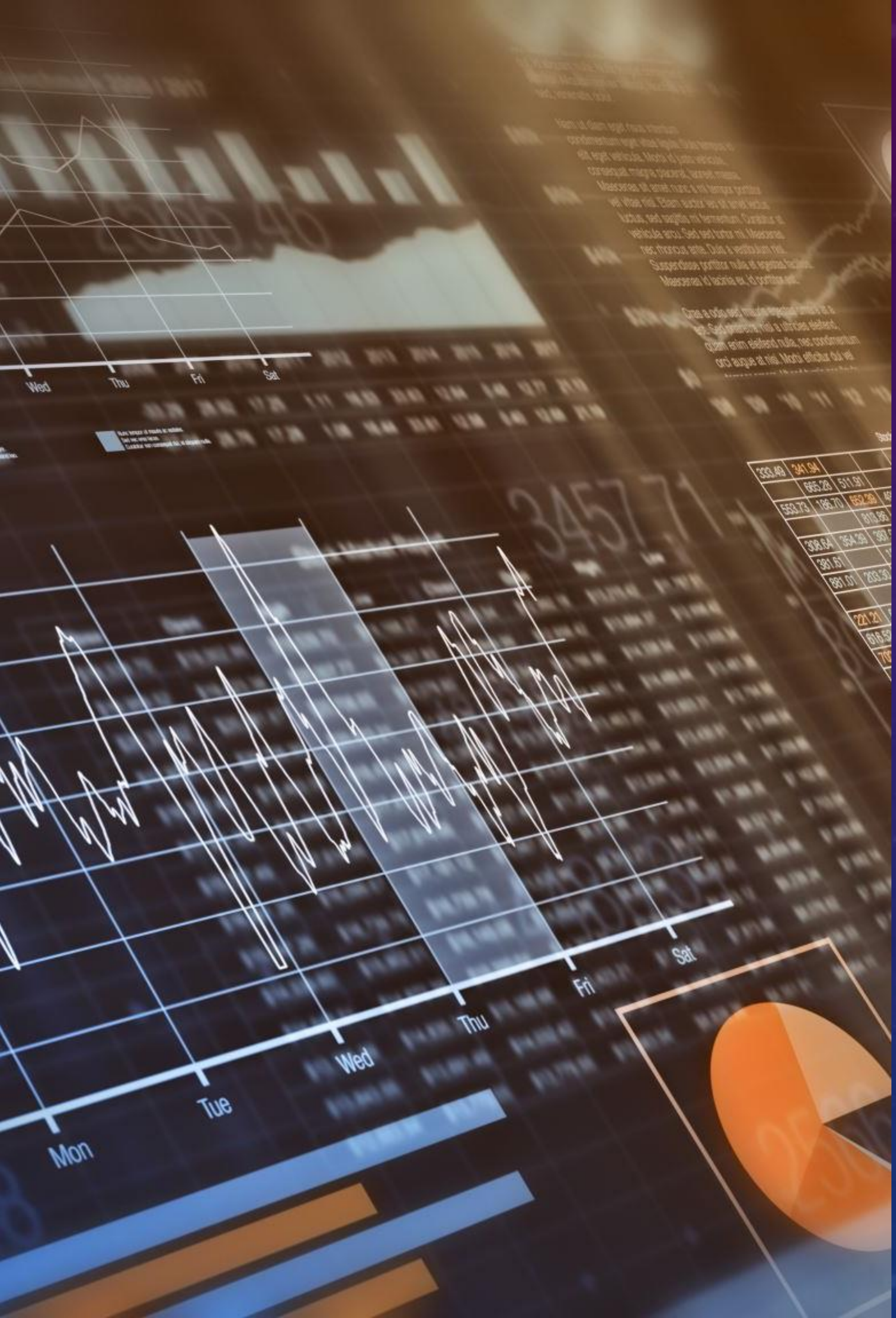
IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ali Dali
3rd November 2022

OUTLINE

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



EXECUTIVE SUMMARY

- Summary of methodologies
- Summary of all results

INTRODUCTION

- 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.
- The successful launch of SpaceX's Falcon 9 rockets
- Does the mass payload affect the success rate?
- Which is the best model we can use to predict the outcome of a launch

Section 1

Methodology

METHODOLOGY

•Executive Summary

- Data collection methodology:
  - Data was collected from digital sources
- Perform data wrangling
  - Data was cleaned, parsed, normalized, classified and subject to 4 different predictive models.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium
- Perform predictive analysis using classification models

DATA COLLECTION

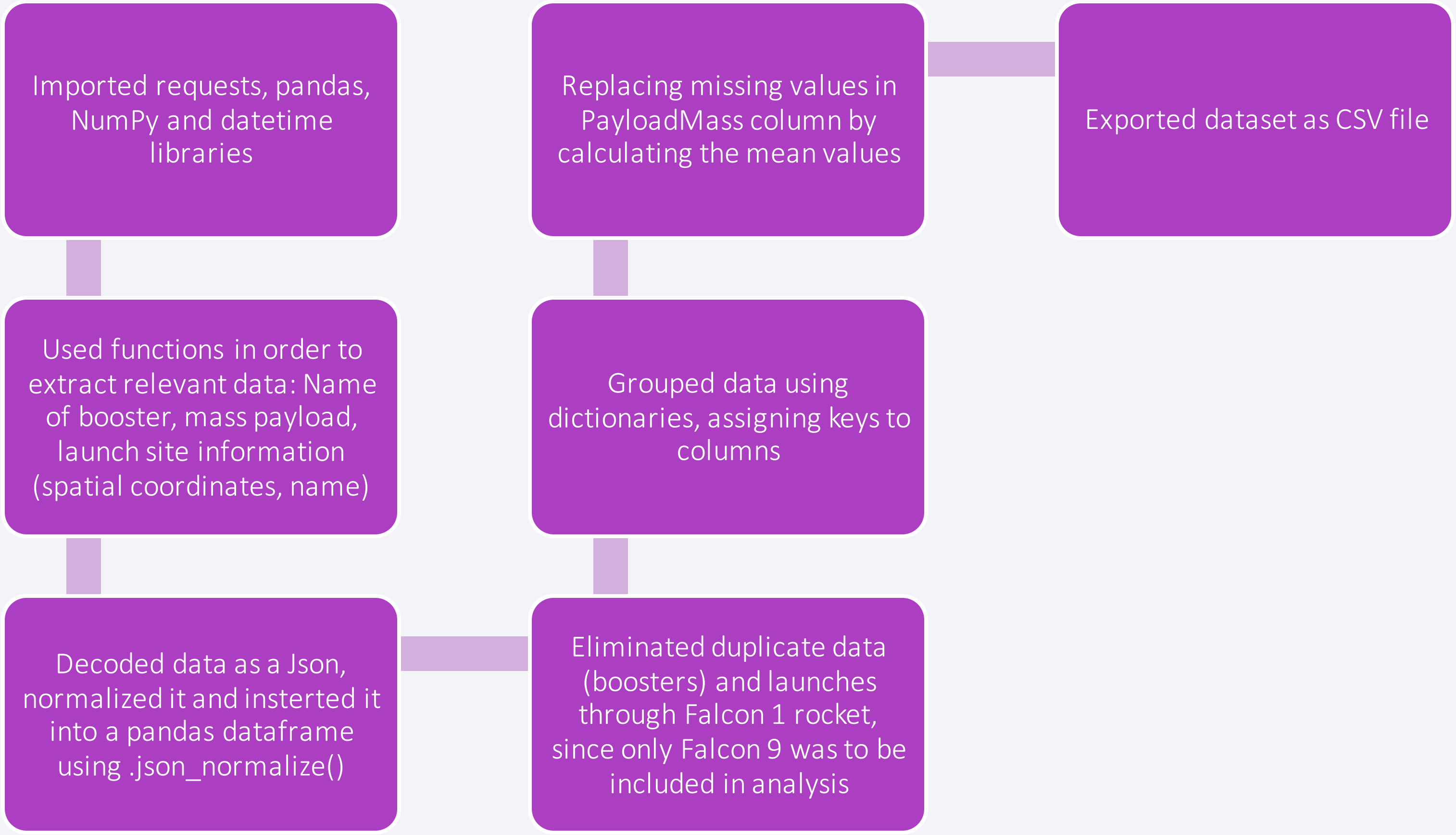
First, we imported various libraries in order to use SpaceX's own API

We were also able to find relevant information on public websites such as Wikipedia

Data Collection – SpaceX API

- We took advantage of Python's rich array of libraries to collect data from SpaceX's API and process it for later use

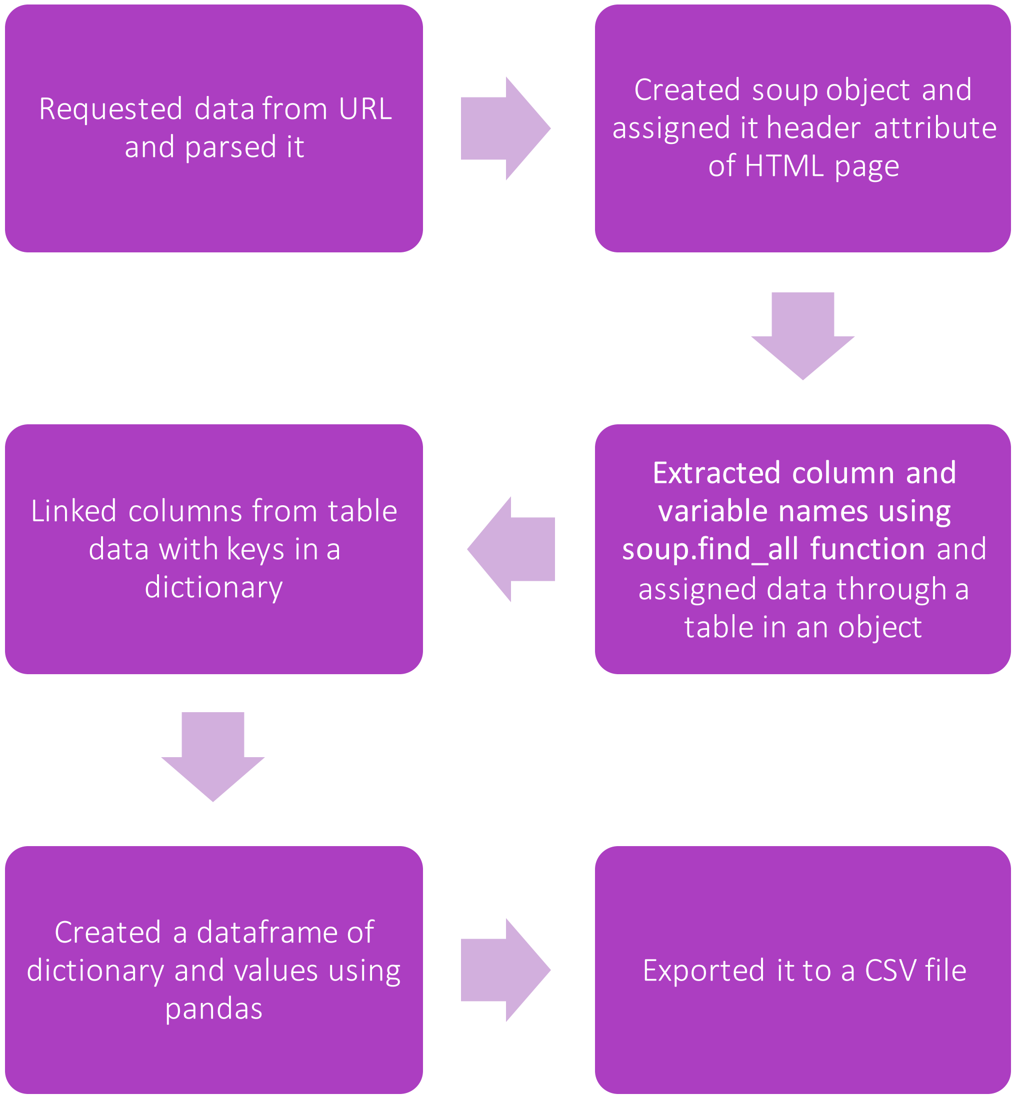
- [Click here to view notebook on Github]()



Data Collection – Web Scraping

As before, we used external libraries such as pandas but also BeautifulSoup, in combination with Python methods to extract data from a Wikipedia page called: List of Falcon 9 and Falcon Heavy launches. It was updated on 9th June 2021.

The Jupyter notebook can be accessed on Github by clicking <u>here</u>



DATA WRANGLING

We used some Exploratory Data Analysis (EDA) techniques in order to transform the scraped data and determine what attributes and values are suited best for achieving the target objective of being able to predict the success or failure of Falcon 9 rocket launches.

Imported NumPy and Pandas libraries

↓

Loaded table with Rocket Launch information from CSV file

↓

Used Pandas functions to find out missing values and datatypes in and of each column

↓

Calculated the number of launches per site

```
In [3]: df.isnull().sum()/df.count()*100

Out[3]: FlightNumber      0.000
        Date              0.000
        BoosterVersion    0.000
        PayloadMass       0.000
        Orbit             0.000
        LaunchSite        0.000
        Outcome           0.000
        Flights           0.000
        GridFins          0.000
        Reused            0.000
        Legs              0.000
        LandingPad       40.625
        Block             0.000
        ReusedCount       0.000
        Serial            0.000
        Longitude         0.000
        Latitude          0.000
        dtype: float64
```

There were a disproportional amount of missing values for the "LandpingPad" attribute,

```
In [4]: df.dtypes

Out[4]: FlightNumber        int64
        Date               object
        BoosterVersion     object
        PayloadMass       float64
        Orbit              object
        LaunchSite         object
        Outcome            object
        Flights             int64
        GridFins             bool
        Reused               bool
        Legs                 bool
        LandingPad         object
        Block             float64
        ReusedCount         int64
        Serial             object
        Longitude         float64
        Latitude          float64
        dtype: object
```

Certain columns did not have the proper data types and had to be changed

```
        df['LaunchSite'].value_counts()

)ut[5]: CCAFS SLC 40    55
        KSC LC 39A      22
        VAFB SLC 4E     13
        Name: LaunchSite, dtype: int64
```

Number of launches for each of the 3 launch pads

DATA WRANGLING

Use the method `.value_counts()` to determine the number and occurrence of each orbit in the column `Orbit`

Calculated the number and occurence of mission outcome per orbit type using .value_counts()

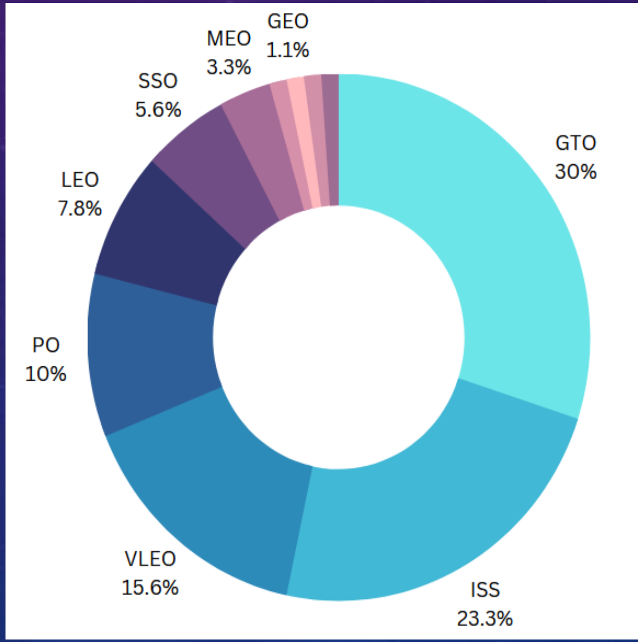
Created a landing outcome label from Outcome column

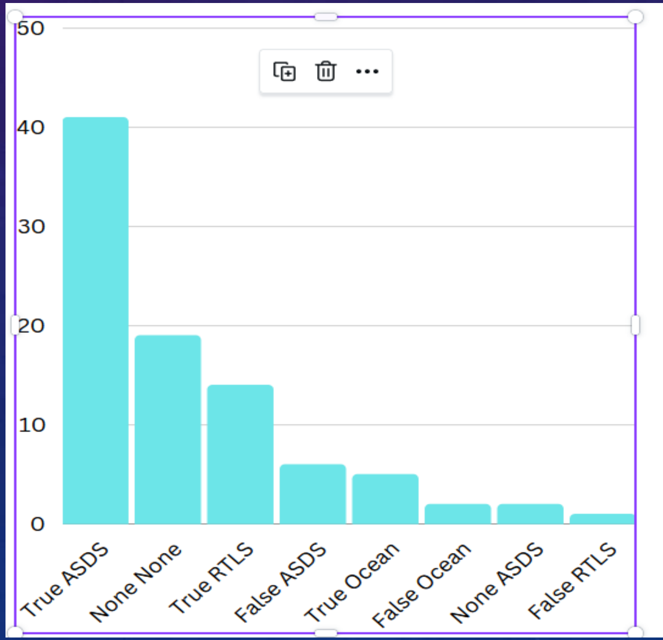
Jupyter notebook on Github [here]()

The orbital target of a rocket launch is essential in find out whether or not it will be successful. Rockets were launched up to different distances from the earth, but the two most common orbits were:

- Geosynchronous orbit (GTO) located at 22,236 miles (35,786 kilometers) above Earth's equator. It was the target of 27 out of 90 launches
- International Space Station (ISS), for the purpose of resupplying astronauts. SpaceX provisioned the ISS through 21 launches.

TRUE ASDS (meaning the mission outcome was successfully landed to a drone ship) was the most common outcome.






```
In [14]: df["Class"].mean()
Out[14]: 0.6666666666666666
```

Using the .mean() function, we found out that the average success rate was 66%

EDA with Data Visualization

Training a machine learning model requires understanding on patterns and trends in relationships among different attributes. The comparisons made were the following:

- Flight Number and Launch Site
- Payload and Launch Site
- FlightNumber and Orbit type
- Payload and Orbit type

The following trends were observed, using bar and line plots:

- Launch success yearly trend
- Success rate of each orbit type

After analyzing the graphs, the columns most important for the analysis were: FlightNumber, PayloadMass, Orbit, LaunchSite, Flights, GridFins, Reused, Legs, LandingPad

Github Jupyter notebook can be found here

EDA WITH SQL

We performed queries in order to retrieve the following information:

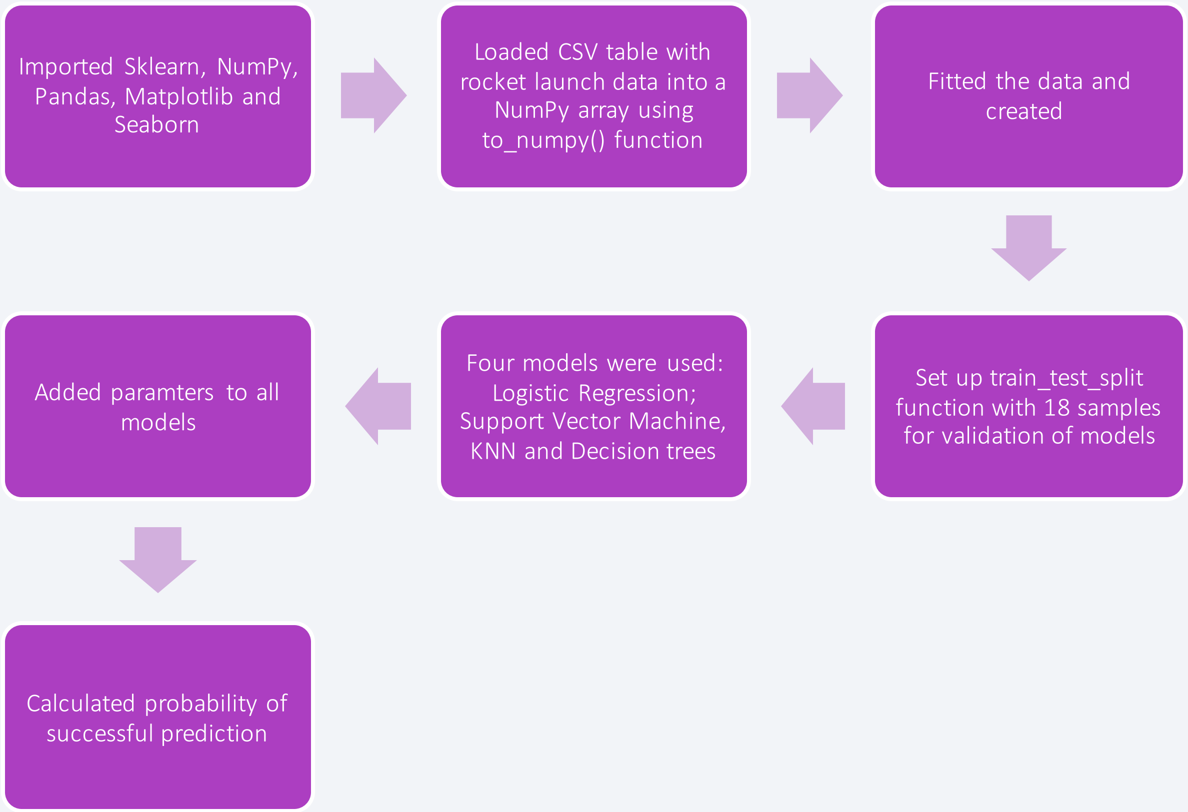
- Names of the unique launch sites in the space mission
- Five records where launch sites begin with the string 'CCA'
- The total payload mass carried by boosters launched by NASA (CRS)
- The average payload mass carried by booster version F9 v1.1
- Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Total number of successful and failure mission outcomes
- Names of the booster_versions which have carried the maximum payload mass. Use a subquery

The Jupyter notebook can be accessed on Github by clicking [here]

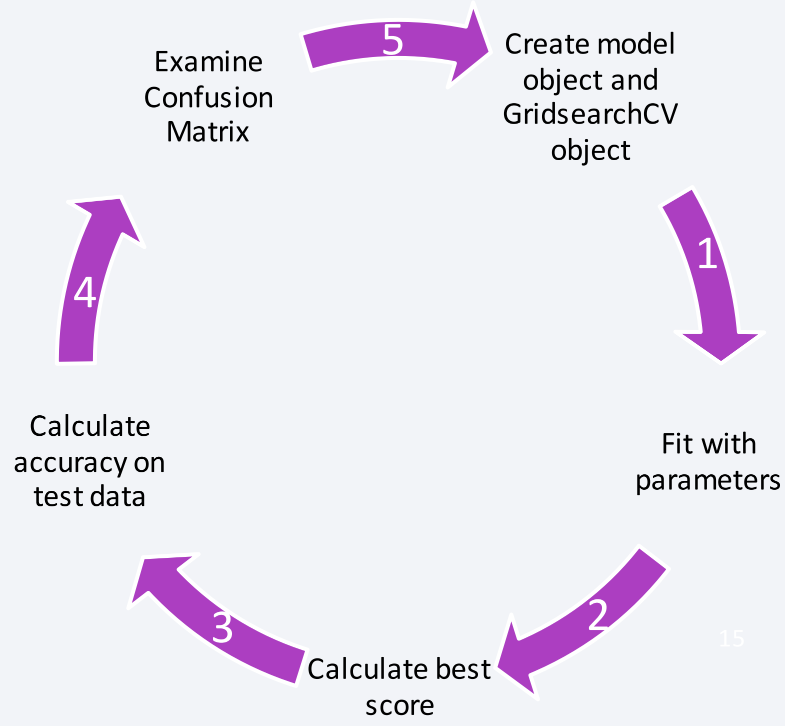
Interactive Map with Folium

- To identify the outcome for each launch site, the positive/negative outcomes were divided and marked with green and red respectively. Furthermore, the outcome marks for each launch pad were clustered.
- Lines were used to point out the distances between locations around the launch sites.
- Circles were used in order to observe where the launch sites actually are on the map
- Click here for the Jupter notebook on Github

Predictive Analysis (Classification)



Lifecycle of Prediction Models



Github link for Jupyter Notebook:
https://github.com/ADGit-cmyk/NR/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

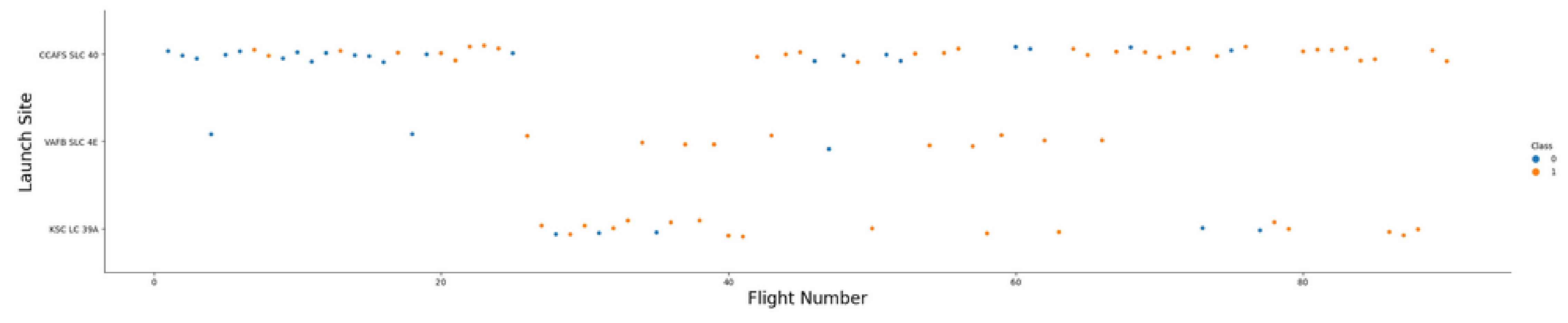
RESULTS

- A total of 90 launches were observed
- The decision-tree classifier method had the highest prediction accuracy at 88.75%
- The orbits ES-L1, GEO, HEO, SSO had success rates of 100%
- All cases where boosters did not land were accurately predicted by all models
- Standard deviation between Regression Model, SVM and K-Nearest Neighbour was ~0.1%
- The success rate has dropped from 2019 to 2020, but it is still significantly higher compared to the first 3 years of launches

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

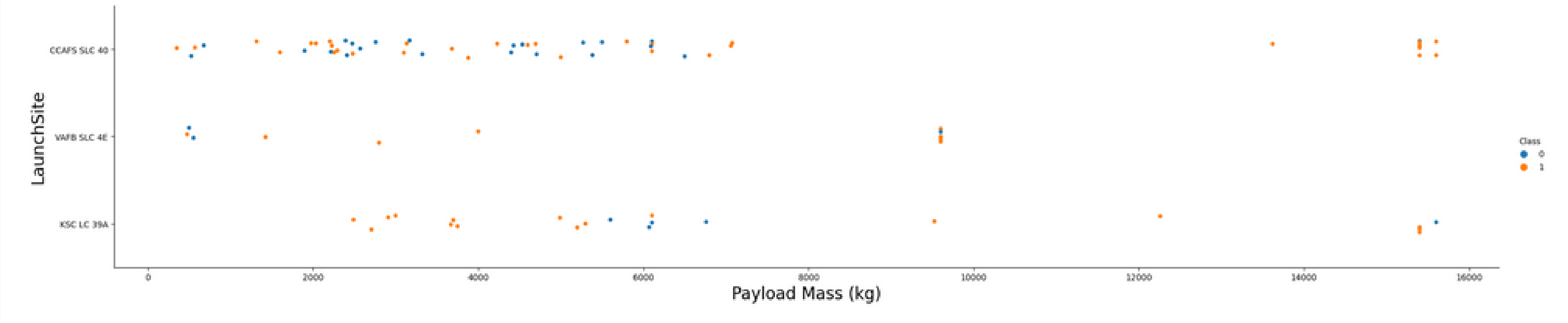


There is an upward trend in successful flights (class 1) as the flight number increases

The trend is positive regardless of the launch site

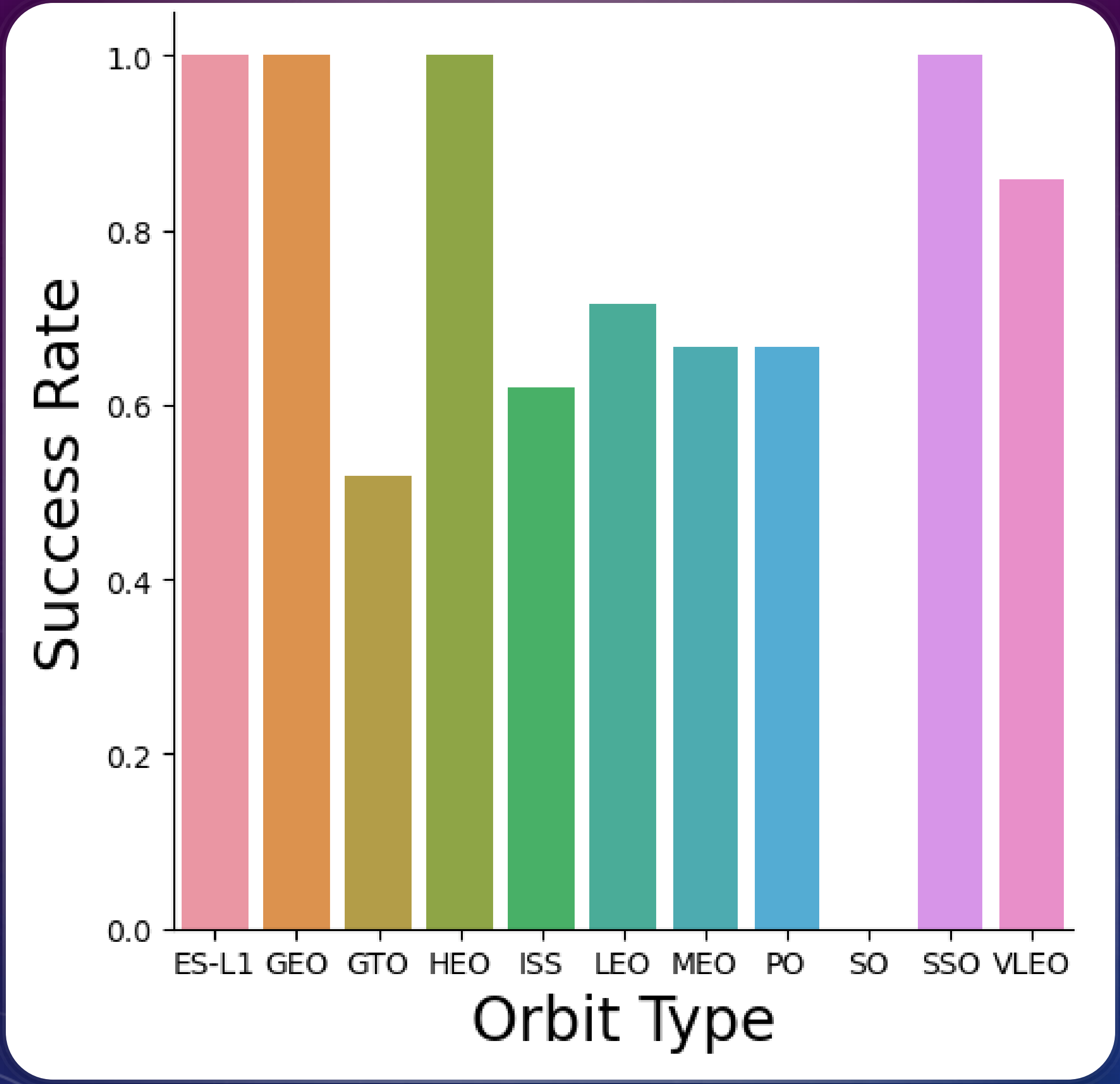
Launch site CCAFS SLC 40 has the largest number of successful launches

Payload vs. Launch Site



- Rockets launched from CCAFS SLC 40 pads with a payload above 8000kg are relatively more successful compared to those launched from KSC LC 39A. The opposite is true for a payload mass of less than 8000kg

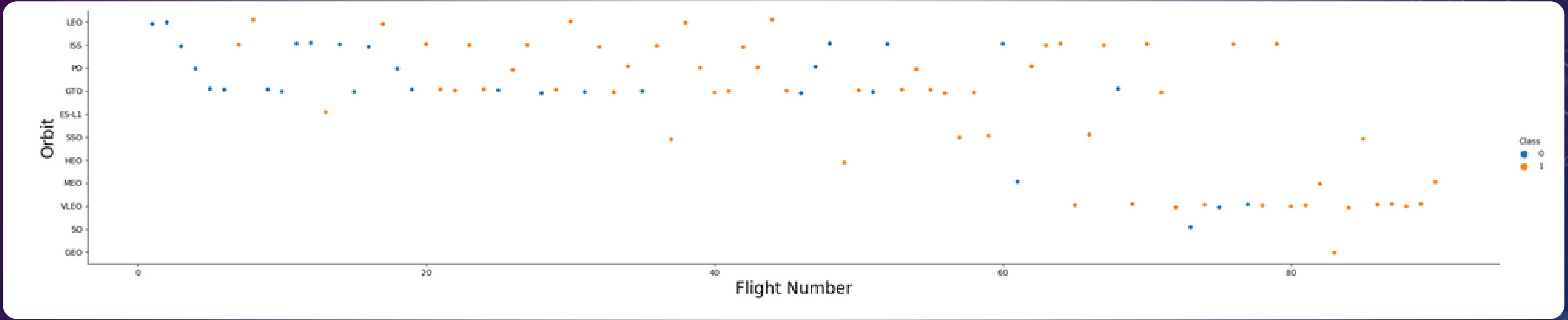
- A correlation exists between higher payloads and successful launches and retrievals of 1st booster.



SUCCESS RATE VS. ORBIT TYPE

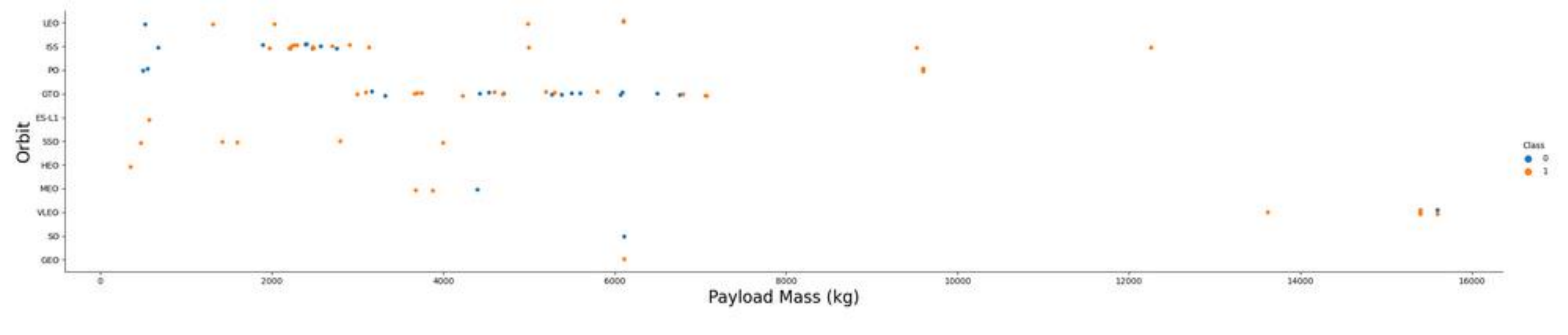
- Only 1 orbit type has a failure rate of 0%, while 4 orbit types have a success rate of 100%
- GEO and SO only have 1 flight sample each which makes it difficult to compare it to other values

FLIGHT NUMBER VS. ORBIT TYPE

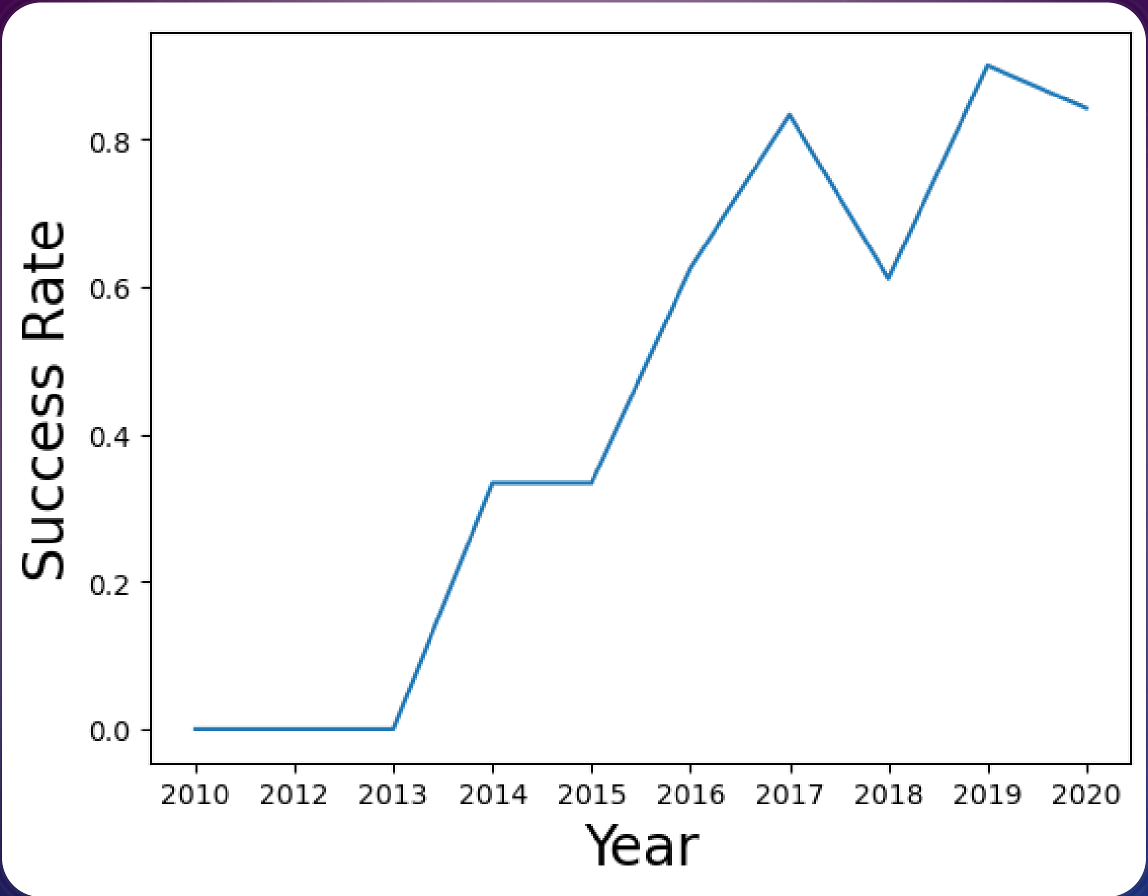


- The low number of flights in SO and GEO orbit can have a negative impact when predicting future success rate for the specific orbits

Payload vs. Orbit Type



- Mass payloads between 14,000kg and 16,000kg are more successful when launched towards VLEO compared to launches with a mass payload between 0kg and 2000kg for PO orbit



LAUNCH SUCCESS YEARLY TREND

- The success rate has dropped from 2019 to 2020, but it is still significantly higher compared to the first 3 years of launches and slightly above 2017 level

ALL LAUNCH SITE NAMES

```
In [7]: %sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL;

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

Out[7]:

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

- According to the query, SpaceX has used 4 launch sites so far:
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40

LAUNCH SITE NAMES BEGIN WITH 'CCA'

```
In [12]: %sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;

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

Out[12]:

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

- Limiting the results to 5, we can see the records from which pads the rockets were launched from, with name beginning with the letters 'CCA'

Total Payload Mass

```
In [17]: %sql SELECT SUM(PAYLOAD_MASS__KG_) as Total_Payload_Mass from SPACEXTBL
 * sqlite:///my_data1.db
Done.
```

Out[17]:

Total_Payload_Mass
619967

- SpaceX has carried almost 620t of payload to space according to the recorded data

Average Payload Mass by F9 v1.1

```
In [18]: %sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';

 * sqlite:///my_data1.db
Done.

Out[18]:
```

AVG_PAYLOAD
2928.4

- The average payload mass for booster version F9 v1.1 is 2928.4kg, almost 4 tons.

Total Number of Successful and Failure Mission Outcomes

```
sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;
```

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

Mission_Outcome	QTY
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- According to the query, there have been 100 successful missions with only 1 failure

Boosters Carried Maximum Payload

```
n [35]: %sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
         * sqlite:///my_data1.db
        Done.
ut[35]:
```

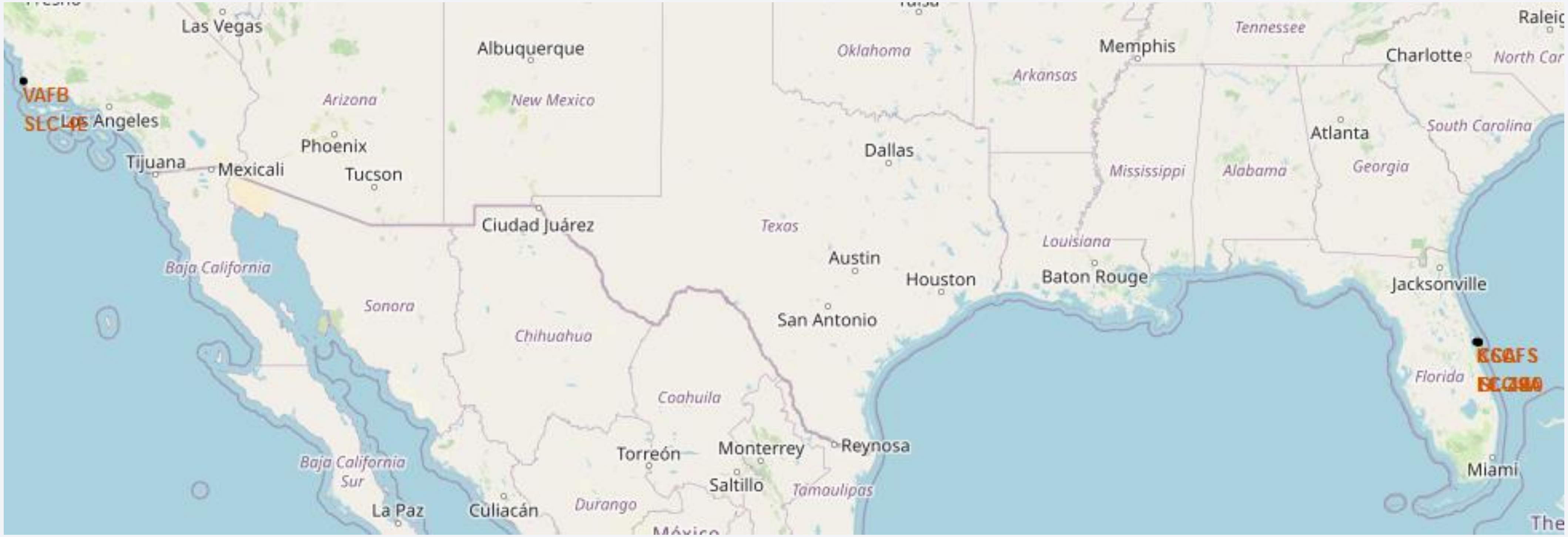
Booster_Version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

- Booster F9 B5 B1048.4 has carried the highest payload to space

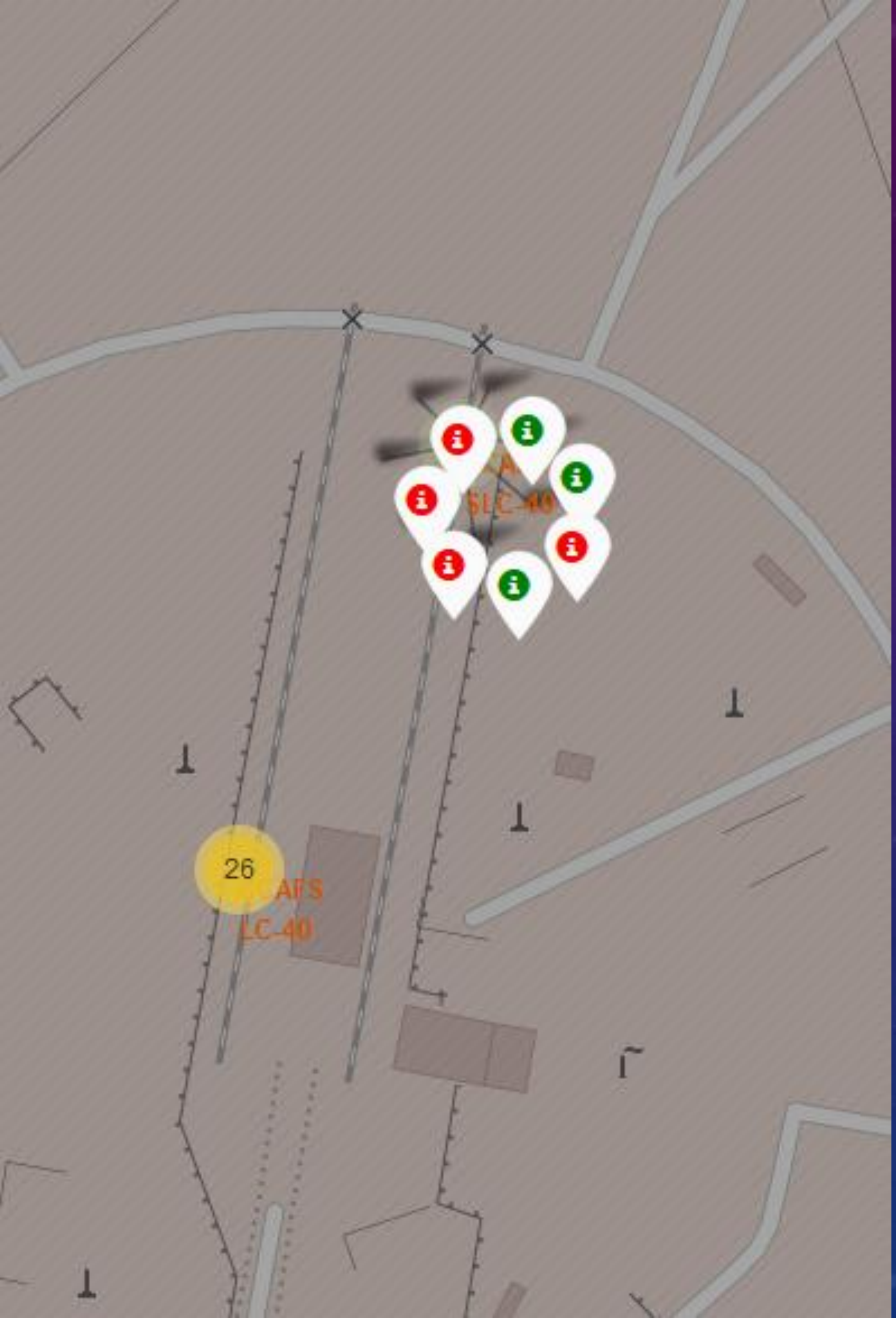
Section 3

Launch Sites Proximities Analysis

Location of Space launch pads – On Map



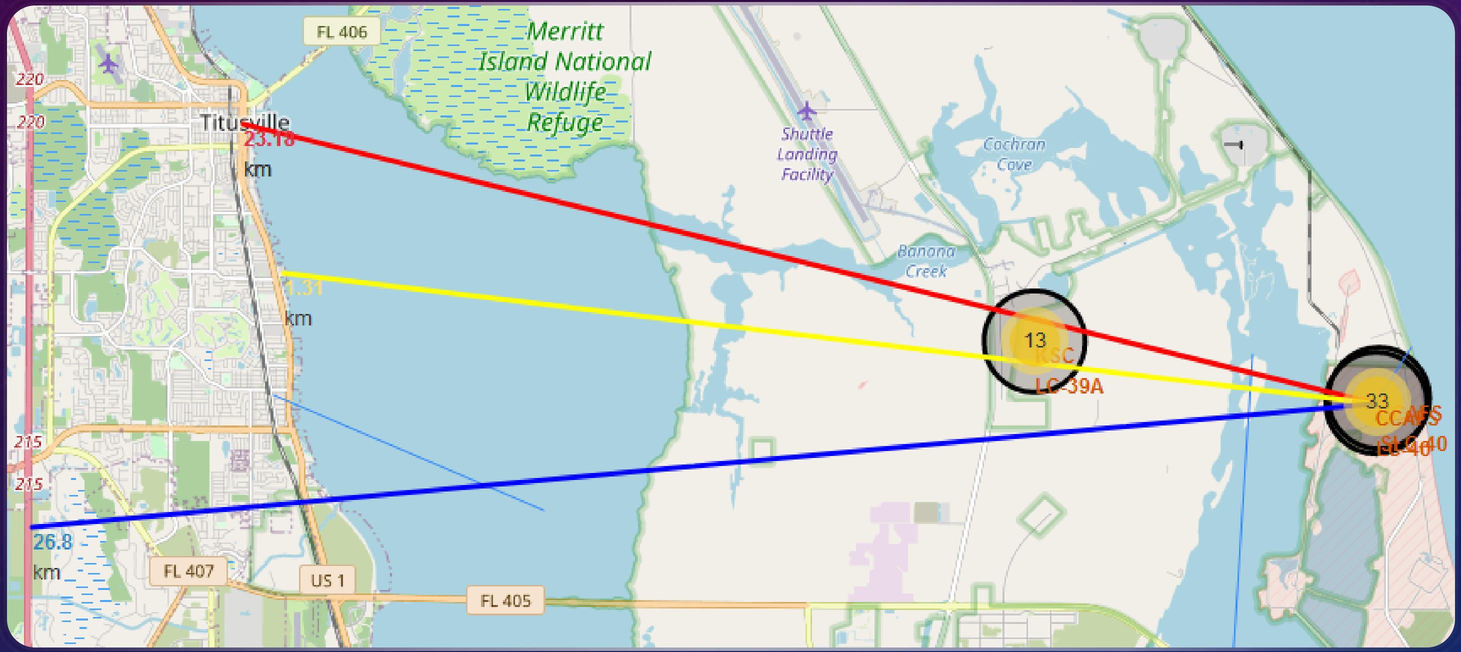
- Both launch pad clusters are on the Eastern and Western coasts respectively



OUTCOMES - ON MAP

- The green-colored marks represent successful launches, while the red represent failures.
- The cluster is tied to the launchpad CCAFS SLC-40

PROXIMITY OF LOCATIONS TO LAUNCHPAD – ON MAP



- The distance between CCAFS SLC-40 launch site and the city of Titusville, I-95 highway and Titus postal office are 23.18km, 26.8km and 21.31km respectively

Section 5

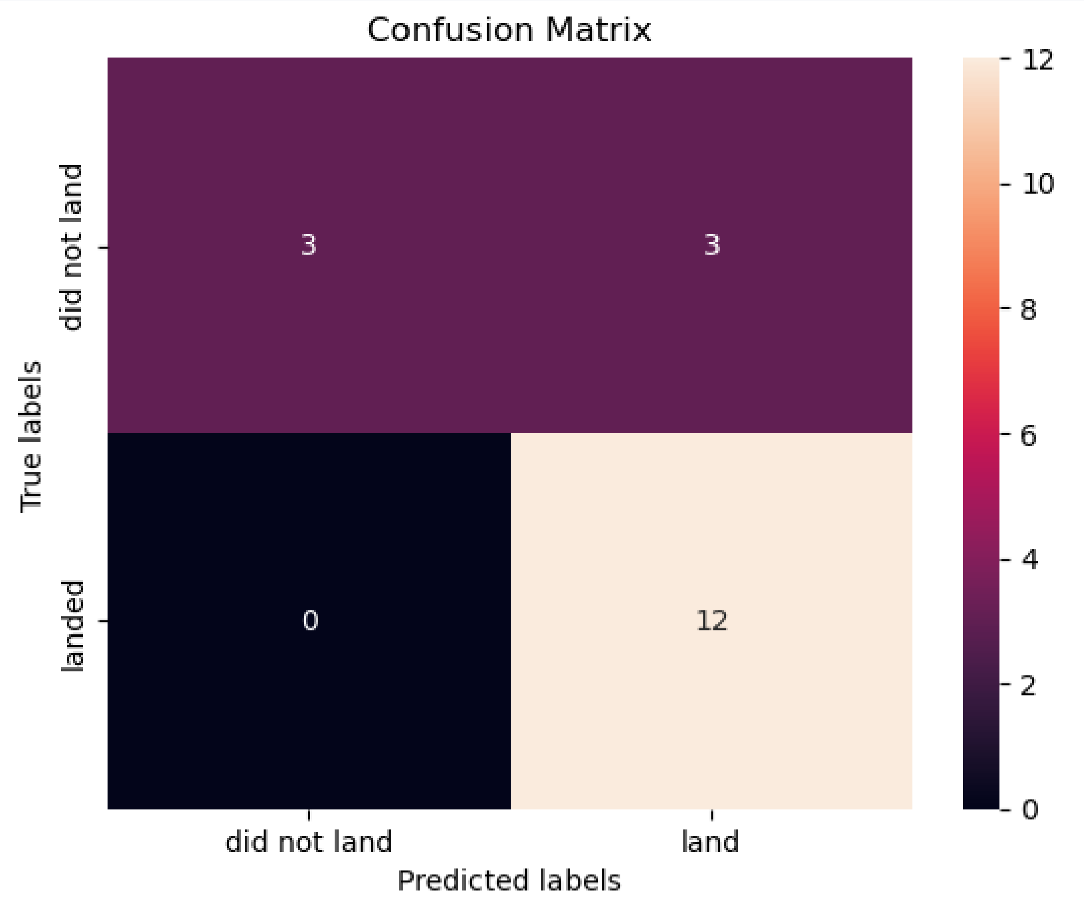
Predictive Analysis (Classification)

CLASSIFICATION ACCURACY

Model	Accuracy	TestAccuracy
LogReg	0.84643	0.83333
SVM	0.84821	0.83333
Tree	0.8875	0.83333
KNN	0.84821	0.83333

- The Logistic Regression model had the lowest overall accuracy while the Decision-Tree Classification model had the highest.

Confusion Matrix



- The matrix shows that the model accurately predicted all launches from which boosters did not land, while it had an 80% accuracy rate in predicting the launches where the boosters did land.

CONCLUSIONS

- The Decision-Tree Classification model is the most optimized for predicting future outcomes based on information gather from this dataset
- Given the annual growth of successful launches since 2013 and a predictive model with a high accuracy, we expect more launches where the boosters will be successfully retrieved
- Launchpad CCAFS SLC 40 has the largest number of successful launches

Thank you!