Winning Space Race with Data Science

Bourhan Barake
October 2022

Outline

ABSTRACT

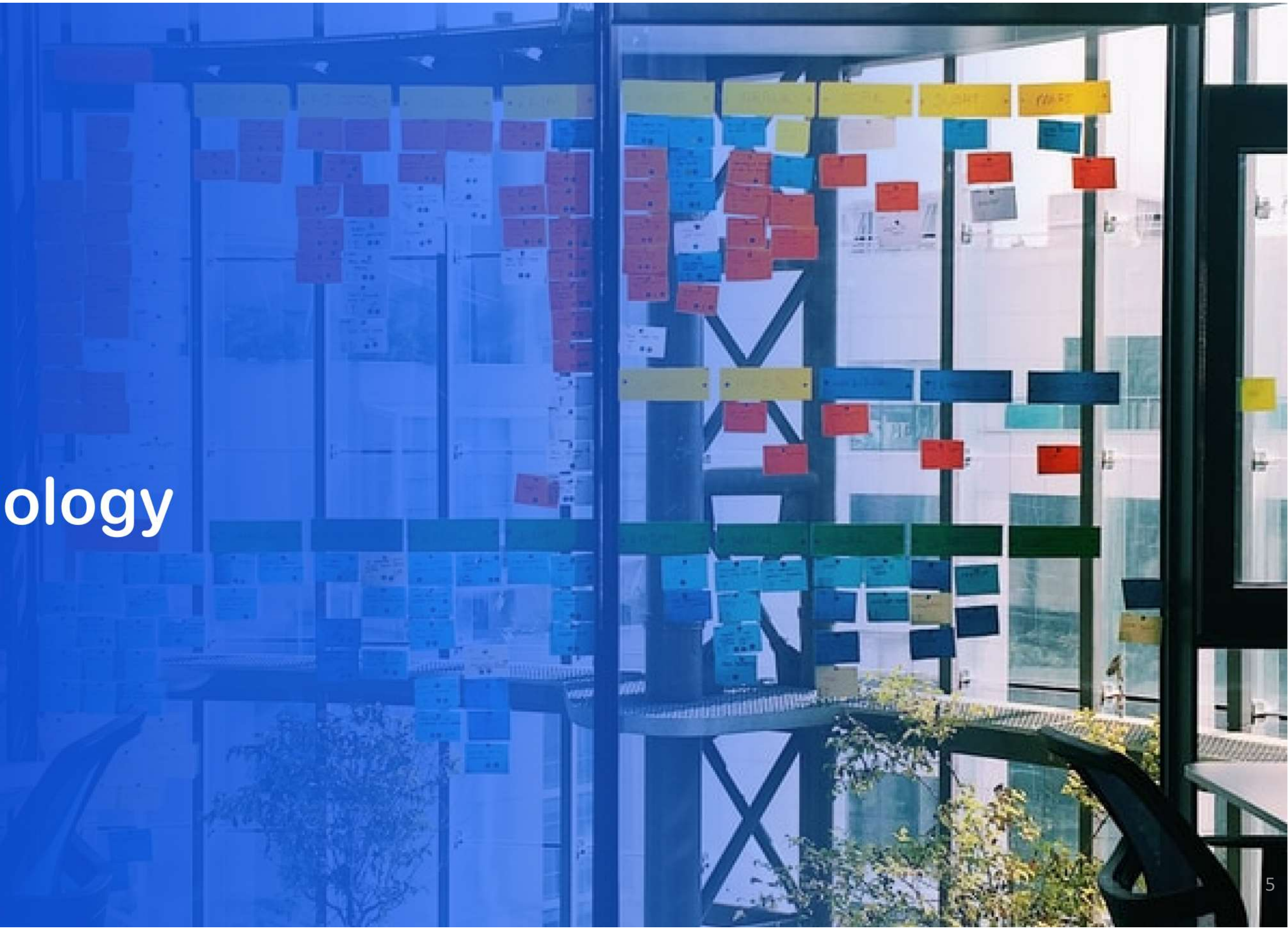
- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

Introduction

- Project background and context
  - Space X Falcon9 rocket cost 62 million dollars; other providers cost upward of 165 million dollars each
  - Much of the savings is because Space X can reuse the first stage.
  - Determining if the first stage will land, we can determine the cost of a launch.
  - The goal of the project is to create a machine learning pipeline to predict if the first stage will land successfully.
- Problems you want to find answers
  - What factors determine if the rocket will land successfully?
  - The interaction amongst various features that determine the success rate of a successful landing.
  - What operating conditions needs to be in place to ensure a successful landing program.

Section 1

Methodology



Methodology

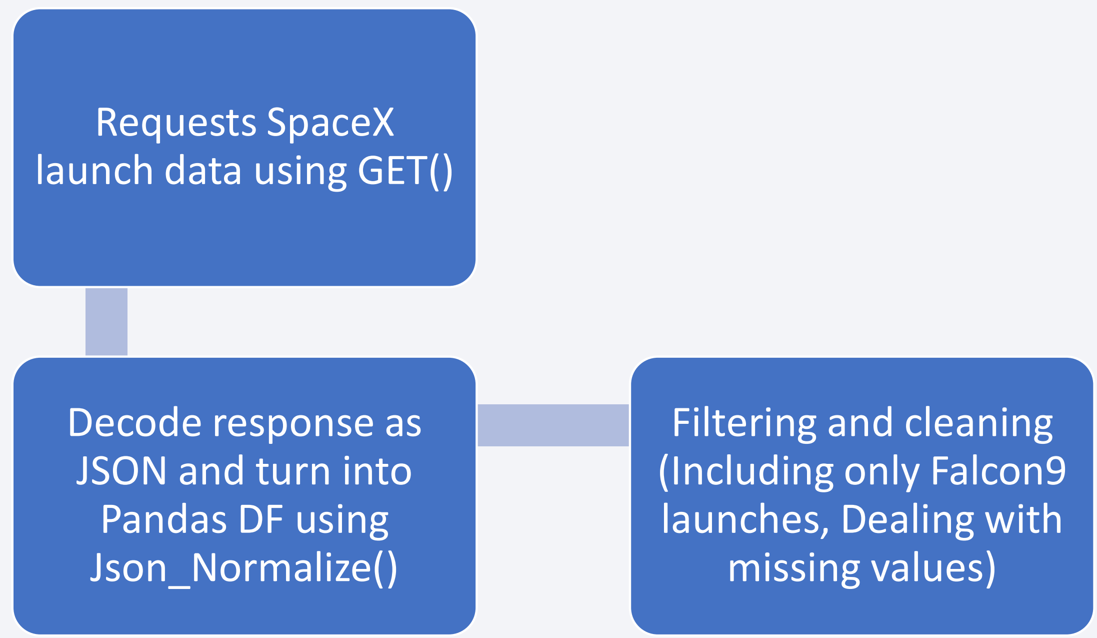
- Executive Summary
- Data collection methodology:
  - Data was collected using SpaceX API and web scrapping from Wikipedia
- Perform data wrangling
  - One-hot encoding was applied to categorical features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Data collection was done using get request to the SpaceX API.
- Decoded the response content as a Json using and turned it into a pandas data frame
- Cleaned the data, checked for missing values and fill in missing values where necessary using various techniques.
- Performed web scrapping from Wikipedia for Falcon 9 launch records using BeautifulSoup.
- The objective was to extract the launch records as HTML table, parse the table and convert it to a pandas data frame for future analysis.

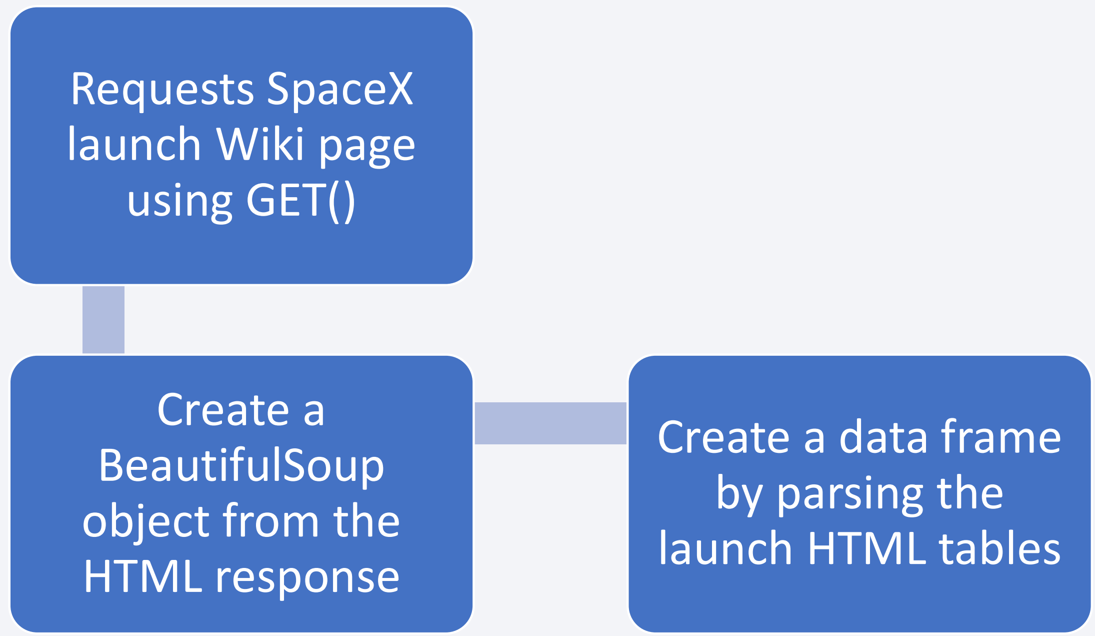
Data Collection – SpaceX API

- We used the get request to the SpaceX API to collect data, clean the requested data and did some basic data wrangling and formatting.
- The link to the notebook is: https://github.com/b-barake/Data-Science-Capstone-Project---IBM/blob/2add858071d586e963baa262fbef7d20fd2946c4/Data%20collection/jupyter-labs-spacex-data-collection-api.ipynb

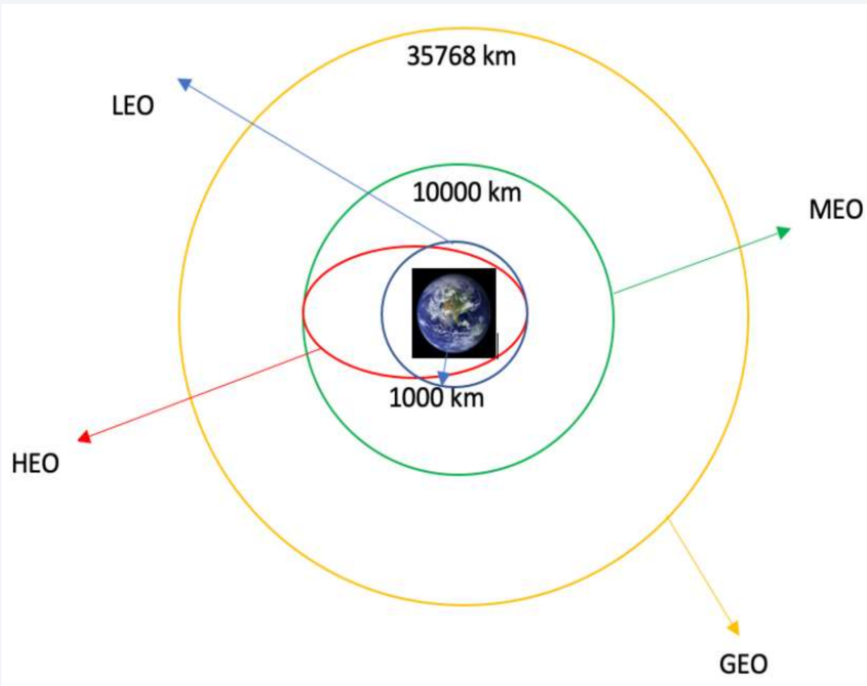


Data Collection - Scraping

- We applied web scrapping to import Falcon 9 launch records using BeautifulSoup
- We parsed the table and converted it into a pandas dataframe.
- The link to the notebook is: https://github.com/b-barake/Data-Science-Capstone-Project---IBM/blob/2add858071d586e963baa262fbef7d20fd2946c4/Data%20collection/jupyter-labs-webscraping.ipynb



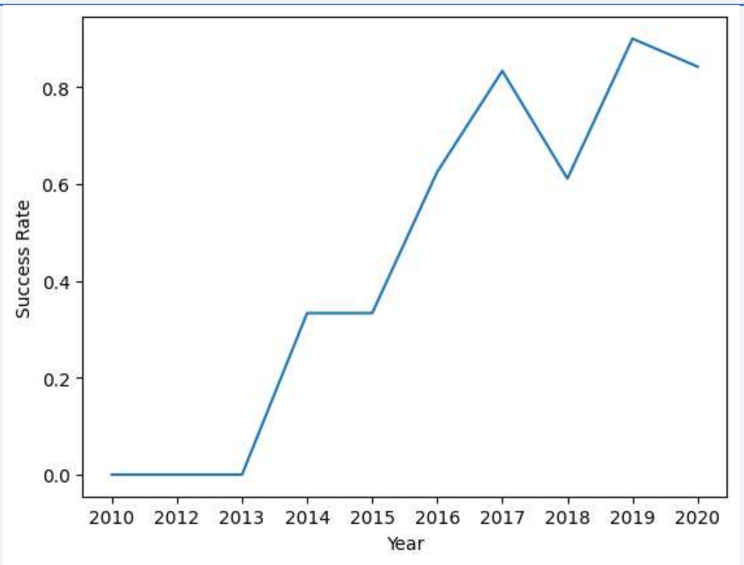
Data Wrangling

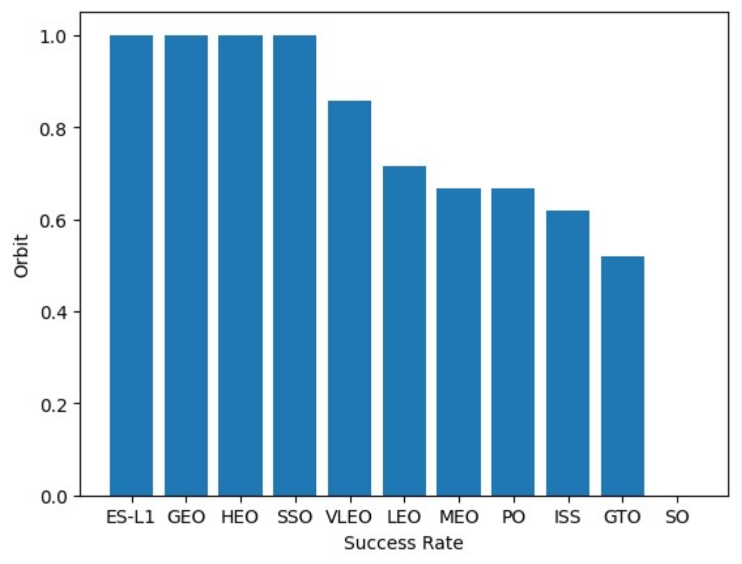


- We performed exploratory data analysis and determined the training labels.
- We calculated the number of launches at each site, and the number and occurrence of each orbits
- We created landing outcome label from outcome column and exported the results to csv.
- The link to the notebook is https://github.com/b-barake/Data-Science-Capstone-Project---IBM/blob/2add858071d586e963baa262fbef7d20fd2946c4/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Explored the data by visualizing the relationship 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.





- The link to the notebook https://github.com/b-barake/Data-Science-Capstone-Project---IBM/blob/2add858071d586e963baa262fbef7d20fd2946c4/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- We applied EDA with SQLite to get insight from the data. We wrote queries to find out for instance:
  - The names of unique launch sites in the space mission.
  - The total payload mass carried by boosters launched by NASA (CRS)
  - The average payload mass carried by booster version F9 v1.1
  - The total number of successful and failure mission outcomes
  - The failed landing outcomes in drone ship, their booster version and launch site names.
- The link to the notebook is: https://github.com/b-barake/Data-Science-Capstone-Project---IBM/blob/2add858071d586e963baa262fbef7d20fd2946c4/jupyter-labs-eda-sql-edx_sqllite.ipynb

Build an Interactive Map with Folium

- We marked all launch sites, and added map objects such as markers, circles, lines to mark the success or failure of launches for each site on the folium map.
- We assigned the feature launch outcomes (failure or success) to class 0 and 1 (failure, success).
- Using the color-labeled marker clusters, we identified which launch sites have relatively high success rate.
- We calculated the distances between a launch site to its proximities. We answered some question for instance:
  - Are launch sites near railways, highways and coastlines.
  - Do launch sites keep certain distance away from cities.

Build a Dashboard with Plotly Dash

- We built an interactive dashboard with Plotly dash, this dashboard contains:
  - A Pie Chart that shows the total launches for selected sites
  - A Scatter plot showing the relationship between Payload Mass and Outcome for different booster versions
- The link to the notebook is: https://github.com/b-barake/Data-Science-Capstone-Project---IBM/blob/2add858071d586e963baa262fbef7d20fd2946c4/spacex_dash_app.py

Predictive Analysis (Classification)

- Loaded the data using NumPy and Pandas, transformed the data, split our data into training and testing.
- Built different machine learning models and tune different hyperparameters using GridSearchCV.
- Accuracy was the metric for our model, improved the model using feature engineering and algorithm tuning.
- Compared different models to find the most accurate one.
- Link to the notebook: https://github.com/b-barake/Data-Science-Capstone-Project---IBM/blob/2add858071d586e963baa262fbef7d20fd2946c4/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

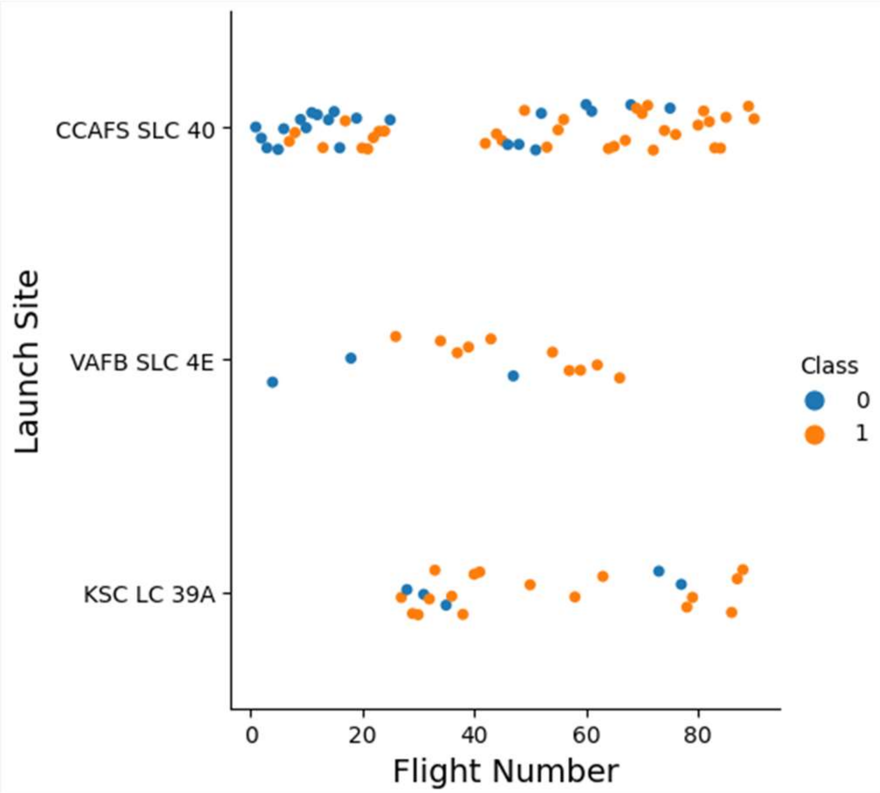
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

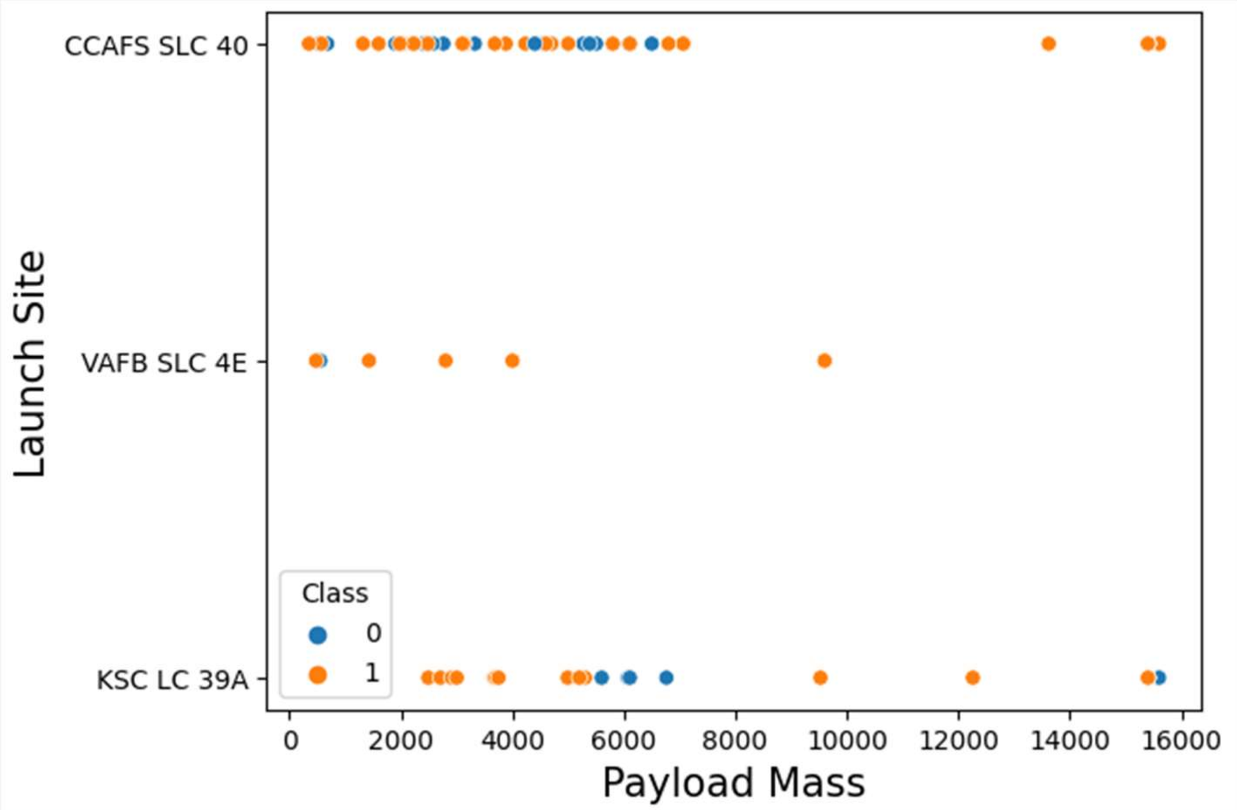
Flight Number vs. Launch Site

- Positive correlation between Flight Number and Success in all Launch Sites
- The larger the flight amount at a launch site, the greater the success rate at a launch site.
- Site KSC was not used for the first 20 flights
- Site VAFB is not used since Flight 70
- CCAFS is the most used site



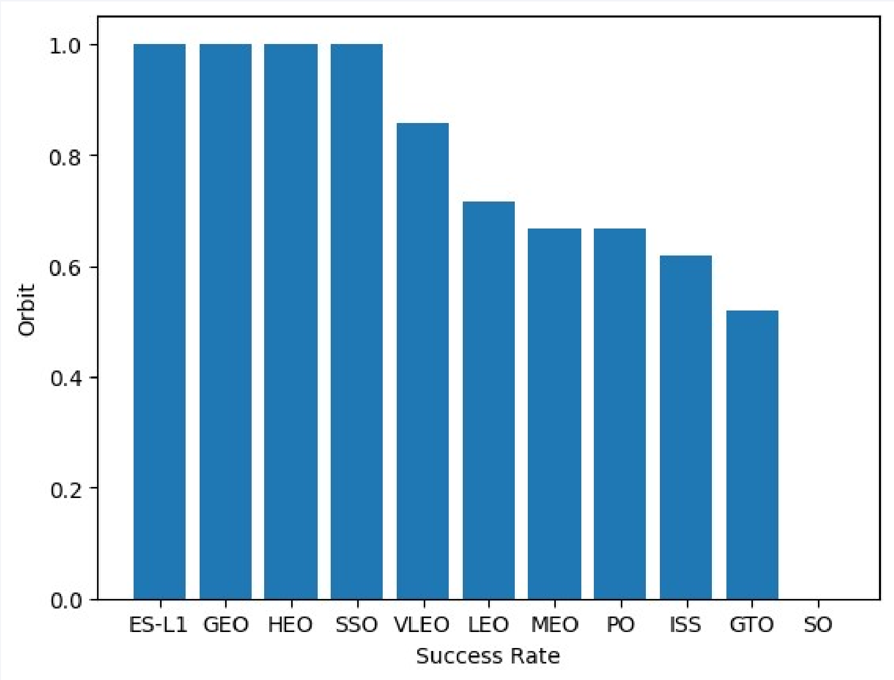
Payload vs. Launch Site

- The greater the Payload the more likely it will successfully land.
- Site VAFB does not have any launches for heavy payload mass (greater than 10,000 kg)



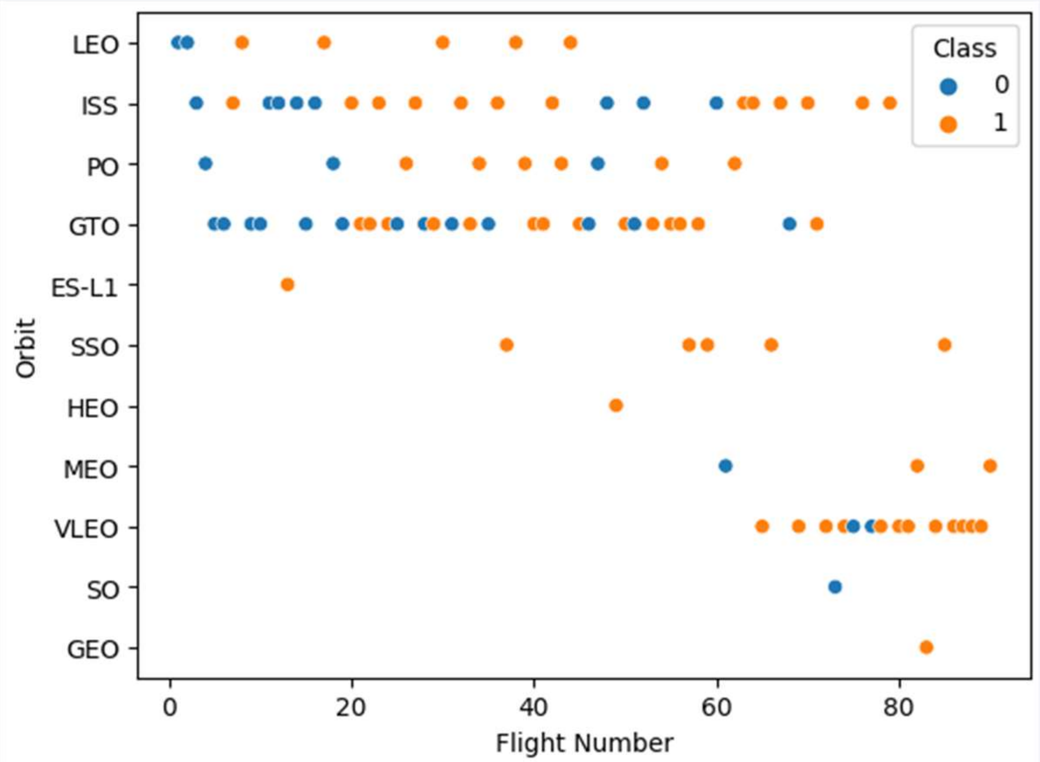
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, SSO and VLEO had the most success rate.
- SO did not have a single successful launch.
- ES-L1, HEO, GEO and SO stats are deceiving with only one launch



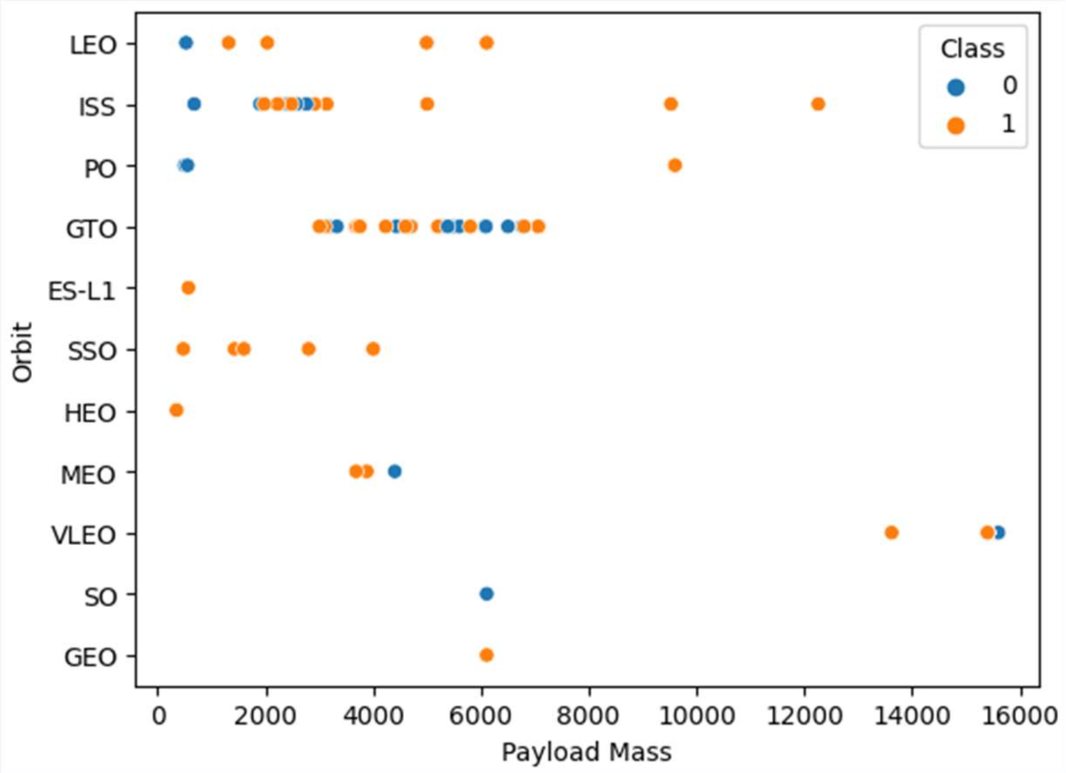
Flight Number vs. Orbit Type

- While viewing this chart along with the one in the previous slide, we can gather important insight
- ES-L1, HEO, GEO and SO stats are deceiving with only one launch
- ISS, GTO and VLEO are the most used sites



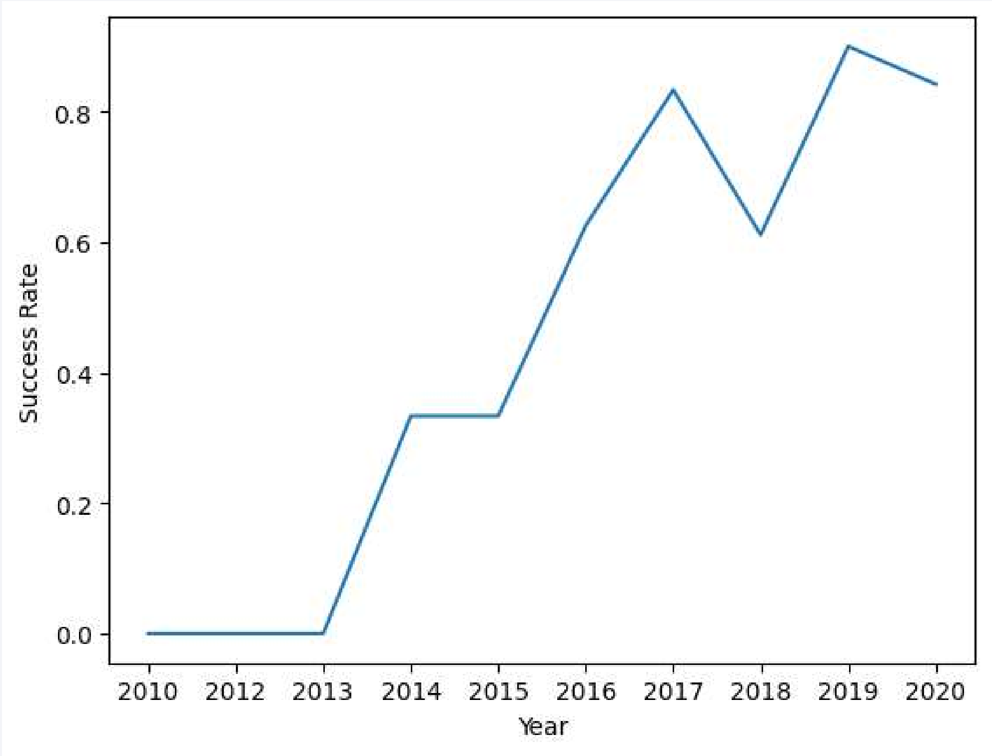
Payload vs. Orbit Type

- ISS and VLEO are the only orbits used for heavy Payloads (more than 10,000 kg)
- Most Launches are in the 2000 – 8000 kg range for Payload mass



Launch Success Yearly Trend

- Until 2013 there were no successful landings for the first stage.
- Positive trend since in the 2013-2020 period for success rate.



All Launch Site Names

- To find the names of the unique launch sites we use DISTINCT.

```
%sql SELECT distinct("Launch_Site") from SPACEXTBL
```

[7] ✓ 0.3s

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

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

Launch Site Names Begin with 'KSC'

- To limit the search for launch sites to begin with KSC we used the query ‘like’
- Limit 5 is used to show only the first 5 records

```
%sql SELECT * from SPACEXTBL where ("Launch_Site") like '%KSC%' limit 5
```

✓ 0.4s

* sqlite:///my_data1.db
Done.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
19-02-2017	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
16-03-2017	6:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
30-03-2017	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
01-05-2017	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
15-05-2017	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

Total Payload Mass

- Similar to the previous query ‘like’ is used to find boosters launched by NASA
- The function SUM is used to find the total Payload

```
%sql SELECT SUM("PAYLOAD_MASS__KG_") from SPACEXTBL where "Customer" like '%NASA (CRS)%'
✓ 0.4s

 * sqlite:///my_data1.db
Done.

SUM(PAYLOAD_MASS__KG_)
                 48213
```

Average Payload Mass by F9 v1.1

- Exact same logic as the previous query, but instead the function SUM we use AVG to find the average Payload

```
%sql SELECT AVG("PAYLOAD_MASS__KG_") from SPACEXTBL where "Booster_Version" like '%F9 v1.1%'
✓ 0.2s

 * sqlite:///my_data1.db
Done.

AVG(PAYLOAD_MASS__KG_)
    2534.6666666666665
```

First Successful Ground Landing Date

- Since the data type of the column Date is date, we can use the function min to extract the first date that had a successful landing

```
%sql SELECT MIN("Date") from SPACEXTBL WHERE [Landing_Outcome] like '%Success (drone ship)%'
✓ 0.4s

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

MIN(Date)
06-05-2016

Successful Drone Ship Landing with Payload between 4000 and 6000

- To query results between two values, we use the query between x and y

```
%sql SELECT * from SPACEXTBL WHERE [Landing_Outcome] like '%Success (ground pad)%' and "PAYLOAD_MASS__KG_" between 4000 and 6000;
✓ 0.4s
 * sqlite:///my_data1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
01-05-2017	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
07-09-2017	14:00:00	F9 B4 B1040.1	KSC LC-39A	Boeing X-37B OTV-5	4990	LEO	U.S. Air Force	Success	Success (ground pad)
08-01-2018	1:00:00	F9 B4 B1043.1	CCAFS SLC-40	Zuma	5000	LEO	Northrop Grumman	Success (payload status unclear)	Success (ground pad)

Total Number of Successful and Failure Mission Outcomes

- The column Mission Outcome is used to check whether the mission failed or not.
- We used the ‘where’ to specify which column we are searching and ‘like’ to specify which outcome we are searching for.

```
%sql SELECT count(*) from SPACEXTBL where "Mission_Outcome" like '%Success%'
✓ 0.2s
* sqlite:///my_data1.db
Done.
count(*)
100
```

```
%sql SELECT count(*) from SPACEXTBL where "Mission_Outcome" like '%Failure%'
✓ 0.3s
* sqlite:///my_data1.db
Done.
count(*)
1
```

Boosters Carried Maximum Payload

- We determined the booster that have carried the maximum payload using a subquery in the WHERE clause and the MAX() function.

```
%sql SELECT("Booster_Version") from SPACEXTBL where "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") from SPACEXTBL)
```

✓ 0.5s

```
 * 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

2015 Launch Records

- In this query we used the function substr() or sub string to specify which part of the date (or string) we want to search for.
- Used the ‘where’ query to find the the successful ground landings and landings in the year 2017

```
%sql SELECT substr(Date,4,2),[Landing_Outcome],[Booster_Version],[Launch_Site] from SPACEXTBL where [Landing_Outcome] like '%Success (ground pad)%' and substr("Date",7,4) = '2017'
✓ 0.4s

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

substr(Date,4,2)	Landing_Outcome	Booster_Version	Launch_Site
02	Success (ground pad)	F9 FT B1031.1	KSC LC-39A
05	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
06	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
08	Success (ground pad)	F9 B4 B1039.1	KSC LC-39A
09	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A
12	Success (ground pad)	F9 FT B1035.2	CCAFS SLC-40

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

- We selected Landing outcomes and the ‘**count**’ of landing outcomes from the data and used the ‘**where**’ clause to filter for landing outcomes ‘**between**’ 2010-06-04 to 2010-03-20.
- We applied the ‘**group by**’ clause to group the landing outcomes and the ‘**order by**’ clause to order the grouped landing outcome in descending order.

```
%sql SELECT [Landing_Outcome], COUNT(*) AS [Count_Launches] FROM SPACEXTBL WHERE "Date" BETWEEN '04-06-2010' AND '20-03-2017' GROUP BY [Landing_Outcome] ORDER BY [Count_Launches] DESC;
```

✓ 0.4s

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

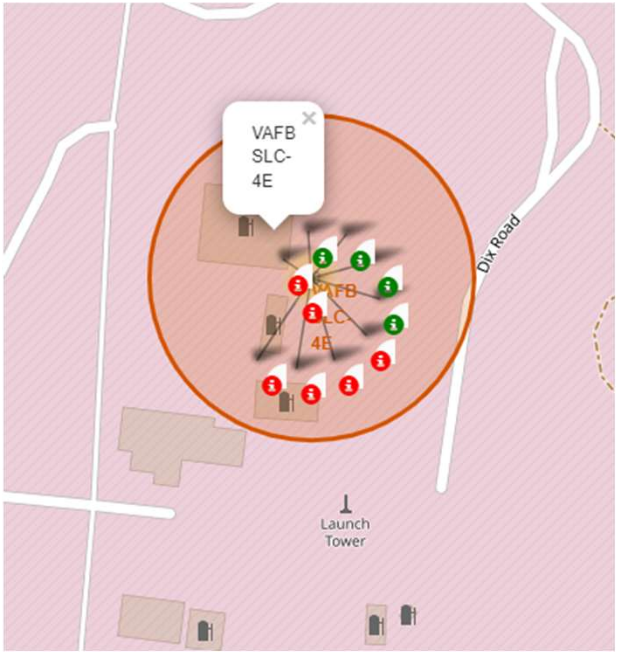
Landing_Outcome	Count_Launches
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

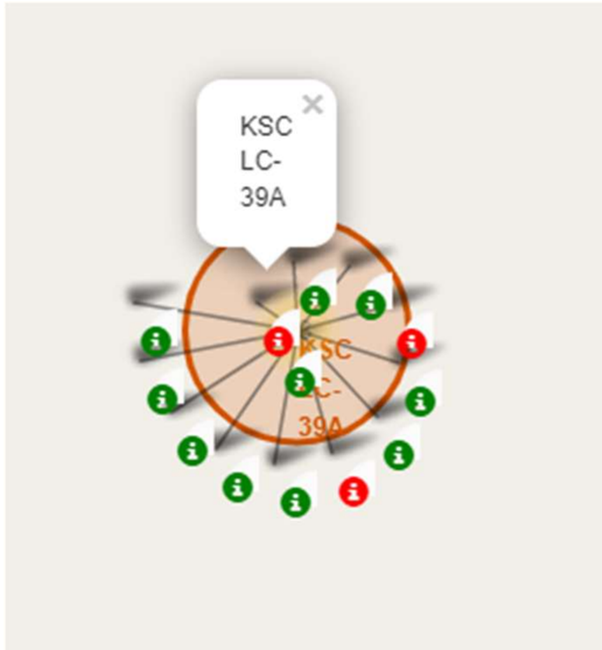
Section 3

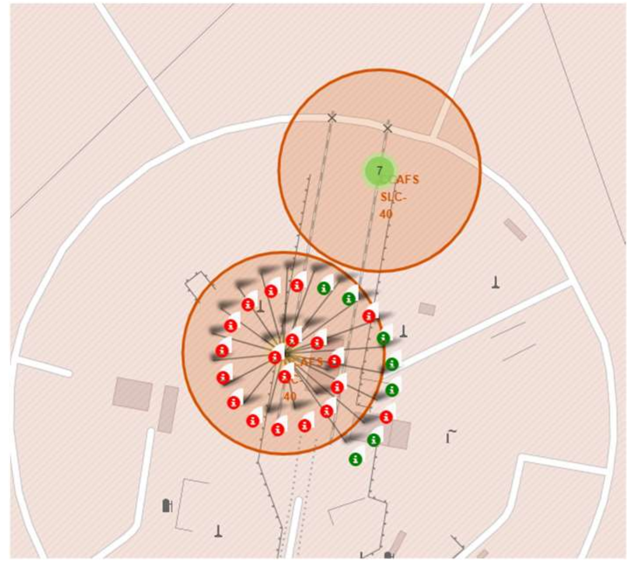
Launch Sites Proximities Analysis

All launch sites global map markers

From the following map, all launch sites from SpaceX are done in US soil, one is in the east coast (Florida) and the other in the west coast (California)

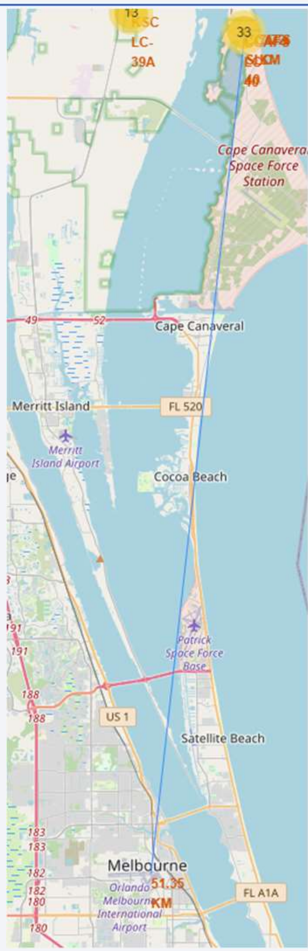


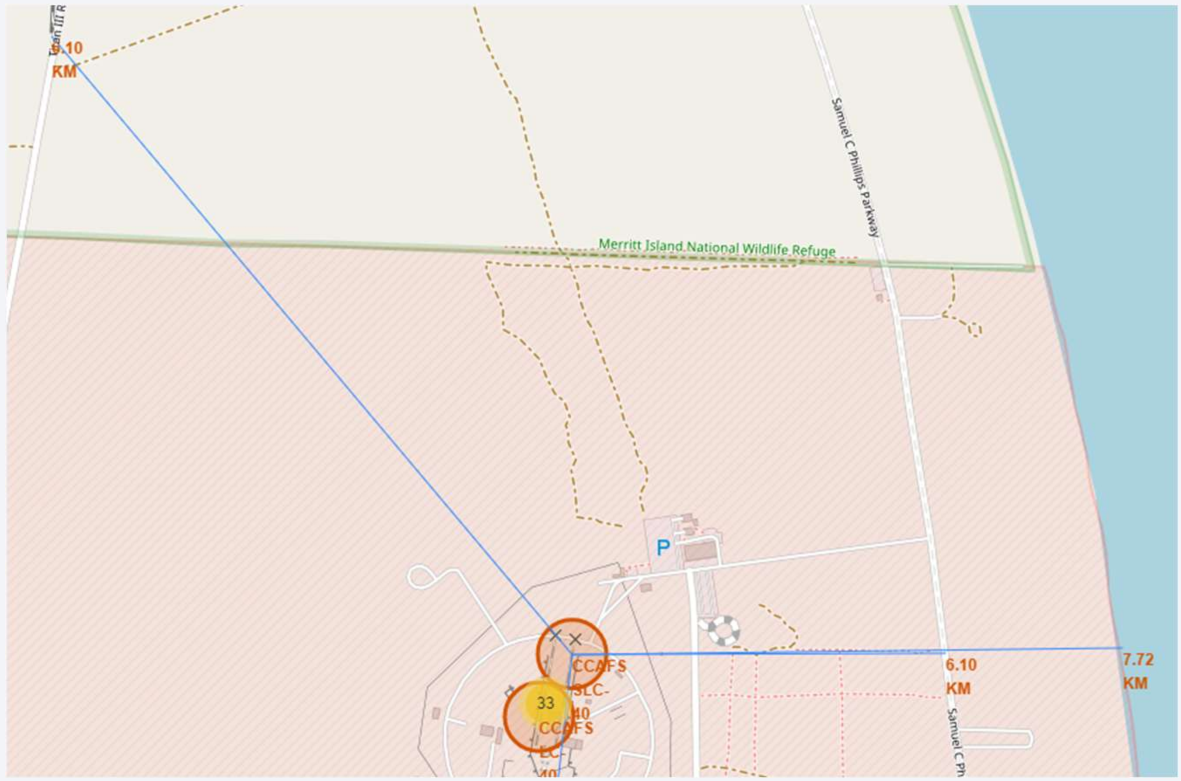


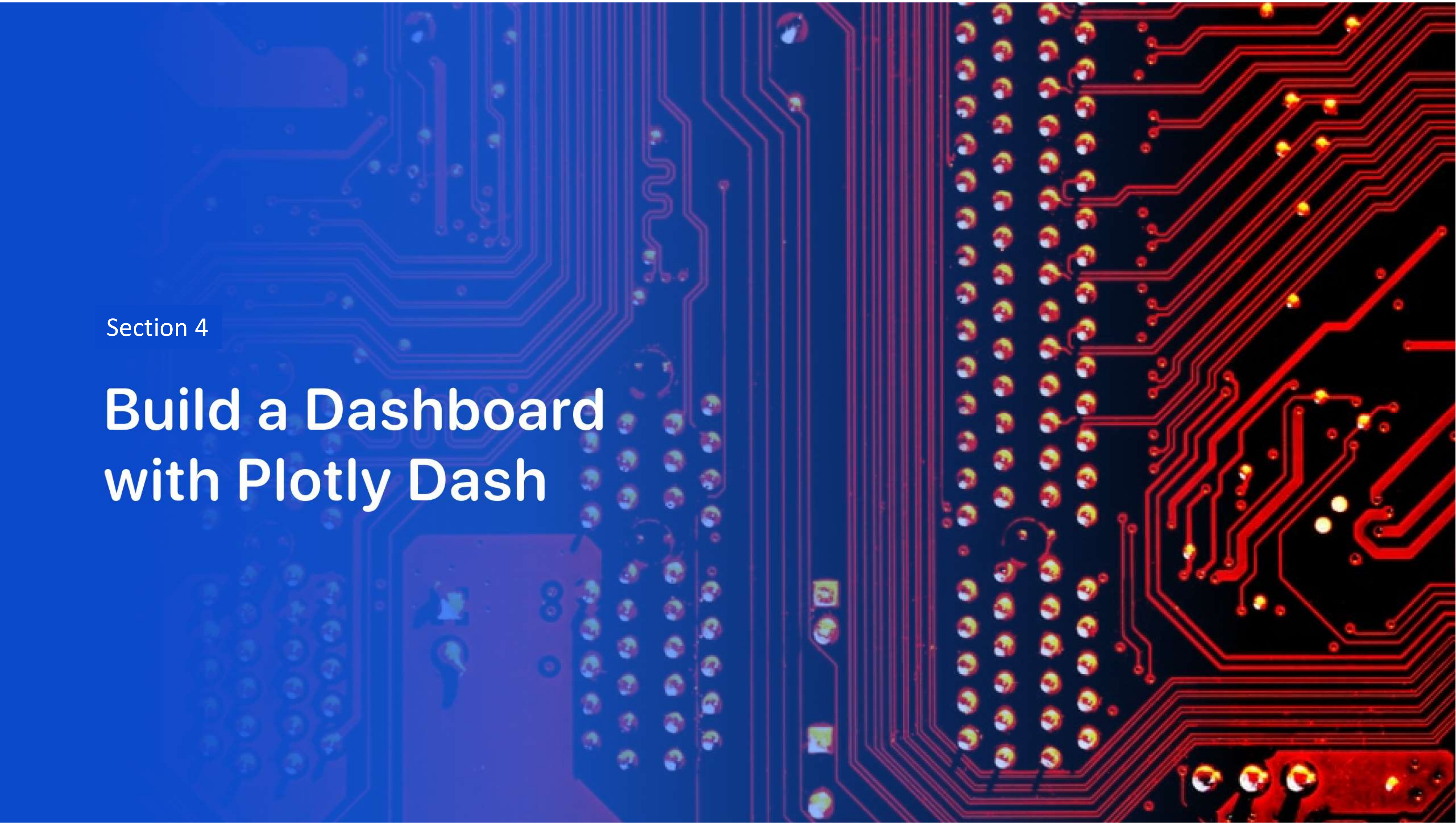


LAUNCH SITES WITH COLOR LABELS

Launch Site distance to landmarks

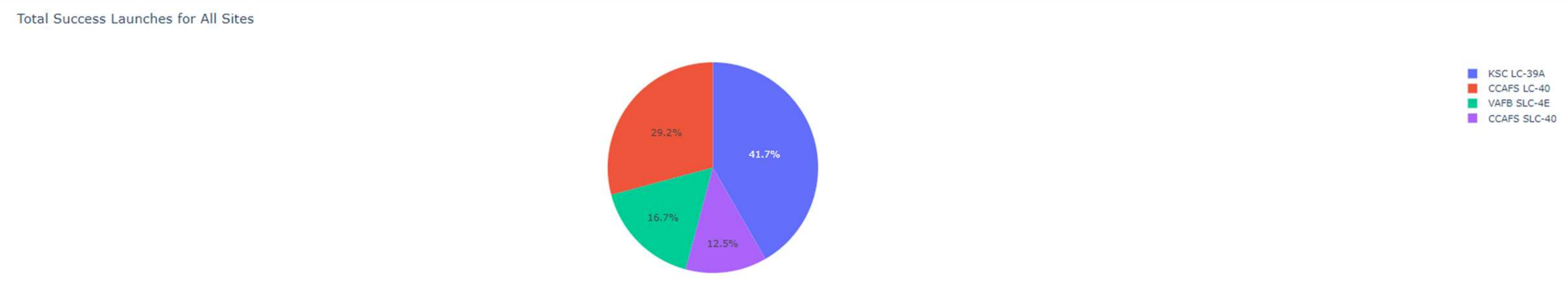





Section 4
Build a Dashboard with Plotly Dash

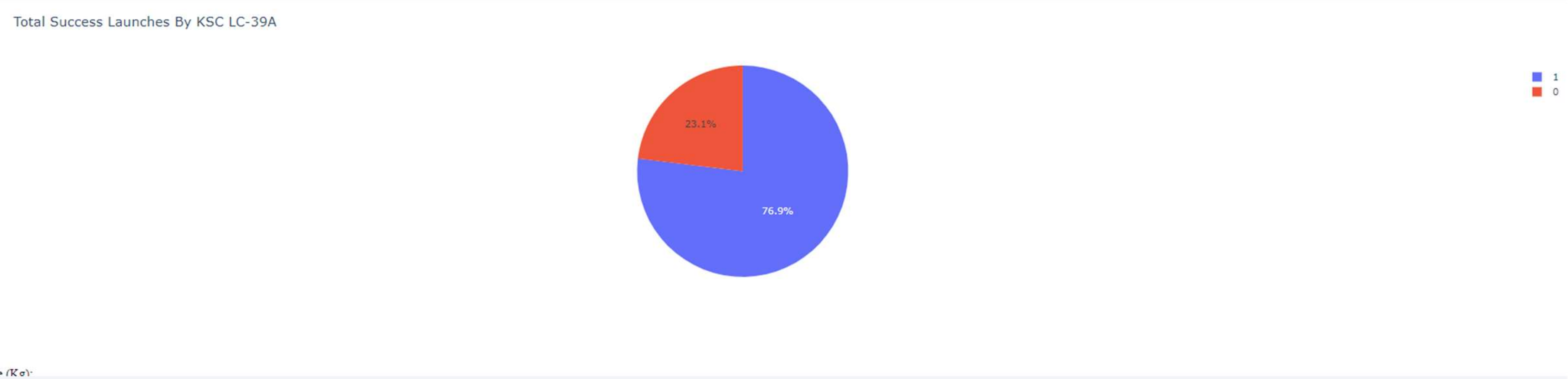
Success rare for all launch sites

- From the pie chart below, site KSC had the most successful launches with a value of 41.7%



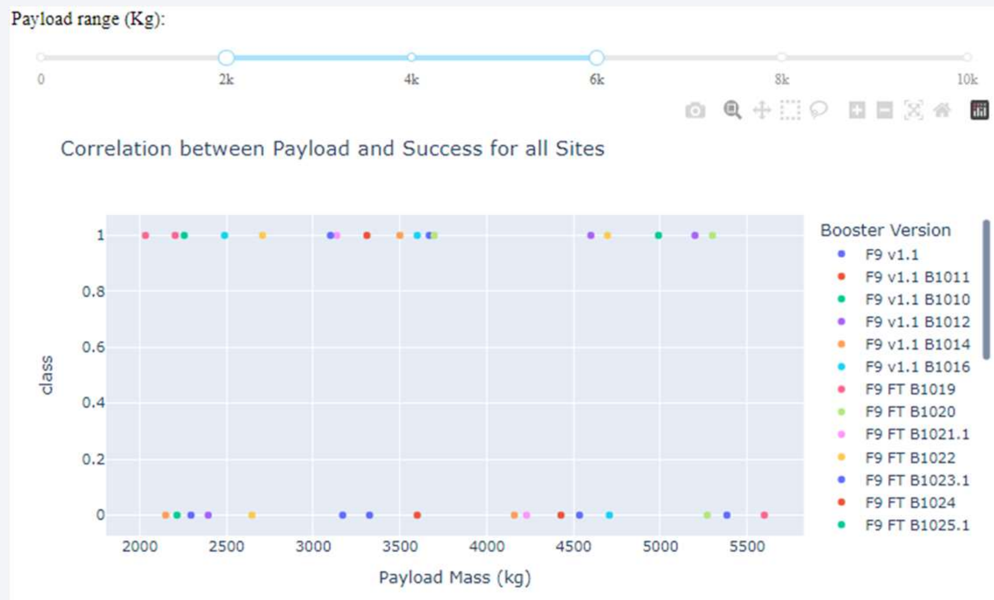
Total Success Launches by KSC LC-39A

- KSC had a success rate of 76.9%



Scatter Plot - Payload vs Success for different Booster Versions

- Most Launches happened in the 2000-6000kg range
- The

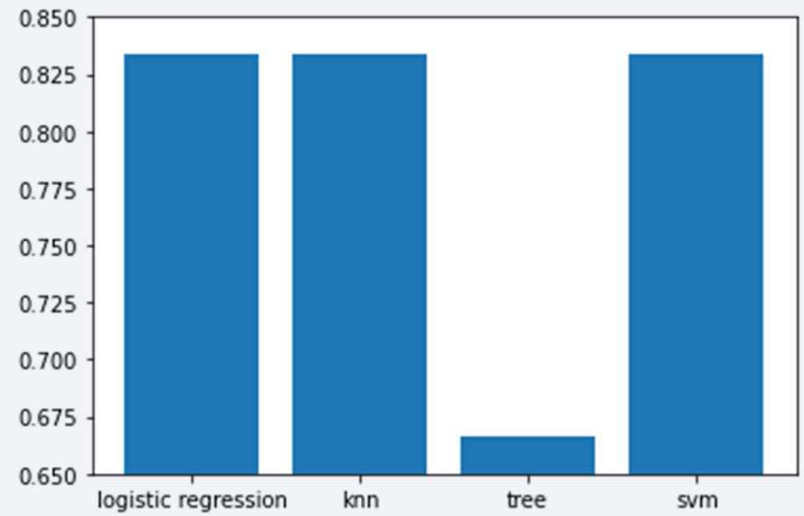


Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The following bar chart shows the accuracy of 4 different classification models, the decision tree model had the lowest accuracy with about 65%
- For log reg, KNN and SVM it is hard to see the difference from the graph, by creating a simple line of code it was found that logistic regression was the most accurate model



Confusion Matrix

- Using a logistic regression model, it was able to correctly predict 100% of the positive landings.
- The problem was with the false negatives with an accuracy of 50%

Conclusions

We can conclude that:

- The larger the flight amount at a launch site, the greater the success rate at a launch site.
- Launch success rate started to increase in 2013 till 2020.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- KSC LC-39A had the most successful launches of any sites.
- Logistic Regression is the best machine learning algorithm for this task.


Thank you!