Winning Space Race with Data Science

B.P. Aravind Arulmoli

25 Sept 2020

Outline

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

Executive Summary

- **Summary of methodologies**
  - Data Collection
    - using SpaceX Rest API
    - Wikipedia Web Scrapping
  - Data Wrangling
  - EDA with SQL
  - EDA with Data Visualization
  - Interactive Visual Analytics with Folium lab
  - Dashboard with Ploty Dash
  - Machine Learning Prediction
- **Summary of all results**
  - Insights drawn from EDA
  - Launch sites Proximity Analysis
  - Predictive Analysis (Classification)

Introduction

- Project background and context
  - The focus of this project is to predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch.

- Problems you want to find answers
  - Can we determine if the first stage will land successfully?
  - What are the factors that affect the outcome of the landing?
  - Can we predict the accuracy of successful landing given certain parameters?

Section 1

Methodology

Methodology

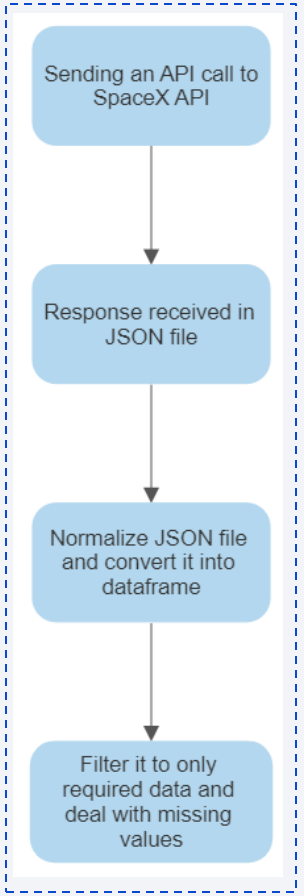
Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX Rest API and using Web Scrapping from Wikipedia.
- Perform data wrangling
  - Data was pre-processed by applying One Hot Encoding features.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Classification models such as Logistic Regression, Support Vector Machine, Decision Tree Classifier, K Nearest Neighbour were built and tuned for analysis.

Data Collection – SpaceX API

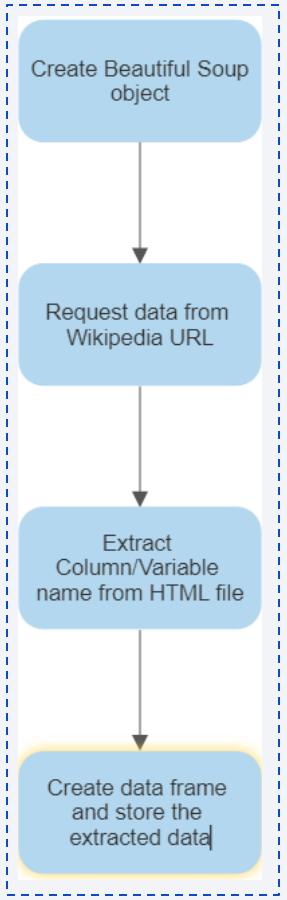
- The data is collected by calling the SpaceX rest API - https://api.spacexdata.com/v4
- The response JSON file that is received is normalized and converted into a data frame.
- The data is filtered and then the missing values are adjusted to the attribute mean value.

- GitHub URL: https://github.com/Arav-01/CapstoneProject-DataScience-IBMSkillsNetwork-Coursera/blob/main/Data%20Collection%20API.ipynb



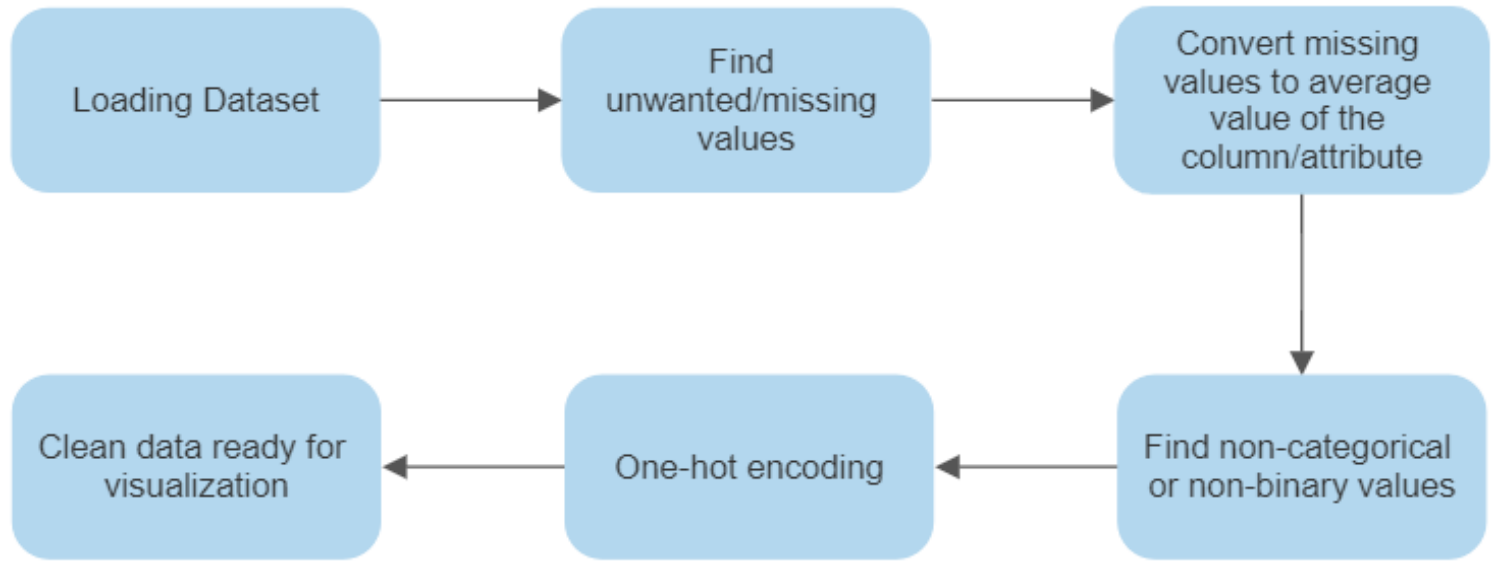
Data Collection - Scraping

- Beautiful soup object is created for HTML web scrapping.
- Data from Wikipedia URL - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches is parsed to Beautiful soup object.
- The data is extracted from HTML tables and stored in a data frame.



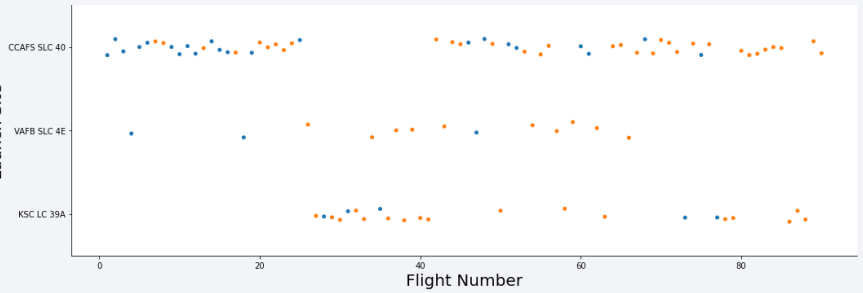
- GitHub URL: https://github.com/Arav-01/CapstoneProject-DataScience-IBMSkillsNetwork-Coursera/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb

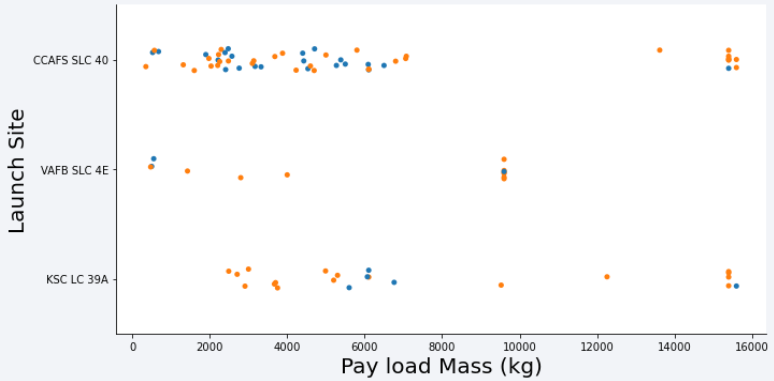
Data Wrangling

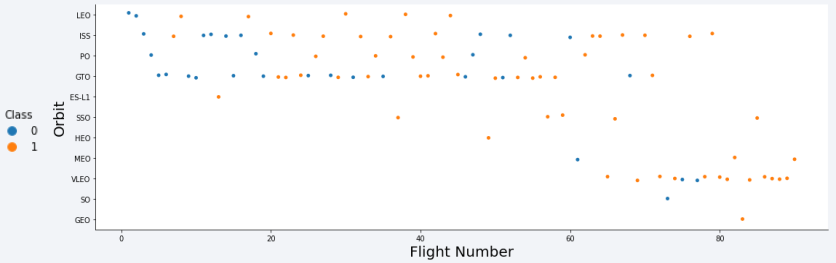


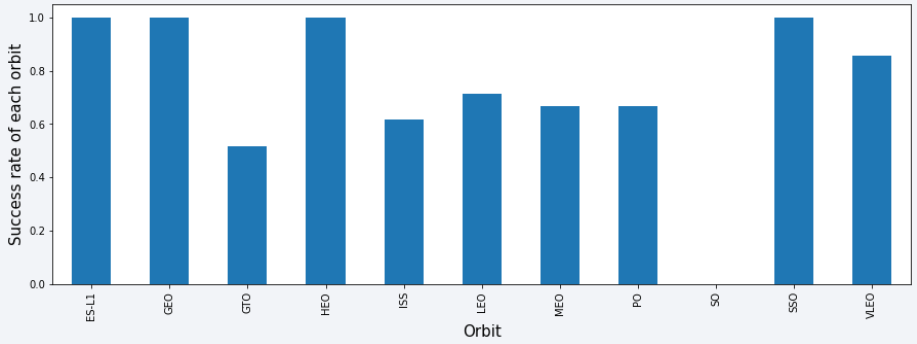
- Missing data is found and converted to the class average.
- Feature data is converted to binary values(0 or 1) by applying One Hot Encoding.
- GitHub URL: https://github.com/Arav-01/CapstoneProject-DataScience-IBMSkillsNetwork-Coursera/blob/main/EDA.ipynb

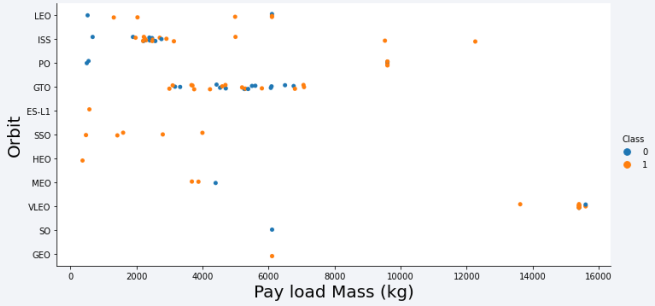
EDA with Data Visualization

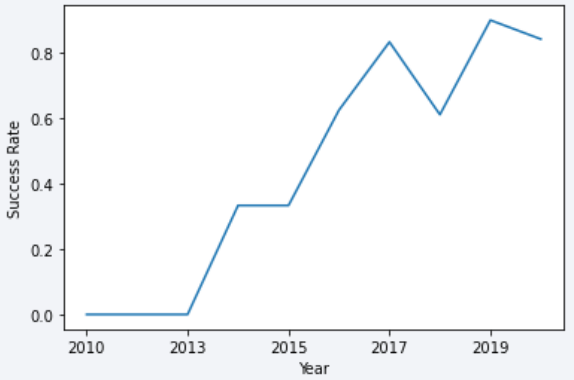












- Several graphs and plots were created using matplotlib and seaborn packages for visualizing the data.
- GitHub URL: https://github.com/Arav-01/CapstoneProject-DataScience-IBMSkillsNetwork-Coursera/blob/main/EDA%20with%20Data%20Visualization.ipynb

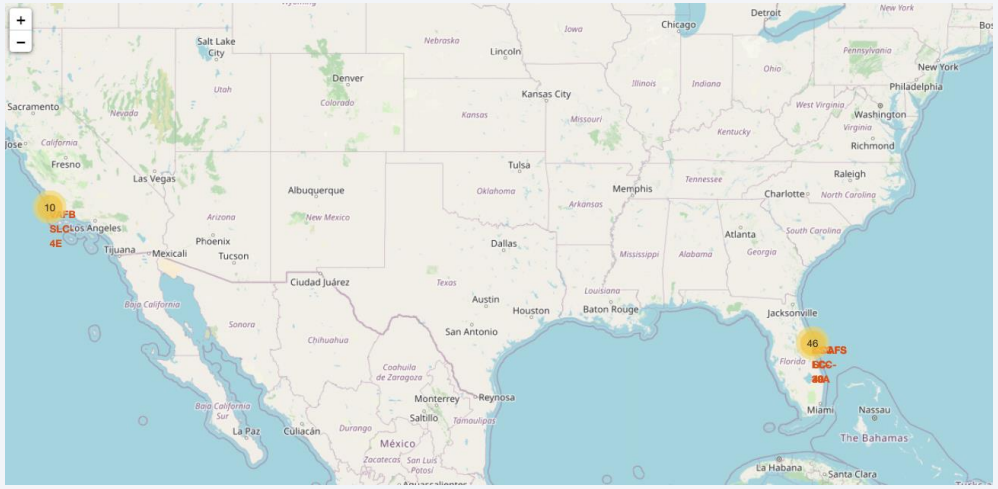
EDA with SQL

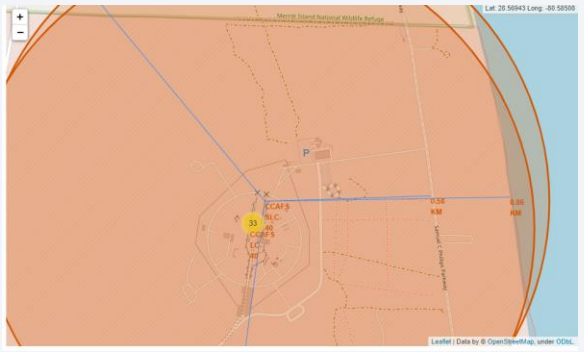
- Summary of SQL Queries performed
  - Displayed names of the unique launch sites.
  - Displayed 5 records of launch sites whose names begin with "CCA".
  - Displayed total payload mass from NASA (CRS)
  - Displayed average payload mass carried by booster version F9 v1.1.
  - Listed the date of first successful landing on a ground pad.
  - Listed total number of successful and failure mission outcomes.
  - Listed the boosters versions that carried maximum payload mass.
  - Listed records for failed drone ship landings in the year 2015.
  - Displayed the count of landing outcomes between 2010-06-04 and 2017-03-20, ranked in descending order

- GitHub URL: https://github.com/Arav-01/CapstoneProject-DataScience-IBMSkillsNetwork-Coursera/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

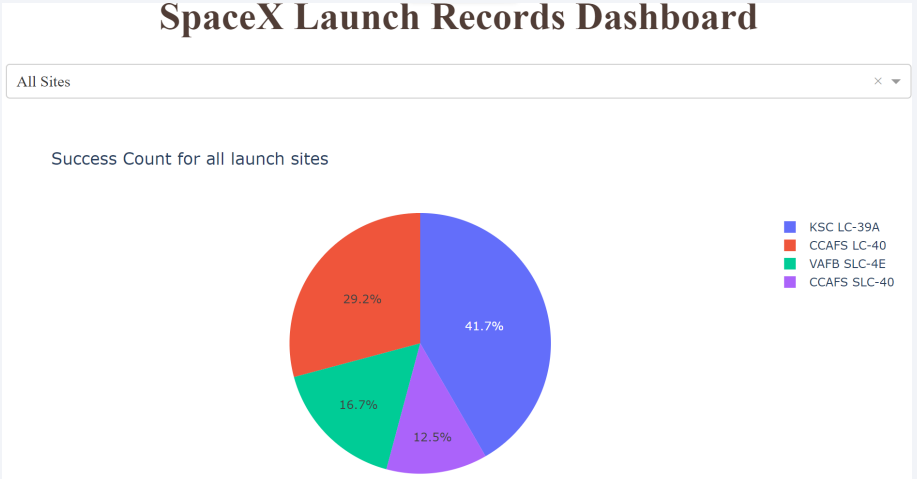
- The world map was created using the folium.Map object and centered to NASA and later the launch site coordinates.
- folium.Marker and folium.Circle objects were used to create a marker on launch sites and a circle around them.
- folium.Cluster object was used to create a group of markers at the same site and folium.Polyline was used to create lines to show distance to proximities.
- GitHub URL: https://github.com/Arav-01/CapstoneProject-DataScience-IBMSkillsNetwork-Coursera/blob/main/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb





Build a Dashboard with Plotly Dash

- Dashboard was build using Ploty and Dash and HTML components
- The plots that were created showed the launch success count and success rate of all the different launch sites.
- Dropdown menus were used to select different launch sites from a list.

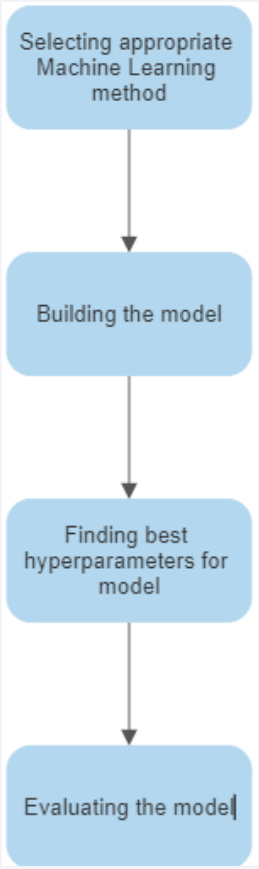


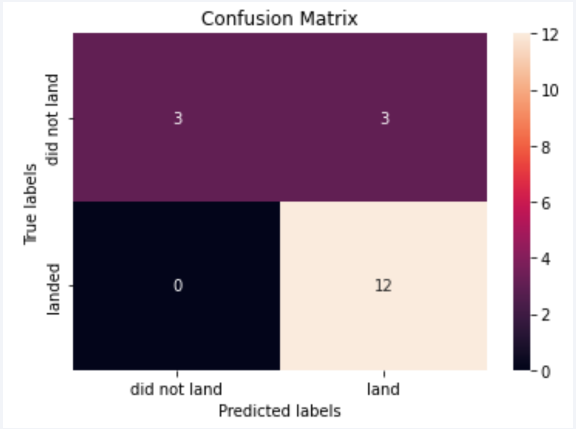
- Range slides were used to visualize payload mass relation with success rate for all the different launch sites.

- GitHub URL: https://github.com/Arav-01/CapstoneProject-DataScience-IBMSkillsNetwork-Coursera/blob/main/Interactive%20Dashboard%20with%20Ploty%20Dash

Predictive Analysis (Classification)

- Several different classification models were built and tuned to find the best performing classification model.
- Logistic Regression, Support Vector Machine, K Nearest Neighbour and Decision Tree Classifier models were used to predict the accuracy of the data.
- You need present your model development process using key phrases and flowchart
- GitHub URL: https://github.com/Arav-01/CapstoneProject-DataScience-IBMSkillsNetwork-Coursera/blob/main/Machine%20Learning%20Prediction.ipynb



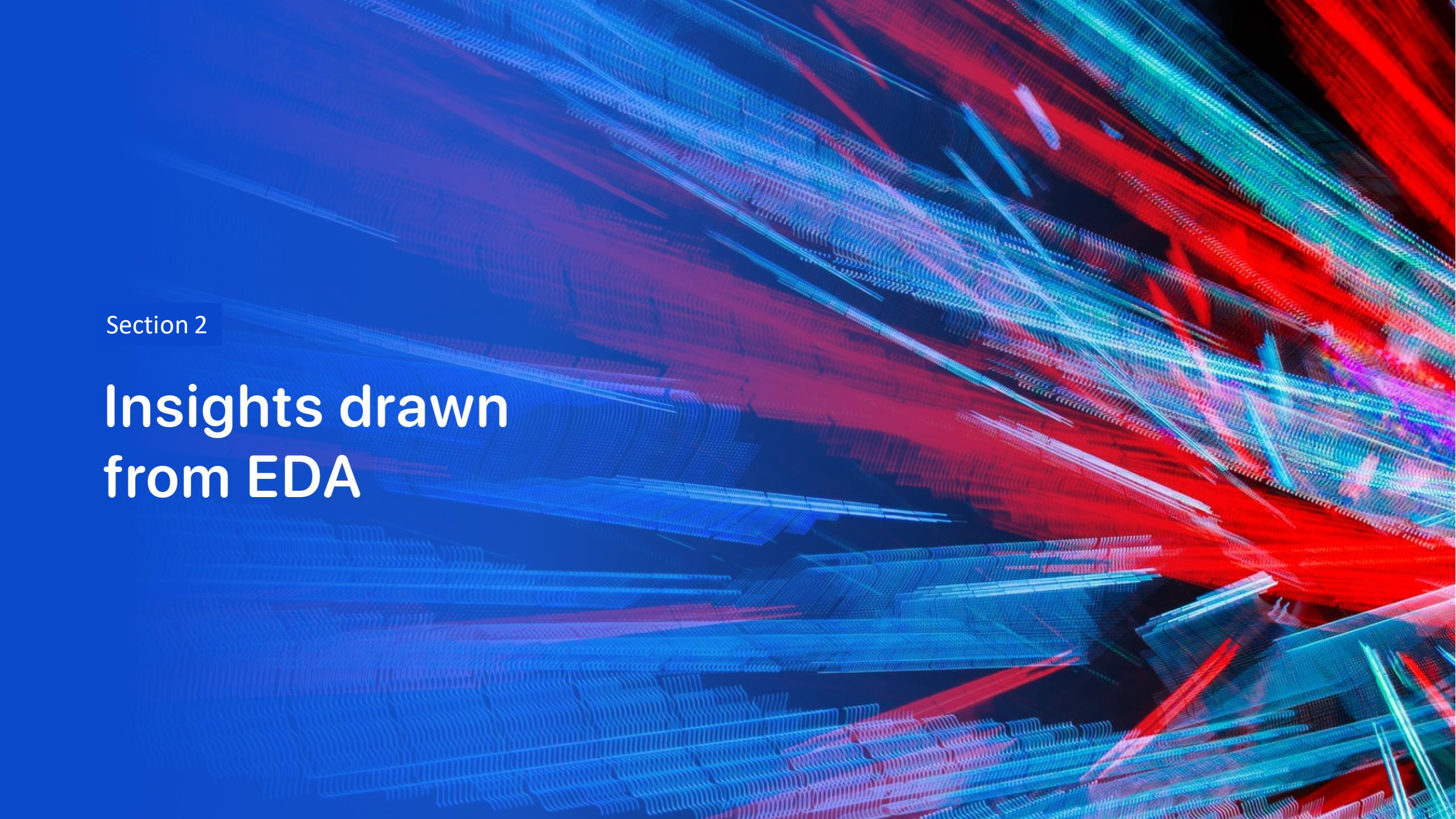


Results

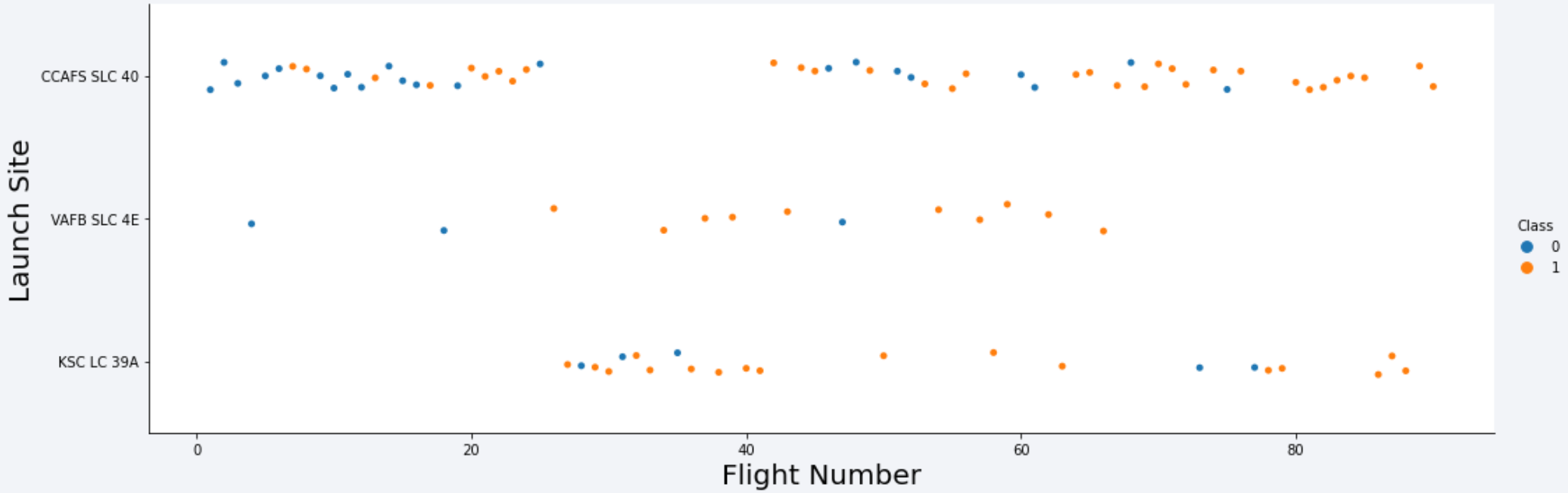
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

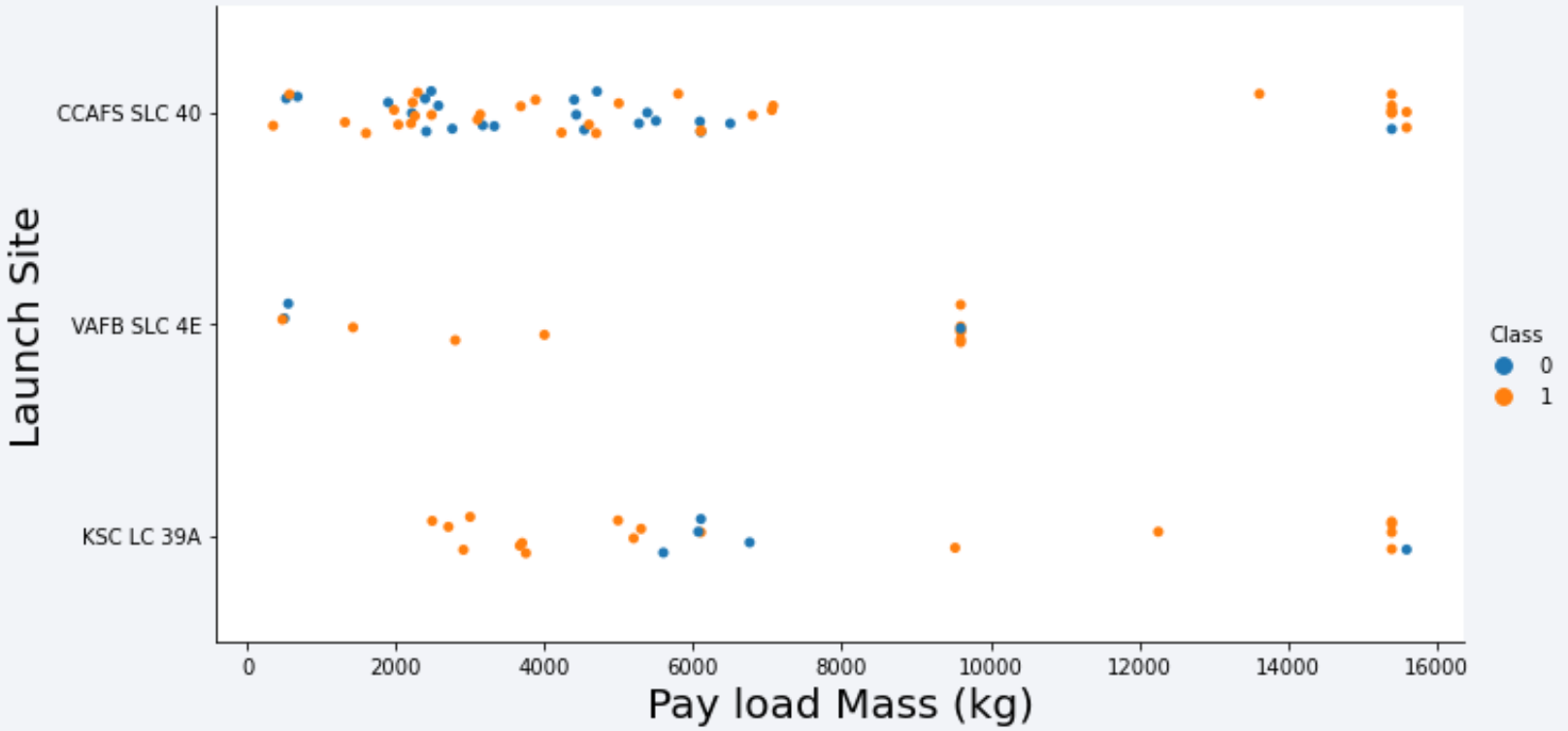


Flight Number vs. Launch Site



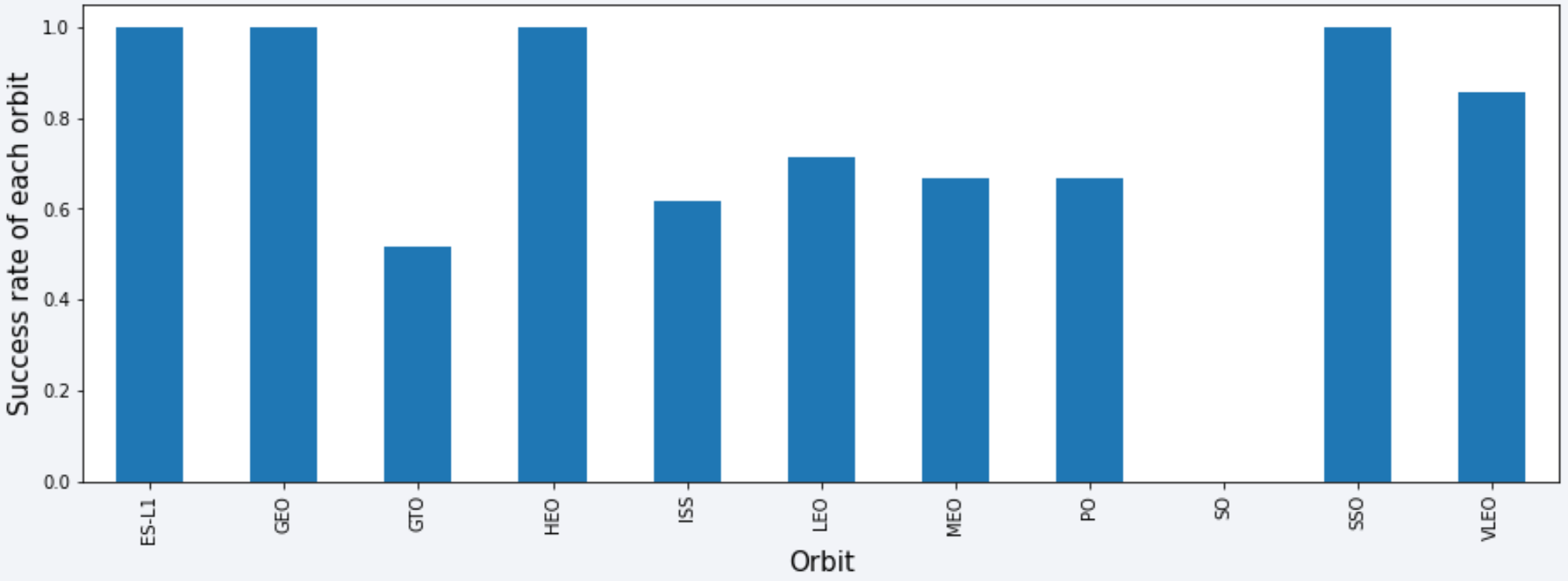
- Launch site CCAFS SLC 40 has had the most number of launches with most successes and also most failures.
- Sites VAFB SLC 4E and KSC LC 39A have lesser launches but comparatively better success rate with the general trend being the higher the flight number, the better the success rate.

Payload vs. Launch Site



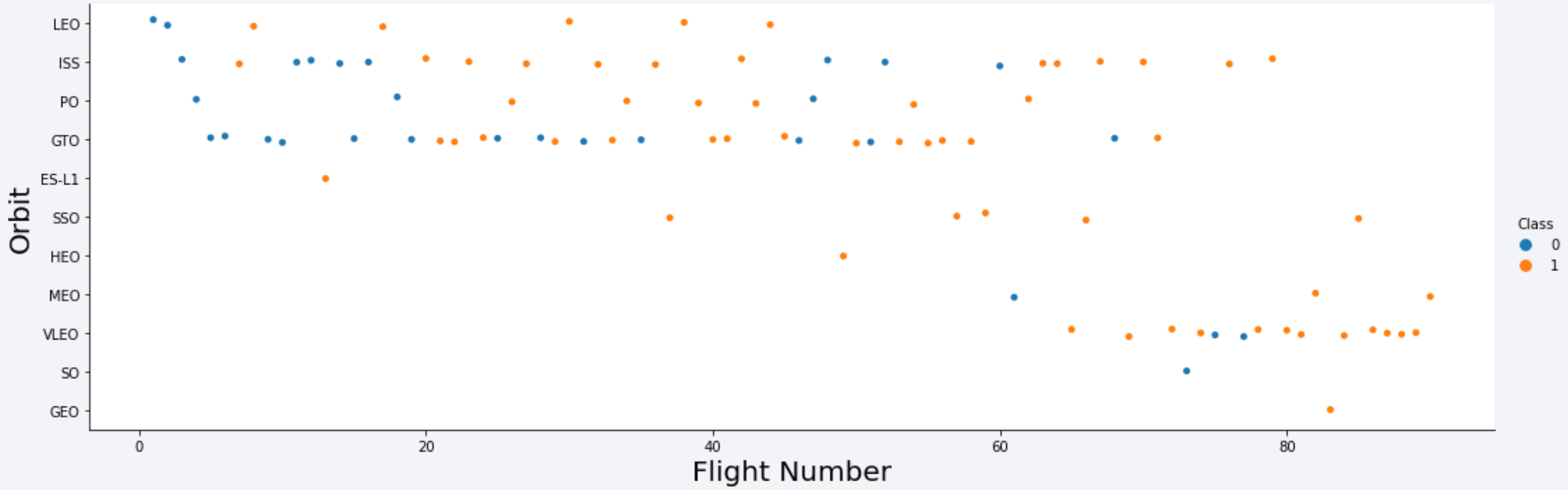
- Most launches from launch site CCAFS SLC 40 have payload mass of <8,000kg.
- Launch site VAFB SLC 4E has not launched anything with payload mass of >10,000kg.

Success Rate vs. Orbit Type



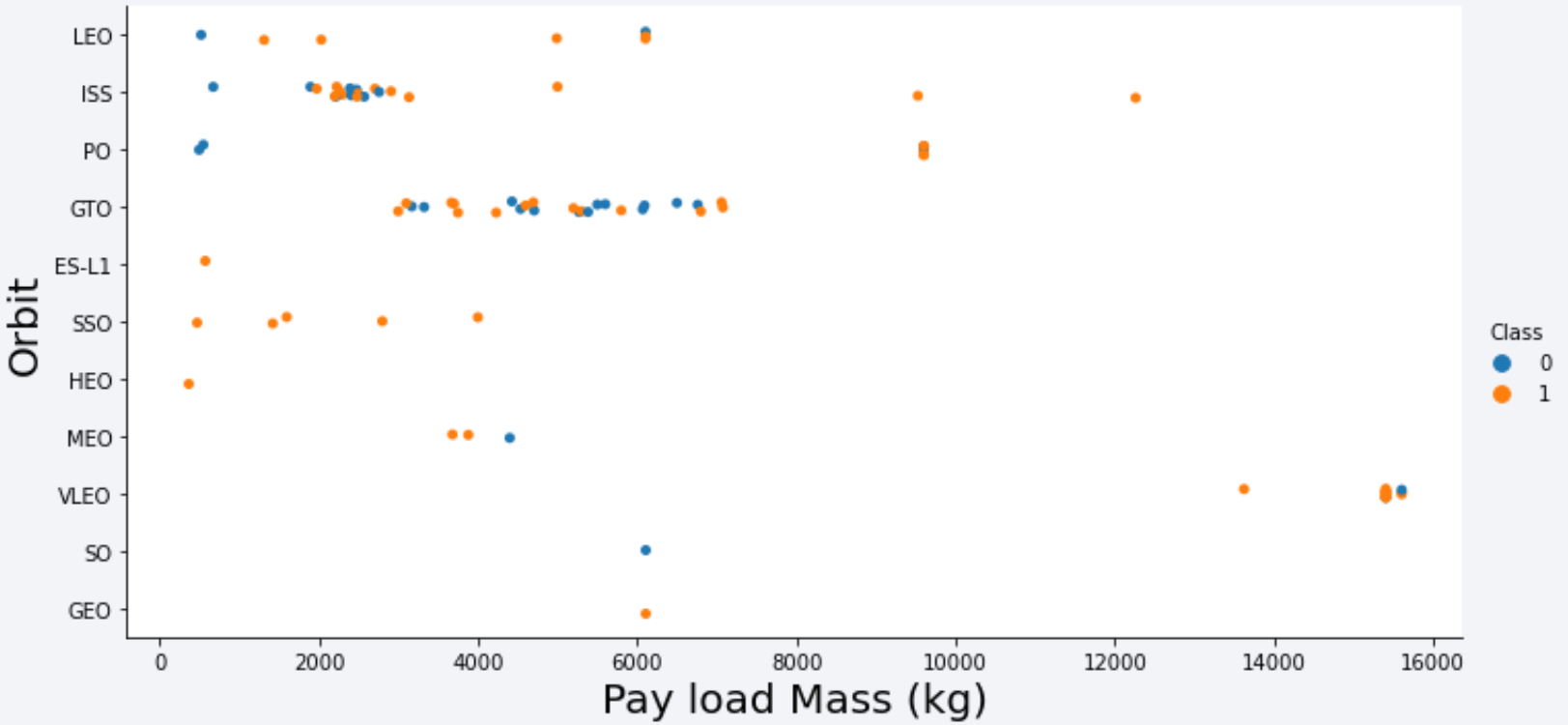
- ES-L1, GEO, HEO and SSO orbit types have a 100% success rate.
- SO orbit type has never had a successful outcome.

Flight Number vs. Orbit Type



- For LEO orbit type launches, the success rate seems to increase with higher flight number.
- The seems to be no relationship between flight number and success for other orbit types, especially for GTO orbit type.

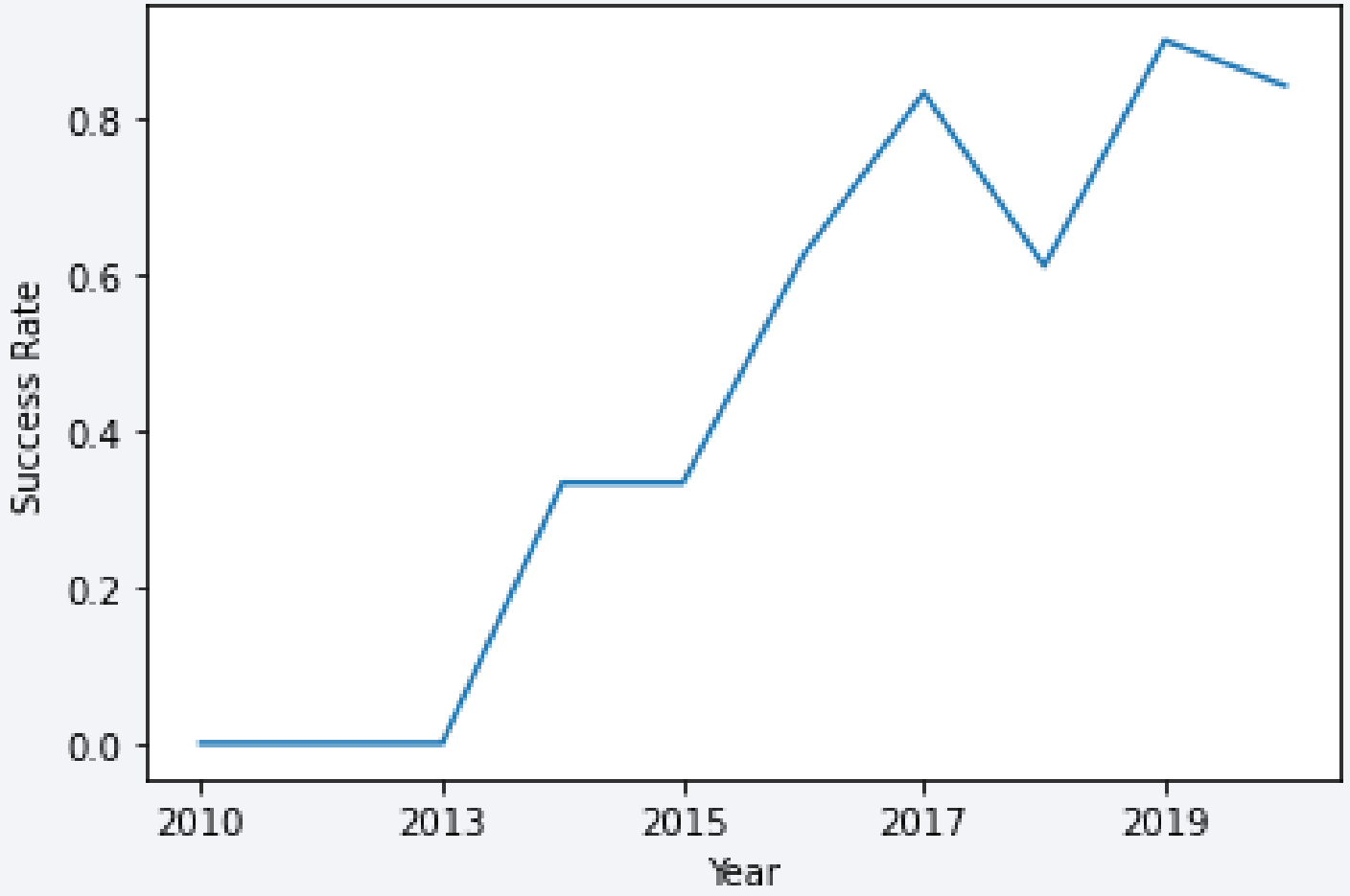
Payload vs. Orbit Type



- With heavy payloads, the successful landing or positive landing rate are more for PO, LEO and ISS orbit types.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing (unsuccessful mission) are both there here.

Launch Success Yearly Trend

- We can clearly observe that the sucess rate since 2013 kept increasing till 2020
- We can also see that there is a small dip in the success rate right after 2017.



All Launch Site Names

In [4]:

```
%%sql
SELECT DISTINCT(launch_site) AS "Unique launch sites"
FROM spacextbl
```

```
 * ibm_db_sa://gdp76693:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

Out[4]:

Unique launch sites
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Launch Site Names Begin with 'CCA'

In [5]:

```
%%sql
SELECT *
FROM spacextbl
WHERE launch_site
    LIKE 'CCA%' LIMIT 5
```

```
 * ibm_db_sa://gdp76693:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

Out[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 of NASA (CRS)

In [6]:

```
%%sql
SELECT SUM(payload_mass__kg_) AS "NASA (CRS) total payload mass"
FROM spacextbl
WHERE customer
    LIKE '%NASA (CRS)%'
```

```
 * ibm_db_sa://gdp76693:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od
8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

Out[6]:

NASA (CRS) total payload mass
48213

Average Payload Mass by F9 v1.1

In [7]:

```
%%sql
SELECT AVG(payload_mass__kg_) AS "F9 v1.1 average payload mass"
FROM spacextbl
WHERE booster_version = 'F9 v1.1'
```

```
 * ibm_db_sa://gdp76693:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od
8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

Out[7]:

F9 v1.1 average payload mass
2928

First Successful Ground Landing Date

In [8]:

```
%%sql
SELECT date AS "First successful ground pad landing"
FROM spacextbl
WHERE landing__outcome = 'Success (ground pad)' LIMIT 1
```

```
 * ibm_db_sa://gdp76693:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od
8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

Out[8]:

First successful ground pad landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

In [9]:

```
%%sql
SELECT booster_version AS "Boosters with drone ship success having payload mass of 4000-6000kg"
FROM spacextbl
WHERE landing__outcome = 'Success (drone ship)'
    AND payload_mass__kg_ BETWEEN 4000 AND 6000
```

```
 * ibm_db_sa://gdp76693:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

Out[9]:

Boosters with drone ship success having payload mass of 4000-6000kg
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

In [10]:

```
%%sql
SELECT mission_outcome, COUNT(mission_outcome) AS "Total outcomes"
FROM spacextbl
GROUP BY mission_outcome
```

```
 * ibm_db_sa://gdp76693:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

Out[10]:

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

Boosters Carried Maximum Payload

In [11]:

```
%%sql
SELECT booster_version AS "Boosters which carried the max payload mass"
FROM spacextbl
WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_)
                           FROM spacextbl)
```

```
 * ibm_db_sa://gdp76693:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.
cloud:31498/bludb
Done.
```

Out[11]:

Boosters which carried the max payload mass
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

In [12]:

```
%%sql
SELECT landing__outcome, booster_version, launch_site, date
FROM spacextbl
WHERE landing__outcome = 'Failure (drone ship)'
    AND YEAR(date) = 2015
```

```
 * ibm_db_sa://gdp76693:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomai
n.cloud:31498/bludb
Done.
```

Out[12]:

landing__outcome	booster_version	launch_site	DATE
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015-01-10
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015-04-14

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

In [13]:

```
%%sql
SELECT landing__outcome, COUNT(landing__outcome) AS "Total outcomes"
FROM (SELECT landing__outcome, date
      FROM spacextbl
      WHERE date BETWEEN '2010-06-04' AND '2017-03-20')
GROUP BY landing__outcome
ORDER BY COUNT(landing__outcome) DESC
```

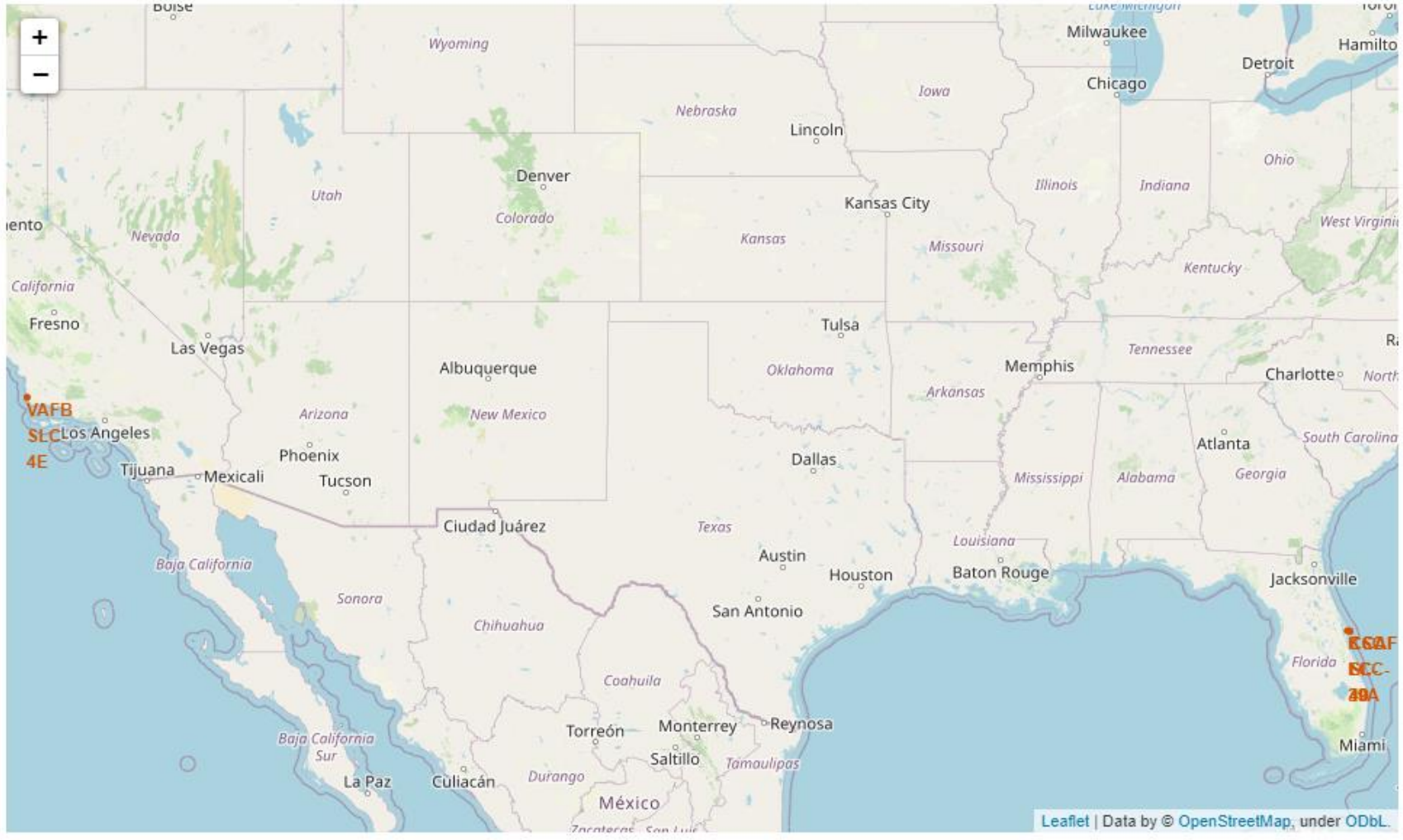
```
 * ibm_db_sa://gdp76693:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.
cloud:31498/bludb
Done.
```

Out[13]:

landing__outcome	Total outcomes
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

Marker for each Launch Site

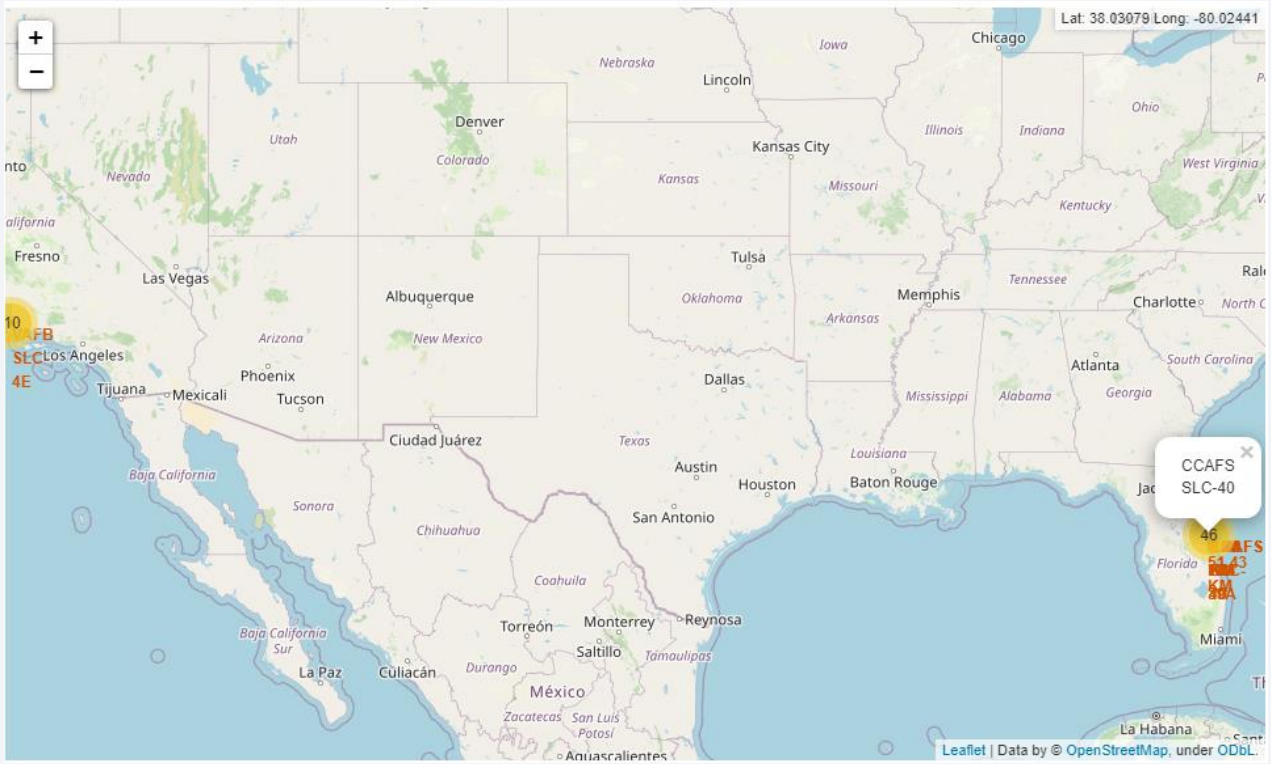


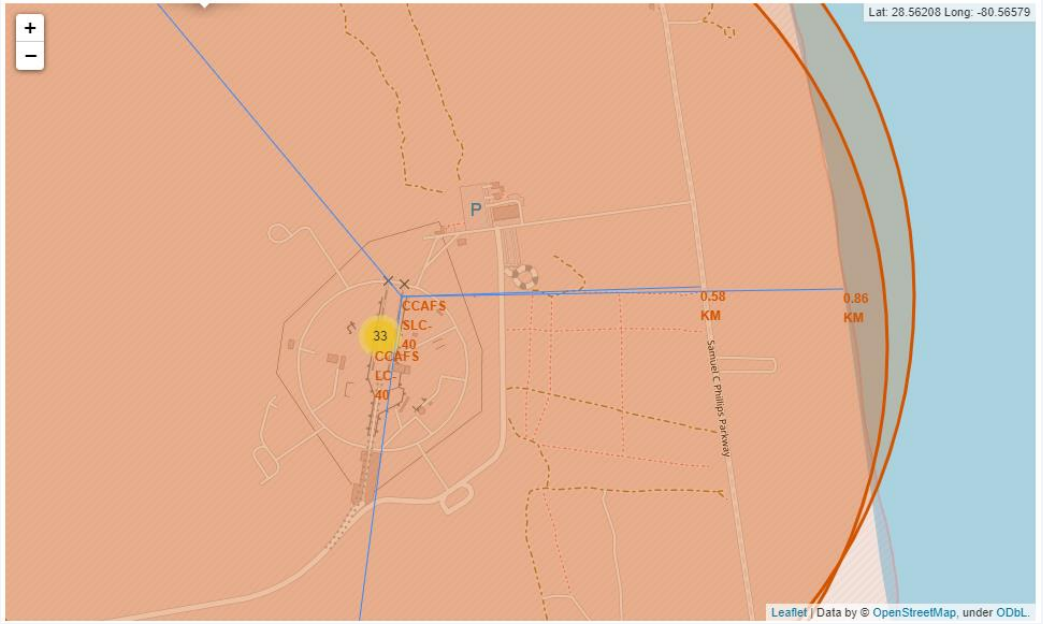
- The launch sites are either near the east coast or the west coast.
- 3 launch sites are in the state of Florida.
- 1 launch site is in the state of California.

Success and Failure for each Launch Site

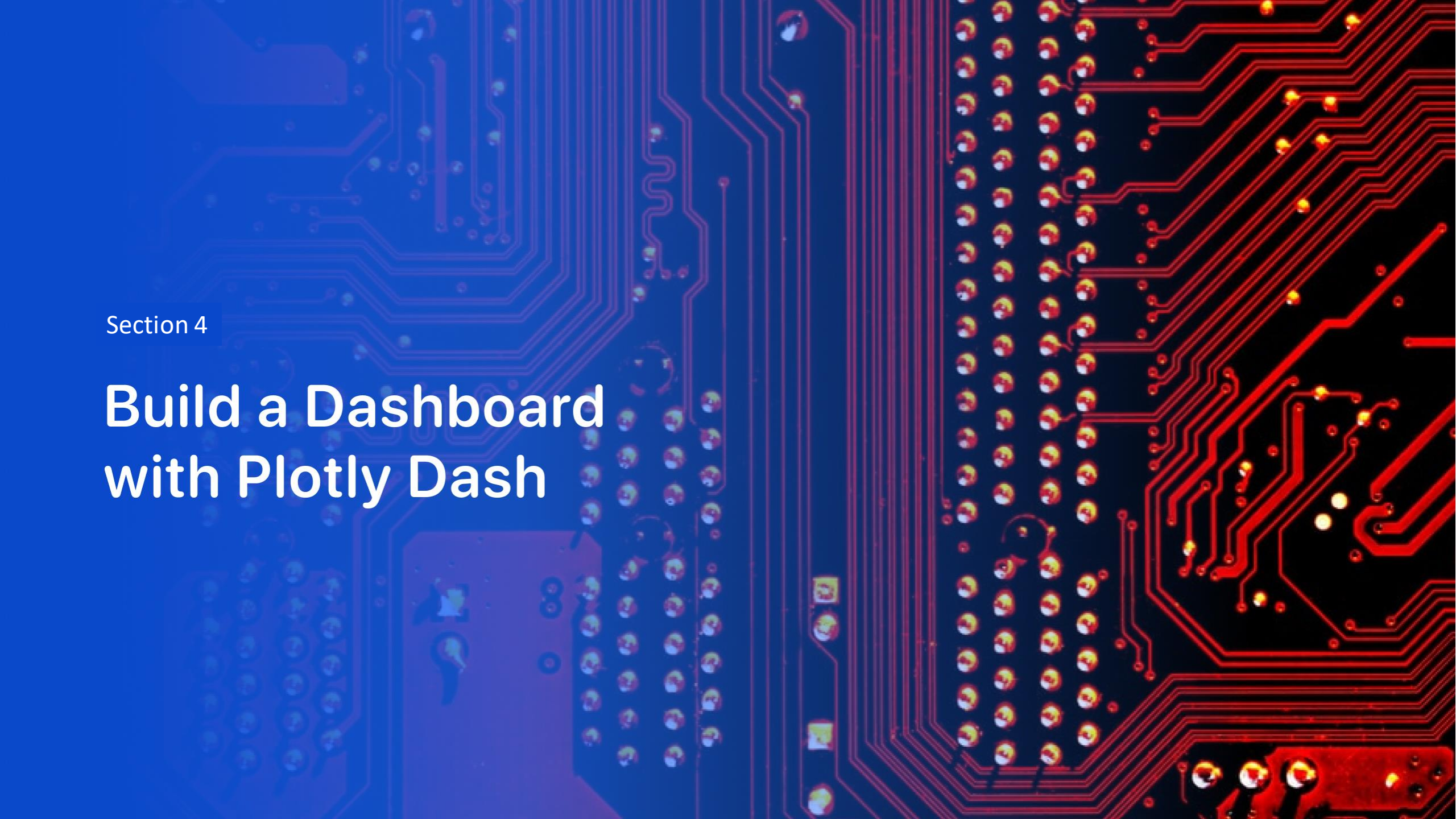
- Successful launches marked with green marker.
- Failures marked with red markers.

Distance between Launch Site to its Proximities

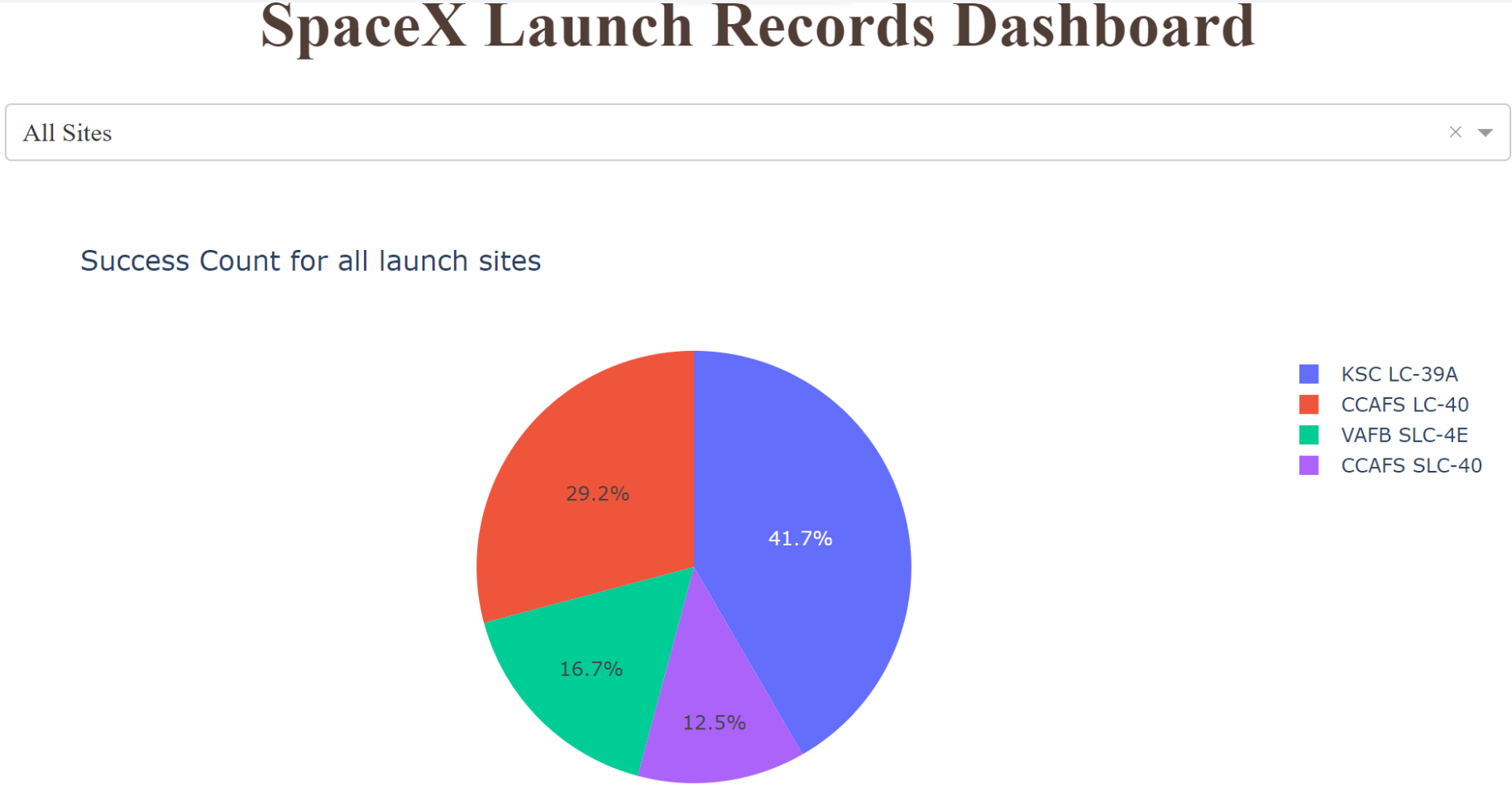




- The distances to the proximities such as railroad, highways and coastlines are marked with a blue line to indicate how far/close the launch sites to them.

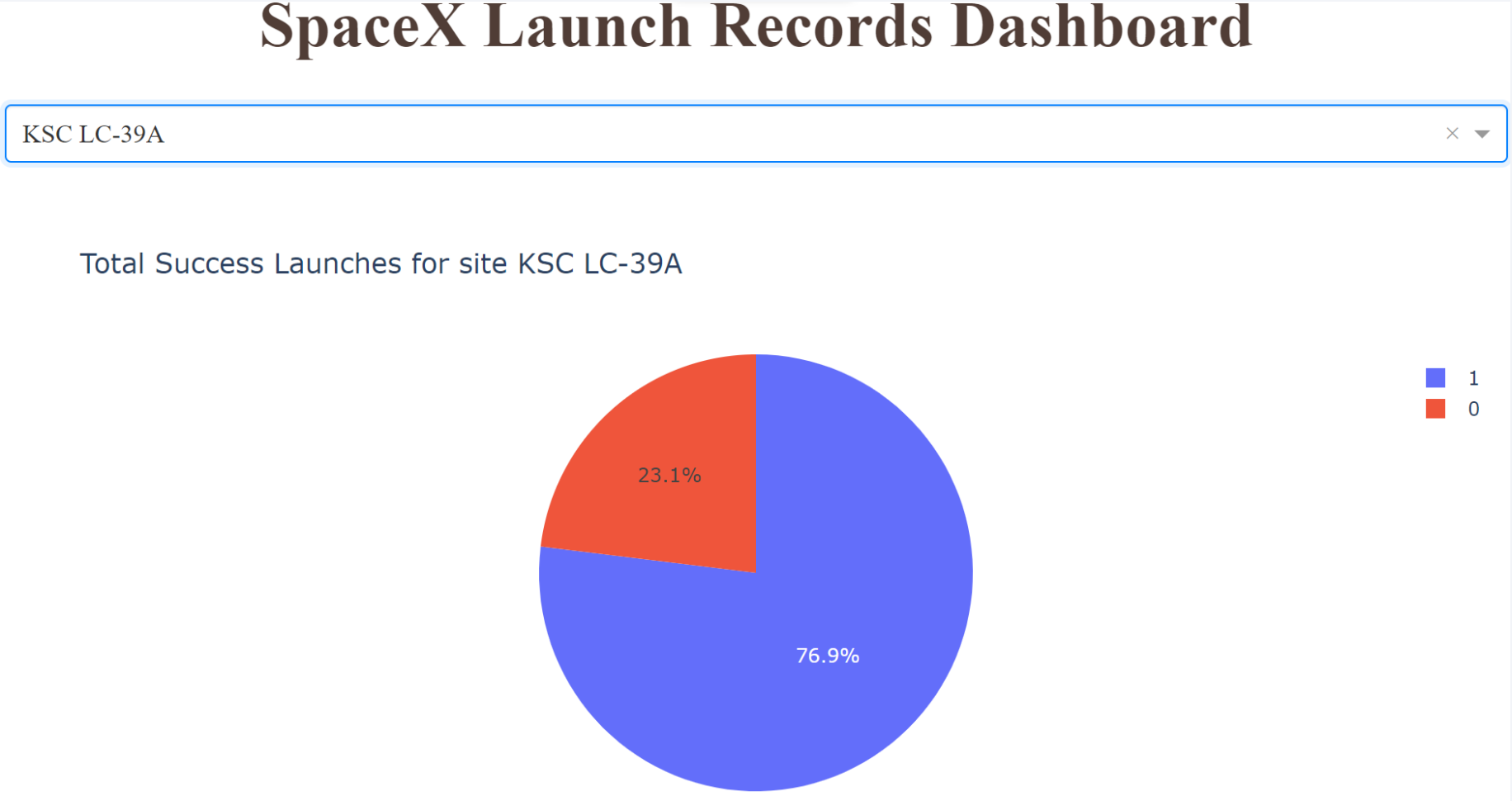

Section 4
Build a Dashboard with Plotly Dash

Pie Chart depicting success count of all Launch Sites



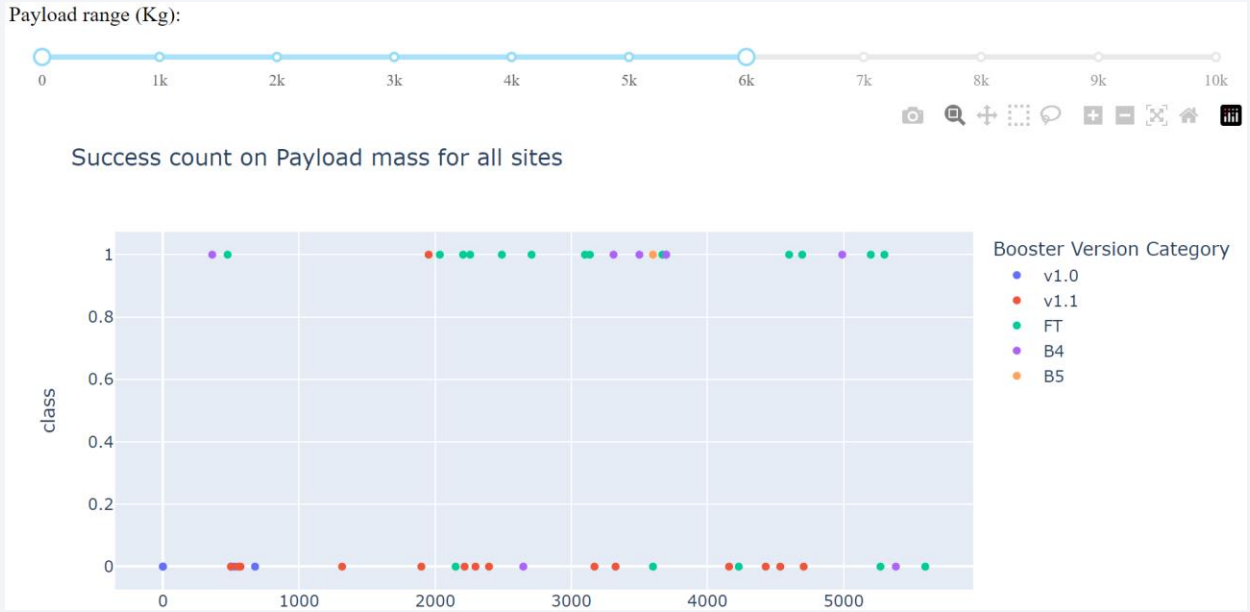
- Launch site KSC LC-39A is the most successful by far having a success count of 41.7%.
- All other launch sites have lower percent of success count, each of them having less than 30%.

Success Ratio of Lauch Site with Highest Success Rate



- The most successful launch site, KSC LC-39A has a success rate of 76.9%.
- Around 3/4th of all launches that take place here are successful.
- It has a failure rate of 23.1%.

Payload vs. Launch outcome for all Launch Sites



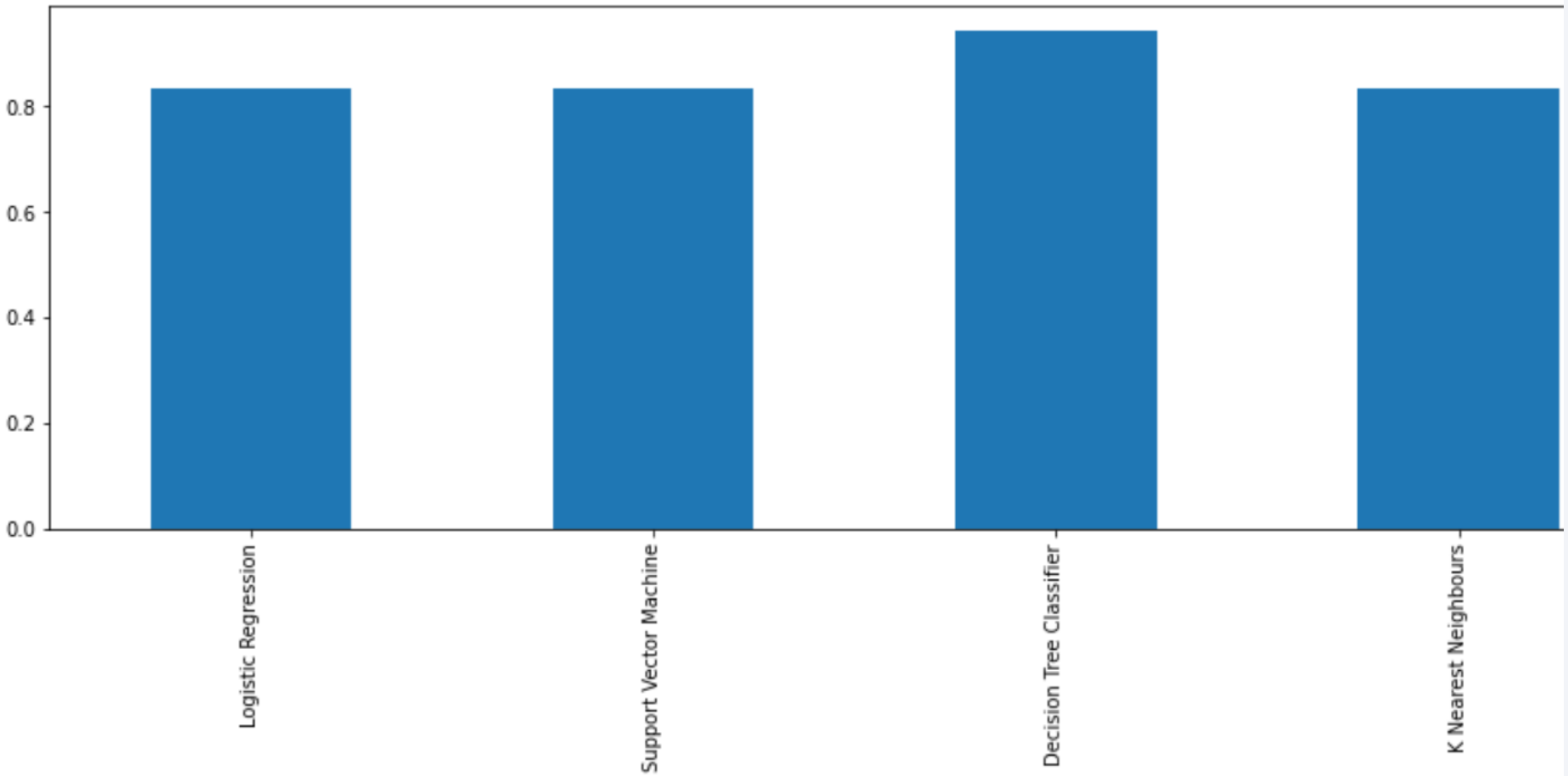
- From this scatter plot, we can observe that there are lot of launches with a payload lesser than 6000kg.
- Even though there seems to be a greater number of successes occurring, an equal number of failures also occur.

- From this scatter plot, we can see there are a lot less launches with a payload over 6000kg.
- Launches with payload of 6000-7000kg are almost always unsuccessful.

Section 5

Predictive Analysis (Classification)

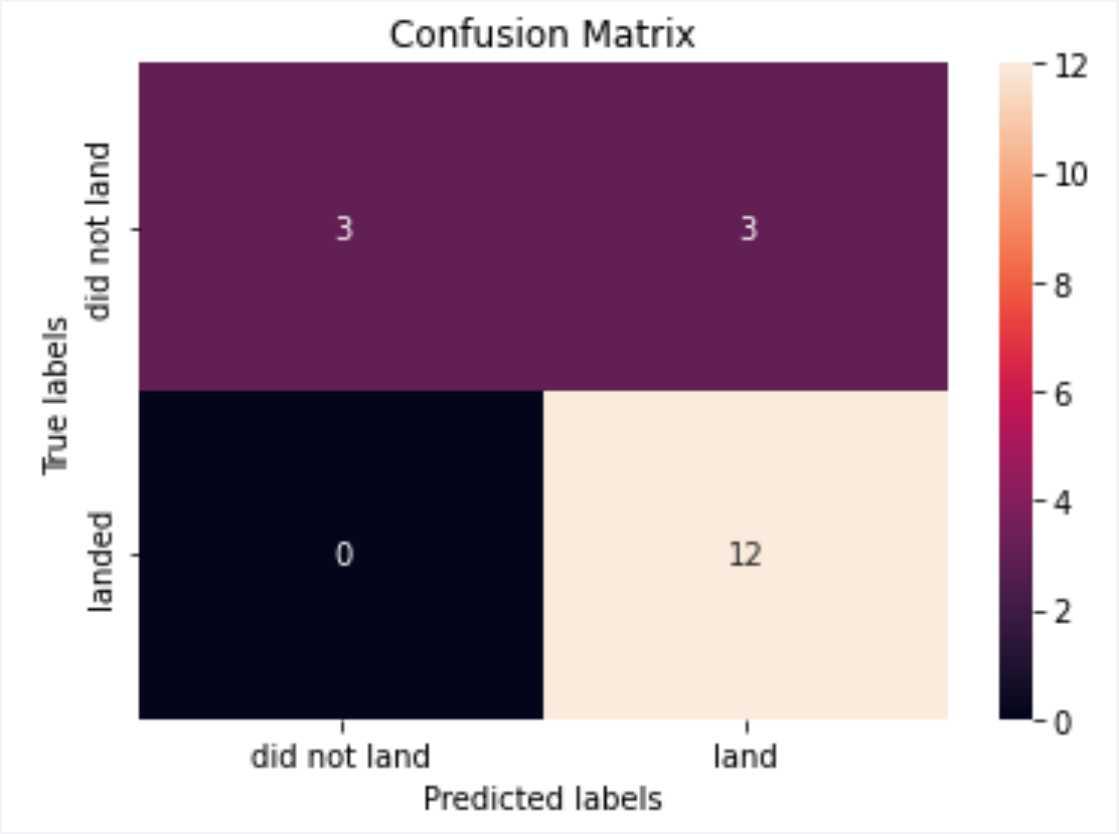
Classification Accuracy



- All 4 classification models have an accuracy rate of 0.83 or higher
- Decision Tree Classifier Model has the highest accuracy rate of ~0.9.

Confusion Matrix

- The following observations can be made from this confusion matrix
  - True Negative (TN) = 3
  - False Positive (FP) = 3
  - False Negative(FN) = 0
  - True Positive (TP) = 12



$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\text{-}score = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- F1-score = 0.89

Conclusions

- The Decision Tree Classifier model is the best model for prediction with a value of accuracy about ~0.90.
- Launches to certain orbits like Geo, HEO type orbits seem to have better success rate while launches to SO type orbit result in more failures.
- This launches that have had a lower payload mass, i.e., <6000kg have a slightly lower success rate than launches having heavier payloads, however launches with heavy payloads (>6000kg) happen less often.
- The success rate of the launches increases every year, as in every year the launches are more successful than the previous year.

Appendix

- All relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that have been created during this project can be found in the GitHub repository provided – https://github.com/Arav-01/CapstoneProject-DataScience-IBMSkillsNetwork-Coursera

Thank you!