IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Nicholas J. Monroe
December 27, 2024

Outline

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

Executive Summary

Summary of methods and methodologies

A variety of methods and methodologies were used in the development of this report. Those methods and methodologies include:

- Data Collection through API
- Data Collection with Web Scraping
- Data Wrangling
- Exploratory Data Analysis (EDA) with SQL
- EDA with Data Visualization
- Interactive Visual Analytics with Folium and Plotly Dash
- Predictive Analysis through Grid Search, Logistic Regression, SVM, Decision Tree Classification, and K-nearest neighbor

Summary of all results

The result produce through the methods and methodologies are as follows:

- EDA Results
- Predictive Analysis Results

Introduction

- Project background and context
  - SpaceX is a space technology company that provides transport to the International Space Station, access to a satellite internet constellation, and private spaceflight. The Falcon 9 is a partially reusable, two-stage-to-orbit launch vehicle that costs $62 million per launch. This is relatively cheap when compared to SpaceX's competitors whose launch cost can reach upward of $165 million.
- Problems you want to find answers
  - Much of SpaceX's launch cost savings is due to the potential reuse of the first stage components. If the success rate first stage landings can be determined, the total cost of a launch can be estimated. The variables of potential first stage landings include parameters such as the payload, orbit, and customer preferences. Thus, data science is needed to assess these parameters and determine the potential cost.

Section 1

Methodology

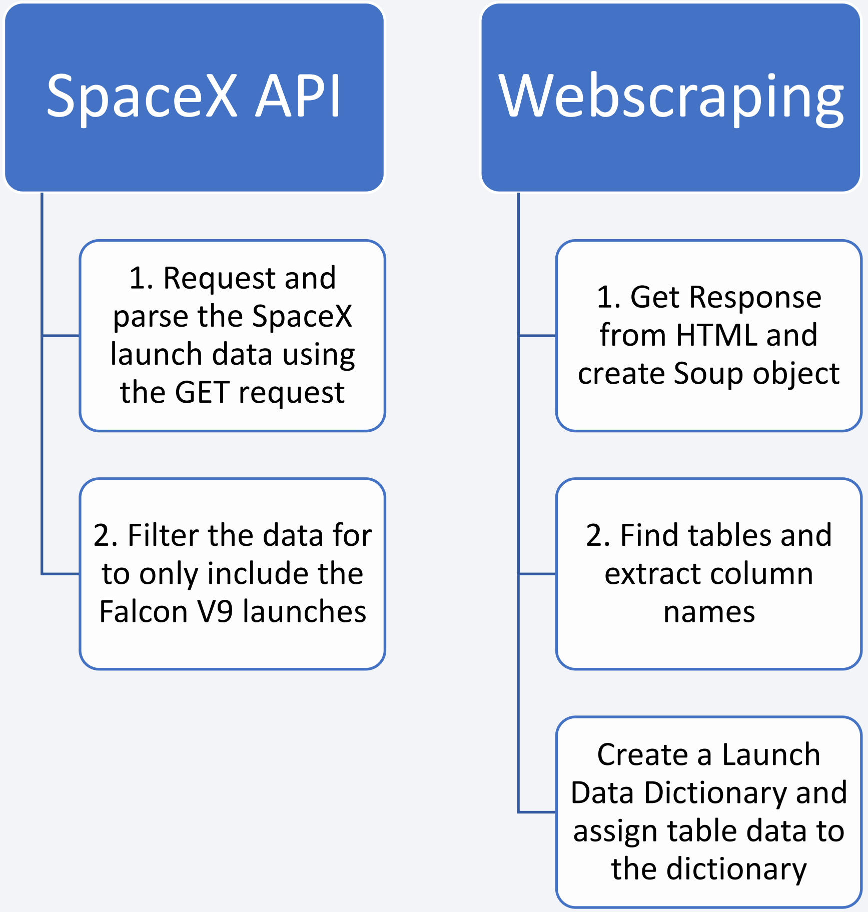
Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX API and web scraping from Wikipedia.
- Perform data wrangling
  - One Hot encoding was used to transform categorical variables into factors
- 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

- The data sets were collected using SpaceX REST API (https://api.spacexdata.com/v4) and additional data on Falcon 9 rocket launches was collected through webscraping the Wikipedia page, List of Falcon 9 and Falcon Heavy launches (https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922) with BeautifulSoup
- The data collection steps for the ScapeX API and Wikipedia webscraping included:



Data Collection – SpaceX API

GitHub URL of the completed SpaceX API calls notebook: https://github.com/nchlsmnro/IBM_Data_Sci_Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Step 1.

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

```
response = requests.get(spacex_url)
```

Step 2.

```
# Call getBoosterVersion
getBoosterVersion(data)
```

```
# Call getLaunchSite
getLaunchSite(data)
```

```
# Call getPayloadData
getPayloadData(data)
```

```
# Call getCoreData
getCoreData(data)
```

Step 3.

```
# Hint data['BoosterVersion']!='Falcon 1'
data_falcon9 = df[df['BoosterVersion'] != 'Falcon 1']
```

```
data_falcon9.to_csv('dataset_part_1.csv', index=False)
```

Data Collection - Scraping

GitHub URL of the completed web scraping notebook: https://github.com/nchlsmnro/IBM_Data_Sci_Capstone/blob/main/jupyter-labs-webscraping.ipynb

Step 1.

```python
# use requests.get() method with the provided static_url\
# assign the response to a object
response = requests.get(static_url)
```

```python
# Use BeautifulSoup() to create a BeautifulSoup object from a response text content
soup = BeautifulSoup (response.content)
```

Step 2.

```python
html_tables = soup.find_all('table')
len(html_tables)
```

```python
first_launch_table = html_tables[2]
print(first_launch_table)
```

```python
column_names = []
for row in first_launch_table.find_all('th'):
    name = extract_column_from_header(row)
    if (name != None and len(name) > 0):
        column_names.append(name)
```

Step 3*.

```python
extracted_row = 0
for table_number,table in enumerate(soup.find_all('table',"wikitable plainrowheaders collapsi
    for rows in table.find_all("tr"):
        if rows.th:
            if rows.th.string:
                flight_number=rows.th.string.strip()
                flag=flight_number.isdigit()
        else:
            flag=False
        row=rows.find_all('td')
        if flag:
            extracted_row += 1
            launch_dict['Flight No.'].append(flight_number)
            datatimelist=date_time(row[0])
            date = datatimelist[0].strip(',')
            launch_dict['Date'].append(date)
            time = datatimelist[1]
            launch_dict['Time'].append(time)
            bv=booster_version(row[1])
            if not(bv):
                bv=row[1].a.string
                launch_dict['Version Booster'].append(bv)
```

* The full code for Step 3 did not fit on the slide

Data Wrangling

- Exploratory Data Analysis (EDA) was performed to find some patterns in the data and determine what would be the label for training supervised models. The performed tasks includes:
  - Calculating the number of launches on each site
  - Calculating the number and occurrence of each orbit
  - Calculating the number and occurrence of mission outcome of the orbits
  - Creating a landing outcome label from Outcome column
- Other wrangling tasks includes:
  - Screening for missing values
  - Replacing "NA" inputs for Payload mass with the sample mean

- GitHub URL of your completed data wrangling related notebooks: https://github.com/nchlsmnro/IBM_Data_Sci_Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Several visualizations were created for the purposes of EDA including:
  - Scatterplots comparing Flight Number and Payload Mass, Flight Number and Launch Site, Payload Mass and Launch Site, Flight Number and Orbit Type, and Payload Mass and Orbit
  - A bar chart visualizing the success rate (quantity of successful landings) of different orbit types
  - A line chart visualizing the yearly trend of launch success

- GitHub URL of your completed EDA with data visualization notebook: https://github.com/nchlsmnro/IBM_Data_Sci_Capstone/blob/main/edadataviz.ipynb

EDA with SQL

- Various SQL queries were executed on the Db2 database for EDA including:
    1. Display the names of the unique launch sites in the space mission
    2. Display 5 records where launch sites begin with the string 'CCA'
    3. Display the total payload mass carried by boosters launched by NASA (CRS)
    4. Display average payload mass carried by booster version F9 v1.1
    5. List the date when the first succesful landing outcome in ground pad was achieved
    6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
    7. List the total number of successful and failure mission outcomes
    8. List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
    9. 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.
    10. 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.

- GitHub URL of your completed EDA with SQL notebook:

https://github.com/nchlsmnro/IBM_Data_Sci_Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Using Folium, the launch sites were marked with circles via their latitude and longitude. Successful launches received a secondary green marker while failures received a red marker.
- The distance between the points indicating launches and points of interest like cities, highways, coastlines, and railway were measured.
    - From theses measurements, it was determined that launch sites are generally place near transportation routes and are 12.62 to 52.12 kilometers away from cities
- GitHub URL of your completed interactive map with Folium map: https://github.com/nchlsmnro/IBM_Data_Sci_Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Following the use of Folium for interactive map making, Ploty Dash was used to build an interactive dashboard.
- Pie charts were plotted which demonstrated the total launches by site.
- Scatter plots were produced exhibiting the relationship between launch outcome and payload mass in kilograms by the different booster versions.
- GitHub URL of your completed Plotly Dash lab: https://github.com/nchlsmnro/IBM_Data_Sci_Capstone/blob/main/plotlydashboardapp_for_spacexlaunches.ipynb

Predictive Analysis (Classification)

GitHub URL of completed predictive analysis lab:
https://github.com/nchlsmnro/IBM_Data_Sci_Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Building

- Utilized data was loaded and assigned to NumPy array. Next, the data was standardized to a uniform scale and split into training and test sets
- Logistic Regression, SVM, Decision Tree, and KNN models were produced

Evaluation

- Confusion Matrix

Improvement

- Accuracy improvement using feature engineering and algorithm tuning

Selection

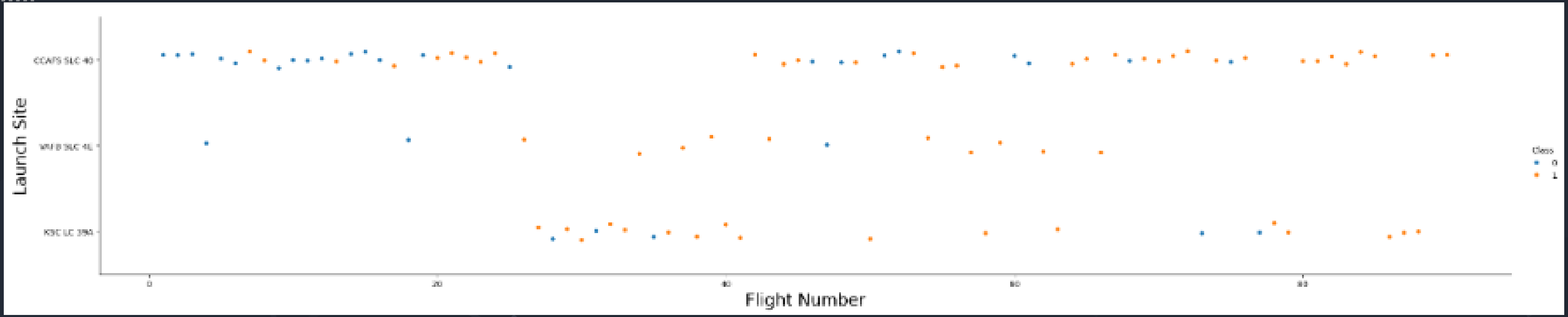
- The Decision Tree showed the best accuracy for the training and testing data and is selected as the best model

Results

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

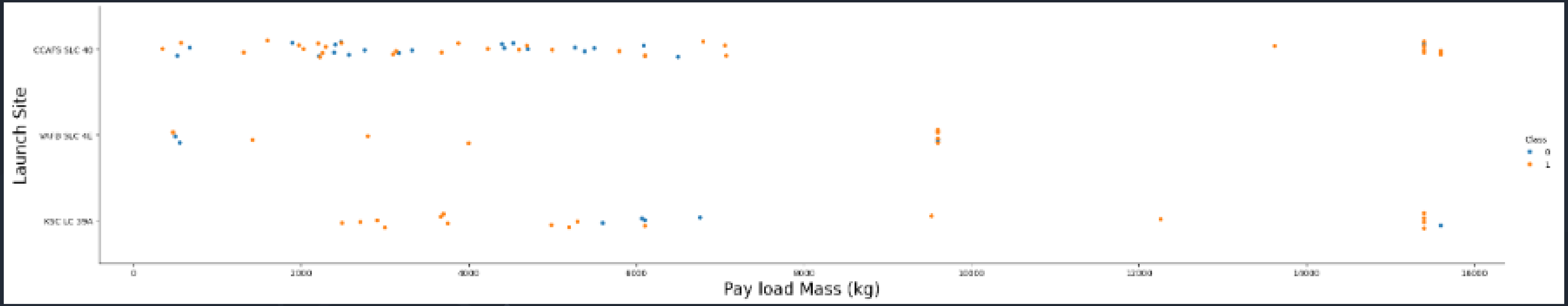
Section 2

Insights drawn from EDA



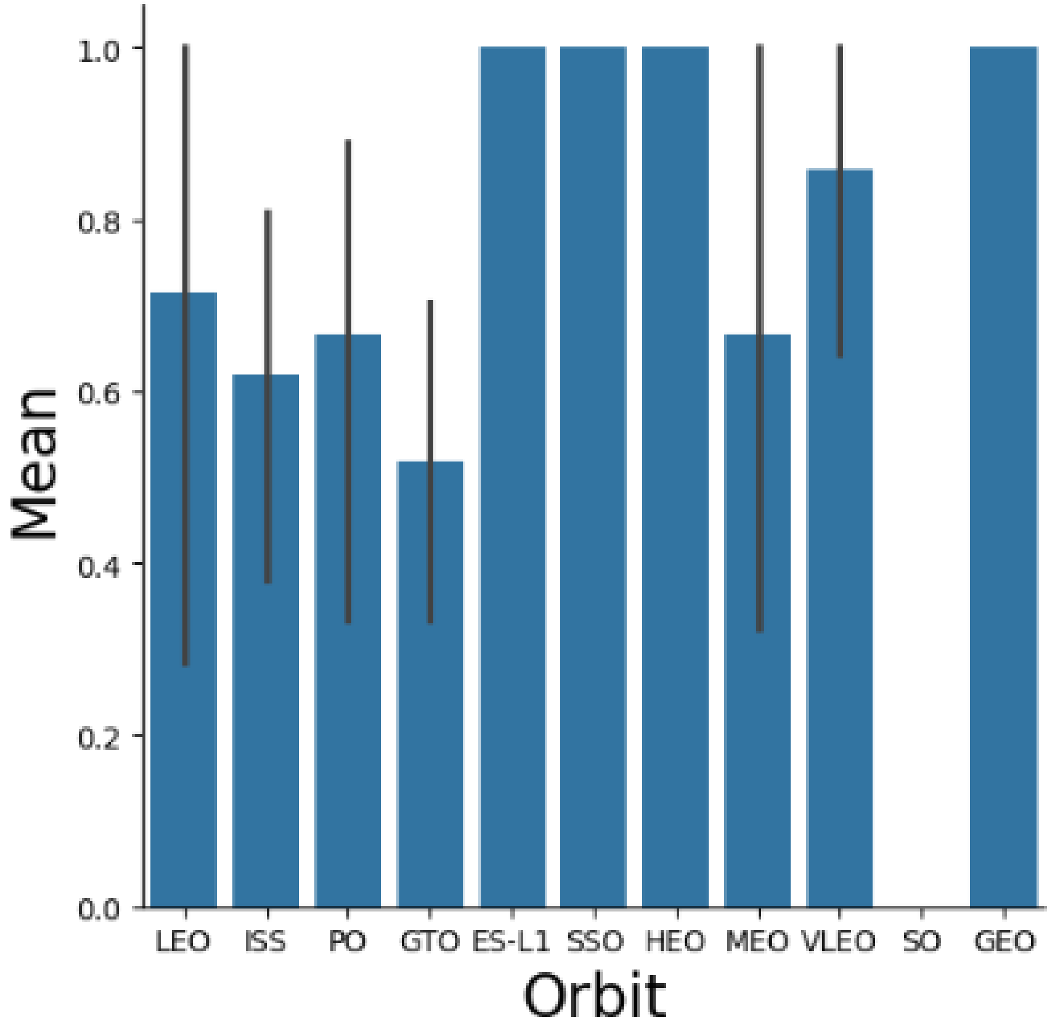
Flight Number vs. Launch Site

The scatter plot above demonstrates that the flight number increases the number of successful landings increases and that the number of successful landings increase as more flights are complete at a launch site



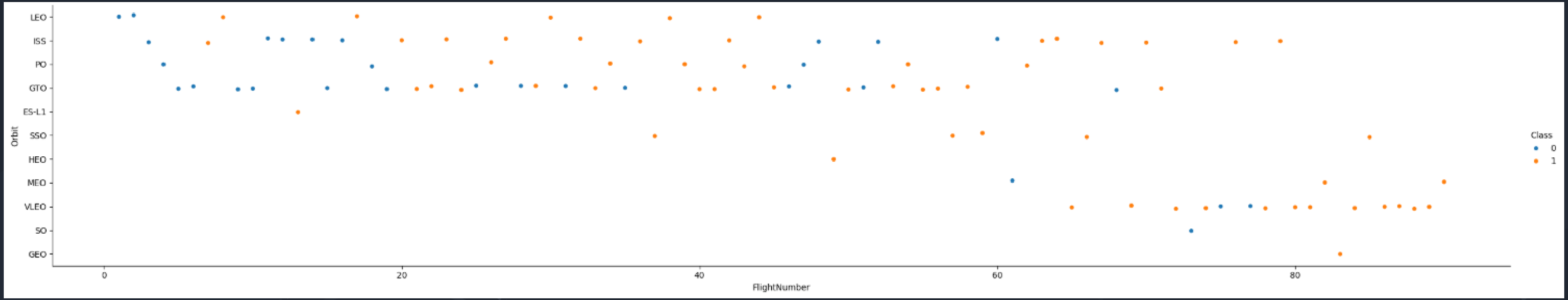
Payload (kg) vs. Launch Site

The scatter plot above demonstrates that the probability of a successful landing increases as the payload increases for Site CCAFS SLC 40.



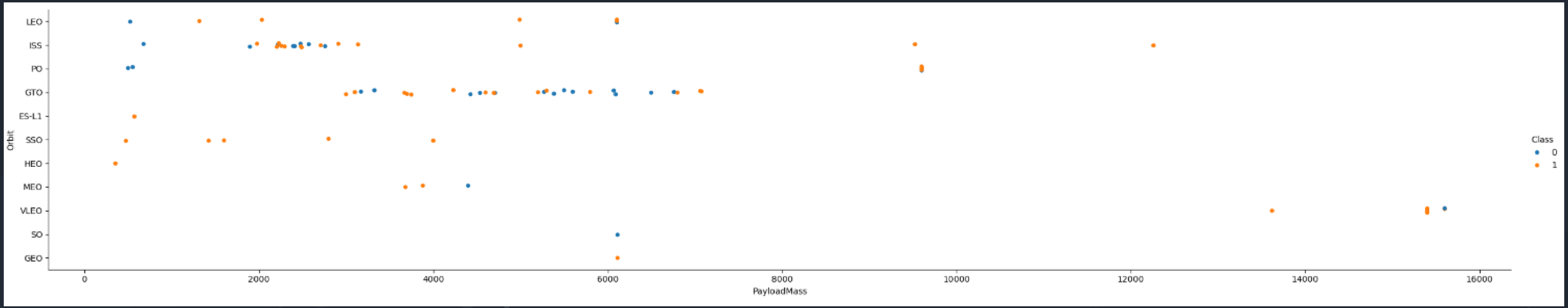
Success Rate vs. Orbit Type

The plot demonstrates that ES-L1, SSO, HEO, and GEO have 100% success rates, LEO, ISS, PO, GTO, MEO, and VLEO have middling success rates, and SO has a concerning 0% success rate.



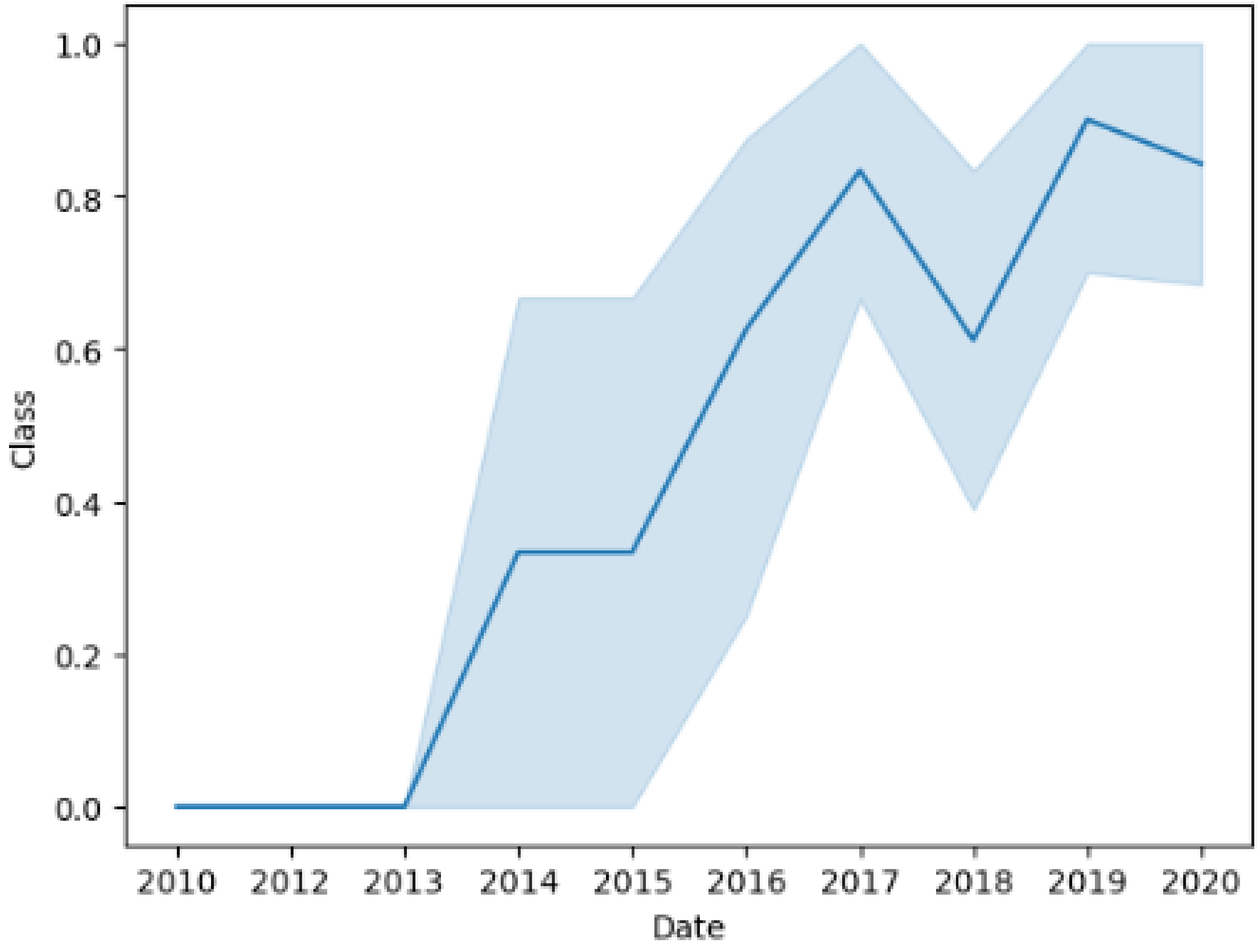
Flight Number vs. Orbit Type

The scatter point of Flight number vs. Orbit type shows that there is little or no distinguishable relationship between flight number and launch success when in ISS or GTO orbit



Payload vs. Orbit Type

This scatter plot shows that the GTO orbit missions have no correlation between payload and orbit type, PO, LEO and ISS orbit missions appear to have a positive correlation between payload and success rate. All other orbit types lack sufficient data for assessment.



Launch Success Yearly Trend

The line plot demonstrates a general trend of success rate improvement over time with a minor decrease in success rate in 2018.

All Launch Site Names

```
%sql select distinct Launch_Site from SPACEXTABLE
```

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

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

Using the SQL query DISTINCT the unique launch site names were called and outputted

Launch Site Names Begin with 'CCA'

- With the SQL query in the screenshot, five records where the launch site name began with ‘CCA’ was produced

In [11]:
```
%sql select Launch_Site from SPACEXTABLE where Launch_Site like 'CCA%' Limit 5
%sql select * from SPACEXTABLE
```

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

Out[11]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_(
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (pa
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 (pa
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	Nc
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	Nc
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	Nc

Total Payload Mass

With the query displayed left, the total payload carried by boosters from NASA is calculated: 48213 kg

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer like '%NASA (CRS)%'
```

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

sum(PAYLOAD_MASS__KG_)
48213

Average Payload Mass by F9 v1.1

With the query displayed left, the average payload mass carried by booster version F9 v1.1 is calculated: 2534.67 kg

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version like 'F9 v1.1%'
```

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

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

With the displayed query, the dates of the first successful landing outcome on ground pad are found: December 22, 2015

```
%sql select min(Date) from SPACEXTABLE where Landing_Outcome like '%success (ground pad)%'
```

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

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

With the query, a list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 is produced

```
%%sql SELECT BOOSTER_VERSION, payload_mass_kg, Landing_Outcome  FROM SPACEX
where 4000 < payload_mass_kg and payload_mass_kg < 6000 and Landing_Outcome = 'Success (drone ship)'
```

```
 * ibm_db_sa://xww99973:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/BLUDB
Done.
```

booster_version	payload_mass_kg	landing_outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

The query calculates the total number of successful and failure mission outcomes

```
%%sql SELECT Mission_Outcome, count(Mission_Outcome) as "Total" FROM SPACEX
Group by Mission_Outcome
```

```
 * ibm_db_sa://xww99973:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/BLUDB
Done.
```

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

Boosters Carried Maximum Payload

The query lists the names of the booster which have carried the maximum payload mass

```
%%sql SELECT Unique Booster_version, payload_mass_kg FROM SPACEX
where payload_mass_kg = (Select max(payload_mass_kg) from SPACEX)
```

```
 * ibm_db_sa://xww99973:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/BLUDB
Done.
```

booster_version	payload_mass_kg
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

2015 Launch Records

The query lists the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%%sql Select landing_outcome, booster_version, launch_site, DATE from SPACEX
where landing_outcome = 'Failure (drone ship)' and Year(DATE) =  2015
```

```
 * ibm_db_sa://xww99973:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/BLUDB
Done.
```

landing_outcome	booster_version	launch_site	DATE
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015-01-10
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015-04-14

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

The query ranks 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

```
%%sql select landing_outcome, count(landing_outcome) as "Total" from Spacex
where DATE between '2010-06-04' and '2017-03-20'
group by landing_outcome
order by "Total" desc
```

```
 * ibm_db_sa://xww99973:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/BLUDB
Done.
```

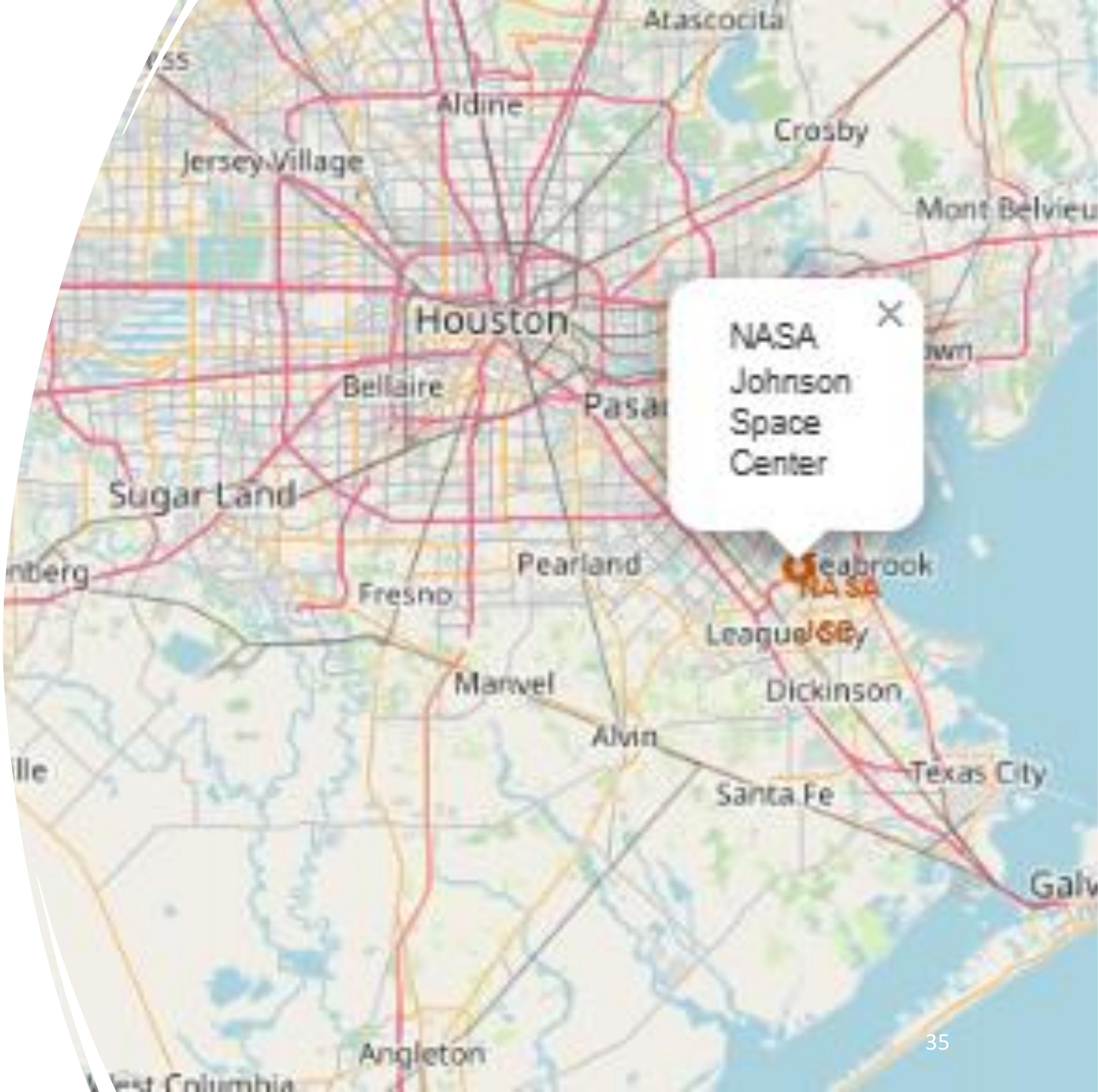
landing_outcome	Total
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

Houston, TX & the NASA Johnson Space Center

The initial map depicts Houston, Texas and the location of the NASA Johnson Space Center



Launch Sites VAFB SLC-4E, KSC LC-SSA, CCAFS LC-40 & CCAFS SLC-40

In the next map, the launch sites VAFB SLC-4E, KSC LC-SSA, and CCAFS SLC-40 are depicted

CCAFS SLC-40 Launch Site Proximity Map

In the map above, the green icons indicate successful launches and the red icons indicate failed launches that have occurred at CCAFS SLC-40. The blue line depicts the distance from the launch site to the coast (0.86 km)

The table shows the longitude and latitude of the launch site (CCAFS SLC-40) and the longitude and latitude and distance to the nearest coastline, railway, highway, and city.

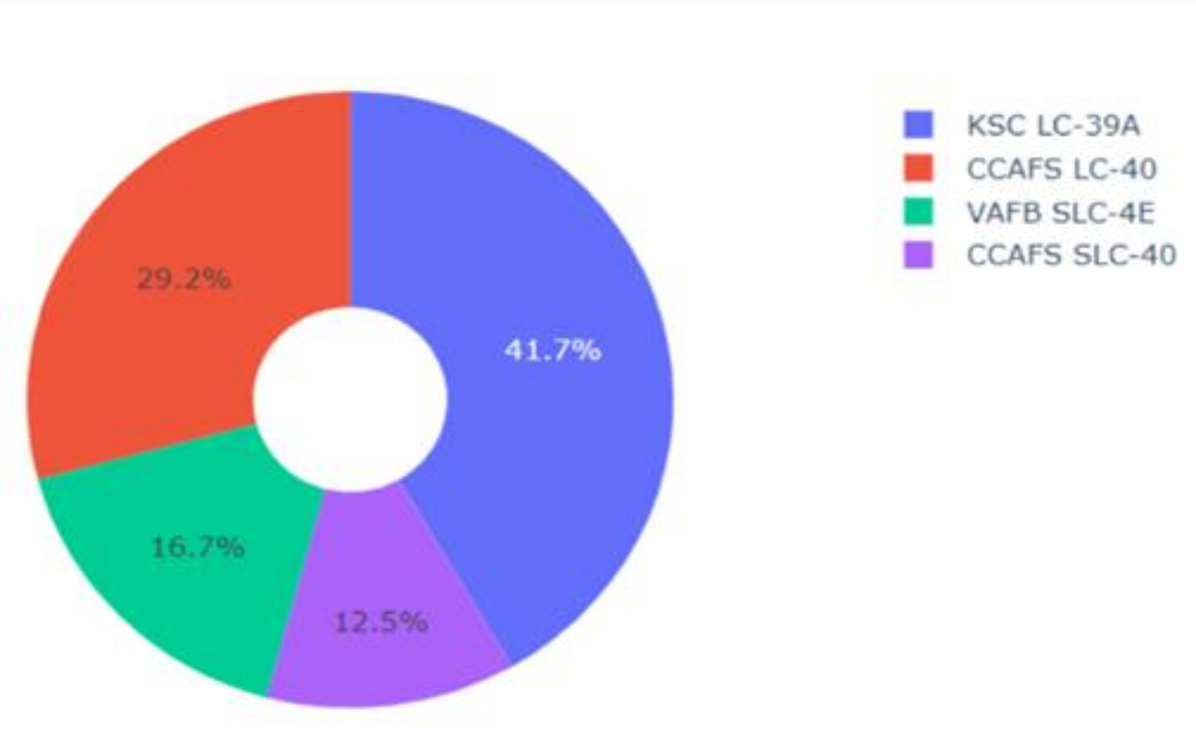
Launch Site	Proximity	LatLaunchSite	LongLaunchSite	Lat	Long	distance
CCAFS SLC-40	coastline	28.563197	-80.576820	28.56409	-80.56806	0.861525
	railway	28.563197	-80.576820	28.57217	-80.58528	1.295798
	highway	28.563197	-80.576820	28.56385	-80.57088	0.584818
	city	28.563197	-80.576820	28.10469	-80.64784	51.471476

Section 4

Build a Dashboard with Plotly Dash

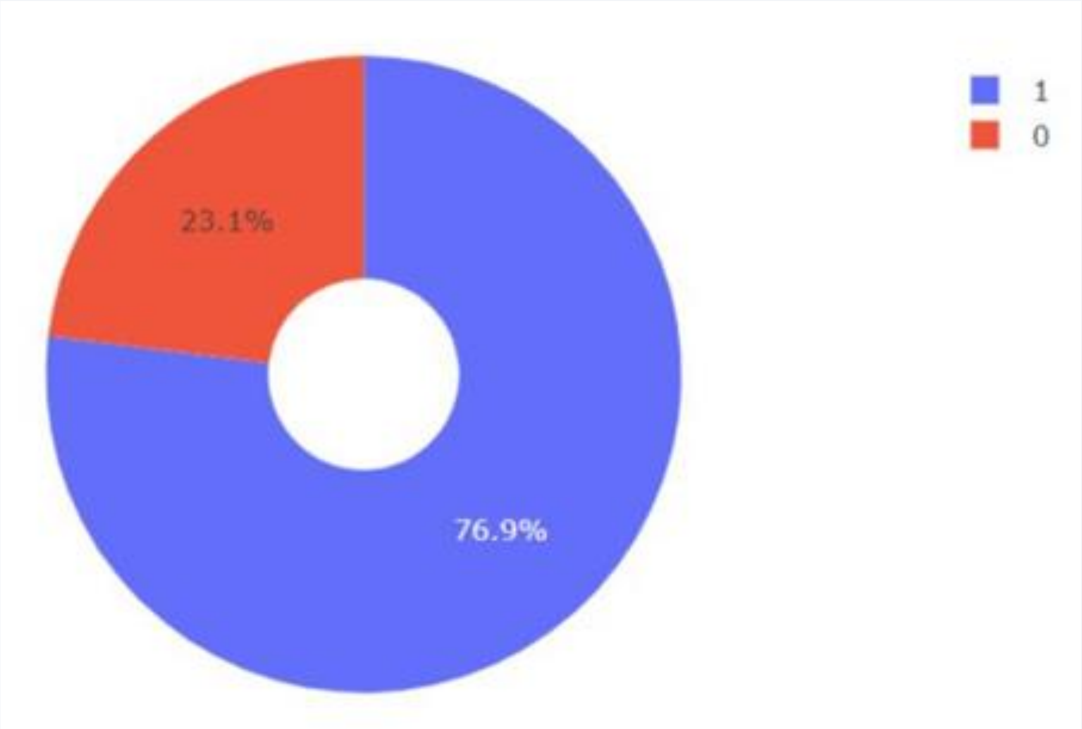
Percentage of Successful Launches by Site

The pie chart depicts the percentage of successful launches that are achieved at each site. The chart was generated and displayed on a dashboard using Plotly Dash.



Launch Site with highest success ratio

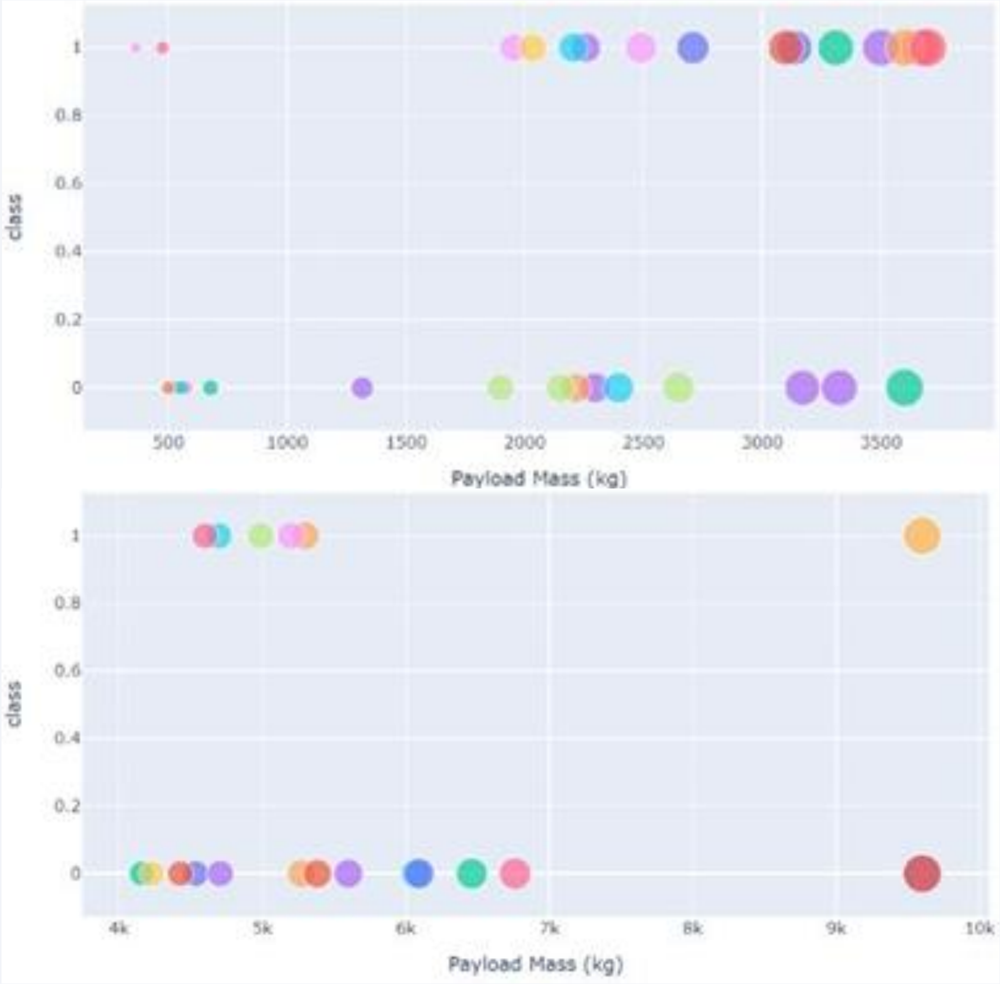
The pie chart depicts the launch site with the highest success ratio. The launch site is KSC LC-39A.



Payload and Launch Outcomes across launch site

The scatter plots depict Payload vs. Launch Outcome for all sites, with different payloads selected in the range slider.

The top plot depicts launches with payloads of between 0 and 4000 kilograms and the bottom plot depicts launches with payloads of between 4000 and 10,000 kilograms.



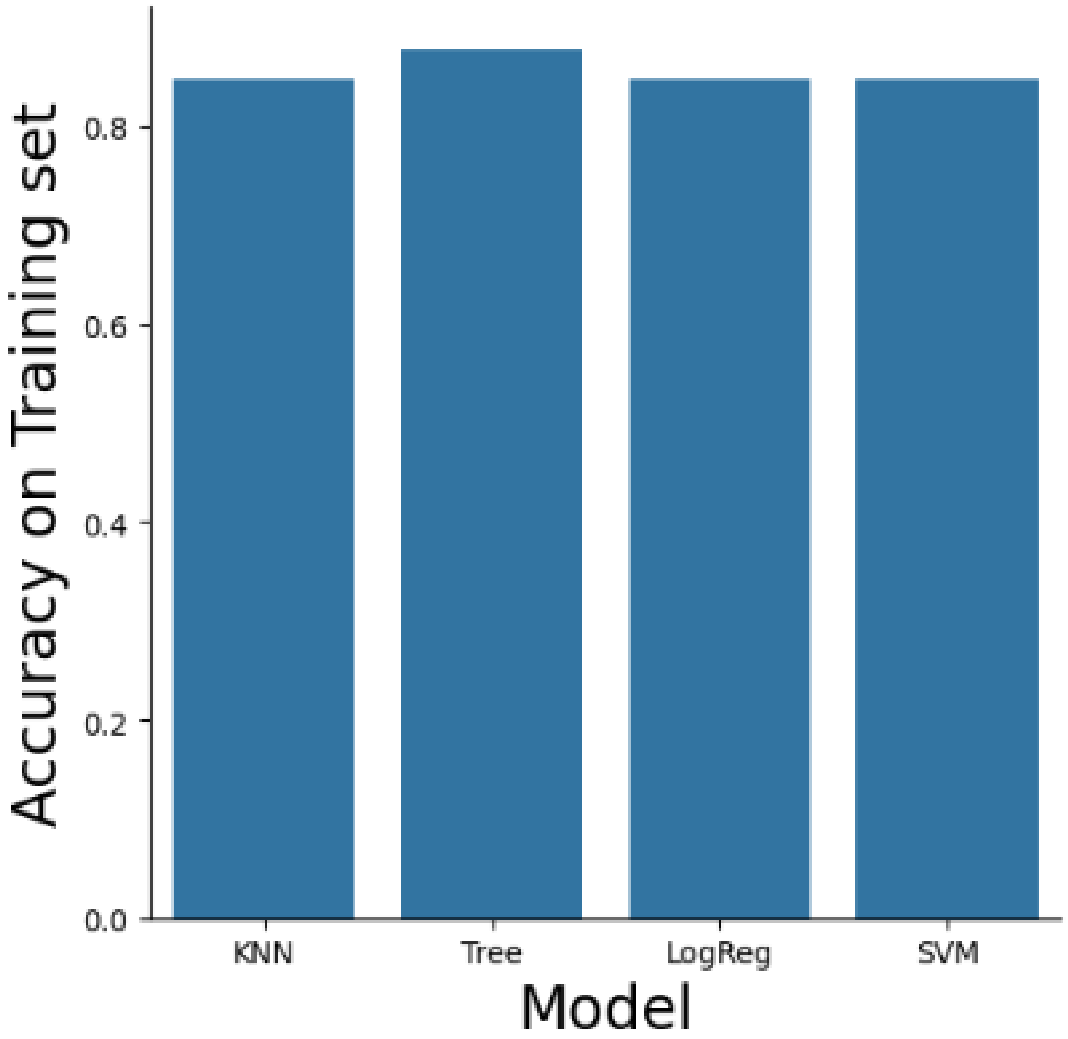
Section 5

Predictive Analysis (Classification)

Classification Accuracy

To the right is a bar chart illustrating the accuracy of each model in predicting the training set

The Decision Tree model is chosen as the best model with 94% accuracy



Confusion Matrix

The confusion matrix for the Decision Tree model shows that only a single launch was incorrectly predicted with a predicted successful landing when the launch failed to land

Conclusions

In this project, we set out to produce a model that would accurately predict the success of SpaceX's Falcon 9 rocket launches. Success was defined as the rocket's ability to land intact as this is a selling point for SpaceX due to the potential cost savings.

We determine that factors like payload mass and orbit type impact the outcome of the launches. We found that SpaceX has been able to consistently improve on the success rate of landings with minor hiccups along the way. Also, we assert that the KSC LC-39A has the highest success rate of the launch sites.

Finally, we built a decision tree model that predicted our test data with 94.44% accuracy.

Thank you!