
# Winning Space Race with Data Science

László György Szilas
28. 02. 2022

# Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

# Executive Summary

- Methodologies
  - Data collected from SpaceX API and from Wikipedia
  - Data transformed with standard techniques
  - EDA performed using
    - SQL
    - basic data visualization (matplotlib)
    - advanced visualization (plotly and dash)
  - Predictive analysis made with basic classification models
- Results

# Introduction

## Background

- Companies are making space travel affordable for everyone
    - SpaceX is probably the most successful company on the market
    - the rocket launches are relatively affordable because SpaceX can reuse the first stage
    - for SpaceX Falcon 9 rocket launches are the most commonly used

## Goal of the project

- We want to determine if the first stage will land, so that we could determine the cost of a launch

Section
1

# Methodology

# Methodology

## Executive Summary

- Data collection methodology:
  - Data was collected through the publicly available SpaceX API's rocket launch database
- Perform data wrangling
  - Data was processed using the pandas python library
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - 4 basic classification model evaluated and compared

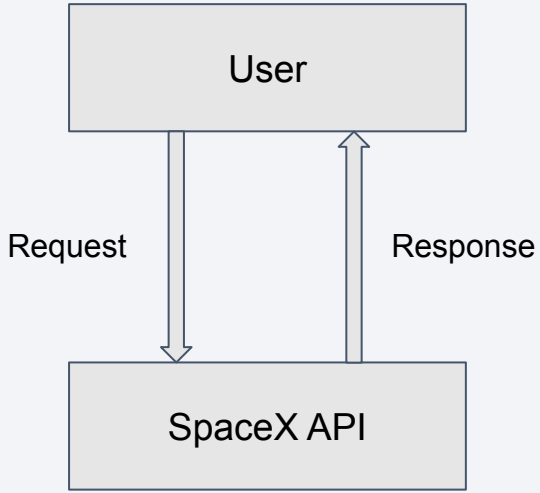
# Data Collection

- Data sets were collected
  - from the SpaceX API by API calls and
  - from Wikipedia by web scraping

# Data Collection – SpaceX API

- Flowchart for the data collection from the SpaceX API can be seen on the right hand side

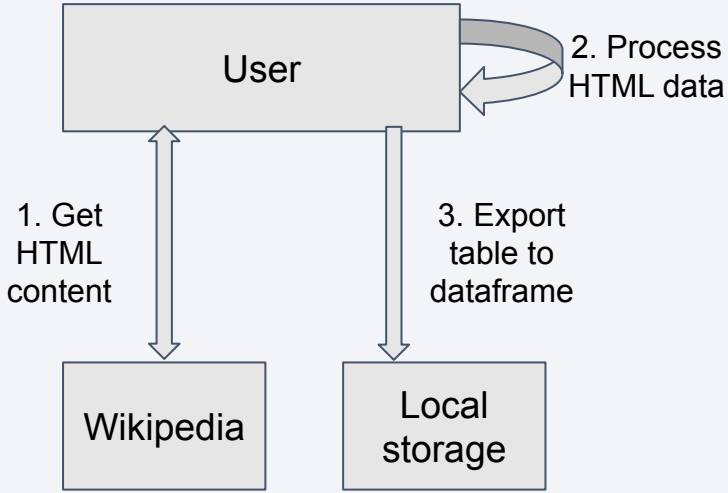
- GitHub URL of the completed notebook



# Data Collection - Scraping

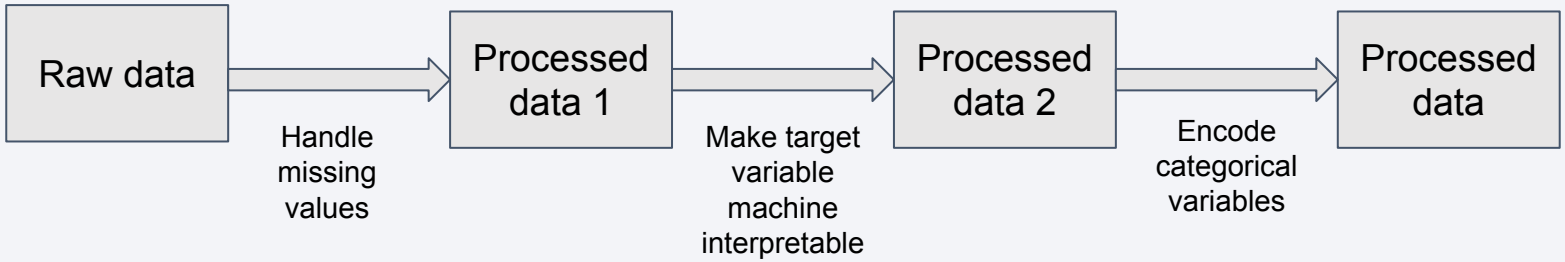
- Flowchart for the data collection from the Wikipedia by web scraping can be seen on the right hand side

- GitHub URL of the completed scraping notebook



# Data Wrangling

- Handling of missing data
  - for the payload mass, NaN values were replaced with the mean
  - for landing pad, NaN values were not replaced
- The outcome of the landing converted to numerical variable with 0 and 1 values
- Categorical variables encoded with one-hot method
- GitHub URL of the completed data wrangling notebook



# EDA with Data Visualization

- Categorical scatter plots created to visualize the correlation between some independent variables and the launch outcome
- The relationship between success rate and orbit type visualized via bar chart
- The yearly trend of the launch successes was visualized via line plot

- GitHub URL of the completed data visualization notebook

# EDA with SQL

- The following queries were made:
  - The names of the unique launch sites in the space mission
  - 5 records where launch sites begin with the string ‘CCA’
  - The total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - The date of the first successful landing outcome in ground pad
  - The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - The total number of successful and failure mission outcomes
  - Names of the booster_versions with the maximum payload mass carried
  - The failed landing_outcomes in drone ship for year 2015
  - The count of landing outcomes between 2010-06-04 and 2017-03-20

- GitHub URL of the completed EDA with SQL notebook

# Build an Interactive Map with Folium

- Circle map markers with text added to the map to mark all launch sites
- Green and red markers added to the map for each site to indicate the succeeded and failed launch attempts
- Text marker to the closest coast line point added which is showing the distance from the launch site
    - line is added to connect the two points
- Text marker to the closest highway point added which is showing the distance from the launch site
    - line is added to connect the two points

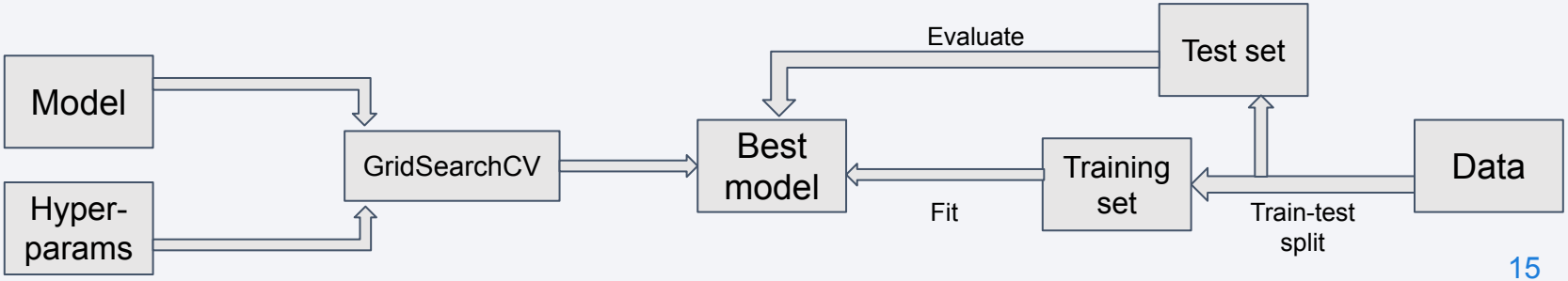
- GitHub URL of the completed Folium data visualization notebook

# Build a Dashboard with Plotly Dash

- Dropdown menu added to be able to filter the launch sites to be visualized
- Range slider added to be able to filter the Payload mass range
- Pie chart added to visualize the success rate of the launch sites
- Interactive categorical scatter plot added to visualize the correlation between the payload mass and the success rate for each booster version categories
- GitHub URL of the completed Plotly Dashboard lab

# Predictive Analysis (Classification)

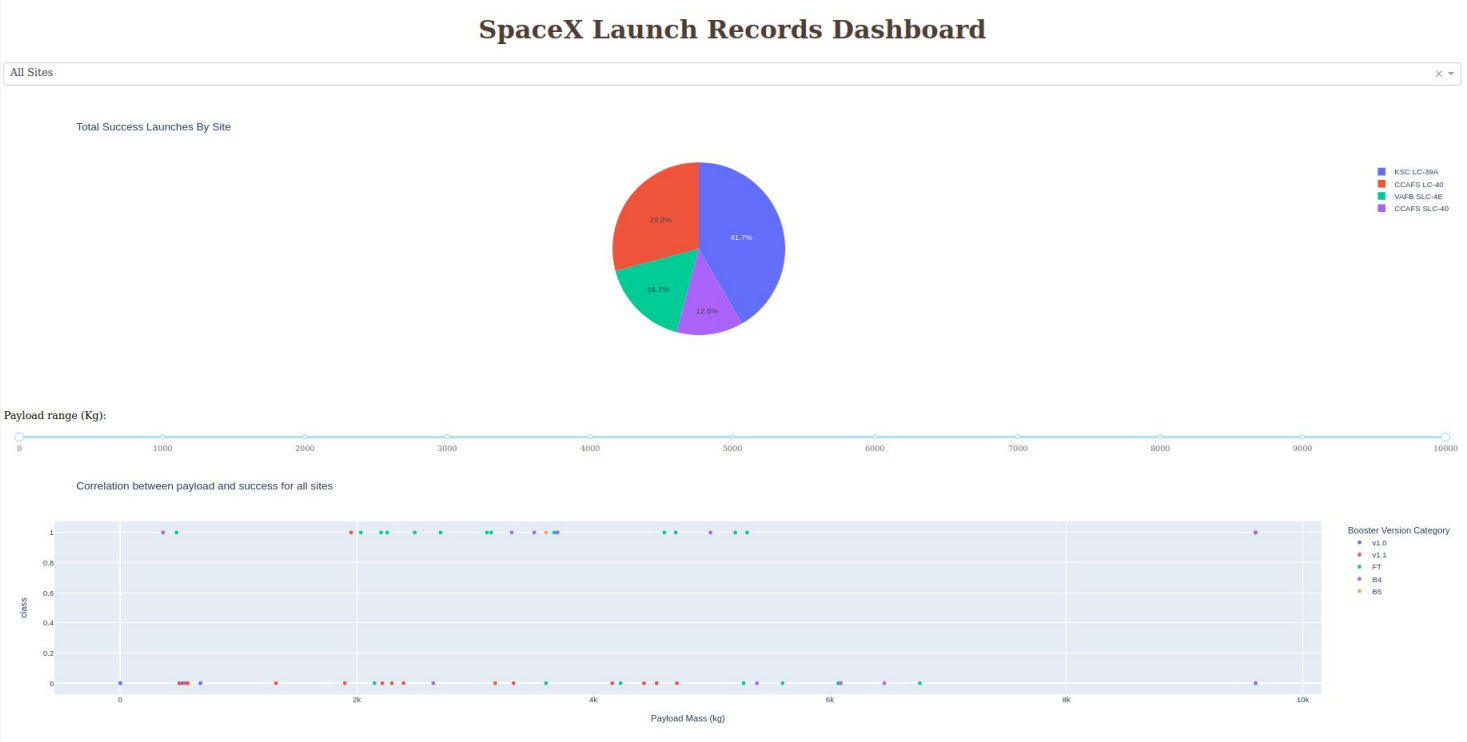
- Four classic classification model used for predictive analysis
  - K-nearest neighbour
  - Decision tree
  - Support Vector Machine
  - Logistic regression
- 80%-20% Train-test split applied on the data
- Grid search with cross-validation used to find the best parameters for each model
- GitHub URL of the completed Predictive Analysis lab



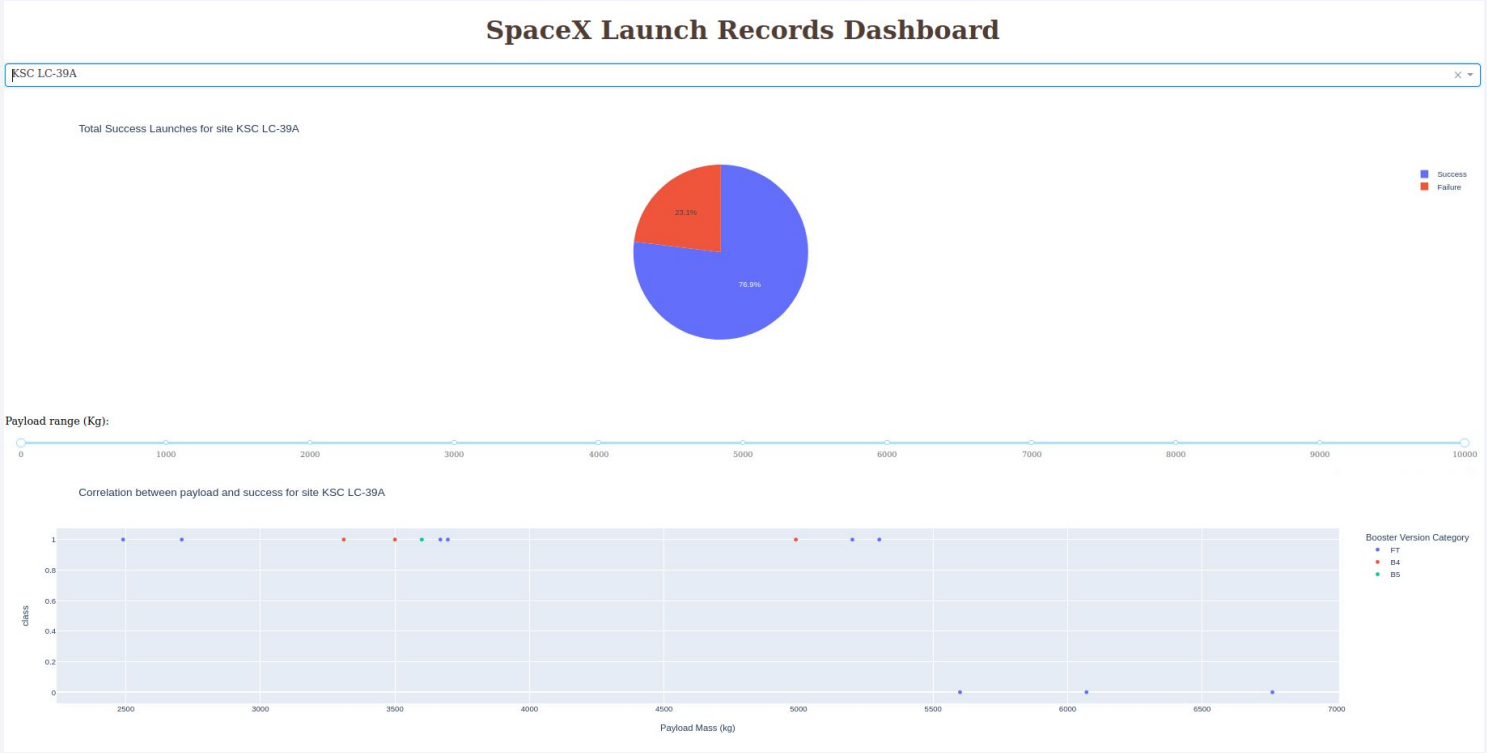
# Results - Exploratory data analysis

- Different launch sites have different success rates
  - CCAFS LC-40 - 60 %
  - KSC LC-39A, VAFB SLC 4E - 77%.
- Positive correlation between launch success and
  - Flight number
  - Payload mass
- For the VAFB-SLC launch site there are no launches for heavy payload mass (greater than 10000)
- ES-L1, GEO, HEO and SSO Orbits has 100% Success rate, for SO it is 0%
- In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit
- With heavy payloads the successful landing rate are more for Polar, LEO and ISS Orbits, However for GTO we cannot distinguish this well
- The sucess rate since 2013 kept increasing till 2020 (except for 2018)

# Results - Dashboard

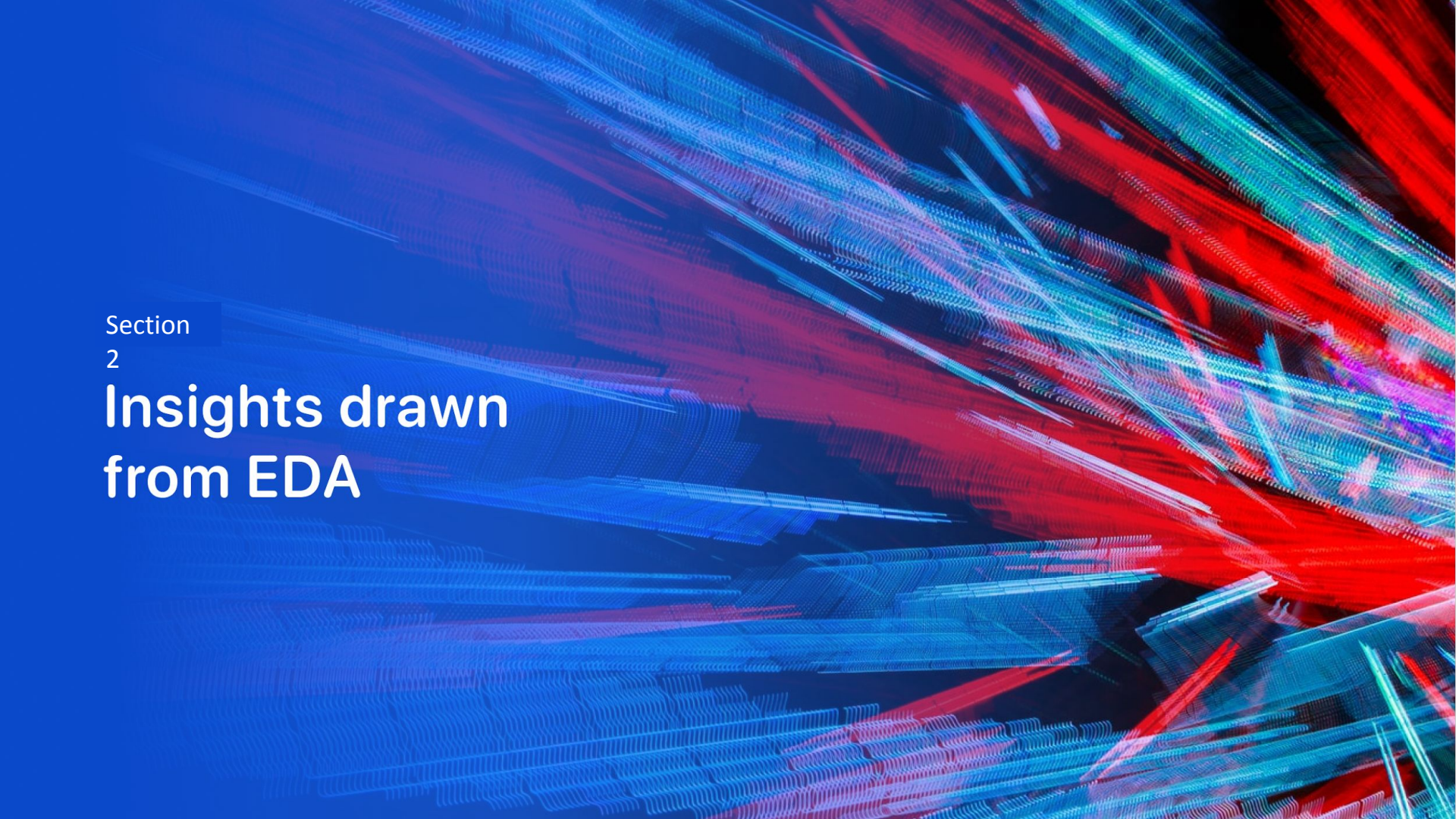


# Results - Dashboard



# Results - Predictive analysis

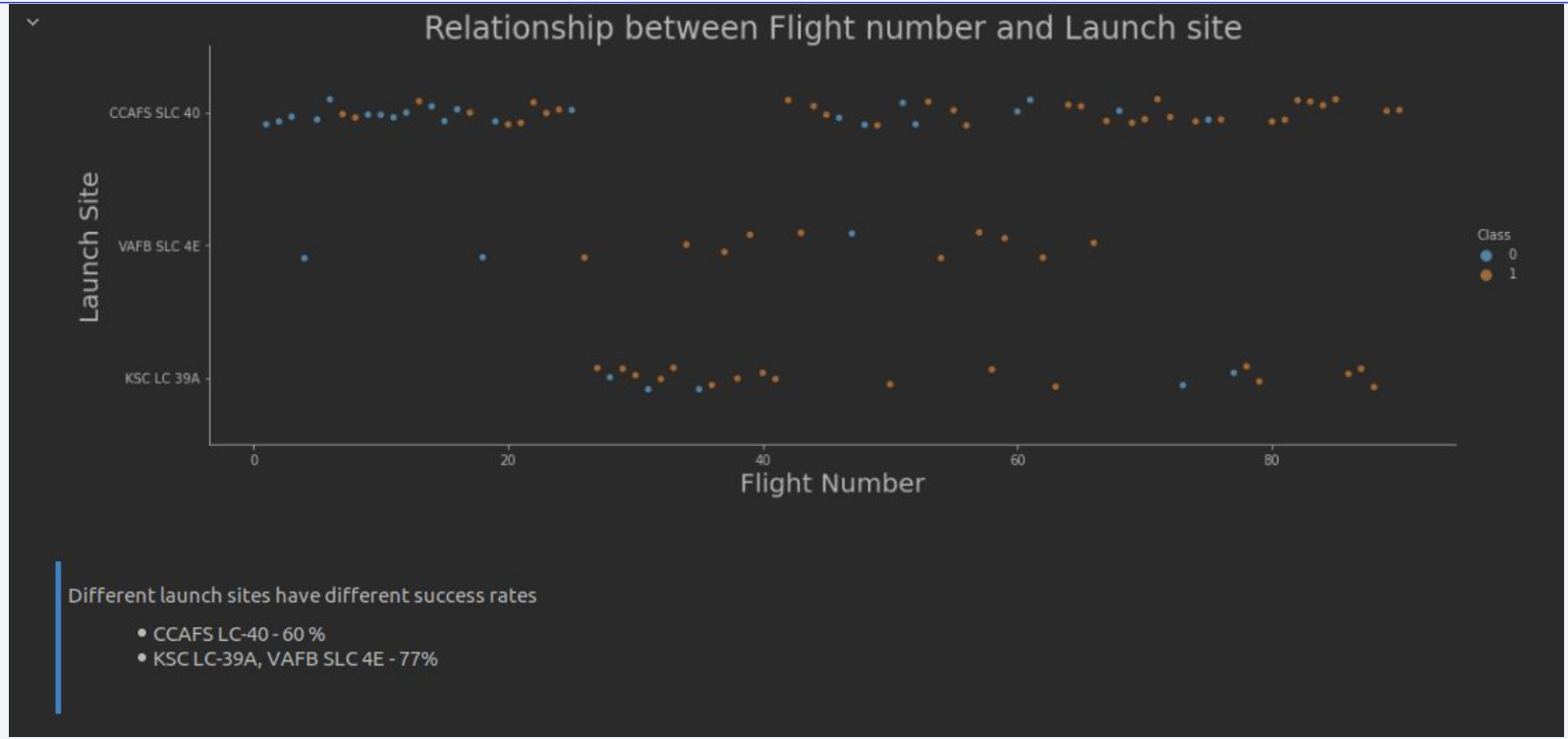
- Predictive analysis
    - The best models has approximately 83.33% accuracy



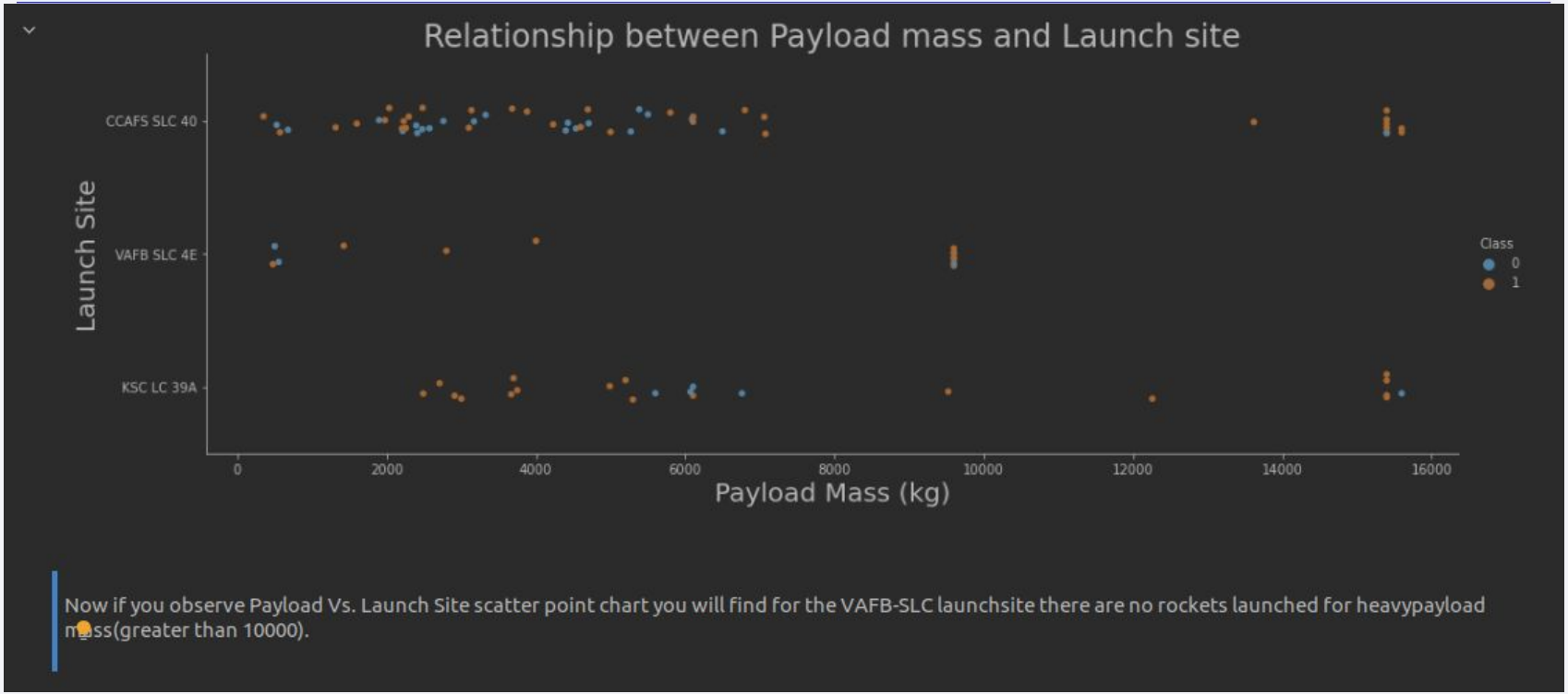
Section 2

# Insights drawn from EDA

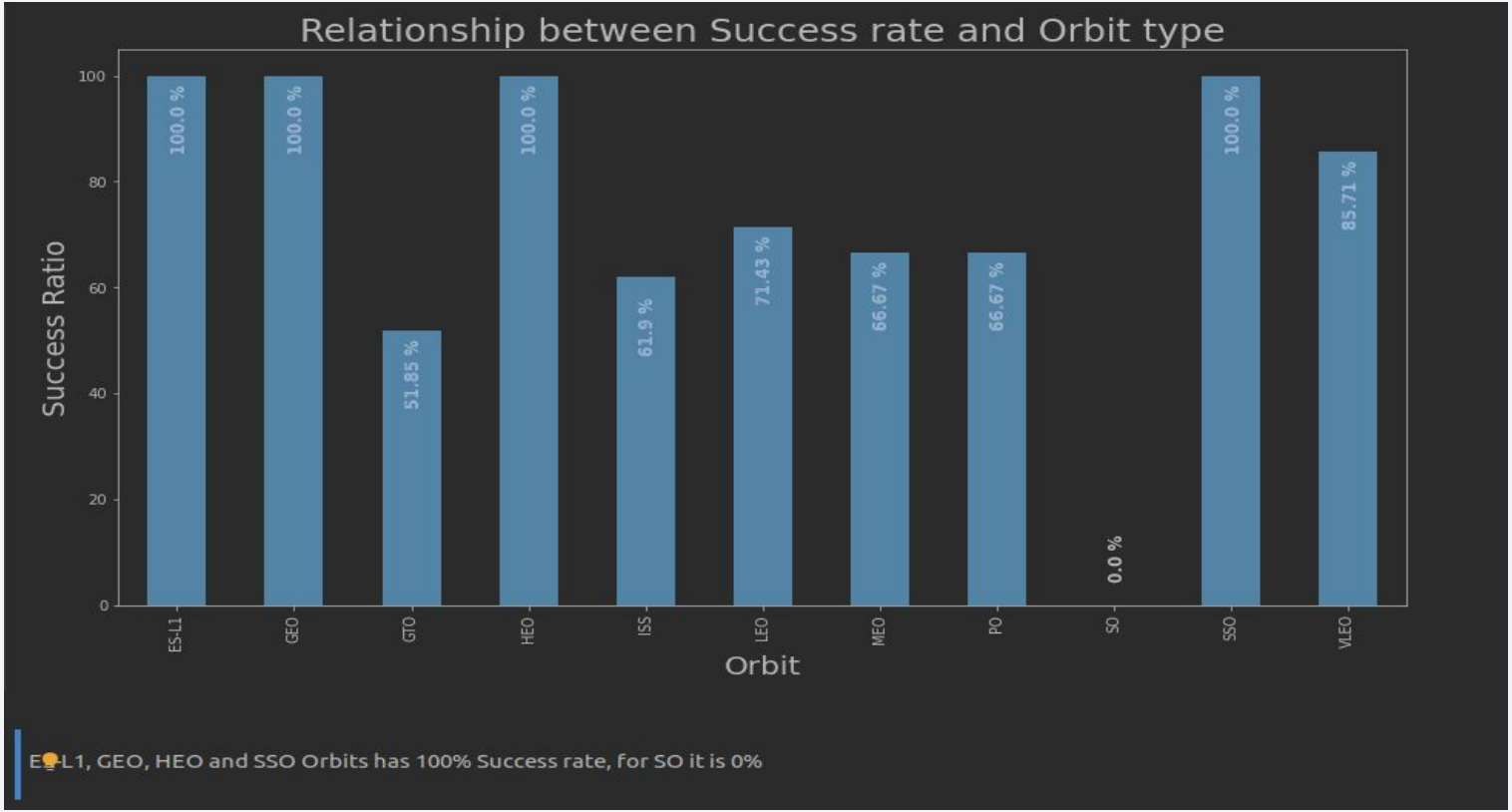
# Flight Number vs. Launch Site



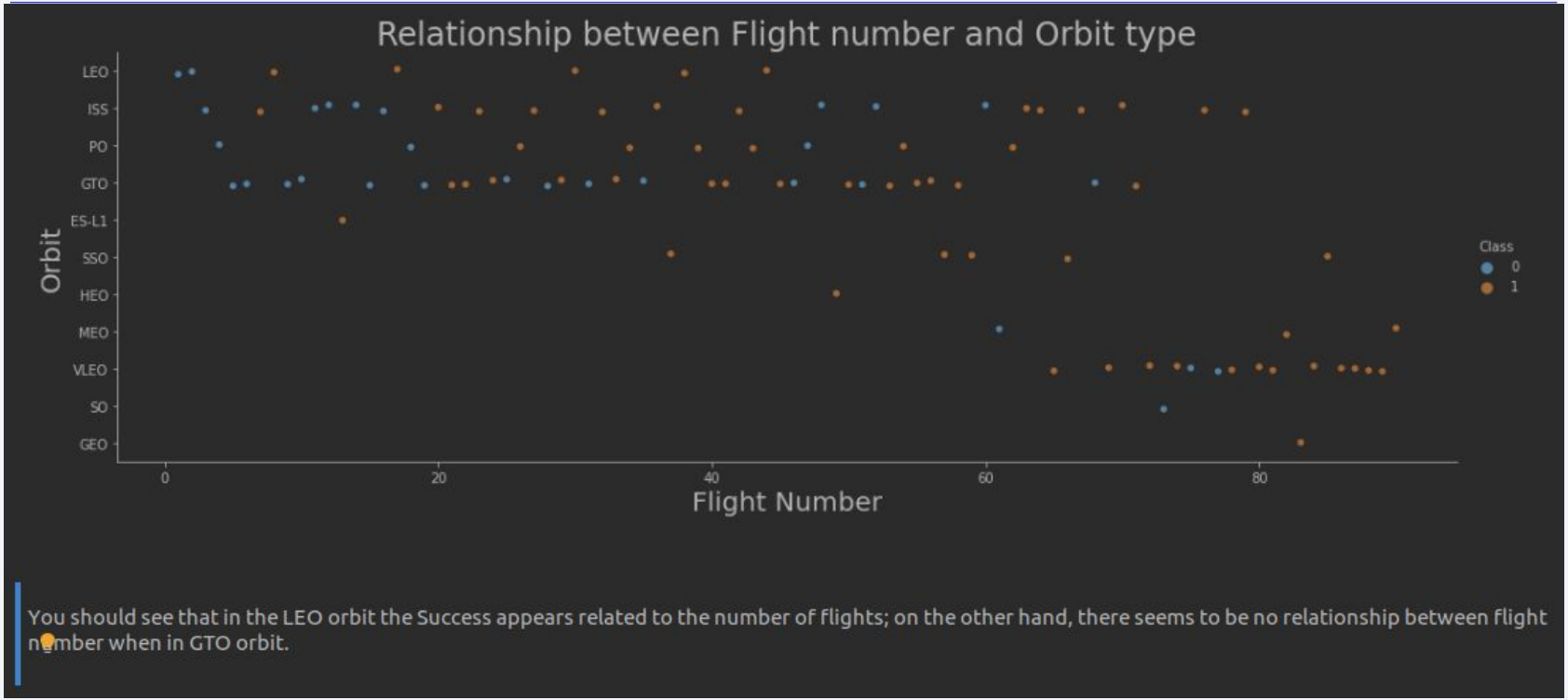
# Payload vs. Launch Site



# Success Rate vs. Orbit Type

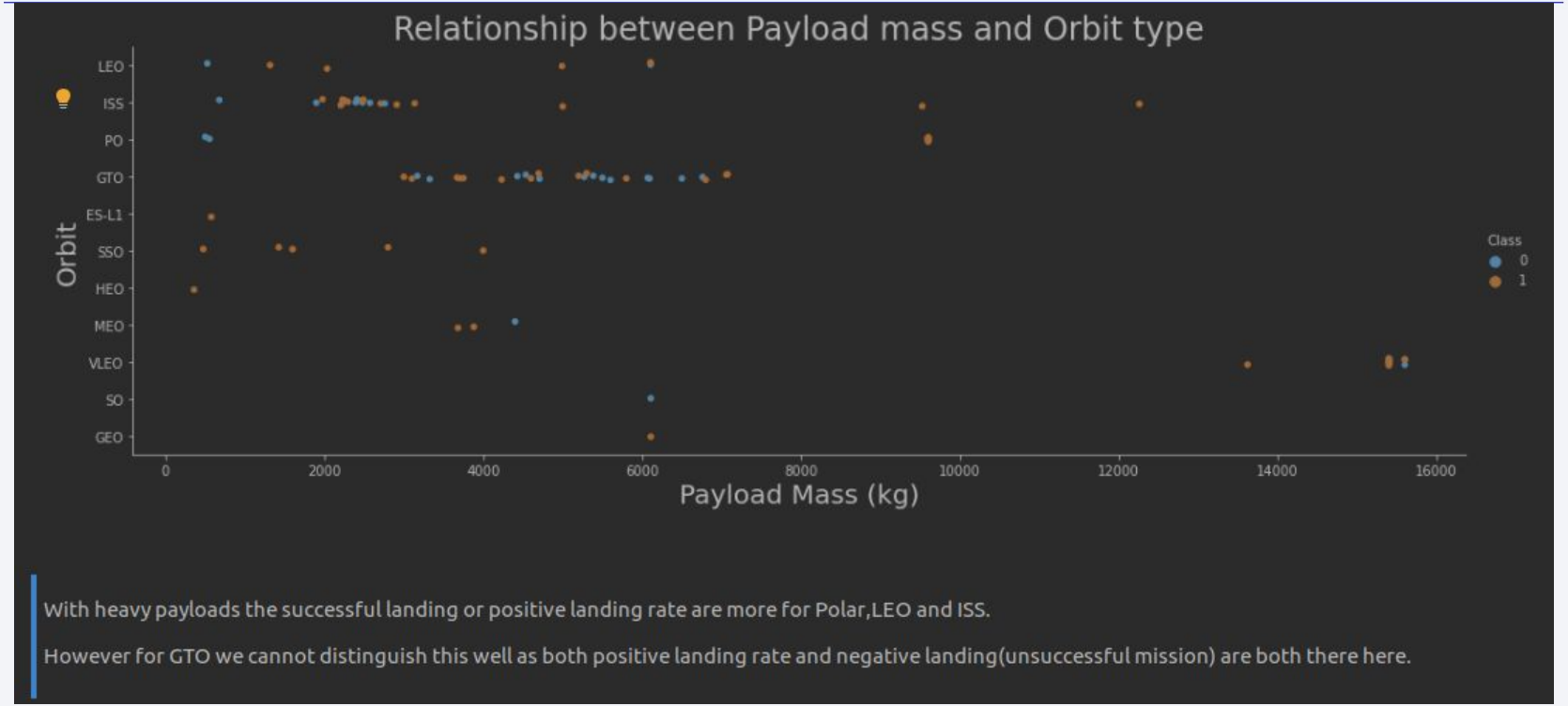


# Flight Number vs. Orbit Type




You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

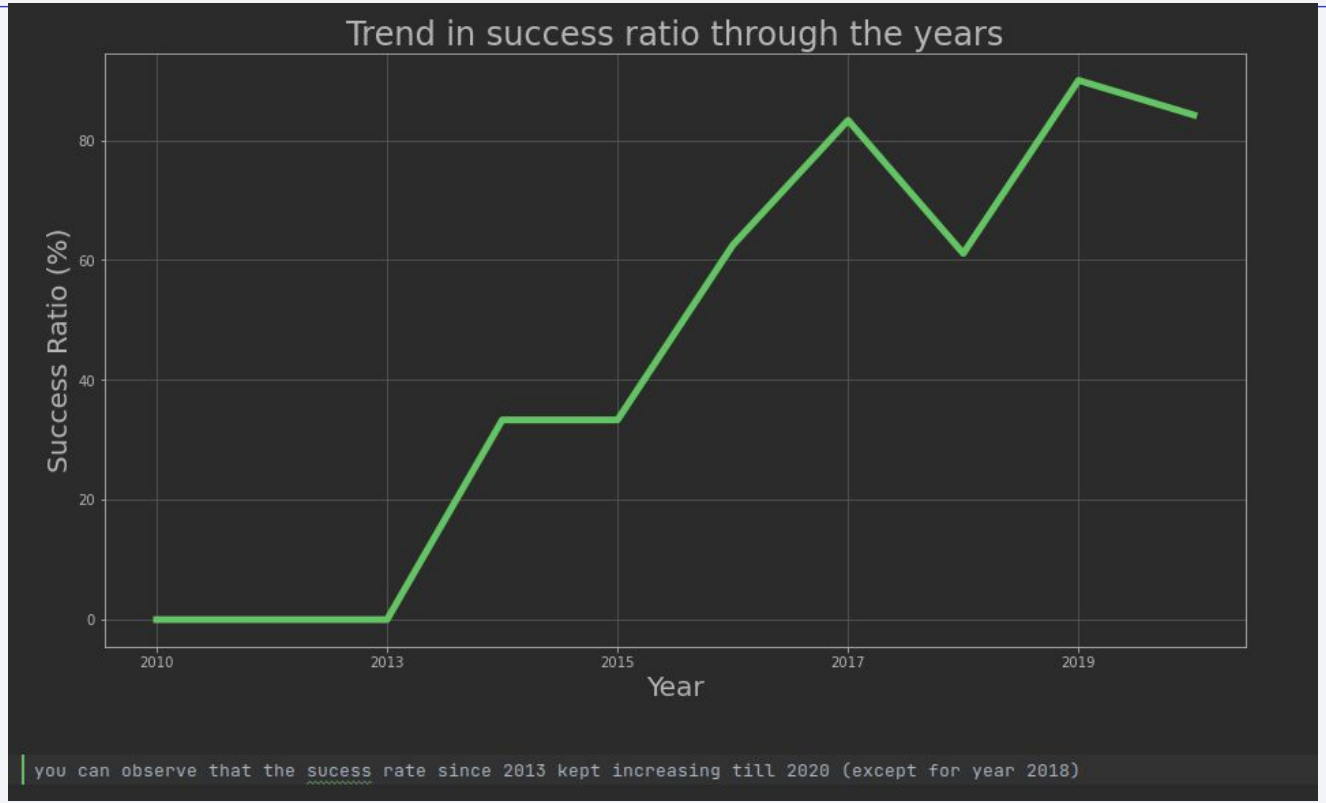
# Payload vs. Orbit Type




With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

# Launch Success Yearly Trend



# All Launch Site Names

- Find the names of the unique launch sites
- There are 4 launch sites in the space mission:
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40

*Display the names of the unique launch sites in the space mission*

```
%sql SELECT DISTINCT(Launch_Site) FROM capstone.spacex;
```

```
 * mysql+mysqldb://root:***@localhost:3306/coursera
4 rows affected.
```

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

# Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- The first 5 row stands for the ‘CCAFS LC-40’ Launch site

*Display 5 records where launch sites begin with the string ‘CCA’*

```
%sql SELECT * FROM capstone.spacex WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
```

```
 * mysql+mysqldb://root:***@localhost:3306/coursera
5 rows affected.
```

| 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

- Calculate the total payload carried by boosters from NASA
- The total payload carried by boosters from NASA is 45,596 kg

*Display the total payload mass carried by boosters launched by NASA (CRS)*

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS total_payload FROM capstone.spacex WHERE Customer = 'NASA (CRS)';
```

```
 * mysql+mysqldb://root:***@localhost:3306/coursera
1 rows affected.
```

| total_payload |
| --- |
| 45596 |

# Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- The average payload mass carried by F9 v1.1 booster is 2,534.67 kg

*Display average payload mass carried by booster version F9 v1.1*

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS avg_payload FROM capstone.spacex WHERE Booster_Version LIKE 'F9 v1.1%';
```

```
 * mysql+mysqldb://root:***@localhost:3306/coursera
1 rows affected.
```

| avg_payload |
| --- |
| 2534.6667 |

# First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- The first successful landing outcome on ground pad was at 2015-12-22

***List the date when the first successful landing outcome in ground pad was acheived.***

*Hint:Use min function*

```
%sql SELECT MIN(Date) AS first_ground_pad_success_date FROM capstone.spacex WHERE `Landing _Outcome` = 'Success (ground pad)';
```

```
 * mysql+mysqldb://root:***@localhost:3306/coursera
1 rows affected.
```

| first_ground_pad_success_date |
| --- |
| 2015-12-22 |

# Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- The boosters with the specified criteria:
    - F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2

*List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000*

```
%%sql
SELECT Booster_Version FROM capstone.spacex
    WHERE `Landing _Outcome` = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

```
 * mysql+mysqldb://root:***@localhost:3306/coursera
4 rows affected.
```

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

# Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- The number of successes: 61
- The number of failures: 10

*List the total number of successful and failure mission outcomes*

```
%%sql

SELECT SUM(CASE WHEN `Landing _Outcome` LIKE 'Success%' THEN 1 ELSE 0 END) AS success_count,
       SUM(CASE WHEN `Landing _Outcome` LIKE 'Failure%' THEN 1 ELSE 0 END) AS failure_count FROM capstone.spacex;
```

```
 * mysql+mysqldb://root:***@localhost:3306/coursera
1 rows affected.
```

| success_count | failure_count |
|---|---|
| 61 | 10 |

# Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- There are 12 booster versions with the criteria

*List the names of the booster_versions which have carried the maximum payload mass. Use a subquery*

```
%%sql
SELECT DISTINCT(Booster_Version) FROM capstone.spacex
    WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM capstone.spacex);
```

```
 * mysql+mysqldb://root:***@localhost:3306/coursera
12 rows affected.
```

| 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 |

# 2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- In 2015 there were 2 failed outcomes, both at CCAFS LC-40, one with booster version F9 v1.1 B1012 and the other with F9 v1.1 B1015

*List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015*

```
%%sql
SELECT Booster_Version, Launch_Site FROM capstone.spacex
    WHERE `Landing _Outcome` = 'Failure (drone ship)' AND YEAR(Date) = 2015;
```

```
 * mysql+mysqldb://root:***@localhost:3306/coursera
2 rows affected.
```

| Booster_Version | Launch_Site |
|---|---|
| F9 v1.1 B1012 | CCAFS LC-40 |
| F9 v1.1 B1015 | CCAFS LC-40 |

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

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order
- The most frequent landing outcome was ‘No attempt’ in the specified interval

*Rank the count of landing outcomes (such as Failure (drone ship) or Success(ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order*

```
%%sql

SELECT `Landing _Outcome`, COUNT(*) AS count_of_outcomes FROM capstone.spacex
    WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
    GROUP BY `Landing _Outcome` ORDER BY count_of_outcomes DESC;
```

```
 * mysql+mysqldb://root:***@localhost:3306/coursera
8 rows affected.
```

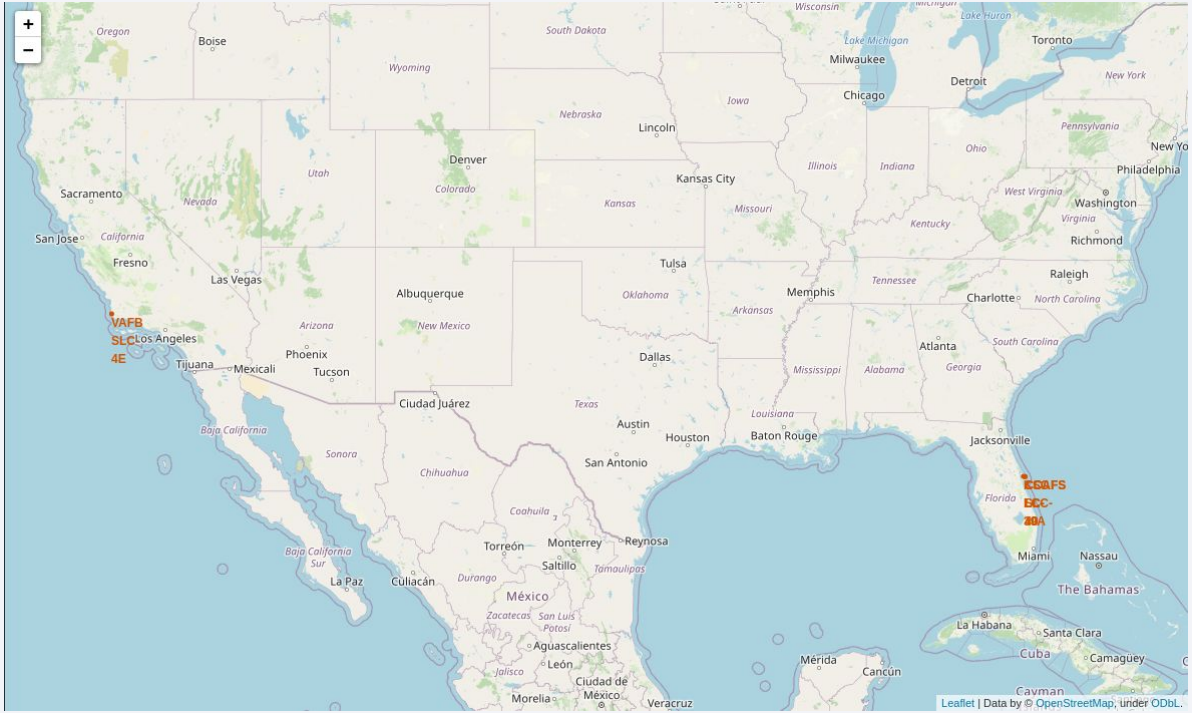
| Landing _Outcome | count_of_outcomes |
|---|---|
| No attempt | 10 |
| Failure (drone ship) | 5 |
| Success (drone ship) | 5 |
| Controlled (ocean) | 3 |
| Success (ground pad) | 3 |
| Failure (parachute) | 2 |
| Uncontrolled (ocean) | 2 |
| Precluded (drone ship) | 1 |

Section 3

# Launch Sites Proximities Analysis

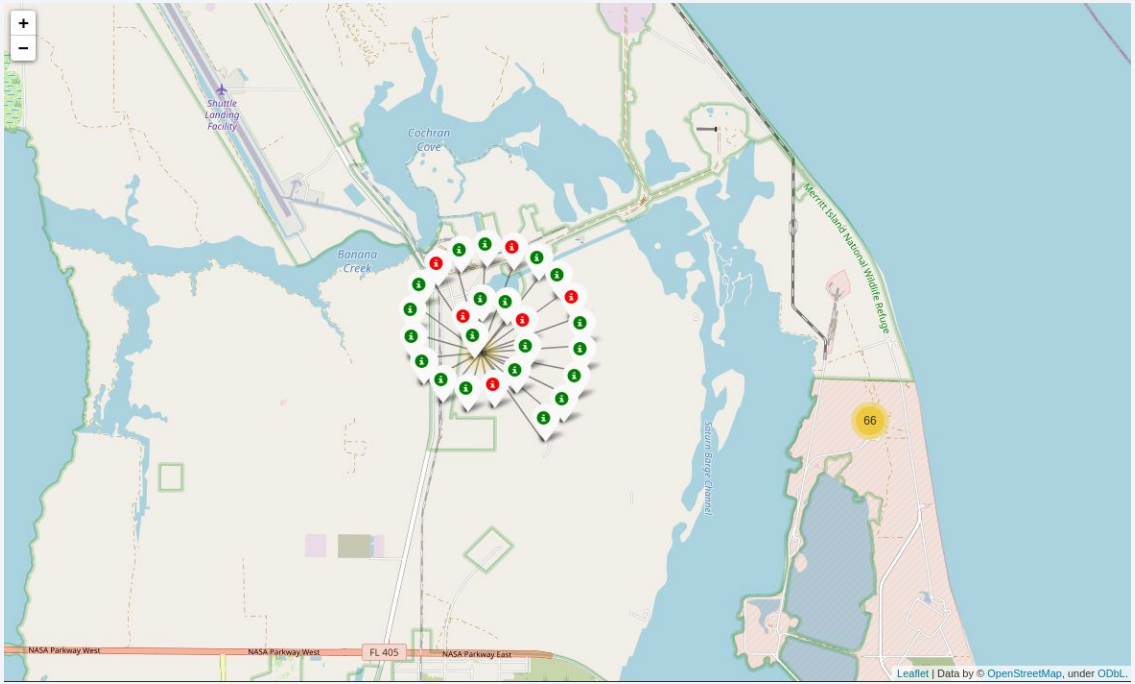
# Launch site locations

- There are launch sites near Los Angeles and also near Orlando, Florida



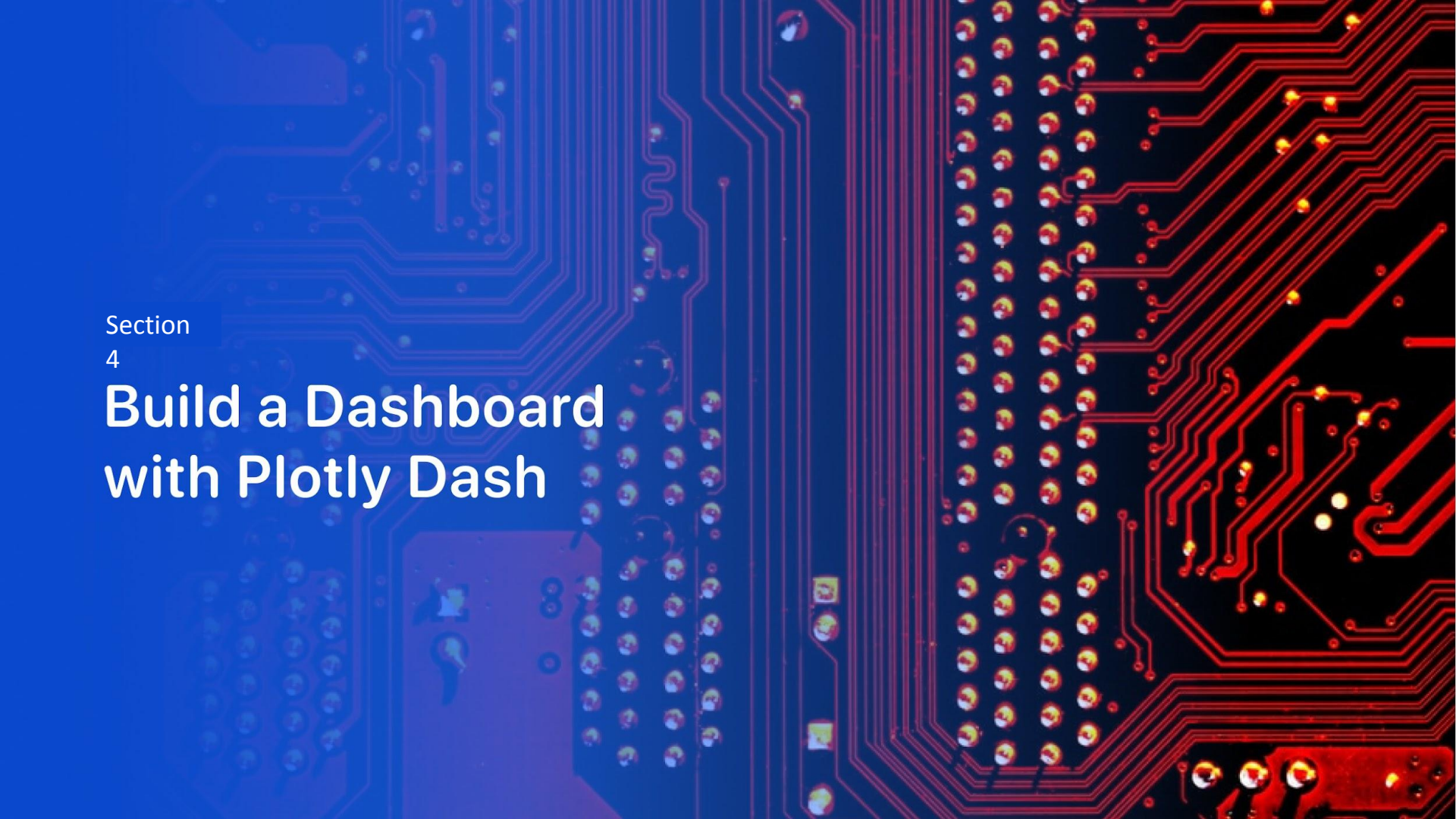
# Succeeded and failed launches

- For KSC LC-39A site, 6 out of 26 attempts failed



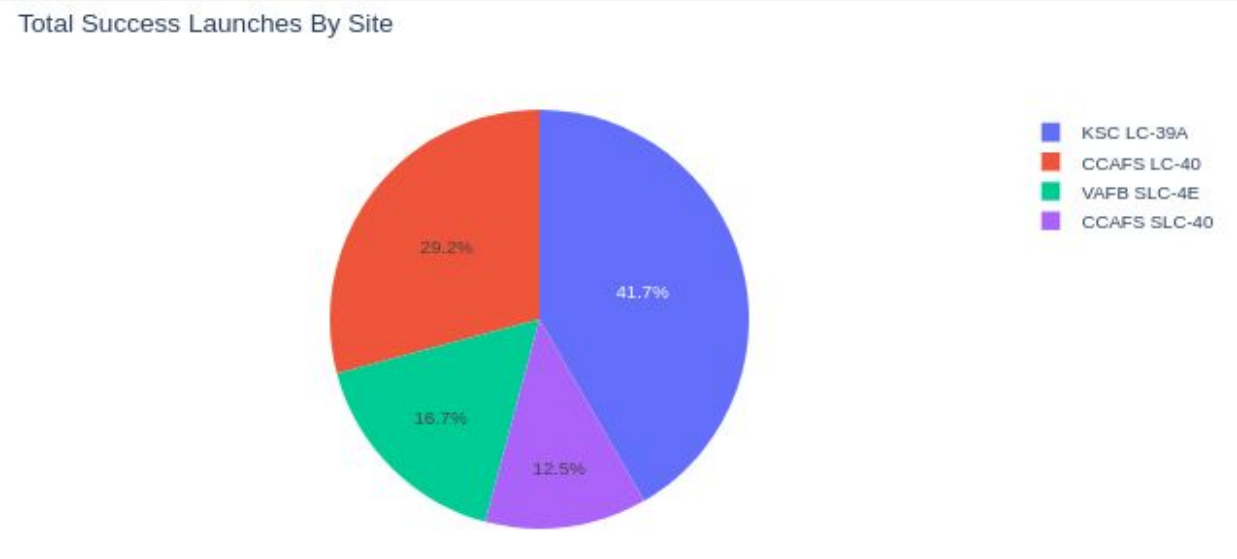
# Nearest coastline point

- The distance to the nearest coastline point from VAFB SLC-4E site is 1.36 km


Section
4
Build a Dashboard
with Plotly Dash

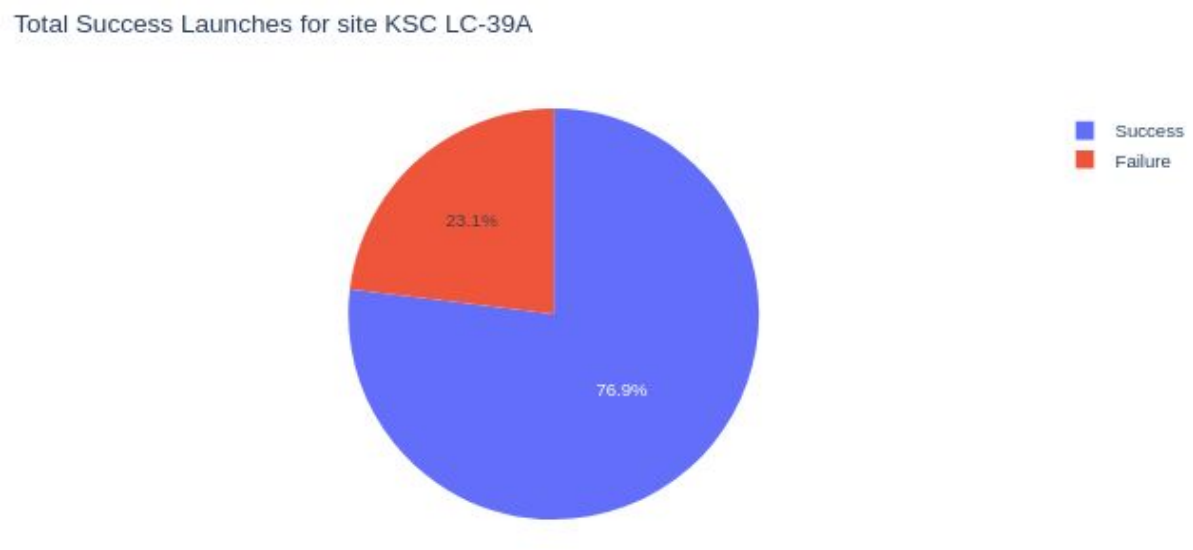
# Success rate for launch sites

- KSC LC-39A has the most succeeded launches among the sites
- CCAFS SLC-40 has the least succeeded launches among the sites



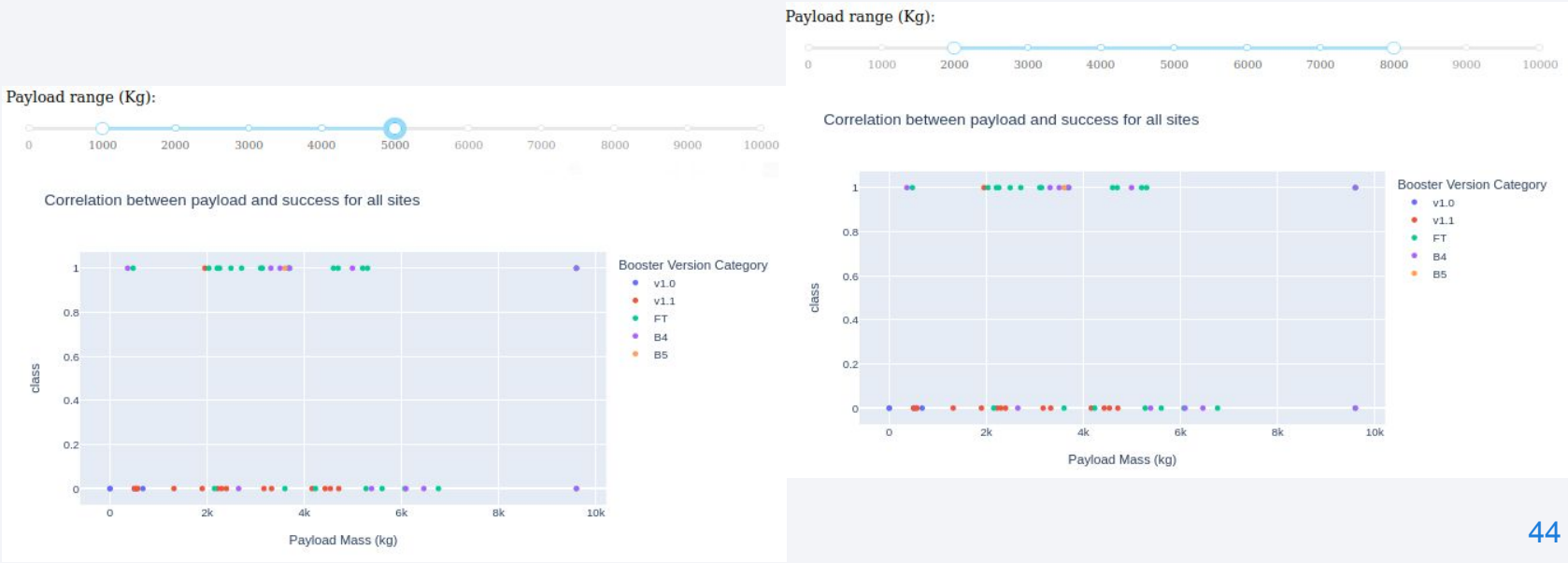
# KSC LC-39A success rate

- KSC LC-39A has a 76.9% success rate which is the highest among the sites



# Correlation between payload and success

- Payload range between 2000 and 6000 has the highest success rate
- Booster version FT has the most successes

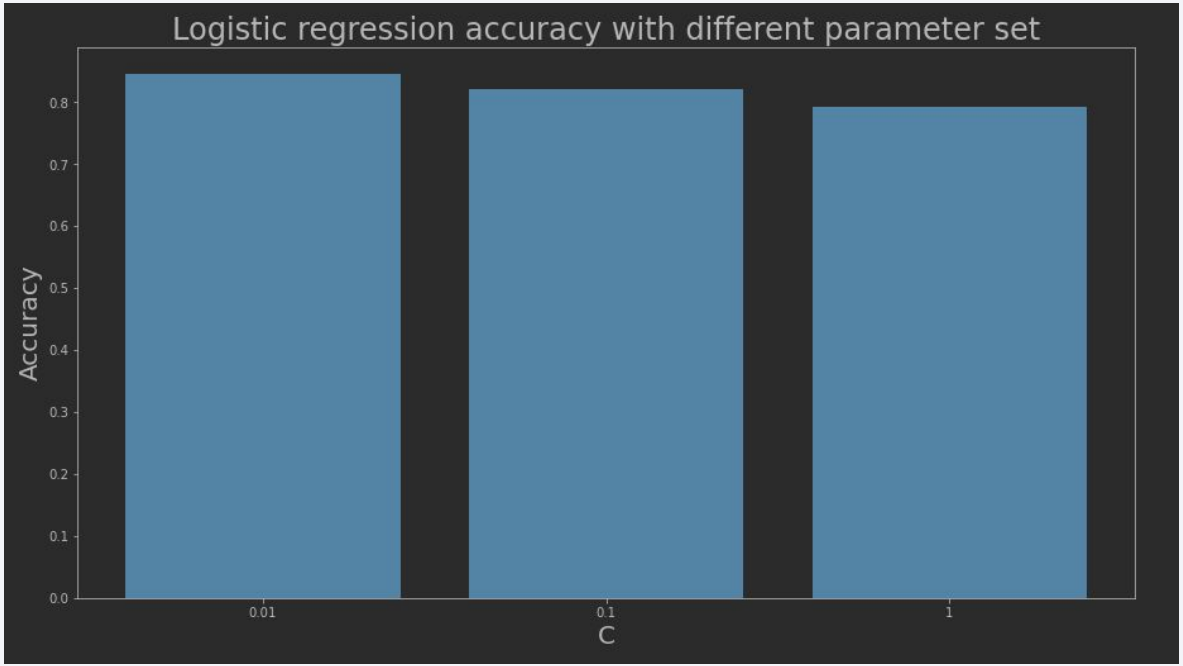


Section
5

# Predictive Analysis (Classification)

# Classification Accuracy - Logistic regression

- For logistic regression, the highest classification accuracy comes with C=0.01



# Classification Accuracy - SVM

- For SVM, the highest classification accuracy comes with C=1, gamma=0.032 and sigmoid kernel

# Classification Accuracy - Decision tree

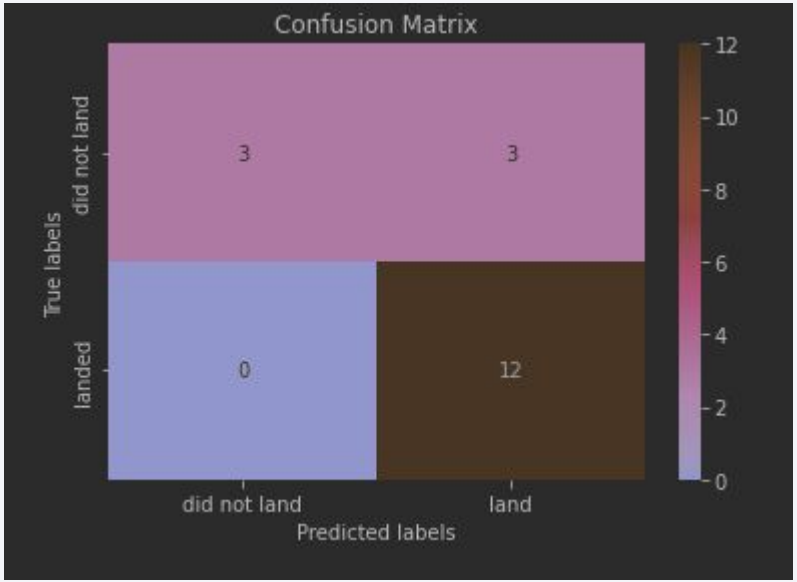
- For decision tree, the best accuracy comes with "entropy" criterion, "sqrt" max features, 4 minimum samples per leaf, 5 minimum samples per split and "random" splitter hyperparameters

# Classification Accuracy - KNN

- For KNN, the highest classification accuracy comes with ‘auto’ algorithm’, 10 neighbours and p=1

# Confusion Matrix

- The decision tree model has the best model accuracy
    - there are 3 false positives for the test set and no false negatives



# Conclusions

- With the developed models we are able to predict if a given landing attempt will be successful or not with an accuracy between 83% and 89%
- The models can be improved with
    - collecting additional training data and/or
    - extending the hyperparameter set of the models
    - using other techniques like neural networks or other ML algorithms

# Appendix

- Link to the GitHub repository which contains all the notebooks and all the other assets: Repository

# Thank you!