IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

July 19, 2022

Outline

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

Executive Summary

- Predictions about whether or not SpaceX rockets' first stages can be reused is something that can be predicted using machine learning techniques.
- Using classification techniques such as Support Vector Machines & K Nearest Neighbors, we can accurately predict if the first stage of a SpaceX rocket will successfully land and be able to be reused.
- We have achieved an accuracy of prediction of 83.3%.

Introduction

- The goal is to determine if Falcon 9 first stage will land successfully.
- Falcon 9 rocket launches by SpaceX are at a cost 62 million dollars (vs 165 million dollars for its competitors). This is because savings made possible by reusing the first stage.
- By deciding on and feeding important parameters into a machine learning algorithm, we can determine if the first stage can land successfully and be reusable.
- This is information that can be useful for the company to predict rocket success, as well as if an alternate company wishes to bid against SpaceX for a rocket launch.

Section 1

Methodology

Methodology

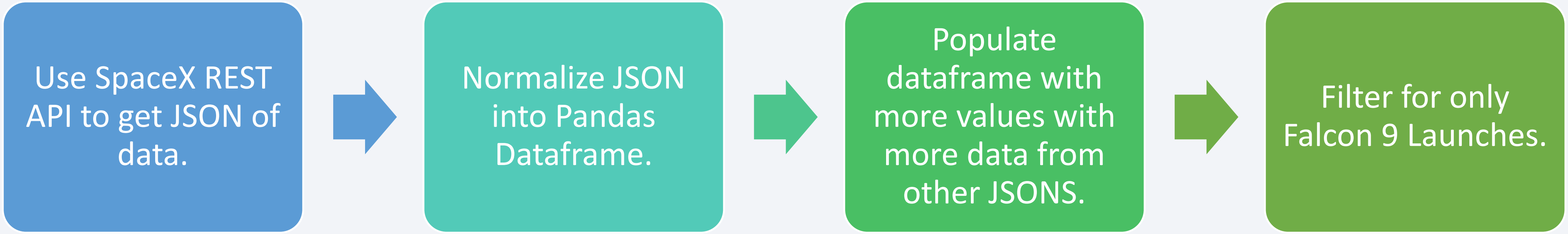
Executive Summary

- Data collection methodology:
    - Data was acquired with the SpaceX REST API (official data about launches) as well as web scraping of Wikipedia pages for more useful information.
- Perform data wrangling
    - Data was processed by checking the larger trends and success rates of various launches and parameter combinations.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Build, tune, and evaluate classification models such as SVM and Decision Trees

Data Collection

- Data was collected through two main ways
- SpaceX REST API provides access to data about launches, such as the type of rocket, payload delivered, launch and landing specifications, and outcome of landing.
- Web scraping with BeautifulSoup was done on Wikipedia pages about SpaceX Falcon 9 launches to gather additional information.

Data Collection – SpaceX API



Data Collection - Scraping

Use BeautifulSoup library to web scrape Wikipedia pages.

→

Extract all column/variable names from the HTML table header

→

Create a data frame by parsing the launch HTML tables

Data Wrangling

- The data was processed by analyzing various trends in the data
- I calculated the total number of successes and the success rate of the launch outcomes
- I removed the data points with null values

EDA with Data Visualization

- Charts were plotted between various features to determine if there were any relationships that could affect a prediction of a successful launch outcome
- Examples of charts included scatter plots of features such as payload weight, launch site, orbit type, etc.
- Examples of graphs can be seen later in the slideshow

EDA with SQL

- SQL Queries that were performed:
  - Average payload mass
  - Date of first successful launch
  - Landing outcomes based on certain payloads
  - Landing outcomes based on certain dates
- More examples of SQL queries can be found later in the slideshow

Build an Interactive Map with Folium

- Map objects such as lines, markers, icons were all added to mark the geographical locations and additional information about the launch sites of the rockets
- For example, by marking the locations, we could see that a majority of the launch sites were near the coastlines
- We could also visualize how the proximity to certain important landmarks such as highways would influence the outcome of the launch site’s ability to have successful launches

Build a Dashboard with Plotly Dash

- I built scatterplots and pie charts that summarized data for all types of launches, as well as launch site specific
- This provided an easy to use tool to see the success rate of the launch sites individually and provide a greater understanding of the effect of each location of the launch site on the final outcome of the rocket launch itself

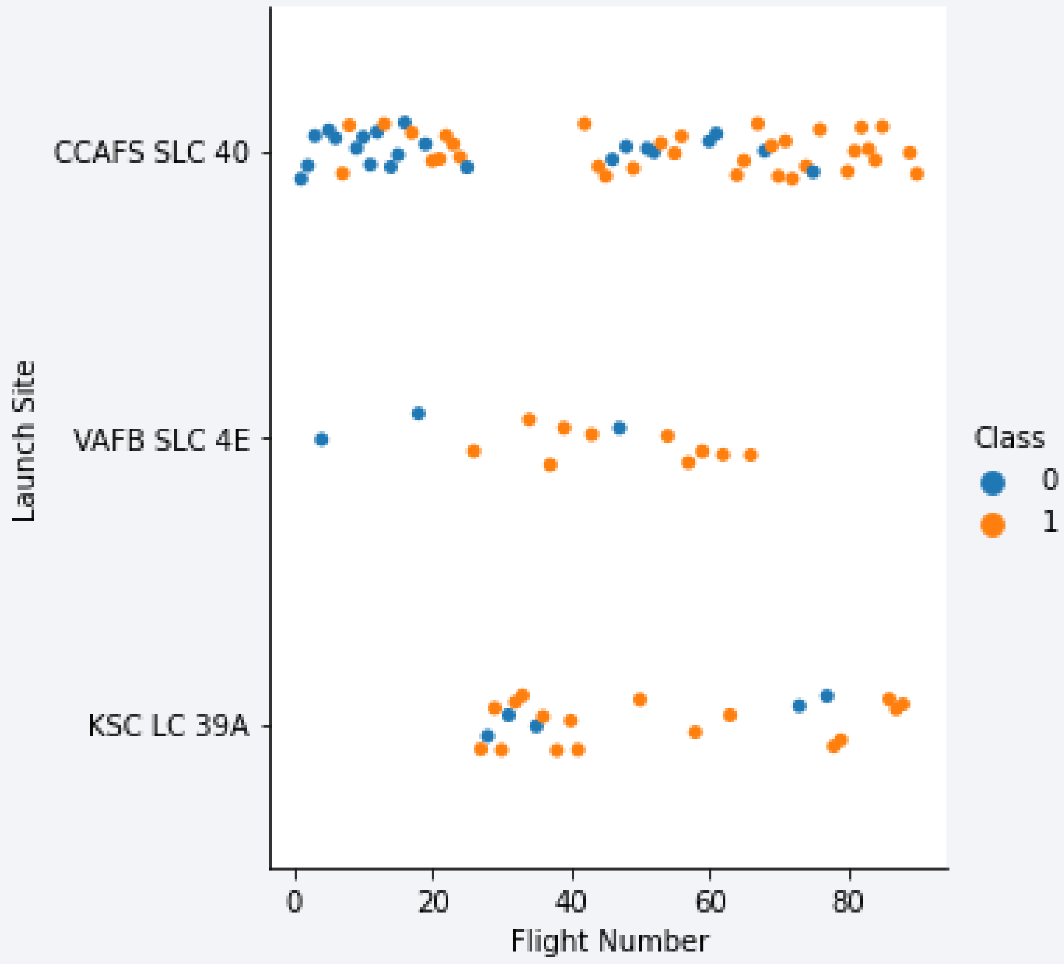
Predictive Analysis (Classification)

- I built, evaluated, improved, and found the best performing classification model to be either SVM, Logistic Regression, or K Nearest Neighbor
- I used the sklearn library to build the models and used Grid Search to effectively train the models, after splitting the training and testing data
- I used important metrics such as the Jaccard score to determine the accuracy of each classification model

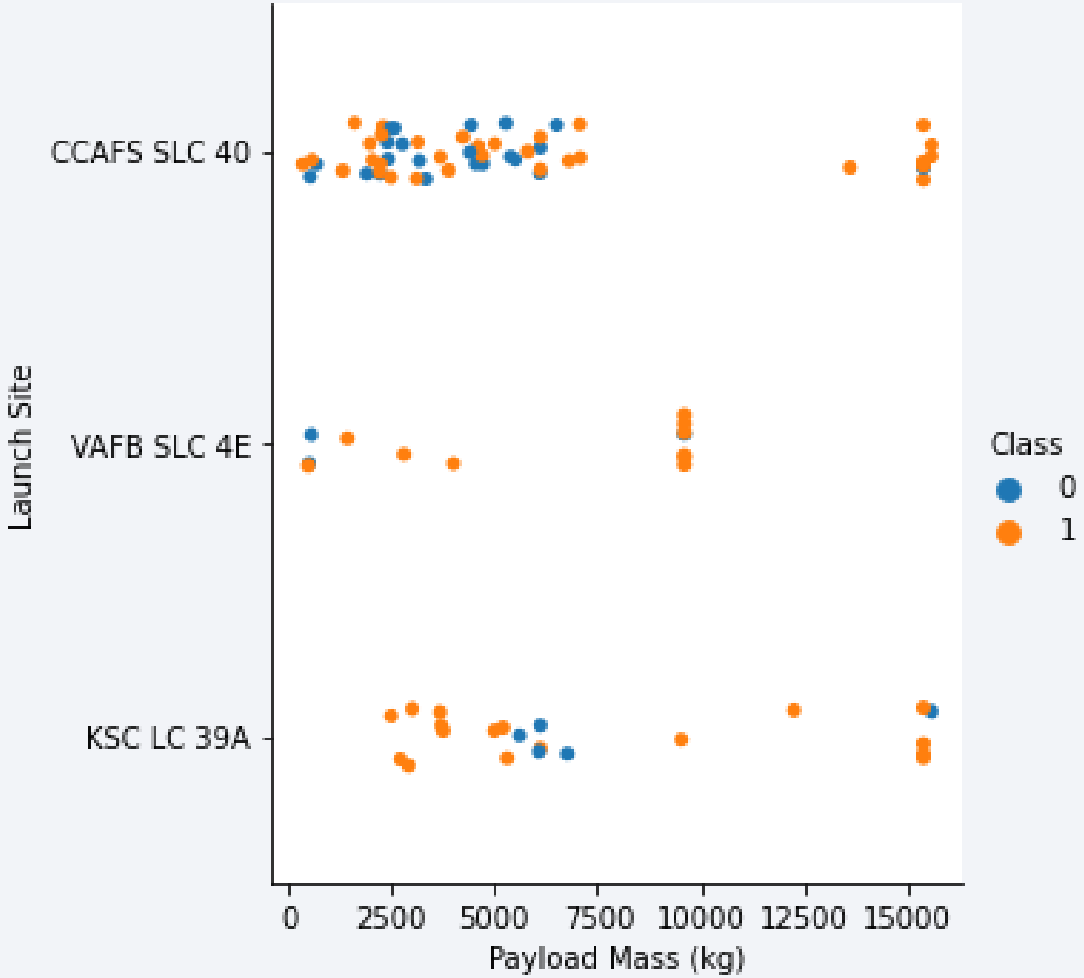
Section 2

Insights drawn from EDA

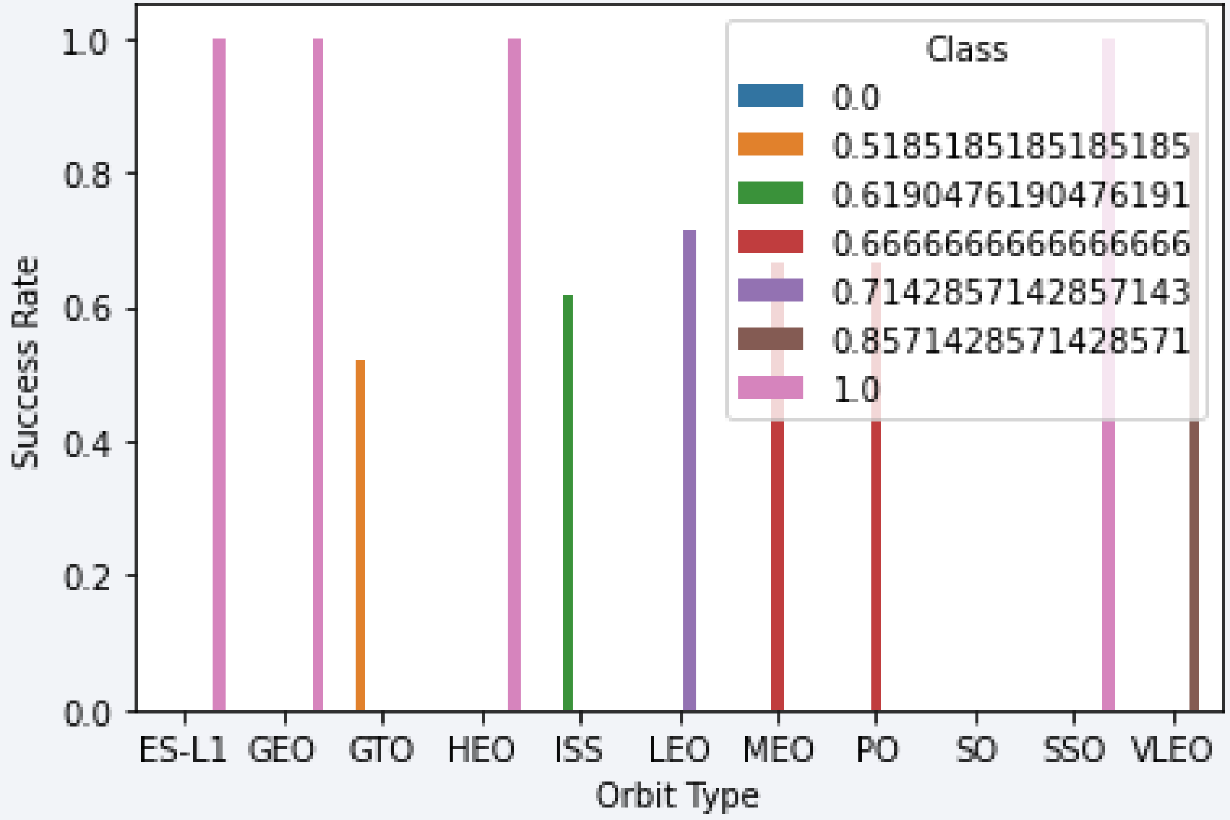
Flight Number vs. Launch Site



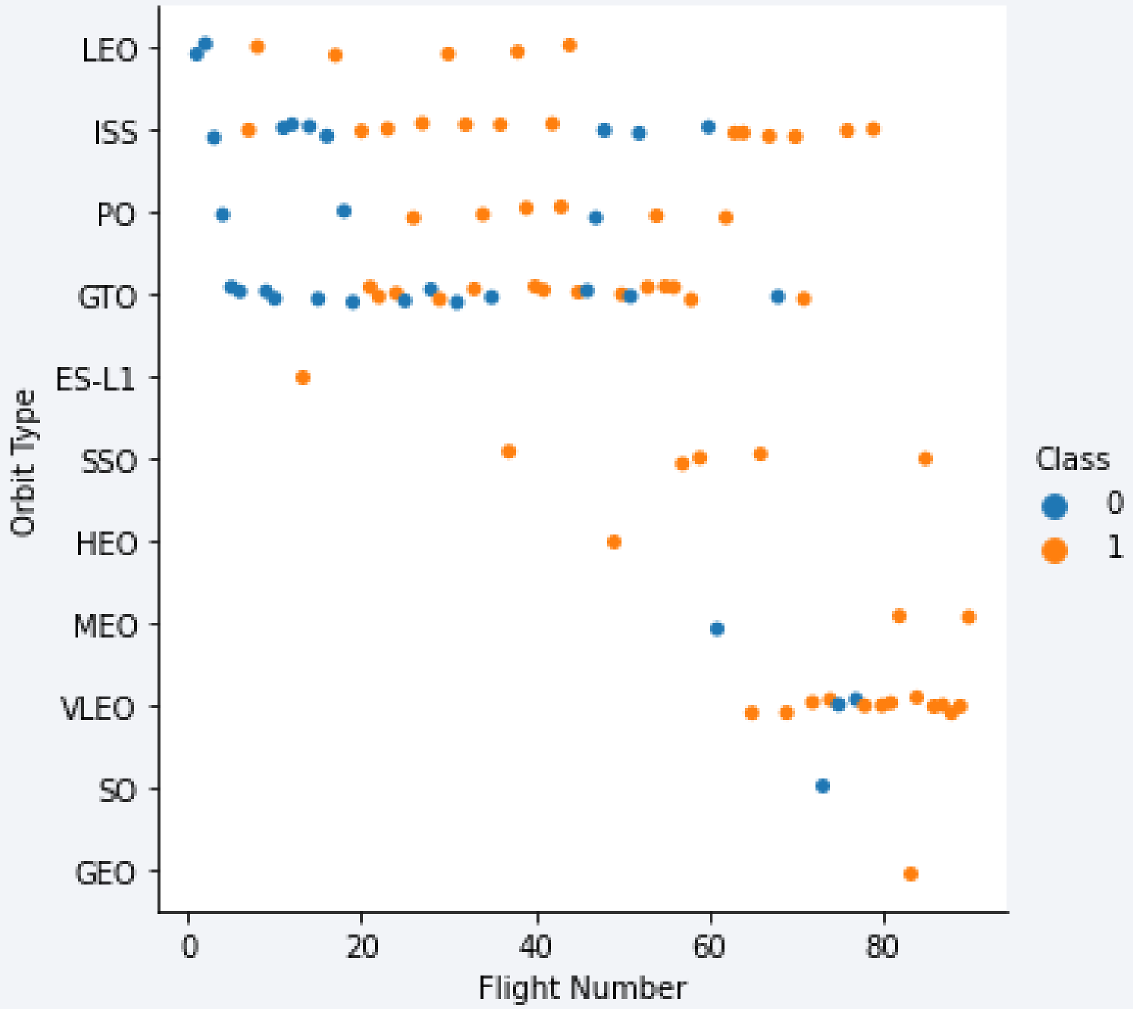
Payload vs. Launch Site



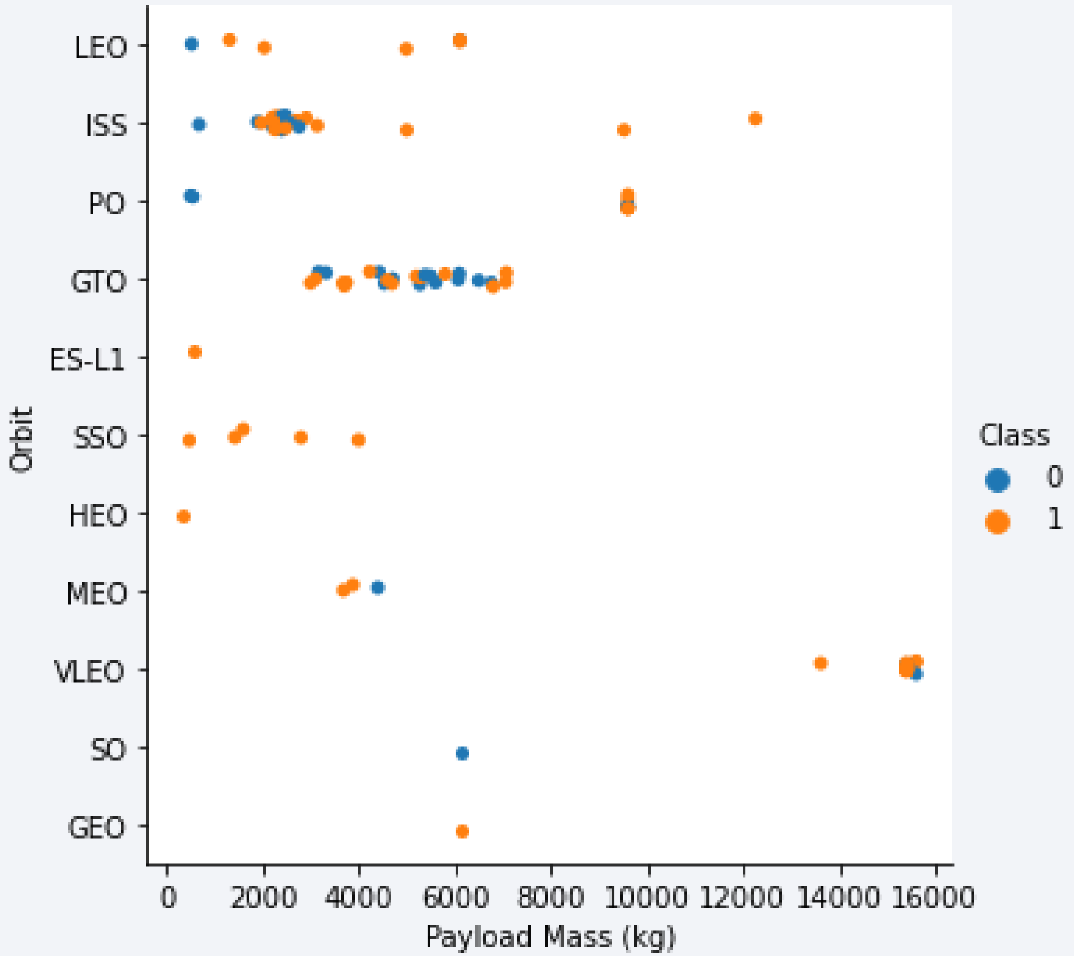
Success Rate vs. Orbit Type



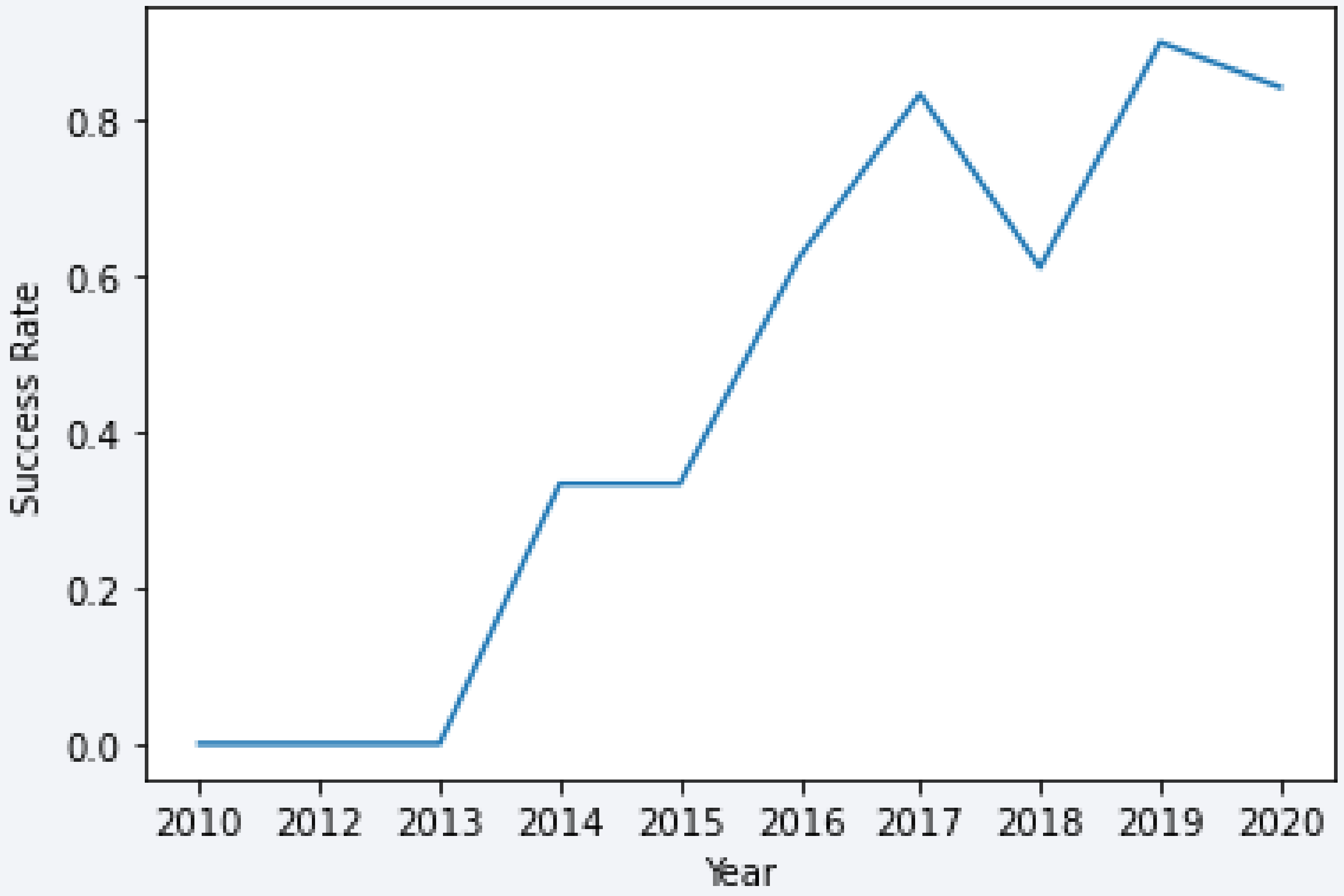
Flight Number vs. Orbit Type



Payload vs. Orbit Type



Launch Success Yearly Trend



All Launch Site Names

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

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEXTBL WHERE "Launch_Site" LIKE "CCA%" LIMIT 5
```

```
 * 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

Total Payload Mass

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

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

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

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

```
%sql SELECT MIN("Date") FROM SPACEXTBL WHERE "Mission_Outcome" LIKE "Success"
```

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

MIN("Date")
01-03-2013

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT "Booster_Version" FROM SPACEXTBL WHERE "Mission_Outcome" LIKE "SUCCESS" AND "PAYLOAD_MASS__KG_" BETWEEN 4000 AND 600
0
```

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

Booster_Version
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 FT B1032.2
F9 B4 B1040.2
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1046.3
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT COUNT(*) FROM SPACEXTBL WHERE "Mission_Outcome" LIKE "%Success%" UNION SELECT COUNT(*) FROM SPACEXTBL WHERE "Mission_Outcome" LIKE "%Failure%"
```

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

COUNT(*)
1
100

Boosters Carried Maximum Payload

```
%sql SELECT "Booster_Version","PAYLOAD_MASS__KG_" FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_=(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPA
CEXTBL)
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

```
%sql SELECT substr(Date, 4, 2), "Mission_Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTBL WHERE substr(Date,7,4)='2015'
```

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

substr(Date, 4, 2)	Mission_Outcome	Booster_Version	Launch_Site
01	Success	F9 v1.1 B1012	CCAFS LC-40
02	Success	F9 v1.1 B1013	CCAFS LC-40
03	Success	F9 v1.1 B1014	CCAFS LC-40
04	Success	F9 v1.1 B1015	CCAFS LC-40
04	Success	F9 v1.1 B1016	CCAFS LC-40
06	Failure (in flight)	F9 v1.1 B1018	CCAFS LC-40
12	Success	F9 FT B1019	CCAFS LC-40

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

```
%sql SELECT "Landing_Outcome" FROM SPACEXTBL WHERE "Mission_Outcome" LIKE "SUCCESS"
```

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

"Landing_Outcome"
Landing_Outcome
Landing_Outcome
Landing_Outcome
Landing_Outcome
Landing_Outcome
Landing_Outcome
Landing_Outcome
Landing_Outcome
Landing_Outcome
Landing_Outcome
Landing_Outcome
Landing_Outcome
Landing_Outcome
Landing_Outcome
Landing_Outcome
Landing_Outcome

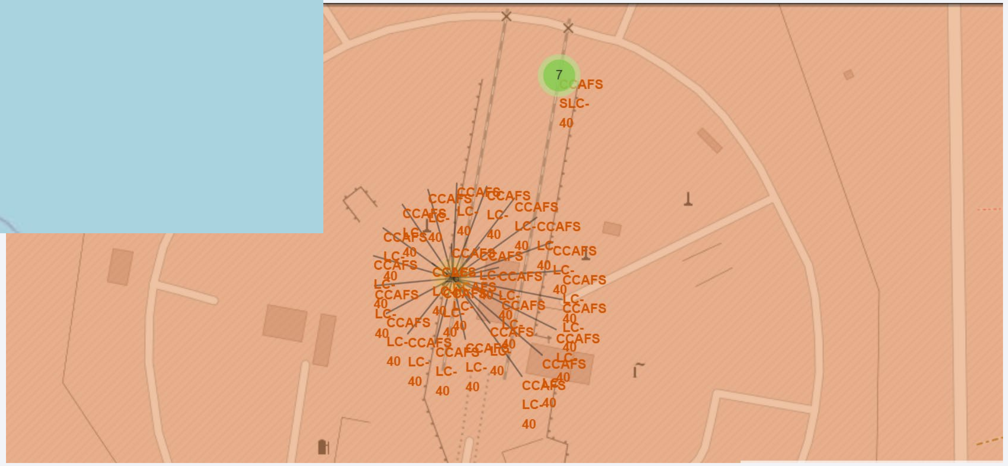
Section 3

Launch Sites Proximities Analysis

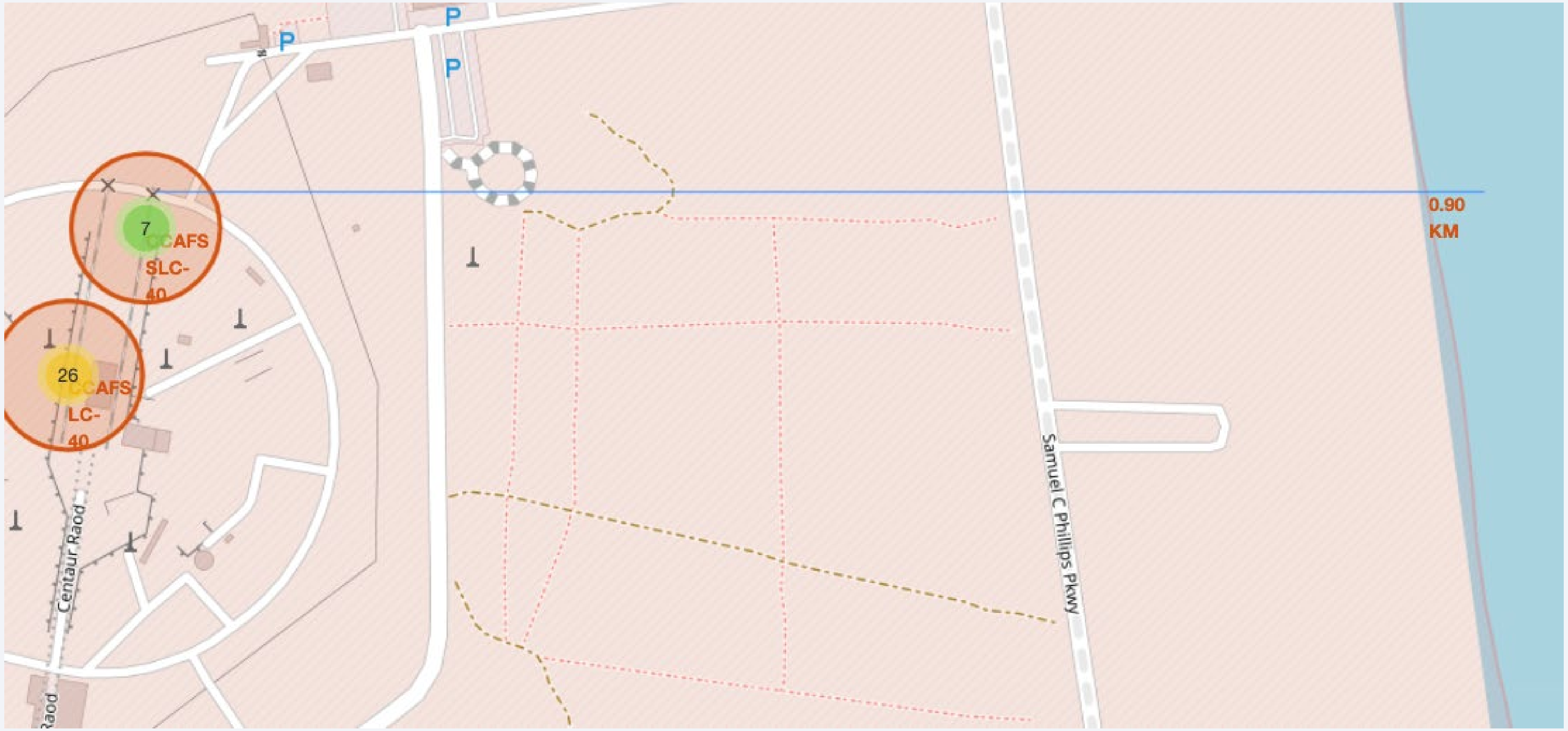
All Launch Sites on Global Map

- Most launch sites are near coastlines

Labelled Launch Outcomes



Proximities of Launch Site to Important Locations

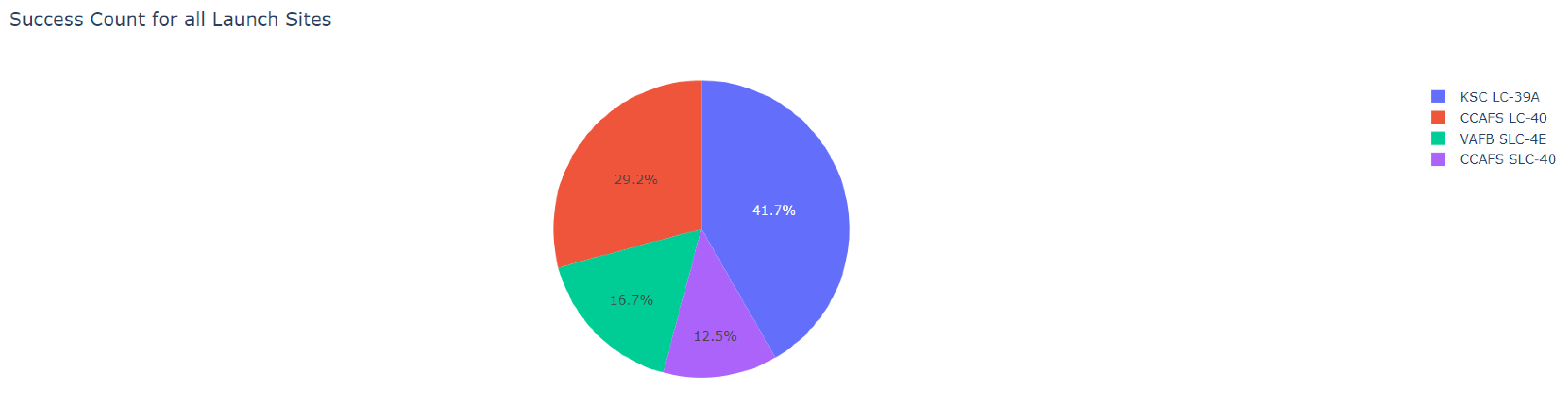


IBM Example image

Section 4

Build a Dashboard with Plotly Dash

Launch Success Count for All Sites



Piechart for the launch site with highest launch success ratio

- Unavailable

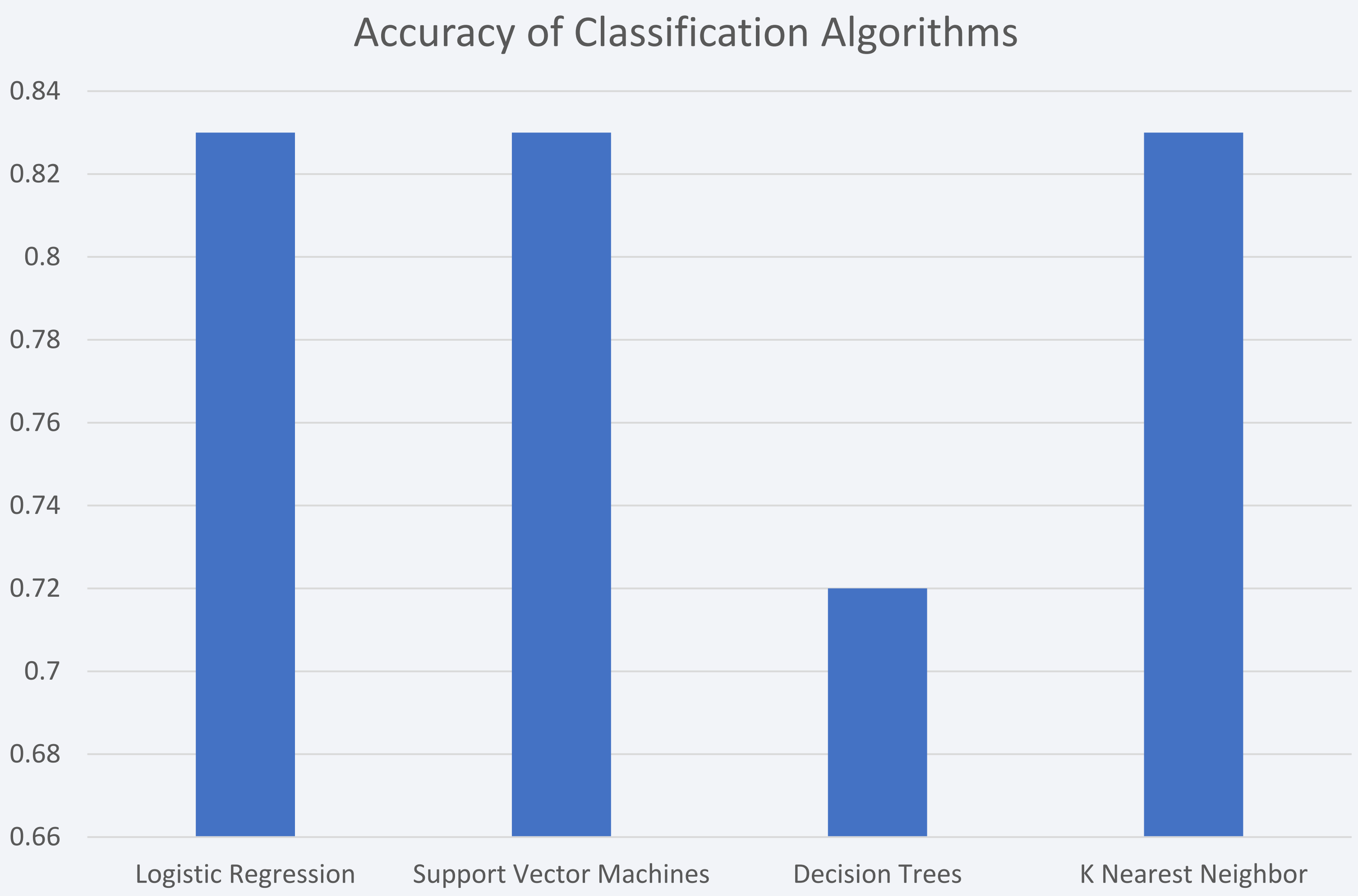
Payload vs. Launch Outcome Scatter Plot for all sites

- Unavailable

Section 5

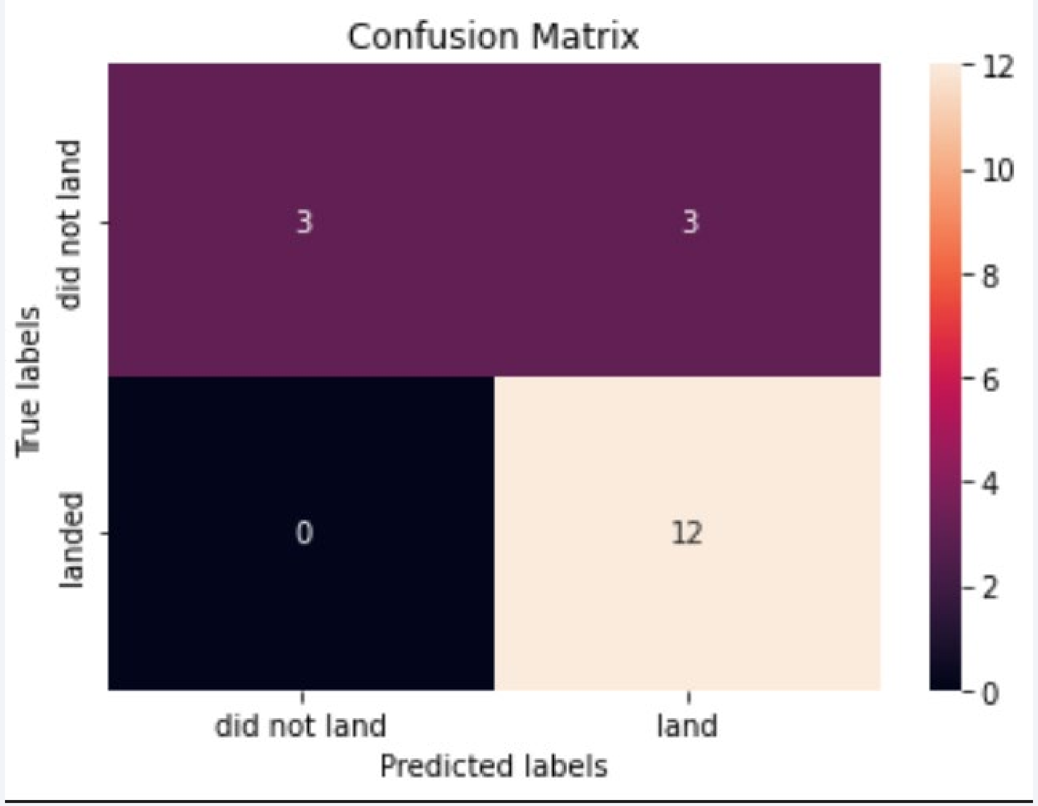
Predictive Analysis (Classification)

Classification Accuracy



Confusion Matrix

- 12/12 of “landed” labels were successfully labelled as “landed”. About 50% of the time, the model was correct about the “did not land” label.



Conclusions

- Logistic Regression, Support Vector Machines, and K Nearest Neighbors provided the highest accuracy with predictions, at about 83.3%.
- Decision Trees provided the lowest accuracy about predictions, at about 72%.

Thank you!