IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

YingCong Chen

21-03-2023

Outline

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

Executive Summary

- **using methodologies in our project:**
- Data collecting
- Data Wrangling
- EDA
  - Feature Engineering by visualisation and SQL
  - interactive visual analytics using Folium and Plotly Dash
- Data prediction classification by Machine Learning
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree
  - K -Nearest Neigbours

- **All results**
- After 2013, technology has made great progression, the success reate has greatly improved.
- launch site should be close to coastal line or equator of th earth

Introduction

- Being a newly setup Space development company, we want to achieve our dream to launch space rocket.

- Problem :
- Can we have less cost to launch a rocket than that of SpaceX? or in other words, can we reuse the first stage ?

Section 1

Methodology

Methodology

Executive Summary

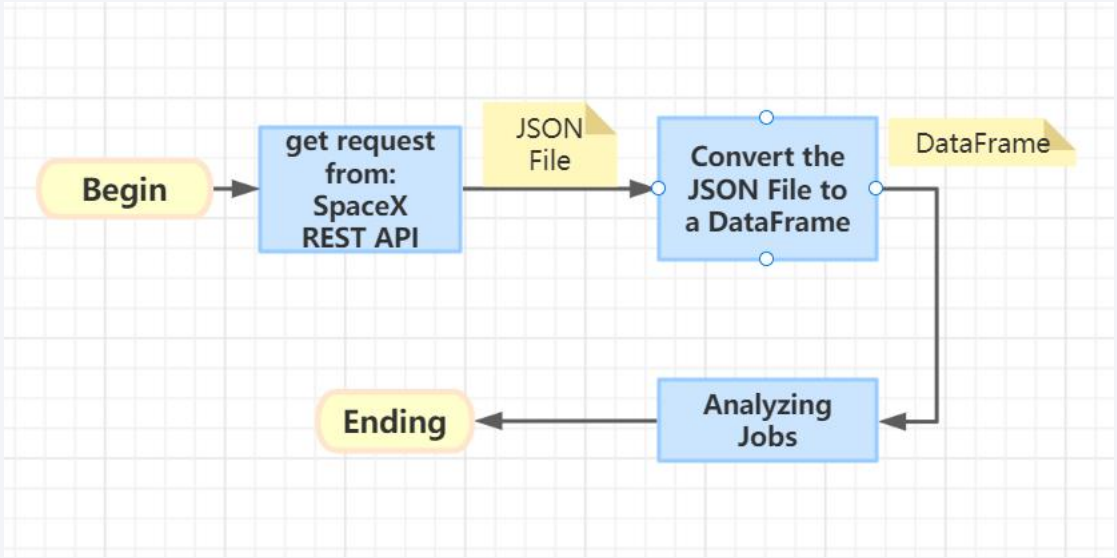
- Data collection methodology:
  - SpaceX REST calls
  - web scraping in WiKi.
- Perform data wrangling
  - standardized , dummy , and check the null value.
  - convert the diverse outcomes into simply numerical 0 or 1
- Perform exploratory data analysis (EDA) using visualization and SQL
  - using various visualisation methods, such as scatter, line, pie charts.
  - using the SQL to analize the trends.
- Perform interactive visual analytics using Folium and Plotly Dash
  - using Folium to analyze the geographic information
  - using Plotly Dash to interactively analyze the trends.
- Perform predictive analysis using classification models
  - using 4 different models, such as Logistic regression, KNN, Decision Tree and Support Vector Machine to predict the result.
  - calculating the accuracy, ie, compute the difference distance from the actal to the predicted value by the module

Data Collection

- Data Collection
  - SpaceX API calls
  - web scraping

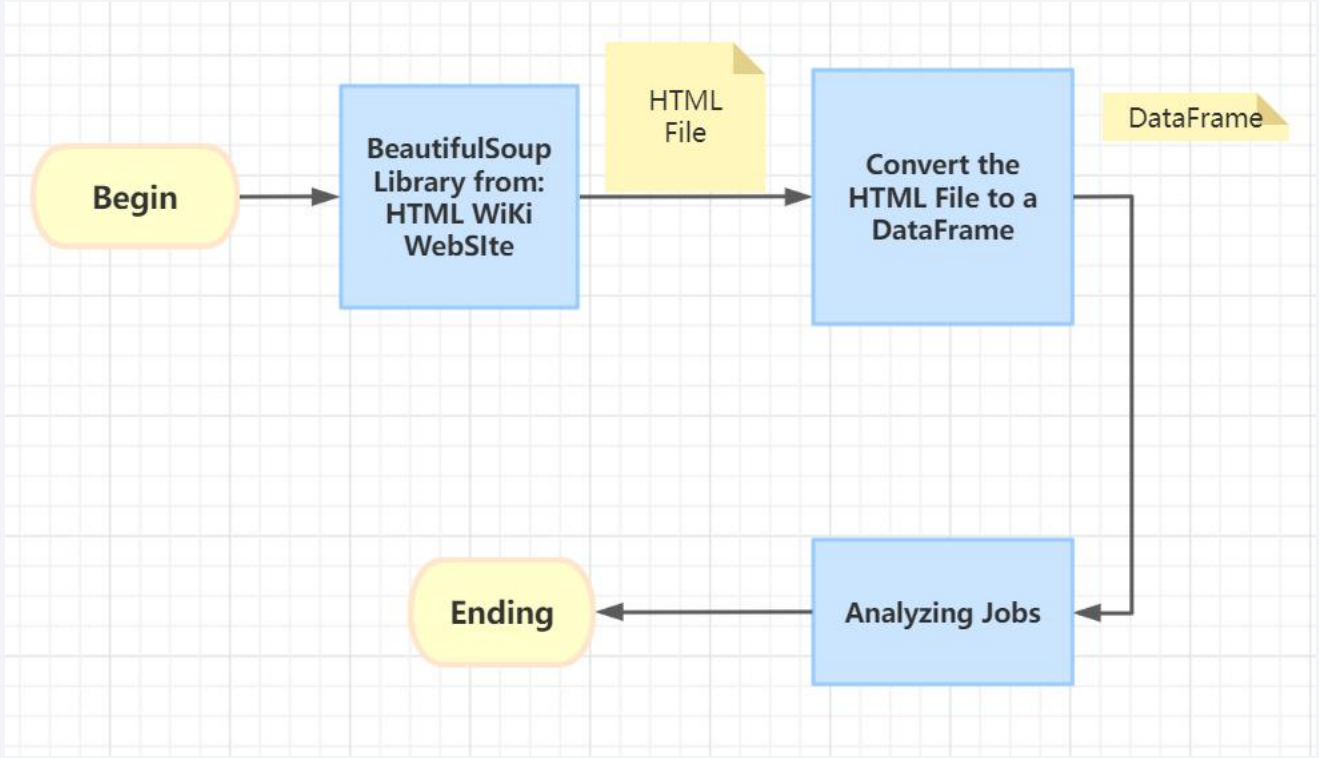
Data Collection – SpaceX API

- The GitHub URL of the completed SpaceX API calls notebook:
- https://github.com/abcgz133/CapStoneIBMDataScience/blob/master/jupyter-labs-spacex-data-collection-api.ipynb

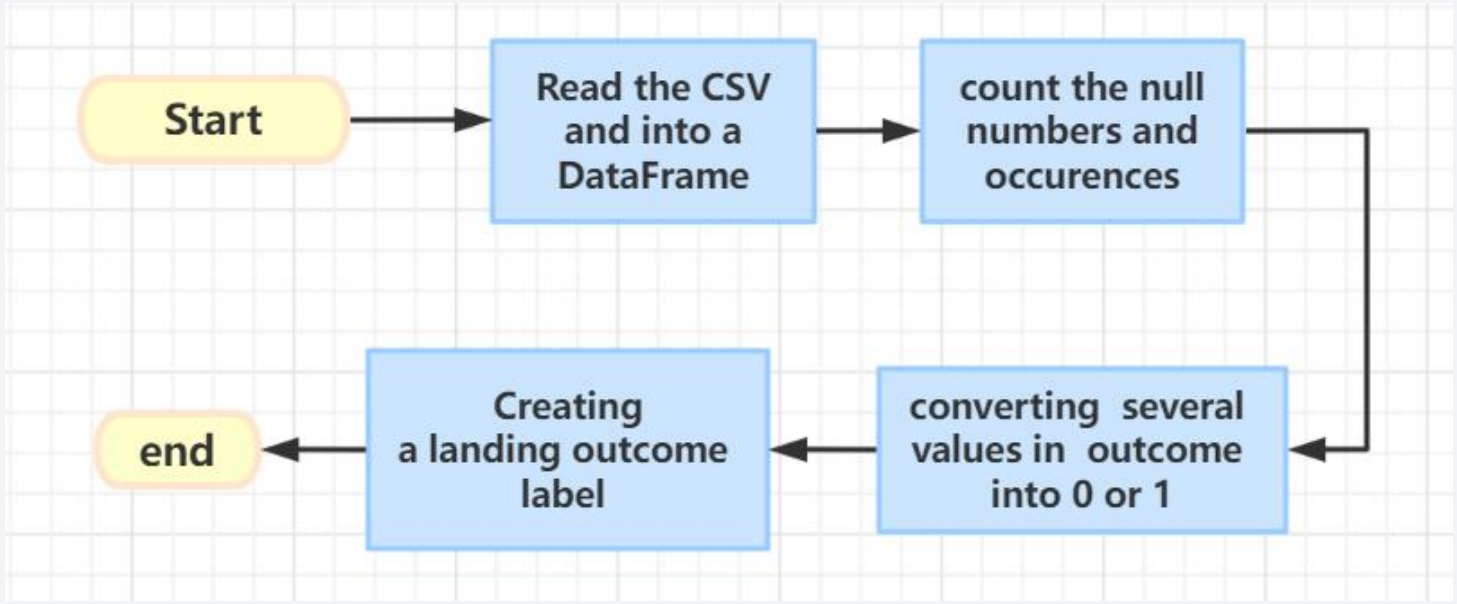


Data Collection - Scraping

- the GitHub URL of the completed web scraping notebook:
- https://github.com/abcgz133/CapStoneIBMDataScience/blob/master/jupyter-labs-webscraping.ipynb



Data Wrangling



- The GitHub URL of completed data wrangling related notebooks:
- https://github.com/abcgz133/CapStoneIBMDataScience/blob/master/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Here are the summary of the charts:
  - using the line chart to visualize the launch success yearly trend.
  - using the scatter point chart to visualize the relationship between features, such as Payload and Orbit type, such as Payload and Launch Site
  - using the bar chart to visualize the sucess rate of each orbit

- the GitHub URL of completed EDA with data visualization notebook:
- https://github.com/abcgz133/CapStoneIBMDataScience/blob/master/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

the SQL queries I kave performed:

- 1. Display the names of the unique launch sites
- 2. Display 5 records where launch sites begin with the string 'CCA'
- 3. Display the total payload mass carried by NASA (CRS)
- 4. Display average payload mass carried by version F9 v1.1
- 5. List the first succesful landing outcome in ground pad
- 6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- 7. List the total number of successful and failure mission outcomes
- 8. List the names of the booster_versions which have carried the maximum payload mass.
- 9. List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the year 2015.
- 10. rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20

- GitHub URL of completed EDA with SQL notebook:
- https://github.com/abcgz133/CapStoneIBMDataScience/blob/master/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Summarize the map objects created and added to the folium map:
  - markers- the distance object, the cluster objects, etc.
  - circles- the circle of launch sites.
  - lines-- line between the city orlando and the launch site/
- Explain why added those objects
  - adding these objects, readers can observe the result apparently.

- the GitHub URL of completed interactive map with Folium map:
- https://github.com/abcgz133/CapStoneIBMDataScience/blob/master/lab_jupyter_launch_site_location%20.ipynb

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions have added to a dashboard

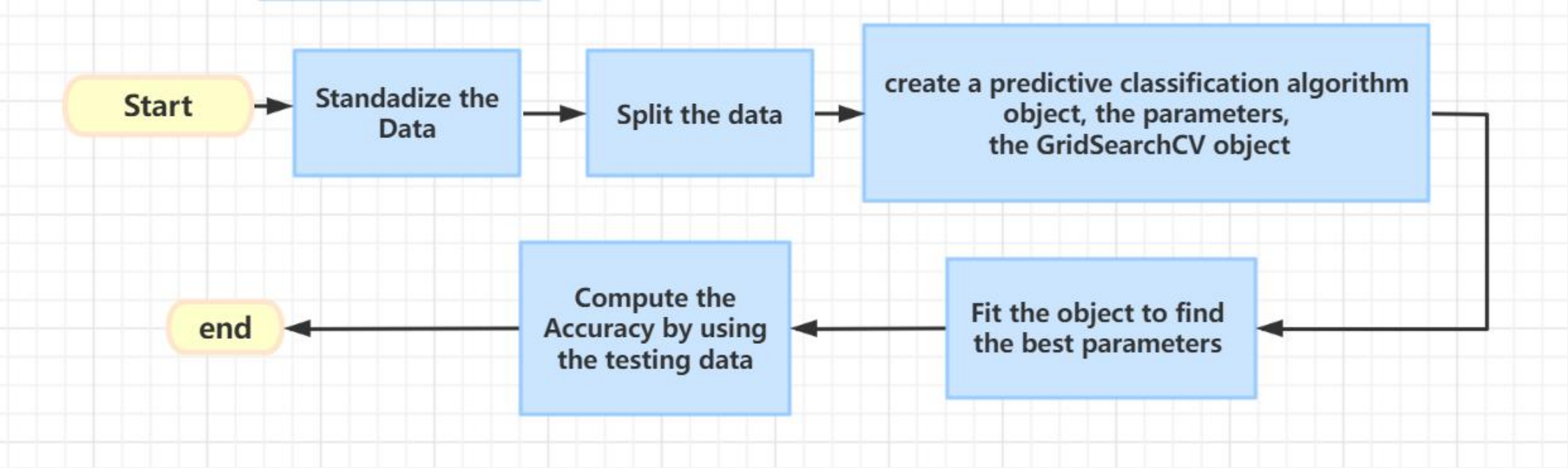
  Plots/graphs: scatter pot chart and pie chart

  interactions: slider and dropdown selections

- Reason:
  - adding these plots and interactions can easily disclose the insight in the data

- GitHub URL of completed Plotly Dash lab
- https://github.com/abcgz133/CapStoneIBMDataScience/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)

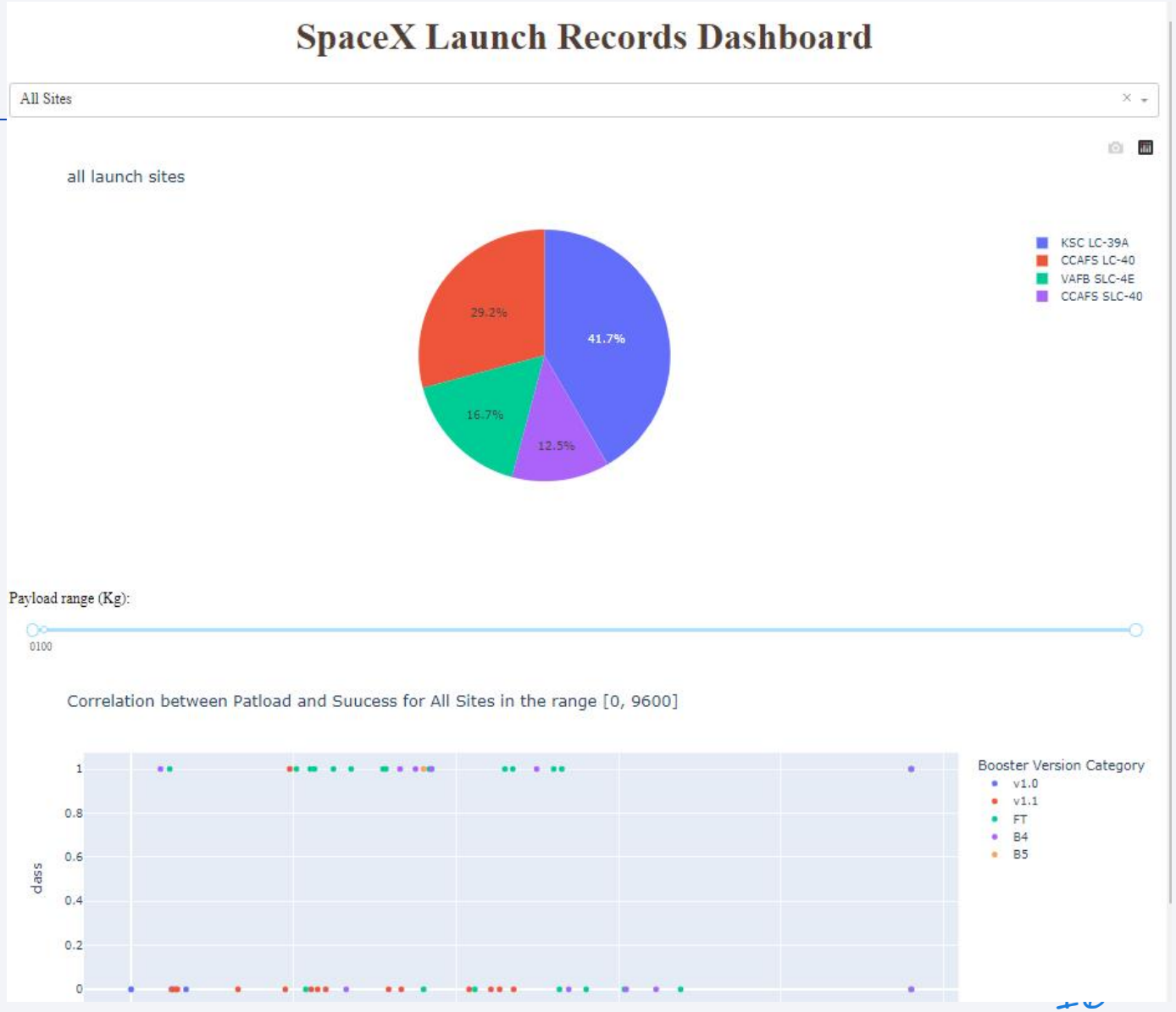


- the GitHub URL of completed predictive analysis lab:
- https://github.com/abcgz133/CapStoneIBMDataScience/blob/master/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

- Exploratory data analysis results
  - after year of 2013, the success rate has been upwards.

- Predictive analysis results
- in 18 cases, there are 12 landed, 3 did not landed which are 100% definity sure. but there are 3 cases are not sure

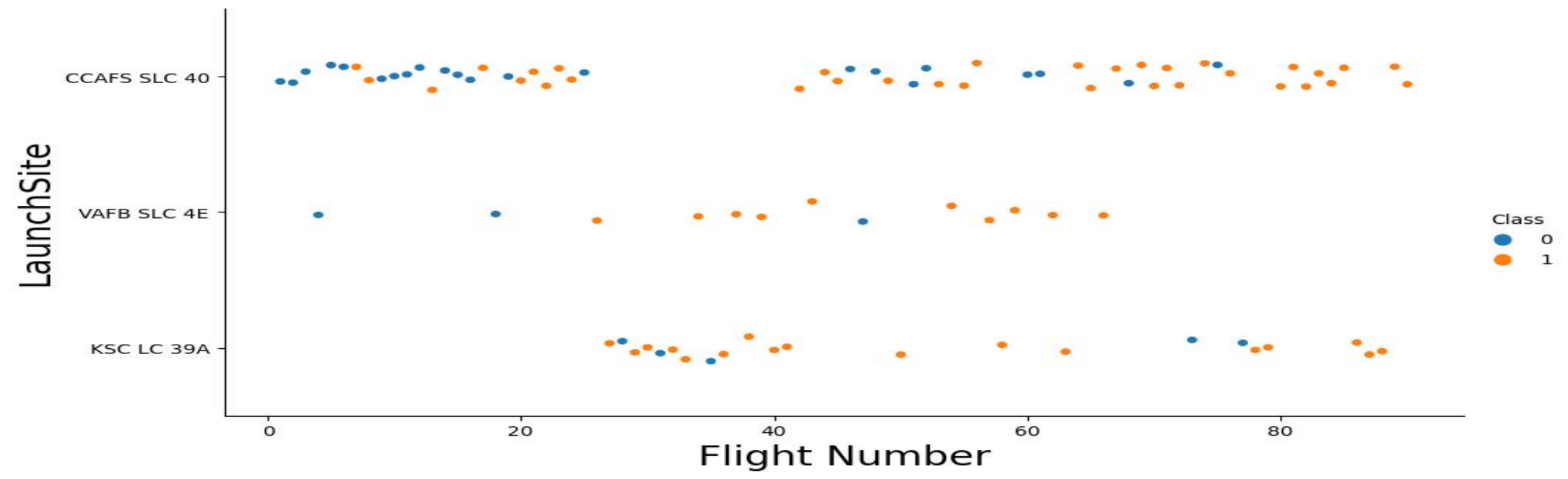


Section 2

Insights drawn from EDA

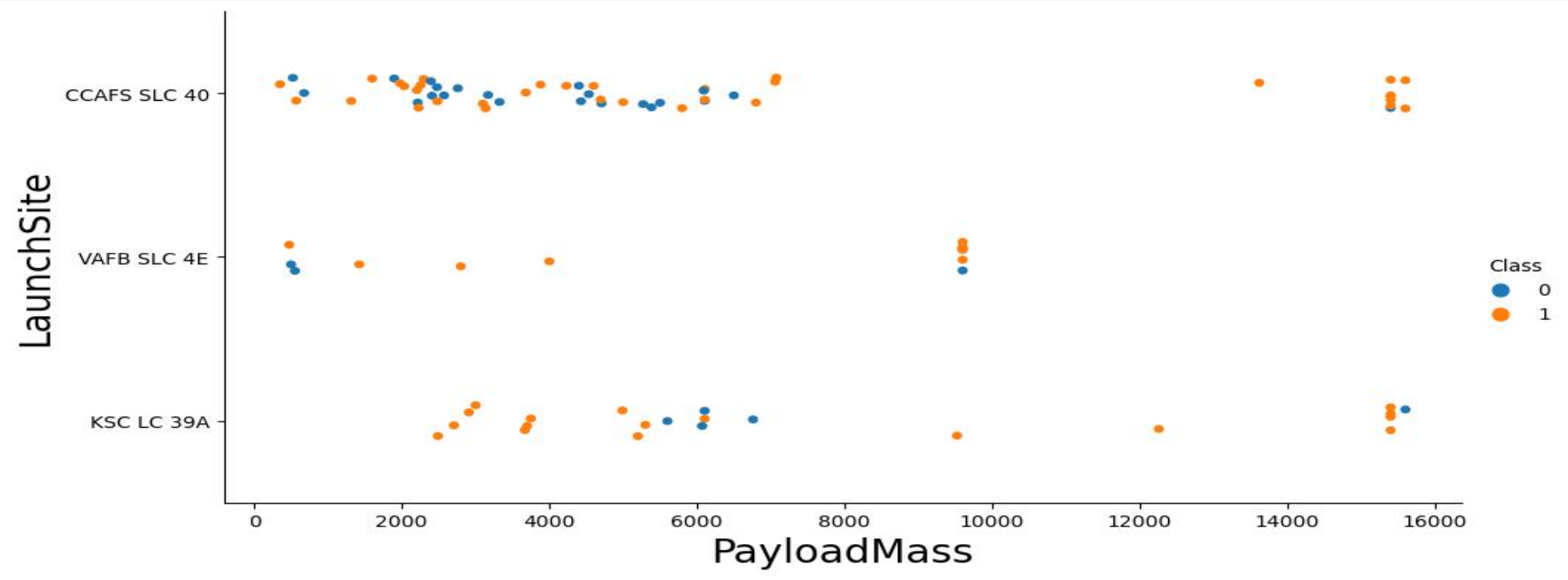
Flight Number vs. Launch Site

- the screenshot of the scatter plot



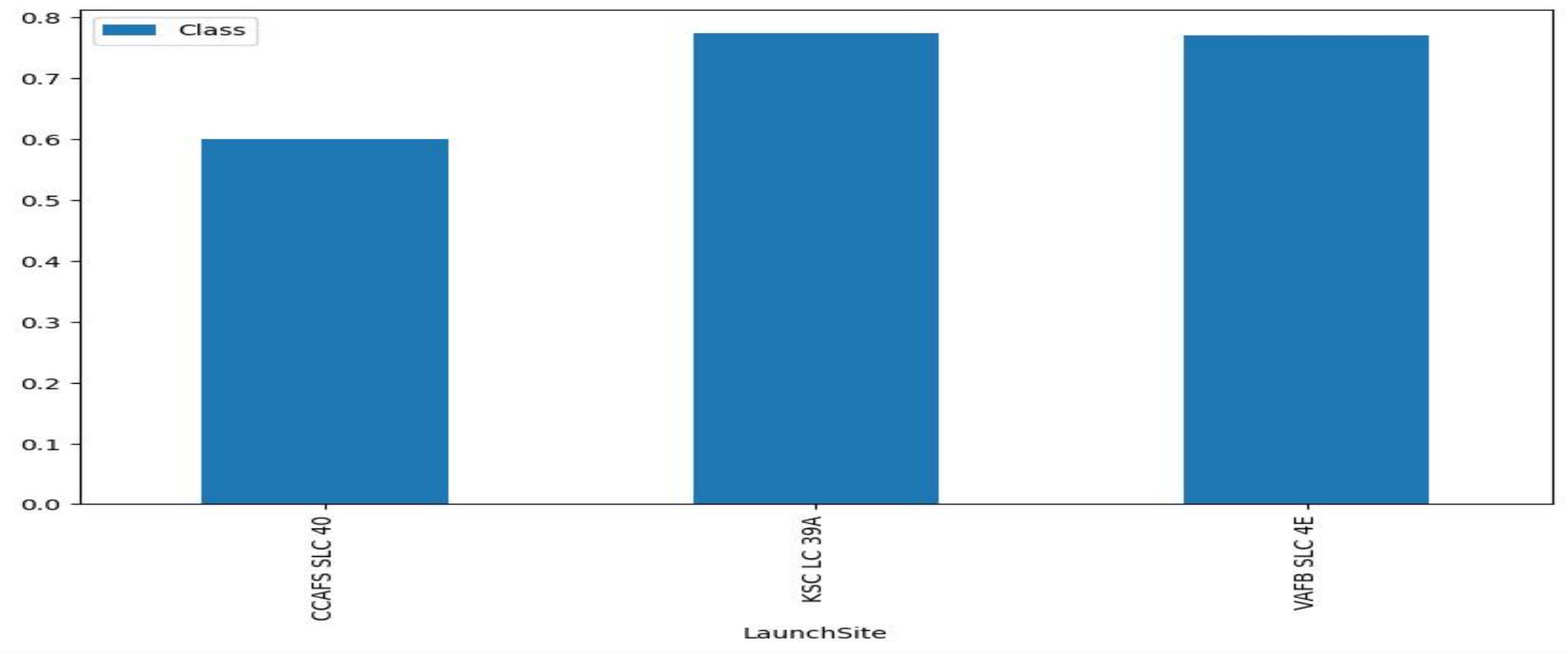
- as the flight number increases,it seems that the experice increases, the landing is ore likely being succed.

Payload vs. Launch Site



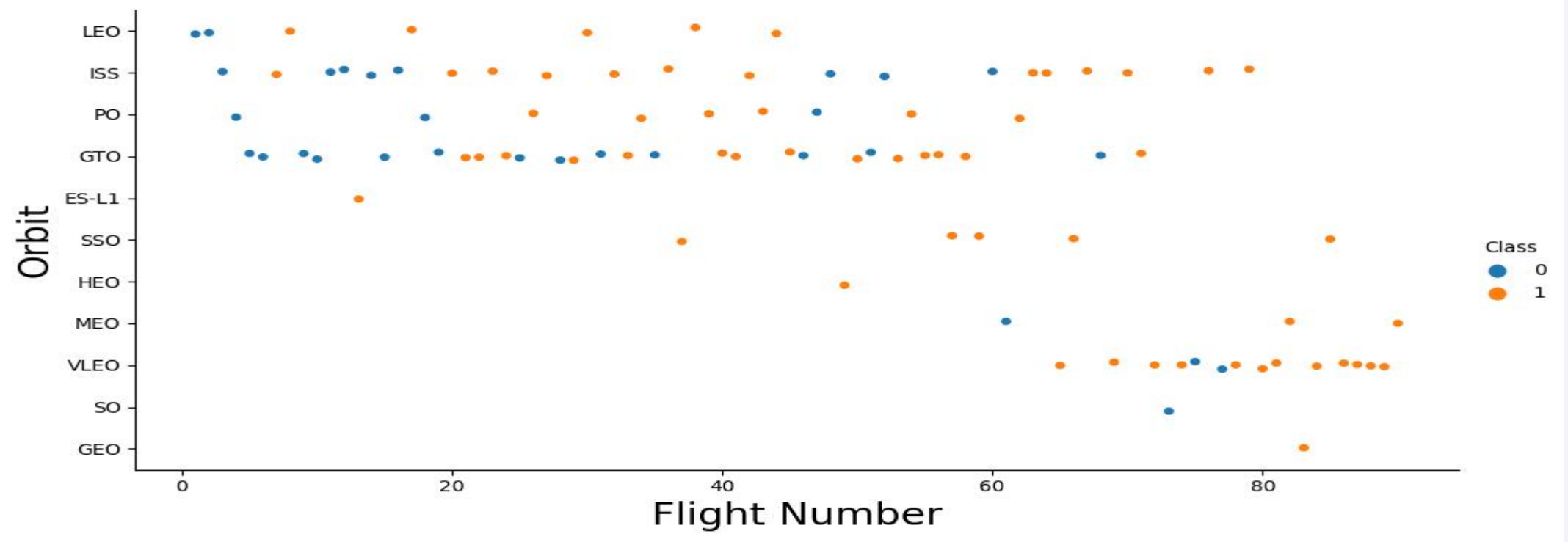
it seems that as the payload increasing(larger than 8,000kg), the more probabilty of success in all sites

Success Rate vs. Orbit Type



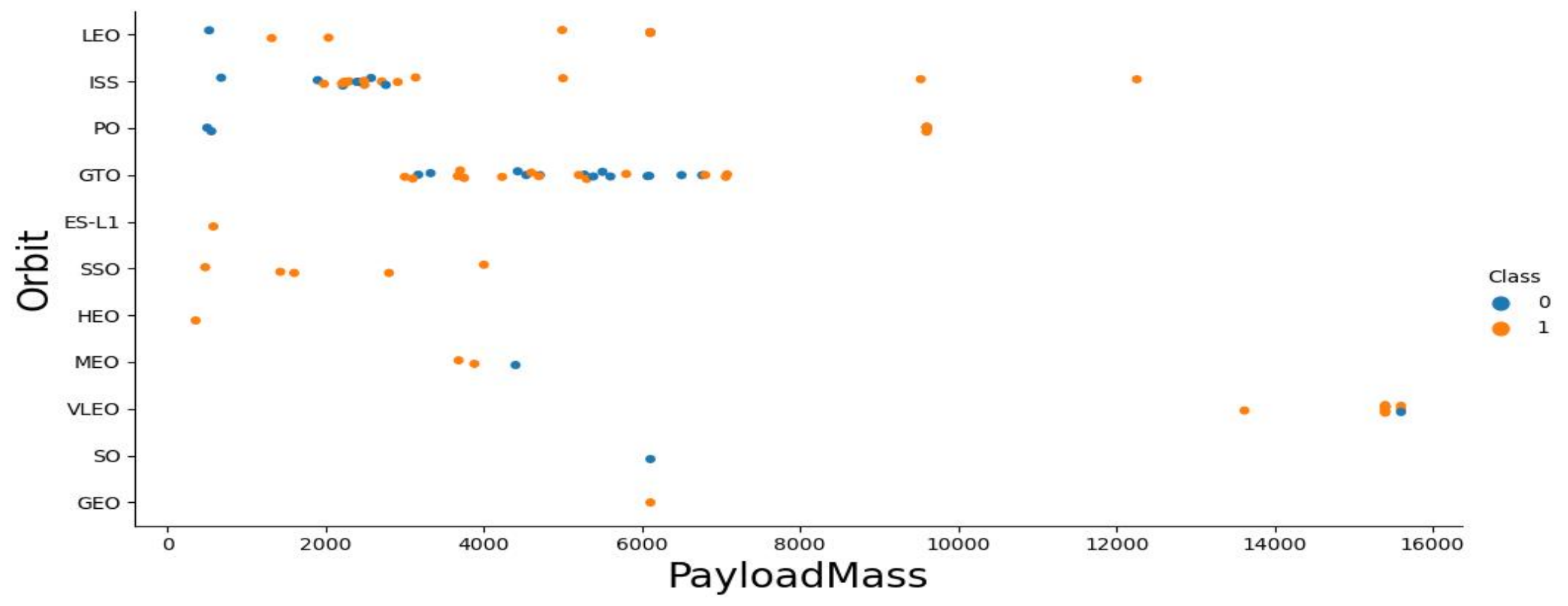
- different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.

Flight Number vs. Orbit Type



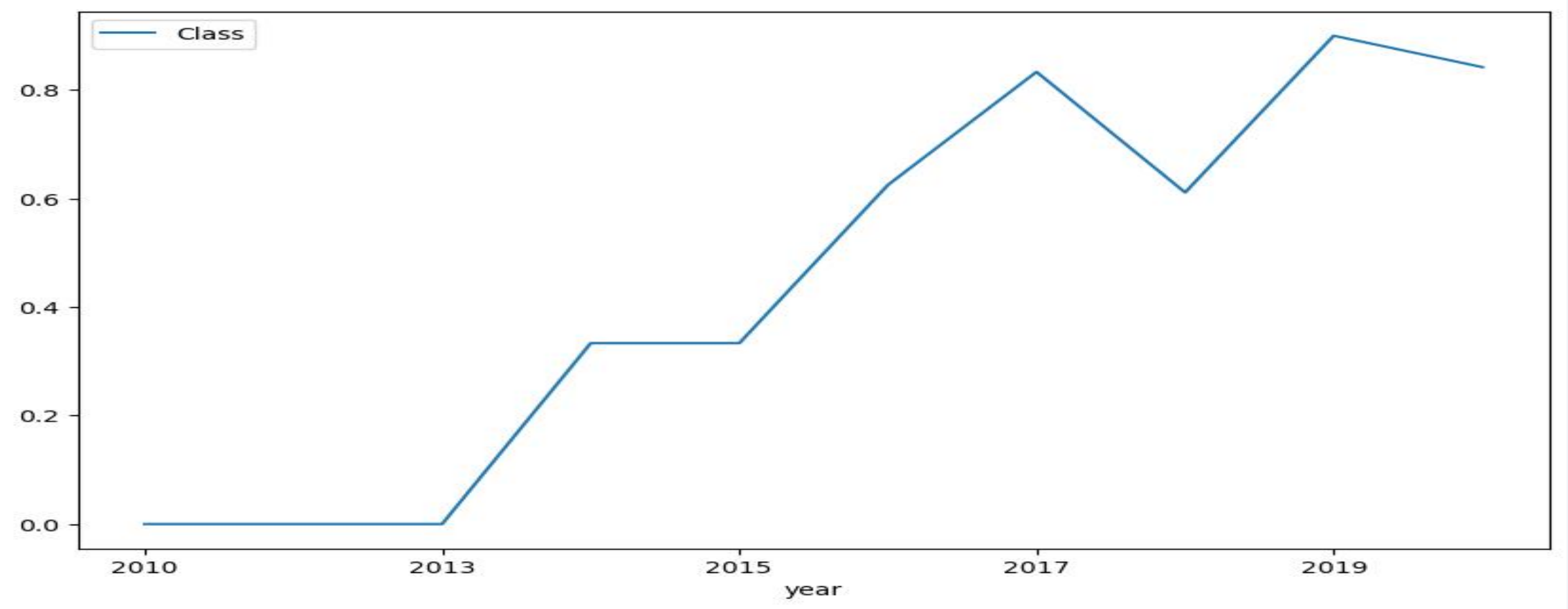
- it seems that with LEO, when number larger than 15 , it has 100% succeed rate. with SSO, it has 100% succeed rate.

Payload vs. Orbit Type



- Obviously, in the LEO orbit, when the Flight number is bigger than approximately 10, the Success rate is kept 100%.

Launch Success Yearly Trend



- we can notice that,except the year of 2017, the trend of yearly success rate is usually upwards after 2013 .

All Launch Site Names

- Launch_Site's names:

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

- There are four launch site located in U.S.A. They are very close to the coast line and the earth equator.

Launch Site Names Begin with 'CCA'

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

they are about 5 records.

Total Payload Mass

- the total payload carried by boosters from NASA

```
In [18]: %sql select sum(PAYLOAD_MASS__KG_) total_payload from spacextbl \
         where  Customer like "NASA (CRS)";
```

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

Out[18]:

total_payload
45596

- the total is about 45,000 kg from NASA

Average Payload Mass by F9 v1.1

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

```
In [19]: %sql select Round(AVG(PAYLOAD_MASS__KG_),2)  total_payload from spacextbl \
         where Booster_Version like  "F9 v1.1%"
 * sqlite:///my_data1.db
Done.
```

Out[19]:

total_payload
2534.67

- It seems that the average payload of F9 V1.1 is approximately 2,500kg

First Successful Ground Landing Date

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

```
In [23]: %sql select MIN(Date)  from spacextbl \
         where "Landing _Outcome" like "Success (ground pad)%"

          * sqlite:///my_data1.db
         Done.

Out[23]:  MIN(Date)
          01-05-2017
```

- since May, 2015, the launch has the succeed landing.

Successful Drone Ship Landing with Payload between 4000 and 6000

- the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Out[24]:

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

- there are 4 records which successfully landed and have a higher payloads.

Total Number of Successful and Failure Mission Outcomes

```
In [26]: %sql select count(*) ,  Mission_outcome   from \
         spacextbl group by Mission_Outcome
```

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

Out[26]:

count(*)	Mission_Outcome
1	Failure (in flight)
99	Success
1	Success (payload status unclear)

- there are 3 groups. the total number of success is 99, the failure is only 1. And there is 1 case is 'payload status unclear'.

Boosters Carried Maximum Payload

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

```
In [30]: %sql select  Booster_Version, PAYLOAD_MASS__KG_ from spacextbl where PAYLOAD_MASS__KG_ =  \
         (select max(PAYLOAD_MASS__KG_) max_payload from spacextbl limit 1)
```

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

Out[30]:

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

- there are 12 boosters. the maximum is 15,600kg.

2015 Launch Records

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

```
In [33]: %sql select substr(date, 7,4) year, substr(date, 4,2) month ,  \
         "Landing _Outcome" , Booster_Version, Launch_site   from spacextbl \
         where "Landing _Outcome" =  "Failure (drone ship)" \
         and substr(date, 7,4 ) = "2015"
```

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

Out[33]:

year	month	Landing _Outcome	Booster_Version	Launch_Site
2015	01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015	04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- in 2015, there were 2 records.

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

- 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

In [35]:
```
%sql SELECT "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 LANDING_OUTCOME_COUNT DESC
```

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

Out[35]:

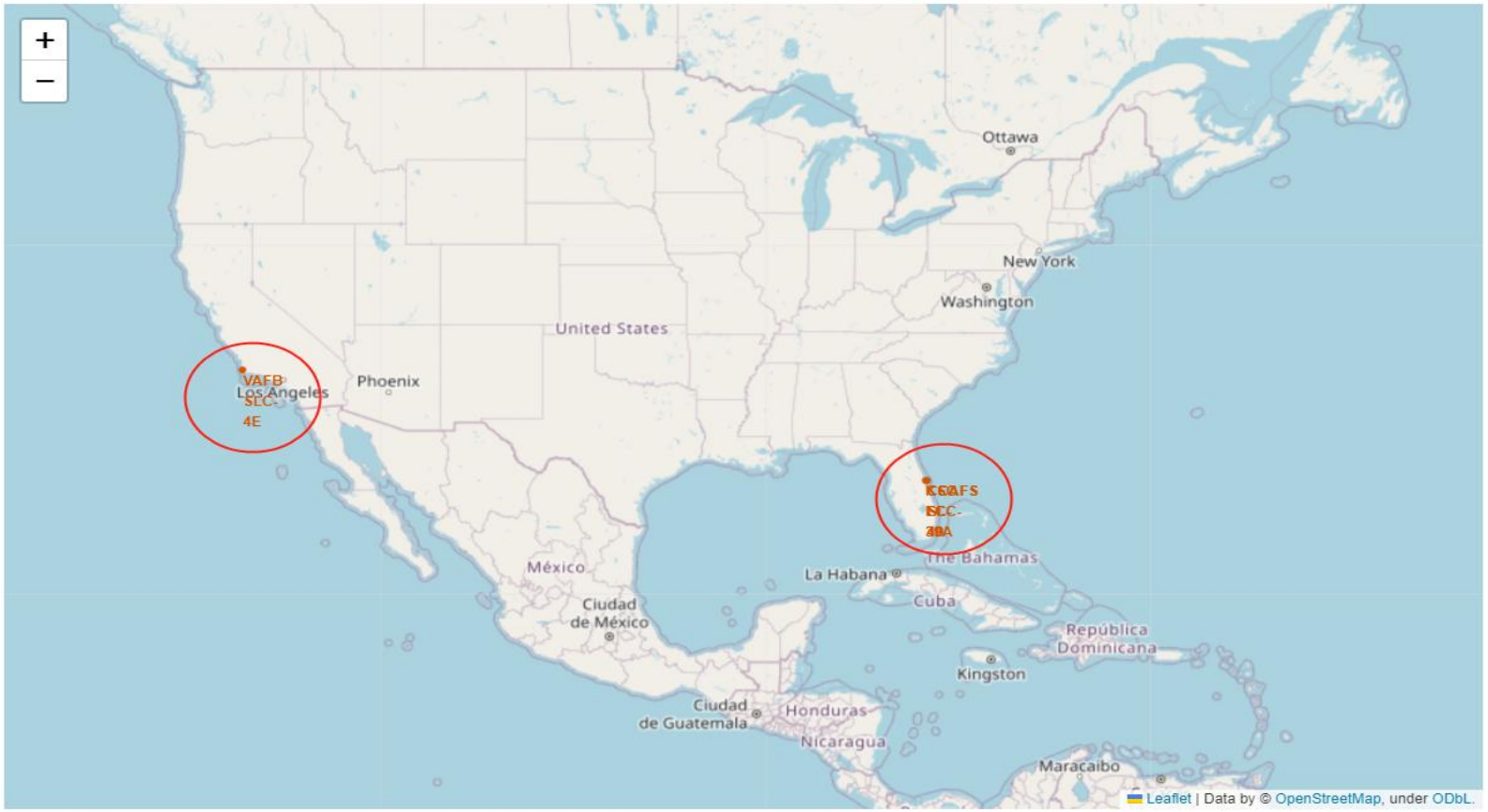
Landing _Outcome	LANDING_OUTCOME_COUNT	Date
No attempt	10	22-05-2012
Success (drone ship)	5	08-04-2016
Failure (drone ship)	5	10-01-2015
Success (ground pad)	3	22-12-2015
Controlled (ocean)	3	18-04-2014
Uncontrolled (ocean)	2	29-09-2013
Failure (parachute)	2	04-06-2010
Precluded (drone ship)	1	28-06-2015

- there were total 8 success which included 5 in drone ship and 3 in ground pad.


Section 3
Launch Sites
Proximities Analysis

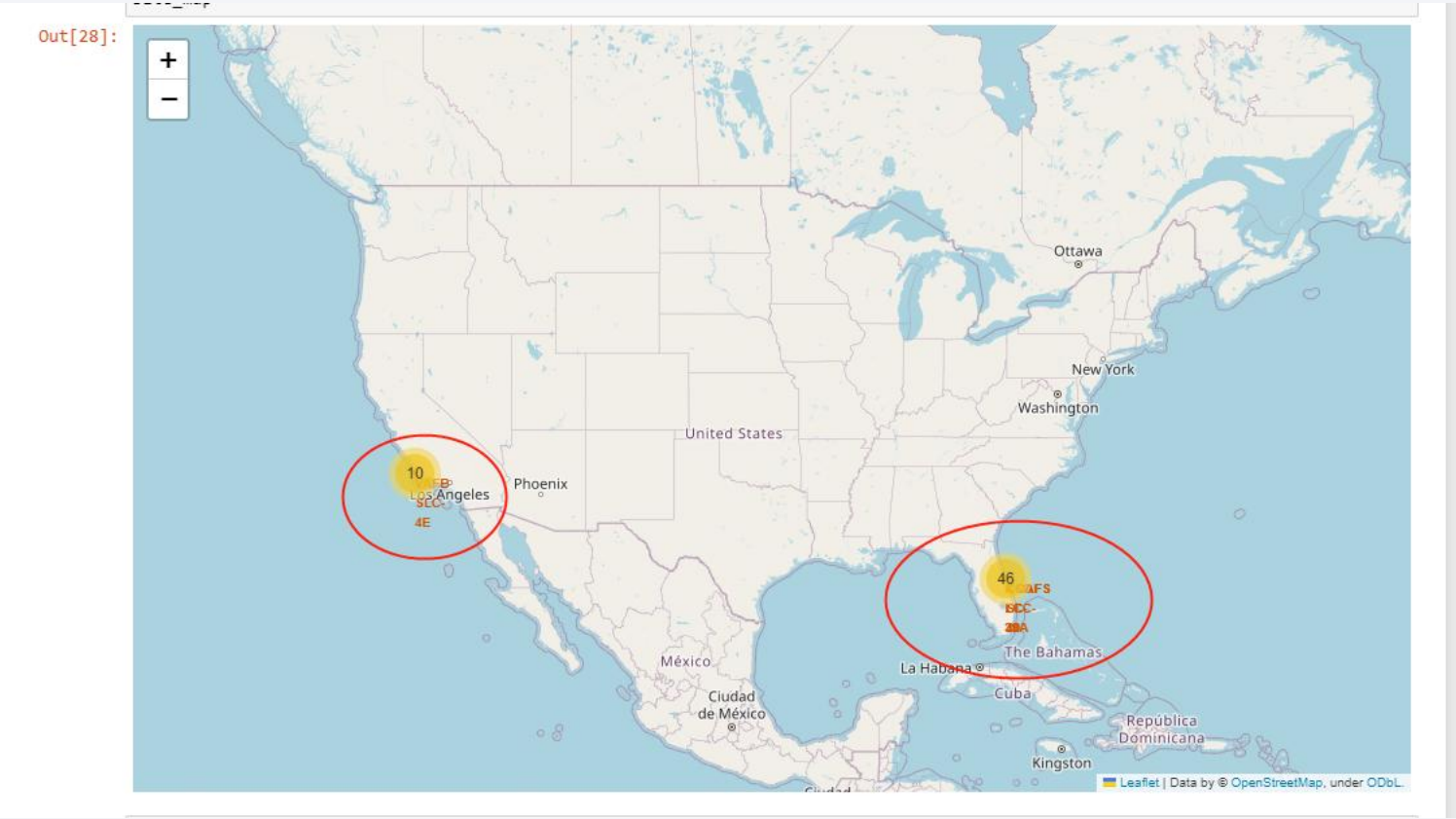
Find the optimal locations by using Folium

Out[22]:



- from this map, we can notice that these Sites are very close to coast line and the equator of the earth.

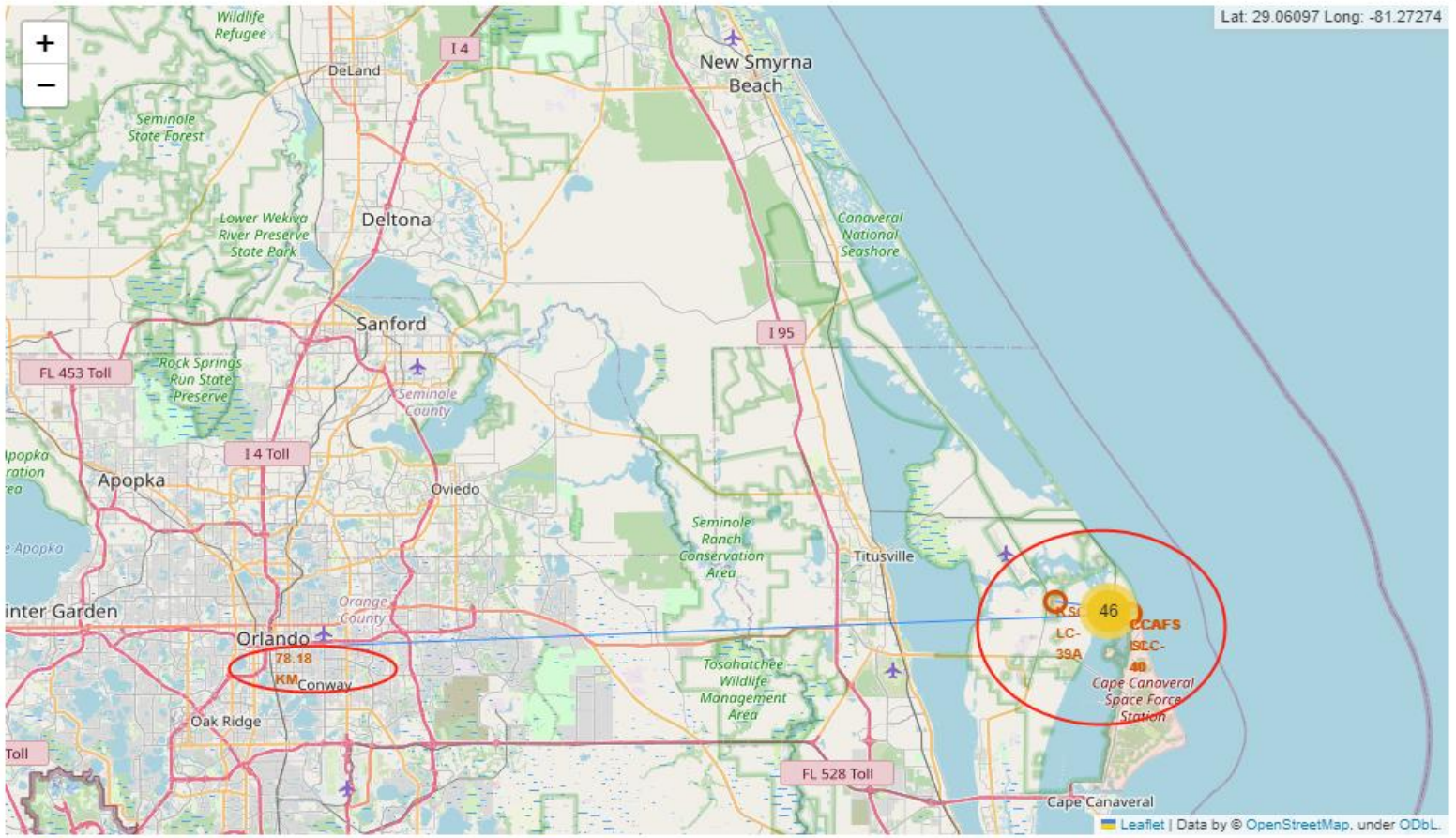
different launch times in Florida and California



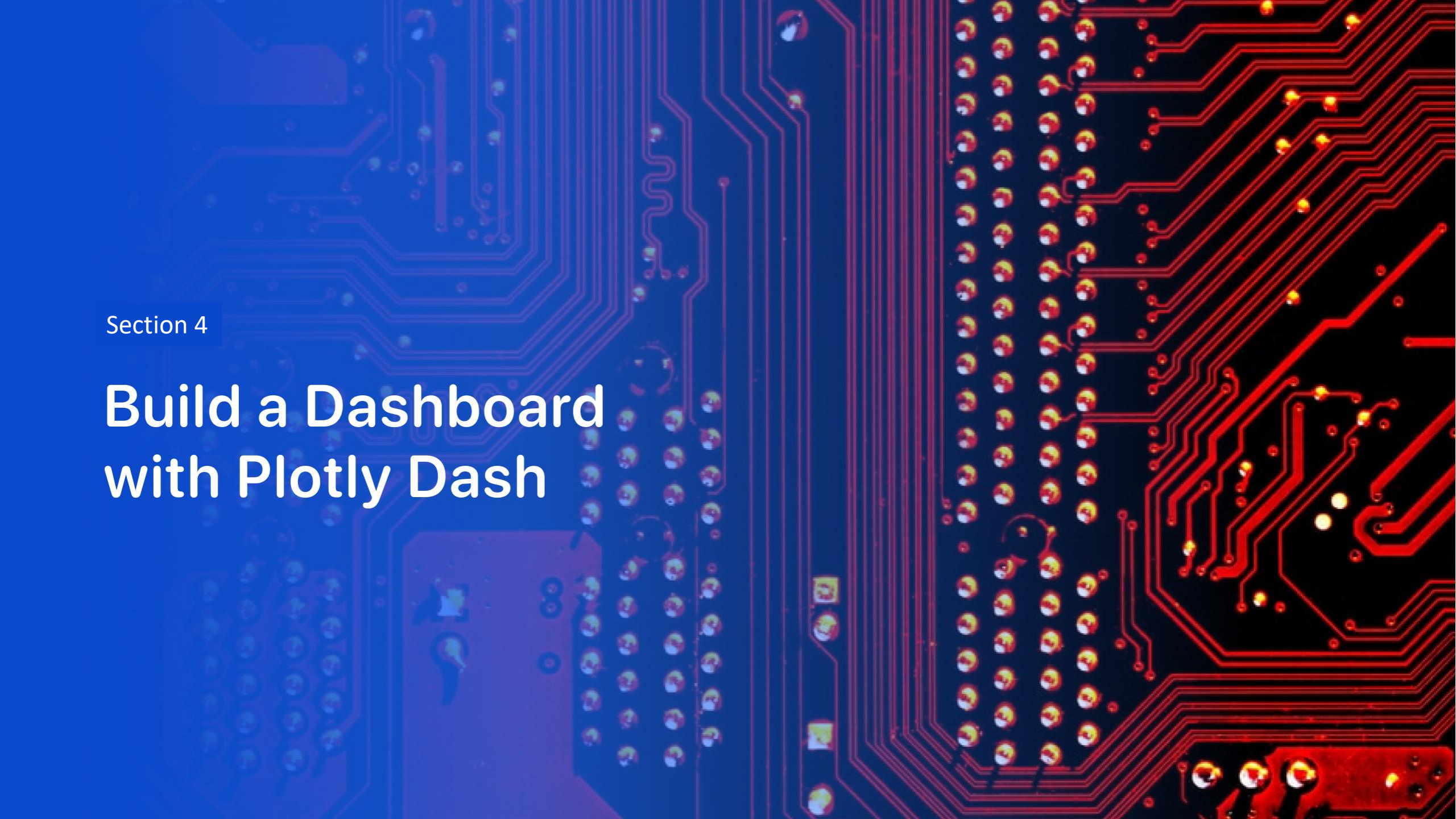
- there are 46 launch times in Florida and 10 in California. The Florida is more close to equator of the earth is one reason.

launch sites are very close to the coast line and city

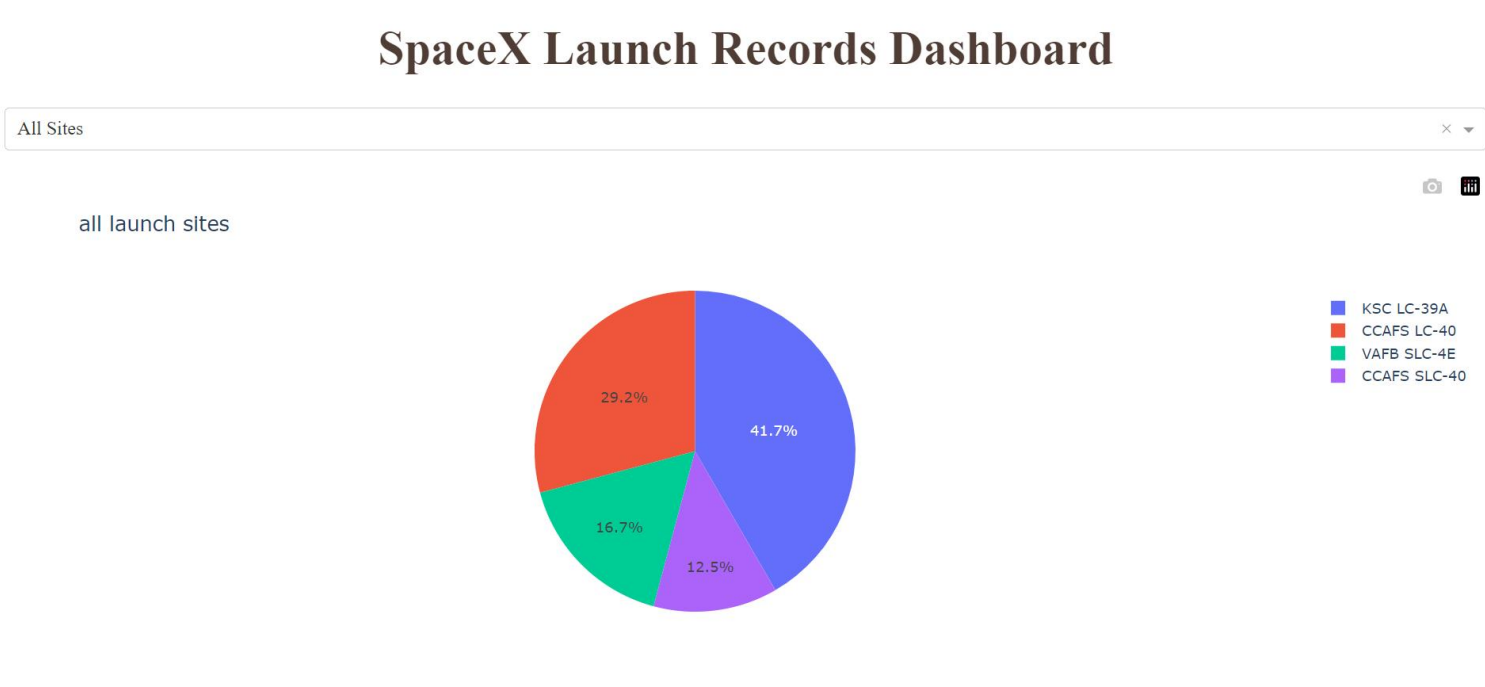
Out[25]:



- the 2 sites in Florida which have the most lanuch numbers are only have about 78 kms to Orlando city and close to coast line.

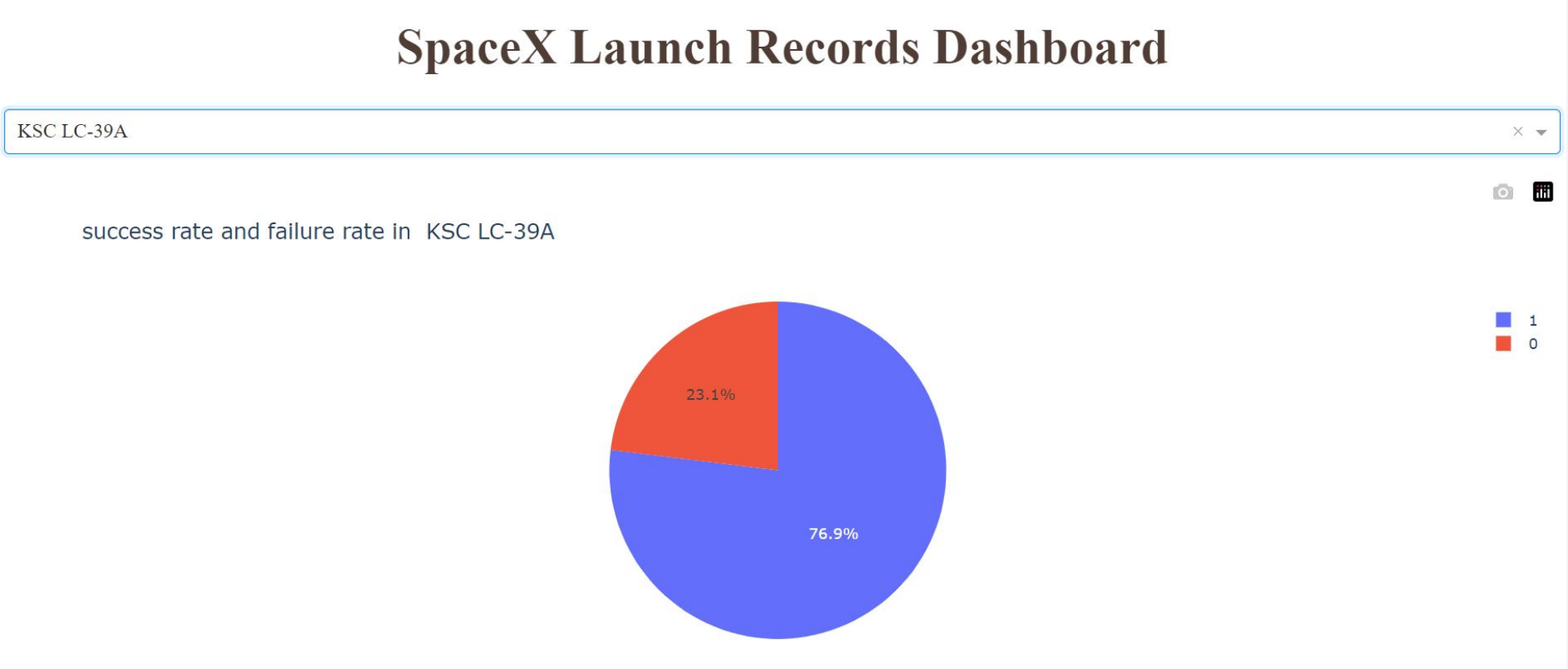

Section 4
Build a Dashboard with Plotly Dash

KSC LC-39A has the largest launch success ratio



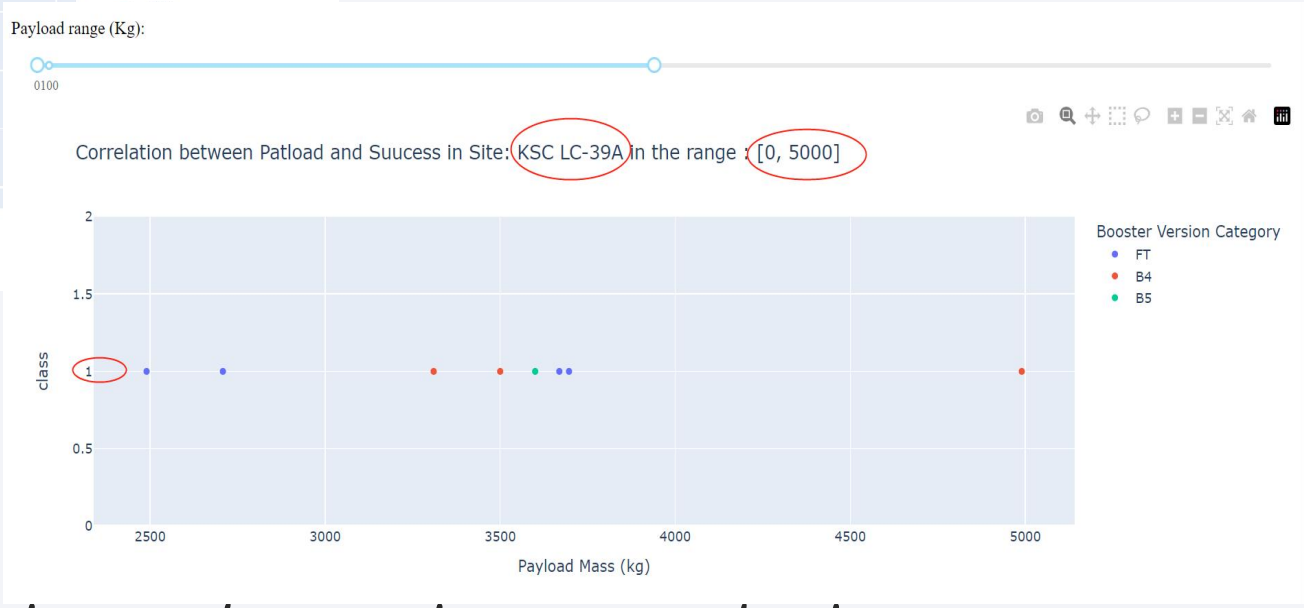
- Comparing other sites, In all success launches, KSC LC-39A occupies largest launch ratio.

KSC LC-39A has the largest ratio



- In KSC LC-39A, it's success ratio is 76.9%

what payload was proper? in which site?

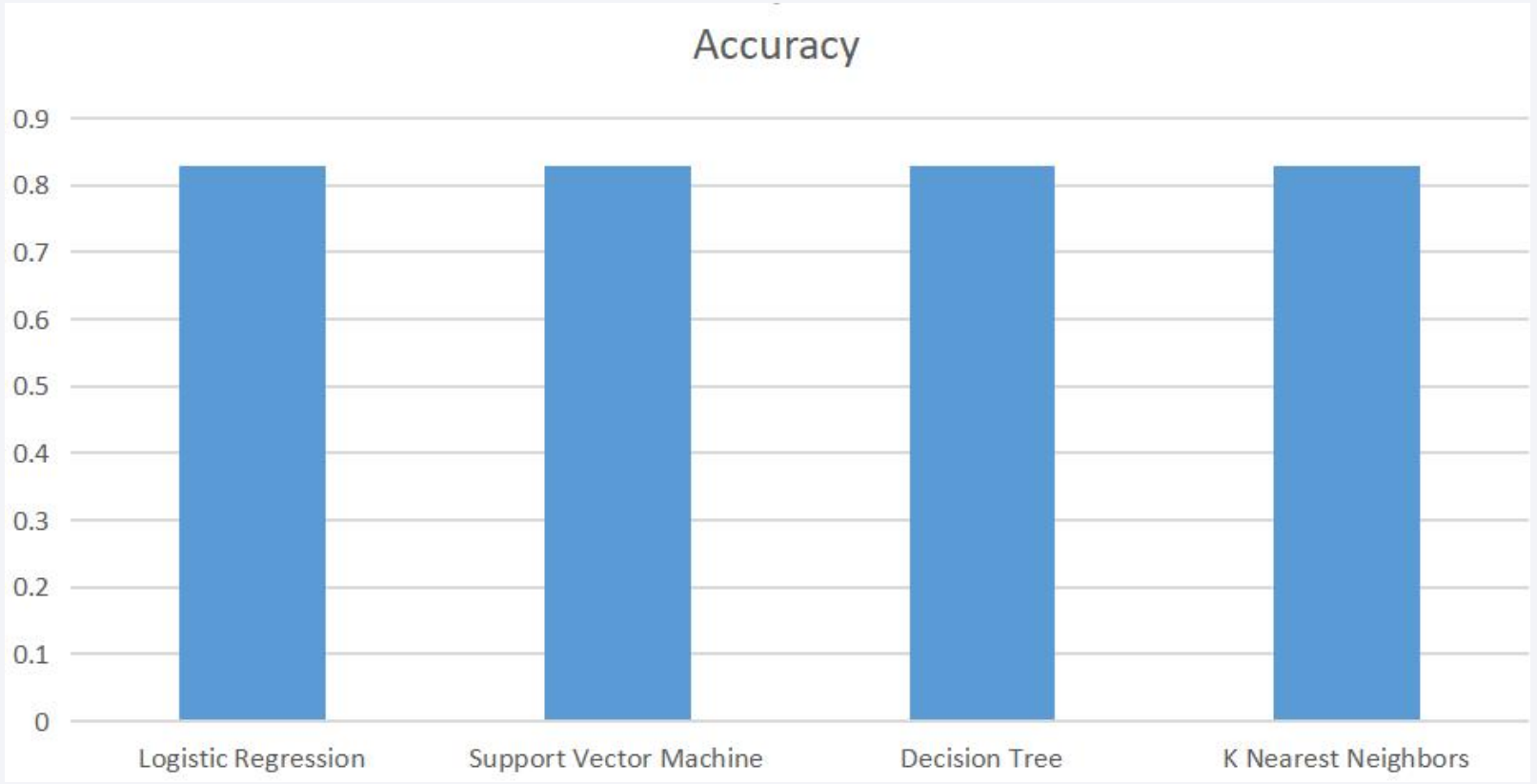


- in site KSC LC-39A, payload which was lower than 5000kg has 100% success rate.

Section 5

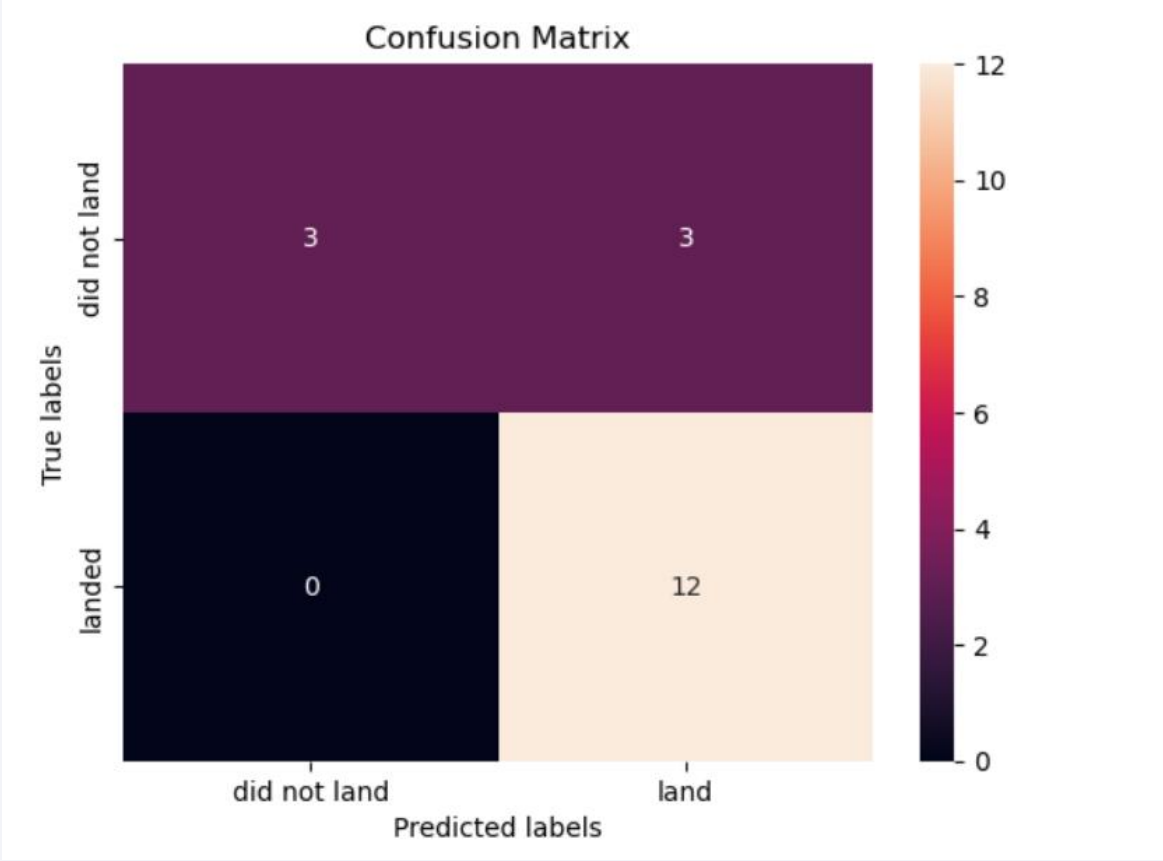
Predictive Analysis (Classification)

Classification Accuracy



- All 4 models have the same accuracy. This is because the samples are too small.

Confusion Matrix



- All 4 models have the same Confusion Matrix. This is because the samples are too small.

Conclusions

- in 18 cases, there are 12 landed, 3 did not landed which were 100% definity sure. but there are 3 cases are not sure

- these samples are too small to identify which one has the best performance accuracy.

- the features are too less. the more important features are collected , the more accuracy of these models.

Appendix

- some comments. Being a data analyst, who have more than 14 years working in the credit card department in a bank, I totally understand the power of Data Science.

- But in this case, 4 models have the same accuracy. This is because the features of the data are too less. the more important features are collected , the more accuracy of these models. so being a Data Scientist , the most important work is to search and collect more and more data from diversity sources to support the analyzing job.

Thank you!