IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

SpaceX Falcon 9 Landing Analysis

PRAVEENKUMAR S
June 23, 2023

Outline

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

Executive Summary

- **Summary of methodologies**

- Data has been collected from the SpaceX API and the SpaceX Wiki Page.

- Data was explored through SQL, Visualizations, Folium Maps and Dashboards.

- Data was standardized and used GridSearchCV to find the parameters used for Machine Learning models and created a visual accuracy score of all models.

- **Summary of all the results**

- Four Machine learning models were produced, namely, Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors.

- All produced similar results with accuracy rate of about 83.33%. All models predicted successful landings.
More data is needed for better model determination and accuracy.

Introduction

- SpaceX launches Falcon 9 rockets at a cost of around $62m. This is considerably cheaper than other providers (which usually cost upwards of $165m), and much of the savings are because SpaceX can land, and, then re-use the first stage of the rocket.
- If we can make predictions on whether the first stage will land, we can determine the cost of a launch, and, use this information to assess whether or not an alternate company should bid and SpaceX for a rocket launch.
- This project will ultimately predict if the Space X Falcon 9 first stage will land successfully.

Section 1

Methodology

Methodology

- Executive Summary
- Data collection methodology:
  - Making GET requests to the SpaceX API and web scraping from the Wiki Page.
- Perform data wrangling
  - Classifying true landings as successful and unsuccessful otherwise
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Used sckit-learn to standardize the data
  - Split the data into training & testing and Train different classification models
  - Tuned models using GridSearchCV

Data Collection

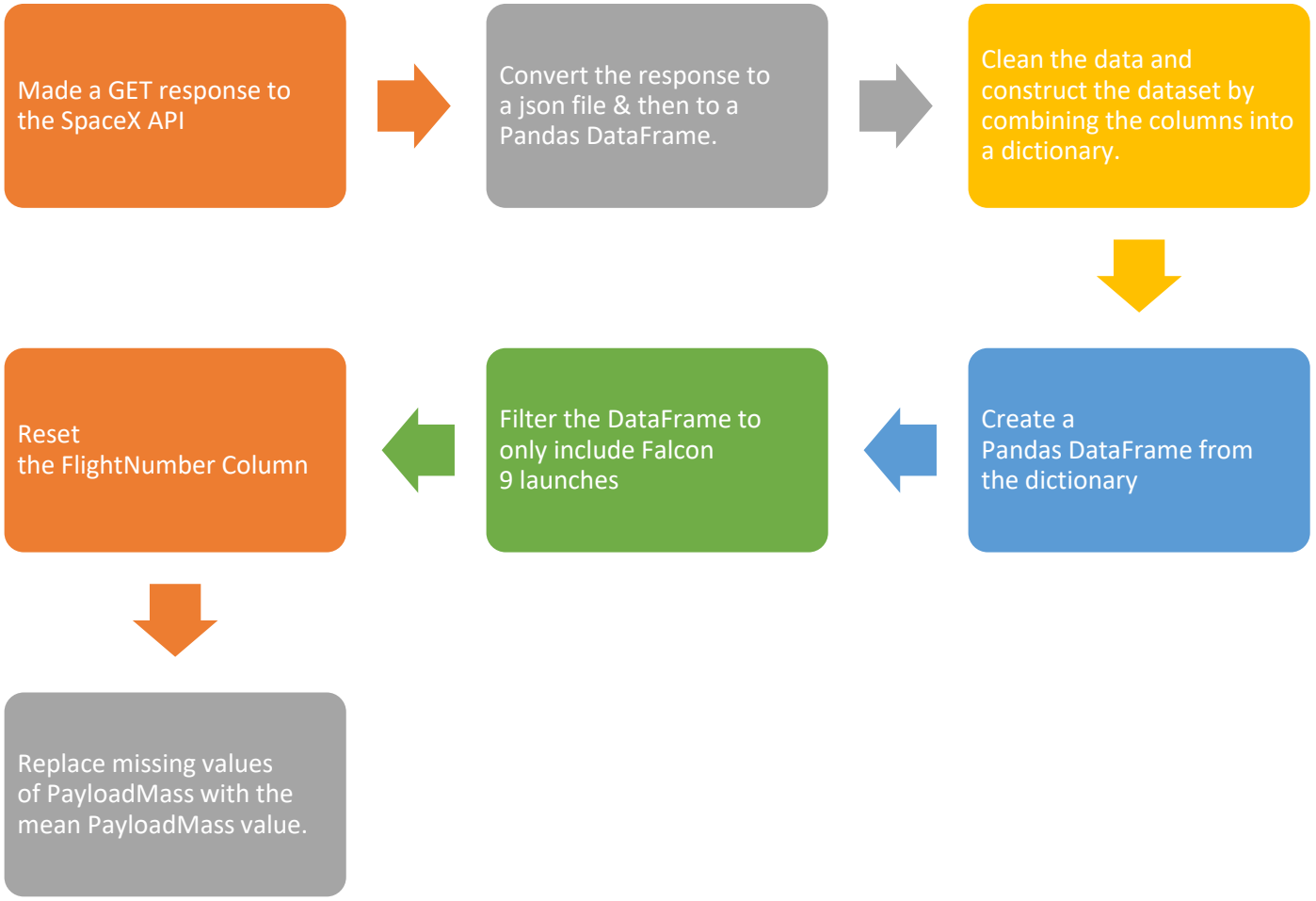
- Data collection process involved a combination of API requests from SpaceX API and web scraping data from a table in Space X's Wikipedia entry.
- The following slide will show the flowchart of data collection from API and the one after will show the flowchart of data collection from web-scraping.
- **Space X API Data Columns:**

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

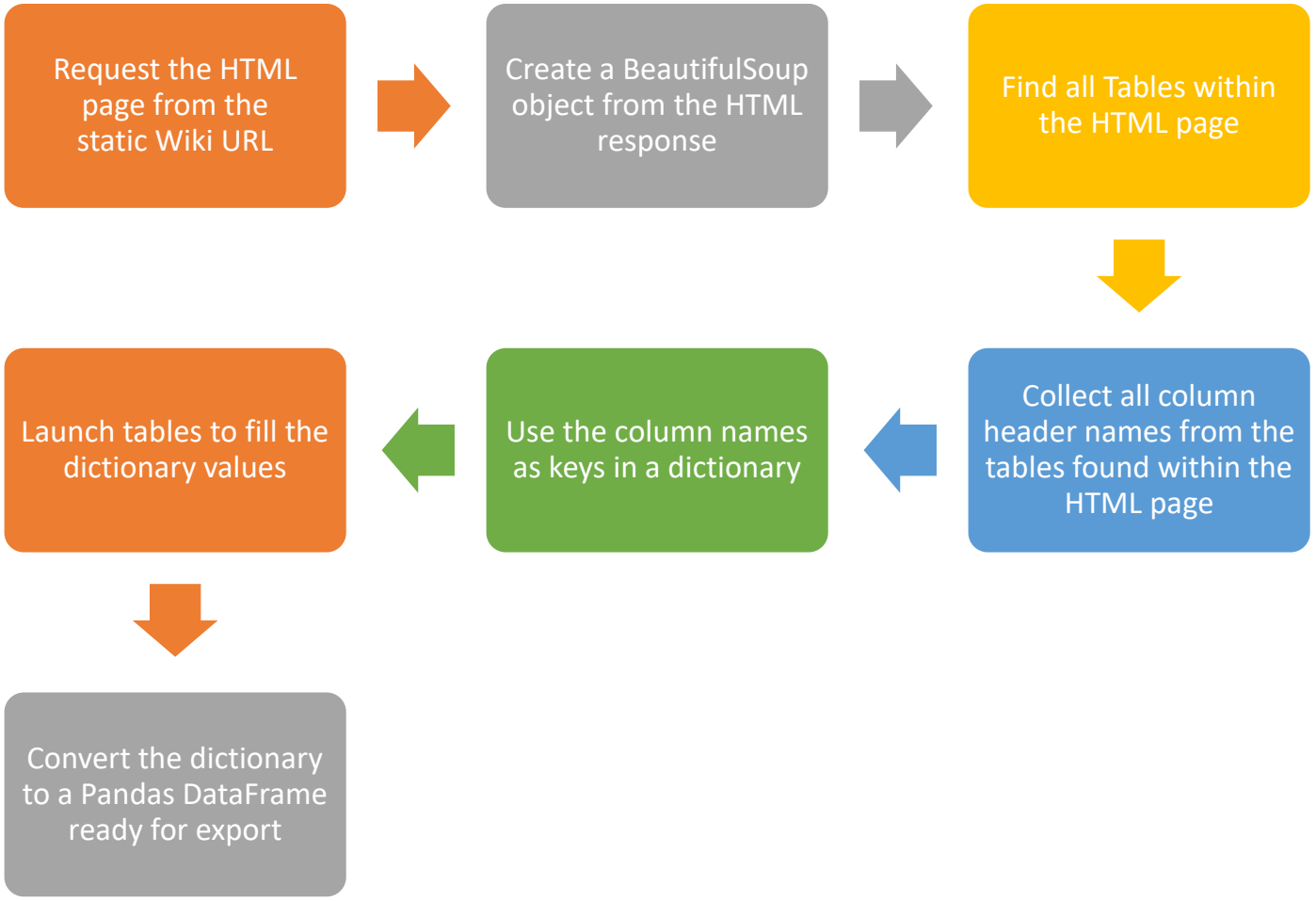
- **Wikipedia Web Scrape Data Columns:**

Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



Data Collection - Scraping



Data Wrangling

Created a training label with the landing outcomes, successful --> 1 and failure --> 0.

Outcome column has two components: ‘Mission Outcome’ ‘Landing Location’

New training label column ‘class’ with a value of 1 if ‘Mission Outcome’ is True and 0 otherwise.

Value Mapping:

- True Ocean – mission outcome was successfully landed to a specific region of the ocean
- False Ocean – mission outcome was unsuccessfully landed to a specific region of the ocean
- True RTLS – mission outcome was successfully landed to a ground pad
- False RTLS – mission outcome was unsuccessfully landed to a ground pad.
- True ASDS – mission outcome was successfully landed to a drone ship
- False ASDS – mission outcome was unsuccessfully landed to a drone ship.
- None ASDS and None None – these represent a failure to land.

EDA with Data Visualization

Scatter charts were produced to analyze the relation between

- Flight Number and Launch Site
- Payload and Launch Site
- Orbit Type and Flight Number
- Payload and Orbit Type

Bar Charts were produced to visualize the relationship between

- Success Rate and Orbit Type

Line Charts were produced to visualize the relationship between

- Success Rate and Year (i.e. the launch success yearly trend)

EDA with SQL

- Loaded data in IBM Db2 Instance
- Queried SQL in Notebook through SQL Alchemy
- The SQL queries performed on the data set were used to:
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string ‘CCA’
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display the average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome on a ground pad was achieved
  - List the names of the boosters which had success on a drone ship and a payload mass between 4000 and 6000 kg
  - List the total number of successful and failed mission outcomes
  - List the names of the booster versions which have carried the maximum payload mass
  - List the failed landing outcomes on drone ships, their booster versions, and launch site names for 2015
  - 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

Build an Interactive Map with Folium

Marked all the Launch sites on a map

- Initialise the map using a Folium 'Map' object
- Added a folium.Circle and folium.Marker for each launch site on the launch map

Mark the success/failed launches for each site on a map

- As many launches have the same coordinates, it makes sense to cluster them together
- Before clustering them, assign a marker colour of successful (class = 1) as green, and failed (class = 0) as red
- To put the launches into clusters, for each launch, add a folium.Marker to the MarkerCluster() object
- Create an icon as a text label, assigning the icon_color as the marker_colour determined previously

Calculate the distances between a launch site to its proximities

- To explore and analyze the proximities of launch sites, calculations of distances between points can be made using the Lat and Long values
- After marking a point using the Lat and Long values, create a folium.Marker object to show the distance
- To display the distance line between two points, draw a folium.PolyLine and add this to the map

Build a Dashboard with Plotly Dash

Pie chart (px.pie()) showing the Total Successful Launches per site

This can be selected to show the distribution of successful landings across sites.

The chart could also be filtered (using a dcc.Dropdown() object) to see the success/failure ratio for an individual site

Scatter graph (px.scatter()) to show the Correlation between outcome (success or not) and Payload mass (kg)

This could be filtered (using a RangeSlider() object) by ranges of payload masses

It could also be filtered by booster version

Predictive Analysis (Classification)

We split the data into training and testing data using the function train_test_split

The training data is divided into validation data, a second set used for training data

The models are trained and hyperparameters are selected using the function GridSearchCV on:

Logisitic Regression

Support Vector Machine

Decision Tree

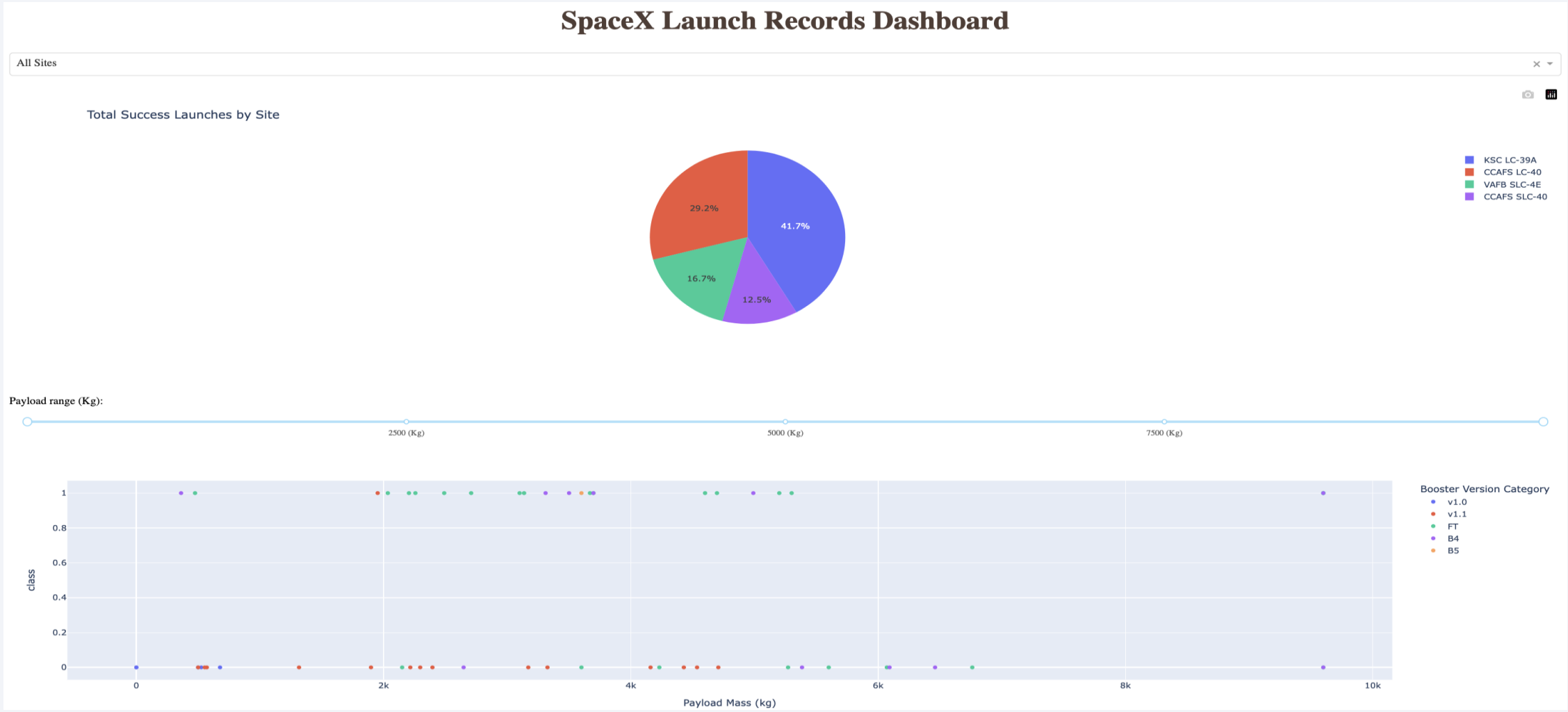
K-Nearest Neighbour

Confusion Matrix for all models is rendered and examined

Reviewing the accuracy of the chosen algorithms

The model with the highest accuracy score is determined as the best performing model

Results

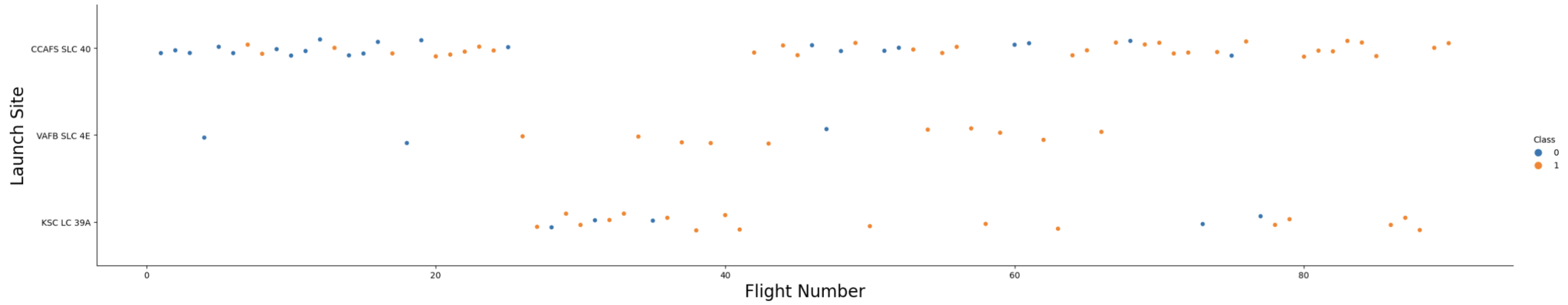

SpaceX Launch Records Dashboard
All Sites
Total Success Launches by Site
29.2%
41.7%
16.7%
12.5%
KSC LC-39A
CCAFS LC-40
VAFB SLC-4E
CCAFS SLC-40
Payload range (Kg):
2500 (Kg)
5000 (Kg)
7500 (Kg)
class
Payload Mass (kg)
Booster Version Category
v1.0
v1.1
FT
B4
B5

Results

- The previous slide shows the dashboard created using Plotly Dash
- The next section will show the results of EDA with Data Visualization, EDA with SQL, Interactive Map using Folium and finally the results from our model with about 83% accuracy.

Section 2

Insights drawn from EDA



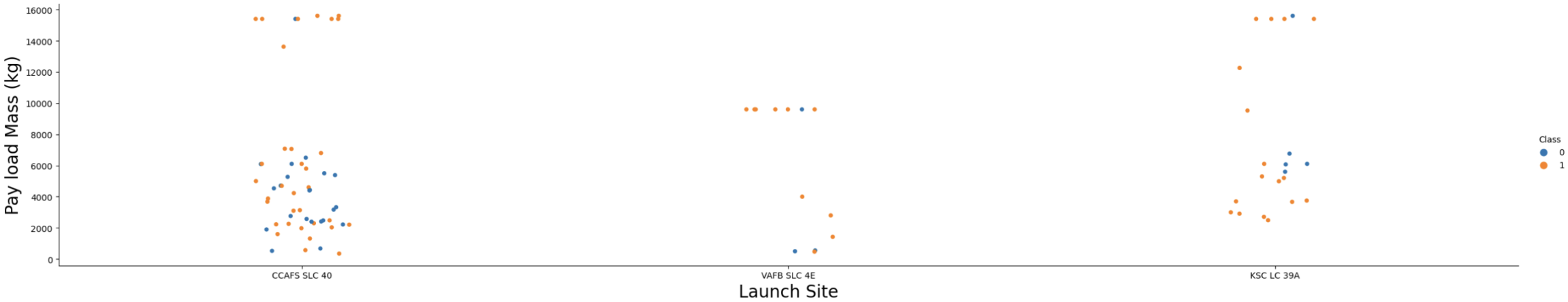
Flight Number v/s Launch Site

Orange indicates a successful launch while blue denotes a failure.

- The scatter plot suggests an increase in Success rate over time
- The earlier flights were launched from CCAFS SLC 40, and were mostly unsuccessful
- The flights from VAFB SLC 4E also show this trend, that, earlier flights were less successful
- No early flights were launched from KSC LC 39A, so the launches from this site are more successful
- Class 1 (Successful Launches) are greater above a flight number of around 30

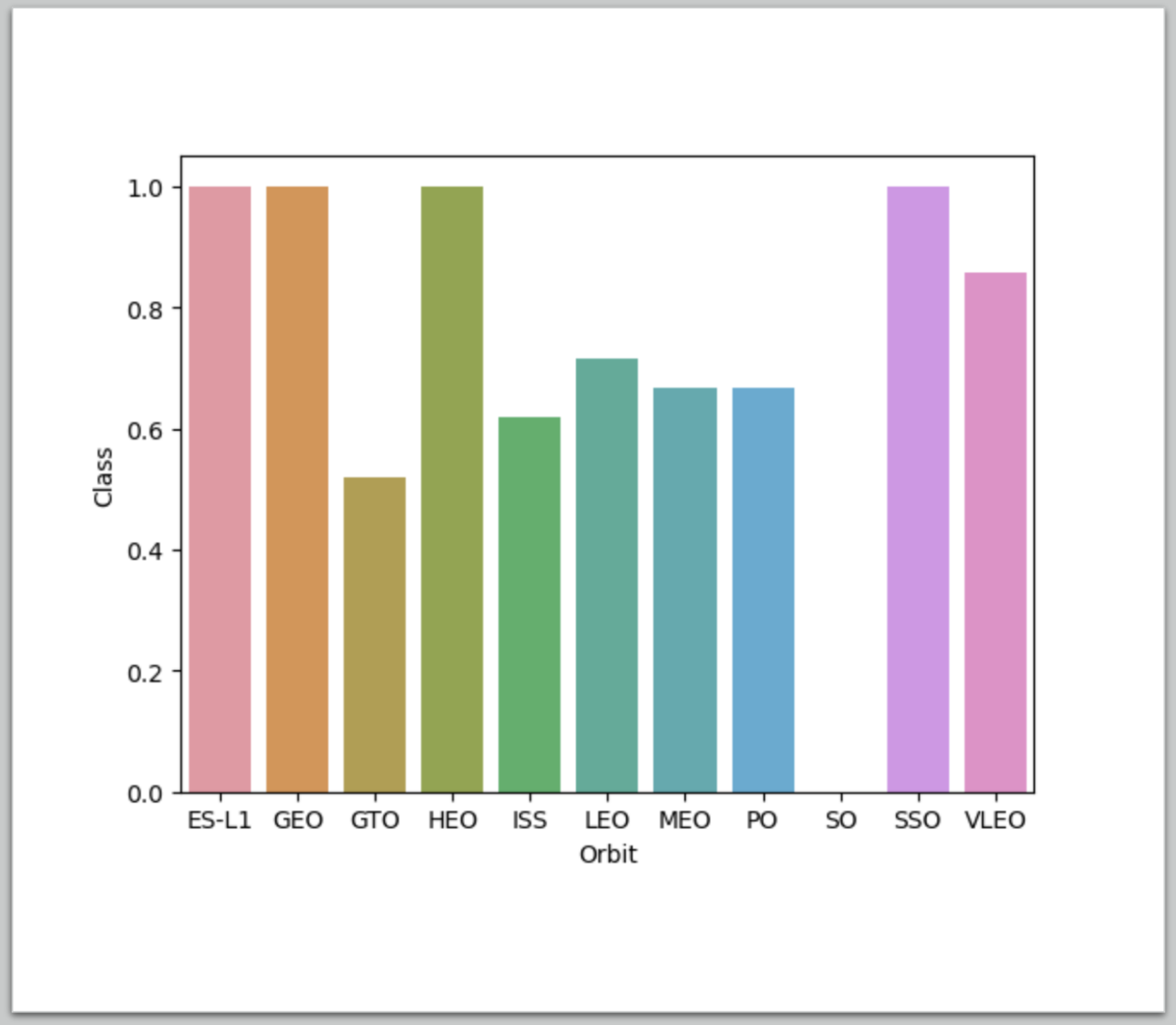
Payload vs. Launch Site

- Payload mass appears to fall mostly between 0 to 7000 kg
- All sites launched a variety of payload masses, with most of the launches from CCAFS SLC 40 with comparatively lighter payloads (with some outliers having payload around 16000 kg)



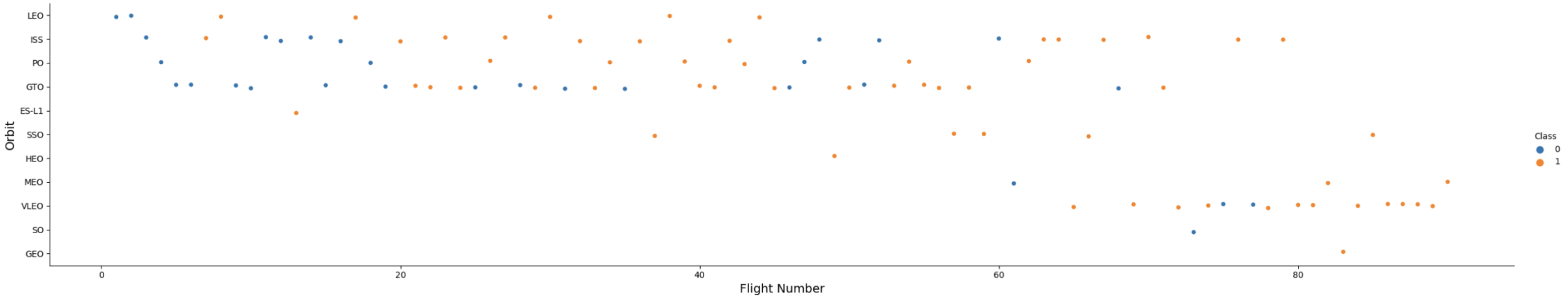
Success Rate v/s Orbit Type

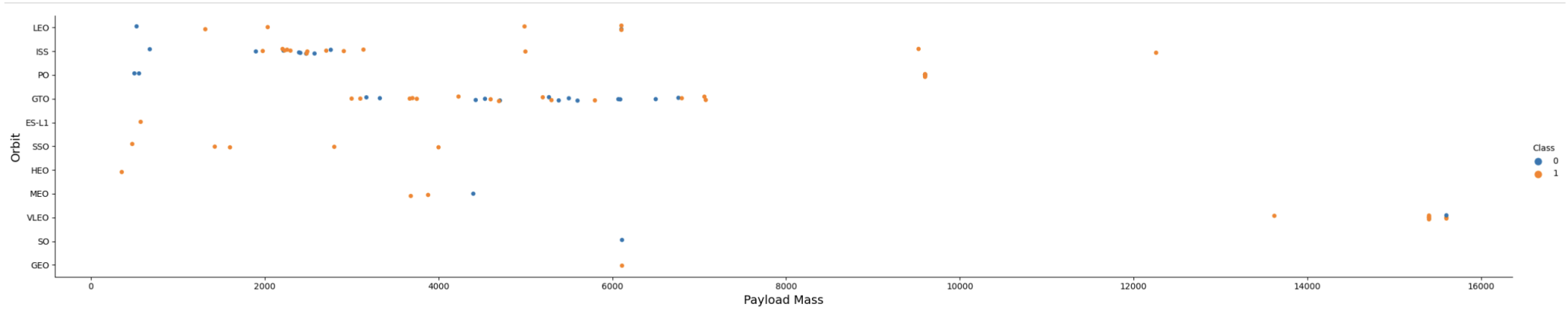
- The bar chart of Success Rate v/s Orbit Type shows that the following orbits have the highest (100%) success rate:
  - ES-L1
  - GEO
  - HEO
  - SSO
- The orbit with the lowest (0%) success rate is:
  - SO



Flight Number v/s Orbit Type

- SpaceX started with LEO orbits which saw moderate success LEO and returned to VLEO in recent launches SpaceX appears to perform better in lower orbits or Sun-synchronous orbits.
- SSO has a 100% success rate over 5 Flights.
- Generally, as Flight Number increases, the success rate increases.



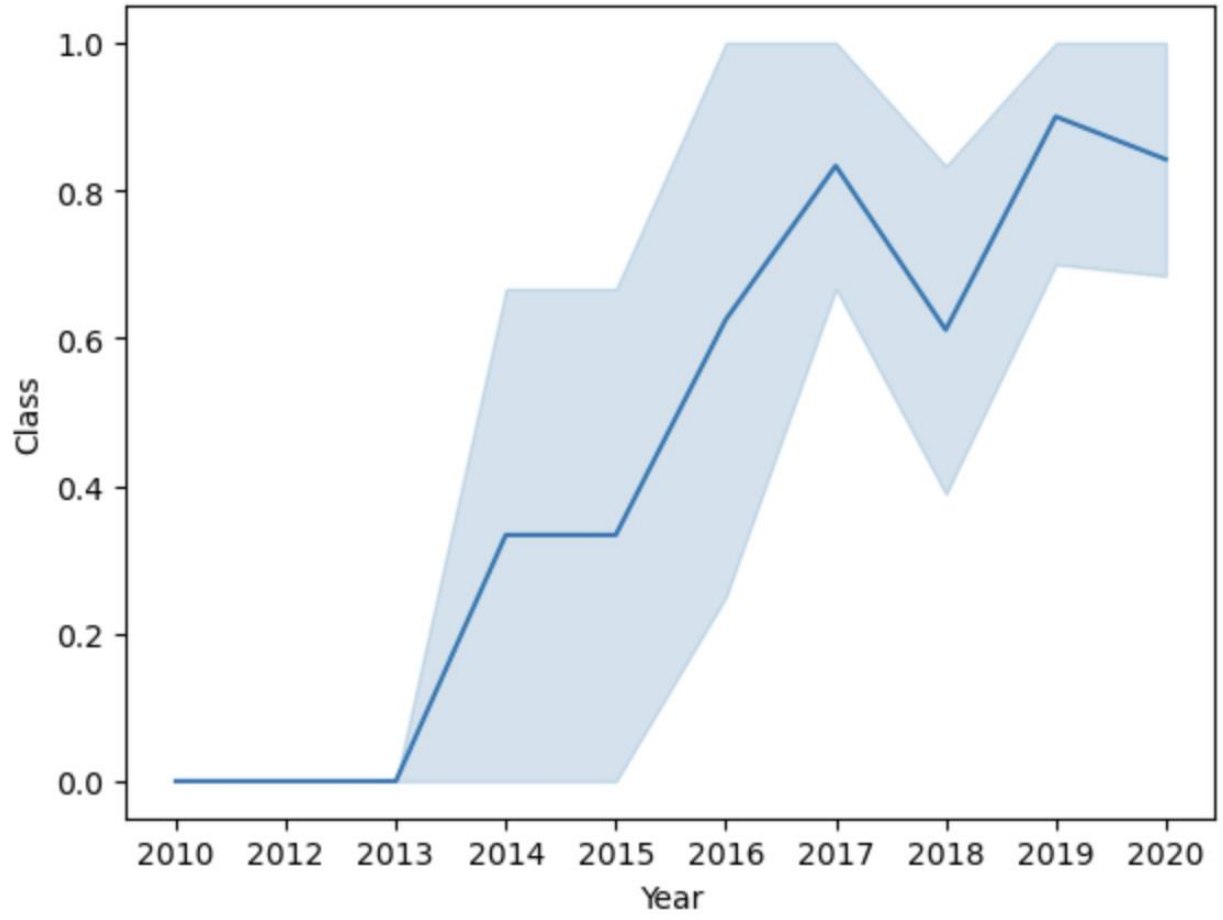


Payload vs. Orbit Type

- The following orbit types have more success with heavy payloads:
  - ISS
  - PO
  - LEO
- VLEO (Very Low Earth Orbit) launches are associated with heavier payloads
- As such, payload mass seems to correlate with the Orbit Type

Launch Success Yearly Trend

- The light blue shading represents the 95% confidence interval.
- Success rate has improved over time starting with the year 2013 and prior to that it was 0.
- Success rate dipped in 2018 marginally.
- After 2016, there has been at least a 50% chance of success.



Average launch success trend over the Years

```
In [10]: %sql SELECT UNIQUE(LAUNCH_SITE) FROM SPACEXTBL;
 * ibm_db_sa://gzj90026:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[10]:

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

All Launch Site Names

Find the names of the unique launch sites

- The query UNIQUE returns only the unique values from the Table which were shown above as the output of the query.

```
In [11]: %sql SELECT LAUNCH_SITE FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
 * ibm_db_sa://gzj90026:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[11]:

launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Launch Site Names Begin with 'CCA'

Find 5 records where launch sites begin with `CCA`

- LIMIT 5 fetches only 5 records, and the LIKE keyword is used with the wild card ‘CCA%’ to retrieve string values beginning with ‘CCA’.

Total Payload Mass

Calculate the total payload carried by boosters from NASA

The SUM keyword is used to calculate the total of the LAUNCH column and the SUM keyword then filters the results to only boosters from "NASA (CRS)"

CRS stands for Commercial Resupply Services which indicates that these payloads were sent to the International Space Station.

In [12]:
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD_MASS FROM SPACEXTBL \
    WHERE CUSTOMER = 'NASA (CRS)';
```

```
 * ibm_db_sa://gzj90026:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[12]:

total_payload_mass
45596

Average Payload Mass by F9 v1.1

Calculate the average payload mass carried by booster version F9 v1.1

- The AVG keyword is used to calculate average of the PAYLOAD_MASS__KG_ column, and WHERE keyword filters the results to only the F9 v1.1 booster version.
- Average payload mass of F9 1.1 is on the low end of our payload mass range

```
In [13]: %sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVERAGE_PAYLOAD_MASS FROM SPACEXTBL \
             WHERE BOOSTER_VERSION = 'F9 v1.1';
 * ibm_db_sa://gzj90026:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
Out[13]: average_payload_mass
         2928
```

First Successful Ground Landing Date

Find the dates of the first successful landing outcome on ground pad

- The MIN keyword is used to calculate the minimum of the DATE column, i.e. the first date, and the WHERE keyword filters the results to only the successful ground landings.
- First ground pad landing wasn't until the end of 2015.

```
In [14]: %sql SELECT MIN(DATE) AS FIRST_SUCCESSFUL_GROUND_LANDING FROM SPACEXTBL \
             WHERE LANDING__OUTCOME = 'Success (ground pad)';

          * ibm_db_sa://gzj90026:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
         Done.

Out[14]: first_successful_ground_landing
                                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 WHERE keyword is used to filter the results to include only those that satisfy both of the conditions in the brackets. The BETWEEN keyword allows for values between 4000 and 6000 values to be selected in this case.

In [15]:
```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL \
    WHERE (LANDING__OUTCOME = 'Success (drone ship)') AND (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000);
```

```
 * ibm_db_sa://gzj90026:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[15]:

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 COUNT keyword is used to calculate the total number of mission outcomes, and the GROUP BY keyword is used to group these results by the mission outcome.
- 1 launch had an unclear payload status.
- SpaceX achieves its mission outcomes 98% (99/101) of the times

```
In [16]: %sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER FROM SPACEXTBL GROUP BY MISSION_OUTCOME;
 * ibm_db_sa://gzj90026:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[16]:

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

Boosters Carried Maximum Payload

List the names of the booster which have carried the maximum payload mass

- The SELECT statement within the brackets finds the maximum payload, and this value is used in the WHERE condition. The DISTINCT keyword is then used to retrieve only unique booster versions.

```
In [17]: %sql SELECT DISTINCT(BOOSTER_VERSION) FROM SPACEXTBL \
             WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
 * ibm_db_sa://gzj90026:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0
Done.
```

Out[17]:

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

2015 Launch Records

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

- The WHERE keyword is used to filter the results for failed landing outcomes, AND for the year 2015.
- There are 2 such occurrences

```
In [18]: %sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL \
             WHERE (LANDING__OUTCOME = 'Failure (drone ship)') AND (EXTRACT(YEAR FROM DATE) = '2015');
          * ibm_db_sa://gzj90026:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases
         Done.
```

Out[18]:

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

In [19]:
```
%sql SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS TOTAL_NUMBER FROM SPACEXTBL \
    WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' \
    GROUP BY LANDING__OUTCOME \
    ORDER BY TOTAL_NUMBER DESC;
```

```
 * ibm_db_sa://gzj90026:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[19]:

landing__outcome	total_number
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

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

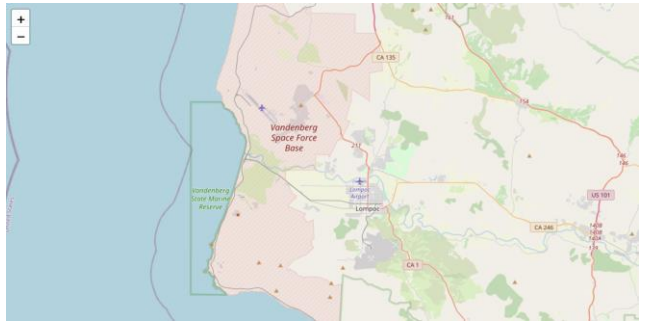
- The WHERE keyword is used with the BETWEEN keyword to filter the results to dates only within those specified. The results are then grouped and ordered, using the keywords GROUP BY and ORDER BY, respectively and DESC is used to specify descending order.
- There were 8 successful launches during this time period.

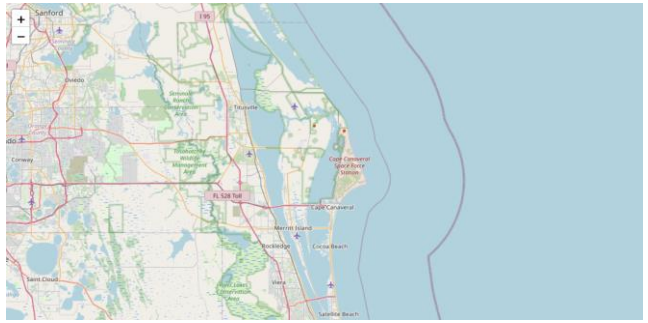
Section 3

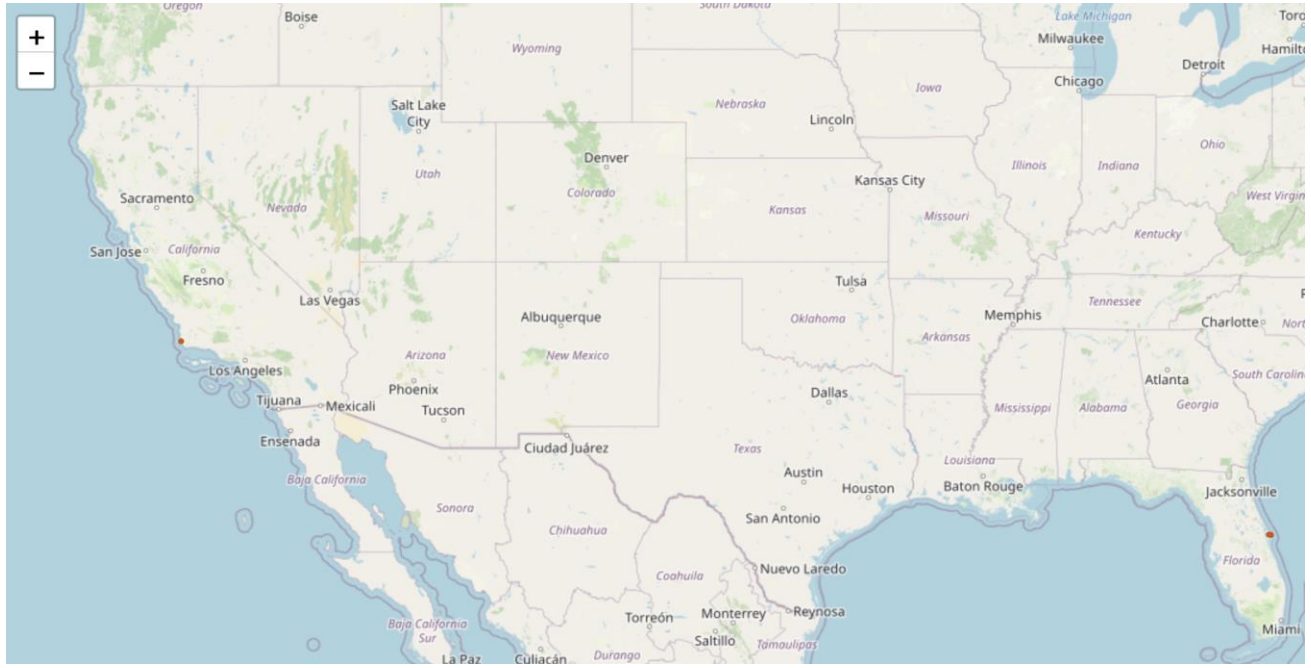
Launch Sites Proximities Analysis

SpaceX Launch Sites in United States

- The first map shows the launch sites in the Vandenberg Space Force Base.
- The second map shows the launch sites in the Cape Canaveral Space Force Base.
- The last map shows the all US launch sites.
- All these launch sites are near the coast
- The launch sites are in pinpointed in red

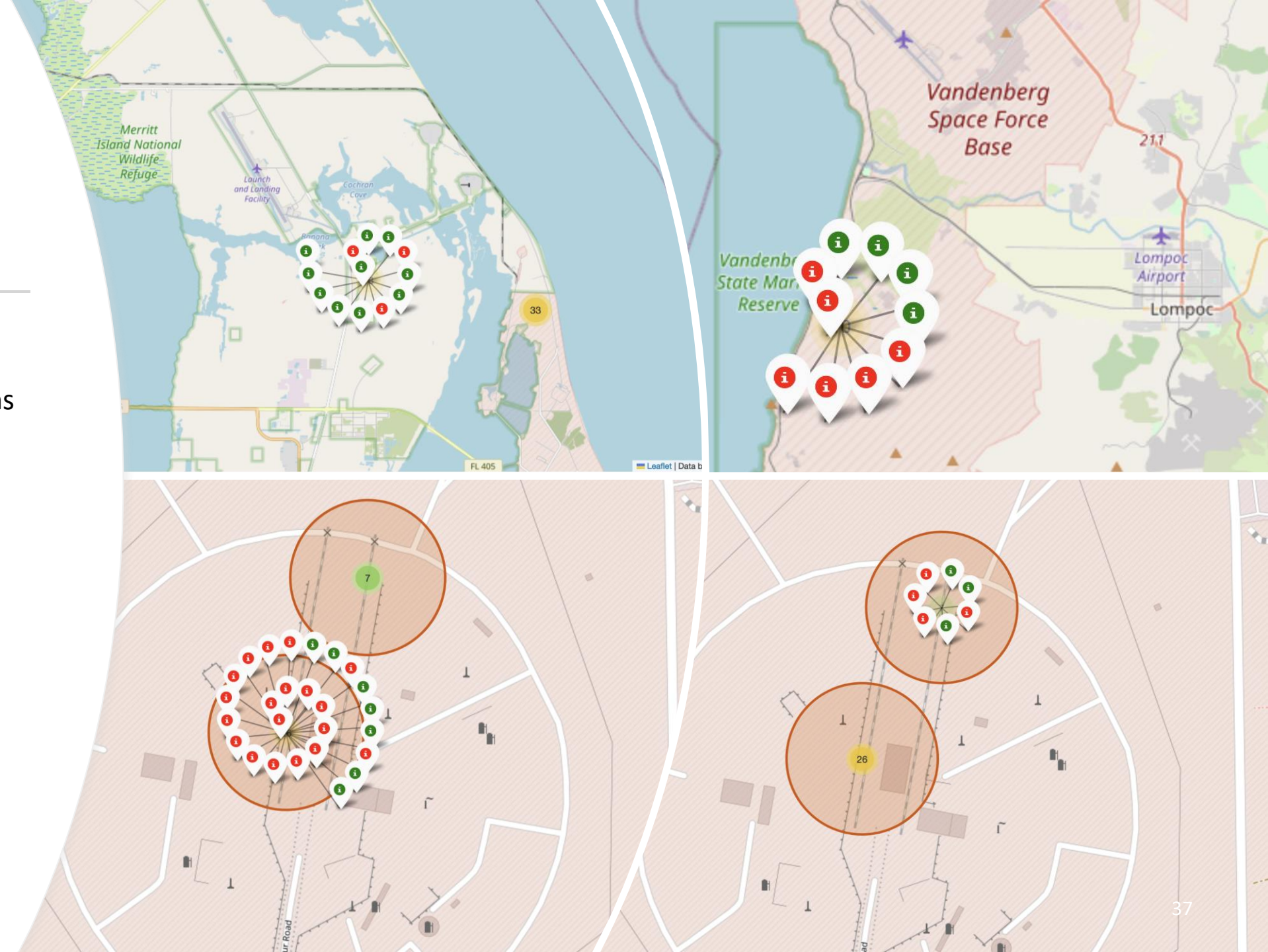






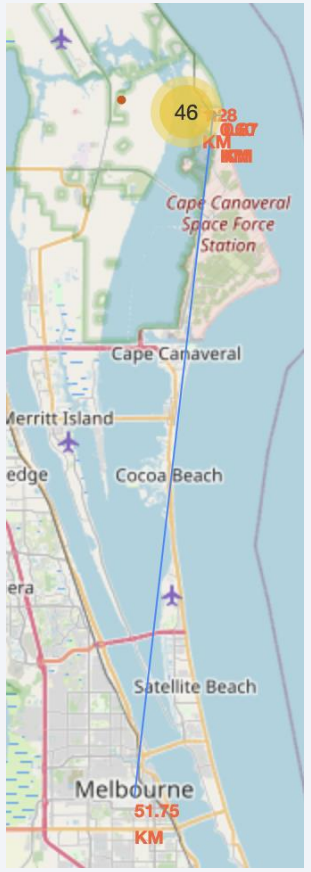
Success/Failed launches for each site on the map

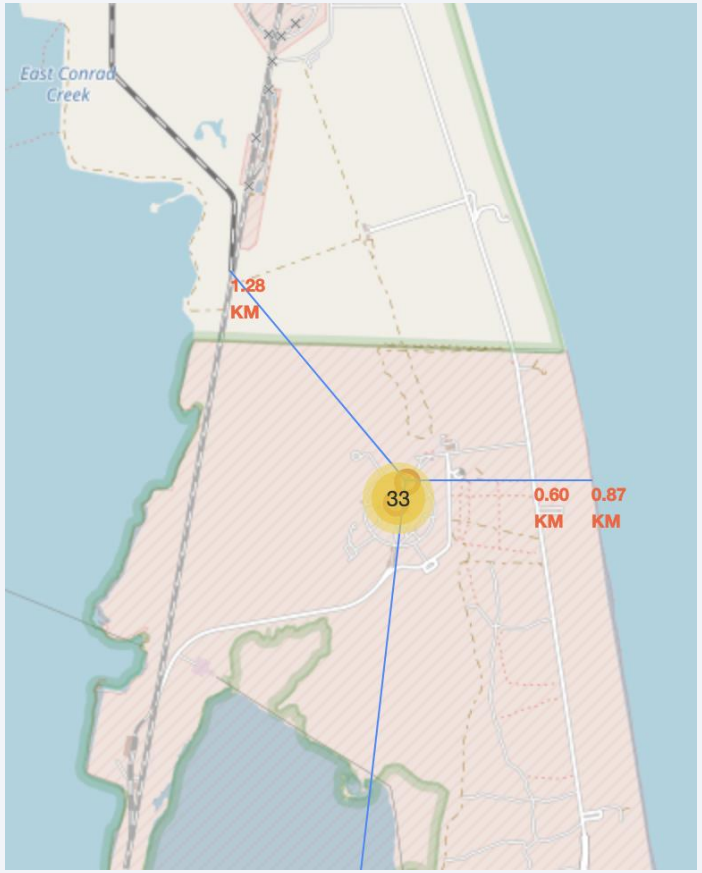
- Launches have been grouped into clusters, and annotated with green icons for successful launches, and red icons for failed launches.
- Launches for both the bases have been depicted in the pictures to the right.



Distances between a launch site to its proximities

- Here, the CCAFS SLC-40 launch site has been used as an example site from which we can understand more about the placement of launch sites and it's proximity with the Coastline (0.87 km), Highway (0.6 km), railway (1.28 km), and the nearest city is 51.75 km away.

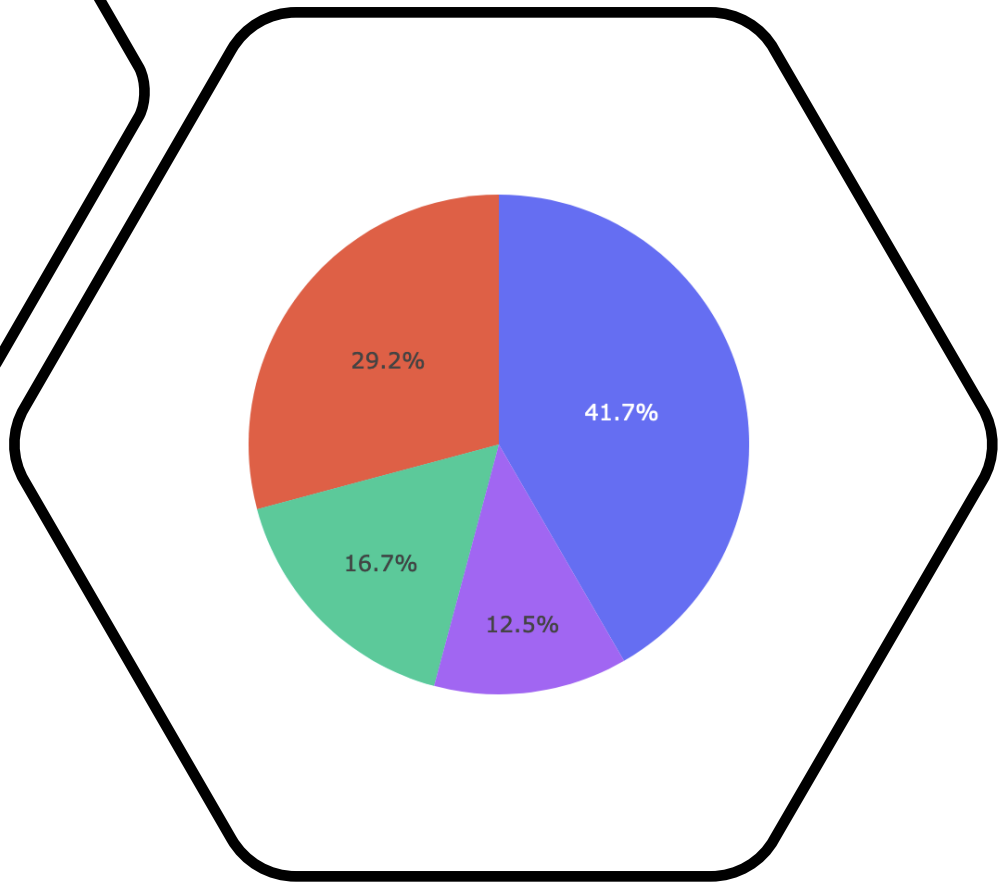




Section 4

Build a Dashboard with Plotly Dash

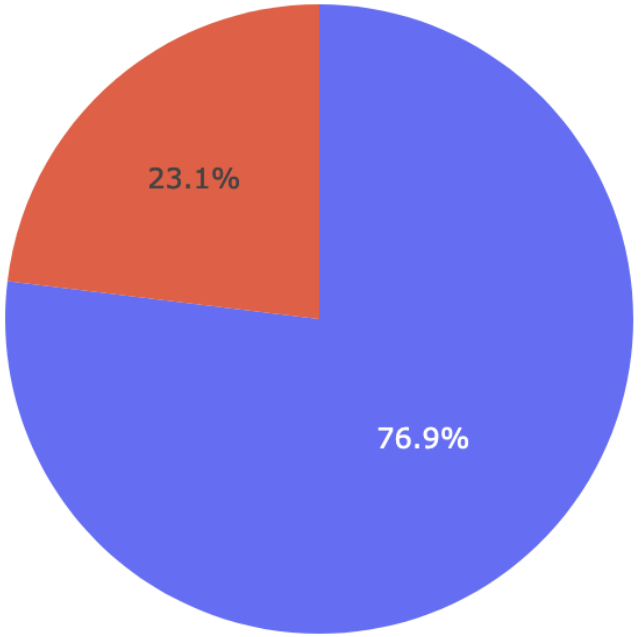
Successful Launches across Launch Sites



- This is the distribution showing the successful landings across all launch sites.
- CCAFS LC-40 is the old name of CCAFS SLC-40 so CCAFS and KSC have the same amount of successful landings, but a majority of the successful landings where performed before the name change.
- VAFB has the smallest share of successful landings which could be due to the difficult in launching the rocket from the US West Coast.

Pie chart for the launch site with the highest launch success ratio

- The launch site KSC LC-39 A has the highest rate of successful launches, with a 76.9% success rate.
- There are 10 successful landings and 3 failures.



Payload vs. Launch Outcome Scatter Plot

- After plotting all the payloads, it can be seen that around 4000 kg there is a gap and as such, I have made the split around 4000 kg.

Less than 4000 kg are low payloads while 4000 – 10000 kg are massive payloads.

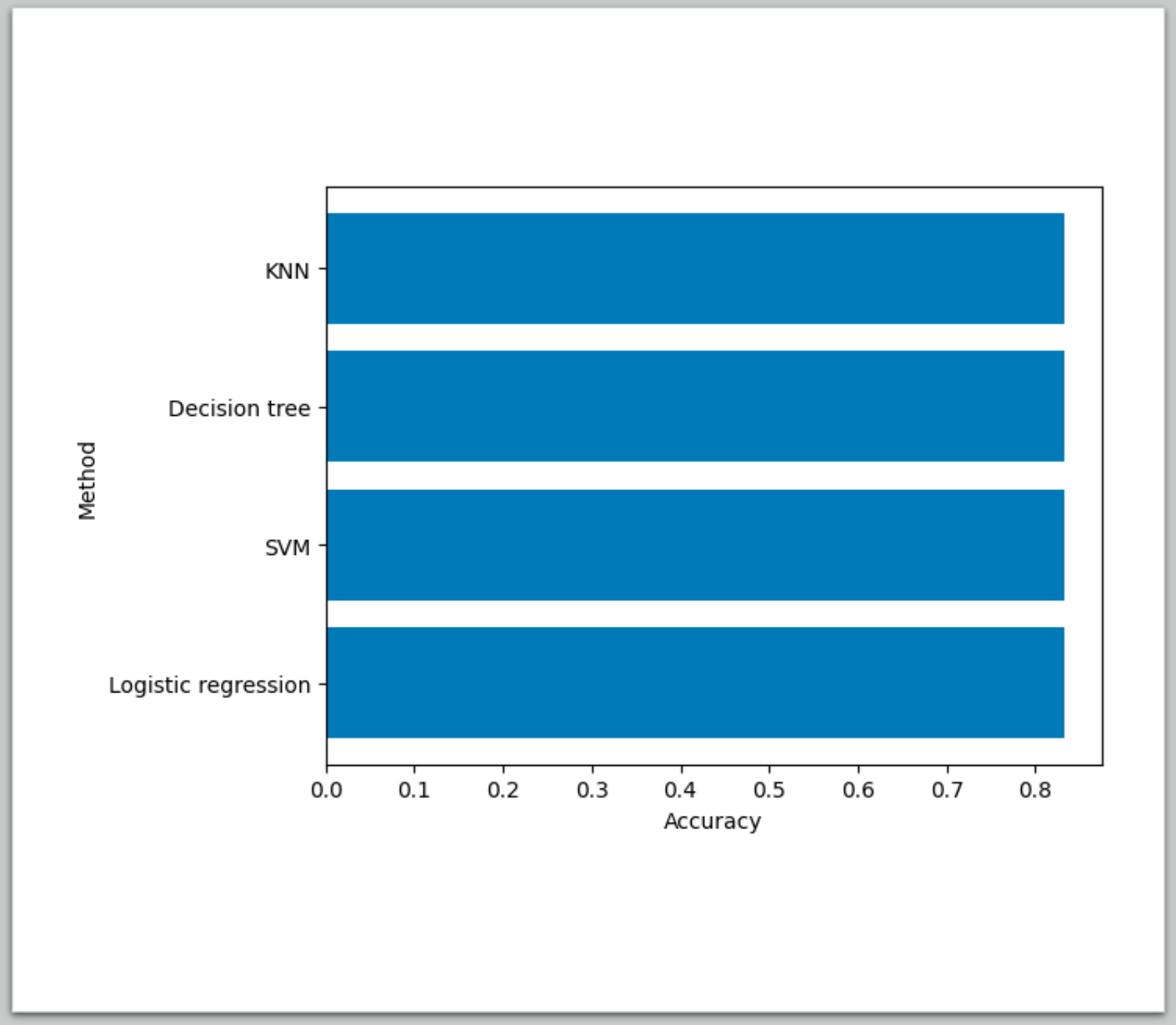
- From the 2 plots to the right, it can be observed that success for massive payloads is lower than for low payloads.
- Further, 2 booster types v1.0 and B5 have not been launched with massive payloads.

Section 5

Predictive Analysis (Classification)

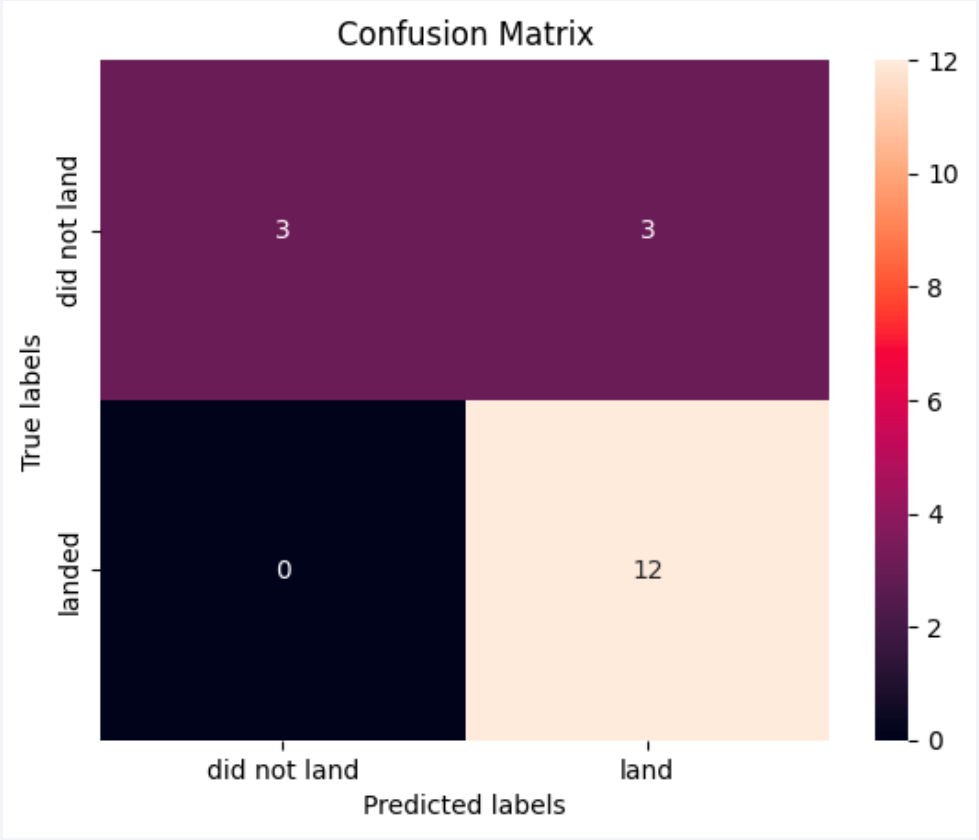
Classification Accuracy

- All the models have the same accuracy score of 0.8333.
- It is to be noted that we are a smaller sample size for this problem and which may cause a variance in the classification accuracy.



Confusion Matrix

- Since all models have similar accuracy across the same test set, the confusion matrix is the same across all models.
- The models predicted 12 successful landings when the True label was a successful landing.
- The models predicted 3 unsuccessful landings when the True label was a unsuccessful landing.
- The models predicted 3 successful landings when the true label was unsuccessful landings thereby creating false positives.



Conclusions

The project was undertaken to build a machine learning model for an alternative company to SpaceX.

We have observed the following:

- As the number of flights increases, the rate of success at a launch site increases, with most early flights being unsuccessful. i.e. with more experience, the success rate increases.
- Between 2010 and 2013, all landings were unsuccessful (as the success rate is 0).
- After 2013, the success rate generally increased, despite small dips in 2018 and 2020.
- After 2016, there was always a greater than 50% chance of success.
- Orbit types ES-L1, GEO, HEO have the highest (100%) success rate as they only have 1 flight in the orbit.
- The 100% success rate in SSO orbit type even better as it is with 5 successful flights from the sample observed.
- The orbit types PO, ISS, and LEO, have more success with heavy payloads.
- VLEO launches are associated with heavier payloads.
- The success for massive payloads (over 4000kg) is lower than that for low payloads.
- The model has an accuracy of 83.33%.
- The alternative company can now use this model to predict with a relatively high rate of accuracy whether the launch and landing of the Falcon 9 rocket would be successful to determine whether or not the launch should be made.
- The machine learning model may be further improved by increasing the sample data size to determine the better machine learning model and improving the accuracy.

Appendix

Github Repository of My Project:

https://github.com/ramsrii/DataScience_Practice.git

A special thanks to all the instructors across the Professional Certification as well as all my peers who reviewed my code.

Thank you!