IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Daniel Martínez
22/08/2024

Outline

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

Executive Summary

- **Methodology Overview**
- Data was sourced from the SpaceX public API and publicly available information on Wikipedia.
- Data wrangling involved extracting launch outcome details to use as the dependent variable in the Machine Learning models.
- SQL queries and data visualizations (including static plots, interactive maps, and a dashboard) were developed to uncover insights and address key questions about the dataset.
- Predictive analysis was conducted using Logistic Regression, Support Vector Machine (SVM), Decision Tree, and k-Nearest Neighbors (KNN) Machine Learning models.
- **Results Overview**
- The launch data includes details such as flight number, launch date, payload mass, orbit type, launch site, mission outcome, and other variables.
- Logistic Regression, SVM, and KNN models all performed equally well in predicting outcomes within this dataset.

Introduction

A competing rocket launch company aims to forecast the success or failure of SpaceX Falcon 9 rocket first stage landings. To achieve this, several key questions must be addressed:

- What is the scope and quality of the available data on SpaceX Falcon 9 first stage landings?
- Which machine learning model would provide the highest accuracy in predicting the outcome of a Falcon 9 first stage landing from a future launch?
- Can we accurately predict whether a future Falcon 9 first stage landing will be successful?

Section 1

Methodology

Methodology

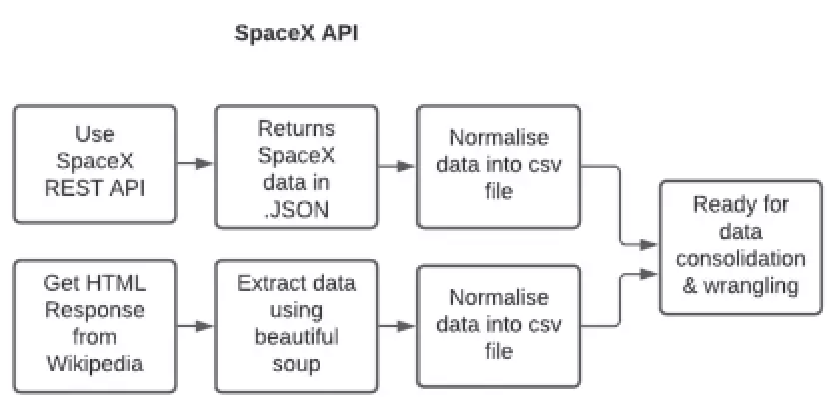
Executive Summary

- Data collection methodology:
  - Data was collected gathering SpaceX API and web scraping from Wikipedia
- Perform data wrangling
  - Data were transforming with one hot encoding to categorical features
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Exploratory data analysis (EDA) was done using visualization and SQL.
- Perform interactive visual analytics using Folium and Plotly Dash
  - Interactive visual analytics were developed using Folium and Plotly Dash.
- Perform predictive analysis using classification models
  - Predictive analysis was conducted using classification models.

and turn it into a Pandas dataframe using .json_normalize()

Data Collection

- Data were collected from the SpaceX API in json format and turn it into a Pandas dataframe using .json_normalize().
- The data from these requests will be stored in lists and will be used to create a new dataframe to construct our dataset using the data we have obtained. Then combine the columns into a dictionary.
- Then we dealing with missing values use the mean and the .replace() function to replace np.nan values in the data with the mean you calculated.



Data Collection – SpaceX API

```python
# Takes the dataset and uses the rocket column to call the API and append the data to the list
def getBoosterVersion(data):
    for x in data['rocket']:
       if x:
        response = requests.get("https://api.spacexdata.com/v4/rockets/"+str(x)).json()
        BoosterVersion.append(response['name'])
```

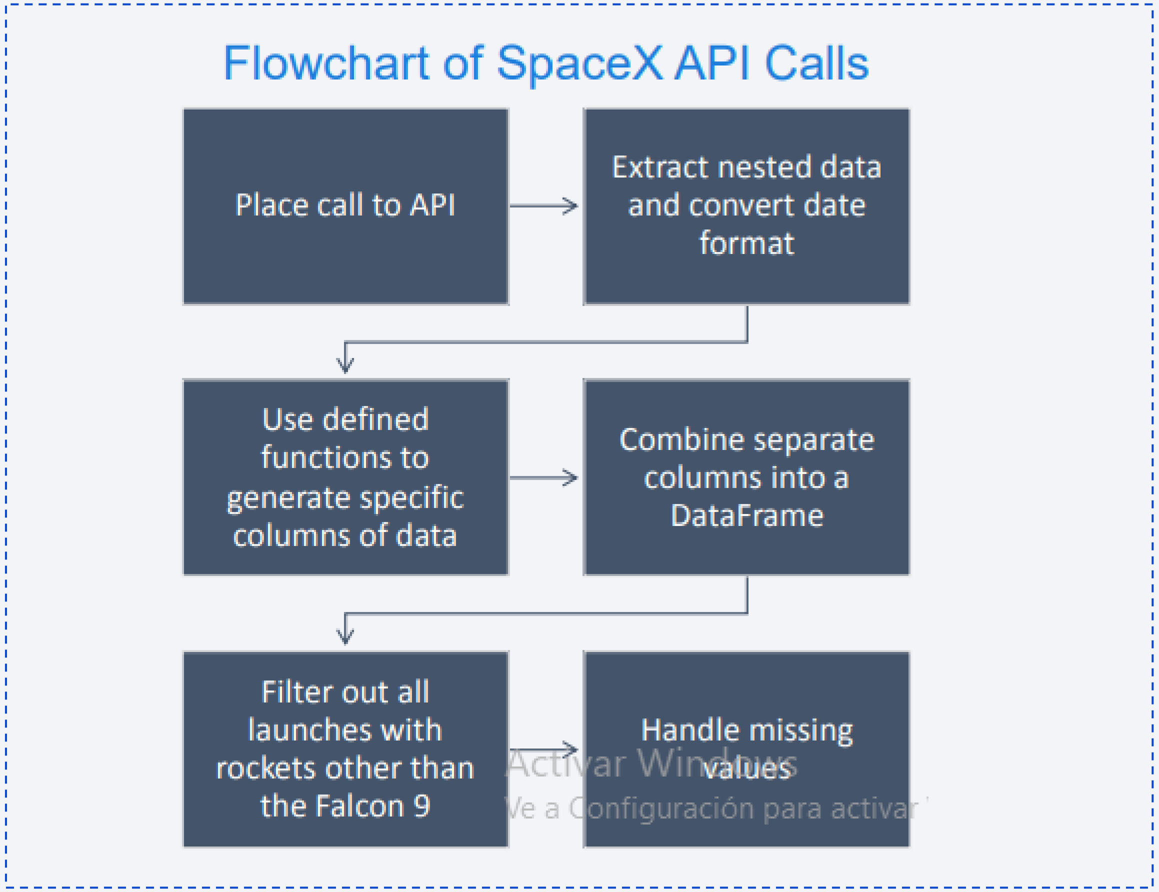
From the `launchpad` we would like to know the name of the launch site being used, the logitude, and the latitude.

```python
# Takes the dataset and uses the launchpad column to call the API and append the data to the list
def getLaunchSite(data):
    for x in data['launchpad']:
       if x:
         response = requests.get("https://api.spacexdata.com/v4/launchpads/"+str(x)).json()
         Longitude.append(response['longitude'])
         Latitude.append(response['latitude'])
         LaunchSite.append(response['name'])
```

From the `payload` we would like to learn the mass of the payload and the orbit that it is going to.

```python
# Takes the dataset and uses the payloads column to call the API and append the data to the lists
def getPayloadData(data):
    for load in data['payloads']:
       if load:
        response = requests.get("https://api.spacexdata.com/v4/payloads/"+load).json()
        PayloadMass.append(response['mass_kg'])
        Orbit.append(response['orbit'])
```

- https://github.com/lazarox10/IBM-Data-Science-Capstone-SpaceX/blob/main/1_jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

```python
# use requests.get() method with the provided static_url
# assign the response to a object
page = requests.get(static_url)
page.status_code
```

```
200
```

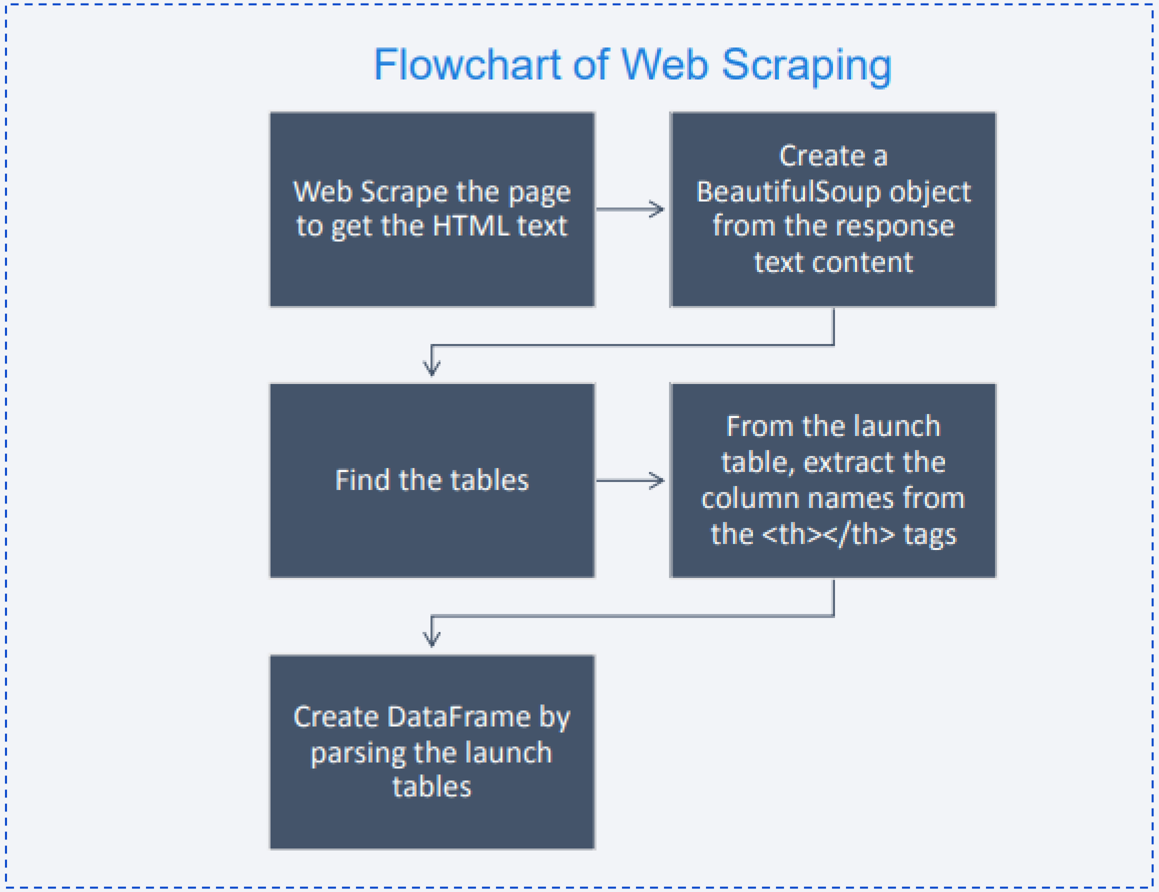
Create a `BeautifulSoup` object from the HTML `response`

```python
# Use BeautifulSoup() to create a BeautifulSoup object from a response text content
soup = BeautifulSoup(page.text, 'html.parser')
```

Print the page title to verify if the `BeautifulSoup` object was created properly

```python
# Use soup.title attribute
soup.title
```

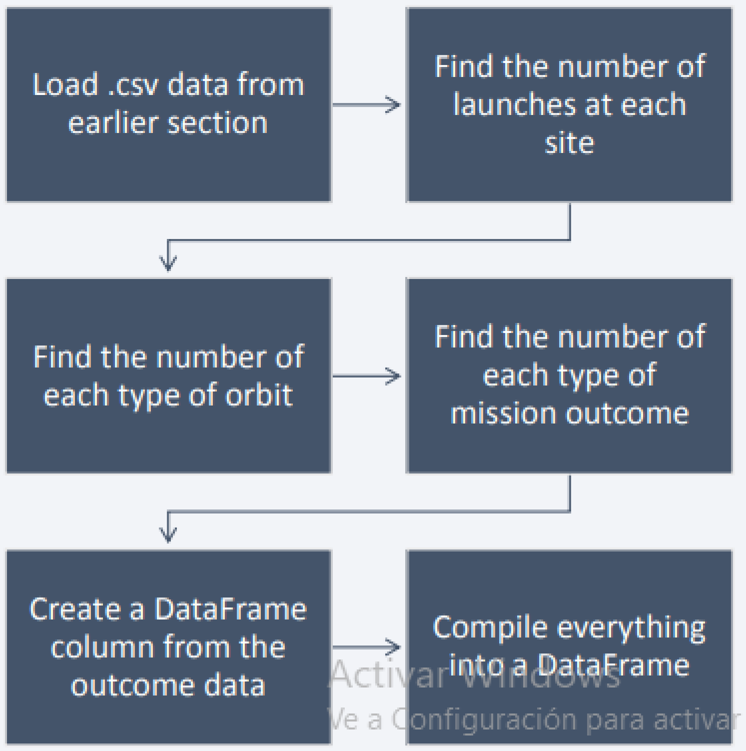
- https://github.com/lazarox10/IBM-Data-Science-Capstone-SpaceX/blob/main/2_jupyter-labs-webscraping.ipynb



Data Wrangling

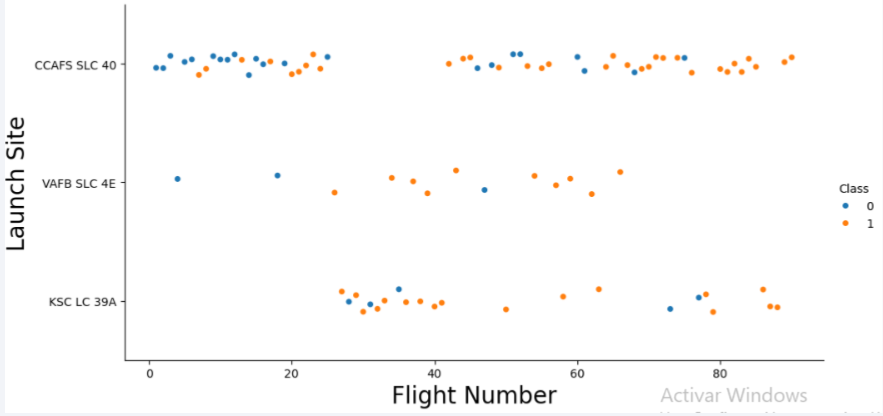
- The initial dataset was stored in a .csv file, which required preprocessing.
- Data cleanup focused on refining launch locations, orbital classifications, and mission results.
- Mission outcomes were simplified into a binary system: successful Falcon 9 first stage landings were assigned a value of 1, while unsuccessful attempts were marked as 0.
- This new binary classification was incorporated into the DataFrame to facilitate subsequent analysis.
- https://github.com/lazarox10/IBM-Data-Science-Capstone-SpaceX/blob/main/3_labs-jupyter-spacex-Data%20wrangling.ipynb

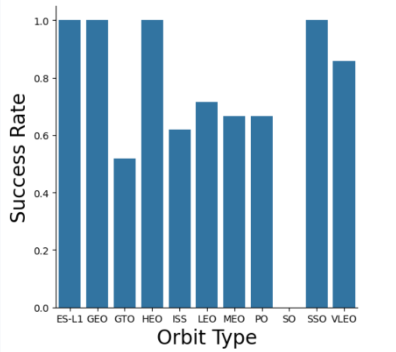
Flowchart of Data Wrangling

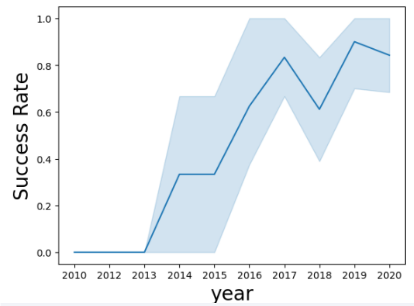


EDA with Data Visualization

- Scatterplot to see mission outcome relationship split by Launch Site and Flight Number.
- Scatterplot to see mission outcome relationship split by Launch Site and Payload.
- Bar chart to see mission outcome relationship with Orbit Type.
- Scatterplot to see mission outcome relationship split by Orbit Type and Flight Number.
- Scatterplot to see mission outcome relationship split by Orbit Type and Payload.
- Line plot to see mission outcome trend by year
- https://github.com/lazarox10/IBM-Data-Science-Capstone-SpaceX/blob/main/5_jupyter-labs-eda-dataviz.ipynb







EDA with SQL

- Summarize of queries:
  - Launch sites
  - Payload masses
  - Dates
  - Booster types
  - Mission outcomes
- https://github.com/lazarox10/IBM-Data-Science-Capstone-SpaceX/blob/main/4_jupyter-labs-eda-sql-coursera_sqllite.ipynb

```
%sql select DISTINCT LAUNCH_SITE from SPACEXTBL;
```

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

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

```
%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

Build an Interactive Map with Folium

- **Geospatial Visualization Elements Incorporated into Folium Map:**

1. Point Data Representation: • Launch site locations denoted by markers • NASA Johnson Space Center indicated with a distinct marker
2. Area of Interest Demarcation: • Launch sites emphasized using circular overlays
3. Proximity Analysis Visualization: • Linear features added to illustrate distances to critical infrastructure:
    1. CCAFS LC-40 to coastline
    2. CCAFS LC-40 to railway network
    3. CCAFS LC-40 to perimeter access road

- This cartographic representation integrates multiple spatial data types to provide a comprehensive overview of launch site locations and their spatial relationships to key geographical features.
- https://github.com/lazarox10/IBM-Data-Science-Capstone-SpaceX/blob/main/6_lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- **Interactive Data Visualization Components:**

1. User Input Controls: • Dropdown menu: Facilitates selection of individual or aggregate launch site data • Slider interface: Enables filtration of payload mass range
2. Pie Chart Visualization: • Aggregate view: Illustrates the distribution of successful Falcon 9 first stage landings across all sites • Site-specific view: Depicts the ratio of successful to failed Falcon 9 first stage landings for the selected site
3. Scatterplot Analysis: • X-axis: Payload mass (filtered via slider input) • Y-axis: Mission outcome (binary success/failure classification) • Data points: Categorized by booster version • Visualization: Illustrates the relationship between payload mass, mission outcome, and booster version across the filtered dataset

This suite of interactive visualizations allows for dynamic exploration of Falcon 9 launch data, facilitating multi-dimensional analysis of mission outcomes in relation to key variables such as launch site, payload mass, and booster version.

- https://github.com/lazarox10/IBM-Data-Science-Capstone-SpaceX/blob/main/spacex_dash_app.py

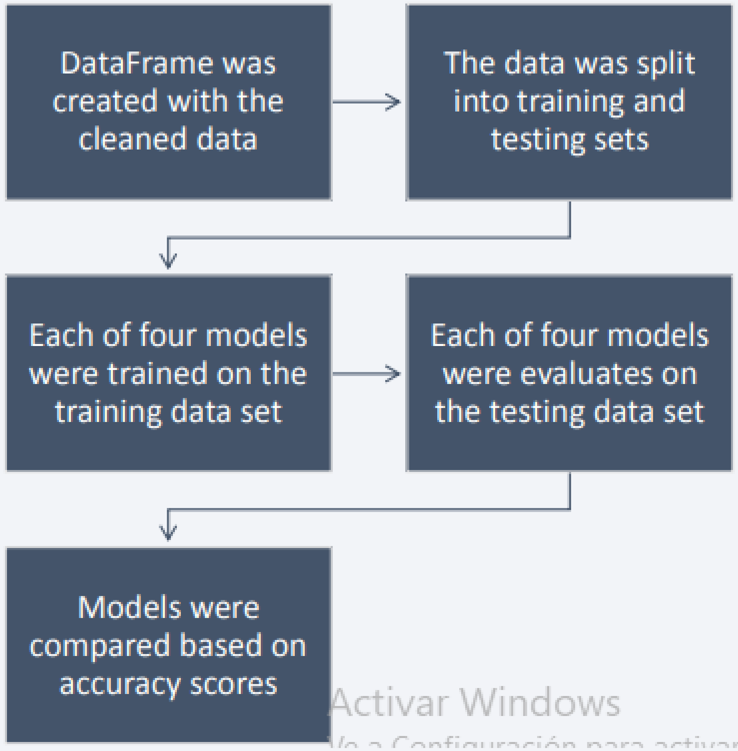
Predictive Analysis (Classification)

- **Machine Learning Model Development and Evaluation Process**:

1. Data Partitioning: • Implementation of train-test split methodology to create distinct datasets for model training and evaluation
2. Model Selection and Training: • Deployment of multiple supervised learning algorithms:
    1. Logistic Regression
    2. Support Vector Machine (SVM)
    3. Decision Tree
    4. k-Nearest Neighbors (KNN) • Training of selected models using the designated training dataset
3. Hyperparameter Optimization: • Utilization of GridSearchCV() for exhaustive hyperparameter tuning • Identification of optimal hyperparameter configurations via '.best_params_' attribute
4. Model Performance Assessment: • Application of optimized models to the hold-out test dataset • Evaluation of model performance using accuracy as the primary metric • Comparative analysis of predictive capabilities across all four algorithmic approaches

- This systematic approach to model development and evaluation ensures robust performance assessment and facilitates the selection of the most appropriate predictive model for the given dataset and problem domain.

Flowchart of Machine Learning



https://github.com/lazarox10/IBM-Data-Science-Capstone-SpaceX/blob/main/7_SpaceX%20Machine%20Learning%20Prediction%20v2.ipynb

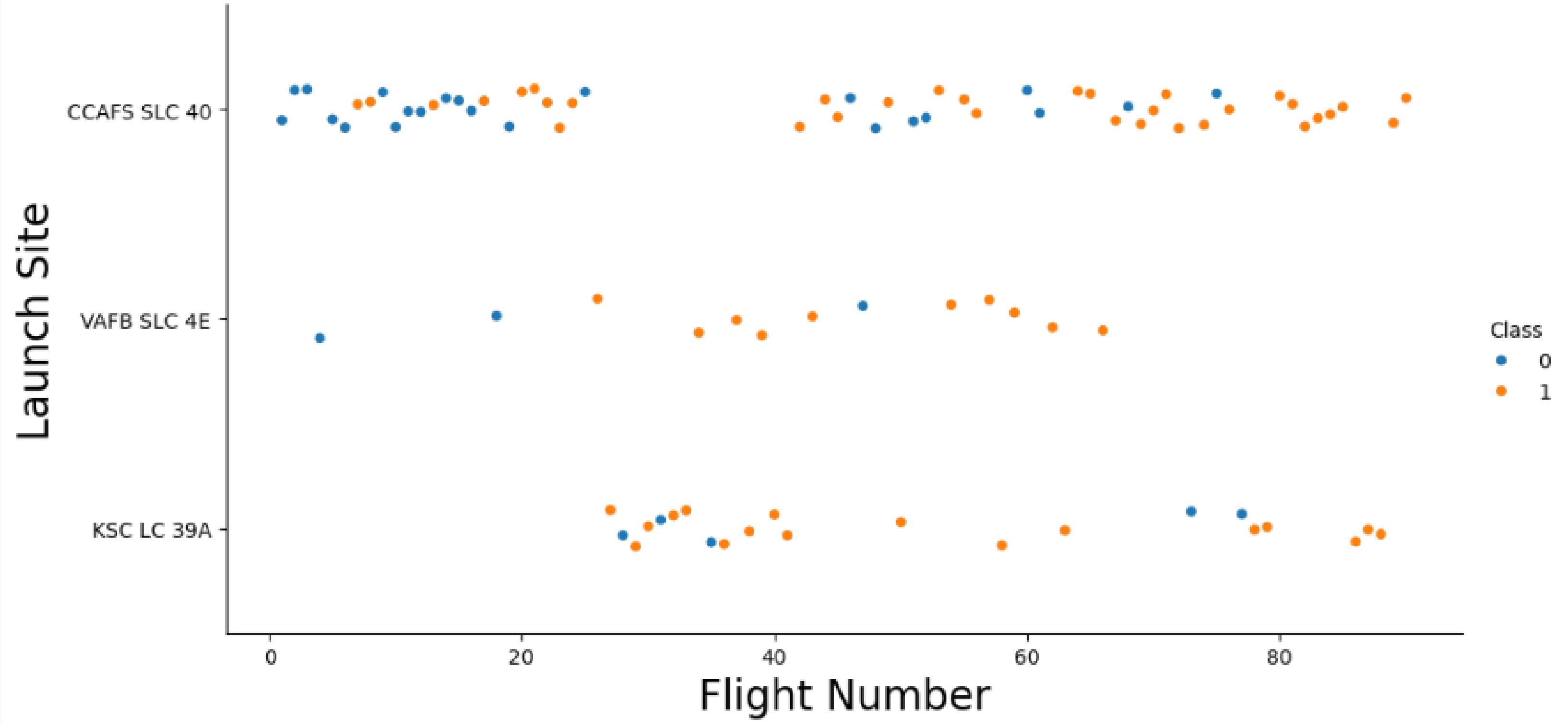
Results

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

Section 2

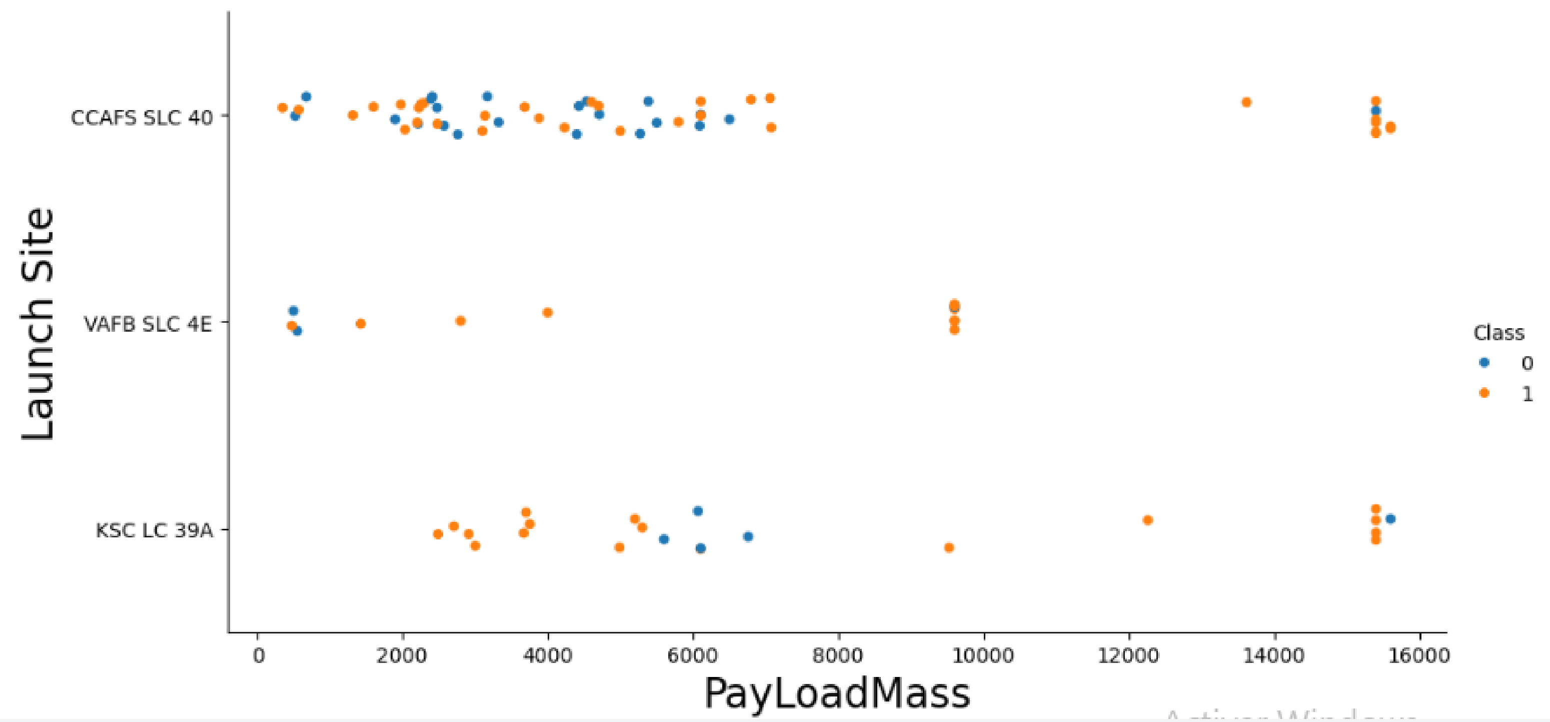
Insights drawn from EDA

Flight Number vs. Launch Site



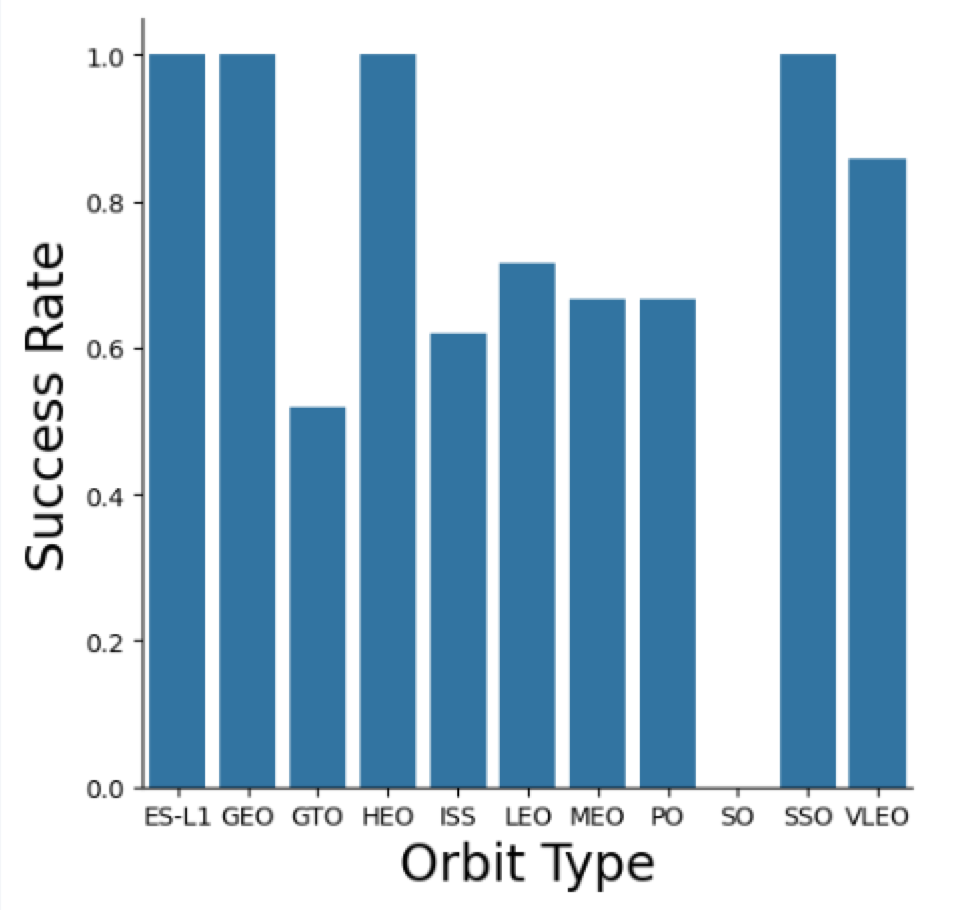
- Success rate (1) increases with flight numbers
- CCAFS SLC 40 has a higher rate of success

Payload vs. Launch Site



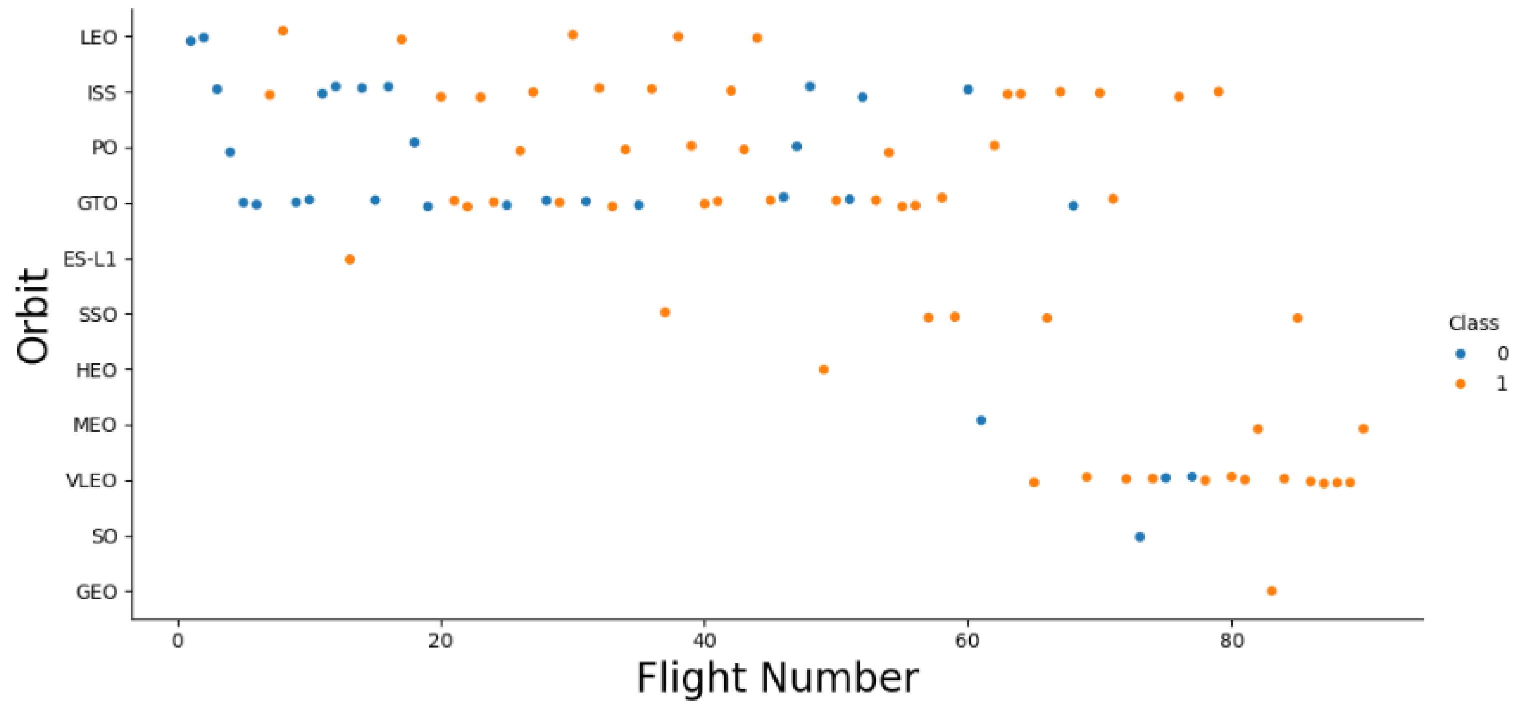
- When the Payload mass increases, the success rate is increasing

Success Rate vs. Orbit Type



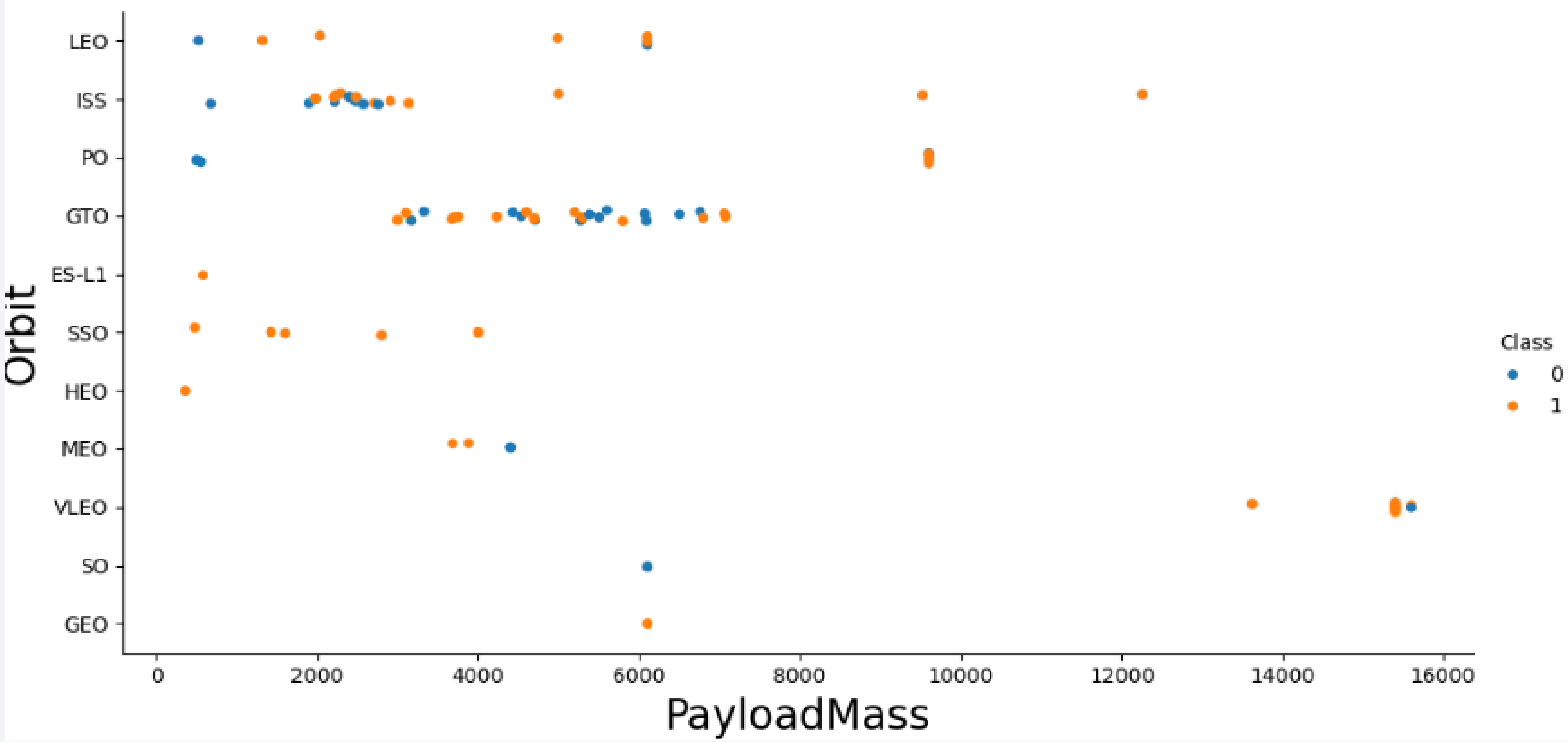
- ES-L1, GEO, HEO and SSO orbits has an 100% success rate of each orbit type
- SO Orbit has an 0% of success rate

Flight Number vs. Orbit Type



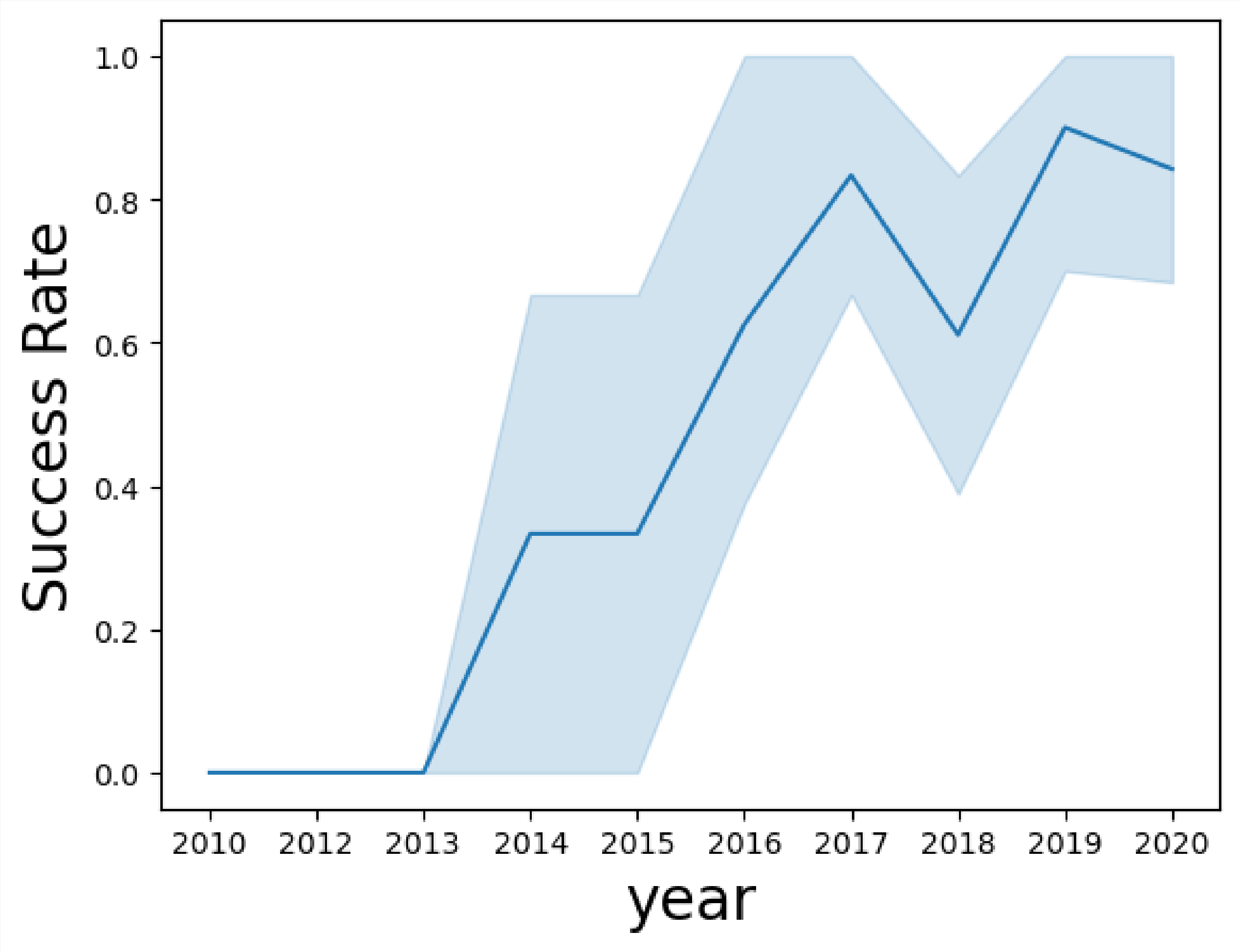
- There is a correlation between flight numbers and landing success, however is better a barchart to determinate the most success orbits.

Payload vs. Orbit Type



- There is not a obvious correlation between PayloadMass and Orbit

Launch Success Yearly Trend



- Success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing.

All Launch Site Names

```
%sql select DISTINCT LAUNCH_SITE from SPACEXTBL;
```

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

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

- Results shows four unique launch sites

Launch Site Names Begin with '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

- This query shows the first five sites contained in database that start with ‘CCA’

Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) as PAYLOADMASS from SPACEXTBL where (CUSTOMER) like 'NASA (CRS)';
```

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

PAYLOADMASS
45596

- Total payload carried by NASA (CRS) is 48,213 kg.

Average Payload Mass by F9 v1.1

```
%sql SELECT sum(payload_mass__kg_) / count(payload_mass__kg_) AS 'Average Payload Mass (kg)' FROM SPACEXTBL WHERE booster_version LIKE 'F
```

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

Average Payload Mass (kg)
2928

- The average payload mass carried by booster version F9 v1.1is 2928 kg

First Successful Ground Landing Date

```
%sql SELECT min(DATE) AS 'First Successful Landing Outcome Date' FROM SPACEXTBL WHERE Landing_Outcome LIKE 'Success (ground pad)';
```

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

First Successful Landing Outcome Date
2015-12-22

- First successful landing outcome on ground pad was on 2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%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.
```

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

- There are four booster versions that have successfully landed on drone ship with a payload mass greater than 4,000 kg but less than 6,000 kg

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
%sql SELECT (SELECT count(*) FROM SPACEXTBL WHERE landing_outcome LIKE '%success%') AS 'Success', count(*) AS "Failure" FROM SPACEXTBL WH
```

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

Success	Failure
61	40

- Accord the query there were 61 successful and 40 failed mission outcomes

Boosters Carried Maximum Payload

```
%sql select BOOSTER_VERSION as boosterversion from SPACEXTBL where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

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

boosterversion
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

- There are 12 Falcon 9 boosters carried the maximum payload mass

2015 Launch Records

```
%sql SELECT strftime('%m', DATE) AS 'Month', landing_outcome, booster_version, launch_site FROM SPACEXTBL WHERE landing_outcome = 'Failu
```

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

- Two drone ship landing attempts in 2015 resulted in failure, both of which were launched from CCAFS LC-40

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

```
%sql SELECT landing_outcome, count(landing_outcome) AS "Count" FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY 1
```

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

- The most frequent landing outcome was ‘not attempted’.

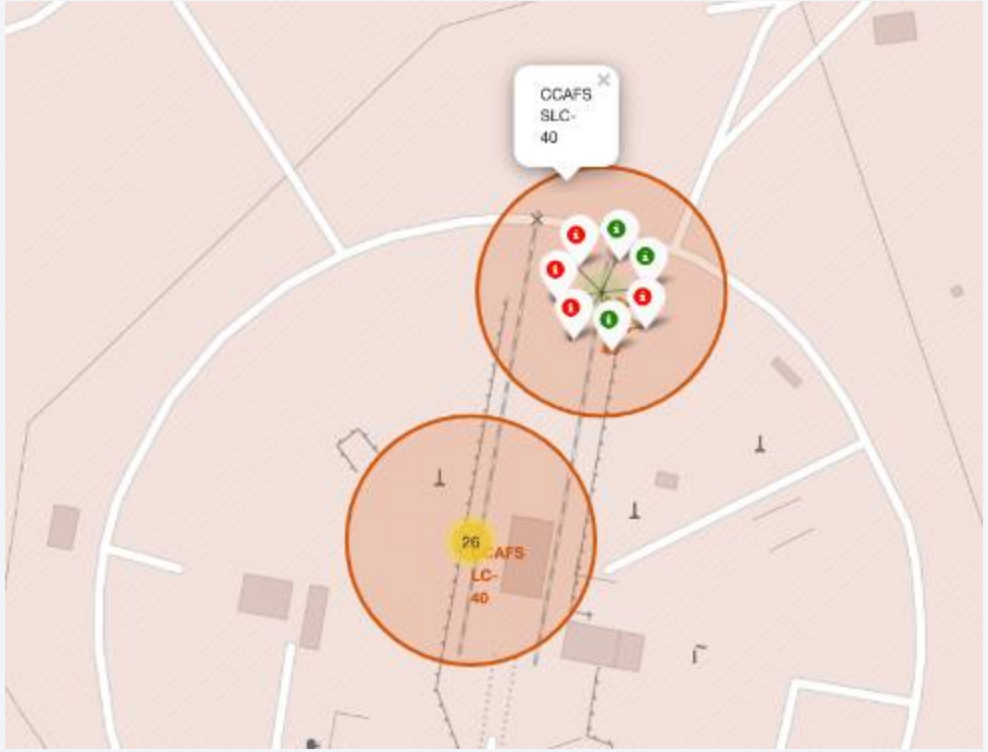
Section 3

Launch Sites Proximities Analysis

All launches Sites markers on global maps

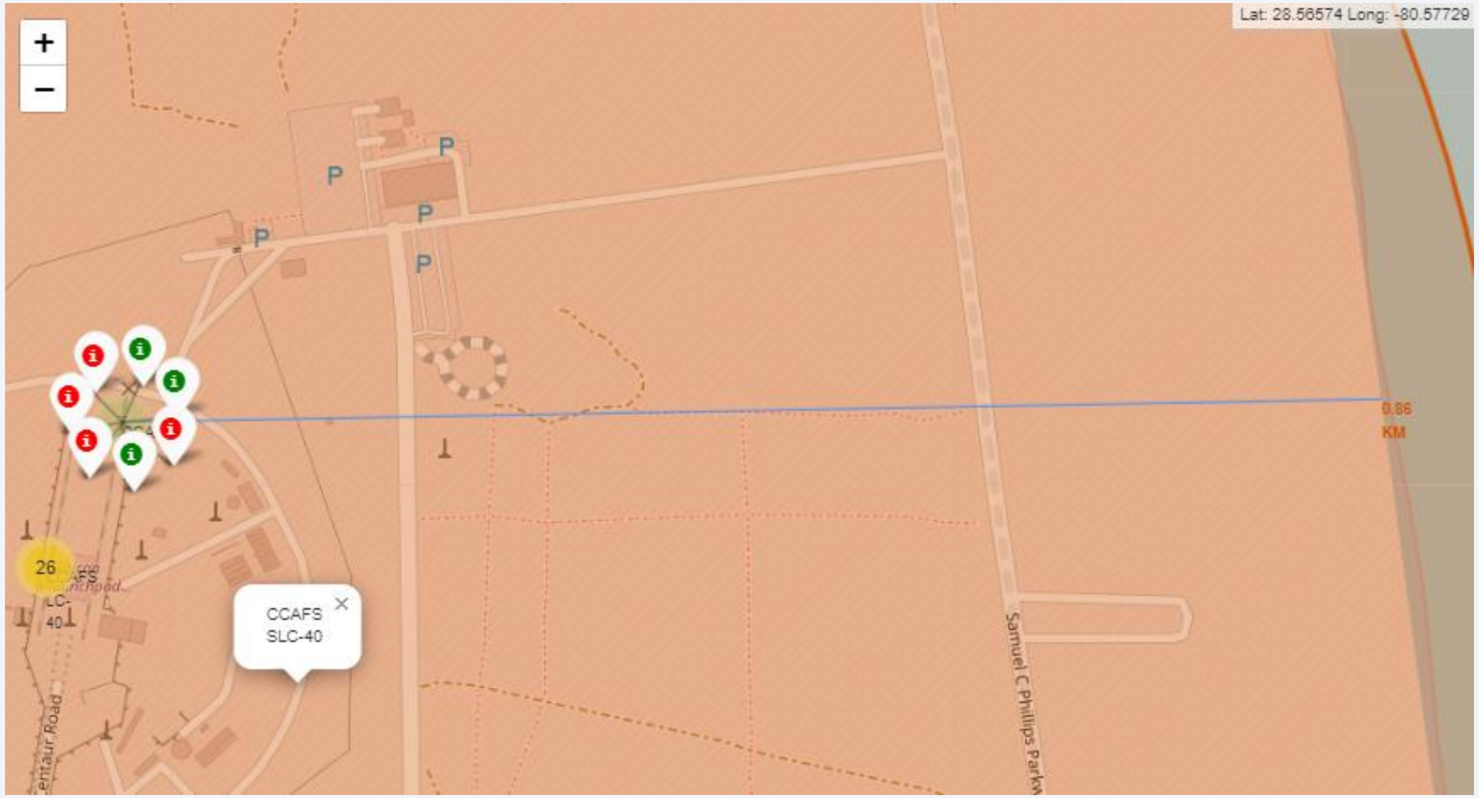
- Launchers are near coasts in USA

Success/Failed launches on the map



- The graphic shows in green marker if a launch was successful, otherwise in red if was a failure

Distances between launch sites

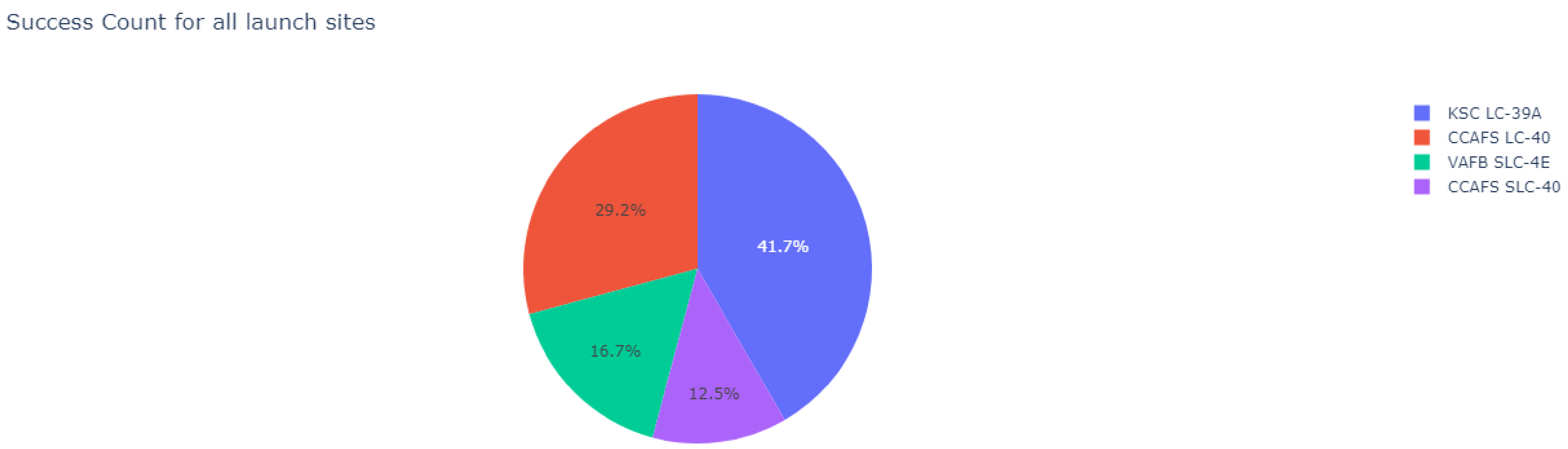


- Launch sites are near to railways

Section 4

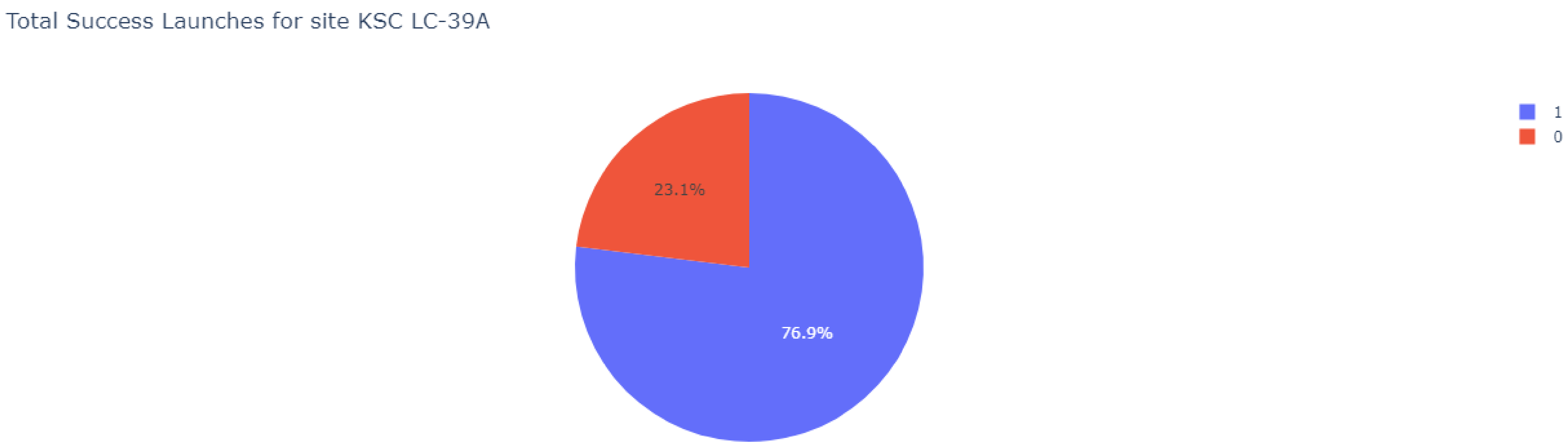
Build a Dashboard with Plotly Dash

Launch Success Count



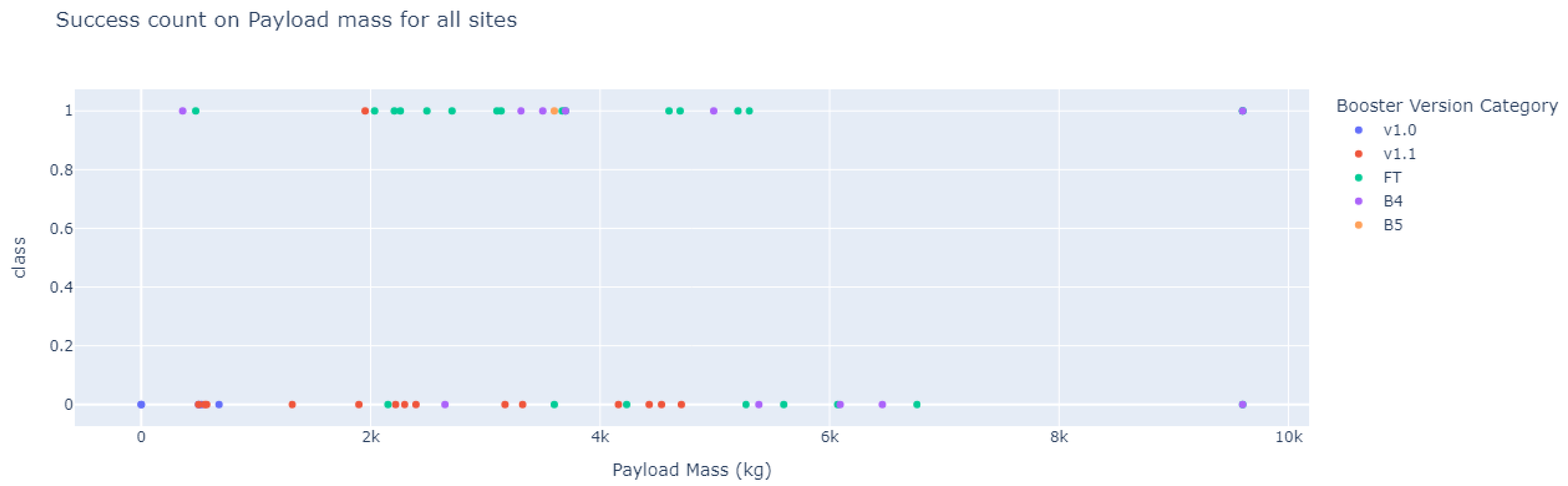
- SC LC-39A has the highest score with 41.7%, next CCAFS LC-40 with 29.2%, then VAFB SLC-4E with 16.7% and finally CCAFS SLC-40 with 12.5%

Launch Site with Higher Score



- KSC LC -39A achieved a 76.9% of success landing while his failure rate is on 23.1%

Payload vs Launch Outcome

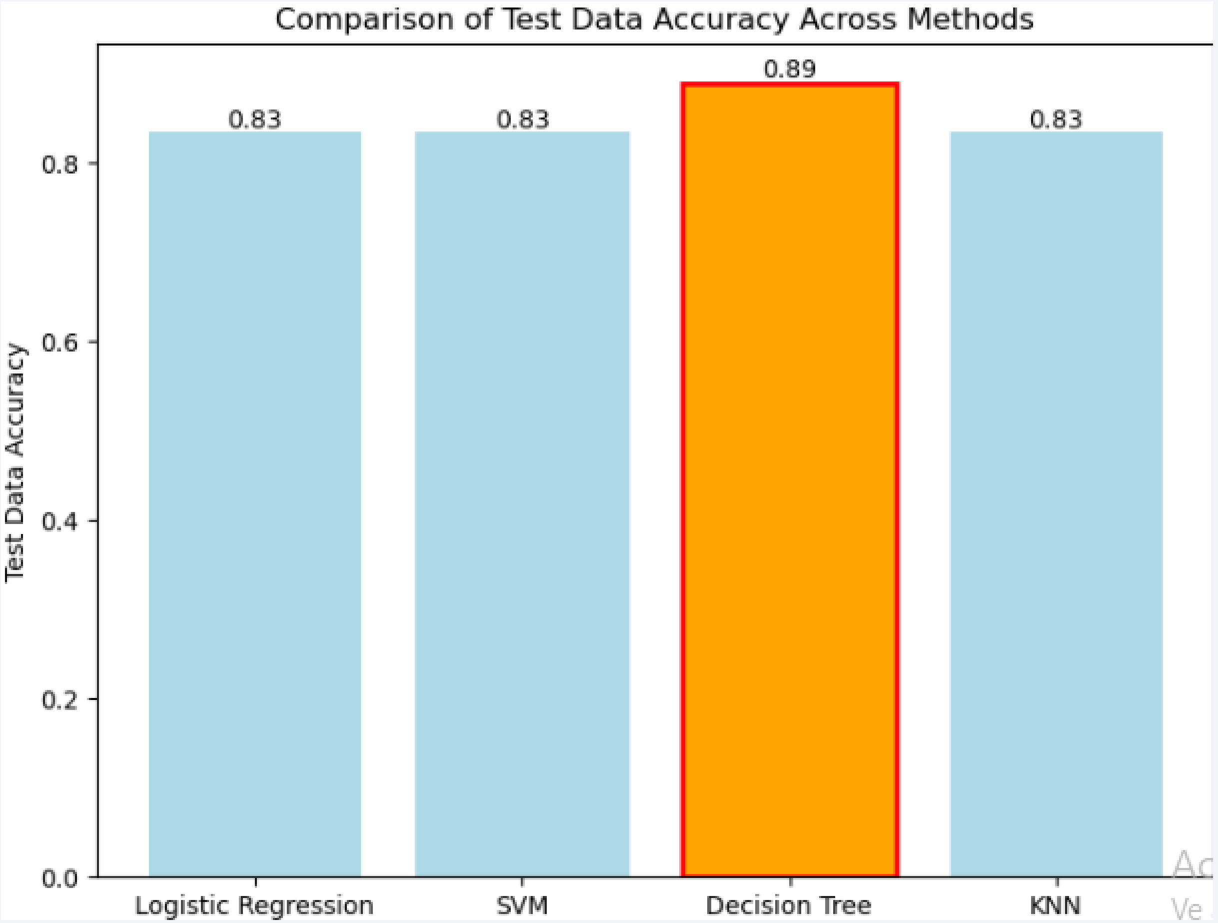


- Payloads ranging from approximately 2,000 kg to 5,000 kg have the highest success rate.
- The 'FT' booster version category boasts the highest success rate among all versions

Section 5

Predictive Analysis (Classification)

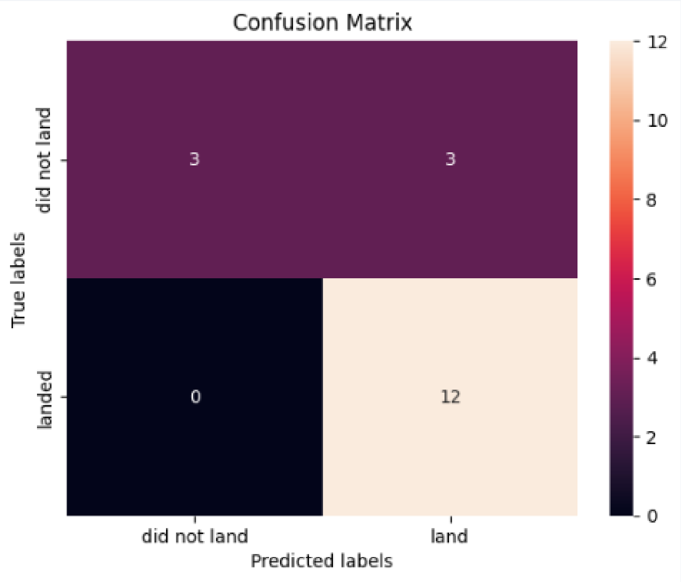
Classification Accuracy



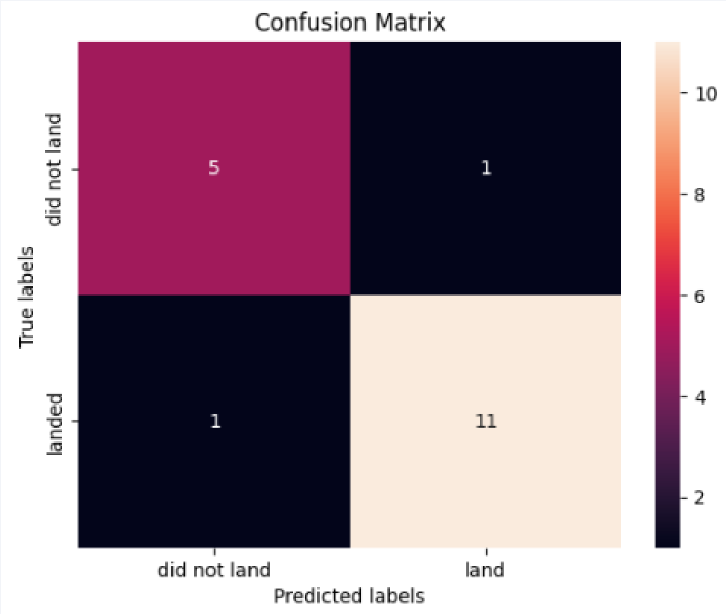
- The model with the highest classification accuracy is Decision Tree with a value of = 0.89

Confusion Matrix

LOG_REG, SVM, KNN



Decision Tree



- According to the confusion matrix, it allows us to evaluate the quality of the classification model, showing better results with the Decision Tree.

Conclusions

- **Model Performance:** The Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Logistic Regression models demonstrated the highest prediction accuracy for this dataset, making them the most reliable models for predicting outcomes in this context.
- **Payload Impact:** Launches with lower payload weights tend to have higher success rates compared to those carrying heavier payloads, suggesting that lighter payloads contribute to better overall performance.
- **Launch Success Over Time**: The success rates of SpaceX launches show a positive correlation with the passage of time, indicating that continuous improvements and experience over the years are leading to increasingly successful missions.
- **Launch Site Success:** The Kennedy Space Center Launch Complex 39A (KSC LC 39A) stands out as the most successful launch site among all SpaceX facilities, with the highest number of successful missions.
- **Orbit Success Rates**: Orbits such as GEO (Geostationary Earth Orbit), HEO (Highly Elliptical Orbit), SSO (Sun-Synchronous Orbit), and ES L1 (Earth-Sun Lagrange Point 1) have exhibited the highest success rates, suggesting that these orbits are particularly favorable for successful launches.

Appendix

- All the code related to this project you can found on: https://github.com/lazarox10/IBM-Data-Science-Capstone-SpaceX

Thank you!