



IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Carolina Herrera
08/12/2024



Outline

| | |
|-------------------|----|
| Executive Summary | 3 |
| Introduction | 4 |
| Methodology | 5 |
| Results | 15 |
| Conclusion | 42 |



Executive Summary

This research focuses on identifying the key factors contributing to a successful rocket landing. The following methodologies were employed:

Data Collection: Utilized SpaceX REST API and web scraping to gather data.

Data Wrangling: Processed and cleaned the data to create a success/fail outcome variable.

Exploratory Data Analysis (EDA):

Visualized data to explore the impact of factors such as payload, launch site, flight number, and yearly trends.

Examined launch site success rates and their proximity to geographical markers.

SQL Analysis: Calculated key statistics, including total payload, payload range for successful launches, and total counts of successful and failed outcomes.

Data Visualization: Highlighted launch sites with the highest success rates and identified successful payload ranges.

Predictive Modeling: Built models to predict landing outcomes using logistic regression, support vector machines (SVM), decision trees, and K-nearest neighbor (KNN).

Exploratory Data Analysis Insights:

Rocket launch success rates have improved significantly over time.

The launch site *KSC LC-39A* has the highest success rate.

Certain orbits (ES-L1, GEO, HEO, and SSO) consistently achieve a 100% success rate.

Visualization & Analytics: Most launch sites are near the equator and close to the coast, aligning with optimal launch conditions.

Predictive Analytics: All models demonstrated similar performance on the test set, with the decision tree model slightly outperforming the others.

Introduction

SpaceX, a leader in the space industry, is committed to making space travel more accessible and affordable. Its groundbreaking achievements include sending spacecraft to the International Space Station, deploying a satellite constellation for global internet access, and conducting manned space missions. A major factor behind SpaceX's success is its innovative reuse of the Falcon 9 rocket's first stage, reducing launch costs to approximately \$62 million, compared to over \$165 million for competitors unable to reuse their rockets.

This project aims to address the following questions:

- What are the key factors that determine the success of a first-stage rocket landing?
- How accurately can machine learning models predict the reusability of the first stage?
- How can these predictions influence the overall cost-efficiency of space launches?
- By answering these questions, the project seeks to provide actionable insights that could benefit SpaceX and other space industry players.



Section 1

Methodology

Methodology

- 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
- 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 – SpaceX 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

https://github.com/CaroHL92/Universidad/blob/d3569629604b76ef3547ceb8c131b265a2bafb79/Applied%20Data%20Science%20Capstone/01_Data_Collection_API.ipynb

Data Collection - 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
- Create dataframe from the dictionary
- Export data to csv file

https://github.com/CaroHL92/Universidad/blob/d3569629604b76ef3547ceb8c131b265a2bafb79/Applied%20Data%20Science%20Capstone/02_Web_Scraping.ipynb

Data Wrangling

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 Cont.

- 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

https://github.com/CaroHL92/Universidad/blob/d3569629604b76ef3547ceb8c131b265a2bafb79/Applied%20Data%20Science%20Capstone/03_Data_Wrangling.ipynb

EDA with Data Visualization

Charts

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

EDA with Visualization

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

Queries

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
- Count of landing outcomes between 2010-06-04 and 2017-03-20 (desc)

https://github.com/CaroHL92/Universidad/blob/d3569629604b76ef3547ceb8c131b265a2bafb79/Applied%20Data%20Science%20Capstone/04_EDA_SQL.ipynb

Build an Interactive 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
- Map with Folium

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

https://github.com/CaroHL92/Universidad/blob/d3569629604b76ef3547ceb8c131b265a2bafb79/Applied%20Data%20Science%20Capstone/06_Visual_Analytics_Folium.ipynb

Build a Dashboard with Plotly Dash

Dropdown List with Launch Sites

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

Dashboard with Plotly Dash

Slider of Payload Mass Range

- Allow user to select payload mass range

Pie Chart Showing Successful Launches

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

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

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

https://github.com/CaroHL92/Universidad/blob/d3569629604b76ef3547ceb8c131b265a2bafb79/Applied%20Data%20Science%20Capstone/07_Visual_Analytics_Dashboard.py

Predictive Analysis (Classification)

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(DecisionTreeClassifier()), K-Nearest Neighbor (KNeighborsClassifier())
- 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 and Accuracy

https://github.com/CaroHL92/Universidad/blob/d3569629604b76ef3547ceb8c131b265a2bafb79/Applied%20Data%20Science%20Capstone/08_Machine_learning.ipynb

The background of the slide is an abstract composition. It features a dark blue base color. Overlaid on this are numerous diagonal streaks in shades of red and cyan. A faint, light blue grid pattern is also visible, particularly in the lower half of the image. The overall effect is dynamic and technological.

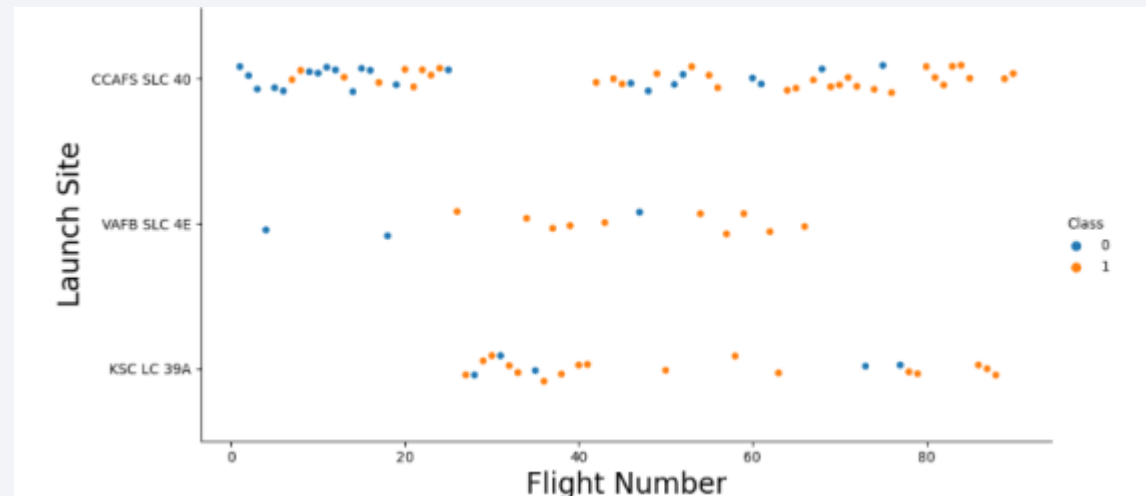
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

Exploratory Data Analysis

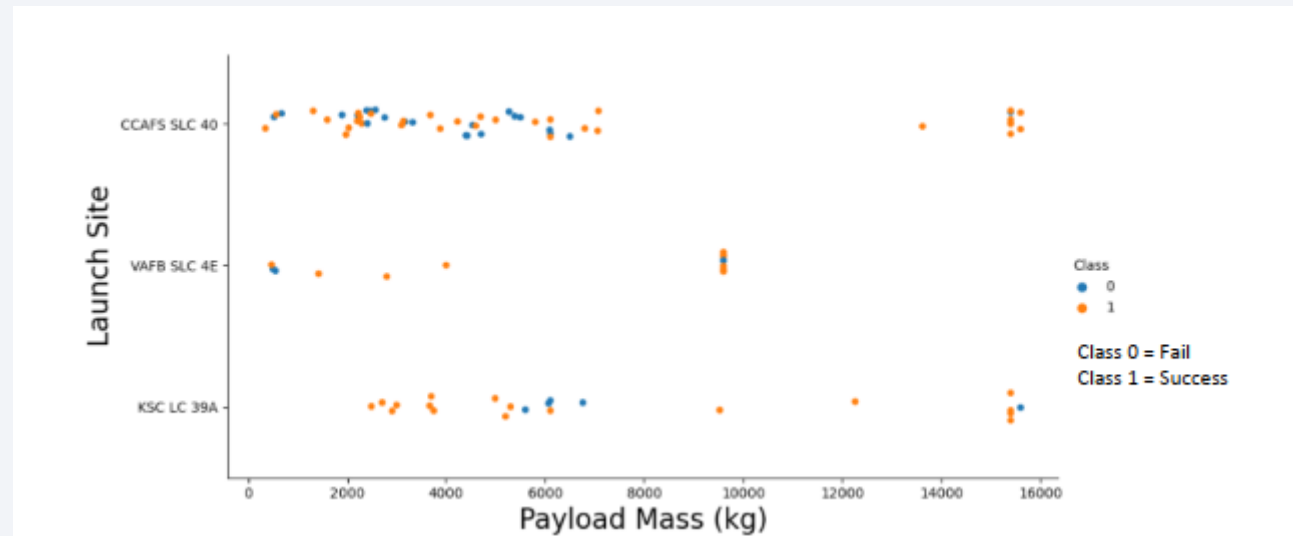
- Earlier flights had a lower success rate (blue = fail)
- Later flights had a higher success rate (orange = success)
- Around half of launches were from CCAFS SLC 40 launch site
- VAFB SLC 4E and KSC LC 39A have higher success rates
- We can infer that new launches have a higher success rate



Payload vs. Launch Site

Exploratory Data Analysis

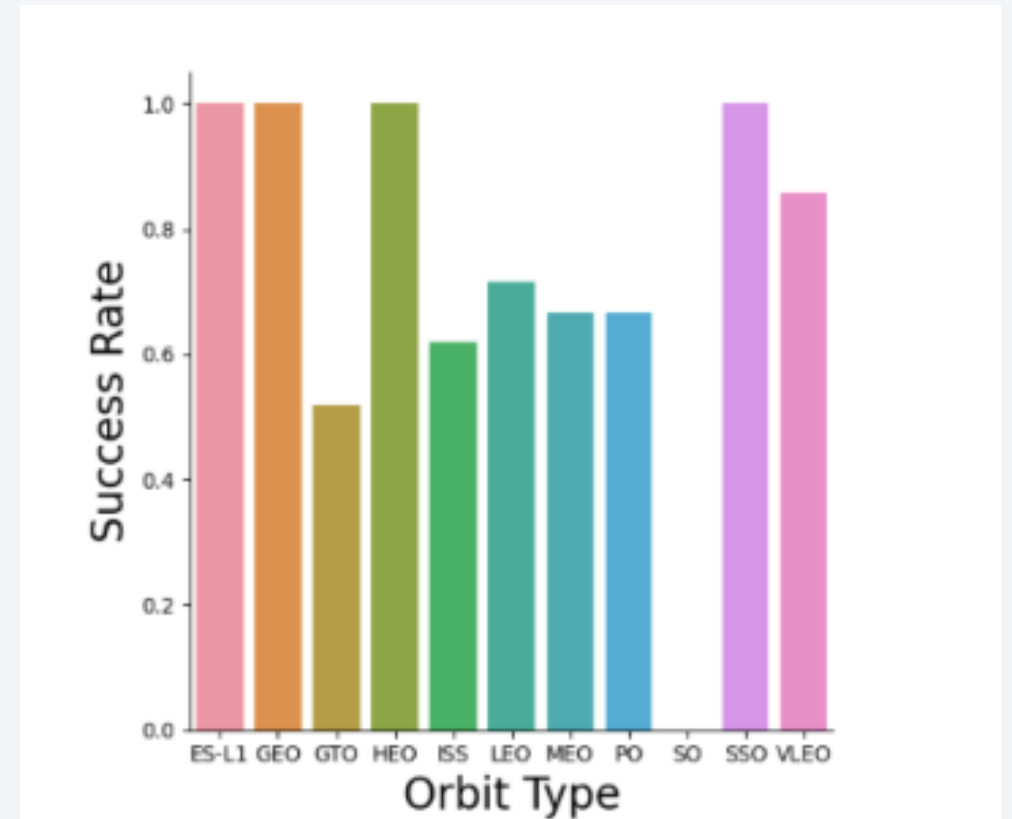
- Typically, the higher the payload mass (kg), the higher the success rate
- Most launches with a payload greater than 7,000 kg were successful
- KSC LC 39A has a 100% success rate for launches less than 5,500 kg
- VAFB SKC 4E has not launched anything greater than ~10,000 kg



Success Rate vs. Orbit Type

Exploratory Data Analysis

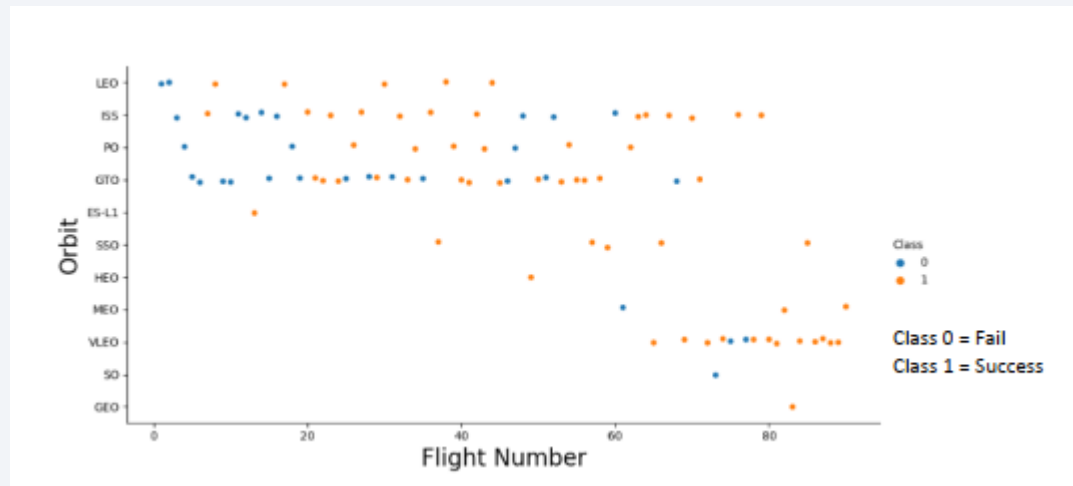
- 100% Success Rate: ES-L1, GEO, HEO and SSO
- 50%-80% Success Rate: GTO, ISS, LEO, MEO, PO
- 0% Success Rate: SO



Flight Number vs. Orbit Type

Exploratory Data Analysis

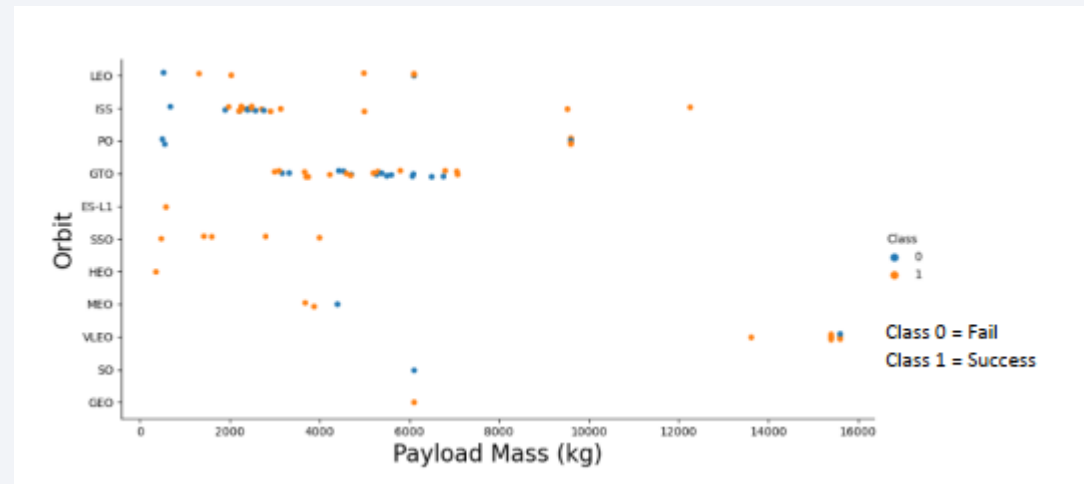
- The success rate typically increases with the number of flights for each orbit
- This relationship is highly apparent for the LEO orbit
- The GTO orbit, however, does not follow this trend



Payload vs. Orbit Type

Exploratory Data Analysis

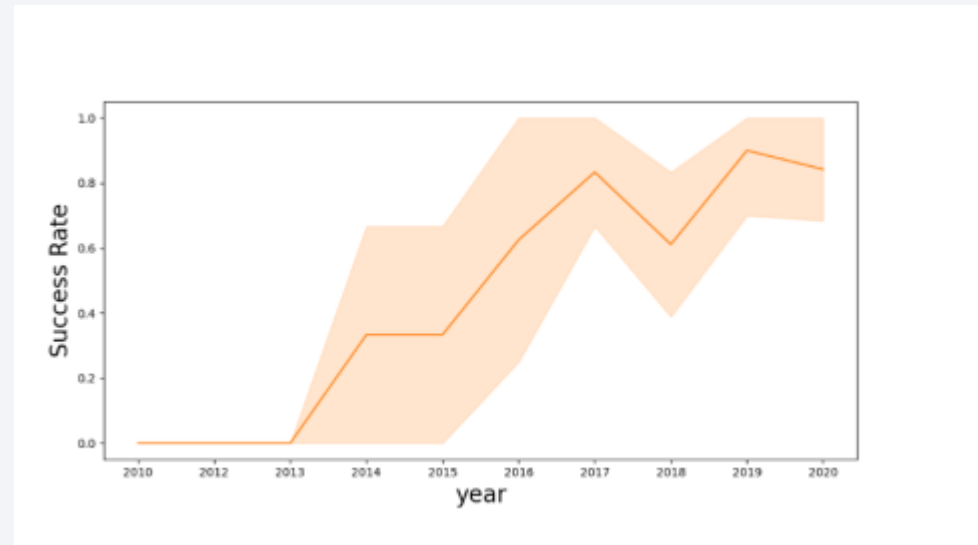
- Heavy payloads are better with LEO, ISS and PO orbits
- The GTO orbit has mixed success with heavier payloads



Launch Success Yearly Trend

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



All Launch Site Names

Launch Site Names

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



Launch Site Names Begin with 'CCA'

Displaying 5 records below

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

* sqlite:///my_data1.db
Done.

| Date | Time (UTC) | Booster_Version | Launch_Site | Payload | PAYLOAD_MASS_KG_ | Orbit | Customer | Mission_Outcome | Landing_Outcome |
|------------|------------|-----------------|-------------|---|------------------|-----------|-----------------|-----------------|---------------------|
| 04-06-2010 | 18:45:00 | F9 v1.0 B0003 | CCAFS LC-40 | Dragon Spacecraft Qualification Unit | 0 | LEO | SpaceX | Success | Failure (parachute) |
| 08-12-2010 | 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) |
| 22-05-2012 | 07:44:00 | F9 v1.0 B0005 | CCAFS LC-40 | Dragon demo flight C2 | 525 | LEO (ISS) | NASA (COTS) | Success | No attempt |
| 08-10-2012 | 00:35:00 | F9 v1.0 B0006 | CCAFS LC-40 | SpaceX CRS-1 | 500 | LEO (ISS) | NASA (CRS) | Success | No attempt |
| 01-03-2013 | 15:10:00 | F9 v1.0 B0007 | CCAFS LC-40 | SpaceX CRS-2 | 677 | LEO (ISS) | NASA (CRS) | Success | No attempt |

Total Payload Mass

Total Payload Mass

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

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

Average Payload Mass

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

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

First Successful Ground Landing Date

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

- 12/22/2015

Successful Drone Ship Landing with Payload between 4000 and 6000

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;
```

Total Number of Successful and Failure 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;
```

Boosters Carried Maximum Payload

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

- 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

2015 Launch Records

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) |

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

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], 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 |

A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The background is a deep blue gradient.

Section 3

Launch Sites Proximities Analysis

Launch Sites

With Markers

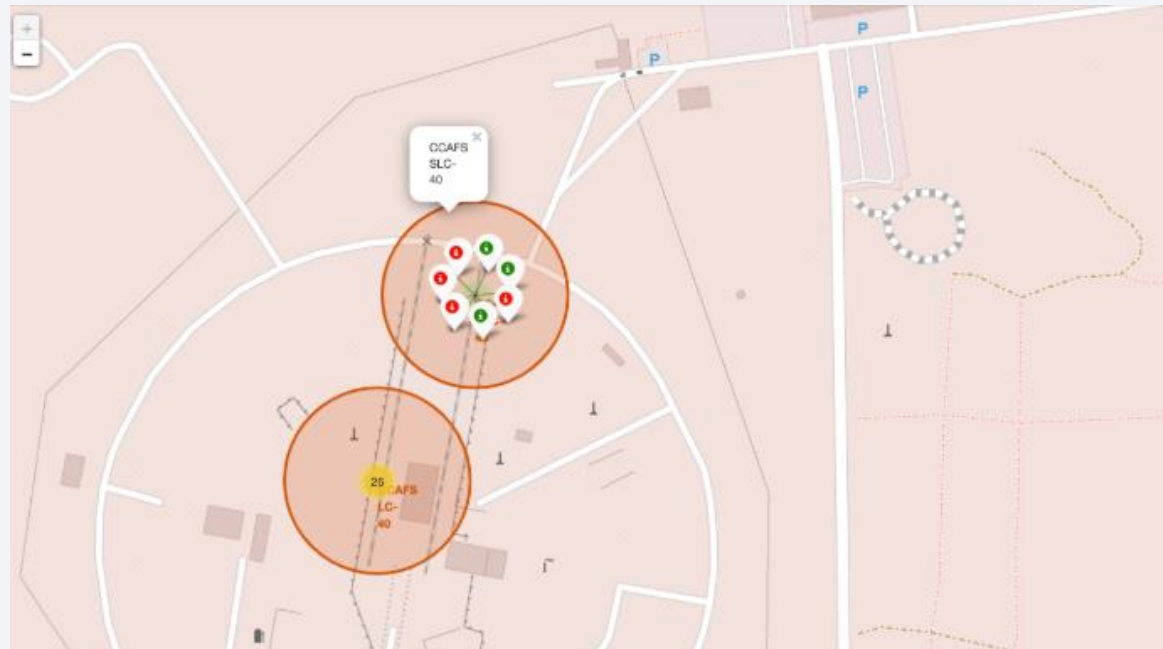
- Near Equator: the closer the launch site to the equator, the easier it is to launch to equatorial orbit, and the more help you get from Earth's rotation for a prograde orbit. Rockets launched from sites near the equator get an additional natural boost - due to the rotational speed of earth - that helps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



Launch outcomes

At Each Launch Site

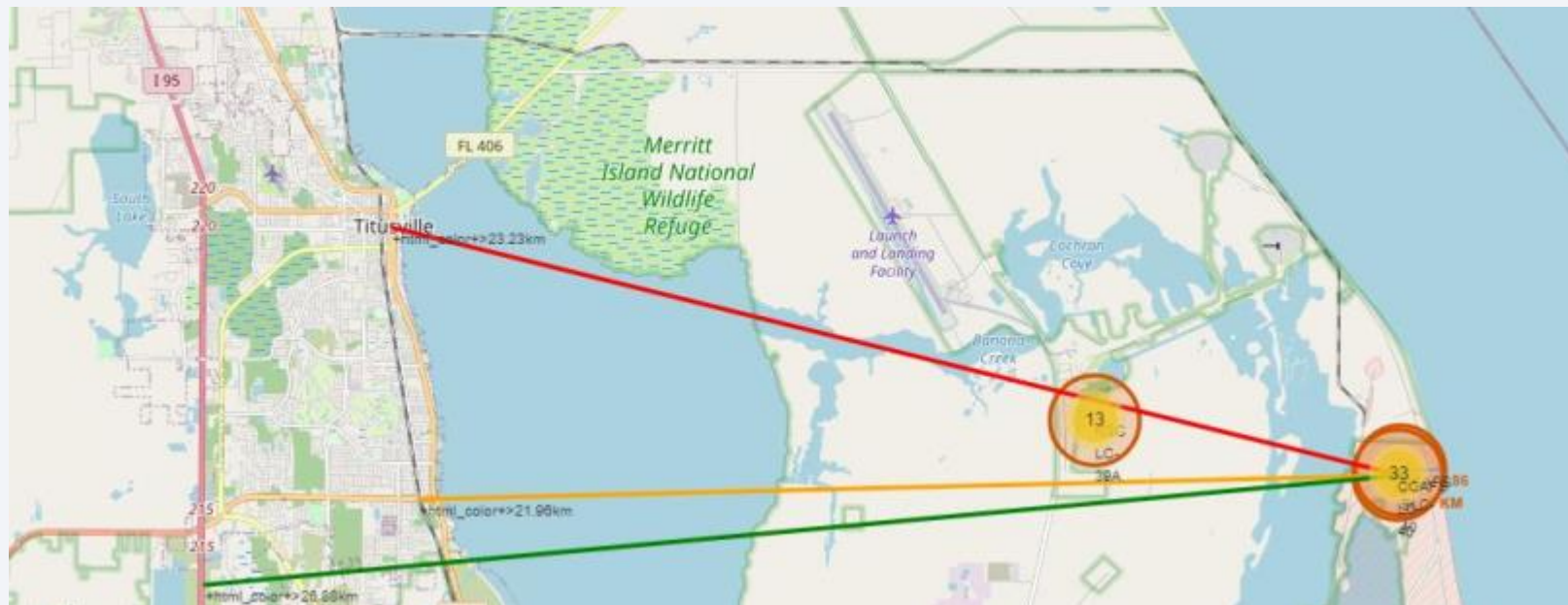
- Outcomes:
- Green markers for successful launches
- Red markers for unsuccessful launches
- Launch site CCAFS SLC-40 has a 3/7 success rate (42.9%)



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



The background of the slide is a close-up, artistic photograph of a printed circuit board (PCB). The board is dark, and the intricate circuit traces are highlighted in a vibrant, glowing red. Numerous small, circular components, likely solder joints or micro-components, are visible along the traces, some of which also appear to be glowing. The overall effect is a high-tech, digital aesthetic.

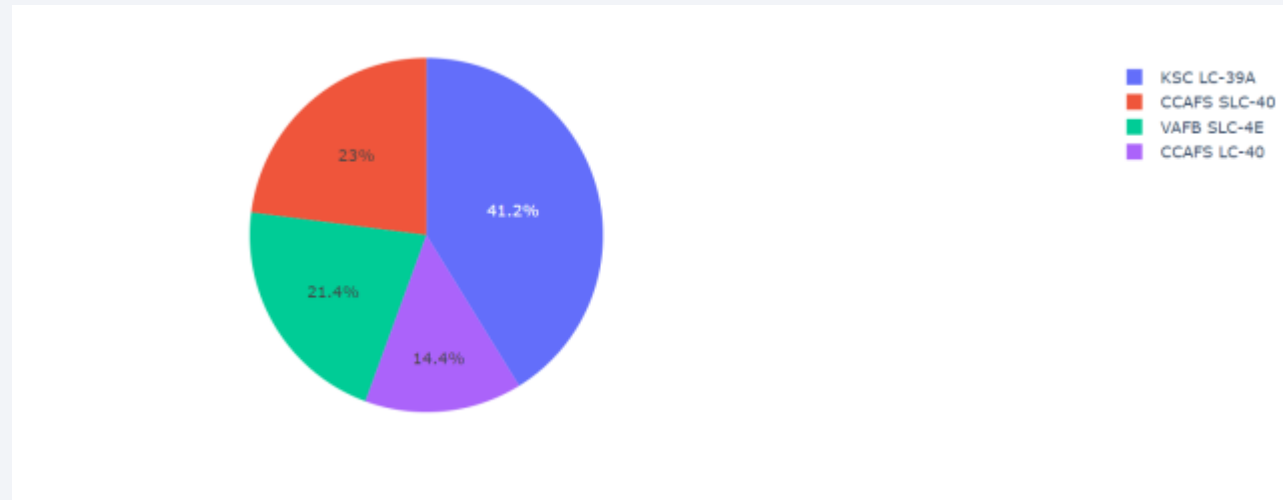
Section 4

Build a Dashboard with Plotly Dash

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

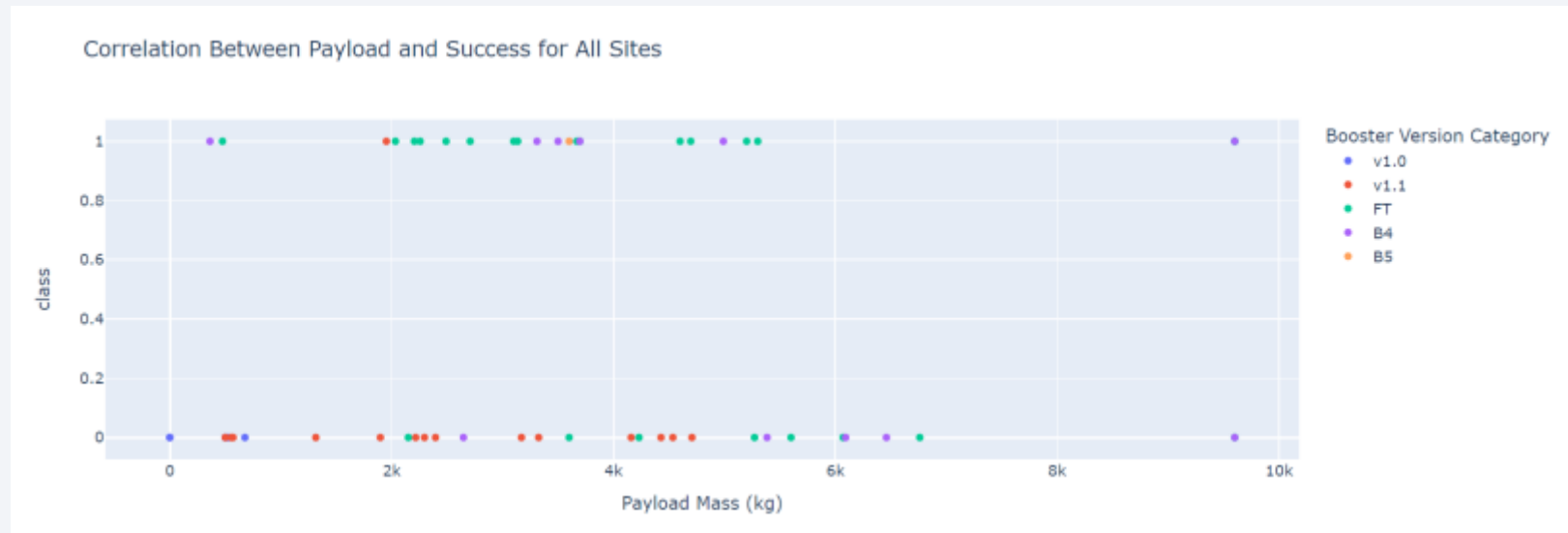
Total Success Launches for Site KSC LC-39A



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



Section 5

Predictive Analysis (Classification)

Classification Accuracy

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

| | 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 |

```
: 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'}
```

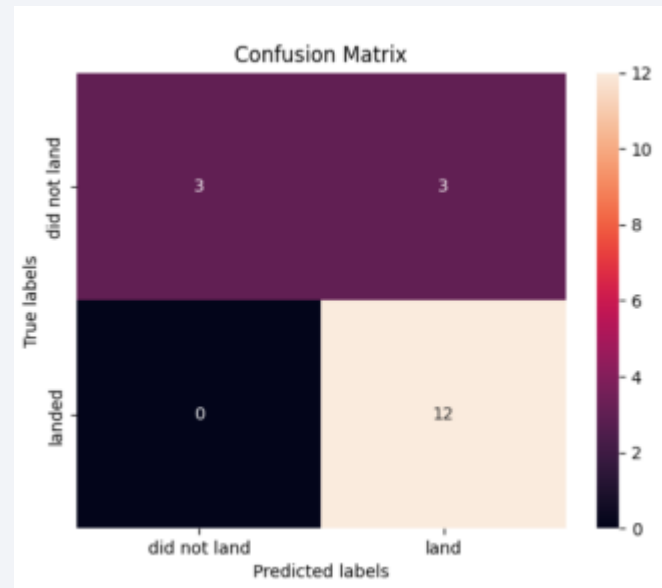

Confusion Matrix

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



Conclusions

- 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

Thank you!

