IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>
<Date>

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

- Project background and context
- Problems you want to find answers

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Describe how data sets were collected.
- Request and parse the SpaceX launch data using the GET request
- Decode the response content as a Json using .json() and turn it into a Pandas data frame using .json_normalize()
- Filter the data frame only to include Falcon 9 launches
- Dealing with Missing Values
- Export it to a CSV for the next section, but to make the answers consistent, we will provide data in a pre-selected date range in the next lab.
- data_falcon9.to_csv('dataset_part_1.csv', index=False)

Data Collection – SpaceX API

- During this process, we use an HTTP GET request to retrieve SpaceX launch data from its API. For easier analysis, we then parse this JSON data into a structured format, such as a Python dictionary or a pandas data frame. A static response object is used for consistent results.

- https://github.com/kingfroglao/Coursera_IBM-Data-Science/blob/main/Applied%20Data%20Science%20Capstone/Data%20Collection%20API%20Lab.ipynb

To make the requested JSON results more consistent, we will use the following static response object for this project:

In [25]:
```
static_json_url='https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/I
```

We should see that the request was successfull with the 200 status response code

In [10]:
```
response.status_code
```

Out[10]:
```
200
```

Now we decode the response content as a Json using `.json()` and turn it into a Pandas dataframe using `.json_normalize()`

In [28]:
```
# Use json_normalize meethod to convert the json result into a dataframe
static_json_df = response.json()
```

Using the dataframe `data` print the first 5 rows

In [30]:
```
# Get the head of the dataframe
data = pd.json_normalize(static_json_df)
data.head()
```

Data Collection - Scraping

- In this step, we send an HTTP GET request to the URL of the Falcon9 Launch Wiki page, which results in an HTTP response containing the HTML content of the requested page.
- https://github.com/kingfroglao/Coursera_IBM-Data-Science/blob/main/Applied%20Data%20Science%20Capstone/Data%20Collection%20with%20Web%20Scraping%20lab.ipynb

```
5]: # use requests.get() method with the provided static_url
    # assign the response to a object
    html_data = requests.get(static_url)
    html_data.status_code

5]: 200
```

Create a `BeautifulSoup` object from the HTML `response`

```
6]: # Use BeautifulSoup() to create a BeautifulSoup object from a response text
    soup = BeautifulSoup(html_data.text, 'html.parser')
```

Print the page title to verify if the `BeautifulSoup` object was created properly

```
7]: # Use soup.title attribute
    soup.title

7]: <title>List of Falcon 9 and Falcon Heavy launches - Wikipedia</title>
```

Data Wrangling

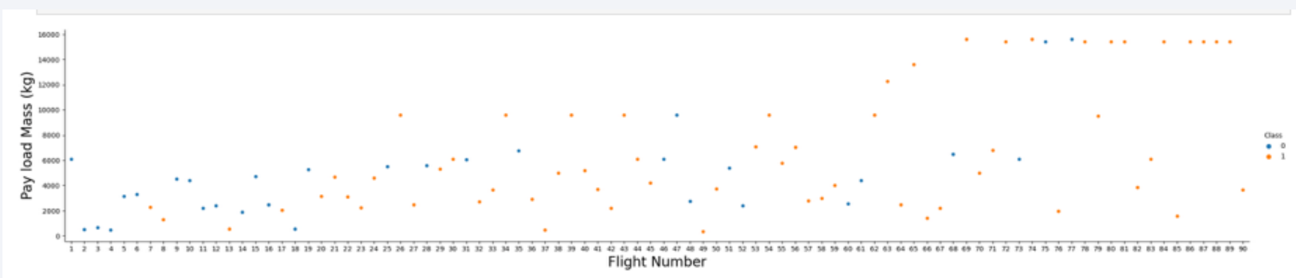
- 1: Calculate the number of launches on each site
- 2: Calculate the number and occurrence of each orbit
- 3: Calculate the number and occurence
mission outcome per orbit type
- 4: Create a landing outcome label from Outcome column

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1

- https://github.com/kingfroglao/Coursera_IBM-Data-Science/blob/main/Applied%20Data%20Science%20Capstone/Data%20wrangling.ipynb

EDA with Data Visualization

- The scatter plot was used to visualize the relationship between the 'FlightNumber' (representing sequential launch attempts) and 'PayloadMass', overlaid with the outcome of each launch. This plot allowed us to infer that with increasing flight numbers and decreasing payload mass, the likelihood of successful first stage landing seems to increase.
- The bar chart was then utilized to show the success rate of each orbit type. This helped us understand the variance in success rate among different orbits, offering insights into which orbits have higher probabilities of successful launches. Both of these charts were instrumental in visualizing key aspects of the launch data and interpreting the factors contributing to launch success.
- https://github.com/kingfroglao/Coursera_IBM-Data-Science/blob/main/Applied%20Data%20Science%20Capstone/EDA%20with%20Data%20Visualization.ipynb



EDA with SQL

- Retrieved unique launch site names.
- Displayed 5 records of launch sites starting with 'CCA'.
- Calculated total payload mass for NASA's CRS launches.
- Found average payload mass for booster version F9 v1.1.
- Determined date of first successful ground pad landing.
- Listed boosters are successful in drone ship landings with payload mass between 4000-6000.
- Calculated total successful and failed mission outcomes.Identified booster versions carrying the maximum payload mass.
- Displayed records with month, failure landing outcomes in drone ships, booster versions, and launch sites for 2015.
- Ranked landing outcomes between 2010-06-04 and 2017-03-20 in descending order.
- https://github.com/kingfroglao/Coursera_IBM-Data-Science/blob/main/Applied%20Data%20Science%20Capstone/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- We added annotations to all launch sites on a folium map, integrating various map elements to represent the success or failure of launches at each location. The 'launch outcomes' were assigned to two classes, 0 for failure and 1 for success, which helped us to distinguish launch sites with relatively high success rates using color-coded marker clusters. Furthermore, we calculated the proximity of each launch site to its surrounding features, investigating their proximity to railways, highways, or coastlines, and their distance from cities.

- https://github.com/kingfroglao/Coursera_IBM-Data-Science/blob/main/Applied%20Data%20Science%20Capstone/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

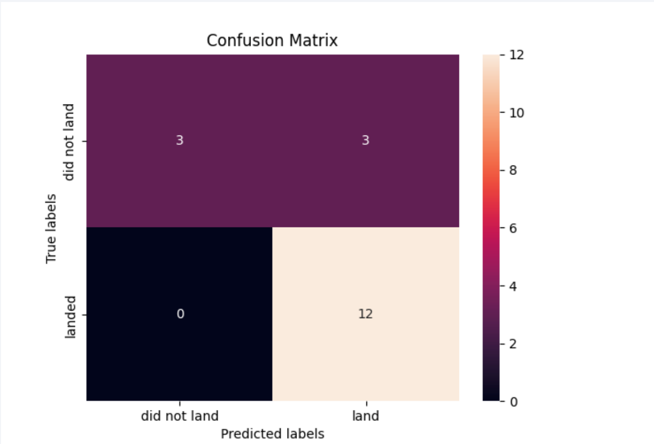
- We developed an interactive dashboard using Plotly Dash, providing a comprehensive view of SpaceX launch data.
- Pie charts were constructed to depict the total number of launches from specific sites, providing a visual breakdown of launch distribution.
- We created scatter plots to illustrate the relationship between the launch outcomes and payload mass (Kg) for various booster versions. These plots offer valuable insights into how these two parameters interact.

Predictive Analysis (Classification)

- We initiated the process by loading the data through Numpy and Pandas, then proceeded with data transformation, followed by a split into training and testing sets. With this prepared data, we constructed various machine learning models and fine-tuned their hyperparameters utilizing GridSearchCV.
- Accuracy served as our key evaluation metric throughout the modeling process. To enhance the performance of our models, we employed feature engineering techniques and refined our algorithm parameters.
- Upon completion of these iterative processes, we successfully identified the best-performing classification model for our dataset.
- https://github.com/kingfroglao/Coursera_IBM-Data-Science/blob/main/Applied%20Data%20Science%20Capstone/%20Machine%20Learning%20Prediction%20lab.ipynb

Results

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

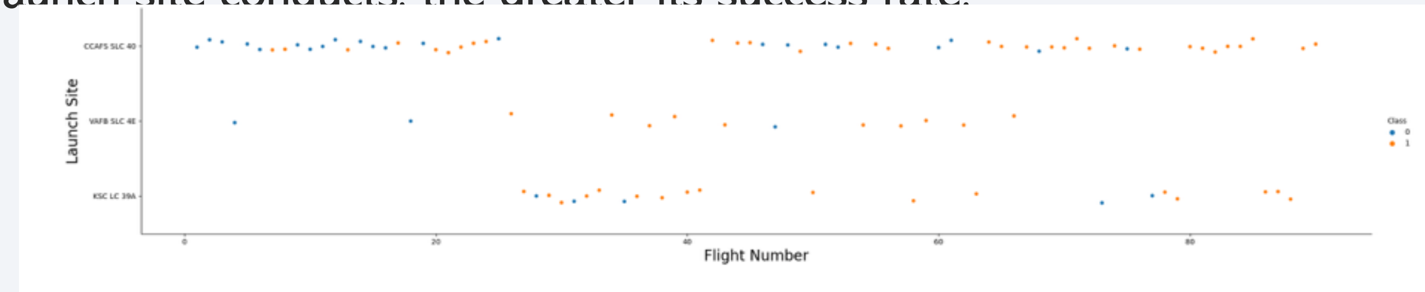


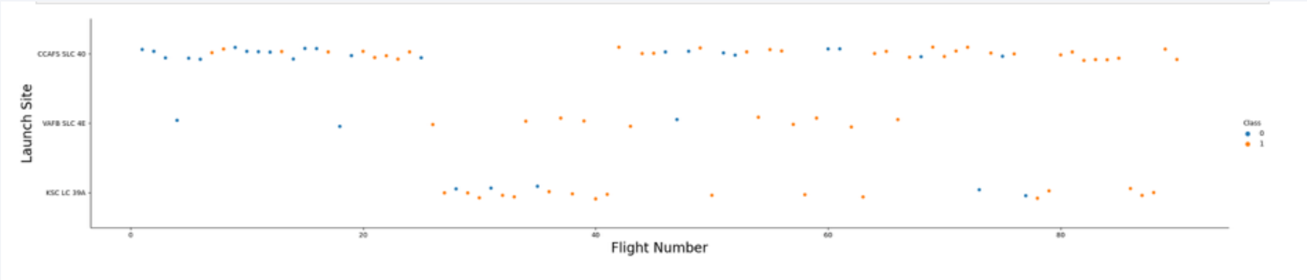
Section 2

Insights drawn from EDA

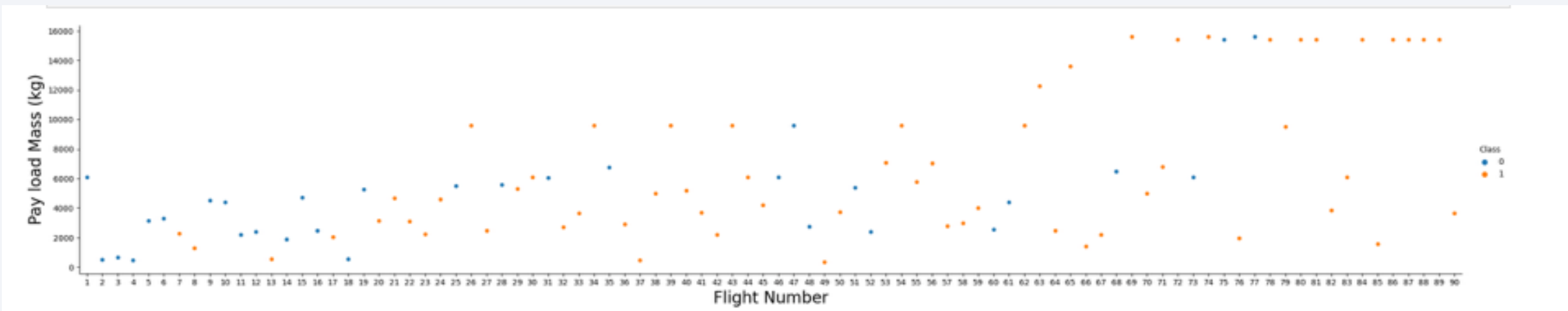
Flight Number vs. Launch Site

- From the visual representation, it is evident that a launch site's success rate is positively correlated with its flight volume - the higher the number of flights a launch site conducts, the greater its success rate.





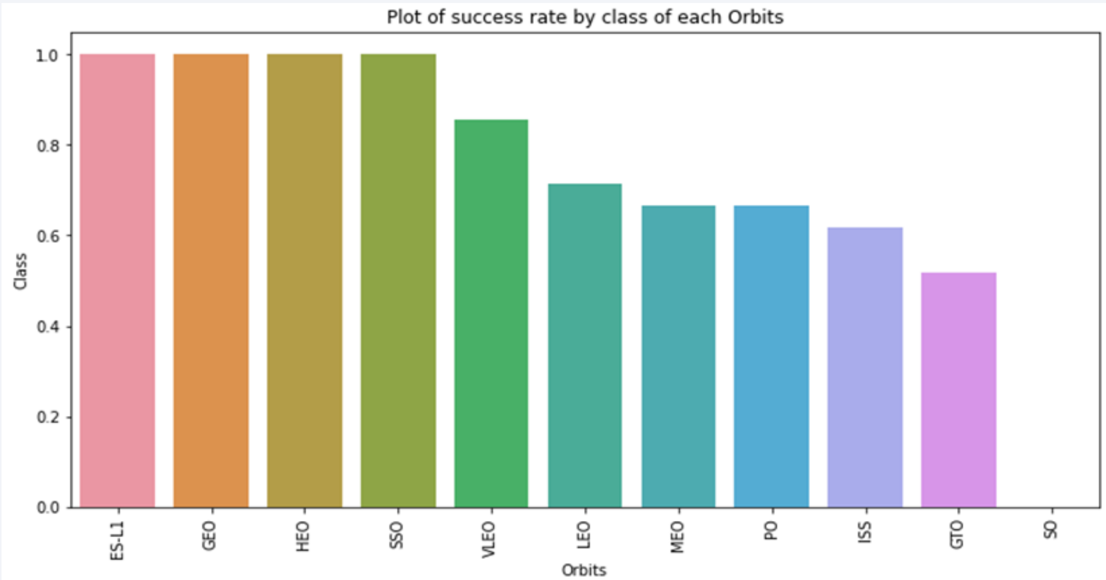
Payload vs. Launch Site



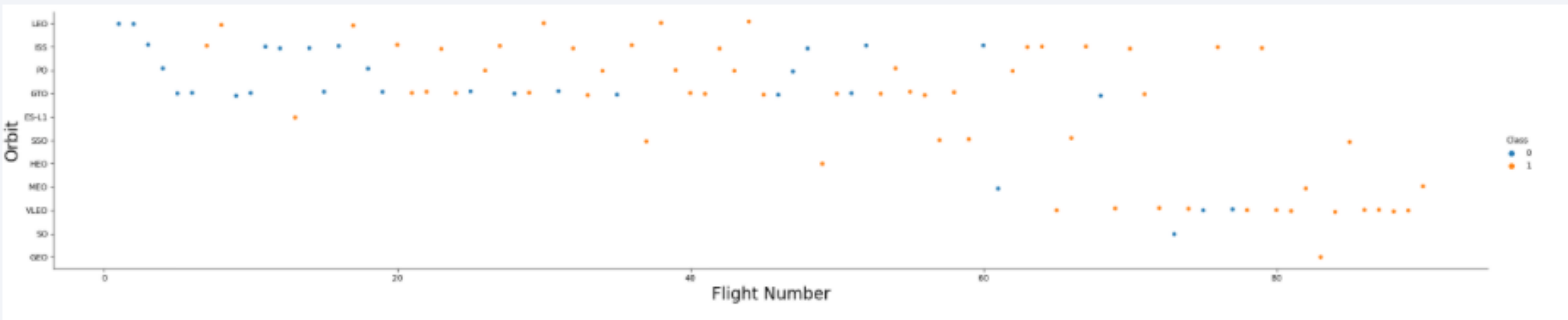
- We see that different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.

Success Rate vs. Orbit Type

- From the visualization, it is evident that the orbits ES-L1, GEO, HEO, SSO, and VLEO have achieved the highest success rates.

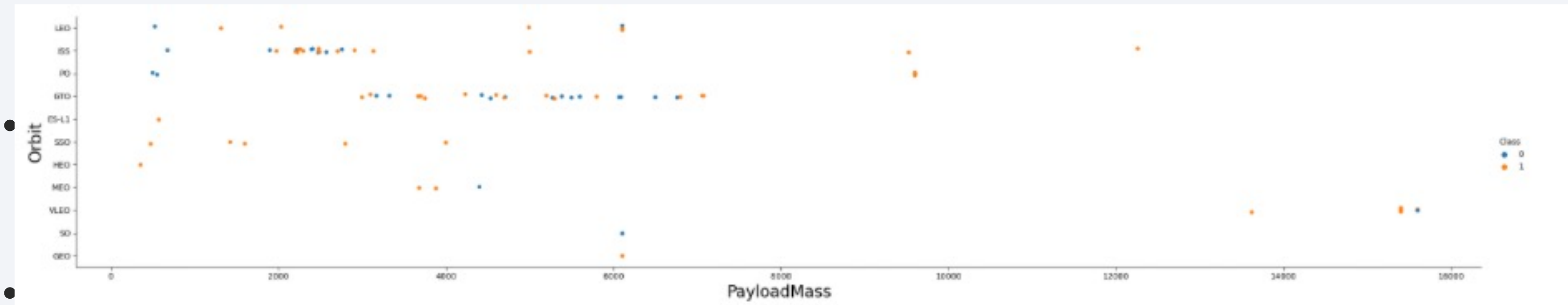


Flight Number vs. Orbit Type



- The following graph illustrates the relationship between Flight Number and Orbit type. It's noticeable that for flights within the Low Earth Orbit (LEO), the success rate seems to correlate with the number of flights. However, for those in the Geosynchronous Transfer Orbit (GTO), no such correlation between the flight number and success rate is evident.

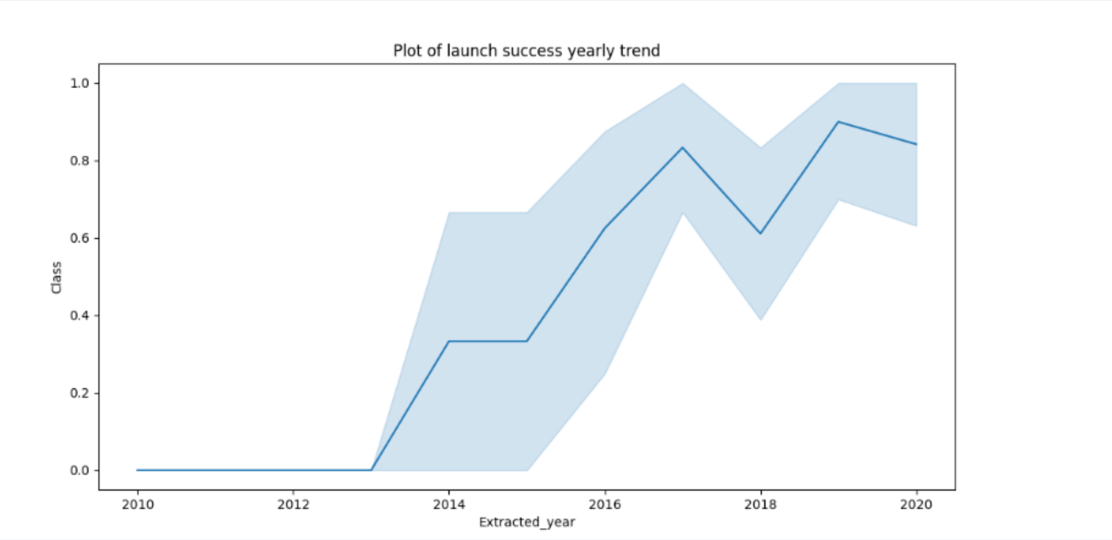
Payload vs. Orbit Type



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

Launch Success Yearly Trend

- Based on the plot, it is evident that the success rate has consistently risen since 2013, up until 2020.



All Launch Site Names

Display the names of the unique launch sites in the space mission

```
%sql select Unique(LAUNCH_SITE) from SPACEX;
```

```
 * ibm_db_sa://crm77791:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnr
75/bludb
   sqlite:///my_data1.db
Done.
```

launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E
None

Launch Site Names Begin with 'CCA'

- We used the query above to display 5 records where launch sites begin with \`CCA\`

Display 5 records where launch sites begin with the string 'CCA'

```
%sql SELECT LAUNCH_SITE from SPACEX where (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5;
```

```
 * ibm_db_sa://crm77791:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u9
75/bludb
   sqlite:///my_data1.db
Done.
```

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

Total Payload Mass

- Display the total payload mass carried by boosters launched by NASA (CRS)

```
In [35]:  %sql select sum(PAYLOAD_MASS__KG_) as payloadmass from SPACEX;

 * ibm_db_sa://crm77791:***@98538591-7217-4024-b027-8baa776ffad1.c3
75/bludb
   sqlite:///my_data1.db
Done.
```

Out[35]:

payloadmass
619967

Task 4

Average Payload Mass by F9 v1.1

- Display average payload mass carried by booster version F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_) as payloadmass from SPACEX;
```

```
 * ibm_db_sa://crm77791:***@98538591-7217-4024-b027-8baa776ffad1.c3
75/bludb
   sqlite:///my_data1.db
Done.
```

payloadmass
6138

First Successful Ground Landing Date

- Use min function to list the date when the first succesful landing outcome in ground pad was acheived.

```
[37]: %sql select min(DATE) from SPACEX;

 * ibm_db_sa://crm77791:***@98538591-7217-4024-
75/bludb
   sqlite:///my_data1.db
Done.
```

[37]:

1
2010-04-06

Successful Drone Ship Landing with Payload between 4000 and 6000

The modified SQL query retrieves the names of boosters that have successfully landed on a drone ship and had a payload mass greater than 4000kg but less than 6000kg.

```
[39]: %sql select BOOSTER_VERSION from SPACEX where LANDING_OUTCOM

 * ibm_db_sa://crm77791:***@98538591-7217-4024-b027-8baa776ffa
75/bludb
   sqlite:///my_data1.db
Done.
```

t[39]:

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

Total Number of Successful and Failure Mission Outcomes

- The SQL query lists the records showing the month names, failure landing outcomes on a drone ship, booster versions, and launch sites for the months in the year 2015, using the substr function to extract the month and year from the date column.

```
The total number of successful mission outcome is:
```

	successoutcome
0	100

```
The total number of failed mission outcome is:
```

	failureoutcome
0	1

Boosters Carried Maximum Payload

- We determined the booster that have carried the maximum payload using a subquery in the WHERE clause and the MAX() function.

```
[42]: %sql select BOOSTER_VERSION as boosterve

 * ibm_db_sa://crm77791:***@98538591-7217-4
75/bludb
   sqlite:///my_data1.db
Done.
```

[42]:

boosterversion
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

2015 Launch Records

- We applied a combination of the WHERE clause, LIKE, AND, and BETWEEN conditions to filter the records for failed landing outcomes on a drone ship, corresponding booster versions, and launch site names specifically for the year 2015.

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

Out[43]:

1	mission_outcome	booster_version	launch_site
10	Success	F9 v1.1 B1012	CCAFS LC-40
11	Success	F9 v1.1 B1013	CCAFS LC-40
2	Success	F9 v1.1 B1014	CCAFS LC-40
4	Success	F9 v1.1 B1015	CCAFS LC-40
4	Success	F9 v1.1 B1016	CCAFS LC-40
6	Failure (in flight)	F9 v1.1 B1018	CCAFS LC-40
12	Success	F9 FT B1019	CCAFS LC-40

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

- Select landing_outcome from spacex where date between '2010-06-04' and '2017-03-20' order by date desc;

```
In [44]: %sql SELECT LANDING_OUTCOME FROM SPACEX WHERE DATE BETWEEN '2010-06-
 * ibm_db_sa://crm77791:***@98538591-7217-4024-b027-8baa776ffad1.c3n41
75/bludb
   sqlite:///my_data1.db
Done.
```

Out[44]:

landing_outcome
No attempt
Success (ground pad)
Success (ground pad)
Success (drone ship)
Success (ground pad)
Success (drone ship)
Success (drone ship)
Success (ground pad)
Failure (drone ship)
Success (drone ship)
Success (drone ship)
Failure (drone ship)
Failure (drone ship)
Success (ground pad)
Controlled (ocean)
Failure (drone ship)
Precluded (drone ship)
No attempt
Failure (drone ship)
No attempt
Uncontrolled (ocean)
Controlled (ocean)
No attempt
No attempt
No attempt
Controlled (ocean)
Uncontrolled (ocean)
No attempt
No attempt
No attempt
No attempt
Failure (parachute)

Section 3

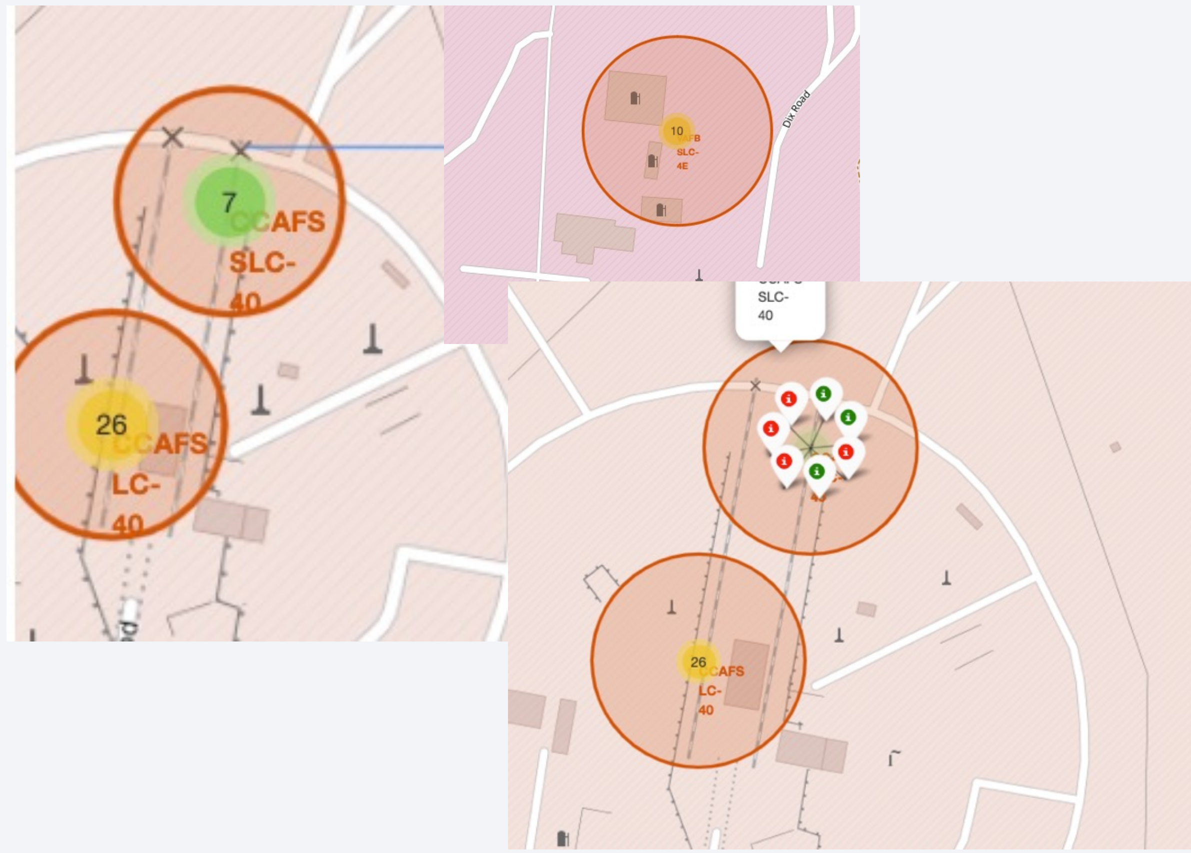
Launch Sites Proximities Analysis

All launch sites global map markers

- All launch sites are off the coast of the southern United States

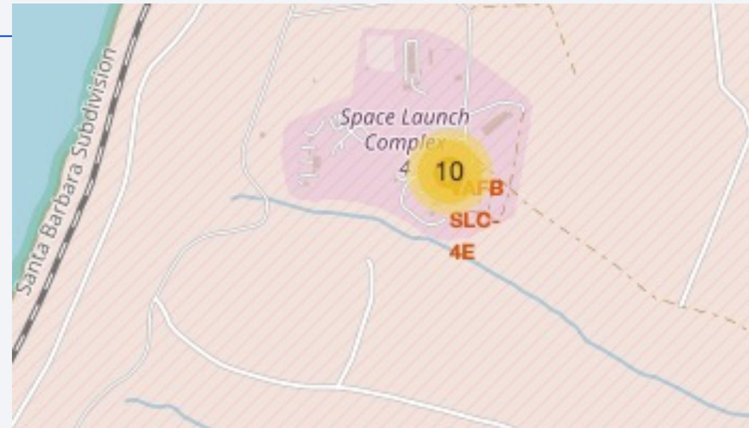
Color-labeled launch outcomes on the map

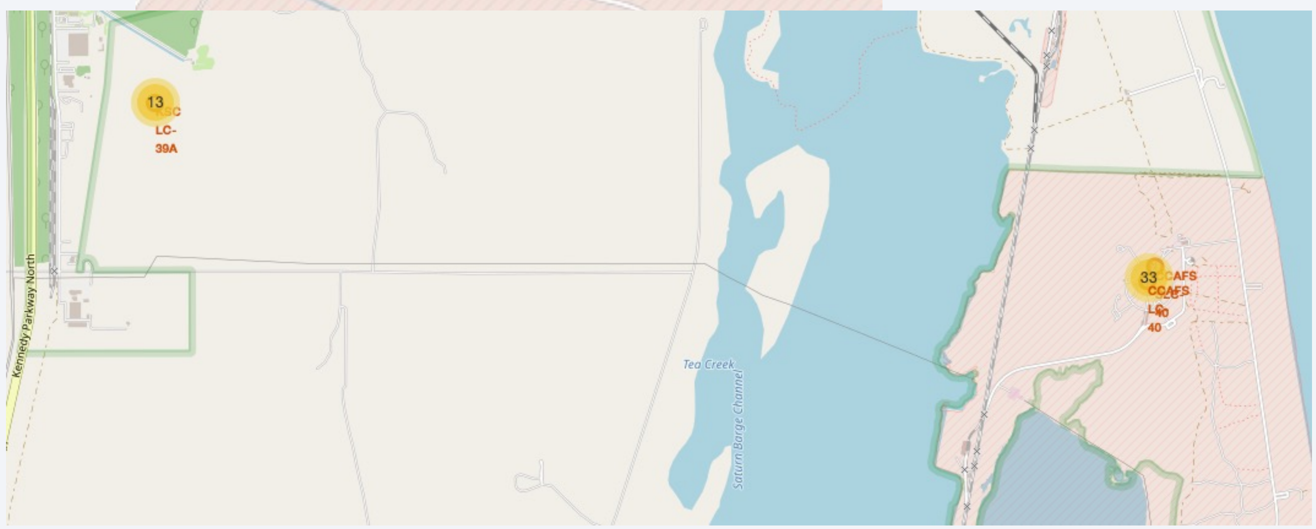
- Green is success, red is failure



Launch Site distance to landmarks

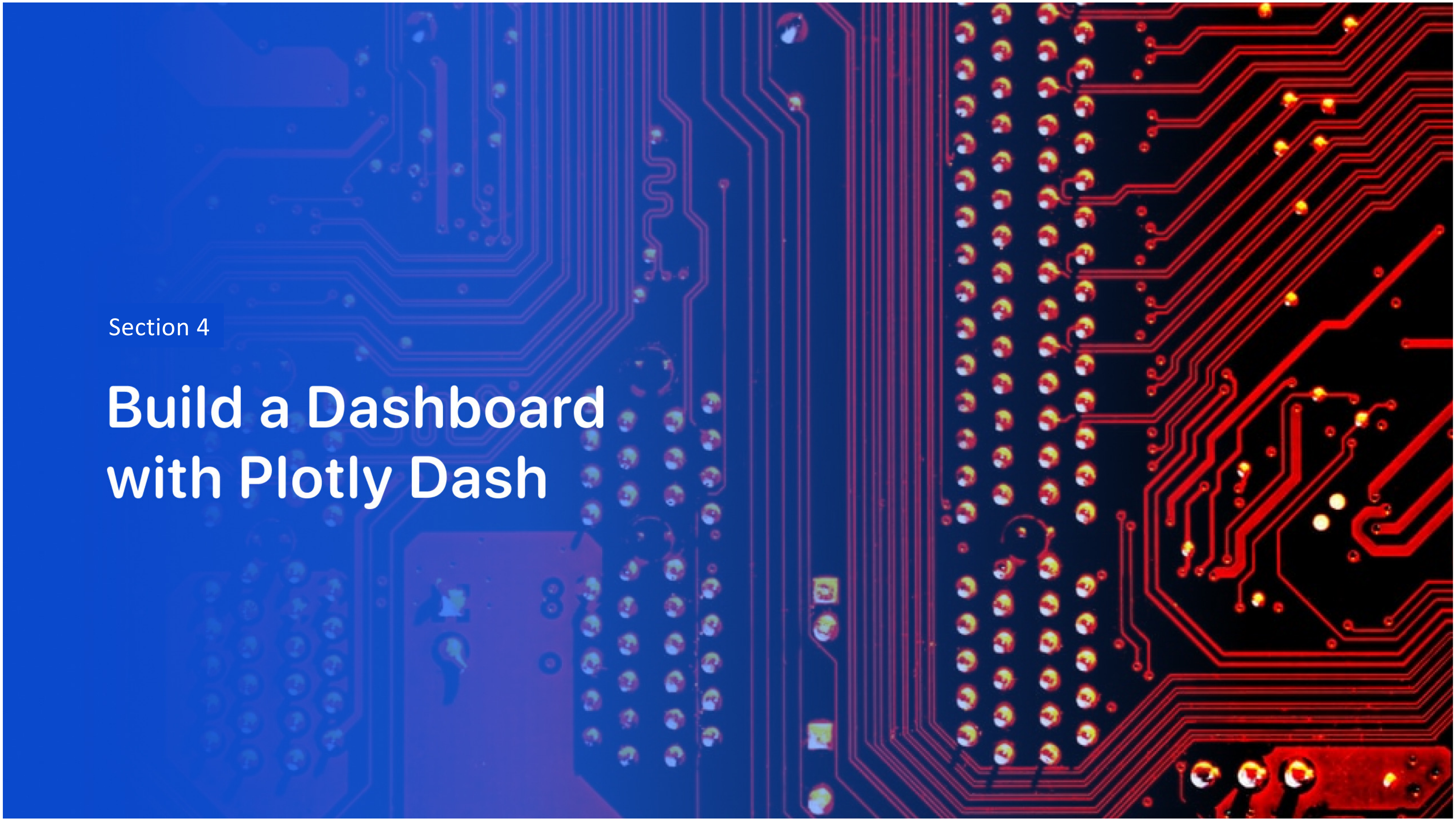
- Launch sites are not in close proximity to railways, they are not in close proximity to highways, they are in close proximity to the coastline, and they maintain a certain distance away from cities.





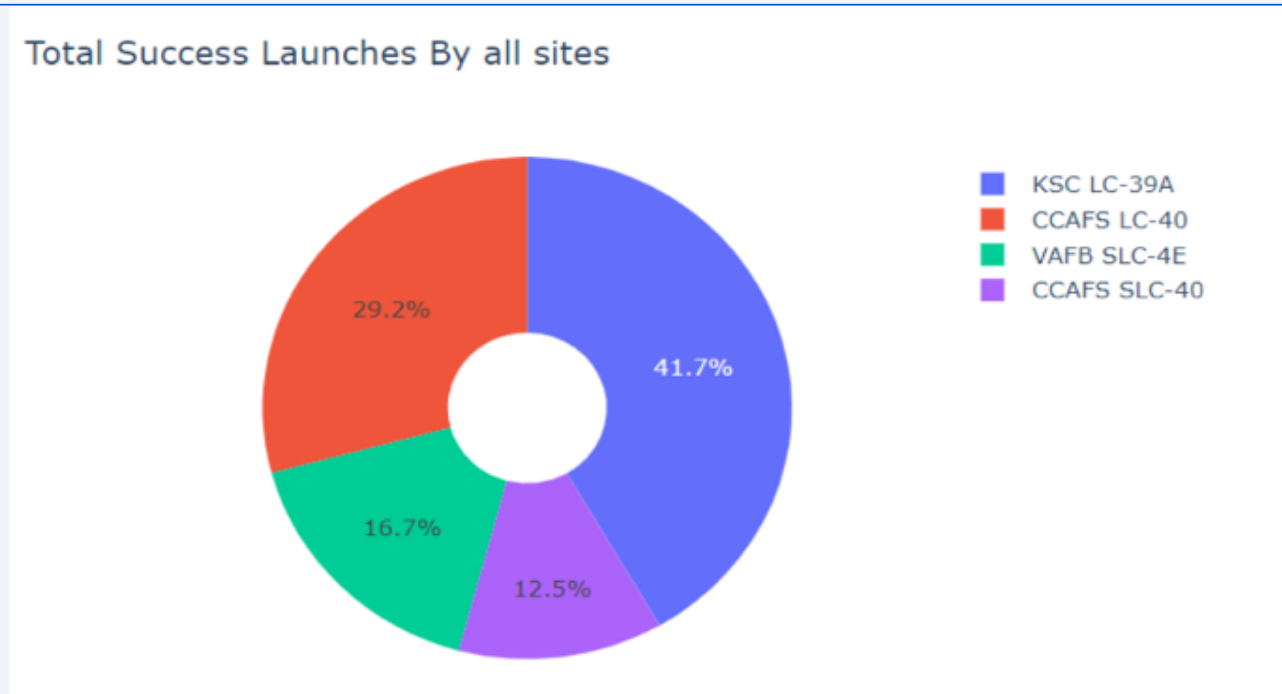
Section 4

Build a Dashboard with Plotly Dash



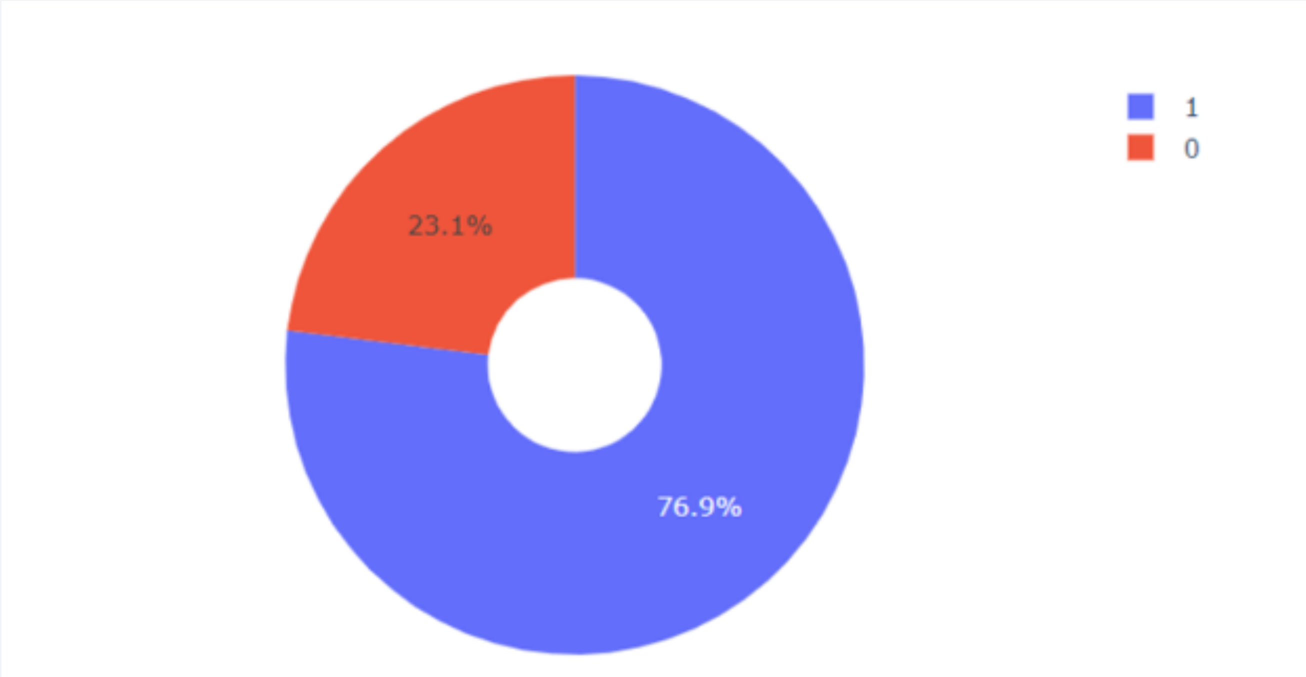
Piechart of launch success count for all sites

KSC LC-39A was the most successful launch.



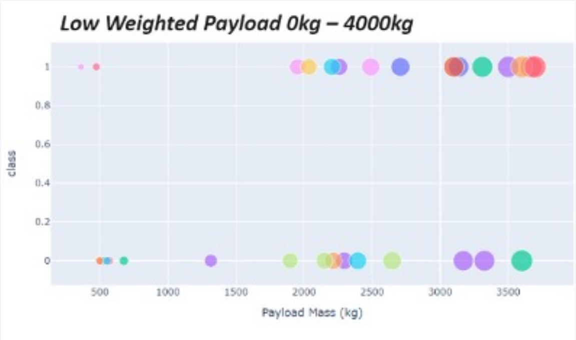
Piechart for the highest launch success ratio

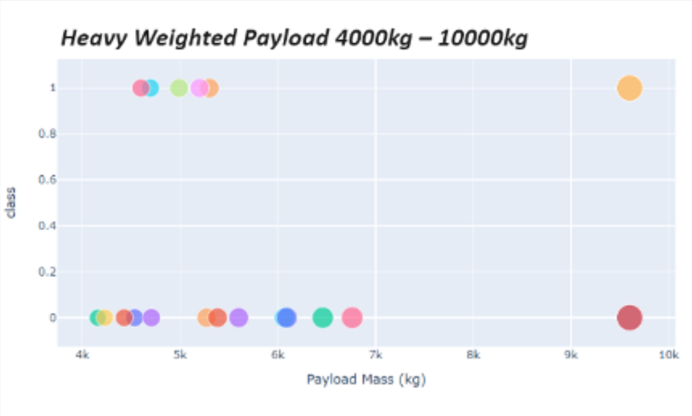
- KSC LC-39A had 76.9% success rate and 23.1% failure rate



Payload vs. Launch

- Low weigted payloads are more successful than high weighted payloads



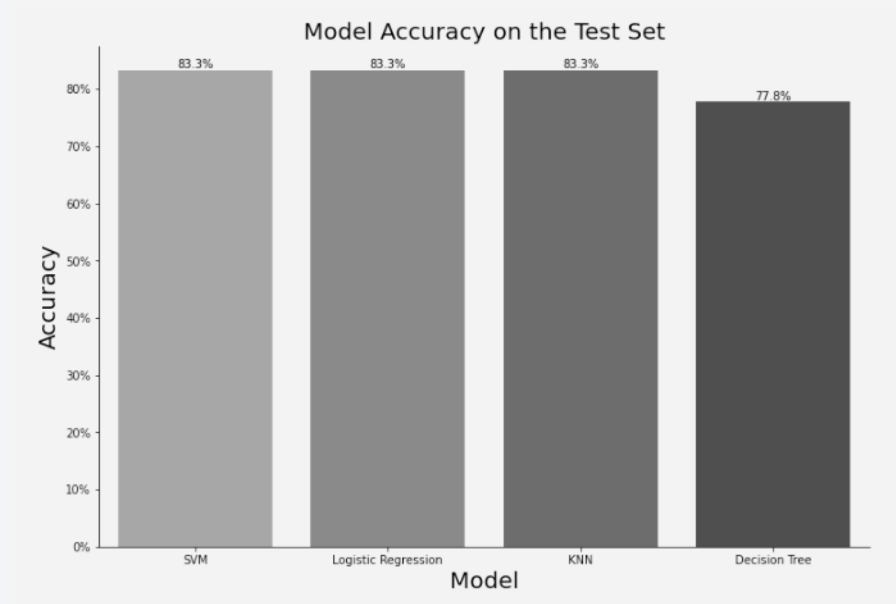


Section 5

Predictive Analysis (Classification)

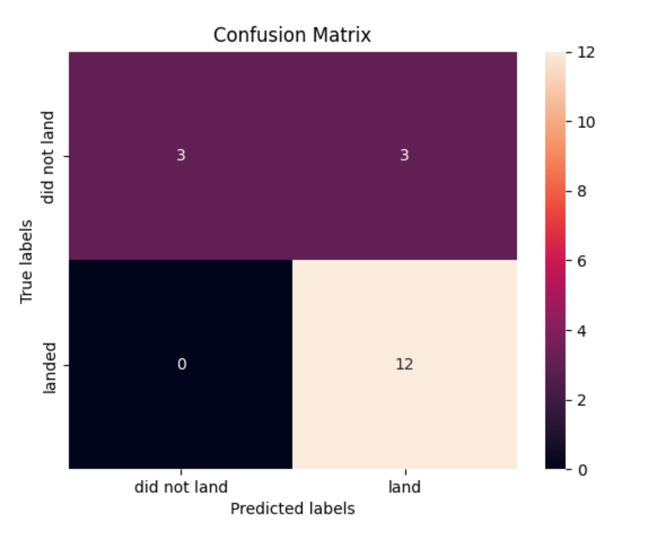
Classification Accuracy

- The decision SVM, Logistic Regreesion and KNN are the model with the highest classification accuracy



Confusion Matrix

- The confusion matrix for a decision tree classifier reveals its ability to differentiate and classify different classes accurately.



Conclusions

- Through the analysis, we obtained the following conclusions:
- positive correlation between flight volume and success rate at the launch site, immediate flight volume exceeded, success rate exceeded.
- During the period from 2013 to 2020, the rate of success has been steadily increasing.
- Highest success rate for ES-L1, GEO, HEO, SSO and VLEO.
- Among the existing launch sites, the KSC LC-39A has the highest launch success rate.
- Finally, after comparing, we decided to present the best machine learning method for this task.


Thank you!