IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Raziq Ramailan
12th February 2023

Outline

Executive Summary

- Summary of methodologies
  - Data collection with web scraping and using API
  - Data wrangling
  - EDA with SQL and data visualization
  - Interactive map with Folium
  - Dashboard with Plotly Dash
  - Predictive analysis
- Summary of all results
  - EDA results
  - Interactive results
  - Predictive analysis results

Introduction

- Project background and context
  - SpaceX has been known worldwide for their ground-breaking reusable rockets technology. But their success only came after countless of failures and millions of money gone. It is exciting to see how far they have came by.
- Problems you want to find answers
  - What are the causes for SpaceX success and failures?
  - How can we predict SpaceX future success/failure based on the data?
  - What are the parameters that can contribute to their success?

Section 1

Methodology

Methodology

Executive Summary

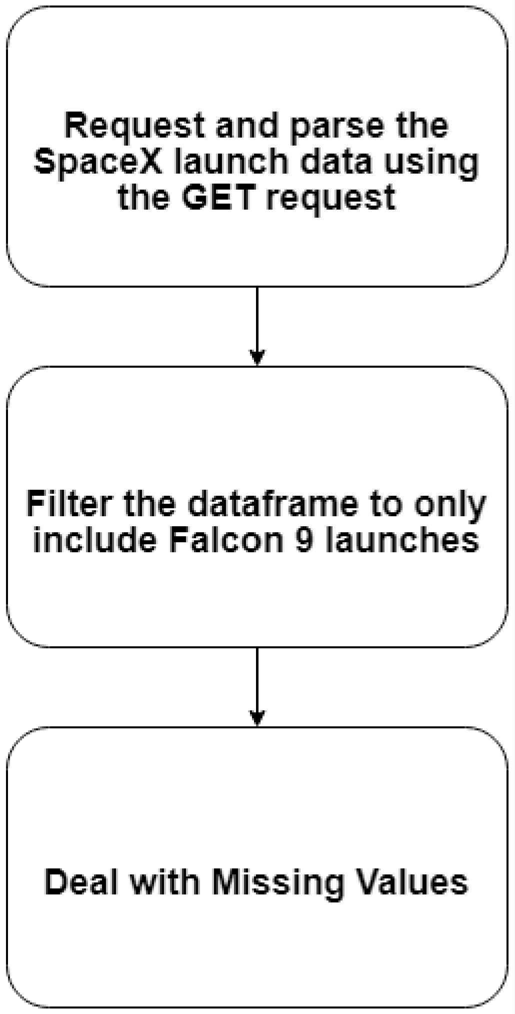
- Data collection methodology:
  - Data collected by web scraping from Wikipedia and using SpaceX API.
- Perform data wrangling
  - Apply one hot encoding and data cleaning to work with null values.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Few methods of machine learning has been applied to train and test the data set to get the best classifier.

Data Collection

- Data has been collected by web scraping from Wikipedia using BeautifulSoup and using SpaceX API calls.
- We retrieved the data from Wikipedia in HTML table format and convert it to Pandas data frame before the data is ready to use.
- Another method is using API. SpaceX has provided the API and retrieved it in json format and normalize the data into Pandas data frame.

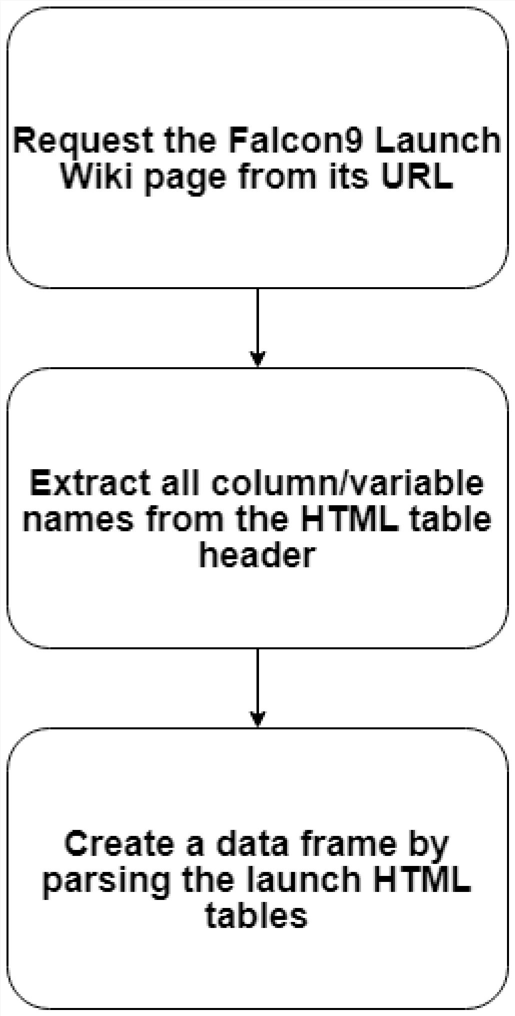
Data Collection – SpaceX API

- Flowcharts of SpaceX API calls process.
- GitHub URL: https://github.com/rzqr/IBM-Data-Science-Projects/blob/master/2-%20Data-collection-api.ipynb



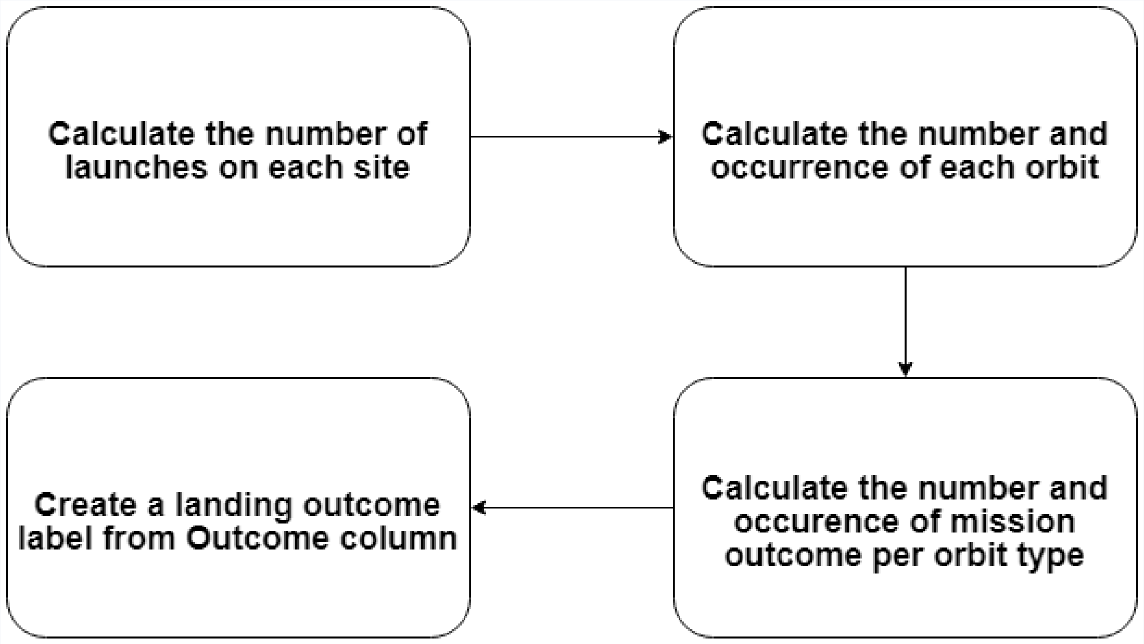
Data Collection - Scraping

- Flowcharts of web scraping of data using BeautifulSoup
- GitHub URL: https://github.com/rzqr/IBM-Data-Science-Projects/blob/master/1%20-%20Data%20collection%20with%20web%20scrapping.ipynb



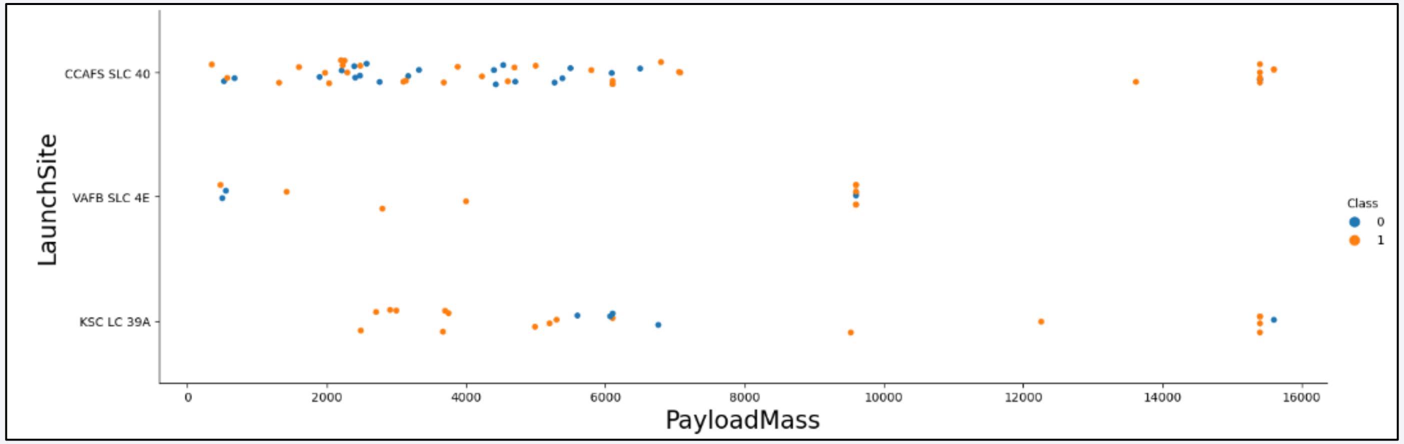
Data Wrangling

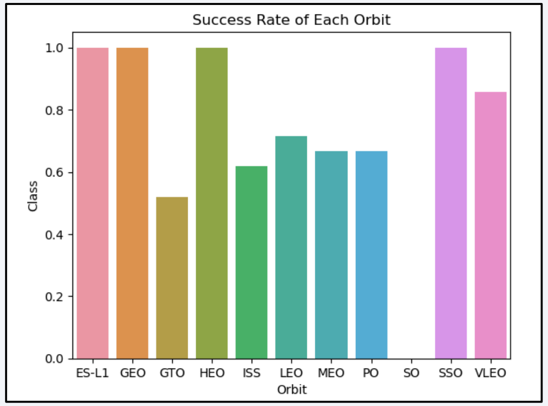
- Perform exploratory Data Analysis and determine Training Labels
- GitHub URL: https://github.com/rzqr/IBM-Data-Science-Projects/blob/master/3%20-%20Data%20wrangling.ipynb

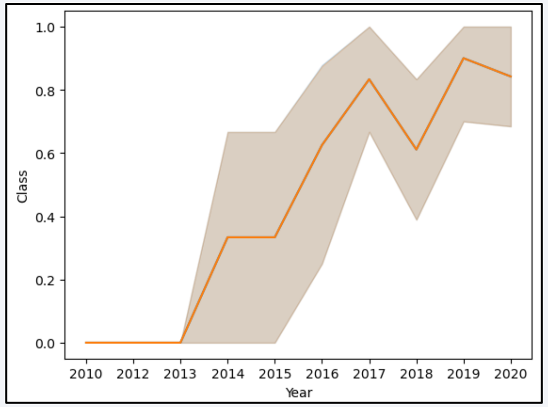


EDA with Data Visualization

- Scatter chart, bar chart and line chart have been plotted to visualize the correlation between categorical such as launch site and numerical variables such as payload mass from SpaceX data.
- GitHub URL: https://github.com/rzqr/IBM-Data-Science-Projects/blob/master/5%20-%20eda-dataviz.ipynb







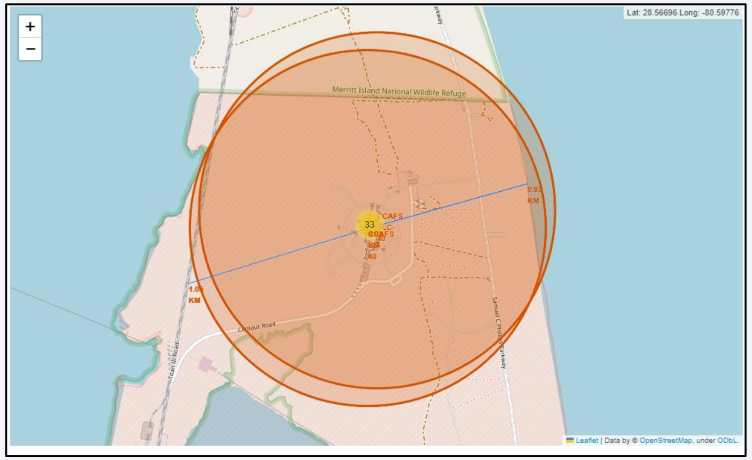
EDA with SQL

- SQL queries performed:
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in ground pad was acheived.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass
  - List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - 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

- GitHub URL: https://github.com/rzqr/IBM-Data-Science-Projects/blob/master/4%20-%20eda-sql-coursera.ipynb

Build an Interactive Map with Folium

- Markers and circles are used to mark and highlight launch sites with success/failed launch outcome.
- Lines are used to show the distance between launch site and nearest coastline and railway locations.
- GitHub URL: https://github.com/rzqr/IBM-Data-Science-Projects/blob/master/6%20-%20Data%20viz%20with%20Folium.ipynb

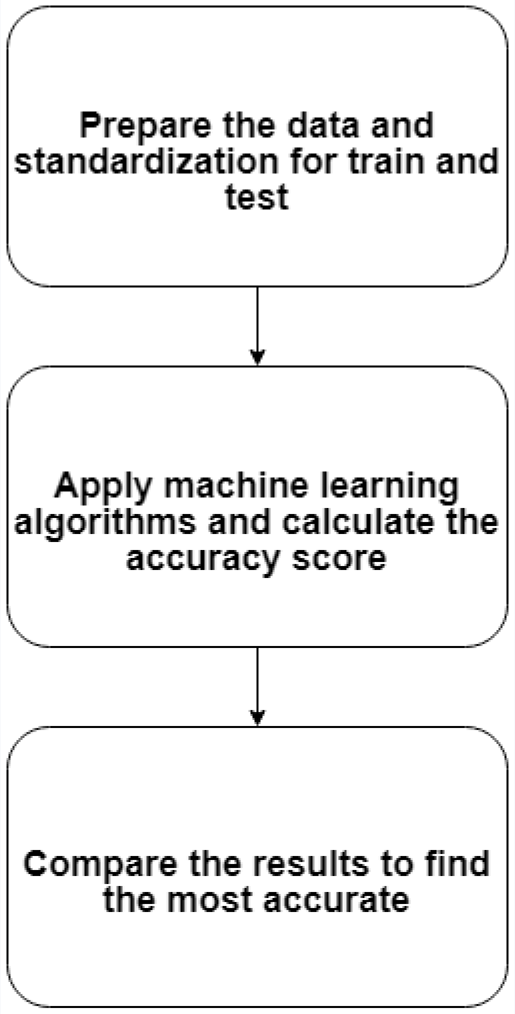


Build a Dashboard with Plotly Dash

- Pie chart has been plotted to show the percentage of launch success between sites.
- Scatter chart has been plotted to show the correlation between launch sites and the payload mass.
- GitHub URL: https://github.com/rzqr/IBM-Data-Science-Projects/blob/master/7%20-%20spacex_dash_app.py

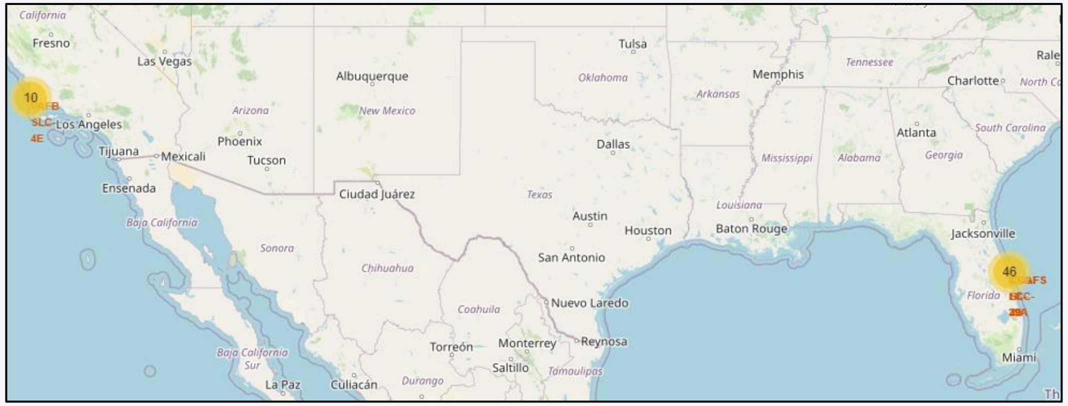
Predictive Analysis (Classification)

- Flowchart of predictive analysis.
- GitHub URL: https://github.com/rzqr/IBM-Data-Science-Projects/blob/master/8%20-%20Machine%20Learning%20Prediction_Part_5.ipynb



Results

- Exploratory data analysis results
  - With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
  - Sucess rate since 2013 kept increasing till 2020
- Interactive analytics demo in screenshots



- Predictive analysis results

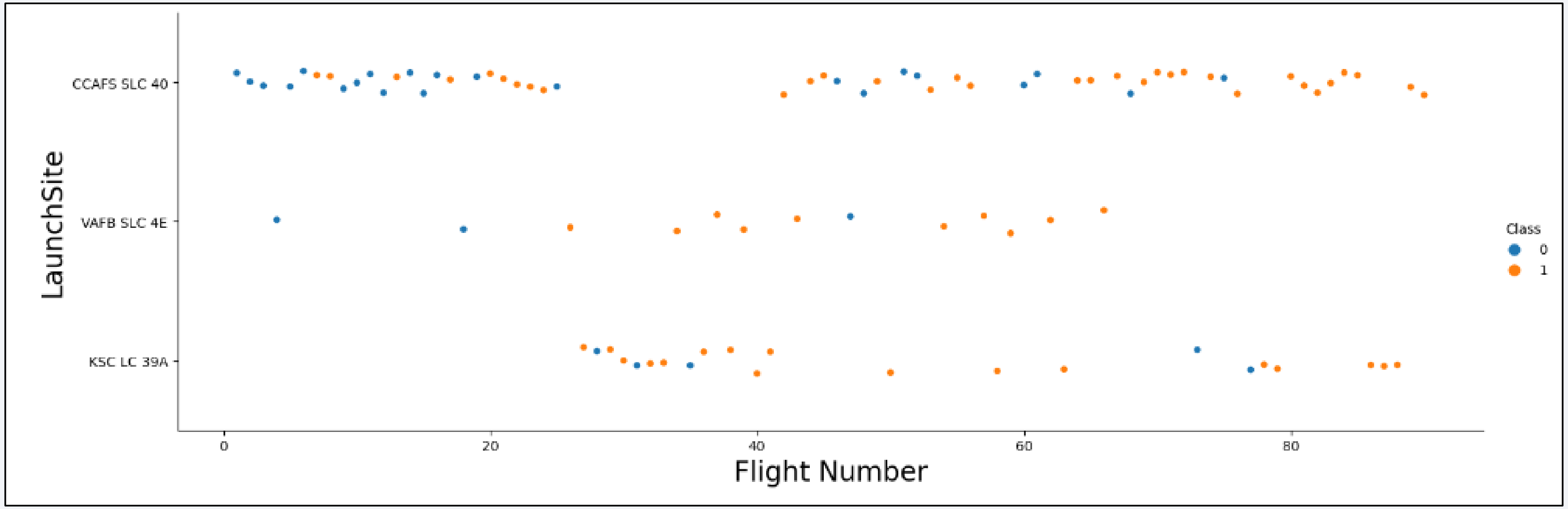
```
models = {'K Nearest Neighbors':knn_cv.best_score_,
              'Decision Tree':tree_cv.best_score_,
              'Logistic Regression':logreg_cv.best_score_,
              'Support Vector Machine': svm_cv.best_score_}

best_method = max(models, key=models.get)
print('Best model is: {} with a score of {}'.format(best_method,models[best_method]))
```

```
Best model is: Decision Tree with a score of 0.8892857142857142
```

Section 2

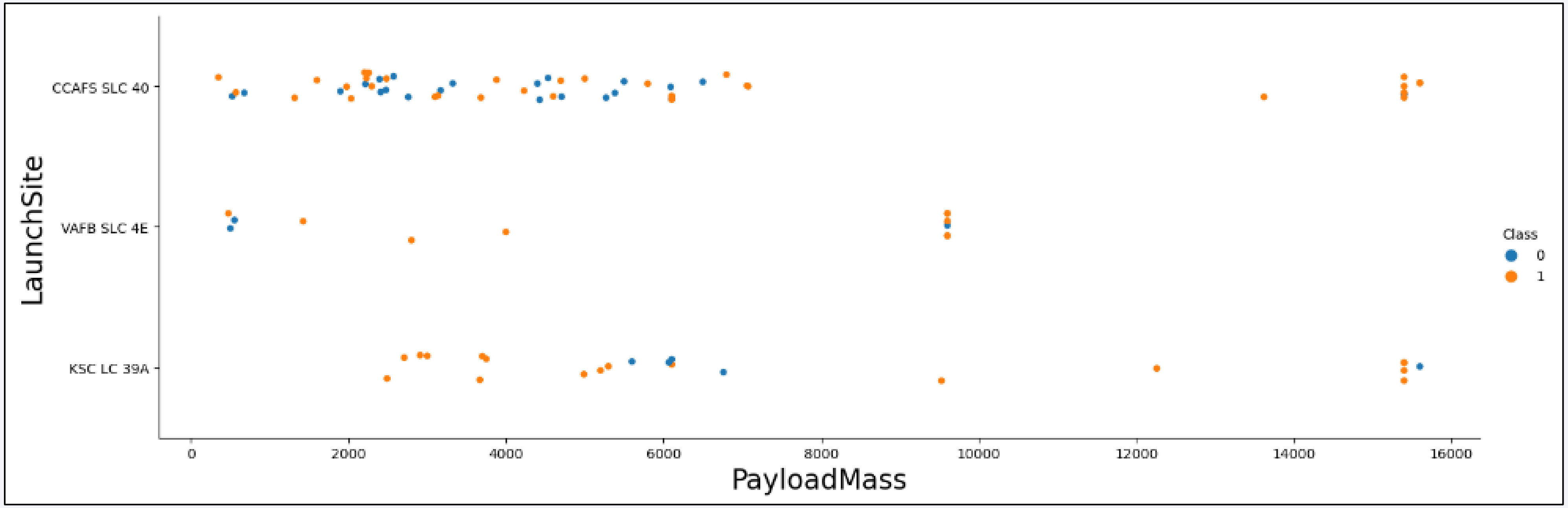
Insights drawn from EDA

Flight Number vs. Launch Site



• Most of the flight launched from CCAFS SLC 40 site

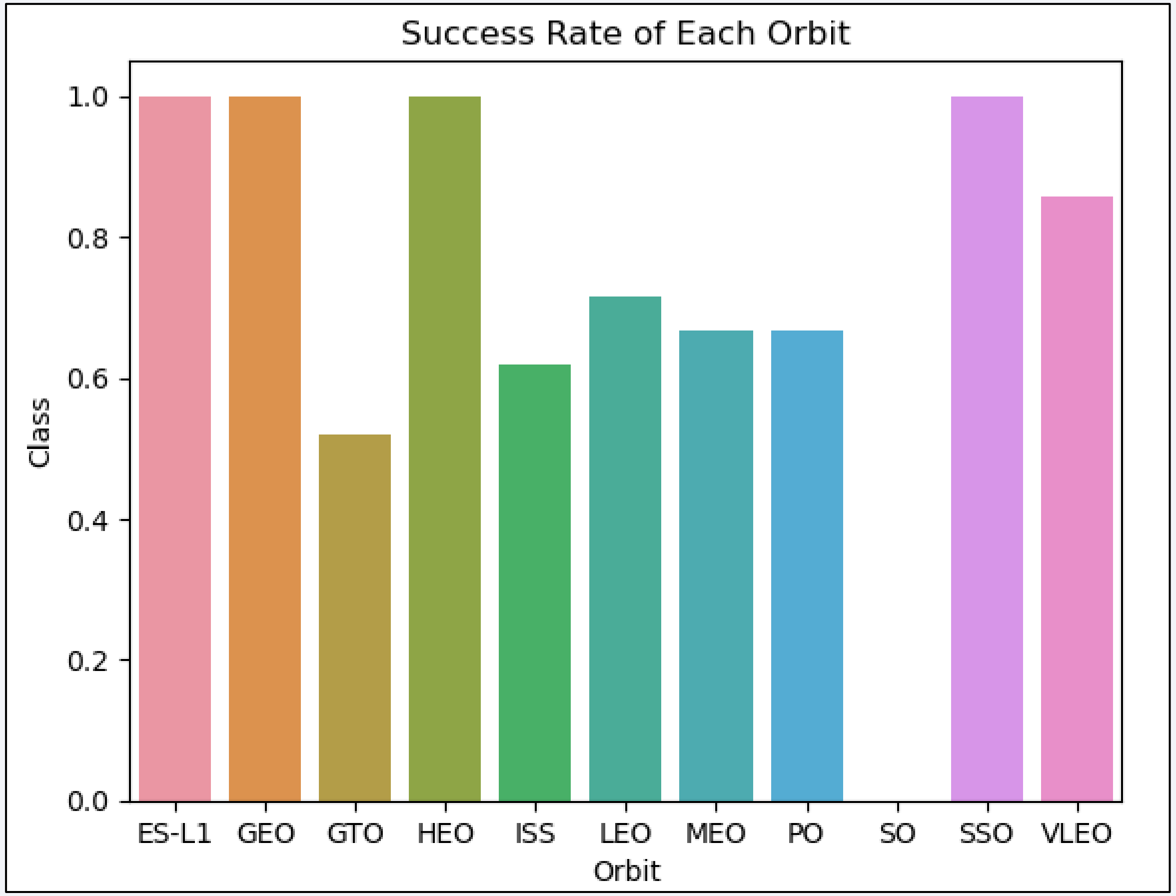
Payload vs. Launch Site



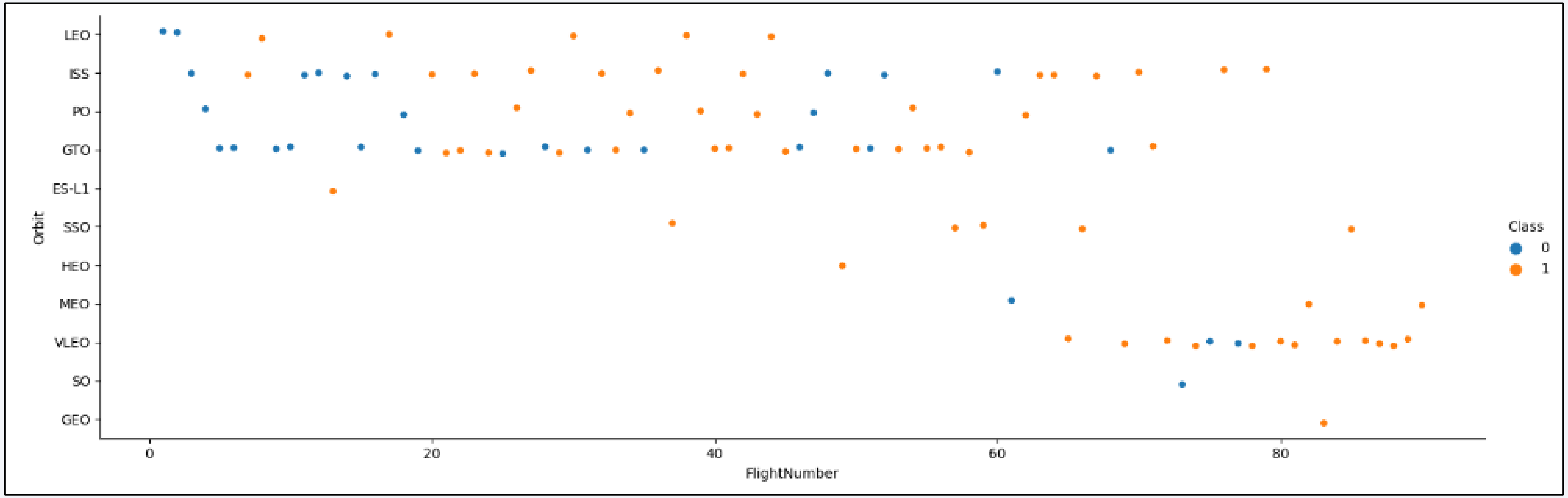
- Most of flight launch with payload less than 8000 kg.
- Only CCAFS SLC 40 and KSC LC 39A site have launched flights with payload mass over 14000 kg.

Success Rate vs. Orbit Type

- Orbits with the most success rate are ES-L1, GEO, HEO, and SSO.
- SO orbit has 0 percent success rate.

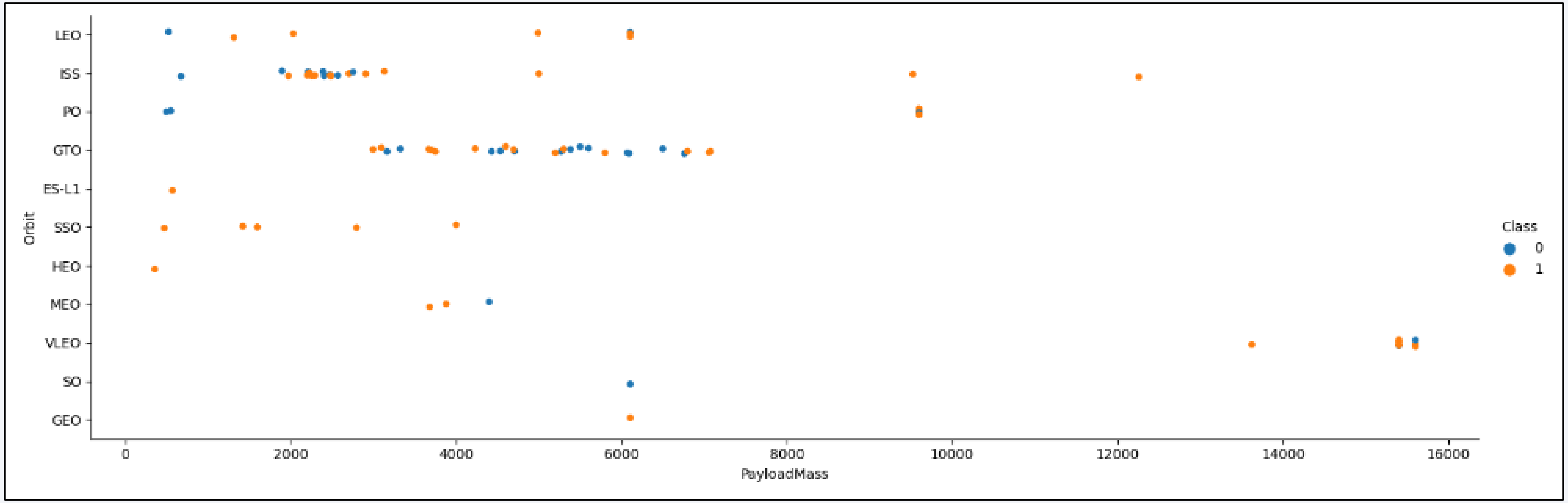


Flight Number vs. Orbit Type



- Flight number above 60 shows more success for ISS.
- LEO, ISS, PO, GTO and VLEO have more flight than the others.

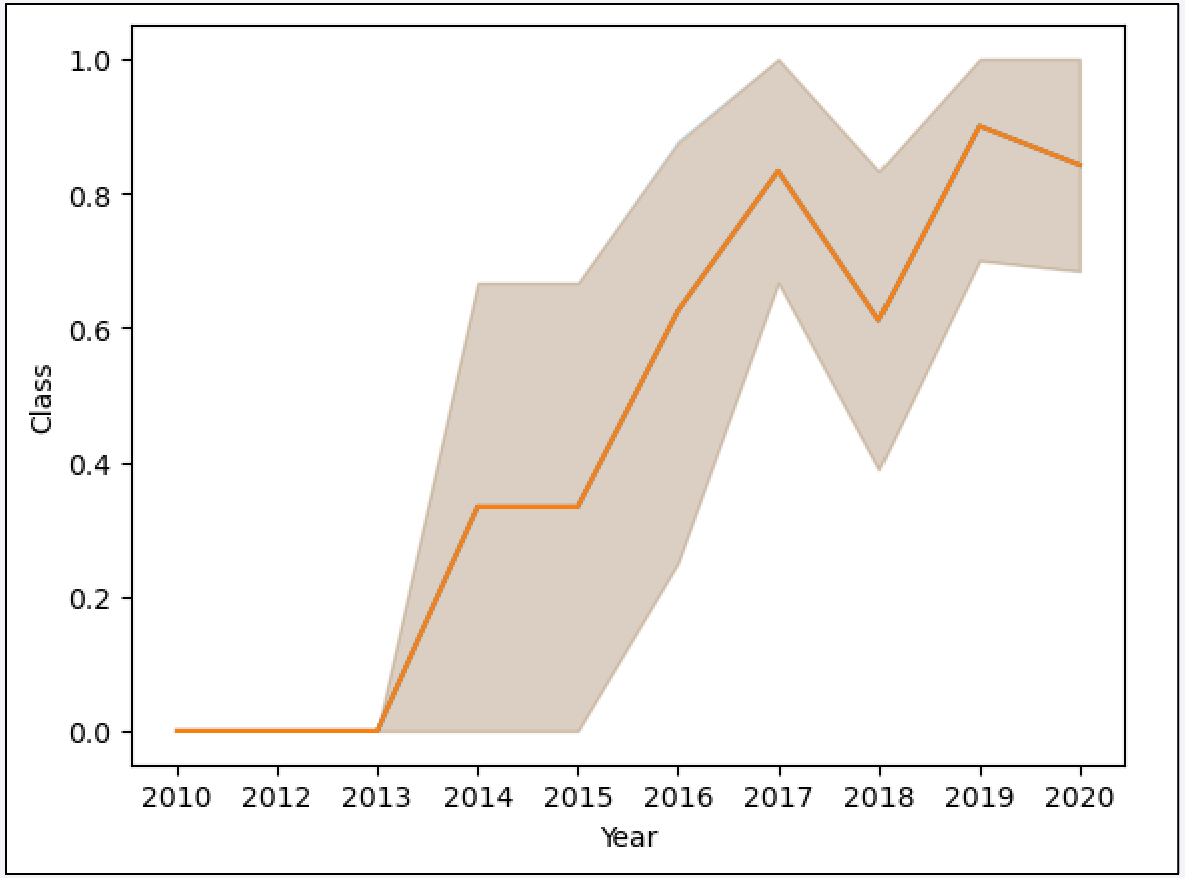
Payload vs. Orbit Type



- GTO orbit only has flights with payload between 2000 kg and 8000 kg.
- Flights with payload over 14000 kg only done on VLEO orbit.
- Most flights of ISS orbit only with payload of 2000 kg to 4000 kg.

Launch Success Yearly Trend

- Success rate increased since 2013 until 2020.



All Launch Site Names

- All launch sites names are obtain by querying a distinct names from SpaceX dataset

```
%%sql
SELECT DISTINCT LAUNCH_SITE FROM SPACEX;
```

```
 * ibm_db_sa://gmp23207:***@824dfd4d-99de
Done.
```

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

Launch Site Names Begin with 'CCA'

- Launch site with names start with CCA are obtain using LIKE method and limit to 5.

```
%%sql
SELECT * FROM SPACEX WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

```
 * ibm_db_sa://gmp23207:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
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	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 has been obtained by querying sum of payload mass but we set the condition to only query results of customer from ‘NASA (CRS)’.

```
%%sql
SELECT SUM(payload_mass__kg_) AS "Total Payload by NASA (CRS)" FROM SPACEX WHERE customer = 'NASA (CRS)';
```

```
 * ibm_db_sa://gmp23207:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.
Done.
```

Total Payload by NASA (CRS)
45596

Average Payload Mass by F9 v1.1

- Calculated using avg function to obtain average of payload mass and apply filter for only booster version of ‘F9 1.1’.

```
%%sql
SELECT AVG(payload_mass__kg_) AS "Average Payload Mass by Booster F9 v1.1"
FROM SPACEX
WHERE booster_version = 'F9 v1.1'
```

```
 * ibm_db_sa://gmp23207:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08
Done.
```

Average Payload Mass by Booster F9 v1.1
2928

First Successful Ground Landing Date

- Obtain the date of first successful landing on ground pad using min function.

```
%%sql
SELECT MIN(DATE) AS "First Successful Landing in Ground Pad" FROM SPACEX
WHERE landing__outcome = 'Success (ground pad)'
```

```
 * ibm_db_sa://gmp23207:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l
Done.
```

First Successful Landing in Ground Pad
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Obtained the list of booster version with payload between 4000 and 6000 kg that have success landing on drone ship by applying WHERE condition.

```
%%sql
SELECT booster_version FROM SPACEX
WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4000 AND 6000;
```

```
 * ibm_db_sa://gmp23207:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databas
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- Calculated the total of each mission outcome using COUNT() function and group them by the mission outcome.

```
%%sql
SELECT mission_outcome, COUNT(mission_outcome) AS "Total" FROM SPACEX GROUP BY mission_outcome;
```

```
 * ibm_db_sa://gmp23207:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.
Done.
```

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

Boosters Carried Maximum Payload

- Obtain the names of boosters that carried maximum payload which is 15600 kg after we applied max() function in sub query.

```
%%sql
SELECT booster_version, payload_mass__kg_ FROM SPACEX
WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM SPACEX)
ORDER BY booster_version;
```

```
 * ibm_db_sa://gmp23207:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2i
Done.
```

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

2015 Launch Records

- Query the booster version and launch site for failed landing at drone ship in 2015 by applying WHERE and YEAR() filter.

```
%%sql
SELECT booster_version, launch_site, landing__outcome FROM SPACEX
WHERE landing__outcome = 'Failure (drone ship)' AND YEAR(DATE) = 2015;
```

```
 * ibm_db_sa://gmp23207:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io9(
Done.
```

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)

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

- Query the landing outcome between given dates by applying COUNT() function and order them in descending order.

```
%%sql
SELECT landing__outcome, COUNT(landing__outcome) AS "Total" FROM SPACEX
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY landing__outcome
ORDER BY "Total" DESC;
```

```
 * ibm_db_sa://gmp23207:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l
Done.
```

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

All Launch Sites Locations

- All launch sites located near the west and east coastal lines

Launch sites marked with circles

Launch sites area are marked with circles and shown by the number of launch site within the area

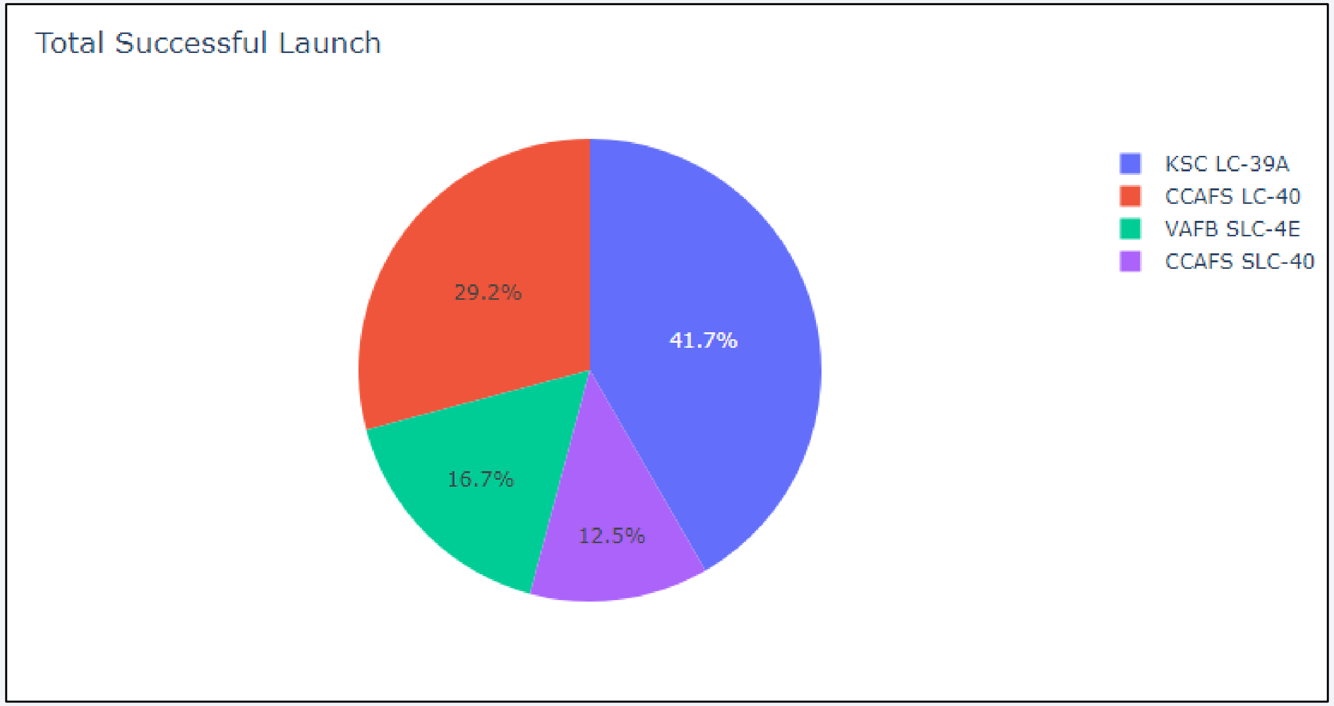
Launch Sites from proximities.

Distance from launch site to its proximities are shown with lines and marked with distance value.

Section 4

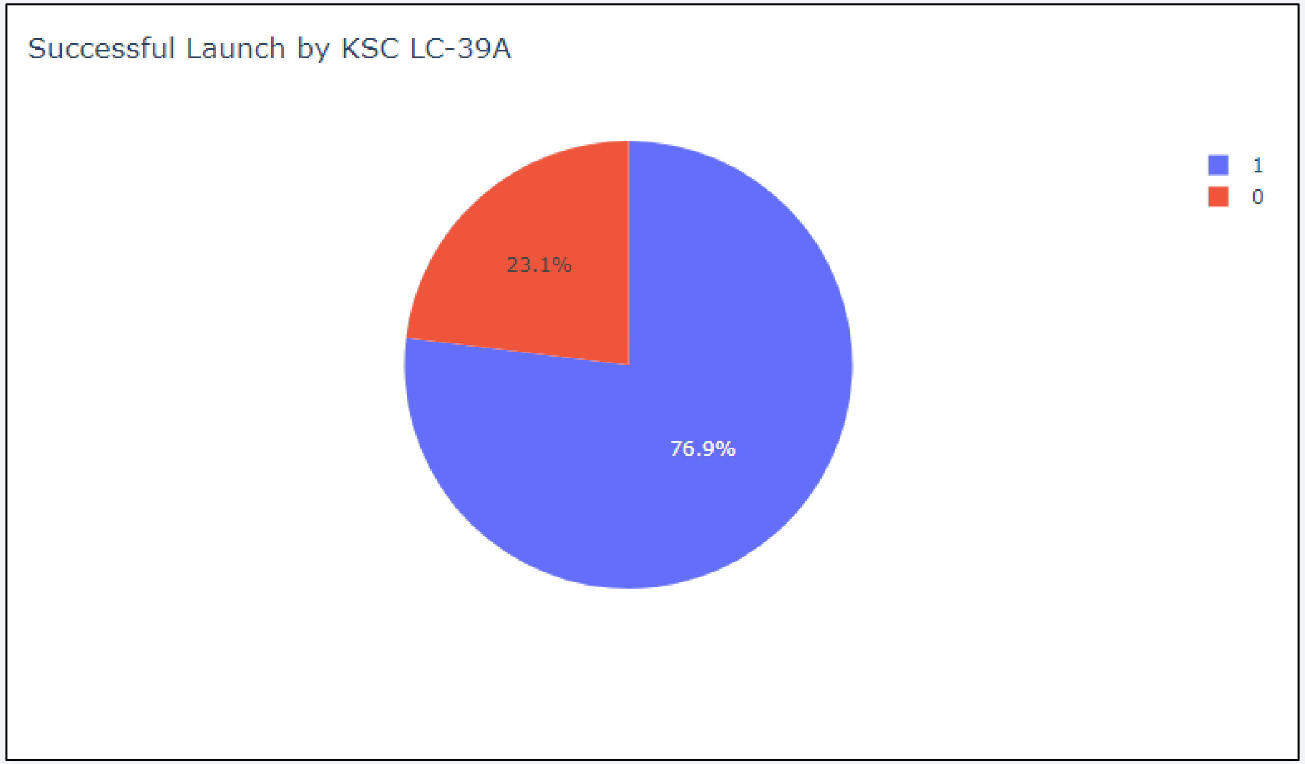
Build a Dashboard with Plotly Dash

Successful Launch for All Sites



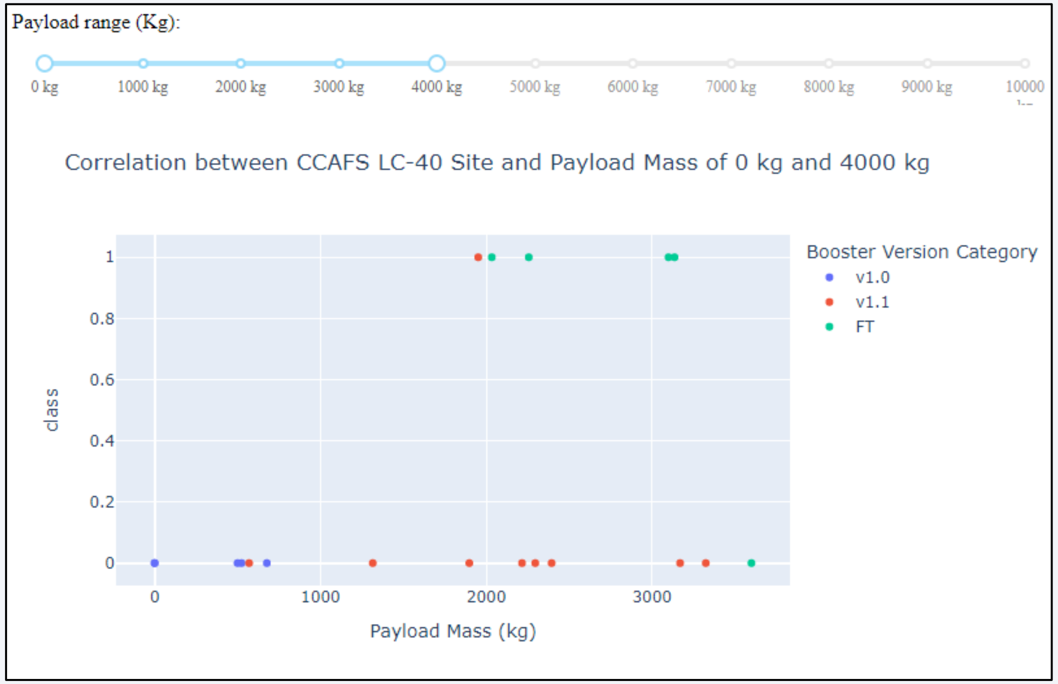
- KSC LC-39A has the most successful launch followed by CCAFS LC-40.

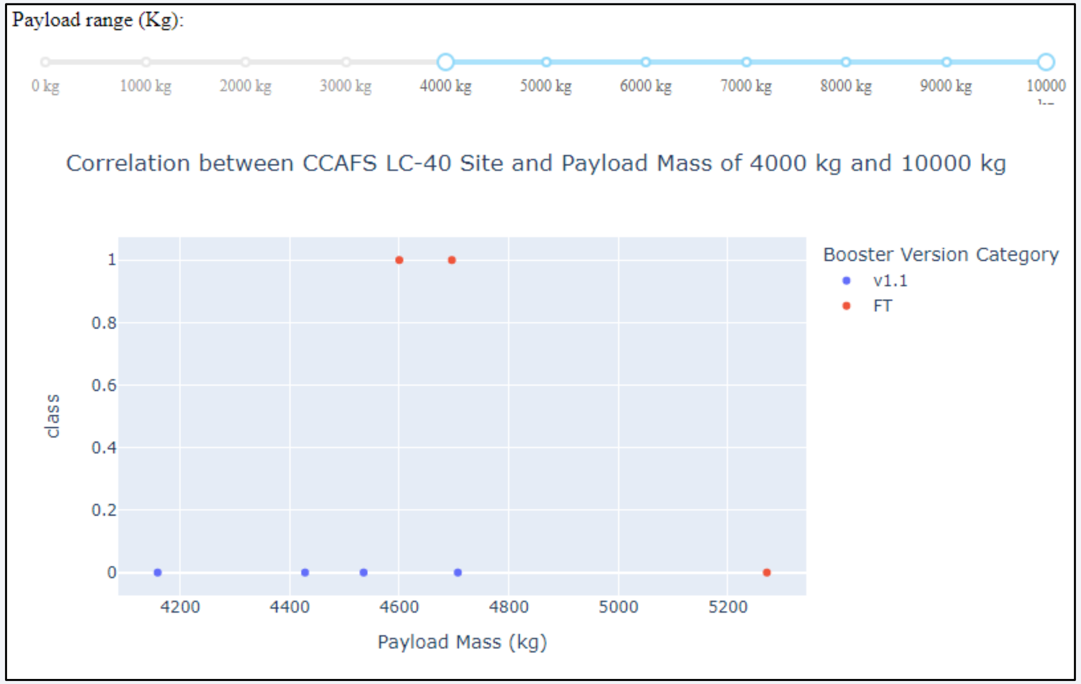
Highest Success Launch Site



- KSC LC-39A has 76.9% success rate which is the highest among other launch site.

Payload vs Launch Outcome





- Booster version FT has higher success rate.
- There is no significance correlation between payload mass and launch outcome for CCAFS LC-40 site.

Section 5

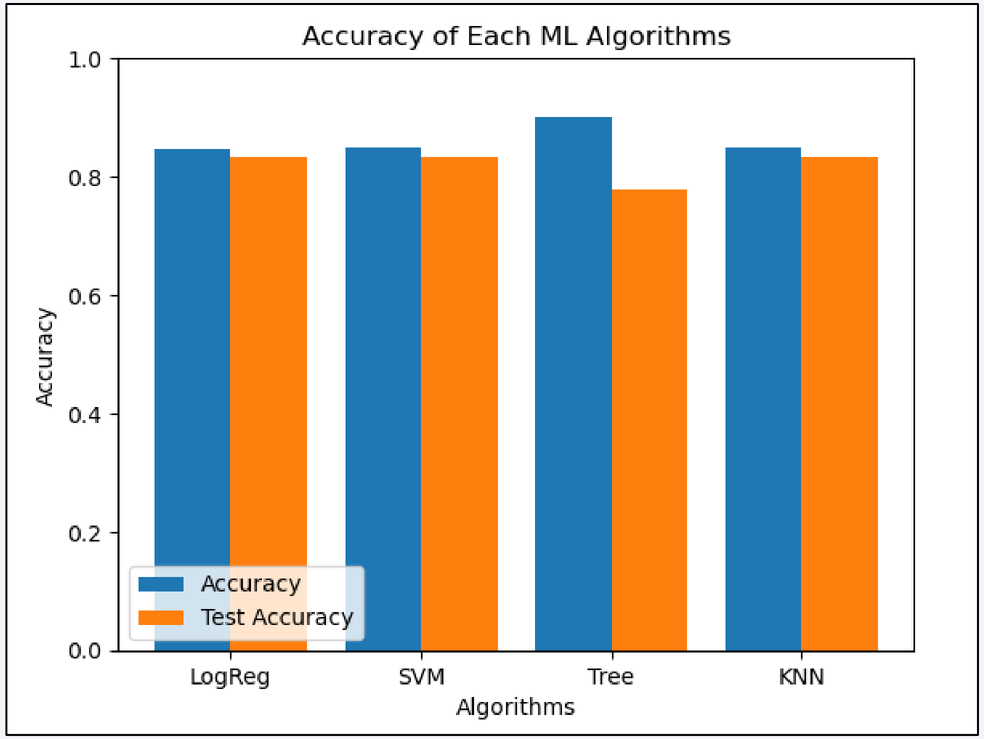
Predictive Analysis (Classification)

Classification Accuracy

Decision tree has the best accuracy with accuracy of 0.9.

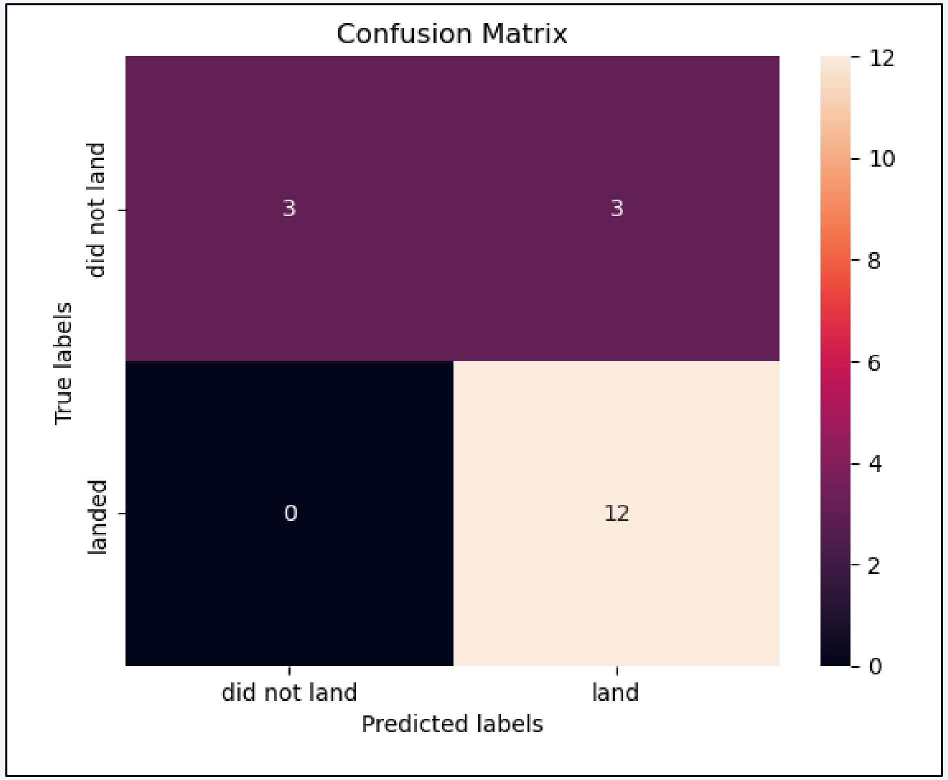
```
models = {'K Nearest Neighbors':knn_cv.best_score_,
              'Decision Tree':tree_cv.best_score_,
              'Logistic Regression':logreg_cv.best_score_,
              'Support Vector Machine': svm_cv.best_score_}

best_method = max(models, key=models.get)
print('Best model is: {} with a score of {}'.format(best_method,models[best_method]))
```

```
Best model is: Decision Tree with a score of 0.9
```



Confusion Matrix

- We can see that decision tree has the highest true positive output.



Conclusions

- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate compared to other orbits.
- Most of flights launched from CCAFS LC-40
- The best predictive analysis method is decision with the highest accuracy.

Thank you!