Data Science Capstone Project

Ilya Elper
21.05.2023

Outline

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

Executive Summary

- Summary of methodologies
  - Collecting Data Using API
  - Collecting Data Using Web Scraping
  - Data Wrangling
  - EDA with SQL
  - EDA with Visualization
  - Interactive Visual Analytics with Folium
  - Predictive Analysis
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

Introduction

- Project background and context

  We will predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars 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.

- Problems you want to find answers
  - What factors determine whether a rocket lands successfully?
  - Interaction between different features that determine the likelihood of success of a successful landing.
  - What operating conditions must be created to ensure a successful planting program.


METHODOLOGY

Methodology

Executive Summary

- Data collection methodology:
  - Data from SpaceX was obtained from 2 sources:
    1. Using SpaceX Rest API
    2. Using Web Scrapping from Wikipedia
- Perform data wrangling
  - Filtering the data
  - Dealing with missing values
  - Using One Hot Encoding to prepare the data to binary classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

  Building, tuning and avaluation of classification models to ensure the best results

Data Collection

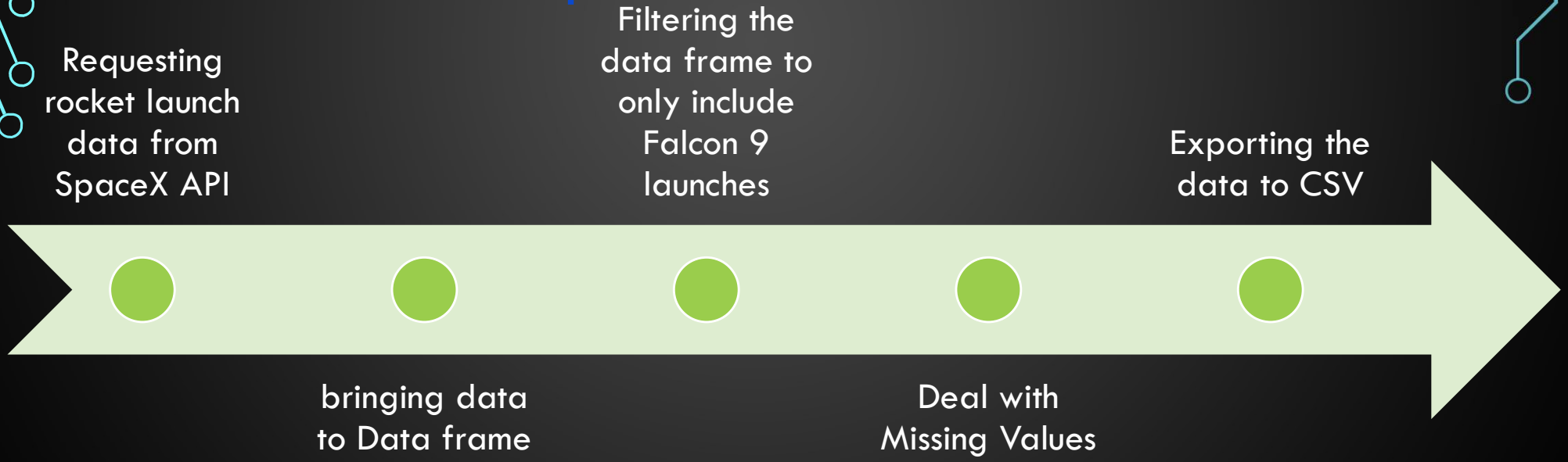
The data collection process involved a combination of API requests from SpaceX REST and a table in the SpaceX Wikipedia article.

We had to use both data collection methods in order to obtain complete information about the launches for a more detailed analysis.

Link to SpaseX API

Link to Wikipedia

Data Collection – SpaceX API



Link to GitHub URL :**Collecting the data** SpaseX API

Data Collection - Scraping

Requesting Falcon 9 launch data from Wikipedia

Creating a BeautifulSoup object from the HTML response

Extracting all column names and Collecting the data by parsing HTML tables

Constructing data we have obtained into a dictionary and further transformation to data frame

Exporting the data to CSV

Link to GitHub URL :**Web scraping from Wikipedia**

Data Wrangling

- Perform exploratory Data Analysis And determine Training Labels
- Calculate the number of launches on each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurrence of mission outcome per orbit type
- Create a landing outcome label from Outcome column
- Exporting the data to CSV

Link to GitHub URL :**Data wrangling**

EDA with Data Visualization

In this EDA, three types of graphs were built:

1. Scatter plots - show the relationship between variables. If a relationship exists, they could be used in machine learning model.
2. Bar charts - show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value.
3. Line charts - show trends in data over time (time series).

Link to GitHub URL : **EDA with Data Visualization**

EDA with SQL

Executed SQL queries:

- Displaying the names of unique launch pads in a space mission
- Display of 5 entries where starting pads begin with the string "CCA".
- A display of the total payload mass carried by NASA Launched Launch Vehicles (CRS)
- Display of the average payload mass of the launch vehicle version F9 v1.1
- An indication of the date of the first successful landing on an unpaved area.
- Listing of names of launch vehicles that are successful in unmanned vehicles and have a payload mass of more than 4000, but less than 6000
- List of total number of successful and unsuccessful mission results
- Listing of launch vehicle version names with maximum payload mass.
- A list of unsuccessful landings on the drone ship, their booster versions and the names of the launch sites for the months in 2015
- Ranking the number of landing results (eg "Failure" (unmanned craft) or "Success" (ground pad)) between date 06/04/2010 and 03/20/2017 in descending order

Link to GitHub URL : **EDA with SQL**

Build an Interactive Map with Folium

- I have created an interactive map using the Folium library. The map includes markers for all launch sites based on their latitude and longitude coordinates. Each marker features a circular area, a popup label, and a text label to identify each launch site.
- For the NASA Johnson Space Center, a separate marker with its coordinates is added, also displaying a circular area, popup label, and text label to indicate its location.
- To show launch results at each site, I used colored markers. Successful launches are marked with green, while unsuccessful ones are marked with red. The markers are clustered to facilitate navigation and data analysis.
- Additionally, I added colored lines to display the distances between a selected launch site (such as KSC LC-39A) and its surroundings, such as railways, highways, coastlines, and the nearest city.
- This interactive map provides a visual representation of the geographical positions of the launch sites and their launch results, as well as demonstrates their proximity to various features and locations.

Link to GitHub URL : Interactive Visual Analytics with Folium

Build a Dashboard with Plotly Dash

- Implemented a dropdown list for selecting launch sites. Pie Chart Showing Successful Launches (All Sites/Specific Site):
- Included a pie chart displaying the total count of successful launches for all sites and the breakdown of success versus failures if a specific launch site is chosen. Slider for Payload Mass Range:
- Integrated a slider to allow the selection of payload mass range. Scatter Chart of Payload Mass vs. Success Rate for Different Booster Versions:
- Created a scatter chart to illustrate the relationship between payload mass and launch success rate across different booster versions.

Link to GitHub URL : SpaseX Dash App

Predictive Analysis (Classification)

Data preparation

- Creating a NumPy array from the columm “Class” in data
- Standardization the data with StandardScaler
- Splitting the data to train and test sets with train_test_split function

Modeling and validation

- Performing cross-validation, each learning model.
- logistic regression
- support vector machines (SVM)
- decision trees
- K-nearest neighbors (KNN)

Comparison of results

- Finding the method performs best by examining the Jaccard_Score , F1_score and Accuracy metrics

Results

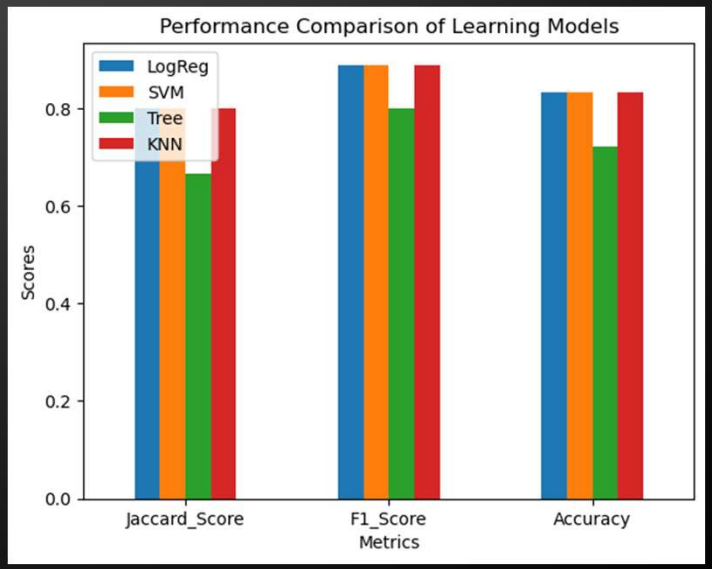
Exploratory data analysis findings:

- Space X operates from four different launch sites.
- Initial launches were conducted for Space X itself and NASA.
- The average payload of the F9 v1.1 booster is 2,928 kg.
- The first successful landing occurred in 2015, five years after the initial launch.
- Several Falcon 9 booster versions achieved successful landings on drone ships, particularly with payloads above the average.
- The mission success rate was nearly 100%.
- Two booster versions, F9 v1.1 B1012 and F9 v1.1 B1015, failed to land on drone ships in 2015.
- The number of successful landing outcomes improved over the years.

Results

- By utilizing interactive analytics, it was possible to observe that launch sites are strategically located in safe areas, near the sea, and have well-developed logistic infrastructure in their vicinity. Furthermore, the majority of launches occur at launch sites situated on the east coast.
- Based on the results of predictive analysis, it was determined that the decision tree classifier performed the poorest in predicting successful landings. On the other hand, the remaining models demonstrated comparable levels of accuracy in their predictions.

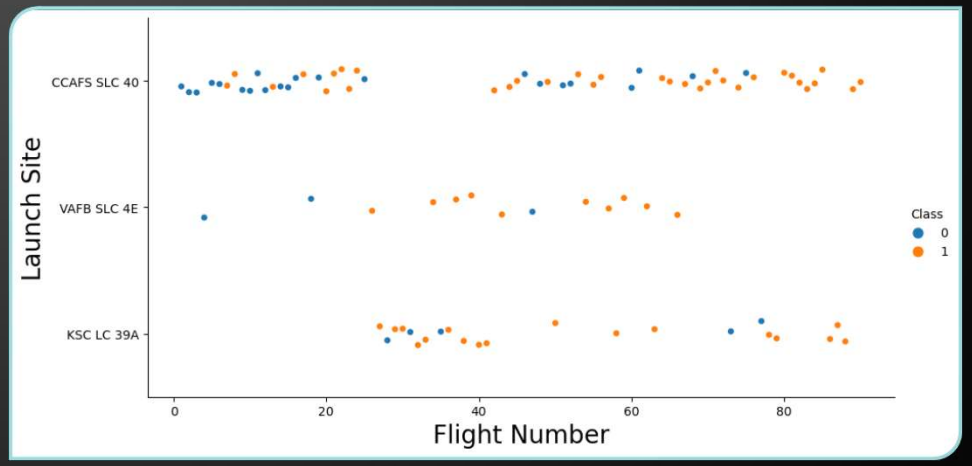
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.666667	0.800000
F1_Score	0.888889	0.888889	0.800000	0.888889
Accuracy	0.833333	0.833333	0.722222	0.833333




INIGHTS DRAWN FROM EDA

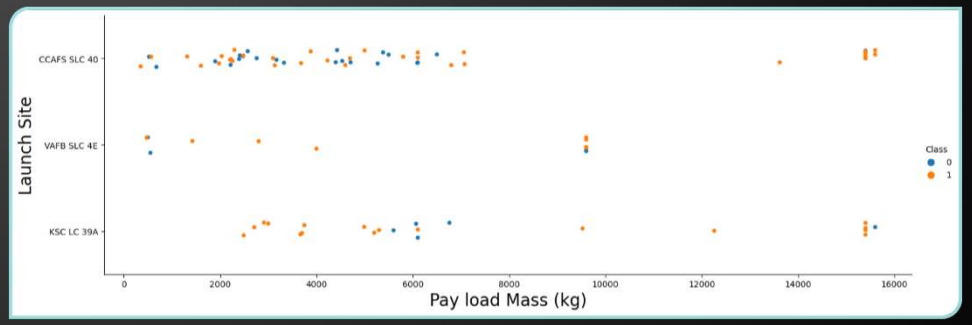
FLIGHT NUMBER VS. LAUNCH SITE

- According to the plot, it is evident that the best launch site currently is CCAFS SLC 40, as it has witnessed a higher rate of successful launches, with a majority of recent launches achieving success. Following closely are VAFB SLC 4E and KSC LC 39A, which also have notable success rates. Additionally, the plot demonstrates an overall improvement in the success rate of launches over time.



PAYLOAD VS. LAUNCH SITE

- CCAFS SLC 40 has achieved a 100% success rate for launches with a payload mass exceeding 7000 Kg. Conversely, KSC LC 39A has maintained a 100% success rate for launches with a payload mass below 5500 kg. VAFB-SLC 4E, on the other hand, has had the fewest launches and none of them involved a payload exceeding 10,000 kg.



SUCCESS RATE VS. ORBIT TYPE

- Orbits with 100% success rate:

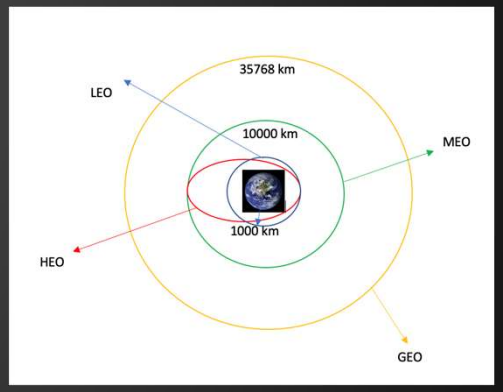
  ES-L1, GEO, HEO, SSO

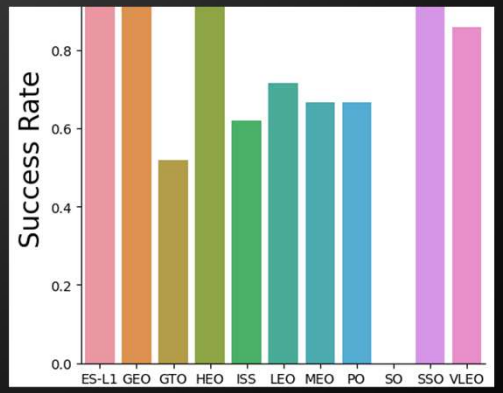
- Orbits with 0% success rate:

  SO

- Orbits with success rate between 50% and 85%:

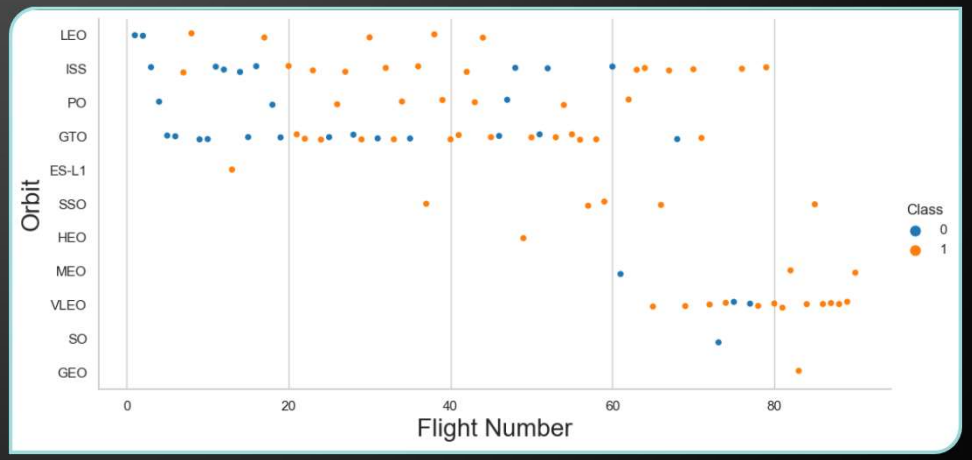
  GTO, ISS, LEO, MEO, PO





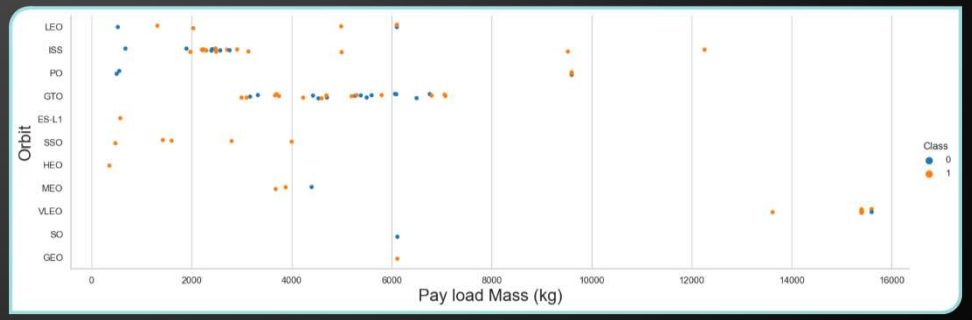
FLIGHT NUMBER VS. ORBIT TYPE

- Success rate improved over time to all orbits
- In the LEO orbit the Success appears related to the number of flights;
- There is no relationship between flight number when in GTO orbit.
- increased flights to VLEO orbit .



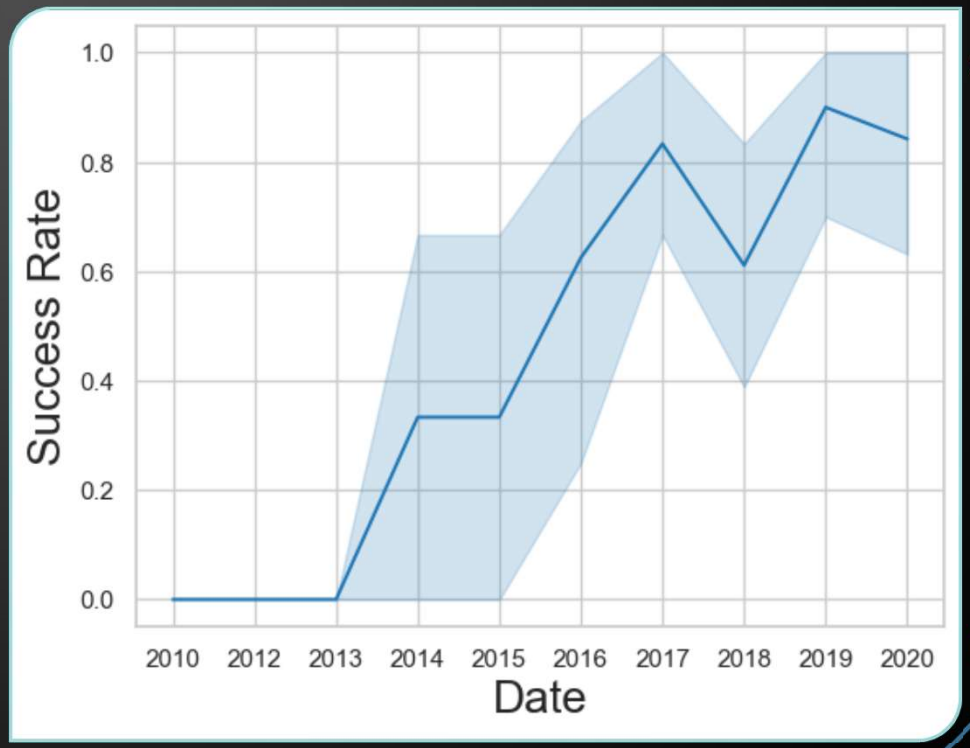
PAYLOAD VS. ORBIT TYPE

- Apparently, there is no relation between payload and success rate to orbit GTO;
- ISS orbit has the widest range of payload and a good rate of success;
- There are few launches to the orbits SO and GEO.



LAUNCH SUCCESS YEARLY TREND

- The level of success started to rise in 2013, and 2014 was a year of stability. The growth in success continued from 2015 to 2017, possibly indicating another breakthrough. In 2017, some problems surfaced, which were addressed in 2019. In 2018, there was a decrease in success, but it was recovered and slightly improved in 2019. In 2020, there was a small decline.



All Launch Site Names

- According to data, there are four launch sites:

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

They are obtained by selecting unique occurrences of " launch_site " values from the dataset.

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`:

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

Here we can see five samples of Cape Canaveral launches

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

Total Payload Mass

- Total payload carried by boosters from NASA equals: 45596 kg

SUM (PAYLOAD_MASS__KG_)
45596

- Total payload calculated above, by summing all payloads whose codes contain 'CRS', which corresponds to NASA.

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

Average Payload Mass by F9 v1.1

- **Average payload mass carried by booster version F9 v1.1 equals:38020 kg**

SUM (PAYLOAD_MASS__KG_)
38020

- **Filtering data by the booster version above and calculating the average**

First Successful Ground Landing Date

- **Fist successful landing outcome on ground pad equal : 01/05/2015**

MIN (Date)
01-05-2017

- **By filtering data by successful landing outcome on ground pad and getting the minimum value for date it's possible to identify the first occurrence.**

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

Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

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

- Selecting distinct booster versions according to the filters above, these 4 are the result.

```
%%sql
SELECT DISTINCT  Booster_Version
FROM SPACEXTBL
WHERE Landing__Outcome='Success (drone ship)' AND
PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
;
```

Total Number of Successful and Failure Mission Outcomes

- **Number of successful and failure mission outcomes:**

Missions_outcome	count
Failure	1
Success	100

- **Grouping mission outcomes and counting records for each group led us to the summary above.**

```
%%sql
SELECT
    CASE
        WHEN Mission_Outcome = 'Failure (in flight)' THEN 'Failure'
        ELSE 'Success'
    END as Missions_outcome,
    COUNT(*) as count
FROM SPACEXTBL
GROUP BY Missions_outcome;
```

BOOSTERS CARRIED MAXIMUM PAYLOAD

- Boosters which have carried the maximum payload mass

- These are the boosters which have carried the maximum payload mass registered in the dataset.

```
%%sql
SELECT BOOSTER_VERSION,PAYLOAD_MASS__KG_
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

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

2015 Launch Records

- List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

month	Launch_Site	Booster_Version	Landing__Outcome
06	CCAFS LC-40	F9 v1.1 B1018	Precluded (drone ship)

- The list above has the only two occurrences.

```
%%sql
SELECT
    substr(Date, 4, 2) as month,
    Launch_Site,
    Booster_Version,
    Landing__Outcome
FROM
    SPACEXTBL
WHERE
    substr(Date,7,4)='2015' AND
    Landing__Outcome LIKE '%drone ship%' AND
    Mission_Outcome LIKE'%Failure%';
```

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

- Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

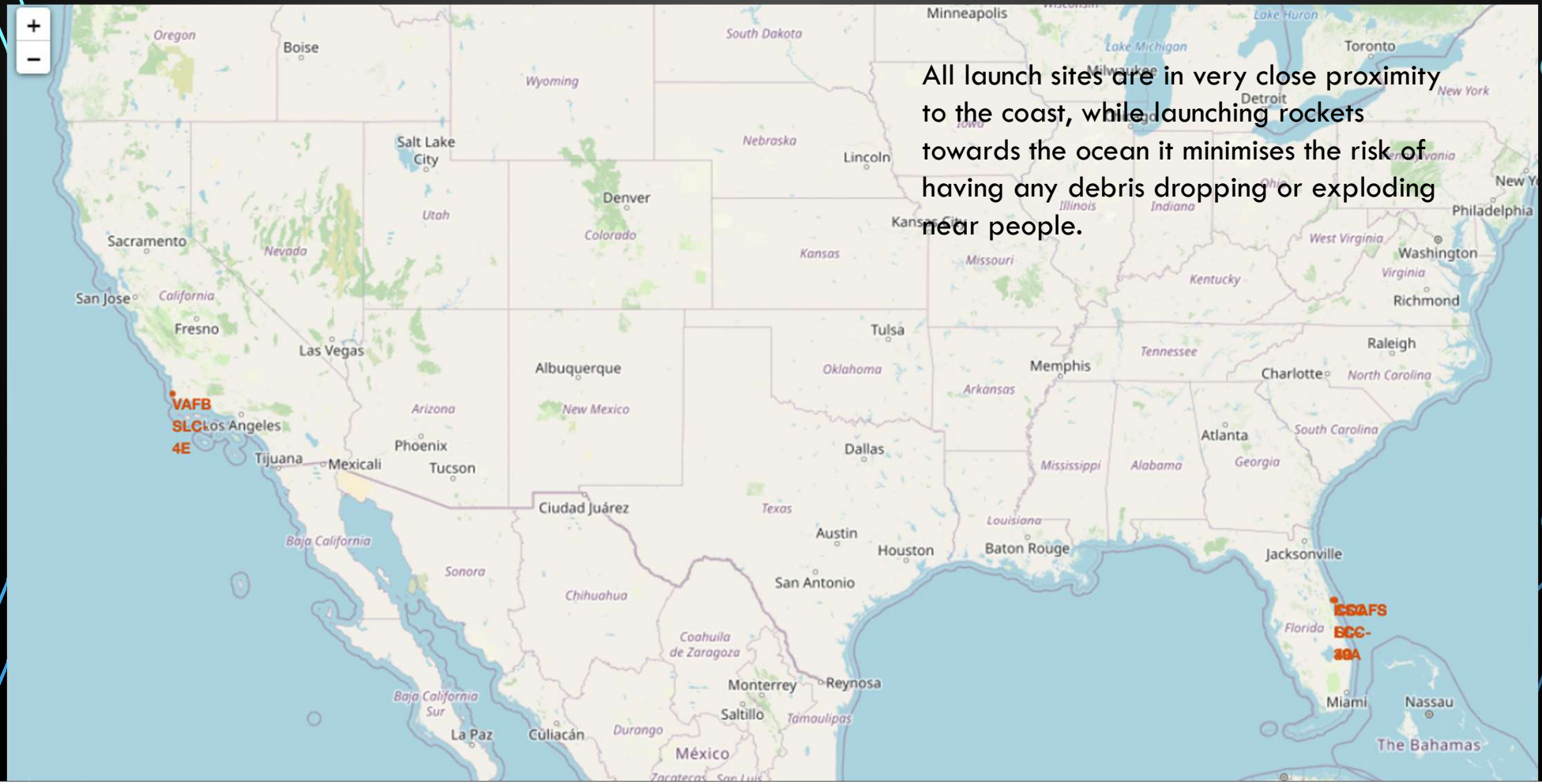
anding__Outcome	Success_Count
Success	20
Success (drone ship)	8
Success (ground pad)	6

- Present your query result with a short explanation here

```
%%sql
SELECT "Landing__Outcome", COUNT(*) AS "Success_Count"
FROM SPACEXTBL
WHERE "Landing__Outcome" LIKE '%Success%' AND "Date" BETWEEN  '04-06-2010' AND '20-03-2017'
GROUP BY "Landing__Outcome"
ORDER BY "Success_Count" DESC;
```

Launch Sites Proximities Analysis

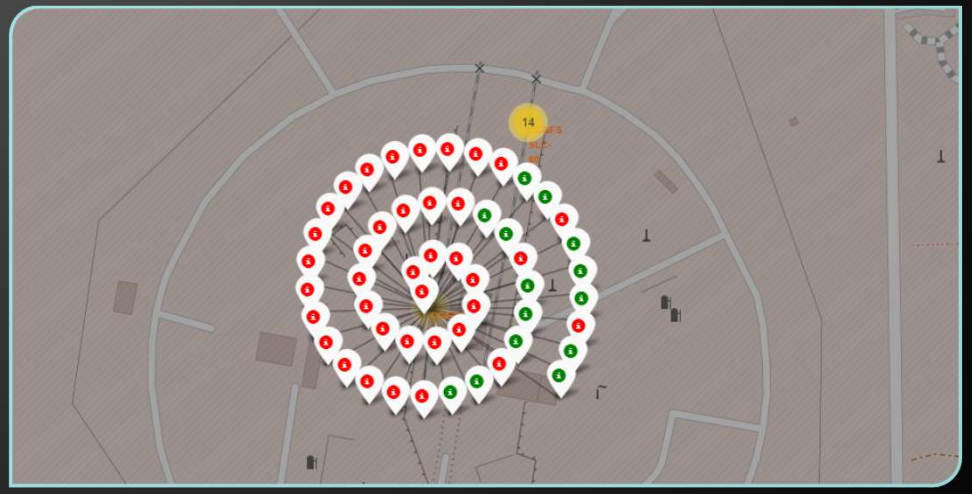
All launch sites



All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimises the risk of having any debris dropping or exploding near people.

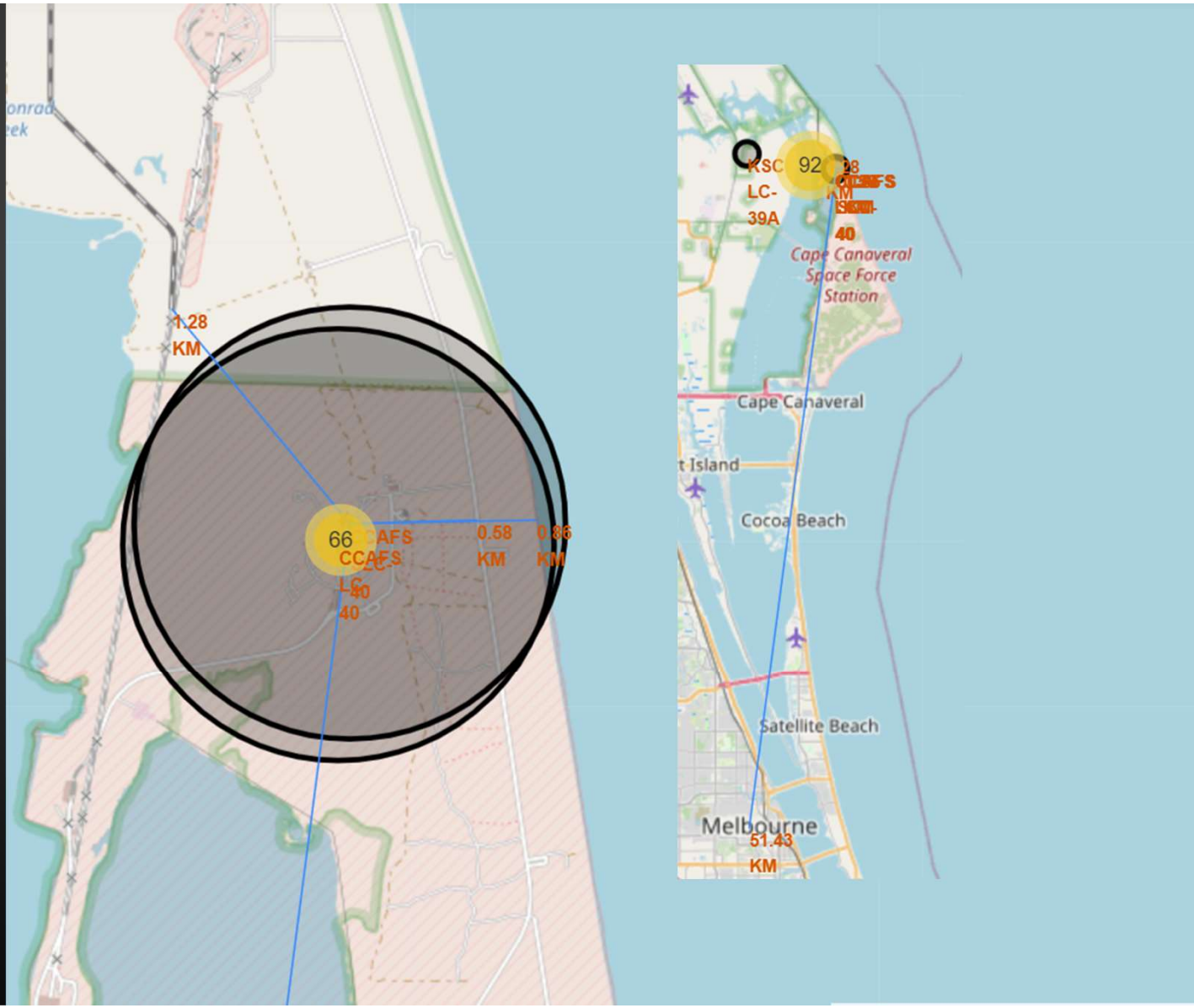
Color-labeled launch records on the map

- From the color-labeled markers we should be able to easily identify which launch sites have relatively high success rates.
  - Green Marker = Successful Launch
  - Red Marker = Failed Launch

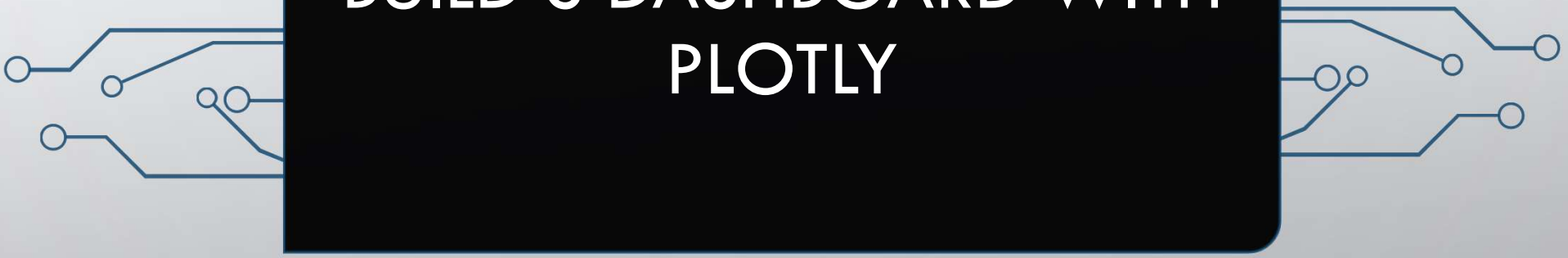


Distance from the launch site CCAFS SLC-40 to its proximities

- Launch site CCAFS SLC-40 has good logistics aspects, being near railroad and road and relatively far from inhabited areas.

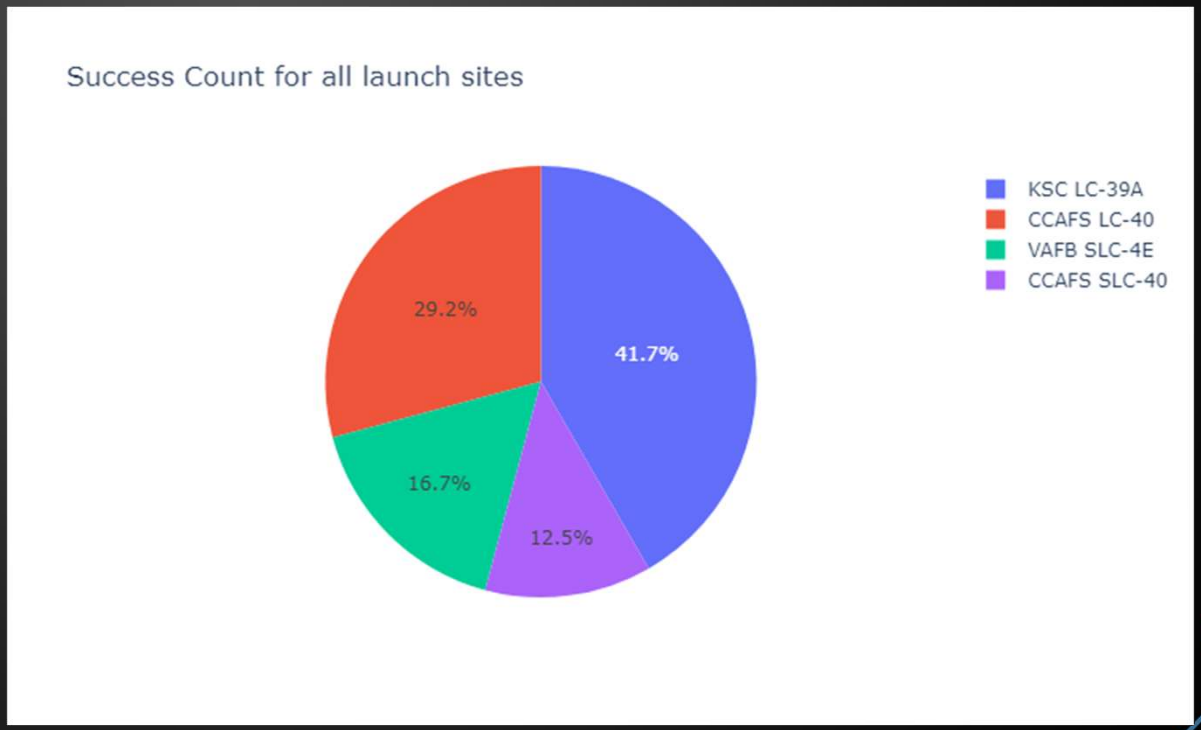


BUILD S DASHBOARD WITH PLOTLY



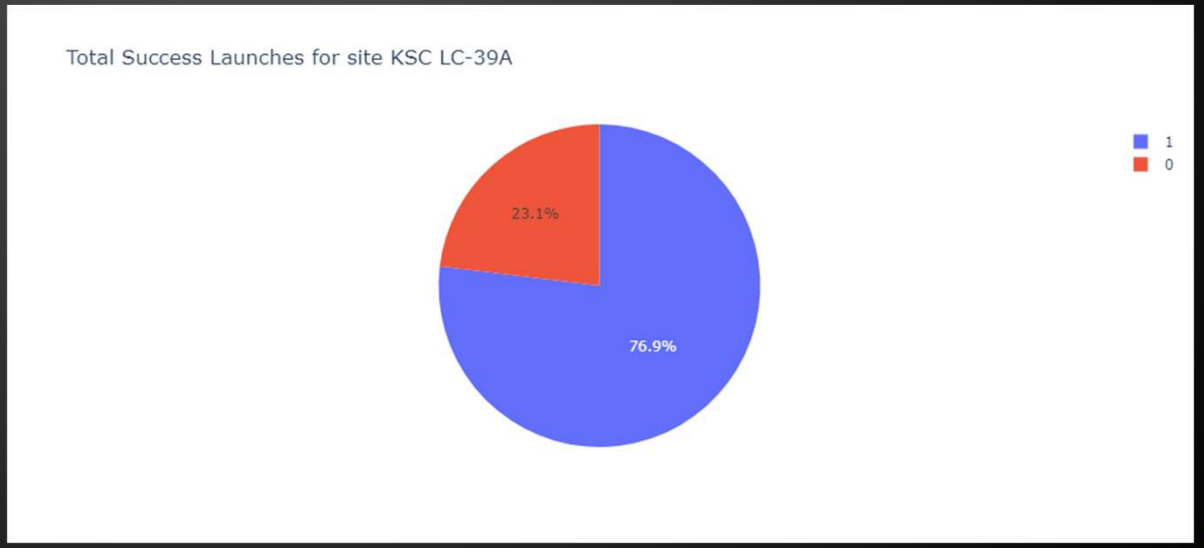
SUCCESSFUL LAUNCHES BY SITE

- The place from where launches are done seems to be a very important factor of success of missions.



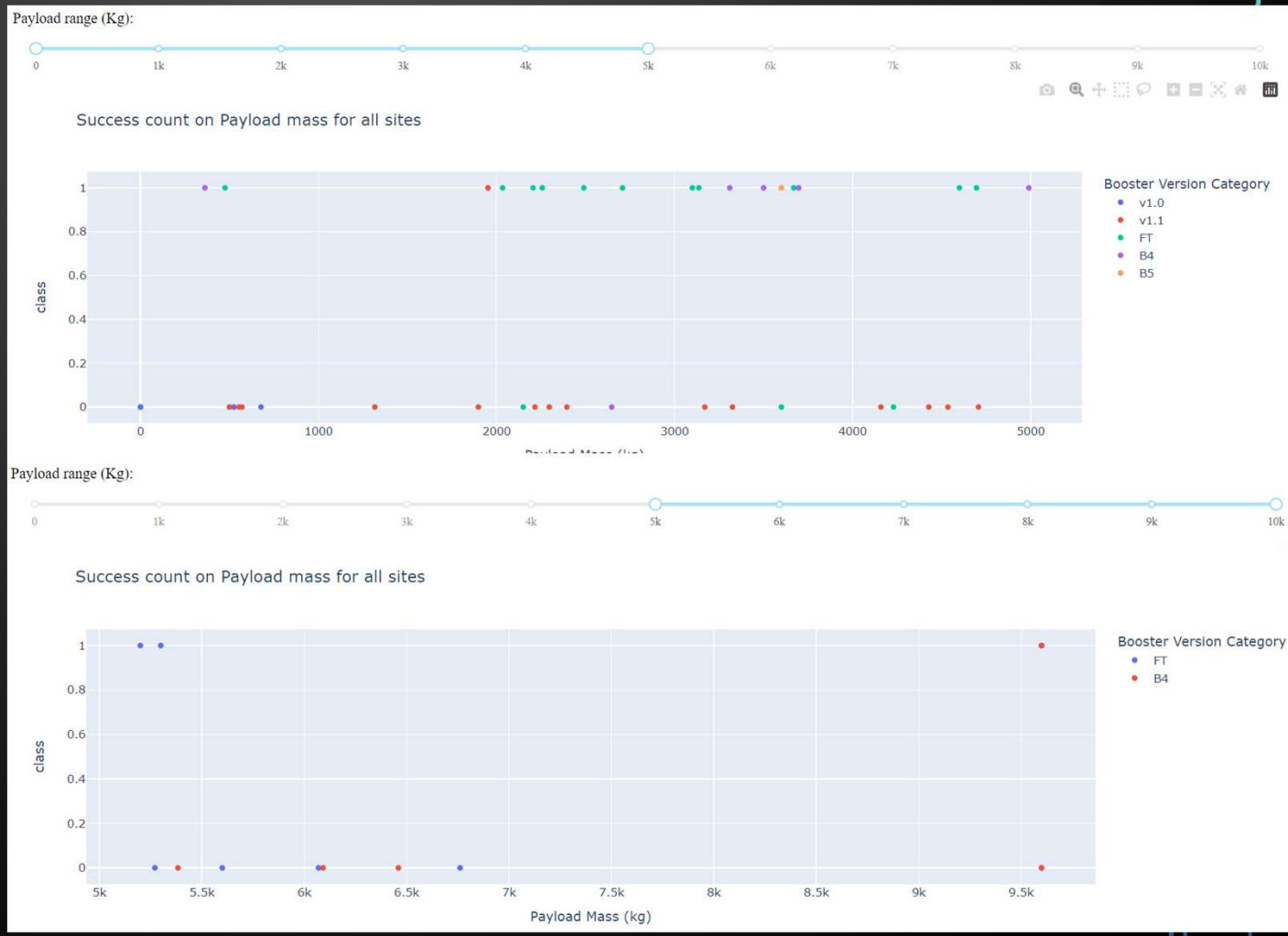
Launch Success Ratio for KSC LC-39A

- 76.9% of launches are successful in this site.

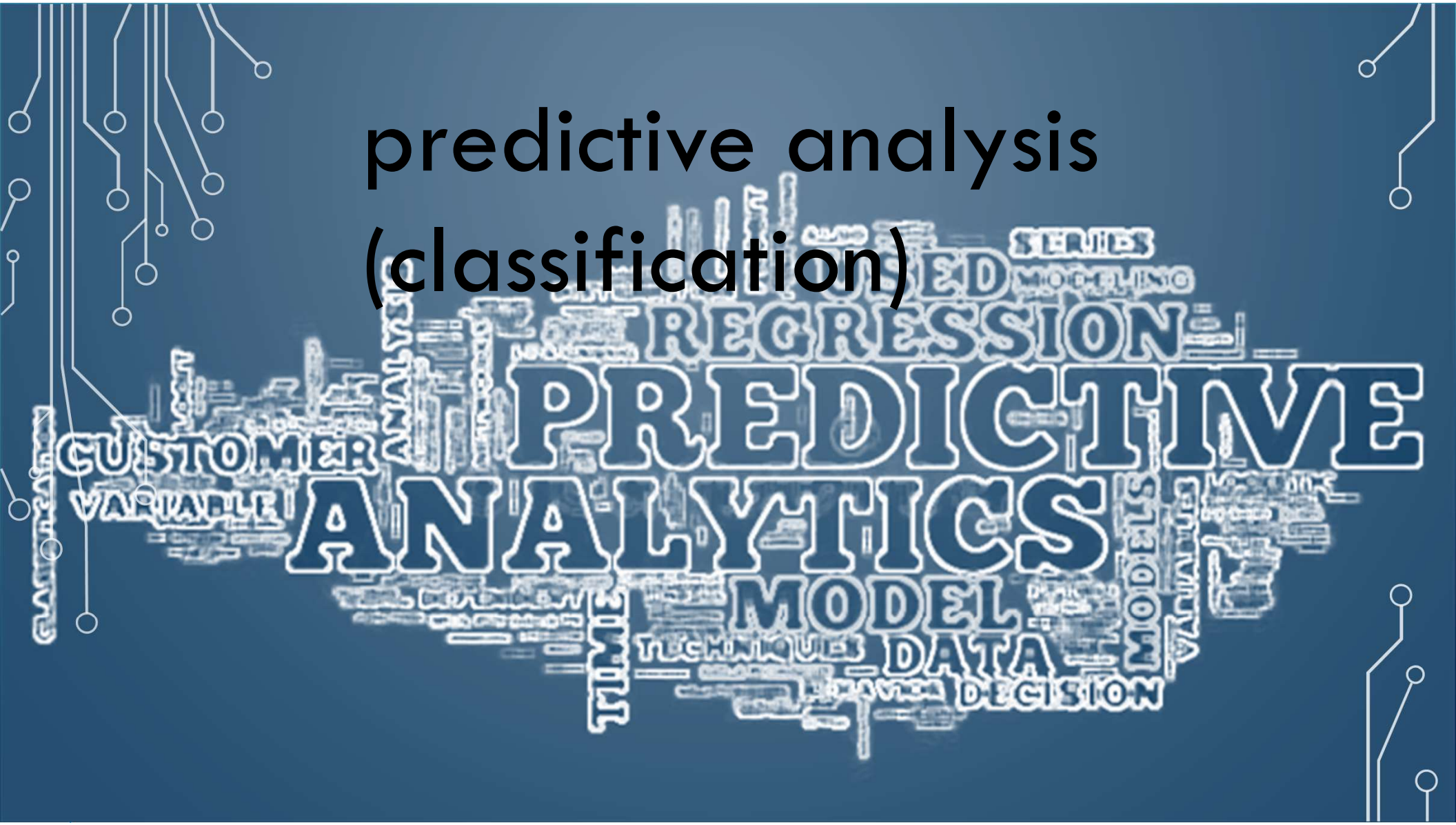


Payload Mass vs. Launch Outcome for all sites

The charts show that payloads between 2000 and 5500 kg have the highest success rate.

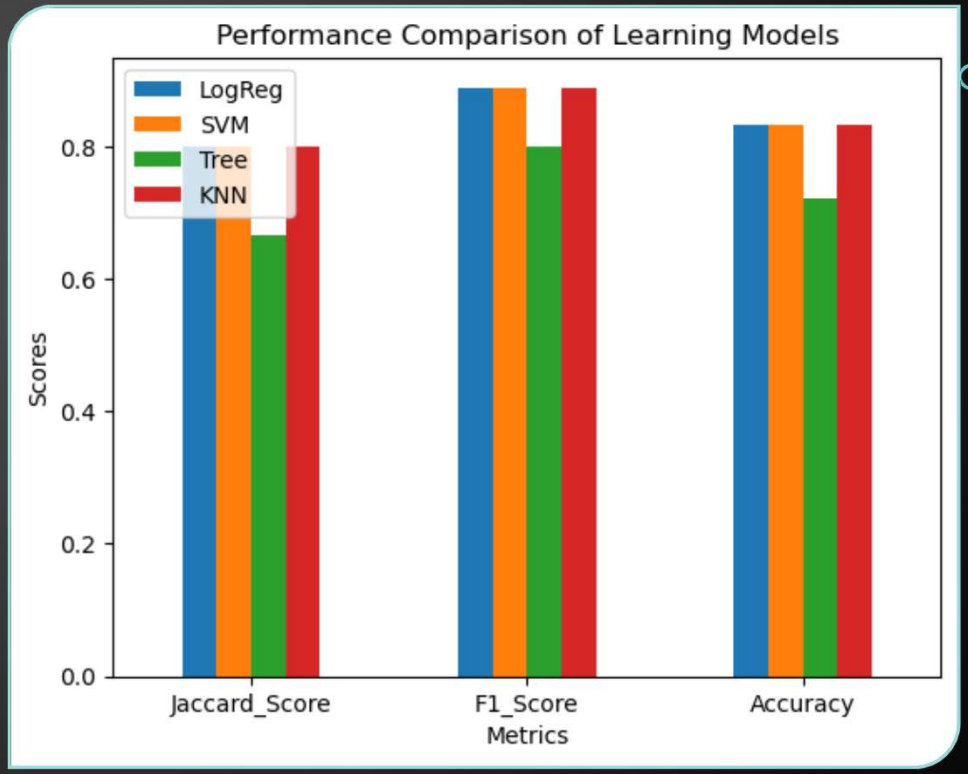


predictive analysis (classification)



CLASSIFICATION ACCURACY

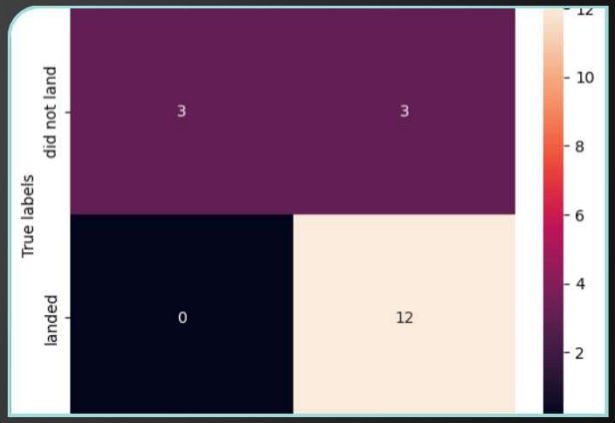
Based on the results of predictive analysis, it was determined that the decision tree classifier performed the poorest in predicting successful landings. On the other hand, the remaining models demonstrated comparable levels of accuracy in their predictions.



	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.666667	0.800000
F1_Score	0.888889	0.888889	0.800000	0.888889
Accuracy	0.833333	0.833333	0.722222	0.833333

CONFUSION MATRIX

- Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.



Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

CONCLUSIONS

Decision Tree Model is the bad algorithm for this dataset.

Most of launch sites are in proximity to the Equator line

All the sites are in very close proximity to the coast.

The success rate of launches increases over the years.

KSC LC-39A has the highest success rate of the launches from all the sites.

•Orbits ES-L1, GEO, HEO and SSO have 100% success rate.