IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Fausto Imamura
06/12/2024

Outline

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

Executive Summary

- This study aims to perform exploratory and predictive data analysis for a successful rocket landing, using the following methodologies:
  - Data collection using SpaceX REST API and webscrapping with BeautifulSoup packages
  - Data wrangling to use as dependent variables for Machine Learning Models
  - SQL queries and Data Visualization, building interactive Follium Maps and Plotly Dashboard
  - Machine Learning Model building and accuracy evaluation (logistic regression, SVM, decision tree and KNN)
- Results
  - SpaceX launch success rate has improved over the years
  - The success rate increases with the number of flights
  - The KSC LC-39A site presents the most successful launch rate (41,2%)
  - All Machine Learning Models performed well to predict successfull landing and SVM has the best results

Introduction

- Project background and context

  Space X is the most successful company that provides affordable space travels. It can do this because its rocket launches are relatively inexpensive – about 62 million dollars vs 165 million dollars from other providers, as it can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. Based on public information and training machine learning models, we will predict if SpaceX can reuse the first stage.

- Problems you want to find answers
  - How can variables such as payload mass, launch site, number of flights and orbits affect the success of first stage landing?
  - How does the rate of successful landings behave over time?
  - What is the best algorithm to predict successful landing (binary classification)?

Section 1

Methodology

Methodology

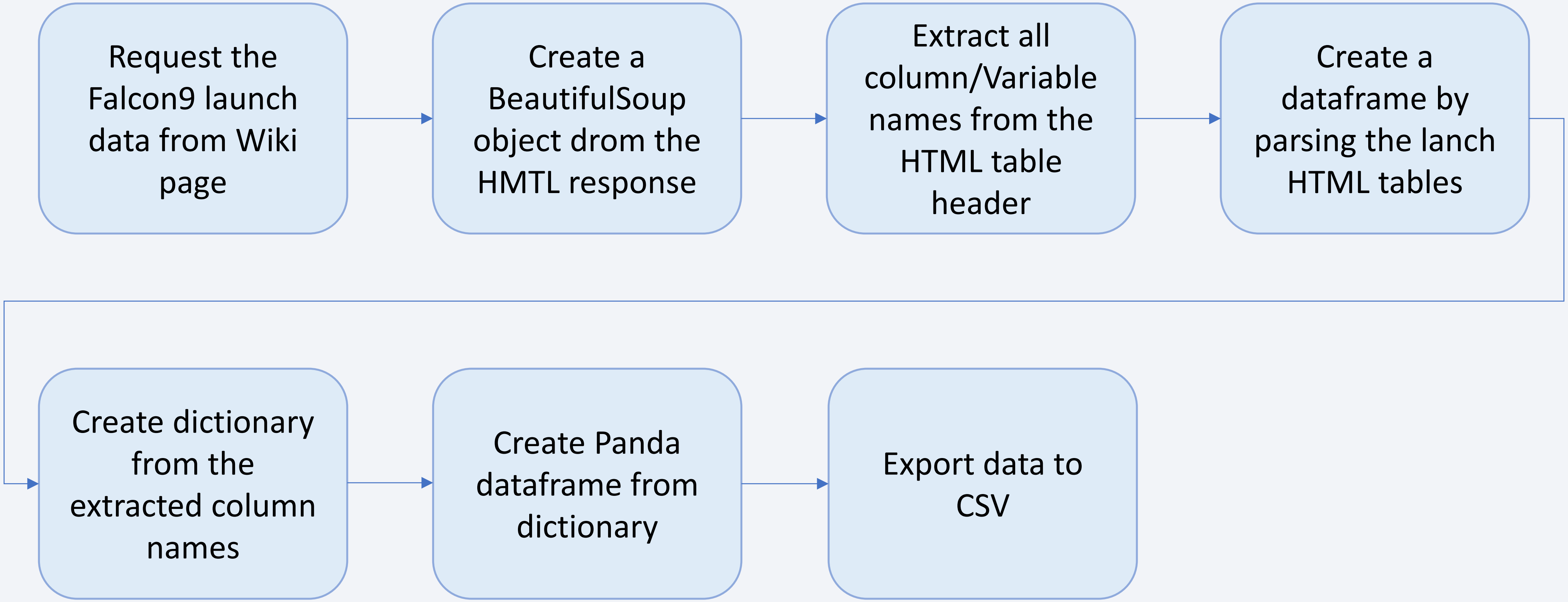
- Data collection methodology:
  - Data gathering with SpaceX REST API
  - Web scrapping HTML tables with BeautifulSoup package
- Performed data wrangling
  - Filtering the data, handling missing values and using one-hot encoding (to binary classification) for analysis and modelling
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - Standardizing data, splitting into training and test data and finding best hyperparameters to evaluate model

Data Collection – SpaceX API

1. Perform **GET request** to SpaceX API
2. Decode response as **.json** and tunrn it into a dataframe using **.json_normalize()**
3. Request launch information from API with custom functions
4. Build dataset with collected data and combine the columns into a dictionary
5. Create Panda dataframe from dictionary
6. Filter dataframe to Only include **Falcon 9** launches
7. Replace missing values of Payload Mass with calculated .mean()
8. Export data to CSV

GitHub URL: https://github.com/fimamu/IBM-Data-Science-SpaceX-Capstone/blob/ff22034a0f999997e9ea325bf8cccd40e6396b23/01-spacex-data-collection-api-fi.ipynb

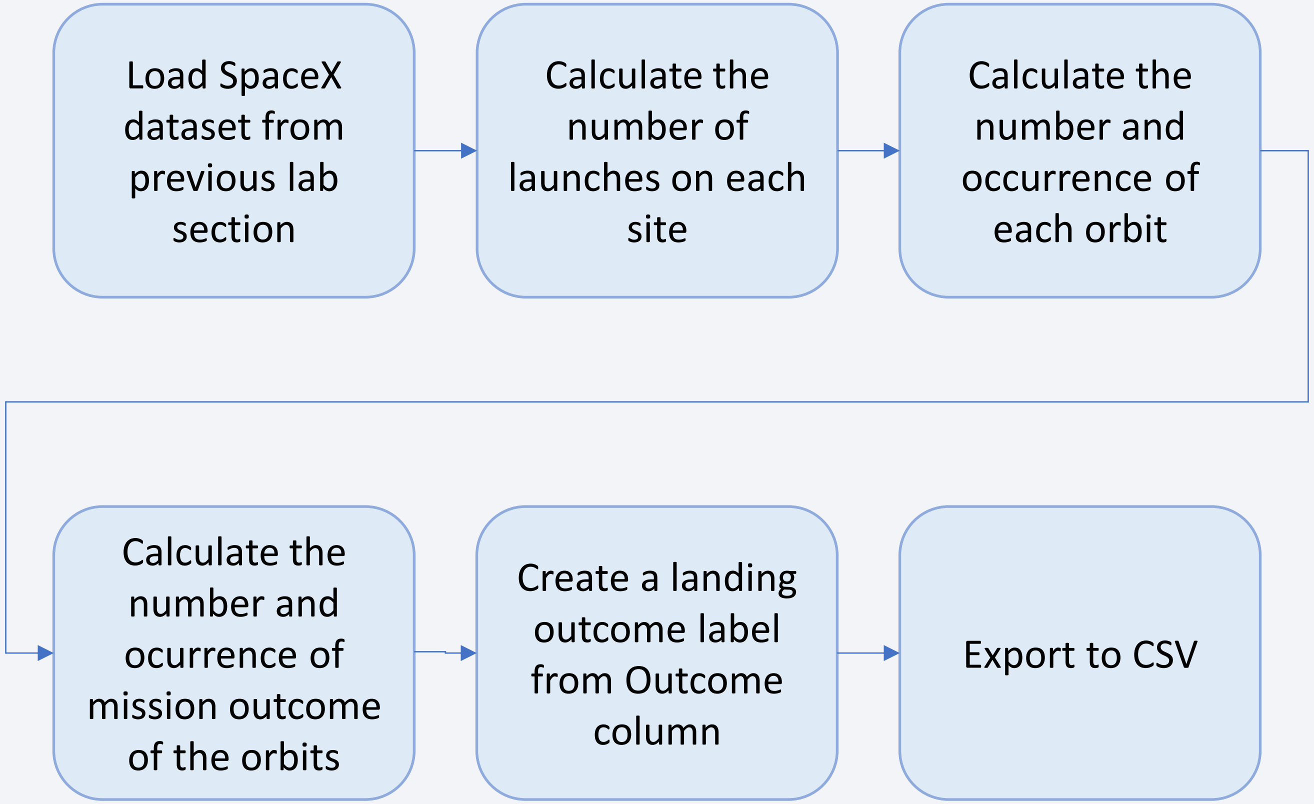
Data Collection - Scraping



GitHub URL: https://github.com/fimamu/IBM-Data-Science-SpaceX-Capstone/blob/ff22034a0f999997e9ea325bf8cccd40e6396b23/02-spacex-webscraping-fi.ipynb

Data Wrangling

- Loaded data from SpaceX dataset from previous section
- Launch sites, orbit types and mission outcomes were calculated
- Created an Outcome list where 1 represented a success landing and 0 represented a failure
- Determined success rate



GitHub URL: https://github.com/fimamu/IBM-Data-Science-SpaceX-Capstone/blob/ff22034a0f999997e9ea325bf8cccd40e6396b23/03-spacex-data-wrangling-fi.ipynb

EDA with Data Visualization

- Scatter plots to see relationship with variables:
  - Flight number vs. Payload
  - Flight number vs Launch site
  - Pay load Mass (kg) vs Launch site
  - Flight number vs. Orbit
  - Pay load Mass (kg) vs. Orbit
- Bar chart of Orbit vs. Class to compare success rate per Orbit type
- Line chart of launch Success Rate yearly to see if there is a trend

GitHub URL: https://github.com/fimamu/IBM-Data-Science-SpaceX-Capstone/blob/486e59e9259d3823afc3c022a0bac3f19edd200c/05-spacex-edadataviz-fi.ipynb

EDA with SQL

Performed SQL queries:

- Displayed the names of the unique launch sites in the space mission
- Displayed 5 records where launch sites begin with the string 'CCA'
- Displayed the total payload mass carried by boosters launched by NASA (CRS)
- Displayed average payload mass carried by booster version F9 v1.1
- Listed the date when the first successful landing outcome in ground pad was achieved
- Listed the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listed the total number of successful and failure mission outcomes
- Listed the names of the booster versions which have carried the maximum payload mass, using a subquery
- Listed the failure landing outcomes in drone ship, booster versions and launch site for the months in year 2015.
- Ranked 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

GitHub URL: https://github.com/fimamu/IBM-Data-Science-SpaceX-Capstone/blob/486e59e9259d3823afc3c022a0bac3f19edd200c/04-spacex-eda-sql-fi.ipynb

Build an Interactive Map with Folium

Map objects created and added to a folium map

- Created a **marker** with circle for NASA Johnson Space Center as start location
- Added **circles** for each launch site
- Added **green markers** for successful and **red markers** for unsuccessful launches at each launch site to illustrate their success rate
- Added **lines** to show distance from launch sites to the coastline

GitHub URL: https://github.com/fimamu/IBM-Data-Science-SpaceX-Capstone/blob/486e59e9259d3823afc3c022a0bac3f19edd200c/06-spacex-launch-site-location-fi.ipynb

Build a Dashboard with Plotly Dash

- Added a drop-down list with all launch sites
- Created Pie Chart based on selected site to show its success rate
- Added a Slider of the Payload Mass range to filter it for the scatterplot
- Included Scatterplot of Payload Mass vs. Success Rate to show the correlation between these variables

GitHub URL: https://github.com/fimamu/IBM-Data-Science-SpaceX-Capstone/blob/ae85d3726d352ba793ac3ef032cf292e30e61005/08-spacex_dash_app.py

Predictive Analysis (Classification)

1. Create NumPy array from the column “Class” of the dataframe
2. Standardize data with .StandardScaler() and reassign it to the variable X
3. Split data into training and testing sets using train_test_split function
4. Create a GridSearch object with cv = 10 to optimize the parameters
5. Calculate accuracy on the test data with method .score()
6. Examine the Confusion Matrix for each model
7. Find the best performing method using Jaccard_score and F1 Score

GitHub URL: https://github.com/fimamu/IBM-Data-Science-SpaceX-Capstone/blob/ae85d3726d352ba793ac3ef032cf292e30e61005/07-spacex-machine-learning-prediction-fi.ipynb

Results

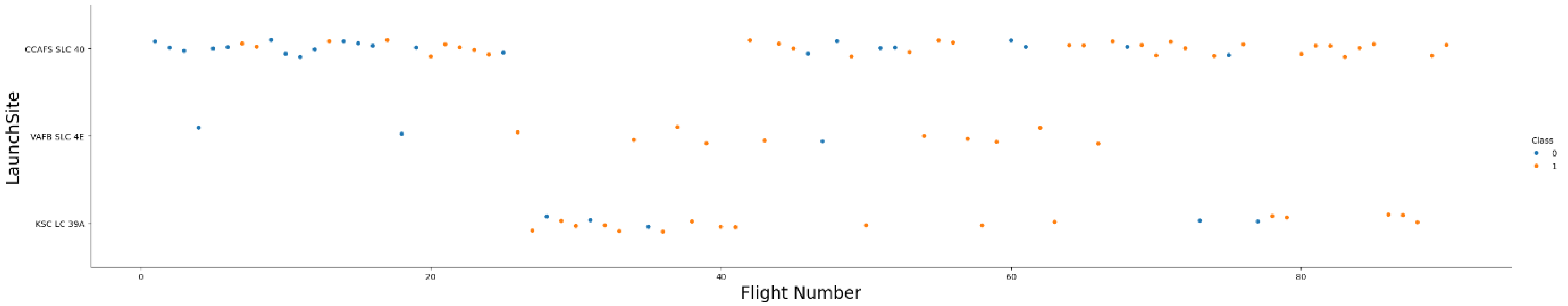
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

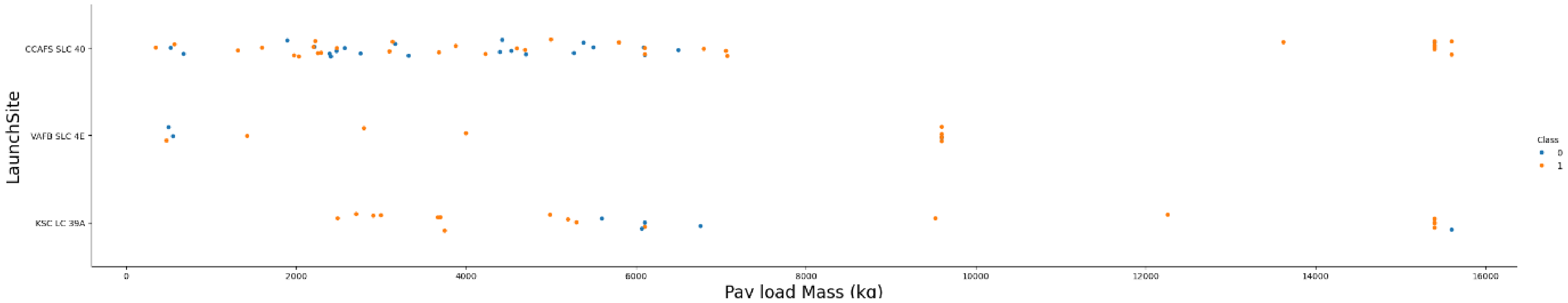
Flight Number vs. Launch Site

- Earlier flights had lower success rate (blue dots) while later flights had higher success rate (orange dots)
- CCAFS SLC 40 site concentrates about half of the launches
- KSC LC 39A site had the highest success rate, followed by VAFB SLC4E
- New launches have a higher success rate



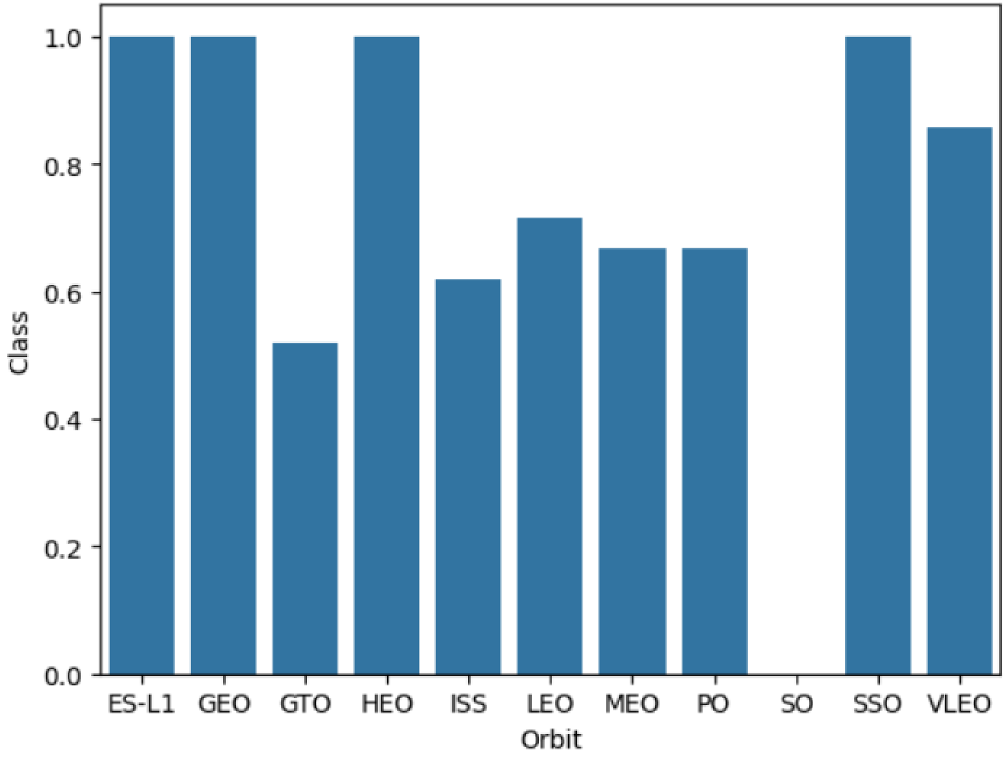
Payload vs. Launch Site

- In general, **higher payload mass (kg)** have **higher success rate**
- Launches with payload mass over 7000 kg were successful, except for one launch from the KSC LC 39A site
- VAFB SKC 4E site has no launch above ~10000 kg
- KSC LC 39A site has a 100% success rate with payload mass below 5500 kg



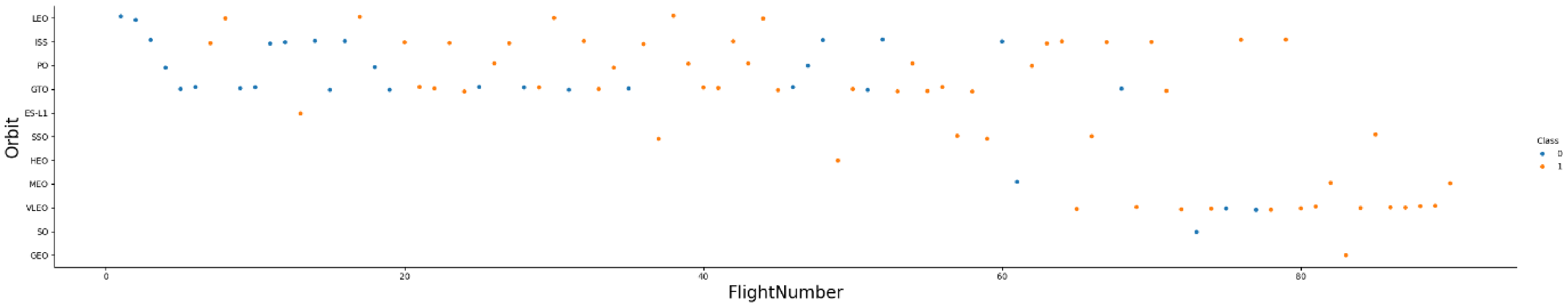
Success rate vs. Orbit type

- Orbits types with 100% success rate:
  - ES-L1, GEO, HEO and SSO
- Orbit types with success rate between 50% and 85%:
  - GTO, ISS, LEO, MEO and PO
- SO Orbit is the only one with 0% success rate



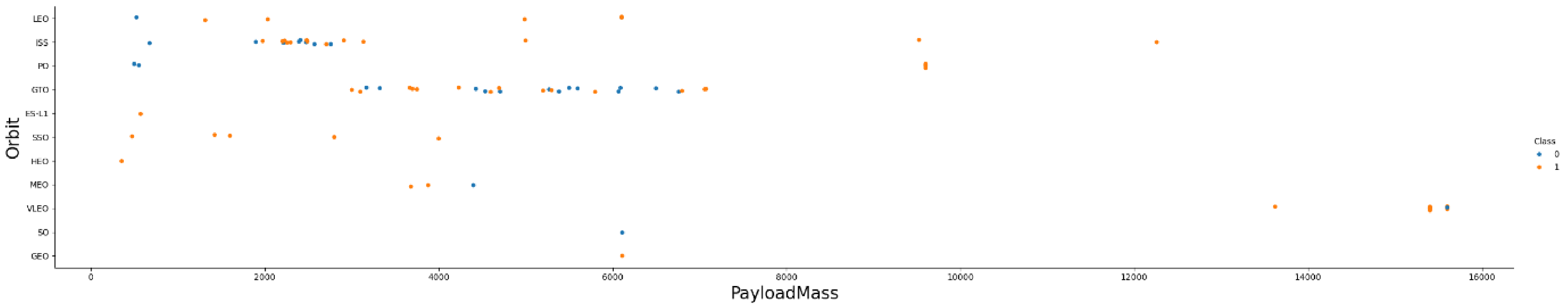
Flight number vs. Orbit type

- In general, **success rate** increases with the **number of flights**, especially in the LEO Orbit
- However, the GTO Orbit does not follow this relationship
- In the VLEO Orbit, failure rate concentrates around 75 flights



Payload vs. Orbit type

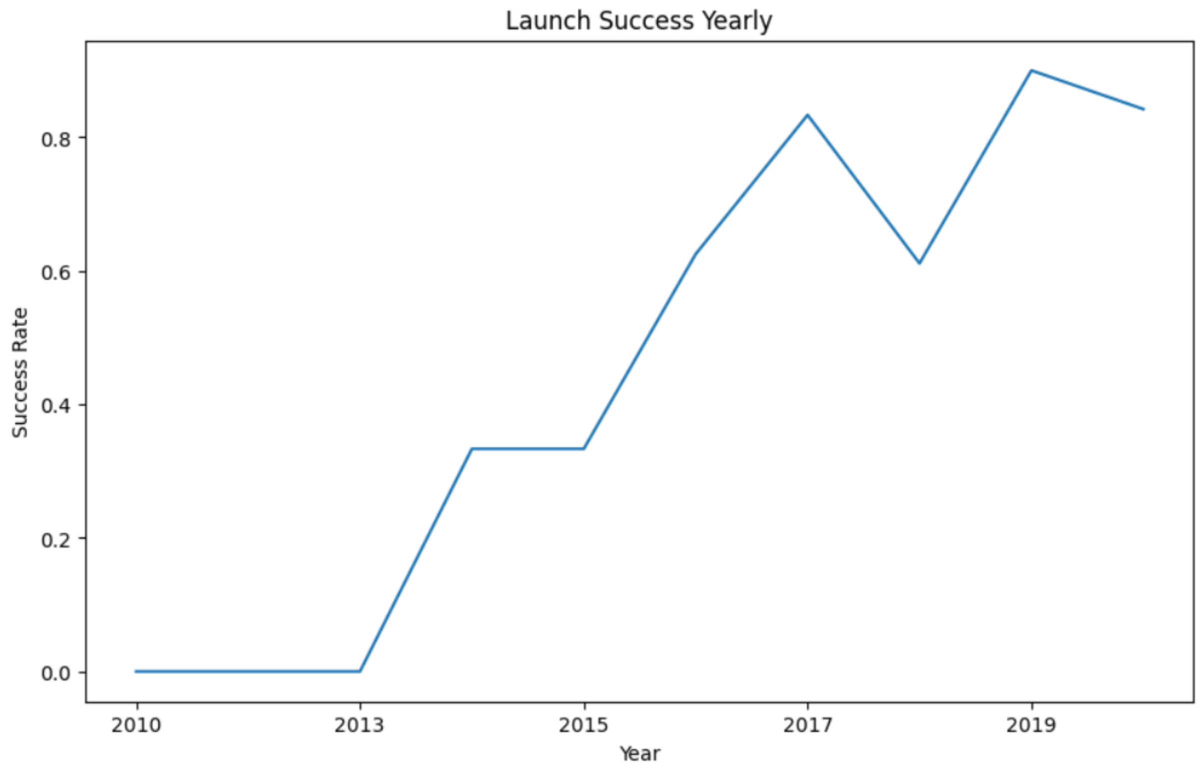
- Heavy payloads have better success rate in LEO, ISS and PO Orbits
- The GTO Orbit does not show a pattern around this relationship (Payload vs. Orbit)



Launch Success Yearly Trend

Exploratory data analysis results

- The success rate has improved since 2013
- However, it has declined from 2017-2018 and from 2019-2020



All Launch Site Names

- Displaying the names of the unique launch sites in the space mission

```
[17]: %sql SELECT DISTINCT(launch_site) FROM SPACEXTBL;
 * sqlite:///my_data1.db
Done.
```

[17]:

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

Launch Site Names Begin with 'CCA'

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

```
[19]: %sql SELECT * FROM SPACEXTBL WHERE (launch_site) LIKE 'CCA%' LIMIT 5;

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

[19]:

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

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

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

```
[21]: %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)';
```

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

[21]:

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Displaying average payload mass carried by booster version F9 v1.1

```
[25]: %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE BOOSTER_VERSION LIKE '%F9 v1.1%';

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

[25]:

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- Listing the date when the first succesful landing outcome in ground pad was achieved

```
[27]: %sql SELECT MIN(DATE) FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (ground pad)';

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

[27]:

MIN(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
[29]: %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.
```

[29]:

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

Total Number of Successful and Failure Mission Outcomes

- Listing the total number of successful and failure mission outcomes

```
[31]: %sql SELECT Mission_Outcome, COUNT(*) AS TOTAL FROM SPACEXTBL \
          GROUP BY Mission_Outcome;

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

[31]:

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

Boosters Carried Maximum Payload

- Listing the names of the booster versions which have carried the maximum payload mass, using a subquery

```
[40]: %sql SELECT Booster_Version, PAYLOAD_MASS__KG_ FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_=(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);

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

[40]:

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

- Listing the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

```
[35]: %sql SELECT substr(Date, 6,2) AS MONTH, Date, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTBL \
WHERE Landing_Outcome='Failure (drone ship)' AND substr(Date,0,5)='2015' ;

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

[35]:

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

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.

[37]:
```
%sql SELECT Landing_Outcome, COUNT(*) AS OUTCOME_COUNT FROM SPACEXTBL \
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY OUTCOME_COUNT DESC;
```

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

[37]:

Landing_Outcome	OUTCOME_COUNT
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

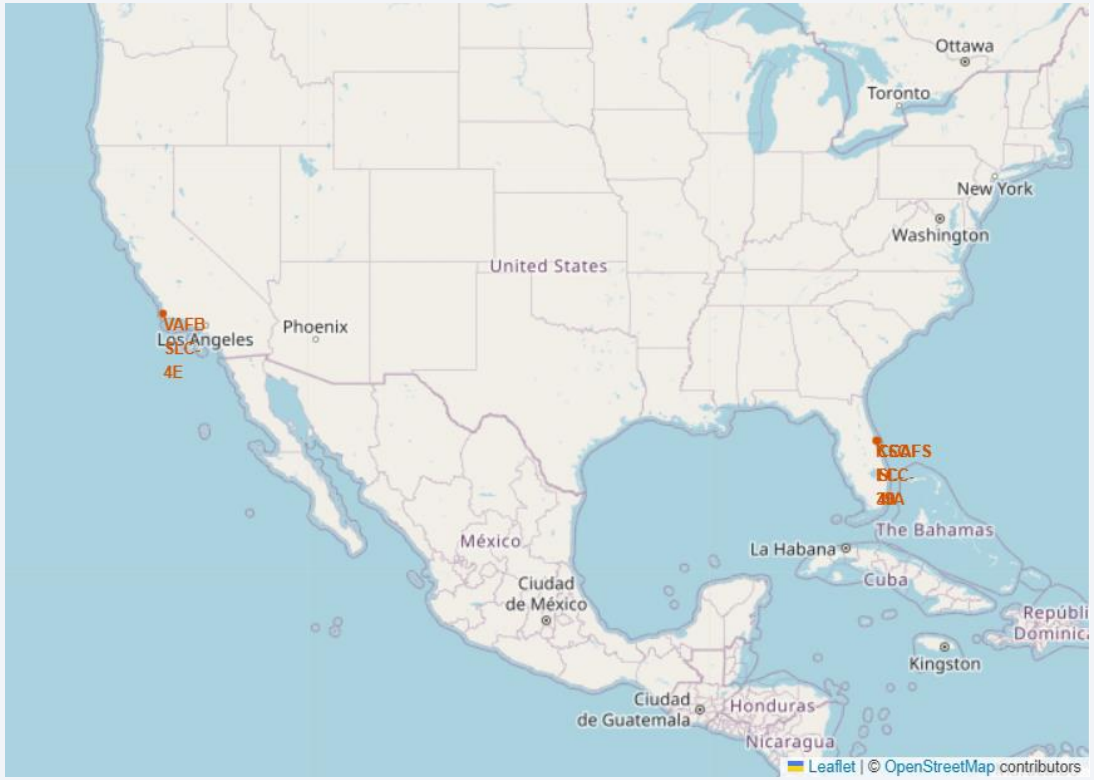
Section 3

Launch Sites Proximities Analysis

Launch sites location markers

- Most launch sites are located near the Equator line. This represents a physical advantage, as rockets launched from areas close to the Equator line can get an extra speed boost due to inertia

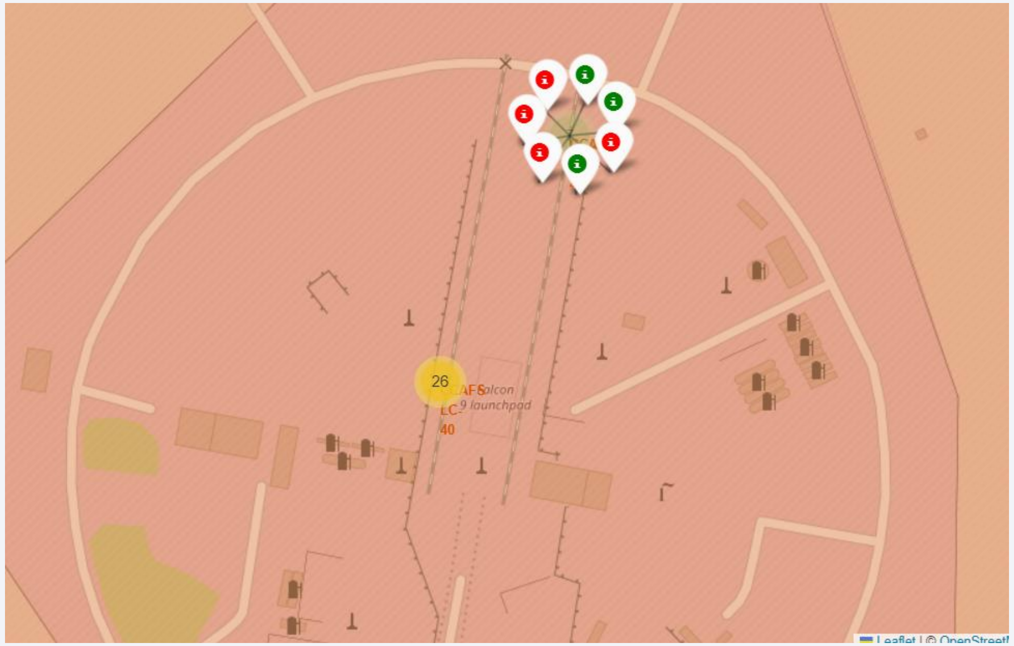
- Besides that, all launch sites are close to coast areas, minimizing the risk of accidents in high-density areas



Launch Outcomes

CCAFS SLC-40 Site

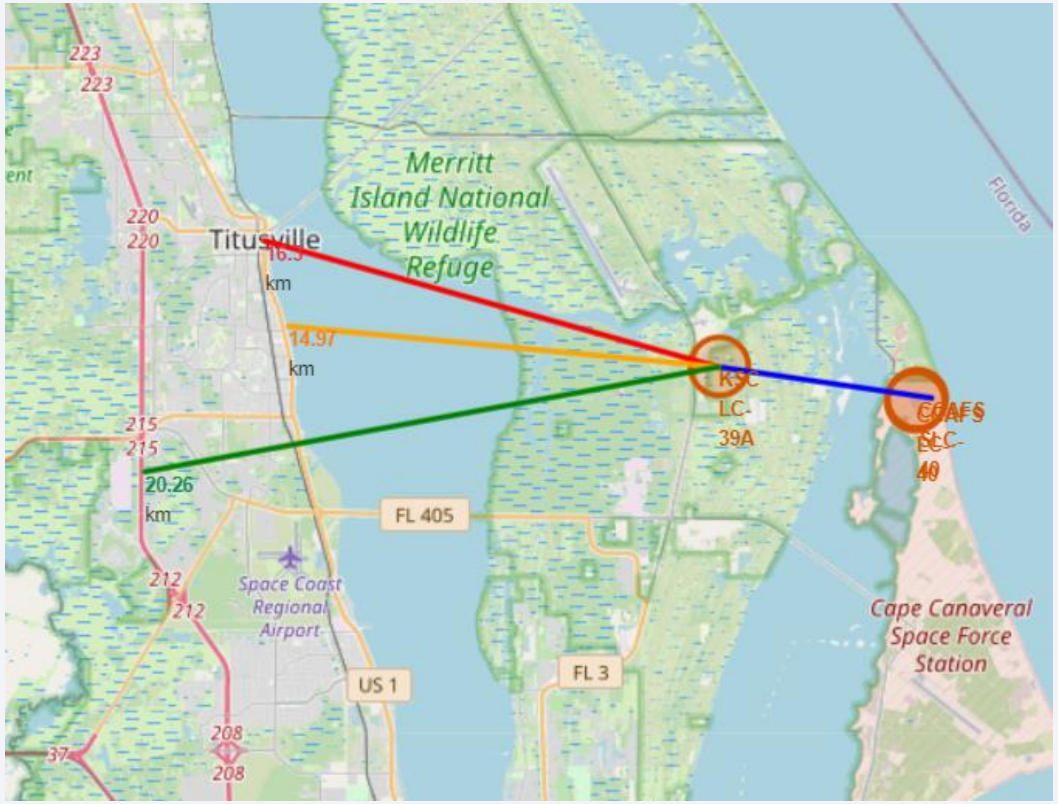
- Green marker for successful launches and Red marker for unsuccessful launches
- Therefore, this site presents a **42,9% success rate** (=3/7)



Distance to Proximities

CCAFS SLC-40 Site

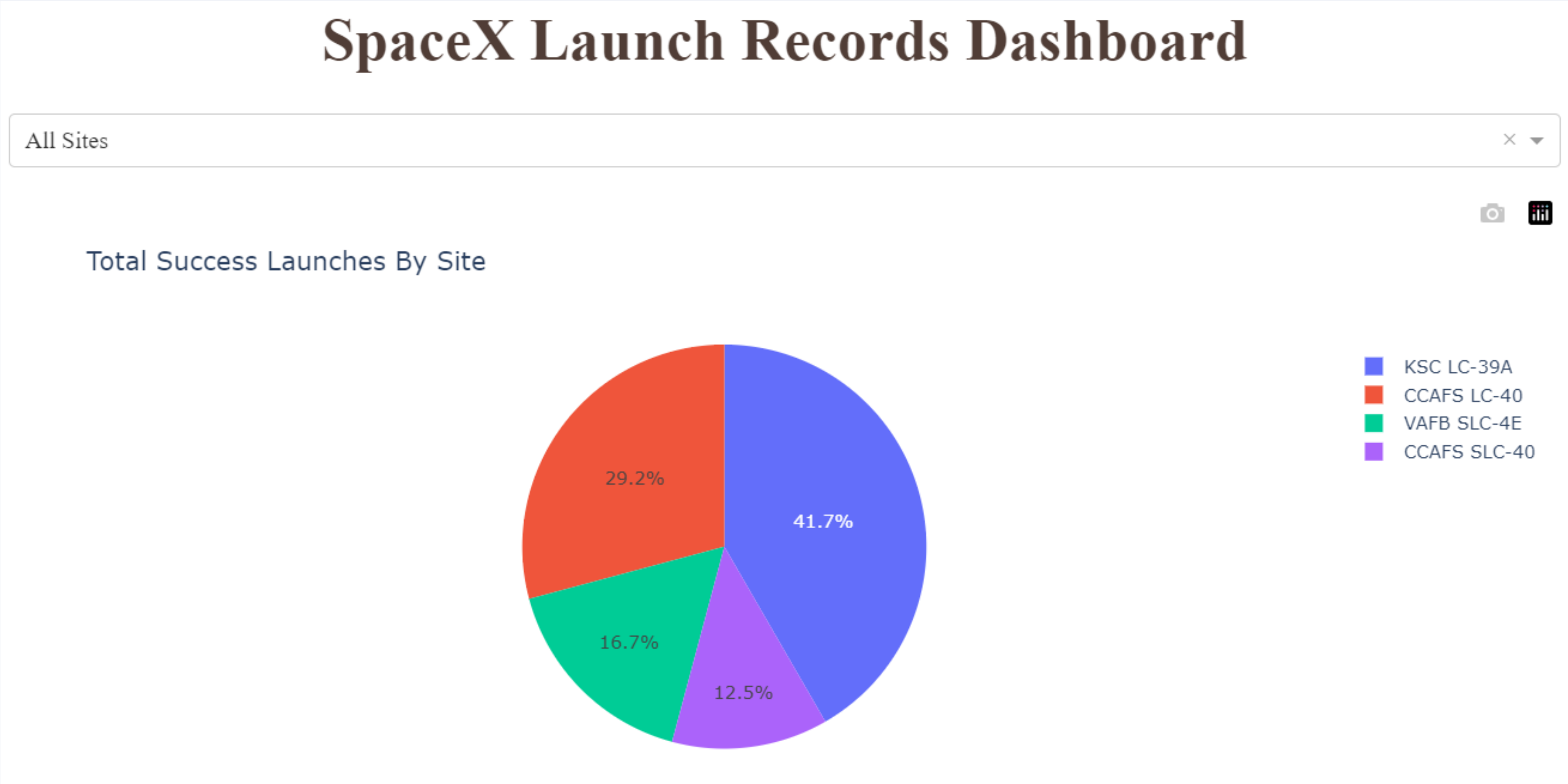
- **16,32 km** from railway
- **14.97 km** from coastline
- **20.26 km** from highway



Section 4

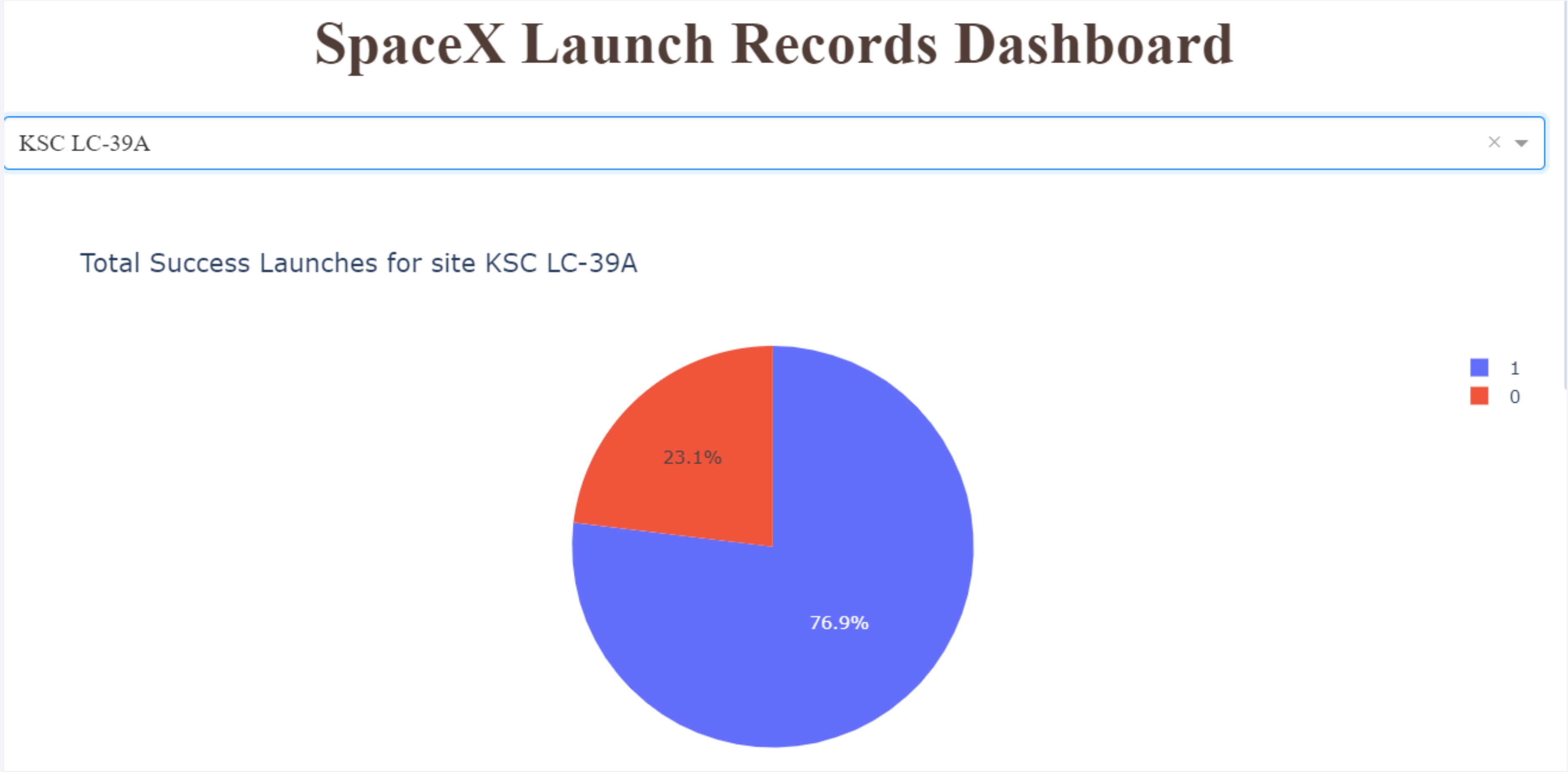
Build a Dashboard with Plotly Dash

Success Launches by Site



- KSC LC-39A site presents the most successful launch rate (41,2%)

Launch site with highest launch rate



- KSC LC-39A success rate (76.9%) consists on 10 successful launches and 3 unsuccessful launches

Payload vs. Launch Outcome for all sites

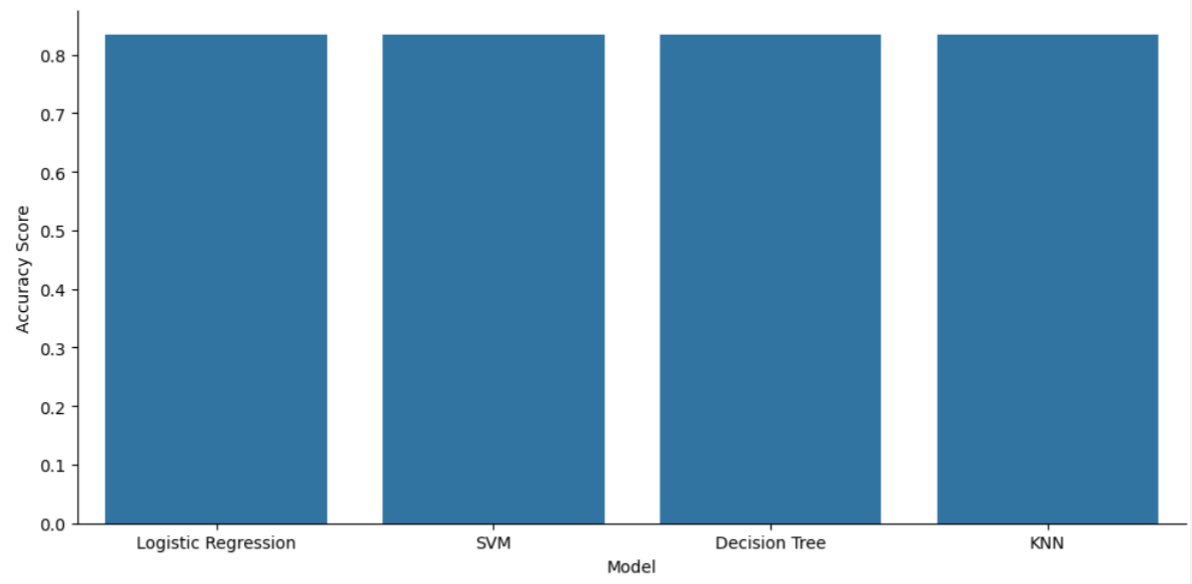
- Most success launches concentrated on payloads between 2k and 6k
- Booster V1.1 category had the worse successful launches, almost all had failed

Section 5

Predictive Analysis (Classification)

Classification Accuracy

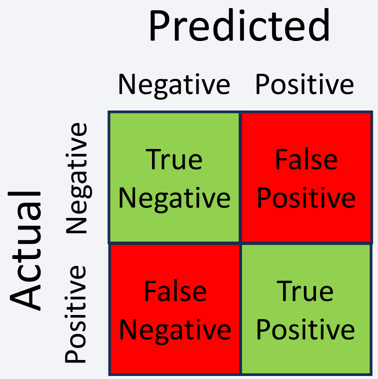
- SVM model had the better results
- All models had good accuracy performance (>80%), although the Decision Tree model presented a slightly lower result
- Jaccard and F1 scores showed the same conclusion



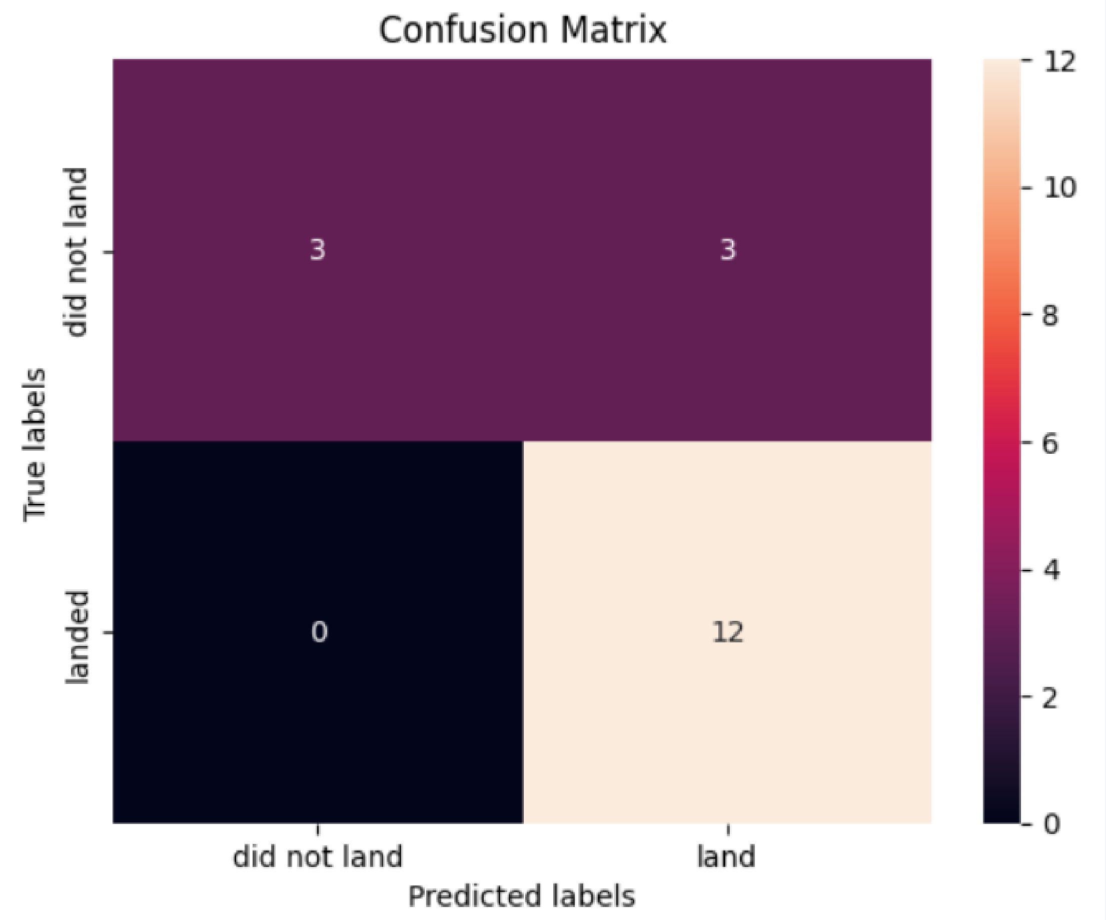
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.808824	0.819444
F1_Score	0.909091	0.916031	0.894309	0.900763
Accuracy	0.866667	0.877778	0.855556	0.855556

Confusion Matrix

- The confusion Matrix was the same for all models, outputting:
  - 12 True Positives
  - 3 False Positives
  - 3 True Negatives
  - 0 False Negatives



- The main problem is having false positives



Conclusions

- SpaceX launch success rate has improved over the years
- The success rate increases with the number of flights
- In general, higher payload mass (kg) have higher success rate
- Orbits ES-L1, GEO, HEO and SSO have a 100% success rate
- The KSC LC-39A site presents the most successful launch rate (41,2%)
- The Machine Learning Models can be used to predict successfull landing and SVM has the best results

Appendix

Github URL for Jupyter notebooks used in this project:

- Data Collection: https://github.com/fimamu/IBM-Data-Science-SpaceX-Capstone/blob/ff22034a0f999997e9ea325bf8cccd40e6396b23/01-spacex-data-collection-api-fi.ipynb
- Web Scrapping: https://github.com/fimamu/IBM-Data-Science-SpaceX-Capstone/blob/ff22034a0f999997e9ea325bf8cccd40e6396b23/02-spacex-webscraping-fi.ipynb
- Data Wrangling: https://github.com/fimamu/IBM-Data-Science-SpaceX-Capstone/blob/ff22034a0f999997e9ea325bf8cccd40e6396b23/03-spacex-data-wrangling-fi.ipynb
- Data Visualization: https://github.com/fimamu/IBM-Data-Science-SpaceX-Capstone/blob/486e59e9259d3823afc3c022a0bac3f19edd200c/05-spacex-edadataviz-fi.ipynb
- EDA with SQL: https://github.com/fimamu/IBM-Data-Science-SpaceX-Capstone/blob/486e59e9259d3823afc3c022a0bac3f19edd200c/04-spacex-eda-sql-fi.ipynb
- Follium Maps: https://github.com/fimamu/IBM-Data-Science-SpaceX-Capstone/blob/486e59e9259d3823afc3c022a0bac3f19edd200c/06-spacex-launch-site-location-fi.ipynb
- Plotly Dashboard: https://github.com/fimamu/IBM-Data-Science-SpaceX-Capstone/blob/ae85d3726d352ba793ac3ef032cf292e30e61005/08-spacex_dash_app.py
- Predictive Analysis: https://github.com/fimamu/IBM-Data-Science-SpaceX-Capstone/blob/ae85d3726d352ba793ac3ef032cf292e30e61005/07-spacex-machine-learning-prediction-fi.ipynb

Thank you!