SpaceY

by Andrea Stevani

21.10.2024

SpaceY

1. Executive Summary
2. Introduction
3. Methodology
4. Results (1/2)
   - EDA with Visualization
   - EDA with SQL
   - Interactive Maps with Folium
   - Plotly Dash Dashboard
   - Predictive Analytics
5. Results (2/2)
6. Conclusion

EXECUTIVE SUMMARY

EXECUTIVE SUMMARY

Summary of Methodologies

The research aims to identify the factors contributing to a successful rocket landing. To determine these factors, the following methodologies were employed:

Data Collection

Data was gathered using the SpaceX REST API and web scraping techniques.

Data Wrangling

Data was organized to create a variable indicating the success or failure of landings.

Data Exploration

Data was analyzed using visualization techniques, considering factors such as payload, launch site, flight number, and yearly trends.

Data Analysis

SQL was used to calculate statistics including total payload, payload range for successful launches, and the total number of successful and failed outcomes.

Launch Site Exploration

Success rates of launch sites and their proximity to geographical markers were examined.

Visualization

Visualizations were created to highlight the most successful launch sites and payload ranges.

Model Building

Predictive models were developed to forecast landing outcomes using logistic regression, support vector machines (SVM), decision trees, and K-nearest neighbor (KNN).

Results

Exploratory Data Analysis

- Launch success has improved over time.
- KSC LC-39A has the highest success rate among landing sites.
- Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate.

Visualization/Analytics

- Most launch sites are near the equator and close to the coast.

Predictive Analytics

- All models performed similarly on the test set, with the decision tree model slightly outperforming the others.

INTRODUCTION

INTRODUCTION

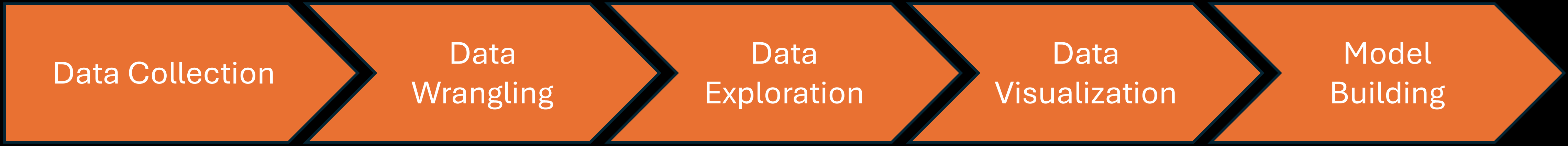
SpaceX, a prominent player in the space industry, aims to make space travel accessible to everyone. Their achievements include sending spacecraft to the International Space Station, deploying a satellite network that provides internet access, and conducting manned space missions. SpaceX manages to keep launch costs relatively low at $62 million per launch, thanks to the innovative reuse of the first stage of their Falcon 9 rocket. In contrast, other providers, who cannot reuse the first stage, face costs upwards of $165 million per launch. By predicting whether the first stage will successfully land, we can estimate the launch cost. This can be achieved using public data and machine learning models to forecast whether SpaceX or a competing company can reuse the first stage.

Explore

- How payload mass, launch site, number of flights, and orbits influence the success of first-stage landings.
- The trend of successful landings over time.
- The most effective predictive model for successful landings (binary classification).

METHODOLOGY

METHODOLOGY



Data Collection

Data Wrangling

Data Exploration

Data Visualization

Model Building

- Gather data using the SpaceX REST API and web scraping techniques.

- Prepare the data for analysis and modeling by filtering it, handling missing values, and applying one-hot encoding.

- Perform exploratory data analysis (EDA) using SQL and data visualization techniques.

- Visualize the data with tools like Folium and Plotly Dash.

- Develop classification models to predict landing outcomes. Tune and evaluate these models to identify the best model and parameters.

DATA COLLECTION - API

Request Data

Obtain rocket launch data from the SpaceX API.

Decode Response

Use *.json()* to decode the response and convert it into a dataframe using *.json_normalize()*.

Request Launch Information

Retrieve detailed information about the launches from the SpaceX API using custom functions.

Create Dictionary

Construct a dictionary from the retrieved data.

Create Dataframe

Convert the dictionary into a dataframe.

Filter Dataframe

Limit the dataframe to include only Falcon 9 launches.

Handle Missing Values

Replace missing values in the Payload Mass column with the calculated mean.

Export Data

Save the data to a CSV file.

DATA COLLECTION - Web Scraping

Request Data	Obtain Falcon 9 launch data from Wikipedia.
Create BeautifulSoup Object	Generate a BeautifulSoup object from the HTML response.
Extract Column Names	Retrieve column names from the HTML table header.
Collect Data	Parse the HTML tables to gather the data.
Create Dictionary	Construct a dictionary from the collected data.
Create Dataframe	Convert the dictionary into a dataframe.
Export Data	Save the dataframe to a CSV file.

DATA WRANGLING

Perform EDA

Conduct exploratory data analysis and identify data labels.

Calculate

- The number of launches for each site.
- The number and frequency of orbits.
- The number and frequency of mission outcomes per orbit type.

Create Binary Column

Create a binary column for landing outcomes (dependent variable)

Export Data

Save the data to a CSV file.

Define Landing Outcomes

Landings were not always successful:
True Ocean: mission outcome mission outcome had a successful landing to a had a successful landing to a specific region of the oceanspecific region of the ocean
False Ocean: Indicates an unsuccessful landing in a specific ocean region.
True RTLS: Indicates a successful landing on a ground pad.
False RTLS: Indicates an unsuccessful landing on a ground pad.
True ASDS: Indicates a successful landing on a drone ship.
False ASDS: Indicates an unsuccessful landing on a drone ship.

Convert Outcomes

Outcomes are converted into 1 for a successful landing and 0 for an unsuccessful landing.

RESULTS (1/2)

EDA with Visualization

Charts

- Flight Number vs. Payload
- Flight Number vs. Launch Site
- Payload Mass (kg) vs. Launch Site
- Payload Mass (kg) vs. Orbit type

Analysis

- View relationships using scatter plots. These variables could be useful for machine learning if a relationship exists.

- Show comparisons among discrete categories using bar charts. Bar charts display the relationships among categories and a measured value.

EDA with SQL

Display

- The names of distinct launch sites.
- Five records where the launch site starts with ‘CCA’.
- The total payload mass carried by boosters launched by NASA (CRS).
- The average payload mass carried by the booster version F9 v1.1.

Analysis

- The date of the first successful landing on a ground pad.
- The names of boosters that successfully landed on a drone ship and have a payload mass between 4,000 and 6,000 kg.
- The total number of successful and failed missions.
- The names of booster versions that have carried the maximum payload.
- The failed landing outcomes on a drone ship, their booster version, and launch site for the months in the year 2015.
- The count of landing outcomes between June 4, 2010, and March 20, 2017 (in descending order).

MAP with Folium

Markers Indicating Launch Sites

- Added a blue circle at the coordinates of NASA Johnson Space Center with a popup label displaying its name using its latitude and longitude coordinates.
- Added red circles at the coordinates of all launch sites with popup labels showing their names using their latitude and longitude coordinates.

Colored Markers of Launch Outcomes

- Added colored markers for successful (green) and unsuccessful (red) launches at each launch site to indicate which sites have higher success rates.

Distances Between a Launch Site and Nearby Locations

- Added colored lines to show the distance between launch site CCAFS SLC-40 and its proximity to the nearest coastline, railway, highway, and city.

DASHBOARD with Plotly Dash

Dropdown List with Launch Sites

- Enable users to select either all launch sites or a specific launch site.

Slider for Payload Mass Range

- Allow users to choose a range for payload mass.

Pie Chart Displaying Successful Launches

- Enable users to view successful and unsuccessful launches as a percentage of the total.

Displaying Payload Mass vs. Success Rate by Booster Version

- Allow users to observe the correlation between payload mass and launch success.

PREDICTIVE ANALYTICS

1. Create a NumPy array from the “Class” column.

2. Standardize the data using StandardScaler. Fit and transform the data.

3. Split the data using train_test_split.

4. Create a GridSearchCV object with cv=10 for parameter optimization.

5. Apply GridSearchCV on various algorithms: logistic regression (LogisticRegression()), support vector machine (SVC()), decision tree (DecisionTreeClassifier()), and k-nearest neighbors (KNeighborsClassifier()).

6. Calculate the accuracy on the test data using .score() for all models.

7. Evaluate the confusion matrix for each model.

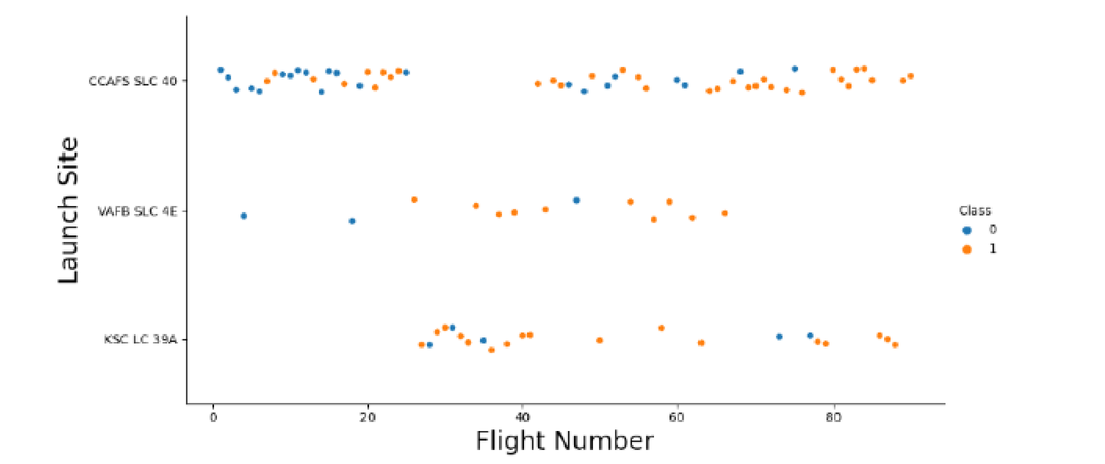
8. Determine the best model using Jaccard_Score, F1_Score, and accuracy.

RESULTS (2/2)

SpaceY

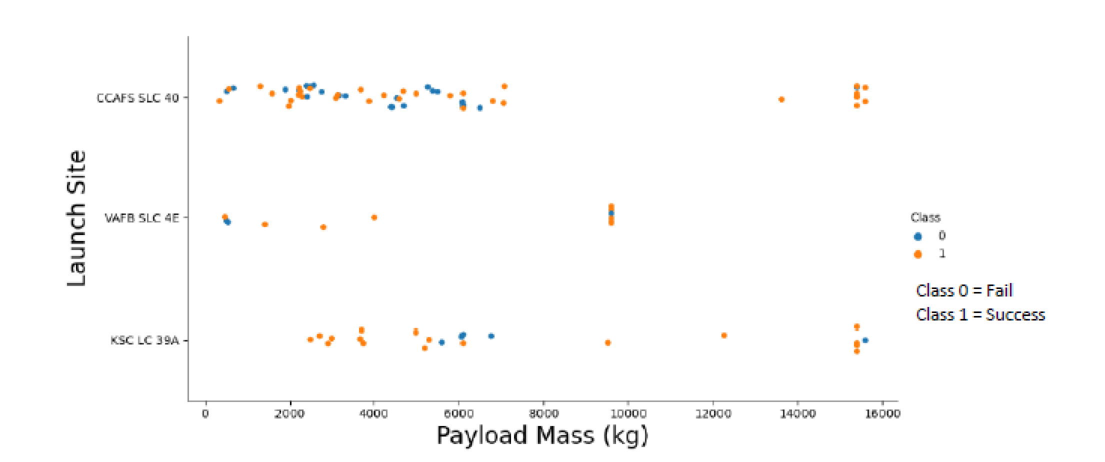
Flight Number vs. Launch Site

- Earlier flights had a lower success rate (blue = fail).
- Later flights had a higher success rate (orange = success).
- Approximately half of the launches were from the CCAFS SLC-40 launch site.
- The VAFB SLC-4E and KSC LC-39A launch sites have higher success rates.
- We can deduce that newer launches tend to have a higher success rate.



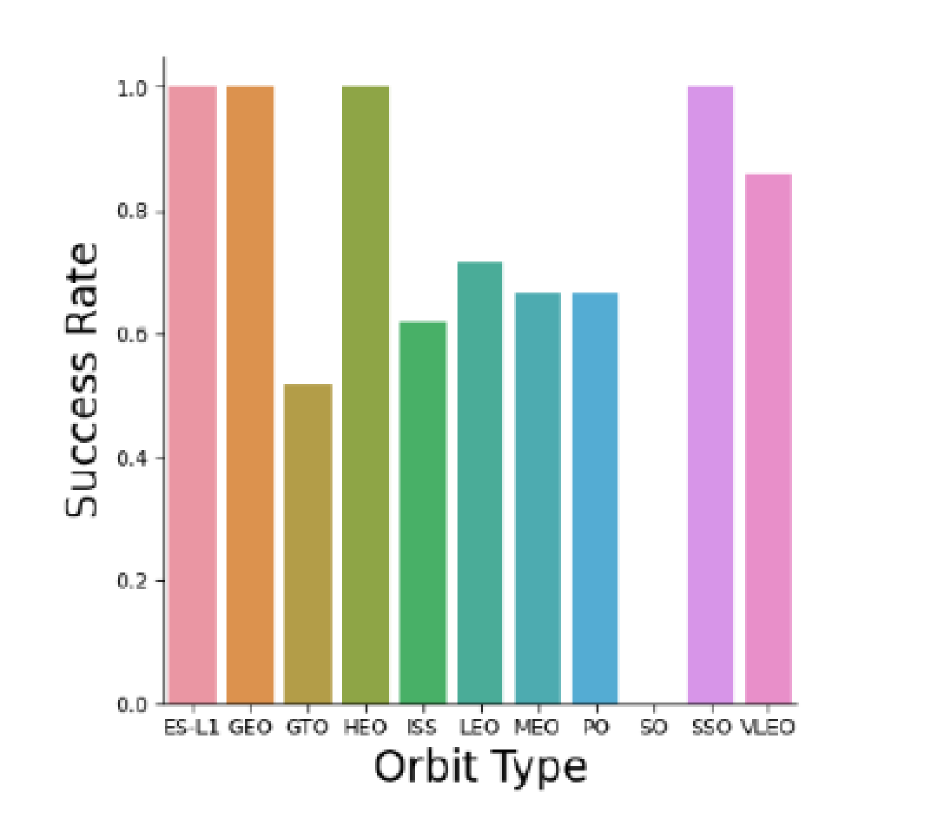
Payload vs. Launch Site

- Generally, the higher the payload mass (kg), the greater the success rate.
- Most launches with a payload exceeding 7,000 kg were successful.
- KSC LC-39A has a 100% success rate for launches with a payload less than 5,500 kg.
- VAFB SLC-4E has not launched any payloads greater than approximately 10,000 kg.



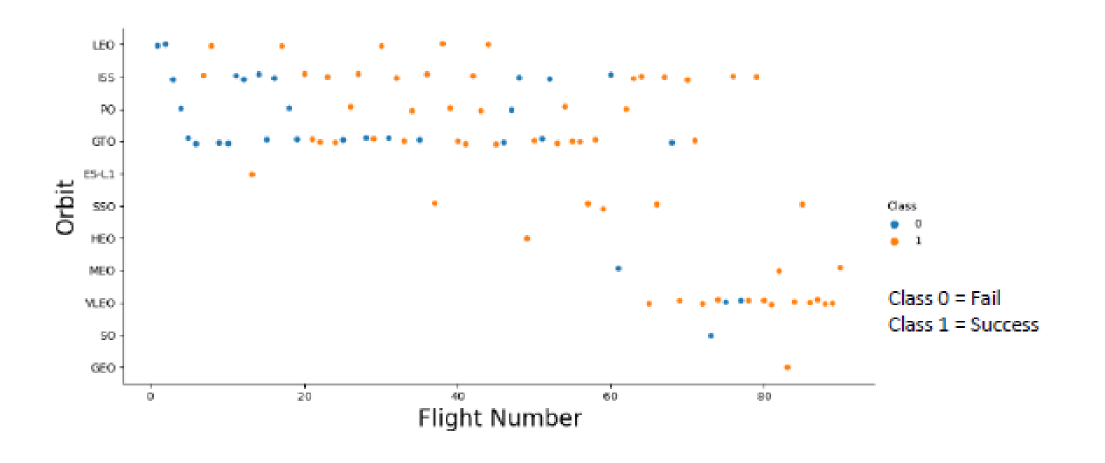
Success Rate by Orbit

- **100% Success Rate**: ES-L1, GEO, HEO, and SSO.
- **50%-80% Success Rate**: GTO, ISS, LEO, MEO, and PO.
- **0% Success Rate**: SO.



Flight Number vs. Orbit

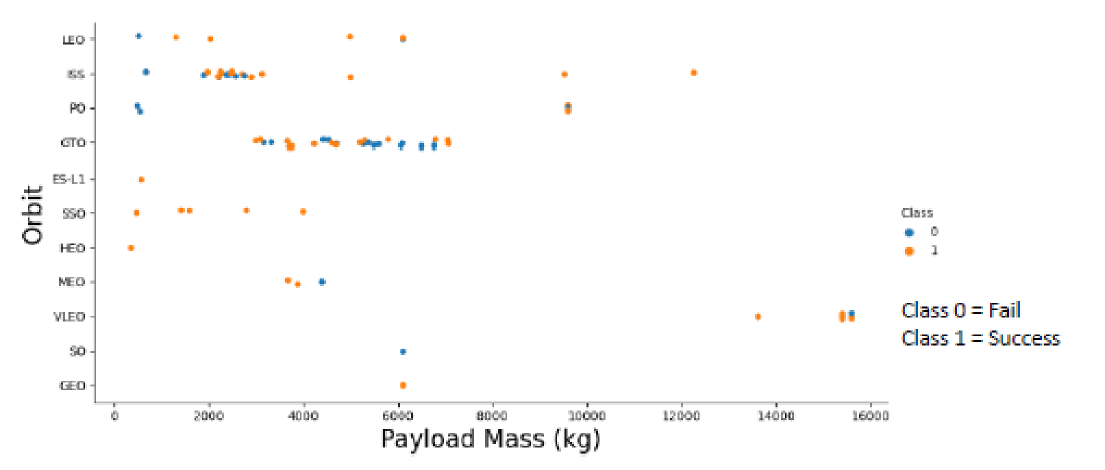
- The success rate generally rises with the number of flights for each orbit.
- This pattern is particularly evident for the LEO orbit.
- However, the GTO orbit does not exhibit this trend.



Class 0 = Fail
Class 1 = Success

Payload vs. Orbit

- Heavy payloads perform better with LEO, ISS, and PO orbits.
- The GTO orbit shows varied success with heavier payloads.

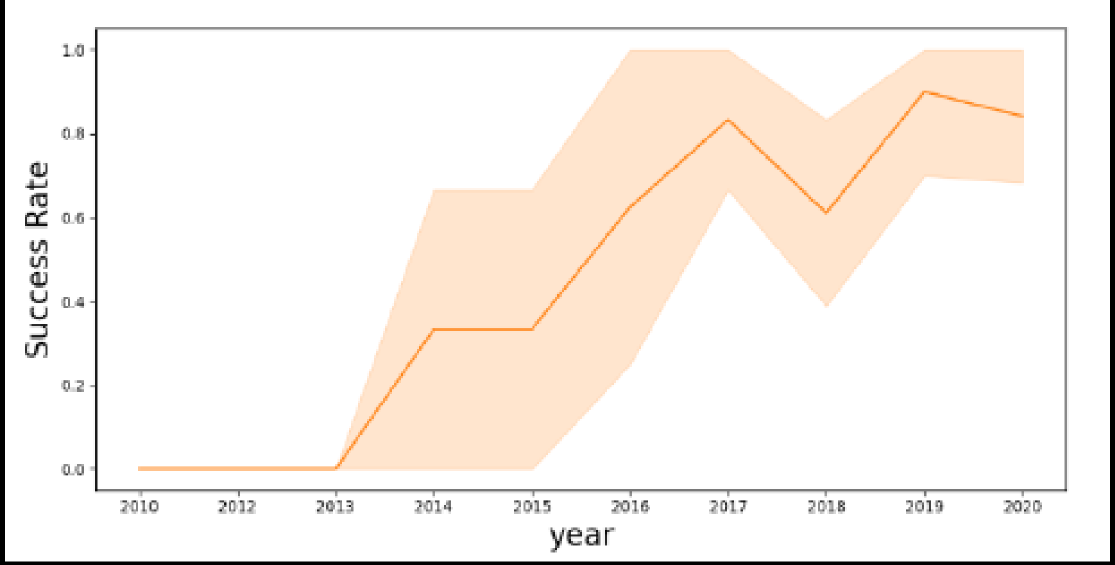


Class 0 = Fail
Class 1 = Success

SpaceY

Launch Success over Time

- The success rate increased from 2013 to 2017 and from 2018 to 2019.
- The success rate declined from 2017 to 2018 and from 2019 to 2020.
- Overall, the success rate has shown improvement since 2013.

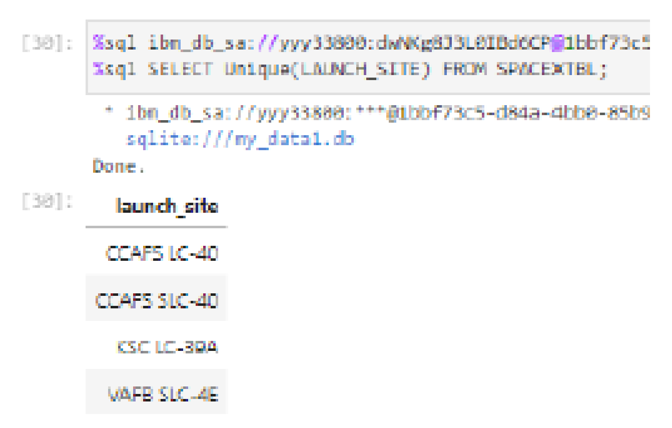


SpaceY

Launch Site Information

Launch Site Names:

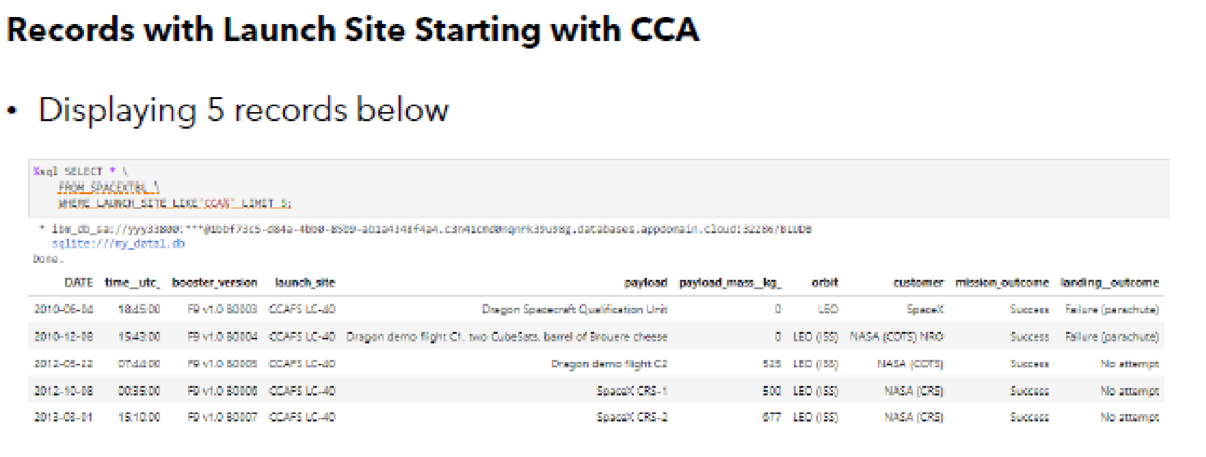
- CCAFS LC-40
- CCAFS SLC-40
- KSC LC-39A
- VAFB SLC-4E



```
[30]: %sql ibm_db_sa://yyy33800:dwNKg8J3L0IBd6CP@1bbf73c5
      %sql SELECT Unique(LAUNCH_SITE) FROM SPACEXTBL;

       * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4bb0-85b9
         sqlite:///my_data1.db
      Done.
```

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



Records with Launch Site Starting with CCA

- Displaying 5 records below

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

 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:32286/BLUDB
   sqlite:///my_data1.db
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

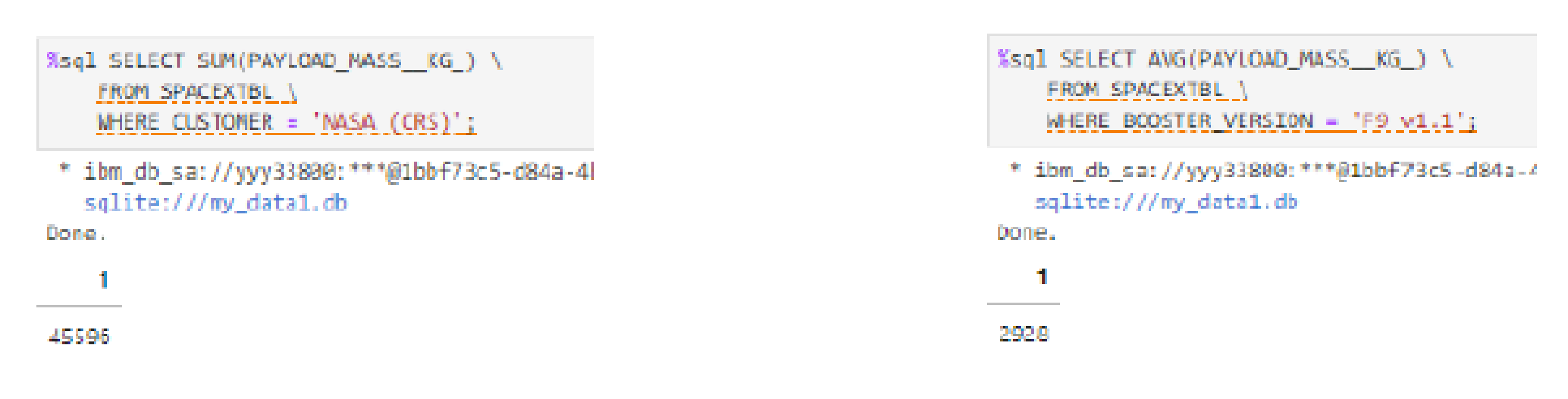
Payload Mass

Total Payload Mass:

45,596 kg (total) carried by boosters launched by NASA (CRS)

Total Payload Mass:

2,928 kg (average) carried by booster version F9 v1.1



```
%sql SELECT SUM(PAYLOAD_MASS__KG_) \
    FROM SPACEXTBL \
    WHERE CUSTOMER = 'NASA (CRS)';
```

```
 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4
   sqlite:///my_data1.db
Done.

   1
45596
```

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) \
    FROM SPACEXTBL \
    WHERE BOOSTER_VERSION = 'F9 v1.1';
```

```
 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4
   sqlite:///my_data1.db
Done.

   1
2928
```

SpaceY

Landing & Mission Info

First Successful Landing on Ground Pad

Date: 12/22/2015

Boosters with Mass Between 4,000 and 6,000 kg

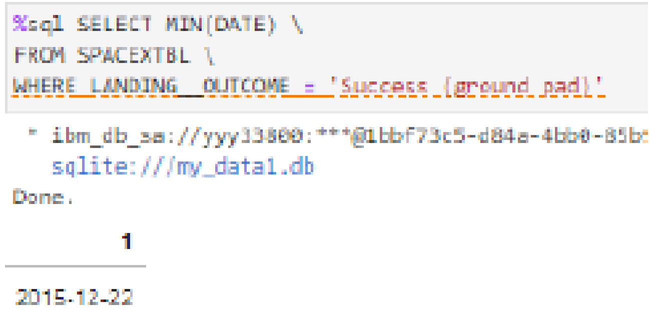
JCSAT-14, JCSAT-16, SES-10, SES-11 / EchoStar 105

Total Number of Successful and Failed Mission Outcomes

1 Failure in Flight

99 Successes

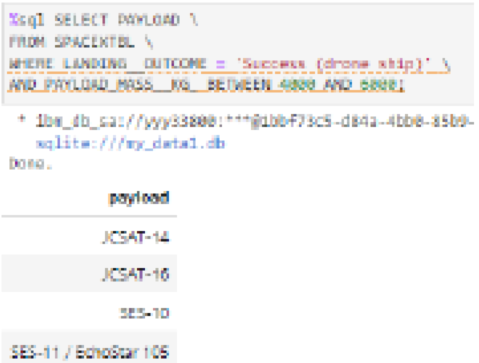
1 Success (payload status unclear)



```
%sql SELECT MIN(DATE) \
FROM SPACEXTBL \
WHERE LANDING__OUTCOME = 'Success (ground pad)'
```

```
 * ibm_db_sa://yyy13800:***@1bbf73c5-d84a-4bb0-85b
   sqlite:///my_data1.db
Done.
```

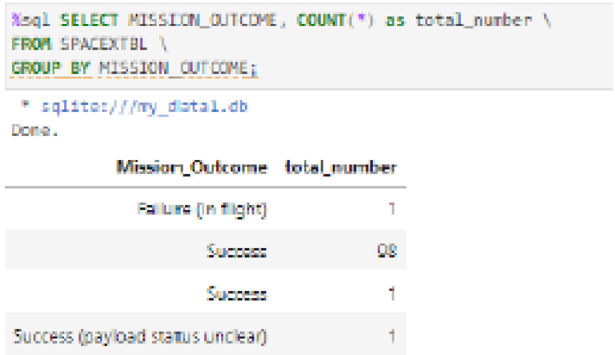
1
2015-12-22



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

```
 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4bb0-85b0-
   sqlite:///my_data1.db
Done.
```

payload
JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105



```
%sql SELECT MISSION_OUTCOME, COUNT(*) as total_number \
FROM SPACEXTBL \
GROUP BY MISSION_OUTCOME;
```

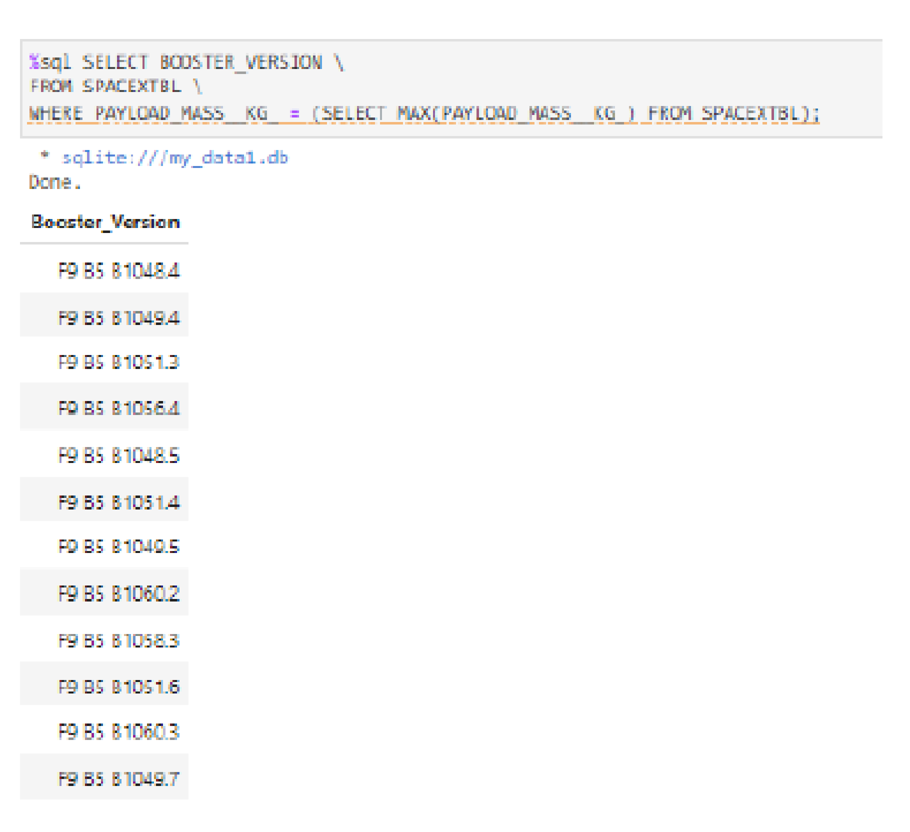
```
 * sqlite:///my_data1.db
Done.
```

Mission_Outcome	total_number
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters

Boosters Carrying Maximum Payload

- 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



```
%sql SELECT BOOSTER_VERSION \
FROM SPACEXTBL \
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

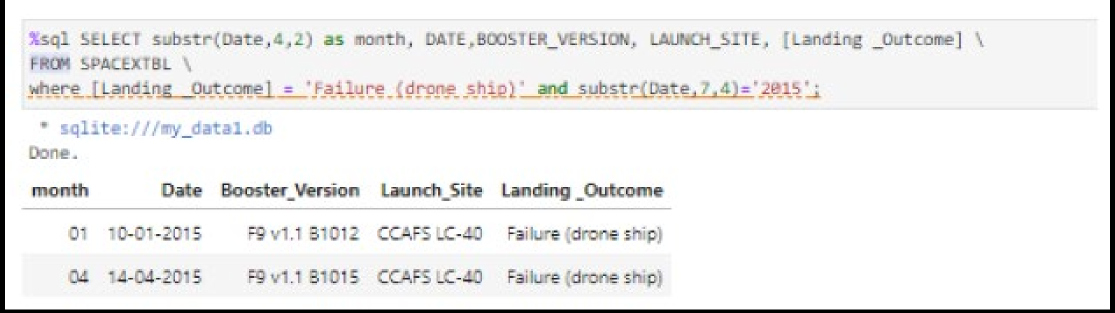
```
 * sqlite:///my_data1.db
Done.
```

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

Failed Landings on Drone Ship

In 2015

- Showing month, date, booster version, launch site and landing outcome



```
%sql SELECT substr(Date,4,2) as month, DATE,BOOSTER_VERSION, LAUNCH_SITE, [Landing _Outcome] \
FROM SPACEXTBL \
where [Landing _Outcome] = 'Failure (drone ship)' and substr(Date,7,4)='2015';
```

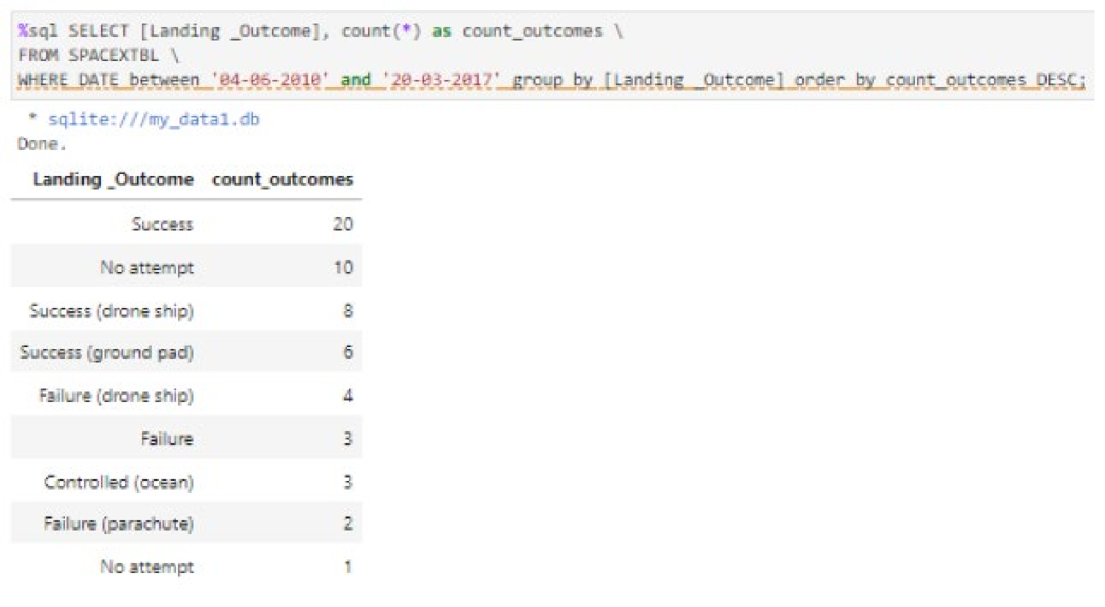
```
 * sqlite:///my_data1.db
Done.
```

month	Date	Booster_Version	Launch_Site	Landing _Outcome
01	10-01-2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	14-04-2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Count of Successful Landings

Ranked in Descending Order

- Count of landing outcomes between June 4, 2010, and March 20, 2017, listed in descending order.



```
%sql SELECT [Landing _Outcome], count(*) as count_outcomes \
FROM SPACEXTBL \
WHERE DATE between '04-06-2010' and '20-03-2017' group by [Landing _Outcome] order by count_outcomes DESC;
```

```
 * sqlite:///my_data1.db
Done.
```

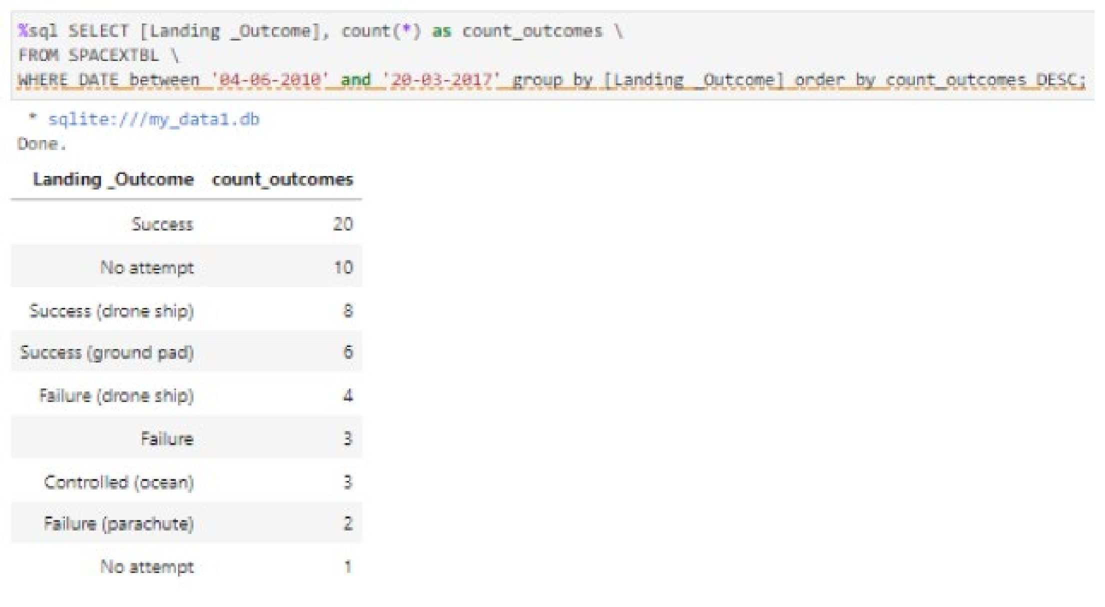
Landing _Outcome	count_outcomes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

SpaceY

Count of Successful Landings

Ranked in Descending Order

- Count of landing outcomes between June 4, 2010, and March 20, 2017, listed in descending order.

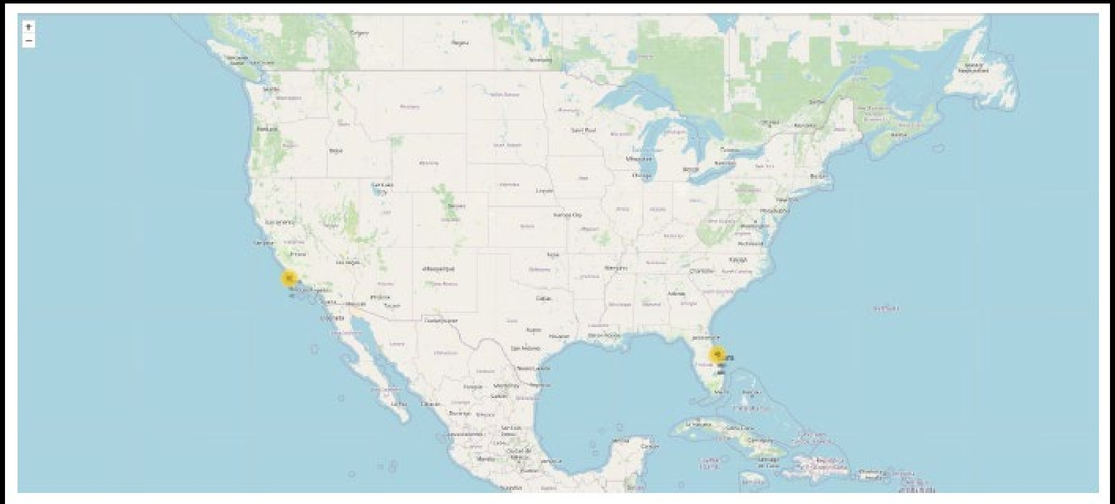


```
%sql SELECT [Landing _Outcome], count(*) as count_outcomes \
FROM SPACEXTBL \
WHERE DATE between '04-06-2010' and '20-03-2017' group by [Landing _Outcome] order by count_outcomes DESC;
```

```
 * sqlite:///my_data1.db
Done.
```

Landing _Outcome	count_outcomes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

Launch Sites

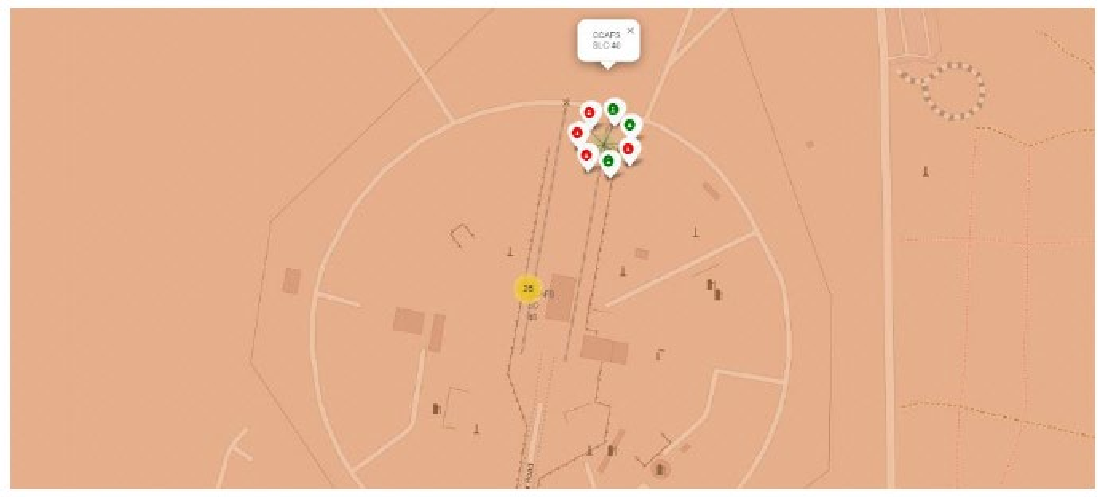


Near the Equator:

- The closer a launch site is to the equator, the easier it is to launch into an equatorial orbit. This is because the Earth’s rotation provides additional assistance for a prograde orbit. Rockets launched from equatorial sites benefit from the Earth’s rotational speed, which offers a natural boost. This boost helps reduce the need for extra fuel and boosters, thereby saving costs.

SpaceY

Launch Outcomes



At Each Launch Site

Outcomes:

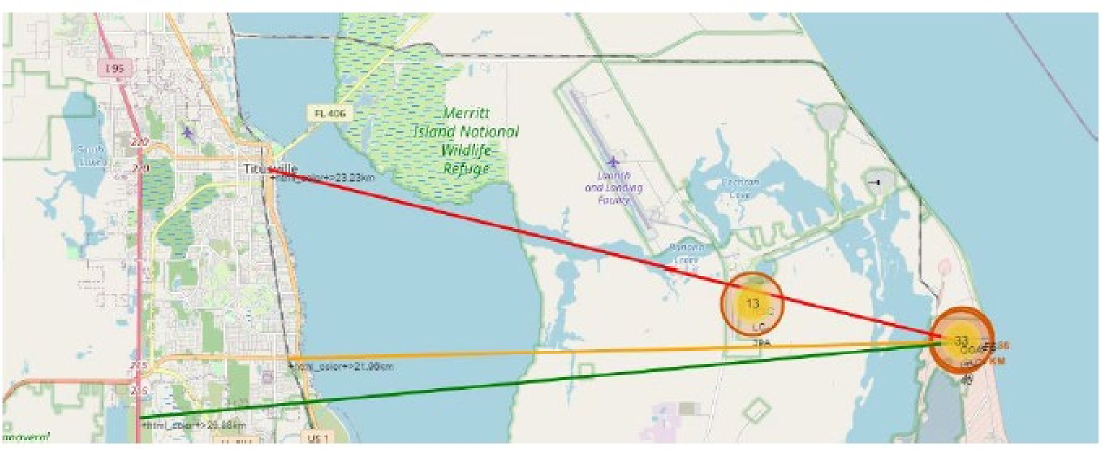
- Green markers for successful launches.
- Red markers for unsuccessful launches.
- Launch site CCAFS SLC-40 has a success rate of 3 out of 7 launches (42.9%).

SpaceY

Distance to Proximities

CCAFS SLC-40

- 0.86 km from the nearest coastline.
- 21.96 km from the nearest railway.
- 23.23 km from the nearest city.
- 26.88 km from the nearest highway.



I 95
FL 406
Merritt Island National Wildlife Refuge
220
Titusville
23.23km
Launch and Landing Facility
13
LC
39A
33
21.96km
405
26.88km
US 1
KM

- **Coasts**: Ensure that spent stages or failed launches do not fall on people or property by dropping them along the launch path over the ocean.

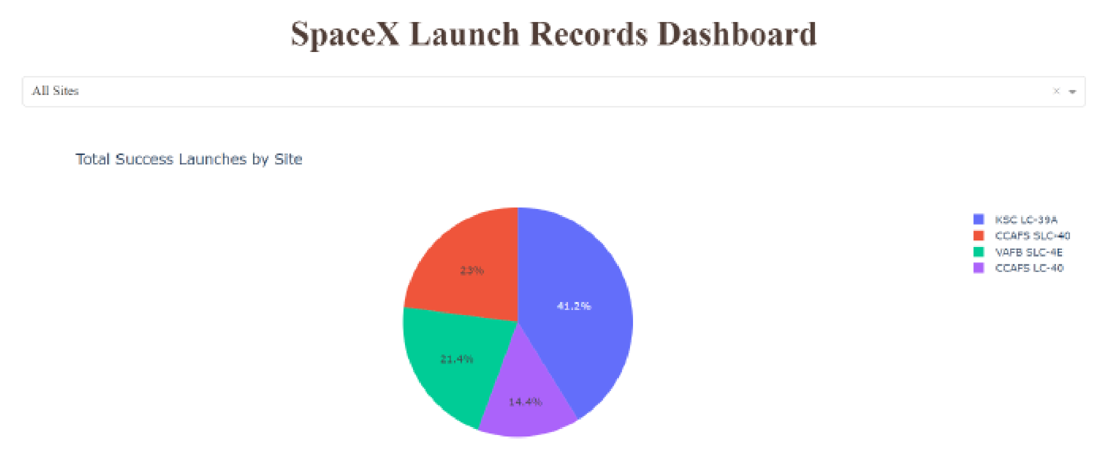
- **Safety / Security**: An exclusion zone around the launch site is necessary to keep unauthorized individuals away and ensure the safety of people.

- **Transportation / Infrastructure and Cities**: Launch sites need to be far enough from anything that could be damaged by a failed launch, but still close enough to roads, railways, and docks to facilitate the transport of people and materials for launch activities.

Launch Success by Site

Success as a Percentage of Total

- KSC LC-39A has the highest number of successful launches among all launch sites (41.2%).



SpaceX Launch Records Dashboard

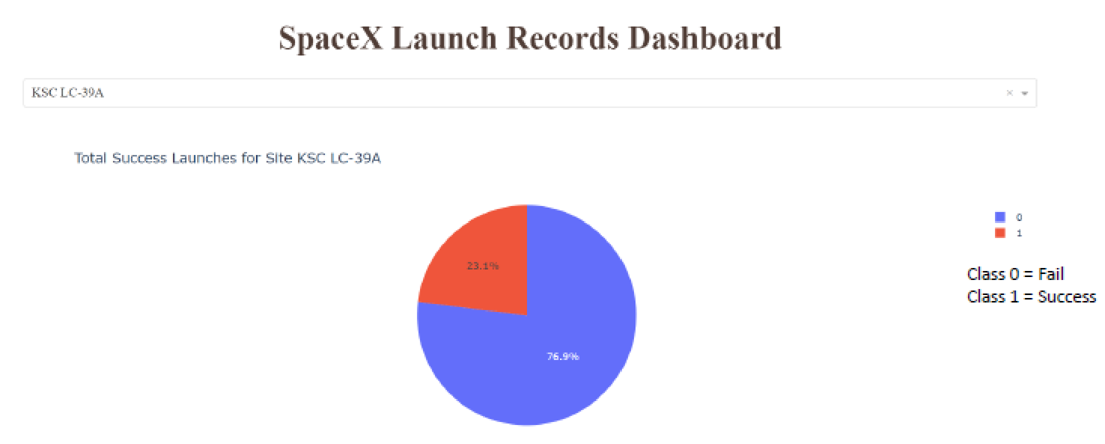
All Sites

Total Success Launches by Site

Launch Success (KSC LC-29A)

Success as a Percentage of Total

- KSC LC-39A has the highest success rate among all launch sites (76.9%).
- There have been 10 successful launches and 3 failed launches.



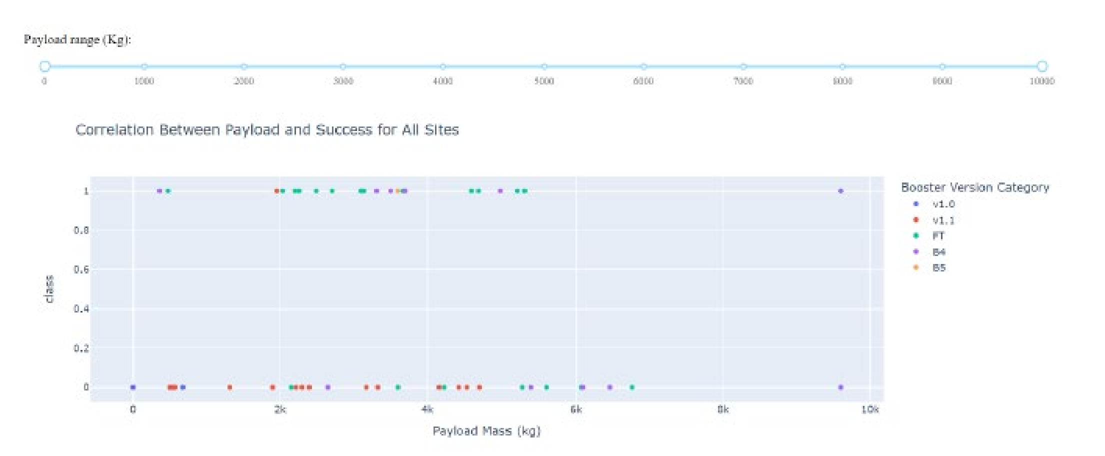
SpaceX Launch Records Dashboard

KSC LC-39A

Payload Mass and Success

By Booster Version

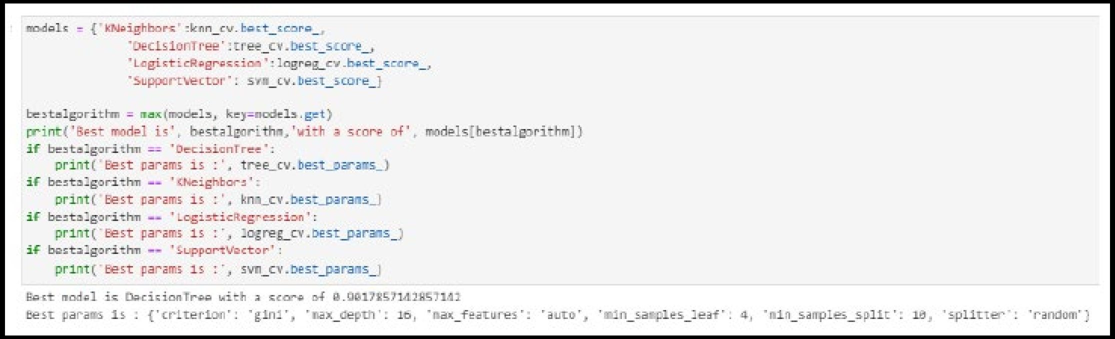
- Payloads ranging from 2,000 kg to 5,000 kg have the highest success rate.
- A value of 1 indicates a successful outcome, while a value of 0 indicates an unsuccessful outcome.



Classification

Accuracy

- All models performed at a similar level, achieving comparable scores and accuracy. This is likely due to the small dataset. The Decision Tree model slightly outperformed the others when considering the *.best_score_*.
- *.best_score_* represents the average of all cross-validation folds for a single combination of parameters.



```
models = {'KNeighbors':knn_cv.best_score_,
              'DecisionTree':tree_cv.best_score_,
              'LogisticRegression':logreg_cv.best_score_,
              'SupportVector': svm_cv.best_score_}

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

```
Best model is DecisionTree with a score of 0.9017857142857142
Best params is : {'criterion': 'gini', 'max_depth': 16, 'max_features': 'auto', 'min_samples_leaf': 4, 'min_samples_split': 10, 'splitter': 'random'}
```

SpaceY

Confusion Matrices

Performance Summary

- A confusion matrix provides a summary of a classification algorithm's performance.
- All the confusion matrices were identical.
- The presence of false positives (Type 1 errors) is problematic.
- Confusion Matrix Outputs:

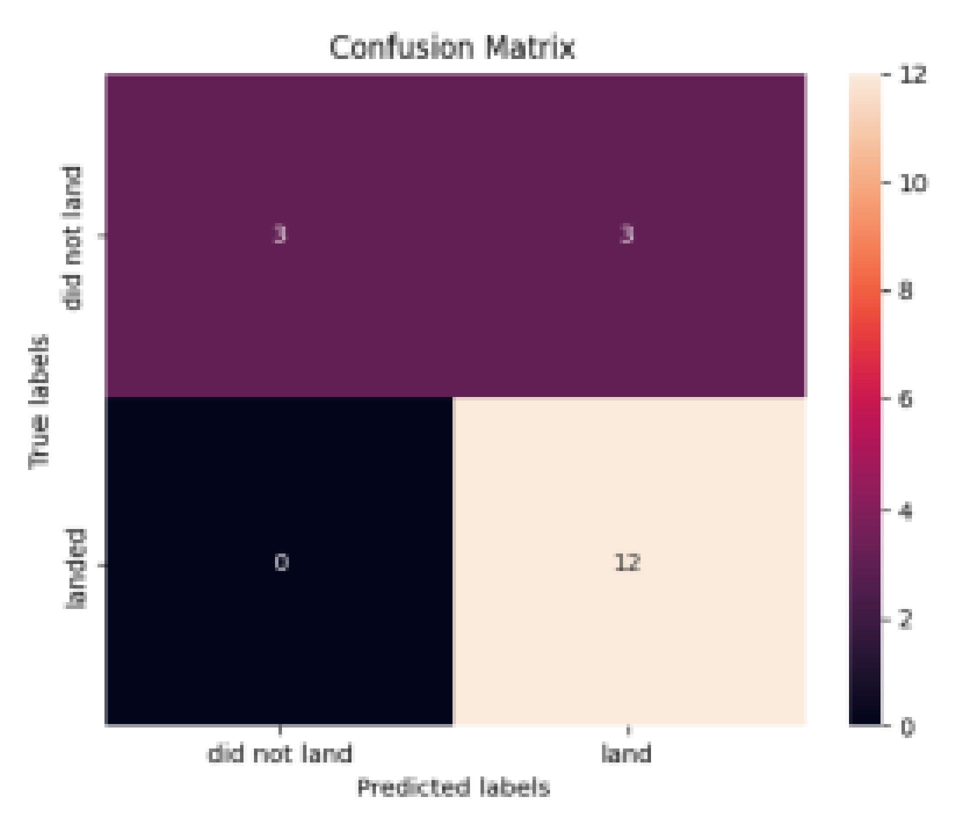
- 12 True Positives
- 3 True Negatives
- 3 False Positives
- 0 False Negatives
- Precision:

- Precision = TP / (TP + FP)
- 12 / 15 = 0.80
- Recall:

- Recall = TP / (TP + FN)
- 12 / 12 = 1
- F1 Score:

- F1 Score = 2 * (Precision * Recall) / (Precision + Recall)
- 2 * (0.8 * 1) / (0.8 + 1) = 0.89
- Accuracy:

- Accuracy = (TP + TN) / (TP + TN + FP + FN)
- (12 + 3) / (12 + 3 + 3 + 0) = 0.833



CONCLUSION

CONCLUSION

Model Performance

The models showed similar performance on the test set, with the decision tree model slightly outperforming the others.

Equator

Most launch sites are located near the equator to take advantage of the Earth’s rotational speed, which provides a natural boost. This helps reduce the cost of additional fuel and boosters.

Coast

All launch sites are situated close to the coast.

Launch Success

The success rate has increased over time.

KSC LC-39A

This site has the highest success rate among all launch sites, with a 100% success rate for launches carrying less than 5,500 kg.

Orbits

ES-L1, GEO, HEO, and SSO orbits have a 100% success rate.

Payload Mass

Across all launch sites, the higher the payload mass (kg), the higher the success rate.

THANK YOU!