IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Author: Tsunghao Chen
Date: 05/01/2024
Email: tsunghaochen@gmail.com

Outline

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

Executive Summary

- In this survey, we collected data from calling SpaceX API and used web scraping from wikipedia. We applied data wrangling to clean the collected data. Then, we used visualizations like bar plots, scatter plots, and even Maps and dashboard to gain insight from the data, and decide which attributes we should select for prediction model. Before modeling, we use one hot coding to turn categorical attributes into numerical ones.
- For predicting models, we tried with logistic regression, support vector machine, tree classification, and KNN, and see which one have the best performance. We used GridSearchCV to find out the best parameters for each model. The results show that in this case, all four models have the same score on the test set, but tree classification model has the best accuracy in train sets.

Introduction

- SpaceX is a private rocket company whose goal is to make commercial space travel achievable and affordable. They initiated a recyclable rockets that helps to reduce the cost significantly. Ever since 2013, the company had first success in recycling all the rockets and boosters. The success rate had grown over 50% since 2016.
- In this study, we would like to review the history data of all the SpaceX launches, and investigate if there is any significant attributes that has high correlations with the success rate. Furthermore, we would also like to build a model that will be able to predict the success rate for future launches based on the machine learning results from history data.

Section 1

Methodology

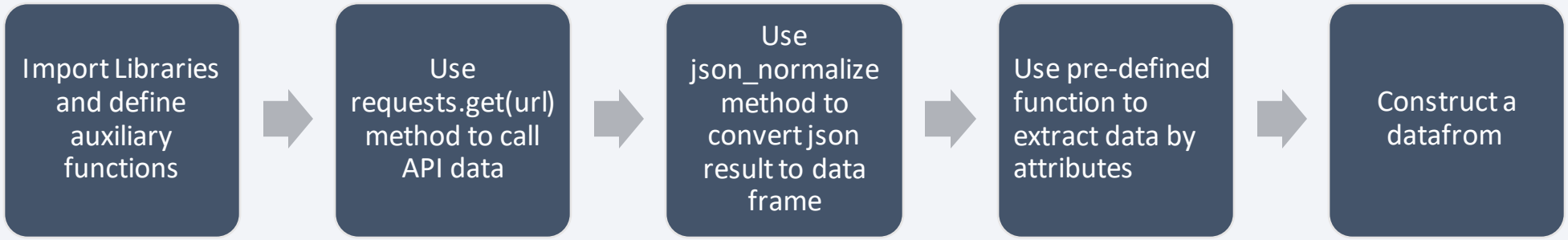
Methodology

Executive Summary

- Data collection methodology:
  - In this study, we collect rocket launch data from SpaceX API (https://api.spacexdata.com/v4/launches/past) and then used Pandas to compile in tabular data frame.
  - We also scraped the launch records of Falcon 9 from wikipedia with BeautifulSoup.
- Perform data wrangling
  - We also scraped the launch records of Falcon 9 from wikipedia with BeautifulSoup.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Used GridSearchCV method in Scikitlearn to find out the best parameters in the following preditive models: Logistic regression, KNN, SVM, and Tree Classification.

Data Collection – SpaceX API

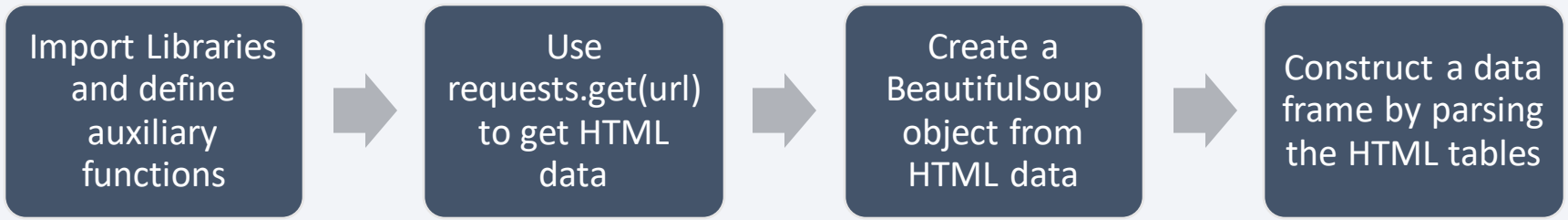
- Flowchart of SpaceX API calls



- GitHub URL: https://github.com/Tsunghao-C/IBM_DS_Capstone/blob/0b86590b0e7f227ced3ee76d07b26c351cf91c35/1_jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

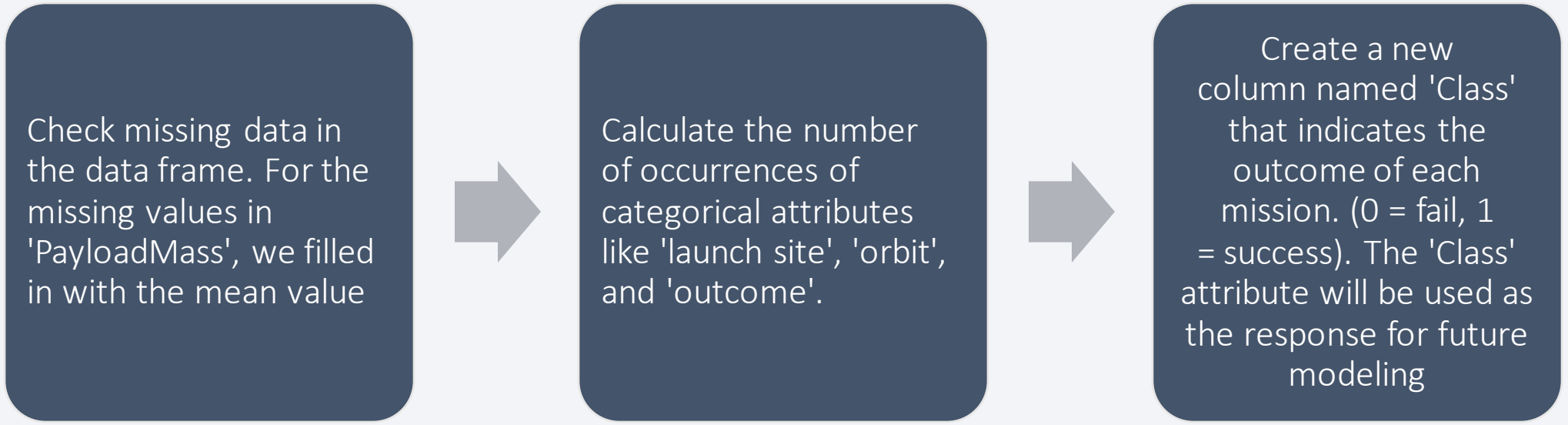
- Flowchart of Web Scraping



- GitHub URL: https://github.com/Tsunghao-C/IBM_DS_Capstone/blob/0b86590b0e7f227ced3ee76d07b26c351cf91c35/2_jupyter-labs-webscraping.ipynb

Data Wrangling

- Flowchart of data wrangling



- GitHub URL: https://github.com/Tsunghao-C/IBM_DS_Capstone/blob/0b86590b0e7f227ced3ee76d07b26c351cf91c35/3_jupyter-labs-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- First of all, we plotted several scatter plots to see the correlations among 'flight number', 'launch site', 'payload mass' and 'orbit'. The results are color coded with 'class' (mission outcome) so we can see how the attributes will impact the mission outcome.
- Secondly, we also plotted bar chart to see what the success rate looks like in different years or different orbits.
- GitHub URL: https://github.com/Tsunghao-C/IBM_DS_Capstone/blob/0b86590b0e7f227ced3ee76d07b26c351cf91c35/5_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

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 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 6000kg
- 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 records which displays the month names, failure landing_outcomes in drone ship, booster versions, launch_site for the months in year 2015
- GitHub URL: https://github.com/Tsunghao-C/IBM_DS_Capstone/blob/0b86590b0e7f227ced3ee76d07b26c351cf91c35/4_jupyter-labs-eda-sql-coursera_sqllite.ipynb

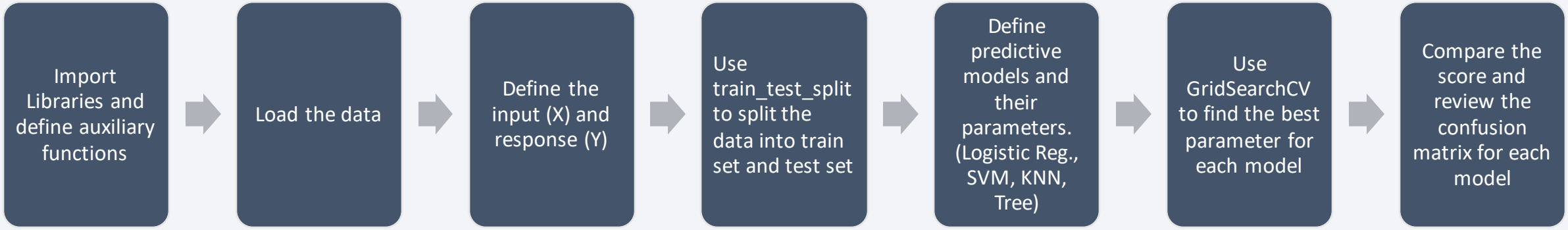
Build an Interactive Map with Folium

- Used Folium library to plot a map that marks the launch sites.
- Used Marker Cluster to plot all launches on the map by launch sites, and color coded by the launch outcome.
- Drew lines to indicate the distance to neighboring objects near launch sites.
- GitHub URL: https://github.com/Tsunghao-C/IBM_DS_Capstone/blob/0b86590b0e7f227ced3ee76d07b26c351cf91c35/6_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- Use Plotly Dash to create an interactive dashboard. There are two adjustable parameters: launch sites and the range of payload.
- There are two plots in the dashboard. The first one is the pie chart indicating the total success launces by launch site. If a specific site is chosen, the pie chart will represent the total success and failure launches of that particular site.
- The second plot is a scatter plot indicating the correlations of payload (in x axis) and launch outcome (in y axis). The results will be filtered according to the selected launch site and payload range.
- GitHub URL: https://github.com/Tsunghao-C/IBM_DS_Capstone/blob/0b86590b0e7f227ced3ee76d07b26c351cf91c35/7_spacex_dash_app.py

Predictive Analysis (Classification)



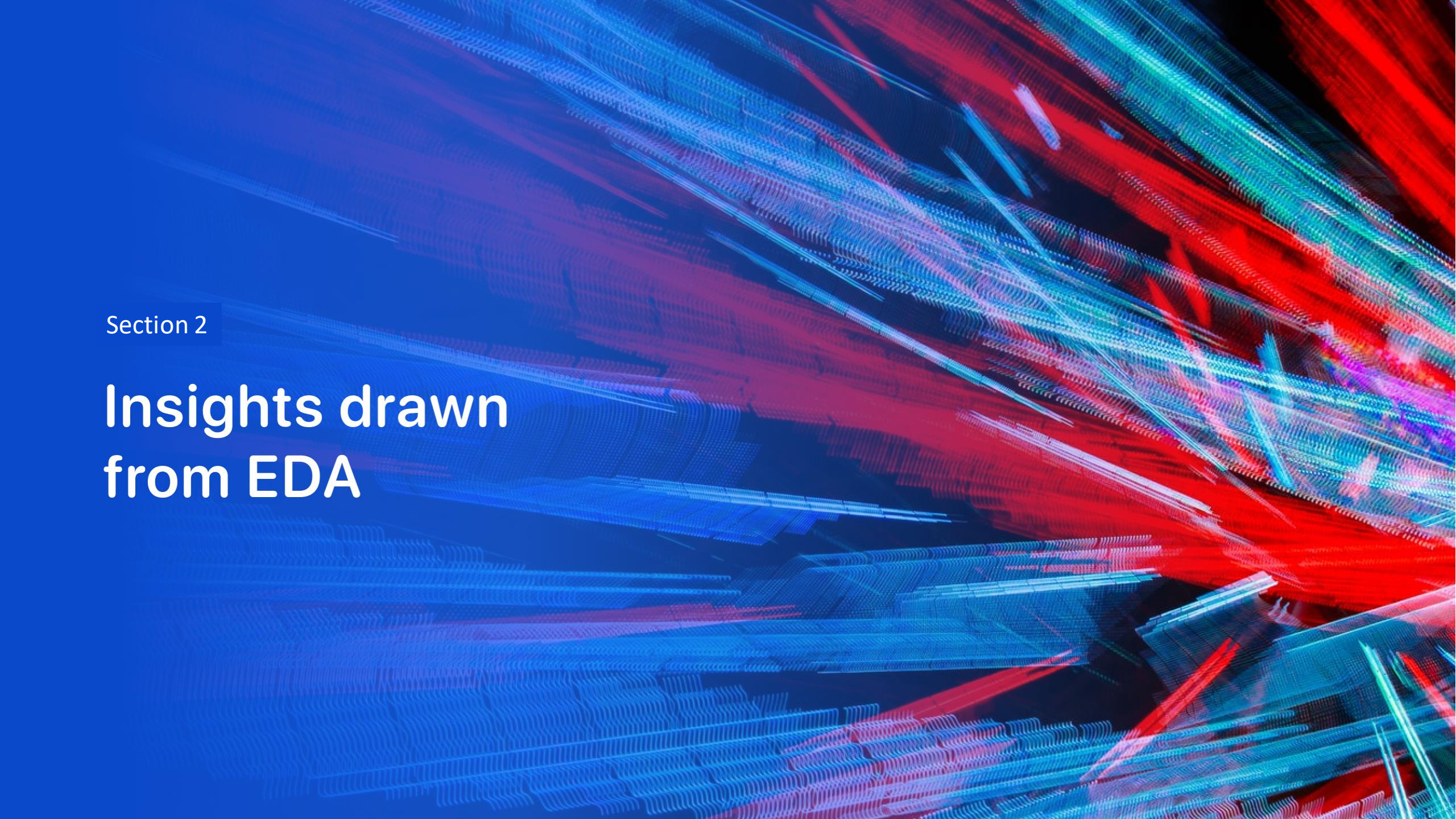
- GitHub URL: https://github.com/Tsunghao-C/IBM_DS_Capstone/blob/0b86590b0e7f227ced3ee76d07b26c351cf91c35/8_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

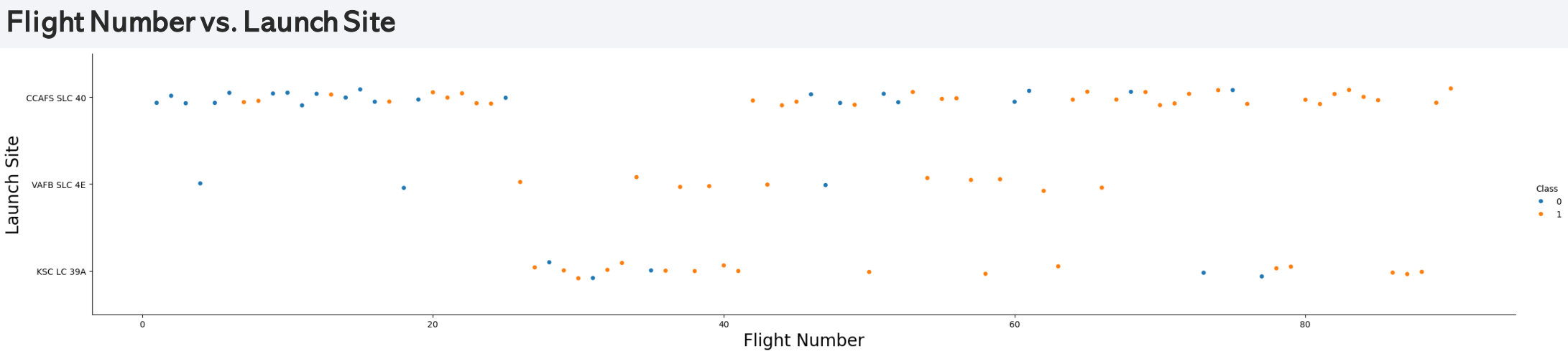
- Exploratory data analysis results (See section 2)
- Interactive analytics demo in screenshots (See section 3)
- Predictive analysis results (See section 4)

Section 2

Insights drawn from EDA

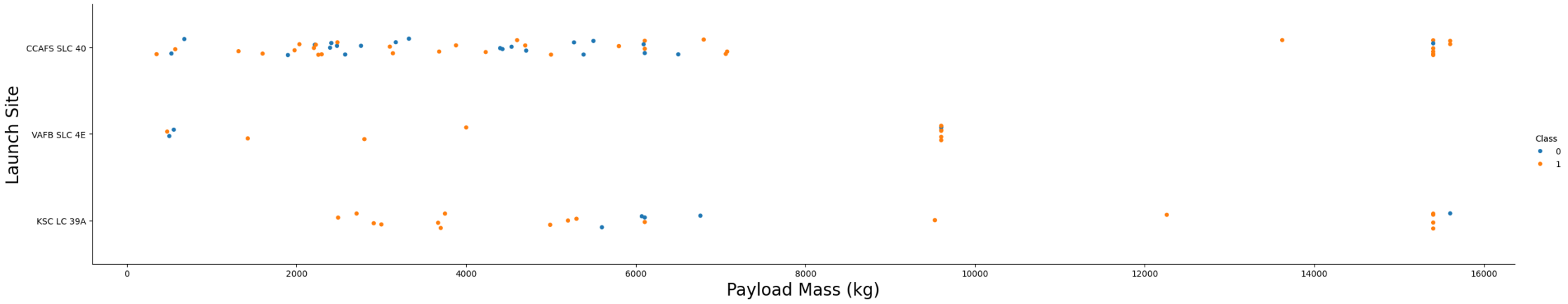


Flight Number vs. Launch Site



- We can see that 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%.
- The initial launches happened in CCAFS LC-40, and it did have quite a few failures in the early launches. We can see that the success rate started to improve after the 20th launch.
- The VAFB has the highest success rate, but the number of launches is also the smallest among all sites.

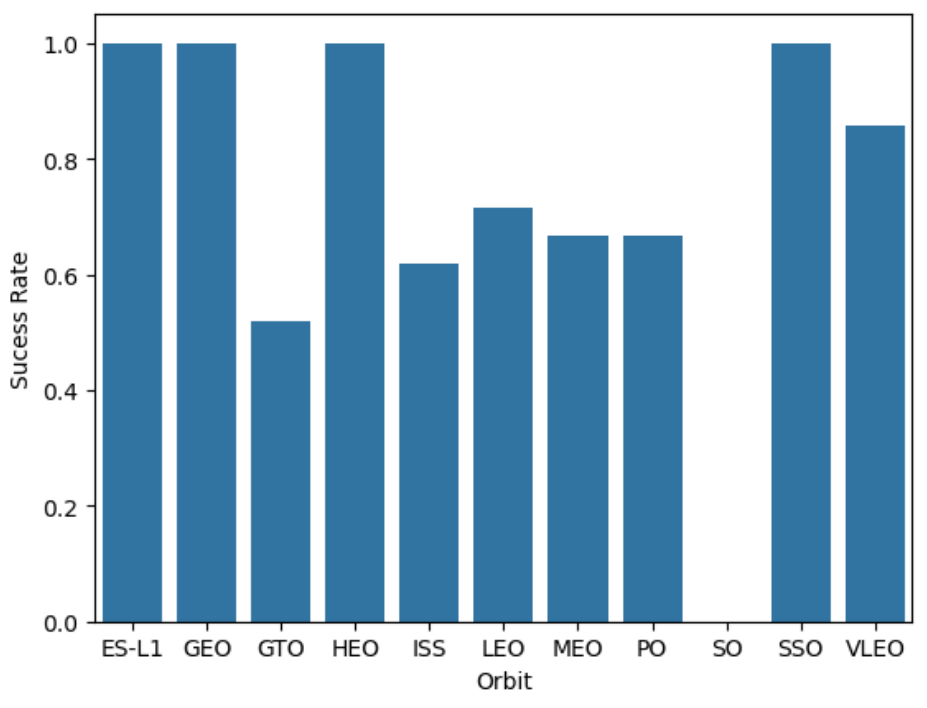
Payload vs. Launch Site



- We can see that there is a positive correlation between payload and success rate. The success rates are greater than 70% for all sites if the payload mass is greater than 8000kg.
- We can find that for the VAFB-SLC launch site, there are no rockets launched for heavy payload mass (greater than 10000kg)

Success Rate vs. Orbit Type

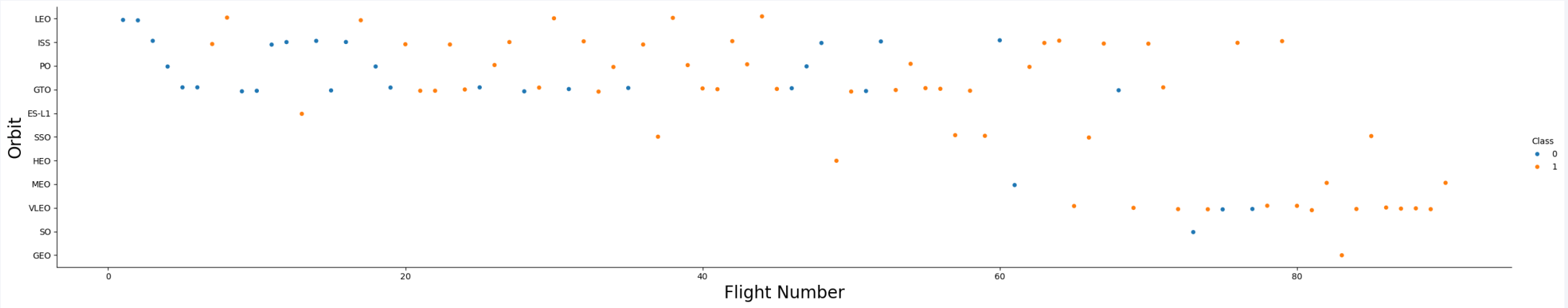
Success rate of each orbit type



- The success rate varies by different orbit type.
- We can see that the orbit types of ES-L1, GEO, HEO, and SSO have the highest success rate so far. However, there is only one launch record for HEO and GEO, and ES-L1.
- On the other hand, SSO has 5 launching records without any failure.

Flight Number vs. Orbit Type

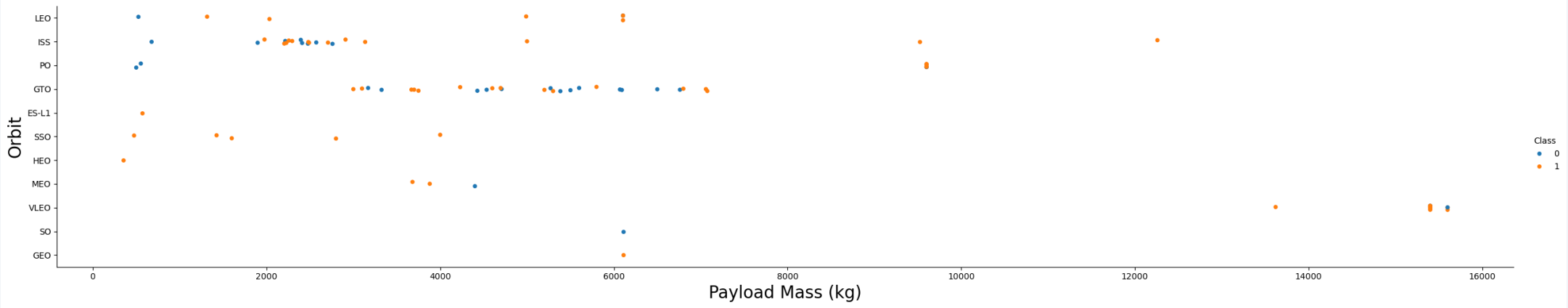
Scatter point of Flight number vs. Orbit type



- In the LEO orbit, the success appears to relate to the number of flights. On the other hand, there seems to be no relationship between flight number when in GTO orbit.
- The SSO orbit has 5 launches without a single failure.

Payload vs. Orbit Type

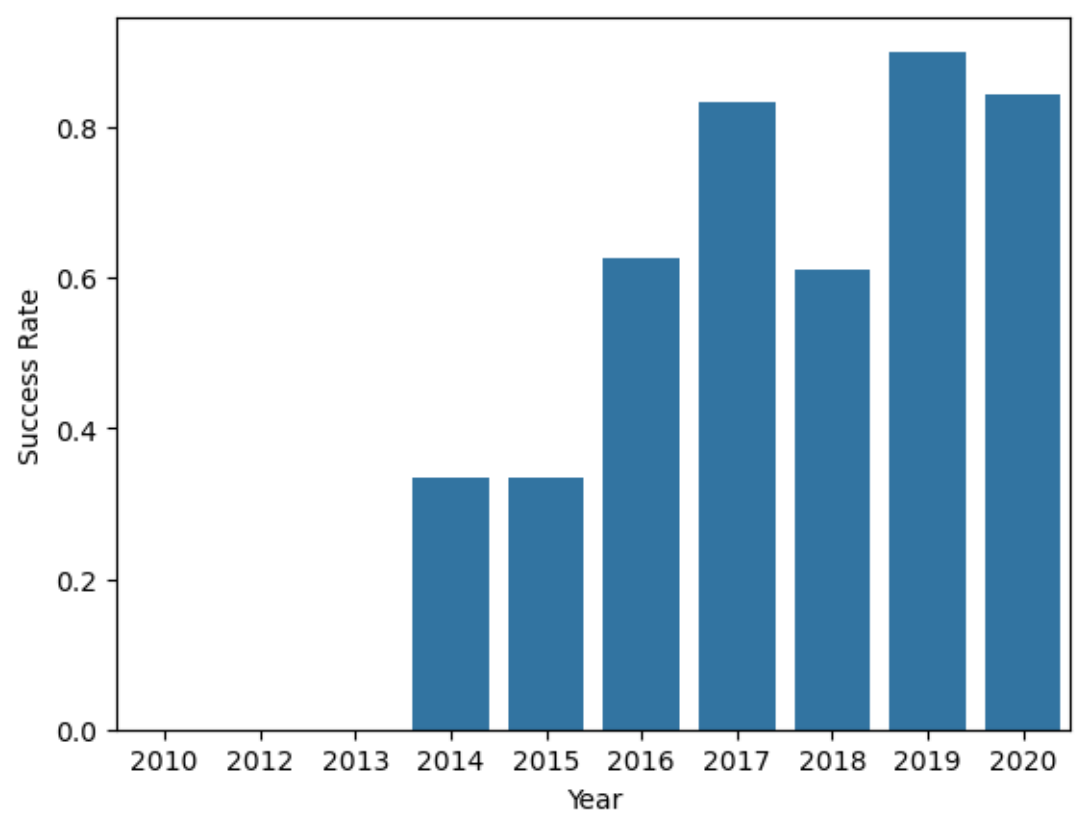
Scatter plot of payload vs. orbit type



- 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.

Launch Success Yearly Trend

Line chart of yearly average success rate



- The success rate is in general increasing over the years.
- The success rate surpassed 50% since 2016.

All Launch Site Names

- Find the names of the unique launch sites
- Select distinct launch site from table "SPACEXTBL"

Task 1

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

In [11]:
```
%sql select distinct "Launch_Site" from SPACEXTBL
```

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

Out[11]:

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Use 'LIKE' to search launch site data that starts with CCA;

Task 2

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

In [15]:

```
%%sql
select * from SPACEXTBL
where "Launch_Site" LIKE 'CCA%'
limit 5
```

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

Out[15]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (p
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 (p
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	N
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	N
201 0 01	Screenshot	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	N

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- Use 'SUM' to calculate the total payload and use 'LIKE' to find all NASA customers

Task 3

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

In [22]:
```
%%sql
select "Customer", SUM("PAYLOAD_MASS__KG_") as "Total_Payload" from SPACEXTBL
where "Customer" LIKE "NASA%"
group by "Customer"
order by "Total_Payload" DESC
```

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

Out[22]:

Customer	Total_Payload
NASA (CRS)	45596
NASA (CCDev)	12530
NASA (CCP)	12500
NASA (CCD)	12055
NASA (CTS)	12050
NASA (CRS), Kacific 1	2617
NASA / NOAA / ESA / EUMETSAT	1192
NASA (LSP) NOAA CNES	553
NASA (COTS)	525
NASA (LSP)	362
NASA (COTS) NRO	0

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Use 'AVG' to calculate the average payload and use 'LIKE' to find all F9 v1.1 booster versions

Task 4

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

In [25]:
```
%%sql
select "Booster_Version", AVG("PAYLOAD_MASS__KG_") as "AVG_Payload" from SPACEXTBL
where "Booster_Version" LIKE "%F9 v1.1%"
group by "Booster_Version"
```

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

Out[25]:

Booster_Version	AVG_Payload
F9 v1.1	2928.4
F9 v1.1 B1003	500.0
F9 v1.1 B1010	2216.0
F9 v1.1 B1011	4428.0
F9 v1.1 B1012	2395.0
F9 v1.1 B1013	570.0
F9 v1.1 B1014	4159.0
F9 v1.1 B1015	1898.0
F9 v1.1 B1016	4707.0
F9 v1.1 B1017	553.0
F9 v1.1 B1018	1952.0

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Use 'MIN' to find the first date of each landing outcome

Task 5

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

Hint:Use min function

In [30]:
```
%%sql
select MIN("Date") as "First_Record_Date", "Landing_Outcome" from SPACEXTBL
Group by "Landing_Outcome"
Order by "First_Record_Date"
```

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

Out[30]:

First_Record_Date	Landing_Outcome
2010-06-04	Failure (parachute)
2012-05-22	No attempt
2013-09-29	Uncontrolled (ocean)
2014-04-18	Controlled (ocean)
2015-01-10	Failure (drone ship)
2015-06-28	Precluded (drone ship)
2015-12-22	Success (ground pad)
2016-04-08	Success (drone ship)
2018-07-22	Success
2018-12-05	Failure
2019-08-06	No attempt

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- Use 'and' to combine all three conditions in the 'Where' clause

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 [34]:
```
%%sql
select "Booster_Version", "Landing_Outcome", "PAYLOAD_MASS__KG_" from SPACEXTBL
where ("Landing_Outcome" == "Success (drone ship)") and (4000 < "PAYLOAD_MASS__KG_") and (6000 > "PAYLOAD_MASS__K
```

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

Out[34]:

Booster_Version	Landing_Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Use 'count' to calculate the number of each outcome and 'Group by' mission outcome

Task 7

List the total number of successful and failure mission outcomes

In [44]:
```
%%sql
select "Mission_Outcome", count("Date") as "COUNT" from SPACEXTBL
Group by "MIssion_Outcome"
```

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

Out[44]:

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Use a subquery 'select MAX(payload_mass__kg_) from SPACEXTBL' as a condition in 'where' clause.

In [59]:
```
%%sql
select "Booster_Version", "PAYLOAD_MASS__KG_" from SPACEXTBL
where "PAYLOAD_MASS__KG_" = (select MAX(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

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

Out[59]:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Use a subquery in the 'from' clause to filter out the data in year 2015

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 [62]:
```
%%sql
select substr("Date",6,2) as "Month", "Landing_Outcome", "Booster_Version", "Launch_Site"
from (select * from SPACEXTBL where substr("Date",0,5) = '2015')
where "Landing_Outcome" = "Failure (drone ship)"
```

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

Out[62]:

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

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

- 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
- Use subquery in 'from' clause to filter the data between desired time range and use 'order by' to display results in descending order

Task 10

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.

In [65]:

```
%%sql
select "Landing_Outcome", COUNT("Landing_Outcome") as "RANK"
from (select * from SPACEXTBL where ("Date" > '2010-06-04') and ("Date" < '2017-03-20'))
group by "Landing_Outcome"
order by "RANK" DESC
```

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

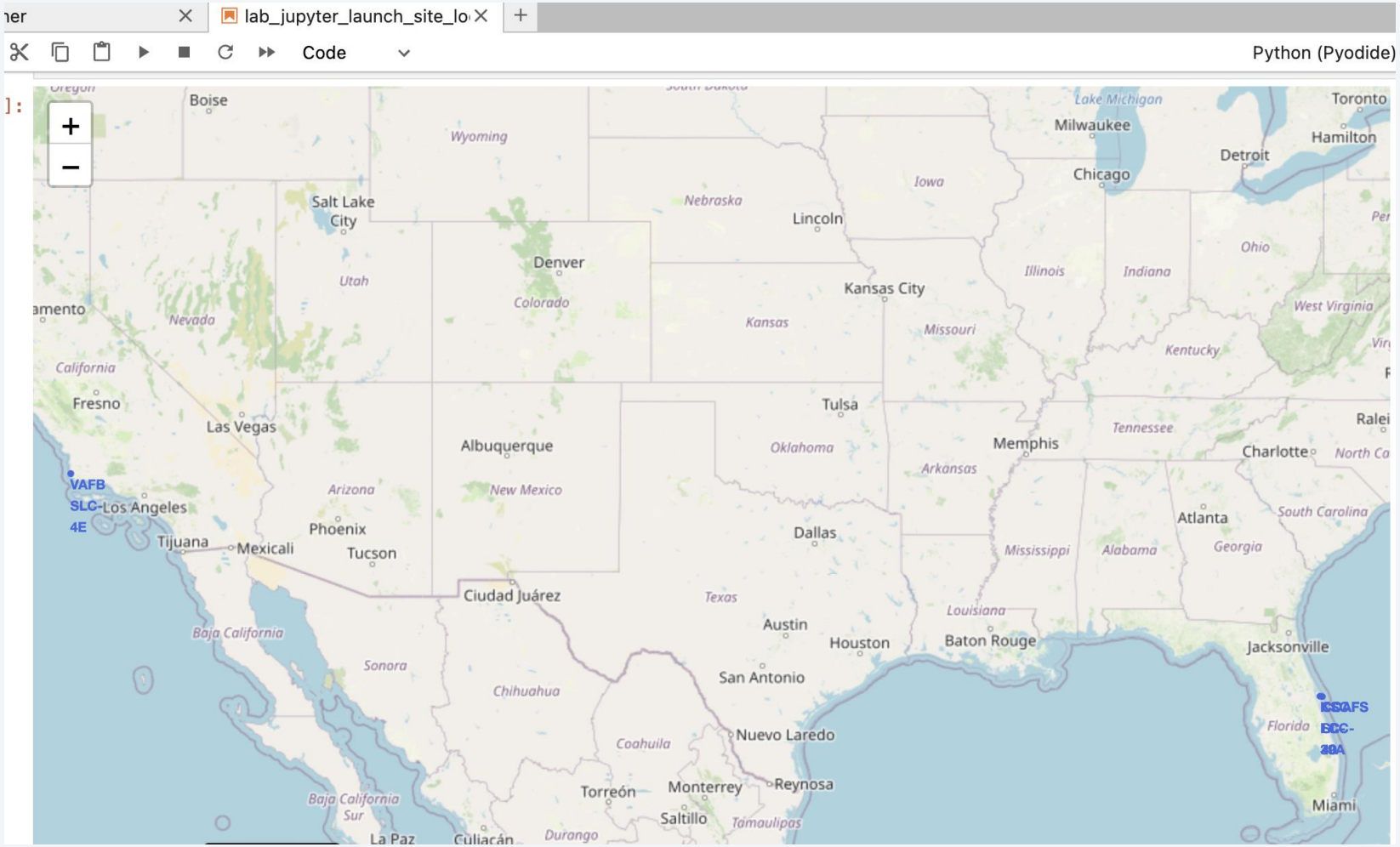
Out[65]:

Landing_Outcome	RANK
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1


Section 3
Launch Sites
Proximities Analysis

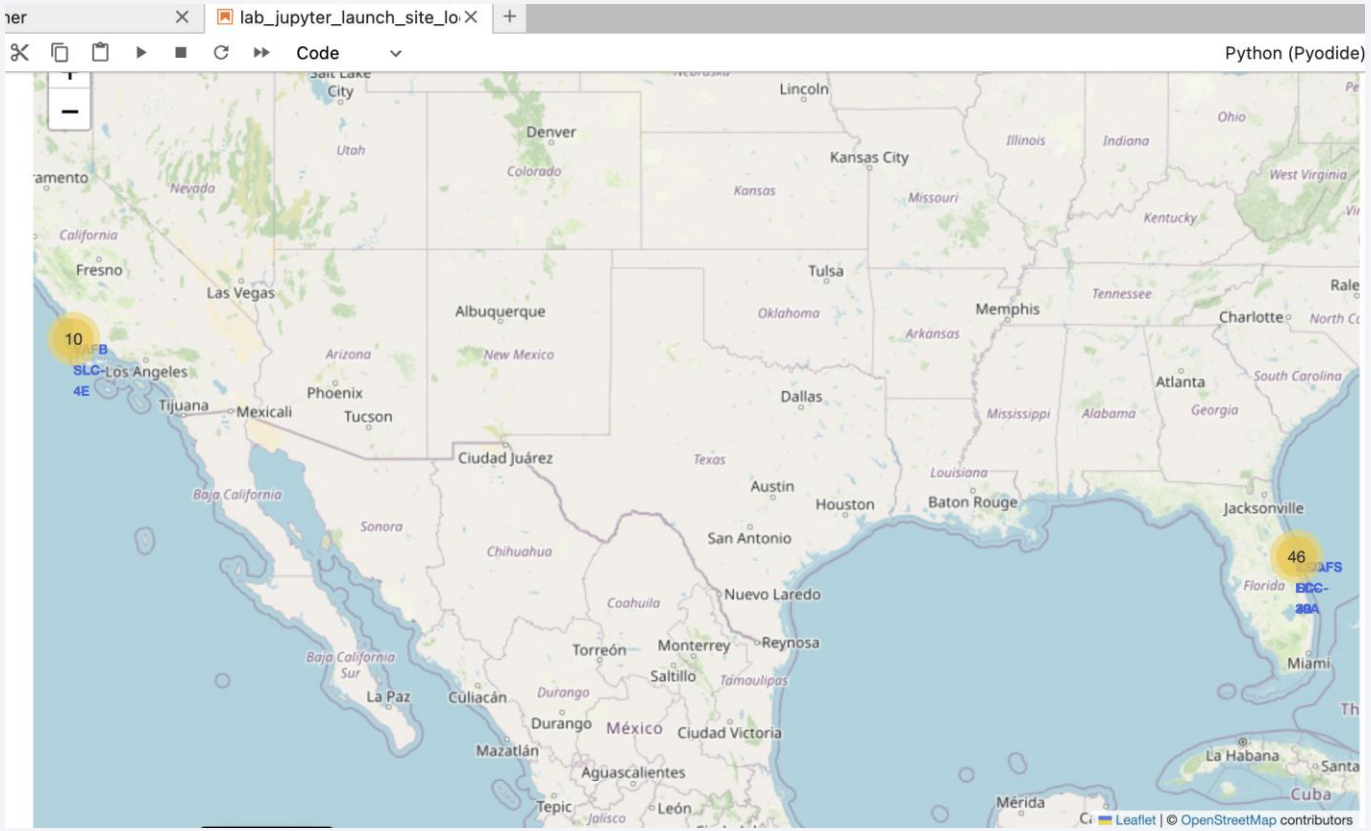
Folium Map – Launch Sites

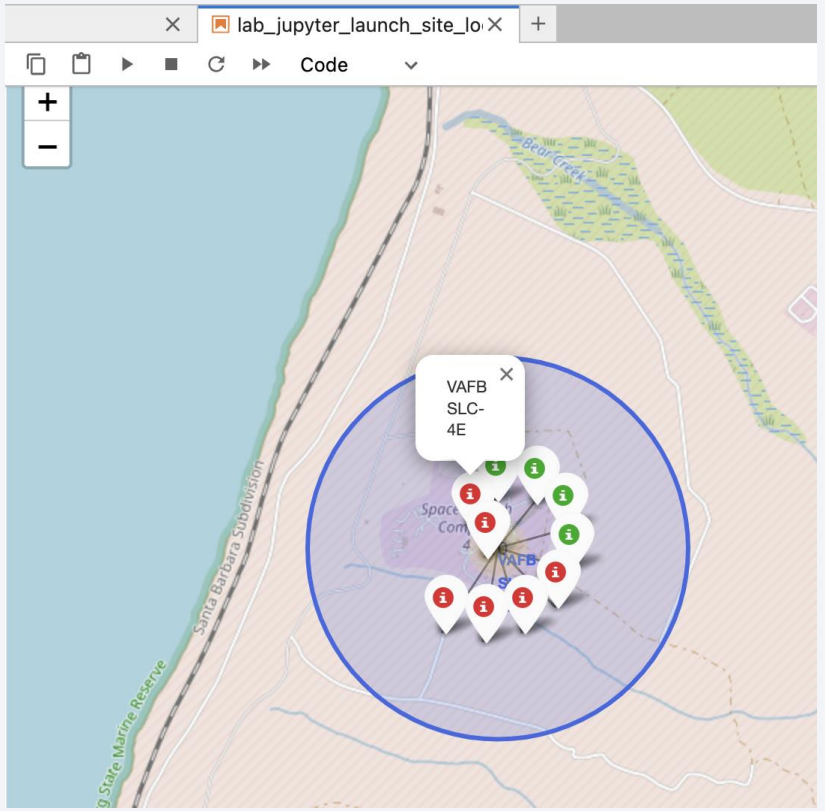
- The launch sites are all near the coasts.



Folium Map – Added successful and failed launches

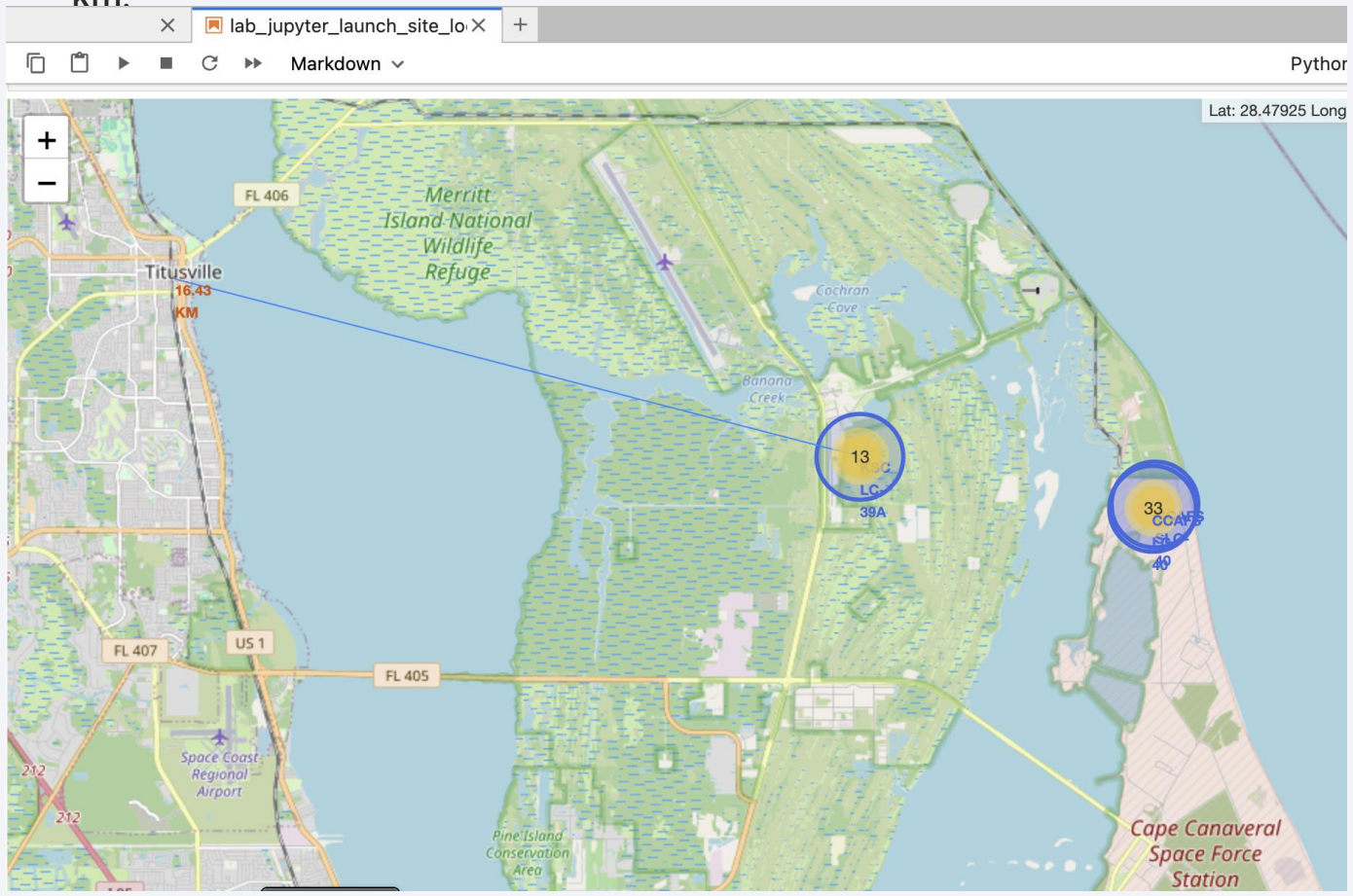
- Adding all launch results in the map with mark clusters, we can see that the east coast has more launches than the west.

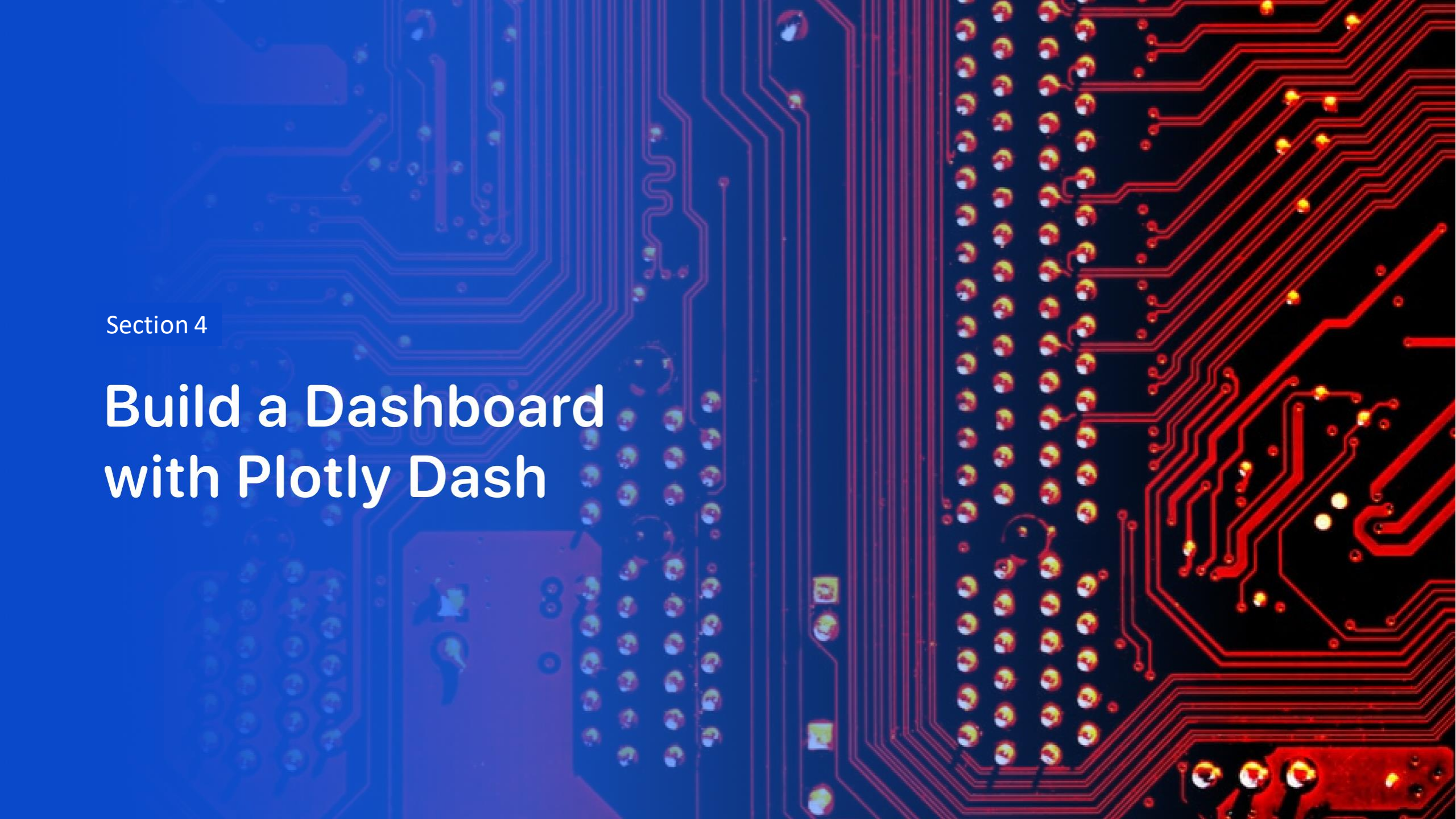




Folium Map – Distance to Proximities

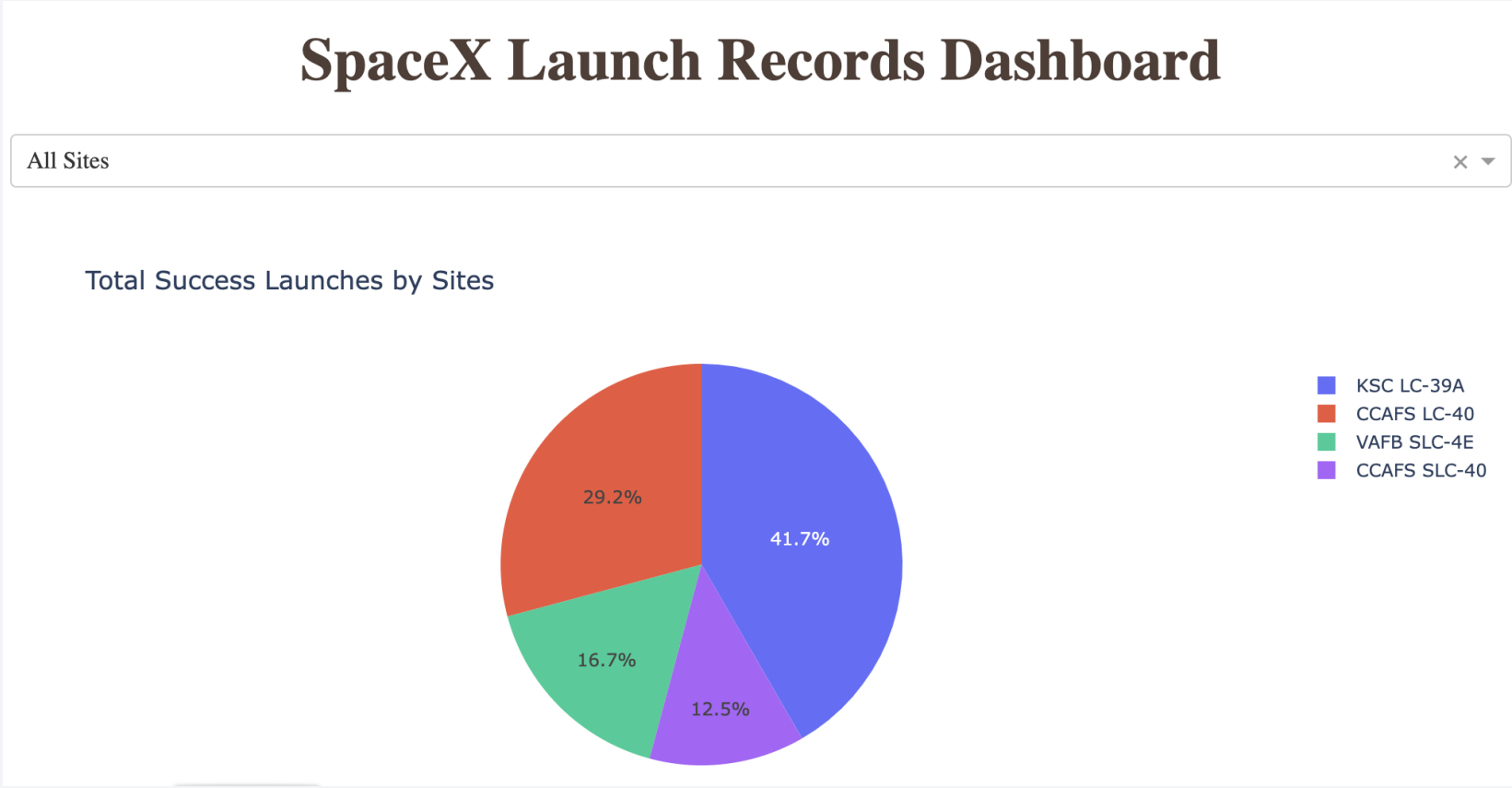
- We can see that the distance of site KSC-SC 39A to the nearest village (Titusville) is 16.43 km.




Section 4
Build a Dashboard with Plotly Dash

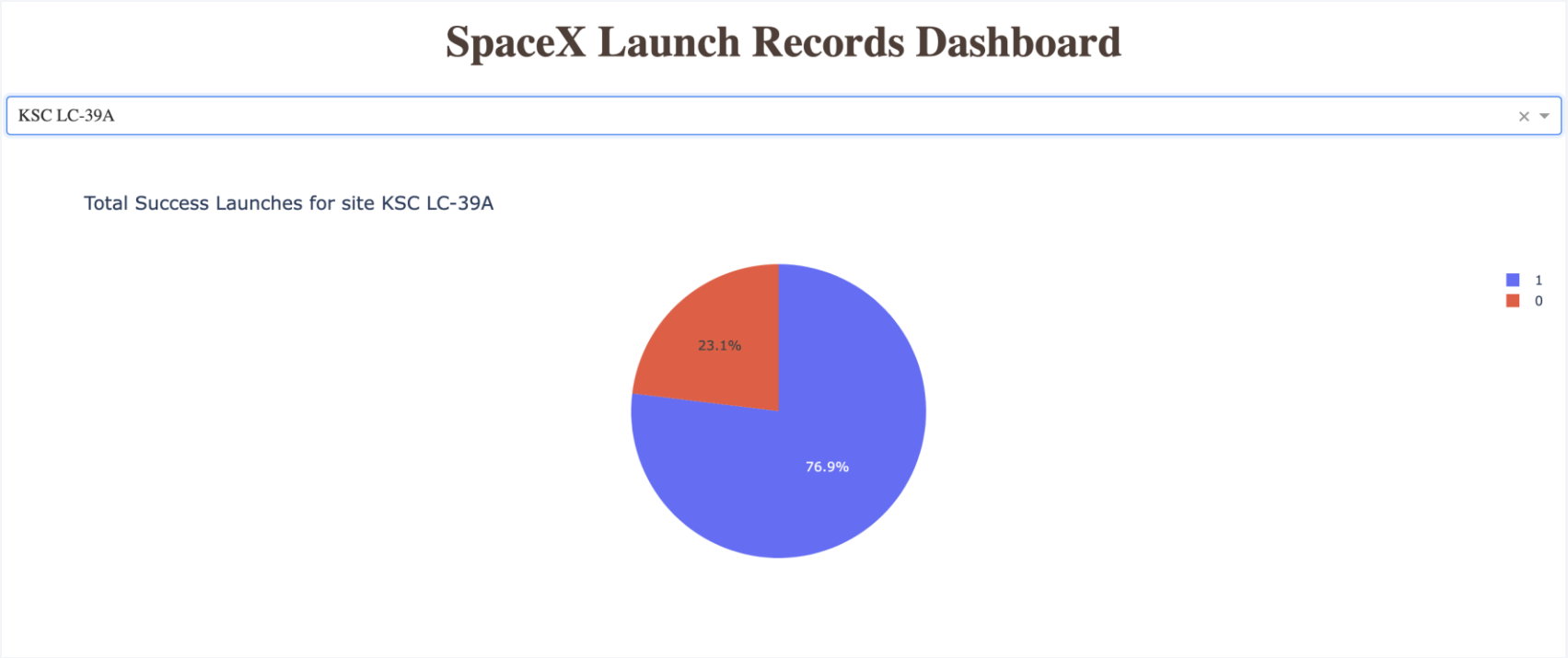
Dashboard – Pie Chart (All Sites)

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



Dashboard – Pie Chart (KSC LC-39A)

- The success rate of launches at KSC LC-39A



Dashboard – Payload vs Booster V. Scatter Chart

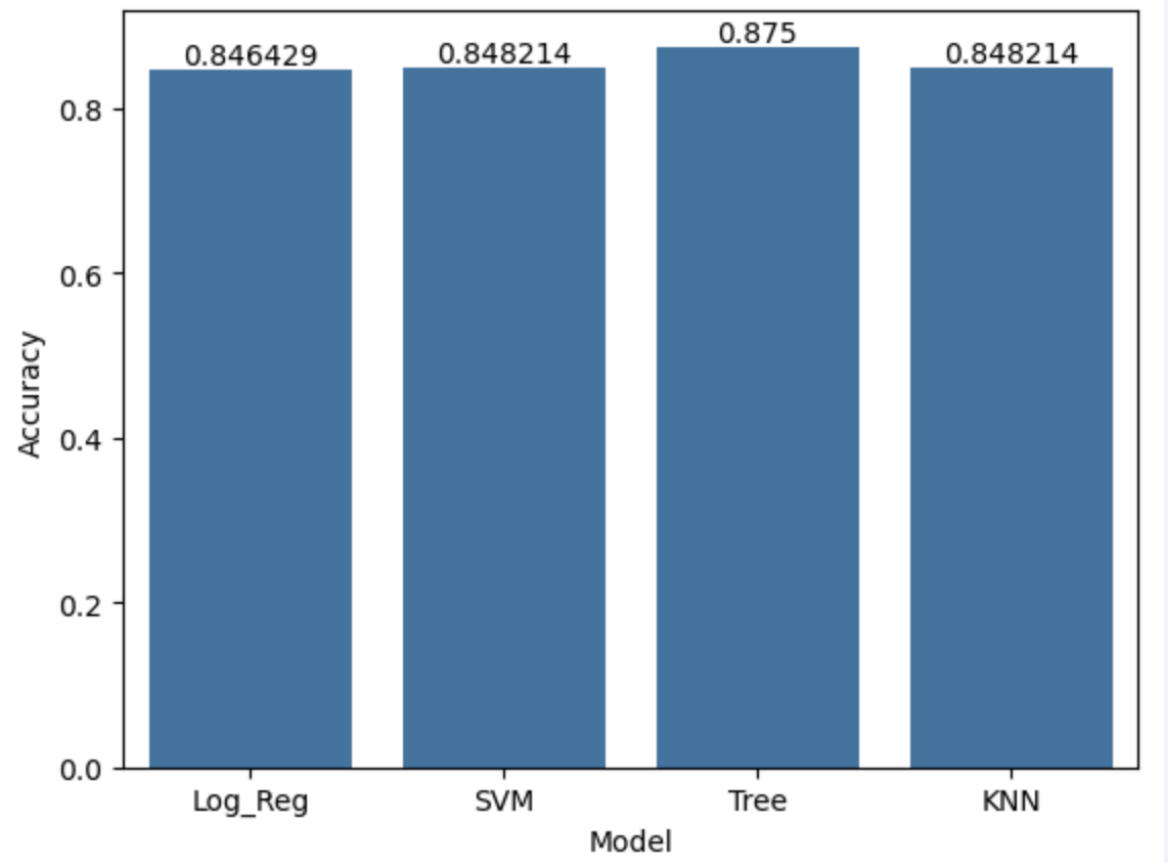
- The booster version of v1.1 has low success rate. So far no successes with payload greater than 2000kg.
- The booster version of FT has high success rate with payload between 2000 and 6000kg.

Section 5

Predictive Analysis (Classification)

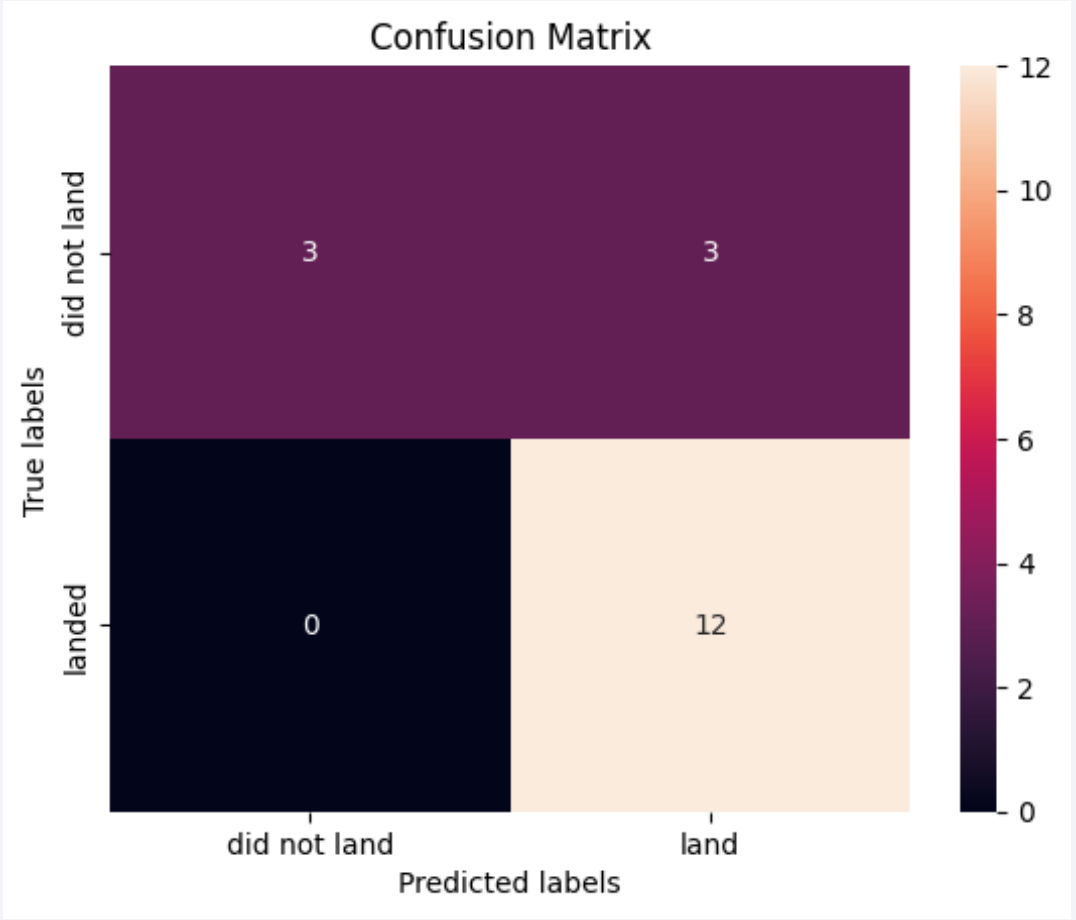
Classification Accuracy

- We can see from the bar plot on the right, the Tree Classification Model has the highest accuracy among all prediction models.



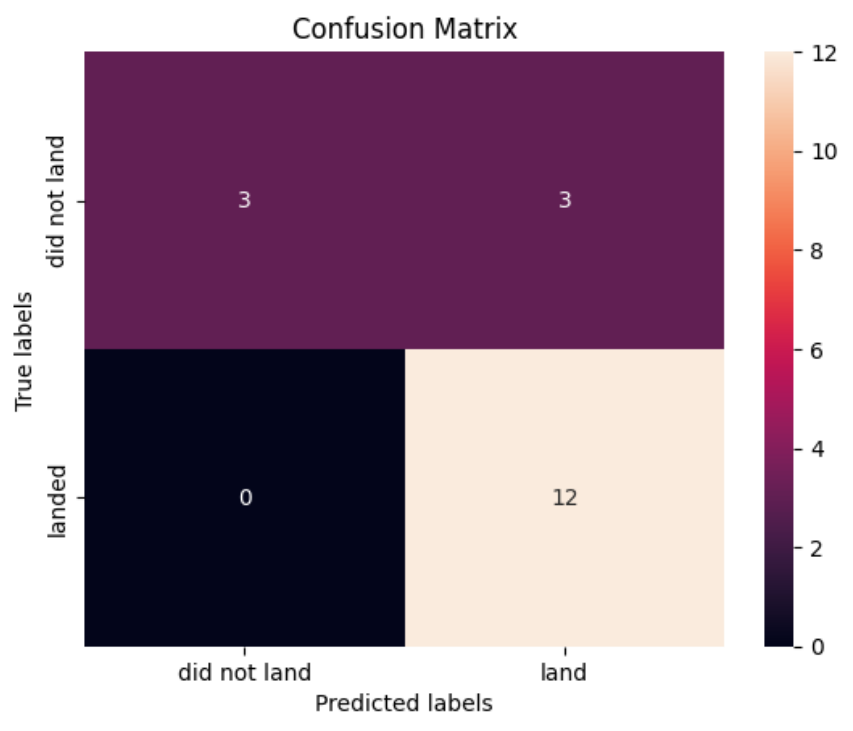
Confusion Matrix

- The four models (logistic regression, SVM, Tree classification, KNN) have the same score on test set, which is 0.833.



Conclusions

- From the EDA, we can see from the scatter plots that the attributes like payload weight, orbit type, and launch site have certain correlations to the mission results, success or fail.
- We can see from the map that all the launching sites are near the coast side.
- After using one hot coding method on the categorical attributes and feed it to the classification models. Among the four models used in this survey, we got Tree Classification model with 0.875 accuracy, and a 0.833 score on test sets.



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!