IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Adithya kumar
02/07/2023

Outline

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

Executive Summary

Summary of methodologies

- Data collection
- Data wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis (Classification)

Summary of all results

- Exploratory Data Analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Introduction

- **Project background and context**
- we will be predicting if the Falcon 9 first stage will land successfully.
- Space offers Falcon 9 rocket launches for $62 million, much cheaper than competitors' prices of $165 million or more. The secret to this affordability lies in Space's ability to reuse the first stage of the Falcon 9 rocket. Unlike traditional rockets that are discarded after launch, the Falcon 9's first stage can detach and safely return to Earth, ready to be used again. This successful landing and reuse process saves a significant amount of money, making Space's rocket launches more cost-effective.

- **Problems you want to find answers**
- With what factors, the rocket will land successfully?
- The effect of each relationship of rocket variables on outcome.
- Conditions which will aid Space have to achieve the best results.

Section 1

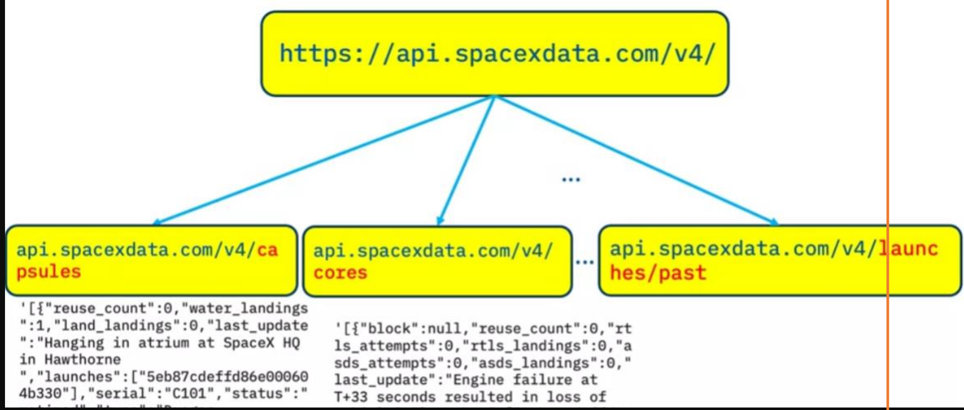
Methodology

Methodology

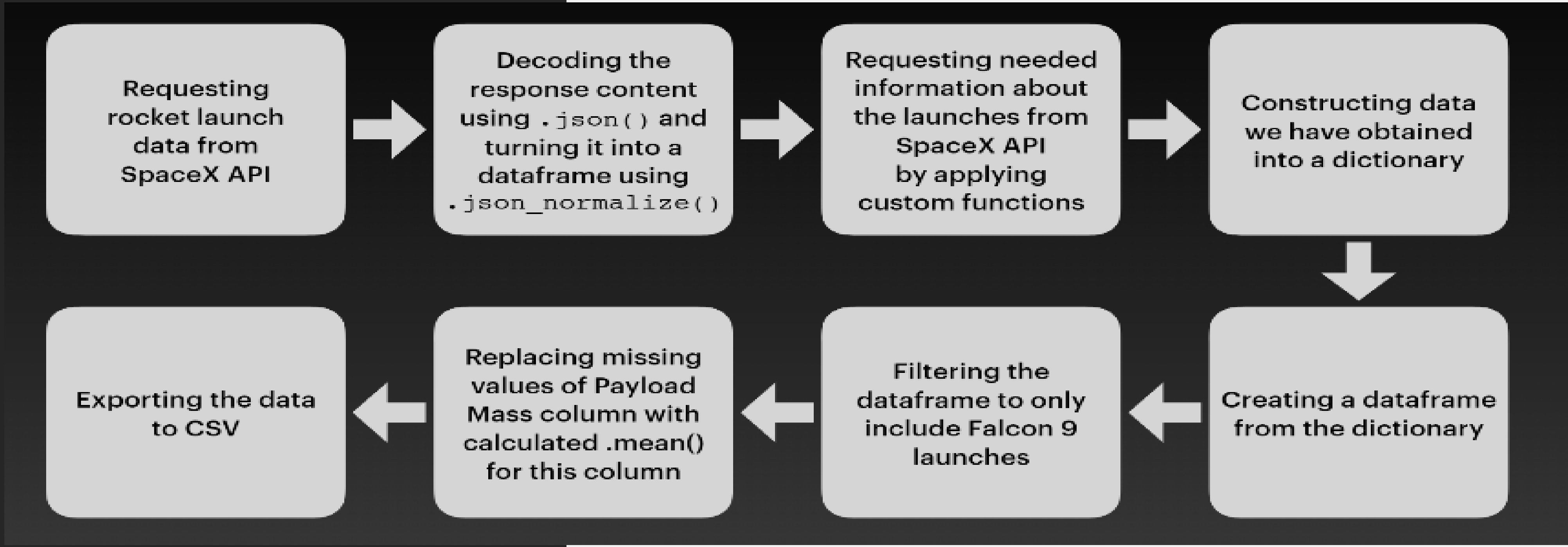
- Executive Summary
- Data collection methodology:
  - Via SpaceX rest API
- Perform data wrangling
  - One hot encoding data fields for machine learning and dropping irrelevant columns via machine learning
- Perform exploratory data analysis (EDA) using visualization and SQL
- scatter and bar graphs show patterns between da ta using Folium and plotly dash
- Perform interactive visual analytics
- using Folium and plotly dash
- Perform predictive analysis using classification models
- Build and evaluate classification models

Data Collection

- Data collection is the process of gathering and measuring information on targeted variables in a relevant way which helps one to answer relevant questions and evaluate outcomes
- Here we gather data Via SpaceX API

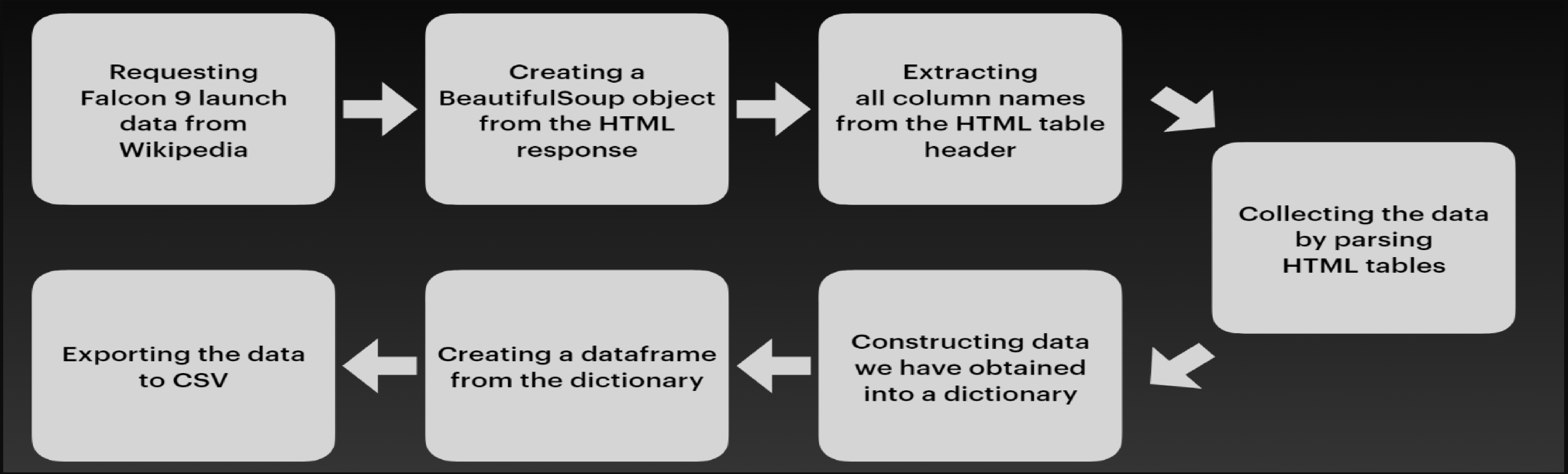


Data Collection – SpaceX API



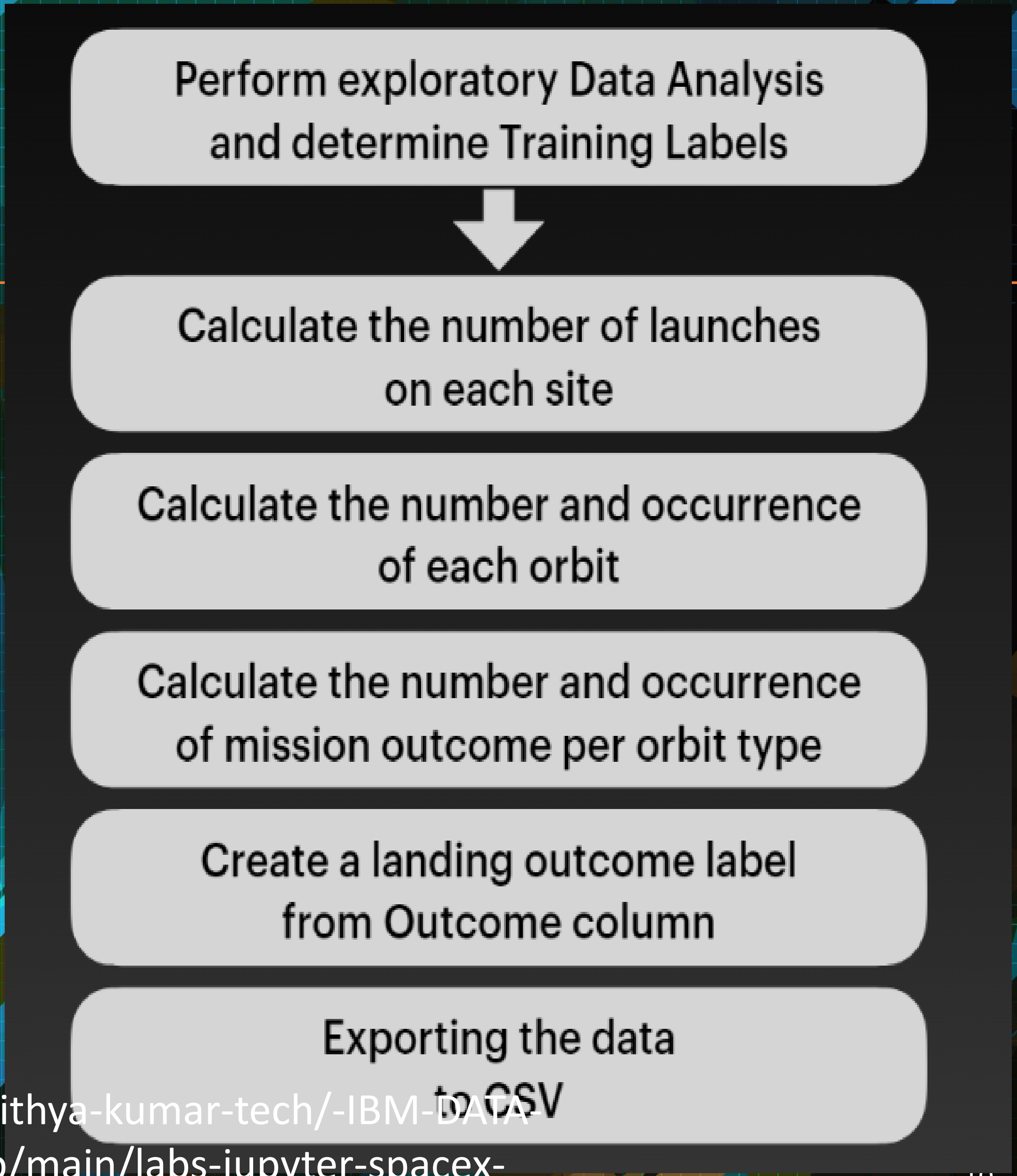
https://github.com/Adithya-kumar-tech/-IBM-DATA-SCIENCE-PROJECT/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection -Web Scraping



Data Wrangling

- In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad.True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship .We mainly convert those outcomes into Training Labels with a" means the booster successfully landed, "O" means it was unsuccessful.



https://github.com/Adithya-kumar-tech/-IBM-DATA-SCIENCE-PROJECT/blob/main/labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

- Charts were plotted:

Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type and Success Rate Yearly Trend Scatter plots show the relationship between variables. If a relationship exists, they could be used in machine learning model .Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value.Line charts show trends in data over time (time series).

https://github.com/Adithya-kumar-tech/-IBM-DATA-SCIENCE-PROJECT/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

Performed SQL queries:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'CCA
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- 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 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•
- Listing the names of the booster versions which have carried the maximum payload mass•
- 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 date2010-06-04 and 2017-03-20 in descending order

https://github.com/Adithya-kumar-tech/-IBM-DATA-SCIENCE-PROJECT/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

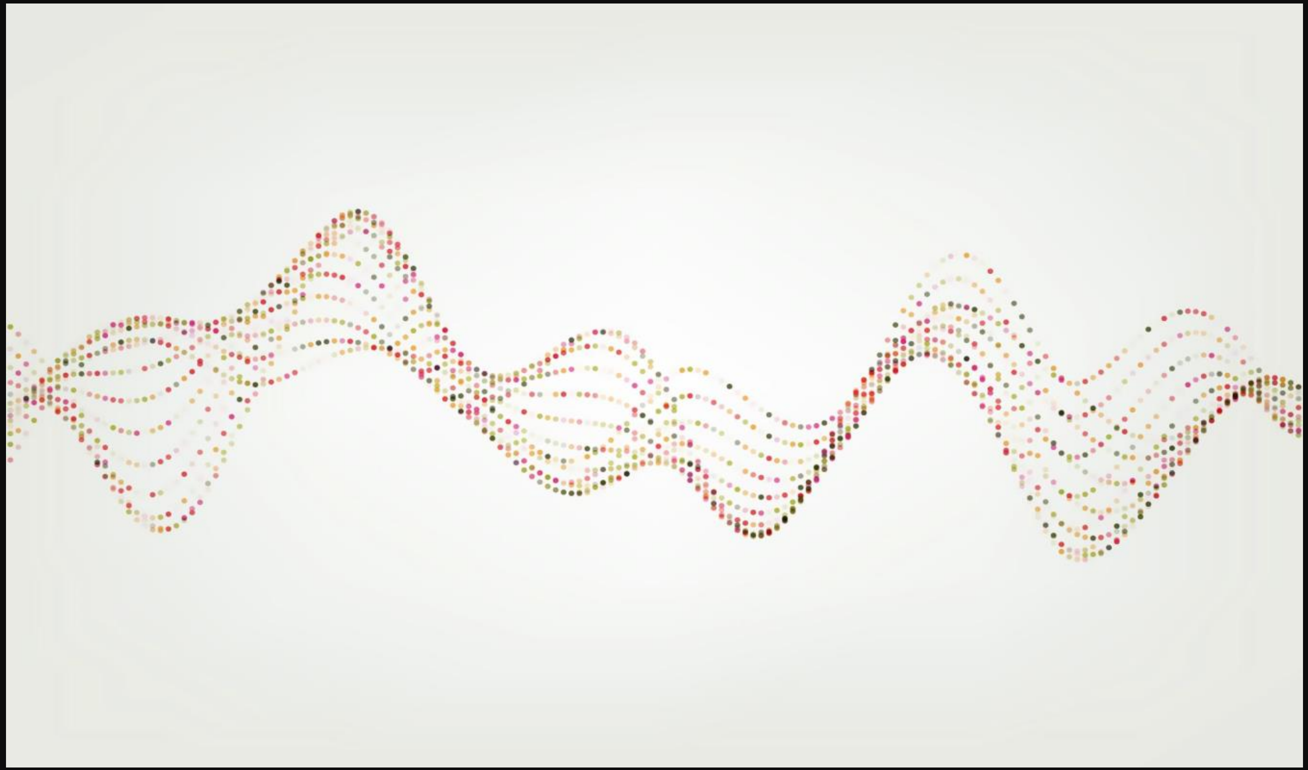
Markers of all Launch Sites:-

- Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.-

- Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts. Colored Markers of the launch outcomes for each Launch Site:-

- Added colored Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates. Distances between a Launch Site to its proximities:-

- Added colored Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City.



Build a Dashboard with Plotly Dash

- Launch Sites Dropdown List:-

❑ Added a dropdown list to enable Launch Site selection.

- Pie Chart showing Success Launches (All Sites/Certain Site):-

❑ Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.

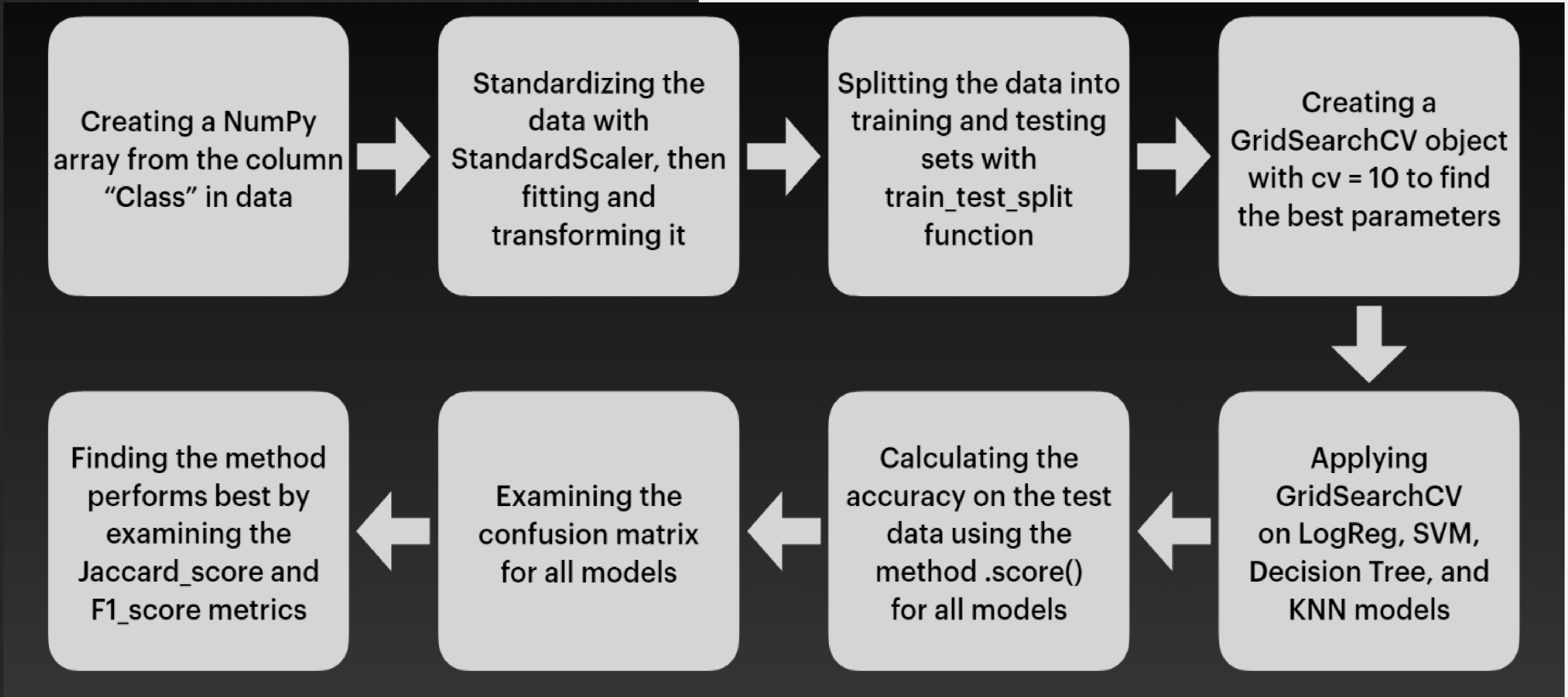
- Slider of Payload Mass Range:-

❑ Added a slider to select Payload range.Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:-

❑ Added a scatter chart to show the correlation between Payload and Launch Success.

Predictive Analysis (Classification)

https://github.com/Adithya-kumar-tech/-IBM-DATA-SCIENCE-PROJECT/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb



Results

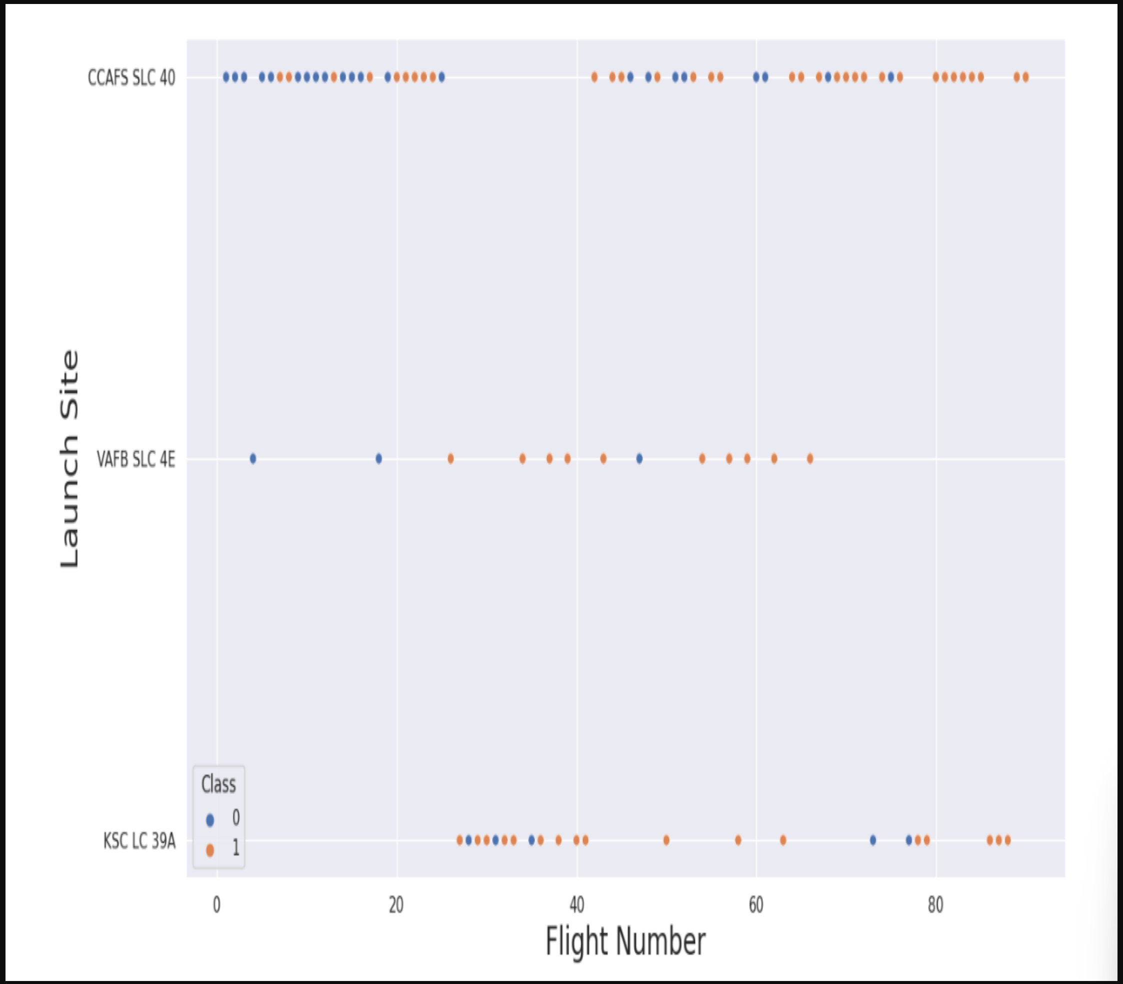
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

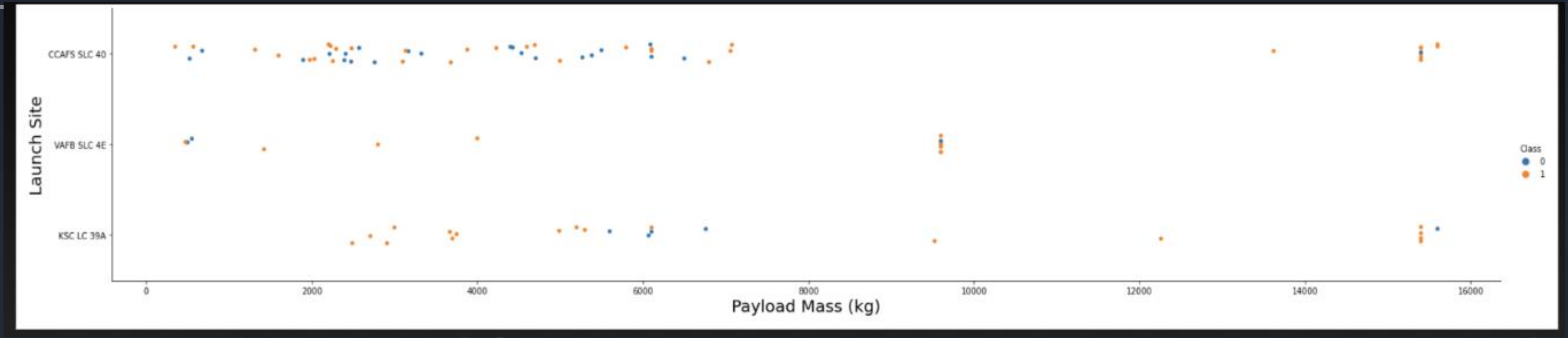
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

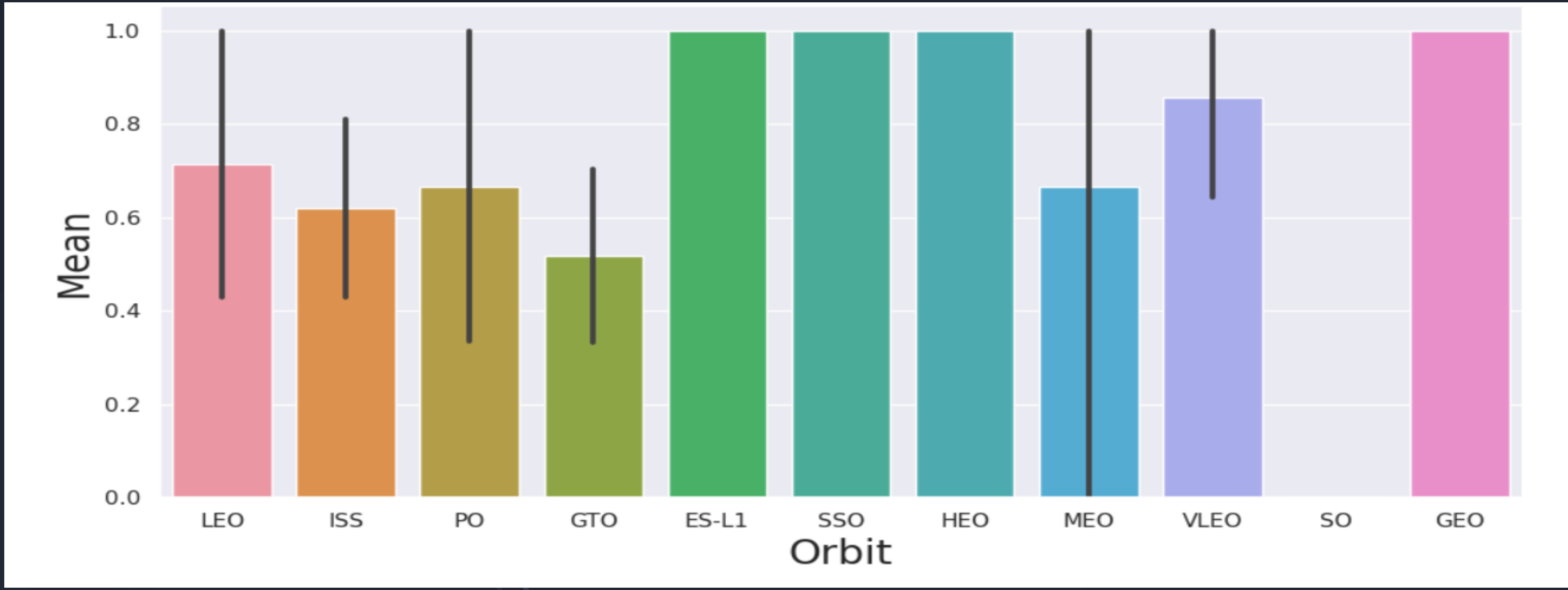
- The earliest flights all failed while the latest flights all succeeded
- The CCAFS SLC launch site has about a half of al launches
- VAFB SLC 4E AND KSC LC 39 a have higher success rates
- It can be assumed that each new launch has a higher rate of success





Payload vs. Launch Site

- For every launch site the higher the payload mass the higher the success rate
- Most of the launches with payload mass over 7000kg were successful
- KSA LC 39A HAS A 100% SUCCESS RATE FOR PAYLOAD MASS UNDER 5500 KG TOO



Success Rate vs. Orbit Type

- Orbit with the 100%success rate

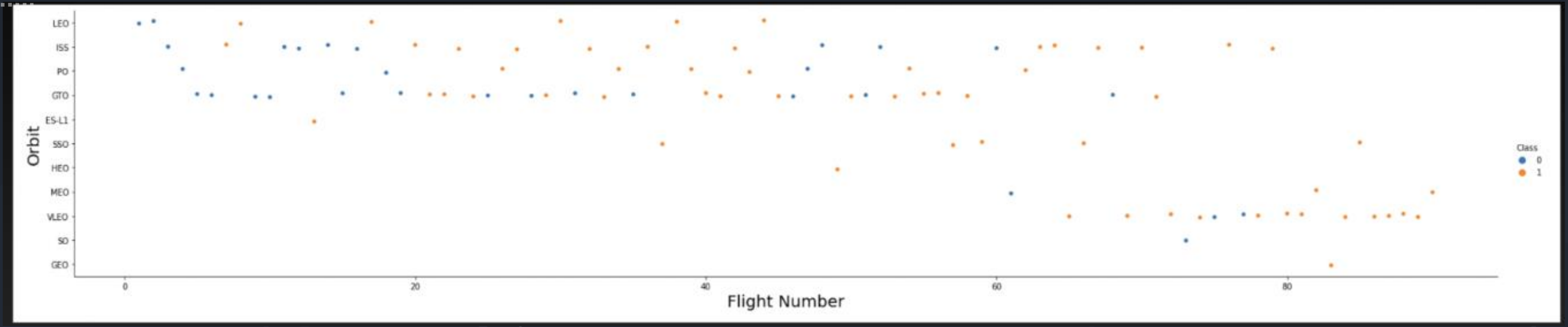
  - ES-L1, GEO,HEO, SSO

- Orbits with 0% success rate

  -SO

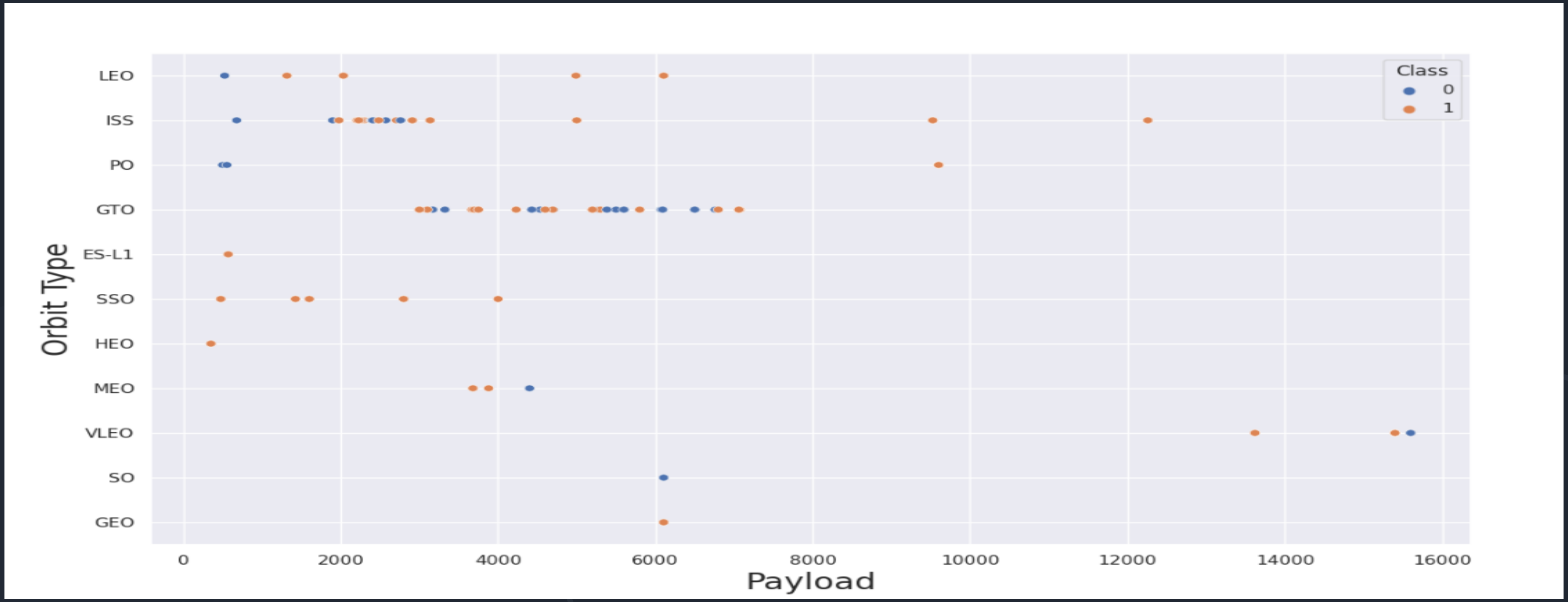
- Orbits with success rate between 50% and 85%

  -GTO, ISS, LEO, MEO, PO



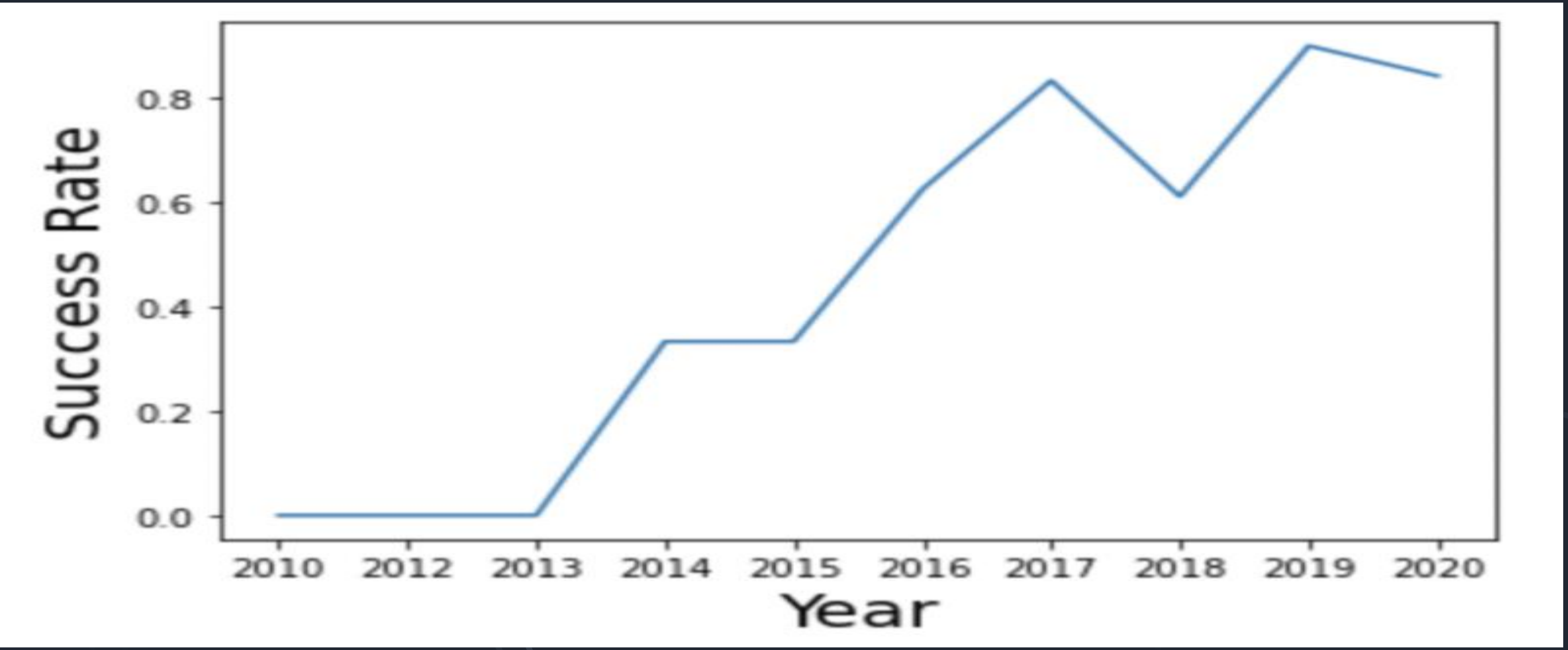
Flight Number vs. Orbit Type

- RANKING THE COUNT OF LANDING OUTCOMES OR SUCCESS BETWWEN THE DATE 2010-06-04 AND 2017-03-20 IN DESENDING ORDER



Payload vs. Orbit Type

- Heavy payloads have a negative influence on GTO orbits and positive on GTO and polar LEO (ISS) orbits



Launch Success Yearly Trend

- Success rate since 2013 kept increasing till 2020

EDA WITH SQL

```
In [19]: %sql SELECT DISTINCT LAUNCH_SITE as "Launch_Sites" FROM SPACEXTBL;
 * sqlite:///my_data1.db
Done.
```

Out[19]:

Launch_Sites
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

All Launch Site Names

Displaying all unique launch sites

```
In [20]: %sql SELECT LAUNCH_SITE FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;

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

Out[20]:

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Launch Site Names Begin with 'CCA'

Displaying 5 records where launch sites begin with the string CCA

```
In [21]: %sql SELECT SUM (PAYLOAD_MASS__kg_) FROM SPACEXTBL WHERE CUSTOMER = 'NASA(CRS)' ;

 * sqlite:///my_data1.db
Done.

Out[21]:
```

SUM (PAYLOAD_MASS__kg_)
None

Total Payload Mass

Displaying total payload mass kg

```
In [41]: %sql SELECT AVG (PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
 * sqlite:///my_data1.db
Done.
```

Out[41]:

AVG (PAYLOAD_MASS__KG_)
2928.4

Average Payload Mass by F9 v1.1

```
In [22]: %sql SELECT MIN (DATE) AS "First Successful Landing" FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (ground pad)';
 * sqlite:///my_data1.db
Done.
```

Out[22]:

First Successful Landing
01/08/2018

First Successful Ground Landing Date

First successful ground landing date is 01/08/2018

```
In [23]: %sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOA

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

Out[23]:

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

Successful Drone Ship Landing with Payload between 4000 and 6000

Listing the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
In [31]: %sql SELECT COUNT(MISSION_OUTCOME) AS "successful mission" FROM SPACEXTBL WHERE MISSION_OUTCOME LIKE 'Success%';

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

Out[31]:

successful mission
100

```
In [33]: %sql SELECT COUNT(MISSION_OUTCOME) AS "failure mission" FROM SPACEXTBL WHERE MISSION_OUTCOME LIKE 'Fail%';

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

Out[33]:

failure mission
1

```
In [34]: %sql SELECT sum(case when MISSION_OUTCOME LIKE '%Success%' then 1 else 0 end) AS "Successful Mission", \
             sum(case when MISSION_OUTCOME LIKE '%Failure%' then 1 else 0 end) AS "Failure Mission" \
         FROM SPACEXTBL;

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

Out[34]:

Successful Mission	Failure Mission
100	1

Total Number of Successful and Failure Mission Outcomes

The total number of successful is 100 and failure mission outcome is 1

```
In [26]: %sql SELECT DISTINCT BOOSTER_VERSION AS "Booster Versions which carried the Maximum Payload Mass" FROM SPACEXTBL \
WHERE PAYLOAD_MASS__KG_ =(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);

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

Out[26]:

Booster Versions which carried the Maximum 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

Boosters Carried Maximum Payload

Listing the names of the booster which have carried the maximum payload mass

```
In [12]: %%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET
               where landing__outcome = 'Failure (drone ship)' and year(date)=2015;

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[12]:
```

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

2015 Launch Records

The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
In [13]: %%sql select landing__outcome, count(*) as count_outcomes from SPACEXDATASET
         where date between '2010-06-04' and '2017-03-20'
         group by landing__outcome
         order by count_outcomes desc;

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[13]:

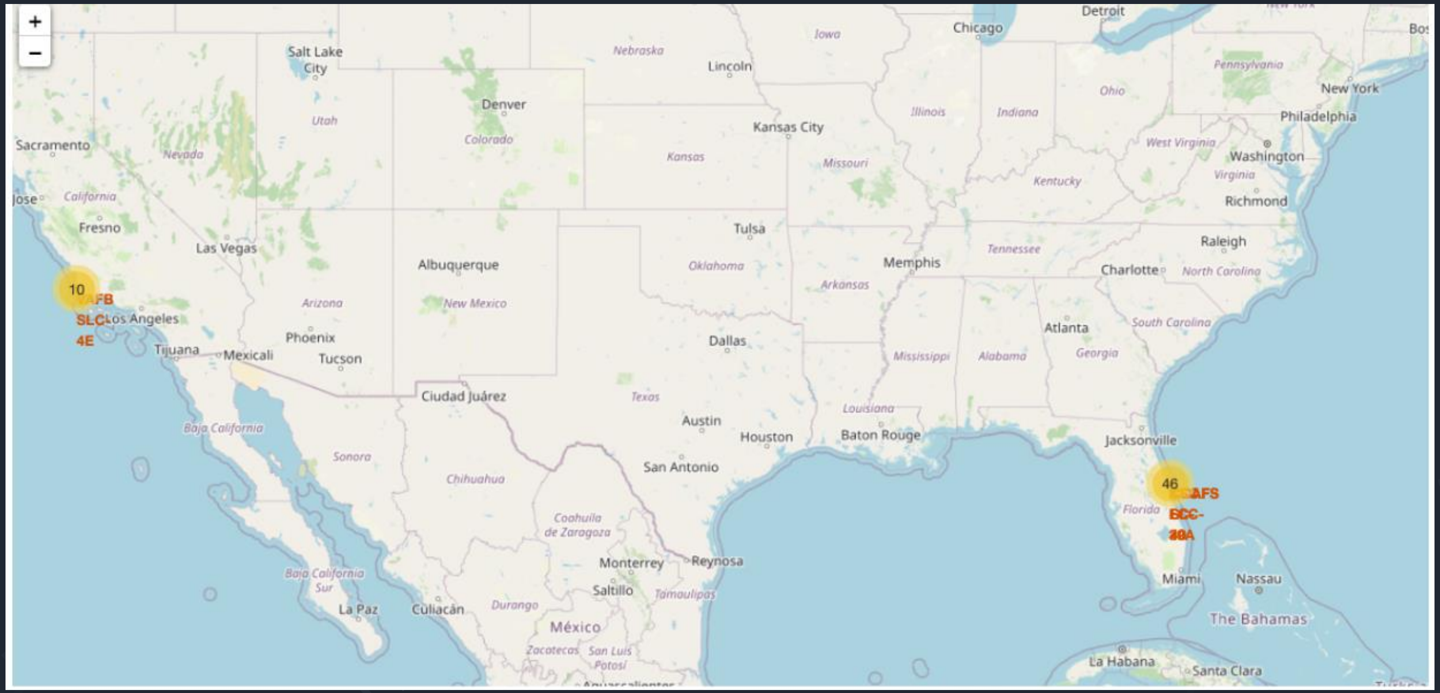
landing__outcome	count_outcomes
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

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

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

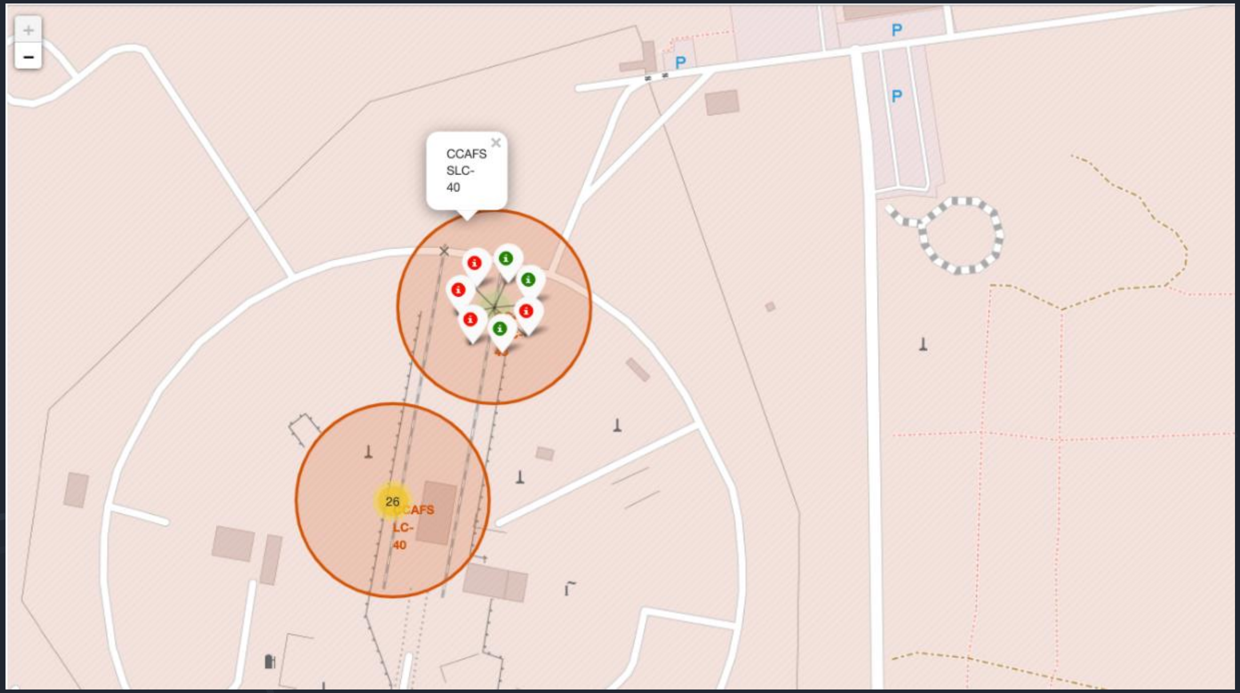
Section 3

Launch Sites Proximities Analysis



All launch sites' location markers on global map

- Most of Launch sites are in proximity to the Equator line. The land is moving faster at the equator than any other place on the surface of the Earth. Anything on the surface of the Earth at the equator is already moving at 1670 km/hour. If a ship is launched from the equator it goes up into space, and it is also moving around the Earth at the same speed it was moving before launching. This is because of inertia. This speed will help the spacecraft keep up a good enough speed to stay in orbit.
- All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimizes the risk of having any debris dropping or exploding near people



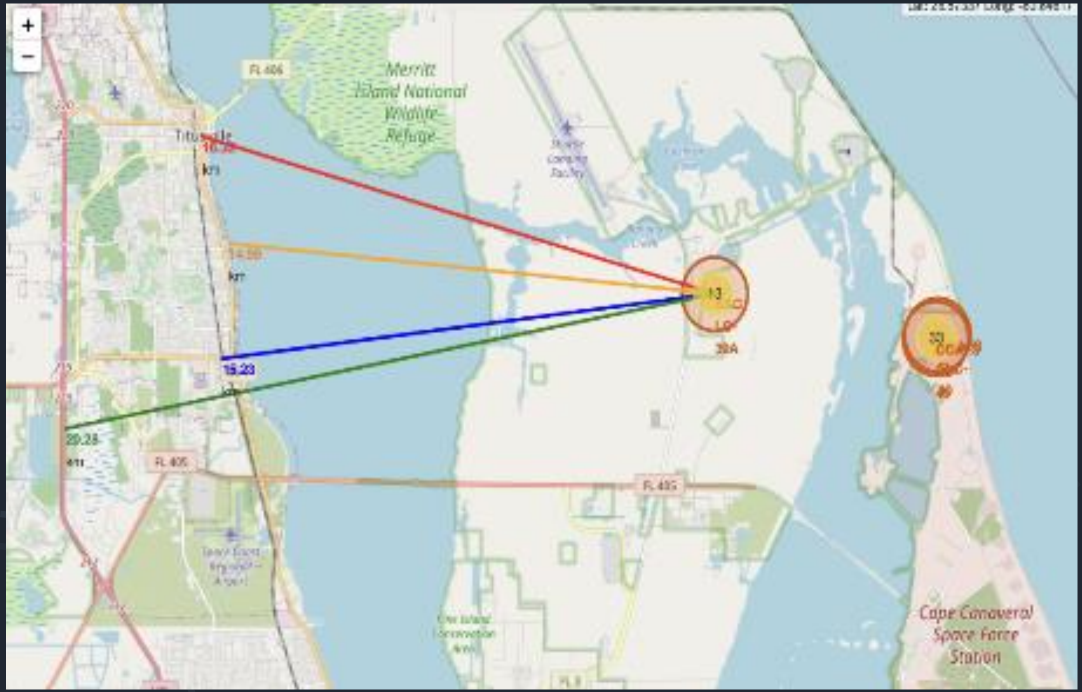
Colour-labeled launch records on the map

- • From the colour-labeled markers we should be able to easily identify which launch sites have relatively high success rates.

  Green Marker = Successful Launch

  Red Marker = Failed Launch•

- Launch Site KSC LC-39A has a very high Success Rate.



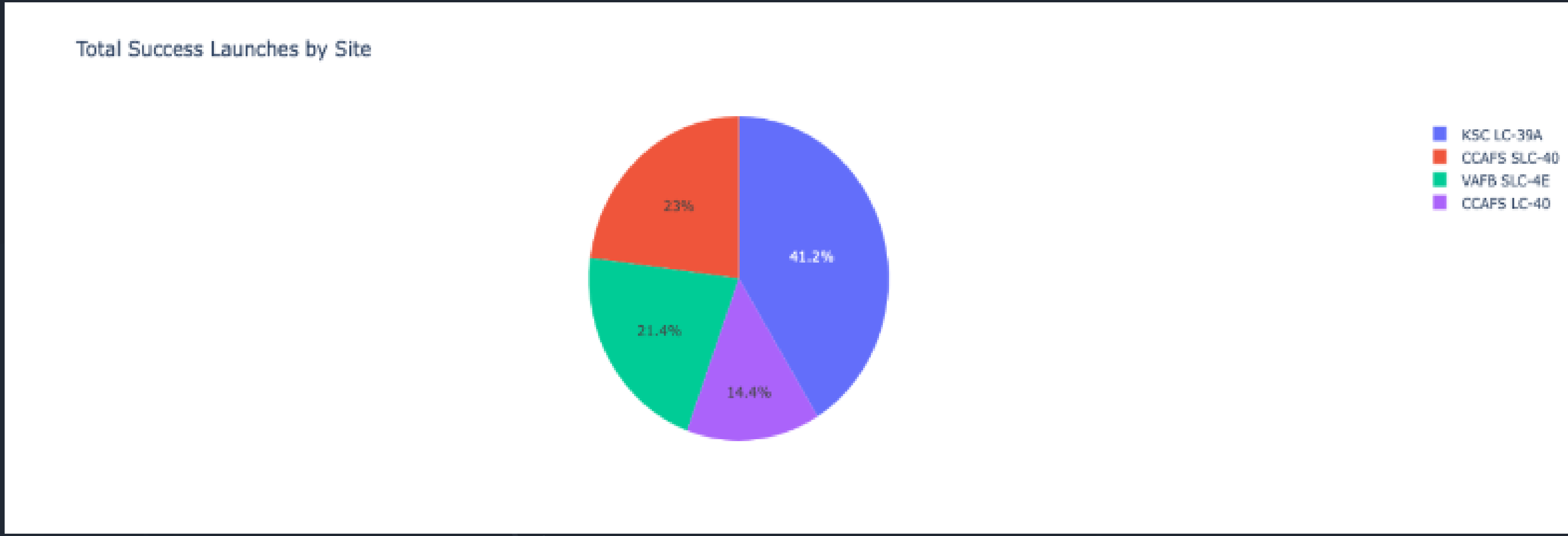
Distance from the launch site to its proximities

• From the visual analysis of the launch site KSC LC-39A we call clearly see that it is

- relatively close to railway (15.23 km)•
- relatively close to highway (20.28 km).
- relatively close to coastline (14.99 km)•
- Also the launch site KSC LC-39A is relatively close to its closest city
- Titusville (16.32 km).• Failed rocket with its high speed can cover distances like 15-20 km in few seconds. It could be potentially dangerous to populated areas.

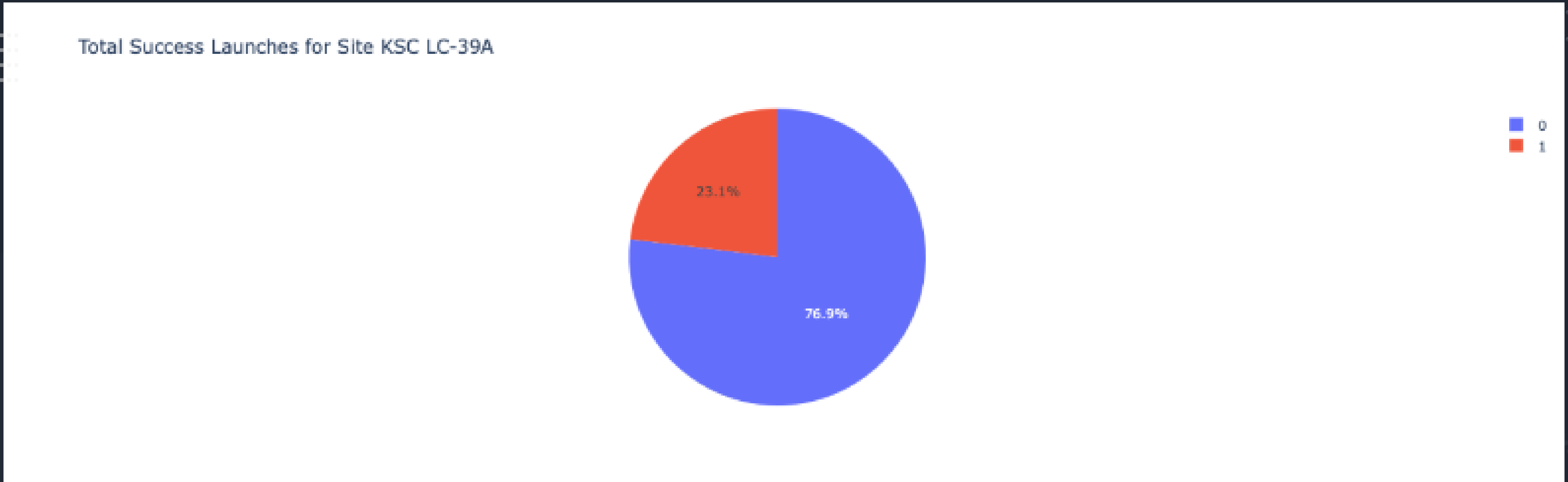
Section 4

Build a Dashboard with Plotly Dash



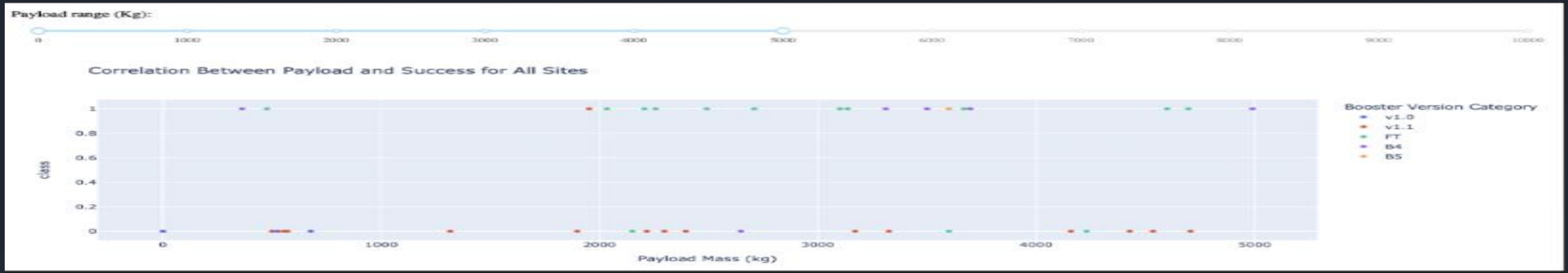
Launch success count for all sites

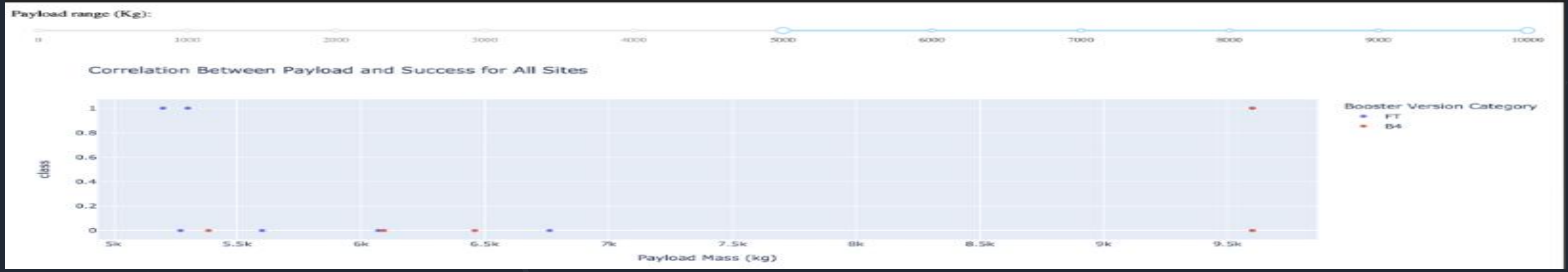
- The chart clearly shows the from all the sites KLC LC-39A has the most successful launches



Launch site with highest launch success ratio

- KSC LC- 39A has the highest launch rate (79.9%) with 10 successful and only 3 failed landings





Payload Mass vs launch Outcome for all sites

The charts show that payloads between 2000 and 5500 kg have the highest success rate

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Based on the scores of the Test Set, we can not confirm which method performs best.
- Same Test Set scores may be due to the small test sample size (18 samples). Therefore, we tested all methods based on the whole Dataset.
- The scores of the whole Dataset confirm that the best model is the Decision Tree Model. This model has not only higher scores, but also the highest accuracy.ch model has the highest classification accuracy

Scores and Accuracy of the Test Set

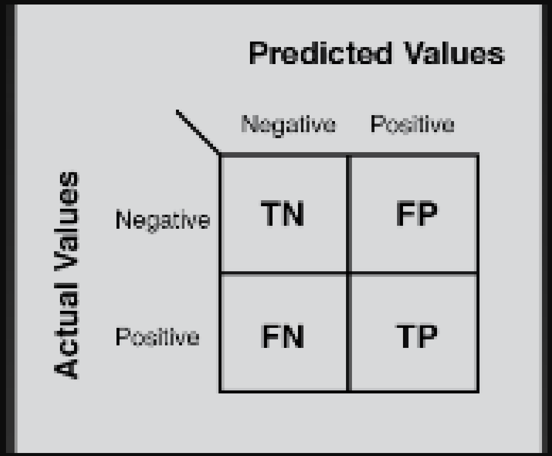
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

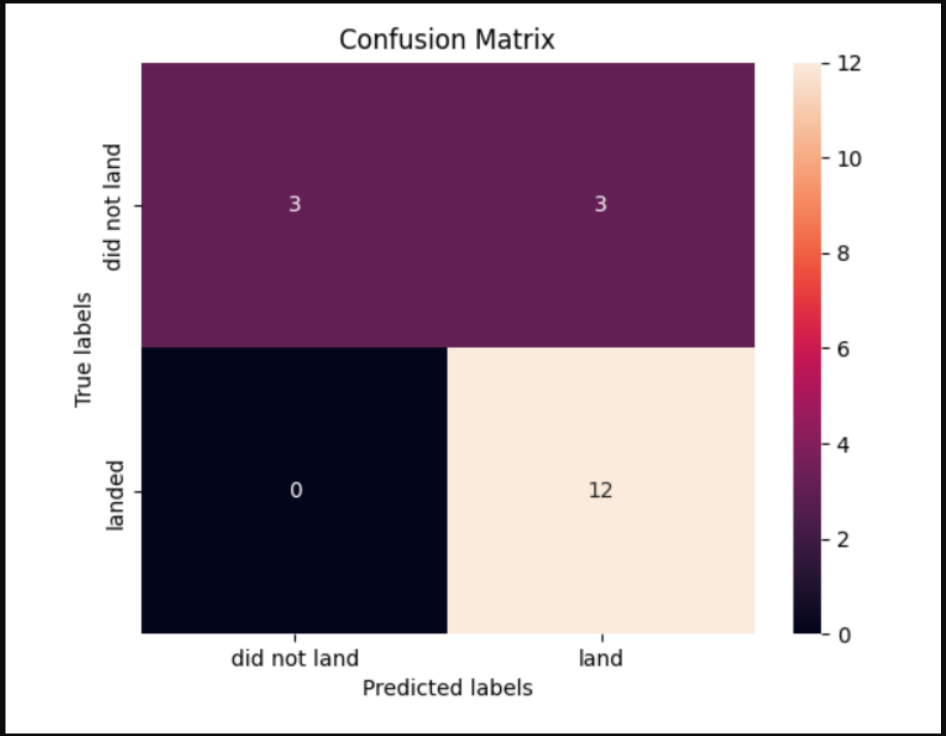
Scores and Accuracy of the Entire Data Set

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

Confusion Matrix

- Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.





Conclusions

- Decision Tree Model is the best algorithm for this dataset
- Launches with a low payload mass show better results than launches with a larger payload mass.
- Most of launch sites are in proximity to the Equator line and 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.
- Orbits ES-LI, GEO, HEO and SSO have 100% success rate.

Appendix

- INTERACTIVE PLOTY
- FOLIUM MEASURECONTROL PLUGIN TOOL
- FOLIUM CUSTOM TITLE LAYERS WITH LABLES
- IBM COGNOS VISUALIZATION TOOL
- BASIC DECISION TREE CONTRUCTION

Thank you!