IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Pui Fung Wong>
<22/9/2025>

Outline

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

Executive Summary

Summary of methodologies

- - Data collection
- - Data wrangling
- - Exploratory Data Analysis with Data Visualization
- - Exploratory Data Analysis with SQL
- - Building an interactive map with Folium
- - Building a Dashboard with Plotly Dash
- - Predictive analysis (Classification)
- Summary

Summary of methodologies

- - Exploratory Data Analysis results
- - Interactive analytics demo in screenshots
- - Predictive analysis results

Introduction

Project background and context

SpaceX is the most successful company of the commercial space age, making space travel affordable. The company 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. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.

Questions to be answered

1. How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
2. Does the rate of successful landings increase over the years?
3. What is the best algorithm that can be used for binary classification in this case?

Section 1

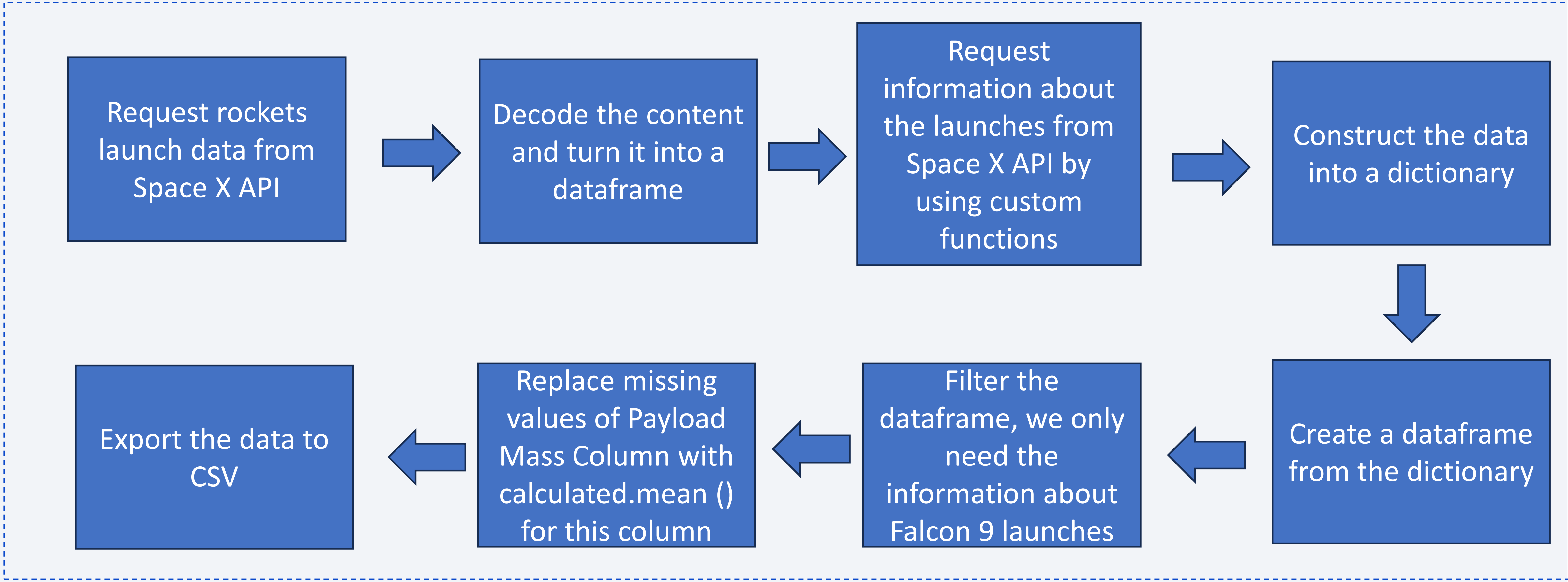
Methodology

Methodology

Executive Summary

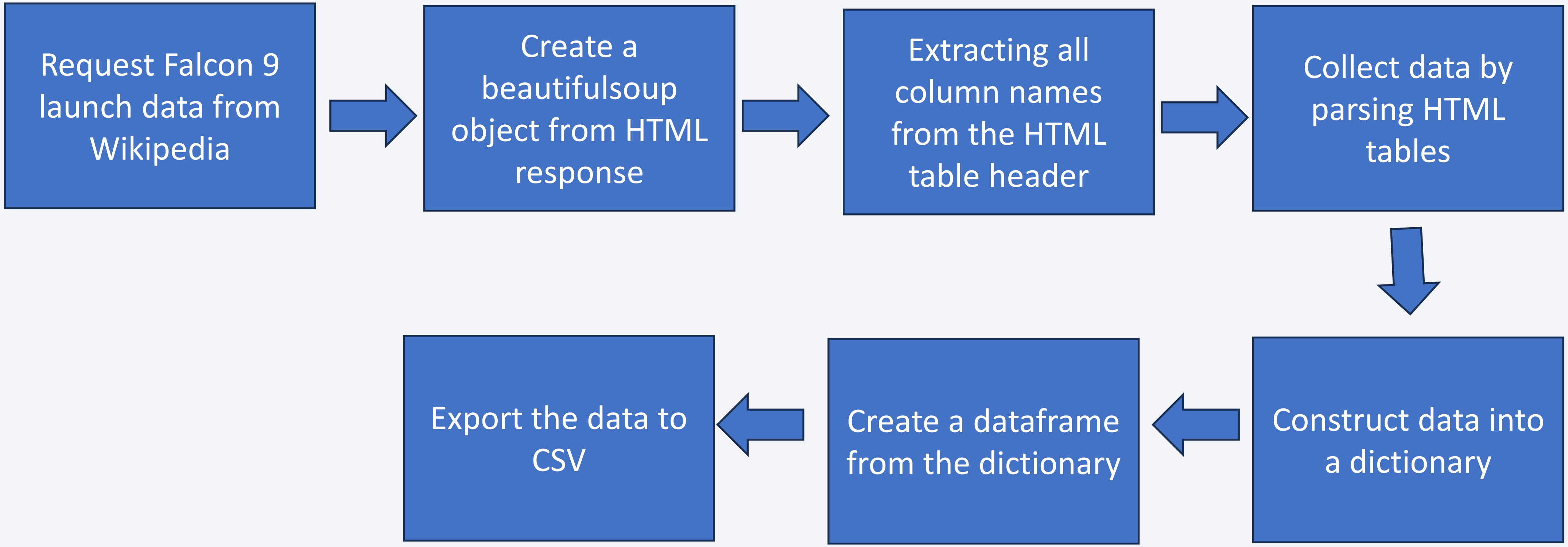
- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection – SpaceX API



https://github.com/wpf-ian/Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping



Http://github.com/wpf-ian/Applied-Data-Science-Capstone/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling

- A training label with landing outcomes is created, where successful = 1 and failure = 0. The outcome column has two components; “Mission Outcome” and “Landing Location”. New training label column class with a value of q if “Mission Outcome” is True and 0 if otherwise.

https://github.com/wpf-ian/Applied-Data-Science-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- We performed the exploratory data analysis on variables such as Flight Number, Payload Mass, Launch Site, Orbit, Class and Year.

We have

Flight number vs Payload Mass

Flight number vs Launch Site

Payload Mass vs Launch Site

Orbit vs Success Rate

Flight Number vs Orbit

Payload vs Orbit

Success Yearly Trend

http://github.com/wpf-ian/Applied-Data-Science-Capstone/blob/main/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

https://github.com/wpf-ian/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Performed SQL queries:

• Displaying the names of the unique launch sites in the space mission

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

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

• Displaying 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 the months in

year 2015

• Ranking 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

Build an Interactive Map with Folium

- Launch Sites Locations Analysis with Folium
- Folium maps mark Launch Sites, successful and unsuccessful landings, and a proximity example to key locations: Railway, Highway, Coast and City.
- Allow us to understand why the launch sites may be located where they are.

https://github.com/wpf-ian/Applied-Data-Science-Capstone/blob/main/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

https://github.com/wpf-ian/Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py

- Launch Sites Dropdown List:
- - Added a dropdown list to enable Launch Site selection.
- Pie Chart showing Success Launches (All Sites/Certain Site):
- - Added a pie chart to show the total successful launches count for all sites and the
- Success vs. Failed counts for the site, if a specific Launch Site was selected.
- Slider of Payload Mass Range:
- - Added a slider to select Payload range.
- Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:
- - Added a scatter chart to show the correlation between Payload and Launch Success.

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
- 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

https://github.com/wpf-ian/Applied-Data-Science-Capstone/blob/main/Machine%20Learning%20Prediction.ipynb

Results

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

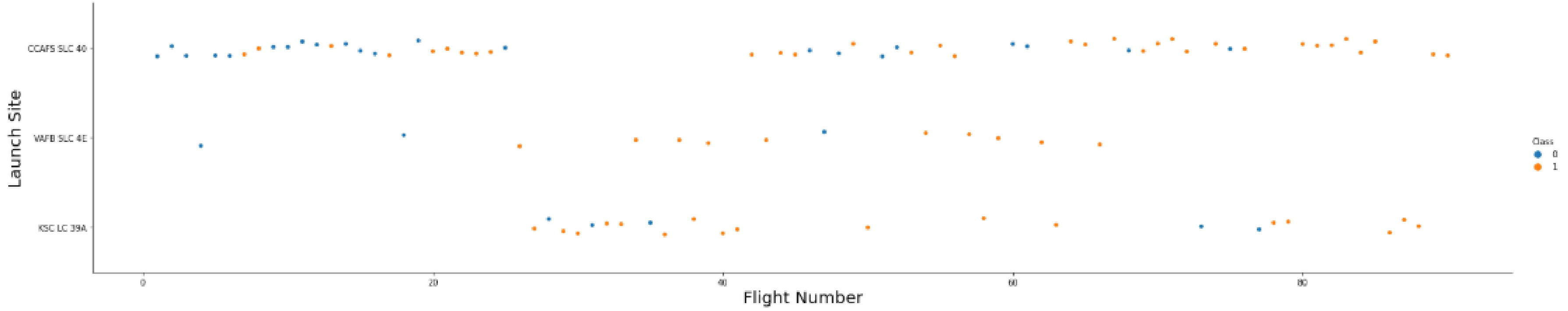
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

In [4]:

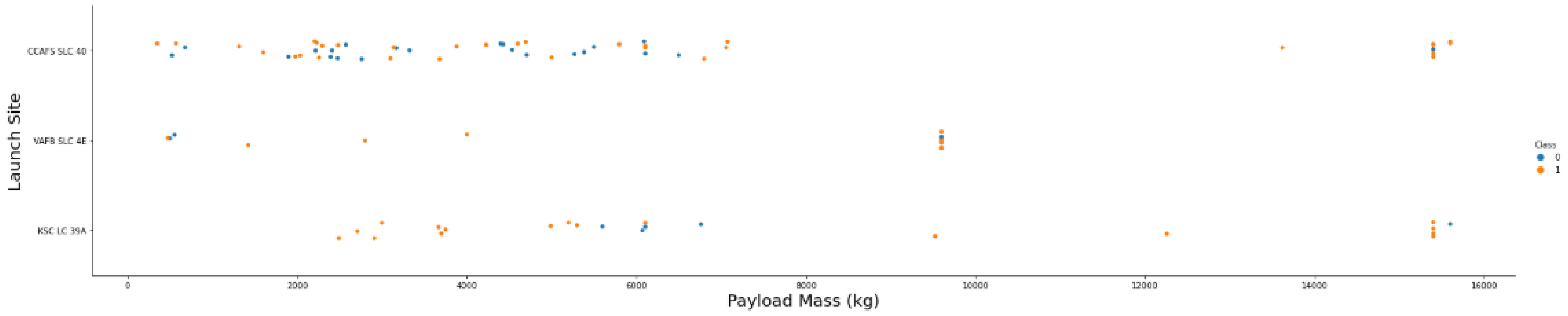
```
# Plot a scatter point chart with x axis to be Flight Number and y axis to be the launch site, and hue to be the class va
sns.catplot(x='FlightNumber', y='LaunchSite', hue='Class', data=df, aspect=5)
plt.xlabel('Flight Number', fontsize=20)
plt.ylabel('Launch Site', fontsize=20)
plt.show()
```



Orange indicates success while blue indicates failure. The graph suggests that most of the launches in the early stage were unsuccessful while the success rate increases over time. CCAFS appears to be the main launch site due to its large volume.

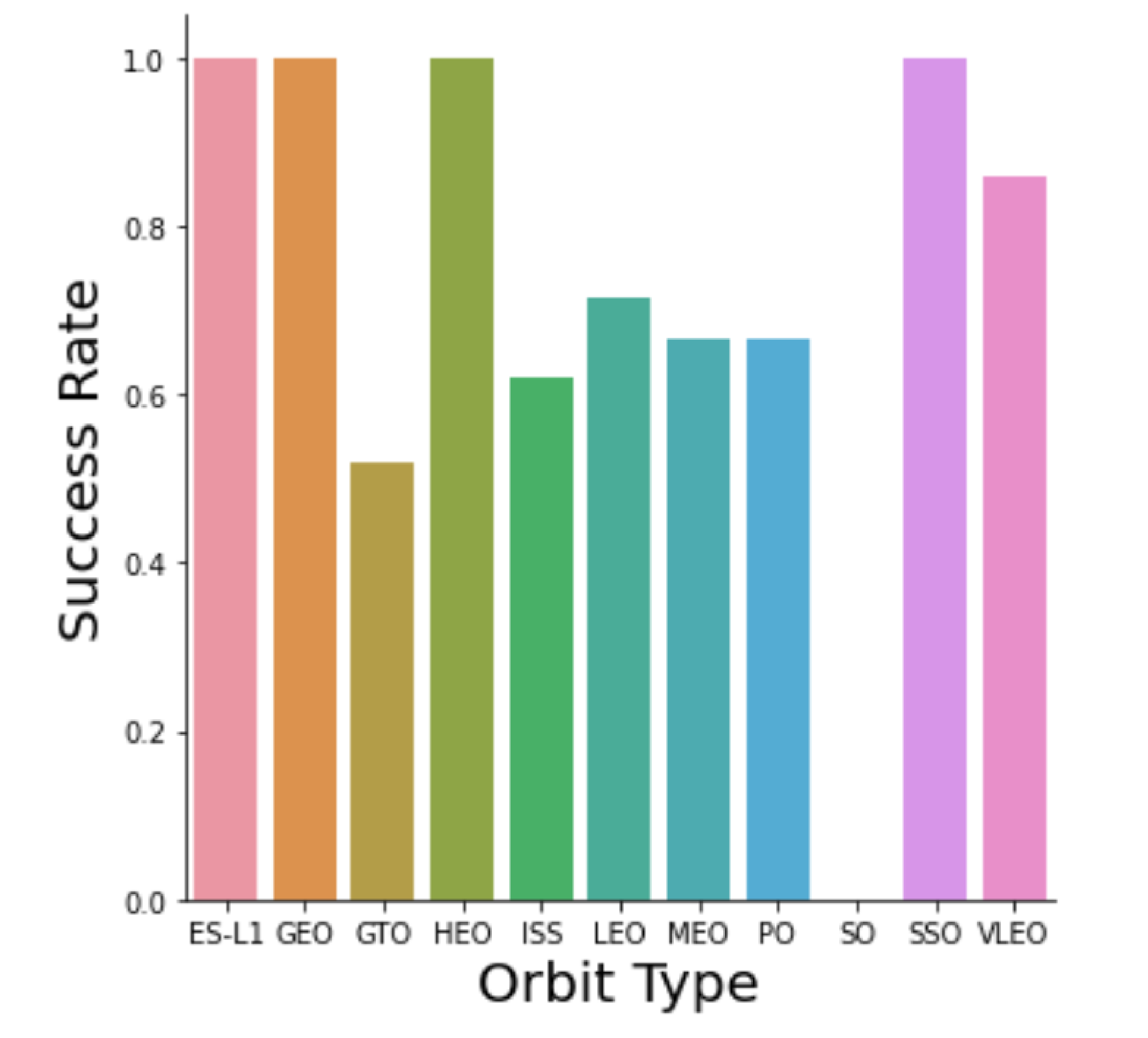
Payload vs. Launch Site

```
# Plot a scatter point chart with x axis to be Pay Load Mass (kg) and y axis to be t
sns.catplot(x='PayloadMass', y='LaunchSite', hue='Class', data=df, aspect = 5)
plt.xlabel('Payload Mass (kg)',fontsize=20)
plt.ylabel('Launch Site',fontsize=20)
plt.show()
```



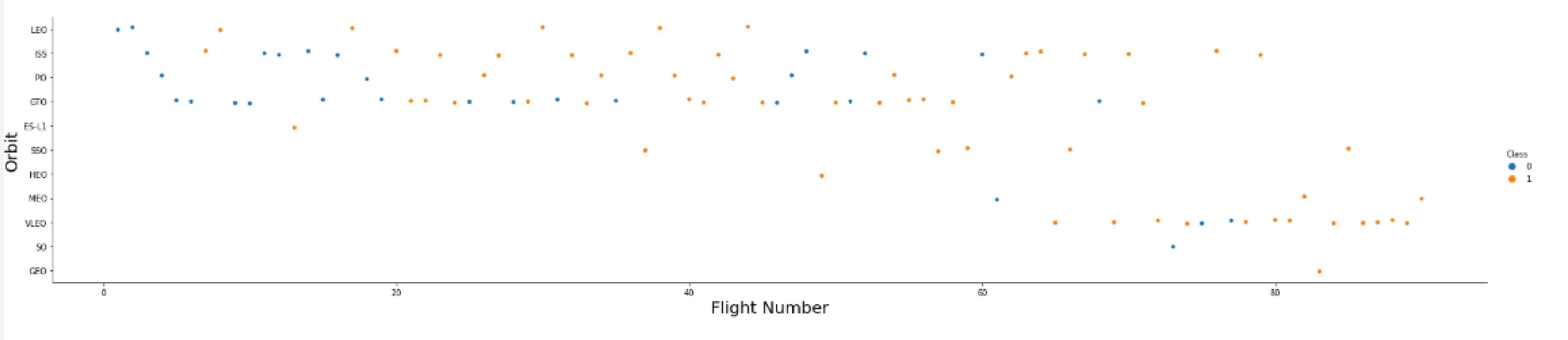
The higher the payload mass, the higher the success rate.
Most of the launches with payload mass over 7000 kg were successful.

Success Rate vs. Orbit Type



- Orbits with 100% success rate:
- ES-L1, GEO, HEO, SSO
- Orbits with 0% success rate:
- SO
- Orbits with success rate between 50% and 85%:
- GTO, ISS, LEO, MEO, PO

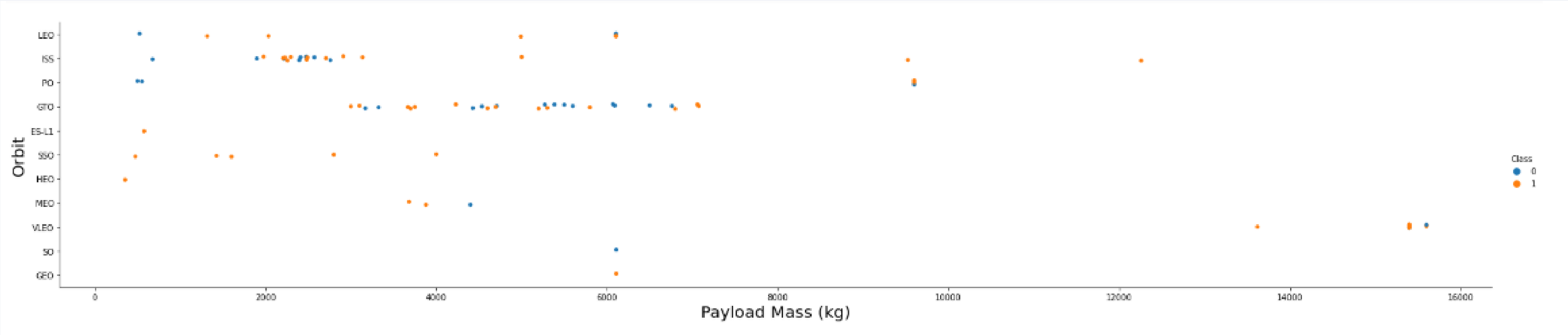
Flight Number vs. Orbit Type



Explanation:

• 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.

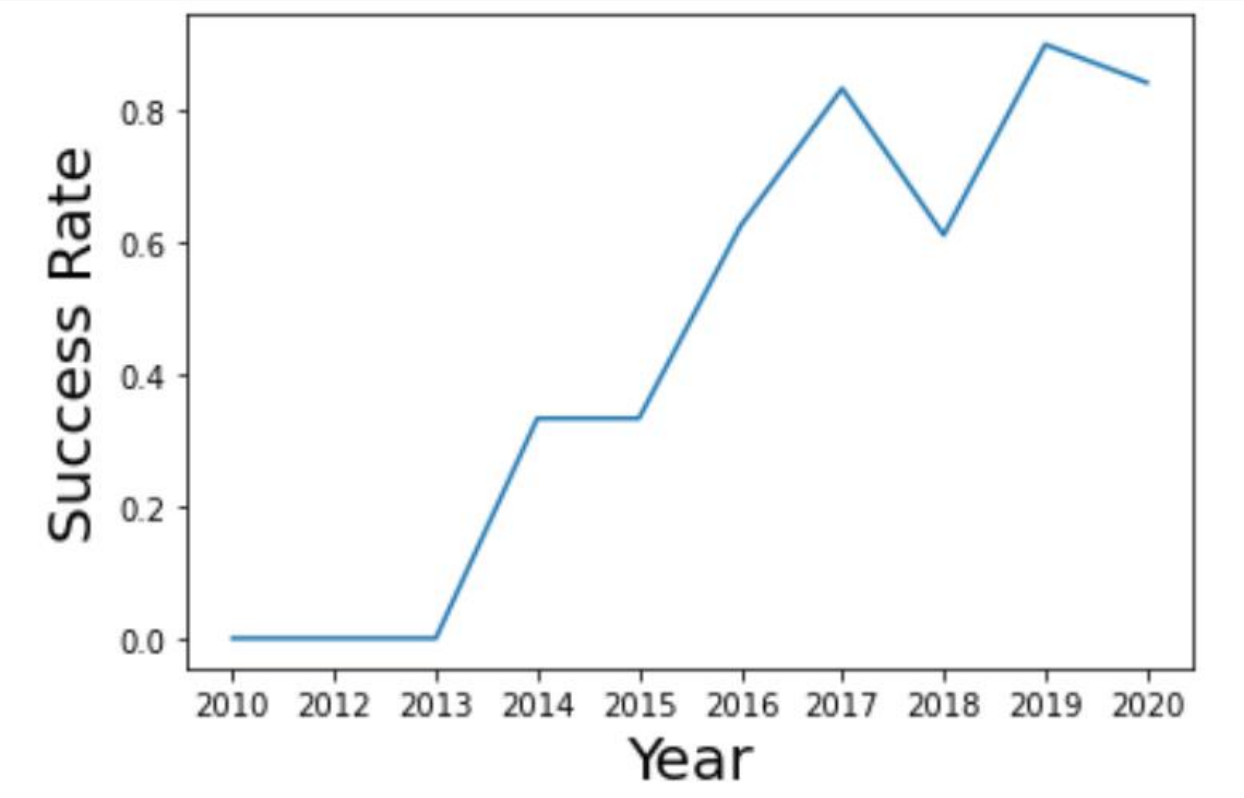
Payload vs. Orbit Type



Explanation:

- Heavy payloads have a negative influence on GTO orbits.

Launch Success Yearly Trend



Explanation:
The success rate keep increasing since 2013.

All Launch Site Names

In [17]:
```
%sql select distinct(LAUNCH_SITE) from SPACEXTBL
```

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

Out[17]:

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

Explanation:
Displaying the names of the unique launch sites in the space mission.

Launch Site Names Begin with 'CCA'

```
In [18]: %sql SELECT Launch_Site FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5;

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

Out[18]:

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

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

Total Payload Mass

```
In [19]: %sql SELECT SUM(PAYLOAD_MASS__KG_) AS total_payload FROM SPACEXTBL WHERE Customer LIKE 'NASA (CRS)';

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

Out[19]:

total_payload
45596

Explanation:
Displaying the total payload mass carried by boosters launched by NASA (CRS).

Average Payload Mass by F9 v1.1

In [20]:
```
%sql SELECT avg(PAYLOAD_MASS__KG_) AS Avg_Payload FROM SPACEXTBL WHERE Booster_Version LIKE 'F9 v1.1';
```

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

Out[20]:

Avg_Payload
2928.4

Explanation:

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

First Successful Ground Landing Date

In [21]:
```
%sql SELECT min(date) AS Early_Date from SPACEXTBL where Landing_Outcome LIKE 'Success (ground pad)'
```

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

Out[21]:

Early_Date
2015-12-22

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

Successful Drone Ship Landing with Payload between 4000 and 6000

```
In [22]: %sql SELECT DISTINCT Customer, Landing_Outcome,PAYLOAD_MASS__KG_ FROM SPACEXTBL WHERE Landing_Outcome ='Success (drone ship)

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

Out[22]:

Customer	Landing_Outcome	PAYLOAD_MASS__KG_
SKY Perfect JSAT Group	Success (drone ship)	4696
SKY Perfect JSAT Group	Success (drone ship)	4600
SES	Success (drone ship)	5300
SES EchoStar	Success (drone ship)	5200

Explanation:
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

In [23]:
```
%sql select count(MISSION_OUTCOME) from SPACEXTBL where MISSION_OUTCOME = 'Success' or MISSION_OUTCOME = 'Failure (in flight
```

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

Out[23]:

count(MISSION_OUTCOME)
99

Explanation:
Listing the total number of successful and failure mission outcomes.

Boosters Carried Maximum Payload

```
In [24]: %sql select BOOSTER_VERSION from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)

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

Out[24]:

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

Explanation:
Listing the names of the booster versions which have carried the maximum payload mass.

2015 Launch Records

```
In [12]: %%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET
              where landing__outcome = 'Failure (drone ship)' and year(date)=2015;

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[12]:

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Explanation:
Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015.

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

In [13]:

```
%%sql select landing__outcome, count(*) as count_outcomes from SPACEXDATASET
      where date between '2010-06-04' and '2017-03-20'
      group by landing__outcome
      order by count_outcomes desc;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[13]:

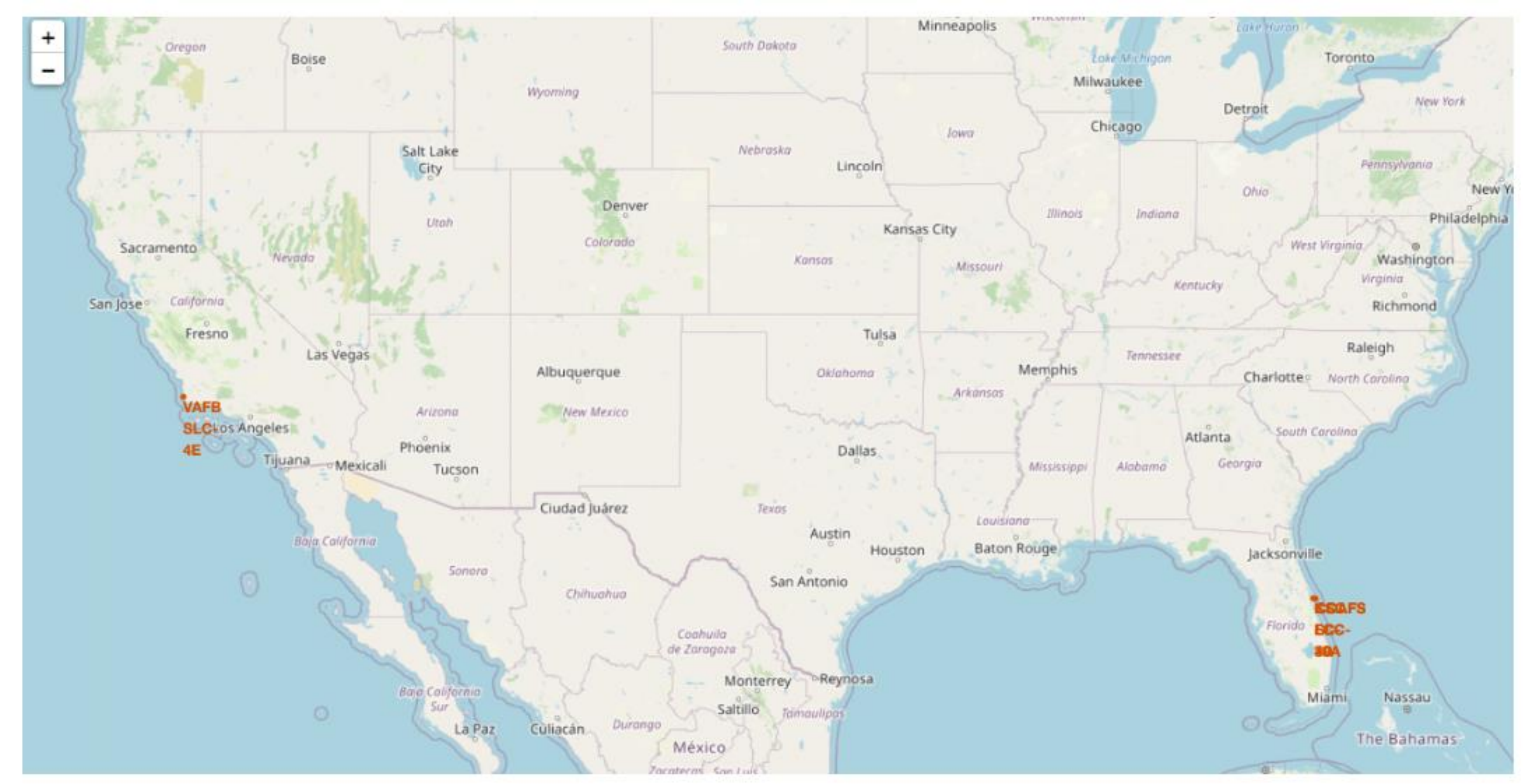
landing__outcome	count_outcomes
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Explanation:
Ranking 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.

Section 3

Launch Sites Proximities Analysis

<Launch Site Map>

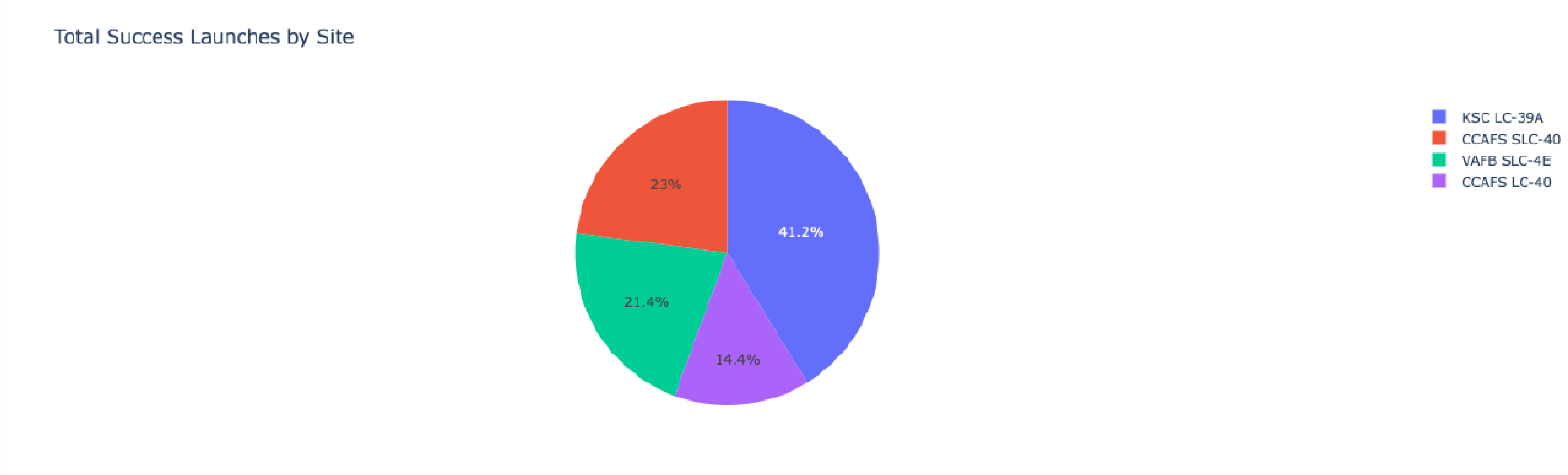


The map shows all launch sites in the US map.

Section 4

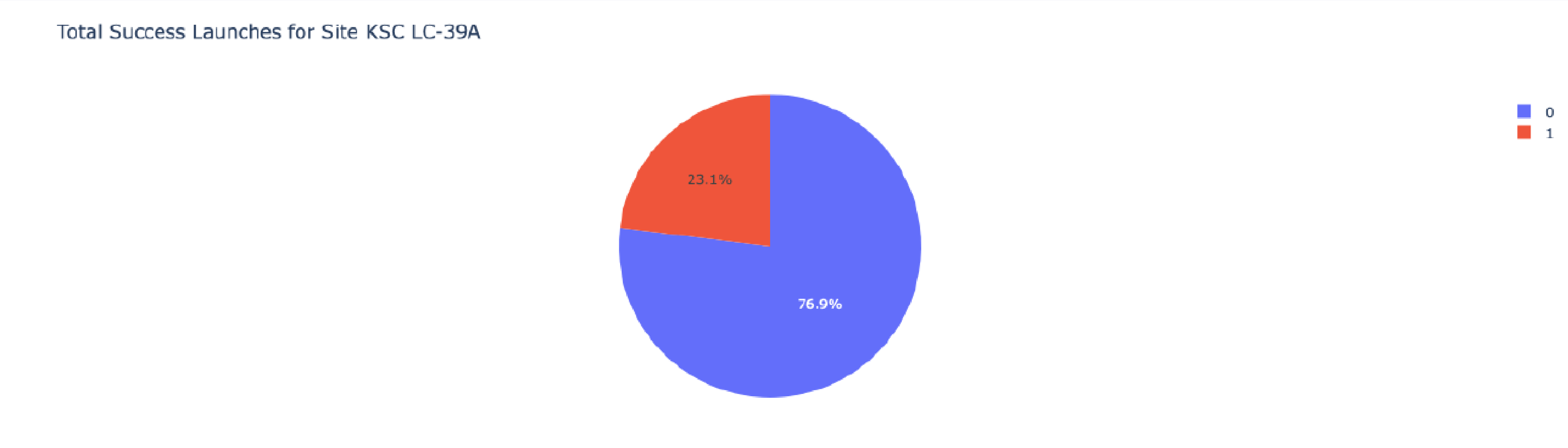
Build a Dashboard with Plotly Dash

Successful Launches across all sites



Explanation:
The chart clearly shows that KSC LC-39A has the most successful launches.

Launch Site with the highest successful ratio



Explanation:
KSC LC-39A has the highest launch success rate (76.9%).

Payload Launch vs Success for all sites

Explanation:

The charts show that payloads between 2000 and 5500 kg have the highest success rate.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

```
print("tuned hpyerparameters :(best parameters) ",logreg_cv.best_params_)
print("accuracy :",logreg_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}
accuracy : 0.8464285714285713
```

```
print("tuned hpyerparameters :(best parameters) ",svm_cv.best_params_)
print("accuracy :",svm_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}
accuracy : 0.8482142857142856
```

```
print("tuned hpyerparameters :(best parameters) ",tree_cv.best_params_)
print("accuracy :",tree_cv.best_score_)
```

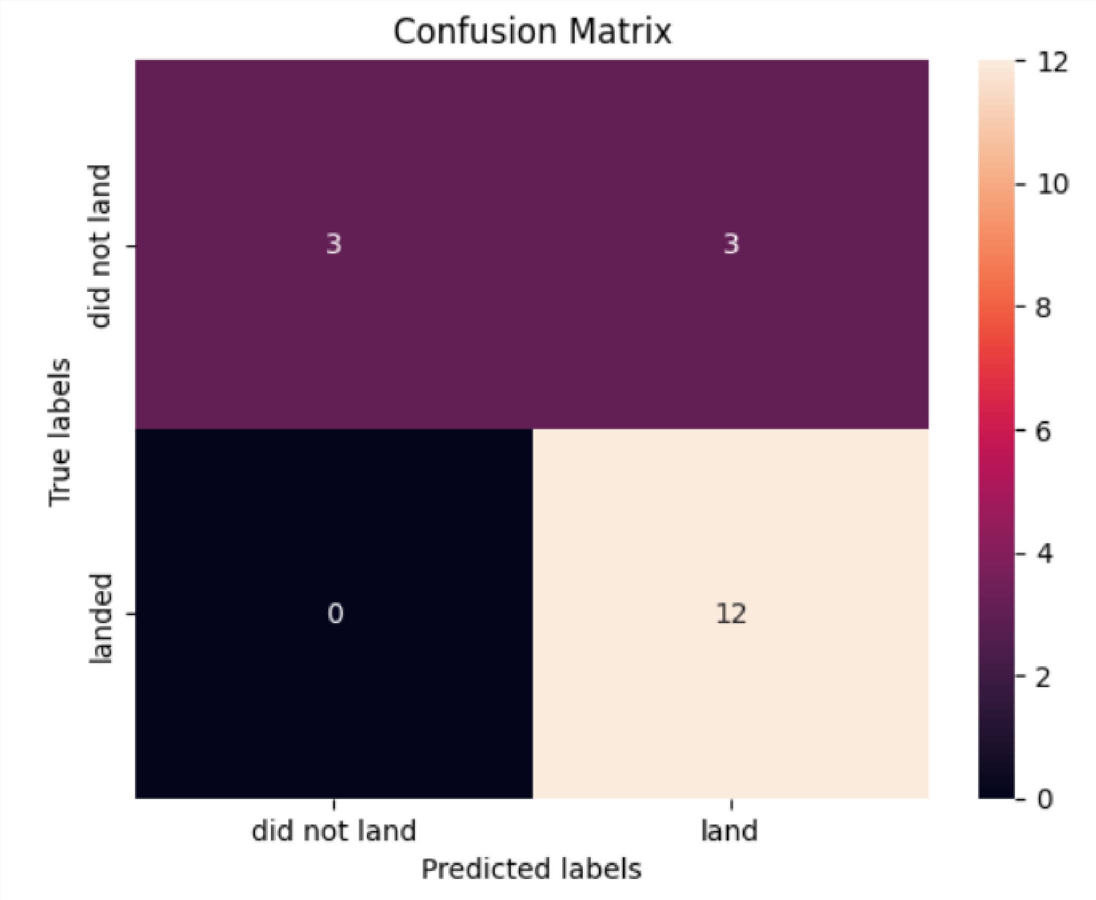
```
tuned hpyerparameters :(best parameters)  {'criterion': 'gini', 'max_depth': 10, 'max_features': 'sqrt',
'min_samples_leaf': 2, 'min_samples_split': 10, 'splitter': 'random'}
accuracy : 0.875
```

```
print("tuned hpyerparameters :(best parameters) ",knn_cv.best_params_)
print("accuracy :",knn_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}
accuracy : 0.8482142857142858
```

Tree model has the best accuracy among all machine learning model

Confusion Matrix



- The models predicted 12 successful landings when the true label was successful landing.
- The models predicted 3 unsuccessful landings when the true label was unsuccessful landing.
- The models predicted 3 successful landings when the true label was unsuccessful landings(false positives).

Conclusions

Decision Tree Model is the best algorithm for this dataset.

- Launches with a low payload mass show better results than launches with a larger payload mass.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- The success rate of launches increases over the years.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.

Appendix

- Instructors
- Coursera
- IBM

Thank you!