Data Science Capstone project

Elnur Shahbalayev

8/29/2021

Outline

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

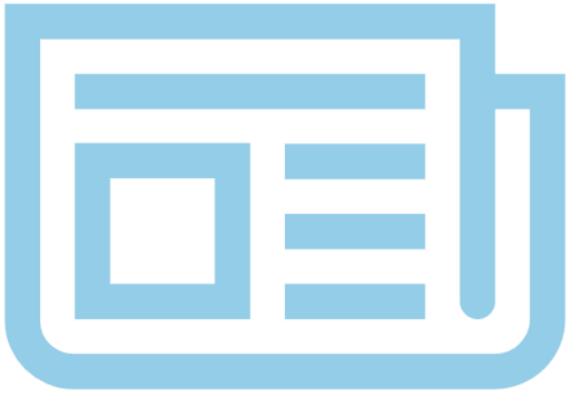
Executive Summary

- In this given report, data is used from SpaceX to construct new predictions for its next flights.

Introduction

- In the website of SpaceX there are advertisements of Falcon 9 rocket with cost of 62 million dollars while competitors in the industry does the same with cost of 165 million dollars, and this big difference if all because of the reusability of first stage in SpaceX rockets. Therefore if there is a way to determine the first stage landing, the cost of launch can be determined too.

Methodology



- Data collection methodology:
  - Describe how data were collected
- Perform data wrangling
  - Describe how data were processed
- 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

Methodology

Data collection

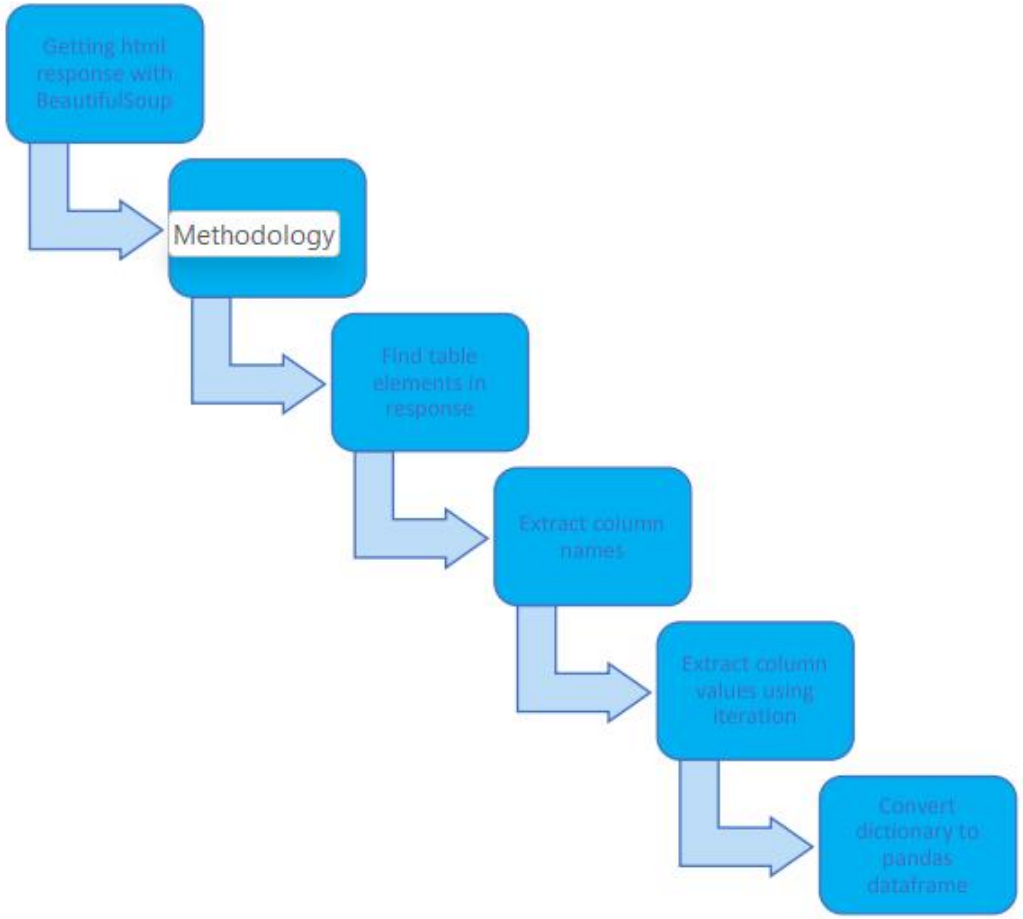
- Data was collected using two methods. First one was API request using python by getting json file and normalizing it using pandas. The other one was webscraping using BeautifulSoup module in python.

Data collection – SpaceX API

https://github.com/elnursahbalayev/Data-Science-Projects/blob/90c401edeb47321ef1352358dea19b07c4d7b19f/IBM%20DS%20projects/course%2010%20-%20capstone%20project/jupyter-labs-spacex-data-collection-api.ipynb

Data collection – Web scraping

https://github.com/elnursahbalayev/Data-Science-Projects/blob/e02057213f4cdbeba3a73b65c23f66ae80c74763/IBM%20DS%20projects/course%2010%20-%20capstone%20project/jupyter-labs-webscraping.ipynb



Data wrangling

- After the data was collected it was processed. Missing values and data types, unordered columns were checked.

- https://github.com/elnursahbalayev/Data-Science-Projects/blob/e02057213f4cdbeba3a73b65c23f66ae80c74763/IBM%20DS%20projects/course%2010%20-%20capstone%20project/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with data visualization

- In this EDA, cat plot, bar chart were used and correlations found between columns.

- https://github.com/elnursahbalayev/Data-Science-Projects/blob/e02057213f4cdbeba3a73b65c23f66ae80c74763/IBM%20DS%20projects/course%2010%20-%20capstone%20project/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- This SQL queries were performed:
  - DISTINCT
  - DATE
  - SUM
  - LIMIT
  - LIKE
- https://github.com/elnursahbalayev/Data-Science-Projects/blob/e02057213f4cdbeba3a73b65c23f66ae80c74763/IBM%20DS%20projects/course%2010%20-%20capstone%20project/jupyter-labs-spacex-data-collection-api.ipynb

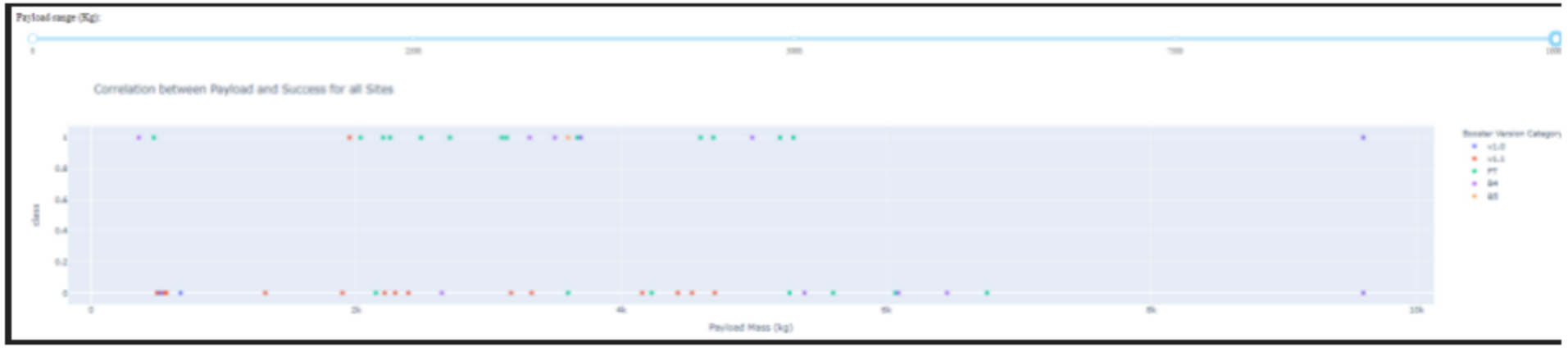
Build an interactive map with Folium

In this part marker, circle, and icon elements were used to show the points in the map.

- https://github.com/elnursahbalayev/Data-Science-Projects/blob/e02057213f4cdbeba3a73b65c23f66ae80c74763/IBM%20DS%20projects/course%2010%20-%20capstone%20project/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- In this part, I included a pie-chart, site-dropdown, and slider.

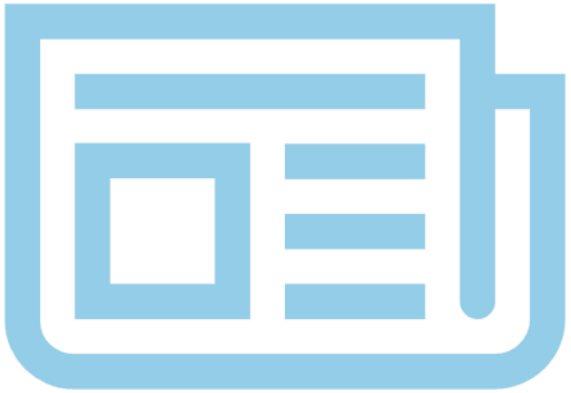


Predictive analysis (Classification)

- In this part I have tried different classification models. I got my best result using Logistic Regression with these hyperparameters.

```
tuned hpyerparameters :(best parameters)  {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}
accuracy : 0.8222222222222222
```

Results



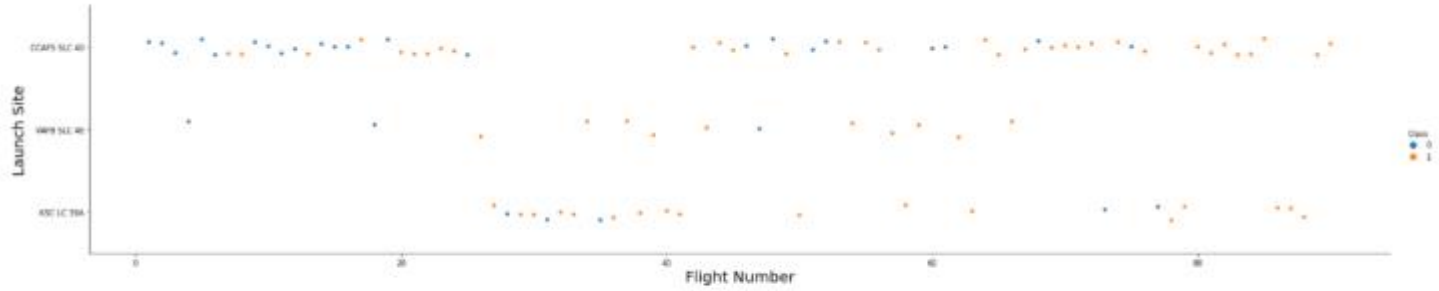
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

EDA with Visualization

Flight Number vs. Launch Site

Show a scatter plot of Flight Number vs. Launch Site

Show the screenshot of the scatter plot with explanations

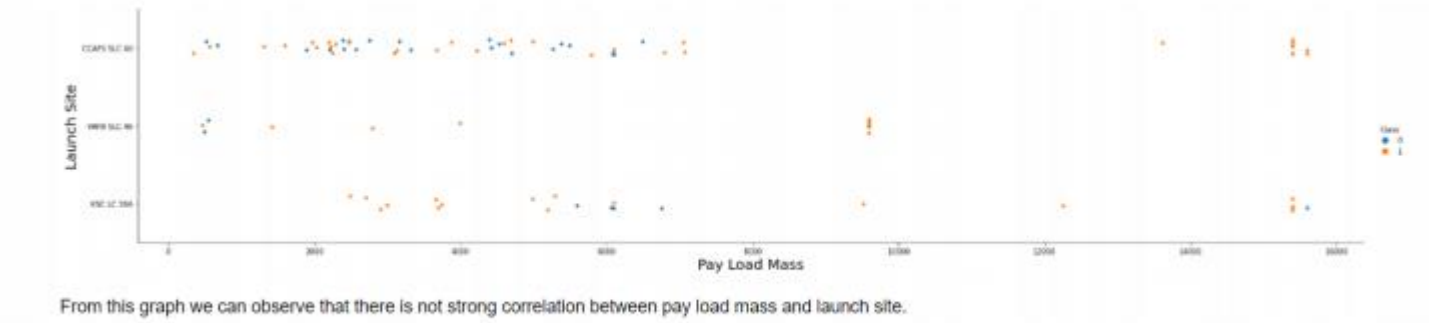


We can observe that CCAFS SLC 40 is most used launch site. But its similar success and unsuccess rate. Other sites are less used however they have high success rate.

Payload vs. Launch Site

Show a scatter plot of Payload vs. Launch Site

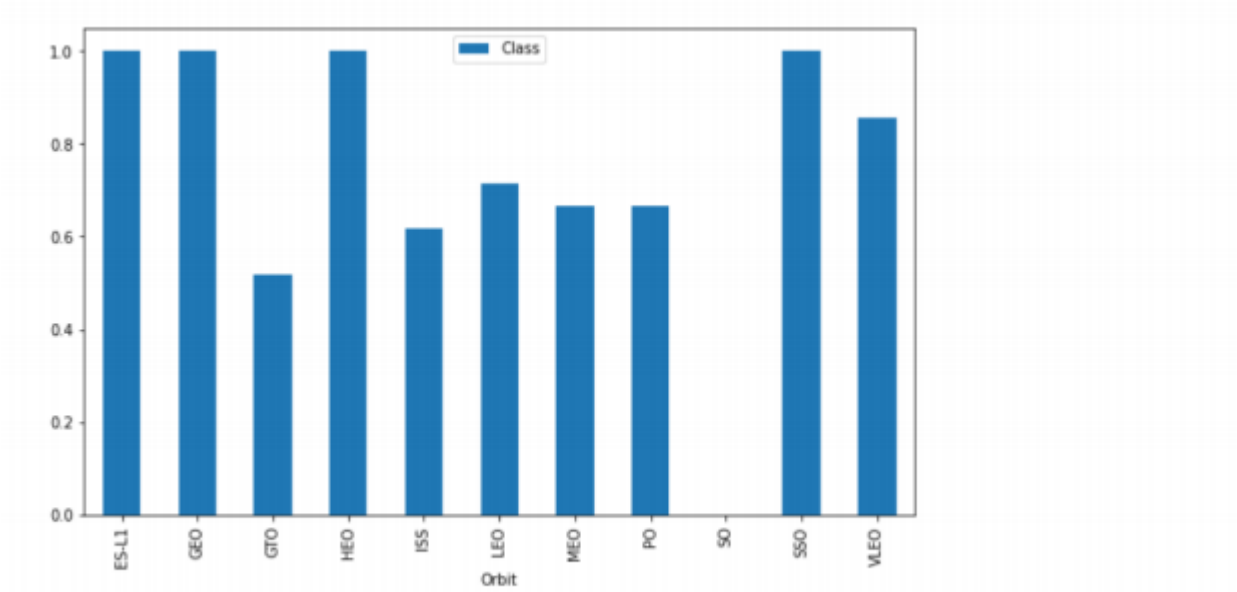
Show the screenshot of the scatter plot with explanations



Success rate vs. Orbit type

Show a barchart for the success rate of each orbit type

Show the screenshot of the scatter plot with explanations

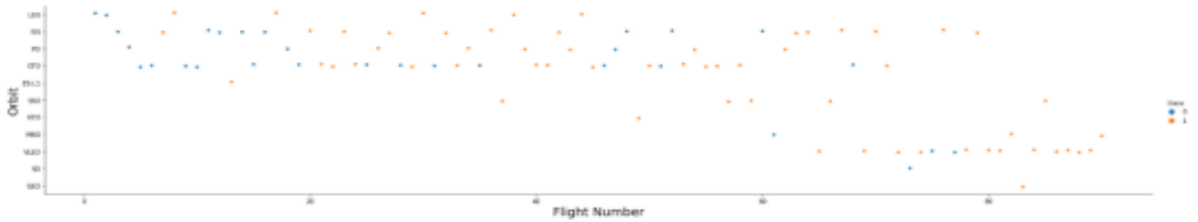


Analyze the ploted bar chart try to find which orbits have high sucess rate. ES-l1, GEO, HEO, SSO has high success rate.

Flight Number vs. Orbit type

Show a scatter point of Flight number vs. Orbit type

Show the screenshot of the scatter plot with explanations

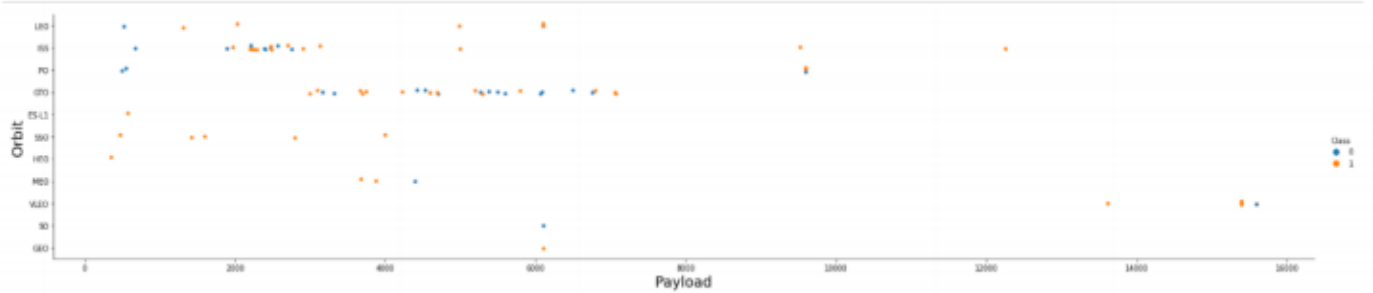


You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit type

Show a scatter point of payload vs. orbit type

Show the screenshot of the scatter plot with explanations



You should observe that Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

EDA with SQL

All launch site names

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

```
%sql  SELECT DISTINCT launch_site FROM SPACEXTABLE
```

```
 * ibm_db_sa://kql94202:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

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

Launch site names begin with `CCA`

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

```
]: %sql SELECT DATE FROM SPACEXTABLE WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
```

```
 * ibm_db_sa://kql94202:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.d
Done.
```

]:

DATE
2010-06-04
2010-12-08
2012-05-22
2012-10-08
2013-03-01

Total payload mass

Task 3

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

```
5]: %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE CUSTOMER='NASA (CRS)'
 * ibm_db_sa://kql94202:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud
Done.
```

5]:

1
45596

Average payload mass by F9 v1.1

Task 4

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

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

```
 * ibm_db_sa://kql94202:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

1
2928

First successful ground landing date

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

Hint:Use min function

```
]: %sql SELECT DATE FROM SPACEXTABLE WHERE LANDING__OUTCOME='Success (ground pad)' LIMIT 1
```

```
 * ibm_db_sa://kql94202:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

]:

DATE
2015-12-22

Successful drone ship landing with payload between 4000 and 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 FROM SPACEXTABLE WHERE LANDING__OUTCOME='Success (drone ship)' AND PAYLOAD_MASS__KG_>4000 AND PAYLOAD_MASS__K
G_<6000
```

```
 * ibm_db_sa://kql94202:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
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

List the total number of successful and failure mission outcomes

```
%sql SELECT COUNT(MISSION_OUTCOME) FROM SPACEXTABLE
```

```
 * ibm_db_sa://kql94202:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

1
101

Boosters carried maximum payload

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 SPACEXTABLE WHERE PAYLOAD_MASS__KG_=(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

```
 * ibm_db_sa://kql94202:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/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 the in year 2015

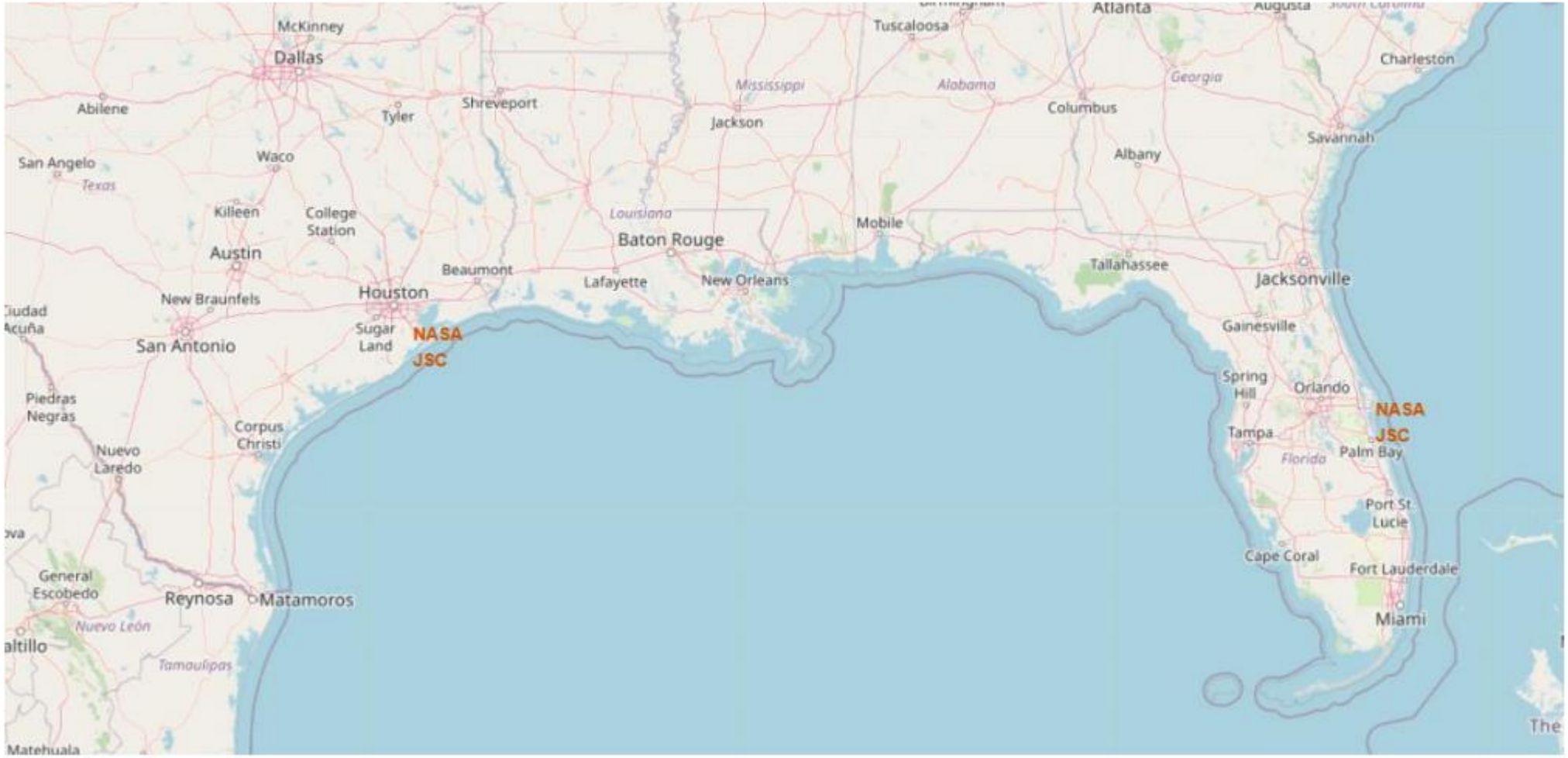
```
%sql SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTABLE WHERE LANDING__OUTCOME='Failure (drone ship)' AND YEAR(DATE)=
2015
```

```
 * ibm_db_sa://kql94202:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

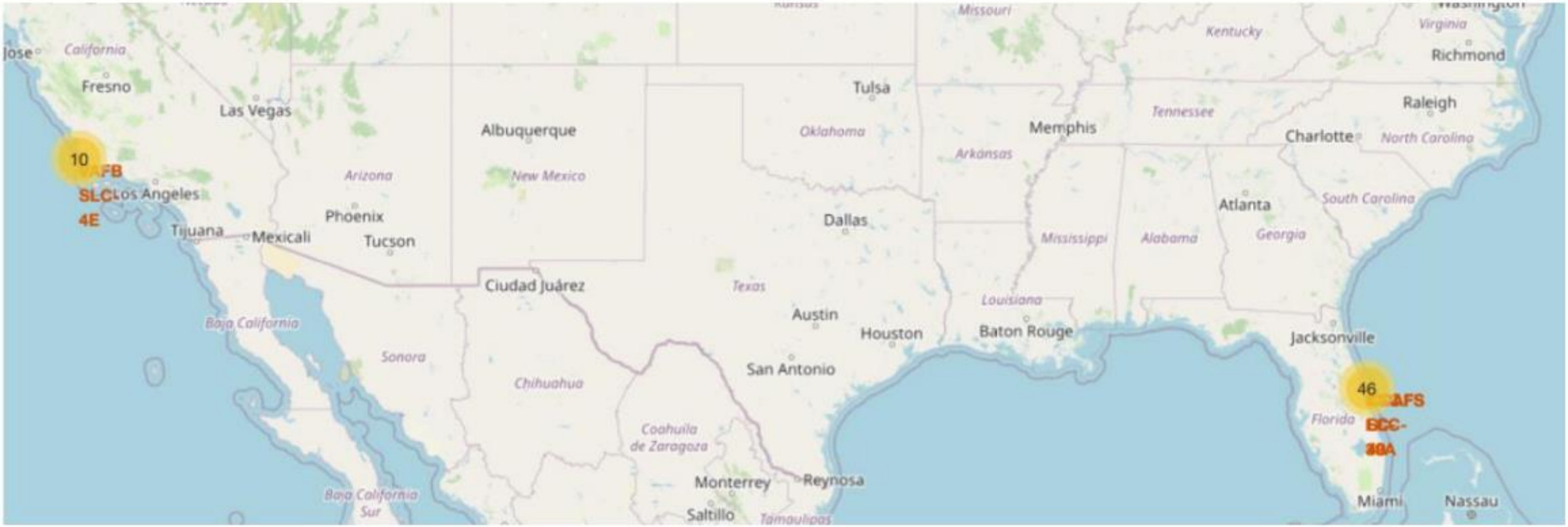
landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Interactive map with Folium

<Folium map screenshot 1>



<Folium map screenshot 2>

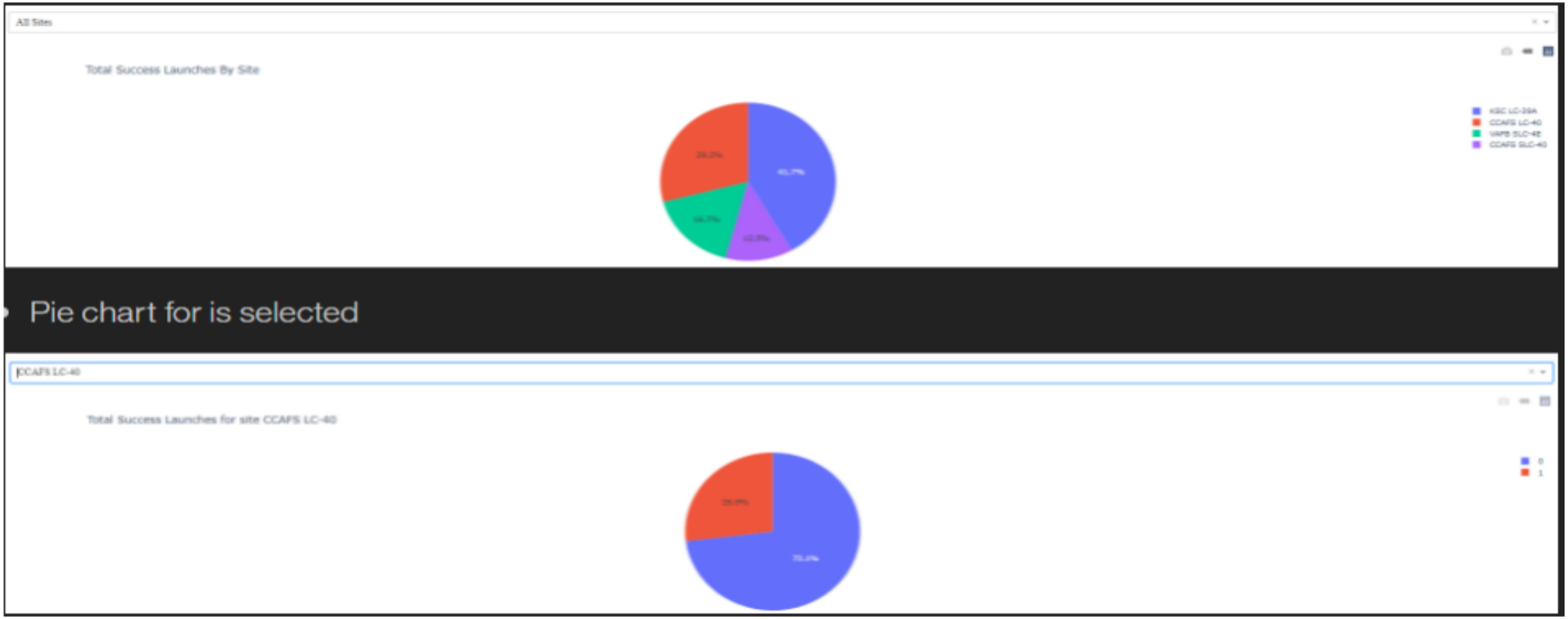


<Folium map screenshot 3>

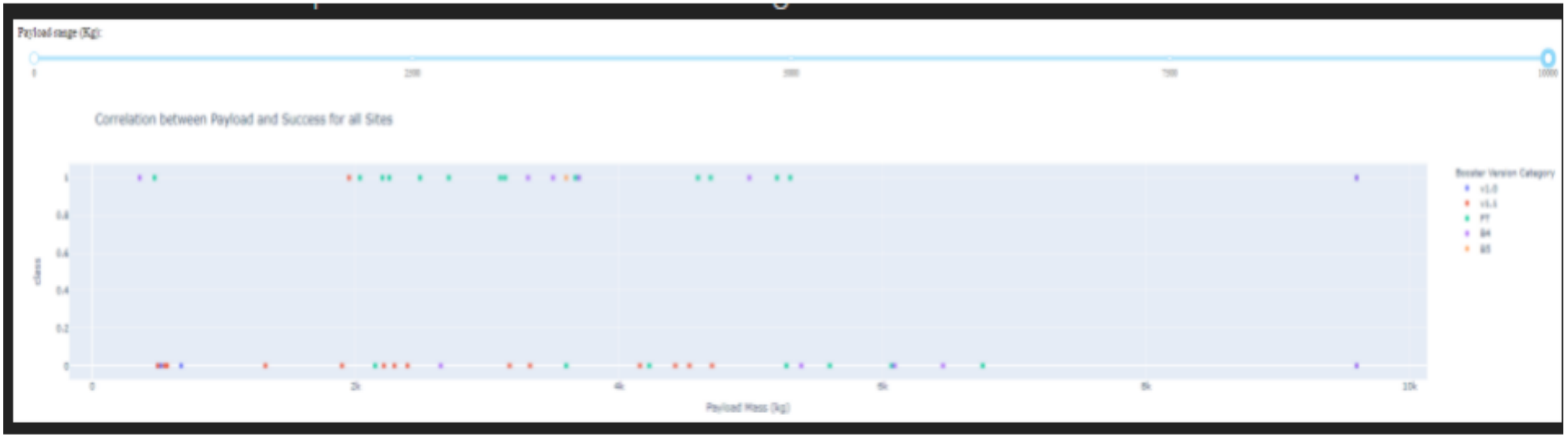
Build a Dashboard with Plotly Dash

<Dashboard screenshot 1>

<Dashboard screenshot 2>



<Dashboard screenshot 3>

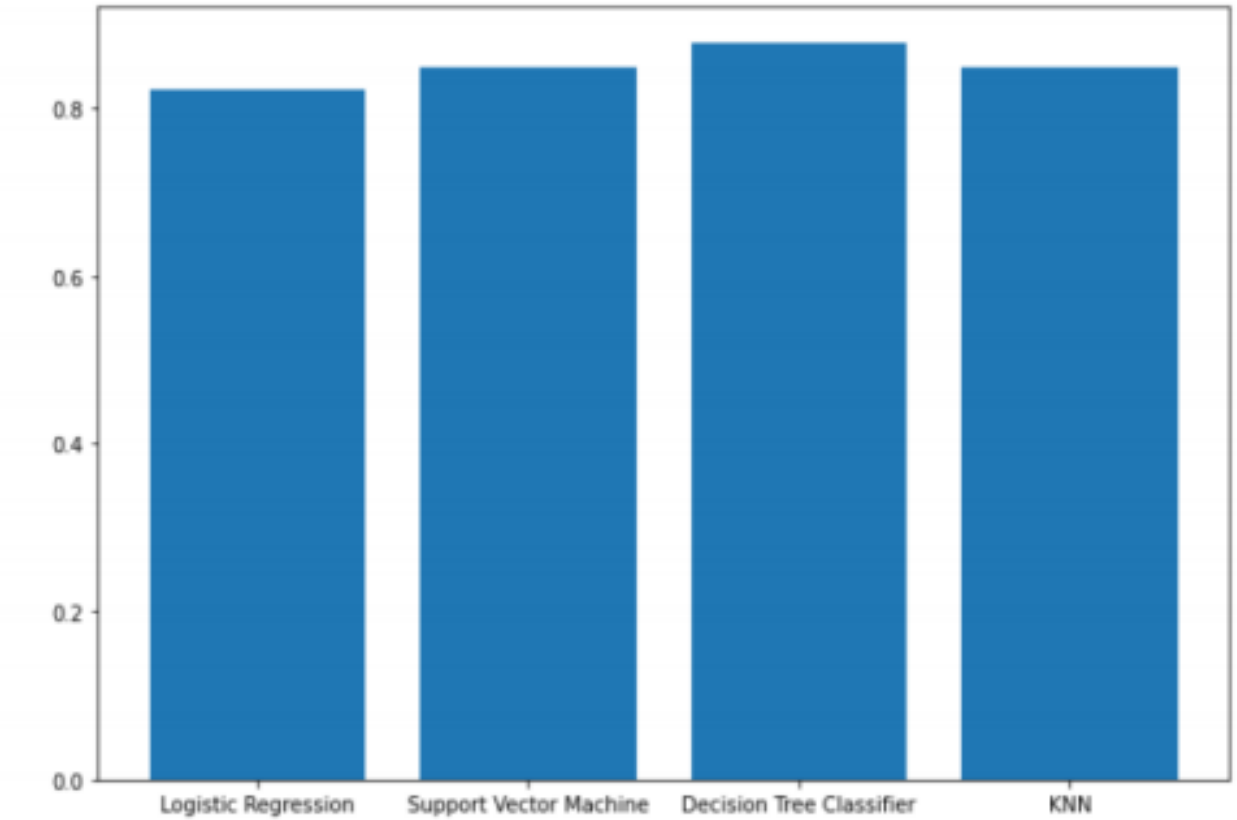


Predictive analysis (Classification)

Classification Accuracy

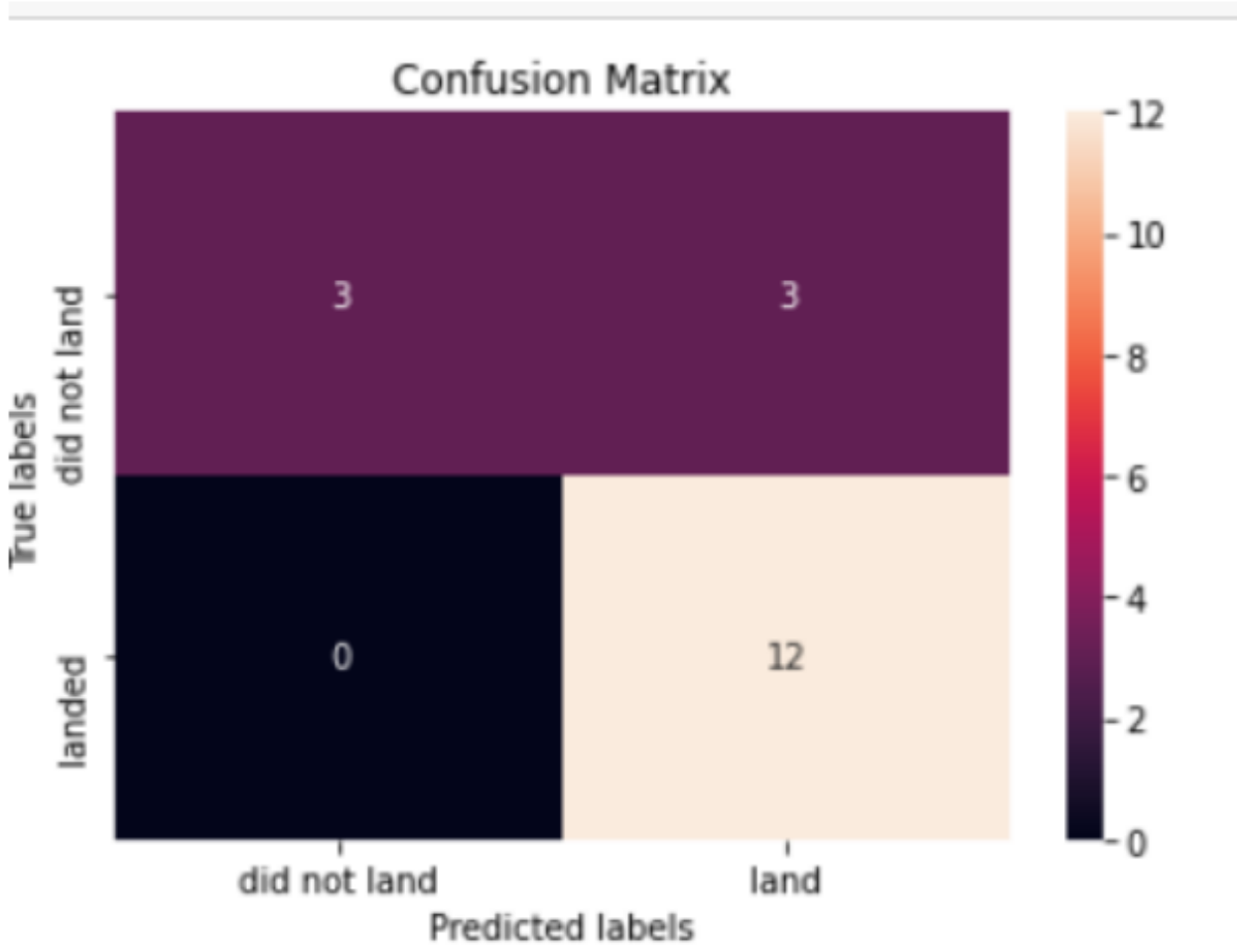
Visualize all the built model accuracy for all built models, in a barchart

The best model: Decision Tree Classifier



Confusion Matrix

From the confusion matrix we see that there are 15 correctly predicted values and 3 wrong. It shows that our model is good.



CONCLUSION

- Data collection
- Data wrangling
- EDA with SQL and Visualization
- Modelling
- …

APPENDIX

- https://github.com/elnursahbalayev/Data-Science-Projects