

Winning Space Race with Data Science

José Daniel Gutiérrez Blázquez
Jun 2th, 2024



Outline

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

Executive Summary

- SpaceX has the business problem of estimating the cost of a Falcon 9 launch, which is determined if the first stage of Falcon 9 will land.
- The aim of this analysis is characterized by the principal features of one successful landing considering the historical data of landing outcomes.
- The Data Collection was made with SpaceX API, subsequently, an exploratory data analysis was driven with pandas, SQL, and Folium for the identification of principal features.
- The result of this analysis was the selection of a suitable classification model that has the capacity to estimate if a launch will be successful considering the orbit, launch site, landing pad, and serial of the Falcon 9.

Introduction

- Currently, the need has arisen to provide affordable space traveling.
- In this scenario, the world known company SpaceX has the business problem of estimating the cost of a Falcon 9 launch, which is determined if the first stage of Falcon 9 will land.

This aim of this analysis it's characterized the principal features of one successful landing considering the historical data of landing outcomes.

The present study is organized as follows. In the first section, we describe the Data Collection process. The subsequent section contains the exploratory data analysis made with pandas, SQL, and Folium for the ~~Identification, classification, and principal~~ features. In the classification model that has the capacity to estimate if a launch will be successful is presented.

Section 1

Methodology

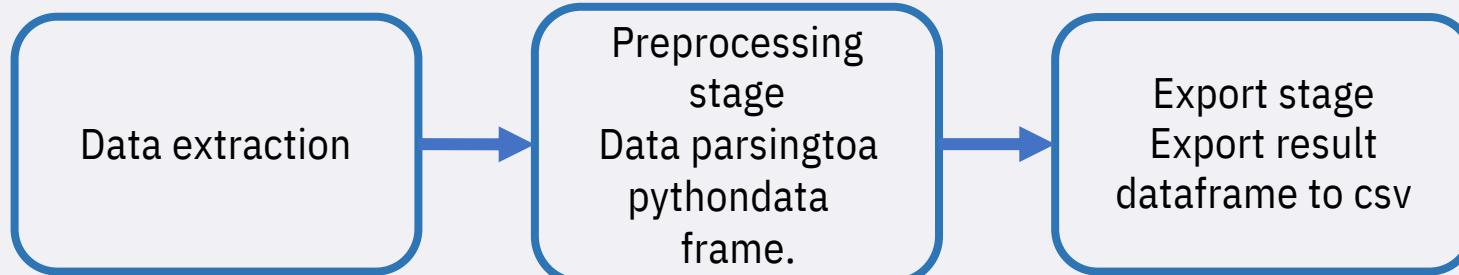
Methodology

Executive Summary

- Data collection methodology:
 - Primary data source: SpaceX API.
 - Secondary data source: data available on Wikipedia.
- Perform datawrangling
 - SpaceX API data has a relatively easy data wrangling, only filtering Falcon 9 data and missing data quick handling. Data available on Wikipedia required HTML syntaxis cleaning, and a HTML parse to Python data frame.
- Perform exploratory data analysis (EDA) using visualization andSQL
- Perform interactive visual analytics using Folium andPlotlyDash
- Perform predictive analysis using classification models
 - A standard mythology for fitting and testing the classification models were carried out.

Data Collection

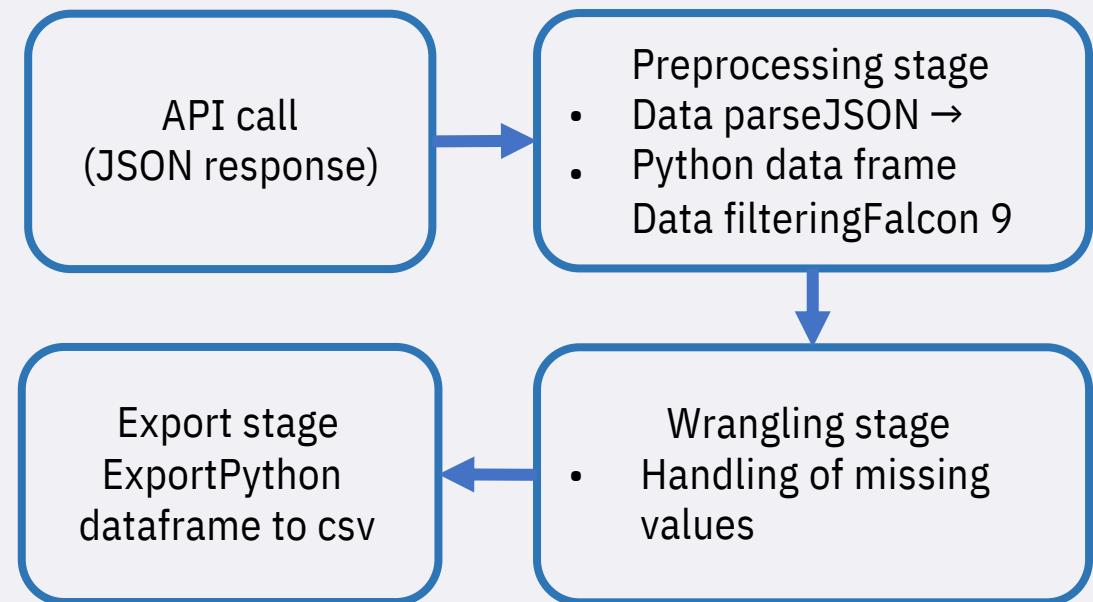
- The SpaceX API gave the primary data resource for the analysis, but an initial data collection process was made using Falcon 9 data available on Wikipedia.
- Although the data collection processes were significantly different, a general schema can be considered, which is described in the following flowchart.



Data Collection –SpaceX API

- The outcome of the data collection was a data frame with critical features that allowed further analysis which included date, longitude, latitude, launch site, and outcome.
- The notebook with the SpaceX API data collection process can be found[here](#).

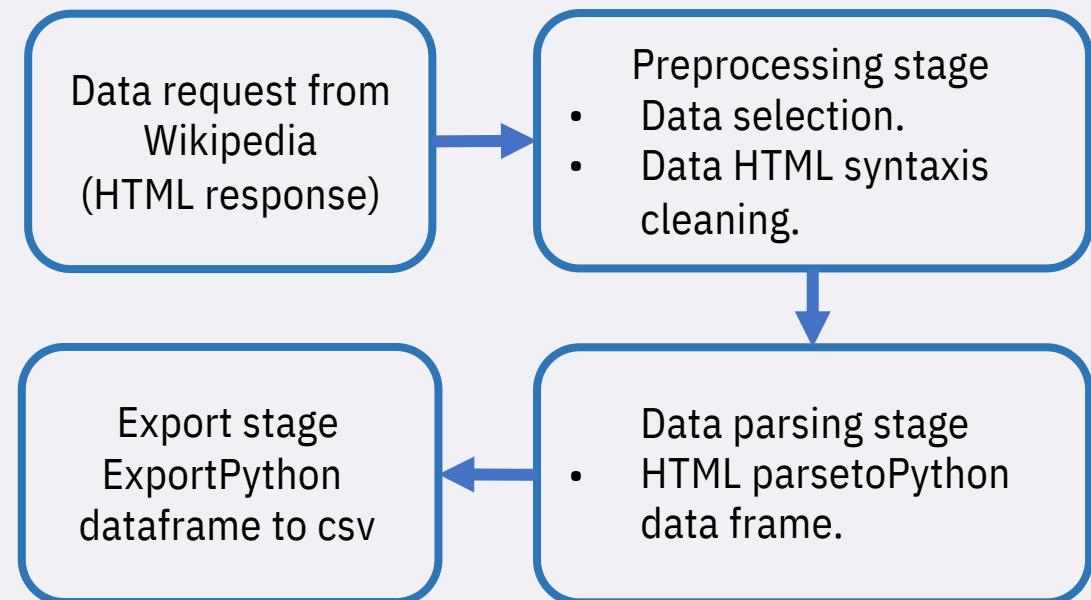
SpaceX API Data Collection general schema



Data Collection -Scraping

- The recollecting data process from Wikipedia required several lines of code for a correct noise-cleaning.
- The notebook with the web scraping data process can be found here ([Wikipedia web scraping](#)).

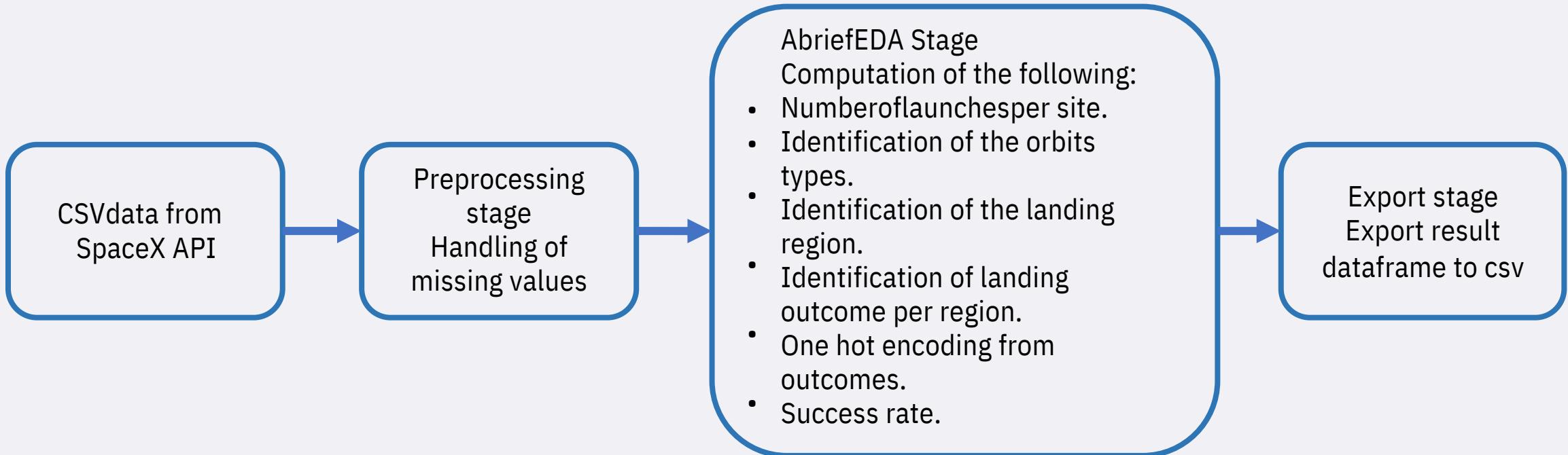
Web scraping Data Collection general schema



Data Wrangling

- Given the csv data from the SpaceX API data collection, analyses to launch facilities, orbits, and mission outcomes features were made.
- The aim goal of the analysis was to provide accurate labeling of the outcomes in a 1/0 fashion, where 1 means the booster successfully landed, and 0 meaning unsuccessful landing.
- Also, from the analysis was possible to identify the number of launches per site, the name and number of orbits in which the launch was made, and the types of possible outcomes for a Falcon 9 mission.
- A 66% success rate was calculated as well.
- The notebook with the data wrangling process can be found [here](#).

Data Wrangling general schema



EDA with Data Visualization

- A total of seven charts were plotted as part of the Exploratory Data Analysis: five scatter plots, one line plot, and one bar plot.
- Generally speaking, the graph type selected for each feature relationship analysis was directly related to the necessity of understanding the connection among three or two features.
- The scatter plots give us an insight of the relationship among the landing outcome (1/0 feature) and the duos (Payload Mass, Flight Number), (Flight Number, Launch Site), (Payload Mass, Launch Site), (Flight Number, Orbit), and (Payload Mass, Orbit).
- The bar plot lets us analyze the relationship between success rate and orbit type. Lastly, a line chart shows the relationship between year and average success rate.
- The EDA with Data Visualization notebook can be found [here](#).

EDA withSQL

- Further Exploratory Data Analyses were made by means of SQL-queries in DB2 to a Falcon 9 data set stored in IBMDB2.
 - The names of the booster_versions which have carried the maximum payload mass.
 - The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015.
 - The ranking of the 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 queries results show us the following information.
 - The names of the unique launch sites in the space mission.
 - Showed 5 records where launch sites begin with the string 'CCA'.
 - The total payload mass carried by boosters launched by NASA (CRS).
 - The average payload mass carried by booster version F9 v1.1.
 - The date when the first successful landing outcome in ground pad was achieved.
 - 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.
 -
- The EDA with SQL notebook can be found [here](#).

Build an Interactive Map with Folium

- The main goal plotting Falcon9data with Folium was identify launch sites, success and failed launches for site, and calculate the distances between launch sites and its proximities.
- A Folium-object marker allowed us to identify launch sites, a Folium-object circle made it possible to see the count of successful and failed launches in situ, and lines show us a visual representation between the launch sites and proximities as railways, cities, etc.
- The Folium notebook can be found [here](#).

Build a Dashboard with PlotlyDash

- A plotlydashboard with two interactivechartsweremade: a pie chart and a scatter plot gave us insight between the duos (landing outcome, site) and (payload, launch success), respectively.
- The chart interactivity was achieved through a site dropdown list and a payload range slider, allowing us to obtain data insights quickly and efficiently.
- The Dash notebook can be found [here](#).

Predictive Analysis (Classification)

- SVM, classification tree, logistic regression and KNN were tested in order to find the best 1/0-outcomes feature classifier.
- A standard mythology for fitting and testing the classification models were carried out. The detailed procedure is explained in the following schema.
- The Predictive Analysis notebook can be found [here](#).

ImportFalcon 9
cvs data

Preprocessing stage 1

- Target feature: 1/0-outcome → Y
- Predictive features: one hot encoding of [orbit, launch site, landing pad and serial] → X

Preprocessing stage 2

- Standardization of X.
- Splitting X, and Y in the training set and testing set.

Models fitting and testing (SVM/ CT / LR/ KNN)

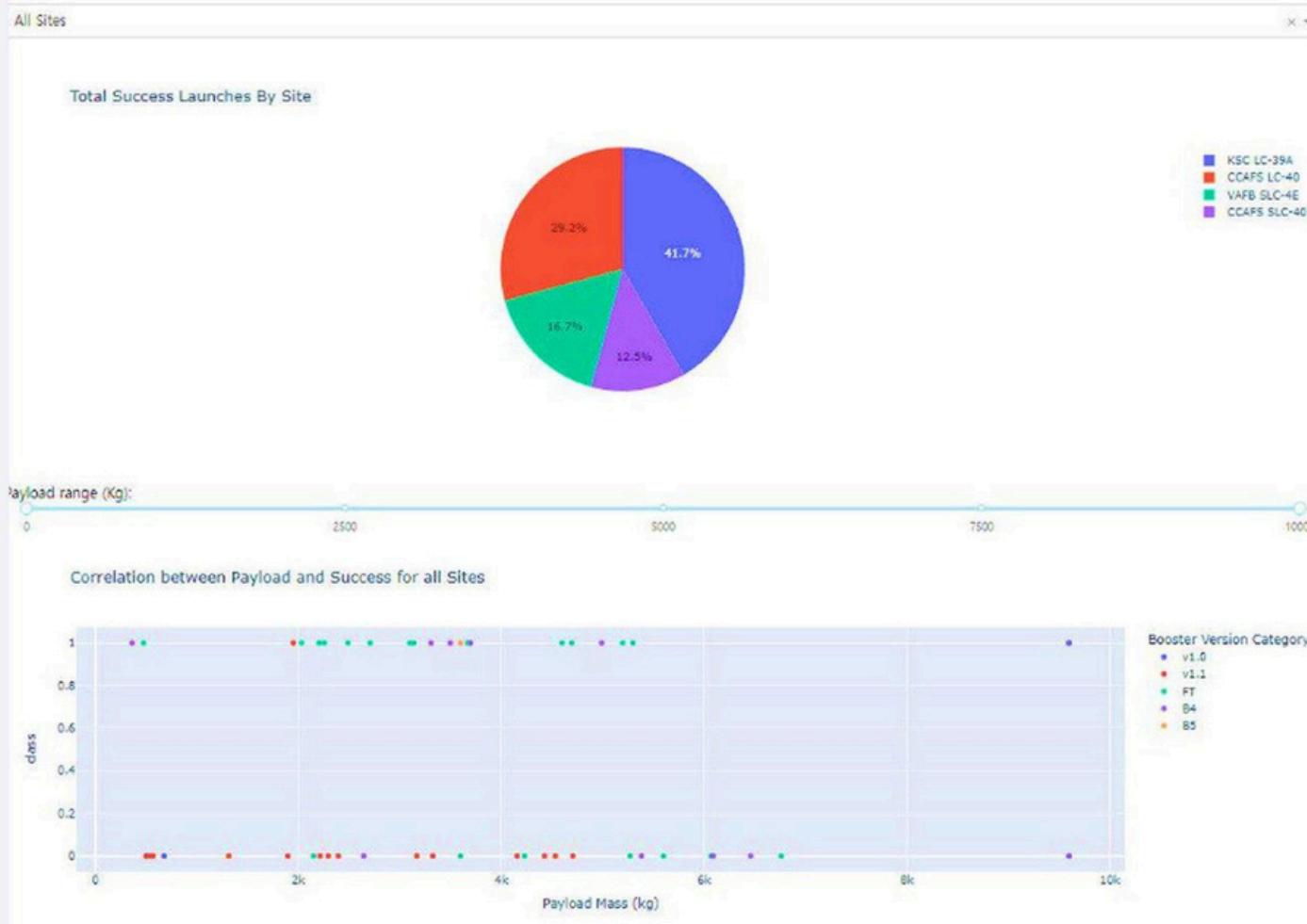
1. Model GridSearch.
2. Model fitting.
3. Finding best hyperparameters.
4. Computation of accuracy model.
5. Model Prediction.
6. Confusion matrix analysis.

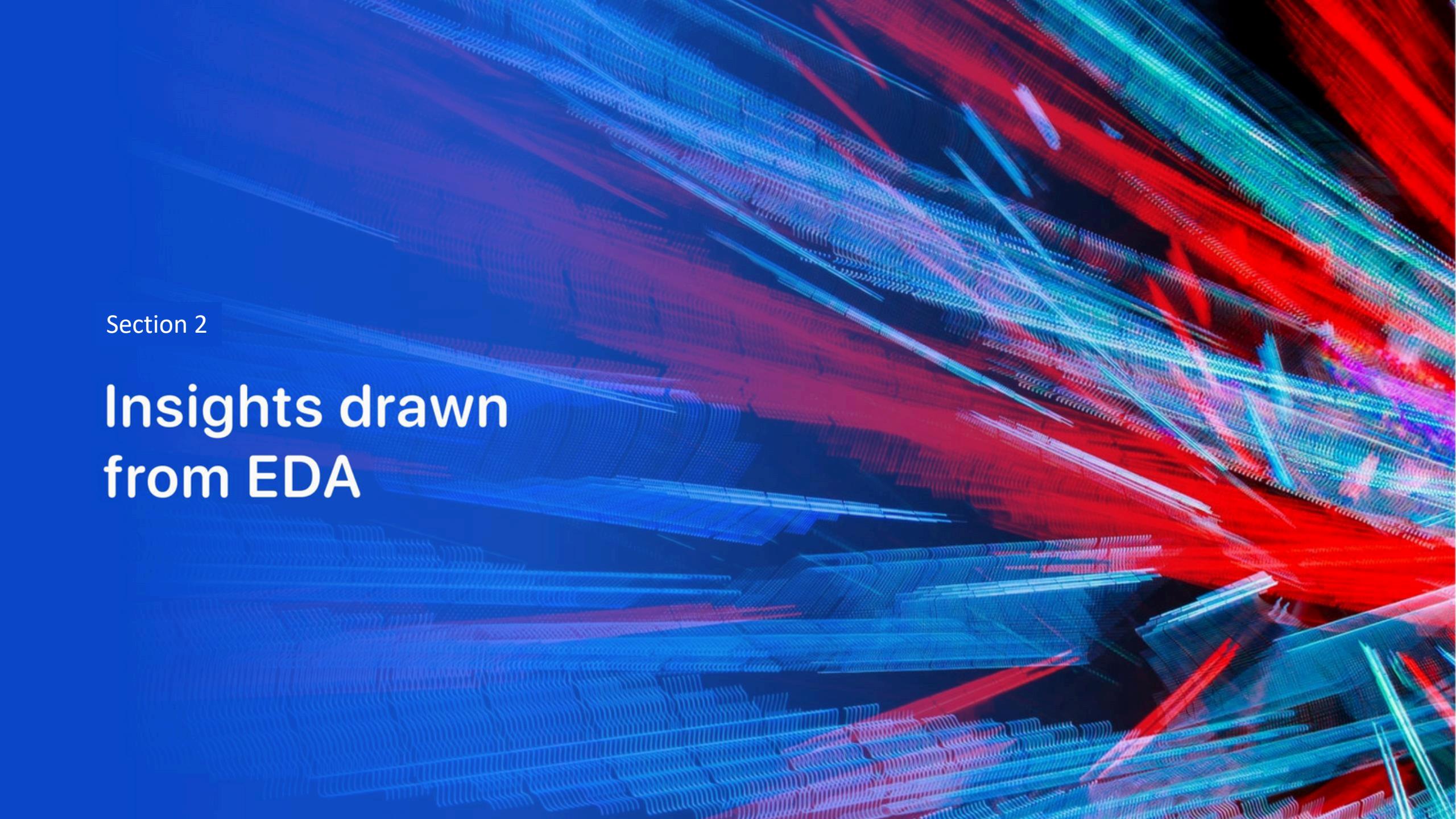
Best performance model selection

Results

- Broadly speaking, the results of the exploratory data analysis shows the following:
 - There's a yearly increasing trend in the success landing outcome; There's a 66% general success rate;
 - Different launch sites have different success rates;
 - VAFB-SLC launch site hasn't launched rockets for heavy payload mass;
 - The orbits ES-L1, GEO, HEO, and SSO have an average success rate of 100%; excluding SO landing sites with a 0% of success rate, generally speaking, all landing sites have an average success greater than 50%.
- The best performance models classifier were the logistic regression, SVM, and KNN, each with a score of 83%.

Results: interactive analytics demo

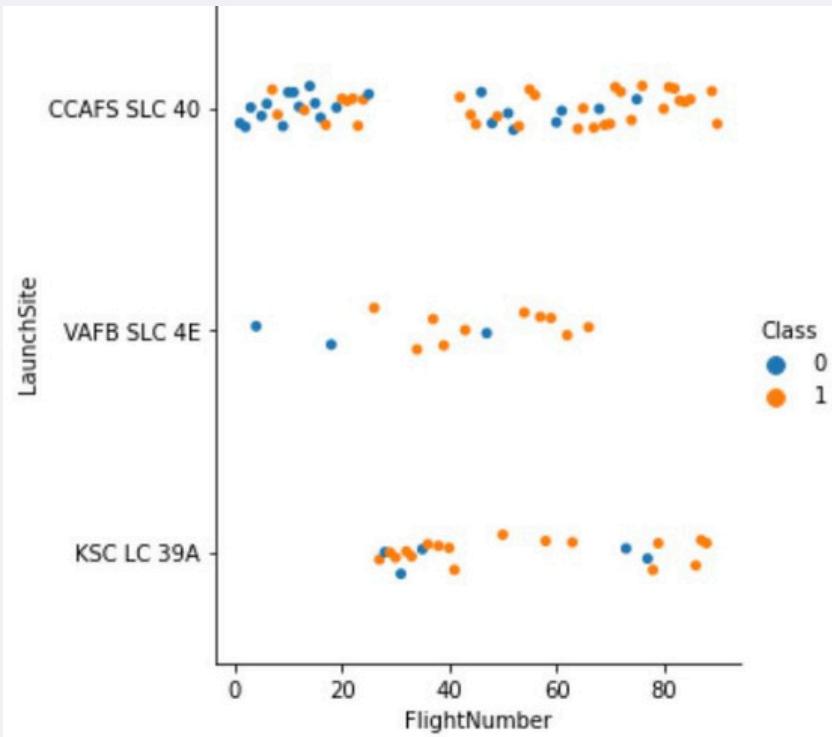


The background of the slide features a complex, abstract digital visualization. It consists of numerous small, glowing particles that form a dense, three-dimensional grid-like structure. The colors of these particles are primarily shades of blue, red, and green, creating a vibrant, futuristic, and dynamic appearance. The grid is not uniform; it has various depths and angles, giving it a sense of depth and movement. Some particles are more prominent than others, appearing as bright streaks or lines that intersect to create a complex web of light.

Section 2

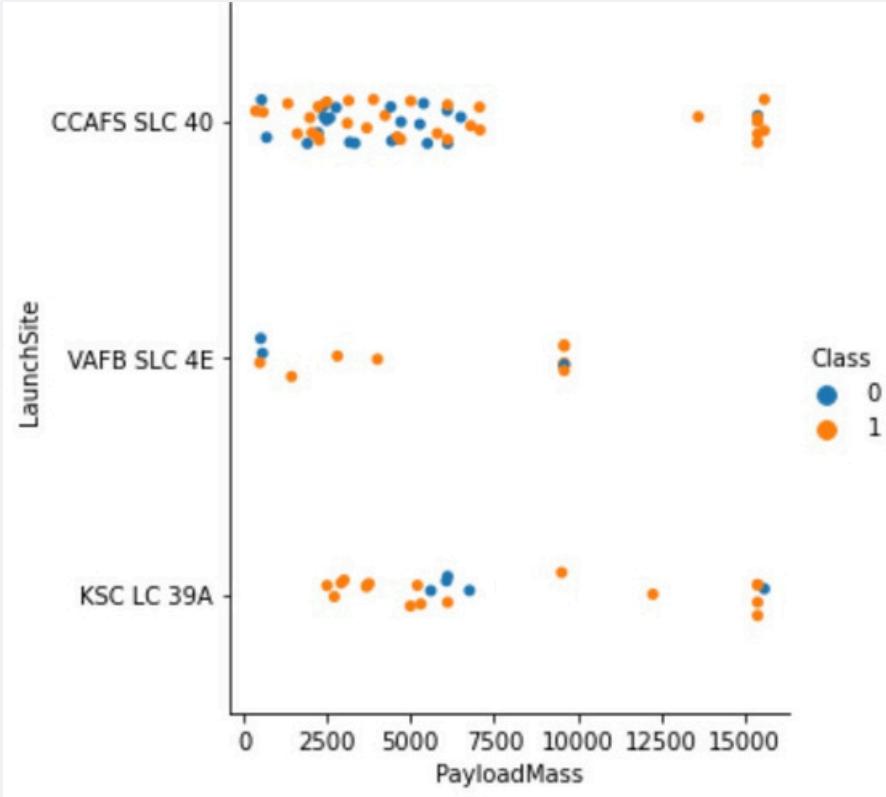
Insights drawn from EDA

Flight Number vs. LaunchSite



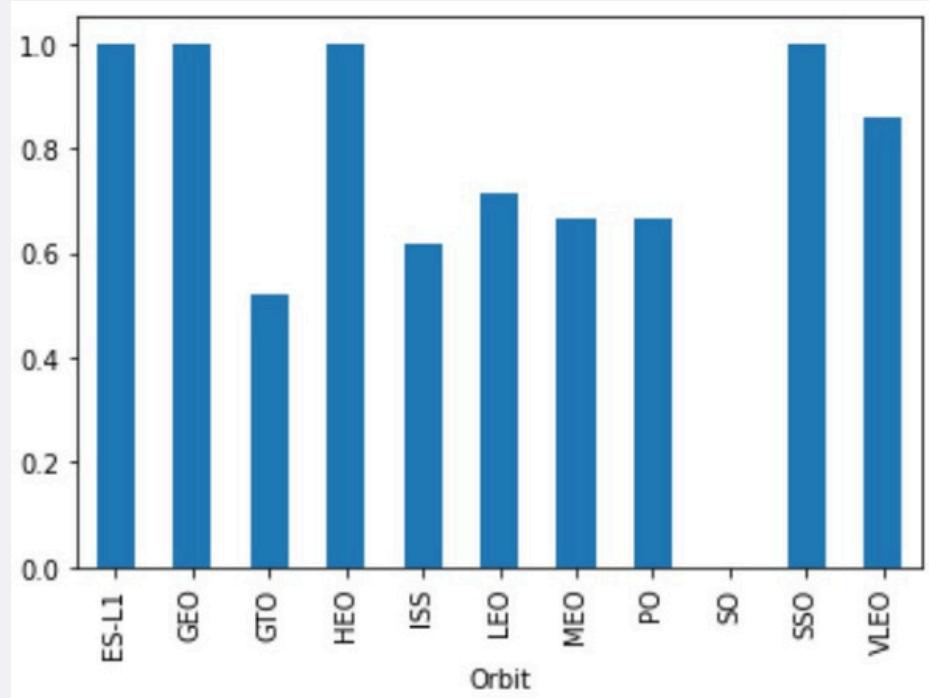
- In the case of the Launch Site CCAFS SLC 40 it's possible to note a rise in the number of successful landings corresponding with an increase in the flightnumber.
- In contrast, we can notice a general failure outcome in the Launch Sites VAFB SLC 4E and KSC LC 39A.

Payload vs. Launch Site



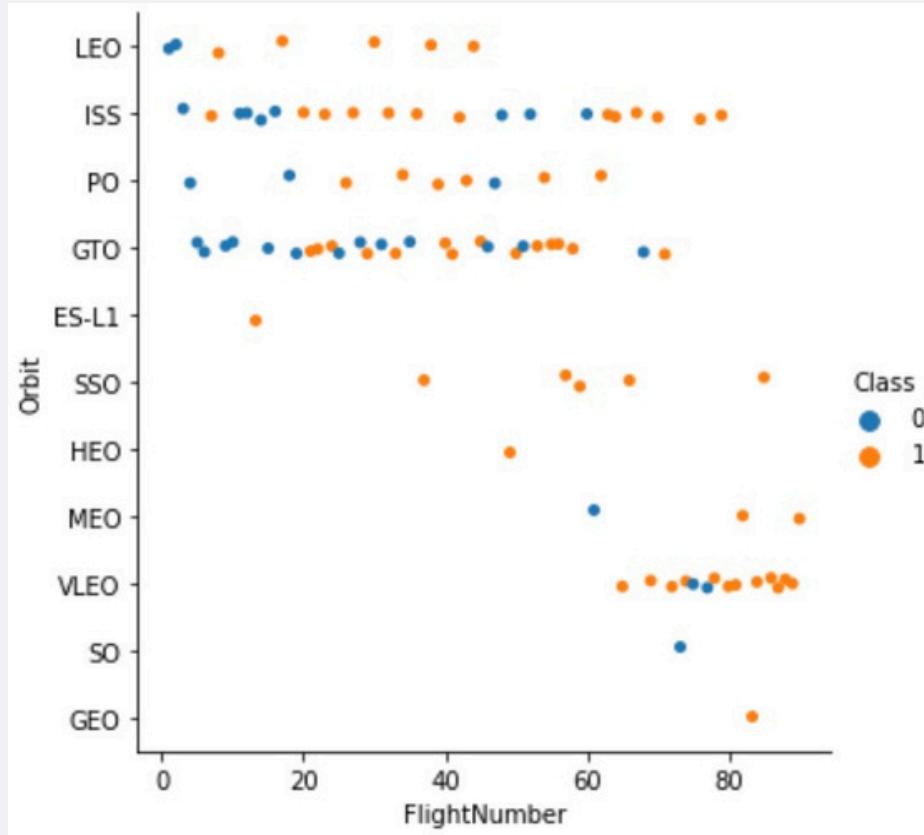
- For the VAFB-SLC 4E launch site there are no rockets launched for heavy pay load mass (greater than 10000).

Success Rate vs. Orbit Type



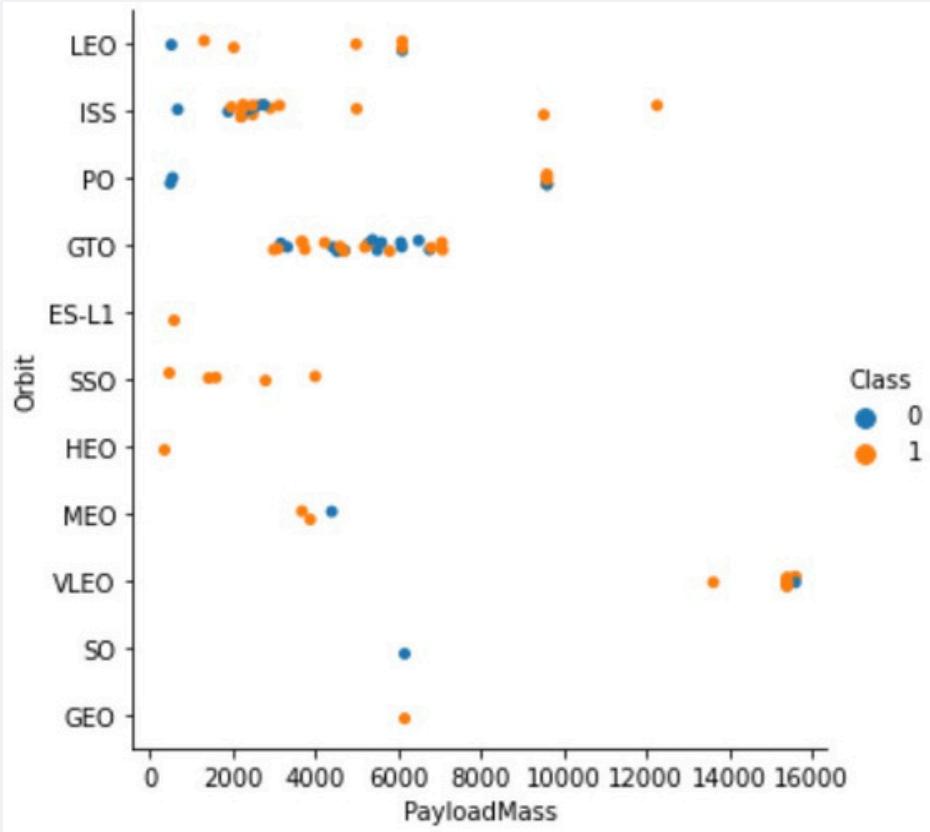
- The orbits ES-L1, GEO, HEO, and SSO have an average success rate of 100%; excluding SO landing sites with a 0% of success rate, generally speaking, all landing sites have an average success greater than 50%.

Flight Number vs. Orbit Type



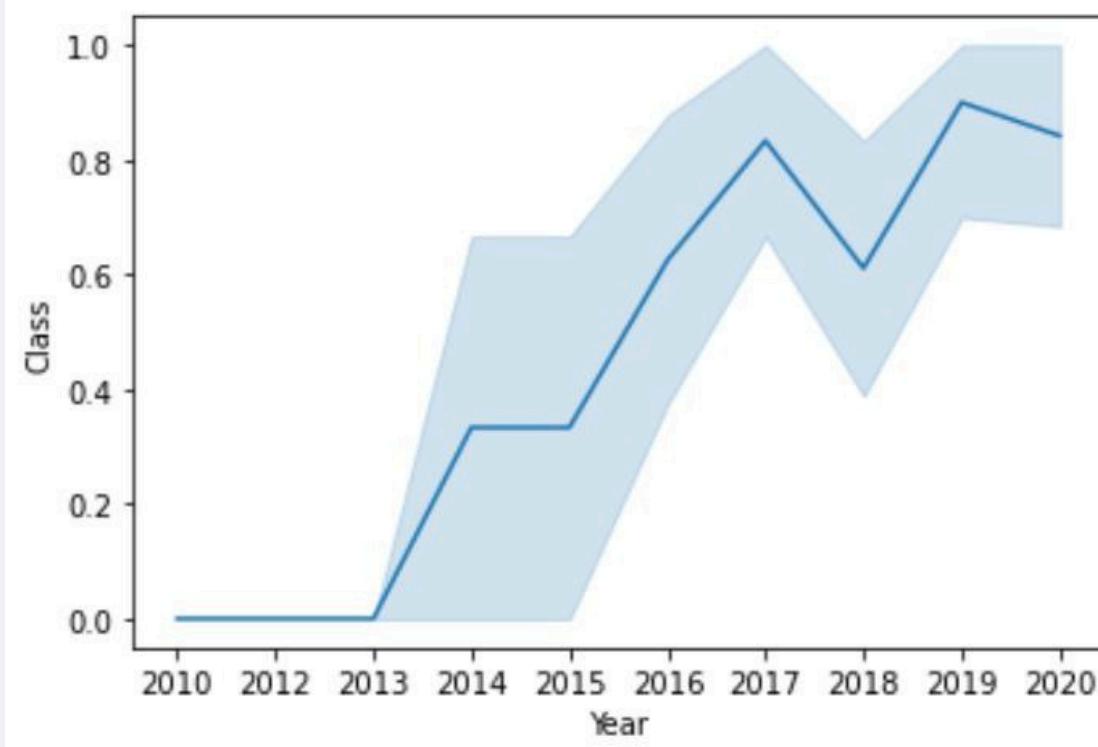
- The LEO orbit the Success appears related to the number of flights.
- There seems to be no relationship between flight number when in GTO orbit.

Payload vs. OrbitType



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



- The success rate since 2013 kept increasing till 2020.

All Launch Site Names

```
%%sql
SELECT DISTINCT LAUNCH_SITE
FROM SPACEXTBL

* ibm_db_sa://fvz79191:***@6667d8e9-9d4d-4ccb-ba32-
21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.

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

- The names of the unique launch sites are CCAFS LC-40, CCAFS SLC- 40,KSC LC-39A, VAFB SLC-4E.

Launch Site Names Begin with 'CCA'

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

Python

```
* ibm_db_sa://fvz79191:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
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

- The image shows 5 records where launch sites begin with 'CCA'.

Total Payload Mass

```
%%sql
SELECT SUM(PAYLOAD_MASS_KG )
FROM SPACEXTBL
WHERE CUSTOMER LIKE 'NASA (CRS)'

* ibm_db_sa://fvz79191:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.

1
45596
```

- The total payload carried by boosters from NASA is 45,596.

Average Payload Mass by F9 v1.1

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

* ibm_db_sa://fvz79191:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

1

2928

- The average payload mass carried by booster version F9 v1.1 is 2,928 kg.

First Successful Ground Landing Date

```
%%sql
SELECT MIN(DATE)
FROM SPACEXTBL
WHERE LANDING_OUTCOME LIKE 'Success (ground pad)'
```

```
* ibm_db_sa://fvz79191:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

1

2015-12-22

- The date of the first successful landing outcome on ground pad is 2015-12-22.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
SELECT DISTINCT(BOOSTER_VERSION)
FROM SPACEXTBL
WHERE LANDING_OUTCOME LIKE 'Success (drone ship)' AND (PAYLOAD_MASS_KG BETWEEN 4000 AND 6000)

* ibm_db_sa://fvz79191:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.

booster_version
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1022
F9 FT B1026
```

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are F9 FT B1021.2, F9 FT B1031.2, F9 FT B1022, and F9FTB1026.

Total Number of Successful and Failure Mission Outcomes

```
%>%sql
SELECT MISSION_OUTCOME,COUNT(MISSION_OUTCOME) AS total
FROM SPACEXTBL
GROUP BY MISSION_OUTCOME

* ibm_db_sa://fvz79191:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tg
Done.

mission_outcome  total
Failure (in flight)    1
Success          99
Success (payload status unclear) 1
```

- The total number of successful and failure mission outcomes is the following.
Failure 1, success 99, and success (payload status unclear) 1.

Boosters Carried Maximum Payload

```
%%sql
SELECT DISTINCT BOOSTER_VERSION
FROM SPACEXTBL
WHERE PAYLOAD_MASS_KG = (SELECT MAX(PAYLOAD_MASS_KG) FROM SPACEXTBL)

* ibm_db_sa://fvz79191:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde
Done.

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3
```

- The names of the booster which have carried the maximum payload mass are the following:

F9 B5 B1048.4, F9 B5 B1048.5, F9 B5 B1049.4, F9 B5 B1049.5, F9 B5 B1049.7, F9 B5 B1051.3, F9 B5 B1051.4, F9 B5 B1051.6, F9 B5 B1056.4, F9 B5 B1058.3, F9 B5 B1060.2, and F9 B5 B1060.3.

2015 Launch Records

```
%%sql
SELECT BOOSTER_VERSION, LAUNCH_SITE
FROM SPACEXTBL
WHERE LANDING_OUTCOME LIKE 'Failure (drone ship)' AND YEAR(DATE) = 2015

* ibm_db_sa://fvz79191:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqd
Done.

booster_version      launch_site
F9 v1.1 B1012      CCAFS LC-40
F9 v1.1 B1015      CCAFS LC-40
```

- The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015 are the F9 v1.1 B1012 with launch site CCAFS LC-40 and F9 v1.1 B1015 CCAFS LC-40, respectively.

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

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

* ibm_db_sa://fvz79191:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1o:15
Done.

   landing_outcome  total
   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
```

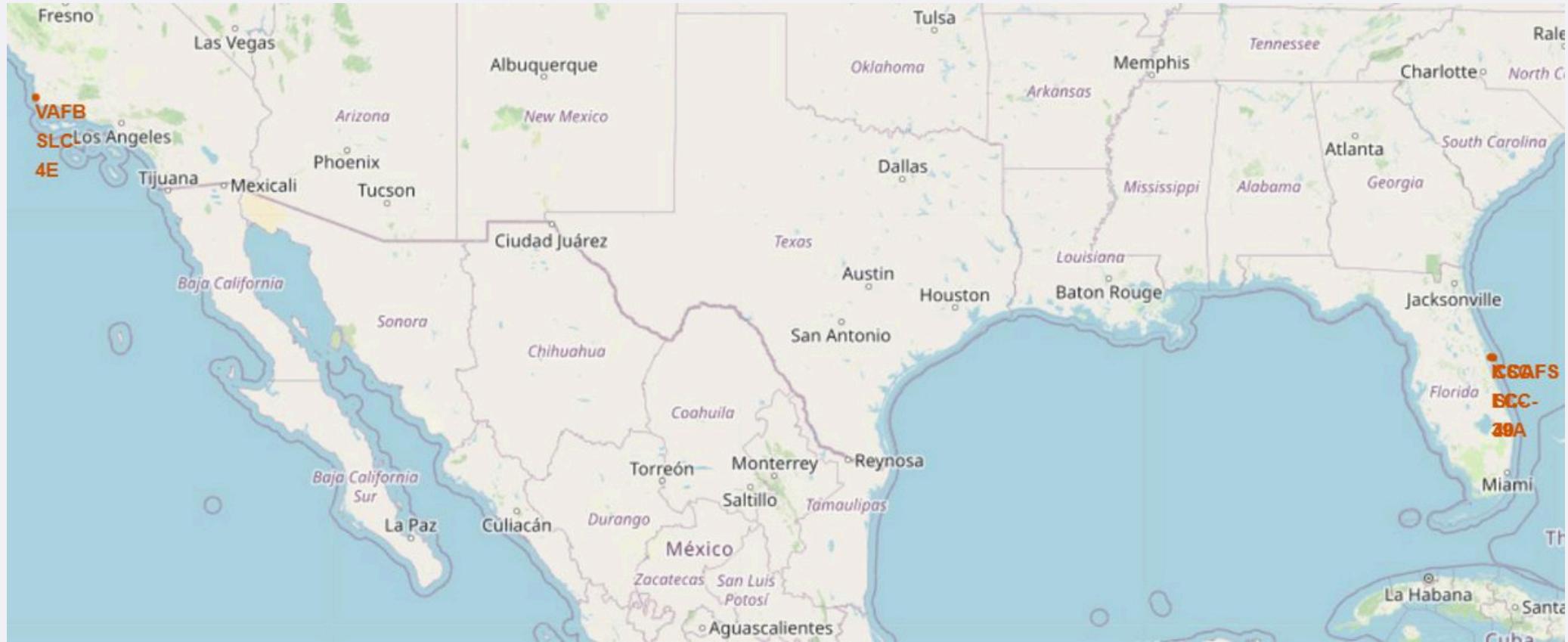
- The rank count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order are shown in the image.

The background of the slide is a photograph taken from space at night. It shows the curvature of the Earth against the dark void of space. City lights are visible as numerous small white and yellow dots, primarily concentrated in the lower right quadrant where the United States and Mexico would be. In the upper left quadrant, the green and blue glow of the aurora borealis is visible in the upper atmosphere.

Section 3

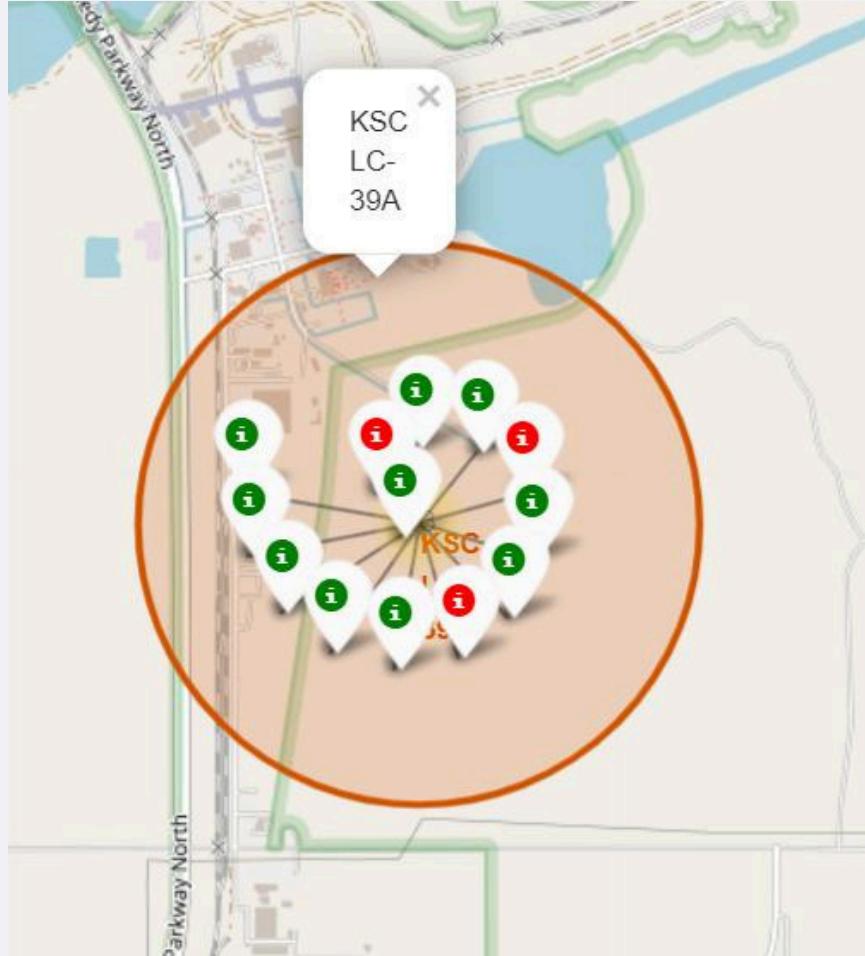
Launch Sites Proximities Analysis

Launch sites in Folium Map



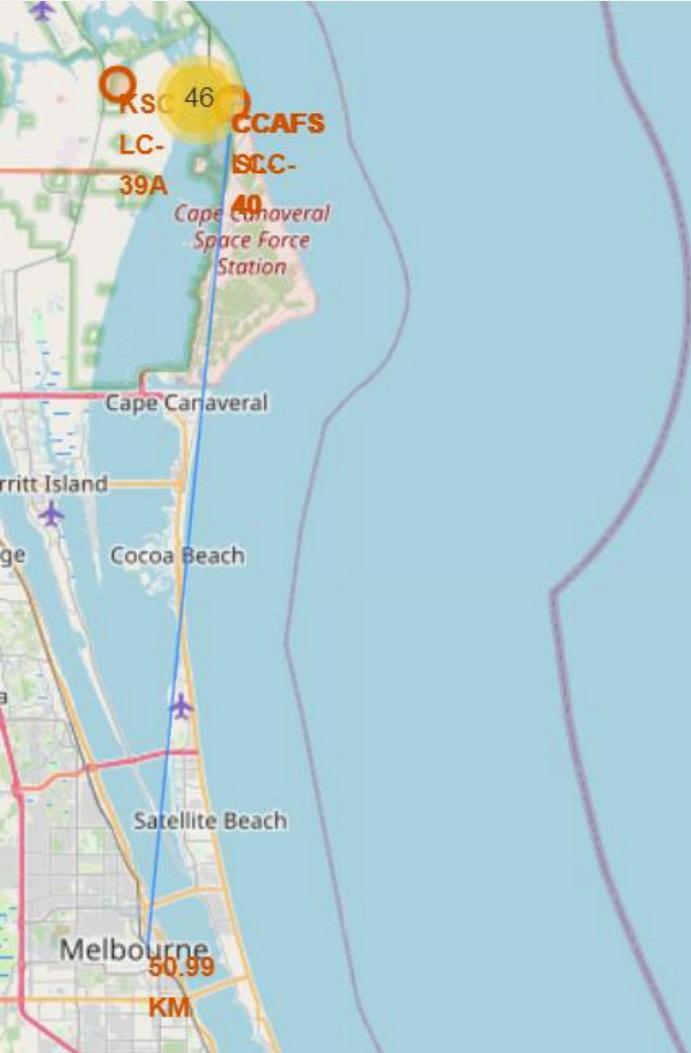
It's possible to notice two cluster launch sites in Florida and California.

Outcomefeatureper launch site with Folium Map



- Folium allows us to visually analyze the outcomefeatureper launchsite.
- For instance, for the launch site KSC LC-39A in Florida, there are 10 successful landingsand3 failedlandings.

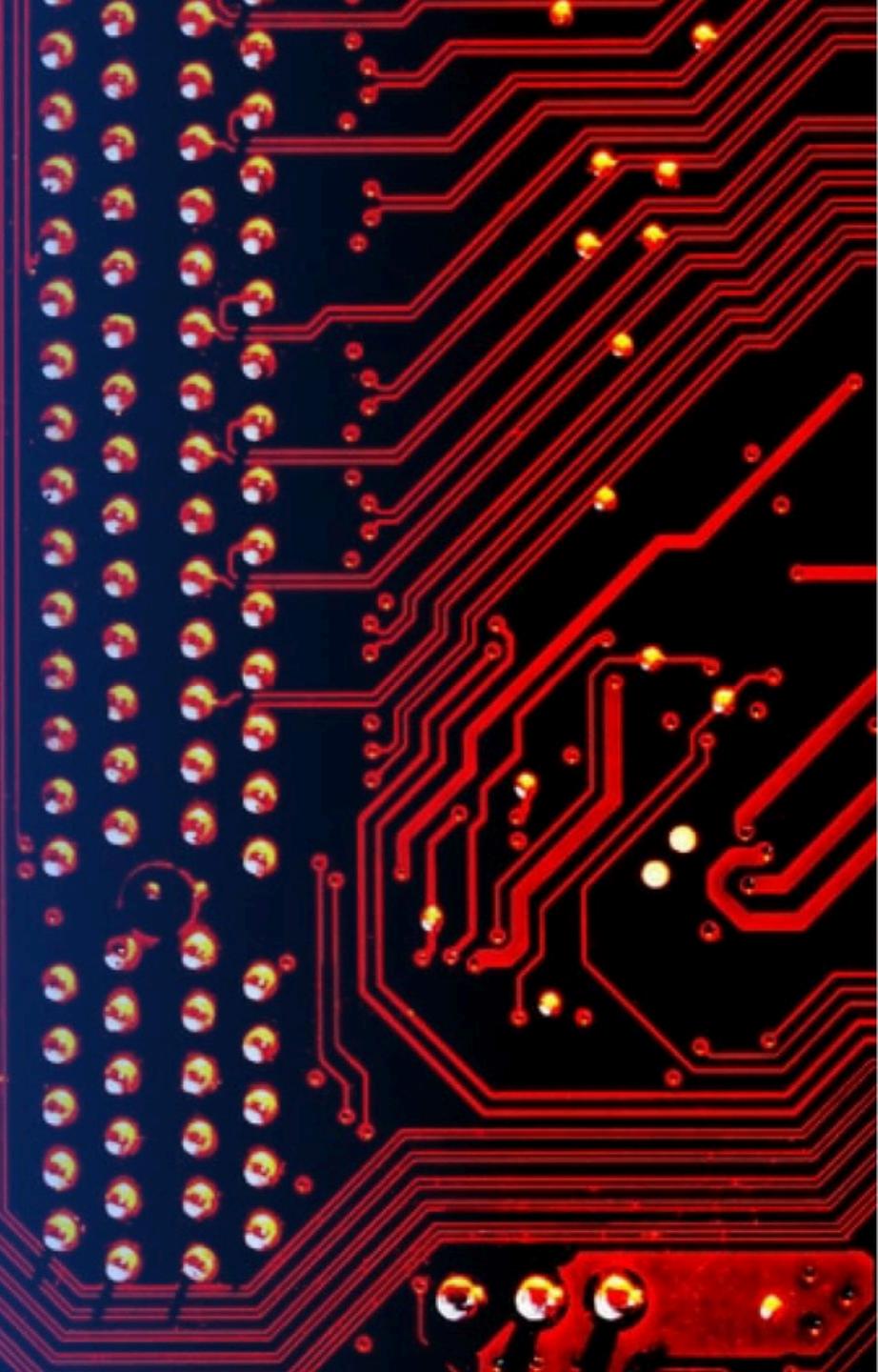
Distance between CCAFSSLC-40 launch site and its proximities



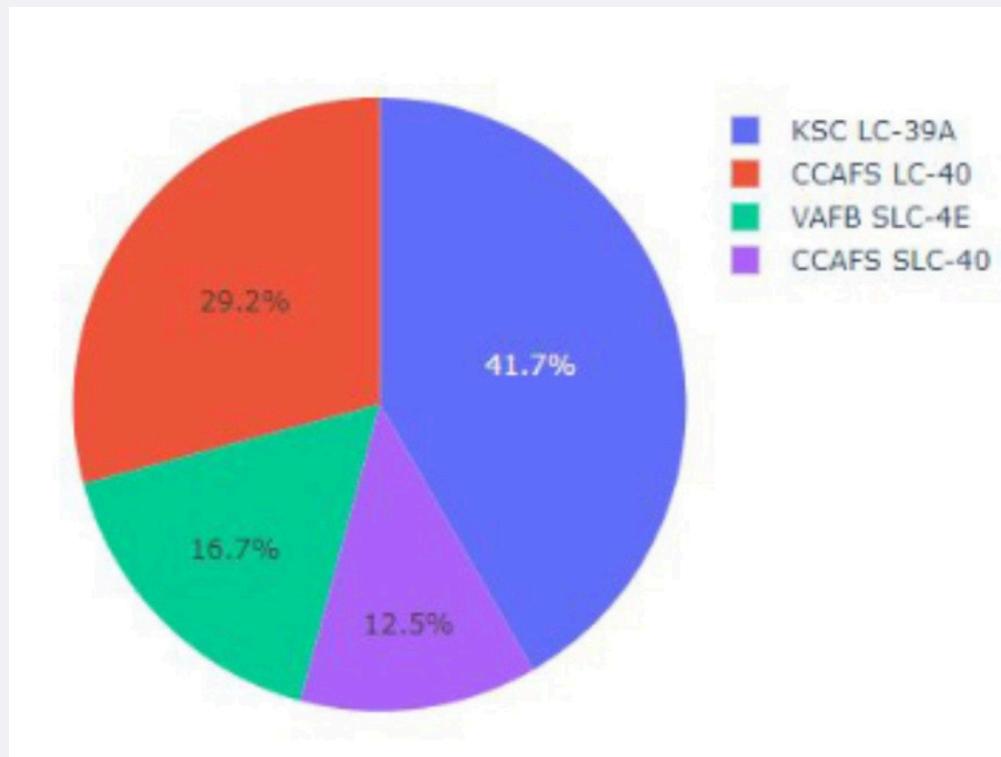
- The ~~distalated~~ ~~proximities~~ between the CCAFSSLC-40 launch site was Melbourne city.
- The distance between CCAFSSLC-40 and Melbourne city is 50.99km.

Section 4

Build a Dashboard with Plotly Dash

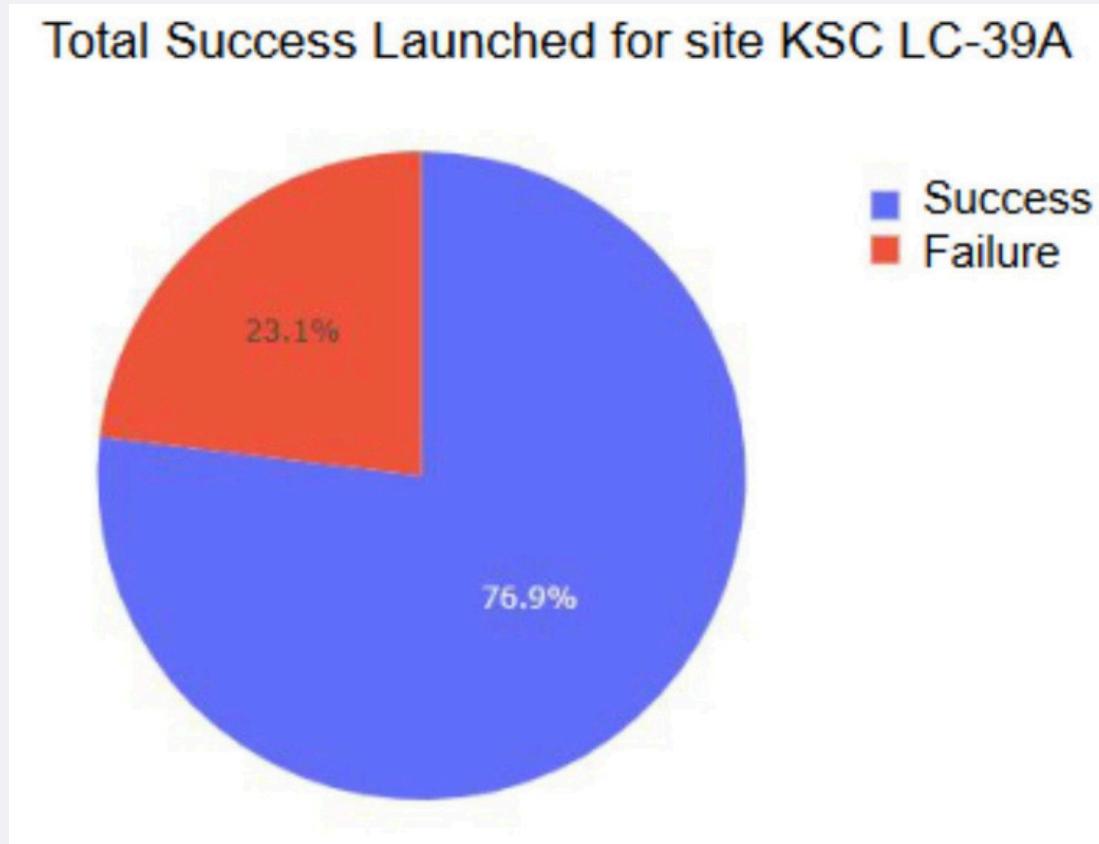


Total Success Launches by site



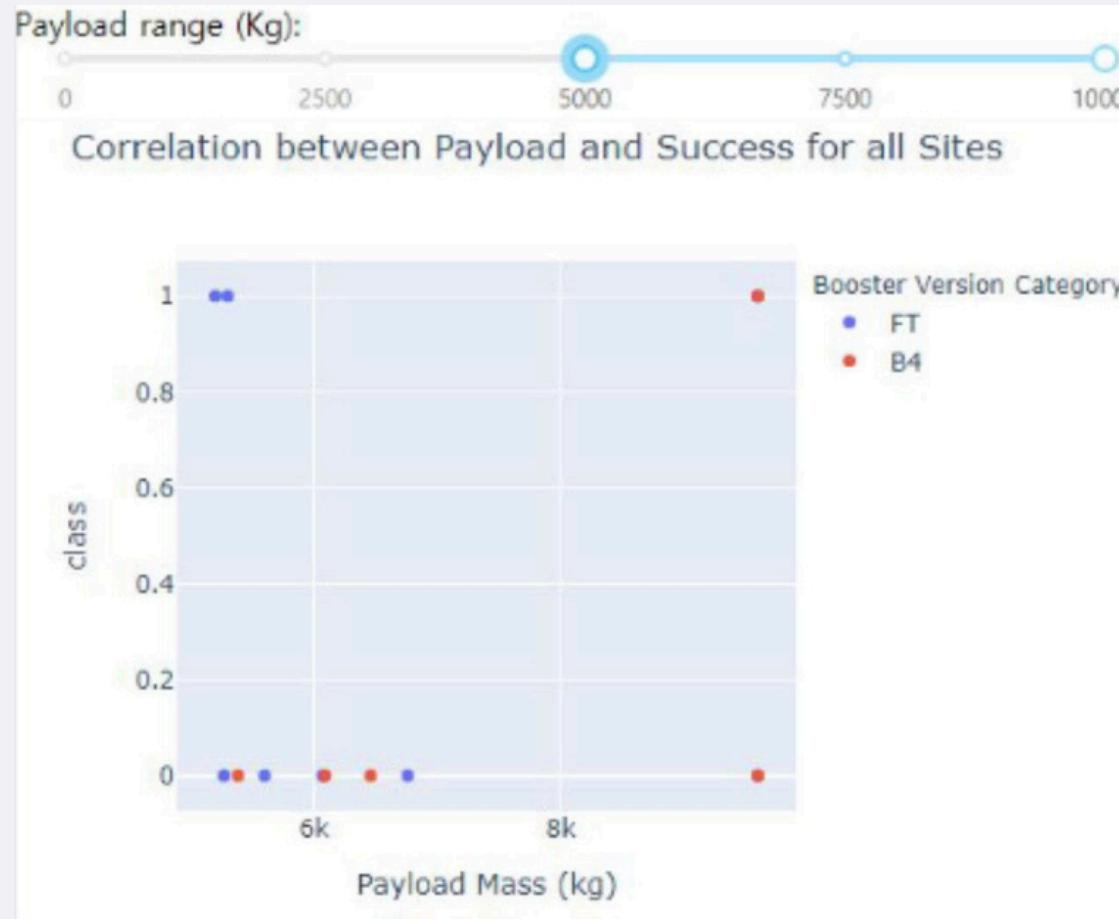
- Through the pie chart, it's possible to determine that the highest launch success rate is KSC LC-39 A with 41.7%, followed by CCAFS LC-40 with 29.2% and at the bottom the VAFB SLC-4E, and CCAFS SLC-40 with 16.7% and 12.5%, respectively.

KSC LC-39A total Success Launches



- Through the pie chart is possible to determine that KSC LC-39A has a 76.9% success rate.

Correlation between payload and success for all sites



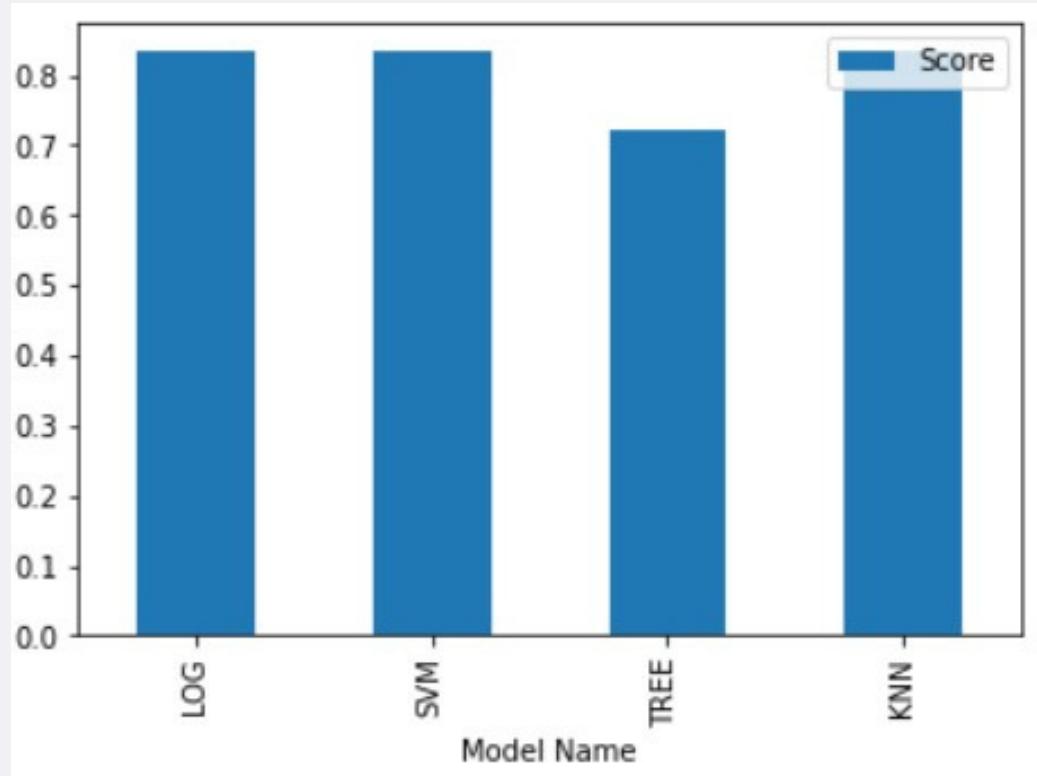
- The plot shows the correlation between payload and success for all sites with a [5000,10000] range.
- With heavier payload mass it appears to end in an unsuccessful landing.

The background of the slide features a dynamic, abstract design. It consists of several thick, curved lines in shades of blue and yellow, creating a sense of motion and depth. The lines curve from the top left towards the bottom right, with some lines being more prominent than others. The overall effect is reminiscent of a tunnel or a high-speed journey through a digital space.

Section 5

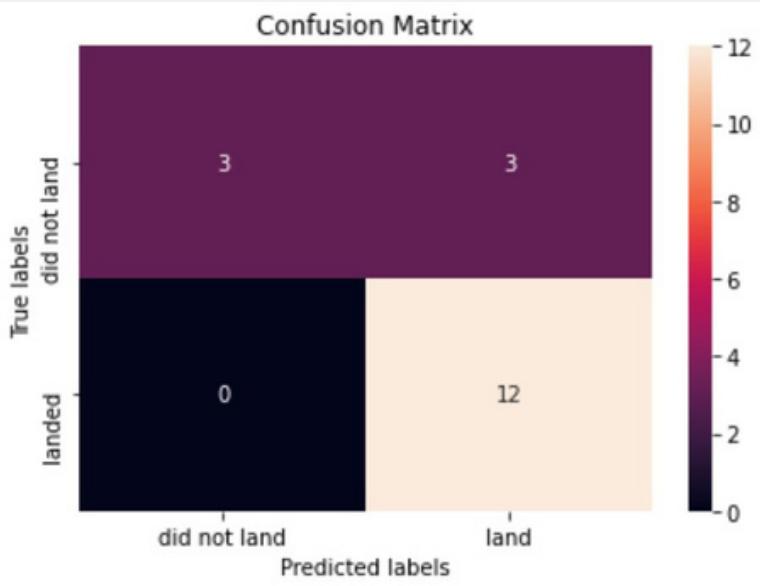
Predictive Analysis (Classification)

Classification Accuracy

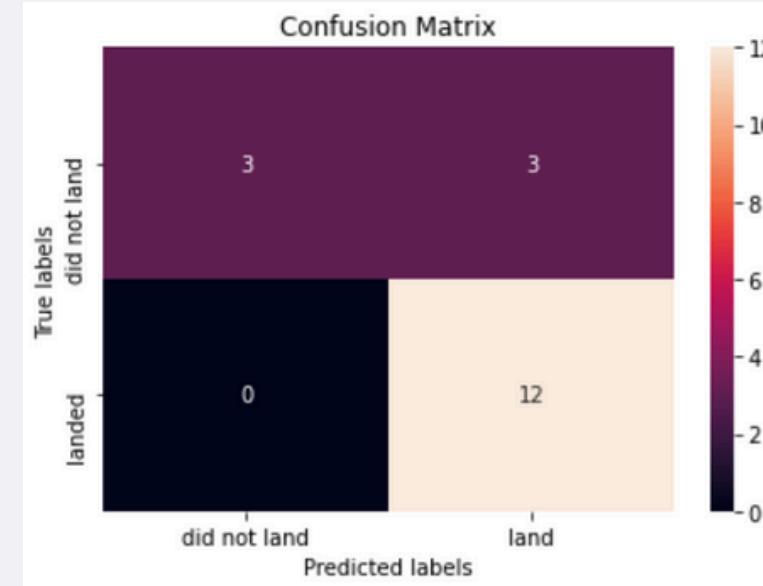


- The models best classifier performance the logistic regression, SVM, and KNN, each with a score of 83%.

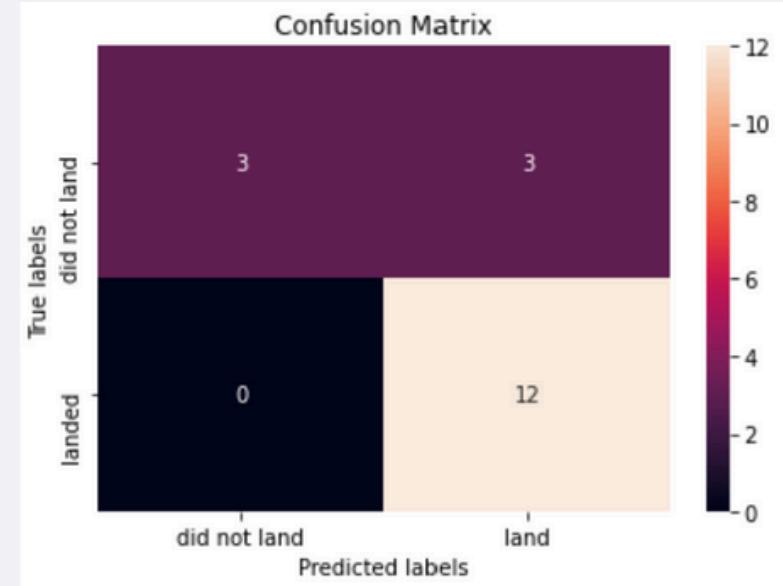
Confusion Matrix



LR results



SVM results



KNN results

- The confusion matrix for the models LR, SVM, and KNN are the same, with 3 TP, 3 FP, 0 FN, and 12 TN.

Conclusions

- Except in very specific cases, payload mass isn't related with a successful outcome.
- The orbit and the flight number appear to have a strong correlation in the analysis of an unsuccessful outcome.
- There are orbits with a tendency for a successful landing, specifically, ES-L1, GEO, HEO, and SSO have an average success rate of 100%.
- Considering the information displayed in Folium and the dashboard, a visual inspection allow us to notice a relationship between the launch site and the landing outcome.
- Logistic regression, SVM, and KNN are suitable classification model for estimation of a successful landing considering the orbit, launch site, landing pad, and serial of the Falcon 9.

Appendix

- The repository for this project is available [here.](#)

Thank you!

