Winning Space Race with Data Science

Nguyen Cao Long
26/09/2021

Outline

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

Executive Summary

- Use of machine learning algorithm to build a predictive model to help business determine whether a rocket will land successfully based on SpaceX Falcon 9 data.
- Results:
  - Support Vector Machines (SVM): 88.89% (accuracy score), 91.67% (f1 score)
  - Logistic Regression: 88.89% (accuracy score), 92.31% (f1 score)
  - K-nearest Neighbors (KNN): 96% (accuracy score), 94.44% (f1 score)
  - Decision Tree Classifier: 66.67% (accuracy score), 80% (f1 score)
- Achieved the main goal as having models to predict future rocket launches.

Introduction

- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars, other providers cost upwards of 165 million dollars each.
- Savings are due to the reuse the first stage.
- Problem: Determine whether the first stage will land as determine the cost of a launch.
- Main goal: Implement machine learning to build a predictive model to help business solve problem efficiently.
- Secondary goal: Explore and analyze SpaceX Falcon 9 competition.

Section 1

Methodology

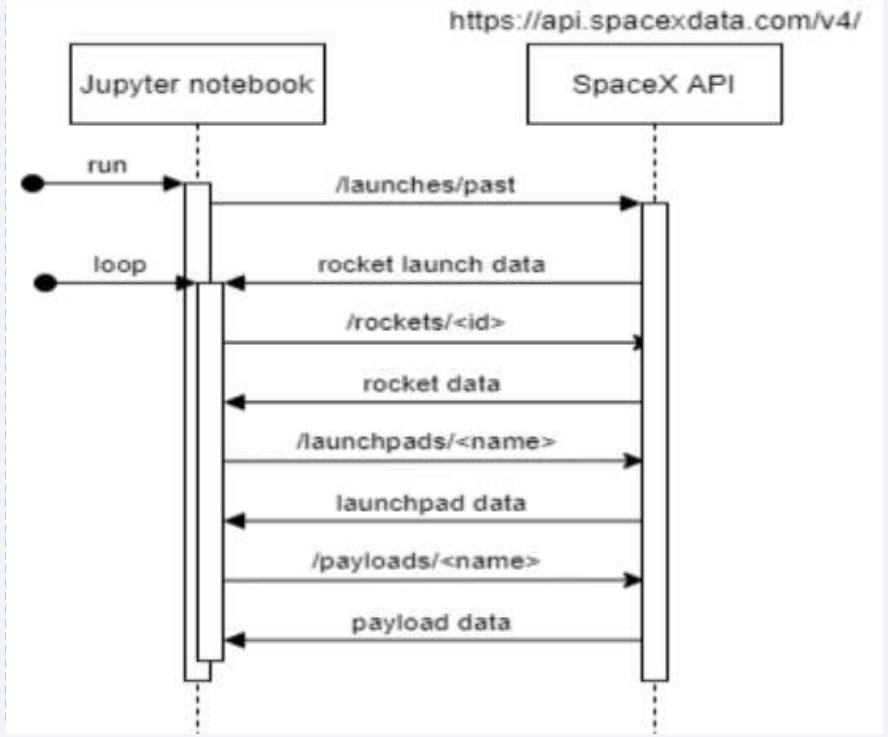
Methodology

Executive Summary

- Data collection methodology
  - Data collected from SpaceX API and Wikipedia web-scraping.
- Perform data wrangling
  - Clean-up dates, null values and filter only for Falcon 9 rockets
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Algorithms: SVM, KNN, Logistic Regression, Decision Tree
  - Hyperparameter tuning: Grid Search CV
  - Evaluation metrics: F1 score, accuracy score

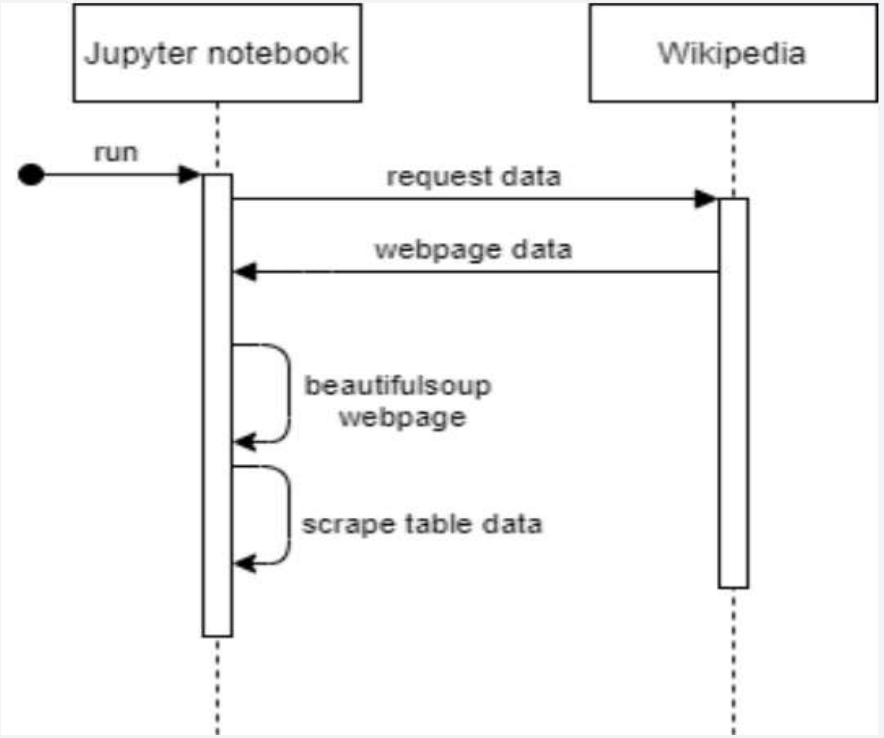
Data Collection – SpaceX API

- Call to the main SpaceX API to gather data about previous launches.
- Enrich data with specific calls to other API endpoints.
- https://github.com/123olala/IBM-Data-Science-Professional-Certificate-Applied-Data-Science-Capstone/blob/main/DataCollectionAPI.ipynb



Data Collection - Scraping

- Call Wikipedia webpage
  - https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches
- Scrape table data from past launches
- https://github.com/123olala/IBM-Data-Science-Professional-Certificate-Applied-Data-Science-Capstone/blob/main/DataCollectionwithWebScraping.ipynb



Data Wrangling

- Landing outcomes (figure) need to be converted into 2 values:
  - 0 if landing was not successful
  - 1 if landing was successful
- First word in rows show if landing was successful or not:
  - True: successful
  - False or None: unsuccessful
- A new column to the dataset was added with 0 or 1 values
- https://github.com/123olala/IBM-Data-Science-Professional-Certificate-Applied-Data-Science-Capstone/blob/main/DataWrangling.ipynb

	Outcome	Values Count
0	True ASDS	41
1	None None	19
2	True RTLS	14
3	False ASDS	6
4	True Ocean	5
5	False Ocean	2
6	None ASDS	2
7	False RTLS	1

EDA with Data Visualization

- Flight Number vs. Payload Mass: Visualize the increase trend on mass and the flight number when more launches were performed (scatterplot).
- Flight Number vs. Launch Site: Observe if the launch site had effect on successful launches and how the sites changed over time (scatterplot).
- Payload vs. Launch Site: In VAFB SLC 4E launch site, the payload mass has a greater impact, and shows that a higher mass tends to a higher success rate. The launch site CCAFS SLC 40 and VAFB SLC 4E seem have smaller pay load mass in general than KSC LC 39A (scatterplot and bar chart).
- Success rate vs. Orbit: ES-L1, GEO, HEO and SSO orbits have a perfect success rate. VLEO falls close with a +80% score (bar chart).

EDA with Data Visualization

- Flight Number vs. Orbit: LEO orbit the success appears related to number of flights, on the other hand, there seems to be no relationship between flight number and success rate when in GTO orbit (scatterplot).
- Payload vs. Orbit: Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits (scatterplot).
- Launch success yearly trend: The success rate since 2013 kept increasing till 2020 (line chart).
- https://github.com/123olala/IBM-Data-Science-Professional-Certificate-Applied-Data-Science-Capstone/blob/main/EDAwithDataVisualization.ipynb

EDA with SQL

- Explore different launch site locations.
- Determine which day happened the first success mission.
- Explore what boosters had success missions.
- Summarize the outcomes of missions from different landing sites.
- Analyze the impact of payload mass on boosters and missions.
- https://github.com/123olala/IBM-Data-Science-Professional-Certificate-Applied-Data-Science-Capstone/blob/main/EDAwithSQL.ipynb

Build an Interactive Map with Folium

- For each launch site, a Circle object based on its coordinate (Lat, Long) values is added with a popup label for site name. A marker is added to show the site name.
- Clusters are used to group launches from same site and they are also to show color of green for success and red for failure.
- Lines are used to show distances from launch to railway and to a city.
- https://github.com/123olala/IBM-Data-Science-Professional-Certificate-Applied-Data-Science-Capstone/blob/main/InteractiveVisualAnalyticswithFolium.ipynb

Build a Dashboard with Plotly Dash

- The first graph is a pie chart to show success rates of a single location or all locations using a dropdown menu.
- The second graph is a scatter plot to visually observe how payload may be correlated with mission outcomes for selected site(s). It has two inputs:
  - Site to be displayed (with ALL option) using a dropdown.
  - Payload mass range filter with a slider.
- https://github.com/123olala/IBM-Data-Science-Professional-Certificate-Applied-Data-Science-Capstone/blob/main/Dashboard.ipynb

Predictive Analysis (Classification)

- Process of obtaining a good model for classification:
  - Load data, select features (X) and target (y)
  - Data preparation, split data into training and test set (80/20)
  - Train different models with tuning of hyperparameters using GridSearchCV
  - Test different models with score function and visualize evaluation with confusion matrix
  - Select best model which has the highest score
- https://github.com/123olala/IBM-Data-Science-Professional-Certificate-Applied-Data-Science-Capstone/blob/main/PredictiveModelBuilding.ipynb

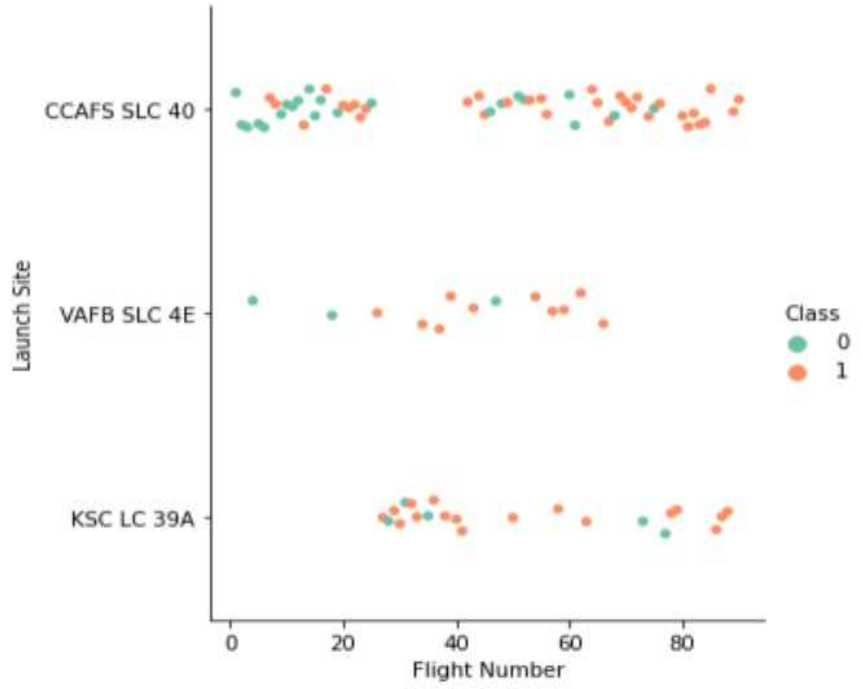
Results

- Having a unique dataset with enriched data of SpaceX Falcon 9 launches.
- Results:
  - Support Vector Machines (SVM): 88.89% (accuracy score), 91.67% (f1 score)
  - Logistic Regression: 88.89% (accuracy score), 92.31% (f1 score)
  - K-nearest Neighbors (KNN): 96% (accuracy score), 94.44% (f1 score)
  - Decision Tree Classifier: 66.67% (accuracy score), 80% (f1 score)
- Achieved the main goal as having trustworthy models to predict future rocket launches.

Section 2

Insights drawn from EDA

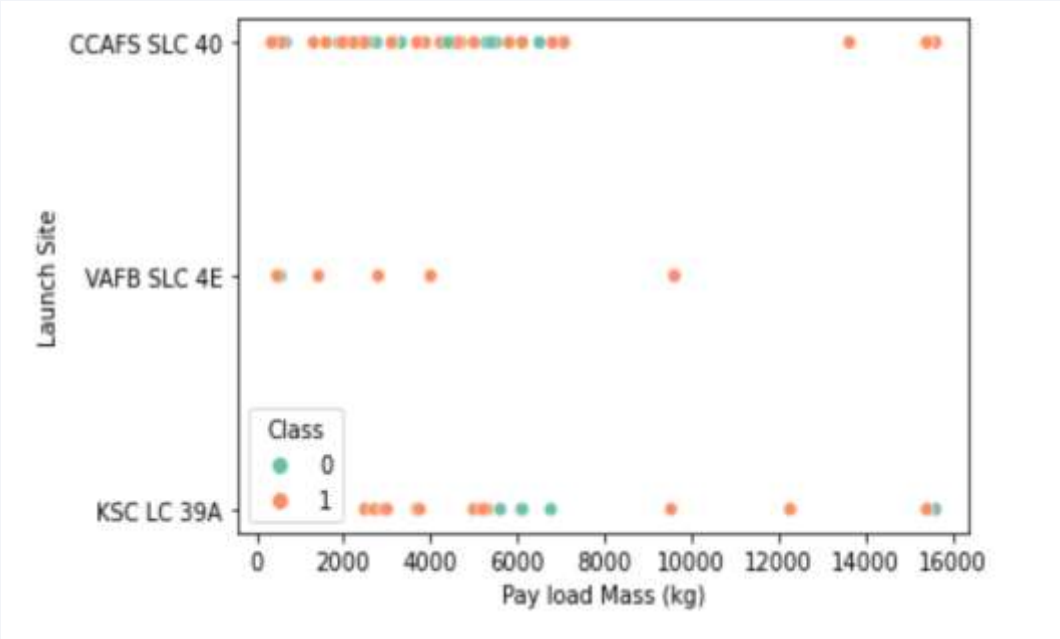
Flight Number vs. Launch Site

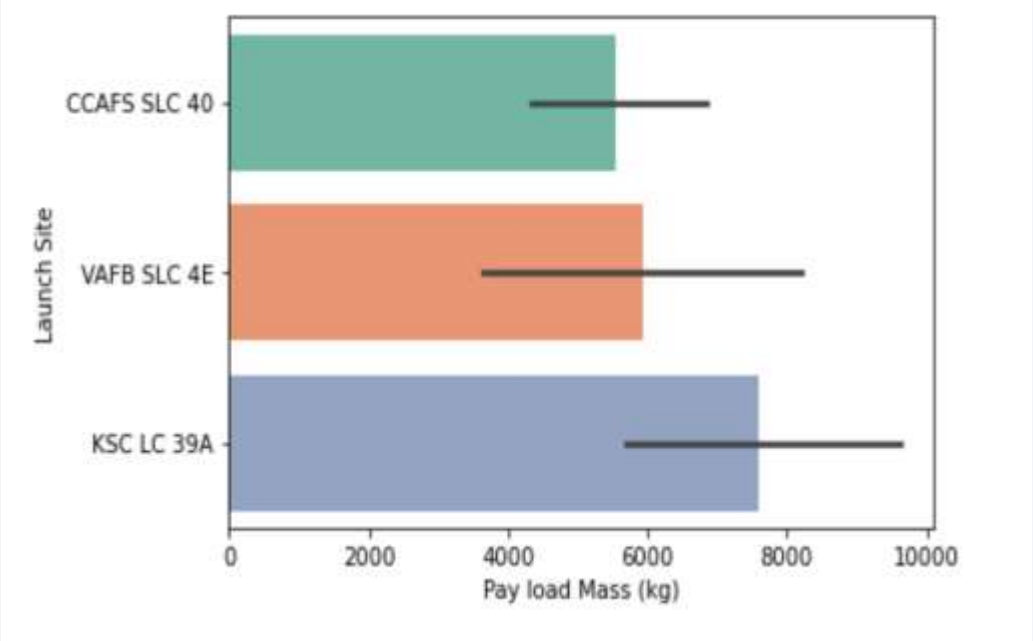


Now try to explain the patterns you found in the Flight Number vs. Launch Site scatter point plots.

- We see that as the flight number increases, the first stage is more likely to land successfully.
- It seems launch site CCAFS SLC 40 has a lower success rate than others

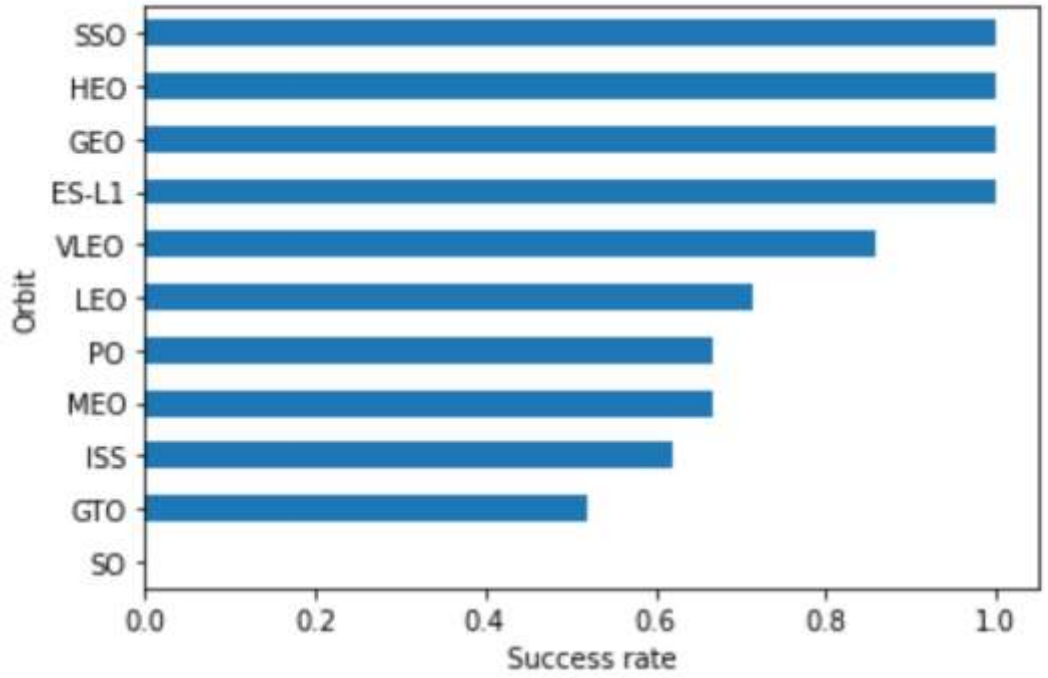
Payload vs. Launch Site





- In VAFB SLC 4E launch site, the payload mass has a greater impact, and shows that a higher mass tends to a higher success rate. The launch site CCAFS SLC 40 and VAFB SLC 4E seem have smaller pay load mass in general than KSC LC 39A

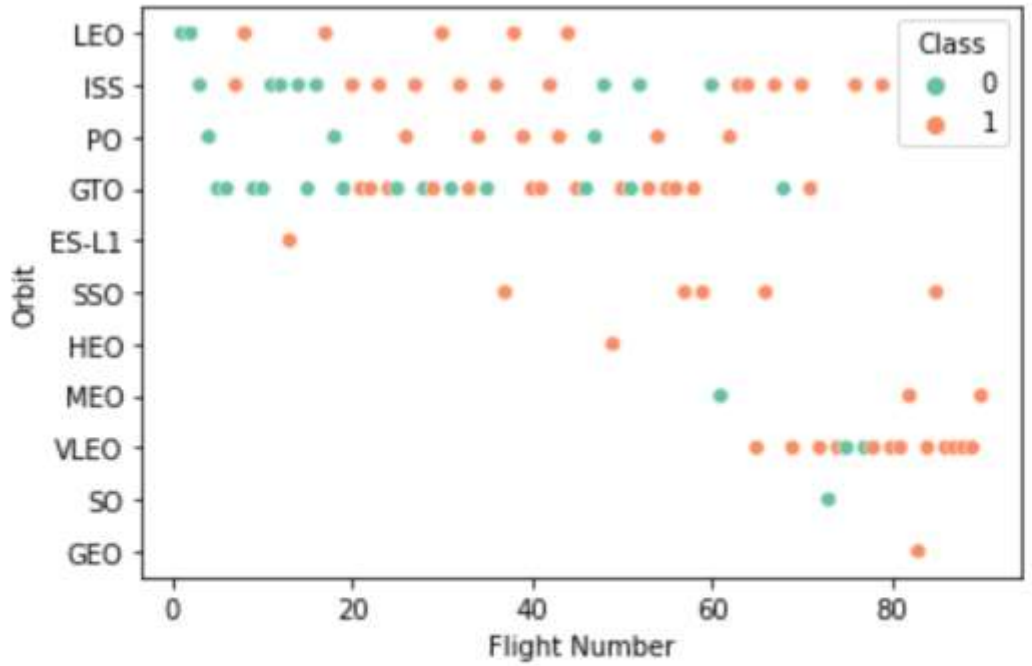
Success Rate vs. Orbit Type



Analyze the ploted bar chart try to find which orbits have high sucess rate.

- ES-L1, GEO, HEO and SSO orbits have a perfect success rate. VLEO falls close with a +80% score

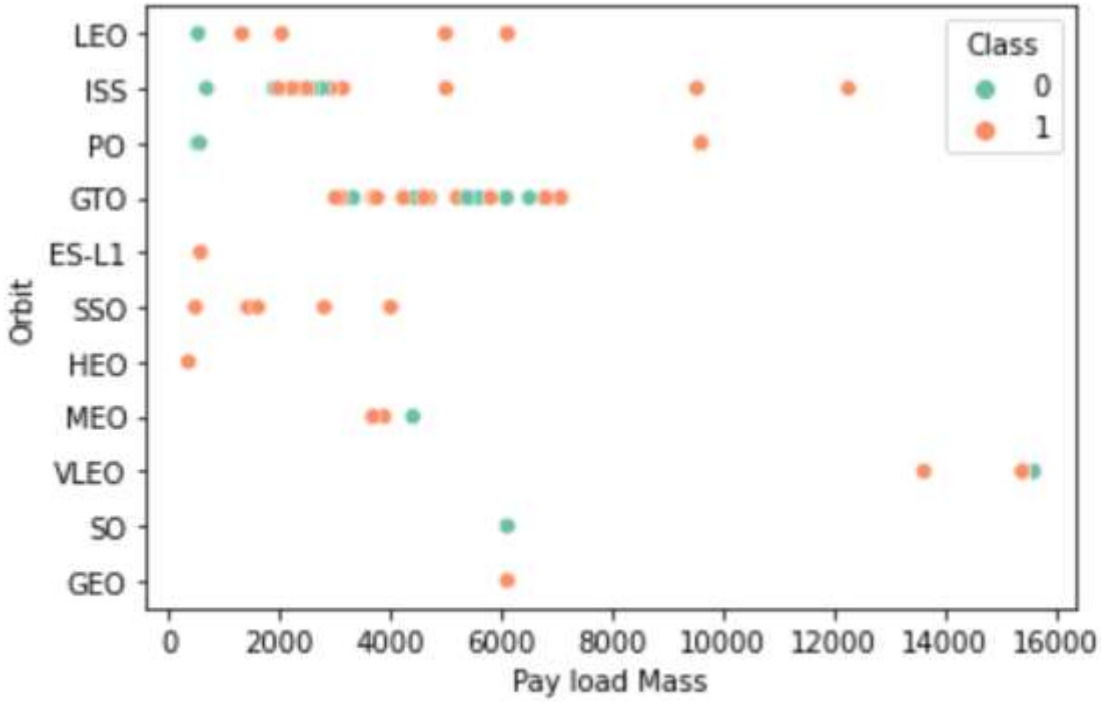
Flight Number vs. Orbit Type



The LEO orbit the Success appears related to the number of flights;

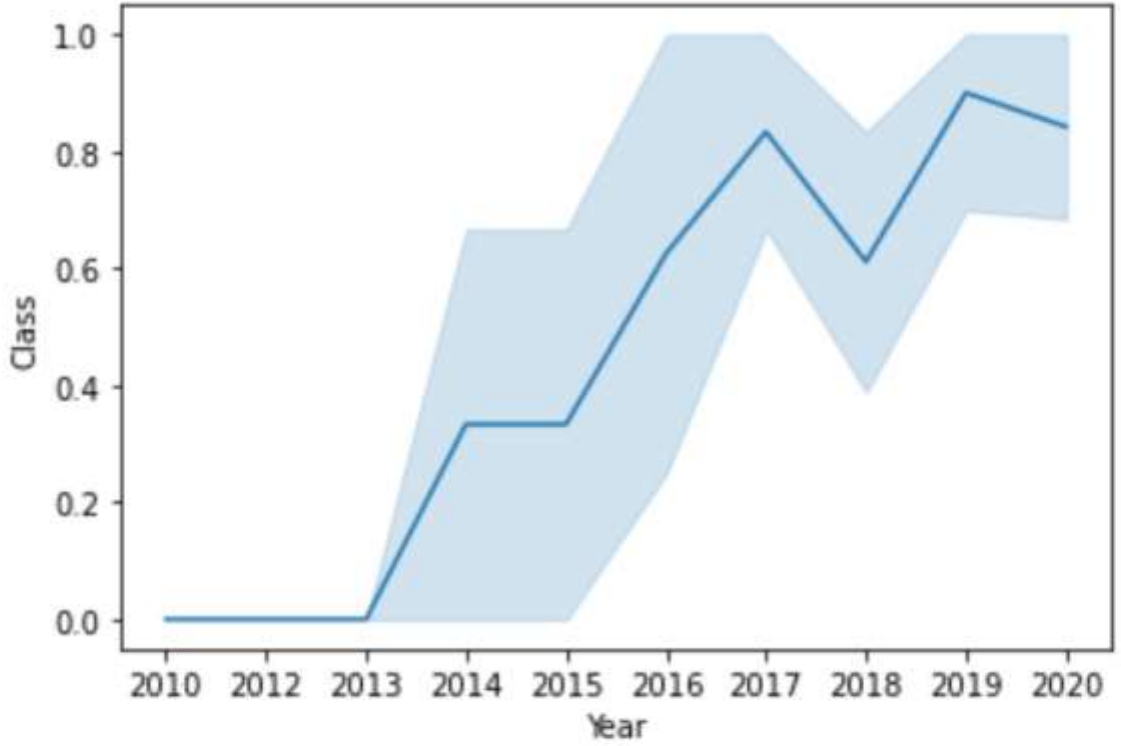
on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



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

Launch Success Yearly Trend



The sucess rate since 2013 kept increasing till 2020

All Launch Site Names

- Use the DISTINCT to get unique launch sites names.

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

In [5]:
```
%%sql
SELECT UNIQUE(LAUNCH_SITE) FROM SPACEXTBL
```

```
 * ibm_db_sa://zvc38333:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/BLUDB
Done.
```

Out[5]:

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

Launch Site Names Begin with 'CCA'

- LIKE is used to filter with beginning of ‘CCA’ with % symbol as wildcard.
- LIMIT 5 to show only 5 records.

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

In [10]:
```
%%sql
SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
```

```
 * ibm_db_sa://zvc38333:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/BLUDB
Done.
```

Out[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	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

- SUM allows us to add all the payload masses into one single value as a total.

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

In [17]:
```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLPOAD_MASK FROM SPACEXTBL
WHERE CUSTOMER = 'NASA (CRS)'
```

```
 * ibm_db_sa://zvc38333:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/BLUDB
Done.
```

Out[17]:

total_paylpoad_mask
45596

Average Payload Mass by F9 v1.1

- AVG is used to compute average of payload mass.
- WHERE to filter the booster version to F9 v1.1.

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

In [22]:
```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLPOAD_MASK FROM SPACEXTBL
WHERE BOOSTER_VERSION = 'F9 v1.1'
```

```
 * ibm_db_sa://zvc38333:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/BLUDB
Done.
```

Out[22]:

avg_paylpoad_mask
2928

First Successful Ground Landing Date

- Look for MIN(DATE) where landing outcome is Success.

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

Hint:Use min function

In [32]:
```
%%sql
SELECT MIN(DATE) FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Success (ground pad)'
```

```
 * ibm_db_sa://zvc38333:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/BLUDB
Done.
```

Out[32]:

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Payload filtering with ‘BETWEEN’ and ‘AND’.
- Mission and landing outcomes have been successful and filtered to show only drone ship.

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [36]:
```
%%sql
SELECT UNIQUE(BOOSTER_VERSION) FROM SPACEXTBL
WHERE (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000)
AND LANDING__OUTCOME = 'Success (drone ship)'
```

```
 * ibm_db_sa://zvc38333:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/BLUDB
Done.
```

Out[36]:

booster_version
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1022
F9 FT B1026

Total Number of Successful and Failure Mission Outcomes

- GROUP BY allows to count by different outcomes

List the total number of successful and failure mission outcomes

In [46]:
```
%%sql
SELECT COUNT(MISSION_OUTCOME) AS VALUE_COUNT,MISSION_OUTCOME FROM SPACEXTBL GROUP BY MISSION_OUTCOME
```

```
 * ibm_db_sa://zvc38333:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/BLUDB
Done.
```

Out[46]:

value_count	mission_outcome
1	Failure (in flight)
99	Success
1	Success (payload status unclear)

Boosters Carried Maximum Payload

- The subquery in WHERE allows to filter with maximum payload.

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

In [52]:
```
%%sql
SELECT UNIQUE(BOOSTER_VERSION) FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

```
 * ibm_db_sa://zvc38333:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/BLUDB
Done.
```

Out[52]:

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

- Retrieve the year from a date using YEAR function.
- Filtered by the desired failed drone ship.

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

In [63]:
```
%%sql
SELECT DATE,LANDING__OUTCOME,BOOSTER_VERSION,LAUNCH_SITE FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND YEAR(DATE) = 2015
```

```
 * ibm_db_sa://zvc38333:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/BLUDB
Done.
```

Out[63]:

DATE	landing__outcome	booster_version	launch_site
2015-01-10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- GROUP BY allows to count by landing outcome.
- ORDER BY DESC with count allows to rank from higher to lower.

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

In [69]:
```
%%sql
SELECT COUNT(LANDING__OUTCOME) AS VALUE_COUNT,LANDING__OUTCOME FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING__OUTCOME
ORDER BY VALUE_COUNT DESC
```

```
 * ibm_db_sa://zvc38333:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/BLUDB
Done.
```

Out[69]:

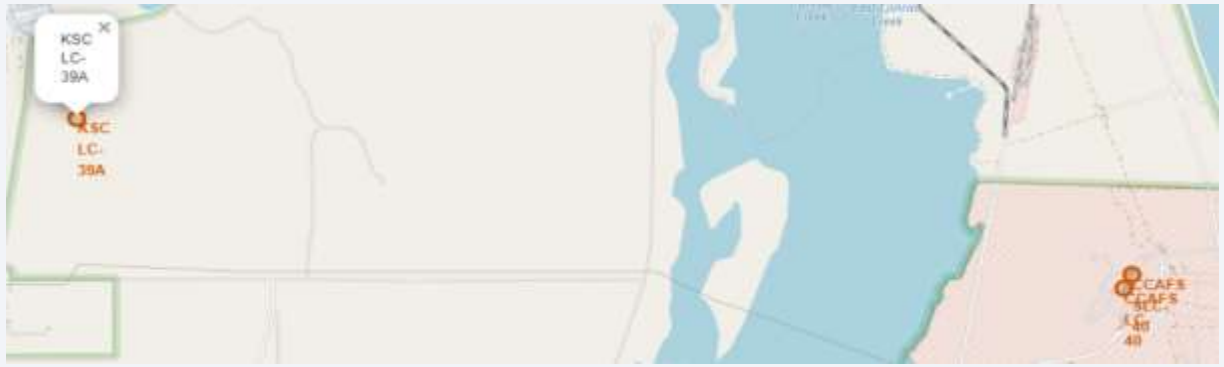
value_count	landing__outcome
10	No attempt
5	Failure (drone ship)
5	Success (drone ship)
3	Controlled (ocean)
3	Success (ground pad)
2	Failure (parachute)
2	Uncontrolled (ocean)
1	Precluded (drone ship)

Section 4

Launch Sites Proximities Analysis

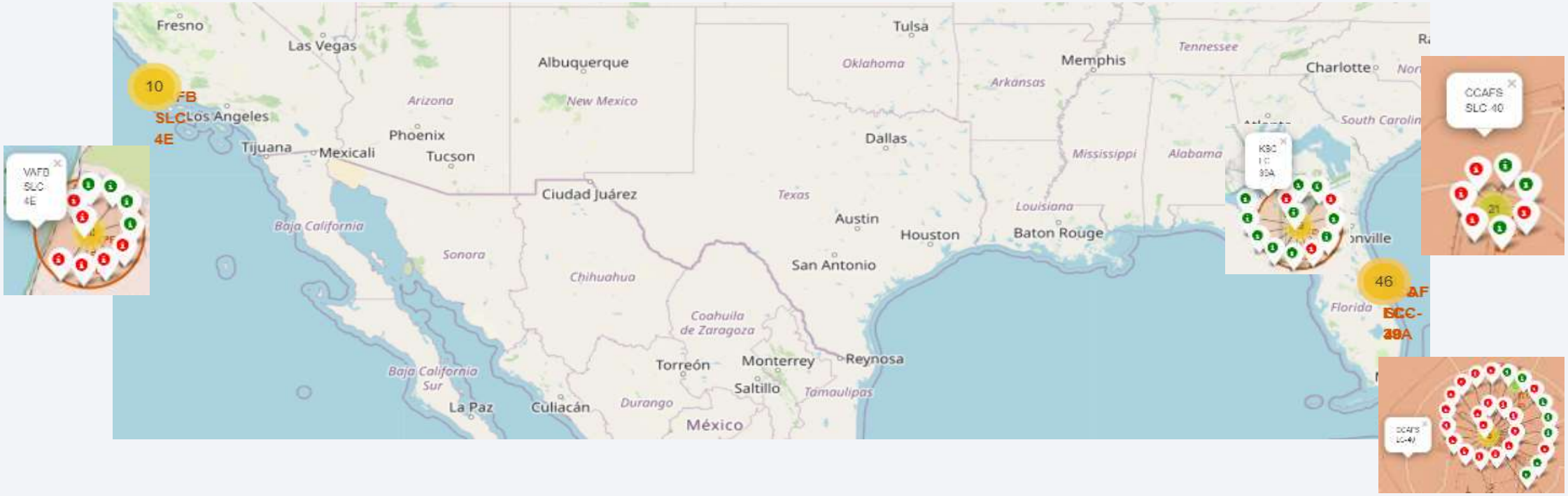
Folium Map – Launch locations

- Explore the different launch locations.
- They are all close to the ocean and the equator line.



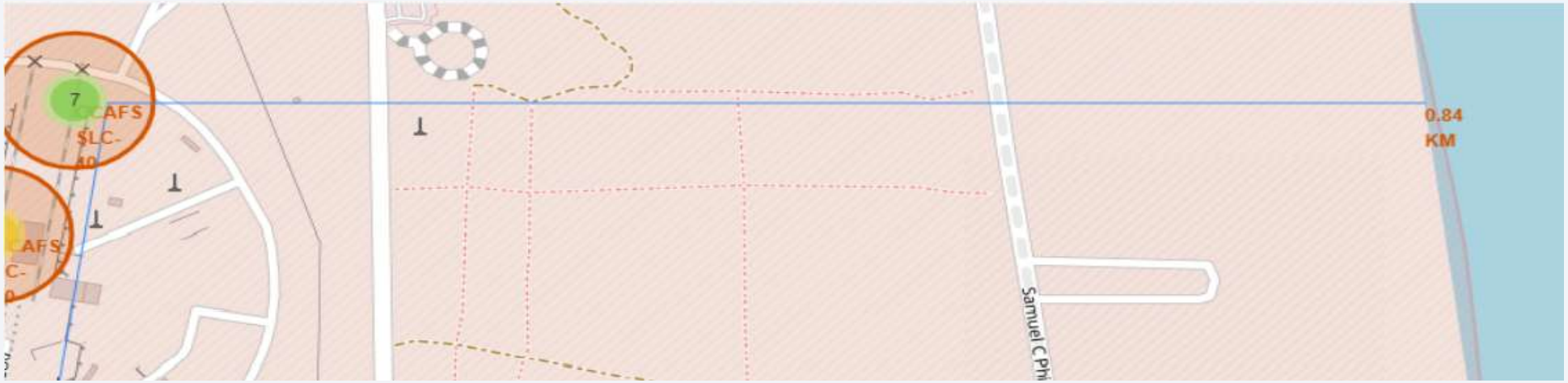
Folium Map – Success and Failure locations

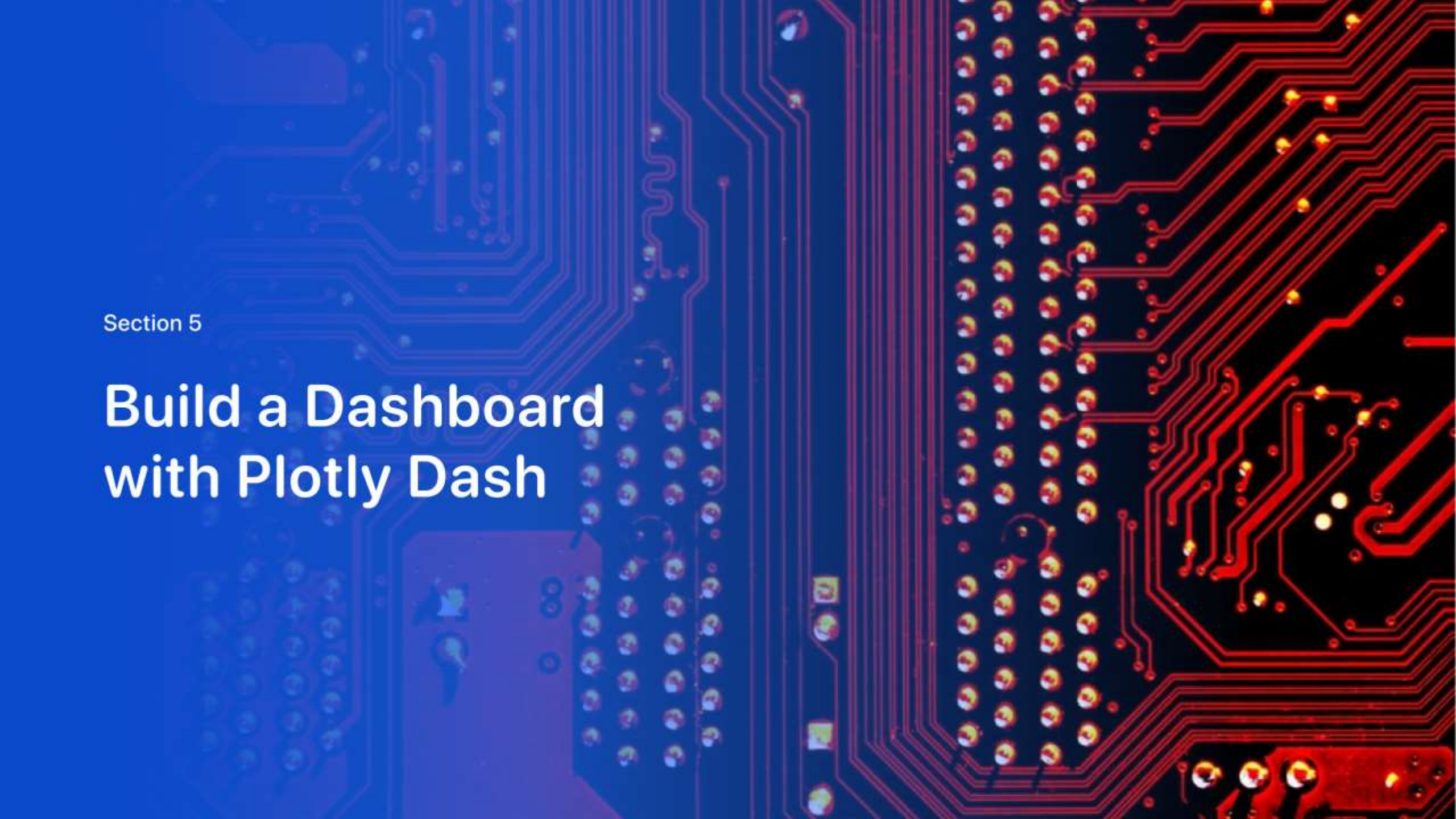
- Most locations have mixed success rate except for one shown in the figure named CCAFS SLC-40.



Folium Map – Distance to proximities

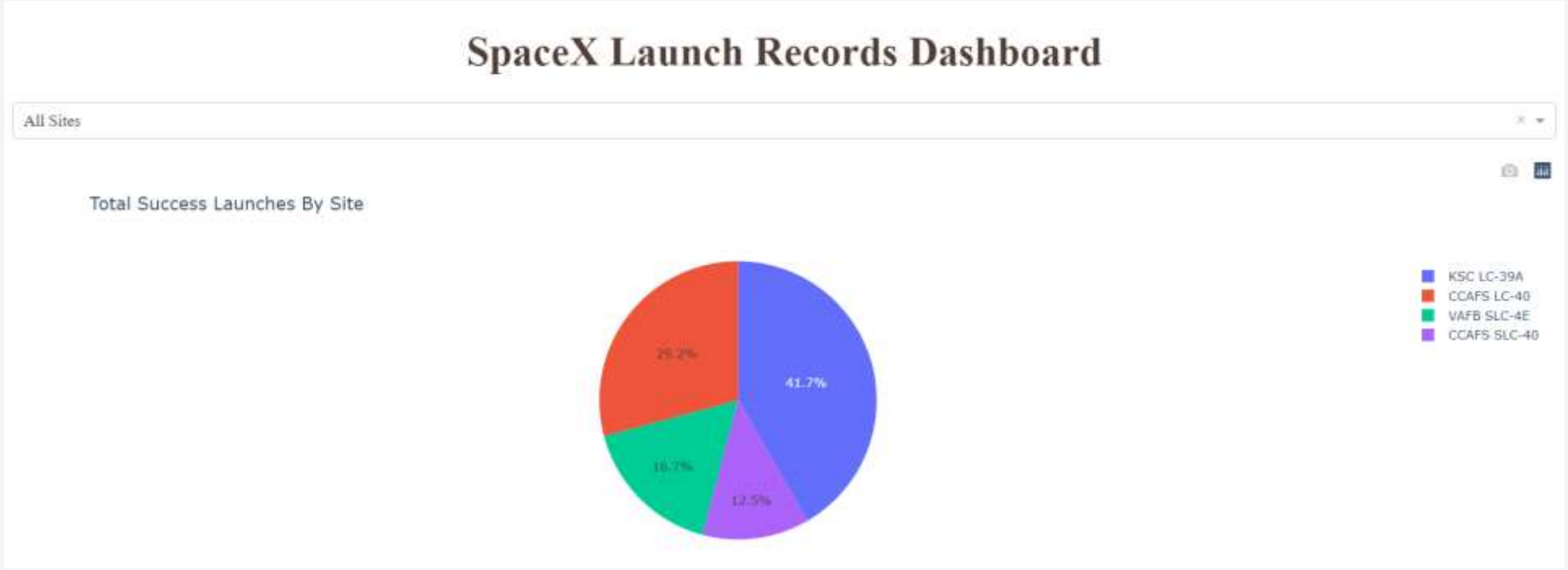
- Launch sites are close to railways but not from highways or coastlines. This is due to safety reasons, as if something goes wrong and an explosion occurs, no one or nothing gets damaged.
- Also, in case of trajectory deviation, its less likely to impact on inhabited areas.




Section 5
Build a Dashboard with Plotly Dash

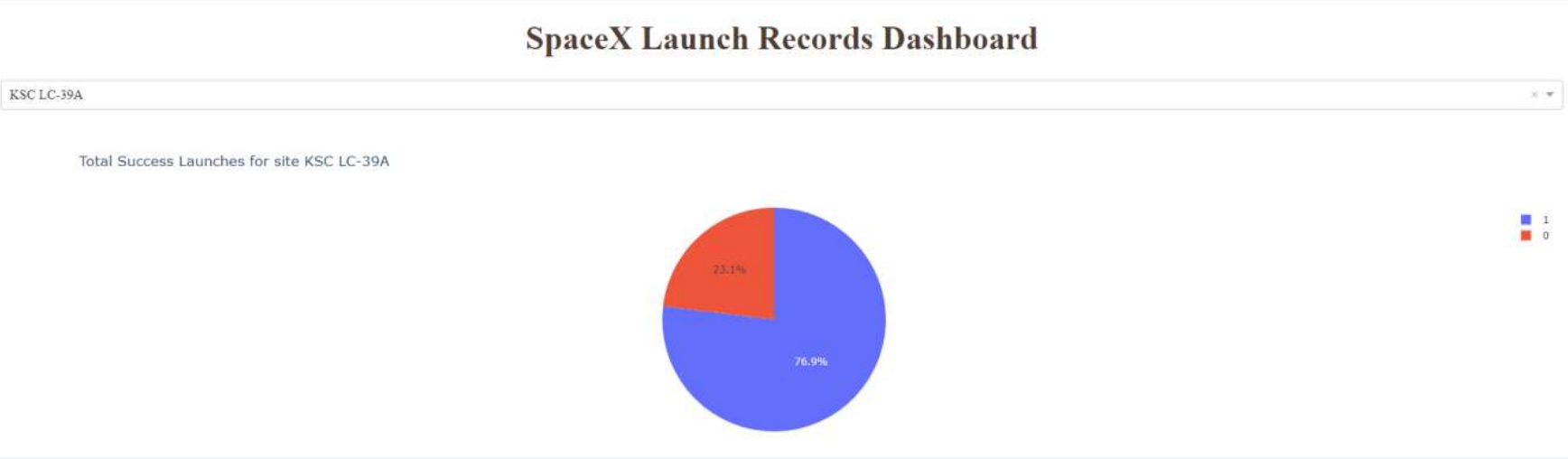
Dashboard – Success ratio of all sites

- KSC LC-39A shows the highest success of all locations.
- CCAFS SLC-40 shows the lowest success of all locations.



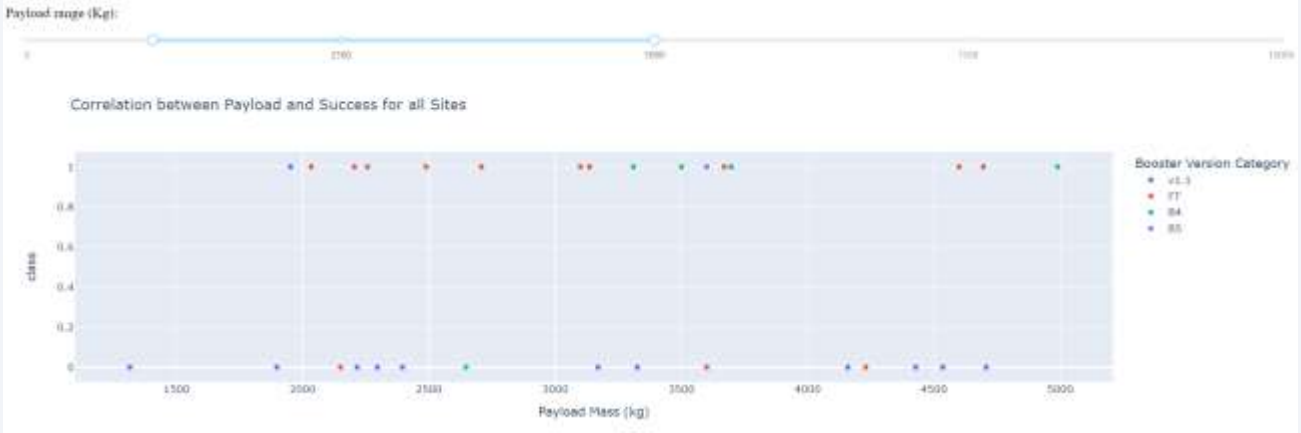
Dashboard – KSC LC-39A success rate

- KSC LC-39A is the highest contributor to success rate of all locations. It has 76.9% success launch rate.



Dashboard – Payload and Booster Version

- B4 achieves the highest payload mass.
- FT has the most launches, with medium values.
- V1.0 and v1.1 show low and mid payload masses.

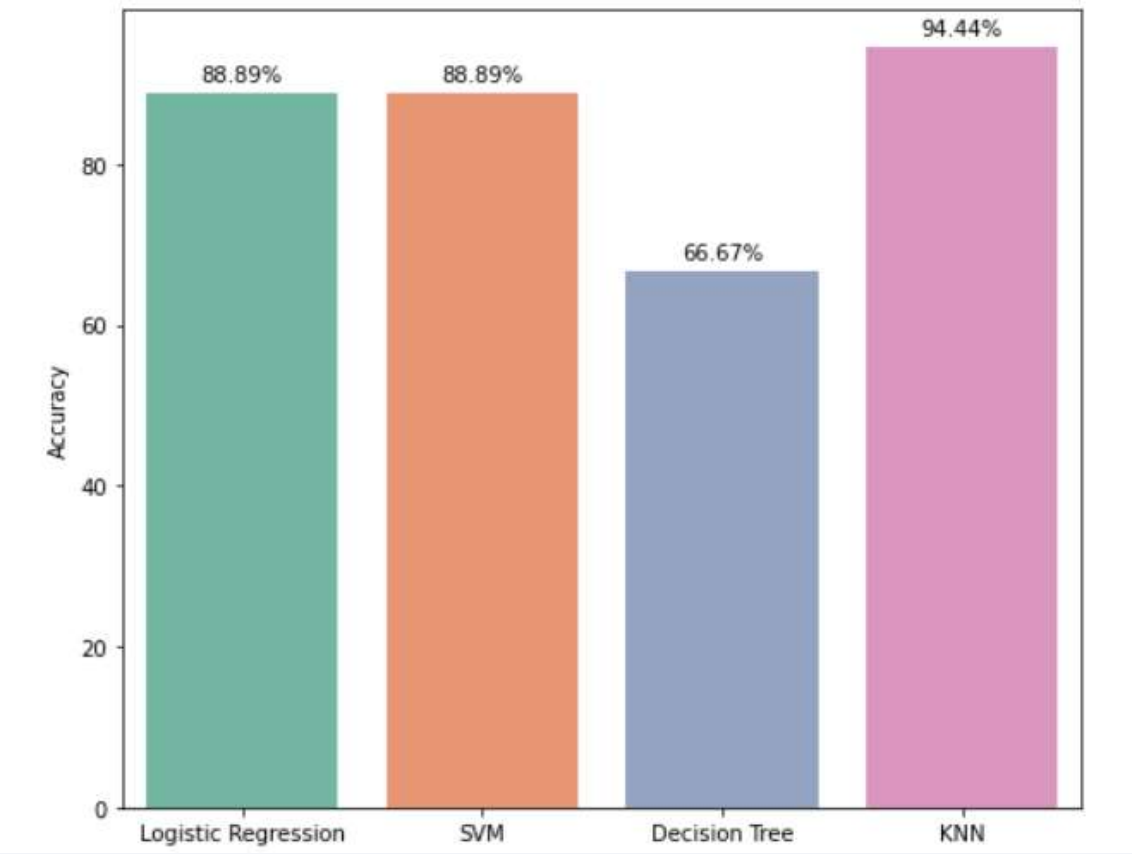


Section 6

Predictive Analysis (Classification)

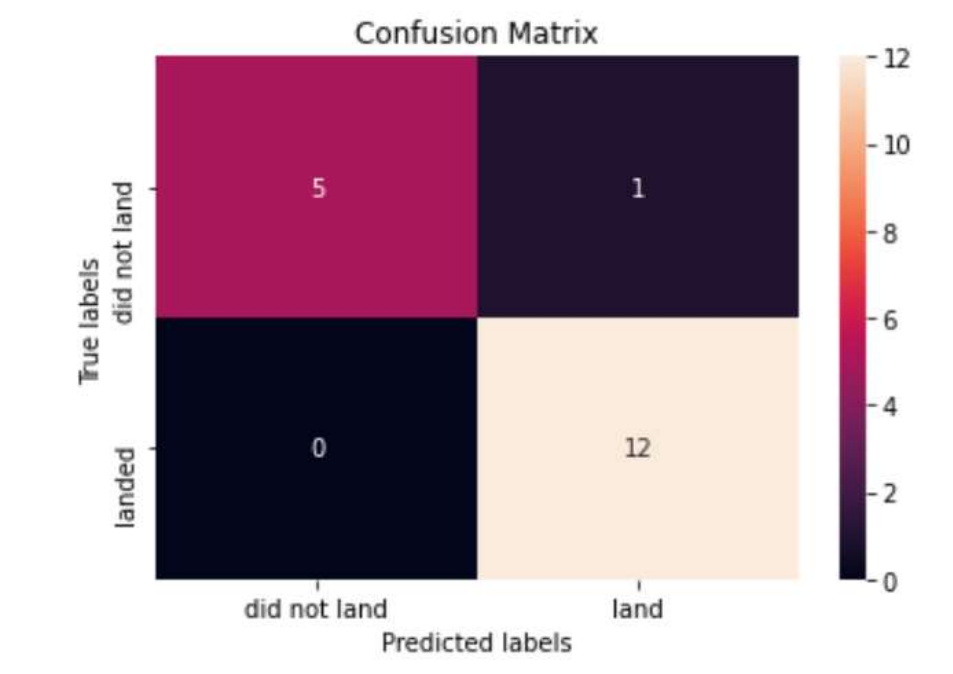
Classification Accuracy

- Logistic Regression and SVM have the same accuracy score that they are very good models, but the most trustworthy models is KNN with the accuracy score – 94.44%



Confusion Matrix

- The models is almost perfect in performance in both predicting success landing and failure landing.



Conclusions

- A SpaceX Falcon 9 dataset has been obtained thanks to data gathering and cleaning.
- Data can be collected from different sources such as: API calls, SQL DBs and web scrapping.
- KSC LC-39A shows the highest success rate of all locations.
- The KNN model achieves the highest accuracy (94.44%) when predict landing outcomes.

Appendix

- Full github repository: https://github.com/123olala/IBM-Data-Science-Professional-Certificate-Applied-Data-Science-Capstone
- This project is a part of the Applied Data Science Capstone from the IBM Data Science professional certificate. (https://www.coursera.org/professional-certificates/ibm-data-science)

Thank you!