IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Todd Kurtz
July 2, 2023

Outline

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

Executive Summary

- This capstone will try to predict if Space X Falcon 9 first stage will land successfully using machine learning and Data Collection, Data Wrangling and Formatting
- The process for the project used the following steps:
    - Data Collection
    - Exploratory data analysis
    - Visualization
    - Machine Learning Prediction
- Based on different aspects of the rocket launch, we will try and determine whether the outcome will be a success or failure

Introduction

- At a cost of 62 million dollars per launch, Space X provides a substantial savings over its competitors. The savings is achieved mostly by reusing the first stage of the rocket. By predicting if the first stage will land, we can therefore determine the cost of each launch.
- We are hoping to solve the main problem of determining if the first stage of the rocket will land successfully. Given a set of features, is the outcome predictable?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using the SpaceX API and Web scraping
- Perform data wrangling
  - Data was cleansed looking for nulls and normalization
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Decision Tree
  - Logistic Regression
  - Support vector machine
  - K-nearest neighbors

Data Collection

- Data sets were collected via the SpaceX API and web scraping
  - SpaceX API: https://api.spacexdata.com/v4/rockets/
  - Web scraping was performed from Wikipedia
    - https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

Data Collection – SpaceX API

- Using the GET request, we used the API to import Booster Version, Launchpad, Payload, and Rocket Core information
- Data was parsed from the json output, Falcon 9 Boosters were filtered and null values were replaced using the mean of the column
- Total data set is 90 rows and 17 columns.

1. Get Rocket
2. Get Launchpad
3. Get Payloads
4. Get Cores
5. Parse and normalize the response
6. Filter for Falcon 9 only
7. Replace nulls with column mean

- [Applied-Data-Science-Capstone/01 Capstone Lab 1 Collecting the Data.ipynb at main · mberringer/Applied-Data-Science-Capstone (github.com)]

Data Collection - Scraping

- Using Beautiful Soup to extract HTML tables from Wikipedia

- Applied-Data-Science-Capstone/01a Capstone Lab 1a Complete the Data Collection with Web Scraping lab.ipynb at main · mberringer/Applied-Data-Science-Capstone (github.com)

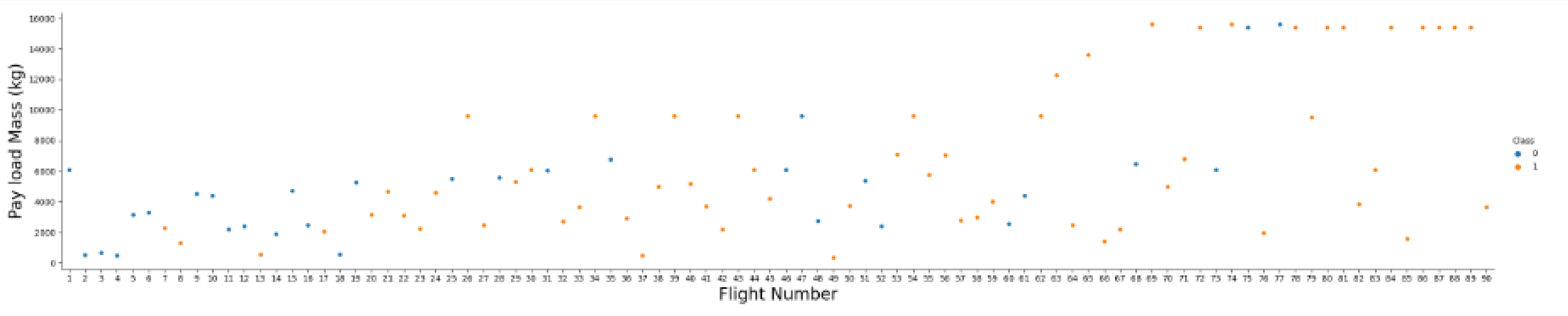
1. Extract the HTML tables from Wikipedia
2. Parse the table and convert to Pandas data frame

Data Wrangling

- The object is to determine if the booster successfully landed or if it failed
- We performed exploratory Data analysis and determined training labels
- We first imported out data set from the prior steps
- We identified missing values for each attribute
- Calculated number of launches per site, the number of occurrences of each orbit, number and occurrence per orbit type and mission outcome
- Created the landing outcome label
- Applied-Data-Science-Capstone/02 Capstone Lab 2 Data Wrangling.ipynb at main · mberringer/Applied-Data-Science-Capstone · GitHub

EDA with Data Visualization

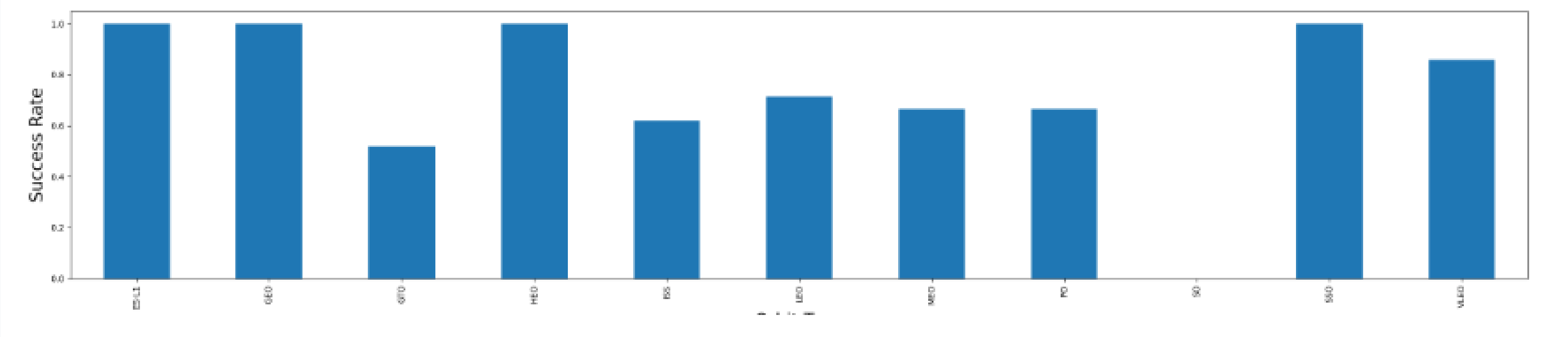
- We plotted scatter plots of Flight Number and Payload Mass to determine success rate



- [Applied-Data-Science-Capstone/04 Capsone Lab 4 EDA with Visualization Lab.ipynb at main · mberringer/Applied-Data-Science-Capstone · GitHub]()

EDA with Data Visualization

- We plotted bar charts to determine success rate of each orbit



- [Applied-Data-Science-Capstone/04 Capsone Lab 4 EDA with Visualization Lab.ipynb at main · mberringer/Applied-Data-Science-Capstone · GitHub]

EDA with SQL

- Names of unique launch sites

```
%%sql
SELECT DISTINCT LAUNCH_SITE
FROM SPACEXTBL;
```

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

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

- Applied-Data-Science-Capstone/03 Capsone Lab 3 Complete the EDA with SQL.ipynb at main · mberringer/Applied-Data-Science-Capstone · GitHub

EDA with SQL

- CCA launch sites

```
%%sql
SELECT LAUNCH_SITE
FROM SPACEXTBL
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```

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

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

- Applied-Data-Science-Capstone/03 Capsone Lab 3 Complete the EDA with SQL.ipynb at main · mberringer/Applied-Data-Science-Capstone · GitHub

EDA with SQL

- Total payload mass for NASA boosters

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) as PayloadMass
FROM SPACEXTBL
WHERE Customer = 'NASA (CRS)';
```

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

PayloadMass
45596.0

- [Applied-Data-Science-Capstone/03 Capsone Lab 3 Complete the EDA with SQL.ipynb at main · mberringer/Applied-Data-Science-Capstone · GitHub]()

EDA with SQL

- Average payload mass for booster Falcon 9

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

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) as AvgPayloadMass
FROM SPACEXTBL
WHERE Booster_Version LIKE 'F9 v1.0%';
```

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

AvgPayloadMass
340.4

- [Applied-Data-Science-Capstone/03 Capsone Lab 3 Complete the EDA with SQL.ipynb at main · mberringer/Applied-Data-Science-Capstone · GitHub]

EDA with SQL

- First successful landing outcome on the ground pad
- Name of boosters which have success in drone ship and have a payload mass between 4,000 and 6,000
- Total Number of successful and failure mission outcomes
- Booster version names that have carried the maximum payload mass
- Landing outcome rank between 6/4/2010 and 3/20/2017
- Applied-Data-Science-Capstone/03 Capsone Lab 3 Complete the EDA with SQL.ipynb at main · mberringer/Applied-Data-Science-Capstone · GitHub

Build an Interactive Map with Folium

- We marked all launch sites on a map and added the success rate for each site
- We calculated the distance between a launch site to its proximities
- We used circles and labels to display launch sites
- We used markers and lines to display the success rate of each site along with its proximity to other sites
- Applied-Data-Science-Capstone/05 Capsone Lab 5 Interactive Visual Analytics with Folium.ipynb at main · mberringer/Applied-Data-Science-Capstone · GitHub

Build a Dashboard with Plotly Dash

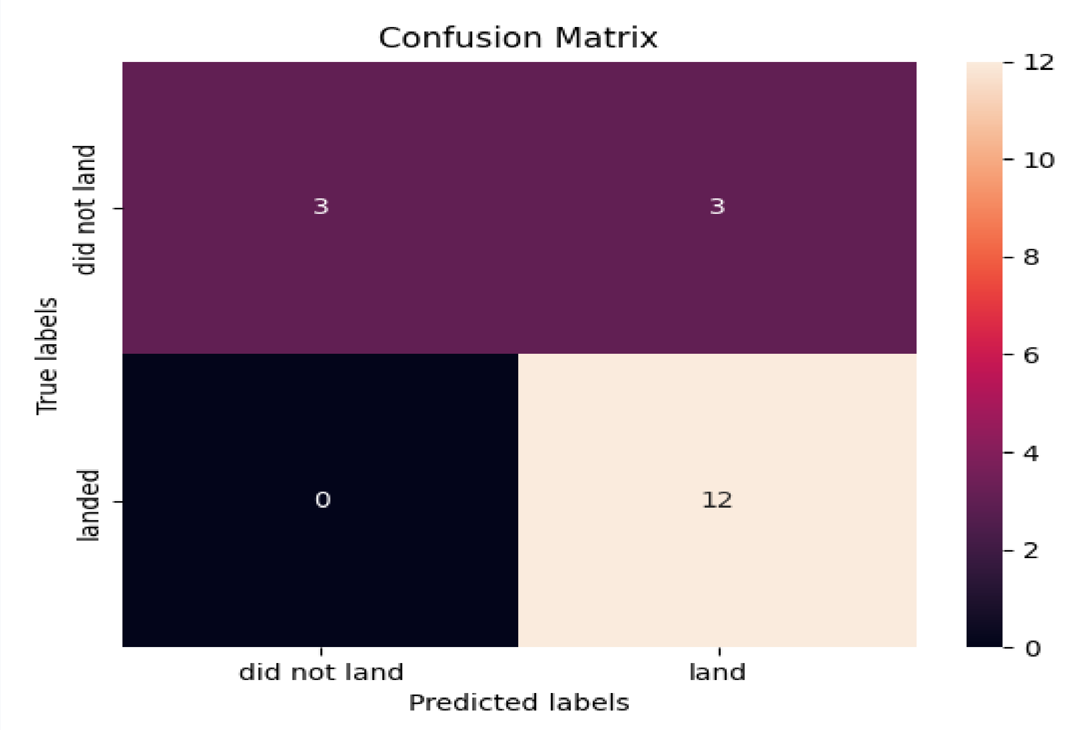
- Dash functions were used to generate an interactive app to toggle input using dropdown menus and sliders
- Pie charts and scatterplots were used to show total launch success rate from each site and the correlation of payload mass to mission outcome per site
- Applied-Data-Science-Capstone/06 Captstone Lab 6 Build and Interactive Dashboard with Ploty Dash.py at main · mberringer/Applied-Data-Science-Capstone · GitHub

Predictive Analysis (Classification)

- Sklearn library was used to create the machine learning models
- We plotted the confusion matrix for each of the models to determine effectiveness
- Data was first standardized then split into training and testing data sets
- Models created included:
  - Logistic regression
  - Support Vector machine (SVM)
  - Decision tree
  - K nearest neighbors (KNN)
- Models were fit on the training set and the best combination of parameters for each model was determined
- Model accuracy was based on the confusion matrix and accuracy scores
- Applied-Data-Science-Capstone/07 Capsone Lab 7 Complete the Machine Learning Prediction lab.ipynb at main · mberringer/Applied-Data-Science-Capstone · GitHub

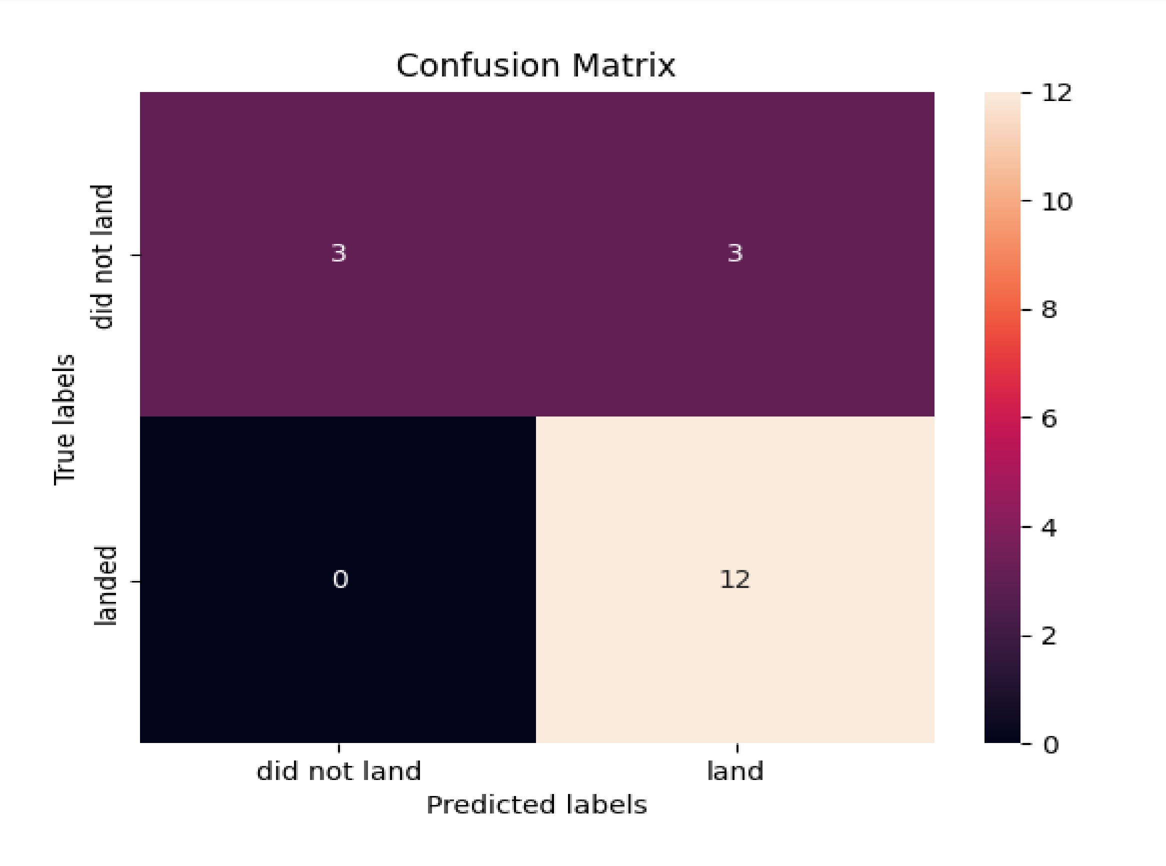
Results

- Logistic regression
  - Major problem with false positives
  - Accuracy: 0.848



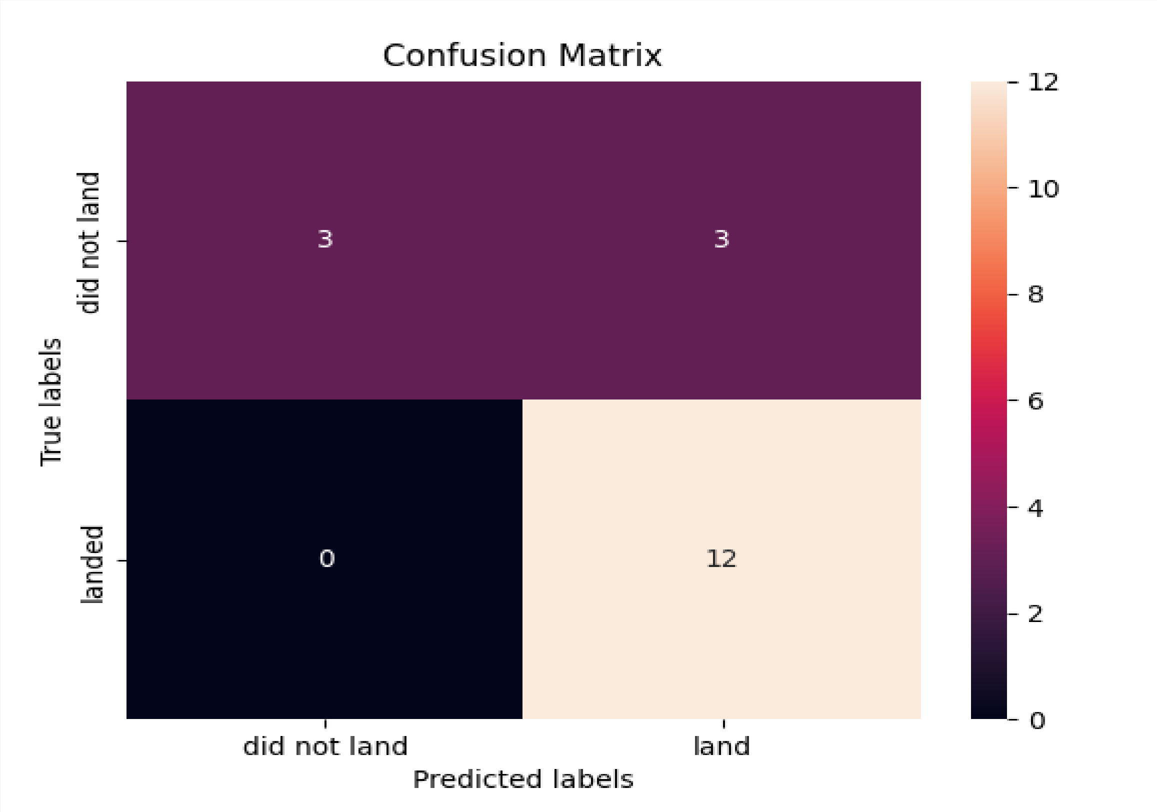
Results

- Support Vector Machine
  - Accuracy: 0.833



Results

- Decision Tree
  - Accuracy: 0.875



Results

- K nearest neighbors (KNN)
  - Accuracy: 0.833

Results

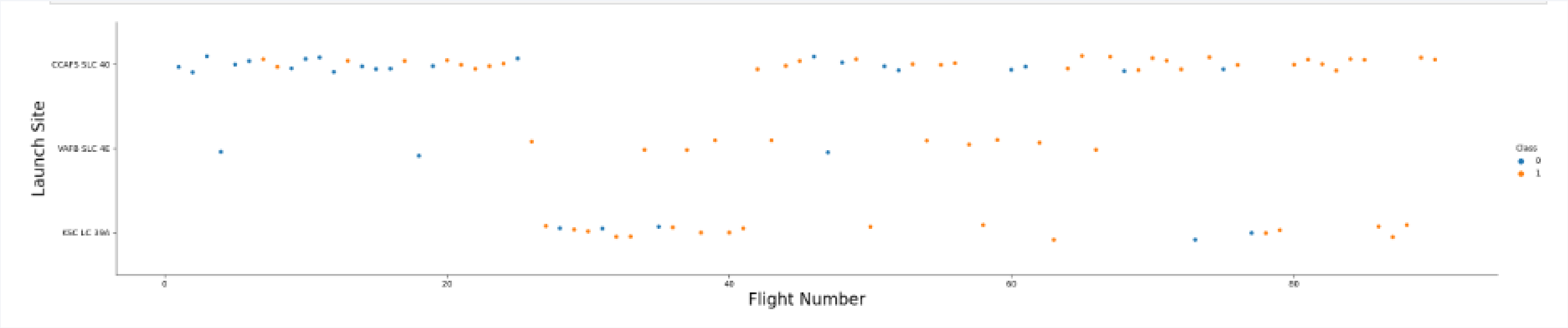
- Using each model GridSearchCV, the best scores are used to rank each model
- The models are ranked in order below, best to worst
  - Decision Tree
  - K nearest neighbors
  - Support Vector machine
  - Logistic regression

Section 2

Insights drawn from EDA

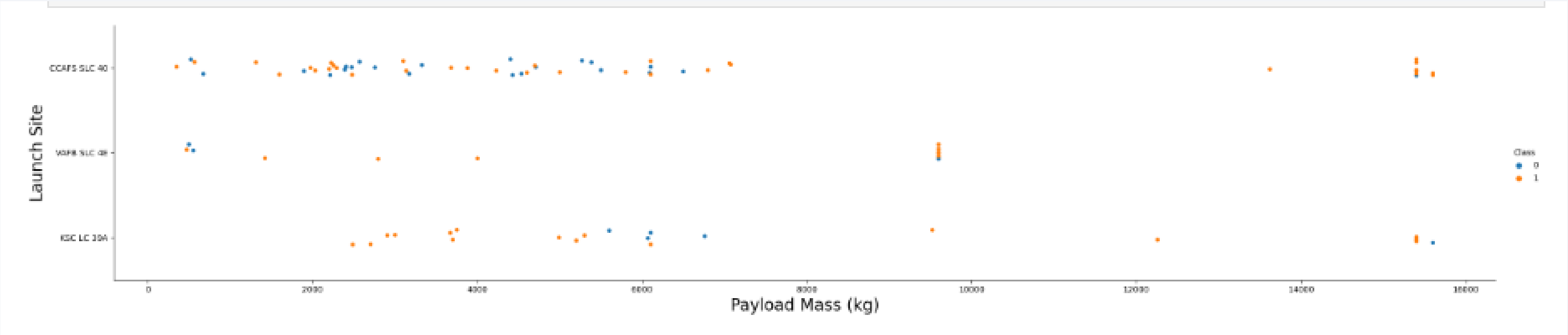
Flight Number vs. Launch Site

- Show a scatter plot of Flight Number vs. Launch Site



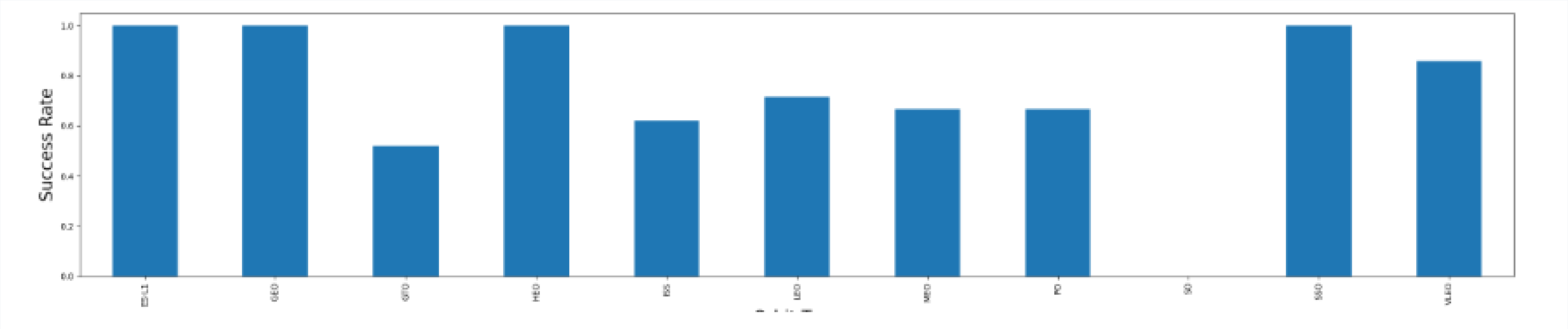
Payload vs. Launch Site

- Show a scatter plot of Payload vs. Launch Site

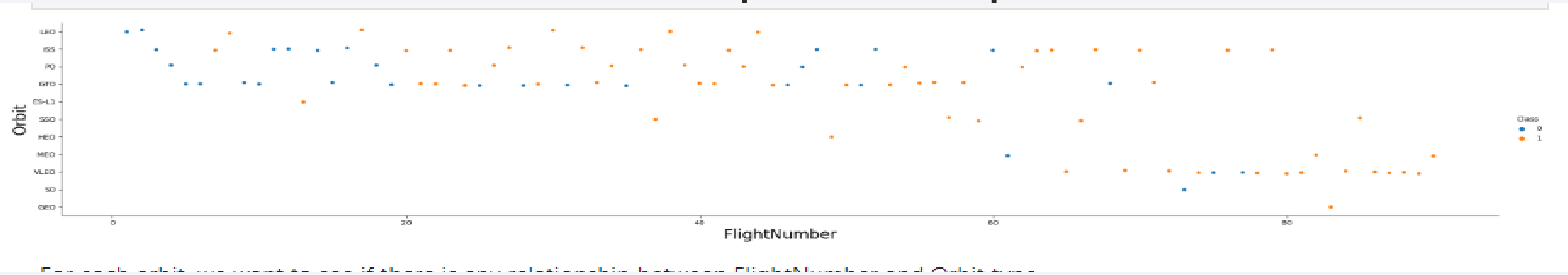


Success Rate vs. Orbit Type

- Show a bar chart for the success rate of each orbit type

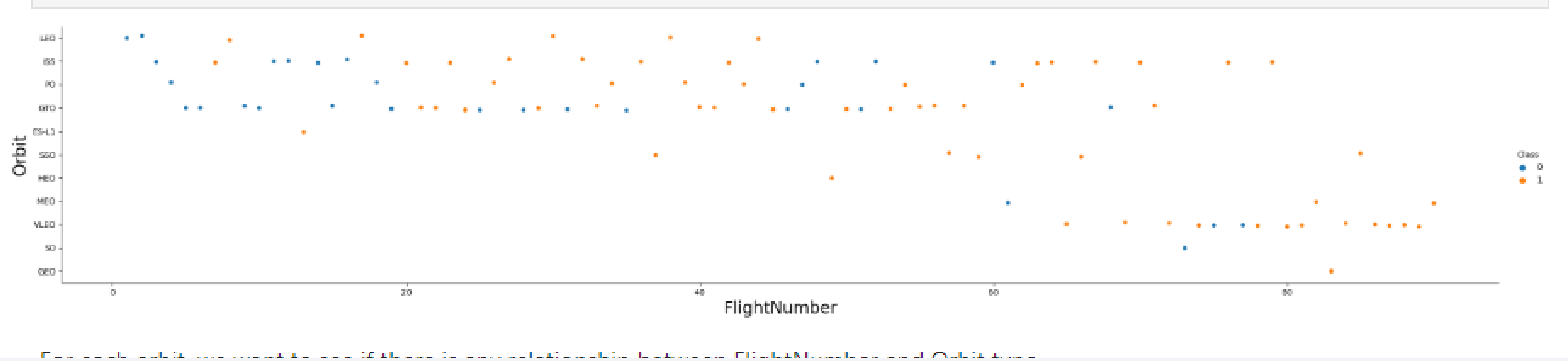


- Show the screenshot of the scatter plot with explanations



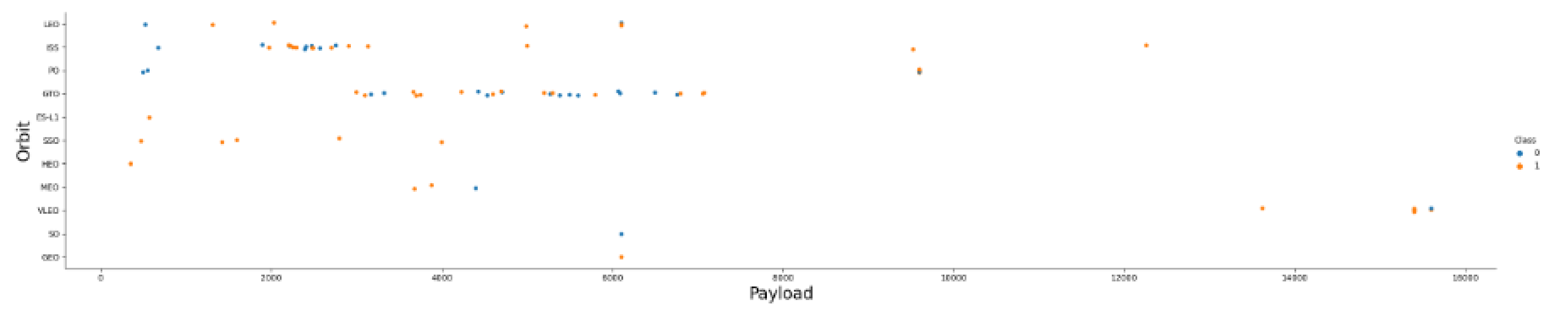
Flight Number vs. Orbit Type

- Show a scatter point of Flight number vs. Orbit type



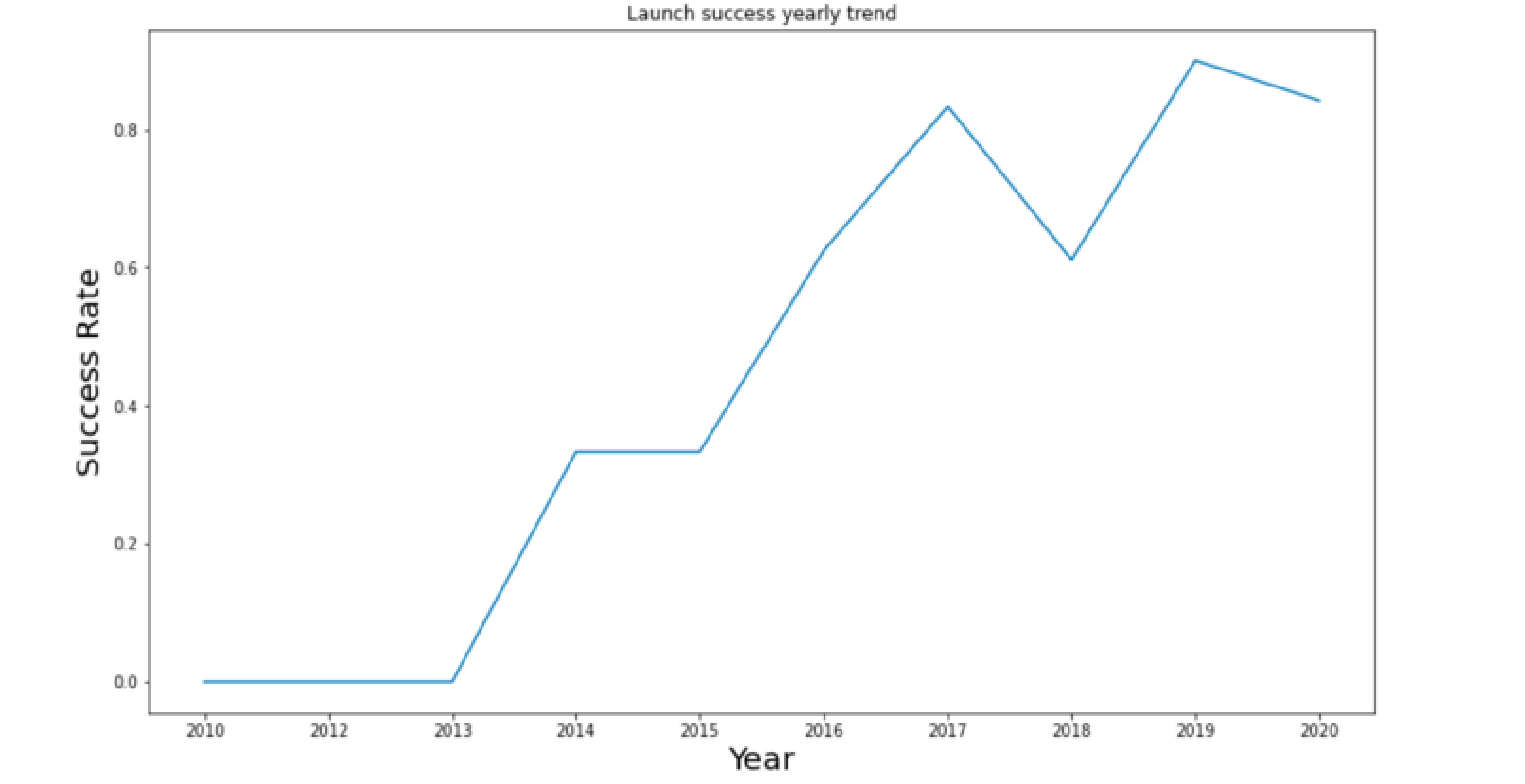
Payload vs. Orbit Type

- Show a scatter point of payload vs. orbit type



Launch Success Yearly Trend

- Show a line chart of yearly average success rate



All Launch Site Names

- Find the names of the unique launch site

```
%%sql
SELECT DISTINCT LAUNCH_SITE
FROM SPACEXTBL;
```

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

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

Display 5 records where launch sites begin with

```
%%sql
SELECT LAUNCH_SITE
FROM SPACEXTBL
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```

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

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

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

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) as PayloadMass
FROM SPACEXTBL
WHERE Customer = 'NASA (CRS)';
```

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

PayloadMass
45596.0

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

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

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) as AvgPayloadMass
FROM SPACEXTBL
WHERE Booster_Version LIKE 'F9 v1.0%';
```

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

AvgPayloadMass
340.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

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)';
```

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

MIN(Date)
01/08/2018

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

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 LANDING_OUTCOME = 'Success (drone ship)'
AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
```

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

]:

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

List the total number of successful and failure mission outcomes

```
%%sql
SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER
FROM SPACEXTBL
GROUP BY MISSION_OUTCOME;
```

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

Mission_Outcome	TOTAL_NUMBER
None	0
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

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

```
]:
%%sql
SELECT DISTINCT BOOSTER_VERSION
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (
     SELECT MAX(PAYLOAD_MASS__KG_)
     FROM SPACEXTBL);
```

```
 * 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

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

```
%%sql
SELECT LANDING_OUTCOME, COUNT(LANDING_OUTCOME) AS TOTAL_NUMBER
FROM SPACEXTBL
GROUP BY LANDING_OUTCOME
ORDER BY TOTAL_NUMBER DESC
```

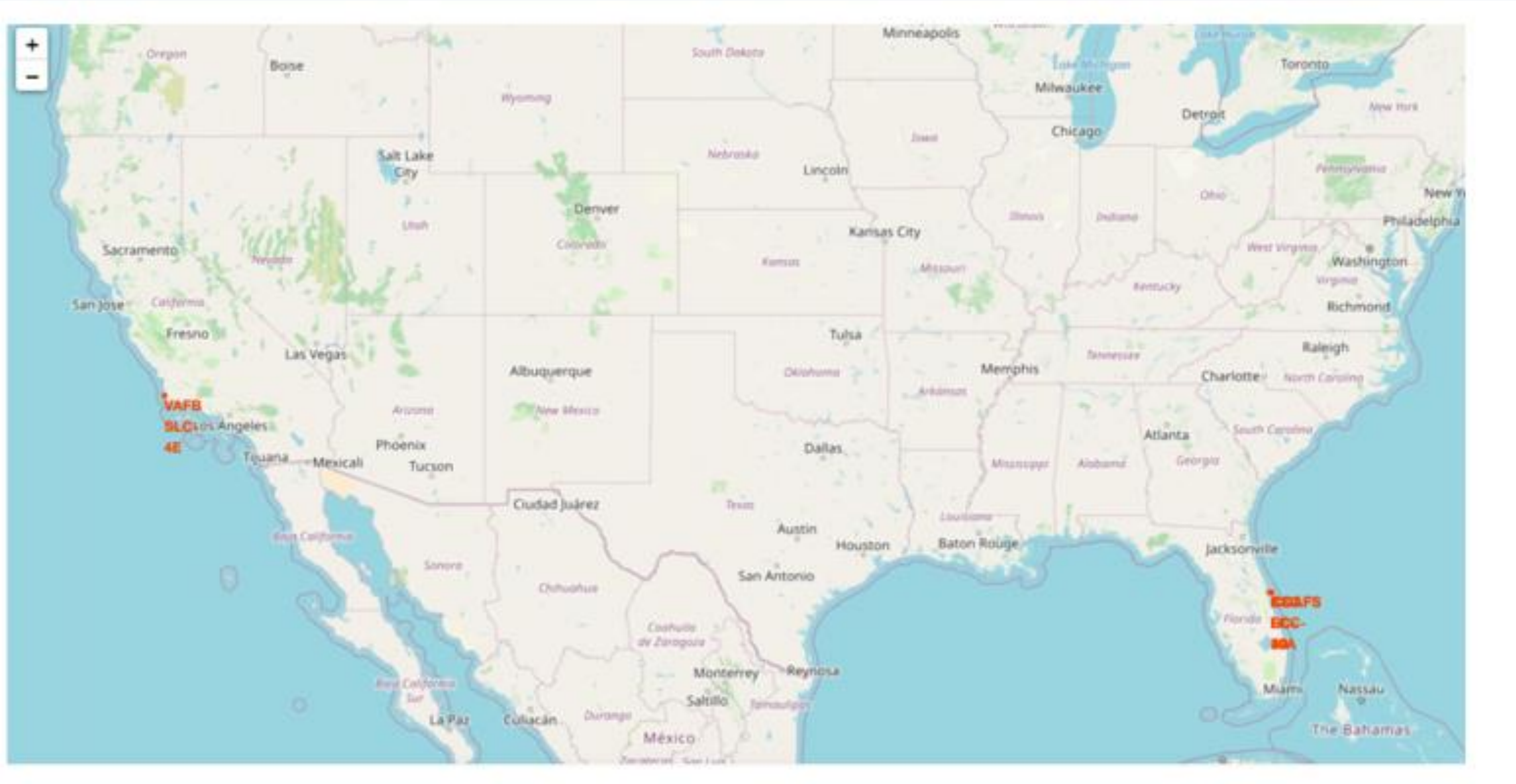
```
* sqlite:///my_data1.db
one.
```

Landing_Outcome	TOTAL_NUMBER
Success	38
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Failure	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1
No attempt	1
None	0

Section 3

Launch Sites Proximities Analysis

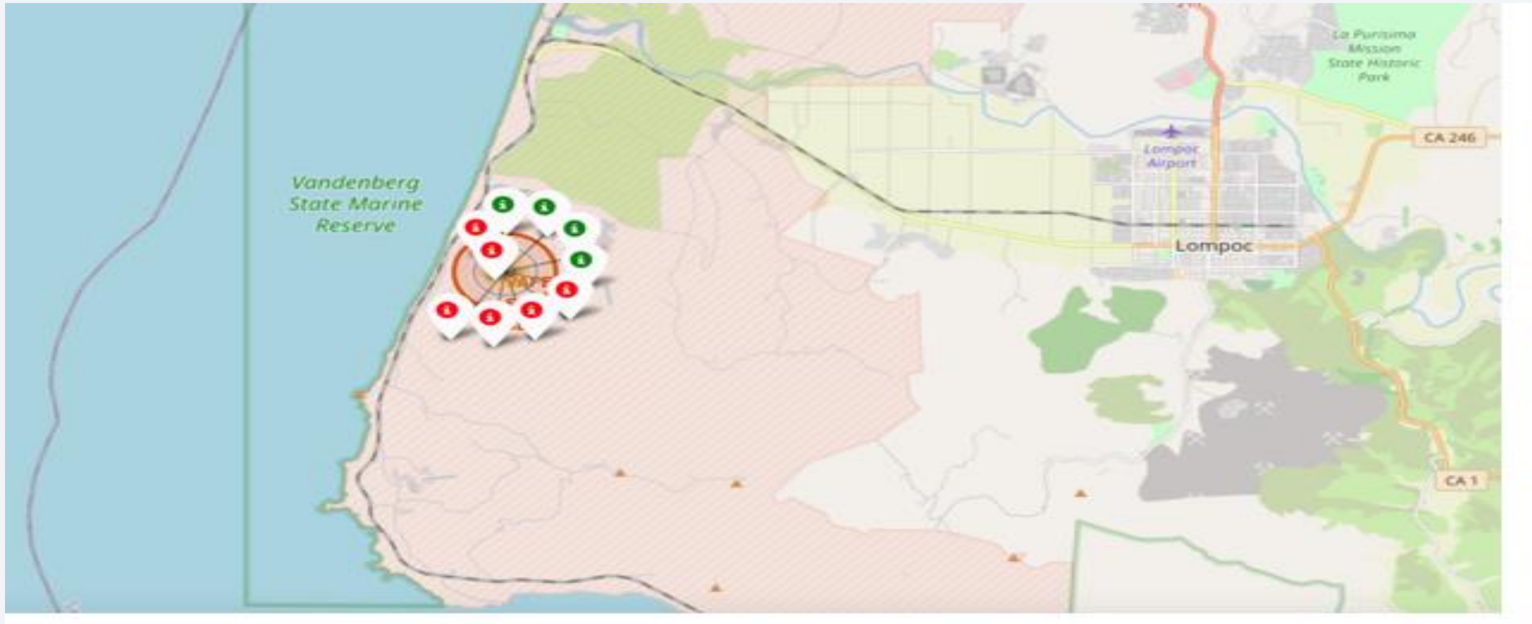
All Launch sites map



Success Rate for each launch site

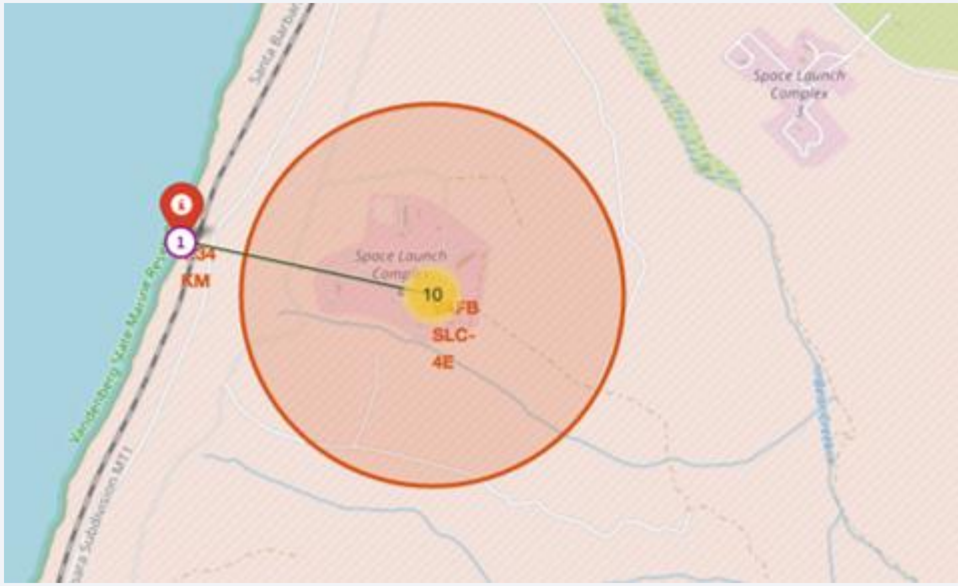
Successful launch's per site

- Zooming in shows green or red with green indicating a success and red indicating a failure



Distance between launch sites and proximities

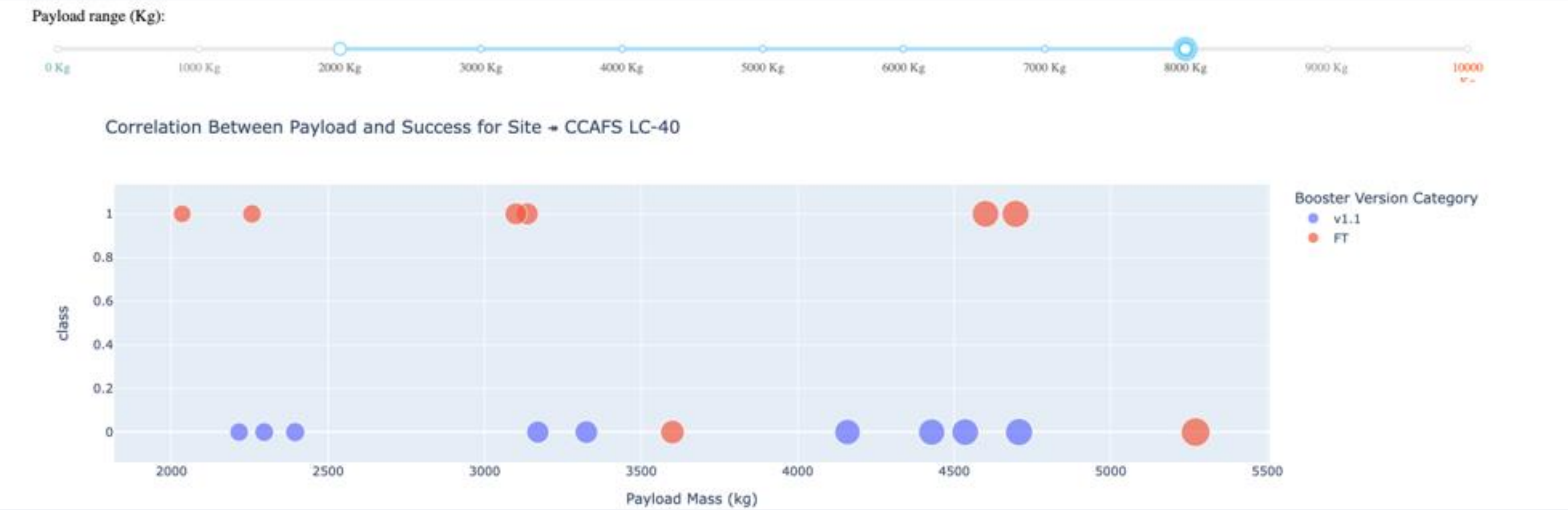
- The distance between a launch site and the proximity including the nearest city, highway or railway.
- Proximity of Site VFAB SLC 4E to nearest coastline



Section 4

Build a Dashboard with Plotly Dash

Payload mass and success rate

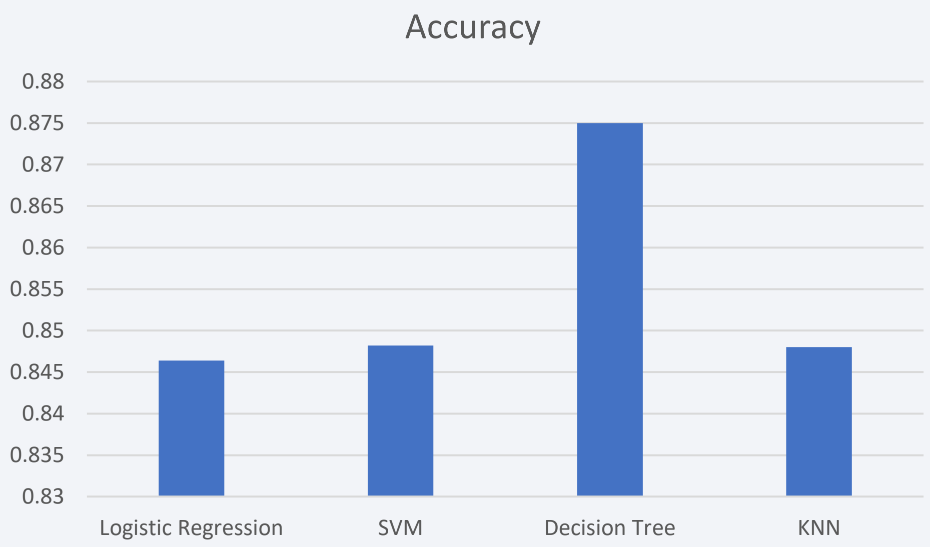


Section 5

Predictive Analysis (Classification)

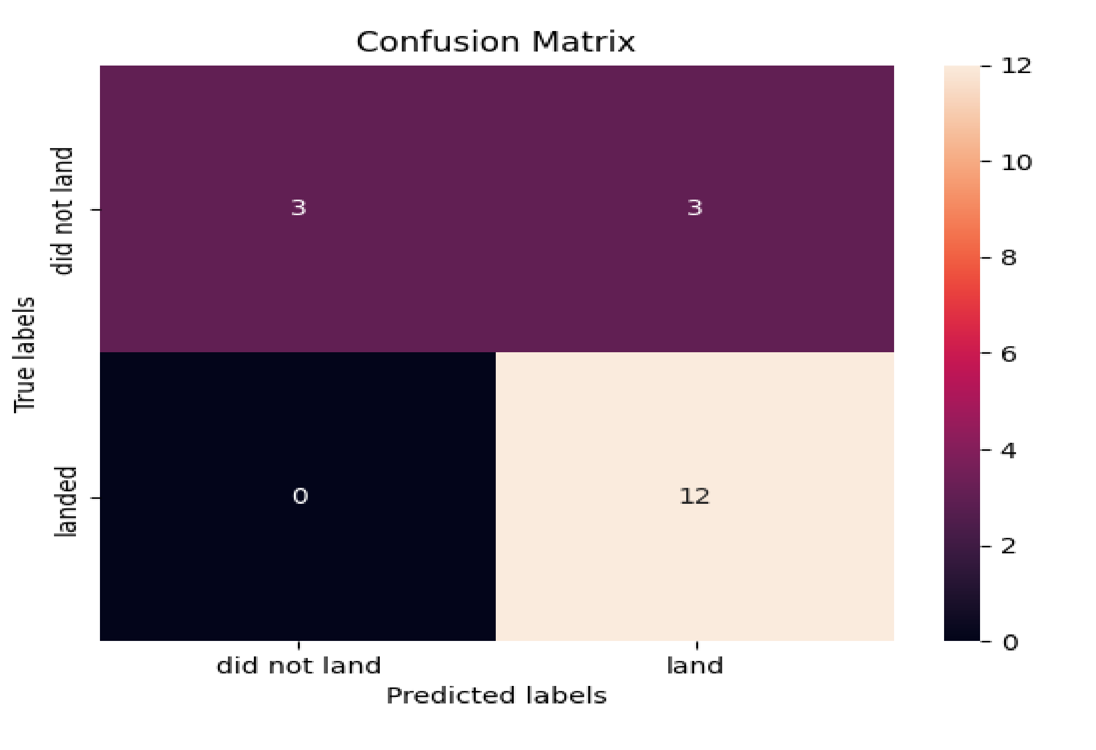
Classification Accuracy

Decision Tree is the most accurate



Confusion Matrix

- Decision tree confusion matrix



Conclusions

- We attempted to predict the success rate of the Falcon 9 booster given a set of parameters to determine the cost of each launch
- Payload Mass, Orbit Type and launch site may impact the success rate
- Using different machine learning models to make the prediction, the decision tree model was the most accurate

Thank you!