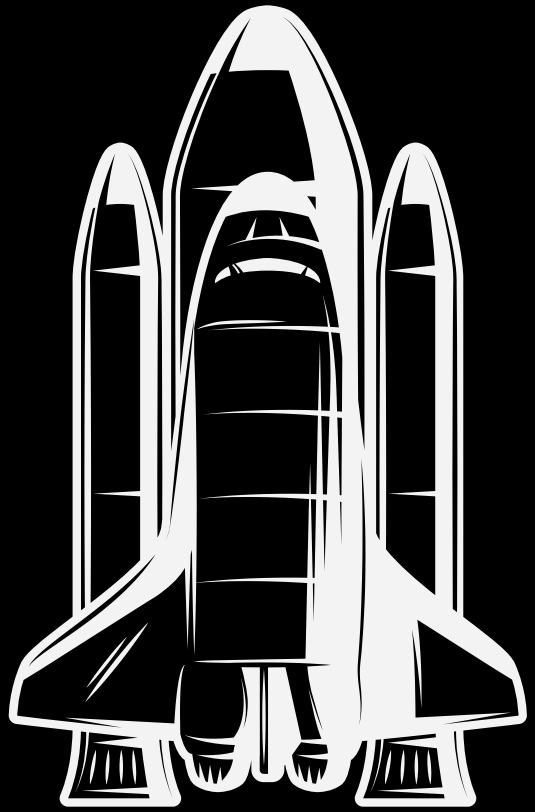
0001



IBM DATA SCIENCE
CAPSTONE
PROJECT

BY NIKHIL PANGAONKAR

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Summary of Methodologies

The research attempts to identify the factors for a successful rocket landing. To make this determination, the following methodologies where used:

- **Collect** data using SpaceX REST API and web scraping techniques.
- **Wrangle** data to create success/fail outcome variable
- **Explore** data with data visualization techniques, considering the following factors: payload, launch site, flight number and yearly trend
- **Analyze** the data with SQL, calculating the following statistics: total payload, payload range for successful launches, and total # of successful and failed outcomes
- **Explore** launch site success rates and proximity to geographical markers
- **Visualize** the launch sites with the most success and successful payload ranges
- **Build** Models to predict landing outcomes using logistic regression, support vector machine (SVM), decision tree and K-nearest neighbor (KNN)

Results

- **Exploratory Data Analysis**
- **Visualization/Analytics**
- **Predictive Analysis**

INTRODUCTION

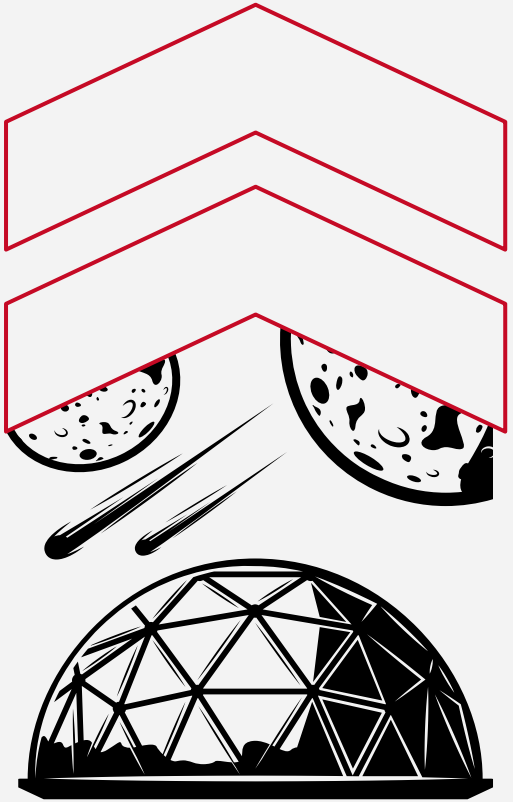
Summary of Methodologies

SpaceX, a leader in the space industry, strives to make space travel affordable for everyone. Its accomplishments include sending spacecraft to the international space station, launching a satellite constellation that provides internet access and sending manned missions to space. SpaceX can do this because the rocket launches are relatively inexpensive ($62 million per launch) due to its novel reuse of the first stage of its Falcon 9 rocket. Other providers, which are not able to reuse the first stage, cost upwards of $165 million each. By determining if the first stage will land, we can determine the price of the launch. To do this, we can use public data and machine learning models to predict whether SpaceX – or a competing company – can reuse the first stage.

Explore

How payload mass, launch site, number of flights, and orbits affect first-stage landing success

- Rate of successful landings over time
- Best predictive model for successful landing (binary classification)



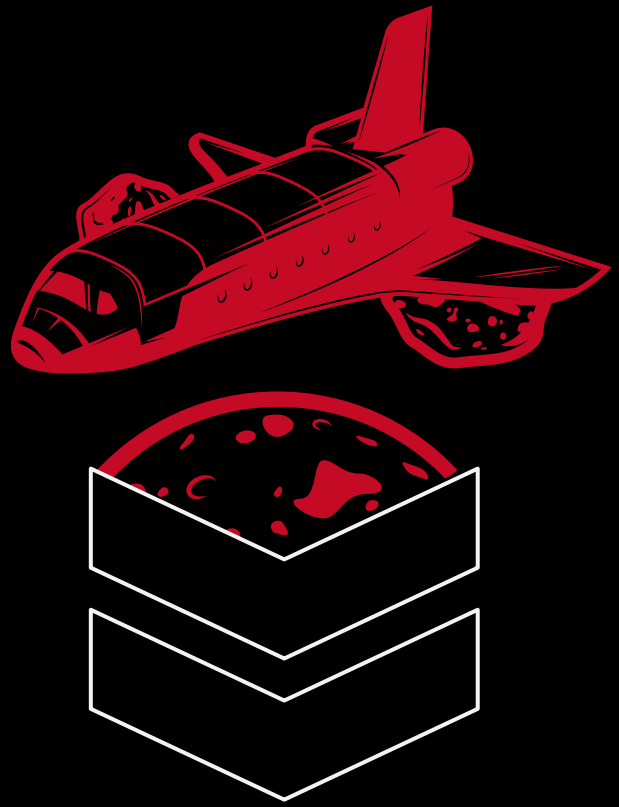
Methodology

Steps

- **Collect** data using SpaceX REST API and web scraping techniques.
- **Wrangle data –** by filtering the data, handling missing values and applying one hot encoding – to prepare the data for analysis and modeling.
- **Explore** data via EDA with SQL and data visualization techniques.
- **Explore** launch site success rates and proximity to geographical markers
- **Visualize** the data using Folium and Plotly Dash.
- **Build Models** to predict landing outcomes using classification models. Tune and evaluate models to find best model and parameters

Data Collection - API

- **Request data** from SpaceX API (rocket launch data)
- **Decode response** using .json() and convert to a dataframe using .json_normalize()
- **Request information** about the launches from SpaceX API using custom functions
- **Create dictionary** from the data
- **Create dataframe** from the dictionary
- **Filter dataframe** to contain only Falcon 9 launches
- **Replace missing values** of Payload Mass with calculated .mean()
- **Export data** to csv file



Data Collection - Web Scraping

- **Request data** (Falcon 9 launch data) from Wikipedia
- **Create BeautifulSoup object** from HTML response
- **Extract column names** from HTML table header
- **Collect data** from parsing HTML tables
- **Create dictionary** from the data
- **Filter dataframe** to contain only Falcon 9 launches
- **Replace missing values** of Payload Mass with calculated .mean()
- **Export data** to csv file

Data Wrangling

Steps

- **Perform EDA** and determine data labels
- **Calculate**:
  - # of launches for each site
  - # and occurrence of orbit
  - # and occurrence of mission outcome per orbit type]
- **Create binary** landing outcome column (dependent variable)
- **Export data** to csv file

Landing Outcome

- **True Ocean** : mission outcome had a successful landing to a specific region of the ocean
- **False Ocean** : represented an unsuccessful landing to a specific region of ocean
- **True RTLS** : meant the mission had a successful landing on a ground pad
- **False RTLS** : represented an unsuccessful landing on a ground pad
- **True ASDS** : meant the mission outcome had a successful landing on a drone ship
- **False ASDS** : represented an unsuccessful landing on drone ship
- **Outcomes converted** into 1 for a successful landing and 0 for an unsuccessful landing

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 relationship** by using **scatter plots**. The variables could be useful for machine learning if a relationship exists
• **Show comparisons** among discrete categories with **bar charts**. Bar charts show the relationships among the categories and a measured value

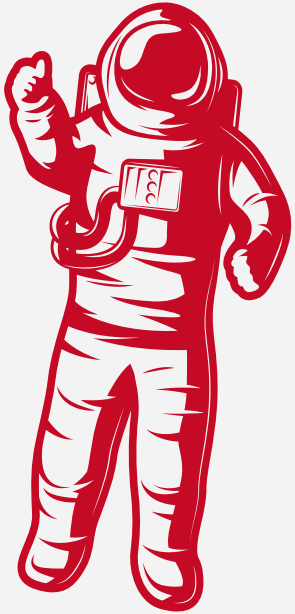
EDA with SQL

Display

- Names of unique launch sites
- 5 records where launch site begins with ‘CCA’
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1.

List

- Date of first successful landing on ground pad
- Names of boosters which had success landing on drone ship and have payload mass greater than 4,000 but less than 6,000
- Total number of successful and failed missions
- Names of booster versions which have carried the max payload
- Failed landing outcomes on drone ship, their booster version and launch site for the months in the year 2015



Map with Folium

Markers Indicating Launch Sites

- Added blue circle at NASA Johnson Space Center's coordinate with a popup label showing its name using its latitude and longitude coordinates
- Added red circles at all launch sites coordinates with a popup label showing its name using its name using its latitude and longitude coordinates

Colored Markers of Launch Outcomes

- Added colored markers of successful (green) and unsuccessful (red) launches at each launch site to show which launch sites have high success rates

Distances Between a Launch Site to Proximities

- Added colored lines to show 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

- Allow user to select all launch sites or a certain launch site

Pie Chart Showing Successful Launches

- Allow user to see successful and unsuccessful launches as a percent of the total

Slider of Payload Mass Range

- Allow user to select payload mass range

Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version

- Allow user to see the correlation between Payload and Launch Success

RESULTS SUMMARY

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

Visual Analytics

- Most launch sites are near the equator, and all are close to the coast
- Launch sites are far enough away from anything a failed launch can damage (city, highway, railway), while still close enough to bring people and material to support launch activities

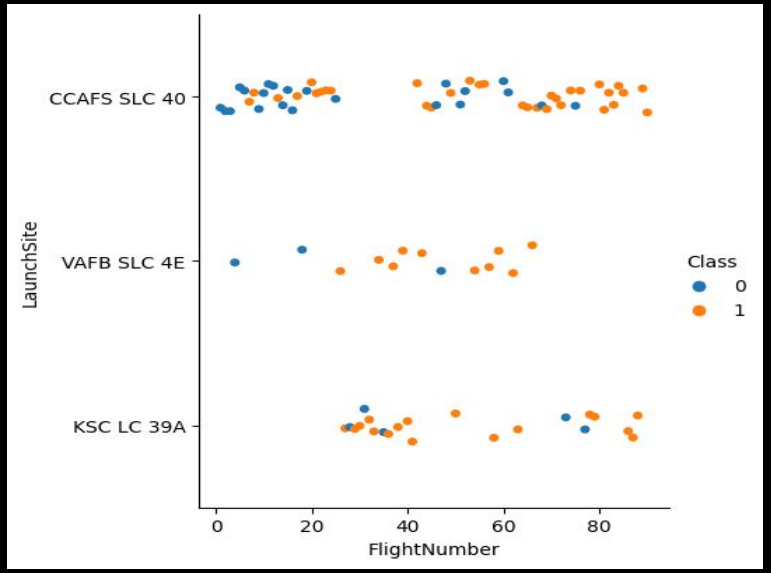
Predictive Analytics

- Decision Tree model is the best predictive model for the dataset

Flight Number vs. Launch Site

Exploratory Data Analysis

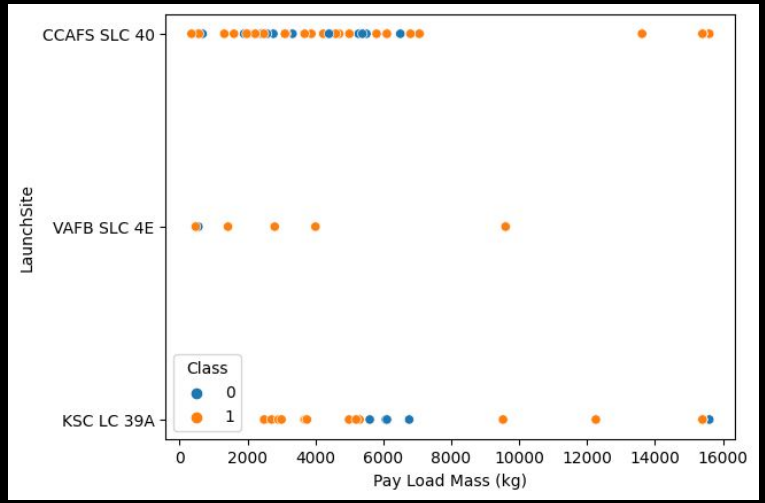
- **Earlier flights** had a **lower success rate** (blue = fail)
- **Later flights** had a **higher success rate** (orange = success)
- Around half of launches were from CCAFS SLC 40 launch site
- VAFB SLC 4E and KSC LC 39A have higher success rates
- We can infer that new launches have a higher success rate



Payload vs. Launch Site

Exploratory Data Analysis

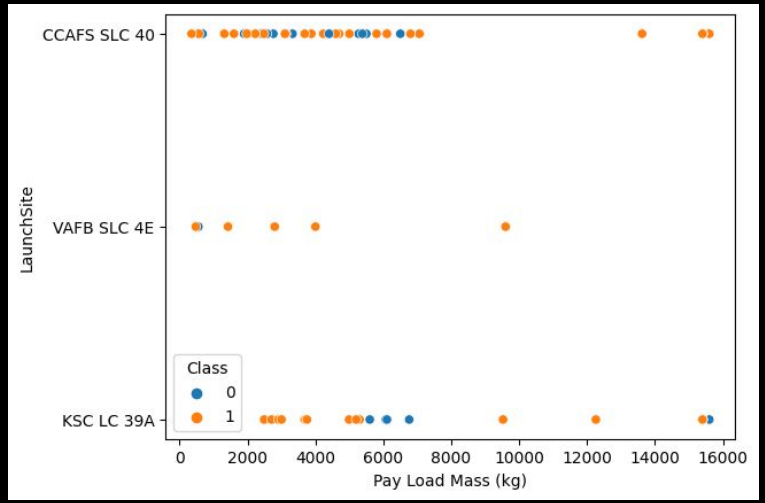
- Typically, the higher the payload mass (kg), the higher the success rate
- Most launches with a payload greater than 7,000 kg were successful
- KSC LC 39A has a 100% success rate for launches less than 5,500 kg
- VAFB SKC 4E has not launched anything greater than ~10,000 kg



Payload vs. Launch Site

Exploratory Data Analysis

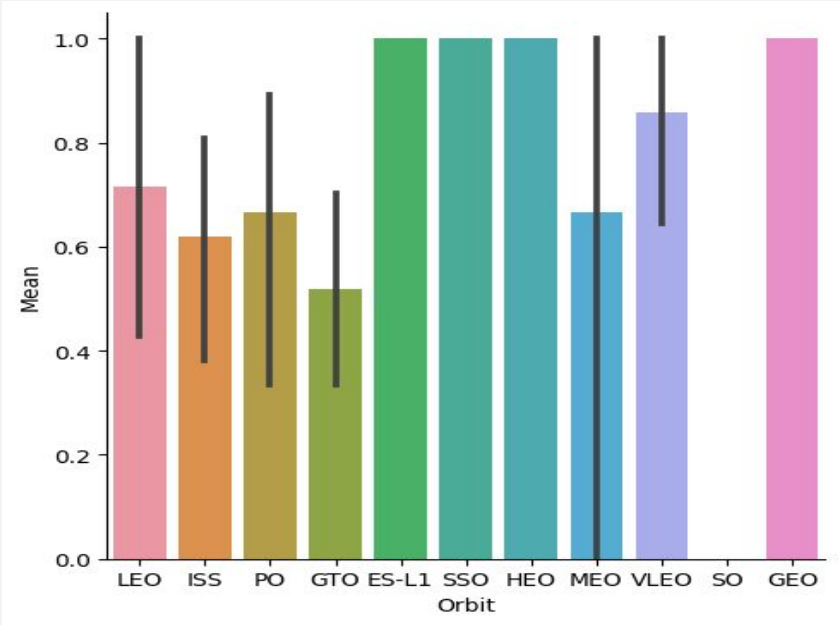
- Typically, the higher the payload mass (kg), the higher the success rate
- Most launches with a payload greater than 7,000 kg were successful
- KSC LC 39A has a 100% success rate for launches less than 5,500 kg
- VAFB SKC 4E has not launched anything greater than ~10,000 kg



Success Rate by Orbit

Exploratory Data Analysis

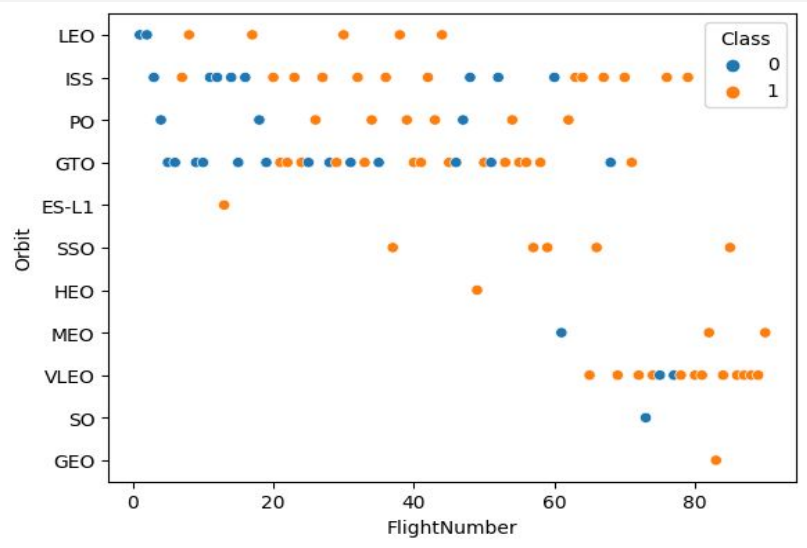
- 100% Success Rate: ES-L1, GEO, HEO and SSO
- .50%-80% Success Rate: GTO, ISS, LEO, MEO, PO
- .0% Success Rate: SO



Flight Number vs. Orbit

Exploratory Data Analysis

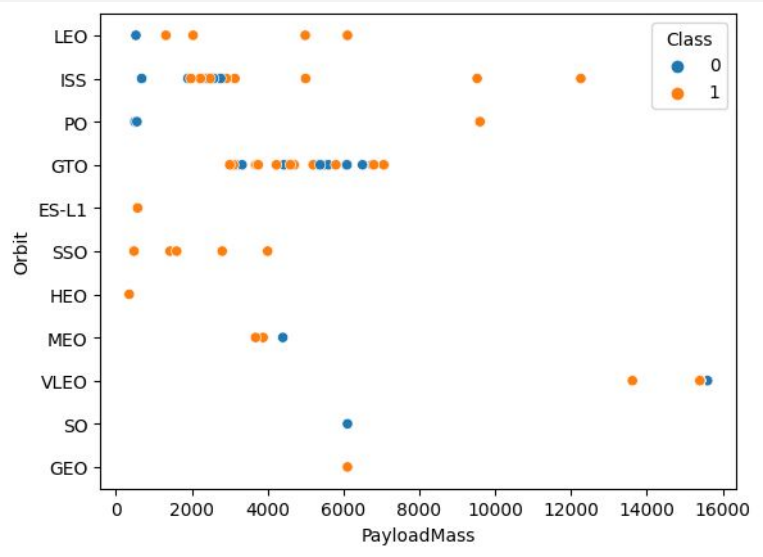
- The success rate typically increases with the number of flights for each orbit
- .This relationship is highly apparent for the LEO orbit
- .The GTO orbit, however, does not follow this trend



Payload vs. Orbit

Exploratory Data Analysis

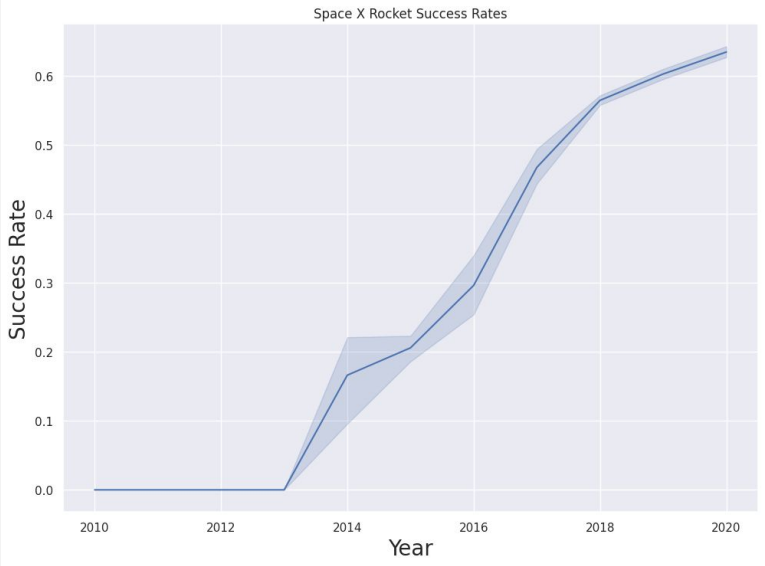
- Heavy payloads are better with LEO, ISS and PO orbits
- .The GTO orbit has mixed success with heavier payloads



Launch Success over Time

Exploratory Data Analysis

- The success rate improved from 2013-2017 and 2018-2019
- .The success rate decreased from 2017-2018 and from 2019-2020
- .Overall, the success rate has improved since 2013



Launch Site Information

Launch Site Names

- CCAFS LC-40
- CCAFS SLC-40
- CCAFS SLC-40
- VAFB SLC-4E

Records with Launch Site Starting with CCA

- Displaying 5 records below

Landing Outcome Cont.

```
%sql ibm_db_sa://nnb49647:BEHttI92iFgf0fTE@764264
%sql SELECT Unique(LAUNCH_SITE) FROM SPACEXTBL
```

```
* ibm_db_sa://nnb49647:***@764264db-9824-4b7c-82df-4
  sqlite:///my_data1.db
Done.
```

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

```
%sql SELECT * FROM SPACEXTBL WHERE launch_site LIKE 'CCA%' LIMIT 5
```

```
* ibm_db_sa://nnb49647:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/BLUDB?security=SSL
  sqlite:///my_data1.db
Done.
```

DATE	time_utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing_outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Landing & Mission Info

1st Successful Landing in Ground Pad

- **12/22/2015**

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

```
* ibm_db_sa://nnb49647:***@764264db-9824-4b7c-82df-
  sqlite:///my_data1.db
Done.
```

1
2015-12-22

Booster Drone Ship Landing

- Booster mass greater than 4,000 but less than 6,000
- JSCAT-14, JSCAT-16, SES-10, SES-11 / EchoStar 105

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

✓ 0.2s

```
* ibm_db_sa://nnb49647:***@764264db-9824-4b7c-82d
  sqlite:///my_data1.db
Done.
```

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

Total Number of Successful and Failed Mission Outcomes

- **1 Failure in Flight**
- **99 Success**
- **1 Success (payload status unclear)**

```
%%sql
SELECT MISSION_OUTCOME,COUNT(*)
FROM SPACEXTBL
GROUP BY MISSION_OUTCOME
```

✓ 0.2s

```
* ibm_db_sa://nnb49647:***@764264db-9824-4b7c-
  sqlite:///my_data1.db
Done.
```

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters

Carrying Max 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)
```

✓ 0.2s

```
* ibm_db_sa://nnb49647:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8
  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,6,2),DATE,
BOOSTER_VERSION,LAUNCH_SITE FROM SPACEXTBL
WHERE substr(DATE,1,4) = '2015'
AND LANDING_OUTCOME ='Failure (drone ship)'
```

✓ 0.2s

```
* ibm_db_sa://nnb49647:***@764264db-9824-4b7c-82df-40d1b13897c2.
  sqlite:///my_data1.db
Done.
```

1	DATE	booster_version	launch_site
10	2015-10-01	F9 v1.1 B1012	CCAFS LC-40
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40

Count of Successful Landings

Ranked Descending

- Count of landing outcomes between 2010-06-04 and 2017-03-20 in descending order

```
%%sql
SELECT COUNT(*) AS RANK,
LANDING_OUTCOME  FROM SPACEXTBL
WHERE DATE BETWEEN '20100604' AND '20170320'
GROUP BY LANDING_OUTCOME
ORDER BY RANK DESC
```

✓ 0.2s

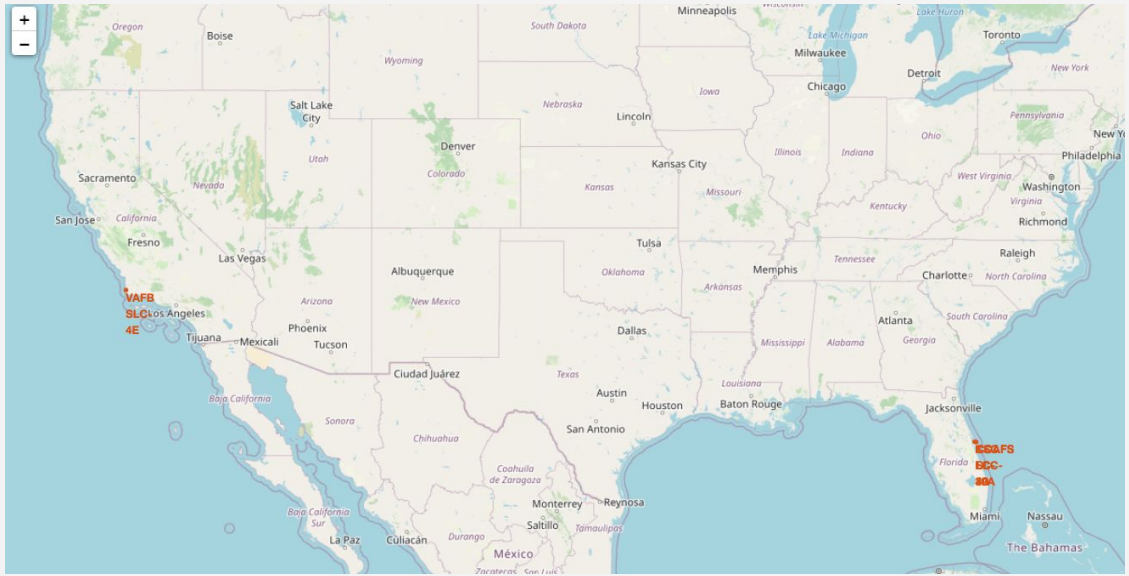
```
* ibm_db_sa://nnb49647:***@764264db-9824-4b7c-82df-
  sqlite:///my_data1.db
Done.
```

RANK	landing_outcome
10	No attempt
5	Failure (drone ship)
5	Success (drone ship)
5	Success (ground pad)
3	Controlled (ocean)
2	Uncontrolled (ocean)
1	Failure (parachute)
1	Precluded (drone ship)

Launch Sites

With Markers

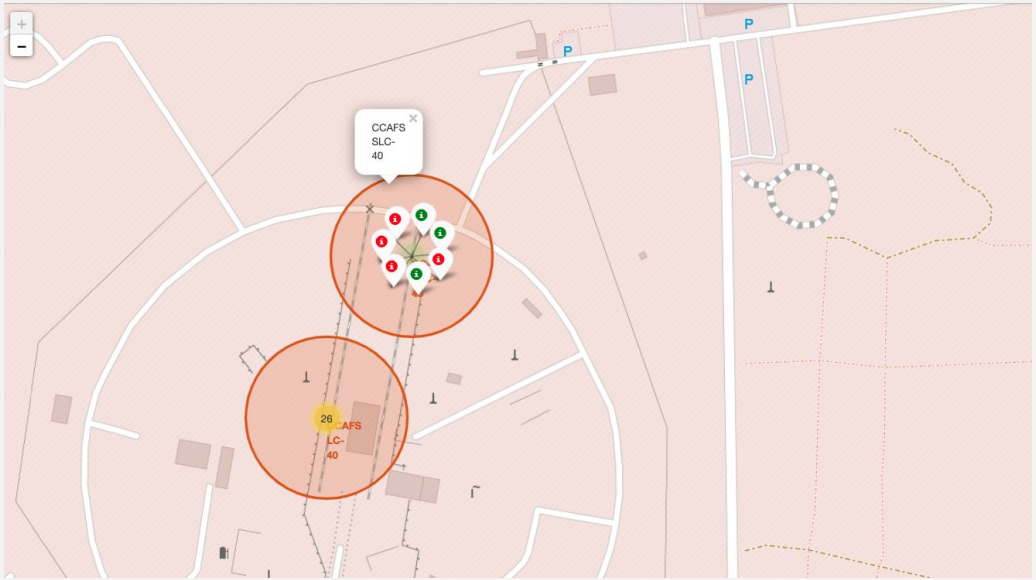
- Near Equator: the closer the launch site to the equator, the easier it is to launch to equatorial orbit, and the more help you get from Earth's rotation for a prograde orbit. Rockets launched from sites near the equator get an additional natural boost - due to the rotational speed of earth - that helps save the cost of putting in extra fuel and boosters.



Launch Outcomes

At Each Launch Site

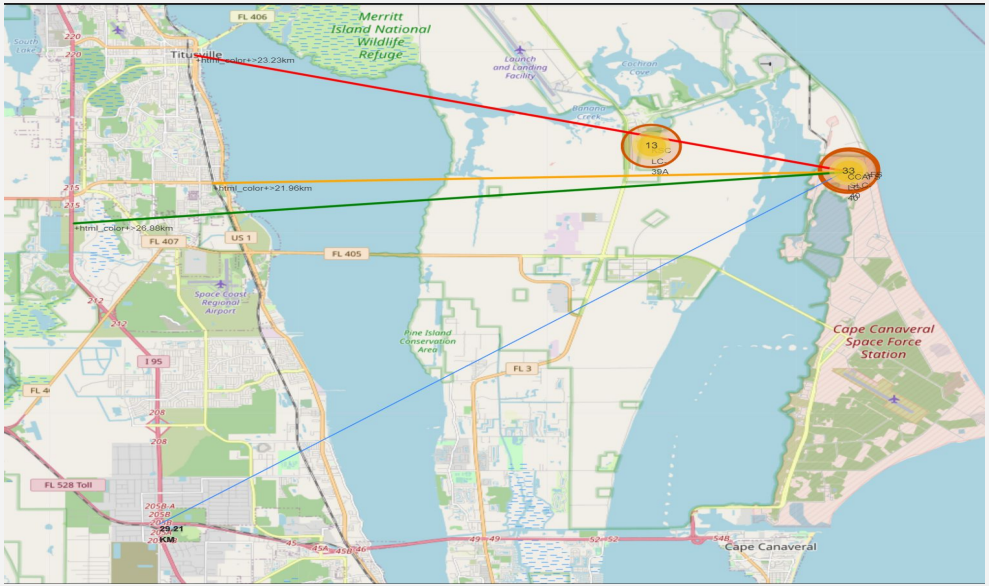
- Green markers for successful launches
- Red markers for unsuccessful launches
- Launch site CCAFS SLC-40 has a 3/7 success rate (42.9%)



Distance to Proximities

CCAFS SLC-40

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

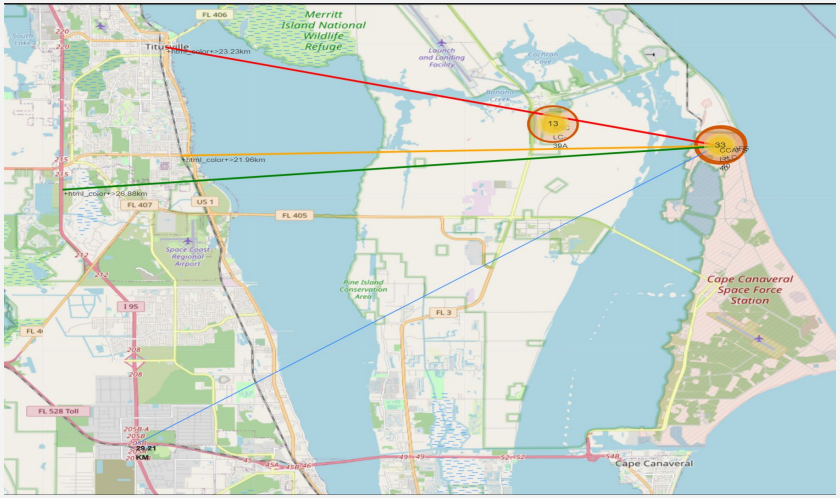


0026

Distance to Proximities

CCAFS SLC-40

- Coasts: help ensure that spent stages dropped along the launch path or failed launches don't fall on people or property.
- Safety / Security: needs to be an exclusion zone around the launch site to keep unauthorized people away and keep people safe.
- Transportation/Infrastructure and Cities: need to be away from anything a failed launch can damage, but still close enough to roads/rails/docks to be able to bring people and material to or from it in support of launch activities.

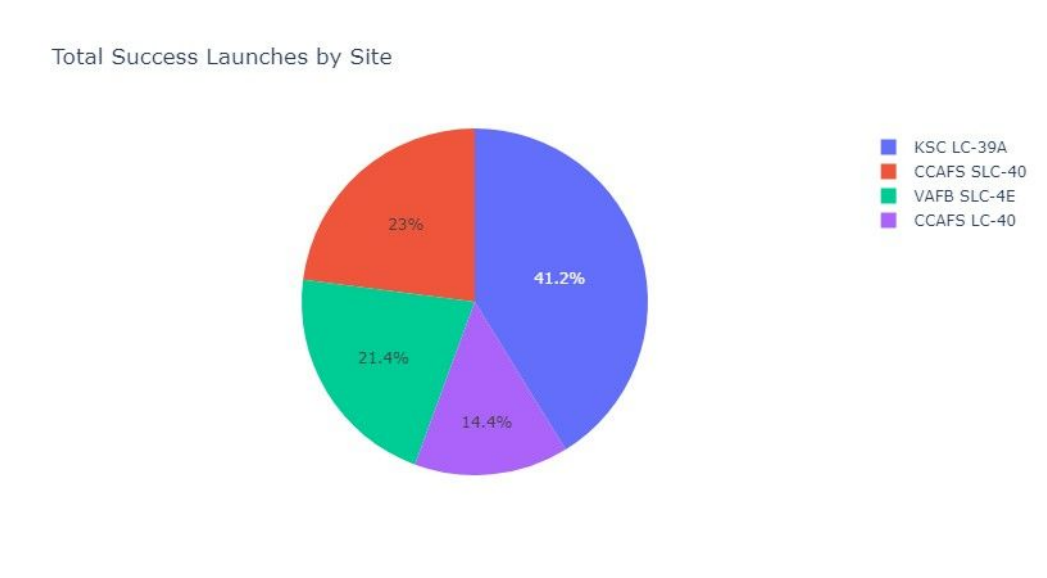


0026

Launch Success by Site

Success as Percent of Total

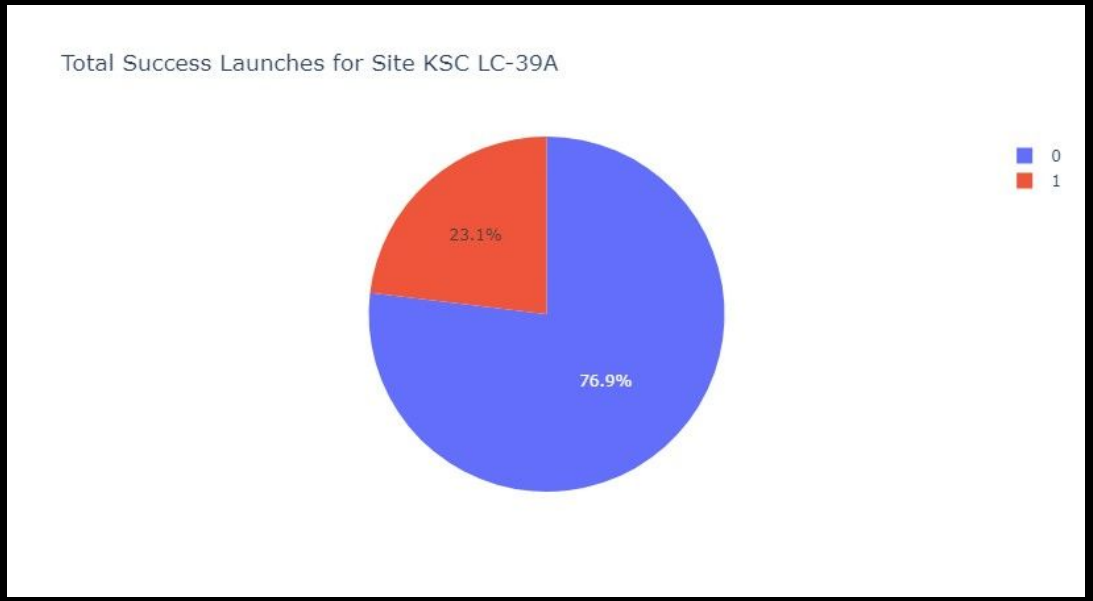
- KSC LC-39A has the most successful launches amongst launch sites (41.2%) .



Launch Success (KSC LC-29A)

Success as Percent of Total

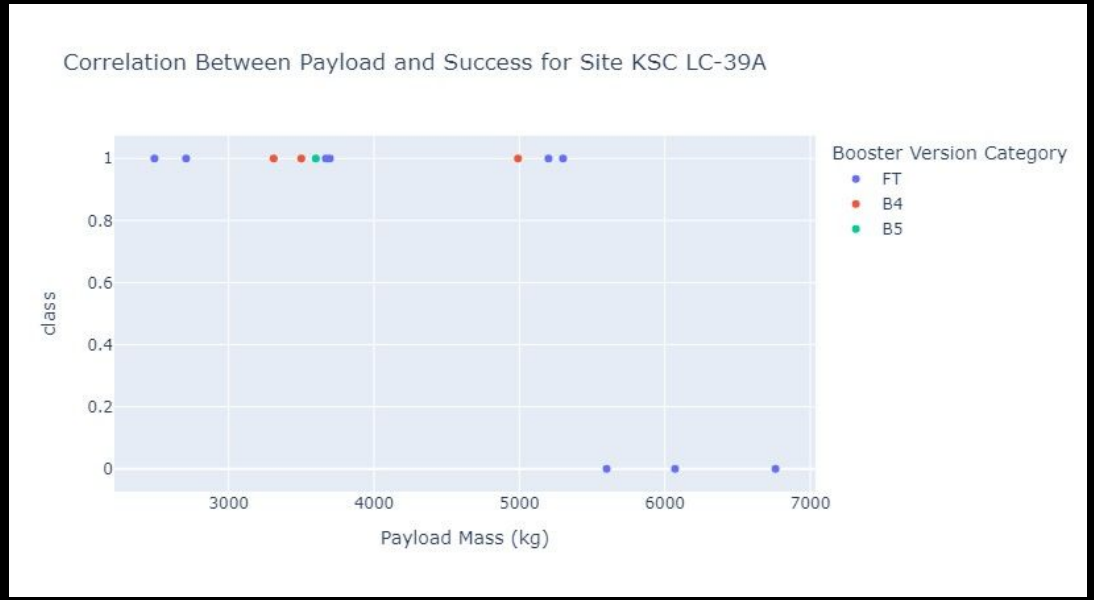
- KSC LC-39A has the highest success rate amongst launch sites (76.9%)
- 10 successful launches and 3 failed launches



Payload Mass and Success

By Booster Version

- Payloads between 2,000 kg and 5,000 kg have the highest success rate
- 1 indicating successful outcome and 0 indicating an unsuccessful outcome



Classification

Accuracy

- All the models performed at about the same level and had the same scores and accuracy. This is likely due to the small dataset. The Decision Tree model slightly outperformed the rest when looking at .best_score_
- .best_score_ is the average of all cv folds for a single combination of the 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_)
```

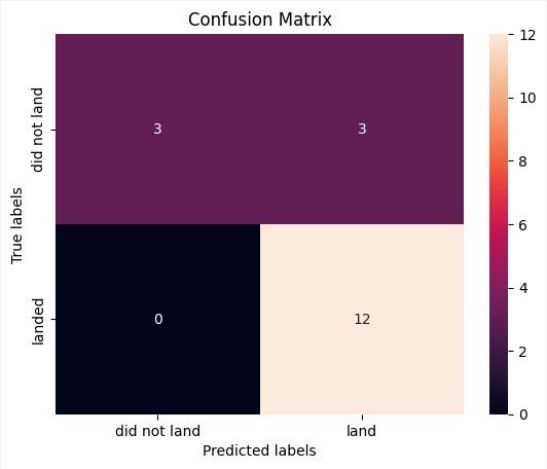
```
Best model is DecisionTree with a score of 0.8767857142857143
Best params is : {'criterion': 'gini', 'max_depth': 18, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'splitter': 'random'}
```

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

Confusion Matrices

Performance Summary

- A confusion matrix summarizes the performance of a classification algorithm
- All the confusion matrices were identical
- The fact that there are false positives (Type 1 error) is not good
- Confusion Matrix Outputs :
  - 12 True Positive
  - 03 True Negative
  - 03 False Positive
  - 12 False Negative



- Precision = TP / (TP + FP)
  - 12 / 15 = .80
- Recall = TP / (TP + FN)
  - 12 / 12 = 1
- F1 Score = 2 * (Precision * Recall) / (Precision + Recall)
  - 2 * (.8 * 1) / (.8 + 1) = .89
- Accuracy = (TP + TN) / (TP + TN + FP + FN) = .833

Conclusion

- **Model Performance** : The models performed similarly on the test set with the decision tree model slightly outperforming
- **Equator** : Most of the launch sites are near the equator for an additional natural boost - due to the rotational speed of earth - which helps save the cost of putting in extra fuel and boosters
- **Coast** : All the launch sites are close to the coast
- **Launch Success** : Increases over time
- **KSC LC-39A** : Has the highest success rate among launch sites. Has a 100% success rate for launches less than 5,500 kg
- **Orbits** : ES-L1, GEO, HEO, and SSO have a 100% success rate
- **Payload Mass** : Across all launch sites, the higher the payload mass (kg), the higher the success rate