Winning Space Race with Data Science

MAYUSHII

12.06.23

IBM Developer
SKILLS NETWORK

Outline

Executive Summary

- Summary of methodologies
  - Data Collection (API & Web Scraping)
  - Data Wrangling
  - Exploratory Data Analysis (SQL & Data Visualization)
  - Interactive Locations Analysis with Folium
  - Dashboard with Plotly Dash
  - ML Predictive Analysis (Classification)
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

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 the first stage lands successfully, the cost of a launch can be determined. This information is beneficial to SpaceX in bid situations with other companies competing against SpaceX for market share in rocket launches.

SECTION 1

Methodology

Methodology

- Data collection was performed using SpaceX API and web scraping from Wikipedia. Data was filtered, cleaned, and one-hot encoded for binary classification.
- EDA was conducted using SQL and visualization. Interactive visual analytics were performed with Folium and Plotly Dash to visualize geospatial launch sites and landing outcomes versus payload mass per booster version.
- Predictive analysis involved fitting data with Logistic Regression, calculating accuracy, and creating confusion matrices to find the best model.

Data Collection

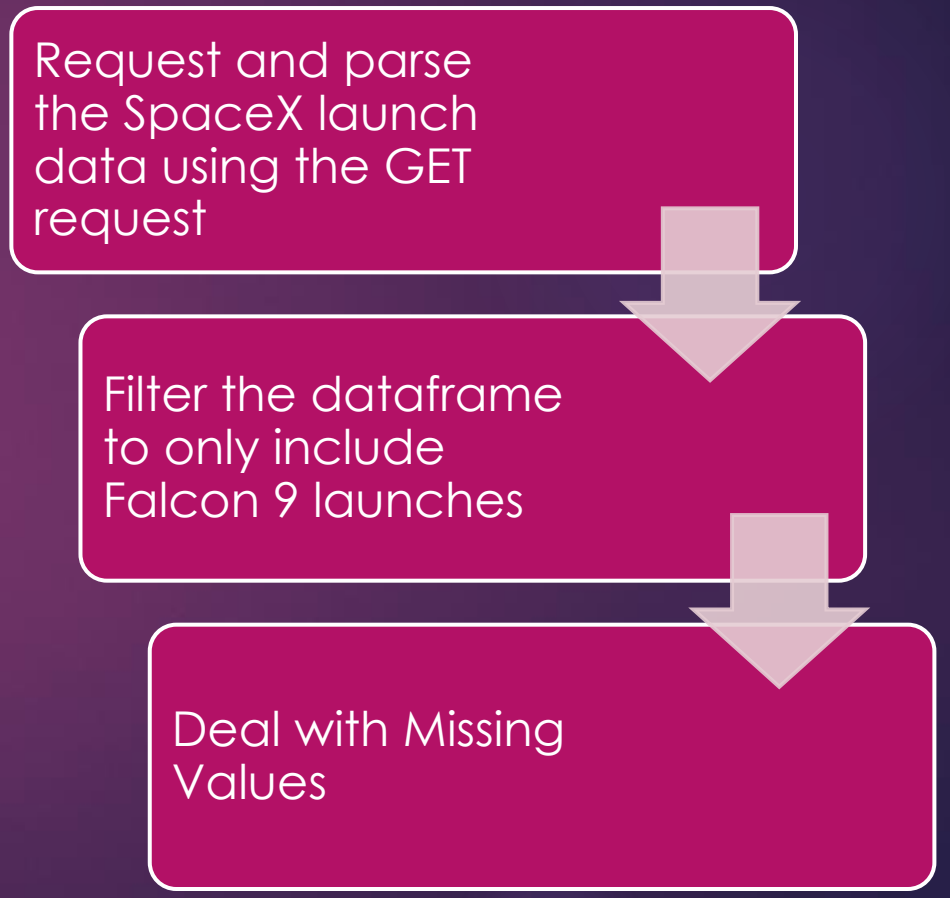
The purpose of the data collection process was to gather information on SpaceX Falcon 9 rockets and launches. Data was collected through a combination of API requests from SpaceX REST API and web scraping a table from Wikipedia's SpaceX page.

The SpaceX API provided data columns such as flight number, date, booster version, payload mass, orbit, launch site, outcome, longitude and latitude, etc.

Web scraping from SpaceX's Wikipedia page provided columns such as flight number, launch site, payload, payload mass, orbit, customer, launch outcome, booster version, booster landing, date, and time.

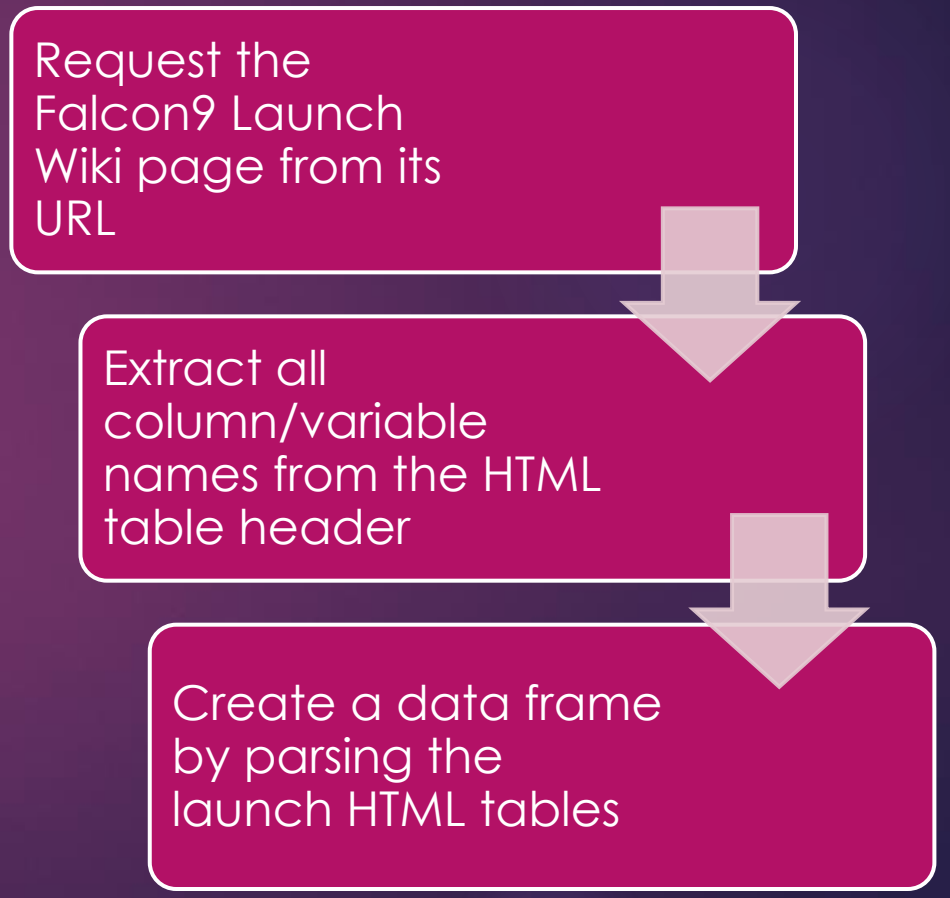
Data Collection - SpaceX API

- A get request to the SpaceX API was used to collect rocket data. This was cleaned, and some basic data wrangling and formatting was performed.
- GitHub



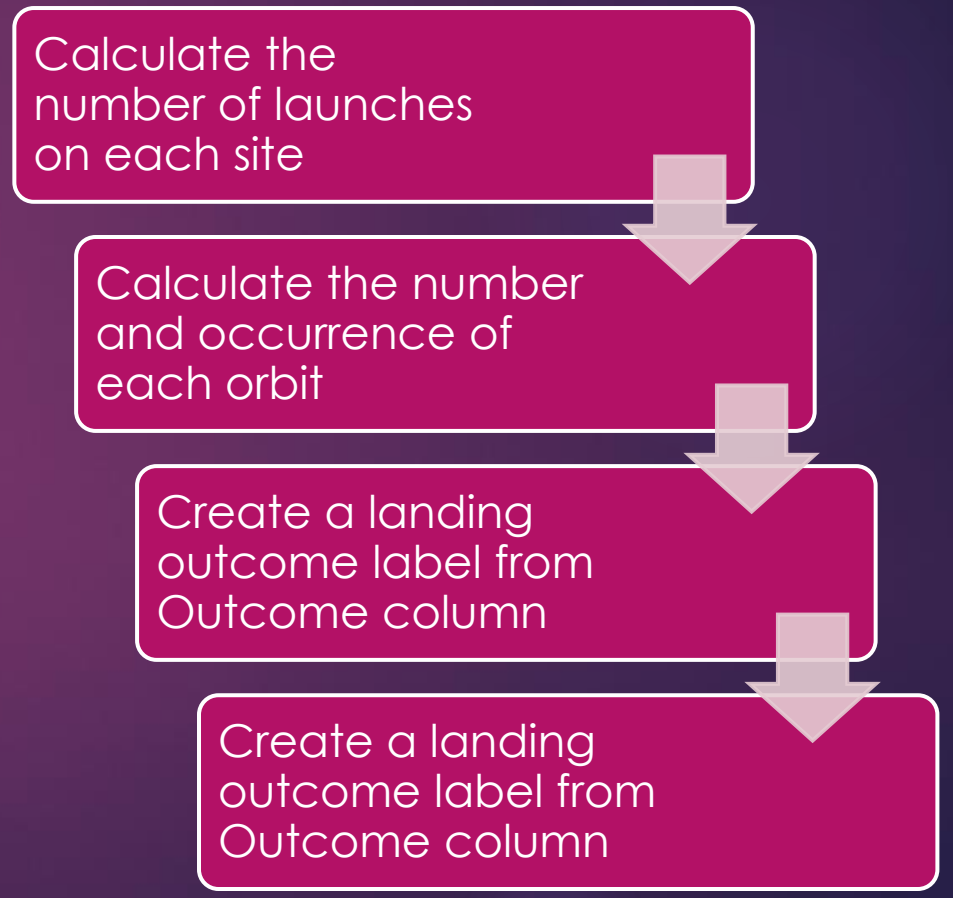
Data Collection - Scraping

- To gather data on Falcon 9 rocket launches from Wikipedia, we employed web scraping techniques using BeautifulSoup. The resulting data was then parsed and converted into a Pandas dataframe.
- GitHub



Data Wrangling

- In order to prepare for training, the data was explored. The number of launches at each site were calculated, as well as the number and occurrence of each orbit. Additionally, we created a landing outcome label based on the outcome column.
- GitHub



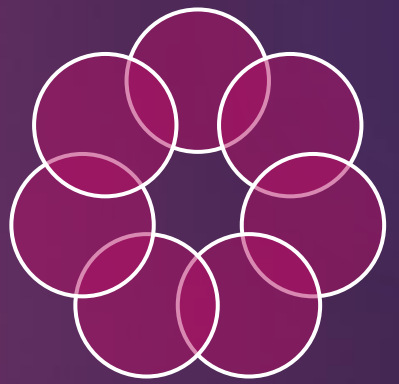
EDA with Data Visualization

- The following charts were plotted to help identify patterns and relationships between variables that can be used for feature engineering to improve model predictions.
- GitHub

a scatter plot of FlightNumber vs. PayloadMass

a line chart of the launch success trend over the years

a scatter plot of FlightNumber vs. LaunchSite



a scatter plot of PayloadMass vs. Orbit type

a scatter plot of PayloadMass vs. LaunchSite

a scatter plot of FlightNumber vs. Orbit type

a bar chart of success rate for each orbit type

EDA with SQL

10 SQL queries were performed:

- The first query selects distinct launch site names from the "spacextbl" table.
- The second query selects records where the launch site name starts with "CCA" and limits the result to 5 records.
- The third query calculates the total payload mass of boosters launched by NASA (CRS).
- The fourth query calculates the average payload mass carried by booster version F9 v1.1.
- The fifth query finds the date when the first successful landing outcome in ground pad was achieved.
- The sixth query lists the names of boosters that had success in drone ship and carried payload mass between 4000 and 6000.
- The seventh query counts the number of successful and failure mission outcomes.
- The eighth query lists the names of booster versions that carried the maximum payload mass using a subquery.
- The ninth query lists the records for month names, failure landing outcomes in drone ship, booster versions, and launch sites in the year 2015.
- The tenth query ranks the count of successful landing outcomes between the dates 04-06-2010 and 20-03-2017 in descending order.

- GitHub

Build an Interactive Map with Folium

We perform a visual analysis of launch sites using the Folium package:

- Markers are added to the map using circles, popup labels, and text labels to show the geographical locations of all launch sites, including NASA Johnson Space Center. The colors of markers are used to indicate the success or failure of launches at each site, with green markers indicating success and red markers indicating failure. Marker clusters are used to identify which launch sites have high success rates.
- Distances between a launch site and its proximities, such as railway, highway, coastline, and closest city, are calculated and marked on the map using colored lines and text labels. This helps illustrate how far the launch sites are from heavily populated areas.
- The visual analysis also includes a mouse position feature to help users find the coordinates of points of interest, such as railway or coastline

- GitHub

Build a Dashboard with Plotly Dash

- The dashboard includes a dropdown list to select a Launch Site, a pie chart to show the total number of successful launches for all sites or per site if selected, a range slider to select the payload range, and a scatter plot to show the correlation between payload mass and launch success.
- The dropdown list is added to enable Launch Site selection, allowing the user to filter data based on specific launch sites. The pie chart shows the total number of successful launches; if a launch site is selected, it displays the count of Success vs. Failed launches for that specific site. This chart provides insights into which launch site has a higher success rate.
- The range slider allows users to select the Payload range for analysis, and the scatter plot shows the correlation between payload mass and launch success. The scatter plot also displays the Booster Version Category as color, enabling the user to identify any patterns in the relationship between launch success, payload mass, and booster version category. This plot helps to understand whether there is a correlation between the payload mass and launch success rate, and how the booster version affects the success rate.
- GitHub

Predictive Analysis (Classification)

Data Preparation:

- Created a NumPy array Y from the column 'Class' in the dataset.
- Standardized the feature data X using the StandardScaler from scikit-learn.
- Split the data into training and test sets using the train_test_split function, with a test size of 0.2 and random state of 2.

Logistic Regression, Support Vector Machine (SVM), Decision Tree, K-Nearest Neighbors:

- Created a logistic regression, a SVM, a decision tree classifier and a KNN object and defined appropriate grids of hyperparameters to search through using GridSearchCV.
- Fit the models using the training data to find the best hyperparameters.
- Retrieved the best parameters and accuracy score on the validation data.

Model Evaluation:

- Calculated the accuracy of each model on the test data using the score method.
- Plotted the confusion matrix to visualize the performance of each model on the test data.

Model Comparison and Selection:

- Compared the accuracy scores of all the models.
- Noted that the decision tree model had a lower accuracy and exhibited high bias and variance.
- Stated that all other models performed similarly, except for the decision tree model.

▶ GitHub

Results

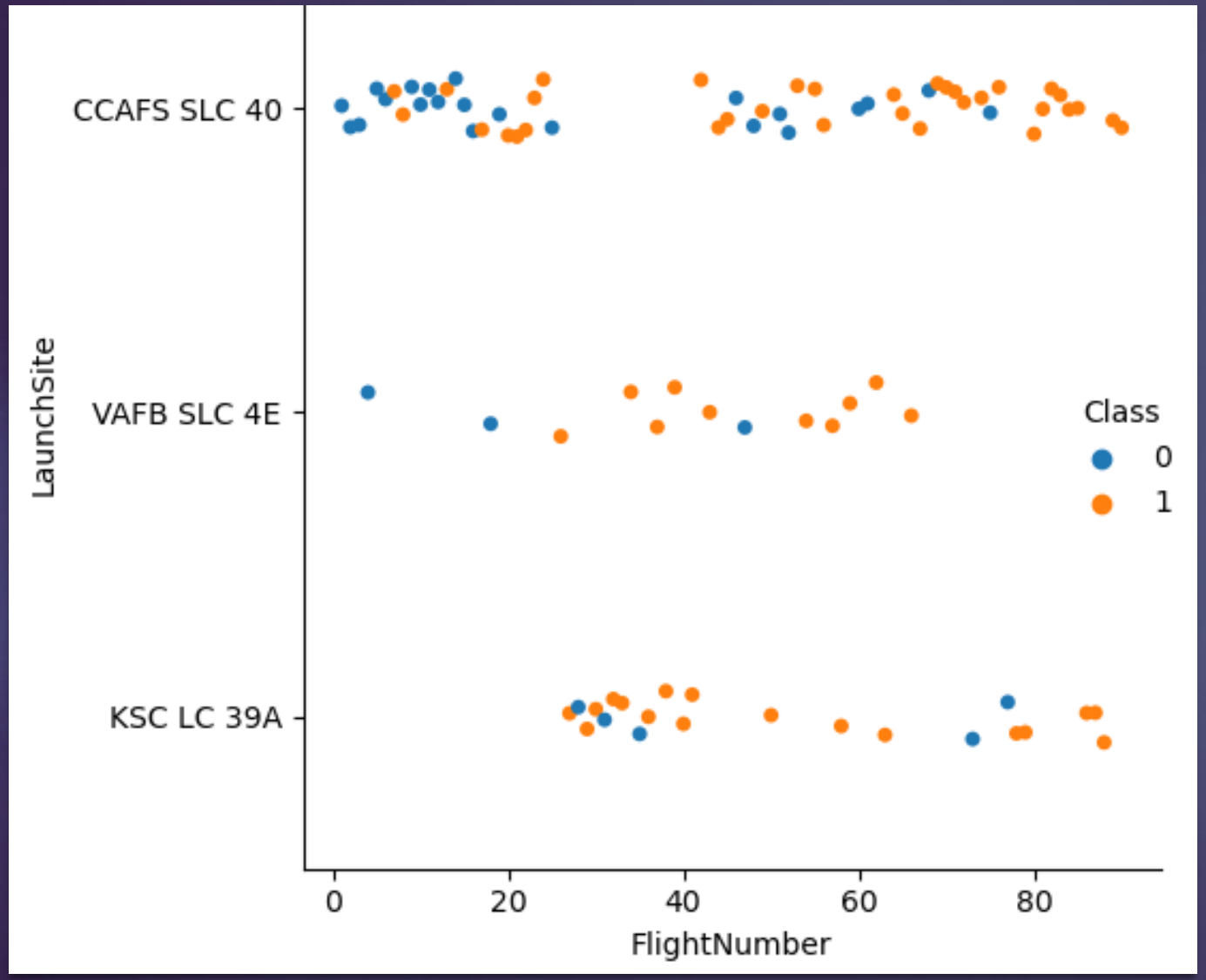
Exploratory data analysis results

Interactive analytics demo in screenshots

Predictive analysis results

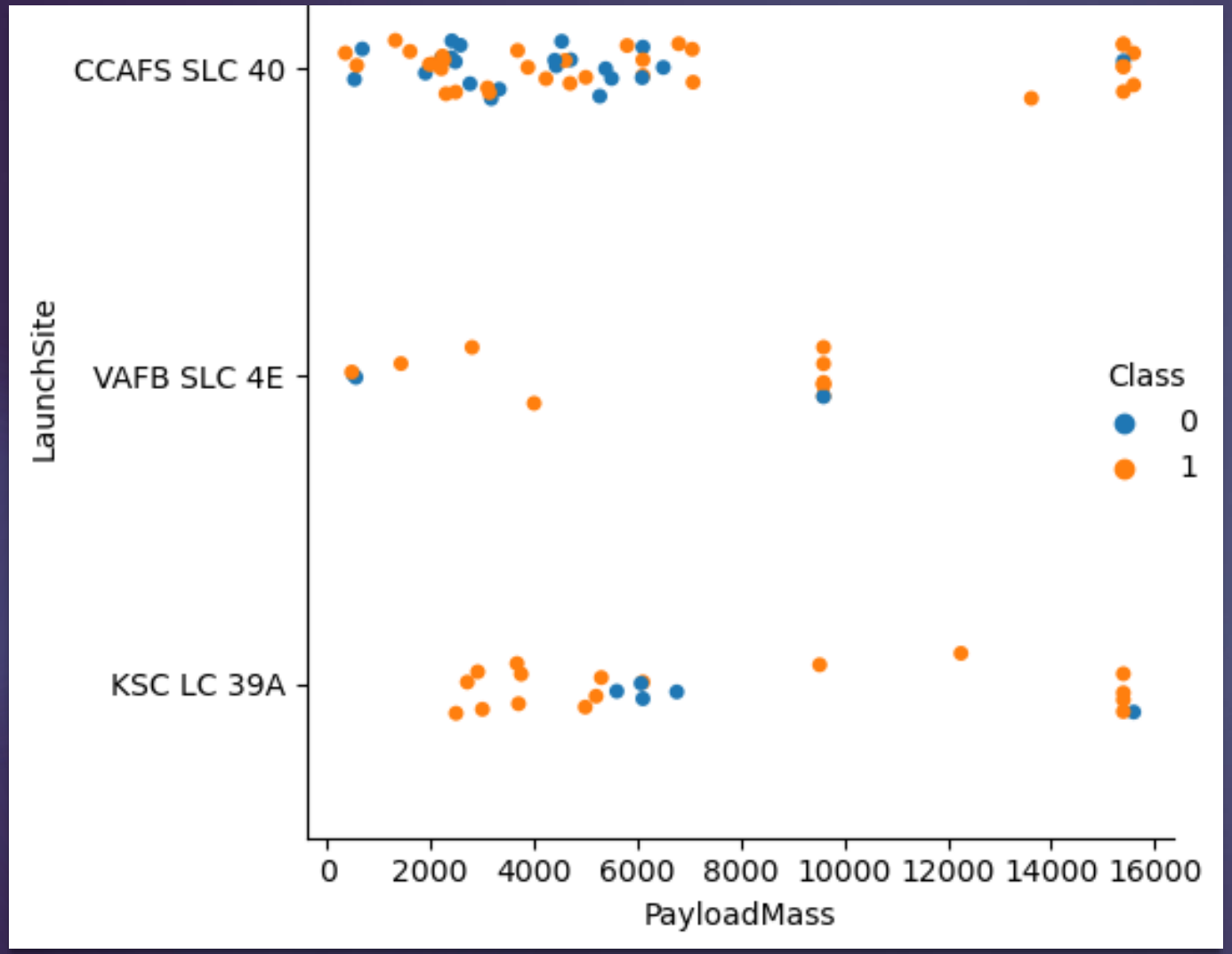
SECTION 2

Insights drawn from EDA



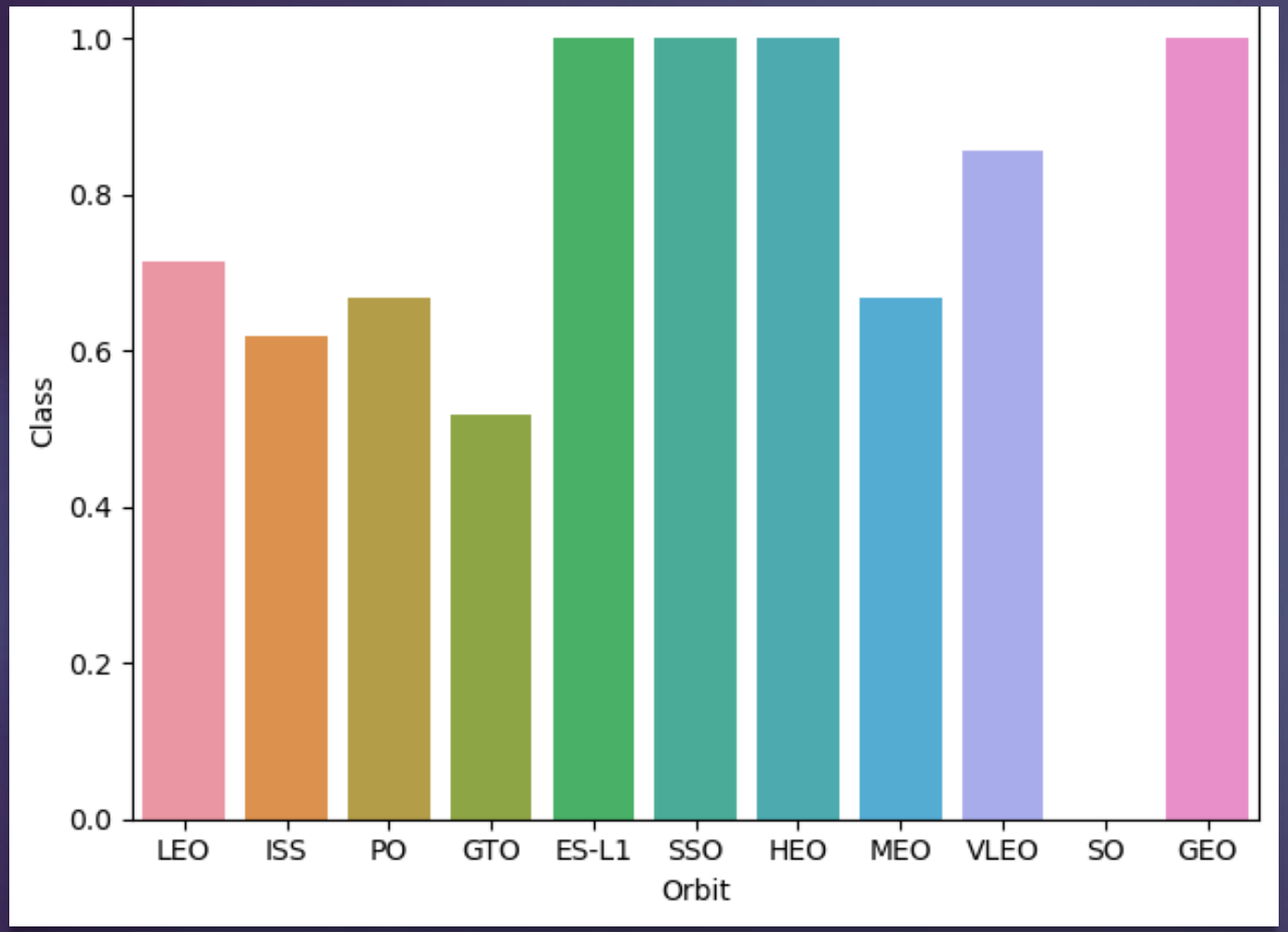
Flight Number vs. Launch Site

- It can be observed, that the success rate for all launch sites increases as the number of flights increases. This can explain why CCAFS SLC has a success rate slightly lower than the other launch sites, because it is the oldest and most of the earliest flights were launched from it. Both VAFB SLC and KSC LC have excellent success rates. Launches from VAFB SLC seem to have stopped at around 70 flights, which may require further investigation, as the sight may have been closed.



Payload vs. Launch Site

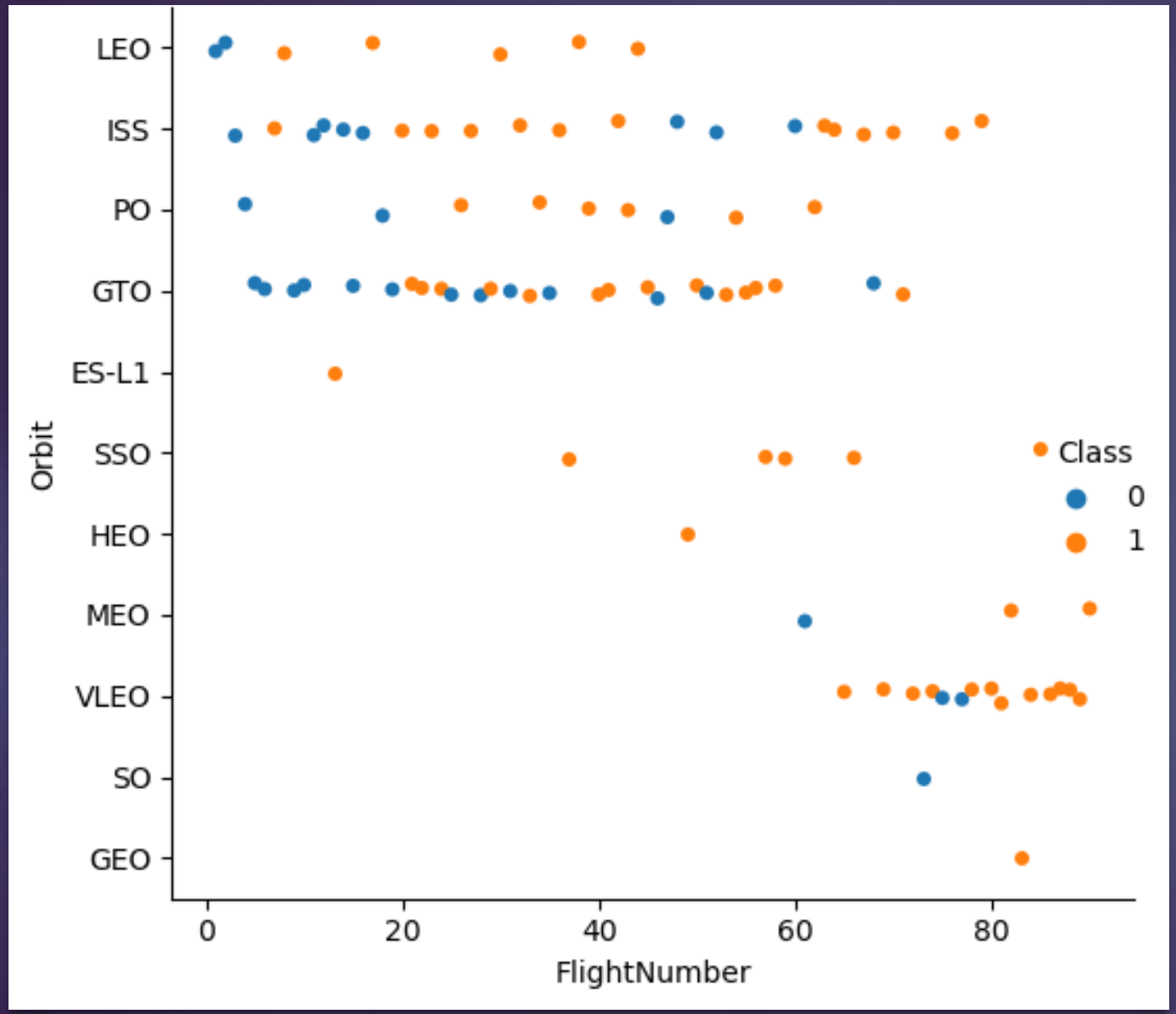
- For the VAFB-SLC launch site there are no rockets launched for heavy payloads (with a mass greater than 10k kg). CCAFS SLC seems to have better success rates with higher payloads, however that may also depend on other factors. No launches with a payload over 16k have been performed.



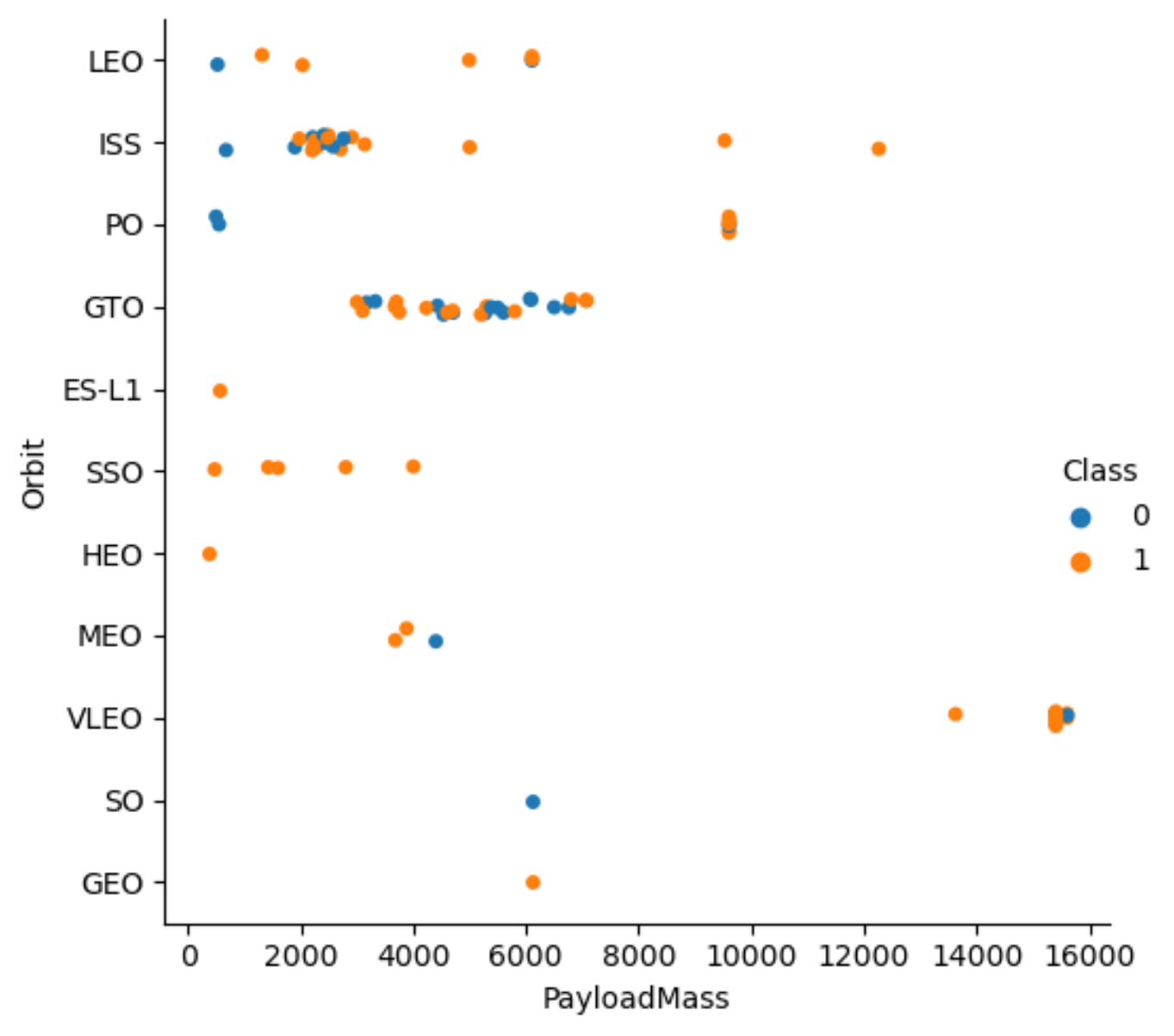
Success Rate vs. Orbit Type

- The GEO, HEO, SSO and ES-L1 orbits have the highest success rate. The SO orbit has no successful outcomes and GTO, ISS and MEO are among the orbits with the lowest success rates.

Flight Number vs. Orbit Type

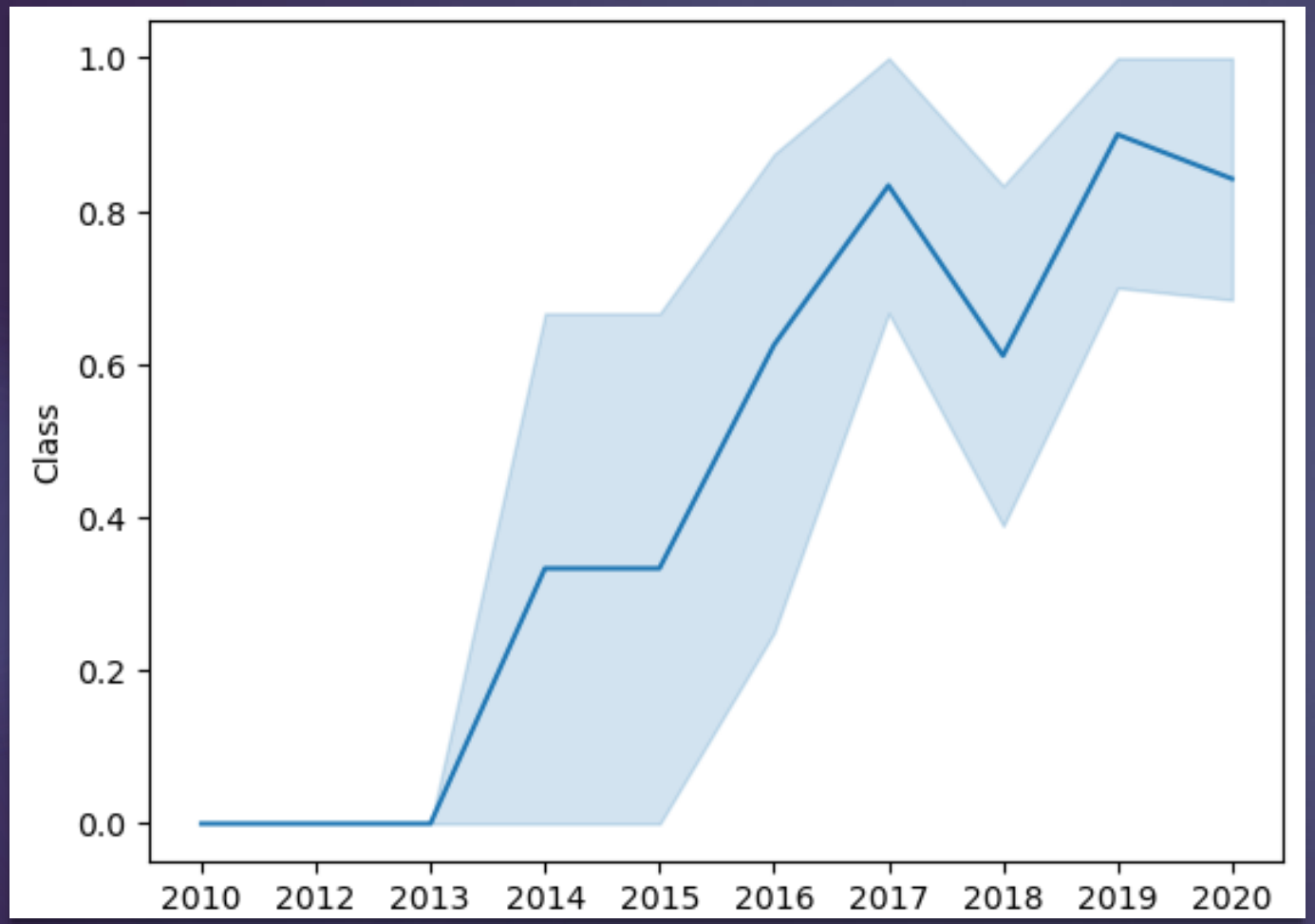


- In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



Payload vs. Orbit Type

- LEO, ISS and VLO orbits are associated with successful landing outcomes and heavy payloads. Launches with the heaviest payloads utilized the ISS and VLEO orbits. There appears to be no relationship between payload mass and the GTO orbit relative to landing success.



Launch Success Yearly Trend

- It is clear, that the success rate has been rising quite steadily (despite some failures in 2018) since 2013.

All Launch Site Names

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

```
%%sql
SELECT DISTINCT "Launch_Site" FROM spacextbl
```

✓ 0.0s

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

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

Launch Site Names Begin with 'CCA'

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

```
%%sql
SELECT * FROM spacextbl WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5
```

✓ 0.0s

* sqlite:///my_data1.db
Done.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

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

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) total_payload FROM spacextbl WHERE "Customer" = "NASA (CRS)"
```

✓ 0.0s

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

total_payload
45596.0

Average Payload Mass by F9 v1.1

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

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) ave_payload FROM spacextbl WHERE "Booster_Version" = "F9 v1.1"
```

✓ 0.0s

* sqlite:///my_data1.db

Done.

ave_payload
2928.4

First Successful Ground Landing Date

List the date when the first succesful landing outcome in ground pad was acheived.

```
%%sql
SELECT MIN("Date") date FROM spacextbl WHERE "Landing_Outcome" = "Success (ground pad)"
```

✓ 0.0s

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

date
01/08/2018

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%%sql
SELECT "Booster_Version" FROM spacextbl
WHERE "Landing_Outcome" = "Success (drone ship)" and PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
```

✓ 0.0s Python Python

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

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

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
%%sql
SELECT CASE WHEN "Mission_Outcome" LIKE "Success%" THEN "Success" ELSE "Failure" END "Outcome",
    COUNT("Mission_Outcome") total FROM spacextbl
WHERE "Mission_Outcome" LIKE "Success%" or "Mission_Outcome" LIKE "Failure%"
GROUP BY "Mission_Outcome" LIKE "Success%"
ORDER BY total DESC
```

✓ 0.0s

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

Outcome	total
Success	100
Failure	1

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass.

```
%%sql
SELECT DISTINCT "Booster_Version" FROM spacextbl
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM spacextbl)
```

✓ 0.0s

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

Booster_Version
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

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

```
%%sql
SELECT substr("Date", 4, 2) Month, "Landing_Outcome", "Booster_Version", "Launch_Site"
FROM spacextbl
WHERE substr("Date", 7, 4) = "2015" AND "Landing_Outcome" = "Failure (drone ship)"
```

✓ 0.0s

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

Month	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

```
%%sql
SELECT "Landing_Outcome", COUNT(*) count FROM spacextbl
WHERE "Landing_Outcome" like "Success%"
    AND substr("Date",7)||substr("Date",4,2)||substr("Date",1,2) BETWEEN "20100604" and "20170320"
GROUP BY "Landing_Outcome"
ORDER BY count DESC
```

✓ 0.0s Python

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

Landing_Outcome	count
Success (ground pad)	5
Success (drone ship)	5

SECTION 3

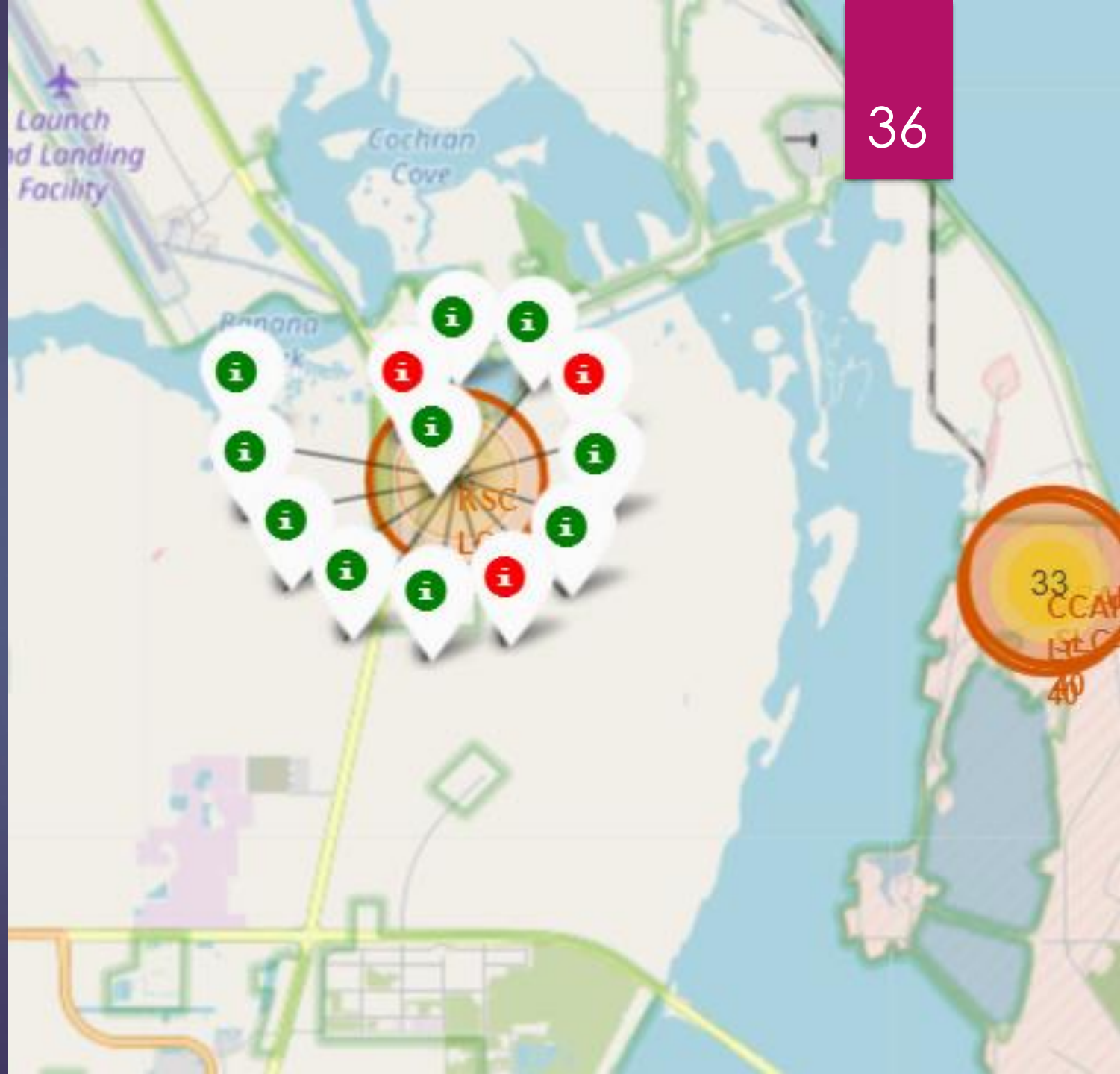
Launch Sites Proximities Analysis

All launch sites on a global map

- All launch sites are situated in the USA in relative proximity to the Equator line and very close to the coasts.

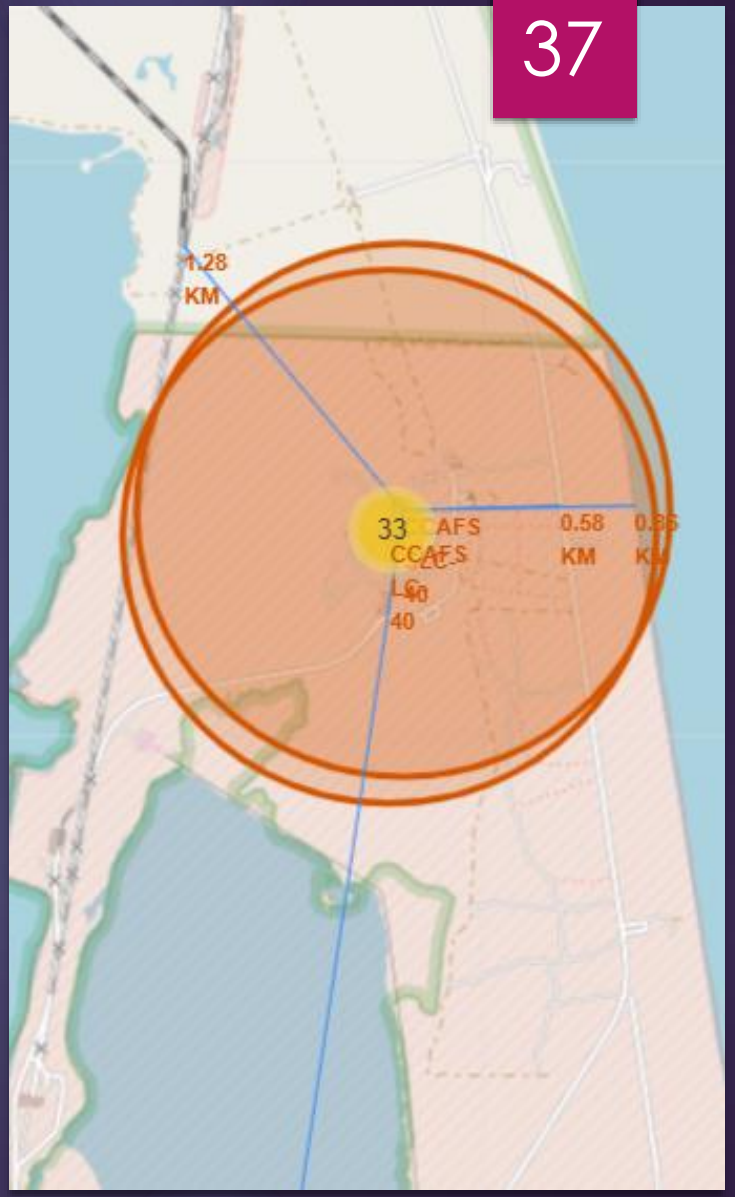
Color-labeled launch outcomes

- From the color-labeled markers in marker clusters, you should be able to easily identify which launch sites have relatively high, and which have low success rates.
- the Green markers indicate Successful launches
- the Red Markers indicate Failed launches
- The following image is of KSC LC-39A, which has a high success rate.



Distance from CCAFS SLC-40 to its proximities

- From the visual analysis of the launch site CCAFS SLC-40 we can clearly see that it is:
  - relatively close to a highway (0.58 km)
  - relatively close to the coastline (0.86 km)
  - relatively close to a railroad (1.28 km)
  - relatively distant from Cape Canaveral (18.15 km)

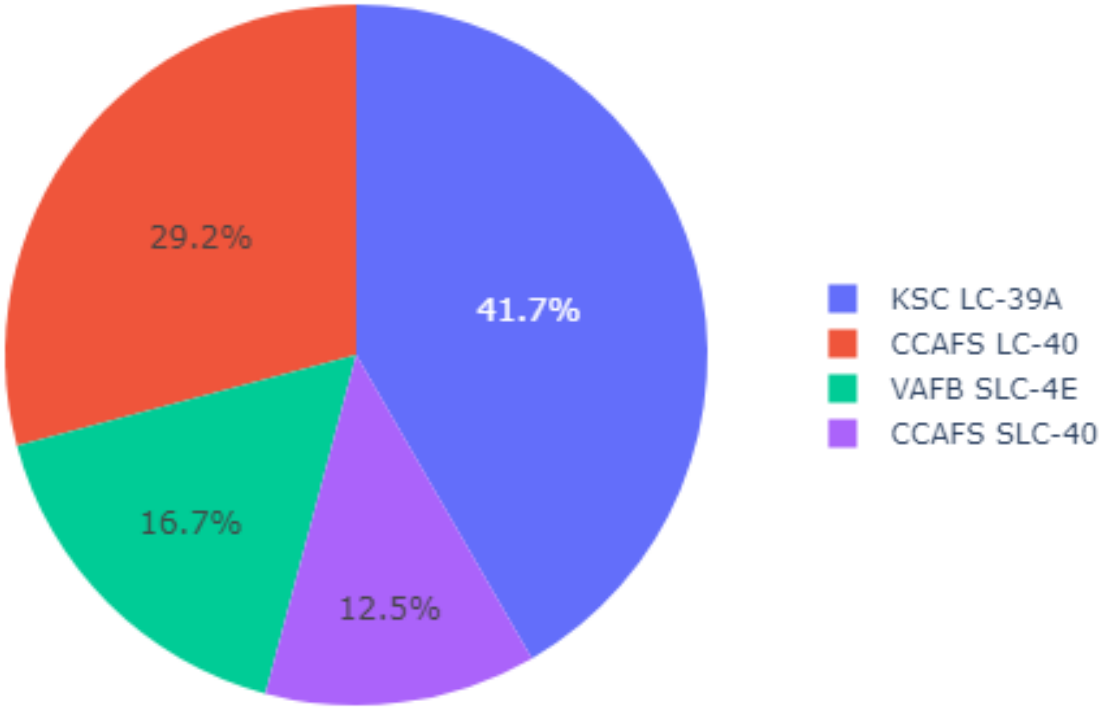


SECTION 4

Build a Dashboard with Plotly Dash

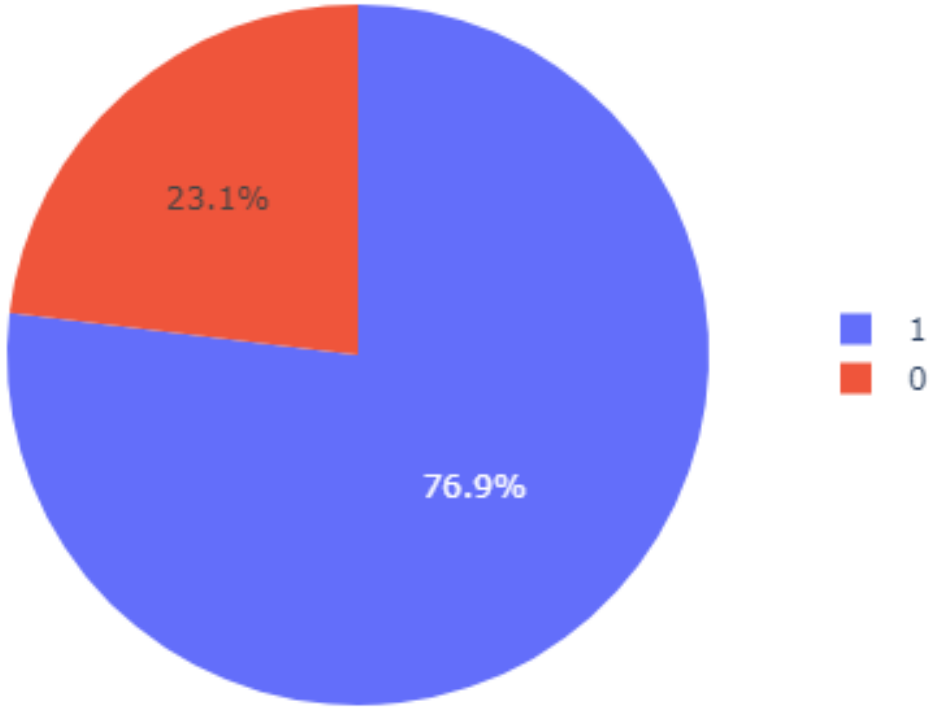
Launch success count for all sites

- From the following pie chart, it is clear, that the launch site with the most successes is KSC LC-39A.



Launch site with highest launch success ratio

- KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.



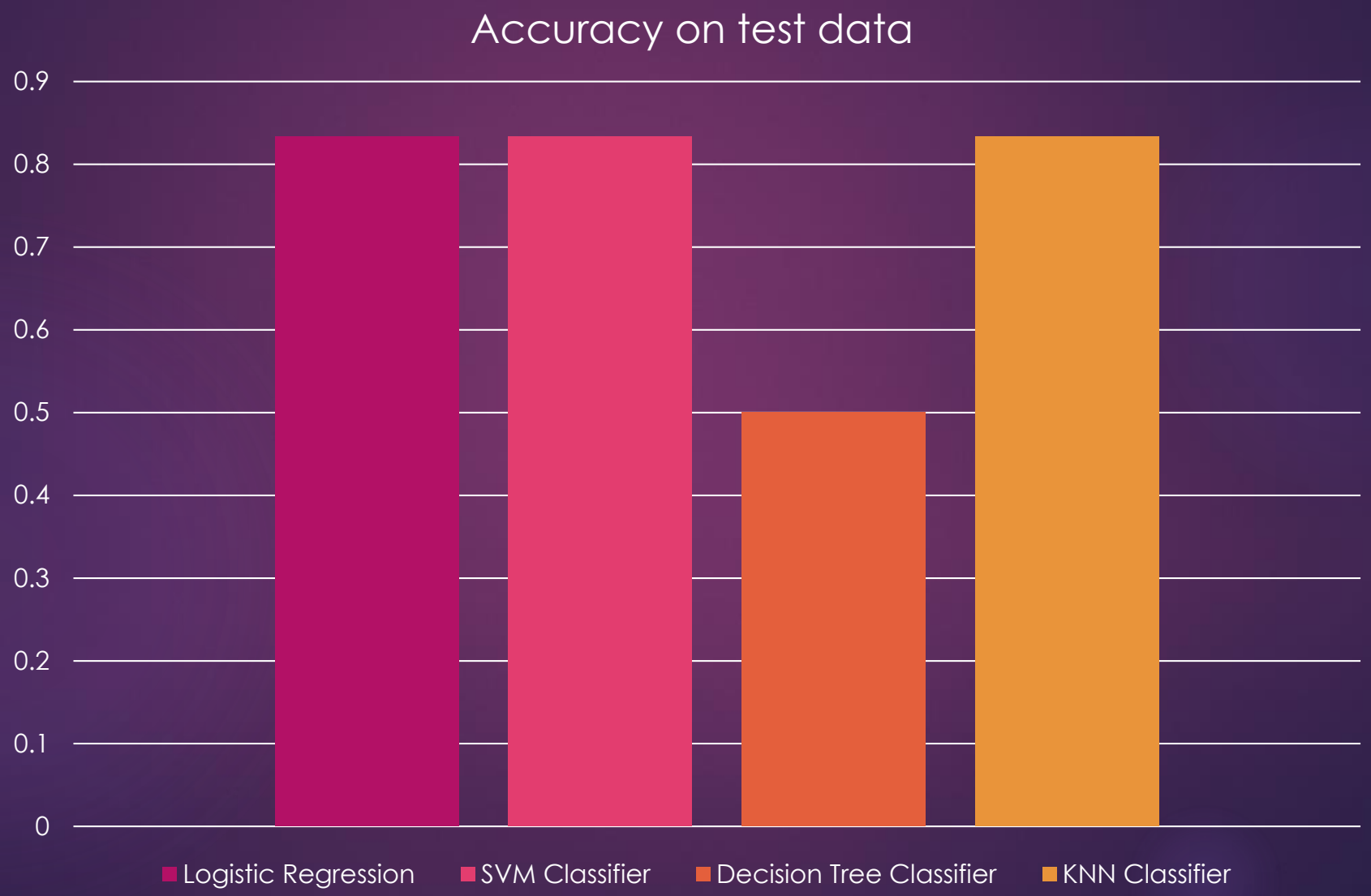
Payload vs. Launch Outcome for all sites

- the success rates for lighter payloads are higher compared to heavier payloads, with the FT and B4 booster versions being successful with payloads up to 2,000 kg. The FT booster version is the most successful for 2,000-4,600 kg payloads, while the B4 is the next most successful. As payload mass increases, there is a decrease in launches, and the B4 booster version is only version that that was successful with a very large payload.

SECTION 5

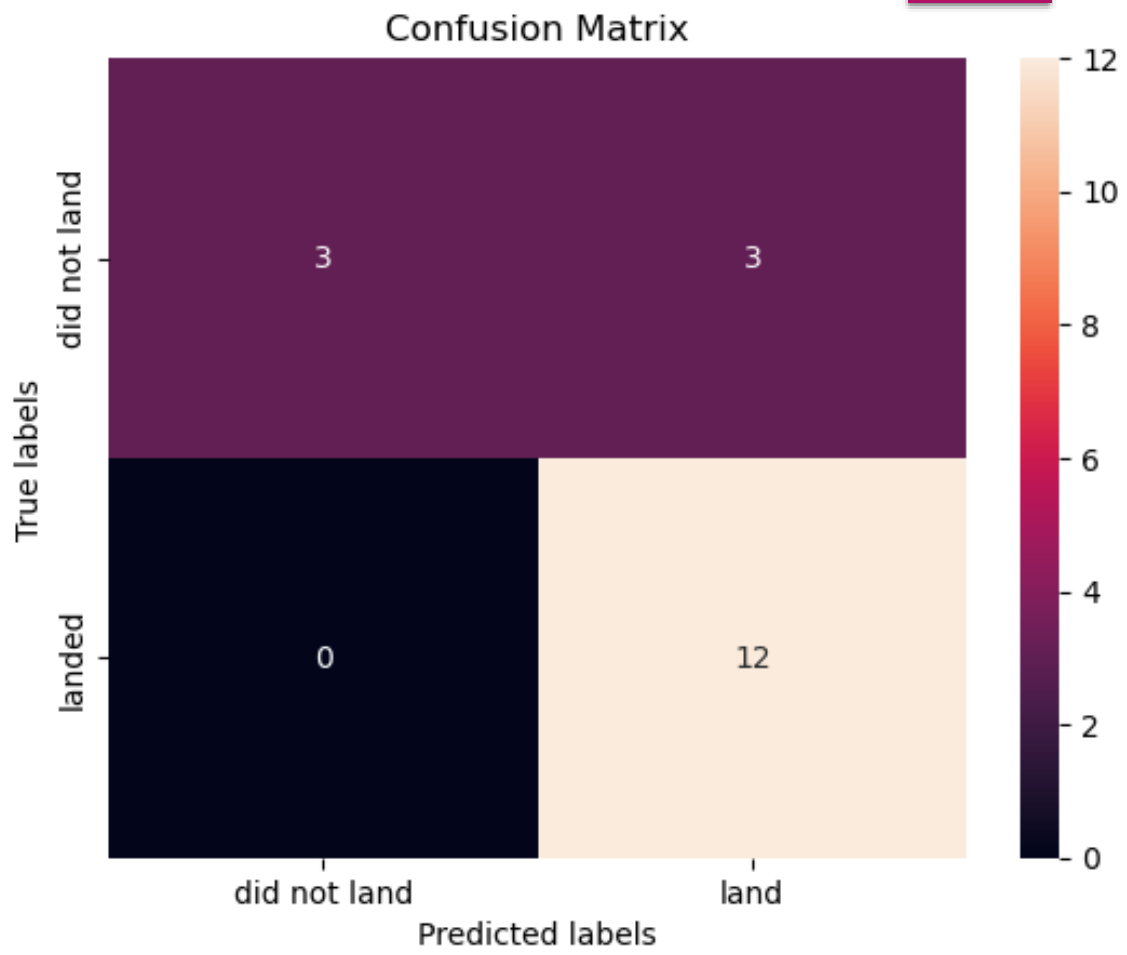
Predictive Analysis

Classification Accuracy



Confusion Matrix

- Due to the small sample size Logistic Regression, SVM and KNN all have the same accuracy, as well as the same confusion matrix. A larger dataset is needed for more accurate results.



Conclusions

- Generally, launches with a low payload mass tend to have a higher success rate than launches with a larger payload mass. However, there is a relationship between booster version, payload mass, and landing outcome.
- The success rate of launches has been increasing (despite some failures in 2018) since 2013.
- Launch site are located near the equator, on coastlines and away from heavily populated areas (likely for launch and safety reasons).
- Some launch sites have higher success rates than others, with KSC LC-39A having the most successful launches among all sites.
- The orbit used also affects the success of the first stage landing, with ES-L1, GEO, HEO, SSO, and VLEO having the highest success rates.
- The sample size is too small for any meaningful comparison of the tried predictive classification models. All models performed similarly, except DT, which resulted in high bias and variance, poorly fitting both the train and the test data.
- More data is needed for a more accurate and thorough analysis of the subject matter.

Thank you!