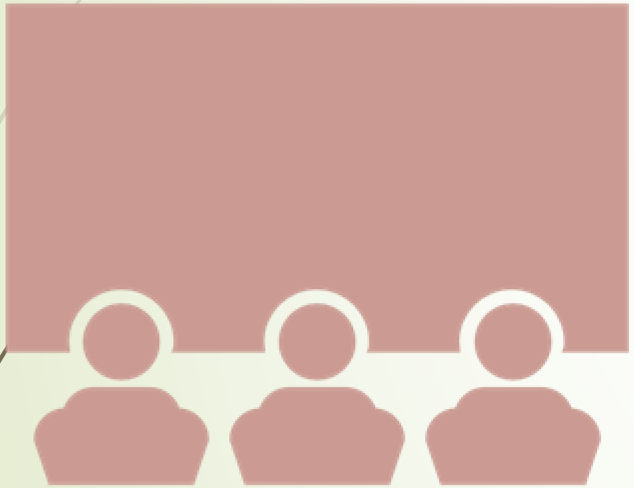
Data Science Capstone project

<Ajith Kumar>

<xx.August.2021>

Outline



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

Executive Summary

Analyzing data for SpaceX rocket launches to predict possible configurations that could lead to a successful recovery of Stage 1 of the rocket. This is made possible via data collection, data wrangling, data visualization and machine learning algorithms for prediction.

- Factors that influence the successful recuperation of Stage 1 are, the Orbit where the payload is delivered to, Payload mass, possibly also the Launch site.
- Major classification methods of machine learning are able to accurately predict successful recuperation.

Introduction

- Space Exploration Technologies Corp. or SpaceX is an American aerospace manufacturer, space transportation services and communications company. We in this project focus on the space transportation side of it mainly the Falcon 9 rockets.

- The goal of the project is to predict if Falcon 9 rocket's first stage will land successfully, helping the company to keep the cost low by enabling re-use.

- Github link: https://github.com/ajithkumars/dspython
- All resources related to the project can be found in the link above

Methodology

- Data is collected via API provided by SpaceX and from Wikipedia page for Falcon 9 launches.
- We perform data wrangling by looking for patterns, simplifying and enriching the data
- Further exploratory data analysis is performed using SQL.
- Interactive analysis is done using Plotly Dash and Folium maps
- Predictive analysis is performed using classification models

Methodology

Data collection

- We primarily use publically available data from
  - url: https://api.spacexdata.com/v4/launches/past
  - Packages used: requests, pandas and numpy

- We also use Wikipedia page link and web scrapping for alternatives for data and comparison
  - url: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
  - Packages used: requests, beautiful soup, pandas

- Helper functions are used to extract exactly the data required

Data collection : SpaceX Api

- We primarily use publicly available data from
  - url: https://api.spacexdata.com/v4/launches/past
  - Packages used: requests, pandas and numpy

- Requests package is used to access the data via REST API (json format).
- Pandas and numpy packages are used to clean and process the data
  - Numpy is used to fill up null values if any.
  - Pandas is for table like processing of data
  - Lastly processed data is saved as CSV file for further processing

- Notebook link : https://github.com/ajithkumars/dspython/blob/main/Data-Collection.ipynb

Requests Package
URL
RESTAPI CALL

Pandas, numpy Package
- Clean
- Process data

Pandas Package
- CSV
- Conversion

CSV File

Data collection : Webscraping

- We also use Wikipedia page link and web scrapping for alternatives for data and comparison
  - url: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
  - Packages used: requests, beautiful soup, pandas

- Requests package is used to access the data via REST API. (html format)
- Beautiful soup package here is used to process html data
- Pandas and numpy packages are used to clean and process the data ()

- Notebook link : https://github.com/ajithkumars/dspython/blob/main/Webscraping.ipynb

Data Wrangling

- Mainly, pandas and numpy packages/libraries are used.
  - We classify landing outcomes mainly as a new column (success : 1 and failure : 0)
  - We count different interesting columns to look for different patterns.
    - Including launch sites, orbit etc.
    - Launch outcomes, success rate for entire data set etc.
  - We identify different possible outcomes

- Notebook link : https://github.com/ajithkumars/dspython/blob/main/Data-Wrangling.ipynb

EDA with data visualization

- We use data visualizations to get insights to correlation/relationship between different features within the data set. We mainly plot scatter plots of:
  - Flight number vs payload mass (with landing outcome as hue)
  - Flight number vs Launch site (with landing outcome as hue)
  - Payload vs Launch site
- Bar chart for Orbit type and their success rates.
- Line plot for success over the years
- Finally, OneHotEncoder is applied to features that we are interested in to create data set that Prediction can be applied on.

- Notebook link : https://github.com/ajithkumars/dspython/blob/main/Exploratory-dataviz.ipynb

EDA with SQL

- SQL is a powerful query language that has been extensively used in accessing data from database in numerous different ways.
- We use the same for some quick summarization of spacex data.
  - Distinct launch sites
  - Summation of payload mass, average mass for given customer etc.
  - Success and failure rates for given booster version etc.
  - First successful landing on particular location (E.g. ground pad)
  - Nested queries with sub queries to get information like booster versions that have successful landing and has payload between 4000 and 6000 kg.
  - Boosters that carried maximum load.

- Notebook link : Has been removed since it needs to link to IBM database which cannot be open.

Build an interactive map with Folium

- We use interactive map from Folium and its packages to get some geographical insights related to rocket launches. These include answering questions like:
  - Where are launch sites located geographically (marked with markers)?
  - Can we mark success and failures on these?
  - Can we check and mark distance between railway, highway, sea and cities and launch sites to see if there is a pattern?

- Notebook link : https://github.com/ajithkumars/dspython/blob/main/Folium-Demo.ipynb

Build a Dashboard with Plotly Dash

- Plotly Dash allows users to create interactive dashboards where we can manipulate and interact with the data to get useful insights.
  - We create dashboard for success rate for different launch sites shown as a pie chart.
  - Which also includes a payload vs success scatter plot which has slider support to check different payloads a bit more closely

- Notebook link for the code of the python application: Data should be obtained from the lab.
  - https://github.com/ajithkumars/dspython/blob/master/plotly/spacex_dash_app.py

Predictive analysis (Classification)

- Using the previously processed dataset we perform some analysis to get the best machine learning algorithm to predict successful outcomes for Falcon 9. In the process we try algorithms like:
  - Logistic Regression
  - Support Vector Machine (SVM)
  - Decision Tree Classifier
  - K Nearest Neighbors
- We perform a grid search across multiple configurations to obtain the best possible parameters for each of these algorithms to get the best possible model.

- Notebook link: https://github.com/ajithkumars/dspython/blob/main/Machine-Learning.ipynb

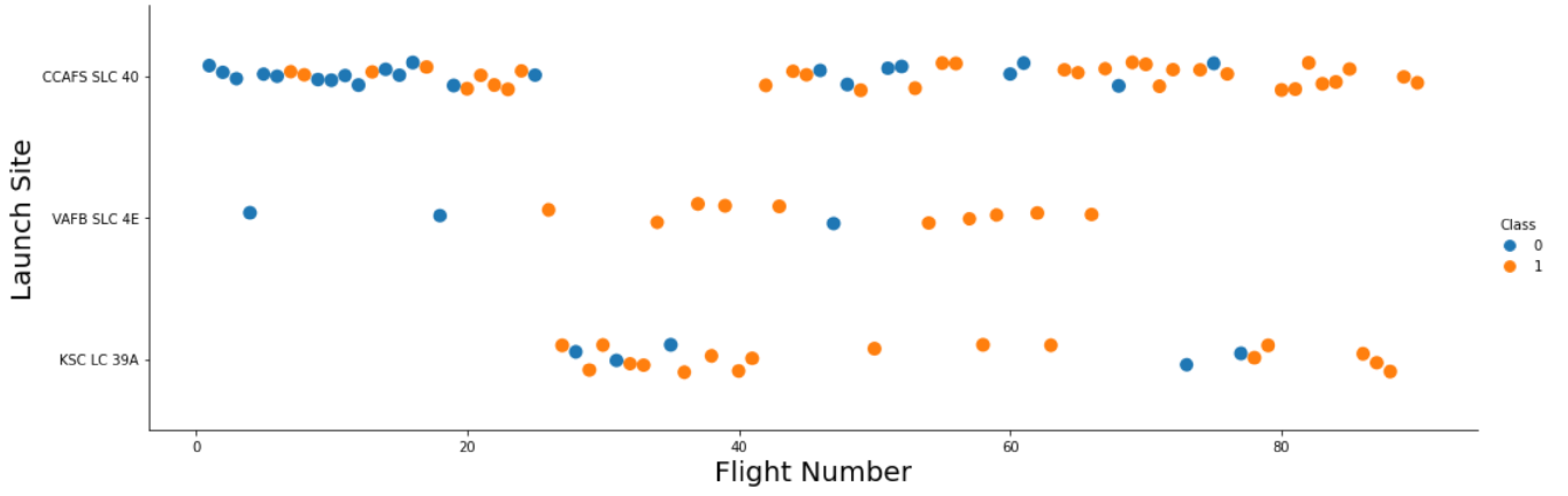
Results

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

EDA with Visualization

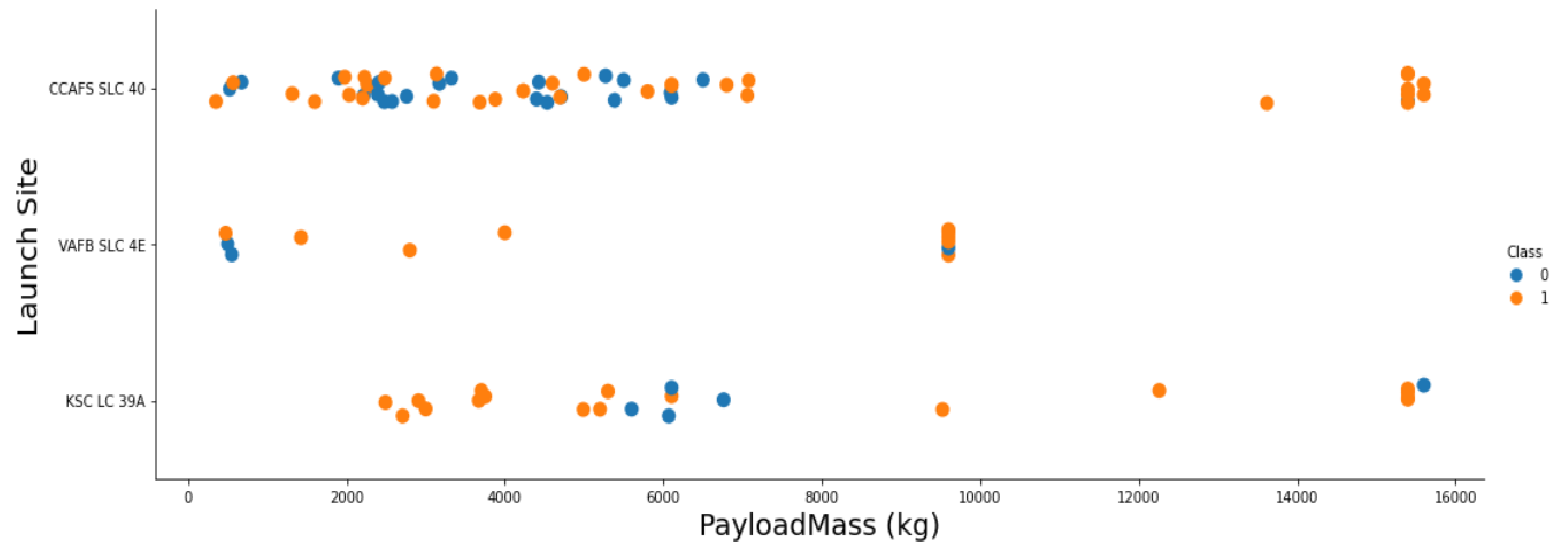
Exploratory Data Analysis with visualizations to back discussions

Flight Number vs. Launch Site



- The latter flight numbers are increasingly successful.

Payload vs. Launch Site



- Heavyier payload are more likely to be successful

	PayloadCategory	SuccessRate
0	Heavy >= 12000 kg	86.666667
1	Light <6000 kg	61.818182
2	Medium 6000-12000 kg	65.000000

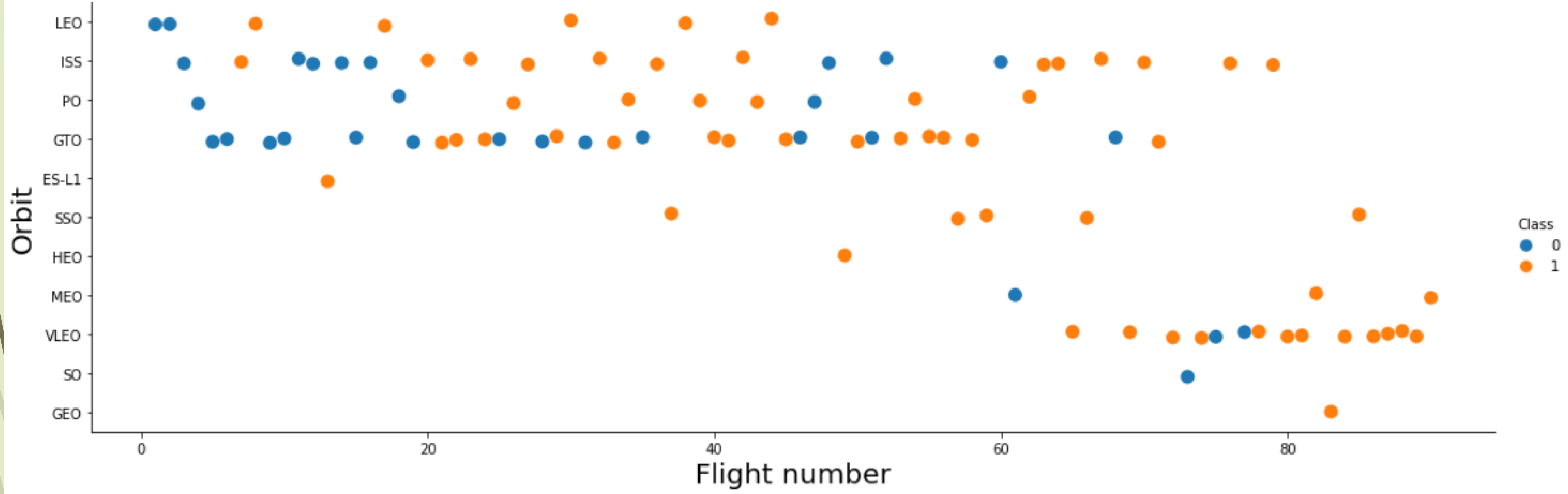
20

Success rate vs. Orbit type

	Orbit	SuccessRate
0	ES-L1	1.000000
1	GEO	1.000000
3	HEO	1.000000
9	SSO	1.000000
10	VLEO	0.857143
5	LEO	0.714286
6	MEO	0.666667
7	PO	0.666667
4	ISS	0.619048
2	GTO	0.518519
8	SO	0.000000

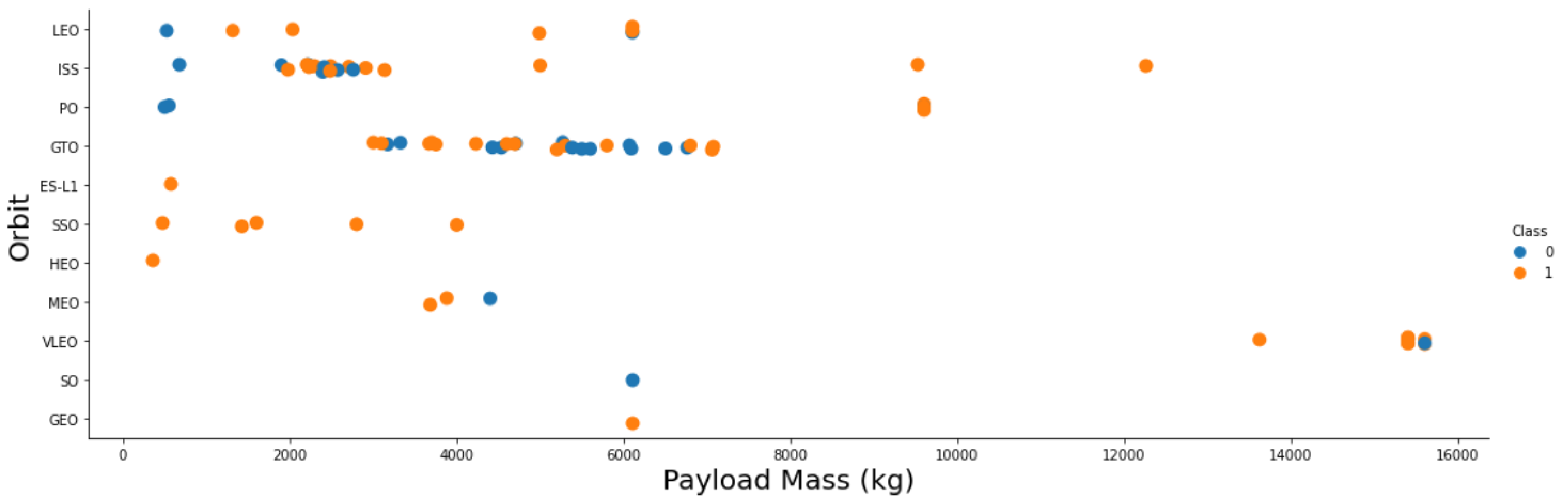
- Launches to orbits ES-L1, GEO, HEO and SSO are more likely to succeed than the others

Flight Number vs. Orbit type



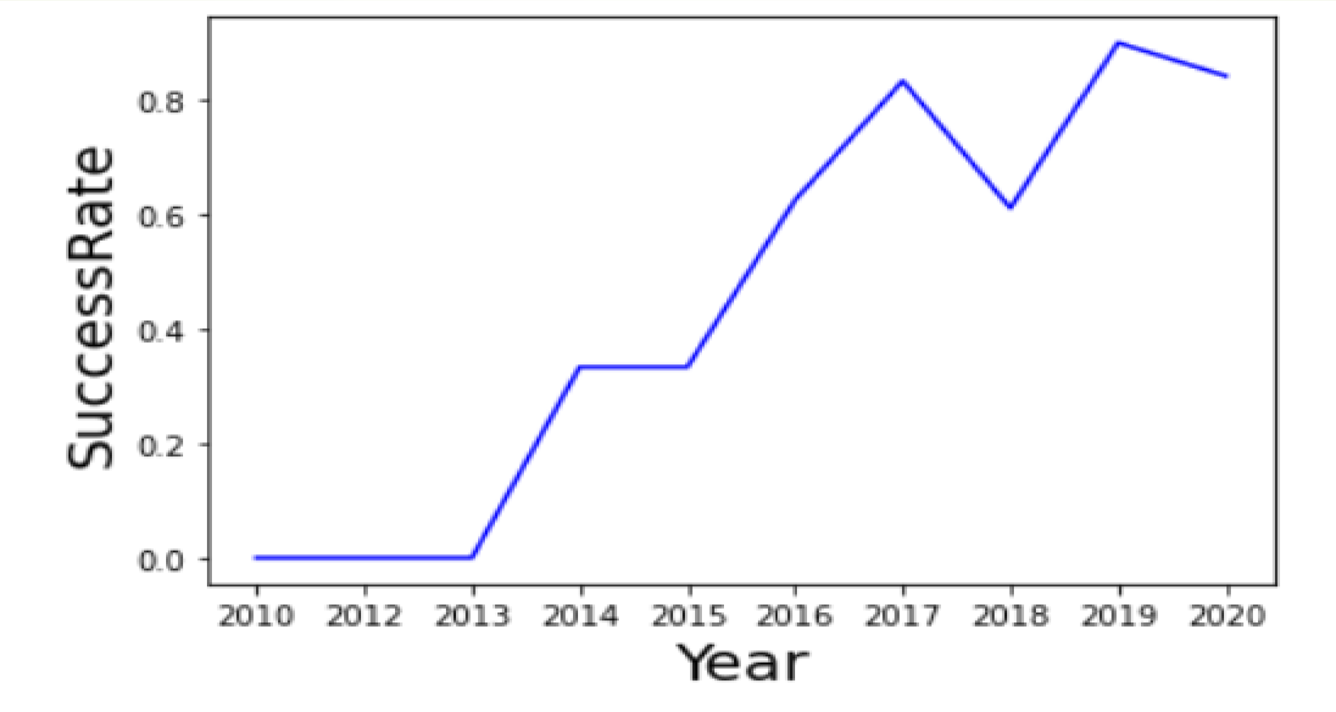
- LEO has more success in latter flight numbers
- Rest of them are rather arbitrary

Payload vs. Orbit type



- Heavy payloads have negative influence on GTO orbits and positive on LEO and ISS

Launch success yearly trend



- Over the year success rates have improved (with a small dip in 2018)

24

EDA with SQL

All launch site names

- Query:
  - %sql select distinct(launch_site) from SPACEXDATASET;
  - Selects distinct values for launch_site column

- Launch sites (listed) – CCAFS SLC-40 has been repeated here due to possible miss entry:

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

Launch site names begin with `CCA`

- Query:
  - %sql select distinct(launch_site) from SPACEXDATASET where launch_site LIKE 'CCA%';
  - Out of the selected distinct columns we match regular expression where launch_site name starts with CCA

- Launch Sites: Same issue as last slide.

CCAFS LC-40
CCAFS SLC-40
CCAFSSLC-40

Total payload mass

- Query:
  - %sql select sum(payload_mass__kg_) from SPACEXDATASET where customer like 'NASA (CRS)';
  - Sums all the elements in column for payload mass where the rows are selected based on customer name
- Total payload mass for all NASA (CRS) related launches : 45596 kg

Average payload mass by F9 v1.1

- Query:
  - %sql select avg(payload_mass__kg_) from SPACEXDATASET where booster_version like 'F9 v1.1%'
  - Sums all the elements in column for payload mass and presents the average, with condition set for booster_version with regular expression matching for F9 v1.1 version.
- Average payload was found to be 2534 (so low payload mass is carried by this booster version)

First successful ground landing date

- Query:
  - %sql select min(DATE) from SPACEXDATASET where landing__outcome like 'Success (ground pad)'
  - Minimum date where the landing_outcome condition of success and on gournd pad is met.

- First successful ground pad landing was achieved on 22nd December 2015.

Successful drone ship landing with payload between 4000 and 6000

- Query:
  - %sql select distinct(booster_version) from SPACEXDATASET where landing__outcome like 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4000 AND 6000;
  - (distinct) Booster versions selected where two conditions based on landing outcome and payload mass are met.

- Result:

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

Total number of successful and failure mission outcomes

- Query:
  - %sql select mission_outcome, count(*) from SPACEXDATASET GROUP BY mission_outcome
  - Count mission_outcome categories

- Result: Almost all mission outcomes have been a success except 1.

Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters carried maximum **payload**

- Query:
  - %sql select distinct(booster_version) from SPACEXDATASET where payload_mass__kg_ in (select max(payload_mass__kg_) from SPACEXDATASET);
  - A nested query, with subquery that finds out the maximum payload mass and uses in the main query to get booster versions(distinct) that carried it.
- Results: ➔

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

2015 launch records

- Query:
  - %sql Select MONTHNAME(DATE) as month,booster_version,landing__outcome,launch_site from SPACEXDATASET where landing__outcome like 'Failure%' AND to_char(DATE,'yyyy')='2015'
  - Selects date which is presented by month name and other expected columns based on failed attempts in year 2015
- Result: 2 failures noticed for the year of 2015 from the same launch site.

MONTH	booster_version	landing__outcome	launch_site
January	F9 v1.1 B1012	Failure (drone ship)	CCAFS LC-40
April	F9 v1.1 B1015	Failure (drone ship)	CCAFS LC-40

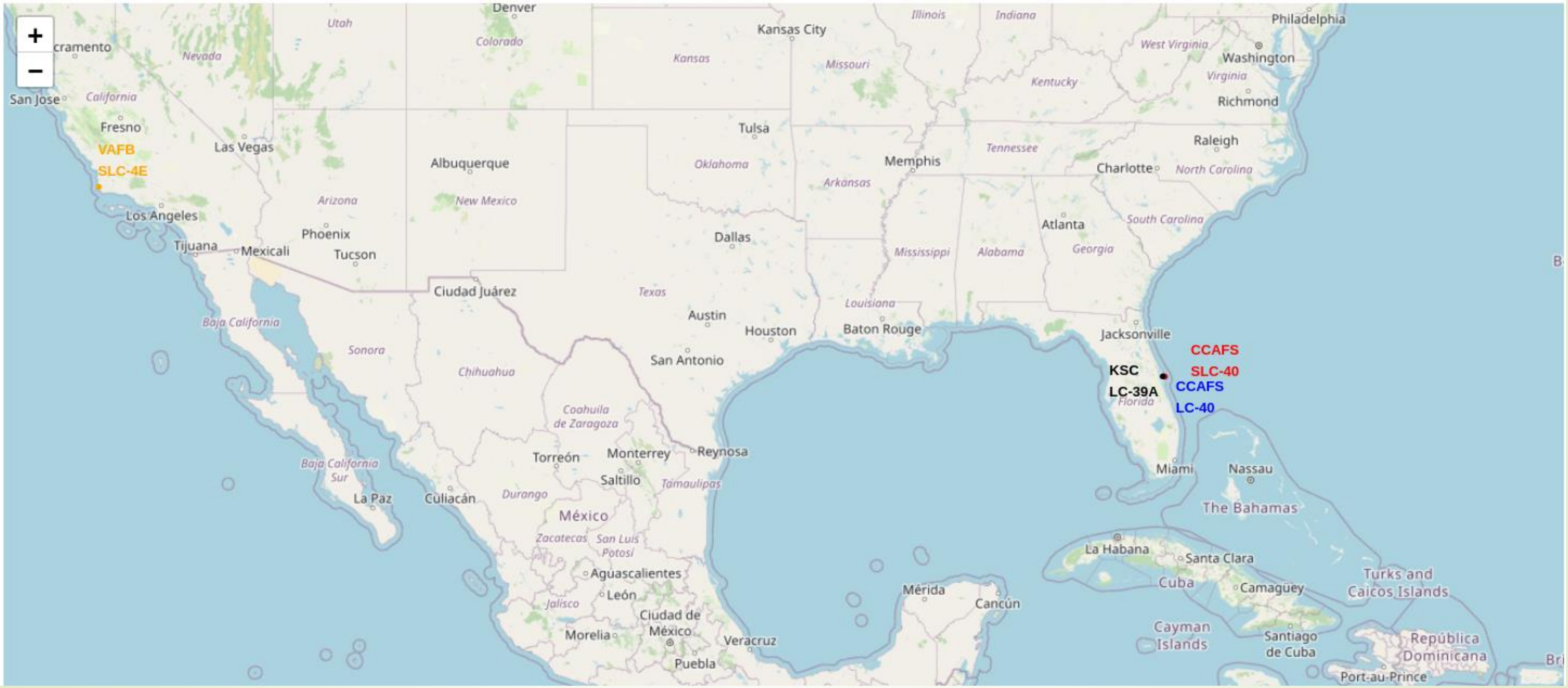
Rank success count between 2010-06-04 and 2017-03-20

- Query:
  - %sql select landing__outcome,count(landing__outcome) as outcomes from SPACEXDATASET where DATE between '2010-06-04' AND '2017-03-20' AND landing__outcome like 'Success%' GROUP BY landing__outcome ORDER BY outcomes desc;
  - Rank landing outcomes between given dates, achieved by counting the outcomes and grouping and ordering them.
- Result: drone ship outcomes have had more successful outcomes

landing__outcome	outcomes
Success (drone ship)	5
Success (ground pad)	3

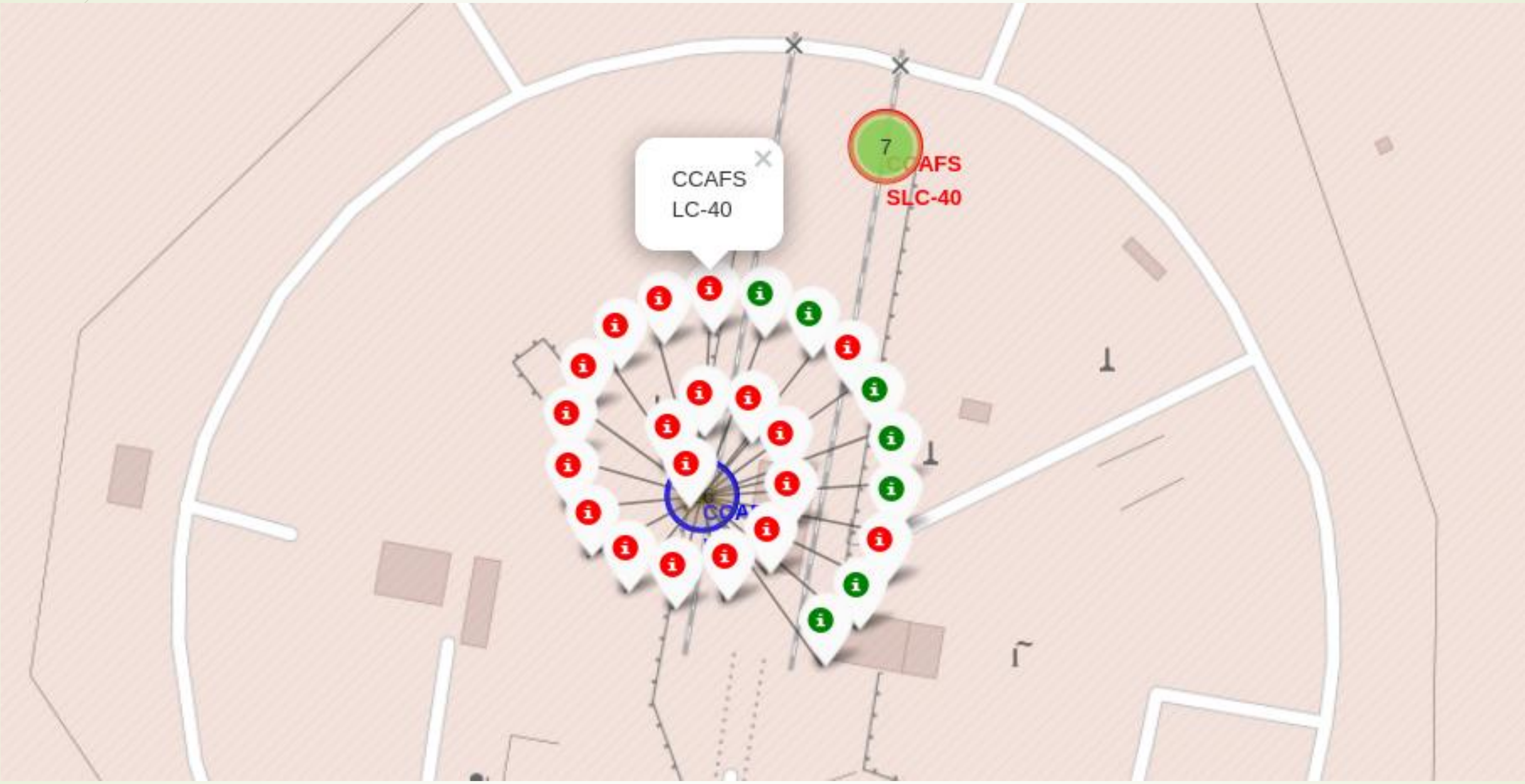
Interactive map with Folium

Launch sites



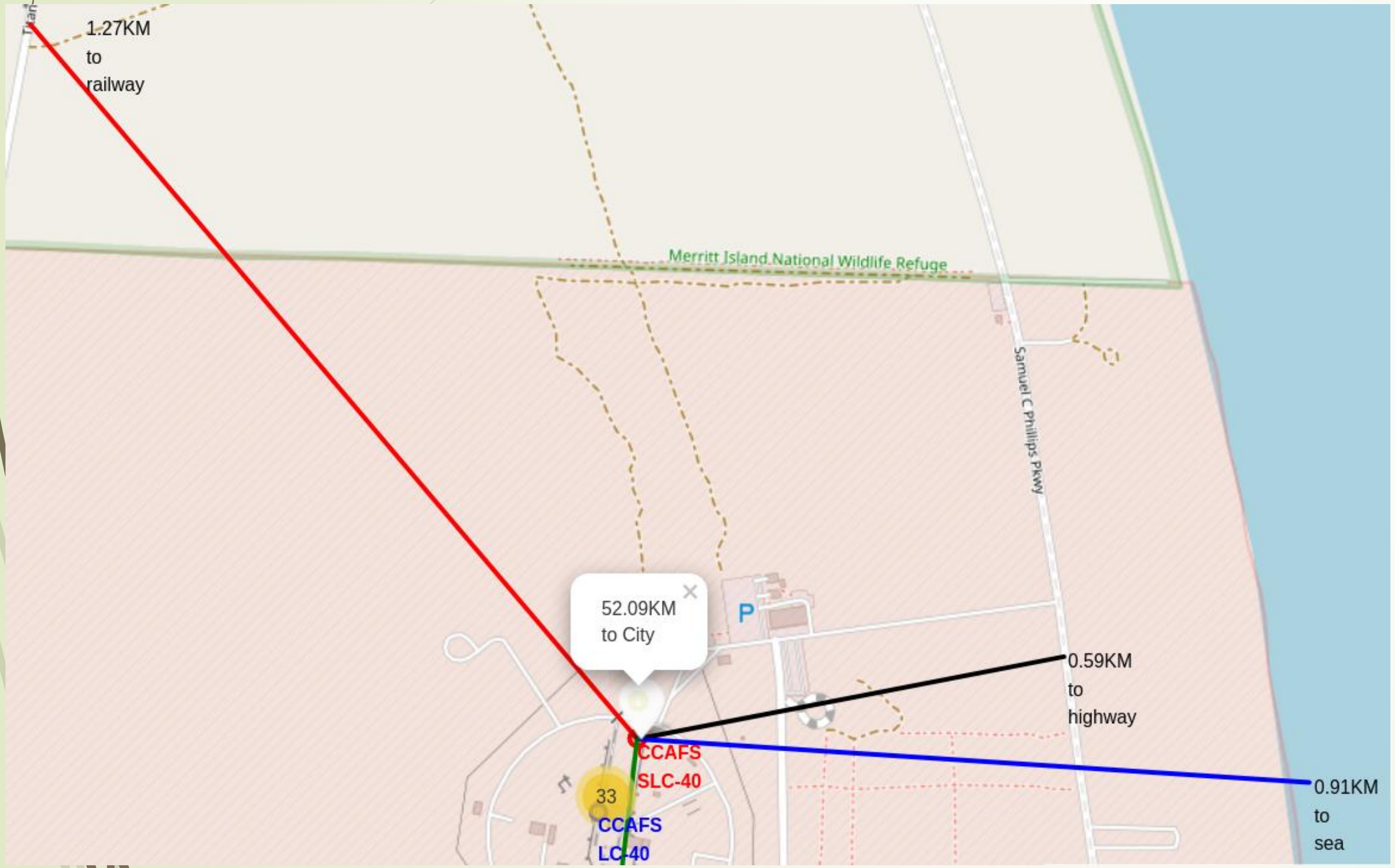
- All launch sites as you see are close to the equator and to the coast.
  - Equator: Due to speed advantages.
  - Coast: Possibly due to safety in case of failures?

Launch records by Launch Site



- Easily visible success rate of launches from the chosen site.

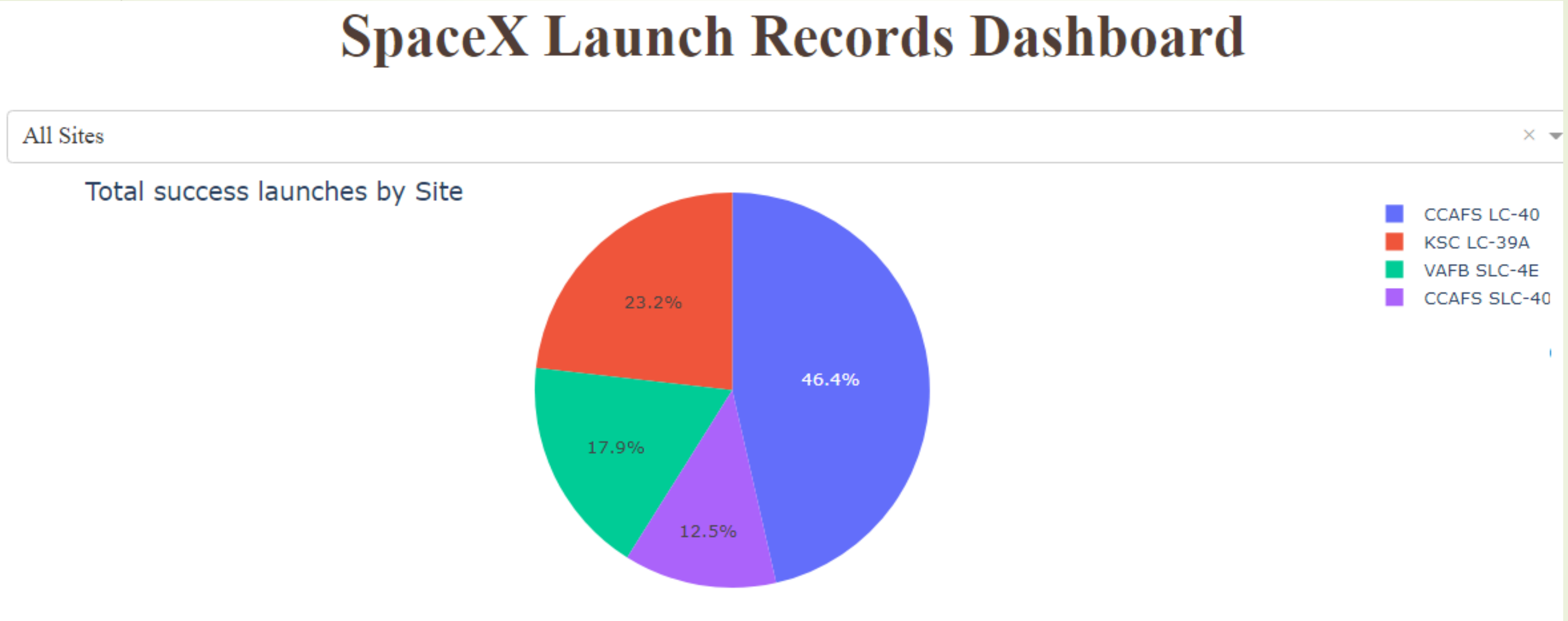
Launch site proximity



- We considered one launch site CCAFS SLC-40 and its proximity to railway, sea, highway and city.
- It does seem to have close proximity to sea, railway and highway and long distance to nearest city as listed in the popup (possibly due to safety reasons

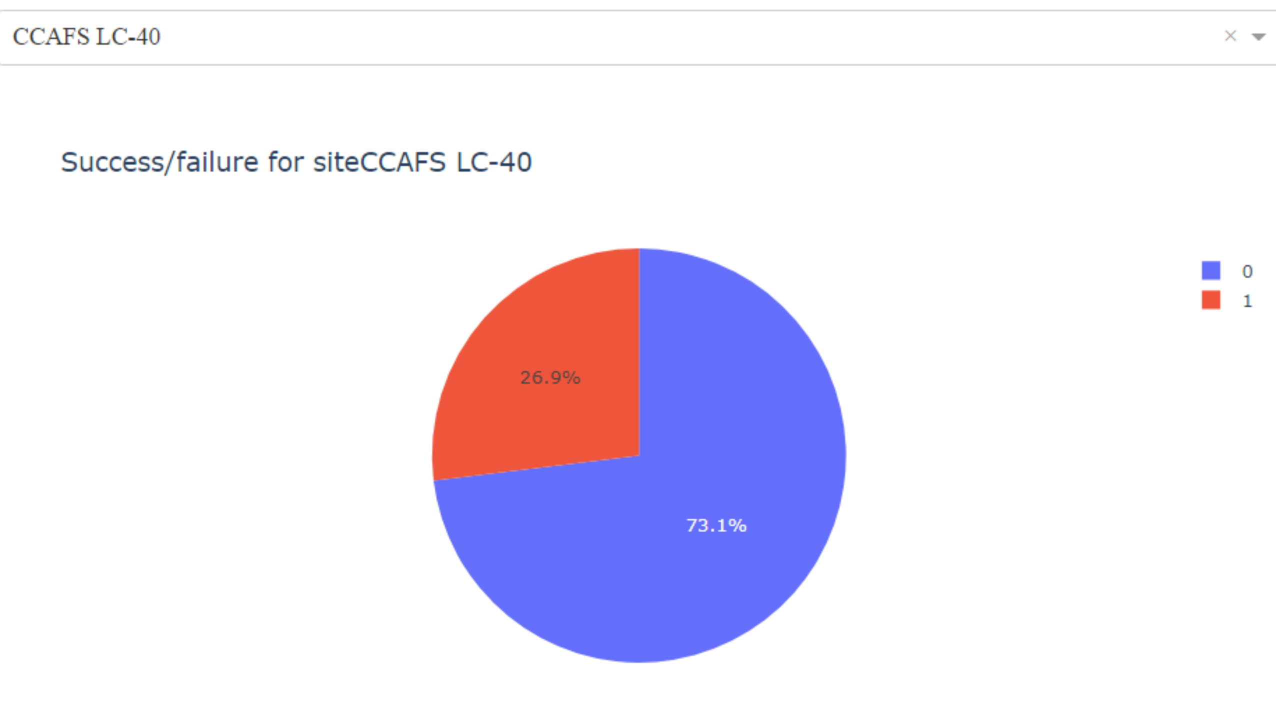
Build a Dashboard with Plotly Dash

Successful launches



- CCAFS LC-40 has the major portion of the pie with successful launches, has the most launches as well
  - CCAFS LC-40 -- 26 (Launches)
  - KSC LC-39A -- 13
  - VAFB SLC-4E -- 10
  - CCAFS SLC-40 -- 7

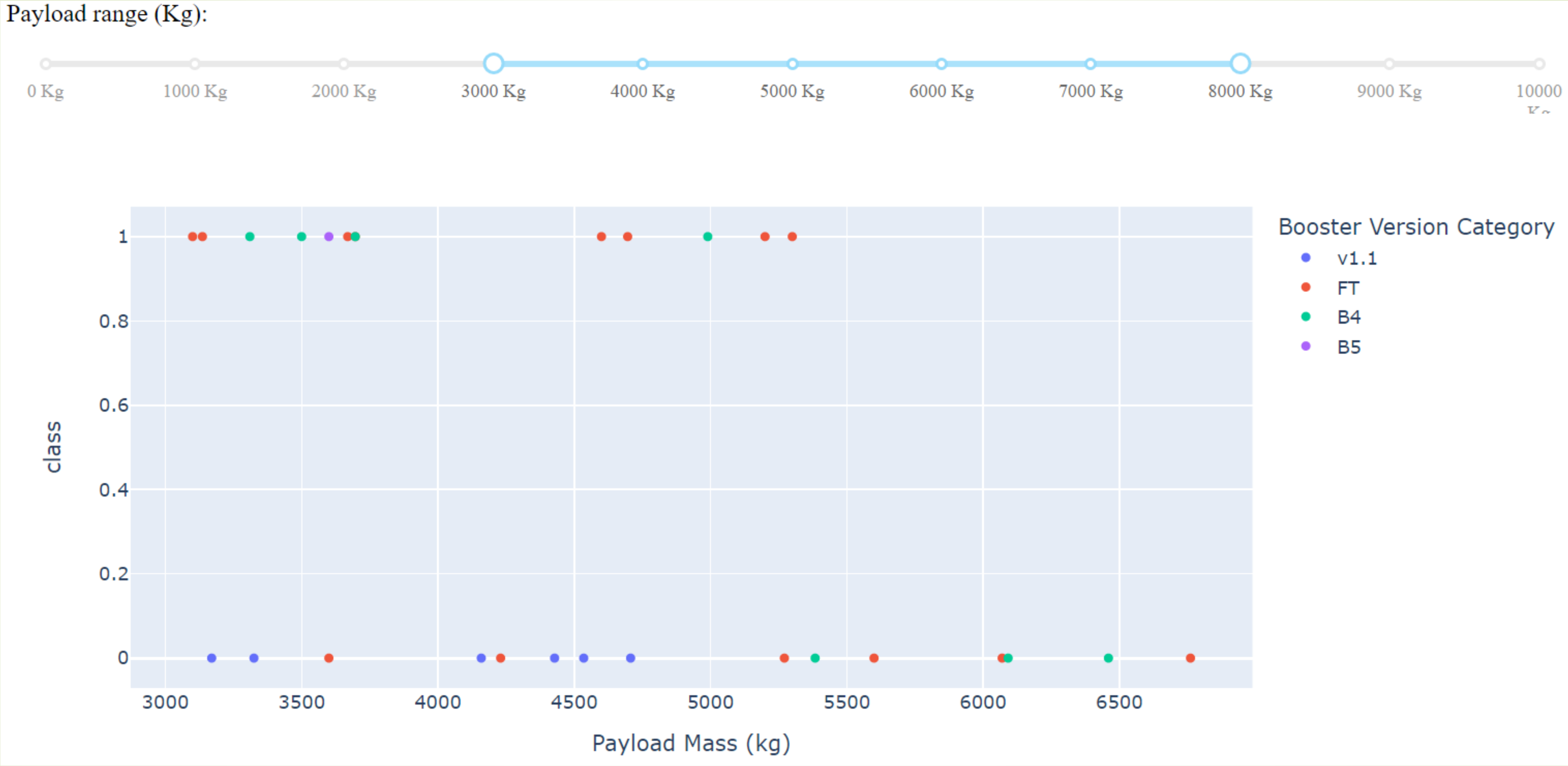
Launch outcomes for site CCAFS LC-40



	Launch Site	SuccessRate
0	CCAFS LC-40	0.269231
1	CCAFS SLC-40	0.428571
2	KSC LC-39A	0.769231
3	VAFB SLC-4E	0.400000

- CCAFS LC-40 has had a very low success rate (1) of 26,9% (unfortunate coloring)
  - 0 represents the failed launches and 1 for successful launches

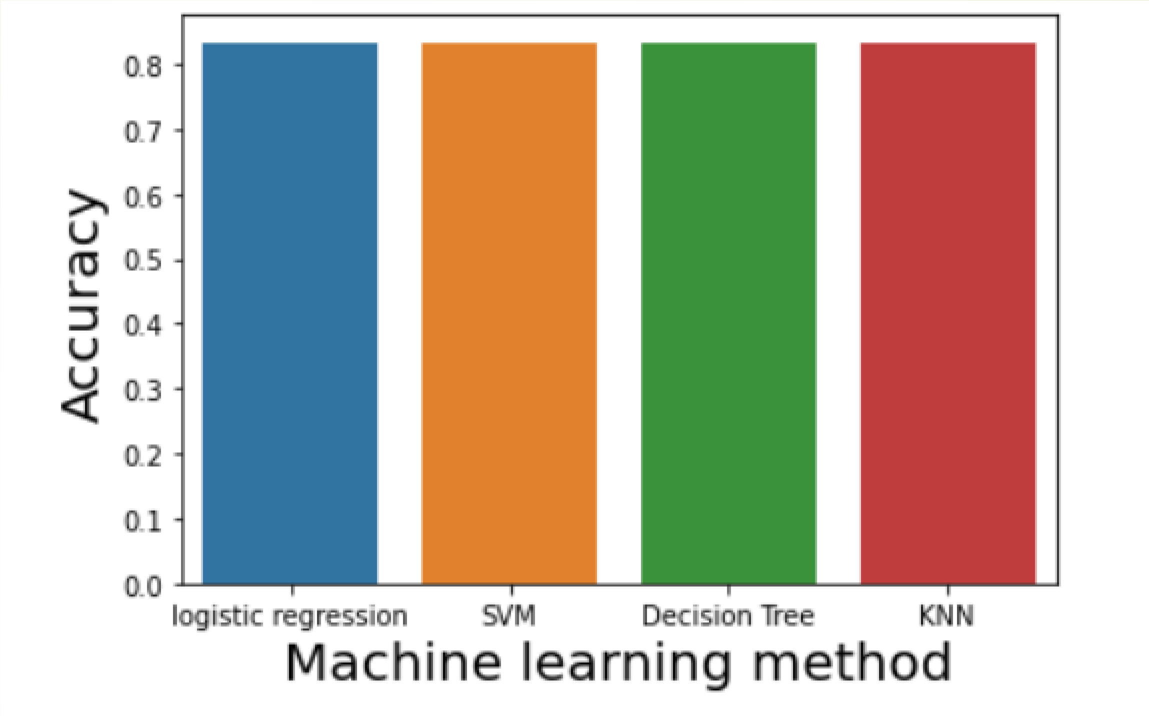
Payload vs Success



- A chart along with slider is provided to provide an interactive experience to check different combinations, here we have a range of 3000 to 8000 kg.

Predictive analysis (Classification)

Classification Accuracy



- All methods used seem to have the same accuracy

Confusion Matrix

- All methods have the same confusion matrix.
- There are three false positives as seen in the matrix (which can be an issue)

CONCLUSION

- Higher payloads have higher success rate
- Certain orbits have higher success rate
- Launch sites should have close proximity to ocean and railways
- It is possible to predict launch outcomes with fairly high accuracy (83%), thus based on machine learning it is possible to create a new company by testing what kind of features they should work to get high probability for successful outcome

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