IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Thomas Farrow
February 2024

Outline

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

Executive Summary

By learning to predict the outcome of a SpaceX rocket launch, we can assign a cost function to apply to our own company SpaceY

- We'll use the following methodologies to achieve this aim:
  - API and Webscraping to acquire data
  - Python to clean the data
  - Analysis through SQL and Visulisation
  - Interactive Visualisation with plotly dash
  - Predictive analysis
- Summary of all results
  - Newer launchers are more successful
  - larger payload mass more successful
  - Orbit choice has impact on success rate

Introduction

- Space Y wants to compete against SpaceX by analysing SpaceX's launch data
- We can determine a price for each launch by using this data

- Understand the cost function by being able to predict the recycling of the first stage rocket booster
- Choose where to launch

Section 1

Methodology

Methodology

- Data collection methodology:
  - Two sources, SpaceX API and web scrapping from Wikipedia launch information
- Perform data wrangling
  - Filtering the data for relevant rockets
  - Handling NAN errors
  - One Hot Encoding to prepare for classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

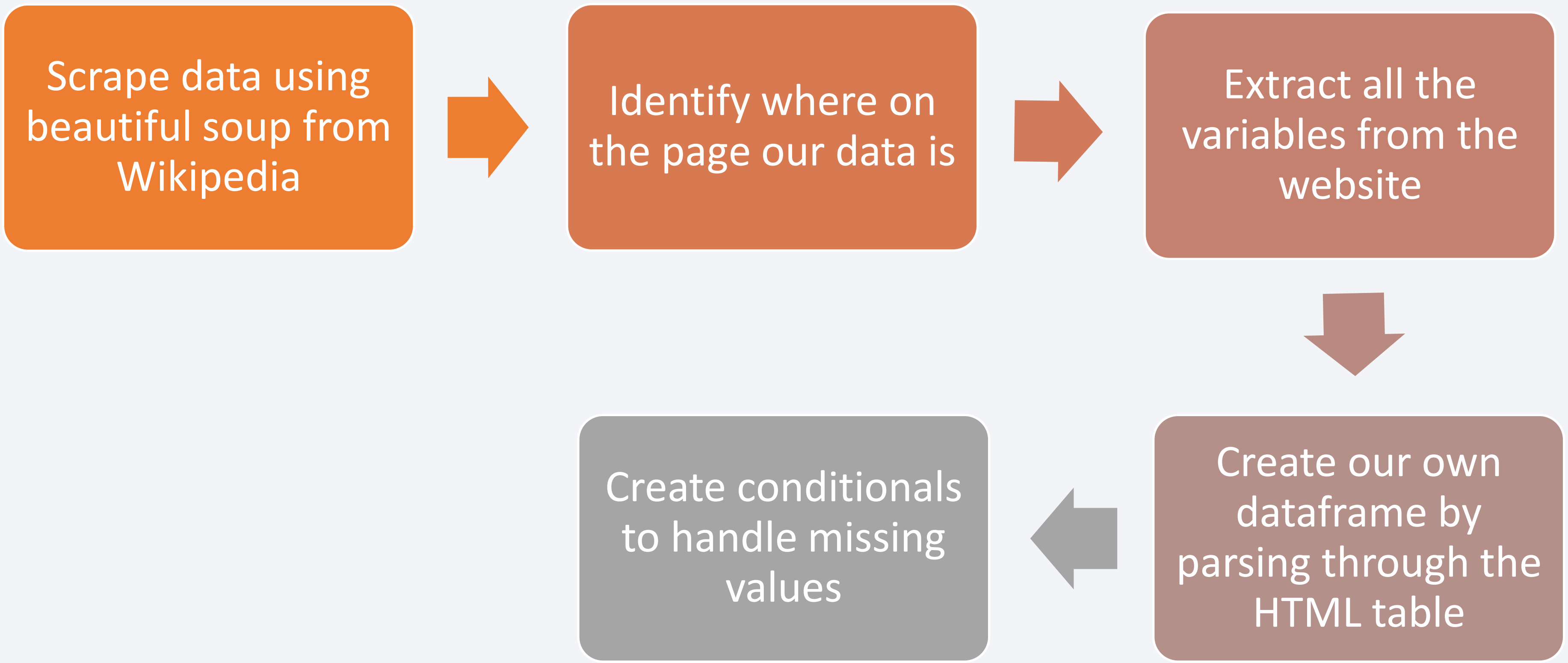
- Data collection involved a combination of API request from SPACEX API and Web scraping from Wikipedia
- Each source contained different information to aid analysis
- API:
  - Flight number, date, booster version, payloadmass, orbit, launch site etc
- Wiki:
  - Customer, launch outcome, launch site etc

Data Collection – SpaceX API

1. Requesting Rocket launch data from SpaceX API
2. Decoding the response Json using .json_normalise to create a dataframe
3. Create a subset new dataframe of specific features that help in the analysis
4. Filter to show only falcon 9 boosters
5. Handle missing values

https://github.com/Jarthz/IBM-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping



https://github.com/Jarthz/IBM-Capstone/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

- The data required several enrichments so as to analyse
- Count the number of launches per site
- Count the number of orbit occurrences
- Create a landing outcome label, having cleaned variable outcome information

https://github.com/Jarthz/IBM-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

Perform initial analysis

Calculate the number of launches on each site

Calculate the number and occurrence of each orbit

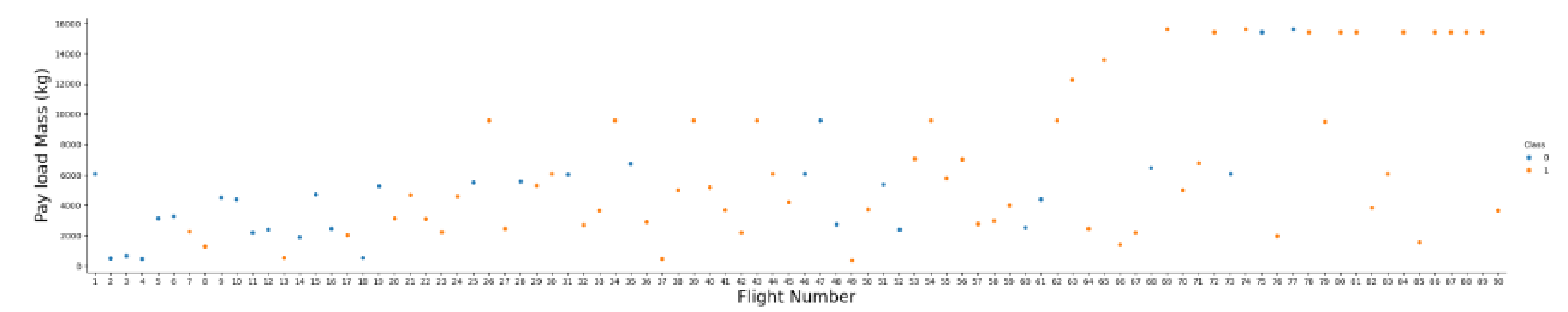
Calculate the number and occurrences of mission outcomes of the orbits

Create a landing outcome label

EDA with Data Visualization

- We created several scatter plots to show the relationship between launch site, payload mass and success over time

https://github.com/Jarthz/IBM-Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb



EDA with SQL

- Using SQL we found:
  - The names of the unique launch sites
  - 5 occurances of launch sites beginning with string CCA
  - Total payload mass carried by boosters launched by NASA
  - Average payload mass carried by booster version F9 V1.1
  - The date when the first successful landing on a ground pad was achieved
  - Names of boosters which successfully landed on a drone ship with a payload between 4000 and 6000
  - Total number of successful and failures missions
  - Names of boosters that have carried the maximum payload mass
  - Listed the failed landing outcomes to drone ships, their booster versions and launch sites names in 2015 by month
  - Ranked and counted the landing outcome between 2010 and 2017

https://github.com/Jarthz/IBM-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

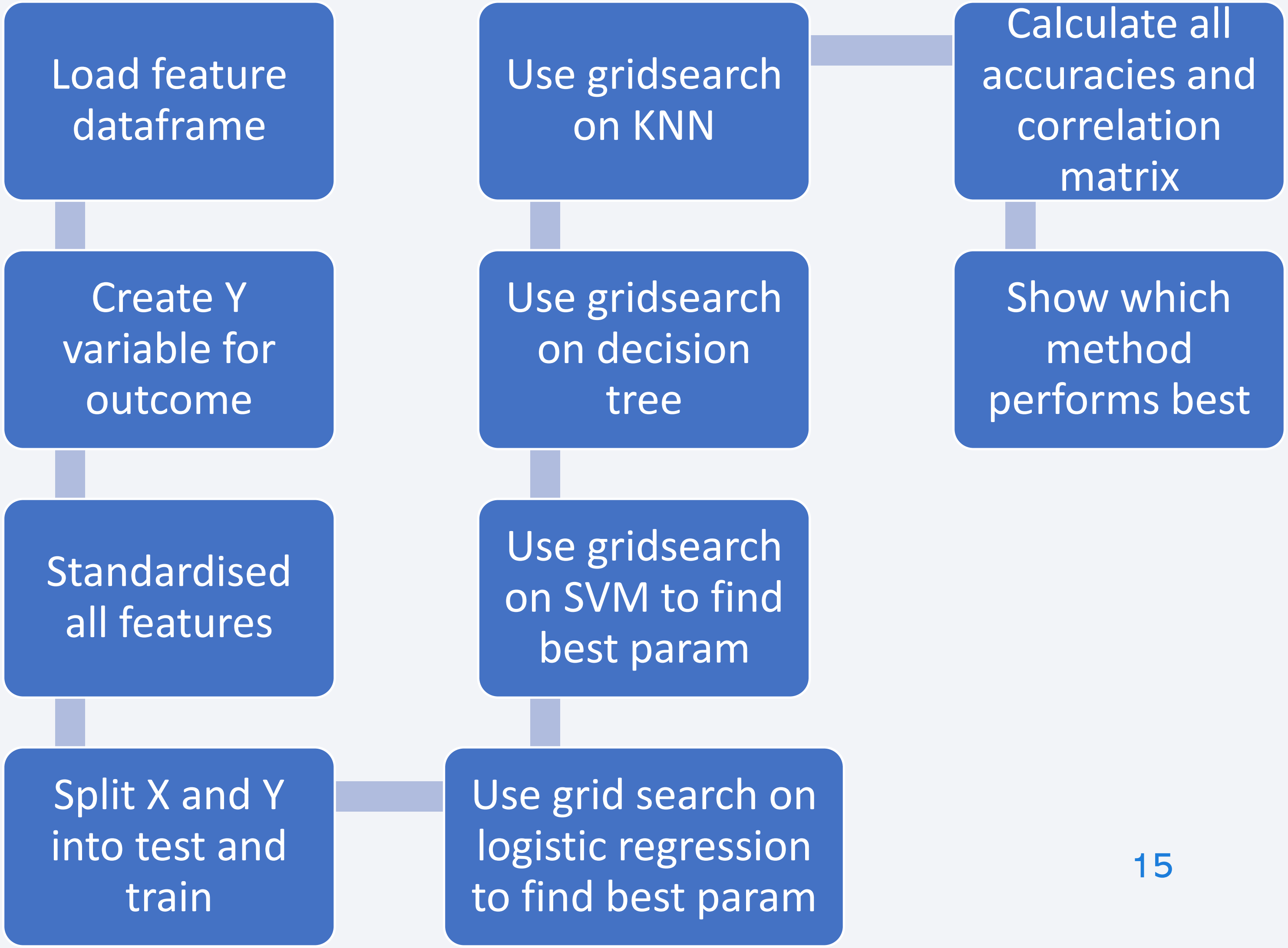
- Markers of all Launch sites:
  - Added markers with circle and popup label of all launch sites using their Lat, long to show geographical locations
- Coloured Markers for mission outcome
  - Added coloured markers for success and failed outcomes to visually identify success by location
- Distances from proximities
  - Identified the distance to the coast, population centres and transit hubs to further understand the choice of launch site location
- https://github.com/Jarthz/IBM-Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Launch site dropdown list:
  - Added a dropdown to select visualization per launch site or ALL sites
  - Created Pie chart showing success of launches per choice
  - Slider for payload mass success
  - Scatter plot showing the success outcome against payload mass choice from above slide and launch site selection
- We can now visualize our analysis to show the best site and ideal mass
- https://github.com/Jarthz/IBM-Capstone/blob/main/spacex_dash_app%20(1).py

Predictive Analysis (Classification)

- [Github]()



Results

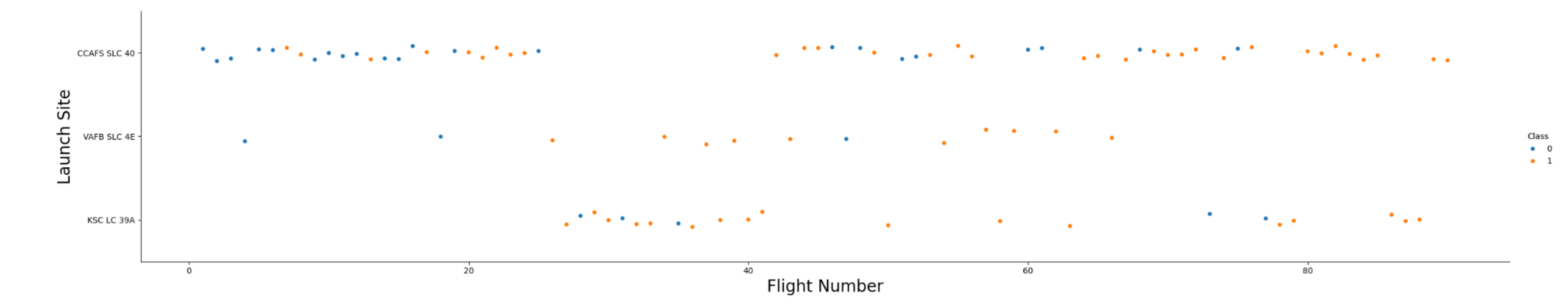
- Exploratory data analysis results

- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

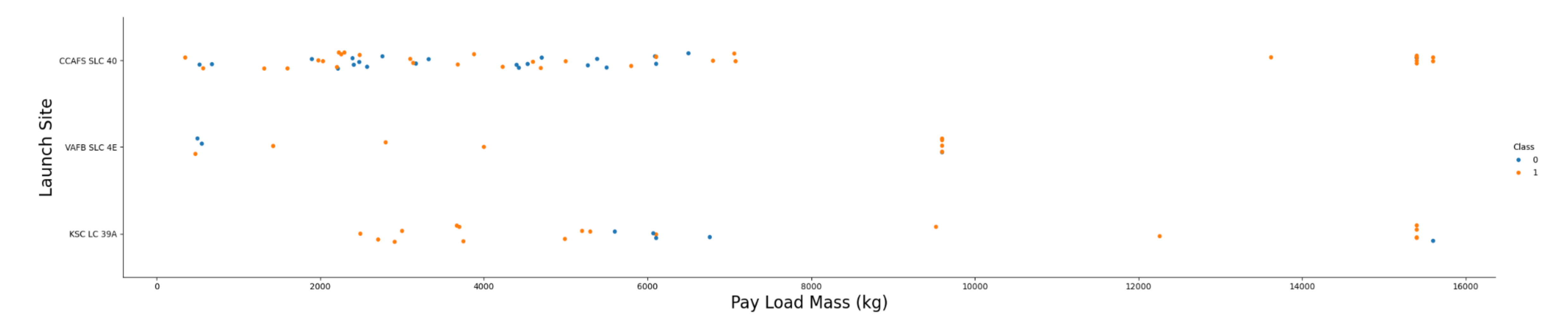
Insights drawn from EDA

Flight Number vs. Launch Site



- There's three main launch sites, with CCAFS being primary, VAFB coming online intermitantly and KSC onlined after approx. 25 launches
- Earlier launches on CCAFS were mixed, with it only being after 60+ launches before success outweighed failures
- All sites became more successful as the number of launches increased

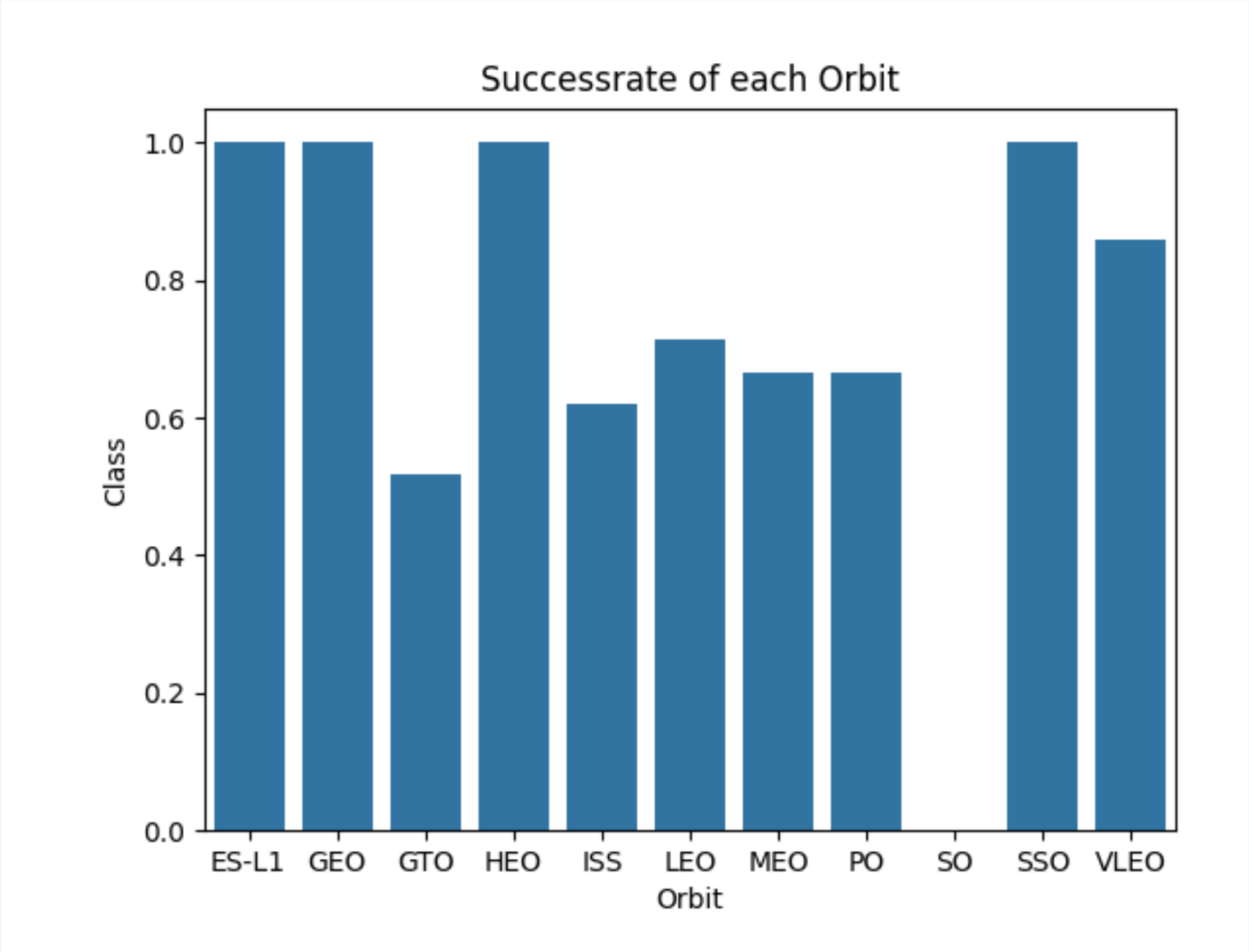
Payload vs. Launch Site



- Higher payloads seem to be more successful
- Payloads > 10000 seem only possible from 2 launch sites
- CCAFs is better at larger payloads
- KSC best with payloads <6000 >2000

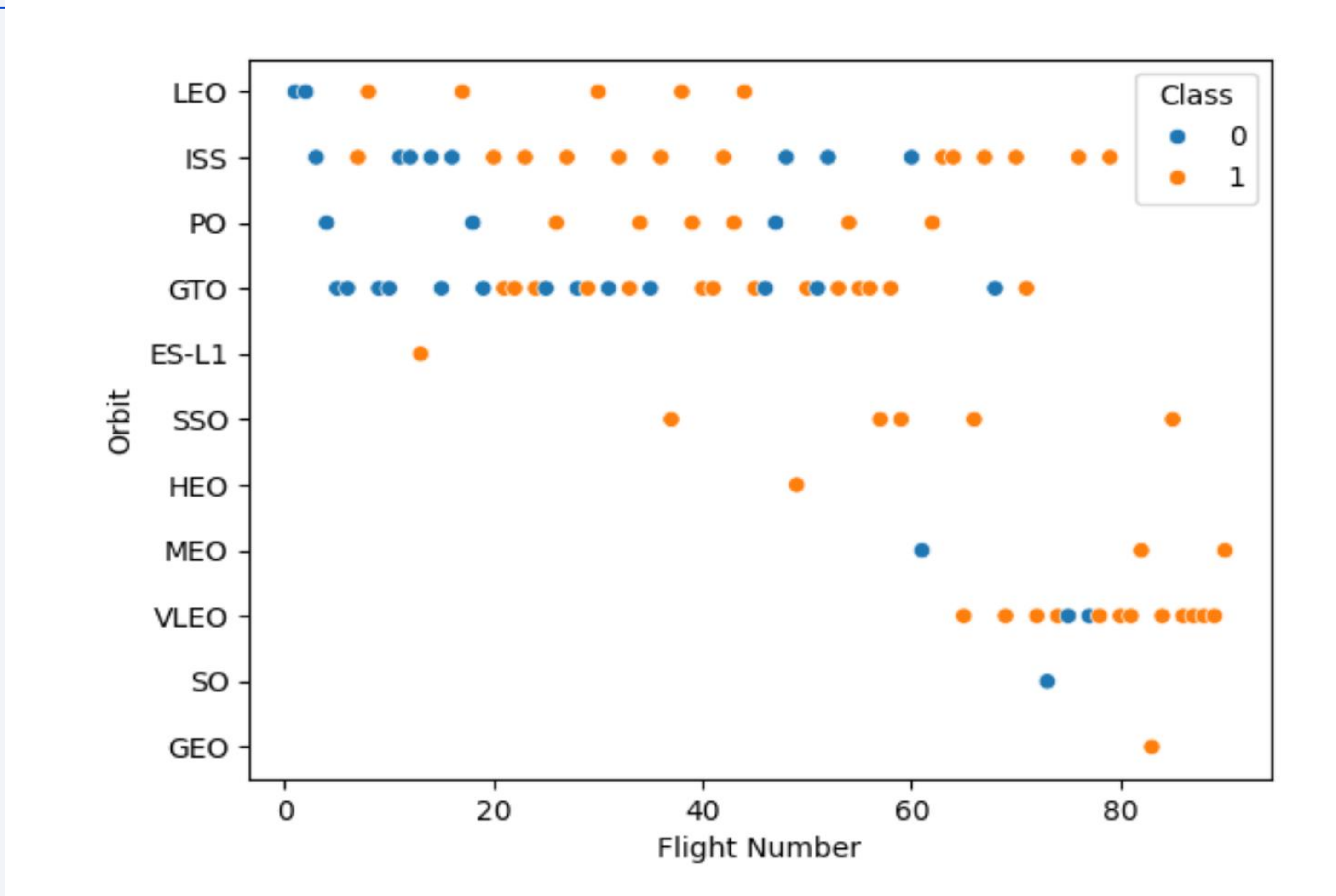
Success Rate vs. Orbit Type

- Orbit type selection has an impact on success rate with ES-L1, GEO, HEO and SSO being optimal with 100% rates
- GTO is the worst type



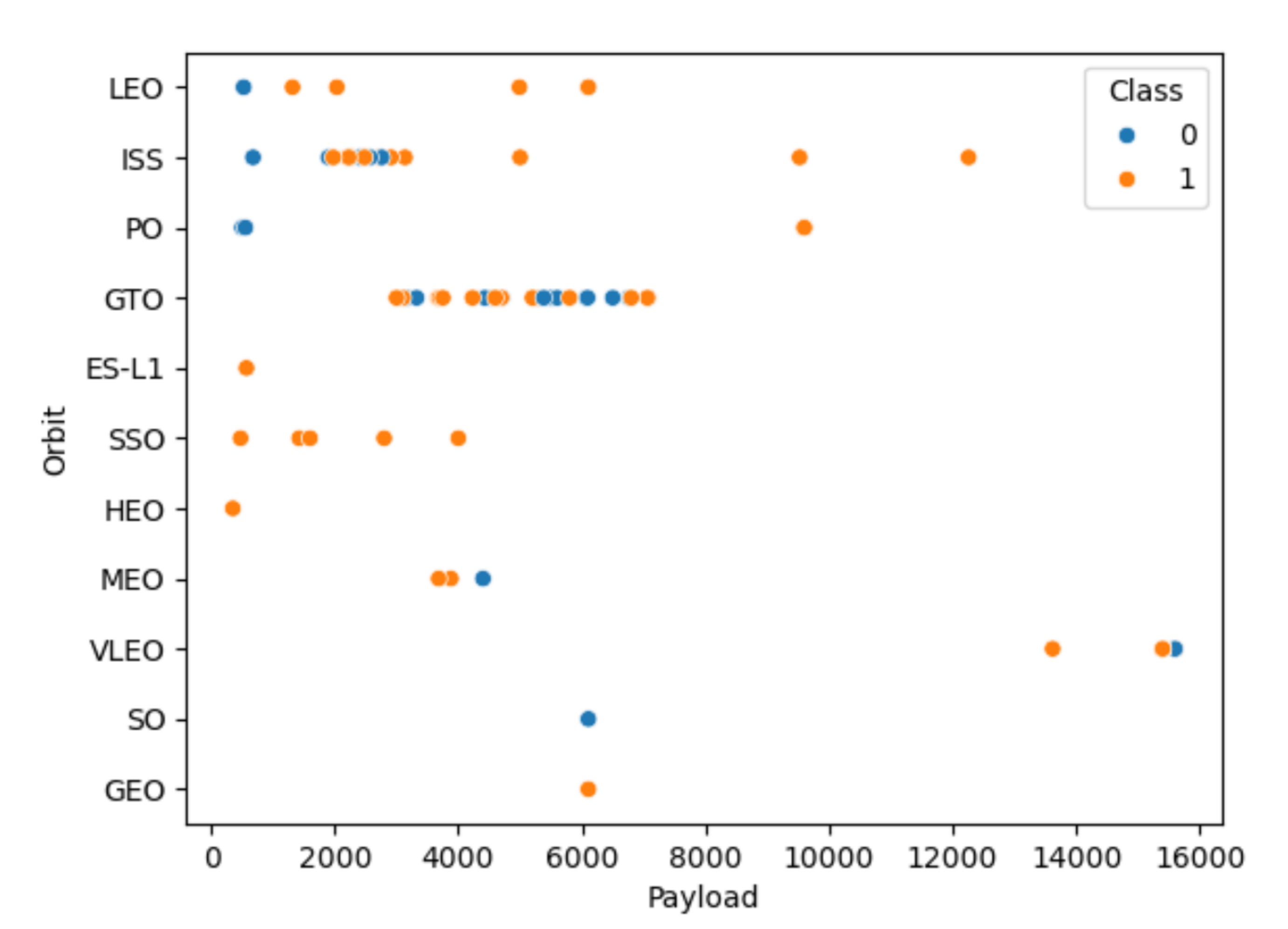
Flight Number vs. Orbit Type

- Some orbits improve with the number of flights, such as LEO
- GTO has no relationship between the number of flights and the success rate
- ES-L1 whilst being 100% successful, only has one launch so need more data



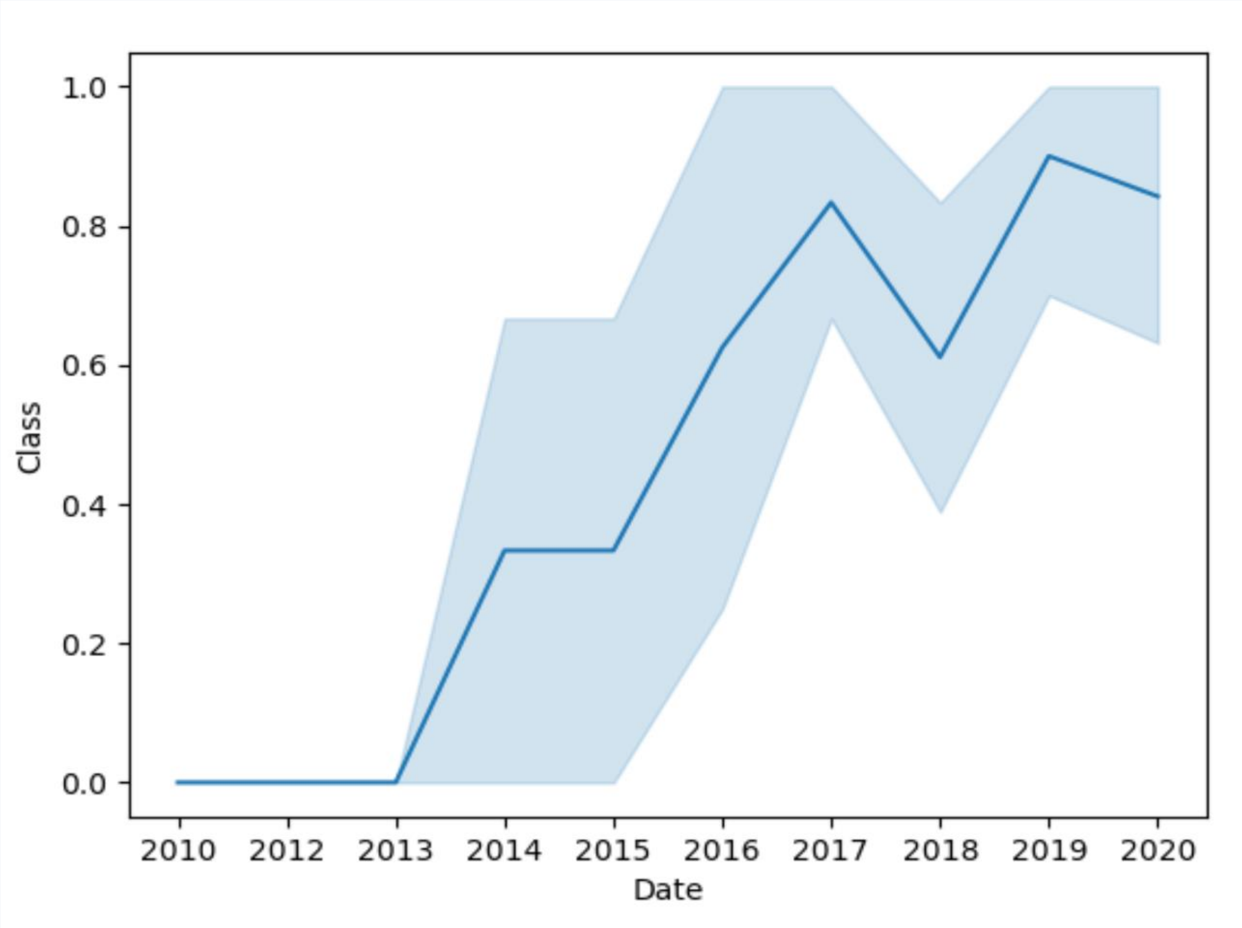
Payload vs. Orbit Type

- LEO improves with increase in payload
- SSO optimal for smaller loads
- GTO has no relationship between load time
- ISS improves with large loads



Launch Success Yearly Trend

- Like a fine wine, success improves with age



All Launch Site Names

- We query our data to show all distinct launch sites

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`

```
In [9]: %sql SELECT Launch_Site FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5

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

Out[9]:

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

Total Payload Mass

- NASA has sent a total of 45596KG through SPACEX. We can use that information to work out how big of a client they could be for us. All we need is the average payload we can send

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer = "NASA (CRS)"
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Booster Version F9 V1.1 is able to send just under an average of 3000kg

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version = "F9 v1.1"
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- The first successful ground pad out come was at the end of 2015
- This is 5 years after the first launch

```
%sql SELECT min(Date) FROM SPACEXTBL WHERE Landing_Outcome = "Success (ground pad)"
```

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

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- These boosters seem to be the most optimal for small/medium sized payloads

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

Total Number of Successful and Failure Mission Outcomes

- SpaceX is very successful, with 100:1 success rate

```
%sql SELECT Mission_Outcome, count(Mission_Outcome) FROM SPACEXTBL GROUP BY Mission_Outcome
```

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

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

Boosters Carried Maximum Payload

- These boosters can carry the maximum payload, we would need something like these to be able to compete

Booster_Version
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

- These are the boosters, launch sites and months for failed outcomes when trying to land on the drone ship

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

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

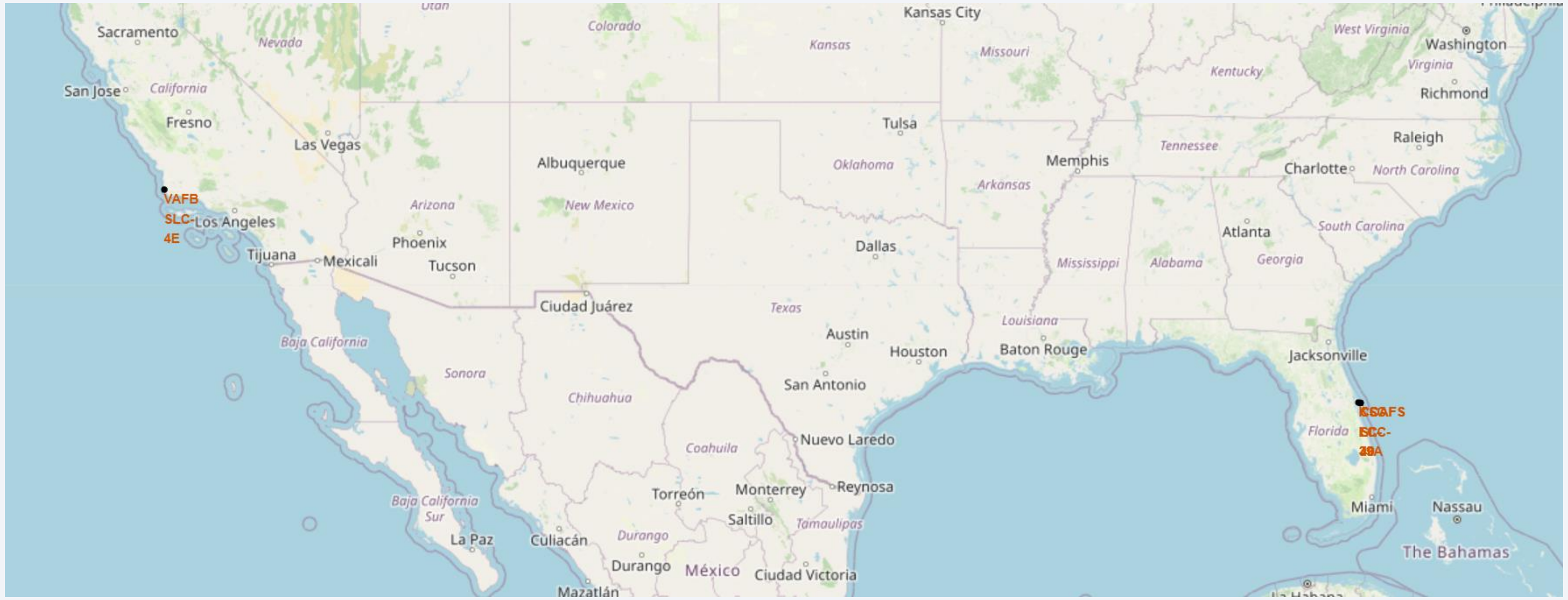
- Landing outcomes between 2010 – 06- 04 and 2017-03-20. More often than not, no attempt to recover or failed landing. There are 8 successful landing vs 23 failures

Landing_Outcome	Count
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

Section 3

Launch Sites Proximities Analysis

SpaceX launch sites



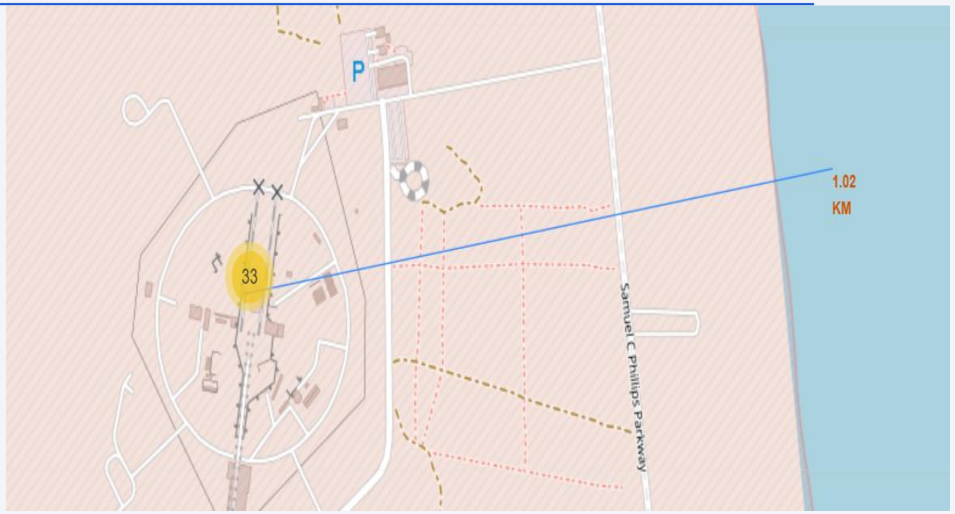
- Launch sites are near both the sea and the equator. Near the sea for safety, equator to improve success rate.
- Launch sites are near metropolises to for transportation links and talent sourcing

Successful launch sites

- Colouring launch sites by outcome allows us to visualize quickly where the optimal location is
- This helps us analyse what features this site has to make it successful

Launch site proximity to key features

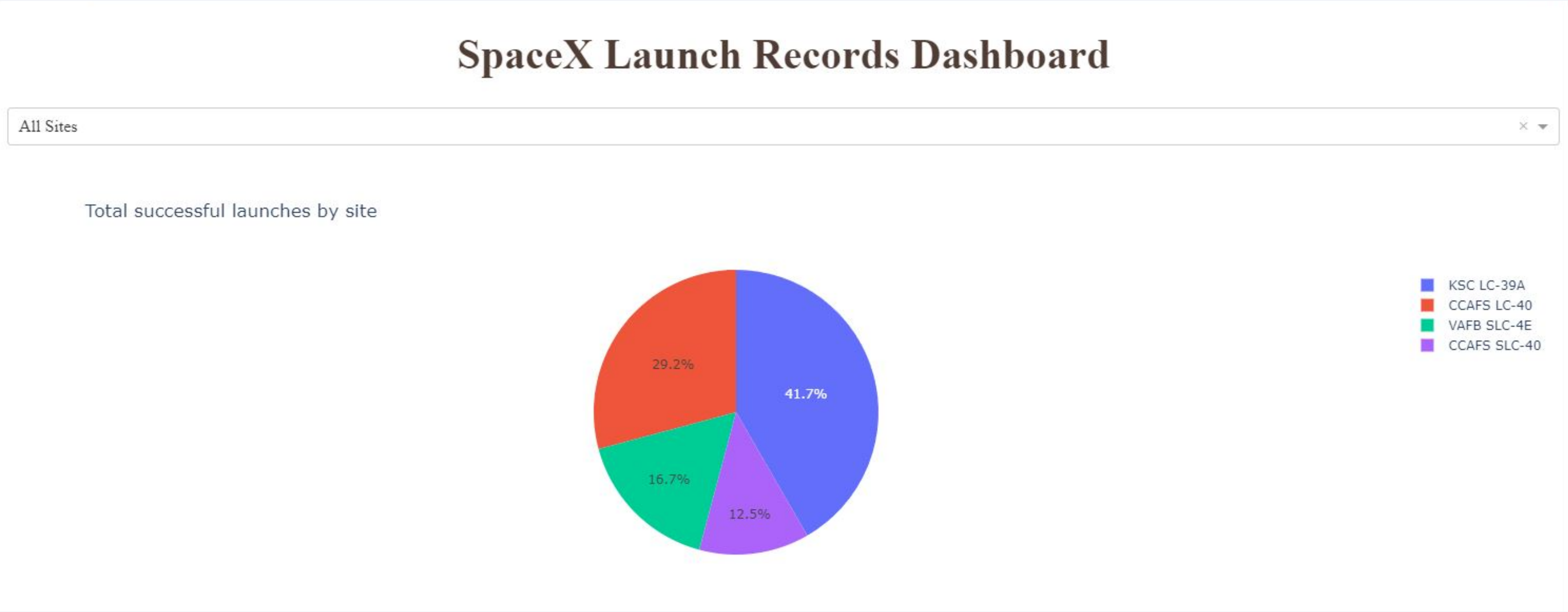
- Clsoe to the sea
- Close metropolitan centre, but far enough not to be a threat
- Close to transport links



Section 4

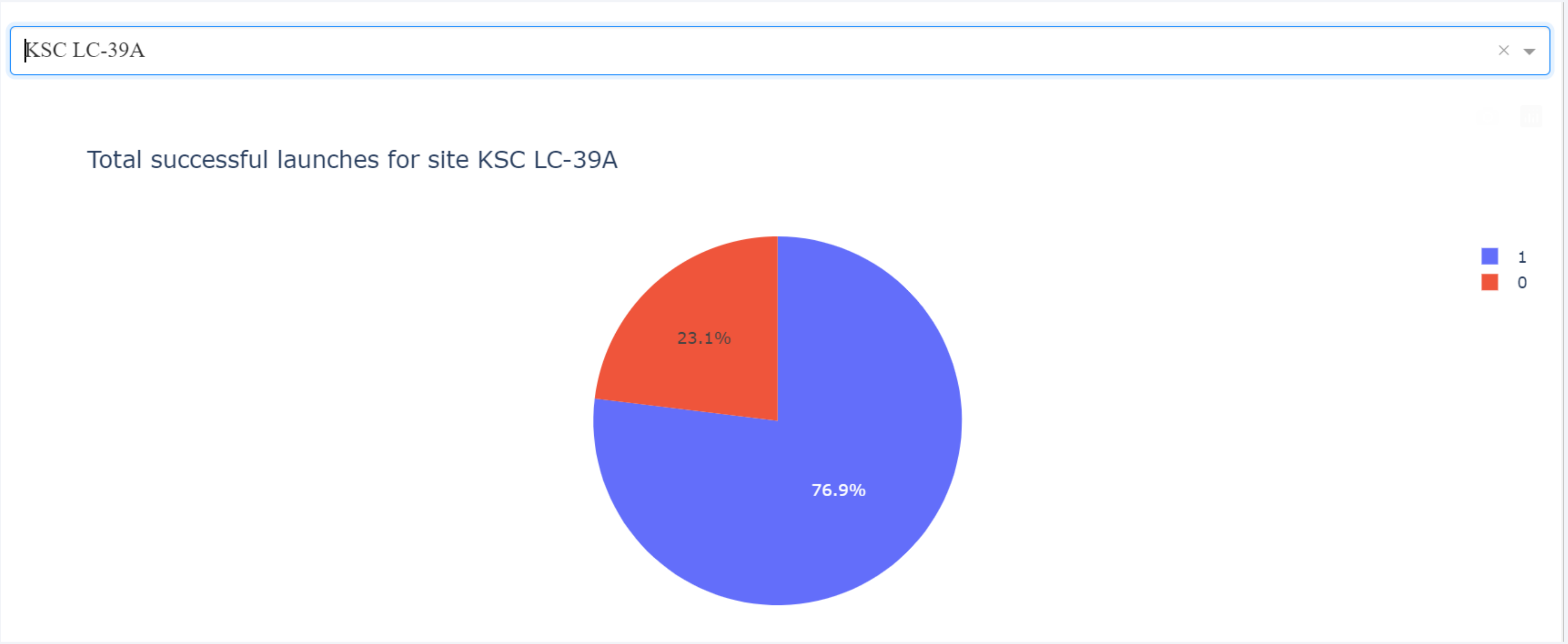
Build a Dashboard with Plotly Dash

Successful launches by site



- KSC is the most successful site with CCAFS being the least

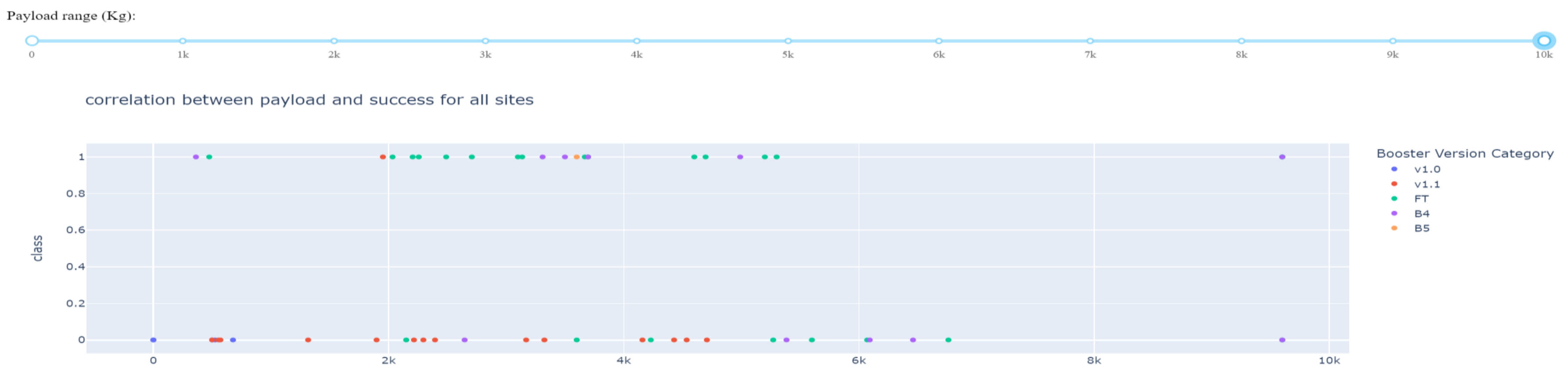
Total successful launches for most successful site



- KSC LC -39A was the most successful site
- As we drill down we see that 76.9% of launches from this site were successful

Correlation between payload and success

- Payloads below 6000 are the most succesful

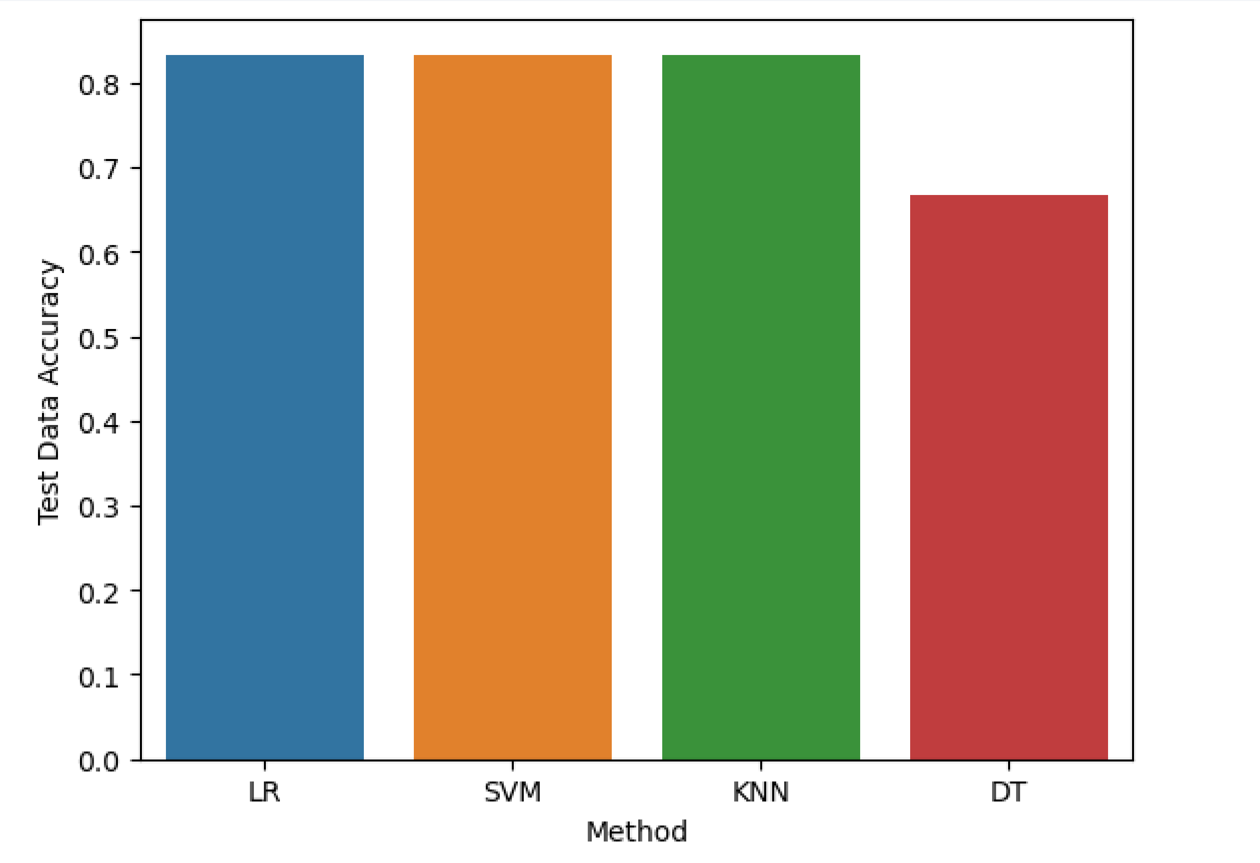


Section 5

Predictive Analysis (Classification)

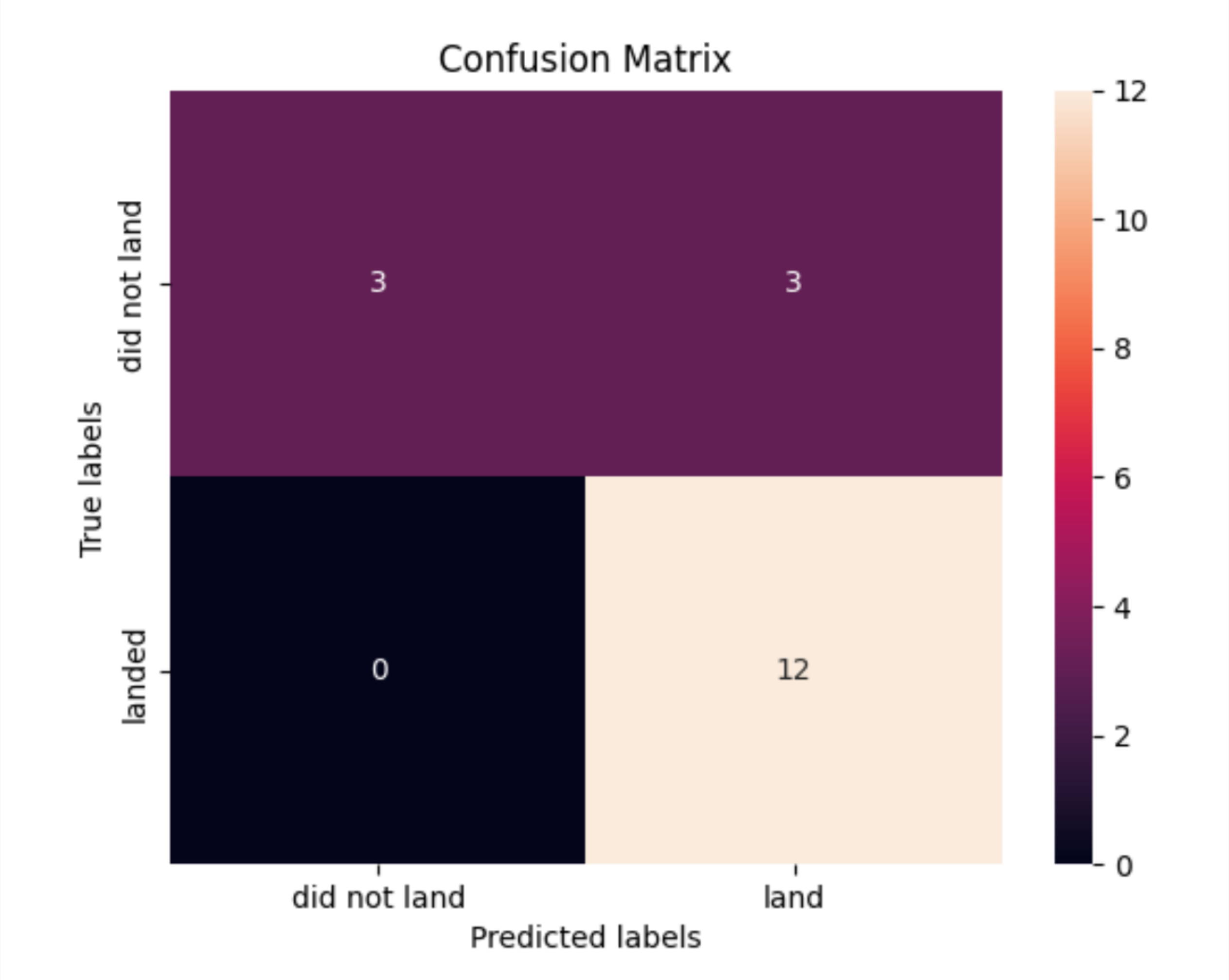
Classification Accuracy

- We used 4 classification methods
- The most accurate model was evenly spread across LR, SVM and KNN. Our decision tree performed the worst.



Confusion Matrix

- This confusion matrix for the SVM model shows that it predicted with 100% accuracy true positives on whether it landed, but 50% with false positives on did not land



Conclusions

- Decission tree was the worst model. All other models showed similar accuracy
- Launches with lower payloads were more succseful
- Success of launches increased over the years
- KSC LC -39A is the most successful launch site
- Orbits ES-L1, GEO, HEO and SSO appear to be the best choices

Appendix

- Repo

Thank you!