IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>
<Date>

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
  - Predictive analysis
- Summary of all results:
  - Exploratory data analysis results
  - Interactive and predictive analysis results

Introduction

- Project background and context:
  In the capstone project, as a data scientist for a startup, Space Y, competing with SpaceX we are using machine learning to analyze data.It highlights the importance of data science in predicting launch costs and reusability of the rocket's first stage.

- Problems you want to find answers:

Space Y tasks us to train a machine learning model to predict successful Stage 1 recovery.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology: Data collected from SpaceX public API and SpaceX Wikipedia page
- Perform data wrangling : Classifying landings as successful and unsuccessful.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models: Tuned models using GridSearchCV to ensure the best results.

Data Collection

- Describe how data sets were collected : Data collection process involved a combination of API requests from Space X public API and web scraping data from a table in Space X's Wikipedia entry.

Space X API Data Columns:

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights,etc.

Wikipedia Webscrape Data Columns:

Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome,etc.

- You need to present your data collection process use key phrases and flowcharts

Data Collection – SpaceX API

- Add the GitHub URL of the completed SpaceX API calls notebook (must include completed code cell and outcome cell), as an external reference and peer-review purpose :https://github.com/Krishitaa/IBM-Applied-Data-Science-Capstone-assignment/tree/main/WEEK%201/DATA%20COLLECTION

Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts
- Add the GitHub URL of the completed web scraping notebook, as an external reference and peer-review purpose : https://github.com/Krishitaa/IBM-Applied-Data-Science-Capstone-assignment/tree/main/WEEK%201/DATA%20WRANGLING

Data Wrangling

- Describe how data were processed:

The dataset categorizes booster landing outcomes as follows:

- **True**: Successful landing (Ocean, RTLS, or ASDS).
- **False**: Failed landing (Ocean, RTLS, or ASDS).

For training, these are labeled as:

- **1**: Successful landing.
- **0**: Unsuccessful landing.

- Add the GitHub URL of your completed data wrangling related notebooks, as an external reference and peer-review purpose: https://github.com/Krishitaa/IBM-Applied-Data-Science-Capstone-assignment/tree/main/WEEK%201/DATA%20WRANGLING

EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts:

Exploratory Data Analysis performed on variables Flight Number, Payload Mass, Launch Site, Orbit, Class and Year.

<u>Plots Used:</u>

Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit vs. Success Rate, Flight Number vs. Orbit, Payload vs Orbit, and Success Yearly Trend,Scatter plots, line charts, and bar plots were used to compare relationships between variables to
decide if a relationship exists so that they could be used in training the machine learning model

- Add the GitHub URL of your completed EDA with data visualization notebook, as an external reference and peer-review purpose : https://github.com/Krishitaa/IBM-Applied-Data-Science-Capstone-assignment/blob/main/WEEK%202/edadataviz.ipynb

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed :
  - Loaded the dataset into the IBM DB2 Database.
  - Queried the database using SQL-Python integration.
  - Explored the dataset to gain insights, including:
    - Launch site names.
    - Mission outcomes.
    - Payload sizes for various customers.
    - Booster versions.
    - Landing outcomes.

- Add the GitHub URL of your completed EDA with SQL notebook, as an external reference and peer-review purpose : https://github.com/Krishitaa/IBM-Applied-Data-Science-Capstone-assignment/blob/main/WEEK%202/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map: Folium maps mark Launch Sites, successful and unsuccessful landings, and a proximity example to key locations: Railway, Highway, Coast, and City.
- Explain why you added those objects: It visualizes successful landings relative to location.
- Add the GitHub URL of your completed interactive map with Folium map, as an external reference and peer-review purpose : https://github.com/Krishitaa/IBM-Applied-Data-Science-Capstone-assignment/blob/main/WEEK%203/lab_jupyter_launch_site_location%20(1).ipynb

Build a Dashboard with Plotly Dash

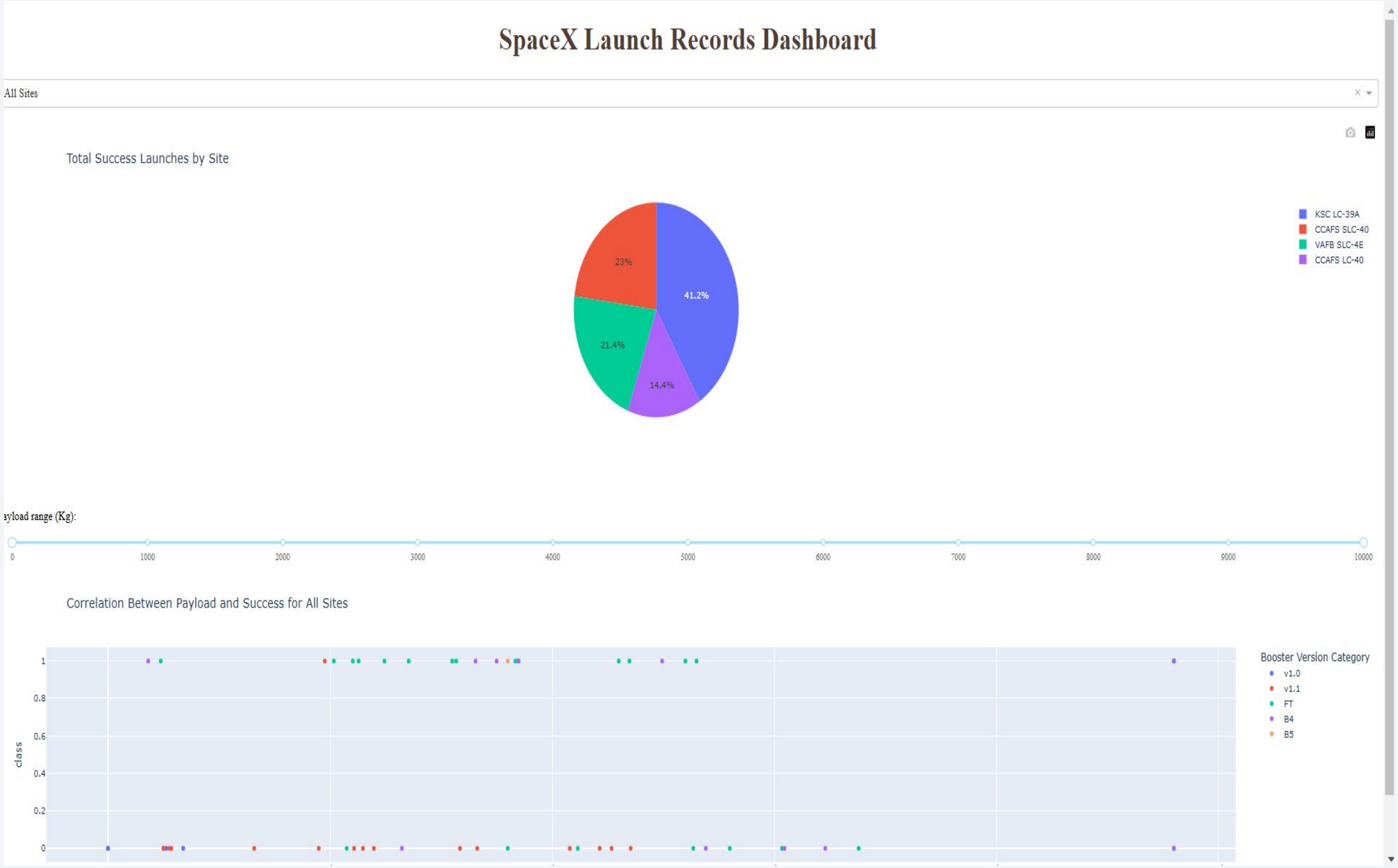
- Summarize what plots/graphs and interactions you have added to a dashboard : Dashboard includes a pie chart and a scatter plot.
- Explain why you added those plots and interactions : The pie chart is used to visualize launch site success rate.The scatter plot can help us see how success varies across launch sites, payload mass, and booster version category
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose : https://github.com/Krishitaa/IBM-Applied-Data-Science-Capstone-assignment/blob/main/WEEK%203/Build%20an%20Interactive%20Dashboard%20with%20Ploty%20Dash

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
- Created a NumPy array from the "Class" column in the dataset.
  - Standardized the data using StandardScaler by fitting and transforming it.
  - Split the data into training and testing sets using the train_test_split function.
  - Created a GridSearchCV object with cv=10 to identify the best parameters.
  - Applied GridSearchCV on multiple models: Logistic Regression, SVM, Decision Tree, and KNN.
  - Calculated accuracy on the test data using the .score() method for all models.
  - Examined the confusion matrix for each model.
  - Evaluated the performance of each model by analyzing Jaccard_score and F1_score metrics to determine the best-performing method.

- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose : https://github.com/Krishitaa/IBM-Applied-Data-Science-Capstone-assignment/blob/main/WEEK%204/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

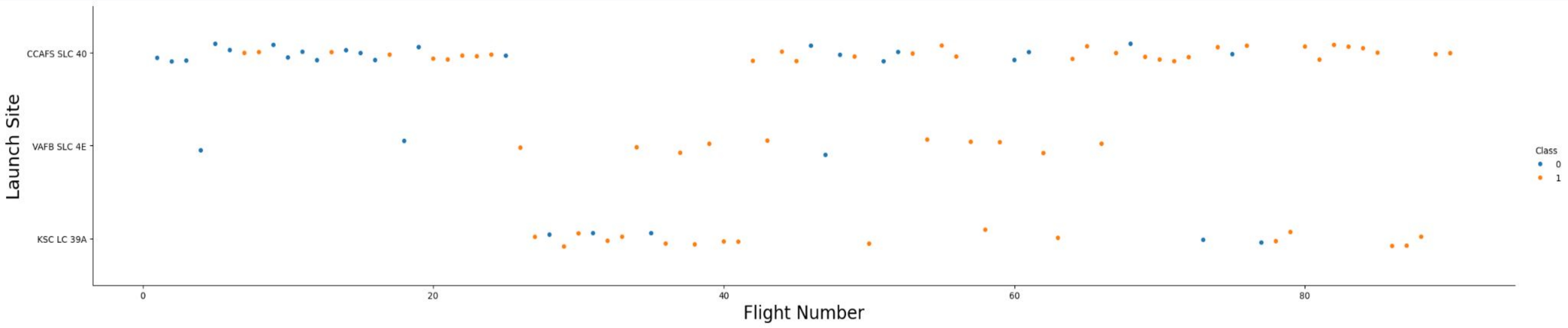


SpaceX Launch Records Dashboard
All Sites
Total Success Launches by Site
41.2%
23%
21.4%
14.4%
KSC LC-39A
CCAFS SLC-40
VAFB SLC-4E
CCAFS LC-40
ayload range (Kg):
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
Correlation Between Payload and Success for All Sites
class
1
0.8
0.6
0.4
0.2
0
Booster Version Category
v1.0
v1.1
FT
B4
B5

Section 2

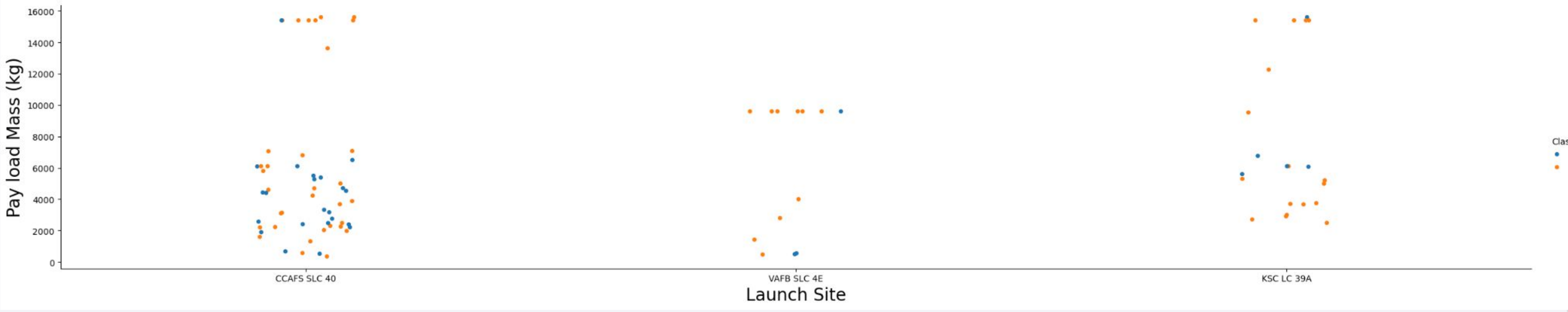
Insights drawn from EDA

Flight Number vs. Launch Site



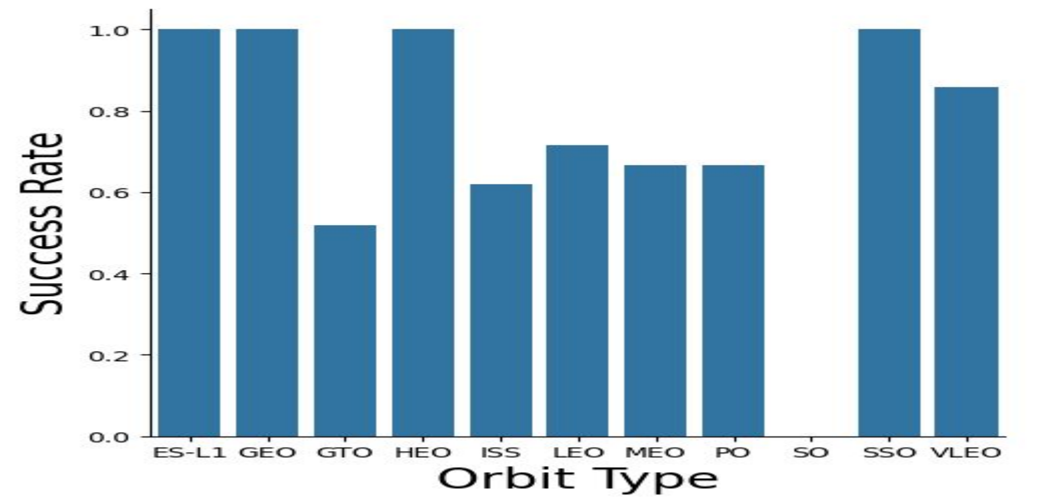
- Early flights failed, while recent flights succeeded.
- CCAFS SLC 40 accounts for about half of all launches.
- VAFB SLC 4E and KSC LC 39A show higher success rates.
- Success rates improve with newer launches.

Payload vs. Launch Site



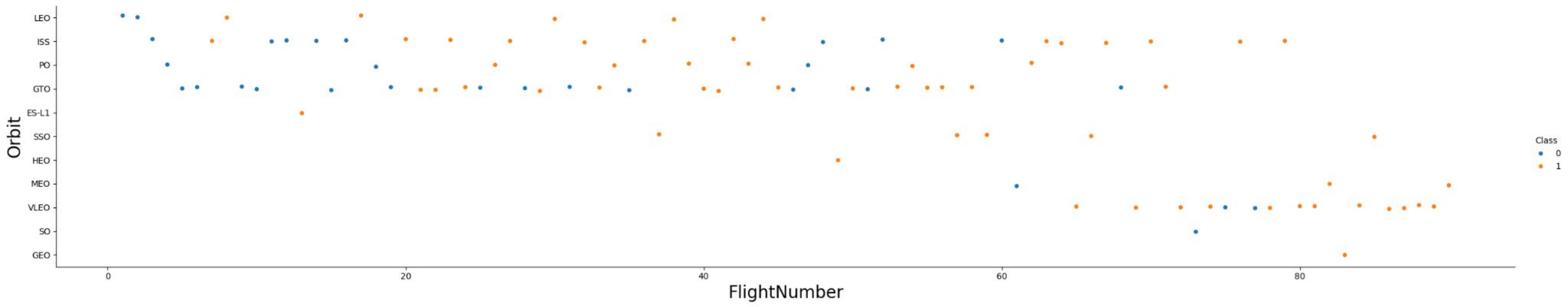
For every launch site the higher the payload mass, the higher the success rate.

Success Rate vs. Orbit Type



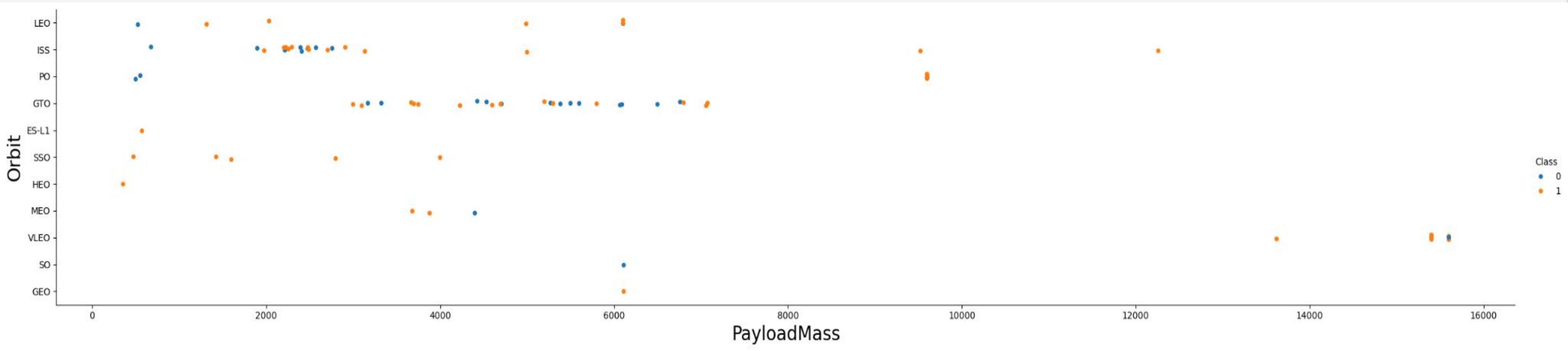
- **100% success rate orbits**: ES-L1, GEO, HEO, SSO
- **0% success rate orbit**: SO
- **50%-85% success rate orbits**: GTO, ISS, LEO, MEO, PO

Flight Number vs. Orbit Type



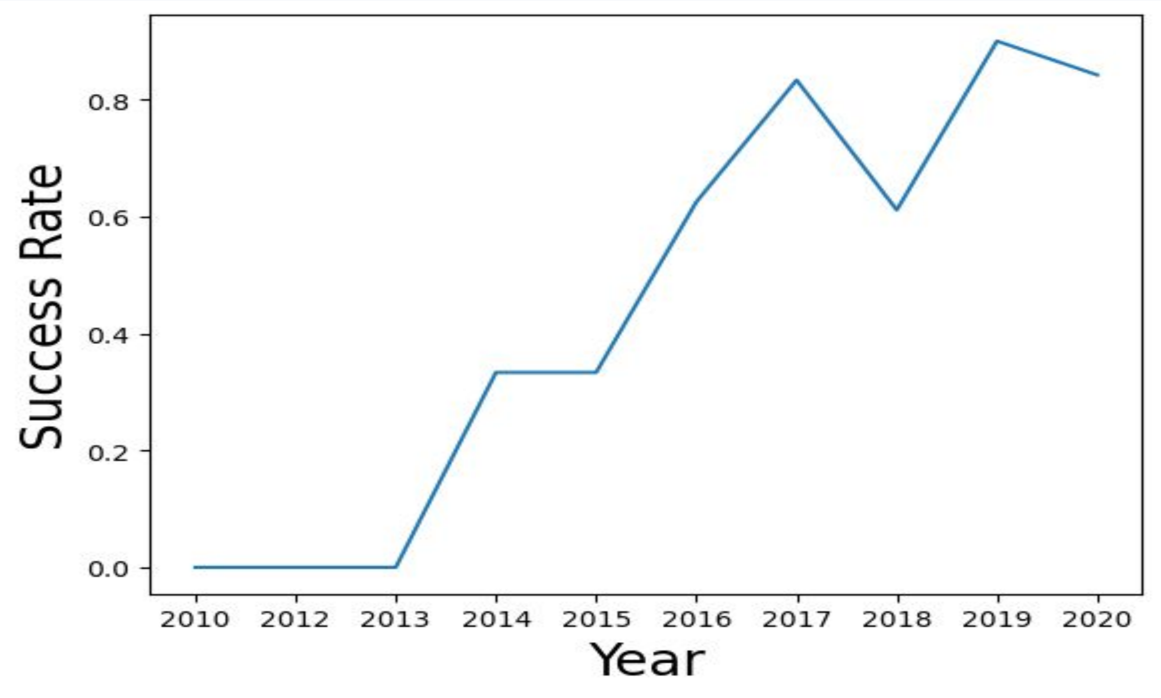
In LEO orbit, success is linked to the number of lights. However, in GTO orbit, there seems to be no correlation.

Payload vs. Orbit Type



GTO and Polar LEO (ISS) orbits are positively impacted by heavy payloads, whereas GTO orbits are negatively impacted.

Launch Success Yearly Trend



The success rate since 2013 kept increasing till 2020.

All Launch Site Names

Task 1

Display the names of the unique launch sites in the space mission

In [10]:
```
%sql select distinct(Launch_Site) from SPACEXTABLE
```

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

Out[10]:

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

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

Launch Site Names Begin with 'CCA'

Task 2

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

In [11]:
```
%sql select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5
```

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

Out[11]:

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

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

Total Payload Mass

Task 3

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

In [12]:
```
%sql select sum("PAYLOAD_MASS__KG_") as TotalPayload from SPACEXTABLE where customer='NASA (CRS)';
```

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

Out[12]:

TotalPayload
45596

Displaying total payload mass.

Average Payload Mass by F9 v1.1

Task 4

Display average payload mass carried by booster version F9 v1.1

In [13]:

```
%sql select avg("PAYLOAD_MASS__KG_") as AverageMass from SPACEXTABLE where Booster_Version='F9 v1.1'
```

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

Out[13]:

AverageMass
2928.4

Display average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

In [14]:
```
%sql select distinct(Landing_Outcome) from SPACEXTABLE
```

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

Out[14]:

Landing_Outcome
Failure (parachute)
No attempt
Uncontrolled (ocean)
Controlled (ocean)
Failure (drone ship)
Precluded (drone ship)
Success (ground pad)
Success (drone ship)
Success
Failure
No attempt

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

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

In [16]:

```
%sql select Booster_Version from SPACEXTABLE where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4(
```

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

Out[16]:

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

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

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

In [17]:
```
%sql select mission_outcome, count(*) as total_number from SPACEXTABLE group by mission_outcome;
```

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

Out[17]:

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

The total number of successful and failure mission outcomes.

Boosters Carried Maximum Payload

Task 8

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

In [18]:
```
%sql select booster_version from SPACEXTABLE  where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTABLE)
```

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

Out[18]:

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

Booster_versions which have carried the maximum payload mass.

2015 Launch Records

Task 9

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

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

In [19]:

```
%%sql select strftime('%m', date) as month, date, booster_version, launch_site, Landing_Outcome from SPACEXTABLE
      where Landing_Outcome = 'Failure (drone ship)' and substr(Date,0,5)='2015';
```

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

Out[19]:

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

2015 launch records

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

```
%%sql
SELECT landing__outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
WHERE landing__outcome LIKE 'Succes%' AND DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY landing__outcome
ORDER BY no_outcome DESC;
```

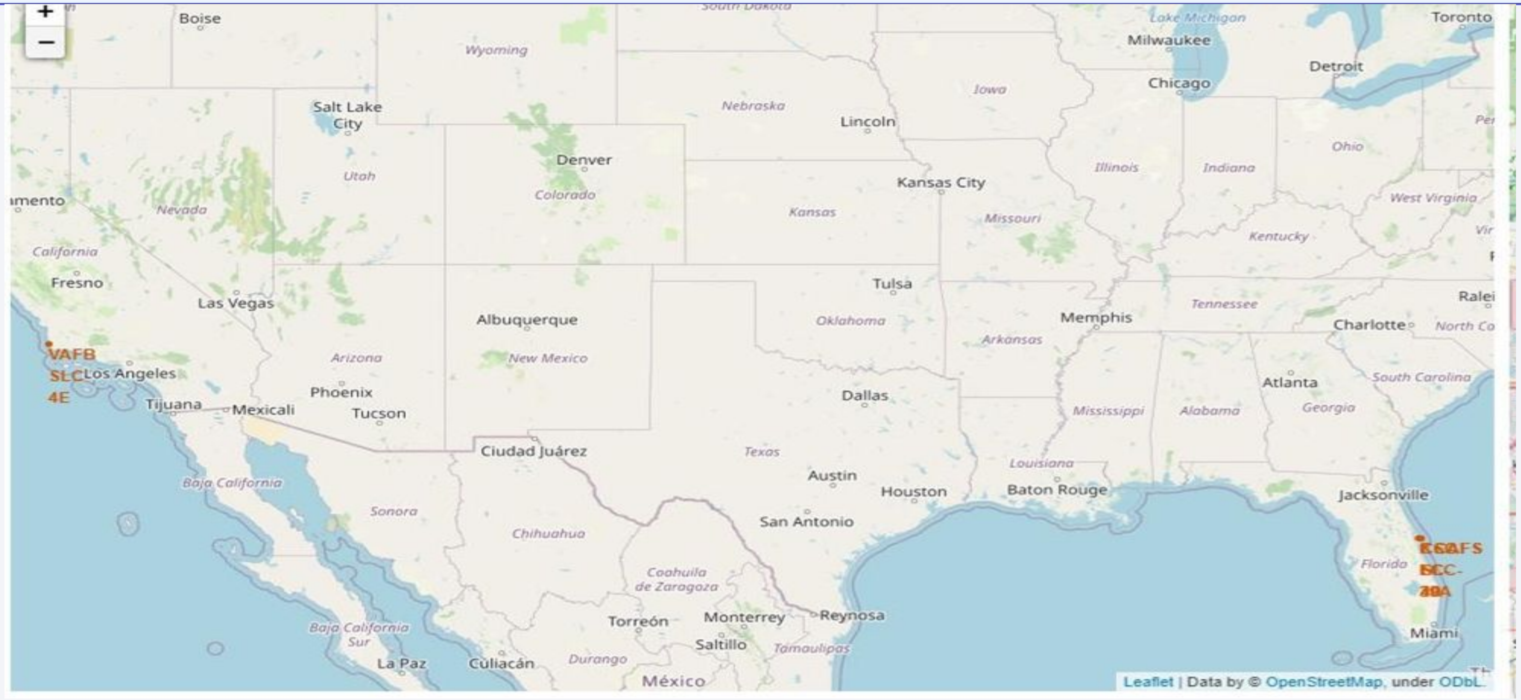
```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg
Done.
```

landing__outcome	no_outcome
Success (drone ship)	5
Success (ground pad)	3

Section 3

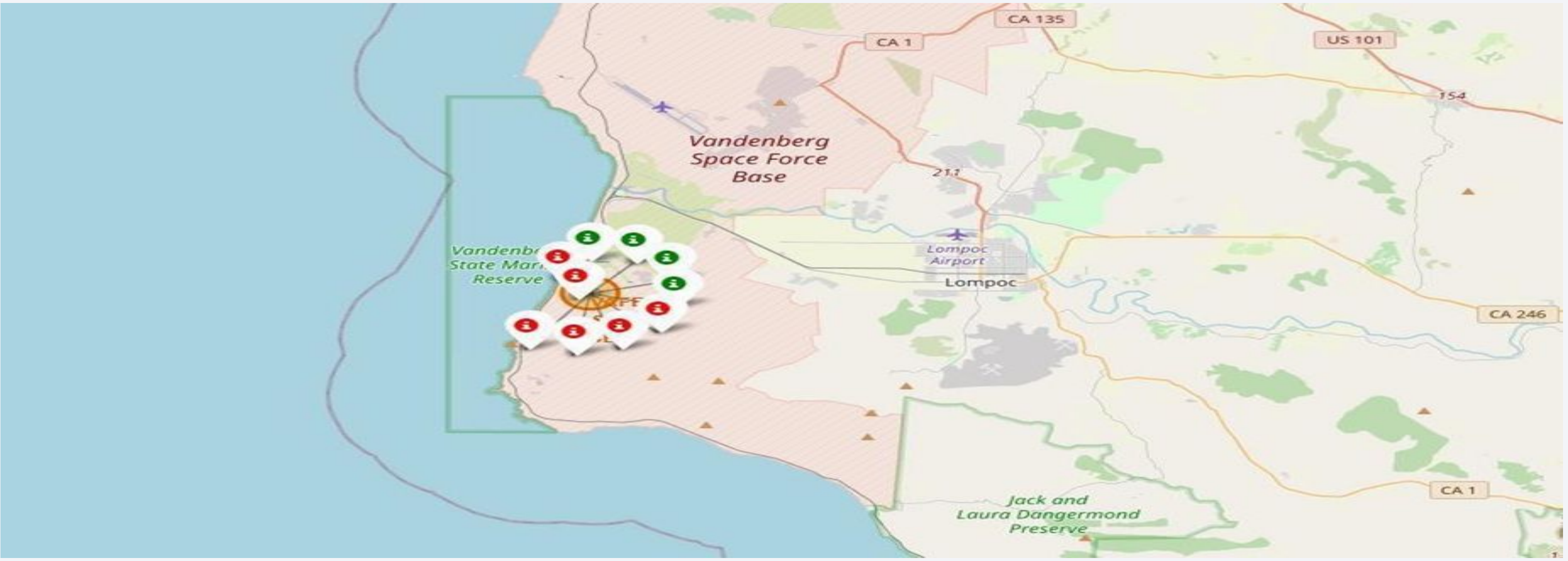
Launch Sites Proximities Analysis

Launch Site Locations



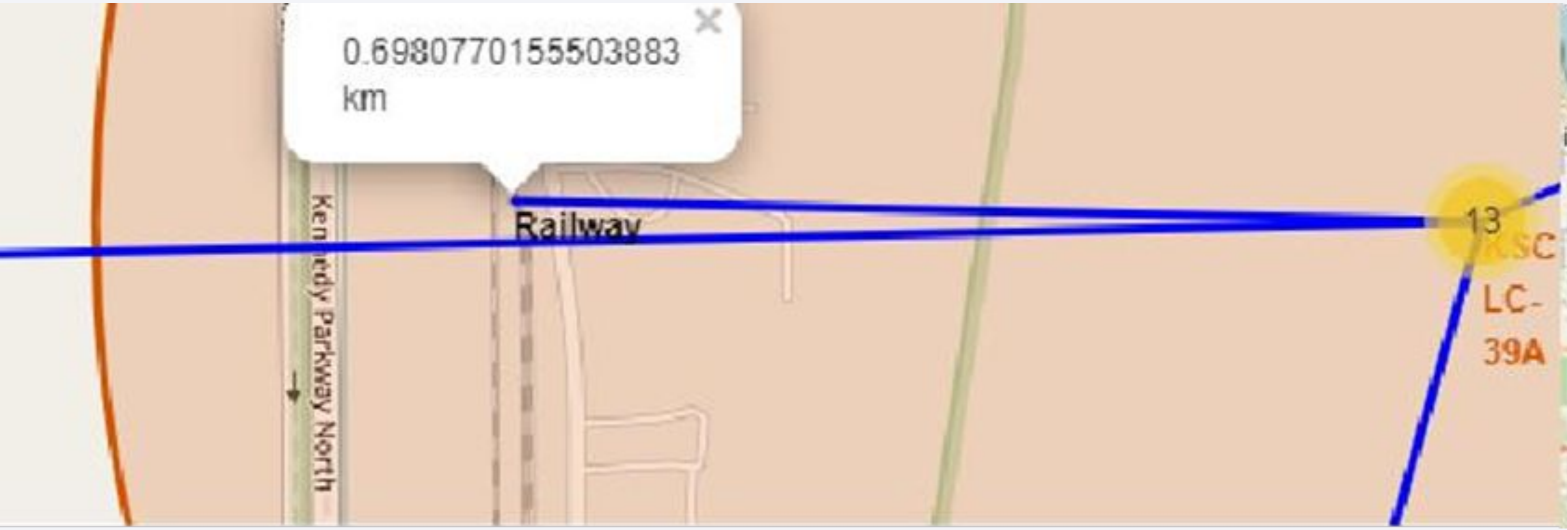
The map shows all launch sites relative to US map.

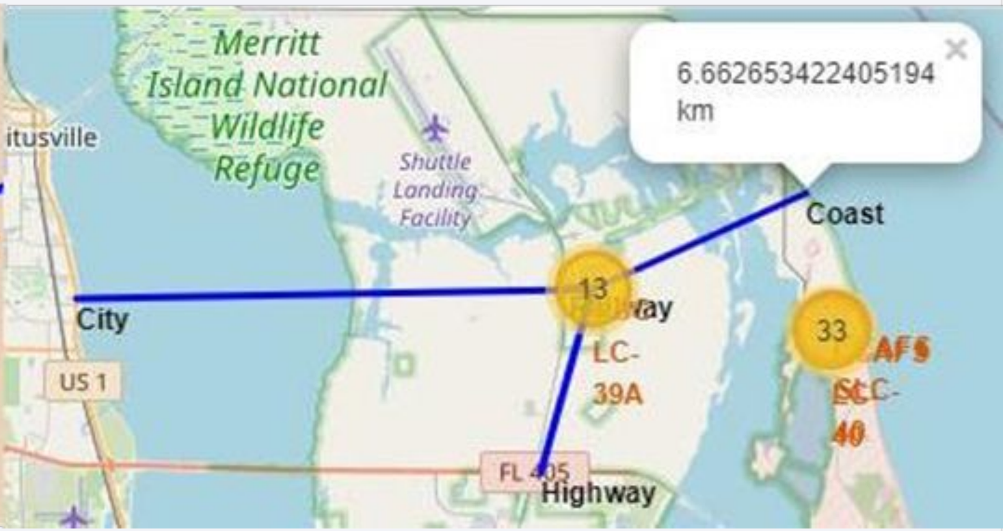
Color-Coded Launch Markers



Clusters on Folium map can be clicked on to display each successful landing (green icon) and failed landing (red icon).

Key Location Proximities



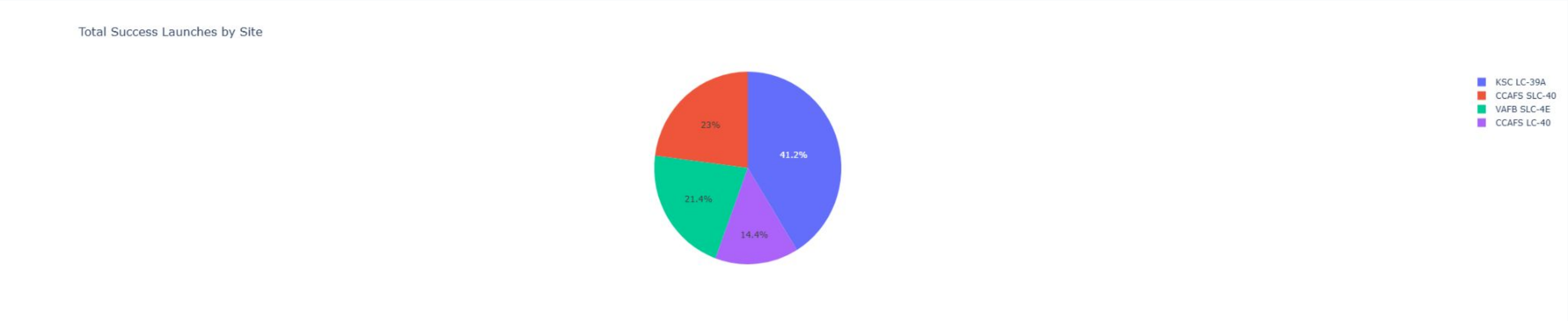


Launch sites are close to highways for human and supply transport. Launch sites are also close to coasts and relatively far from cities.

Section 4

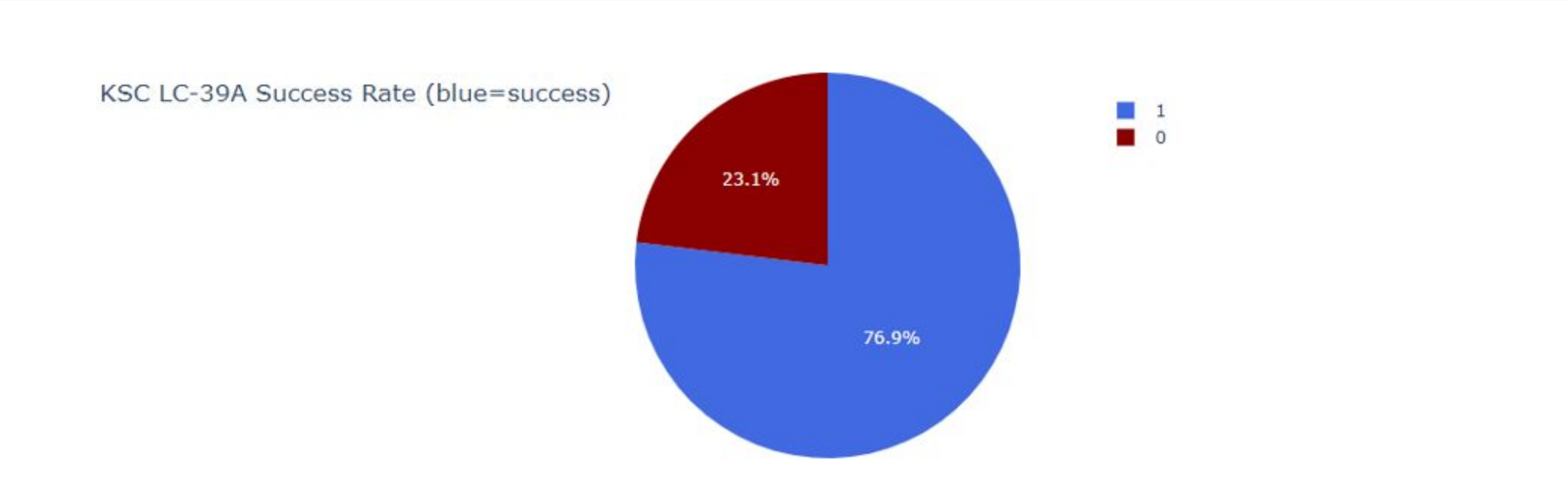
Build a Dashboard with Plotly Dash

Successful Launches Across Launch Sites



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

Highest Success Rate Launch Site



KSC LC-39A has the highest launch success rate.

Payload Mass vs. Success vs. Booster Version Category

The charts shows the payload b/w 1500 to 2000 have highest success rate.

Section 5

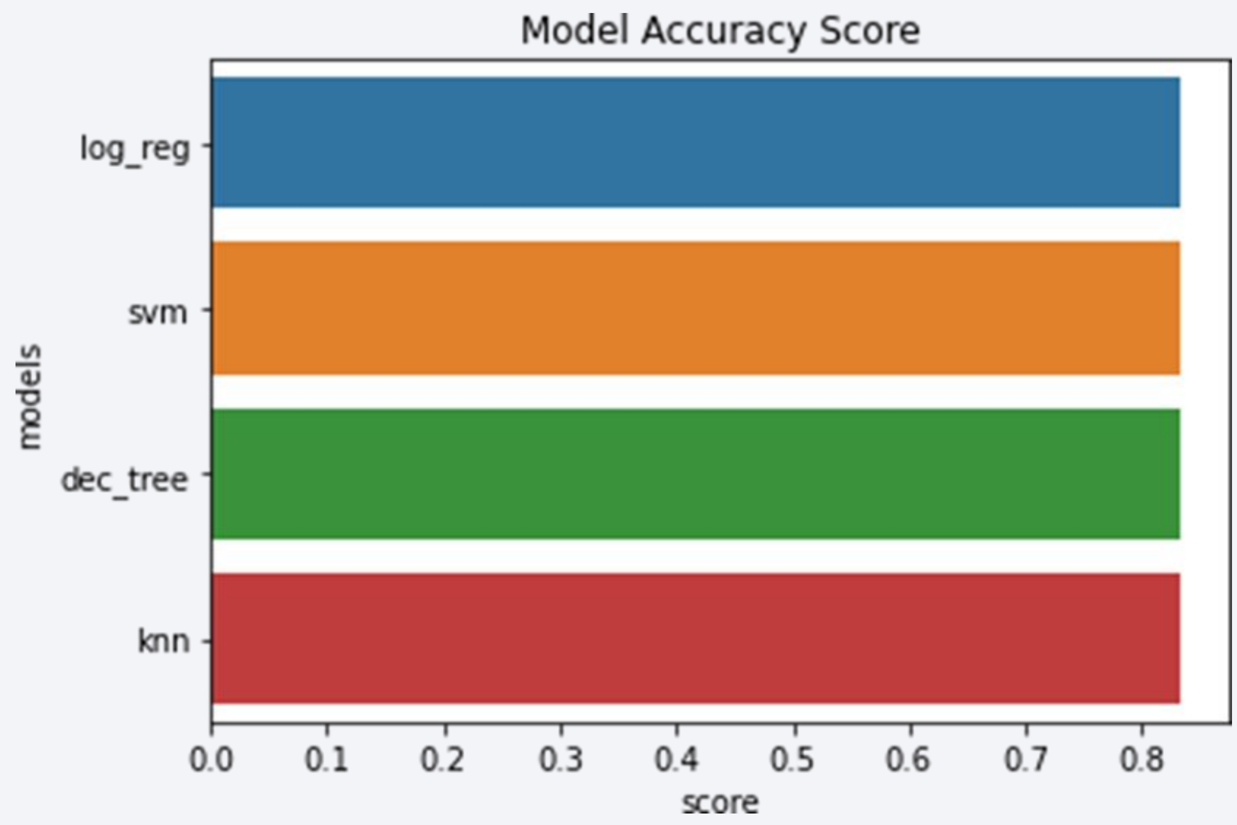
Predictive Analysis (Classification)

Classification Accuracy

Based on the scores of the Test Set,

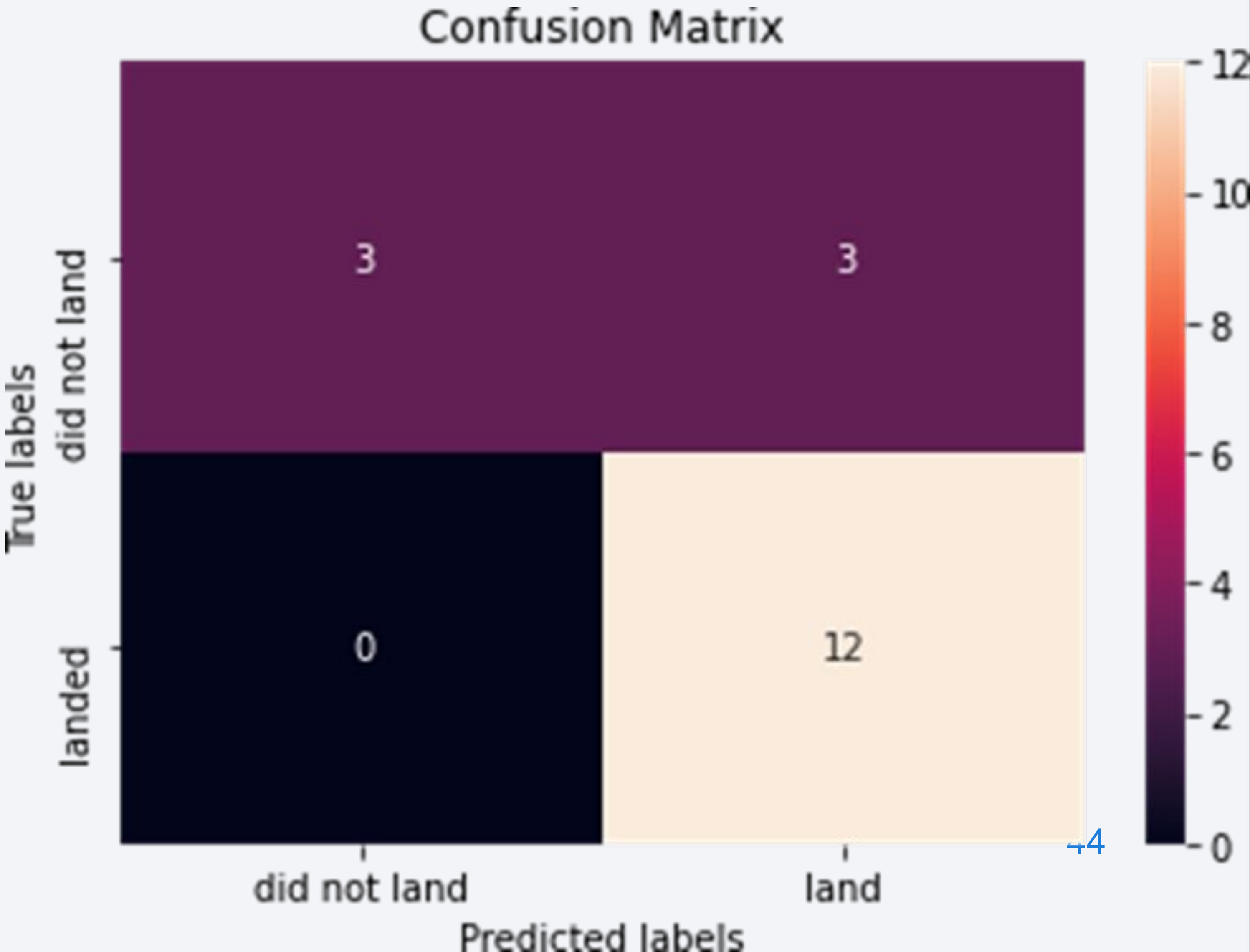
we can not conirm which method

performs best.



Confusion Matrix

Since all models performed the same for the test set, the confusion matrix is the same across all models.



Conclusions

- **Task Overview**: Develop a machine learning model for Space Y to predict successful Stage 1 landings, aiming to save ~$100 million USD.
- **Data Sources**:
  - Used data from a public SpaceX API.
  - Web scraped information from the SpaceX Wikipedia page.
- **Data Processing**:
  - Created data labels for predictive modeling.
  - Stored the processed data in a DB2 SQL database.
- **Visualization**:
  - Developed a dashboard for visualization of the model's predictions and insights.

Appendix

- GITHUB LINK: https://github.com/Krishitaa/IBM-Applied-Data-Science-Capstone-assignment/tree/main

Thank you!