IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Thanasis Chousiadas

April 2022

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

- Project background and context
- Problems you want to find answers

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SPACEX REST API
  - Web scraping Wikipedia’s website about SPACEX launches
- Perform data wrangling
  - Filtering data, sampling data and dealing with Nulls
- 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

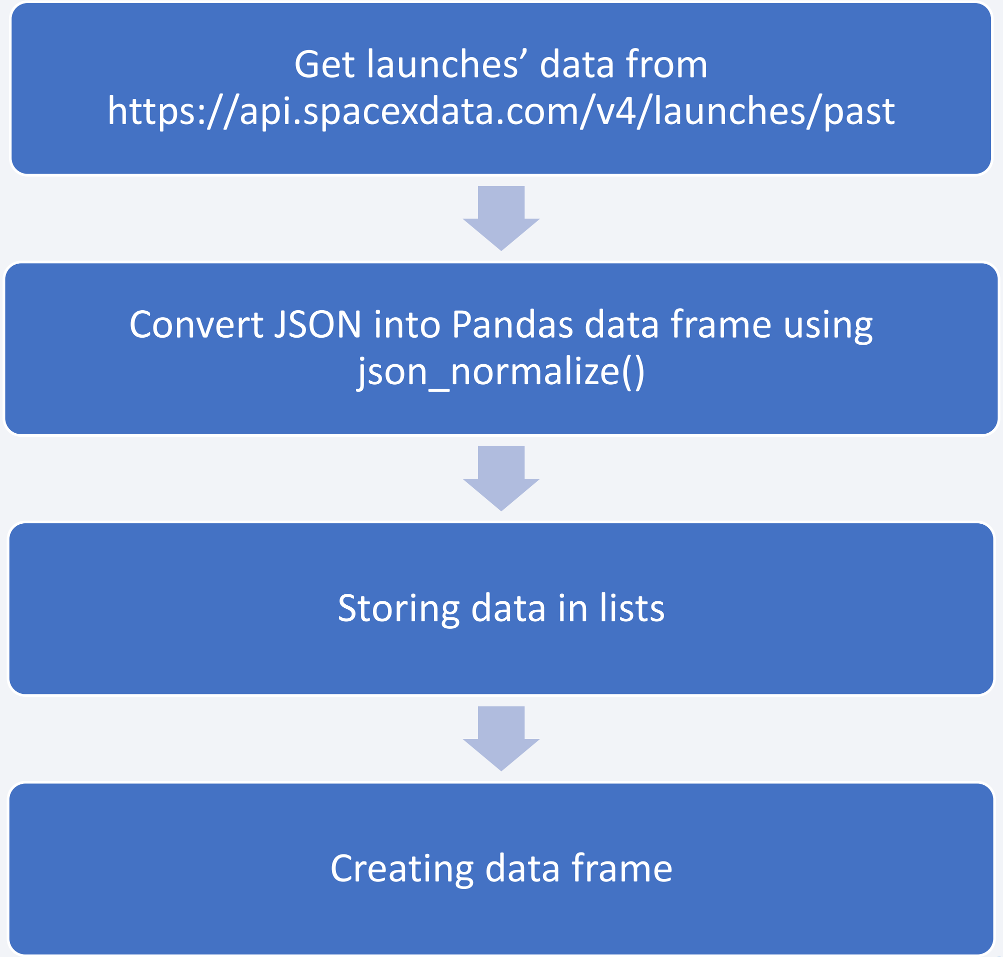
- Datasets were collecting using two different approaches.
  - The first approach is to use the SPACEX REST API and making request for receiving the data.
  - The second approach is to web-scraping public websites for the data. For that purpose, we used Wikipedia’s SPACEX website where data was presented in html tables.
- We performed requests on SPACEX API endpoints to get the data and save them in Panda's data frame.
- We web-scraped Falcon 9 launch records from Wikipedia using Python’s BeautifulSoup package.

Data Collection – SpaceX API

- Get data from SPACEXAPI.
- Convert data from JSON format to data frame.
- Using helper functions we get the data for each launch, such as booster version, launch site, payload data etc.
- Store data in list for each column.
- From these lists we create the dataset.

Notebook's GitHub link:

- https://github.com/SakisHous/spacex-data-science/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Get launches' data from https://api.spacexdata.com/v4/launches/past

Convert JSON into Pandas data frame using json_normalize()

Storing data in lists

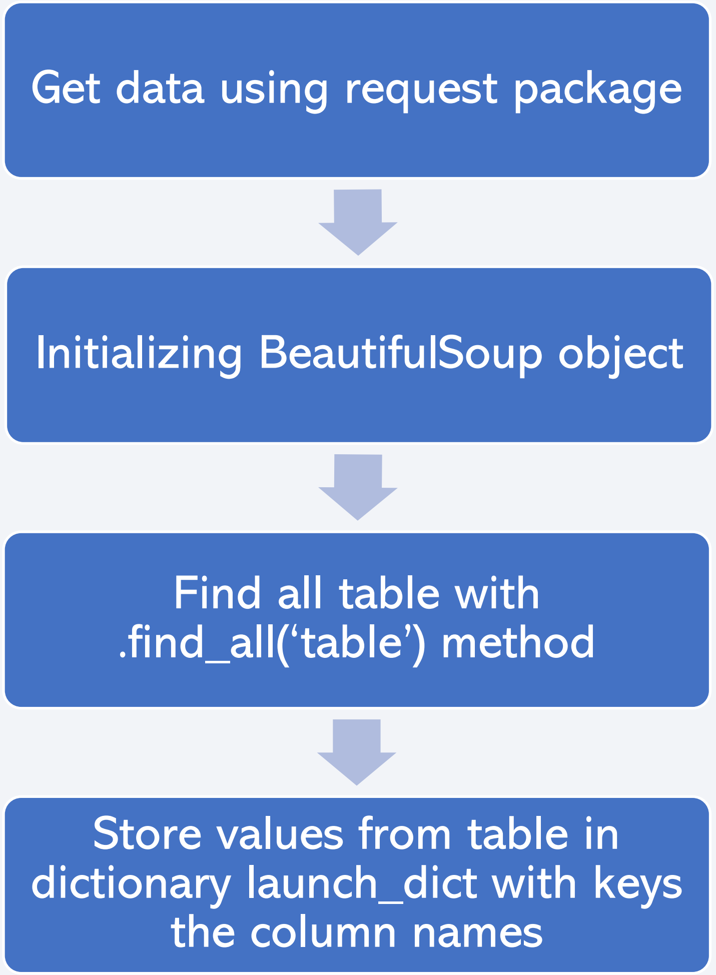
Creating data frame

Data Collection - Scraping

- Extract a Falcon 9 launch records HTML table from Wikipedia
- Parse the table and convert it into a Pandas data frame

Notebook's GitHub link:

- https://github.com/SakisHous/spacex-data-science/blob/main/jupyter-labs-webscraping.ipynb



Get data using request package

Initializing BeautifulSoup object

Find all table with .find_all('table') method

Store values from table in dictionary launch_dict with keys the column names

Data Wrangling

- Data wrangling for dataset received from SPACEX API
  - Find rows with missing values and replacing them with the mean value. We saved the new data set as dataset_part_1.csv
  - GitHub Link: https://github.com/SakisHous/spacex-data-science/blob/main/jupyter-labs-spacex-data-collection-api.ipynb (Data Wrangling Section)
- Data wrangling and Exploratory Data Analysis (EDA) to determine the training labels for dataset_part_1.csv dataset.
  - Calculating the percentage of missing values for each column
  - Identifying which columns are numerical and categorical
  - Calculating the number of launches for each site
  - Determining the number and occurrence of each orbit in the column “Orbit”.
  - Calculating the number and occurrence of mission outcome per orbit type
  - Creating new column, named “Class” which represents the outcome of each launch. If the value is zero, the first stage did not land successfully; one means the first stage landed Successfully
  - GitHub Link: https://github.com/SakisHous/spacex-data-science/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Using seaborn package, we plotted scatter plot with catplot() overlaying the outcome of the launch setting the parameter hue to “Class” :
  - x=‘FlightNumber’ vs y=‘PayloadMass’
  - x=‘FlightNumber’ vs y=‘LaunchSites’
  - x=‘PayloadMass’ vs y=‘LaunchSites’
  - x=‘FlightNumber’ vs y=‘Orbit type’
  - x=‘Payload Mass’ vs y=‘Orbit’
- Bar chart for the success rate of each orbit, using seaborn’s barplot() where x=‘Orbit’ and y=‘Class’
- Plot line for yearly success rate, using seaborn’s lineplot() where x=‘Year’ and y=‘Class’

We used that charts to find some insights about the variables that would affect the success rate and how we will select the features that they will be used in success prediction.

GitHub Url: https://github.com/SakisHous/spacex-data-science/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

SQL queries we performed:

- 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)
- Display 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 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

- GitHub URL: https://github.com/SakisHous/spacex-data-science/blob/main/jupyter-labs-eda-sql-coursera.ipynb

Build an Interactive Map with Folium

- We created with folium.Circle objects to represent the locations of different sites on folium map and folium.Marker objects to add labels for these location. For example, we added NASA Johnson Space Center on folium and a text label. We make the same for each launches that they took place in different launch sites.
- We added MarkerCluster object to denote the success rate (Class column value 1) with green and failure rate (Class column value 0) with red.
- We created a folium.MousePosition object which gives latitude and longitude in the upper left corner. We could mark a location in the map such as railway and calculate the distance for a given launch site.

- GitHub URL: https://github.com/SakisHous/spacex-data-science/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

The graphs that we created in the Dashboard with Plotly Dash, they give us good insights about, how launch sites are correlated with success rate. In addition, we can find which launch site is used most for rockets with different payloads.

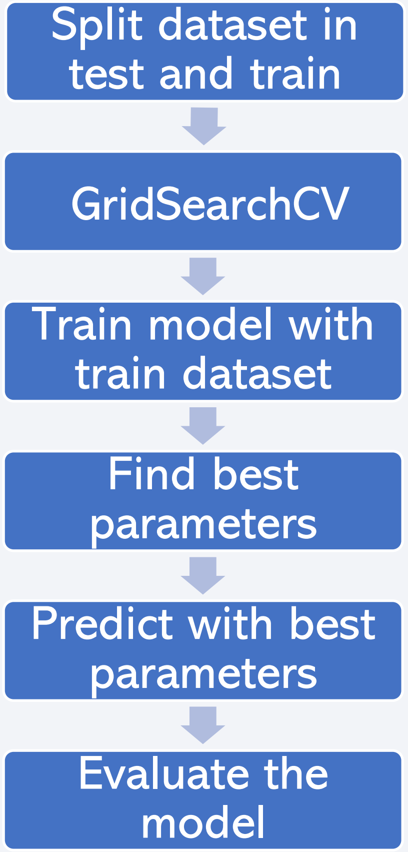
For that purpose, we added a scatter plot which gives this information clearer than the pie chart. In order to take more information and insights we added a slider where we could change payload range and get only launch within the payload range.

GitHub URL for Python code https://github.com/SakisHous/spacex-data-science/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- We split the dataset into train and test dataset.
- We used GridSearchCV to find the best parameters for each model.
- Models were Logistic Regression, Support Vector Machine, Decision Tree and k-Nearest Neighbors.
- With GridSearchCV could find best parameter for each model.
- We predicted the launch outcome for the test dataset.
- We evaluate the model.

GitHub URL: https://github.com/SakisHous/spacex-data-science/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb


Split dataset in test and train
GridSearchCV
Train model with train dataset
Find best parameters
Predict with best parameters
Evaluate the model

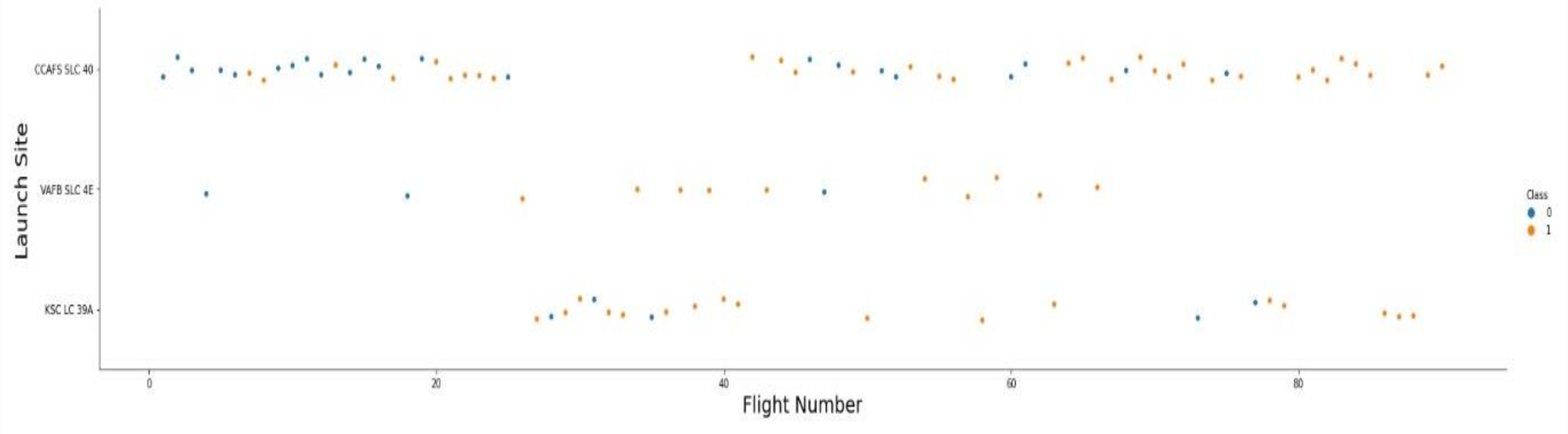
Results

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

Section 2

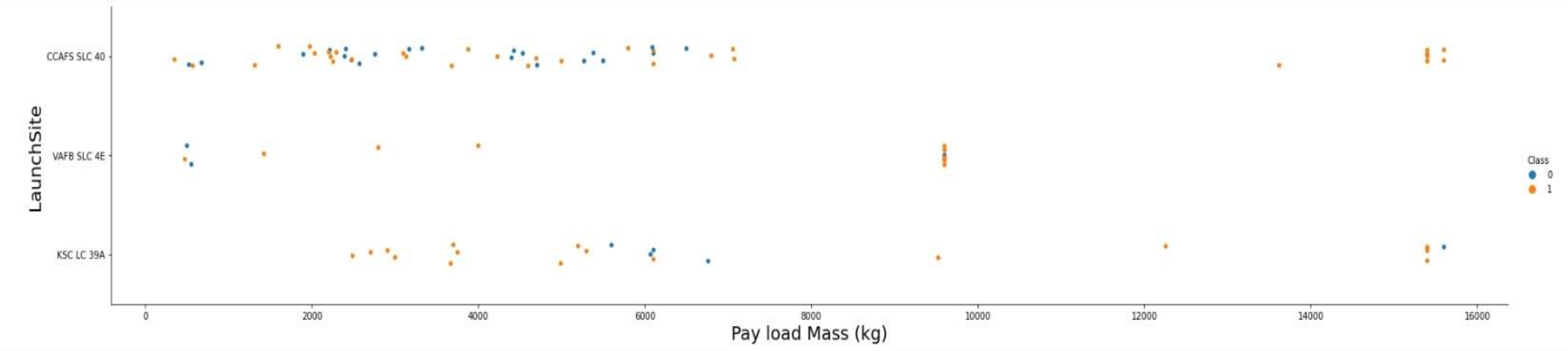
Insights drawn from EDA

Flight Number vs. Launch Site



As Flight Number increases, success rate also increases for all launch sites. Although, we can see that for Vandenberg Space Launch Complex 4 (VAFB-SLC) the flights were stopped. In addition, for the Kennedy Space Center Launch Complex 39 (KSC LC-39A) the flights began later compared to other launch sites. Last, the launch site with the most launches is Cape Canaveral Space Launch Complex 40 (CCAFS LC-40).

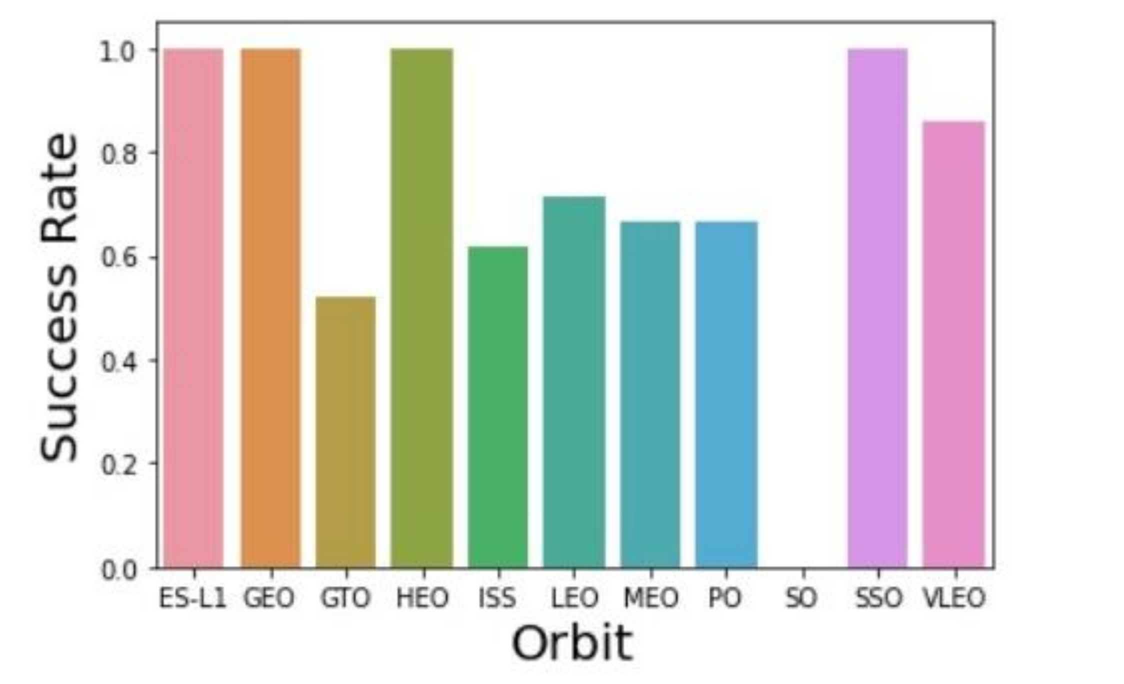
Payload vs. Launch Site



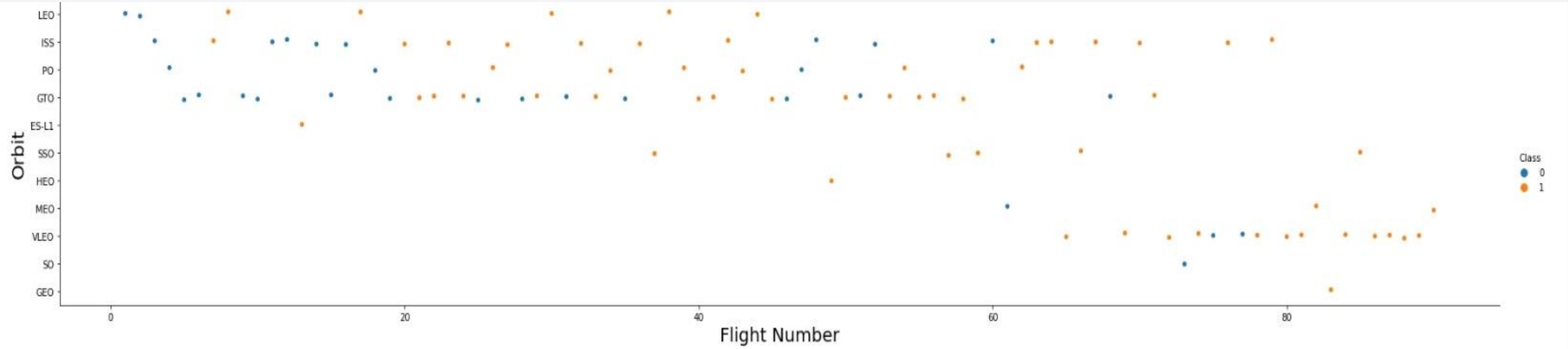
Firstly, we can see for VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000). Most rockets have launched from CCAFS SLC 40 launch site. In addition, many heavy payload rockets have launched from KSC LC 39A.

Success Rate vs. Orbit Type

As we can see the orbits with high success rate are ES-L1, GEO, HEO, SSO and VLEO.



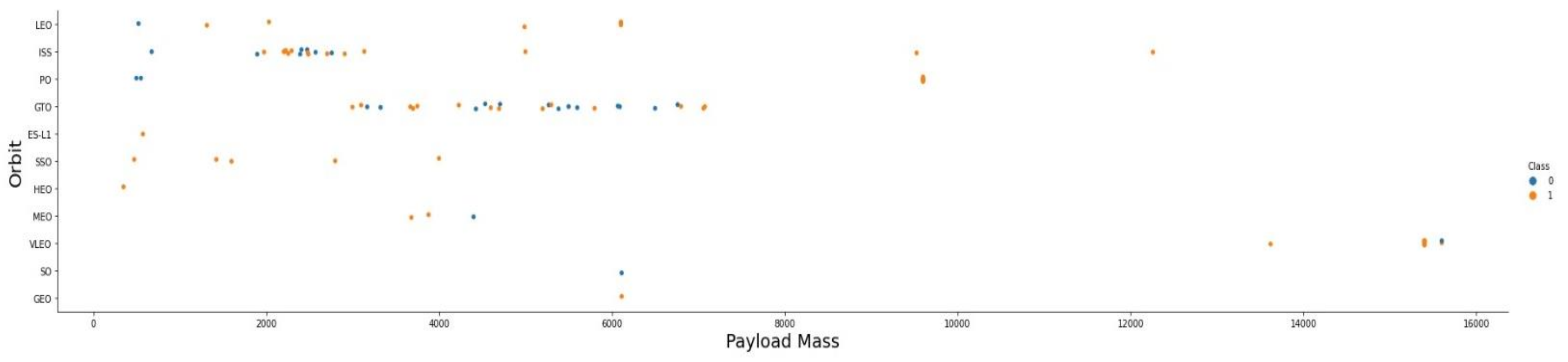
Flight Number vs. Orbit Type



From graph we can see that as Flight number increases success rate also increases in different orbits.

Orbits not close to Earth are tested after some flight.

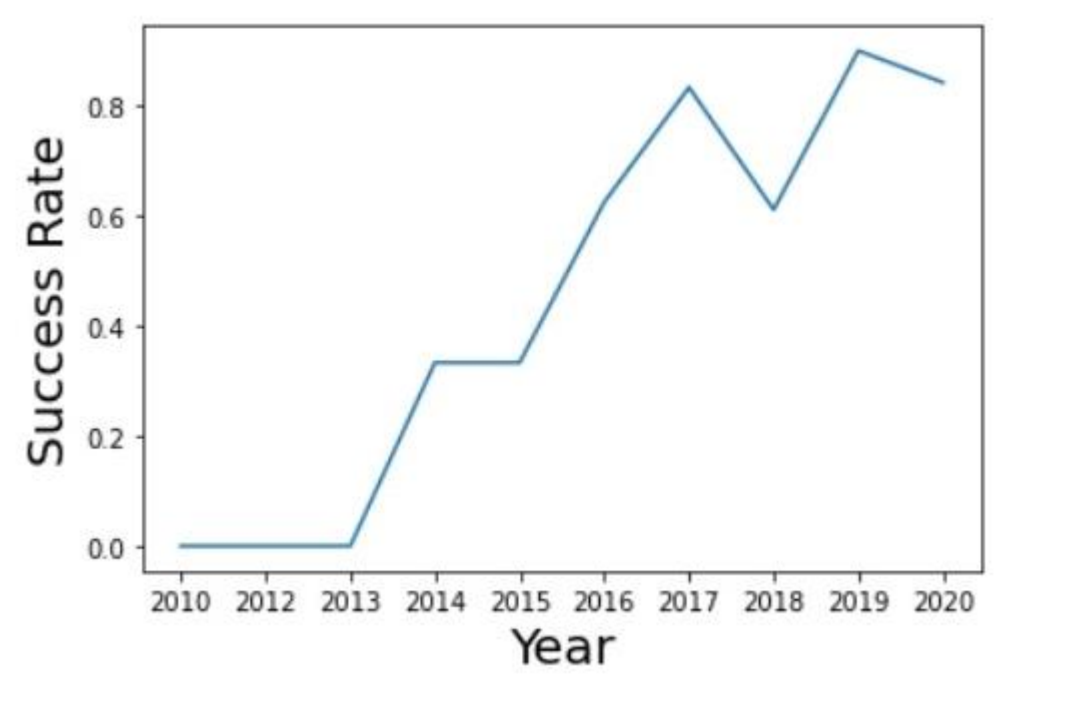
Payload vs. Orbit Type



As we can when payload mass increases, orbits for LEO and VLEO are taken place. In addition, failure is more likely to happen for bigger orbits with heavy payload mass.

Launch Success Yearly Trend

From the graph, we can observe that the success rate since 2013 kept increasing till 2020.

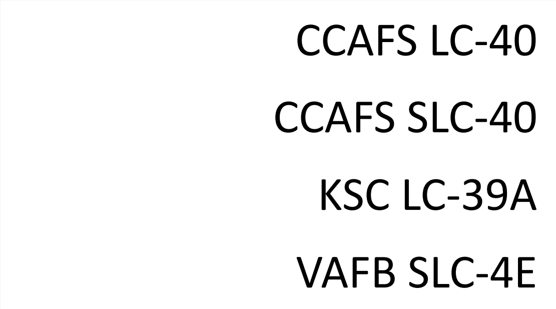


All Launch Site Names

The query command to find all launch site names from the database SPACEXDATASET:

```
select LAUNCH_SITE as SITES from SPACEXDATASET group by LAUNCH_SITE;
```

The results is the 4 different launch sites



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

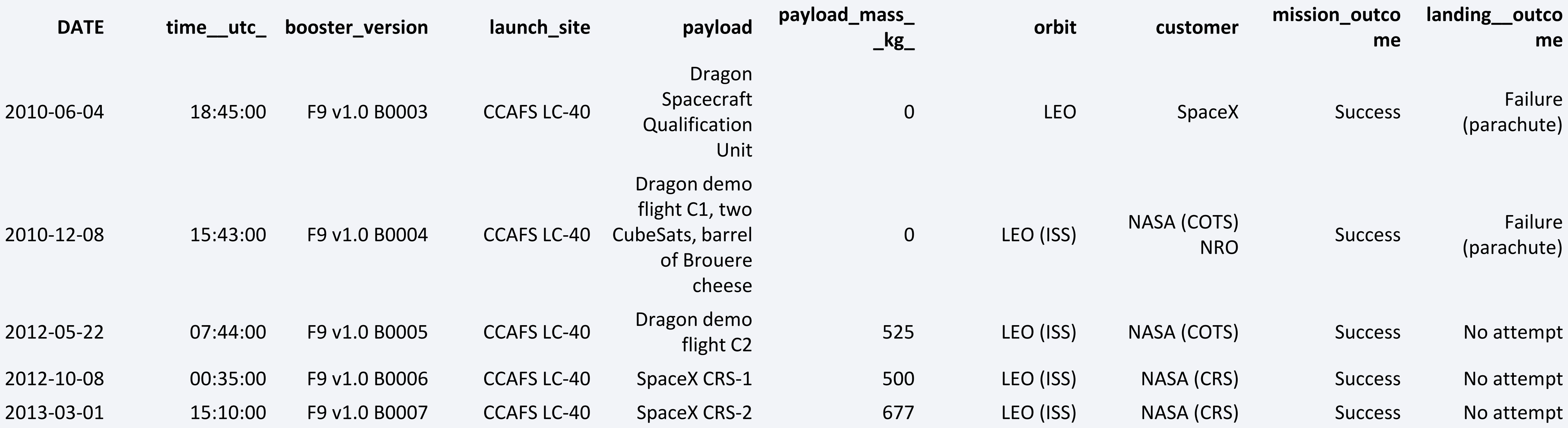
This dataset has CCAFS LC-40 launch site which is very close to CCAFS SLC-40 site and practical similar.

Launch Site Names Begin with 'CCA'

The query string to find 5 records where launch sites begin with `CCA`

```
select * from SPACEXDATASET where LAUNCH_SITE like 'CCA%' limit 5;
```

The result is referred to the first 5 launches that they have taken place.



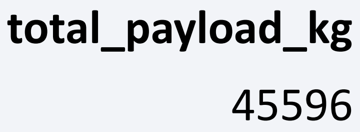
DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

Total Payload Mass

Total payload mass can be found from the query,

```
                select sum(payload_mass__kg_) as total_payload_kg from SPACEXDATASET where customer =
'NASA (CRS)’;
```

We filter the output for the customer NASA. The result is



total_payload_kg
45596

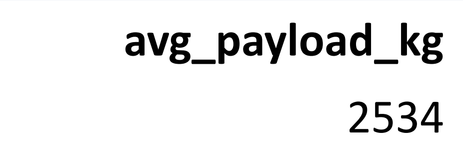
Number is in kilograms as depicted with the column name total_payload_kg.

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1 can be calculated from the query

```
            select avg(payload_mass__kg_) as avg_payload_kg from SPACEXDATASET where
booster_version like 'F9 v1.1%’;
```

The result is,



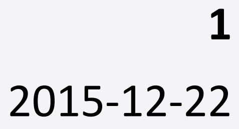
avg_payload_kg
2534

First Successful Ground Landing Date

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

```
select min(date) from SPACEXDATASET where landing__outcome like '%Success%’
```

The result from the query is,



```
         1
2015-12-22
```

And it is referred to the first successful landing outcome for Falcon 9.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
select booster_version, payload_mass__kg_ from SPACEXDATASET \

where landing__outcome = 'Success (drone ship)' and (payload_mass__kg_ >= 4000 and
payload_mass__kg_ <= 6000);
```

booster_version	payload_mass__kg_
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

The query string for the total number of successful and failure mission outcomes

```
select mission_outcome, count(mission_outcome) from SPACEXDATASET group by
mission_outcome;
```

The result of the above query is

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

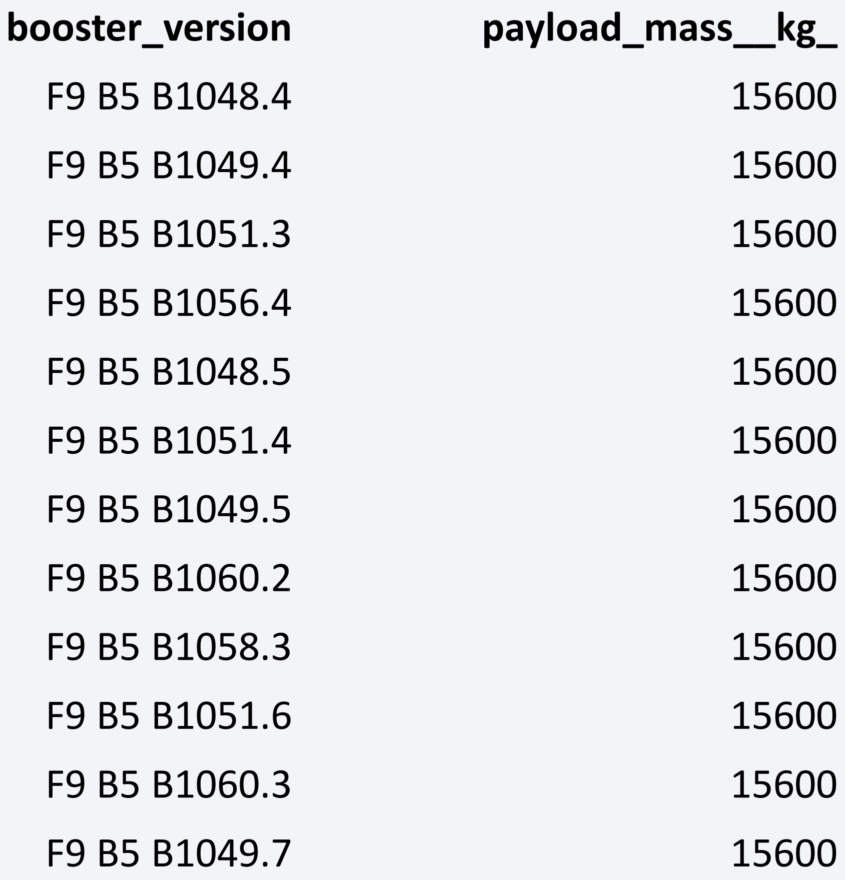
and it is referred to Falcon 9 with all booster’s versions.

Boosters Carried Maximum Payload

Listing the names of the booster which have carried the maximum payload mass can be done

```
            select booster_version,
payload_mass__kg_  from SPACEXDATASET where
payload_mass__kg_=(select
max(payload_mass__kg_) from SPACEXDATASET);
```

The result is shown in the image. The payload is referred to newly booster’s version of Falcon 9.



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

2015 Launch Records

The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
            select booster_version, launch_site, landing__outcome from SPACEXDATASET where landing__outcome =
'Failure (drone ship)' and year(date) = 2015;
```

Where the result is,

booster_version	launch_site	landing__outcome
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

and it shows the two cases with booster version and launch site where rockets failed to land in the drone ship.

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

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,

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

which give us the output in the table.

We can see that we filtered and sorted the result for the given date interval.

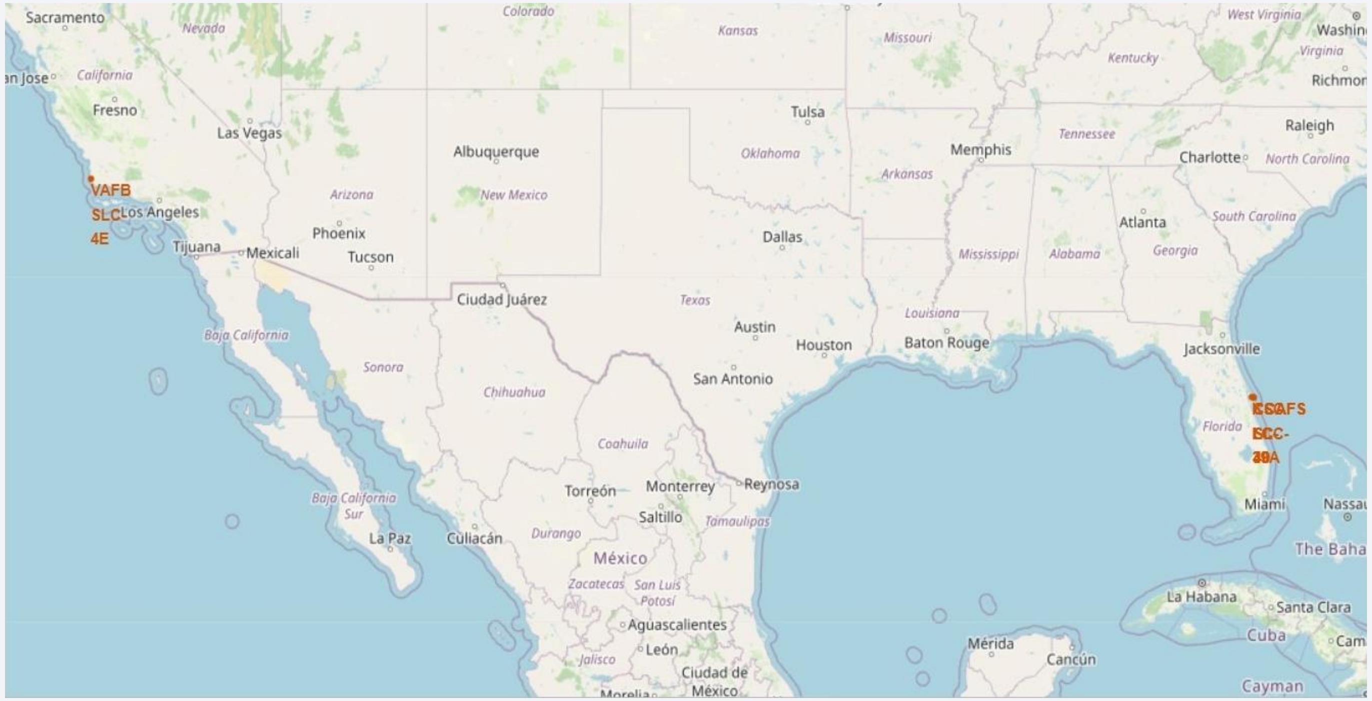
landing__outcome	total
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

Launch Site on the map using Folium

The screenshot which depicts all launch site. As we can see all the launch sites are near the ocean (Atlantic or Pacific) and it is as close as possible to Earth’s equator. This helps the launches.

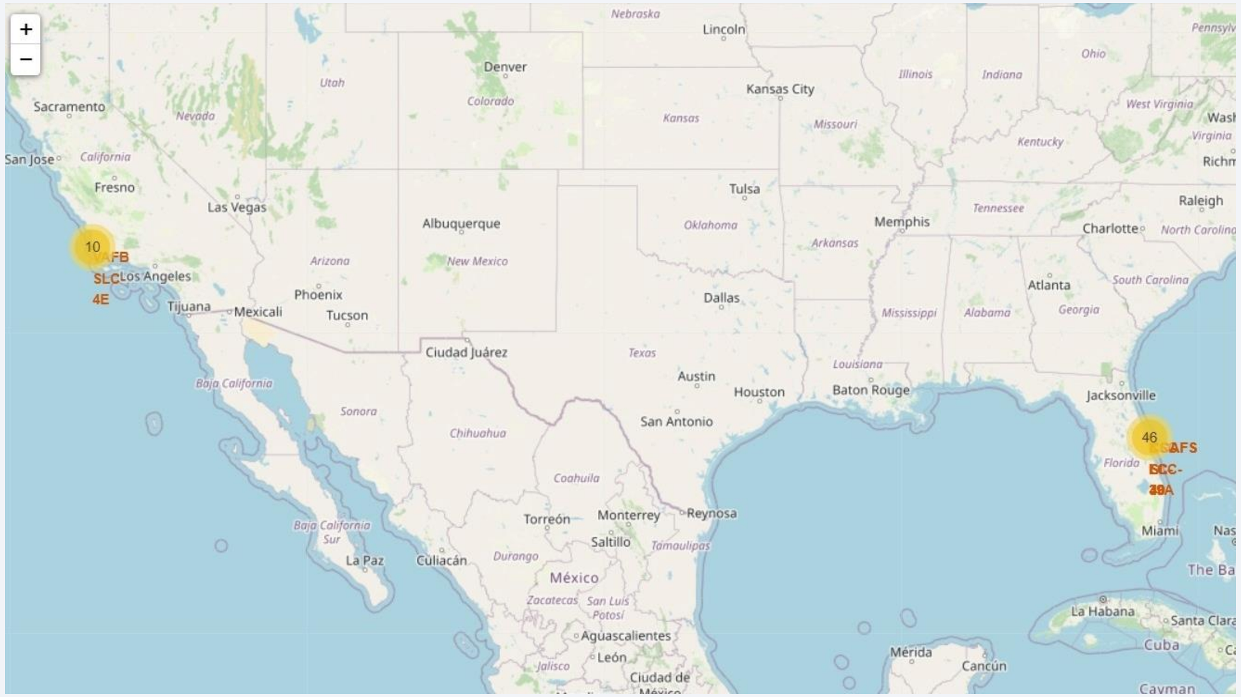


Number of launches for each site

This screenshot depicts the number of launches for each site.

We can see that most of the launches took places in the sites at Cape Canaveral.

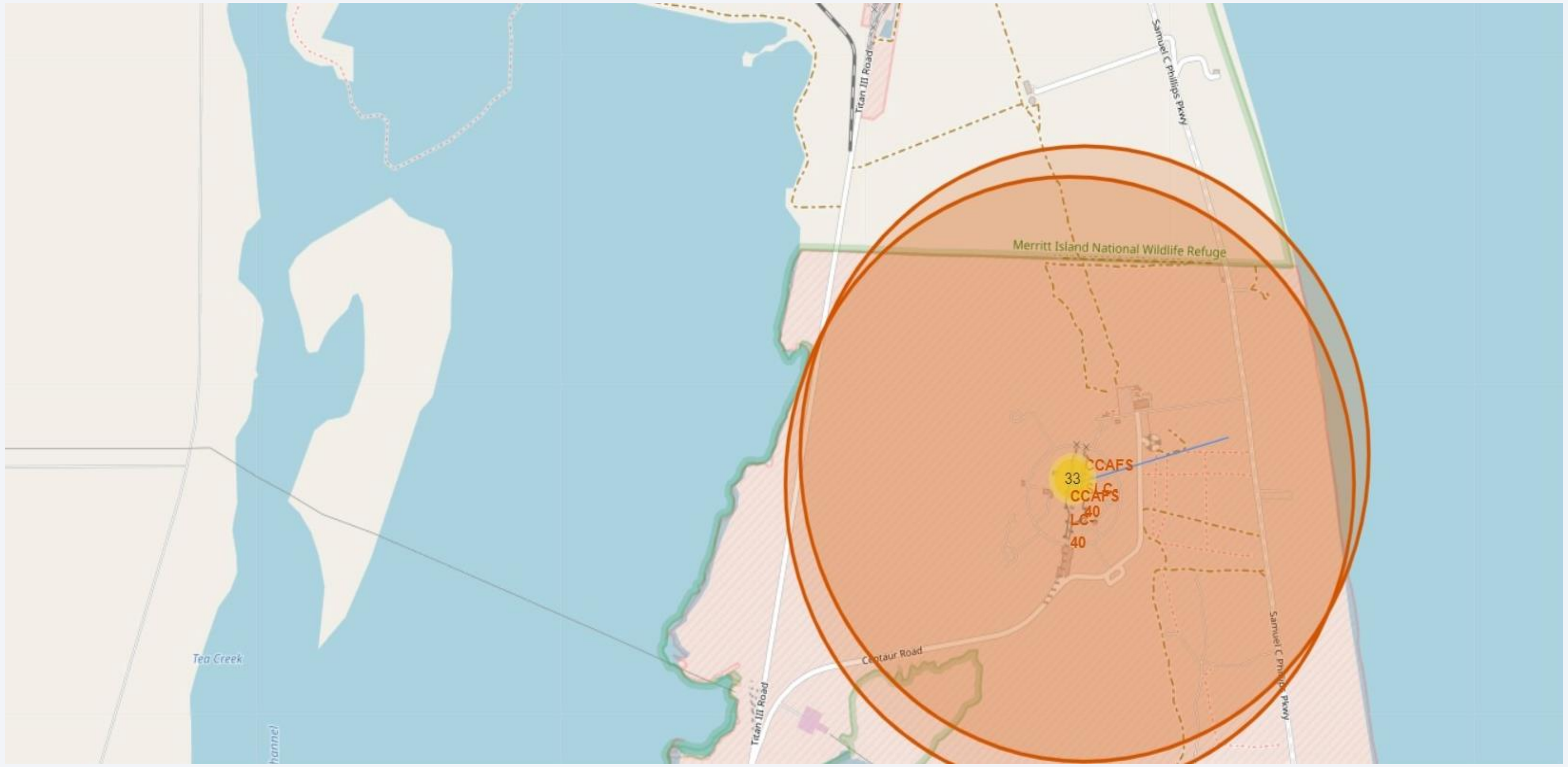
In addition, for the successful launches we added green color for circles and for failure launches we added red color on the folium map.



Calculating distance on Folium maps

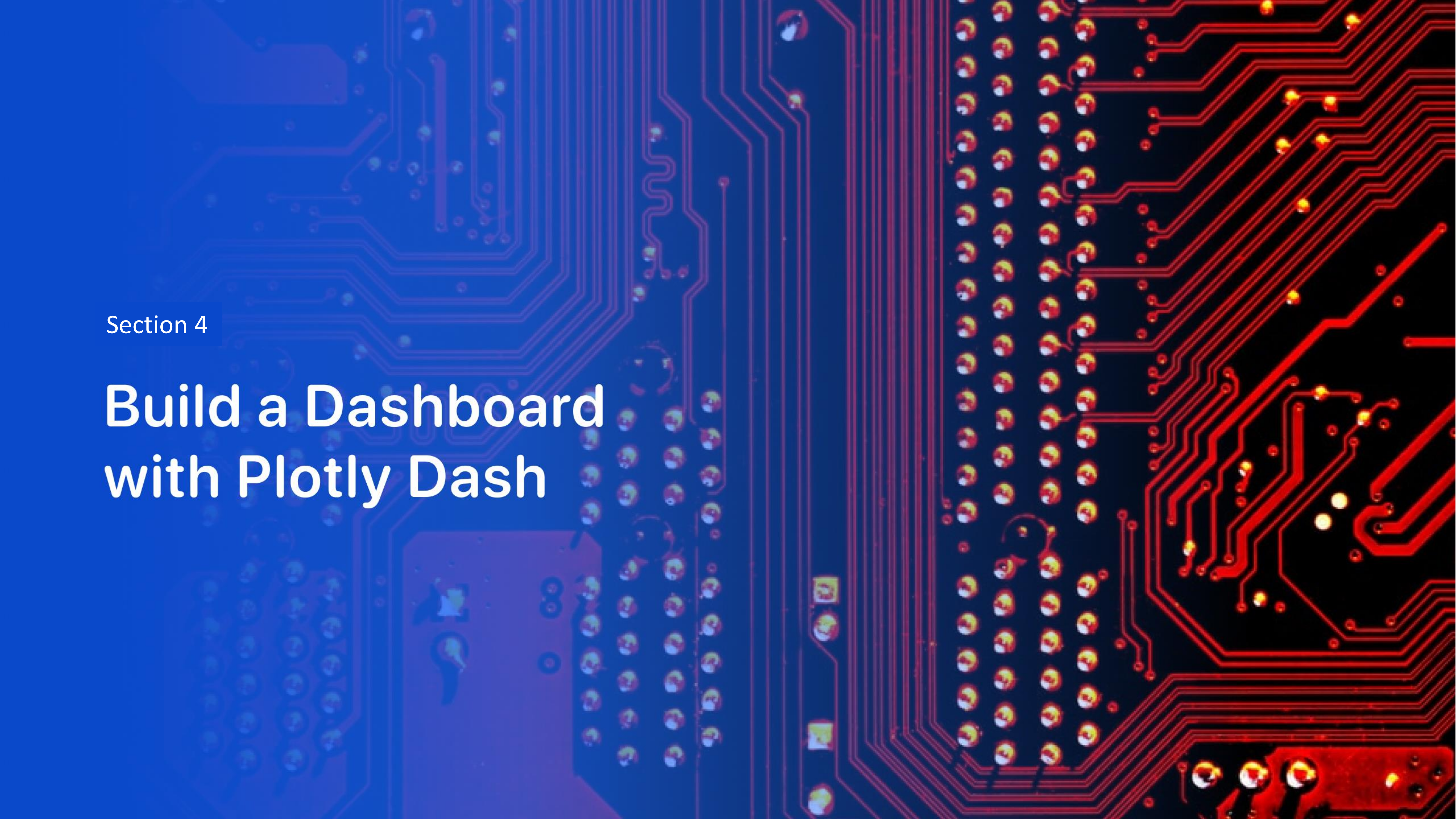
With this screenshot we show for the CCAFS LC-40 and a point near highway how we can calculate distance of two points on a Folium map.

This is very helpful because we can infer how the transportation of rockets is related with the launch site. Easily accessible launch sites are preferred for examples sites near railway, coastline or highway.



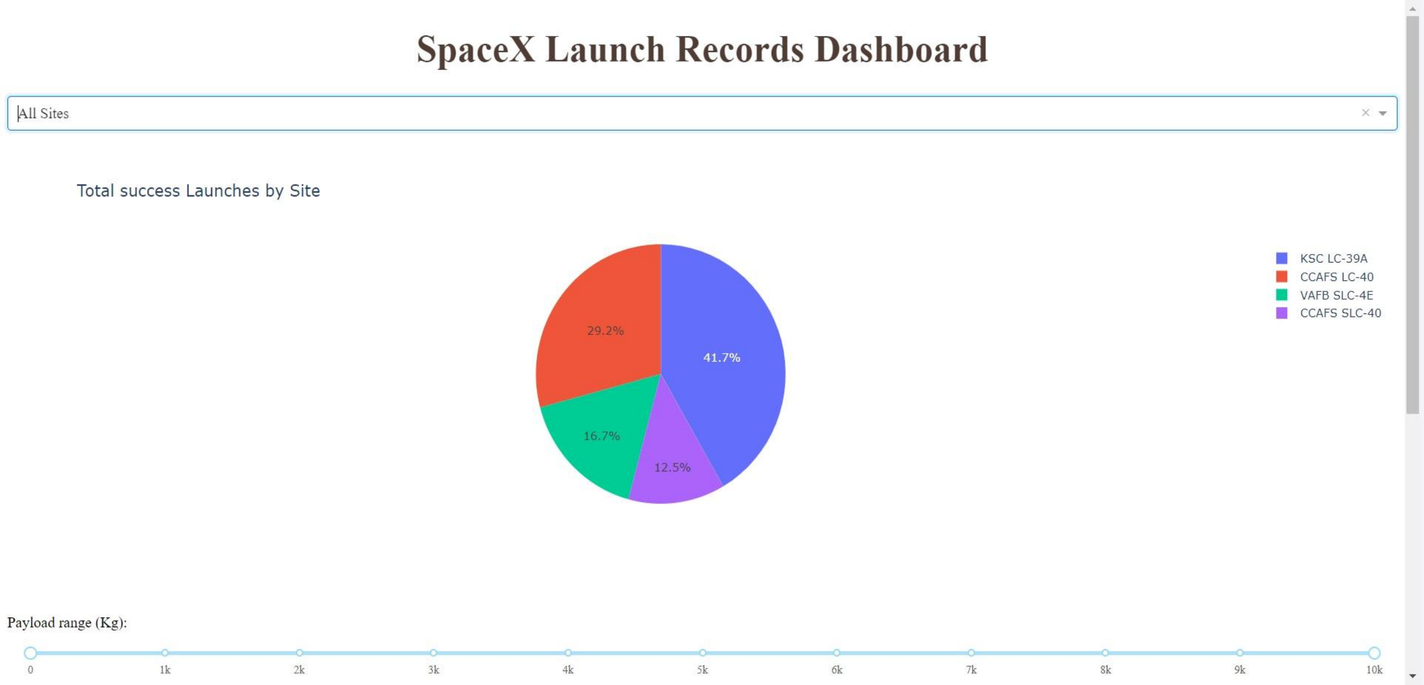
Section 4

Build a Dashboard with Plotly Dash



Successful Launches Pie Chart in Dashboard

- This screenshot shows total success for all launch sites.
- As we can see the KSC LC-39A has the most successes, followed by CCAFS lC-40.



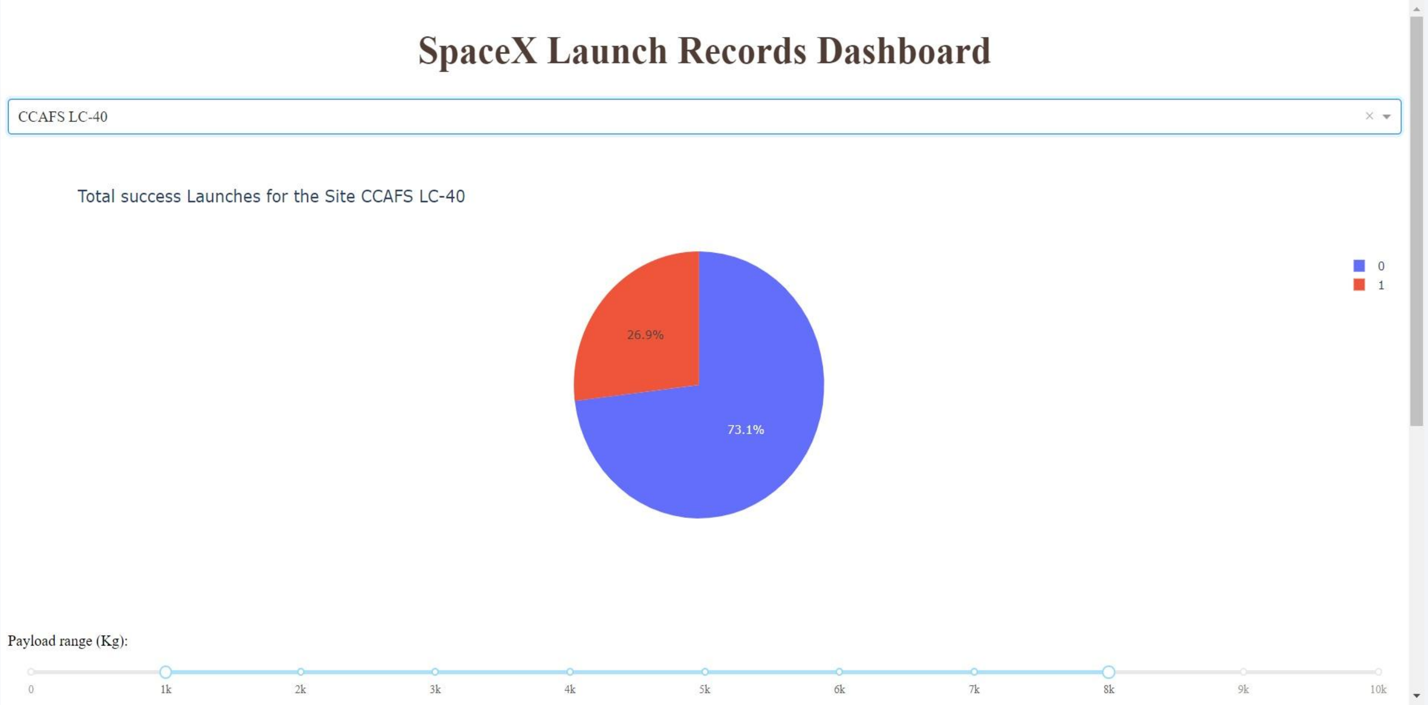
SpaceX Launch Records Dashboard

All Sites

Payload range (Kg):

Pie chart for specific launch site

- Pie chart for specific launch site. We can easily see the total success and failure rate.



SpaceX Launch Records Dashboard

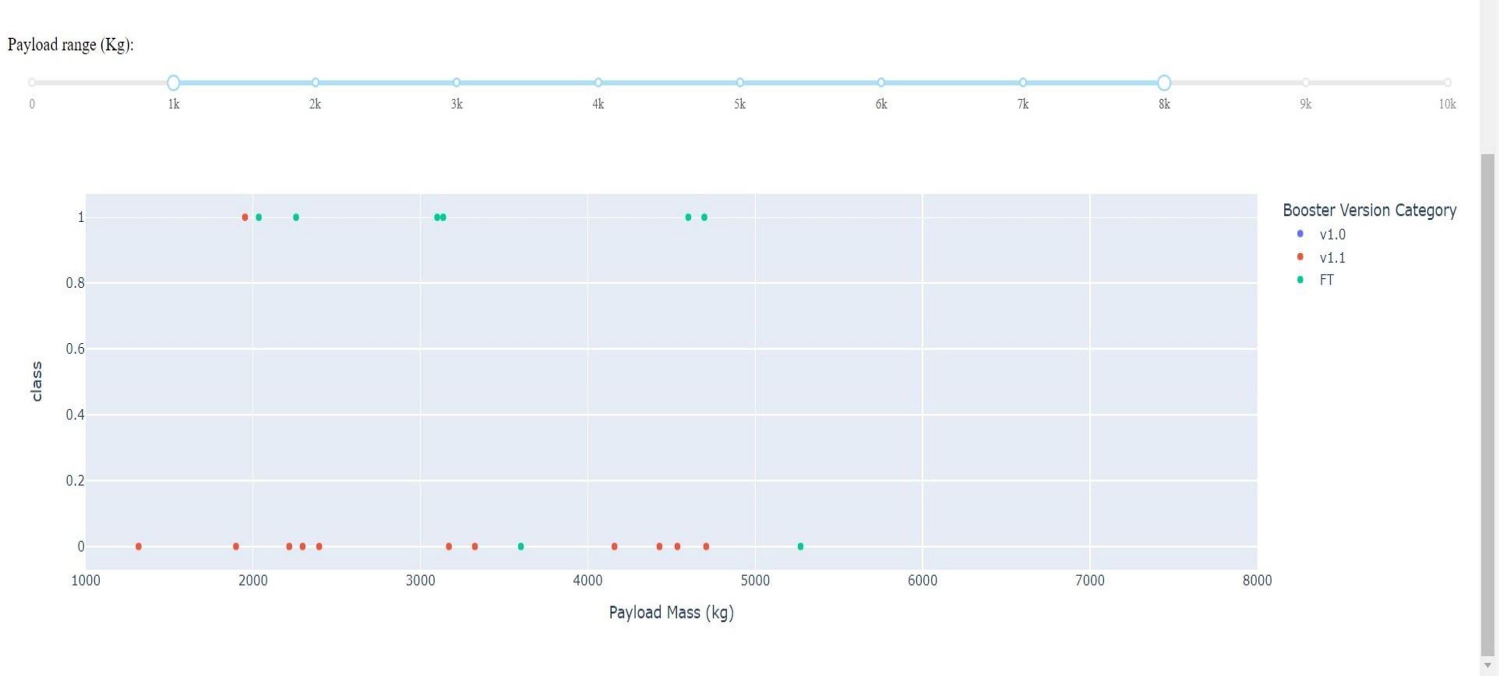
CCAFS LC-40

Total success Launches for the Site CCAFS LC-40

Payload range (Kg):

Scatter plot for payload mass

- This screenshot shows the success and failure rates for a chosen payload range.
- We can change payload range with slide ranger to get more insights about how payload is correlated with successful launch.

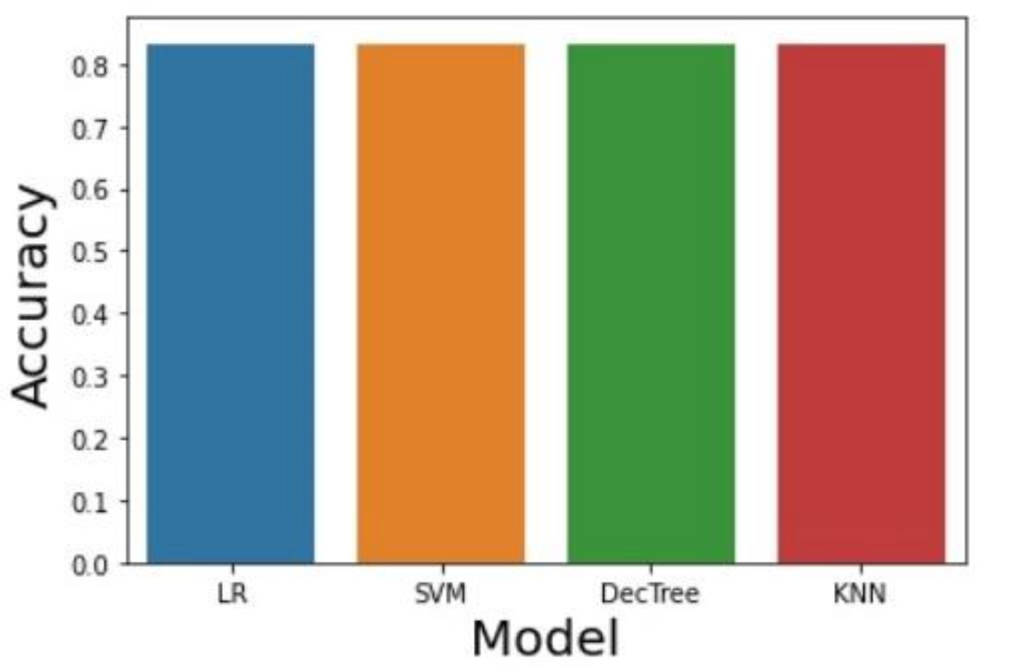


Section 5

Predictive Analysis (Classification)

Classification Accuracy

In the bar chart we see the accuracy for the models we trained with dataset. Models are Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DecTree) and K-Nearest Neighboors (KNN).



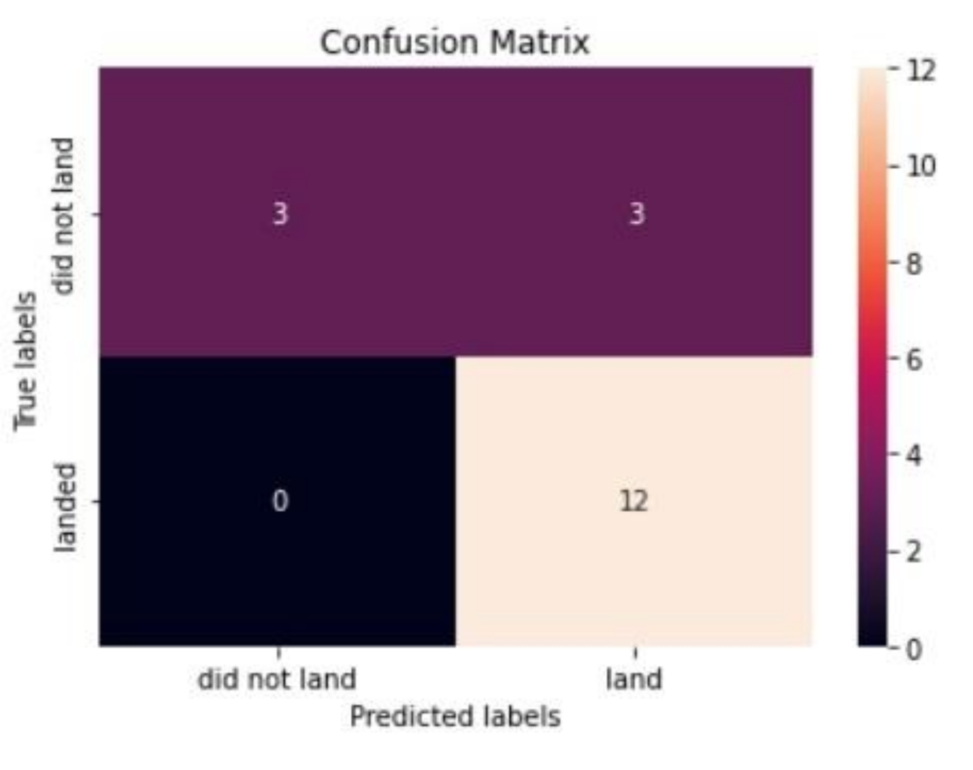
As we can see accuracy is practical the same for all models.

This is due to our dataset is small and models predict the same results.

Confusion Matrix

Confusion matrix is the same for all models.

As we can see all models predict rockets that land successfully, but they cannot predict rockets with failure outcome (they give true positive).



Conclusions

- Success rate increases with Flight number.
- Launch site has impact on success rate.
- Sites with the most launches are those that are close to railways, coastlines and highways.
- Success rate is correlated with launch site. Most successful launches and heavy payload rocket took place from sites near Earth's equator.
- Orbits that are not close to earth, and payload have impact to success rate.
- Good features for prediction are payload mass, orbit type and launch site.

Appendix

- SPACEX API: https://api.spacexdata.com/v4/launches/past
- GitHub URL with all Notebooks: https://github.com/SakisHous/spacex-data-science

Thank you!