IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Babajide Alao
7th July 2022

Outline

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

Executive Summary

- The following methodologies were used to analyze the data:
  - Gathering of data Using Web scraping and SpaceX API
  - Exploratory Data Analysis by wrangling, Visualization, and Interactive dashboard
  - Machine learning Prediction
- The following are the result:
  - EDA helped to identify important features
  - I was able to identify the best model/algorithm for prediction

Introduction

The objective is to evaluate the possibility of the new company SpaceY competing with SpaceX

- Problems you want to find answers:
  - The best location to launch
  - Evaluate the cost of launch by predicting successful landings of the rockets

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was gathered from two sources. i.e.:
    - SpaceX APA
    - Web Scraping
- Perform data wrangling
  - Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing 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
  - Data that was collected until this step were normalized, divided into training and test data sets, and evaluated by four different classification models, the accuracy of each model was evaluated using different combinations of parameters.

Data Collection

- Data sets were collected from Space X API (https://api.spacexdata.com/v4/rockets/)
- Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches), using web scraping technics(Beautiful Soup).

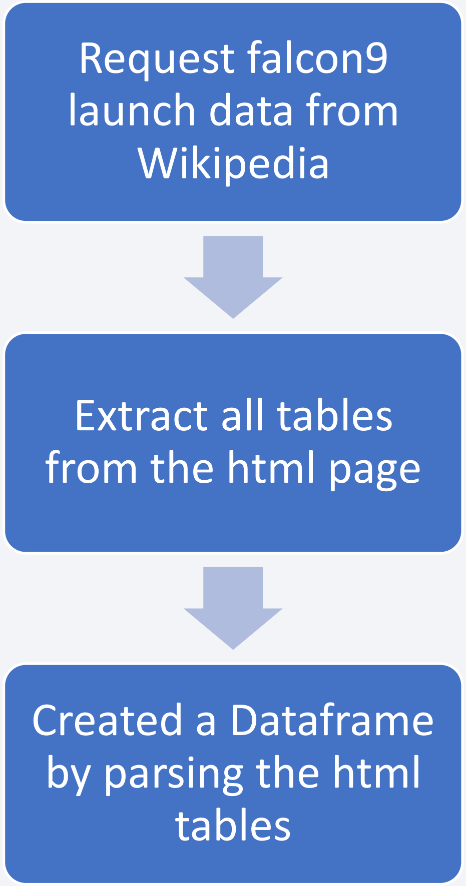
Data Collection – SpaceX API

- SpaceX offers a public API from where data can be obtained and then used
- Source: https://github.com/Mcfoxy/Ibm-Applied-Data-Science-Project/blob/master/Gathering%20from%20Api.ipynb

Data Collection – Scraping

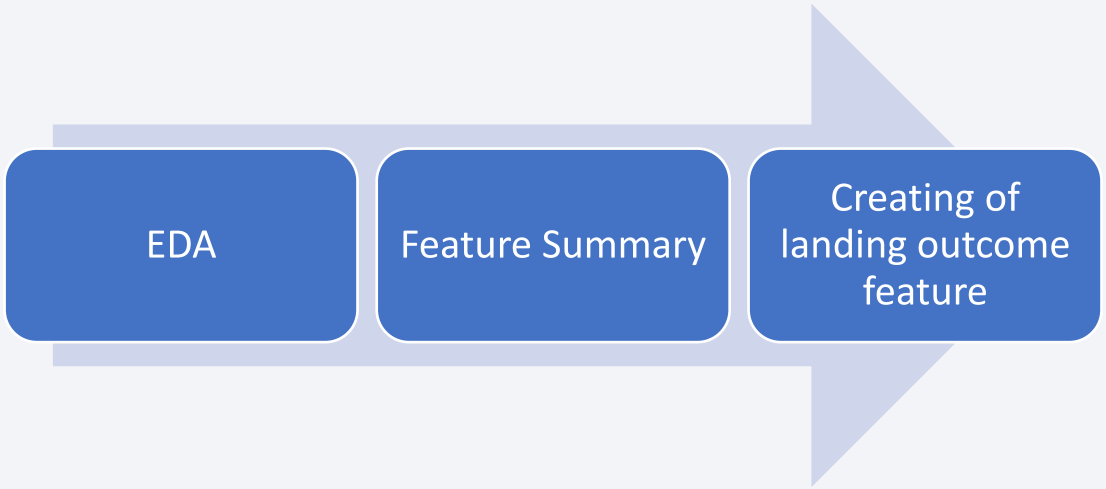
- Data from SpaceX launches can also be obtained from Wikipedia

- Source: https://github.com/Mcfoxy/Ibm-Applied-Data-Science-Project/blob/master/Data%20Collection%20Web%20scraping.ipynb



Data Wrangling

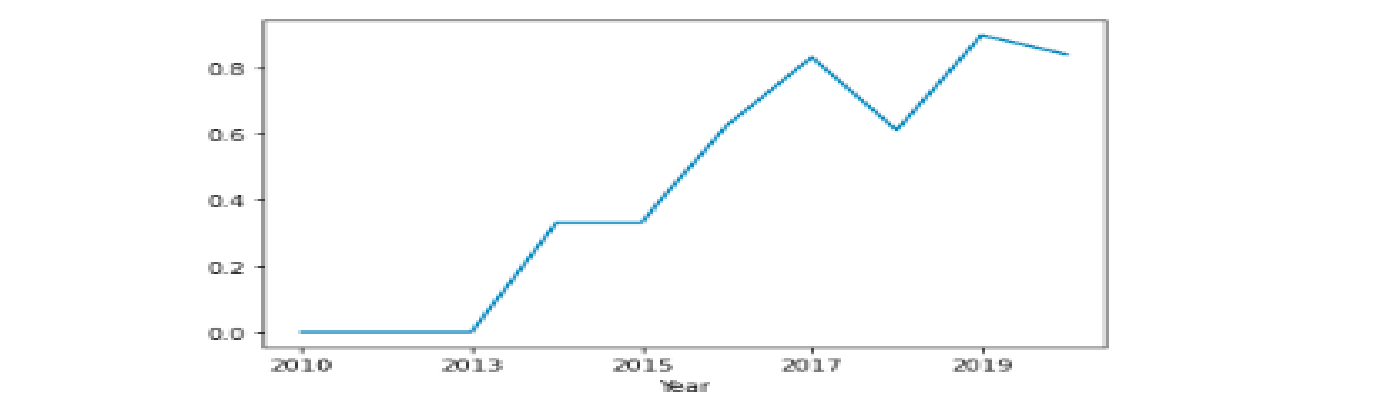
- First I performed exploratory data analysis on some of the columns
- Then summarization of some features
- Lastly, creation of the landing outcome feature



- Source: https://github.com/Mcfoxy/Ibm-Applied-Data-Science-Project/blob/master/Data%20Wrangling%20-%20EDA.ipynb

EDA with Data Visualization

- To explore data, scatterplots and barplots were used to visualize the relationship between pair of features: •
  - Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, Payload and Orbi



- Source: https://github.com/Mcfoxy/Ibm-Applied-Data-Science-Project/blob/master/Data%20Wrangling%20-%20EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- The following SQL queries were performed:
  - ***the names of the unique launch sites in the space mission***
  - ***Top 5 records where launch sites begin with the string 'CCA'***
  - ***the total payload mass carried by boosters launched by NASA (CRS)***
  - ***average payload mass carried by booster version F9 v1.1***
  - ***the date when the first successful landing outcome in ground pad was acheived.***
  - ***the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000***
  - ***the total number of successful and failure mission outcomes***
  - ***the names of the booster_versions which have carried the maximum payload mass. Use a subquery***

  - Source: https://github.com/Mcfoxy/Ibm-Applied-Data-Science-Project/blob/master/Data%20Wrangling%20-%20EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Markers, circles, lines, and marker clusters were used with Folium Maps
  - Markers indicate points like launch sites;
  - Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center;
  - Marker clusters indicate groups of events in each coordinate, like launches in a launch site; and • Lines are used to indicate distances between two coordinates.

- Source: https://github.com/Mcfoxy/Ibm-Applied-Data-Science-Project/blob/master/Data%20Visualization%20With%20Folium.ipynb

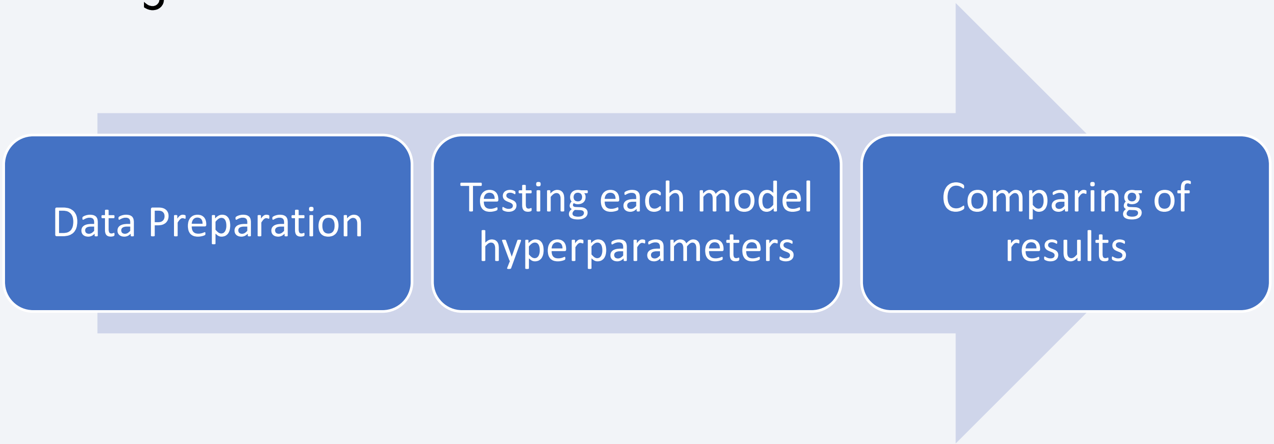
Build a Dashboard with Plotly Dash

- The following graphs and plots were used to visualize data :
    - Percentage of launches by site • Payload range •
    - This combination allowed to quickly analyze the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads.

- Source: https://github.com/Mcfoxy/Ibm-Applied-Data-Science-Project/blob/master/spacex_dash_app%20(1).py

Predictive Analysis (Classification)

- Four classification models were compared: logistic regression, support vector machine, decision tree, and k nearest neighbors.



- Source: https://github.com/Mcfoxy/Ibm-Applied-Data-Science-Project/blob/master/Machine%20Learning%20Prediction.ipynb

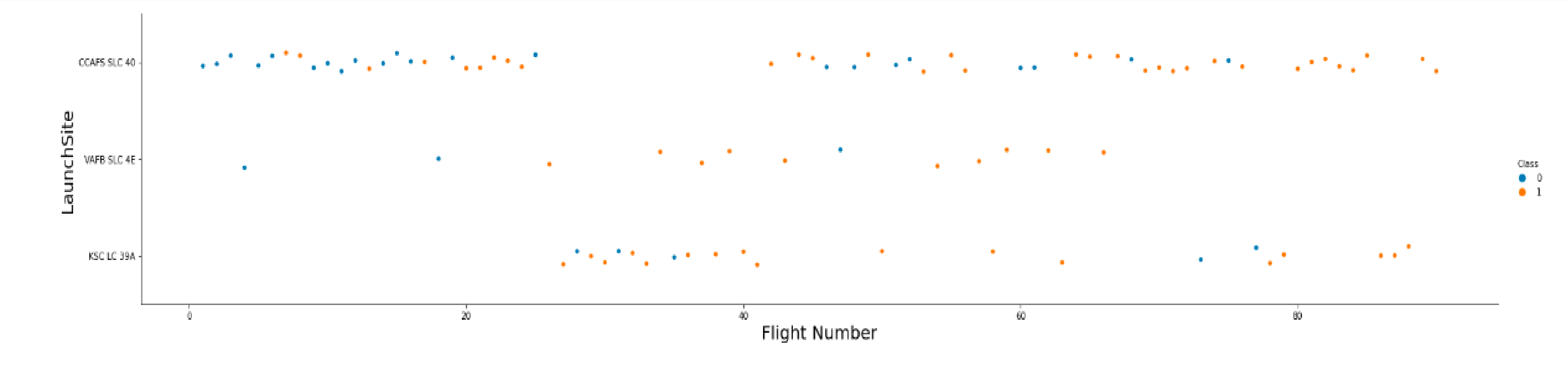
Results

- Exploratory data analysis results:
  - Space X uses 4 different launch sites;
  - The first launches were done to Space X itself and NASA;
  - The average payload of the F9 v1.1 booster is 2,928 kg;
  - The first successful landing outcome happened in 2015 five year after the first launch;
  - Many Falcon 9 booster versions were successful at landing in drone ships having payloads above the average;
  - Almost 100% of mission outcomes were successful;
  - Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015;
  - The number of landing outcomes became as better as years passed

Section 2

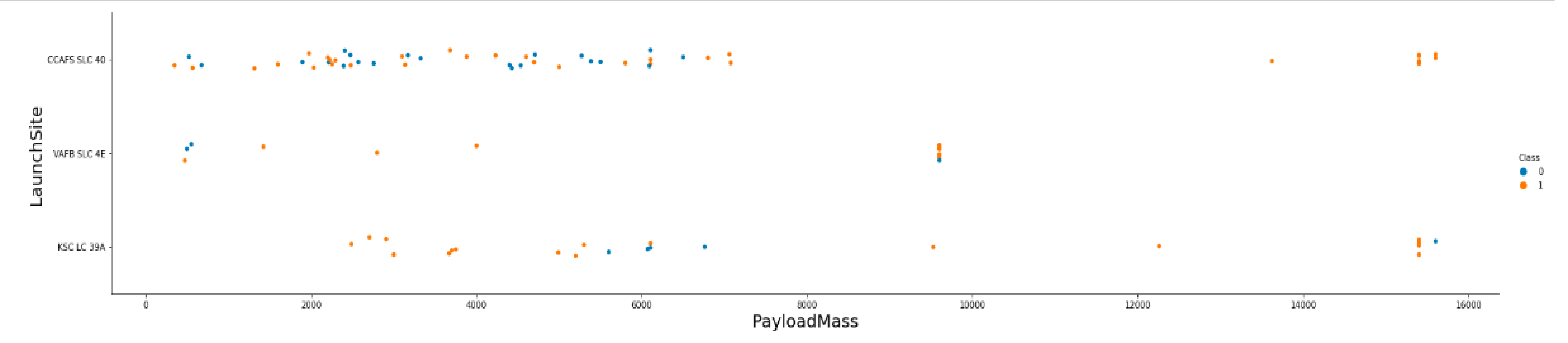
Insights drawn from EDA

Flight Number vs. Launch Site



- According to the plot above, it's possible to verify that the best launch site nowadays is CCAF5 SLC 40, where most of the recent launches were successful;
- In second place VAFB SLC 4E and third place KSC LC 39A;
- It's also possible to see that the general success rate improved over time.

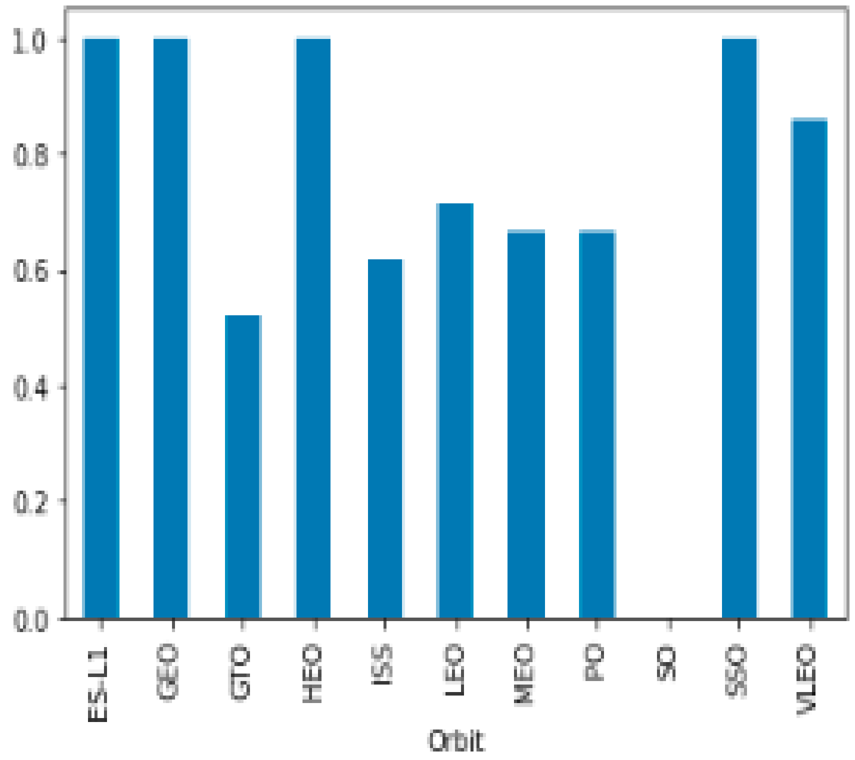
Payload vs. Launch Site



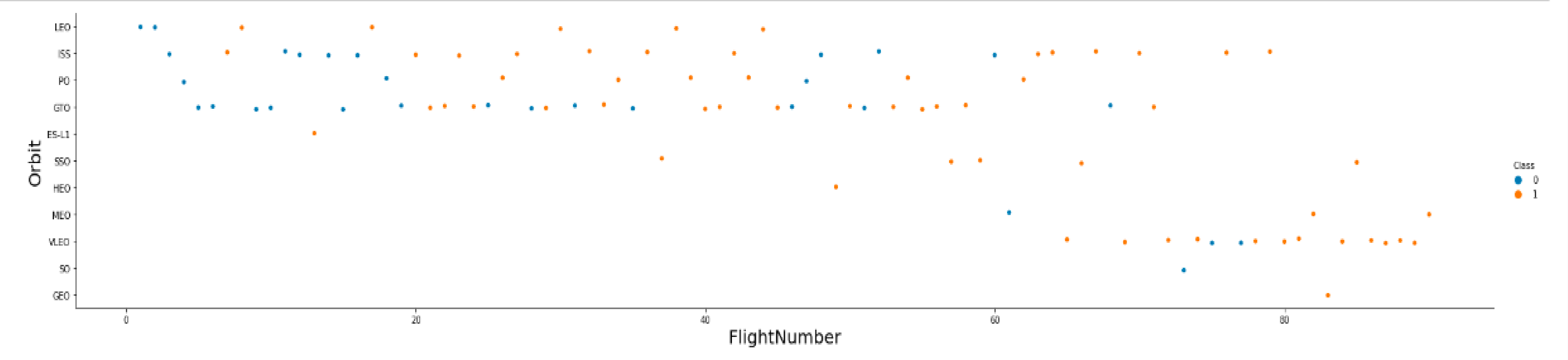
- Payloads over 9,000kg (about the weight of a school bus) havean excellent success rate;
- Payloads over 12,000kg seem to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites

Success Rate vs. Orbit Type

- The biggest success rates happens to orbits:
  - ES-L1;
  - GEO;
  - HEO; and
  - SSO.

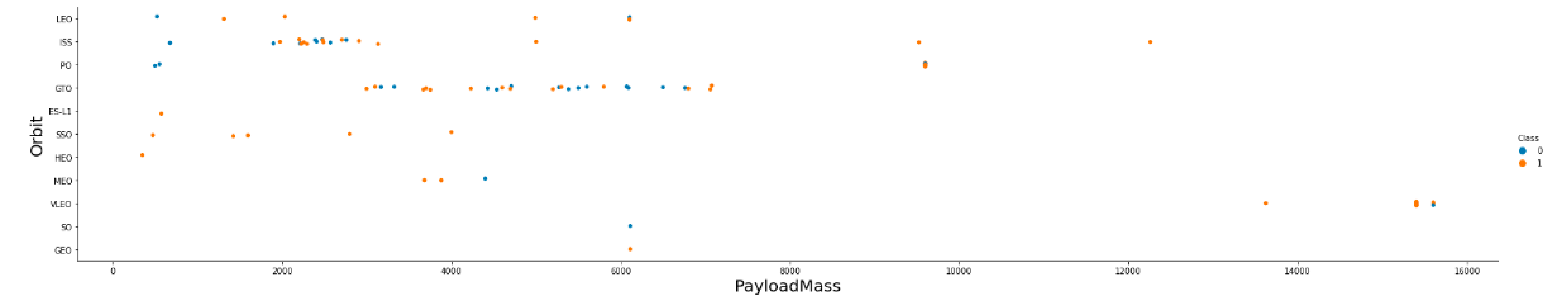


Flight Number vs. Orbit Type



- The success rate improved over time to all orbits

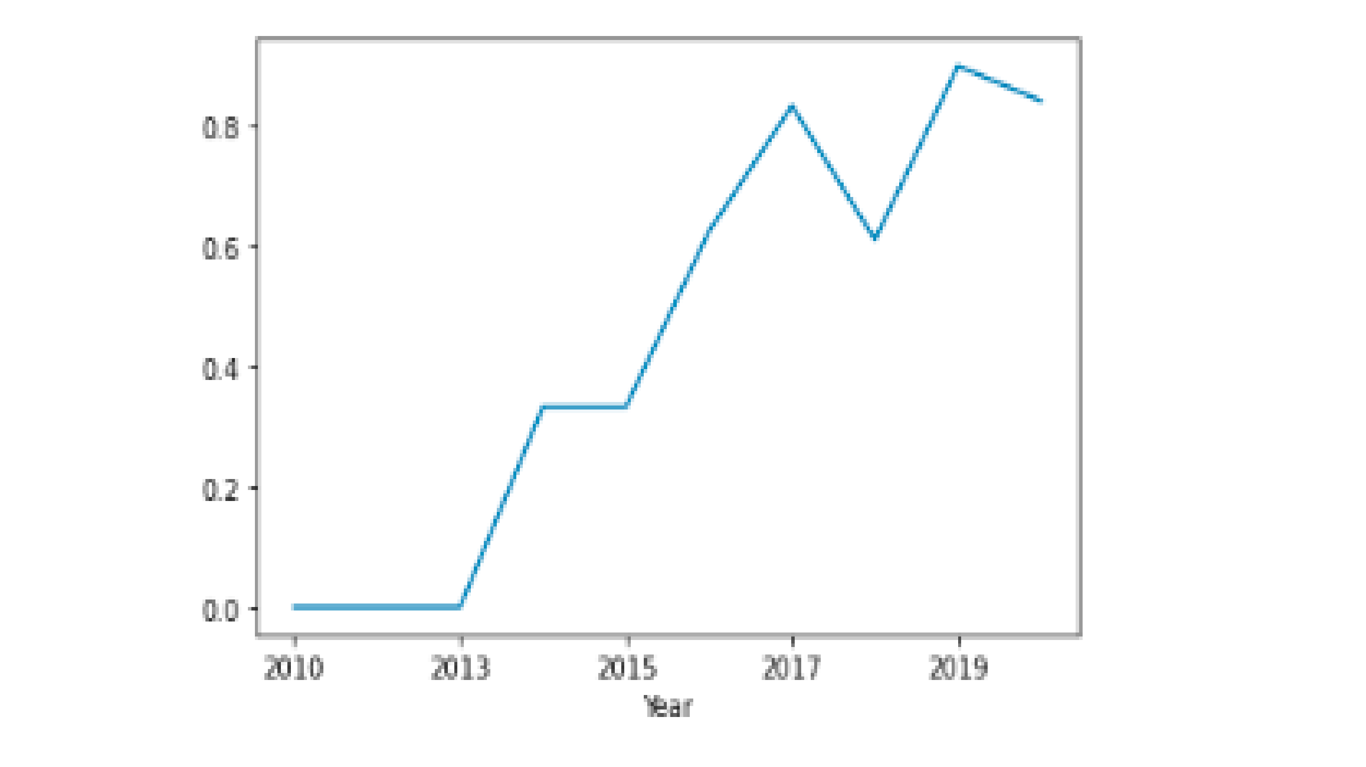
Payload vs. Orbit Type



- there is no relation between payload and success rate to orbit GTO;
- ISS orbit has the widest range of payload and a good rate of success;
- There are few launches to the orbits SO and GEO.

Launch Success Yearly Trend

- Success rate started increasing in 2013 and kept until 2020



All Launch Site Names

```
sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
```

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

Launch Site Names Begin with 'CCA'

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

Out[6]:

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)
2012-08-10	00:35:00	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
2013-03-12	22:41:00	F9 v1.1	CCAFS LC-40	SES-8	3170	GTO	SES	Success	No attempt

Total Payload Mass

```sql
sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE
PAYLOAD LIKE '%CRS%';
```

total_payload
58479

Average Payload Mass by F9 v1.1

```sql
sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL
WHERE BOOSTER_VERSION = 'F9 v1.1';
```

avg_payload
3676

First Successful Ground Landing Date

```sql
sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL WHERE
LANDING__OUTCOME = 'Success (ground pad)';
```

first_success_gp
2017-01-05

Successful Drone Ship Landing with Payload between 4000 and 6000

```
sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE
PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND LANDING__OUTCOME =
'Success (drone ship)';
```

booster_version
F9 FT B1031.2
F9 FT B1022

Total Number of Successful and Failure Mission Outcomes

```
sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL GROUP BY
MISSION_OUTCOME ORDER BY MISSION_OUTCOME;
```

mission_outcome	qty
Success	44
Success (payload status unclear)	1

Boosters Carried Maximum Payload

sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL) ORDER BY BOOSTER_VERSION;

booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1058.3
F9 B5 B1060.2

2015 Launch Records

```
sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE
LANDING__OUTCOME = 'Failure (drone ship)' AND DATE_PART('YEAR', DATE) =
2015;
```

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40

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

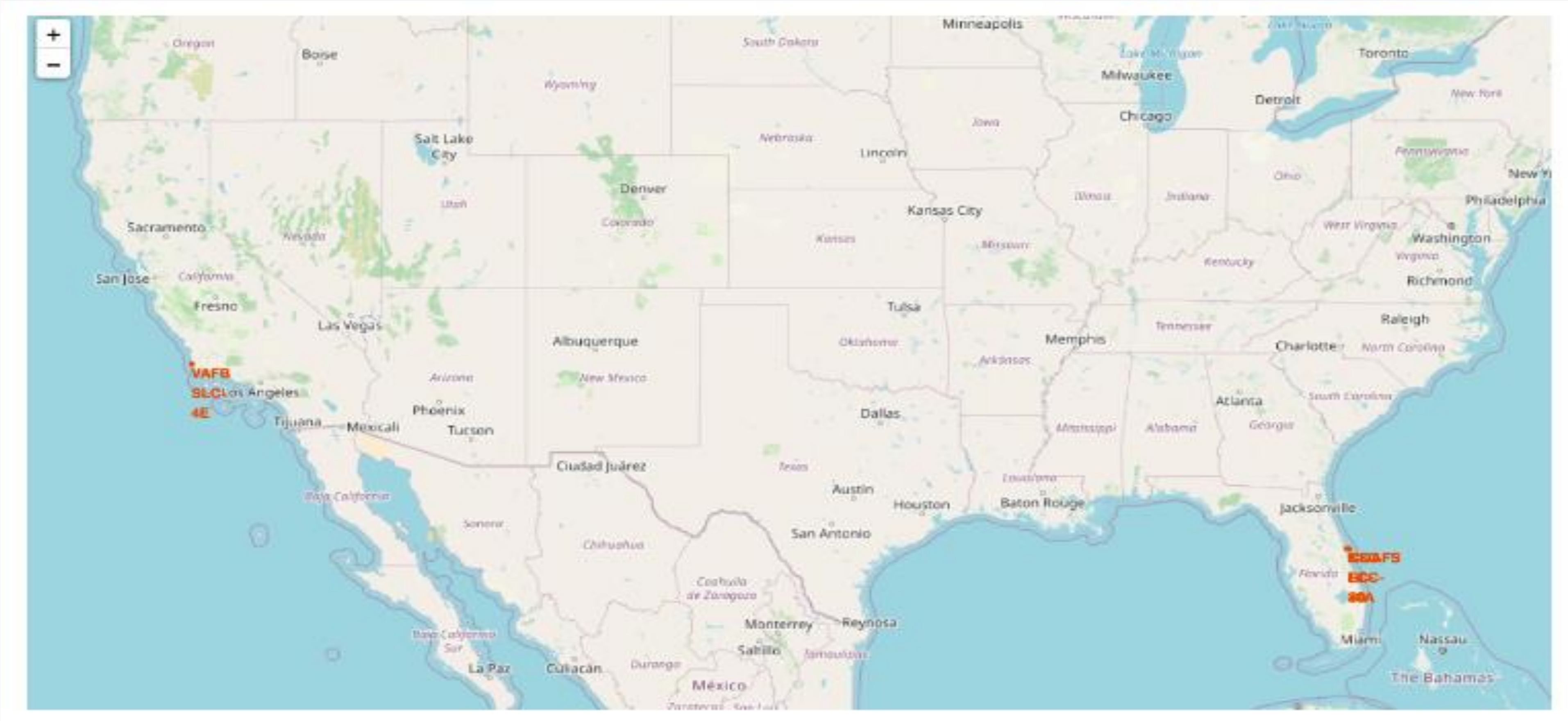
sql SELECT LANDING__OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING__OUTCOME ORDER BY QTY DESC;

landing__outcome	qty
No attempt	7
Failure (drone ship)	2
Success (drone ship)	2
Success (ground pad)	2
Controlled (ocean)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

Map of All the Launch Cites



Map of Success/Failed Launches

Distance Between Launch site and Proximities

Section 5

Predictive Analysis (Classification)

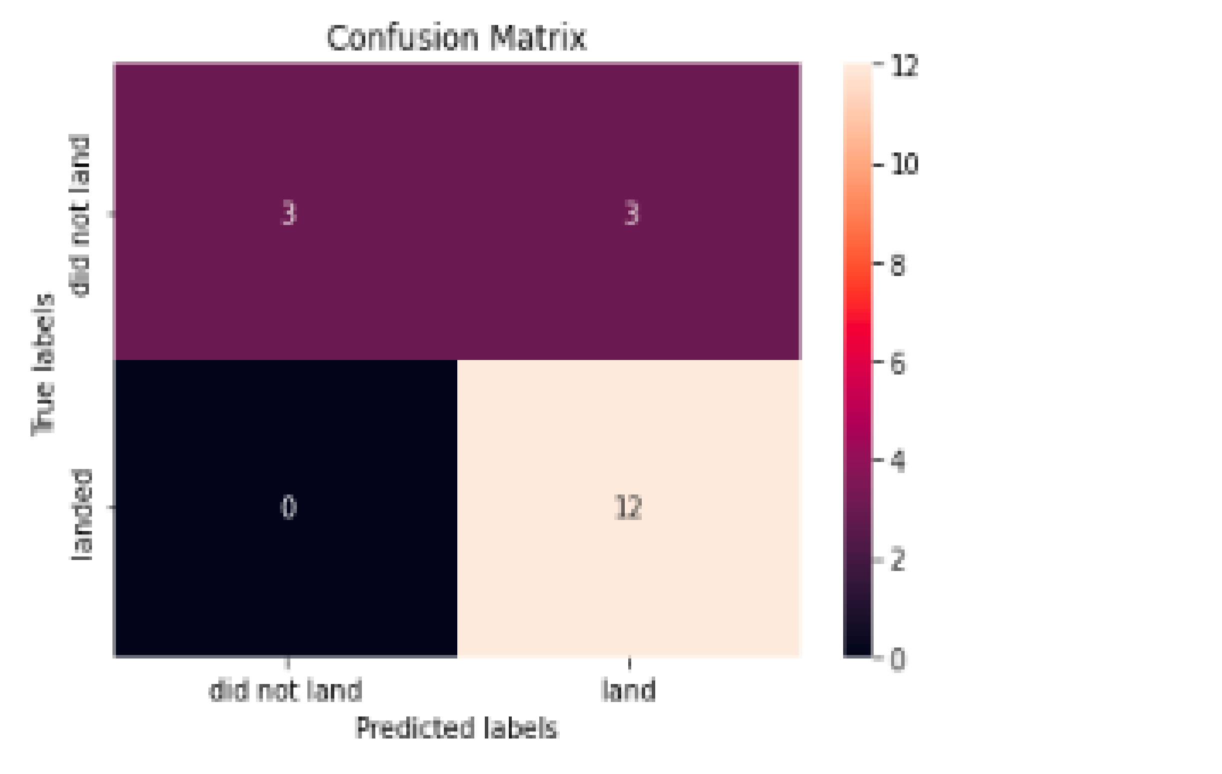
Classification Accuracy

- Four classification models were tested, and their accuracies are plotted beside;
- The logistic regression, SVM and KNN has the highest accuracy with 83%

Model	Accuracy	TestAccuracy
LogReg	0.84643	0.83333
SVM	0.84821	0.83333
Tree	0.875	0.66667
KNN	0.84821	0.83333

Confusion Matrix

- Show the confusion matrix of the best performing model with an explanation



Conclusions

- Different data sources were analyzed, refining conclusions along the process;
- The best launch site is KSC LC-39A;
- Launches above 7,000kg are less risky;
- Although most of the mission outcomes are successful, successful landing outcomes seem to improve over time, according to the evolution of processes and rockets

Thank you!