IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jiakun Pan

Oct 3, 2021

Outline

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

Executive Summary

- We used Python to collect and process data, applied SQL for quantitative insights, visualized the geographical data to understand the proximities, and conducted machine learning to train a predictive model for rocket landing outcomes.
- As a result, we now have a better idea about what really matters and what does not. Furthermore, we have opened a door of forecasting the future landing outcomes.

Introduction

- In this project, we collect data about SpaceX to study the Falcon 9 rockets.

- We hope to find out the trends from the data, and specifically, we wonder
  - How do the factors (payload mass, booster, orbit, etc.) affect the landing outcome?
  - What is the trend of the success rate in the long run?
  - Can we use a machine learning model to predict the outcome of an individual landing?
  - Etc.

Section 1

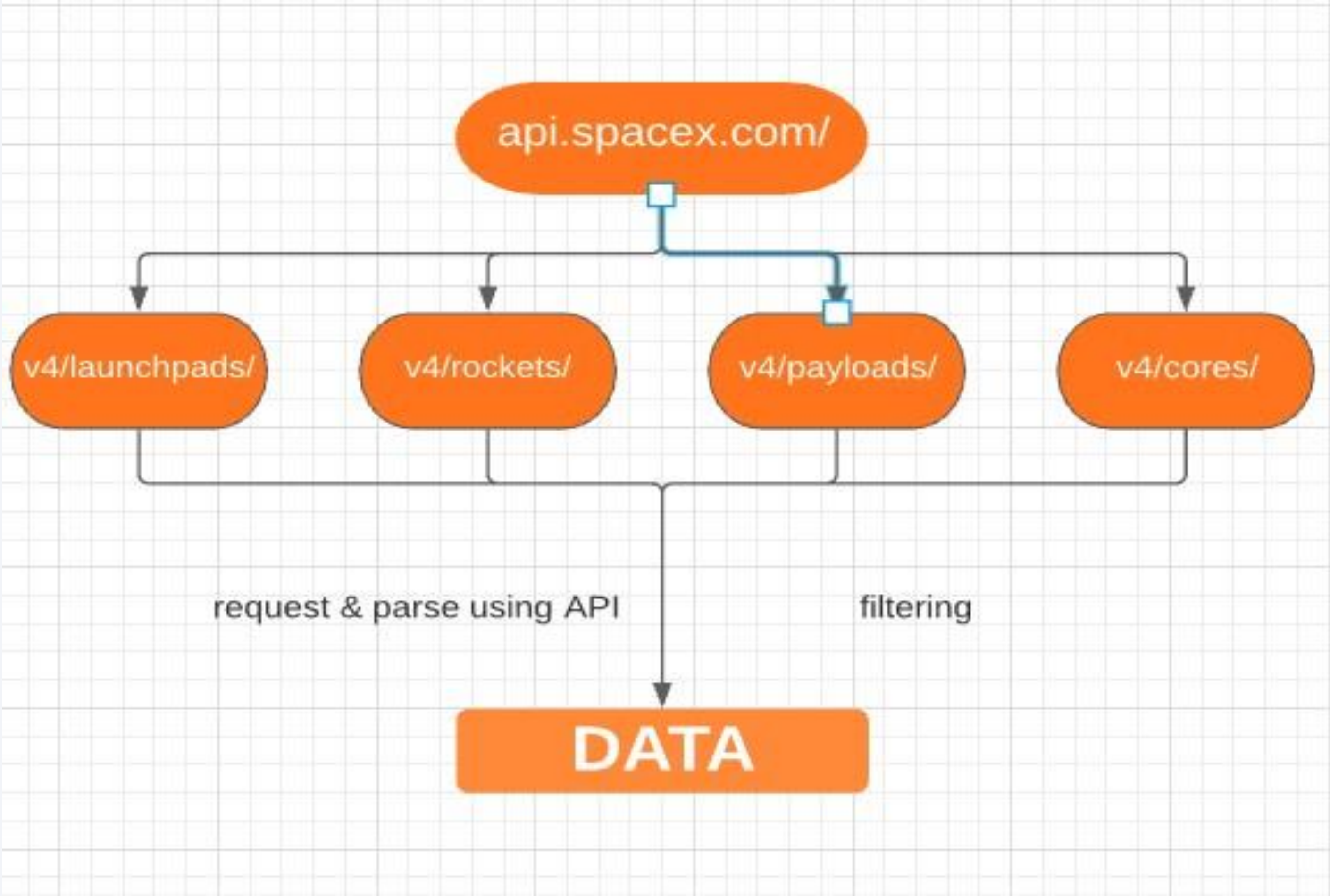
Methodology

Executive Summary

- Data collection methodology:
  - We collected records of 90 rockets launches from *api.spacexdata.com* with web scraping.
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection – API

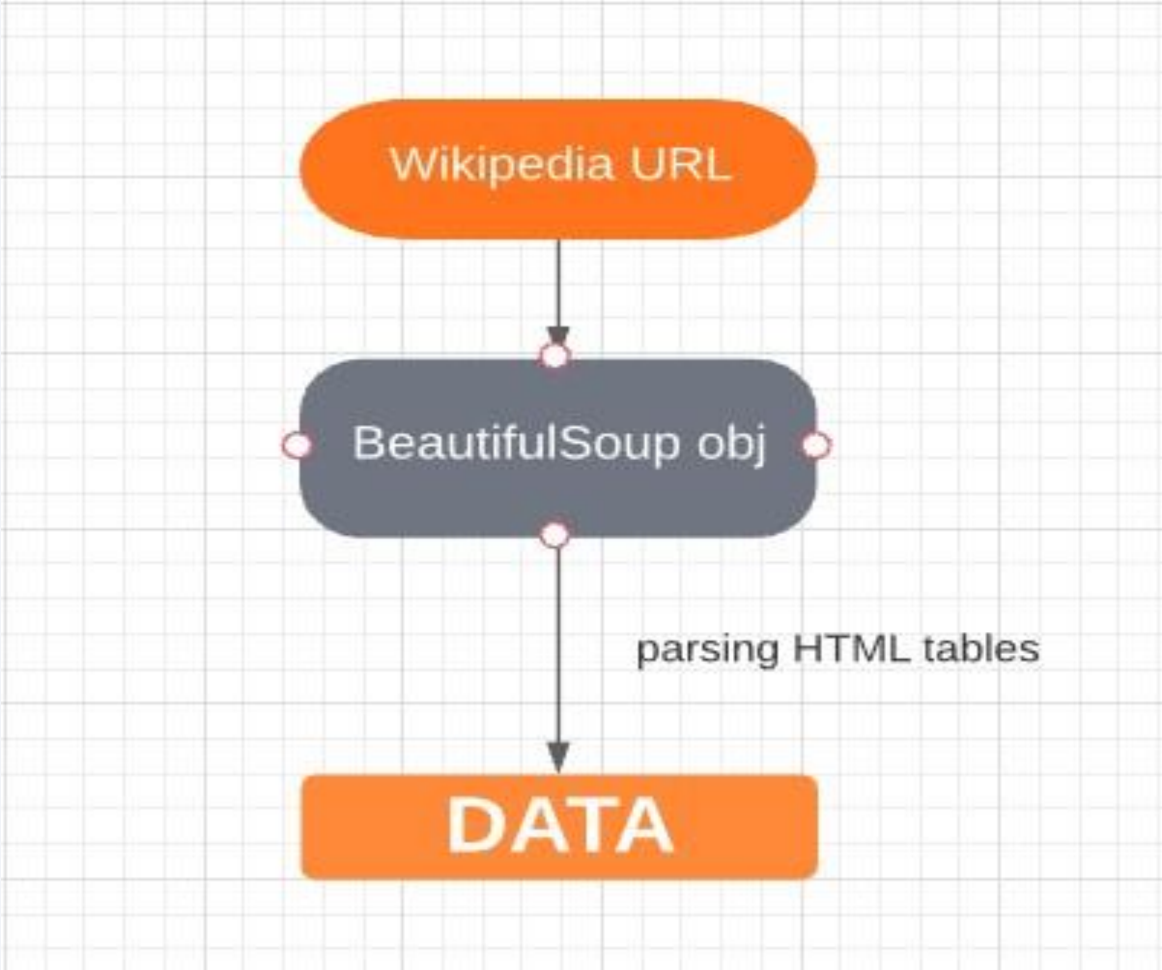
- Data of different aspects are restored in different datasets and formats.
- For each webpage, we performed web scraping using specifically defined REST API function.
- We later filtered the data so that it only has records for the Falcon 9 rockets.



Data Collection – Web Scraping

- Another part of our data come from Falcon 9 and Falcon Heavy Launches Records from Wikipedia.

- To this end we used web scraping to obtain the HTML resource code from Wikipedia and extract its tables to form dataframes.



Data Wrangling

- We first examined the types for the missing data.

- Since all missing data are numerical and regarding two attributes of the rocket launches, we replace each NaN with its attribute mean.

- Next, we converted the categorical data to the 0/1 binary.

EDA with Data Visualization

1. We scatter-plotted the relationship between Payload and Launch Site.
2. We bar-plotted the success rate of each Orbit Type.
3. We scatter-plotted the relationship between Payload and Orbit Type.
4. We visualized the launch success yearly trend with a line chart.

EDA with SQL

- Using SQL queries, we made the following observations from the data:
  - Distinct names of launch sites;
  - The total payload mass carried by boosters launched by NASA (CRS);
  - Dates when the first successful landing outcome in ground pad was achieved;
  - Names of the booster_versions which have carried the maximum payload mass;
  - Total number of successful and failure mission outcomes;
  - Rank the count of landing outcomes under certain restrictions;
  - Etc.

Build an Interactive Map with Folium

- We marked all launch sites on a map with circles and pop-up markers.
- For each launch record, we added details to the corresponding marker.
- With the Mouse Position function, we calculated the distance between each launch site and its proximities (coastline, railway, highway, etc.).

Build a Dashboard with Plotly Dash

- We used a dashboard to compare the success rates between different Launch Sites, Payload Ranges, and F9 Booster Versions.

Predictive Analysis (Classification)

- We adopted four (linear regression, SVM, decision tree, and k-neighbors) models for the prediction, and tune hyperparameters with the sklearn package GridSearchCV.
- We evaluated the accuracy and obtained the confision matrices for each model.

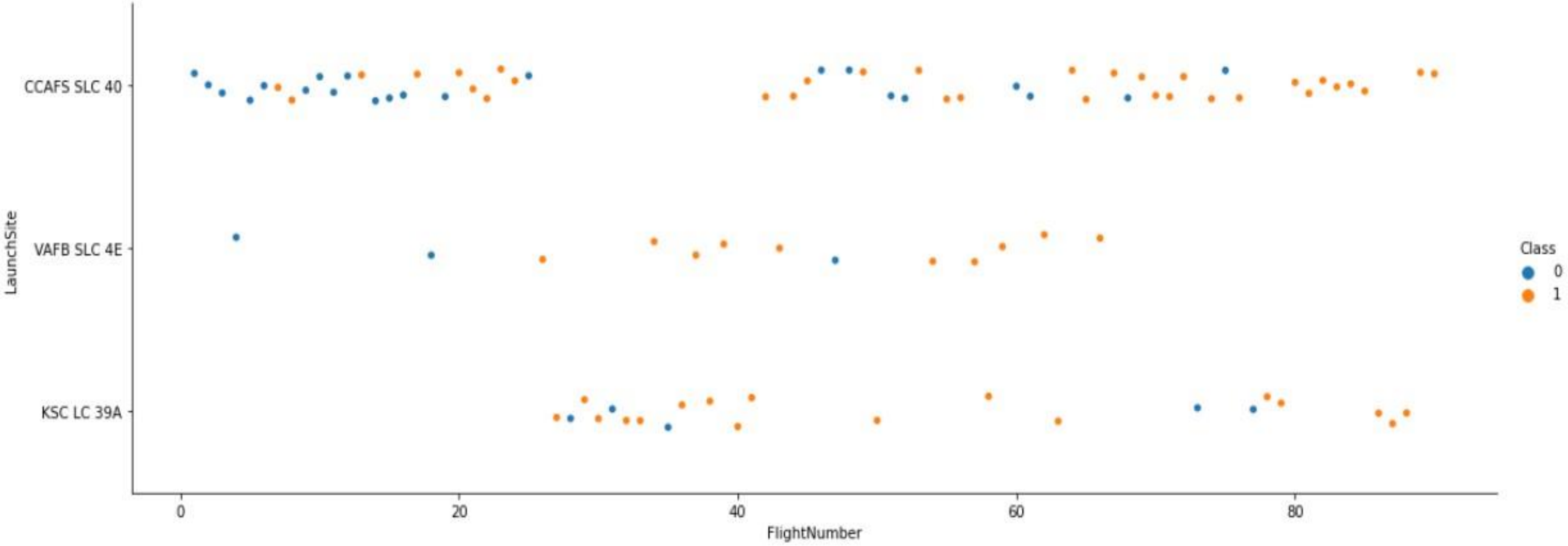
Results

- Exploratory data analysis results
  - Payload Mass affects success rate positively in all launch sites.
  - Success rate distributes highly unevenly among launch sites.
  - Heavy Payload affects success rate significantly on different orbits.
  - The success rate keeps increasing since 2013 in general.
- Predictive analysis results
  - With 72 rows of data in the training set, our four predictive models have similar accuracy for the rest 18 rows in the testing set.

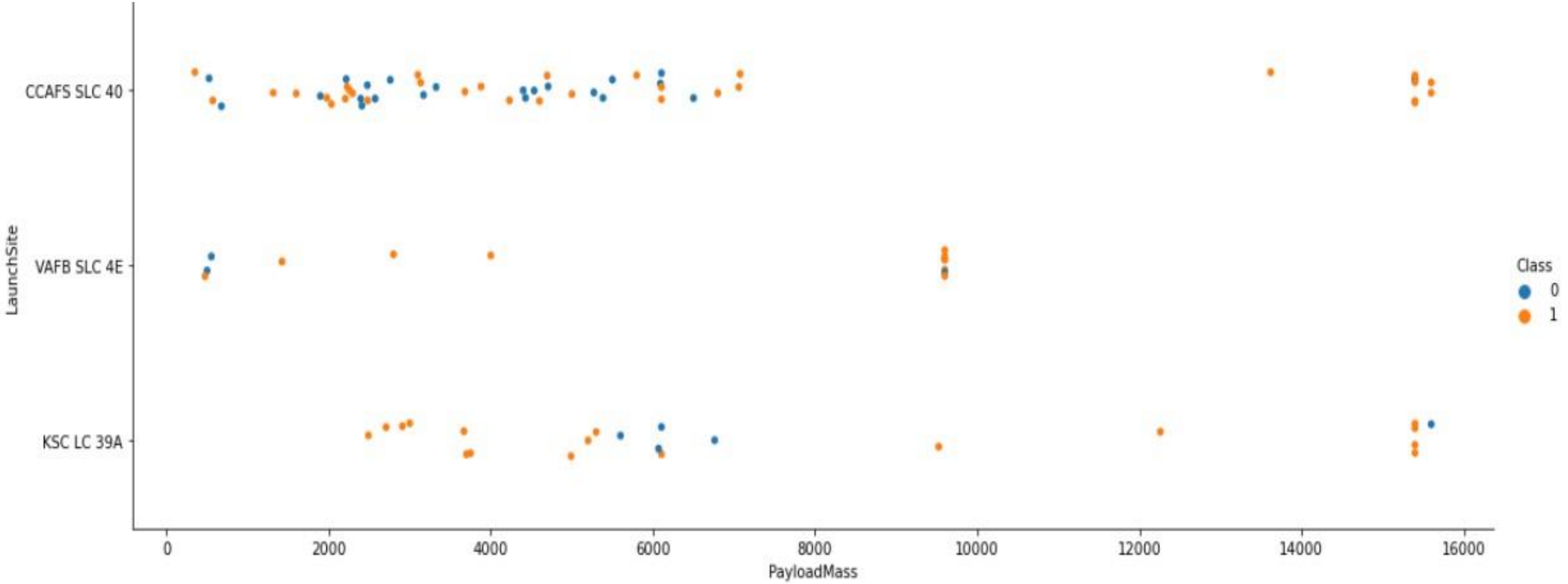
Section 2

Insights drawn from EDA

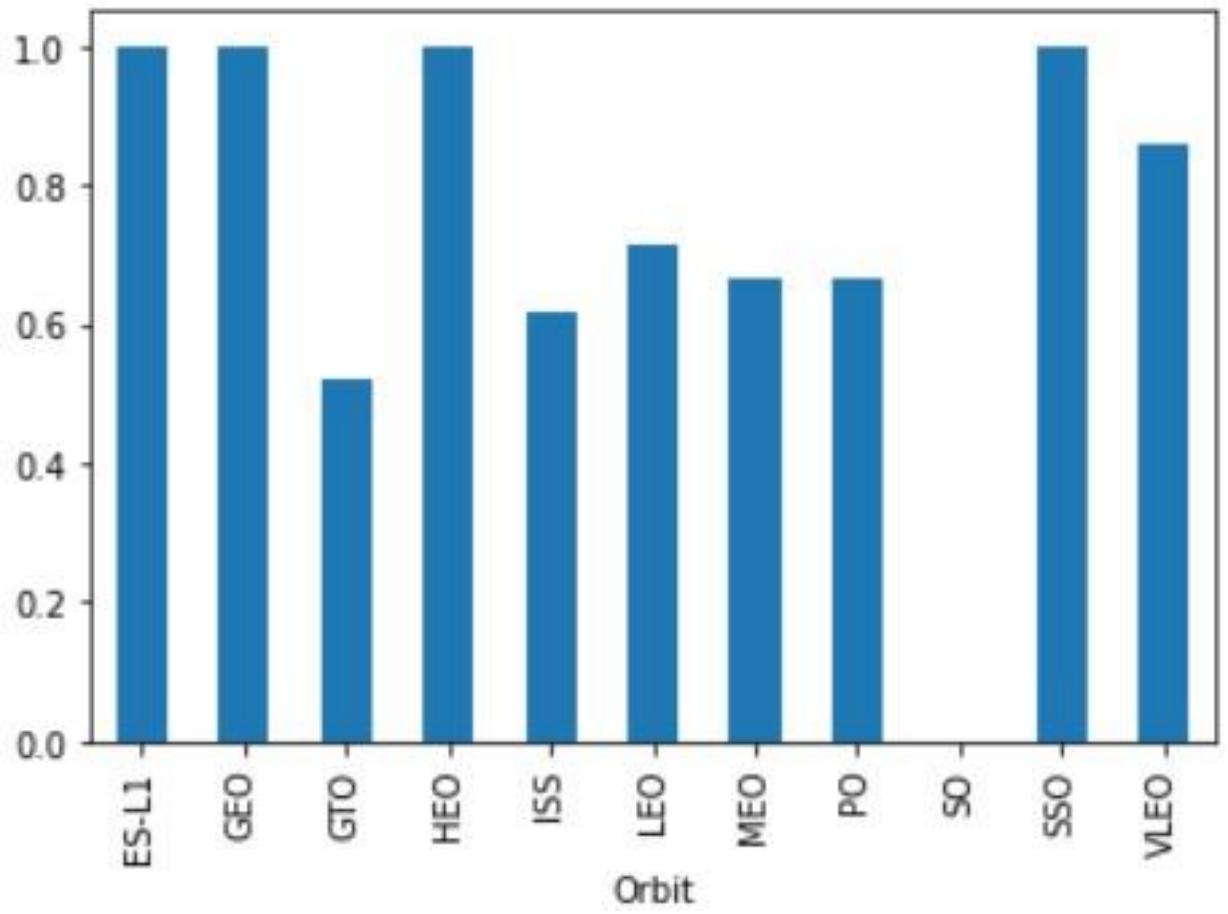
Flight Number vs. Launch Site



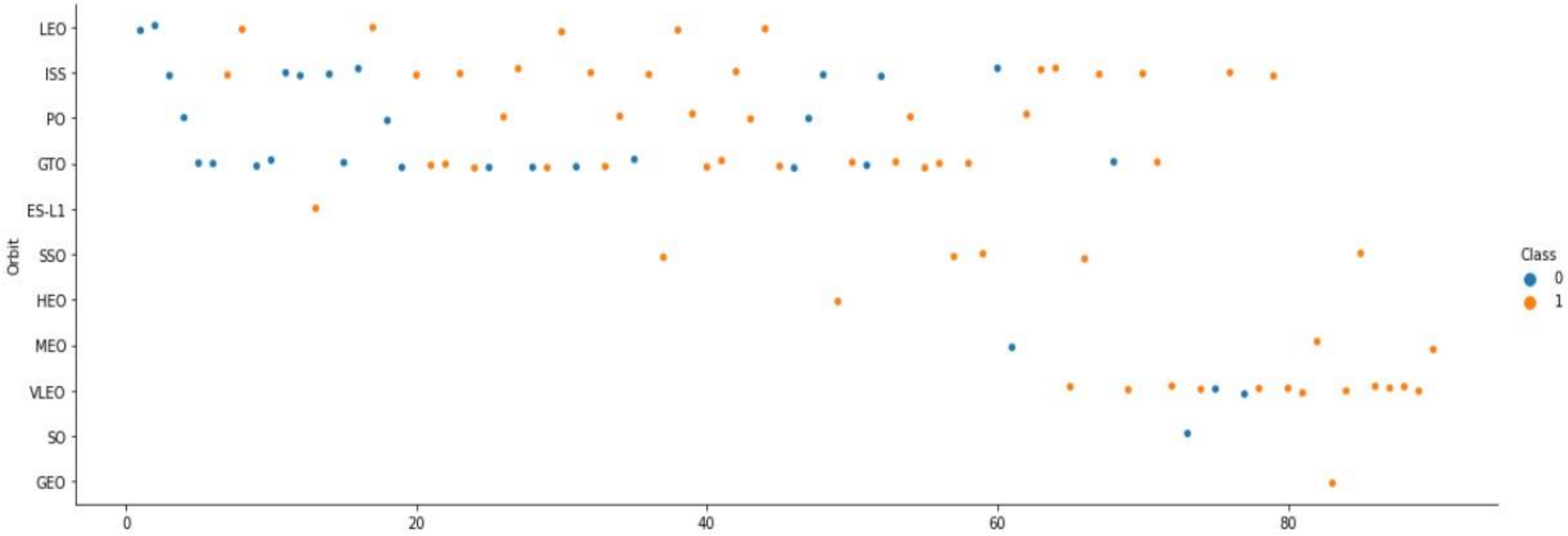
Payload vs. Launch Site



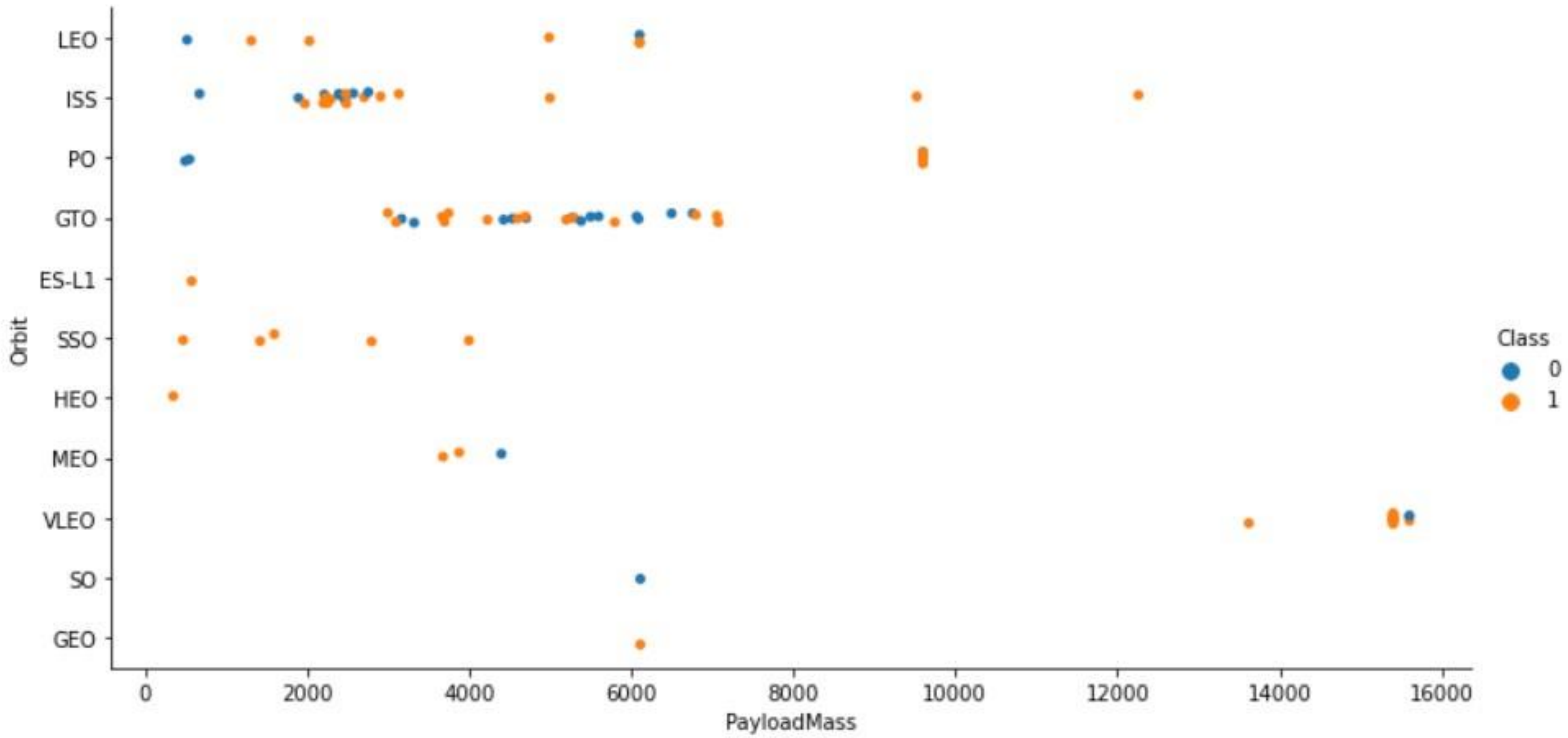
Success Rate vs. Orbit Type



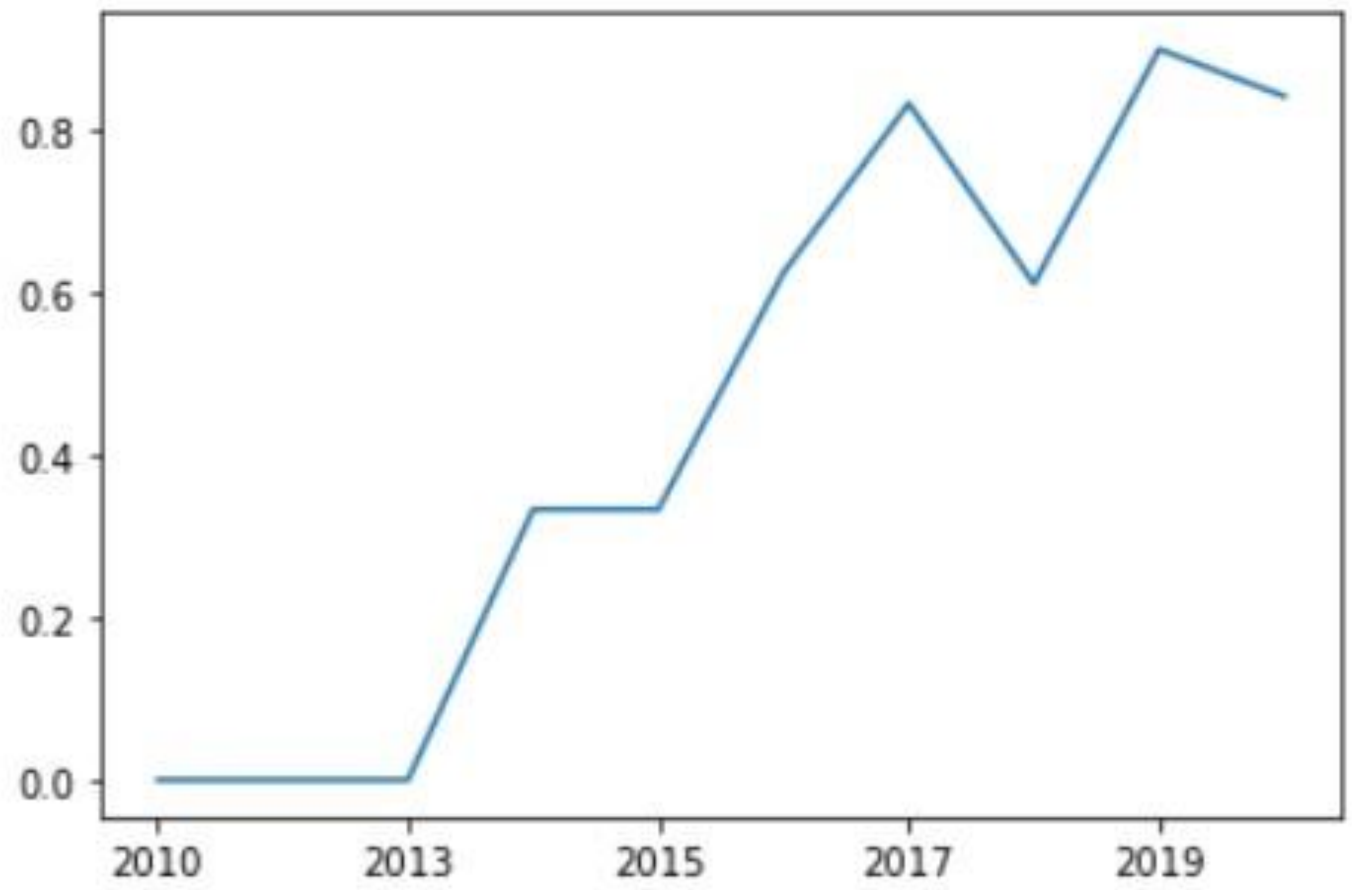
Flight Number vs. Orbit Type



Payload vs. Orbit Type



Launch Success Yearly Trend



All Launch Site Names

```
select distinct Launch_Site from spacex
```

Run time: **0.011 s**

Result set 1

Find

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

Launch Site Names Begin with 'CCA'

select * from spacex where Launch_Site lik... Run time: **0.010 s**

Result set 1

Find

DATE	TIME__UTC_	BOOSTER_VERSION	LAUNCH_SITE
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40

Total Payload Mass

select sum(PAYLOAD_MASS__KG_) from s...

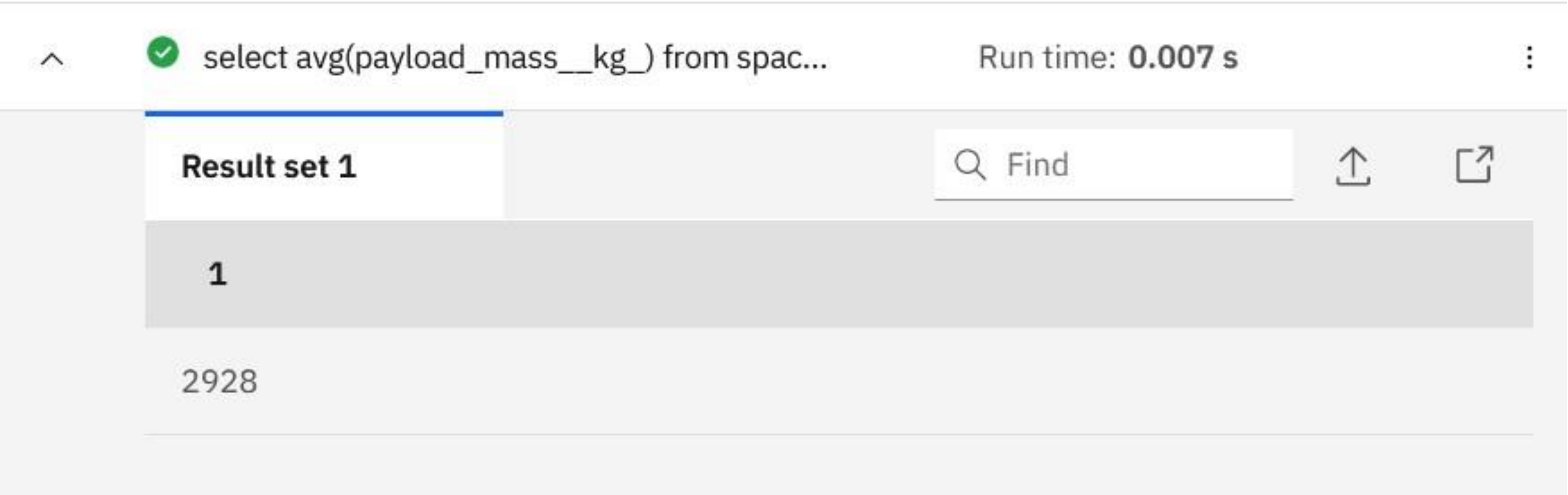
Run time: **0.008 s**

Result set 1

Find

1
45596

Average Payload Mass by F9 v1.1



First Successful Ground Landing Date

select date from spacex where landing__o...

Run time: **0.007 s**

Result set 1

Find

DATE
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

select booster_version from spacex where... Run time: **0.008 s**

Result set 1

Find

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

Total Number of Successful and Failure Mission Outcomes

select count(*) as sux from spacex where l... Run time: **0.006 s**

Result set 1

SUX
61

select count(*) as fails from spacex where ... Run time: **0.006 s**

Result set 1

FAILS
10

Boosters Carried Maximum Payload (use subquery)

Result set 1

Find

BOOSTER_VERSION
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7

Result set is truncated, only the first 12 rows have been loaded. Select "View all loaded data" on the right top of the result to view all loaded rows.

More

Failed Drone Ship Landings in 2015

select landing__outcome, booster_version,... Run time: **0.008 s**

Result set 1

Find

LANDING__OUTCOME	BOOSTER_VERSION	LAUNCH_SITE
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

select landing__outcome, count(*) as outc... Run time: **0.011 s**

Result set 1

Find

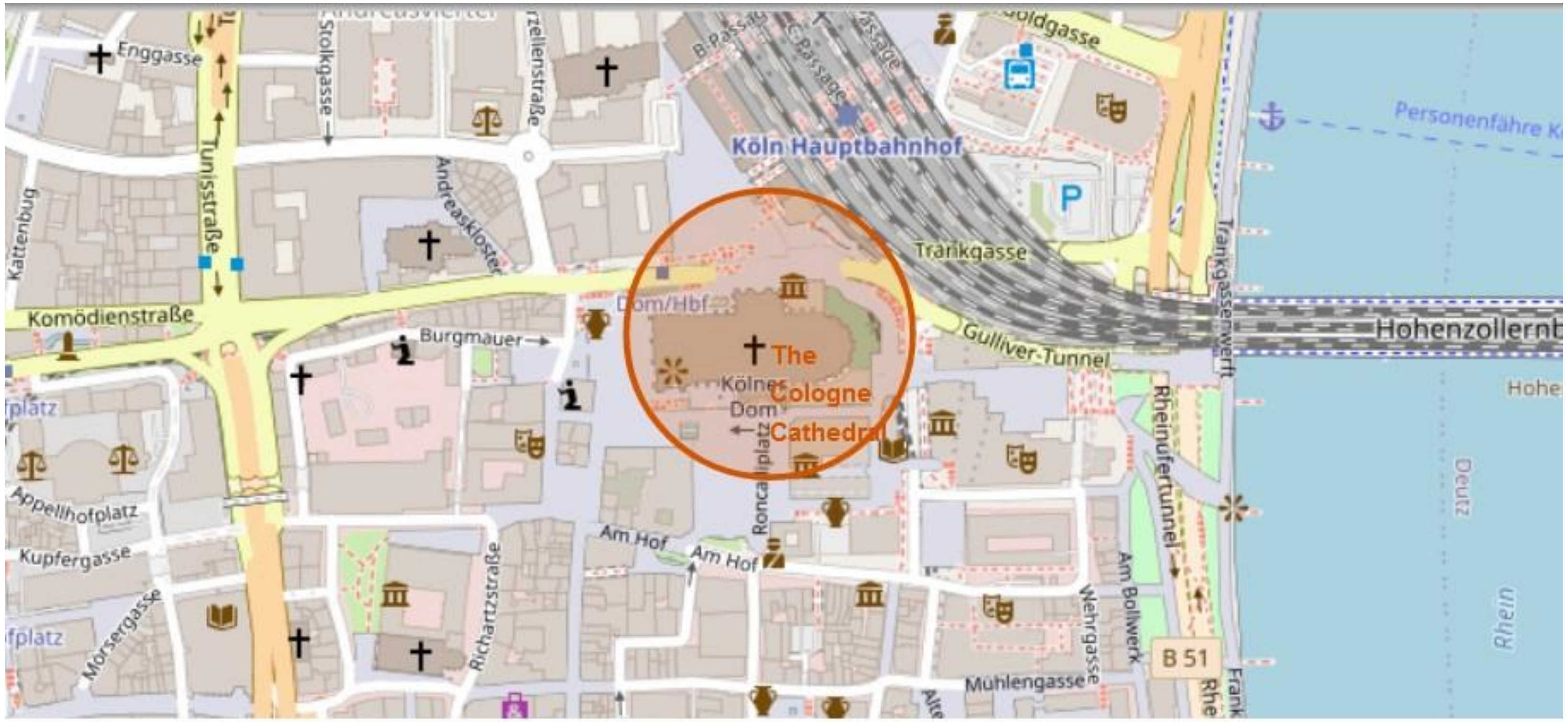
LANDING__OUTCOME	OUTCOME_COUNT
Controlled (ocean)	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	10
Precluded (drone ship)	1

Result set is truncated, only the first 8 rows have been loaded. Select "View all loaded data" on the right top of the result to view all loaded rows. More

Section 4

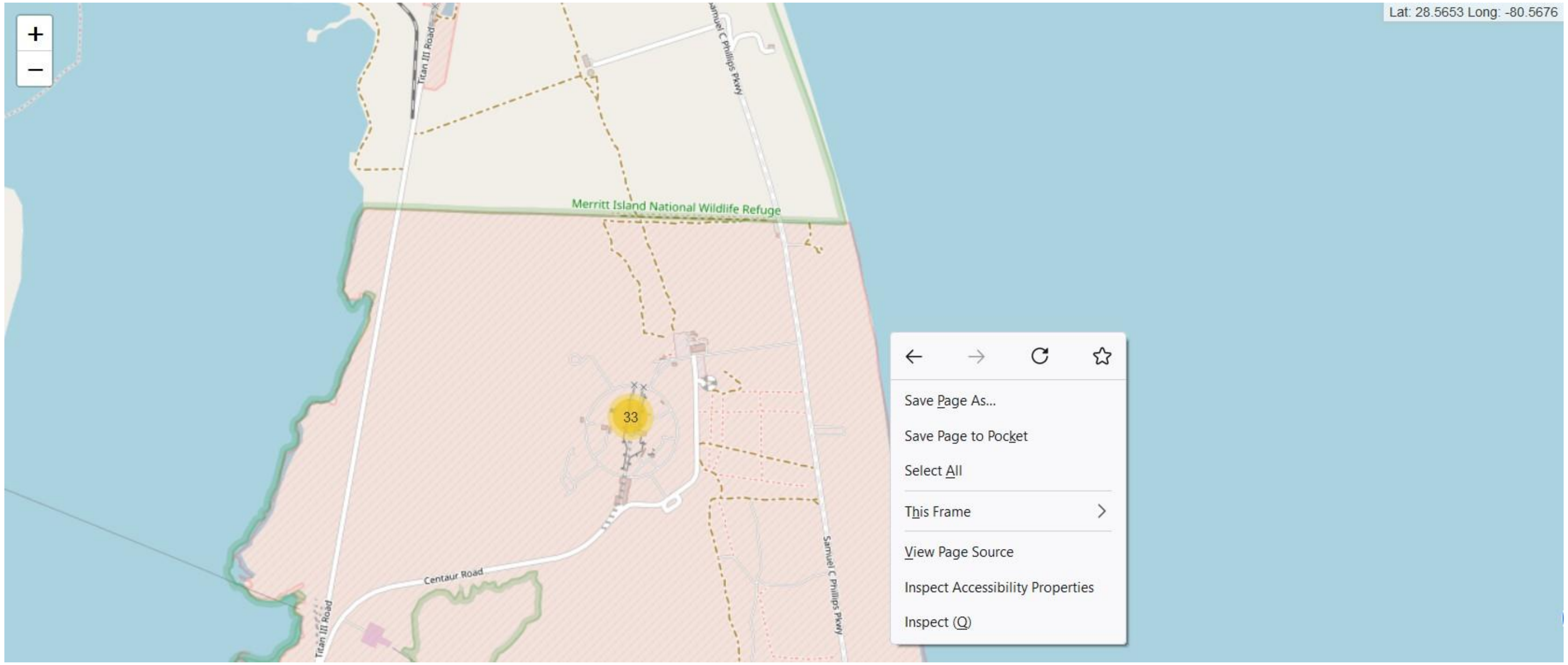
Launch Sites Proximities Analysis

Circle and Mark a place near me



Mark launches with pop-ups (green=sux, red=fail)

Locate mouth position to calculate coastline dist

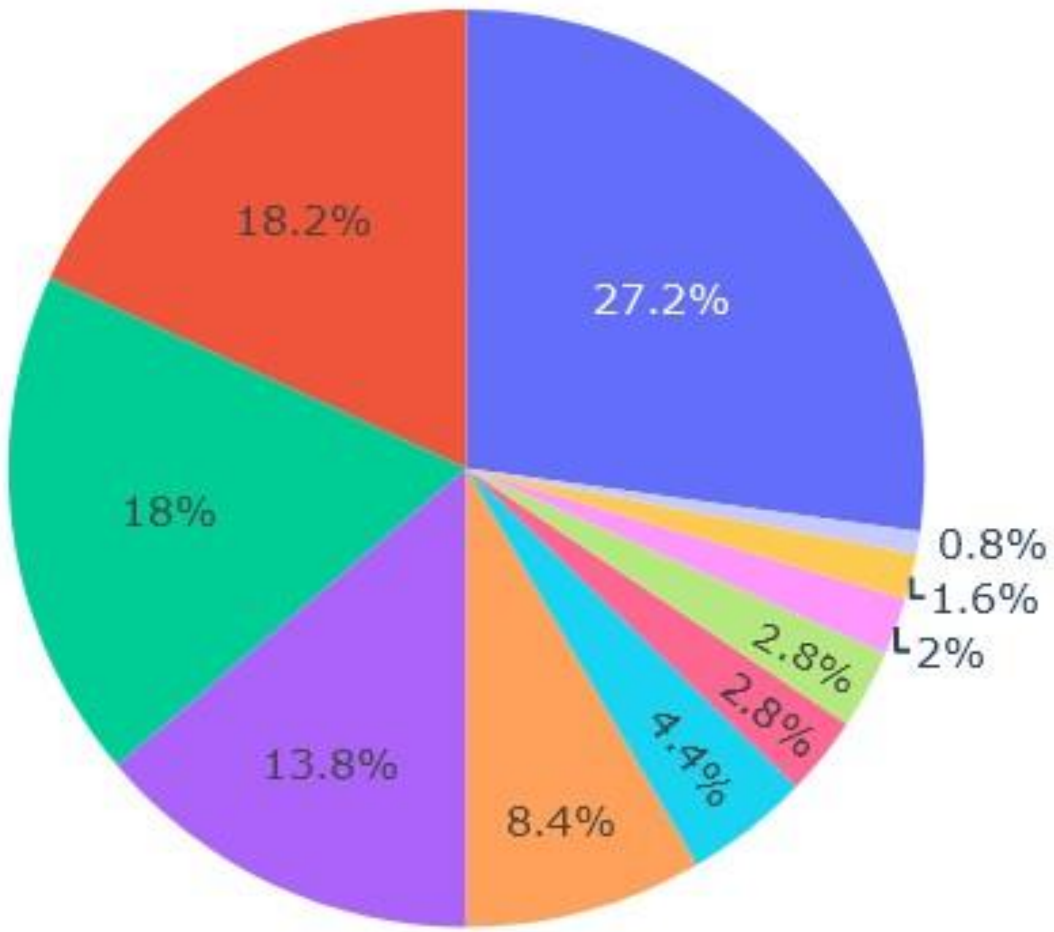


Section 5

Build a Dashboard with Plotly Dash

…n of successful landings by launch sites.

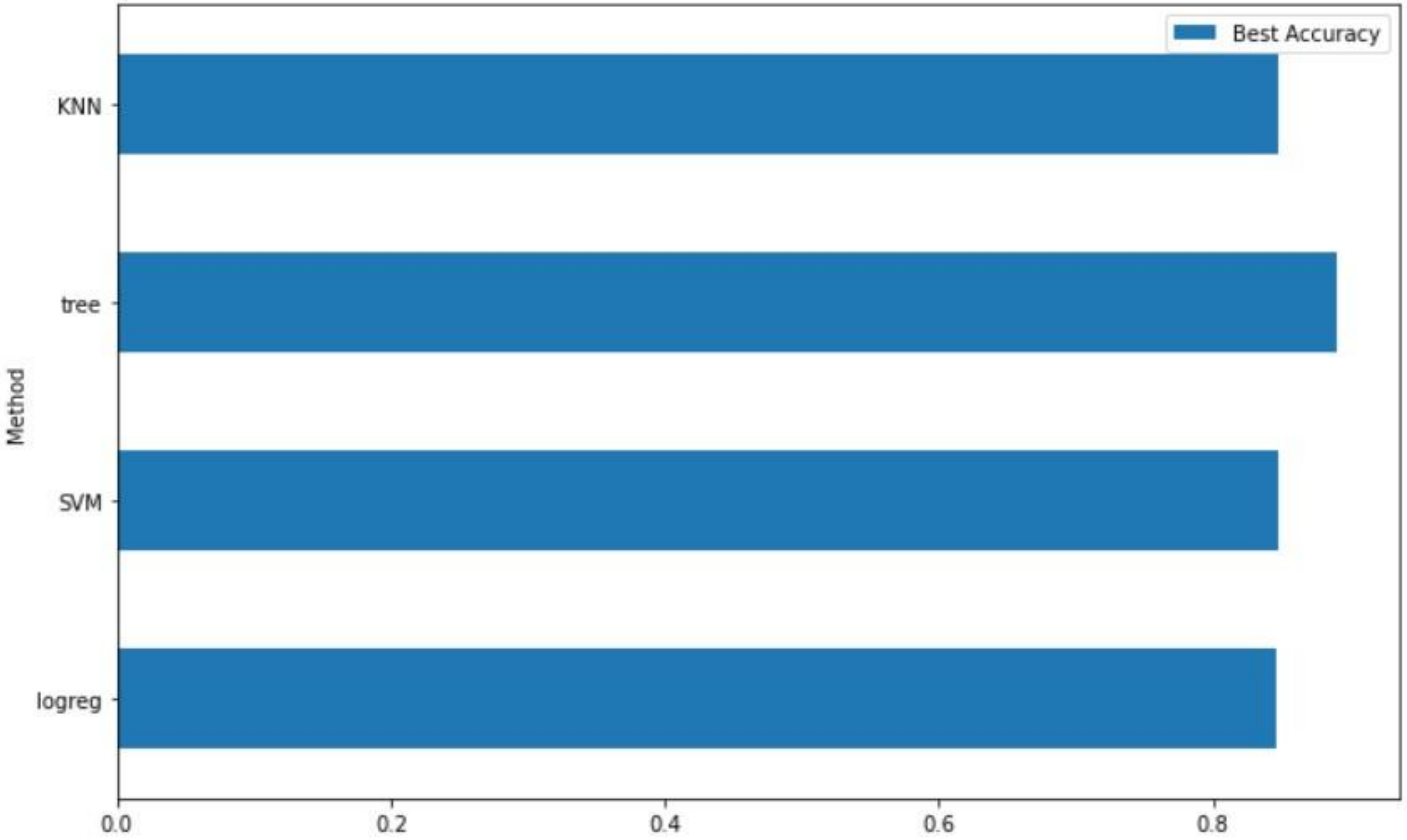
Succesful landings by site



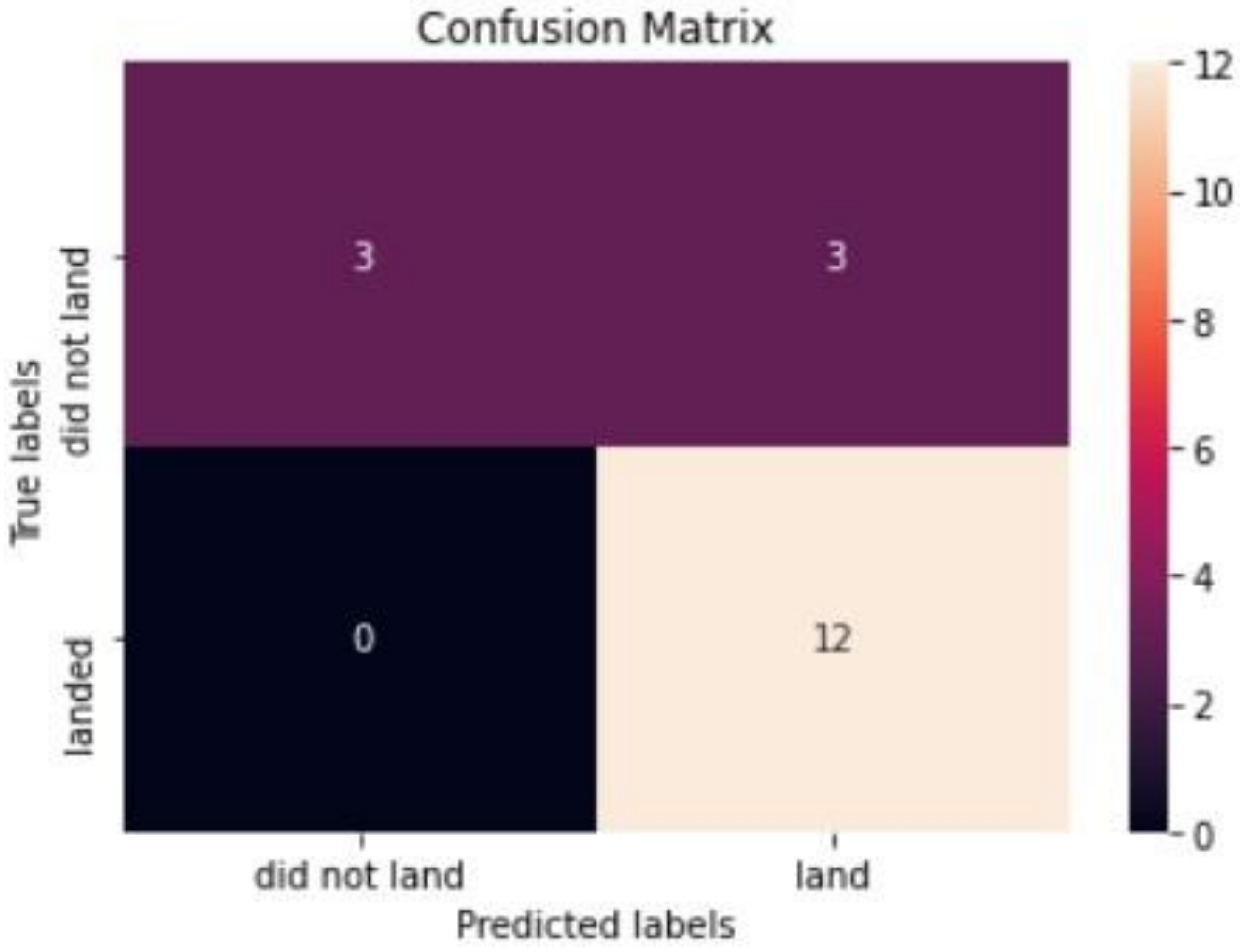
Section 6

Predictive Analysis (Classification)

Classification Accuracy



Confusion Matrix



Conclusions

- Exploratory data analysis results
  - Payload Mass and Orbit affects outcome positively.
  - Heavy Payload affects success rate significantly on different orbits.
  - The success rate keeps increasing since 2013 in general.
- Proximity analysis results
  - Proximity of coastline, railway, or highway has no effect on outcomes!
- Predictive analysis results
  - The four predictive models have similar performance and confusion matrices, and it is too early to decide which model is the best.

Appendix

- The machine learning model is a basic one, and we can expect better prediction accuracy with a deeper neuron network and more data in the training set.
- In a financial angle, we wonder the effect of each landing outcome on the stock market. If we have a good predictive model, then NASDAQ can be more interesting.

Thank you!