Winning Space Race with Data Science

Aryan Mehna
09/09/21
Github Url=https://github.com/aryanmehna/DS-capstone-project.git

Outline

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

Executive Summary

- Collected data from SpaceX API and wiki page.
- Created labels column ‘class’ which classifies successful landings.
- Explore data using SQL.
- Data visualization , folium maps, and dashboards.
- Changed all categorical variables to binary(one hot encoding).
- Standardized data ,gridSearchCV to find best parameters for ML models.
- Accuracy score of all models.

Result Summary

- The process produced the following models: Logistic Regression, SVM, Decision Tree Classifier, and K nearest Neighbors.
- All the models had similar outcomes and showed an accuracy of 83.33%circa.
- All models over predicted successful landings.
- More data is needed for better model determination and accuracy.

Introduction

Background

- Space X has best pricing
- Space Y is competing with Space X

Problem

- Space Y tasks us to train a machine learning model to predict successful stage 1 recovery.

Section 1

Methodology

Methodology

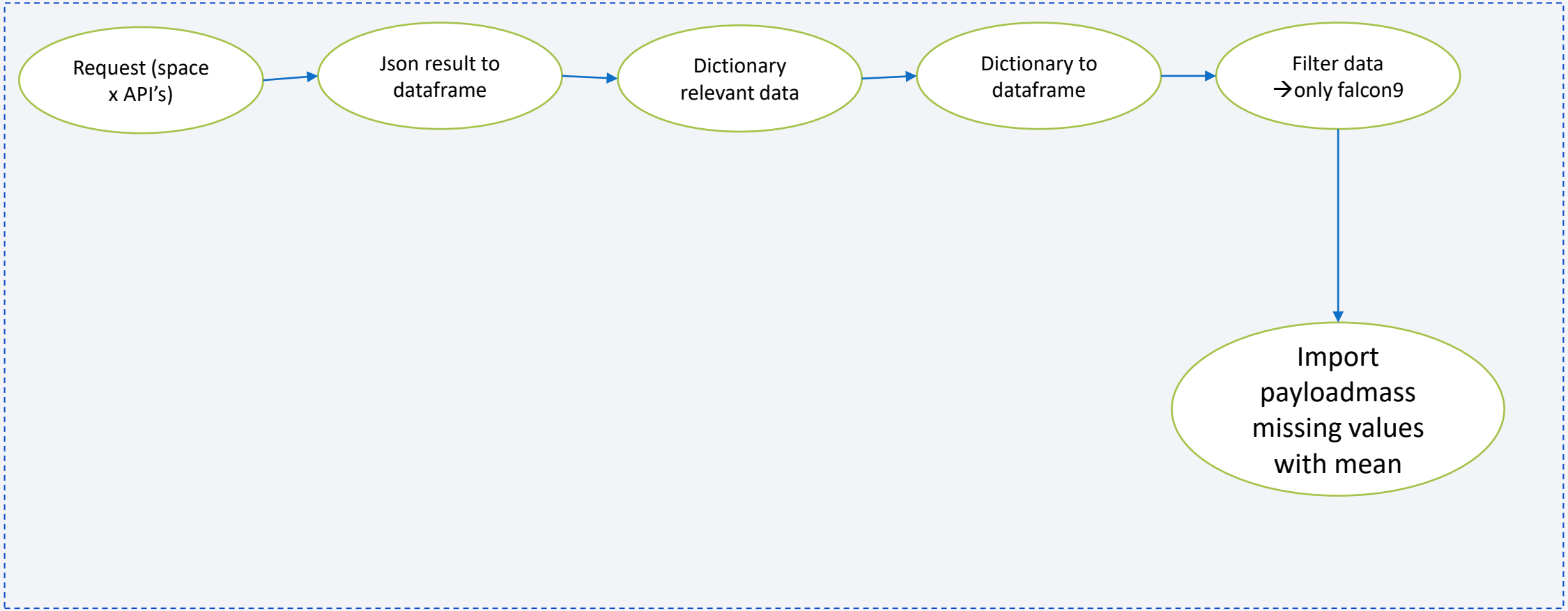
Executive Summary

- Data collection methodology:
  - Combined data from SpaceX public API and SpaceX wiki page.
- Perform data wrangling
  - Classifying True landings and as successful and unsuccessful otherwise
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Tuned models using GridSearchCV

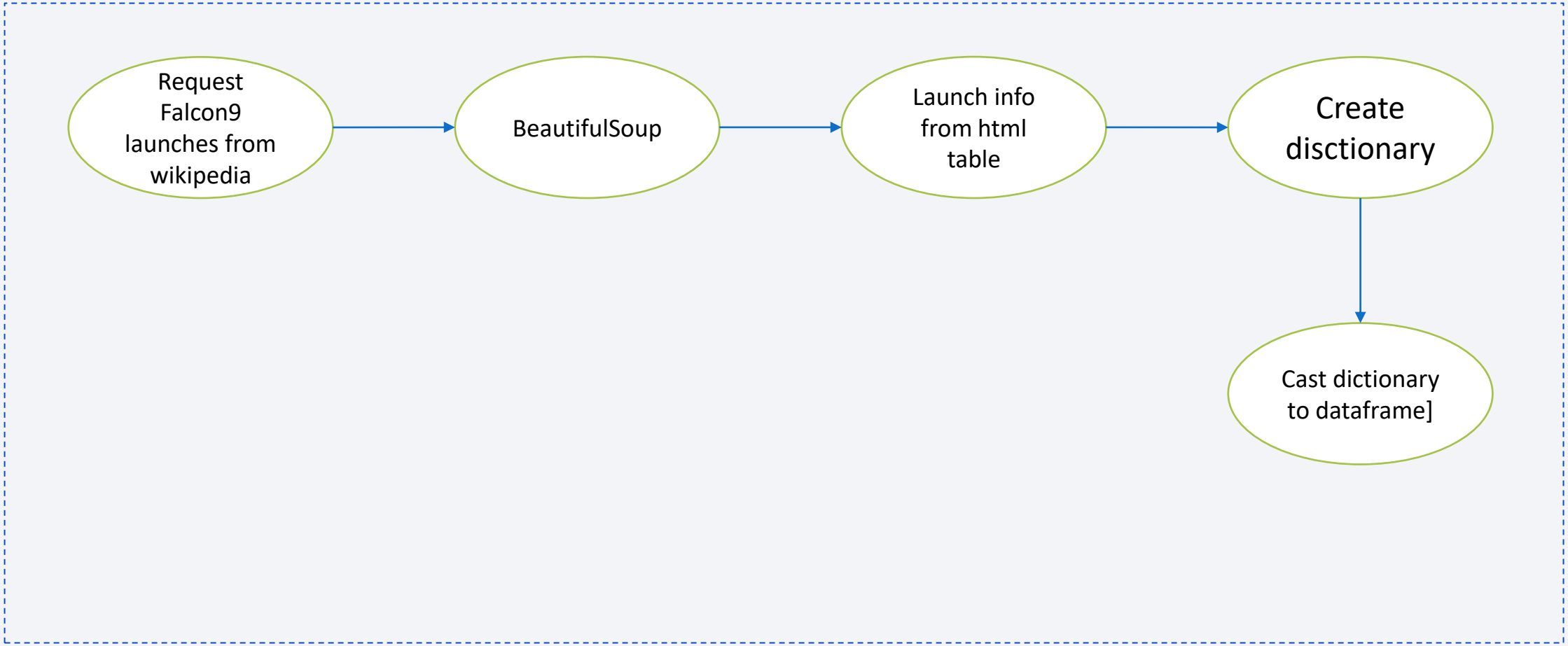
Data Collection Overview

- Data collection process had 2 steps:
    1. API Request from SpaceX public API.
    2. Web scrapping the SpaceX's Wiki page.
- The next slide will show the flowchats of the data collection.
- Git URL=https://github.com/aryanmehna/DS-capstone-project/blob/master/capstone%20project.ipynb

Data Collection – SpaceX API



Data Collection - Scraping



Data Wrangling

- Create a training label with landing outcomes where successful=1& failure = 0.
- Outcome column contains : ‘Mission outcome’ ‘Landing Location’.
- New training label column ‘class’ with a value of 1 if ‘Mission Outcome’ is true and 0 otherwise
- Value Mapping:
  - True ASDS ,True RTLS & True Ocean are set to be 1 because they represent a successful landing
  - None None ,False ASDS, None ASDS, False Ocean , False RTLS are set to be 0 because they represent a failed attempt.

- Git URL=https://github.com/aryanmehna/DS-capstone-project/blob/master/capstone%20project%20part%202.ipynb

EDA with Data Visualization

- Exploratory data analysis performed on variables with the help of some interesting plots in order to understand correlations between variables.
- The following plots were done:
    1. Flight Number vs Payload Mass
    2. Flight Number vs Launch Site
    3. Payload Mass vs Launch Site
    4. Orbit vs Success Rate
    5. Flight Number vs Orbit
    6. Payload vs Orbit
    7. Success Yearly Trend
- Scatter plots , line charts, and bar plots were used to compare relationships between variables to decide if a relation exists so that they could be used in training the machine learning model.

EDA with SQL

1. The dataset was loaded into IBM DB2 database through IBM cloud.
2. Performed 10 queries thanks to the python integration.

The objective of this task was to obtain a better understanding of the dataset.

Queried info regarding launch site names, mission outcomes, various pay load sizes of customers and booster versions and landing outcomes.

Build an Interactive Map with Folium

- Folium maps mark launch sites ,successful and unsuccessful landiongs and a proximity example to key locations: Railway,Highway,Coast and City
- This allows us to understand why launch sites may be located where they are . Also visualizes successful landings relative to location.

Build a Dashboard with Plotly Dash

The dashboard created contains a pile chart and scatterplots.

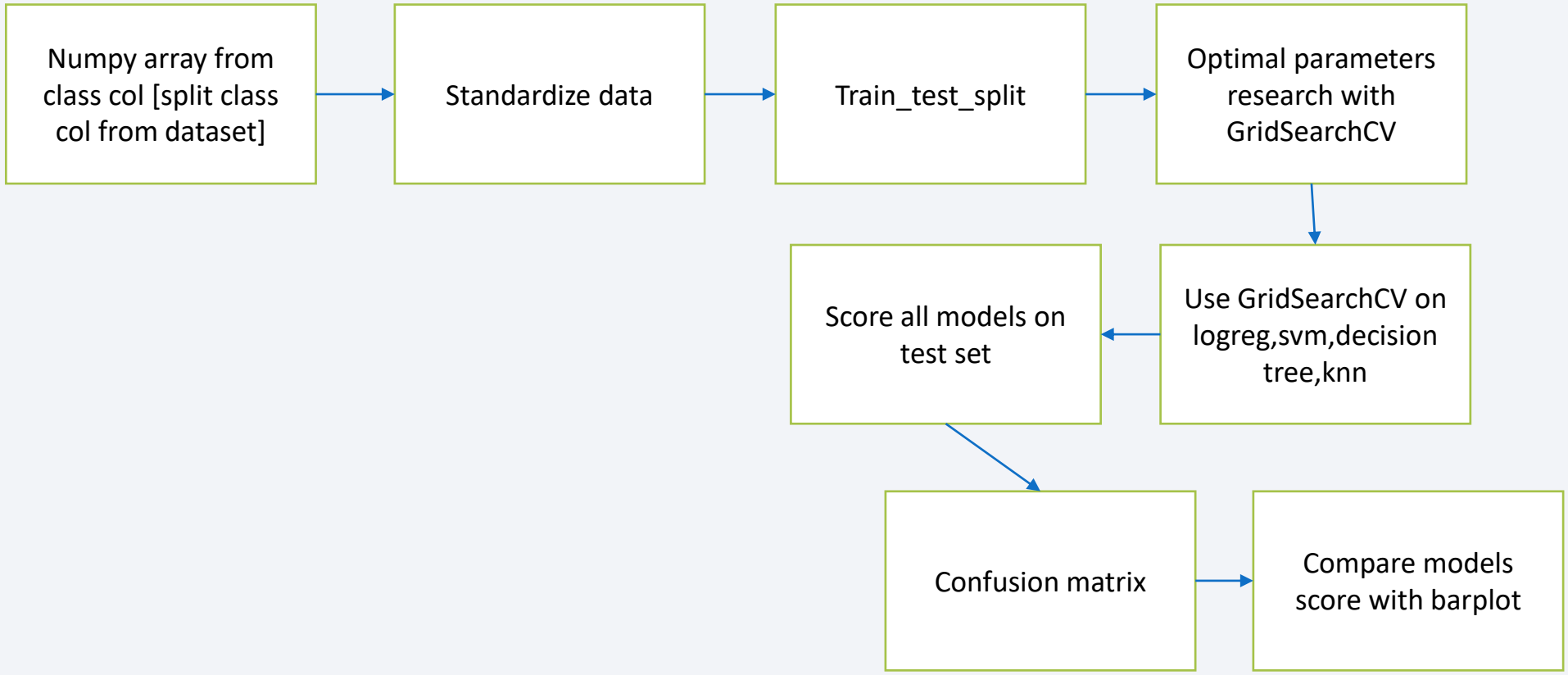
Pie chart can be used to show distribution of successful landings for all launch sites or also for individual launch site.

Scatterplots are made in order to visualize relationships between launch sites and payload mass.

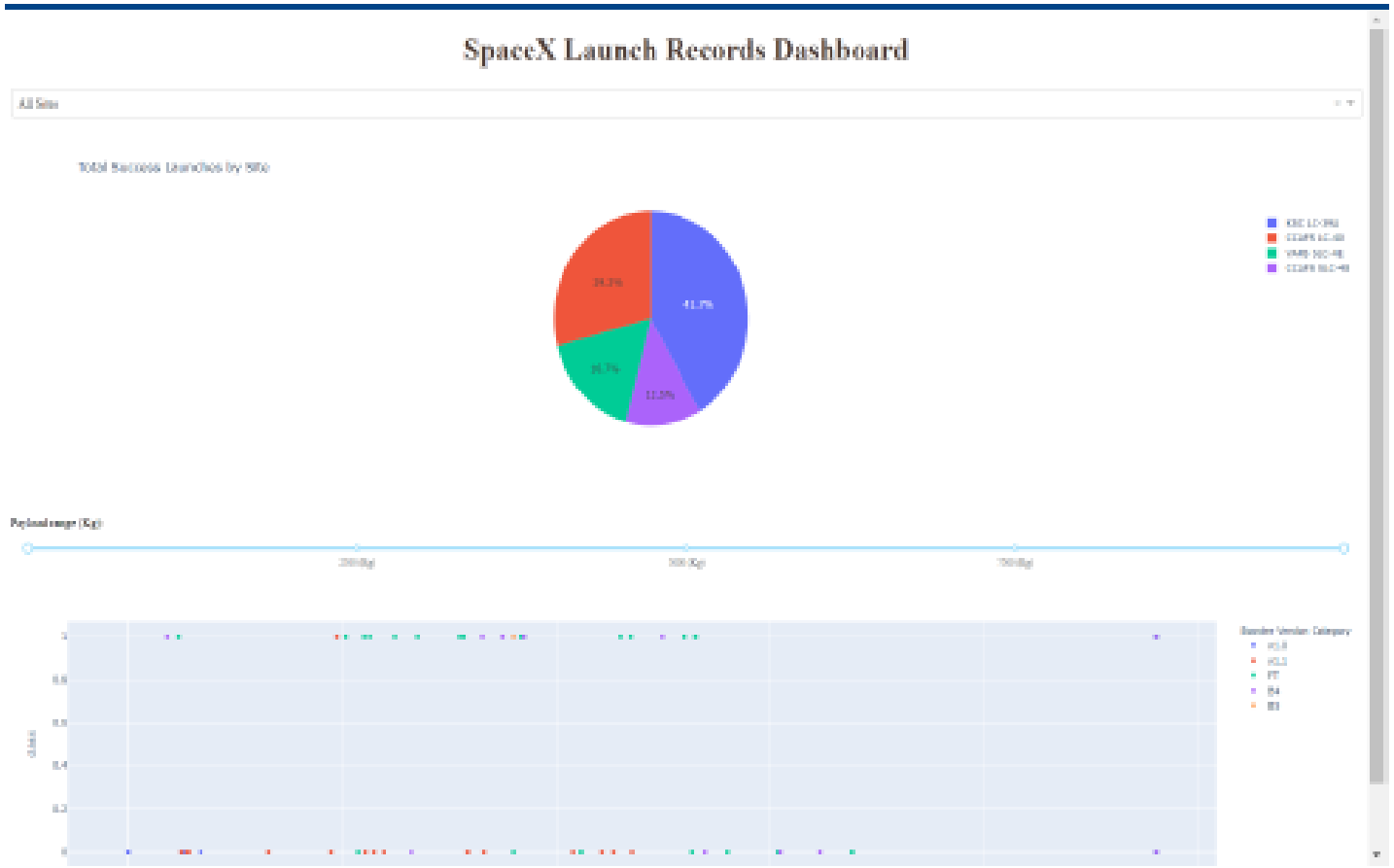
The pie chart is used to visualize launch site success rate.

Predictive Analysis (Classification)

- Flowchart:

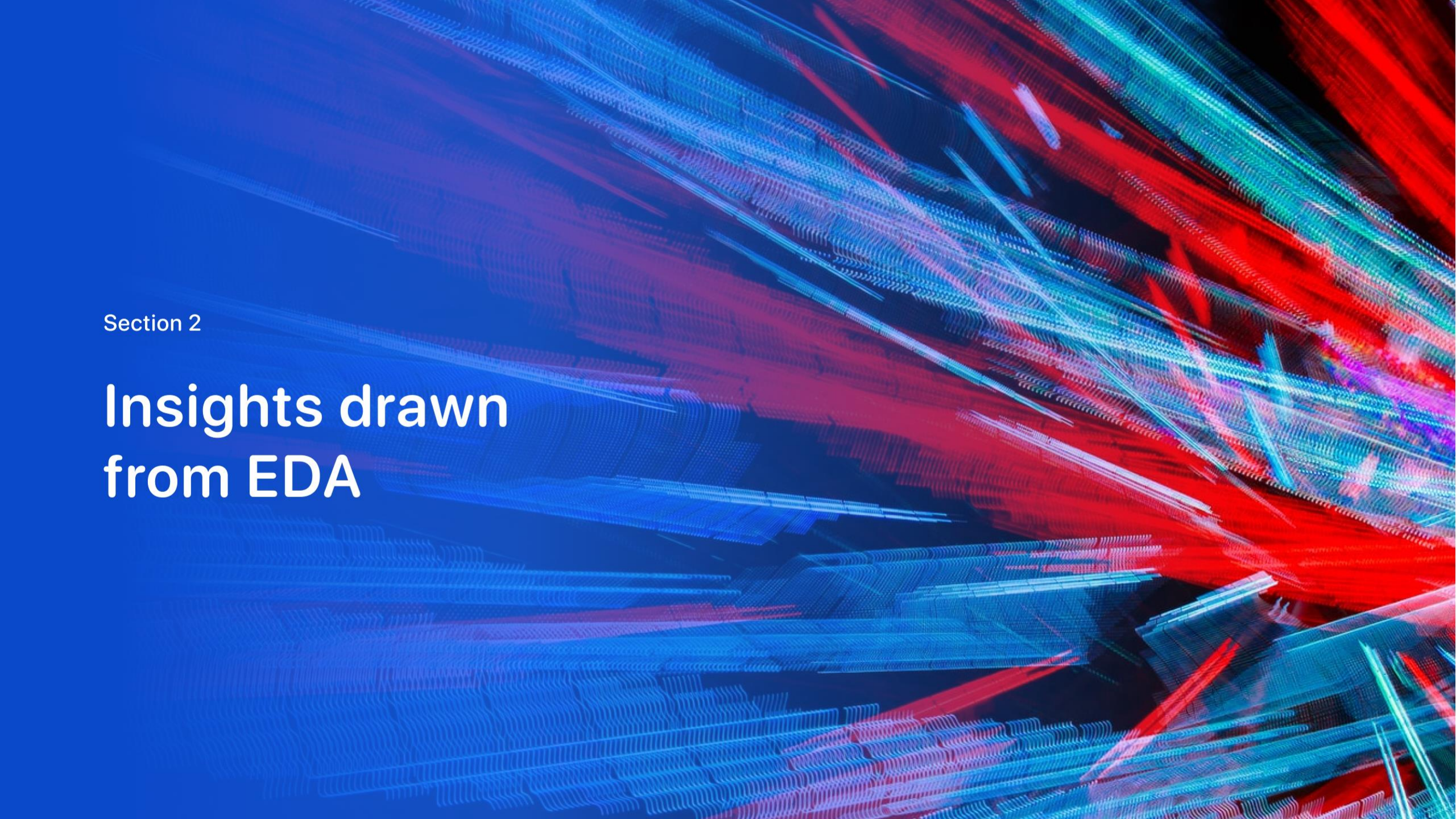


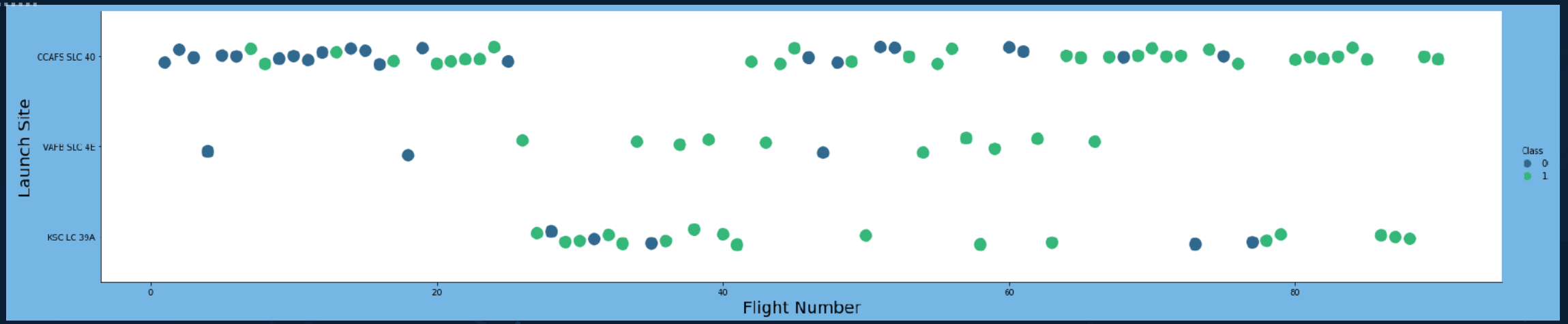
Results



Section 2

Insights drawn from EDA

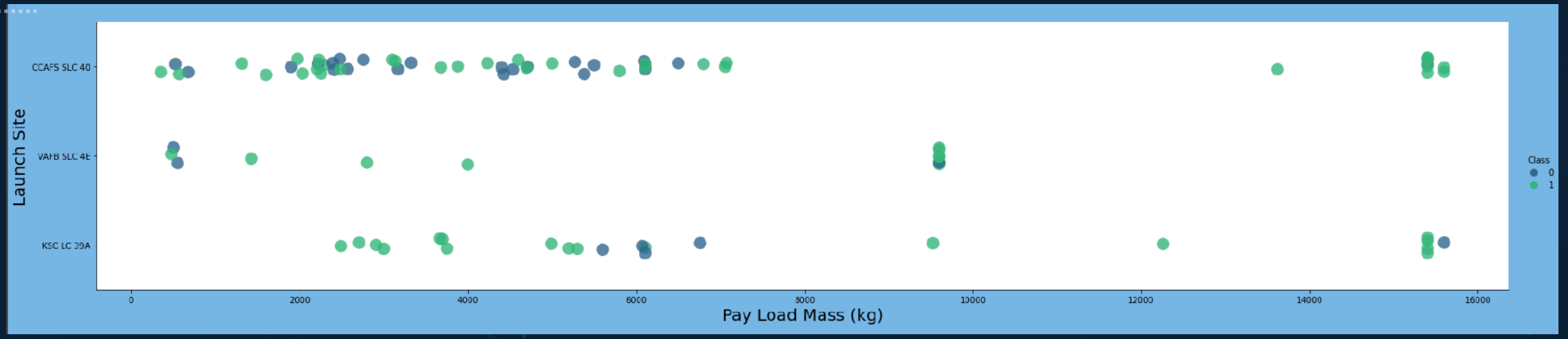




Flight Number vs. Launch Site

The plot can indicate that there was an increase of success over time (Flight Number).

We can observe a breakthrough from flight 20 on which swignificantly increased success rate.

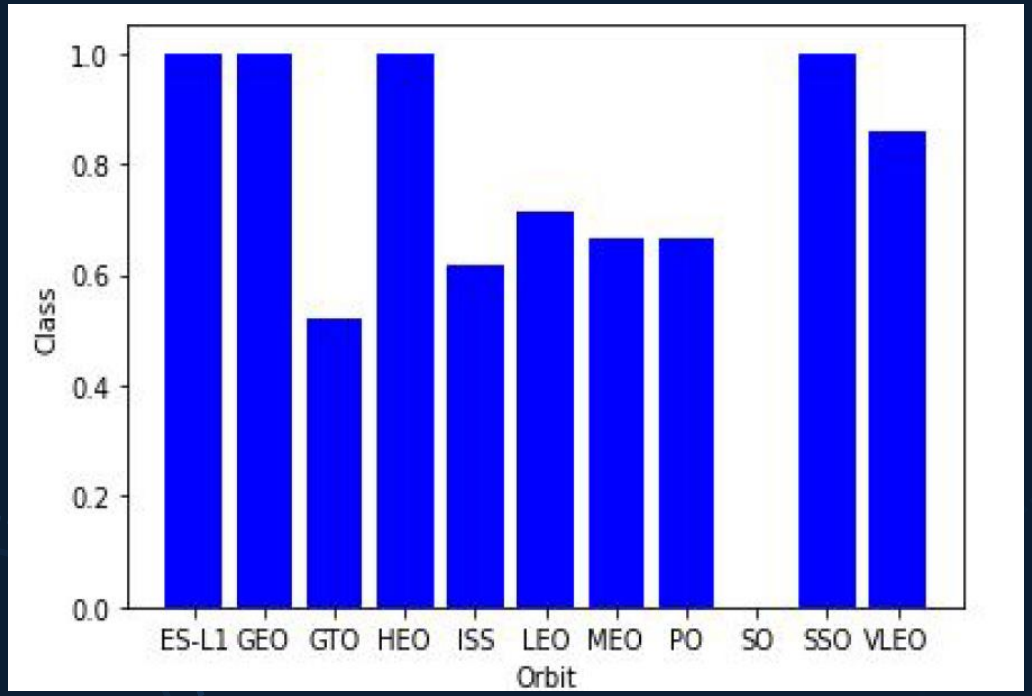


Payload vs. Launch Site

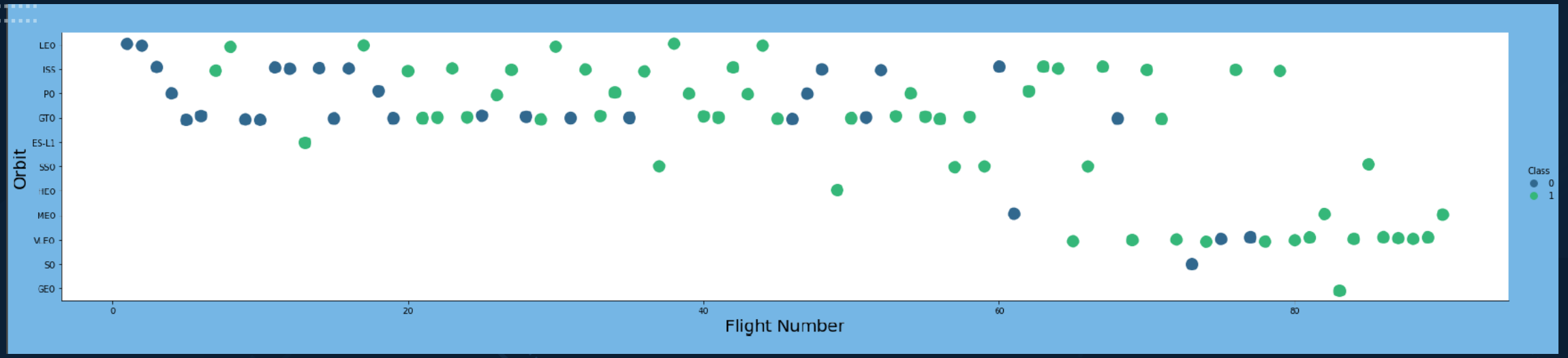
Class 0 = Unsuccessful launches

Class 1 = Successful launches

Success Rate vs. Orbit Type



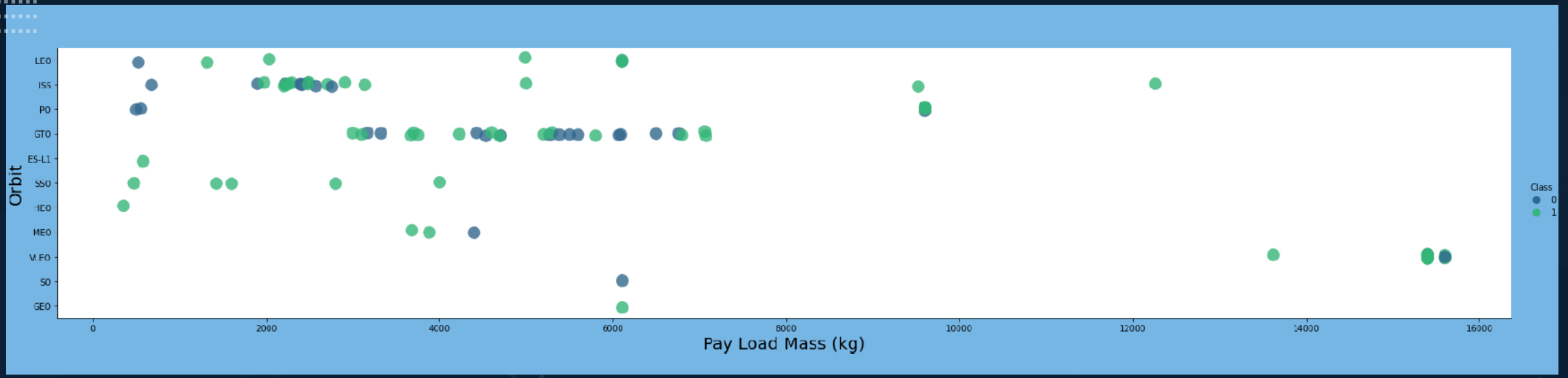
- The plot above shows the different success rate for the various orbit type.
- Most success orbit type : ES-L1,GEO,HEO and SSO



Flight Number vs. Orbit Type

Class 0 = Unsuccessful launches

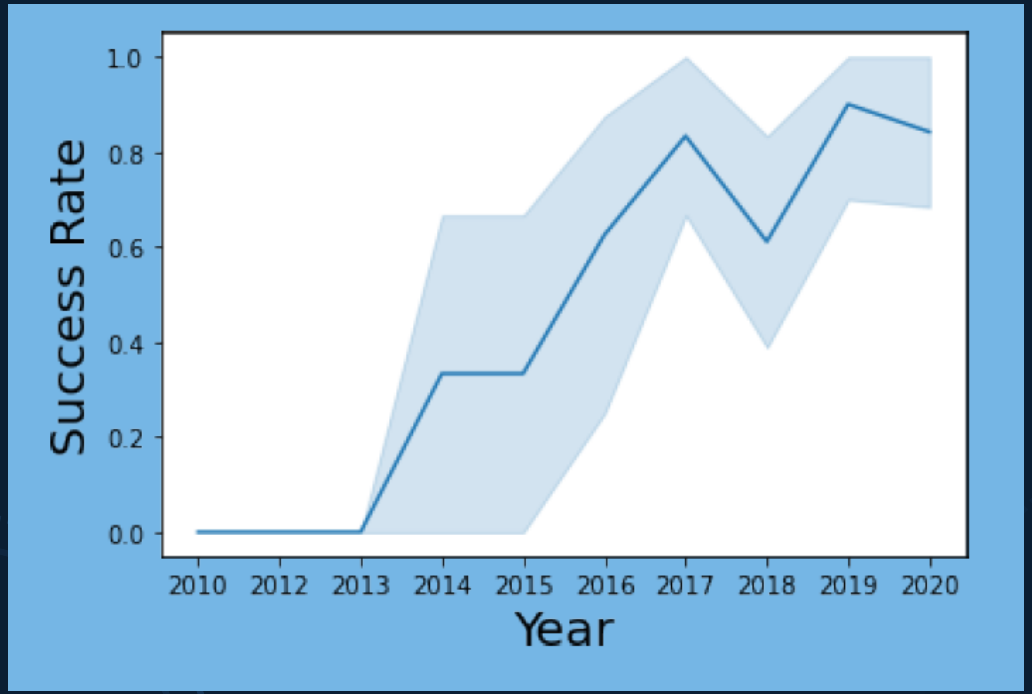
Class 1 = Successful launches



Payload vs. Orbit Type

- Class 0 = Unsuccessful launches
- Class 1 = Successful launches

Launch Success Yearly Trend



- We can observe that success increased continuously from 2013 to 2018when there was a small fall. Then from 2018 to 2019 there was another important increase that then slightly slowed


EDA WITH SQL

All Launch Site Names

- Find the names of the unique launch sites

Task 1

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

```
%sql select DISTINCT LAUNCH_SITE from SPACEXDATASET
```

```
 * ibm_db_sa://nxs27972:***@54a2f15b-5c0f-46df-8954-7e38e61
Done.
```

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

Task 2

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

```
%sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5
```

```
 * ibm_db_sa://nxs27972:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/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

Task 3

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

```
%sql select sum(payload_mass__kg_) as sum from SPACEXDATASET where customer like 'NASA (CRS)'
```

```
 * ibm_db_sa://nxs27972:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases
Done.
```

SUM
45596

Average Payload Mass by F9 v1.1

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

Task 4

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

```
%sql select avg(payload_mass__kg_) as Average from SPACEXTBL where booster_version = 'F9 v1.1'
```

```
 * ibm_db_sa://vwm02263:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases
Done.
```

average
2928

First Successful Ground Landing Date

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

Task 5

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

Hint:Use min function

```
%sql select min(date) as Date from SPACEXTBL \
where landing__outcome like 'Success (ground pad)'
```

```
 * ibm_db_sa://vwm02263:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l0
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

Task 6

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 SPACEXTBL \
where (LANDING__OUTCOME = 'Success (drone ship)') and (PAYLOAD_MASS__KG_  between 4000 and 6000)
```

```
 * ibm_db_sa://vwm02263:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.clou
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

- Calculate the total number of successful and failure mission outcomes

Task 7

List the total number of successful and failure mission outcomes

```
%sql SELECT mission_outcome, count(*) as Count FROM SPACEXTBL GROUP by mission_outcome ORDER BY mission_outcome
```

```
 * ibm_db_sa://vwm02263:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/b
Done.
```

mission_outcome	COUNT
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

Task 8

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

```
%sql select distinct(BOOSTER_VERSION) \
from SPACEXTBL \
where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

```
 * ibm_db_sa://vwm02263:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databa
Done.
```

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

2015 Launch Records

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

Task 9

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

```
%sql select MONTHNAME(DATE) as Month, landing__outcome, booster_version, launch_site \
from SPACEXTBL \
where DATE like '2015%' AND landing__outcome like 'Failure (drone ship)'
```

```
 * ibm_db_sa://vwm02263:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.app
Done.
```

MONTH	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

- 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

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

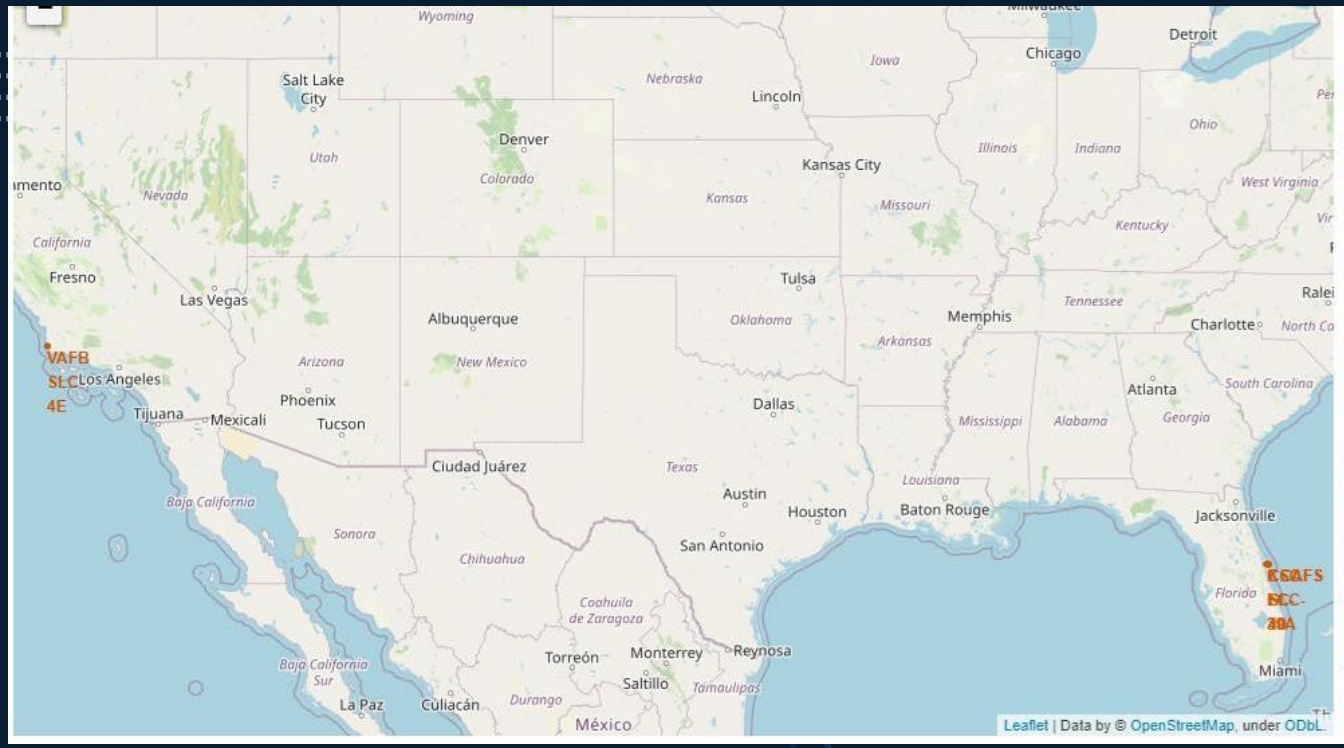
```
%sql select landing__outcome, count(*) as count \
from SPACEXTBL \
where Date >= '2010-06-04' AND DATE <= '2017-03-20' \
GROUP by landing__outcome ORDER BY count Desc
```

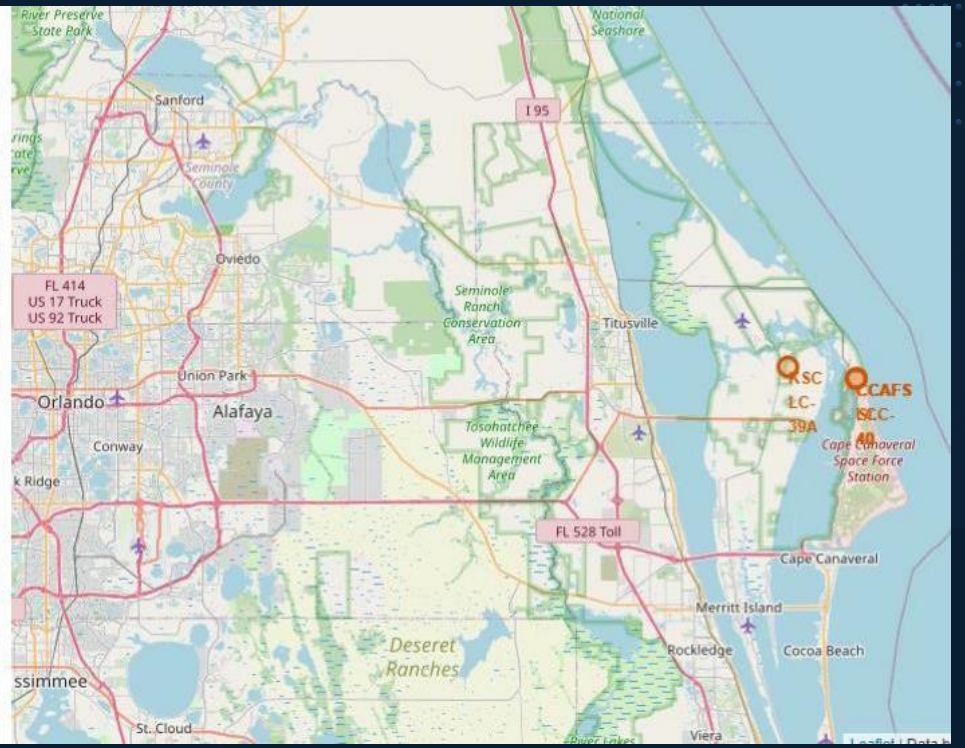
```
 * ibm_db_sa://vwm02263:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

landing__outcome	COUNT
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 4

Launch Sites Proximities Analysis





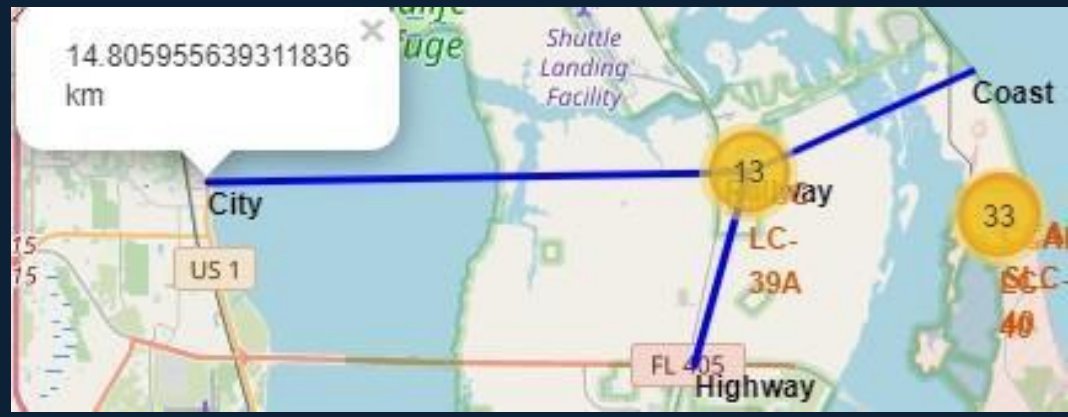
LAUNCH SITE LOCATIONS

- The left map shows all launch sites relative us maps.
- The right maps shows the two florida launch sites.

COLOR CODED LAUNCH MARKERS

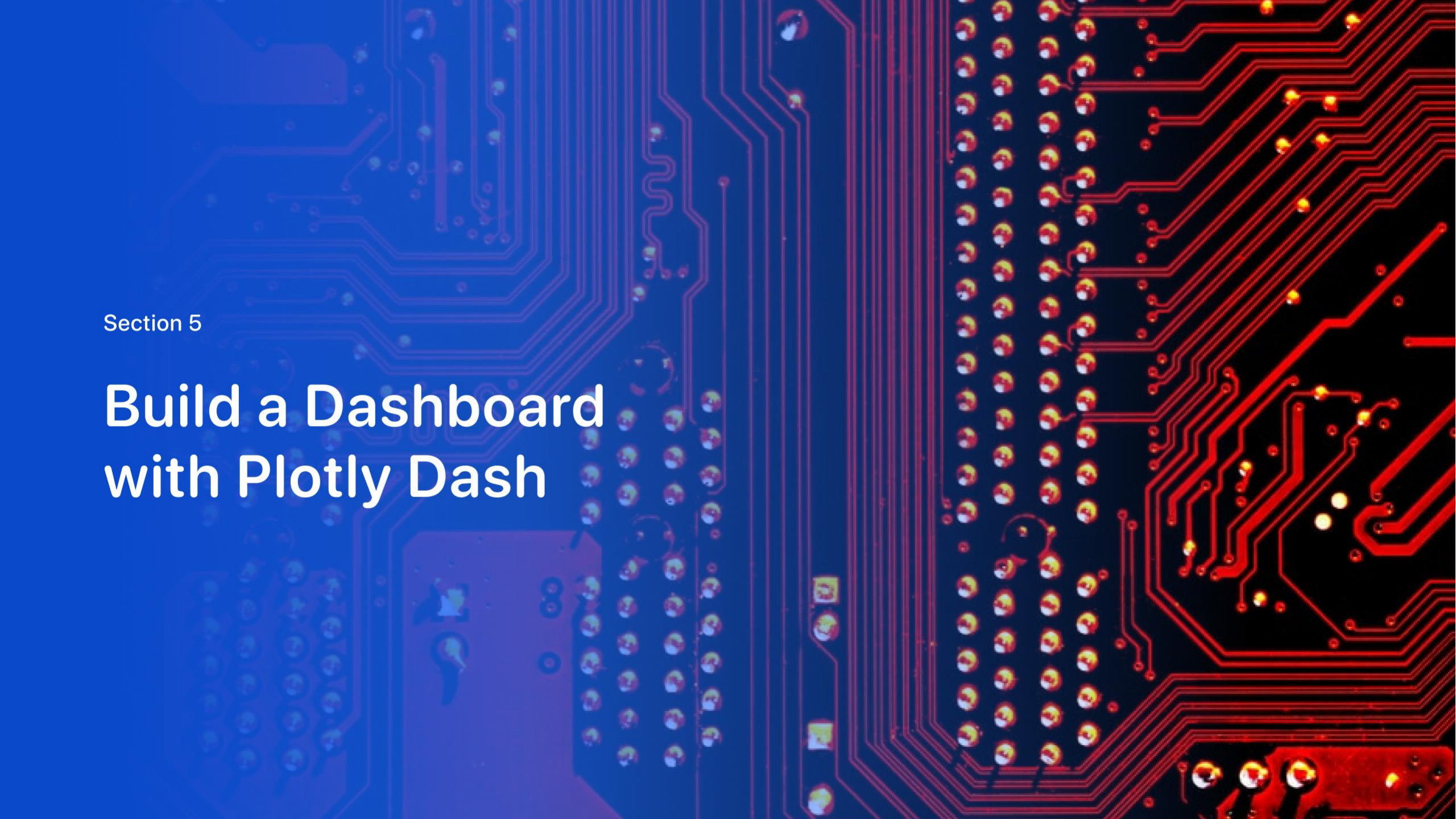
- Cluster on folium map can be clicked on to display each successful landing and failed landing.

KEY LOCATION PROXIMITIES



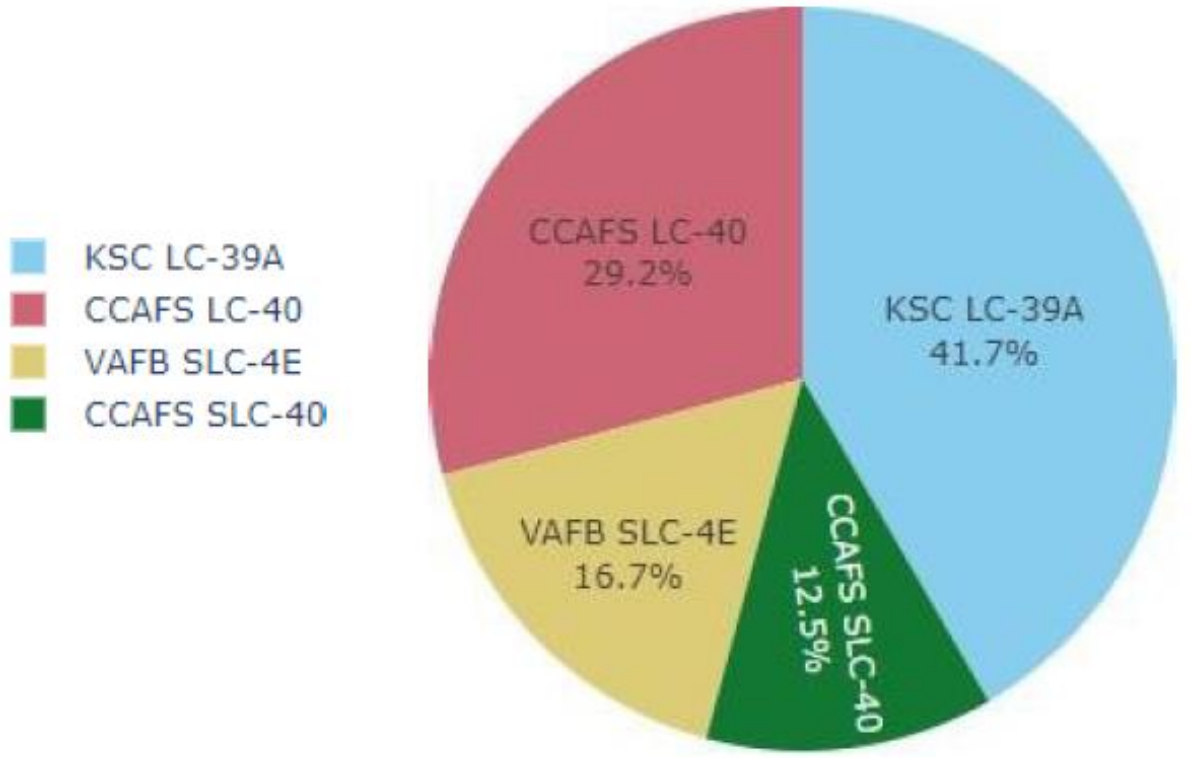
Section 5

Build a Dashboard with Plotly Dash



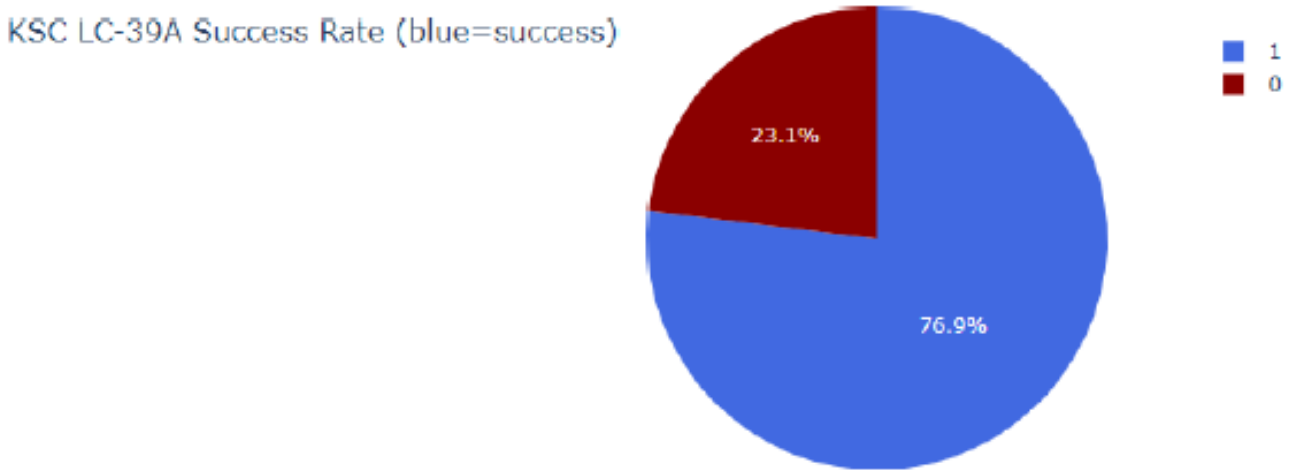
SUCCESSFUL LANDING ACROSS DIFFERENT LAUNCH SITES

- The launch sites with most successful landings ID KSCLC-39A , followed by CCAFS LC-40



HIGHEST SUCCESS RATE LAUNCH SITE

- KSC LC-39A has the highest success rate



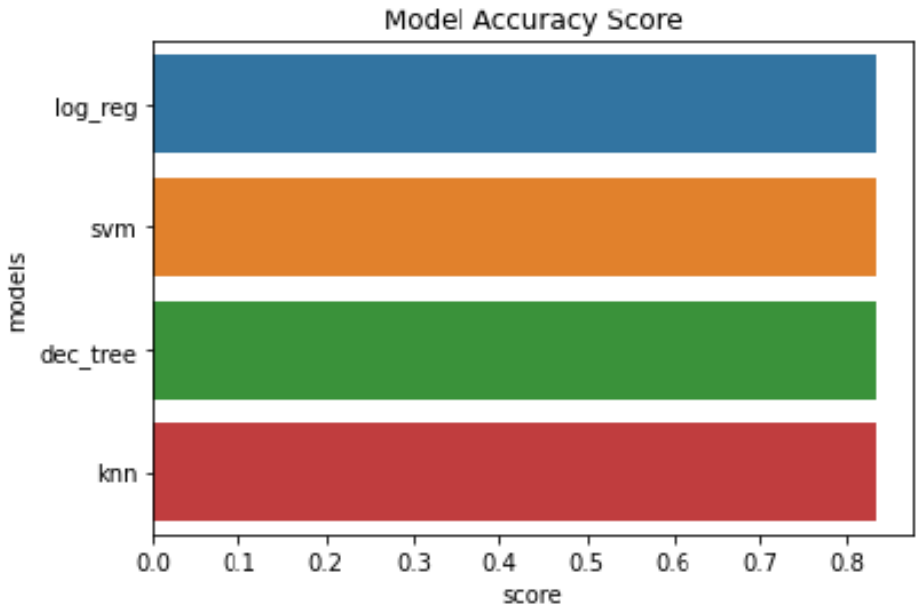
PAYLOAD MASS VS. SUCCESS VS. BOOSTER VERSION CATEGORY

Section 6

Predictive Analysis (Classification)

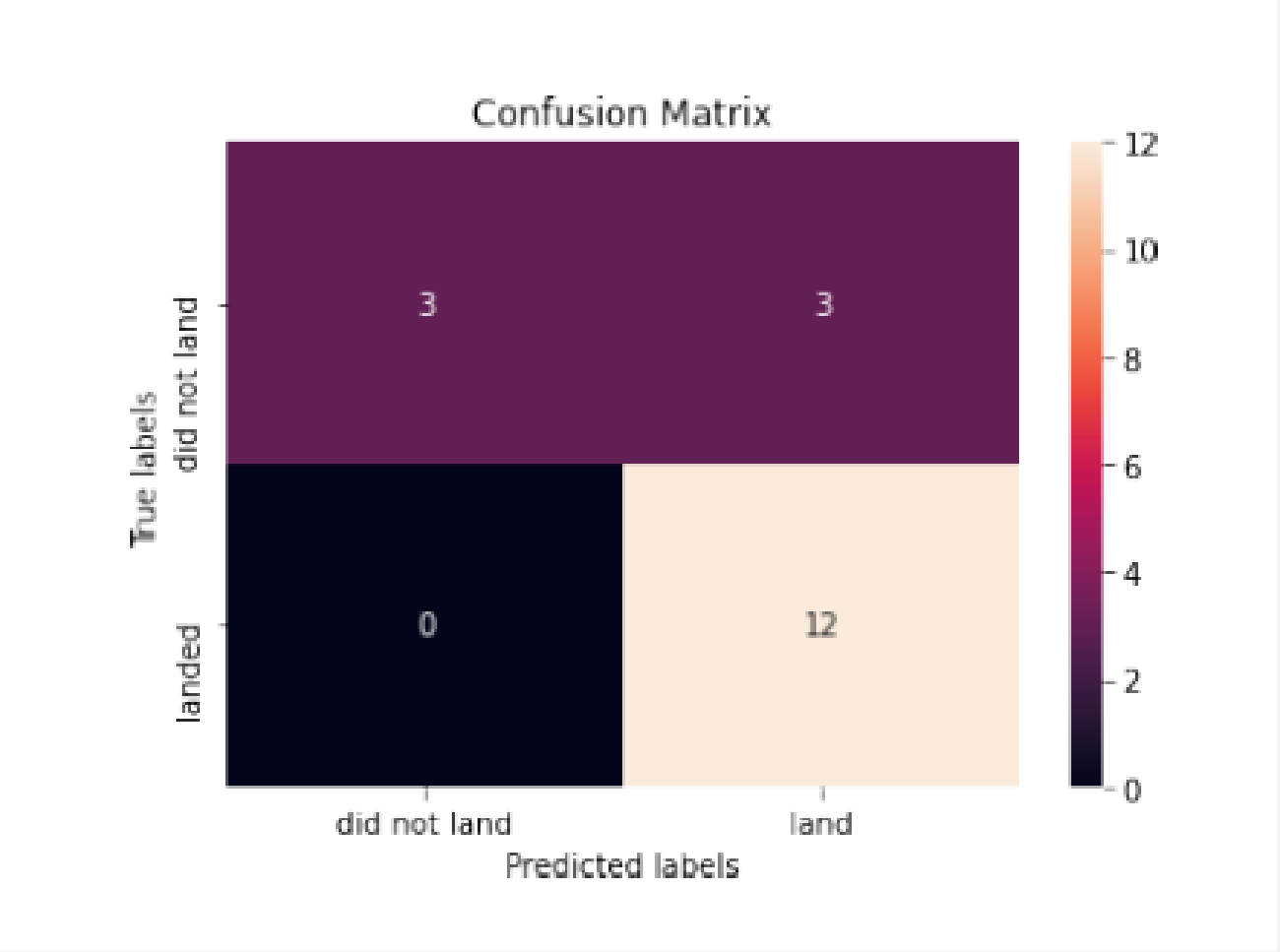
Classification Accuracy

- All models show to have approximately the same accuracy score of 83.33%circa.



Confusion Matrix

- Correct predictions are on a diagnol from top left to bottom right.



Conclusions

1) Our task: to develop a ML model for space Y who wants bid against spaceX

2) The goal of model is to predict when stage 1 will successfully land to save ~ $100 million USD

3) Data source-> public spaceX api and Wikipedia page

4) Stored data into the IBM DB2 databse

5) Done visualization and interactive dashboard

6)Trained a ML model with an accuracy of 83%

7)Space Y can use the Model to predict the high accuracy if stage 1 will land successfully or not and possibly save money

Appendix

- Git url = https://github.com/aryanmehna/DS-capstone-project.git

Thank you!