SpaceX Capstone

Jiawei Sui

7/17/2023

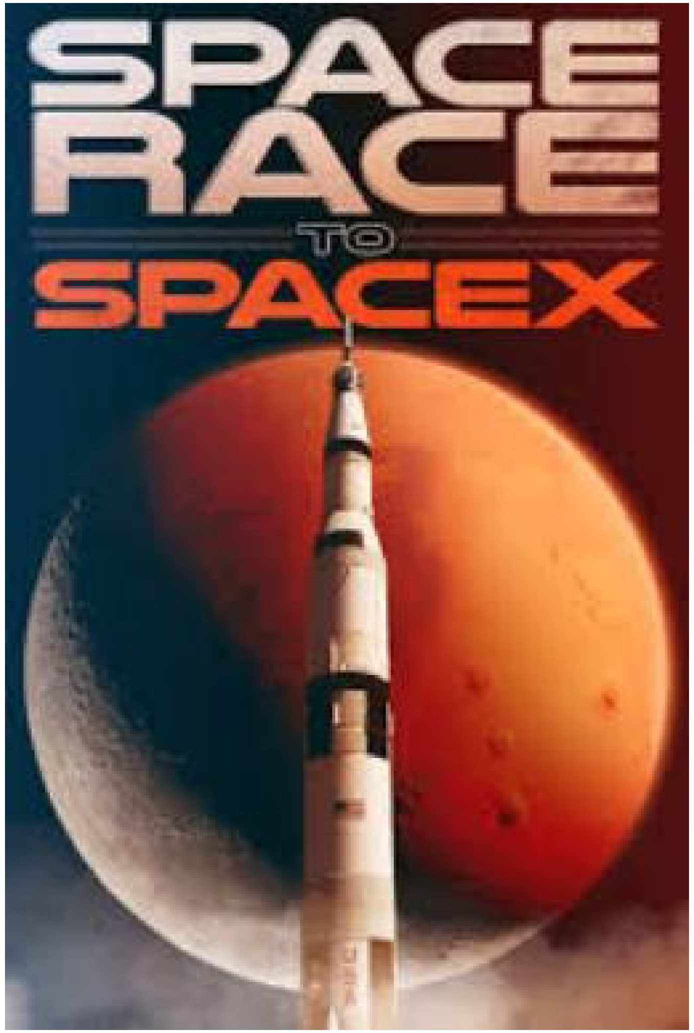
OUTLINE

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

EXECUTIVE SUMMARY

- Methodology Summary
  - Data Collection
  - Data Wrangling
  - Exploratory Data Analysis with SQL and Visualization
  - Interactive Visual Analysis with Folium
  - Machine Learning Prediction
- Result Summary
  - Exploratory Data Analysis Results
  - Interactive Analysis Screenshots
  - Predictive Analysis Results

INTRODUCTION



By ingeniously reusing the first stage of its Falcon 9 rocket, SpaceX has been able to significantly reduce launch costs compared to traditional providers. While other companies face expenditures of over $165 million per launch, SpaceX's innovative approach brings the price down to a relatively affordable $62 million. This game-changing cost advantage positions SpaceX as a disruptive force in the industry, democratizing access to space for both governmental and commercial entities. With its bold vision and unwavering determination, SpaceX continues to push the boundaries of space exploration and redefine what is achievable.

METHODOLOGY

- Collect Data using Space X REST API and web scraping technologies.
- Data Wrangling by filtering the data, handling missing values, preparing the data for exploratory data analysis and machine learning.
- Exploratory Data Analysis using visualization and SQL
- Interactive visualization analysis using Folium and Plotly Dash
- Predictive analysis using various classification and regression models

Data Collection

- Request Data from SpaceX API
- Using .json() and .json_normaliza() to preprocess the data and turn it to a pandas dataframe
- Examine the missing values and replace them with the mean value or just drop them
- Using the BeautifulSoup to scrape data for Falcon 9 launch records from Wikipedia

Data Wrangling

- We calculated the number of launches on each site
- We calculated the number and occurrence of each orbit
- We calculated the number and occurrence of mission outcome per orbit type
- We created a landing outcome label from Outcome column

Use the method `.value_counts()` to determine the number and occurrence of each orbit in th

In [8]:
```
# Apply value_counts on Orbit column
df.Orbit.value_counts()
```

Out[8]:
```
GTO      27
ISS      21
VLEO     14
PO        9
LEO       7
SSO       5
MEO       3
ES-L1     1
HEO       1
SO        1
GEO       1
Name: Orbit, dtype: int64
```

TASK 3: Calculate the number and occurence of mission outcome

Use the method `.value_counts()` on the column `Outcome` to determine the number of `lan` variable landing_outcomes.

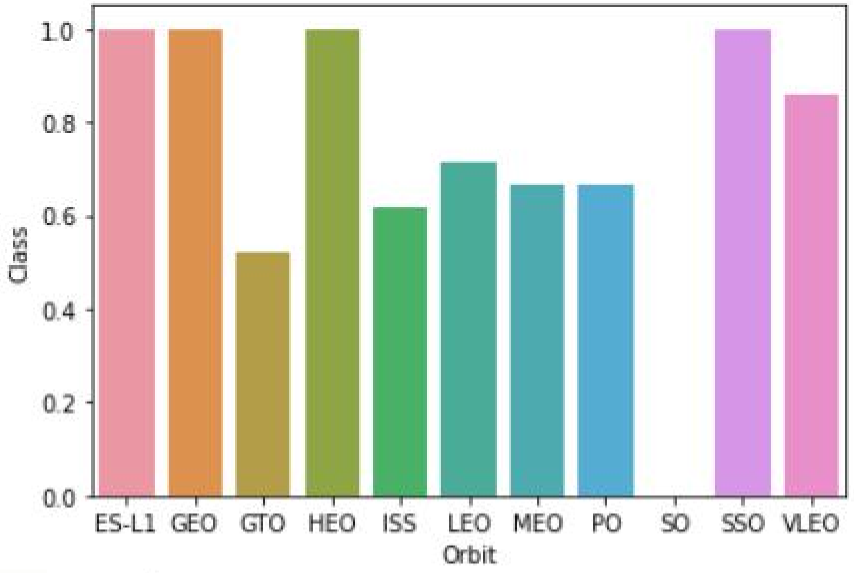
In [9]:
```
# landing_outcomes = values on Outcome column
landing_outcomes = df.Outcome.value_counts()
landing_outcomes
```

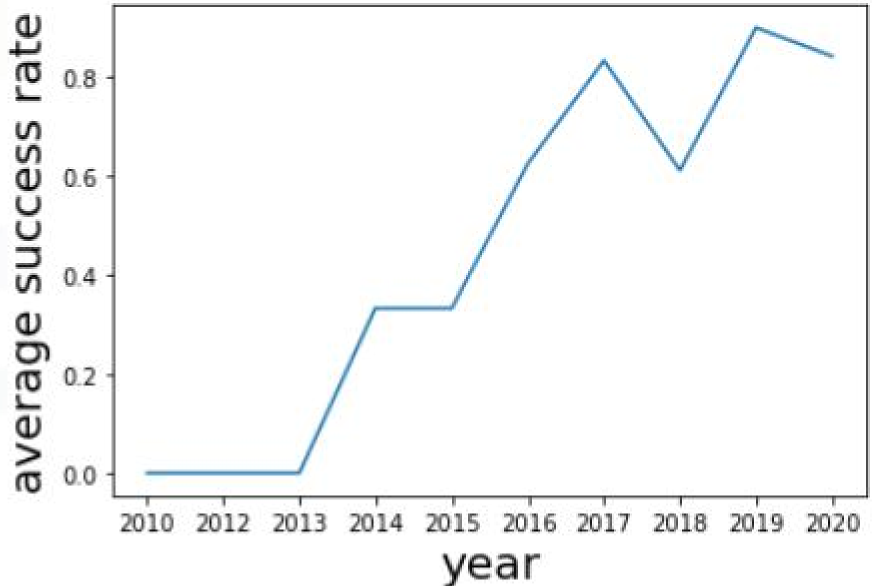
Out[9]:
```
True ASDS      41
None None      19
True RTLS      14
False ASDS      6
True Ocean      5
False Ocean     2
None ASDS       2
False RTLS      1
Name: Outcome, dtype: int64
```

Result Summary

- **Exploratory Data Analysis results**
- **Interactive analysis screenshots**
- **Predictive analysis results**

EDA with Data Visualization



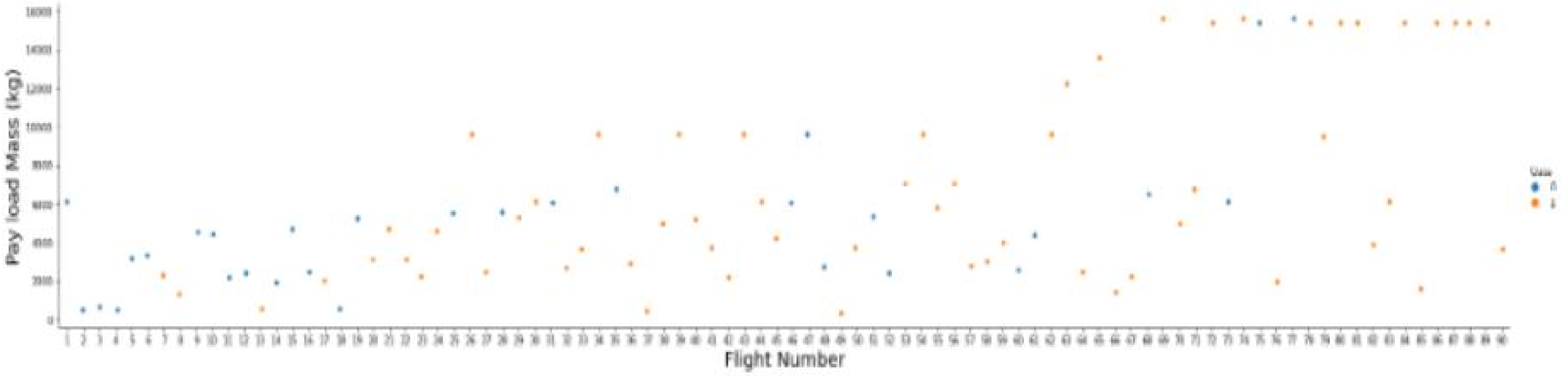


Flight Number and Launch Site

- We use the function catplot to plot flight number and launch site, set the parameter x to flight number, set the parameter y to launch site, and set the parameter hue to class

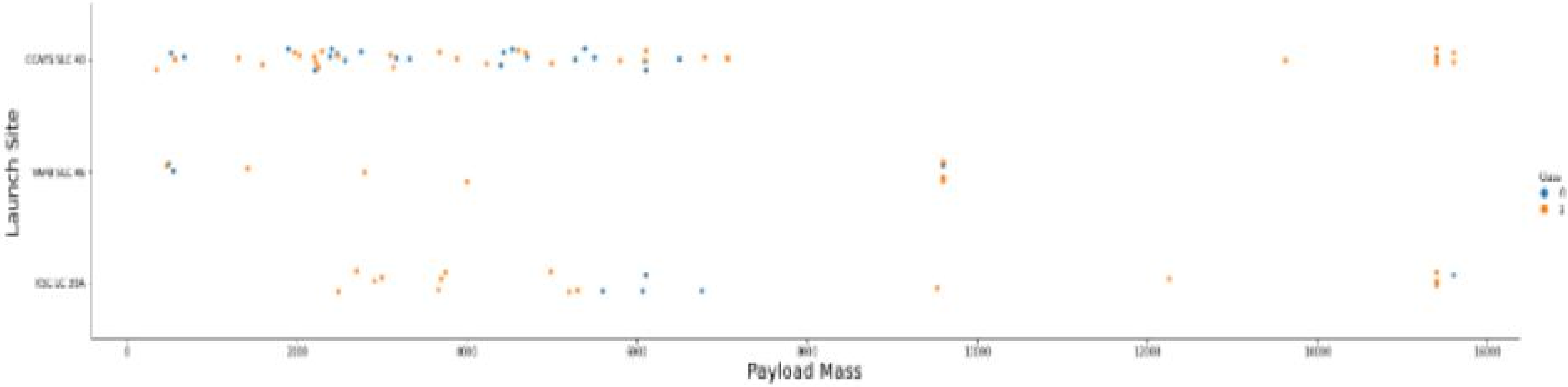
Flight Number and Payload Mass

• We can plot out the flight number and payload mass and overlay the outcome of the launch. We see that as the flight number increases, the first stage is more likely to land successfully. The payload mass is also important. It seems the more massive the payload, the less likely the first stage will return.



Payload and Orbit Type

- We use the function catplot to plot flight number and launch site, set the parameter x to flight number, set the parameter y to launch site, and set the parameter hue to class



EDA with SQL

SQL Screenshots

Total Number of Successful Outcomes

We use where and like functions to generate a column of successful launch outcomes and use % function to minimize the search scope

```
%%sql
select Booster_Version from spacex where Landing_Outcome like 'Success%' and  PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_
```

```
 * mysql+mysqlconnector://root:***@localhost:3306/mydb
14 rows affected.
```

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 B4 B1043.1
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

SQL Screenshots

We list the number of failed landing_outcomes in drone ship, their booster versions and launch sites names for in year 2015

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%%sql
select Landing_Outcome, Booster_Version, Launch_Site, my_date from spacex where Landing_Outcome = 'Failure (drone ship)' anc
```

```
 * mysql+mysqlconnector://root:***@localhost:3306/mydb
2 rows affected.
```

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

SQL Screenshots

Display 5 records where launch sites begin with the string “CCA”

Display 5 records where launch sites begin with the string 'CCA'

```
%%sql
select *
from spacex
where Launch_Site like "CCA%" limit 5
```

```
 * mysql+mysqlconnector://root:***@localhost:3306/mydb
5 rows affected.
```

My_date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX
08-12-2010	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
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)

SQL Screenshots

- **We determined the maximum payload using a subquery in the WHERE clause and the MAX() function.**

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [17]:

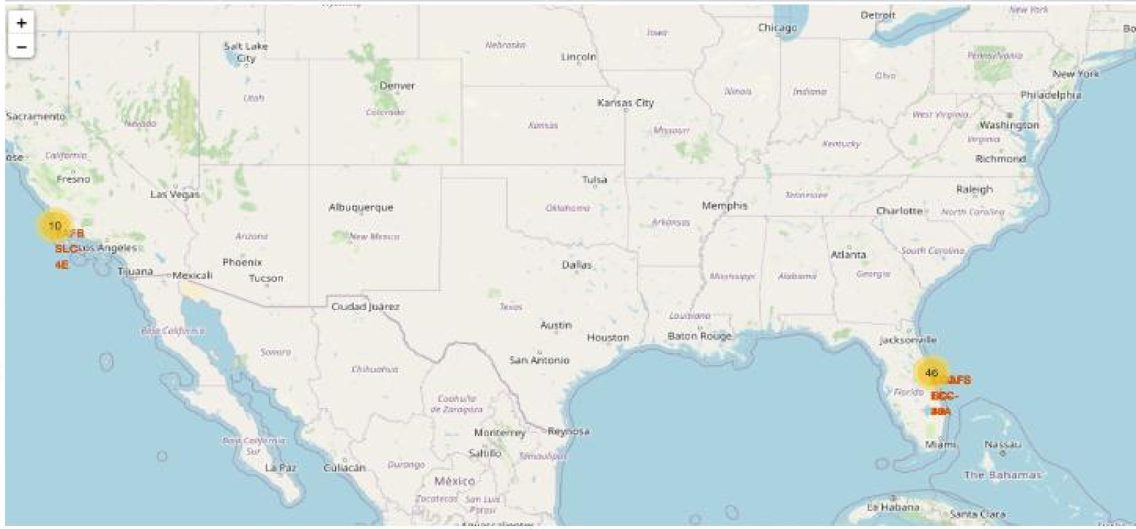
```
task_8 = '''
        SELECT BoosterVersion, PayloadMassKG
        FROM SpaceX
        WHERE PayloadMassKG = (
                                SELECT MAX(PayloadMassKG)
                                FROM SpaceX
                                )
        ORDER BY BoosterVersion
        '''
create_pandas_df(task_8, database=conn)
```

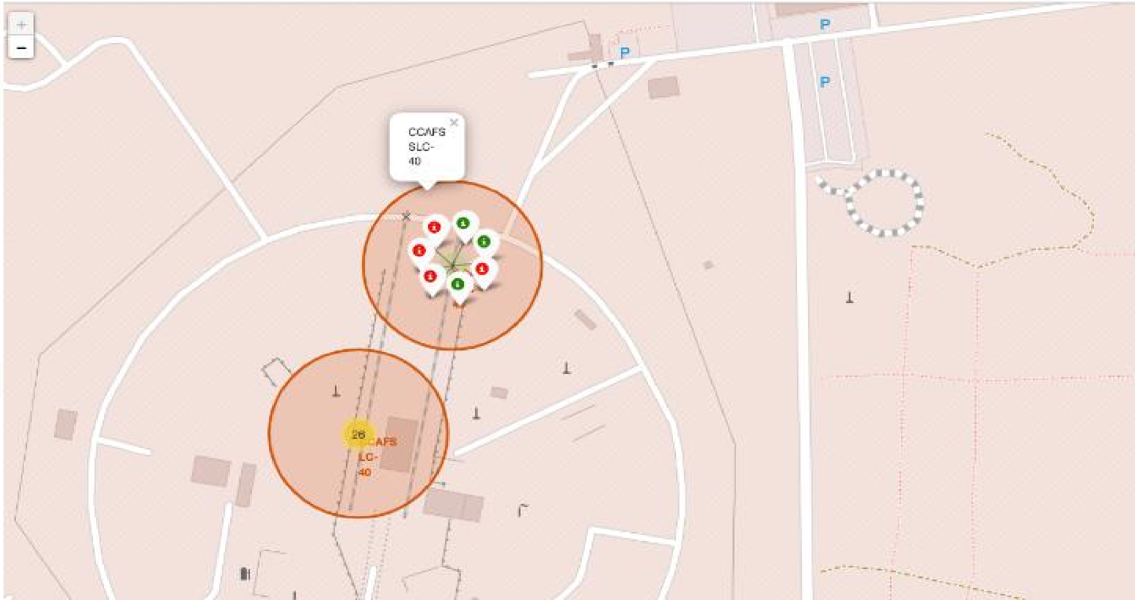
Out[17]:

	boosterversion	payloadmasskg
0	F9 B5 B1048.4	15600
1	F9 B5 B1048.5	15600
2	F9 B5 B1049.4	15600
3	F9 B5 B1049.5	15600
4	F9 B5 B1049.7	15600
5	F9 B5 B1051.3	15600
6	F9 B5 B1051.4	15600
7	F9 B5 B1051.6	15600
8	F9 B5 B1056.4	15600
9	F9 B5 B1058.3	15600
10	F9 B5 B1060.2	15600
11	F9 B5 B1060.3	15600

Interactive map with Folium

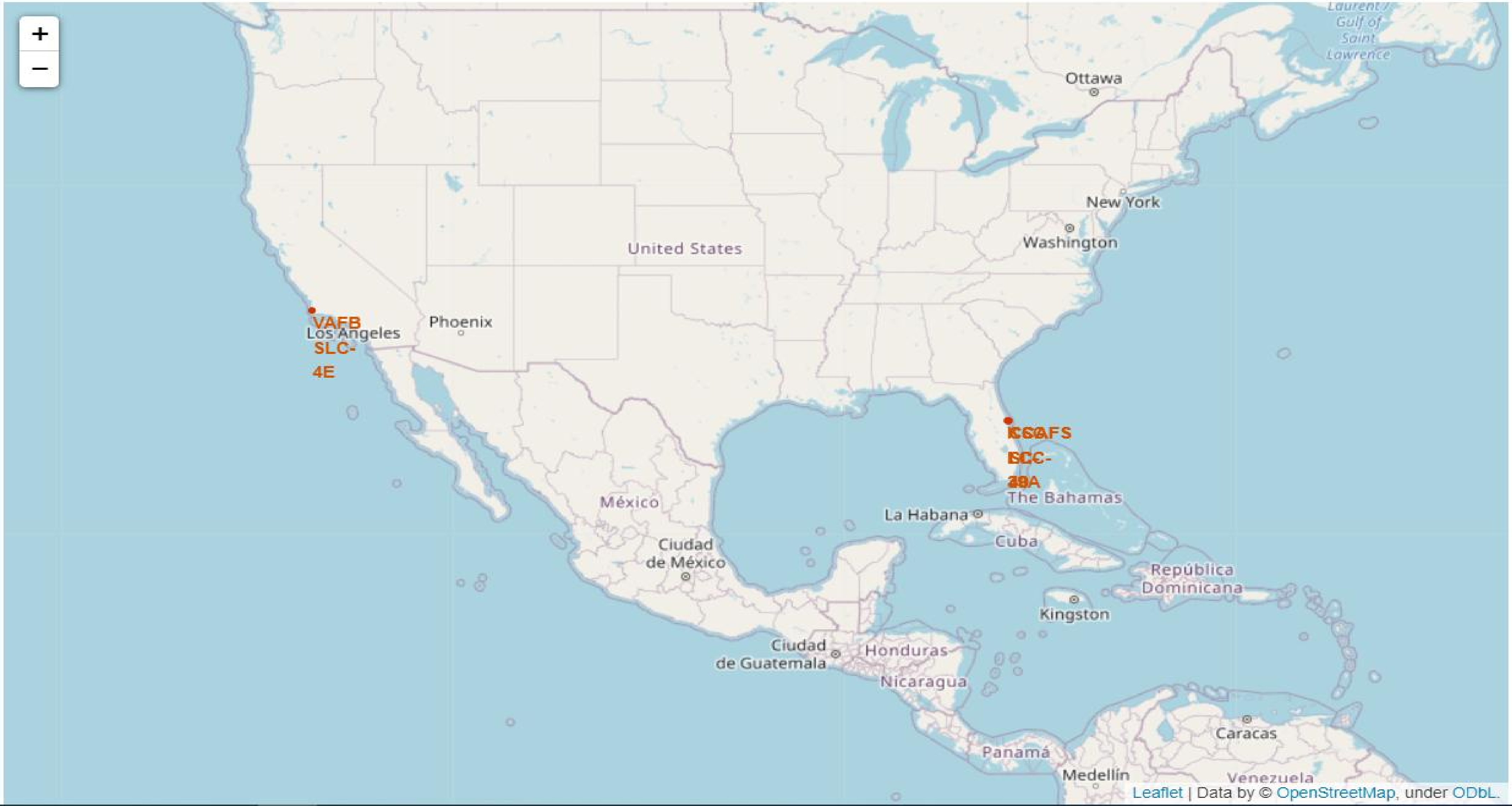
- We use red markers to indicate launch sites in the United States and show its name using its latitude and longitude coordinates.
- We use green to indicate a successful launch and use red to indicate a unsuccessful launch
- The distances from a launch site to its proximities, railway, high way ,and coasts are explained in map.



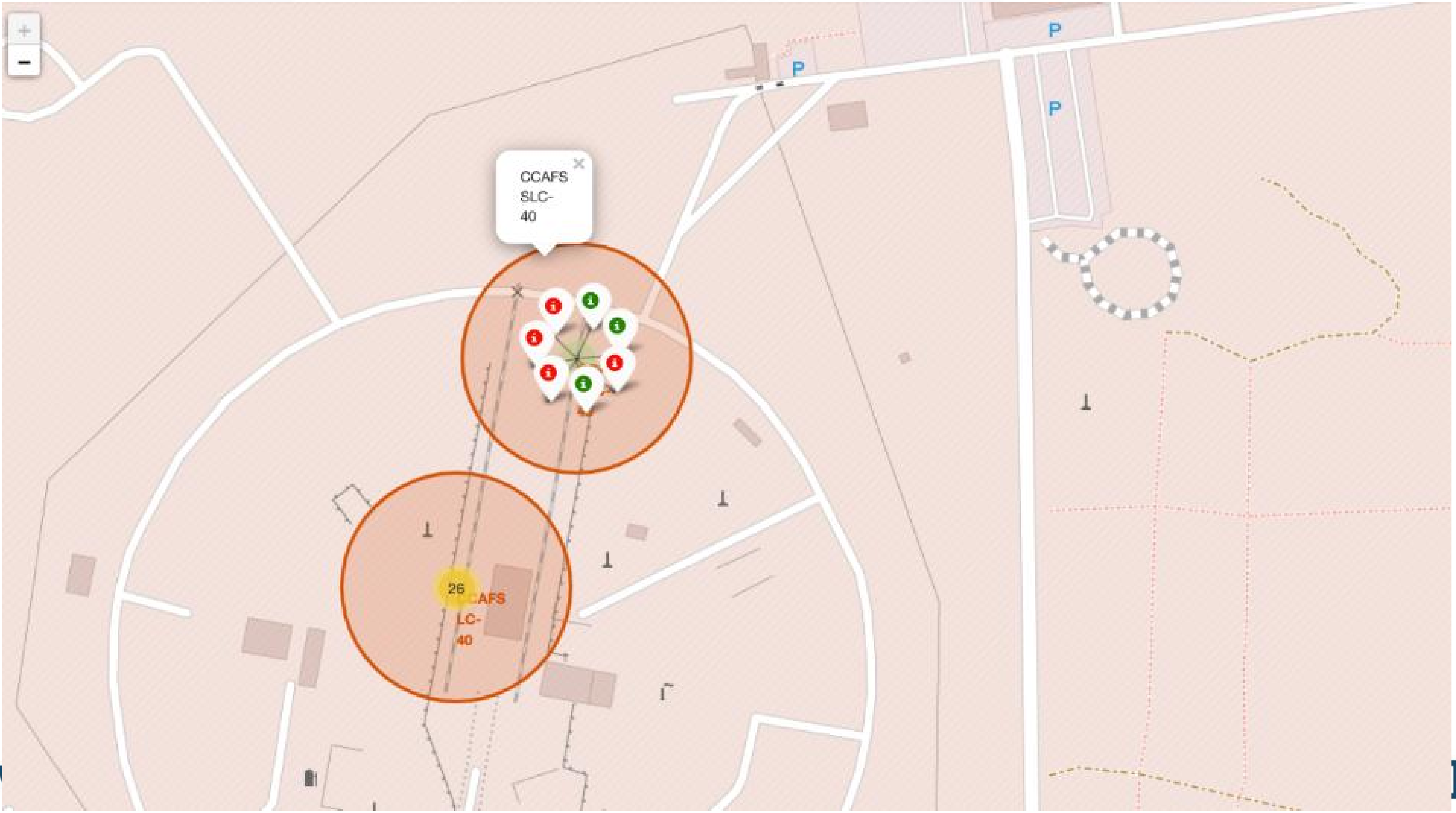


All launch sites on the map

Out[10]:

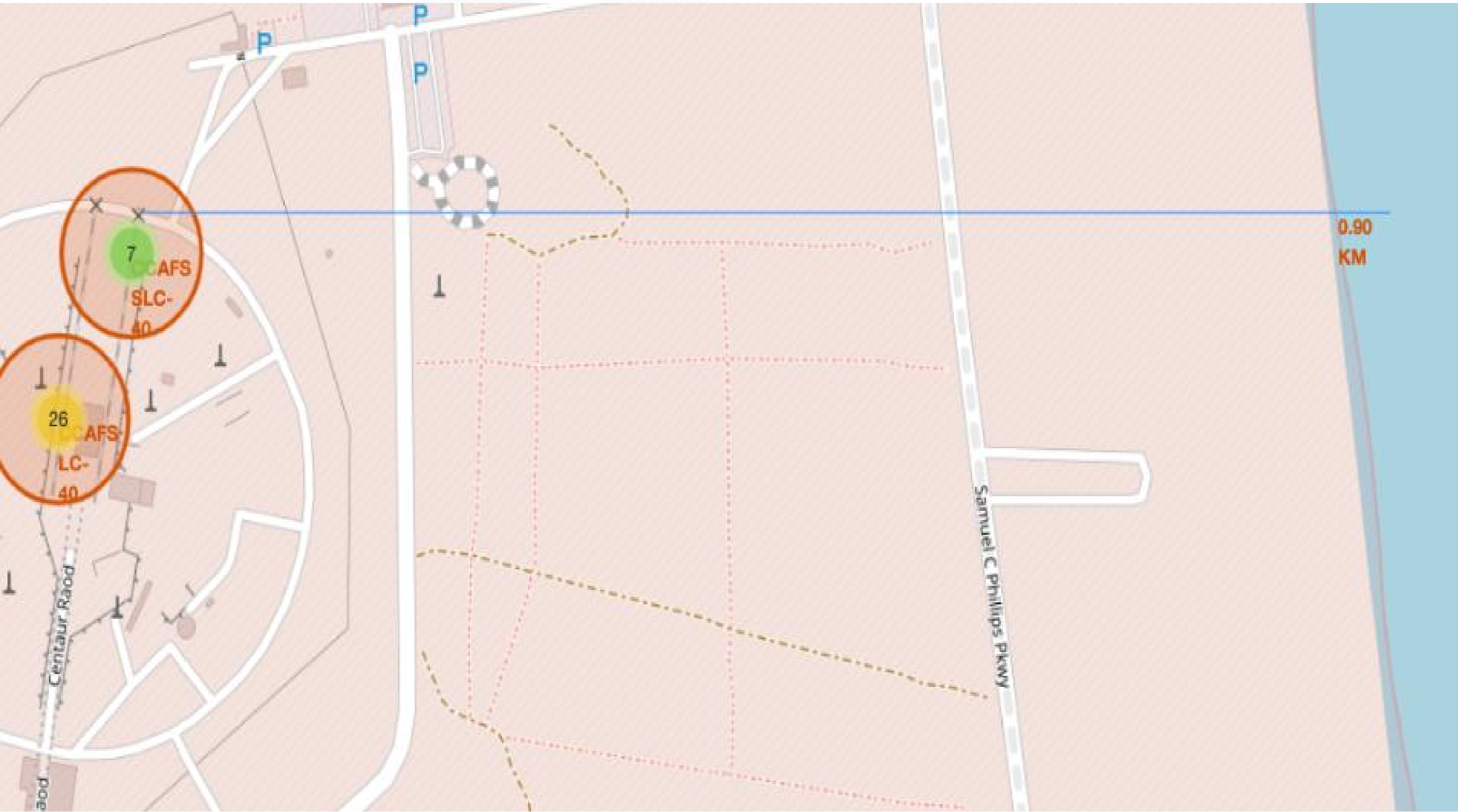


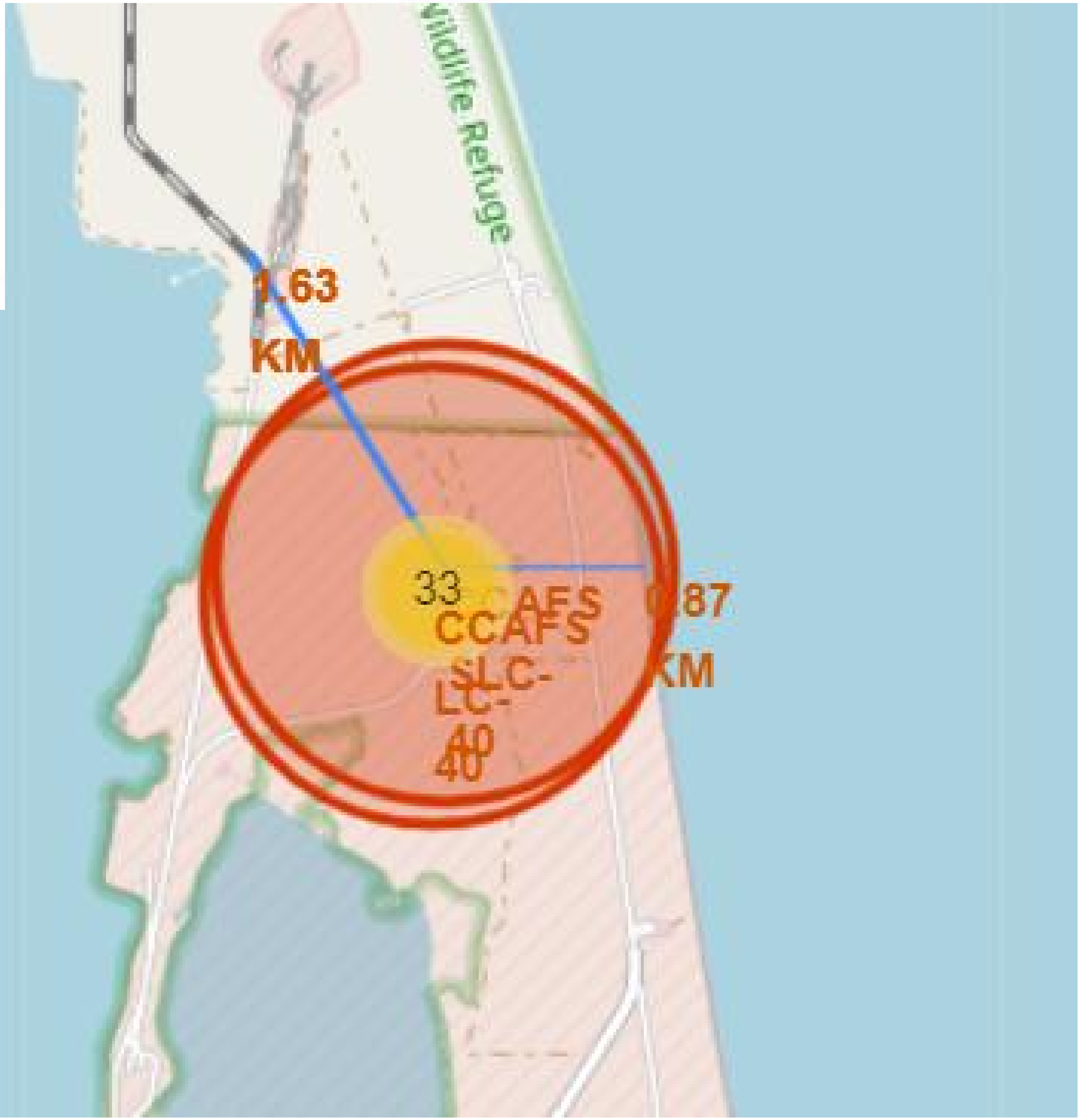
All successful and failed launches



Distance to Proximities

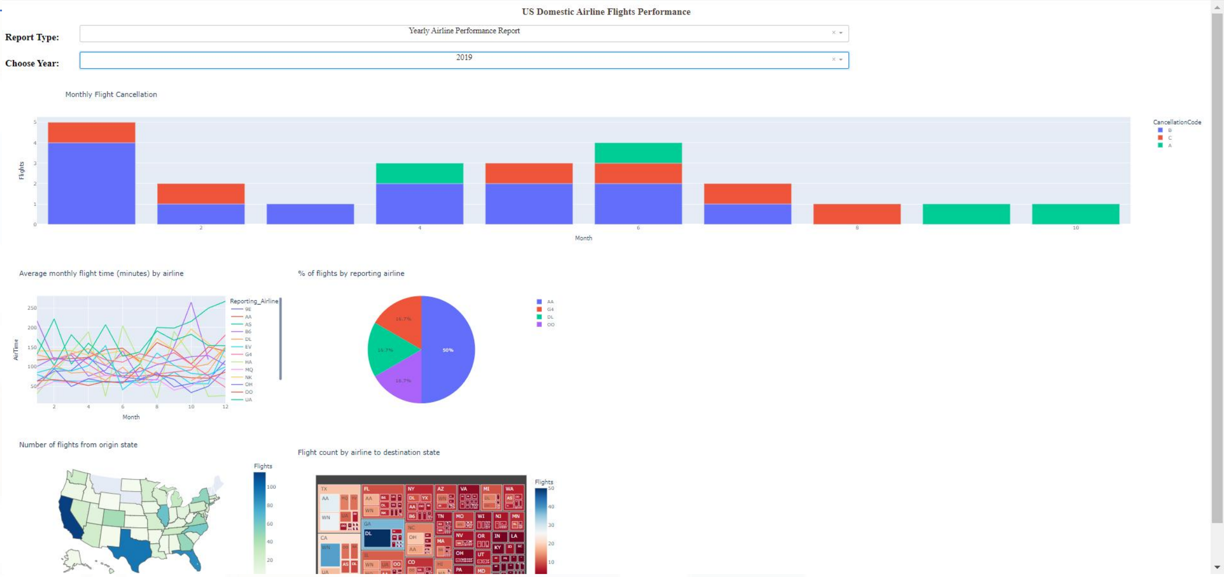
AFS-LC-4
AFS-SLC-40

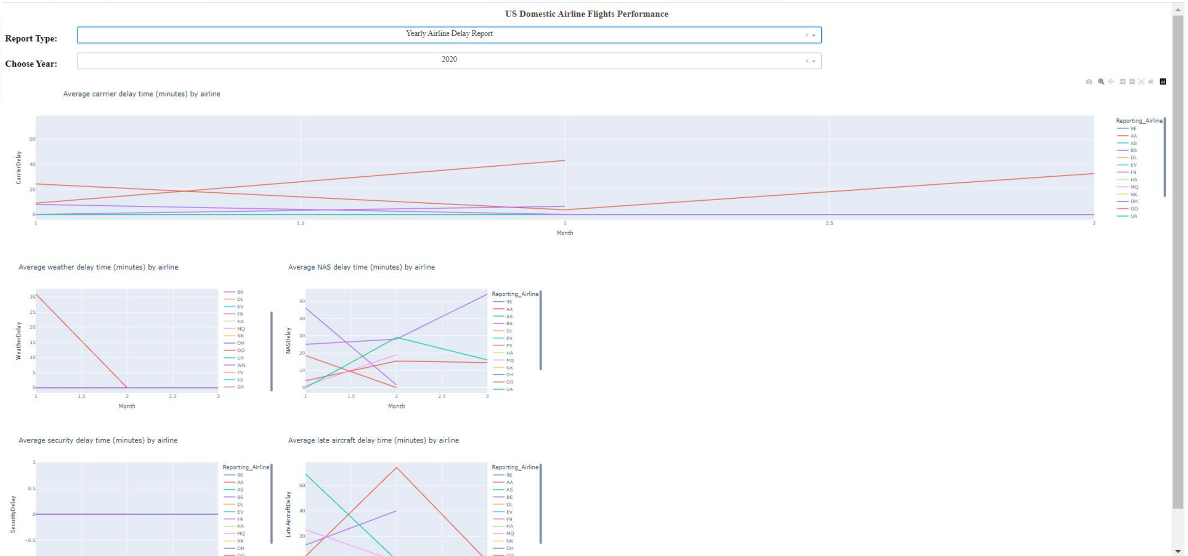




Dashboard with Plotly Dash

- Dropdown list with launch sites allows user to select all launch sites
- Pie charts show successful launches
- Plotly Dash application for users to perform interactive visual analytics on SpaceX launch data in real-time.

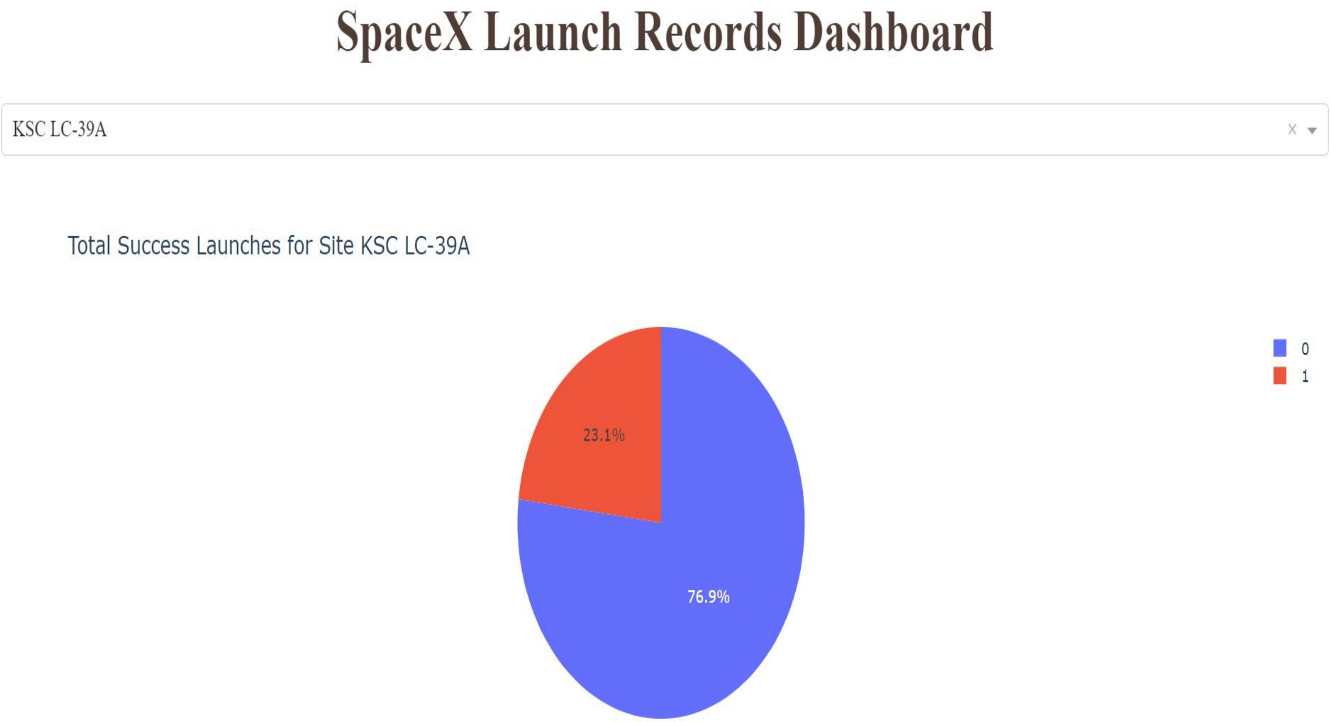


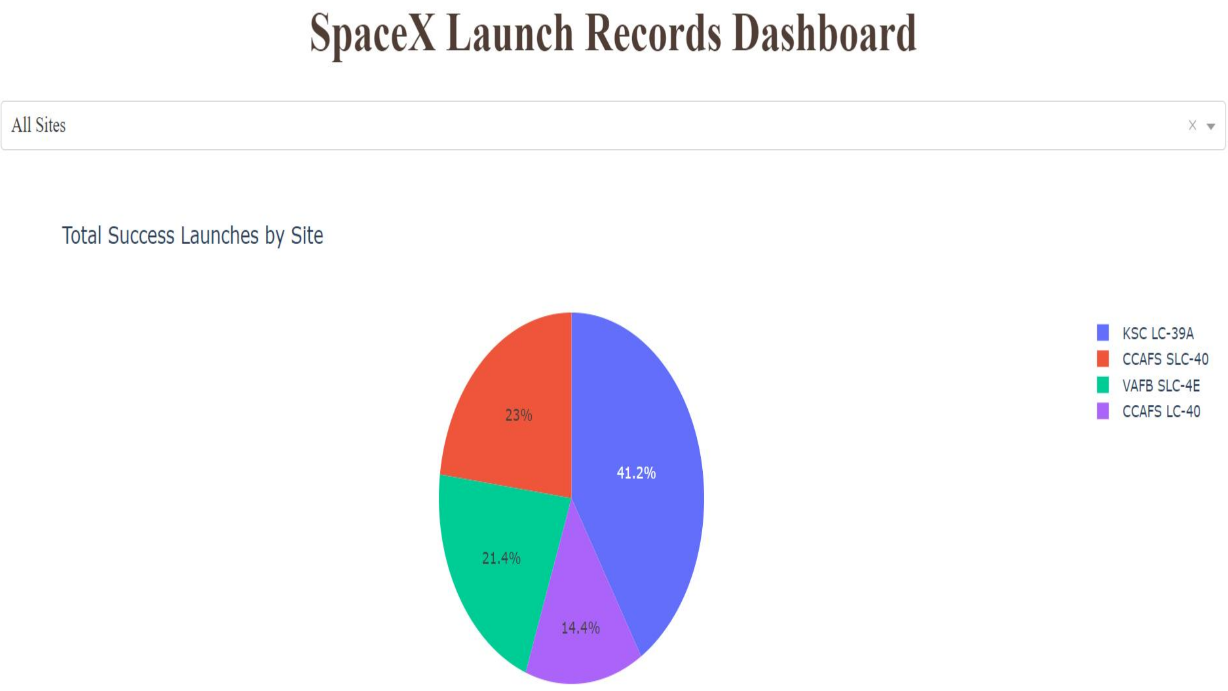


Launch Success

KSC LC-39A has the highest success rate among launch site

Total Success Launches by Sites



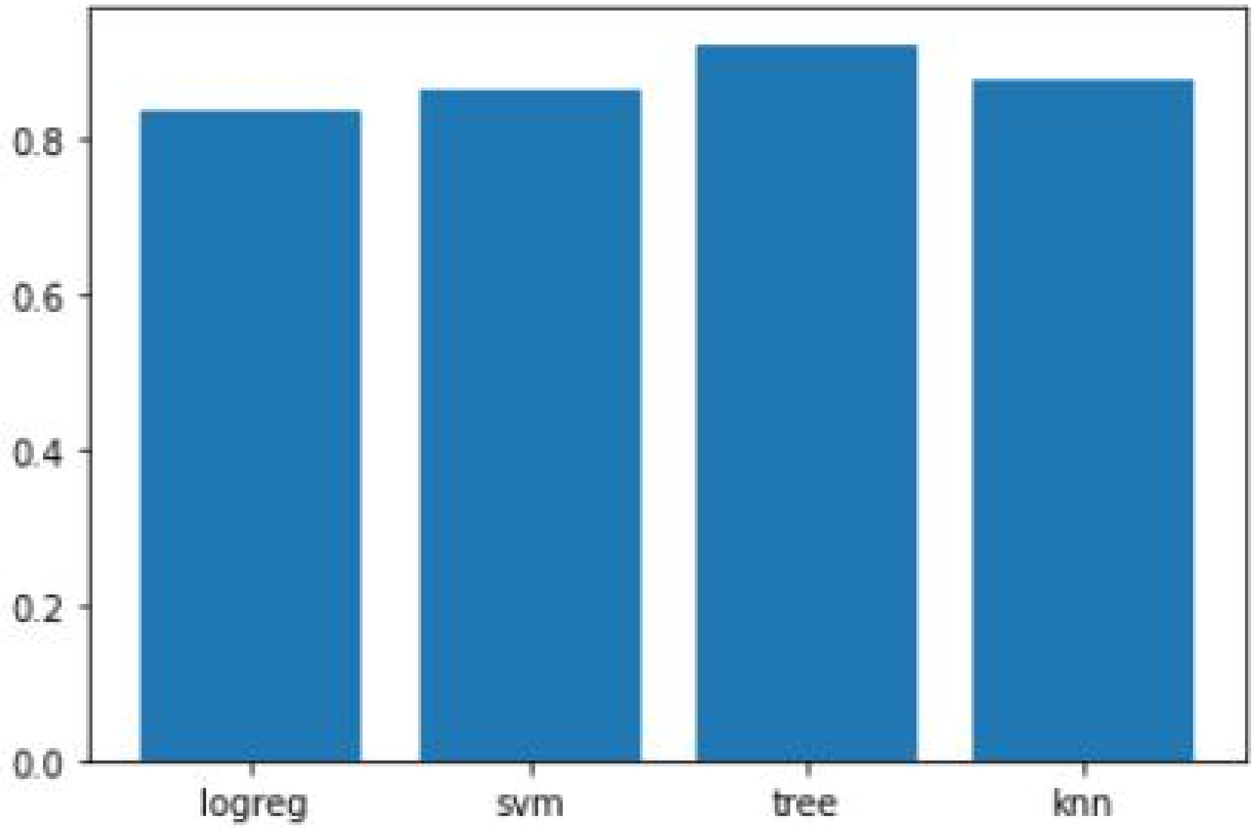


Predictive Analytics

- **The data was loaded, transformed, and split into training and testing sets for machine learning modeling.**
- **Multiple algorithms were tried, including logistic regression, support vector machines, decision trees, and K-nearest neighbors.**
- **Hyperparameter tuning was performed using GridSearchCV to improve the models' performance.**
- **The best performing classification model was identified based on accuracy as the evaluation metric.**
- **The predictive analysis involved importing packages, defining input and output data, and utilizing various algorithms to determine the best prediction results.**

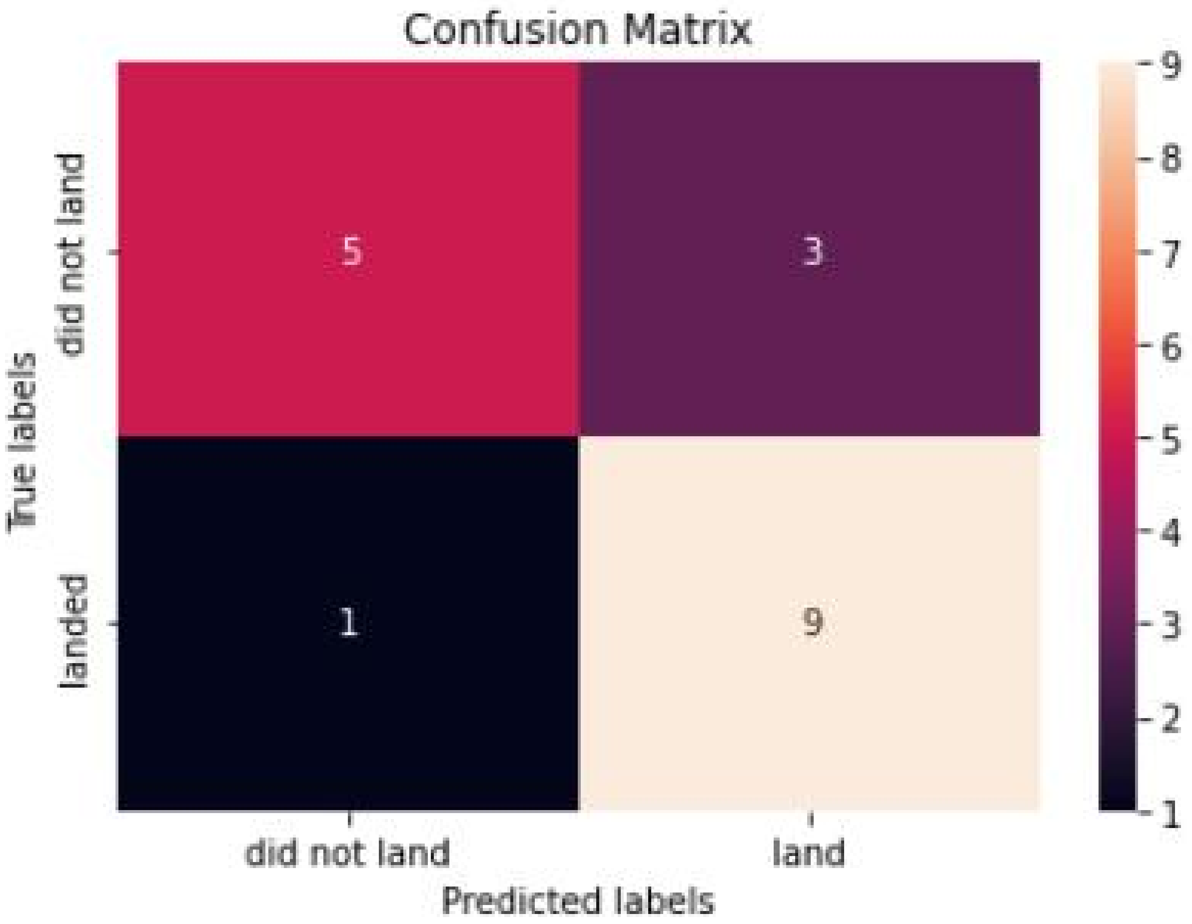
Classification Accuracy

We compared the accuracy score of each model, such as logistic regression, support vector regression, K Nearest Neighbors, and Decision Tree. It can be shown that the Decision Tree has the best accuracy with a score of 93%



Confusion Matrix

- **Examining the confusion matrix, we see that Decision Tree regression can distinguish between the different classes. We see that the major problem is false negatives, aka unsuccessful landing marked as successful landing by the classifier.**



Conclusions

- The success rate of launches at a site tends to increase as the number of flights conducted at that site increases.
- From 2013 to 2020, there was a notable upward trend in the launch success rate.
- Orbits such as ES-L1, GEO, HEO, and SSO exhibited the highest success rates.
- Among all launch sites, KSC LC-39A stood out with the highest number of successful launches.

Conclusions

- **The Decision Tree classifier emerged as the most effective machine learning algorithm for this task.**
- **The models performed comparably on the test set, with the decision tree model slightly outperforming others.**
- **Launch sites are strategically located near the equator to benefit from the natural boost provided by the Earth's rotational speed.**

Conclusions

- **All the launch sites are situated in close proximity to coastlines.**
- **Launch success rates have shown a consistent improvement over time.**
- **Payload mass has a positive correlation with launch success across all launch sites.**


SPACEX