IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Emmanuel Jacob
November 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection using web scraping and SpaceX API
  - Exploratory Data Analysis
  - Data Wrangling
  - Data Visualisation
  - Interactive visual analytics
  - Machine learning prediction
- Summary of all results
  - Exploratory Data Analysis Results
  - Predictive analysis results

Introduction

- The objective is to evaluate the viability of the new company SpaceY to compete with SpaceX
- Question to answer
  - How do variables such as payload mass, launch site, number of flights and orbits affect the success of the first stage landing?
  - Does the rate of successful landing increase over time?
  - What is the best algorithm that can be used for binary classification in this case?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX Rest API
  - Using Web Scrapping from Wikipedia
- Perform data wrangling
  - Filtering the data
  - Dealing with missing values
  - Using One Hot Encoding to prepare the data to a binary classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building, tuning and evaluation of classification models to ensure the best results

Data Collection

- SpaceX Rest API
  - https://api.spacexdata.com/v4/rockets/
- Web Scraping
  - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Data Collection – SpaceX API

- SpaceX offers a public API from where data can be obtained

- Source code: https://github.com/EmJacob/IBM-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

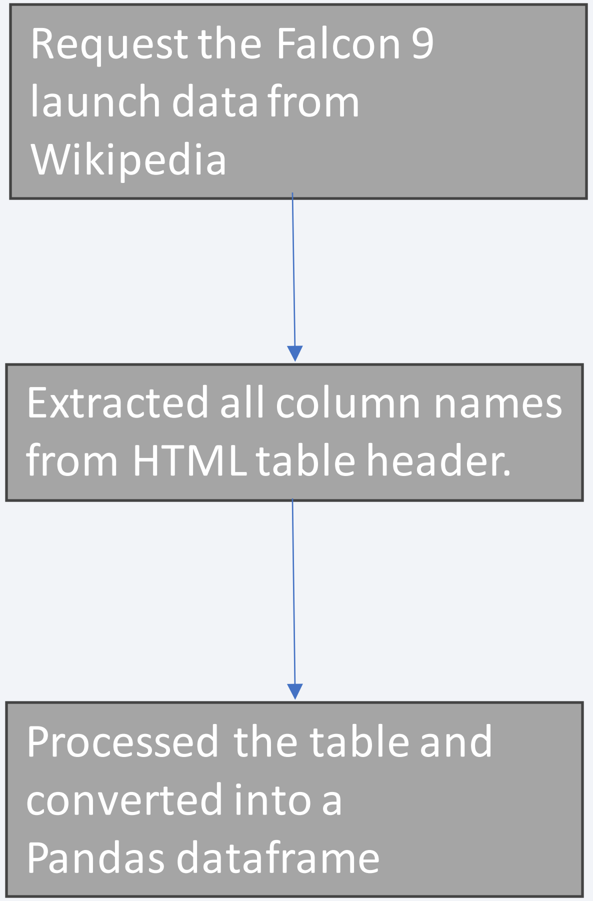
Request rocket launch data form SpaceX REST API

Filter data to only include Falcon 9 launches

Replace Missing data values

Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts
- Add the GitHub URL of the completed web scraping notebook, as an external reference and peer-review purpose
- Source code: https://github.com/EmJacob/IBM-Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Exploratory Data Analysis (EDA) was performed on the dataset.
- Summaries launches per site and amount of each orbits and mission outcomes per orbit type.
- Created landing outcomes column with labels with 1 or 0, with 1 meaning first stage landed successfully and 0 meaning it was unsuccessful.
- Source code: https://github.com/EmJacob/IBM-Data-Science-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Exploratory Data Analysis with Visualisation
  - Move data into a Pandas dataframe
  - Used matplotlib and Seaborn to Visualise data
    - Flight number to Payload mass
    - Flight number to Launch site
    - Payload to Launch site
    - Orbit type to Success rate
    - Flight number to Orbit type
    - Payload to Orbit type
    - Year to Success rate
- Source code: https://github.com/EmJacob/IBM-Data-Science-Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- Performed SQL queries:
  - Display the 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 acheived.
  - 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. Use a subquery
  - List the records which will display 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 (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.
- Source code: https://github.com/EmJacob/IBM-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

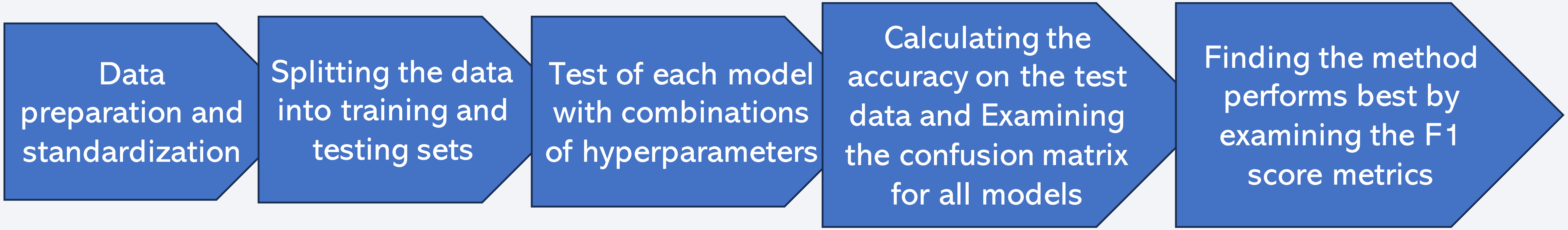
Build an Interactive Map with Folium

- Added Marker with Circle, Popup Label and Text Label to all the launch sites using its latitude and longitude coordinates to show their geographical locations.
- Added markers for success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.
- Added lines to show distances between the Launch Site CCAFS SLC-40 and its proximities like Railway, Coastline and Closest City.
- Source code: https://github.com/EmJacob/IBM-Data-Science-Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- Added a dropdown list to enable Launch Site selection.
- Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for selected site.
- Added a slider to select Payload range.
- Added a scatter chart to show the correlation between Payload and Launch Success.
- Source code: https://github.com/EmJacob/IBM-Data-Science-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)



- Source code: https://github.com/EmJacob/IBM-Data-Science-Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

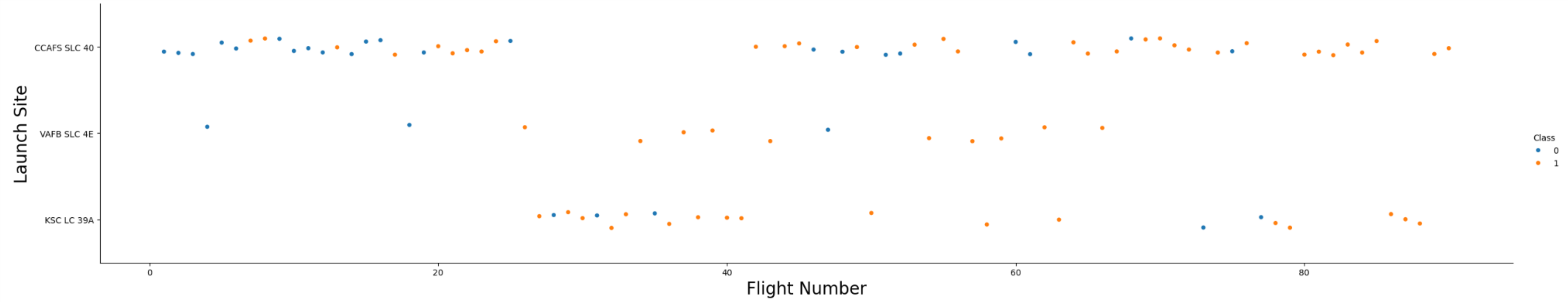
Results

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

Section 2

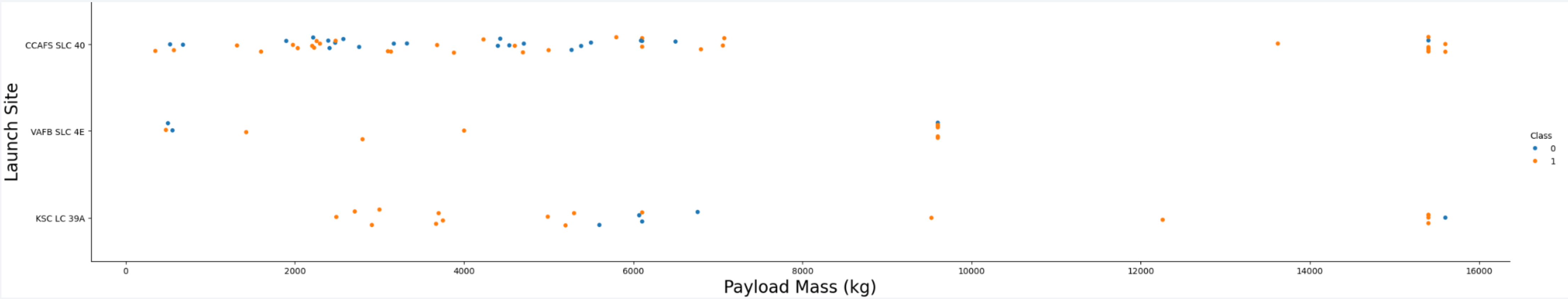
Insights drawn from EDA

Flight Number vs. Launch Site



- Flight number to Launch site shows that most of the early launches was done at CCAF5 SLC-40

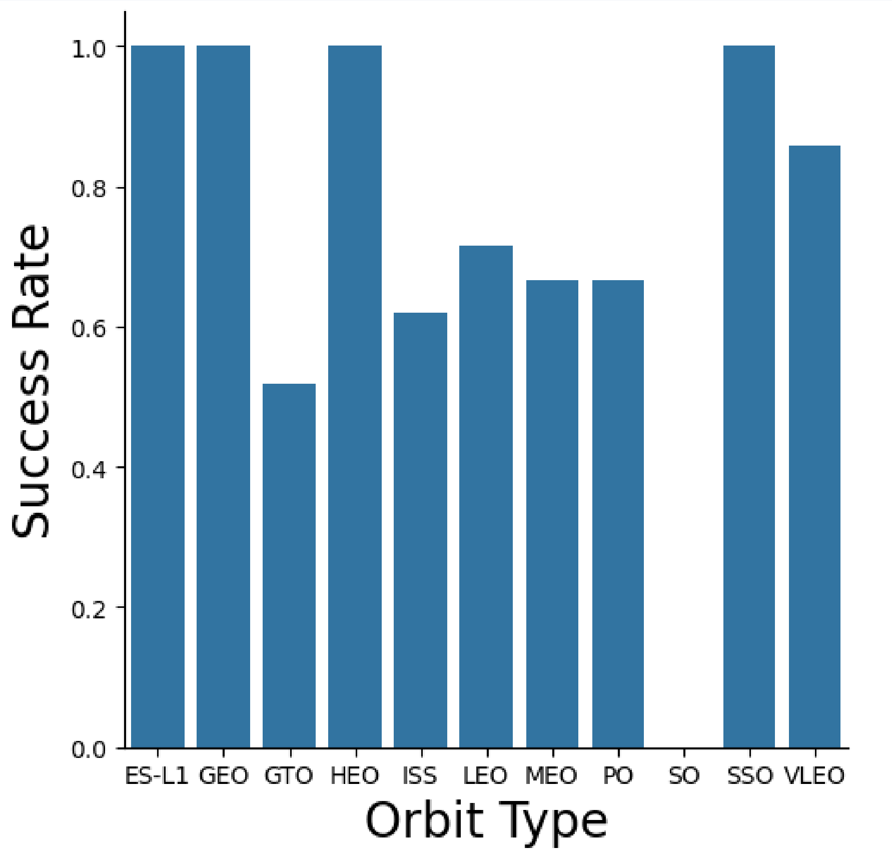
Payload vs. Launch Site



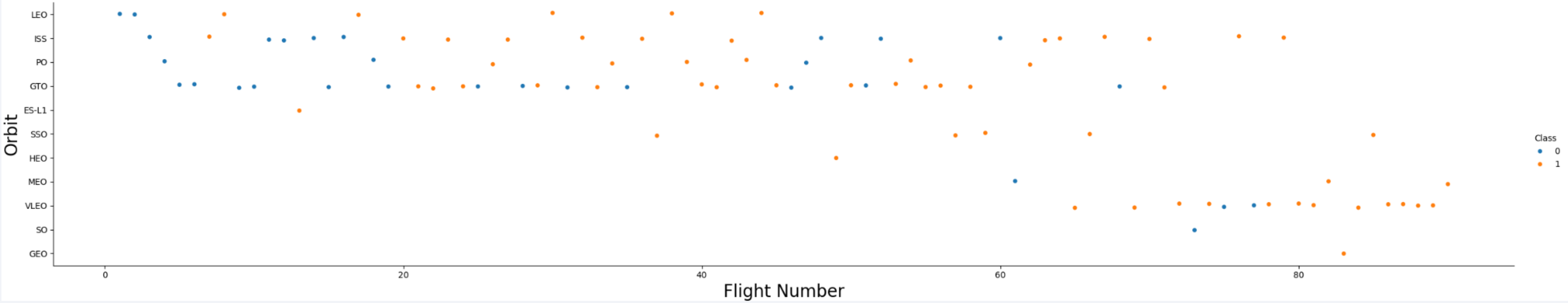
- Payload mass to Launch site. CCAF SLC-40 and KSC LC-39A are the only sites used for the heavy payloads

Success Rate vs. Orbit Type

- Orbit type to Success rate. All orbits except for SO have had successful first stage landings.
- Orbits with 100% success rate are ES-L1, GEO, HEO, SSO



Flight Number vs. Orbit Type



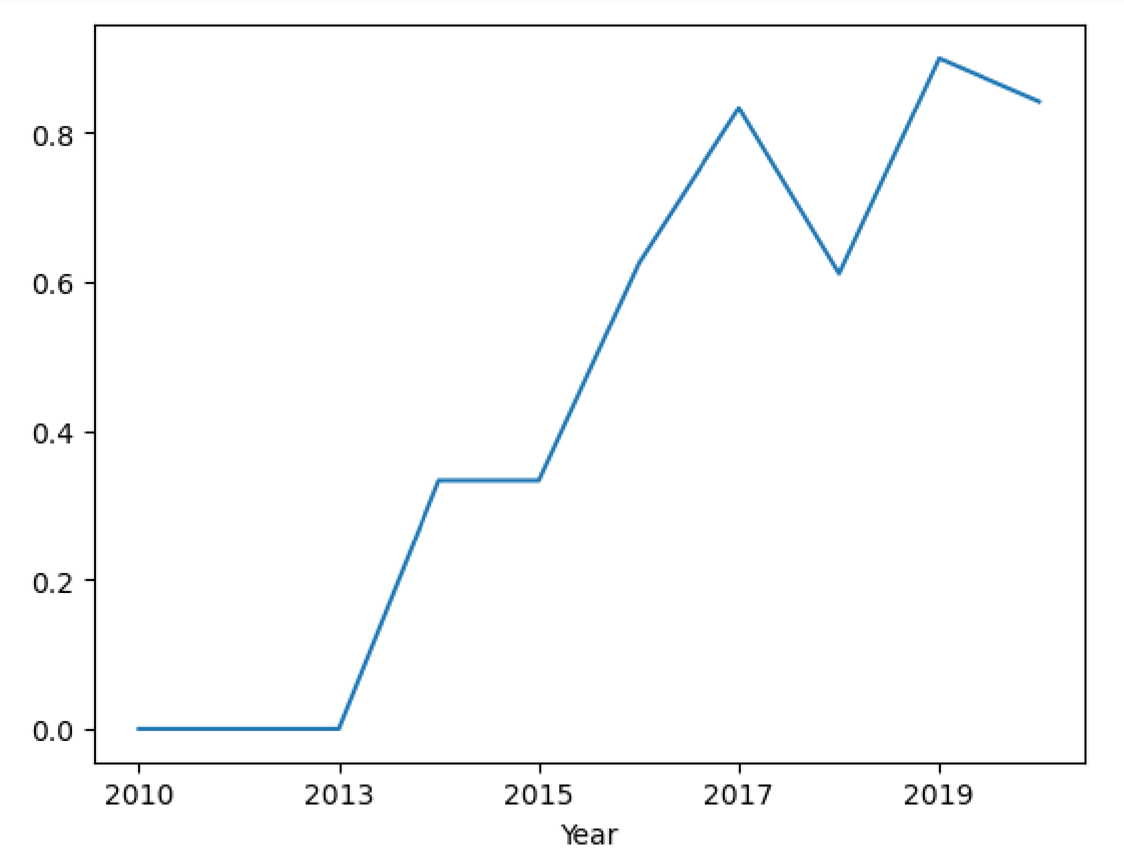
- Success rate improved over time to all orbits.

Payload vs. Orbit Type

- No relation between payload and success rate GTO
- ISS has the widest range of payload

Launch Success Yearly Trend

- Success rate started increasing in 2013.



All Launch Site Names

- All launch sites

```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
```

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

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with the string 'CCA'.

```
%sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

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

Total Payload Mass

- The total payload mass carried by boosters launched by NASA (CRS).

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE PAYLOAD  LIKE '%CRS%';
```

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

TOTAL_PAYLOAD
111268

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1.

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
```

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

AVG_PAYLOAD
2928.4

First Successful Ground Landing Date

- Date when the first successful landing outcome in ground pad was achieved

```
%sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (ground pad)';
```

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

FIRST_SUCCESS_GP
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.

```
%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND LANDING_OUTCOME = 'Suc
```

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

- Total number of successful and failure mission outcomes.

```
sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;
```

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

Mission_Outcome	QTY
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Names of the booster versions which have carried the maximum payload mass.

```
sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

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

- Failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015.

```
%%sql select substr(Date, 6,2) as month, date, booster_version, launch_site, landing_outcome from SPACEXTBL
      where landing_outcome = 'Failure (drone ship)' and substr(Date,0,5)='2015';
```

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

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- Ranking 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_outcomes from SPACEXTBL
      where date between '2010-06-04' and '2017-03-20'
      group by landing_outcome
      order by count_outcomes desc;
```

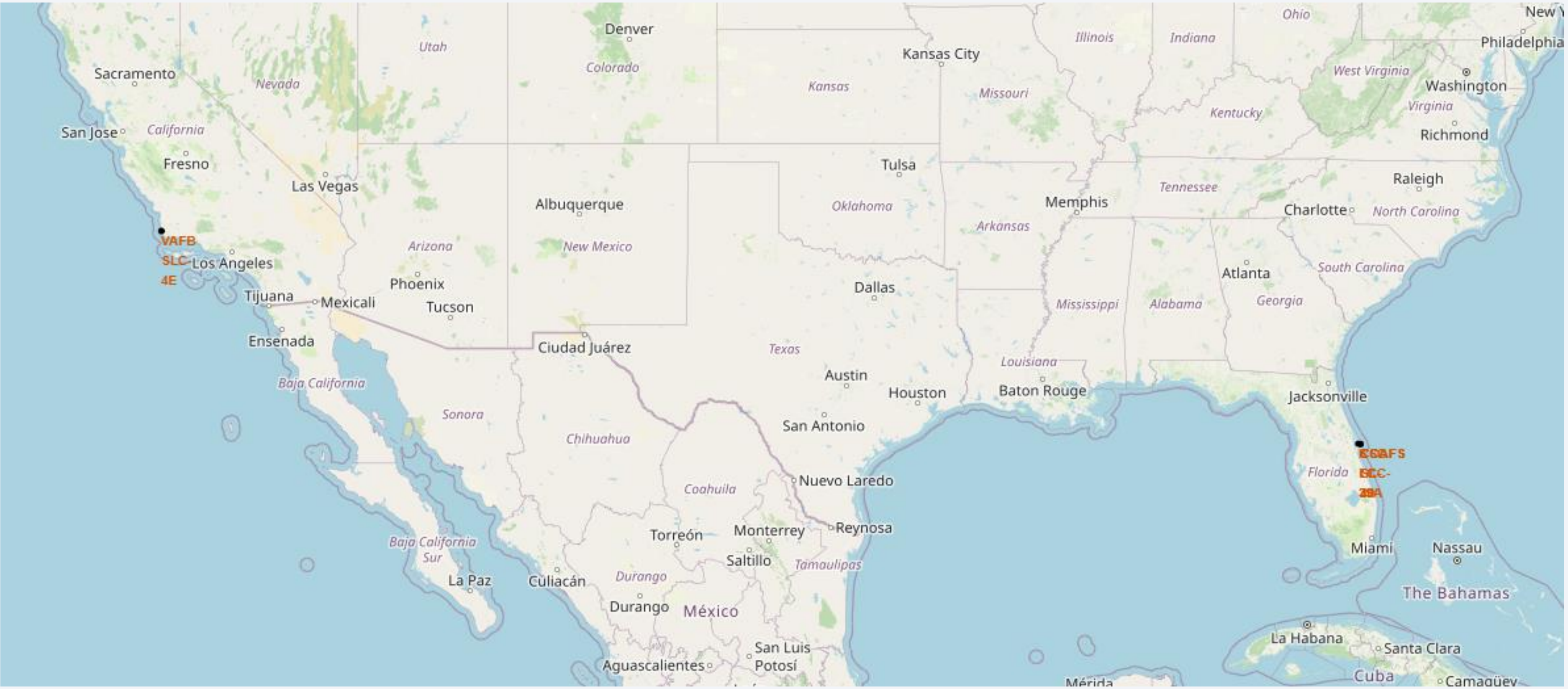
```
* sqlite:///my_data1.db
Done.
```

Landing_Outcome	count_outcomes
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

Section 3

Launch Sites Proximities Analysis

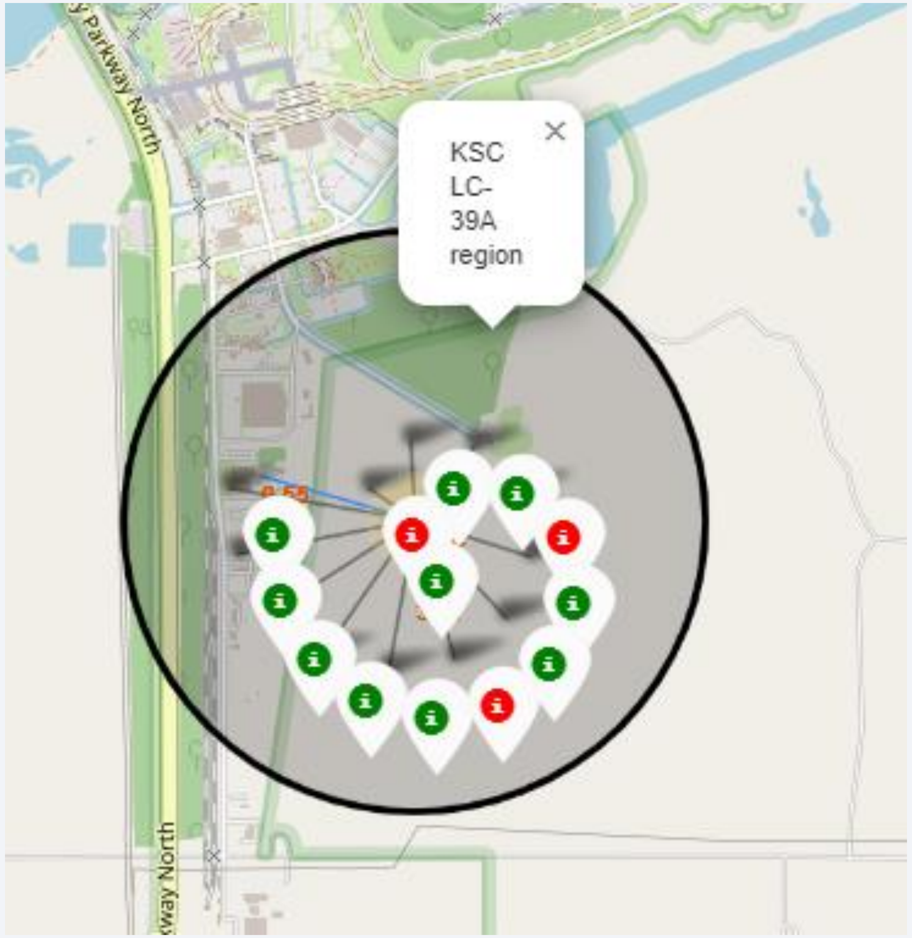
All launch sites



- Launch sites are near sea, possible for safety, but not too far from roads and railroads.

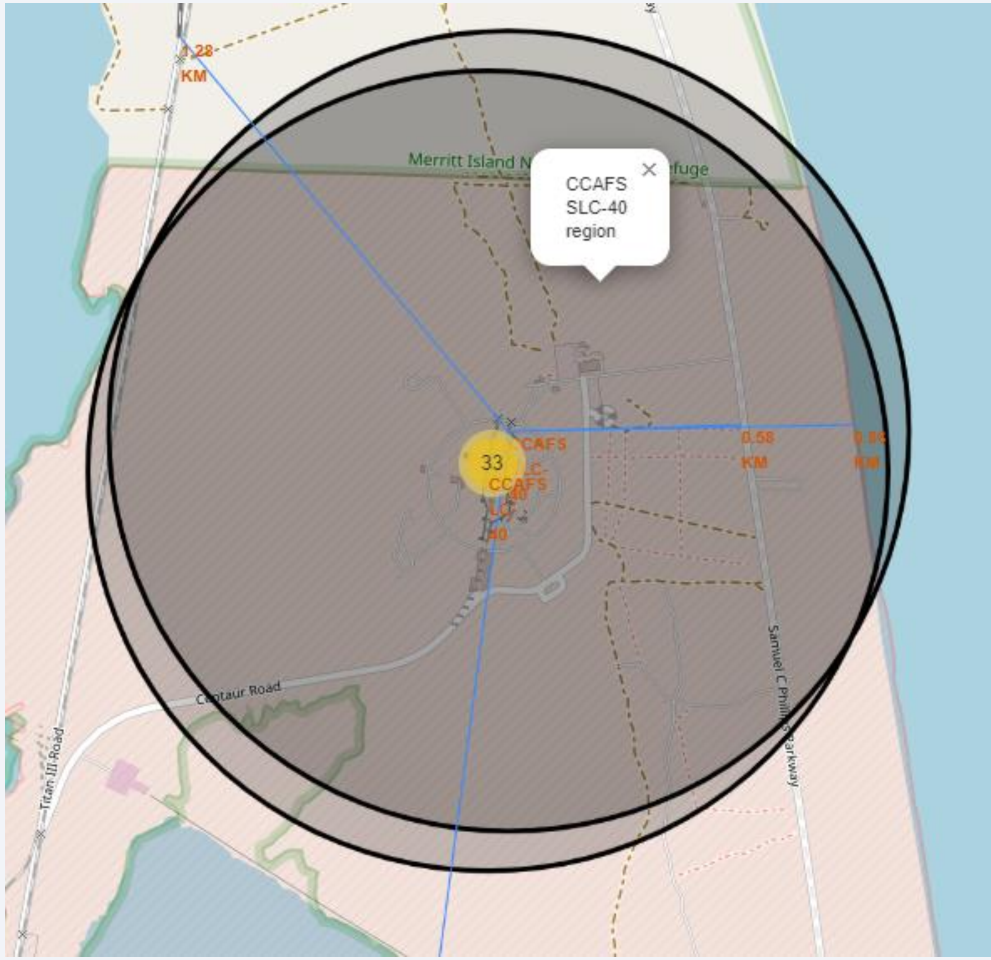
Launch Outcomes by Site

- KSC LC-39A launch site launch outcomes
- Green markers indicate successful and red markers indicate failure.



<Folium Map Screenshot 3>

- Launch site CCAFS SLC-40 we can clearly see the lines to the nearest coast, railway, and city

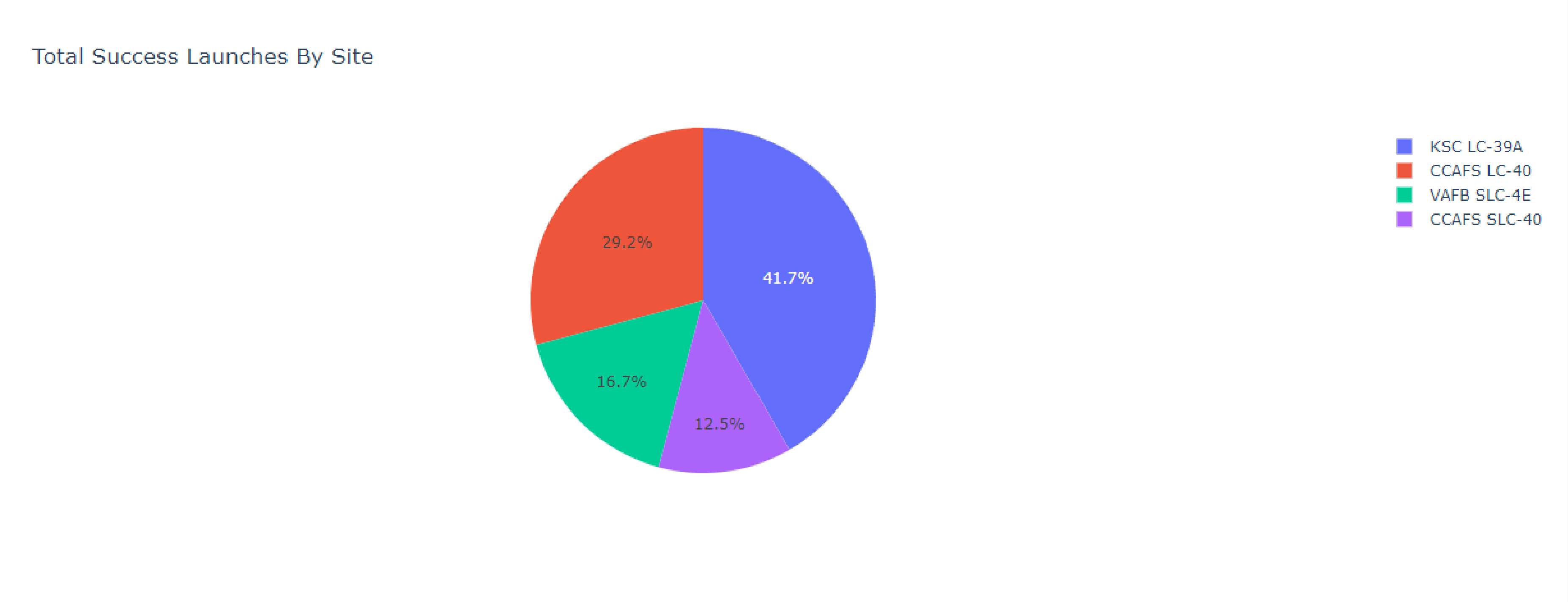


Section 4

Build a Dashboard with Plotly Dash

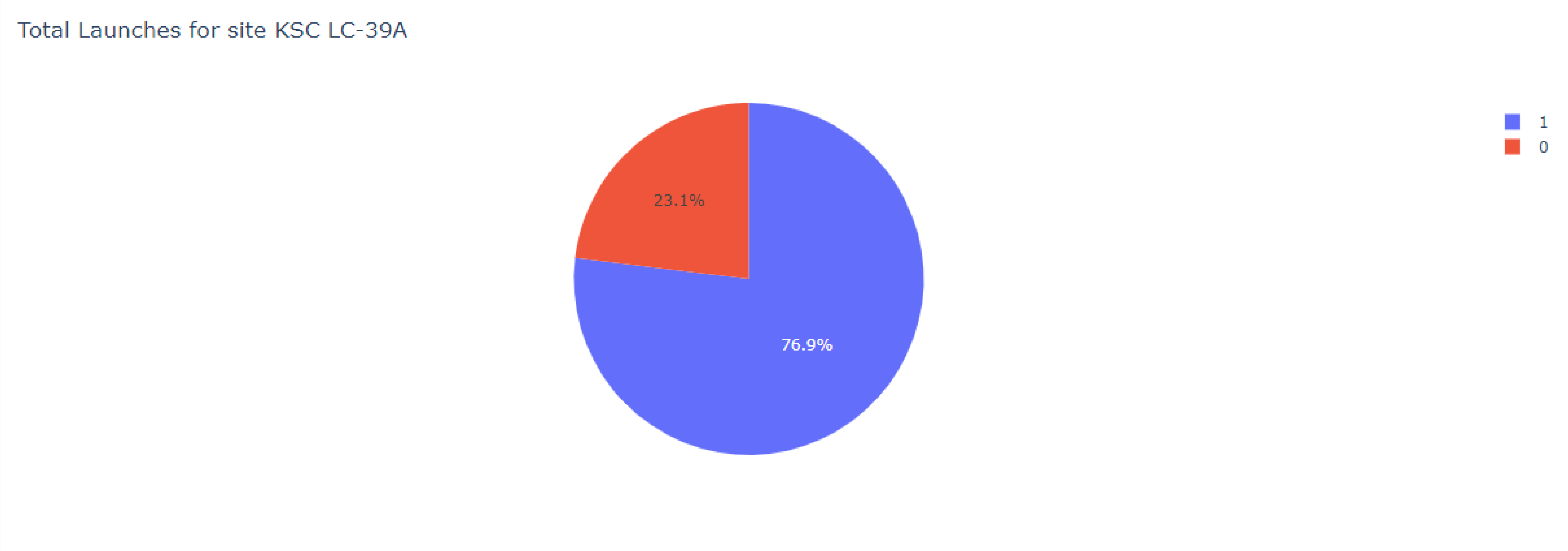
Launch success count for all sites

- The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches.



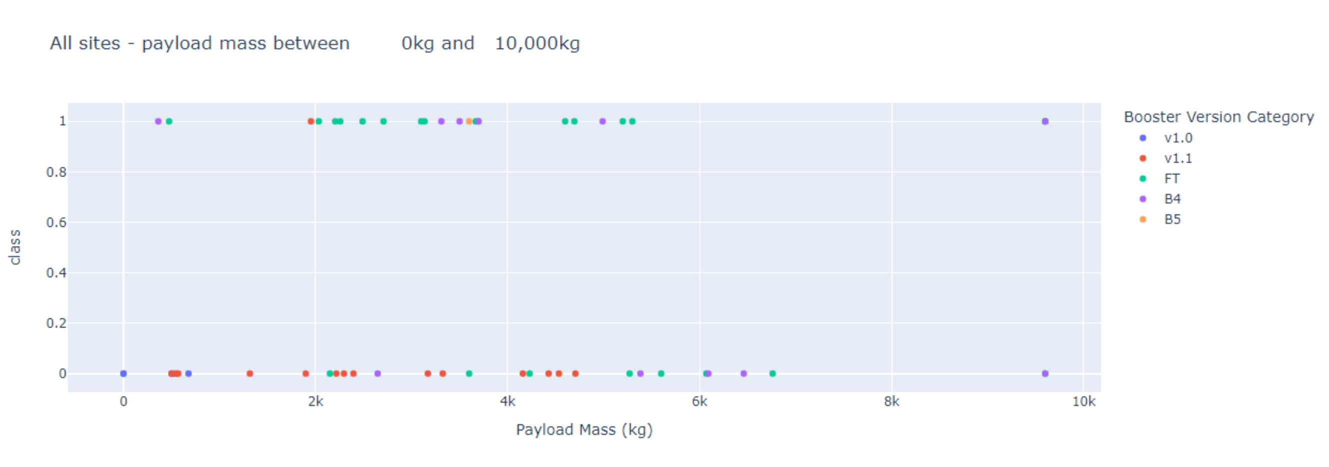
<Dashboard Screenshot 2>

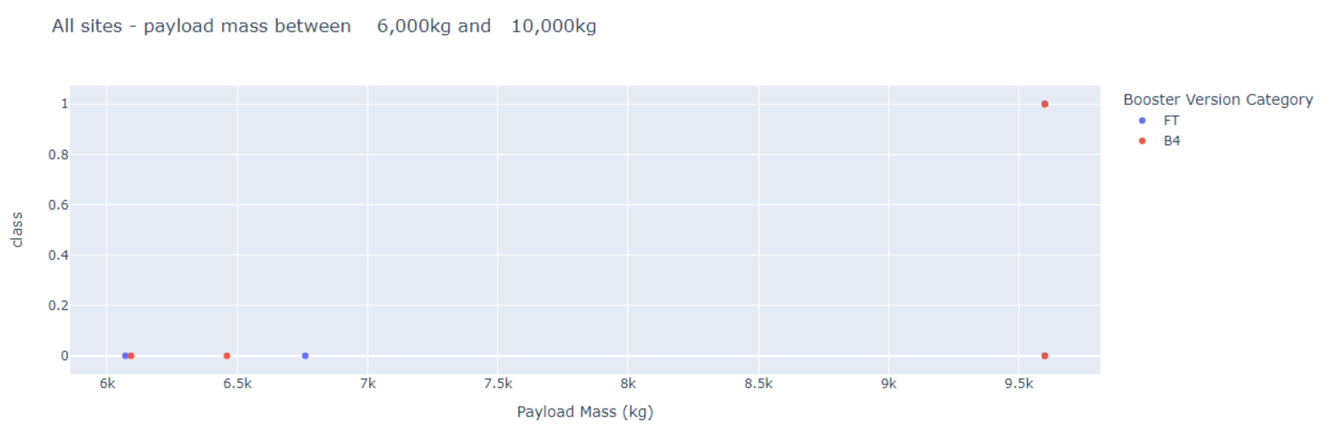
- KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.



<Dashboard Screenshot 3>

- Payloads under 6,000kg and FT boosters are the most successful combination.
- Payloads over 6,000kg is most likely to fail.



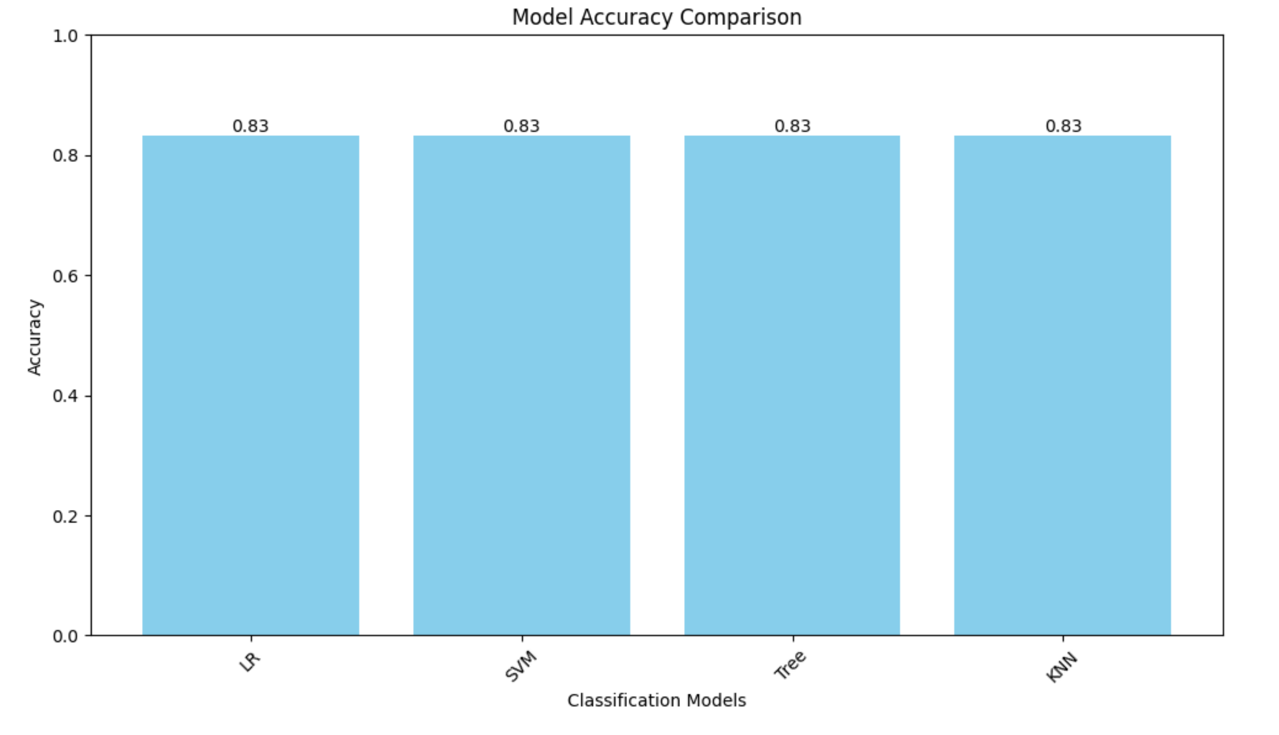


Section 5

Predictive Analysis (Classification)

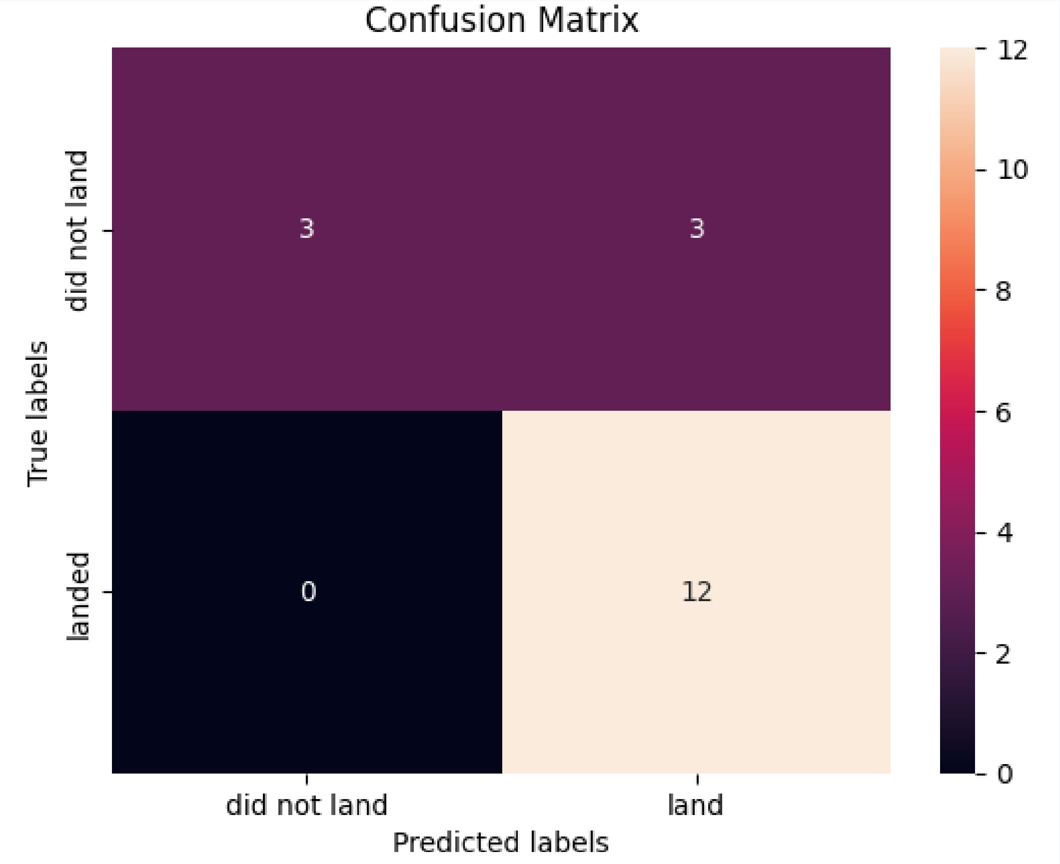
Classification Accuracy

- Used four machine learning models:
  - Logistic Regression
  - SVM
  - Decision Tree
  - KNN.
- All models demonstrated an identical accuracy of 83.33% on the test data.



Confusion Matrix

- The major problem is false positives as evidenced by the models incorrectly predicting the 1st stage booster to land in 3 out of 18 samples in the test set



Conclusions

- Launches with under 6000kg payload mass show better results than launches with a larger payload mass.
- Majority of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast and railway.
- The success rate of launches increases over the years since 2013.
- KSC LC-39A has the highest success rate of the launches from all the sites.

Appendix

- Git hub : https://github.com/EmJacob/IBM-Data-Science-Capstone/tree/main
- Acknowledgments
  - Thank you to Joseph Santarcangelo and Yan Luo

Thank you!