Applied Data Science Capstone

NIHAL. S. KUMAR

05-06-2024

IBM Developer

SKILLS NETWORK

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion
- Appendix

EXECUTIVE SUMMARY

- Methodology:
  - Data collection
  - Data Wrangling
  - EDA with SQL
  - EDA with Data Visualization
  - Interactive maps with Folium
  - Dashboard with Plotly Dash
  - Predictive Analysis
- Results:
  - EDA
  - Interactive analytics
  - Predictive Analytics

INTRODUCTION

- 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.

Questions to be ansered:

- What are the factors that can affect the landing?
- How do these factors interact with the each other to determine the landing result?
- What is the best way to predict the landing result?

Methodology

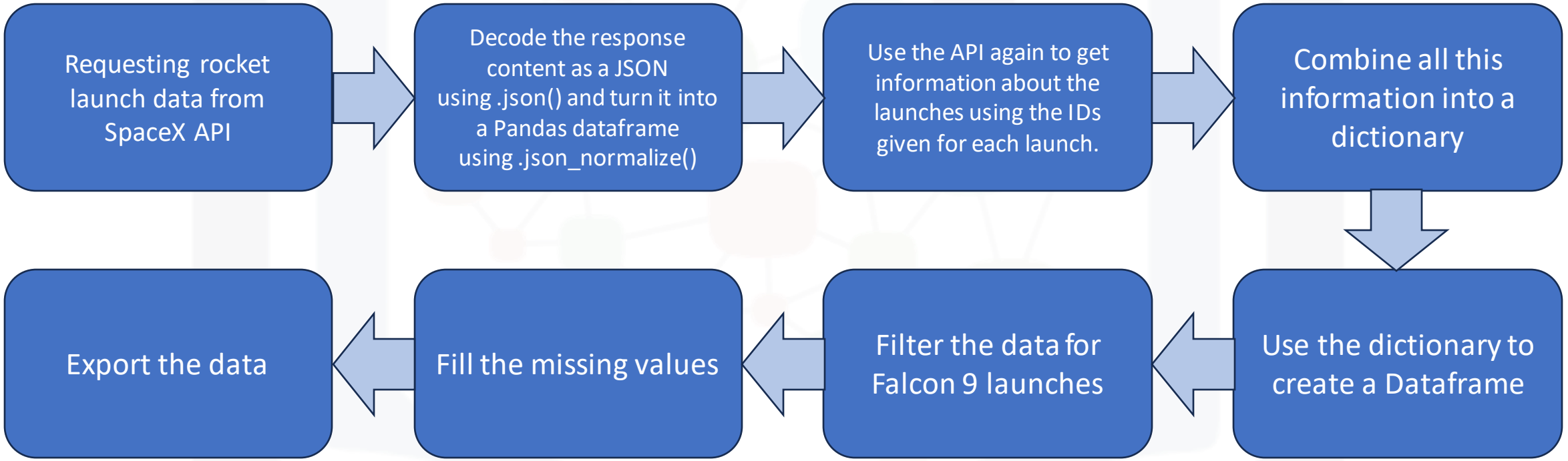
IBM Developer

SKILLS NETWORK

METHODOLOGY

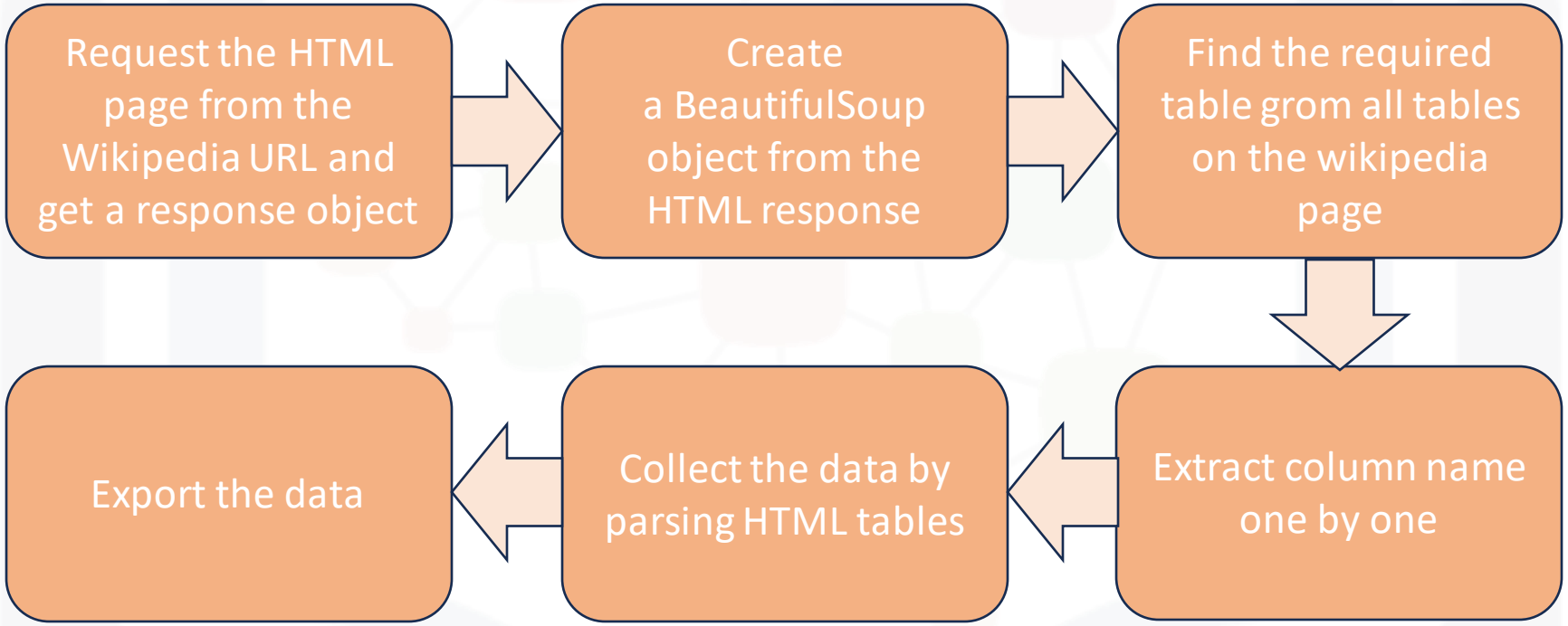
- Data Collection:
  - With SpaceX Rest API
  - Web Scraping from Wikipedia
- Data Wrangling:
  - Filling in missing entries
  - One hot encoding
- Exploratory Data Analysis using SQL and Data Visualization.
- Interactive Visual Analytics:
  - Interactive maps with Folium
  - Interactive dashboard with Plotly Dash
- Predictive analytics using Classification methods.

Data Collection with SpaceX API



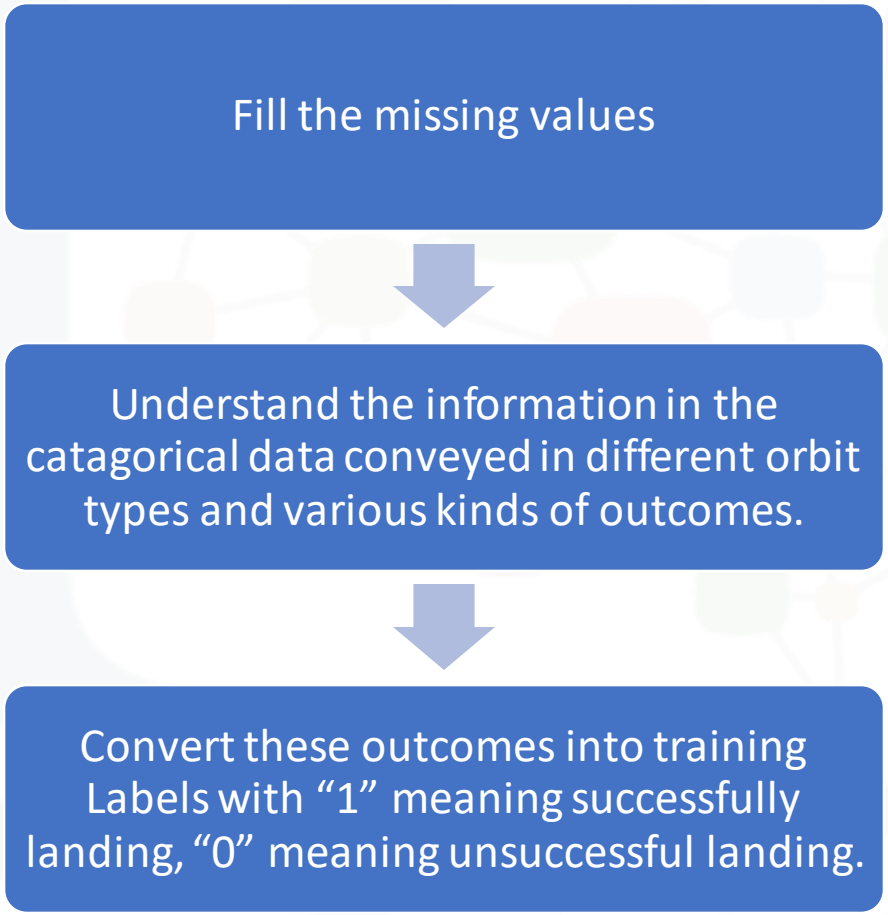
GitHub URL

Data Collection through Web Scraping



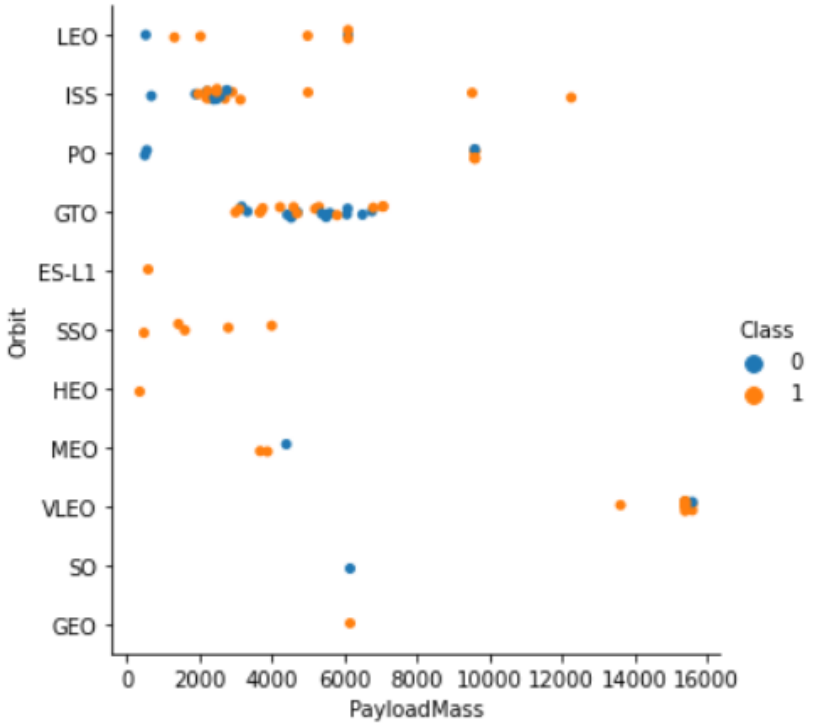
GitHub URL

Data Wrangling

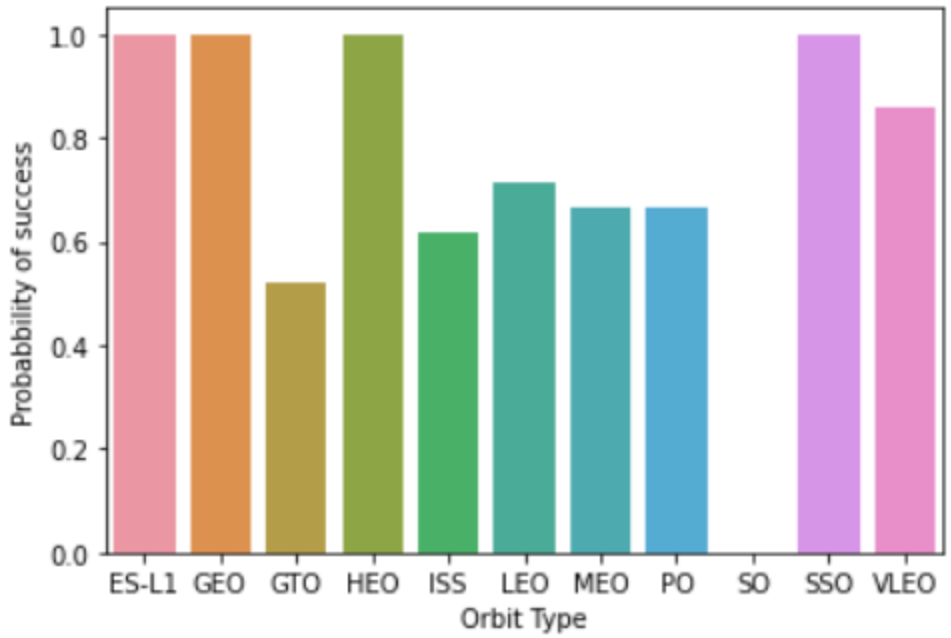


GitHub URL

EDA with Data Visualization



Visualize the relation between flight number and launch Site, payload and launch site, success rate of each orbit type, flight number and orbit type, the launch success yearly trend



GitHub URL: Data Visualization

EDA with SQL

- Apply SQL to get insight from the data.
- Some of the queries:
  - List the names of the booster_versions which have carried the maximum payload mass
  - 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
  - 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 failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
  - Ranking 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

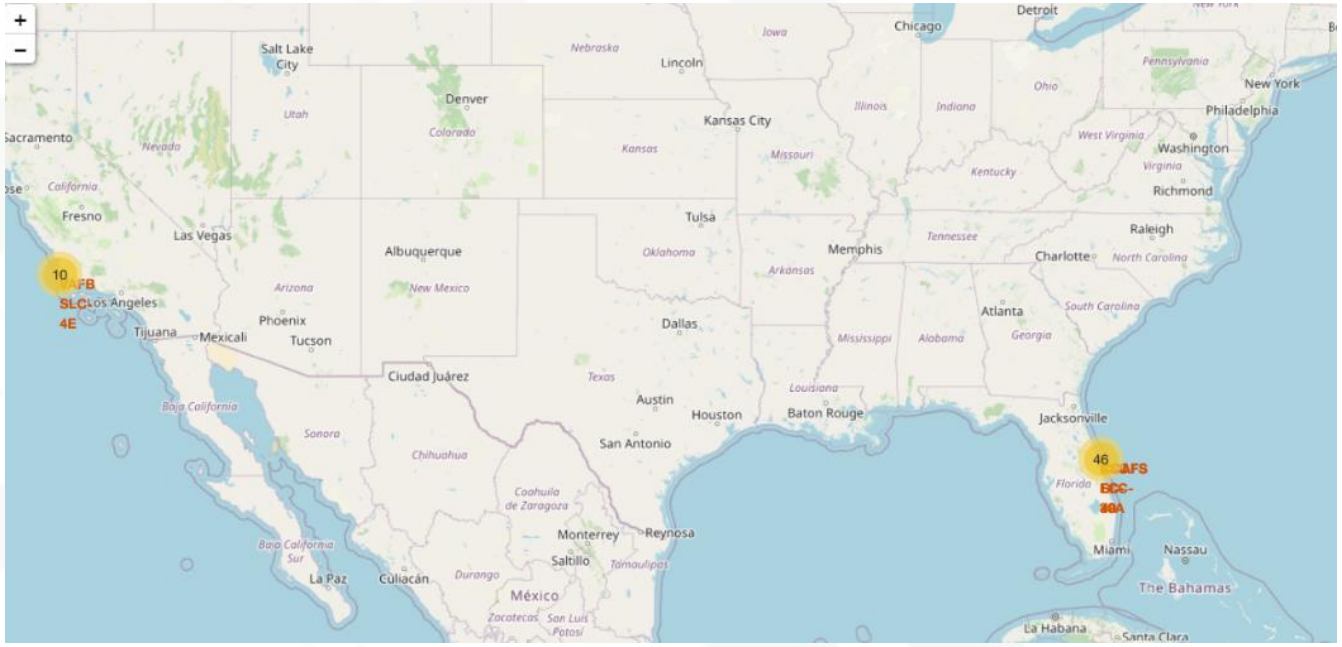
GitHub URL

Interactive maps with Folium

Markers of all Launch Sites

Coloured Markers of the launch outcomes for each Launch Site

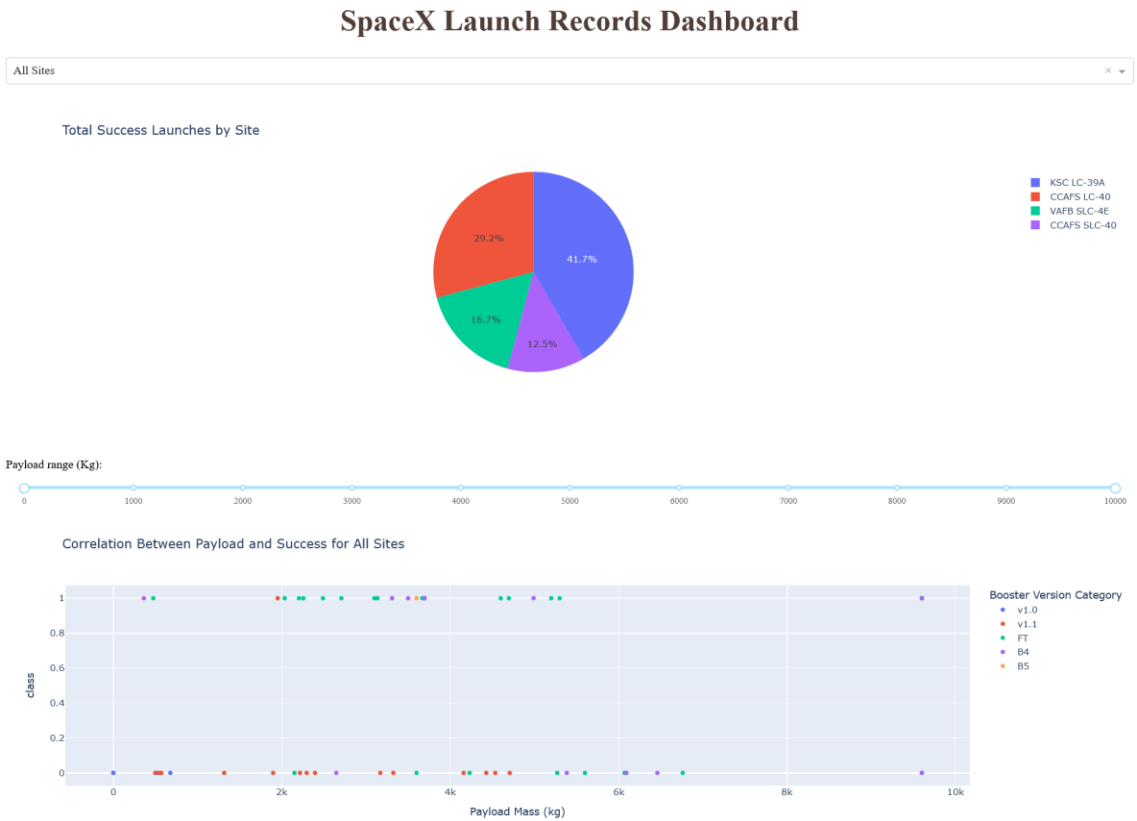
Distances between a Launch Site to its proximities:



GitHub URL: Interactive Maps

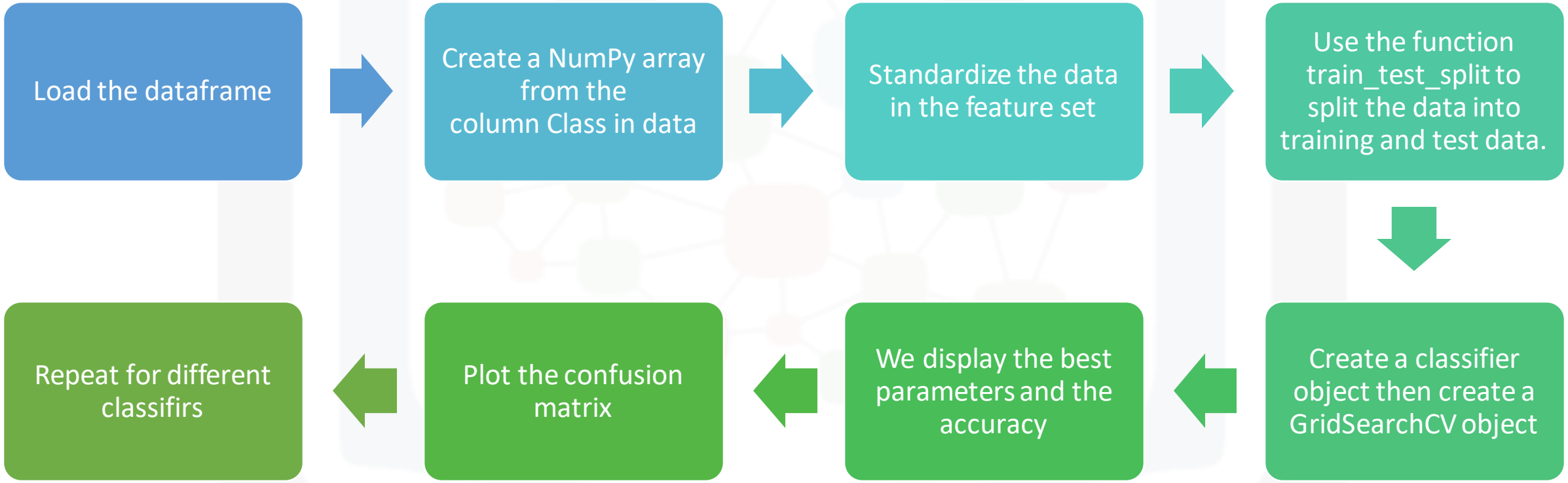
Dashboard with Plotly Dash

- Launch Sites Dropdown List
- Pie Chart showing Successful Launches for All Sites or Specific Site)
- Slider of Payload Mass Range
- Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions



GitHub URL: Dashboard

Predictive Analysis



GitHub URL: Classifiers

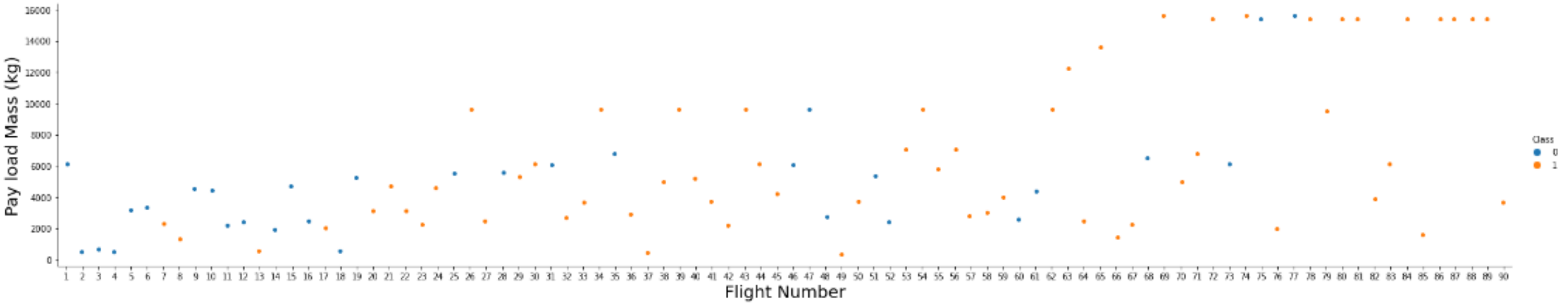
Results

IBM Developer

SKILLS NETWORK

EDA With Data Analysis

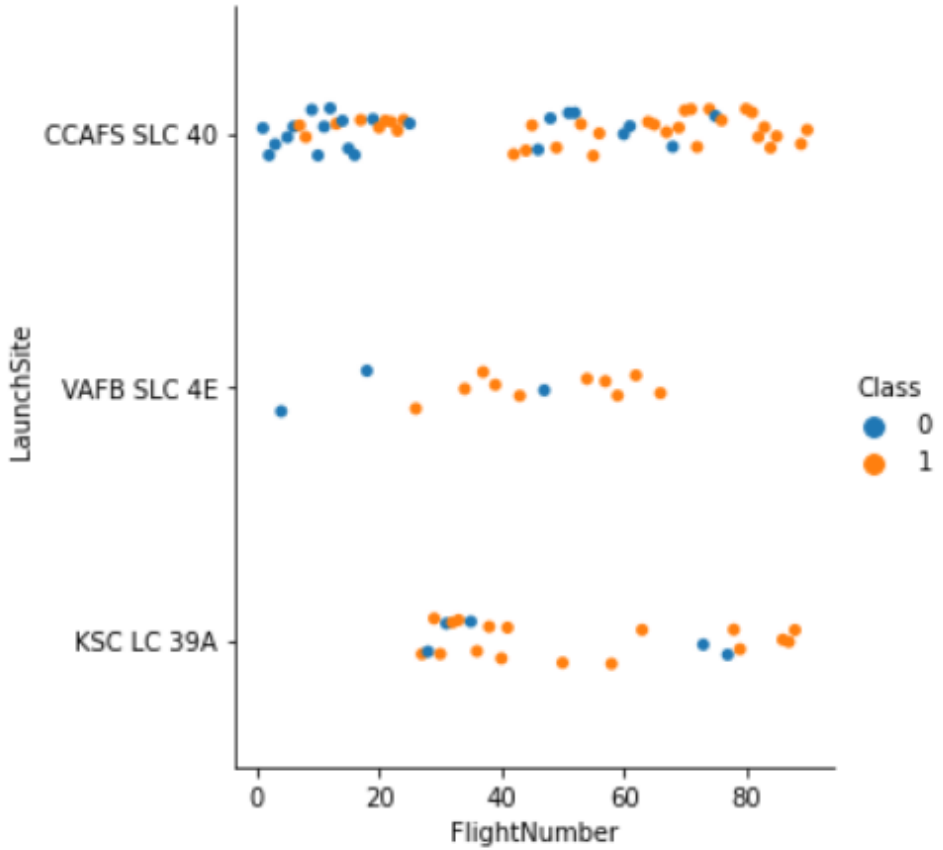
FlightNumber vs. PayloadMass



The initial flights had smaller payloads are were prone to failure.

The payloads gradually increased and more success is observed in the later launches.

FlightNumber vs LaunchSite

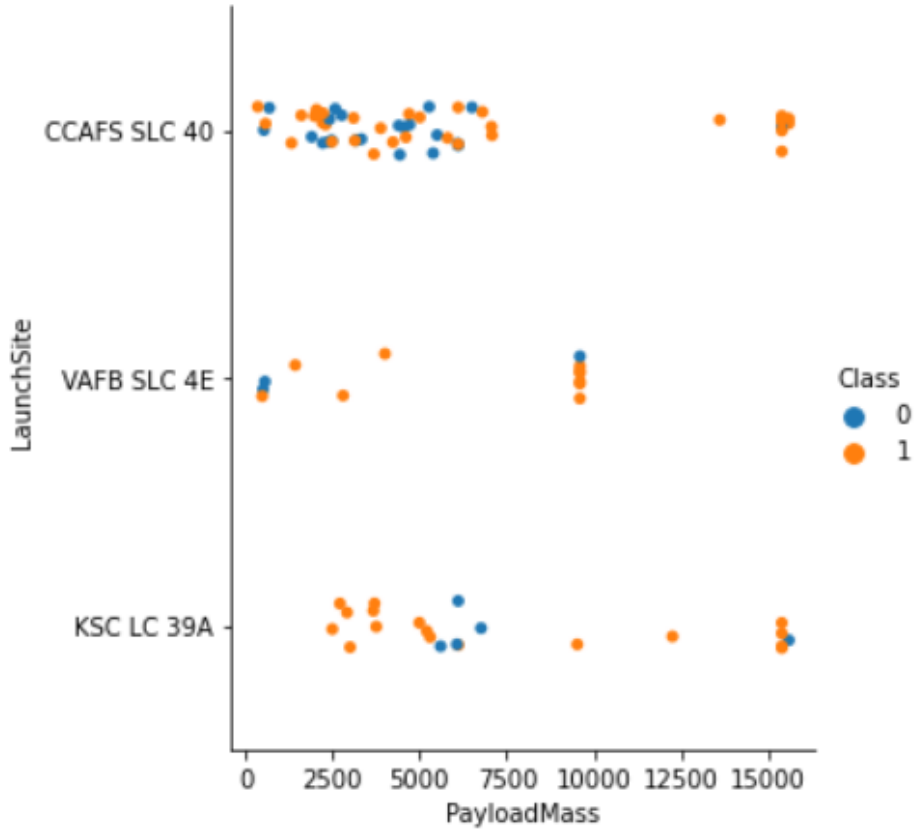


The later launches have higher success rates.

The site 'VAFB SLC 4E' has a lower application rate.

The site 'KSC LC 39A' has a higher success rate.

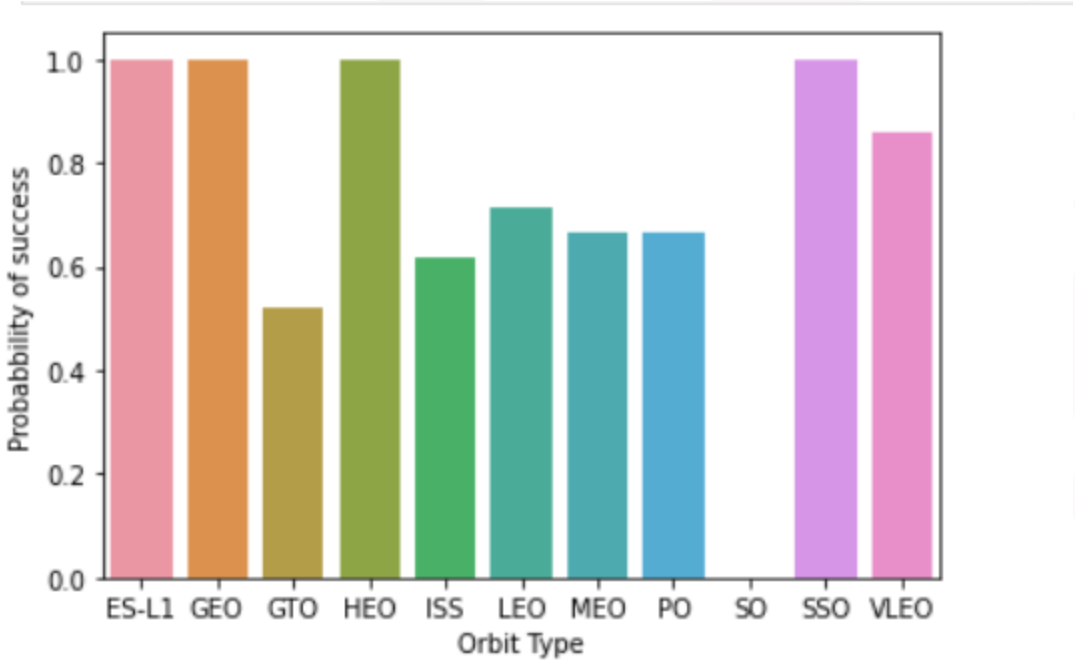
Payload and Launch Site



Large payloads at 'CCAFS SLC 40' and small payloads at 'KSC LC 39A' have large success rates.

Small Payloads at 'CCAFS SLC 40' have a 50% success rate. This could be because most of the initial launches were conducted here.

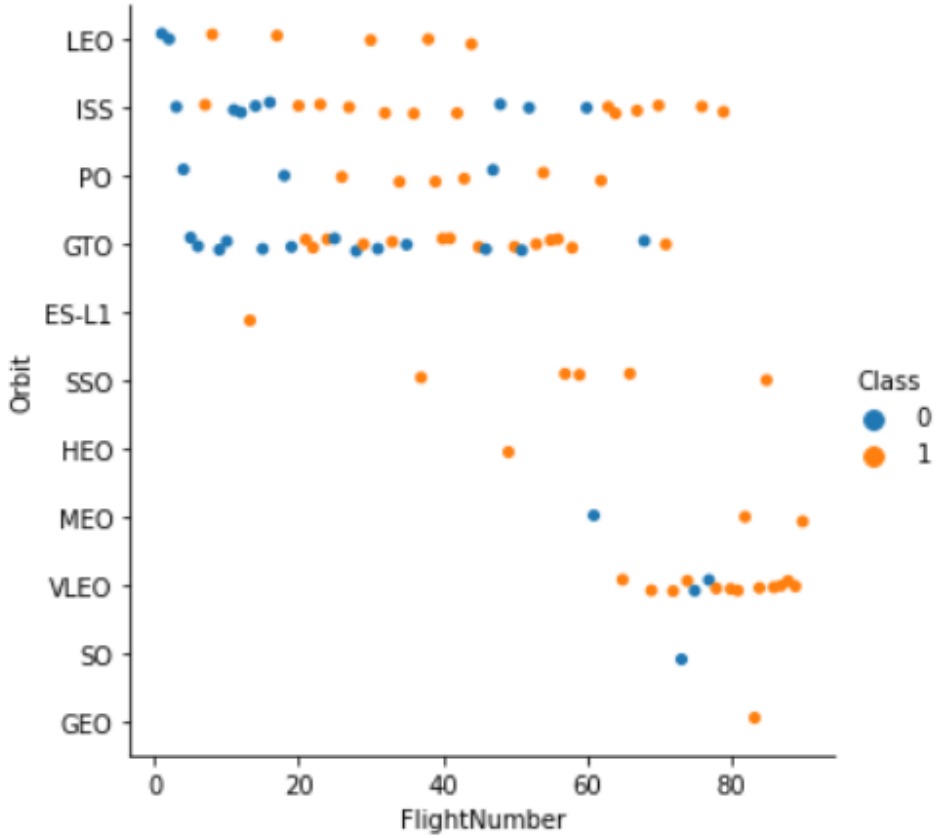
Success rate of each orbit type



'ES-L1', 'GEO', 'HEO' and 'SSO' orbits have 100% success rates.

'SO' orbit has 0% succes rate.

FlightNumber vs. Orbit type

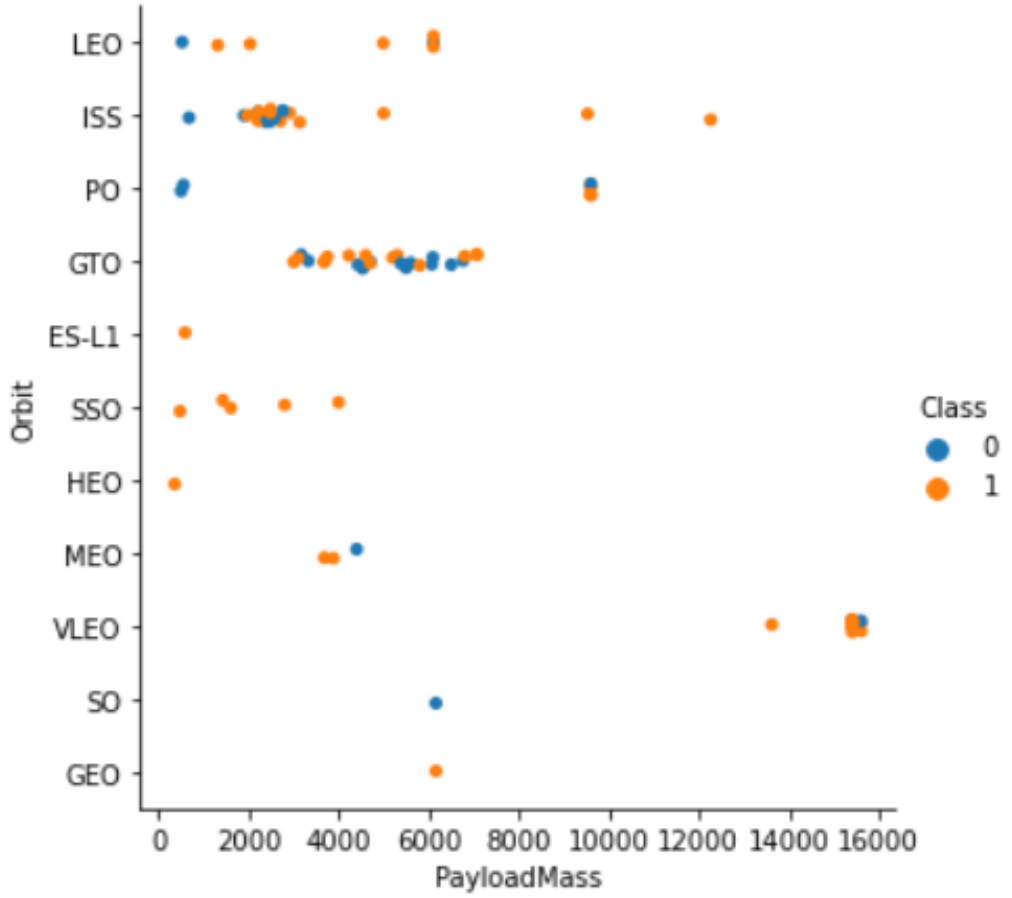


A large number of launches have been in 'ISS', 'PO' and 'GTO'.

'ES-L1', 'HEO', 'SO', 'GEO' and 'MEO' have very few launches.

A large number of recent launches have been in 'VLEO' and a large number of them have been successful.

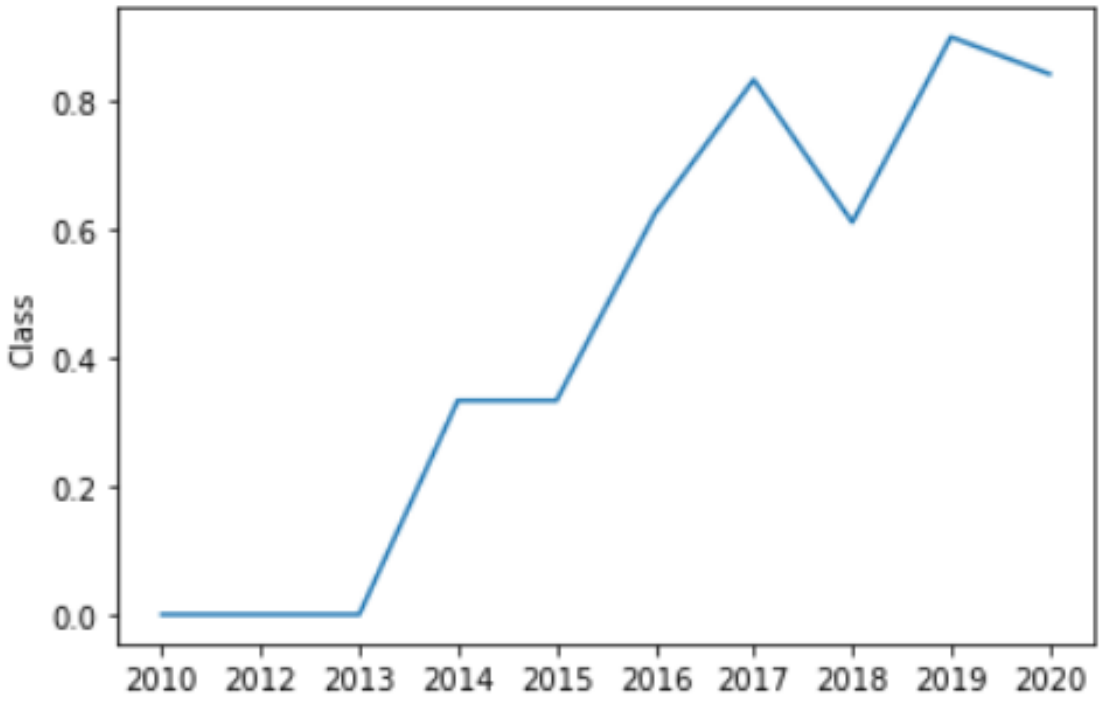
Payload vs. Orbit type



With heavy payloads the successful landing rate is more for VLEO, LEO and ISS.

However for GTO we cannot distinguish this well as both positive landing and negative landing are both there here.

Yearly trend of launch success



The success rate increases with year.

There is a drop during 2018.

EDA with SQL

Names of the unique launch sites in the space mission

In [7]:
```
%sql SELECT DISTINCT("Launch_Site") FROM SPACEXTABLE
```

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

Out[7]:

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

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

In [10]:
```
%sql select * from SPACEXTABLE where Launch_Site like '%CCA%' limit 5
```

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

Out[10]:

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 carried by boosters launched by NASA (CRS) &
The average payload mass carried by booster version F9 v1.1

```
In [14]: %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer='NASA (CRS)'
 * sqlite:///my_data1.db
Done.
Out[14]: sum(PAYLOAD_MASS__KG_)
                          45596
```

```
In [17]: %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version like '%F9 v1.1%'
 * sqlite:///my_data1.db
Done.
Out[17]: avg(PAYLOAD_MASS__KG_)
                  2534.6666666666665
```

The first succesful landing outcome in ground pad was achieved &
Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [20]:
```
%sql select min(Date) as "First_Successful_Landing" from SPACEXTABLE where Landing_Outcome like '%Success%'
```

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

Out[20]:

First_Successful_Landing
2015-12-22

In [22]:
```
%sql select Booster_Version from SPACEXTABLE where Landing_Outcome='Success (drone ship)' and PAYLOAD_MASS__KG_>4000 and PA
```

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

Out[22]:

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

The total number of successful and failure mission outcomes

```
In [25]: %sql select Mission_Outcome, count(*) from SPACEXTABLE group by Mission_Outcome
 * sqlite:///my_data1.db
Done.
```

Out[25]:

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

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

```
In [26]: %sql select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
 * sqlite:///my_data1.db
Done.
```

Out[26]:

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

The records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months 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

In [30]:
```
%sql select substr(Date,6,2) as month, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTABLE where Landing_Outcome=
```

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

Out[30]:

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

In [35]:
```
%sql select "Landing_Outcome", count(*) as 'Count' from SPACEXTABLE \
where "Date" between '2010-06-04' and '2017-03-20' order by 'Count' desc
```

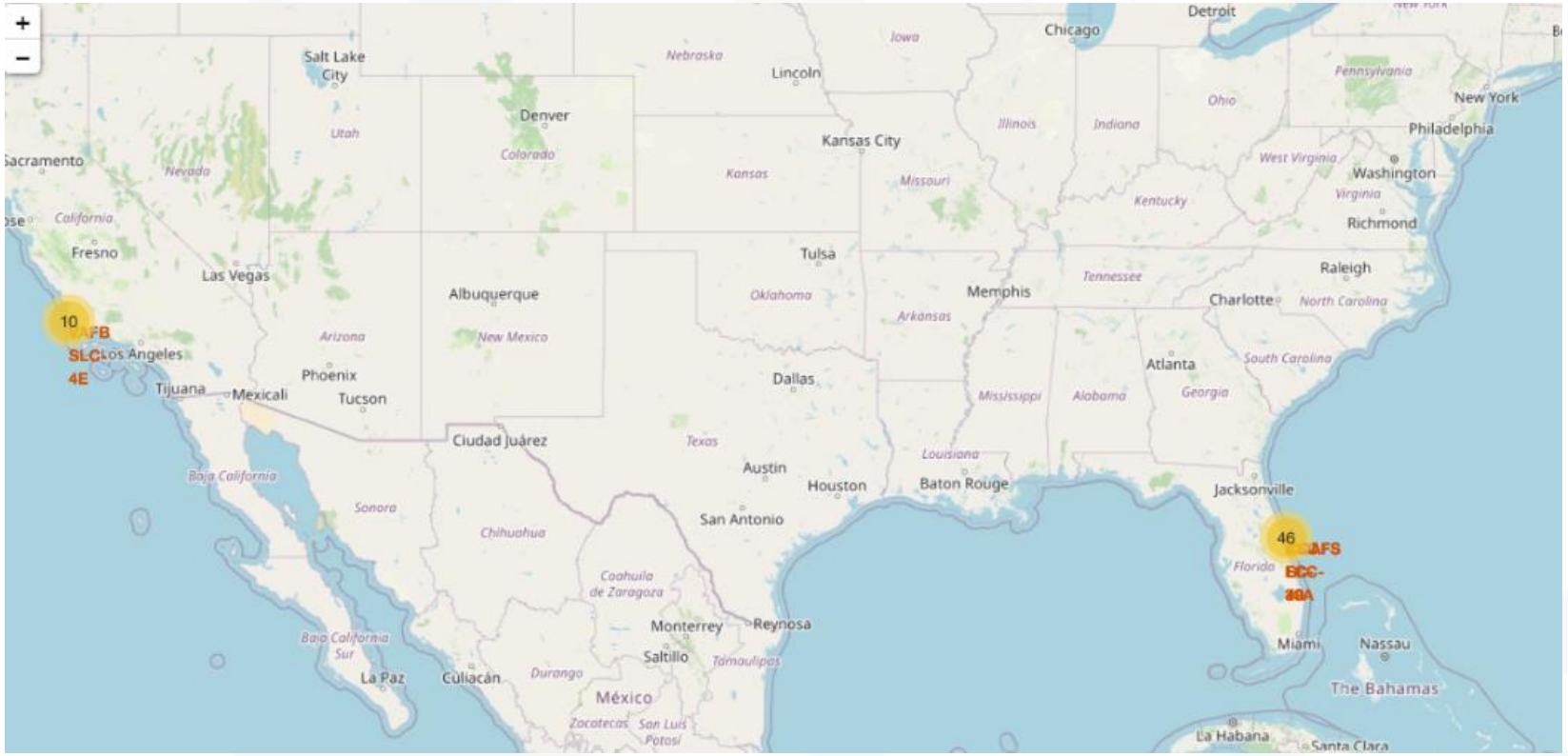
```
 * sqlite:///my_data1.db
Done.
```

Out[35]:

Landing_Outcome	Count
Failure (parachute)	31

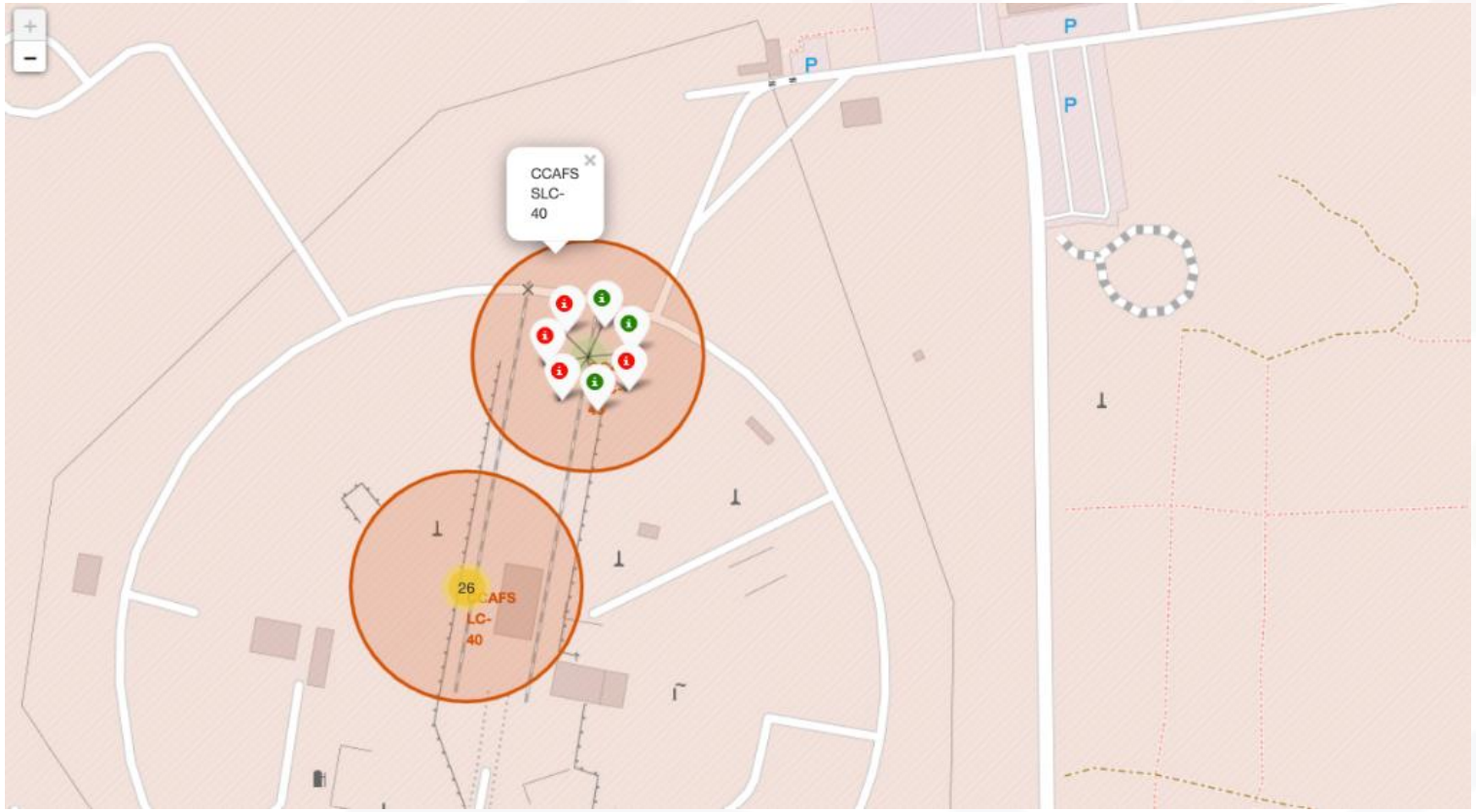
Interactive maps with Folium

All launch sites on a map



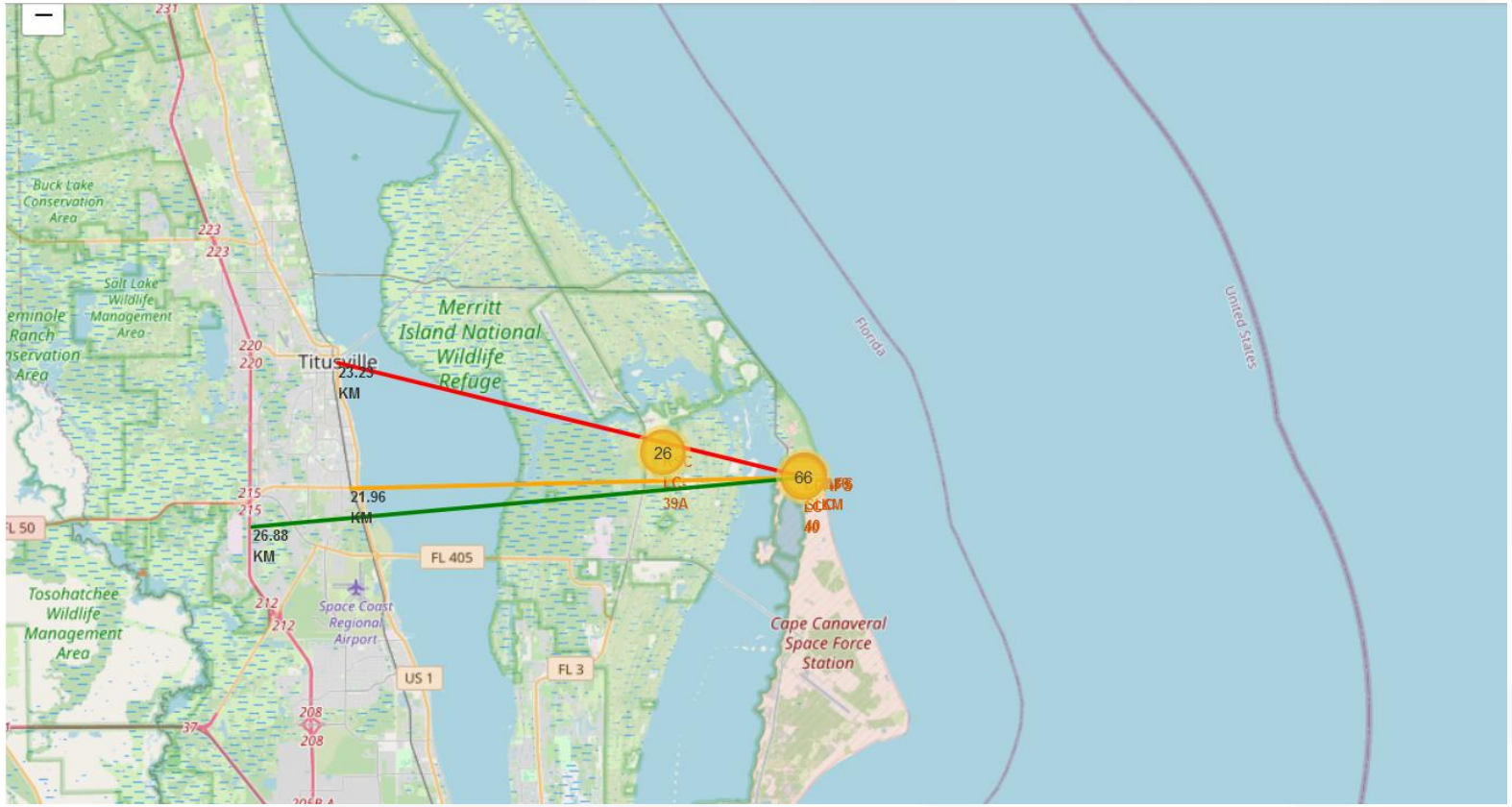
The launch sites appear on the coasts.

Markers showing launch sites with color labels



From the color-labeled markers in marker clusters, we can identify which launch sites have relatively high success rates.

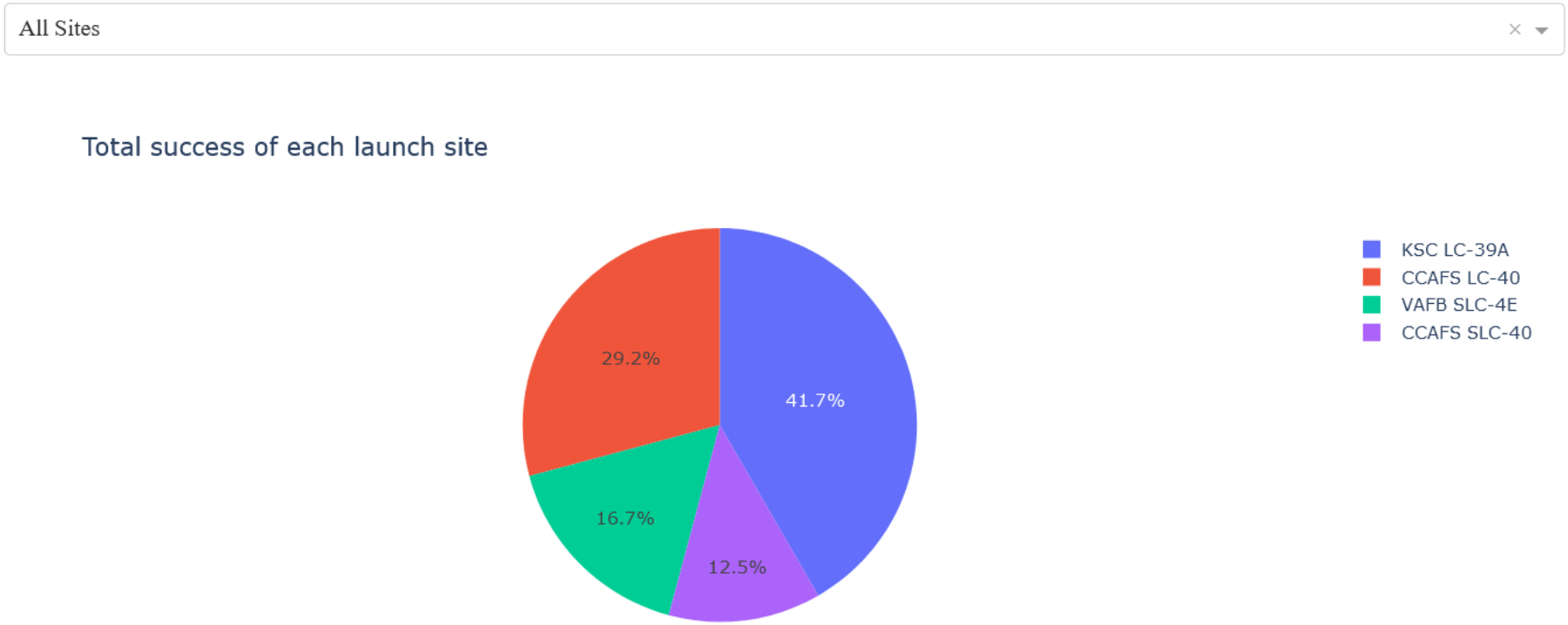
Distance between CCAFS SLC 40 and it's proximities



Launch sites are not close to railways, haighways or cities.

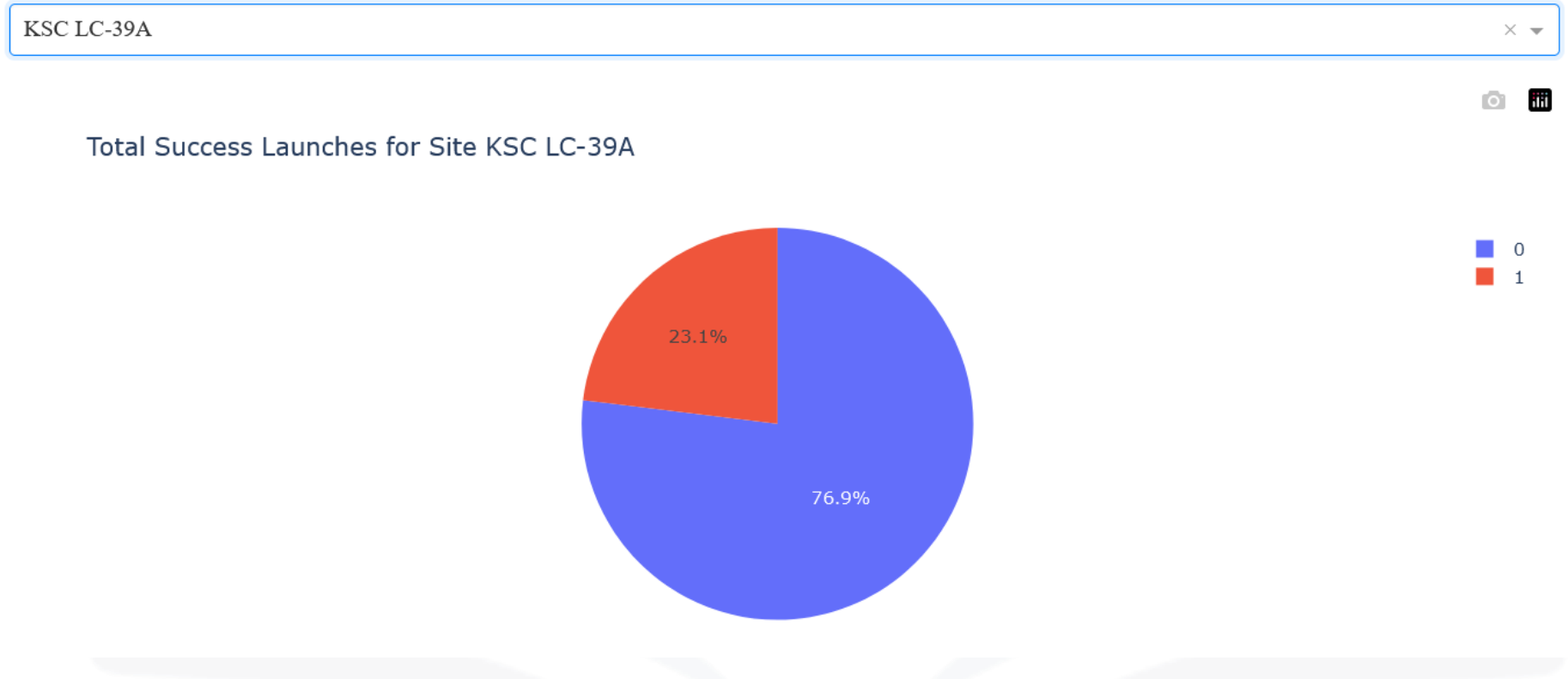
Dashboard with Plotly Dash

Launch success count for all sites

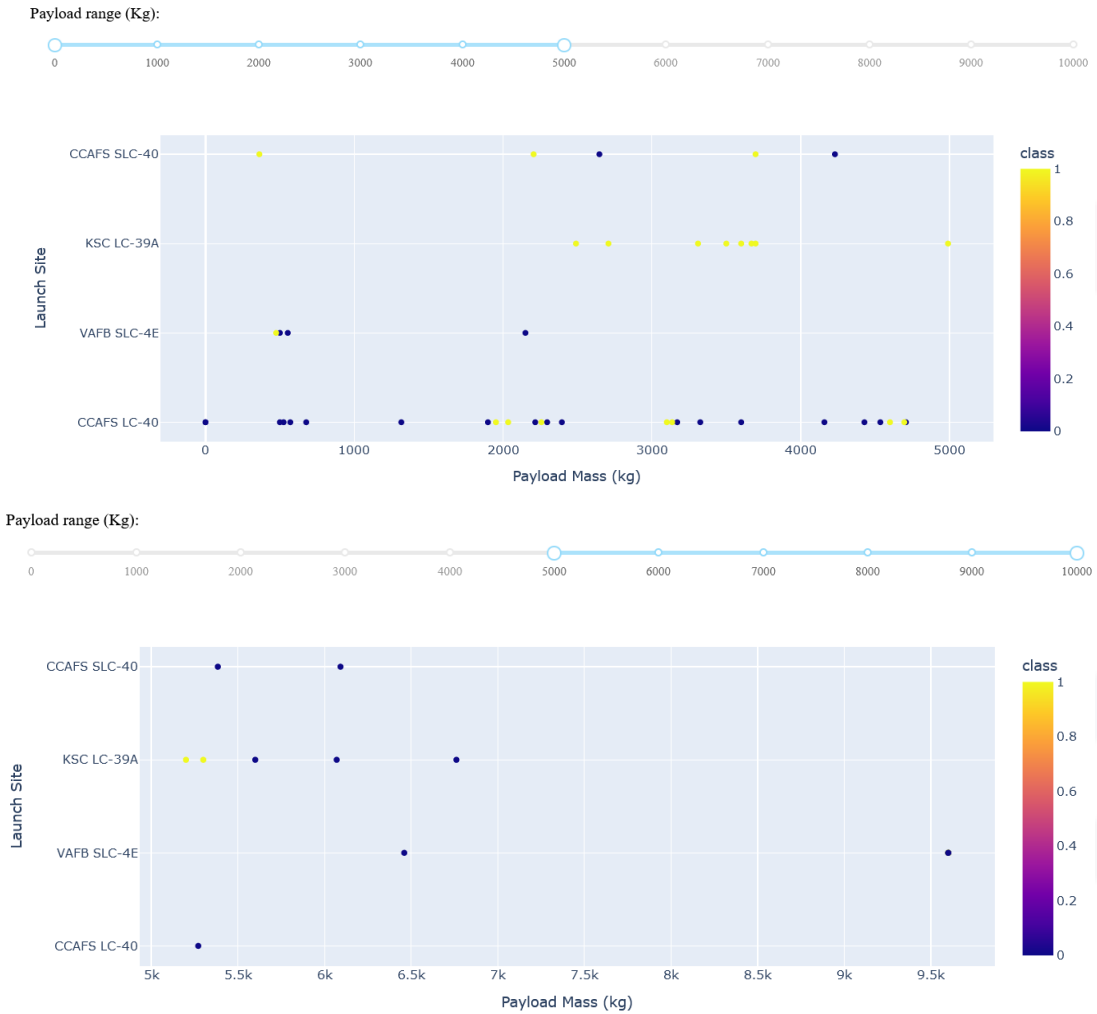


KSC LC-39A has the most successful launches

Launch site with highest launch success ratio



Payload Mass vs. Launch Outcome for all sites




Predictive Analysis
IBM Developer
SKILLS NETWORK

The best score of 0.944 is observed for Decision Tree Classifies.

It also has the best accuracy score.

	Test	Score
0	Logistic Regression	0.833333
1	SVM	0.833333
2	Decision Tree	0.944444
3	KNN	0.833333

The best parameters are {'criterion': 'entropy', 'max_depth': 12, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'splitter': 'random'}

Confusion Matrix for the best Decision Tree Classifier model

CONCLUSION

- Decision Tree Model is the best algorithm for this dataset
- All the sites are in very close proximity to the coast.
- The success rate of launches increases over the years.
- KSC LC-39A has the highest success rate of the launches from all the sites and 'VAFB SLC 4E' has a lower application rate.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.
- Large payloads at 'CCAFS SLC 40' and small payloads at 'KSC LC 39A' have large success rates.

Acknowledgment

Thank you to IBM and the Instructors for all the guidance provided for the course.

THANK YOU!

IBM Developer

SKILLS NETWORK