IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Young Kee Chae

Nov. 15, 2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection through API or 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 result
  - Interactive analytics in screenshots
  - Predictive Analytics result from Machine Learning Lab

Introduction

- In this capstone, we predicted if the Falcon 9 first stage would land successfully. SpaceX advertised 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 could 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. In this module, you will be provided with an overview of the problem and the tools you need to complete the course.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using API or Web Scraping
- Perform data wrangling
  - Convert outcomes either 1 or 0
- 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

- Data collection methods
  - For API, json response was converted to pandas dataframe
  - For Web scraping, BeautifulSoup was used to collect tables

Data Collection – SpaceX API

Getting request for rocket launch data using API → Using json_normalize met hod to convert json result to dataframe → Performed data cleaning and filling the missing va lue

- https://github.com/toolgen/final-assignments/blob/main/lab_jupyter_launch_site_location.ipynb

Data Collection - Scraping

Request the Falcon9 Launch Wiki page from url → Create a BeautifulSoup from the HTML response → Extract all column/variable names from the HTML header

- https://github.com/toolgen/final-assignments/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

- Data were cleaned and prepared for the next stage
- Success or Fail was changed either to 1 or 1

EDA with Data Visualization

- Scatter plots, bar charts, and line graphs were used.
  - Scatter plots: show dependency of attributes on each other.
  - Bar charts: interpret the relationship between the attributes.
  - Line graphs: show a trends or pattern of the attribute

- https://github.com/toolgen/final-assignments/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- %sql select DISTINCT LAUNCH_SITE from SPACEXTBL;
- %sql SELECT LAUNCH_SITE from SPACEXTBL where (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5;
- %sql select sum(PAYLOAD_MASS__KG_) as payloadmass from SPACEXTBL;
- %sql select avg(PAYLOAD_MASS__KG_) as avgpayloadmass from SPACEXTBL;
- %sql select min(DATE) from SPACEXTBL;
- %sql select BOOSTER_VERSION from SPACEXTBL where "Landing_Outcome"='Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000;
- %sql select count(MISSION_OUTCOME) as missionoutcomes from SPACEXTBL GROUP BY MISSION_OUTCOME;
- %sql select BOOSTER_VERSION as boosterversion from SPACEXTBL where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTBL);
- %sql SELECT substr(Date, 4, 2),"Landing_Outcome",BOOSTER_VERSION,LAUNCH_SITE FROM SPACEXTBL where substr(Date,7,4)='2015';
- %sql SELECT LANDING__OUTCOME FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' ORDER BY DATE DESC;

Build an Interactive Map with Folium

- If a launch was successful (class=1), then we use a green marker and if a launch was failed, we use a red marker (class=0)
- https://github.com/toolgen/final-assignments/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- The dashboard application contained input components such as a dropdown list and a range slider to interact with a pie chart and a scatter point chart.
- We could visually observe how payload might be correlated with mission outcomes for selected site(s). In addition, we wanted to color-label the Booster version on each scatter point so that we may observe mission outcomes with different boosters.
- https://github.com/toolgen/final-assignments/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

Perform exploratory Data Analysis and determine Training Labels → Find best Hyperparameter for SVM, Classification Trees and Logistic Regression

- EDA
  - Create a column for the class
  - Standardize the data
  - Split into training data and test data
- Prediction
  - Find the method performs best using test data
- https://github.com/toolgen/final-assignments/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

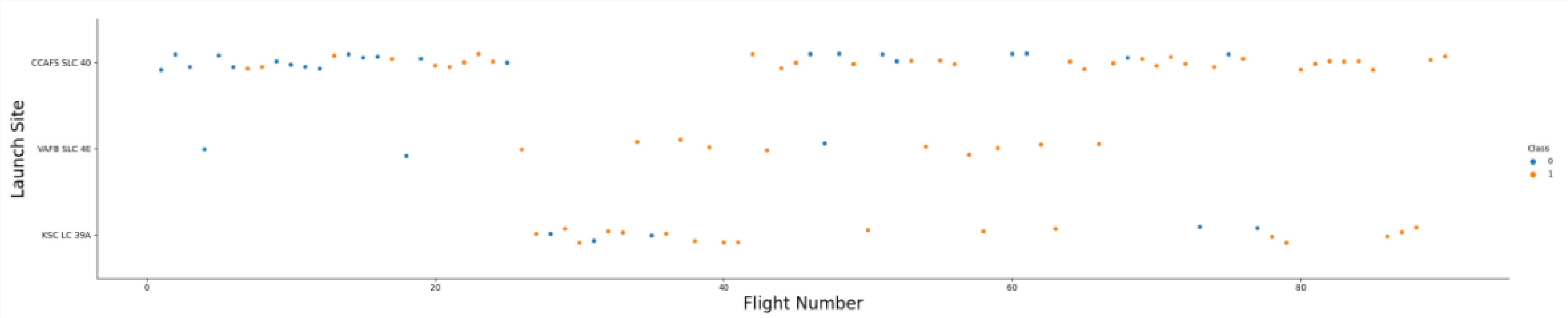
Results

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

Section 2

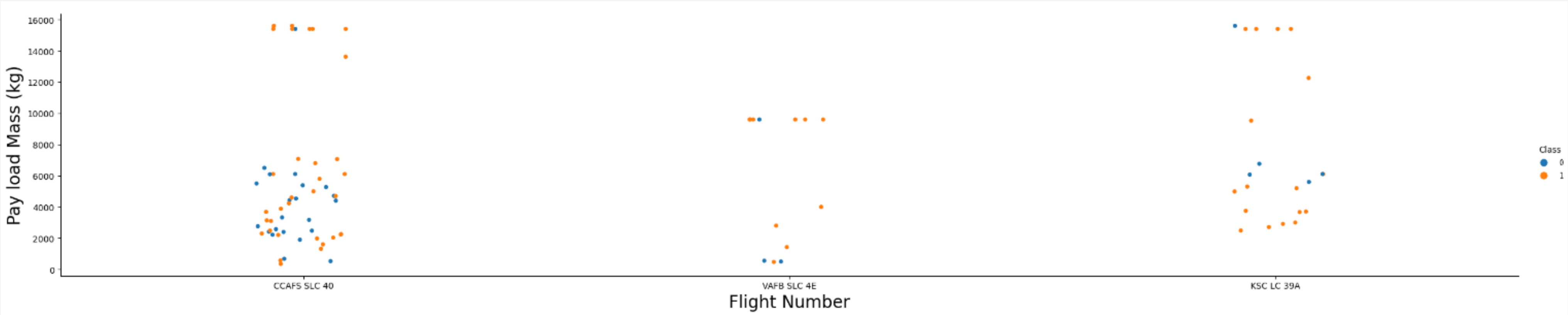
Insights drawn from EDA

Flight Number vs. Launch Site



- SLC 40 has the largest flight numbers

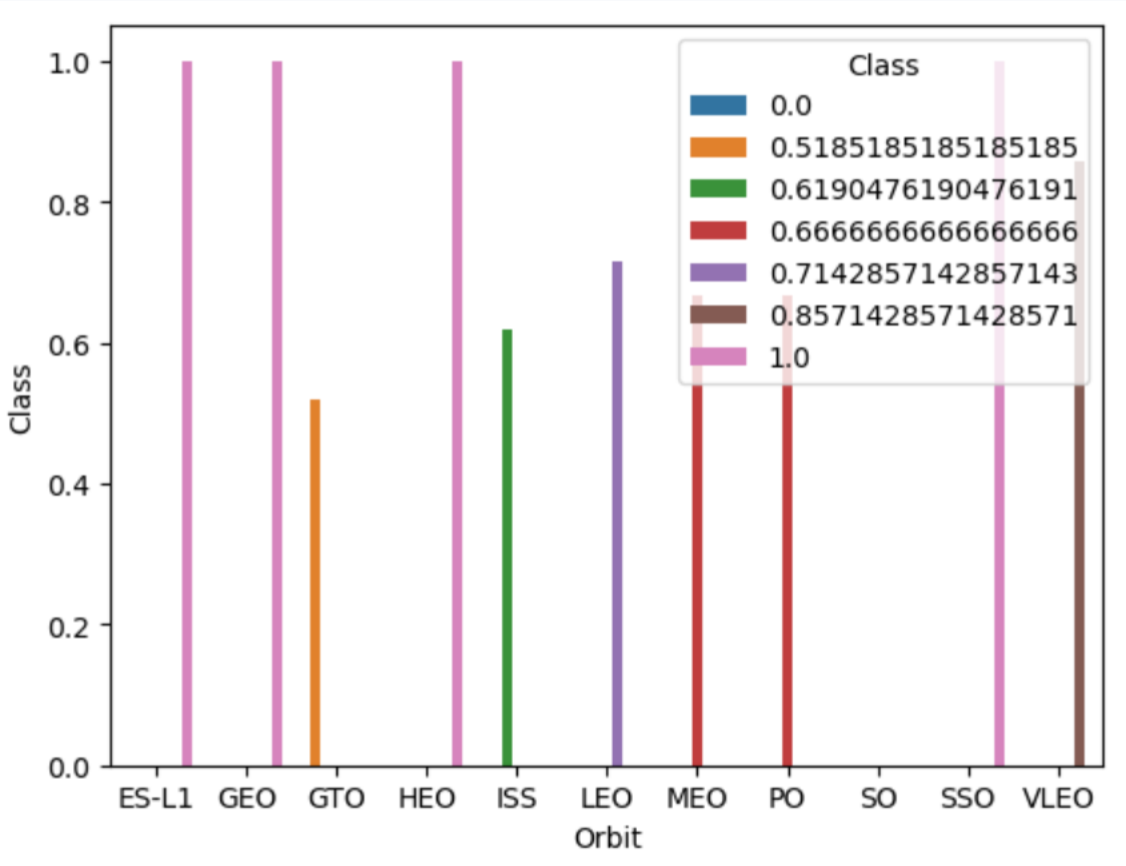
Payload vs. Launch Site



- We see that different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.

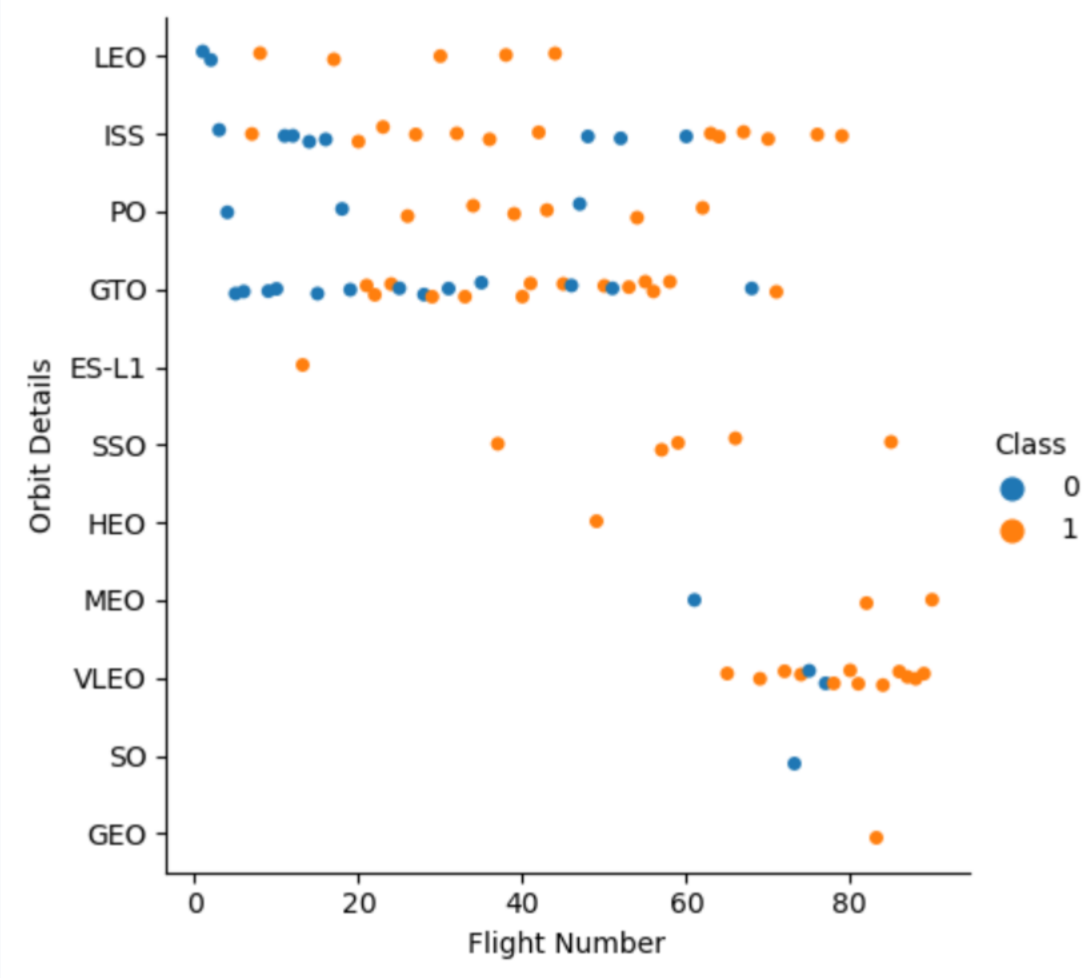
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, SSO have the highest success rate
- SO has zero success rate



Flight Number vs. Orbit Type

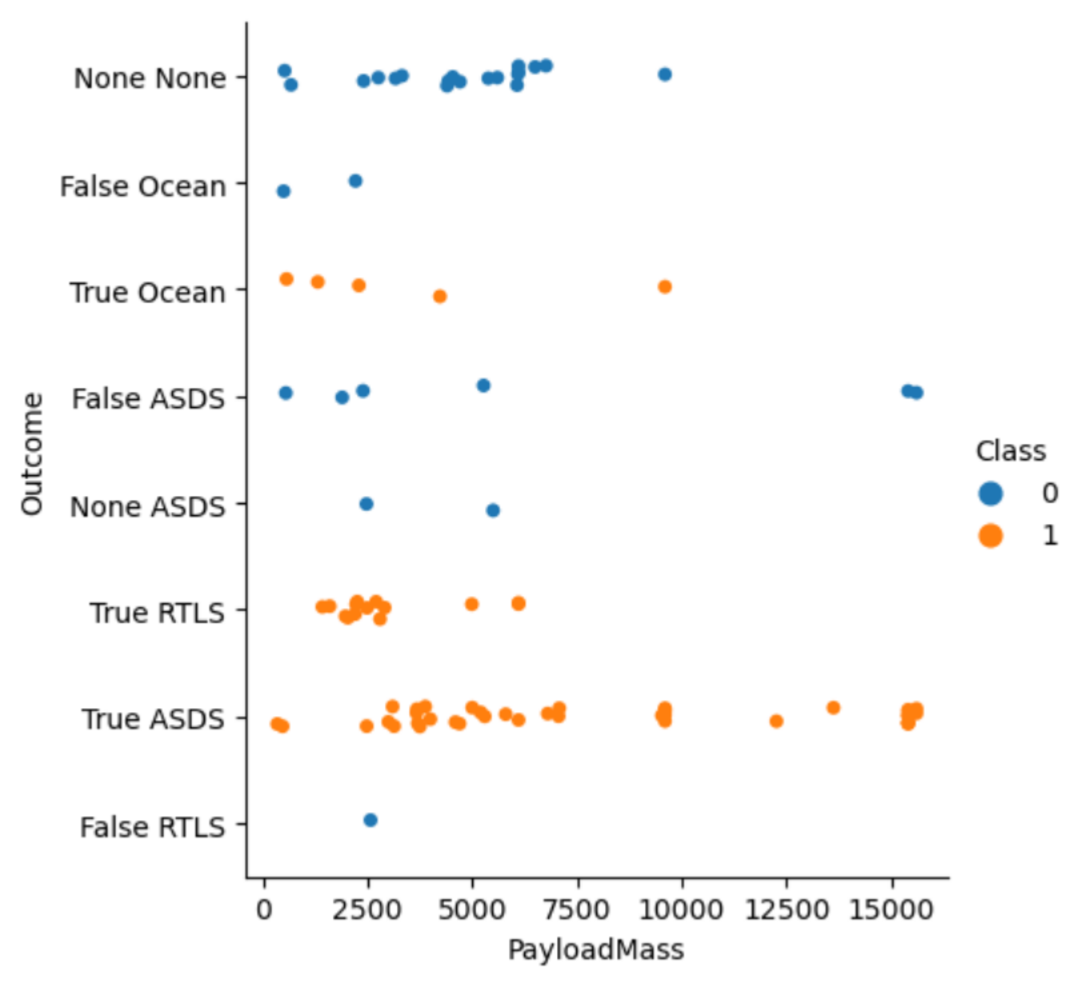
- In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



Payload vs. Orbit Type

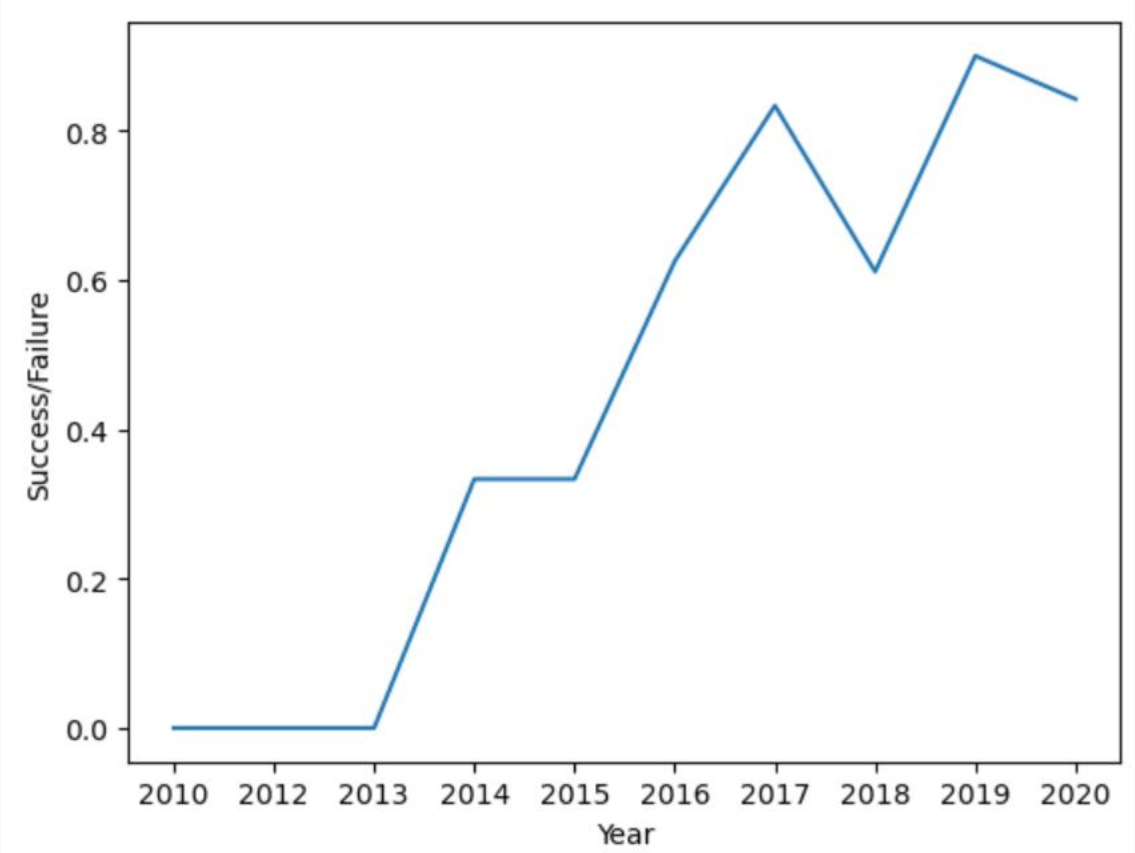
- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.

- 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

- The names of the unique launch sites are displayed in the space mission with DISTINCT option

```
%sql select DISTINCT LAUNCH_SITE from SPACEXTBL;
```

```
 * sqlite:///my_data1.db
Done.
```

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

Launch Site Names Begin with 'CCA'

- There are 5 records where launch sites begin with `CCA`

```
%sql SELECT LAUNCH_SITE from SPACEXTBL where (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5;
```

```
 * sqlite:///my_data1.db
Done.
```

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

```
%sql select sum(PAYLOAD_MASS__KG_) as payloadmass from SPACEXTBL;
```

```
 * sqlite:///my_data1.db
Done.
```

payloadmass
619967

- The total payload carried by boosters from NASA was 619967.

Average Payload Mass by F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_) as avgpayloadmass from SPACEXTBL;
```

```
 * sqlite:///my_data1.db
Done.
```

avgpayloadmass
6138.287128712871

- The average payload mass carried by booster version F9 v1.1 was 6138.

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad was 01-03-2013.

```
%sql select min(DATE) from SPACEXTBL;
```

```
 * sqlite:///my_data1.db
Done.
```

min(DATE)
01-03-2013

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select BOOSTER_VERSION from SPACEXTBL where "Landing_Outcome"='Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000;
```

```
 * sqlite:///my_data1.db
Done.
```

Booster_Version

- There was no booster which has successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

```
%sql select count(MISSION_OUTCOME) as missionoutcomes from SPACEXTBL GROUP BY MISSION_OUTCOME;
```

```
 * sqlite:///my_data1.db
Done.
```

missionoutcomes
1
98
1
1

- The total number of successful and failure mission outcomes was 101.

Boosters Carried Maximum Payload

```
%sql select BOOSTER_VERSION as boosterversion from SPACEXTBL where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

```
 * sqlite:///my_data1.db
Done.
```

boosterversion
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

- MAX() generated maximum payload.

2015 Launch Records

```
%sql SELECT substr(Date, 4, 2),"Landing_Outcome",BOOSTER_VERSION,LAUNCH_SITE FROM SPACEXTBL where substr(Date,7,4)='2015';
```

```
 * sqlite:///my_data1.db
Done.
```

substr(Date, 4, 2)	"Landing_Outcome"	Booster_Version	Launch_Site
01	Landing_Outcome	F9 v1.1 B1012	CCAFS LC-40
02	Landing_Outcome	F9 v1.1 B1013	CCAFS LC-40
03	Landing_Outcome	F9 v1.1 B1014	CCAFS LC-40
04	Landing_Outcome	F9 v1.1 B1015	CCAFS LC-40
04	Landing_Outcome	F9 v1.1 B1016	CCAFS LC-40
06	Landing_Outcome	F9 v1.1 B1018	CCAFS LC-40
12	Landing_Outcome	F9 FT B1019	CCAFS LC-40

- There were 7 records of the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015.

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

```
%sql SELECT LANDING__OUTCOME FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' ORDER BY DATE DESC;
```

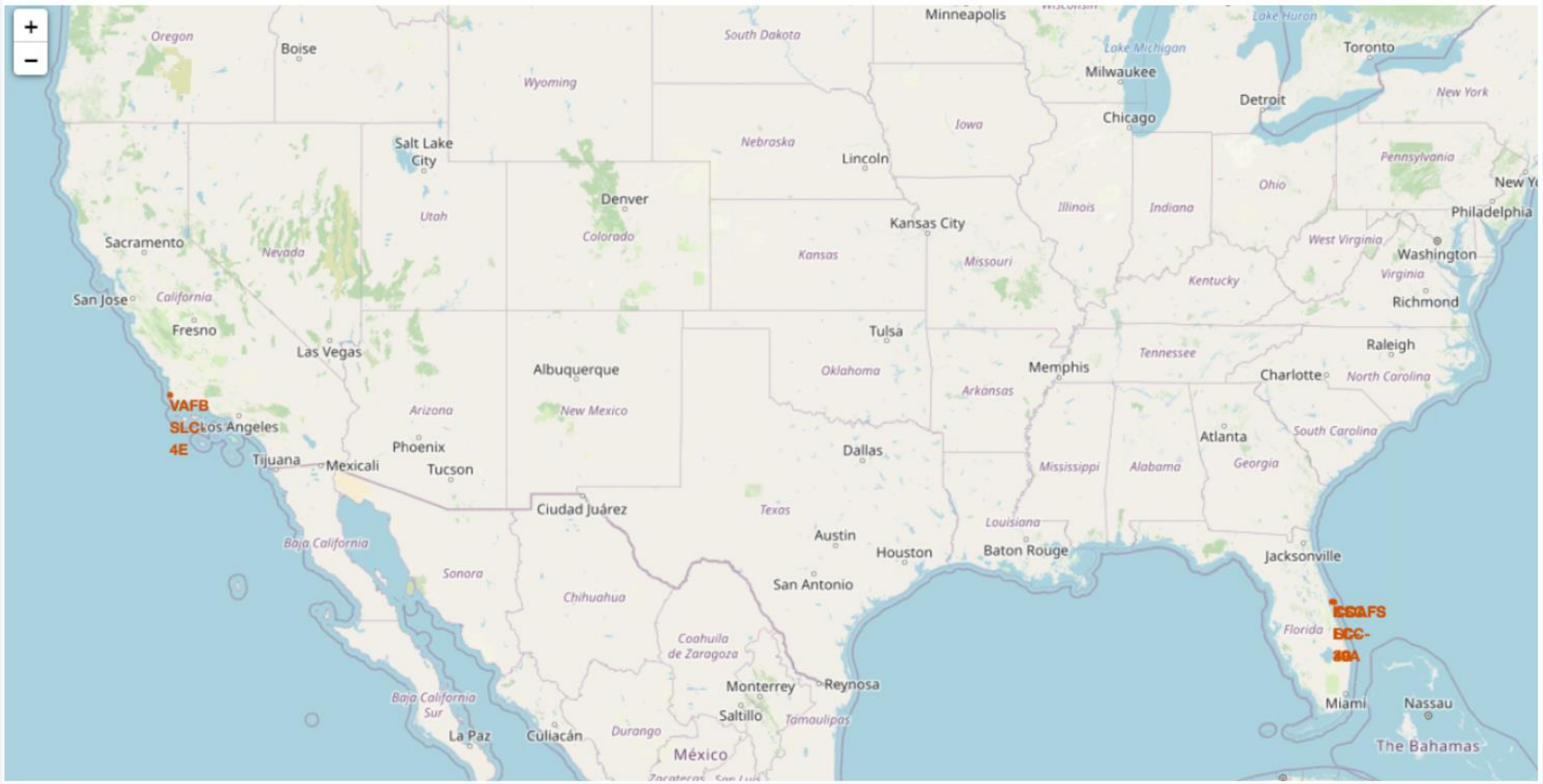
```
 * sqlite:///my_data1.db
(sqlite3.OperationalError) no such column: LANDING__OUTCOME
[SQL: SELECT LANDING__OUTCOME FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' ORDER BY DATE DESC;]
(Background on this error at: http://sqlalche.me/e/e3q8)
```

- 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 was ranked.

Section 3

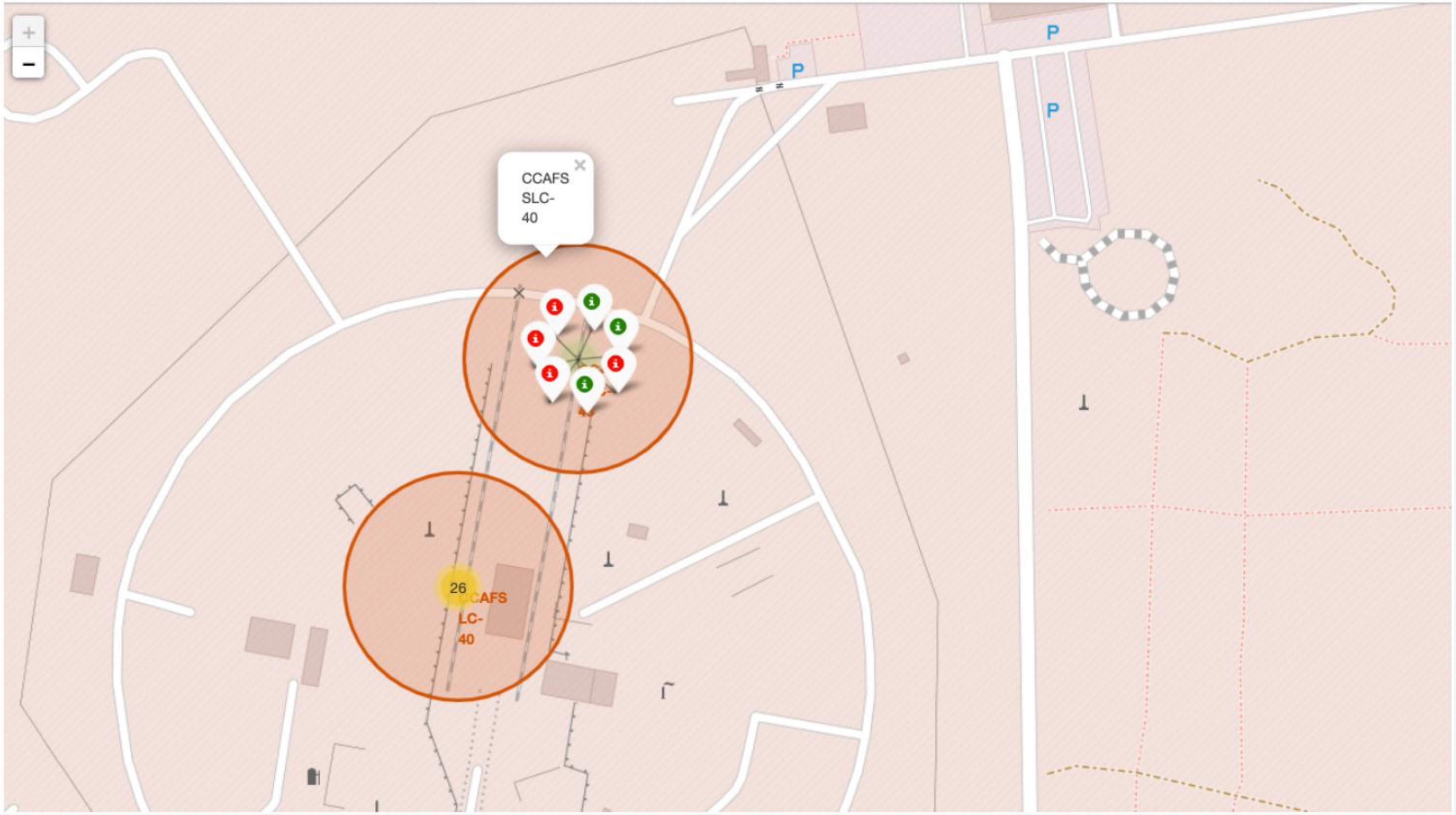
Launch Sites Proximities Analysis

Location of all the Launch Sites



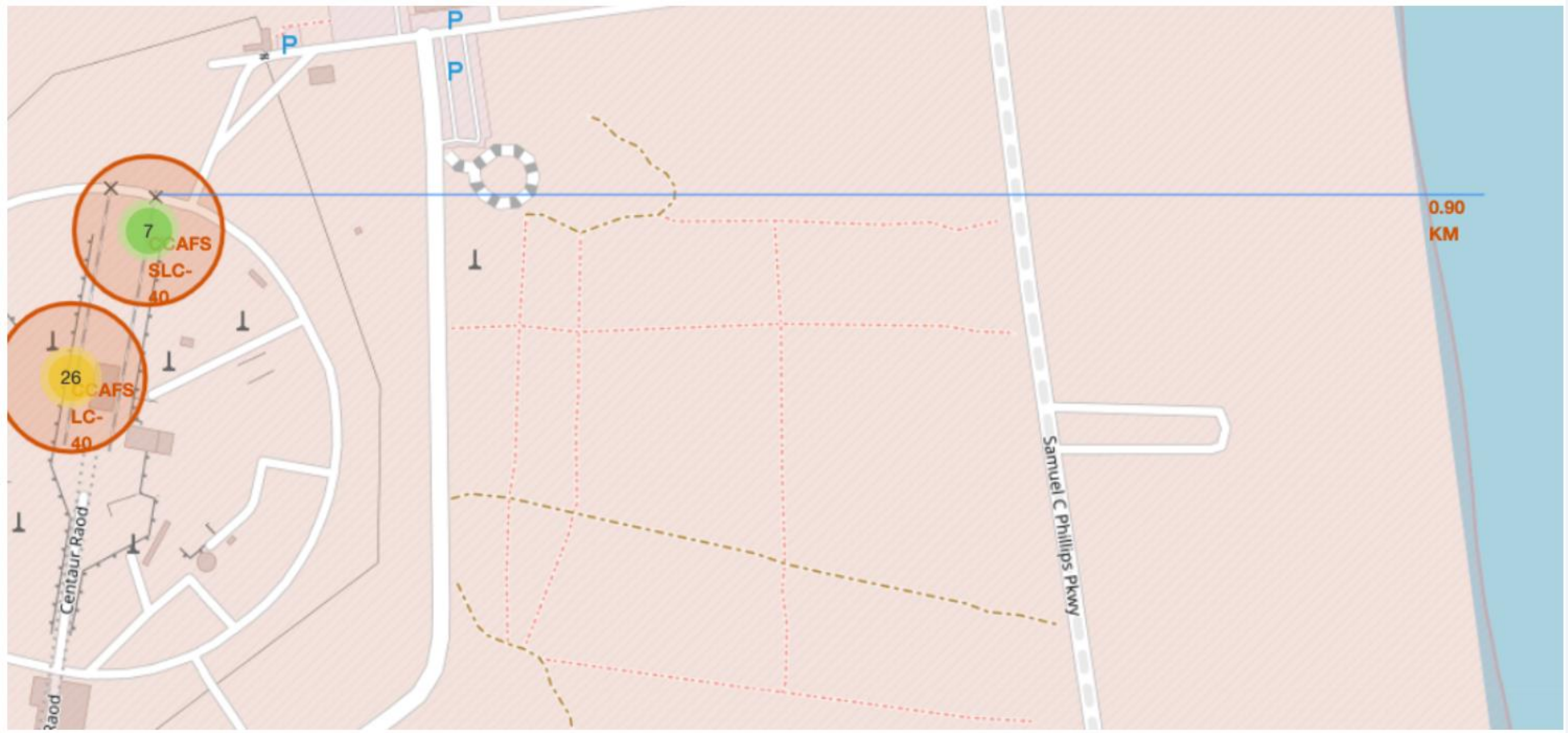
- All launch sites' location markers on a global map were shown

Color-labeled launch outcomes



- The color-labeled launch outcomes on the map are shown.

Launch site to its proximities

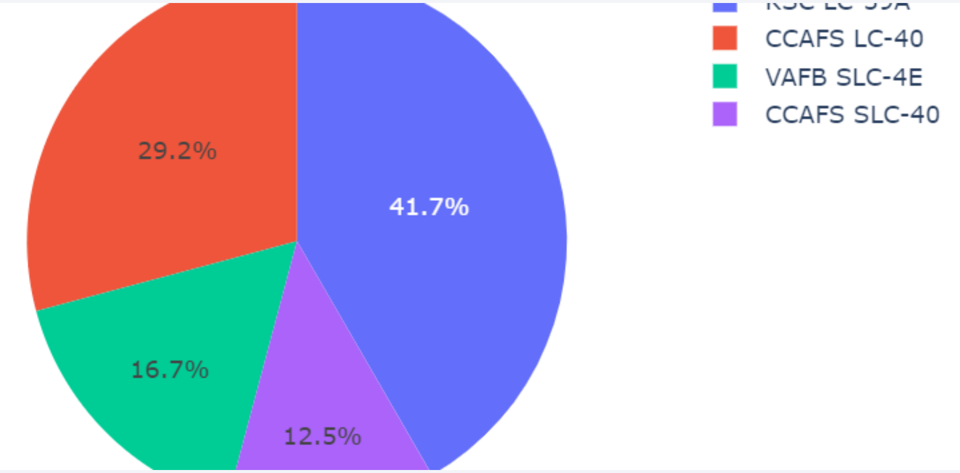


- Launch site to its proximities such as railway, highway, coastline, with distance is calculated and displayed

Section 4

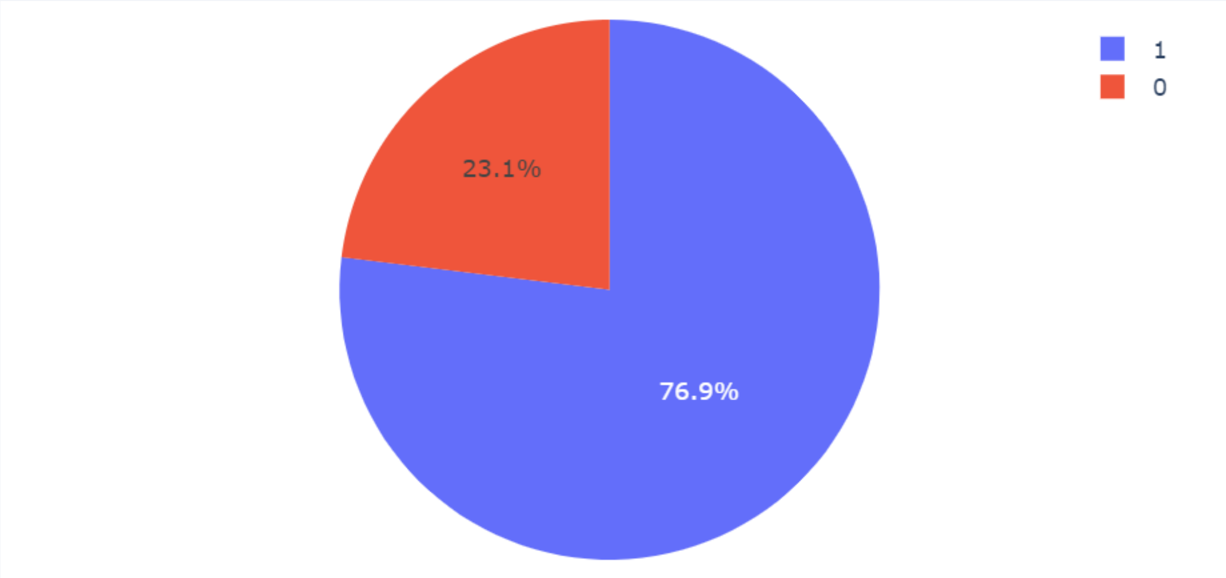
Build a Dashboard with Plotly Dash

Launch success count for all sites



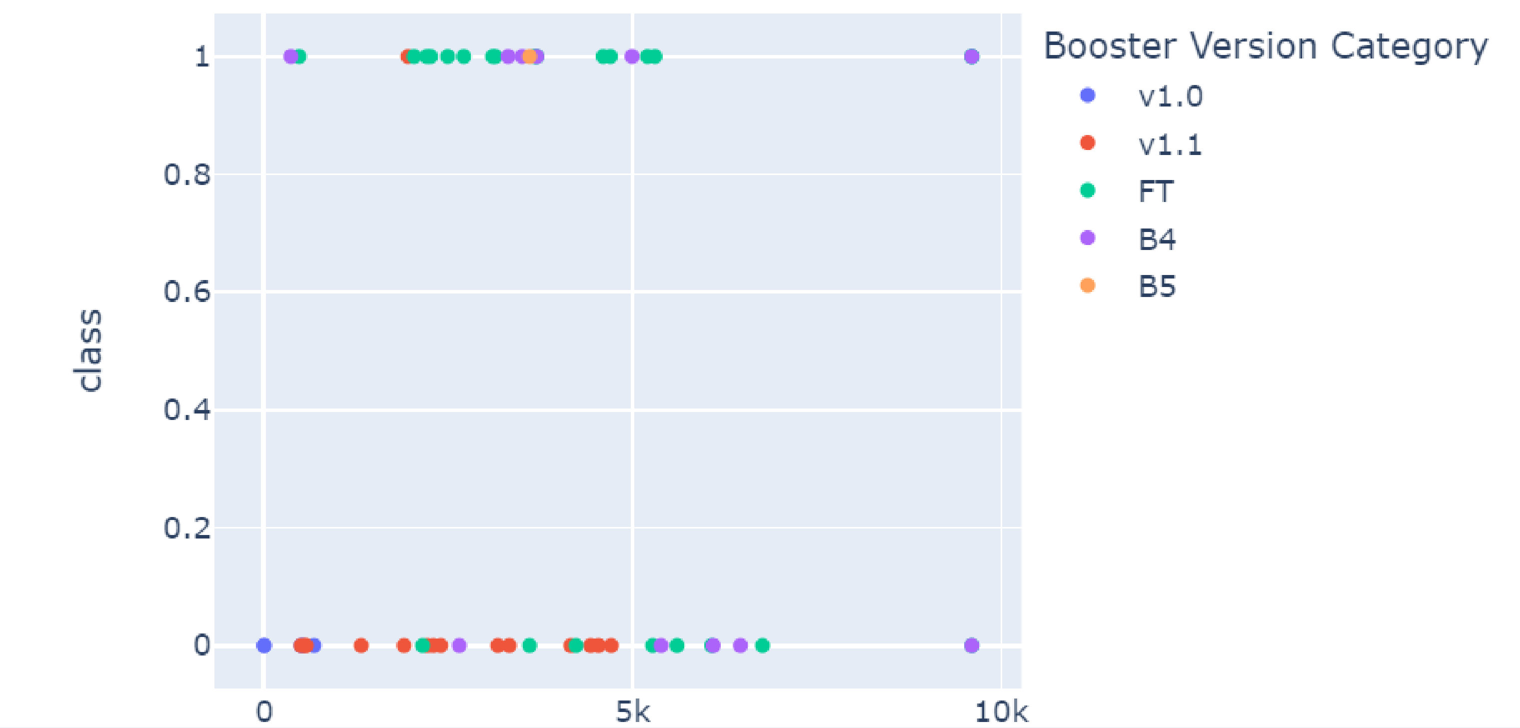
- KSC LC 39A has the highest success rate.

KSC LC 39A site



- The highest success rate was 76.9%.

Payload vs. Launch Outcome



- Success rate is higher with payloads less than 5000 kG.

Section 5

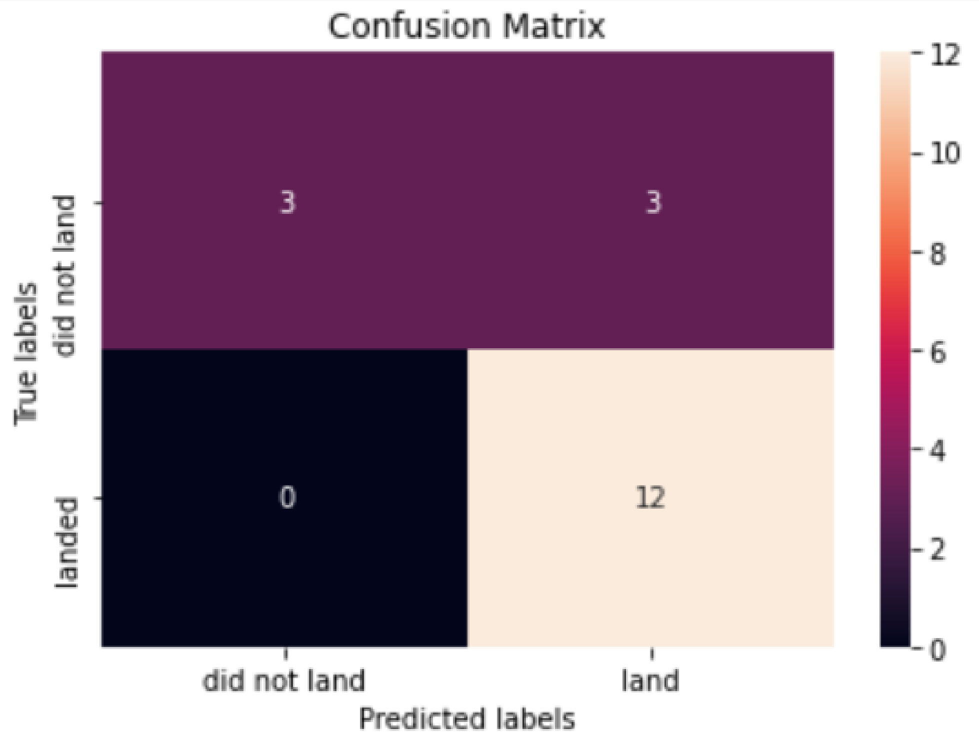
Predictive Analysis (Classification)

Classification Accuracy

- The decision tree has the highest accuracy.

Confusion Matrix

- The major problem is false positives.



Conclusions

- Collected data should be wrangled before analysis
- Appropriate data could be extracted by SQL
- Folium is a great tool for interactive visualization
- Dashboard could be used to update the result in real time
- Several predictive methods could be tested to get the best result

Appendix

- All notebooks were uploaded to GitHub

Thank you!