IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Christian Chukov
2/18/2022

Outline

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

Executive Summary

- Methodologies used to analyze the data:
  - Data Collection for this project was accomplished through the SpaceX API, as well as web scraping the Wikipedia entry for Falcon 9 and Falcon Heavy launches;
  - Exploratory Data Analysis (EDA) methods used include: data wrangling, data visualization, interactive visual dashboards
  - Machine Learning prediction models.

- Summary of all results:
  - A variety of valuable data was collected from the sources listed;
  - EDA helped identify the significant characteristics of a successful launch;
  - The best-performing Machine Learning model was selected for the goals of this project.

Introduction

- Project background:
  - SpaceY, a new company is exploring potential entry to the commercial space transportation market;
  - The goal is to analyze the performance of SpaceX rocket launches based on historical data and assess optimal parameters for SpaceY’s own operation.
- Problem needed to be analyzed:
  - Determine the most cost-effective way to enter the market by analyzing:
    - success rate of SpaceX launches
    - location data of launch sites
    - payload volumes
    - orbital placement

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.
- You need to present your data collection process use key phrases and flowcharts

Data Collection – SpaceX API

- SpaceX API:

https://api.spacexdata.com/v4/launches/past

  - The SpaceX API contains vast amounts of information regarding the launch history of the company
  - For the purposes of this project the data was filtered to only include the relevant specifications
  - The resulting data frame consists of 90 rows and 17 columns
  - The values were carefully inspected, and data wrangling was applied to the data frame in order to replace missing entries

Initial SpaceX API request

Data was requested, parsed and normalized into a Pandas data frame, after identifying the desired columns for the project

Filter data to only include Falcon 9 launch information

Resulting data frame was reduced to 90 entries

Data wrangling identified missing values

Replace missing values

The mean value for Payload Mass was calculated and applied to the missing entries

Landing Pad was allowed to retain its missing values

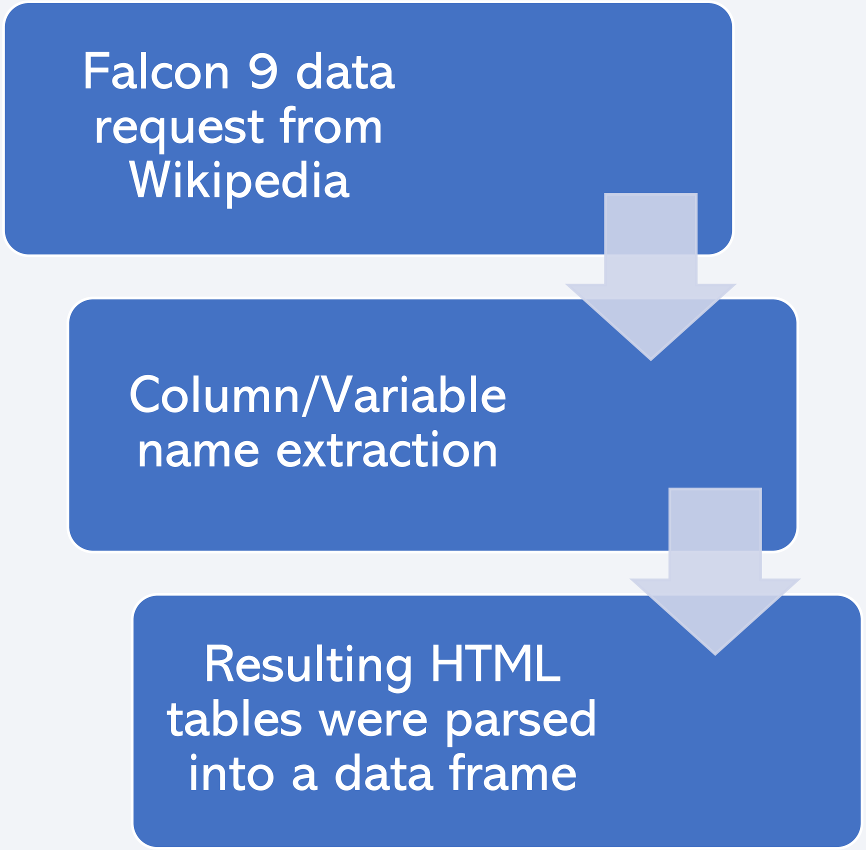
Reference: https://github.com/chrischukov/ds_capstone/blob/c0afdddd65d93188296210b32b61a6acfeac523e/1-jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

- Wikipedia: List of Falcon 9 and Falcon heavy launches

https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

  - Additional data was collected from Wikipedia using web scraping
  - During the extraction process empty columns were excluded
  - A formatting program was applied to handle any formatting errors, noise or reference links present in the original Wikipedia HTML tables
  - A data frame was then created with consistent values



Reference: https://github.com/chrischukov/ds_capstone/blob/79240f326cf9dd23c9ec9b92f378affa71ebb74a/2-webscraping.ipynb

Data Wrangling

- EDA was applied to the dataset and some initial insights were gained
- Main focus on Launch sites and Orbit types
- Overall success rate was calculated at 67%

1.
- Additional missing/null values were identified
- Data types were analyzed (categorical vs. numerical)

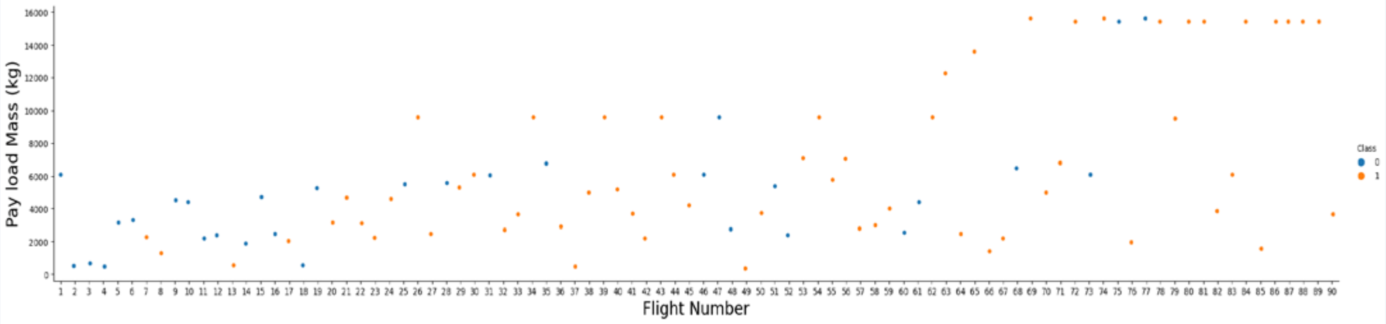
2.
- Launch count was established according to both launch sites and orbit type parameters

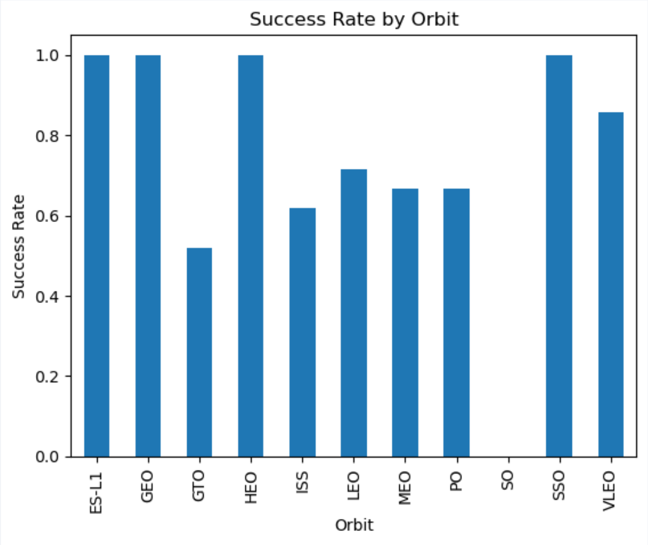
3.
- A mission outcome class label was created with 0/1 values for success/failure

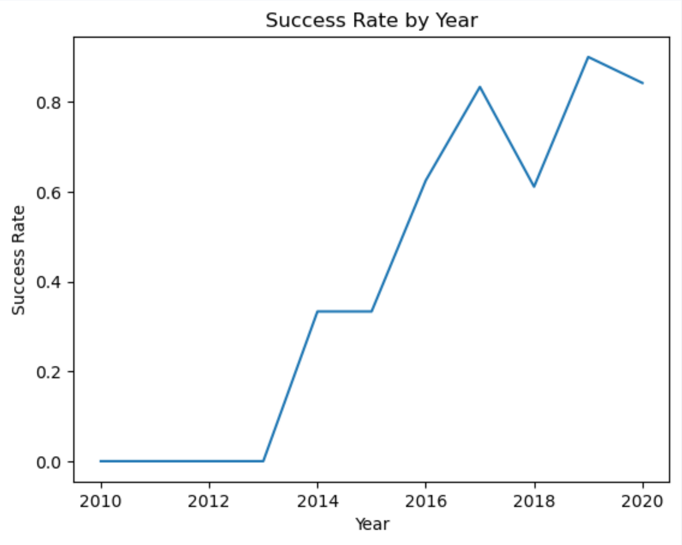
Reference: https://github.com/chrischukov/ds_capstone/blob/c52dedc29a00a2a30dee1cd18ec0b51545e8ddb0/3-labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Various relationships between the dataset variables were explored during the visualization stage, incl.: Flight Number, Launch Site, Payload Mass, Orbit, Success rate
- The primary visualization tools used here are scatter plot, bar chart and line plot utilizing the capabilities of the Seaborn and MatPlotLib libraries







Reference: https://github.com/chrischukov/ds_capstone/blob/e4652fcdd344268aecff72d28eedb57543cd21ae/5-jupyter-labs-eda-dataviz.ipynb

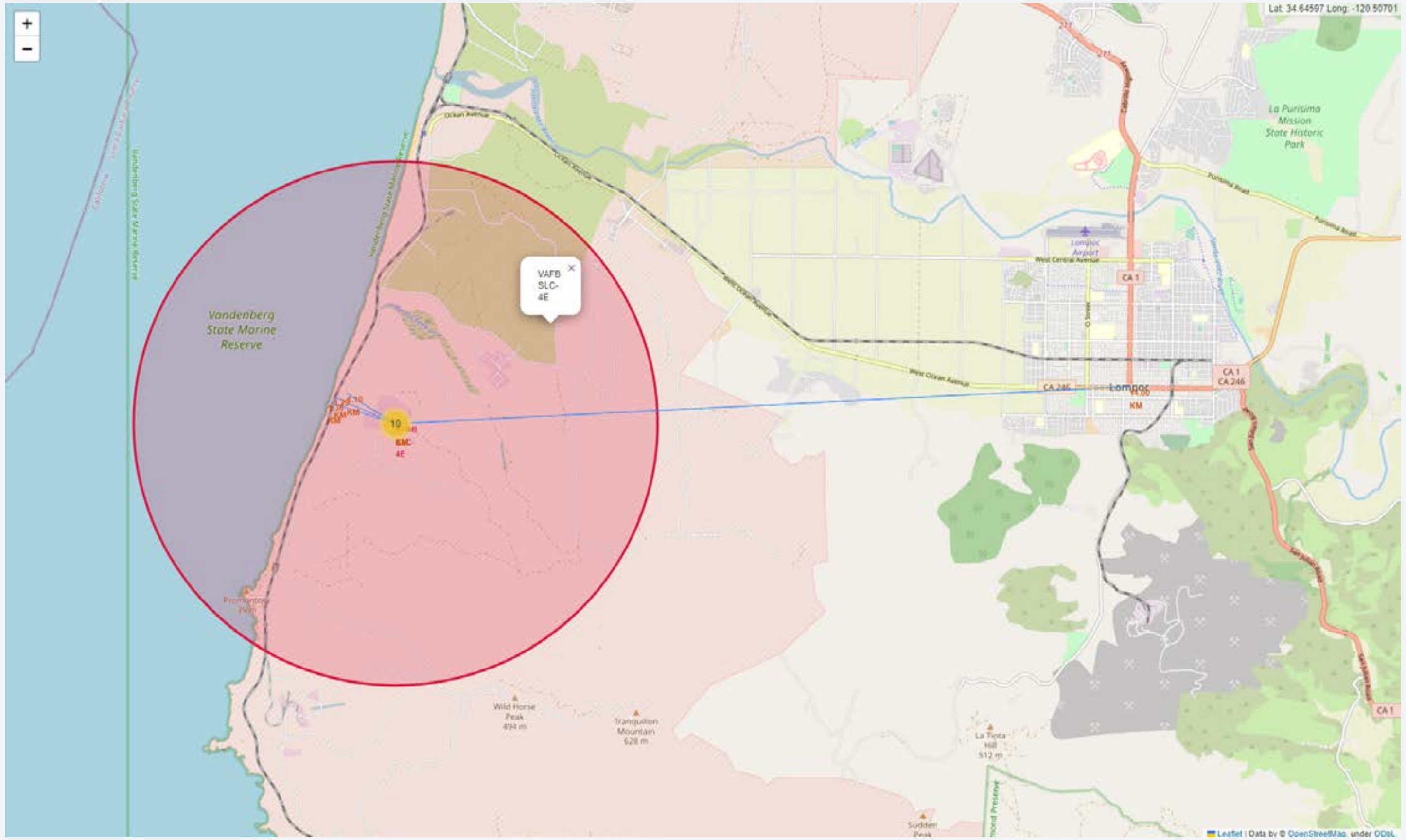
EDA with SQL

- SQL queries were performed on the dataset in order to gain additional insights:
  - Unique launch sites
  - Launch sites beginning with 'CCA'
  - Total payload mass for customer NASA
  - Average payload mass carried by F9 v1.1
  - First successful ground pad landing date
  - Boosters carrying payload between 4000 and 6000kg that landed successfully on drone ship
  - Total number of successful and failed missions
  - Booster version that carried max payload
  - Months of 2015 when boosters failed landing
  - Rank of successful landings between 04-06-2010 and 20-03-2017 (descending order)

Reference: https://github.com/chrischukov/ds_capstone/blob/deb61fe3fc612b873a350ff8ae26d1d2ec043efe/4-jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

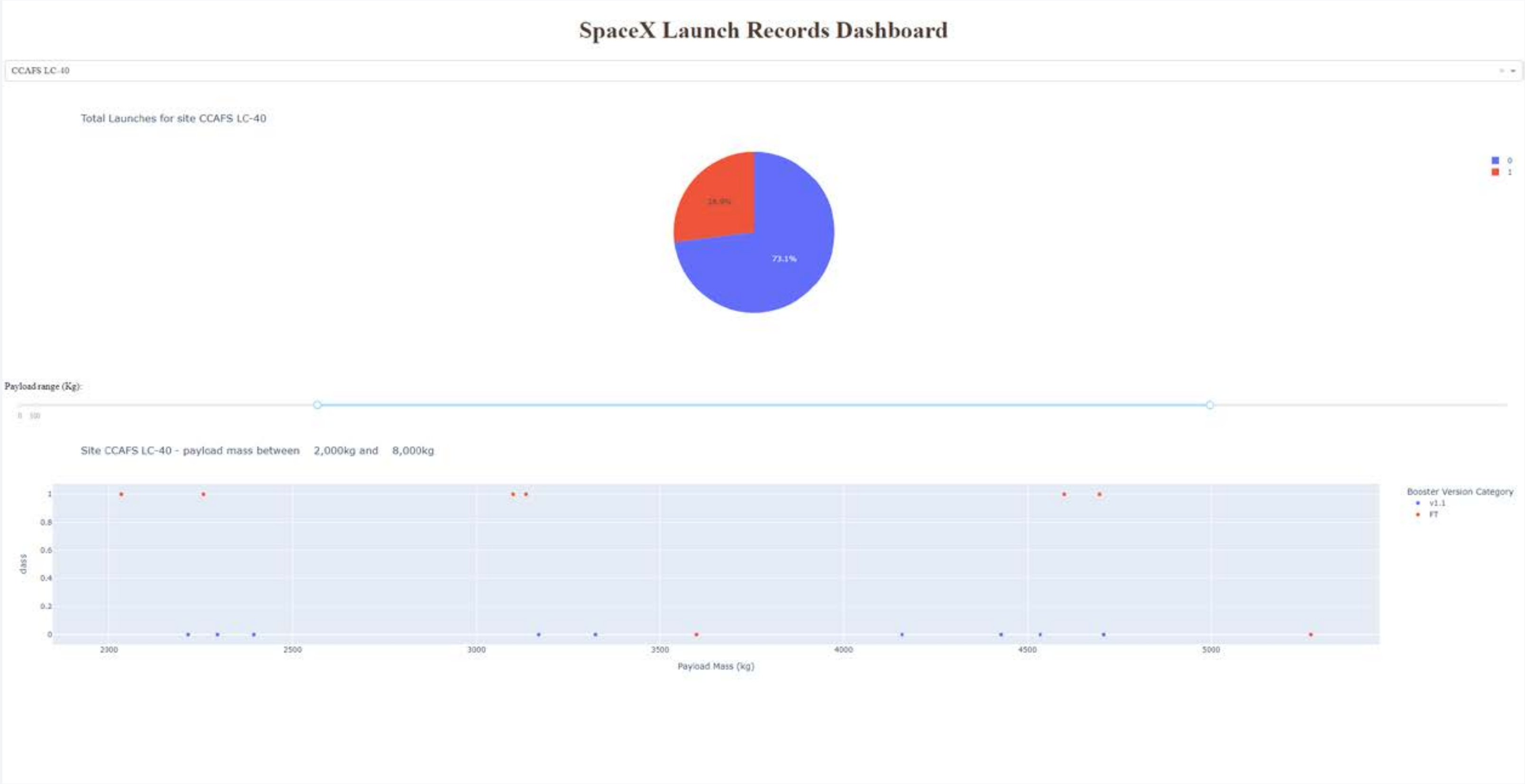
- Utilizing the capabilities of Folium, an interactive map was created with the following objects:
  - Marker (at location of a launch facility)
  - Circle (of the general vicinity of a launch facility
  - Marker cluster (showcasing successful/failed missions)
  - Lines and distance markers (establishing the proximity of a launch facility to the closest city, coastline, railway and highway



Reference: https://github.com/chrischukov/ds_capstone/blob/1f21790e760040743fc8932b978c2f55d637e28b/6-ab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

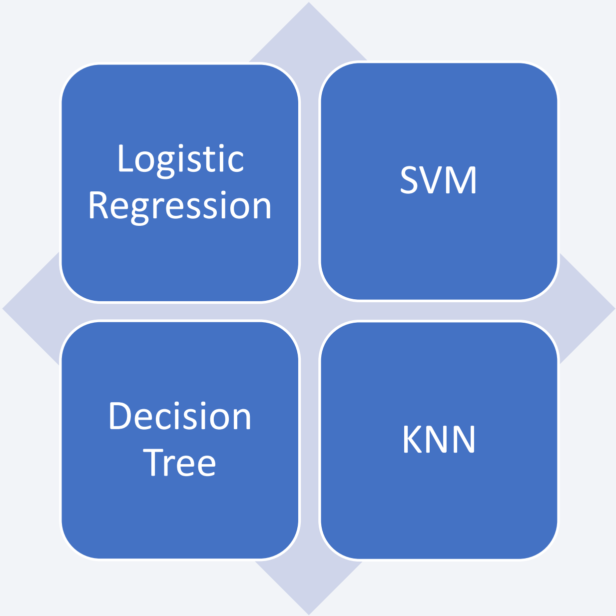
- The interactive dashboard includes:
  - Dropdown menu to toggle between individual launch sites
  - Scatterplot with information about the booster type, payload mass and success/failure class of each mission
  - A slider which to select the payload range to be displayed on the scatterplot
- The flexibility of this dashboard makes it easy to quickly retrieve the desired information for each of the launch sites as well as the parameters for all launch sites combined



Reference: https://github.com/chrischukov/ds_capstone/blob/a66d786124522a1a80af8c0dc8f95e95cba9882c/7-plotly-dash.ipynb

Predictive Analysis (Classification)

- During the predictive analysis stage, the data was split into training and testing sets and four separate Machine Learning models were applied
    - Logistic Regression
    - Support Vector Machine
    - Decision Tree
    - K-nearest neighbors (KNN)
- The accuracy of each model was established, models were ranked by accuracy and the best performing one was selected



	Model	Accuracy	TestAccuracy
2	Tree	0.889286	0.833333
3	KNN	0.848214	0.833333
1	SVM	0.848214	0.833333
0	LogReg	0.846429	0.833333

Reference: https://github.com/chrischukov/ds_capstone/blob/b31eb183cb05655150059de7d959f8b7401820ca/8-SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

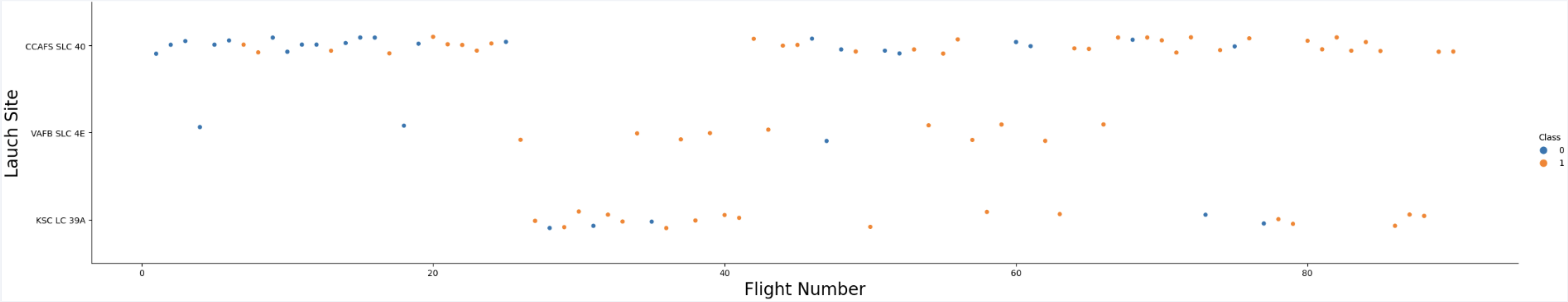
Results

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

Section 2

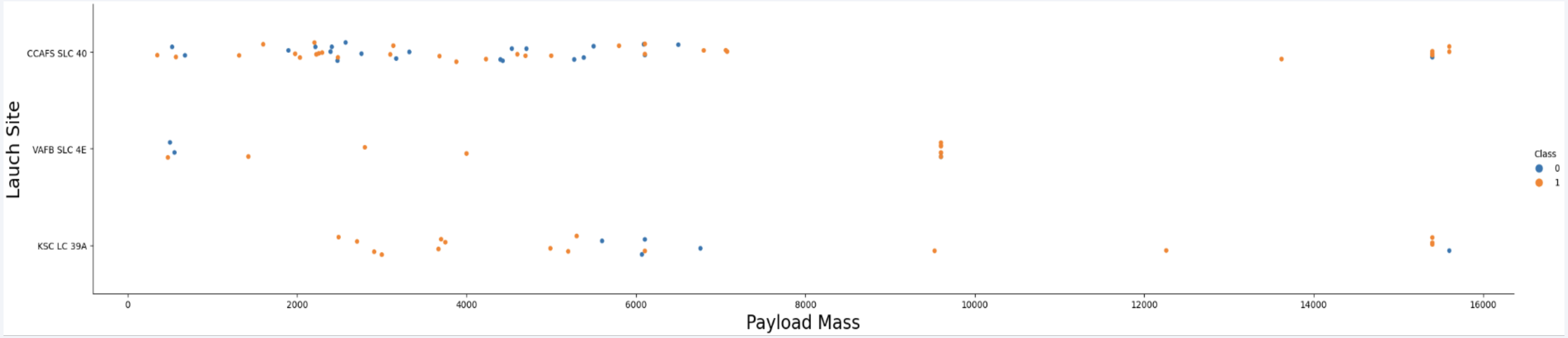
Insights drawn from EDA

Flight Number vs. Launch Site



- This scatter plot shows the relationship between Flight number and Launch site, with respective class labels for each record:
  - Generally, as the flight number increases, so does the success rate for each site
  - The most heavily used launch site is CCASF SLC-40, which also leads to it having the lowest success rate with 60% due to a large number of failed missions in the initial stages of SpaceX's operation
  - The success rates for both KSC LC 39A and VAFB SLC 4E were calculated at 77%

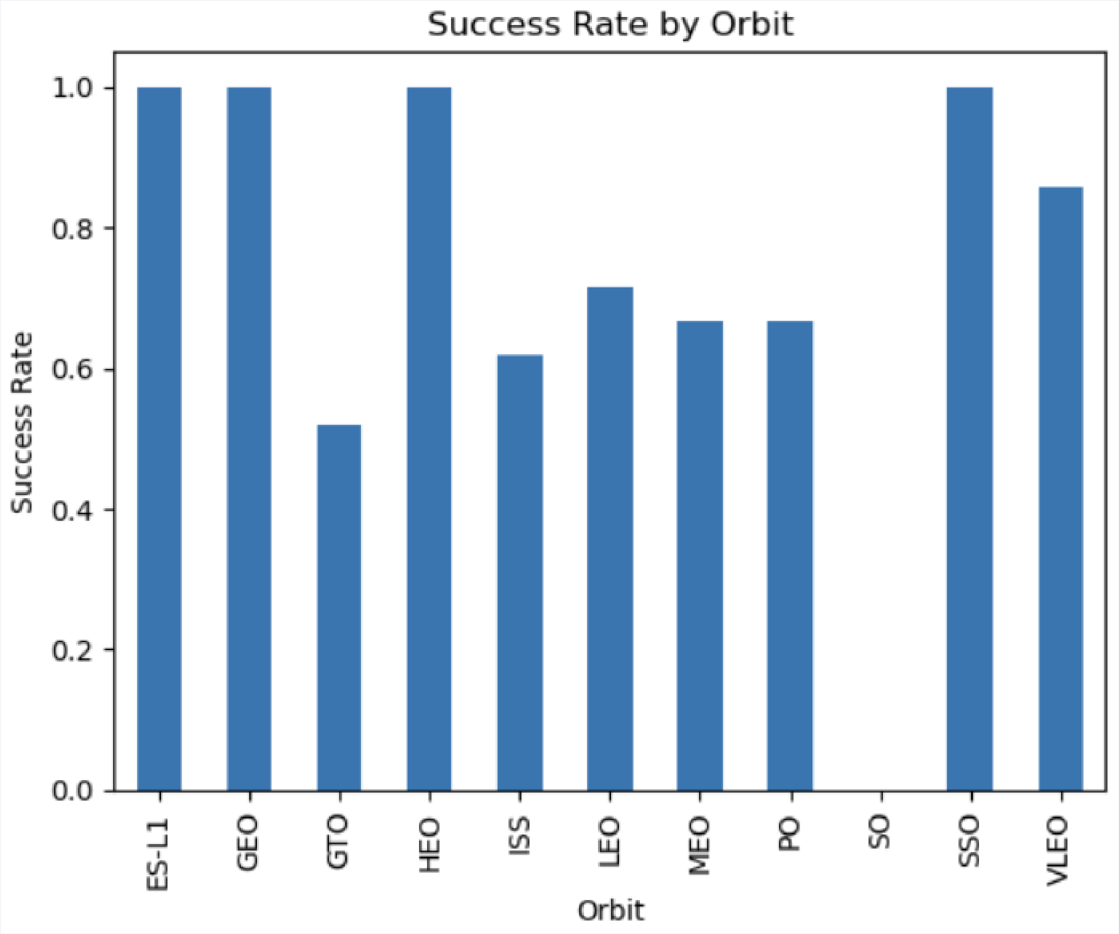
Payload vs. Launch Site



- This scatter plot visualizes the relationship between Payload mass and Launch site:
  - Generally, heavier payloads (above 9000kg) have higher success rate, however there are fewer records for those and only for launch sites CCASF SLC-40 and KSC LC 39A
  - Payloads heavier than 10000kg have not been launched from launch site VAFB SLC 4E
  - Mission outcome for payloads under 7000kg is very mixed and clear relationship can not be established

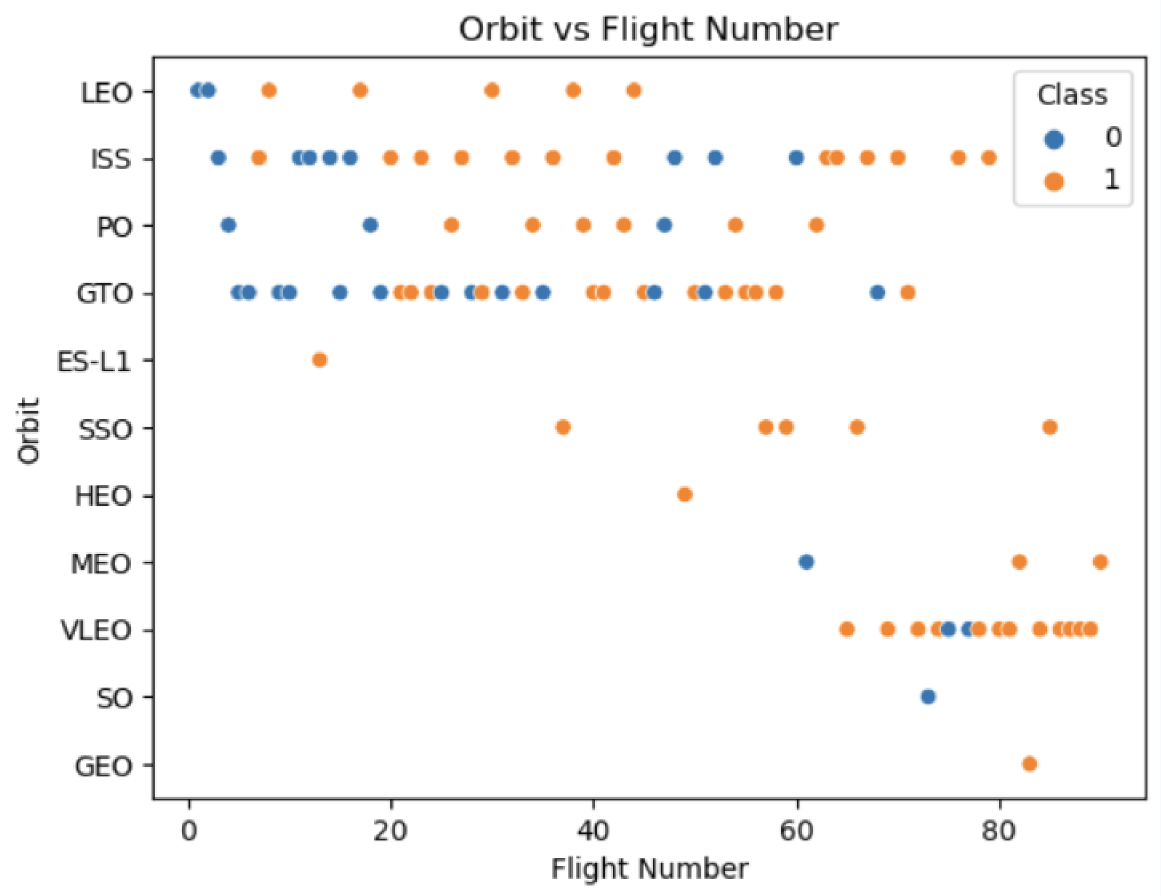
Success Rate vs. Orbit Type

- This bar chart visualizes the relationship between Orbit and Success rate:
  - Missions launched into orbits ES-L1, GEO, HEO and SSO all have 100% success rate
  - VLEO orbit results in a high success rate of over 80%
  - The least dependable orbit placements are SO and GTO with respective success rates of 0% and 50%



Flight Number vs. Orbit Type

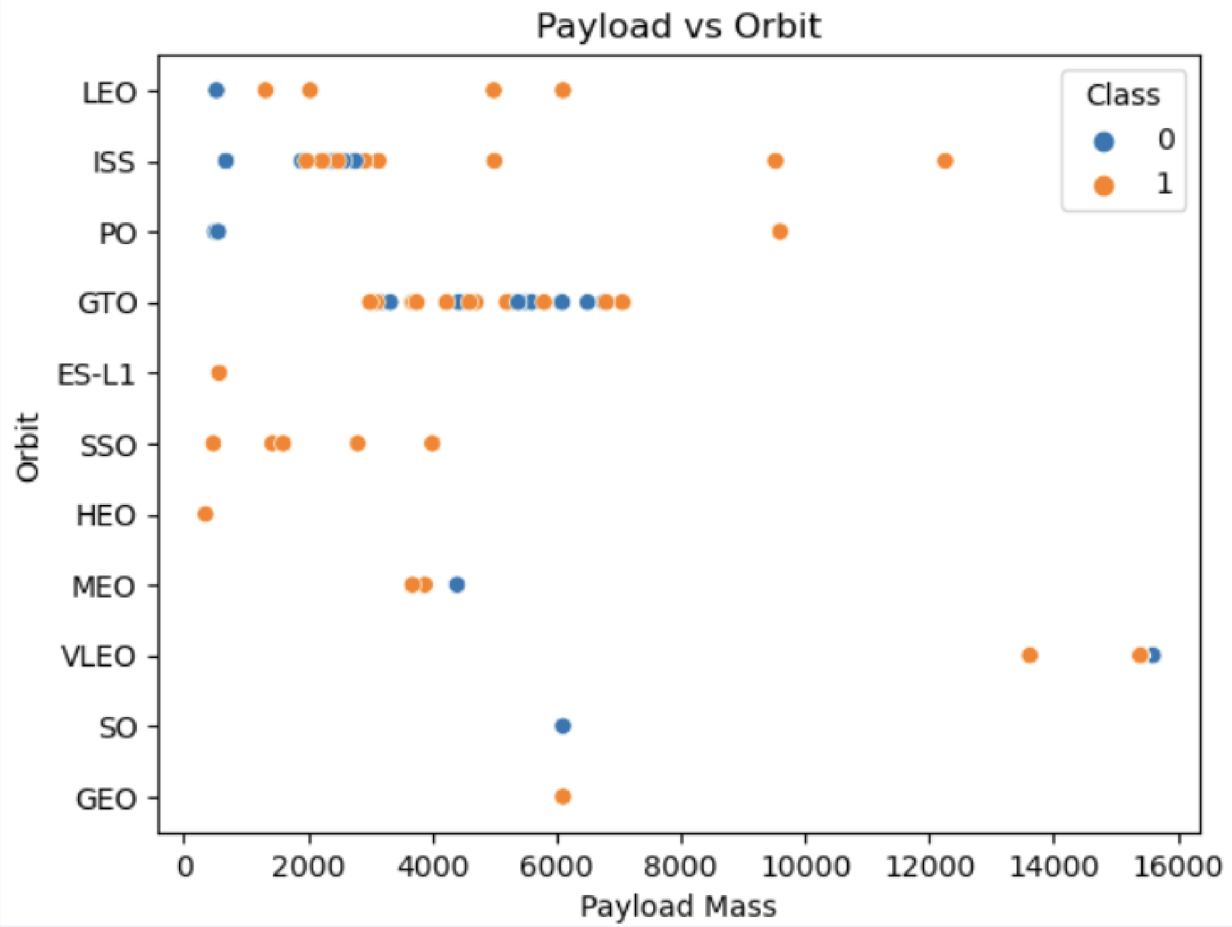
- The following scatter plot shows the relationship between orbit type and flight number:



  - Overall, as the flight number increases, the variety of orbit types grows
  - VLEO seems to be the preferred orbital placement
  - Later flights produce more successful outcomes for all orbits

Payload vs. Orbit Type

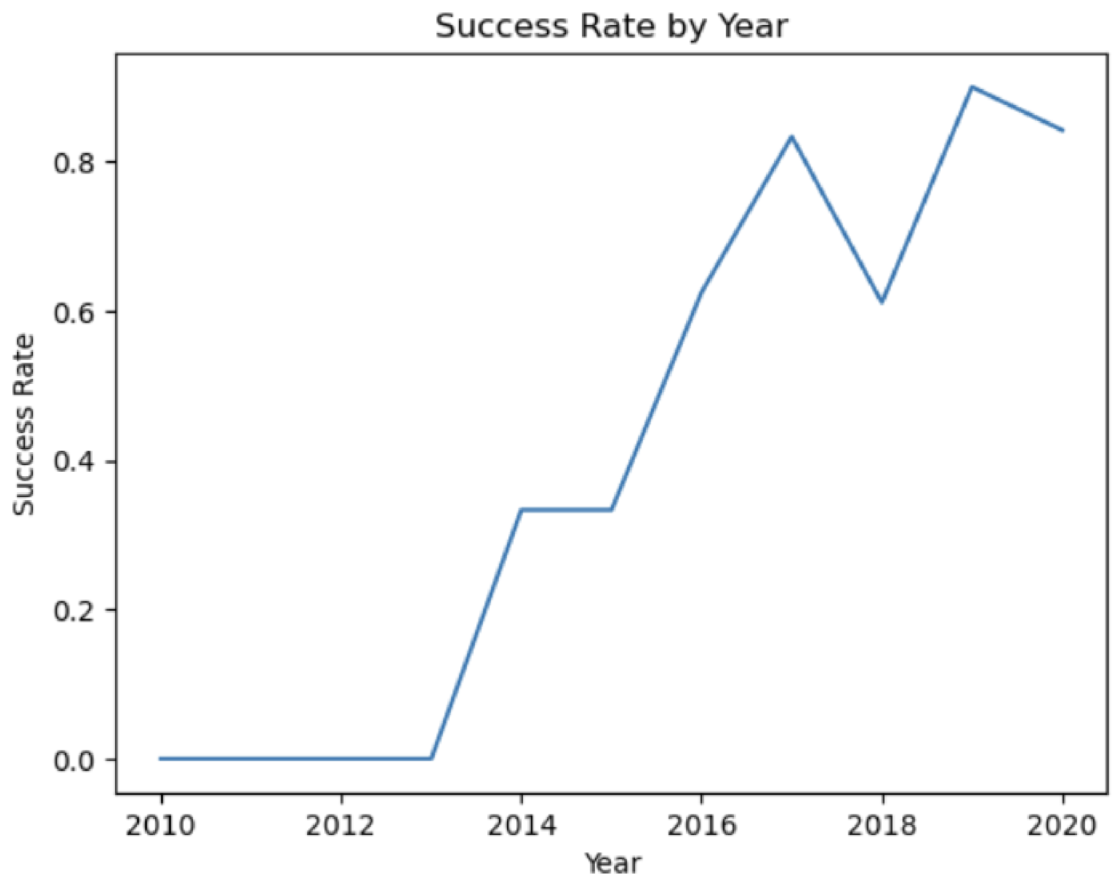
- This scatter plot shows the relationship between Payload and Orbit type:
  - Payloads under 8000kg have been launched into a variety of orbits, GTO and ISS being the most frequently used
  - Payloads heavier than 8000kg have been launched into ISS, PO and VLEO, however there are not enough records to assess viability



Launch Success Yearly Trend

- This line chart represents the success rate of SpaceX's missions since 2010:

  - There is a clear trend of increasing success rate as time goes by and SpaceX scales up and perfects its operation
  - The 20% dip in 2017 was heavily publicized, with a variety of statements from CEO Elon Musk from that time period pointing towards problems with used grid fins as well as a high-profile ZUMA mission failing to achieve desired orbit and being lost



All Launch Site Names

- The displayed query was used to list all four unique launch sites used by SpaceX:
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40

```
query = "SELECT DISTINCT Launch_site FROM SPACEXTBL;"
cur.execute(query)
launch_sites = cur.fetchall()
print("Unique Launch Sites:")
for site in launch_sites:
    print(site[0])
```

```
Unique Launch Sites:
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
```

Launch Site Names Begin with 'CCA'

```
query = "SELECT * FROM SPACEXTBL WHERE Launch_site LIKE 'CCA%' LIMIT 5;"
cur.execute(query)

results = cur.fetchall()
for result in results:
    print(result)
```

```
('04-06-2010', '18:45:00', 'F9 v1.0  B0003', 'CCAFS LC-40', 'Dragon Spacecraft Qualification Unit', 0, 'LEO', 'SpaceX', 'Success', 'Failure (parachut
e)')
('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 (COT
S) 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')
```

- The displayed query returned the first 5 mission records for launch site name beginning with ‘CCA’
    - Each entry has various types of information, including Date, Time, Orbit, Customer, Outcome etc.

Total Payload Mass

```
query = "SELECT sum(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer='NASA (CRS)';"
cur.execute(query)
total_mass = cur.fetchall()
print("Total payload mass carried by boosters launched by NASA (CRS):", total_mass[0][0], "kg")
```

```
Total payload mass carried by boosters launched by NASA (CRS): 45596 kg
```

- The displayed query calculated the total payload mass carried by boosters for SpaceX customer NASA at 45596 kg

Average Payload Mass by F9 v1.1

```
query = "SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version = 'F9 v1.1';"
cur.execute(query)
result = cur.fetchall()
print("Average payload mass carried by F9 v1.1:", result[0][0])
```

```
Average payload mass carried by F9 v1.1: 2928.4
```

- This query returned the average payload mass carried by Falcon 9 v1.1 boosters with a value of 2928.4 kg

First Successful Ground Landing Date

```
query = "SELECT min(Date) FROM SPACEXTBL WHERE Landing_Outcome = 'Success (ground pad)' ORDER BY Date LIMIT 1;"

cur.execute(query)
result = cur.fetchone()
print(result[0])
```

```
01-05-2017
```

- This query found the first successful landing date using a ground pad, 01-05-2017
    - It is notable that it took SpaceX seven years to perfect its automated landing technology when using ground pads

Successful Drone Ship Landing with Payload between 4000 and 6000

```
query = "SELECT Booster_Version FROM SPACEXTBL WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ >= 4000 AND PAYLOAD_MASS__KG_ < 60

cur.execute(query)
result = cur.fetchall()

print("The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000:")
for row in result:
    print(row[0])
```

```
The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000:
F9 FT B1022
F9 FT B1026
F9 FT  B1021.2
F9 FT  B1031.2
```

- This query returned a list of the boosters that landed successfully on a drone ship after carrying payloads between 4000 and 6000 kg

Total Number of Successful and Failure Mission Outcomes

```
query = "SELECT Mission_Outcome, COUNT(Mission_Outcome) FROM SPACEXTBL GROUP BY Mission_Outcome;"

cur.execute(query)
result = cur.fetchall()
for outcome, count in result:
    print("Outcome: {}\tCount: {}".format(outcome, count))
```

```
Outcome: Failure (in flight)	Count: 1
Outcome: Success	Count: 98
Outcome: Success 	Count: 1
Outcome: Success (payload status unclear)	Count: 1
```

- This query returned the total count of successful and failed mission outcomes

Boosters Carried Maximum Payload

```
query = "SELECT Booster_version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT max(PAYLOAD_MASS__KG_) FROM SPACEXTBL)"

cur.execute(query)
result = cur.fetchall()
for r in result:
    print(r[0])
```

```
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
```

- This query presented a list of booster versions that have carried maximum payload

2015 Launch Records

```
query = "SELECT substr(Date, 4, 2) as Month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTBL WHERE Landing_Outcome = 'Failure (drone ship

cur.execute(query)
result = cur.fetchall()
for row in result:
    print(row)
```

```
('01', 'Failure (drone ship)', 'F9 v1.1 B1012', 'CCAFS LC-40')
('04', 'Failure (drone ship)', 'F9 v1.1 B1015', 'CCAFS LC-40')
```

- This query lists all the failed missions that took place during 2015, including month number, type of attempted landing, booster version and launch site

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

```
query = "SELECT Landing_Outcome, COUNT(Landing_Outcome) AS Successful_Landing FROM SPACEXTBL WHERE Landing_Outcome IN ('Success (ground pad)', 'Succes

cur.execute(query)
result = cur.fetchall()
for row in result:
    print(row)

con.close()
```

```
('Success', 20)
('Success (drone ship)', 8)
('Success (ground pad)', 6)
```

- This query returned a list of mission outcomes that took place between 2010-06-04 and 2017-03-20 in descending order

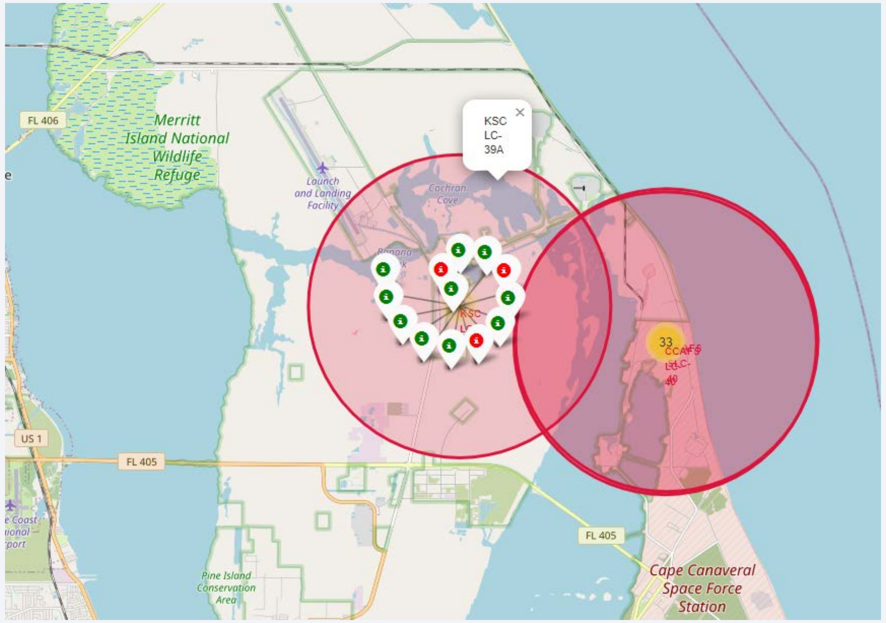
Section 3

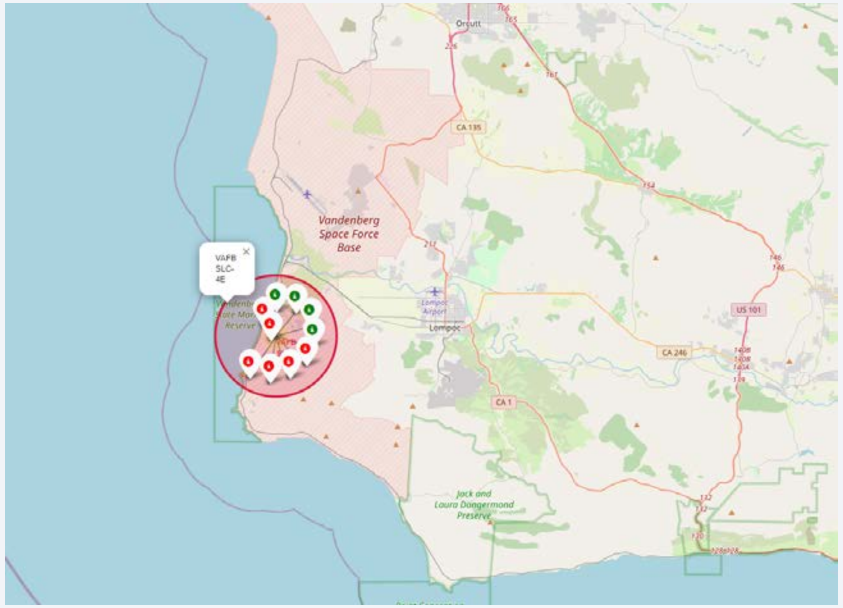
Launch Sites Proximities Analysis

SpaceX launch sites

- On the provided map, the SpaceX launch sites are marked with their names on the east and west coast of the US
  - The locations are carefully selected for safety reasons while also providing access to crucial infrastructure

Launch outcomes by site

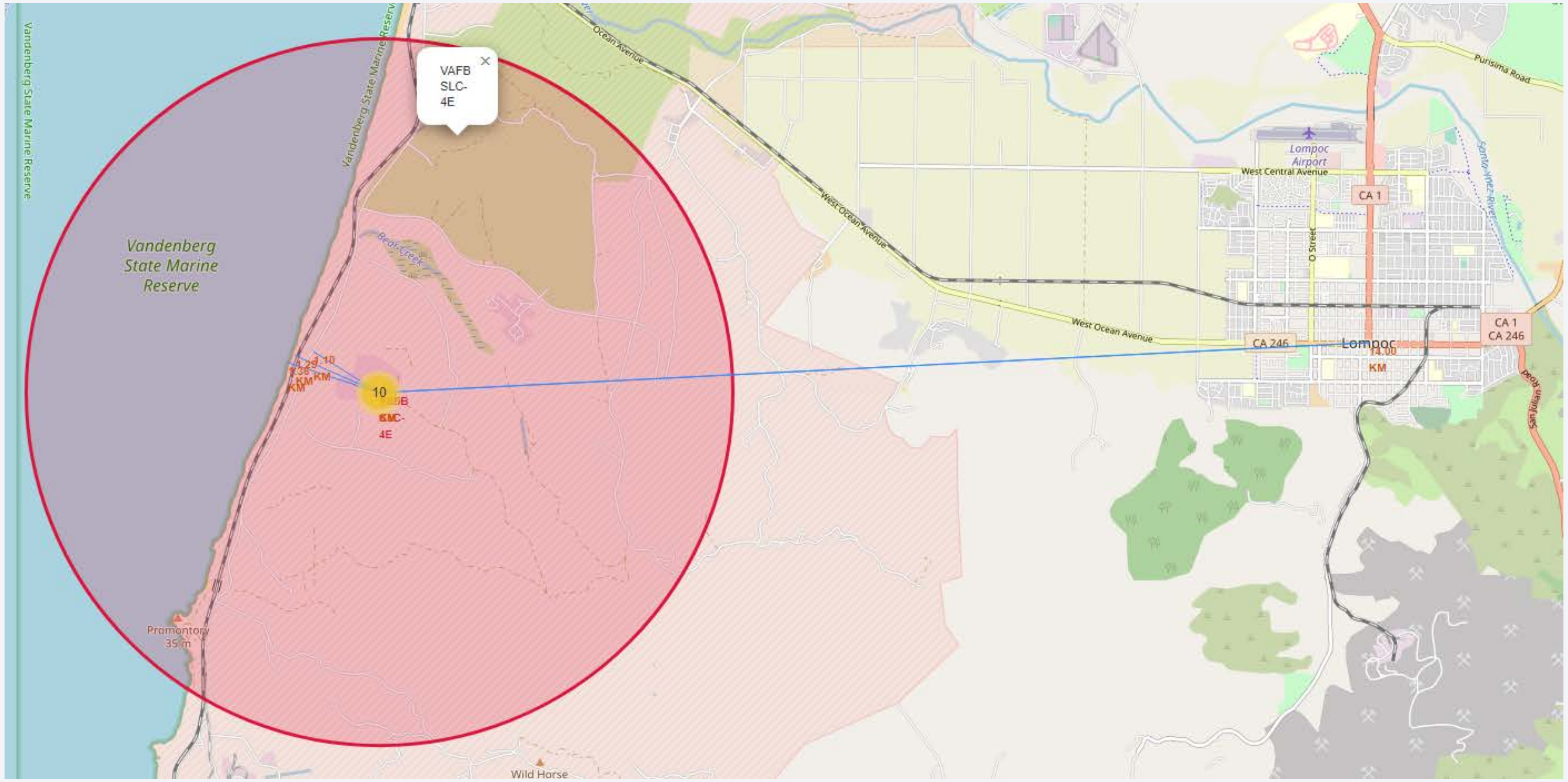




- The provided close-up views of the launch sites show their general areas (marked by the red circles), names as well as mission record with separate labels for outcome class

Launch sites and infrastructure

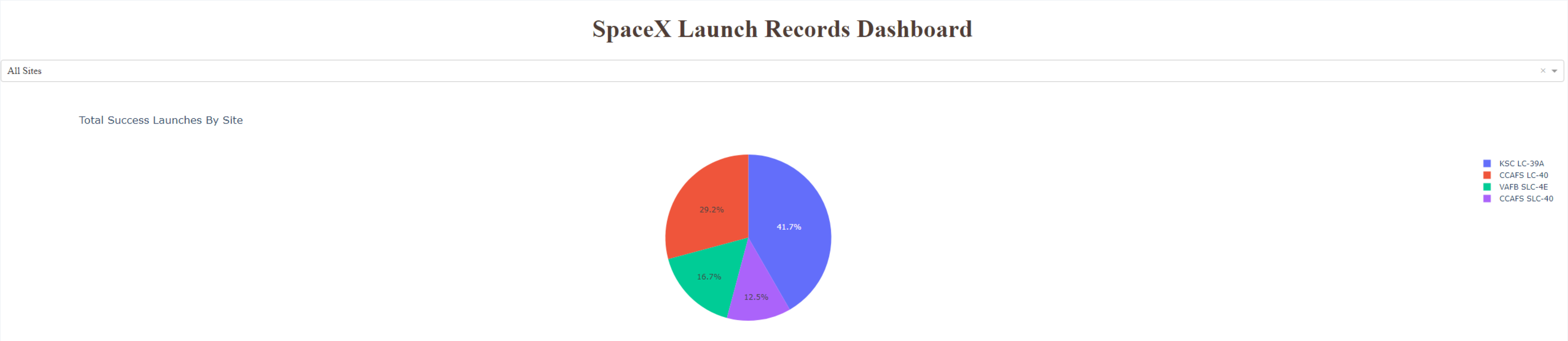
- This close up displays launch site VAFB SLC-4E, along with its respective proximities to coastline, city, railway and highway (blue lines and distance markers measured in kilometers)



Section 4

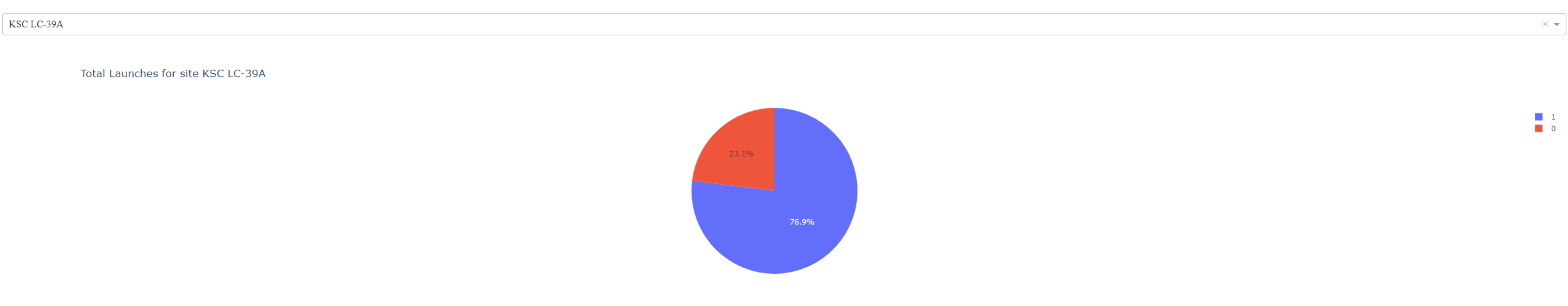
Build a Dashboard with Plotly Dash

Successful launch ratio for all sites



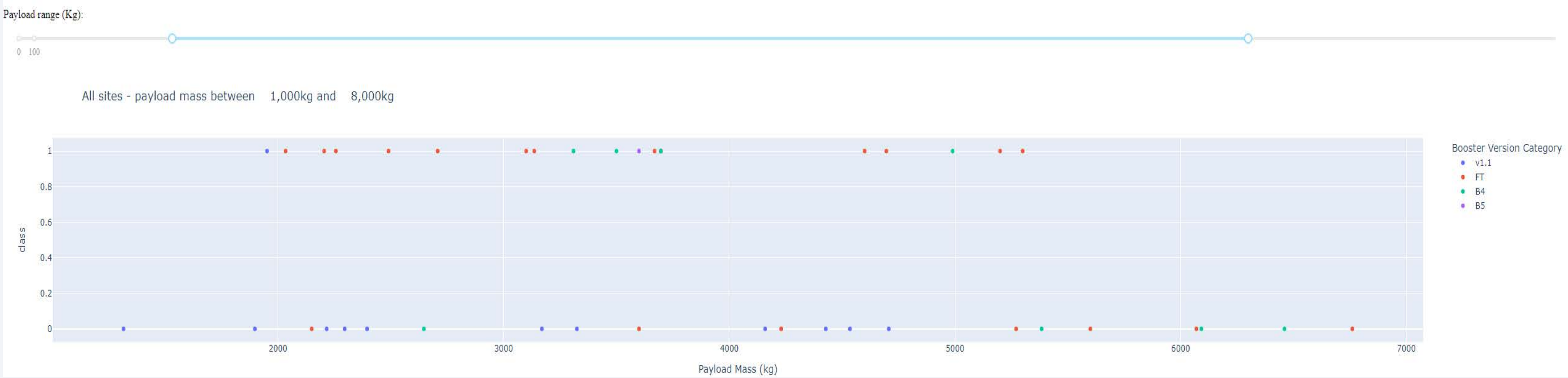
- The interactive dashboard displays the success ratio of each launch site in the form of a pie chart
  - KSC LC-39A has the highest count of successful missions, followed by CCAFS LC-40

Success ratio for KSC LC-39A



- Drilling down into the record for KSC LC-39A, we can observe that 76.9% of missions launched from that site resulted in a successful outcome

Payload/Booster Version relationship



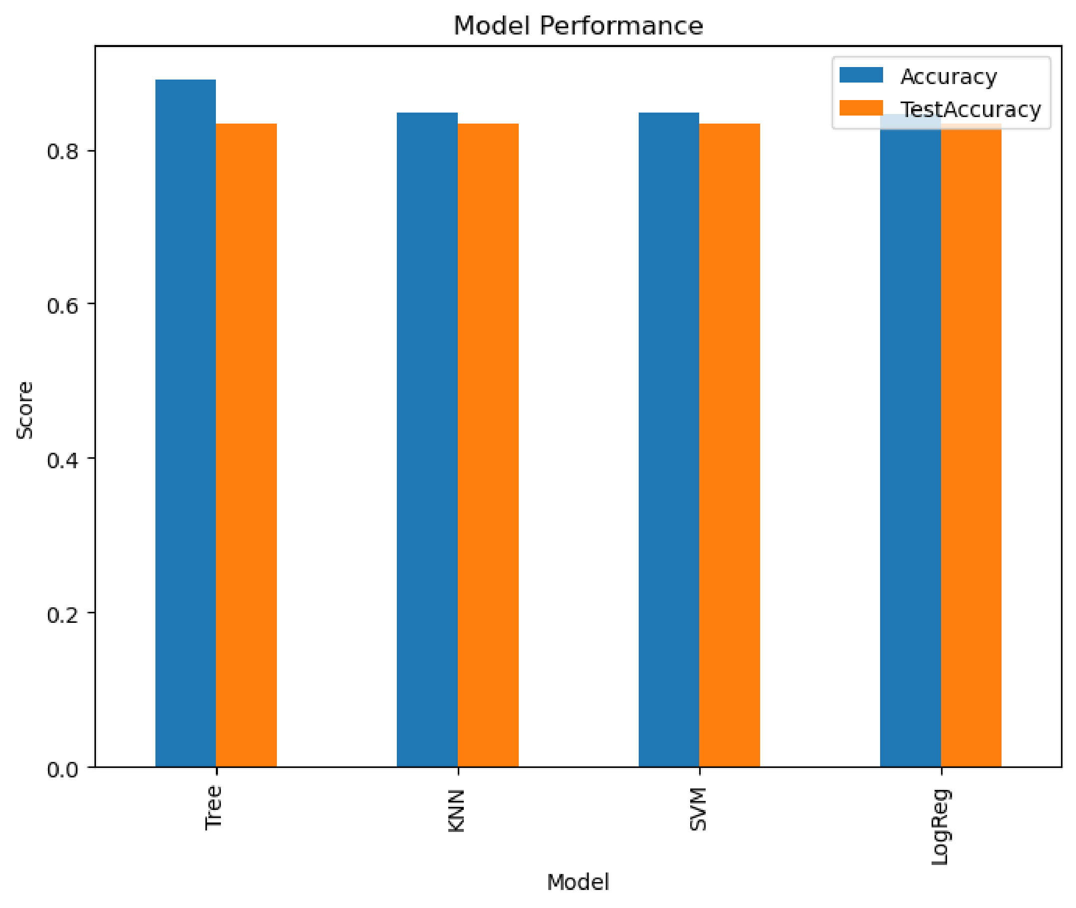
- Exploring the payload mass to booster version relationship, we can notice that in the 1000 – 8000kg range booster version FT seems to be the most dependable, while other versions yield mixed results, with v1.1 having the most failure outcomes

Section 5

Predictive Analysis (Classification)

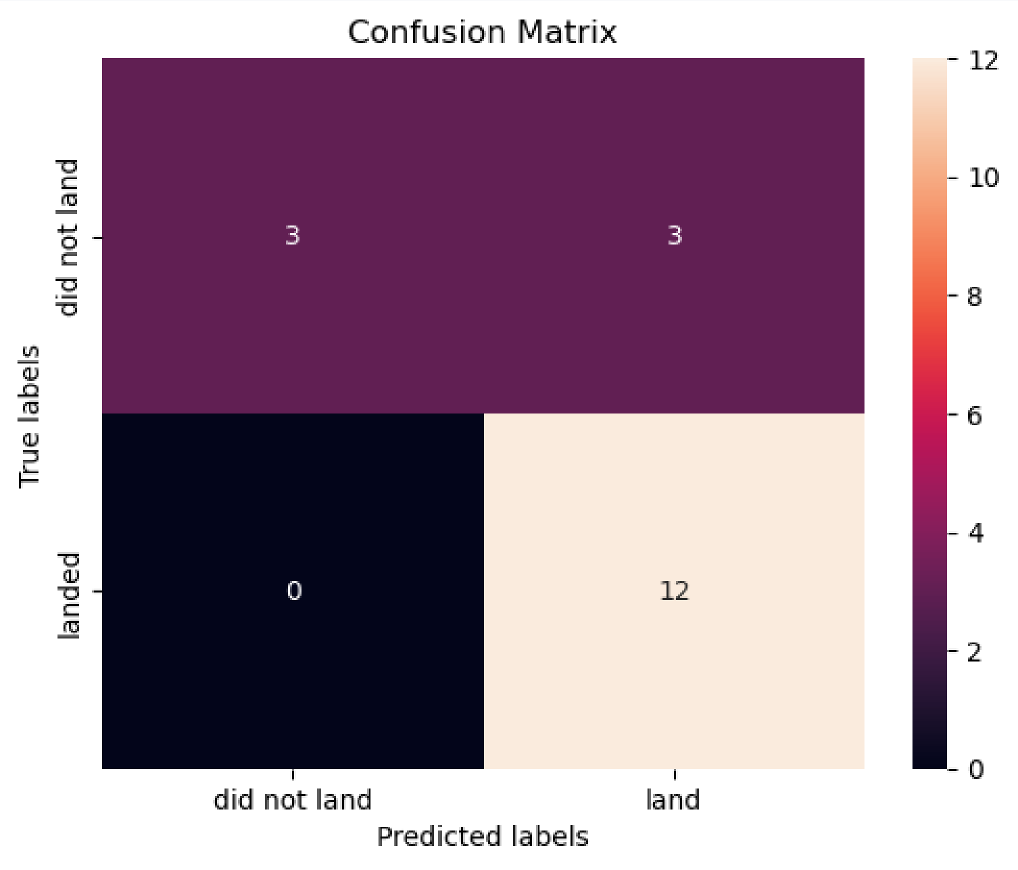
Classification Accuracy

- As previously shown, a variety of predictive models were built for the project
  - While all four models achieved relatively high accuracy scores, Decision Tree is the leader with accuracy of 88.9%



Confusion Matrix

- Analyzing the confusion matrix for the Decision Tree classifier, we can observe correct predictions for class 0 (did not land)
- When predicting class 1 (landed), the model returned a large number of false positives
- All four models have similar results, which has to do with the very limited scope of the dataset



Conclusions

- This project highlights many of the intricacies of setting up a successful commercial space transportation operation:
  - The 67% mission success rate of SpaceX hints at many of the risk and high costs involved
  - We can be confident that as time goes by and the flight number increases, so will the success rate depending on multiple factors such as launch site, booster type, orbit and payload mass
  - The complex relationships between all these parameters are detrimental to the success of a mission
  - A competitor company would need years and a large financial commitment in order to achieve a successful and reliable operation
  - When selecting launch sites, it is crucial that they are in proximity to critical infrastructure such as highways and railways, while also being far enough from large cities and preferably near a coastline for safety reasons
  - Relying on predictive models at this stage would not be recommended as there is not sufficient data available to return dependable positive predictions

Appendix

- https://github.com/chrischukov/ds_capstone

Thank you!