Winning Space Race with Data Science

Pham Hong Duc
21 Aug 2022

Outline

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

Executive Summary

- What factors affect the success of the reproduction of a rocket booster. If we can analyze the influence of factors: launch pad, trajectory, weight, rocket type... on the probability of success, we can improve the booster reuse rate. . This paper uses SVM, Classification Trees and Logistic Regression to evaluate the accuracy of the predictive model.
- As a result, I found the model was able to predict with 83.3% accuracy.

Introduction

The commercial space age is here, companies are making space travel affordable for everyone. Virgin Galactic is providing suborbital spaceflights. Rocket Lab is a small satellite provider. Blue Origin manufactures sub-orbital and orbital reusable rockets. Perhaps the most successful is SpaceX. SpaceX's accomplishments include: Sending spacecraft to the International Space Station. Starlink, a satellite internet constellation providing satellite Internet access. Sending manned missions to Space. One reason SpaceX can do this is the rocket launches are relatively inexpensive. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.

Therefore, if we can determine if the first stage will land, we can determine the cost of a launch.

Section 1

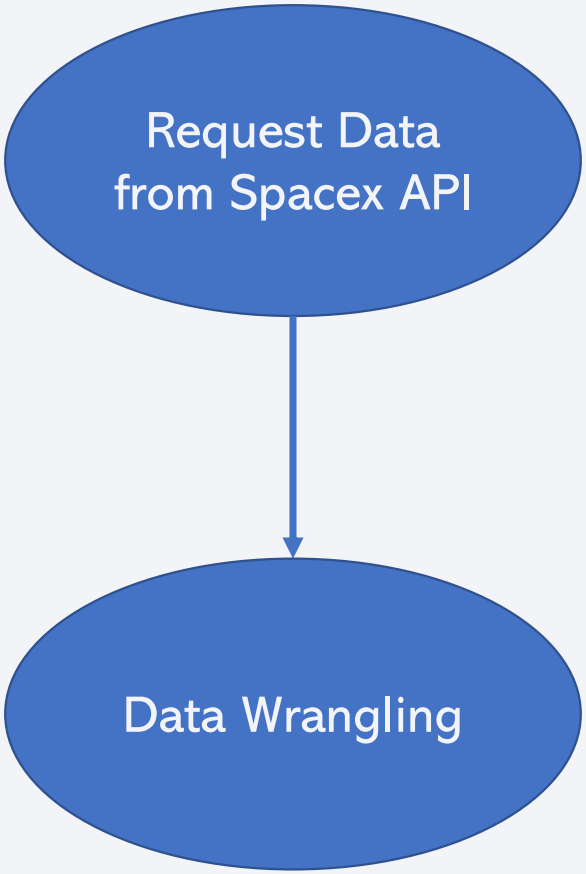
Methodology

Methodology

Executive Summary

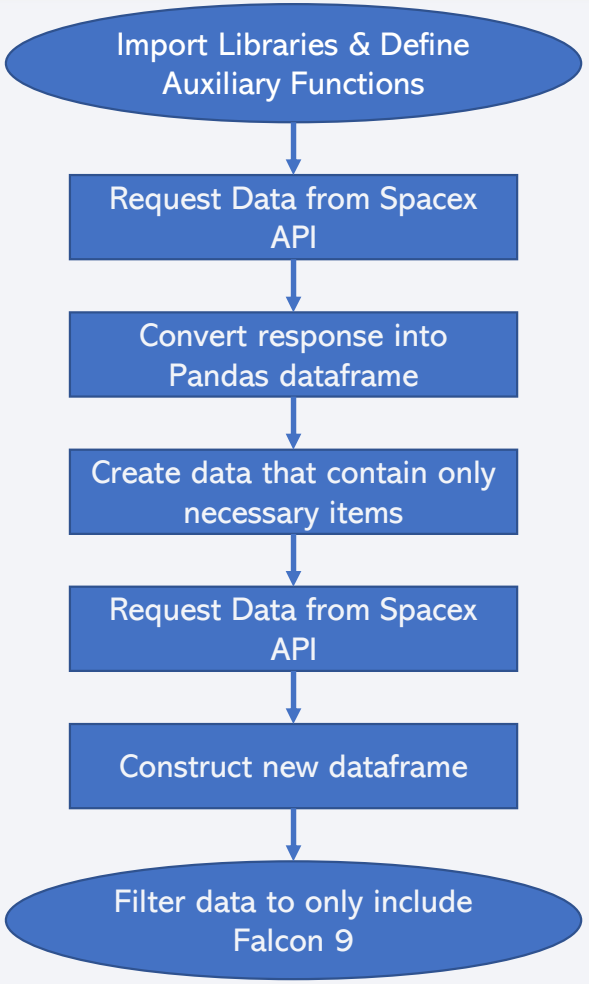
- Data collection methodology:
  - Make a get request to the SpaceX API
- Perform data wrangling
  - Deal with missing values, create label
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Split the data, create a logistic regression, support vector machine, decision tree classifier, k nearest neighbors then calculate the accuracy on the test data using the method score.

Data Collection



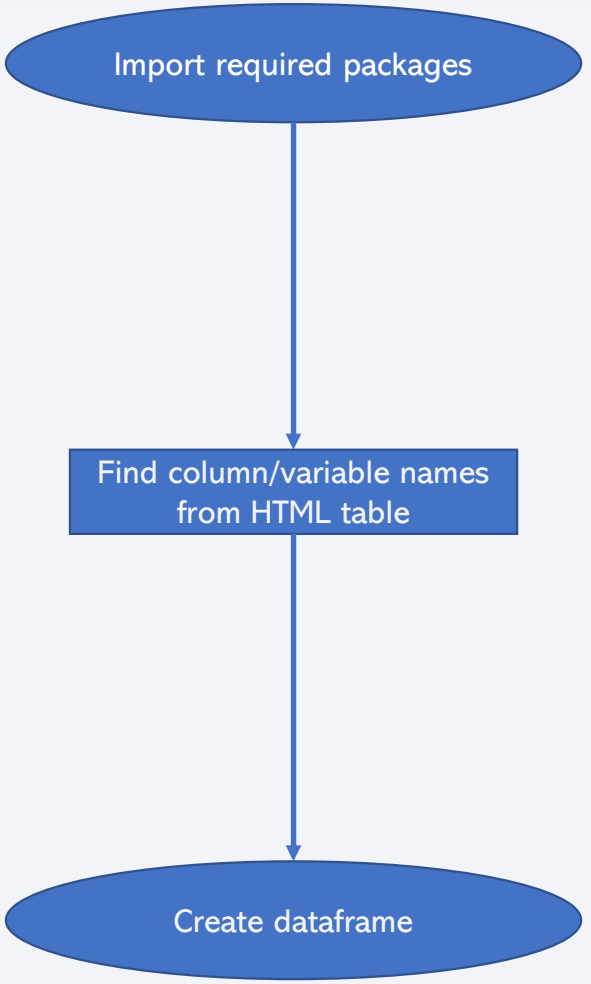
Data Collection – SpaceX API

- GitHub URL: https://github.com/PhamDuc1710/Final-assigment/blob/main/jupyter-labs-spacex-data-collection-api%20(1).ipynb



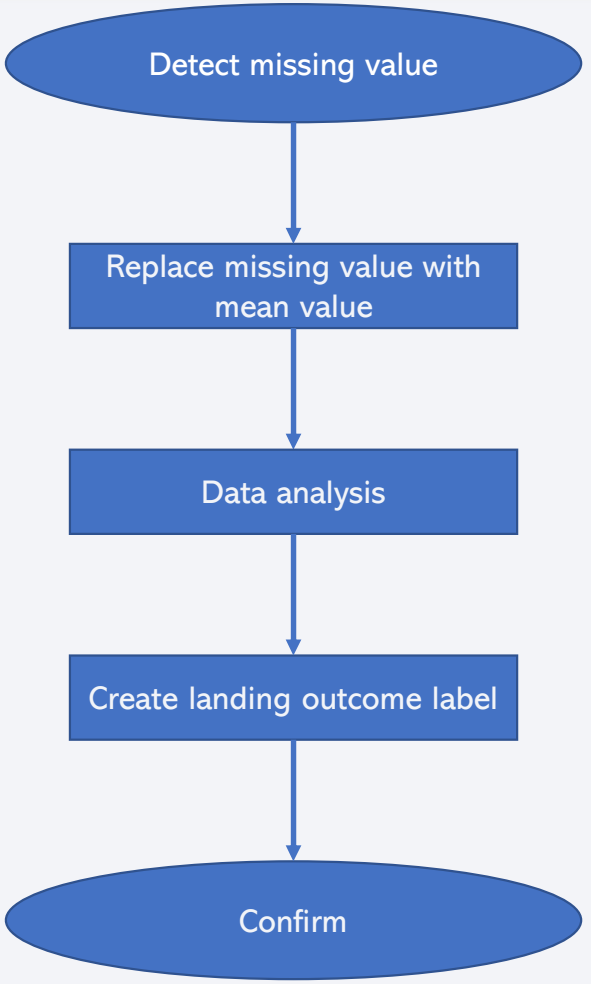
Data Collection - Scraping

- GitHub URL: https://github.com/PhamDuc1710/Final-assigment/blob/main/jupyter-labs-webscraping%20(1).ipynb



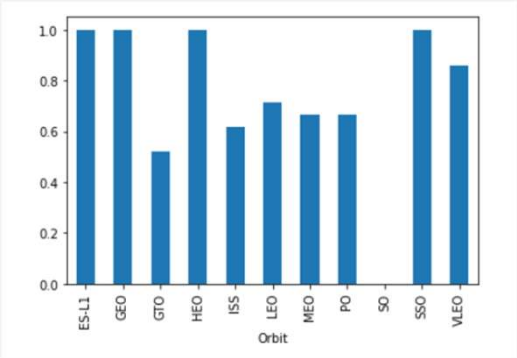
Data Wrangling

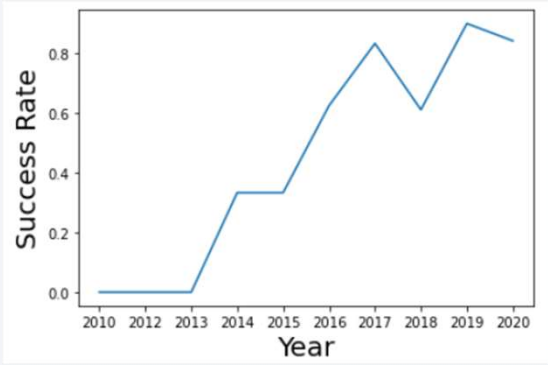
- GitHub URL: https://github.com/PhamDuc1710/Final-assigment/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

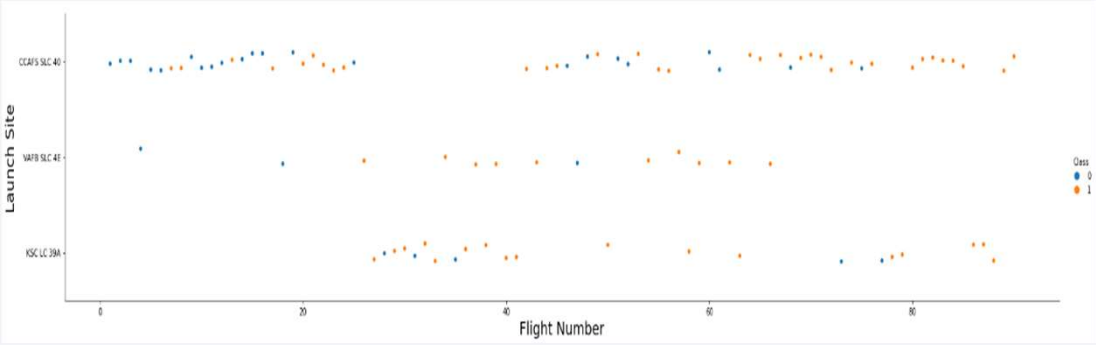


EDA with Data Visualization

- Use scatter plot to study the relationship between factors
- Use bar chart to find which orbit have high success rate
- Use line chart to study success yearly trend
- GitHub URL: https://github.com/PhamDuc1710/Final-assigment/blob/main/jupyter-labs-eda-dataviz.ipynb







EDA with SQL

- %sql sqlite:///my_data1.db
- %sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL
- %sql SELECT LAUNCH_SITE FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
- %sql SELECT TOTAL(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer = 'NASA (CRS)'
- %sql SELECT Avg(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version like 'F9 v1.1'
- %sql SELECT MIN(Date) FROM SPACEXTBL WHERE "Landing _Outcome" = 'Success (ground pad)'
- %sql SELECT "Booster_Version", "PAYLOAD_MASS__KG_" FROM SPACEXTBL WHERE ("Mission_Outcome" like 'Success')and ("PAYLOAD_MASS__KG_" > '4000')and ("PAYLOAD_MASS__KG_" < '6000') order by "PAYLOAD_MASS__KG_"
- %sql SELECT "Mission_Outcome", count("Mission_Outcome") FROM SPACEXTBL group by "Mission_Outcome"
- %sql SELECT "Booster_Version", "PAYLOAD_MASS__KG_" FROM SPACEXTBL WHERE "PAYLOAD_MASS__KG_" = (select MAX("PAYLOAD_MASS__KG_")FROM SPACEXTBL)
- %sql select Date, "Booster_Version", "Launch_Site" from SPACEXTBL WHERE "Landing _Outcome" like "Fail%" and substr(Date,7,4) like "2015%"
- %%sql SELECT "DATE","Landing _Outcome",count("Landing _Outcome")as LANDING_OUTCOME_COUNT,DATE
  - from SPACEXTBL where substr(Date,7,4) || substr(Date,4,2) || substr(Date,1,2) between '20100604' and '20170320' group by "Landing _Outcome" order by count("Landing _Outcome") desc
- GitHub URL: https://github.com/PhamDuc1710/Final-assigment/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

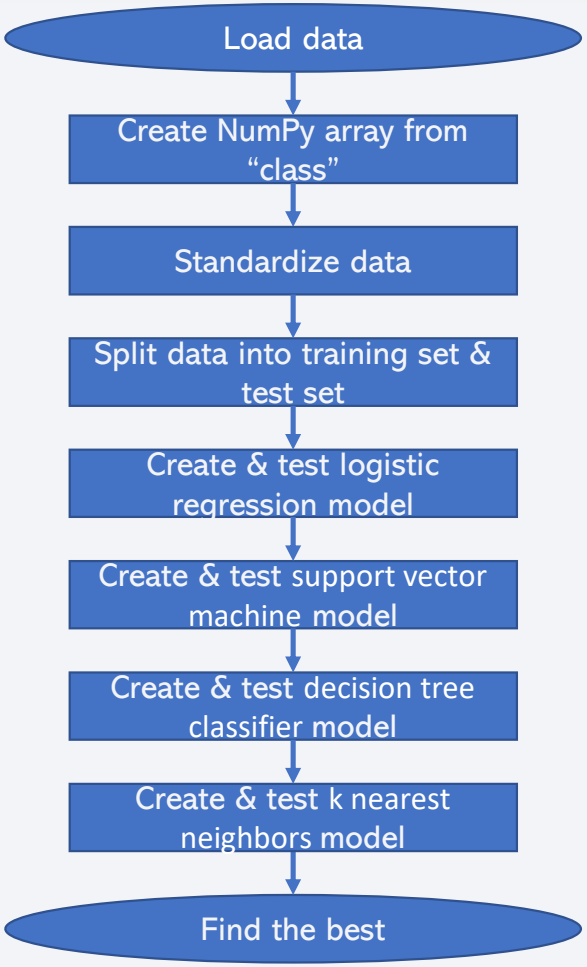
- Add to the map: circle, marker, marker_cluster, mouse_position, distance_marker, line
- Add marker objects to map to highlight positions of location on map, add line and distance marker to highlight the distance between locations.
- GitHub URL: https://github.com/PhamDuc1710/Final-assigment/blob/main/lab_jupyter_launch_site_location%20(1).ipynb

Build a Dashboard with Plotly Dash

- Add launch site drop-down: to select launch site
- Add success pie-chart: to study success rate
- Add range slider to select payload weight
- Add success payload scatter chart to study relationship between payload mass – success
- GitHub URL: https://github.com/PhamDuc1710/Final-assigment/blob/main/spacex_dash_app.py

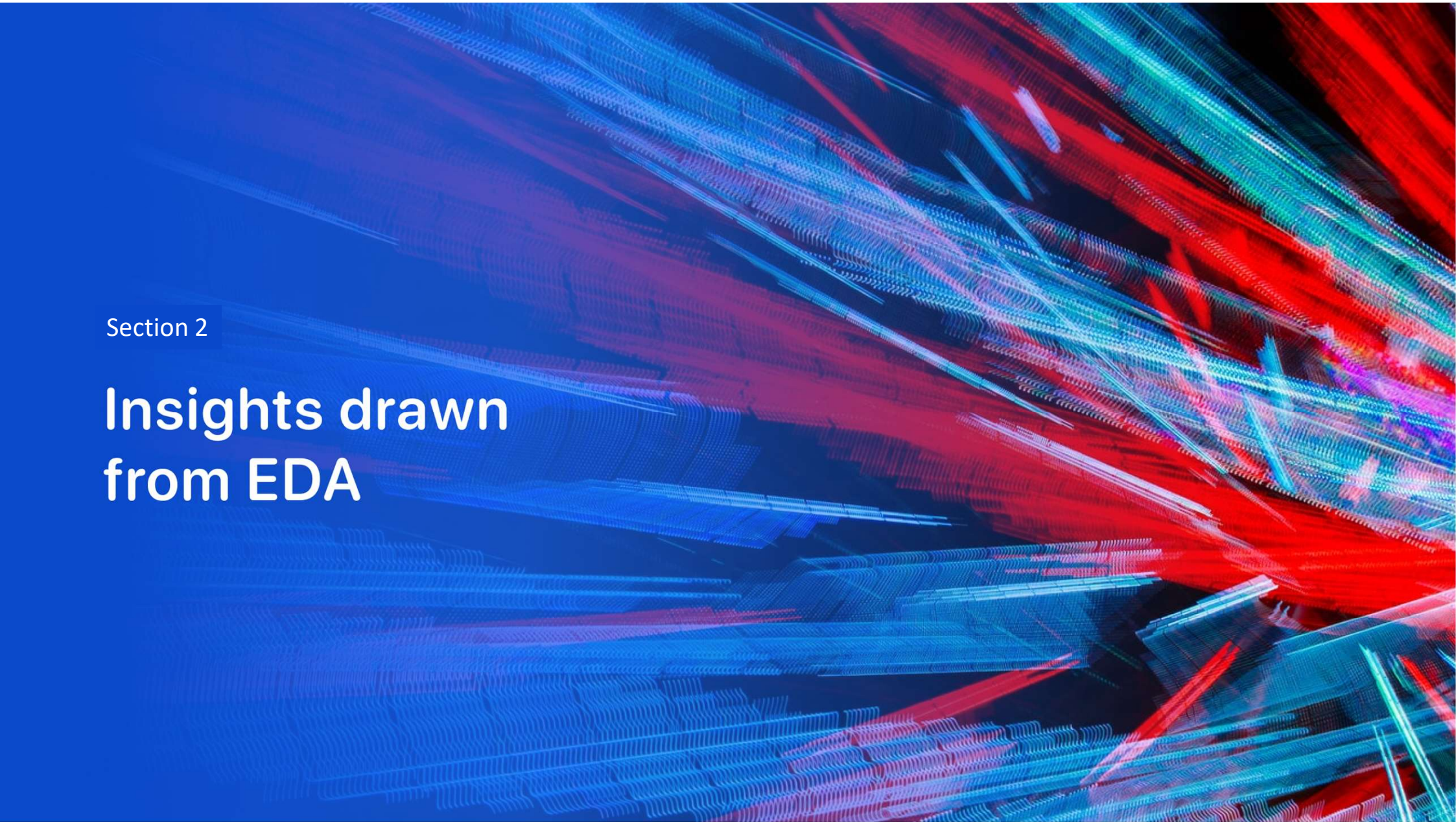
Predictive Analysis (Classification)

- GitHub URL: https://github.com/PhamDuc1710/Final-assigment/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

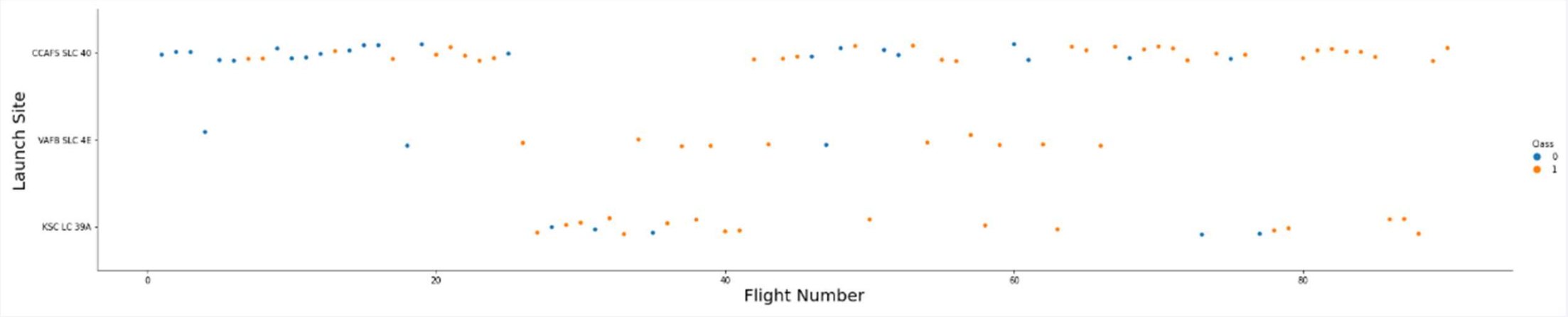


Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

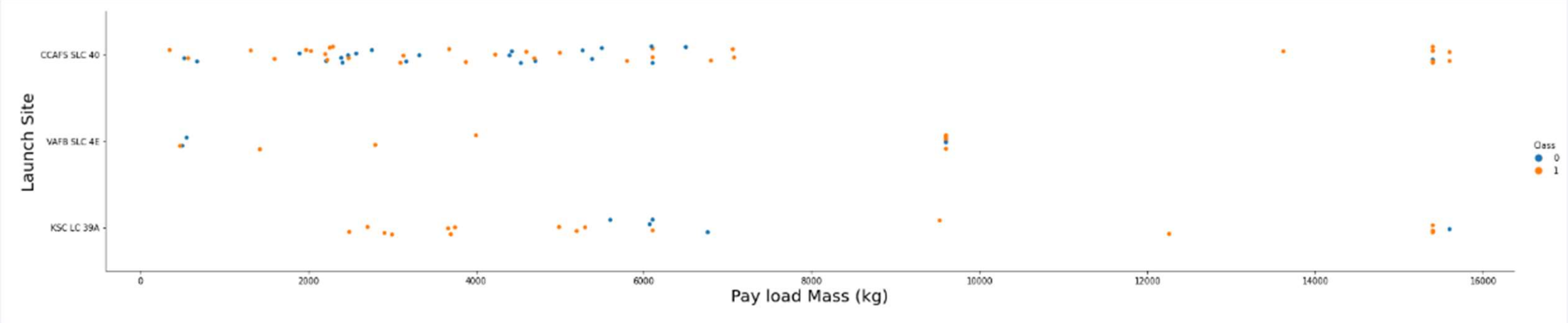

Section 2
Insights drawn
from EDA

Flight Number vs. Launch Site



From scatter plot we found that VAFB SLC 4E and KSC LC-39A have highest success rate. The success rate is low at the beginning (when flight number is small), then become better recently (when flight number > 40).

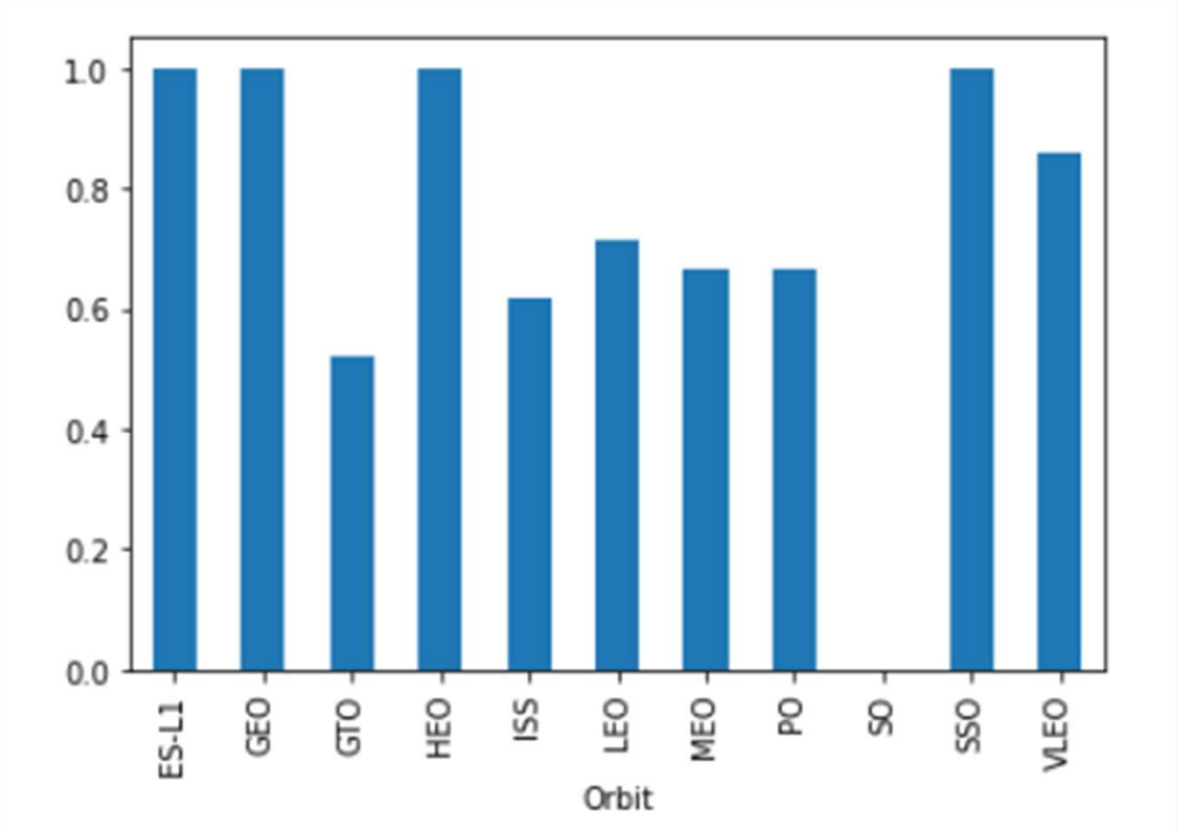
Payload vs. Launch Site



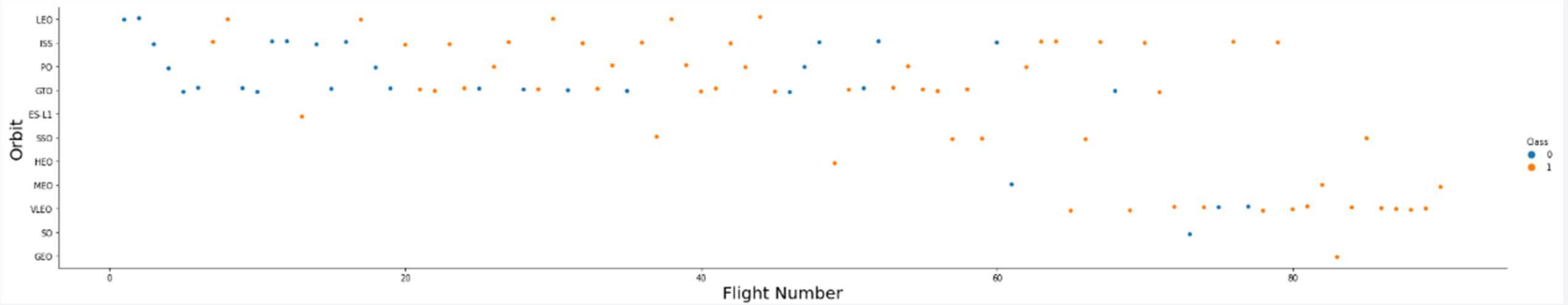
- From scatter plot we found that:
  - VAFB SLC 4E have high success rate with payload mass from 2000 to 10000.
  - KSC LC-39A have high success rate with payload mass from 2000 to 5000 and from 8000 to 15000.
  - CCAFS LC-40 have high success rate with payload mass from 7000 to 16000.

Success Rate vs. Orbit Type

- From bar chart we found that:
  - Orbit ES.L1, GEO, HEO, SSO have highest success rate
  - GTO, SO have lowest success rate

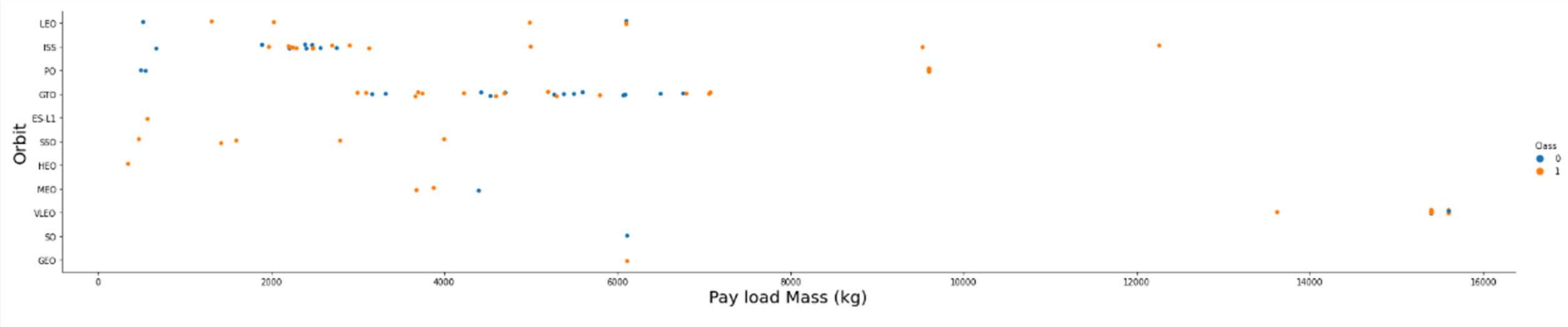


Flight Number vs. Orbit Type



- From scatter plot we found that:
  - LEO orbit: flight number increase ➔ success rate increase
  - ISS, PO, GTO, MEO orbit: no relationship between Flight number and Orbit

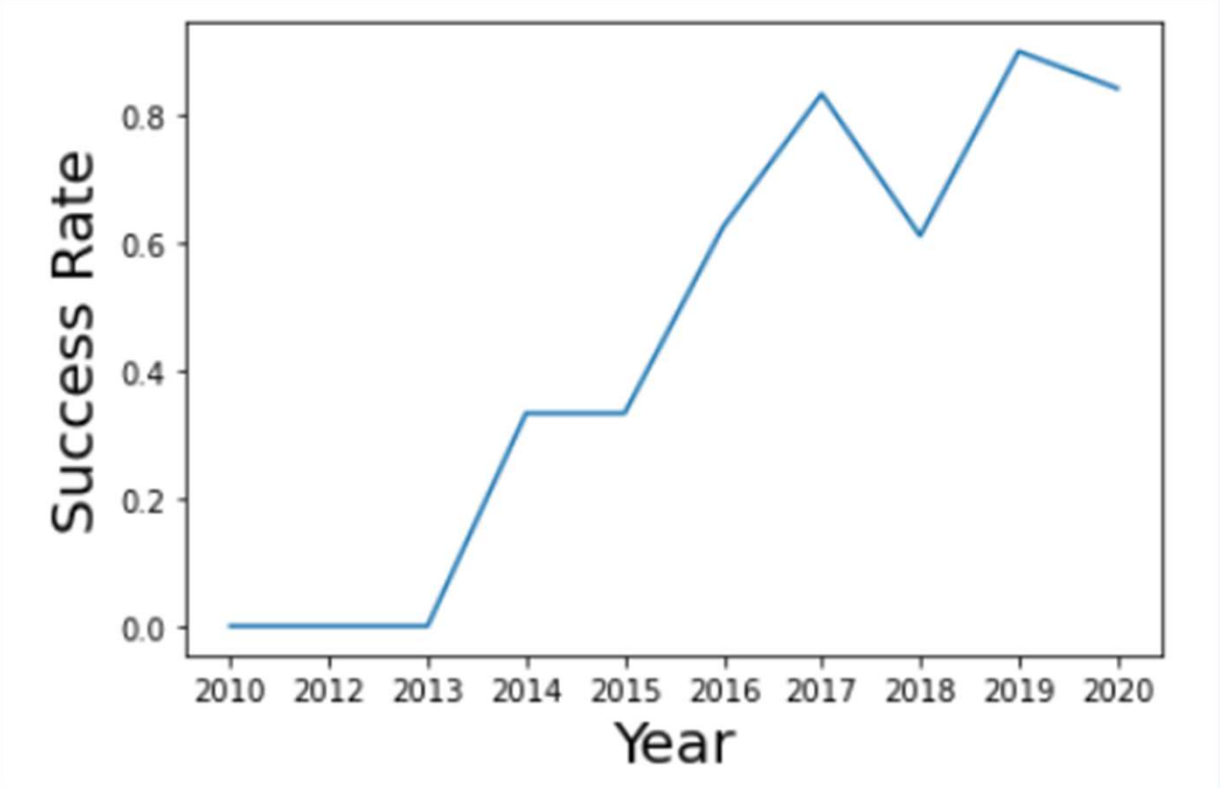
Payload vs. Orbit Type



- From scatter plot we found that:
  - ISS, PO, VLEO have high success rate with heavy payload
  - ES.L1, SSO, HEO, MFO have high success rate with small payload

Launch Success Yearly Trend

- We found that success rate increase from 2013.



All Launch Site Names

- Names of the 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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`

```
%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

- Total payload carried by boosters from NASA

```
%sql SELECT TOTAL(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer = 'NASA (CRS)'
```

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

TOTAL(PAYLOAD_MASS__KG_)
45596.0

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'
```

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

Avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

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

```
%sql SELECT Date FROM SPACEXTBL WHERE "Landing _Outcome" like 'Success%'
```

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

Date
22-12-2015

Successful Drone Ship Landing with Payload between 4000 and 6000

- 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", "PAYLOAD_MASS__KG_" FROM SPACEXTBL WHERE ("Mission_Outcome" like 'Success')and ("PAYLOAD_MASS__KG_" > '4000')and ("PAYLOAD_MASS__KG_" < '6000') order by "PAYLOAD_MASS__KG_"
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 v1.1 B1014	4159
F9 B5 B1051.2	4200
F9 FT B1032.2	4230
F9 B5B1060.1	4311
F9 B5B1062.1	4311
F9 v1.1 B1011	4428
F9 v1.1	4535
F9 FT B1026	4600
F9 FT B1022	4696
F9 v1.1 B1016	4707
F9 B5 B1048.3	4850
F9 B4 B1040.1	4990
F9 FT B1031.2	5200
F9 FT B1020	5271
F9 FT B1021.2	5300

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

```
%sql SELECT "Mission_Outcome", count("Mission_Outcome") FROM SPACEXTBL group by "Mission_Outcome"
```

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

Mission_Outcome	count("Mission_Outcome")
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

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

```
 * 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

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

```
%sql select Date, "Booster_Version", "Launch_Site" from SPACEXTBL WHERE "Landing _Outcome" like "Fail%" and substr(Date,7,4) like "2015%"
```

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

Date	Booster_Version	Launch_Site
10-01-2015	F9 v1.1 B1012	CCAFS LC-40
14-04-2015	F9 v1.1 B1015	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 "DATE","Landing _Outcome",count("Landing _Outcome")as LANDING_OUTCOME_COUNT,DATE
from SPACEXTBL where substr(Date,7,4) || substr(Date,4,2) || substr(Date,1,2) between '20100604'
and '20170320'
group by "Landing _Outcome" order by count("Landing _Outcome") desc
```

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

Date	Landing _Outcome	LANDING_OUTCOME_COUNT	Date_1
22-05-2012	No attempt	10	22-05-2012
08-04-2016	Success (drone ship)	5	08-04-2016
10-01-2015	Failure (drone ship)	5	10-01-2015
22-12-2015	Success (ground pad)	3	22-12-2015
18-04-2014	Controlled (ocean)	3	18-04-2014
29-09-2013	Uncontrolled (ocean)	2	29-09-2013
04-06-2010	Failure (parachute)	2	04-06-2010
28-06-2015	Precluded (drone ship)	1	28-06-2015

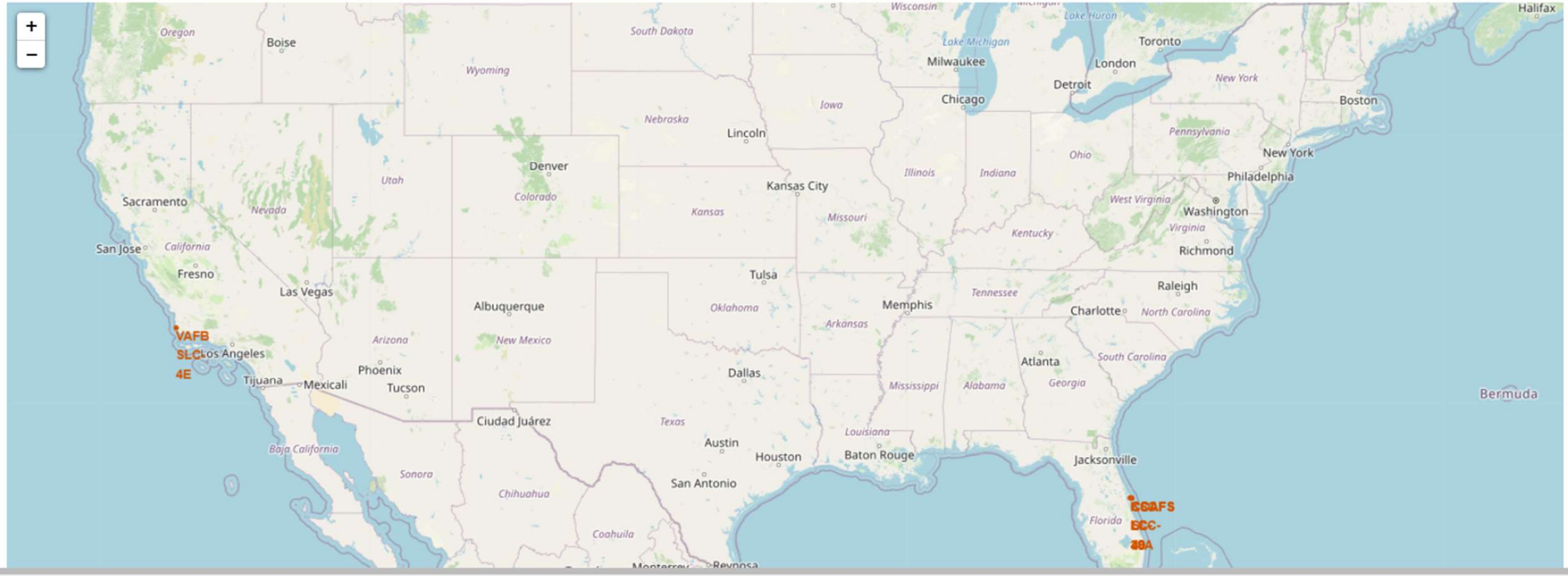

Section 3
Launch Sites
Proximities Analysis

Launch sites' location

- All launch sites placed near the coat

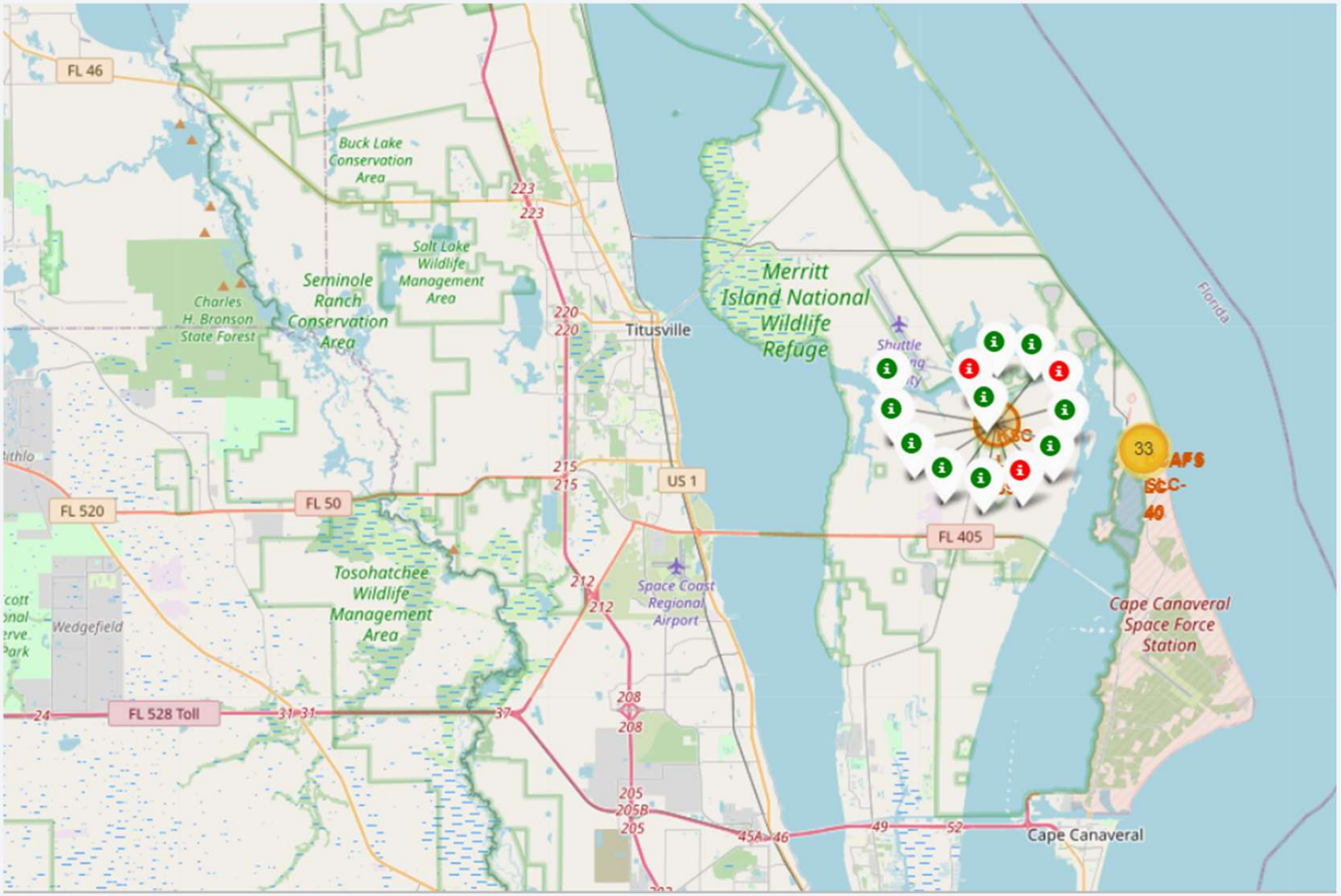
site_map

[8]:



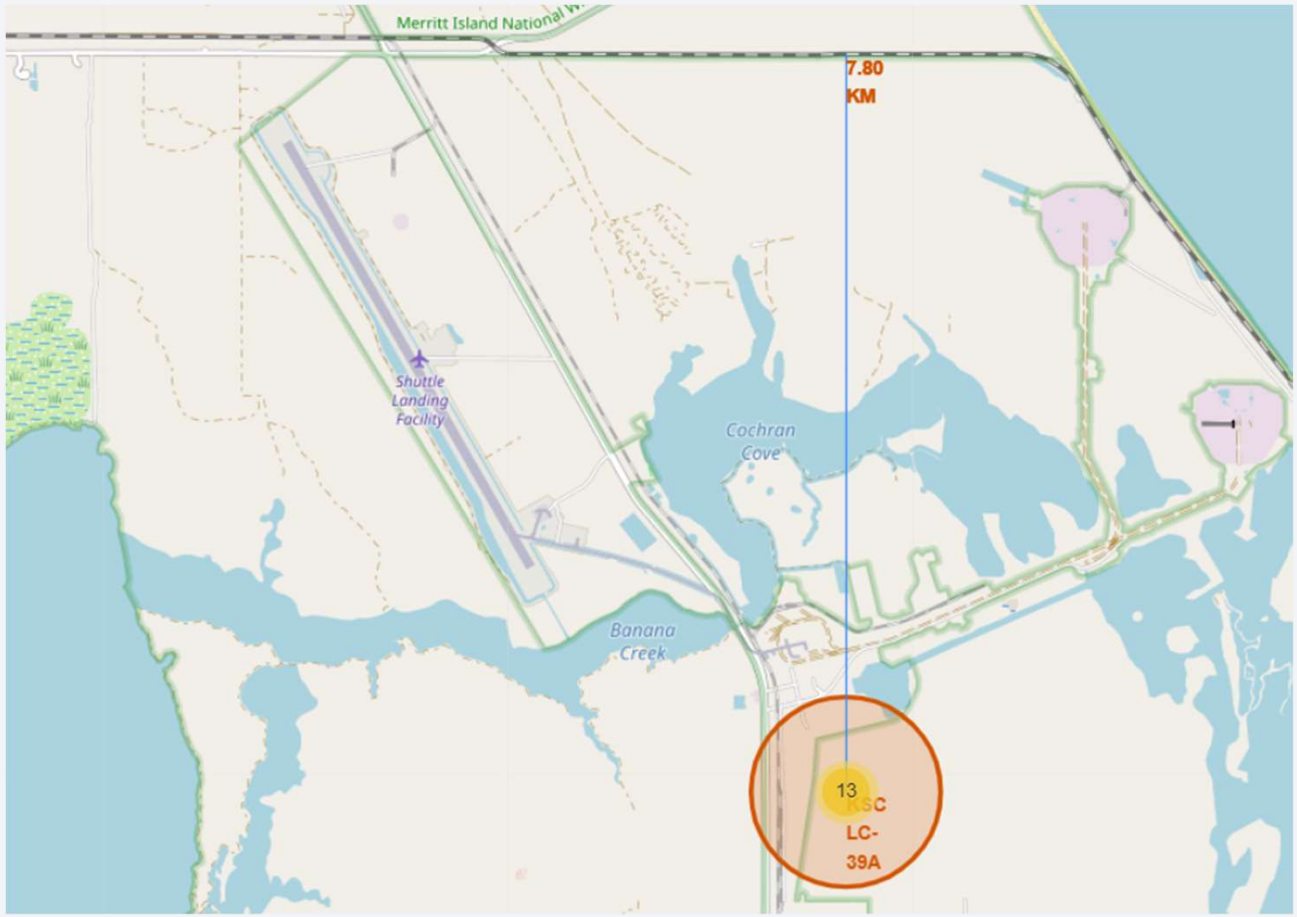
0 1 Fully initialized Python | Idle Mem: 433.07 / 6144.00 MB Mode: Command Ln 1, Col 1 English (United States) lab_jupyter_launch_site_loca

Color – labeled Map

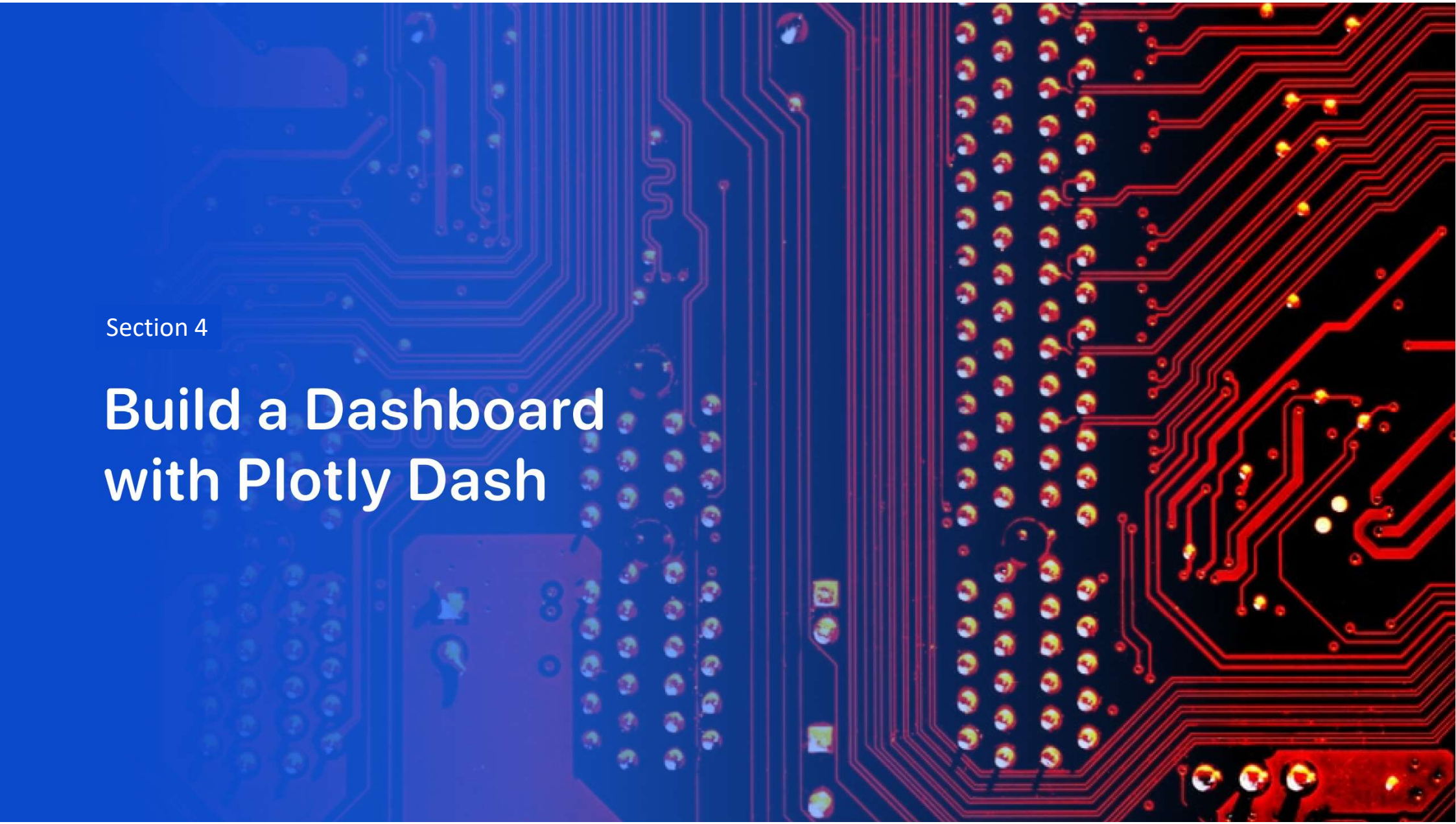


- From color – labeled map we found that KSC LC-39A Launch site has highest success rate

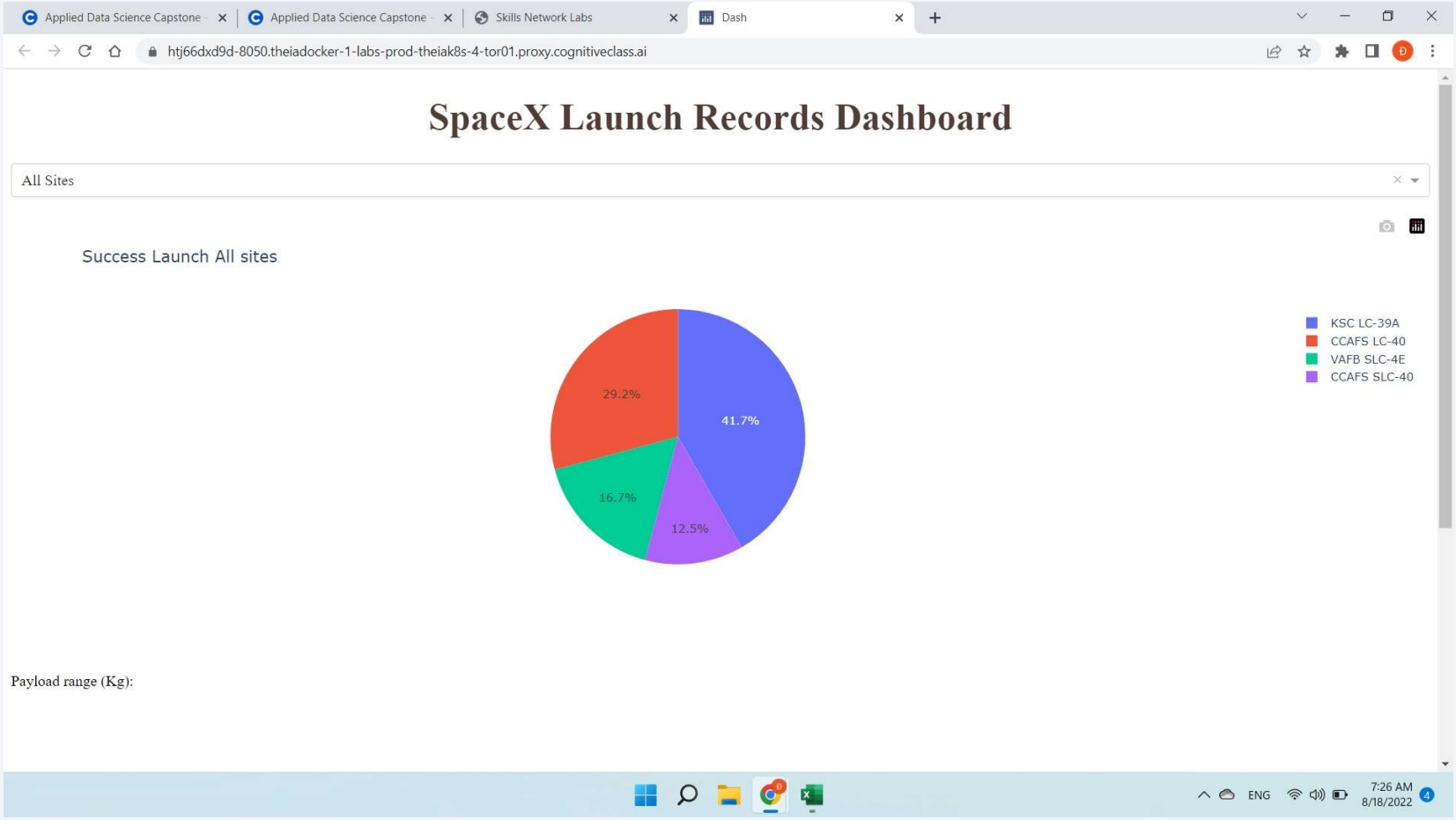
Distance display



- From map, we could find distance from KSC LC-39A to railway equal to 7.8 km

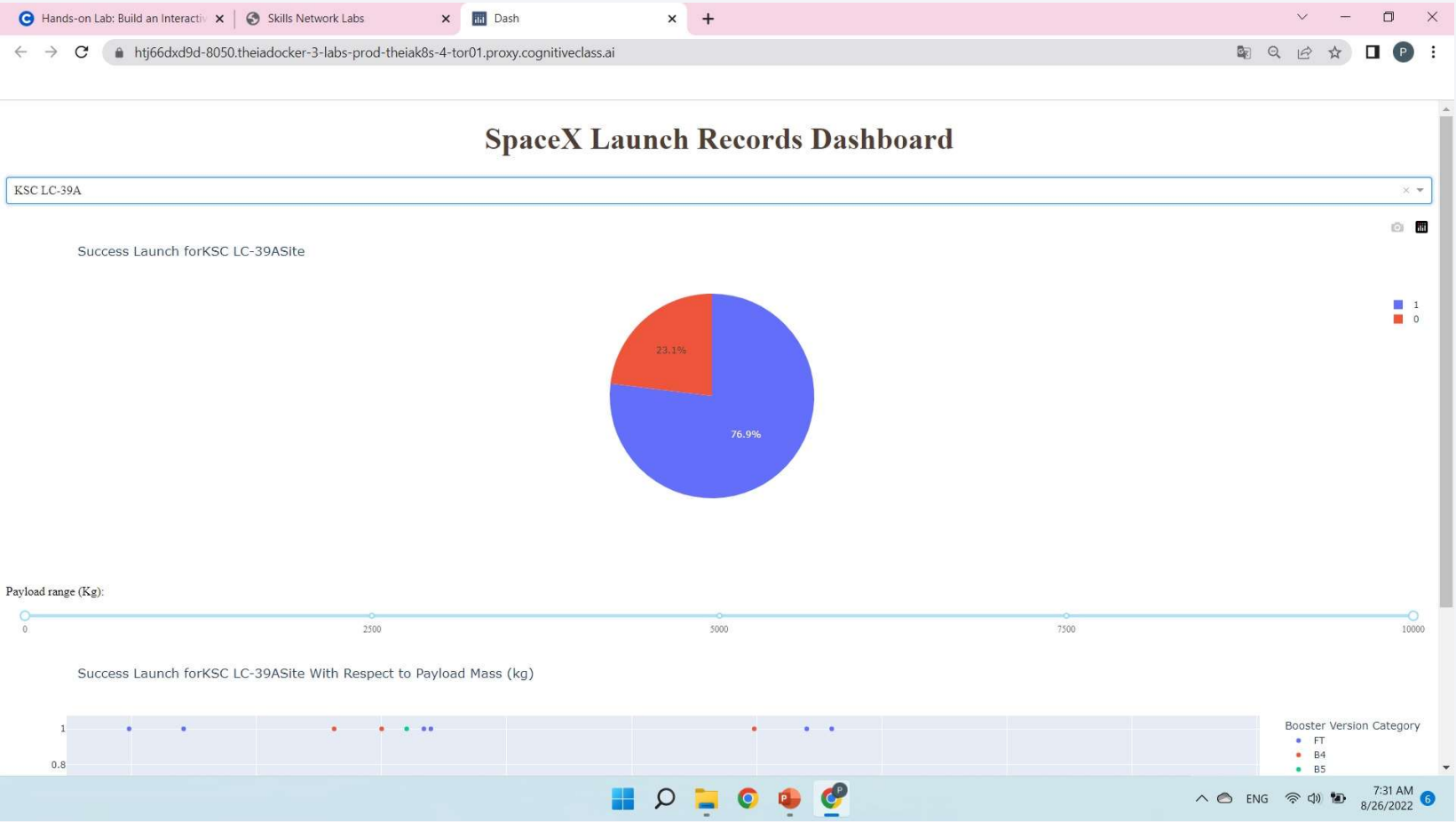

Section 4
Build a Dashboard with Plotly Dash

Pie-chart of success Launch



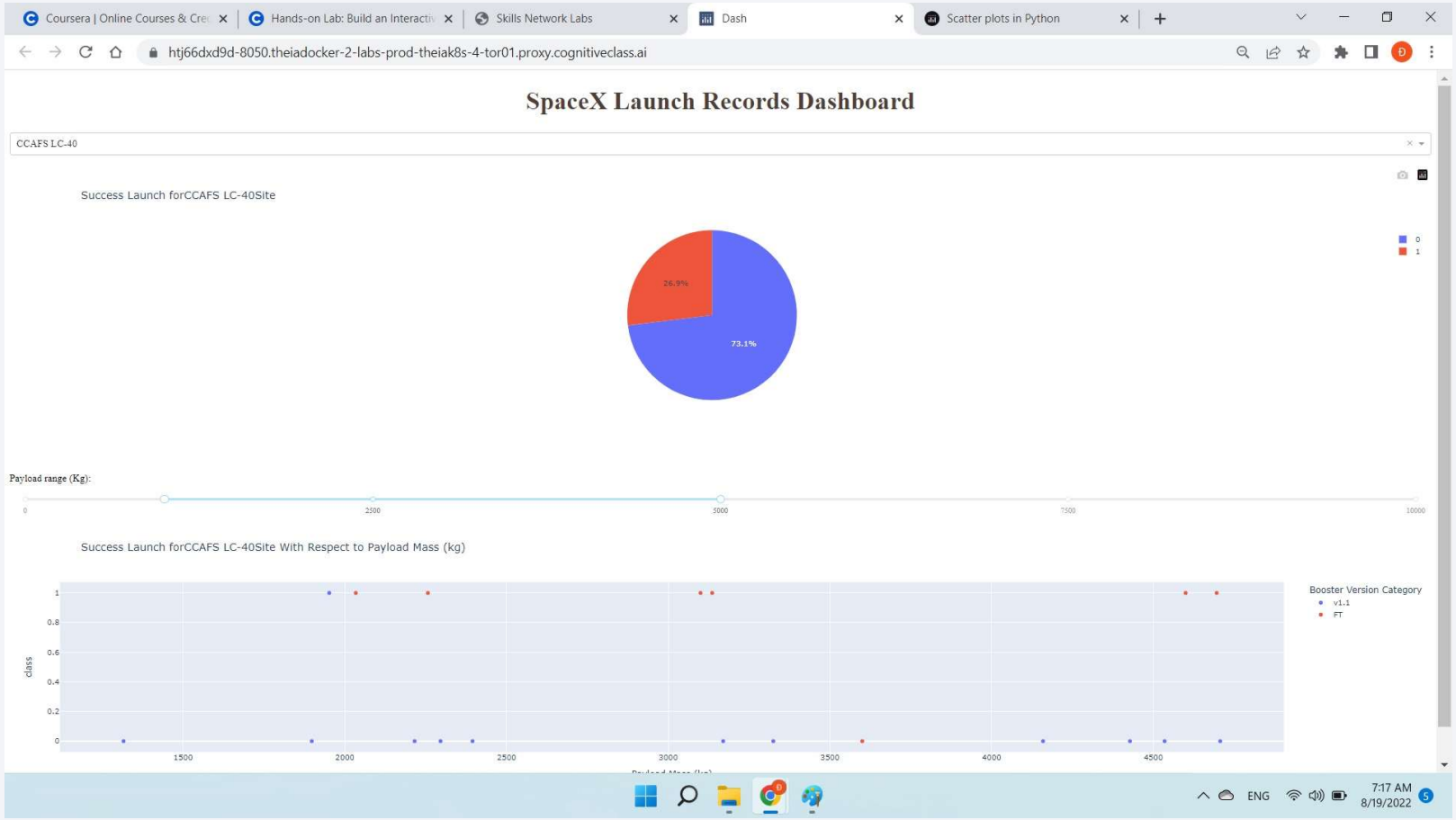
- From pie-chart we found KSC LC-39A have highest success rate

Pie-chart of KSC LC-39A



- We found that KSC LC-39A has success rate nearly 77%

Success scatter



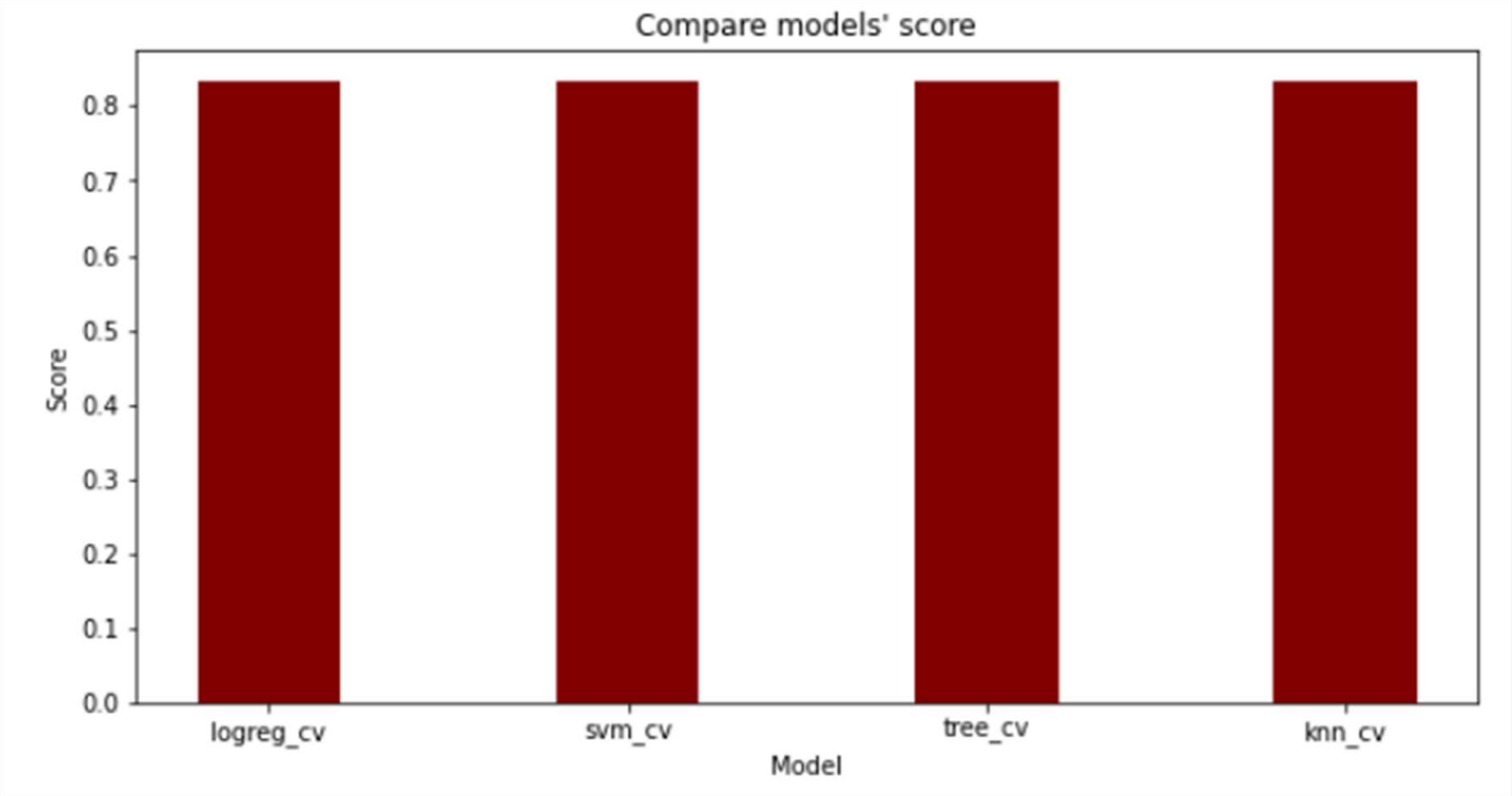
- We found relationship between payload mass and success rate of each launch site

Section 5

Predictive Analysis (Classification)

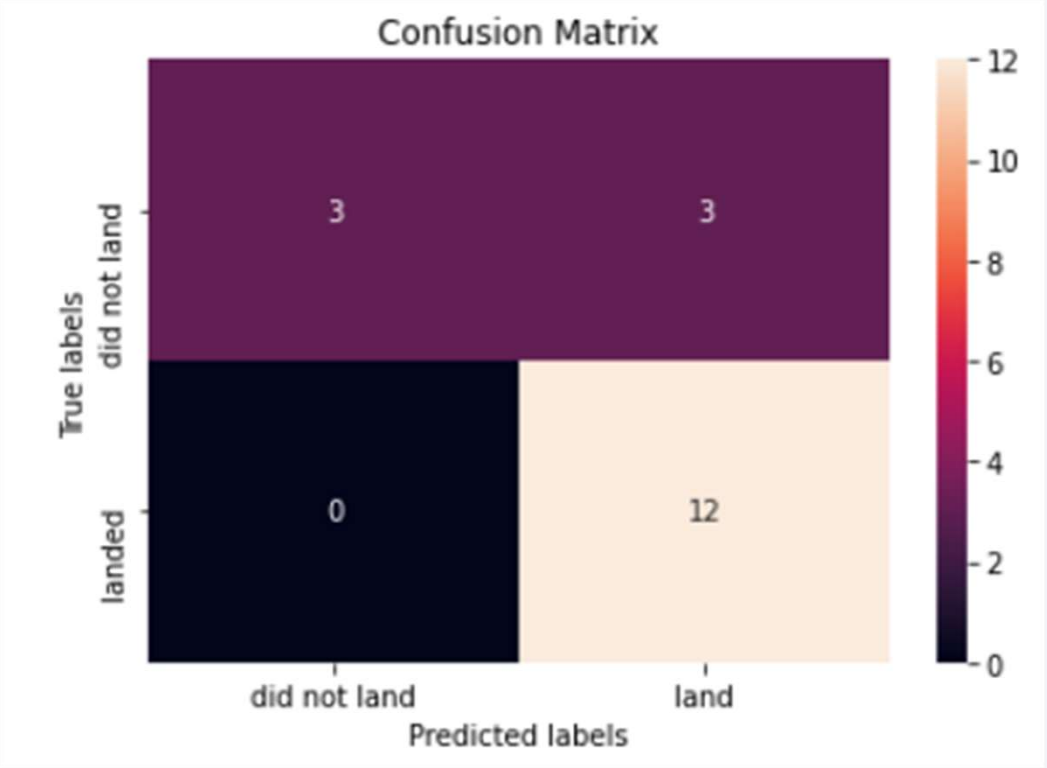
Classification Accuracy

- We found that all models have same score



Confusion Matrix

- Accuracy = (3 +12)/(3+3+0+12) =0.833333



Conclusions

- From this studying we could increase the success the first stage of rocket will land:
  - If the payload mass is heavy & Orbits are ISS, PO, VLEO ➔ we choose launch site KSC LC-39A & CCAFS LC-40
  - If the payload mass is light & Orbits are ES.L1, SSO, HEO, MFO ➔ we choose launch site KSC LC-39A & VAFB SLC 4E

Appendix

```
data = {'logreg_cv':logreg_cv.score(X_test, Y_test), 'svm_cv':svm_cv.score(X_test, Y_test), 'tree_cv':tree_cv.score(X_test, Y_test),
        'knn_cv':knn_cv.score(X_test, Y_test)}
courses = list(data.keys())
values = list(data.values())

fig = plt.figure(figsize = (10, 5))

# creating the bar plot
plt.bar(courses, values, color ='maroon',
        width = 0.4)

plt.xlabel("Model")
plt.ylabel("Score")
plt.title("Compare models' score")
plt.show()
```

Thank you!