IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Phillip Lee
2/27/2023

Overview

Executive Summary

- Summary of methodologies
  - Data Collection through SpaceX API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Display relationships between various features in the data
  - Visualize and analyze launch sites highlighted on interactive map
  - Predict Space X Falcon 9 First Stage Landing.

Introduction

- Project background and context

  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. In this module, you will be provided with an overview of the problem and the tools you need to complete the course.

- Problems

  - What factors determine a successful first-stage rocket landing?
  - What relationship do these factors have with each other?
  - Can we use those factors to predict whether or not a future launch will be successful?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected via the SpaceX API and web scraped from Wikipedia
- Perform data wrangling
  - One-hot encoding was applied to categorical features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build, fit, & evaluate classification models

Data Collection Method

GET request via SpaceX API

Convert to pandas data frame using **.json_normalize()**

Use **BeatifulSoup** to web scrape Falcon 9 launches

Extract as an HTML Table and parse it into a pandas data frame

Clean data

Fill in missing values

Data Collection – SpaceX API

`getBoosterVersion`

`getLaunchSite`

`getPayloadData`

`getCoredata`

https://github.com/phil-lee675/IBMDS0720EN/blob/main/Data%20Collection%20API.ipynb

Data Collection - Scraping

Request Falcon9 Launch Wiki from URL

Extract all column/variable names from HTML header

Create a dataframe by parsing launch HTML tables

https://github.com/phil-lee675/IBMDS0720EN/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling

Calculate number of launches on each site

Calcluate number and occurrence of each orbit

Calculate number and occurrence of mission outcome per orbit type

Create a landing outcome label from Outcome column

https://github.com/phil-lee675/IBMDS0720EN/blob/main/Data%20Wrangling.ipynb

EDA with Data Visualization

We want to visualize the relationships between the following variables:

- Flight Number vs Launch Site
- Payload vs Launch Site
- Success rate of each Orbit
- Flight Number and Orbit type
- Payload and Orbit type
- Launch Success yearly trend

https://github.com/phil-lee675/IBMDS0720EN/blob/main/EDA%20with%20Visualization.ipynb

EDA with SQL

With SQL query commands, we want to display or list the following:

- All Launch Site Names
- Launch Site Names that begin with 'KSC'
- Total Payload Mass carried by boosters launched by NASA (CRS)
- Average Payload Mass carried by Booster F9 v1.1
- The date where the first successful landing via drone ship was achieved.
- The names of the successful boosters that have a payload mass between 4000 & 6000
- Total number of successful and failure mission outcomes
- Names of boosters which have carried max payload
- Records of successful landings for 2017
- Ranking of successful landings between the June 4th 2010 and March 3rd 2020.

https://github.com/phil-lee675/IBMDS0720EN/blob/main/Exploratory%20Data%20Analysis%20with%20SQL.ipynb

Build an Interactive Map with Folium

With the following markings we can find the answer to the following:

- Are launch sites in close proximity to railways?
- Are launch sites in close proximity to highways?
- Are launch sites in close proximity to coastline?
- Do launch sites keep certain distance away from cities?

Mark all launch sites on a map	Calculate distances between a launch site to its proximities	Mark the closest railway
• Mark the success/failed launches for each site	• Have it follow mouse position on map	• Drew a line between marker and launch site

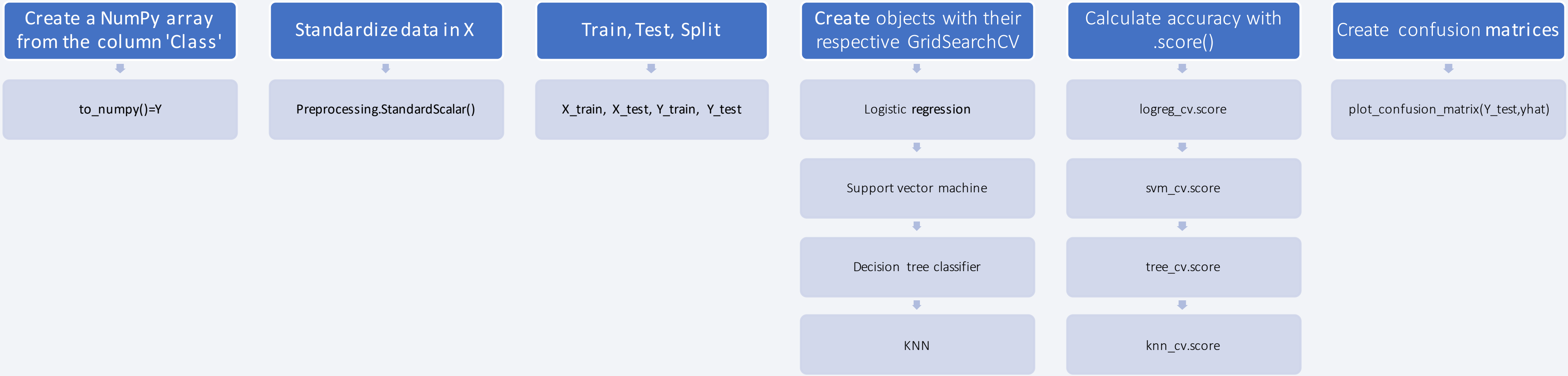
https://github.com/phil-lee675/IBMDS0720EN/blob/main/Launch%20Site%20Locations%20Analysis%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- Added a dropdown list to enable Launch Site selection
- Added a pie chart to show the total successful launches count for all sites
- Added a slider to select payload range
- Added a scatter chart to show the correlation between payload and launch success

https://github.com/phil-lee675/IBMDS0720EN/blob/main/PlotlyDash/spacex_dash_app.py

Predictive Analysis (Classification)



https://github.com/phil-lee675/IBMDS0720EN/blob/main/Machine%20Learning%20Prediction.ipynb

Results

Exploratory data analysis results

Interactive analytics demo in screenshots

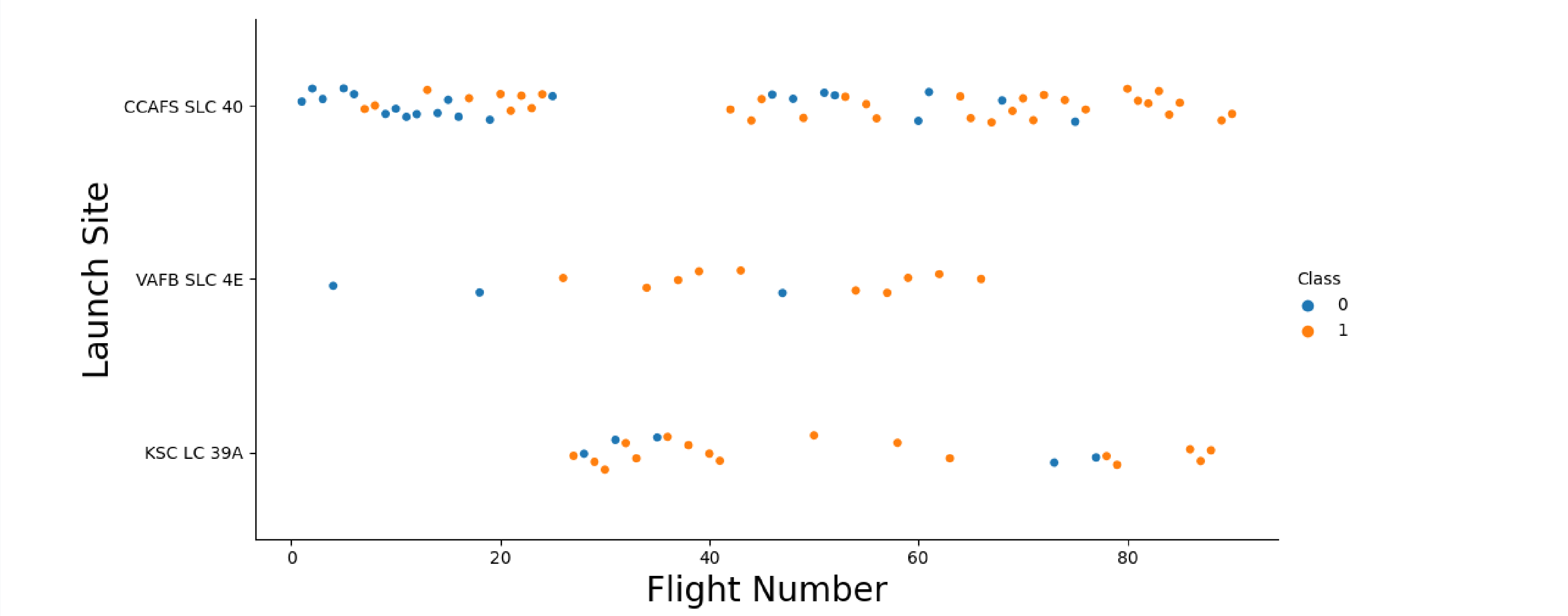
Predictive analysis results

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

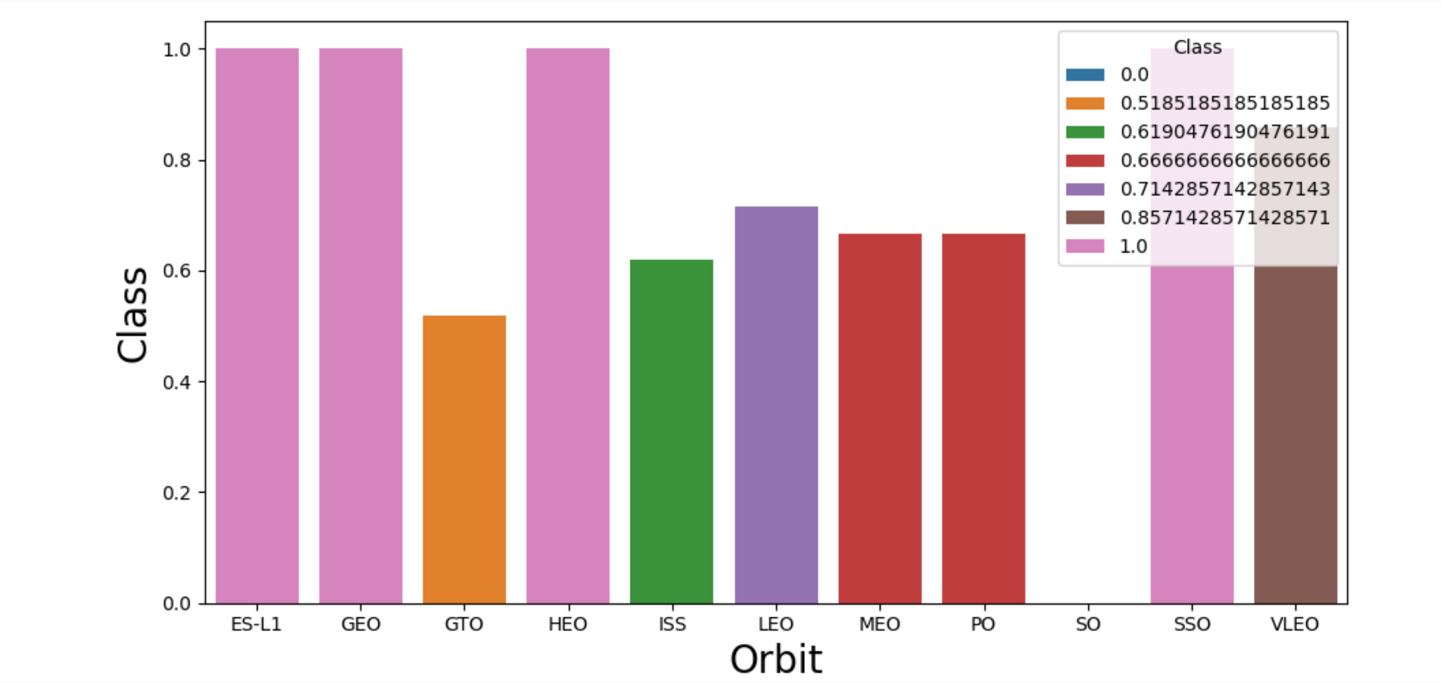
From the scatter plot, we see a relationship where the larger the flight count at a launch site, the greater the success rate.



Success Rate vs. Orbit Type

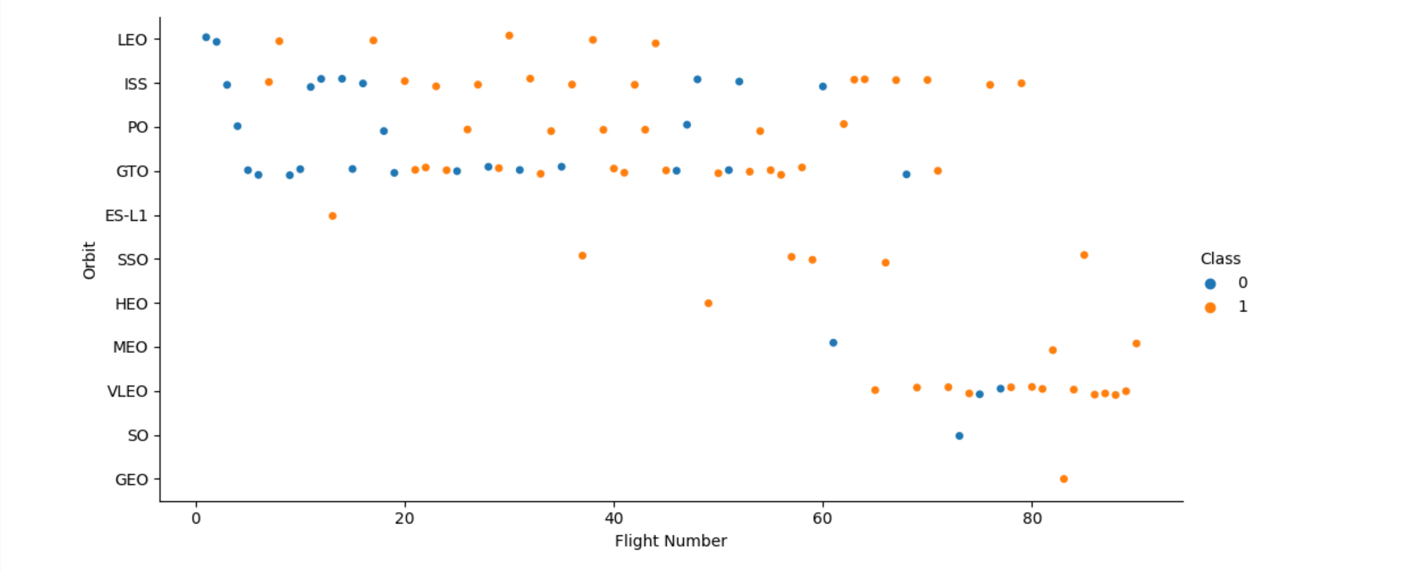
From the bar chart we see that the following orbits have a high success rate:

- ES-L1
- GEO
- HEO
- SSO



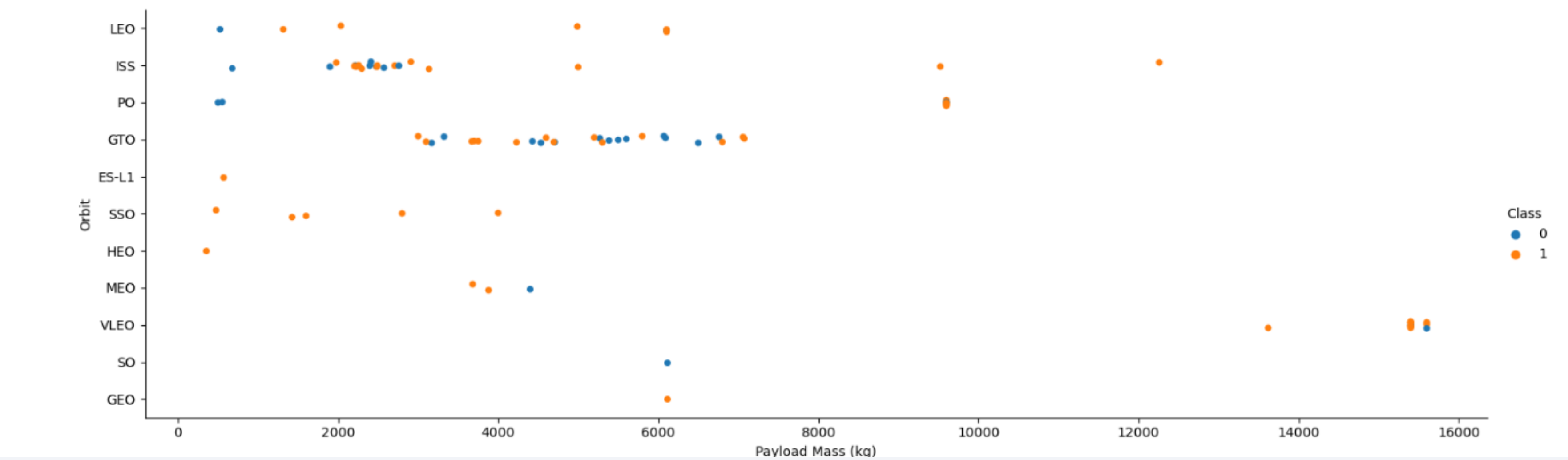
Flight Number vs. Orbit Type

- From the scatter plot, we 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.



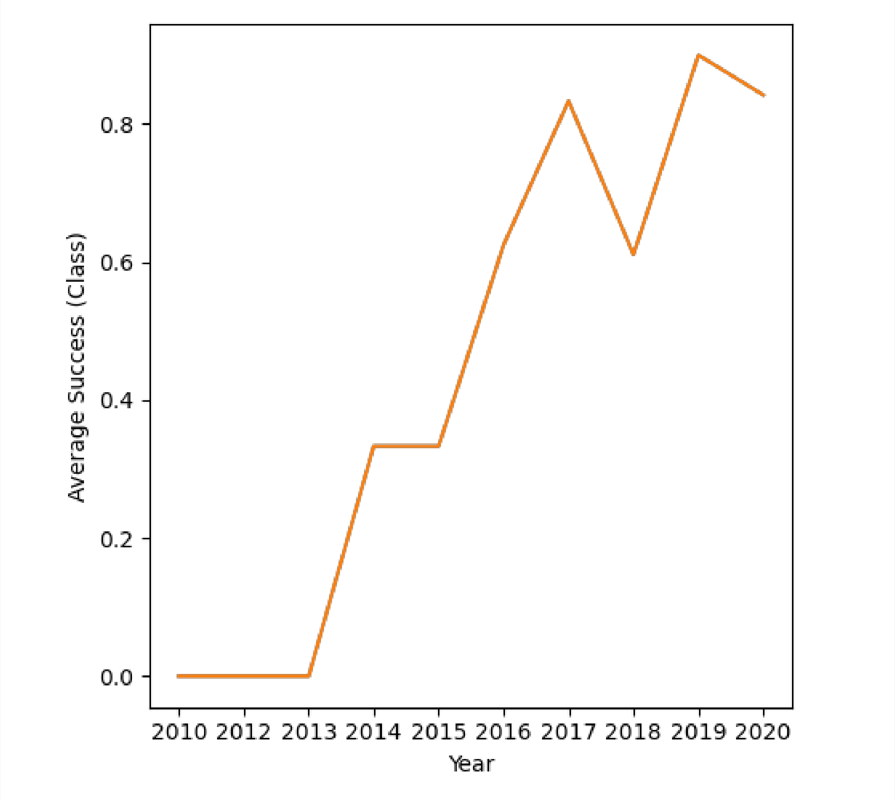
Payload vs. Orbit Type

- From the scatter plot, we see that 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

From the line plot we see
that the success rate since 2013
kept increasing till 2020



All Launch Site Names

```
%sql select distinct launch_site from spacex
```

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

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

Launch Site Names that begin with 'KSC'

```
%%sql select *
    from spacex
        where launch_site like 'KSC%' limit 5
```

- Display 5 records where launch sites begin with the string 'KSC'

DATE	TIME__UTC_	BOOSTER_VERSION	LAUNCH_SITE	PAYLOAD	PAYLOAD_MASS__KG_	ORBIT	CUSTOMER	MISSION_OUTCOME	LANDING__OUTCOME
2017-01-05	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2017-03-06	21:07:00	F9 FT B1035.1	KSC LC-39A	SpaceX CRS-11	2708	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-05-07	23:38:00	F9 FT B1037	KSC LC-39A	Intelsat 35e	6761	GTO	Intelsat	Success	No attempt
2017-07-09	14:00:00	F9 B4 B1040.1	KSC LC-39A	Boeing X-37B OTV-5	4990	LEO	U.S. Air Force	Success	Success (ground pad)
2017-11-10	22:53:00	F9 FT B1031.2	KSC LC-39A	SES-11 / EchoStar 105	5200	GTO	SES EchoStar	Success	Success (drone ship)

Total Payload Mass

```
%%sql select sum(payload_mass__kg_) from spacex
 where customer='NASA (CRS)'
```

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

```
In [67]: %%sql select sum(payload_mass__kg_)
             from spacex
                 where customer='NASA (CRS)'

 * ibm_db_sa://mwr47214:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

Out[67]:

1
22007

Average Payload Mass by F9 v1.1

```
%%sql select avg(payload_mass__kg_)
    from spacex
        where booster_version='F9 v1.1'
```

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

In [66]:
```
%%sql select avg(payload_mass__kg_)
    from spacex
        where booster_version='F9 v1.1'
```

```
 * ibm_db_sa://mwr47214:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

Out[66]:

1
3676

First Successful Ground Landing Date

```
%%sql select min(DATE)
    from spacex
        where landing__outcome='Success (drone ship)'
```

- List the date where the first successful landing outcome in drone ship was achieved.

```
In [65]: %%sql select min(DATE)
             from spacex
                 where landing__outcome='Success (drone ship)'

 * ibm_db_sa://mwr47214:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

Out[65]:

1
2016-06-05

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql select booster_version, payload_mass__kg_, landing__outcome
    from spacex
        where landing__outcome='Success (ground pad)' AND payload_mass__kg_
  between 4000 AND 6000
```

- List the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000

```
In [64]: %%sql select booster_version, payload_mass__kg_, landing__outcome
             from spacex
                 where landing__outcome='Success (ground pad)' AND payload_mass__kg_ between 4000 AND 6000
```

```
 * ibm_db_sa://mwr47214:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

Out[64]:

booster_version	payload_mass__kg_	landing__outcome
F9 FT B1032.1	5300	Success (ground pad)
F9 B4 B1040.1	4990	Success (ground pad)
F9 B4 B1043.1	5000	Success (ground pad)

Total Number of Successful and Failure Mission Outcomes

```
%%sql select
   count(mission_outcome) as
   Successes

     from spacex

          where mission_outcome
   like 'Success%'
```

- List the total number of successful mission outcomes

```
%%sql select
   count(mission_outcome) as
   Failures

     from spacex

          where mission_outcome
   like 'Failure%'
```

- List the total number of failure mission outcomes

In [11]:

```
%%sql select count(mission_outcome) as Successes
    from spacex
        where mission_outcome like 'Success%'
```

```
 * ibm_db_sa://mwr47214:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

Out[11]:

successes
45

In [12]:

```
%%sql select count(mission_outcome) as Failures
    from spacex
        where mission_outcome like 'Failure%'
```

```
 * ibm_db_sa://mwr47214:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

Out[12]:

failures
0

Boosters Carried Maximum Payload

```
%%sql select booster_version, payload_mass__kg_
    from spacex where payload_mass__kg_ =
        (select max(payload_mass__kg_) from spacex)
```

- List the names of the *booster_versions* which have carried the maximum payload mass using a subquery

```
In [62]: %%sql select booster_version, payload_mass__kg_
             from spacex where payload_mass__kg_ =
                 (select max(payload_mass__kg_) from spacex)

 * ibm_db_sa://mwr47214:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

Out[62]:

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600

2017 Launch Records

```
%%sql select to_char(to_date(month(DATE), 'MM'), 'MONTH') as
  month, landing__outcome, booster_version, launch_site

    from spacex

        where landing__outcome = 'Success (ground pad)' and
  DATE like '2017%'
```

- List the records which will display the month names, successful *landing_outcomes* in ground pad, *booster versions, launch_site* for the months in year 2017

```
[91]: %%sql select to_char(to_date(month(DATE), 'MM'), 'MONTH') as month, landing__outcome, booster_version, launch_site
          from spacex
              where landing__outcome = 'Success (ground pad)' and DATE like '2017%'
```

```
 * ibm_db_sa://mwr47214:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

[91]:

MONTH	landing__outcome	booster_version	launch_site
JANUARY	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
MARCH	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
JULY	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A

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

```
%%sql select DATE,
   count(landing__outcome) as count
   from spacex

        where DATE between '2010-06-04'
   and '2017-03-20' and
   landing__outcome like '%Success%'

        group by DATE

        order by
   count(landing__outcome) DESC
```

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

```
%%sql select
   count(mission_outcome) as
   Failures

     from spacex

          where mission_outcome
   like 'Failure%'
```

- Display all the Successful Landing Outcomes

In [103…

```
%%sql select DATE, count(landing__outcome) as count from spacex
        where DATE between '2010-06-04' and '2017-03-20' and landing__outcome like '%Success%'
        group by DATE
        order by count(landing__outcome) DESC
```

* ibm_db_sa://mwr47214:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.

Out[103…

DATE	COUNT
2016-06-05	1
2016-08-04	1
2017-01-05	1
2017-03-06	1

In [101…

```
%%sql select DATE,landing__outcome
    from spacex
        where DATE between '2010-06-04' and '2017-03-20' and landing__outcome like '%Success%'
```

* ibm_db_sa://mwr47214:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.

Out[101…

DATE	landing__outcome
2016-08-04	Success (drone ship)
2016-06-05	Success (drone ship)
2017-01-05	Success (ground pad)
2017-03-06	Success (ground pad)

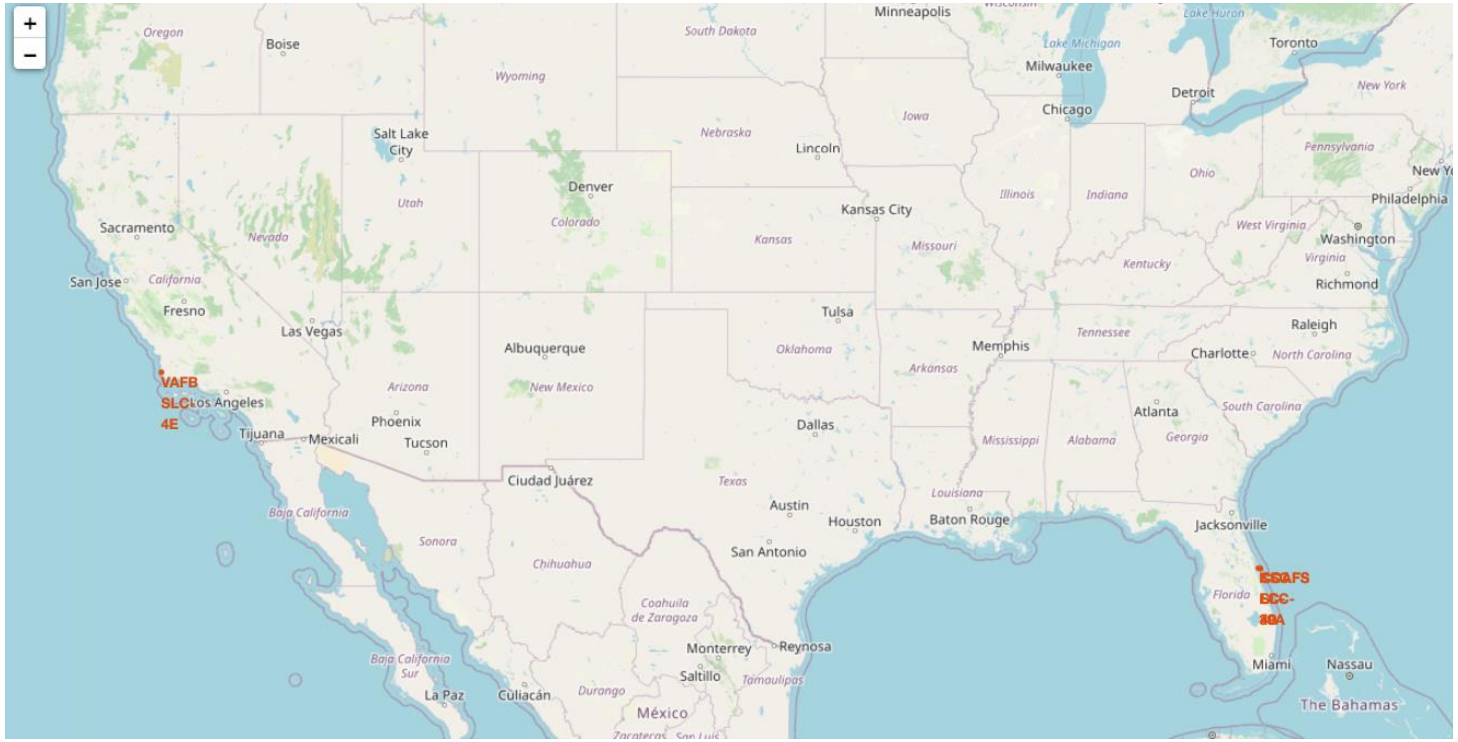
Section 3

Launch Sites Proximities Analysis

Mark all Launch Sites on a Map

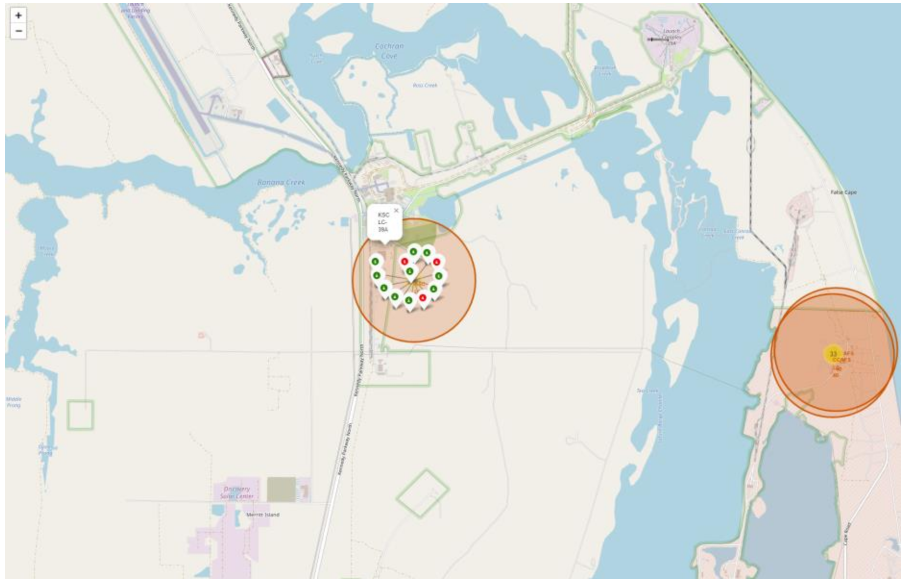
From the map, we're able to see that

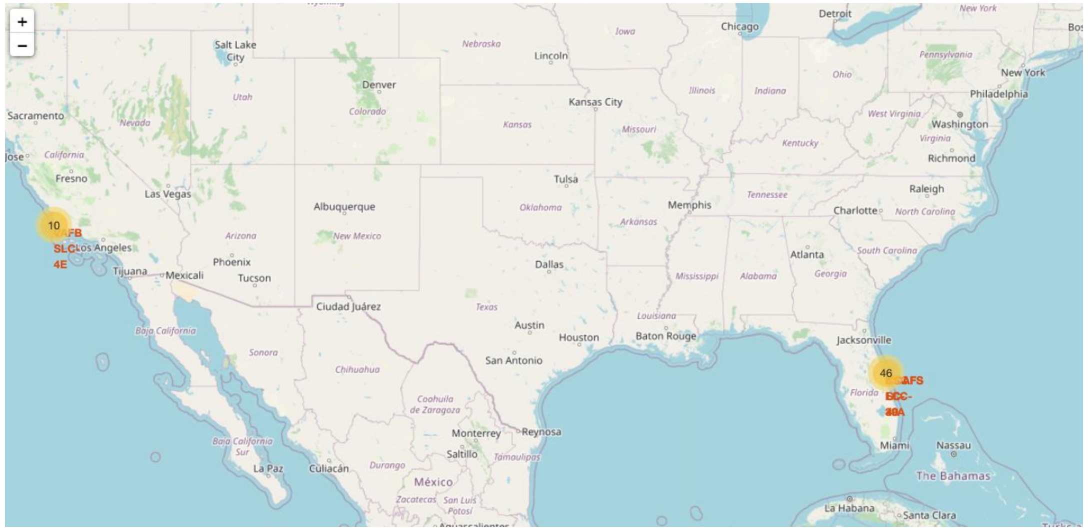
- All launch sites are in proximity to the equator line
- All launch sites are in very close proximity to the coast



Mark the success/failed launches for each site on the map

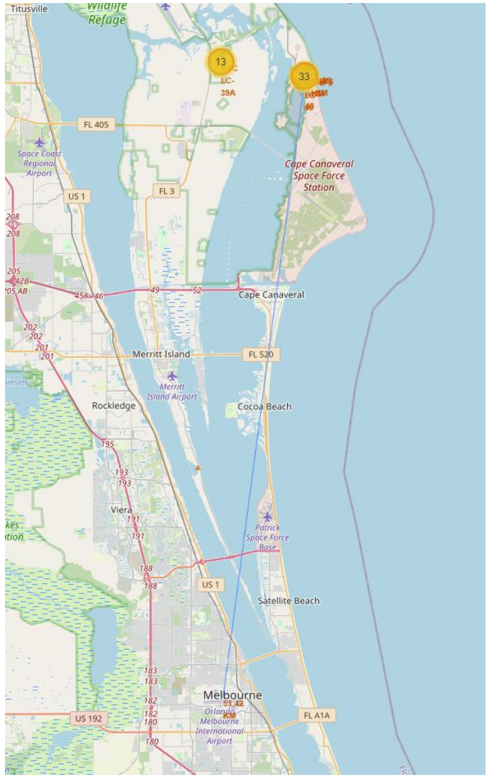
- From the color-labeled markers in marker clusters, we can easily identify which launch sites have relatively high success rates.
- Green Markers show successful launches
- Red Markers show failures

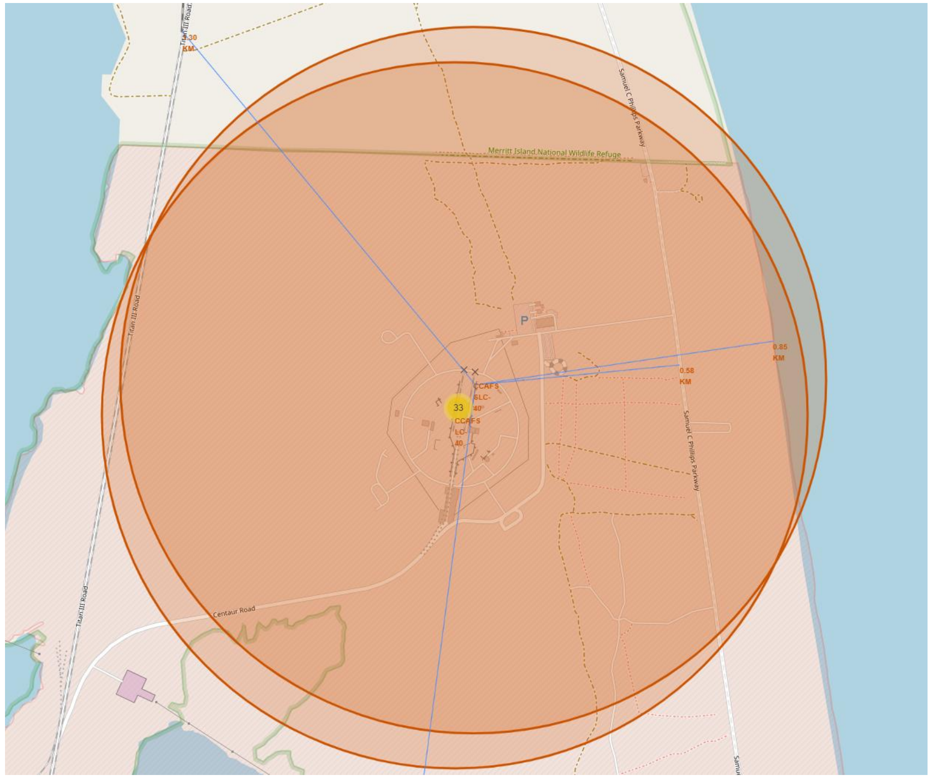




Calculate the distances between a launch site to its proximities

- Launch Sites are in close proximity to railways
- Launch Sites are in close proximity to highways
- Launch Sites are in close proximity to the coastline
- Launch Sites keep far away from cities



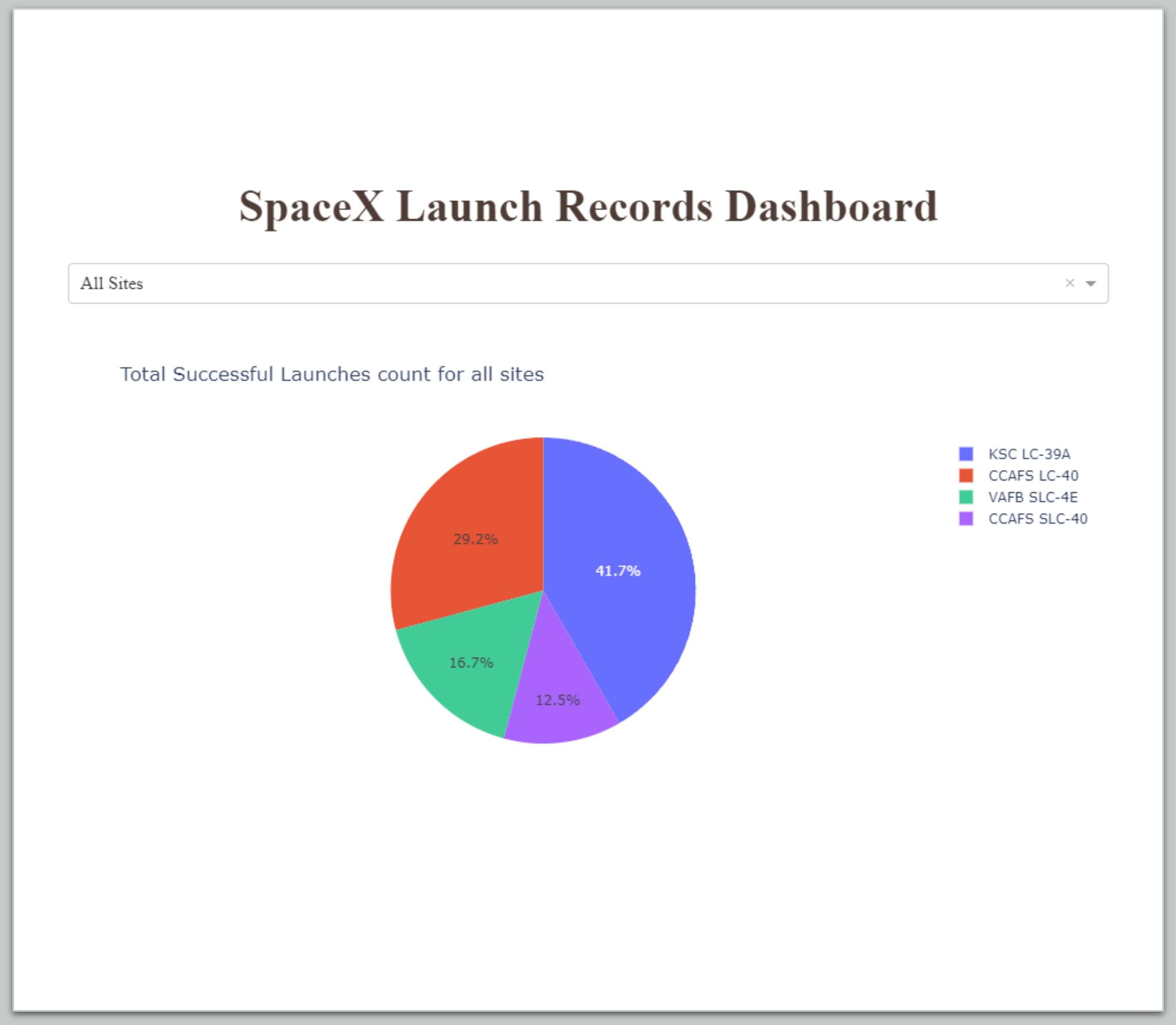


Section 4

Build a Dashboard with Plotly Dash

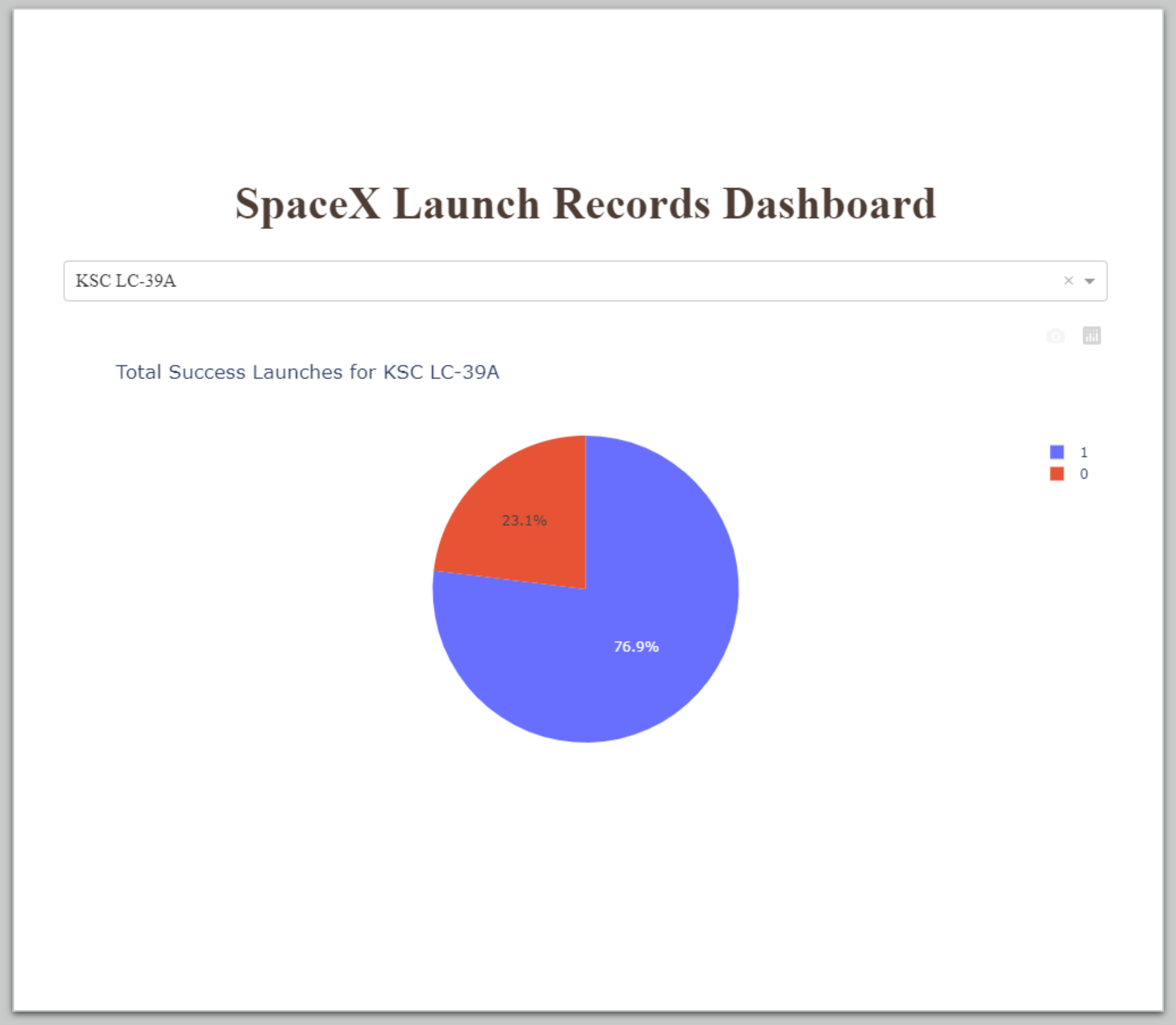
SpaceX Falcon9 Interactive Launch Dashboard

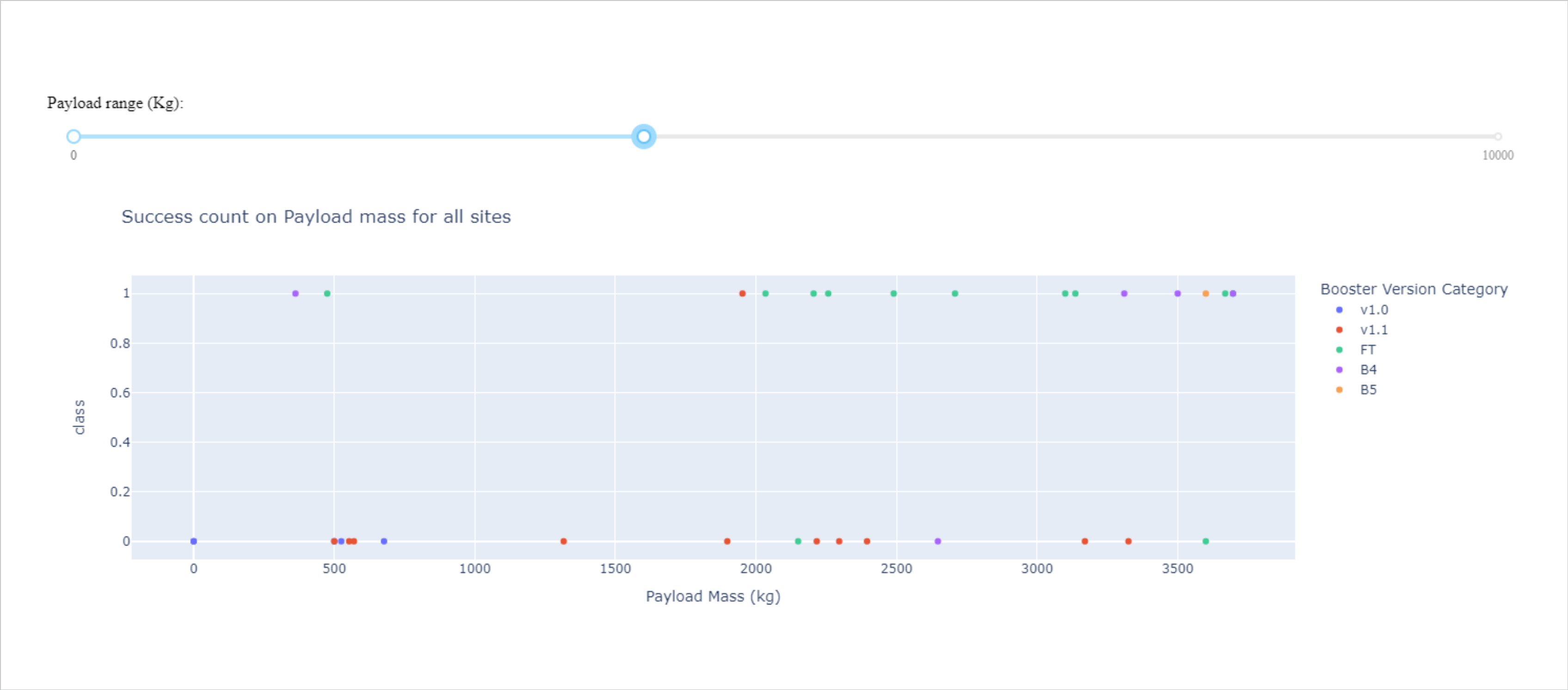
- From the pie chart, we can see all the successful launches from each site.
- KSC LC-39A has the most successful launches.
- CCAFS SLC-40 has the least.



Launch Site with the Highest Launch Success Ratio

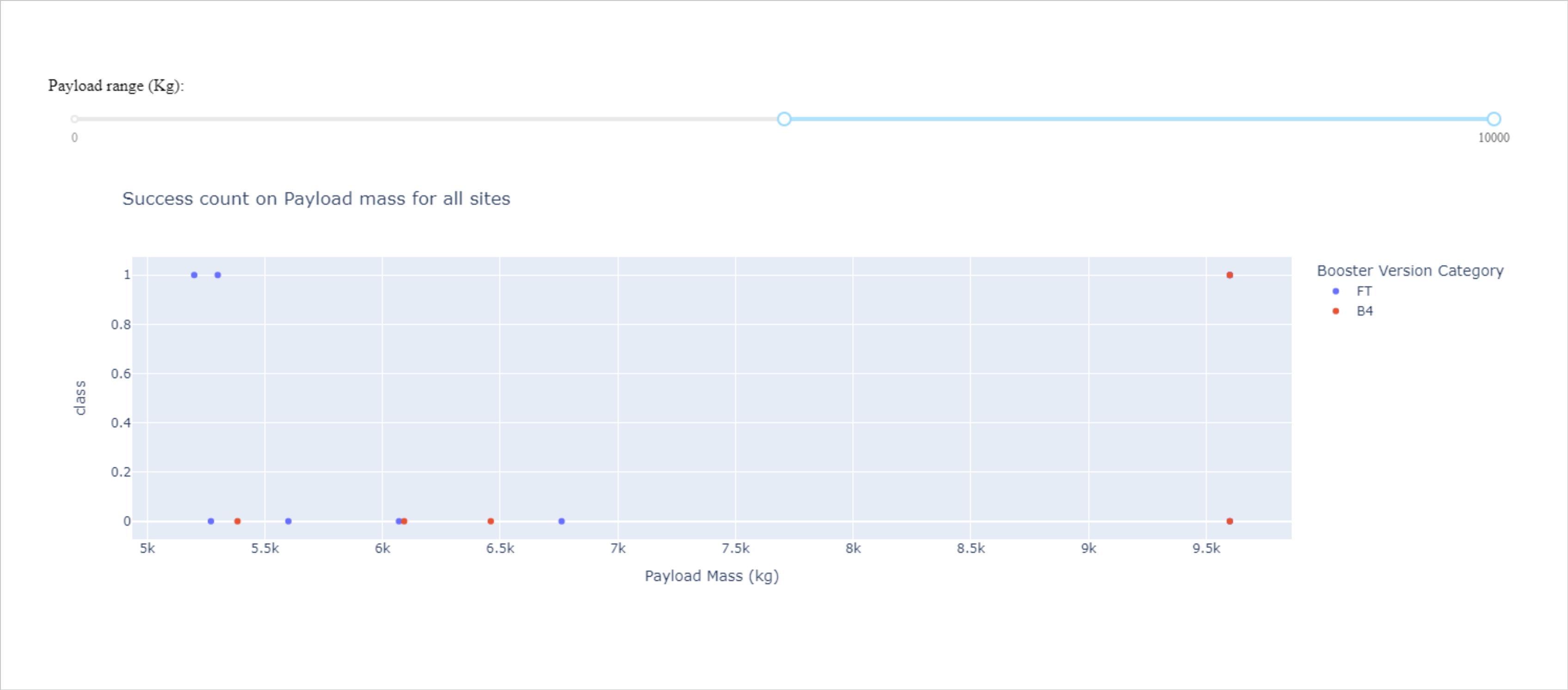
- KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate





Payload vs Launch Outcome Interactive Visual Analysis: Low Payload Mass (kg)

- A scatter plot is created to display the Success Count vs Payload Mass for all sites
- Adjust range slider to display low payload mass (0-4k)
- Note the high success count
- We also see that the FT booster has the highest launch success rate



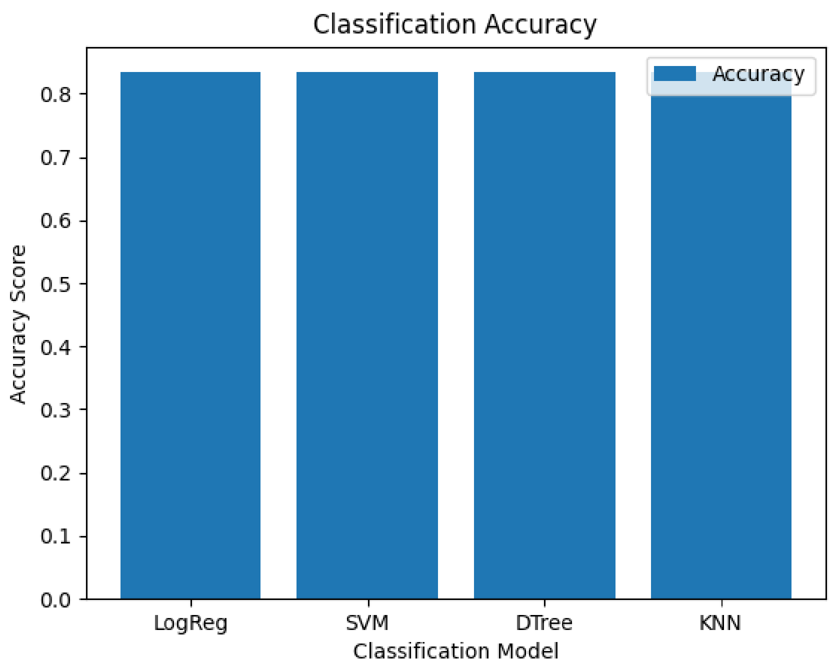
Payload vs Launch Outcome Interactive Visual Analysis: High Payload Mass (kg)

- Adjust range slider to display high payload mass (5k-10k)
- Note the low success count

Section 5

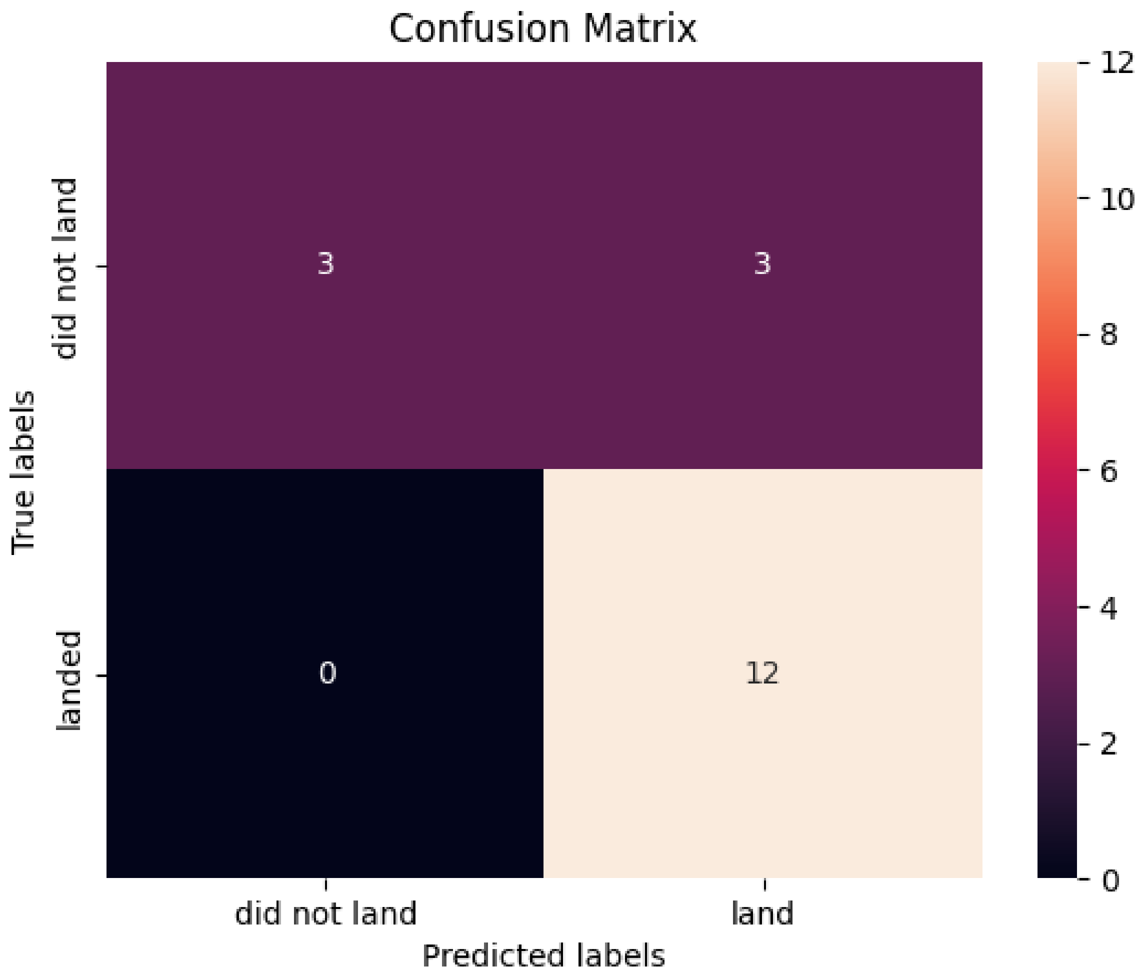
Predictive Analysis (Classification)

Classification Accuracy



- Classification Models:
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree Classifier
  - K-Nearest Neighbors

- All models had a test data accuracy of around .83%



Confusion Matrix

- All models shared the same confusion matrix
- The confusion matrix shows that the classifier can distinguish between the different classes.
- The major problem is the false positives (unsuccessful landing marked as successful landing)

Conclusions

- The larger the flight count at a launch site, the greater the success rate at said launch site.
- Launch success rate started to increase in 2013 until 2020.
- Orbits ES-L1, GEO, HEO, & SSO had the highest success rate.
- KSC LC-39A had the most successful launches of any sites with a 76.9% success rate and 23.1% failure rate
- Any of the aforementioned classification algorithms work well for this task.

Appendix

Cross-Validation data using GridSearchCV:

- LogReg
  - `tuned hyperparameters :(best parameters) {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}`
  - **`accuracy : 0.8464285714285713`**
- SVM
  - `tuned hyperparameters :(best parameters) {'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}`
  - **`accuracy : 0.8482142857142856`**
- Decision Tree
  - `tuned hyperparameters :(best parameters) {'criterion': 'entropy', 'max_depth': 6, 'max_features': 'auto', 'min_samples_leaf': 4, 'min_samples_split': 10, 'splitter': 'random'}`
  - **`accuracy : 0.8910714285714286`**
- KNN
  - `tuned hyperparameters :(best parameters) {'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}`
  - **`accuracy : 0.8482142857142858`**

Thank you!