IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Xinyi Lou
2022-03-15

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies
  - Data Collection (API/Scraping)
  - Data Wrangling
  - Exploratory Data Analysis with Data Visualization
  - Exploratory Data Analysis with SQL
  - Interactive Visual Analytics with Folium
  - Predictive Analysis with Machine Learning
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

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.

- Problems you want to find answers

  - What factors determine if the rocket will land successfully?
  - Are there interaction amongst various features that could determine a successful landing?
  - Which operating or geographical precondition needed to be full filled so that a successful landing could be ensured?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected by using SpaceX API and web scraping 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
  - How to build, tune, evaluate classification models

Data Collection

- Data Collection including two parts:
  - Get request to the SpaceX API to get Falcon 9 launch data
  - Web scrap Falcon 9 launch records with "BeautifulSoup"

GET request to the SpaceX API

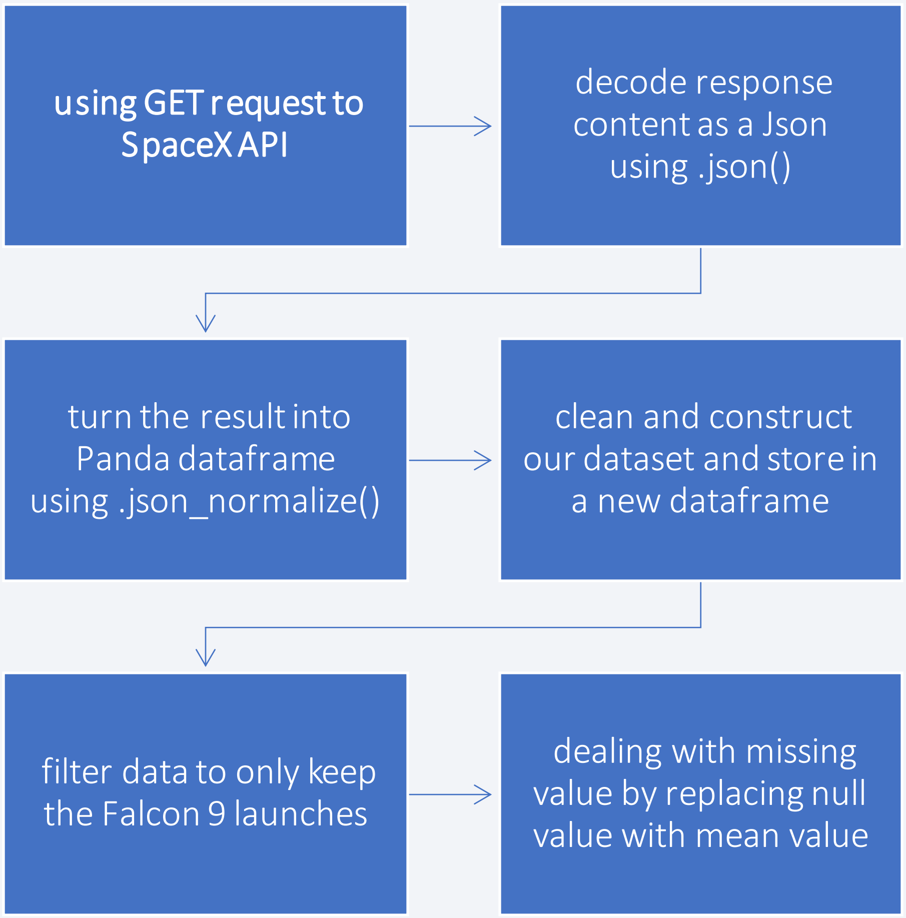
- Request and parse the SpaceX launch data using GET request
- Clean the request data: Filter the dataframe to only include Falcon 9 launches

Web scrap Falcon 9 launch records with "BeautifulSoup"

- Extract a Falcon 9 launch records HTML table from Wikipedia
- Parse the table and convert it into a Panda dataframe for further analysis

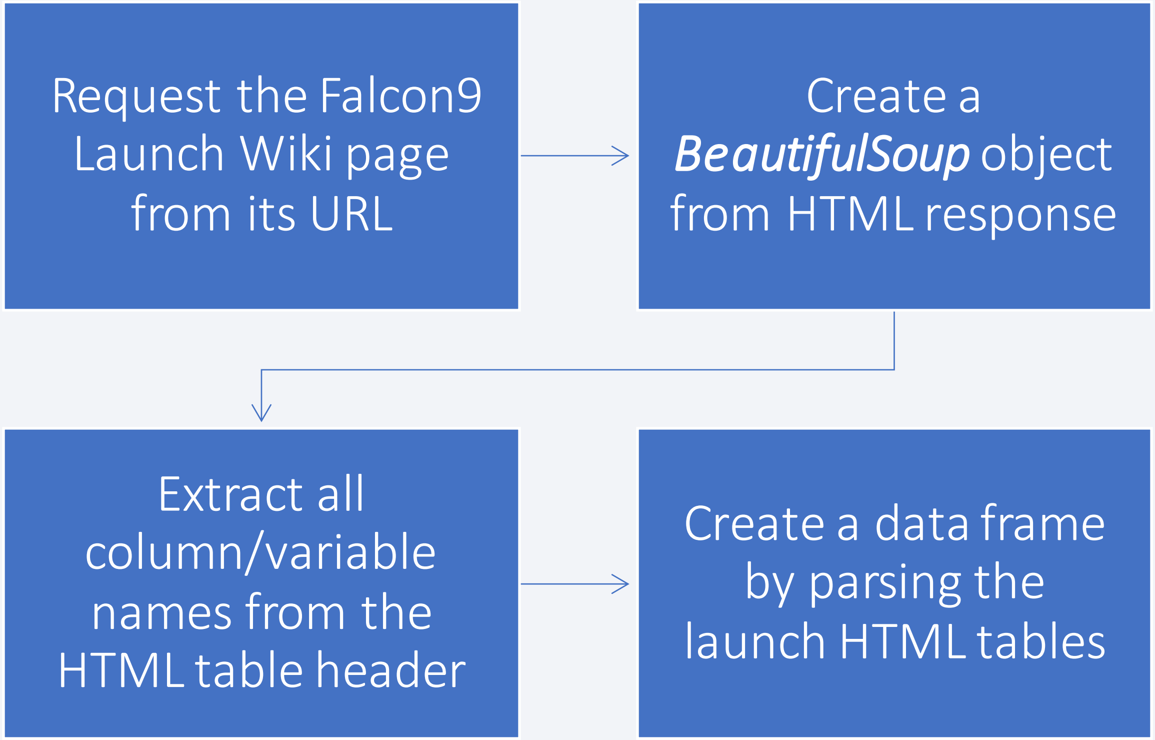
Data Collection – SpaceX API

- To collect data, we used the GET request to SpaceX API and then we cleaned the response dataset and did some data wrangling and formatting. Finally we stored it in a data frame.
- GitHub link to the notebook: https://github.com/JaneeMiao/Applied-Data-Science-Capstone/blob/main/Week1/1.1_jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

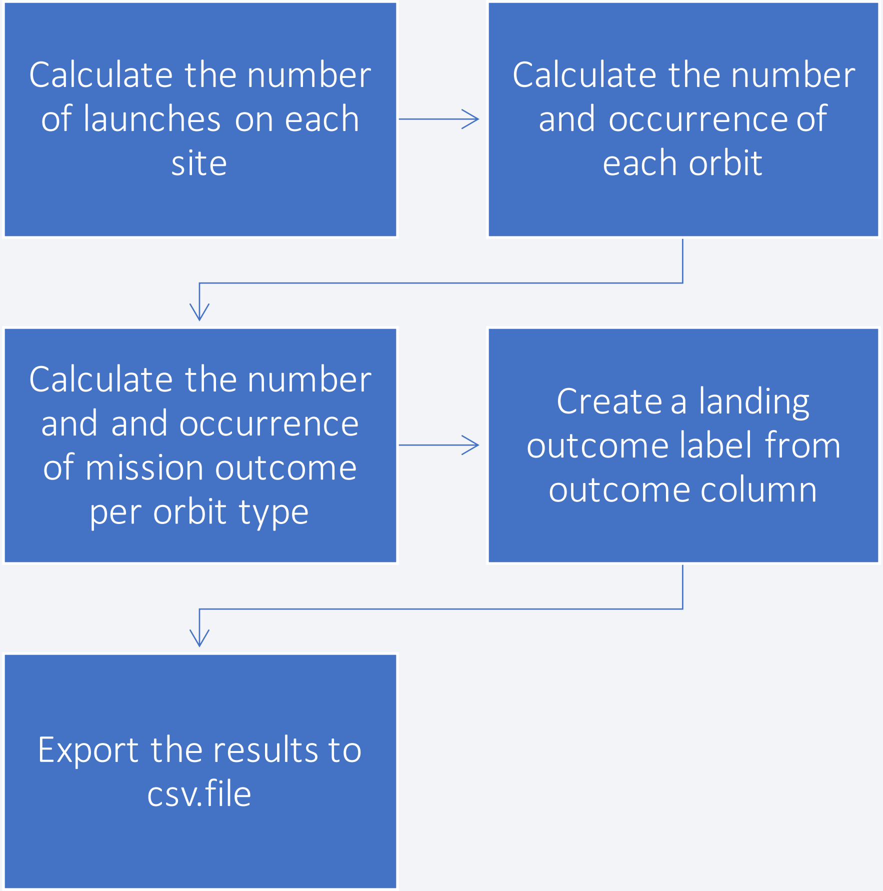
- We applied web scraping to get Falcon 9 launch
  records with BeautifulSoup
  - Extract a Falcon 9 launch records HTML table from Wikipedia with BeautifulSoup
  - Parse the table and convert it into a Pandas data frame
- GitHub link to web scraping notebook:
  https://github.com/JaneeMiao/Applied-Data-Science-Capstone/blob/main/Week1/1.2_jupyter-labs-webscraping.ipynb



Data Wrangling

- We performed some Exploratory Data Analysis (EDA) to find some patterns in the data and determined what would be the label for training supervised models.

- GitHub link to data wrangling related notebooks: https://github.com/JaneeMiao/Applied-Data-Science-Capstone/blob/main/Week1/1.3_Data%20Wrangling.ipynb



EDA with Data Visualization

- We explored the data by visualizing the relationship between:
  - Flight number and launch site by using scatter plot
  - Payload and launch site by using scatter plot
  - Success rate of each orbit type by using bar chart
  - Flight number and orbit type by using scatter plot
  - Payload and orbit type by using scatter plot
  - Launch success yearly trend by using line chart

- GitHub link of EDA with data visualization notebook: https://github.com/JaneeMiao/Applied-Data-Science-Capstone/blob/main/Week2/2.2.Complete%20the%20EDA%20with%20Visualization%20lab.ipynb

EDA with SQL

- We loaded the SpaceX dataset into corresponding table in a IBM db2 database and applied EDA with SQL to get insight from the dataset, for instance, we wrote SQL queries to find out:

  - The names of the unique launch sites in the space mission
  - The total payload mass carried by boosters launched by NASA (CRS)
  - The average payload mass carried by booster version F9 v1.1
  - The date when the first successful landing outcome in ground pad was achieved
  - The booster versions which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - The total number of successful and failure mission outcomes
  - The booster versions which have carried the maximum payload mass
  - The failed landing outcomes in drone ship, their booster version and launch site names

  ....

- GitHub link to EDA with SQL notebook: https://github.com/JaneeMiao/Applied-Data-Science-Capstone/blob/main/Week2/2.1.Complete%20the%20EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- We marked all launch sites and added map objects such as markers, circles, lines to mark succeeded and failed launches for each site on a folium map

  • We assigned the feature launch outcomes (failure and success) to class 0 and 1.
  0 for failure, 1 for success

  • We used colour labeled (red and green) marker to identify the succeeded and failed launches on a folium map.

  • We also calculated the distances between a launch site to its proximities and answered some questions like:
  "are launch sites in close proximity to railways / highways / coastline and do launch sites keep certain distance away from cities ? "

- GitHub link to interactive map with Folium map:
  https://github.com/JaneeMiao/Applied-Data-Science-Capstone/blob/main/Week3/3.1.Complete%20the%20Data%20Visualization%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- We built an interactive dashboard with Plotly Dash to enable to perform interactive visual analytics on SpaceX launch data in real-time.
- The dashboard contains input components such as a dropdown list and a range slider to interact with a pie chart and a scatter point chart.
    - Pie charts show the total launches by a certain site.
    - Scatter point charts show the relationship with Outcome and Payload Mass for different booster versions.
- GitHub link of code in Plotly Dash lab: https://github.com/JaneeMiao/Applied-Data-Science-Capstone/blob/main/Week3/spacex_dash_app.py

Predictive Analysis (Classification)

- We found the best performing classification model by performing following steps:
  - We performed exploratory data analysis and determined training labels:

Create a column for the class → Standardize the data → Split into training data and test data

  - We built different machine learning models (SVM, Classification Trees and Logistic Regression) and found best hyperparameter for these models
  - We calculated the accuracy of different models on the test data and found the best performing model for our use case.

- GitHub link to completed predictive analysis notebook:
  https://github.com/JaneeMiao/Applied-Data-Science-Capstone/blob/main/Week4/4.1.Machine%20Learning%20Prediction.ipynb

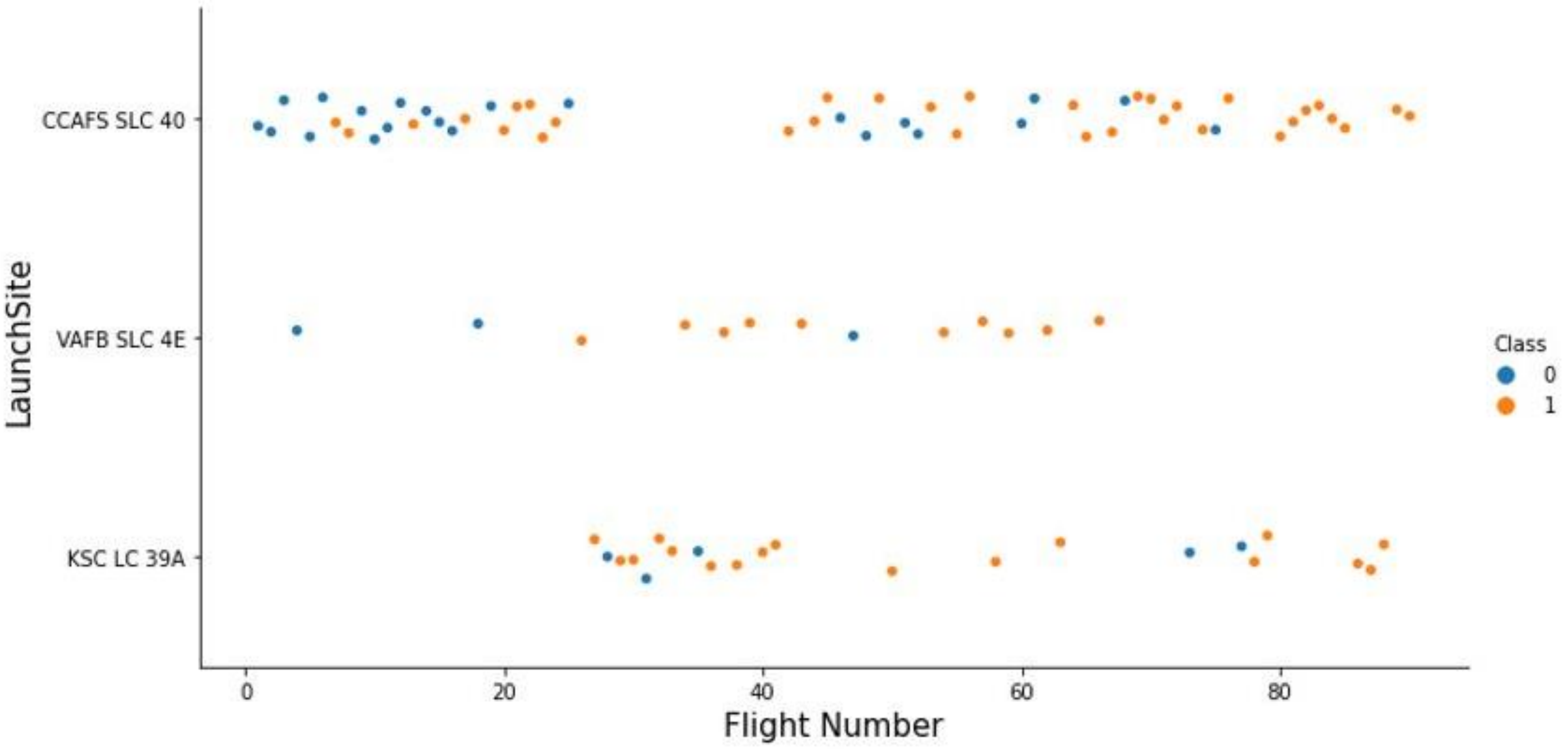
Results

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

Section 2

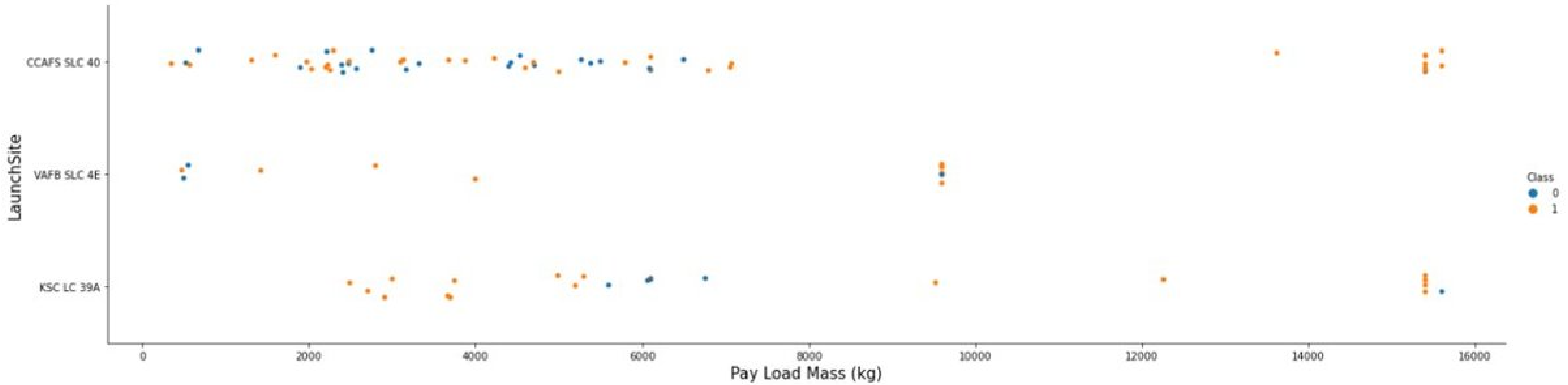
Insights drawn from EDA

Flight Number vs. Launch Site



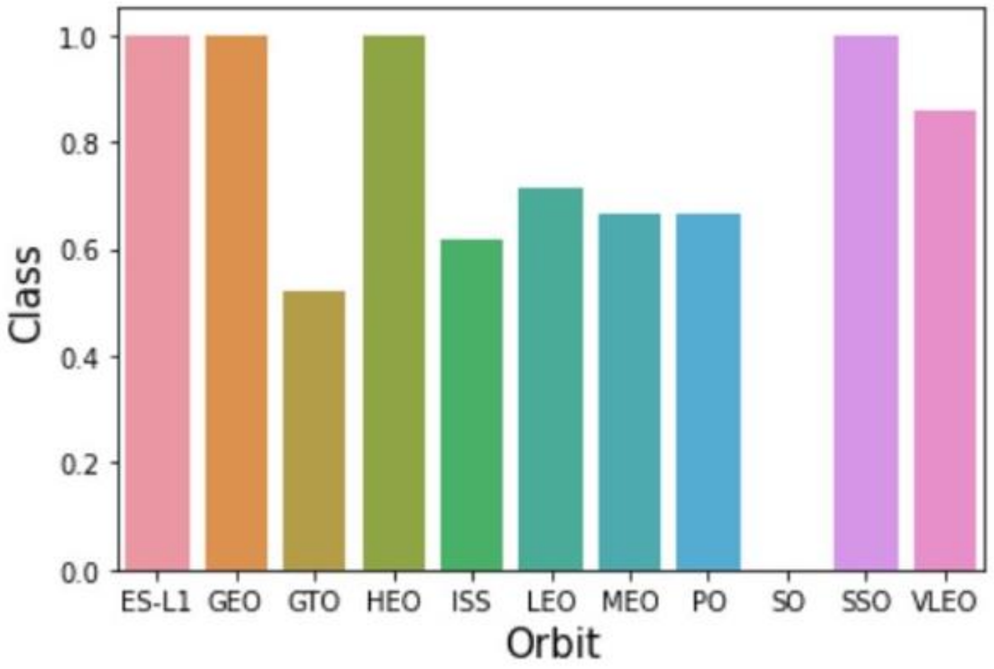
The larger the flight amount at a launch site, the higher the success rate at a launch site

Payload vs. Launch Site



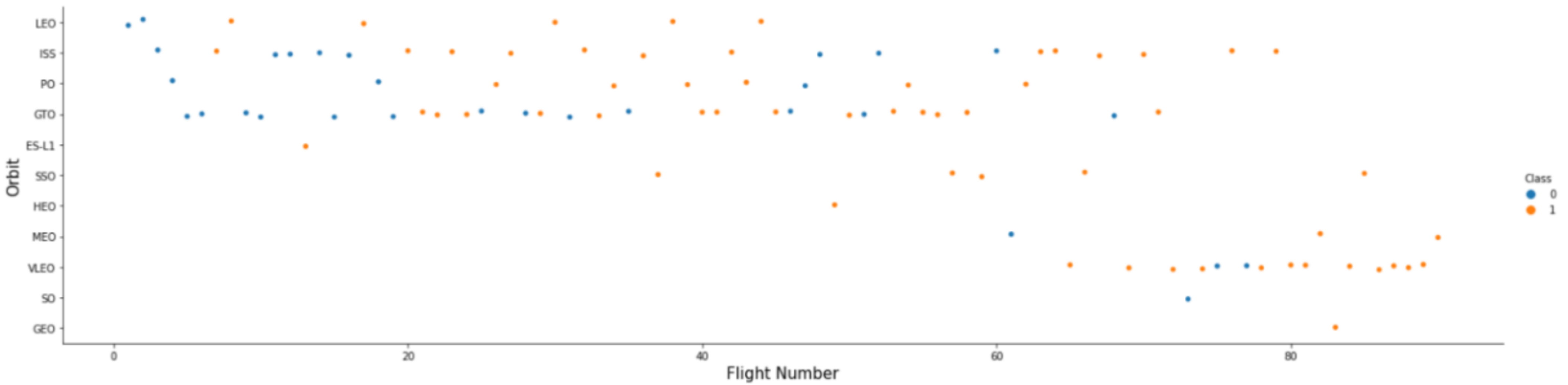
- For the VAFB-SLC launchsite there are no rockets launched for heavypayload mass (greater than 10000 kg).
- The greater the payload mass for launch site CCAFS SLC40, the higher the success rate for the rocket.

Success Rate vs. Orbit Type



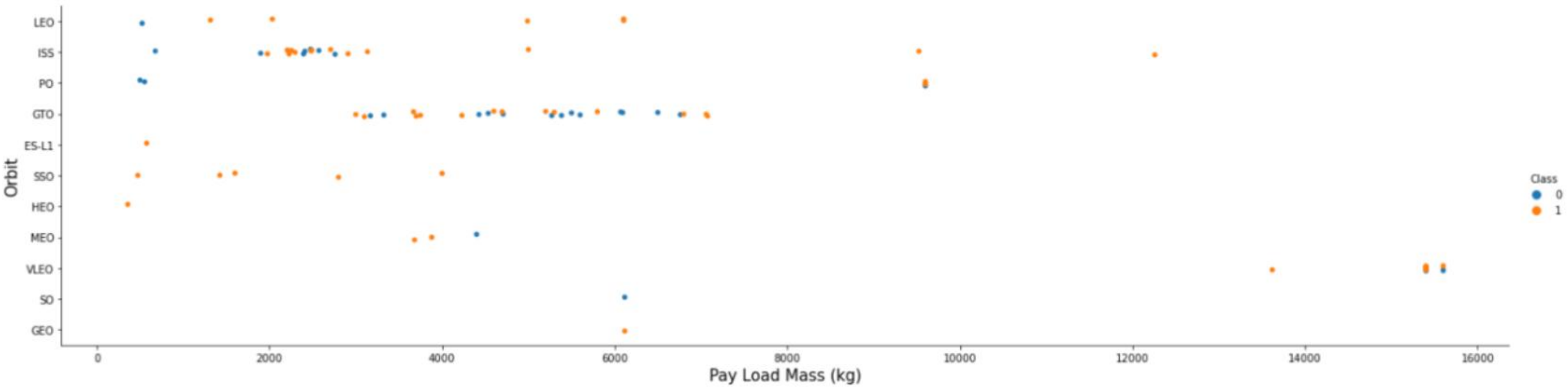
- The Orbits ES-L1, GEO, HEO, SSO have high success rate.

Flight Number vs. Orbit Type



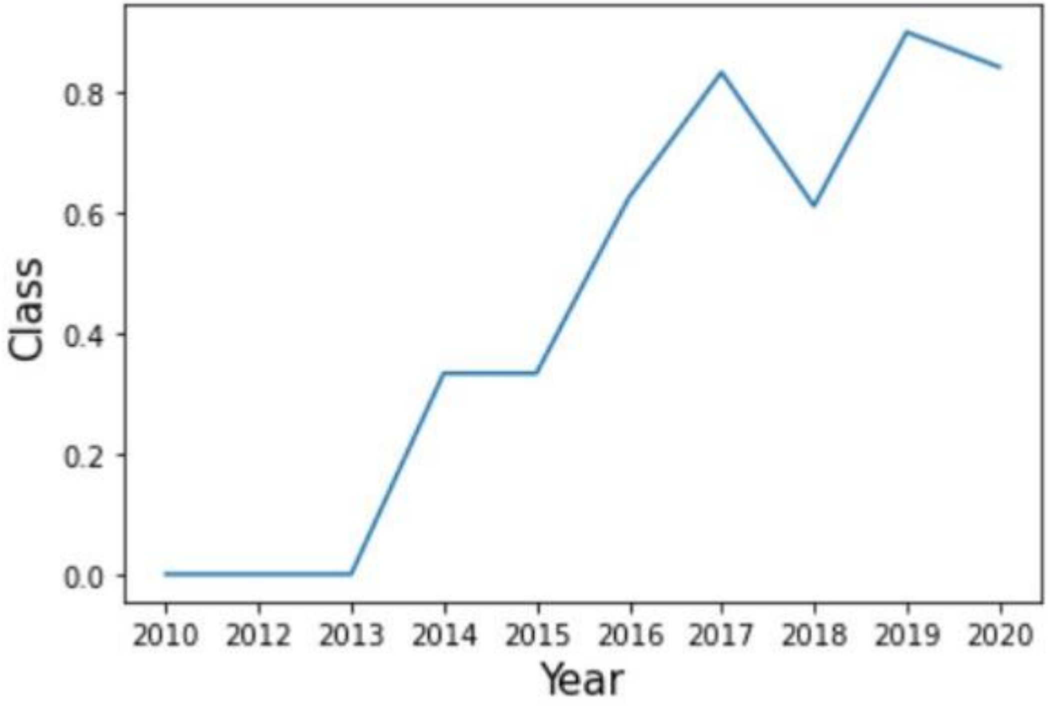
We could see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- 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



- We can observe that the sucess rate since 2013 kept increasing till 2020.

All Launch Site Names

- We used DISTINCT to show unique launch sites from SpaceX dataset.

In [5]:
```
%%sql
select distinct LAUNCH_SITE
from SPACEXDATASET;
```

Out[5]:

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

Launch Site Names Begin with 'CCA'

- We used the query below to display 5 records where launch sites begin with the string 'CCA'

In [10]:
```
%%sql
select * from SPACEXDATASET
where LAUNCH_SITE like 'CCA%'
limit 5;
```

Out[10]:

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

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

In [11]:
```
%%sql
select sum(PAYLOAD_MASS__KG_)
from SPACEXDATASET
where Customer = 'NASA (CRS)';
```

Out[11]:

1
45596

Average Payload Mass by F9 v1.1

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

In [21]:
```
%%sql
select AVG(payload_mass__kg_) as avg from SPACEXDATASET
where booster_version like 'F9 v1.1%'
```

Out[21]:

AVG
2534

First Successful Ground Landing Date

- We used DISTINCT to find the right value representing successful ground landing and then used MIN-Function found the dates of the first successful landing outcome on ground pad

In [22]:
```
%%sql
select distinct landing__outcome from SPACEXDATASET
```

Out[22]:

landing__outcome
Controlled (ocean)
Failure
Failure (drone ship)
Failure (parachute)
No attempt
Precluded (drone ship)
Success
Success (drone ship)
Success (ground pad)
Uncontrolled (ocean)

In [23]:
```
%%sql
select min(date) from SPACEXDATASET where landing__outcome = 'Success (ground pad)'
```

Out[23]:

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- We used the WHERE Clause to list the names of boosters which have successfully landed on drone ship and applied the AND condition to determine successful landing with payload mass greater than 4000 but less than 6000

In [32]:
```
%%sql
select booster_version, payload_mass__kg_ from SPACEXDATASET
where landing__outcome = 'Success (drone ship)' and 4000 < payload_mass__kg_ and payload_mass__kg_ < 6000
group by booster_version, payload_mass__kg_
```

Out[32]:

booster_version	payload_mass__kg_
F9 FT B1021.2	5300
F9 FT B1031.2	5200
F9 FT B1022	4696
F9 FT B1026	4600

Total Number of Successful and Failure Mission Outcomes

- We used COUNT calculated the total number of successful and failure mission outcomes

In [42]:
```
%%sql
select mission_outcome, count(mission_outcome) as total_nr
from SPACEXDATASET
group by mission_outcome
```

Out[42]:

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

Boosters Carried Maximum Payload

- We used a subquery in the WHERE clause and the MAX function to list the names of the booster which have carried the maximum payload mass

In [43]:
```
%%sql
SELECT DISTINCT booster_version
FROM SPACEXDATASET
WHERE payload_mass__kg_ = (
    SELECT max(payload_mass__kg_)
    FROM SPACEXDATASET
)
```

Out[43]:

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

- We used WHERE clause with AND condition to list the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

In [52]:
```
%%sql
select landing__outcome, booster_version,launch_site
from SPACEXDATASET
where landing__outcome = 'Failure (drone ship)' and year(date) = 2015
group by landing__outcome, booster_version,launch_site
```

Out[52]:

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

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

- We used GROUP BY and ORDER BY to 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 a descending order.

In [54]:

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

Out[54]:

landing__outcome	total_nr
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

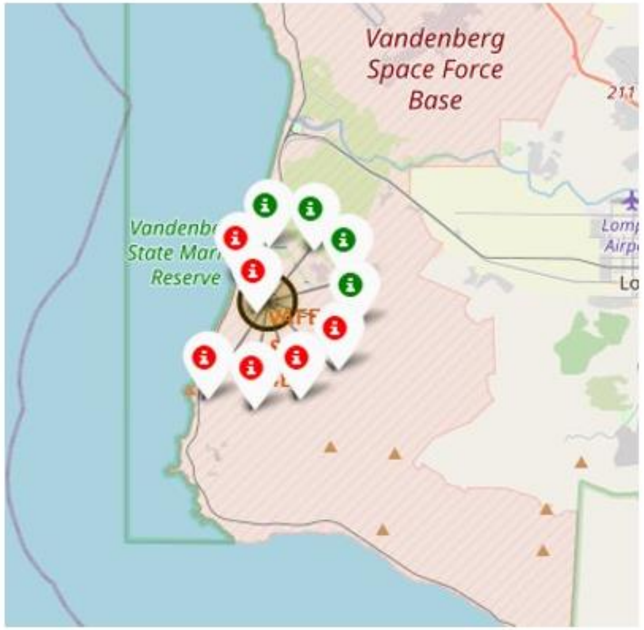
Section 4

Launch Sites Proximities Analysis

All launch sites' location on a global map

As we can see, all SpaceX launch sites are in the United States of America coasts, Florida and California.

Color labeled launch sites on the map

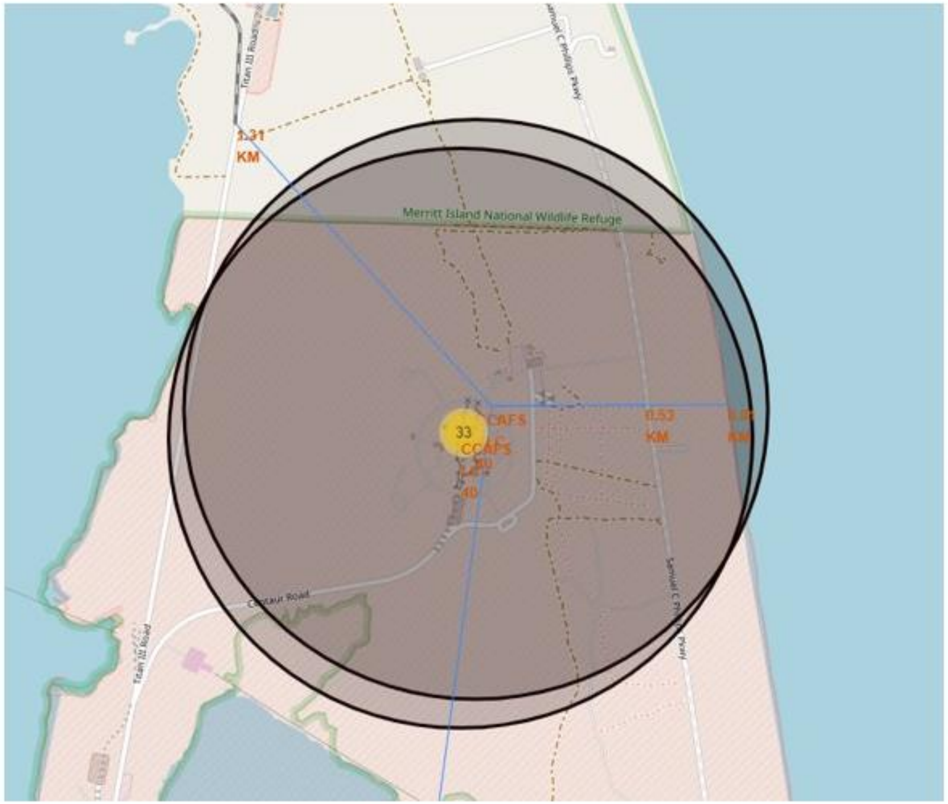


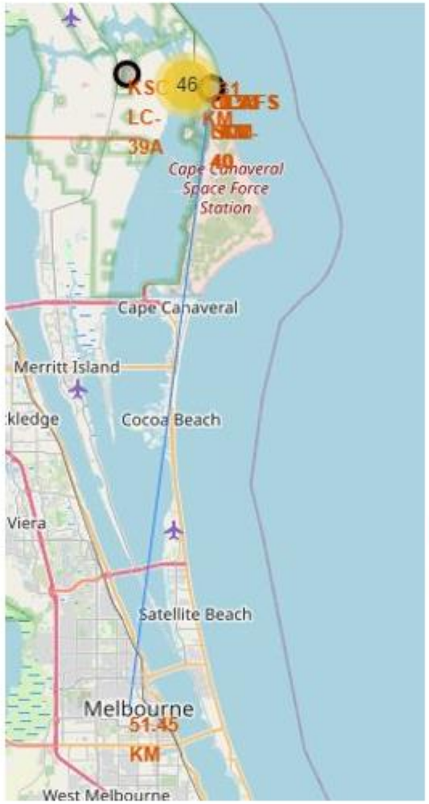
California Launch Site

Florida Launch Sites

Green Marker: successful launches
Red Marker: failures

Launch site distance to railway, highway, coastline, city





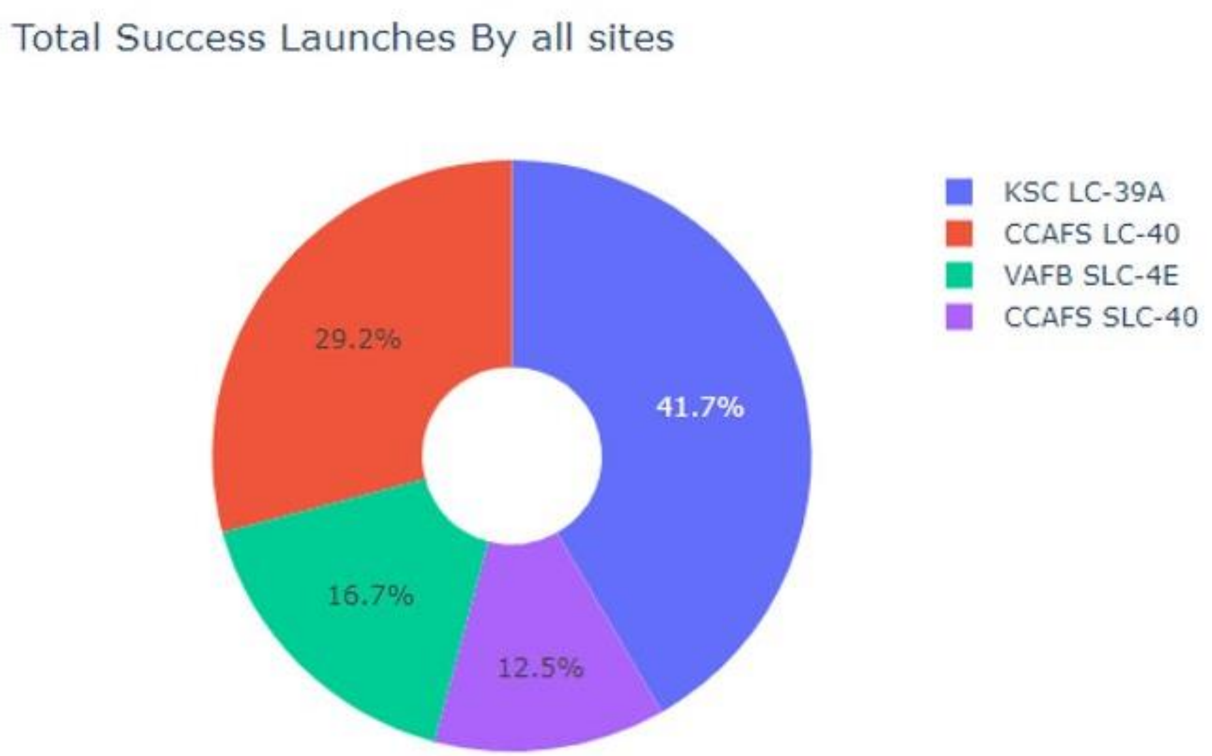
Distance to railway station: 1.31 km
Distance to closest highway: 0.53 km
Distance to coast: 0.81 km
Distance to city: 51.45 km

- As rockets may crash so that in order to minimize people at risk from falling debris it is quite often that launch sites are in close proximity to coastline but not in close proximity to cities
- Launch sites are relatively close to railway and highway for transport reasons

Section 5

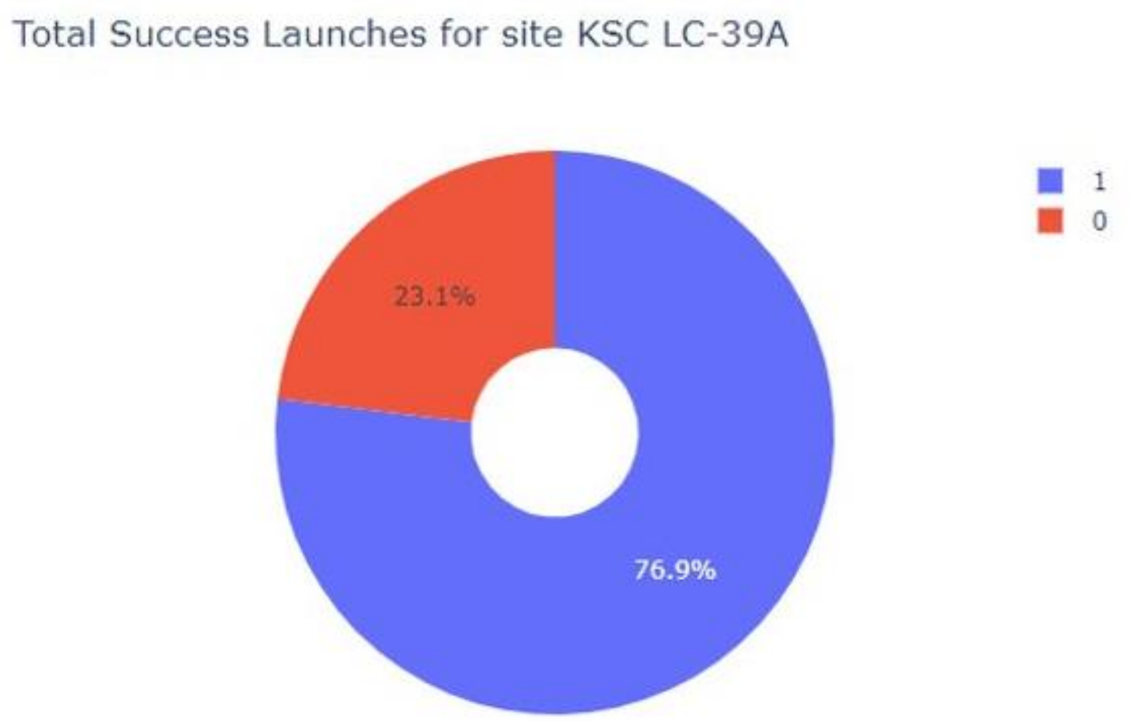
Build a Dashboard with Plotly Dash

Total Success Launches by All Sites



As we can see in the pie chart, KSC LC-39A has the most successful launches among all sites.

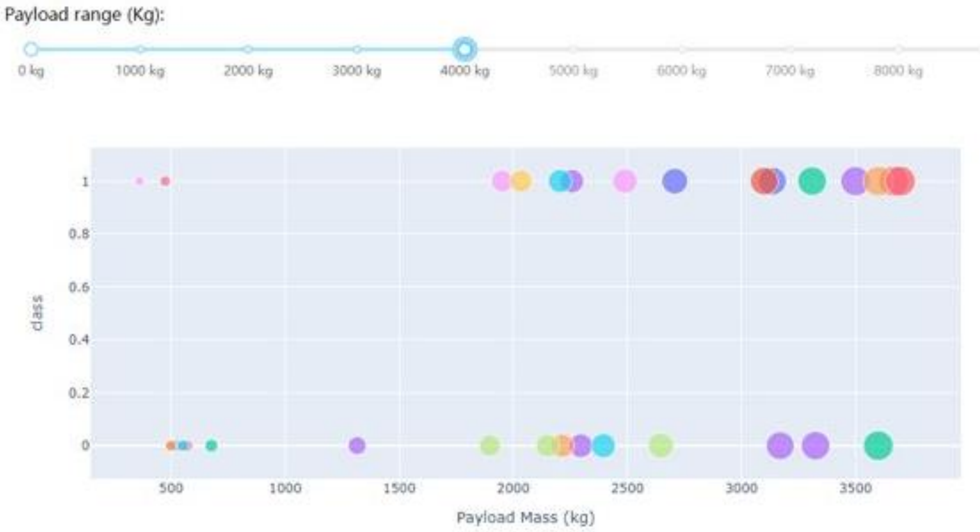
Launch site with the highest launch success ratio



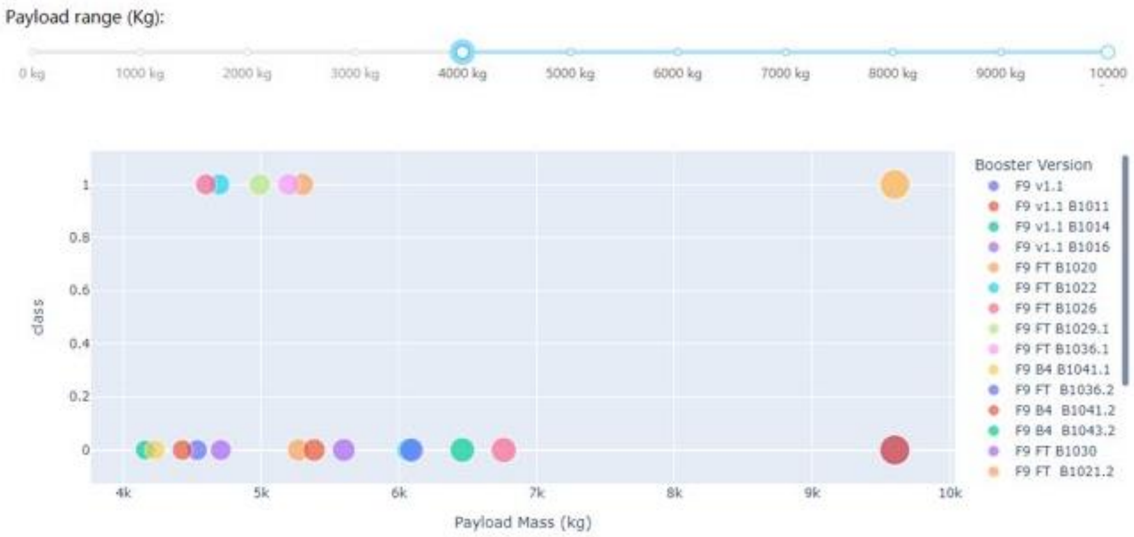
KSC LC-39A has a success rate at 76.9% while having a 23.1% failure rate.

Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider

Low weighted payload: 0 - 4000 kg



High weighted payload: 4000 - 10000 kg



Success rates for low weighted payload is higher than heavy weighted payloads

Section 6

Predictive Analysis (Classification)

Classification Accuracy

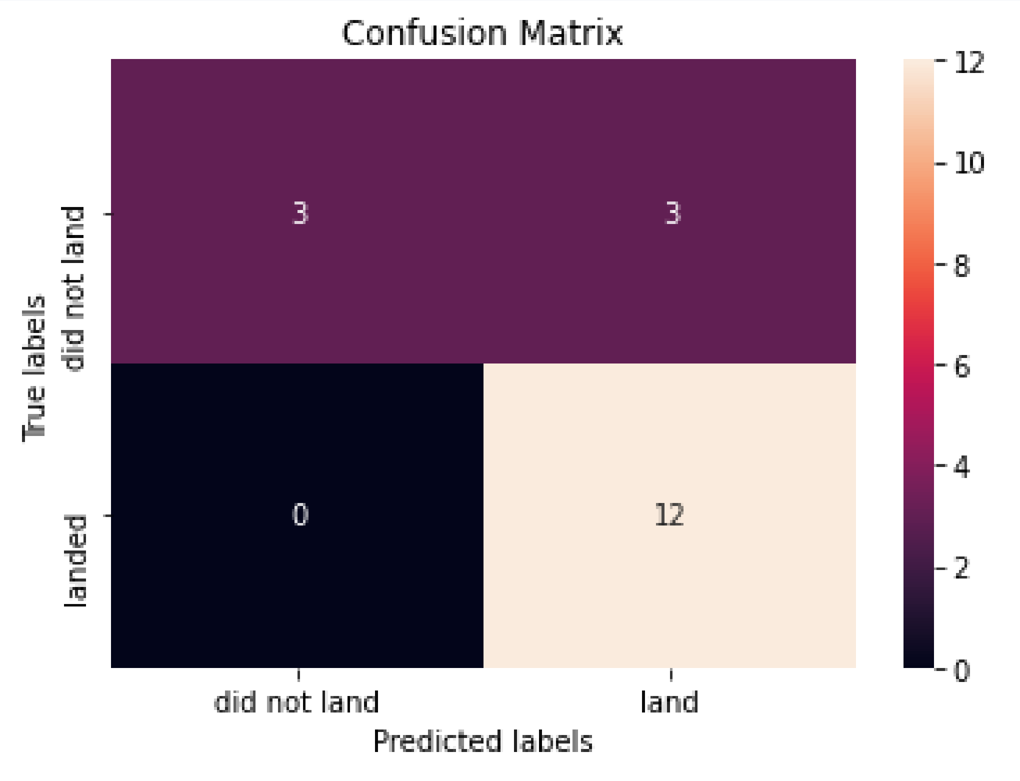
- The decision tree is the model with the highest classification accuracy

```
models = {'LogisticRegression':logreg_cv.best_score_,
          'SupportVectorMachine': svm_cv.best_score_,
          'DecisionTree':tree_cv.best_score_,
          'KNeighbours':knn_cv.best_score_
         }
bestalgorithm = max(models, key=models.get)
print('Best model is', bestalgorithm,'with a score of', models[bestalgorithm])
if bestalgorithm == 'LogisticRegression':
    print('Best params is :', logreg_cv.best_params_)
if bestalgorithm == 'SupportVectorMachine':
    print('Best params is :', svm_cv.best_params_)
if bestalgorithm == 'DecisionTree':
    print('Best params is :', tree_cv.best_params_)
if bestalgorithm == 'KNeighbours':
    print('Best params is :', knn_cv.best_params_)
```

```
Best model is DecisionTree with a score of 0.875
Best params is : {'criterion': 'gini', 'max_depth': 4, 'max_features': 'auto', 'min_samples_leaf': 2, 'min_samples_split': 5, 'splitter': 'r
andom'}
```

Confusion Matrix

- As we can see the confusion matrix of decision tree classifier showed on the right side, there is false positive issue occurred .i.e. there are three unsuccessful landing marked as successful landing by the classifier



Conclusions

- The larger the flight amount at a launch site, the higher the success rate at a launch site.
- The Orbits ES-L1, GEO, HEO, SSO have high success rate.
- Launch sucess rate since 2013 kept increasing till 2020.
- KSC LC-39A has the most successful launches among all sites.
- Success rates for low weighted payload is higher than heavy weighted payloads.
- The decision tree classifier is the best machine learning algorithm for our use case.

Thank you!