IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Cong Cheng
2023-11-13

Outline

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

Executive Summary

Summary of Methodologies

- Data Collection
- Webscraping
- Data Wrangling
- Exploratory Data Analysis(EDA)
- Geospatial Analysis
- Dashboard
- Machine Learning Prediction

Summary of all results

- After applying 4 different methods of machine learning, the results on the test set are almost the same among these 4 method. Further study is need for accuracy improvment.

Introduction

Project background and context

- SpaceX has pioneered the technology with its Falcon 9 rocket. The ability for the first stage of a rocket to land is a significant advancement in aerospace technology, primarily because it represents a leap towards reusable launch vehicles, which can drastically reduce the cost of access to space.

Problems you want to find answers

- Predict the success rate of first stage landing.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Request rocket launch data from SpaceX API
  - Webscrap from the Wikipage
- Perform data wrangling
  - Define what bad outcomes are and create a new column "Class" to store the result
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Pre-processing
  - Split the data
  - Apply models
  - Tune hyperparameters with GridSearchCV

Data Collection

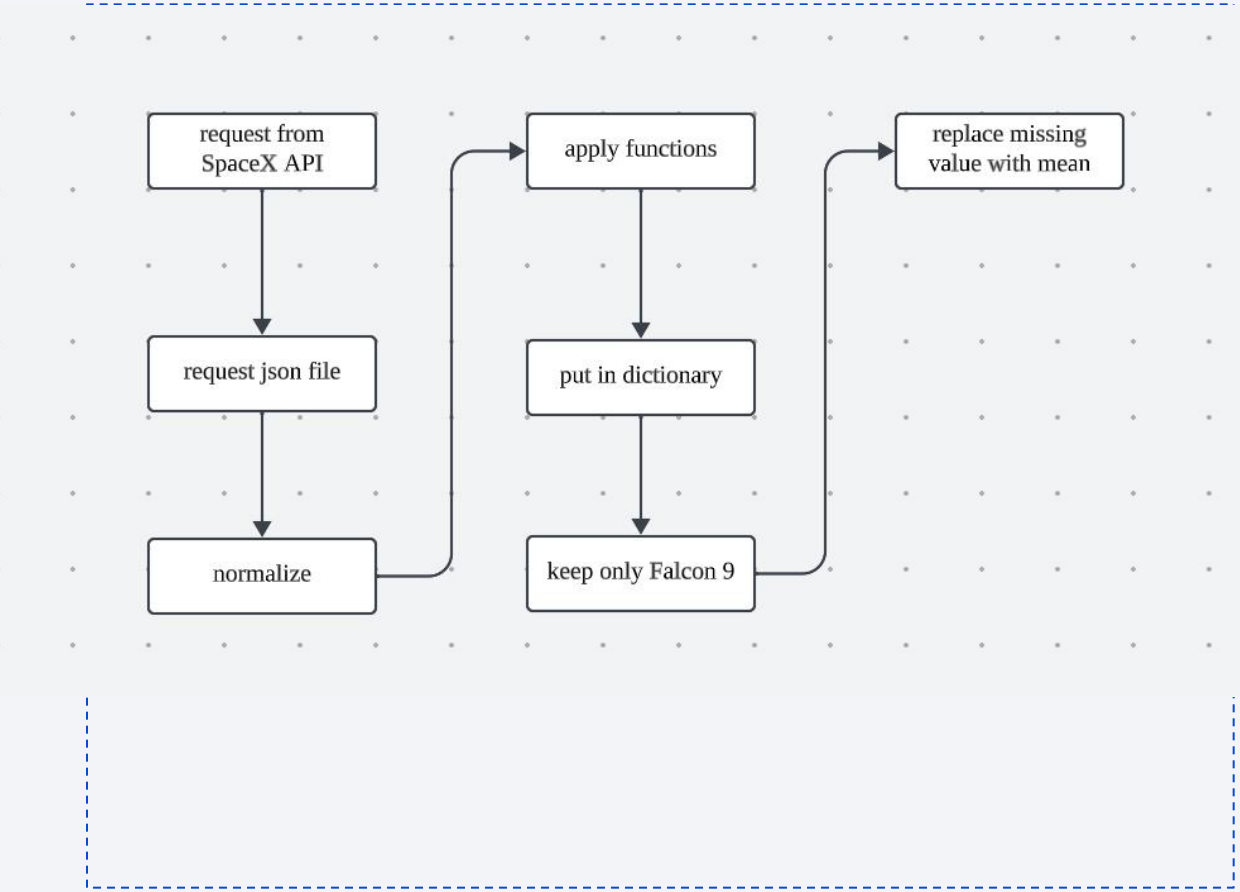
- Request rocket launch data from SpaceX API

- Webscrap from the Wikipage

Data Collection – SpaceX API

- Make request from SpaceX API with url.
- Define functions to get corresponse data
- Filter the data of interest and replace the missing value with mean value

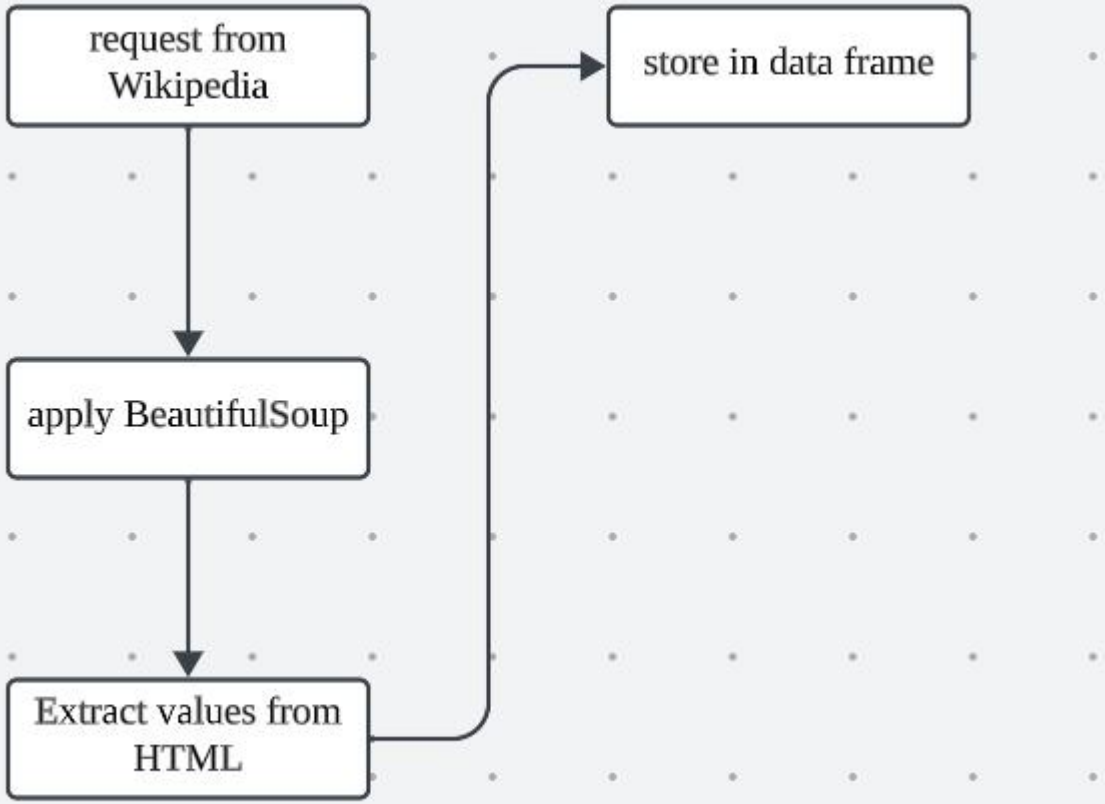
- GitHub URL:https://github.com/Csquaree/applied-data-science-capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Extract a Falcon 9 launch records HTML table from Wikipedia
- Parse the table and convert it into a Pandas data frame

- GitHub URL: https://github.com/Csquaree/applied-data-science-capstone/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Determine Training Labels:

  Define what bad outcomes are

  create a new column "Class" to store the result

- GitHub URL:https://github.com/Csquaree/applied-data-science-capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Scatter plot (observe relationship)
  - Flight Number v.s. Launch Site
  - Payload Mass v.s. Launch Site
  - Orbit Type v.s. Flight Number
  - Payload v.s. Orbit Type
- Bar chart (comparision)
  - Success rate v.s. Orbit Type
- Line chart (understand the change)
  - Success rate v.s. Year

- GitHub URL:https://github.com/Csquaree/applied-data-science-capstone/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- The SQL queries performed:

1. Display the names of the unique launch sites in the space mission
2. Display 5 records where launch sites begin with the string 'CCA'
3. Display the total payload mass carried by boosters launched by NASA (CRS)
4. Display average payload mass carried by booster version F9 v1.1
5. List the date when the first succesful landing outcome in ground pad was acheived.
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. Use a subquery
9. List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015
10. 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.

- GitHub URL: https://github.com/Csquaree/applied-data-science-capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- what map objects are created and added to a folium map:
  - Mark all launch sites on a map
  - Mark the success/failed launches for each site on the map
  - Calculate the distances between a launch site to its proximities
- The reason why these objects are added:
  - To help understand why they are here

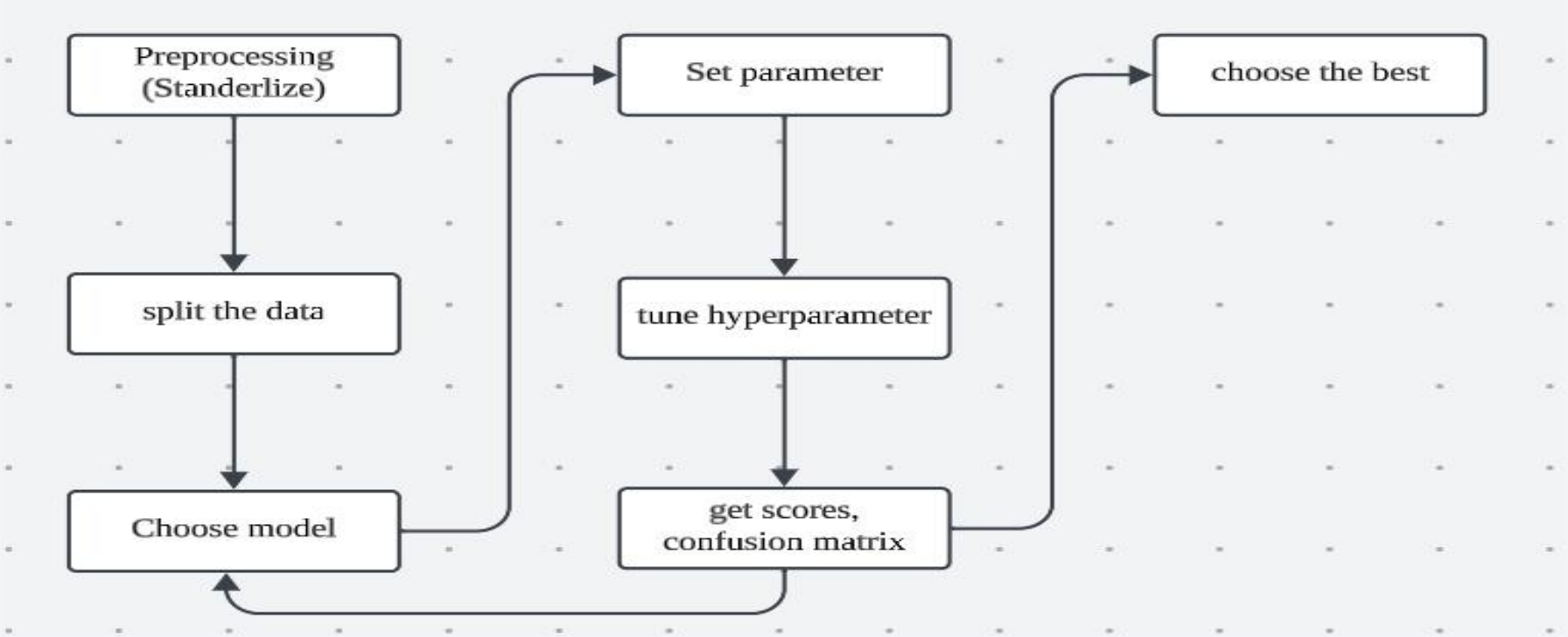
- GitHub URL: https://github.com/Csquaree/applied-data-science-capstone/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- Pie chart
  - All Sites: to see the proportion of success rate in all locations.
  - Certain Site: to see the success rate in the certain site.
- Scatter chart (with adjustable Payload Mass)
  - relationship between these two variables

- GitHub URL: https://github.com/Csquaree/applied-data-science-capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)



- GitHub URL: https://github.com/Csquaree/applied-data-science-capstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

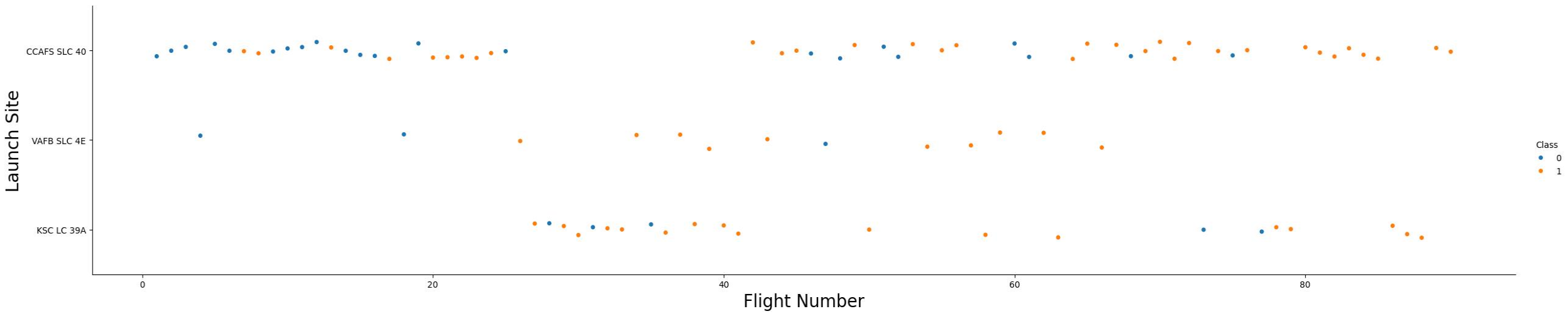
Results

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

Section 2

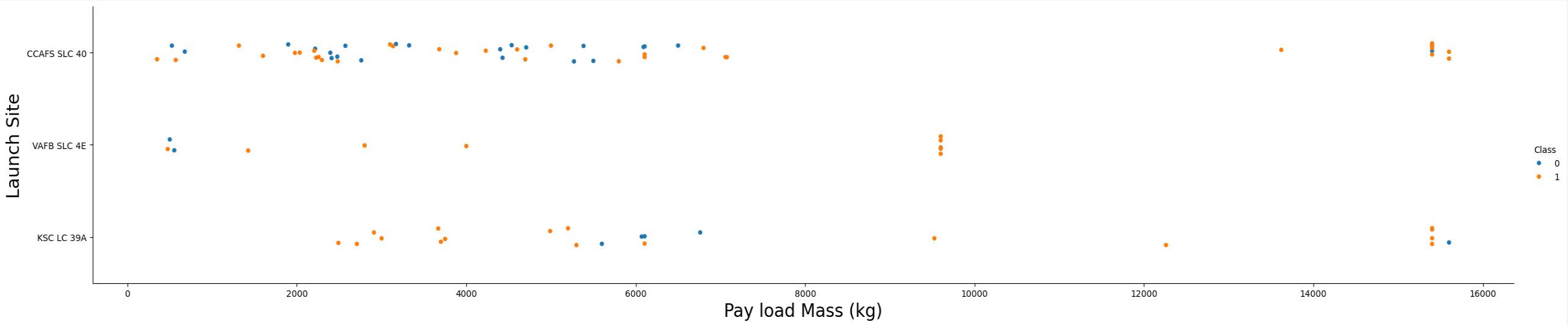
Insights drawn from EDA

Flight Number vs. Launch Site



- Class 0 indicate failure, Class 1 indicate success.
- After Flight Number 20, the count of success land becomes majority. CCAFS SLC 40 launched most flights from 1 to 25 and from 40 to 80+. VAFB SLC 4E launched the least flights. KSC LC 39A launched most flights from 25 to 40.

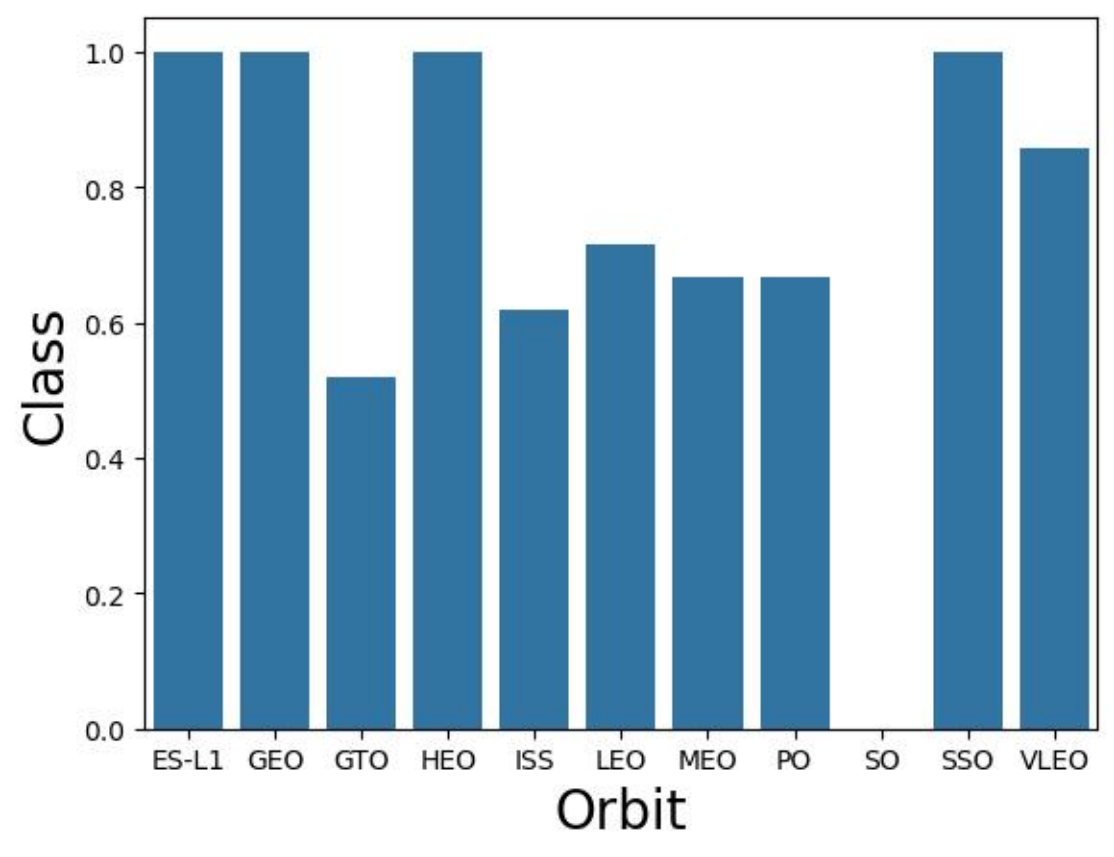
Payload vs. Launch Site



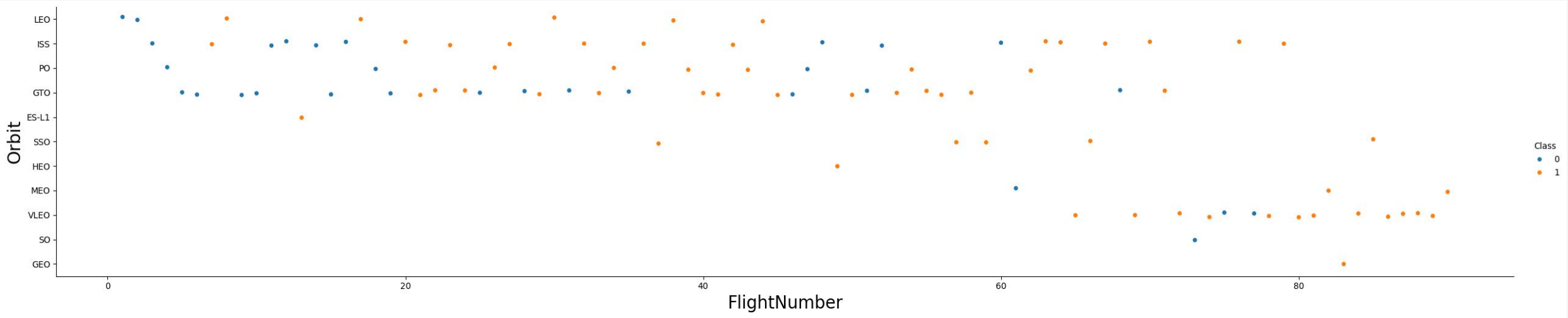
- When the Payload Mass is haveier than 8000kg, only a few flights fail to land.
- Most failure cases are have Payload Mass between 0 to 8000kg in CCAFS SLC 40

Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, SSO have 100% success rate.
- SO has 0% success rate
- VLEO has about 90% success rate
- The rest orbit types have around 50%-70% success rate

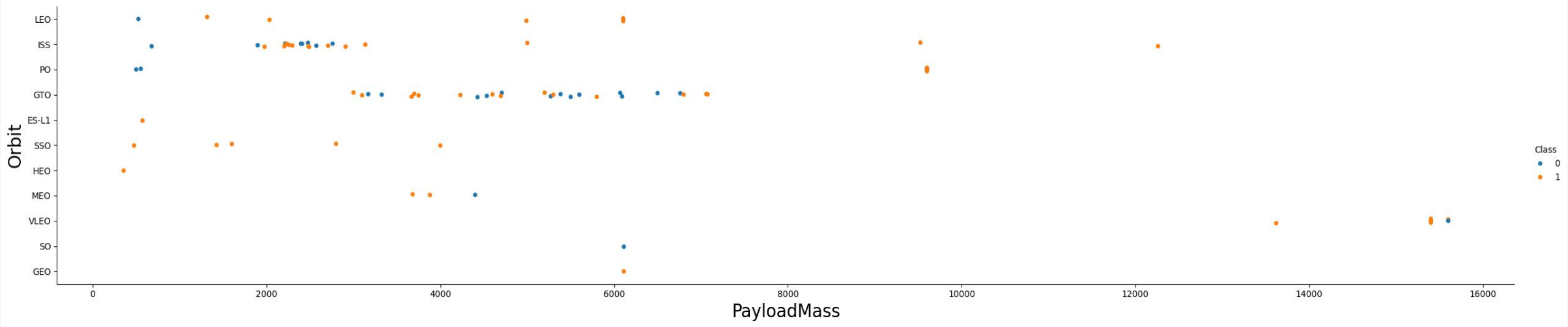


Flight Number vs. Orbit Type



- Most blue points are located in the left side of the plot indicating the successful rate is related to the flight number.
- One thing about those have 100% success rate is that they all only have a small number of flights.

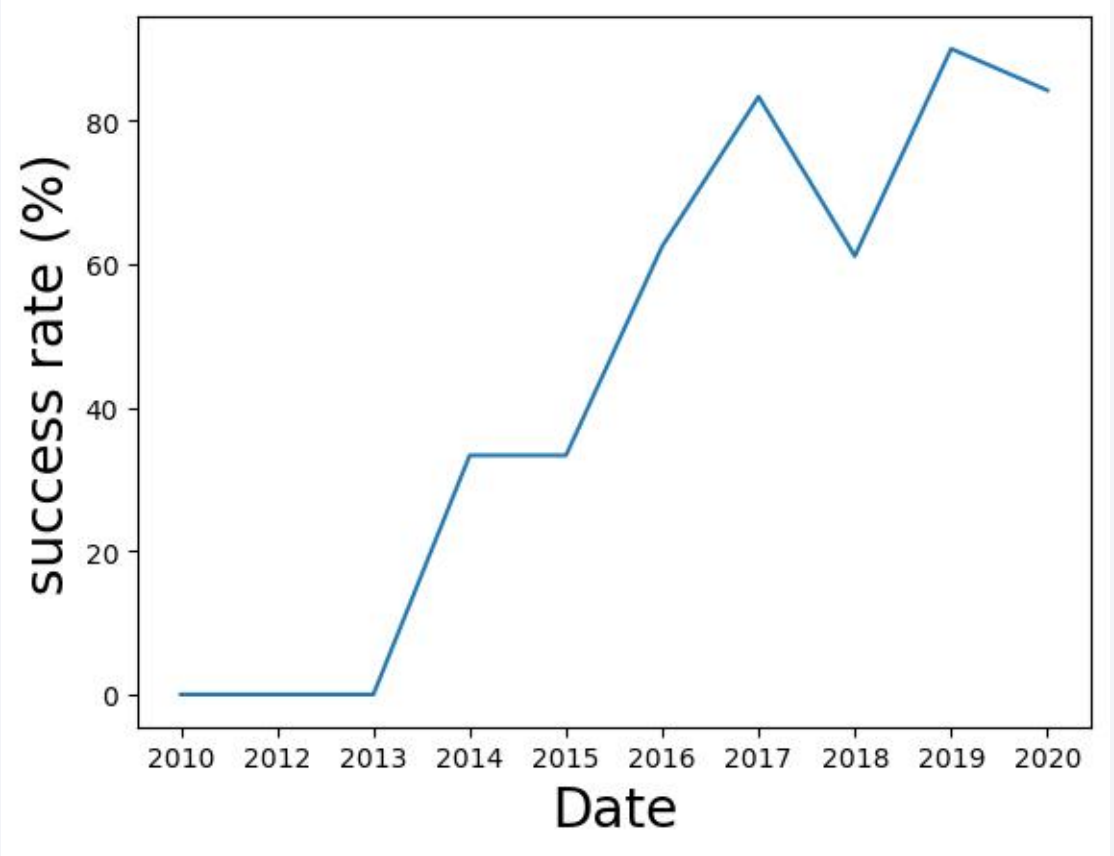
Payload vs. Orbit Type



- The relationship between mass and successful rate is not clear for ISS and GTO, and the rest of orbit types are have a few number of flights which is not enough to support any conclusions.

Launch Success Yearly Trend

- The success rate is 0% from 2010 to 2013
- The success rate is increasing over the time.
- There is a significant decrease from 2017 to 2018.



All Launch Site Names

Task 1

Display the names of the unique launch sites in the space mission

In [23]:
```
%sql SELECT DISTINCT(LAUNCH_SITE) FROM SPACEXTABLE
```

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

Out[23]:

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

- Use Distinct to avoid repeated values

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Present your query result with a short explanation here

Total Payload Mass

Task 3

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

In [25]:
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD_MASS FROM SPACEXTABLE WHERE CUSTOMER = "NASA (CRS)"
```

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

Out[25]:

TOTAL_PAYLOAD_MASS
45596

- Use 'sum' to add the value up and 'as' to assign a new name

Average Payload Mass by F9 v1.1

Task 4

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

In [26]:
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVERAGE_PAYLOAD_MASS FROM SPACEXTABLE WHERE BOOSTER_VERSION = "F9 v1.1"
```

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

Out[26]:

AVERAGE_PAYLOAD_MASS
2928.4

- Use 'avg' to find the mean

First Successful Ground Landing Date

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

In [27]:
```
%sql SELECT MIN(DATE) AS FIRST_SUCCESSFUL_LANDING_GROUND FROM SPACEXTABLE WHERE LANDING_OUTCOME = "Success (ground pad)"
```

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

Out[27]:

FIRST_SUCCESSFUL_LANDING_GROUND
1/8/2018

- Use 'min' to find the smallest value

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [28]:
```
%sql SELECT DISTINCT(BOOSTER_VERSION) FROM SPACEXTABLE WHERE LANDING_OUTCOME = "Success (drone ship)" AND PAYLOAD_MASS__KG_ > 4000 AND
```

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

Out[28]:

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

- Use 'Where' to filter values

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

In [29]:
```
%sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS NUMBER FROM SPACEXTABLE GROUP BY MISSION_OUTCOME
```

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

Out[29]:

Mission_Outcome	NUMBER
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- Use 'Count' to find the number of a certain mission outcome
- Use 'Group by' to group the rows with the same mission outcome up

Boosters Carried Maximum Payload

- Use subquery to find the max payload mass and use it to compare

Task 8

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

In [30]:
```
%sql SELECT BOOSTER_VERSION FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

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

Out[30]:

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

2015 Launch Records

Task 9

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

In [31]:
```
%sql SELECT SUBSTR(DATE, -7, 2) AS MONTH, LANDING_OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTABLE WHERE LANDING_OUTCOME = "Fail
```

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

Out[31]:

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

Task 10

- Use 'substr' to get the value and use it to compare

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

- make the date format same first
- filter out the date we want

Task 10

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.

In [47]:

```
%%sql
SELECT LANDING_OUTCOME, COUNT(LANDING_OUTCOME)
FROM SPACEXTABLE
WHERE
  substr('0000' || substr(DATE, -4), -4) || '-' ||
  substr('00' || substr(DATE, instr(DATE, '/') + 1, instr(substr(DATE, instr(DATE, '/') + 1), '/') - 1), -2) || '-' ||
  substr('00' || substr(DATE, 1, instr(DATE, '/') - 1), -2)
  BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING_OUTCOME
ORDER BY COUNT(LANDING_OUTCOME) DESC
```

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

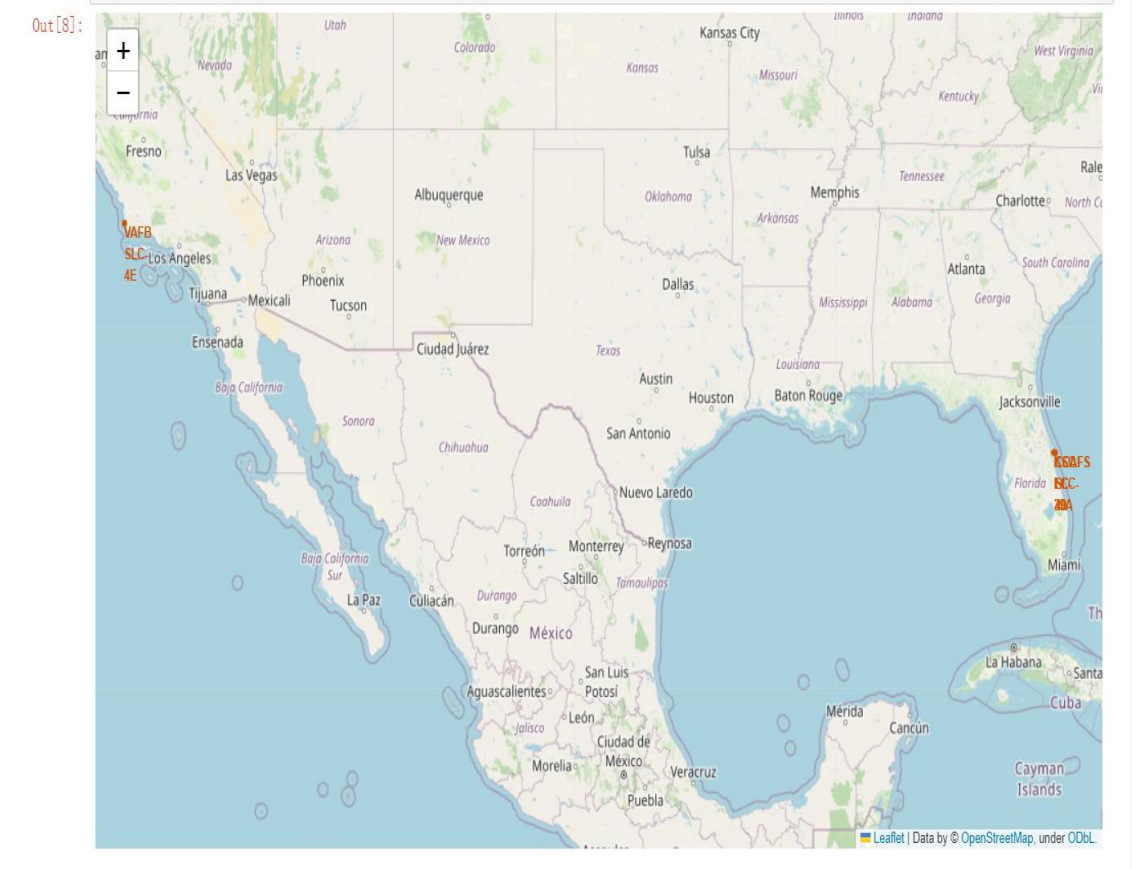
Out[47]:

Landing_Outcome	COUNT(LANDING_OUTCOME)
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

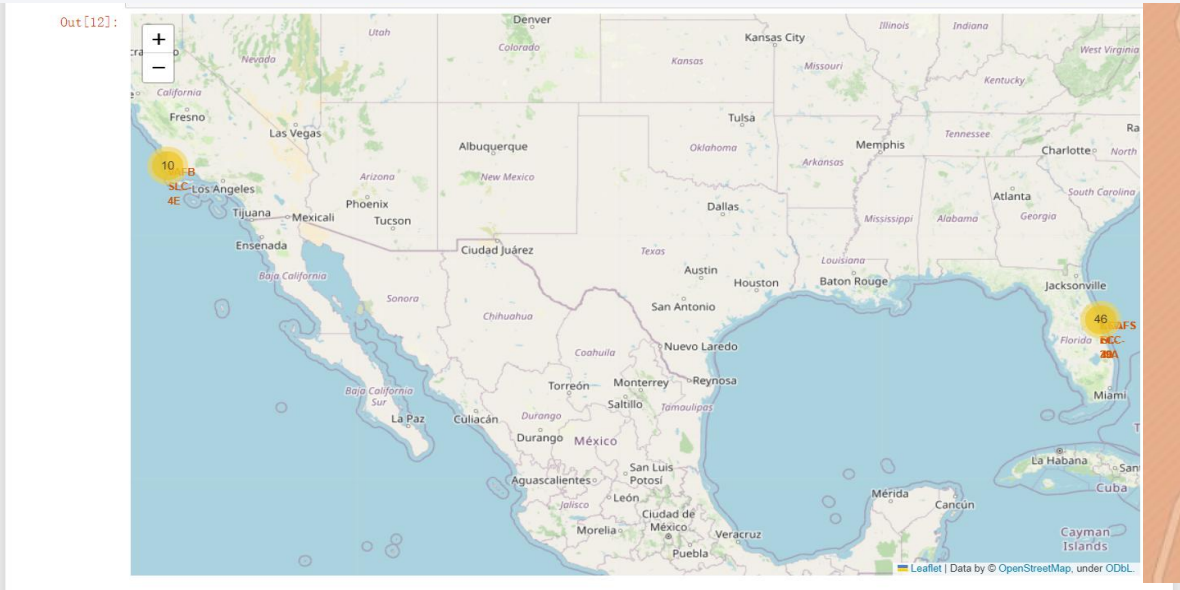

Section 3
Launch Sites
Proximities Analysis

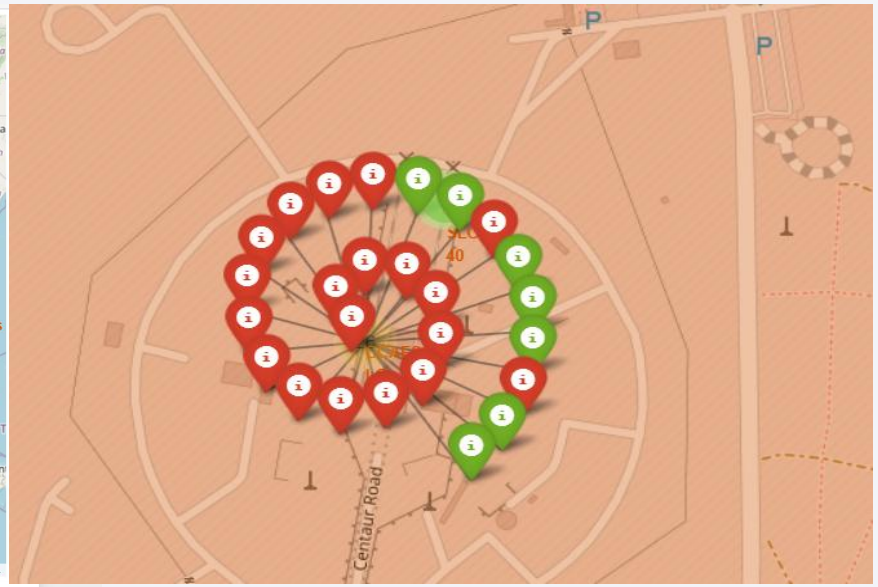
All launch sites' location markers on a global map

- All launch sites are near the ocean.



The color-labeled launch outcomes on the map

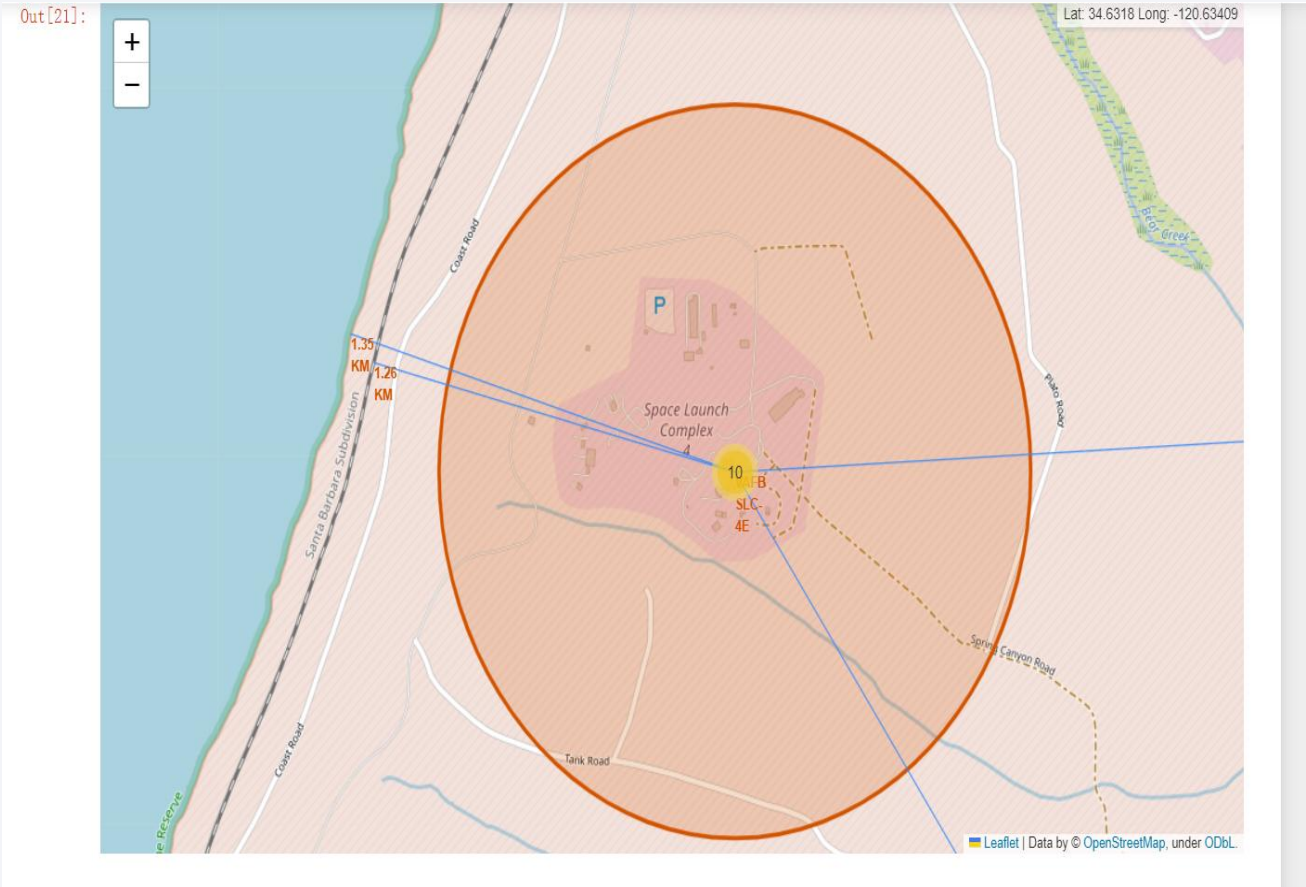


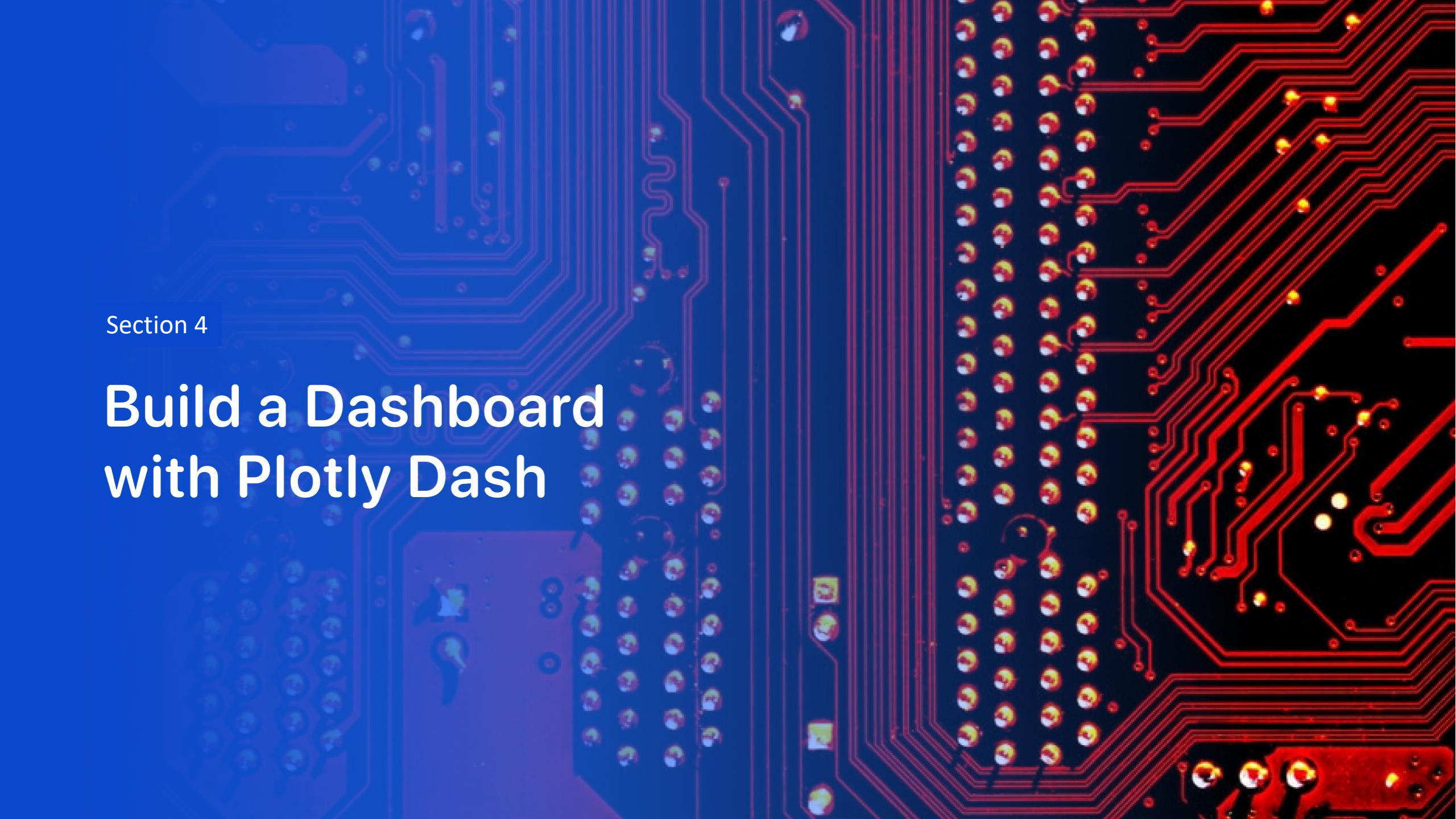


- can see the number of flights in the location
- can see the detail of the location by clicking it

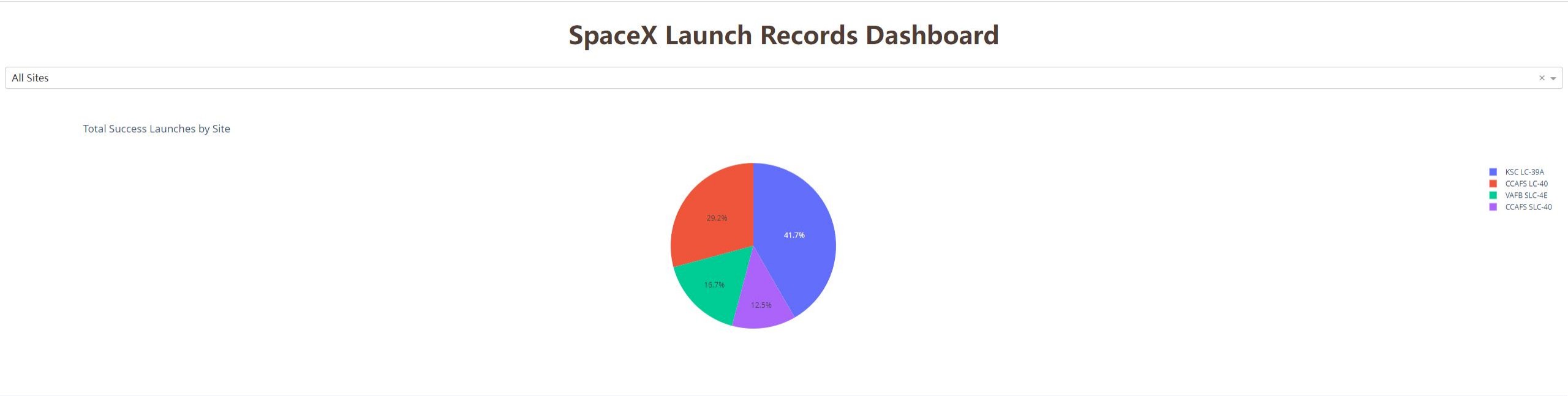
Launch site VAFB SLC-4E to its proximities

- For VAFB SLC-4E, it is close to coastline and rail way, far away from city and highway.
- Close to coastline can avoid too much damage when failure
- Close to railway can improve the efficiency of transporting supply
- Far from city and highway can minimize the damage to human



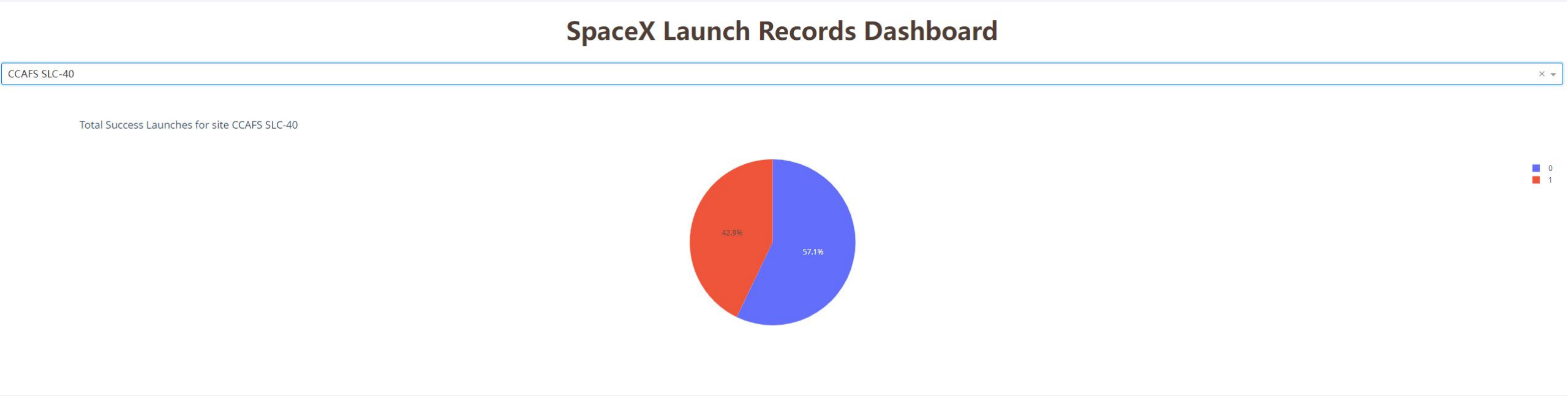

Section 4
Build a Dashboard with Plotly Dash

Launch success count for all sites



- KSC LC-39A has the most success count and CCAFS SLC-40 has the least.

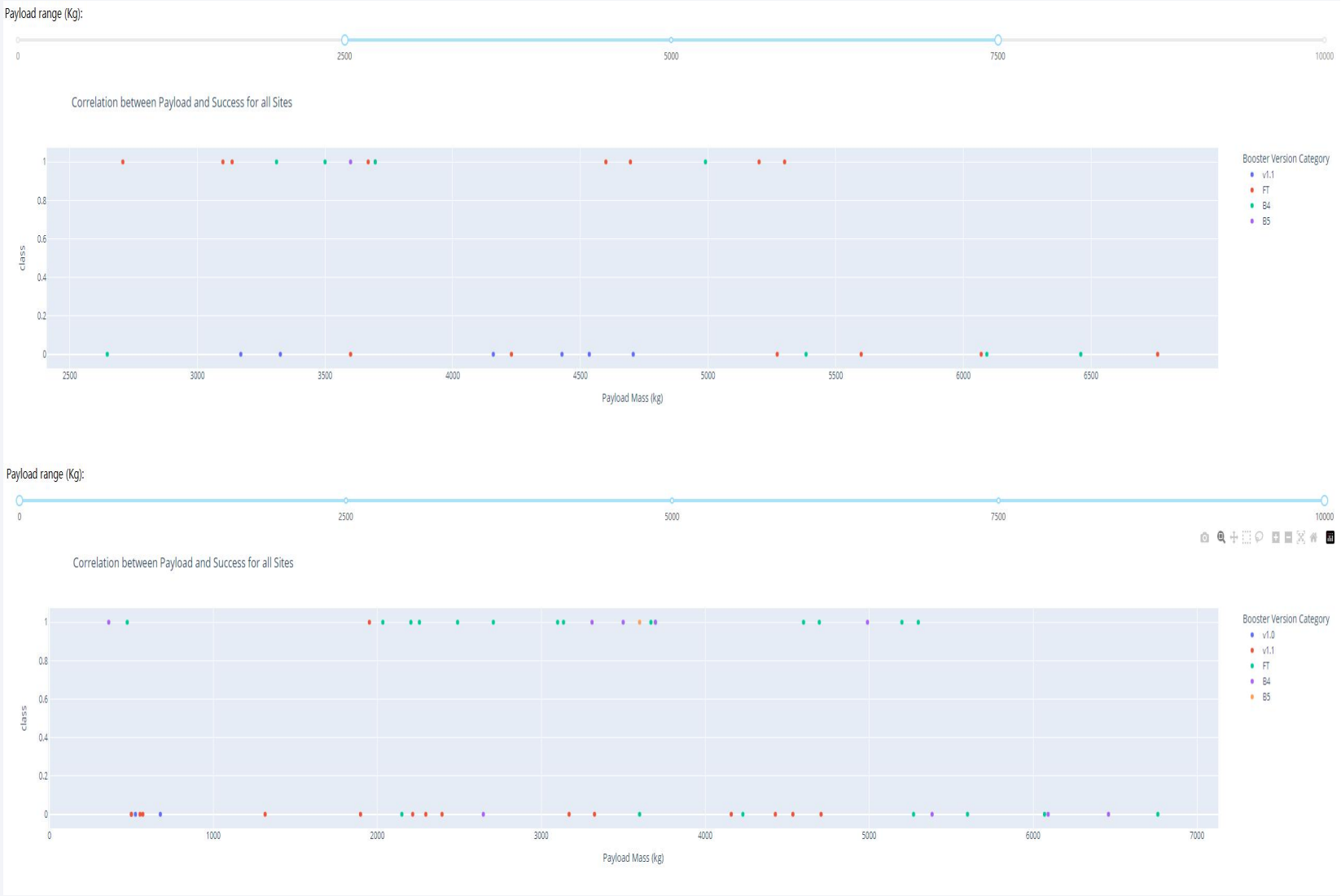
The launch site with highest launch success ratiothe launch site with highest launch success ratio



- The launch site with highest launch success ratiothe launch site with highest launch success ratio is CCAFS SLC-40

Payload vs. Launch Outcome scatter plot for all sites

- Using this functionality can easily compare the success rate for different mass group

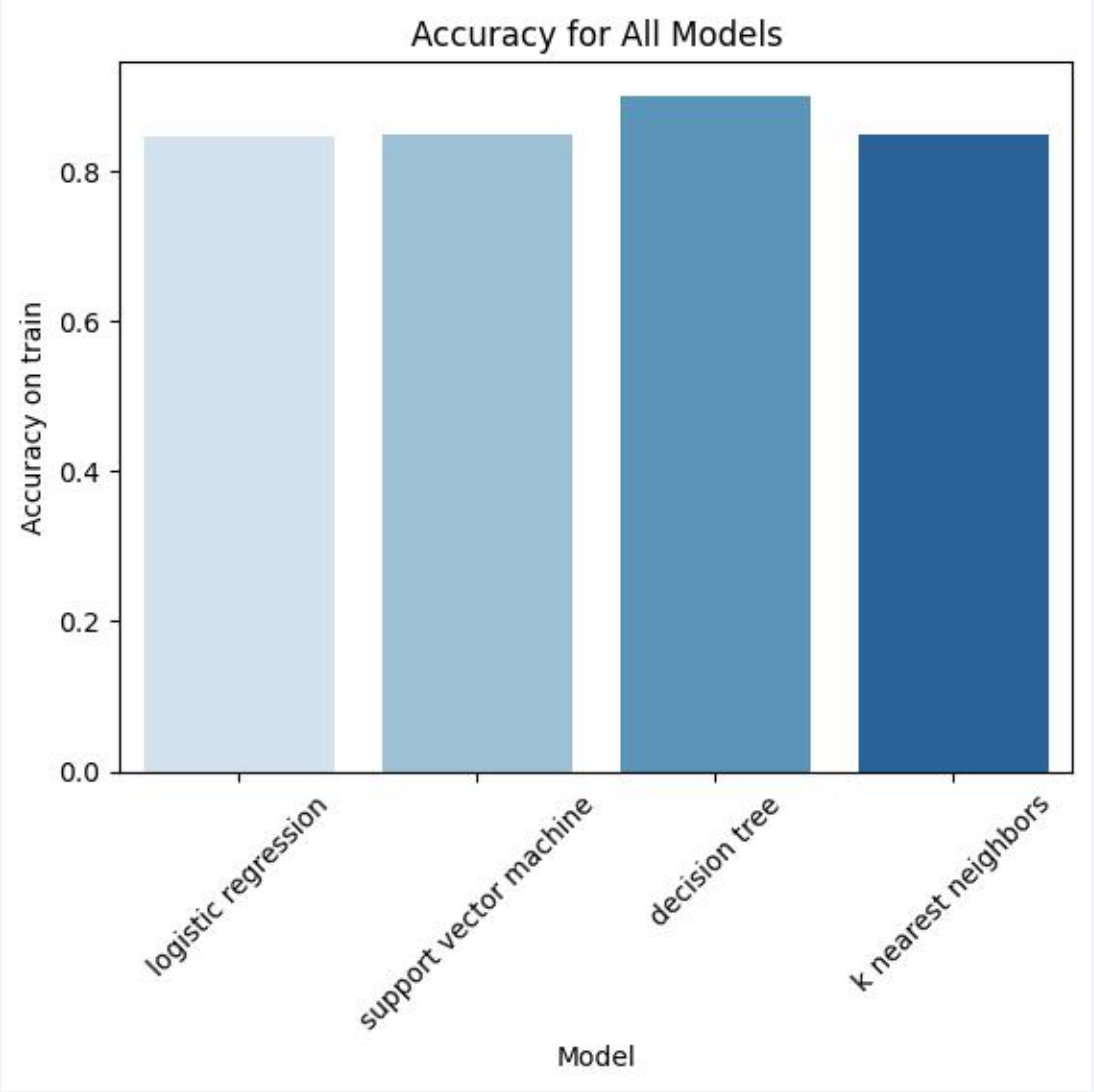


Section 5

Predictive Analysis (Classification)

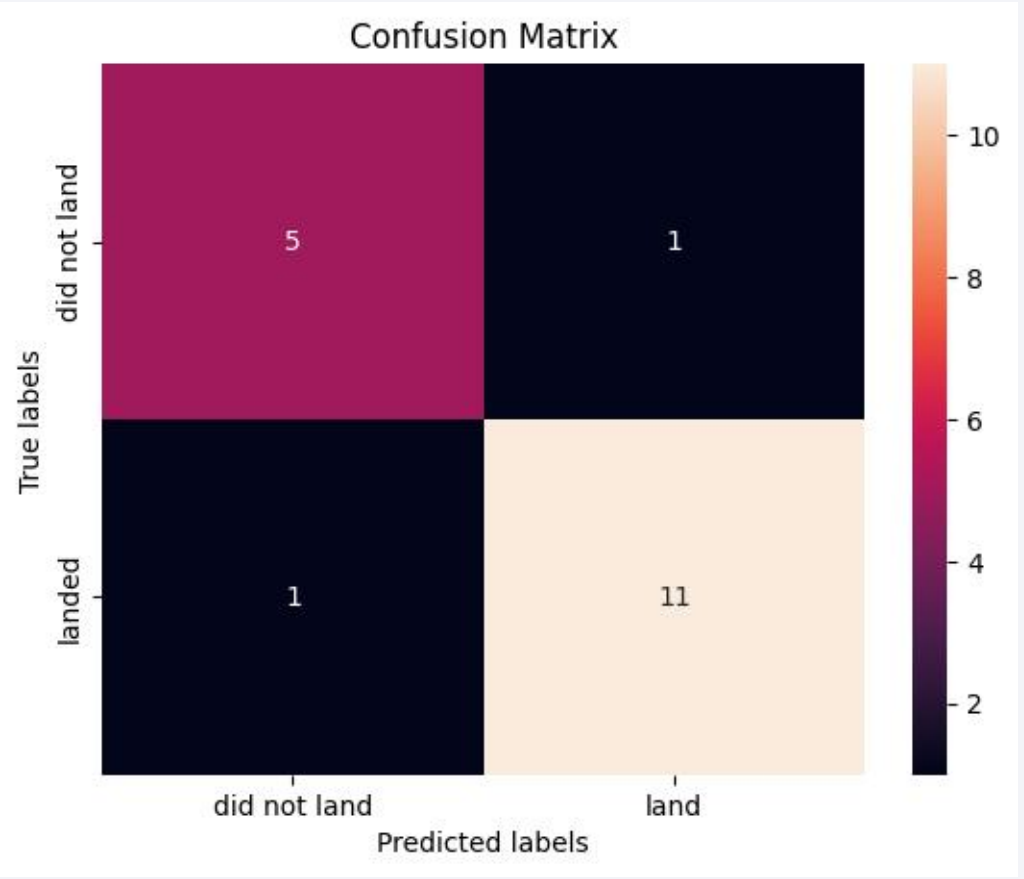
Classification Accuracy

- The decision tree model has the best accuracy.



Confusion Matrix

- In the confusion matrix, we can see only 2/18 cases are labeled incorrectly.



Conclusions

- AS the number of flights increase, the success rate increases
- There are some Orbit has 100% success rate with relatively small number of samples
- KSC LC-39 has the most count of success flights
- CCAFS SLC-40 has the highest success rate
- Best model in this case is decision tree model

Appendix

- Thank you for all the code templates and labs provided by IBM teaching team

Thank you!