IBM DATA SCIENCE CAPSTONE

Trenton Brasel
5/24/23

Outline

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

Executive Summary

- The data was acquired from the public Space API and the Space Wikipedia page.
- Explored the data through SQL queries, visualization techniques, folium maps, and dashboards.
- Extracted relevant columns as features for analysis.
- Converted categorical variables into binary format using one-hot encoding.
- Standardized the data and employed GridSearchCV to identify optimal parameters for the machine learning models.
- Visualized the accuracy scores of all the models.
- Four machine learning models were generated: Logistic Regression, Support Vector Machine, Decision Tree Classifier and K Nearest Neighbors.
- All of the machine learning models yielded similar results, achieving an accuracy rate of approximately 83%.
- However, all models tended to over-predict successful landings.
- Obtaining more data would enhance model determination and improve accuracy.

Introduction

• Space X (Falcon 9) offers the most competitive pricing at $62 million compared to the higher cost of $165 million USD for other providers.

• This pricing advantage is largely attributed to Space X's capability to recover a portion of the rocket (Stage 1).

• In This Project Space Y aims to rival Space X's market position.

Challenge:

• Space Y has approached us with the following tasks:

- Estimating the launch price for each mission.
- Collecting publicly available information about Space X and creating dashboards for the Space Y team.
- Developing a machine learning model to predict the success of Stage 1 recovery

Section 1

Methodology

Methodology

Data Collection Method

- Merged data obtained from the Space public API and the Space Wikipedia page.
- Conducted data cleaning and preprocessing.
- Classified landings as successful if true, and otherwise as unsuccessful.
- Conducted exploratory data analysis (EDA) utilizing visualization techniques

SQL.

- Utilized Folium and Plotly Dash for interactive visual analytics.
- Performed predictive analysis using classification models.
- Optimized the models using GridSearchCV for parameter tuning.

-

Data Collection

- The data collection process involved a combination of API requests made to the Space public API and web scraping data from a table within Space X's Wikipedia entry. The following slide will illustrate the sequence of processing the data retrieved from the Space public API, while the subsequent slide will depict the sequence of processing the data obtained through web scraping.

- Wikipedia Web Scraping Data Columns:
  - Flight No.
  - Launch site
  - Payload
  - PayloadMass
  - Orbit
  - Customer
  - Launch outcome
  - Version Booster
  - Booster landing
  - Date
  - Time

- Space X API Data Columns:
  - FlightNumber
  - Date
  - BoosterVersion
  - PayloadMass
  - Orbit
  - LaunchSite
  - Outcome
  - Flights
  - GridFins
  - Reused
  - Legs
  - LandingPad
  - Block
  - ReusedCount
  - Serial
  - Longitude
  - Latitude

Data Collection – SpaceX API

- https://github.com/Tcby04/Applied-Data-Science-Capstone/blob/main/Data_Collection_API.ipynb

- Request Api’s
- .JSON File+ List
- Json_Normalize To DF
- Dictionary Data That’s Relevant
- Dictionary To DF
- Filter Data For only Falcon 9
- Replace Missing Payloads with Mean

Data Collection - Scraping

- https://github.com/Tcby04/Applied-Data-Science-Capstone/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb

1) Sent a request to retrieve the HTML content of the Wikipedia page.

2) Utilized BeautifulSoup with the html5lib parser to parse the HTML content.

3) Located the HTML table containing the launch information.

4) Converted the table data into a dictionary format.

5) Iterated through the cells of the table to extract the data and populate the dictionary.

6) Create Dictionary

Data Wrangling

https://github.com/Tcby04/Applied-Data-Science-Capstone/blob/main/Data%20Wrangling.ipynb

- Generate a training label based on the landing outcomes, assigning a value of 1 for successful landings and 0 for failures. The 'Outcome' column consists of two components: 'Mission Outcome' and 'Landing Location'. Create a new column called 'class' which will have a value of 1 if 'Mission Outcome' is True, and 0 otherwise.

- For the value mapping:

- - If 'Mission Outcome' is True and 'Landing Location' is either 'ASDS', 'RTLS', or 'Ocean', assign a value of 1 to the 'class' column.

- - If 'Mission Outcome' is either 'None' or False, and 'Landing Location' is either 'None', 'ASDS', 'Ocean', or 'RTLS', assign a value of 0 to the 'class' column.

- Furthermore, in cases where 'Mission Outcome' is False and 'Landing Location' is either 'ASDS' or 'Ocean', assign a value of -3 to the 'class' column.

EDA with Data Visualization

https://github.com/Tcby04/Applied-Data-Science-Capstone/blob/main/EDA%20with%20Data%20Visualization.ipynb

- An Exploratory Data Analysis (EDA) was conducted on the following variables: Flight Number, Payload Mass, Launch Site, Orbit, Class, and Year. The objective was to examine the relationships between these variables and determine if they can be utilized in training the machine learning model. Several types of plots were employed during the analysis, including scatter plots, line charts, and bar plots. The specific plots used are as follows:

- 1. Flight Number vs. Payload Mass: A scatter plot was created to visualize the relationship between Flight Number and Payload Mass.

- 2. Flight Number vs. Launch Site: A scatter plot was generated to explore the relationship between Flight Number and Launch Site.

- 3. Payload Mass vs. Launch Site: A scatter plot was utilized to examine the relationship between Payload Mass and Launch Site.

- 4. Orbit vs. Success Rate: A bar plot was employed to illustrate the success rate for different Orbit types.

- 5. Flight Number vs. Orbit: A scatter plot was employed to analyze the relationship between Flight Number and Orbit.

- 6. Payload vs. Orbit: A scatter plot was created to investigate the relationship between Payload and Orbit.

- 7. Success Yearly Trend: A line chart was generated to visualize the yearly trend in the success rate.

- By studying these plots, it was determined whether any relationships existed between the variables, thereby assisting in the selection of relevant features for the machine learning model training.

EDA with SQL

https://github.com/Tcby04/Applied-Data-Science-Capstone/blob/main/EDA%20with%20SQL.ipynb

- The dataset was successfully loaded into the IBM DB2 Database. SQL Python integration was utilized to execute queries and gain a comprehensive understanding of the dataset. The following queries were executed to extract relevant information:
- 1. Launch site names: A query was made to retrieve the names of the launch sites.
- 2. Mission outcomes: Information about mission outcomes, such as successful or unsuccessful launches, was queried.
- 3. Various payload sizes of customers: Queries were executed to obtain details about payload sizes for different customers.
- 4. Booster versions: Information about the versions of boosters used in the launches was queried.
- 5. Landing outcomes: Queries were performed to retrieve data on landing outcomes, including successful landings or any failures.

Build an Interactive Map with Folium

https://github.com/Tcby04/Applied-Data-Science-Capstone/blob/main/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

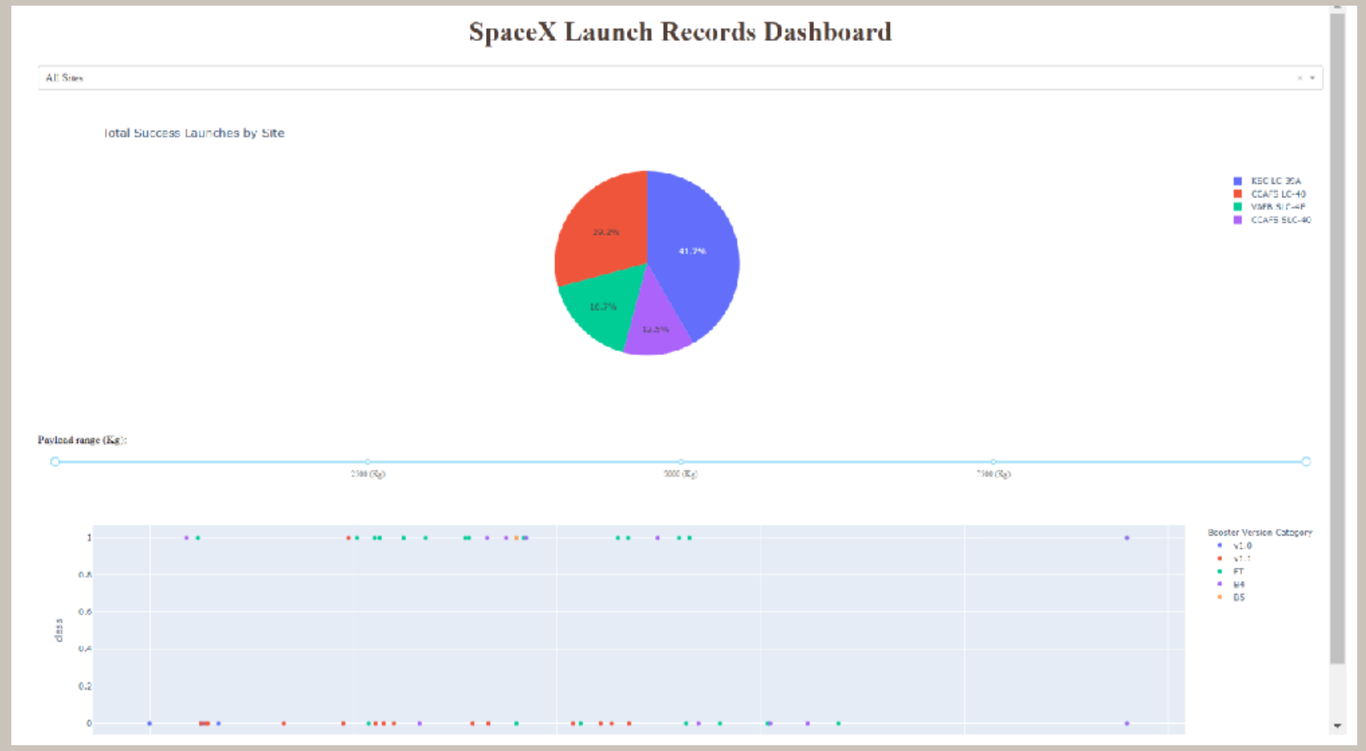
- Analysis of Launch Site Locations with Folium Maps:
- - Folium maps used to analyze launch site locations
- - Launch sites, successful and unsuccessful landings marked on the maps
- - Proximity examples shown: Railway, Highway, Coast, and City
- Insights gained:
- - Understanding of why launch sites are located where they are
- - Visualization of successful landings in relation to location
- Benefits of the analysis:
- - Identifying patterns and factors influencing launch site selection
- - Insight into the importance of accessibility to transportation infrastructure
- - Consideration of proximity to coastlines and nearby cities

Predictive Analysis (Classification)

https://github.com/Tcby04/Applied-Data-Science-Capstone/blob/main/Machine%20Learning%20Prediction.ipynb

- 1. Split the label column 'Class' from the dataset
- 2. Fit and transform the features using Standard Scaler to standardize the data
- 3. Train, test, and split the data into training and testing sets
- 4. Utilize GridSearchCV with a cross-validation of 10 folds to find optimal parameters for each model
- 5. Apply GridSearchCV on Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN) models.
- 6. Construct confusion matrices for all models to evaluate their performance.
- 7. Create a bar plot to compare the scores of the different models.

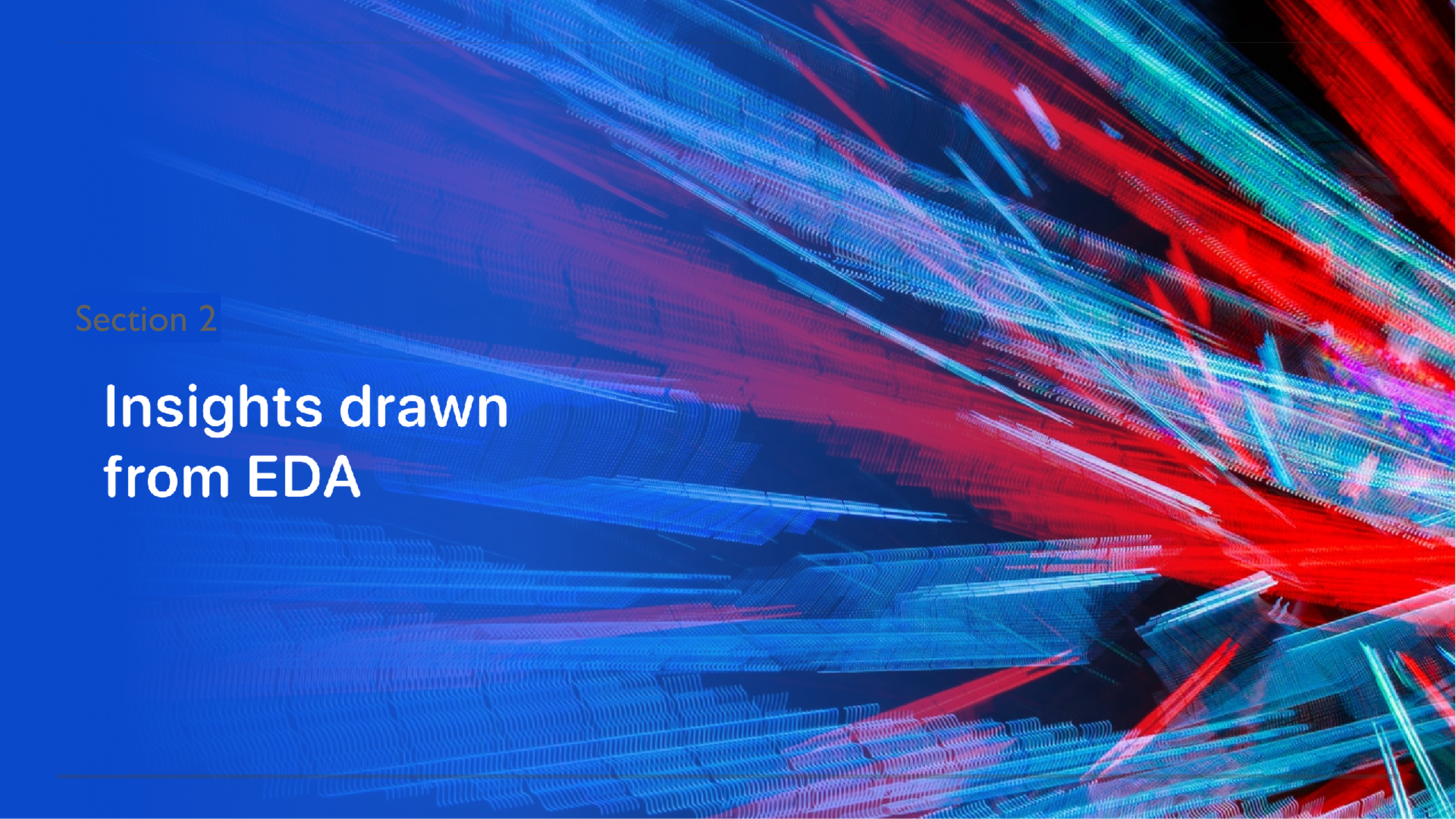
Results



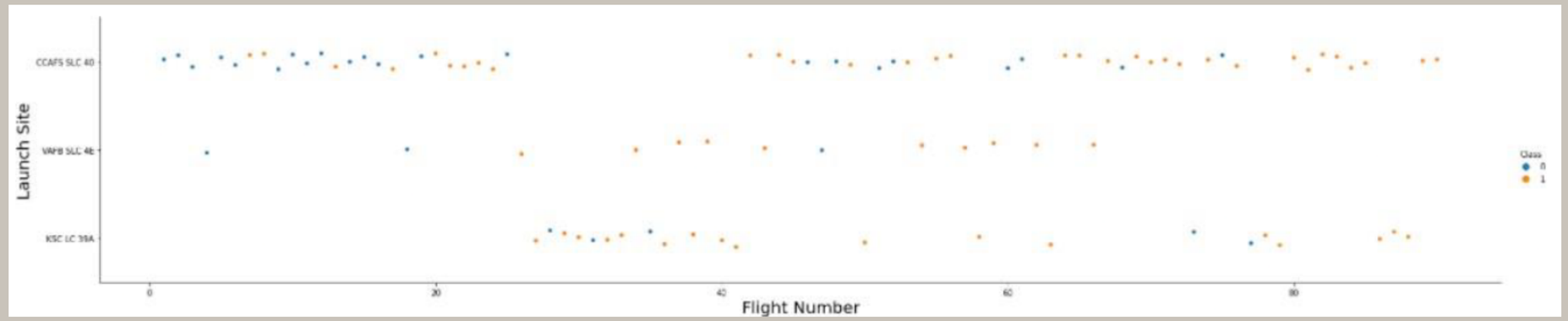
Here is a preview of the Plotly dashboard. The upcoming slides will showcase the results of the Exploratory Data Analysis (EDA) with visualizations, EDA with SQL integration, an interactive map created with Folium, and the outcomes of our model with an accuracy rate of approximately 83%.

Section 2

Insights drawn from EDA



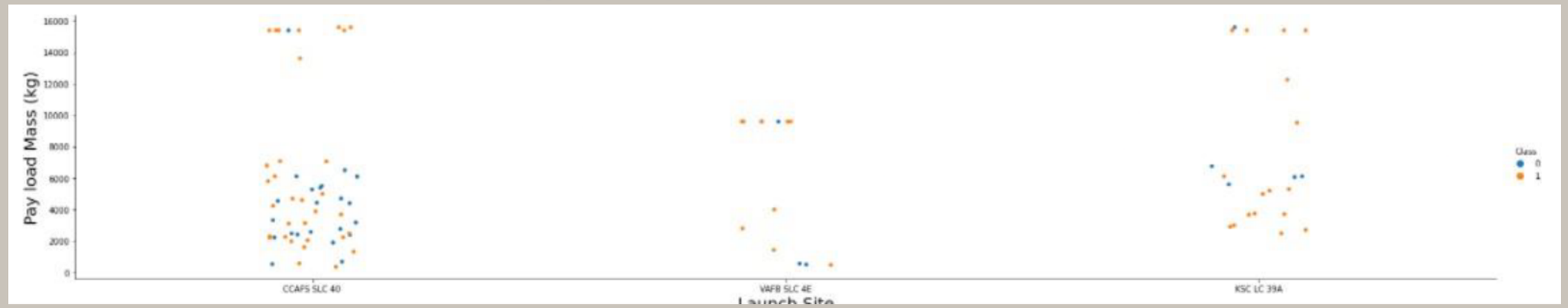
Flight Number vs. Launch Site



- The scatter plot indicates a notable trend of increasing success rate over time, as indicated by the Flight Number. There seems to be a significant breakthrough around flight 20, which resulted in a substantial increase in the success rate.

- Furthermore, based on the scatter plot, it appears that Cape Canaveral Space Launch Complex (CCAFS) is the primary launch site, considering its higher volume compared to other launch sites.

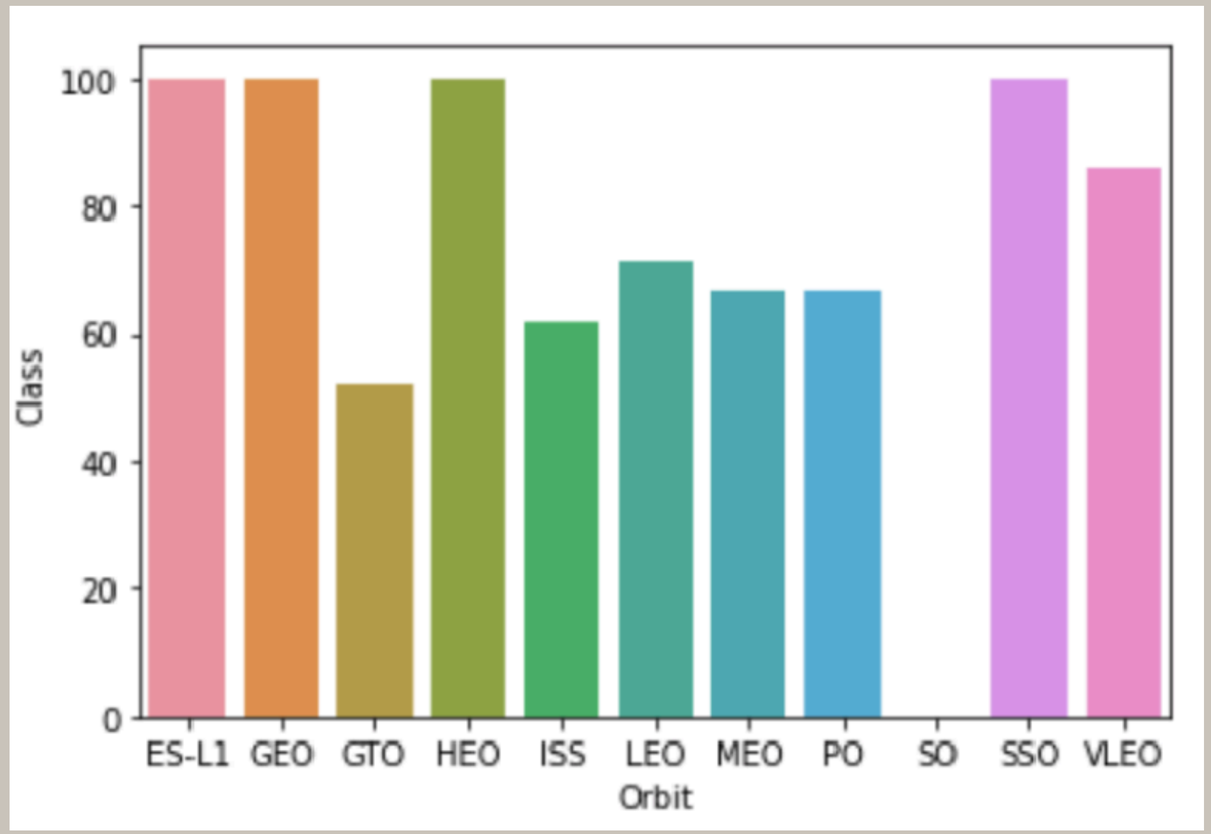
Payload vs. Launch Site



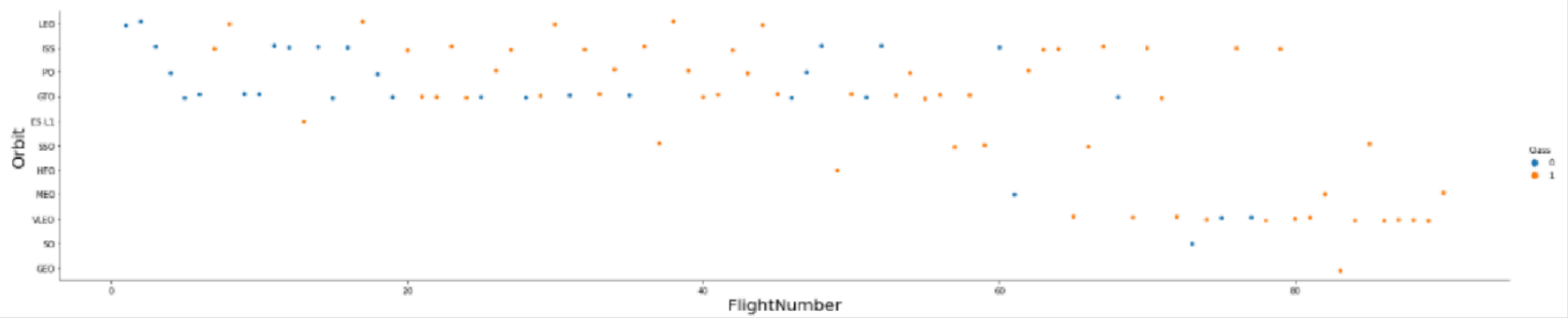
- Payload Mass Falls Between 0-7000Kg
- Different Launch sites Als seem to Have Differing payloads

Success Rate vs. Orbit Type

ES-L1 (1), GEO (1), HEO (1) have 100% success rate (sample sizes in parenthesis)
SSO (5) has 100% success rate
VLEO (14) has decent success rate and Attempts
SO (1) has 0% success rate
GTO (27) has the around 50% success rate but largest sample



Flight Number vs. Orbit Type

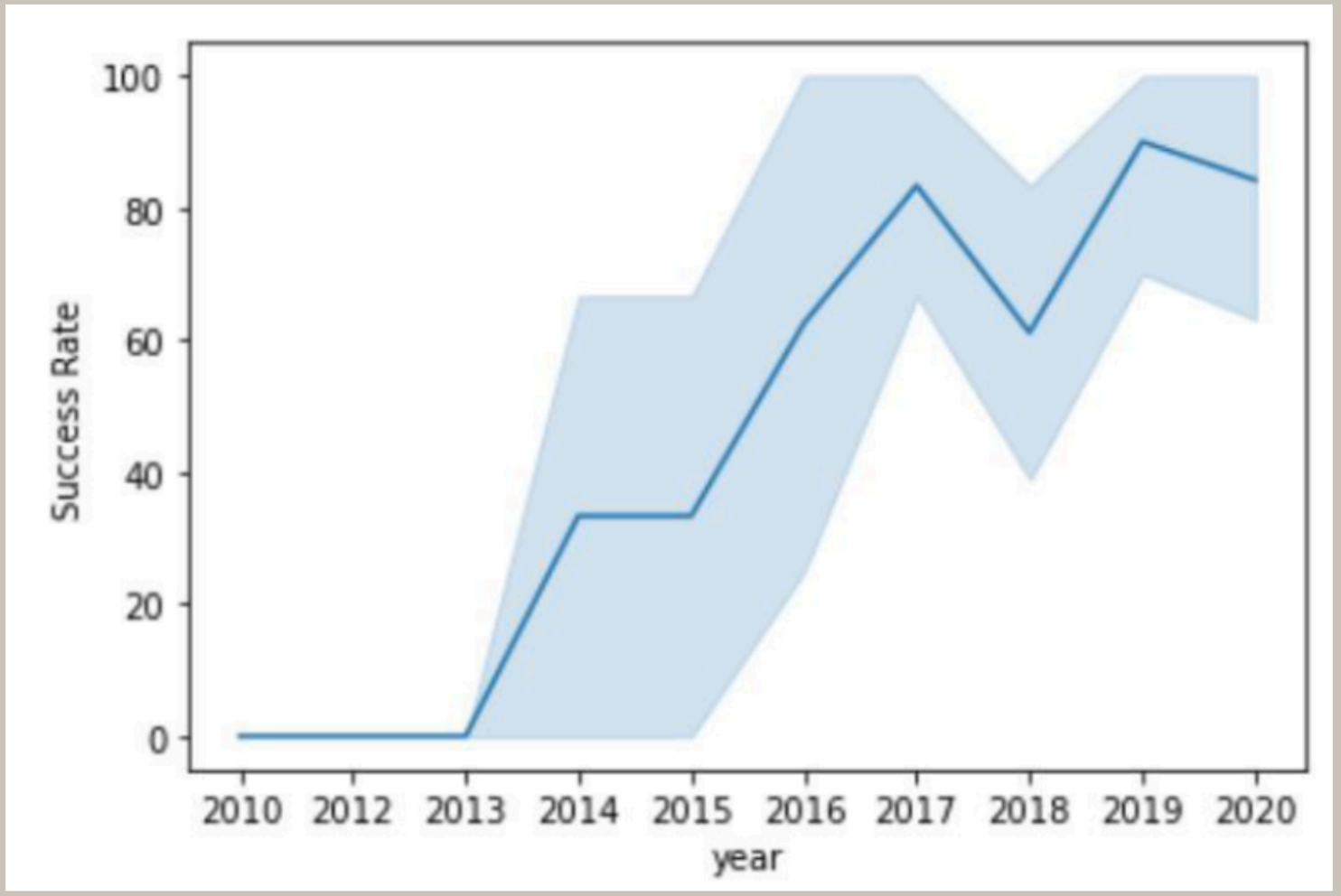


Launch Orbit preferences changed over Flight Number.
Launch Outcome seems to correlate with this preference.
Space started with LEO orbits which saw moderate success LEO and returned to VLEO in recent launches
Space appears to perform better in lower orbits or Sun-synchronous orbits

Payload vs. Orbit Type

Payload mass seems to correlate with orbit
LEO and SSO seem to have relatively low payload mass
The other most successful orbit VLEO only has payload mass values in the higher end of the range

Launch Success Yearly Trend



Success generally increases over time since 2013 with a slight dip in 2018

Success in recent years at around 80%

All Launch Site Names

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

```
 * ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-8
d:32733/BLUDB
Done.
```

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

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5
```

```
 * ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.clou
d:32733/BLUDB
Done.
```

DATE	Time (UTC)	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	Landing _Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-12	22:41:00	F9 v1.1	CCAFS LC-40	SES-8	3170	GTO	SES	Success	No attempt

Total Payload Mass

```
%sql select sum(payload_mass__kg_) as sum from SPACEXTBL
where customer like 'NASA (CRS)'
```

```
 * ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-8954-7e38
e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:
32733/BLUDB
Done.
```

SUM
22007

Average Payload Mass by F9 v1.1

```
%sql select avg(payload_mass__kg_) as Average from SPACE
XTBL where booster_version like 'F9 v1.1%'
```

```
 * ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-8954-7e38
e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:
32733/BLUDB
Done.
```

average
3226

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
SELECT booster_version
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4001 AND 5999;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.database
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

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

- This Shows That Space X has A High probability of a mission success

Boosters Carried Maximum Payload

```
maxm = %sql select max(payload_mass__kg_) from SPACEXTBL
maxv = maxm[0][0]

%sql select booster_version from SPACEXTBL where payload
_mass__kg_=(select max(payload_mass__kg_) from SPACEXTB
L)
```

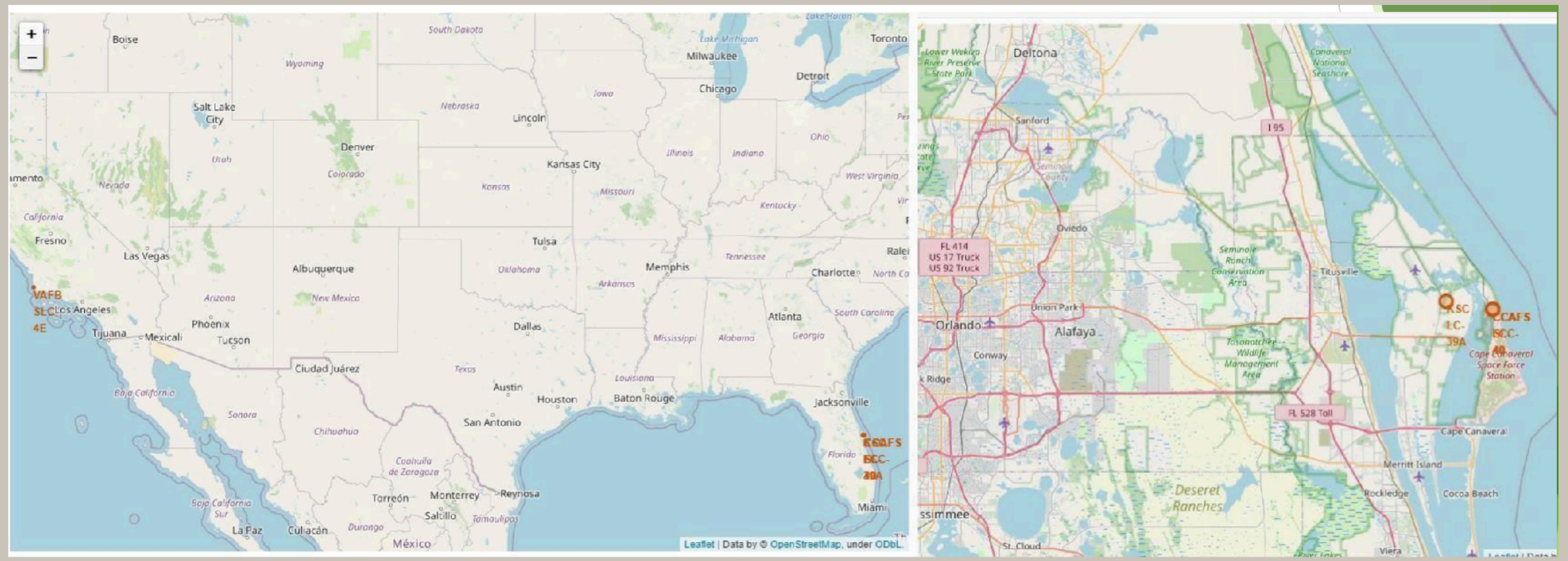
```
 * ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-8954-7e38
e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:
32733/BLUDB
Done.
 * ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-8954-7e38
e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:
32733/BLUDB
Done.
```

booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3

- This query reveals that the highest payload mass carried was 1560C kg.
- Furthermore, it appears that the booster versions used, namely F9 B5 B10xx, are very similar. This suggests that there is a correlation between the payload mass and the specific booster version employed.

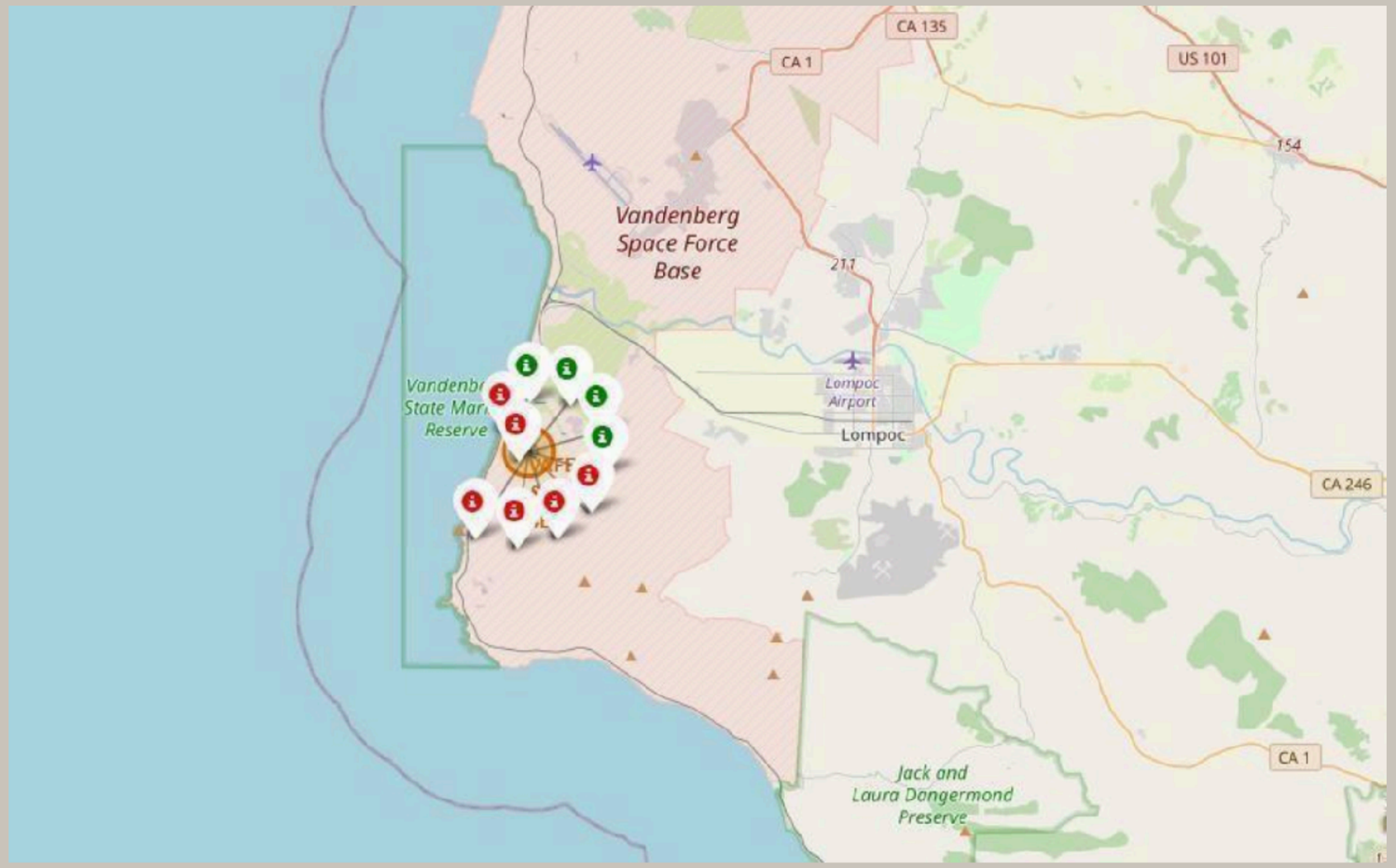

Section 3
Launch Sites
Proximities Analysis

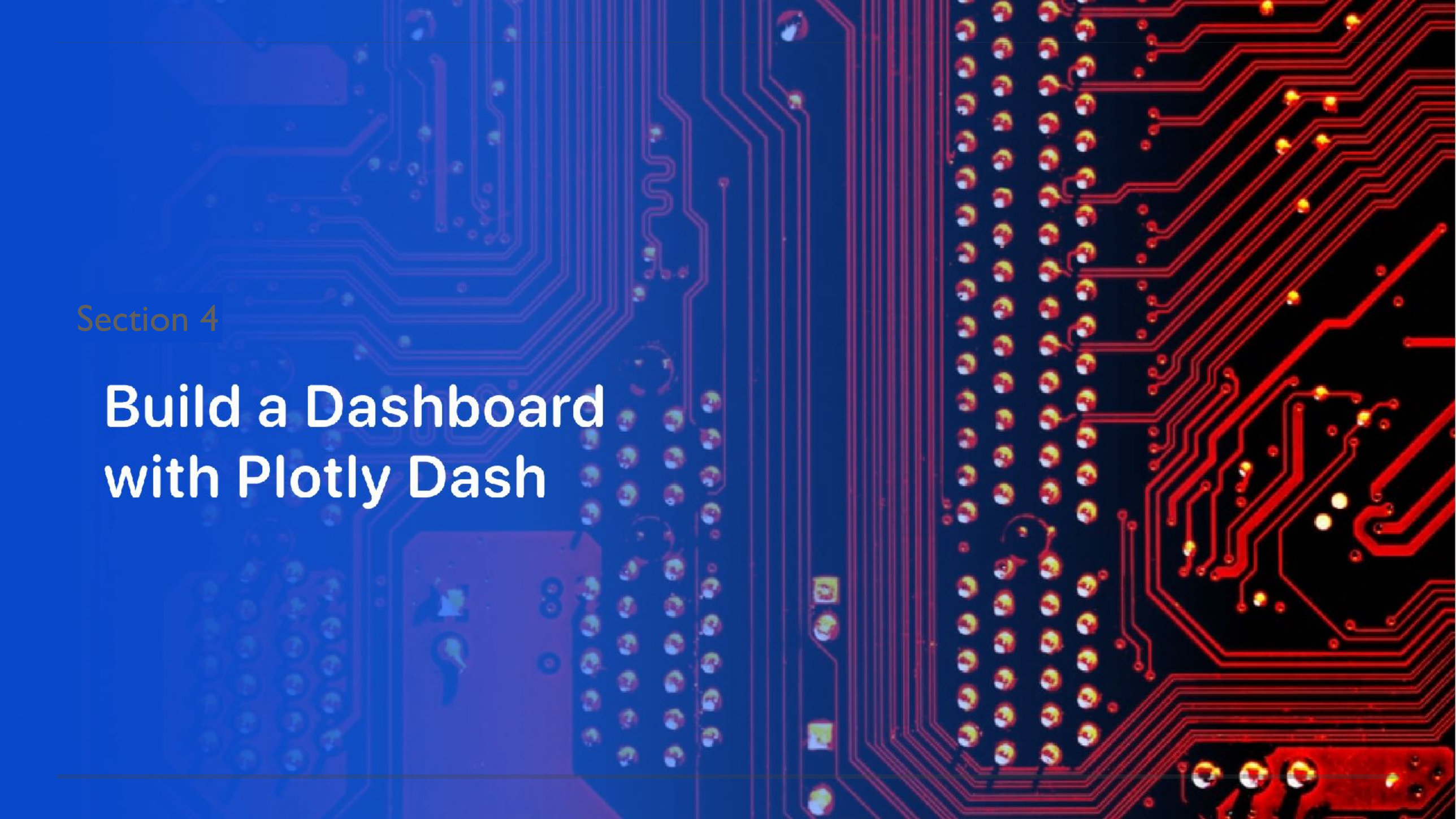
<Folium Map Screenshot 1>



The Left Shows all Launch Locations While the Right Just Shows Florida
All Sites are near the ocean

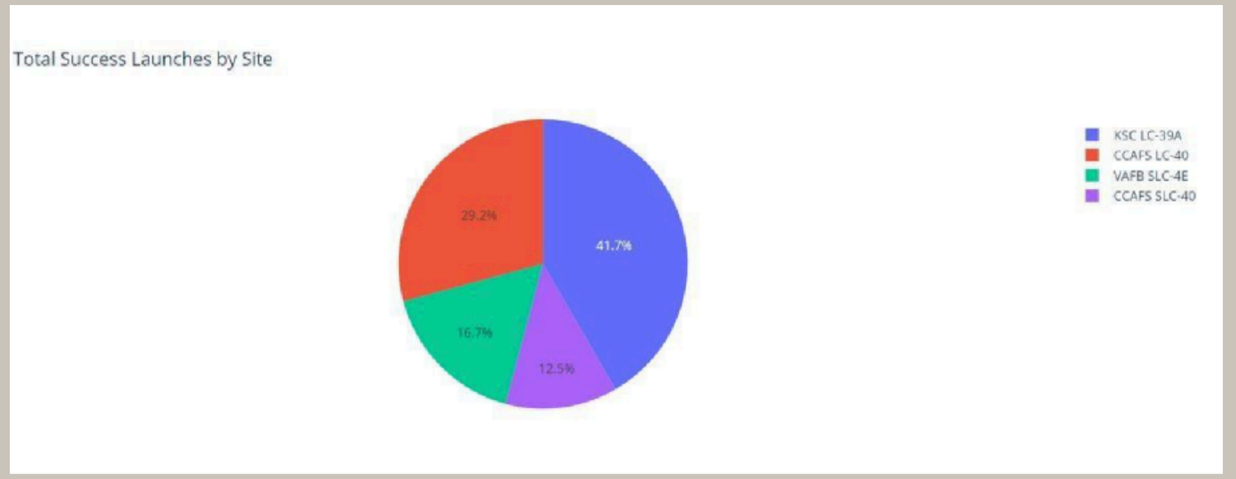
<Folium Map Screenshot 3>



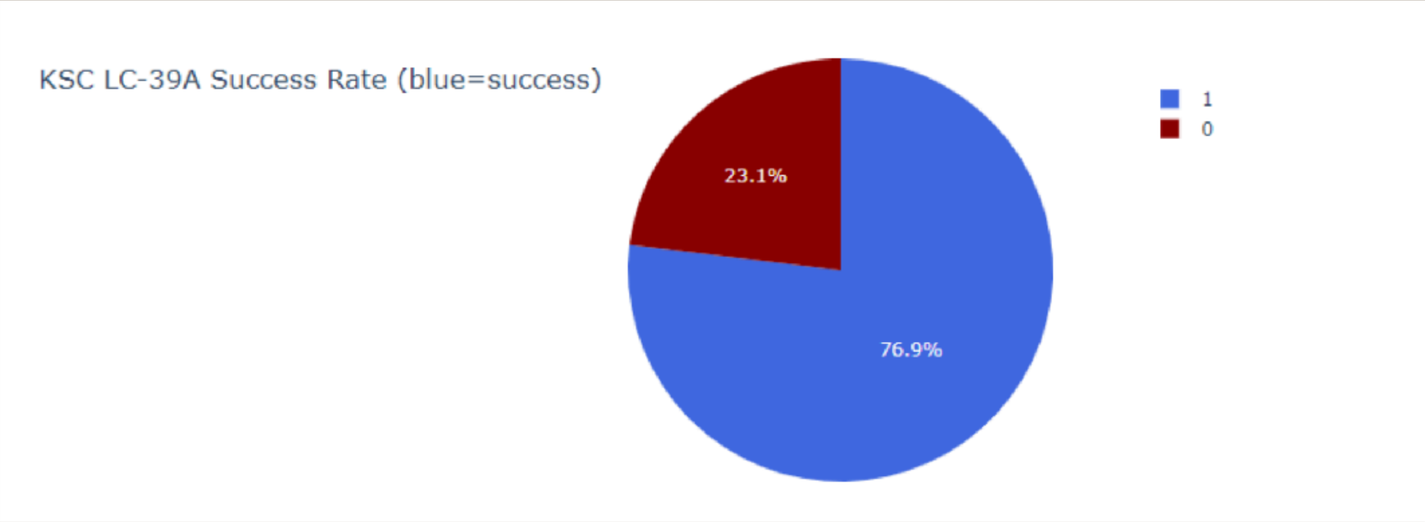

Section 4
Build a Dashboard with Plotly Dash

Successful Launch Sites

- Distribution of successful landings across launch sites:
- - CCAFS LC-40 (old name) and KSC have the same number of successful landings.
- - Majority of successful landings occurred before the name change.
- - VAFB has the smallest share of successful landings, possibly due to smaller sample and west coast challenges.



Highest Success Rate



- KSC-LC-39A Hash the Highest Success Rate of Just 3 Failures

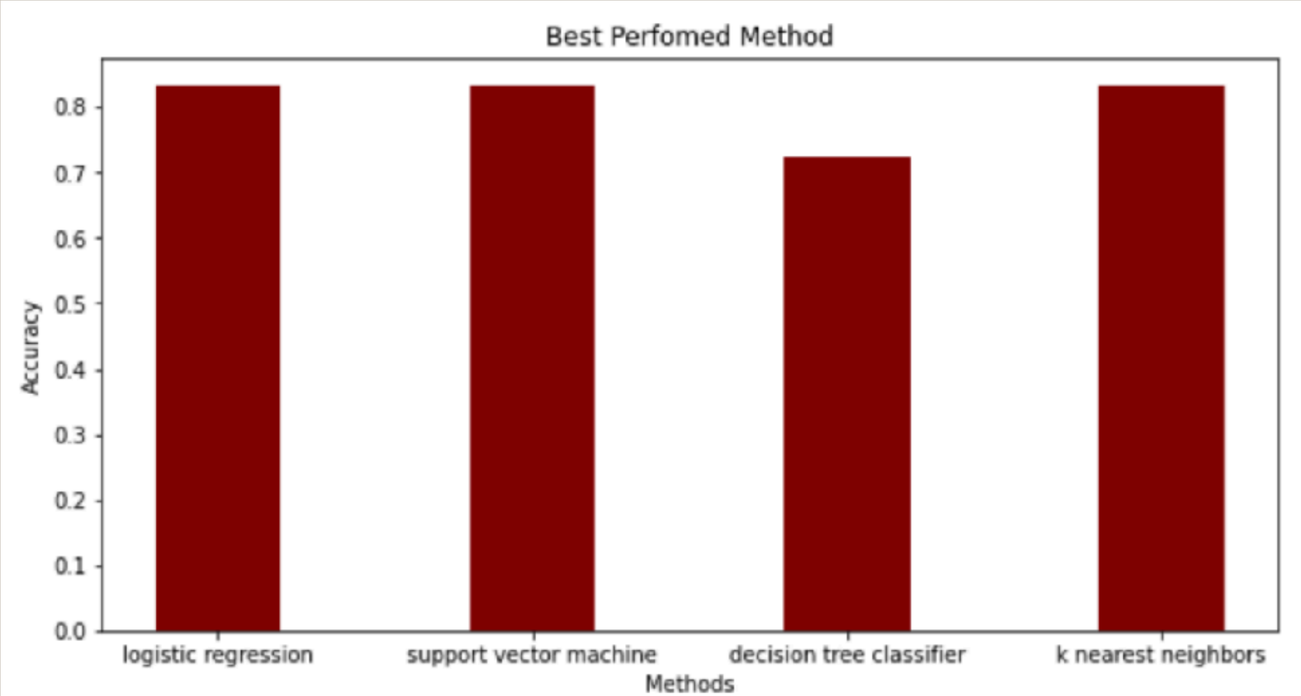
Payload Mass Vs Success

- The Plotly dashboard features a Payload range selector, which is currently set from 0 to 10000. However, the maximum Payload observed in the dataset is 15600.
- The scatter plot within the dashboard considers the Class column, where 1 represents a successful landing and 0 indicates a failure. Additionally, the scatter plot incorporates the booster version category, represented by color, and the number of launches, depicted by point size.
- Interestingly, within the selected range of 0-7500, there are two instances of failed landings with payloads recorded as zero kilograms. This highlights a noteworthy observation within the dataset.

Section 5

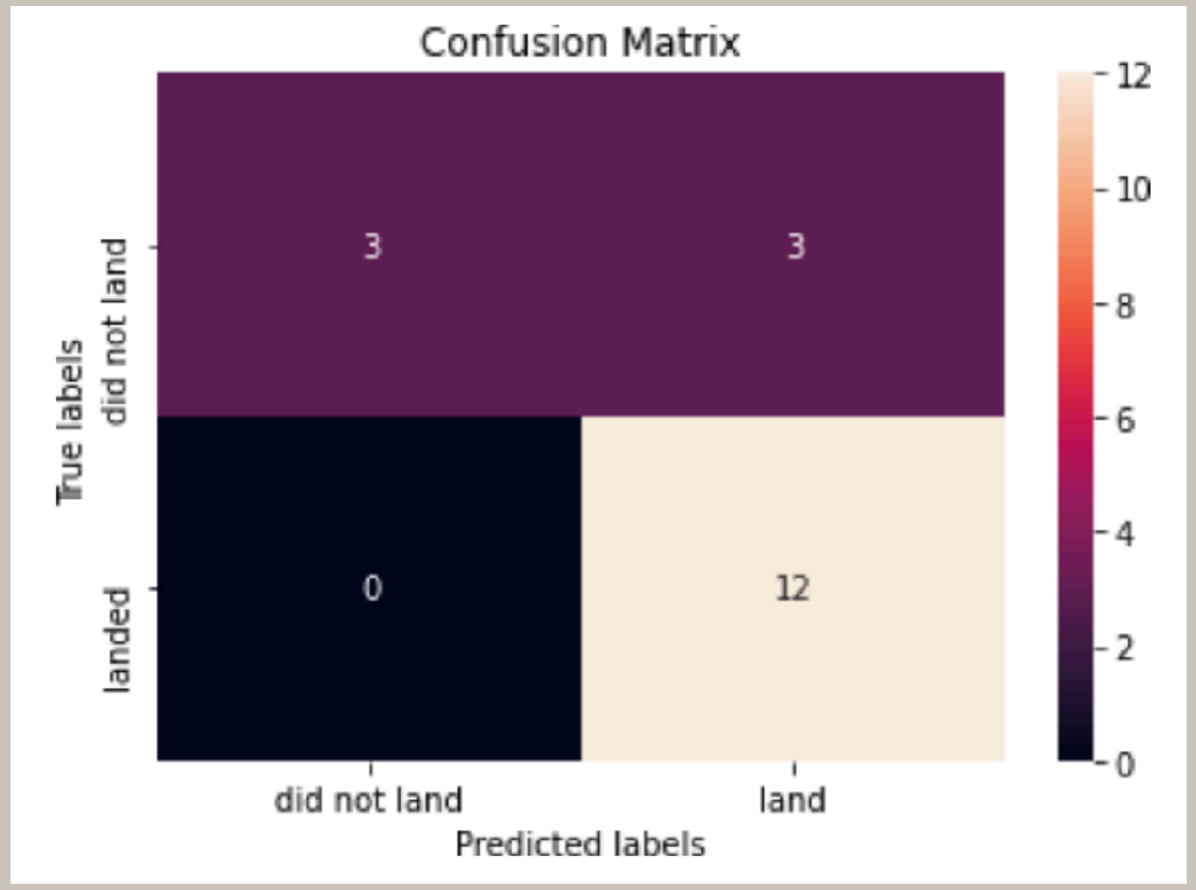
Predictive Analysis (Classification)

Classification Accuracy



- The models exhibited similar levels of accuracy on the test set, with an overall accuracy of 83.33% except for the Decision Tree Classifier model, which achieved 72.23% accuracy.
- It is important to note that the test size is relatively small, with only 18 samples. This limited sample size can result in larger variance in accuracy results, as seen in the Decision Tree Classifier model across repeated runs.

Confusion Matrix



- Since all models performed similarly on the test set, the confusion matrix is consistent across all models. The following are the results of the confusion matrix:
- - The models correctly predicted 12 successful landings when the true label was a successful landing.
- - The models correctly predicted 3 unsuccessful landings when the true label was an unsuccessful landing.
- - However, the models also predicted 3 successful landings when the true label was an unsuccessful landing, resulting in false positives.

Conclusions

- Our task was to develop a machine learning model for Space Y, enabling them to compete with Space X by predicting successful Stage 1 landings and potentially saving around $100 million USD. Here's a summary of our process and findings:

- 1. Data Collection: We gathered data from a public Space API and performed web scraping on the Space Wikipedia page.
- 2. Data Labeling and Storage: We created data labels and stored the collected data in a DB2 SQL database.
- 3. Visualization Dashboard: A dashboard was created to visualize the data, allowing for better insights and analysis.
- 4. Machine Learning Model: We developed a machine learning model with an accuracy of 83% to predict the success of Stage 1 landings.
- 5. Practical Application: The model can be utilized by Allon Musk and Space Y to predict, with relatively high accuracy, whether a launch will have a successful Stage 1 landing prior to the launch. This information can assist in determining whether the launch should proceed or not.
- 6. Need for More Data: Collecting additional data is recommended to further refine and improve the accuracy of the machine learning model and identify the best model for the task at hand.

- Overall, our work provides a valuable tool for Space Y in making informed decisions regarding their launches, but further data collection and analysis will enhance the effectiveness of the model.

Thank you!