IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Yuki Nakayama>
<1st,MAR,2024>

in process

Outline

- EXECUTIVE SUMMARY
- INTRODUCTION
- METHODOLOGY
- RESULTS
- CONCLUSION
- APPENDIX

Executive Summary

Summary of methodologies

1. Data Collection:
2. Data Preprocessing:
3. Exploratory Data Analysis (EDA):
4. Feature Engineering:
5. Model Selection:
6. Model Training:
7. Model Evaluation:

Summary of all results

･Developed predictive models to forecast the success or failure of future SpaceX launches based on historical data, achieving a high level of accuracy and reliability. Utilized machine learning

･algorithms such as logistic regression and random forests to predict launch outcomes and assess the likelihood of mission success.

Introduction

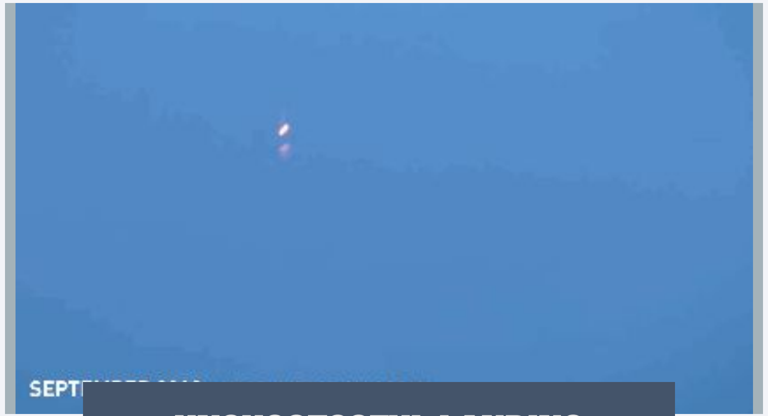
Project background and context

Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch.

Problems you want to find answers

・What factors contribute to the success of landing?

SUCCESSFUL LANDING



UNSUCCESSFUL LANDING

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Accessing SpaceX data using the SpaceX API.
  - Retrieving information about launches, rockets, payloads, and other relevant data.
- Perform data wrangling
  - Cleaning and formatting the retrieved data. Handling missing values, duplicates, and outliers. Transforming data into a suitable format for analysis.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Create classification models with algorithms like logistic regression, decision trees, or random forests, fine-tune parameters, and evaluate accuracy using metrics such as accuracy and F1-score.

Data Collection

Access Data Sources

Utilize APIs to access real-time data from SpaceX, ensuring up-to-date information on launches, rockets, and payloads.

Data Cleaning and Preprocessing

Perform data cleaning to remove duplicates, handle missing values, and address inconsistencies in the dataset.

Data Integration

Merge and integrate multiple datasets obtained from different sources to create a unified dataset for analysis.

Data Validation

Validate the integrity and accuracy of the collected data by cross-referencing with trusted sources and performing data quality checks.

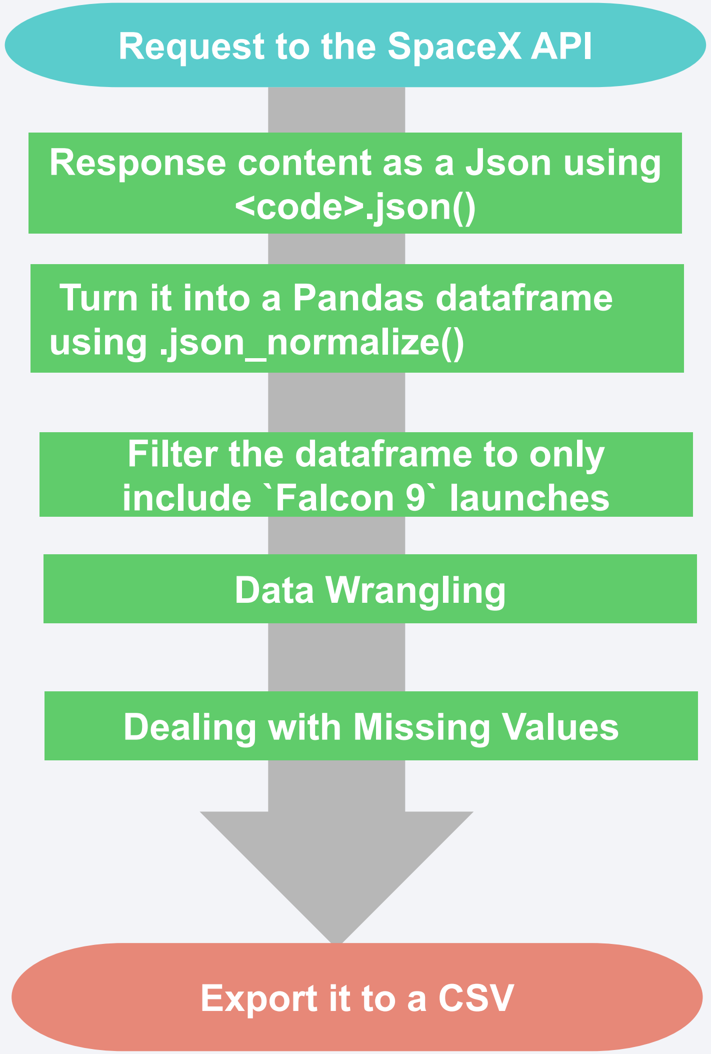
Data Collection – SpaceX API

Now let's start requesting rocket launch data from SpaceX API with the following URL:

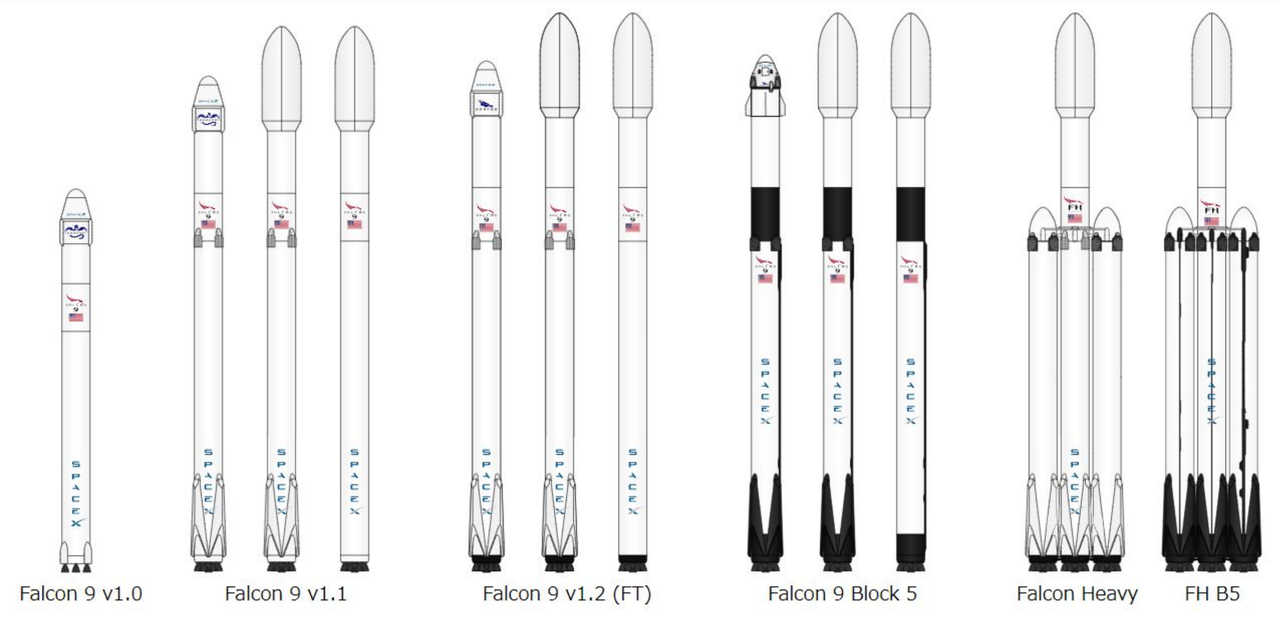
```
[6]: spacex_url="https://api.spacexdata.com/v4/launches/past"

[7]: response = requests.get(spacex_url)
```

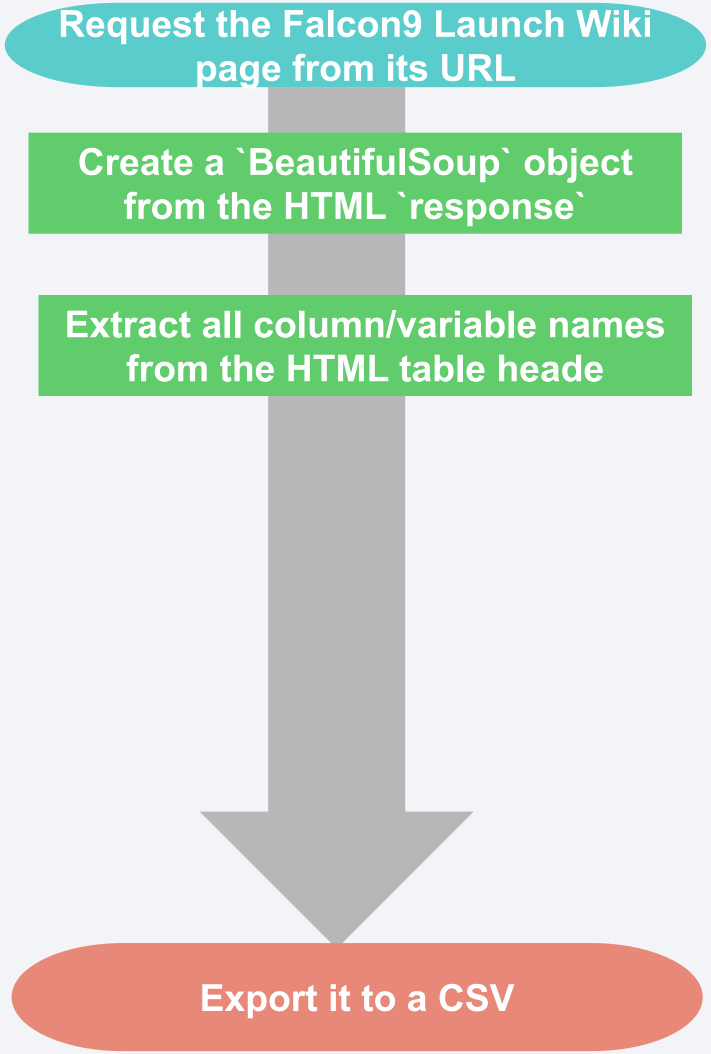
https://github.com/YukiG16/IBM-Data-Science-Certificate/blob/main/Course10_applied-data-science-capstone/Data-science-using-SpaceX-API/week1-1_jupyter-labs-spacex-data-collection-api.ipynb



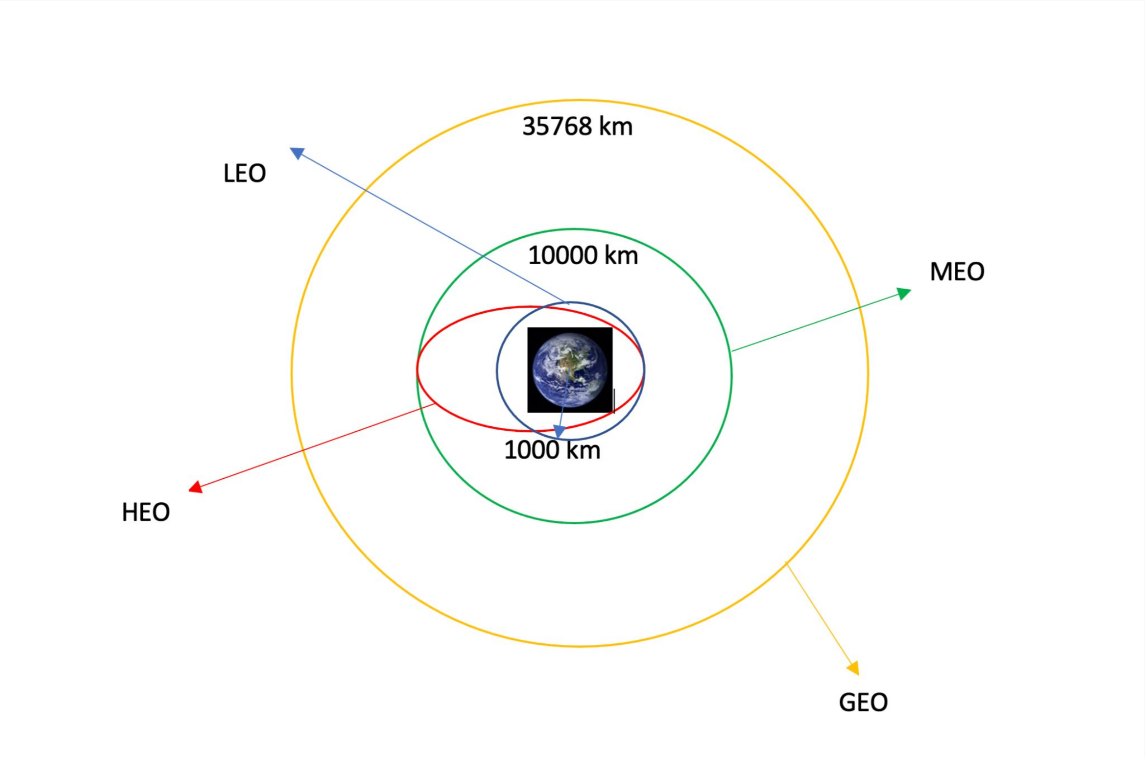
Data Collection - Scraping



https://github.com/YukiG16/IBM-Data-Science-Certificate/blob/main/Course10_applied-data-science-capstone/Data-science-using-SpaceX-API/week1-2_jupyter-labs-webscraping.ipynb



Data Wrangling



Each launch aims to an dedicated orbit

https://github.com/YukiG16/IBM-Data-Science-Certificate/blob/main/Course10_applied-data-science-capstone/Data-science-using-SpaceX-API/week1-3_labs-jupyter-spacex-Data%20wrangling.ipynb

Calculate the number of launches on each site

Calculate the number and occurrence of each orbit

Calculate the number and occurence of mission outcome of the orbits

Create a landing outcome label from Outcome column

Export it to a CSV

EDA with Data Visualization

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

This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

This dataset includes a record for each payload carried during a SpaceX mission into outer space.

https://github.com/YukiG16/IBM-Data-Science-Certificate/blob/main/Course10_applied-data-science-capstone/Data-science-using-SpaceX-API/week2-1_jupyter-labs-eda-sql-coursera_sqllite.ipynb

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed

https://github.com/YukiG16/IBM-Data-Science-Certificate/blob/main/Course10_applied-data-science-capstone/Data-science-using-SpaceX-API/week2-1_jupyter-labs-eda-sql-coursera_sqllite.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

https://github.com/YukiG16/IBM-Data-Science-Certificate/blob/main/Course10_applied-data-science-capstone/Data-science-using-SpaceX-API/week3-1_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
- Explain why you added those plots and interactions
-

https://github.com/YukiG16/IBM-Data-Science-Certificate/blob/main/Course10_applied-data-science-capstone/Data-science-using-SpaceX-API/week3-2_dash_completed.png

Predictive Analysis (Classification)

- Find best Hyperparameter for SVM, Classification Trees and Logistic Regression

https://github.com/YukiG16/IBM-Data-Science-Certificate/blob/main/Course10_applied-data-science-capstone/Data-science-using-SpaceX-API/week4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

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

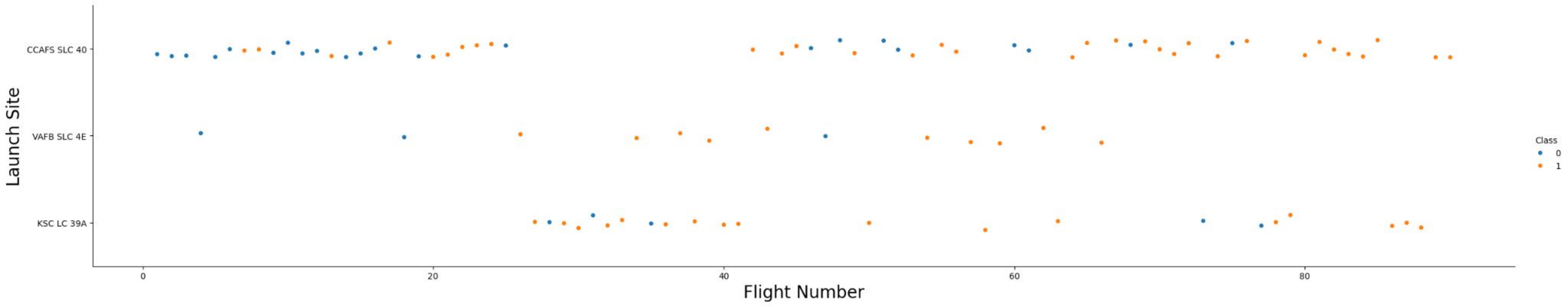
in process

Section 2

Insights drawn from EDA

Yuki Nakayama 1st,MAR,2024

Flight Number vs. Launch Site

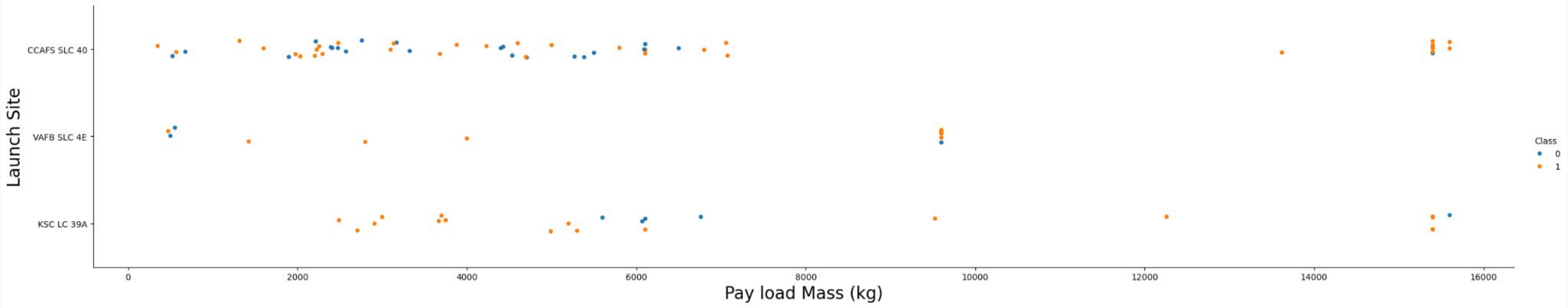


●:Success

●:failed

As the flight number increases, the first stage is more likely to land successfully.
This trend seems to be present across all launch sites.

Payload vs. Launch Site

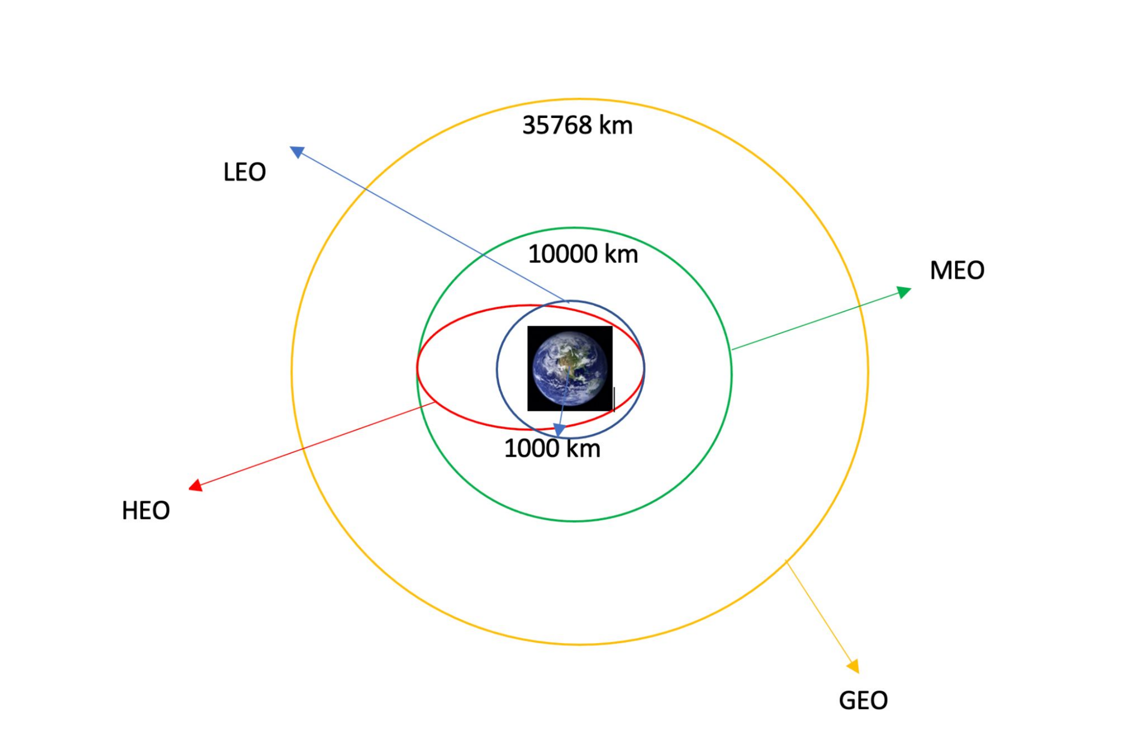


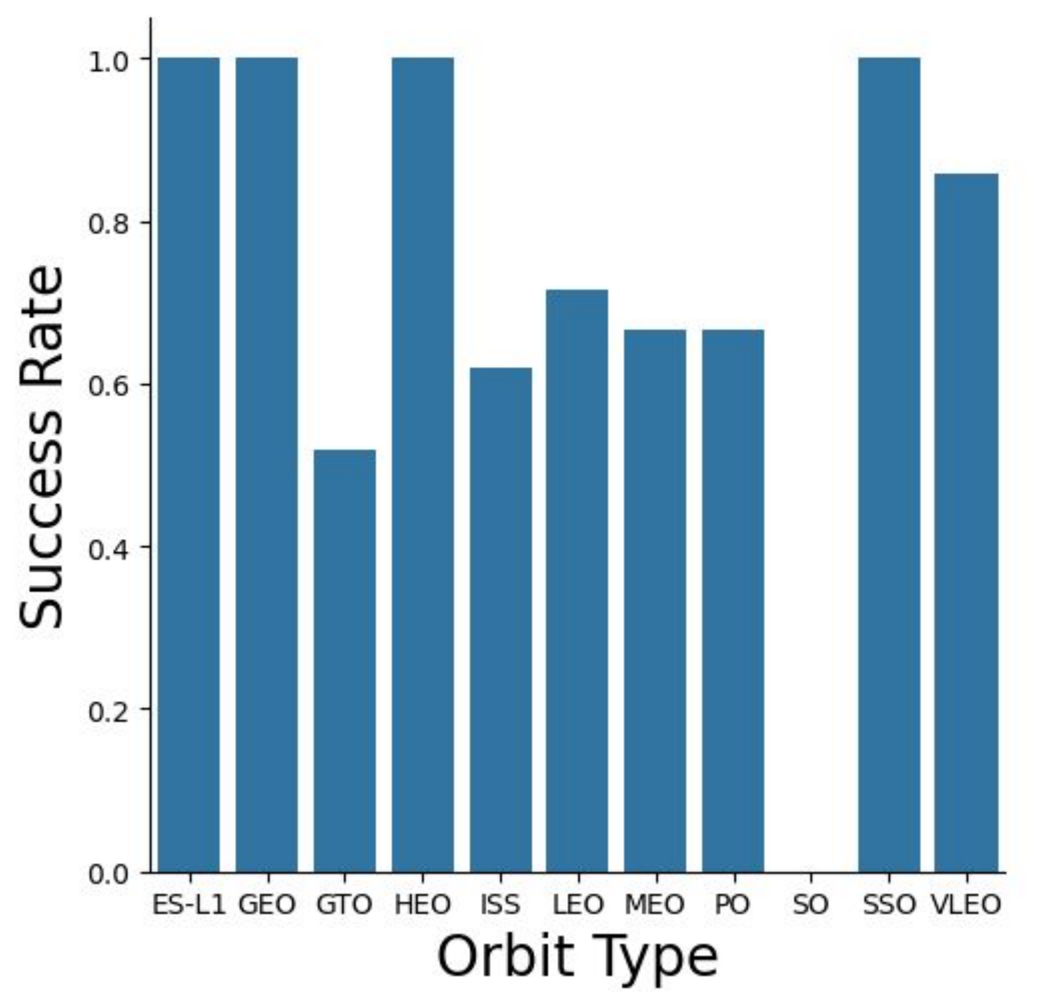
●:Success

●:failed

It appears that as the payload increases, so does the flight number.

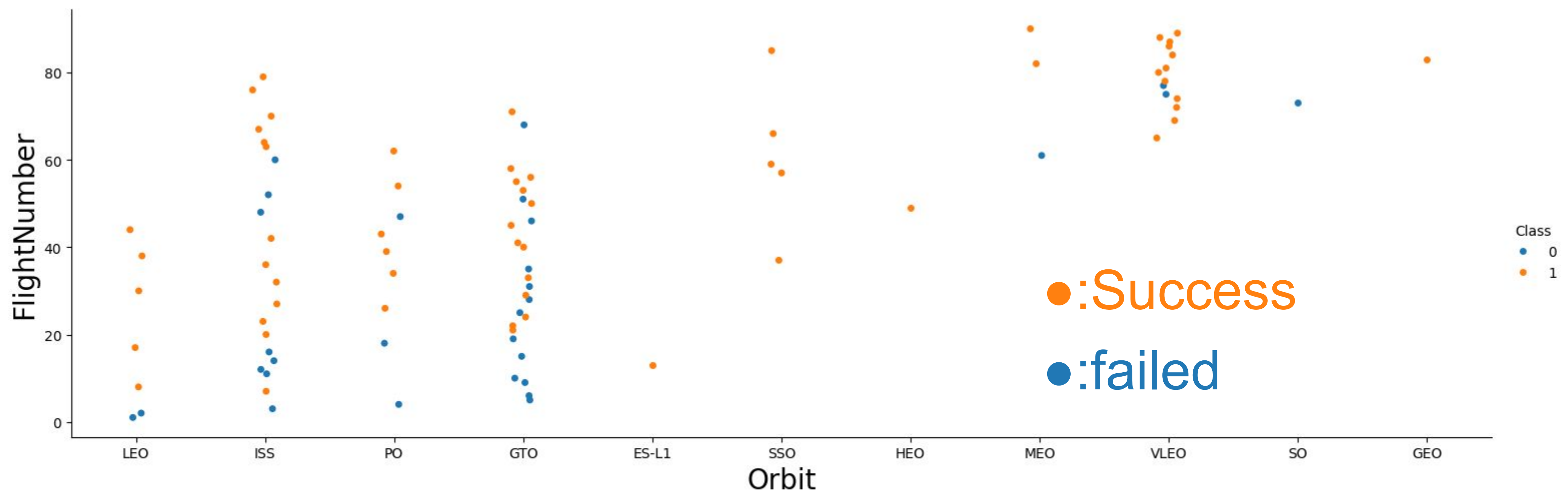
Success Rate vs. Orbit Type





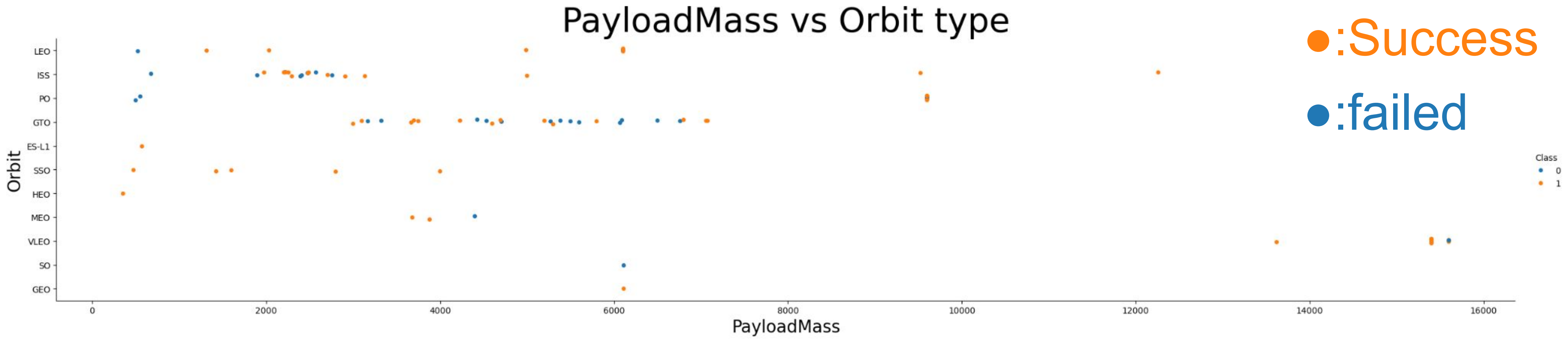
ES-L1, GEO, HEO, and SSO have a high success rate.

Flight Number vs. Orbit Type



LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

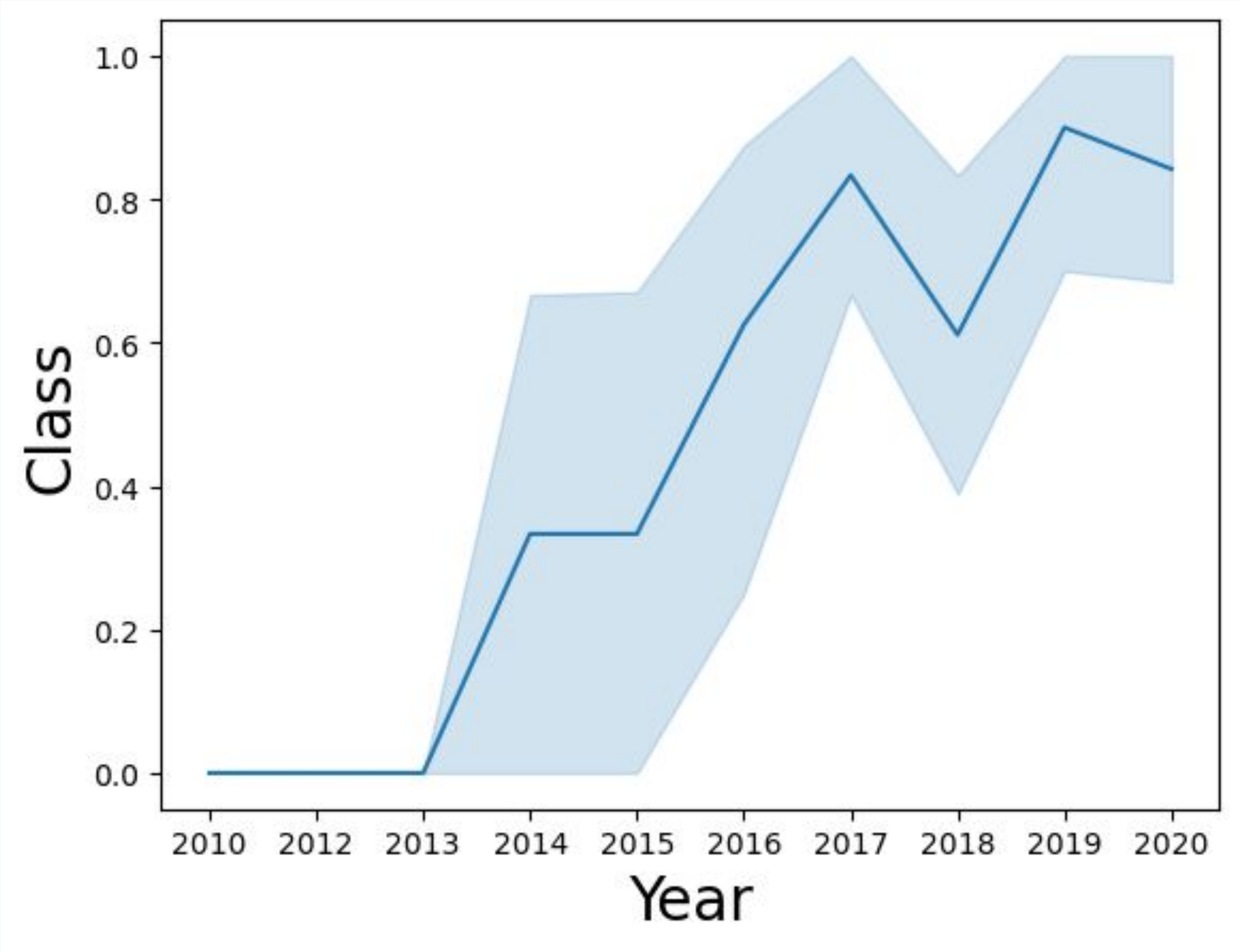
Payload vs. Orbit Type



For heavy payloads, it appears that the successful landing rate or positive landing rate is higher for Polar, LEO, and ISS.

However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend



It can be noted that the success rate has shown a consistent increase from 2013 to 2020.

All Launch Site Names

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

```
%sql select distinct(LAUNCH_SITE) from SPACEXTBL
```

SQL QUERY WITH A SHORT EXPLANATION

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

Launch Site Names Begin with 'CCA'

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

```
%sql select * from SPACEXTBL where LAUNCH_SITE like "CCA%" limit 5
```

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

Total Payload Mass

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

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where CUSTOMER = 'NASA (CRS)'
```

SQL QUERY WITH A SHORT EXPLANATION

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

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

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where BOOSTER_VERSION = 'F9 v1.1'
```

SQL QUERY WITH A SHORT EXPLANATION

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

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

```
%sql select min(DATE) from SPACEXTBL where LANDING_OUTCOME = 'Success (ground pad)'
```

SQL QUERY WITH A SHORT EXPLANATION

min(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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 SPACEXTBL where LANDING_OUTCOME = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000
```

SQL QUERY WITH A SHORT EXPLANATION

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

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
%sql \
SELECT \
    Landing_Outcome, \
    COUNT(*) AS Outcome_Count \
FROM \
    SPACEXTBL \
WHERE \
    Landing_Outcome IN ('Success', 'Failure') \
GROUP BY \
    Landing_Outcome;
```

SQL QUERY WITH A SHORT EXPLANATION

Landing_Outcome	Outcome_Count
Failure	3
Success	38

Boosters Carried Maximum Payload

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

```
%sql SELECT DISTINCT Booster_Version \
FROM SPACEXTBL \
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL LIMIT 1);
```

SQL QUERY WITH A SHORT EXPLANATION

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

Falcon 9 Block 5

2015 Launch Records

Month	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)

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) AS Month, \
    Booster_Version, \
    Launch_Site, \
    Landing_Outcome \
FROM \
    SPACEXTBL \
WHERE \
    substr(Date, 0, 5) = '2015' \
    AND Landing_Outcome LIKE 'Failure (drone ship)%';
```


SQL QUERY WITH A SHORT EXPLANATION

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 Outcome_Count \
FROM \
    SPACEXTBL \
WHERE \
    Date BETWEEN '2010-06-04' AND '2017-03-20' \
GROUP BY \
    Landing_Outcome \
ORDER BY \
    Outcome_Count DESC;
```

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

SQL QUERY WITH A SHORT EXPLANATION

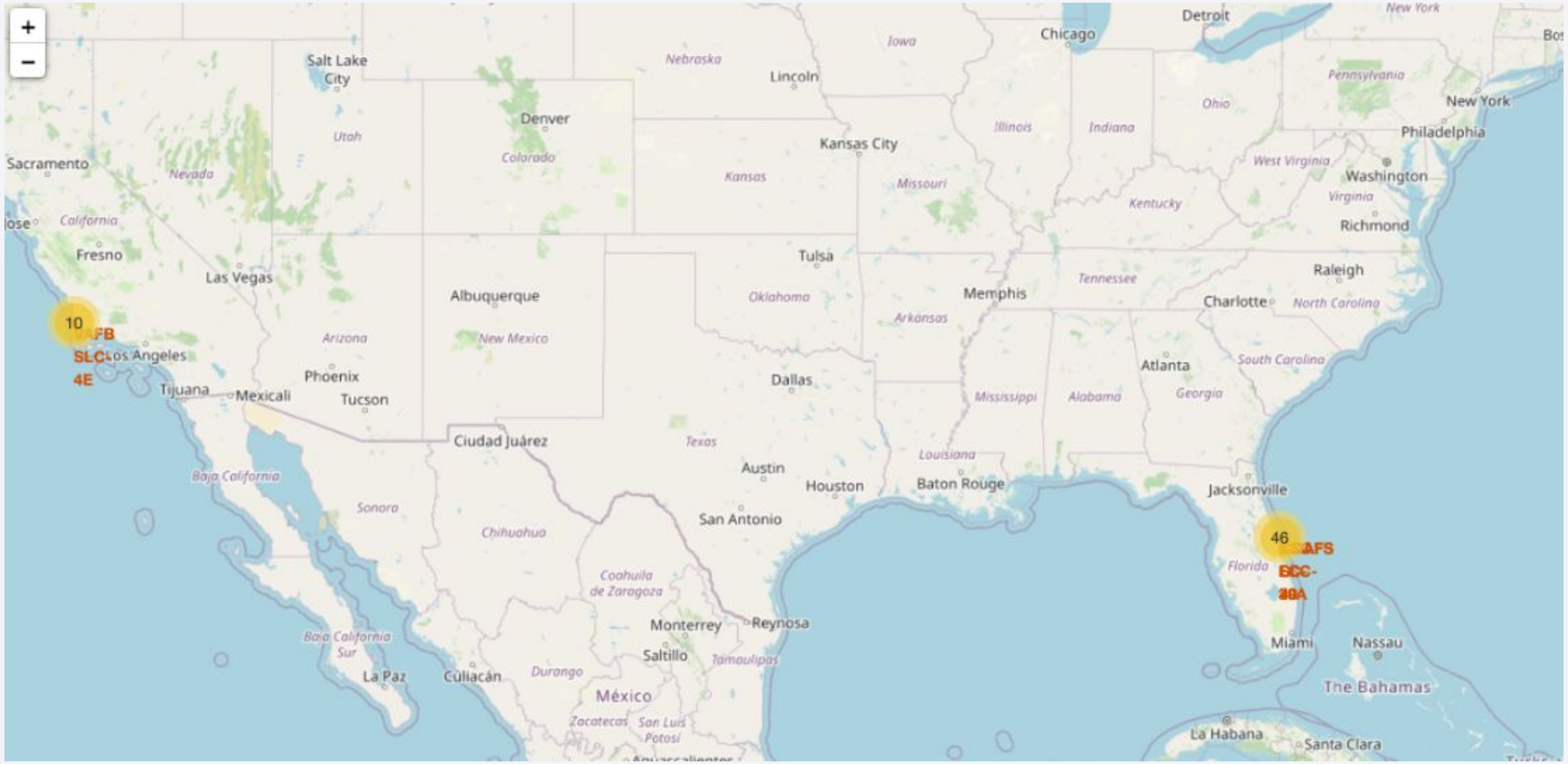
Section 3

Launch Sites Proximities Analysis

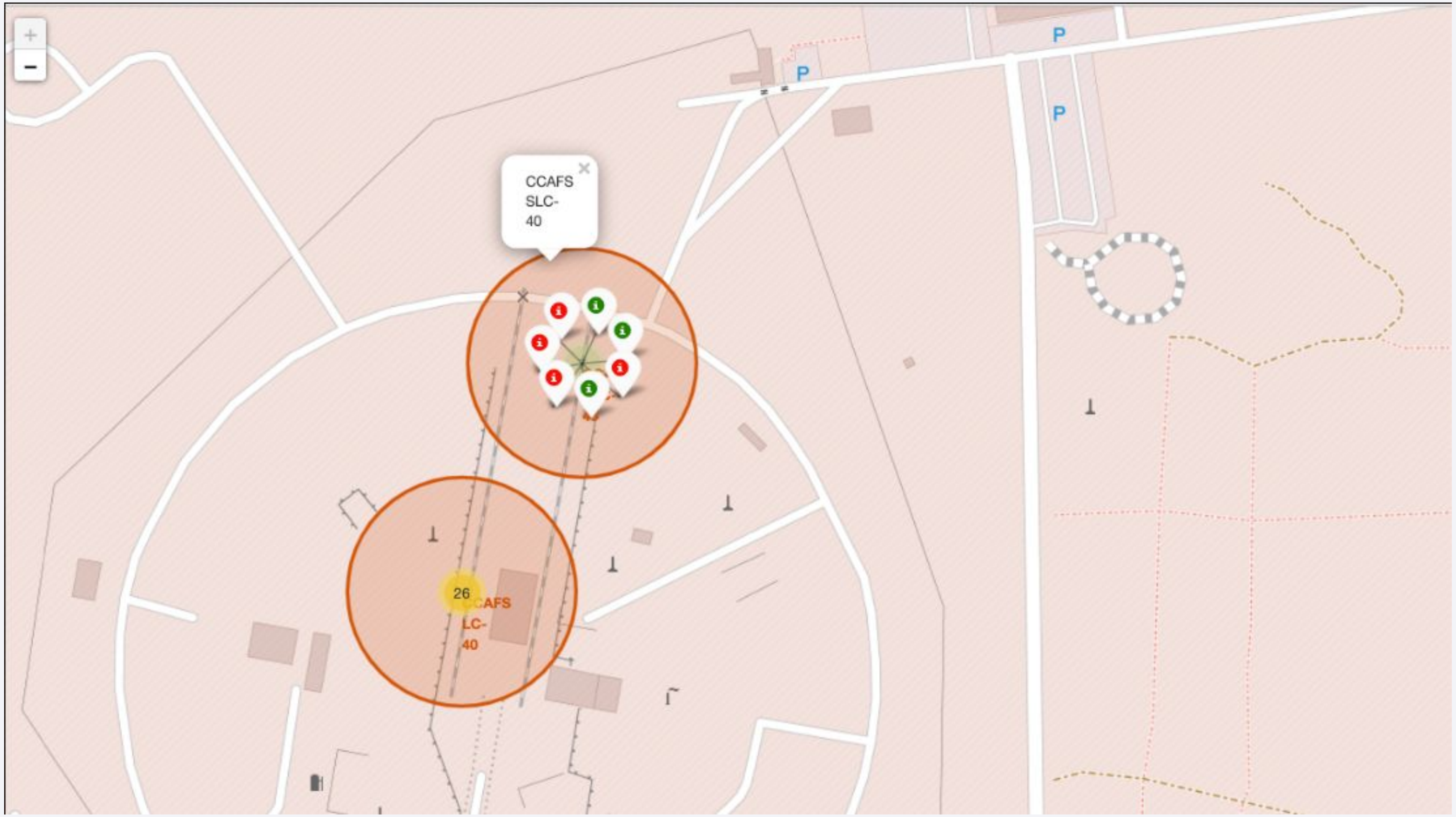
Markers all launch sites on a map

All launch sites are typically located near the Equator due to the benefits of Earth's maximum rotational speed, which results in reduced fuel needs and lower costs for reaching orbit.

All launch sites are strategically situated in close proximity to the coast to ensure safe paths over water, thereby minimizing risks to populated areas during launch failures.
In addition, it is worth noting that these systems can aid in the effective removal of rocket debris, while also providing logistical benefits for both operations and support.



an overview of the outcomes of launches at each site on the map



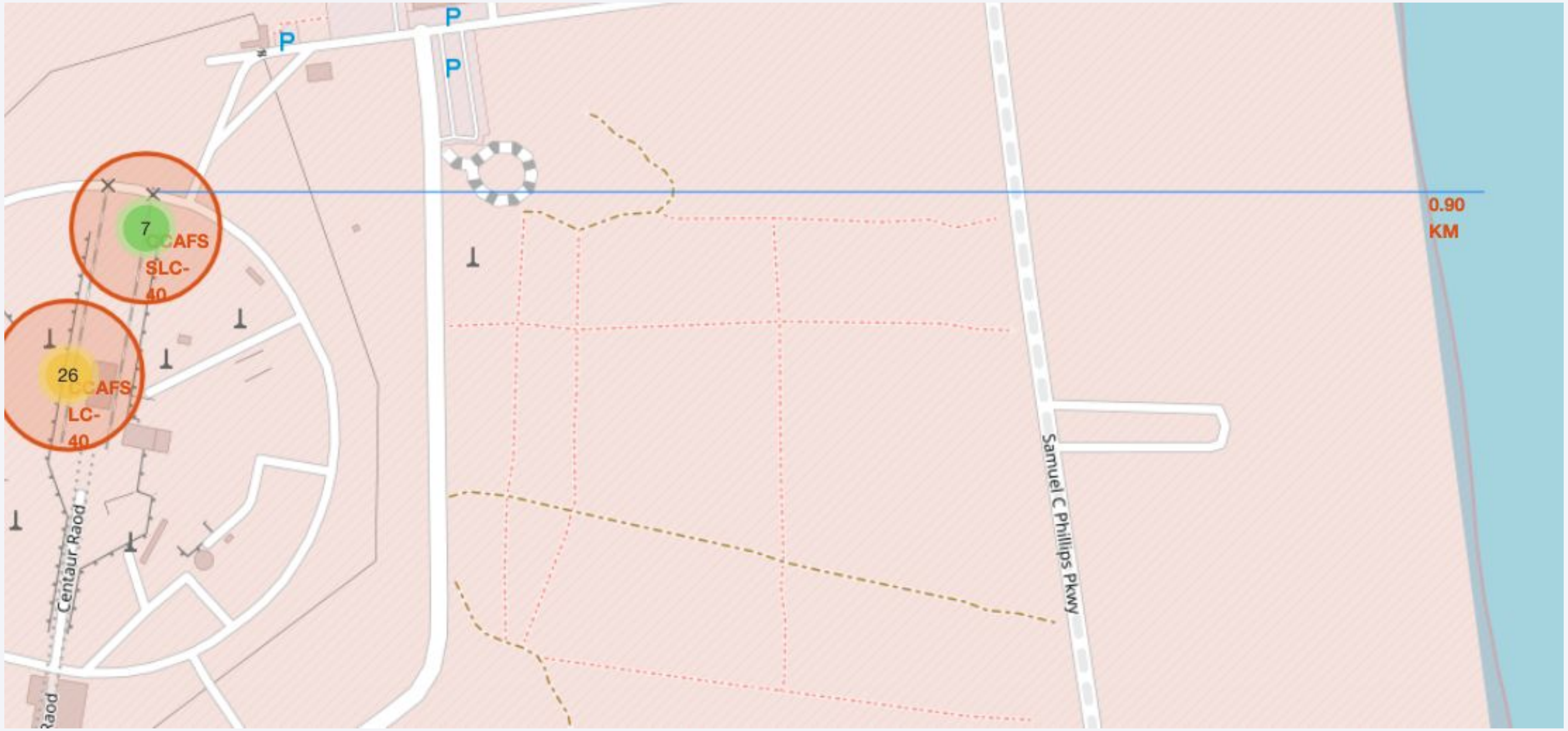
Distances between a launch site to its proximities

Launch sites are in close proximity to railways : Efficient transportation of heavy equipment.

Launch sites are in close proximity to highways: Easy access for personnel and support vehicles.

Launch sites are in close proximity to coastline: Clear trajectory over water, minimal risks to populated areas.

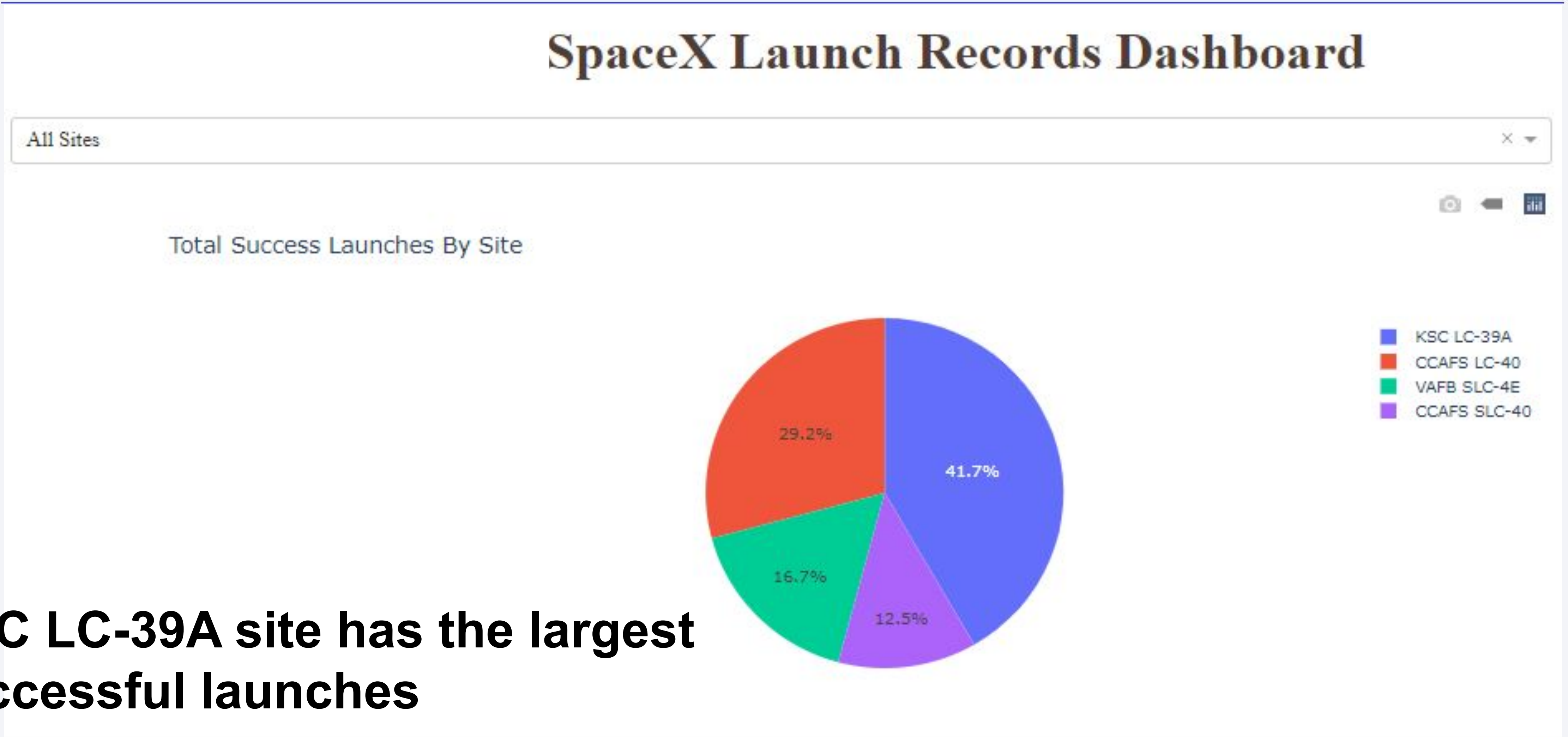
launch sites keep certain distance away from cities: Minimize risks to human life and property.



Section 4

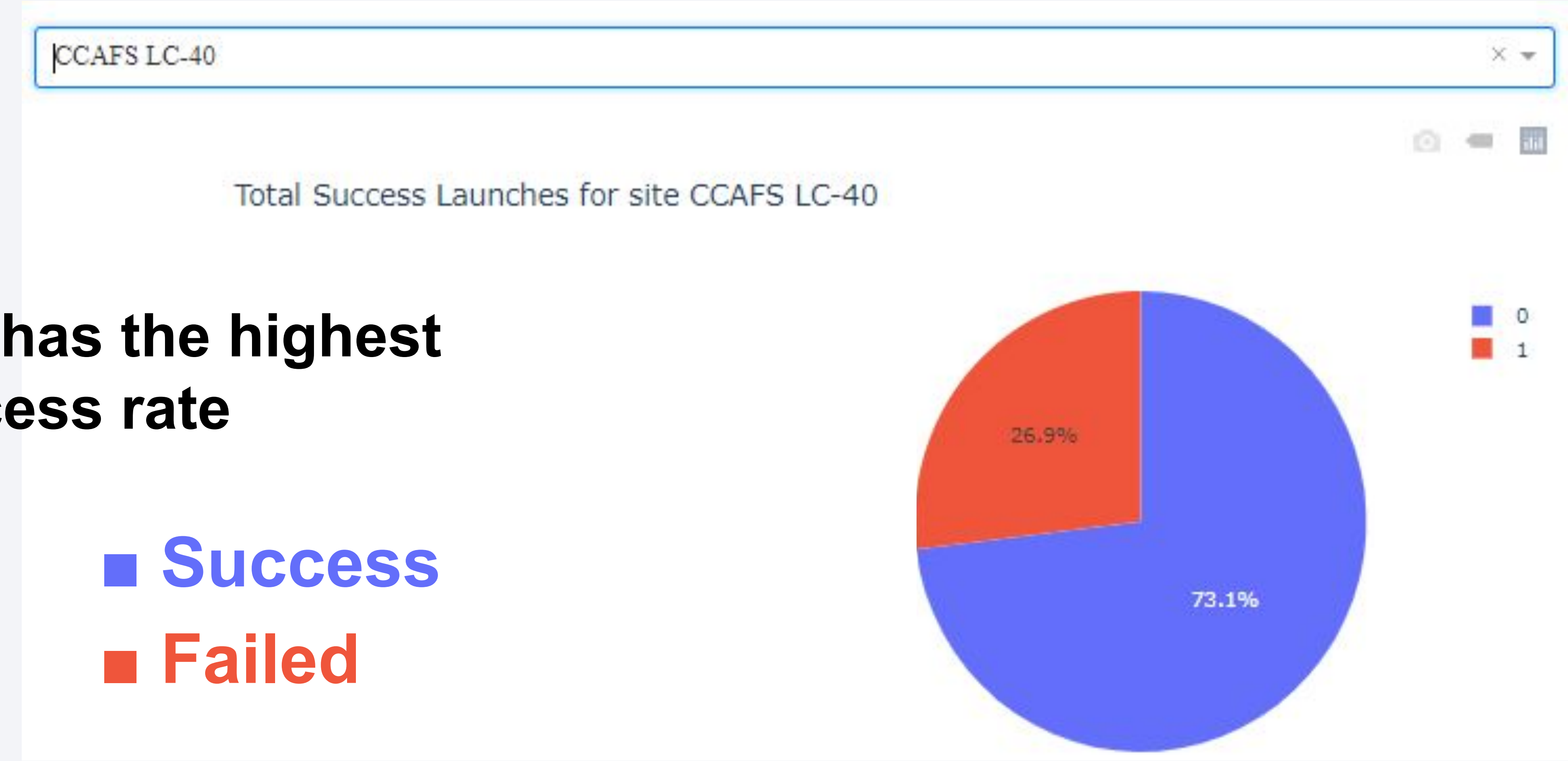
Build a Dashboard with Plotly Dash

Total Success Launches By Site



KSC LC-39A site has the largest successful launches

Total Success Launches for site CCAFS LC-40



CCAFS site has the highest launch success rate

- Success
- Failed

Correlation between payload and Success for all Sites

- Payload from 2000kg to 5000kg has the highest launch success rate.
- Payload more than 6000kg has the lowest launch success rate.
- F9 Booster version FT has the highest launch success rate.

Section 5

Predictive Analysis (Classification)

Yuki Nakayama 1st,MAR,2024

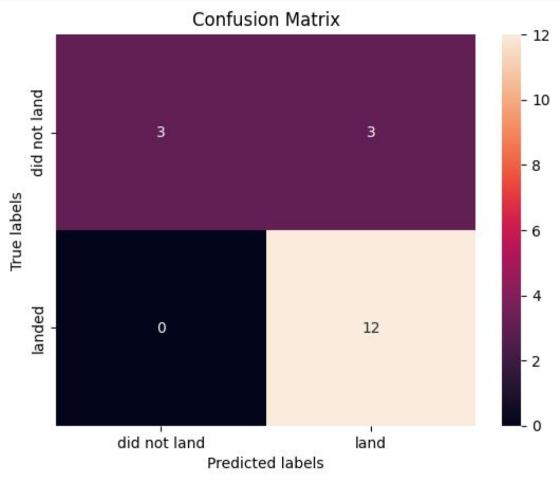
Classification Accuracy

- Visualize the built model accuracy for all built classification models, in a bar chart

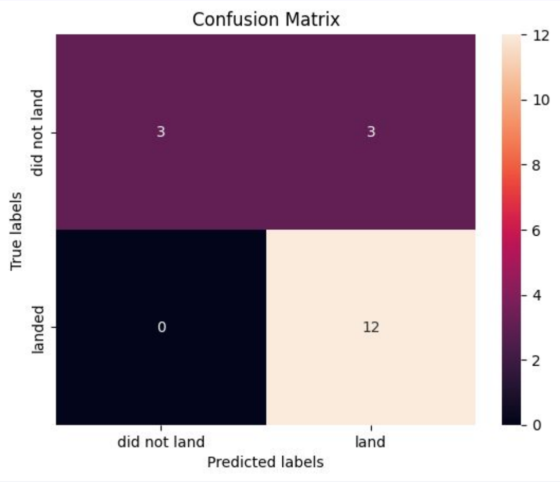
- Logistic regression decision tree, SVM and KNN model has the same classification accuracy

ML Method	Accuracy Score (%)
Support Vector Machine	83.333333
Logistic Regression	83.333333
K Nearest Neighbour	83.333333
Decision Tree	83.333333

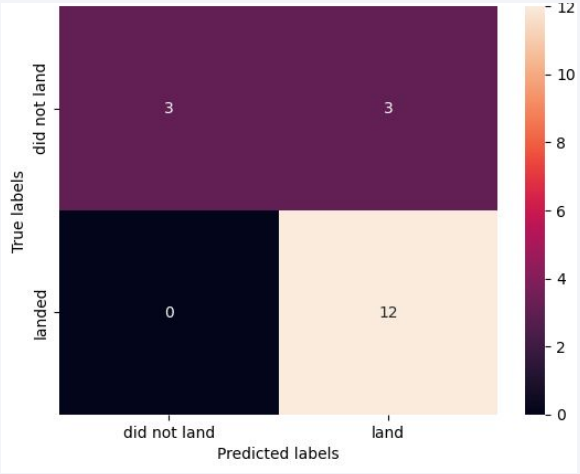
Confusion Matrix



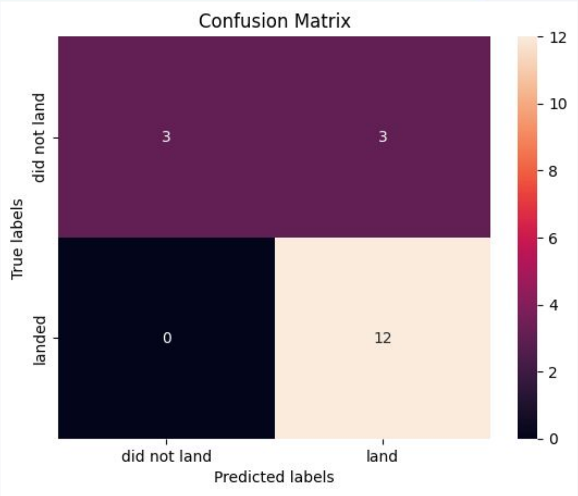
LOGISTIC REGRESSION



DECISION TREE

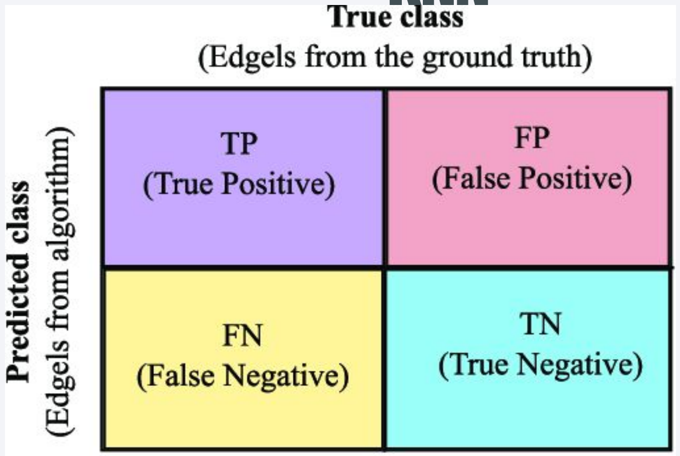


SVM



KNN

Upon examining the confusion matrix, it is clear that logistic regression decision tree, SVM and KNN are capable of distinguishing between the different classes. However, false positives remain a significant issue.



Tariq, N., Hamzah, R. A., Ng, T. F., Wang, S. L., & Ibrahim, H. (2021). Quality assessment methods to evaluate the performance of edge detection algorithms for digital image: A systematic literature review. IEEE Access, 9, 87763-87776.

Conclusions

1 Flight Number

s

2 Orbit Type

ES-L1, GEO, HEO, and SSO have a high success rate.

3 a

s

4 a

s

5 Launch site

・sKSC LC-39A site has the largest successful launches

・CCAFS site has the highest launch success rate

6 Classification

Upon examining the confusion matrix, it is clear that logistic regression decision tree, SVM and KNN are capable of distinguishing between the different classes

Appendix

https://github.com/YukiG16/IBM-Data-Science-Certificate/blob/main/Course10_applied-data-science-capstone/Data-science-using-SpaceX-API/

- week1-1_jupyter-labs-spacex-data-collection-api.ipynb
- week1-2_jupyter-labs-webscraping.ipynb
- week1-3_labs-jupyter-spacex-Data wrangling.ipynb
- week2-1_jupyter-labs-eda-sql-coursera_sqllite.ipynb
- week2-2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb
- week3-1_lab_jupyter_launch_site_location.jupyterlite.ipynb
- week3-2_dash_completed.png
- week4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite

Thank you!