Winning Space Race with Data Science

Yen Po Liu
10/12/2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through REST API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis (EDA) result
  - Interactive analytics in screenshots
  - Predictive Analytics result

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. The goal of this project is to develop a machine learning prediction pipeline to predict if the first stage will land successfully.
- Problems you want to find answers
  - ➢ What are the critical factors that effect landing outcome of first stage?
  - ➢ How to make the best prediction to first stage landing?
  - ➢ What operating conditions needs to be in place to ensure a successful landing?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - REST API and Web scraping
- Perform data wrangling
  - One-hot encoding was applied to categorical 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 - API

- Data collected through API to connect endpoint. (https://api.spacexdata.com/v4/launches/past)

Request

- Request the endpoint to get launch data

Get

- Received raw data in .json format

Normalized

- Normalized json format data into data frame structure.

Data Collection – Web Scraping

- Data collected from Wikipedia (https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922)

Request

- Requested webpage to get text data in HTML format.

Parse

- Applied beautiful soup module as a parser to extract launch table.

Transform

- Transformed HTML data into data frame through pandas module.

Data Collection – SpaceX API

- GitHub URL: https://github.com/YbLiu1127/IBM-Data-Science-Capstone/blob/main/Data%20Collection.ipynb

Request

- Request to endpoints.
- Extract launch data into lists through API functions.

Wangling

- Normalize data structure, extract columns to lists respectively.
- Import lists in to dictionary, transform into data frame.

Cleaning

- Filter that data to only include “Falcon 9 ”launches.
- Replace null in “Payload column” with mean value.

Data Collection - Scraping

- GitHub URL: https://github.com/YbLiu1127/IBM-Data-Science-Capstone/blob/main/Web%20Scraping.ipynb

Request
- Request to Falcon9 launch Wikipedia webpage.

Parse
- Parse the HTML with beautiful soup module, find out the Falcon 9 launch table.

Extract
- Implement “for loop”, extract each cell from table, fill up new launch dictionary.

Data Wrangling

- Perform exploratory data analysis and determine training labels
- GitHub URL: https://github.com/YbLiu1127/IBM-Data-Science-Capstone/blob/main/Data%20Wrangling.ipynb

Analysis

- Use “value counts” to determine the number and occurrence of orbits and launch sites.

Identify

- Exploring landing outcomes feature, classify all outcome situation into binary stage(1 for succeed ,0 for failed)

Add column

- Create a landing outcome label from outcome column that is going to be predicted in the next step.

EDA with Data Visualization

- In order to explore the whether features would affect the launch outcome.
- **Scatter plots**: plot out relationship between "Flight number, Launch Site, Orbit Type and Payload Mass", trying to find trends that might have higher success landing rate.
- **Line chart**: display the trend of success rate increasing.
- **Bar chart**: check if there are any relationship between success rate and orbit type.
- GitHub URL: https://github.com/YbLiu1127/IBM-Data-Science-Capstone/blob/main/EDA%20with%20Vusualization.ipynb

EDA with SQL

- 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 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

EDA with SQL

- 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 failed landing outcomes in drone ship, their booster versions, and launch site names for 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: https://github.com/YbLiu1127/IBM-Data-Science-Capstone/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Objective: try to observe in geographical view and explore implicit relationship that affect landing outcome.
- Markers: launch places and launch events (succeed or failed)
- Lines: display the nearest public facility or landmark and its distance from the launch sites.
- GitHub URL: https://github.com/YbLiu1127/IBM-Data-Science-Capstone/blob/main/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

- Objective: to understand success rate of each site and landing result that base on different payload mass and booster versions
- Display each site's successful landing rate in pie chart and payload mass within booster versions in scatter plot.
- GitHub URL: https://github.com/YbLiu1127/IBM-Data-Science-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Import data with numpy and pandas module, use fix_transform function to normalized and then use train test split function to split data with 20% test size.
- Build machine learning models and test them with different hyperparameters using GridSearchCV.
- Display models' accuracy metric.
- GitHub URL: https://github.com/YbLiu1127/IBM-Data-Science-Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction.ipynb

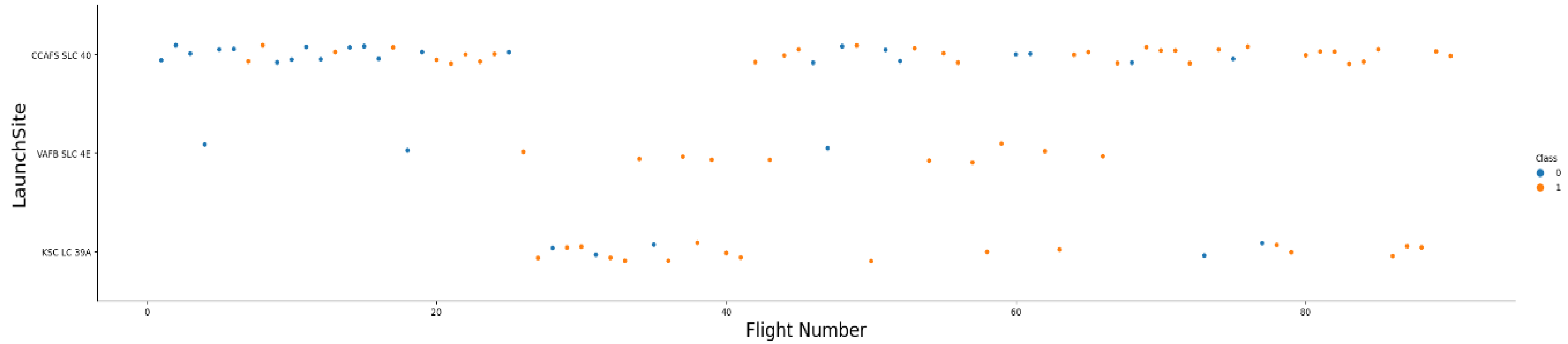
Results

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

Section 2

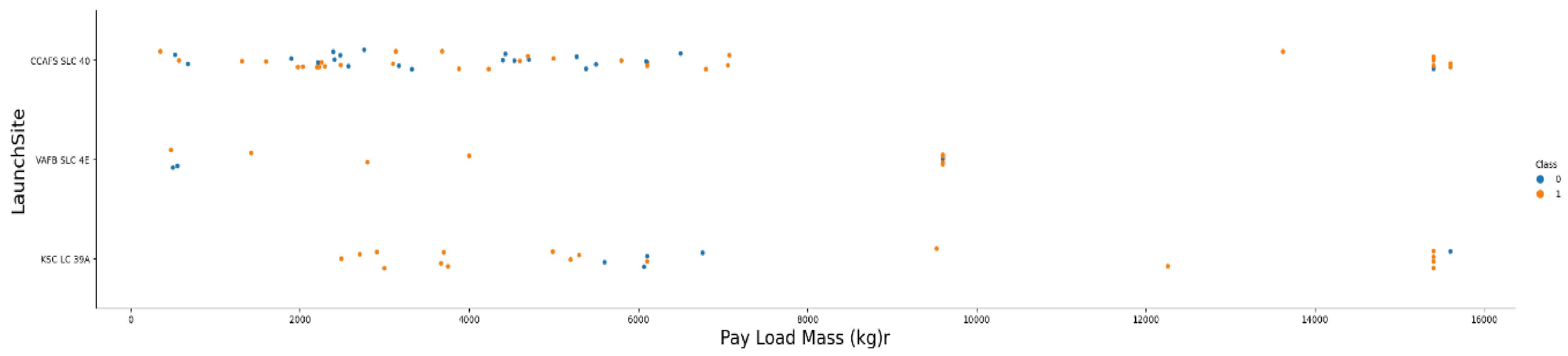
Insights drawn from EDA

Flight Number vs. Launch Site



- According to the plot, we found that in recent launches, it seems to have the most successful landings and highest success rate at CCAFS SLC 40.

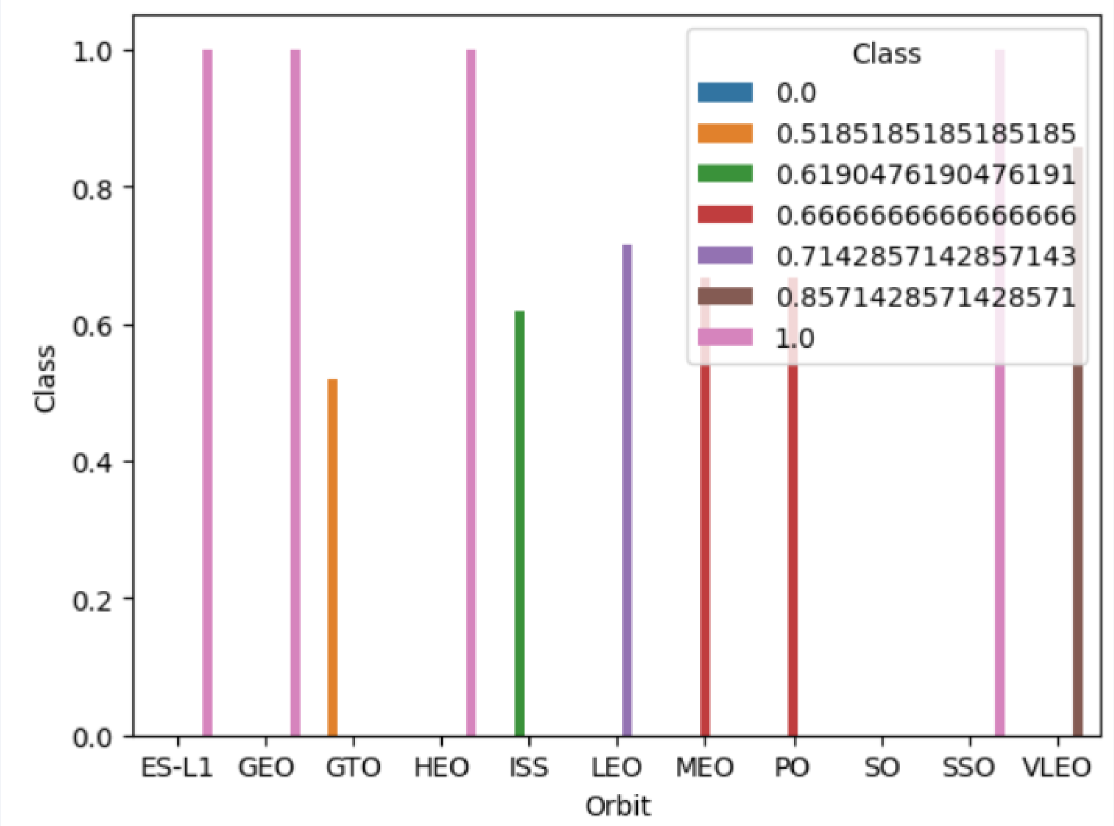
Payload vs. Launch Site



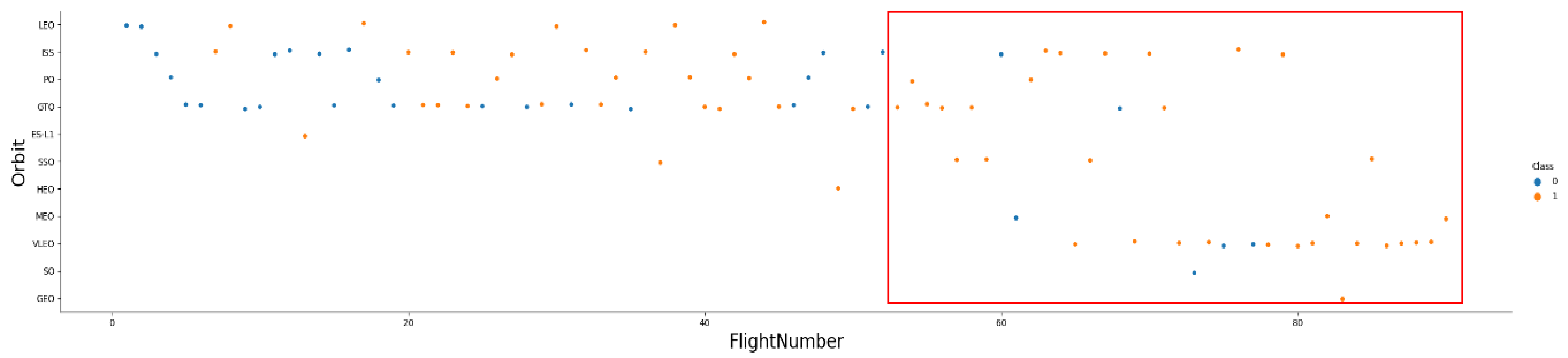
- The greater payload mass the higher success rate for launches.

Success Rate vs. Orbit Type

- We found out that “ES-L1, GEO, HEO, SSO” orbit type have the highest success rate.

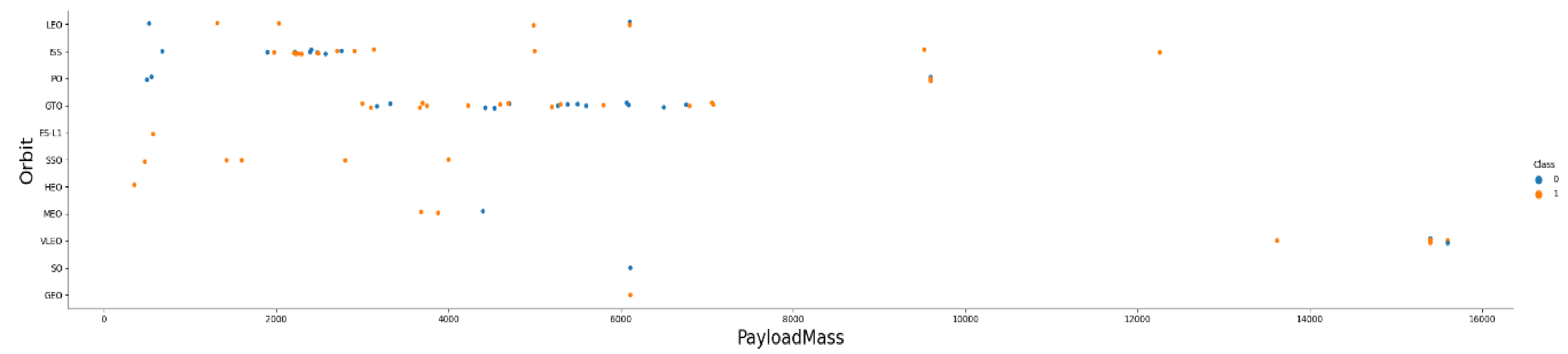


Flight Number vs. Orbit Type



- We can tell that according to the red box, it seems that last 5 of orbit type had significantly launched recently, and most of them have higher success rate.

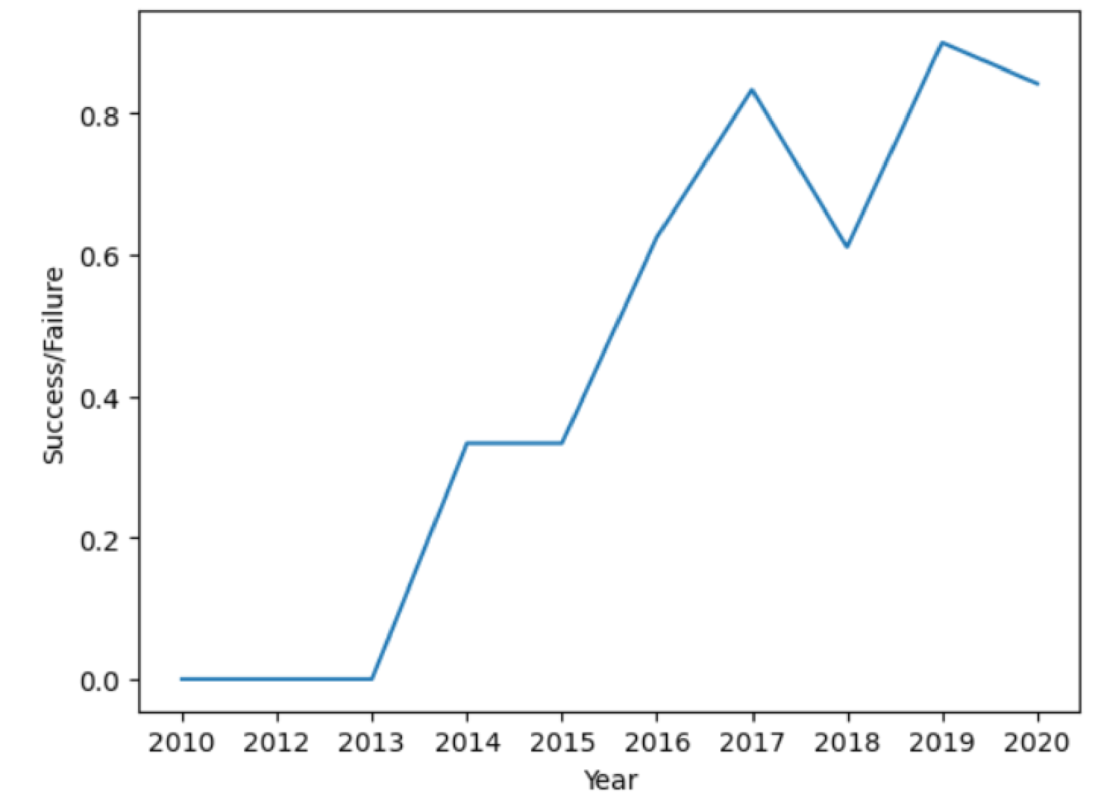
Payload vs. Orbit Type



- We can found that with heavy payloads the successful landing or positive landing rate are more for PO, VLEO and ISS.

Launch Success Yearly Trend

- According to the plot, we can tell that success rate has generally increased in last decade, however, the rate performed in 2018 and 2020 decreased, further analysis is recommend to find out reasons that cause the result.



All Launch Site Names

- DISTINCT was used when querying unique launch site name from spacex data.

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

```
%sql SELECT DISTINCT(LAUNCH_SITE) FROM SPACEXTBL
```

```
 * ibm_db_sa://ygk14970:***@21fecfd8-47b7-4937-840d-d791d02
Done.
```

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

Launch Site Names Begin with 'CCA'

- We used the query below to display 5 records where launch sites begin with `CCA`

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

```
%sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
```

```
 * ibm_db_sa://ygk14970:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/bludb
Done.
```

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

- Calculate the total payload carried by boosters from NASA as 45596 using the query below

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

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

```
 * ibm_db_sa://ygk14970:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od
Done.
```

1
45596

Average Payload Mass by F9 v1.1

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

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

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE BOOSTER_VERSION LIKE 'F9 v1.1%'
```

```
 * ibm_db_sa://ygk14970:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.dat
Done.
```

1
2534

First Successful Ground Landing Date

- Found the dates of the first successful landing outcome on ground pad was 2015-12-22

List the date when the first successful landing outcome in ground pad was acheived.

Hint:Use min function

```
%sql SELECT MIN(DATE) FROM SPACEXTBL WHERE "Landing _Outcome" LIKE 'Success (ground pad)'
```

```
 * ibm_db_sa://ygk14970:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.datal
Done.
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- We 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,"Landing _Outcome", PAYLOAD_MASS__KG_  FROM SPACEXTBL WHERE "Landing _Outcome" LIKE 'Success (drone ship)' and PAYLOAD_MAS
```

```
 * ibm_db_sa://ygk14970:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/bludb
Done.
```

booster_version	Landing _Outcome	payload_mass__kg_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

Total Number of Successful and Failure Mission Outcomes

- Calculated the total number of successful and failure mission outcomes by using count(MISSION_OUTCOME)

List the total number of successful and failure mission outcomes

```
%sql SELECT count(MISSION_OUTCOME) FROM SPACEXTBL GROUP BY MISSION_OUTCOME
```

```
 * ibm_db_sa://ygk14970:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08k
Done.
```

1
1
99
1

Boosters Carried Maximum Payload

- We extracted the names of the booster which have carried the maximum payload mass by using subquery.

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

```
%sql SELECT BOOSTER_VERSION, PAYLOAD_MASS__KG_ FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

```
 * ibm_db_sa://ygk14970:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/bludb
Done.
```

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

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015, set query date between 2015/01/01 and 2015/12/31

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

```
%sql SELECT BOOSTER_VERSION,LAUNCH_SITE FROM SPACEXTBL WHERE "Landing _Outcome" = 'Failure (drone ship)' AND DATE BETWEEN '2015-01-01' AND '2015-12-31
```

```
 * ibm_db_sa://ygk14970:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/bludb
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

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("Landing _Outcome") AS FREQ FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY "Landing _O
```

```
 * ibm_db_sa://ygk14970:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/bludb
Done.
```

Landing _Outcome	freq
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

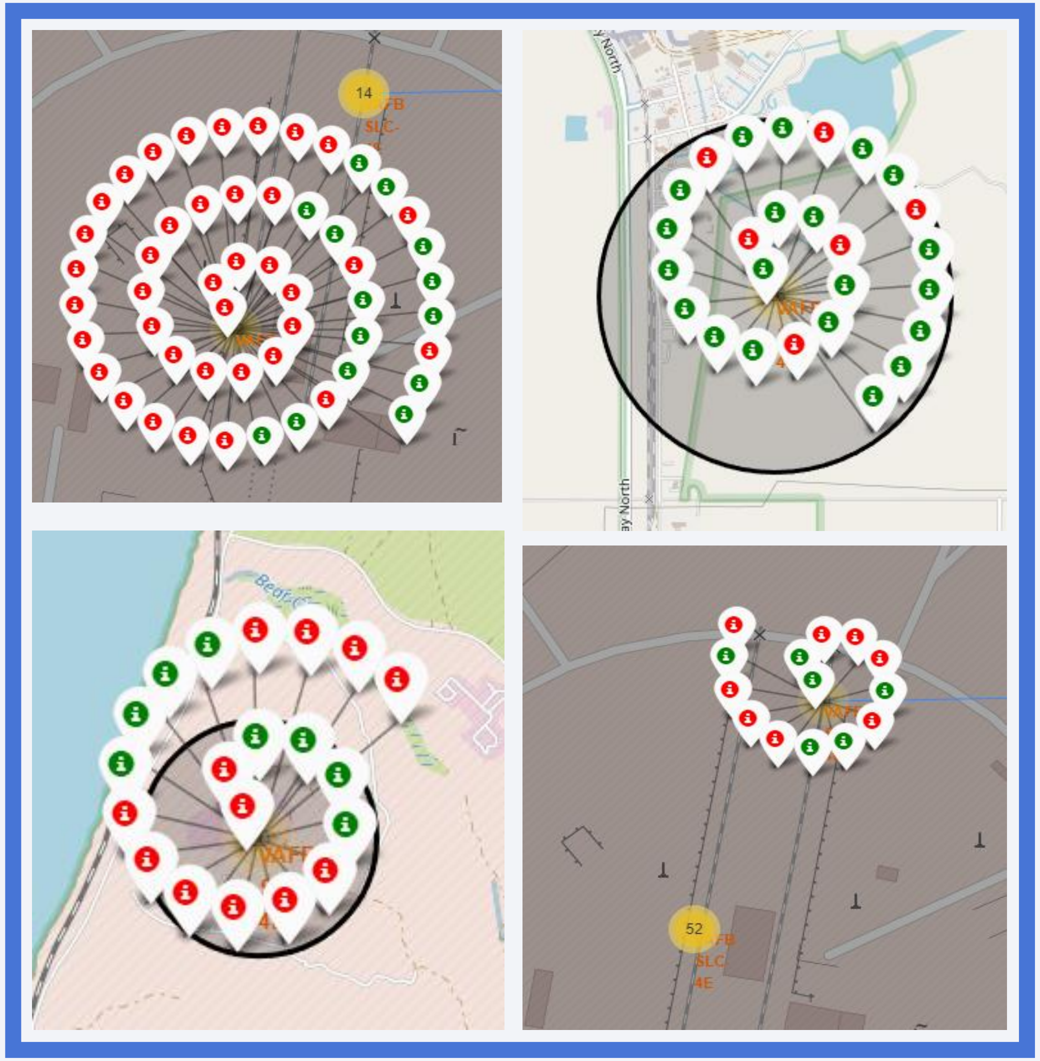
Section 3

Launch Sites Proximities Analysis

Global map markers – all launch sites

- We can discover that all SpaceX launch sites are basically located in coast of United States

Marker Clusters – launch events



- Marker clusters show every launch at each launch site.

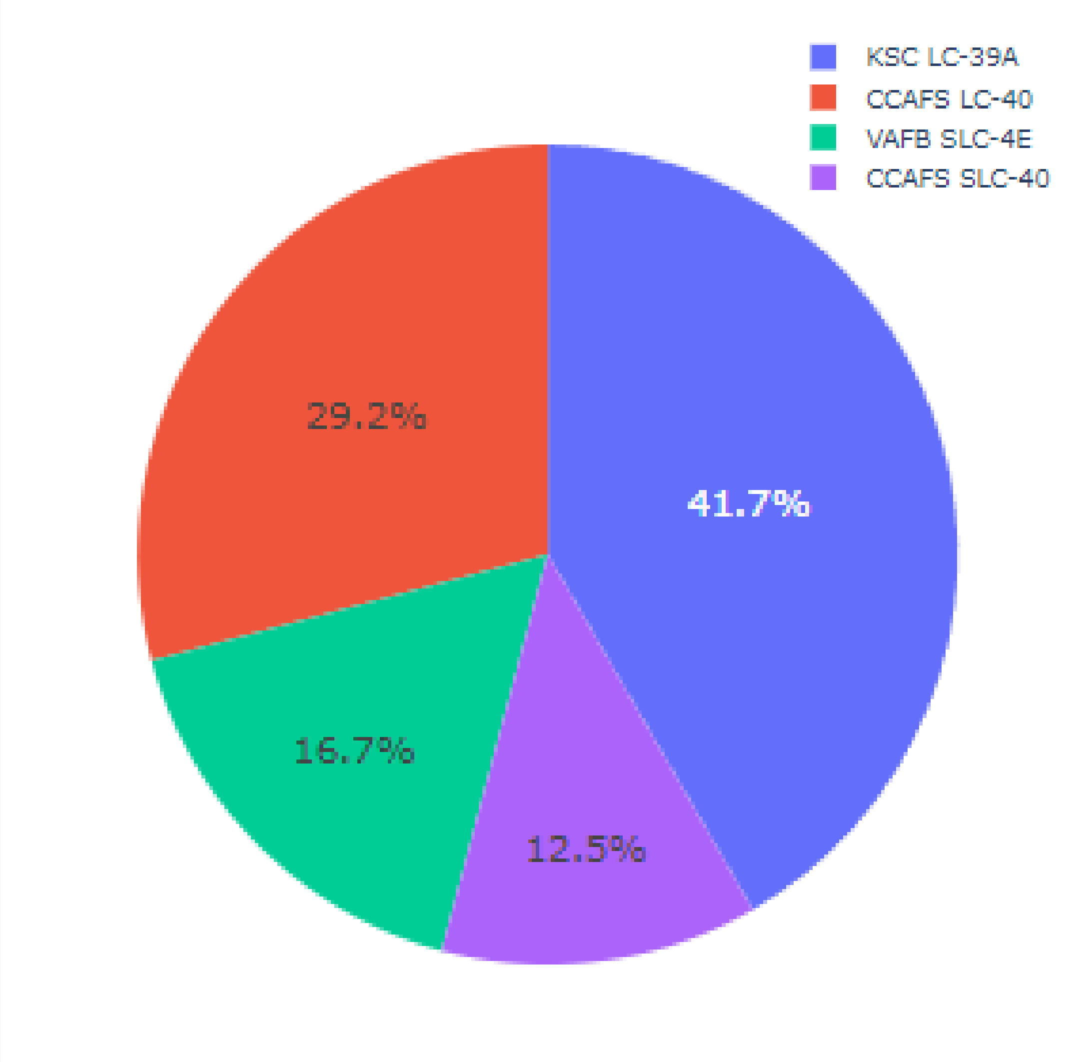
- Green marker: landing success
- Red marker: landing failed

Distance to landmarks

Section 4

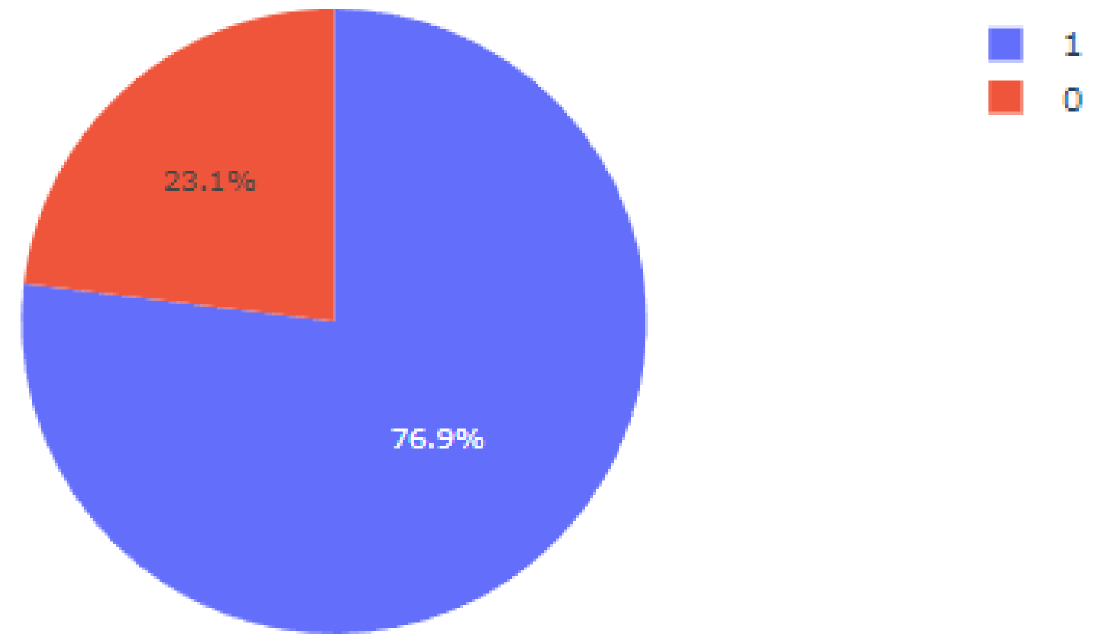
Build a Dashboard with Plotly Dash

Success count for all launch sites



- We can see that launch site KSC LC-39A has the most successful launches.

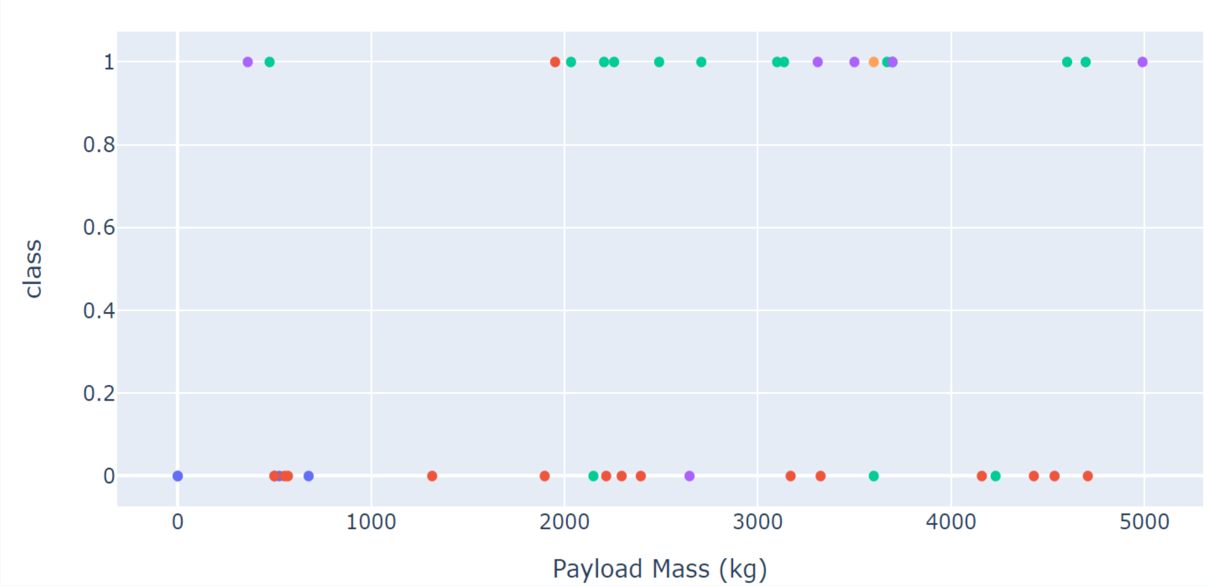
Total success launches for KSC LC-39A

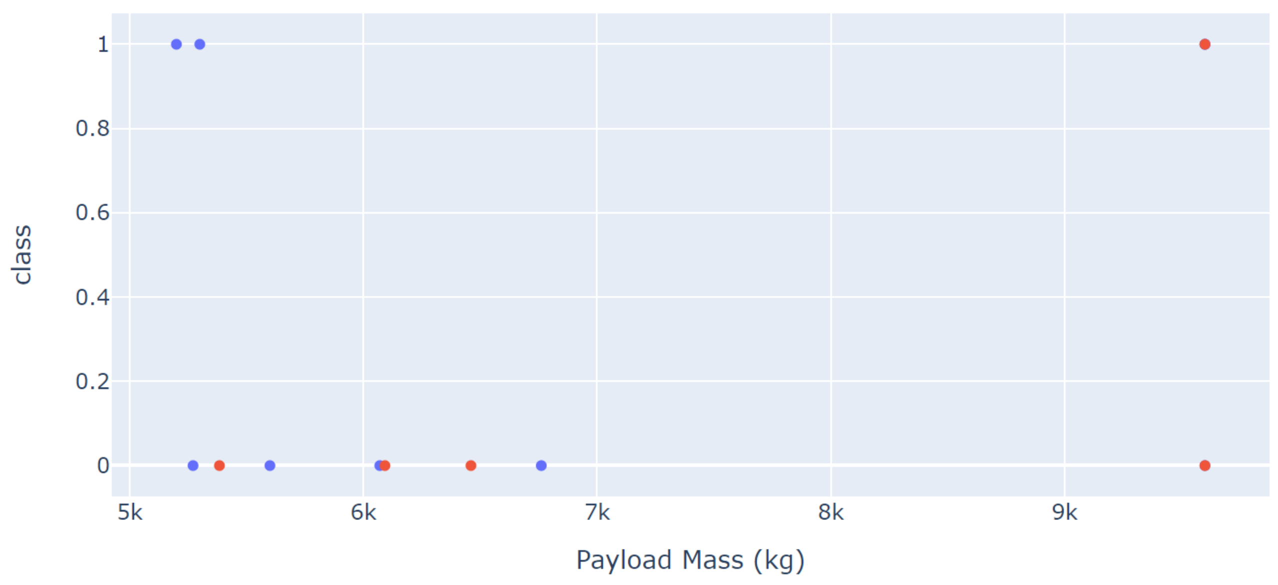


- KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate.

Payload vs. Launch Outcome scatter plot for all sites

- We can see that success rate for low payload is higher that heavy payload.

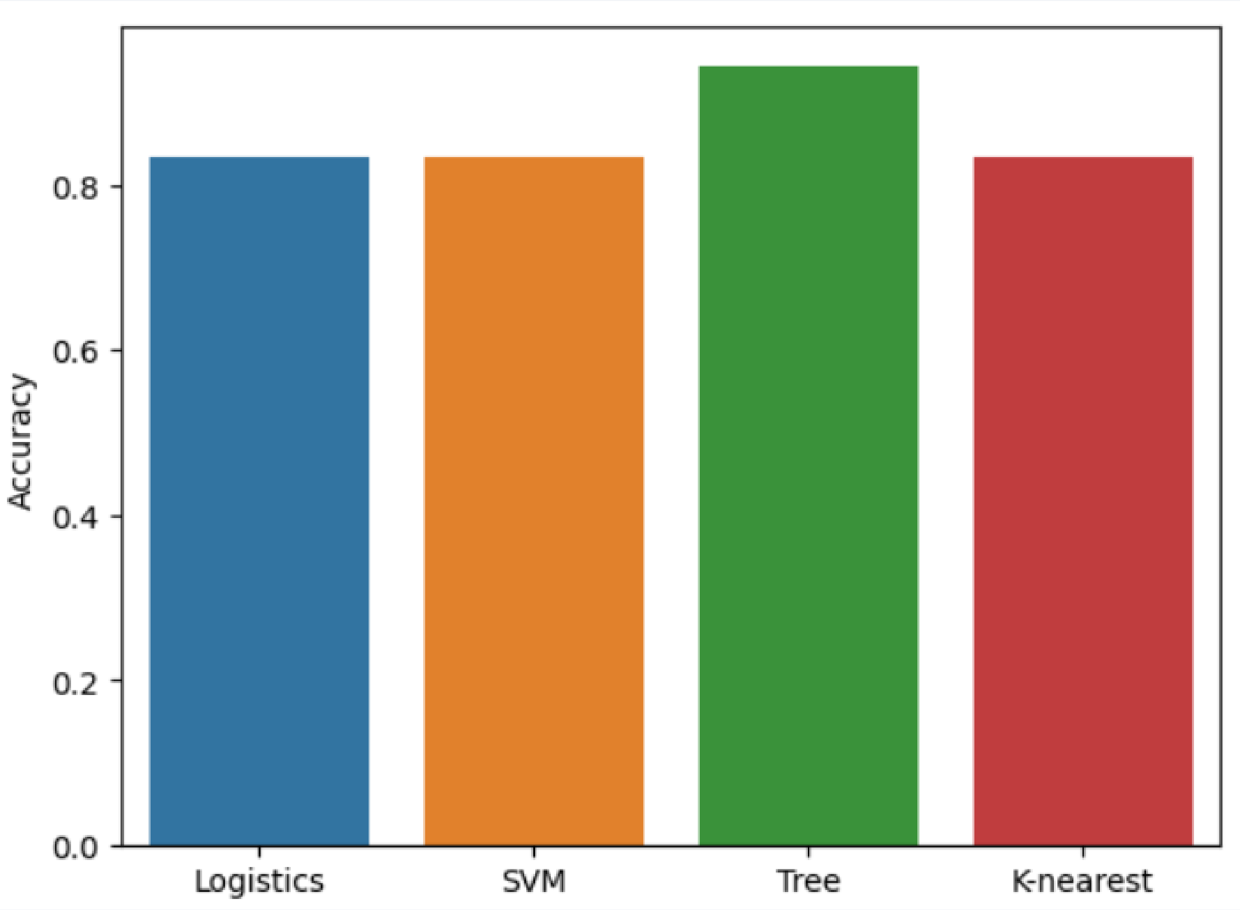




Section 5

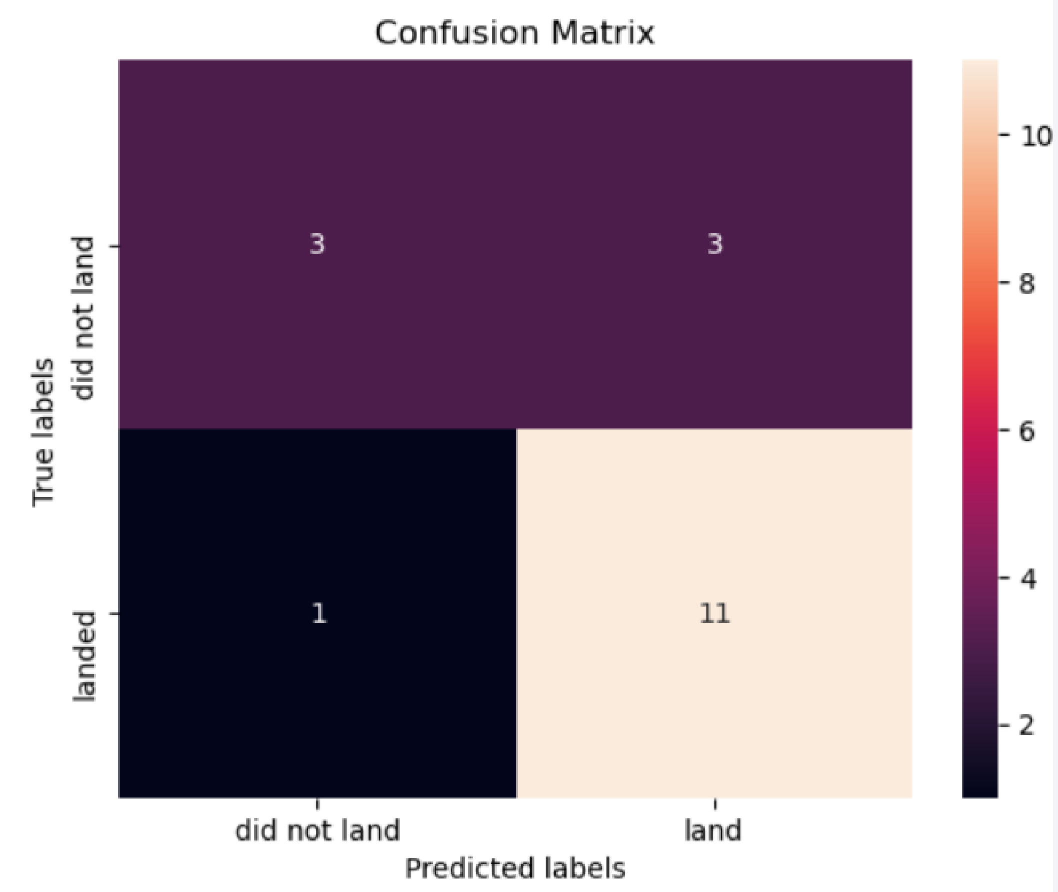
Predictive Analysis (Classification)

Classification Accuracy



- After applying grid search to models, it shows that decision tree model perform the highest accuracy to the prediction.

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

We can conclude that:

- The larger the flight amount at a launch site, the greater the success rate at a launch site.
- Generally, launch success rate increased in the last decade.
- Orbits ES-L1, GEO, HEO, SSO had the most success rate.
- KSC LC-39A had the most successful launches of any sites.
- The Decision tree classifier is the best machine learning algorithm for this task.

Thank you!