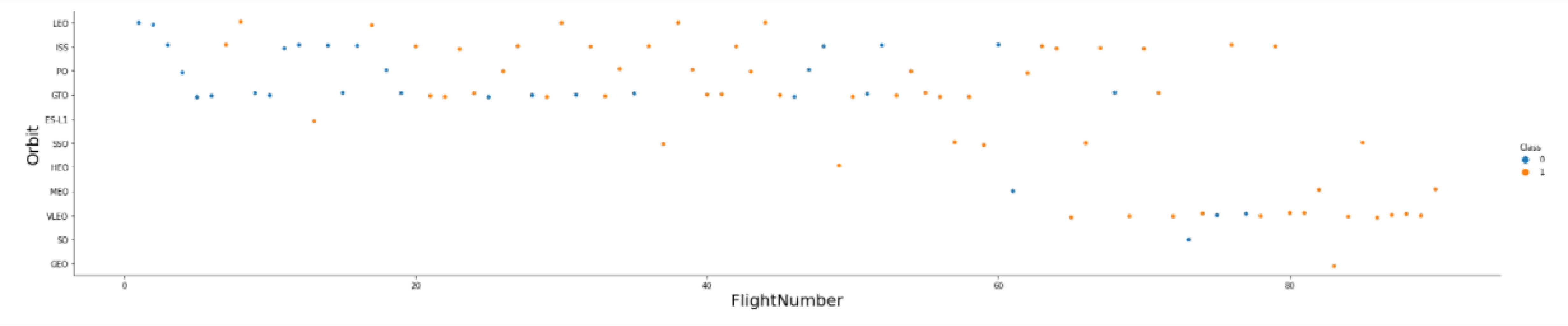
- No rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO have the highest success rate.

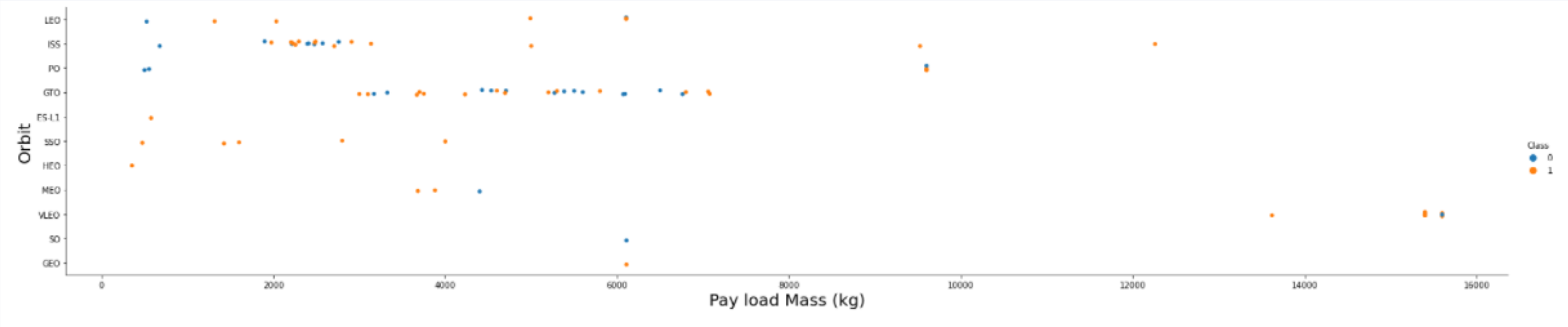


Flight Number vs. Orbit Type



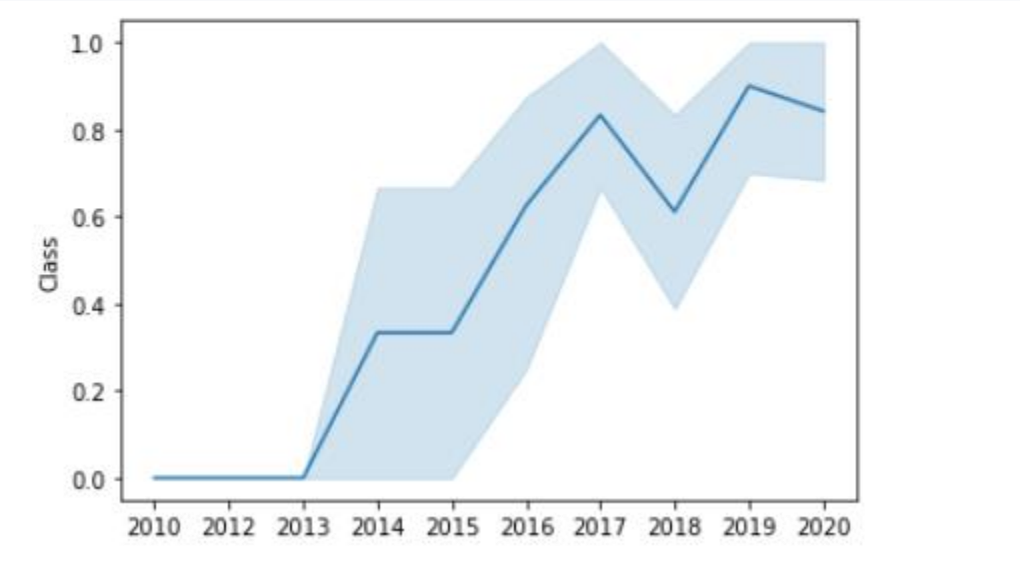
- the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



- 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



- the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

- SELECT DISTINCT(launch_site) FROM HTB11006.SPACEXTBL
- Distinct(launch_site) only shows unique names from the table

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

Launch Site Names Begin with 'CCA'

- SELECT DISTINCT(launch_site) FROM HTB11006.SPACEXTBL WHERE launch_site LIKE 'CCA%’
- Adding “LIKE ‘CCA%’ chooses launch site names starting with CCA. In order to choose 5 records, remove “DISTINCT” and add LIMIT=5 after WHERE.

launch_site
CCAFS LC-40
CCAFS SLC-40

Total Payload Mass

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

```
In [14]: %sql SELECT SUM(payload_mass__kg_) as total_payload_mass FROM HTB11006.SPACEXTBL WHERE CUSTOMER='NASA (CRS)'

 * ibm_db_sa://htb11006:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32459/bludb
Done.
```

Out[14]:

total_payload_mass
45596

- Find customer = NASA (CRS) and sum the payload to get the total

Average Payload Mass by F9 v1.1

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

```
]: %sql SELECT AVG(payload_mass__kg_) as avg_payload_mass FROM HTB11006.SPACEXTBL WHERE BOOSTER_VERSION='F9 v1.1'
```

```
 * ibm_db_sa://htb11006:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32459/bludb
Done.
```

]:

avg_payload_mass
2928

- Find where Booster is ‘F9 v1.2’ and then average the payload mass of the results

First Successful Ground Landing Date

- Look for when “Success (ground pad)” was the LANDING_OUTCOME, and find the earliest date.

```
%sql SELECT min(DATE) as min_date FROM HTB11006.SPACEXTBL WHERE LANDING__OUTCOME = 'Success (ground pad)'
```

```
 * ibm_db_sa://htb11006:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32459/bludb
Done.
```

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 6000

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 BOOSTER_VERSION FROM HTB11006.SPACEXTBL WHERE LANDING__OUTCOME = 'Success (drone ship)' AND payload_mass__kg_ > 4000 AND payload_mass__kg
```

```
 * ibm_db_sa://htb11006:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32459/bludb
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

- Group by mission outcome and then count the amounts to get total numbers

List the total number of successful and failure mission outcomes

```
%sql SELECT MISSION_OUTCOME, COUNT(*) as tot_num_missions FROM  HTB11006.SPACEXTBL GROUP BY MISSION_OUTCOME
```

```
 * ibm_db_sa://htb11006:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32459/bludb
Done.
```

mission_outcome	tot_num_missions
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

%sql SELECT BOOSTER_VERSION, PAYLOAD_MASS__KG_ FROM HTB11006.SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from HTB11006.SPACEXTBL)

Create a subquery for maximum payload mass and select the booster version from subquery to see which boosters have carried the max payload.

Task 8

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

```
5]: %sql SELECT BOOSTER_VERSION, PAYLOAD_MASS__KG_ FROM HTB11006.SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from HTB11006.SPACEXT
```

```
 * ibm_db_sa://htb11006:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32459/bludb
Done.
```

5]:

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

2015 Launch Records

- Choose 2015 as date failure as outcome, and list the outcome and booster version and launch site

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

```
%sql SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM HTB11006.SPACEXTBL WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND DATE LIKE '201
```

```
 * ibm_db_sa://htb11006:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32459/bludb
Done.
```

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- Create sub query of dates between June 4 2010 and March 20 2017, and then rank the counts of each outcome.

Task 10

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 num_outcomes FROM (SELECT * FROM HTB11006.SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20') GROUP
```

```
 * ibm_db_sa://htb11006:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32459/bludb
Done.
```

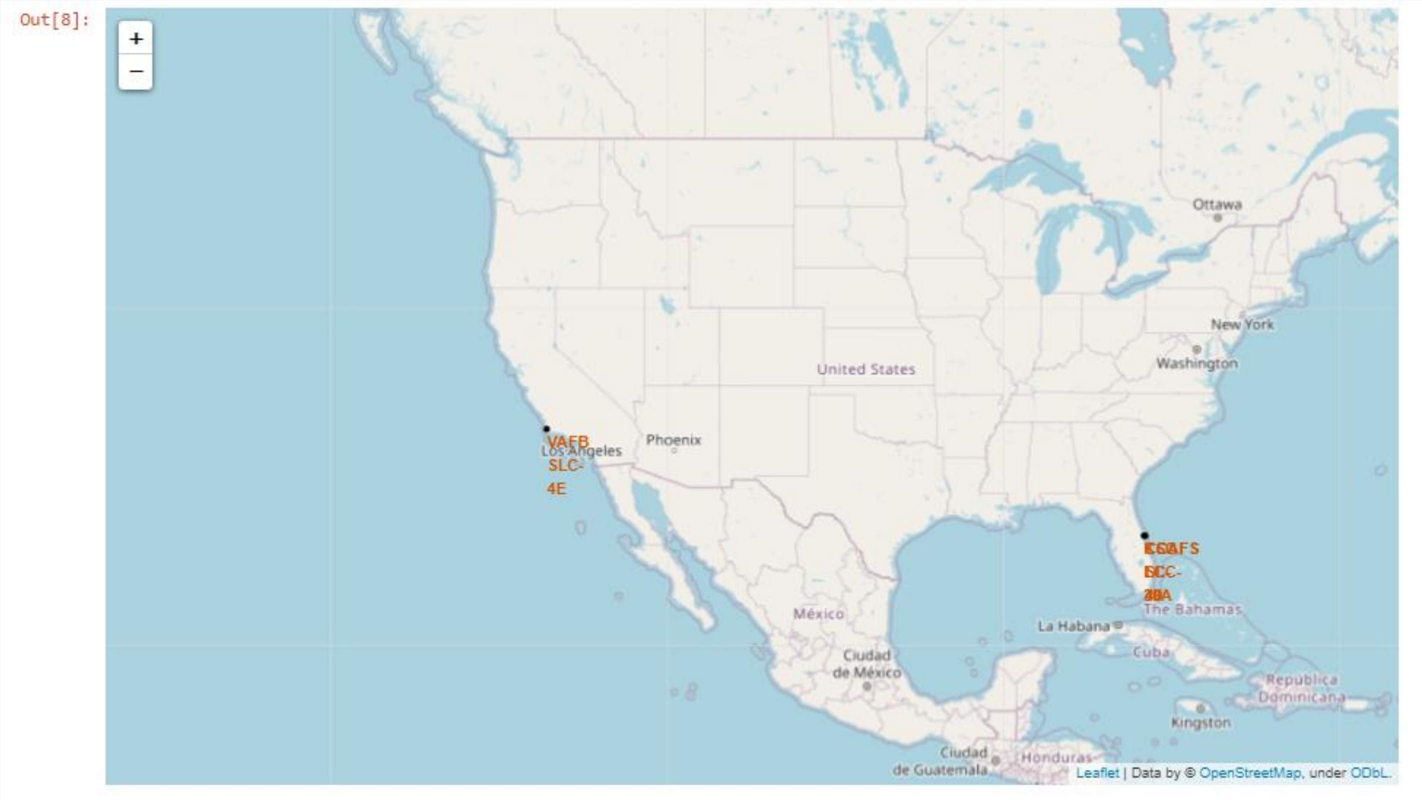
landing__outcome	num_outcomes
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

Launch Sites Locations on Global Map

- Are all launch sites in proximity to the Equator line?
  - No, none of the sites are in proximity to the Equator line. Florida has the lowest ones and that isnt near the equator.
- Are all launch sites in very close proximity to the coast?
  - Yes, all launch sites are in proximity to the coast, they are on teh border of Florida and California.

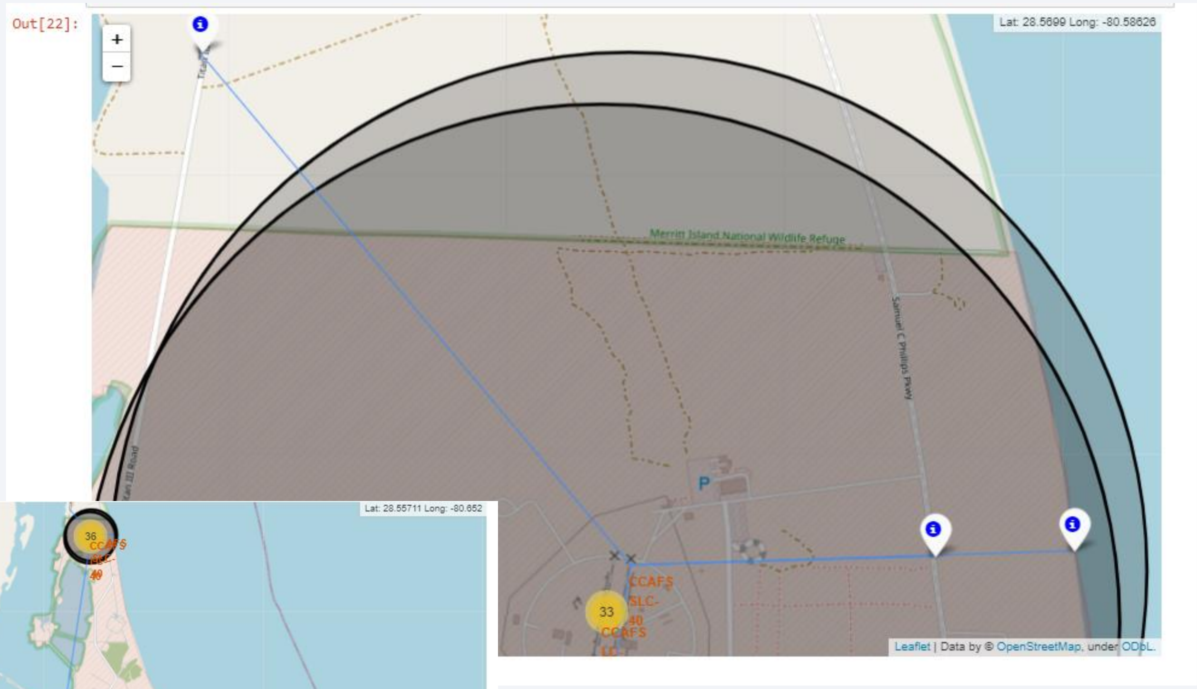


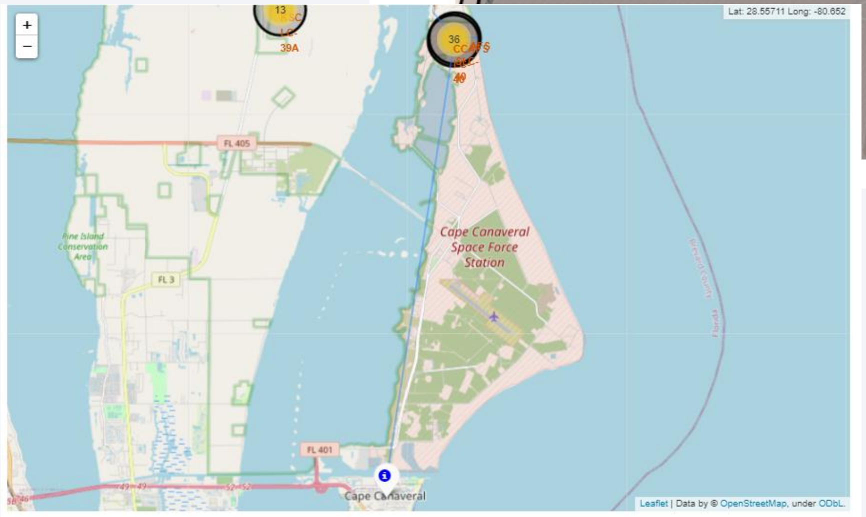
Color Labled Launch Outcomes

- Green is successful, Red is failure

CCAFS Distance to Railway, Highway, Coastline, and City

- Distance to
  - Railway: 1.291
  - Highway: 05824
  - Coastline: 0.8513
  - City: 18.2044
- Close to railway, highway, coastline, FAR from city

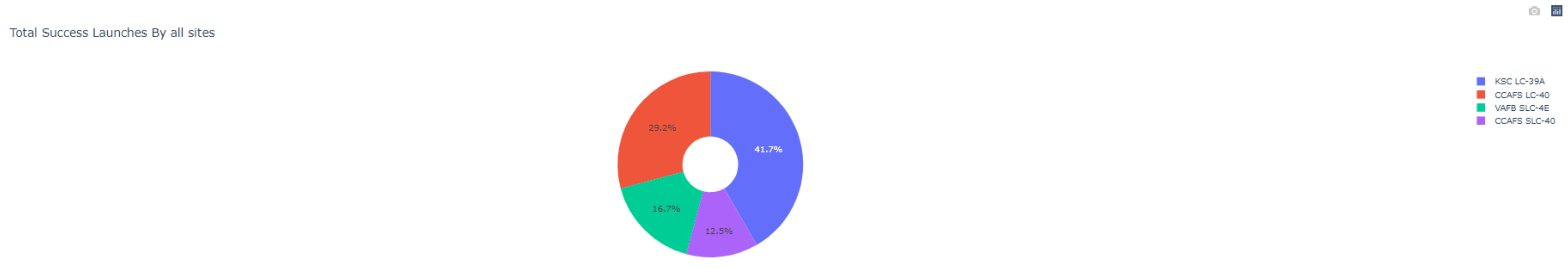




Section 4

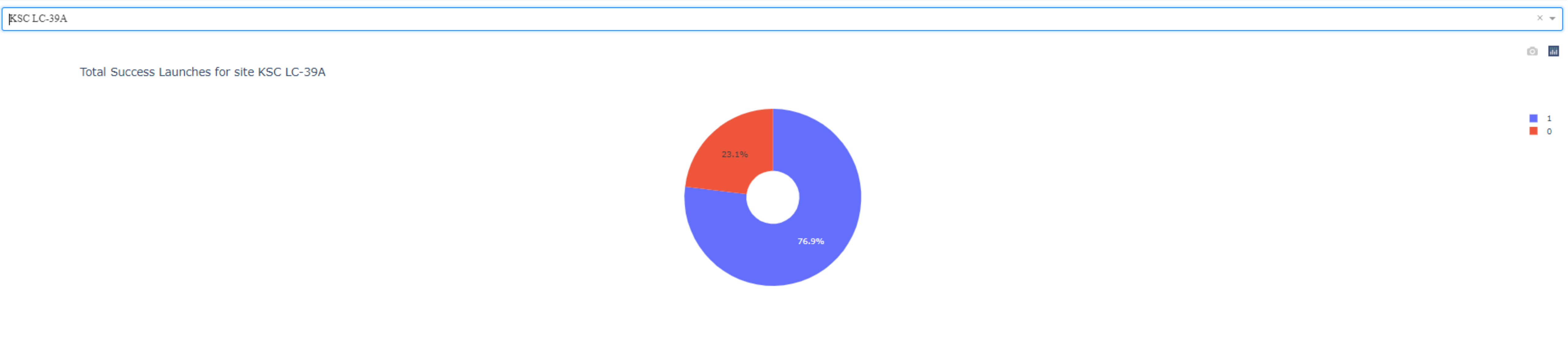
Build a Dashboard with Plotly Dash

Launch Success Count (PieChart)



- KSC LC-39A has highest success rate

Launchsite with Highest Success Ratio (KSC LC 39A)



- KSC LC 39A had 23.1% Failures and 76.9% Successes

Payload vs Launch Outcome scatter plots

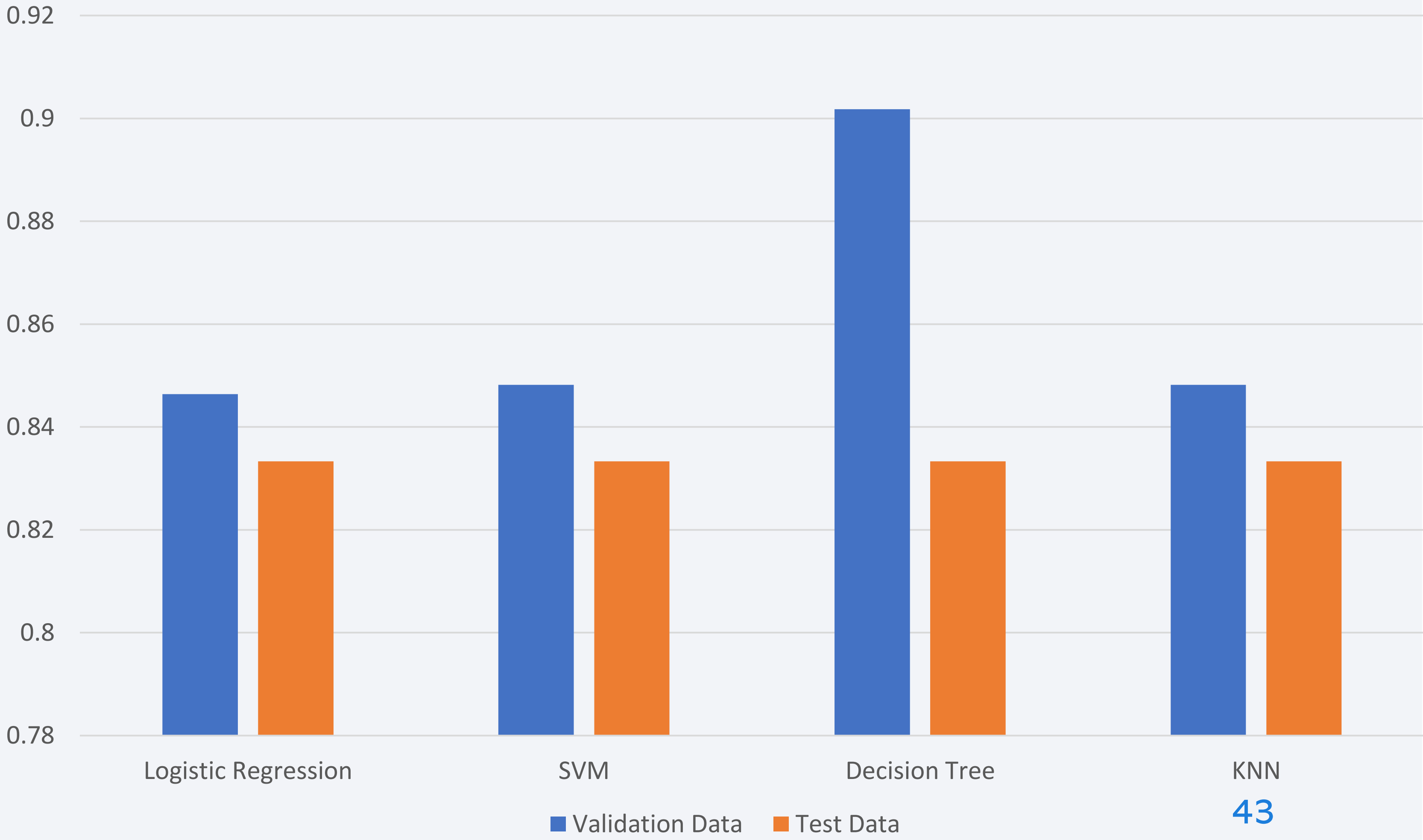
- TOP: 0kg-5000kg
- Right: 5000kg-10000kg
- Shows better success rate for lower payload mass

Section 5

Predictive Analysis (Classification)

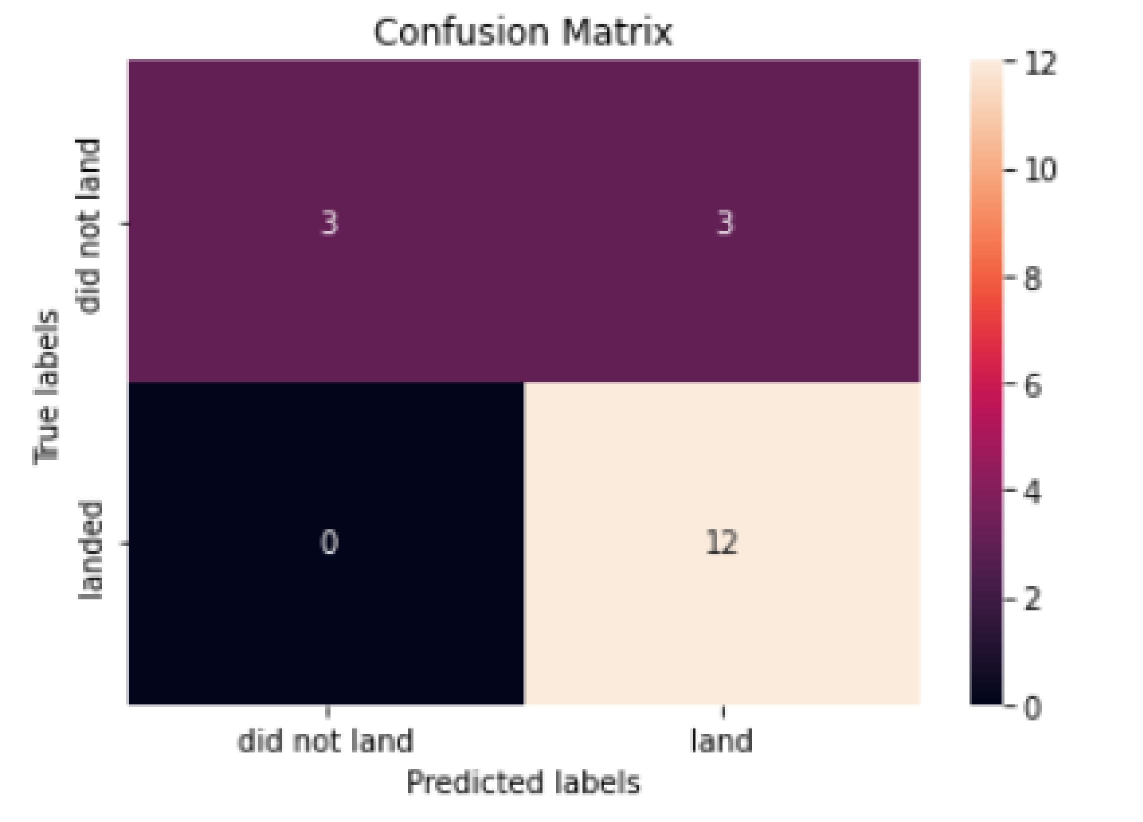
Classification Accuracy

- Logistic Regression Accuracy
  - Validation data: 0.8464
  - Test data: 0.8333
- SVM Accuracy
  - Validation Data: 0.84821
  - Test data: 0.8333
- Decision Tree Classifier Accuracy
  - Validation Data: 0.901785
  - Test data: 0.83333
- KNN Accuracy
  - Validation Data: 0.84821
  - Test data: 0.8333



Confusion Matrix for Decision Tree

- Although all models performed the same on the test data, the Decision Tree had the best on the validation step.
- Predicted accurately
  - 3/6 did not land
  - 12/12 landed



Conclusions

- KSC LC-39A had the best success rate
- Orbits ES-L1, GEO, HEO, SSO, VLEO had best success rate
- Lighter payloads have higher success rate
- All models predict the same, but the decision tree classifier worked best on the validation data

Thank you!