IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Fuya Koshiro
03/12/2023

Outline

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

Executive Summary

- Objective of this research
  - To predict whether a new launch in the future would success or fail
- Summary of methodologies
  - The data used in this research were collected from SpaceX Rest API and Wikipedia
  - Train model (Logistic Regression, SVM, KNN, Tree) and compare the accuracy
- Summary of all results
  - Decision-Making Tree model has the highest accuracy of 83%

Introduction

- Project background and context
  - Space X Falcon 9 costs around1/3 of other companies because of the reuse of the first stage
  - To keep this competence, return of the first stage is key
- Problem
  - It is important to know preferred conditions for the first stage to land to return

Section 1

Methodology

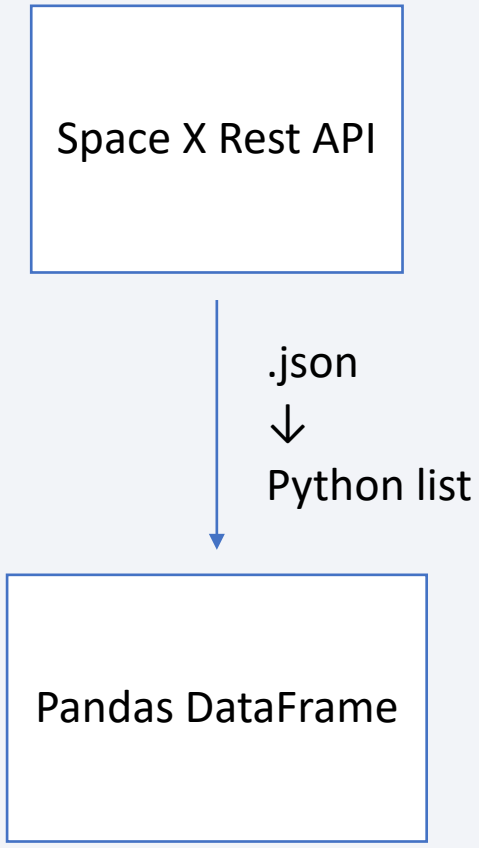
Methodology

Executive Summary

- Data collection methodology:
  - Space X Rest API
  - Wikipedia using bs4
- Perform data wrangling
  - Convert outcome into class: 1 for successful landing and 0 for unsuccessful landing
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Split the data into train (80%) and test (20%) data
  - Perform grid research with 10 folds for cross validation
  - Models are Logistic Regression, SVM, KNN, and Decision Tree
  - Perform evaluation using accuracy

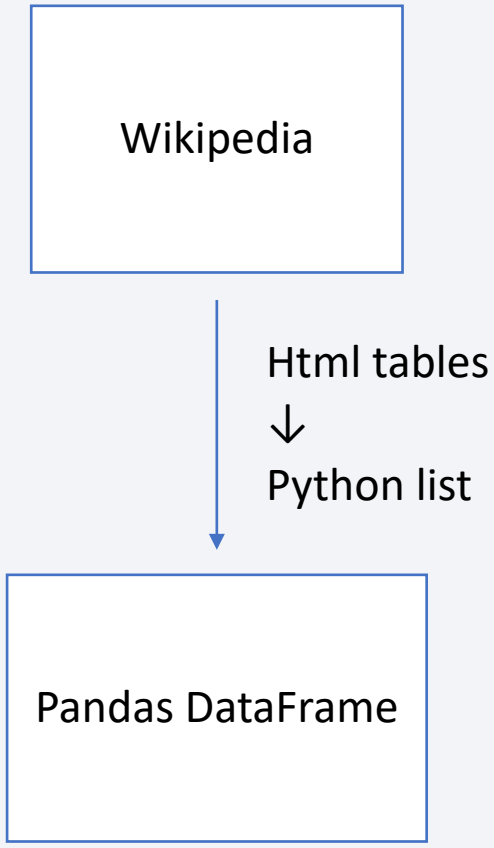
Data Collection – SpaceX API

- Access Space X Rest API and extract important columns with .json format
- The important columns are something that could affects the result of landing of the first stage
- Append each columns in the .json response to python lists
- Concatenate the lists to make a data frame
- GitHub URL of the completed SpaceX API calls notebook (https://github.com/FuyaKoshiro/coursera_ds_ch10/blob/main/jupyter-labs-spacex-data-collection-api.ipynb)



Data Collection - Scraping

- Used a Wikipedia page to scrape tables using bs4
- Since “Customer” column has a missing value, np.nan was used to fill the empty value
- Concatenate the lists to make a data frame
- GitHub URL of the completed web scraping notebook (https://github.com/FuyaKoshiro/coursera_ds_ch10/blob/main/jupyter-labs-webscraping.ipynb)



Data Wrangling

- “Outcome” column has 8 different expressions to indicate whether the landing was successful or unsuccessful
- Extract the keys of the unsuccessful outcomes and make it a list
- Classify the values in the “Outcome” column into 1 (successful) and 0 (unsuccessful) referring the list
- the GitHub URL of your completed data wrangling related notebooks (https://github.com/FuyaKoshiro/coursera_ds_ch10/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb)

Successful
1

Unsuccessful
0

EDA with Data Visualization

- Visualized some combinations of columns which could affect the outcome, and visually check if there is any relationship between the selected values
    - Scatter plot of “Payload Mass vs Flight Number”
    - Scatter plot of “Launch Site vs Flight Number”
    - Scatter plot of “Launch Site vs Payload Mass”
    - Bar chart of “Success Rate vs Orbit”
    - Scatter plot of “Flight Number vs Orbit”
    - Scatter plot of “Payload Mass vs Orbit”
    - Line plot of “Date and Success Rate”
- Created dummy variables for the categorical columns to enable machine learning
- Changed the data type of numerical columns to float64
- the GitHub URL of your completed EDA with data visualization notebook (https://github.com/FuyaKoshiro/coursera_ds_ch10/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb)

EDA with SQL

- Using the SQL queries, following tasks were performed:
  - Displaying the unique launch site names
  - Displaying the total number of failure and success of mission outcomes
  - Displaying the maximum payload mass of each booster version
  - Displaying the info of failure of the drone ship

- the GitHub URL of your completed EDA with SQL notebook (https://github.com/FuyaKoshiro/coursera_ds_ch10/blob/main/jupyter-labs-eda-sql-course.ipynb)

Build an Interactive Map with Folium

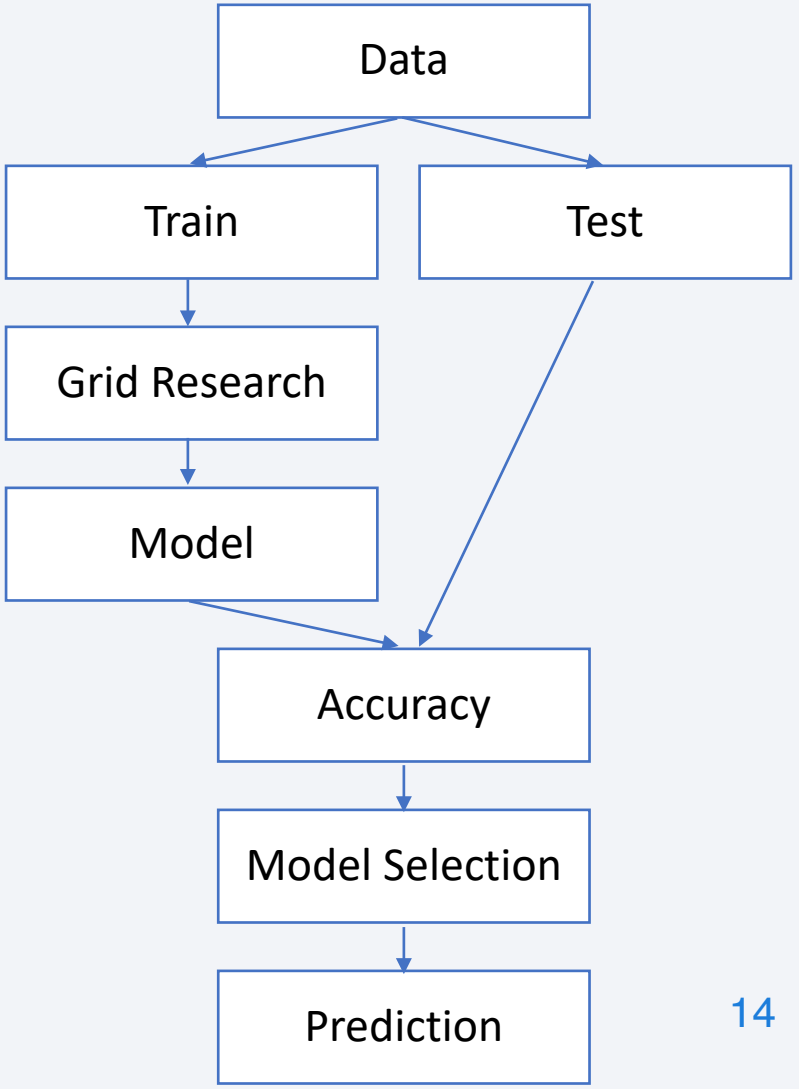
- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
- Explain why you added those objects
- Add the GitHub URL of your completed interactive map with Folium map, as an external reference and peer-review purpose

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
- Plots/graphs added are including:
  - Pie charts of success count & rate corresponding with a location
    - Location can be chosen from the dropdown menu
  - Scatter plot of “class vs payload mass” with booster engines indicated with circles
    - X axis can be zoomed in/out
- the GitHub URL of your completed Plotly Dash lab (https://github.com/FuyaKoshiro/coursera_ds_ch10/blob/main/spacex_dash_app.py)

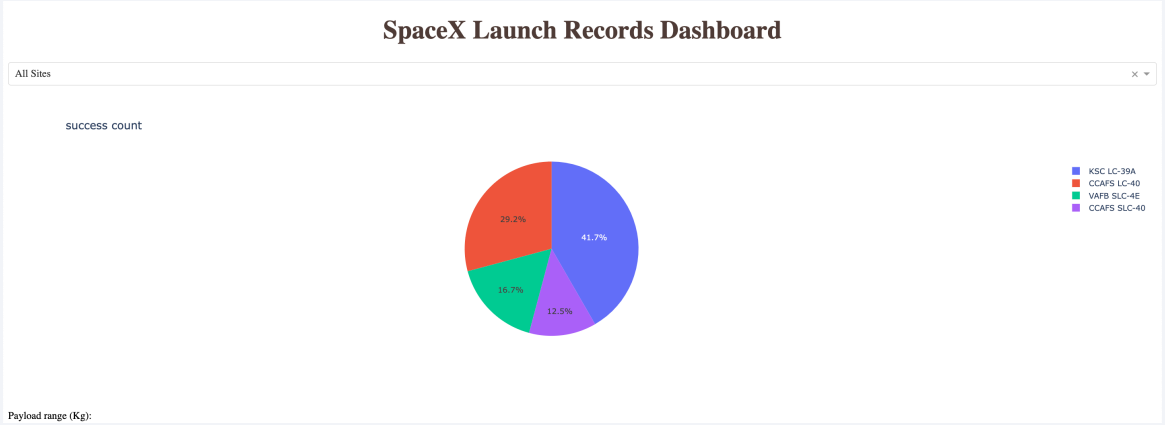
Predictive Analysis (Classification)

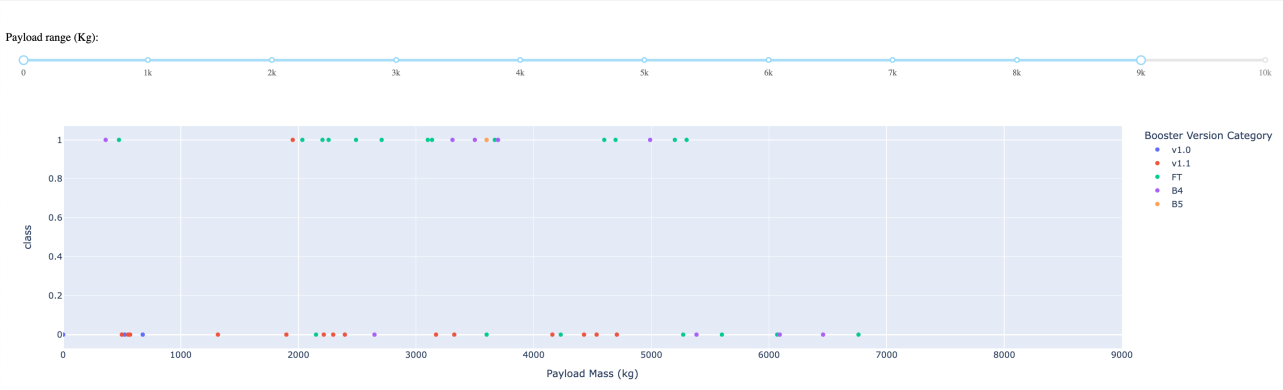
- Summarize how you built, evaluated, improved, and found the best performing classification model
- The data was split into train and test data
- Grid research was conducted using the cross validation with 10 folds.
- Models were LR, SVM, KNN, Tree.
- the GitHub URL of your completed predictive analysis lab (https://github.com/FuyaKoshiro/coursera_ds_ch10/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb)



Results

- Exploratory data analysis results
  - KSC LC-39A has the highest successful counts with the highest successful ratio
  - There were no successful case with over 5,000 kg payload mass
  - FT booster engine has the most successful cases
- Predictive analysis results
  - Tree model has the highest accuracy of 83%





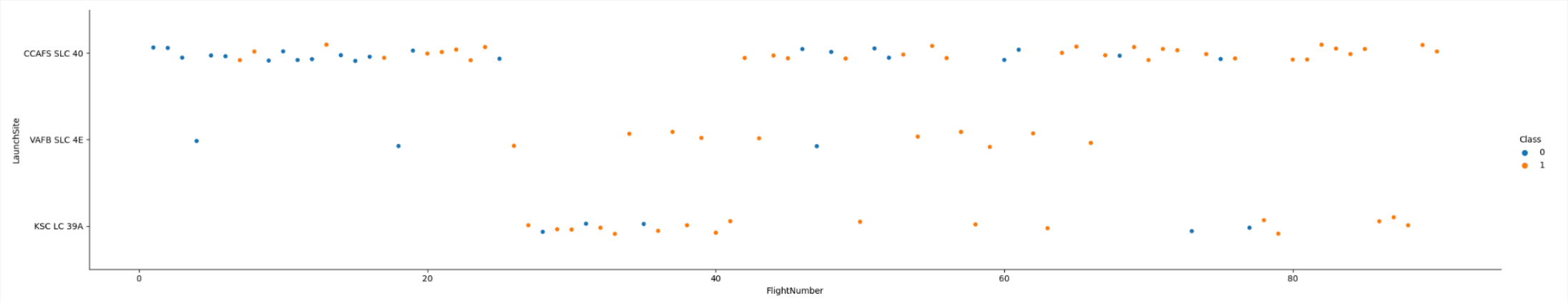
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

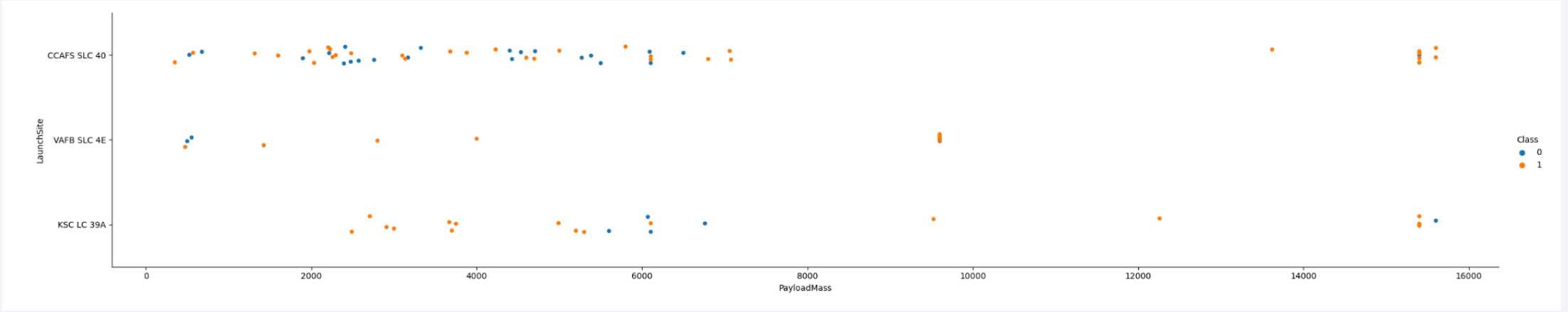
Overall, as the trials go by, the success rate increases

CAFS SLC 40 are used the most, and recent launches displayed the successful outcomes.

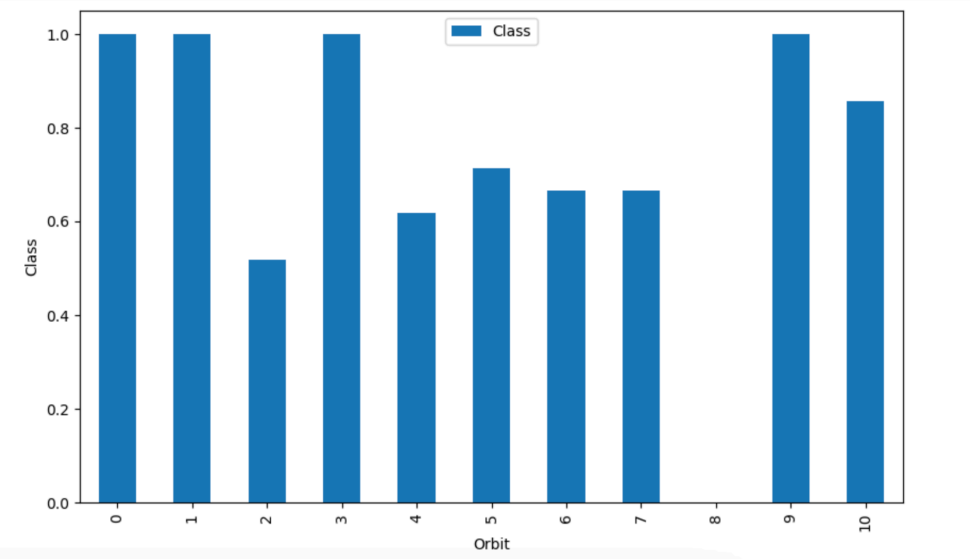


Payload vs. Launch Site

For the VAFB-SLC launch site, there are no rockets launched for heavy payload mass (greater than 10000 kg)



Success Rate vs. Orbit Type

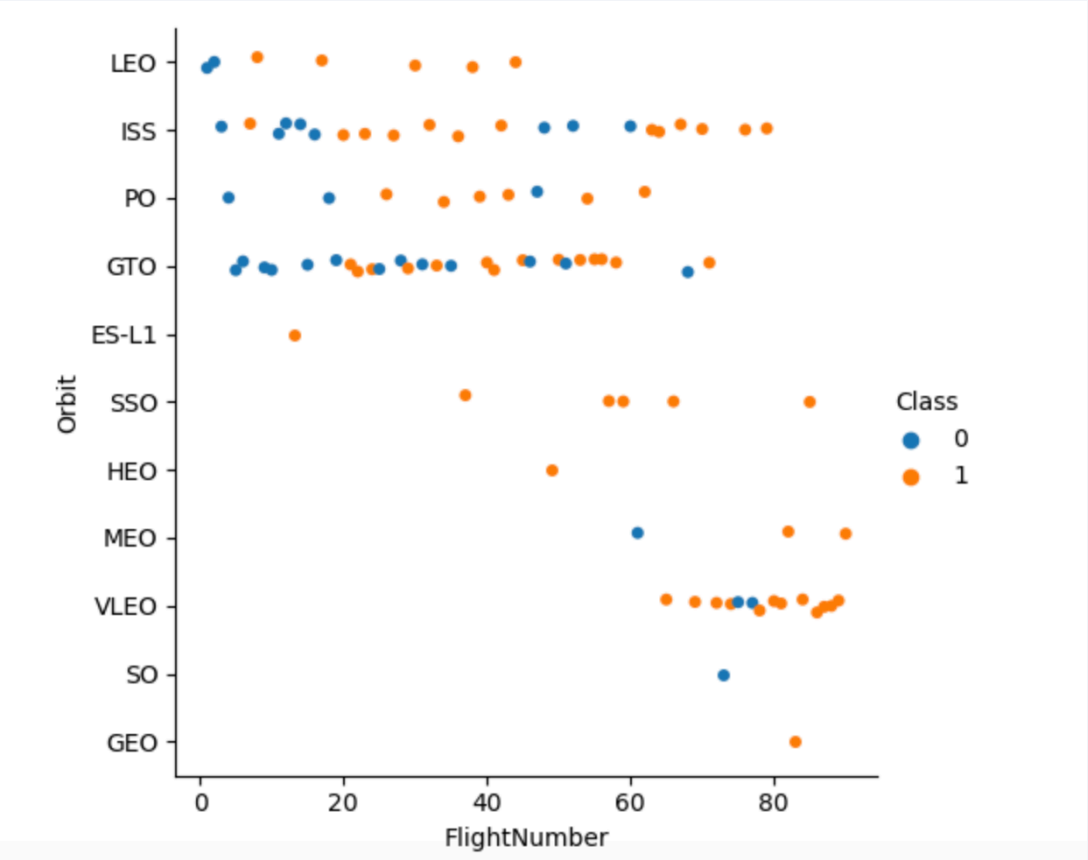


	Orbit	Class
0	ES-L1	1.000000
1	GEO	1.000000
2	GTO	0.518519
3	HEO	1.000000
4	ISS	0.619048
5	LEO	0.714286
6	MEO	0.666667
7	PO	0.666667
8	SO	0.000000
9	SSO	1.000000

ES-L1, GEO, HEO, SSO have the highest successful rate.

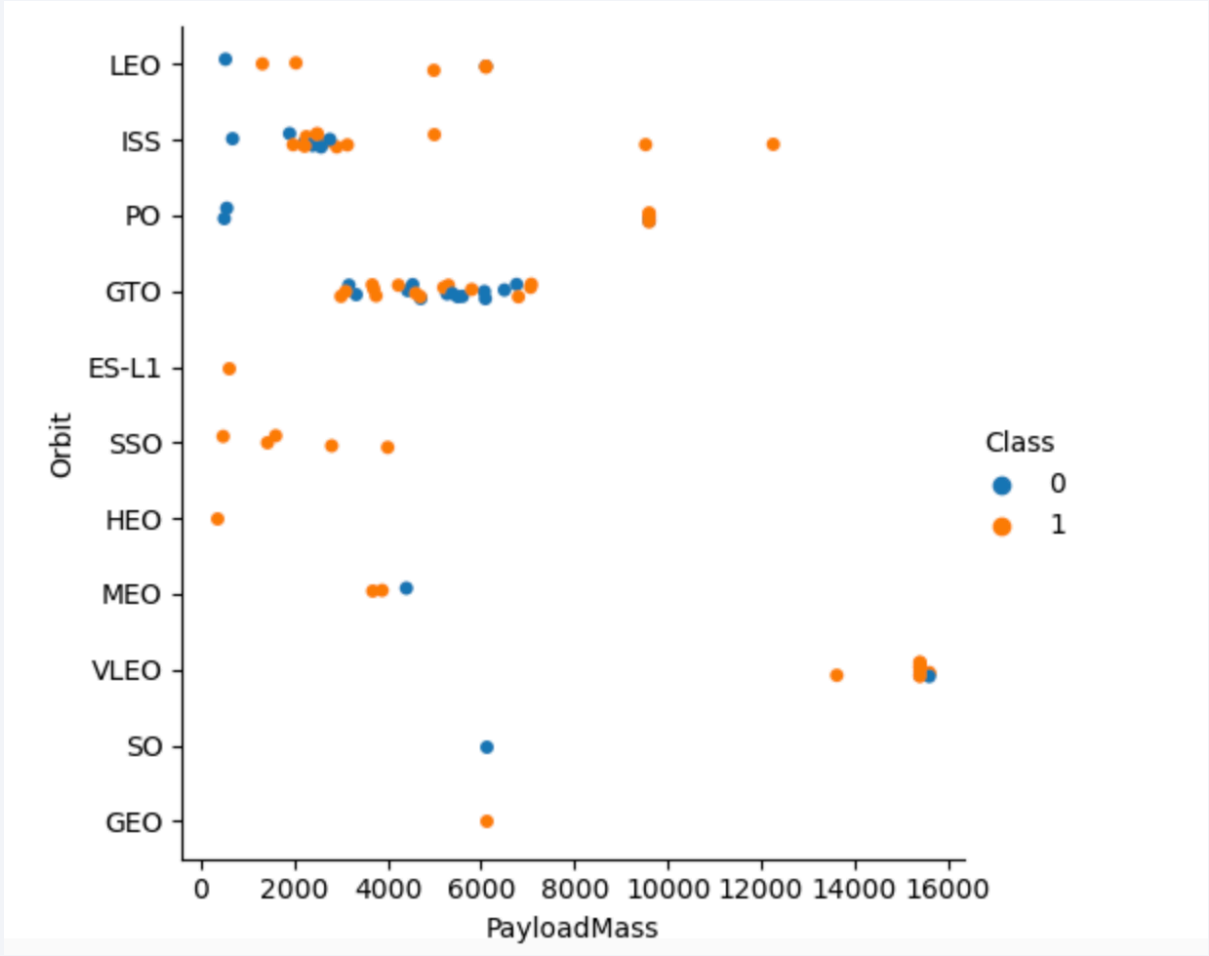
Flight Number vs. Orbit Type

- In the LEO orbit the Success appears related to the number of flights
- There seems to be no relationship between flight number when in GTO orbit



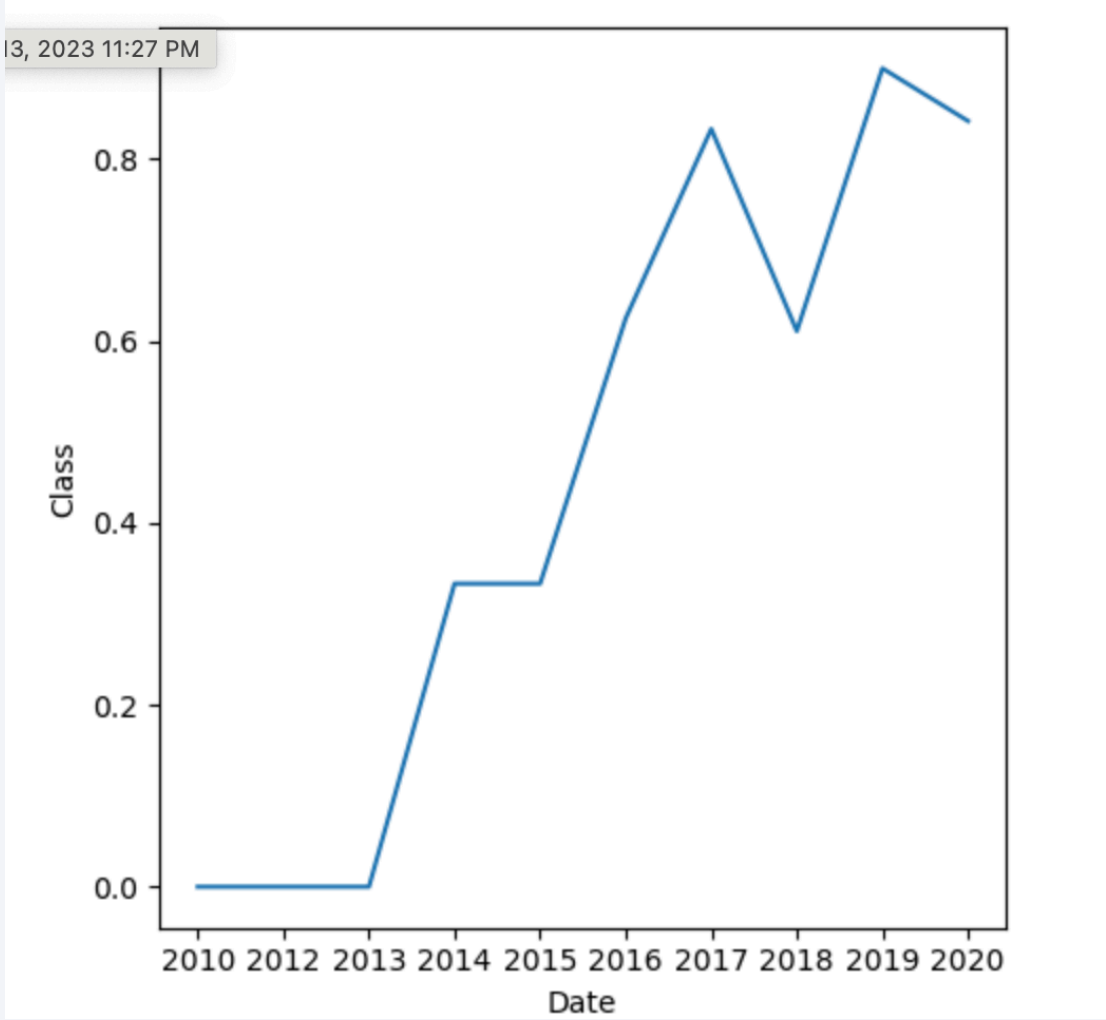
Payload vs. Orbit Type

- In GTO, the class is not affected by payload mass because the both points were distributed closely
- For LEO, ISS, PO, when the payload mass is small, it likely fails



Launch Success Yearly Trend

- Since 2013, the success rate is increasing.



All Launch Site Names

- All the launch site are located close to the coasts.

```
%sql select distinct Launch_Site from SPACEXTBL
```

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`
- All of them are CCAFS LC-40

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

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

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

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

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

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

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

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

min(Date)
2015-12-22 00:00:00

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

```
%%sql
select Booster_Version
from spacextbl
where "Landing _Outcome" like "%Precluded%"
and 4000 < PAYLOAD_MASS__KG_ < 6000;
```

Booster_Version
F9 v1.1 B1018

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

```
%%sql
select count(Mission_Outcome)
from spacextbl
group by Mission_Outcome;
```

count(Mission_Outcome)
1
98
1
1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

```
%%sql
select Booster_Version,
max(PAYLOAD_MASS__KG_) as "maximum payload mass"
from spacextbl
group by Booster_Version;
```

Booster_Version	maximum payload mass
F9 B4 B1039.2	2647
F9 B4 B1040.2	5384
F9 B4 B1041.2	9600
F9 B4 B1043.2	6460
F9 B4 B1039.1	3310
F9 B4 B1040.1	4990
F9 B4 B1041.1	9600

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%%sql
select substr(Date, 6, 2) as month,
"Landing _Outcome",
Booster_Version,
Launch_Site
from spacextbl
where substr(Date, 1, 4) = '2015'
and "Landing _Outcome" like "%drone ship%";
```

month	Landing _Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
06	Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40

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(*) as count
from spacextbl
where Date between'2010-04-06' AND '2017-03-20'
and "Landing _Outcome" like "%Success%"
GROUP by "Landing _Outcome" ORDER BY count Desc
```

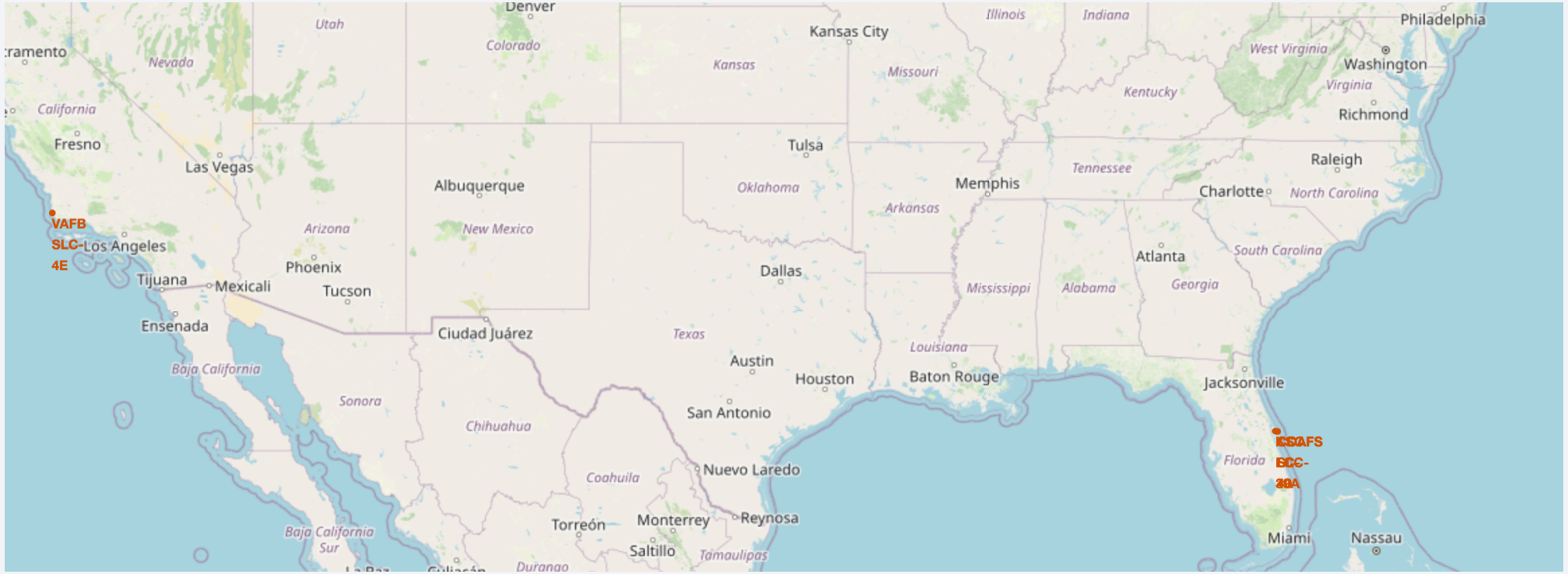
Landing _Outcome	count
Success (ground pad)	5
Success (drone ship)	5

Section 3

Launch Sites Proximities Analysis

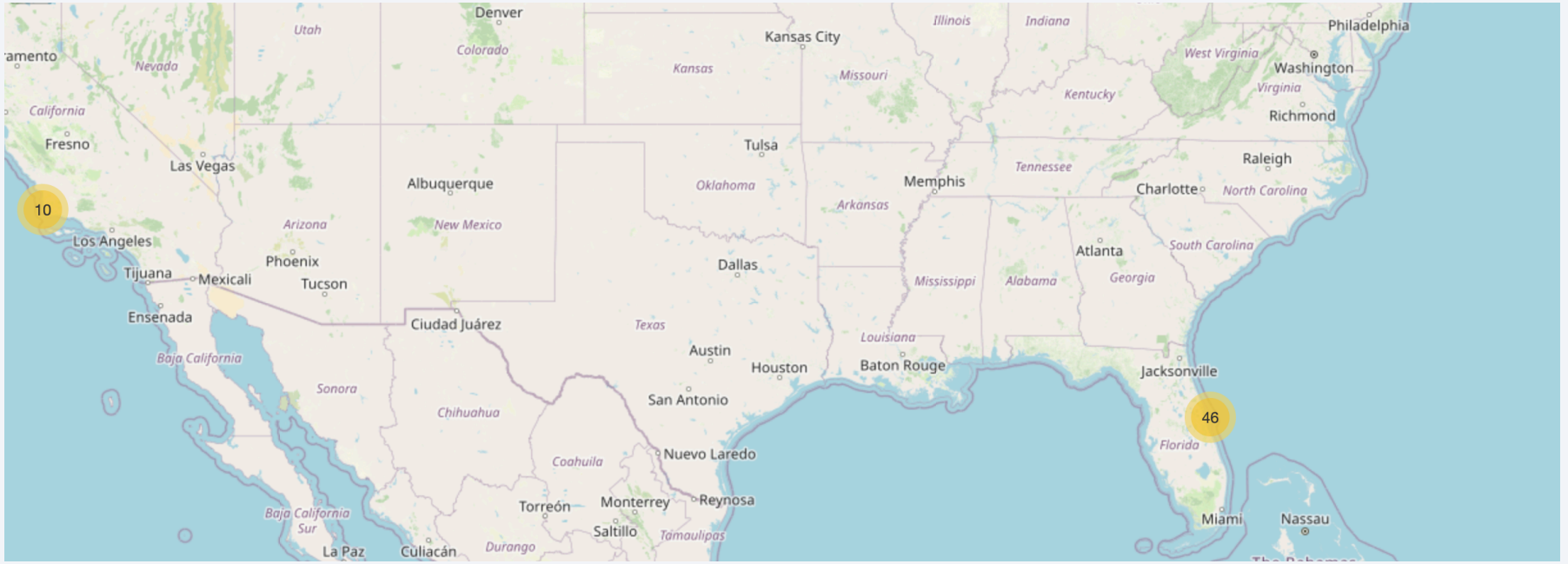
Location of the launch sites

- The circles indicate the location, and texts indicate the name of the sites
- All the launch sites are located close to the coasts



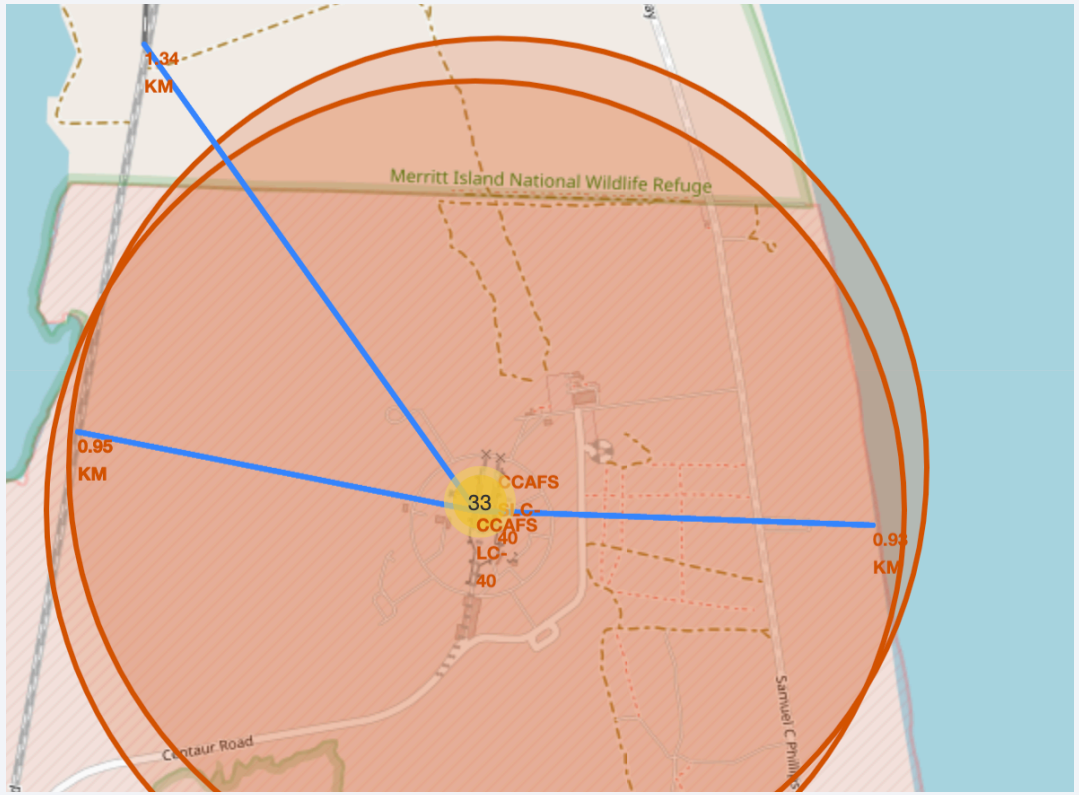
Launch sites and outcomes

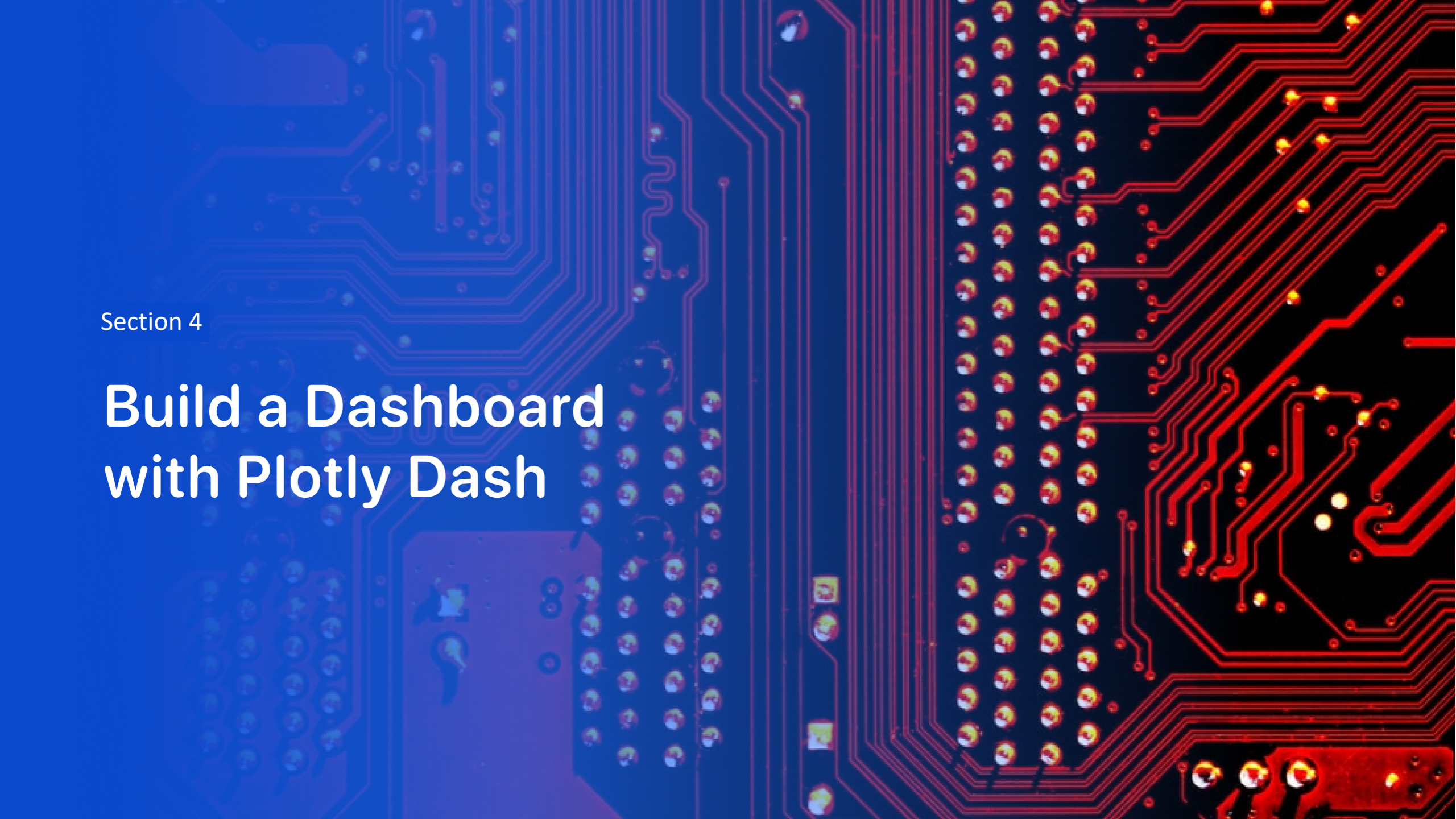
- The yellow circles indicates the launch sites, and the numbers indicates the number of successful outcomes
- The launch sites in the east coast has higher number of successful outcomes.



Distance between the proximities

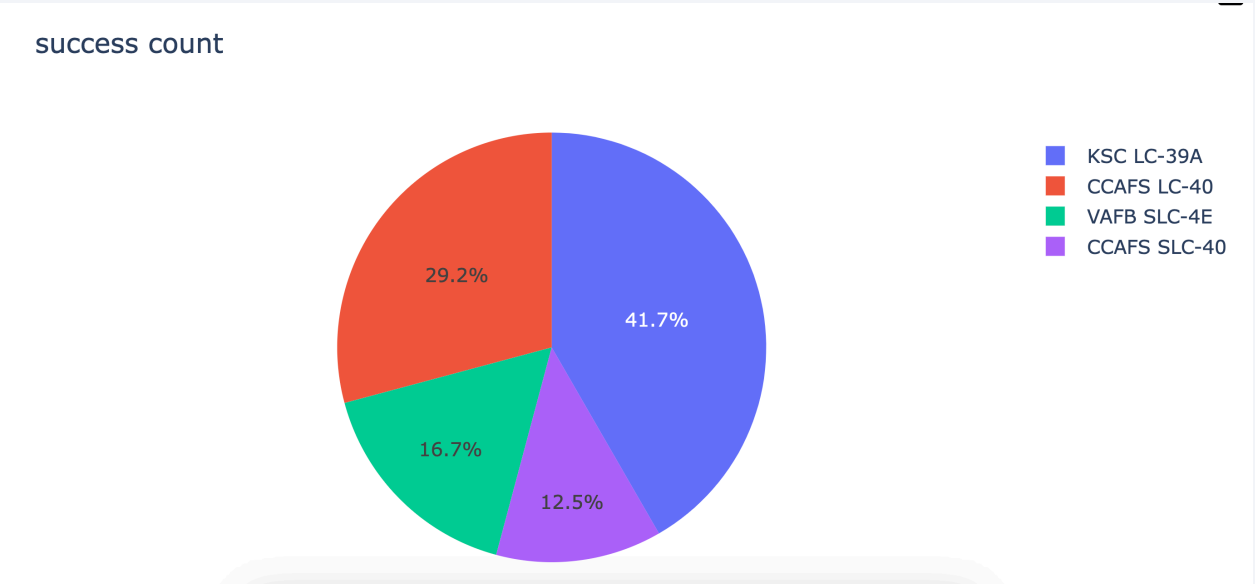
- The blue lines show the distance between CCAFS LC-40 and the proximities with the number indicated by red strings
- The launch site has access to the proximities' within 1.5 km which would be close.
- The proximities include the coast, highway, and station.




Section 4
Build a Dashboard with Plotly Dash

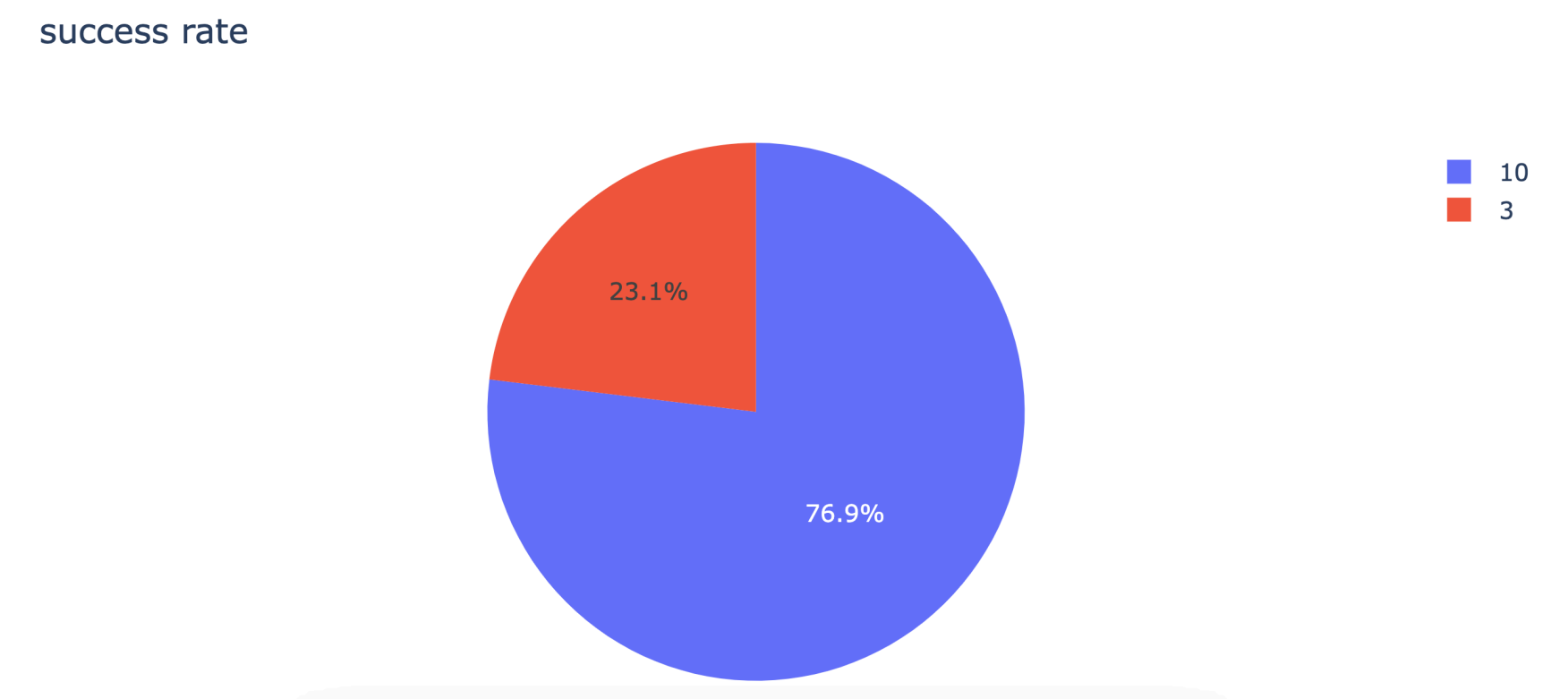
Success count of all launch sites

- KSL LC039A has the largest count of successful landing outcome



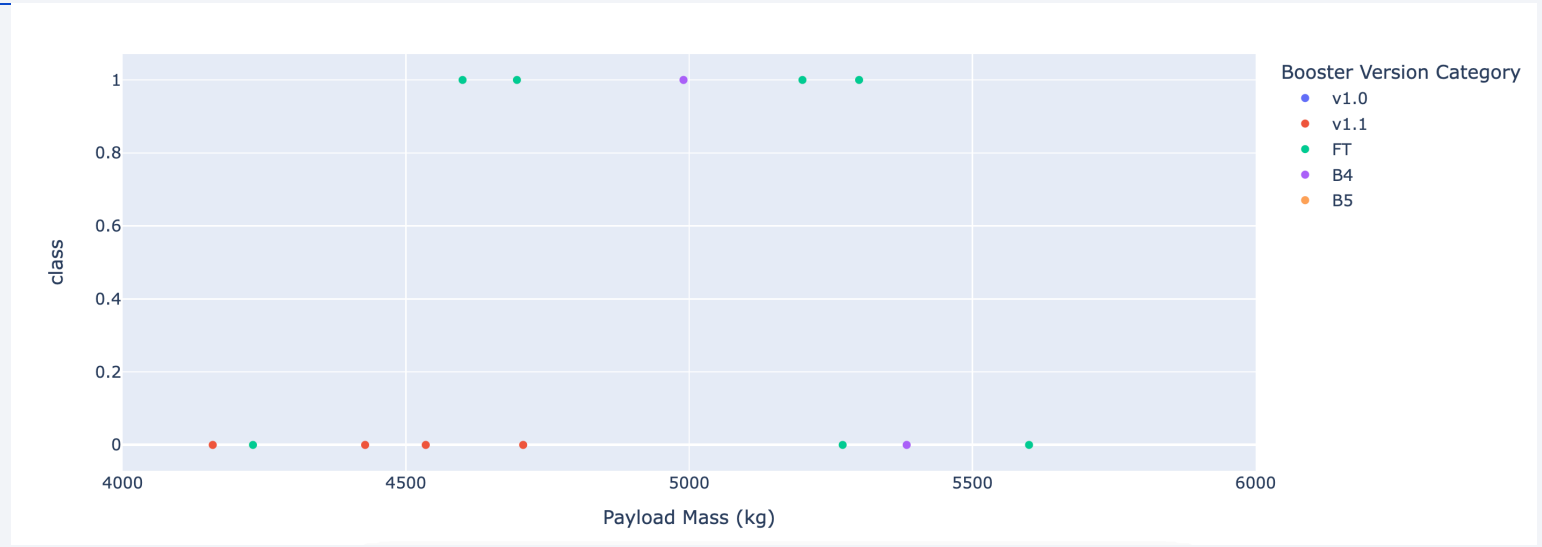
Success ratio of KSC LC-39A

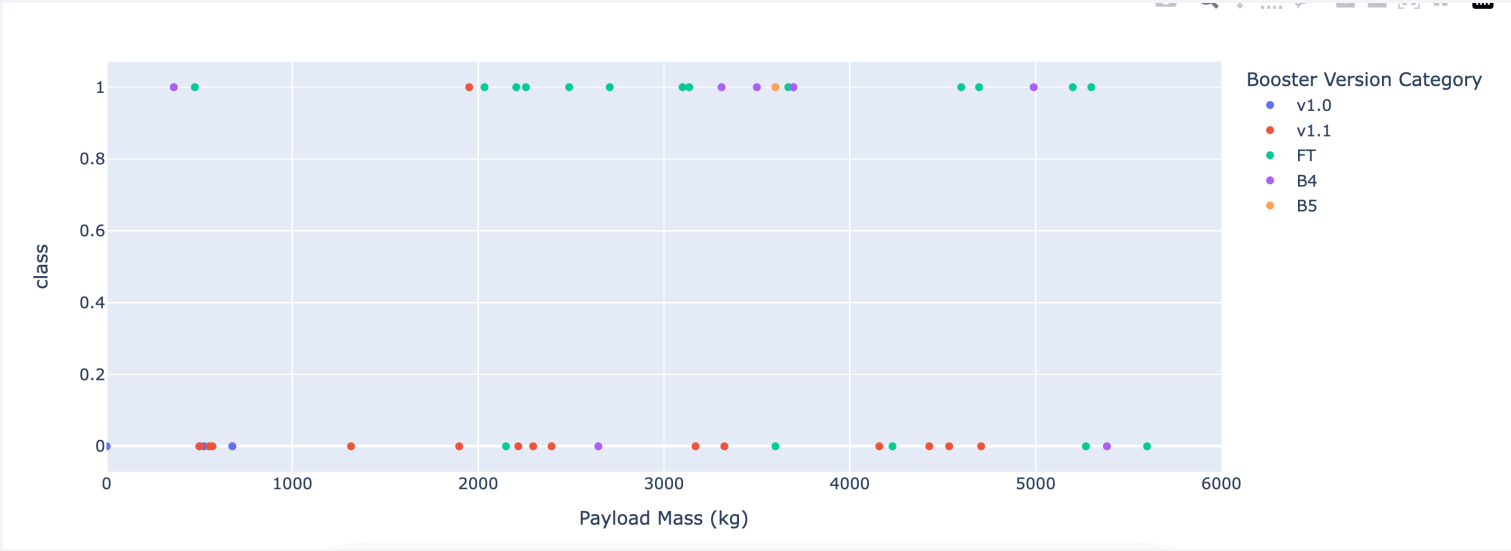
- KSC LC-39A has the highest success ratio.



<Dashboard Screenshot 3>

- Two graphs with the different payload mass scale
- When payload mass is between 4000 and 6000, success ratio is the highest



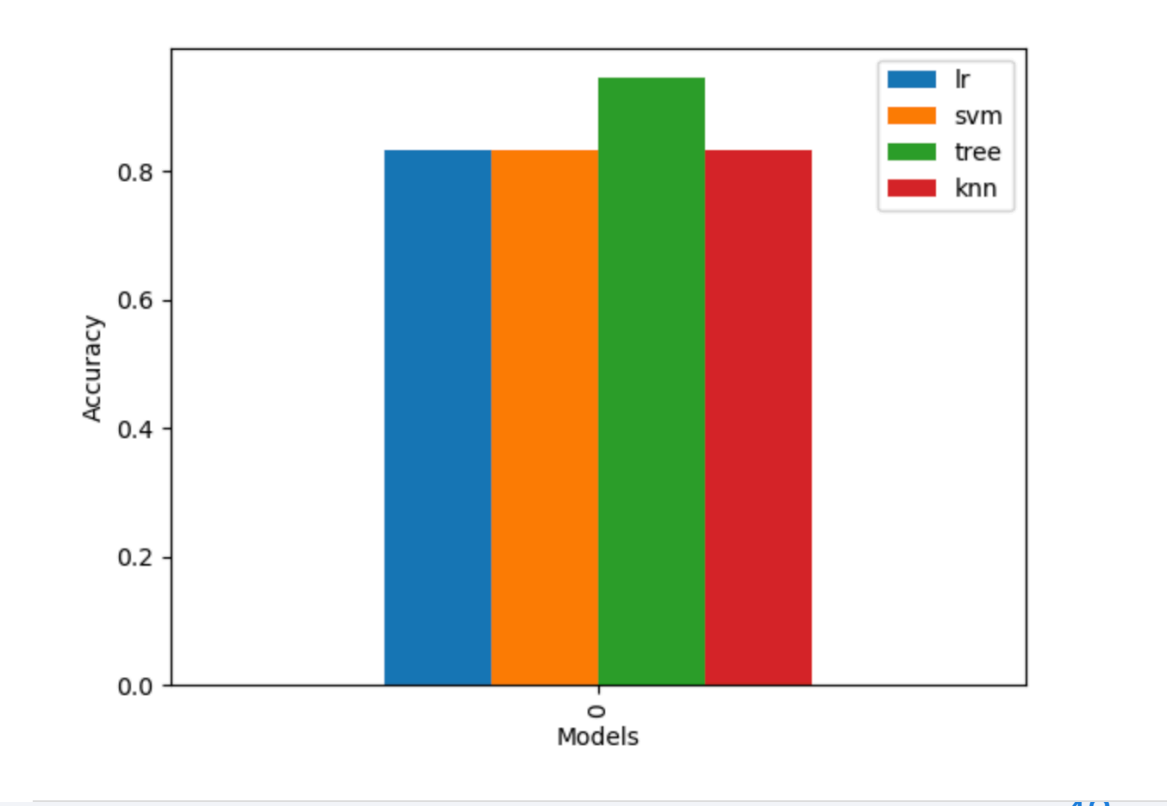


Section 5

Predictive Analysis (Classification)

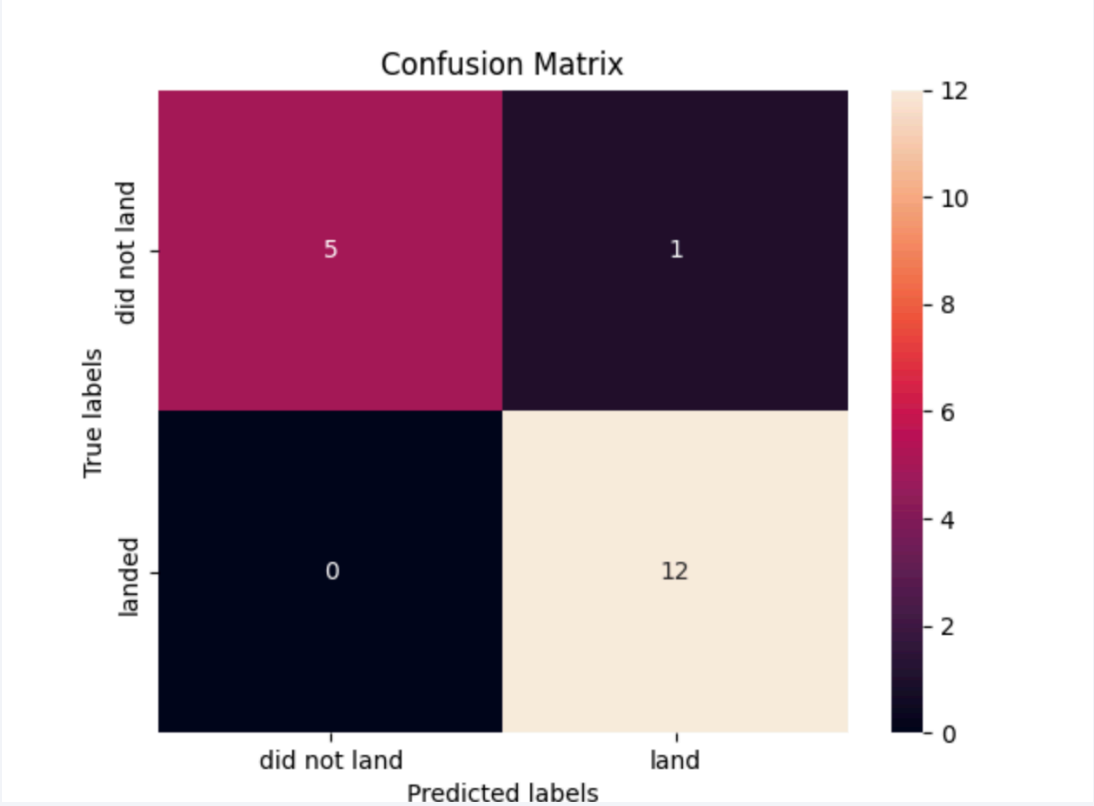
Classification Accuracy

- Visualization of the built model accuracy for all built classification models
- Tree model has the highest classification accuracy of 94.4%



Confusion Matrix

- the confusion matrix of the best performing model (Tree)
- It is a 2 ✖ 2 matrix
- Ex: the top-left cell indicates “the count when actual outcome is unsuccessful and predicted to be unsuccessful”



Conclusions

- Space X’s cost competency relies on successful landings of the first stage
- Data collected from Space X Rest API and Wikipedia
- 4 different classification models were trained and Tree model shows the highest accuracy

Thank you!