SpaceX
first stage reuse

Joshua B. Buttery

Presentation Contents

- Executive Summary
- Introduction
- Methodology
- Results

-EDA with visualisations

-EDA with SQL

-Interactive Maps with Folium

-Plotly Dash Dashboard

-Predictive Analytics

- Conclusion

Executive summary

Executive Summary

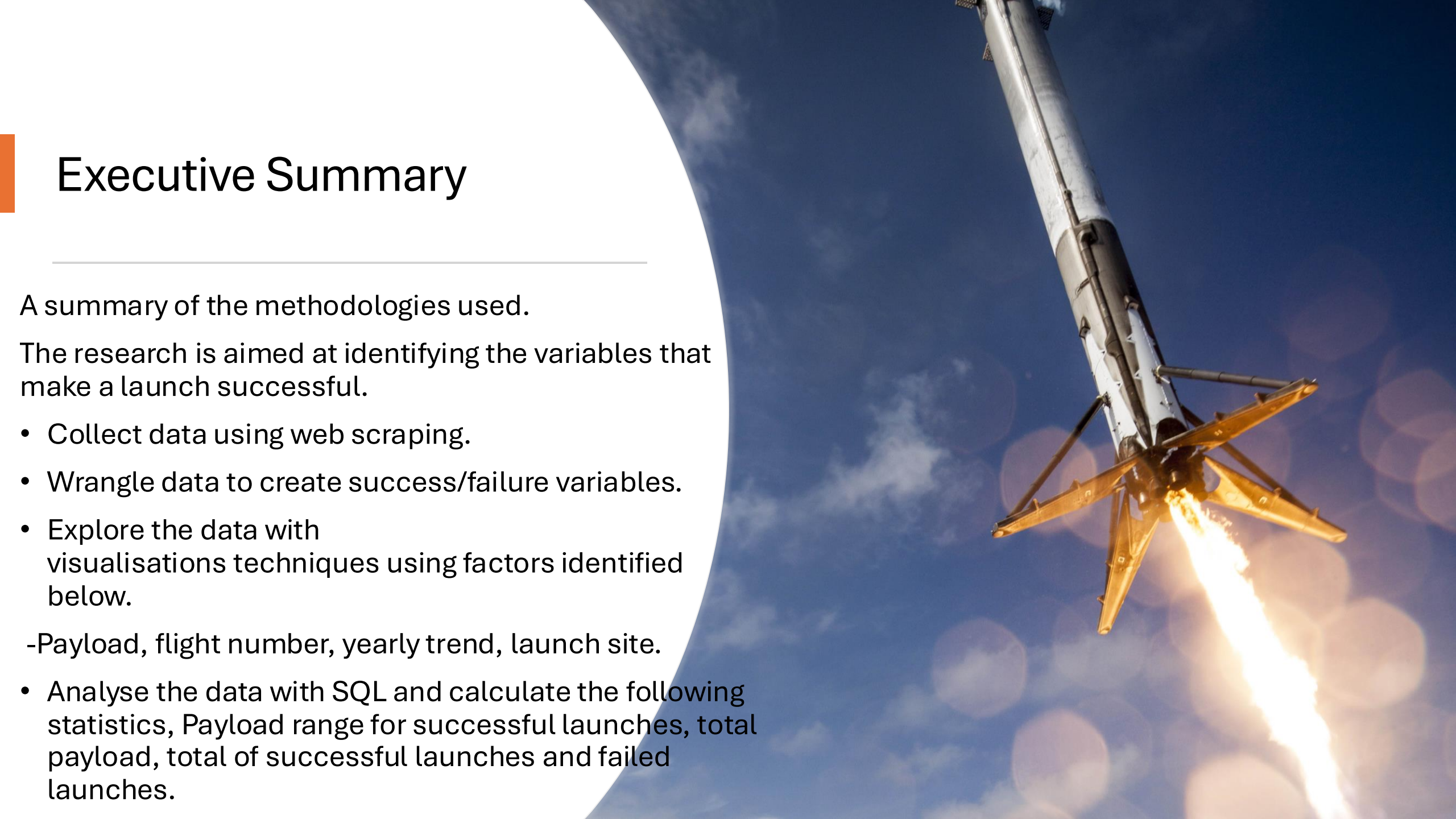
A summary of the methodologies used.

The research is aimed at identifying the variables that make a launch successful.

- Collect data using web scraping.
- Wrangle data to create success/failure variables.
- Explore the data with visualisations techniques using factors identified below.

-Payload, flight number, yearly trend, launch site.

- Analyse the data with SQL and calculate the following statistics, Payload range for successful launches, total payload, total of successful launches and failed launches.



Executive Summary

- Explore launch site success rate with geographical markers.
- Visualize the launch sites with the most successful payload launches and ranges.
- Build models to analyze the data and predict landing outcomes using logistic regression, decision tree, k-nearest Neighbour (KNN) and support vector machine (SVM).

Executive Summary Results

- Exploratory data analysis

-KSC LC-39A has the highest success rate among all landing sites.

-Launch success has improved over time.

-Orbits ES-L1, HEO, SSO, GEO have 100% success rates.

- Visualization and Analytics

-Most launch sites are situated near the equator and near the coast.

- Predictive Analytics

-The decision tree model outperformed a bit. All models performed similar on the test set.

Introduction

Introduction

In the past few decades SpaceX has become a leader in innovation in the space sector. It is known for sending up multiple satellites into orbit and sending payloads to the ISS. Some of its launches have been manned also and using multiple satellites to also internet access worldwide.
The average to launch a rocket with SpaceX in $62 million per launch which is achieved by reusing the first stage of the falcon 9 rocket. Other competitors cost upwards of $165 million per launch. By using data and factors involved in the launch we can determine the cost of the launch. We can use open source data to create machine models to predict whether SpaceX or indeed a competitor can reuse the first stage.

- To do this we will explore how payload mass, number of flights, launch site and orbit affects the success of the first stage landing intact.
- We will determine the success of landing over a range of time.
- Create a predictive binary model for landings that have been successful.

Methodology

Methodology

Steps

- Collect data using web scraping and SpaceX REST API.
- Wrangle the data by handling missing values, filtering data and applying one hot coding for the next steps which involves analysis and modelling.
- Explore the data with SQL and EDA and data visualization techniques.
- Visualize the data with Plotly Dash and Folium.
- Build models to predict the landing success of the first stage and tune the models to best fit the parameters.

Data Collection

Steps

- Request data from SpaceX API. The rocket launch data.
- Decode response using .json() and convert to a data frame using .json_normalize()
- Request information about the launches from SpaceX API using custom functions.
- Create dictionary from the data.
- Create data frame from the dictionary.
- Filter data frame to show only falcon 9 launches.
- Replace missing values of payload mass with the calculated .mean()
- Export data to CSV file.

Data Collection (Web scraping)

Steps

- Request data for falcon 9 launches 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.
- Create data frame from the dictionary.
- Export data to CSV file.

Data Wrangling

Steps

- Perform EDA and determine data labels.
- Calculate, launches from each site, mission outcome per orbit type, occurrence of orbit.
- Create binary landing outcome column. (dependent variable).
- Export data to CSV file.

Landing outcome (Wrangling continued)

- True Ocean- Mission outcome was a successful landing to a specific region of the ocean.
- False Ocean- Represents an unsuccessful landing in a region of the ocean.
- True RTLS- Means the mission had a successful landing on a ground pad.
- False RTLS- Means the mission has an unsuccessful landing on a ground pad.
- True ASDS- Means the mission outcome had a successful landing on a drone ship.
- False ASDS- Means the mission outcome had an unsuccessful landing on a drone ship.
- Outcomes Converted into 1 for successful landing and 0 for unsuccessful landing.

Results

EDA with visualizations

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 relationship among the categories and a measured value.

EDA with SQL

Queries

- Names of the launch sites.
- 5 records where launch site begins with "CCA".
- Total payload mass carried by the boosters launched by NASA (CRS).
- Average payload mass carried by booster version Falcon 9 Vs. 1.1.

EDA with SQL

List

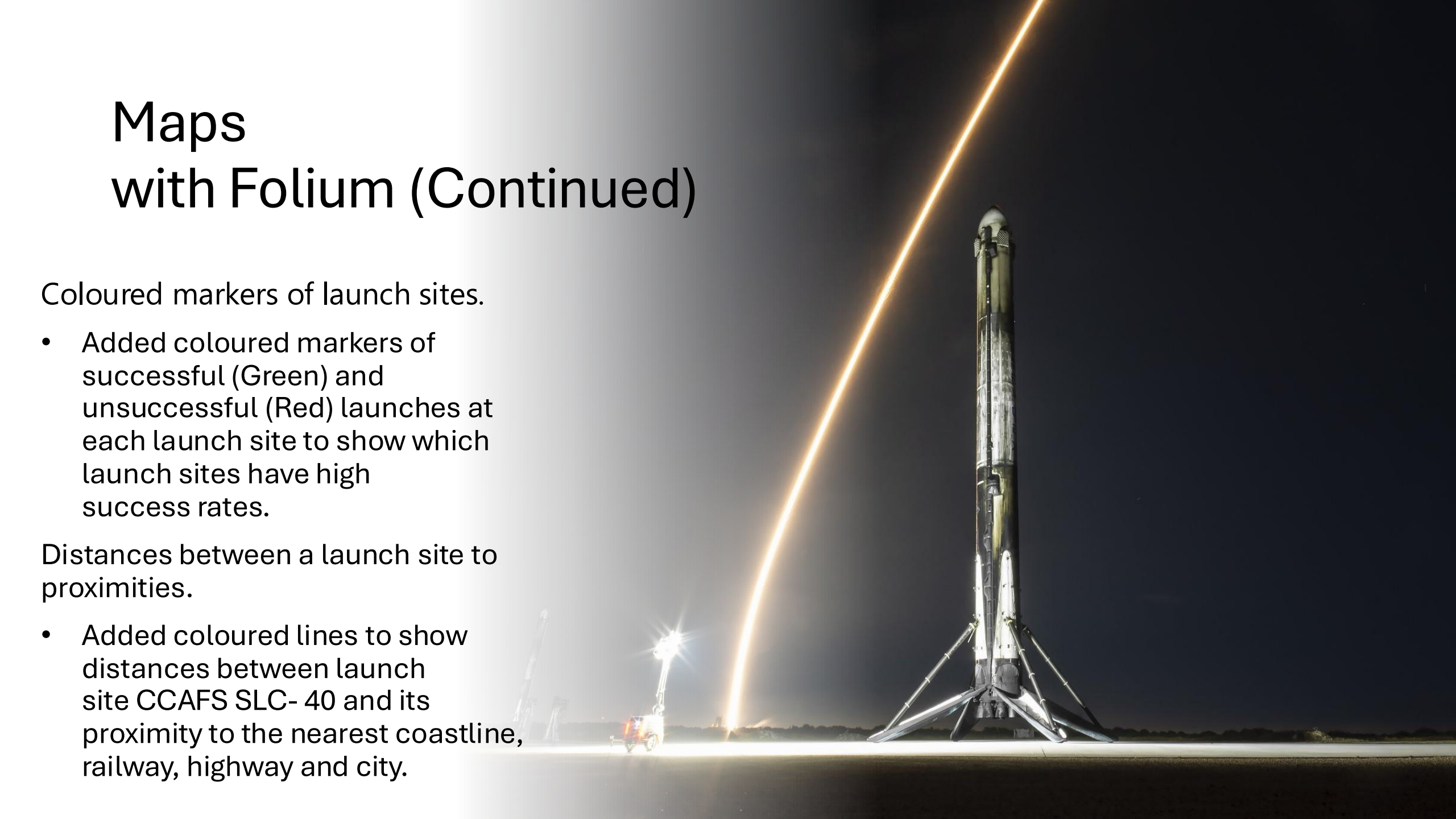
- Date of the first successful landing on a ground pad.
- Names of boosters which had successful landings on drone ships and have a payload mass of greater than 4000 but less than 6000.
- Total number of successful and failed missions.
- Names of booster versions which have carried the max payload.
- Failed landing outcomes on drone ship, their booster version and launch site for the months in the year 2015.
- Count of landing outcomes between 2010/06/04 and 2017/03/20.

Maps with Folium

Markers indicating land sites

- Added Blue markers at NASA Johnson Space Centre's Co-ordinates with a pop up label showing its name using its latitude and longitude co-ordinates.
- Added red markers at all launch sites co-ordinates with a pop up label showing its name using it's co-ordinates of latitude and longitude.

Maps
with Folium (Continued)



Coloured markers of launch sites.

- Added coloured 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 coloured lines to show distances 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.

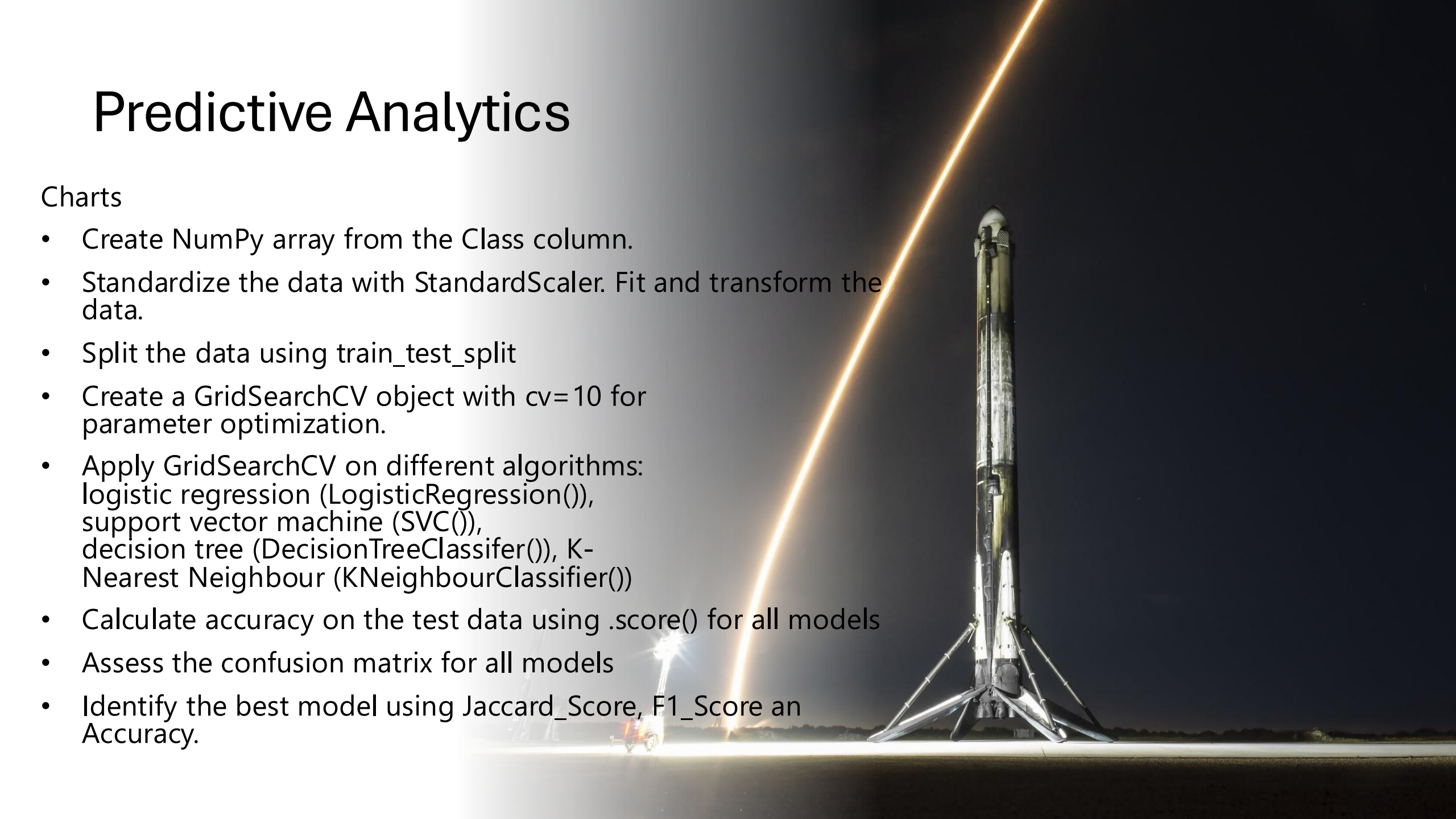
Scatter chart showing payload mass Vs. Success rate by booster version

- Allow user to see the correlation between payload and launch success.

Predictive Analytics

Charts

- Create NumPy array from the Class column.
- Standardize the data with StandardScaler. Fit and transform the data.
- Split the data using train_test_split
- Create a GridSearchCV object with cv=10 for parameter optimization.
- Apply GridSearchCV on different algorithms: logistic regression (LogisticRegression()), support vector machine (SVC()), decision tree (DecisionTreeClassifer()), K-Nearest Neighbour (KNeighbourClassifier())
- Calculate accuracy on the test data using .score() for all models
- Assess the confusion matrix for all models
- Identify the best model using Jaccard_Score, F1_Score an Accuracy.



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.

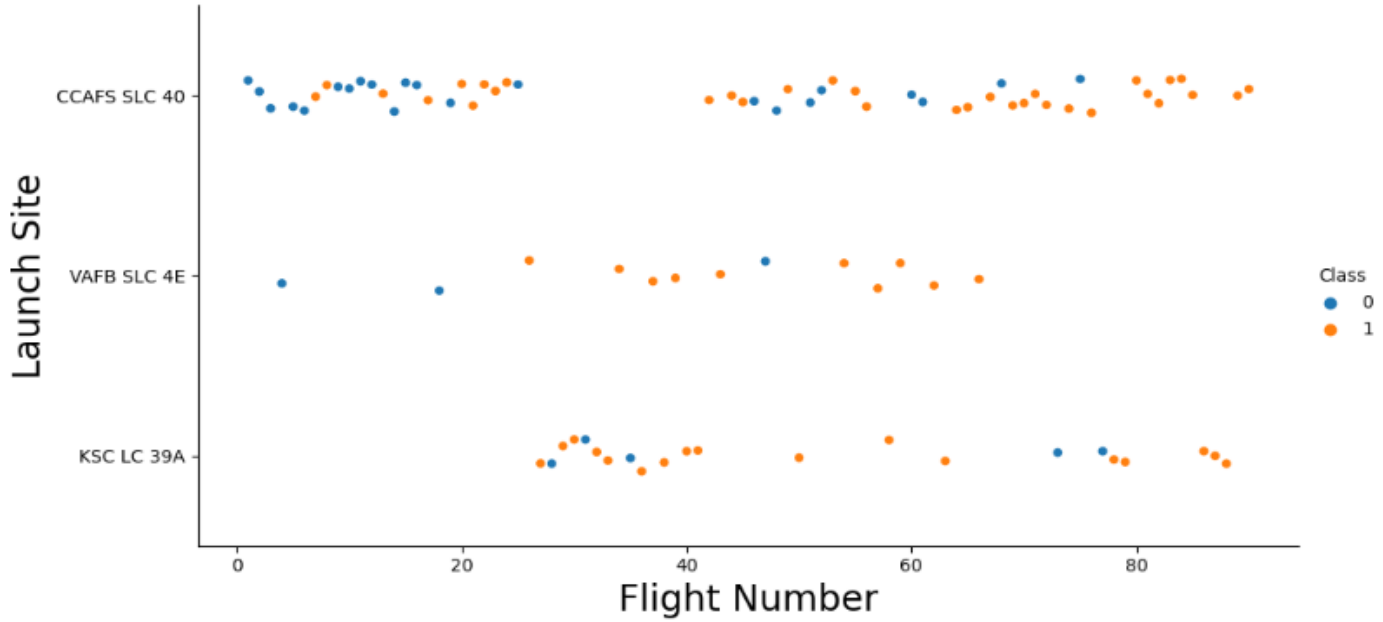
Predictive Analytics

- Decision tree model is the best predictive model for the dataset.

Flight number Vs. Launch site

Exploratory data analysis

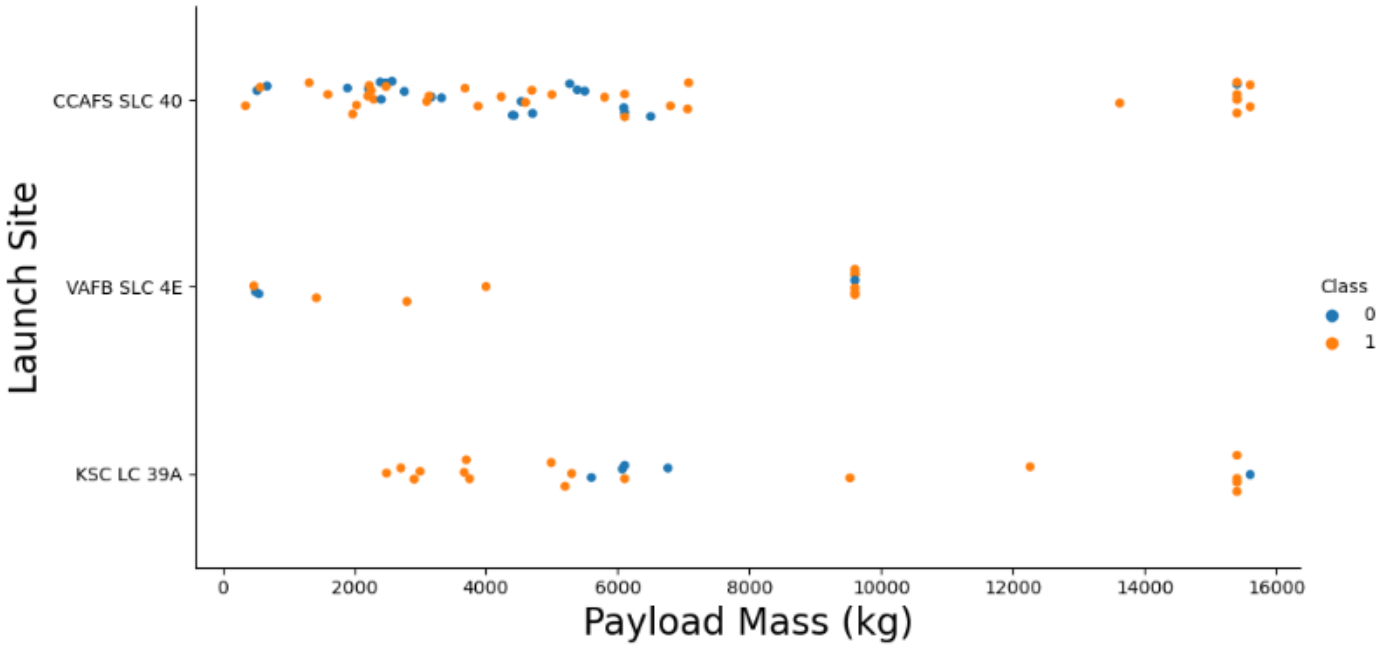
- Earlier flights had a lower success rate (blue=fail).
- Later flights had a higher success rate (orange=success).
- Around half of launches were from CCAFS SLC 40 launch site.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- We can infer that new launches have a higher success rate.



Payload Vs. Launch site

Exploratory data analysis

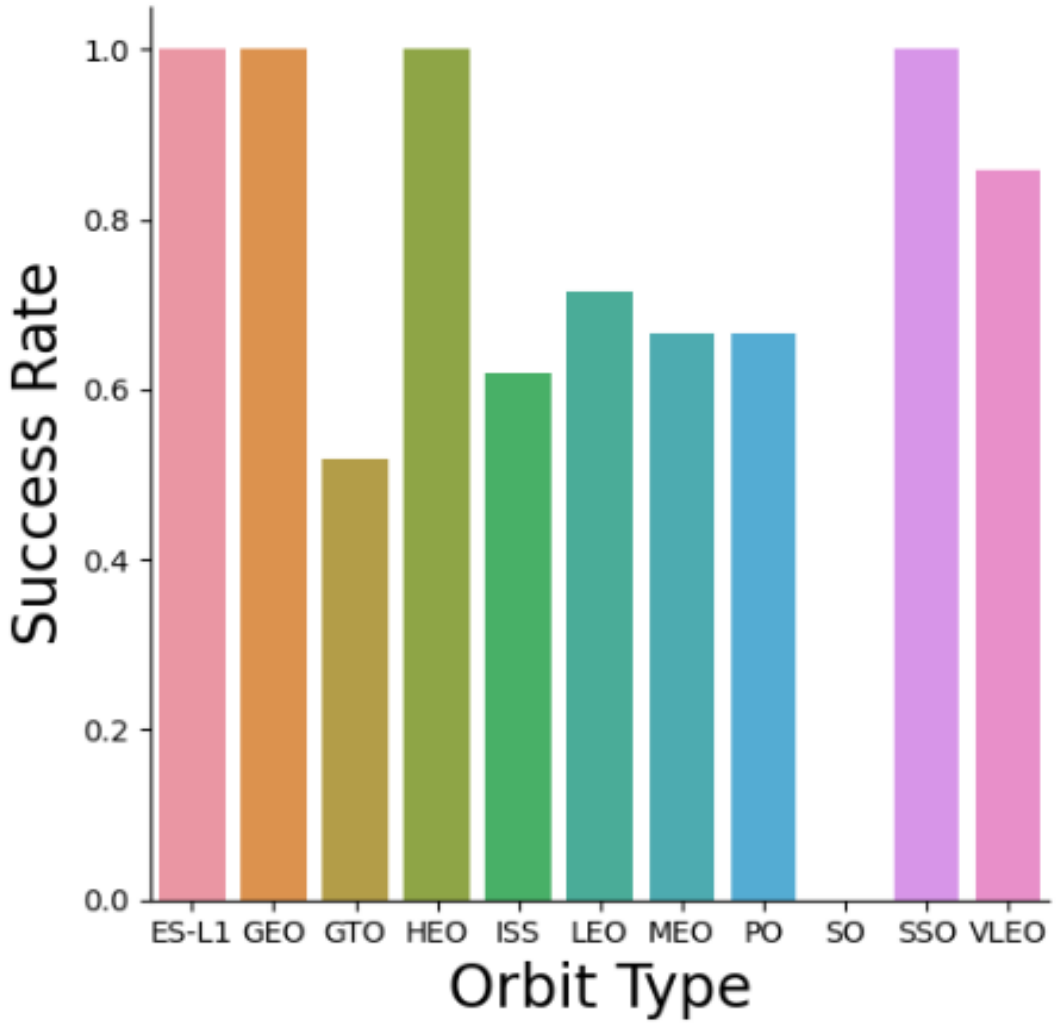
- Typically, the higher the payload mass (KG), the higher the success rate.
- Most launches with a payload greater than 7,000 KG were successful.
- KSC LC 39A has a 100% success rate for launches less than 5,000 KG.
- VAFB SKC 4E has not launched anything greater than ~ 10,000 KG.



Success rate by orbit

Exploratory data analysis

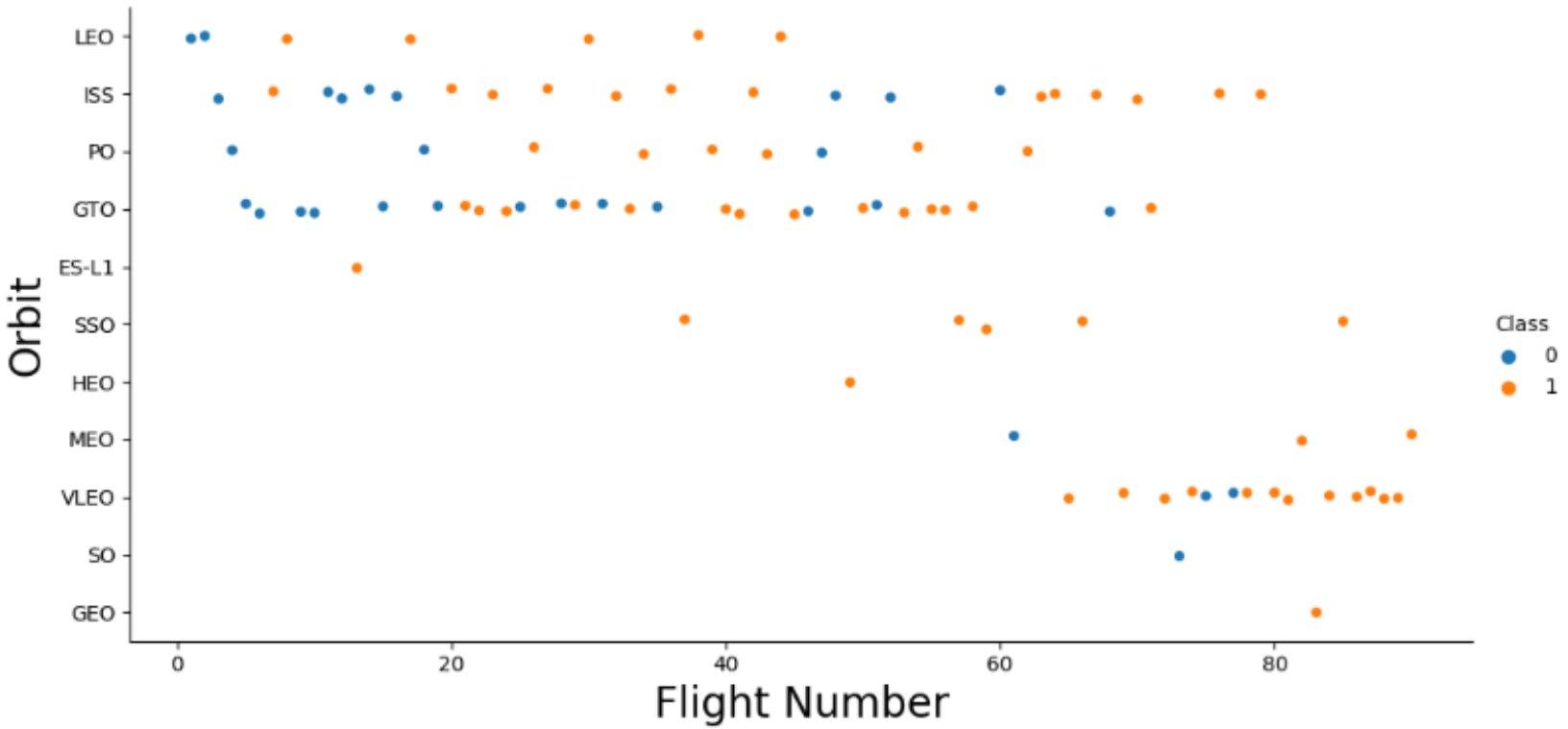
- 100% Success Rate : ES-L1, GEO, HEO and SSO.
- 50% - 80% Success Rate : GTO, ISS, LEO, MEO, PO.
- 0% Success Rate ; SO.



Flight number Vs. Orbit

Exploratory data analysis

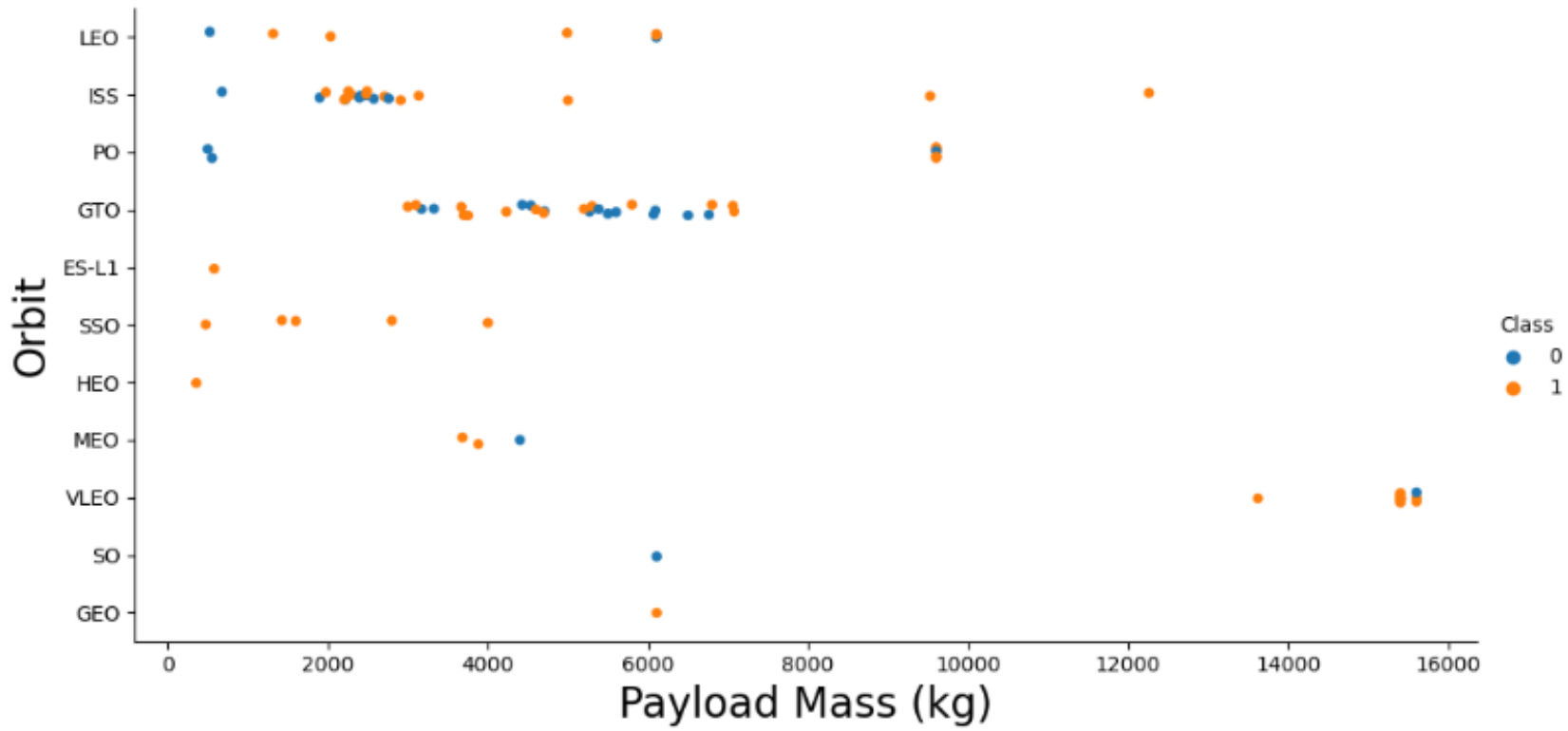
- The success rate typically increases with the number of flights for each orbit.
- This relationship is highly apparent for the LEO orbit.
- The GTO orbit, however, does not follow this trend.



Payload Vs. Orbit

Exploratory data analysis

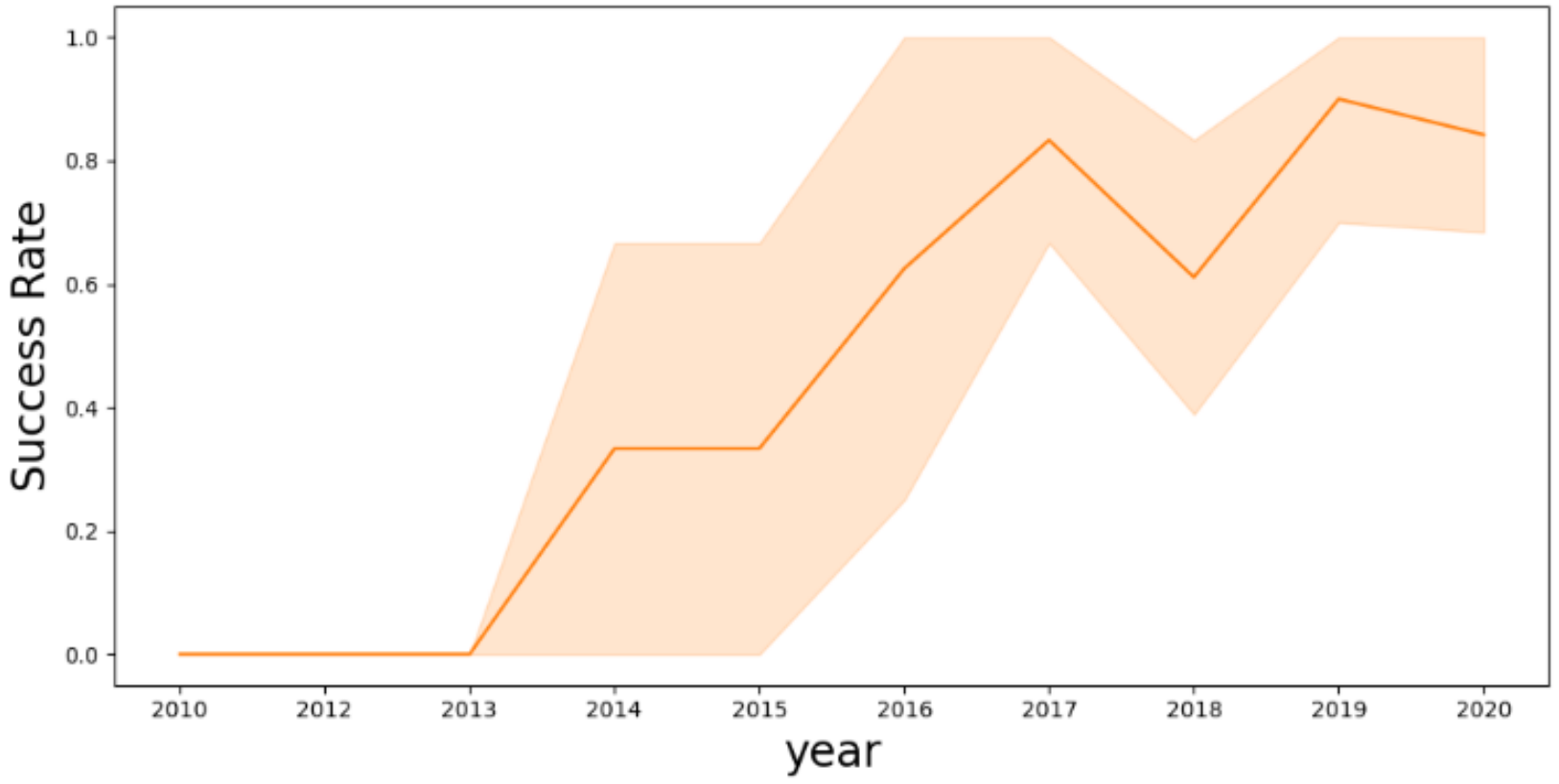
- Heavy payloads are better with LEO, ISS and PO orbits.
- The GTO orbit has mixed success with heavier pa yloads.



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

Records with launch sites starting with CCA
-Displaying 5 records below.

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

```
 * ibm_db_sa://yyy10586:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c1ogj3sd0tgtu0lqde00.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).

```
%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

Average payload mass

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

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

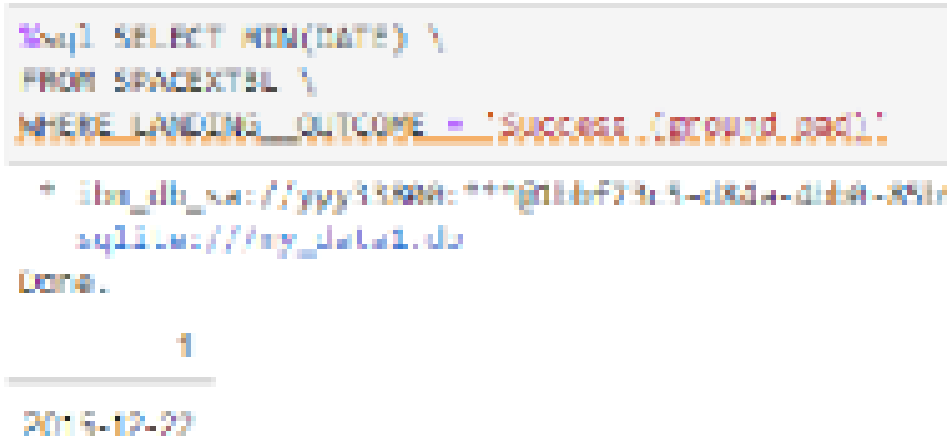
```
 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4
   sqlite:///my_data1.db
Done.
```

1
2928

Landing and mission information

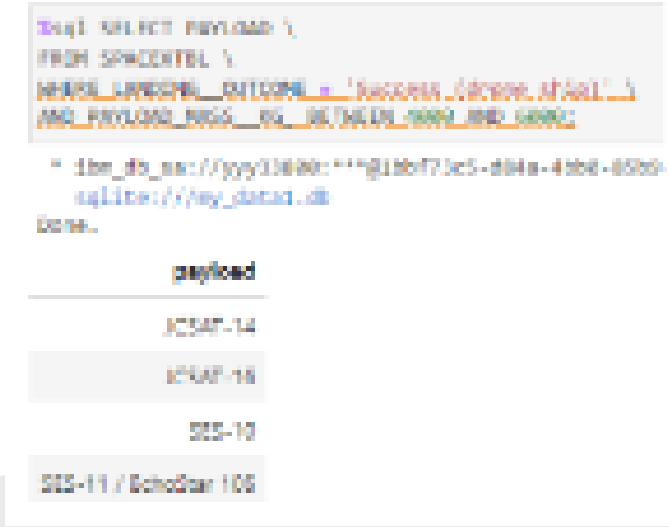
First successful landing on a ground pad

- 12/22/2015



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



Landing and mission information

Total number of successful and failed mission outcomes

- 1 Failure in flight.
- 99 Success.
- 1 Success (payload status unclear).

```
%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

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);
```

```
 * 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 descending

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

```
%sql SELECT [Landing _Outcome], ccunt(*) 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

With markers

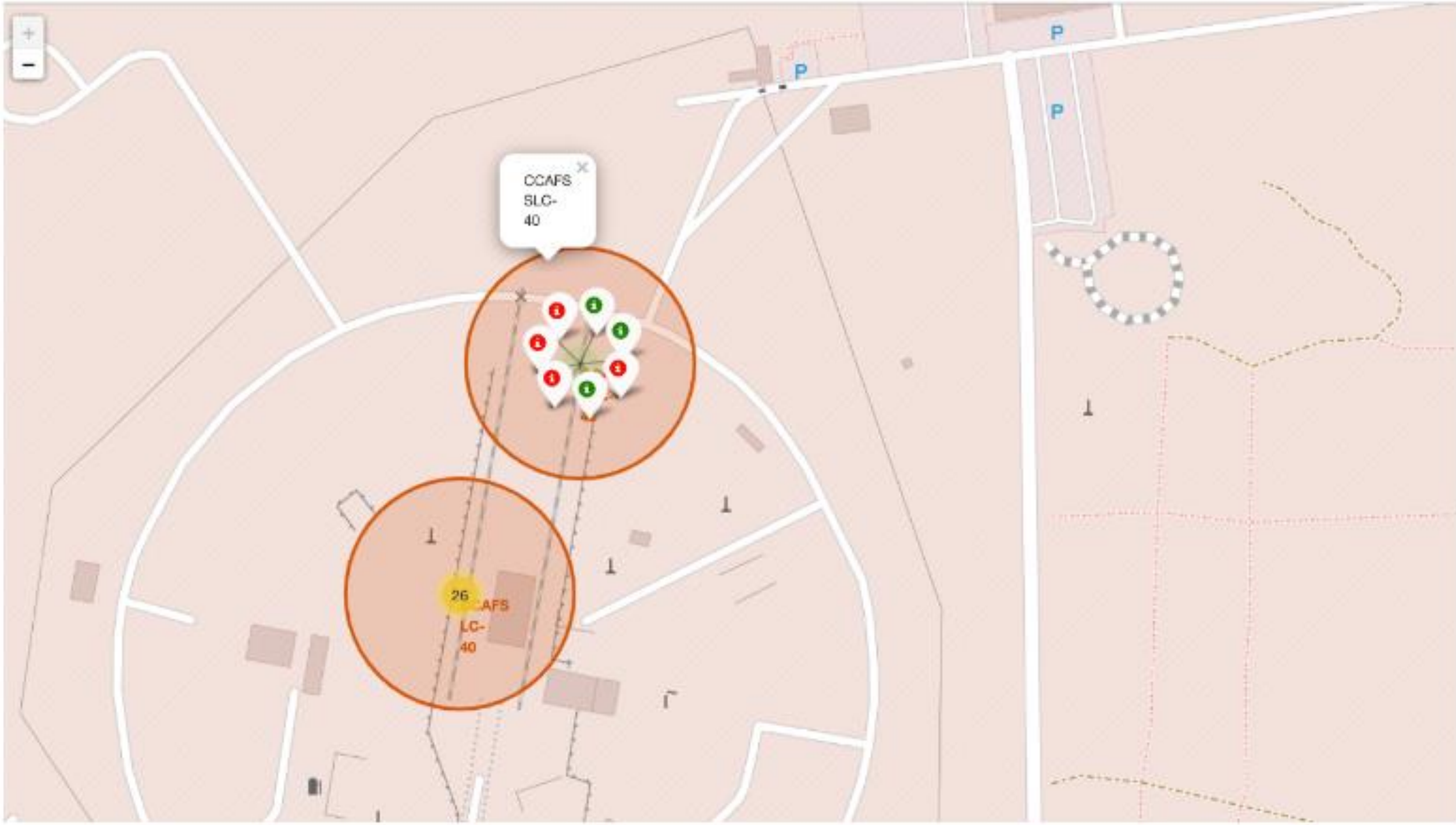
- Near equator: the closer the launch site is to the equator the easier it is to launch to equatorial orbit and the more help you get from earth's rotation prograde orbit.
- Rockets launched from sites near the equator get an additional natural boost - due to the rotational speed of the earth – that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


+
−
Oregon
Boise
South Dakota
Minneapolis
Toronto
Wyoming
Milwaukee
Detroit
Chicago
New York
Salt Lake City
Nebraska
Iowa
Lincoln
Pennsylvania
Ohio
Denver
Utah
Colorado
Illinois
Indiana
Philadelphia
Kansas City
Sacramento
Nevada
Kansas
Missouri
West Virginia
Washington
Kentucky
Virginia
San Jose
California
Richmond
Fresno
Tulsa
Las Vegas
Albuquerque
Tennessee
Raleigh
Oklahoma
Memphis
Charlotte
North Carolina
Arkansas
VAFB
SLC
4E
Los Angeles
Arizona
New Mexico
Atlanta
South Carolina
Phoenix
Dallas
Tijuana
Mexicali
Tucson
Mississippi
Alabama
Georgia
Ciudad Juárez
Texas
Louisiana
Austin
Houston
Baton Rouge
Jacksonville
Baja California
Sonora
San Antonio
Chihuahua
ECAFS
ECC-
40A
Florida
Coahuila de Zaragoza
Monterrey
Reynosa
Baja California Sur
Saltillo
Tamaulipas
Miami
Nassau
La Paz
Culiacán
Durango
México
The Bahamas

Launch outcomes

At each launch site

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


CCAFS
SLC-
40
26
AFS
LC-
40
P
P
P

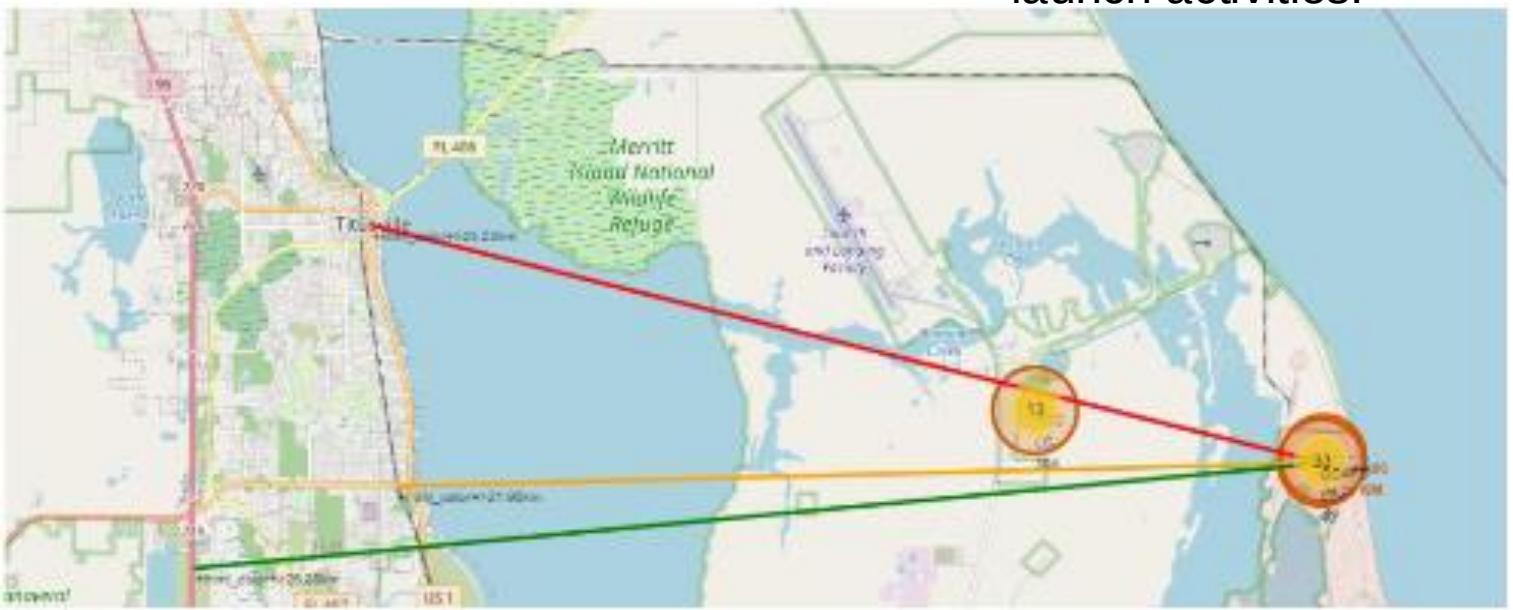
Distance to proximities

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.

CCAFS SLC-40

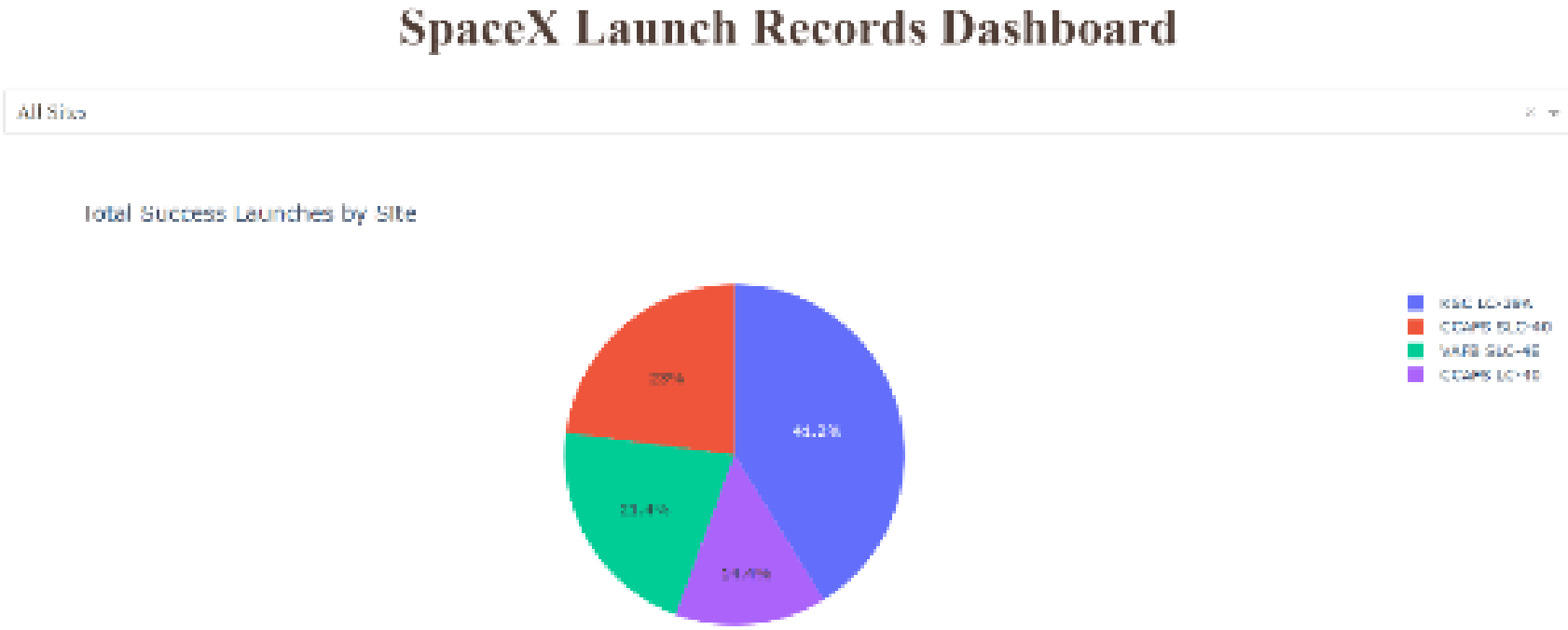
- Coasts: help ensure that spent stages dropped along the launch path or failed launches don't fall on people's property.
- Safety/Security: needs to be an exclusion zone around the launch site to keep unauthorized people away and keep people safe.
- Transportation/infrastructure and cities: need to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



Launch success by site

Success as percent of total

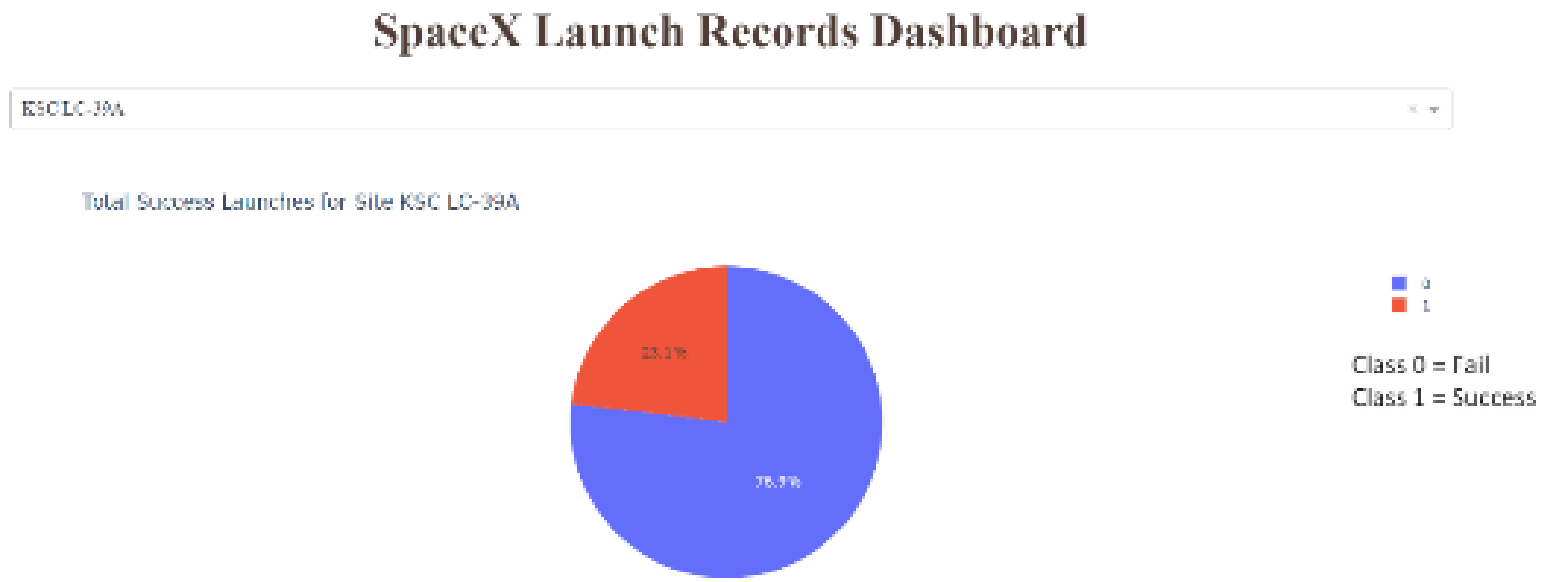
- KSC LC-39A has the most successful launches amongst launch sites (41.2%).



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.

Predictive analytics

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 data set. 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.

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

In [32]:

```
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 0.9017857142857144
Best params is : {'criterion': 'gini', 'max_depth': 18, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_spl
it': 2, 'splitter': 'best'}
```

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

- 3 True negative

- 3 False positive

- 0 False negative

- Precision = TP/(TP+FP)

-12/15=80

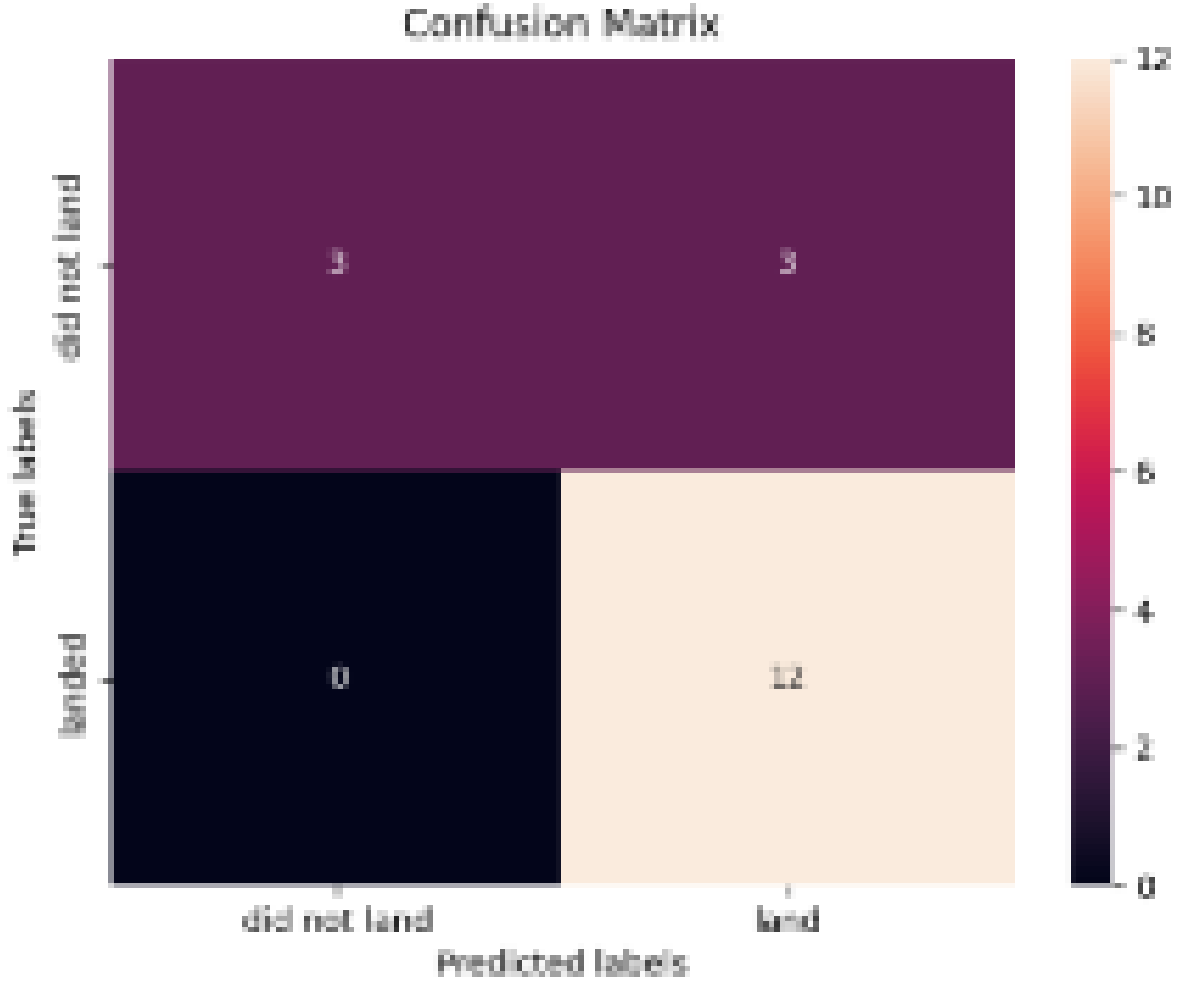
- Recall =TP/(TP+FN)

-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

Conclusion

Research

- 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 the 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.

Conclusion

Things to consider

- Dataset: A larger dataset will help build on the predictive analytics results to help understand if the findings can be generalized to a larger data set.
- Feature Analysis/
  PCA: Additional feature analysis or principal component analysis should be conducted to see if it can help improve accuracy.
- XGBoost: Is a powerful model which was not utilized in this study. It would be interesting to see if it outperforms the other classification models.