Winning Space Race with Data Science

Falcon 9 First Stage Landing Prediction

Do Thanh Pham

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection: Gather through SpaceX public API and Web Scraping the data from Wikipedia.
  - Data wrangling: Extract launch outcome details, which served as the dependent variable in the Machine Learning models.
  - Exploratory Data Analysis: SQL queries and data visualizations including static plots, interactive maps.
  - Predictive analysis and ML model: Logistic Regression, Support Vector Machine (SVM), Decision Tree, and k-Nearest Neighbors (KNN) ML models.
- Summary of all results
  - Launch data include info about flight number, date of launch, payload mass, orbit type, launch site, mission outcome and other variables.
  - Logistic Regression, SVM (Support Vector Machine), Decision Tree and KNN (k-Nearest Neighbors) all perform equally well for Machine Learning models on this dataset.

Introduction

- Project background and context
    - Goal: Develop a machine learning pipeline to predict the successful landing of the first stage of SpaceX's Falcon 9 rocket launches.
    - Importance: SpaceX offers launches at $62 million, much lower than competitors, due to first stage reusability.
    - Objective: Predicting first stage landings helps estimate launch costs, aiding companies in bidding against SpaceX.
- Problems I want to find answers
    - What is the nature and extent of the data of SpaceX Falcon 9 first stage landings?
    - Which machine learning model is most suitable for accurately predicting the outcome of a Falcon 9 first stage landing in future launches?

Section 1

Methodology

Methodology

Executive Summary

- Data was collected using SpaceX API and web scraping from Wikipedia.
- Data was cleaned in preparation for visualizations, queries and machine learning model creation.
- Exploratory data analysis (EDA) using visualization and SQL.
- Interactive visual analytics were created using Folium.
- Predictive analysis using classification models.

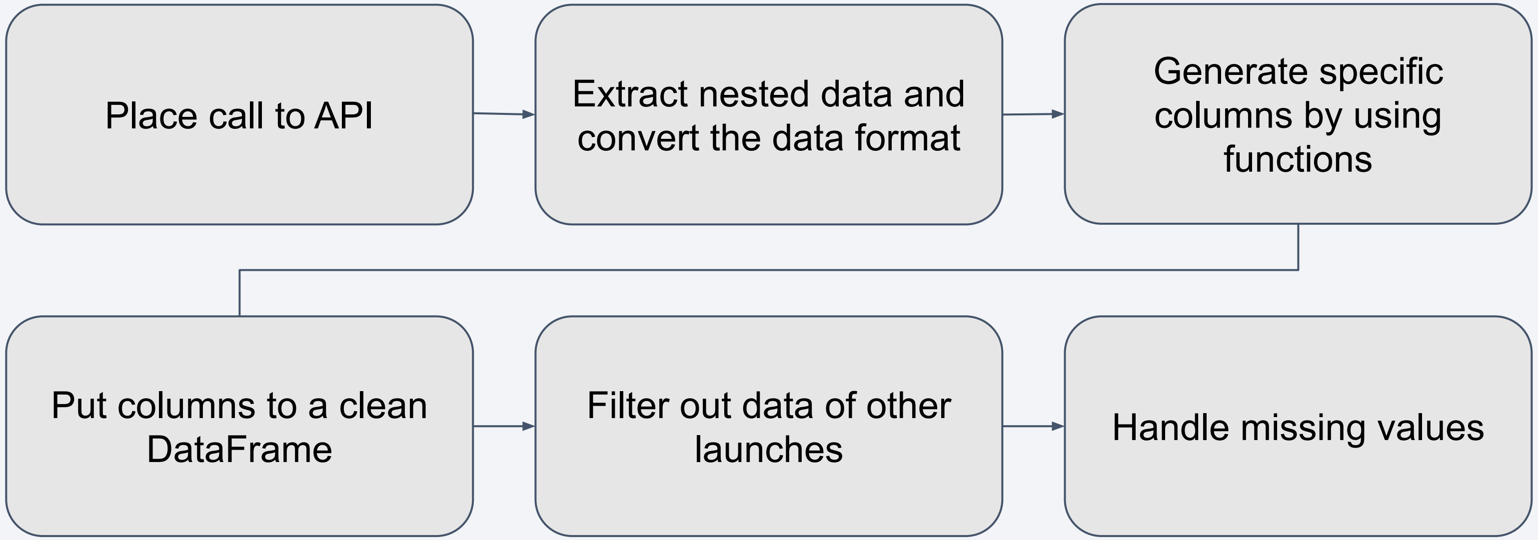
Data Collection

- Data collection involved sending GET requests to the SpaceX API.
- The response content was decoded as JSON using the .json() function and converted into a pandas dataframe using .json_normalize().
- Data cleaning was conducted, including checking for and filling in missing values as needed.
- Web scraping was performed on Wikipedia to obtain Falcon 9 launch records using BeautifulSoup.
- The goal was to extract launch records from an HTML table, parse the table, and convert it into a pandas dataframe for subsequent analysis.

Data Collection – SpaceX API

- The SpaceX API offers public data.
- After retrieving a response from a GET request to the SpaceX API, the data is loaded into a Pandas DataFrame.
- GitHub URL: [Link_here]

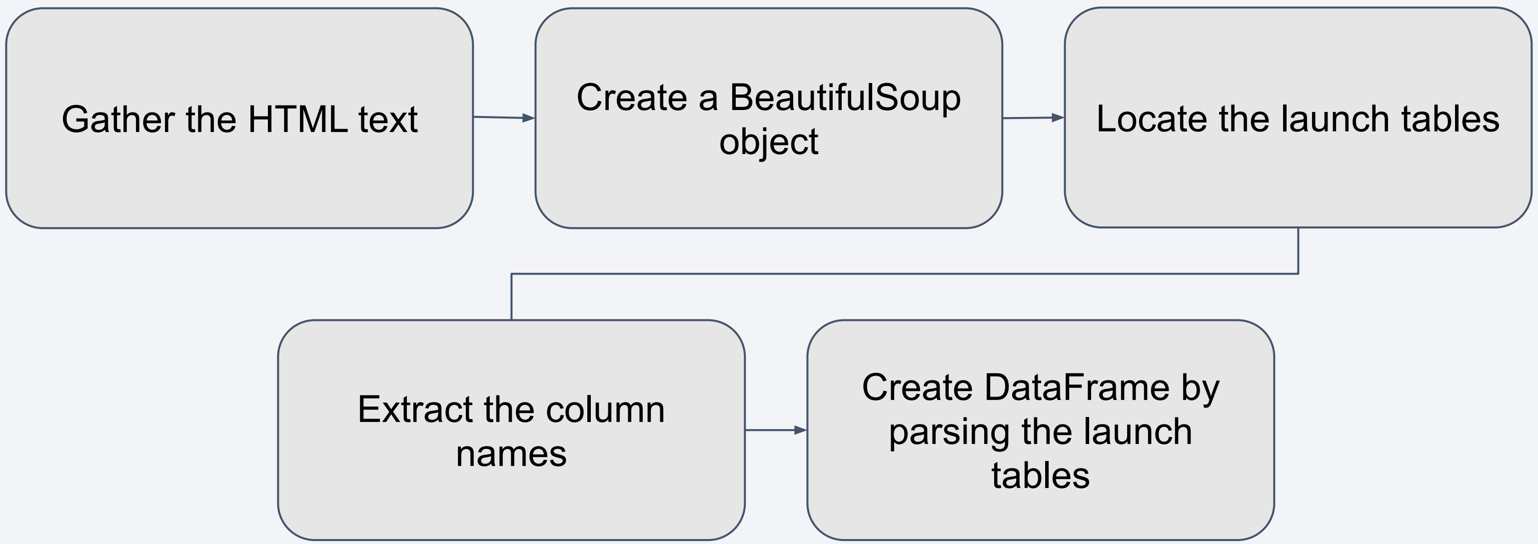
Flowchart of SpaceX API Calls



Data Collection - Scraping

- After scraping data from Wikipedia, the extracted data can be structured into a Pandas DataFrame for detailed analysis.
- GitHub URL (Web Scraping): [Link_here]

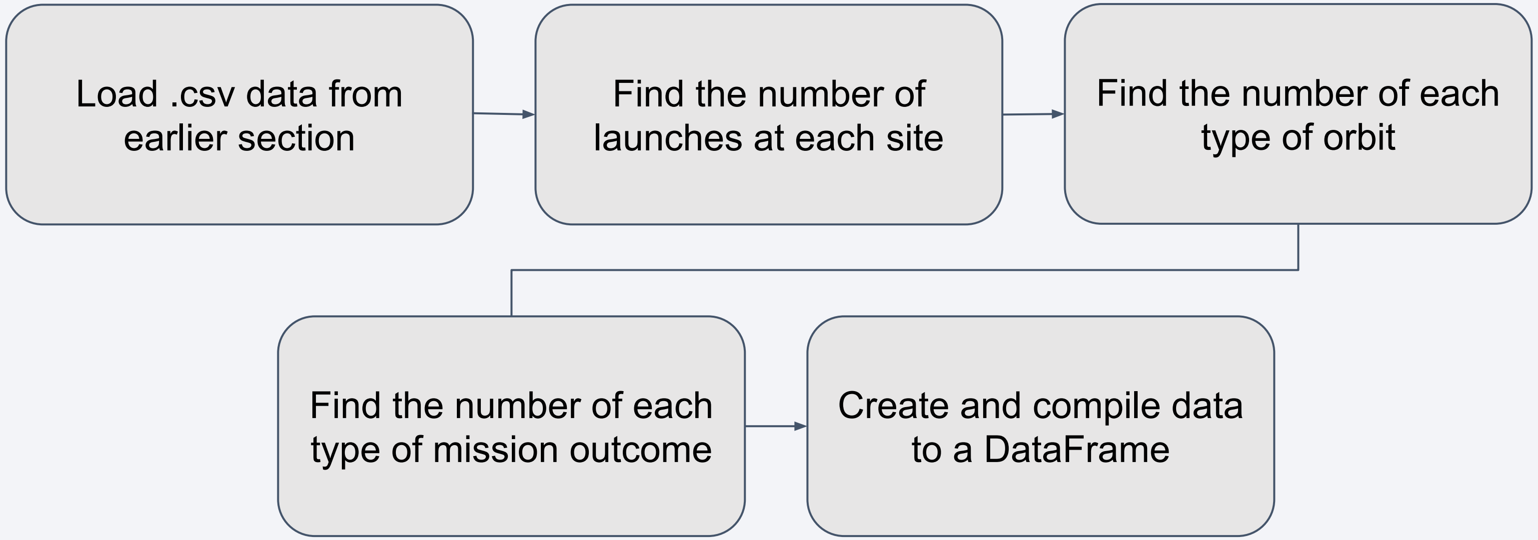
Flowchart of Web Scraping



Data Wrangling

- The data from the initial section is stored in a .csv file and requires cleaning.
- Cleanup involved refining launch sites, orbit types, and mission outcomes.
- GitHub URL (Data Wrangling): [Link_here]

Flowchart of Data Wrangling



EDA with Data Visualization

- The following charts were created to look at Launch Site trends
  - Scatterplot to see mission outcome relationship split by Launch Site and Flight Number.
  - Scatterplot to see mission outcome relationship split by Launch Site and Payload.
- The following charts were created to look at Orbit Type trends
  - Bar chart to see mission outcome relationship with Orbit Type.
  - Scatterplot to see mission outcome relationship split by Orbit Type and Flight Number.
  - Scatterplot to see mission outcome relationship split by Orbit Type and Payload.
- The following chart was created to look at trends based on time.
- Line plot to see mission outcome trend by year.
- GitHub URL(EDA with Data Visualization): [Link_here]

EDA with SQL

- Queries were written to extract information about:
  - Launch sites
  - Payload masses
  - Dates
  - Booster types
  - Mission outcomes
- GitHub URL (EDA with SQL): [Link_here]

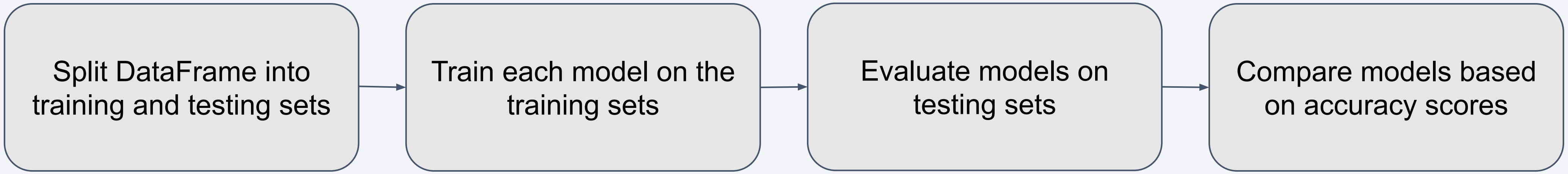
Build an Interactive Map with Folium

- Here's a summary of the map objects, including markers, circles, lines, etc., that I created and added to the Folium map:
  - Markers: added for launch sites and for the NASA Johnson Space Center
  - Circles: added for the launch sites.
  - Lines: added to show the distance to the nearby features:
    - Distance from CCAFS LC-40 to the coastline.
    - Distance from CCAFS LC-40 to the rail line.
    - Distance from CCAFS LC-40 to the perimeter road.
- GitHub URL (Folium Maps): [Link_here]

Predictive Analysis (Classification)

- Logistic Regression, SVM, Decision Tree, and KNN machine learning models were trained using the training data.
- Hyperparameters were optimized using GridSearchCV(), and the best parameters were selected based on '.best_params_'.
- Each model's accuracy was assessed using the testing data set with the best hyperparameters.
- GitHub URL (Machine Learning): [Link_here]

Flowchart of Machine Learning



Results

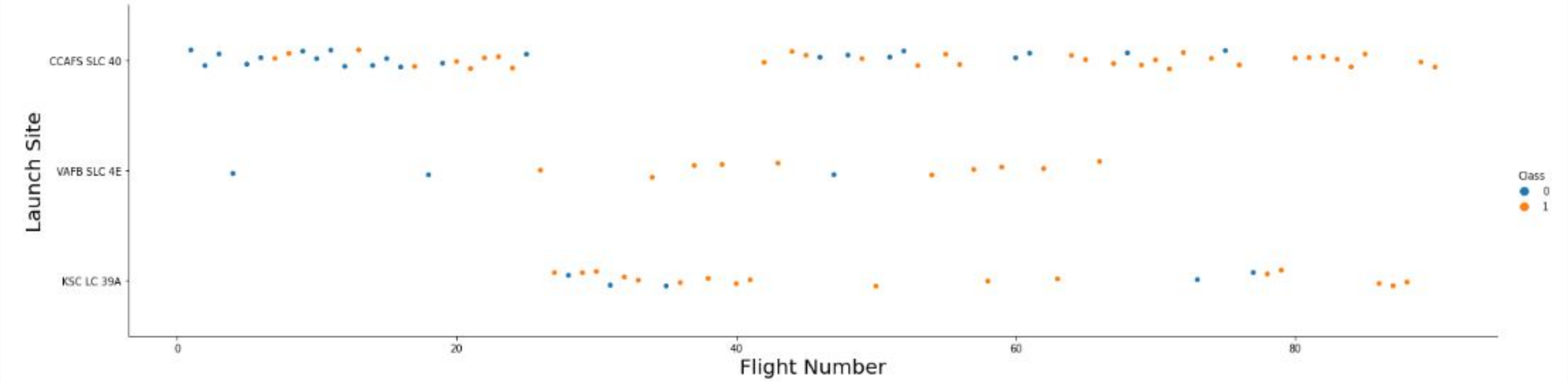
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

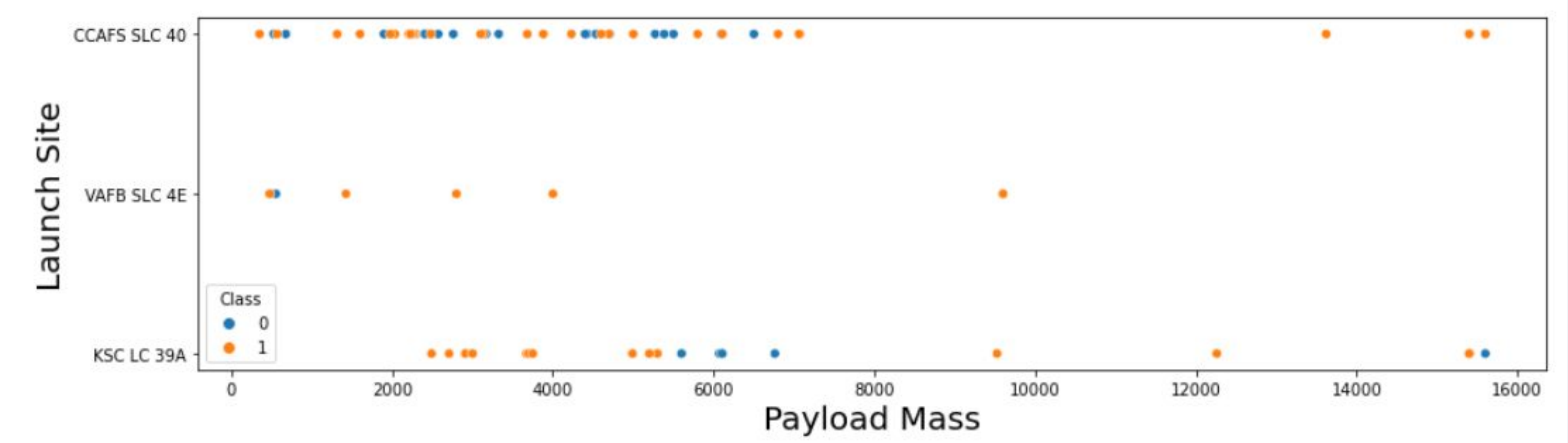
Flight Number vs. Launch Site

- Success rate varies noticeably with launch site.
- Successful Falcon 9 first stage landings appear to become more prevalent as the flight number increases.
- Failed landings are indicated by the ‘0’ Class (blue markers) and successful landings by the ‘1’ Class (orange markers).



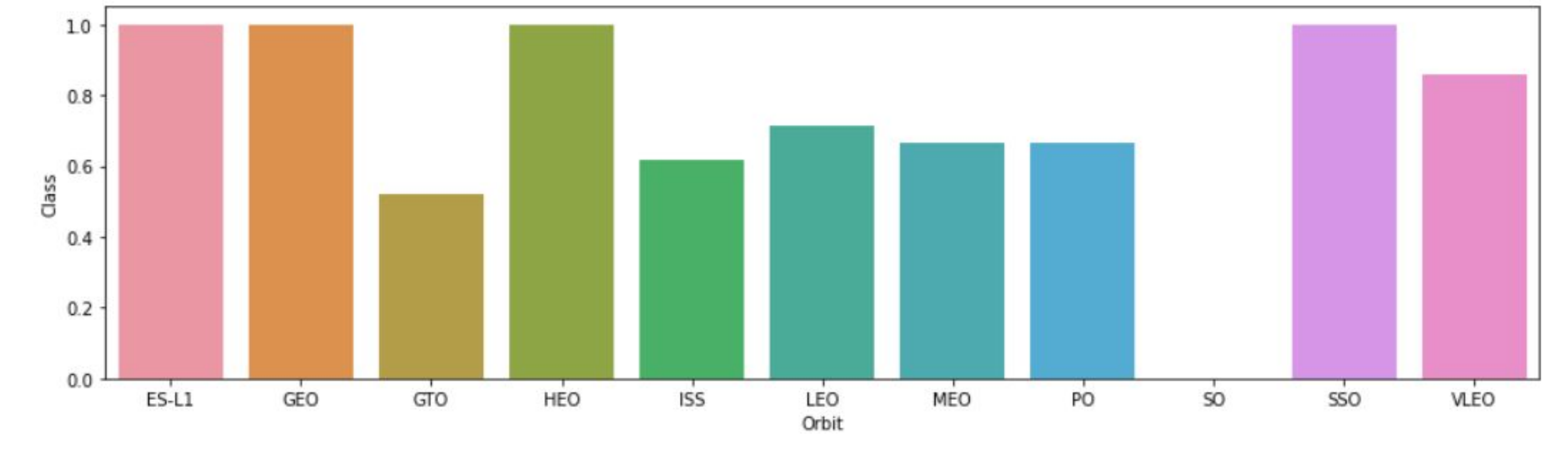
Payload vs. Launch Site

- CCAFS SLC 40 launch site: Payload mass and landing outcome correlation is weak.
- KSC LC 39A launch site: Failed landings cluster around specific payload mass range (6000).
- VAFB-SLC launchsite there are no rockets launched for heavy payload mass (greater than 10000).



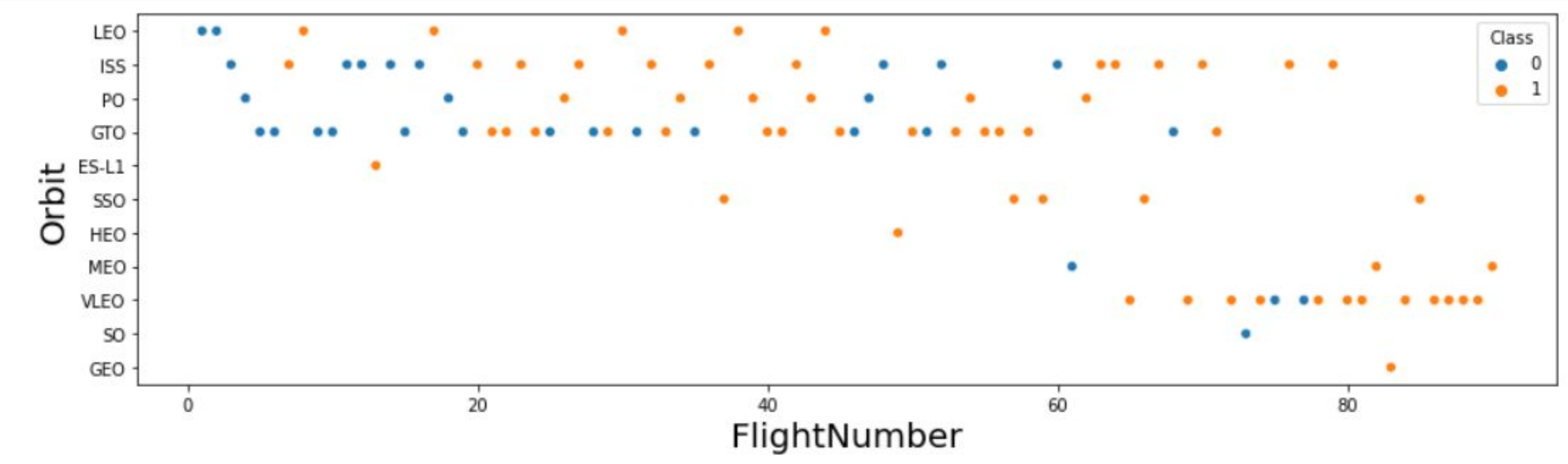
Success Rate vs. Orbit Type

- ES-L1, SSO, HEO, and GEO orbits: No failed first stage landings.
- SO orbits: No successful first stage landings.



Flight Number vs. Orbit Type

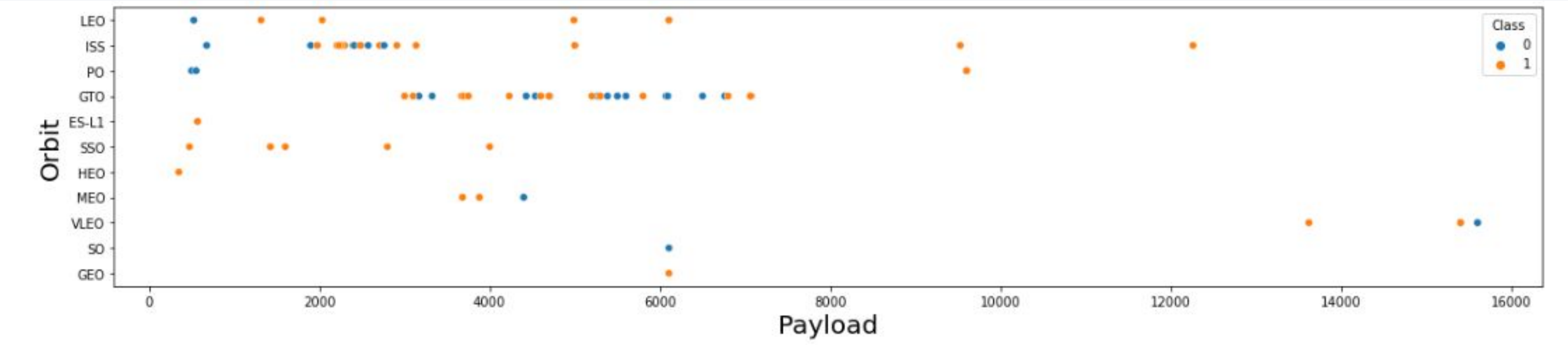
- There is no apparent relationship between flight number and success.
- In the LEO orbit, success is related to the number of flights whereas in the GTO orbit, there is no relationship between flight number and the orbit.



Payload vs. Orbit Type

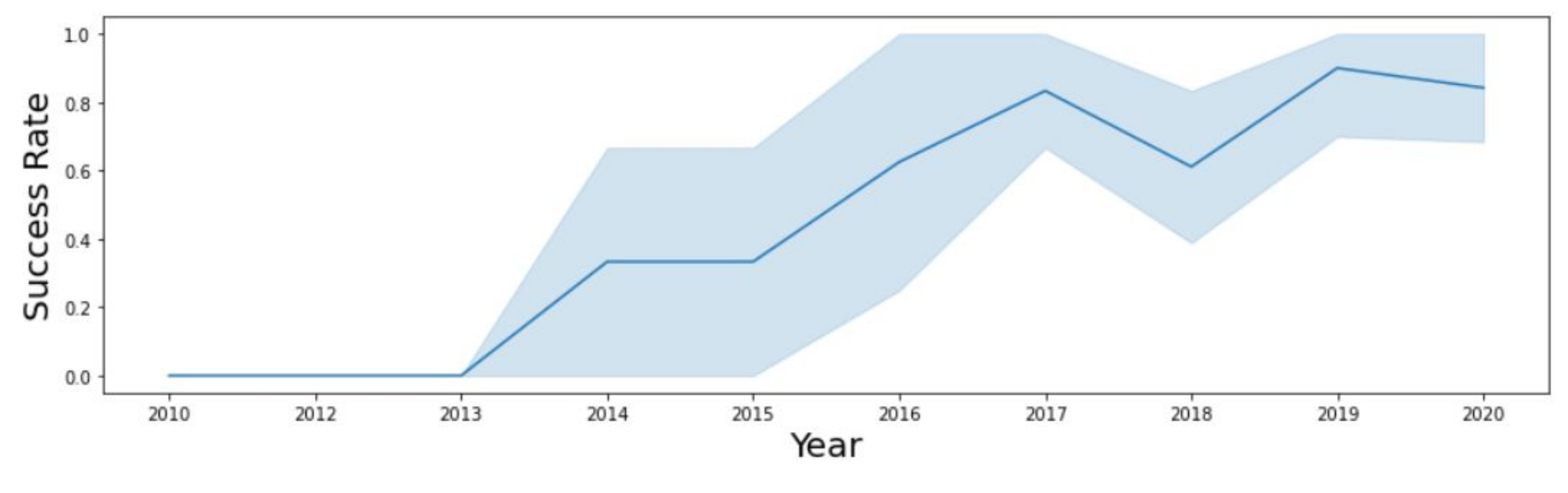
Certain orbit types (PO, LEO, ISS) exhibit higher success rates than others. In the contrast, GTO, distinguishing between successful and unsuccessful landings is challenging due to their frequent occurrence.

-> No clear correlation between success rate and payload mass.



Launch Success Yearly Trend

The success rate has increased significantly from 2013 to 2020.



All Launch Site Names

The query is to display the unique launch site names.

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

```
%%sql
select DISTINCT(LAUNCH_SITE) from SPACEX
```

```
 * ibm_db_sa://bqn92294:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:31249/bludb
Done.
```

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

Launch Site Names Begin with 'CCA'

The query is used to display 5 records where launch sites begin with 'CCA'.

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

```
%%sql
select LAUNCH_SITE from SPACEX
where LAUNCH_SITE LIKE 'CCA%'
limit 5
```

```
 * ibm_db_sa://bqn92294:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.c
loud:31249/bludb
Done.
```

launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

The total payload carried by boosters from NASA is 321400.

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

In [10]:

```
%%sql
select SUM(PAYLOAD_MASS__KG_) from SPACEX
where LAUNCH_SITE LIKE 'CCA%'
```

```
 * ibm_db_sa://bqn92294:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:31249/bludb
Done.
```

Out[10]:

1
321400

Average Payload Mass by F9 v1.1

The query is to display average payload mass carried by booster version F9. v1.1

Task 4

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

```
%%sql
select AVG(PAYLOAD_MASS__KG_) from SPACEX
where BOOSTER_VERSION LIKE 'F9 v1.1'
```

```
 * ibm_db_sa://bqn92294:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.c
loud:31249/bludb
Done.
```

1
2928

First Successful Ground Landing Date

The query to list the date of the first successful landing outcome on ground pad.

```
%%sql
select min(DATE) from SPACEX
where LANDING__OUTCOME LIKE 'Success%'
```

```
 * ibm_db_sa://bqn92294:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:31249/bludb
Done.
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

This query is to list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.

```
%%sql
select BOOSTER_VERSION, PAYLOAD_MASS__KG_  from SPACEX
where (LANDING__OUTCOME LIKE 'Success (drone ship)') AND (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000)
```

```
 * ibm_db_sa://bqn92294:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.c
loud:31249/bludb
Done.
```

booster_version	payload_mass__kg_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

Total Number of Successful and Failure Mission Outcomes

This query is to calculate the total number of successful and failure mission outcomes.

List the total number of successful and failure mission outcomes

```
%%sql
select COUNT(MISSION_OUTCOME) FROM SPACEX
```

```
 * ibm_db_sa://bqn92294:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:31249/bludb
Done.
```

1
202

Boosters Carried Maximum Payload

The query is to list the names of the booster which have carried the maximum payload mass.

```
%%sql
SELECT DISTINCT(BOOSTER_VERSION), PAYLOAD_MASS__KG_ FROM SPACEX
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEX)
```

```
 * ibm_db_sa://bqn92294:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:31249/bludb
Done.
```

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

2015 Launch Records

The query is to list the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015.

```
%%sql
SELECT BOOSTER_VERSION, LAUNCH_SITE, DATE, LANDING__OUTCOME FROM SPACEX
WHERE LANDING__OUTCOME LIKE 'Failure (drone ship)' AND YEAR(DATE)=2015
```

```
 * ibm_db_sa://bqn92294:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:31249/bludb
Done.
```

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

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 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 count FROM SPACEX
WHERE DATE BETWEEN '2010-06-04'and'2017-03-20'
GROUP BY landing__outcome
ORDER BY COUNT(landing__outcome) DESC
```

```
 * ibm_db_sa://bqn92294:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:31249/bludb
Done.
```

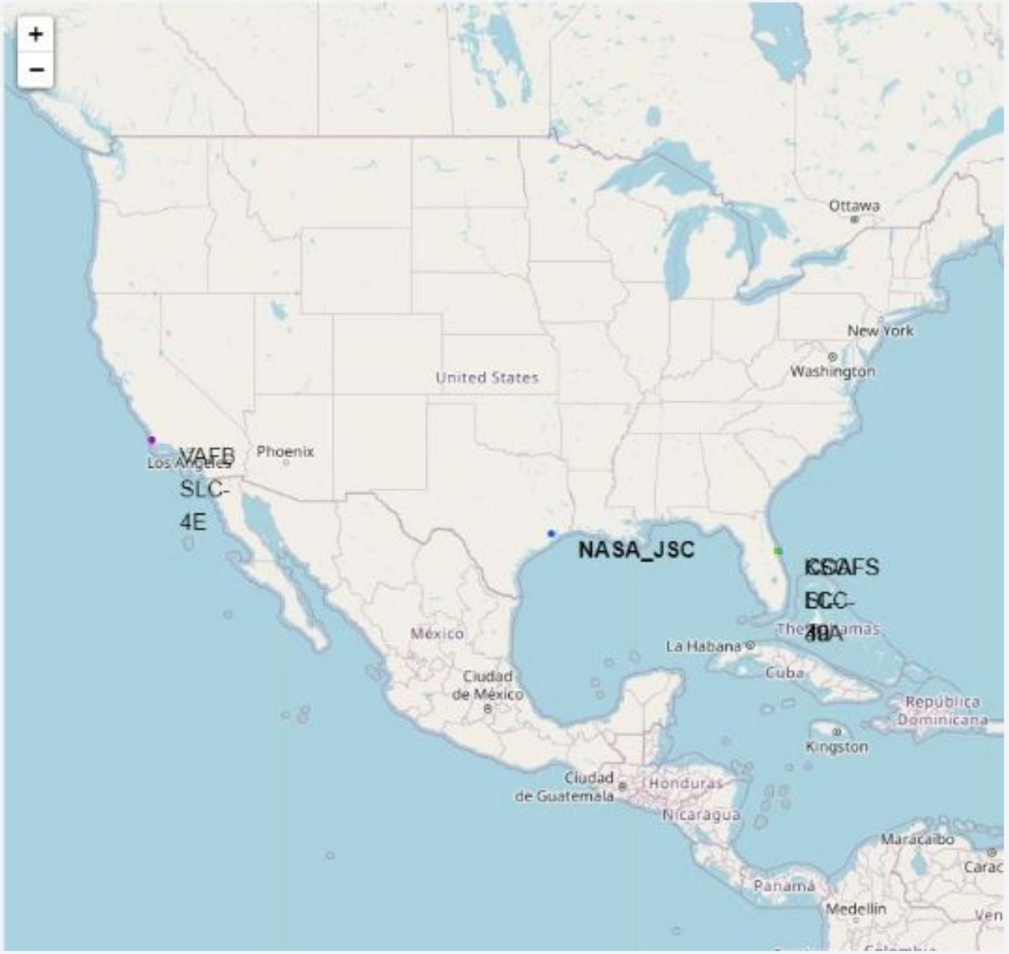
landing__outcome	COUNT
No attempt	20
Failure (drone ship)	10
Success (drone ship)	10
Controlled (ocean)	6
Success (ground pad)	6
Failure (parachute)	4
Uncontrolled (ocean)	4
Precluded (drone ship)	2

Section 3

Launch Sites Proximities Analysis

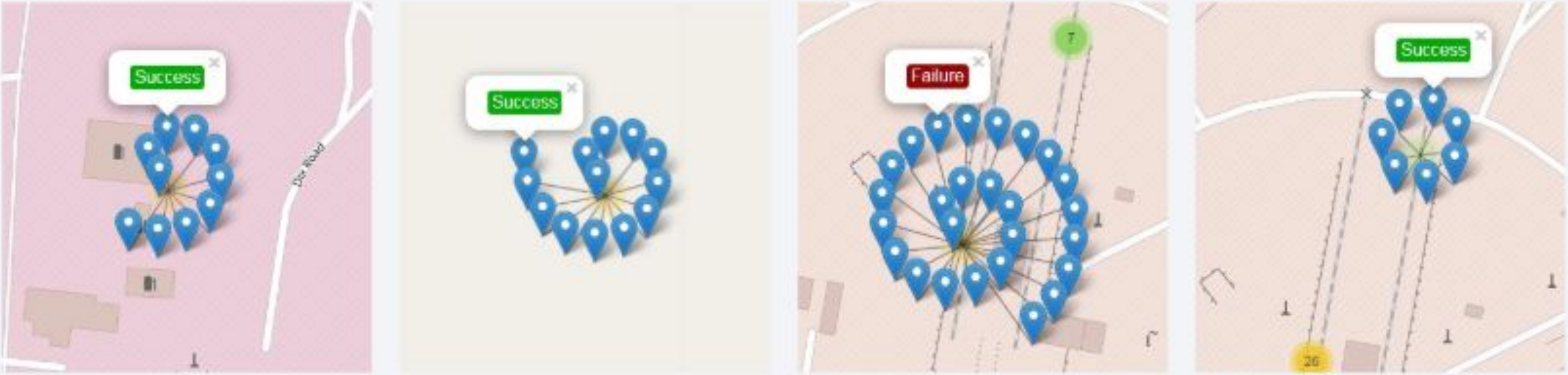
Falcon 9 Launch Site Locations

- VAFB SLC-4E (California, USA)
  - Vandenberg Air Force Base Space Launch Complex 4E
- KSC LC-39A (Florida, USA)
  - Kennedy Space Center Launch Complex 39A
- CCAFS LC-40 (Florida, USA)
  - Cape Canaveral Air Force Station Launch Complex 40
- CCAFS SLC-40 (Florida, USA)
  - Cape Canaveral Air Force Station Space Launch Complex 40



Map Markers of Success/Failed Landings

Markers on the map represent Falcon 9 first stage landing outcomes (Success/Failure) grouped by launch site coordinates, providing an overview of success rates.

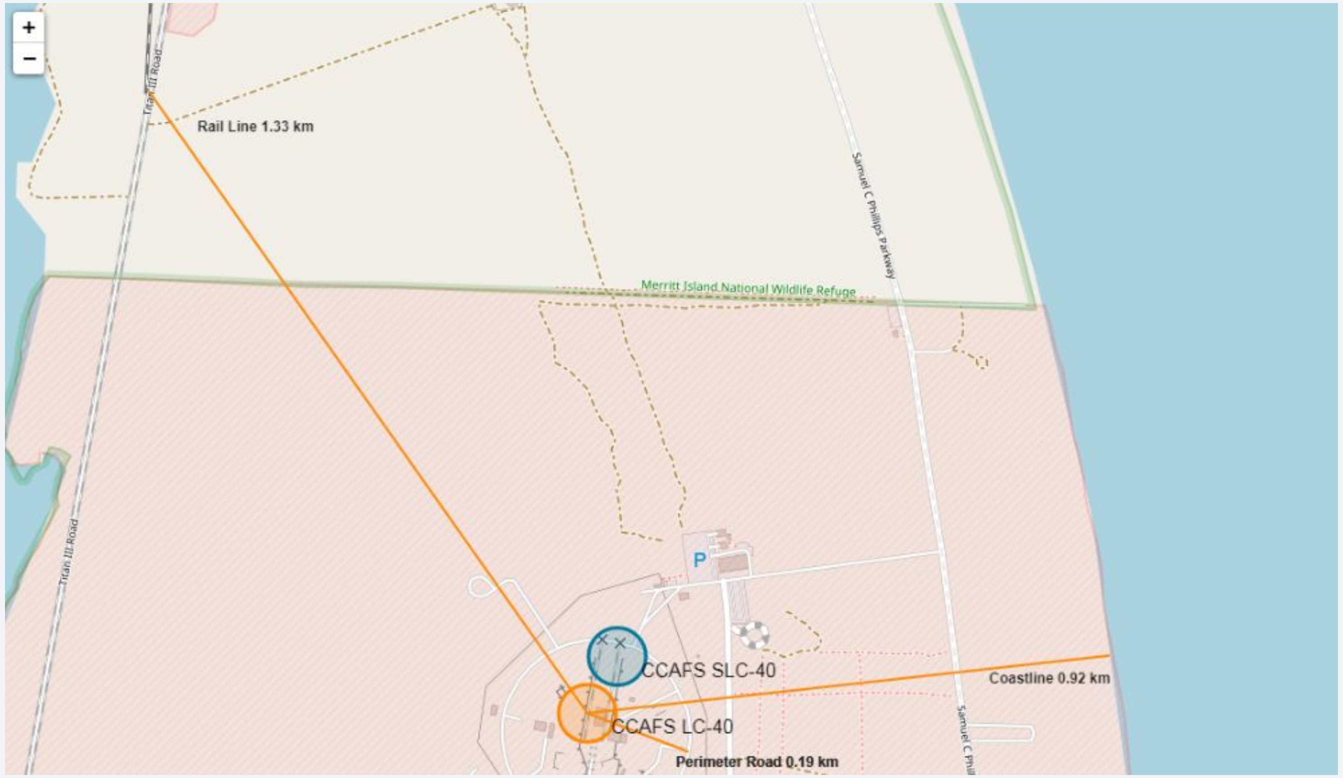


VAFB SLC-4E KSC LC-39A CCAFS LC-40 CCAFS SLC-40

Distance from Launch Site to Proximities

The CCAFS LC-40 and CCAFS SLC-40 launch sites are close but not exactly aligned.

- The perimeter road around CCAFS LC-40 is 0.19 km away.
- The coastline is 0.92 km away.
- The rail line is 1.33 km away from CCAFS LC-40.



Section 4

Predictive Analysis (Classification)

Classification Accuracy

All models performed equally well.

```
model=["knn_cv", "tree_cv", "svm_cv","logreg_cv"]
funct=[knn_cv, tree_cv, svm_cv, logreg_cv]
acc=[]
for x in funct:
    acc.append(x.score(X_test, Y_test))
perf={'Model':model,'Score':acc}
performance= pd.DataFrame.from_dict(perf, orient='columns').set_index('Model')
performance
```

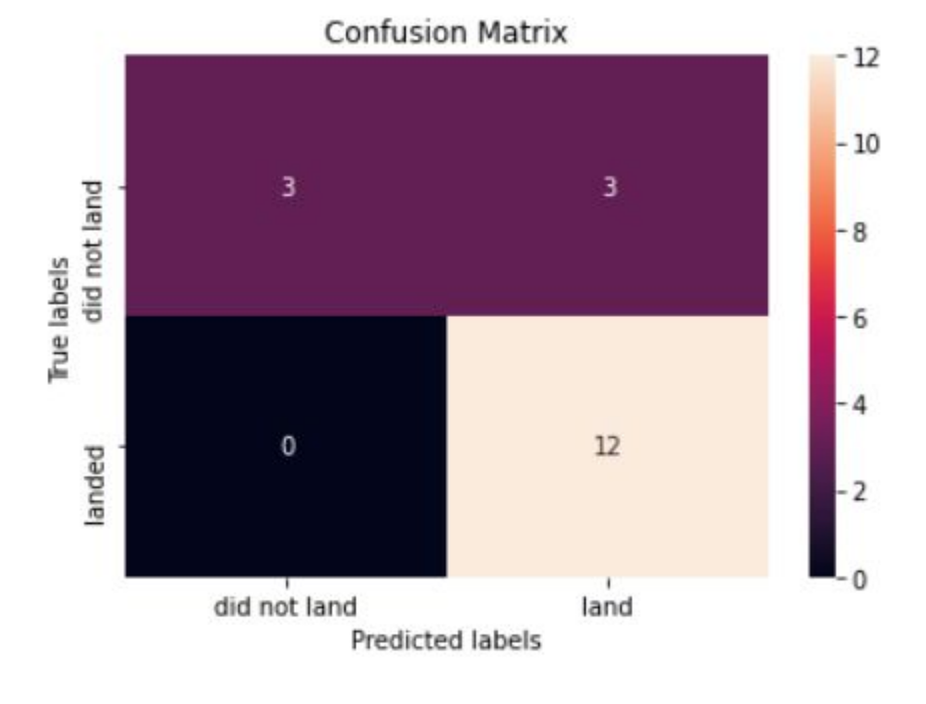
	Score
Model	
knn_cv	0.833333
tree_cv	0.833333
svm_cv	0.833333
logreg_cv	0.833333

Confusion Matrix

The confusion matrix for the Logistic Regression can be read as:

True Negative	False Positive
False Negative	True Positive

Prediction Breakdown:
- 12 True Positives and 3 True Negatives
- 3 False Positives and 0 False Negatives



Conclusions

- The larger the flight amount at a launch site, the greater the success rate at a launch site.
- Launch success rate started to increase in 2013 till 2020.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate, while SO orbits had no successful first stage landings.
- 4 (Logistic Regression, SVM, Decision Tree and KNN) ML model performed equally well on the provided data set.

Thank you!