IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Lukas Niebler
11th April 2023

Outline

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

Executive Summary

- **Summary of methodologies:**

  The research target is to identify the key-factors for a successful rocket landing for cost reduction.
  Hence following approaches were used:

  - **Collect** data using SpaceX REST-API in combination with web scraping
  - **Wrangle** data to create a binary outcome variable (success/fail)
  - **Visualize** the launch outcome focusing on payload, launch site, flight number and yearly trend
  - **Analyze** the data based on Pandas DataFrame and SQL calculating total payload, factor ranges for successful launches and the success/failure rates
  - **Explore** launch site success rate and their geographic distribution
  - **Visualize** the factor depended success rates using an interactive Folium Dashboard
  - **Building Models** to predict landing outcomes using logistic regression, support vector machine (SVM), decision tree and K-nearest neighbors (KNN)

- **Results:**

Analysis:

- Improving launch success over time
- Kennedy Space Center LC-39A highest local success rate

Interactive Visualization:

- Success rate by launch site, 100% for orbits ES-L1, GEO, HEO and SSO

Prediction:

- Comparable performance of the examined models, decision tree model slightly best performance

Introduction

Background

SpaceX advertises Falcon 9 rocket launches on its website, with a cost of **$62 million**; other providers cost upward of **$165 million** each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

Targets

- Factors influencing a successful landing using accessible Data
- Factor dependencies to describe the landing success rate
- Increasing success probability through model-based prediction

Section 1

Methodology

Methodology

- Data collection:
  - SpaceX REST-API
  - Web Scraping – List of Launches
- Perform data wrangling
  - Binary Classification Outcome of Landing
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - logistic regression, support vector machine (SVM), decision tree and K-nearest neighbors (KNN)

Data Collection

REST-API

Request Data
form SpaceX (rocket launch data)

Decode Response
.Json in pandas DataFrame using .json_normalize()

Request further Information
using SpaceX API request by calling for 24char IDs

Consolidate Data
combine data in a Dictionary and transform into a DataFrame

Filter and Clean Data
filter for Falcon 9 (reusable) and replacing missing Payload Mass by .mean()

Export Data
to .csv file

https://github.com/luni6445/Capstone_Project/blob/main/Data%20Collection%20API%20Lab.ipynb

Web Scraping

Request Data
from Wikipedia (Falcon 9 launch data till June 2021)

BeautifulSoup Object
by parsing trough HTML response

Extract Columns
by iterating through the first listed launch data table header (2010-2013)

Collect Data
by parsing through HTML tables (starting form 2010-2013)

Consolidate Data
combine data in a Dictionary and transform into a DataFrame

Export Data
to .csv file

https://github.com/luni6445/Capstone_Project/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

- Check Data Consistency
  Missing values, data types, launch site and orbit distribution
- Binary Landing Outcomes Classification
  1 – Positive: True Ocean, True RTLS, True ASDS
  0 – Negative: None None, False ASDS, None ASDS, False RTLS
- Export Data to .csv file

➤ https://github.com/luni6445/Capstone_Project/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Plotted Charts:
  - flight number vs. payload mass
  - flight number vs. launch site
  - payload mass vs. launch site
  - payload mass vs. orbit type
  - orbit type vs success rate
  - flight number vs. orbit type
  - year vs success rate

➤ https://github.com/luni6445/Capstone_Project/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

Queries:

- **Display**
    - names of the unique launch sites
    - 5 records where launch sites begin with the string 'KSC’
    - total payload mass carried by boosters launched by NASA (CRS)
    - average payload mass carried by booster version F9 v1.1
- **List**
    - First successful landing date on ground pad
    - Booster versions successfully landed on ground pad (payload: 4000-6000kg)
    - Total number of successful and failure mission outcomes
    - Names of the booster versions carried the maximum payload mass
    - Successful landing outcome on ground pad, booster versions and launch site for the months in 2017
    - Amount of successful landing outcome (04-06-2010 - 20-03-2017, desc.)

➤ https://github.com/luni6445/Capstone_Project/blob/main/jupyter-labs-eda-sql-edx_sqllite.ipynb

Build an Interactive Map with Folium

- Mark all launch sites on a map to show their location
- Mark the success/failed launches for each site on the map to display the success rates of each site
- Calculate the distances between a launch site to its proximities to check the distance to next railway, highway and city.

➤ https://github.com/luni6445/Capstone_Project/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

Dashboard Content:

- **Dropdown List** to enable launch site selection
- **Pie Chart** to show the total successful launches count for all sites
- **Slider** to select payload range
- **Scatter Chart** to show the correlation between payload and launch success

➤ https://github.com/luni6445/Capstone_Project/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

1. Split the data in label Y-array and standardized feature X-array
2. Split the data X and Y into training and test data using train_test_split()
3. Apply GridSearchCV(cv=10) for parameter optimization on:
   - logistic regression
   - support vector machine
   - decision tree
   - k-nearest neighbour
4. Evaluate and identified best preforming models by
   - accuracy
   - precision
   - recall
   - f1 score

- https://github.com/luni6445/Capstone_Project/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

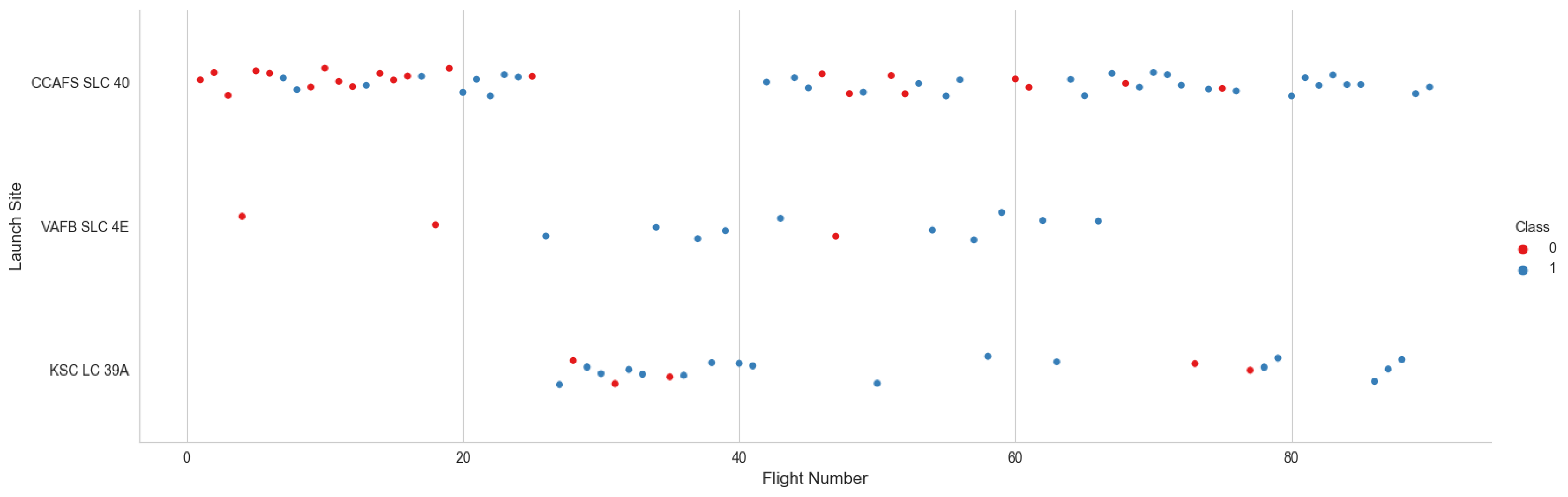
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

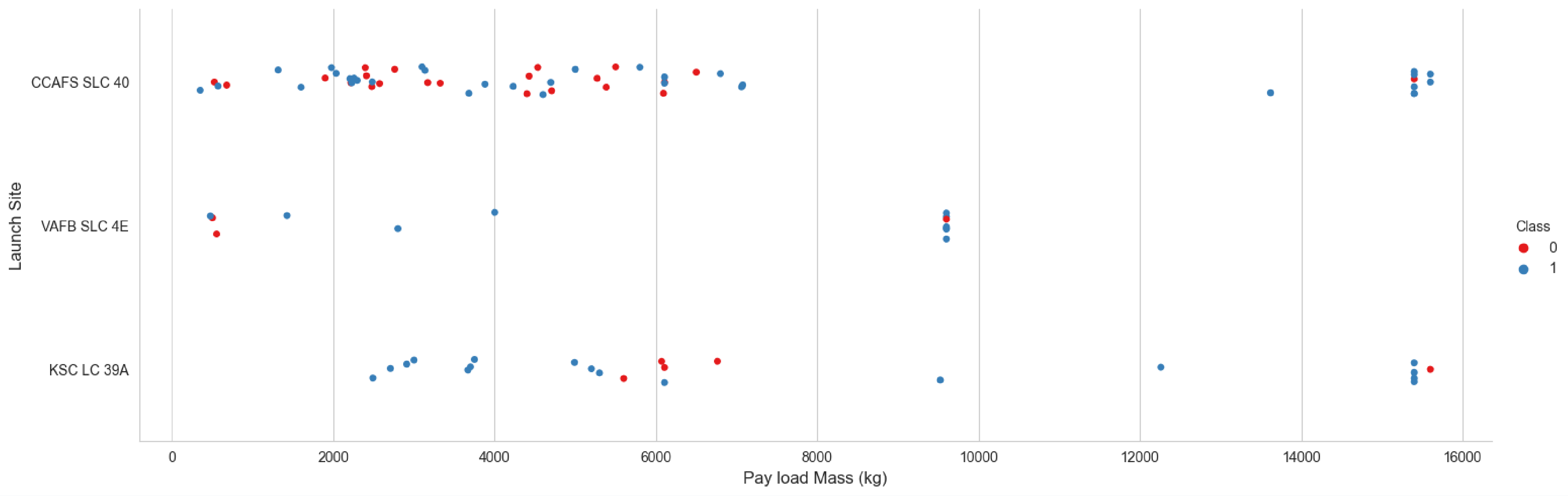
Flight Number vs. Launch Site

- The success rate increasing with flight number (1= success | 0 = failure)
- Most of the launches were from CCAFS
- VAFB and KSC have higher success rate



Flight Number vs. Launch Site

- Success rate increasing with payload mass
- Uneven Distribution of Payload
- KSC were completely successful under 5.500kg

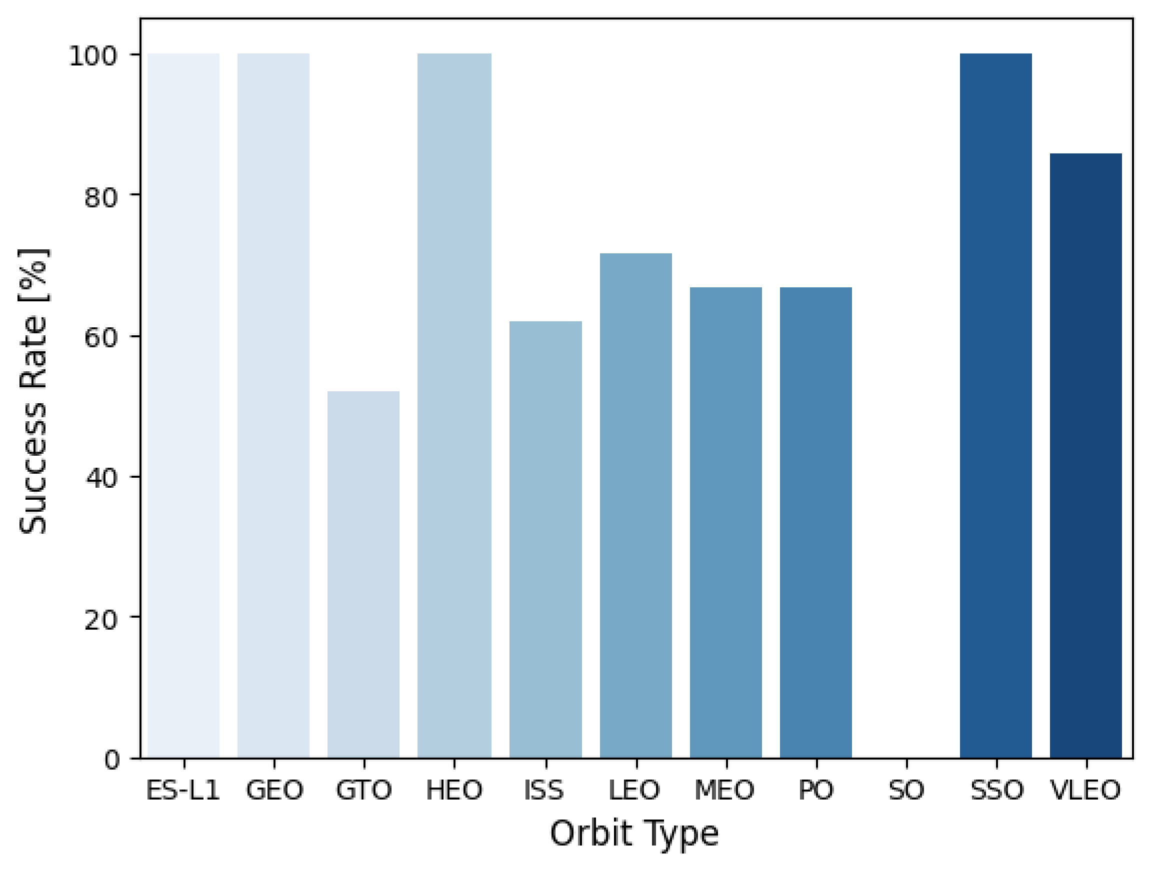


Success Rate vs. Orbit Type

100% Success Rate:
ES-L1, GEO, HEO & SSO

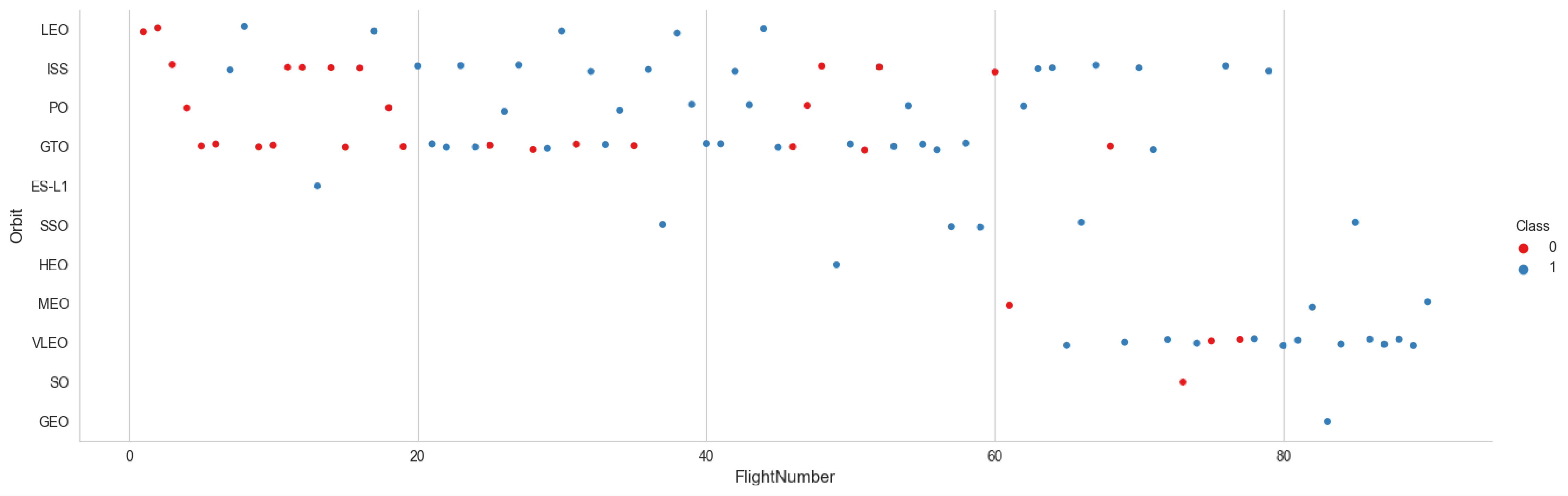
50-85% Success Rate:
GTO, ISS, LEO, MEO, PO & VLEO

0% Success Rate:
SO



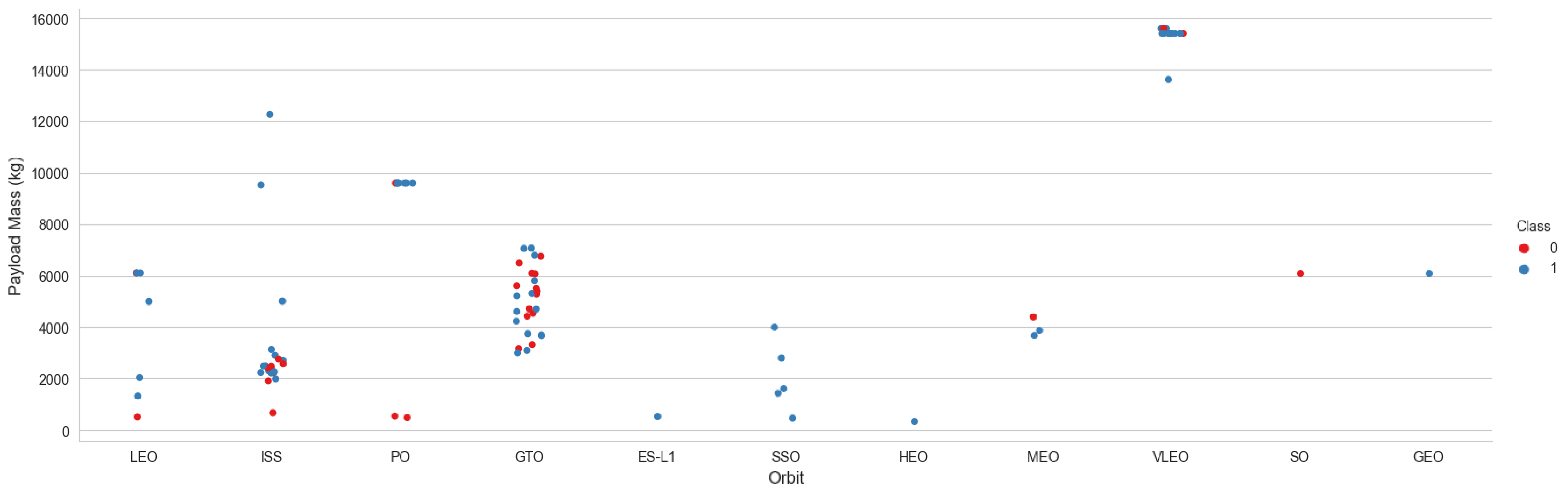
Flight Number vs. Orbit Type

- Orbits with 100 or **0%** Success Rate have the lowest amount of flights
  ES-L1, GEO, HEO, SSO & SO
- Success rate is increasing with the number of flights for all orbits
- LEO orbit shows success strongly linked to flight count; less relation found for GTO orbit.



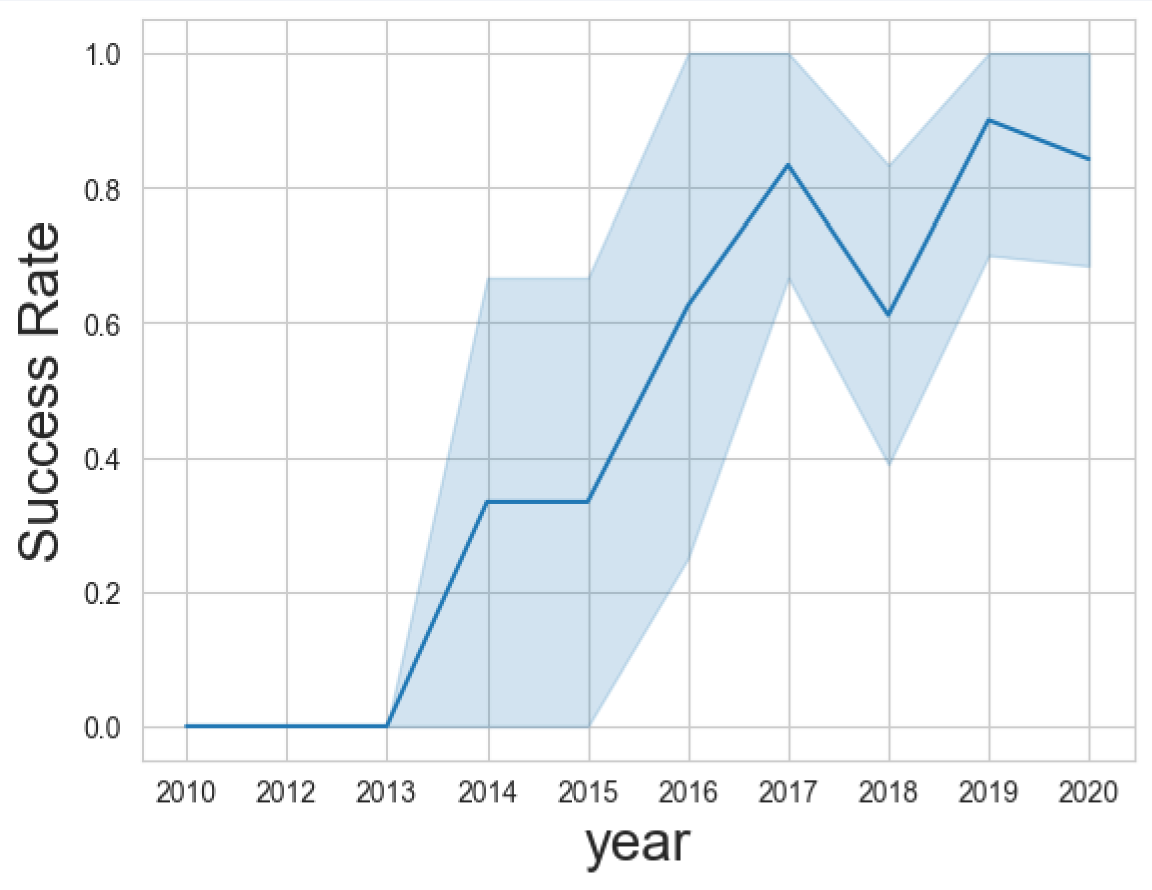
Payload vs. Orbit Type

- Increased Landing success with heavier payloads
- Higher success rate PO, LEO and ISS orbits with increasing payload mass
- Relation not existing for GTO



Launch Success Yearly Trend

- Increasing success rate for the years 2013-2017 & 2018-2019
- Decreasing success rate for the years 2018 & 2020
- Overall increasing success rate



All Launch Site Names

- **SQL-queries**

  ```
  SELECT DISTINCT "Launch_Site“
  FROM SPACEXTBL;
  ```

- Used DISTINCT to show only unique entries

- **All launch sites SpaceX starts off**

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

Launch Site Names Begin with 'KSC'

- **SQL-queries**

```
SELECT *
FROM SPACEXTBL
WHERE "Launch_Site" LIKE 'KSC%'
LIMIT 5;
```

- **5 records where launch sites begin with the string 'KSC'**

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
19-02-2017	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
16-03-2017	06:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
30-03-2017	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
01-05-2017	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
15-05-2017	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

Payload Mass

- **SQL-queries**

  SELECT SUM("PAYLOAD_MASS__KG_") AS "TOTAL_PAYLOAD_MASS__KG_“
  FROM SPACEXTBL
  WHERE "Customer" LIKE 'NASA (CRS)’;

- **Result: <u>45596 Kg</u>**

- The sum of the Payload is Calculated in Total Payload Mass, with WHERE clause

Payload Mass

Total Payload Mass (NASA)

45 596 Kg

Carried by boosters launched by NASA (CRS)

```
%%sql
SELECT SUM("PAYLOAD_MASS__KG_") AS "TOTAL_PAYLOAD_MASS
FROM SPACEXTBL
WHERE "Customer" LIKE 'NASA (CRS)';
```

Python

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

TOTAL_PAYLOAD_MASS__KG_
45596

Average Payload Mass (F9 v1.1)

45 596 Kg

Carried by booster version F9 v1.1

```
%%sql
SELECT AVG("PAYLOAD_MASS__KG_")
AS "AVG_PAYLOAD_MASS__KG_"
FROM SPACEXTBL
WHERE "Booster_Version" LIKE 'F9 v1.1';
```

Python

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

AVG_PAYLOAD_MASS__KG_
2928.4

Interesting Mission Information

1st Landing Success on Ground Pad

01-05-2017

```
%%sql
SELECT MIN("Date") AS "First_Successful_Landing_Date"
FROM SPACEXTBL
WHERE "Landing_Outcome" LIKE 'Success (ground pad)'
```

Python

* sqlite:///SPACEXDB.db
Done.

First_Successful_Landing_Date
01-05-2017

Drone Ship Landing 4000-6000 kg

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

```
%%sql
SELECT "Booster_Version"
FROM SPACEXTBL
WHERE "Landing_Outcome" LIKE 'Success (drone ship)'
    AND "Payload_Mass__kg_" > 4000
    AND "Payload_Mass__kg_" < 6000;
```

Python

* sqlite:///SPACEXDB.db
Done.

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

Total Number of

Success 100

Failure 1


```
%%sql
SELECT "Booster_Version"
FROM SPACEXTBL
WHERE "Landing_Outcome" LIKE 'Success (drone ship)'
    AND "Payload_Mass__kg_" > 4000
    AND "Payload_Mass__kg_" < 6000;
```

Python

* sqlite:///SPACEXDB.db
Done.

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

Boosters Carried Maximum Payload

- Boosters carrying 15600kg
  - 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", "PAYLOAD_MASS__KG_"
FROM SPACEXTBL
WHERE "PAYLOAD_MASS__KG_" = (
    SELECT MAX("PAYLOAD_MASS__KG_")
    FROM SPACEXTBL)
```

Python

```
 * sqlite:///SPACEXDB.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

2017 Launch Records

```
%%sql
SELECT
    CASE
        WHEN substr(Date, 4, 2) = '01' THEN 'January'
        WHEN substr(Date, 4, 2) = '02' THEN 'February'
        WHEN substr(Date, 4, 2) = '03' THEN 'March'
        WHEN substr(Date, 4, 2) = '04' THEN 'April'
        WHEN substr(Date, 4, 2) = '05' THEN 'May'
        WHEN substr(Date, 4, 2) = '06' THEN 'June'
        WHEN substr(Date, 4, 2) = '07' THEN 'July'
        WHEN substr(Date, 4, 2) = '08' THEN 'August'
        WHEN substr(Date, 4, 2) = '09' THEN 'September'
        WHEN substr(Date, 4, 2) = '10' THEN 'October'
        WHEN substr(Date, 4, 2) = '11' THEN 'November'
        WHEN substr(Date, 4, 2) = '12' THEN 'December'
    END AS Month,
    Landing_Outcome,
    booster_version,
    launch_site
FROM SPACEXTBL
WHERE substr(Date, 7, 4) = '2017'
    AND landing_outcome LIKE '%Ground%'
    AND Mission_Outcome = 'Success';
```

Month	Landing_Outcome	Booster_Version	Launch_Site
February	Success (ground pad)	F9 FT B1031.1	KSC LC-39A
May	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
June	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
August	Success (ground pad)	F9 B4 B1039.1	KSC LC-39A
September	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A
December	Success (ground pad)	F9 FT B1035.2	CCAFS SLC-40

Showing Month, Outcome, Booster Version and corresponding Launch Site for the year 2017

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

```
%%sql
SELECT "Landing_Outcome", "Date"
FROM SPACEXTBL
WHERE Landing_Outcome LIKE '%Success%'
    AND substr(Date, 7, 4) || '-' || substr(Date, 4, 2) || '-' || substr(Date, 1, 2) BETWEEN '2010-06-04' AND '2017-03-20'
```

Landing_Outcome	Date
Success (ground pad)	22-12-2015
Success (drone ship)	08-04-2016
Success (drone ship)	06-05-2016
Success (drone ship)	27-05-2016
Success (ground pad)	18-07-2016
Success (drone ship)	14-08-2016
Success (drone ship)	14-01-2017
Success (ground pad)	19-02-2017

Section 3

Launch Sites Proximities Analysis

SpaceX Launch Sites in America

The launch facilities are located near the **equator**, where the maximum **centrifugal force** resulting from the Earth's rotation **supports** the **rocket launch** and helps to **reduce costs** for overcoming the earth's gravity.

Lompoc

VAFB SLC-4E

Orlando

KSC LC-39A

CCAFS SLC 40

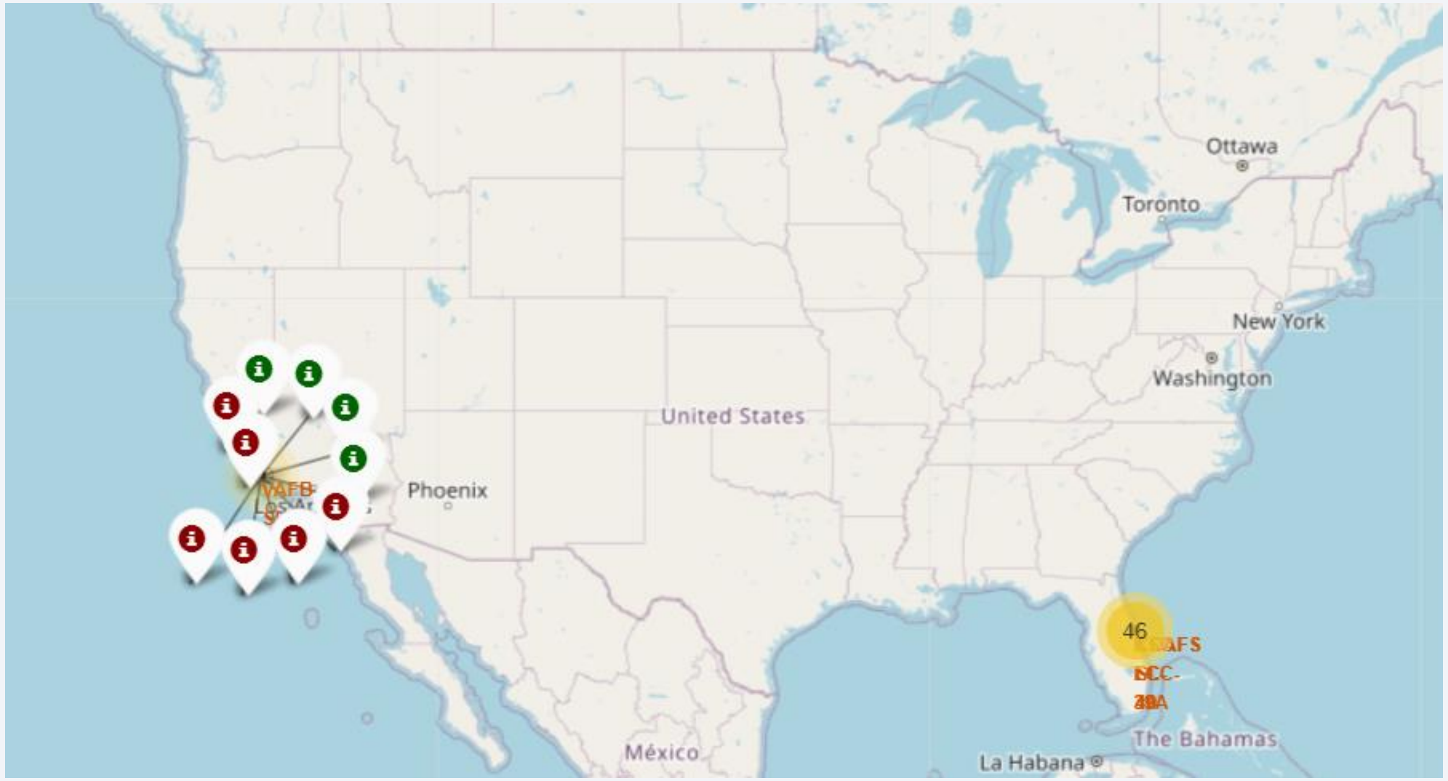
Launch Outcomes

Displayed for each launchsite

Marks:

- Successful Landing
- Unsuccessful Landing

CCAFS SLC-40
Success Rate:
43%



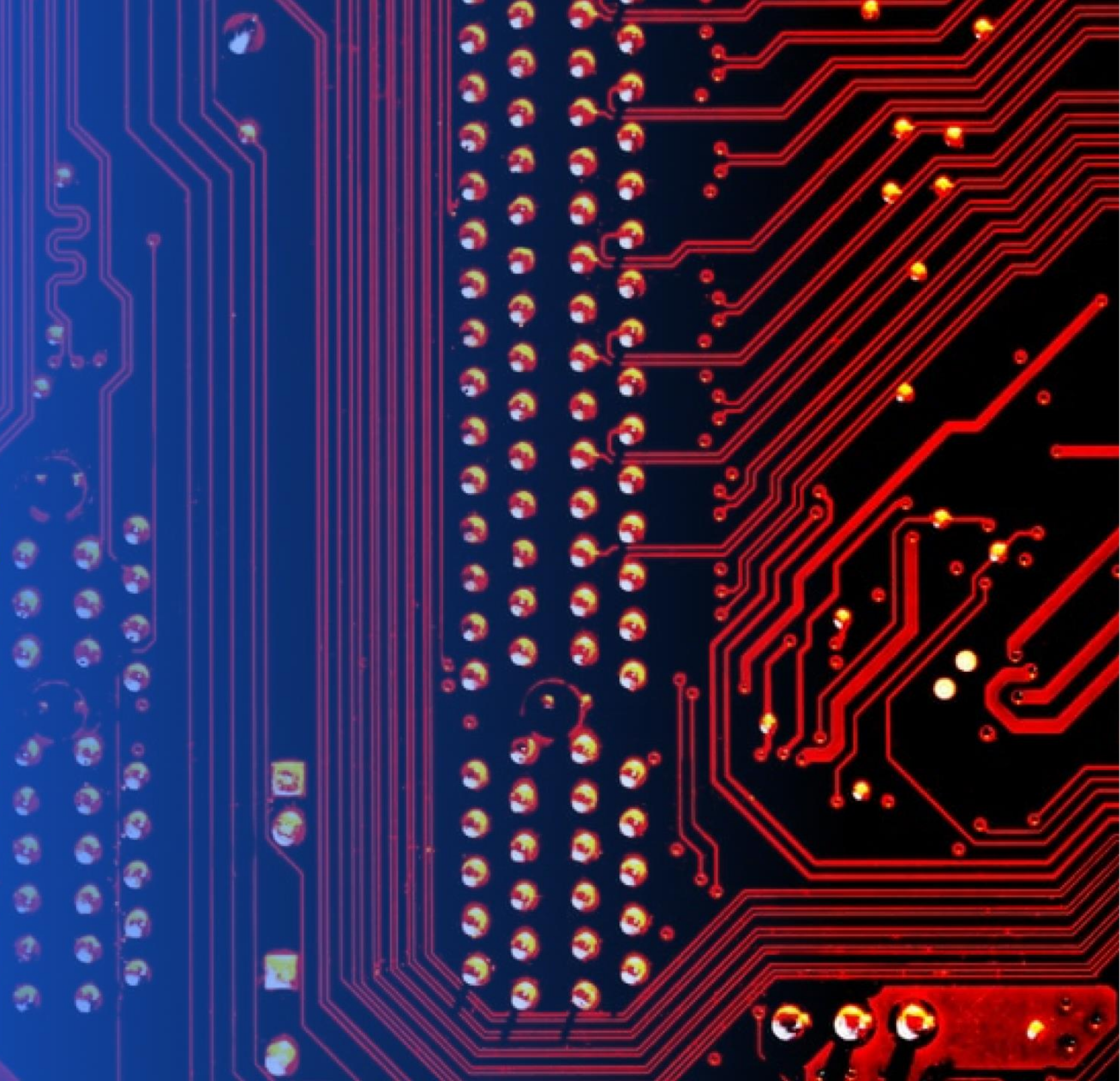
Distamce to Proximities

Next Coastline	Next Highway	Next Railway	Next City
0,93 km	**0,65 km**	**0,95 km**	**25,8 km**

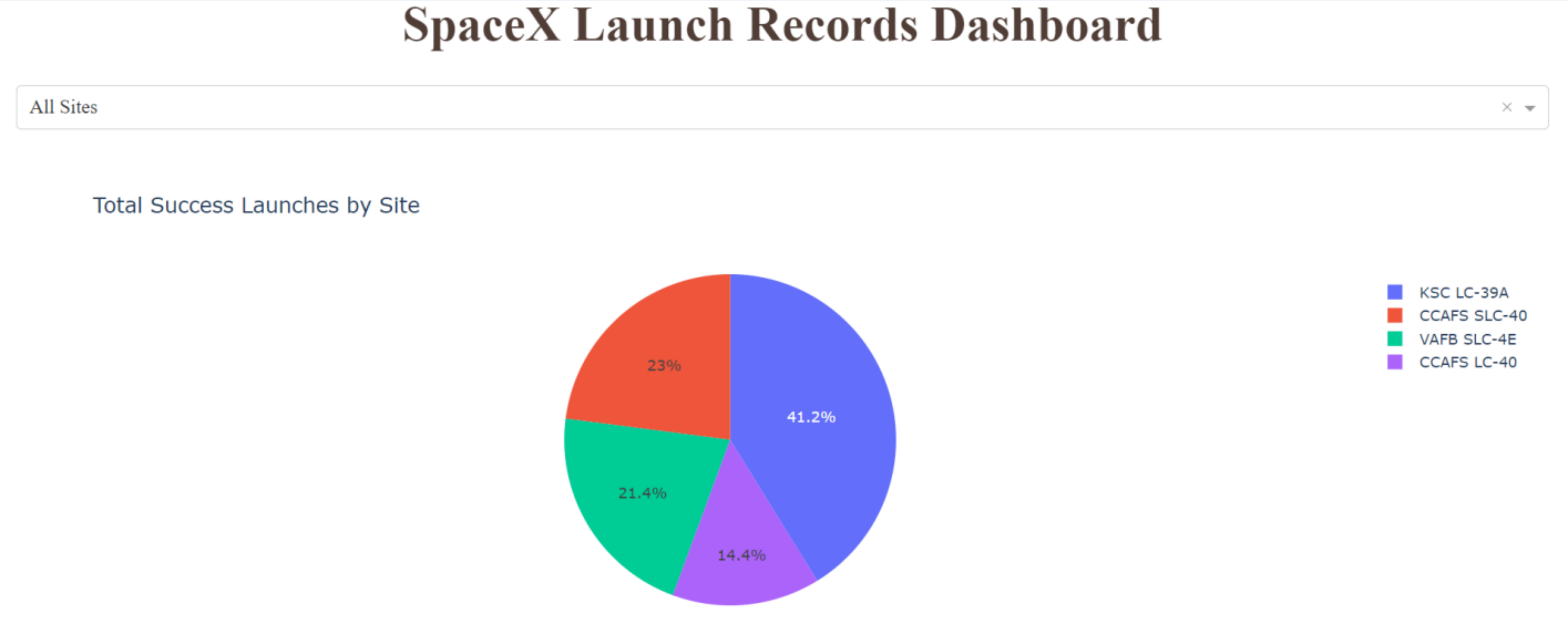
Near Cost in case of crash landing of Vehicles in the ocean
Near Public transportation to bring Workers to the Facilities

Section 4

Build a Dashboard with Plotly Dash



<Dashboard Screenshot 1>



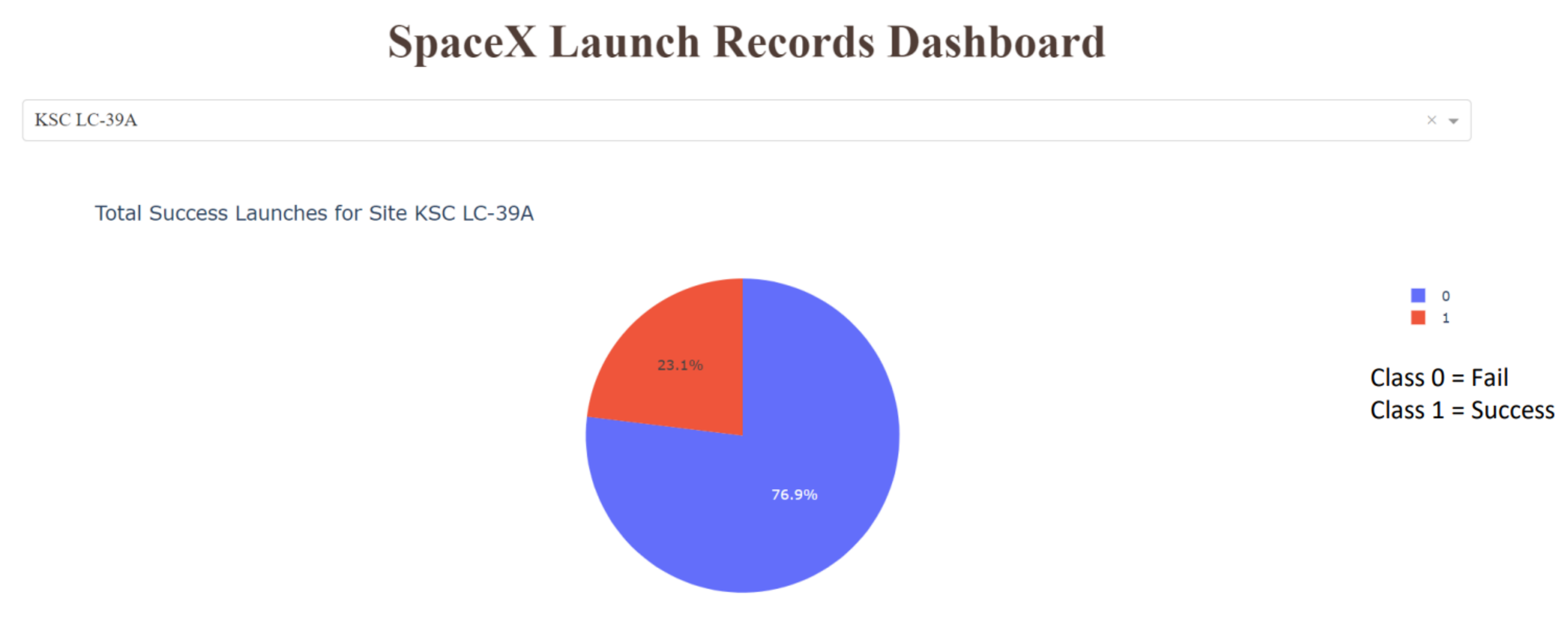
41,2 %

of all successful launches

located in

KSC LC-39A

<Dashboard Screenshot 2>



KSC LC-39A
highest success rate amongst all sites
10 out of 13 launches

76,9 %

Success vs Payload

Payloads between **2000-5000** kg have the highest success rate

Section 5

Predictive Analysis (Classification)

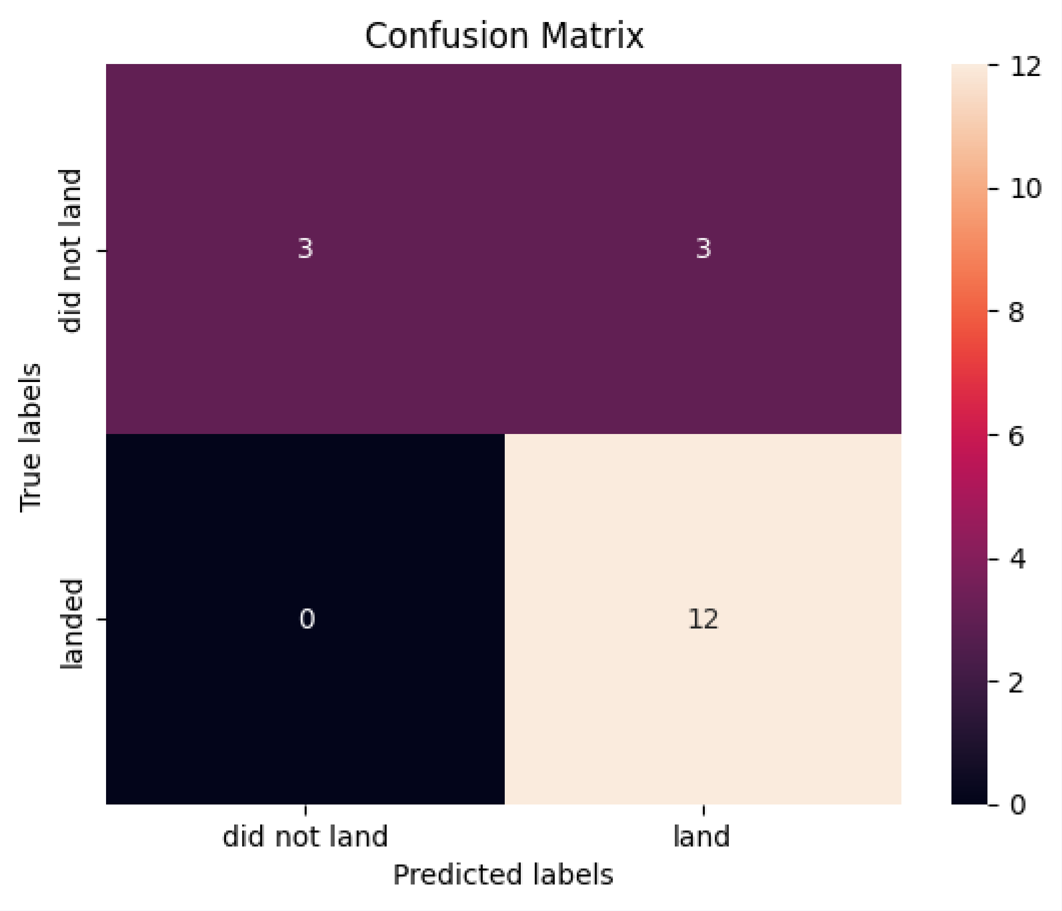
Classification Accuracy

	Model	Test Accuracy	Test Precision	Test Recall	Test F1 Score
0	KNN_cv	0.833333	0.866667	0.833333	0.814815
1	tree_cv	0.833333	0.844156	0.833333	0.836120
2	lr_cv	0.833333	0.866667	0.833333	0.814815
3	svm_cv	0.833333	0.866667	0.833333	0.814815
4	rf_cv	0.833333	0.866667	0.833333	0.814815

All models perform at the same level decision tree performed a bit worse in case of Presison

Confusion Matrix - KNN

- 12 true positive
- 3 true negative
- 3 false positive
- 0 false negativ



Conclusions

- Success rate increase over time
- Orbits ES-L1, GEO, HEO & SSO have 100% success rate
- Launch site KSC LC 39 have the highest success rate
- Increasing success rates with increasing payload mass
- All launch sites are located near the equator and to the ocean
- Model performing similar with good predictability

Thank you!