Data Science Capstone project

Yi Liu

2021/09/02

Outline

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

Executive Summary

- **Summary of methodologies**
  - Define the business context and question to answer
  - Data collection from API and web scraping
  - Exploratory Data Analysis using SQL, data visualization
  - Interactive Data Analytics with Folium and Plotly Dash
  - Build, tune and evaluate the machine learning models
  - Conclusion and future developments

- **Summary of all results**
- Exploration:
  - Four unique launch sites: CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E, all are in close proximity to railways, highways, coastline and equator line, while keeping some distance away from cities.
- Important features and relationship to launch outcome:
  - Flight number, booster version, payload mass, orbit type, launch site, year of launch are important features that can be used to predict launch outcomes, while there are also interactions between different variables
- Predictive models
  - Similar test set accuracy (83.3%) and confusion matrix
  - Main issue is False Positive (the model predicted outcome is success but true outcome is failure)

Introduction

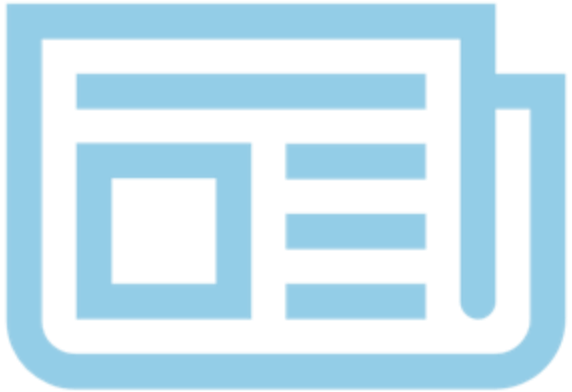
- Project background and context

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 want to find answers

Predict if Space X Falcon 9 first stage will land successfully

Methodology



- Data collection methodology:
  - Requests from SpaceX API and clean the requested data
  - Web scraping Falcon 9 and Falcon Heavy Launches Records from Wikipedia
- Perform data wrangling
  - Perform exploratory Data Analysis and determine Training Labels
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Create response variable Y, standardize features X -> train test split
  - Classification models: Logistic Regression, Support Vector Machine, Decision Tree Classifier, K Nearest Neighbors
  - Model Training and evaluation: Hyperparmeter tuning with GridSearch CV to find best parameters using training set, evaluate with accuracy score and confusion matrix using test set

Methodology

Data collection

- Data Source
  - Requests from SpaceX API and clean the requested data
  - Web scraping Falcon 9 and Falcon Heavy Launches Records from Wikipedia

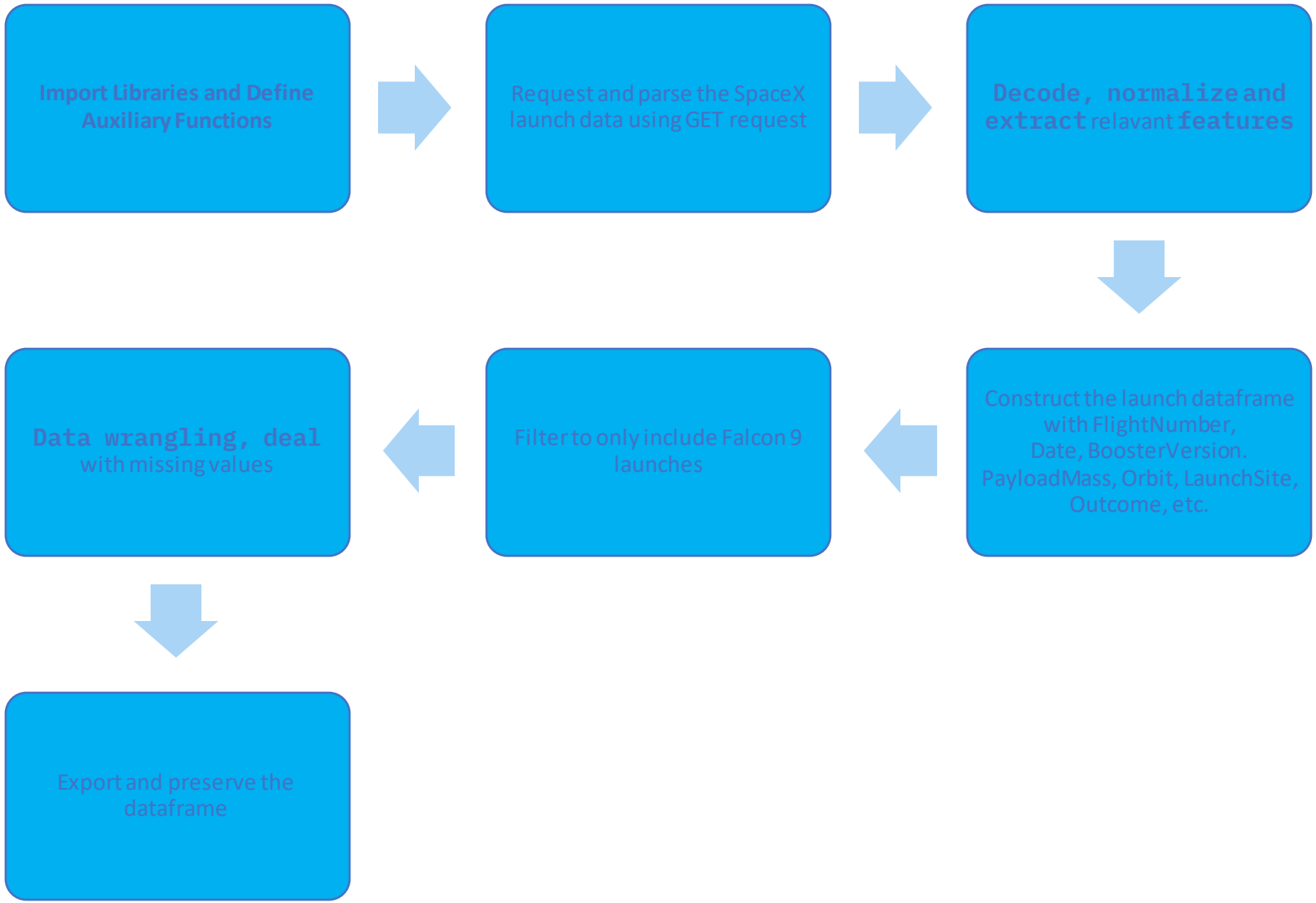
- Data Extraction and Preprocessing

Data collection – SpaceX API

- Request to the SpaceX API
- Clean the requested data

GitHub URL of the completed SpaceX API calls notebook for reference:

https://github.com/liuyiemily/IBM_Data-Science-Project/blob/main/Data%20Collection%20API.ipynb

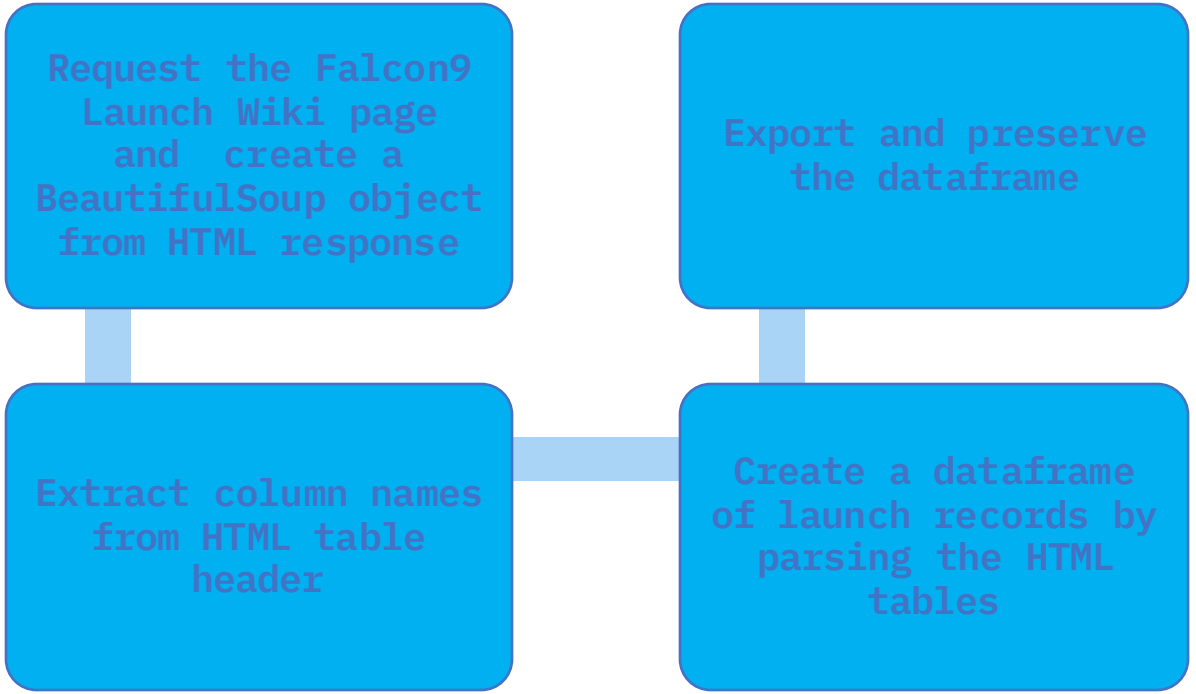


Data collection – Web scraping

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

GitHub URL of the completed SpaceX Web scraping notebook for reference:

https://github.com/liuyiemily/IBM_Data-Science-Project/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb

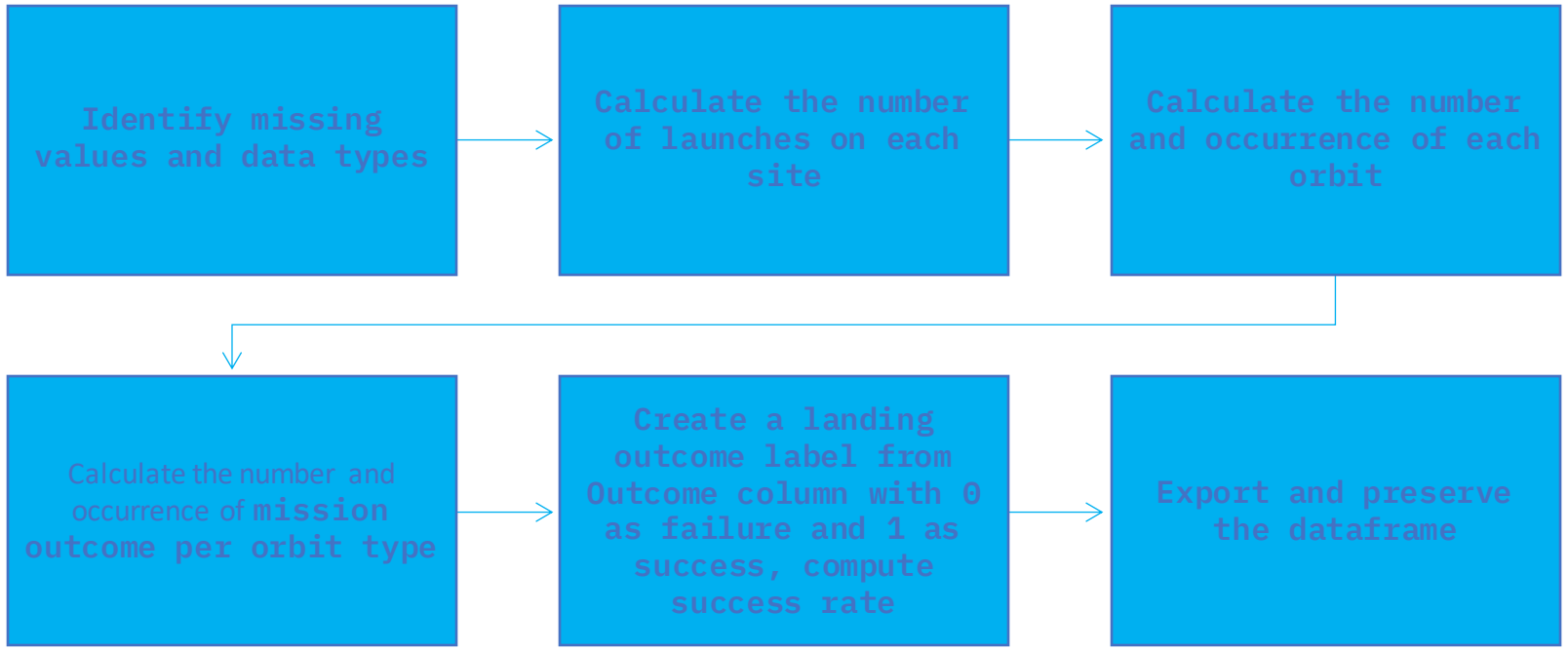


Data wrangling

- Data Processing and Analysis:
  - Exploratory Data Analysis
  - Determine Training Labels

- GitHub URL of data wrangling notebooks for reference:

https://github.com/liuyiemily/IBM_Data-Science-Project/blob/main/EDA.ipynb



EDA with data visualization

1. EDA: Summary of charts and rationale
   - Scatter point chart: show relationship between numerical and one or more categorical variables
     - FlightNumber vs PayloadMass -> launch outcome
     - FlightNumber vs LaunchSite -> launch outcome
     - PayloadMass vs LaunchSite -> launch outcome
     - FlightNumber vs Orbit type -> launch outcome
     - PayloadMass vs Orbit type -> launch outcome
   - Bar chart: Show magnitude on y-axis and categories on x-axis, good to check relationship between success rate and orbit type (categorical variable)
   - Line chart: Show the relationship between two numerical variables, good to visualize trend of time series data, here used to visualize the launch success yearly trend
2. Feature Selection and Feature Engineering
   - Select predictive features / variables
   - One-hot-encoding categorical variables
   - Cast all numeric variables to float64
3. Export and preserve data

- GitHub URL of EDA with data visualization notebook for reference:
  - https://github.com/liuyiemily/IBM_Data-Science-Project/blob/main/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- Summary of SQL queries
  1. *Display the names of the unique launch sites in the space mission*
     - %sql SELECT DISTINCT(LAUNCH_SITE) FROM SPACEXDATASET;
  2. *Display 5 records where launch sites begin with the string 'CCA'*
     - %sql SELECT * FROM SPACEXDATASET WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
  3. *Display the total payload mass carried by boosters launched by NASA (CRS)*
     - %sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_NASA_PAYLOAD FROM SPACEXDATASET WHERE CUSTOMER = 'NASA (CRS)';
  4. *Display average payload mass carried by booster version F9 v1.1*
     - %sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_F911_PAYLOAD FROM SPACEXDATASET WHERE BOOSTER_VERSION = 'F9 v1.1';
  5. *List the date when the first succesful landing outcome in ground pad was achieved*
     - %sql SELECT MIN(DATE) AS FIRST_GROUND_PAD FROM SPACEXDATASET WHERE LANDING__OUTCOME = 'Success (ground pad)';
  6. *List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000*
     - %sql SELECT BOOSTER_VERSION FROM SPACEXDATASET WHERE LANDING__OUTCOME = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;

EDA with SQL

- Summary of SQL queries (continued)

7. List the total number of successful and failure mission outcomes

- %sql SELECT COUNT(*) AS SUCCESS_MISSION FROM SPACEXDATASET WHERE mission_outcome LIKE '%Success%';
- %sql SELECT COUNT(*) AS FAILURE_MISSION FROM SPACEXDATASET WHERE mission_outcome LIKE '%Failure%';

8. List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

- %sql SELECT DISTINCT(BOOSTER_VERSION) FROM SPACEXDATASET WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXDATASET);

9. List the failed landing_outcomes in drone ship, their booster versions, and launch site names for the in year 2015

- %sql SELECT MONTH(DATE) AS MONTH_2015, LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXDATASET WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND YEAR(DATE) = 2015;

10. 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 descending order

- %sql SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS FREQ FROM SPACEXDATASET WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING__OUTCOME ORDER BY COUNT(LANDING__OUTCOME) DESC;

- GitHub URL of EDA with SQL notebook for reference:
  - https://github.com/liuyiemily/IBM_Data-Science-Project/blob/master/EDA%20with%20SQL.ipynb

Build an interactive map with Folium

- Summary of map objects created and added to a folium map
  1. Create and add Circle and Marker for each launch site on map
     - Want to visualize the location of launch site and its proximities on map
  2. Create a MarkerCluster and add a Marker to MarkerCluster for each launch
     - Launch only happen at one of the four launch sites. Marker clusters is a good way to simplify a map containing many markers having the same coordinate
  3. MousePosition: to get corordinate of proximity on map
  4. PolyLine: to draw a line between a launch site and its proximity on map

- GitHub URL of interactive map with Folium map for reference:
  - https://github.com/liuyiemily/IBM_Data-Science-Project/blob/main/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- Summary of graphs and interactions added to the dashboard
  - Title of the dashboard
  - Launch Site Drop-down Input Component
  - `'success-pie-chart'` based on selected site dropdown
    - Visualize the relationship between success launches and launch site
  - Range Slider to Select Payload
  - `'success-payload-scatter-chart'` scatter plot
    - Visualize the relationship between success launches and payload for selected site

- Rationale for the plots and interactions

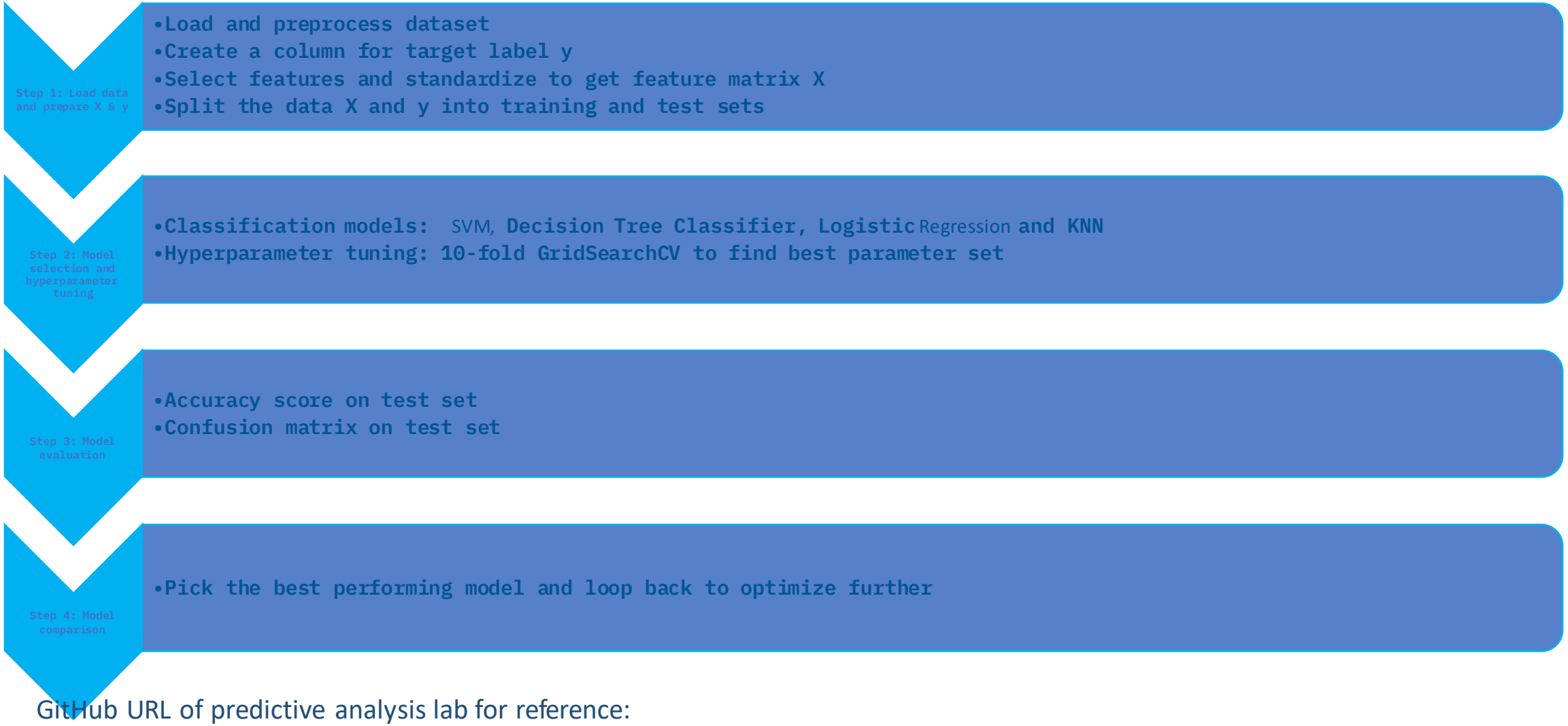
  Use visualization to answer the following questions:

  1) Which site has the largest successful launches?
  2) Which site has the highest launch success rate?
  3) Which payload range(s) has the highest launch success rate?
  4) Which payload range(s) has the lowest launch success rate?
  5) Which F9 Booster version (`v1.0`, `v1.1`, FT, B4, B5, etc.) has the highest launch success rate?

- GitHub URL of Plotly Dash lab for reference:
  - https://github.com/liuyiemily/IBM_Data-Science-Project/blob/main/7.%20spacex_dash_app.py

Predictive analysis (Classification)

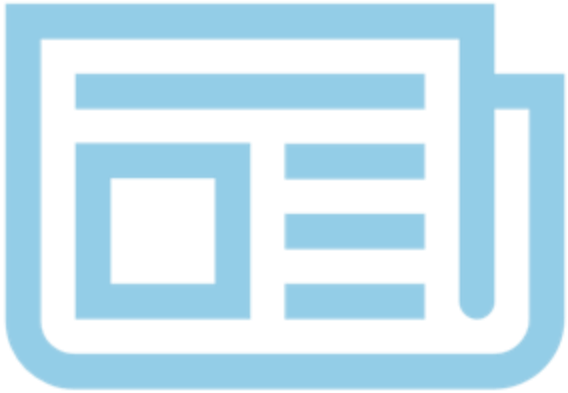
- Summary of model development, hyperparameter tuning and evaluation



GitHub URL of predictive analysis lab for reference:

- https://github.com/liuyiemily/IBM_Data-Science-Project/blob/main/8.%20Machine%20Learning%20Prediction.ipynb

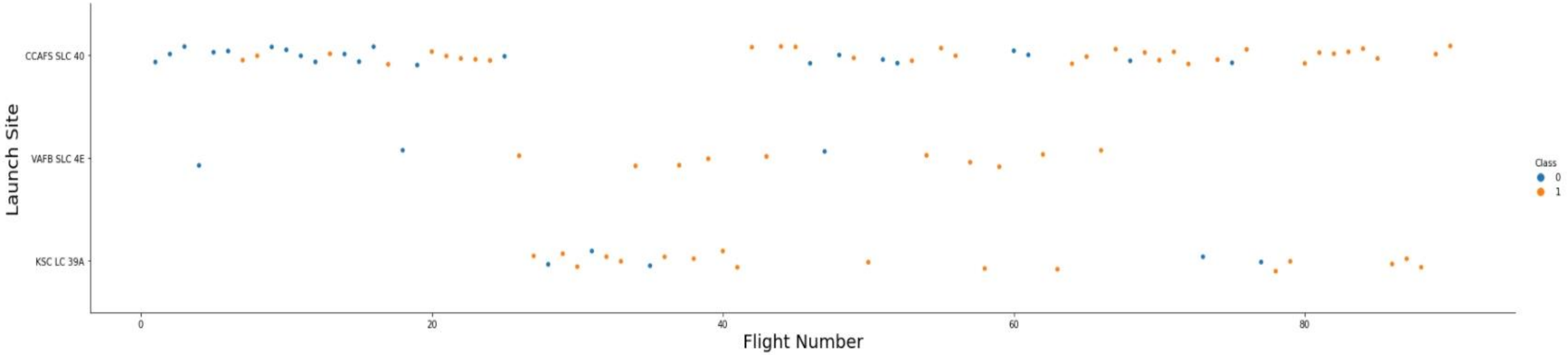
Results



- Exploratory data analysis results
  - EDA with Data Visualization
  - EDA with SQL
- Interactive analytics demo in screenshots
  - Interactive visual analytics with Folium
  - Interactive dashboard with Plotly Dash
- Predictive analysis results
  - Train and test set accuracy
  - Confusion matrix

EDA with Visualization

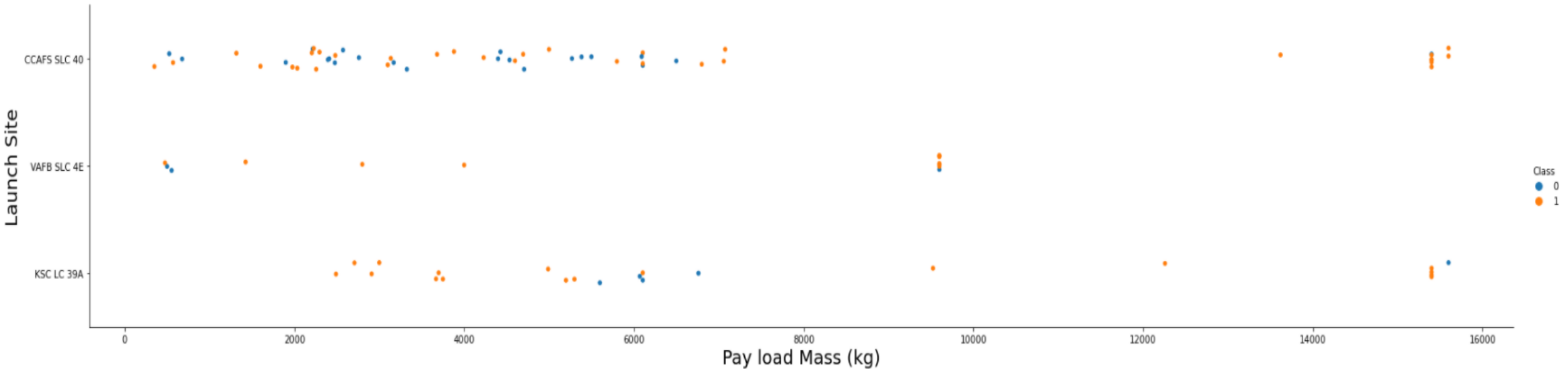
Flight Number vs. Launch Site



Explanation: From the plot, we can see that

- As the flight number increases, the first stage is more likely to land successfully.
- Different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A has a success rate of around 77%.

Payload vs. Launch Site



Explanation:

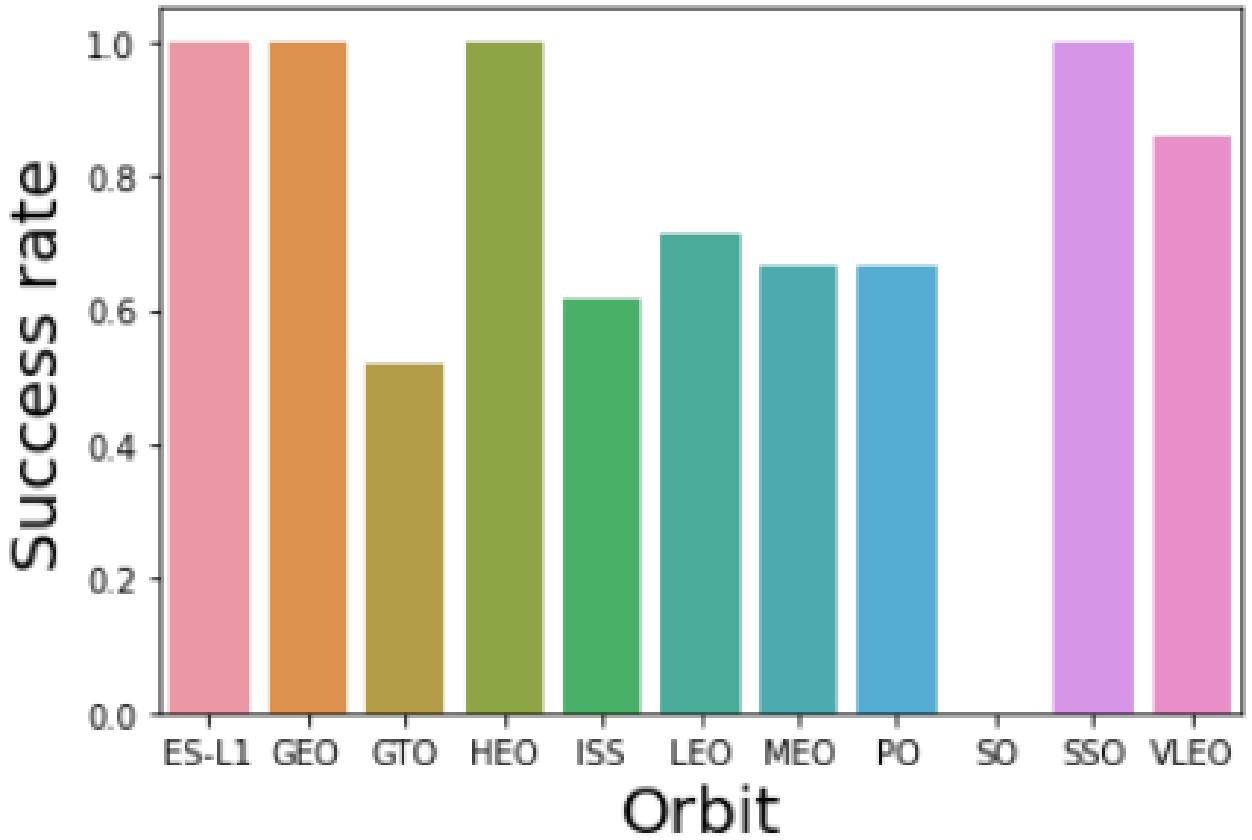
- VAFB SLC 4E tends to have small pay load mass, KSC LC 39A is more spread out and skewed towards heavy payload mass.
- CCAFS SLC 40 majorly associates with small payload mass, but also has heavy payload mass, and it seems that heavy payload mass in CCAFS SLC 40 has positive impact in success.

Success rate vs. Orbit type

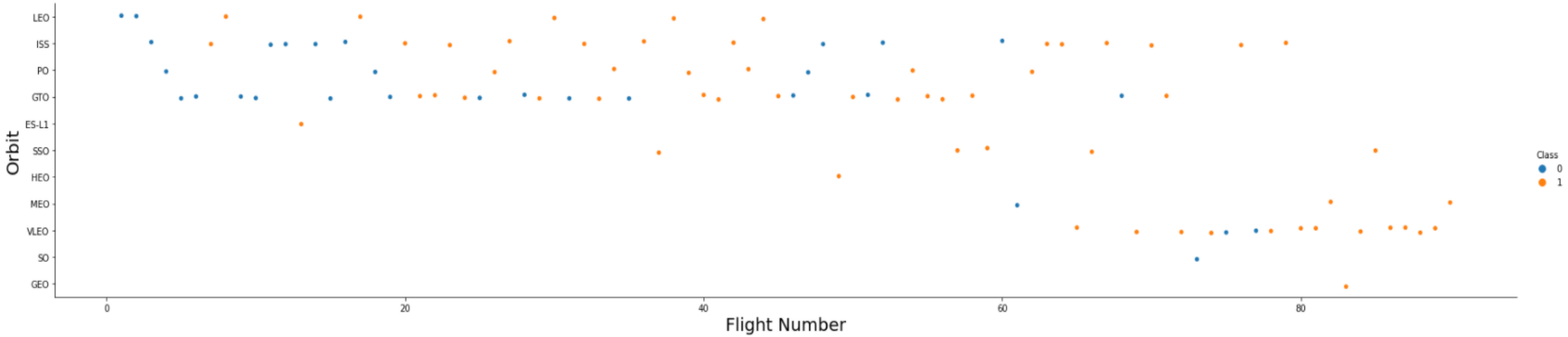
Explanation:

- ES-L1, GEO, HEO, SSO have the highest success rate of 100%, while SO have success rate of 0, and GTO success rate below 60%

	Orbit	Success Rate
0	ES-L1	1.000000
1	GEO	1.000000
3	HEO	1.000000
9	SSO	1.000000
10	VLEO	0.857143
5	LEO	0.714286
6	MEO	0.666667
7	PO	0.666667
4	ISS	0.619048
2	GTO	0.518519
8	SO	0.000000



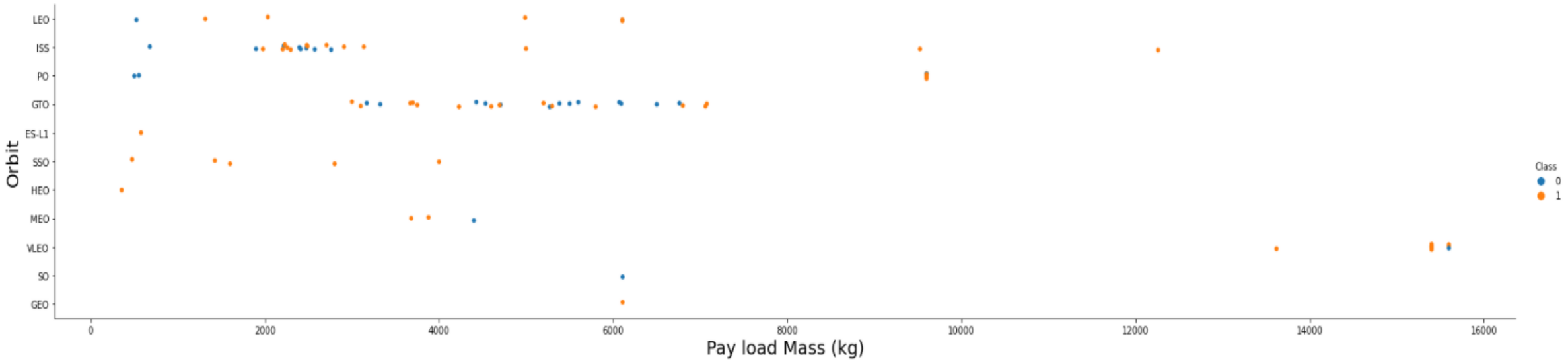
Flight Number vs. Orbit type



Explanation:

- in the LEO orbit, the Success rate appears to be positively 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



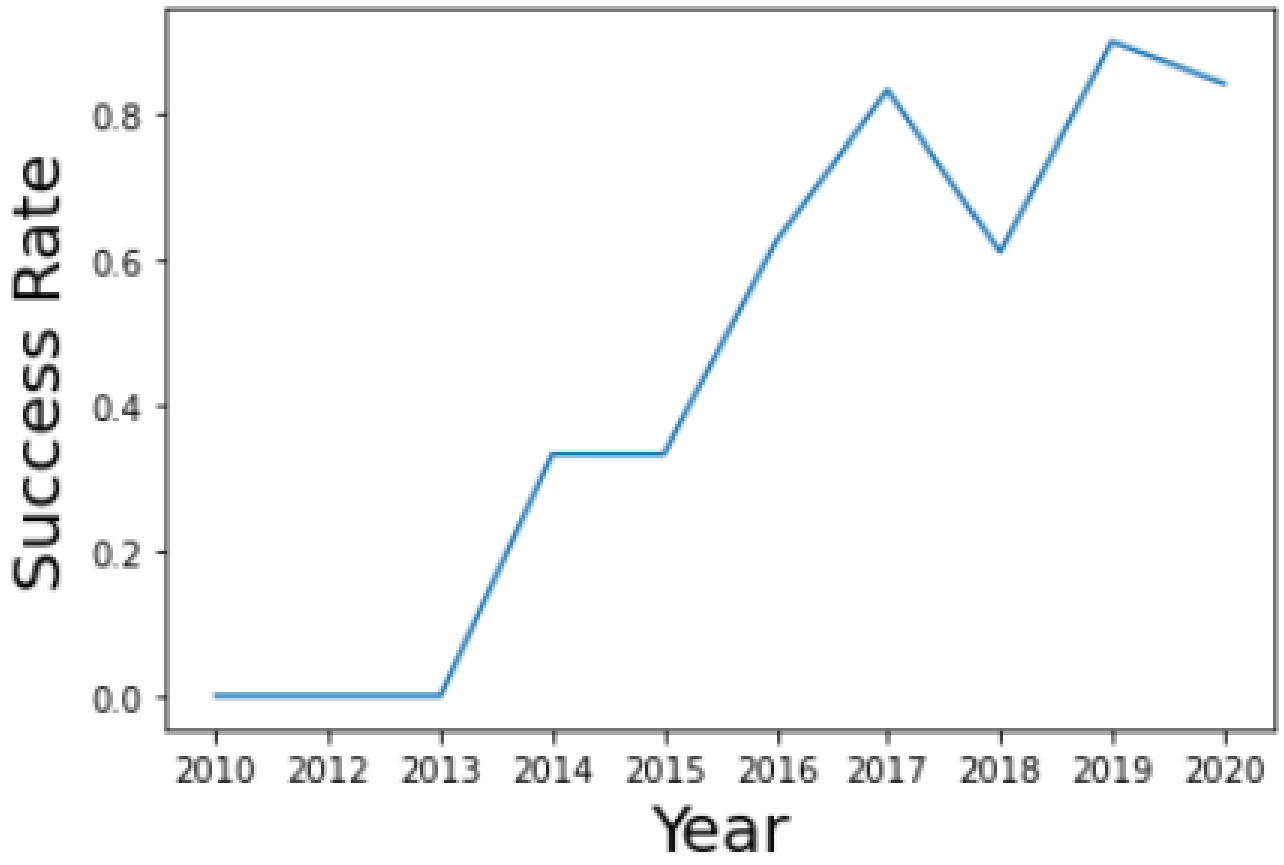
Explanation:

- Heavy payloads have a negative influence on VLEO and MEO orbits and positive on GTO, PO, Polar LEO (ISS) and LEO orbits.

Launch success yearly trend

Explanation:

- observe that the success rate since 2013 kept increasing till 2020, with a small dip in 2018



EDA with SQL

All launch site names

- Find the names of the unique launch sites
  - Four unique launch sites: CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E

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

```
%sql SELECT DISTINCT(LAUNCH_SITE) FROM SPACEXDATASET;
```

```
 * ibm_db_sa://rgj46413:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

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

Launch site names begin with `CCA`

- Display 5 records where launch sites begin with the string 'CCA'
  - The first 5 records are with launch sites 'CCAFS LC-40'

Display 5 records where launch sites begin with the string 'CCA'

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

```
 * ibm_db_sa://rgj46413:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/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

- Calculate the total payload carried by boosters from NASA
  - total payload carried by boosters from NASA is 45,596 KG

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

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_NASA_PAYLOAD FROM SPACEXDATASET WHERE CUSTOMER = 'NASA (CRS)';
```

```
 * ibm_db_sa://rgj46413:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

total_nasa_payload
45596

Average payload mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
  - average payload mass carried by booster version F9 v1.1 is 2,928 KG

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

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_F911_PAYLOAD FROM SPACEXDATASET WHERE BOOSTER_VERSION = 'F9 v1.1';
```

```
 * ibm_db_sa://rgj46413:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

avg_f911_payload
2928

First successful ground landing date

- Find the date when the first successful landing outcome in ground pad
  - the first successful landing in ground pad was on '2015-12-22'

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%sql SELECT MIN(DATE) AS FIRST_GROUND_PAD FROM SPACEXDATASET WHERE LANDING__OUTCOME = 'Success (ground pad)';
```

```
 * ibm_db_sa://rgj46413:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

first_ground_pad
2015-12-22

Successful drone ship landing with payload between 4000 and 6000

- List the names of boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - 'F9 FT B1022', 'F9 FT B1026', 'F9 FT B1021.2', 'F9 FT B1031.2'

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%%sql
SELECT BOOSTER_VERSION
FROM SPACEXDATASET
WHERE LANDING__OUTCOME = 'Success (drone ship)'
    AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

```
 * ibm_db_sa://rgj46413:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
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

- Calculate the total number of successful and failure mission outcomes
  - total number of successful mission outcomes is 100
  - total number of successful mission outcomes is 1

List the total number of successful and failure mission outcomes

```
%sql SELECT COUNT(*) AS SUCCESS_MISSION FROM SPACEXDATASET WHERE mission_outcome LIKE '%Success%';
```

```
 * ibm_db_sa://rgj46413:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

success_mission
100

```
%sql SELECT COUNT(*) AS FAILURE_MISSION FROM SPACEXDATASET WHERE mission_outcome LIKE '%Failure%';
```

```
 * ibm_db_sa://rgj46413:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

failure_mission
1

Boosters carried maximum payload

- List the names of the booster which have carried the maximum payload mass
  - 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

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%%sql
SELECT DISTINCT(BOOSTER_VERSION)
FROM SPACEXDATASET
WHERE PAYLOAD_MASS__KG_ =
(SELECT MAX(PAYLOAD_MASS__KG_)
FROM SPACEXDATASET);
```

```
 * ibm_db_sa://rgj46413:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databas
Done.
```

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

- List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015
  - Two records, one in 2015-01, Failure (drone ship), F9 v1.1 B1012, CCAFS LC-40; the other in 2015-04, Failure (drone ship), F9 v1.1 B1015, CCAFS LC-40

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

```
%%sql
SELECT MONTH(DATE) AS MONTH_2015, LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE
FROM SPACEXDATASET
WHERE LANDING__OUTCOME = 'Failure (drone ship)'
    AND YEAR(DATE) = 2015;
```

```
 * ibm_db_sa://rgj46413:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

]:

month_2015	landing__outcome	booster_version	launch_site
1	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
4	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Rank success count between 2010-06-04 and 2017-03-20

- Rank the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order.
  - 10 no attempt, followed by 5 failure (drone ship), 5 success (drone ship), see details below

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 descending order

```
%%sql
SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS FREQ
FROM SPACEXDATASET
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING__OUTCOME
ORDER BY COUNT(LANDING__OUTCOME) DESC;
```

```
 * ibm_db_sa://rgj46413:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

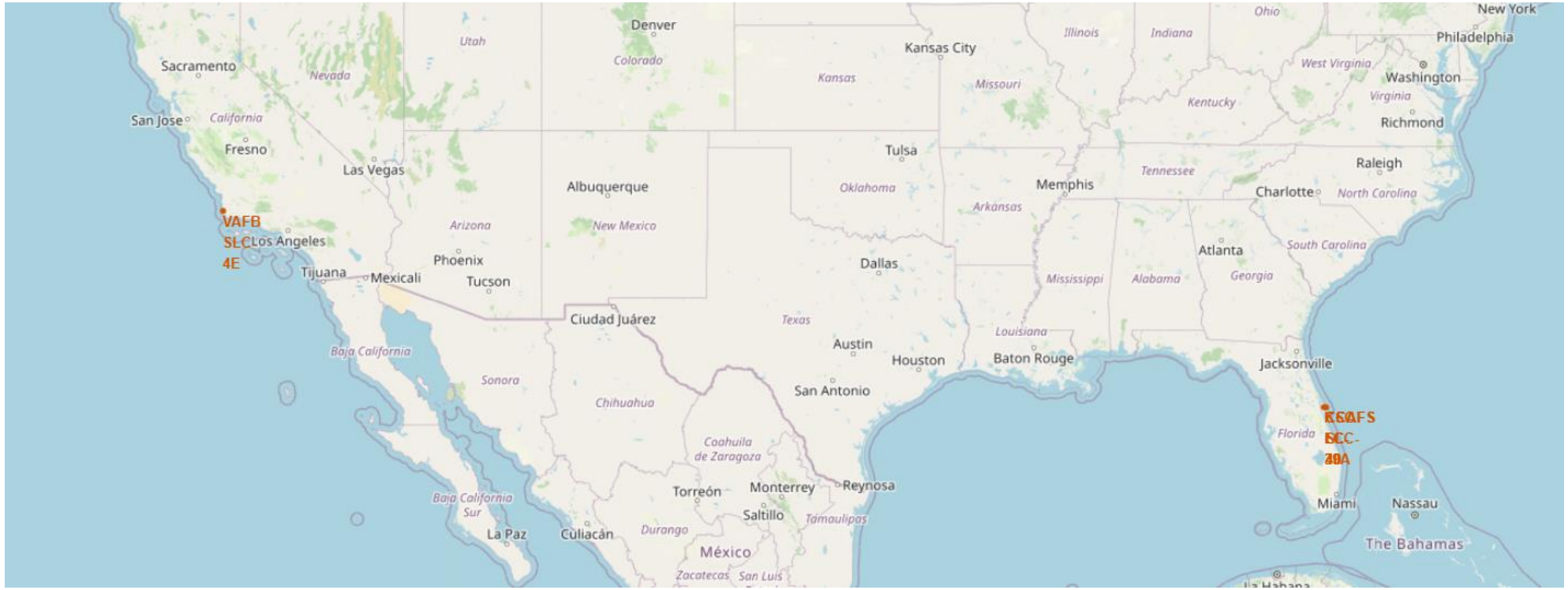
]:

landing__outcome	freq
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

Interactive map with Folium

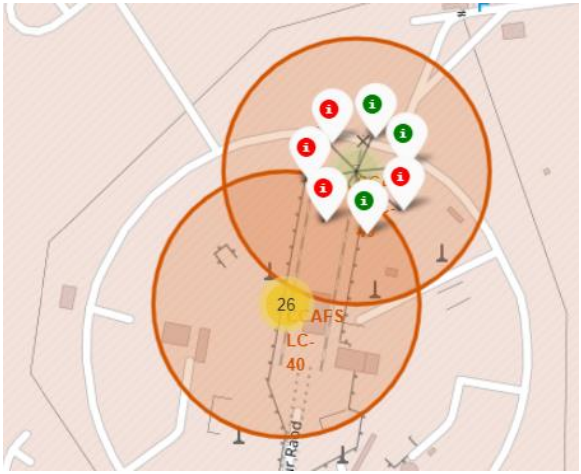
Launch site Locations

- See the screenshot of all launch sites’ location markers on a global map below
- Finding: all 4 launch sites are close to coast and Equator line, 3 ('CCAFS LC-40', 'CCAFS SLC-40', 'KSC LC-39A') are close to each other, as in the bottom right of the map, and 'VAFB SLC-4E' is in the left side



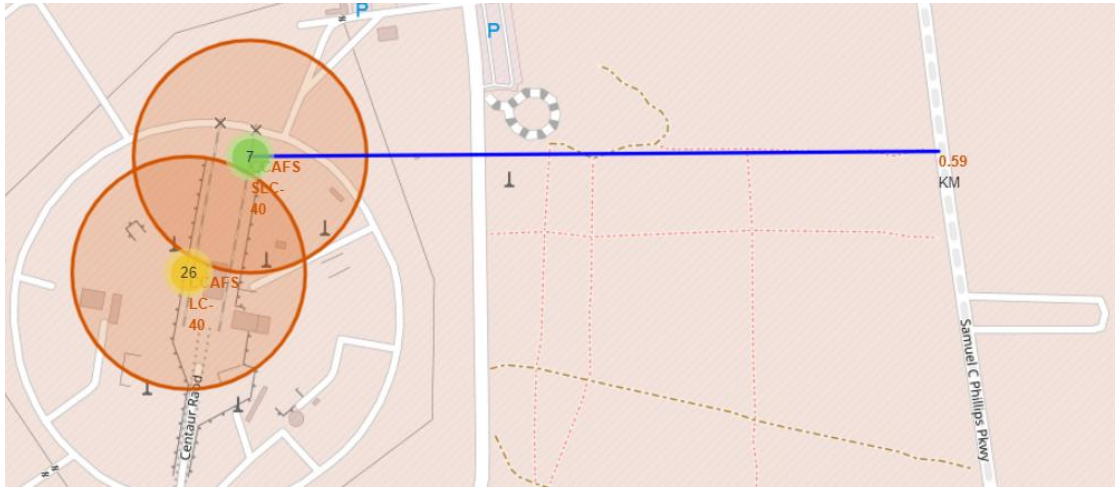
Mark the success/failure for each launch site

- See the screenshot of success/failure marks for each launch site
- Finding: 'KSC LC-39A' has relatively high success rate, with 10 successes and 3 failures, success rate of 77%, while 'CCAFS SLC-40' has 3 successes and 4 failures, success rate of 43%, followed by 'VAFB SLC-4E' with 4 sucesses and 6 failures and success rate of 40%. The worst one is 'CCAFS LC-40', which has 7 successes and 19 failures and success rate of 27%



Distances between a launch site to its proximities

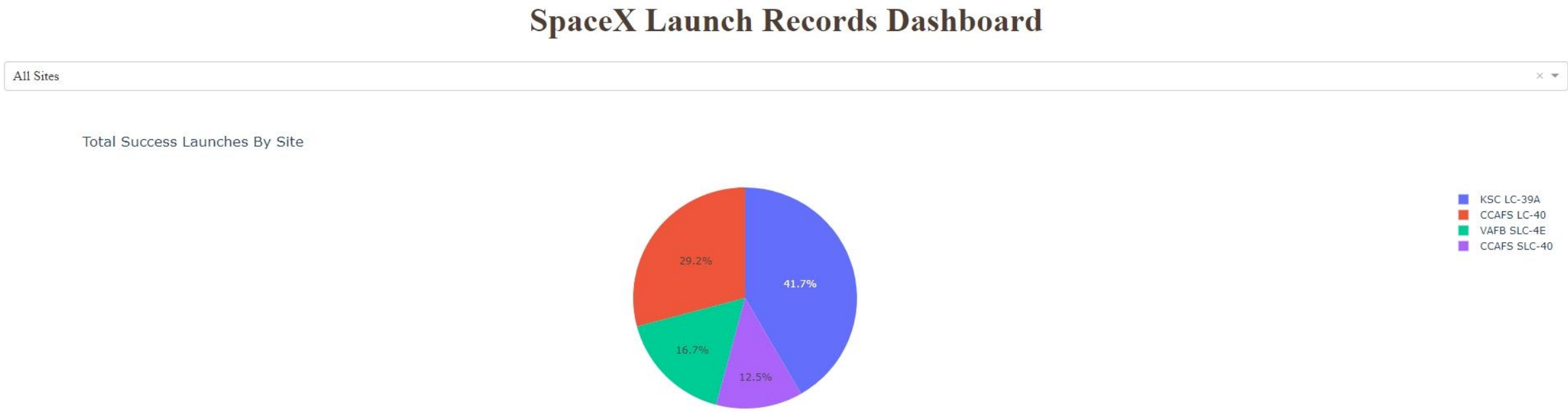
- See the screenshot of 'CCAFS SLC-40' to its closet railway (left graph) and 'CCAFS LC-40' to its cloest coastline (right graph), with distance calculated displayed
- Finding:
  - As shown from the map, all launch sites are in close proximity to railways, highways and coastline, while keeping some distance away from cities.
  - My guess is that, it's easier to ship required materials when it's close to transportation, while it helps to control the damage in case of failures in launching when it's close to coastline and away from cities.



Build a Dashboard with Plotly Dash

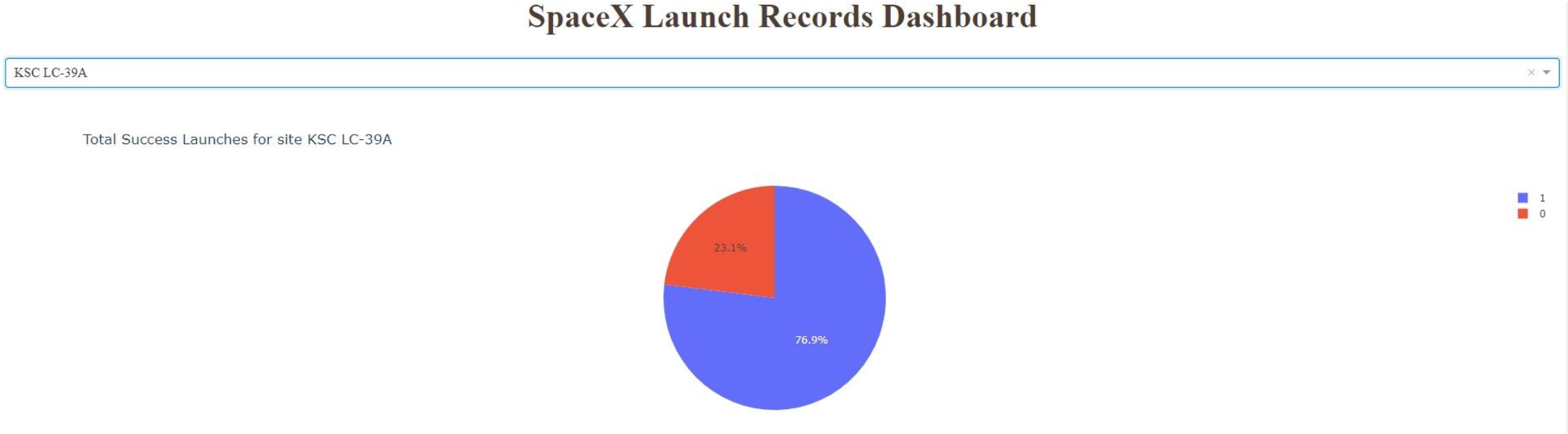
Total Success Launches By Site

- Screenshot of launch success count for all sites in a pie chart
- Findings: 'KSC LC-39A' has the largest successful launches (~42%), followed by 'CCAFS LC-40' (~29%), 'VAFB SLC-4E' (~17%), lastly 'CCAFS SLC-40' (12.5%)



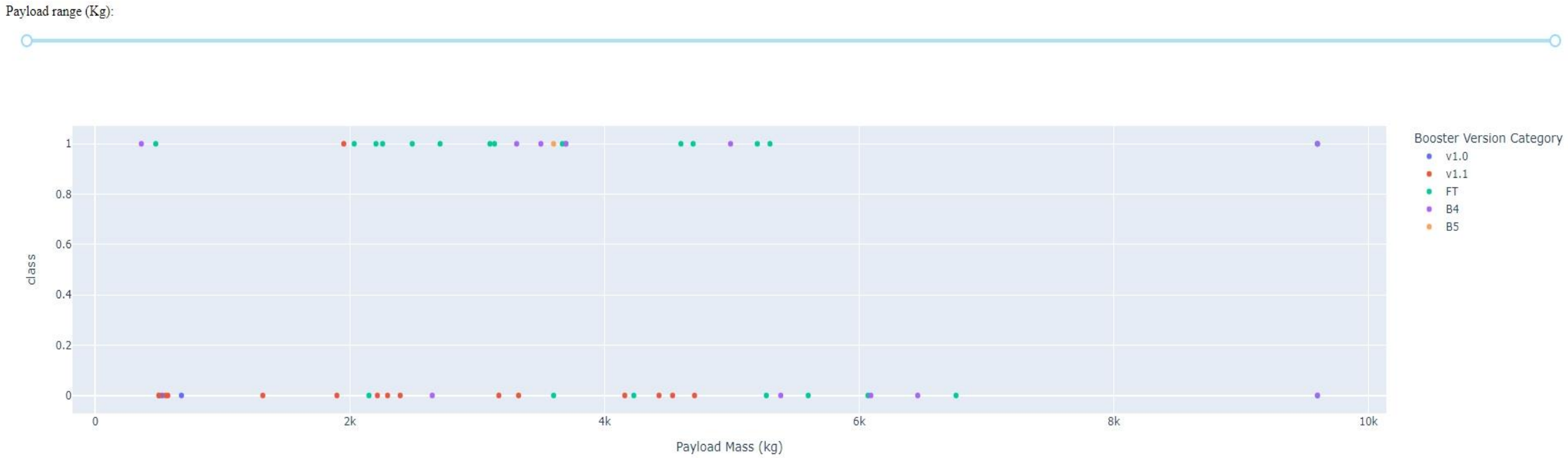
Launch Success Rate for each site

- Screenshot of the piechart for the launch site with highest launch success ratio
- Finding: 'KSC LC-39A' has the highest launch success rate of 77%, followed by 'CCAFS SLC-40' (~43%), 'VAFB SLC-4E' (40%), lastly 'CCAFS LC-40' (~27%)



Payload vs. Launch Outcome for All Sites

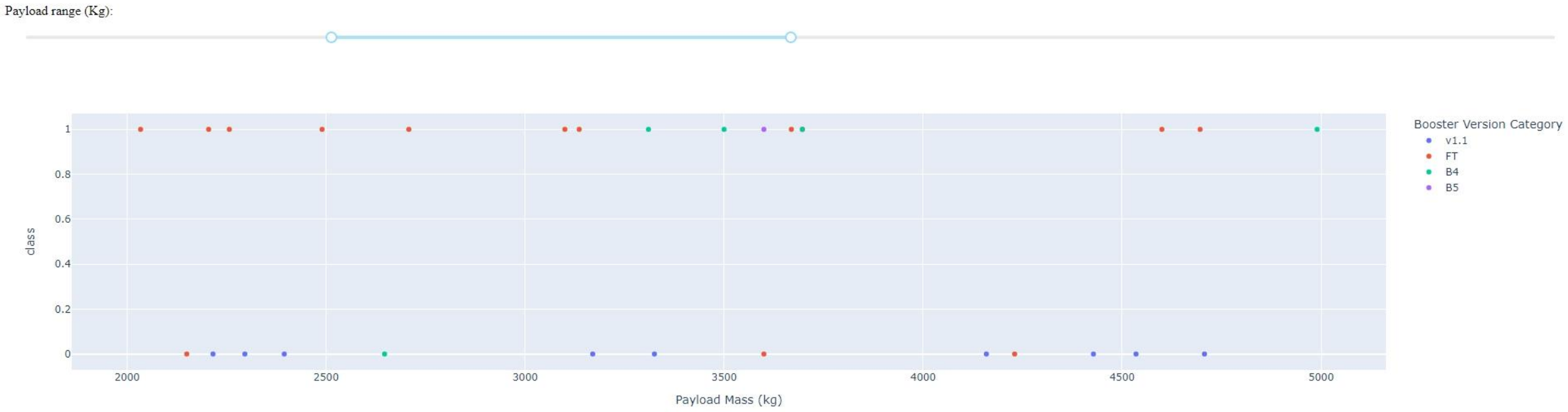
- Screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider
- Findings: mid payload range from 1900 to 5300 KG has the highest launch success rate, while light payload range within 500-1500 KG and heavy payload above 5500 KG have lowest success rate (almost 0)



Payload vs. Launch Outcome for All Sites (Highest Success Range)

- Screenshots of Payload vs. Launch Outcome scatter plot for all sites, with highest success range

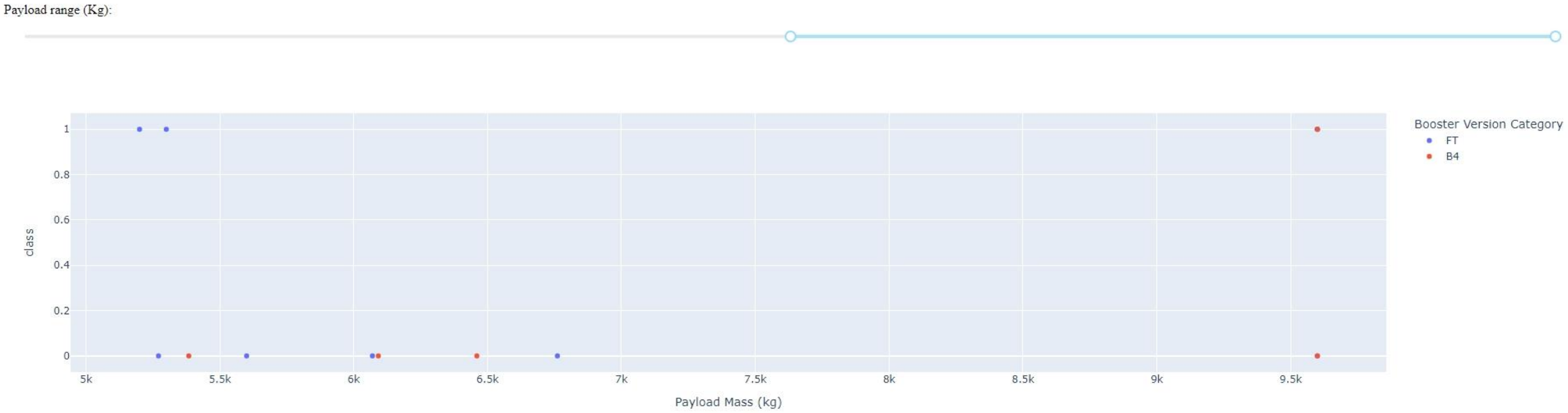
Mid payload range from 1900 to 5300 KG (high success rate)



Payload vs. Launch Outcome for All Sites (Lowest Success Range)

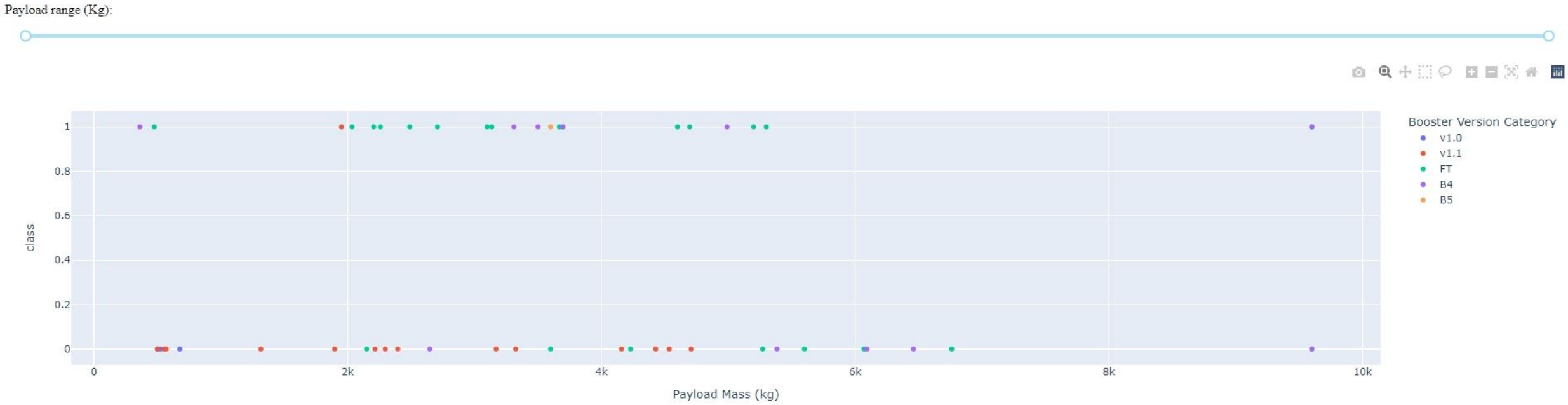
- Screenshots of Payload vs. Launch Outcome scatter plot for all sites, with lowest success range

Heavy payload above 5500 KG (low success rate)



Payload vs. Launch Outcome for All Sites (F9 Booster Version with Highest Success Rate)

- Screenshot of Payload vs. Launch Outcome scatter plot for all sites, with FT (green dots) booster version has the highest success rate



Predictive analysis (Classification)

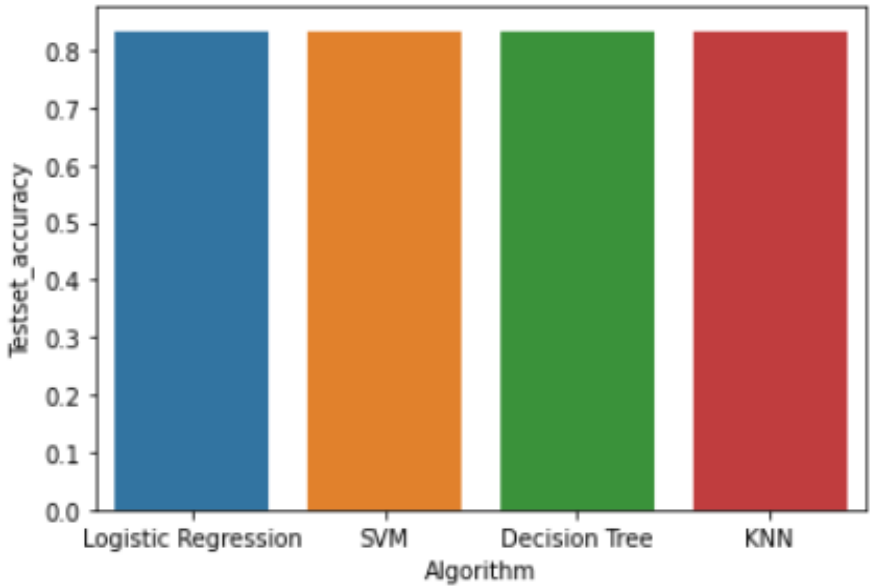
Classification Accuracy

Visualize the train set and test set accuracy for all the built models: Logistic Regression, SVM, Decision Tree and K Nearest Neighbors

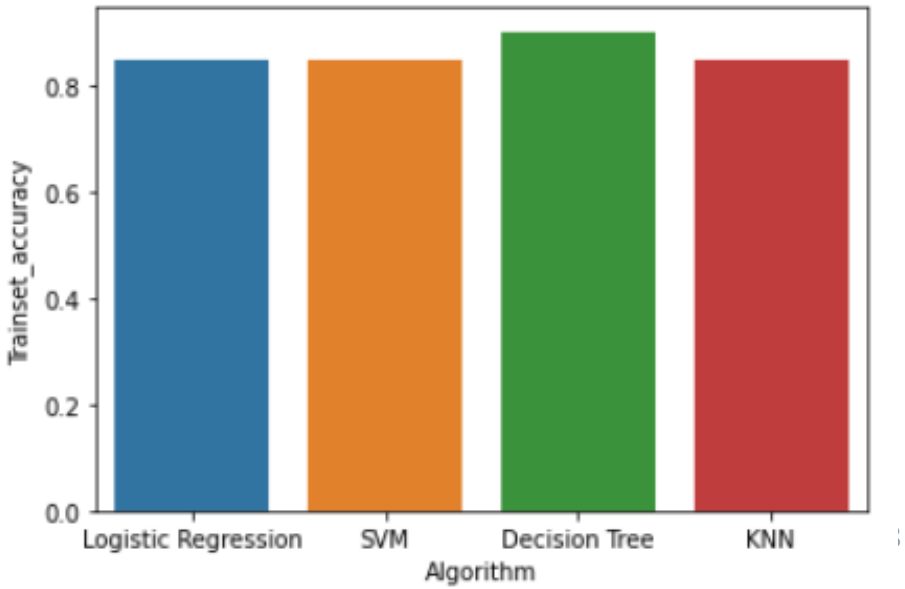
Findings:

1. Logistic Regression, SVM, Decision Tree Classifier and K Nearest Neighbors all have similar performance in terms of test set accuracy (83.3%) and confusion_matrix
2. Decision Tree Classifier performs slightly better than other 3 models in the training set

Test set accuracy



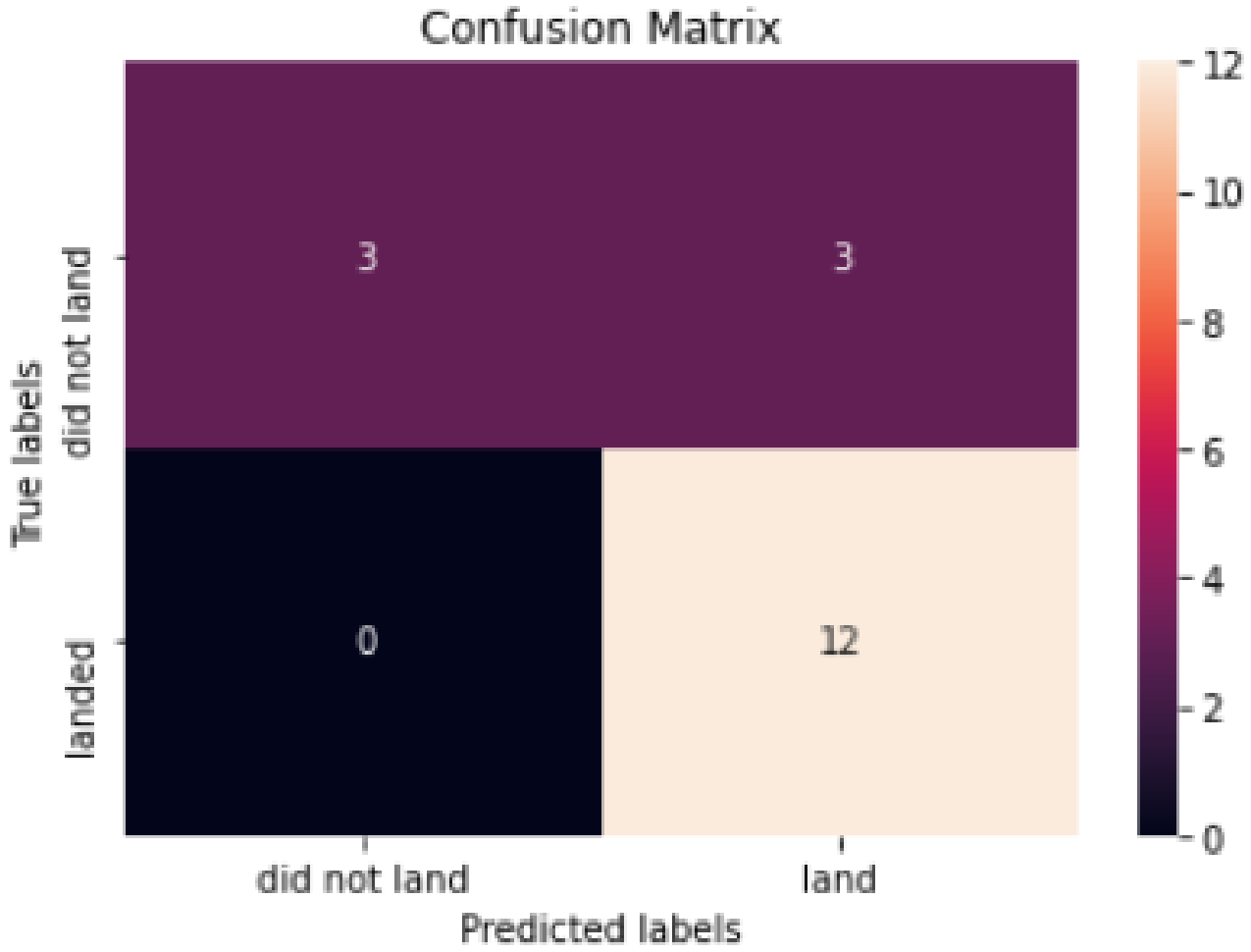
Train set accuracy



Confusion Matrix

From the confusion matrix of logistic regression, one can observe that:

1. The model can distinguish between different classes
2. The predicted accuracy for 'land' class is 100%, with all predictions correct
3. The problem is with the 'did not land' class', among them the model incorrectly predict 50% to 'land', which is known as 'False Positives'



CONCLUSION

- **Exploration:**

✓ Four unique launch sites: CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E , all are in close proximity to railways, highways, coastline and equator line, while keeping some distance away from cities.

- **Important features and relationship to launch outcome:**

✓ Flight number: positively correlated to the first stage land success.

✓ Launch site: different launch sites have different success rates. 'KSC LC-39A' has the largest successful launches (~42%) and highest success rate (77%), while 'CCAFS LC-40' has the second largest successful launches (~29%) but lowest success rate (~27%).

✓ Orbit type: different orbits have different success rates. ES-L1, GEO, HEO, SSO have the highest success rate of 100%, while SO have success rate of 0, and GTO success rate below 60%

✓ Payload mass: mid payload range from 1900 to 5300 KG has the highest launch success rate, while light payload range within 500-1500 KG and heavy payload above 5500 KG have lowest success rate (almost 0)

✓ Booster version: FT has the highest success rate

✓ Year of launch: success rate since 2013 kept increasing till 2020, with a small dip in 2018

✓ Interaction between different variables: Payload vs Orbit type, Flight number vs Orbit type, Payload vs Launch site

- **Predictive models:**

✓ Logistic Regression, SVM, Decision Tree Classifier and K Nearest Neighbors all have similar performance in terms of test set accuracy score (83.3%) and confusion_matrix

✓ Decision Tree Classifier performs slightly better than other 3 models in the training set

✓ Major issue is with False Positives, while the model predicted outcome is success but true outcome is failure.

APPENDIX

- See the Github repository for all codes involved in the project:
  - https://github.com/liuyiemily/IBM_Data-Science-Project