IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Narges Bagheri
January 17, 2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection with Webscraping and API
  - Data wrangling
  - Exploratory Data Analysis (EDA) using SQL
  - EDA data visualization using Pandas and Matplotlib libraries
  - Analyze launch records interactively with Plotly
  - Build an interactive map to analyze the launch site proximity with Folium python Library
  - Machine learning landing prediction
- Summary of all results
  - EDA results
  - Interactive Visual Analytics
  - Predictive Analysis(Classification)

Introduction

- Project background and context

SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.
Unlike other rocket providers, SpaceX's Falcon 9 Can recover the first stage. Sometimes the first stage does not land. Sometimes it will crash. Other times, Space X will sacrifice the first stage due to the mission parameters like payload, orbit, and customer.

- Problems you want to find answers

We will predict if the Falcon 9 first stage will land successfully

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API
  - Lunch records with BeautifulSoup
- Perform data wrangling
  - Create a landing outcome label
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Split data into test and training data.
  - Find the best Hyperparameter for SVM, Classification Trees, and Logistic Regression
  - Calculate the accuracy on the test data

Data Collection

- Define a series of helper functions that will help us use the API to extract information using identification numbers in the launch data:

  To learn the booster name

  To know the name of the launch site being used

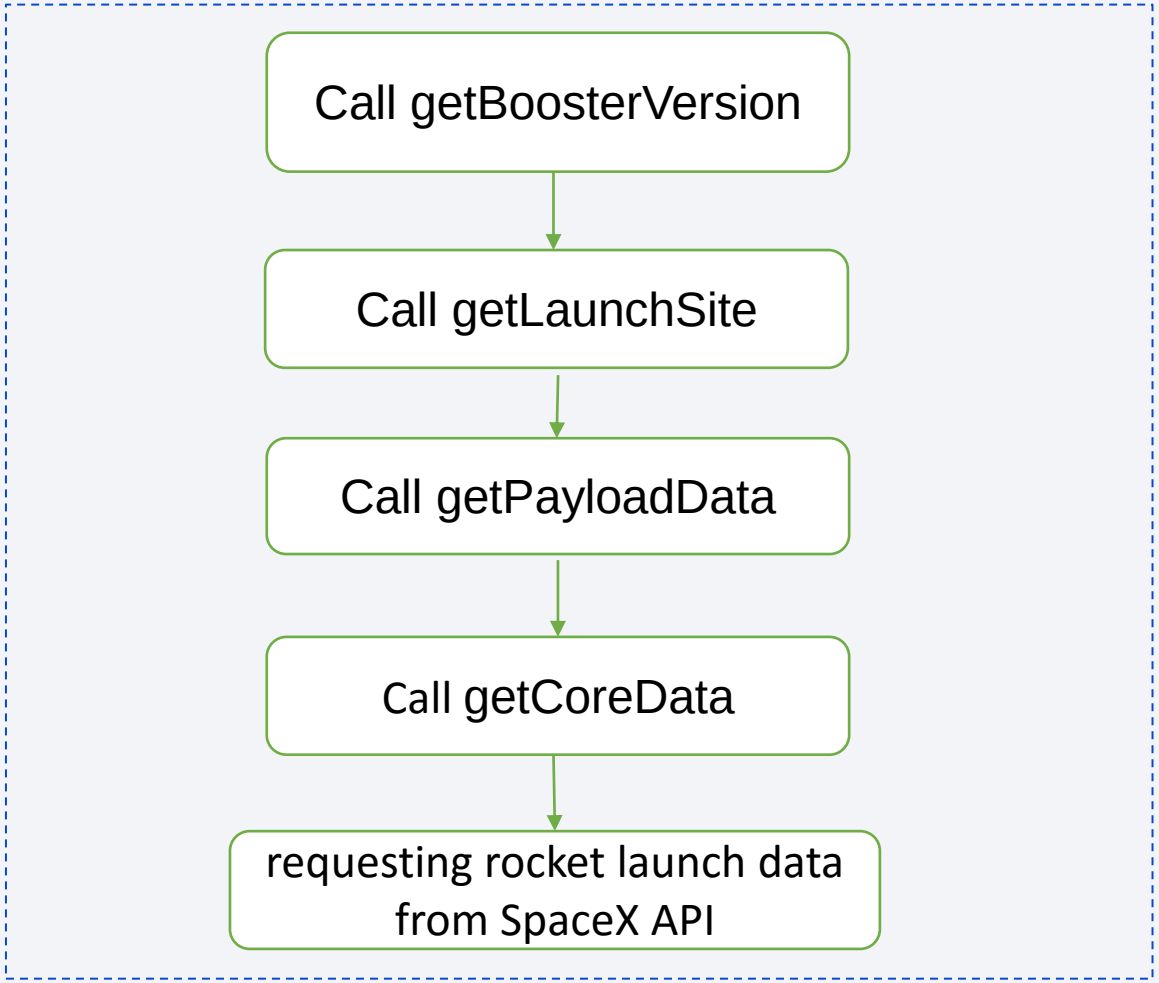
  To learn the mass of the payload and the orbit that it is going to

  To learn the outcome of the landing, the type of the landing, number of flights with that core, whether gridfins were used, wheter the core is reused, wheter legs were used, the landing pad used, the block of the core which is a number used to seperate version of cores, the number of times this specific core has been reused, and the serial of the core.

- Request and parse the SpaceX launch data using the GET request
- Filter the dataframe to only include Falcon 9 launches

Data Collection – SpaceX API

- Define a series of helper functions that will help us use the API to extract information using identification numbers in the launch data.
- Request and parse the SpaceX launch data using the GET request



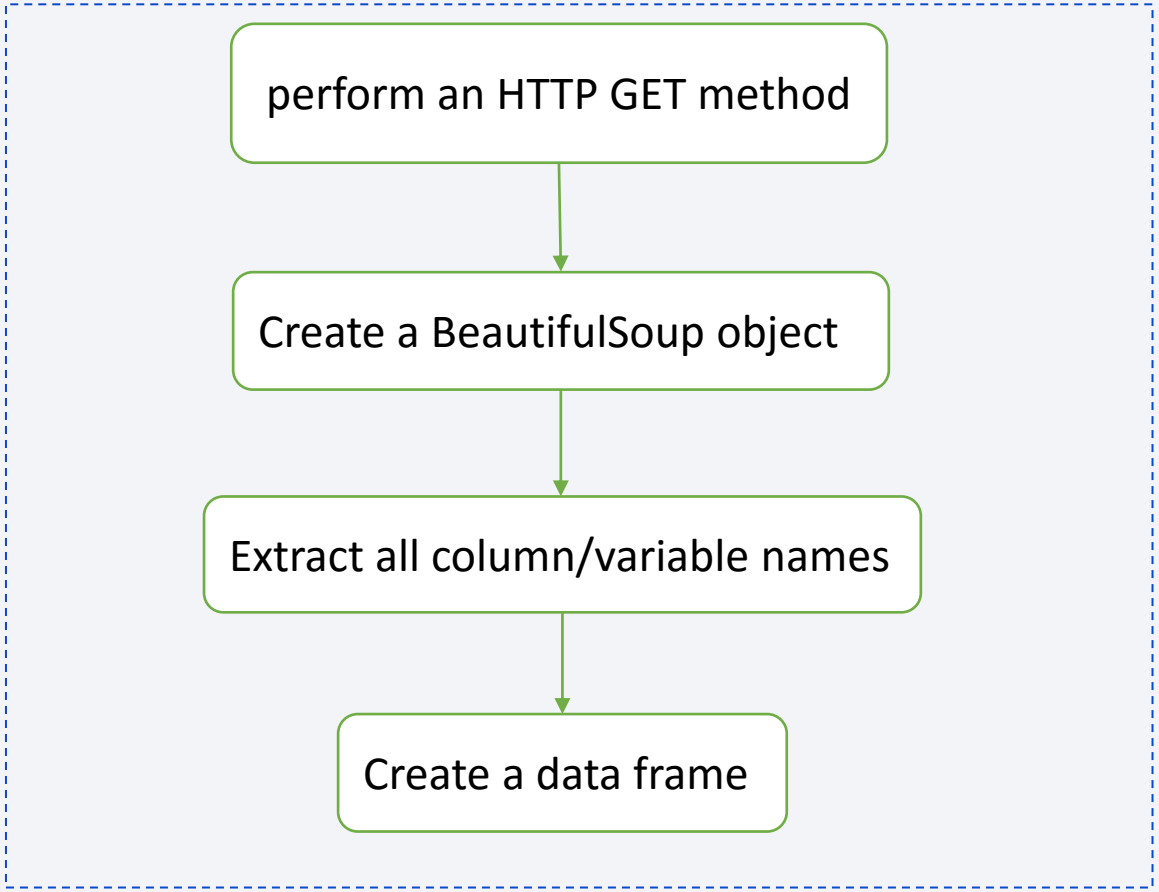
https://github.com/NarBagheri/IBM_Capstone/blob/main/1-1-spacex-data-collection-api.ipynb

Data Collection - Scraping

Web scrap Falcon 9 launch records with BeautifulSoup:

- Extract a Falcon 9 launch records HTML table from Wikipedia

- Parse the table and convert it into a Pandas data frame



https://github.com/NarBagheri/IBM_Capstone/blob/main/1-2-webscraping.ipynb

Data Wrangling

- Dealing with Missing Values
    - Calculate the mean for the PayloadMass
    - replace np.nan values in the data with the mean
- Calculate the number of launches on each site
- Calculate the number and occurence of mission outcome per orbit type
- Create a landing outcome label from Outcome column

https://github.com/NarBagheri/IBM_Capstone/blob/main/1-3-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Scatter plot

To visualize the relationship between FlightNumber and PayloadMass, Flight Number and Launch Site, Payload and Launch Site, FlightNumber and Orbit type, Payload and Orbit type

- Bar chart

To visualize the relationship between success rate of each orbit type

- Line chart

To visualize the launch success yearly trend

https://github.com/NarBagheri/IBM_Capstone/blob/main/2-2-exploratory-analysis-using-pandas.ipynb

EDA with SQL

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved.
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster_versions which have carried the maximum payload mass
- Listing the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015
- Ranking the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017

https://github.com/NarBagheri/IBM_Capstone/blob/main/2-1-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Mark all launch sites on a map to add a highlighted circle area with a text label on a specific coordinate.
- Mark the success/failed launches for each site to see which sites have high success rates
- Plot distance lines to the proximities to answer the following questions:
  - Are launch sites in close proximity to railways?
  - Are launch sites in close proximity to highways?
  - Are launch sites in close proximity to coastline?
  - Do launch sites keep certain distance away from cities?

https://github.com/NarBagheri/IBM_Capstone/blob/main/3_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
- Explain why you added those plots and interactions
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

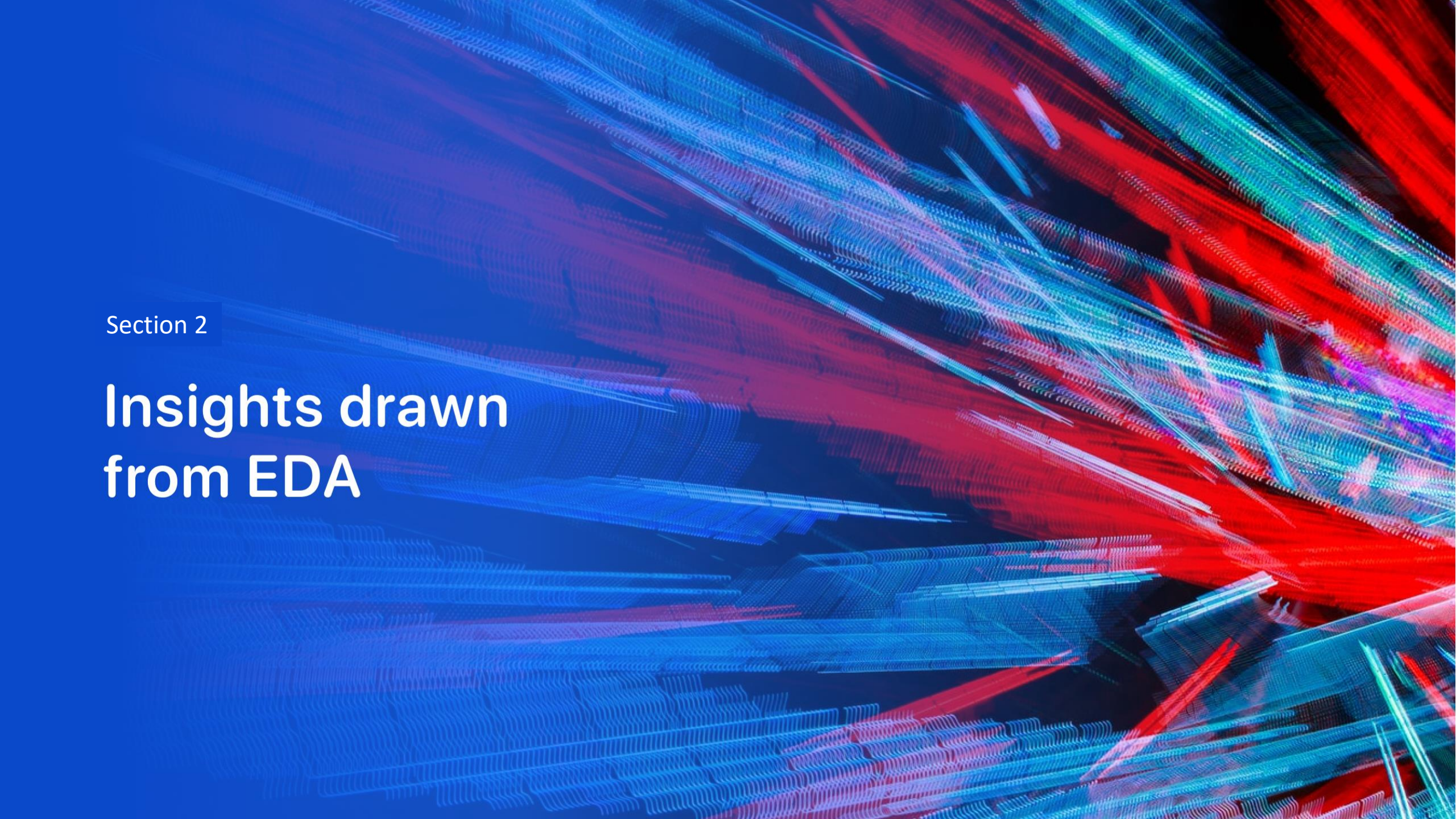
Predictive Analysis (Classification)

- Creating a NumPy array from the column Class in data and Standardize the data in X
- Useing the function train_test_split to split the data X and Y into training and test data
  - Set the parameter test_size to 0.2 and random_state to 2
- Creating a logistic regression object then creating a GridSearchCV object logreg_cv with cv = 10
- Creating a support vector machine object then creating a GridSearchCV object svm_cv with cv - 10
- Creating a decision tree classifier object then creating a GridSearchCV object tree_cv with cv = 10
- Creating a k nearest neighbors object then creating a GridSearchCV object knn_cv with cv = 10
- Calculating the accuracy of the above models on the test data

https://github.com/NarBagheri/IBM_Capstone/blob/main/4_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

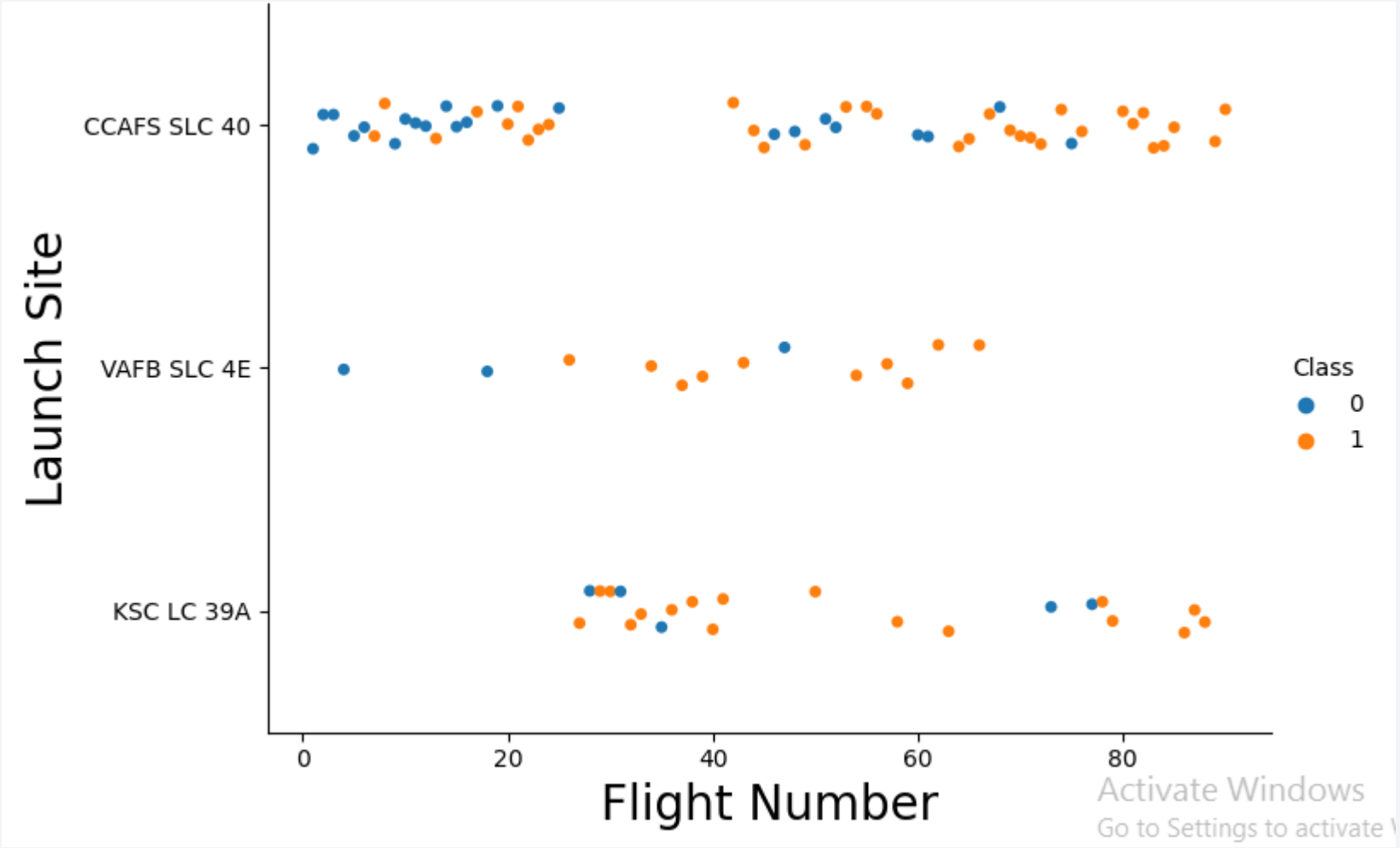
Results

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

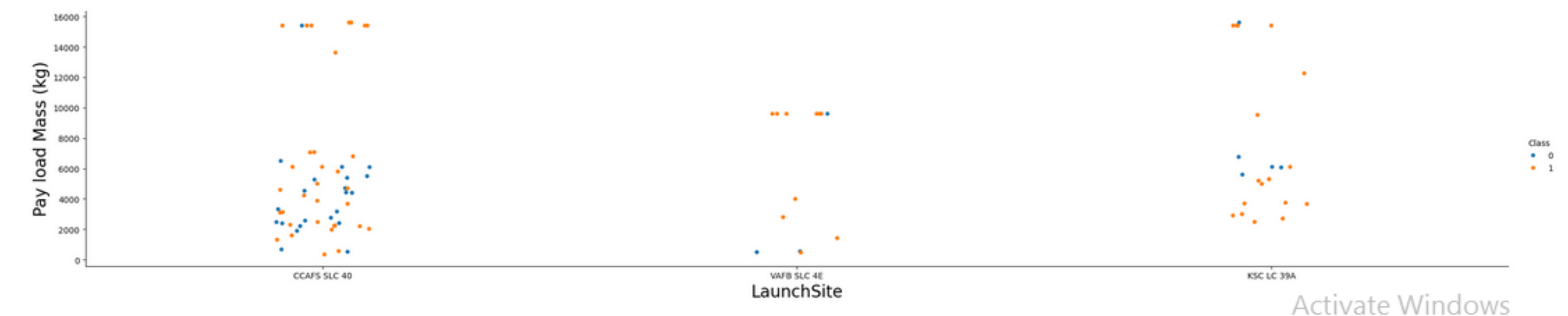

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

As the flight number increases in each of the 3 launcg sites, so does the success rate.



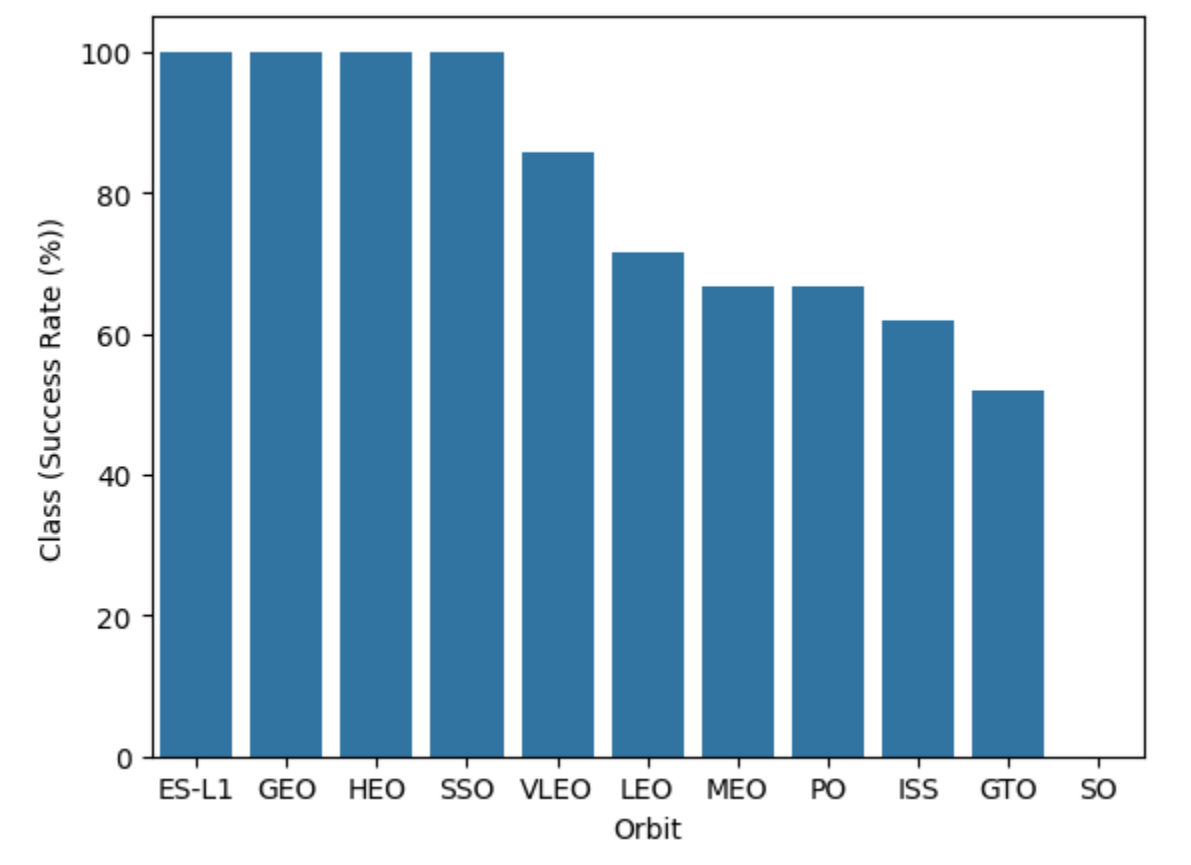
Payload vs. Launch Site



For the VAFB-SLC launchsite there are no rockets launched for heavypayload mass (greater than 10000).

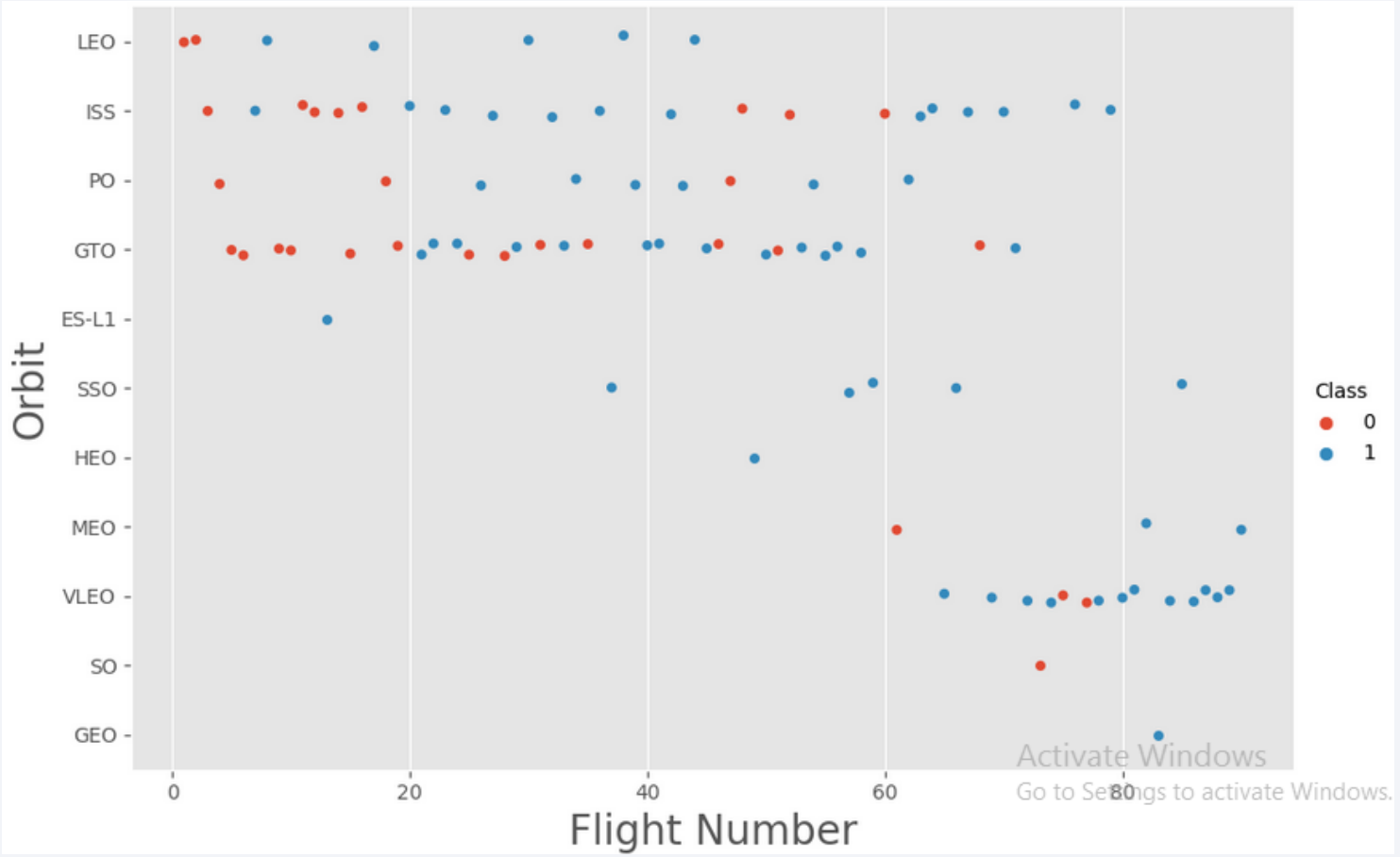
Success Rate vs. Orbit Type

Orbits ES-L1, GEO, HEO & SSO have the highest success rates at 100%, with SO orbit having the lowest success rate at ~50%. Orbit SO has 0% success rate.



Flight Number vs. Orbit Type

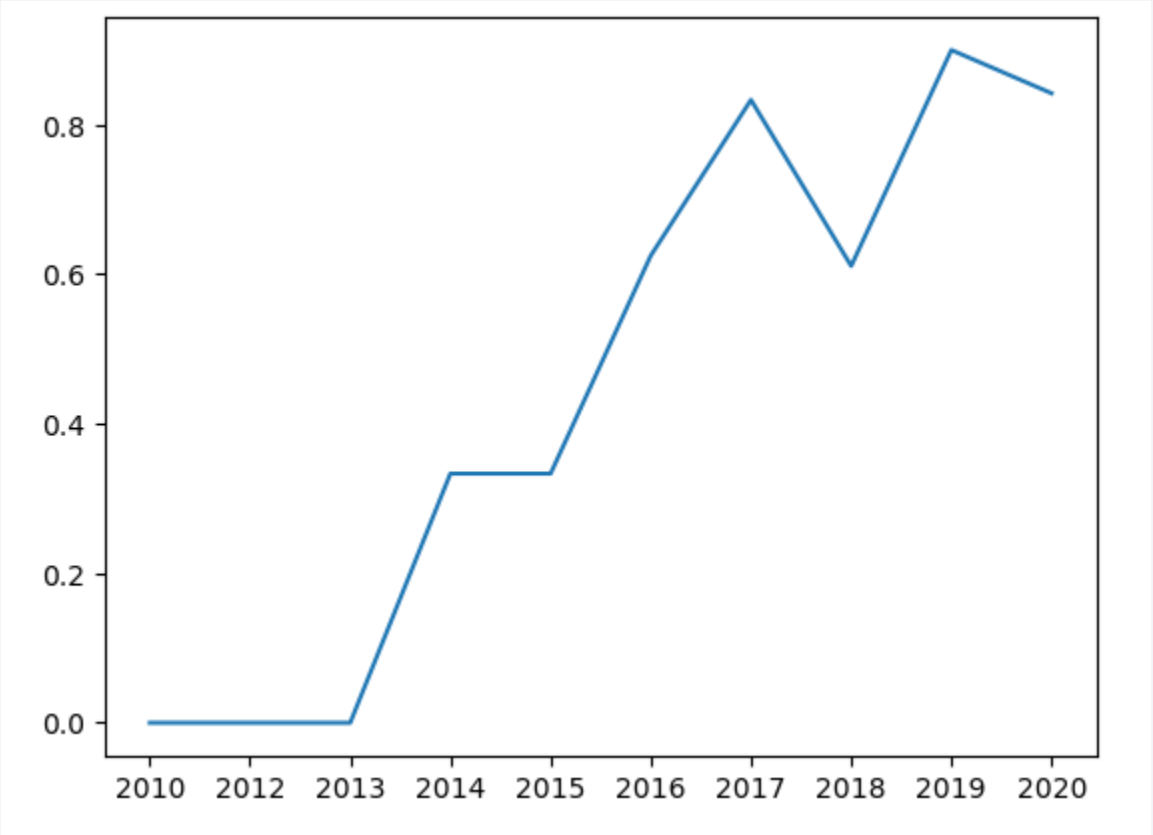
In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



Payload vs. Orbit Type

With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS

Launch Success Yearly Trend



The success rate since 2013 kept increasing till 2020

All Launch Site Names

To find the names of the unique launch SELECT Unique(LAUNCH_SITE) FROM SPACEXTBL is used.

The names of the unique launch sites in the space mission

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

Launch Site Names Begin with 'CCA'

Used ‘LIKE’ command with ‘%’ wildcard in ‘WHERE’ clause to select and display a table of 5 records where launch sites begin with `CCA`

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

The total payload carried by boosters from NASA is calculated by SUM function

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) \
    FROM SPACEXTBL \
    WHERE CUSTOMER = 'NASA (CRS)';
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- To calculate the average payload mass carried by booster version F9 v1.1

```
%sql SELECT AVG (PAYLOAD_MASS__KG_) \
    FROM SPACEXTBL \
    WHERE Booster_Version = "F9 v1.1";
```

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

AVG (PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad

```
%sql SELECT MIN (Date)\
      FROM SPACEXTBL \
     WHERE Landing_Outcome = 'Success (ground pad)';
```

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

MIN (Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

Total Number of Successful and Failure Mission Outcomes

- To calculate the total number of successful and failure mission outcomes

```
%sql SELECT Mission_Outcome, COUNT(*) as Total_number \
     FROM SPACEXTBL \
     GROUP BY Mission_Outcome;
```

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

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

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass

```
%sql SELECT Booster_Version \
     FROM SPACEXTBL \
     WHERE PAYLOAD_MASS__KG_ = (SELECT MAX (PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

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

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

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

- The count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) are ranked between the date 2010-06-04 and 2017-03-20, in descending order

```
%sql SELECT Landing_Outcome, COUNT (*) as Total_number\
     FROM SPACEXTBL \
     WHERE Date between '2010-06-04' and '2017-03-20'\
     GROUP BY Landing_Outcome \
     ORDER BY Total_number DESC;
```

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

Landing_Outcome	Total_number
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

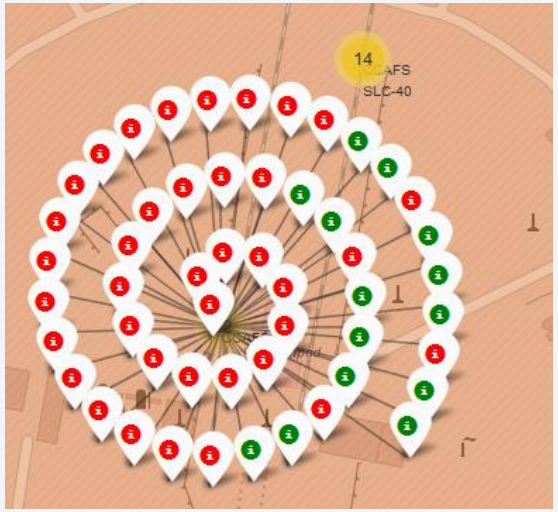
Launch Sites Proximities Analysis

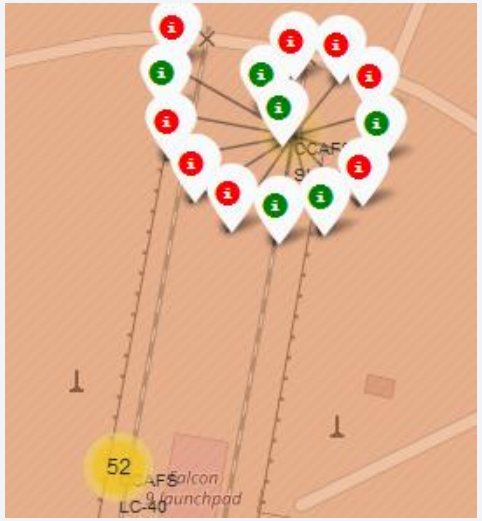
Markers of all launch sites on a map

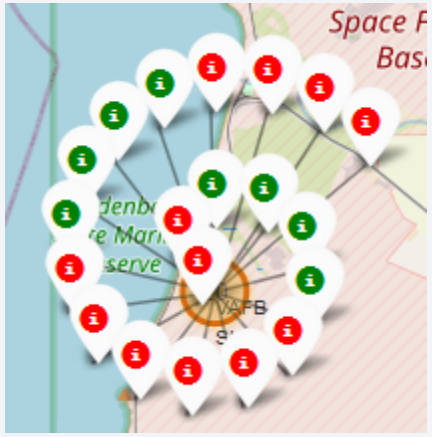
All launch sites are in proximity to the Equator, (located southwards of the US map). Also all the laumch sites are in very close proximity to the coast.

The success/failed launches for each site on the map

In the Eastern coast (Florida) Launch site KSC LC-39A has relatively high success rates compared to CCAFS SLC-40 & CCAFS LC-40.

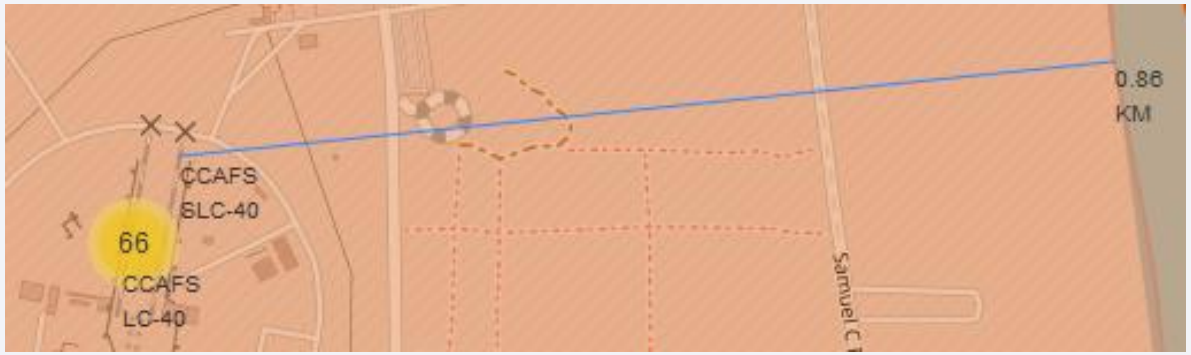




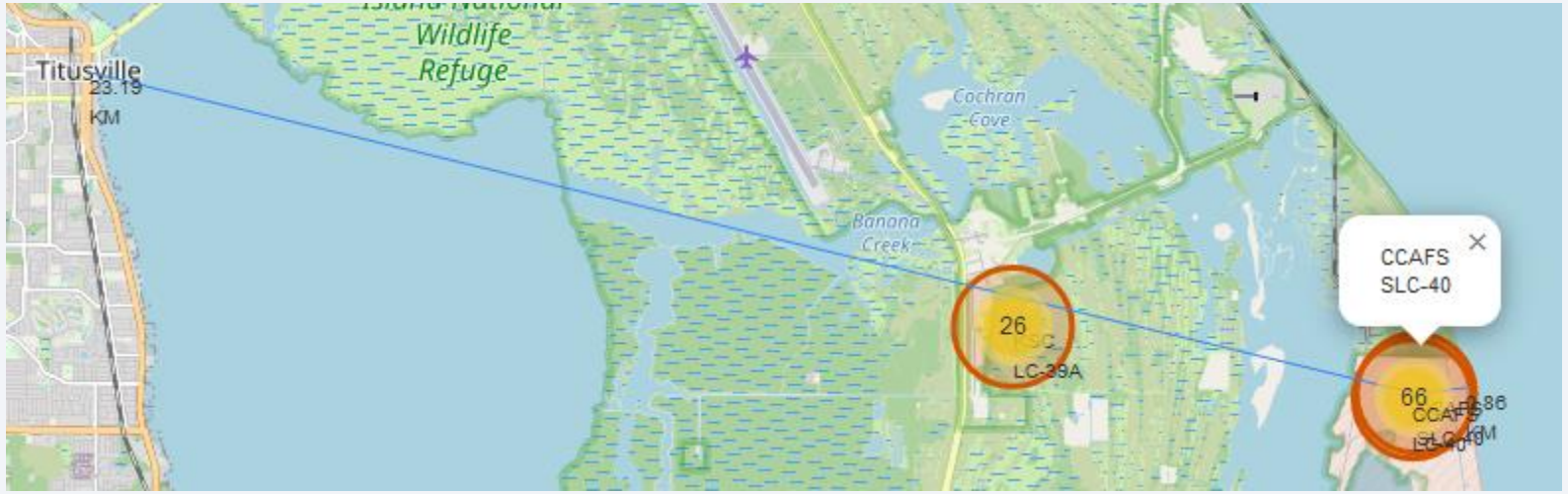


In the West Coast (Californai) Launch site VAFB SLC-4E has relatively lower success rates 8/20 compared to KSC LC-39A launch site in the Eastern Coast of Florida.

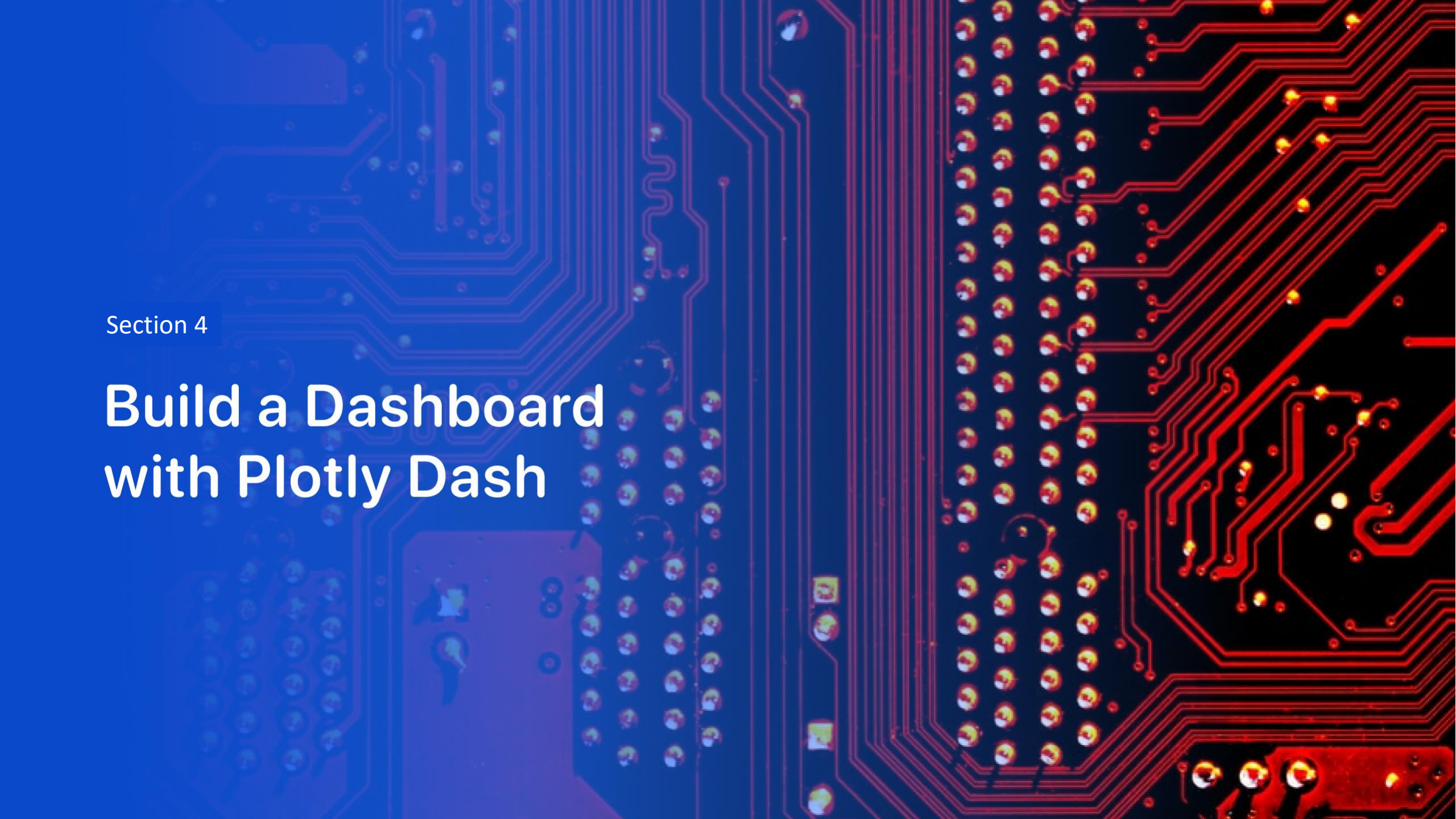
Distances between a launch site to its proximities



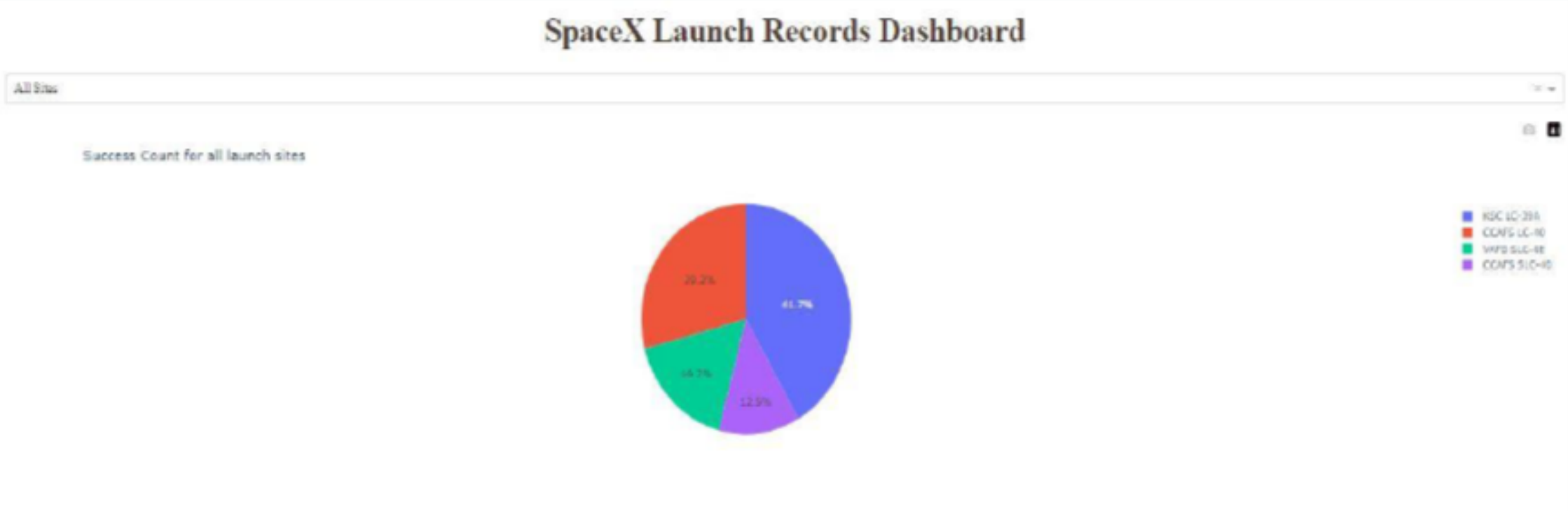
- Launch site CCAFS SLC-40 proximity to coastline is 0.86km



Launch site CCAFS SLC-40 closest to highway (Washington Avenue) is 23.19km

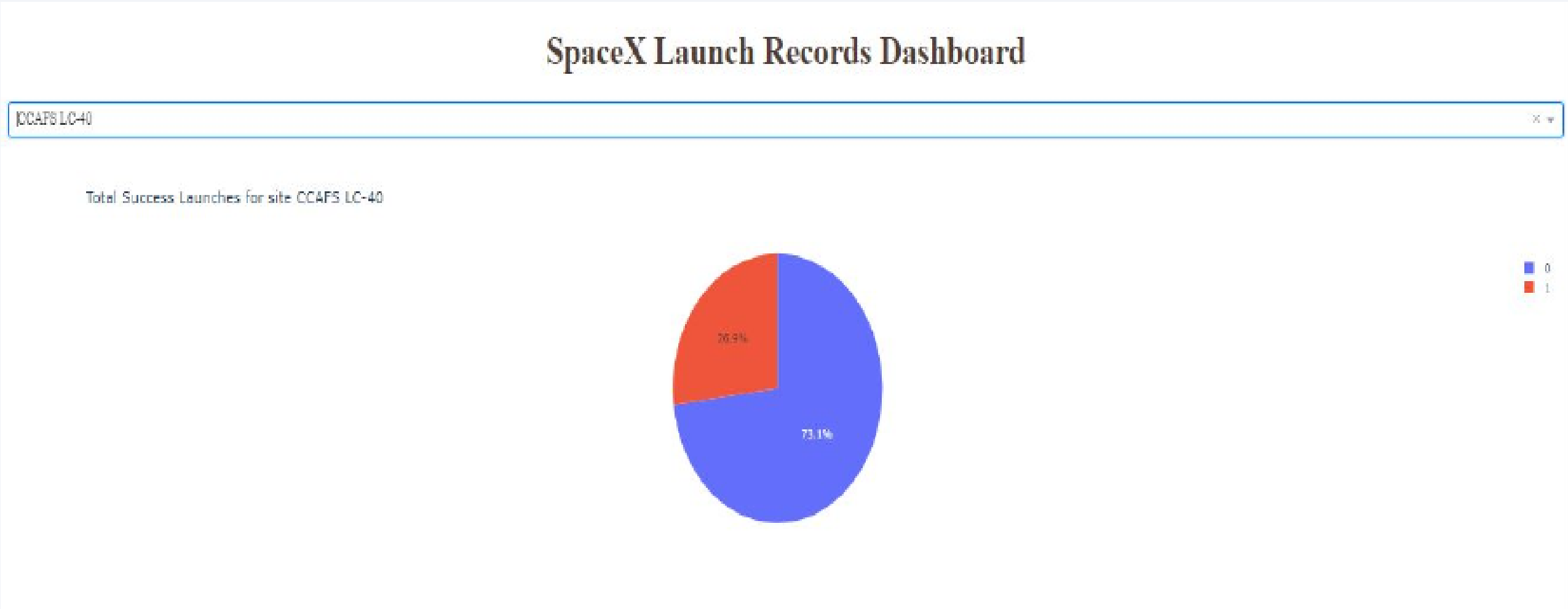

Section 4
Build a Dashboard with Plotly Dash

Pie-Chart for launch success count for all sites



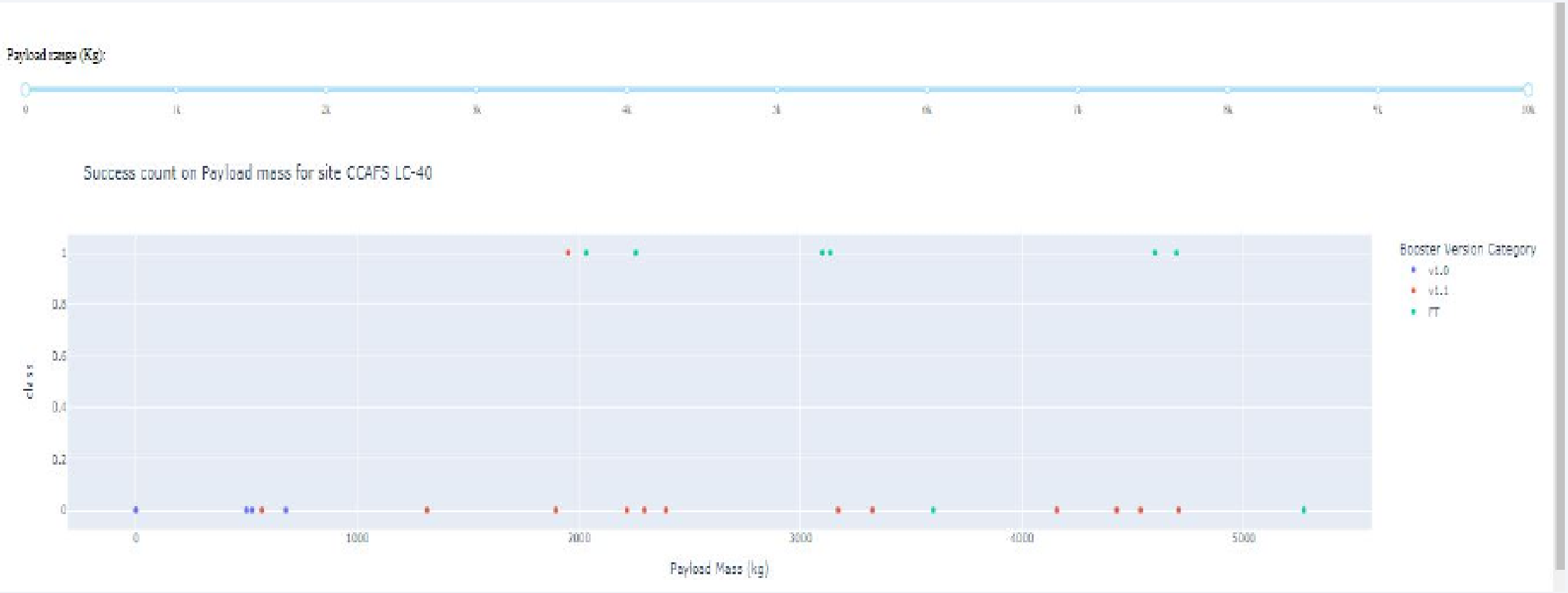
- Launch site KSC LC-39A has the highest launch success rate at 42% followed by CCAFS LC-40 at 29%, VAFB SLC-4E at 17% and lastly launch site CCAFS SLC-40 with a success rate of 13%

Pie chart for the launch site with 2nd highest launch success ratio



- Launch site CCAFS LC-40 had the 2nd highest success ratio of 73% success against 27% failed launches

Payload vs. Launch Outcome scatter plot for all sites

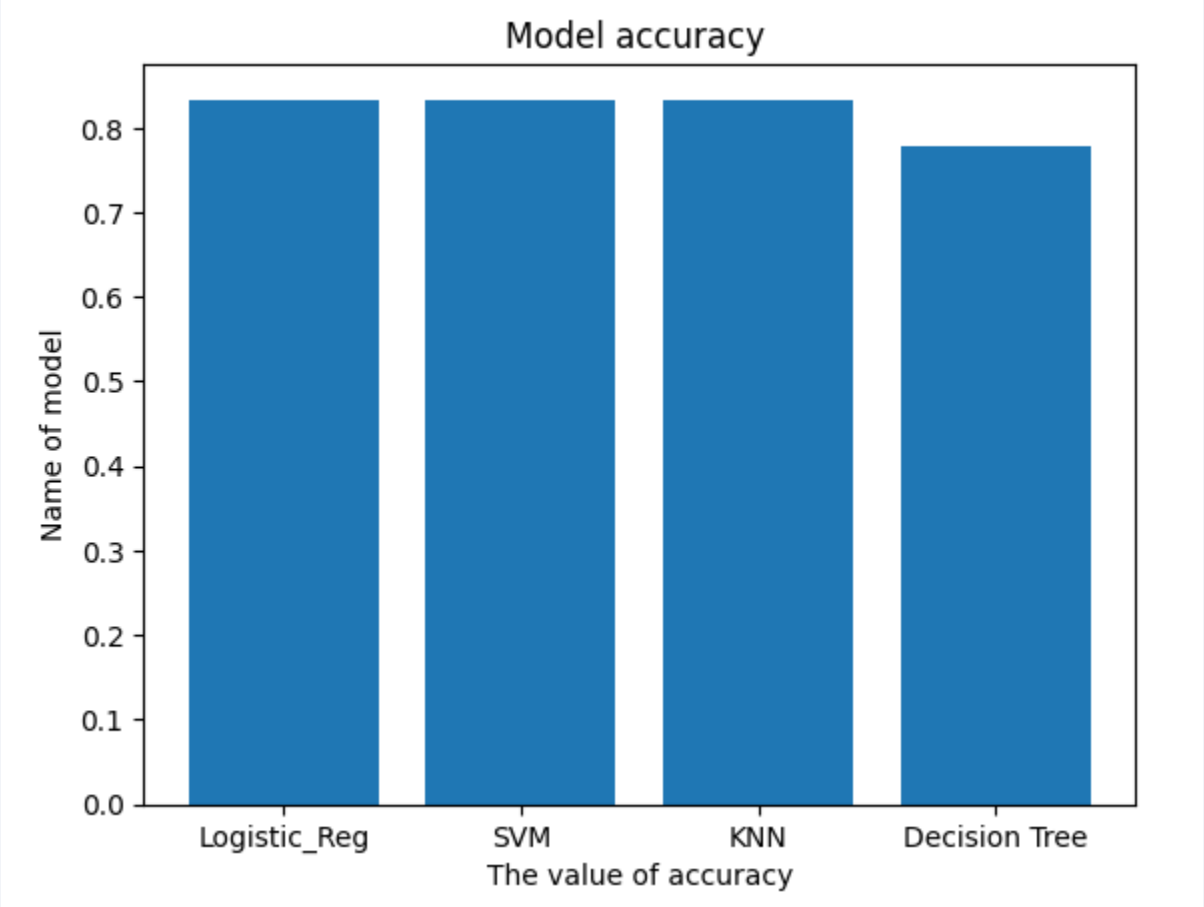


- For Launch site CCAFS LC-40 the booster version FT has the largest success rate from a payload mass of >2000kg

Section 5

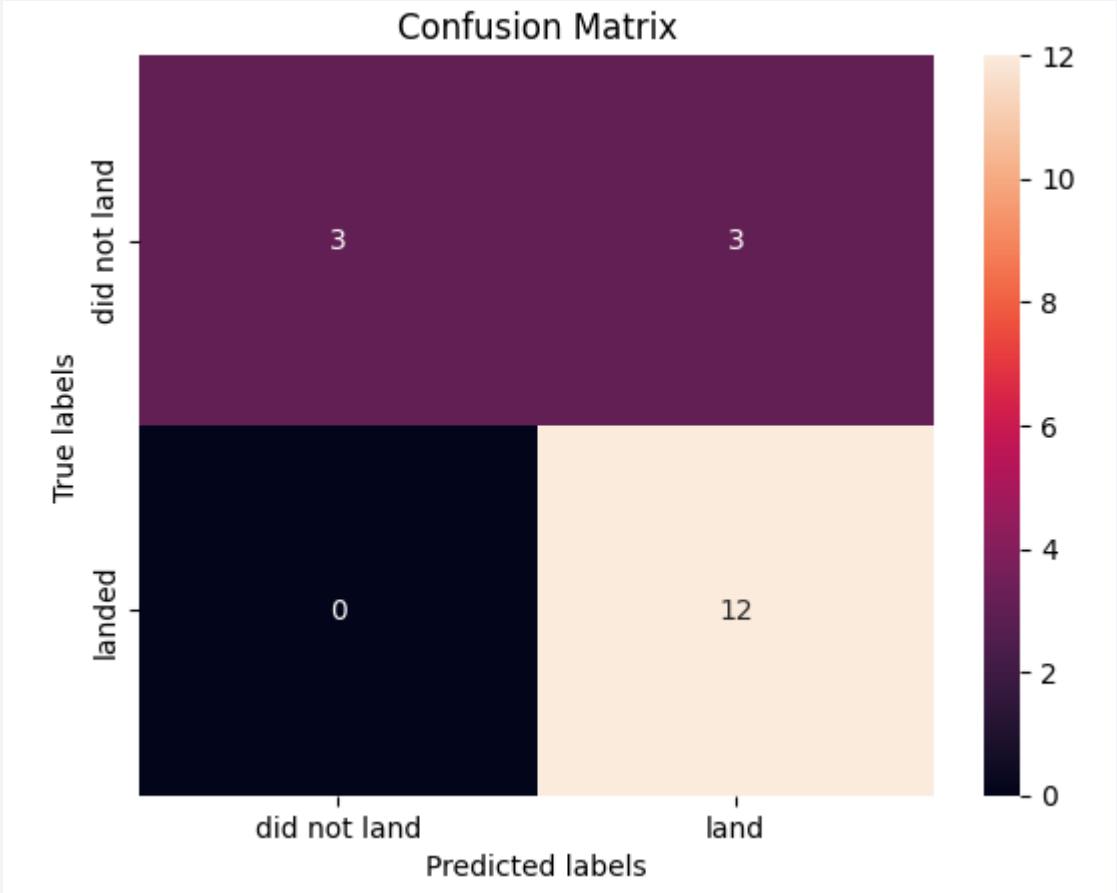
Predictive Analysis (Classification)

Classification Accuracy



- Three of the model has the same of accuracy 0.833333. But,

Confusion Matrix



Three of the classification model had the same confusion matrixes and were able equally distinguish between the different classes. The major problem is false positives for all the models.

Conclusions

- Different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
- We can deduce that, as the flight number increases in each of the 3 launch sites, so does the success rate. For instance, the success rate for the VAFB SLC 4E launch site is 100% after the Flight number 50. Both KSC LC 39A and CCAFS SLC 40 have a 100% success rates after 80th flight
- If you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).
- Orbits ES-L1, GEO, HEO & SSO have the highest success rates at 100%, with SO orbit having the lowest success rate at ~50%. Orbit SO has 0% success rate.
- LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit

- With heavy payloads the successful landing or positive landing rate are more for Polar LEO and ISS. However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here
- Finally the success rate since 2013 kept increasing till 2020.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!