



IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Mingyu Zhang>
<03/18/2023>



Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

- **Project background and context**
- As a data scientist working for a new rocket company. Space Y that would like to compete with SpaceX founded by Billionaire industrialist Allon Musk. We will predict if the Falcon 9 first stage will land successfully. SpaceX 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 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. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

- **Problems you want to find answers**

Determine the price of each launch.

- gathering information about Space X
- creating dashboards for the team.

Determine if SpaceX will reuse the first stage.

- train a machine learning model and use public information to predict if SpaceX will reuse the first stage.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology: Collecting Falcon 9 historical launch records from a Wikipedia page Web Scraping or SpaceX API.

Make a get request to the SpaceX API for data wrangling and formating.	Web scrap Falcon 9 launch Wikipedia records with BeautifulSoup:
1. Request to the SpaceX API	1. Extract a Falcon 9 launch records HTML table from Wikipedia
2. Perform data wrangling Clean the requested data	2. Parse the table and convert it into a Pandas data frame

- **Exploratory Data Analysis:** perform some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models
- **Determine Training Labels:** Convert outcomes (True/False Ocean, True/False ASDS) into Training Labels with `1` means the booster successfully landed `0` means it was unsuccessful.
- 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 machine learning to determine if the first stage of Falcon 9 will land successfully
 - 1. split your data into training data and test data
 - 2. Train different classification models
 - 3. find the best Hyperparameter for SVM, Classification Trees, and Logistic Regression.

Data Collection

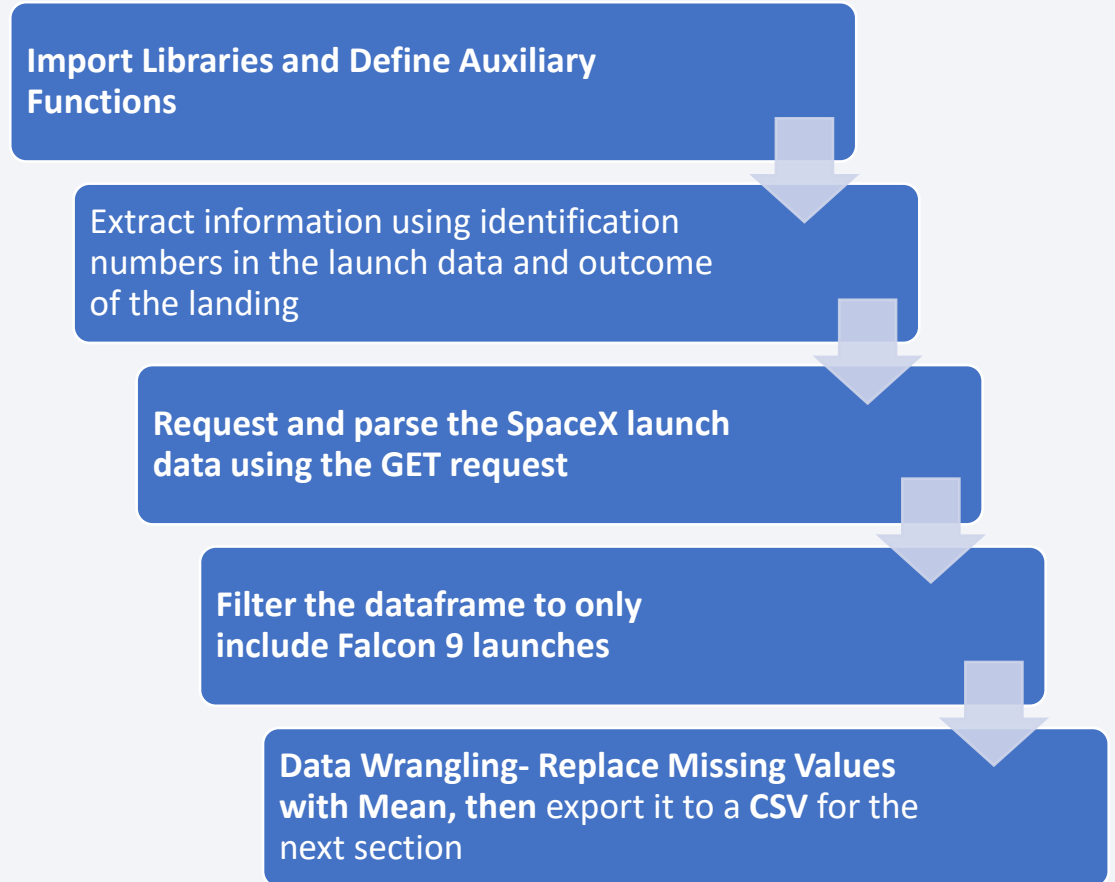
- Describe how data sets were collected.

The dataset was collected by REST API and Web Scrapping from Wikipedia

- Present your data collection process use key phrases and flowcharts
 1. REST API Data Collection--See page 8
 2. Web scrapping Data Collection-with Beautiful Soup—See Page 9

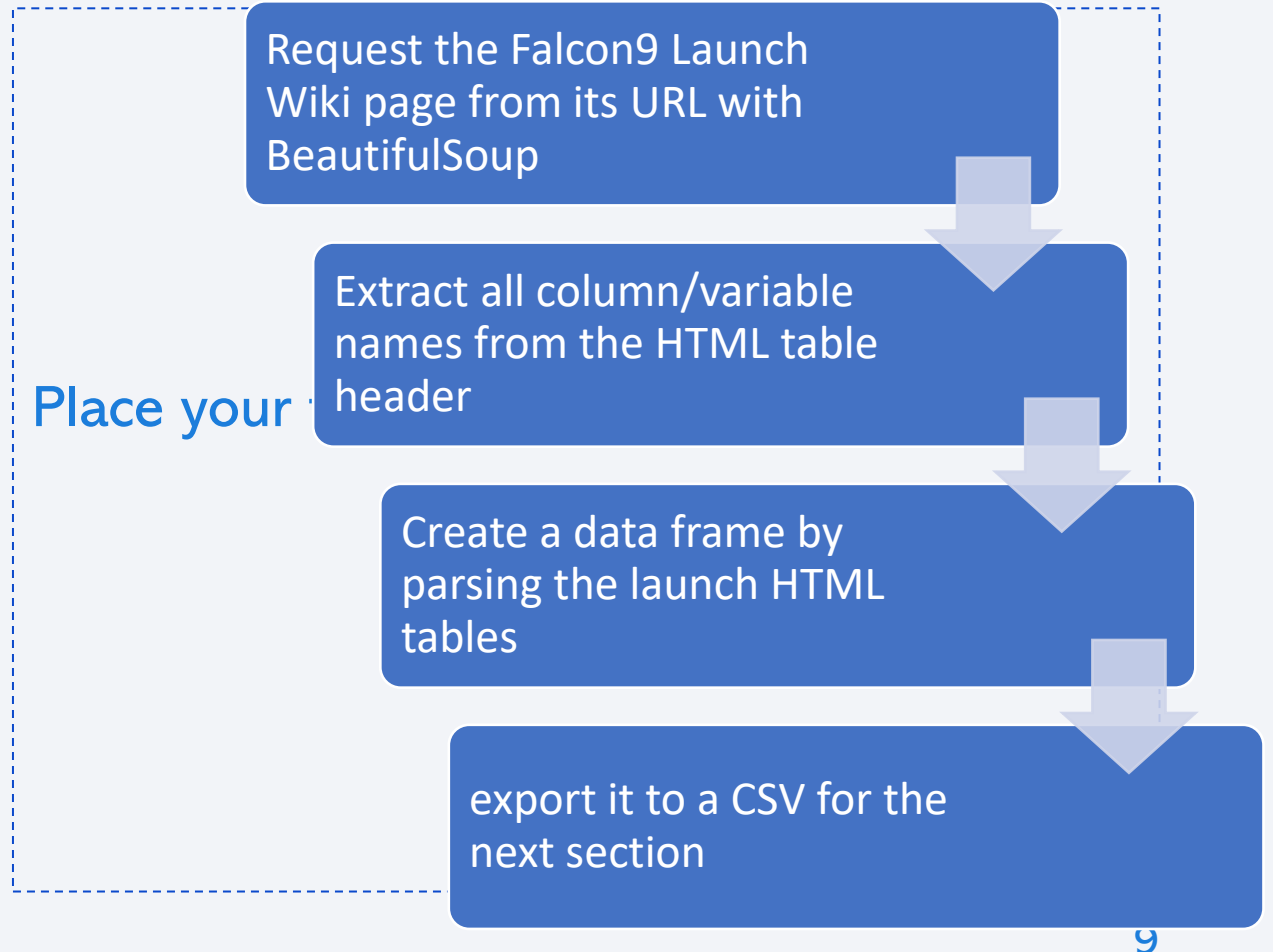
Data Collection – SpaceX API

- Data collection with SpaceX REST calls using key phrases and flowcharts
- the completed SpaceX API calls notebook
<https://github.com/zmy2338/IBM.github.io/blob/8ca04155c5f5946a49e38ac9013b4ae6582f52a6/1.1a%20Collecting%20the%20data.ipynb> ---an external reference and peer-review purpose



Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts
- GitHub URL of the completed web scraping notebook:
[https://github.com/zmy2338/IBM.github.io/blob/8ca04155c5f5946a49e38ac9013b4ae6582f52a6/jupyter-labs-eda-sql-coursera_sqlite%20\(1\).ipynb](https://github.com/zmy2338/IBM.github.io/blob/8ca04155c5f5946a49e38ac9013b4ae6582f52a6/jupyter-labs-eda-sql-coursera_sqlite%20(1).ipynb)



Data Wrangling

- Describe how data were processed

Import Libraries and Define Auxiliary Functions

Calculate the number of launches on each site

Calculate the number and occurrence of each orbit

Calculate the number and occurrence of mission outcome per orbit type

Create a landing outcome label from Outcome column, then export to a CSV for the next section

- GitHub URL:

<https://github.com/zmy2338/IBM.github.io/blob/8ca04155c5f5946a49e38ac9013b4ae6582f52a6/1.2%20labs-jupyter-spacex-Data%20wrangling.ipynb>

EDA with Data Visualization

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

When FlightNumber vs. PayloadMass and plot overlay the outcome of the launch. We see that as the flight number increases, the first stage is more likely to land successfully. The payload mass is also important; it seems the more massive the payload, the less likely the first stage will return.

- Relationship between Flight Number and Launch Site: no obvious finding.
- Relationship between Payload and Launch Site: Payloads that greater than 9000kg tended to launch from CCAFS SLC 40 & KSC LC 39A. Payloads less than 6000 kg tended to fail at a higher rate when launched from CCAFS SLC 40.
- A bar chart for the success rate of each orbit---relationship between success rate of each orbit type: find which orbits have high success rate.
- A scatter point chart with x axis to be FlightNumber and y axis to be the Orbit, and hue to be the class value: you should see that in the 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 scatter point charts to reveal the relationship between Payload and Orbit type: With heavy payloads the successful landing or positive landing rate are more 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.
- The launch success yearly trend you can observe that the success rate since 2013 kept increasing till 2020
- GitHub URL <https://github.com/zmy2338/IBM.github.io/blob/24996ceb436de98d7c740aa626f49dd812374ae5/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb>

EDA with SQL

Using bullet point format, summarize the SQL queries you performed

- Displaying the names of the launch sites.
- Displaying 5 records where launch sites begin with the string 'CCA'.
- Displaying the total payload mass carried by booster launched by NASA (CRS).
- Displaying the average payload mass carried by booster version F9 v1.1.
- Listing the date when the first successful landing outcome in ground pad was achieved.
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- Listing the total number of successful and failure mission outcomes.
- Listing the names of the booster_versions which have carried the maximum payload mass.
- Listing the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015.
- Rank the count of landing outcomes or success between the date 2010-06-04 and 2017-03-20, in descending order.
- GitHub URL [https://github.com/zmy2338/IBM.github.io/blob/8ca04155c5f5946a49e38ac9013b4ae6582f52a6/jupyter-labs-eda-sql-coursera_sqlite%20\(1\).ipynb](https://github.com/zmy2338/IBM.github.io/blob/8ca04155c5f5946a49e38ac9013b4ae6582f52a6/jupyter-labs-eda-sql-coursera_sqlite%20(1).ipynb)

Build an Interactive Map with Folium

Summarize what map objects such as markers, circles, lines added to the folium map and discovery geographical patterns about launch sites:

1. Add latitude and longitude coordinates at each launch site
2. Mark all launch sites on a map: use `folium.Circle`` to add a highlighted circle area with a text label on a specific coordinate
3. Mark the success/failed launches for each site on the :use `folium.Circle`` to add a highlighted circle area with a text label on a specific coordinate.
4. Assigned the dataframe `launch_outcomes(failure,success)` to classes 0 and 1 with Red and Green markers on the map in `MarkerCluster()`.
5. Calculate the distances between a launch site to its proximities

Add various mark and objects helping us explain the questions of:

1. Are all launch sites in proximity to the Equator line?
2. Are all launch sites in very close proximity to the coast?
3. • How close the launch sites with railways, highways and coastlines?
4. • How close the launch sites with nearby cities?

GitHub URL:

[https://github.com/zmy2338/IBM.github.io/blob/6bdc39e3ad35a5ed71445335d8d7054f408431a6/lab_jupyter_launch_site_location.jupyterlite%20\(1\).ipynb](https://github.com/zmy2338/IBM.github.io/blob/6bdc39e3ad35a5ed71445335d8d7054f408431a6/lab_jupyter_launch_site_location.jupyterlite%20(1).ipynb)


Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
 - **Launch Site Drop-down Input**
 - **pie chart visualizing launch success counts**
 - **Range Slider to Select Payload**
 - **a scatter plot with the x axis to be the payload and the y axis to be the launch outcome**
- Explain why you added those plots and interactions
 - **Interactive dashboard allowing the user to play around with the data as they need**
 - **Pie charts can show the total success launches by selected site.**
 - **Scatter plot can show the relationship with Outcome and Payload Mass (Kg) for the different booster version**
- GitHubURL:
https://github.com/zmy2338/IBM.github.io/blob/e16a10489e522f4222c06af75a02e887be62ab20/spacex_dash_app.py



Predictive Analysis (Classification)

Model selection: Select the appropriate machine learning model that can solve the problem



Model training: Train the model using the historical data collected splitting the data into training and testing sets, training the model on the training data, and evaluating its performance on the testing data.



Model evaluation: Evaluate the model's performance using metrics such as accuracy, precision, recall.

Results

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



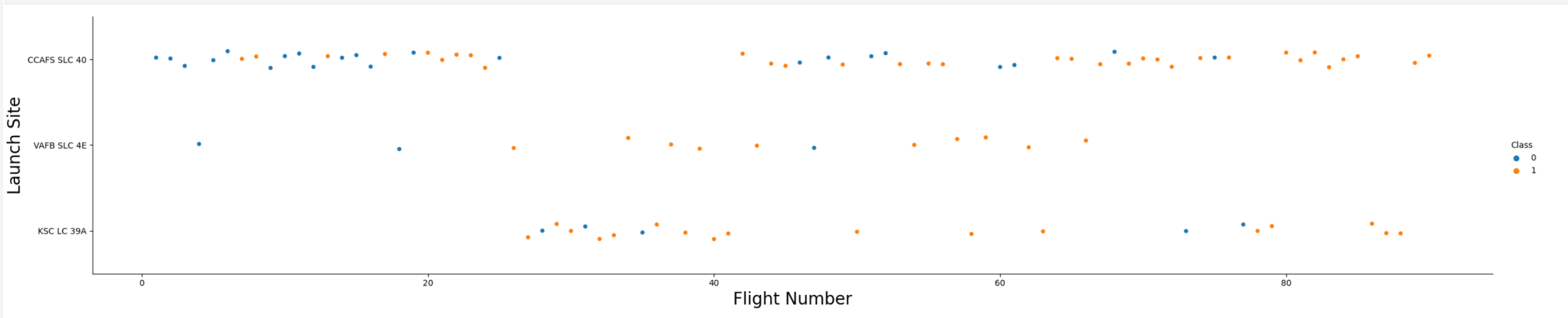
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- scatter plot of Flight Number vs. Launch Site

```
# Plot a scatter point chart with x axis to be Flight Number and y axis to be the Launch site, and hue to be the class value
sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect=5)
plt.xlabel("Flight Number", fontsize=20)
plt.ylabel("Launch Site", fontsize=20)
plt.show()
```



Explanations:

Payloads that great than 9000kg tended to launch from CCAFS SLC 40 & KSC LC 39A

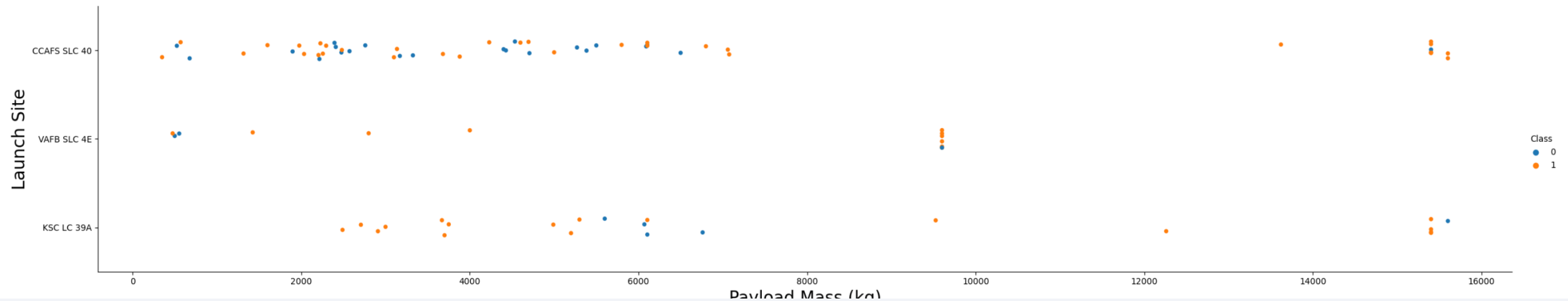
Payloads less than 6000 kg tended to fail at a higher rate when launched from CCAFS SLC 40

Payload vs. Launch Site

- Show a scatter plot of Payload vs. Launch Site

We also want to observe if there is any relationship between launch sites and their payload mass.

```
# Plot a scatter point chart with x axis to be Pay Load Mass (kg) and y axis to be the Launch site, and hue to be the class value
sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("Payload Mass (kg)",fontSize=20)
plt.ylabel("Launch Site",fontSize=20)
plt.show()
```

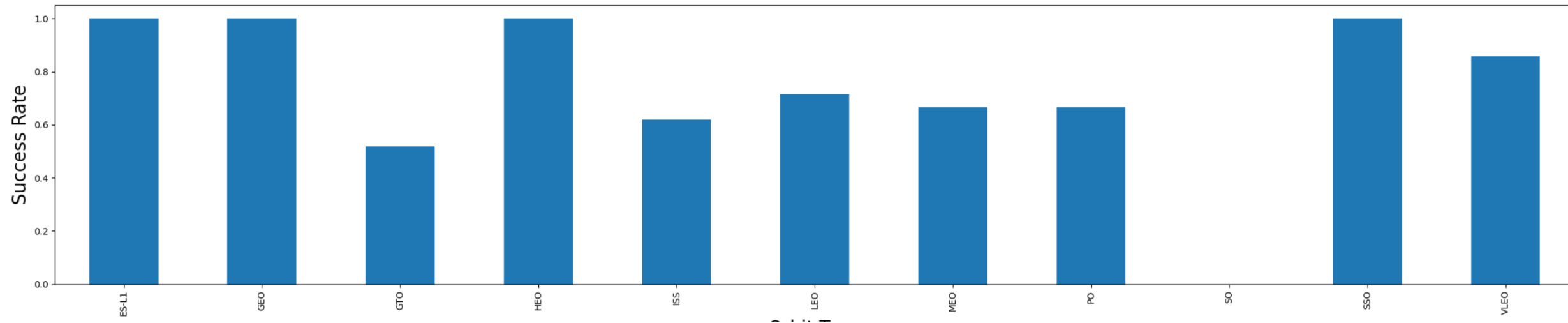


Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type

- Show a bar chart for the success rate of each orbit type

```
# HINT use groupby method on Orbit column and get the mean of Class column
df.groupby("Orbit").mean()["Class"].plot(kind='bar')
plt.xlabel("Orbit Type",fontsize=20)
plt.ylabel("Success Rate",fontsize=20)
plt.show()
```

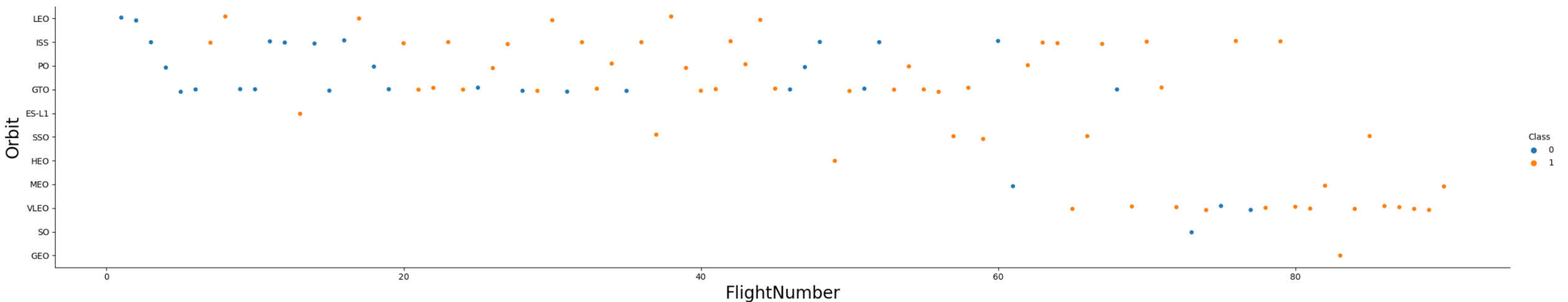


We want to visually check if there are any relationship between success rate and orbit type. ES-L1,GEO, SSO and HEO have the highest success rate.

Flight Number vs. Orbit Type

- Show a scatter point of Flight number vs. Orbit type

```
# Plot a scatter point chart with x axis to be FlightNumber and y axis to be the Orbit, and hue to be the class value
sns.catplot(y="Orbit", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("FlightNumber", fontsize=20)
plt.ylabel("Orbit", fontsize=20)
plt.show()
```

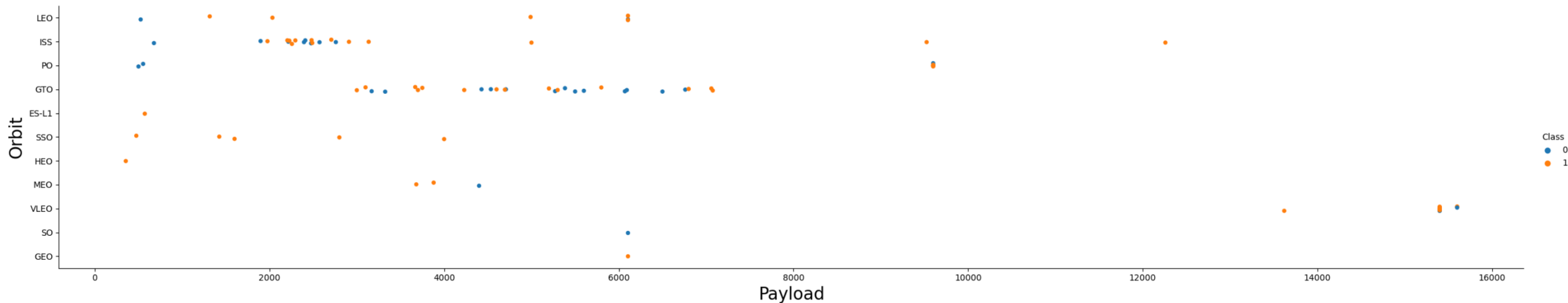


You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type

- Show a scatter point of payload vs. orbit type

```
# Plot a scatter point chart with x axis to be Payload and y axis to be the Orbit, and hue to be the class value
sns.catplot(y="Orbit", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("Payload",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```



With heavy payloads the successful landing or positive landing rate are more 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.

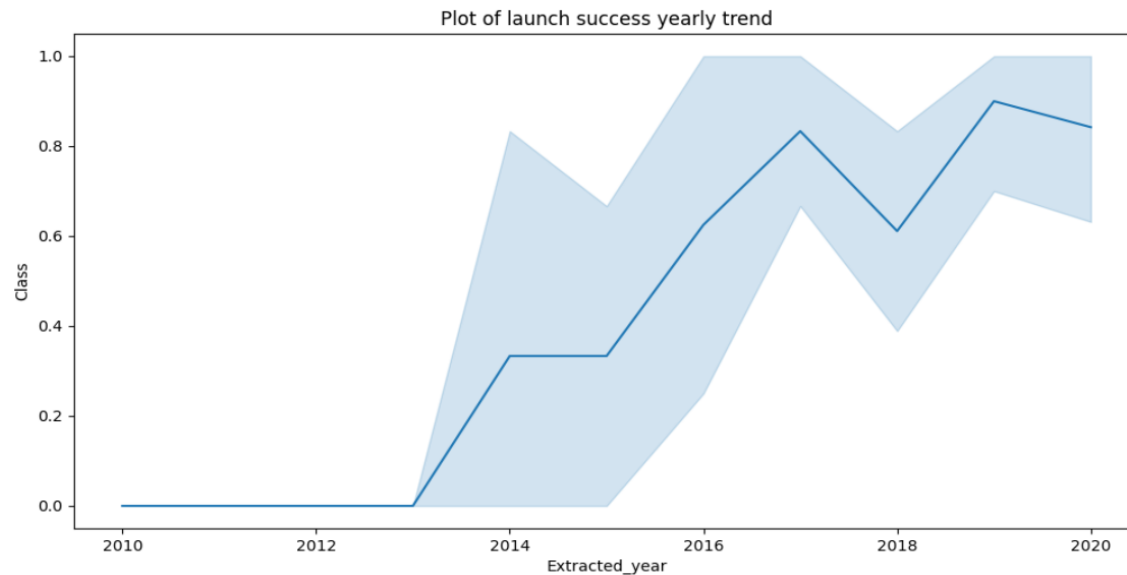
Launch Success Yearly Trend

- Show a line chart of yearly average success rate

```
# Plot a line chart with x axis to be the extracted year and y axis to be the success rate

df_copy = df.copy()
df_copy['Extracted_year'] = pd.DatetimeIndex(df['Date']).year

# plot line chart
fig, ax = plt.subplots(figsize=(12,6))
sns.lineplot(data=df_copy, x='Extracted_year', y='Class')
plt.title('Plot of launch success yearly trend')
plt.show()
```



You can observe that the success rate since 2013 kept increasing till 2020.

All Launch Site Names

- Find the names of the unique launch site

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

```
%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

Displaying the names of the launch sites.

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with 'CCA'

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

Displaying 5 records where launch sites begin with the string 'CCA'.

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

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

```
%%sql
SELECT SUM(PAYLOAD_MASS_KG_)
FROM SPACEXTBL
WHERE Customer = 'NASA (CRS)';
```

* sqlite:///my_data1.db

Done.

SUM(PAYLOAD_MASS_KG_)
45596

Displaying the total payload mass carried by booster launched by NASA (CRS).

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version 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 LIKE 'F9 v1.0%';
```

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

```
AVG(PAYLOAD_MASS_KG_)
```

```
340.4
```

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

First Successful Ground Landing Date

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

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

Hint: Use min function

```
%%sql
SELECT MIN(Date)
FROM SPACEXTBL
WHERE "Landing _Outcome" = 'Success (ground pad)';
```

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

MIN(Date)

01-05-2017

Listing the date when the first successful landing outcome in ground pad was achieved.

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

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 (ground pad)'
AND (PAYLOAD_MASS_KG_ < 6000 and PAYLOAD_MASS_KG_ > 4000);
```

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

Booster_Version
F9 FT B1032.1
F9 B4 B1040.1
F9 B4 B1043.1

Listing 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

- Calculate the total number of successful and failure mission outcomes

```
List the total number of successful and failure mission outcomes

: %%sql
SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER
FROM SPACEXTBL
GROUP BY MISSION_OUTCOME;

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

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

Listing the total number of successful and failure mission outcomes.

Boosters Carried Maximum Payload

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

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);
```

```
* 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

Listing the names of the booster_versions which have carried the maximum payload mass

2015 Launch Records

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

```
%%sql
SELECT "Landing_Outcome", "Booster_Version", "Launch_Site"
FROM SPACEXTBL
where "Landing_Outcome" in ('Failure (drone ship)')
and substr(Date,7,4) = '2015'
```

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

Landing_Outcome	Booster_Version	Launch_Site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Listing the failed landing_outcomes in drone ship, their booster versions, and launch sites names for in year 2015.

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

Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

```
%%sql
SELECT "Landing _Outcome" , COUNT("Landing _Outcome") AS TOTAL_NUMBER
FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY "Landing _Outcome"
ORDER BY TOTAL_NUMBER DESC
```

* sqlite:///my_data1.db

Done.

Landing _Outcome	TOTAL_NUMBER
------------------	--------------

Rank the count of landing outcomes or success between the date 2010-06-04 and 2017-03-20, in descending order.

A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The background is a deep blue gradient.

Section 3

Launch Sites Proximities Analysis

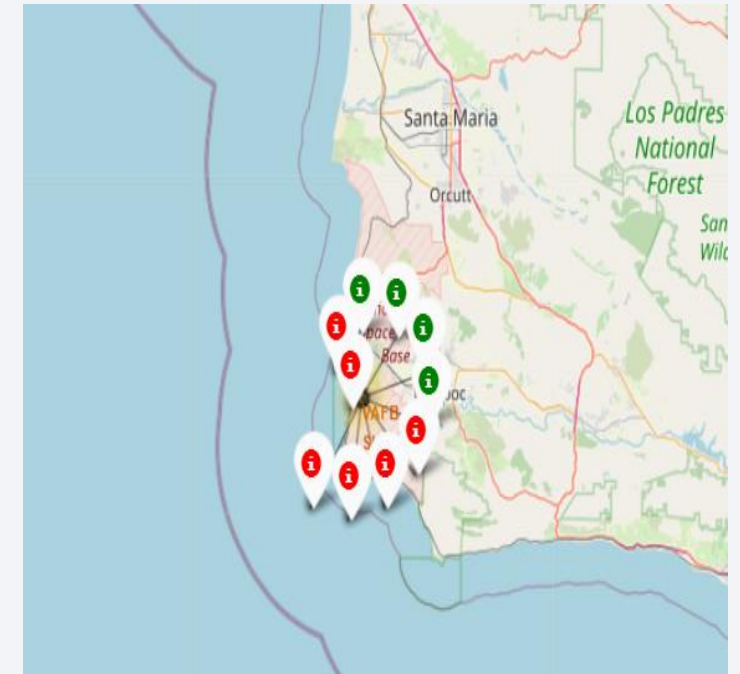


Lunch Sites location on Map

- all the SpaceX launch sites are located in FL or CA.

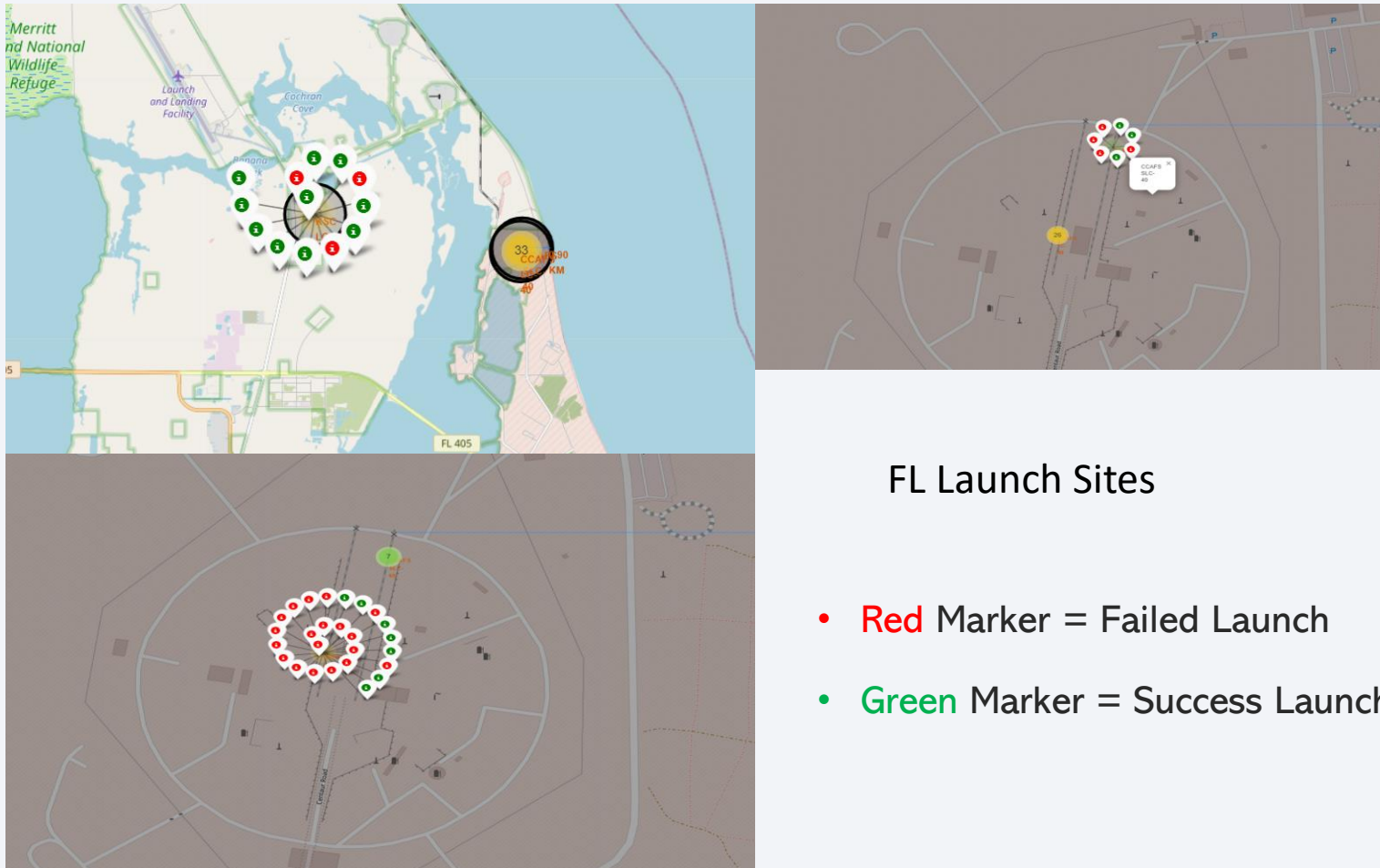
The Success/Failed Launches for Each Site

CA Launch Sites

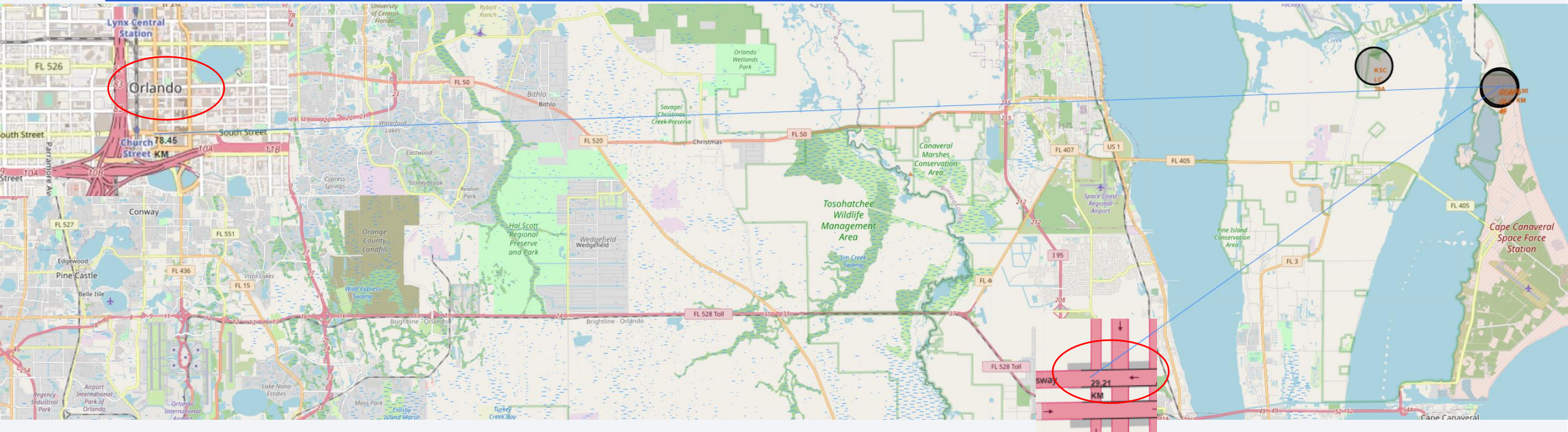


FL Launch Sites

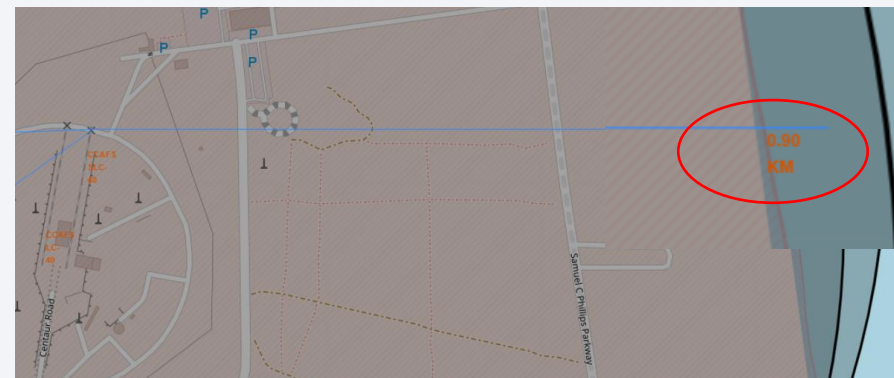
- Red Marker = Failed Launch
- Green Marker = Success Launch



Distance Lines



- * Are launch sites in close proximity to railways? No
- * Are launch sites in close proximity to highways? No
- * Are launch sites in close proximity to coastline? Yes
- * Do launch sites keep certain distance away from cities? Yes





Section 4

Build a Dashboard with Plotly Dash

The Success Lunches by Each Site

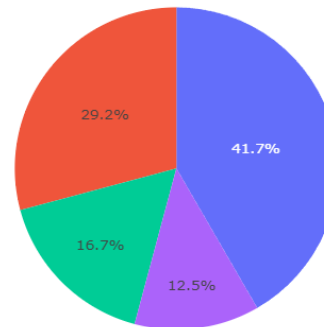
- KSC LC-39 had the most volume of success launches

SpaceX Launch Records Dashboard

All Sites

x

Success Count for all launch sites



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

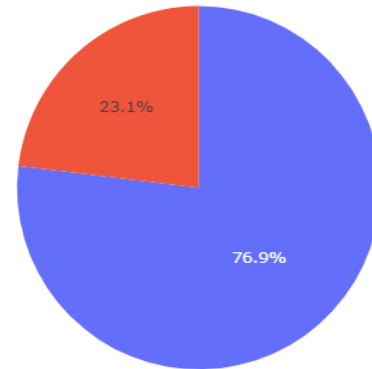
Launch Site with Highest Launch Success Ratio

SpaceX Launch Records Dashboard

KSC LC-39A

×

Total Success Launches for site KSC LC-39A



■ 1
■ 0

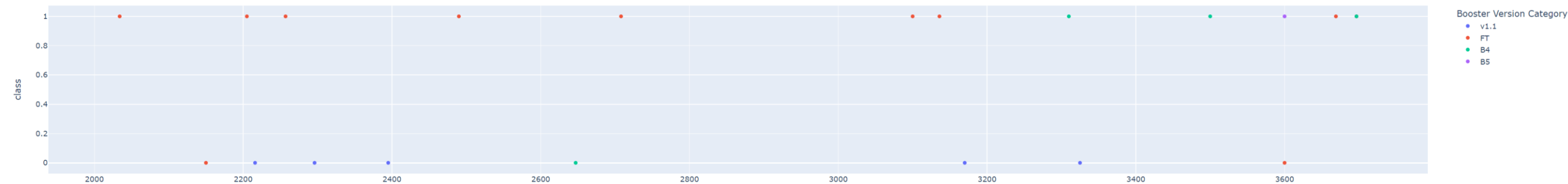
KSC L-39A have a success launch rate of 76.9%, which is the highest amount all the launch sites.

Payload vs. Launch Outcome scatter plot for all sites

Payload range (Kg):



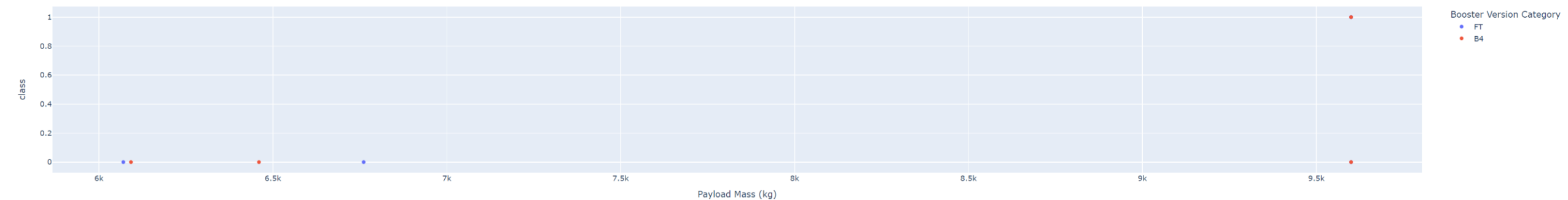
Success count on Payload mass for all sites



Payload range (Kg):



Success count on Payload mass for all sites



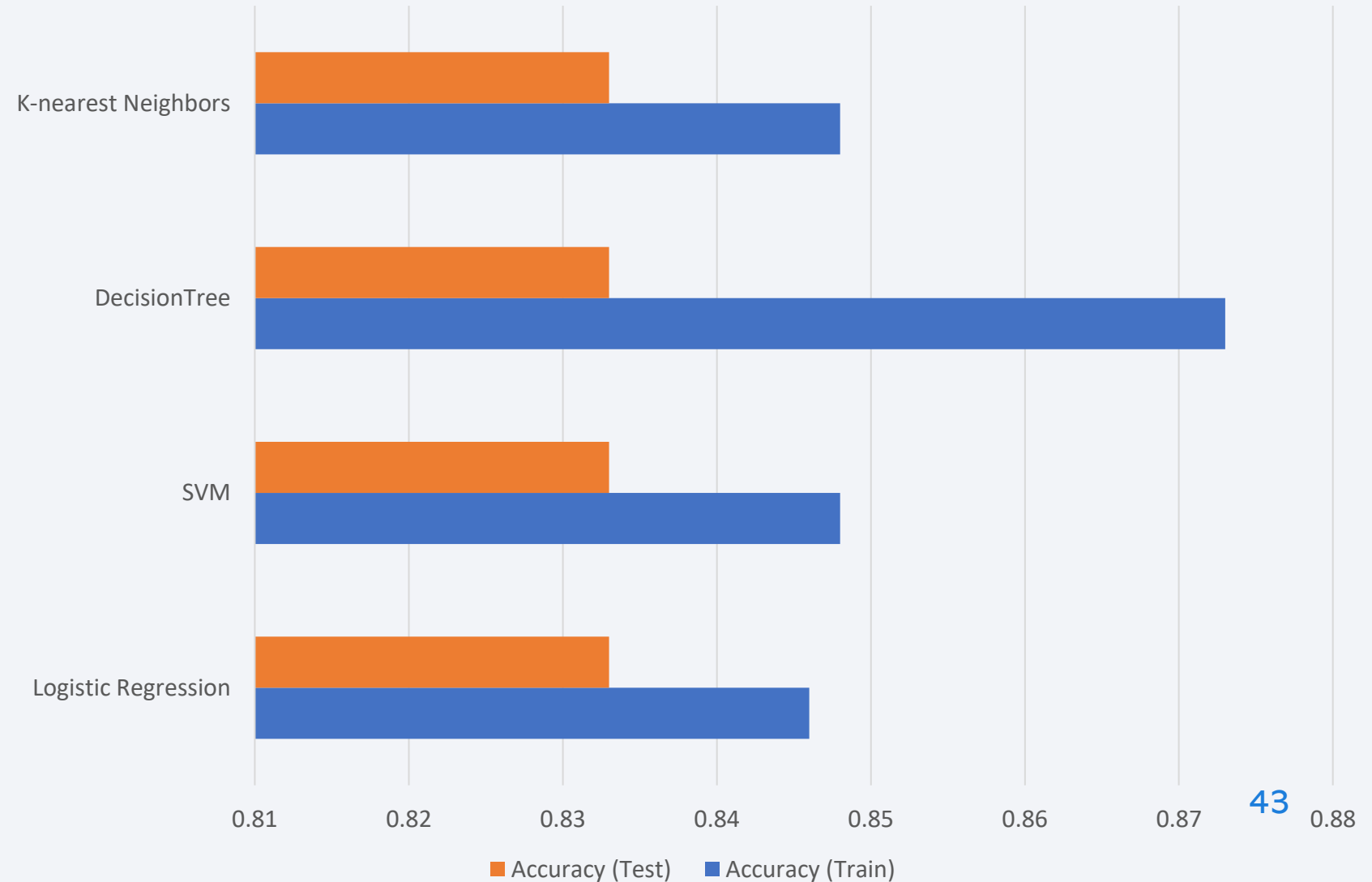
1. Payload range from 2k to 4k have the largest success rate.
2. Payload range from 6k to 10k lowest success rate.
3. B4 booster have one success record with the highest payload mass.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- After comparing accuracy of the models for test data, they all performed very similar, except for decision tree which fit train data slightly better (0.873).



Confusion Matrix

Confusion Matrix is a performance measurement for machine learning classification. It is extremely useful for measuring Recall, Precision, Specificity, Accuracy, and most importantly AUC-ROC curves.

True Positive: (12)

Interpretation: You predicted positive and it's true.

True Negative: (3)

Interpretation: You predicted negative and it's true.

False Positive (Type 1 Error): (3)

Interpretation: You predicted positive and it's false.

False Negative (Type 2 Error): (0)

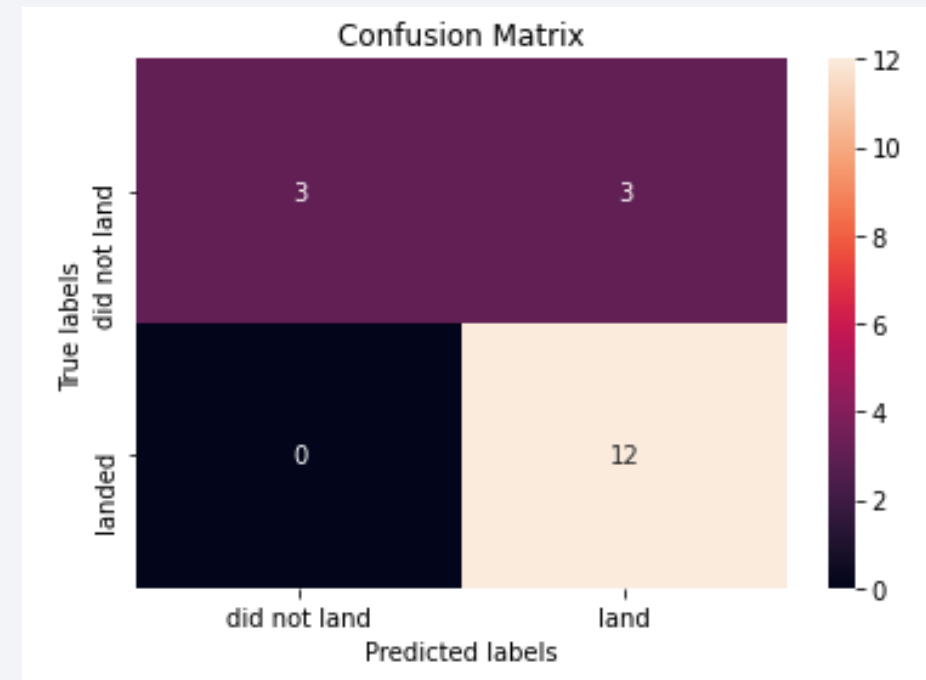
Interpretation: You predicted negative and it's false.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Precision
 $12/(12+3)=0.8$

$$\text{Recall} = \frac{TP}{TP + FN}$$

Recall
 $12/(12+0)=1$



Recall can be explained by saying, from all the positive classes, how many we predicted correctly.

Precision can be explained by saying, from all the classes we have predicted as positive, how many are actually positive.

Conclusions

Based on the information provided, we can conclude that:

- It is interesting to note that low weighted payloads perform better than heavy weighted payloads, indicating that the launch success rate is influenced by the payload's weight. This information can be used to optimize the payload's weight for future launches.
- The increase in SpaceX's launch success rate over time, starting from the year 2013, is a positive trend that suggests the company is improving its launch capabilities. This trend indicates that the company may continue to perfect its launches in the future, leading to higher success rates and greater reliability.
- KSC LC-39A has the most successful launches of any sites, with a success rate of 76.9%. This information can be used to optimize the selection of launch sites for future launches.
- It seems that the Tree Classifier algorithm is a suitable choice for building a machine learning model for this dataset. This algorithm is commonly used for classification tasks and is capable of handling both numerical and categorical data.
- Finally, SSO orbit has the highest success rate, with a success rate of 100% and more than one occurrence. This information can be used to optimize the selection of orbits for future launches.

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

Thank you!

