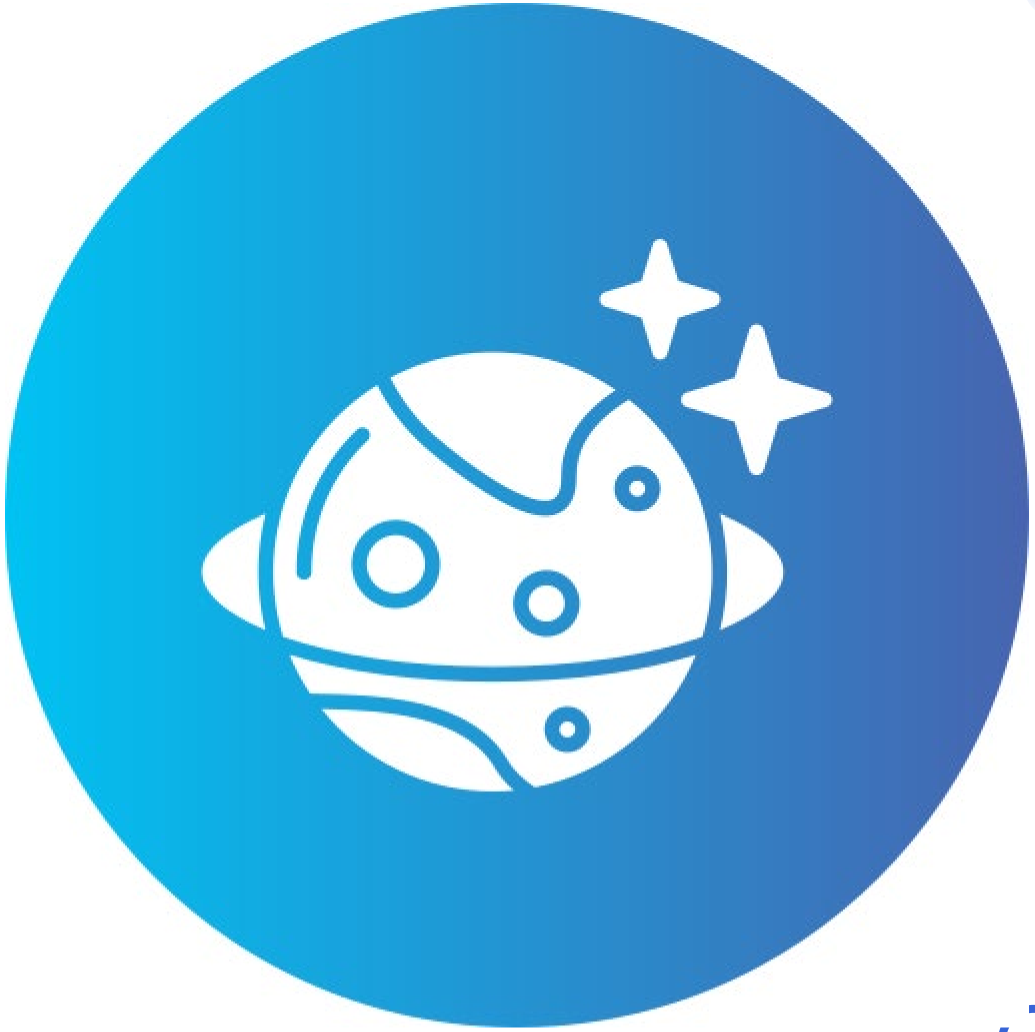
SpaceX
in the sights of Data Science

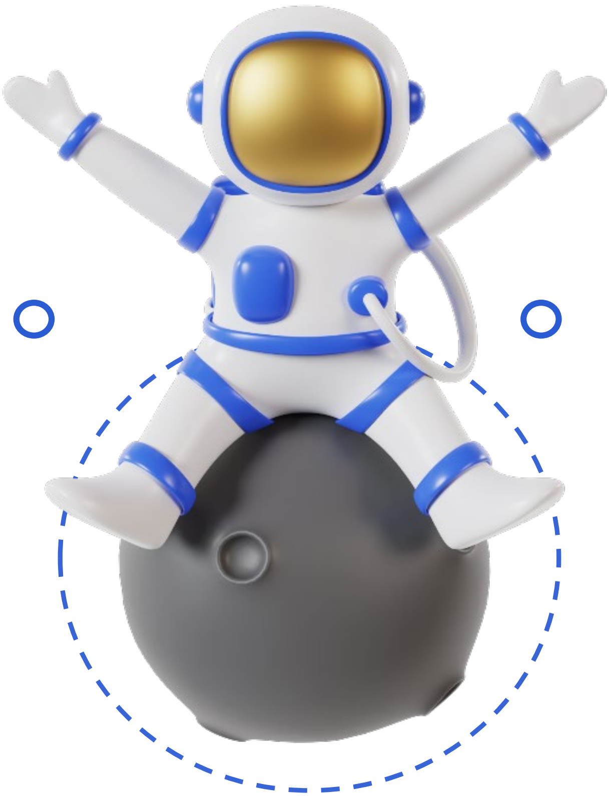
Yevheniia Volovatova, student Coursera

31/12/2024

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)

INTRODUCTION

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.

Problems that needed solving:

- What influences if the rocket will land successfully?
- The effect each relationship with certain rocket variables will impact in determining the success rate of a successful landing.
- What conditions does SpaceX have to achieve to get the best results and ensure the best rocket success landing rate.

DETAILED CONTENTS

METHODOLOGY

Required data collection and wrangling 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
- **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

The results of laboratory work are available at the link:
https://github.com/evgeniaekologia/my-projekt-SpaceX-Falcon-9

DETAILED CONTENTS

METHODOLOGY

EDA and interactive visual analytics methodology

Scatter Graphs being drawn:

- Flight Number VS. Payload Mass
- Flight Number VS. Launch Site
- Payload VS. Launch Site
- Orbit VS. Flight Number
- Payload VS. Orbit Type
- Orbit VS. Payload Mass

Scatter plots show how much one variable is affected by another. The relationship between two variables is called their correlation .Scatter plots usually consist of a large body of data.

Bar Graph being drawn:

- Mean VS. Orbit

A bar diagram makes it easy to compare sets of data between different groups at a glance. The graph represents categories on one axis and a discrete value in the other. The goal is to show the relationship between the two axes. Bar charts can also show big changes in data over time.

Line Graph being drawn:

- Success RateVS. Year

Line graphs are useful in that they show data variables and trends very clearly and can help to make predictions about the results of data not yet recorded

DETAILED CONTENTS

METHODOLOGY

Required predictive analysis methodology

Predictive Analysis (Classification):

- Loaded the data using numpy and pandas, transformed the data, split our data into training and testing.
- Built different machine learning models and tune different hyperparameters using GridSearchCV.
- Used accuracy as the metric for our model, improved the model using feature engineering and algorithm tuning.
- Found the best performing classification model.

Data Wrangling:

In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad. True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship.

We mainly convert those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.

DETAILED CONTENTS

RESULTS AND DISCUSSION SECTION

Results: Results Summary

ExploratoryData Analysis

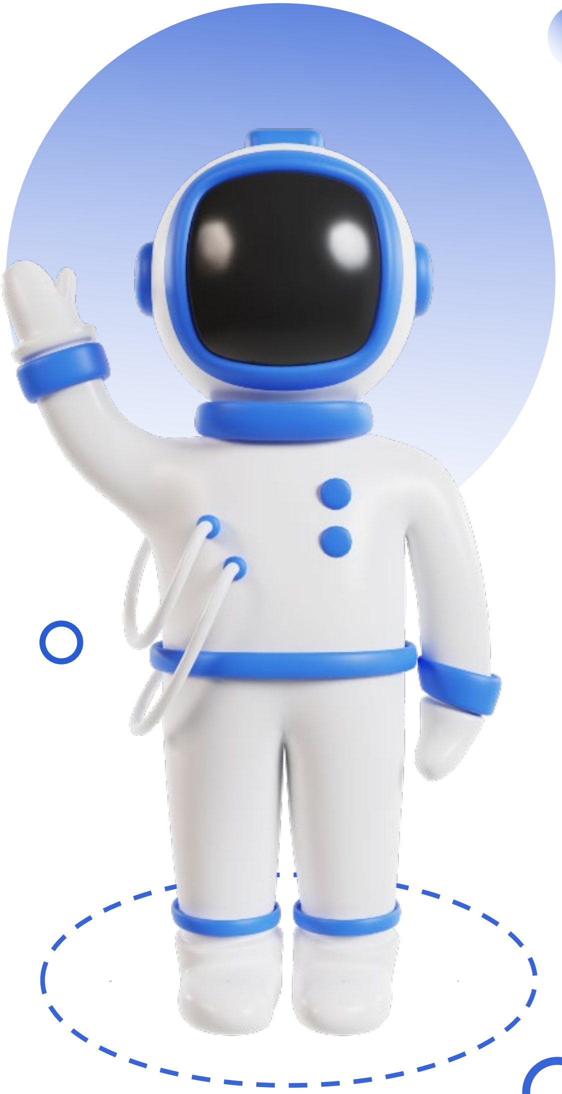
- Launch success has improved over time
- KSC LC-39A has the highest success rate among landing sites
- Orbits ES-L1, GEO, HEO and SSO have a 100% success rate

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

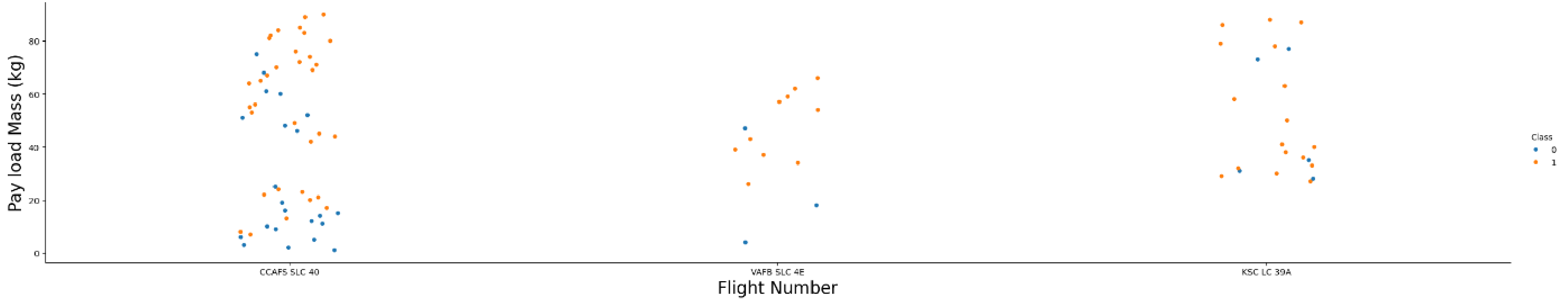


DETAILED CONTENTS

RESULTS AND DISCUSSION SECTION

Results: Flight Number vs. Launch Site

- **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



DETAILED CONTENTS

RESULTS AND DISCUSSION SECTION

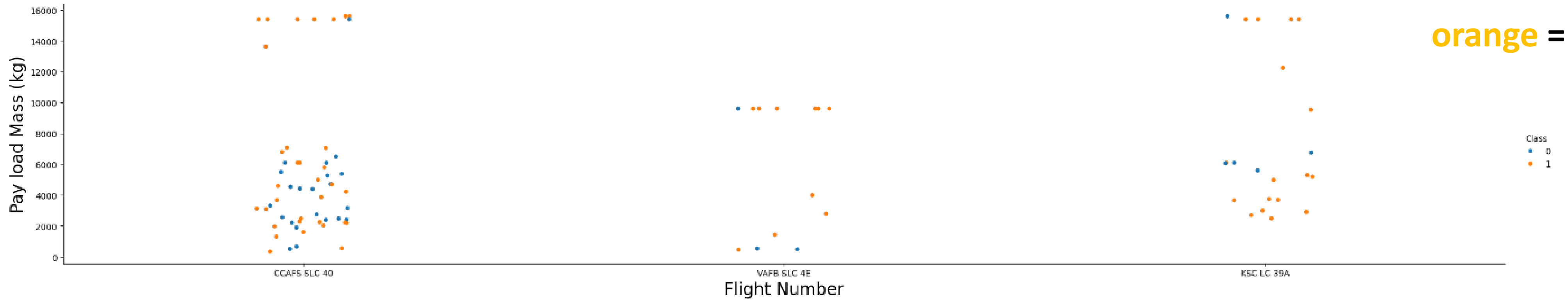
Results: Payload vs. Launch Site

Exploratory Data Analysis

- Typically, the **higher** the **payload mass** (kg), the **higher** the **success rate**
- Most launces 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

blue = **fail**

orange = **success**



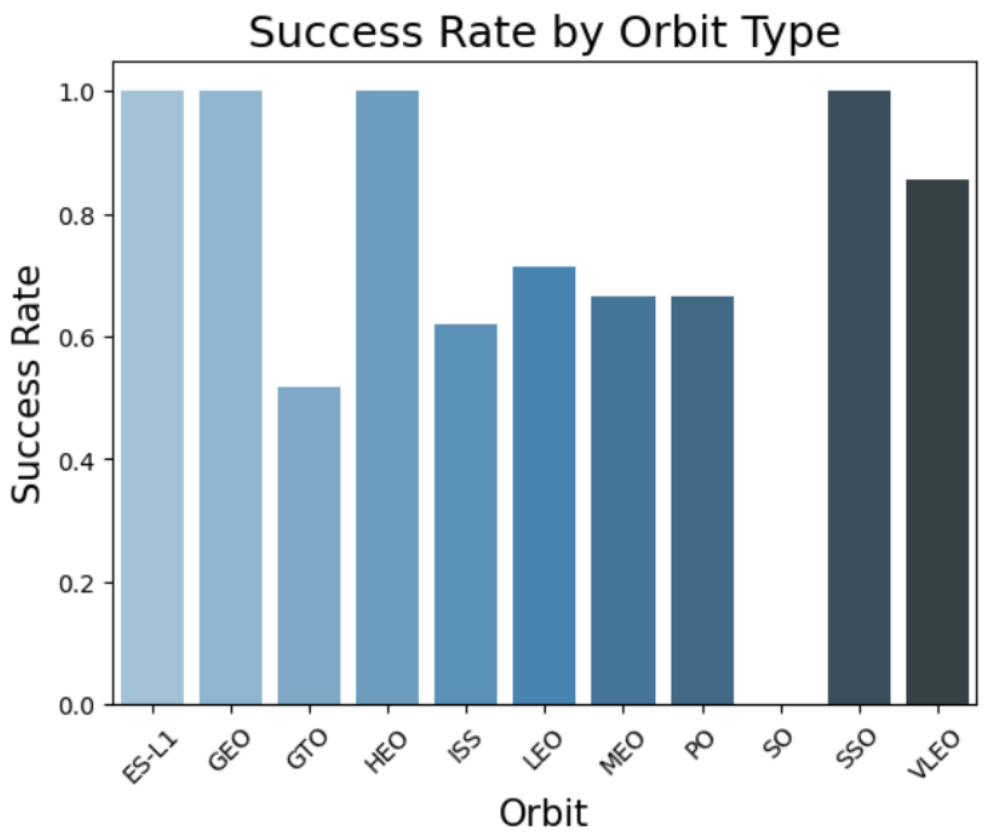
DETAILED CONTENTS

RESULTS AND DISCUSSION SECTION

Results: 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 and VLEO
- **0% Success Rate**: SO



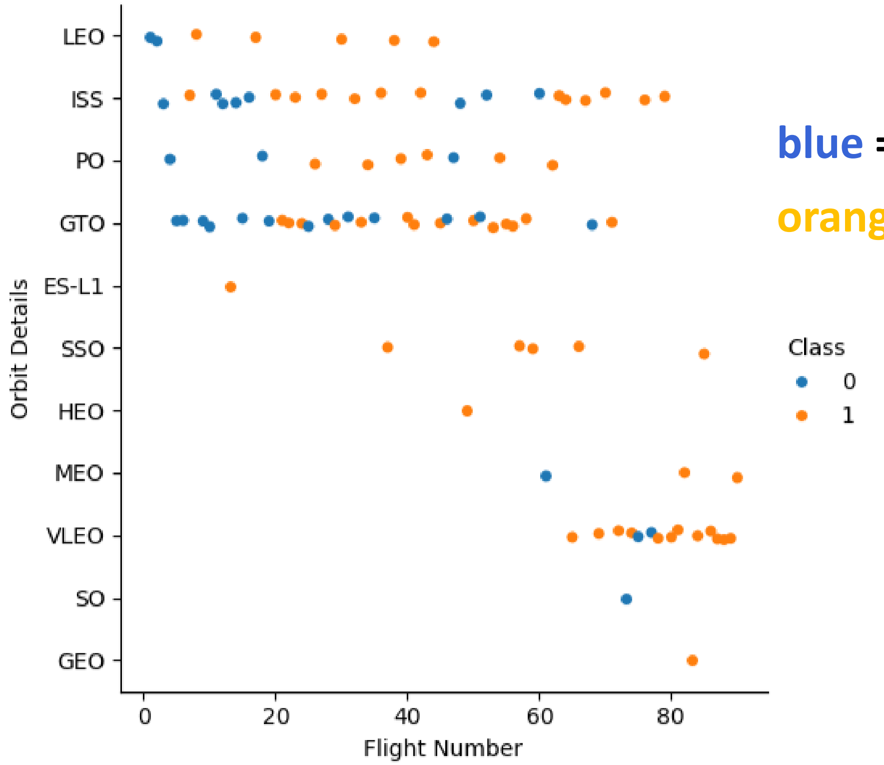
DETAILED CONTENTS

RESULTS AND DISCUSSION SECTION

Results: 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



blue = fail

orange = success

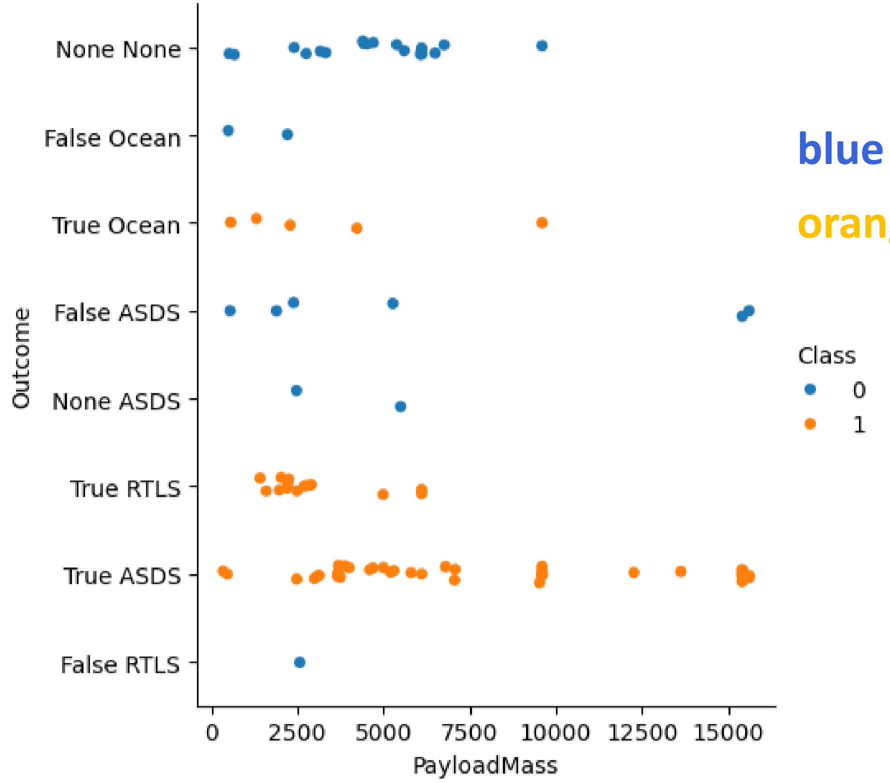
DETAILED CONTENTS

RESULTS AND DISCUSSION SECTION

Results: 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



blue = fail

orange = success

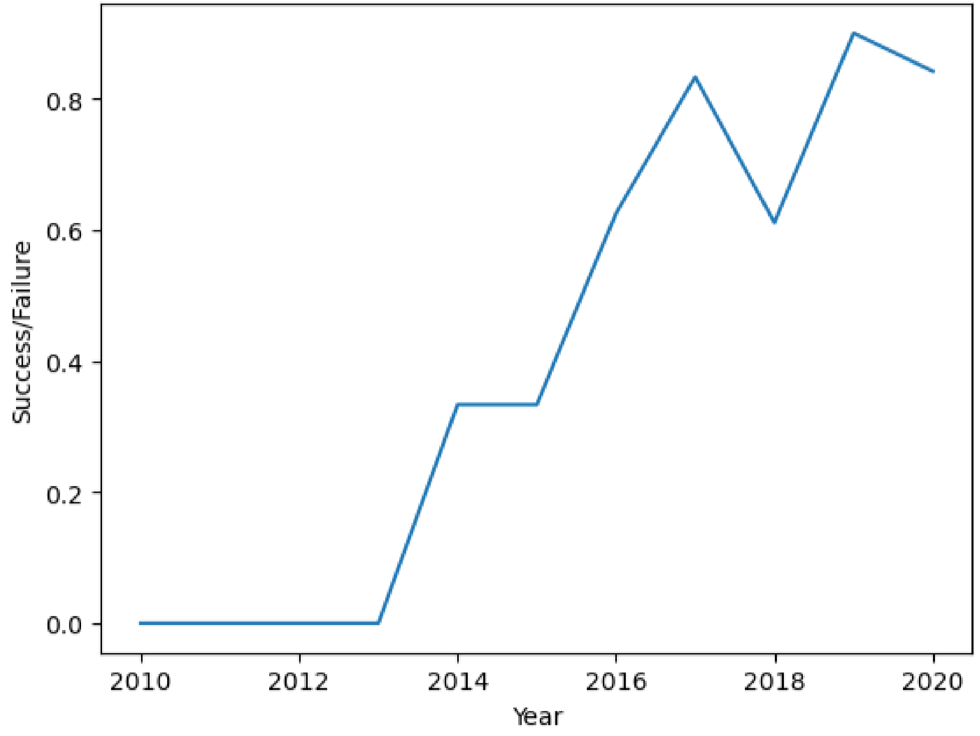
DETAILED CONTENTS

RESULTS AND DISCUSSION SECTION

Results: 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



DETAILED CONTENTS

RESULTS AND DISCUSSION SECTION

Results: Launch Site Information

Launch Site Names

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

Landing Outcome Cont

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

* sqlite:///my_data1.db
Done.

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

Records with Launch Site Starting 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
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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

DETAILED CONTENTS

RESULTS AND DISCUSSION SECTION

Results: Payload Mass

Total Payload Mass

- **45596 kg** (total) carried by boosters launched by NASA (CRS)

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

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass

- **2534 kg** (average) carried by booster version F9 v1.1

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

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

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

DETAILED CONTENTS

RESULTS AND DISCUSSION SECTION

Results: Landing & Mission Info

1stSuccessful Landing in Ground Pad

- 22/12/2015

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

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

MIN(DATE)
2015-12-22

Booster Drone Ship Landing

- Booster mass greater than 4,000 but less than 6,000

```
%sql select BOOSTER_VERSION from SPACEXTBL where Landing_Outcome='Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000;
```

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

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failed Mission Outcomes

```
%%sql
SELECT COUNT(Landing_Outcome) AS SUCCESSFUL_MISSIONS
FROM SPACEXTBL
WHERE Landing_Outcome LIKE 'Success%';
```

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

SUCCESSFUL_MISSIONS
61

```
%%sql
SELECT COUNT(Landing_Outcome) AS FAILURE_MISSIONS
FROM SPACEXTBL
WHERE Landing_Outcome LIKE 'Failure%';
```

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

FAILURE_MISSIONS
10

DETAILED CONTENTS

RESULTS AND DISCUSSION SECTION

Results: Boosters

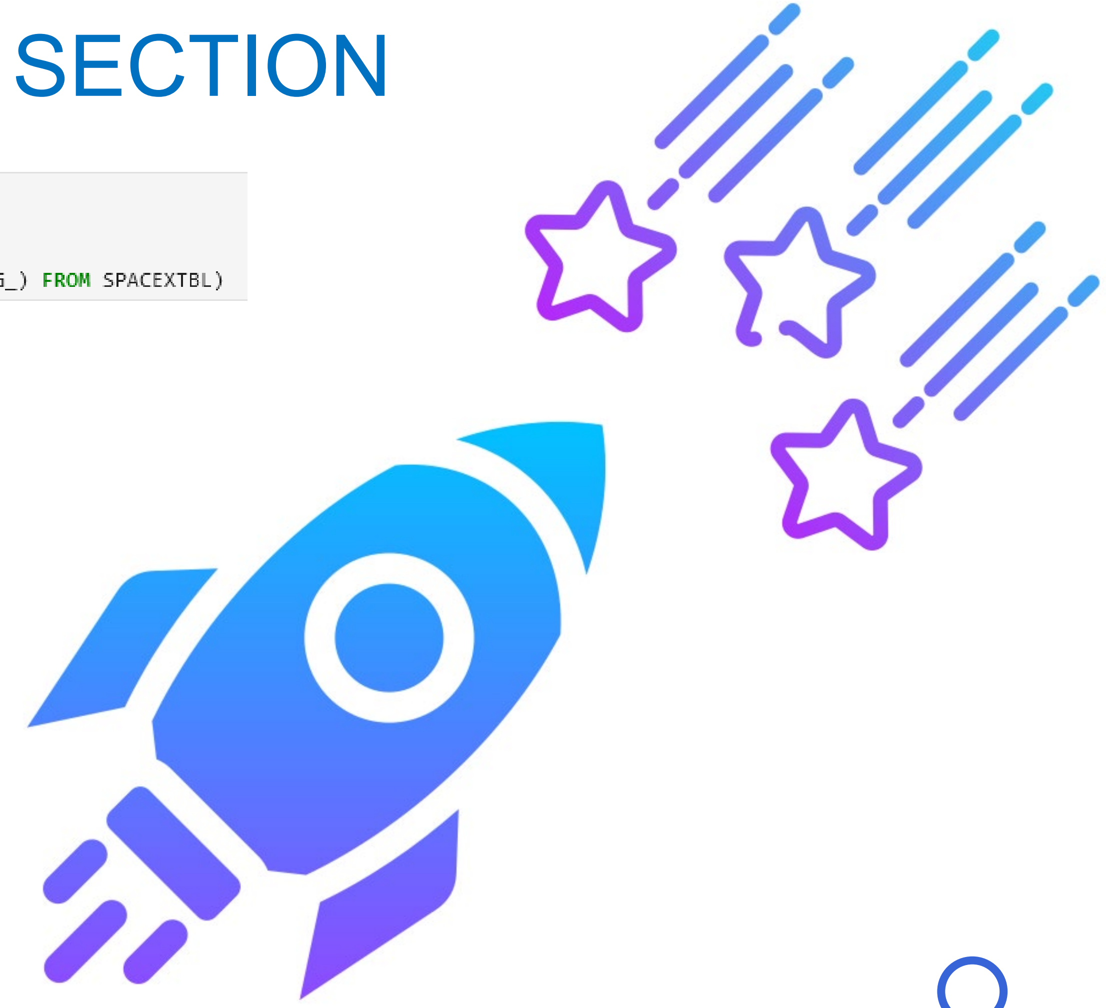
Carrying Max Payload

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 DISTINCT(BOOSTER_VERSION), PAYLOAD_MASS__KG_
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600



DETAILED CONTENTS

RESULTS AND DISCUSSION SECTION

Results: Failed Landings on Drone Ship

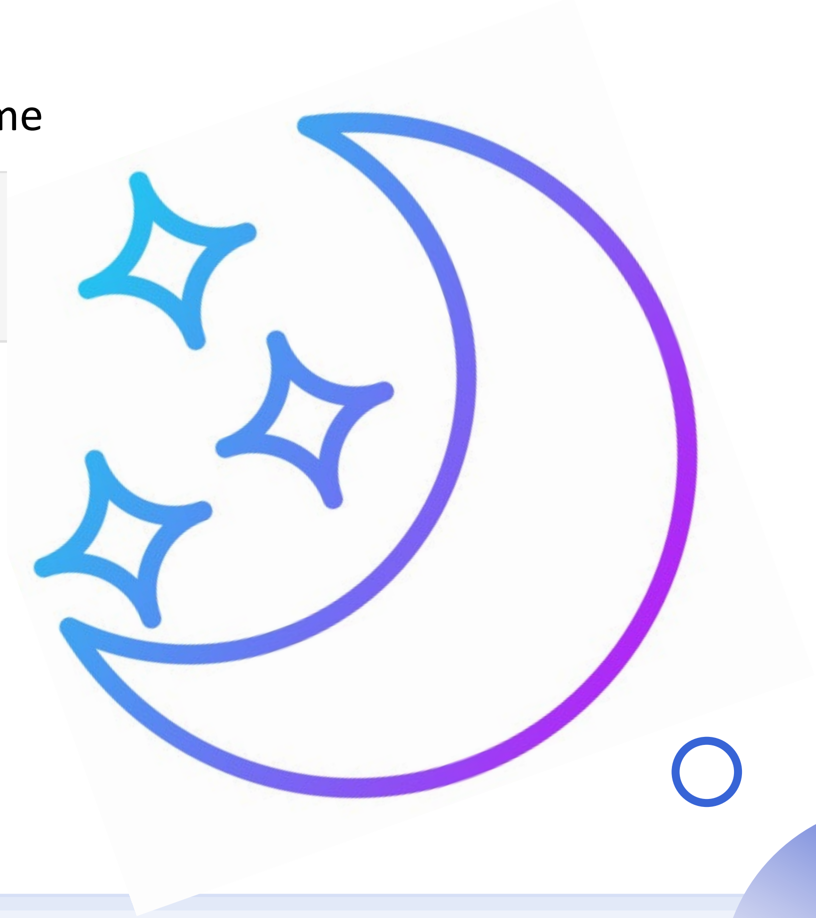
In 2015

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

```
%%sql
SELECT Landing_Outcome, Booster_Version, Launch_Site, strftime('%Y', DATE) AS DATE_YEAR
FROM SPACEXTBL
WHERE Landing_Outcome = 'Failure (drone ship)' AND strftime('%Y', DATE) = '2015';
```

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

Landing_Outcome	Booster_Version	Launch_Site	DATE_YEAR
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015



DETAILED CONTENTS

RESULTS AND DISCUSSION SECTION

Results: Count of Successful Landings

Ranked Descending

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

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

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

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

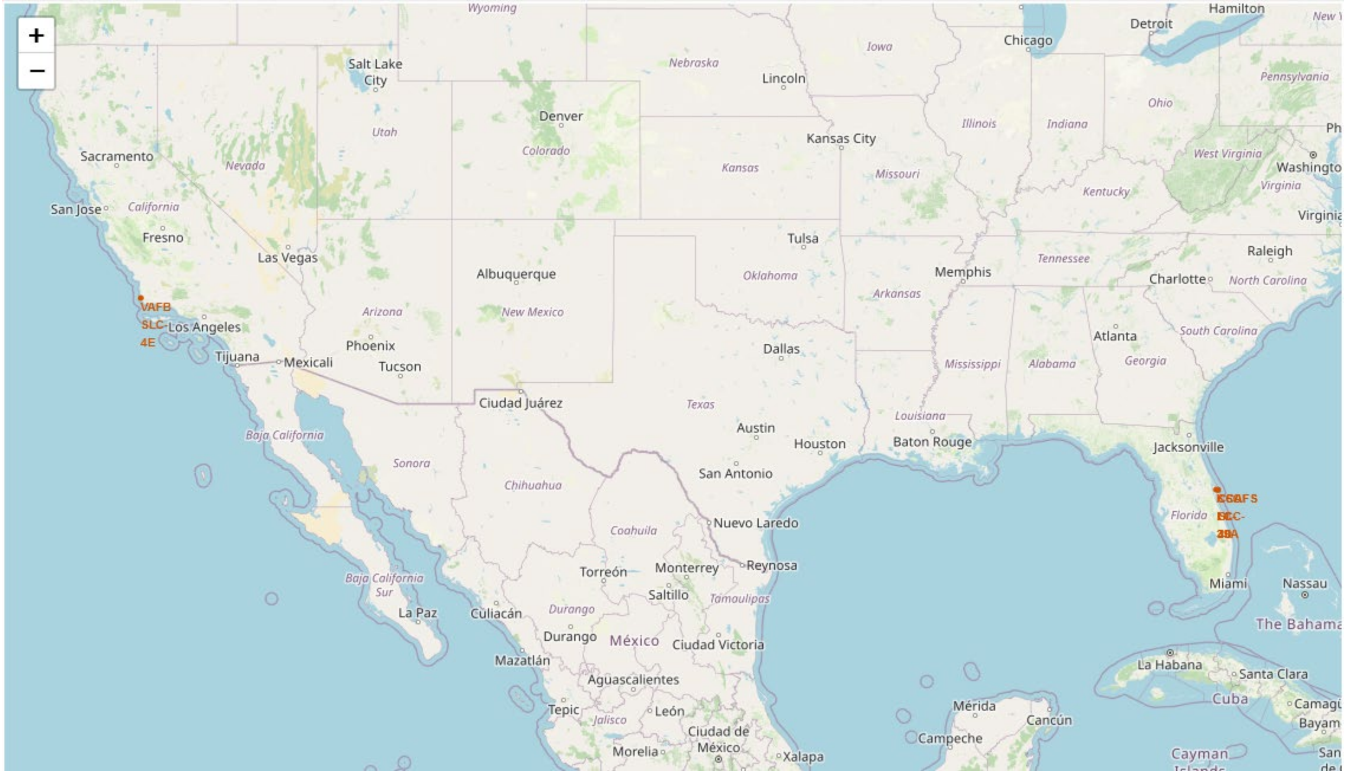
DETAILED CONTENTS

RESULTS AND DISCUSSION SECTION

Launch Site 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.



DETAILED CONTENTS

RESULTS AND DISCUSSION SECTION

Launch Site Analysis: Launch Outcomes

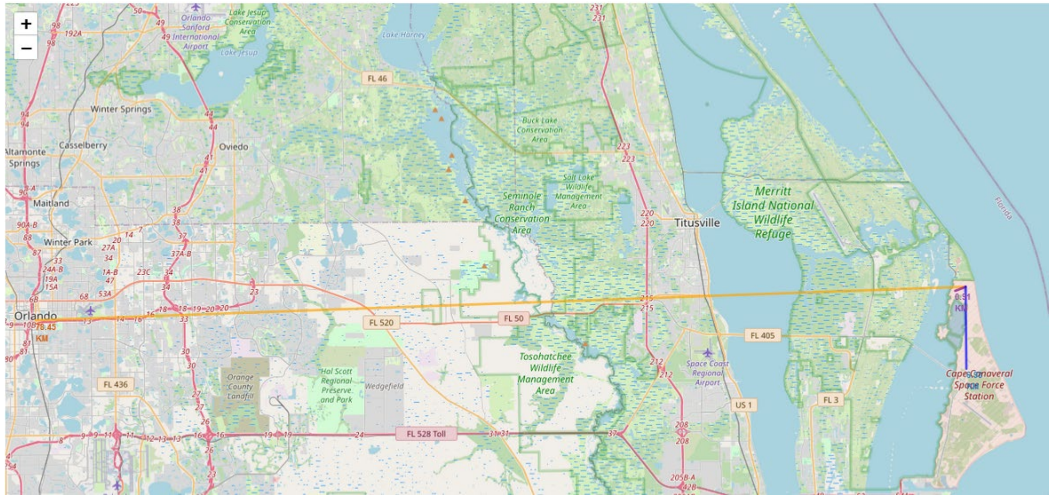
At Each Launch Site

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

DETAILED CONTENTS

RESULTS AND DISCUSSION SECTION

Launch Site Analysis: Distance to Proximities



- **21.96 km** from nearest railway (purple)
- **23.23 km** from nearest city (orange)
- **26.88 km** from nearest highway (blue)

- **Safety / Security:** needs to be an exclusion zone around the launch site to keep unauthorized people away and keep people safe; ensure that spent stages dropped along the launch path or failed launches don't fall on people or property.
- **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.

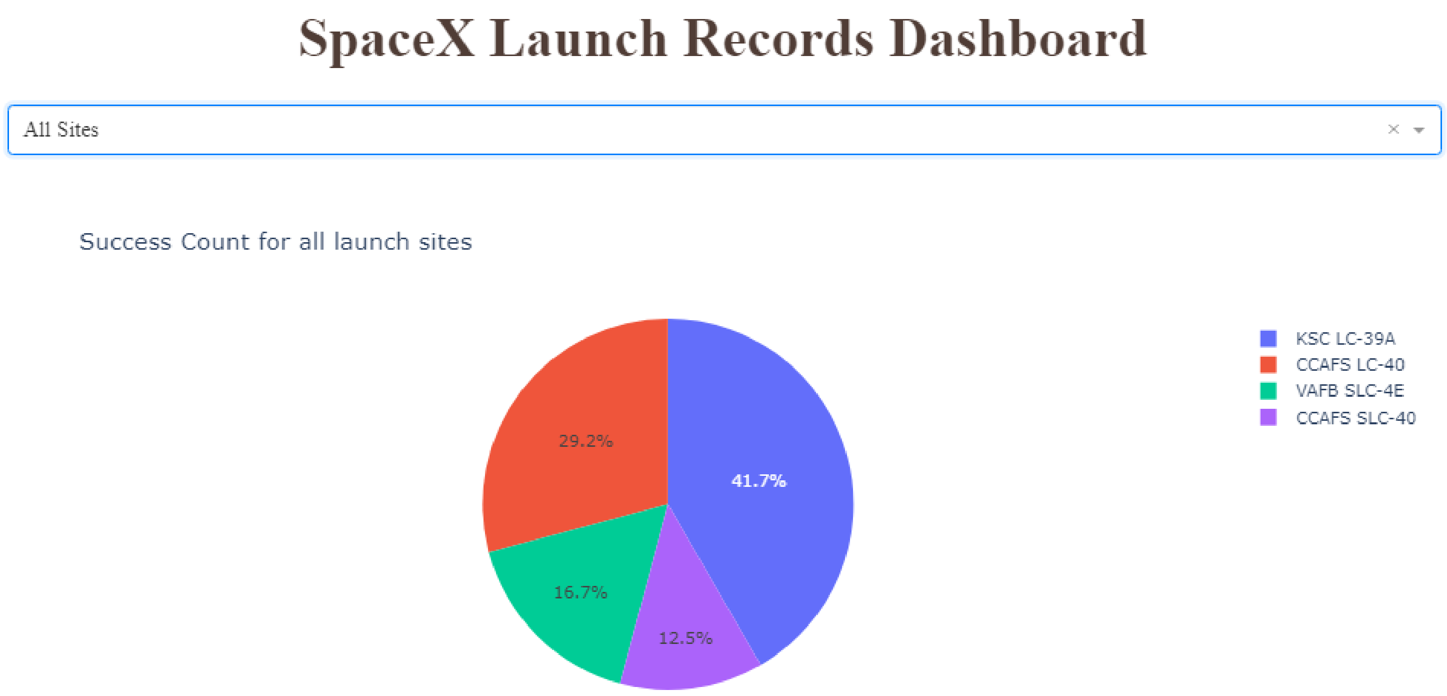
DETAILED CONTENTS

RESULTS AND DISCUSSION SECTION

Dashboard with Plotly: Launch Success by Site

Success as Percent of Total

- **KSC LC-39A** has the **most successful launches** amongst launch sites (**41.7%**)



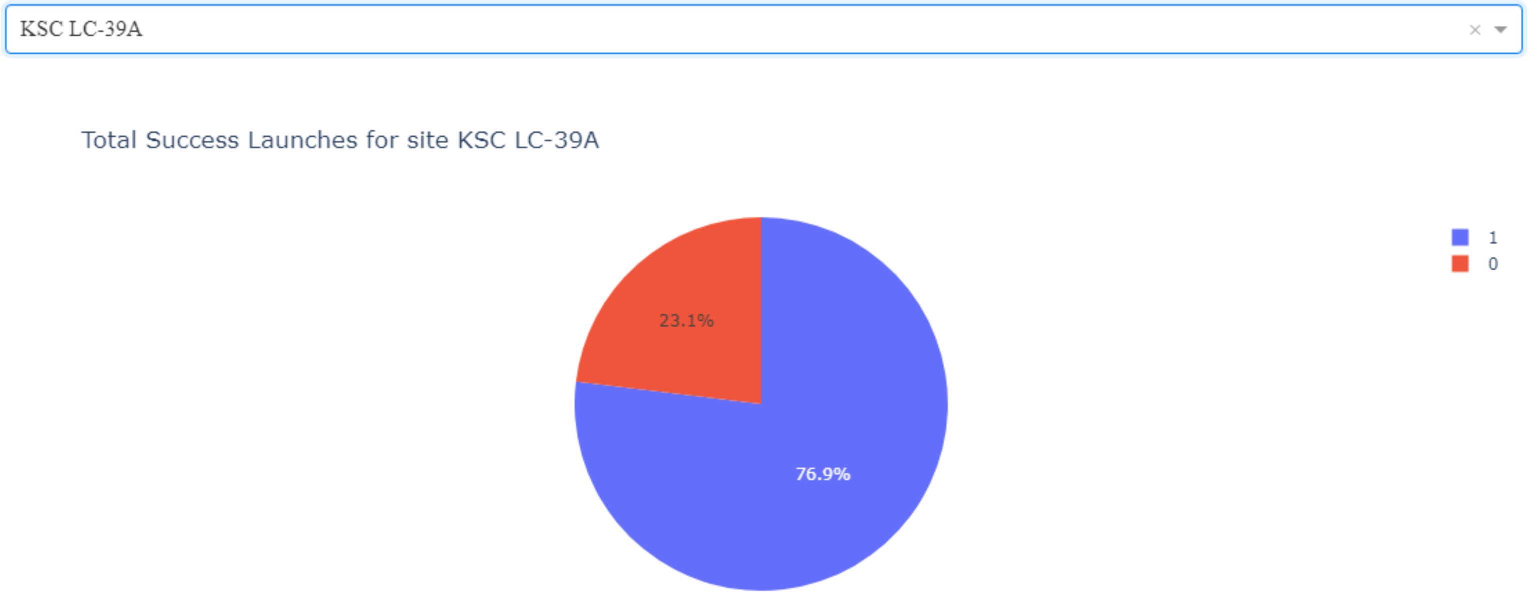
DETAILED CONTENTS

RESULTS AND DISCUSSION SECTION

Dashboard with Plotly: 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



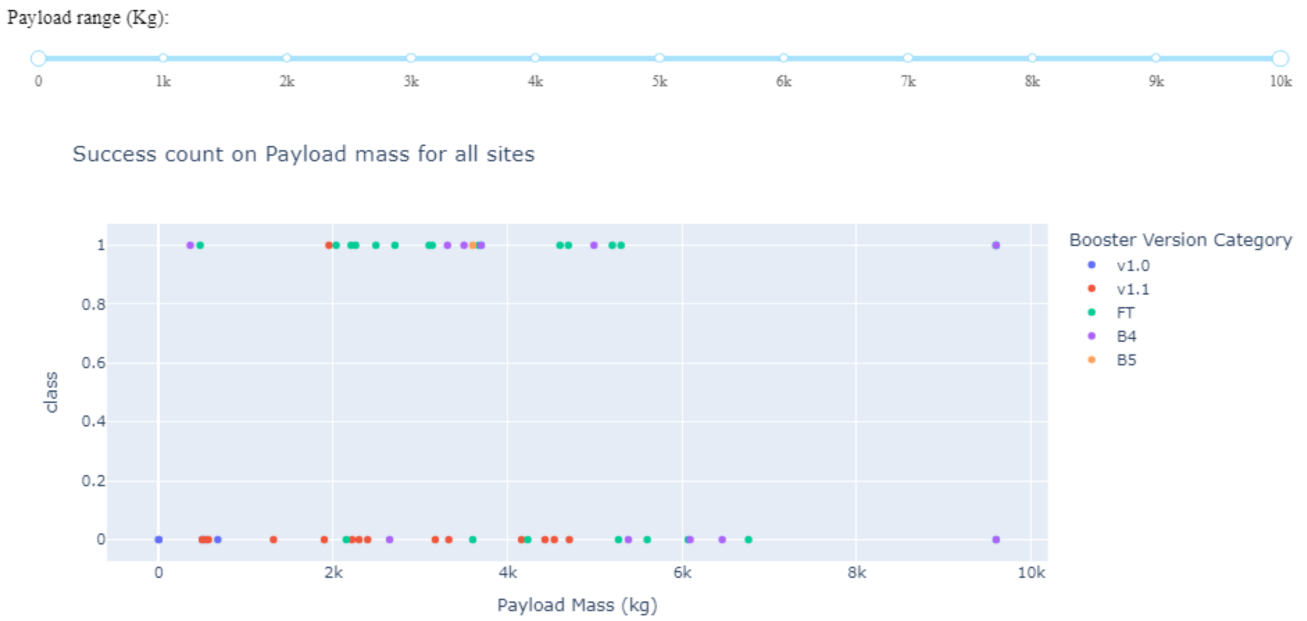
DETAILED CONTENTS

RESULTS AND DISCUSSION SECTION

Dashboard with Plotly: 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



DETAILED CONTENTS

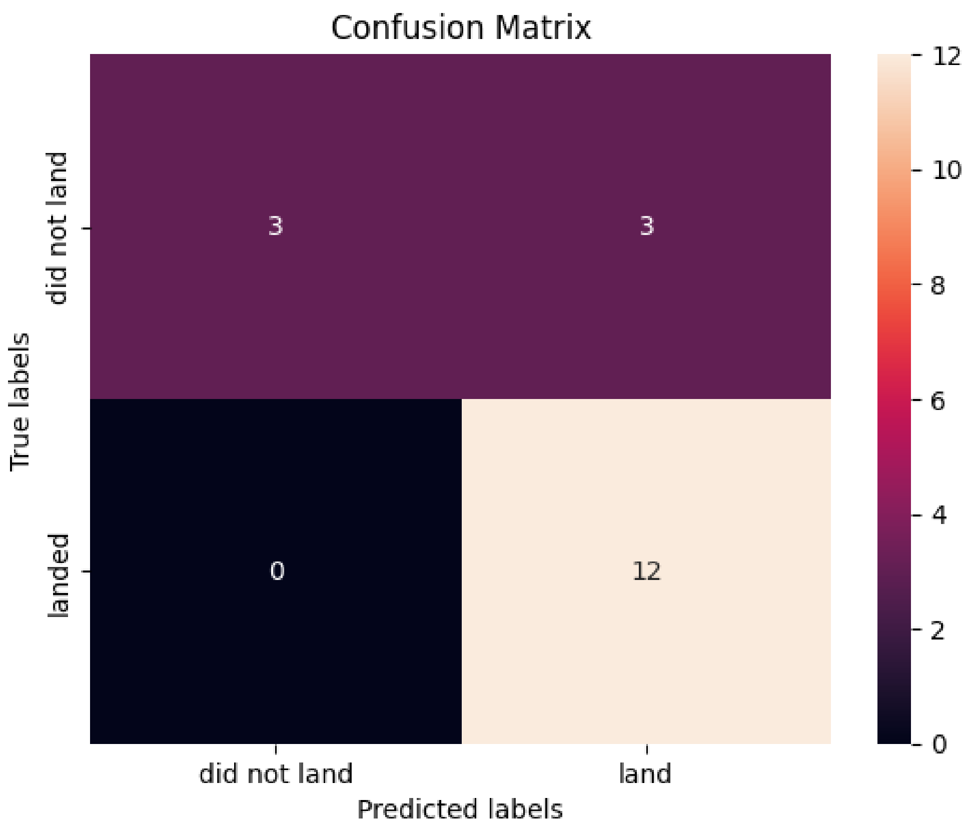
RESULTS AND DISCUSSION SECTION

Predictive Analytics: 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



DETAILED CONTENTS

RESULTS AND DISCUSSION SECTION

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 dataset**.

	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

```
accuracy = [svm_cv_score, logreg_score, knn_cv_score, tree_cv_score]
accuracy = [i * 100 for i in accuracy]

method = ['Support Vector Machine', 'Logistic Regression', 'K Nearest Neighbour', 'Decision Tree']
models = {'ML Method':method, 'Accuracy Score (%)':accuracy}

ML_df = pd.DataFrame(models)
ML_df
```

```
from sklearn.metrics import jaccard_score, f1_score

# Examining the scores from Test sets
jaccard_scores = [
                  jaccard_score(Y_test, logreg_yhat, average='binary'),
                  jaccard_score(Y_test, svm_yhat, average='binary'),
                  jaccard_score(Y_test, tree_yhat, average='binary'),
                  jaccard_score(Y_test, knn_yhat, average='binary'),
                 ]

f1_scores = [
             f1_score(Y_test, logreg_yhat, average='binary'),
             f1_score(Y_test, svm_yhat, average='binary'),
             f1_score(Y_test, tree_yhat, average='binary'),
             f1_score(Y_test, knn_yhat, average='binary'),
            ]

accuracy = [logreg_score, svm_cv_score, tree_cv_score, knn_cv_score]

scores_test = pd.DataFrame(np.array([jaccard_scores, f1_scores, accuracy]), index=['Jaccard_Score', 'F1_Score', 'Accuracy'] ,
scores_test
```

DETAILED CONTENTS

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 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

ACKNOWLEDGMENTS

Thanks to everyone who supported me in this difficult study. Thanks to the developers of the training program and to those who read my work.

To all colleagues, good luck in your studies and happy new year!