IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Yastene Ait Kheddache
19th July, 2022

Outline

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

Executive Summary

- Summary of Our aim was to find the factors that affected the success of SpaceX missions, such as the rocket design, payload, etc
- For Data collection, we use SpaceX web API to retrieve raw Data as well as SpaceX wikipedia page.
- We fit data into classification models: Logistic Regression, Decision Tree, Support Vector Machine, k-Nearest Neighbors and find which model is best for predicting landing outcome.

Introduction

- SpaceX has the aim to make space flights more cheaper by first stage landings.
- We will predict landing outcomes on data of previous launces

Section 1

Methodology

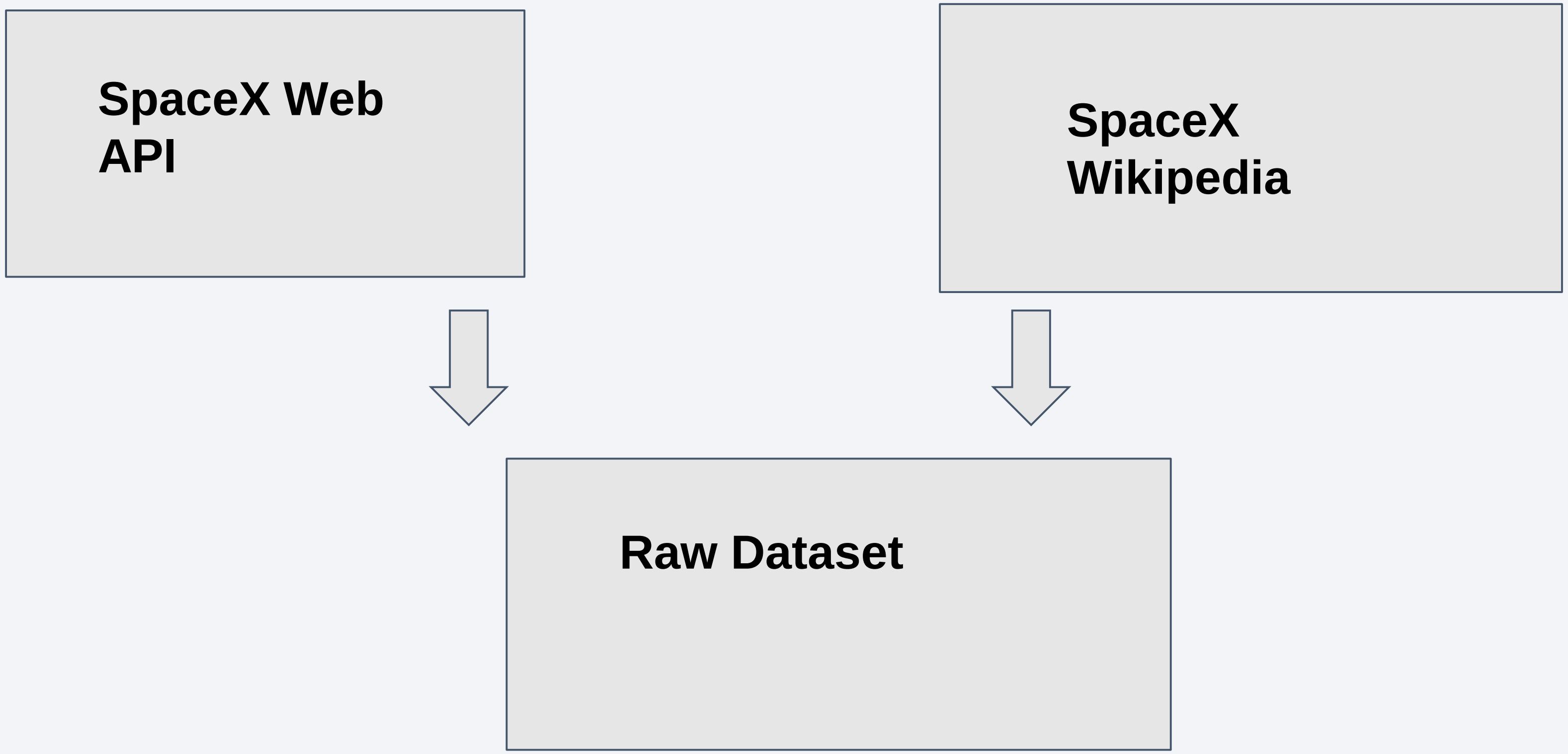
Methodology

Executive Summary

- Data collection methodology:
  - • SpaceX web API or SpaceX Wikipedia web scrapping
- Perform data wrangling
  - Dealt with missing values and created landing category variable
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Grid search to train, test and tune classification models

Data Collection

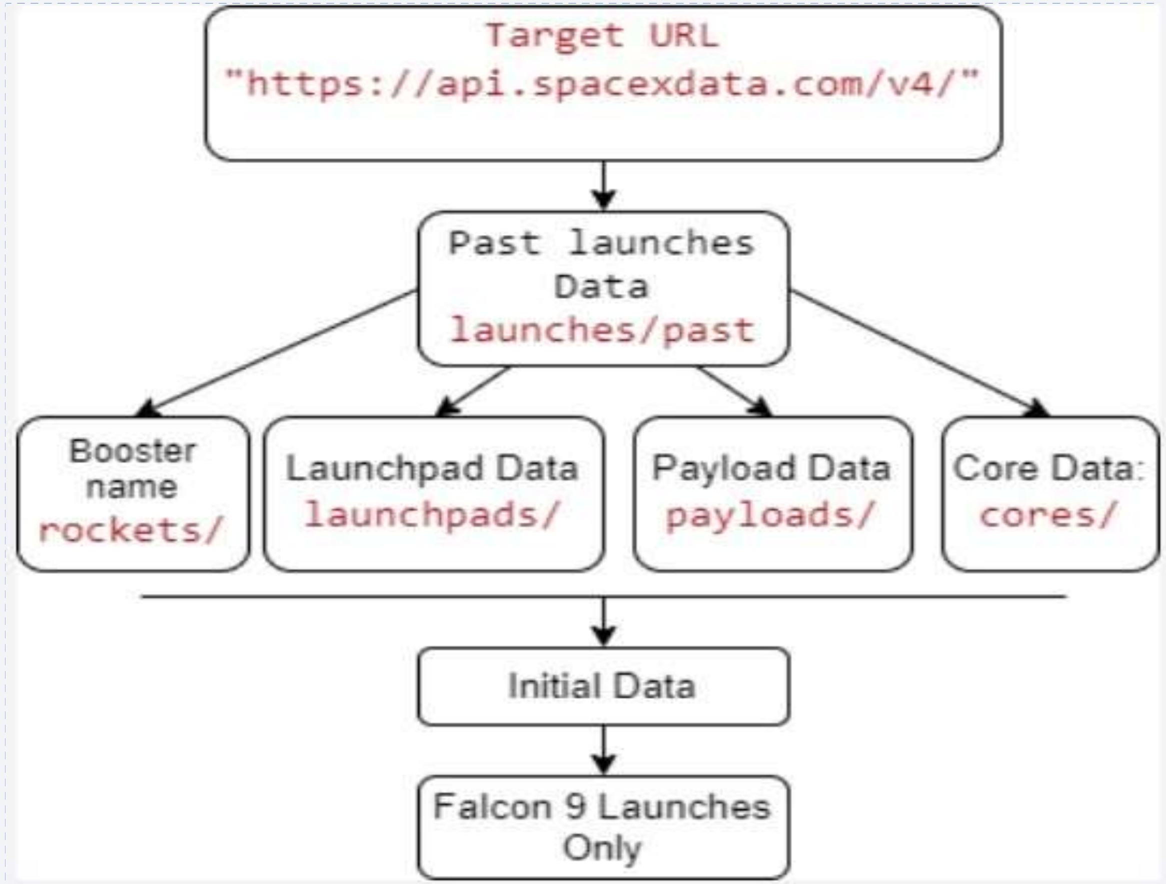
- We use SpaceX web API or SpaceX Wikipedia web scrapping for data collection.



Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

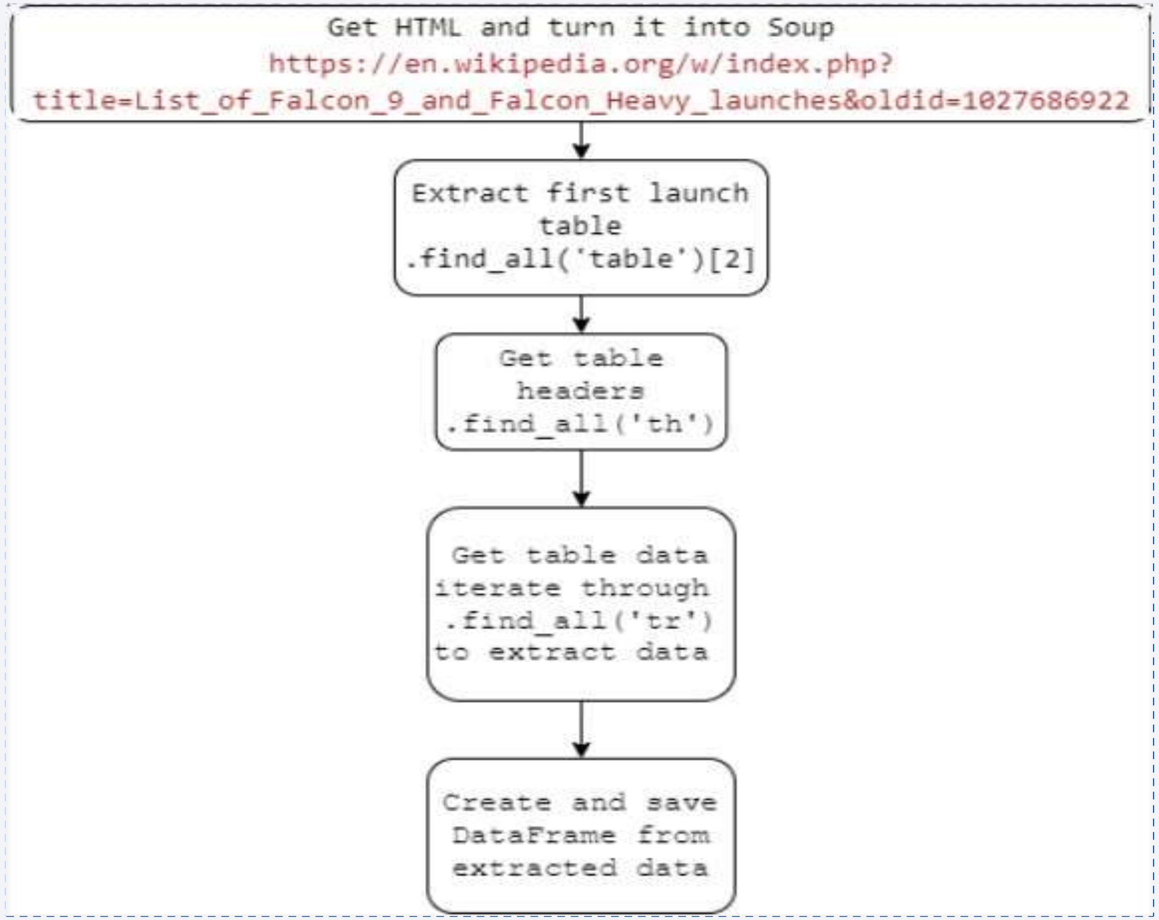
- https://github.com/YasteneAitKheddache/capstone/blob/950fdd3d061baf35dd8e1382a90a32f4791a2caa/Data%20Collection.ipynb



Data Collection - Scraping

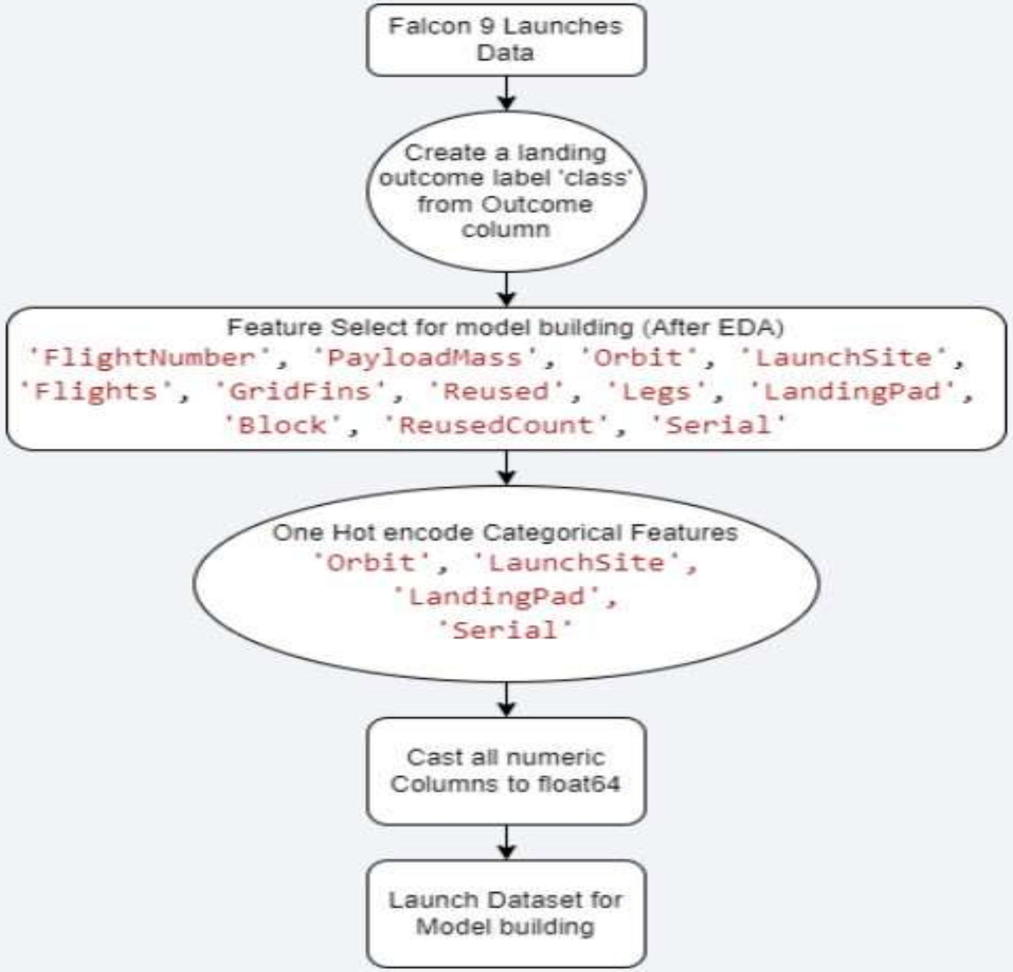
- Present your web scraping process using key phrases and flowcharts

- https://github.com/YasteneAitKheddache/capstone/blob/950fdd3d061baf35dd8e1382a90a32f4791a2caa/jupyter-labs-webscraping.ipynb



Data Wrangling

- Describe how data were processed
- You need to present your data wrangling process using key phrases and flowcharts
- https://github.com/YasteneAitKheddache/capstone/blob/950fdd3d061baf35dd8e1382a90a32f4791a2caa/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts

Payload Mass vs. Flight Number vs Launch Outcome Scatter Plot:

- First stage is more likely to land in later launches
- Seems the more massive the payload, the less likely First stage returns

Launch Site vs. Flight Number vs. Launch Outcome:

- Successful First Stage return improves over launches for all 3 sites
- CCAFS SLC 40 site is used for launches more often than the other 2 sites

https://github.com/YasteneAitKheddache/capstone/blob/950fdd3d061baf35dd8e1382a90a32f4791a2caa/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

https://github.com/YasteneAitKheddache/capstone/blob/950fdd3d061baf35dd8e1382a90a32f4791a2caa/jupyter-labs-eda-sql-coursera_sqllite.ipynb

- Task 1: Get Launch Site names
- Task 2: Get data with launch sites names starting with CCA
- Task 3: Get total Payload Mass carried by booster launched by NASA (CRS)
- Task 4: Get average Payload Mass carried by booster version F9 v1.1
- Task 5: Get the day of the first successful landing outcome in the ground pad
- Task 6: Get Booster Versions with a payload range of 4000kg to 6000kg and successfully land on a drone ship.
- Task 7: Get the mission outcome counts
- Task 8: Get booster versions that carried the maximum payload
- Task 9: Get data on failed landing onto a drone ship in the year 2015
- Task 10: Rank the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order

Build an Interactive Map with Folium

- Added objects to Map:
- Launch Sites, and their names
- Marker Cluster for each launch site to illustrate the successful and failed landings, for easily visualizing the number of launches and their landing success rate
- Distance from site CCAFS LC – 40 to the nearest City, Highway, Railroad, and Coast, to understand the importance of launch site location. Launch sites are generally far from cities and near the coast for safety reasons. They are near highways and railroads for efficient building materials transport

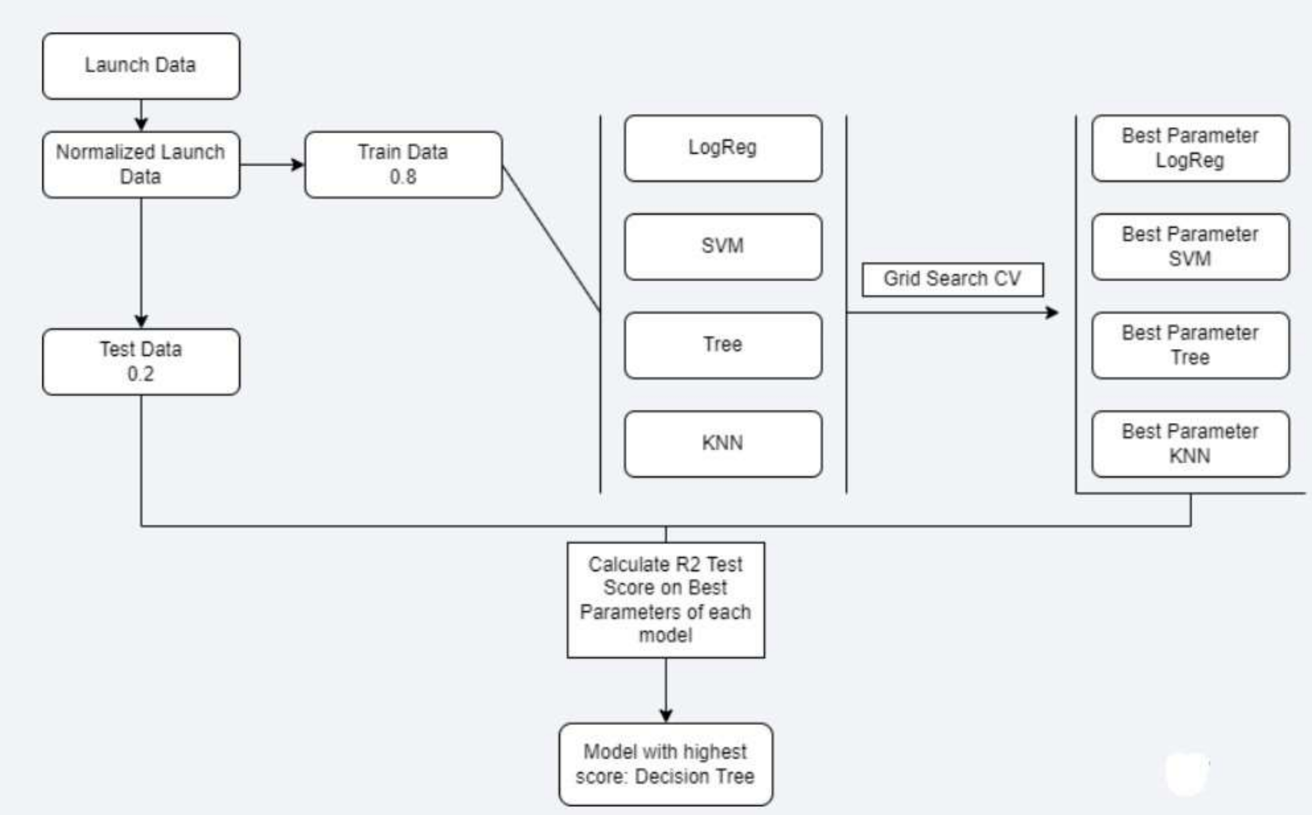
- https://github.com/YasteneAitKheddache/capstone/blob/950fdd3d061baf35dd8e1382a90a32f4791a2caa/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- A drop-down menu for choosing a site’s successful landing rate on a pie chart. If the all-sites option is selected, the percentage of successful landings for each site is displayed on a pie chart
- A payload range slider as input
- A graph that displays the correlation between Payload and Success for selected sites/site with a selected payload range. The interactive graph lets the user see how payload mass and site location affects the landing success rate
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose
- https://github.com/YasteneAitKheddache/capstone/blob/950fdd3d061baf35dd8e1382a90a32f4791a2caa/dash_interactive.py

Predictive Analysis (Classification)

- Using Grid Search CV for Logistic Regression, Decision Tree Classification, Support Vector Machine, and K-nearest Neighbors to find which model has the best Test Score
- https://github.com/YasteneAitKheddache/capstone/blob/950fdd3d061baf35dd8e1382a90a32f4791a2caa/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



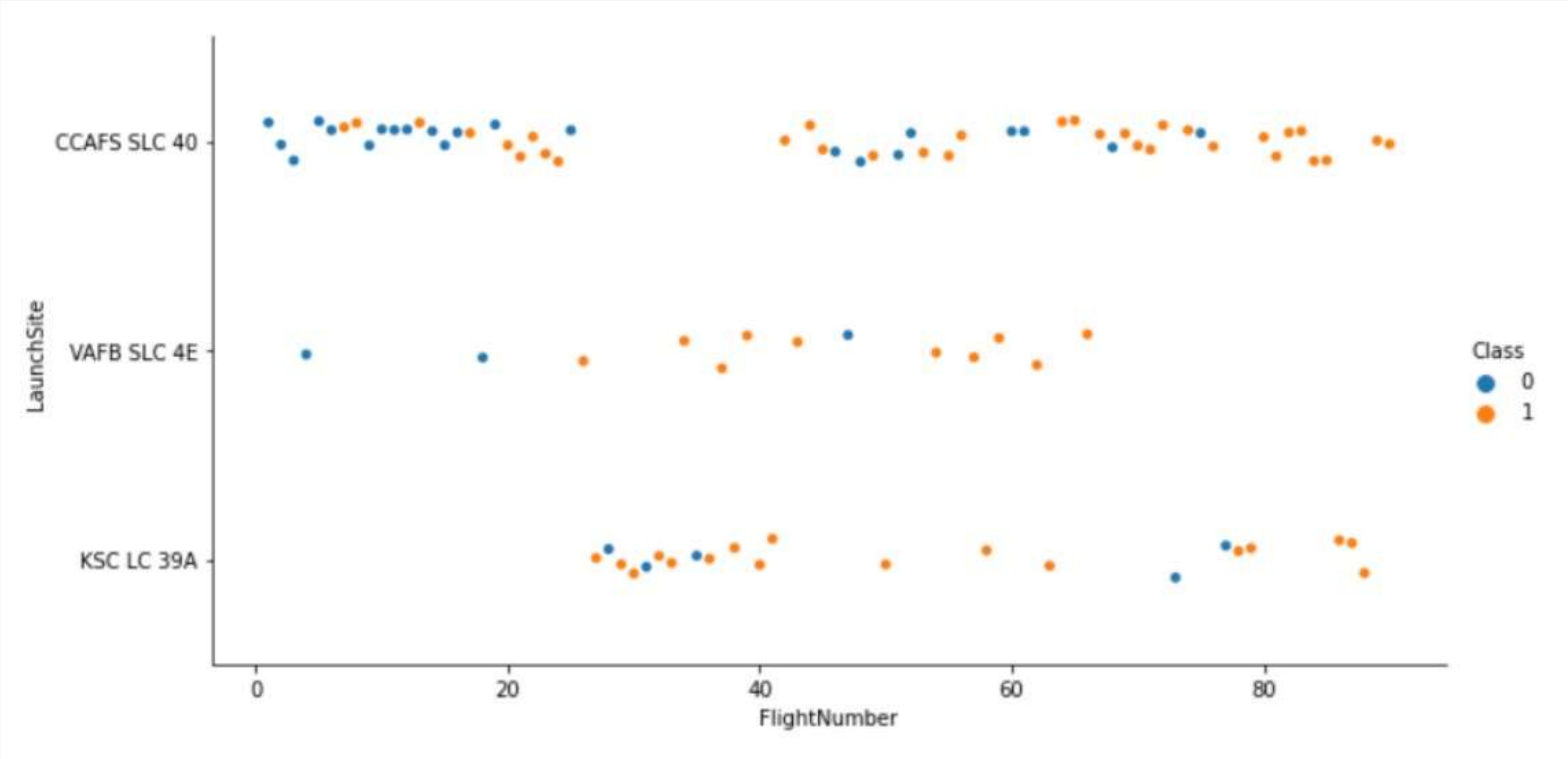
Results

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

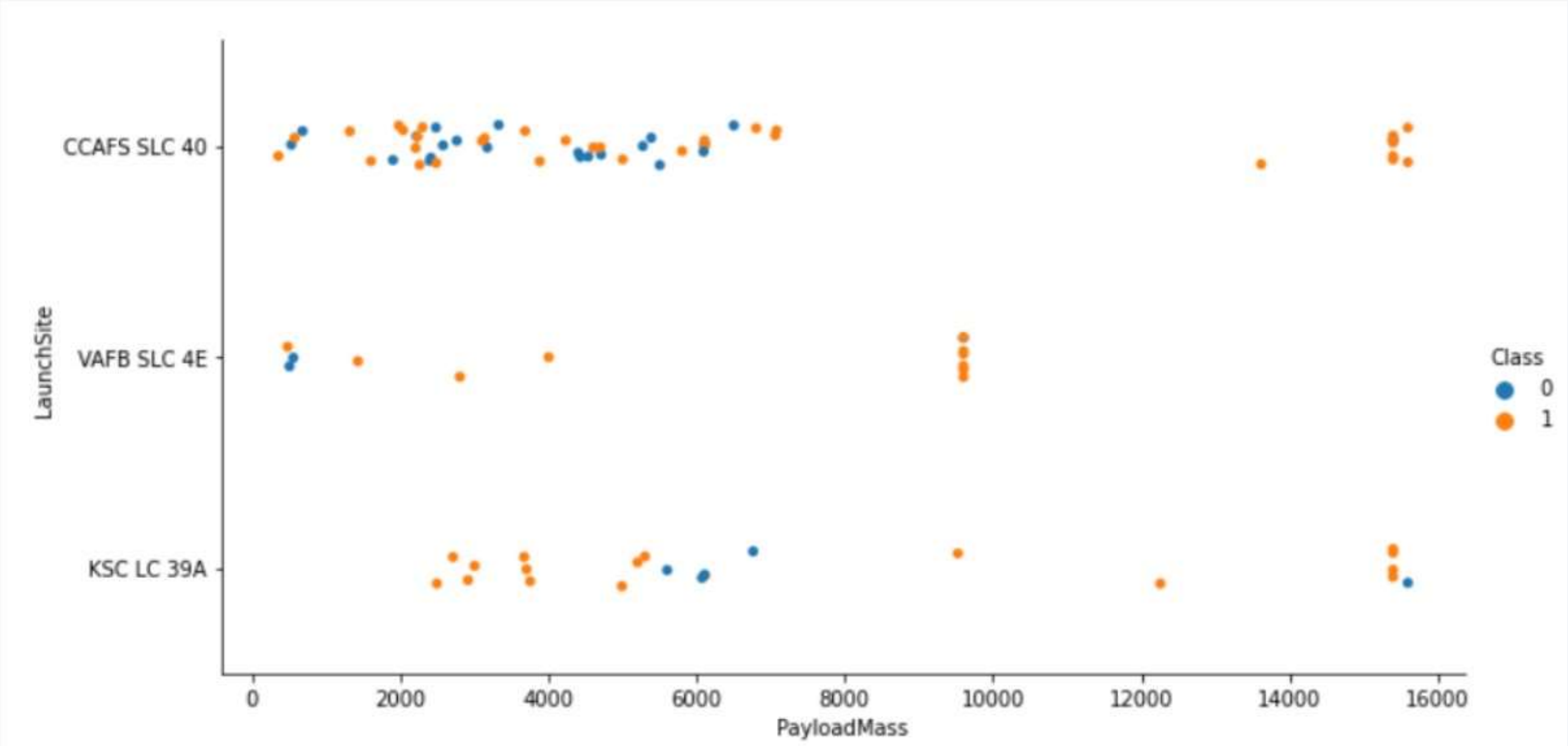
Section 2

Insights drawn from EDA

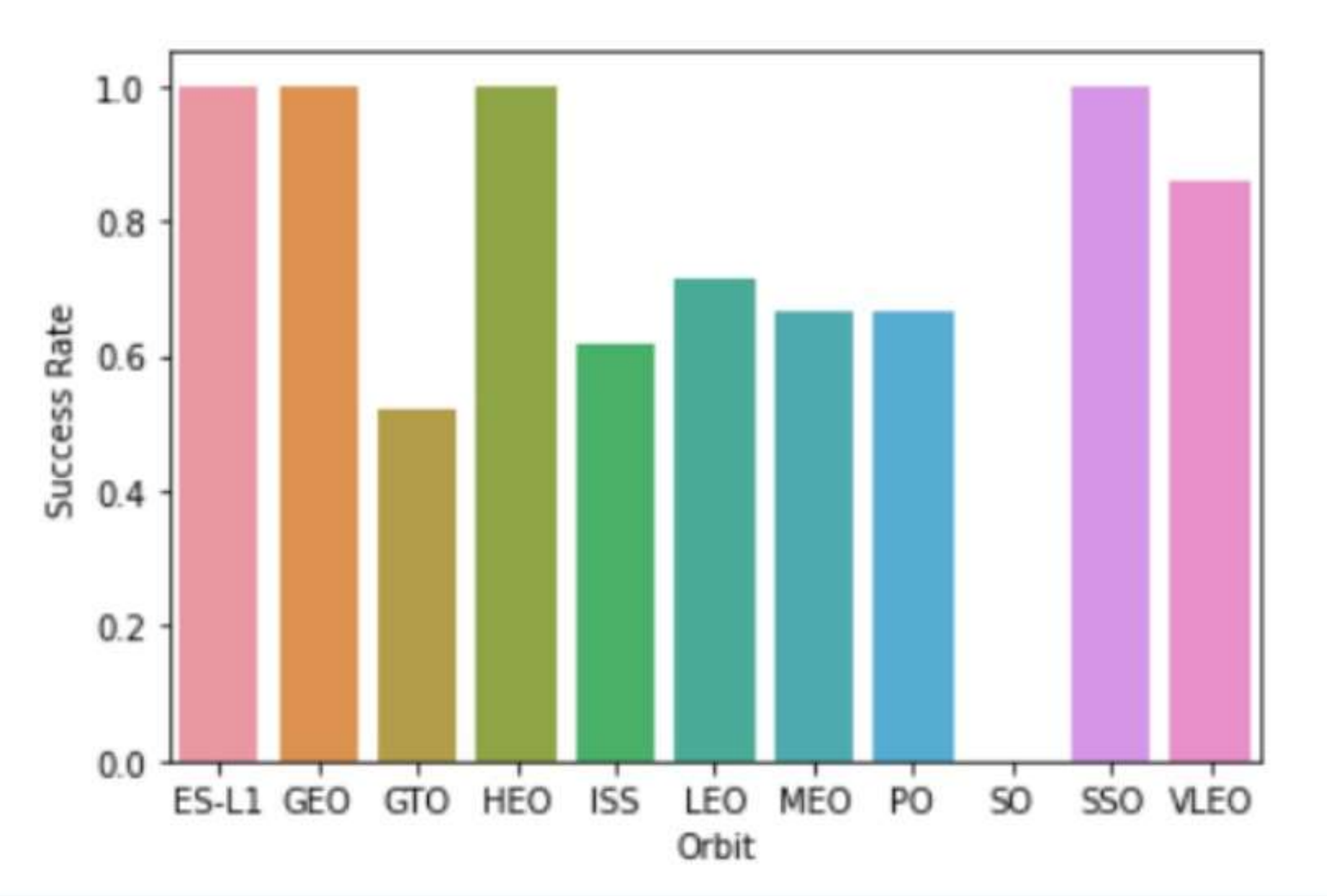
Flight Number vs. Launch Site



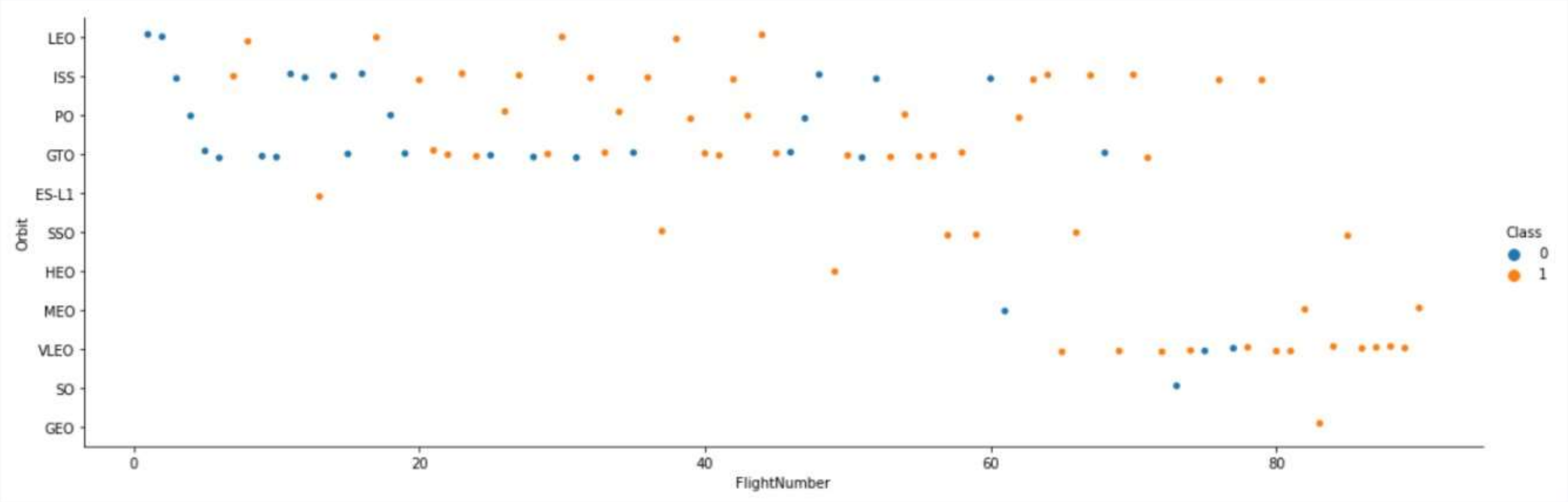
Payload vs. Launch Site



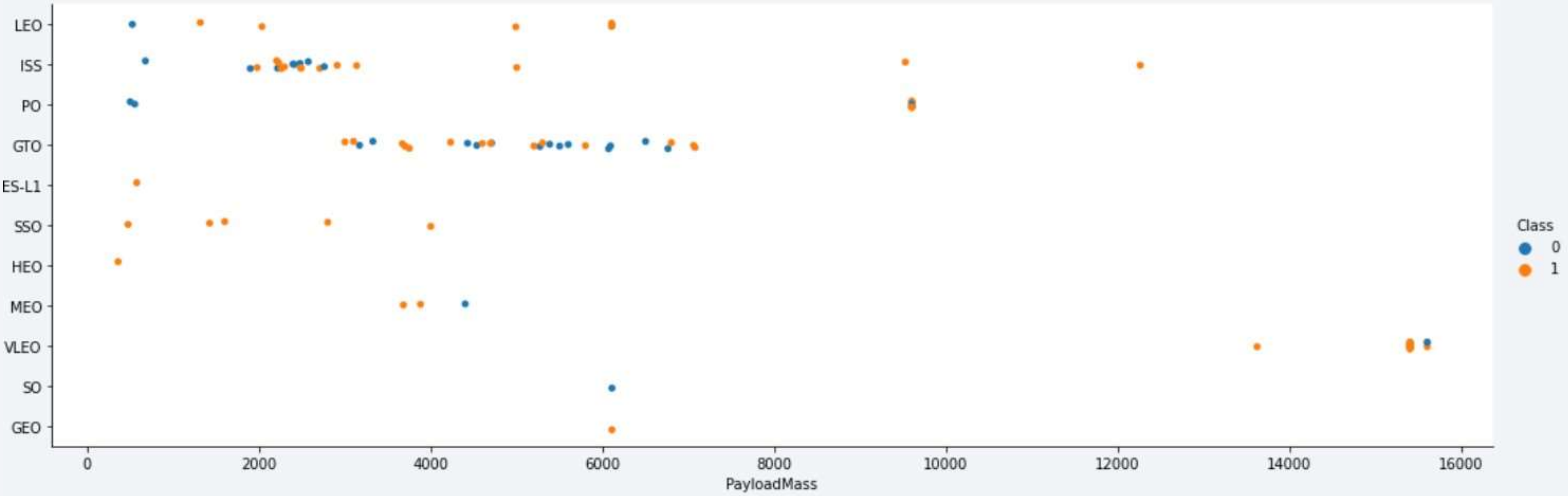
Success Rate vs. Orbit Type



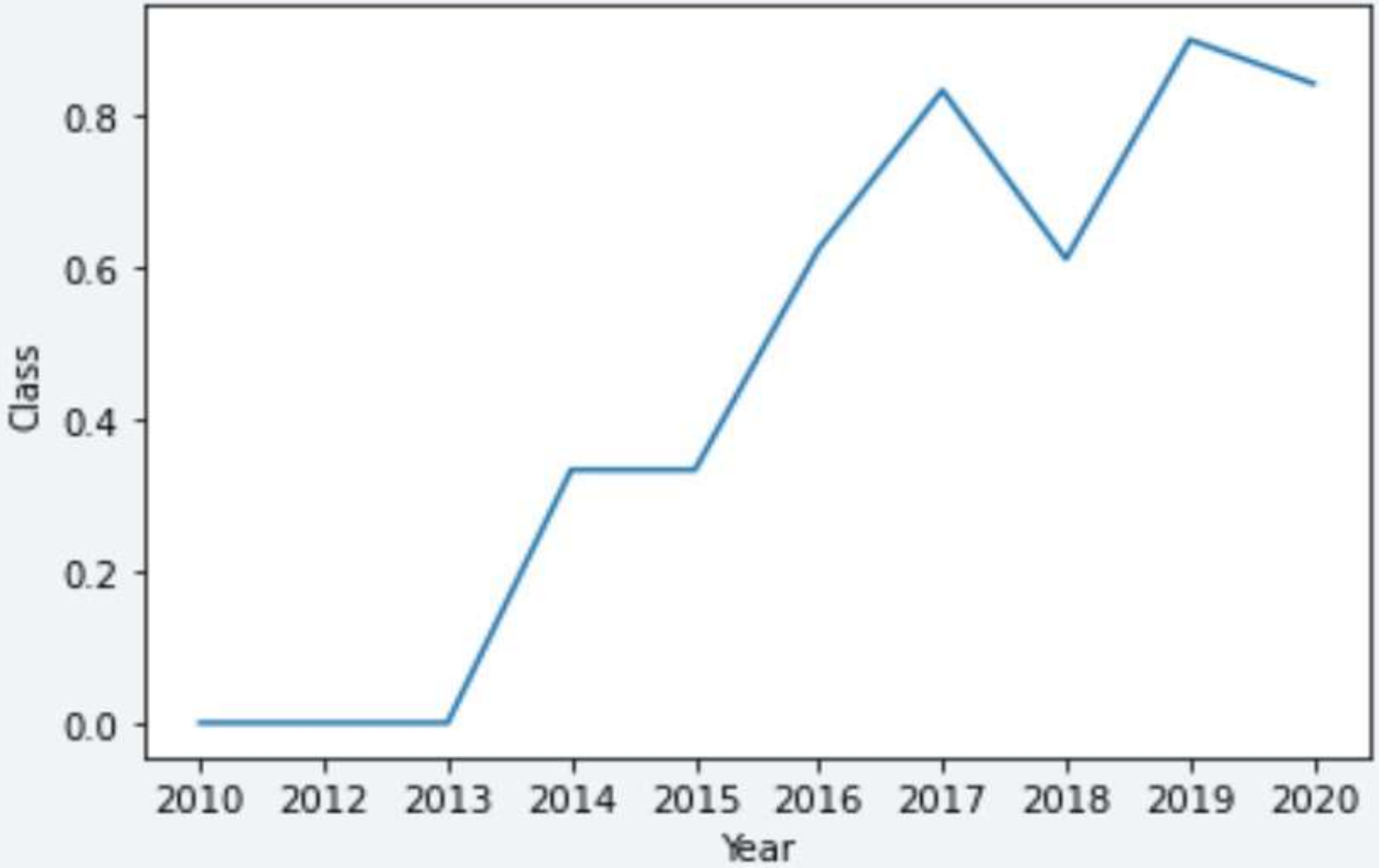
Flight Number vs. Orbit Type



Payload vs. Orbit Type



Launch Success Yearly Trend



All Launch Site Names

Task 1

Display the names of the unique launch sites in the space mission

```
%%sql

select distinct "Launch_Site" from SPACEXTBL;
```

✓ 0.1s

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

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

Use distinct call on “Launch Site” column

Launch Site Names Begin with 'CCA'

Task 2

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

```
%%sql
select * from SPACEXTBL where "Launch_Site" like 'CCA%' limit 5
```

✓ 0.1s Python

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Use LIKE to find word string ‘CCA%’ in “Launch Site” column

Total Payload Mass

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%%sql
select sum("PAYLOAD_MASS__KG_") 'Payload Mass' from SPACEXTBL where "Customer" == "NASA (CRS)"
```

✓ 0.3s

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

Payload Mass
45596

Use sum() method on “Payload mass kg” column with filter customer == NASA (CRS)

Average Payload Mass by F9 v1.1

Task 4

Display average payload mass carried by booster version F9 v1.1

```
%%sql
select avg("PAYLOAD_MASS__KG_") 'Average Mass' from SPACEXTBL where "Booster_Version" == 'F9 v1.1'
```

✓ 0.1s

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

Average Mass
2928.4

- Use avg() method on “Payload mass kg” column with filter booster version == F9 v1.1

First Successful Ground Landing Date

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%%sql
select min("Date") from SPACEXTBL where "Landing _Outcome" == "Success (ground pad)"
```

✓ 0.1s

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

min("Date")
01-05-2017

- Use min() function on “Date” column with filter landing outcome == Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%%sql
select "Booster_Version" from SPACEXTBL where ("PAYLOAD_MASS__KG_" between 4000 and 6000) and "Landing _Outcome" == "Success (drone ship)"
```

✓ 0.3s

* sqlite:///my_data1.db
Done.

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

- Get “Booster Version” column with filter “payload mass kg” between 4000 and 6000, and landing outcome == Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

```
%%sql
select "Mission_Outcome", count("Mission_Outcome") 'Count' from SPACEXTBL group by "Mission_Outcome"
```

✓ 0.2s

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

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

- Use Count() on Mission Outcome column, with group by methods on Mission Outcome

Boosters Carried Maximum Payload

Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%%sql
select "Booster_Version" from SPACEXTBL where "PAYLOAD_MASS__KG_" == (select max("PAYLOAD_MASS__KG_") from SPACEXTBL)
```

✓ 0.1s

```
* 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

- Use subquery max() to filter Booster Versions where they have max payload

2015 Launch Records

Task 9

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

```
%%sql
select substr("Date",4,2) 'Month', "Landing _Outcome","Booster_Version","Launch_Site" from SPACEXTBL where ("Landing _Outcome" == 'Failure (drone ship)') and (substr("Date",7,4) == '2015')
```

✓ 0.2s Python

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

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

- Use substr() method to extract month and year for filtering. Filters landing outome == ‘Failure (drone ship)’, year. Display booster version and the month it is used

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

Task 10

Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

```
%%sql
select "Landing _Outcome", count("Disinct _Outcome") 'Count' from SPACEXTBL where ("Landing _Outcome" in ('Success (ground pad)','Success (drone ship)','Success' )) and ("Date" between '04-04-2010' and '20-03-2017')
```

✓ 0.1s Python

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

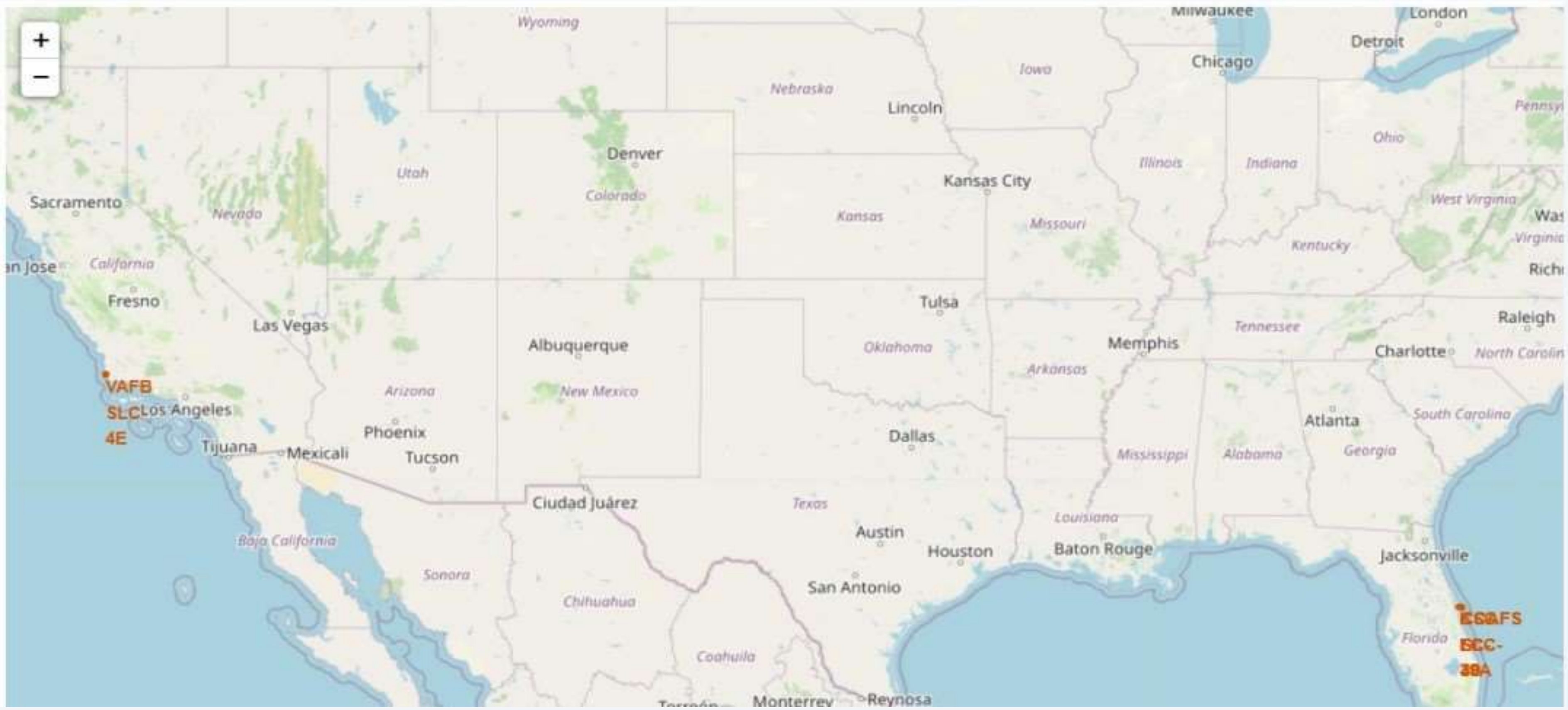
Landing _Outcome	Count
Success (drone ship)	35

Use a combination of count(), group by, filtering in (), and filtering date between

Section 3

Launch Sites Proximities Analysis

Launch Site

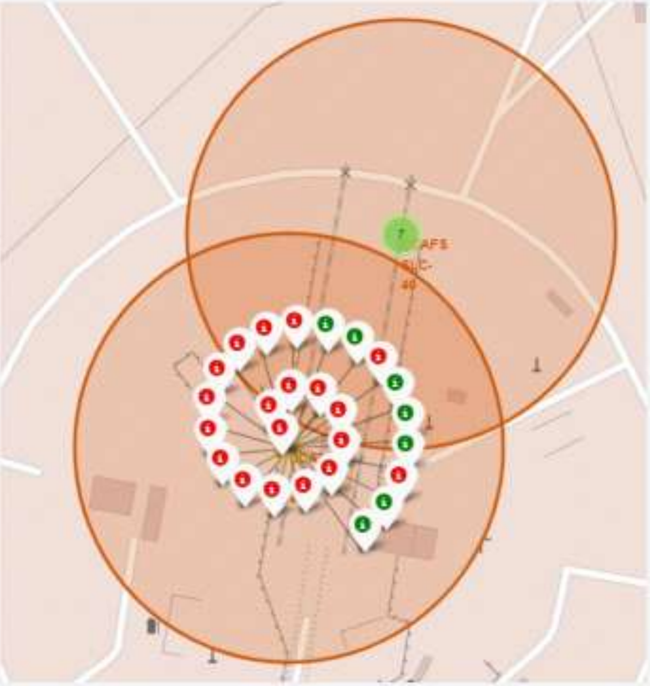


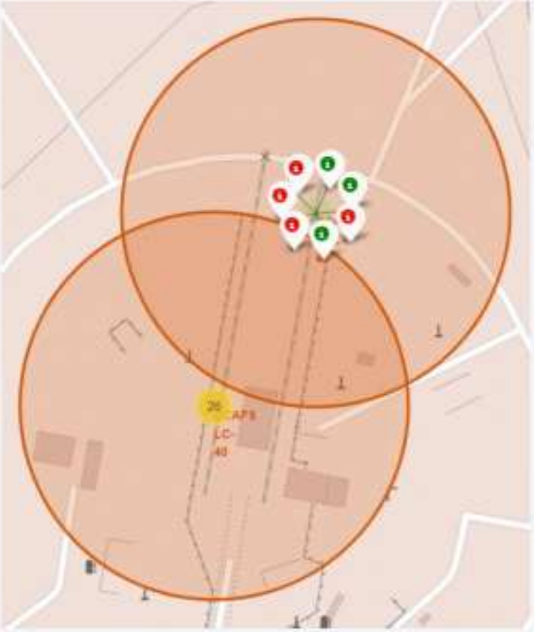
Launch Sites are near the coast for safety reasons, and near the equator line for better launch

Launch Site Landing Success Rate

- There are more launches on the west coast, as it is nearer to the equator line

Launch Sites Landing Success rate



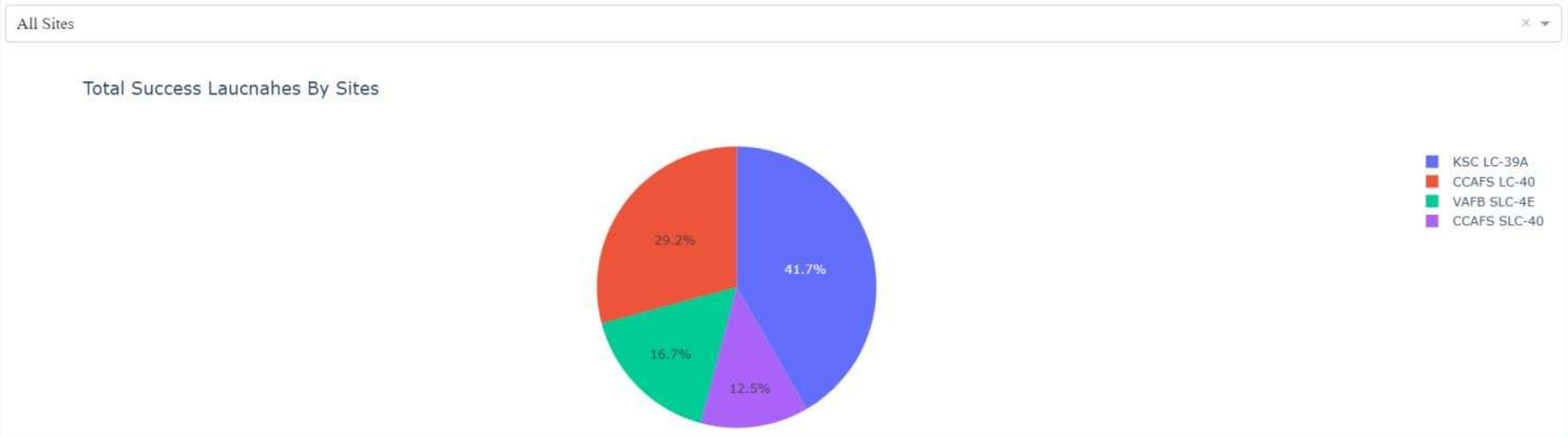


- Launch Site KSC LC 39A has the highest success rate

Section 4

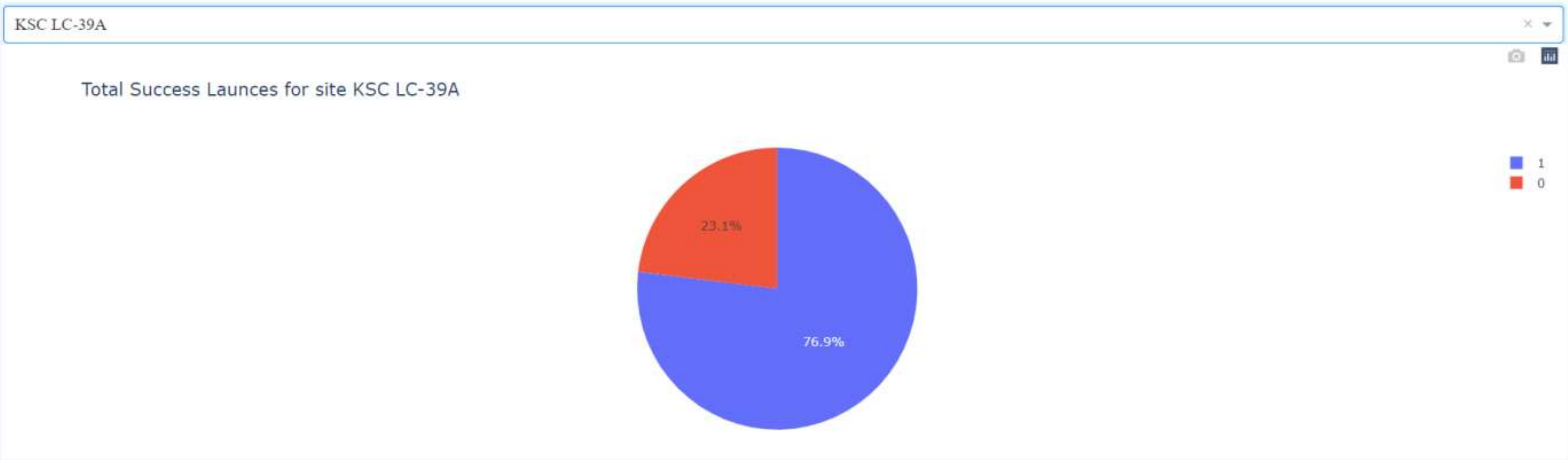
Build a Dashboard with Plotly Dash

All Sites Contribution of Success Landings



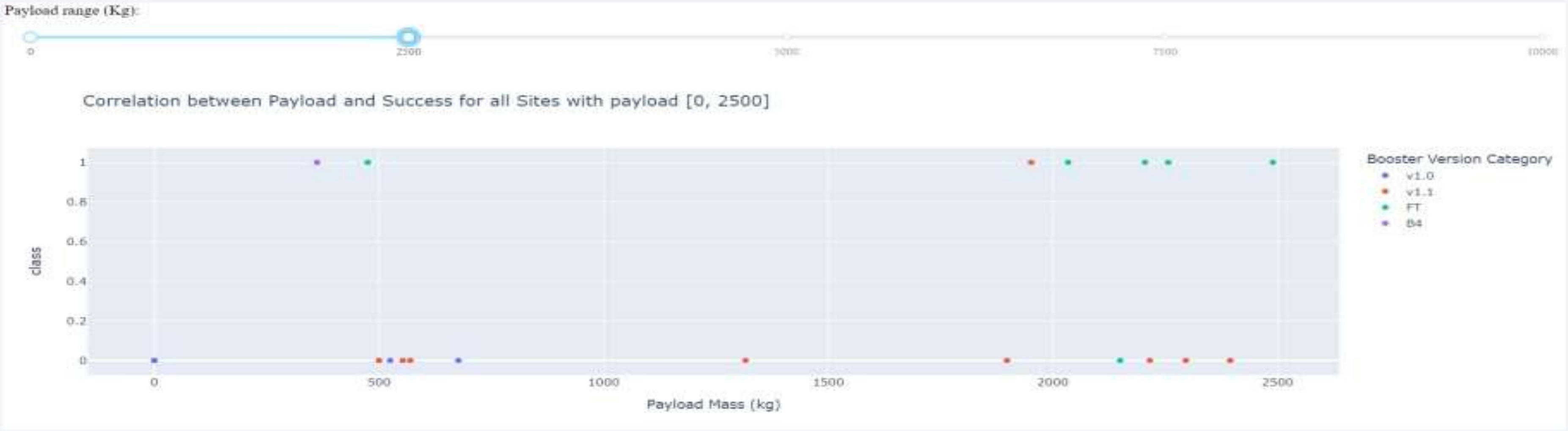
- KSC LC-39A has the most number of success landings

KSC LC-39A Success Landing Percentage



- KSC LC-39A launching site has the highest success landing percentage

Success Landings with different Booster Versions and Payloads



Success Landings with different Booster Versions and Payloads

Success Landings with different Booster Versions and Payloads

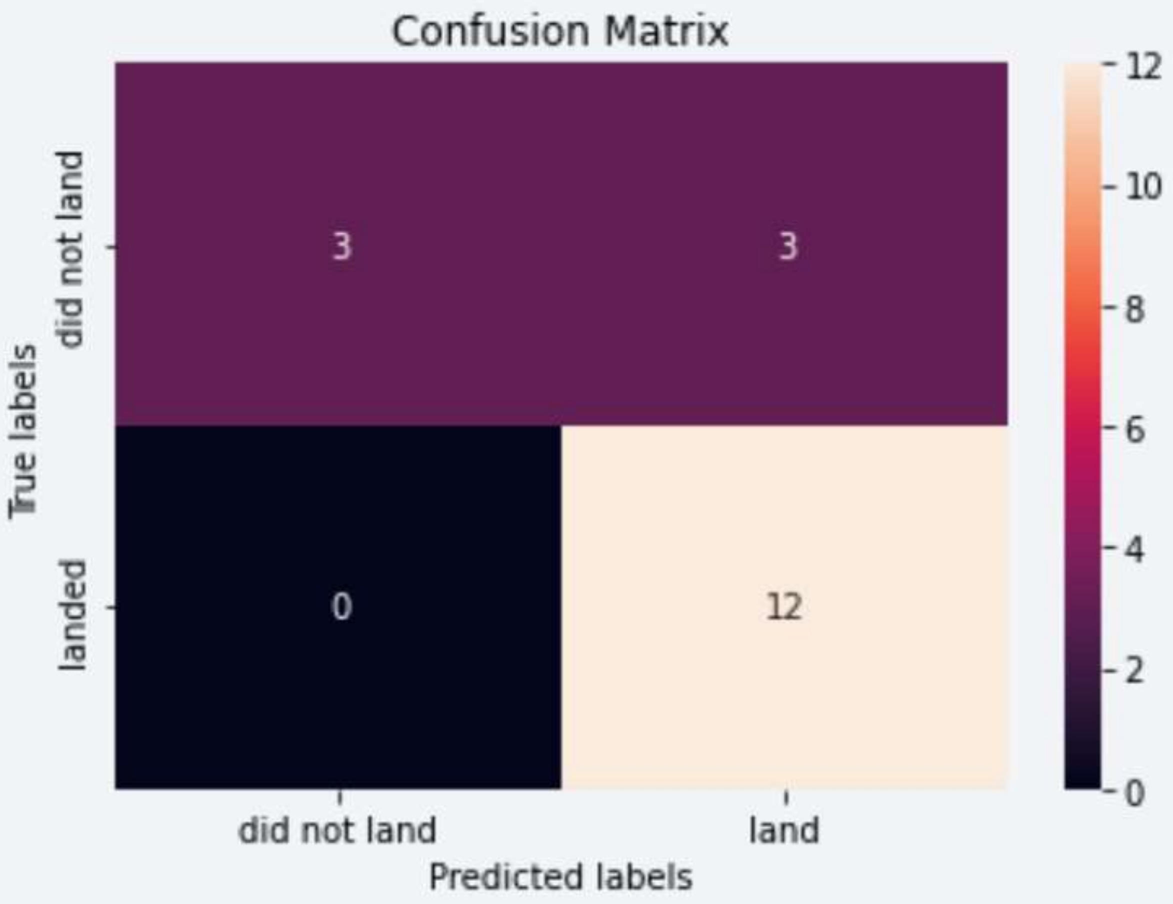
- Payload range [0,2500]: Highest number of success landings Booster Version: FT
- Payload range [2500,5000]: Highest number of success landings Booster Version: FT
- Payload range [5000,7500]: Highest number of success landings Booster Version: FT
- Payload range [7500,10000]: Highest number of success landings Booster Version: B4
- Conclusion: Version Booster FT has the highest success rate

Section 5

Predictive Analysis (Classification)

Confusion Matrix

- Decision Tree can distinguish between different classes. However the True Positive Rate (TRP) is only 0.5



Conclusions

- Launch Site with highest success landing rate and highest number and success landings: KSC LC-39A
- Highest Landing success rate for various payload masses: FT
- Launch Site position: Far from cities and near the coast for safety reasons (In case when rocket fails, it is better for it to fall into the water and far from population-dense areas). Near railroads, highways, and coast for better rocket-building-materials transportation
- Best model for Landing predictions: Decision Tree. However, TPR is low (0.5). More data is needed for a better fit. Example: Weather conditions (wind, temperature, humidity, etc..)

Appendix

- Raw Data Set:

  https://github.com/YasteneAitKheddache/capstone/blob/950fdd3d061baf35dd8e1382a90a32f4791a2caa/dataset_part_1.csv

https://github.com/YasteneAitKheddache/capstone/blob/950fdd3d061baf35dd8e1382a90a32f4791a2caa/dataset_part_2.csv

https://github.com/YasteneAitKheddache/capstone/blob/950fdd3d061baf35dd8e1382a90a32f4791a2caa/dataset_part_3.csv

Thank you!