IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Eva Ananda
27 January, 2023

# Outline

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

# Executive Summary

- **Methodologies**
  - Data collection using SpaceX API and Wikipedia web scraping & data wrangling in python
  - Exploratory Data Analysis (EDA) using python and SQL (IBM db2)
  - Visual analysis with python matplotlib, folium, and dash
  - Machine learning prediction
- **Summary**
  - We can use various source to obtain data.
  - Visualization helps us to understand about the information better.
  - Good prediction method could help us to predict the outcome in the future.

# Introduction

- Background: SpaceY, a new rocket launch company would like to provide affordable space travel for everyone. SpaceY has consider to make SpaceX as a benchmark since they're proven to be able to make much more efficient rocket launch compared to other company.

- Problem: we want to investigate what are the factors that affecting SpaceX's efficiency using EDA, visualization and machine learning prediction.

Section 1

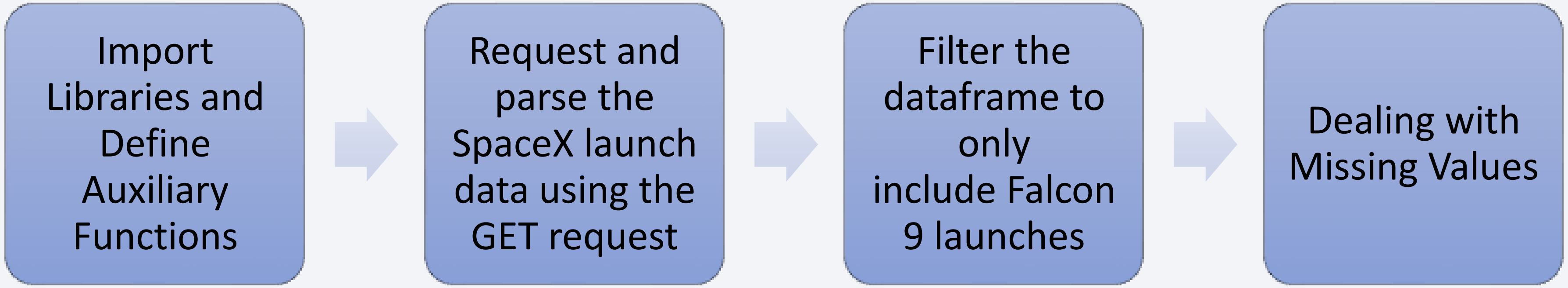
# Methodology

# Methodology

## Executive Summary

- Data collection methodology:
  - Data is collected using SpaceX API (https://api.spacexdata.com/v4/rockets/) and Wikipedia web scraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- Perform data wrangling
  - Checking missing values and data type, processing the data, adding important columns
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistic regression, decision tree, k nearest neighbor, and support vector machine

# 1. Data Collection - SpaceX API



Import Libraries and Define Auxiliary Functions

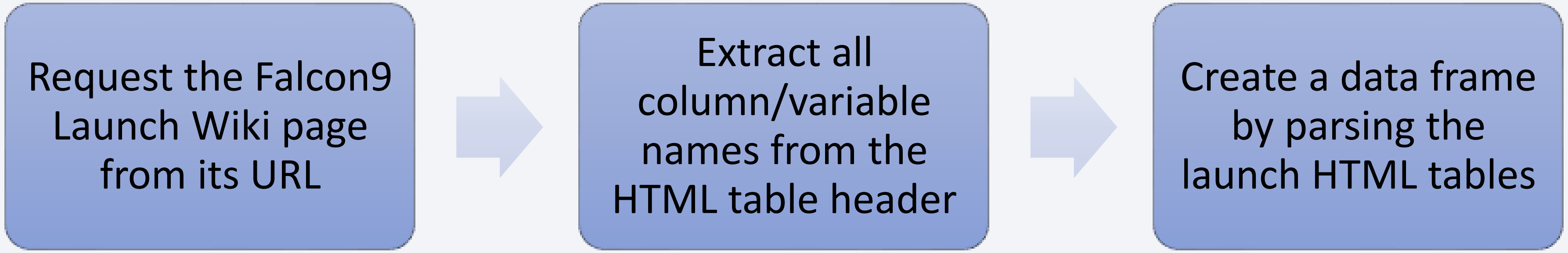
Request and parse the SpaceX launch data using the GET request

Filter the dataframe to only include Falcon 9 launches

Dealing with Missing Values

Click here for GitHub URL

# 2. Data Collection - WebScraping



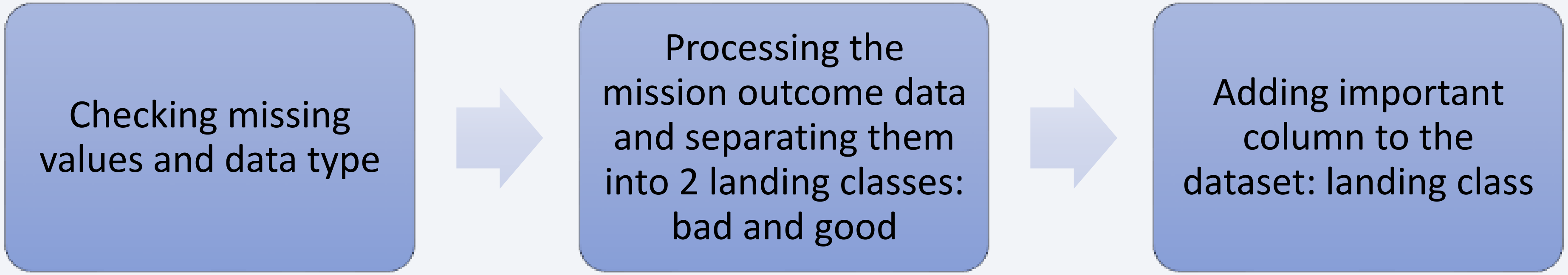
Request the Falcon9 Launch Wiki page from its URL

Extract all column/variable names from the HTML table header

Create a data frame by parsing the launch HTML tables

Click here for GitHub URL

# 3. Data Wrangling



Checking missing values and data type

Processing the mission outcome data and separating them into 2 landing classes: bad and good

Adding important column to the dataset: landing class

Click here for GitHub URL

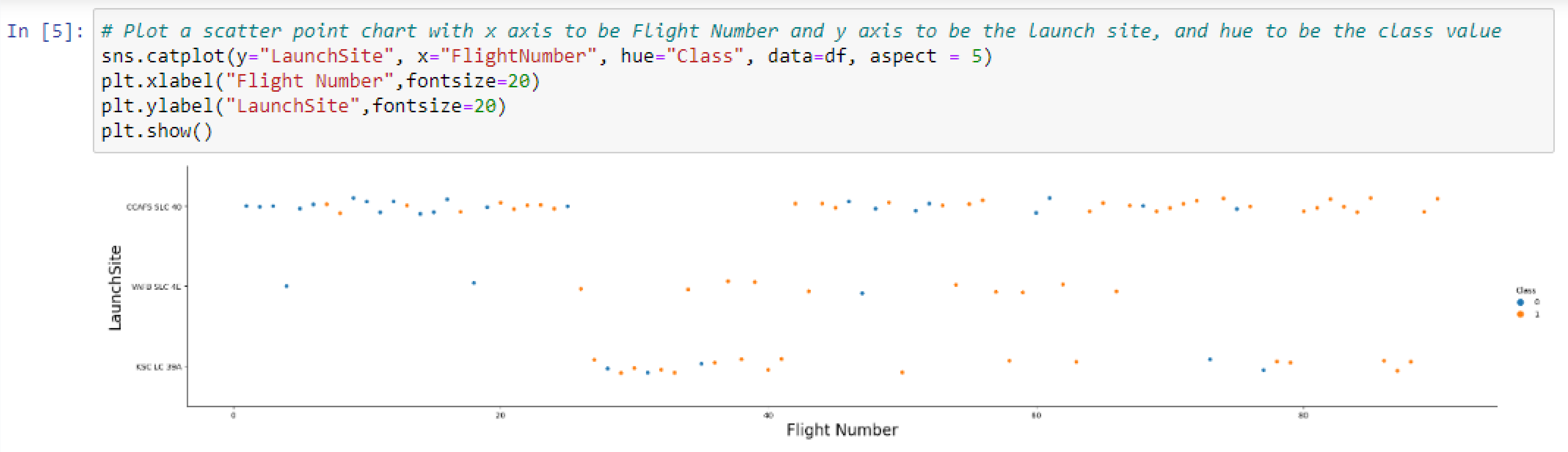
# 4. EDA with SQL

- Display the names of the unique launch sites in the space mission.
- Display 5 records where launch sites begin with the string 'CCA‘
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass
- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

Click here for GitHub URL

# 5. EDA with Data Visualization

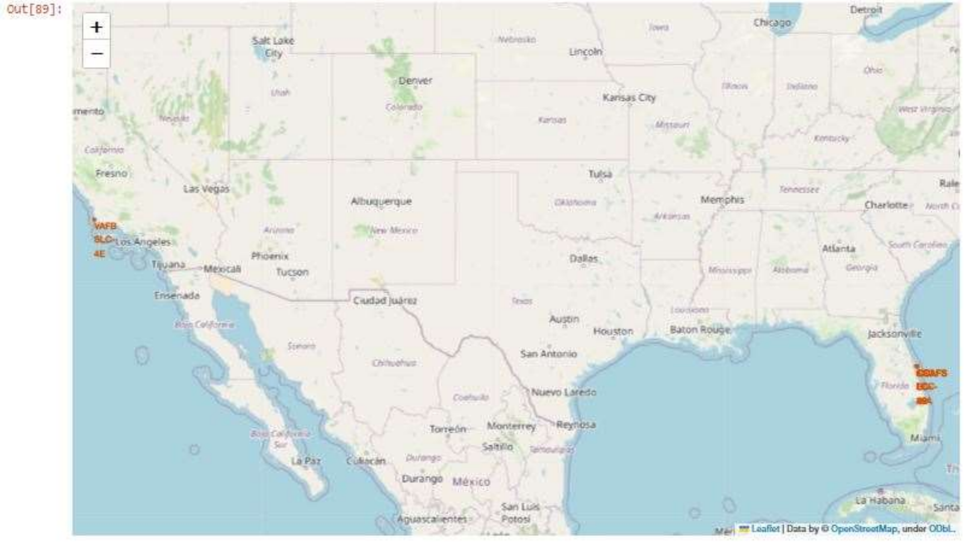
For data visualization, **scatterplot and barplot** are used to show the relationship between : **flight number & launch site, payload & launch site, success rate of each orbit type, flight number & orbit type, payload & orbit type**. Also, we use the barplot to show **yearly trend** of the launch success.

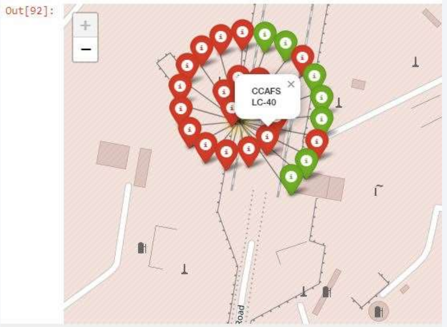


```
In [5]: # Plot a scatter point chart with x axis to be Flight Number and y axis to be the launch site, and hue to be the class value
        sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)
        plt.xlabel("Flight Number",fontsize=20)
        plt.ylabel("LaunchSite",fontsize=20)
        plt.show()
```

Click here for GitHub URL

# 6. Build an Interactive Map with Folium





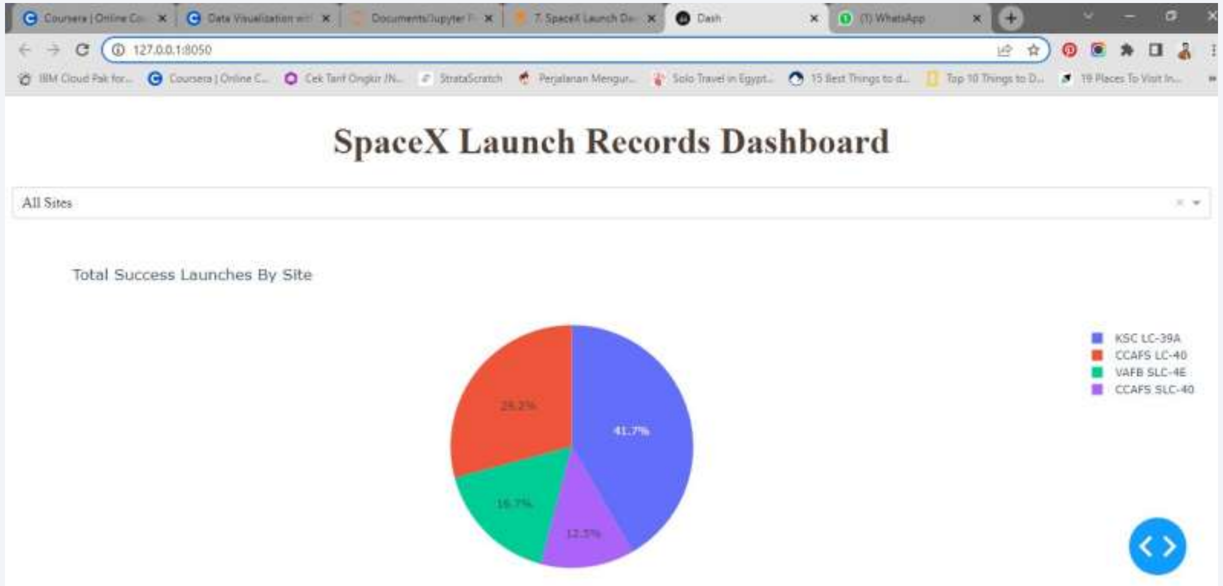
+
−
Space Launch
Complex
4

On this section, we draw **circle** for each launch site, measuring distance between launch site and other object with **polyline**, and differentiating **marker colors** for the successful and failed launch.

We use folium to get the general idea of how launch sites are supposed to be located, in this case, we found that it is better to build rocket launch sites near the coastline and far from residential area.

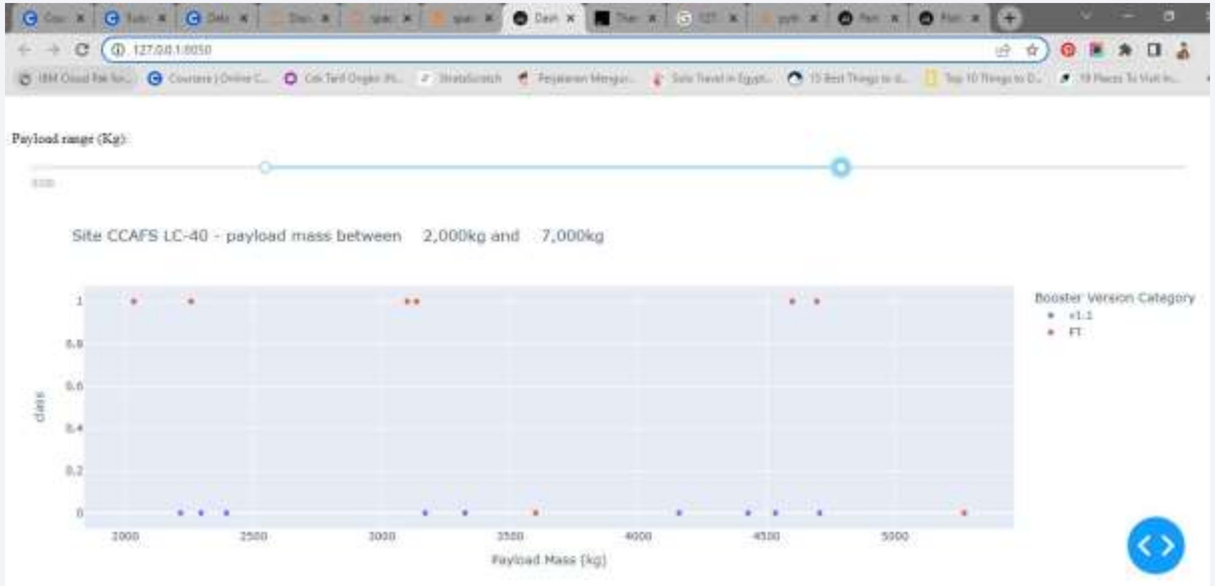
[Click here for GitHub URL]

# 7. Build a Dashboard with Plotly Dash



SpaceX Launch Records Dashboard

All Sites



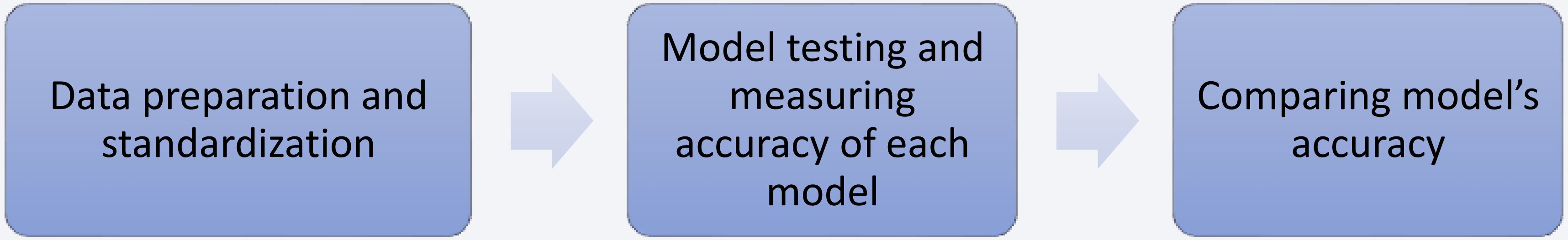
Payload range (Kg):

Site CCAFS LC-40 - payload mass between 2,000kg and 7,000kg

Interactive dashboard is used to show percentage of launch success rate per launch site and also the relationship between payload mass and booster version.

Click here for GitHub URL

# 8. Predictive Analysis (Classification)


Data preparation and standardization

Model testing and measuring accuracy of each model

Comparing model's accuracy

Machine learning methods used are logistic regression, support vector machine (SVM), decision tree, and k nearest neighboor (KNN).

Click here for GitHub URL

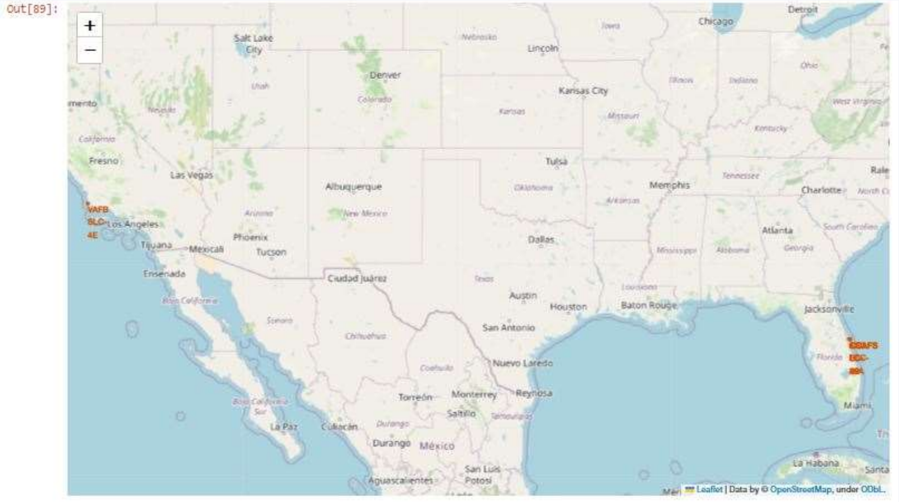
# Results

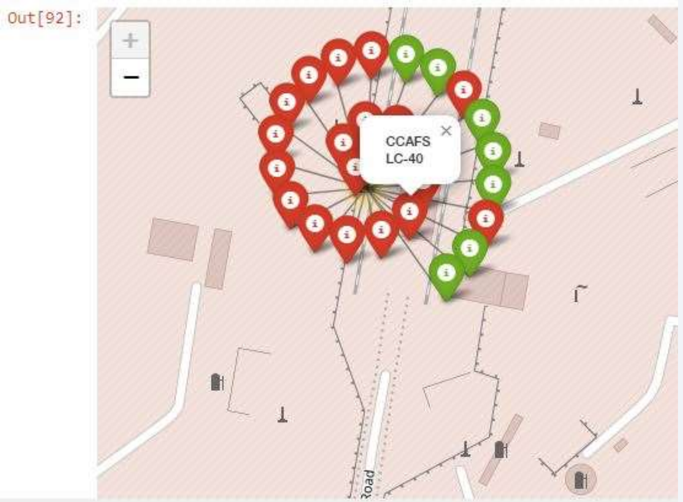
Exploratory data analysis results:

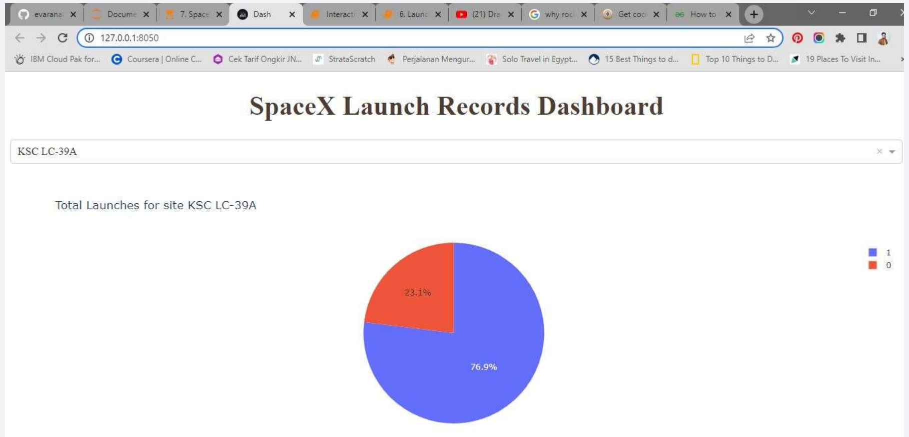
- SpaceX has 4 launch sites: CCAFS LC-40, CCAFS SLC-40, KSC LC-39A and VAFB SLC-4E.
- Almost 100% mission outcome is succesful, from the data, we can see that they have only 1 failure.
- The first successful launch is in 22 December 2015.
- Average payload for F9 v1.1 booster is 2.928 kg.
- Total payload from booster launched by NASA is 111.268 kg.
- Important factors that affect the successful rate are: launch site, orbit, and payload mass.

# Results

- Interactive analytics demo in screenshots





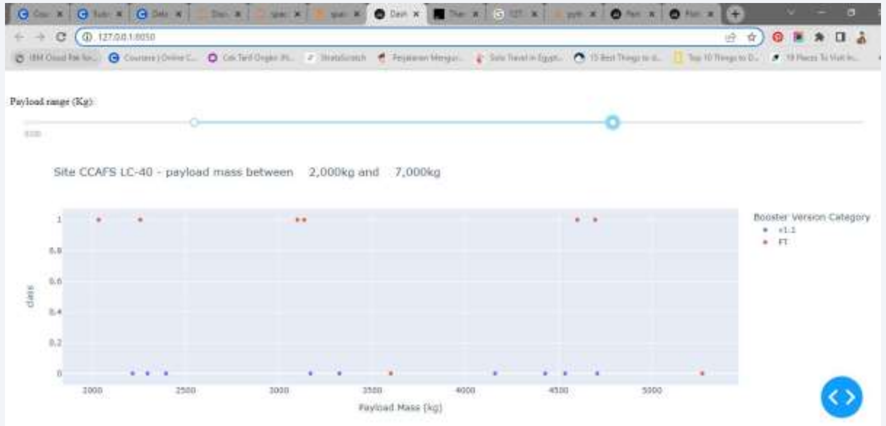


127.0.0.1:8050

SpaceX Launch Records Dashboard

KSC LC-39A

Total Launches for site KSC LC-39A



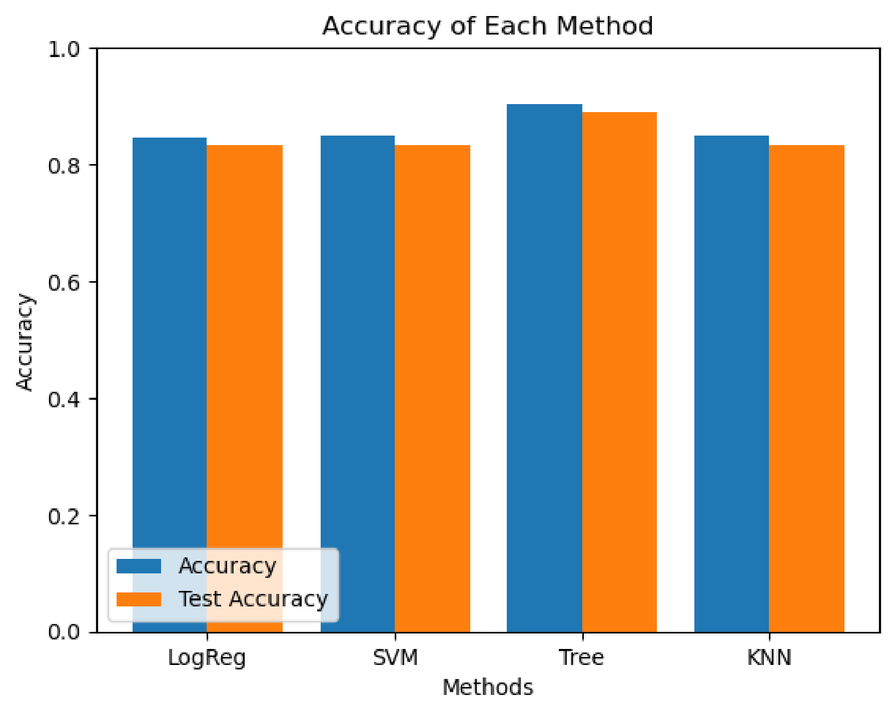
127.0.0.1:8050

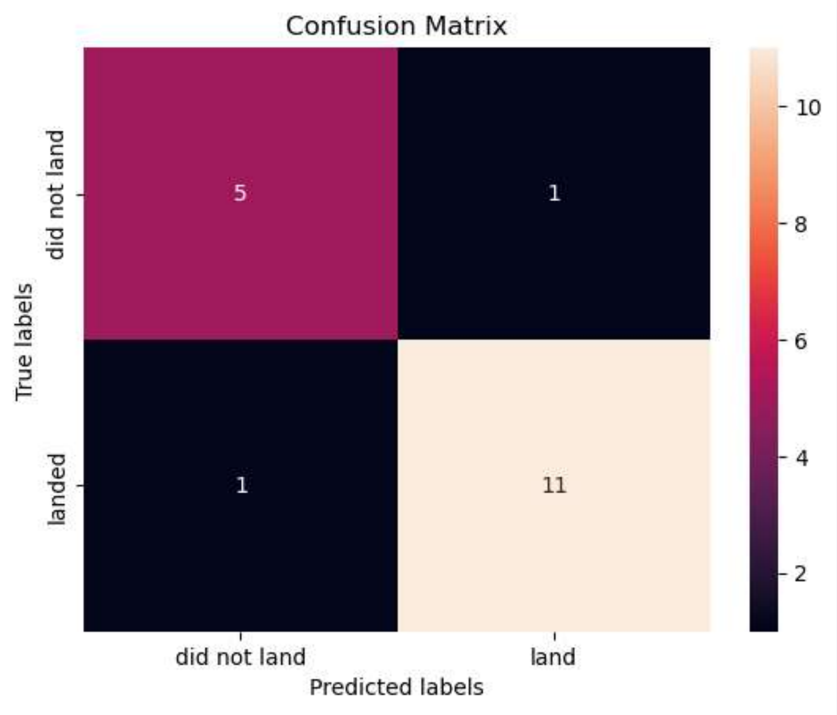
Payload range (Kg):

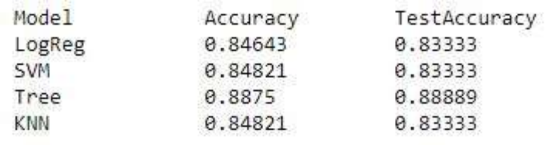
Site CCAFS LC-40 - payload mass between 2,000kg and 7,000kg

# Results

- Predictive analysis results







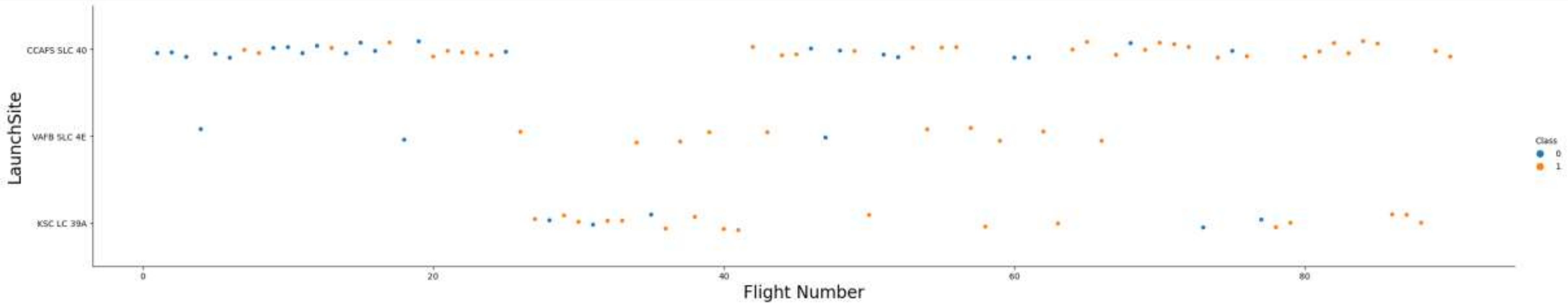
```
Model           Accuracy        TestAccuracy
LogReg          0.84643         0.83333
SVM             0.84821         0.83333
Tree            0.8875          0.88889
KNN             0.84821         0.83333
```

The most accurate prediction method is decision tree, we can see from its confusion matrix that it has high true positive and true negative.

Section 2

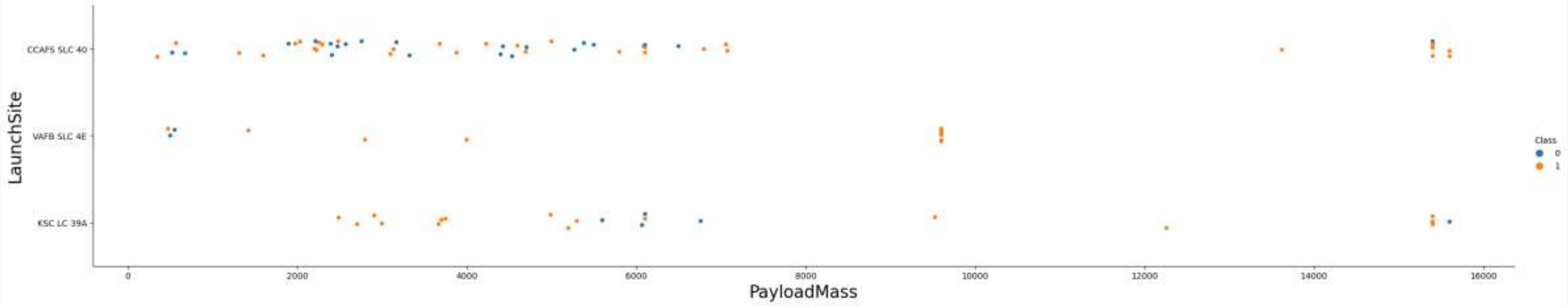
# Insights drawn from EDA

# Flight Number vs. Launch Site



- Most launch is taking place in CCAFS SLC-40, probably because the launch site has better facilities or anything related to the needs.

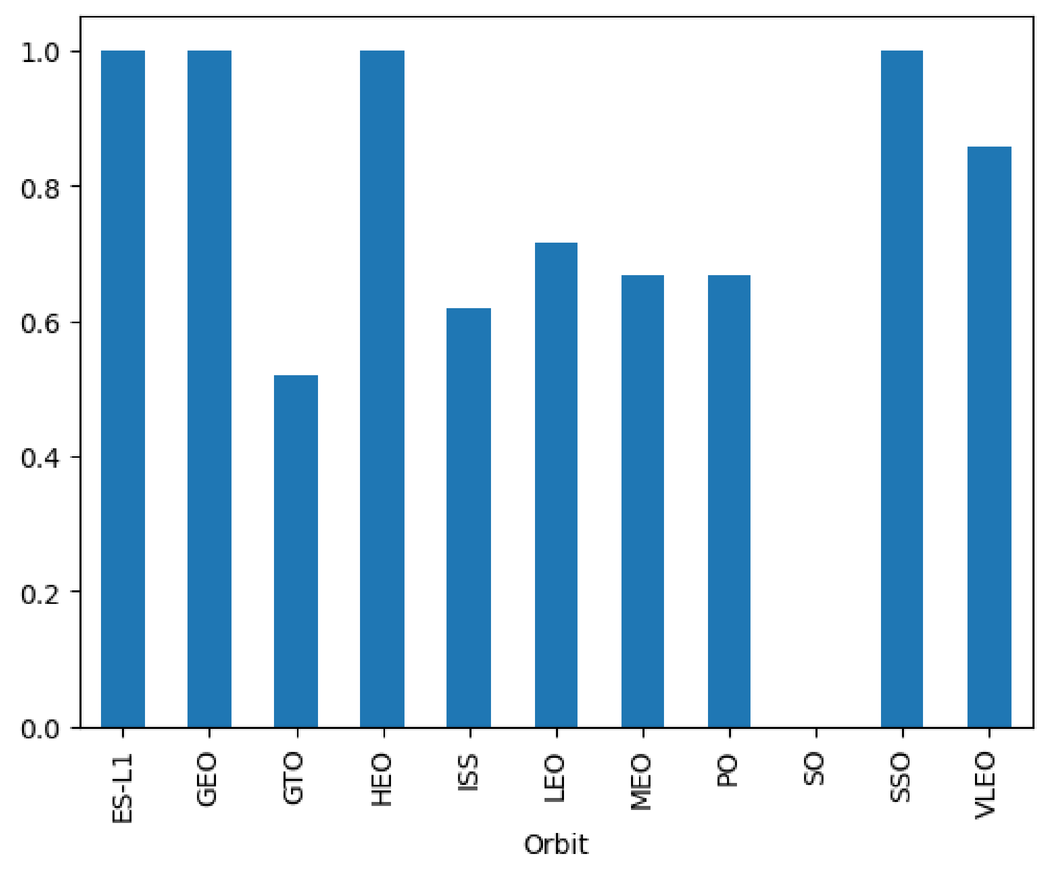
# Payload vs. Launch Site



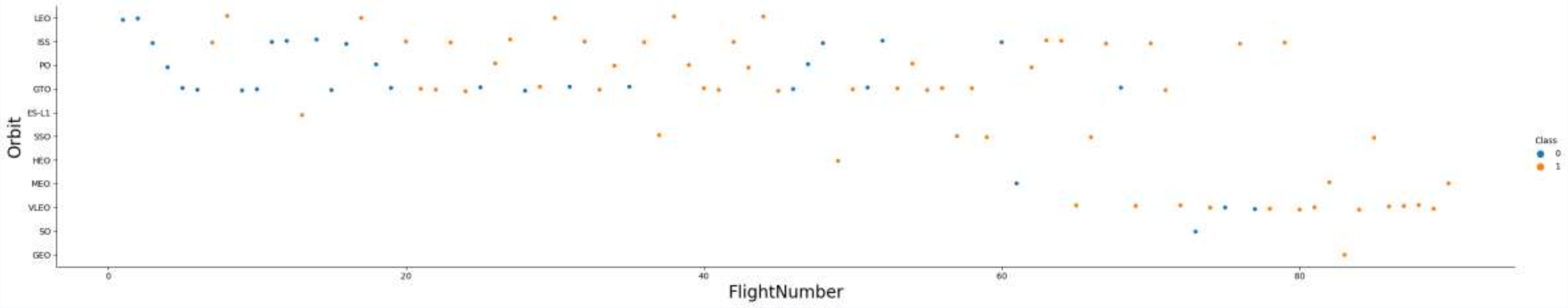
- Success rate for payload mass below 8.000kg and between 15.000kg and 16.000kg happened in these launch sites : CCAFS SLC-40 and KSC LC-39A so most probably those 2 launch sites are suitable for those payload mass.
- For launch site VAFB SLC-4E, payload mass close to 10.000 kg have high success rate.

# Success Rate vs. Orbit Type

- Orbit ES-L1, GEO, HEO, and SSO have high success rate
- The lowest success rate is from GTO

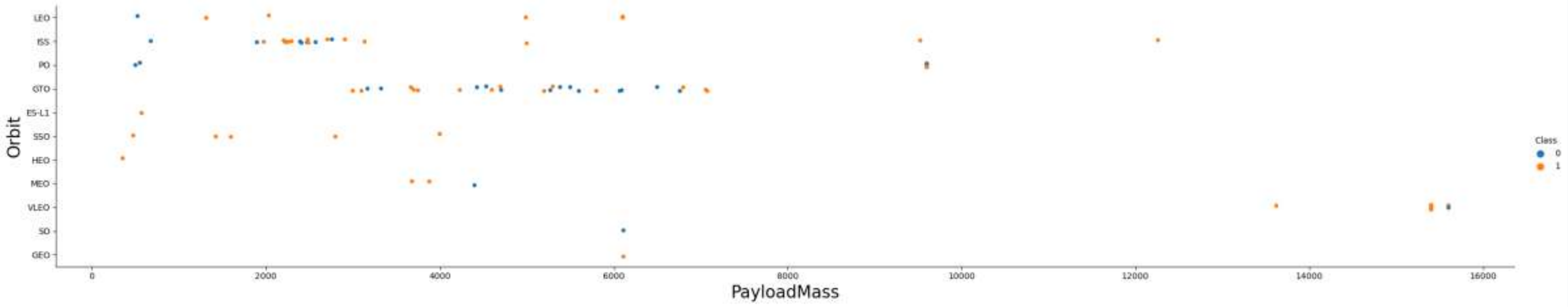


# Flight Number vs. Orbit Type



- Orbit SSO and VLEO have high success rate.
- Most flight happened in ISS, GTO, and VLEO, however, the failed rate in ISS and GTO are higher than others.

# Payload vs. Orbit Type



- Payload mass below 8.000 kg and GTO orbit are a good combination.
- VLEO has the highest payload among all.

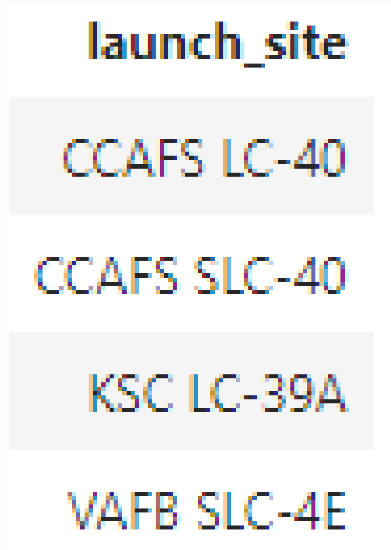
# Launch Success Yearly Trend

The chart doesn’t appear in my python notebook, however, it was written in the notebook that the trend is getting higher. It means that SpaceX is committed to improve the technologies and maximizing all those factors that affecting rocket launch efficiency.

# All Launch Site Names

```
SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL
```

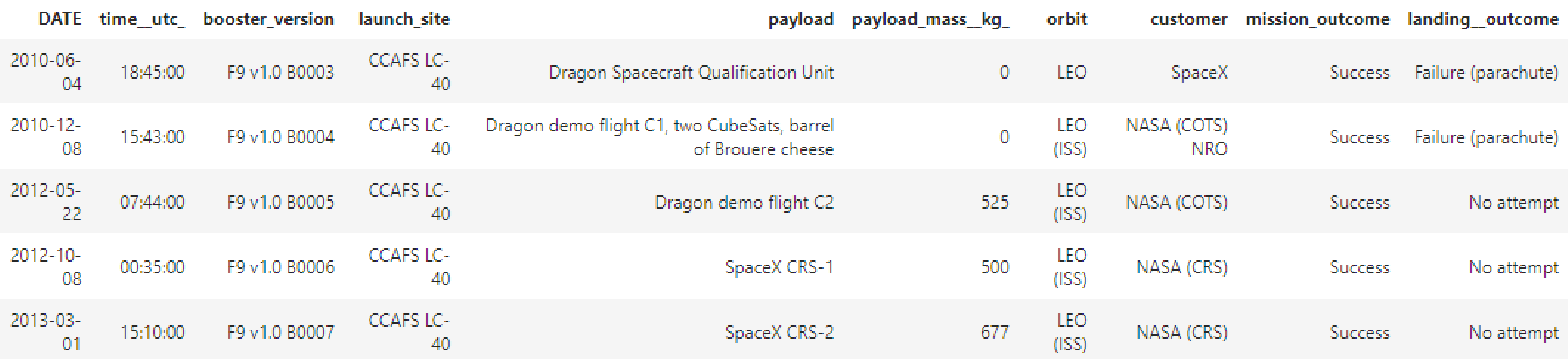
Launch site names are CCAFS LC-40, CCAFS SLC-40, KSC LC-39A and VAFB SLC-4E



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

# Launch Site Names Begin with 'CCA'

SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5

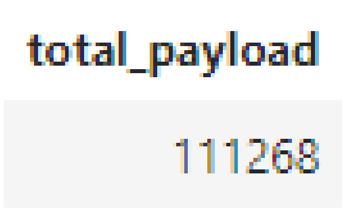


| DATE | time__utc_ | booster_version | launch_site | payload | payload_mass__kg_ | orbit | customer | mission_outcome | landing__outcome |
|---|---|---|---|---|---|---|---|---|---|
| 2010-06-04 | 18:45:00 | F9 v1.0 B0003 | CCAFS LC-40 | Dragon Spacecraft Qualification Unit | 0 | LEO | SpaceX | Success | 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 | 07:44:00 | F9 v1.0 B0005 | CCAFS LC-40 | Dragon demo flight C2 | 525 | LEO (ISS) | NASA (COTS) | Success | No attempt |
| 2012-10-08 | 00:35:00 | 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

```
SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM
SPACEXTBL WHERE PAYLOAD  LIKE '%CRS%‘
```

Total payload mass carried by boosters launched by NASA (CRS) is 111.268 kg

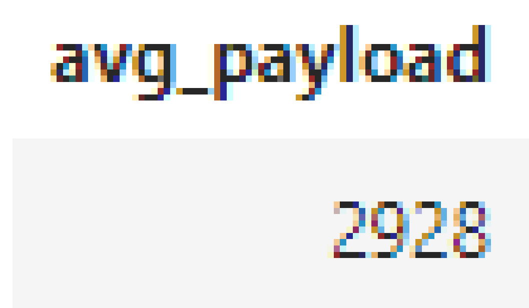


| total_payload |
|---|
| 111268 |

# Average Payload Mass by F9 v1.1

```
SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM
SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1'
```

Average payload mass carried by booster version F9 v1.1 is 2.928kg

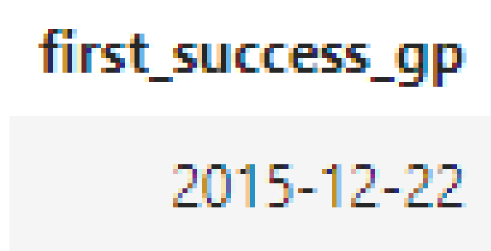


| avg_payload |
|---|
| 2928 |

# First Successful Ground Landing Date

```
SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Success (ground pad)‘
```

First successful ground landing date is 22 December 2015

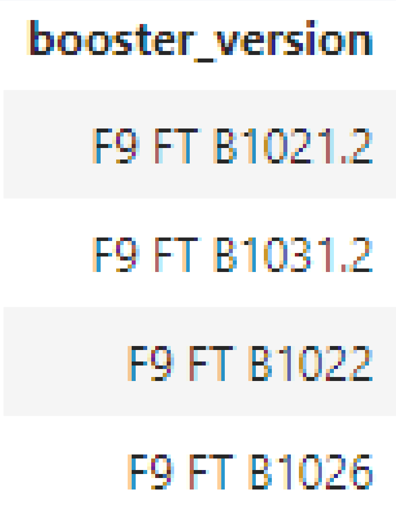


| first_success_gp |
|---|
| 2015-12-22 |

## Successful Drone Ship Landing with Payload between 4000 and 6000

```
SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND
LANDING__OUTCOME = 'Success (drone ship)'
```

List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are F9 FT B1021.2, F9 FT B1031.2, F9 FT B1022, and F9 FT B1026

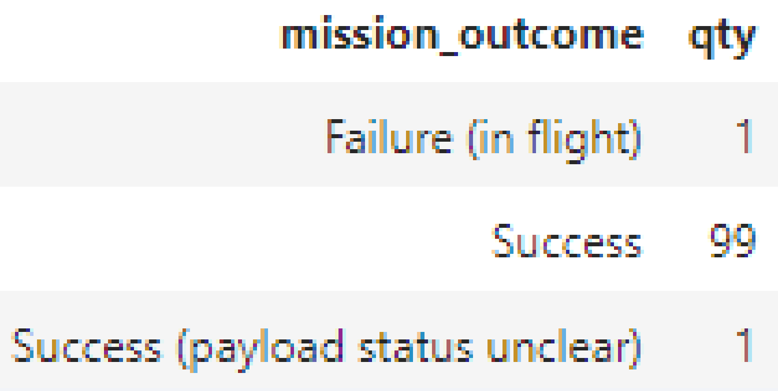


| booster_version |
|---|
| F9 FT B1021.2 |
| F9 FT B1031.2 |
| F9 FT B1022 |
| F9 FT B1026 |

# Total Number of Successful and Failure Mission Outcomes

SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME

Mission outcome success is 99 times, failure (in flight) is 1 time, and success but payload status unclear is 1 time

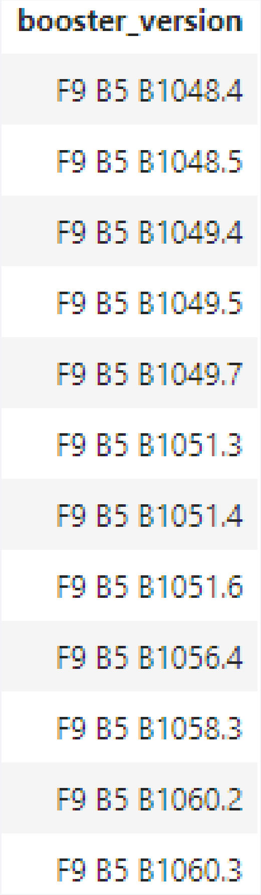


| mission_outcome | qty |
|---:|---:|
| Failure (in flight) | 1 |
| Success | 99 |
| Success (payload status unclear) | 1 |

# Boosters Carried Maximum Payload

SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL) ORDER BY BOOSTER_VERSION

Here are the list of the booster_versions which have carried the maximum payload mass

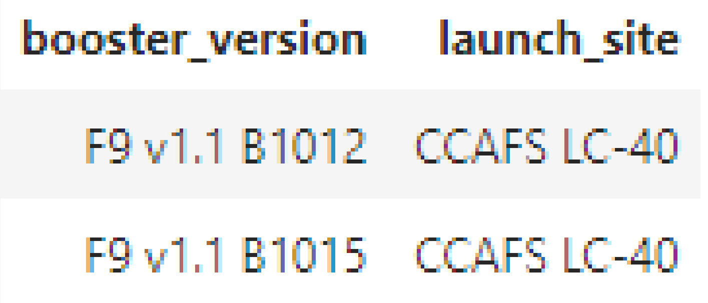


| booster_version |
|---|
| F9 B5 B1048.4 |
| F9 B5 B1048.5 |
| F9 B5 B1049.4 |
| F9 B5 B1049.5 |
| F9 B5 B1049.7 |
| F9 B5 B1051.3 |
| F9 B5 B1051.4 |
| F9 B5 B1051.6 |
| F9 B5 B1056.4 |
| F9 B5 B1058.3 |
| F9 B5 B1060.2 |
| F9 B5 B1060.3 |

# 2015 Launch Records

```
SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND
DATE_PART('YEAR', DATE) = 2015
```

Here is the list of failed landing_outcomes in drone ship with their booster versions, and launch site names for in year 2015

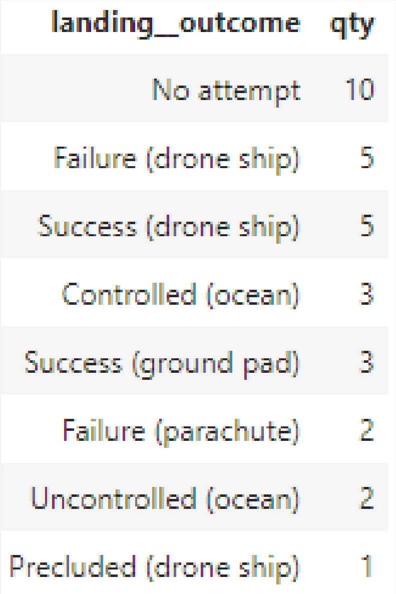


| booster_version | launch_site |
|---|---|
| F9 v1.1 B1012 | CCAFS LC-40 |
| F9 v1.1 B1015 | CCAFS LC-40 |

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

```
SELECT LANDING__OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING__OUTCOME ORDER BY QTY DESC
```

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

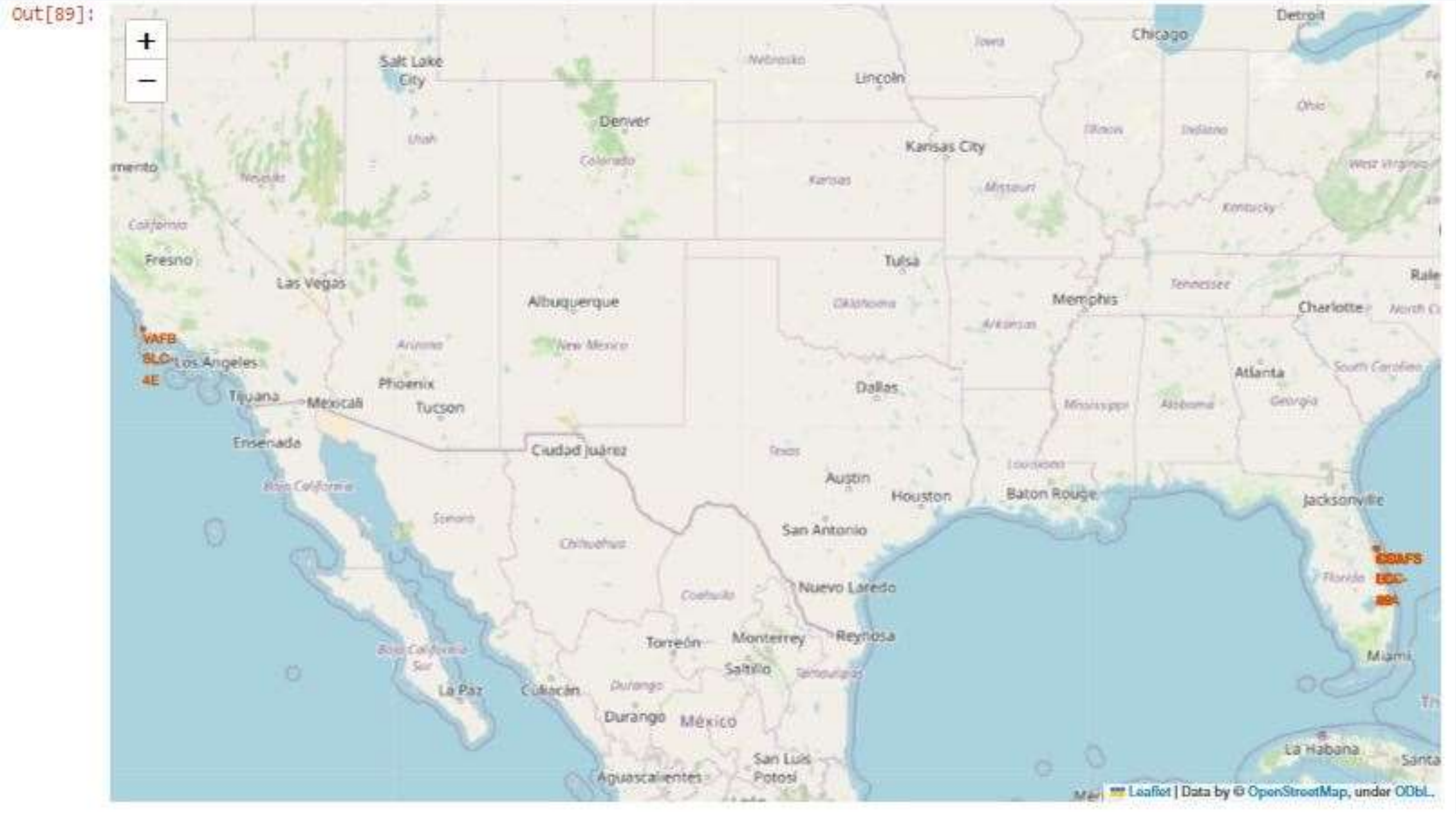


| landing__outcome | qty |
|---|---|
| No attempt | 10 |
| Failure (drone ship) | 5 |
| Success (drone ship) | 5 |
| Controlled (ocean) | 3 |
| Success (ground pad) | 3 |
| Failure (parachute) | 2 |
| Uncontrolled (ocean) | 2 |
| Precluded (drone ship) | 1 |

Section 3

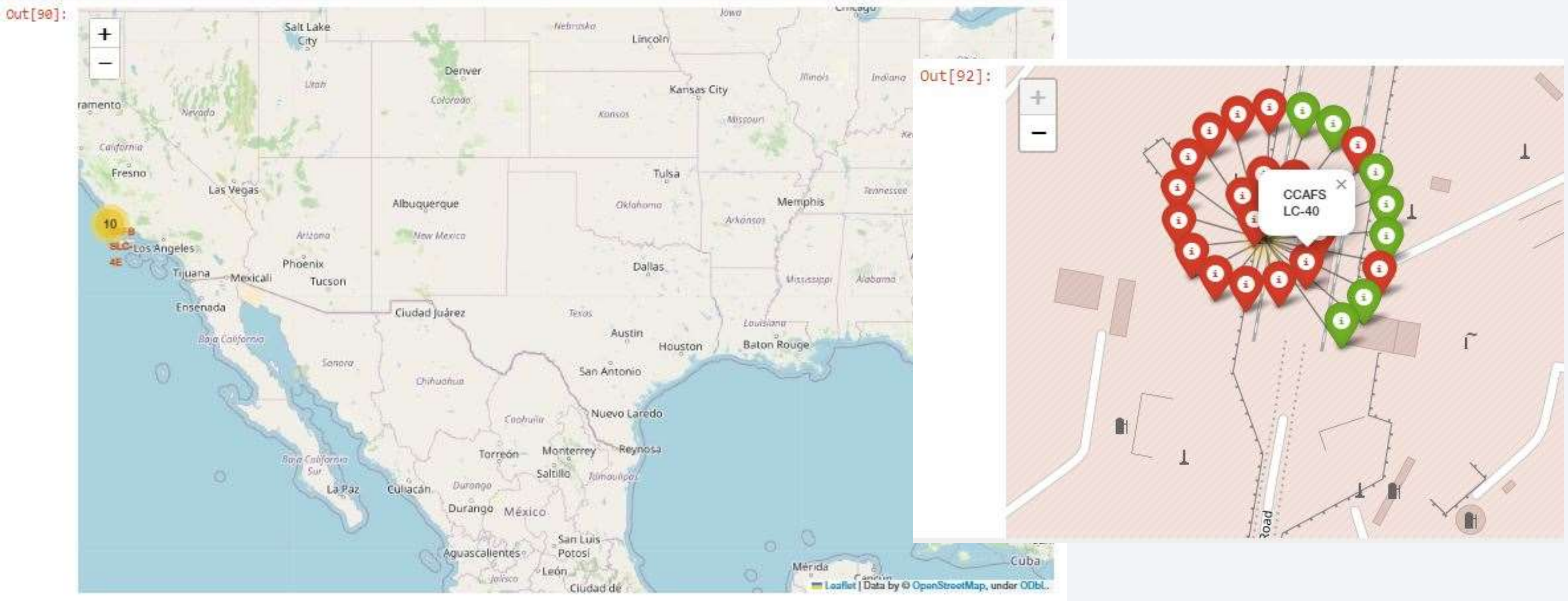
# Launch Sites Proximities Analysis

# All Launch Sites with Markers



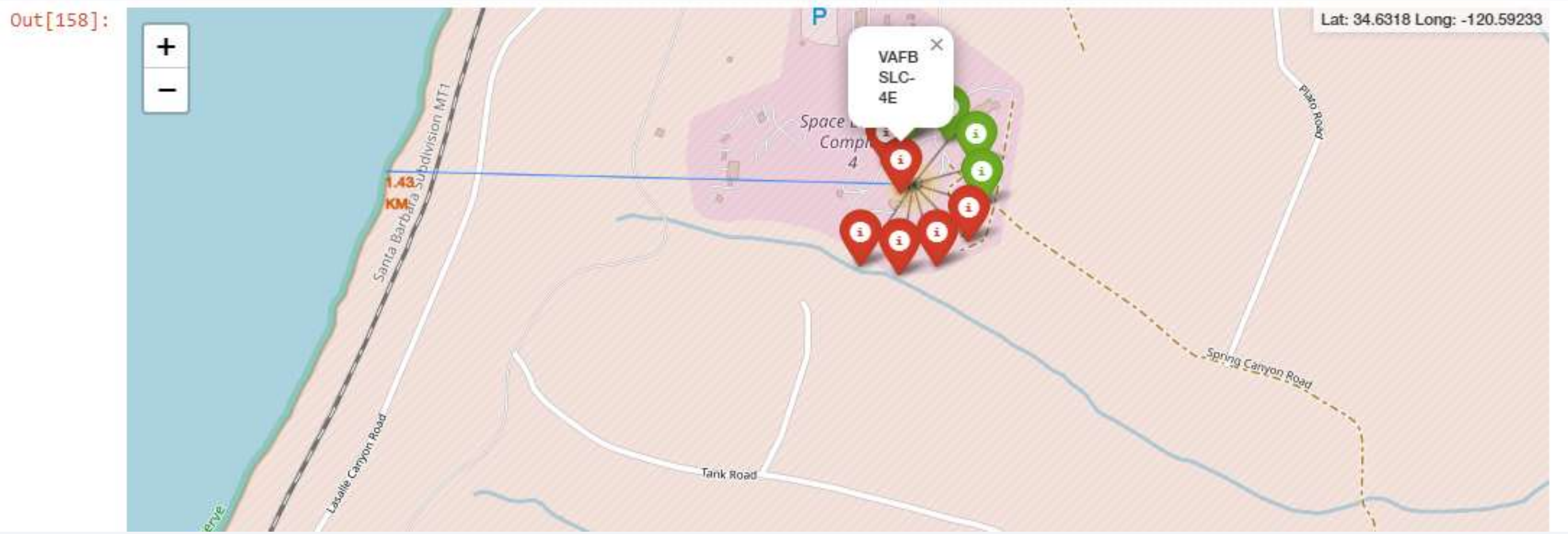
From the map above we can see that all launch sites are located near the ocean to reduce the risk of rocket crashes. If they are located in the city center, the likelihood of harming the outer parties are quite high.

# Launch Sites with Color Label

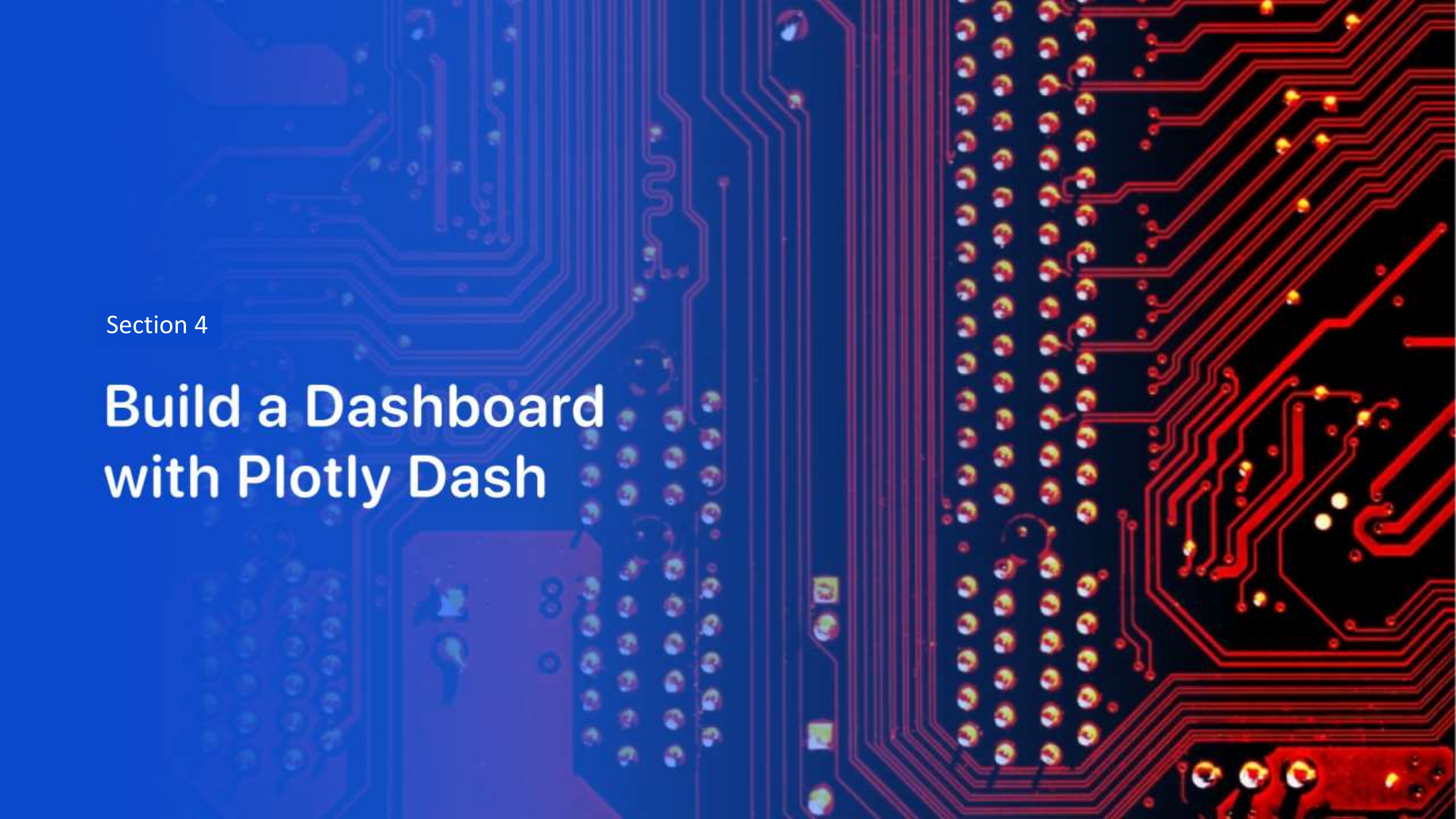


Red marker indicating launch failures and green marker is indicating the successful one

# Distance from Launch Sites



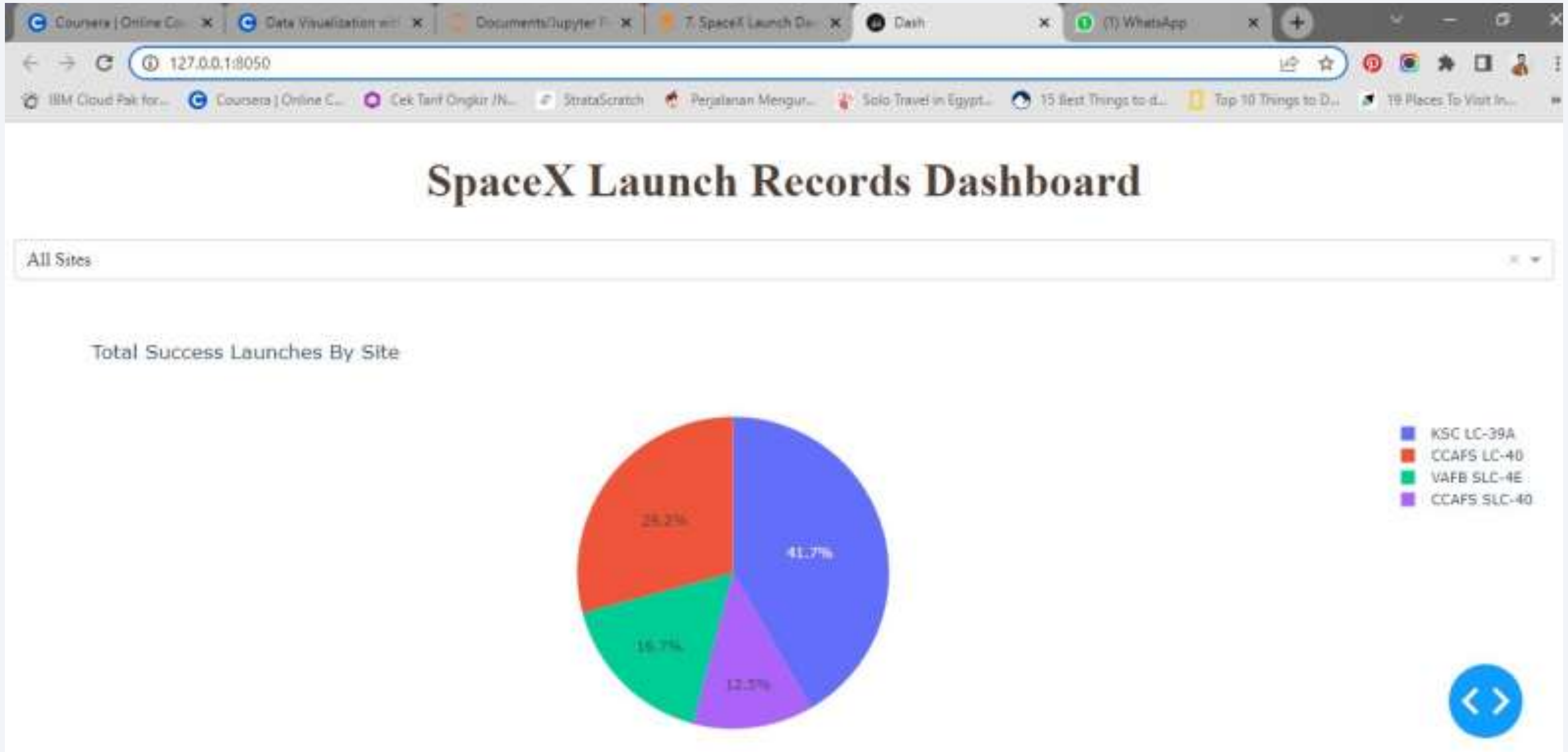
The distance from launch sites (we take VAFB SLC-4E as an example) is quite close from the coastline and we also can see from the map that launch sites are far from airport, highway, and residential to prevent the danger. However, they are located near the railway, probably to make logistic distribution easier.



Section 4

# Build a Dashboard with Plotly Dash

# Total Success Launches by Sites



SpaceX Launch Records Dashboard

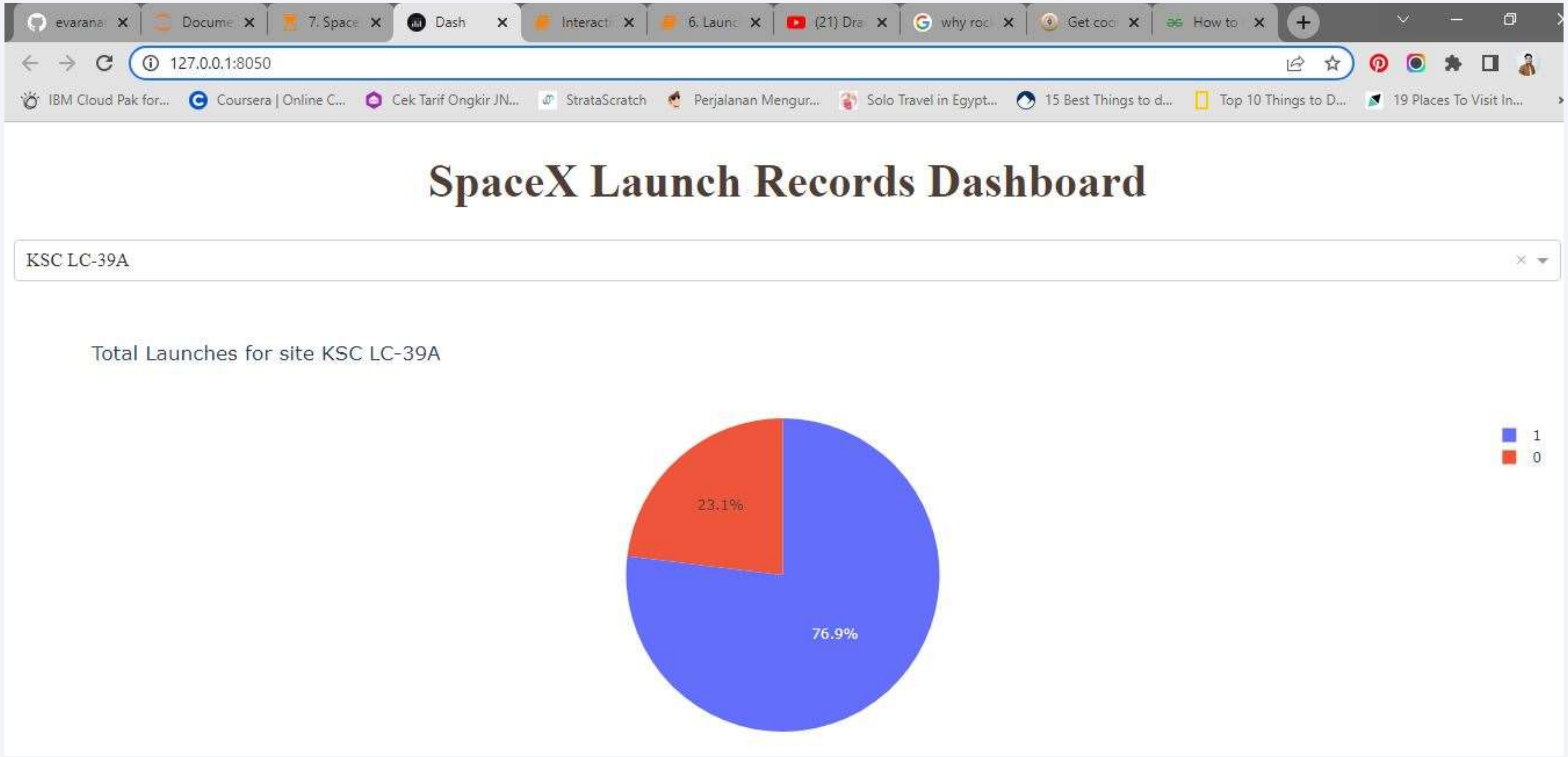
All Sites

Total Success Launches By Site

From the dashboard we can conclude that the highest success rate is from KSC LC-39A which occupy almost half of the chart. Second position is CCAFS LC-40 and then VAFB SLC-4E has almost the same rate with CCAFS SLC-40.

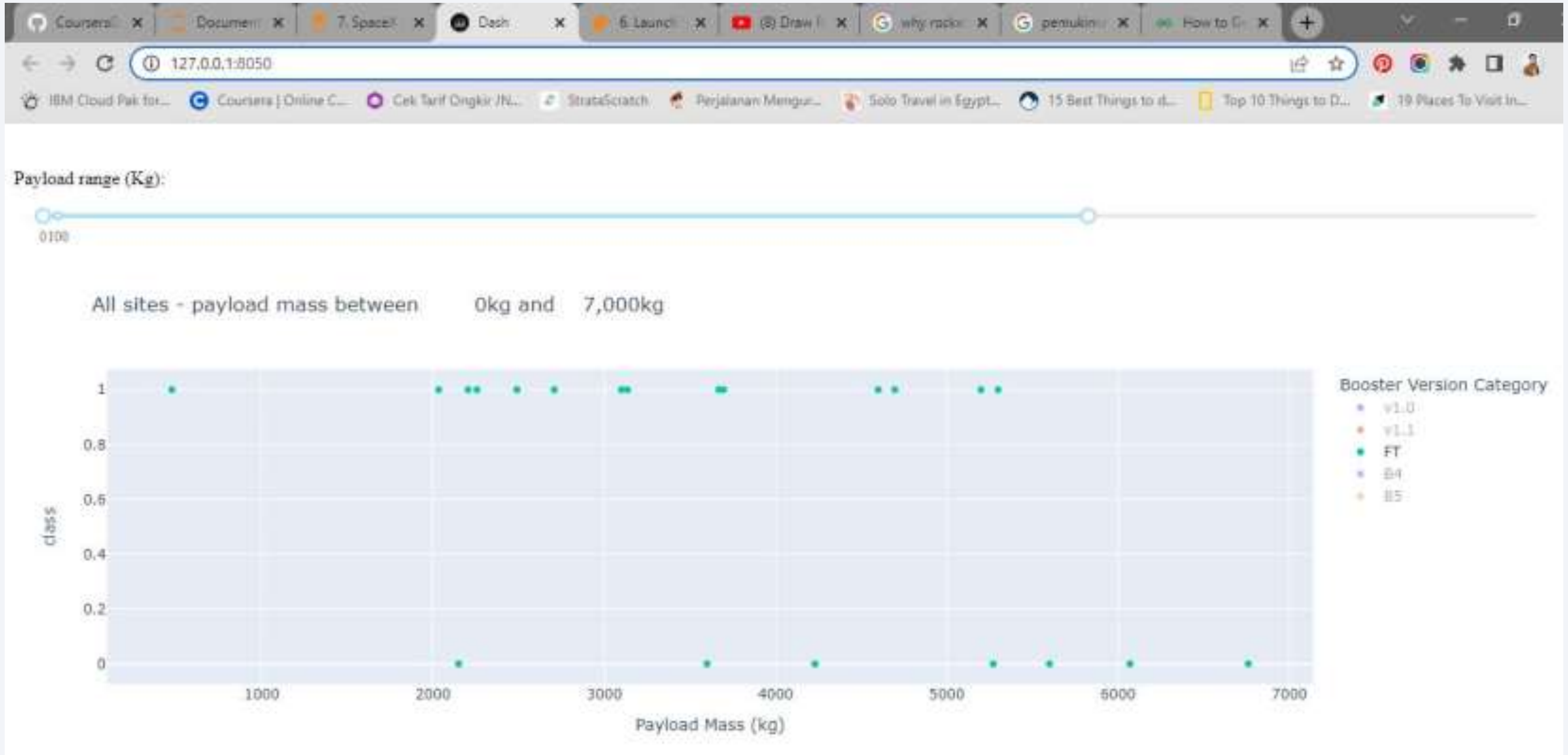
We still need to observe more whether the high success rate is affected by launch site location only (in this case KSC LC-39A) or there are another factors to consider.

# Site with Highest Launch Success Ratio


127.0.0.1:8050
SpaceX Launch Records Dashboard
KSC LC-39A
Total Launches for site KSC LC-39A
23.1%
76.9%
🟦 1
🟥 0

This site (KSC LC-39A) has high number of success launch (76,9%) compared to the failed one (23,1%).

# Payload vs Launch Outcome



Payload range (Kg):

All sites - payload mass between 0kg and 7,000kg

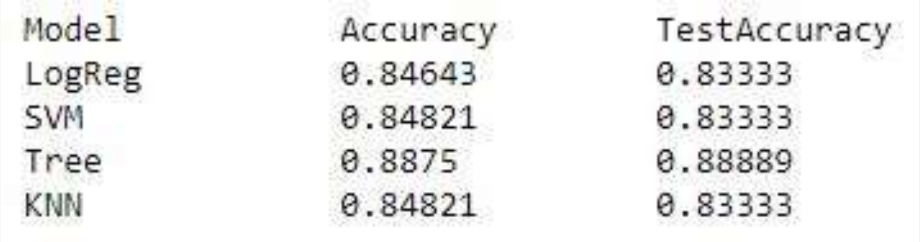
FT booster works well with payload mass under 7.000kg

Section 5

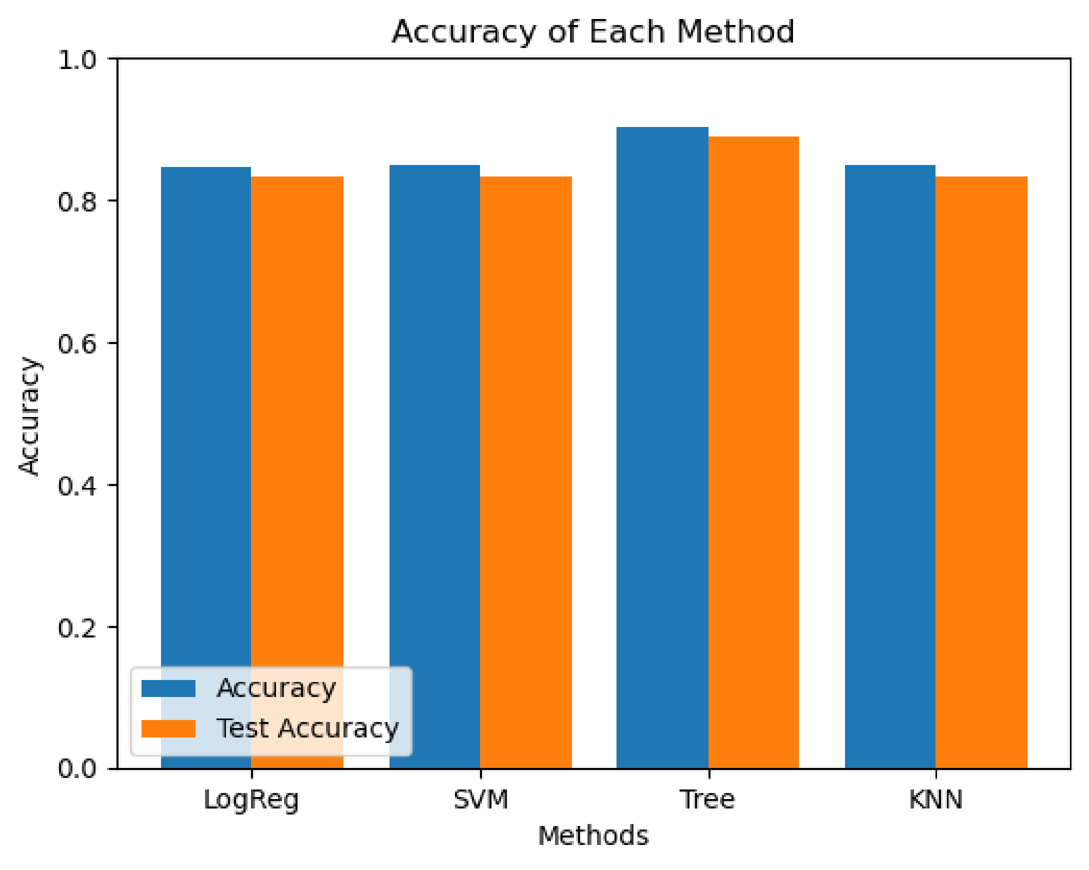
# Predictive Analysis (Classification)

# Classification Accuracy

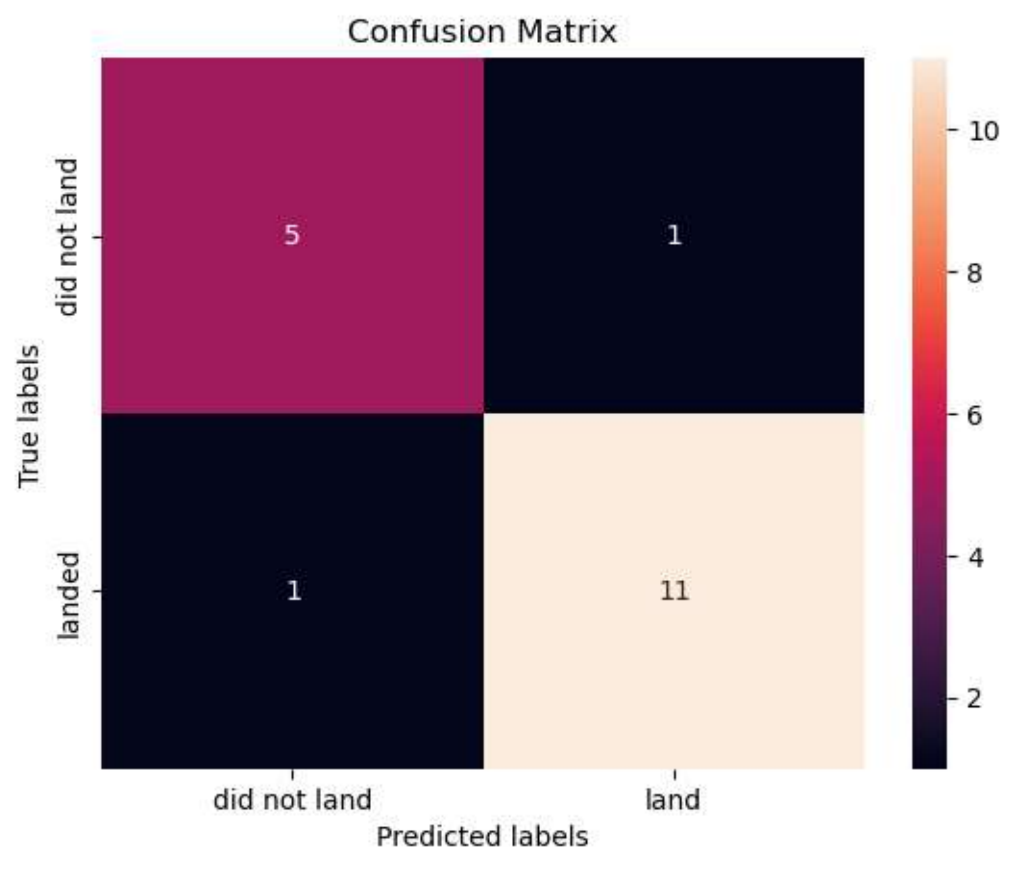
From the chart and calculation, we can see that decision tree method have the highest accuracy compared to the other methods (almost 90%).



```
Model         Accuracy        TestAccuracy
LogReg        0.84643         0.83333
SVM           0.84821         0.83333
Tree          0.8875          0.88889
KNN           0.84821         0.83333
```



# Confusion Matrix



High true positive (5 compared to 1) and high true negative (11 compared to 1) is a sign of good accuracy of a prediction method, in this case, it’s the confusion matrix of decision tree.

# Conclusions

- There are important factors that affect rocket launch success rate such as launch site, payload mass, and orbit. We can use this information to develop efficiency in SpaceY new business.
- The best launch site is KSC LC-39A.
- Launch sites are located near the coastline to prevent danger to third parties when unexpected things happened.
- Decision tree method can be used to predict future successful launch.
- Technology improvement is always needed to keep the business growing.

# Appendix

- Some visual doesn’t appear on GitHub so I put screenshots in folder.

Thank you!