IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Rahul Mitra
04.11.2022

Outline

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

Executive Summary

- Summary of methodologies
    - Data collection
    - Data wrangling
    - EDA with data visualization
    - EDA with SQL
    - Building an interactive map with Folium
    - Building a Dashboard with Plotly Dash
    - Predictive analysis (Classification)
- Summary of all results
    - EDA results
    - Interactive analytics
    - Predictive analysis

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.
- Problems you want to find answers
  - The project task is to predict if the first stage of the SpaceX Falcon 9 rocket will land successfully or not.

Section 1

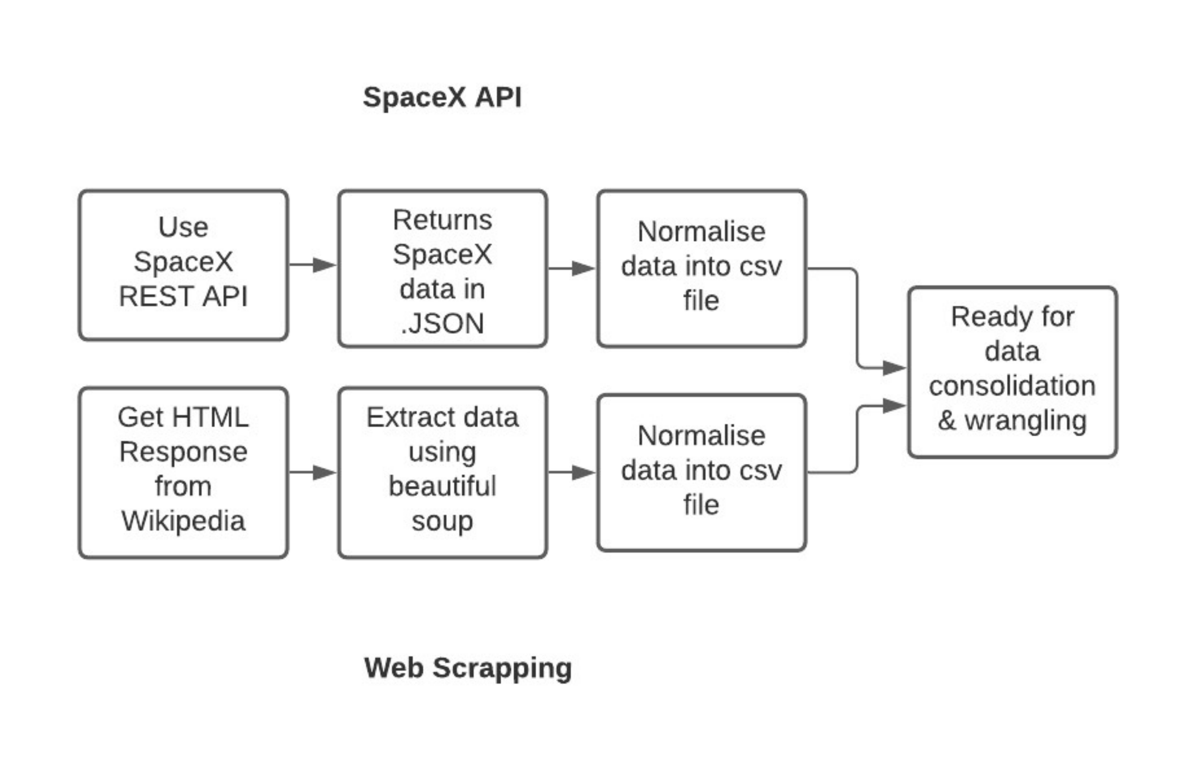
Methodology

Methodology

Executive Summary

- **Data collection methodology:**
  - SpaceX Rest API
  - Web Scrapping from Wikipedia
- **Perform data wrangling**
  - One Hot Encoding data fields for Machine Learning and data cleaning of null values and irrelevant columns
- **Perform exploratory data analysis (EDA) using visualization and SQL**
- **Perform interactive visual analytics using Folium and Plotly Dash**
- **Perform predictive analysis using classification models**
  - Logistic Regression, KNN, SVM, Decision Tree models have been built and evaluated for the best classifier

Data Collection



- Data was collected from the following dataset:
  - SpaceX launch data gathered from the SpaceX REST API.
  - Data about launches includes information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.
  - BeautifulSoup can be used for web scraping Wikipedia to gather Falcon 9 Launch data.

Data Collection - SpaceX API

- Data collection with SpaceX REST calls

[Link to notebook]

1 .Getting Response from API

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
response = requests.get(spacex_url).json()
```

2. Converting Response to a .json file

```
response = requests.get(static_json_url).json()
data = pd.json_normalize(response)
```

3. Apply custom functions to clean data

```
getLaunchSite(data)
getPayloadData(data)
getCoreData(data)
```

```
getBoosterVersion(data)
```

4. Assign list to dictionary then dataframe

```
launch_dict = {'FlightNumber': list(data['flight_number']),
'Date': list(data['date']),
'BoosterVersion':BoosterVersion,
'PayloadMass':PayloadMass,
'Orbit':Orbit,
'LaunchSite':LaunchSite,
'Outcome':Outcome,
'Flights':Flights,
'GridFins':GridFins,
'Reused':Reused,
'Legs':Legs,
'LandingPad':LandingPad,
'Block':Block,
'ReusedCount':ReusedCount,
'Serial':Serial,
'Longitude': Longitude,
'Latitude': Latitude}
```

```
df = pd.DataFrame.from_dict(launch_dict)
```

5. Filter dataframe and export to flat file (.csv)

```
data_falcon9 = df.loc[df['BoosterVersion']!="Falcon 1"]
```

```
data_falcon9.to_csv('dataset_part_1.csv', index=False)
```

Data Collection - Web Scraping

- Web Scrapping from Wikipedia

[Link to notebook]

1 .Getting Response from HTML

```
page = requests.get(static_url)
```

2. Creating BeautifulSoup Object

```
soup = BeautifulSoup(page.text, 'html.parser')
```

3. Finding tables

```
html_tables = soup.find_all('table')
```

4. Getting column names

```
column_names = []
temp = soup.find_all('th')
for x in range(len(temp)):
    try:
     name = extract_column_from_header(temp[x
     if (name is not None and len(name) > 0):
        column_names.append(name)
    except:
     pass
```

5. Creation of dictionary

```
launch_dict= dict.fromkeys(column_names)

# Remove an irrelvant column
del launch_dict['Date and time ( )']

launch_dict['Flight No.'] = []
launch_dict['Launch site'] = []
launch_dict['Payload'] = []
launch_dict['Payload mass'] = []
launch_dict['Orbit'] = []
launch_dict['Customer'] = []
launch_dict['Launch outcome'] = []
launch_dict['Version Booster']=[]
launch_dict['Booster landing']=[]
launch_dict['Date']=[]
launch_dict['Time']=[]
```

6. Appending data to keys (refer) to notebook block 12

```
In [12]: extracted_row = 0
         #Extract each table
         for table_number,table in enumer
            # get table row
             for rows in table.find_all("
                   #check to see if first t
```

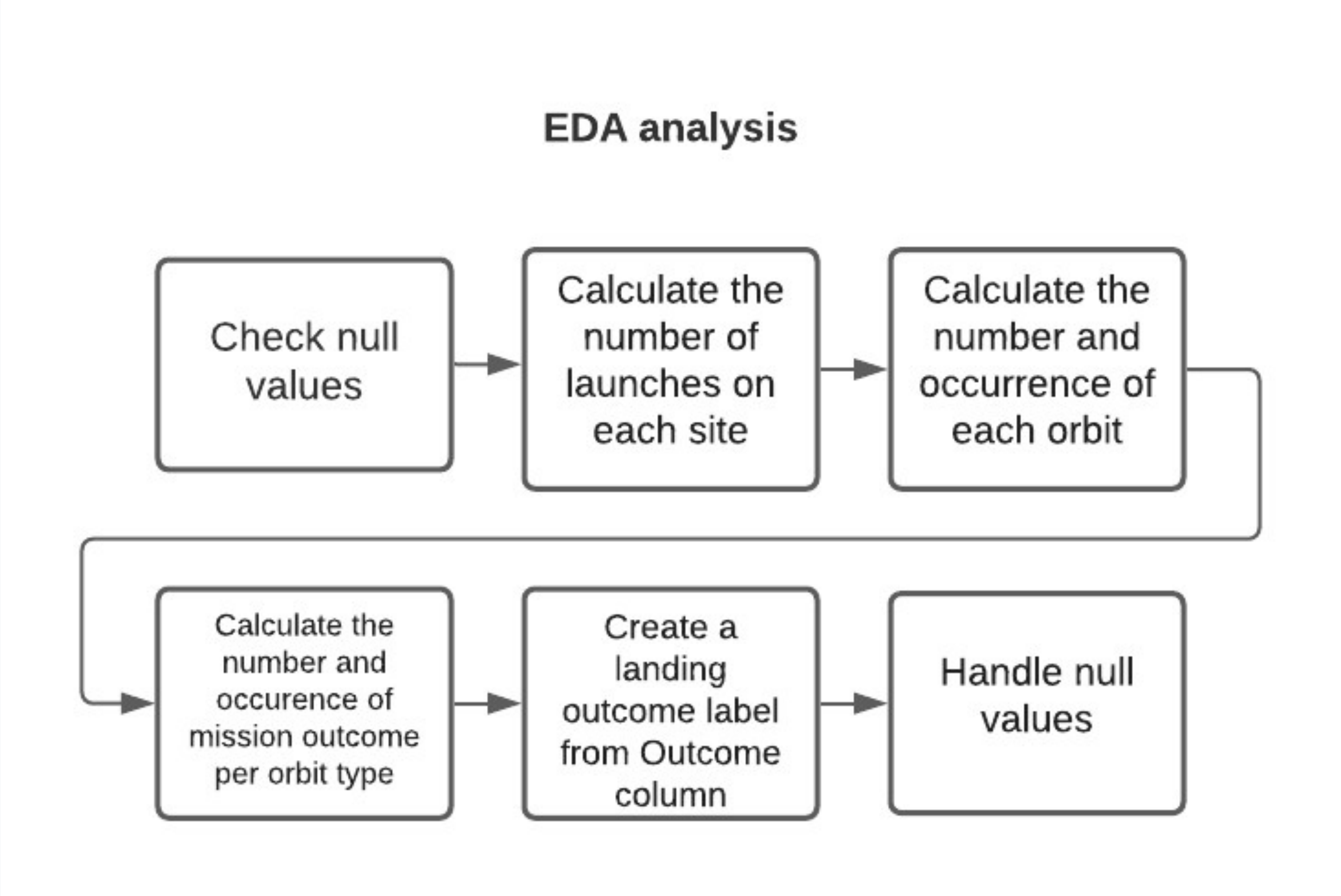
7. Converting dictionary to dataframe

```
df = pd.DataFrame.from_dict(launch_dict)
```

8. Dataframe to .CSV

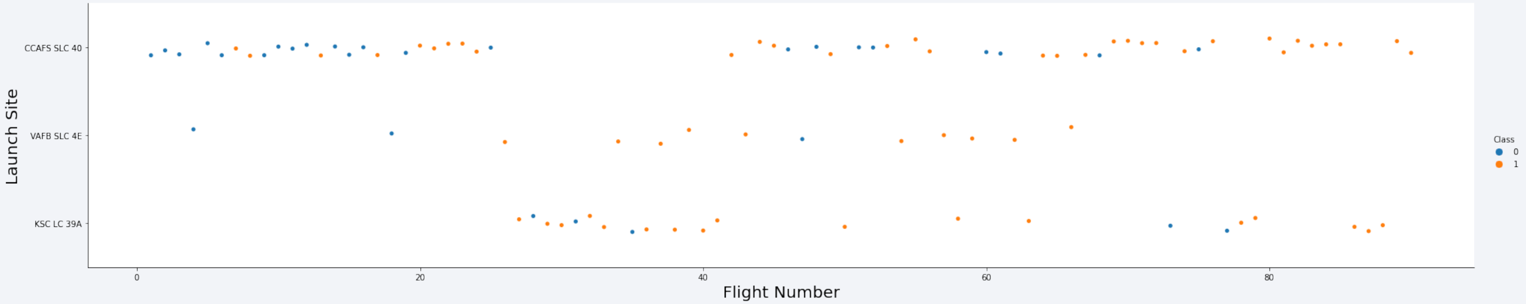
```
df.to_csv('spacex_web_scraped.csv', index=False)
```

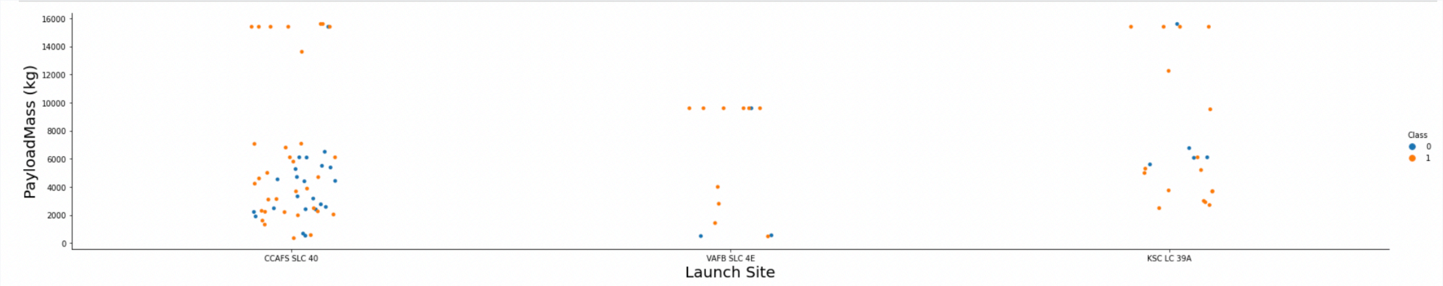
Data Wrangling

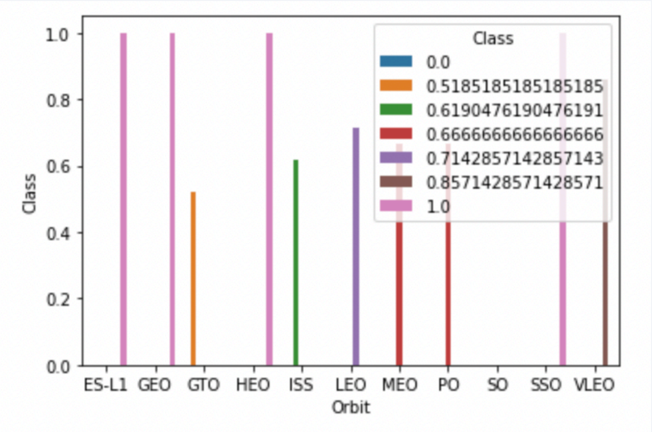


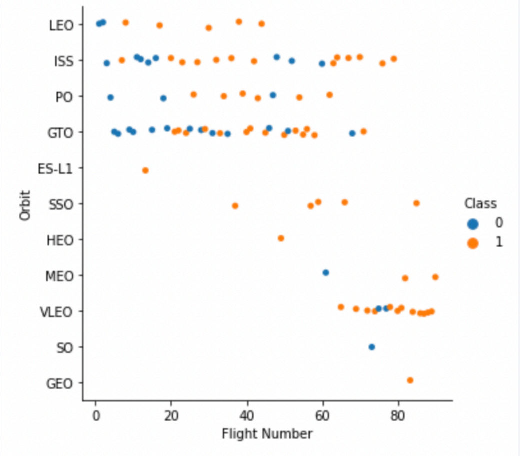
[Link to notebook]

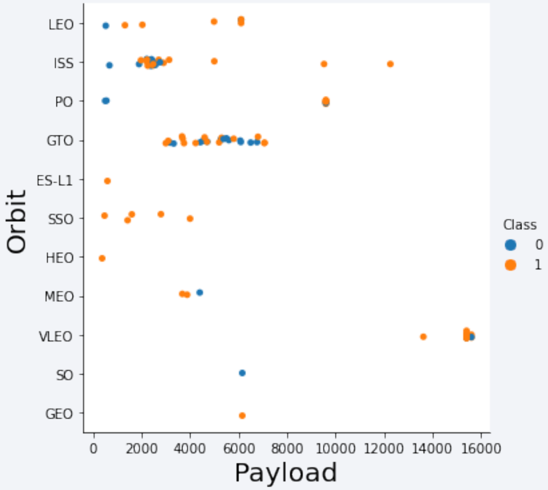
EDA with Data Visualisation

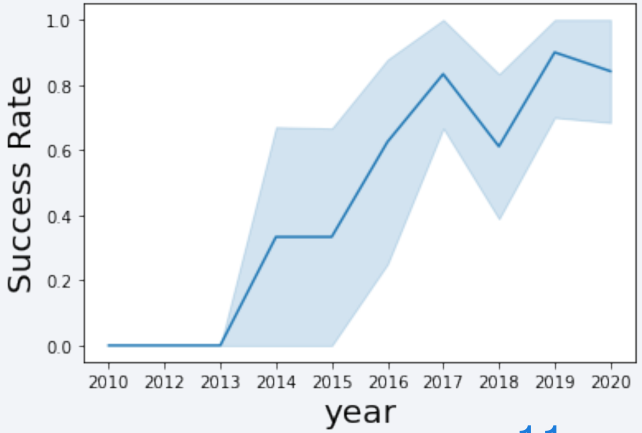








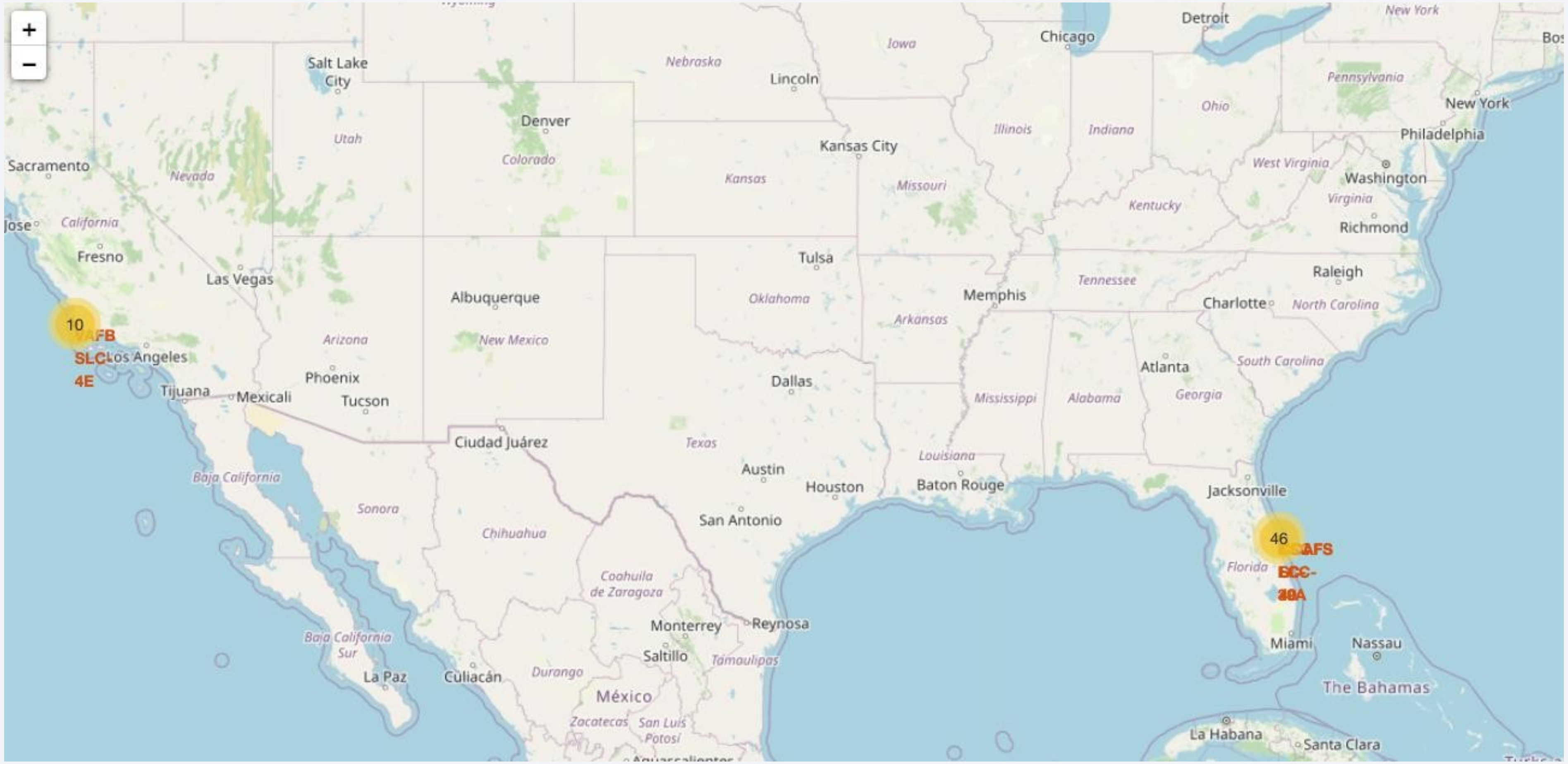




EDA with SQL

- Tasks performed:
  - Displayed the names of the unique launch sites in the space mission
  - Displayed 5 records where launch sites begin with the string 'KSC'
  - Displayed the total payload mass carried by boosters launched by NASA (CRS)
  - Displayed average payload mass carried by booster version F9 v1.1
  - Listed the date where the successful landing outcome in drone ship was achieved.
  - Listed the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
  - Listed the total number of successful and failure mission outcomes
  - Listed the names of the booster_versions which have carried the maximum payload mass.
  - Listed the records which will display the month names, successful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017
  - Ranked the count of successful landing_outcomes between the date 2010/06/04 and 2017/03/20 in descending order.

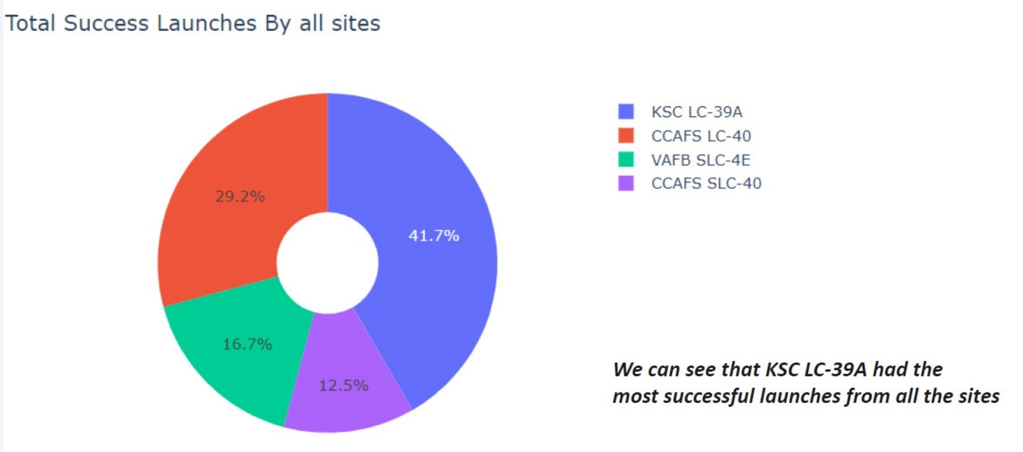
Build an Interactive Map with Folium



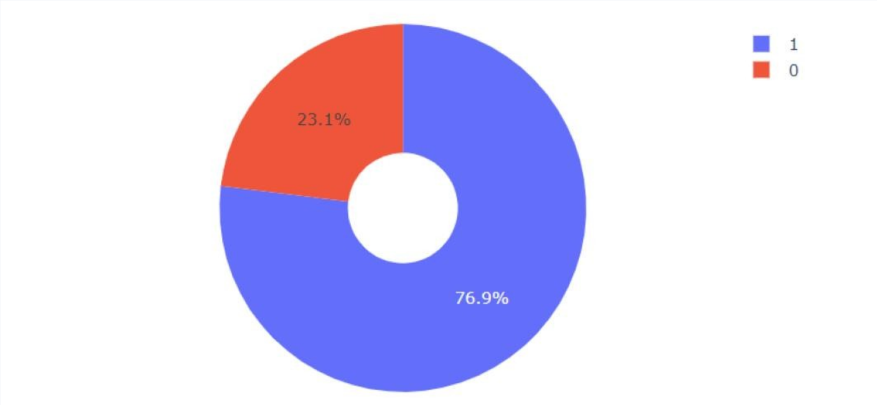
Map markers have been added to the map with aim to finding an optimal location for building a launch site

Link to notebook

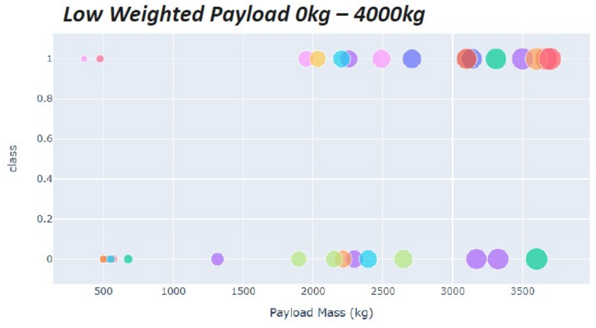
Build a Dashboard with Plotly Dash

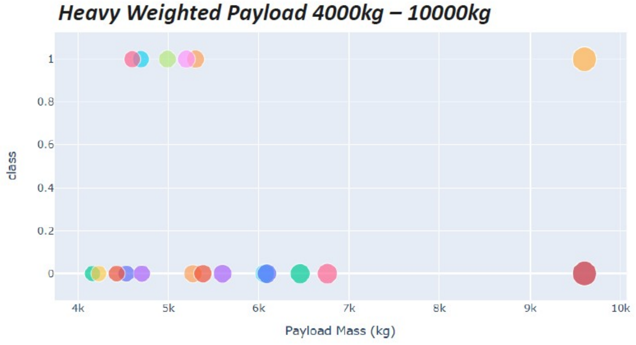


We can see that KSC LC-39A had the most successful launches from all the sites



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





We can see the success rates for low weighted payloads is higher than the heavy weighted payloads

[Link to notebook]

Predictive Analysis (Classification)

- The SVM, KNN, and Logistic Regression model achieved the highest accuracy at 94.4%.

Link to notebook

Results

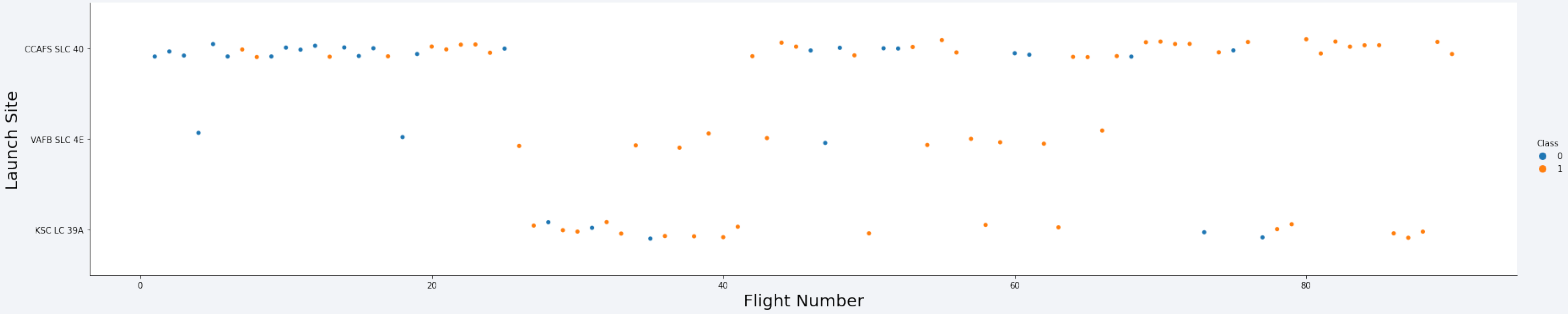
- KSC LC 39A had the most successful launches from all the sites.
- Orbits which had the highest success rate were GEO, HEO, SSO, ES L1.
- The SVM, KNN, and Logistic Regression models are the best in terms of prediction accuracy for this dataset, but not Decision Trees.
- Low weighted payloads perform better than the heavier ones.
- The success rates for SpaceX launches increased with the years since 2013.

Section 2

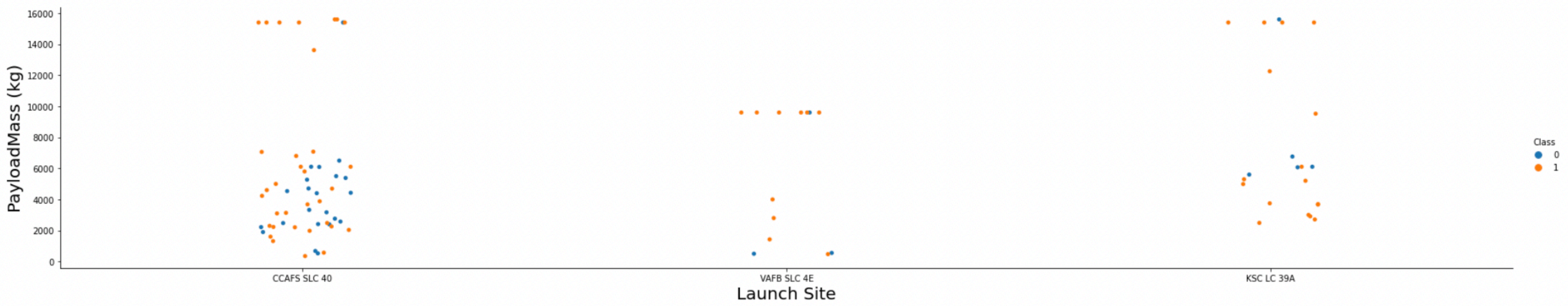
Insights drawn from EDA

Flight Number vs. Launch Site

- Launches from the site of CCAFS SLC 40 are much higher than launches form other sites.

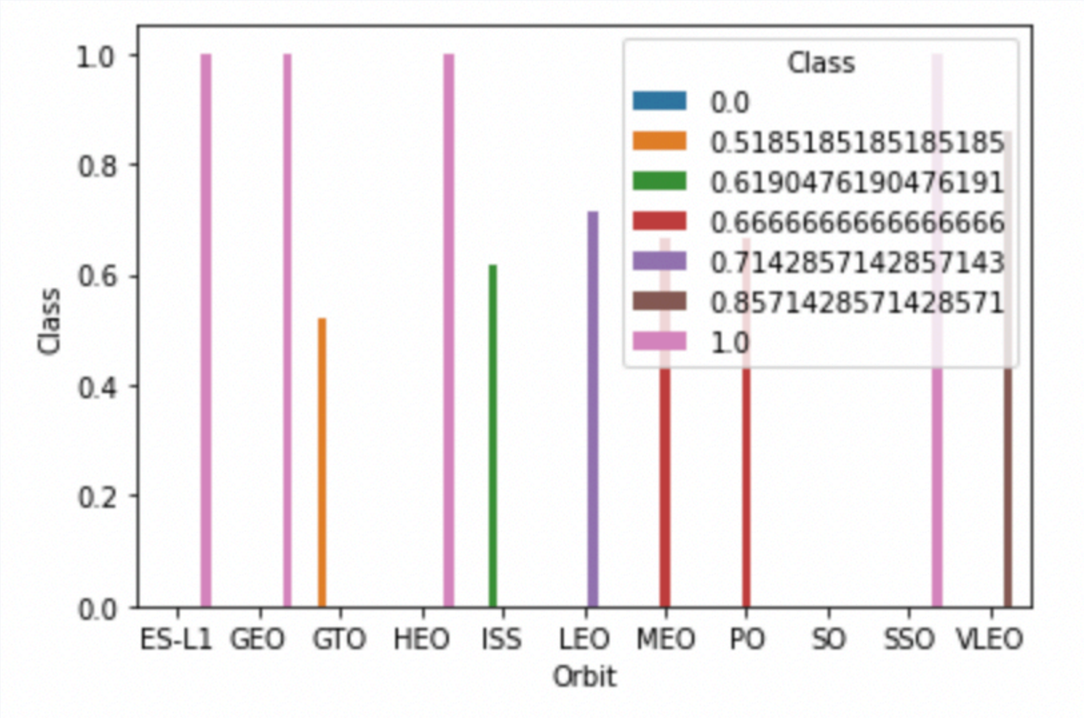


Payload vs. Launch Site



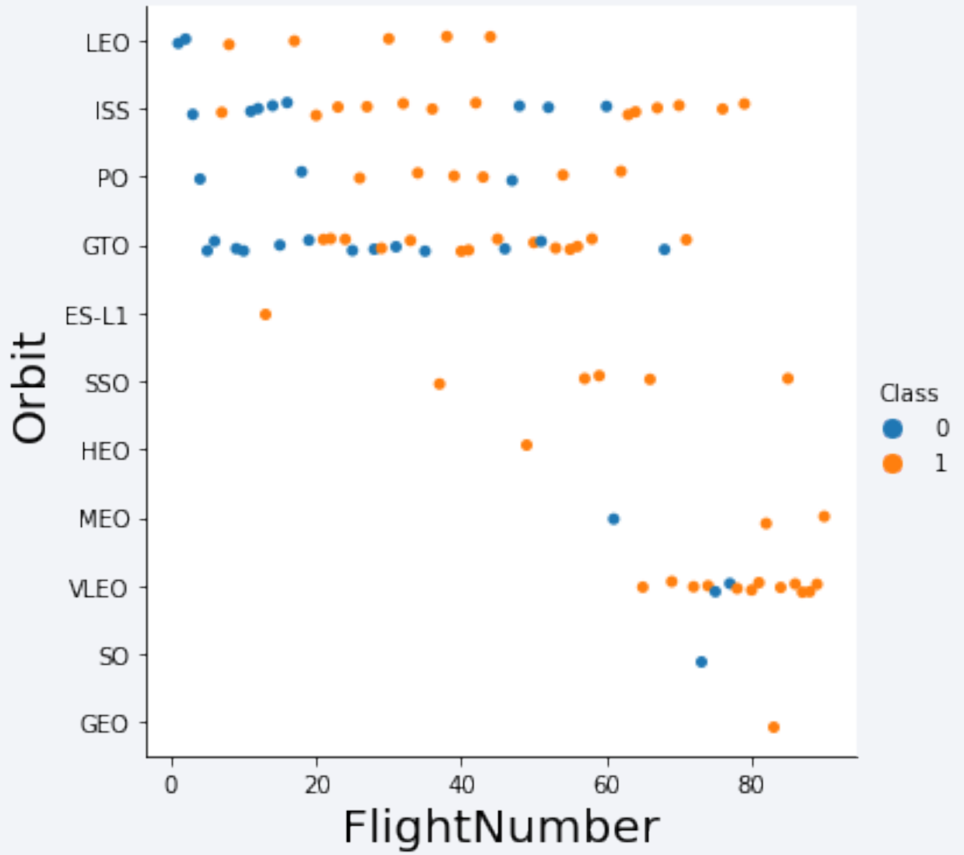
- The majority of launches with lower mass have been from CCAFS SLC 40.
- For VAFB, there is no launch above 10000 kg.

Success Rate vs. Orbit Type



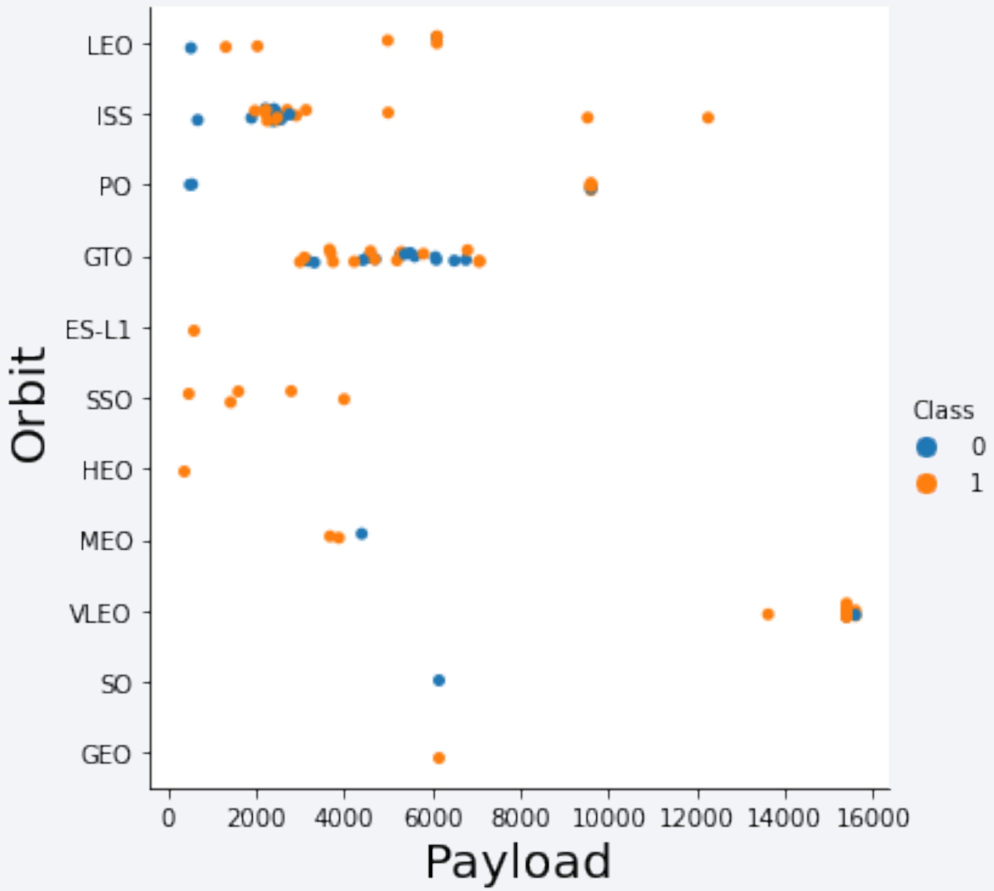
- The orbit types of ES-L1, GEO, HEO, SSO have one the highest success rates.

Flight Number vs. Orbit Type



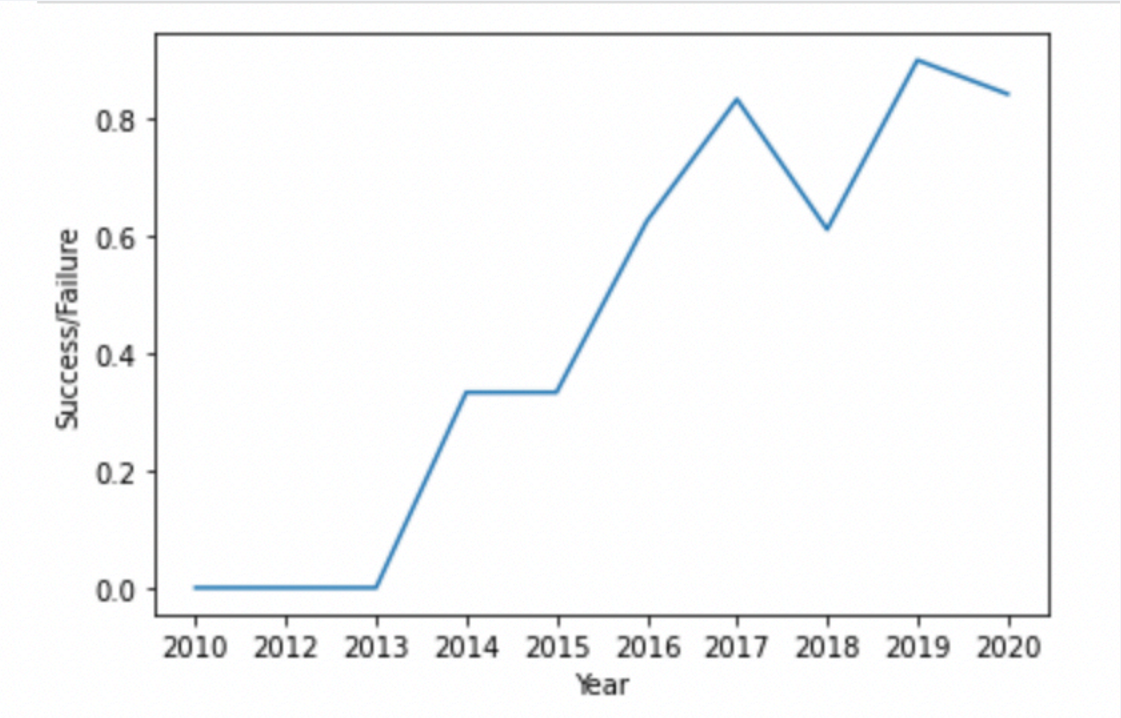
- For LEO, the later launches were successful.
- Towards the end, more launches can be seen from VLEO.

Payload vs. Orbit Type



- A strong correlation can be seen between GTO and Payload at the range around 4000 - 7500, as well as between ISS and Payload at 2000.

Launch Success Yearly Trend



- Launch success rate has increased significantly since 2013, potentially due to advancement in the field of technology.

All Launch Site Names

- %sql select distinct(LAUNCH_SITE) from spacex

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

Launch Site Names Begin with 'CCA'

- %sql select * from spacex where LAUNCH_SITE like 'CCA%' limit 5

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

- %sql select sum(PAYLOAD_MASS__KG_) from spacex where CUSTOMER = 'NASA (CRS)'

45596

Average Payload Mass by F9 v1.1

- %sql select avg(PAYLOAD_MASS__KG_) from spacex where BOOSTER_VERSION = 'F9 v1.1'

2928.400000

First Successful Ground Landing Date

- %sql select min(DATE) from spacex where Landing__Outcome = 'Success (ground pad)'

2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- %sql select BOOSTER_VERSION from spacex where Landing__Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000

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

Total Number of Successful and Failure Mission Outcomes

- %sql select count(MISSION_OUTCOME) from spacex where MISSION_OUTCOME = 'Success' or MISSION_OUTCOME = 'Failure (in flight)'

100

Boosters Carried Maximum Payload

- %sql select BOOSTER_VERSION from spacex where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)

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

2015 Launch Records

- %sql select * from spacex where Landing__Outcome like 'Success%' and (DATE between ‘2015-01-01' and ‘2015-12-31') order by date desc

time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
17:54:00	F9 FT B1029.1	VAFB SLC-4E	Iridium NEXT 1	9600	Polar LEO	Iridium Communications	Success	Success (drone ship)
05:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
04:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
21:39:00	F9 FT B1023.1	CCAFS LC-40	Thaicom 8	3100	GTO	Thaicom	Success	Success (drone ship)
[illegible]	[illegible]	CCAFS LC-	[illegible]	[illegible]	[illegible]	SKY Perfect JSAT	[illegible]	[illegible]

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

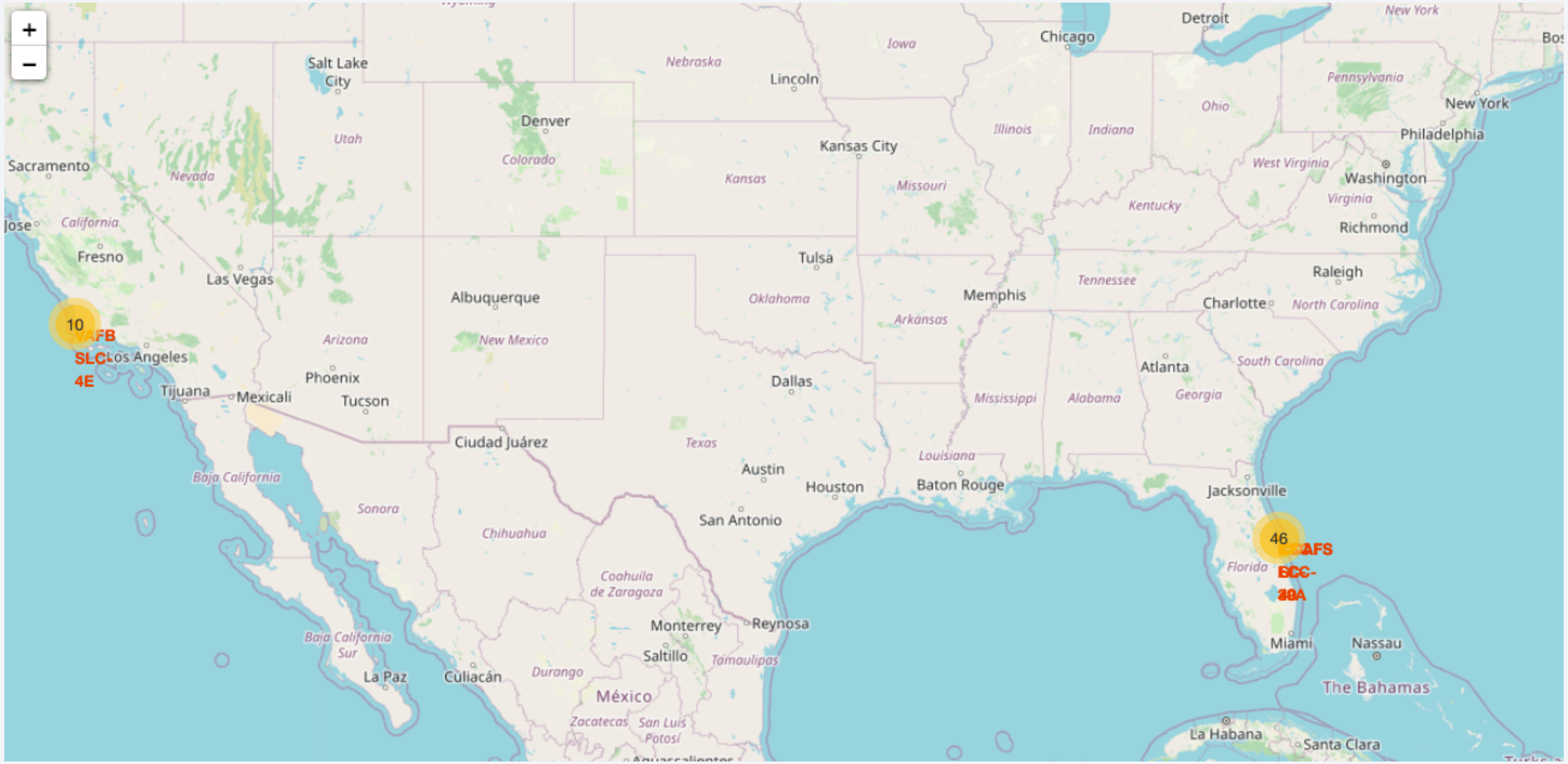
- %sql select * from spacex where Landing__Outcome like 'Success%' and (DATE between '2010-06-04' and '2017-03-20') order by date desc

2016-05-27	21:39:00	F9 FT B1023.1	CCAFS LC-40	Thaicom 8	3100	GTO	Thaicom	Success	Success (drone ship)
2016-05-06	05:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-04-08	20:43:00	F9 FT B1021.1	CCAFS LC-40	SpaceX CRS-8	3136	LEO (ISS)	NASA (CRS)	Success	Success (drone ship)
2015-12-22	01:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

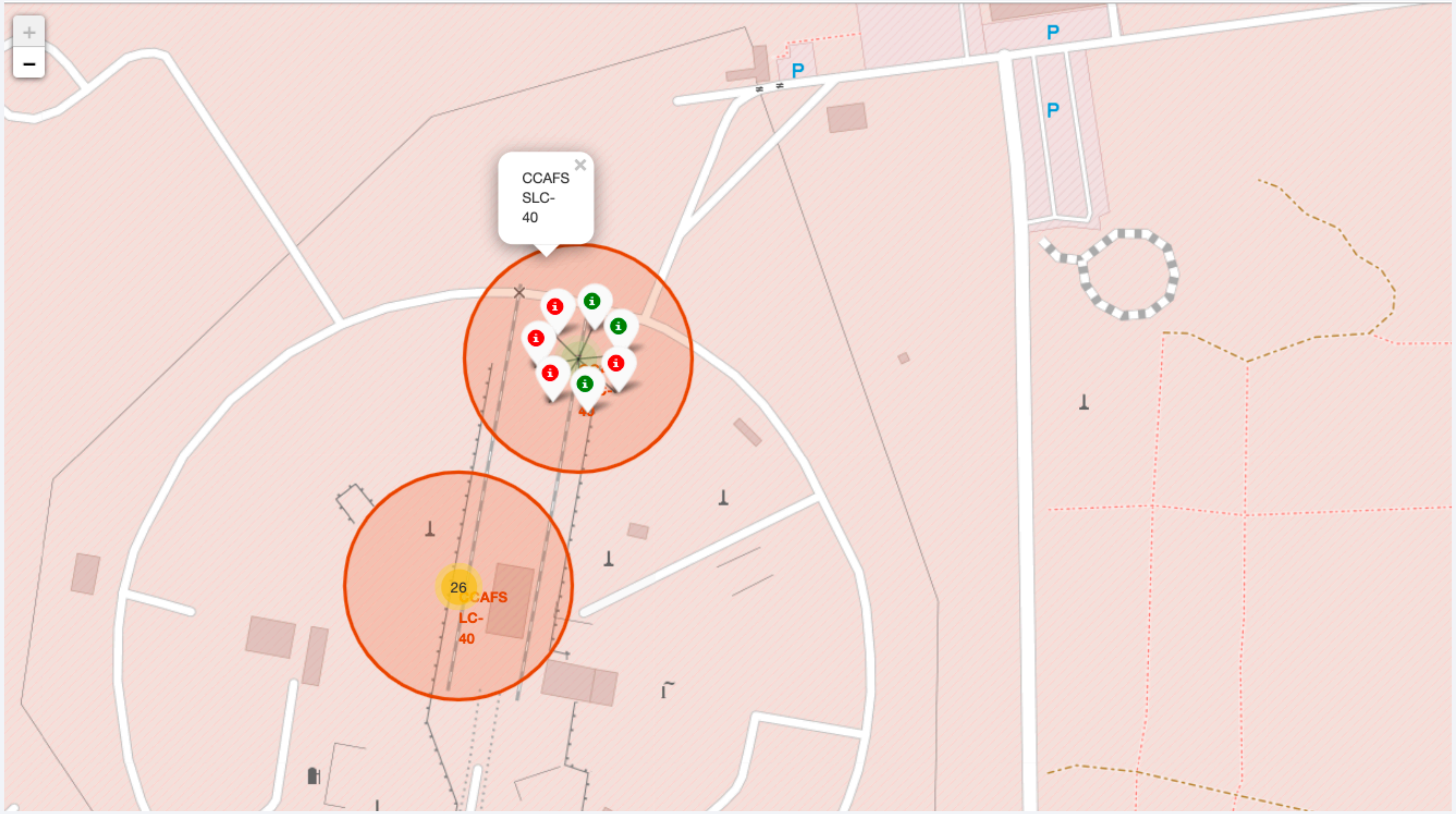
Section 4

Launch Sites Proximities Analysis

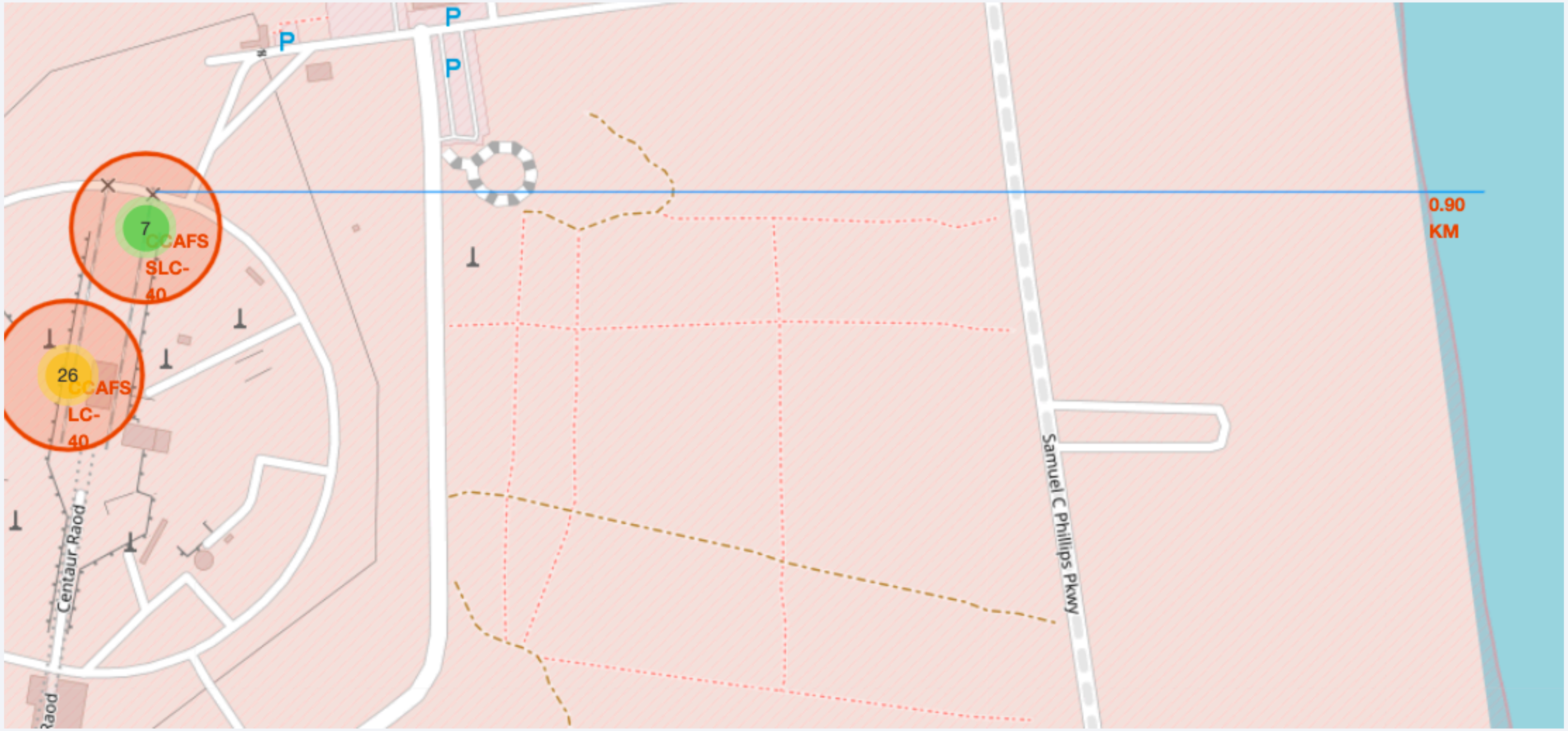
All launch sites marked on a map



Success/failed launches marked on the map



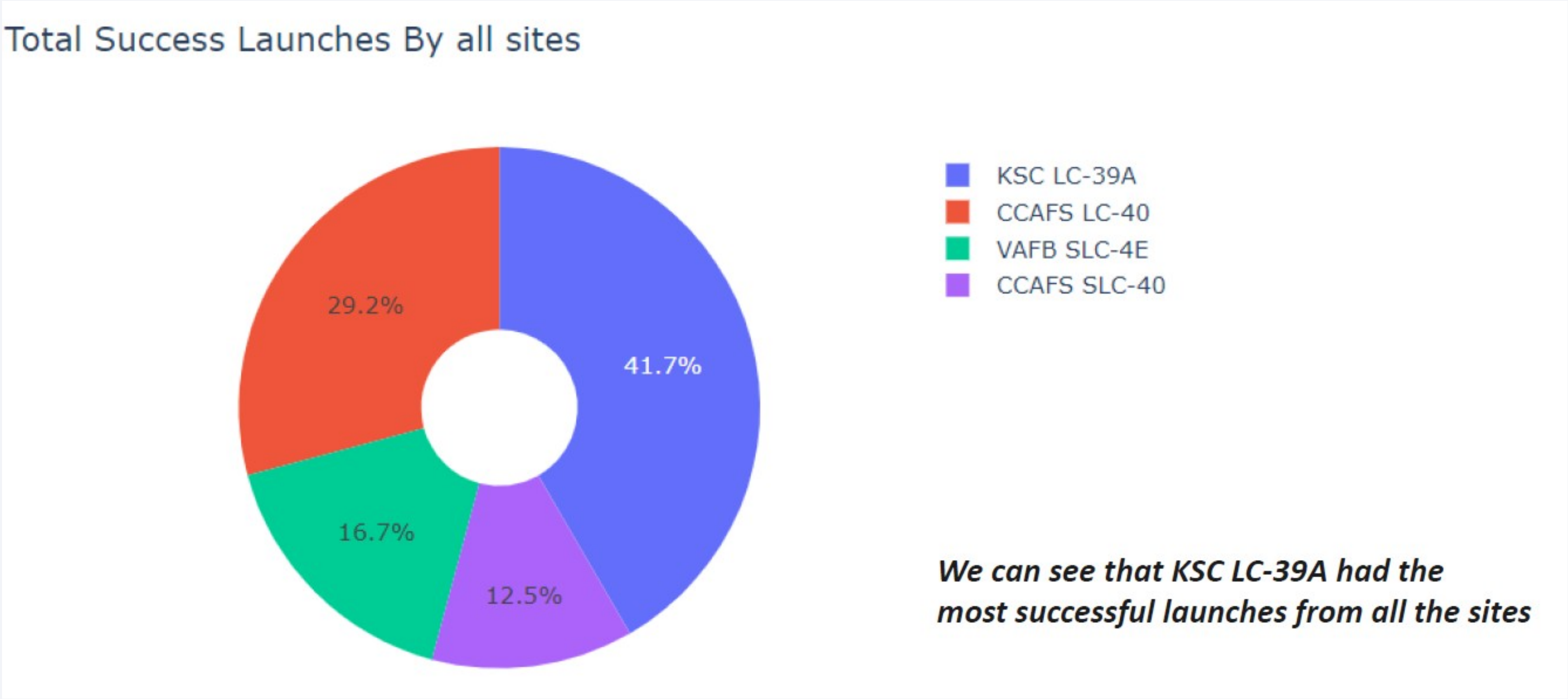
Distances between a launch site to its proximities



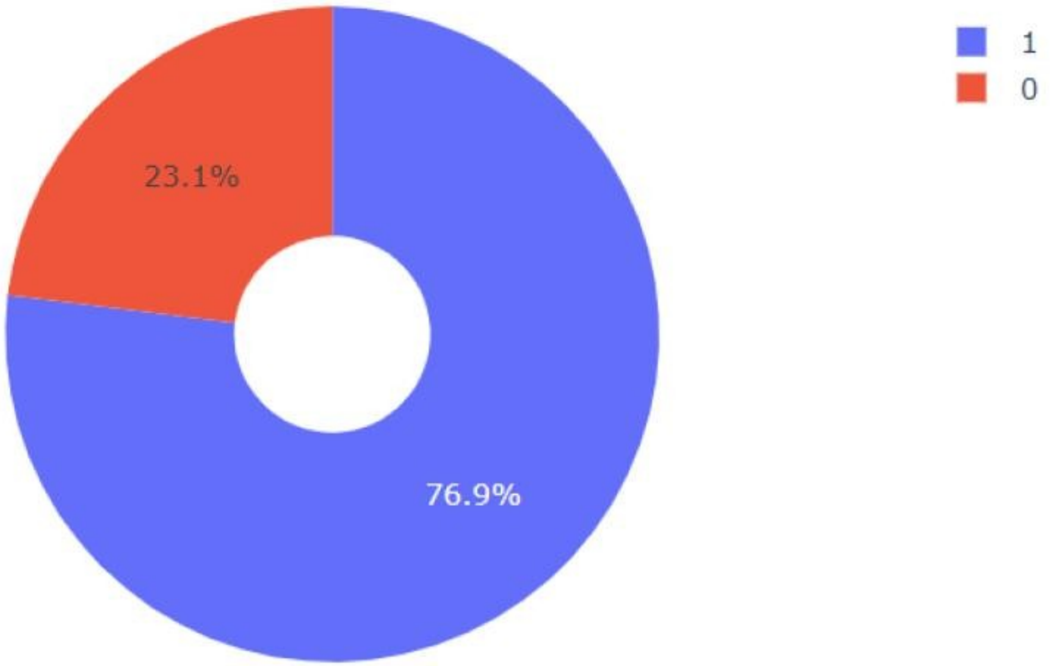
Section 5

Build a Dashboard with Plotly Dash

Total success launches by all sites

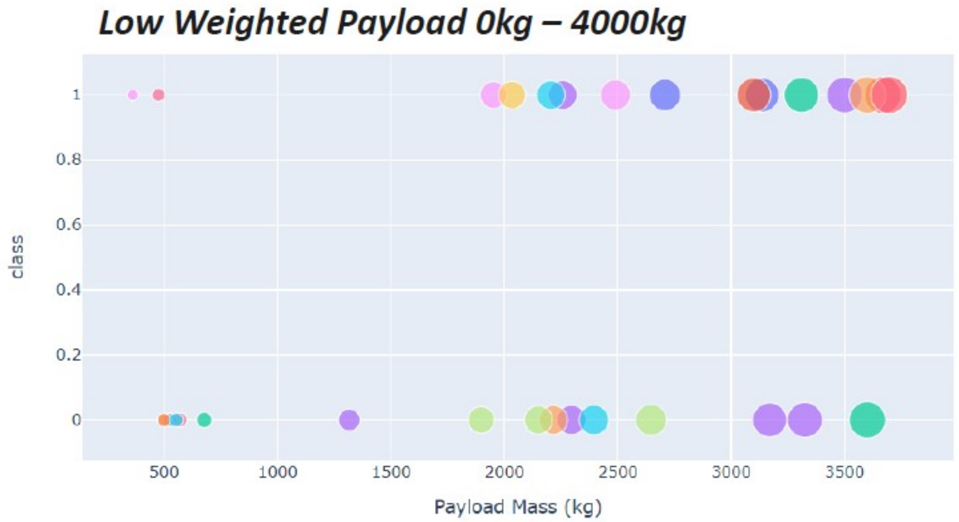


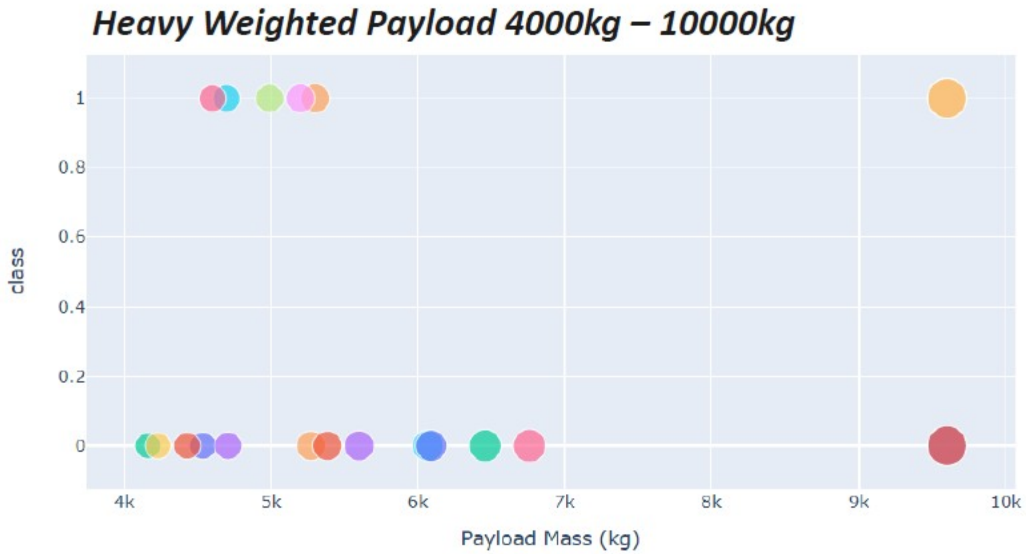
Success rate by site



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

Payload vs launch outcome





We can see the success rates for low weighted payloads is higher than the heavy weighted payloads

Section 6

Predictive Analysis (Classification)

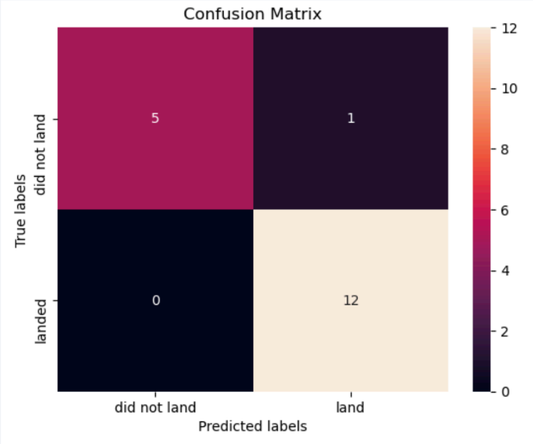
Classification Accuracy

Scores on test data for each method

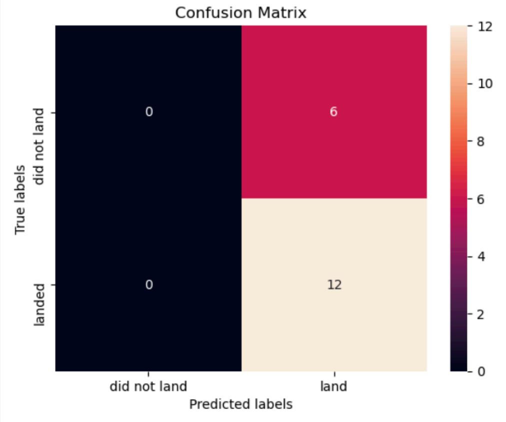
Logistic Regression: 0.944
SVM: 0.944
Decision Tree: 0.667
KNN: 0.944

So, except for Decision Tree, all other methods performed equally better.

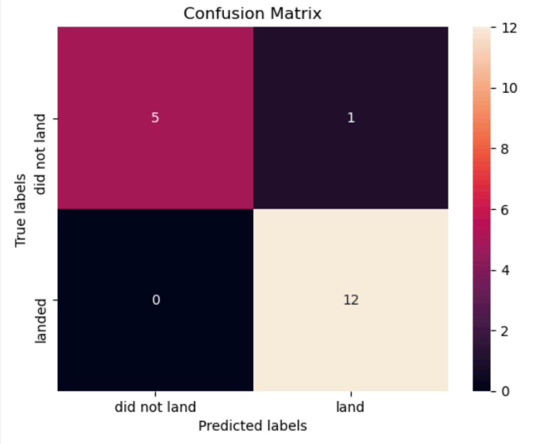
Confusion Matrix



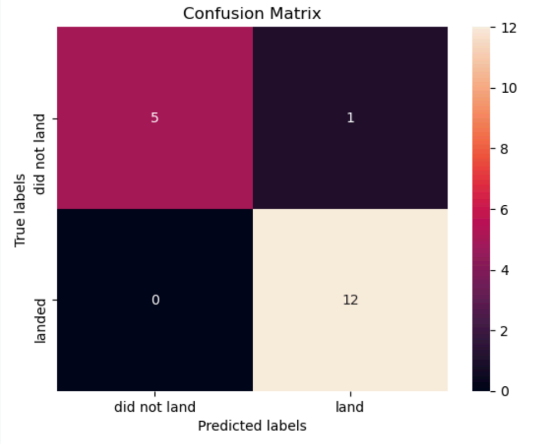
LR



Tree



SVM



KNN

Conclusions

- KSC LC 39A had the most successful launches from all the sites.
- Orbits which had the highest success rate were GEO, HEO, SSO, ES L1.
- The SVM, KNN, and Logistic Regression models are the best in terms of prediction accuracy for this dataset, but not Decision Trees.
- Low weighted payloads perform better than the heavier ones.
- The success rates for SpaceX launches increased with the years since 2013.

Thank you!