



IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sergio Díaz Suárez

17 June 2024



Outline

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

Executive Summary

- Summary of methodologies
 - Perform data wrangling after data collection
 - Perform exploratory data analysis using visualization and SQL
 - Perform interactive visual analytics using Folium and Plotly Dash
 - Perform predictive analysis using classification models
- Summary of all results
 - Some orbits are more successful than others
 - Generally, launch sites are far away from cities and well communicated and in the coast
 - Decision tree is the best model to solve this problem

Introduction

- SpaceX Falcon 9 rocket launches **cost: 62 millions dollars**
- Other companies: 165 million dollars or more
- SpaceX can reuse the first stage of a rocket launch
- **Aim: Determining if the first stage will land as costs will decrease**



Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
 - Webscraping and Application Programming Interface (API)
- Perform data wrangling
 - Filtering data and dealing with missing values
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
 - Support Vector Machine, K-Nearest Neighbors, Decision Tree, and Logistic Regression
 - Evaluation models through Jaccard index and confusion matrix

Data Collection

- Two datasets were collected. One dataset was obtained using webscraping from:

https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

- The other dataset was obtained using SpaceX API

<https://api.spacexdata.com/v4/launches/past>

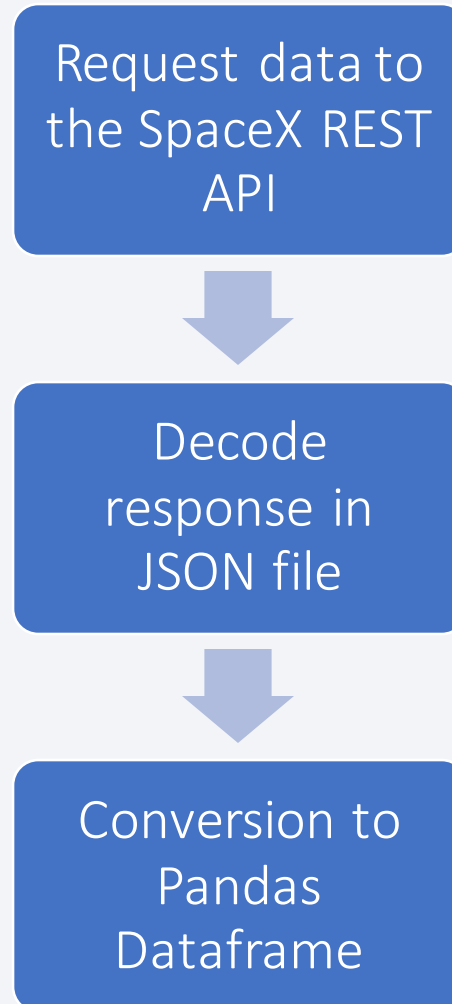
Data Collection – SpaceX REST API

- URL webpage

<https://api.spacexdata.com/v4/launches/past>

- Full information about data extraction can be found on:

<https://github.com/xyonhart/Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb>



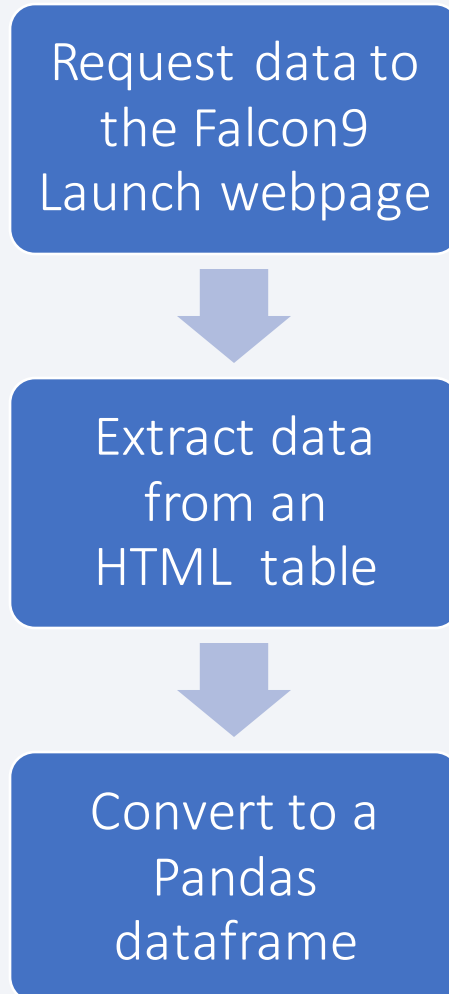
Data Collection – Scraping

- URL webpage


https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

- Full information about data extraction can be found on:

<https://github.com/xyonhart/Applied-Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb>



Data Wrangling

- In this step, we preprocess the data. The operations done are:
 1. Filtering data so that Falcon9 is the only rocket in our datasets
 2. Dealing with missing values  Use average value
 3. Create a landing outcome based on types of (un)successful landing
- Full information about data wrangling can be found on:

<https://github.com/xyonhart/Applied-Data-Science-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb>

EDA with Data Visualization

- The following plots are done to study their influence on the launch outcome:
 - Catplot of the payload mass against the flight number
 - Catplot of the launch site against the flight number
 - Bar plot of the payload mass and the launch site
 - Catplot of the orbit against the flight number
 - Catplot of the orbit against the payload mass
 - Line plot of the average success rate along the years
- Full information about EDA with Data Visualization can be found on:

<https://github.com/xyonhart/Applied-Data-Science-Capstone/blob/main/edadataviz.ipynb>

EDA with SQL

- **SQL queries that were performed**
 - Display 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
 - List the total number of successful and failure mission outcomes
 - List the names of the booster_versions which have carried the maximum payload mass
 - List the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
 - Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order
- Full information about EDA with SQL can be found on:

https://github.com/xyonhart/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- On a folium map, to find geographical patterns, the following objects were added:
 1. Circles with origin in the launch center location
 2. Markers to show the name of the location
 3. Using MarkerCluster objects, markers to show success and failed launches on each site
 4. Lines to calculate distance between the launch center location and a point of interest such as a city
- Full information can be found on:

https://github.com/xyonhart/Applied-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

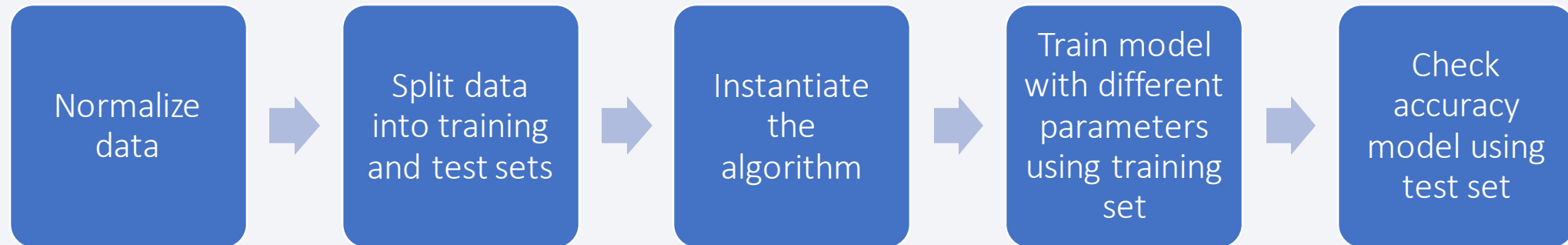
Build a Dashboard with Plotly Dash

- The Dashboard can show:
 1. Pie chart showing the total successful launches count for all launch sites
 2. Pie chart showing the success vs failed counts for the site in a specific launch site
 3. Scatter plot showing the success and failures of rockets against the payload mass for all sites
 4. Scatter plot showing the success and failures of rockets against the payload mass in a specific launch site
- Full information about the Plotly Dash lab can be found on:

https://github.com/xyonhart/Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Different algorithms: KNN, SVM, LR, and decision trees



- Full information can be found on:

https://github.com/xyonhart/Applied-Data-Science-Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

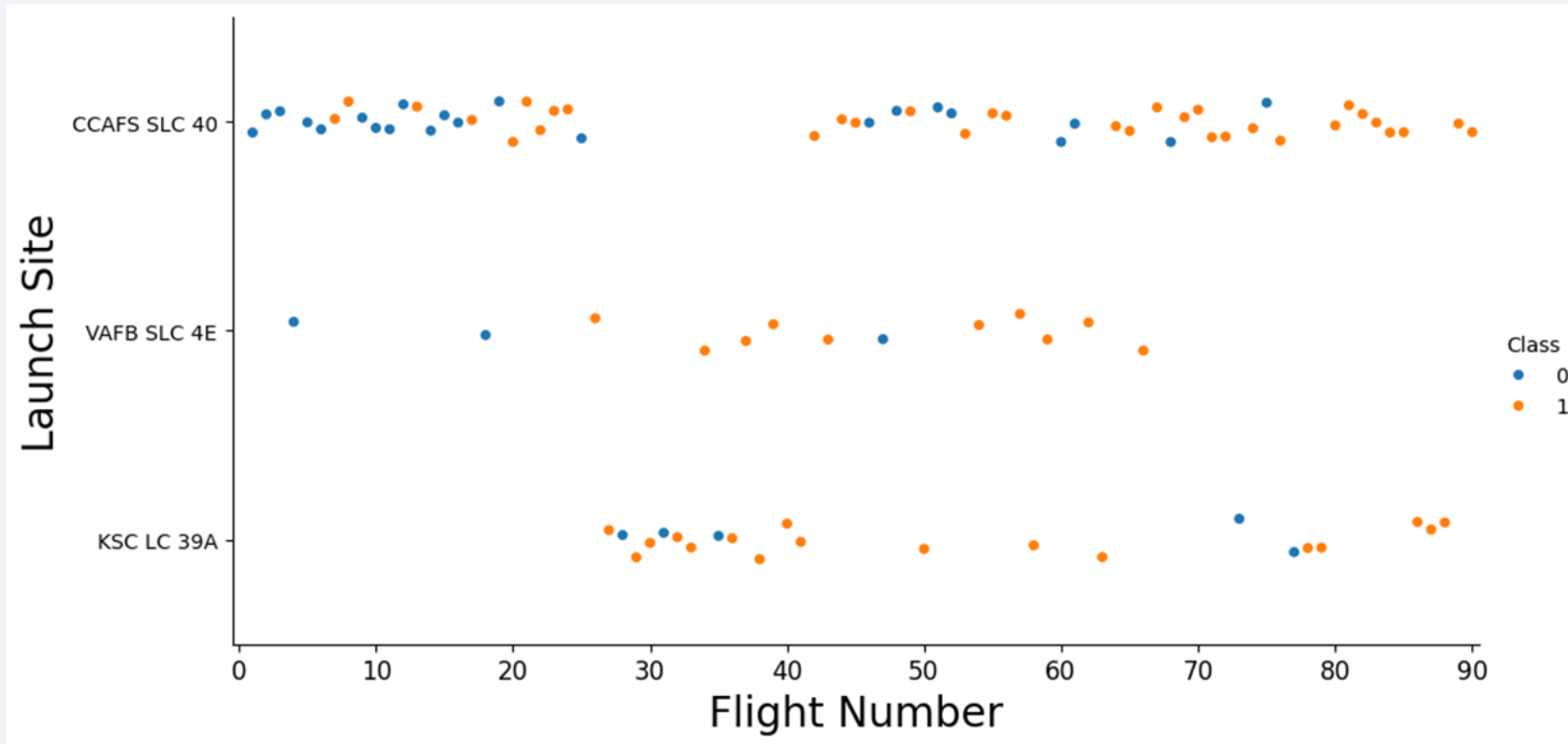
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

The background of the slide is an abstract composition. It features a dark blue base color. Overlaid on this are numerous diagonal streaks in shades of red and cyan. A faint, light blue grid pattern is also visible, particularly in the lower-left quadrant. The overall effect is dynamic and technological.

Section 2

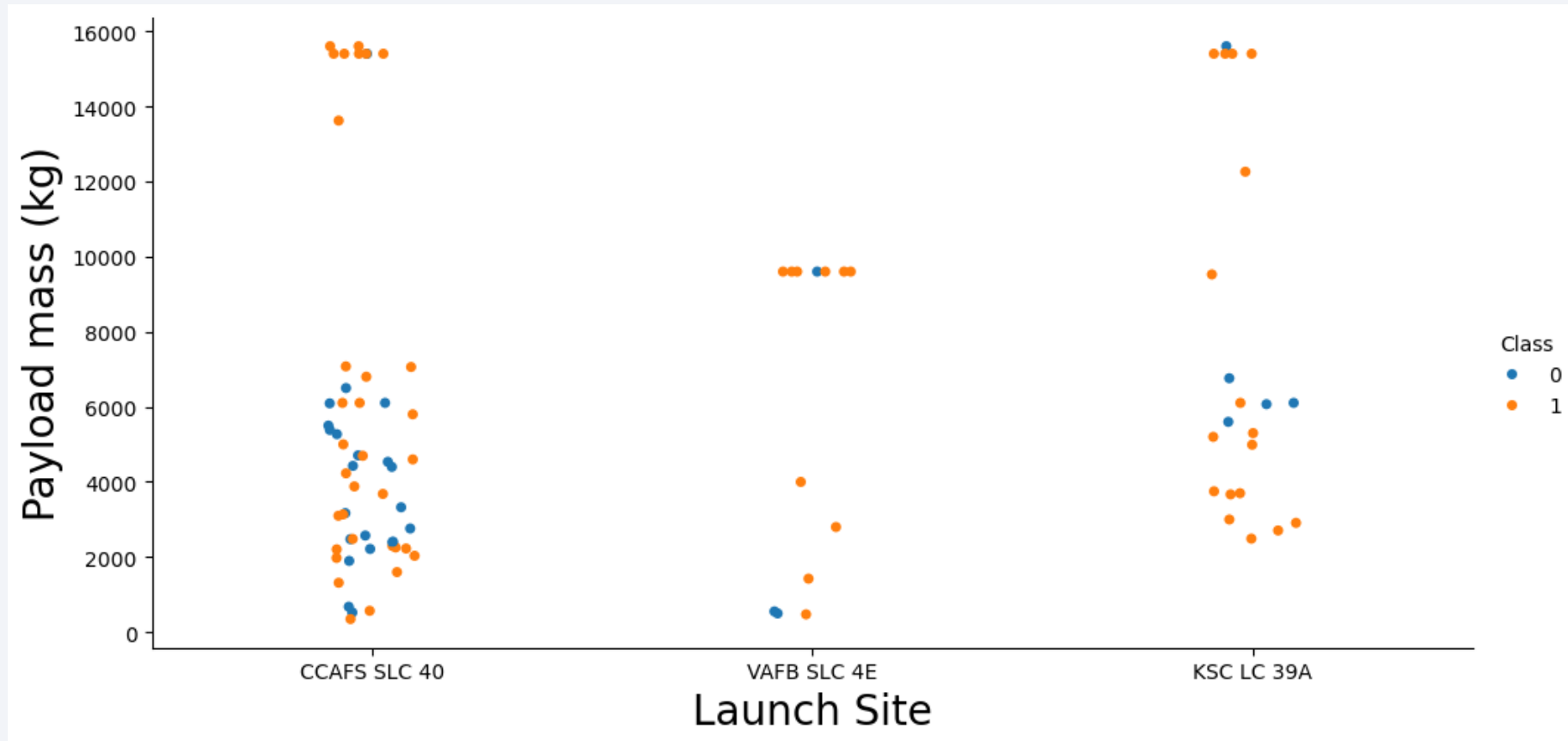
Insights drawn from EDA

Launch Site vs Flight Number



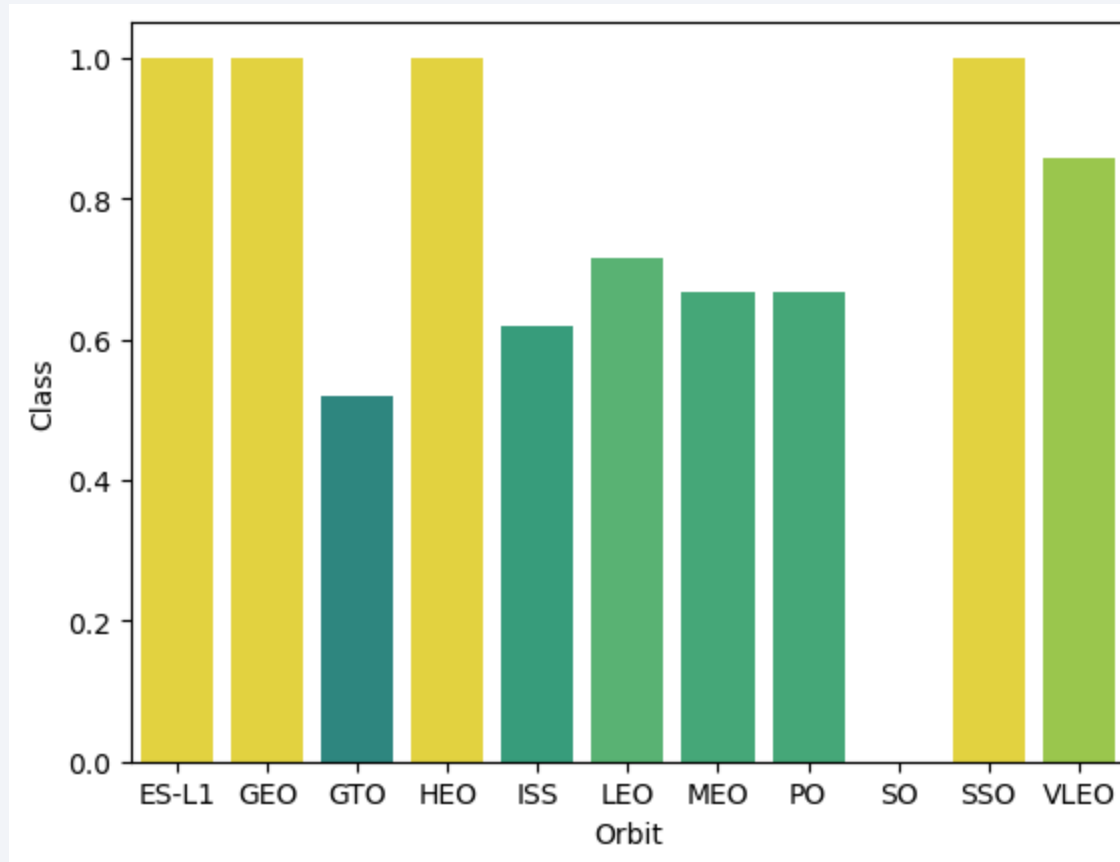
- Most of launches are done in CCAFS SLC 40, but VAFB SLC 4E and KSC LC 39A have highest ratio of success

Payload vs. Launch Site



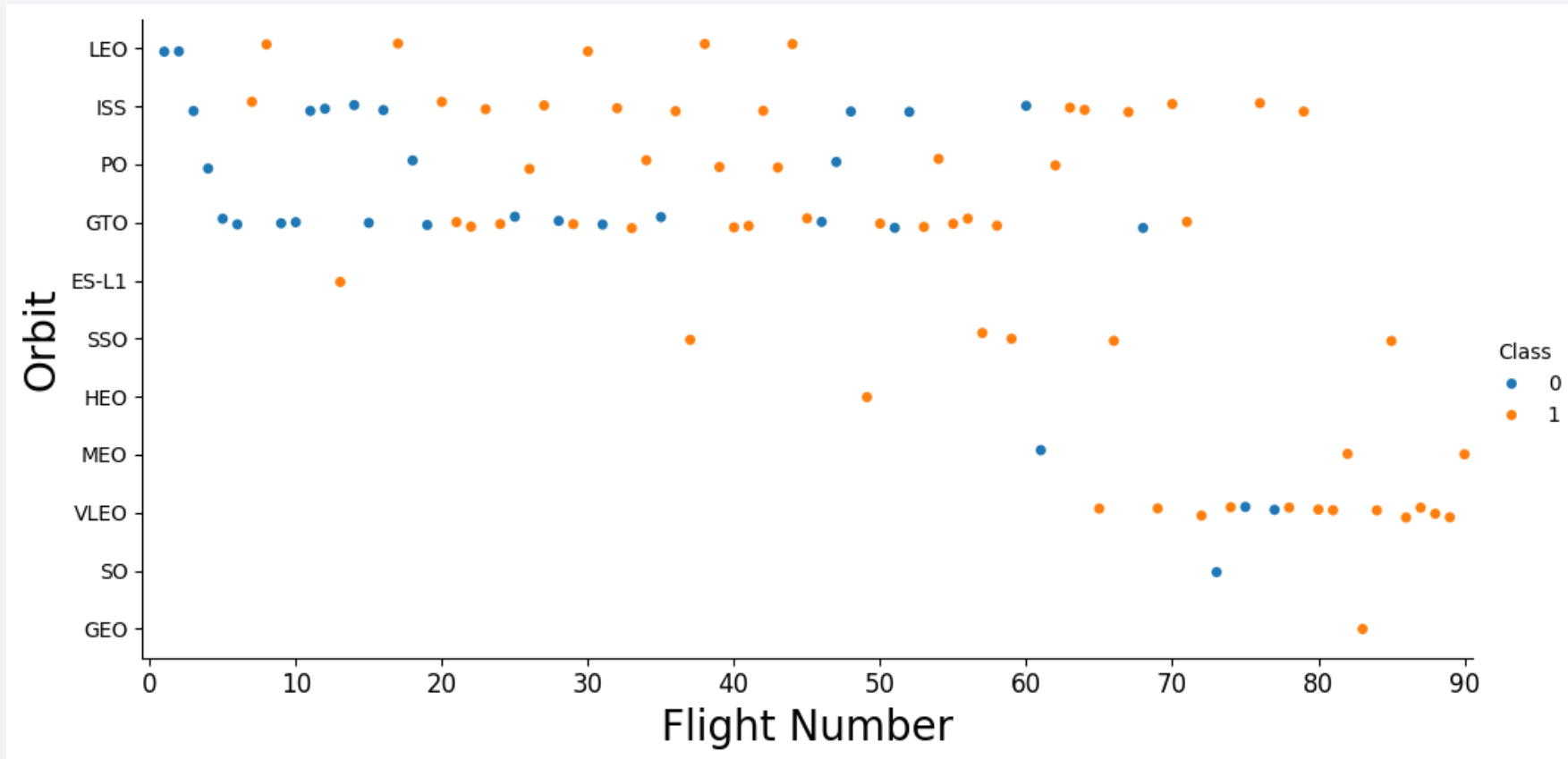
- There is no launches in VAFB SLC 4E with payload mass larger than 10000 kg

Success Rate vs. Orbit Type



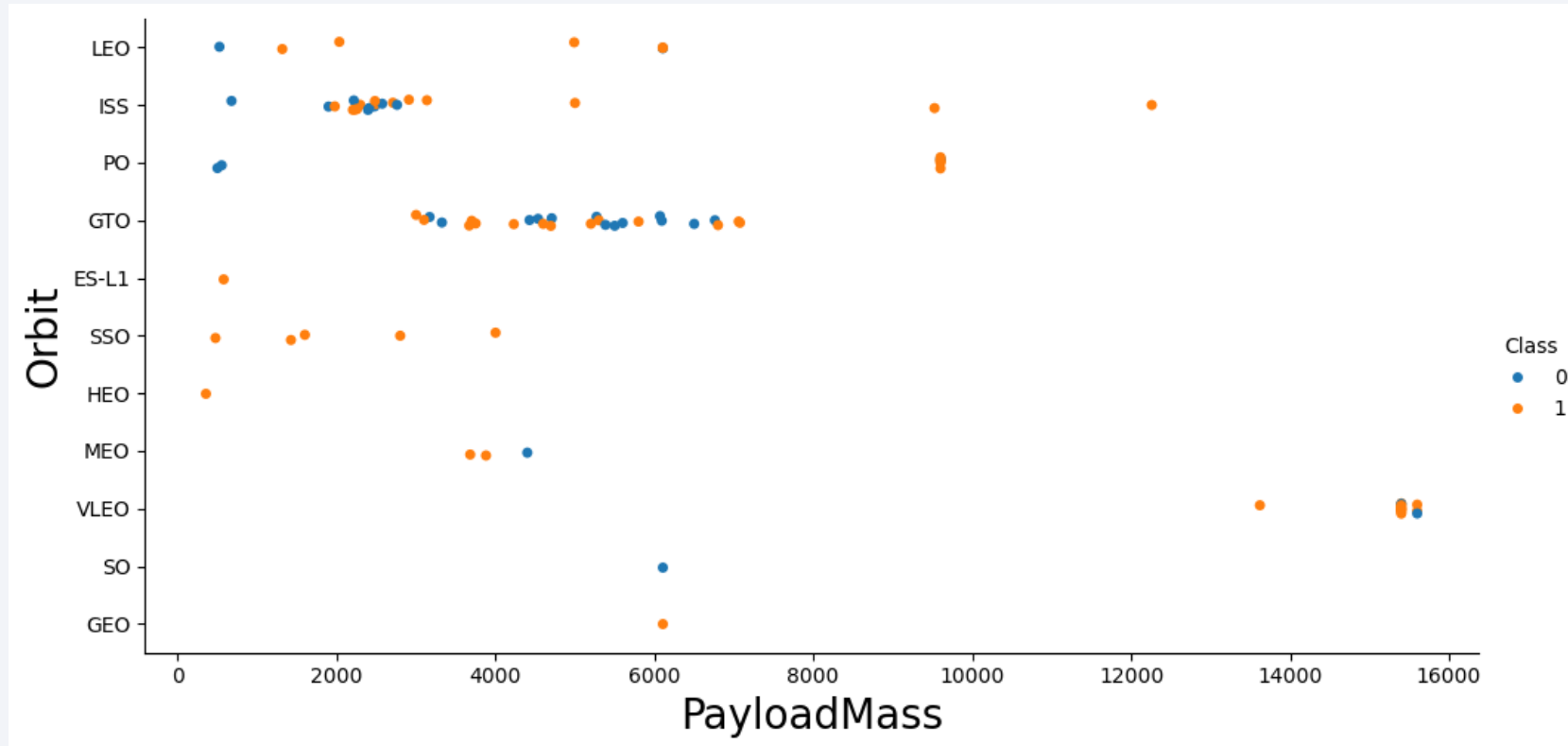
- SO is totally unsuccessful. This is contrary to what happens in Geostationary Orbits, Orbits in L1 point, High Earth Orbit, or SSO.

Orbit Type vs Flight Number



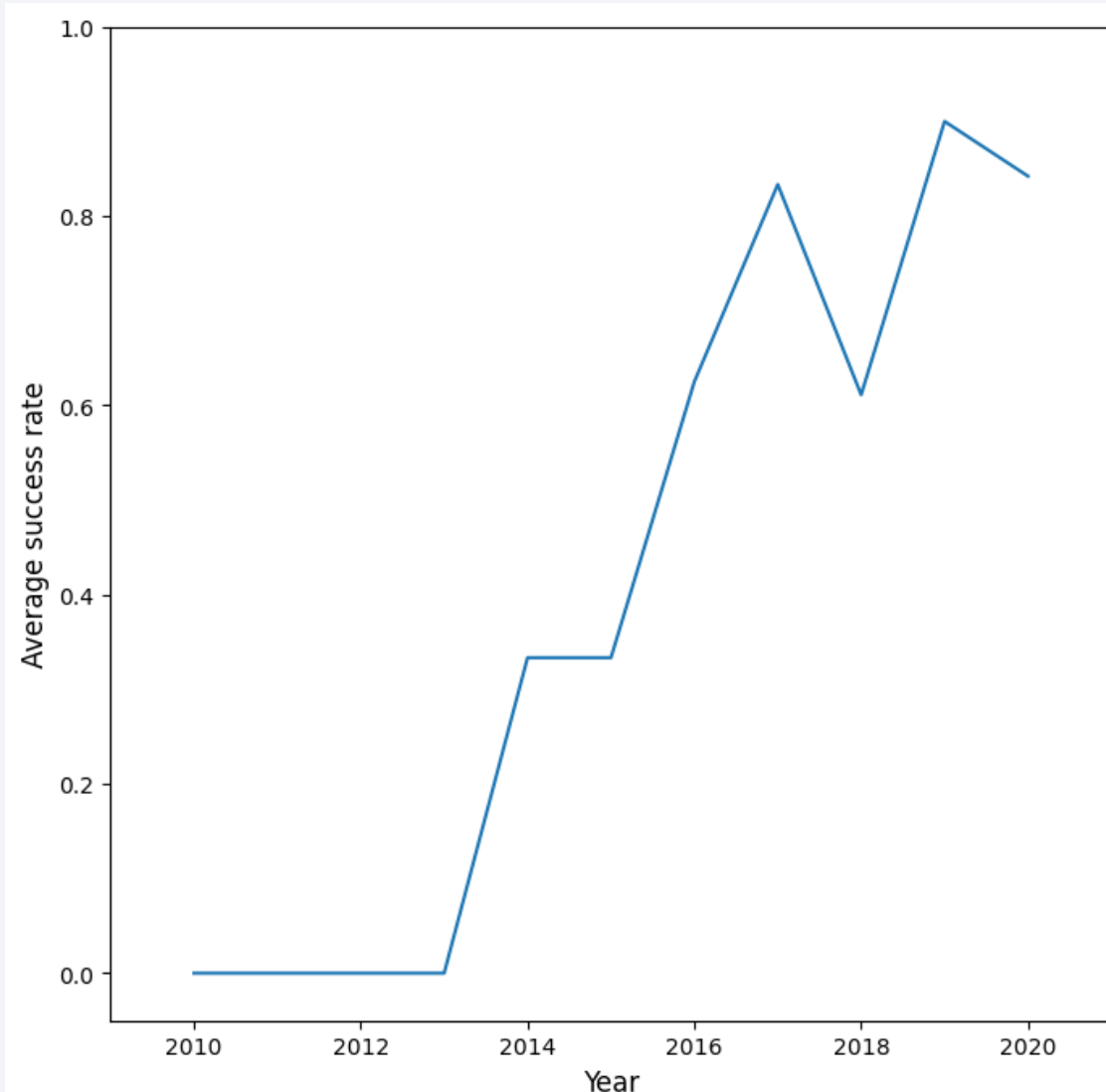
- The trend shown before is fully deglosed here

Orbit Type vs Payload



- SSO orbits are successful at low payloads, while VLEO are frequently successful at high payloads

Launch Success Yearly Trend



- The average success of rocket launch increases from 2013

All Launch Site Names

- The result of the query is:

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

- These are the sites where Falcon9 rocket were launched at least once

Launch Site Names Begin with 'CCA'

[10]:

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

- Records of the first five places where the name starts with CCA. Includes CCAFS LC-40 only but CCAFS SLC-40 records are valid also

Total Payload Mass

```
Out[11]: SUM_OF_PAYLOAD_MASS_KG  
45596
```

- Here a sum is applied to payloadmass when customer is NASA (CRS)

Average Payload Mass by F9 v1.1

```
Out[12]: AVERAGE_PAYLOAD_MASS_KG  
2928.4
```

- Here the average is applied to payloadmass when the booster version is F9 v1.1

First Successful Ground Landing Date

- The first successful landing outcome on ground pad was on:

22-12-2015

- A minimum function is applied on date and the LANDING_OUTCOME column should be like '%ground%'

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and their payload mass greater than 4000 but less than 6000 are shown here

```
Out[14]: Booster_Version
```

```
F9 FT B1020
```

```
F9 FT B1022
```

```
F9 FT B1026
```

```
F9 FT B1021.2
```

```
F9 FT B1031.2
```

Total Number of Successful and Failure Mission Outcomes

- We found that the number of successful missions outcomes was 100
- We found that the number of failed missions outcomes was 1
- In this case, it was necessary to select cases as succesful or notsuccesful and count each of them

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

Out[16]: **Booster_Version**

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

2015 Launch Records

- Failed landing_outcomes in drone ship, booster versions, and launch site names for in year 2015 are:

Out[17]:	MONTH	YEAR	Booster_Version	Landing_Outcome	Launch_Site
	01	2015	F9 v1.1 B1012	Failure (drone ship)	CCAFS LC-40
	04	2015	F9 v1.1 B1015	Failure (drone ship)	CCAFS LC-40

- To extract months and years from Date, we used SUBSTR(Date,x,y) function where x and y are positive integers

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

Out[18]:

TYPE_LANDING	COUNTING
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

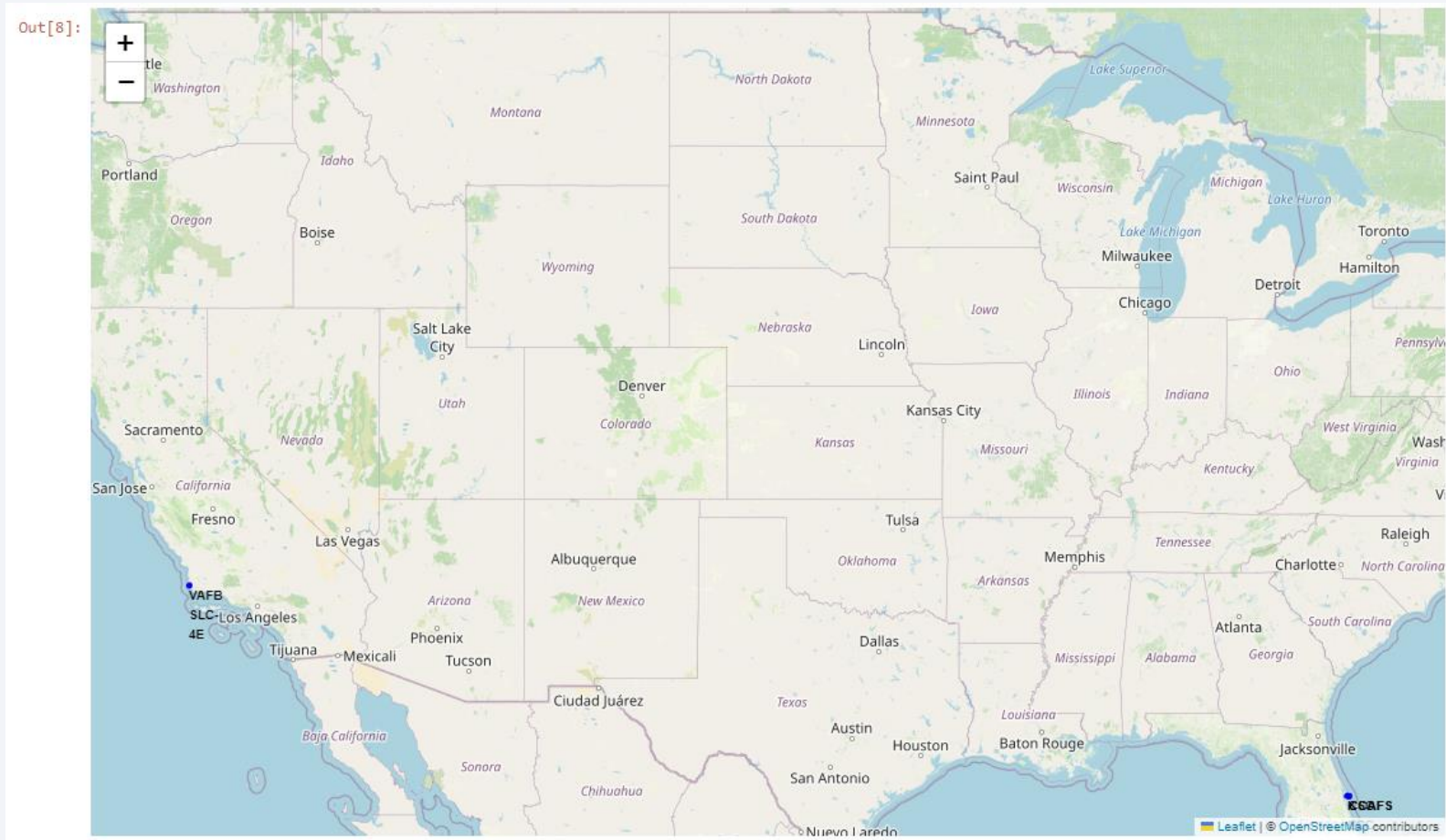
- Different types of landing although they can be englobed in failure and succes

A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The image is dark blue with a thin white line representing the horizon. The city lights are visible as bright yellow and orange spots against the dark blue background of the night sky.

Section 3

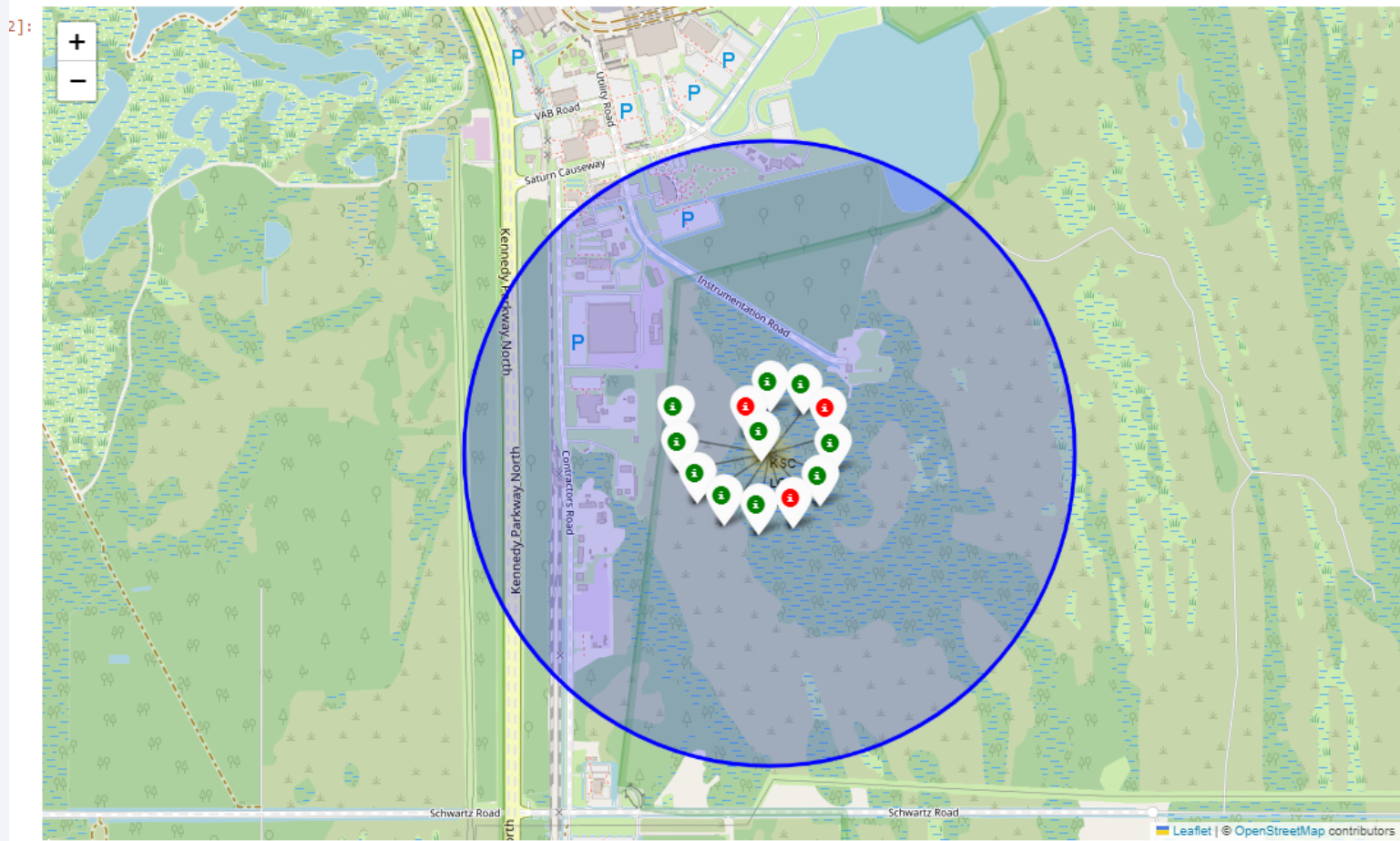
Launch Sites Proximities Analysis

Location of the launch sites



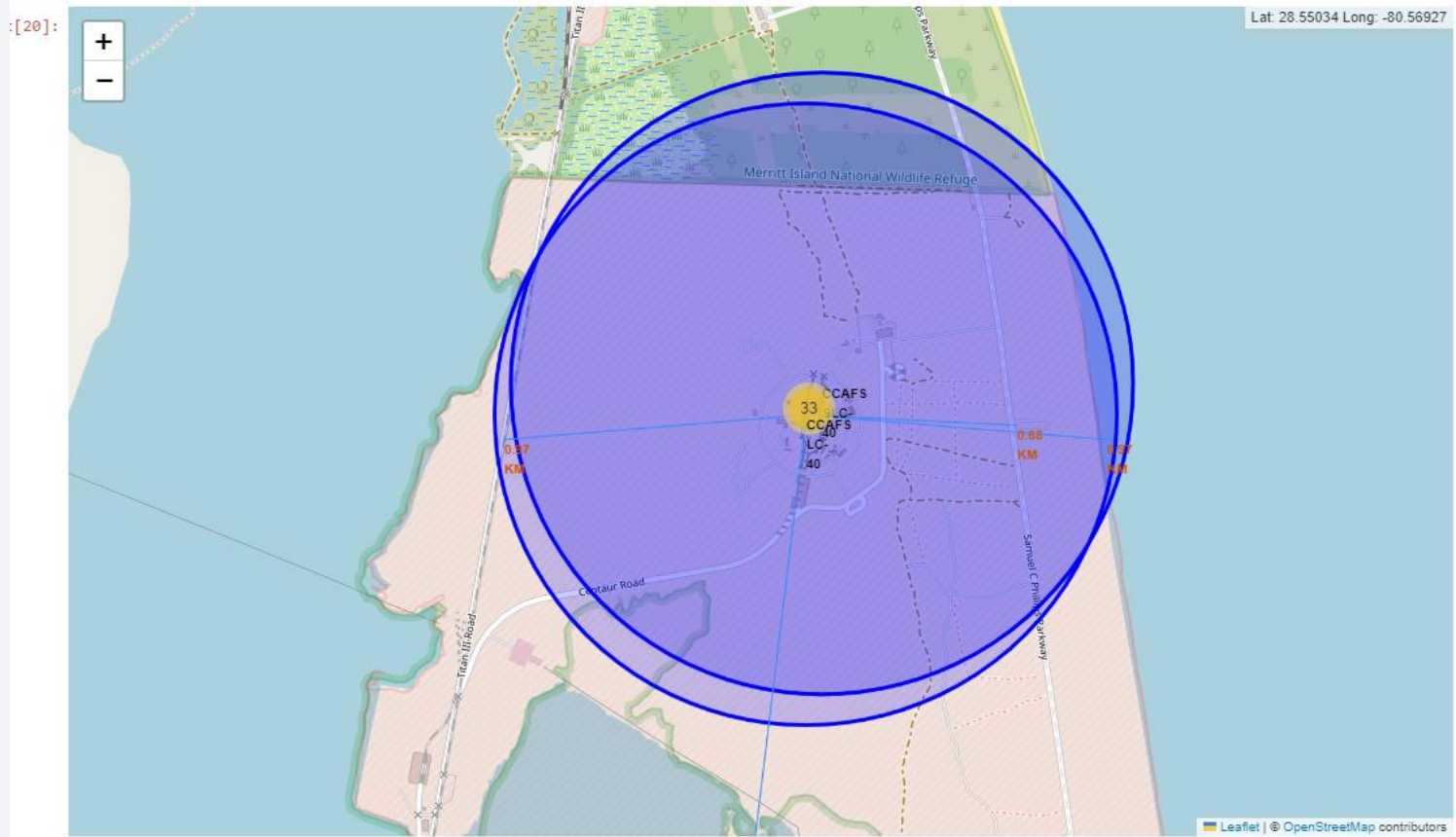
- Map of USA showing the locations of launch in blue
- Label is also included in each case

Location of KSC and statistical of launches



- The blue circle has a radius of 1 km and the center has the coordinates of KSC
- Red label means a failed mission
- Green label means a successful mission

CCAFS LC-40 site: well communicated, but isolated



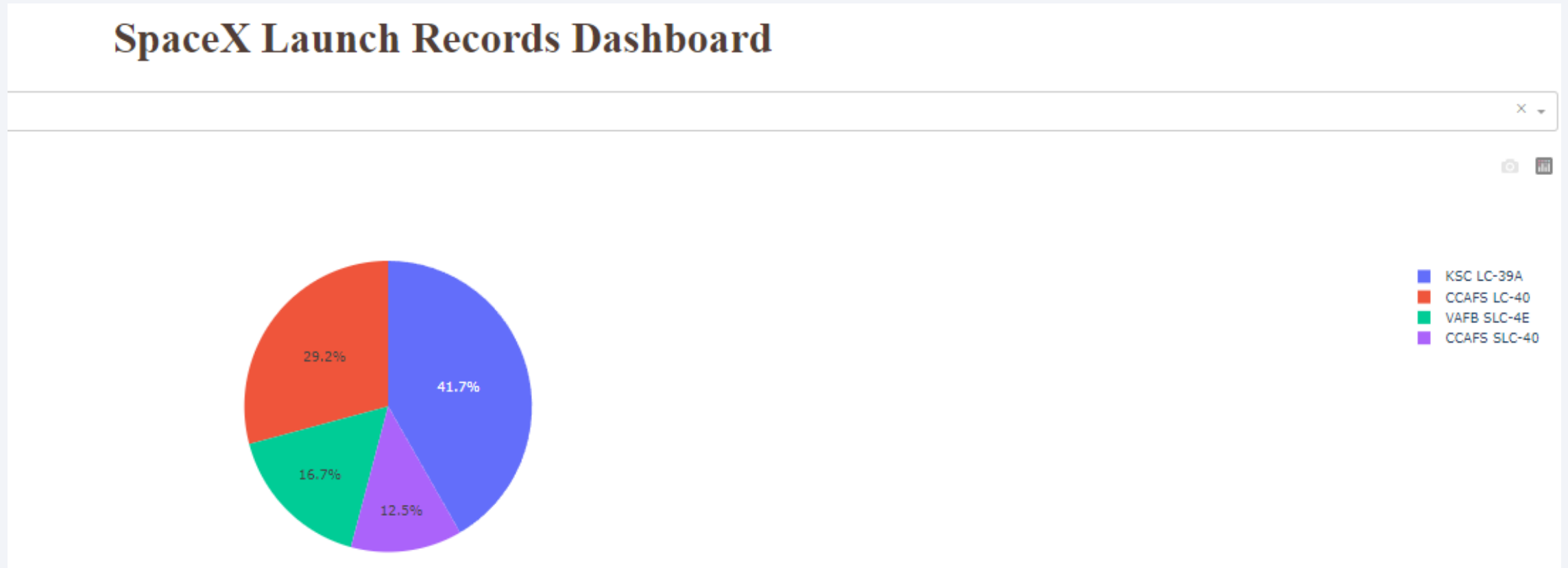
- Blue circle centers as before, but for CCAFS LC-40 and CCAFS SLC 40
- Cyan lines
 1. West: distance from CCAFS LC-40 to nearest railway
 2. East: same but distances to coast and highway
 3. South: same but minimum distance to Cape Canaveral distance (17 km)
- Orange labels: distances



Section 4

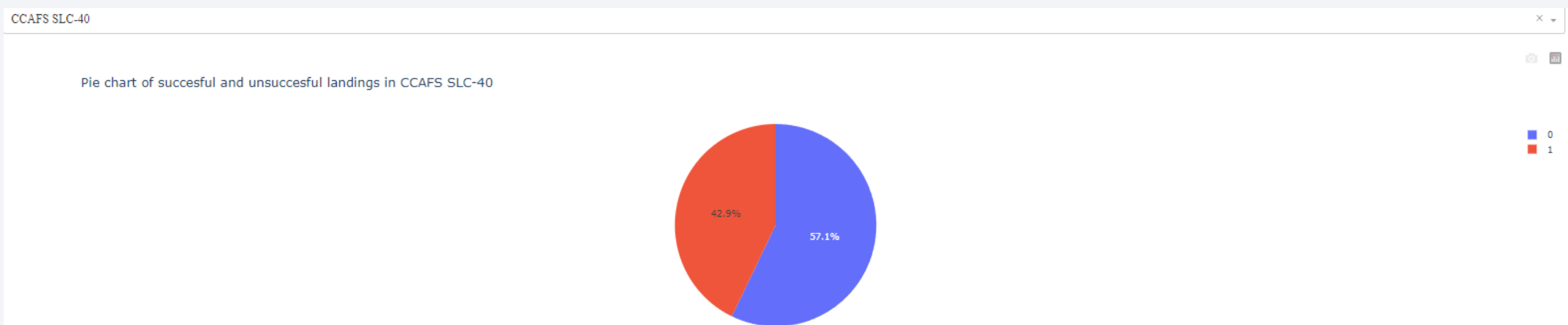
Build a Dashboard with Plotly Dash

Piechart of successful launches by launch site



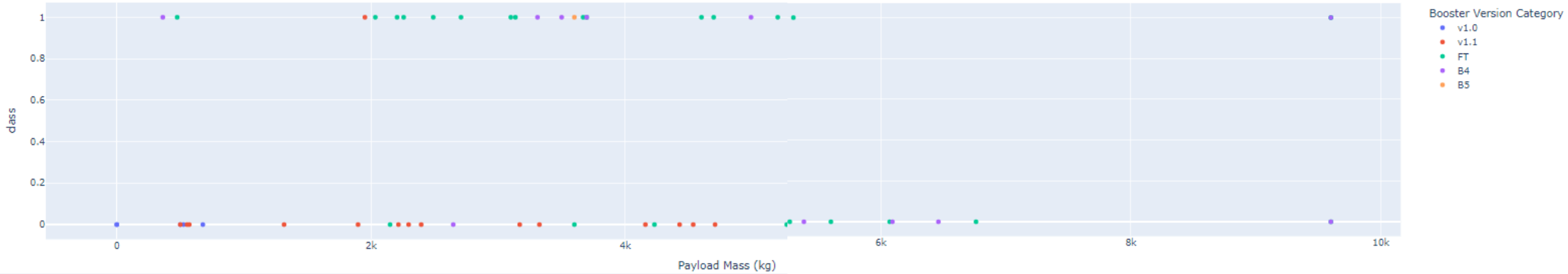
- KSC LC-39A has the largest number of successful launches. CCAFS SLC-40 has the lowest number of successful launches

Piechart of the highest ratio of successful launches



- Nonetheless, CCAFS SLC-40 has the highest ratio of successful launches

Scatterplot of successful/ failed mission against payload mass



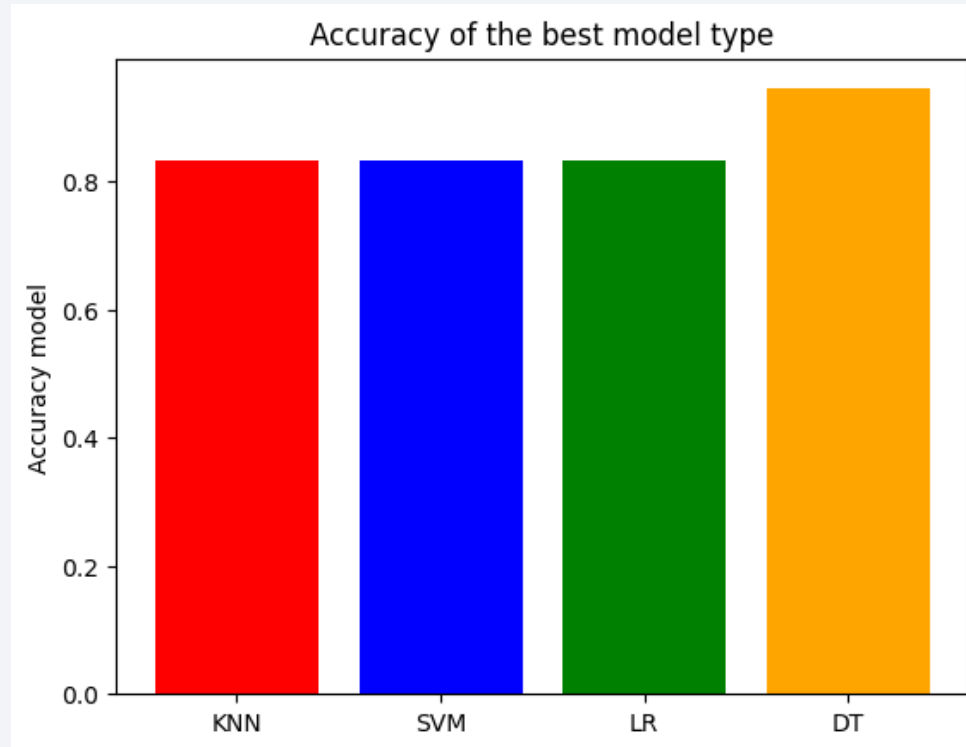
- The FT Booster Version seems the optimal version for a rocket launch



Section 5

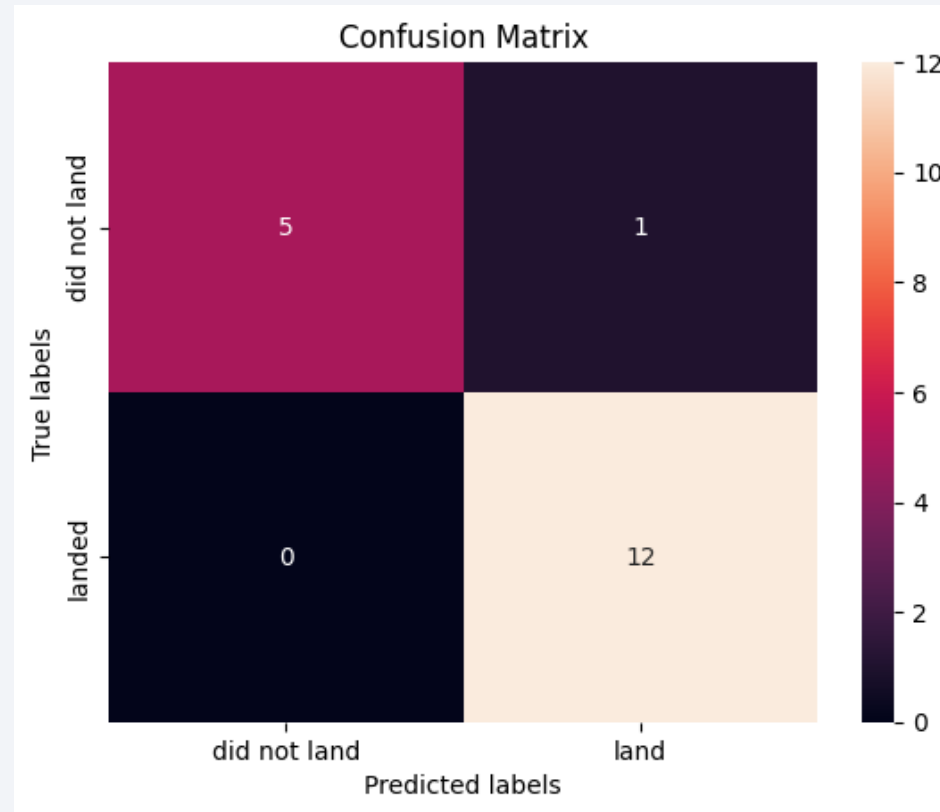
Predictive Analysis (Classification)

Classification Accuracy



- Decision Tree is the best model for this problem

Confusion Matrix



- Decision Tree is the best model for this problem. Only 1 False Positive is found. Other models found 3 False Positives.

Conclusions

- SO is totally unsuccessful unlike GEO, ES-L1, or HEO
- On average, the successful landing is improved
- CCAFS SLC-40 has the highest ratio of successful launches
- CCAFS LC-40 is well communicated by train or highway, but far away from cities. Also close to coast.
- The FT booster version has the highest ratio of successful launches
- Decision tree is the best model to determine if the first stage of Falcon9 rocket will land successfully or not

Appendix

In order to improve performance on Decision Tree, parameters have been modified so that:

- 1) 'min_samples_leaf': [1, 2, 4, 6, 8]
- 2) 'min_samples_split': [2, 5, 7, 8, 10]
- 3) 'auto' has been disabled due to model's divergence

Additionally, GitHub does not show correctly maps. Instead I used the following webpage to show results:

https://nbviewer.org/github/xyonhart/Applied-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Thank you!

