IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Hashini Liyanage
10th Nov 2022

Outline

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

Executive Summary

- **Summary of methodologies**
  - Data Collection
    - API
    - Web scraping
  - Data Wrangling
  - Exploratory Data Analysis
    - SQL
    - Data Visualization
  - Interactive Visual Analytics
    - Map with Folium
    - Dashboards with dash
  - Predictive Analysis
    - Machine Learning

- **Summary of all results**
  - Exploratory Data Analysis
  - Interactive Visual Analytics
  - Predictive Analysis

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. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

  The goal of project is to determine if the first stage will land successfully.

- **Problems you want to find answers**
  - Factors that determine successfully landing
  - Relationship amongst various features that determine the success rate of landing
  - Conditions needs to be in place to ensure a successful landing

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX REST API
  - Web Scraping from Wikipedia
- Perform data wrangling
  - Drop unwanted columns and replace missing values
  - One-hot encoding for categorical fields (ML)
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash

Methodology

Executive Summary

- Perform predictive analysis using classification models
  - LR, KNN, SVM, DT models build and evaluated
  - Data was normalized
  - Divided in to training and test data sets
  - Evaluated by four different classification models
  - Each model evaluated using different combinations of parameters.

Data Collection

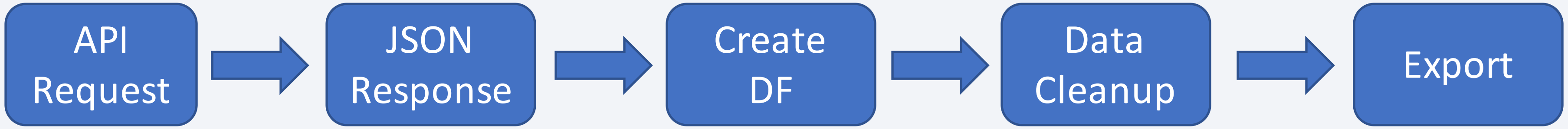
Datasets are collected from API and Web Scraping are rocket, launches, payload

1. Rest SpaceX API

   https://api.spacexdata.com/v4/

2. Web scrapping from Wikipedia
   https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Data Collection – SpaceX API



[Link to Code]

1. Requesting data from SpaceX API

```
response = requests.get(spacex_url)
```

2. Decode response content as a Json turn it into a Pandas dataframe

```
data = pd.json_normalize(response.json())
```

3. Apply custom functions > get usefull information > assign to list

```
getBoosterVersion(data)
getLaunchSite(data)
getCoreData(data)
getPayloadData(data)
```

```
BoosterVersion = []
PayloadMass = []
Orbit = []
LaunchSite = []
Outcome = []
Flights = []
GridFins = []
Reused = []
Legs = []
LandingPad = []
Block = []
ReusedCount = []
Serial = []
Longitude = []
Latitude = []
```

4. Create Dictionary

```
launch_dict = {'FlightNumber': list(data['flight_number']),
'Date': list(data['date']),
'BoosterVersion':BoosterVersion,
'PayloadMass':PayloadMass,
'Orbit':Orbit,
'LaunchSite':LaunchSite,
'Outcome':Outcome,
'Flights':Flights,
'GridFins':GridFins,
'Reused':Reused,
'Legs':Legs,
'LandingPad':LandingPad,
'Block':Block,
'ReusedCount':ReusedCount,
'Serial':Serial,
'Longitude': Longitude,
'Latitude': Latitude}
```

5. Create new Dataframe

```
data1 = pd.DataFrame(launch_dict)
```

6. Filter the dataframe

```
data_falcon9 = data1[data1['BoosterVersion'] != 'Falcon 1']
```

7. Export

```
data_falcon9.to_csv('dataset_part_1.csv', index=False)
```

Data Collection - Scraping

HTTP GET Request → BeautifulSoup object from the HTML response → Create DF → Data Cleanup → Export

[Link to Code]

1. HTTP GET method to request the data

```
data = requests.get(static_url).text
```

2. Create a BeautifulSoup object from the HTML response

```
soup = BeautifulSoup(data, 'html5lib')
```

3. Extract all column names from the HTML table header

```
html_tables = soup.find_all('table')
```

4. Create dictionary

```
launch_dict= dict.fromkeys(column_names)

# Remove an irrelvant column
del launch_dict['Date and time ( )']

# Let's initial the launch_dict with each
launch_dict['Flight No.'] = []
launch_dict['Launch site'] = []
launch_dict['Payload'] = []
launch_dict['Payload mass'] = []
launch_dict['Orbit'] = []
launch_dict['Customer'] = []
launch_dict['Launch outcome'] = []
# Added some new columns
launch_dict['Version Booster']=[]
launch_dict['Booster landing']=[]
launch_dict['Date']=[]
launch_dict['Time']=[]
```

5. Extract data & append to dictionary

```
extracted_row = 0
fligtnum = []
#Extract each table
for table_number,table in enumerate(soup.
   # get table row
    for rows in table.find_all("tr"):
        #check to see if first table head
        if rows.th:
            if rows.th.string:
                flight_number=rows.th.str
                flag=flight_number.isdigi
        else:
            flag=False
        #get table element
        row=rows.find_all('td')
```

6. Create DF

```
df = pd.DataFrame(launch_dict)
df
```

7.Export

```
df.to_csv('spacex_web_scraped.csv', index=False)
```

Data Wrangling

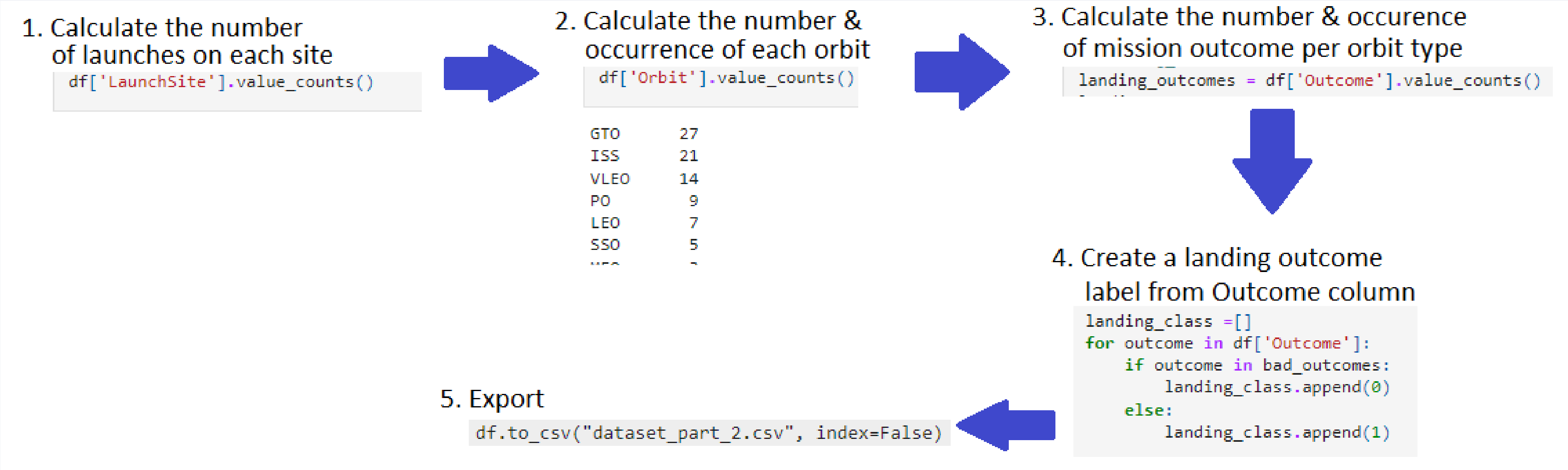
In the data set, there are several different cases where the booster did not land successfully.

- True Ocean, True RTLS, True ASDS - successful
- False Ocean, False RTLS, False ASDS - Unsuccessful

Added classification variable that represents the outcome of each launch.

- 1 means the booster successfully landed
- 0 means it was unsuccessful.

<u>**Link to Code**</u>

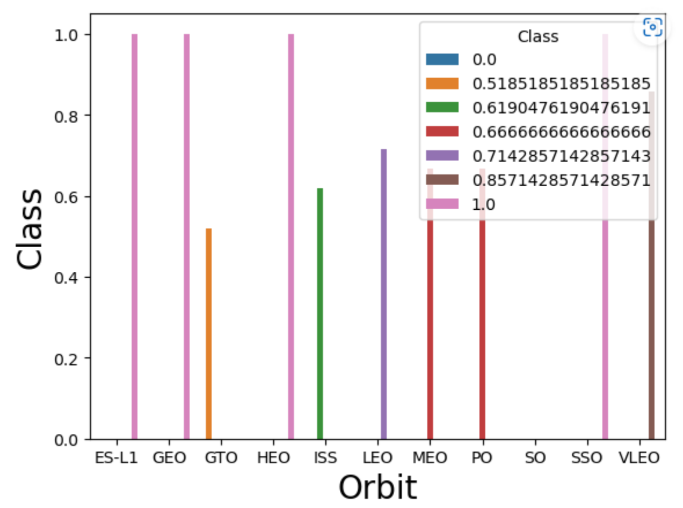


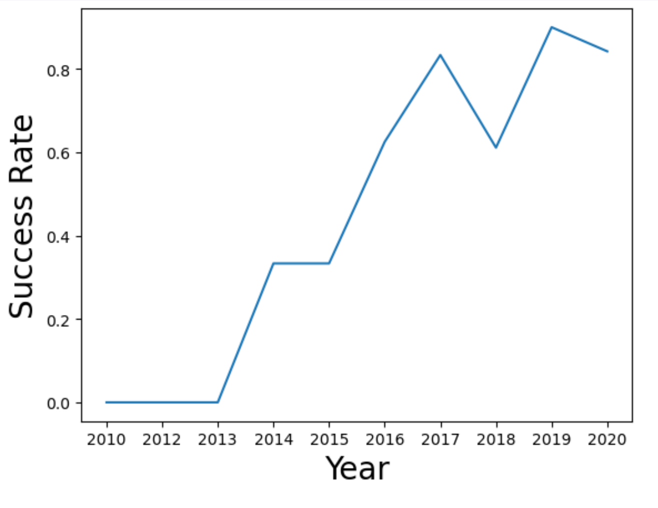
EDA with Data Visualization

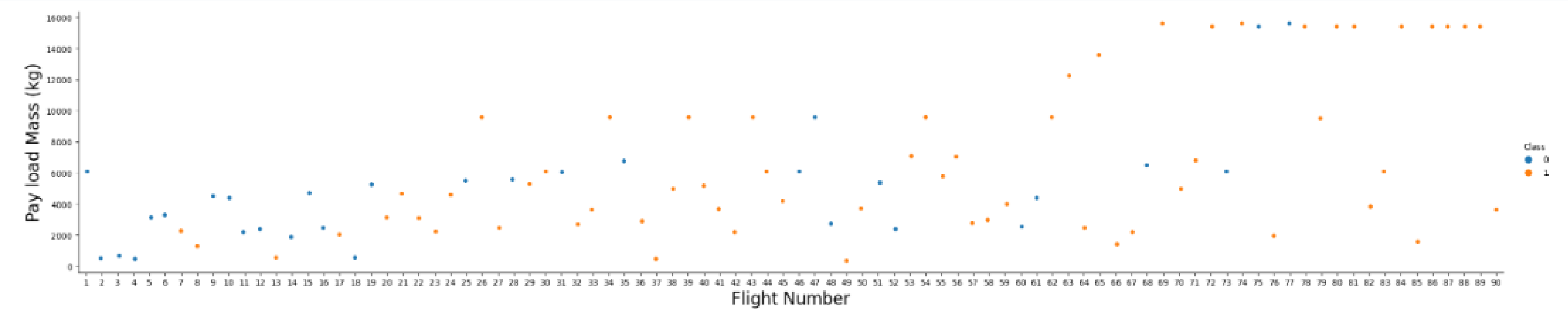
Visualized relationship between;

1. Flight number vs Payload mass
2. Payload vs Launch Site
3. Flight number vs launch site
4. Success rate of each orbit type
5. Flight number vs orbit type
6. Payload vs orbit type
7. Launch success yearly trend

[Link to Code]()







EDA with SQL

- Performed SQL queries to gather and understand data from dataset

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 acheived
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. Use a subquery
9. List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
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

[Link to Code]()

Build an Interactive Map with Folium

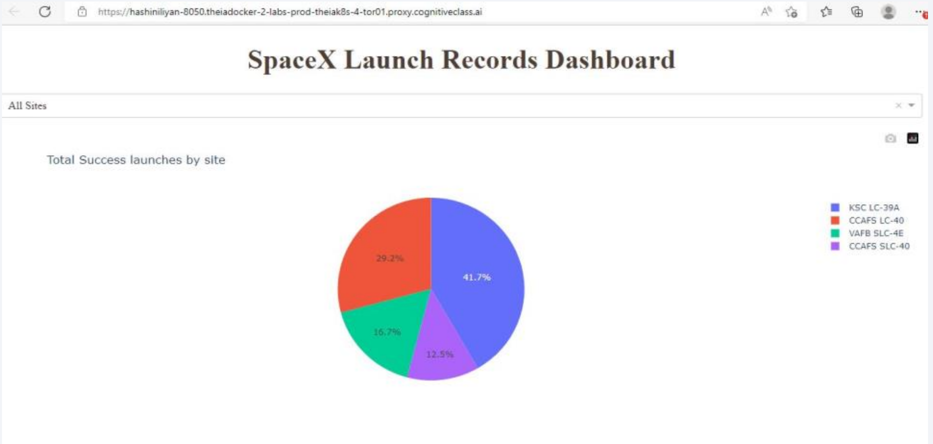
- Marked all launch sites on a map
- Marked the success/failed launches for each site on the map
- Calculate the distances between a launch site to its proximities and answered some question for instance:
    - Are launch sites near railways, highways and coastlines.
    - Do launch sites keep certain distance away from cities.
    -
    -

These objects were created in order to understand data and to find an optimal location for building a launch site.
We can easily show all launch sites, their surroundings and the number of successful and unsuccessful landings on map. So that we can discover some of the factors by analyzing the existing launch site locations that may cause success rate.
Link to Code
Link to Images (Zip uploaded since Git does not open map)

Build a Dashboard with Plotly Dash

- Built an interactive dashboard with Plotly dash
- Plotted pie charts showing the total launches by a certain sites
- Plotted scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster version.

[Link to Code]()



Predictive Analysis (Classification)

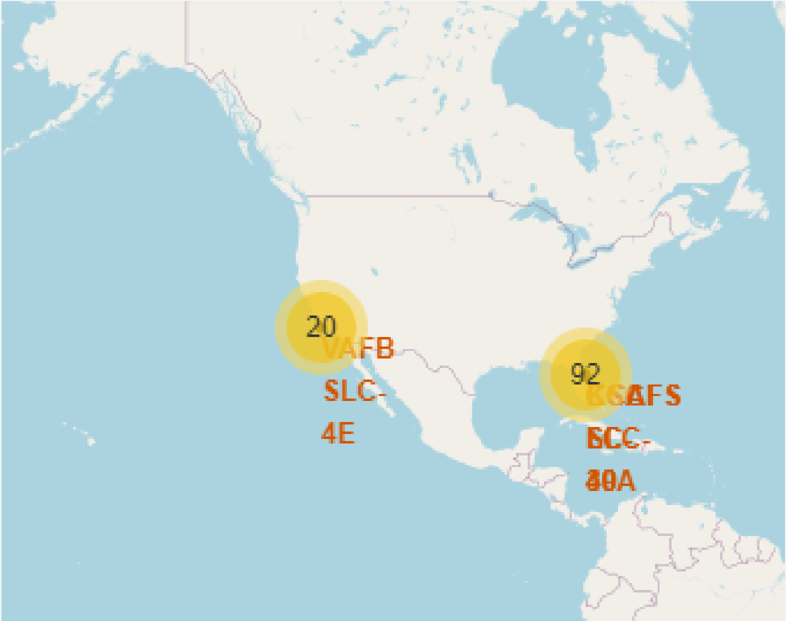
- Perform exploratory Data Analysis and determine Training Labels
- create a column for the class
- Standardize the data
- Split into training data and test data
- Find best Hyperparameter for SVM, Classification Trees and Logistic Regression
- Find the method performs best using test data

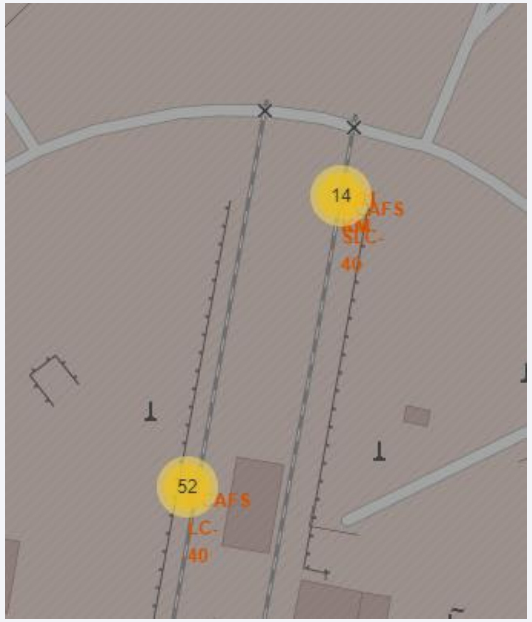
Results

- Exploratory data analysis results
  - Success rate since 2013 kept increasing till 2020
  - Positive landing rate are more for Polar, LEO and ISS
  - Different launch sites have different success rates.
    - CCAFS LC-40 - 60 %
    - KSC LC-39A - 77%
    - VAFB SLC 4E - 77%
  - The average payload of F9 v1.1 booster is 2,928 kg;
  - Space X uses 4 different launch sites

Results

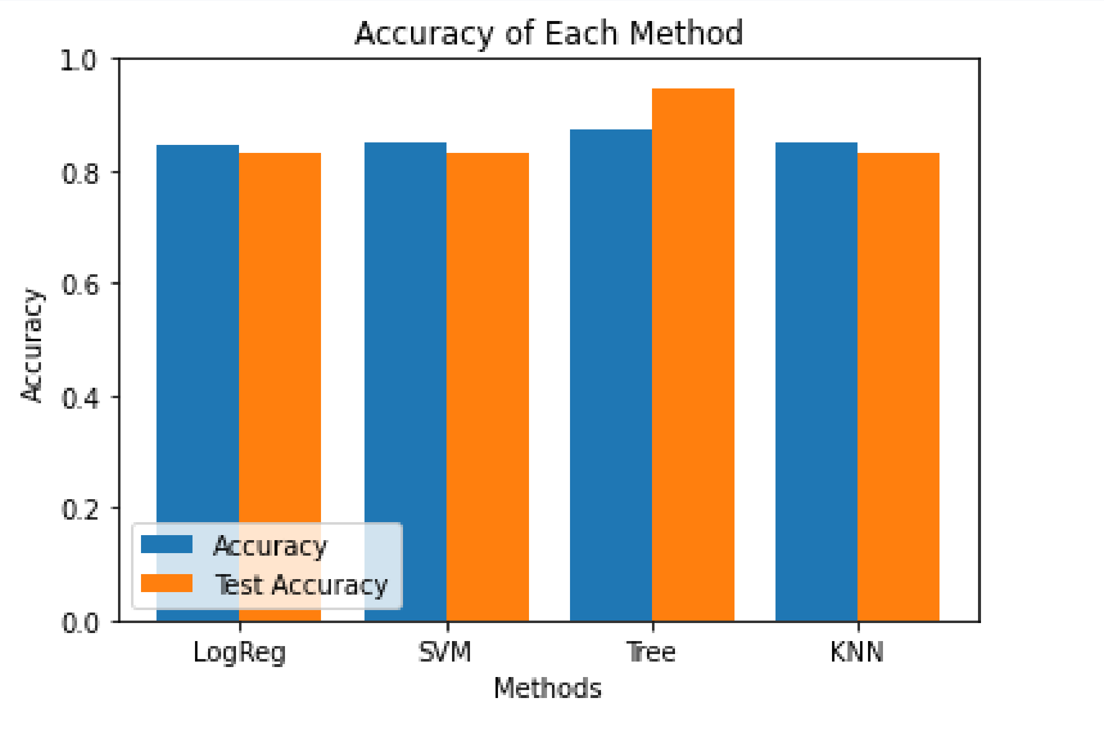
- Interactive analytics demo in screenshots
  - Most launches happens at east costal launch sites.





Results

- Predictive analysis results
  - Decision Tree is the best model

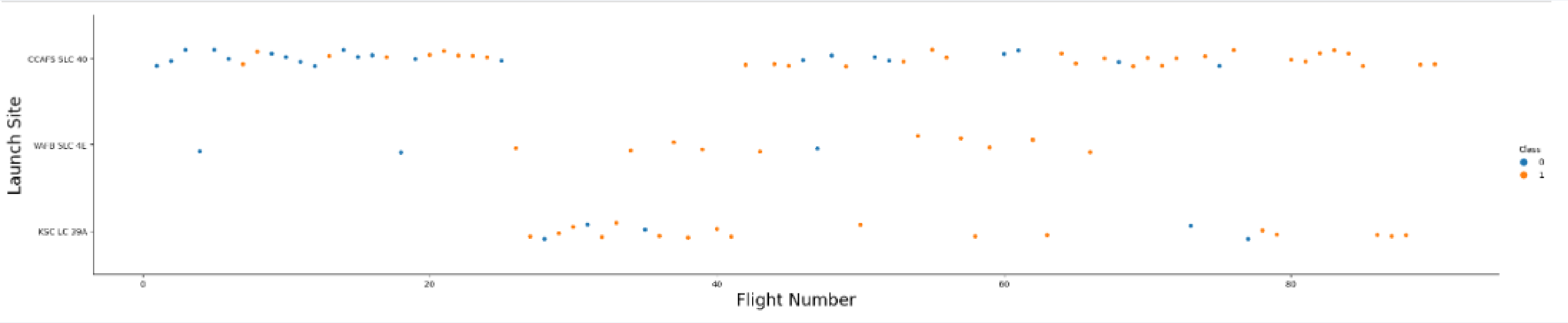


Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

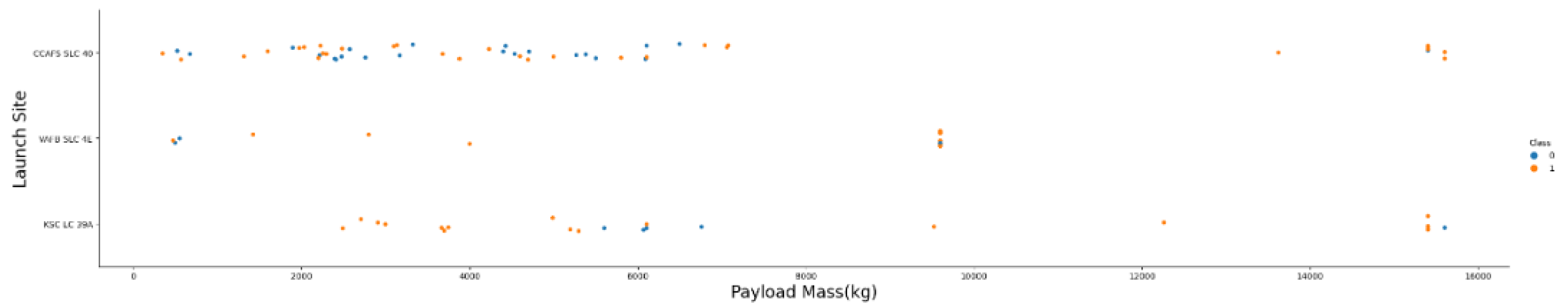
- Flight Number vs. Launch Site



1. Best launch site is CCAF5 SLC 40, where most of recent launches were successful;
2. Success rate improved over time.
3. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.

Payload vs. Launch Site

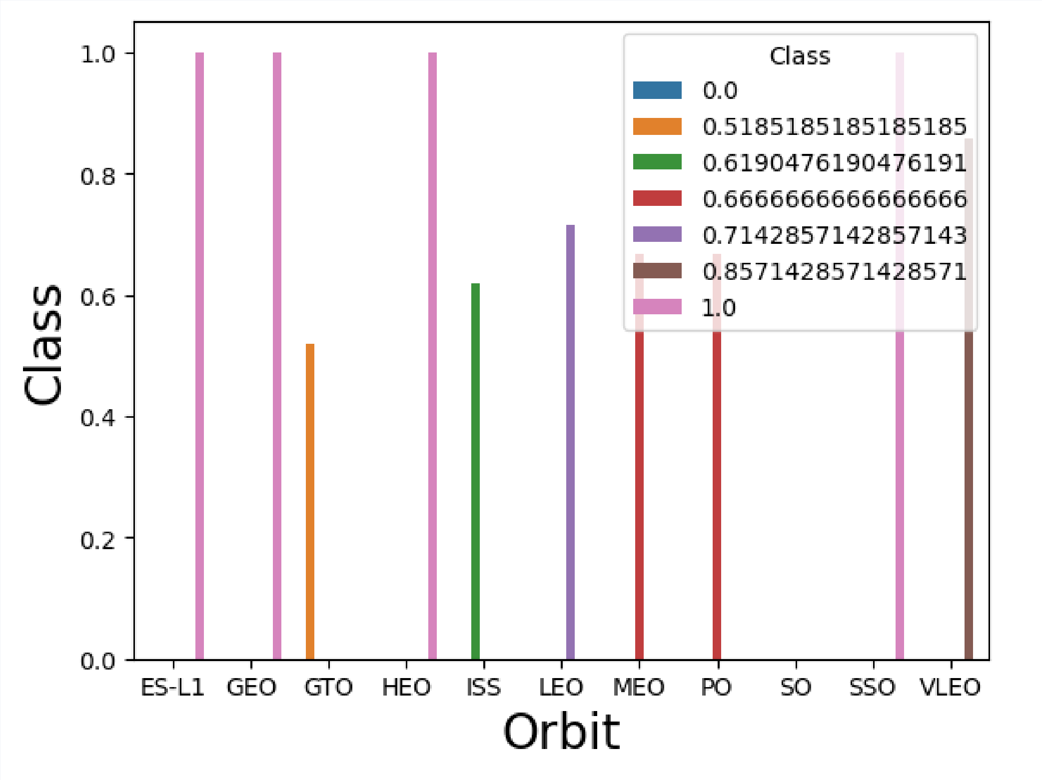
- Payload vs. Launch Site



1. More success rate for Payload more than 9,000kg
2. Launch site CCAFS LC-40 and KSC LC-39A support payload greater than 12,000 kg
3. VAFB-SLC launch site launched no rockets for heavy payload mass(greater than 10000).

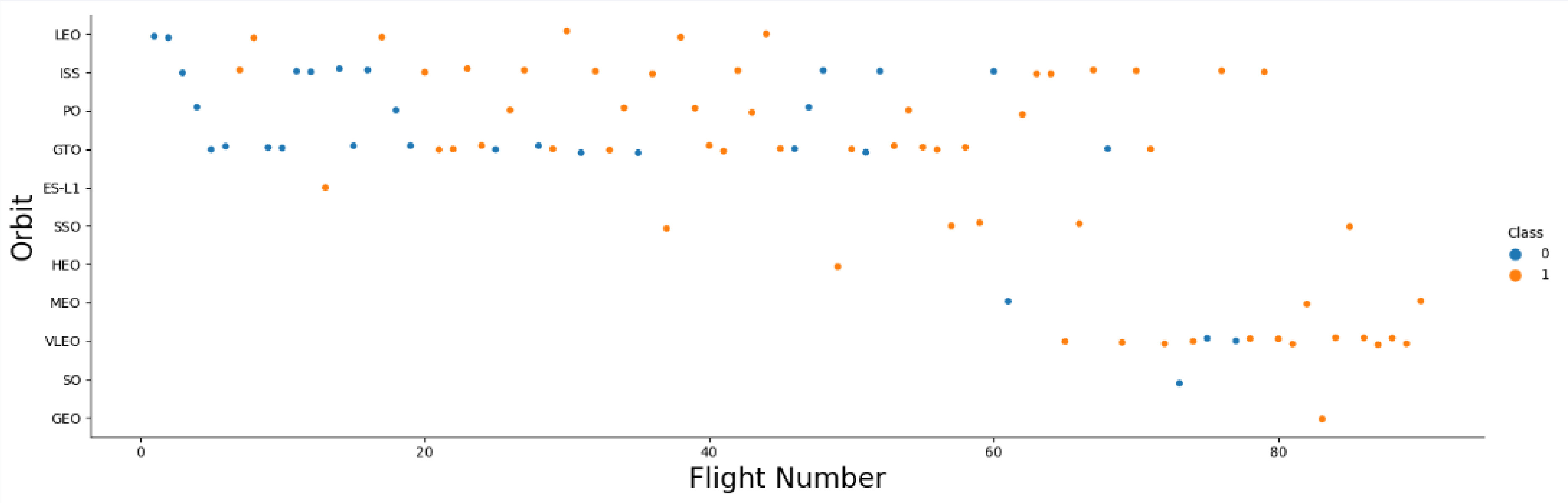
Success Rate vs. Orbit Type

- Success rate of each orbit type
  1. The biggest success rates nearly 100% happens to orbits:
     - ES-L1
     - GEO
     - HEO
     - SSO



Flight Number vs. Orbit Type

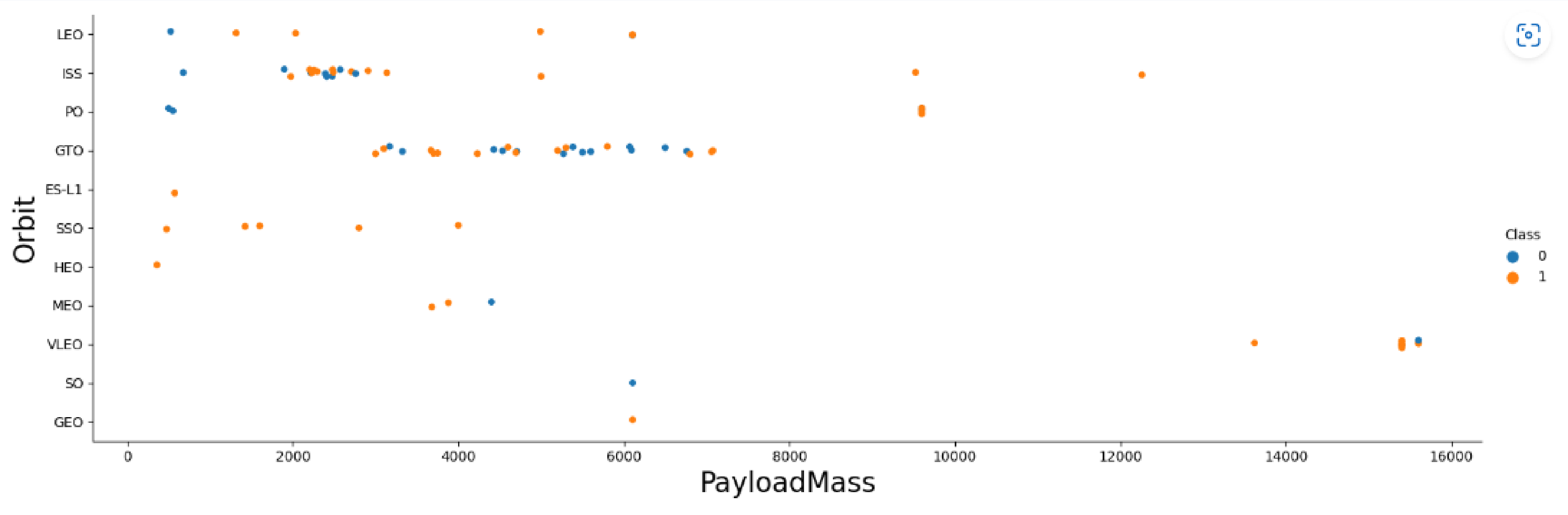
- Flight number vs. Orbit type



1. Success rate improved over time to all orbits
2. LEO orbit the Success appears related to the number of flights
3. On the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type

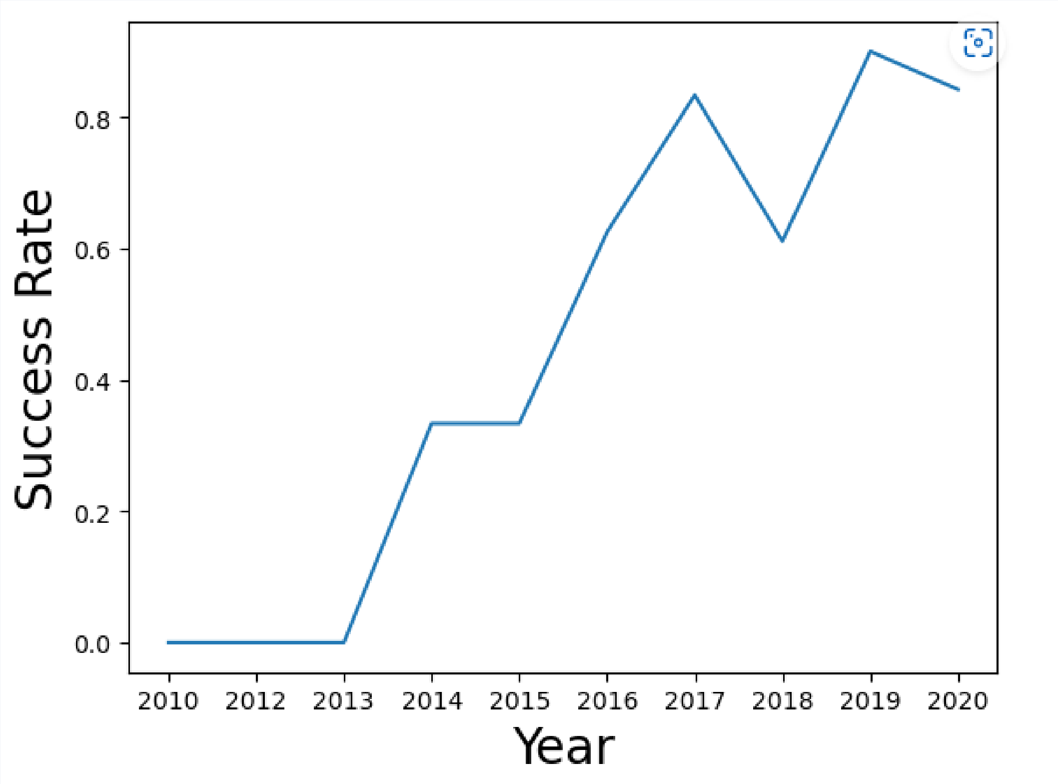
- Payload vs. Orbit type



1. With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
2. 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

- Yearly average success rate
    1. Success rate started increasing in 2013 and kept until 2020



All Launch Site Names

- Find the names of the unique launch sites
  - There were 4 sites used by SpaceX
  - Result obtained by taking distinct values of 'launch_site' from data set

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

```
%sql select DISTINCT(LAUNCH_SITE) from FINAL_DATA
```

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
    1. Result obtained by filtering records where 'launch_site' starting with 'CCA'
    2. Result display 5 records from data set which fulfill the conditions

```
%sql select * from FINAL_DATA WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

```
 * ibm_db_sa://wyz78043:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/BLUDB
Done.
```

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	0:35:00	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
  1. Result obtained by sum up all payload carried from Customer, 'NASA (CRS)'
  2. The total payload carried by boosters from NASA is **49,596 Kg**

```
%sql select SUM(PAYLOAD_MASS__KG_) from FINAL_DATA WHERE CUSTOMER = 'NASA (CRS)';
```

```
 * ibm_db_sa://wyz78043:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.d
Done.
```

1
45596

Average Payload Mass by F9 v1.1

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

    1. Result obtained by calculating average payload for booster version, 'F9 V1.1'
    2. The average payload mass carried by booster version F9 v1.1 is **2928 Kg**

```
%sql select AVG(PAYLOAD_MASS__KG_) from FINAL_DATA WHERE BOOSTER_VERSION = 'F9 v1.1';
```

```
 * ibm_db_sa://wyz78043:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.dat
Done.
```

1
2928

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

  1. Result obtained by filtering dataset by successful landing outcome on ground pad and getting the minimum value for date.
  2. First successful landing outcome on ground pad was on **12/22/2015**

```
%sql select DATE from FINAL_DATA WHERE LANDING__OUTCOME = 'Success (ground pad)' ORDER BY DATE ASC limit 1;
```

```
 * ibm_db_sa://wyz78043:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud
Done.
```

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

1. Result obtained by querying for 'landing_outcome' is 'Success (drone ship)' and Payload between 4,000 kg and 6,000 Kg

2. There are 4 Booster versions which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
%sql select BOOSTER_VERSION, PAYLOAD_MASS__KG_ from FINAL_DATA \
WHERE LANDING__OUTCOME = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ >4000 AND PAYLOAD_MASS__KG_<6000 ;
```

```
 * ibm_db_sa://wyz78043:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cl
Done.
```

booster_version	payload_mass__kg_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

1. Result obtained by grouping data set by 'Mission_outcome' and counting result set to each group.

```
%sql select MISSION_OUTCOME, COUNT(*) as TOTAL from FINAL_DATA GROUP BY MISSION_OUTCOME
```

```
 * ibm_db_sa://wyz78043:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.datab
Done.
```

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

    1. Result obtain by filtering payload mass with maximum payload
    2. Maximum payload calculated using sub query
    3. There are 12 booster versions which is having maximum payload

```
%sql select BOOSTER_VERSION from FINAL_DATA \
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) from FINAL_DATA);
```

```
 * ibm_db_sa://wyz78043:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1o
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, their booster versions, and launch site names for in year 2015

1. Result obtained by filtering landing_outcome is 'Failure (drone ship)' in 2015
2. There are 2 booster versions which are failed landing outcome in drone in Year 2015

```
%sql select BOOSTER_VERSION, LAUNCH_SITE from FINAL_DATA \
WHERE LANDING__OUTCOME = 'Failure (drone ship)' and YEAR(DATE) = 2015;
```

```
 * ibm_db_sa://wyz78043:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l0
Done.
```

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
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

1. Result obtained by grouping landing out come in between 2010-06-04 to 2017-03-20

```
%sql select LANDING__OUTCOME, COUNT(*) as TOTAL from FINAL_DATA \
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' \
GROUP BY LANDING__OUTCOME \
ORDER BY TOTAL DESC;
```

```
 * ibm_db_sa://wyz78043:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2
Done.
```

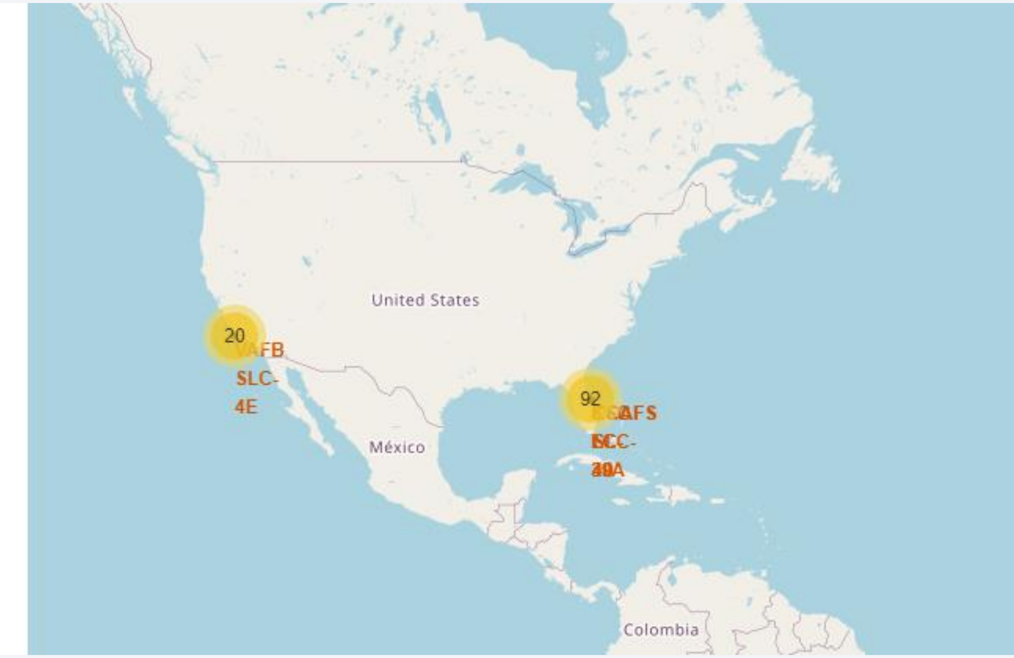
landing__outcome	total
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Success (ground pad)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	1
Precluded (drone ship)	1

Section 3

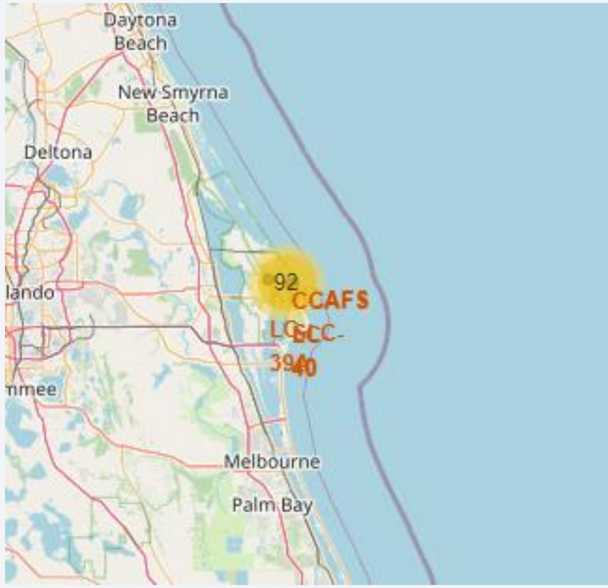
Launch Sites Proximities Analysis

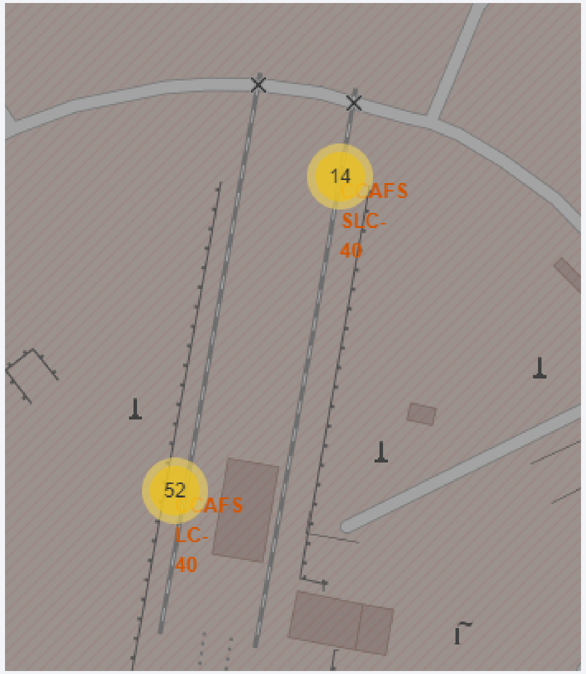
<All launch sites>

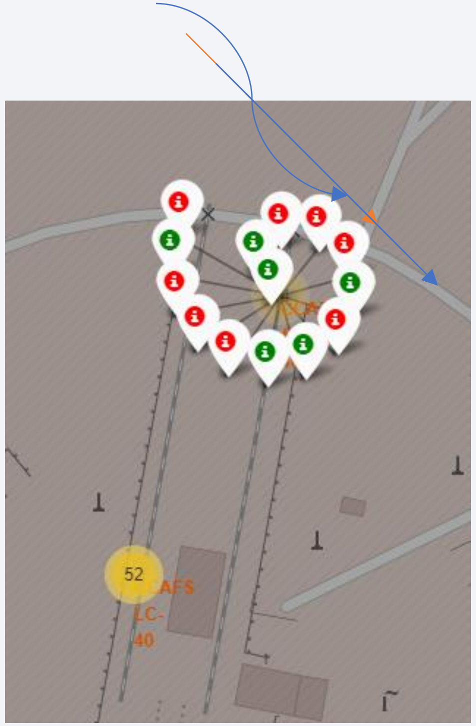
- All sites are located near costal area
- Most launches happened near east costal area
- But sites are not far from Roads, Railways

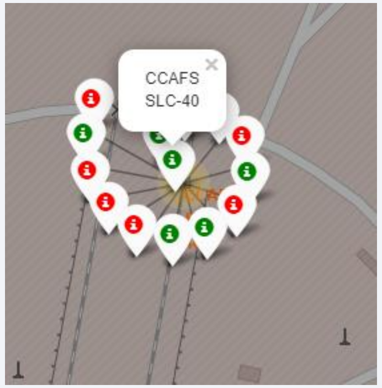


Launches and success







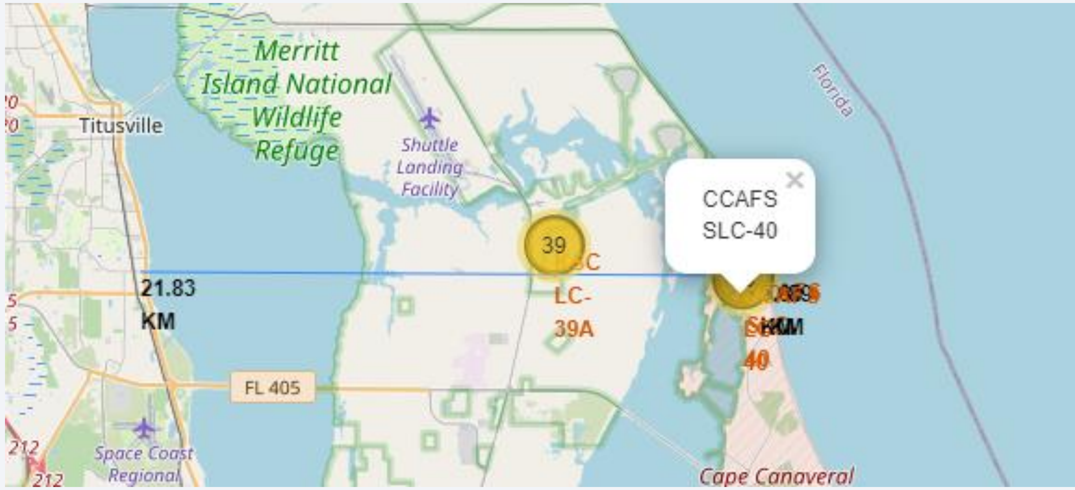


Green marker represents successful launches. Red marker represents unsuccessful launches.

Distance lines to the proximities

CCAFS SLC-40 :

- Close to Railway, Highway - Bad
- Far from Cities - Good
- Near to costal area - Good

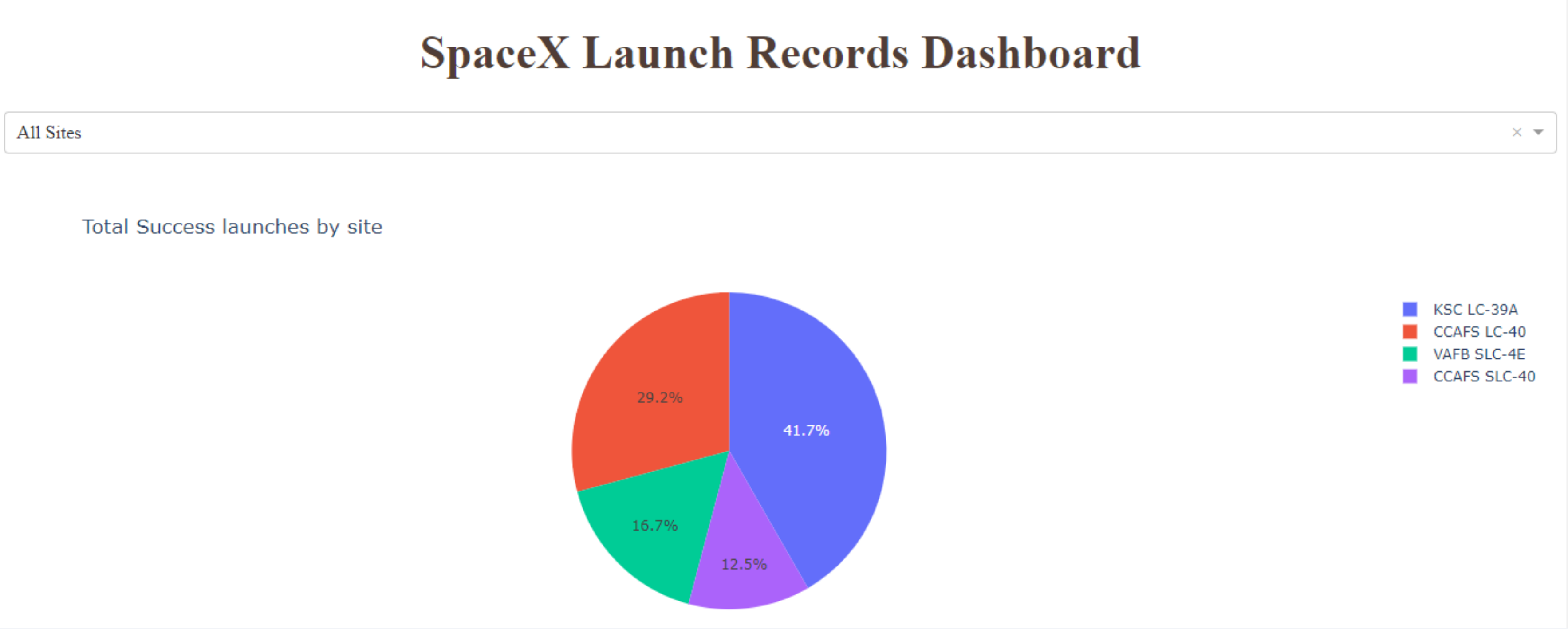


Section 4

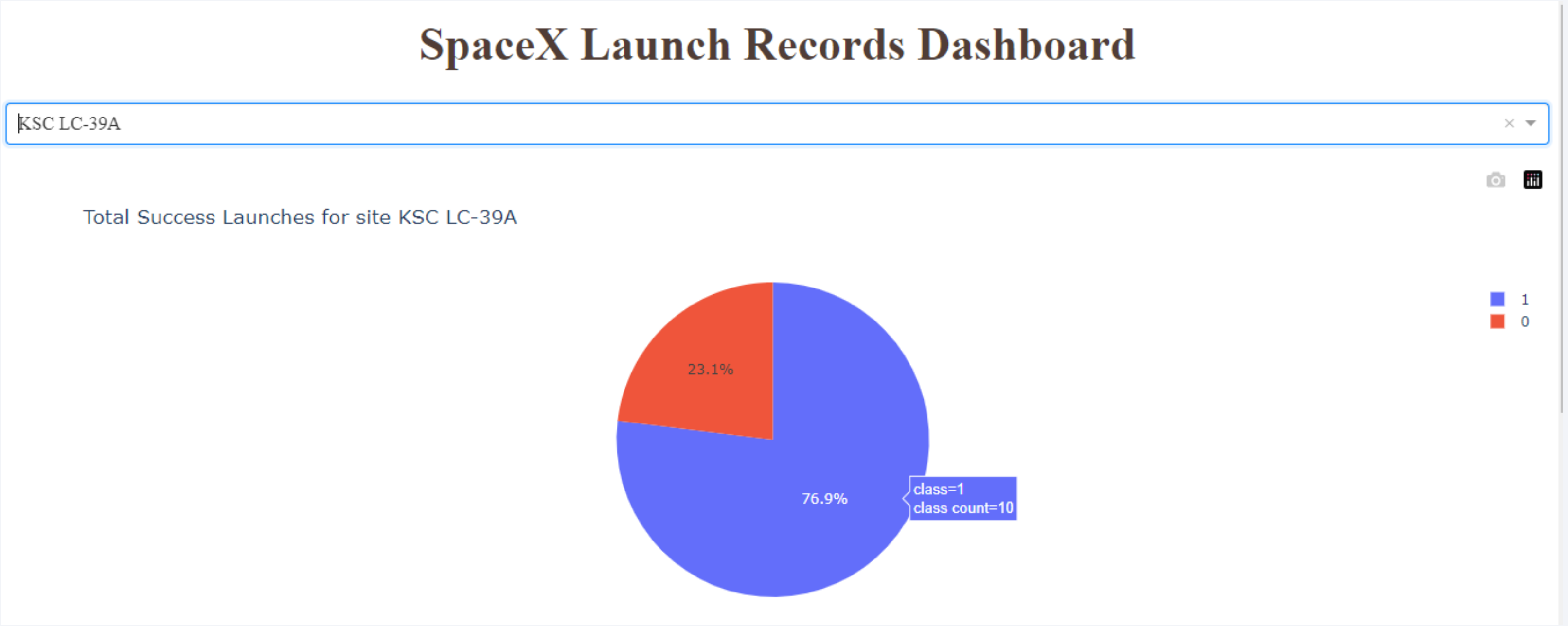
Build a Dashboard with Plotly Dash

Dashboard – Total success by Site

- KSC LC-39A has the best success rate of launches



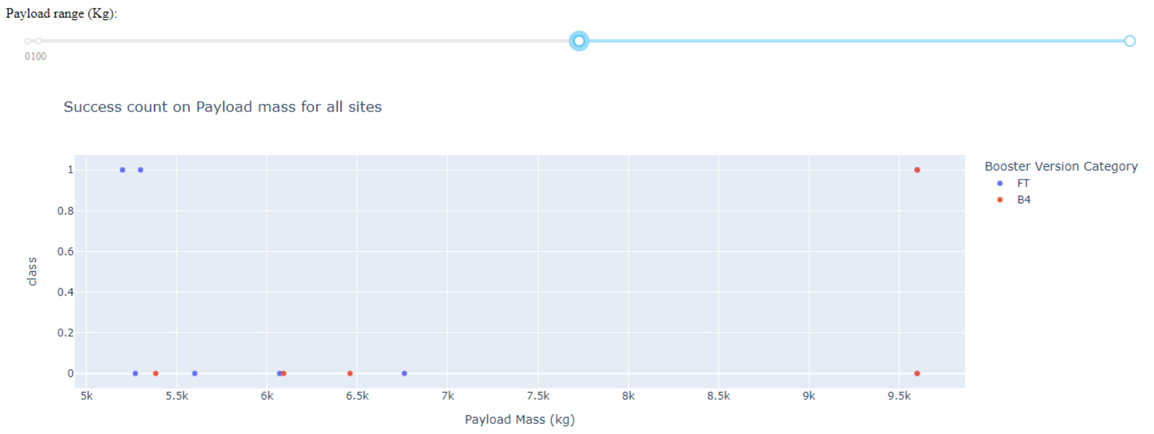
Total success launches – Site KSC LC 39A



- 76.9% of launches are successful

Payload vs. Launch Outcome

- Low weighted payloads success rate is higher than the heavy weighted payloads.



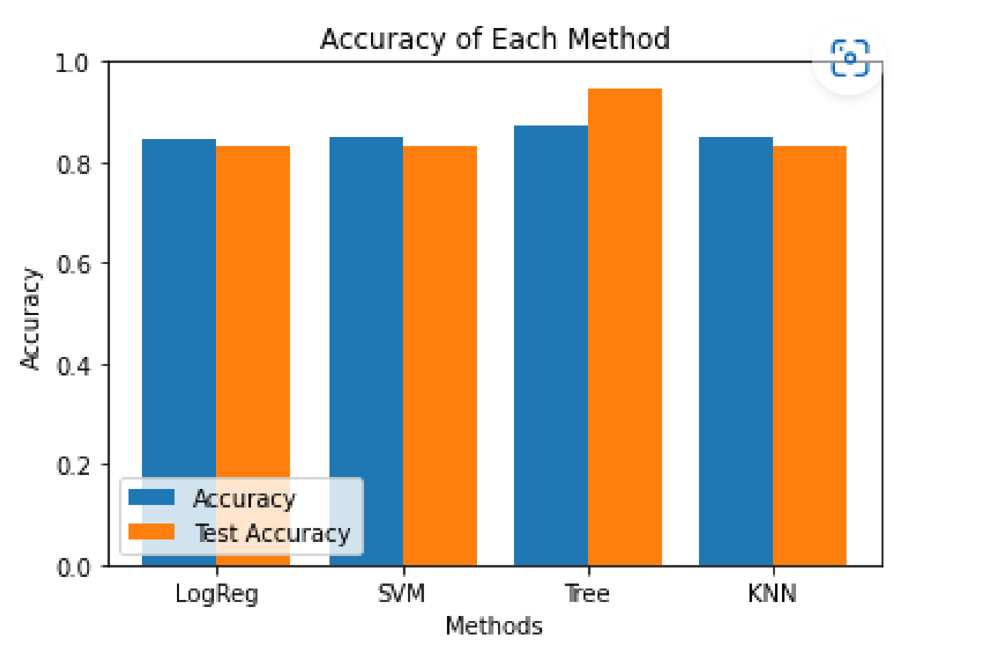
Section 5

Predictive Analysis (Classification)

Classification Accuracy

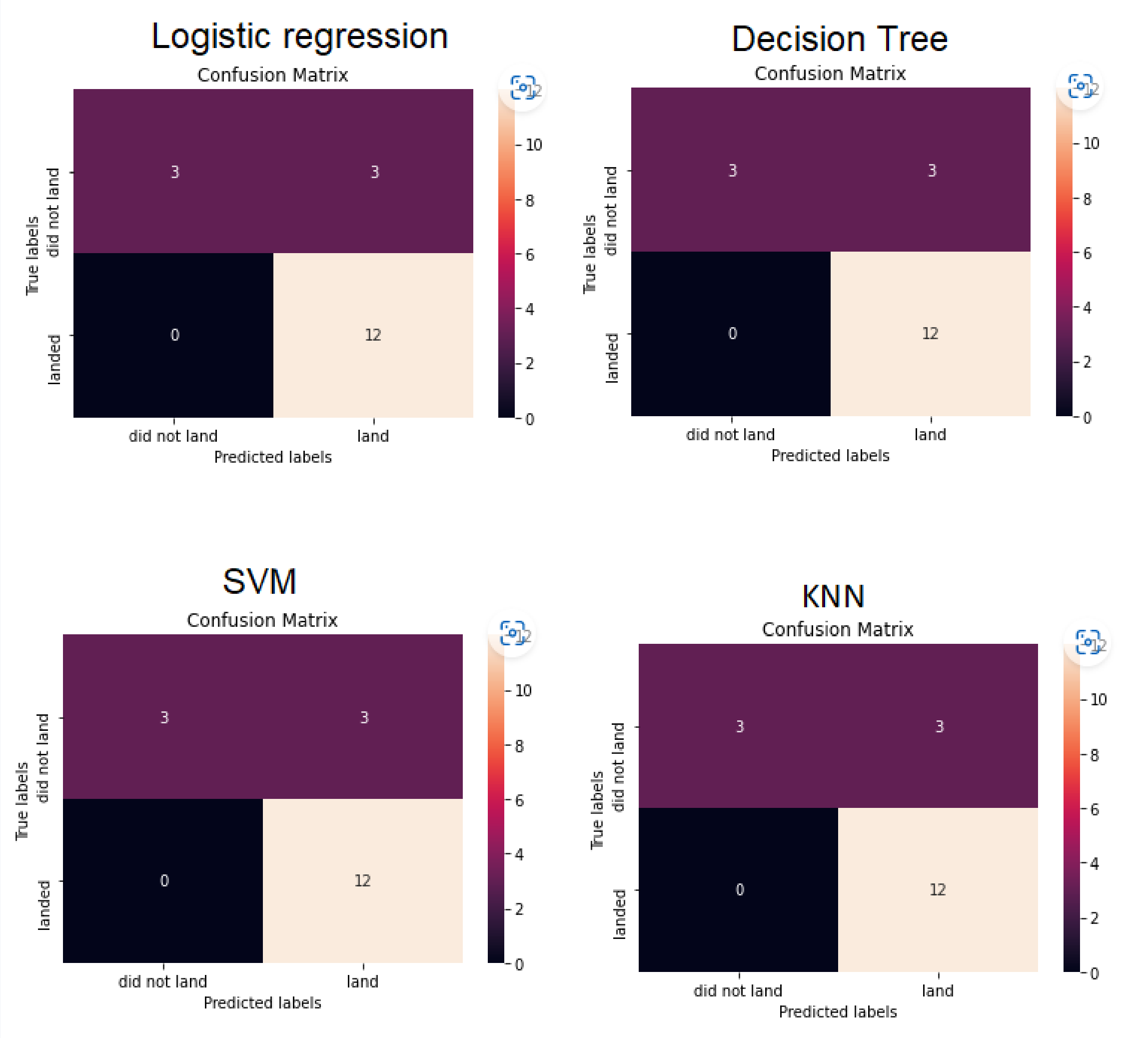
- Four classification models were tested
- Highest classification accuracy is for:
  - Decision Tree Classifier
  - Accuracies greater than 87%

Model	Accuracy	TestAccuracy
LogReg	0.84643	0.83333
SVM	0.84821	0.83333
Tree	0.875	0.94444
KNN	0.84821	0.83333



Confusion Matrix

- Confusion matrices are identical.



Conclusions

- Low weighted payloads success rate is higher than the heavy weighted payloads.
- KSC LC-39A has the best success rate among launch sites
- The biggest success rates nearly 100% happens to orbits:(ES-L1, GEO, HEO and SSO)
- Landing success rate is improving over time
- Decision Tree Classifier can be used to predict successful landings and increase profits.

Appendix

- Folium maps are not loading on GIT. Screen shots uploaded.

Thank you!