Winning Space Race with Data Science

Abhishek Kumar
May 30, 2024

Outline

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

Executive Summary

- Summary of methodologies

The following report deals with SpaceX's Falcon 9 rocket launches. It aims to understand the factors contributing to a successful SpaceX launch. The dataset is first collected through a CSV file. Web scraping methods such as Beautiful Soup were also used to obtain the data. After obtaining the right data and cleaning it using Pandas, Exploratory Data Analysis was performed to gather further insights into the data. Data Visualization tools such as Folium, Matplotlib, Seaborn, and Plotly Dash were used to display data in a human-readable format. In conclusion for the research, Machine Learning prediction models offered by Scikit-Learn were used to build models that best-predicted success rate for Launches

- Summary of all results

Introduction

Project background and context:

- The Falcon9 is one of SpaceX's most popular rockets.
- SpaceX advertises Falcon9's cost to be around 62 Million USD, whereas other providers quote 165 Million dollars for their market equivalents.
- Much of these savings are because Falcon9 is able to reuse the first stage of its launch.

Problems you want to find answers

- If we can determine if the first stage will land, we can determine the cost of the launch.
- Therefore, this study aims to study the causing factors behind a successful Falcon9 launch for SpaceX.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using 3 different methods: 1)SpaceX API, 2)Web scraping
- Perform data wrangling
  - Data was processed using 2 different methods: 1)The Pandas Library, 2) 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
  - Scikit-Learn Library: LogisticRegression, Support Vector Machines, Decision Tree Classifiers etc.

Data Collection

- Data was collected using 2 different methods/sources:

1) SpaceX API
2) Web Scraping

Data Collection – SpaceX API

- The requests library was used to obtain data from the SpaceX API, followed by json normalization using the Pandas library.
- Furthermore, new columns were created using user-defined functions, as can be seen in the notebook below.
- https://github.com/Abhi-61/SpaceX_Rocket_Launch_Data_Analysis/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

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

```
response = requests.get(spacex_url)
```

```
# Use json_normalize meethod to convert the json result into a dataframe
data = pd.json_normalize(response.json())
```

```
# Call getBoosterVersion
getBoosterVersion(data)
```

```
# Call getLaunchSite
getLaunchSite(data)
```

```
# Call getPayloadData
getPayloadData(data)
```

```
# Call getCoreData
getCoreData(data)
```

Data Collection - Scraping

- The response library was used again to access the Wikipedia page of SpaceX, which was then converted into a BeautifulSoup object, and then finally transformed into a pandas Dataframe.
- https://github.com/Abhi-61/SpaceX_Rocket_Launch_Data_Analysis/blob/main/jupyter-labs-webscraping.ipynb

```
static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"
```

```
# use requests.get() method with the provided static_url
# assign the response to a object
response = requests.get(static_url)
```

```
# Use BeautifulSoup() to create a BeautifulSoup object from a response text content
soup = BeautifulSoup(response.content, 'html.parser')
```

```
# Use the find_all function in the BeautifulSoup object, with element type `table`
# Assign the result to a list called `html_tables`
html_tables = soup.find_all('table')
```

```
df= pd.DataFrame({ key:pd.Series(value) for key, value in launch_dict.items() })
```

Data Wrangling

- Once the data was converted into a Pandas dataframe, it had to be further analyzed for cleaning.
- You need to present your data wrangling process using key phrases and flowcharts
- First and foremost, the datatypes of each attribute in the dataframe had to be determined in order to understand what methods can be applied to the dataset.
- Moreover, null_counts and the value_counts of the important attributes had to be determined before moving forward.

```
df.isnull().sum()
```

```
df.dtypes
```

```
# Apply value_counts() on column LaunchSite
df['LaunchSite'].value_counts()
```

- https://github.com/Abhi-61/SpaceX_Rocket_Launch_Data_Analysis/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- What data had to be visualized and why?

1) Cat plot of PayloadMass vs FlightNumber, to see if there was any change in Payload Mass as FlightNumber increased. (Note that Flight Number was sequential ranging from 1 to 90).

2) Scatterplot of FlightNumber vs LaunchSites to determine in which sites were a majority of the rockets were launched.

3) Scatterplot of LaunchSite and PayloadMass to determine if certain LaunchSites were used to launch rockets only of a specific Payload Mass.

4) Bar Chart of success rate with regards to the orbit in which the rocket was launched.

5) Scatterplot representing the number of Flights in each orbit, and colored based on whether the flight was a success or failure.

EDA with Data Visualization (Contd.)

- What data had to be visualized and why?

6) Scatterplot of Orbit vs Payload Mass of each Flight to gather insight on whether certain orbit rockets had a specific payload mass.

7) Line plot of the average success rate of launches in each year to study the general trend of SpaceX's Falcon9 Launches.

- https://github.com/Abhi-61/SpaceX_Rocket_Launch_Data_Analysis/blob/main/edadataviz.ipynb

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
- Select the distinct Launch sites in the dataset to determine the unique launch sites used for Falcon9 Launches
- Select the first 5 records whose Launch Sites began with characters “CCA”
- Select the total payload mass carried by boosters launched by NASA.
- Display the average payload mass carried by booster version F9 v1.1
- List the Date when the first successful ground pad rocket was launched
- List boosters of payload mass between 4000 and 6000kg which have success in drone ship
- Display boosters that have carried the maximum payload mass

EDA with SQL (Contd.)

- List Records of launches that failed by drone ship in 2015.
- List landing outcomes between 2010-06-04 and 2017-03-20, ranked in descending order
- https://github.com/Abhi-61/SpaceX_Rocket_Launch_Data_Analysis/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Objects Added to Folium Map:

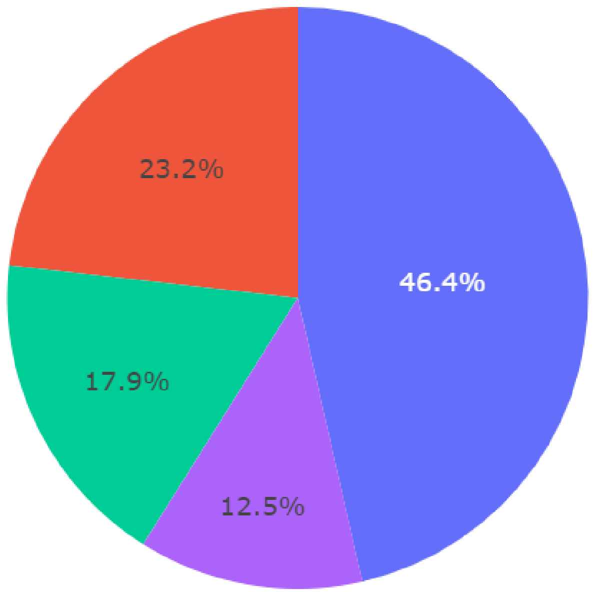
1) Marker and Circle representing the NASA Johnson Space Center at Houston, Texas. This was used as the center of the Map.
2) A circle and marker for each distinct launch site in the dataset, to identify the sites better.
3) Marker Cluster for each launch site where each marker represented either a successful launch(in green) or failed launch(in red). This made it easier to identify launch sites with higher successful launches.
4) A marker and polyline connecting the launchsite to the marker representing the nearest coastline, highway, city etc. to the launch site VAFB SLC-4E.

- https://github.com/Abhi-61/SpaceX_Rocket_Launch_Data_Analysis/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Interactions and Plots that were added to the Dash Application:
  - Success Rate of All Sites compared with each other. (Dropdown Menu)
  - Success Rate of each Launch site Taken one at a time. (Dropdown Menu)
  - Scatterplot of Success Rate of the selected launch site with regards to flights of different payload mass. A slider was used to filter the minimum and maximum payload range

https://github.com/Abhi-61/SpaceX_Rocket_Launch_Data_Analysis/blob/main/plotly-dash-application/spacex_dash_app.py



Predictive Analysis (Classification)

- Steps used in determining the best predictive model:

1) Standardize data using StandardScaler(), and then fit and transform the data.
2) Split the data into train and test data using train_test_split.
3) Use Grid Search on 4 different predictors, namely Logistic Regression, Support Vector Machines, Decision Tree Classifiers, K Nearest Neighbors.
4) Determine which predictor is best based on F1_Score

- https://github.com/Abhi-61/SpaceX_Rocket_Launch_Data_Analysis/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

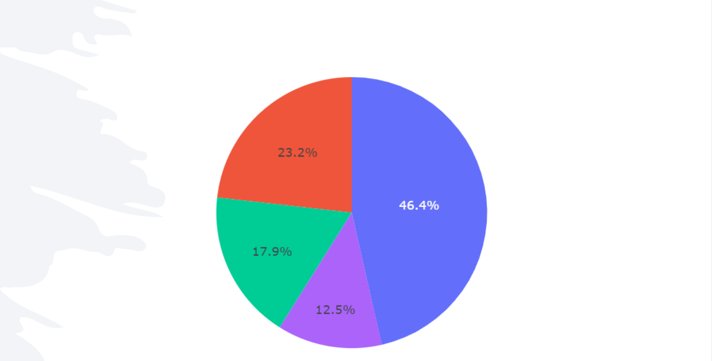
- Exploratory data analysis results

1) Most Flights were launched at CCAFS SLC-40
2) Most Flights launched had a payload mass below 8000kg, especially for CCAFS SLC-40
3) Flights launched into orbits LEO, ISS, GTO and SO have significantly higher success rates.
4) Success rates kept increasing since 2013 until 2020.

- Interactive analytics demo in screenshots

- Predictive analysis results

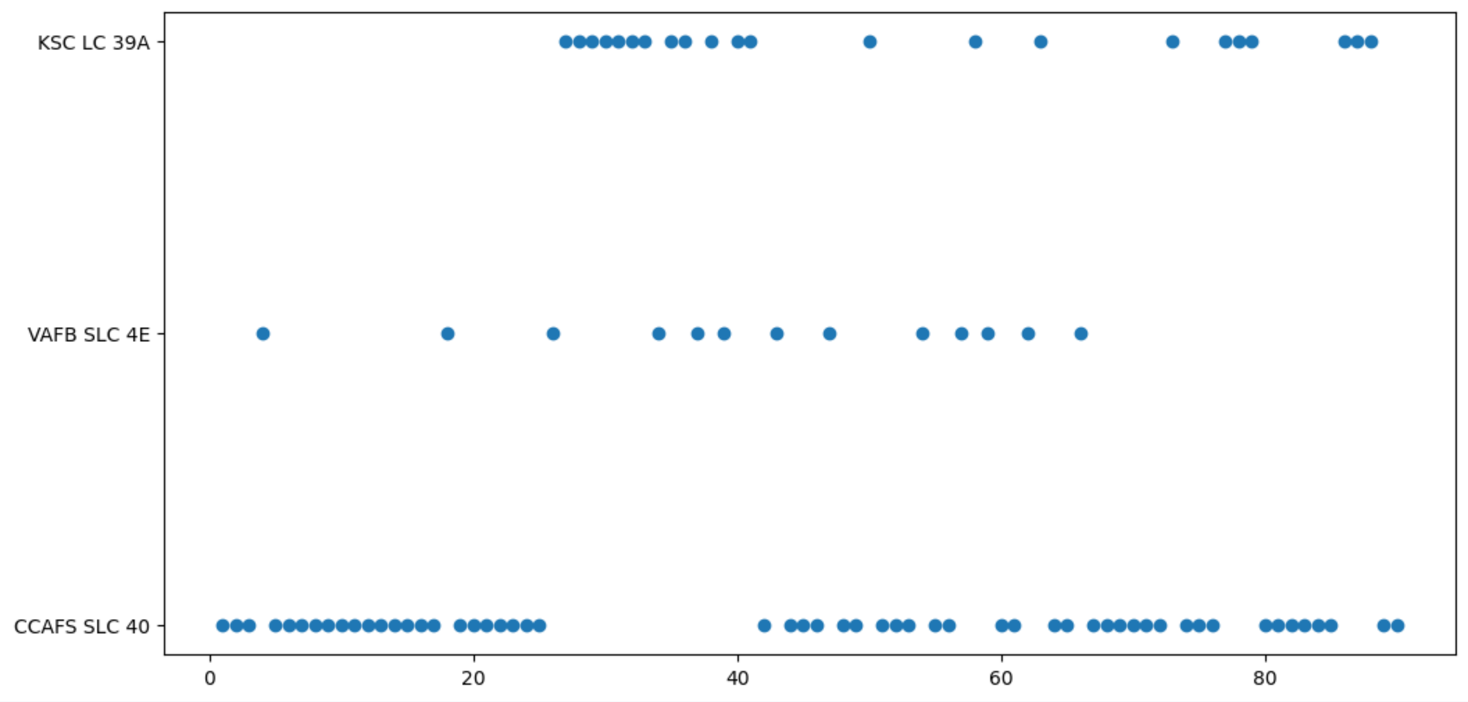
```
LogisticRegression
f1_score: 0.8148148148148149

SVM
f1_score: 0.8148148148148149

DecisionTree
f1_score: 0.5282051282051283

KNN
f1_score: 0.8148148148148149

#Therefore, KNN, SVM and Logistic Regression all tend to be really good prediction models for the given dataset
```



Section 2

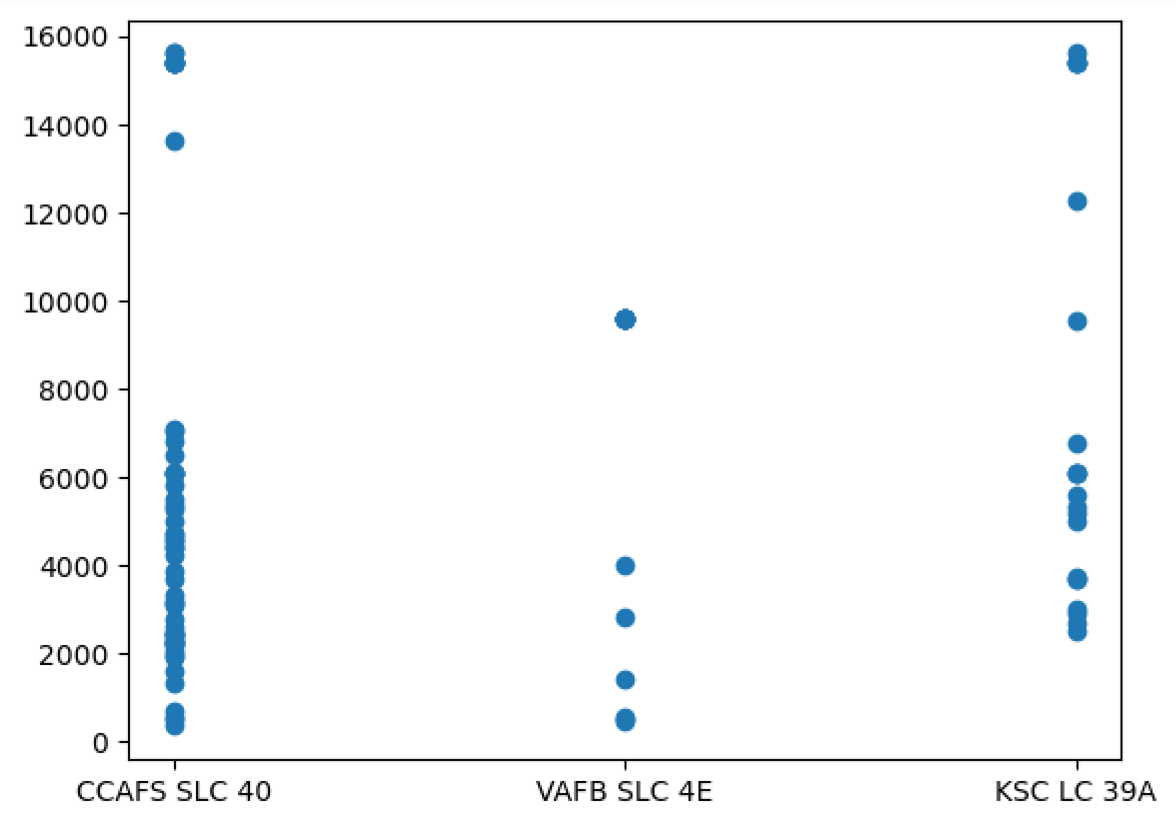
Insights drawn from EDA

Flight Number vs. Launch Site



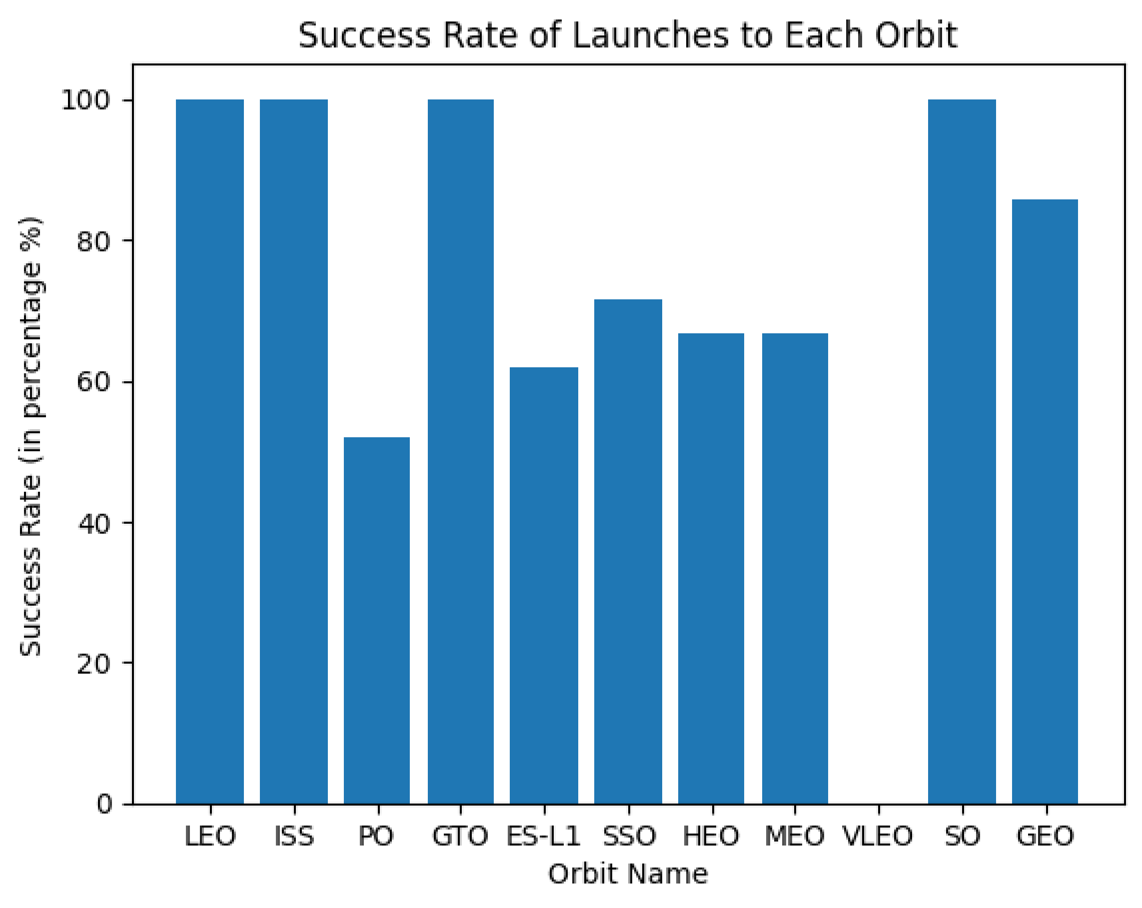
A majority of Falcon9 Launches were from the CCAFS SLC-40 Launch Site.

Payload vs. Launch Site



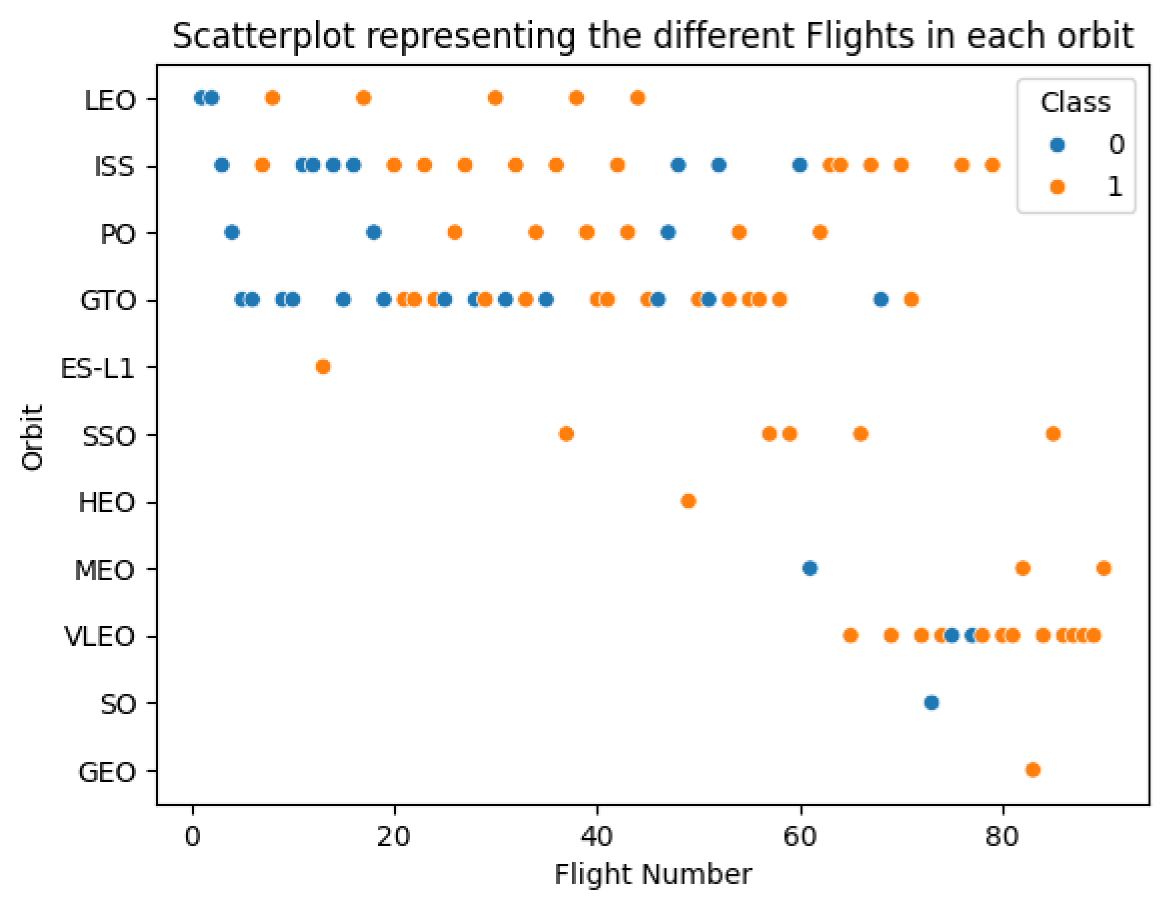
- Most of the flights launched from all launch sites seem to have a payload mass of below 8000kg

Success Rate vs. Orbit Type



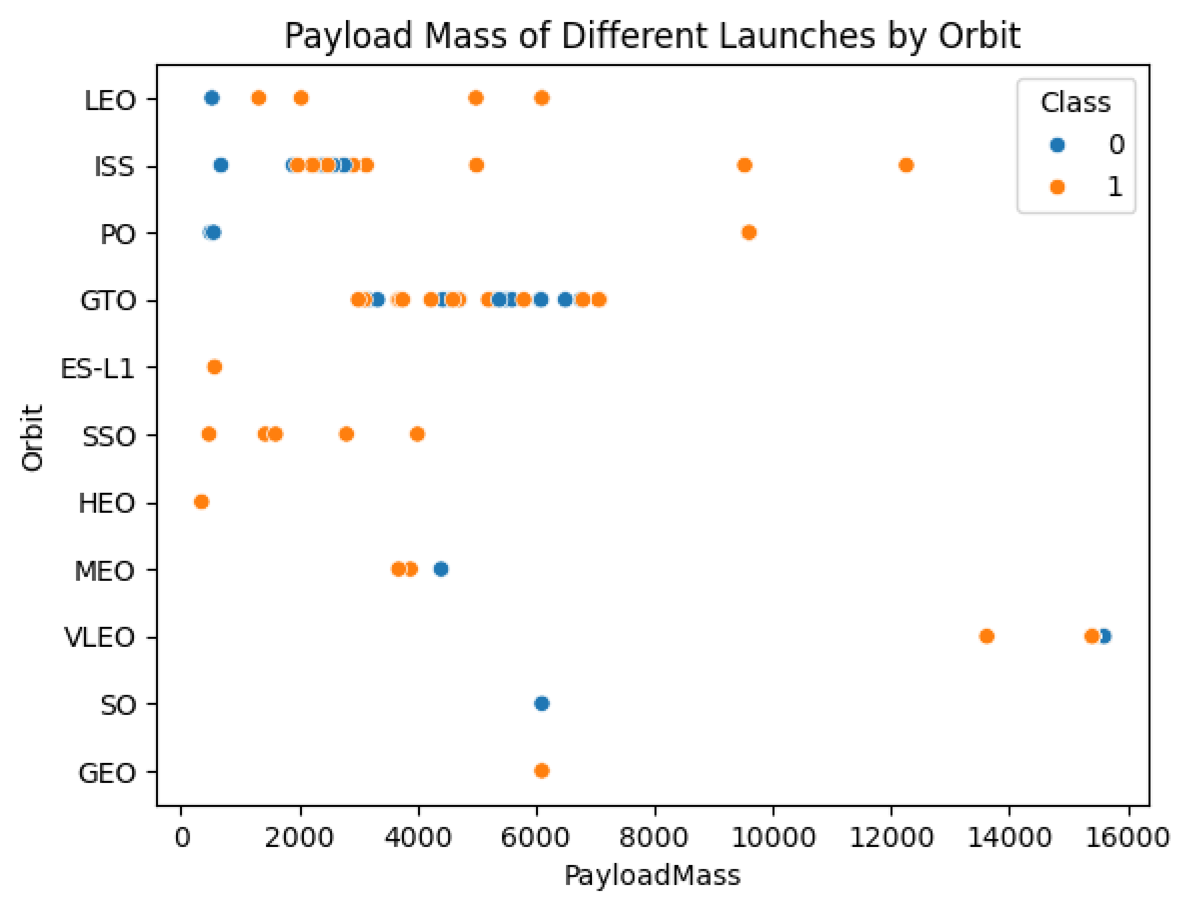
Flights launched into orbits LEO, ISS, GTO and SO have 100% success rates.

Flight Number vs. Orbit Type



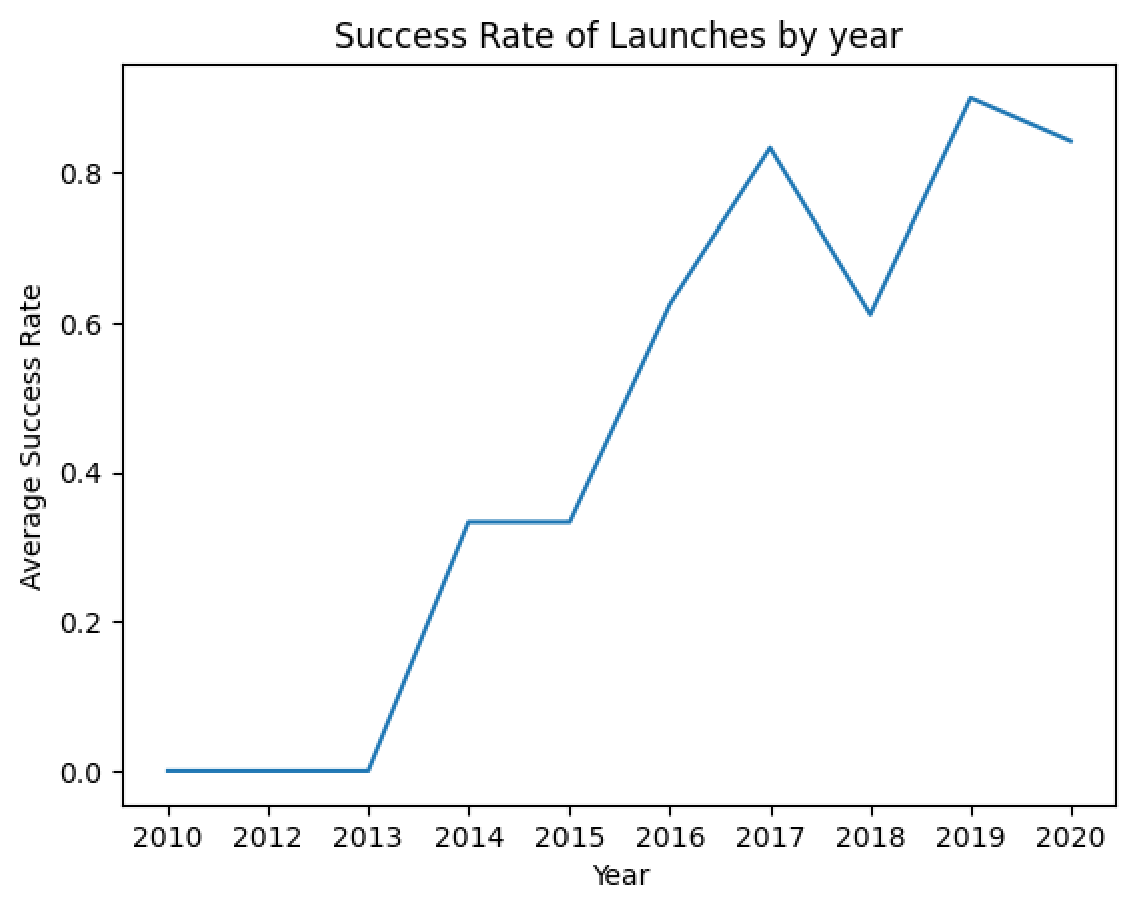
- Orbits SSO, HEO, MEO, VLEO, SO, GEO had Falcon9 flights launched into them only in the later launches of SpaceX.

Payload vs. Orbit Type



Most flights were launched into orbits GTO and ISS, with payload mass below 8000kg.

Launch Success Yearly Trend



- Success rates for Falcon9 launches showed an upward trend since 2013 until 2020 with a slight dip in 2018.

All Launch Site Names

The Unique launch sites of Falcon 9 space launches can be extracted from the dataset as shown in the query below.

```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE;
```

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

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

Therefore, the 4 unique launch sites in the dataset are CCAFS LC-40, CCAFS SLC-40, VAFB SLC-4E, KSC LC-39A.

Launch Site Names Begin with 'CCA'

- The query to display the first records of launch sites beginning with CCA is as shown below:

```
%sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE "CCA%" LIMIT 5;
```

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

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

It can be seen that all the missions in the retrieved result are successes.

Total Payload Mass

The total payload mass of boosters launched by NASA is retrieved using the query:

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE "Customer" LIKE "NASA%"
```

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

SUM(PAYLOAD_MASS__KG_)
99980

As seen above, NASA has sent boosters with a combined payload mass of 99980 Kg

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1is given by the query:

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version LIKE "F9 v1.1%"
```

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

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

The average payload mass as shown above carried by booster version F9 v1.1 is 2534.67 Kg

First Successful Ground Landing Date

The first successful ground pad launch is given by the query:

```
%sql SELECT * FROM SPACEXTABLE WHERE Date = (SELECT MIN(DATE) FROM SPACEXTABLE WHERE Landing_Outcome LIKE "%Success%")
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2015-12-22	1:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

As shown in the result above, the launch was conducted on 22nd December 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 is given by the query:

SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = "Success (drone ship)" AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000

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

- As shown in the result set above, the successful drone ship missions with payload mass between 4000 and 6000 contain the above-given boosters.

Total Number of Successful and Failure Mission Outcomes

The total number of successful and unsuccessful missions is given by the query:

```
%sql SELECT COUNT(Mission_Outcome),Mission_Outcome FROM SPACEXTABLE GROUP BY Mission_Outcome
```

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

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

A total of 100 successful missions and 1 failure.

Boosters Carried Maximum Payload

The names of the booster that have carried the maximum payload mass is given by the query:

```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)

 * 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

The boosters that have carried the maximum payload is given above.

2015 Launch Records

The launch records of all failed landing outcomes in 2015 is given by the query:

SELECT substr(Date, 6,2) as Month, "Landing_Outcome" FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE "%Failure (drone ship)%" AND substr(Date, 0,5) = '2015’

Month	Landing_Outcome
01	Failure (drone ship)
04	Failure (drone ship)

2 Drone ship failures occurred in 2015, in January and April as shown above.

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

The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 is given by the query:

```
%sql SELECT "Landing_Outcome","Date" FROM SPACEXTABLE WHERE "Date" BETWEEN "2010-06-04" AND "2017-03-20" ORDER BY "Date" DESC
```

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

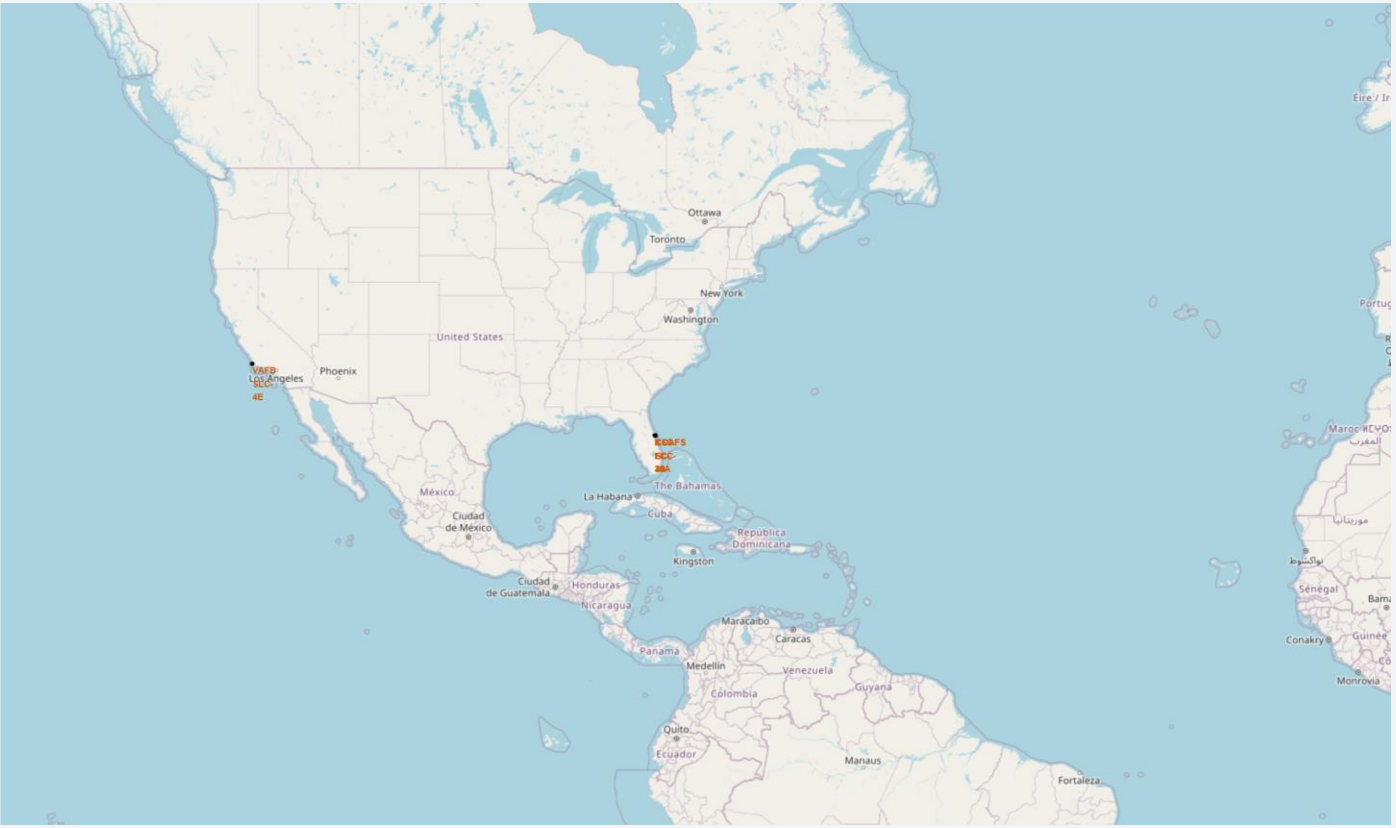
Landing_Outcome	Date
No attempt	2017-03-16
Success (ground pad)	2017-02-19
Success (drone ship)	2017-01-14
Success (drone ship)	2016-08-14
Success (ground pad)	2016-07-18
Failure (drone ship)	2016-06-15
Success (drone ship)	2016-05-27
Success (drone ship)	2016-05-06
Success (drone ship)	2016-04-08
Failure (drone ship)	2016-03-04
Failure (drone ship)	2016-01-17
Success (ground pad)	2015-12-22
Precluded (drone ship)	2015-06-28
No attempt	2015-04-27
Failure (drone ship)	2015-04-14
No attempt	2015-03-02
Controlled (ocean)	2015-02-11
Failure (drone ship)	2015-01-10
Uncontrolled (ocean)	2014-09-21
No attempt	2014-09-07
No attempt	2014-08-05
Controlled (ocean)	2014-07-14
Controlled (ocean)	2014-04-18
No attempt	2014-01-06
No attempt	2013-12-03
Uncontrolled (ocean)	2013-09-29
No attempt	2013-03-01
No attempt	2012-10-08
No attempt	2012-05-22
Failure (parachute)	2010-12-08
Failure (parachute)	2010-06-04

Section 3

Launch Sites Proximities Analysis

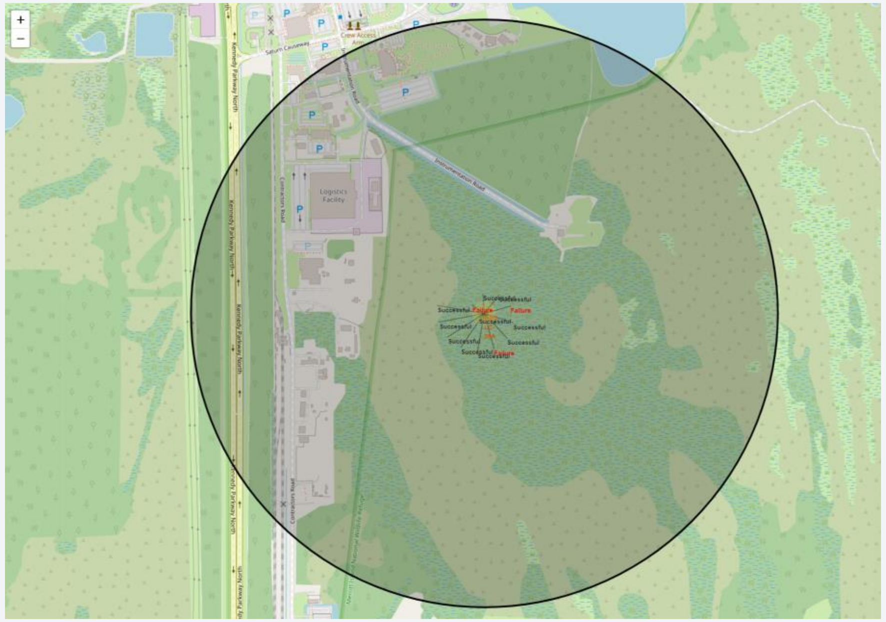
Launch Site Locations across the US

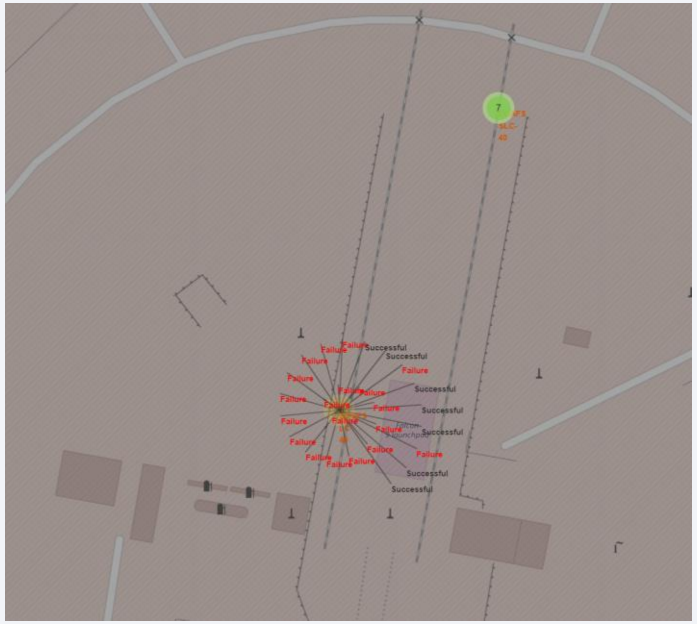
There is one launch site on the eastern coast and 2 on wester coast. The 2 sites on the western coast are extremely close to each other which is why it looked overlapped.



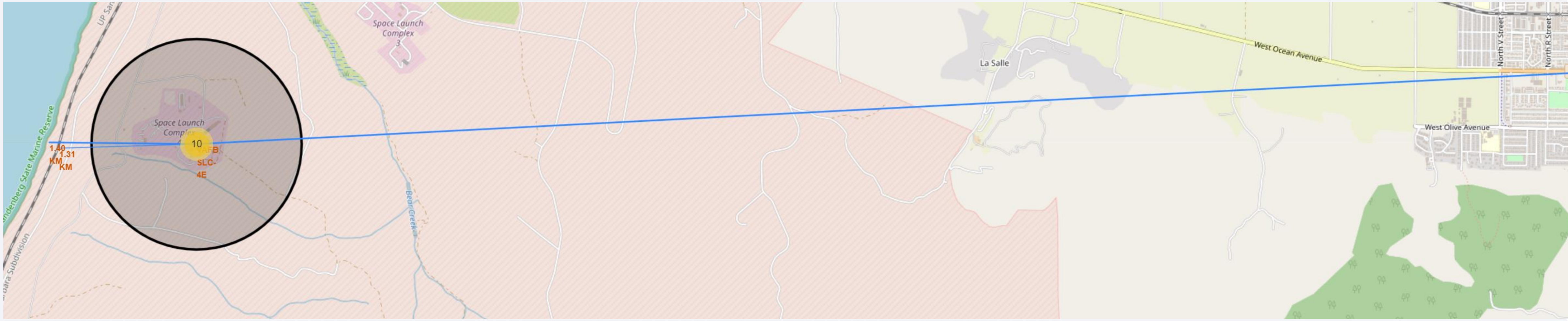
Missions completed at each Launch Site

- Explore the folium map and make a proper screenshot to show the color-labeled launch outcomes on the map





Proximities to the VAFB SLC-4E Site

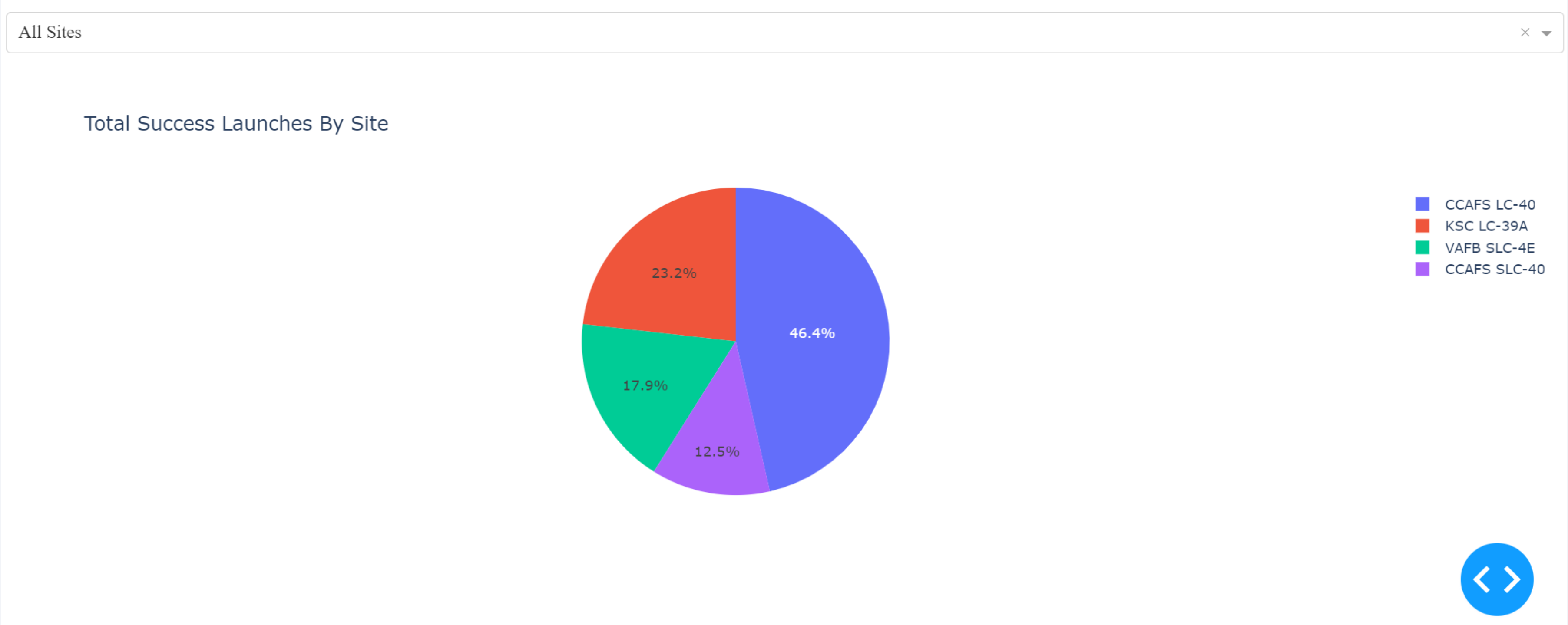


There is a coastline, and railway around 1.3 km from the site and the town of Lompoc about 13.12km away.

Section 4

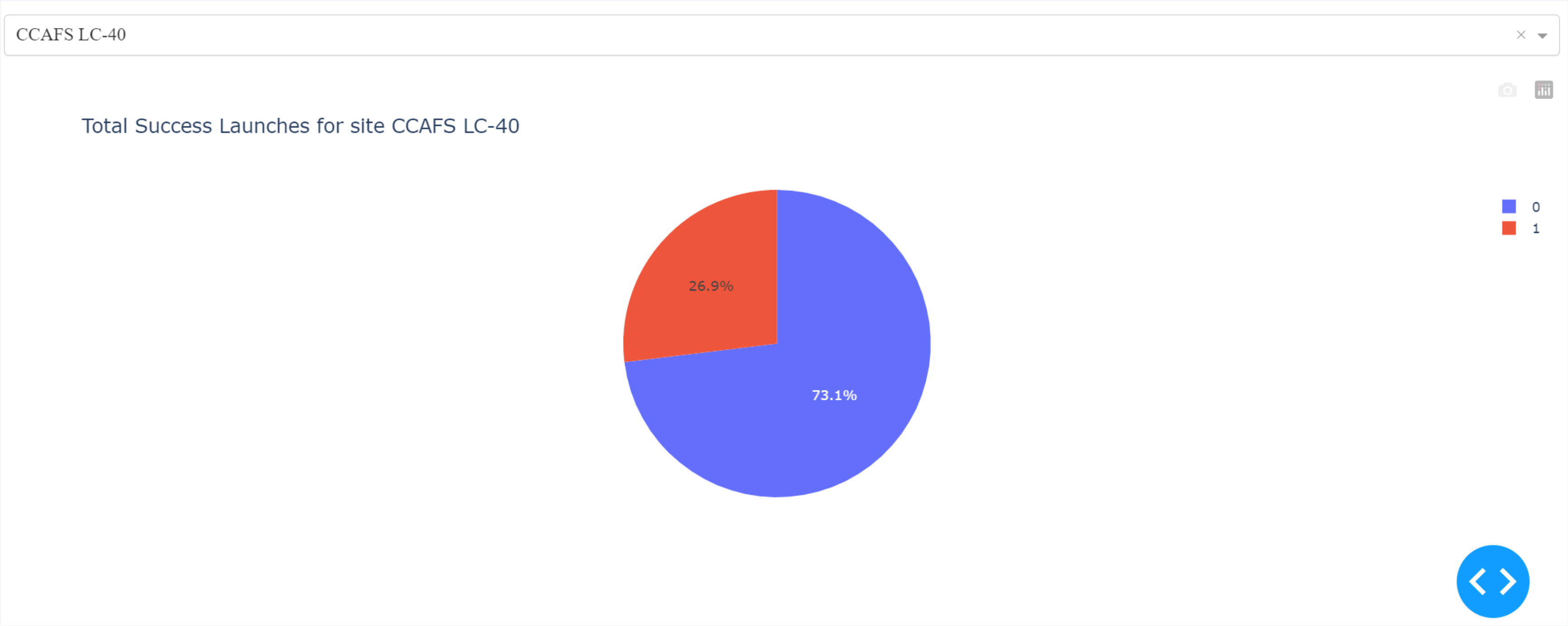
Build a Dashboard with Plotly Dash

Success Rates of All Launch Sites



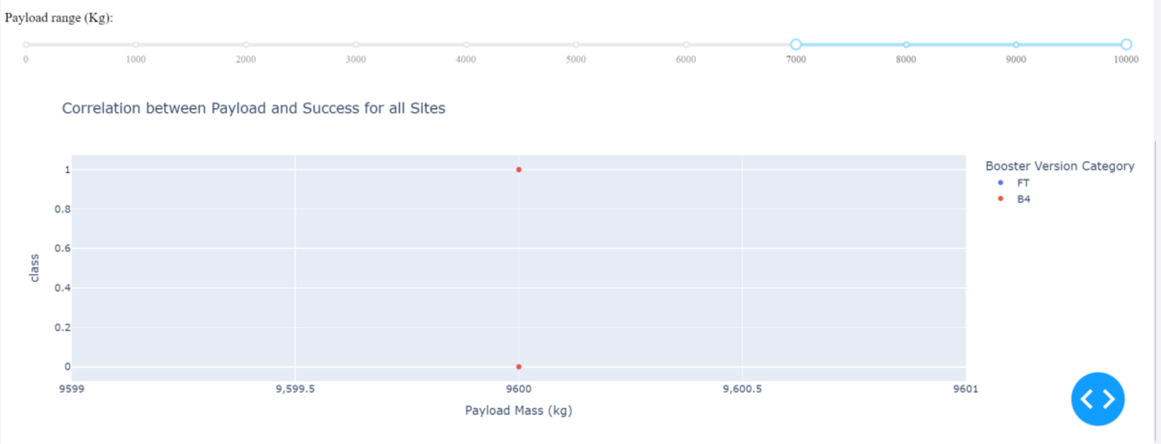
CCAFS LC-40 had the highest success rate followed by KSC LC-39A

Success Count of CCAFS LC-40



73.1% of all the launches from the CCAFS LC-40 Launch Site were a success.

Payload Mass vs Launch Outcomes



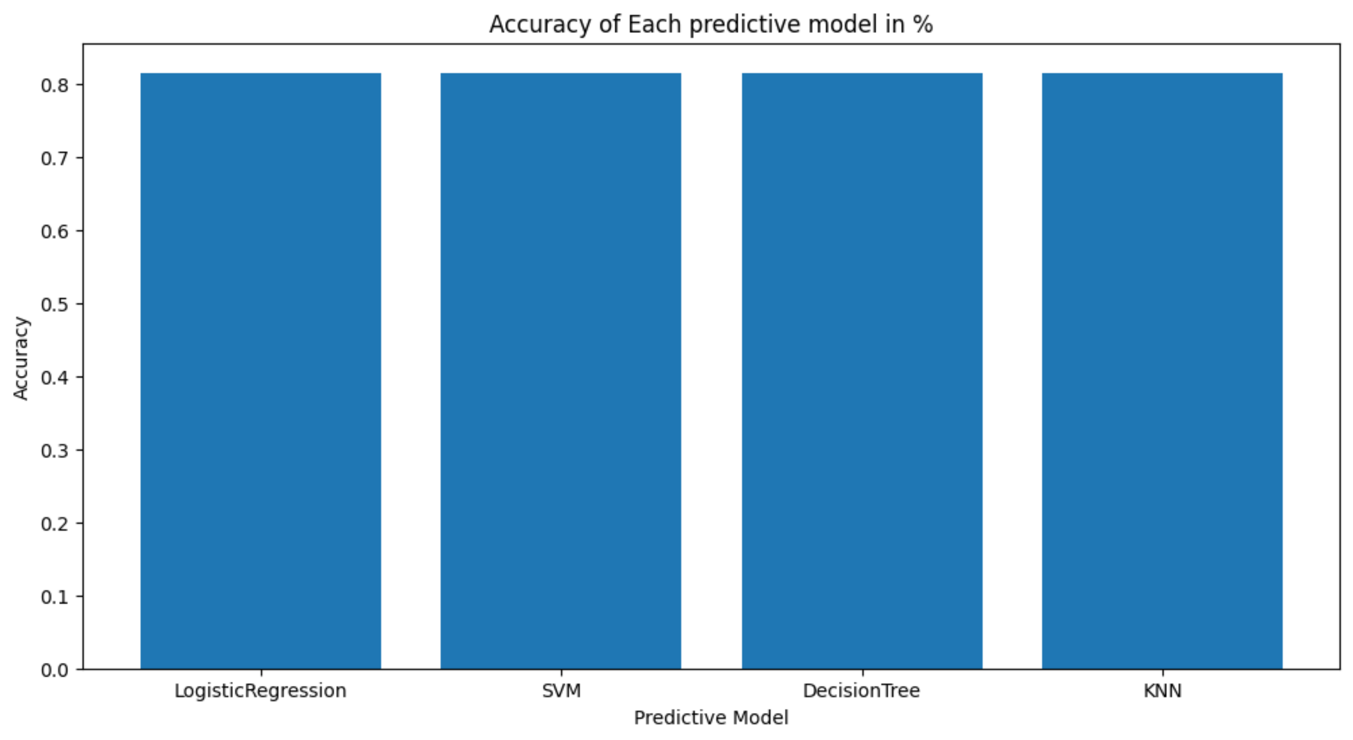
Most successful launches had a booster with payload mass between 1000kg and 7000kg.

Section 5

Predictive Analysis (Classification)

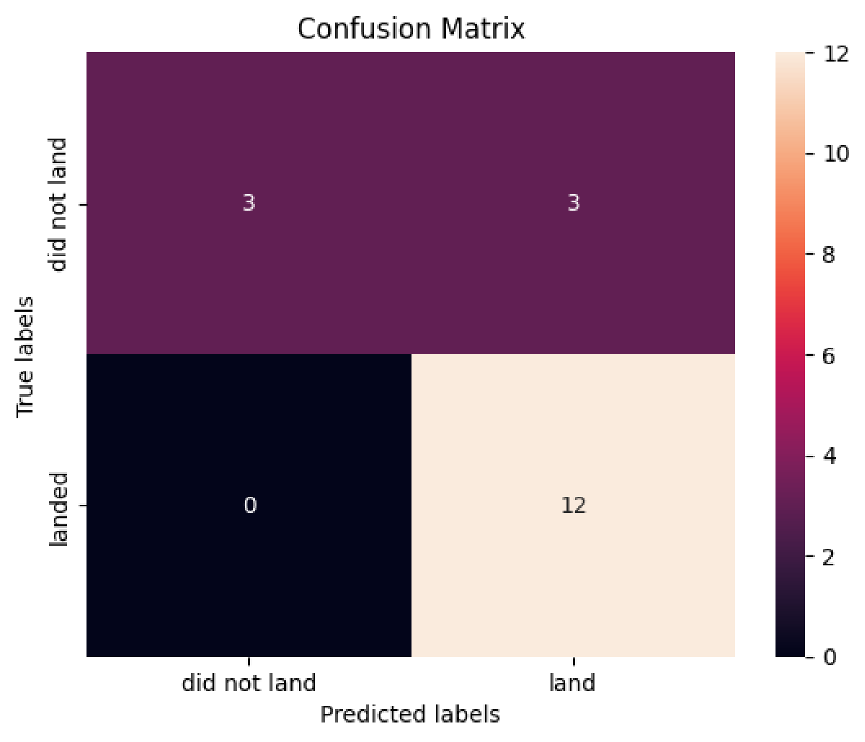
Classification Accuracy

As seen below, all predictive models have the same accuracy rate based on their F1_scores (81.48) and .score() accuracy values(83.33).



Confusion Matrix of the Best Performing Model

The Confusion Matrix of the Logistic Regression Model is shown below:



3 True Negatives, 3 False Positives, 0 True Negatives, 12 True Positives, from a total of 18 samples.

Conclusions

- Most Successful Falcon9 launches originate from the CCAFS LC-40 Launch Site.
- Most Successful Falcon9 launches have a payload mass between 1000kg and 7000kg
- Almost all Launch Sites are by the coastline and have either a city/town or a railroad next to them.
- All 4 predictive models used proved to be equally efficient in predicting the success of future Falcon9 Missions
- Thus, we can with a significant amount of confidence, predict the success of future Falcon9 missions, given that we are provided with relevant launch data such payload mass, launch site,

Appendix

Most of the python notebooks included in the links have the following imports:

```
# Pandas is a software library written for the Python programming language for data manipulation and analysis.
import pandas as pd
# NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, a
import numpy as np
# Matplotlib is a plotting library for python and pyplot gives us a MatLab like plotting framework. We will use this in our
import matplotlib.pyplot as plt
#Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attracti
import seaborn as sns
# Preprocessing allows us to standarsize our data
from sklearn import preprocessing
# Allows us to split our data into training and testing data
from sklearn.model_selection import train_test_split
# Allows us to test parameters of classification algorithms and find the best one
from sklearn.model_selection import GridSearchCV
# Logistic Regression classification algorithm
from sklearn.linear_model import LogisticRegression
# Support Vector Machine classification algorithm
from sklearn.svm import SVC
# Decision Tree classification algorithm
from sklearn.tree import DecisionTreeClassifier
# K Nearest Neighbors classification algorithm
from sklearn.neighbors import KNeighborsClassifier
```

```
import folium
import pandas as pd
```

```
# Import folium MarkerCluster plugin
from folium.plugins import MarkerCluster
# Import folium MousePosition plugin
from folium.plugins import MousePosition
# Import folium DivIcon plugin
from folium.features import DivIcon
```

Thank you!