IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Gabriel Pinho

19-mar-2024

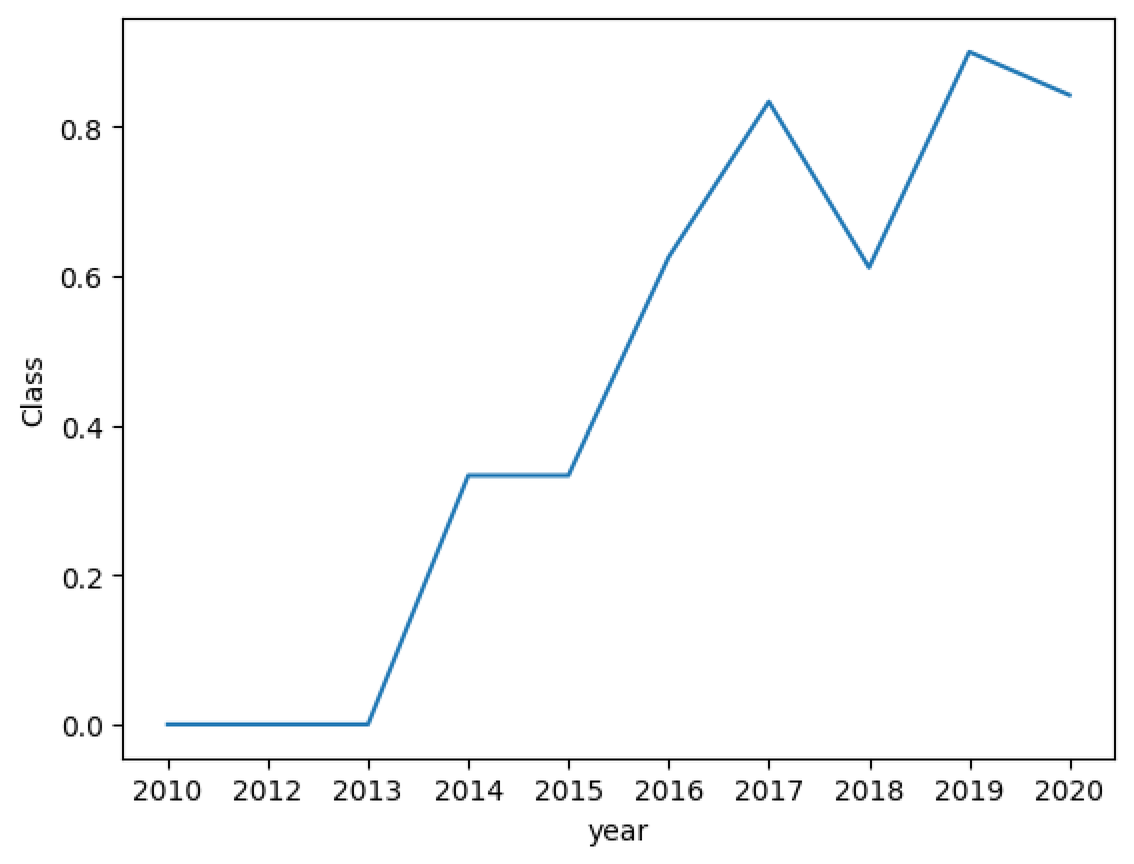
Outline

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

Executive Summary

- Summary of methodologies
  - EDA with SQL
  - EDA with data visualization
  - Maps with Folium
  - DashBoard with Plotly Dash

The main factor of success in launches is the experience gain in the firsts attempts, like we can below:



Introduction

- Project background and context
  - SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.
- Problems you want to find answers
  - if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch. In this project, we will predict if the Falcon 9 first stage will land successfully

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collecting using spacex API
- Perform data wrangling
  - Several functions were created to process the data
- 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 was collecting using spacex API with python library request and the URL https://api.spacexdata.com/v4/launches/past
- Several functions were created to process the data, mainly to:
  - filter only to include Falcon9 launches
  - Replace null entries in PayloadMass with its mean value
  - and to transform cathegorical data into numerical ones using one hot encoder (Pandas get_dummies). Details of the functions in the flow chart next slide.

Data Collection – SpaceX API

Notebook with Data Collection

- Estudos.IBM_Data_Science/10_Applied_Data_Science_Capstone/notebooks/01_jupyter-labs-spacex-data-collection-api.ipynb at main · gaugustop/Estudos.IBM_Data_Science (github.com)

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

```
response = requests.get(spacex_url)
```

```
json = response.json()
```

```
data = pd.json_normalize(json)
```

```
getBoosterVersion
getLaunchSite
getPayloadData
getCoreData
```

Filter to only Falcon 9

dataset_part_1.csv

Data Collection - Scraping

Notebook with Data Collection

- Estudos.IBM_Data_Science/10_Applied_Data_Science_Capstone/notebooks/01_jupyter-labs-spacex-data-collection-api.ipynb at main · gaugustop/Estudos.IBM_Data_Science (github.com)

```python
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

↓

```python
response = requests.get(spacex_url)
```

```python
json = response.json()
```

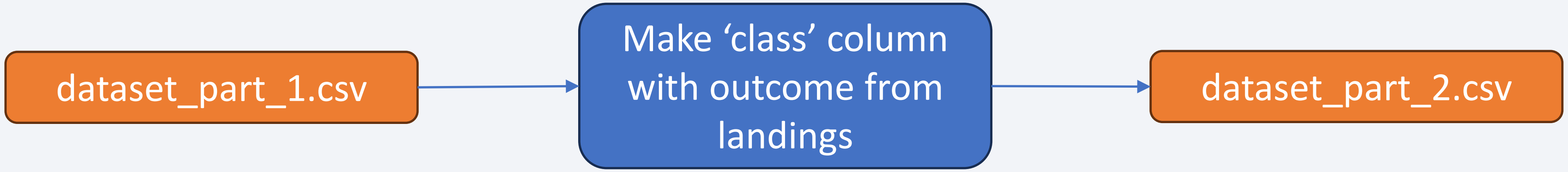
↓

```python
data = pd.json_normalize(json)
```

Data Wrangling

Github Link to data Wrangling Jupyter Notebook

- Estudos.IBM_Data_Science/10_Applied_Data_Science_Capstone/notebooks/02_labs-jupyter-spacex-Data wrangling.ipynb at main · gaugustop/Estudos.IBM_Data_Science (github.com)



EDA with Data Visualization

- Charts ploted (to see if there is correlation between those variables)
  - Scatter plot: Pay load mass vs Flight Number
  - Scatter plot: Launch Site vs Flight Number
  - Scatter plot: Launch Site vs Payload mass
  - Bar plot: Class vs Orbit
  - Scatter plot: Orbit vs Flight Number
  - Scatter plot: Orbit vs Payload mass
  - Line chart: Class vs Year

Github Jupyter Notebook:

Estudos.IBM_Data_Science/10_Applied_Data_Science_Capstone/notebooks/04_jupyter-labs-eda-dataviz.ipynb at main · gaugustop/Estudos.IBM_Data_Science (github.com)

EDA with SQL

- SQL queries performed:
  - select distinct(Launch_Site) from spacextable
  - select * from spacextable where Launch_Site like 'CCA%' limit 5
  - select Customer, sum(PAYLOAD_MASS__KG_) as 'total payload' from spacextable where Customer = "NASA (CRS)“
  - select distinct(Booster_Version) from spacextable where Booster_Version like 'F9 v1.1%'
  - select round(avg(PAYLOAD_MASS__KG_),3) from spacextable where Booster_Version like 'F9 v1.1%'
  - select min(Date) as 'Date of first succesful landing' from spacextable where Landing_Outcome like '%Success%'
  - select Booster_Version from spacextable where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000
  - select Mission_Outcome, count(Mission_Outcome) as 'Count' from spacextable group by Mission_Outcome
  - select Booster_Version from spacextable where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from spacextable)
  - select Date, Booster_Version, Launch_Site, Landing_Outcome, substr(Date, 6, 2) as month from spacextable where Landing_Outcome = 'Failure (drone ship)' and substr(Date,0,5)='2015’
  - select Landing_Outcome, count(Landing_Outcome) as Count from spacextable group by Landing_Outcome having Date between '2010-06-04' and '2017-03-20' order by Count desc

EDA with SQL

Github Jupyter Notebook:

Estudos.IBM_Data_Science/10_Applied_Data_Science_Capstone/notebooks/03_jupyter-labs-eda-sql-coursera_sqllite.ipynb at main · gaugustop/Estudos.IBM_Data_Science (github.com)

Build an Interactive Map with Folium

- Interactive maps were created using Folium, a Python Library. In order to better understand the geographic influence in launch success there were added to the maps: **markers, marker clusters, distance from coastline and lines.**

Github Jupyter Notebook:
Estudos.IBM_Data_Science/10_Applied_Data_Science_Capstone/notebooks/05_lab_jupyter_launch_site_location.ipynb at main · gaugustop/Estudos.IBM_Data_Science (github.com)

Build a Dashboard with Plotly Dash

- In the Dashboard with Plotly Dash, we have one dropdown to chose the Lauch Site, then a Pie chart showing the successful count per launch site (if one site is chosen then we see the proportion of success/failure of that site). Below that we have a payload mass range the user can chose and a scatter plot with class vs payload mass. This scatter plot also interacts with the dropdown.

Github Python file:

[Estudos.IBM_Data_Science/10_Applied_Data_Science_Capstone/app/spacex_dash_app.py at main · gaugustop/Estudos.IBM_Data_Science (github.com)]

Predictive Analysis (Classification)

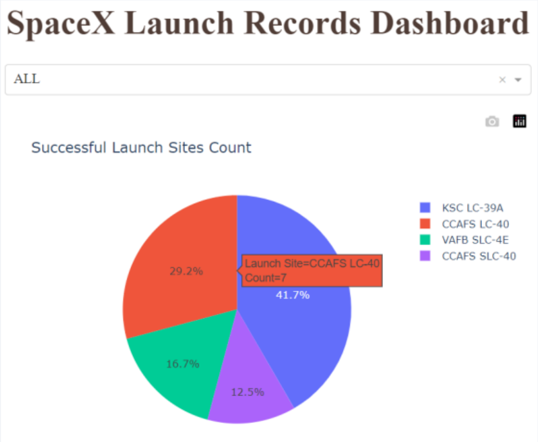
- Summarize how you built, evaluated, improved, and found the best performing classification model
- The data were splitted into train and test (test size = 20%). For each model, there were created a GridSearch with cv = 10, some parameters were tested and the score was taken with the test set, also a confusion matrix was created for each model.
- Some classification models were tested:
  - Logistic regression
  - Support Vector Machine
  - Decision Tree Classifier
  - K Nearest Neighbors

Github Jupyter Notebook:

[Estudos.IBM_Data_Science/10_Applied_Data_Science_Capstone/notebooks/06_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb at main · gaugustop/Estudos.IBM_Data_Science (github.com)]

Results

- Exploratory data analysis results
  - Exploratory data analysis shown that we have a correlation between the launch site, number of the flight and payload mass in the success launch
- Interactive analytics with Plotly Dash



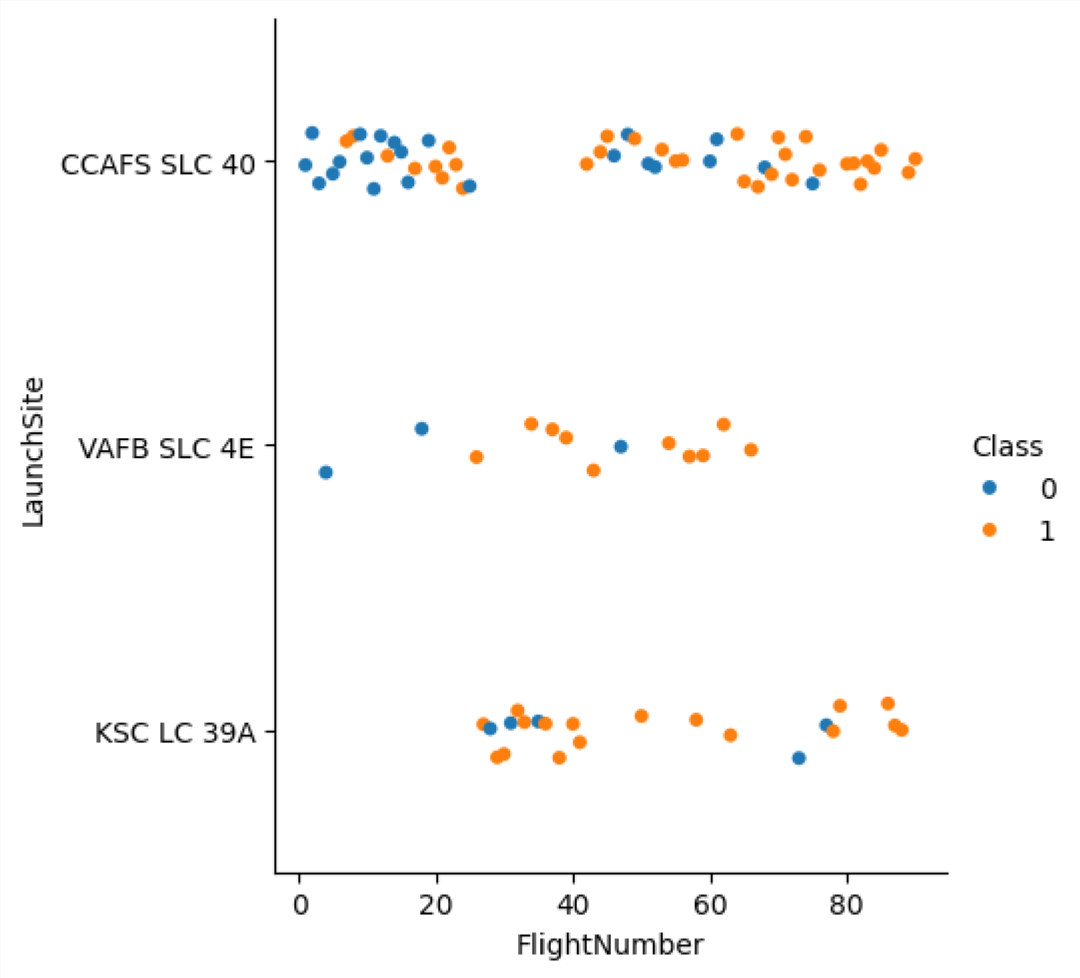
- Predictive analysis results
  - The best models tested were Logistic Regression, Support Vector Machine and K-Nearest Neighbors

Section 2

Insights drawn from EDA

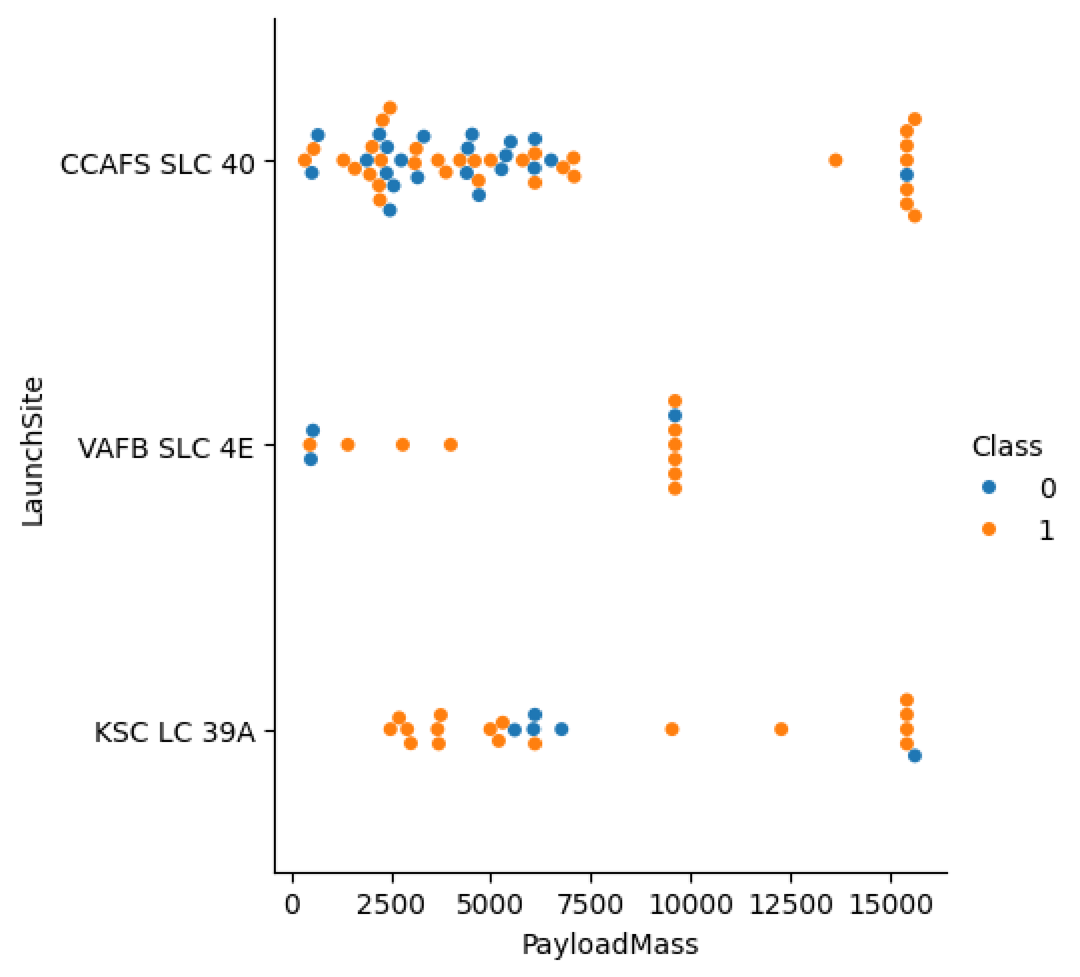
Flight Number vs. Launch Site

- Scatter plot of Flight Number vs. Launch Site



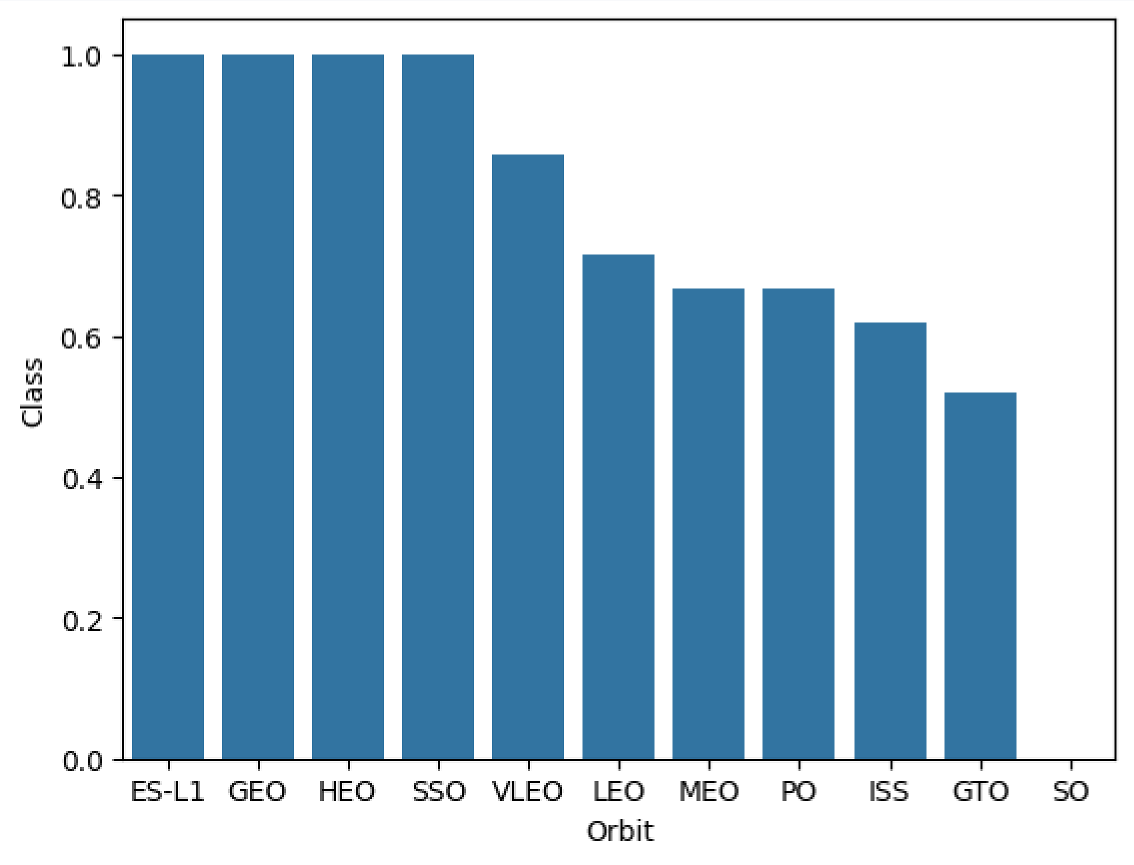
Payload vs. Launch Site

- Scatter plot of Payload vs. Launch Site



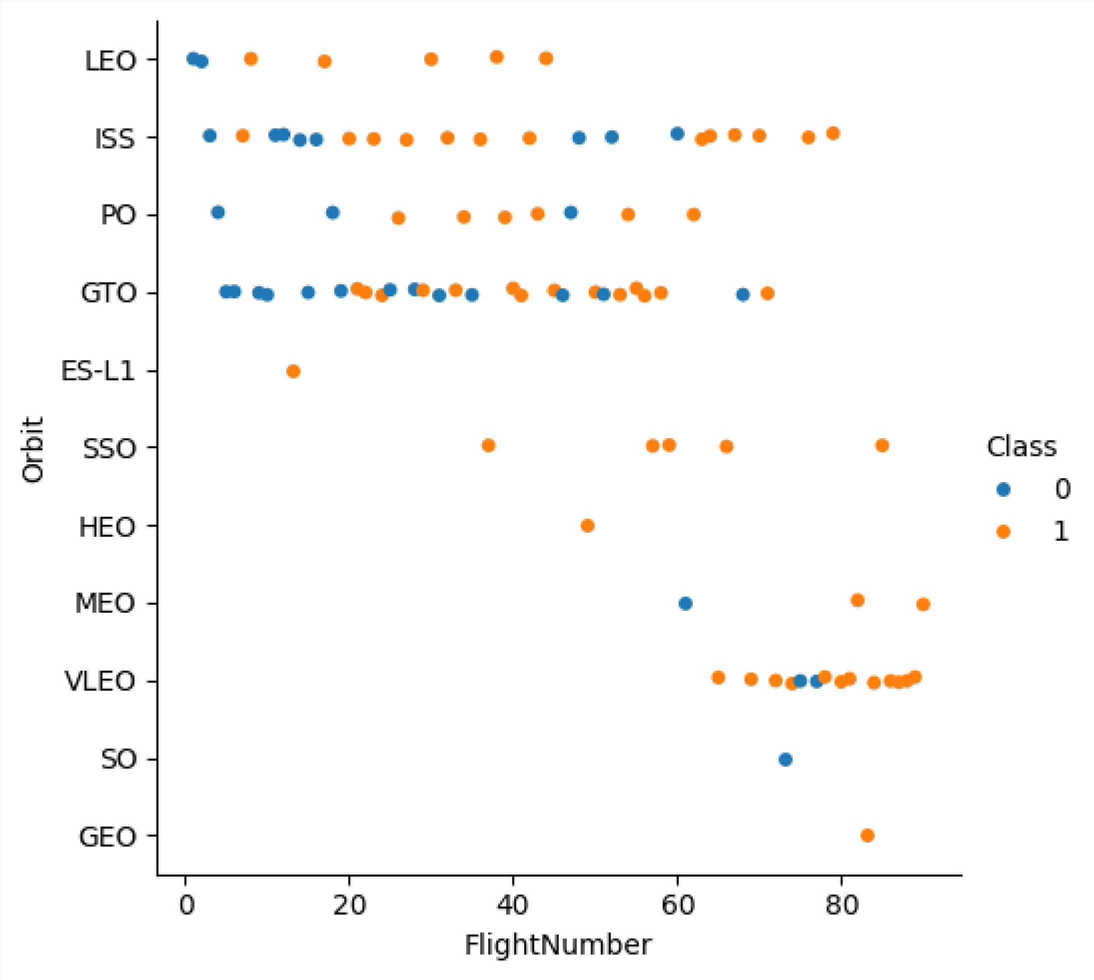
Success Rate vs. Orbit Type

- Bar chart for the success rate of each orbit type



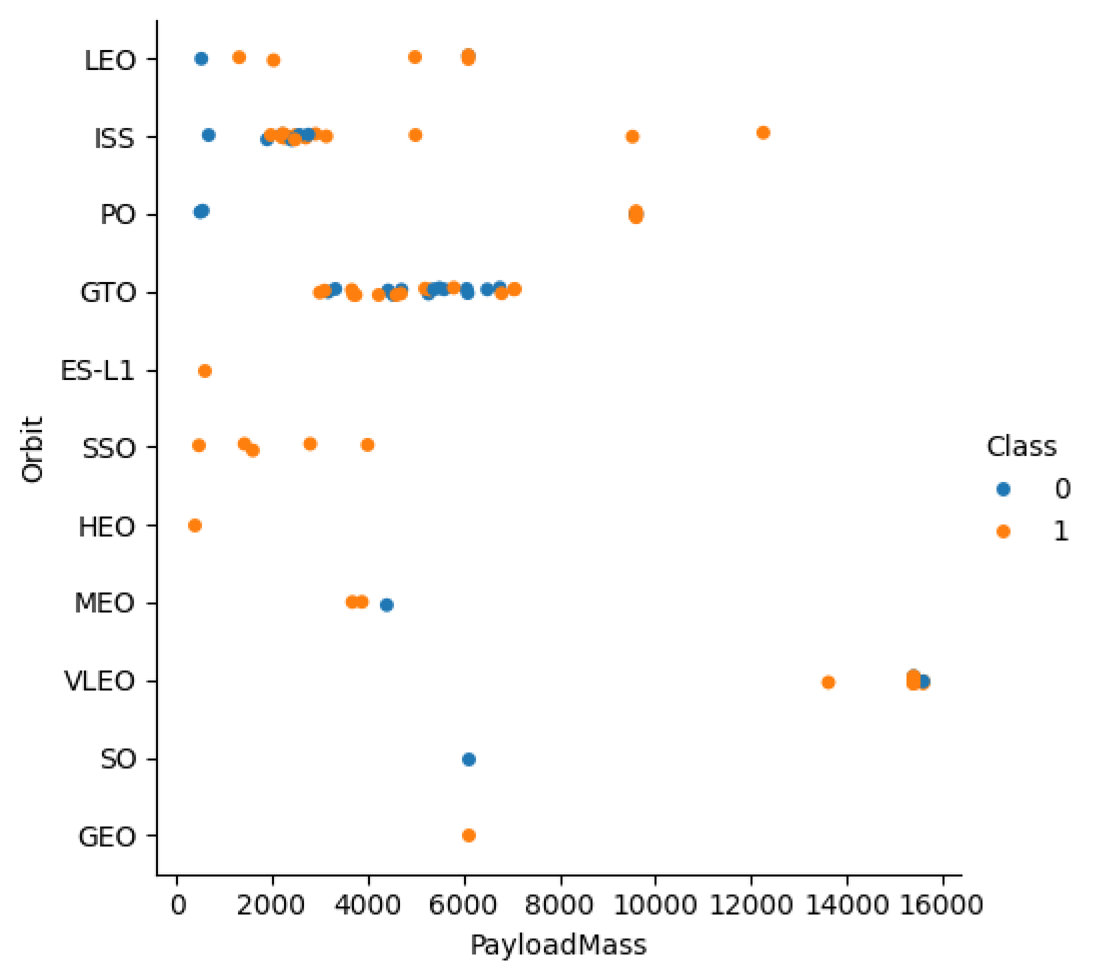
Flight Number vs. Orbit Type

- Scatter plot of Flight number vs. Orbit type



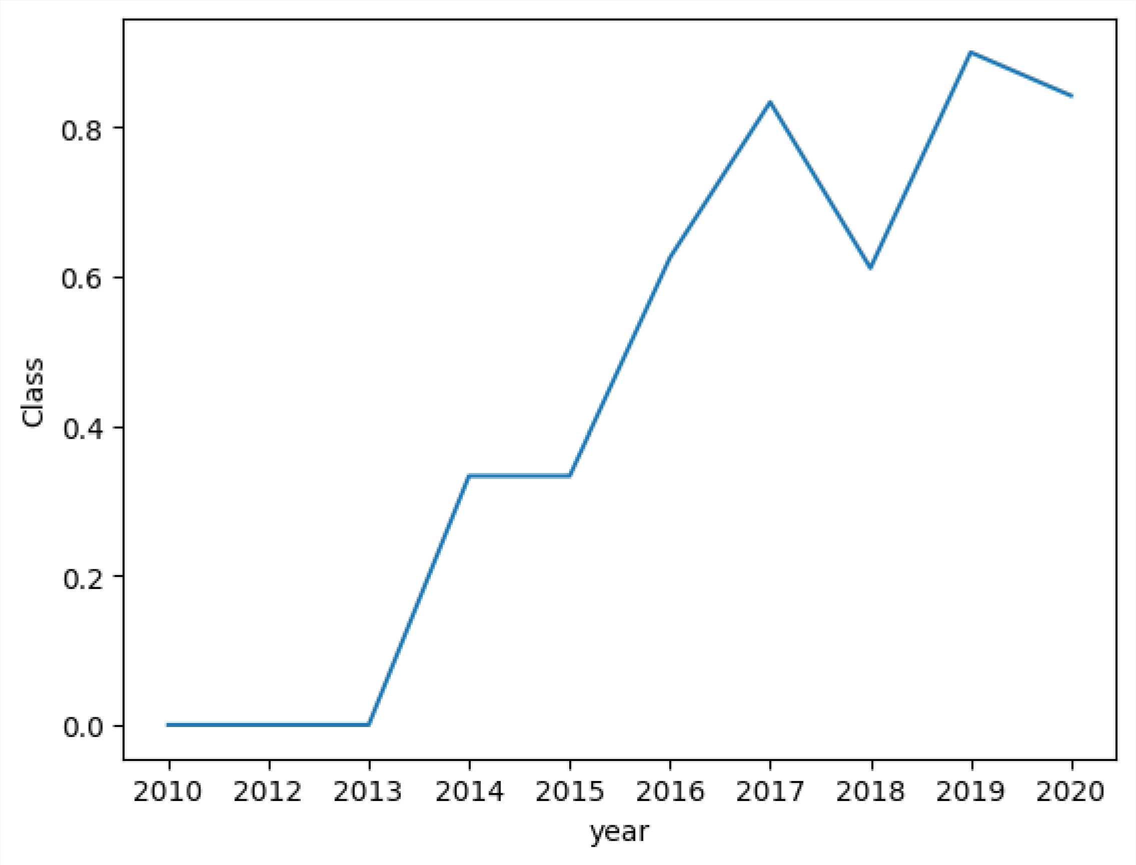
Payload vs. Orbit Type

- Scatter plot of payload vs. orbit type



Launch Success Yearly Trend

- Line chart of yearly average success rate



All Launch Site Names

- Names of the unique launch sites
- *select distinct(Launch_Site) from spacextable*

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`
- *select * from spacextable where Launch_Site like 'CCA%' limit 5*

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	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)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- select Customer, sum(PAYLOAD_MASS__KG_) as 'total payload' from spacextable where Customer = "NASA (CRS)"

Customer	total payload
NASA (CRS)	45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- select round(avg(PAYLOAD_MASS__KG_),3) from spacextable where Booster_Version like 'F9 v1.1%'

round(avg(PAYLOAD_MASS__KG_),3)
2534.667

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- select min(Date) as 'Date of first succesful landing' from spacextable where Landing_Outcome like '%Success%'

Date of first succesful landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- select Booster_Version from spacextable where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- select Mission_Outcome, count(Mission_Outcome) as 'Count' from spacextable group by Mission_Outcome

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- select Booster_Version from spacextable where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from spacextable)

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

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- select Date, Booster_Version, Launch_Site, Landing_Outcome, substr(Date, 6, 2) as month from spacextable where Landing_Outcome = 'Failure (drone ship)' and substr(Date,0,5)='2015'

Date	Booster_Version	Launch_Site	Landing_Outcome	month
2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)	01
2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)	04

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

- Rank 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
- select Landing_Outcome, count(Landing_Outcome) as Count from spacextable group by Landing_Outcome having Date between '2010-06-04' and '2017-03-20' order by Count desc

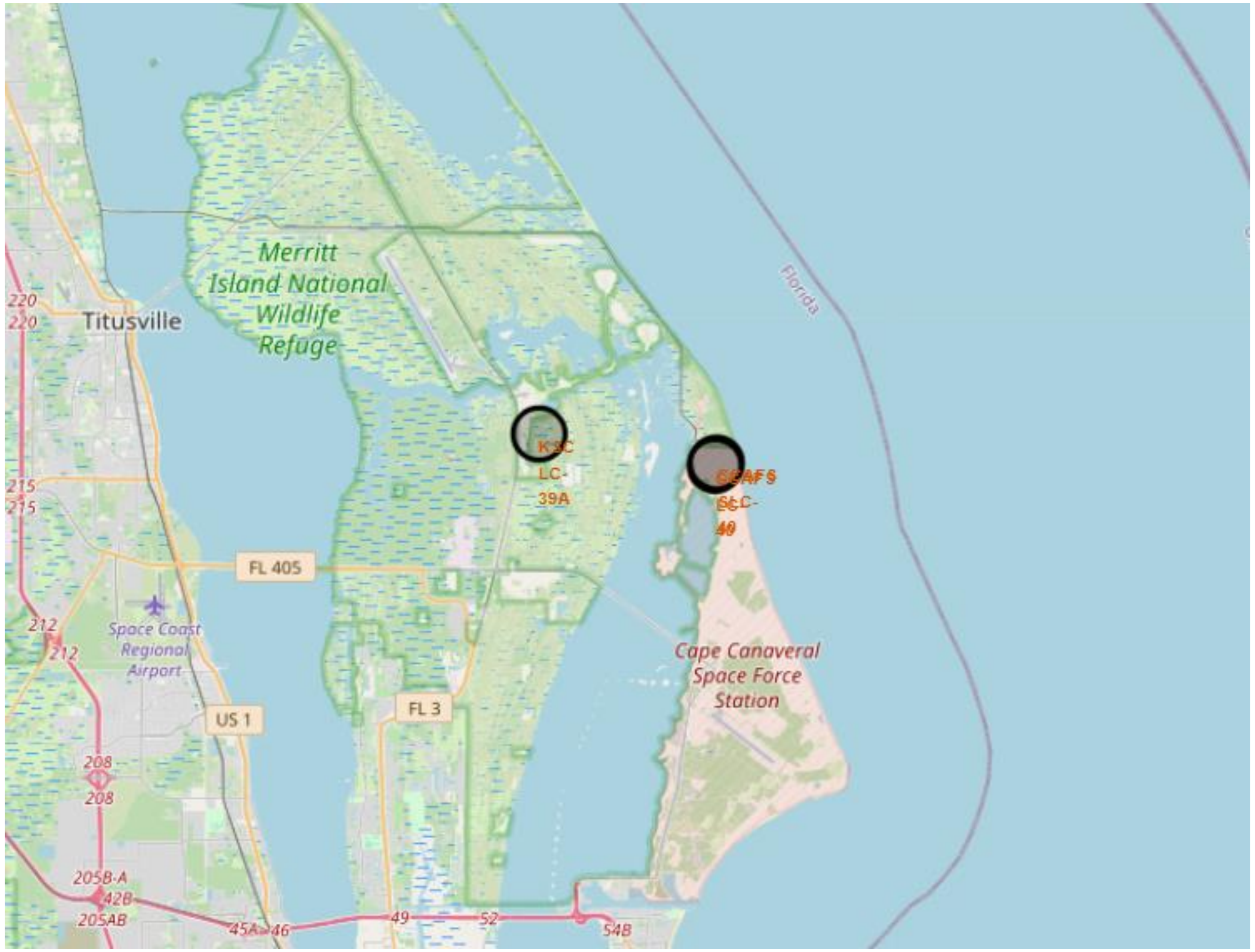
Landing_Outcome	Count
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

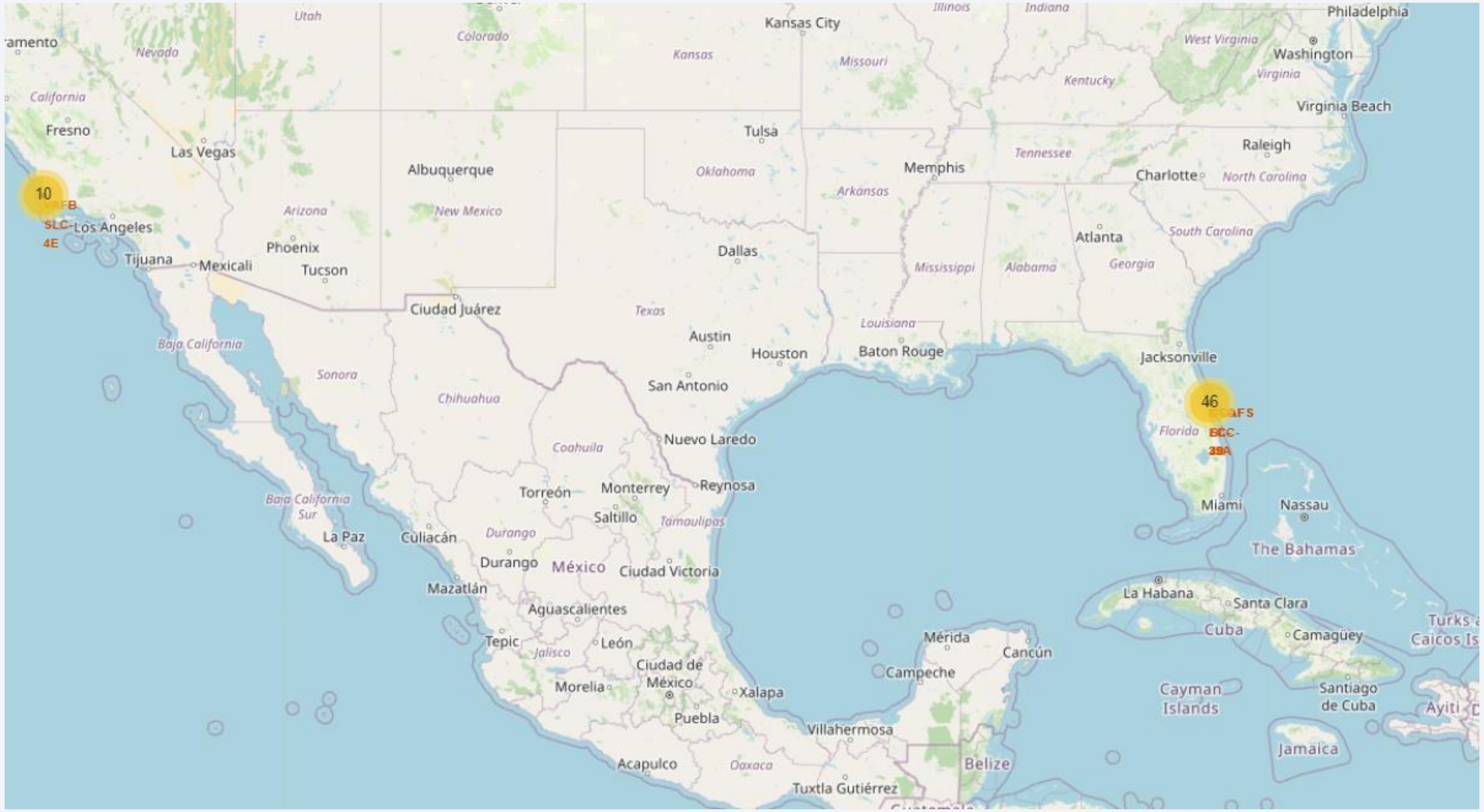
Launch Sites Proximities Analysis

Launch sites map

Launch sites map

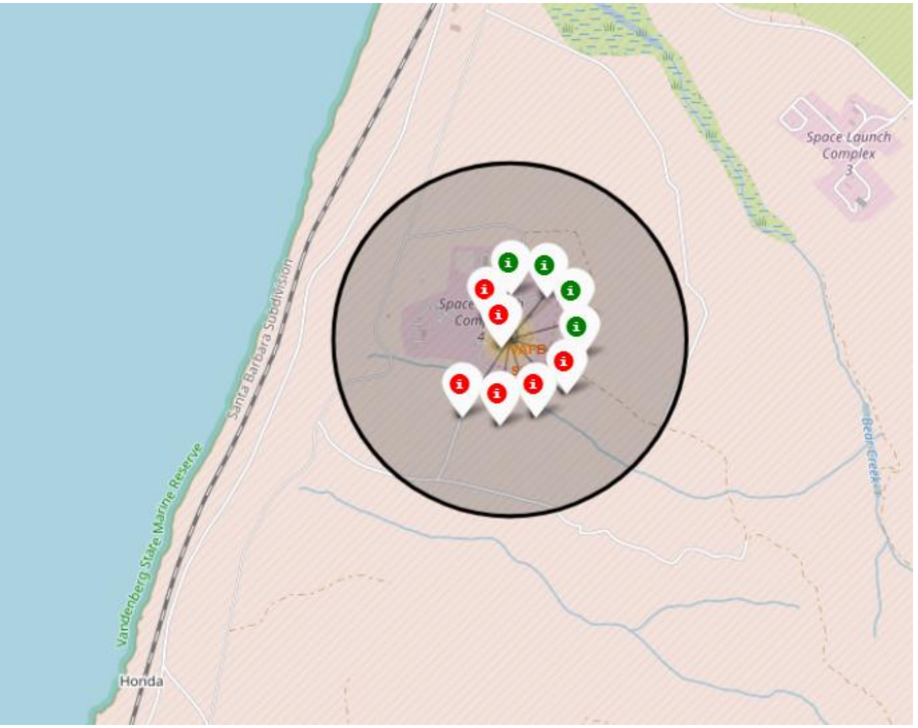


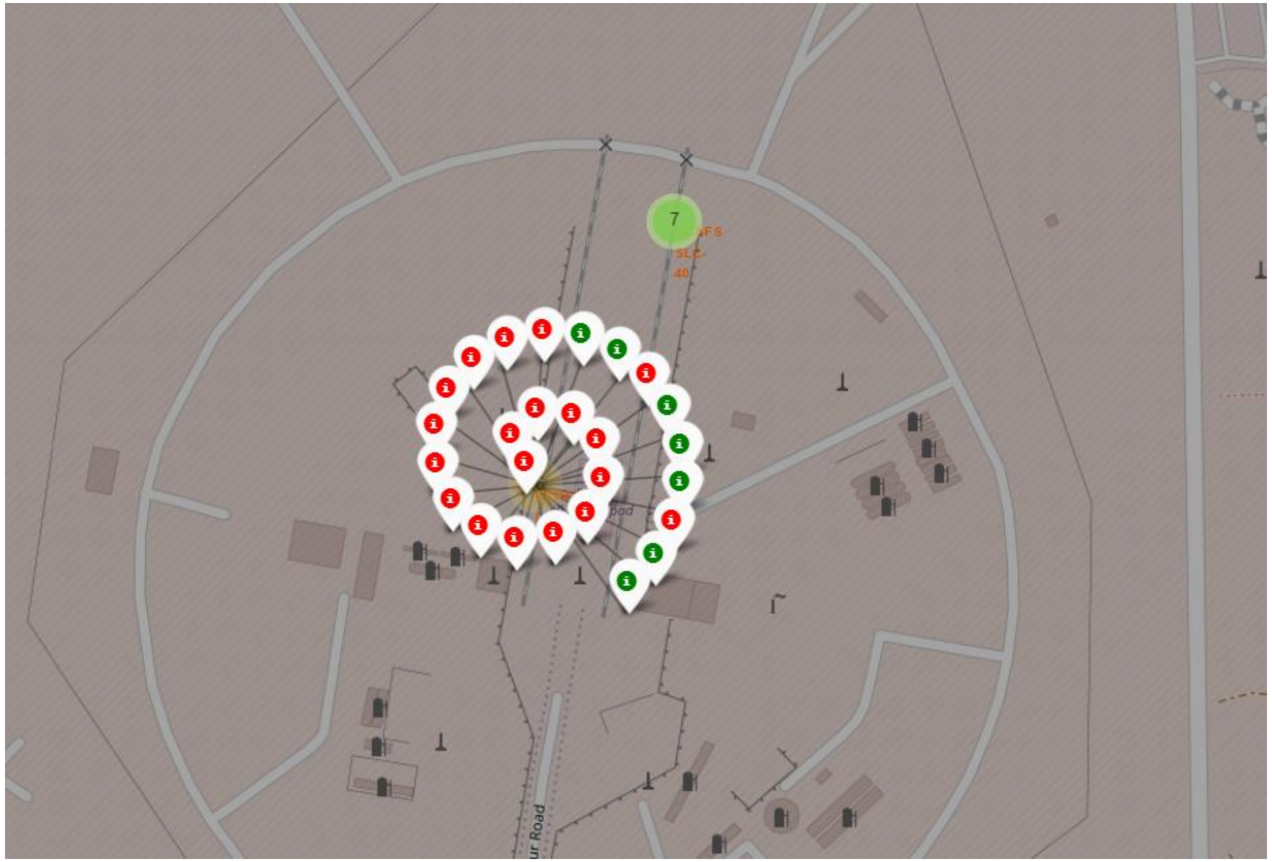
Launch Sites with outcomes



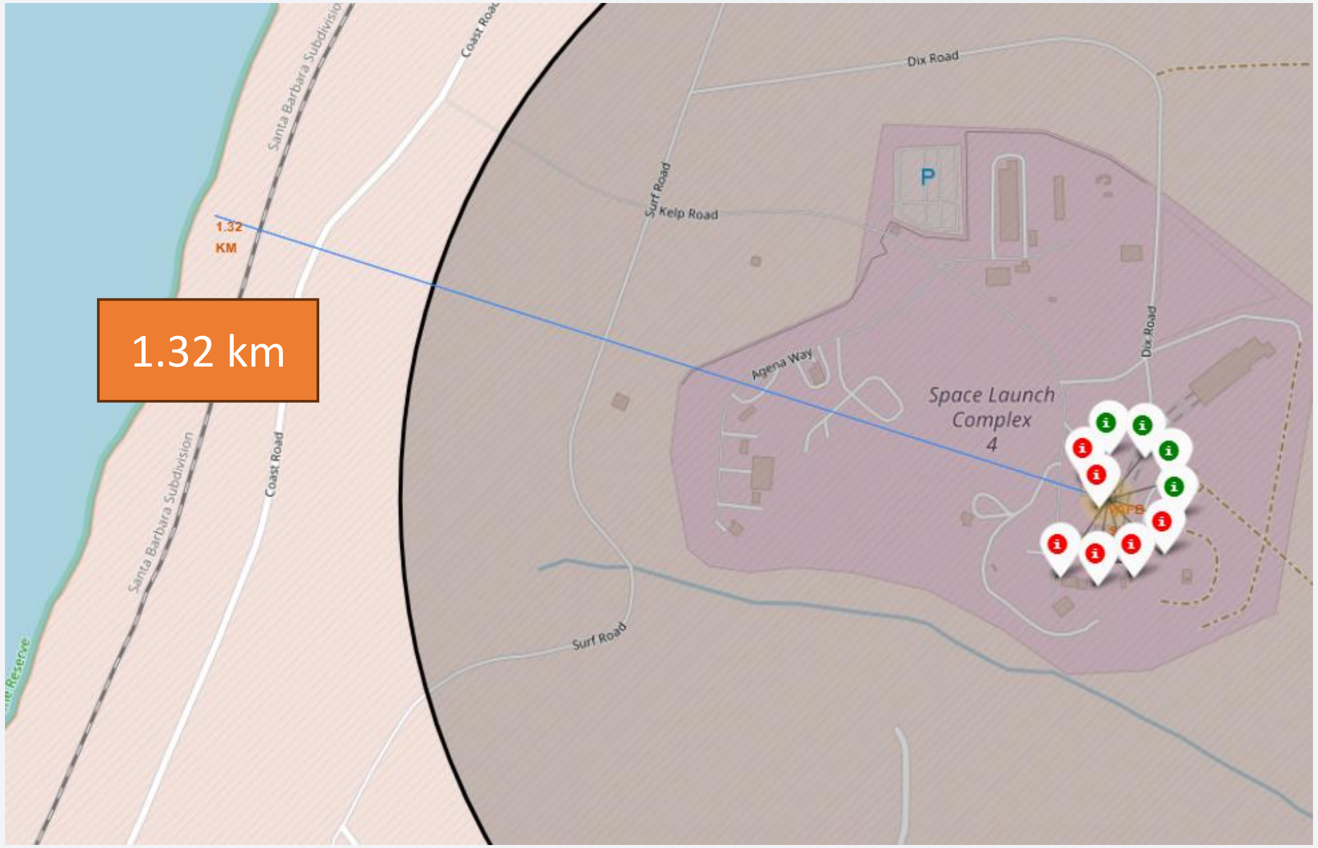
PÚBLICA

Launch Sites with outcomes





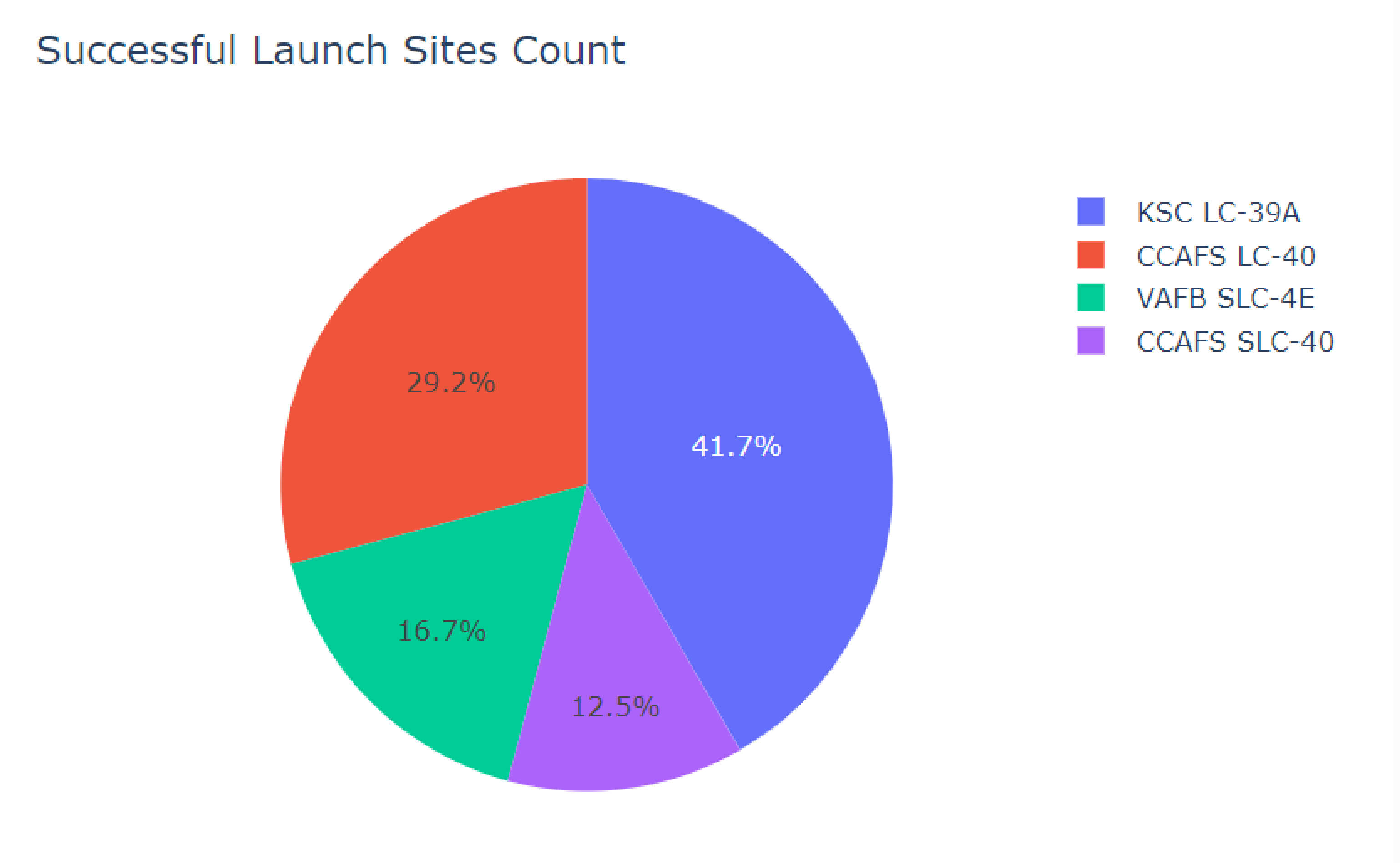
Distance from coastline



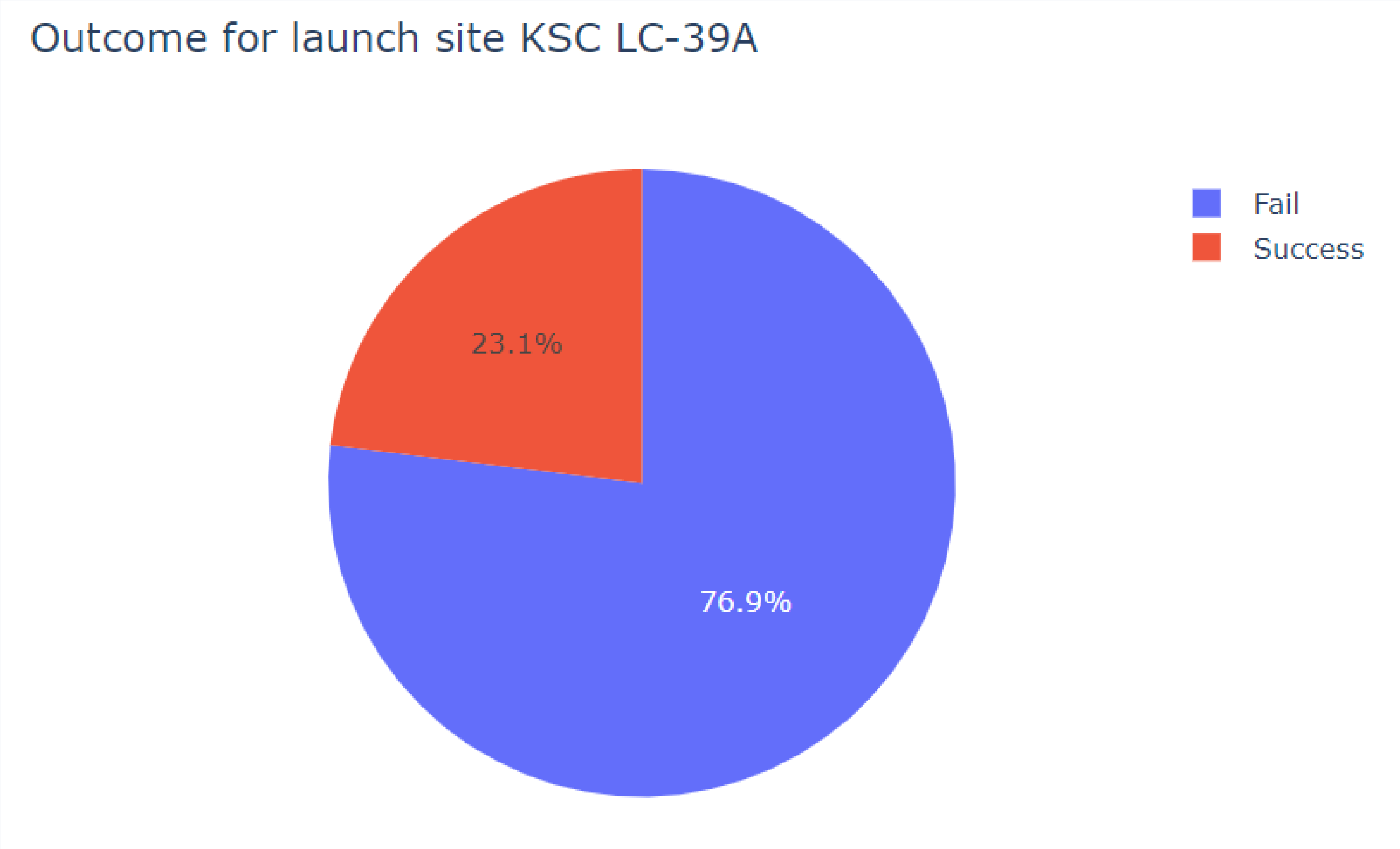
Section 4

Build a Dashboard with Plotly Dash

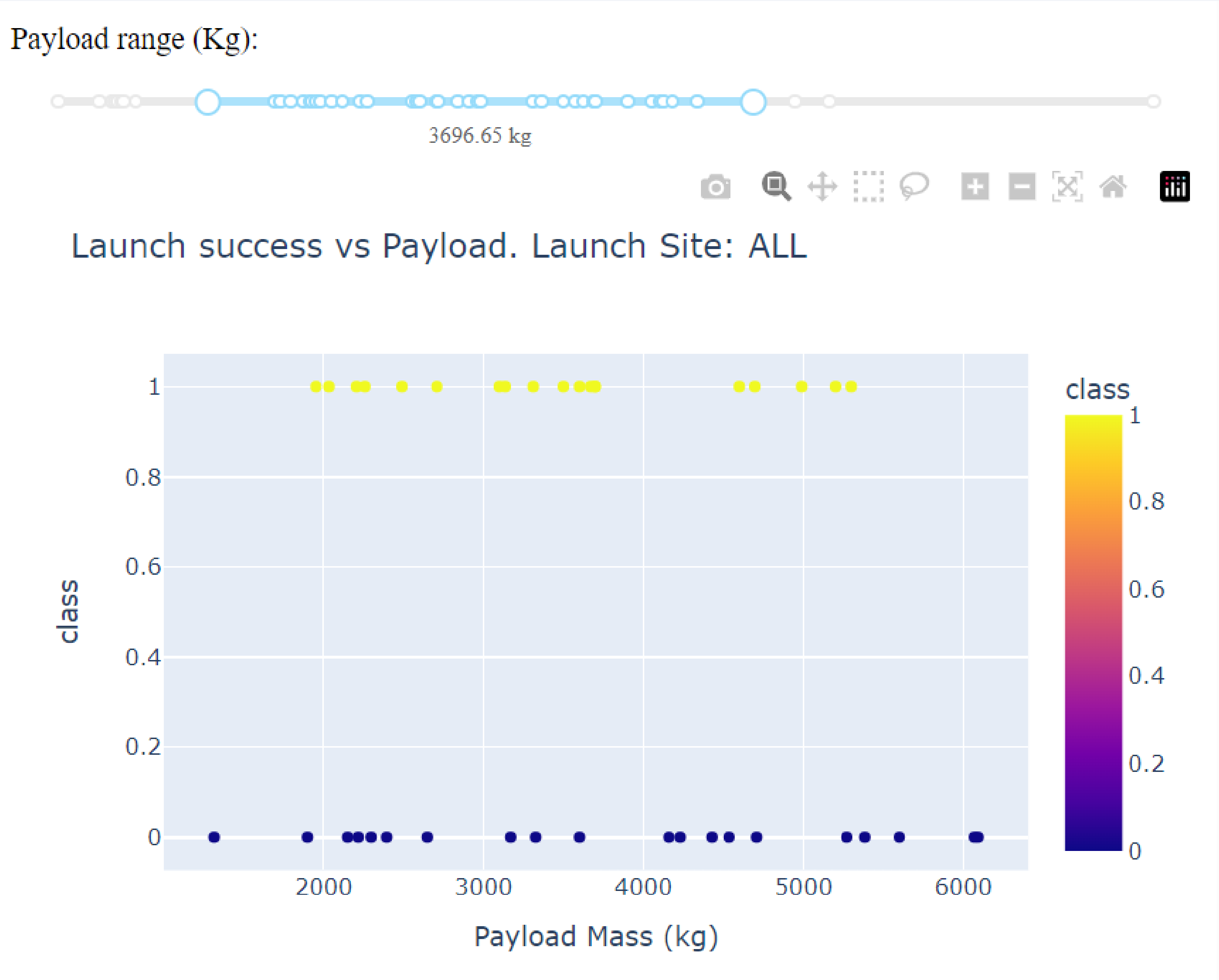
Successful Launch Sites Count



Outcome for launch site with highest success rate



Payload Mass vs Launch outcome (all sites)



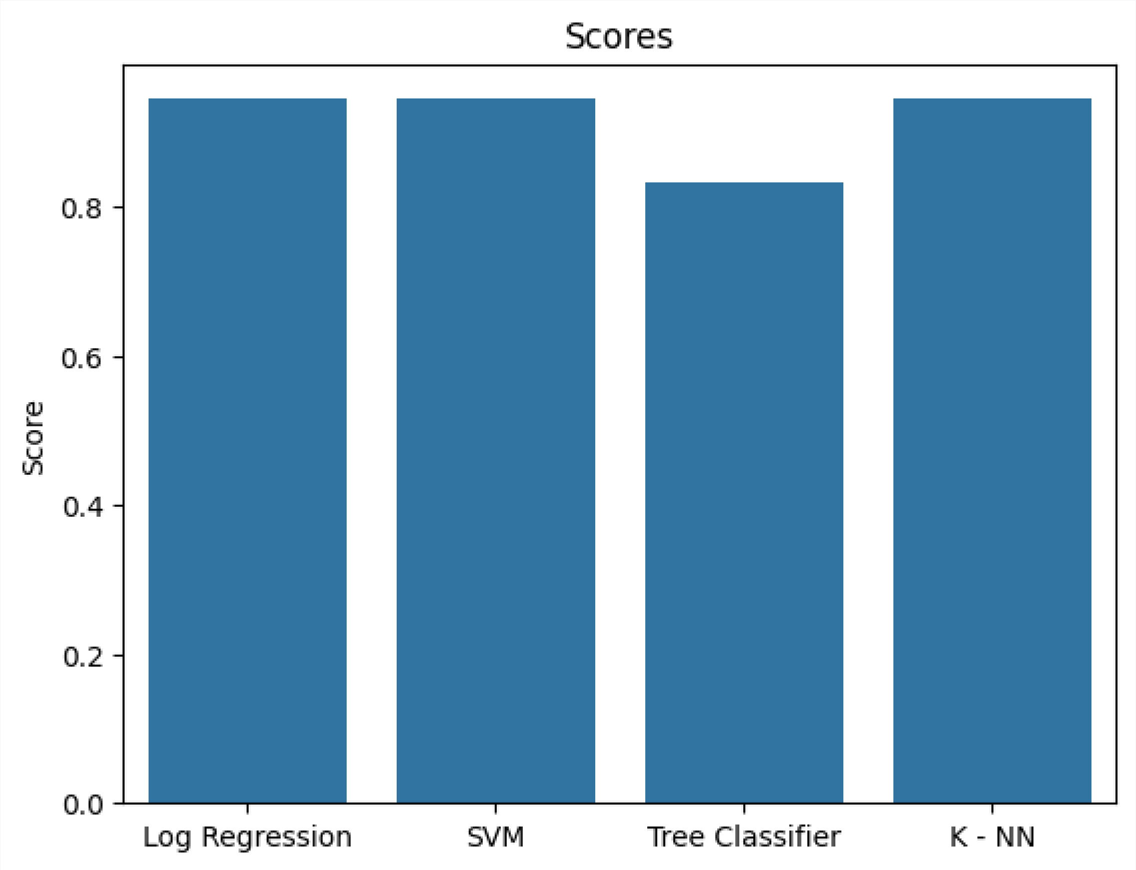
Section 5

Predictive Analysis (Classification)

Classification Accuracy

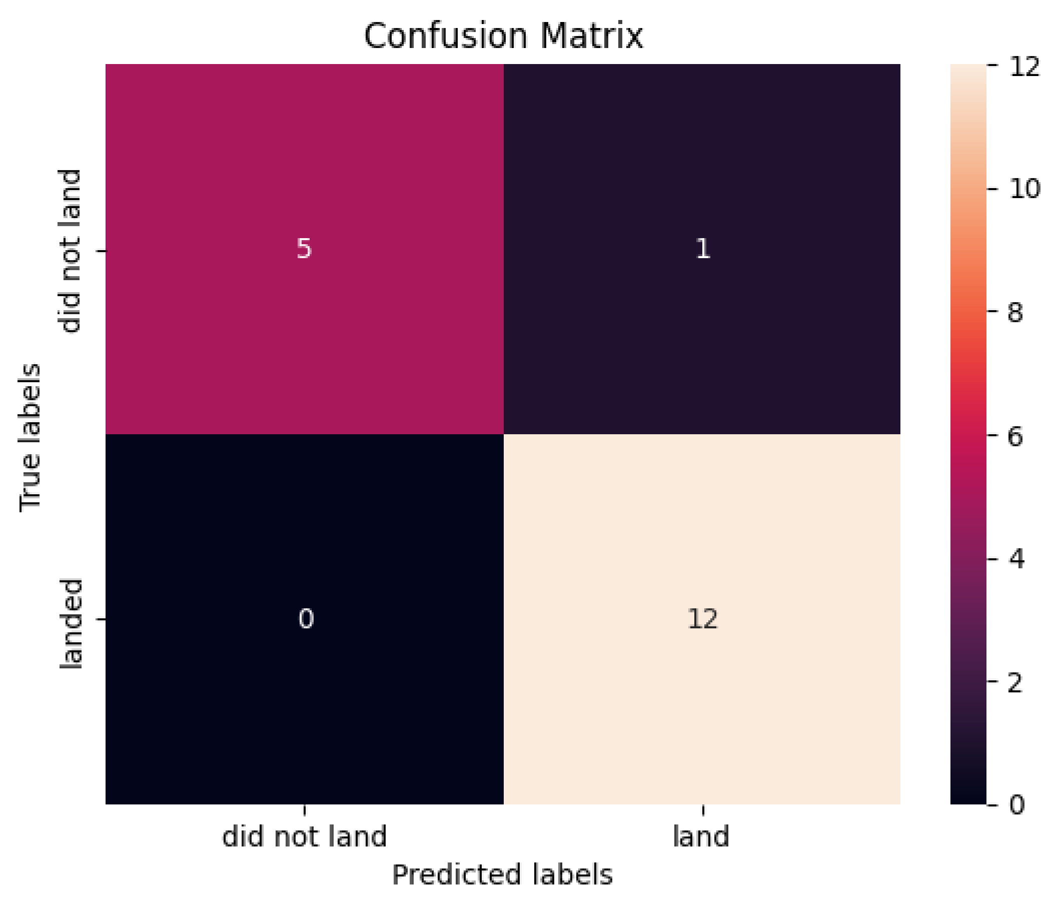
- Visualize the built model accuracy for all built classification models, in a bar chart

- Find which model has the highest classification accuracy: **Log Reg, SVM and KNN**



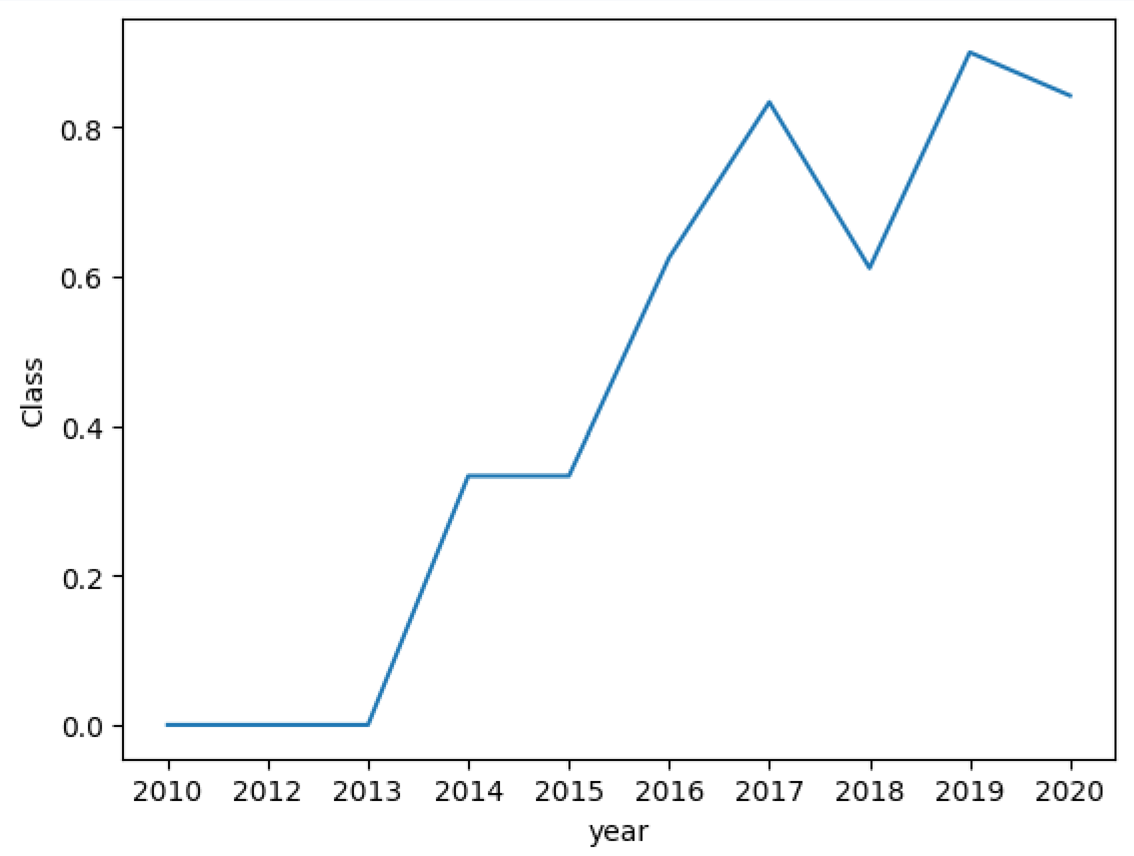
Confusion Matrix

- Show the confusion matrix of the best performing model with an explanation



Conclusions

- The main factor of success in launches is the experience gain in the firsts attempts, we can see that in slide 24:



Thank you!