IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Koray Tarakçı
08/11/2024

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

- Project background and context
- Problems you want to find answers

Section 1

Methodology

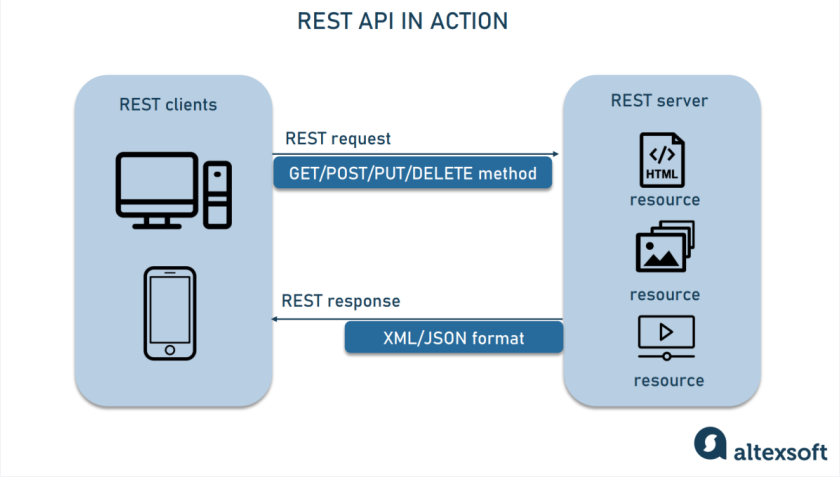
Methodology

Executive Summary

- Data collection methodology:
  - Data supplied via Spacex Rest API
  - Web scraipping applied to Wikipedia tables
- Perform data wrangling
  - Data Filtered
  - Missing values held via using pandas and numpy functions
- 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 skit-learn to apply models. The predictions' fine tuning are ensured via comparison

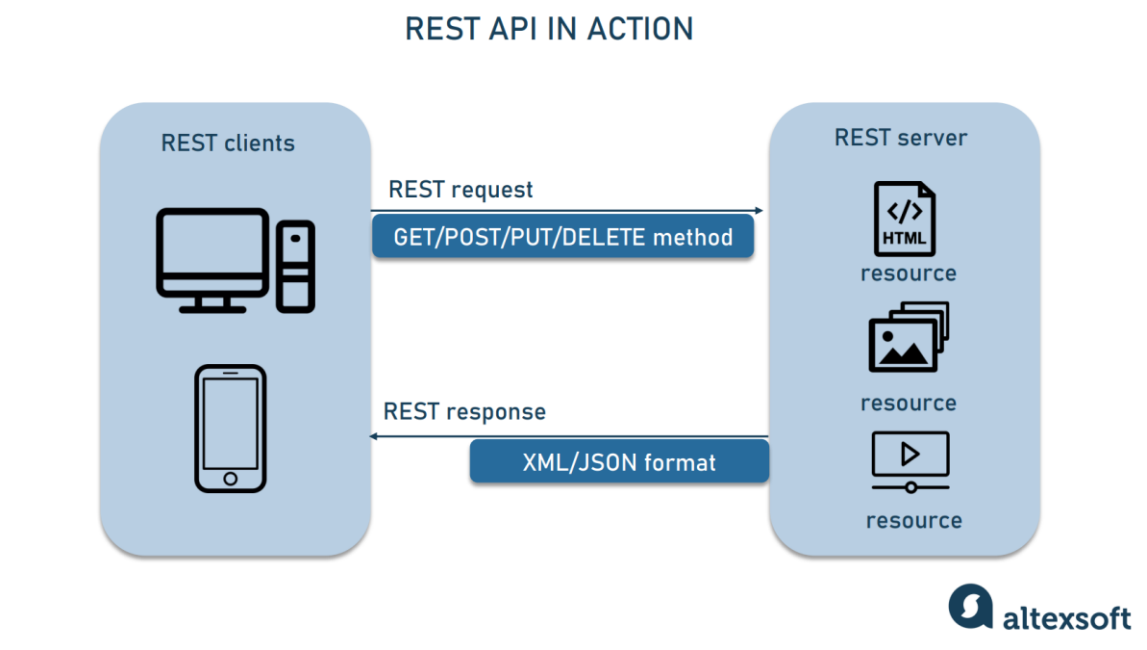
Data Collection

- Describe how data sets were collected.
  - We used SpaceX REST API to contain data over Python. Used requests, pandas, and numpy libraries.
  - Web scrapping tools are applied to Wikipedia tables over Python using BeautifulSoup library.
- You need to present your data collection process use key phrases and flowcharts



Data Collection – SpaceX API

- GitHub URL
- As you can observe via shared link, we have obtained related data with the help of data collection libraries of Python



Data Collection - Scraping

- We have addressed the Wikipedia table via a static url. Then, imported the table by using BeautifulSoup library. Finally, we have reflected the table to a dataframe and export it as a csv file.
- GitHub URL

Data Wrangling

- We have obtained the landing data over a URL.

Then, we have prepared the data for further analysis.

  - Calculate the number of launches on each site
  - Calculated the number and occurrece of each orbit.
  - Calculated the number and occurence of mission outcome of the orbits.
  - Created a landing outcome label from Outcome column and added to the dataframe.

- GitHub URL

EDA with Data Visualization

- We have used listed charts below, in order to examine the relationship between parameters;
  - Dot plot
  - Bar plot
  - Line plot
- GitHub URL

EDA with SQL

- By using SQL, we have analyzed Spacex data for obtain the features listed ;
  - To obtain unique launch site names,
  - To display the total payload mass carried by boosters launched by NASA (CRS),
  - To display average payload mass carried by booster version F9 v1.1,
  - To list the date when the first succesful landing outcome in ground pad was achieved,
  - To list the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000,
  - To list the total number of successful and failure mission outcomes,
  - To list the names of the booster_versions which have carried the maximum payload mass. Use a subquery,
  - To 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.
  - To 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.

- GitHub URL

Build an Interactive Map with Folium

- We have used Folium library and its features listed below;
  - Circle
  - Marker
  - Mouse position
  - PolyLine
- We have added those objects in order to positioned the launch site and examine the relationship between of its success and environmental conditions.
- GitHub URL

Build a Dashboard with Plotly Dash

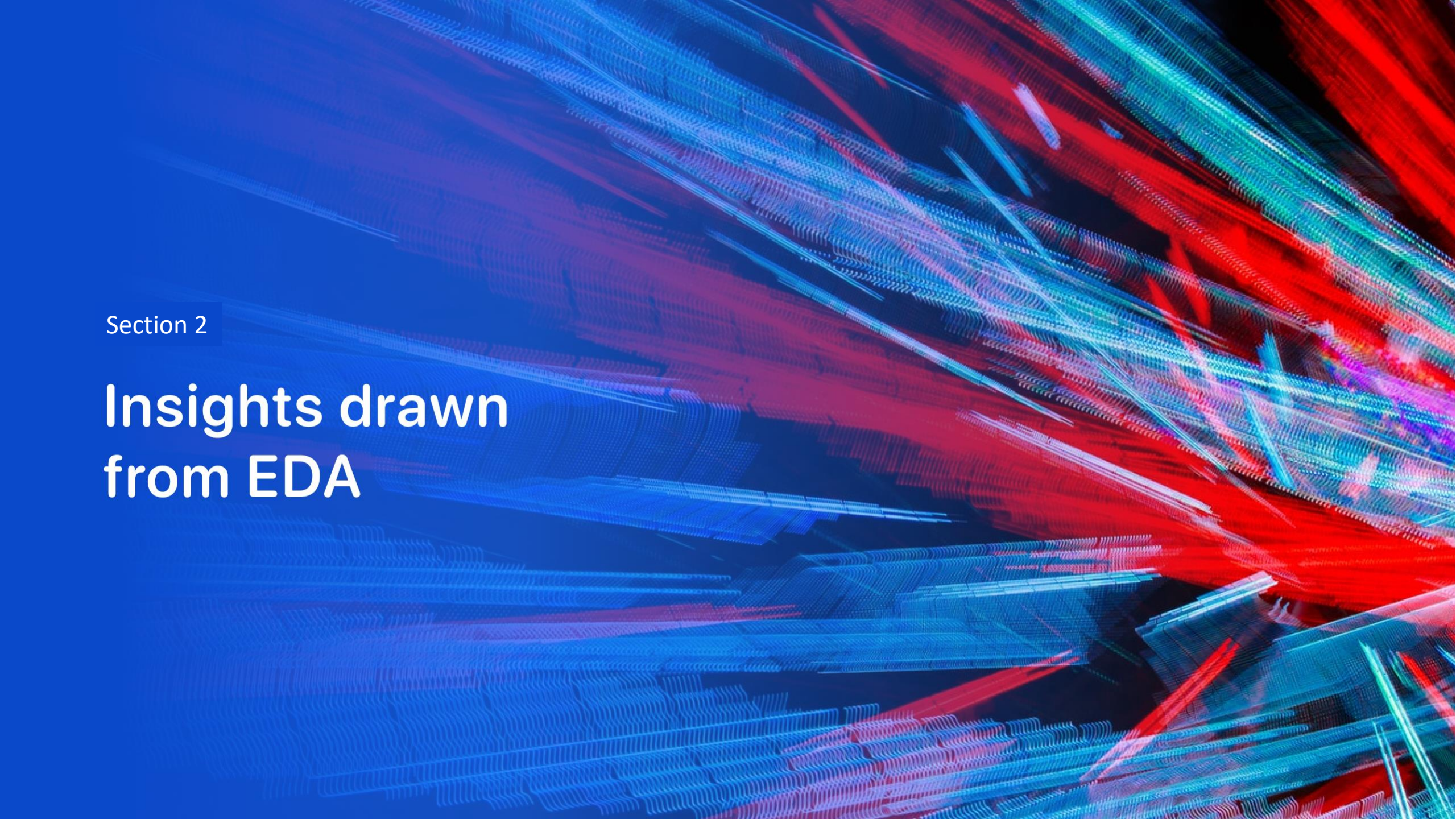
- We haved used ploty dash listed features;
  - Pie Chart,
  - Scatter Plot,
  - Dropdown
  - RangeSlider
- We have used these feautures in order to dynamically feed user's curiosity, and create an environment to make sure the data easily analyzed by the users.
- GitHub URL

Predictive Analysis (Classification)

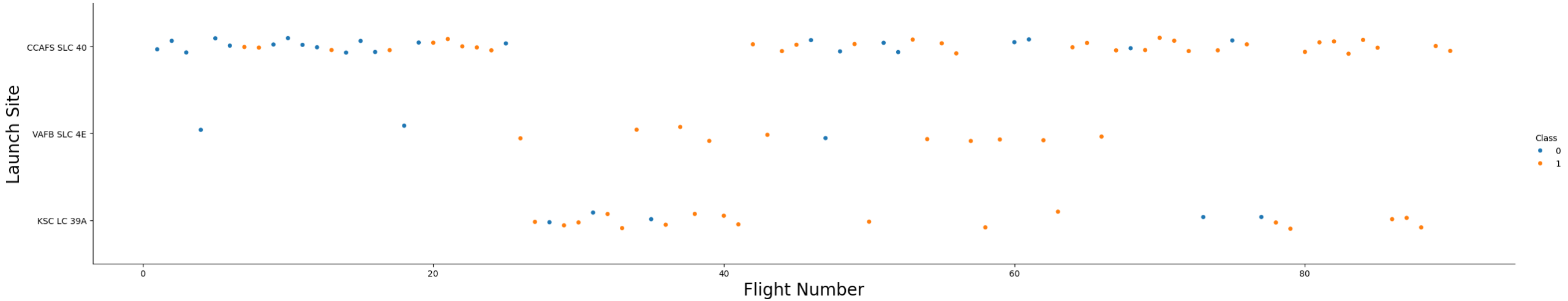
- We have used skit-learn library in order to apply machine learning models, which are;
  - Logistic Regression,
  - KNN
  - SVM
  - Decision Tree
  - Grid Search
  - Etc..
- GitHub URL

Results

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

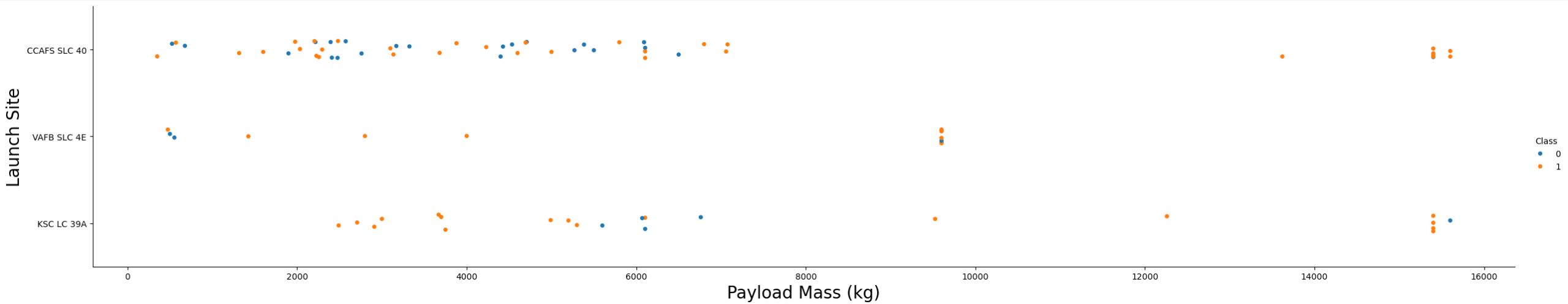

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



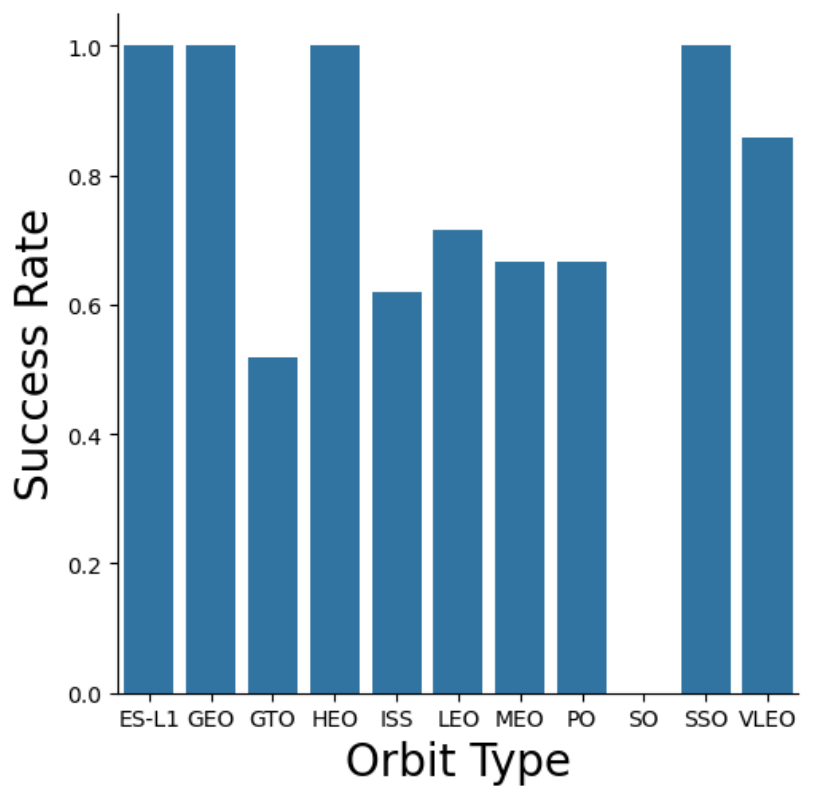
- While the flight numbers are increase, success rate increases as well

Payload vs. Launch Site



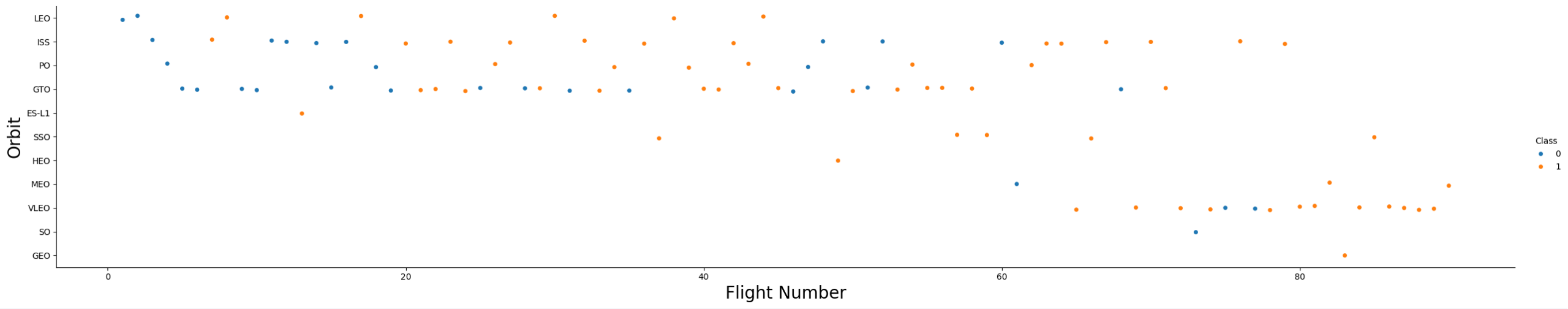
- For launch site CCAF SLC 40, success rate is lower especially for payload mass between 2000 and 5500

Success Rate vs. Orbit Type



- Most successful orbit types;
  - ES-L1
  - GEO
  - HEO
  - SSO

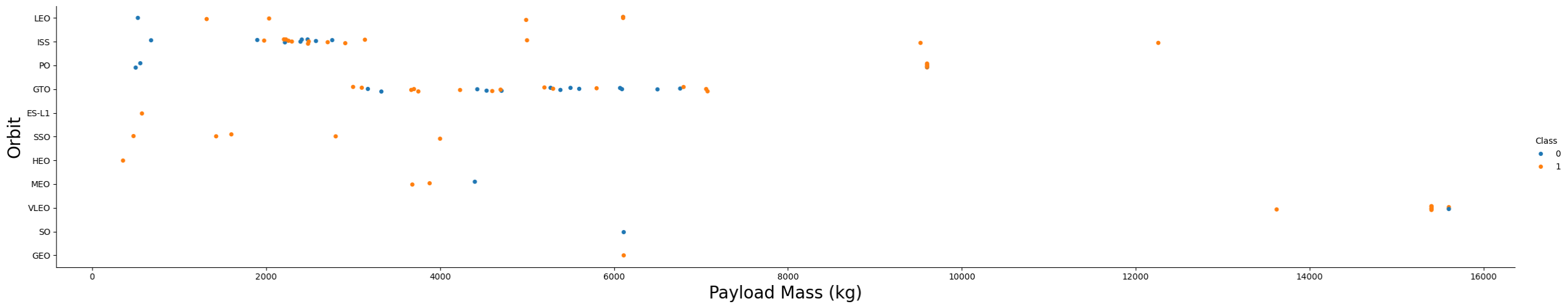
Flight Number vs. Orbit Type



- Show a scatter point of Flight number vs. Orbit type

- Show the screenshot of the scatter plot with explanations

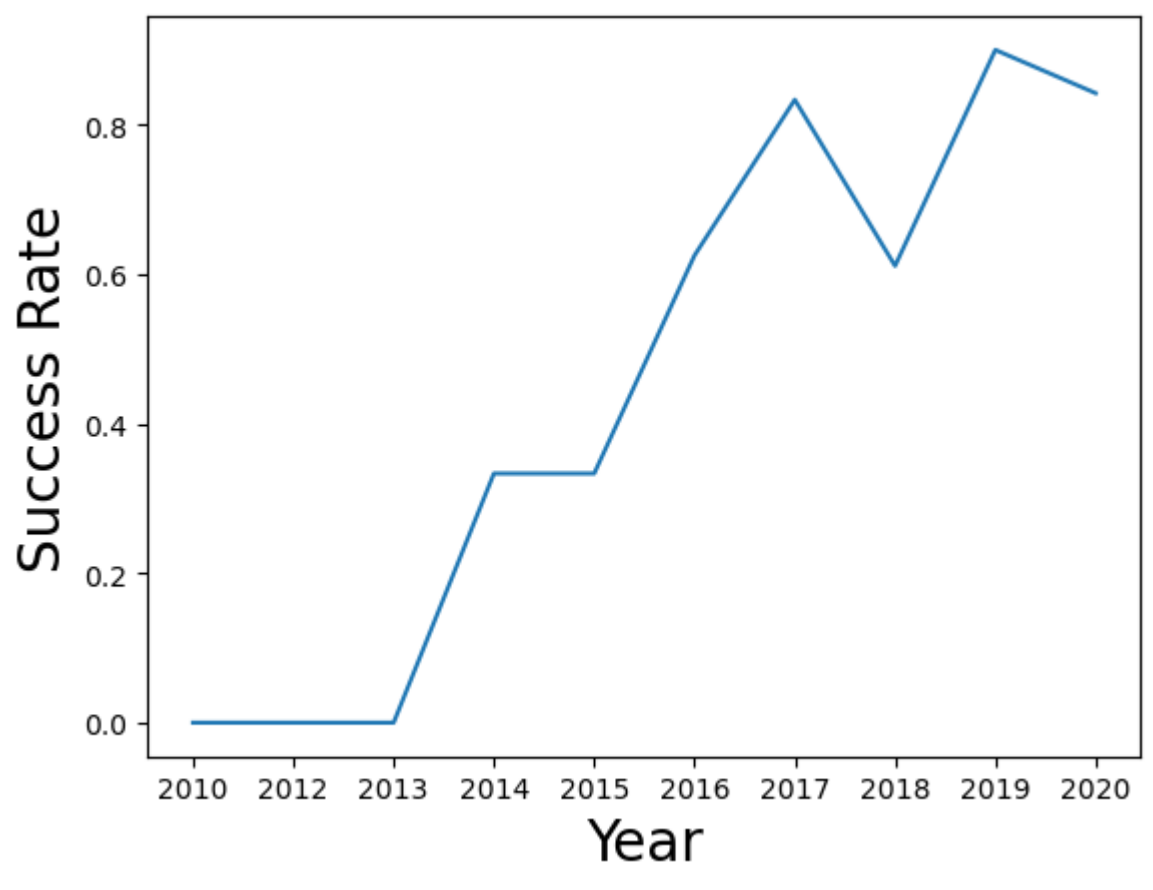
Payload vs. Orbit Type



- For ISS and GTO orbit types, there have been various attempts that shows high variance.

Launch Success Yearly Trend

- Over years of attempt, the success rate seems to increase.



All Launch Site Names

We have choosen launch sites as distinct values.

```
%sql select distinct Launch_Site from SPACEXTABLE
```

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

Launch Site Names 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

```
%sql select * from SPACEXTABLE where Launch_Site like 'CCA%' LIMIT 5
```

- By using LIKE feature we managed to obtain launch sites starts with CCA, then, limited the query with 5 values.

Total Payload Mass

- Our query directly sums payload mass while LIKE feature bounded the data by choosing lines that contains NASA sequence in Customer column.

```
%sql select SUM(PAYLOAD_MASS__KG_) from SPACEXTABLE WHERE Customer LIKE '%NASA%'
```

SUM(PAYLOAD_MASS__KG_)
107010

Average Payload Mass by F9 v1.1

- We have took the average of payload mass while bounded the data lines with LIKE feature which starts with F9 v1.1 in Booster Version column.

```
%sql select AVG(PAYLOAD_MASS__KG_) from SPACEXTABLE WHERE Booster_Version LIKE 'F9 v1.1%'
```

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

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- Instead of using min feature, we have used order by in order to obtain minumum date of successful landing.

```
%sql select * from SPACEXTABLE WHERE Landing_Outcome LIKE '%Success%' ORDER BY Date ASC LIMIT 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2015-12-22	1:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)
2016-04-08	20:43:00	F9 FT B1021.1	CCAFS LC-40	SpaceX CRS-8	3136	LEO (ISS)	NASA (CRS)	Success	Success (drone ship)
2016-05-06	5:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-05-27	21:39:00	F9 FT B1023.1	CCAFS LC-40	Thaicom 8	3100	GTO	Thaicom	Success	Success (drone ship)
2016-07-18	4:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select distinct Booster_Version from SPACEXTABLE WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 and upper(Landing_outcor
```

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

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

- To obtain the info, we have used BETWEEN function that bounds payload mass.

Total Number of Successful and Failure Mission Outcomes

```
%sql select distinct Mission_Outcome, Count(*) from SPACEXTABLE GROUP BY Mission_Outcome
```

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

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

- The query lists the unique misson outcome states and due to this grouping, it counts the number of records for the related mission outcome

Boosters Carried Maximum Payload

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

```
%sql select booster_version from SPACEXTABLE where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTABLE);
```

```
 * 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

- We have obtained the list of boosters that carries maximum payload mass by equilizing related payload mass to maximum value with a subquery

2015 Launch Records

```
%sql select substr(Date, 6,2) as month, substr(Date,0,5), date, booster_version, launch_site, Landing_Outcome from SPACEXTAB
```

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

month	substr(Date,0,5)	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- To import this date, we have crop the part of Date column that relates year info and used it for filtering at WHERE part

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.

```
%%sql select Landing_Outcome, count(*) as count_outcomes from SPACEXTABLE
      where date between '2010-06-04' and '2017-03-20'
      group by Landing_Outcome
      order by count_outcomes desc;
```

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

Landing_Outcome	count_outcomes
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

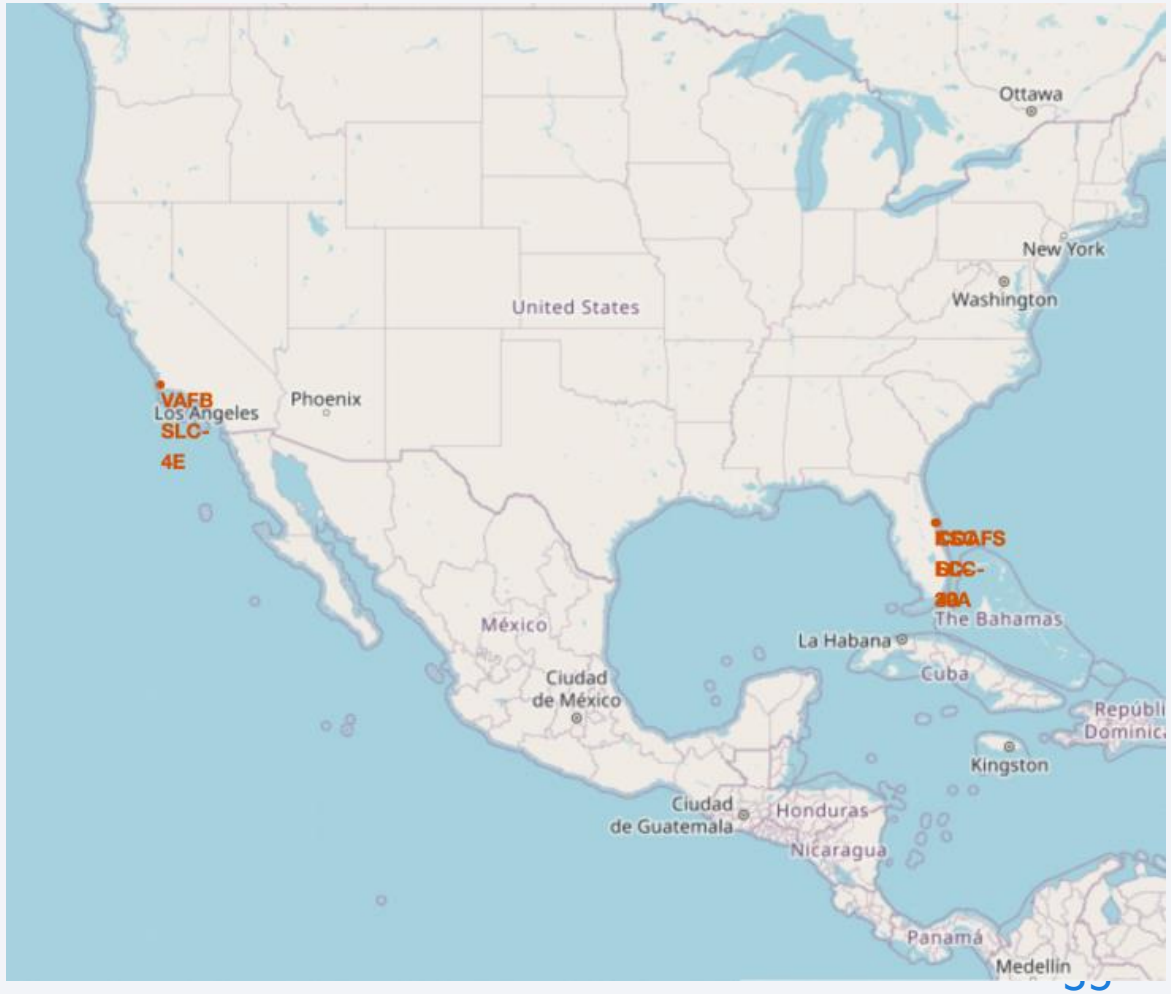
- Between dates of 2010-06-04 and 2017-03-20, the biggest proportion of attemps are not held. For the rest of the landing efforts, the success of the landing outcome is nearly %50

Section 3

Launch Sites Proximities Analysis

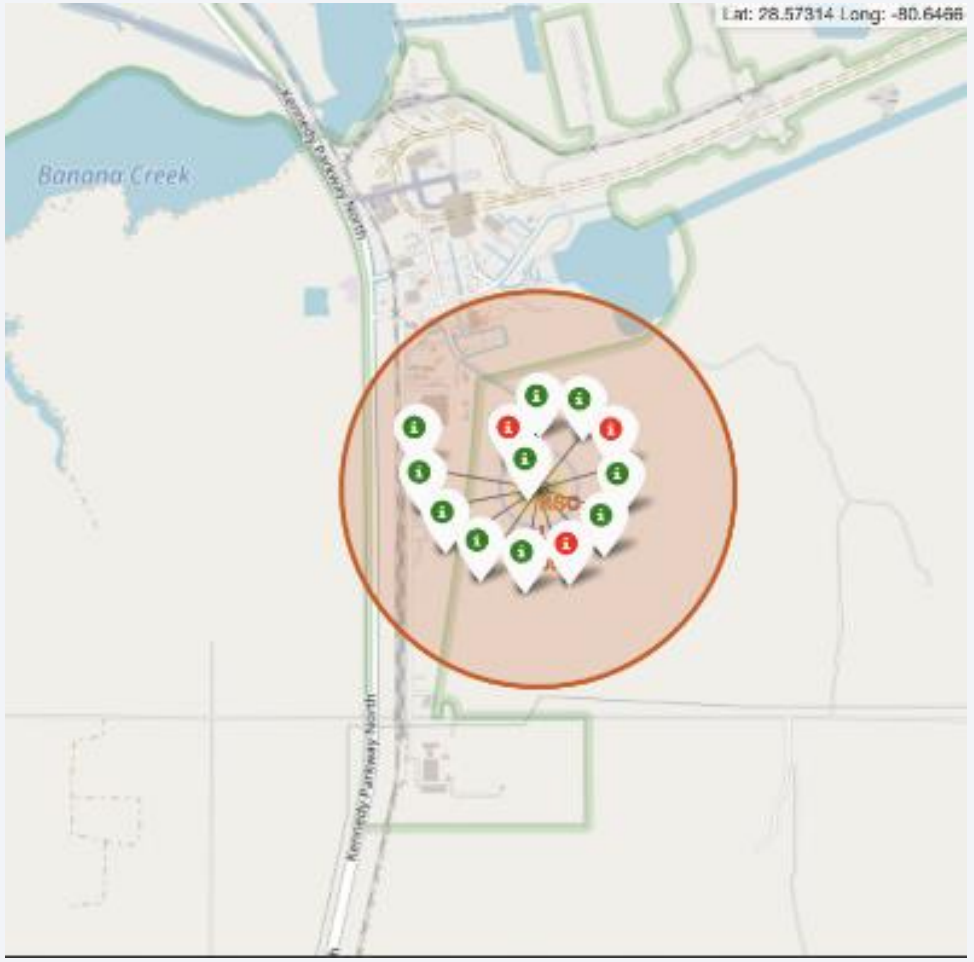
Map of All Launch Sites

- All launch sites are near Ecuador line and placed closer to the coasts.



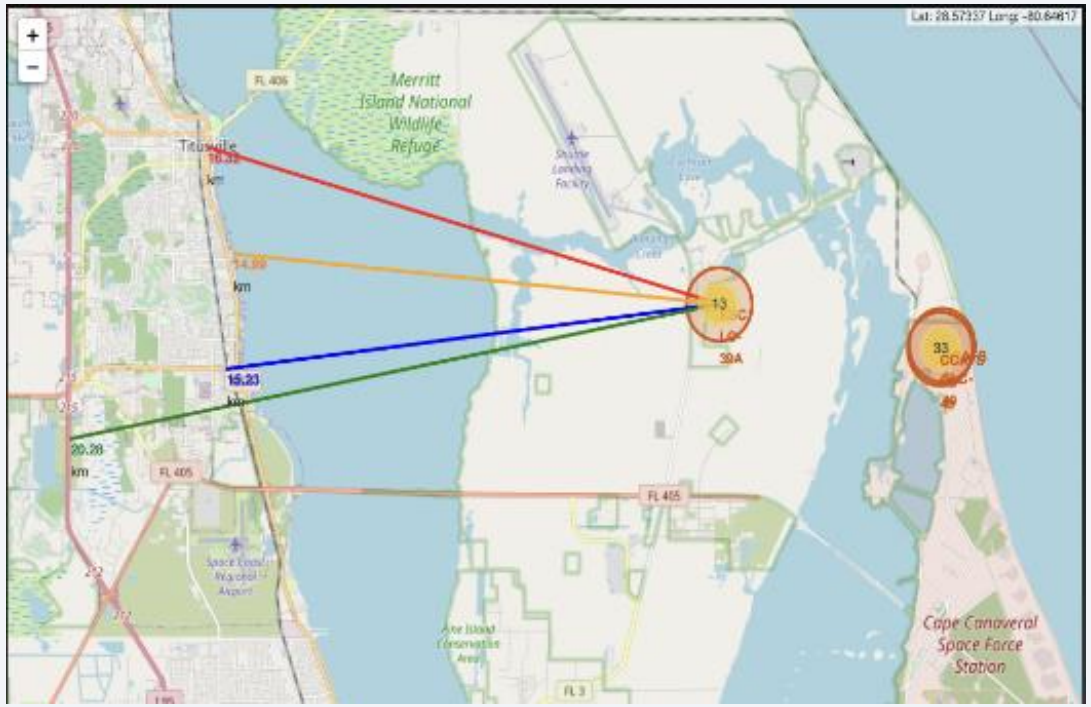
An Example of a Launch Attempt Visulization

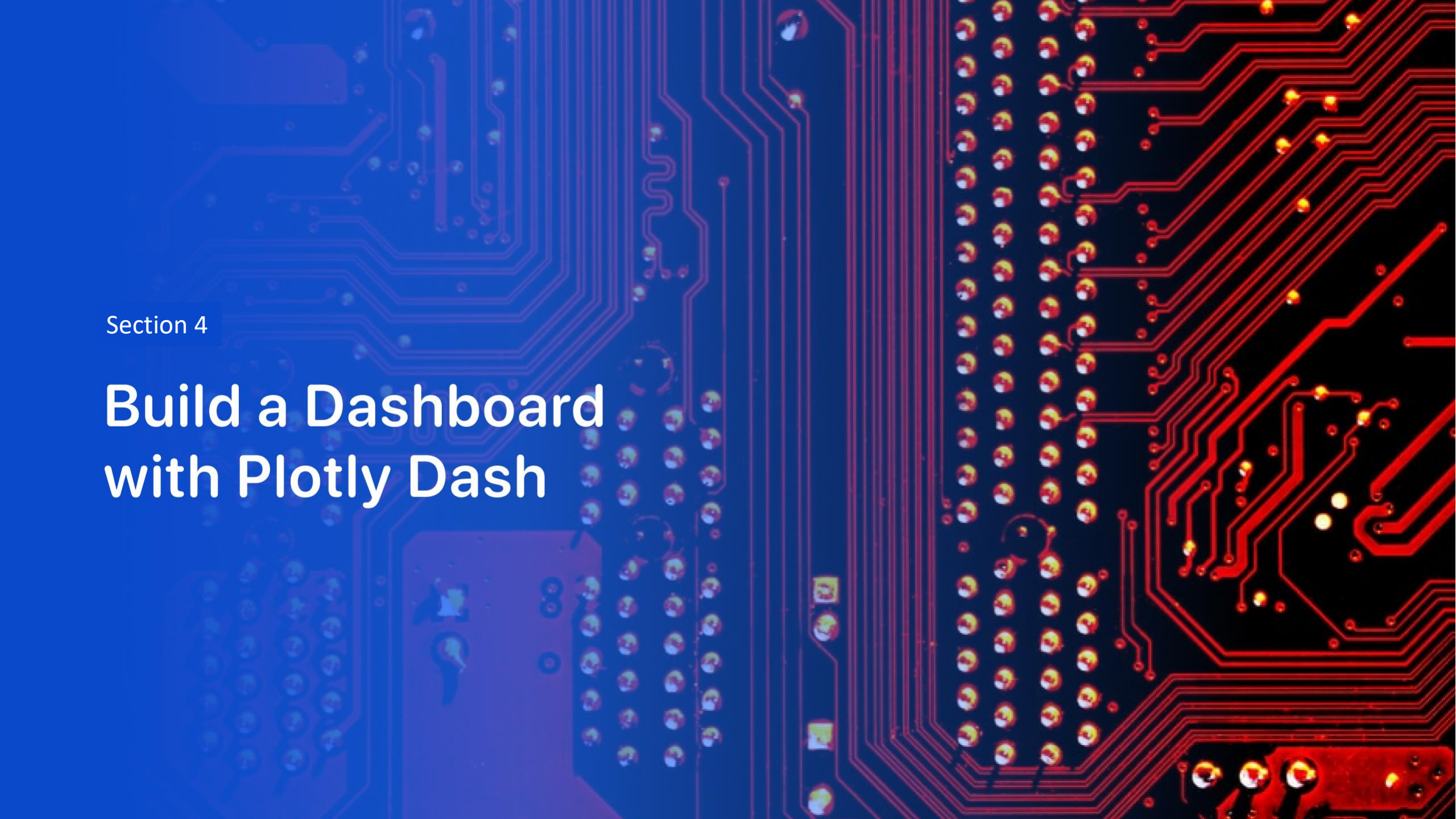
- As it can be observed through map, you may see the launch records of a specific launch pad.
- In this visualization, green and red markers represent the output of related launch attemp which are success and failure respectively.



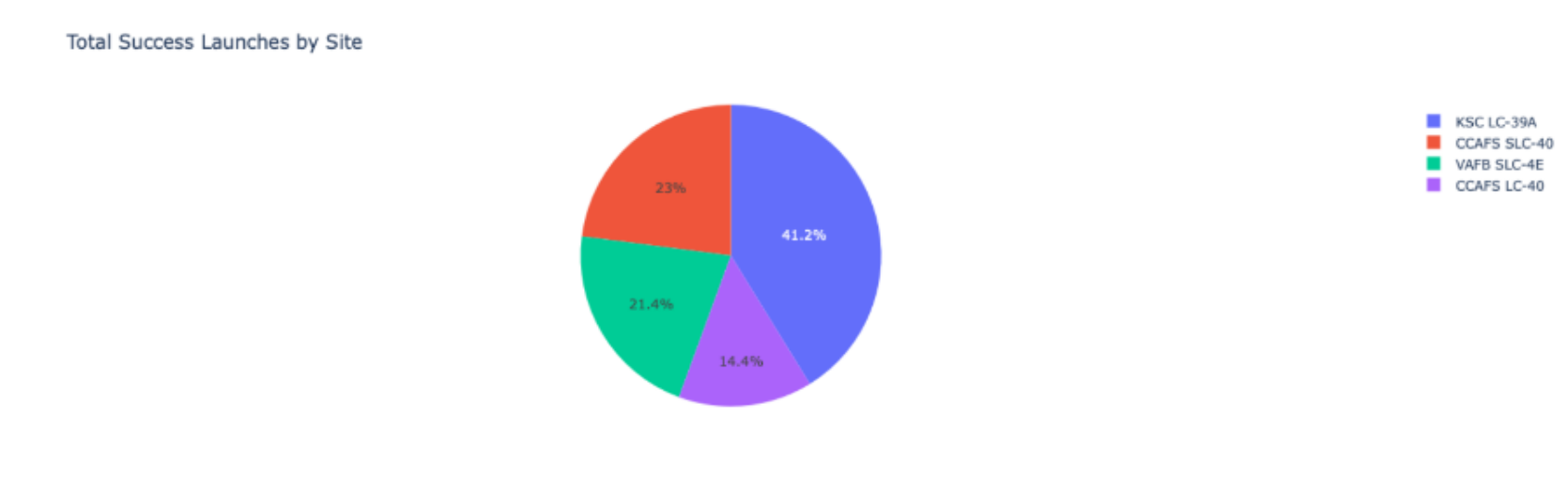
Distance Measurement Between Launch Pads and Proximities

- Folium's marker and mouseposition features have been used to measure between launch sites and proximities
- Which shows how far away the launch pads are placed subject to these proximities. Therefore, such an act ensures the safety of people and valuables in case of catastrophic outputs.



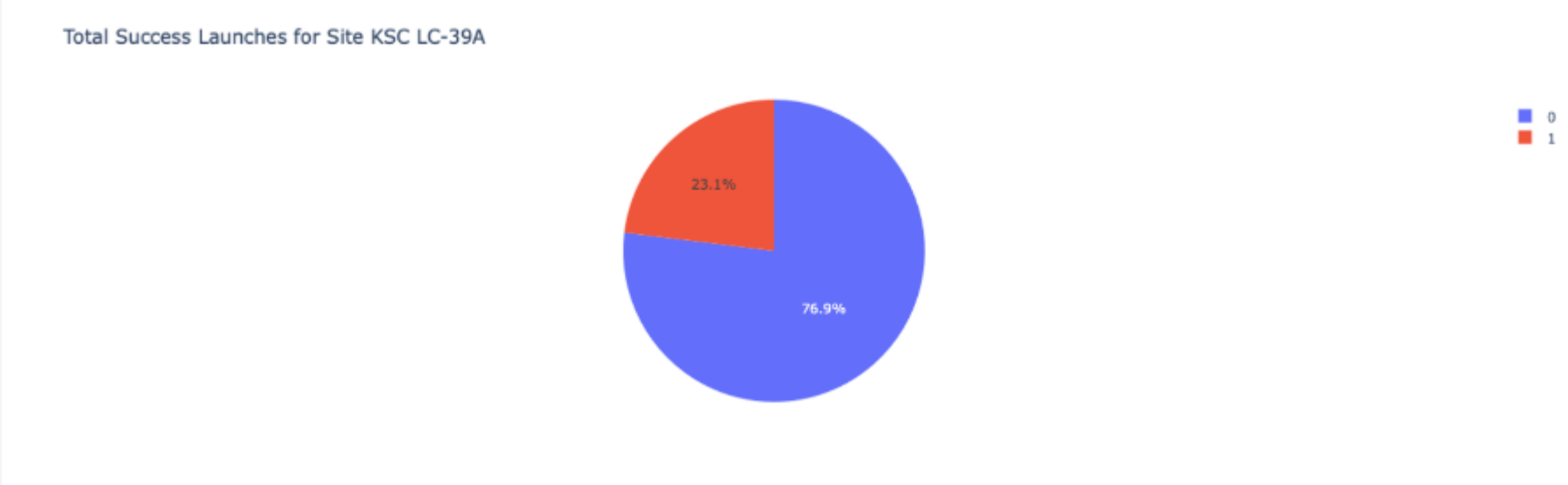

Section 4
Build a Dashboard with Plotly Dash

Launch Site Success Proportions



- While most succesful attempts are held in KSC LC-39A, VAFB SLC-4E has the least success rate.

Launch Site Success and Failure Ratio



- KSC LC-39A Launch Pad has a high success ratio.

Payload and Success Comparison DBs

- These DBs show that highest success ratio for payload mass around between 2000 and 5000 Kg.

Section 5

Predictive Analysis (Classification)

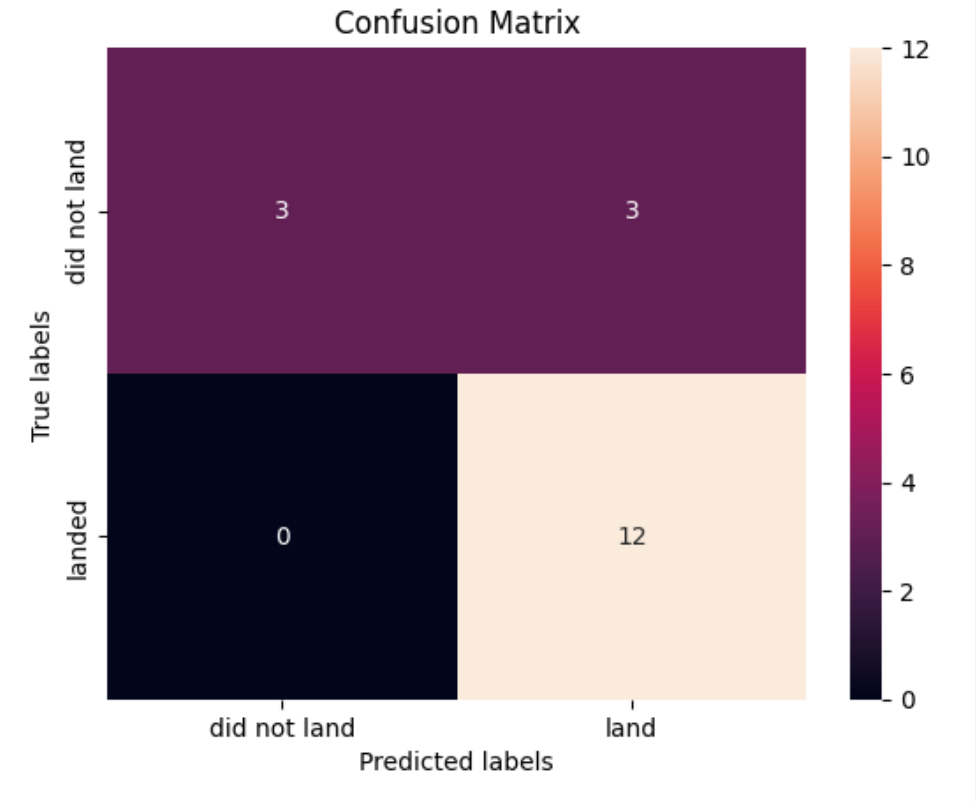
Classification Accuracy

- Due to table shown at the side, highest accuracy rate is SVM, which has best outputs for each score type.

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.830986	0.819444
F1_Score	0.909091	0.916031	0.907692	0.900763
Accuracy	0.866667	0.877778	0.866667	0.855556

Confusion Matrix

- Based on the matrix, you may observe that the model has a high accuracy however, it also does False Positive classifications.



Conclusions

- SVM method is the best option for this dataset
- Around 2000 and 5500 payload mass is good for successful landings.
- Success rate increased over time

Thank you!