



IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Cesar Calderon
10/30/2023



Outline

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

Executive Summary

- In this project, we will predict if the Space X Falcon 9 first stage will land successfully using several machine learning classification algorithms.
- The main steps in this projects are:
 - Data collection, wrangling and formatting
 - EDA
 - Interactive Data Visualization
 - ML Prediction
- Some conclusions are:
 - Some features of the launches have a correlation with the success or failure to land
 - Practically all algorithms for prediction give the same result

Introduction

- Background: SpaceX, a rocket company, launches satellites at lower prices than their competitor since they have developed a technology to land the first stage booster, which is 70% the cost of the rocket. By landing it safely, they can reuse the booster.
- Problem: can we use the previous data of launches of Falcon 9 rocket to predict the probability of the booster landing back to the pad safely?



Section 1

Methodology

Methodology

Executive Summary

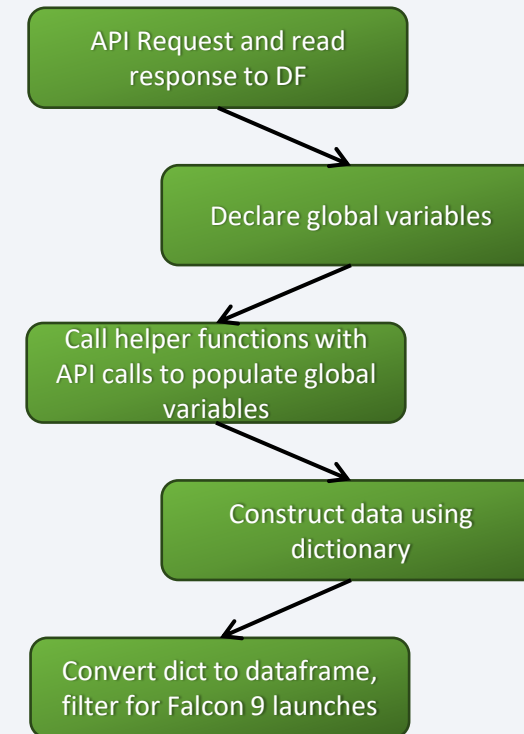
- Data collection methodology:
 - Space X API and Web Scrapping
- Perform data wrangling
 - Pandas, Numpy and SQL
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
 - Using Logistic Regression, SVM, Decision Tree and KNN

Data Collection

- Data collection is the process of gathering data from available sources. This data can be structured, unstructured or semi-structured. For this project, data was collected via SpaceX API and Web Scrapping from wiki pages for relevant launch data.
- SpaceX API: using the REST API, we extract the data in form of JSON and transform it to a dataframe using inbuilt python pandas methods.
- Web Scrapping: web scrapping Space X launches from Wikipedia and converting into a pandas dataframe.

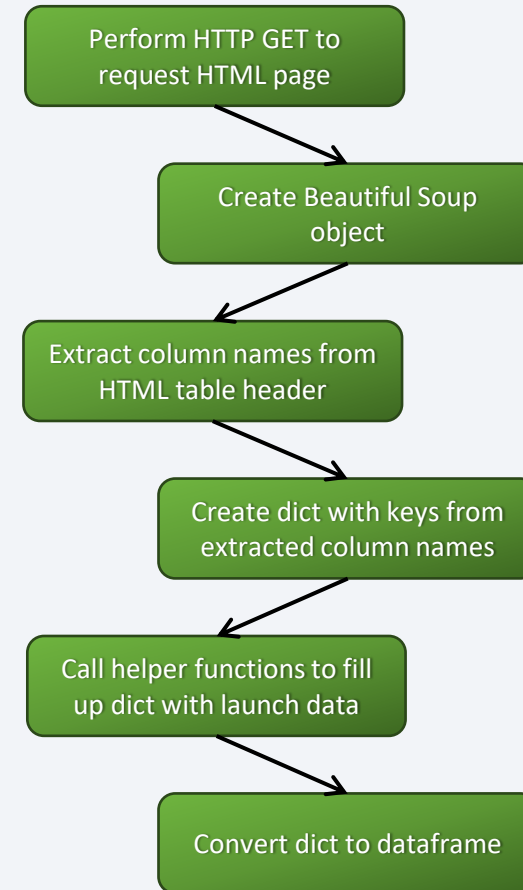
Data Collection – SpaceX API

- The API used is <https://api.spacexdata.com/v4/rockets/>
- The API provides data about many types of rocket launches done by Space X, the data is therefore filtered to include only Falcon 9 launches
- Every missing value in the data is replaced by the mean of the respective column
- We end up with 90 instances and 17 features.
- GitHub page [here](#)



Data Collection - Scraping

- The data is scrapped from [List of Falcon 9 and Falcon Heavy launches – Wikipedia](#)
- The website contains only the data about Falcon 9 launches
- We end up with 121 instances and 11 features
- GitHub page [here](#)



Data Wrangling

- The data is later processed so that there are no missing entries
- An extra column called 'Class' is also added to the data frame. This column contains 0 if a given launch failed and 1 if it is successful.
- GitHub page [here](#)

EDA with Data Visualization

- The following charts were plotted to gain further insights:
 - Scatter Plot: it shows the correlation between two variables. The following charts were visualized:
 - Flight Number and Launch Site
 - Payload Mass and Launch Site
 - Flight Number and Orbit Type
 - Payload Mass and Orbit Type
 - Bar Chart: it compares values for different categories, making it easy to see which groups are highest/common. Length of each bar is proportional to the value of the items that it represents. The following charts were visualized:
 - Success Rate and Orbit Type
 - Line Chart: it depicts trends over time. The following chart was visualized:
 - Average launch success yearly trend
- GitHub page [here](#)

EDA with SQL

- The data is queried using SQL to answer several questions about data:
 1. Display the names of the unique launch sites in the space mission.
 2. Display 5 records where launch sites begin with the string 'CCA'.
 3. Display the total payload mass carried by boosters launched by NASA (CRS).
 4. Display average payload mass carried by booster version F9 v1.1.
 5. List the date when the first succesful landing outcome in ground pad was achieved.
 6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
 7. List the total number of successful and failure mission outcomes.
 8. List the names of the booster_versions which have carried the maximum payload mass. Use a subquery.
 9. 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.
 10. 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.
- GitHub page [here](#)

Build an Interactive Map with Folium

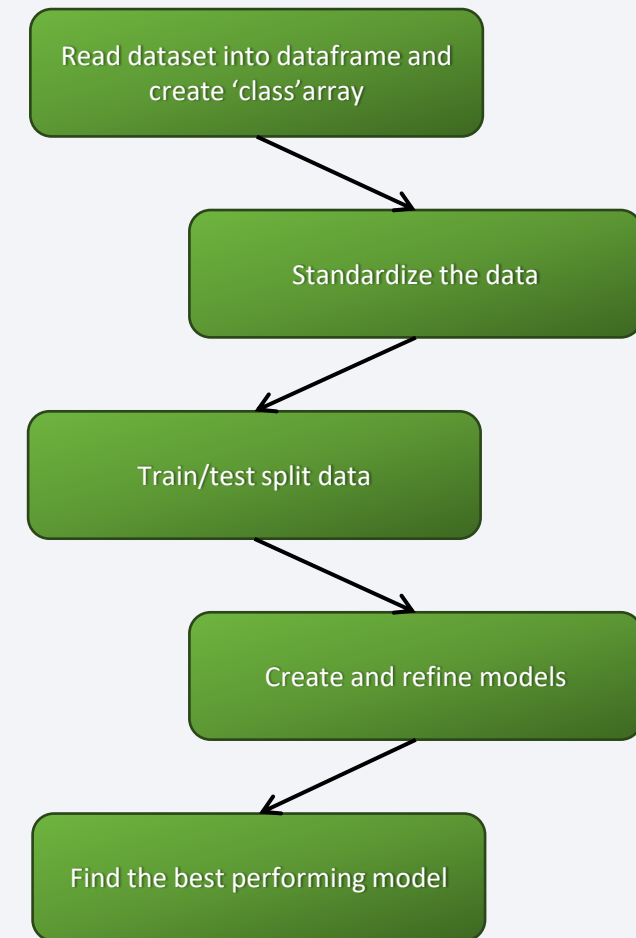
- Folium map helps to analyze geospatial data to perform more interactive visual analytics and better understand factors such location and proximity of launch sites that impact landing success rate
- Building the Interactive Map with Folium helped answering the following questions:
 - Are launch sites in close proximity to railways? YES
 - Are launch sites in close proximity to highways? YES
 - Are launch sites in close proximity to coastline? YES
 - Do launch sites keep certain distance away from cities? YES
- GitHub page [here](#)

Build a Dashboard with Plotly Dash

- Plotly Dash web application assists in the interactive visual analytics on SpaceX launch data in real-time. Added launch site dropdown, pie chart, payload range slide and a scatter chart to the Dashboard.
- This dashboard helped answering the following questions:
 - Which site has the largest successful landing? **KSC LC-39A (10 successful landings)**
 - Which site has the highest landing successful rate? **KSC LC-39A (76.9% success)**
 - Which payload range has the highest landing success rate? **2000-5000 kg**
 - Which payload range has the lowest landing success rate? **0-2000 and 5500-7000 kg**
 - Which F9 booster version has the highest landing success rate? **FT**
- GitHub page [here](#)

Predictive Analysis (Classification)

- Functions from sklearn library are used to create our machine learning models
- The following machine learning algorithms were tested:
 - Logistic Regression
 - SVM
 - Decision Tree
 - K Nearest Neighbors (KNN)
- The best selected model are chosen based on their accuracy score on test data and confusion matrix
- GitHub page [here](#)

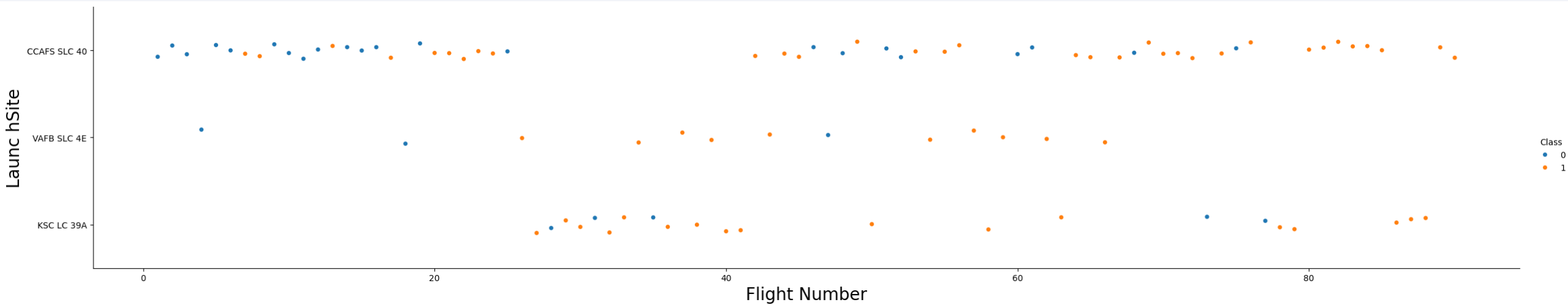




Section 2

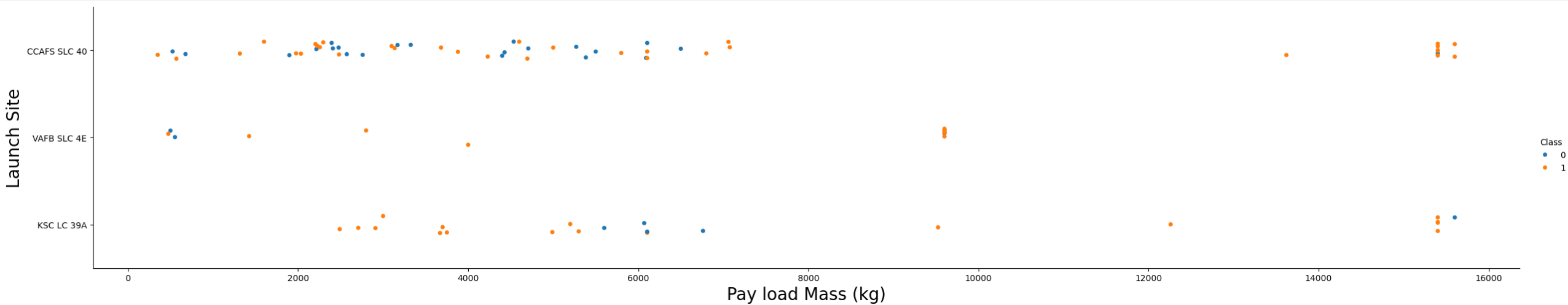
Insights drawn from EDA

Flight Number vs. Launch Site



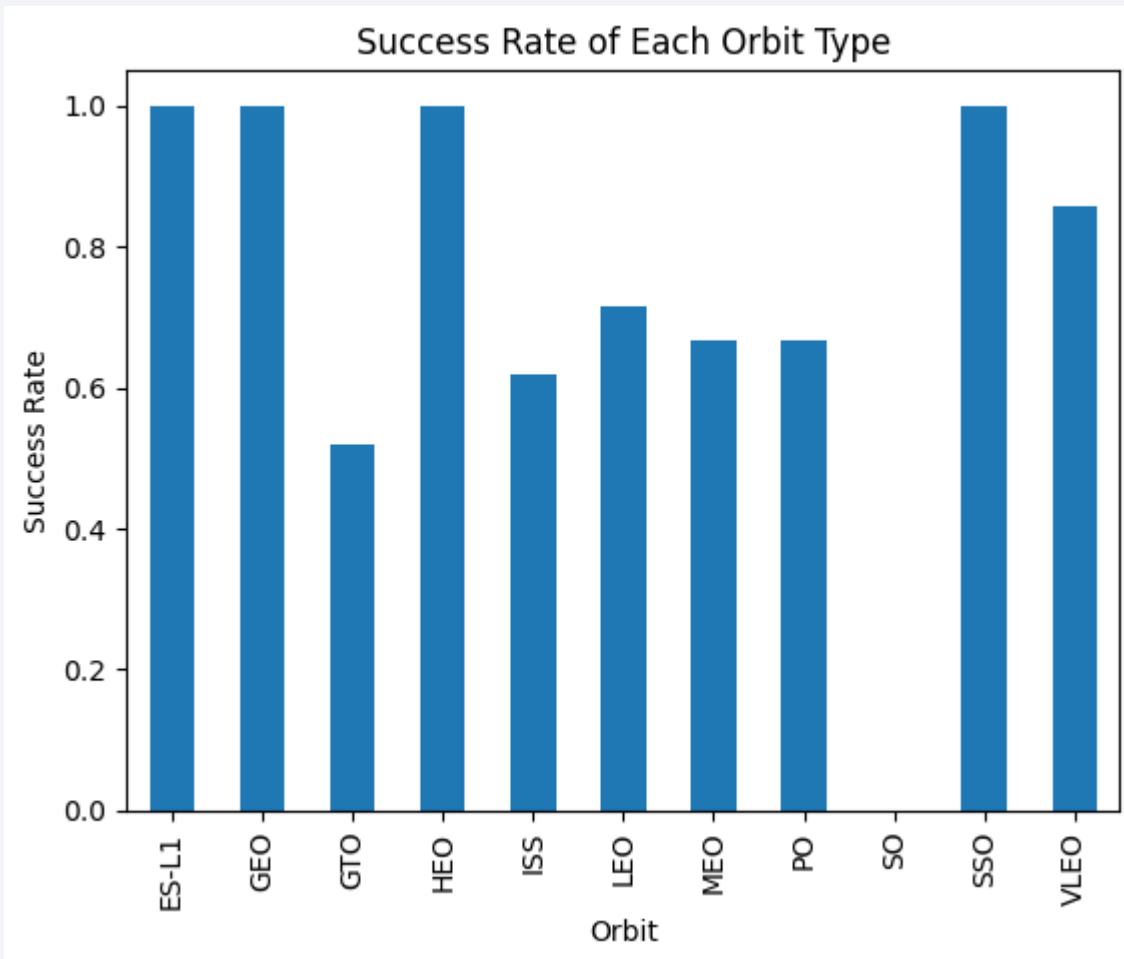
- Success rate (class 1) increases as the number of flights increases
- For launch site KSC LC 39Am it takes at least 25 launches before a first successful launch

Payload vs. Launch Site

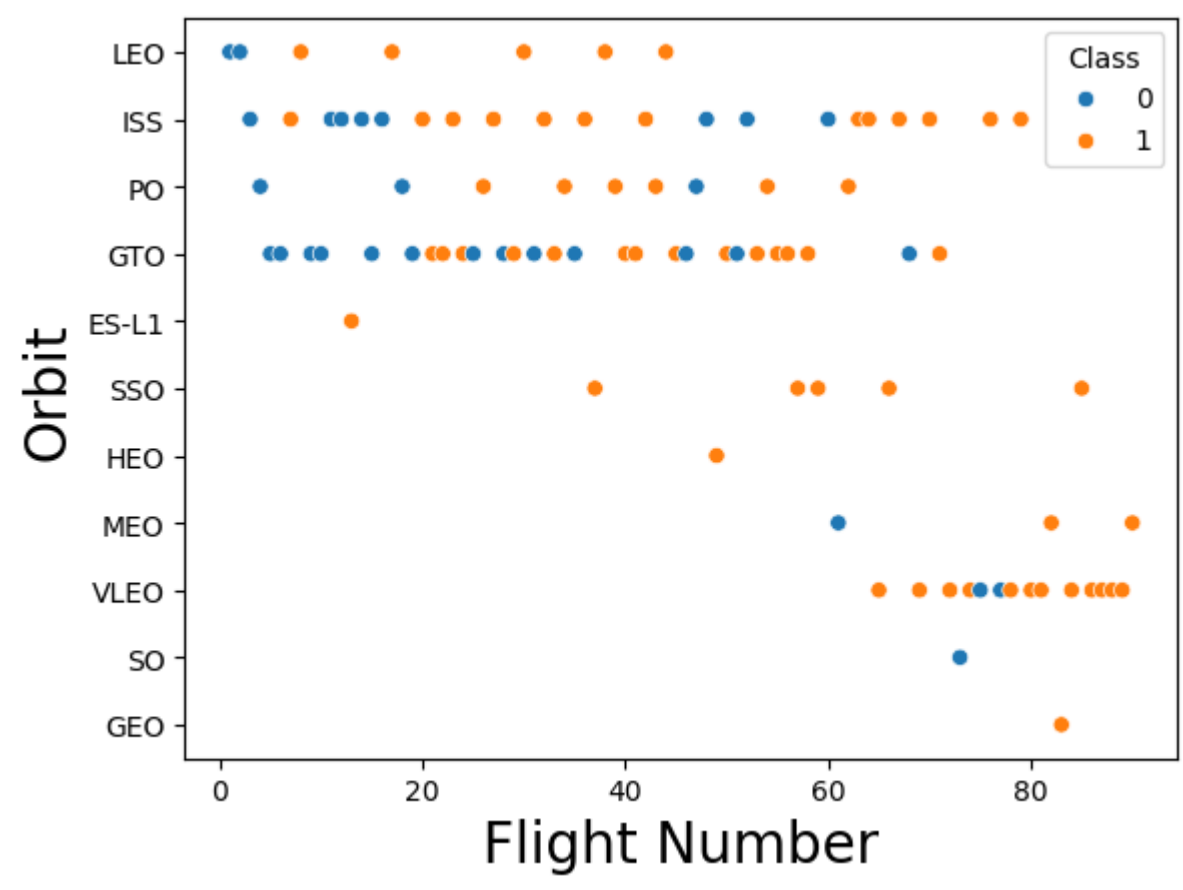


- For launch site VAFB SLC 4E:
 - There are no rockets launched with payload mass greater than 10,000 kg
 - Rate of successful landing (class 1) increases with payload mass
- There are no clear correlation between launch site and payload mass

Success Rate vs. Orbit Type

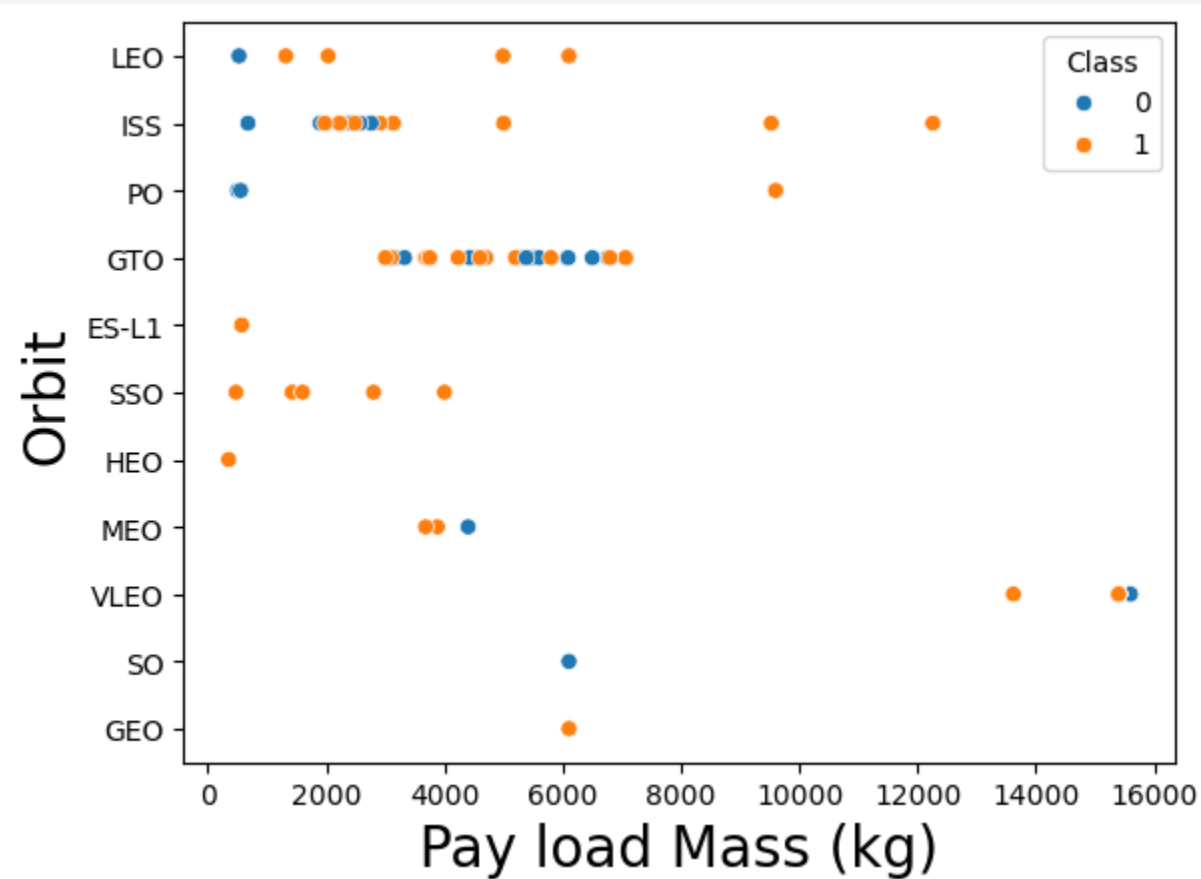


- Orbits ES-L1, GEO, HEO and SSO have the highest successful rates
- GTO orbit has the lowest success rate



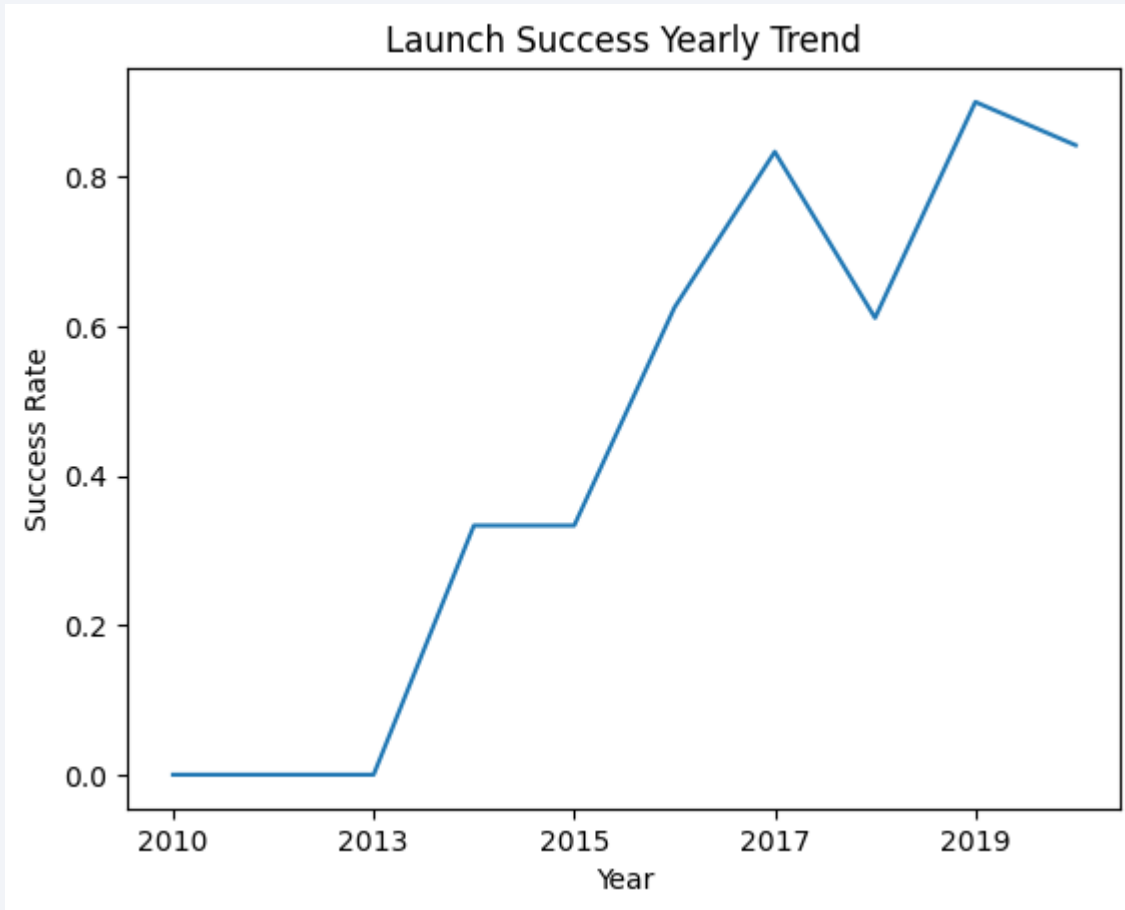
- For VLEO, first successful landing (class 1) doesn't occur until 60+ number of flights
- For most orbits (LEO, ISS, PO, SSO, MEO, VLEO) successful landing rates appear to increase with flight numbers
- There is no correlation between flight number and orbit for GTO

Payload vs. Orbit Type



- Successful landing rates (class 1) appear to increase with Pay load mass for orbits LEO, ISS, PO, SSO
- For GEO orbit, there is no correlation between payload mass and orbit for a successful landing

Launch Success Yearly Trend



- Success rates increased about 80% between 2013 and 2020
- Success rates remained the same between 2010 and 2013 and between 2014 and 2015
- Success rates decreased between 2017 and 2018 and between 2019 and 2020

All Launch Site Names

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

In [8]:

```
%%sql  
  
select distinct Launch_site from spacetable
```

* sqlite:///my_data1.db
Done.

Out[8]:

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

- 'distinct' returns only the unique values from the queries column
- There are four unique launch sites

Launch Site Names Begin with 'CCA'

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

In [9]:

```
%%sql  
  
select * from spacetable where Launch_Site like 'CCA%' limit 5
```

* sqlite:///my_data1.db
Done.

Out[9]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS_KG	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- using 'like' and format 'CCA%', it returns records where Launch_Site column starts with 'CCA'
- 'limit 5' limits the number of returned records to 5

Total Payload Mass

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

```
In [11]: %%sql
          select sum(PAYLOAD_MASS__KG_) as Total_Payload from spacetable where Customer='NASA (CRS)'
          * sqlite:///my_data1.db
          Done.
Out[11]: 

|       |
|-------|
| 45596 |
|-------|


```

- 'sum' adds column 'Payload_mass__kg_' and returns total payload mass for customers named 'NASA (CRS)'

Average Payload Mass by F9 v1.1

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

```
%%  
%%sql
```

```
select avg(PAYLOAD_MASS__KG_) as Average_Payload from spacextbl where Booster_Version = 'F9 v1.1'
```

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

```
Done.
```

Average_Payload

2928.4

- 'avg' returns the average of payload mass in 'PAYLOAD_MASS__KG_' column where booster version is 'F9 v1.1'

First Successful Ground Landing Date

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

Hint: Use min function

```
] : %%sql
select min(Date) as min_date from spacextbl where Landing_Outcome = 'Success (ground pad)'

* sqlite:///my_data1.db
Done.
] : min_date
    min_date
-----
2015-12-22
```

- 'min(Date)' selects the first or the oldest date from 'Date' column where first successful landing on group pad was achieved

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

```
In [18]: %%sql
select Booster_Version from spacetable where (PAYLOAD_MASS_KG_ < 6000 and PAYLOAD_MASS_KG_ > 4000) and (Landing_Outcome = 'Success')
* sqlite:///my_data1.db
Done.
```

Out[18]: **Booster_Version**

F9 FT B1022

F9 FT B1026

F9 FT B1021.2

F9 FT B1031.2

- the query finds the booster version where payload mass is greater than 400 but less than 6000 and the landing outcome is successful in drone ship
- the 'and' in the 'where' clause returns booster version where both conditions are true

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

In [23]:

```
%%sql  
  
select Mission_Outcome, count(Mission_Outcome) as counts from spacetable group by Mission_Outcome
```

* sqlite:///my_data1.db

Done.

Out[23]:

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

- the 'group by' arranges identical data in a column in to groups

Boosters Carried Maximum Payload

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

```
In [26]: %%sql
select Booster_Version, PAYLOAD_MASS_KG_ from spacextbl where PAYLOAD_MASS_KG_ = (select max(PAYLOAD_MASS_KG_) from space)

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

```
Out[26]:
```

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

- the subquery returns the maximum payload mass by using 'max'
- the main query returns booster version and respective payload mass where the last is maximum with value of 15600

2015 Launch Records

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: SQLite 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.

```
In [30]: %%sql
select substr(Date,6,2) as Month, Landing_Outcome, Booster_Version, Launch_Site from spacetable where (substr(Date,0,5)='2015')
* sqlite:///my_data1.db
Done.
```

```
Out[30]:
```

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

- the result identify launch sites as 'CCAFS LC-40' and booster version as 'F9 v1.1 B1012' and 'F9 v1.1 B1015' that had failed landing in drop ship in the 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.

In [31]:

```
%%sql
```

```
select Landing_Outcome, count(*) as LandingCounts from spacextbl where Date between '2010-06-04' and '2017-03-20'
group by Landing_Outcome
order by count(*) desc;
```

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

```
Done.
```

Out[31]:

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

- the 'order by' arranges the counts 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 space with stars. The Earth's surface is dark blue, with bright yellow and orange lights from cities and towns. The lights are concentrated in the lower right quadrant of the image, forming a dense network of glowing points and lines.

Section 3

Launch Sites Proximities Analysis

SpaceX Falcon9 – Launch Sites Map



Figure 1 – Global Map

Figure 1 on the left displays the global map with Falcon 9 launch sites located in the USA (California and Florida). Each launch site contains a circle, a label and a popup to highlight the location and the name of the launch site. It is also evident that all launch sites are near the coastline.

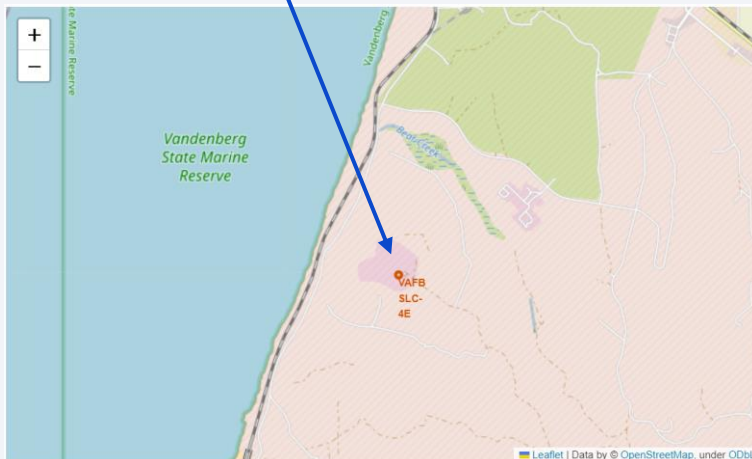


Figure 2 – Zoom 1

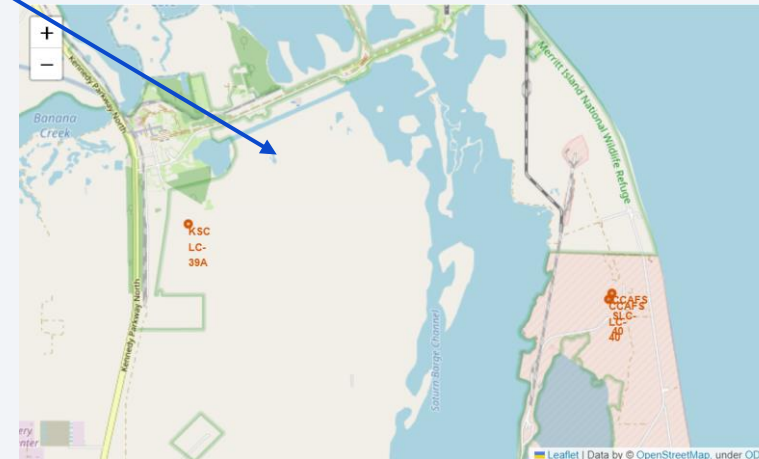


Figure 3 – Zoom 2

SpaceX Falcon 9 – Success/Fail Launch Map

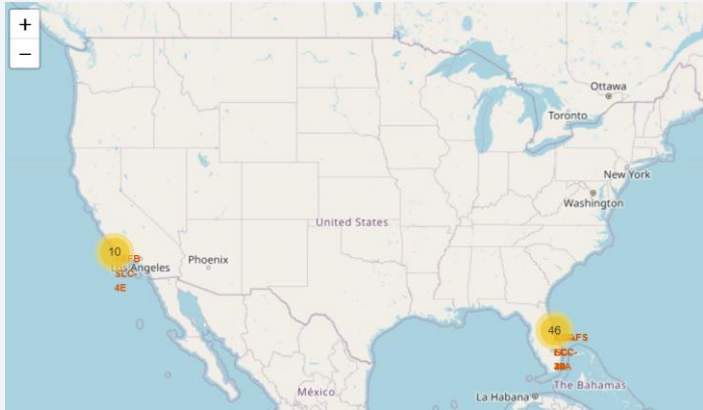


Figure 1 – US map with all launch sites

- Figure 1 is the US map with all the launch sites. The numbers on each site depicts the total number of launches.
- Figure 2, 3, 4 and 5 zoom in each site and displays the success (green) and fail (red) markers.
- By looking at each site map, KSC LC-39A has the greatest number of successful landings.



Figure 2 – VAFB Launch Site with success/fail markers

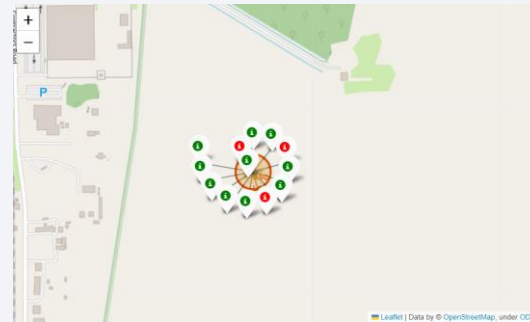


Figure 3 – KSC LC-39A Launch Site with success/fail markers

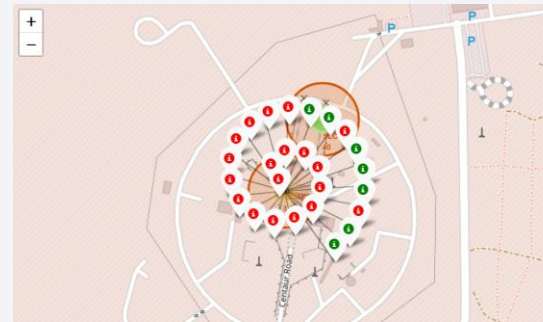


Figure 4 – CCAFS LC-40 Launch Site with success/fail markers

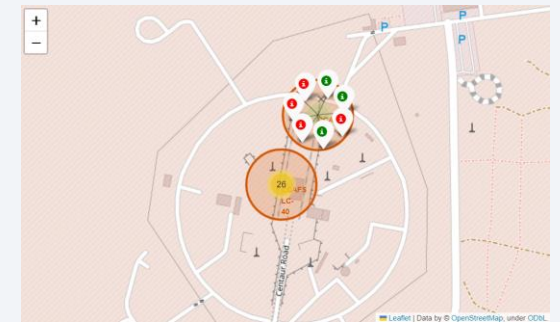


Figure 5 – CCAFS SLC-40 Launch Site with success/fail markers

SpaceX Falcon9 – Launch Site Distance Map

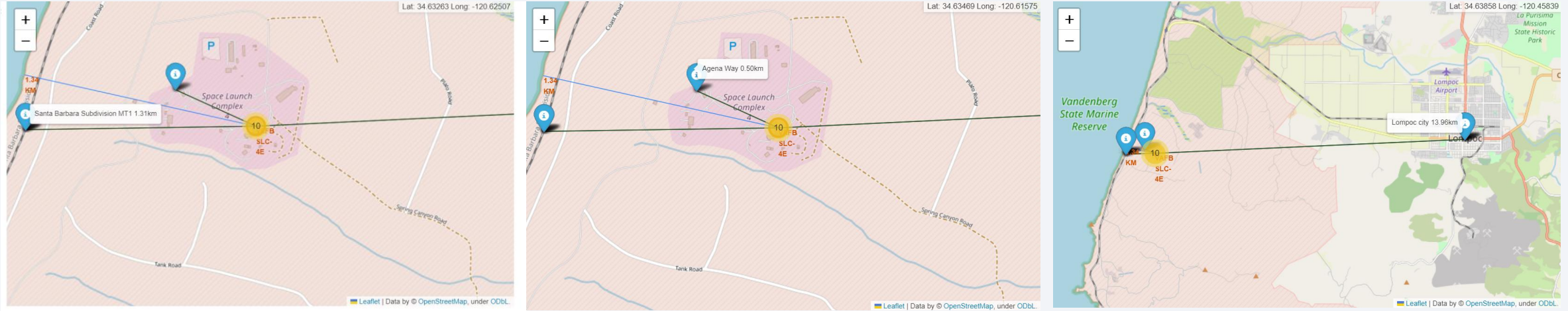


Figure 1 – Proximity site map for VAFB SLC-4E for different sites

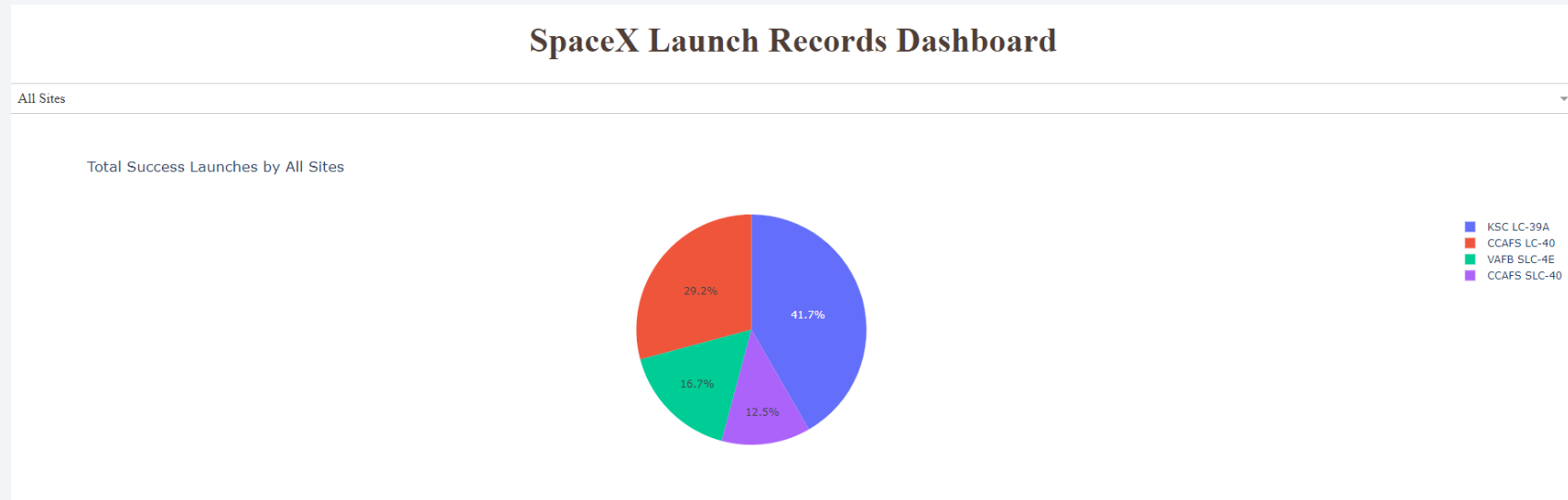
In general, cities are located further away from the launch site to minimize impacts of any accident to the public and infrastructure. Launch sites are strategically located near coastline, railroad and highways to provide easy access to resources.



Section 4

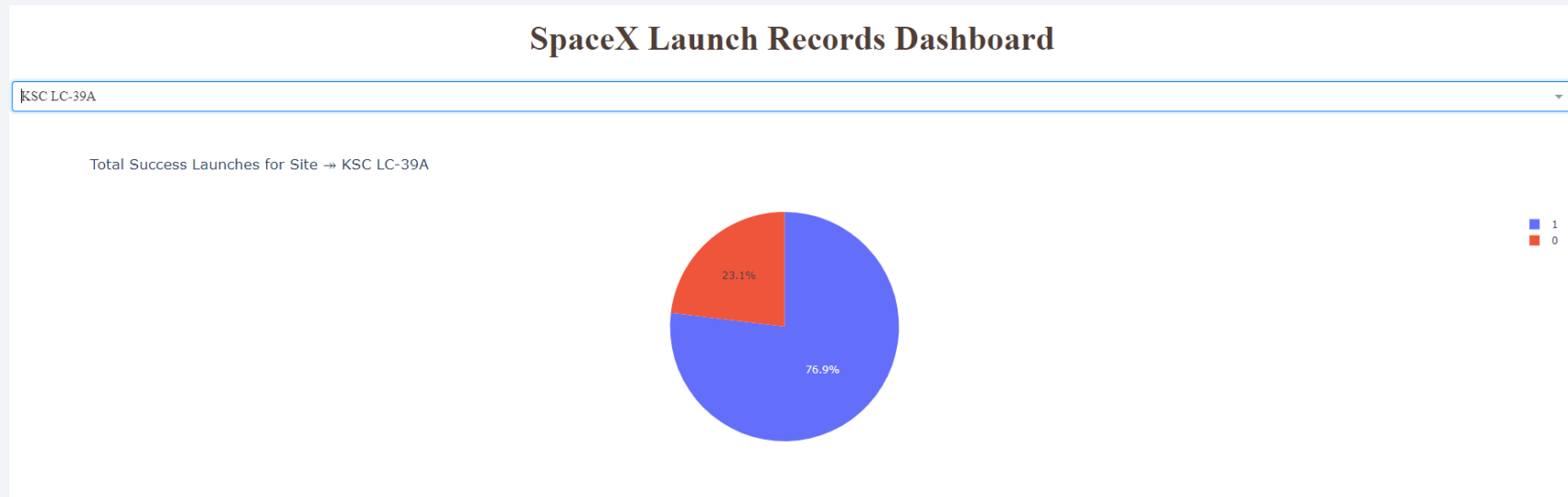
Build a Dashboard with Plotly Dash

Launch Success Counts for all sites



- Launch site KSC LC-39A has the highest launch success rate.
- Launch site CCAFS SLC-40 has the lowest launch success rate.

Launch Site with Highest Launch Success Ratio



- Launch success rate is 76.9%
- Launch failure rate is 23.1%

Payload vs. Launch Outcome Scatter Plot for all sites

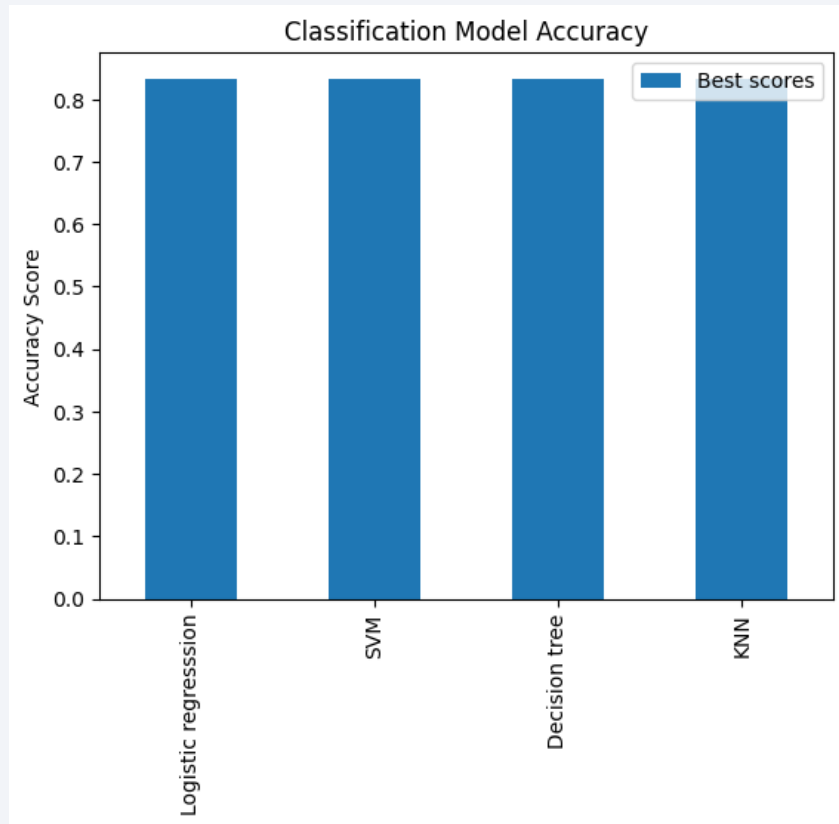


- Most successful launches are in the payload range from 2000 to about 5500 kg
- Booster version category “FT” has the most successful launches
- Only booster with a successful launch when payload is greater than 6000 kg is B4

Section 5

Predictive Analysis (Classification)

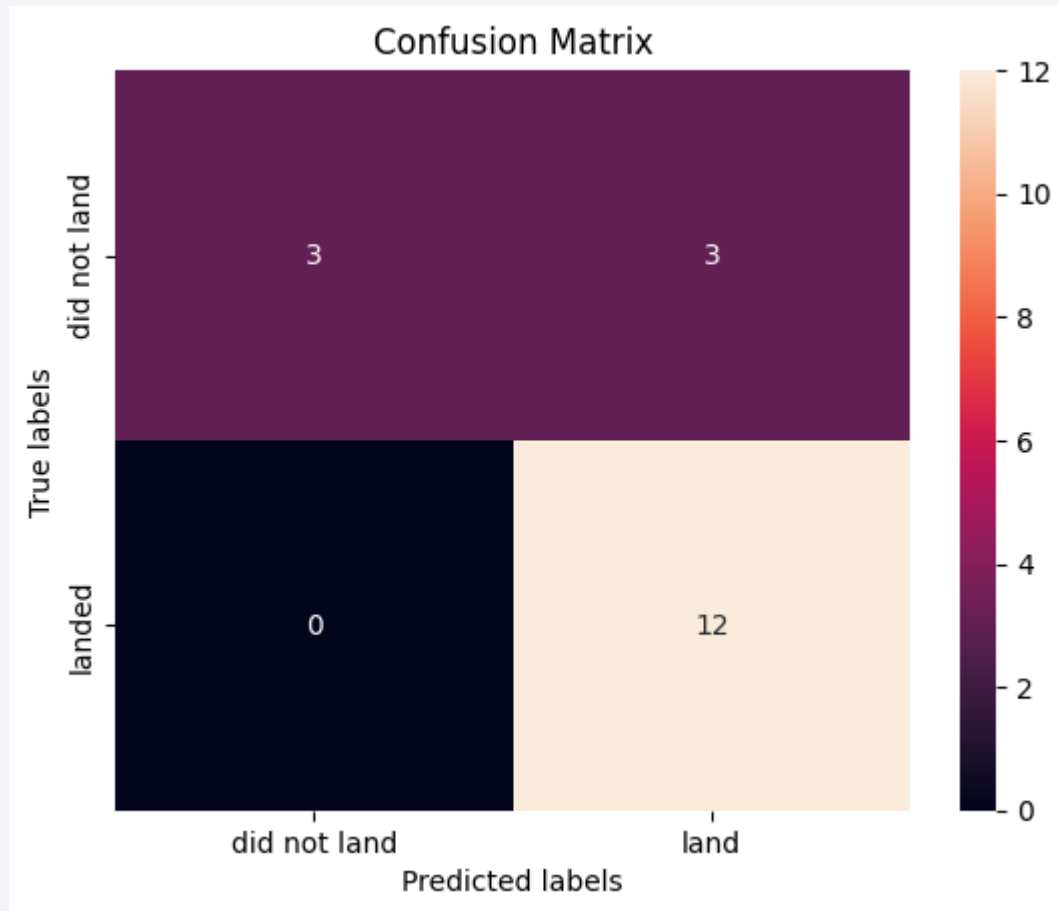
Classification Accuracy



Best scores	
Logistic regresssion	0.833333
SVM	0.833333
Decision tree	0.833333
KNN	0.833333

- Accuracy score on the test data is the same for all the classification algorithms based on the test set with a value of 0.8333
- Given the accuracy scores for classification algorithms are the same, we may need a broader dataset to further tune the models.

Confusion Matrix



- The confusion matrix is same for all the models (LR, SVM, Decision Tree, KNN)
- Per the confusion matrix, the classifier made 18 predictions
- 12 scenarios were predicted Yes for landing, and they did land successfully (True positive)
- 3 scenarios (top left) were predicted No for landing, and they did not land (True negative)
- 3 scenarios (top right) were predicted Yes for landing, but they did not land successfully (False positive)
- Overall, the classifier is correct about 83% of the time $((TP + TN) / Total)$ with a misclassification or error [43](#) rate $((FP + FN) / Total)$ of about 16.5%

Conclusions

- As the numbers of flights increase, the first stage is more likely to land successfully
- Success rates appear to go up as payload increases but there is no clear correlation between payload mass and success rates
- Launch success rate increased by about 80% from 2013 to 2020
- Launch Site KSC LC-39A has the highest launch success rate and Launch Site CCAFS SLC40 has the lowest launch success rate
- Orbits ES-L1, GEO, HEO, and SSO have the highest launch success rates and orbit GTO the lowest
- Launch sites are located strategically away from the cities and closer to coastline, railroads, and highways
- All the models scored the same on the test data, the accuracy score was about 83% for all models. More data may be needed to further tune the models and find a potential better fit.

Thank you!

