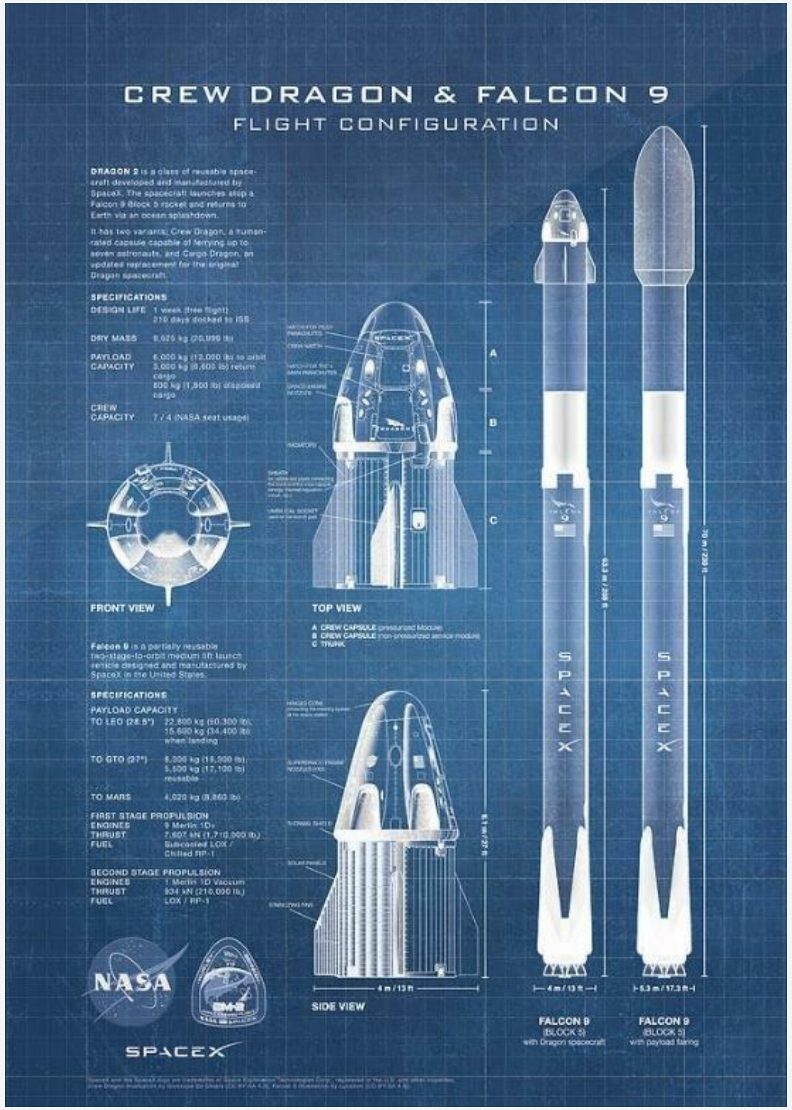
IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Gökhan Elbistan
22.04.2024

Outline

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



Executive Summary

- Summary of methodologies
    - Data collection
    - Data wrangling
    - Exploratory Data Analysis with Data Visualization
    - Exploratory Data Analysis with SQL
    - Building an interactive map with Folium
    - Building a Dashboard with Plotly Dash
    - Predictive analysis (Classification)
- Summary of all results
    - Exploratory Data Analysis results
    - Interactive analytics demo in screenshots
    - Predictive analysis results

Introduction

- Project background and context
  - SpaceX is the most successful company of the commercial space age, making space travel affordable. The company advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX 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 machine learning models, we are going to predict if SpaceX will reuse the first stage.
- Research Questions
  - 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 landings increase over the years?
  - What is the best algorithm that can be used for binary classification in this case?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected with SpaceX Rest API and Web Scraping from Wikipedia with BeautifulSoup
- Perform data wrangling
  - Filters were applied on the data, missing values has been dealt with and one hot encoding has been used to transform categorical variables into integer from SVM.
- 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 through GridSearchCV to ensure the best results

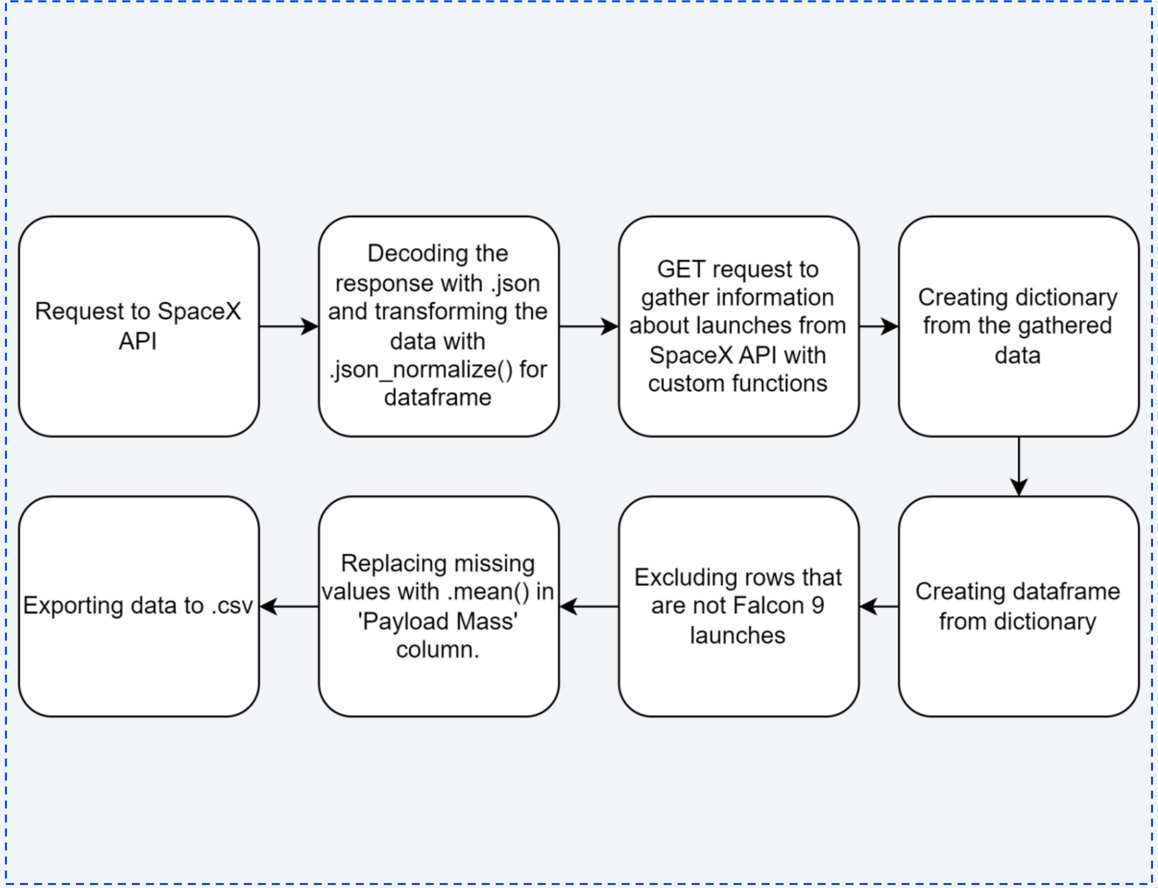
Data Collection

- Data was collected using various methods:
  - Data collection for landing was done with requests to SpaceX API.
  - Raw data has been decoded with .json_normalize(), missing data has been filled with appropriate values.
  - Additional data about Falcon 9 launch records has been gathered from Wikipedia with BeautifulSoup.
  - The objective was to extract information about landing statistics for various models.

Data Collection – SpaceX API

- SpaceX API has been used to gather information about launches of Falcon series. Data was transformed to be suitable for dataframe. Resulting data has been filtered with Falcon 9 launches and missing values has been filled with mean value of the column.

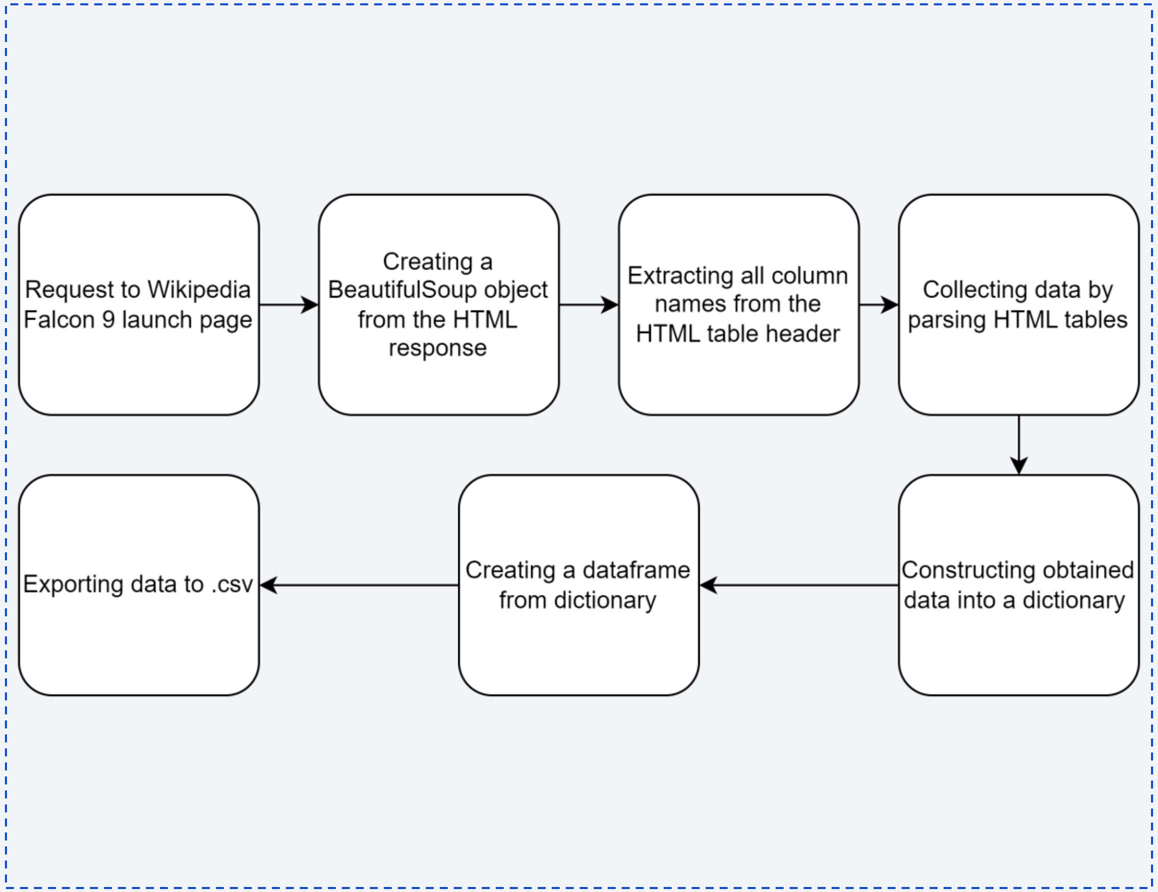
- Link to the Notebook: [Data Collection Notebook for SpaceX API]



Data Collection - Scraping

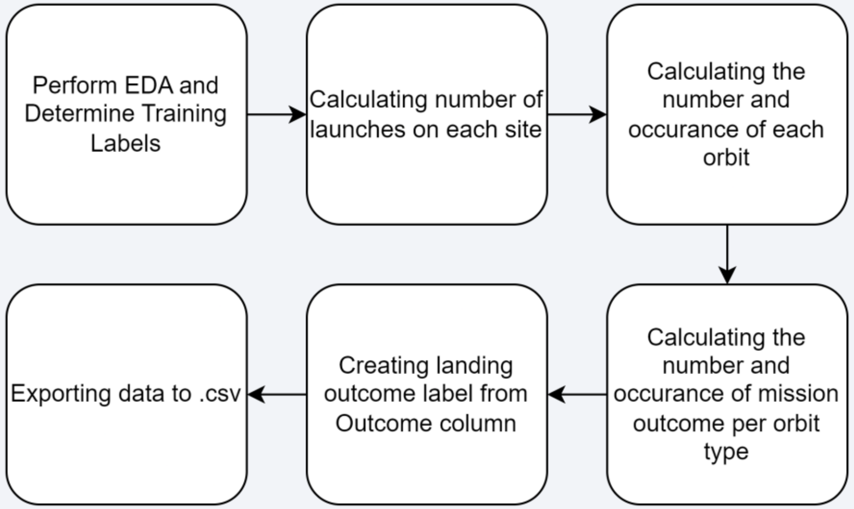
- Launch records has been mined through Wikipedia with BeautifulSoup for gathering Falcon 9 launch records. Results was parsed and turned into a pandas dataframe.

- Link to the Notebook: [Web Scraping Project with BeautifulSoup]



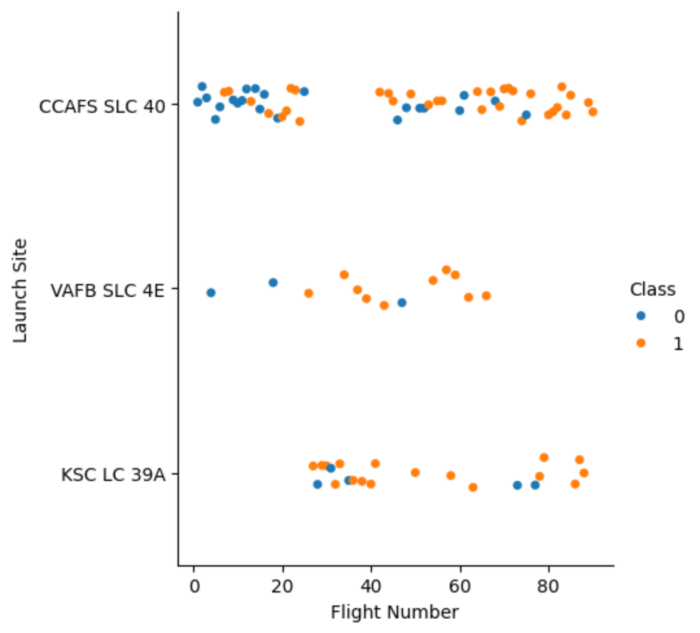
Data Wrangling

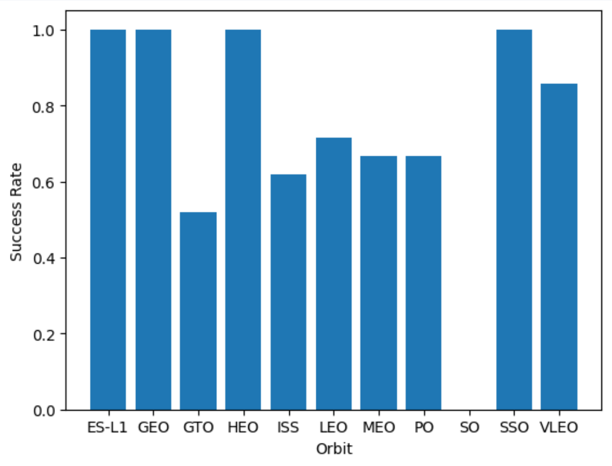
- Our dataset details various booster landing scenarios. These scenarios fall into two categories: successful landings and unsuccessful landings.
- Successful Landings:
    - True Ocean: Booster successfully landed in a designated ocean zone.
    - True RTLS: Booster successfully landed on a designated ground pad.
    - True ASDS: Booster successfully landed on a drone ship in the ocean.
- Unsuccessful Landings:
    - False Ocean: Attempted landing in the ocean zone failed.
    - False RTLS: Attempted landing on the ground pad failed.
    - False ASDS: Attempted landing on the drone ship failed.
- For machine learning purposes, we convert these outcomes into simpler labels: 1 for successful landings and 0 for unsuccessful landings.
- Link to the Notebook: Data Wrangling



EDA with Data Visualization

- Charts were plotted:
    - Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type and Success Rate Yearly Trend
- Scatter plots show the relationship between variables. If a relationship exists, they could be used in machine learning model.
- Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value.
- Line charts show trends in data over time (time series).





Link to the Notebook: [EDA with Data Visualization]

EDA with SQL

- Sqlite was using for queries in SpaceX dataset.
- EDA was applied wih SQL to get insight from the data. Queries were created for these purposes:
  - Names of unique launch sites in the space mission.
  - Total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - Total number of successful and failure mission outcomes
  - Failed landing outcomes in drone ship, their booster version and launch site names.
- Link to the Notebook: EDA with SQL

Build an Interactive Map with Folium

- All launch sites has been marked and markers, circles, lines added to point out success or failure of launches for each site on folium map.
- Feature launch outcomes has been assigned with binary variable (0: Failure, 1: Success)
- Colored markers has been added (Green/Red) to identify which launch sites have relatively high success rates.
- Colored lines has been added to show distances between the launch site KSC LC-39A and its proximities like Railway, Highway, Coastline and Closest City

Link to the Notebook: [Interactive Map with Folium]

Build a Dashboard with Plotly Dash

Interactive dashboard has been created with Plotly Dash.

Dropdown list has been created for launch sites.

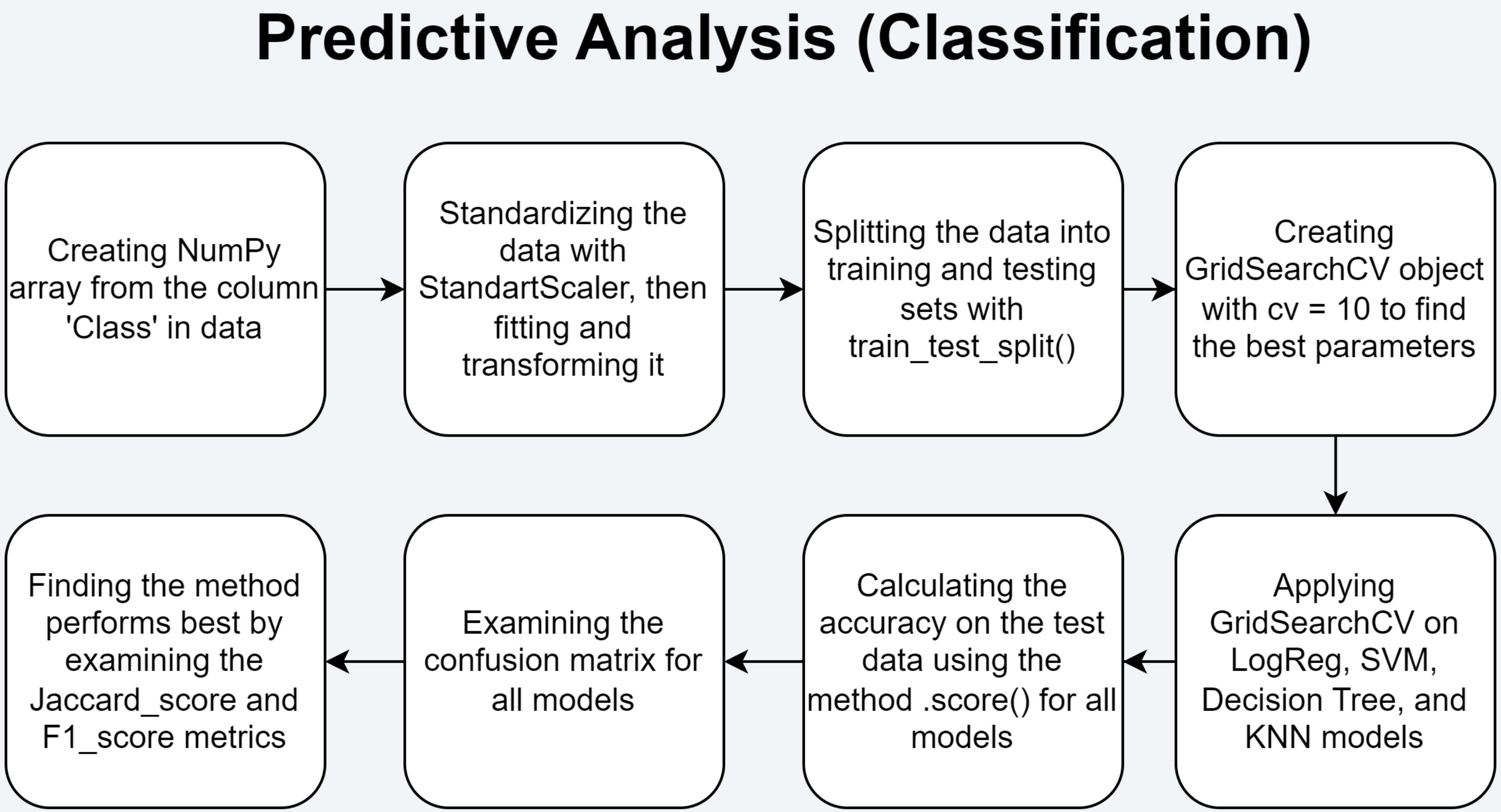
Pie charts has been created to show total launches by each site.

Scatter graph has been created to show relationship between Outcome and Payload Mass (kg) for different booster versions.

Slider has been created for Payload Mass Range.

Link to the Notebook: [Web Scraping Project with BeautifulSoup]

Predictive Analysis (Classification)



Link to the Notebook: Predictive Analysis (Classification)

Results

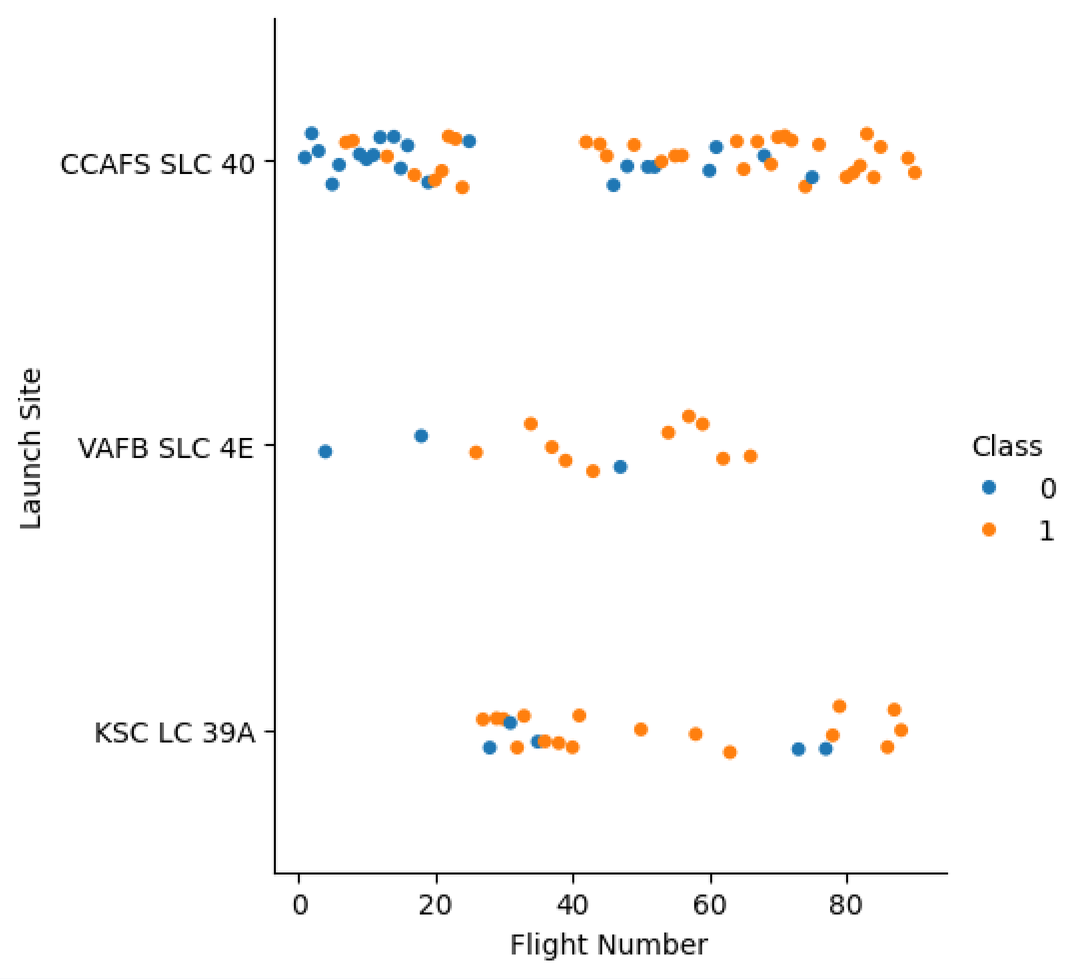
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

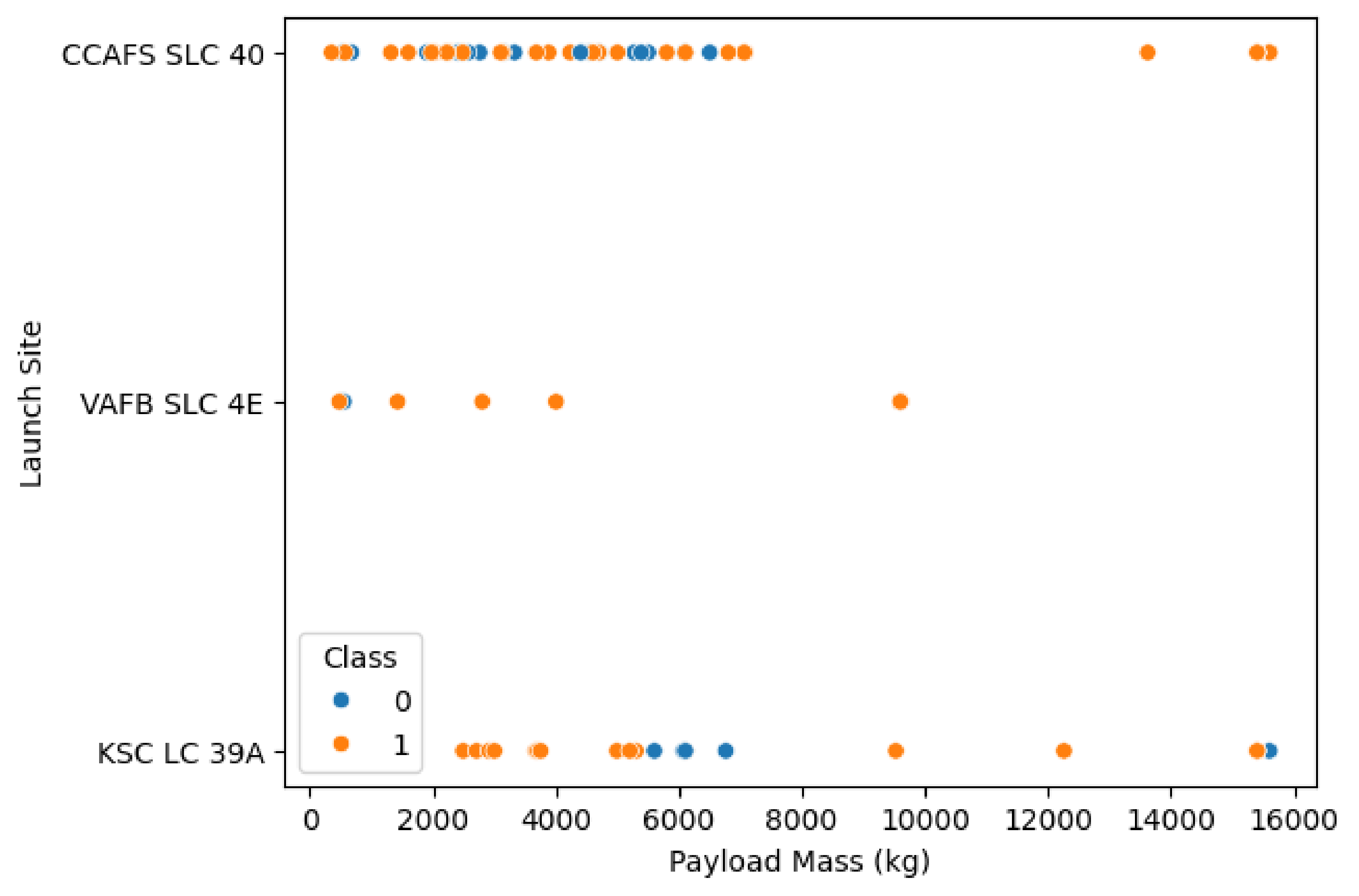
Flight Number vs. Launch Site

- Earlist launch trials have been failed more while latest flights have been more successful.
- CCAFS SLC 40 has about half of the launches.
- VAFB SLC 4e and KSC LC 39A have higher success raters.
- It can be assumed that more flight number can be resulted with higher rate of success.



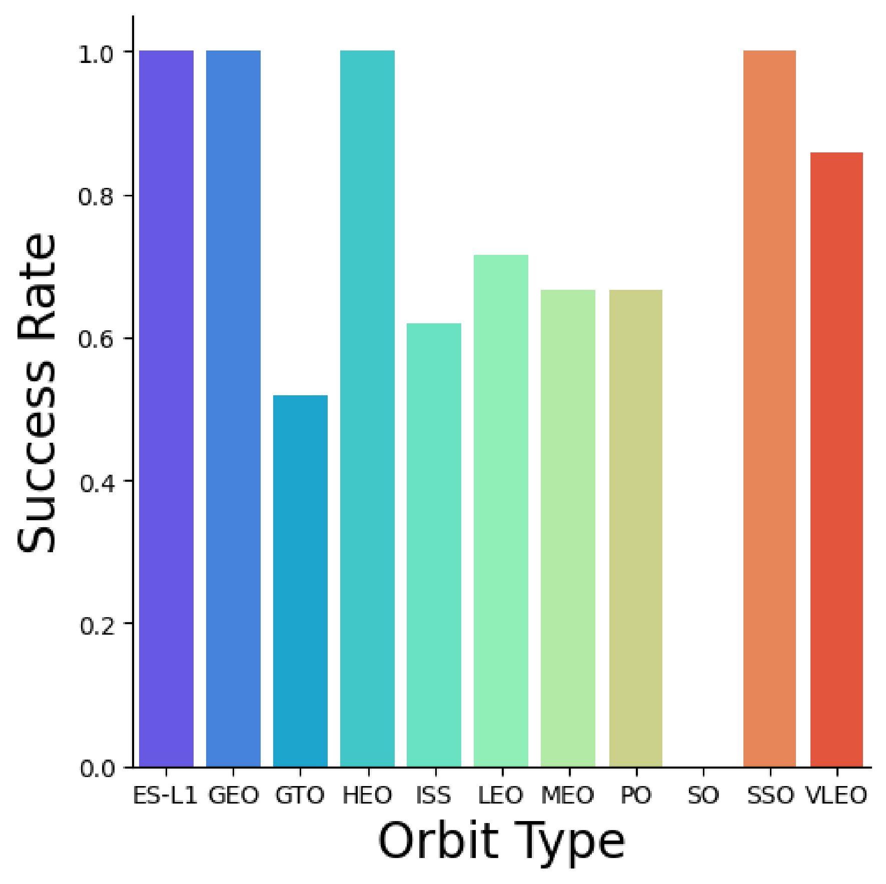
Payload vs. Launch Site

- For each launch site; higher the payload mass, higher the success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.



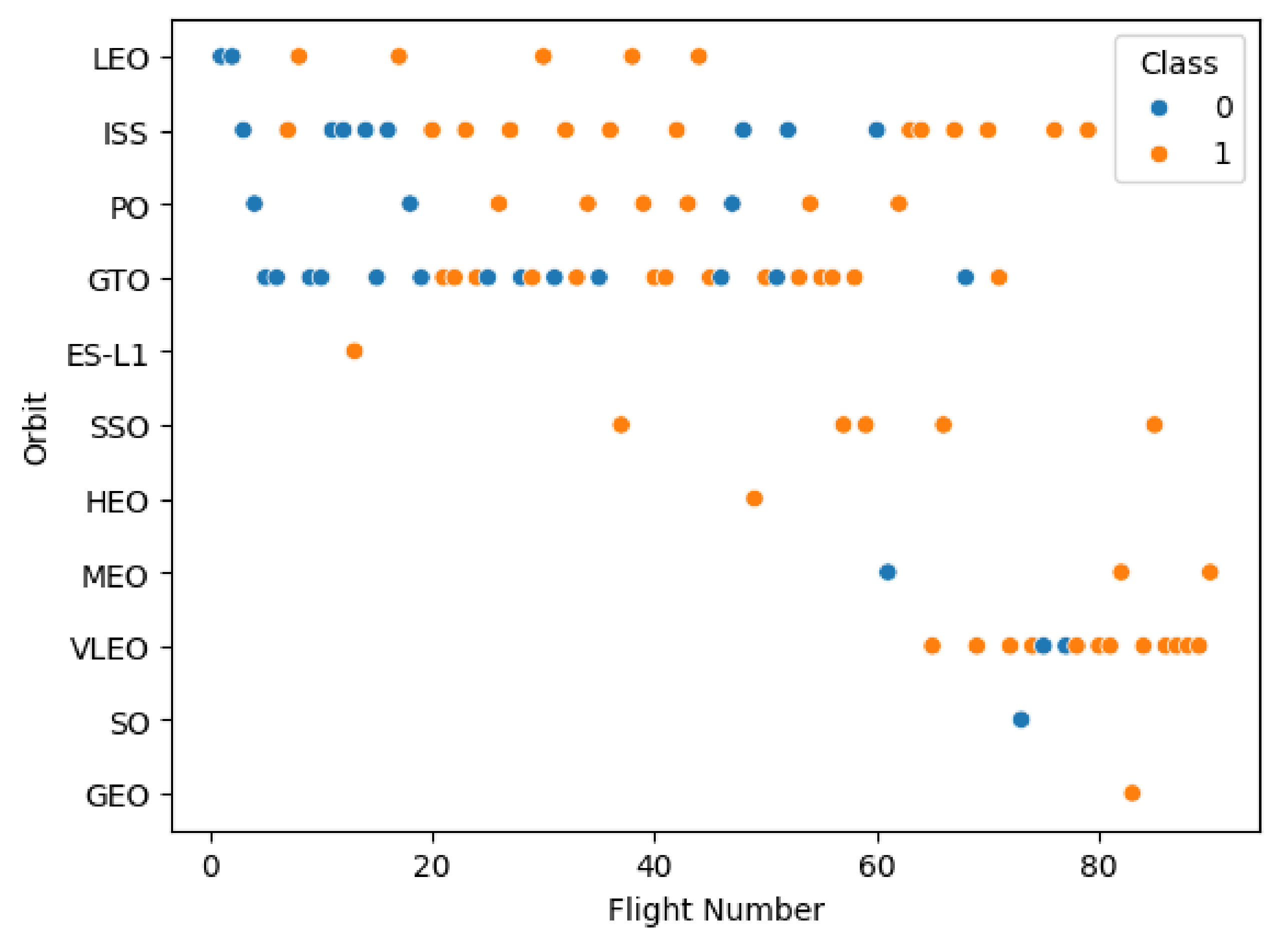
Success Rate vs. Orbit Type

- Orbits with 100% success rate are:
  - ES-L1, GEO, HEO, SSO
- Orbits with 0% success rate:
  - SO
- Orbits with success rate between 50% and 85%:
  - GTO, ISS, LEO, MEO, PO



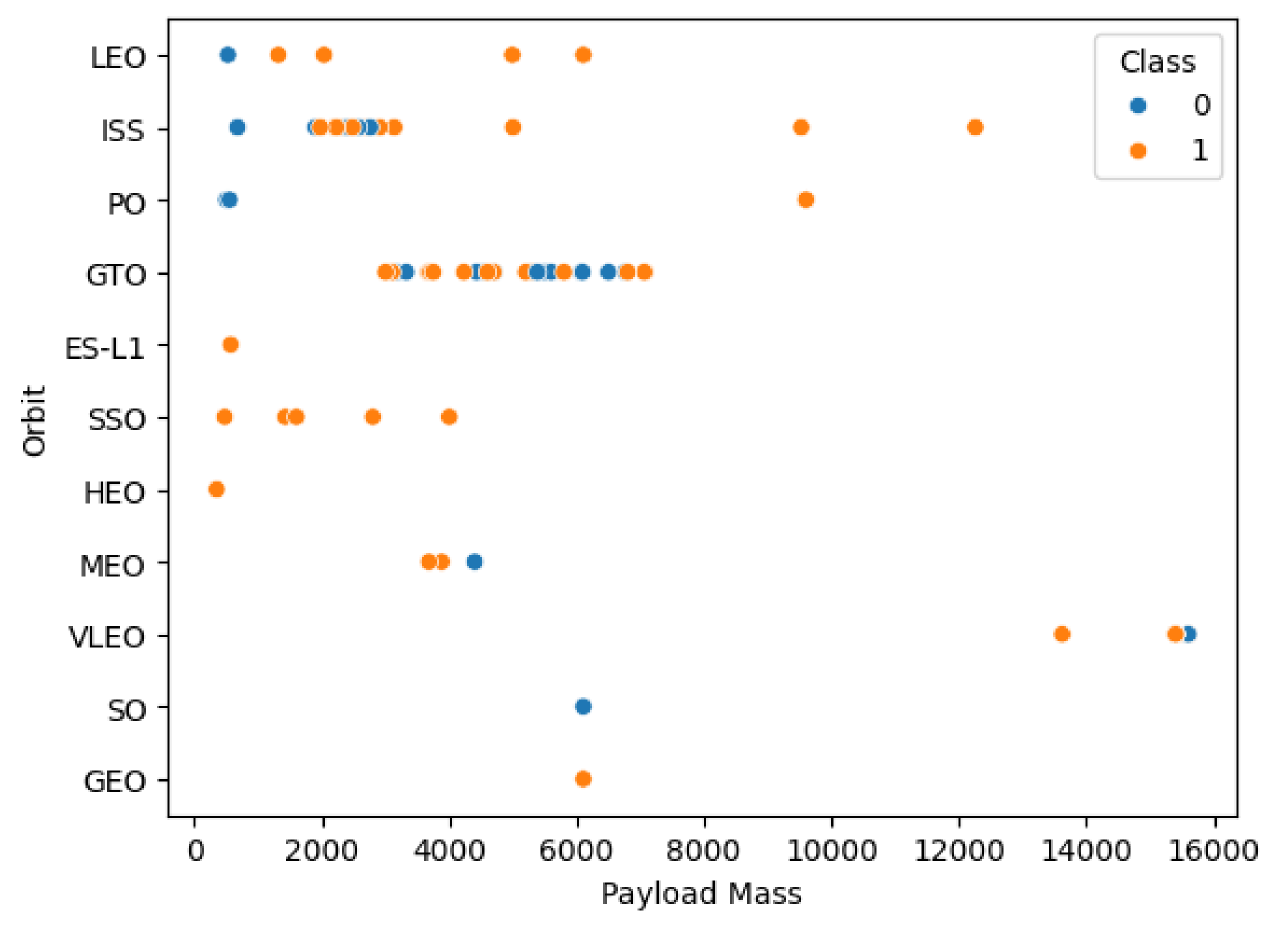
Flight Number vs. Orbit Type

- There is no constant evidence for the relationship between Flight Number and Orbit Type.



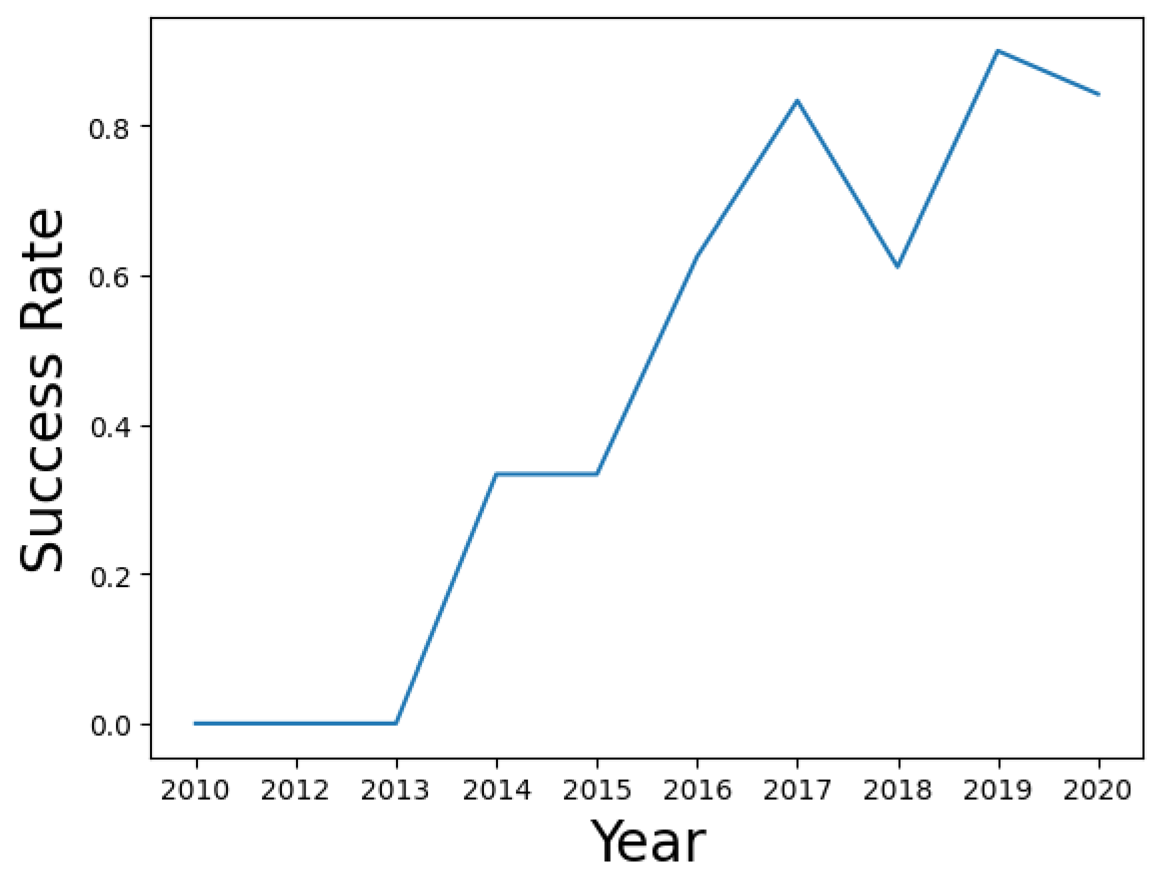
Payload vs. Orbit Type

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



Launch Success Yearly Trend

- Trials has been started at 2013 and linearly increased until 2017. Between 2017 and 2018 was at decrease with a fixation between 2018 and 2019.



All Launch Site Names

- Sqlite was used in Jupyter Notebook for this query.
- From SPACEXTABLE table, each unique value has been shown in launch_site column.

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

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

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

Launch Site Names Begin with 'CCA'

- A query with ‘CCA’ filter was used with a limit of 5 to get records in launch sites that starts with CCA.

```
%sql SELECT * FROM SPACEXTABLE 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

- Total payload mass carried by boosters launched by NASA (CRS) has been calculated with an aggregate function in the query.

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS total_payload_mass FROM SPACEXTABLE WHERE customer = 'NASA (CRS)'
```

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

total_payload_mass
45596

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1 resulted as 2534.67.

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS mean_payload_mass FROM SPACEXTABLE WHERE Booster_Version LIKE '%F9 v1.1%'
```

✓ 0.0s

```
* sqlite:///my_data1.db?check_same_thread=False
Done.
```

mean_payload_mass
2534.6666666666665

First Successful Ground Landing Date

- Date of the first successful landing was achieved.

```
%sql SELECT MIN(Date) AS first_successful_landing_date FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)'
```

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

first_successful_landing_date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

- Present your query result with a short explanation here

Total Number of Successful and Failure Mission Outcomes

- Total number of successful outcome is 100 and failure is 1.

```
%sql select mission_outcome, count(*) as total_number from SPACEXTABLE group by mission_outcome;
```

✓ 0.0s

```
 * sqlite:///my_data1.db?check_same_thread=False
Done.
```

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

Boosters Carried Maximum Payload

- List of the names that has the booster versions which have carried the maximum payload mass.

```
%%sql

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

✓ 0.0s

```
* sqlite:///my_data1.db?check_same_thread=False
Done.
```

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

- Listing the 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,
    Landing_Outcome,
    Booster_Version,
    launch_site
FROM
    SPACEXTABLE
WHERE
    substr(Date, 0, 5) = '2015' AND
    Landing_Outcome LIKE '%Failure (drone ship)%'
```

✓ 0.0s

```
 * sqlite:///my_data1.db?check_same_thread=False
Done.
```

month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	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 between 2010-06-04 and 2017-03-20 in descending order.

```
%%sql

SELECT Landing_Outcome, COUNT(*) AS Count
FROM SPACEXTABLE
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY Count DESC
```

✓ 0.0s

* sqlite:///my_data1.db?check_same_thread=False
Done.

Landing_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

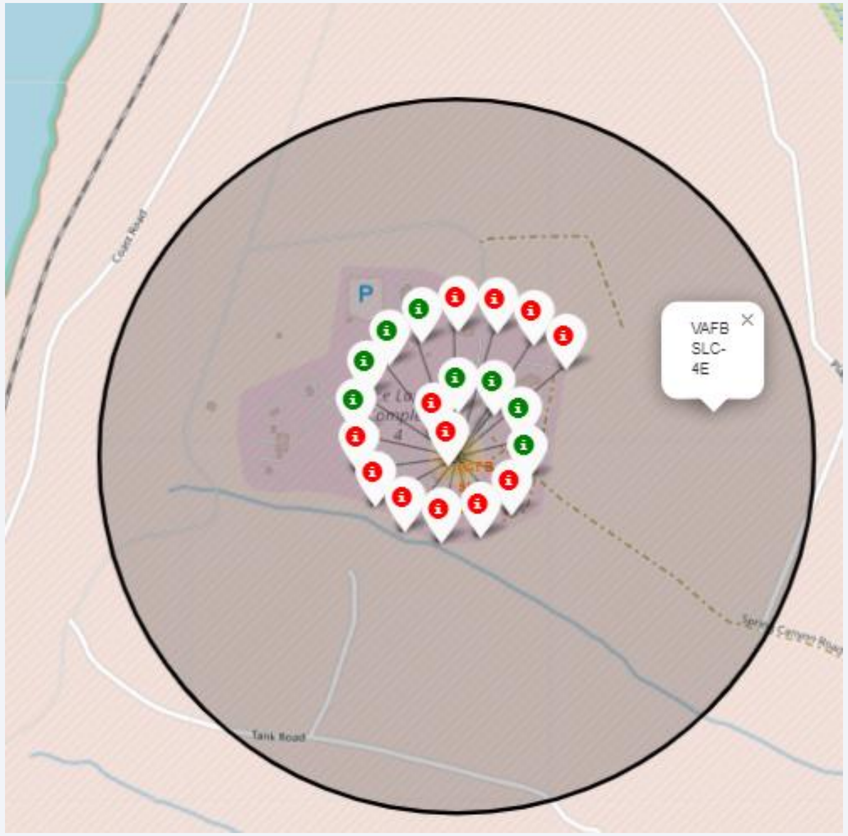
Launch Sites Proximities Analysis

Launch Sites All Around United States

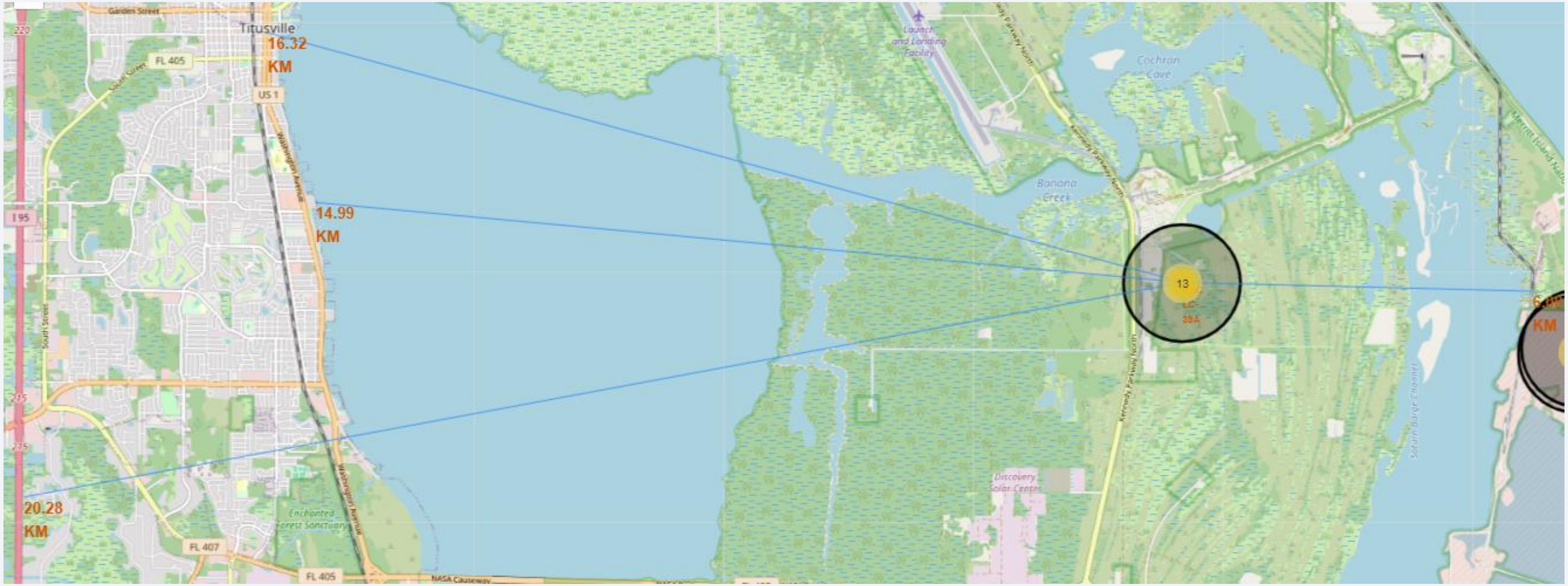
- Launch sites are near the coast lines.
- Launch sites are closer to the equator for more precise prediction.

Success of Launches

- Booster landing outcomes for each launch site can be seen with different color by their success. In addition to that, name of launching site can be seen as a pop-up too.



Distance From the Launch Site KSC LC-39A to its Proximity

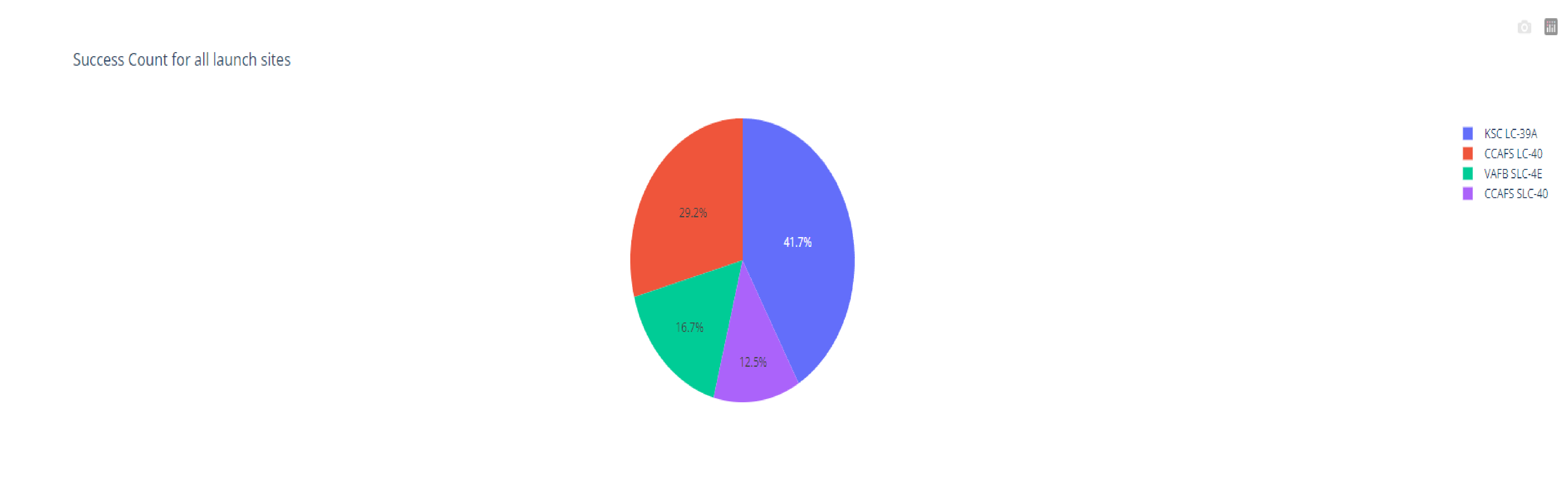


- Relative close to railway is 6 km, relative close to highway is 20.28 km, relative close to coastline is 14.99 km, relative close to its closest city Titusville is 16.32 km.

Section 4

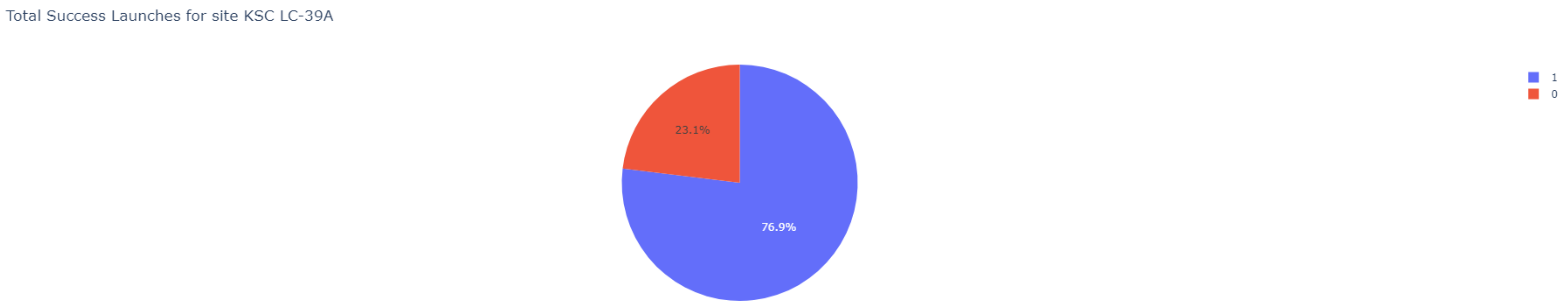
Build a Dashboard with Plotly Dash

Launch success count for all sites



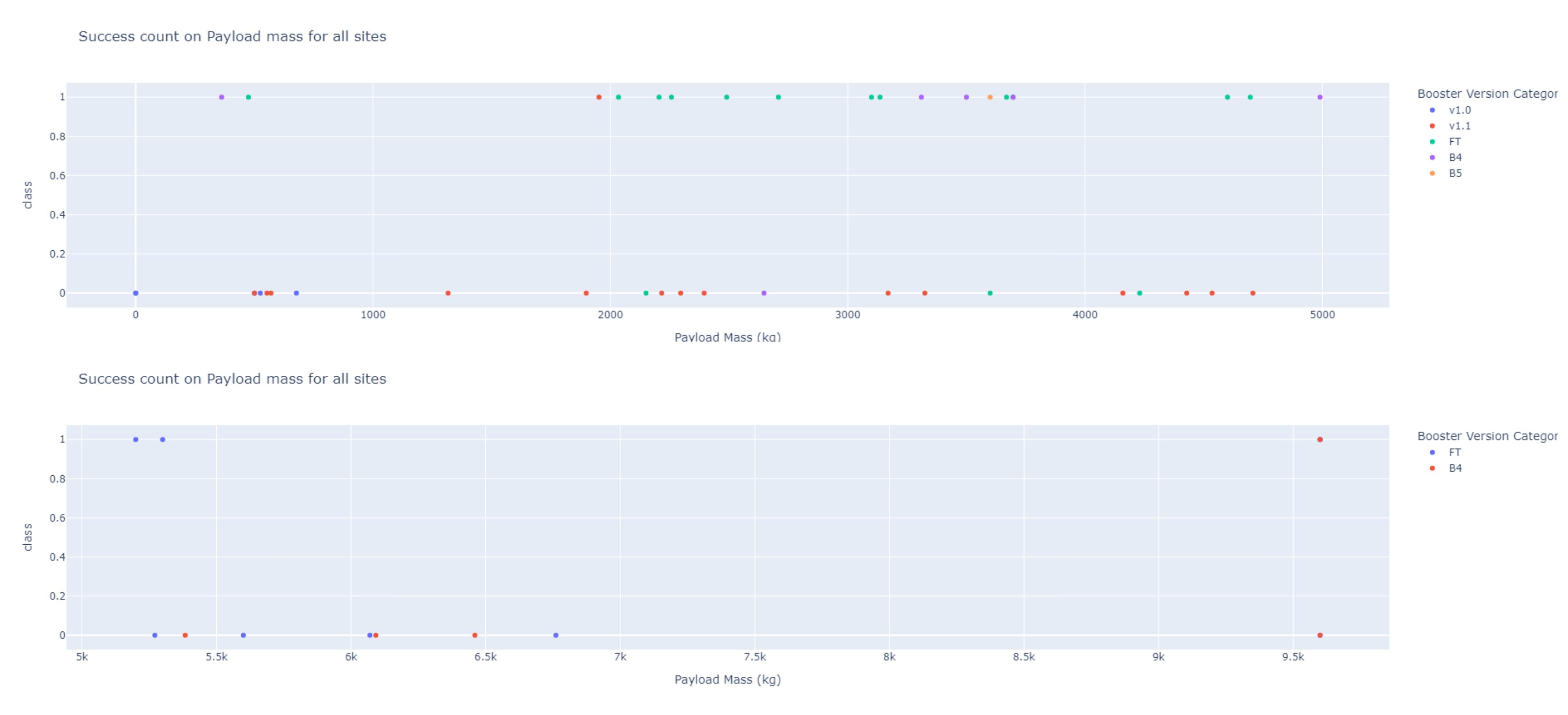
- Pie chart shows that from all sites, KSC LC 39-A has the most successful launches.

Launch site with highest launch success ratio



- KSC LC-39A has the highest launch success rate (76.9%).

Payload Mass vs. Launch Outcome for all sites



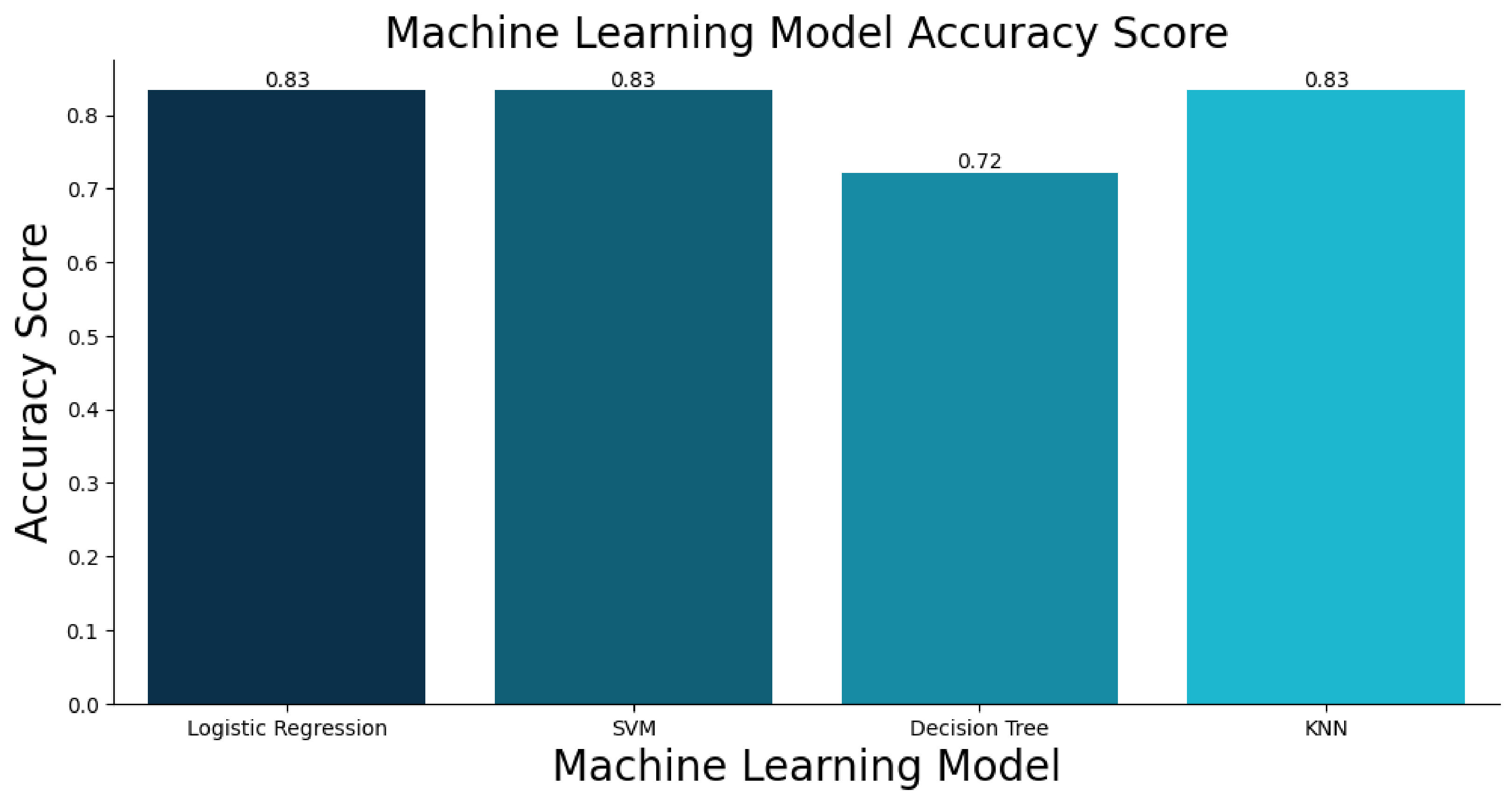
- The charts show that payloads between 2000 and 5500 kg have the highest success rate.

Section 5

Predictive Analysis (Classification)

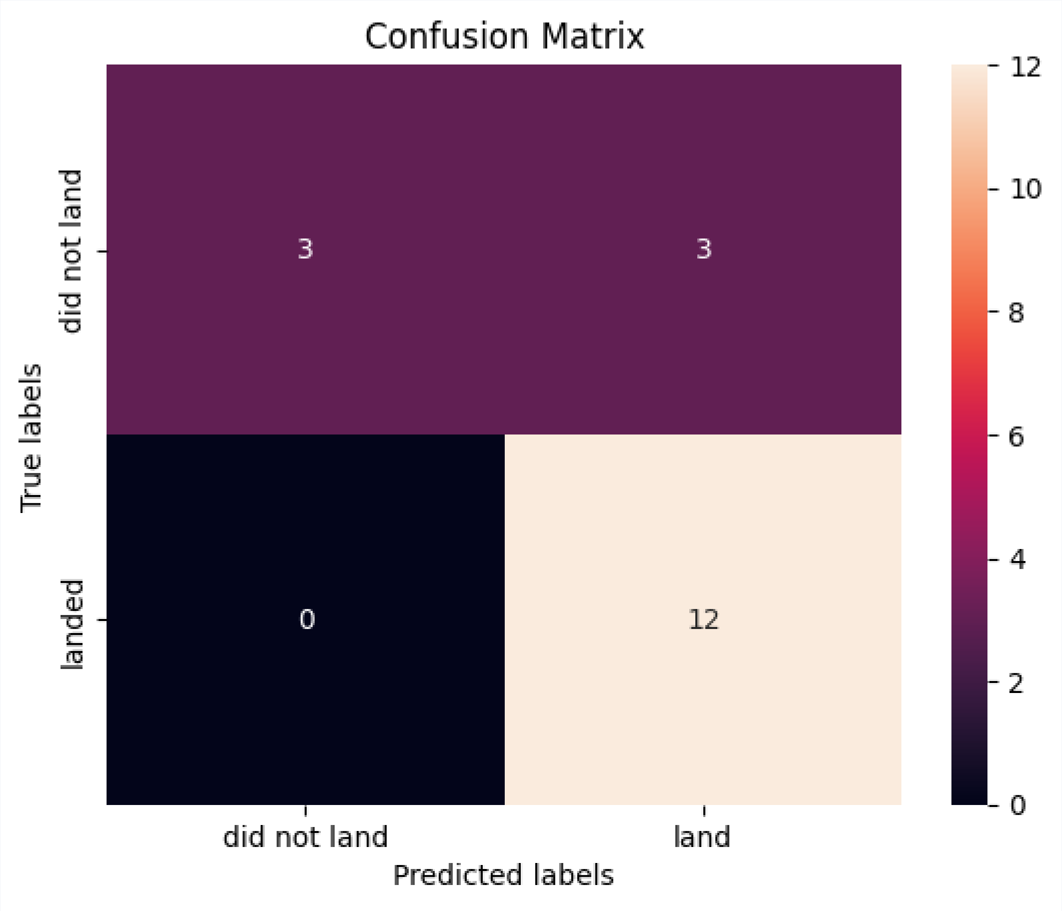
Classification Accuracy

- All models performed equally well except for the Decision Tree model, which has a lower score (0.72) than others.



Confusion Matrix

- This is the confusion matrix for he Logistic Regression model.
- There are 12 True Positives and 3 True Negatives (Correctly Predicted)
- There are 3 False Positives and 0 False Negatives (Falsely Predicted)



Conclusions

- SpaceX doesn't have perfect landing outcomes in Falcon 9 first stage landing trials.
- SpaceX's Falcon 9 trials are getting better with more launches.
- Machine learning models, in our case, Logistic Regression model can be used to predict future trials and their outcomes.

Appendix

Additional Datasets:

- SpaceX API (json): https://cf-courses-data.s3.us.cloud-object storage.appdomain.cloud/IBM-DS0321EN SkillsNetwork/datasets/API_call_spacex_api.json
- Wikipedia (Webpage): https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- Launch Geo (CSV): https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/spacex_launch_geo.csv
- Launch Dash (CSV): https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/spacex_launch_dash.csv

Main Datasets:

- GitHub URL (CSV 1): https://github.com/freudisapothead/ibmcapstoneproject/blob/main/dataset_part_1.csv
- GitHub URL (CSV 2): https://github.com/freudisapothead/ibmcapstoneproject/blob/main/dataset_part_2.csv
- GitHub URL (CSV 3): https://github.com/freudisapothead/ibmcapstoneproject/blob/main/dataset_part_3.csv
- GitHub URL (Launch Geo): https://github.com/freudisapothead/ibmcapstoneproject/blob/main/spacex_launch_geo.csv
- GitHub URL (Launch Dash): https://github.com/freudisapothead/ibmcapstoneproject/blob/main/spacex_launch_dash.csv

Appendix-2 (Continued)

Jupyter Notebooks and Dash .py File:

GitHub URL (Data Collection): https://github.com/freudisapothead/ibmcapstoneproject/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

GitHub URL (Web Scraping): https://github.com/freudisapothead/ibmcapstoneproject/blob/main/jupyter-labs-webscraping.ipynb

GitHub URL (Data Wrangling): https://github.com/freudisapothead/ibmcapstoneproject/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

GitHub URL (EDA with SQL): https://github.com/freudisapothead/ibmcapstoneproject/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

GitHub URL (EDA with Data Visualization): https://github.com/freudisapothead/ibmcapstoneproject/blob/main/jupyter-labs-eda-dataviz.ipynb

GitHub URL (Folium Maps): https://github.com/freudisapothead/ibmcapstoneproject/blob/main/lab_jupyter_launch_site_location.ipynb

GitHub URL (Dashboard File): https://github.com/freudisapothead/ibmcapstoneproject/blob/main/spacex_dash_app.py

GitHub URL (Machine Learning): https://github.com/freudisapothead/ibmcapstoneproject/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Thank you!