IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>WAYNE LI
<Date>2025/1/12

Outline

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

Executive Summary

- Summary of methodologies
  - Data import by API,webscrap
  - Data wrangling, preprocessing using pandas
  - Exploratory data analysis using SQL & matplotlib/seaborn
  - Display geographic data by folium
  - Create dashboard by dash
  - Create and evaluate machine learning category model by sklearn
- Summary of all results

Introduction

- Project background and context:
  - In this project, the main purpose is to predict if the first stage of spaceship can land successfully, which can be reused to reduce the overall cost. This is the main reason why spacex can stand out from others and dominate the market. To complete with it, we have to find out the reason and learn how we could achieve it.
- Problems you want to find answers
  - The factors affect the success rate of landing.
  - If this relate to the launch location or some geographic factors
  - If we want to use machine learning to predict success landing or not, which model would be the most suitable

Section 1

Methodology

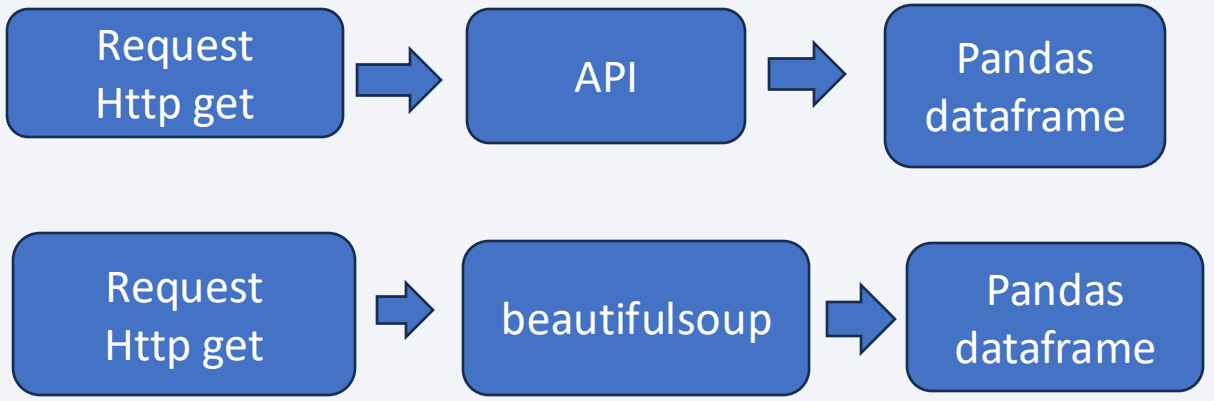
Methodology

Executive Summary

- Data collection methodology:
  - Use requests methods to send a http get command to gather information from spacex API
  - Use requests & Beautifulsoup methods to webscrap data from wikipedia table
- Perform data wrangling
  - Use pandas replace method to deal with the NAN value
  - Use one hot encoding get_dummies to make data suitable for model prediction
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Use sklearn packages to build logistic regression,SVM, decision tree,knn and use accuaracy_score,f1_score,jaccard_score to evaluate efficiency

Data Collection

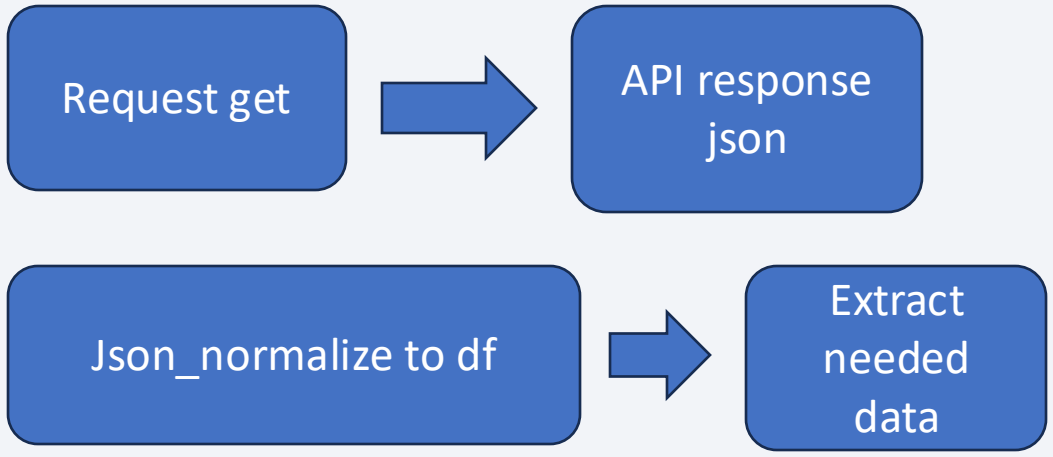
- Describe how data sets were collected.
  - Use requests get method to get spacex data by API json file and use pd.json_normalize()to read data into dataframe
  - Use requests get method and beautifulsoup to get wikipedia spacex html structure and parse html table by webscrap
- You need to present your data collection process use key phrases and flowcharts
  - Request get/API
  - Webscrap



Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

- Add the GitHub URL of the completed SpaceX API calls notebook (must include completed code cell and outcome cell), as an external reference and peer-review purpose

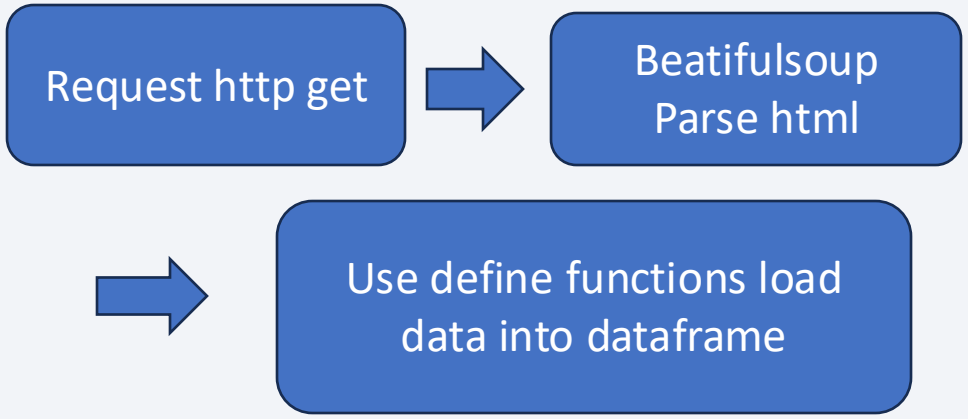


Github url:
https://github.com/Wayne0401/spacex/blob/main/spacex%20part1.ipynb

Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts

- Add the GitHub URL of the completed web scraping notebook, as an external reference and peer-review purpose



Github url:
https://github.com/Wayne0401/spacex/blob/main/spacex%20web%20scrape%20collect%20data.ipynb

Data Wrangling

- Describe how data were processed
  - Filter the data to include only falcon9 data using pandas
  - Replace nan in payload_mass with mean() using pandas
  - Summarize how many launch_site,orbit,mission type are there in the data and the number of each category using value_counts
  - Create a column name class if success:value=1 if not value=0
- You need to present your data wrangling process using key phrases and flowcharts
  - Include only falcon9
  - Deal with missing values with mean
  - Summarize occurrence
  - Create outcome of success and failure

Github url:
https://github.com/Wayne0401/spacex/blob/main/spacex%20data%20analysis.ipynb

- Add the GitHub URL of your completed data wrangling related notebooks, as an external reference and peer-review purpose

EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts

  Compare relationship between variables:

  Category plot(classify by success or failure) :

  - payload mass vs flightnumber
  - Launchsite vs flightnumber
  - Launchsite vs payload mass
  - Orbit vs flightnumber
  - Orbit vs payloadmass

Barplot:
Compare which orbit have higher success rate

Linechart:
Show trendline for success rate over the yesrs

Github url:
https://github.com/Wayne0401/spacex/blob/main/EDA.ipynb

- Add the GitHub URL of your completed EDA with data visualization notebook, as an external reference and peer-review purpose

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
  - Find distinct launch site
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first succesful landing outcome in ground pad was acheived.
  - list the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - total number of successful and failure mission outcomes
  - names of the booster_versions which have carried the maximum payload mass.
  - display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
  - Rank the count of landing outcomes

Github url: https://github.com/Wayne0401/spacex/blob/main/spacex%20sql.ipynb

- Add the GitHub URL of your completed EDA with SQL notebook, as an external reference and peer-review purpose

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
  - Marker,circle of nasa
  - Marker,circle of launchsiite
  - Markers of success and failure location
  - Markercluster to group the markers of success and failure
  - Marker of airport, city,railway
  - Polyline to airport,city,railway

Github url:
https://github.com/Wayne0401/spacex/blob/main/spacex%20folium.ipynb

- Explain why you added those objects
  - To see the location of launchsite, landing area, distances more easily from the map and clusters
- Add the GitHub URL of your completed interactive map with Folium map, as an external reference and peer-review purpose

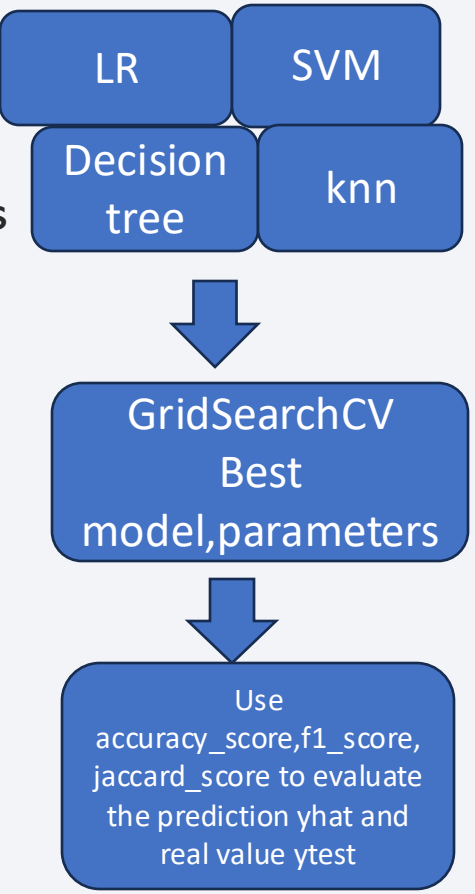
Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
  - Pie plots show the success rate for all launchsite and success and failure for each launchsite repectively depend on. Dropdown values
  - Scatter plot show the success rate vs payload mass control by the range of payload mass and dropdown launchsite values
- Explain why you added those plots and interactions
  - To see how each launchsite performs and see if launchsite will affect the success rate
  - To see how payloadmass affect success result and see if there is a difference between launchsite
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

Github url: https://github.com/Wayne0401/spacex/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

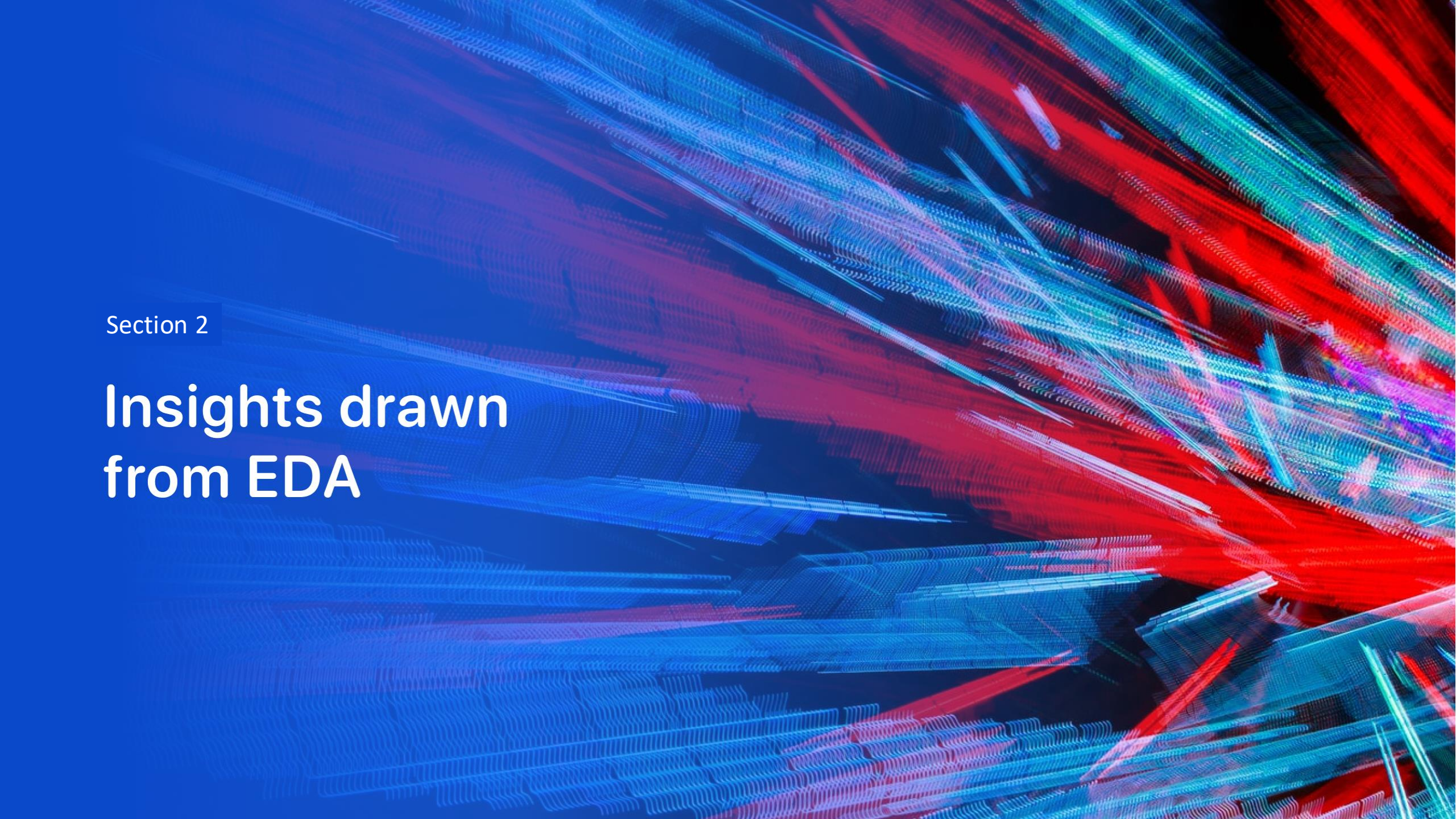
- Summarize how you built, evaluated, improved, and found the best performing classification model
  - Use logistic regression, svm, decision tree, knn machine learning model to predict success or not
  - Use gridsearchcv to find the best parameters for the models and the best model
  - Use accuracy_score,f1_score,jaccard_score to evaluate the efficiency and accuracy of the models
- You need present your model development process using key phrases and flowchart
- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose



Github url:
https://github.com/Wayne0401/spacex/blob/main/machine%20learning.ipynb

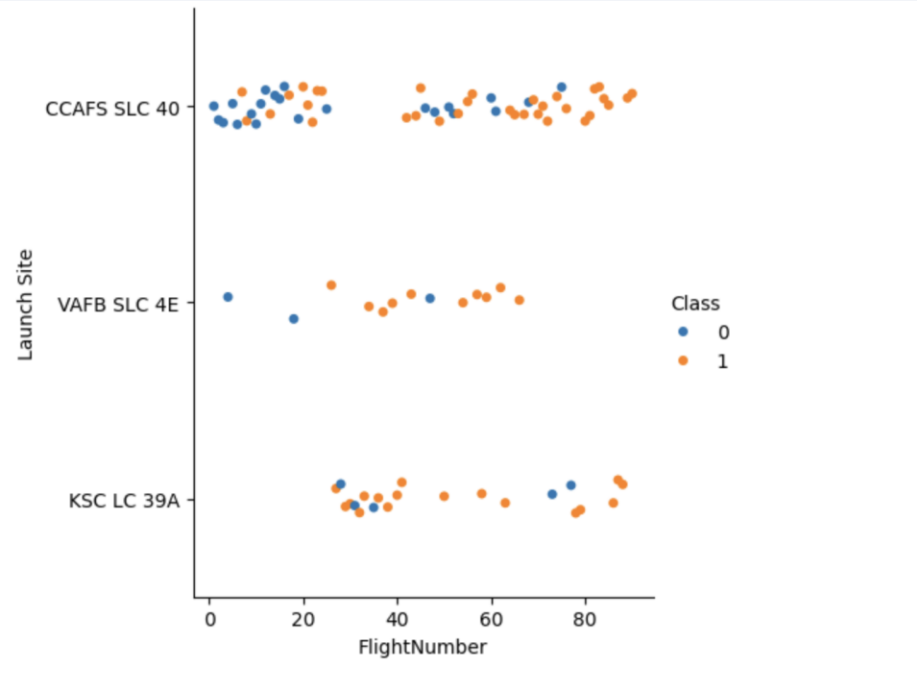
Results

- Exploratory data analysis results
  - There are three launch location in the dataset
  - Of all 101 data, 98 record have successful mission outcome
  - CCAFS SLC 40 has the most launch
  - Most rocket destination are GTO and ISS orbit,GTO is used for weather monitor and ISS is used for space station
  - The most frequent landing outcome is True ASDS(successful landing on drone ship)
  - The success rate might increase as flight number increases(more experience) and payload mass increases
  - The success rate of landing increase dramatically since 2013
  - ES-L1,GEO,HEO,SSO orbit have highest landing rate, but this depends a lot on launch times, and VLEO performs best with landing rate of 0.8 third highest launch times
- Interactive analytics demo in screenshots
- Predictive analysis results
  - Use many ways such as accuracy_score,f1_score,jaccard_score all show the same evaluation values. This means they are all suitable to predict success and have the same accuracy.


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

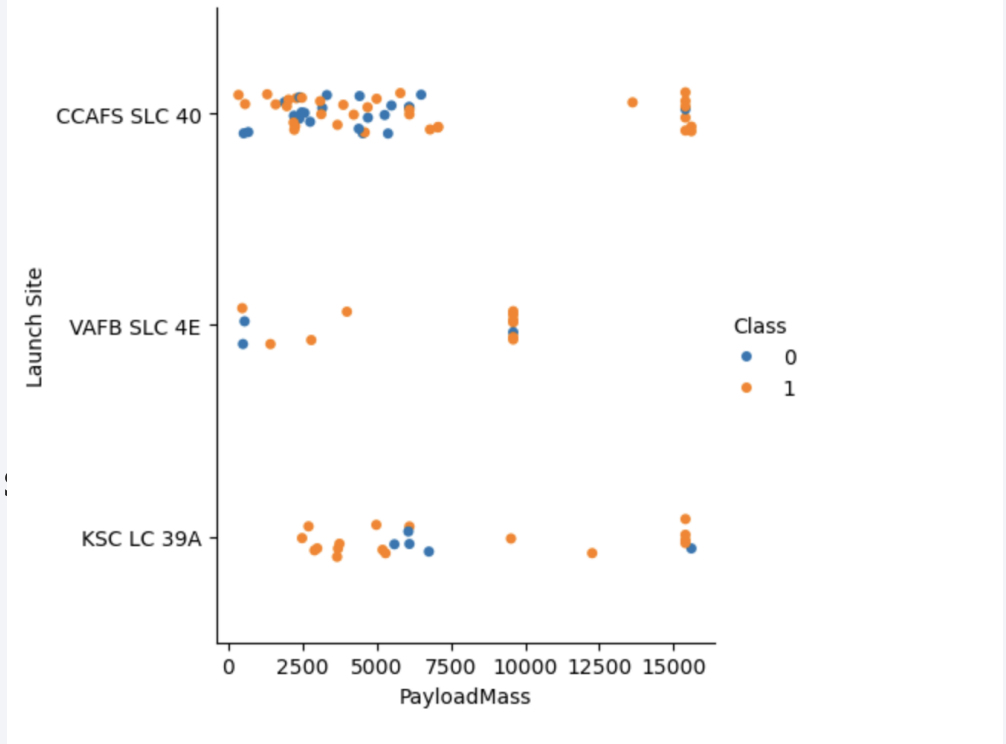
- Show a scatter plot of Flight Number vs. Launch Site
- Show the screenshot of the scatter plot with explanations



Explanation:
As the picture shows, although CCAFS SLC 40 have a little bit unclear, VAFB SLC 4E and KSC LC 39A launchsite show higher flight number more experience cause higher success rate

Payload vs. Launch Site

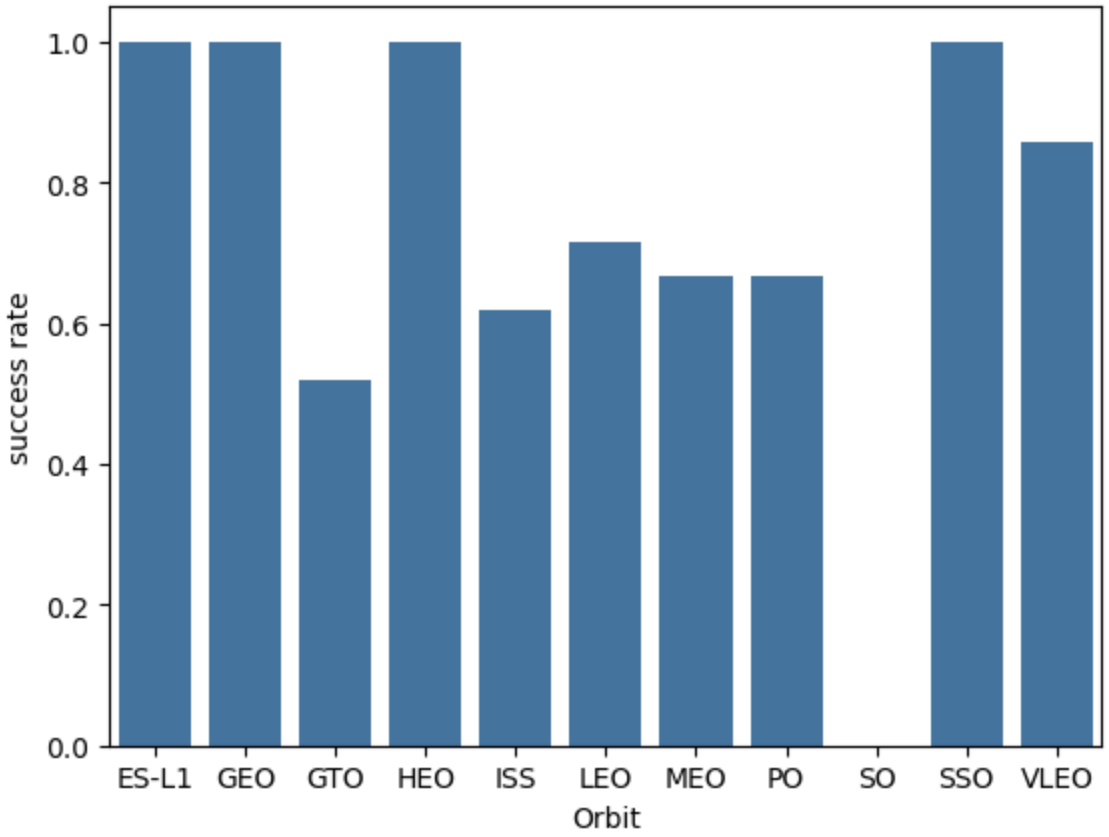
- Show a scatter plot of Payload vs. Launch Site
- Show the screenshot of the scatter plot with explanations



Explanation:
No matter which launchsite all shows higher payload do affect success rate(higher)

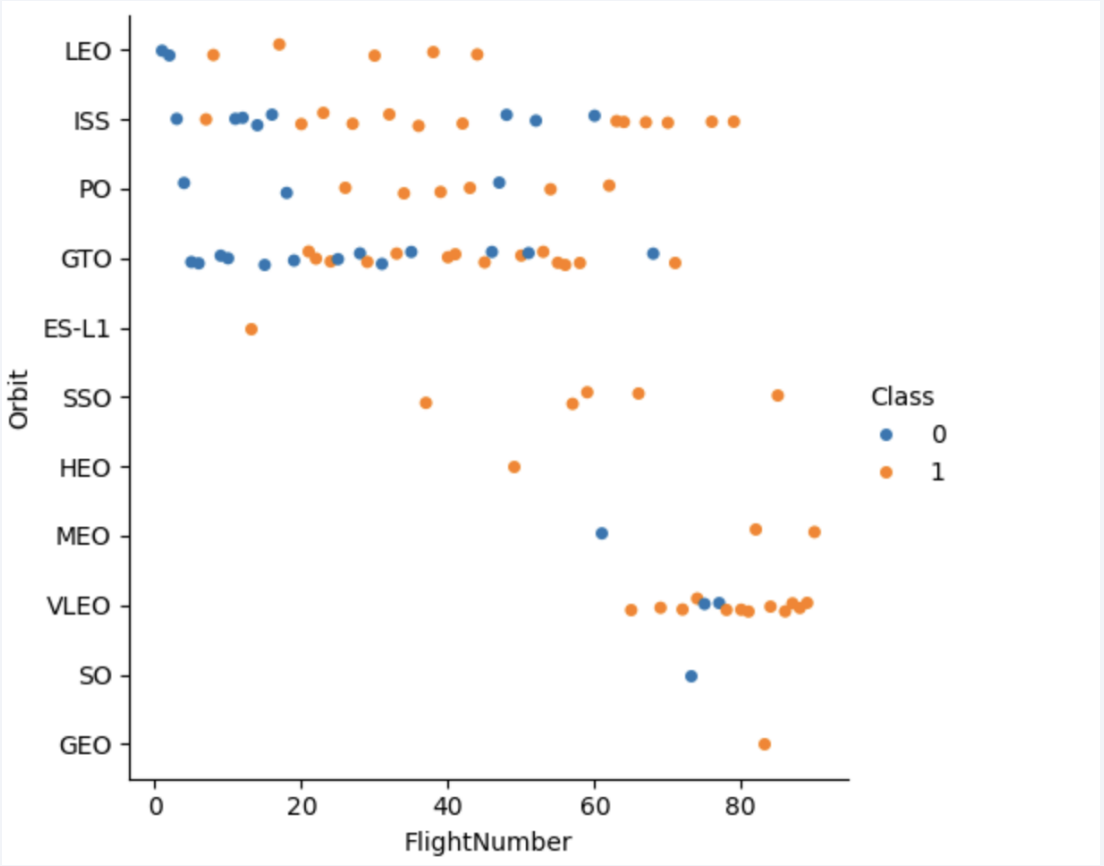
Success Rate vs. Orbit Type

- Show a bar chart for the success rate of each orbit type
- Show the screenshot of the bar plot with explanations
- Explanations:
  - From the graph, it seems that ES-L1, GEO,HEO,SSO have highest success rate, but this depends largely on the launch times ,some of them just have 1 launch time. In my opinion, VLEO have the best performance with 0.8 success rate and third highest launch times.



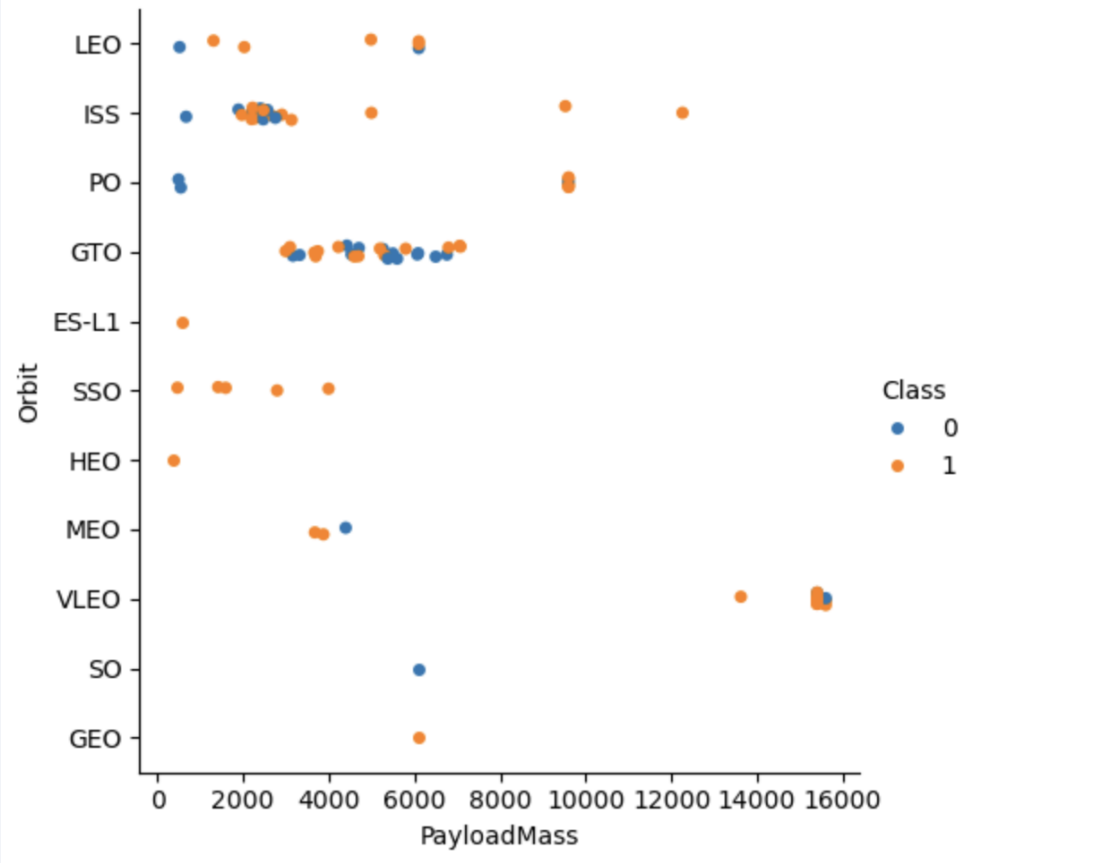
Flight Number vs. Orbit Type

- Show a scatter point of Flight number vs. Orbit type
- Show the screenshot of the scatter plot with explanations
- Explanations:
  - It seems that flight number have impact on success rate. The higher flight number end with higher success rate.



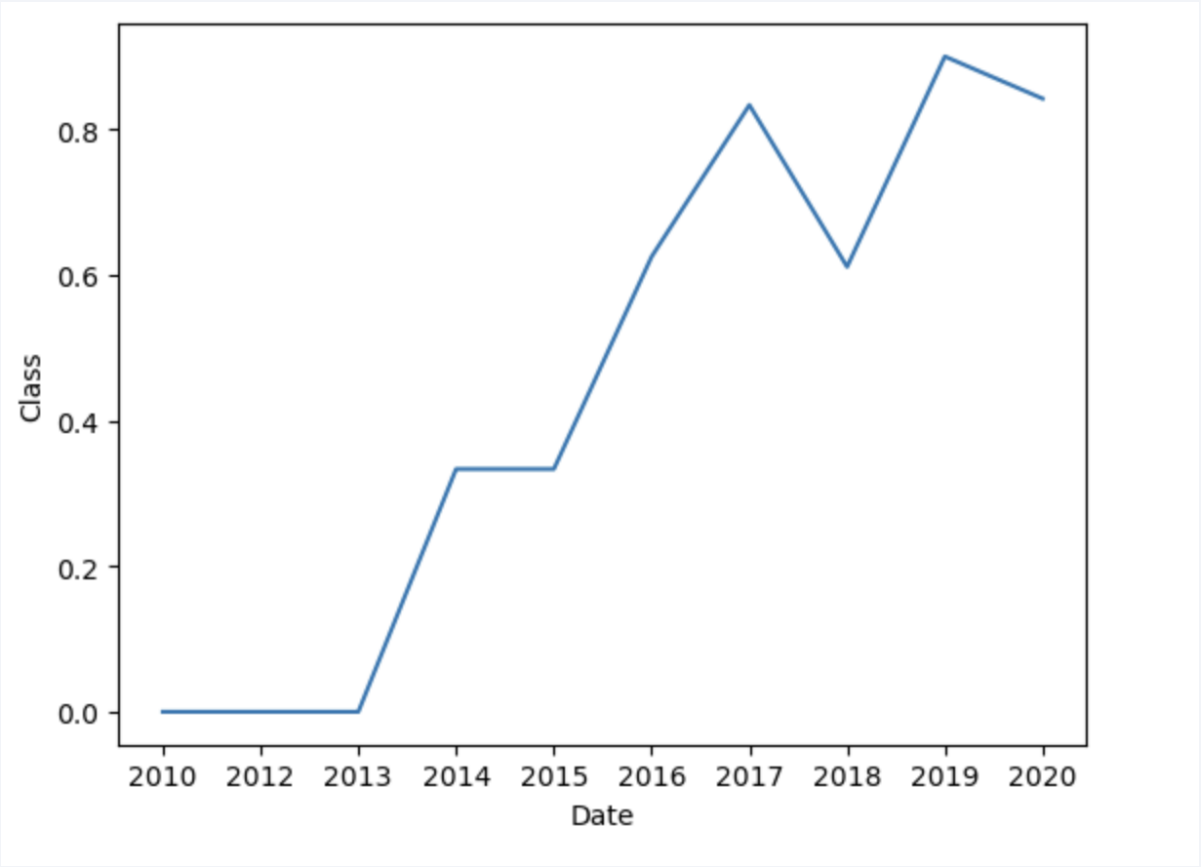
Payload vs. Orbit Type

- Show a scatter point of payload vs. orbit type
- Show the screenshot of the scatter plot with explanations
- Explanations:
  - Higher payload mass may have higher success, although some orbits like GTO might not have such phenomenon.



Launch Success Yearly Trend

- Show a line chart of yearly average success rate
- Show the screenshot of the scatter plot with explanations
- Explanation:
  - Success rate keeps increase since 2013, this might be as a result of technology breakthrough.



All Launch Site Names

- Find the names of the unique launch sites
- Present your query result with a short explanation here

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

```
%sql SELECT DISTINCT("Launch_Site") from SPACEXTABLE
```

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

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

Explanation:
From this result, it shows there are four unique launch_site in the table, the DISTINCT function will choose unique values in the column Launch_Site

Launch Site Names Begin with 'CCA'

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

Display 5 records where launch sites begin with the string 'CCA'

```
%sql SELECT * from  SPACEXTABLE WHERE "Launch_Site" LIKE "CCA%" LIMIT 5;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	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)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Explanations:
The result shows first first rows of the rows with launch site begin with CCA. LIKE function will find something similar and LIMIT will limit the rows to display.

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- Present your query result with a short explanation here

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

```
%sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE Customer='NASA (CRS)'
```

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

SUM("PAYLOAD_MASS__KG_")
45596

Explanations:
The result will find the NASA data and add all the payload mass information together.

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Present your query result with a short explanation here

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

```
%sql SELECT AVG("PAYLOAD_MASS__KG_") from  SPACEXTABLE where Booster_Version LIKE 'F9 v1.1%';
```

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

AVG("PAYLOAD_MASS__KG_")
2534.6666666666665

Explanations:
This similar to find launchsite begin with cca.
Use the LIKE find booster version F9 v1.1
then count the average payload mass

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Present your query result with a short explanation here

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

Hint:Use min function

```
%sql SELECT MIN(Date) FROM SPACEXTABLE WHERE "Landing_Outcome"="Success (ground pad)"
```

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

MIN(Date)
2015-12-22

Explanations:
This will find the success outcome on ground pad and choose the earliest date by using the MIN function on Date

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

- Present your query result with a short explanation here

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 SPACEXTABLE WHERE "Landing_Outcome"="Success (drone ship)"and (PAYLOAD_MASS__KG_ between 4000 and 6000)
```

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

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

Explanations:
This will find booster version which has success on drone ship and payload mass between 4000 and 6000.

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Present your query result with a short explanation here

List the total number of successful and failure mission outcomes

```
%sql SELECT COUNT(*) from SPACEXTABLE WHERE Mission_Outcome='Success';
```

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

COUNT(*)
98

```
%sql SELECT COUNT(*) from SPACEXTABLE WHERE Mission_Outcome='Failure';
```

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

COUNT(*)
0

Explanations:
The result shows of the total 101 data records 98 have successful missions,which is a good news.

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Present your query result with a short explanation here

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

```
%sql select Booster_Version     from SPACEXTABLE where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTABLE);
```

```
 * 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

Explanations:
Use subquery to select the booster version of the max payload mass. This could avoid error since SQL cannot write aggregate functions like max(),min(),AVG() directly in the where conditions.

2015 Launch Records

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

- Present your query result with a short explanation here

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.

```
%sql select  substr(Date, 6,2),Booster_Version,Launch_Site,Landing_Outcome from SPACEXTABLE where  substr(Date,0,5)='2015'and Landing_Outcome='Failure (drone ship)';
```

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

substr(Date, 6,2)	Booster_Version	Launch_Site	Landing_Outcome
01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Explanations:
You can use Year()or Month() function to select year or month of the date in database,but here it says the function cannot work in jupiterlite.

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

- Present your query result with a short explanation here

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(*) from  SPACEXTABLE  group by Landing_Outcome having date between '2010-06-04' and '2017-03-20' order by count(*) DESC;
```

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

Landing_Outcome	count(*)
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Explanations:
Use GROUP BY to group result together by landing outcomes and use having to give conditions then use ORDER BY columnname DESC to order in descending order.

Section 3

Launch Sites Proximities Analysis

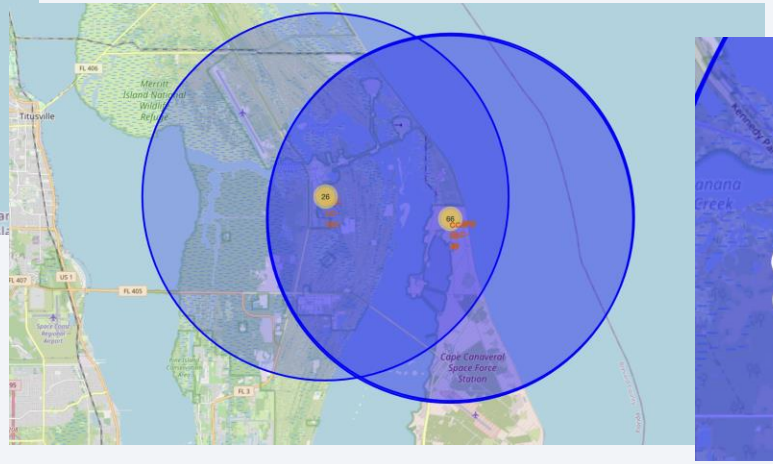
<Launch site location on map>

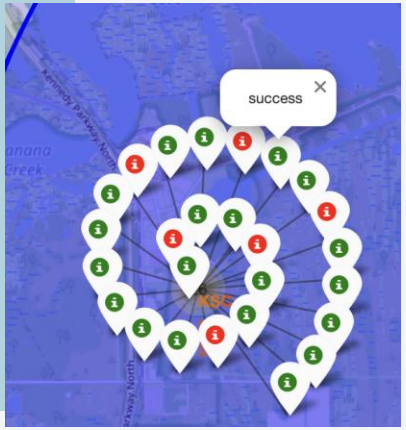
Explanations:
This is the launch site map of the spacex falcon9. There are four sites in total. One at the west coast,and the other three on the east.One at california and the other three at Florida.All of them are along with the coast line.

<success and failure markercluster>

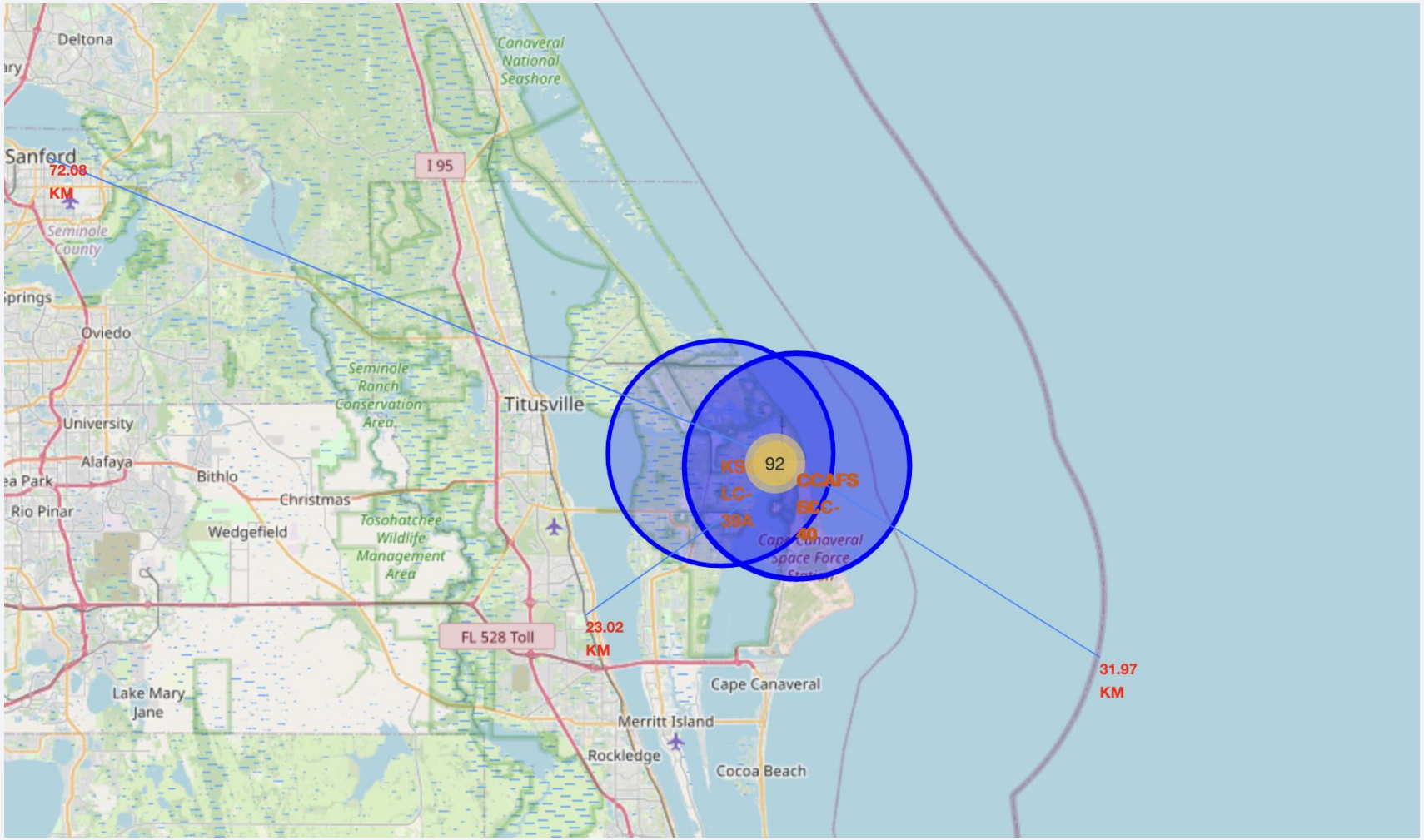
Explanations:

To express the failure and success location more easily. I use the markercluster to the map.This will group the marker together as you zoom in and out the map.The cluster with separate whem zoom in. And it will also shows the success landing result in green marker, failure in red marker.

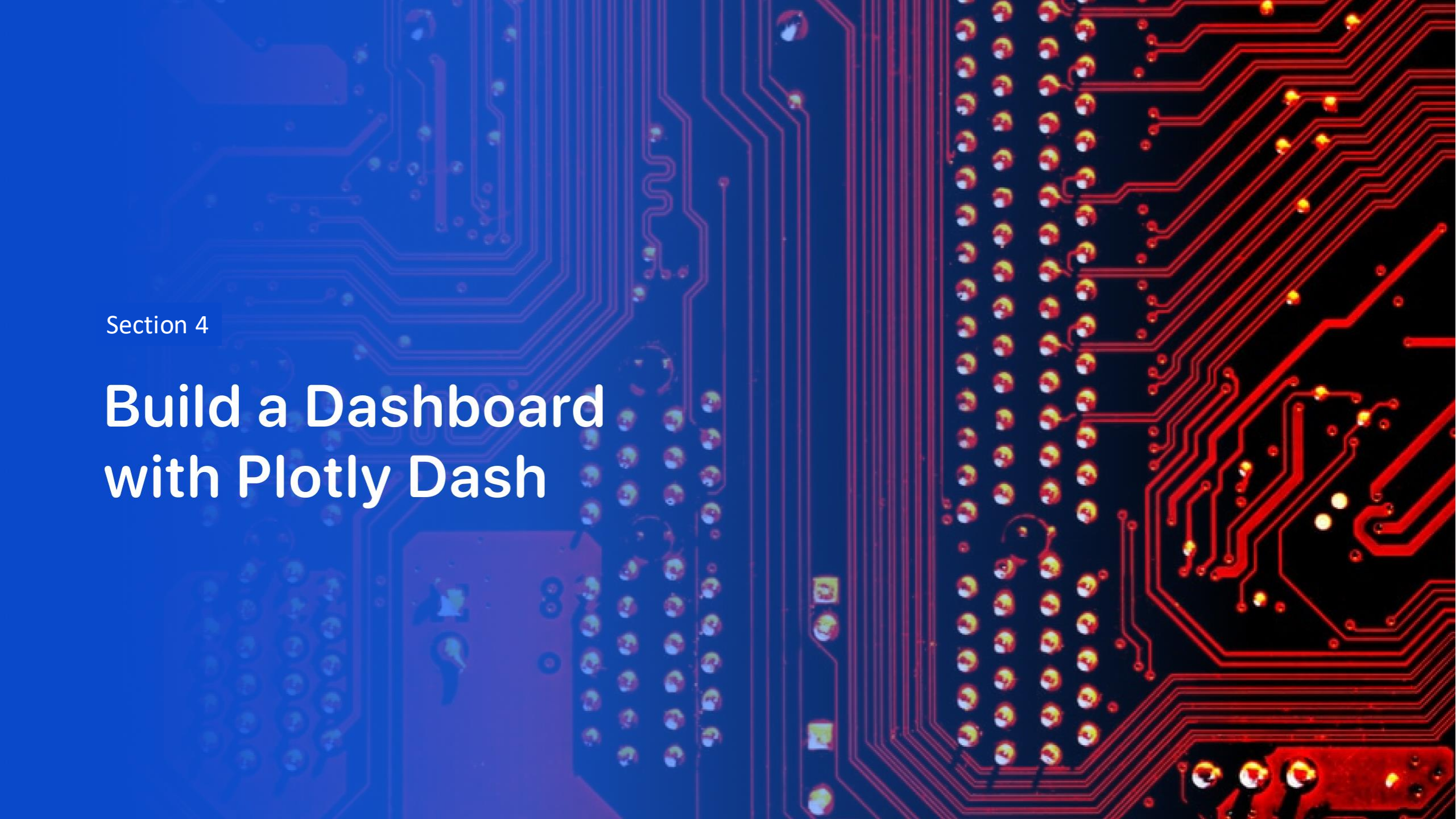




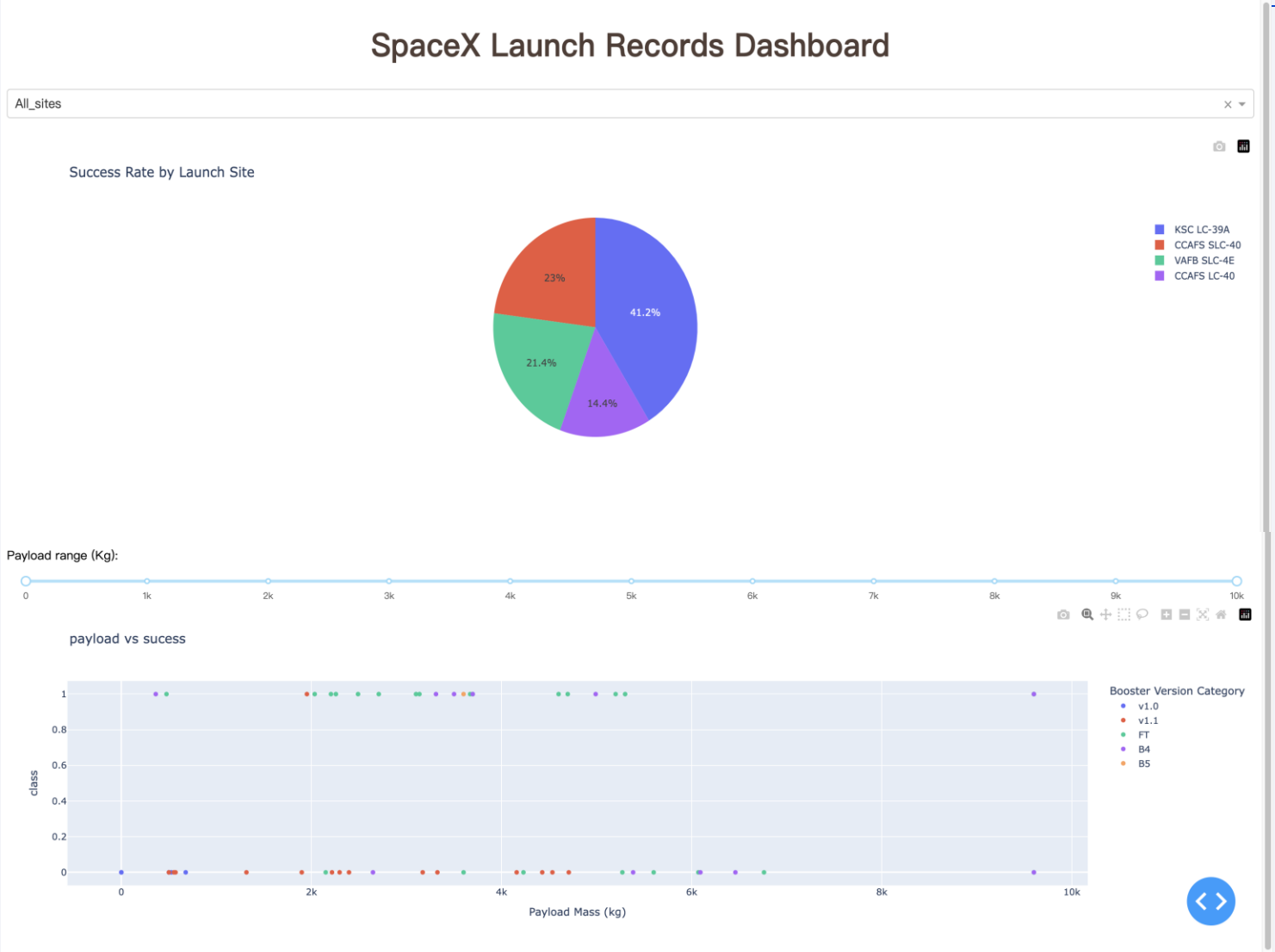
<Distance between locations>



Explanations:
I use the mouseposition method to the the latitude and longtitude of the location. Then define function to marker the location,such as cities,coastline, railway and count the distance between them and launch site. Finally, use the polyline(the blue line in the map) to connect the two locations.


Section 4
Build a Dashboard with Plotly Dash

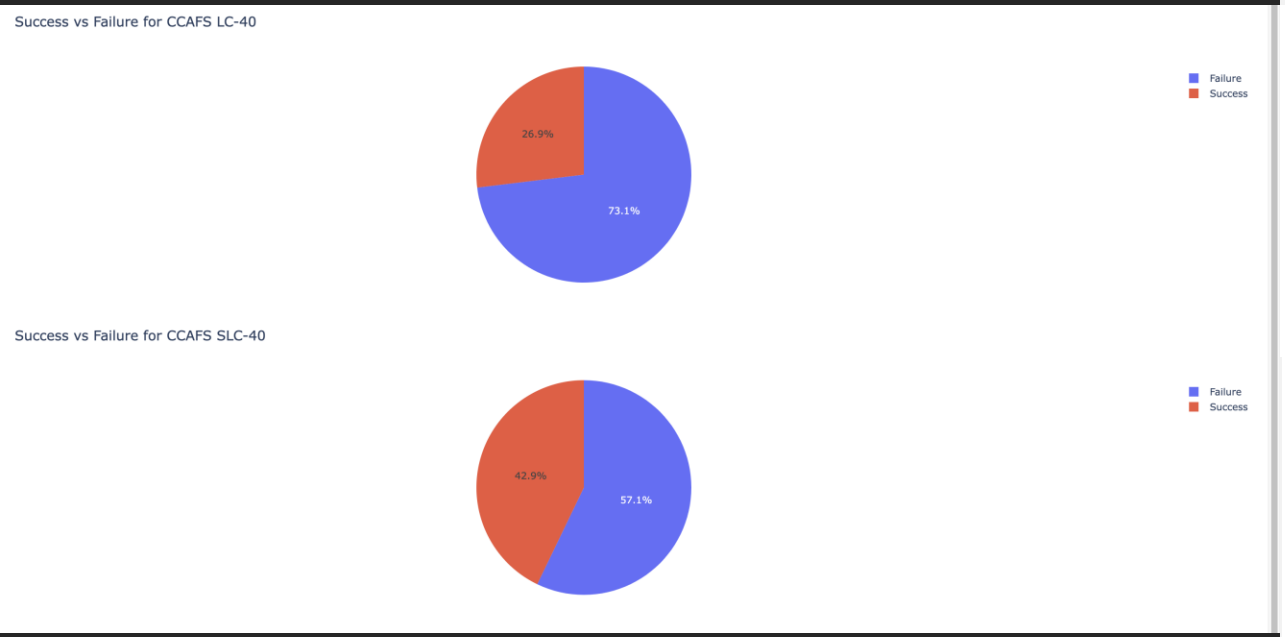
<success rate for all sites and payload mass>

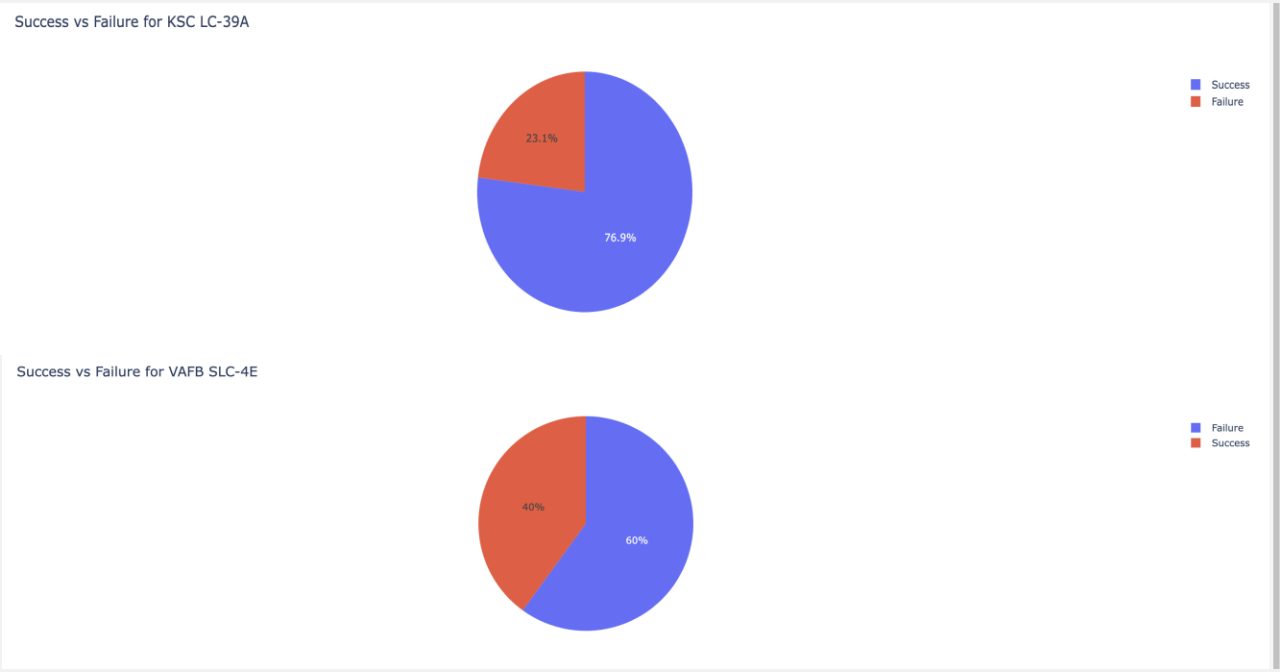


Explainations:
This is a dashboard which shows the pie chart of success rate and the scatter plot of the payload mass and class (success=1,failure=0) . There are two element that you can choose, dropdown to choose the site and range slide to choose the range of payload. The payload graph will change simutaneouly as the site change. From the graph(all site), we see that KSC LC-39A has the highest success rate,follow by CCAFS SLC-40. For the payload mass, we see that the success occur between 2000 and 6000 and it seems that payload mass and success do not have positive relationship.

<success rate for each site respectively>

Explanations:
As the plots show, the site with the highest success rate is KSC LC-39A near 80% launch is successful, the least is CCAFS LC-40. This might be caused by more using frequency for testing which can be seen in next slide, use drone ship rather than solid ground, or the mission difficulties(farer orbit).



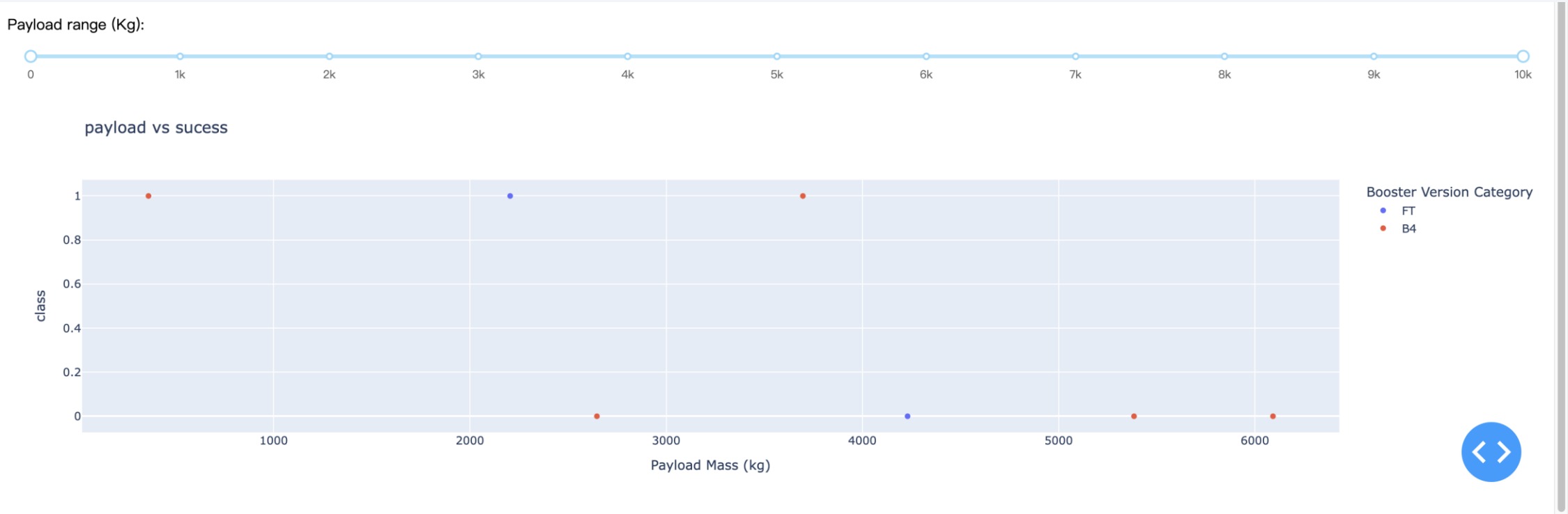


<success, failure and payload mass>

CCAFS LC-40:

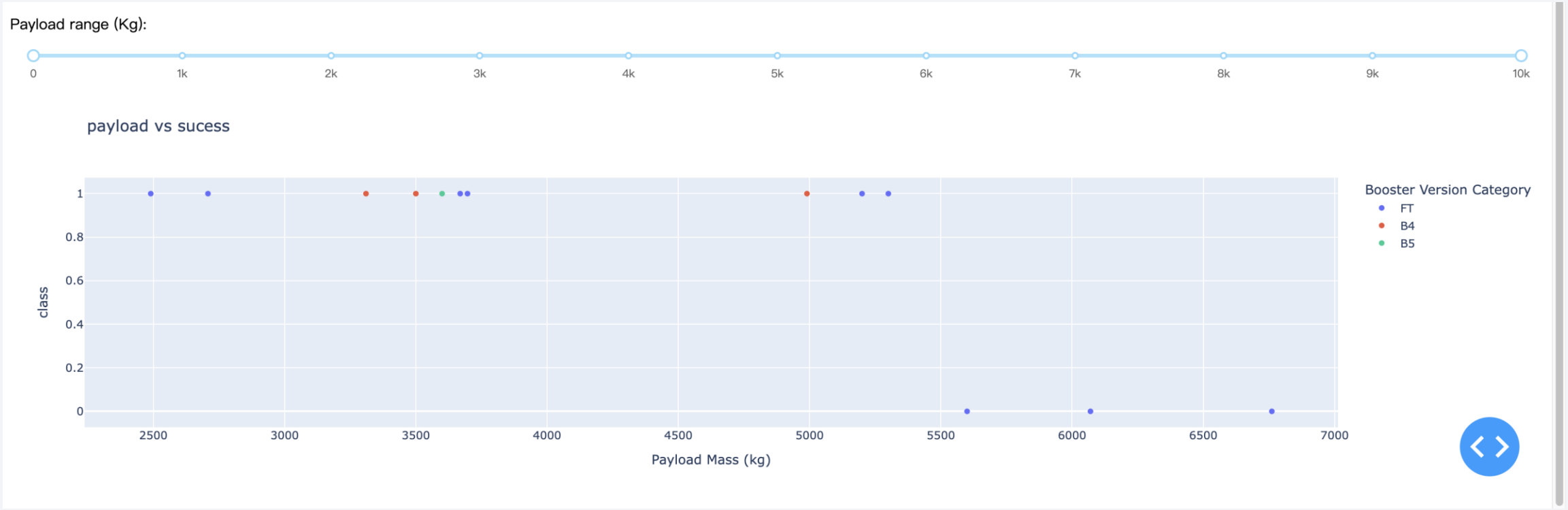
<success, failure and payload mass>

CCAFS SLC-40:



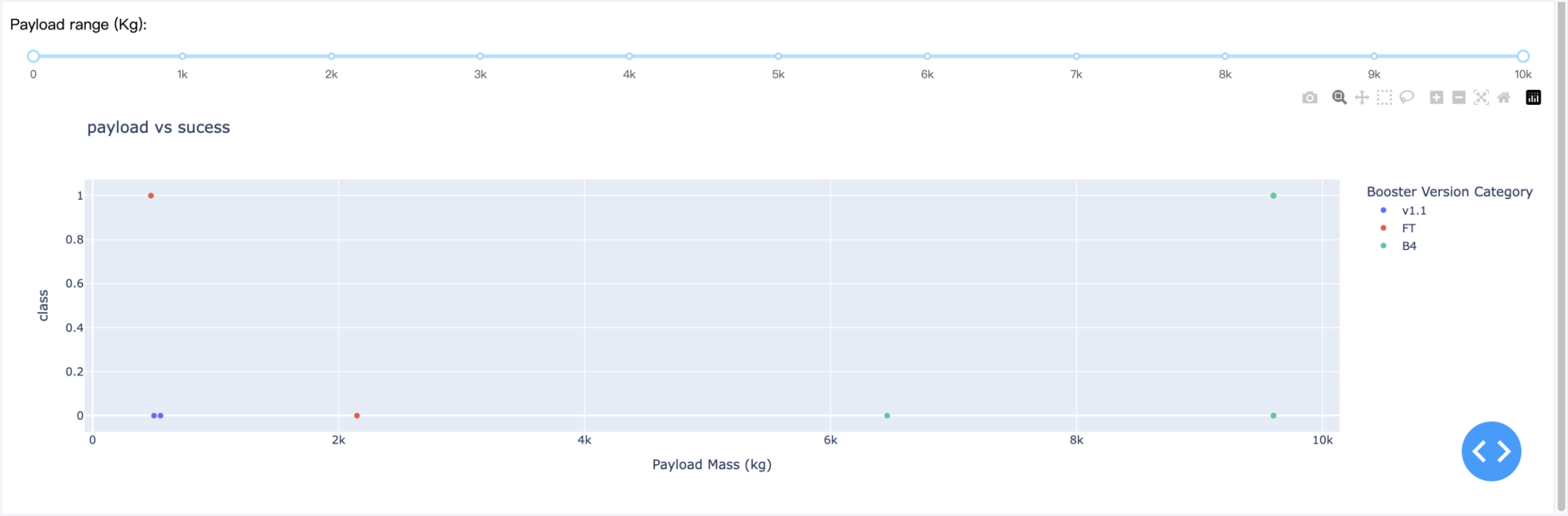
<success, failure and payload mass>

KSC LC-39A



<success, failure and payload mass>

VAFB SLC-4E:



<success, failure and payload mass>

Summary:

1.Just as the reason I mention in the pie chart, the CCAFS LC-40 has the most launch times , this makes sense since more probability to get higher failure as the launch times increase. In addition, CCAFS LC-40 is used by spacex to test landing techology before the breakthrough happens. Other reasons may be using drone ship rather than solid ground, more difficult missions,and so on.
2.From these four payload mass plot, we can see that there are just a few launch with payload mass higher than 6000, and the data above these range shows high failure rate.CCAFS SLC-40,KSC LC-39A,VAFB SLC-4E data points higher than 6000 rarely success. These graph shows success below 5000.It seems that the most suitable range is 2000-5000.This means higher payload might increase landing difficulties. Maybe is because the fuel. Higher payload mass needs higher energy to launch, and it casues not enough fuel to land at particular location.
3.Booster version difference will also affect the result.The success event happens more frequently when the rocket is equipped with FT Booster version.

Section 5

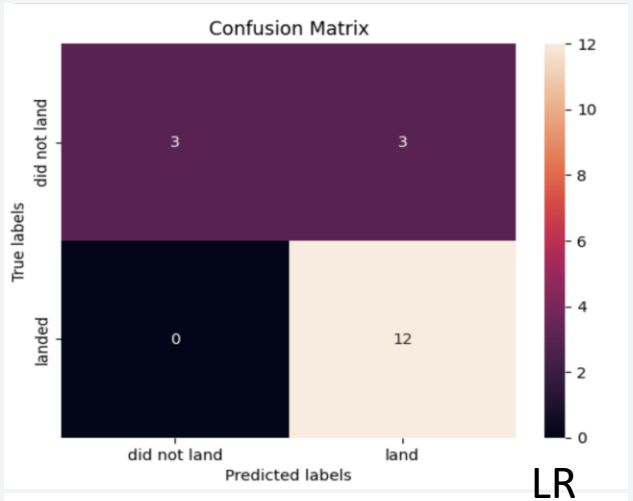
Predictive Analysis (Classification)

Classification Accuracy

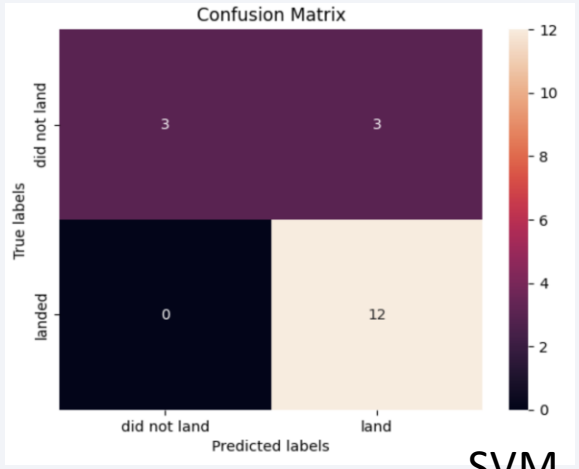
	accuracy_score	f1_score	jaccard_score
lr	0.833333	0.888889	0.8
svm	0.833333	0.888889	0.8
decision_tree	0.833333	0.888889	0.8
knn	0.833333	0.888889	0.8

Explanations:
I use gridsearchCV to find the best model for Logistic Regression, SVM,decision tree,knn category machine learning model, and I split the data into training and testing data. The training part is used to fit the model, and after training I use test data to test the accuracy. I use accuracy_score,f1_score and jaccard_score to evaluate the model. This is the result of these tests. They perform equally in these three methods, which means the result will be very close.I can choose one of them to predict the landing outcome.

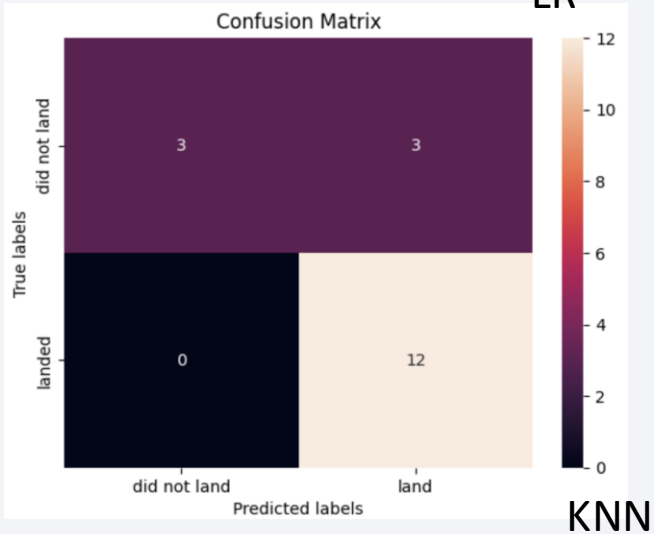
Confusion Matrix



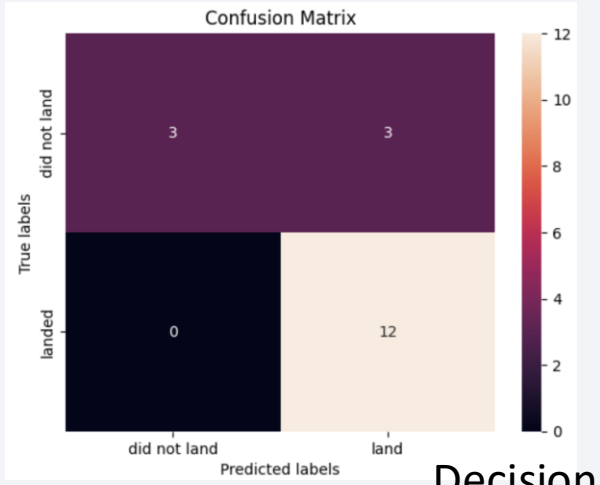
LR



SVM



KNN



Decision tree

Explanations:
The result is the same by using confusion matrix. They all have 3 TN,12TP, and 3FP. This shows they have same accuracy.

Conclusions

- The Flight number(experience) might cause success rate for landing to rise
- The payload mass might increase landing success might not according to the plot in exploratory data analysis process and dash board plot
- The CCAFS LC-40 launch site have least success rate because the launch times, more difficult missions, weather or use drone ship.
- The locations of the launch sites often are at somewhere near the ocean to for safety concerns, as the falling position is easier to control and less people will be affected. Also, it is also good for transporting equipment.
- The booster version and orbit type(difficulties) will also affect the successful rate
- Logistic Regression, SVM, decision tree, KNN models are all suitable for predicting landing success or not(category)

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project
- Data import, data wraggling:
  - Python(pandas,numpy,request,Beautifulsoup)
- Data visualize and exploratory data analysis:
  - Python(matplotlib,seaborn),SQL
- Map,dashboard:
  - Python(folium,dash,plotly)
- Machine learning:
  - Python(sklearn)
- Data source: spacex API, wikipedia tables

Thank you!