IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Christopher Brereton
11th May 2023

Executive Summary

This project aimed to predict the likelihood of success of a SpaceX mission based on various features, such as payload mass, orbit type, booster type, landing type, and more. The purpose was to guide a company's entry into the aerospace industry and enable them to offer competitive prices by assessing the risk of a SpaceX mission.

The data was collected from web scraping and SpaceX APIs, cleaned, and stored in SQL databases and JSON-based data frames. Exploratory analysis was done using SQL queries, Seaborn, Dash, and Folium to visualize relationships between variables. The predictive analysis was performed using decision trees, logistic regression, and support vector machines, and the best-performing model was identified.

The exploratory analysis provided insights into the relationships between various variables and mission outcomes. For example, missions launched from Cape Canaveral had a higher success rate than those launched from Vandenberg Air Force Base, and missions with a payload mass between 2500kg and 5000kg had the highest success rate.

The different models were found to be very similar in accuracy with the decision tree algorithm being the best-performing model for predicting the likelihood of success of a SpaceX mission, with an accuracy score of 83.3% on the testing data.

Table Of Contents

Introduction

The aerospace industry has witnessed significant growth in recent years, with several private companies entering the market to compete with traditional players. One of the most notable new entrants is SpaceX, founded by Elon Musk in 2002, which has disrupted the industry with its reusable rockets and ambitious plans for space exploration. SpaceX's success has inspired other companies to enter the industry, but they face the challenge of competing with a well-established player that has a proven track record of successful missions.

To address this challenge, a new company wishes to enter the aerospace industry and compete with SpaceX by launching satellites and other payloads into orbit. As a guide to the prices they will charge, they want to be able to predict the likelihood of success of a SpaceX mission, given characteristics such as payload mass, orbit type, booster type, landing type, time of year, and other features. The thinking is that if a customer's desired project seems to imply a risky mission for SpaceX, then this competitor can advertise this and justify a price accordingly.

To achieve this goal, the company has conducted a project to predict the likelihood of success of a SpaceX mission based on various features. The project utilized data collected from web scraping and SpaceX APIs, cleaned, and stored in SQL databases and JSON-based data frames. The data was then analyzed using exploratory analysis techniques such as SQL queries, Seaborn,

Dash, and Folium to visualize relationships between variables. Predictive analysis was performed using decision trees, logistic regression, and support vector machines, with the best-performing model identified.

The project's primary goal is to provide valuable insights into the factors that contribute to the success of a SpaceX mission and guide pricing strategies for the new company entering the industry. The specific problems that the project aims to answer include:

1. Can we predict the likelihood of success of a SpaceX mission based on various features such as payload mass, orbit type, booster type, landing type, and time of year?
2. What are the most important features for predicting the success of a SpaceX mission?
3. Can we provide insights into the relationships between various variables and mission outcomes?
4. How can the predictive model be used to guide pricing strategies for the new company entering the industry?

By answering these questions, the project aims to provide valuable information for the new company entering the aerospace industry, allowing them to make informed decisions about pricing and risk assessment.

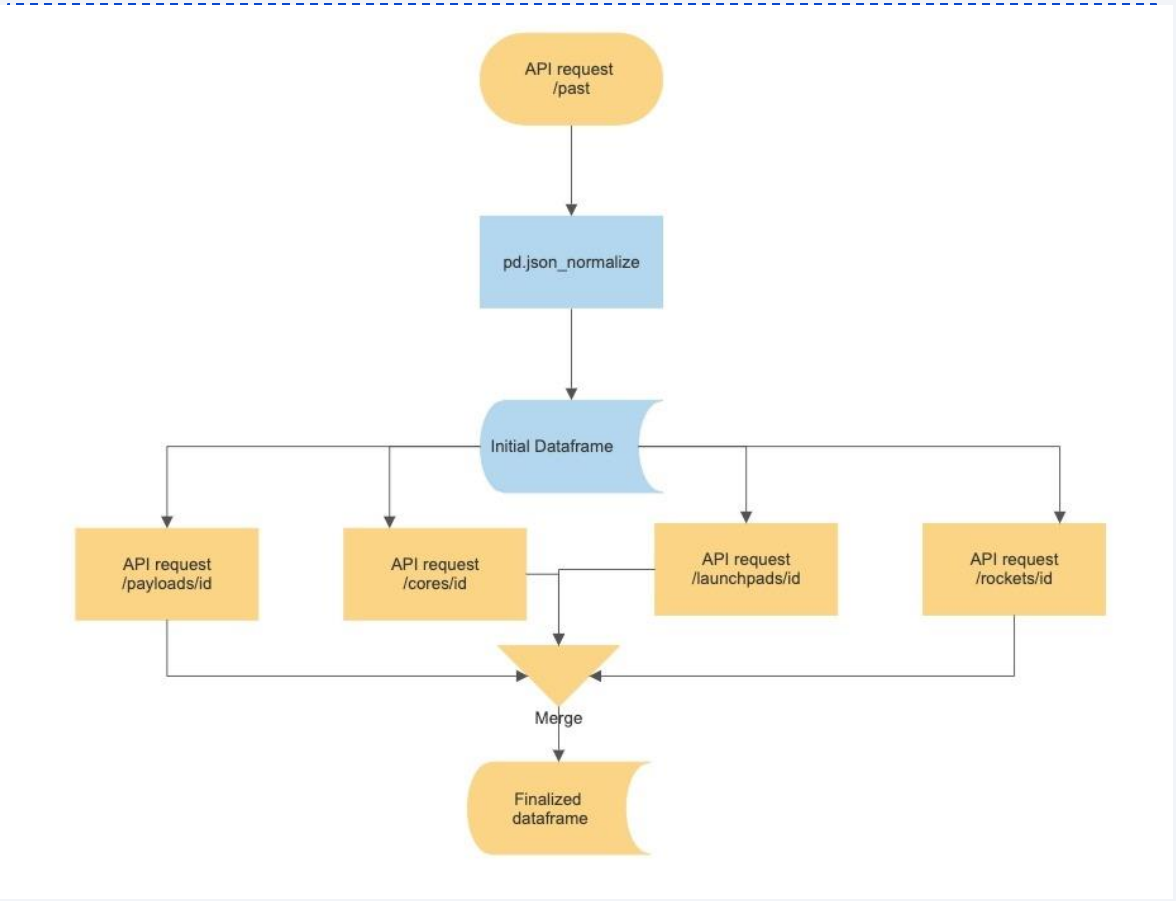
Section 1

Methodology

Data Collection – SpaceX API

All endpoints used start with https://api.spacexdata.com/v4
We made a request from the endpoint /launches/past. The Json data received was converted into a pandas data frame. Many entries consist of the ids of entities not contained in the data.
To complete the data, further API calls were made, using endpoints such as /rockets, /launchpads, /payloads and /cores. This output was used to create an updated and self-contained data frame.

Jupyter Notebook

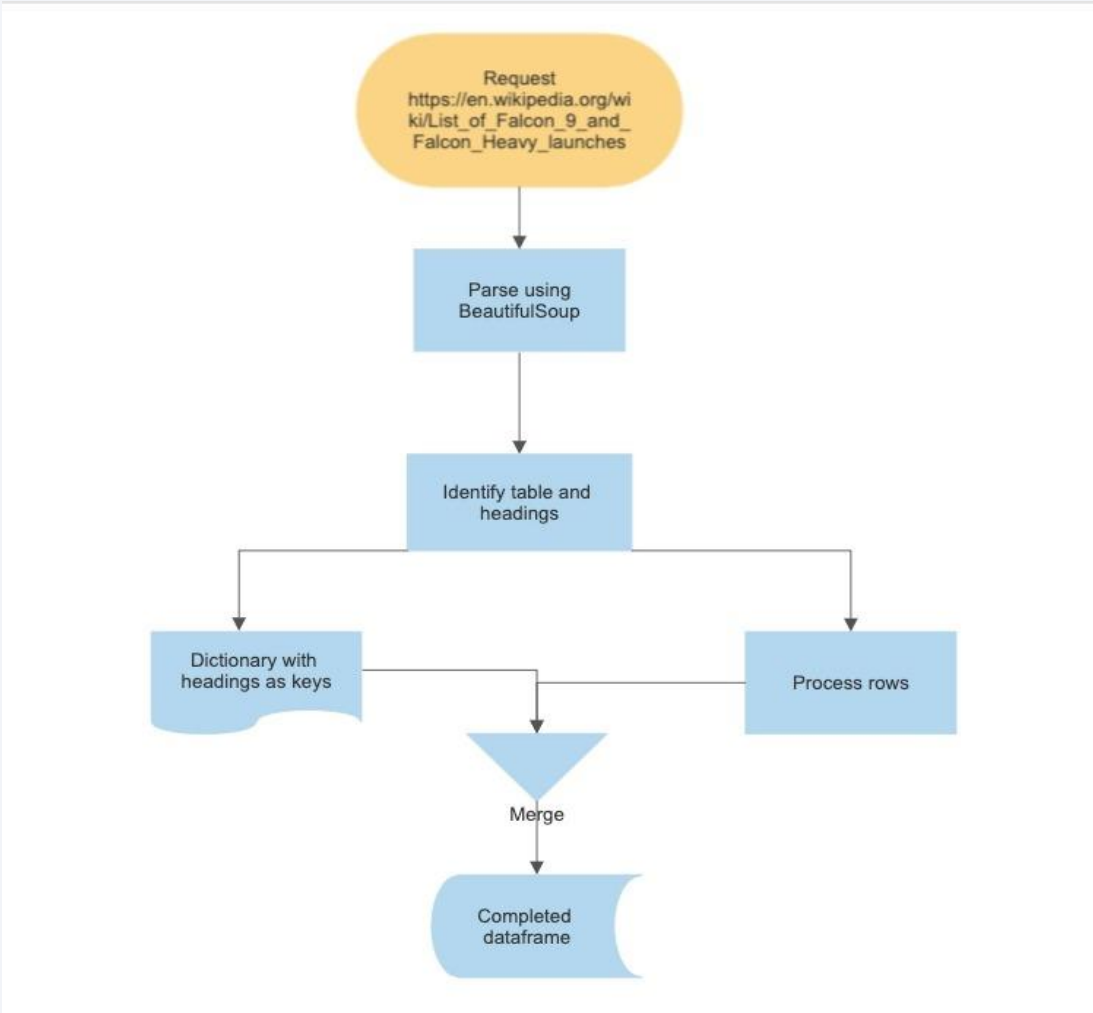


Data Collection - Scraping

More information was gathered from wikipedia at https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

This page was requested and parsed using the BeautifulSoup library. The relevant data was stored in a large table. The headings of this table was extracted and used as the keys for a dictionary. Each row was processed to populate the values of this dictionary. Finally, this was used to create a pandas data frame.

Jupyter Notebook



Data Wrangling

First, launches that used the 'Falcon 1' booster were removed from the dataset, ensuring that only 'Falcon 9' boosters were considered. This was done to make sure that the data was consistent and applicable to the competitor's planned operations.
Next, missing "PayloadMass" entries were replaced with the average of existing values. This was done to ensure that all entries had a value for this important variable, which is likely to have an impact on the mission outcome.
Finally, different "Outcomes" were simplified to binary values of 0 (failure) or 1 (success). Originally, the outcome variable had eight different categories, which would have made it difficult to use in predictive modeling. By simplifying the outcomes to success or failure, the data is easier to work with and can be used to train models to predict the likelihood of mission success.
In preparation for machine learning models, categorical variables were also converted using one hot encoding.

Jupyter Notebook
Jupyter Notebook

EDA with Data Visualization

Using Seaborn as our visualization tool, many charts were created to help identify features that affect the success rate and thus should be included in prediction models.
Scatterplots of Payload Mass, Launch Sites, Orbit Type were are created versus Flight Number.
Success and failure was indicated visually by the color of plotted points.
Similarly, Launch Sites and Orbit typed were plotted versus Payload mass, again with the outcome being represented by color.
Finally the average rate of success was illustrated as trend over time, as well as the average rate of success versus Orbit types.

Jupyter Notebook

EDA with SQL

Using SQL queries, the following questions were answered

- The names of the unique launch sites
- Five records of launch sites starting with 'KSC'
- The total payload mass carried by boosters launched by NASA (CRS)
- The average payload mass carried by booster F9 v1.1
- The date of the first successful landing on a drone ship
- The boosters which resulted in a successful ground pad landing within a range of payload mass.
- The total number of successful and failed mission outcomes
- The names of boosters which carried the maximum payload mass
- The records of successful landing outcomes within a specified range of dates

Jupyter Notebook

Build an Interactive Map with Folium

- All launch sites were marked with opaquely filled in circles: This was done to give the user an overview of the locations of all the launch sites being considered in the analysis. By marking all the launch sites on the map, the user can quickly see the geographic spread of the sites and the proximity to relevant infrastructure.
- Success/failed launches for each site were added as a cluster of marks: This was done to help the user quickly identify the success rate of launches from each site. By marking successful and failed launches with different colors or symbols, the user can easily distinguish between them and visually understand the success rates of each launch site.
- The closest coastline, railway track, highway, and city were marked, and their distances to the Vandenberg Space Launch complex were calculated and displayed with lines: This was done to give the user a sense of the proximity of each launch site to important transportation infrastructure and population centers. By showing the distance to the closest coastline, railway track, highway, and city, the user can get a sense of the accessibility of each launch site and the potential impact of the launch on nearby populations. This information can be important for making decisions about where to launch different types of payloads and under what conditions.

Jupyter Notebook

Build a Dashboard with Plotly Dash

Using Plotly Dash, an interactive dashboard was created to allow for quick and flexible interaction with the data.

- A drop-down list contains each launch site as well as an extra option "All": The drop-down list allows the user to select a launch site or view data for all launch sites. By providing the user with a selection of launch sites, the user can focus on a specific site or compare data across all sites.
- A pie chart follows: If a launch site has been selected in the list, the pie chart shows the success and failures of missions launched from that site. If the "All" option is selected, then the chart will display all launch sites in proportion to the number of successful missions. The pie chart provides a quick visual overview of the success rate for the selected launch site or all launch sites combined.
- A range selector exists to restrict attention to missions with payload mass between selectable limits: The range selector allows the user to focus on missions with a specific payload mass range. By restricting the data to a specific payload mass range, the user can analyze data for a specific payload mass category and identify any patterns or relationships.
- Below this, a scatter plot exists showing mission success (1) or failure (0) vs payload mass as determined by the range selector. If "All" is selected in the drop-down list, then this is for all launch sites; otherwise, it is restricted only to the selected. The scatter plot provides a visual representation of the relationship between mission success and payload mass. By allowing the user to see how payload mass affects mission success or failure, the scatter plot can help the user identify any correlations or patterns.

Jupyter Notebook

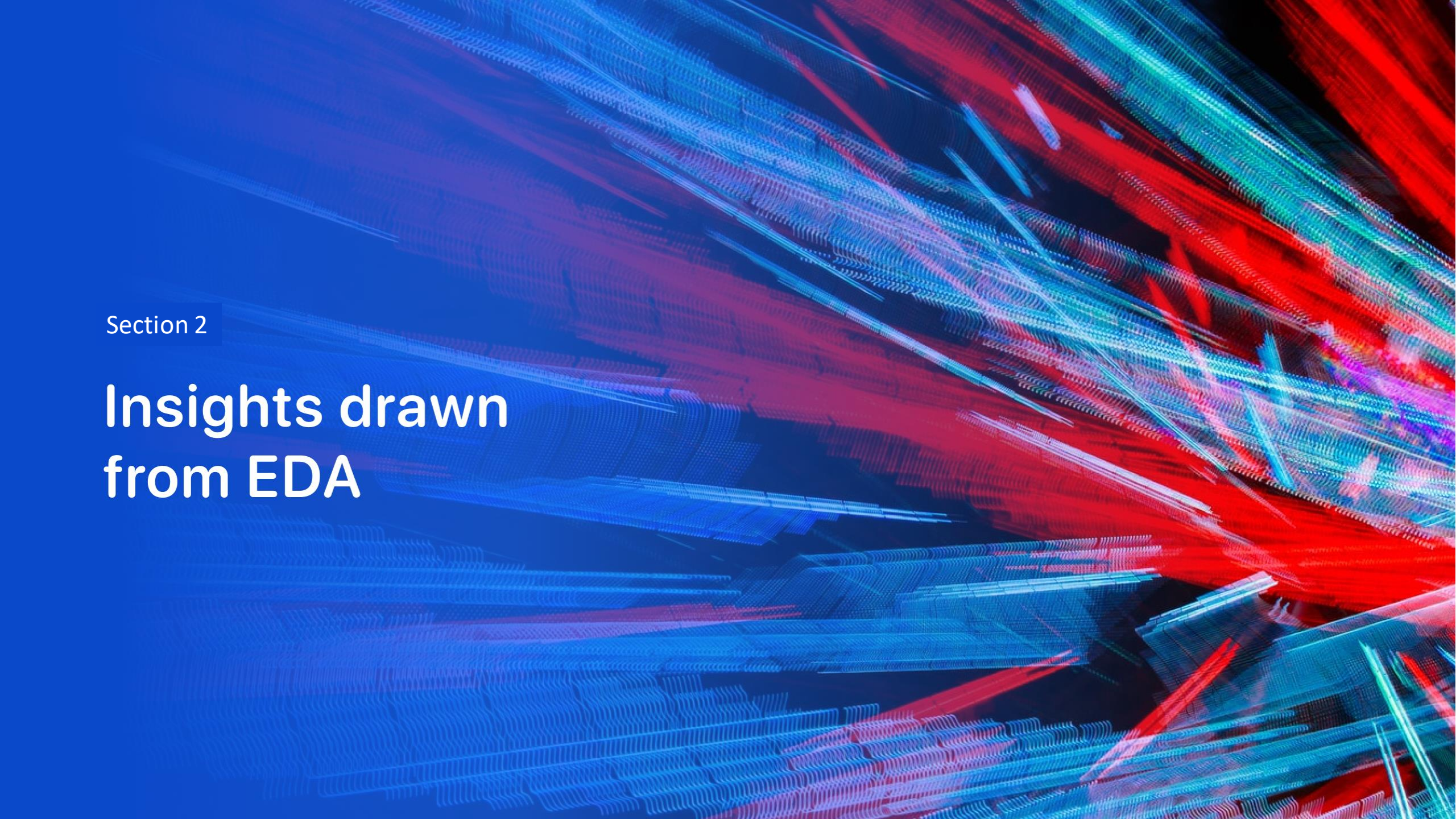
Predictive Analysis (Classification)

To predict the likelihood of mission success, different machine learning classification models were built using the scikit-learn library. The models used were decision trees, support vector machines (SVM), logistic regression, and k-nearest neighbors (KNN).
The dataset was loaded and standardised to ensure that all features had a mean of zero and standard deviation of one. This was done to ensure that each feature contributed equally to the model and that no single feature was overweighted.
The dataset was then split into training and testing sets to ensure that the models were evaluated on data that had not been seen during training. This helped to avoid overfitting, where the models perform well on training data but poorly on unseen data.
GridSearchCV was used to tune different hyperparameters of the models to identify the best parameters for each model. This was done to ensure that the models were optimised and performed to their best ability. Different hyperparameters were adjusted for each model, as they have different requirements.
The scores for each model were noted, and each model was then applied to the testing set to evaluate their performance. Confusion matrices were created to visualise the performance of each model, showing the number of true positives, false positives, true negatives, and false negatives. This helped to evaluate the accuracy of each model in predicting mission success or failure.

Jupyter Notebook

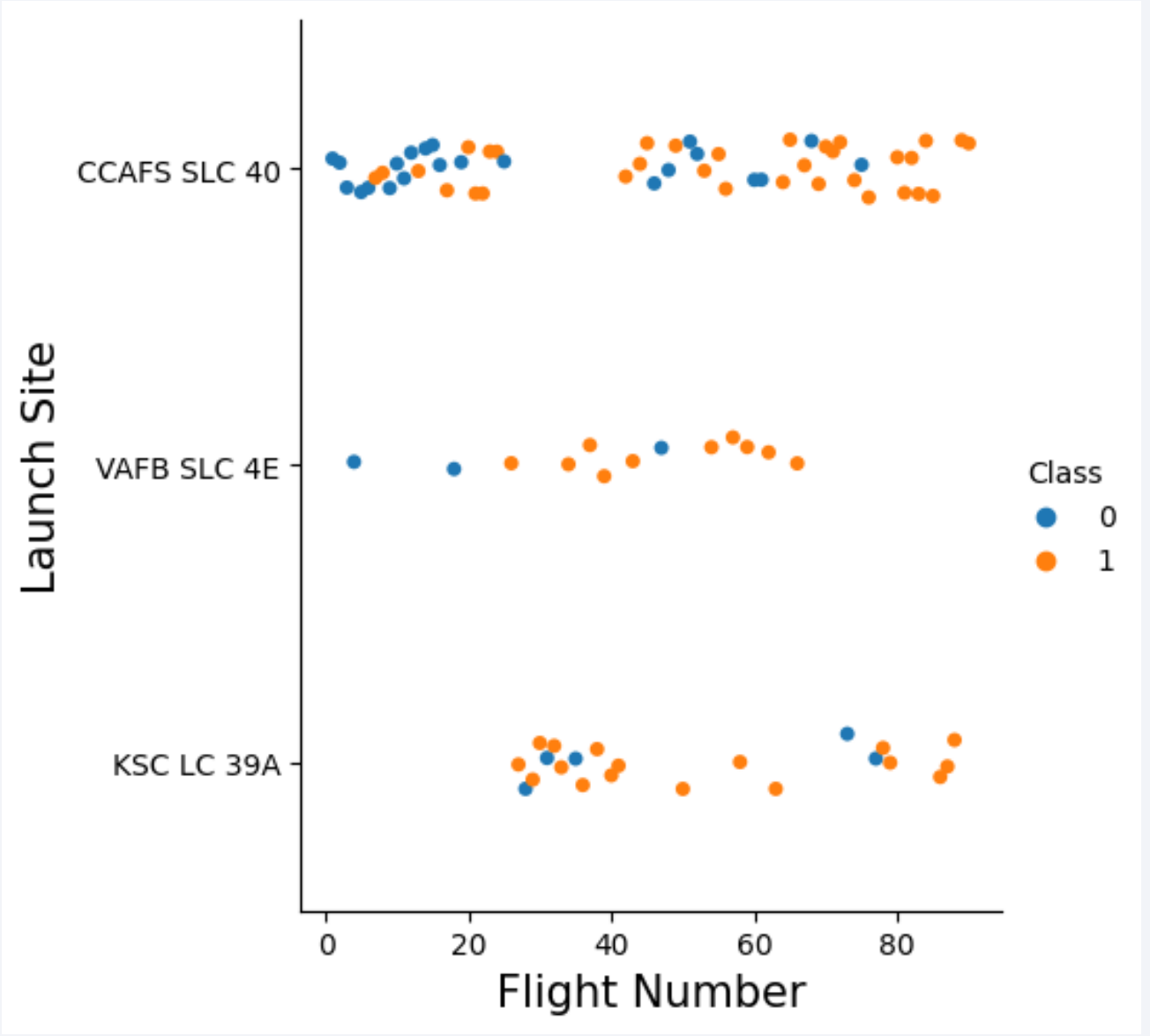
Section 2

Insights drawn from EDA



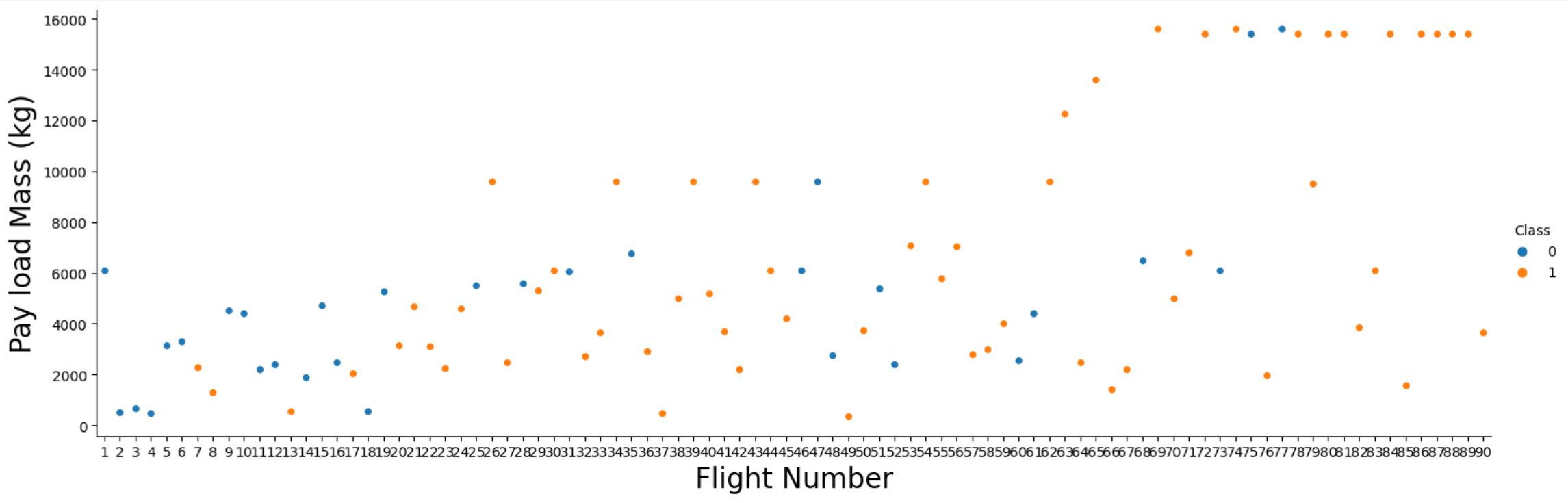
Flight Number vs. Launch Site

- This scatter plot shows the rough trend of improving success rates. It also shows that that the Kennedy Space Center was not used as frequently at the start.



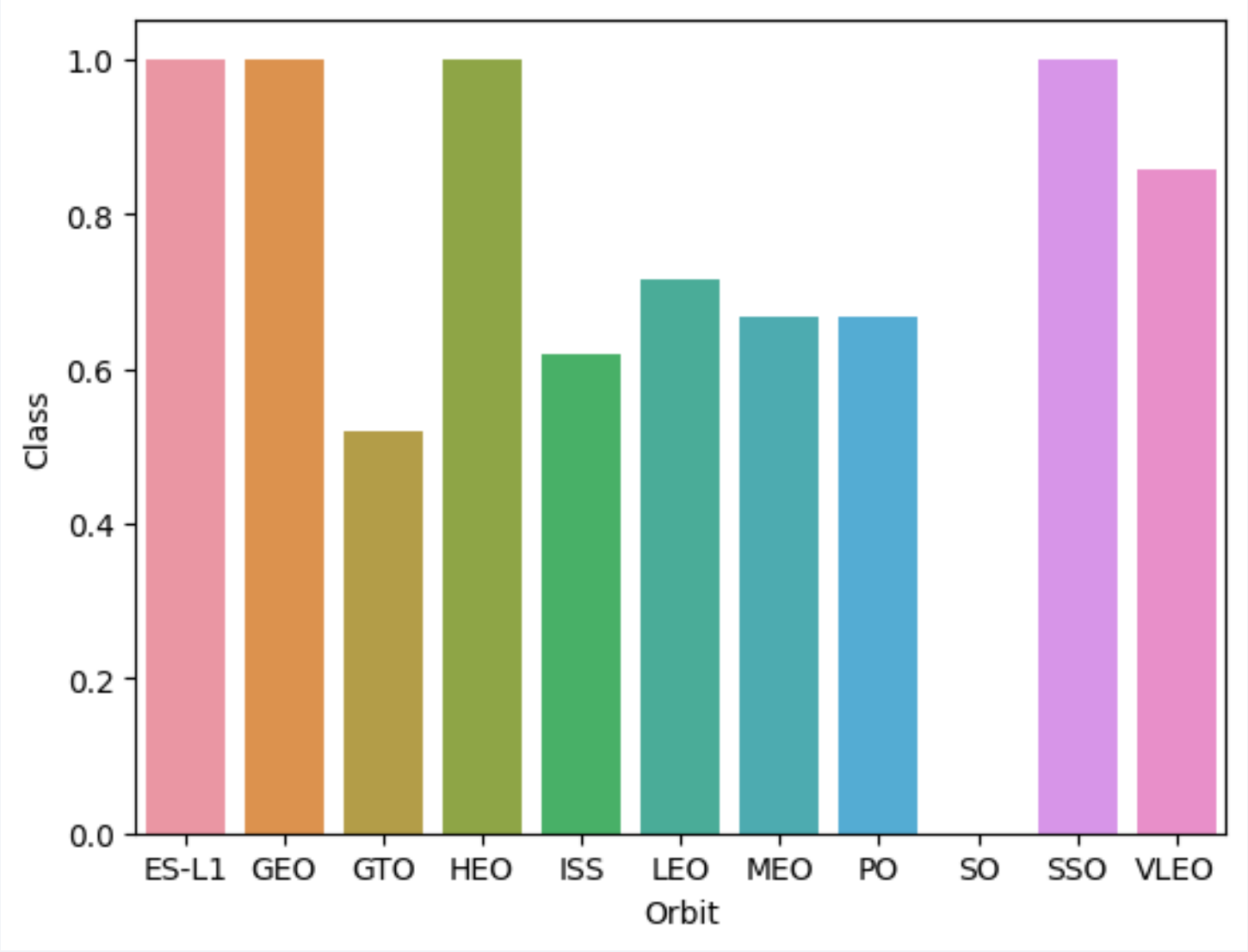
Payload vs. Launch Site

- We can see the same improvement as time goes on (represented by Flight Number). In addition we can see that heavier payloads were introduced much later where they show a good rate of success.



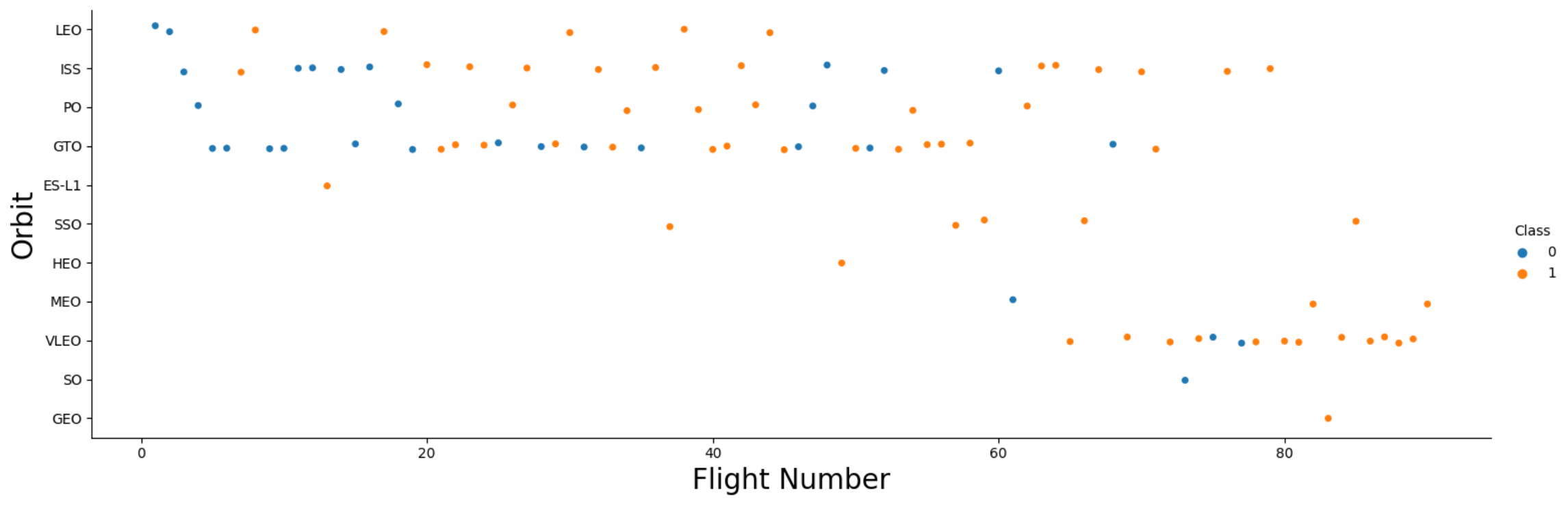
Success Rate vs. Orbit Type

- This shows us the average rate of success for each orbit type.



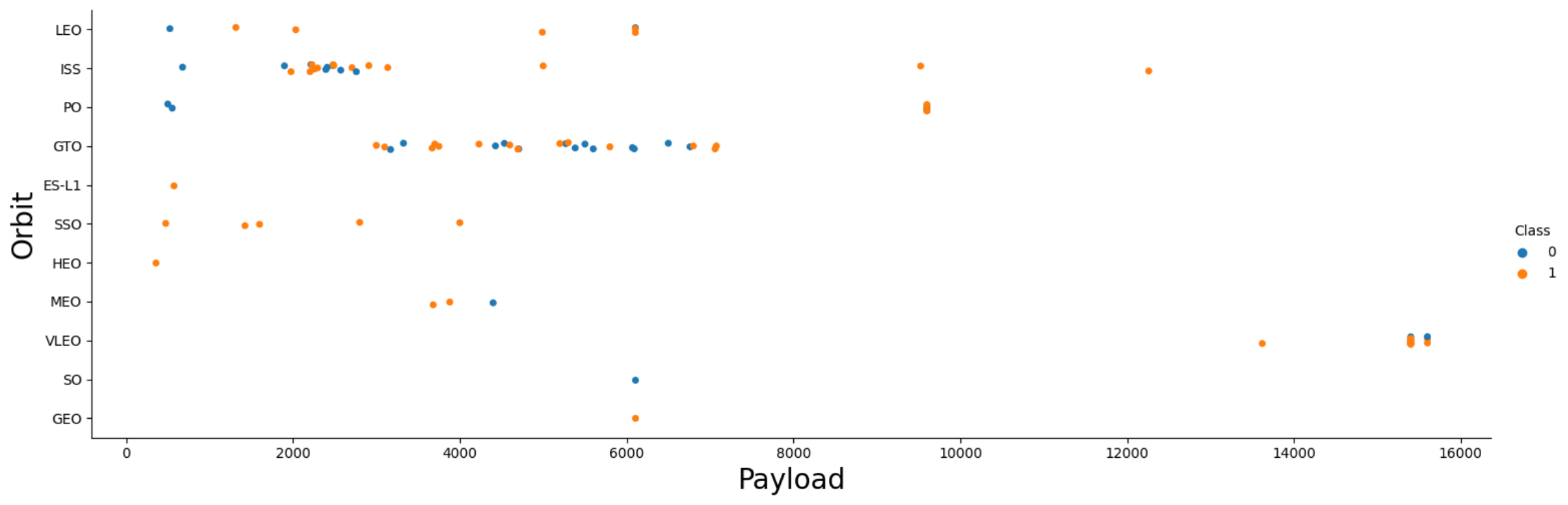
Flight Number vs. Orbit Type

- This shows the same data but allows us to see trends over time thanks to Flight Number. We can see that VLEO was introduced later and immediately became popular.



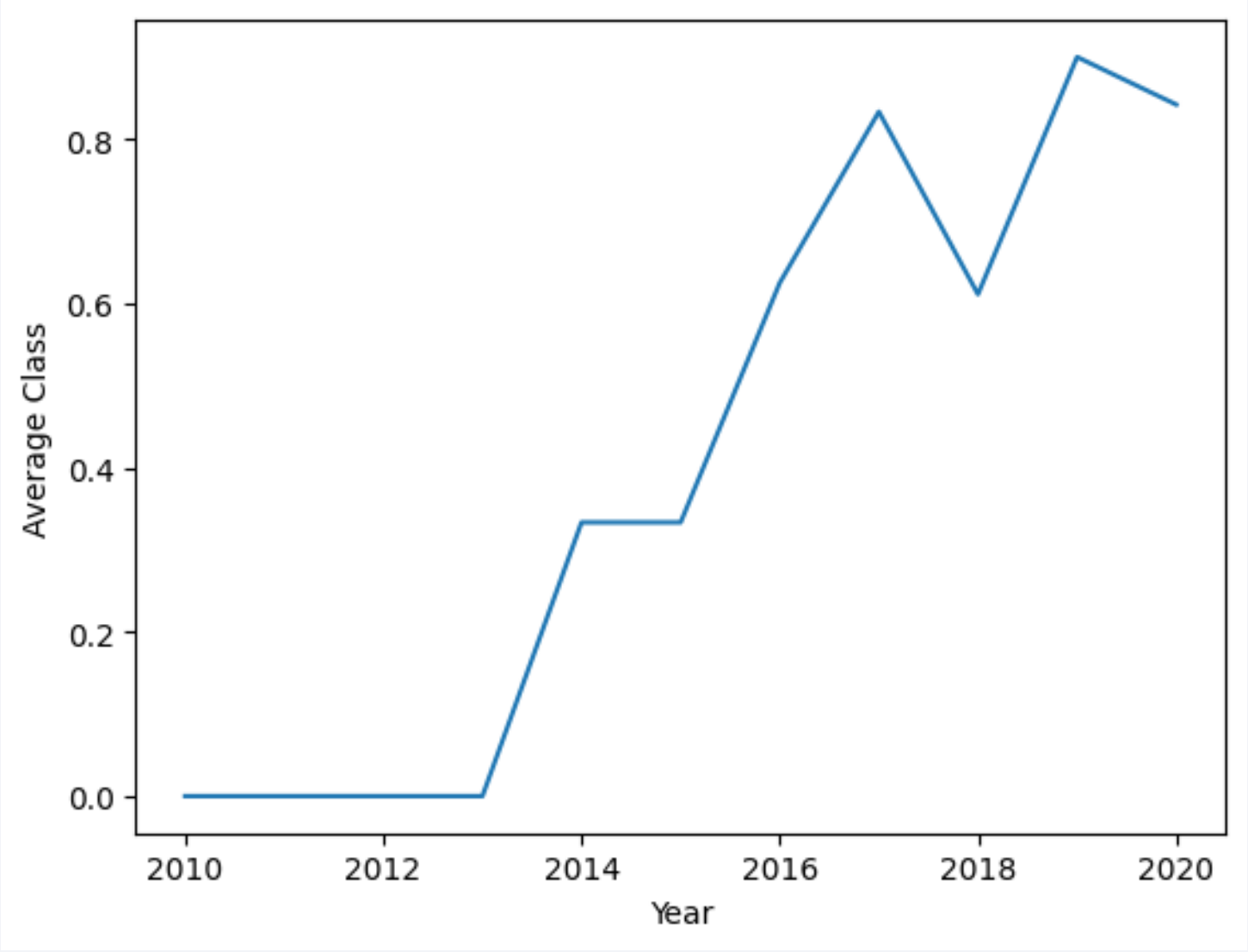
Payload vs. Orbit Type

- Finally, we can compare the payloads of missions along with their orbit types. We can see that VLEO missions are associated with the heaviest of payloads.



Launch Success Yearly Trend

- This shows us the trend of success rate versus time. Overall the rate has increased. More data will be needed to see if the rate will fluctuate around 0.8 or if it will continue to increase.



All Launch Site Names

To get the names of the launch sites, the following sql query was used

select distinct(LAUNCH_SITE) from spacex;

Launch Site Names Begin with 'KSC'

To get five records where the launch site name begins with 'KSC', the following query was executed
select * from spacex where LAUNCH_SITE like 'KSC%' limit 5;

DATE	Time (UTC)	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	Landing _Outcome
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-03-16	06:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2017-05-15	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

Total Payload Mass

The total payload mass carried by boosters launched by NASA (CRS) was found to be 45596kg

This was determined by executing the following query

```
select SUM(PAYLOAD_MASS__KG_) from spacex where CUSTOMER = 'NASA (CRS)';
```

Average Payload for F9 v1.1

The average payload mass carried by booster version F9 v1.1 was found to be 2928kg

This was determined by executing the following query

```
select AVG(PAYLOAD_MASS__KG_) from spacex where BOOSTER_VERSION = 'F9 v1.1';
```

First Successful Ground Landing Date

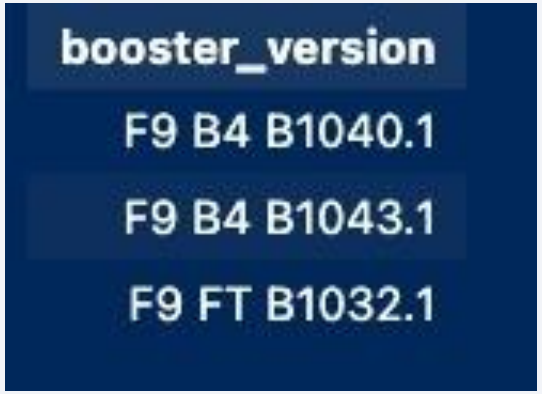
The date of the first successful drone ship landing outcome was found to be 2016-04-8

This was determined by executing the following query

```
select min(DATE) from spacex where "Landing _Outcome"='Success (drone ship)';
```

Successful Ground Pad Landing with Payload between 4000 and 6000

The names of the boosters which resulted in a successful ground pad landing whose payload mass was between 4000kg and 6000kg was determined by executing the following query




```
select distinct(BOOSTER_VERSION) from spacex where "Landing_Outcome"='Success (ground pad)' and PAYLOAD_MASS__KG_ between 4000 and 6000;
```

Total Number of Successful and Failure Mission Outcomes

The total number of successful and failed missions was determined by executing the following query

```
select MISSION_OUTCOME, count(*) as "Count" from spacex group by MISSION_OUTCOME;
```

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

Boosters Carried Maximum Payload

The list of boosters that carried the maximum payload mass was calculated by executing the following query

```
select BOOSTER_VERSION from spacex where PAYLOAD_MASS__KG_ =
(select MAX(PAYLOAD_MASS__KG_) from spacex);
```

This query required the use of a subquery

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

2017 Launch Records

To get records that display the month names, booster versions and launch sites of missions that ended with a successful ground pad landing in the year 2017, the following query was executed

```
select MONTHNAME(DATE) as "Month Name", BOOSTER_VERSION,LAUNCH_SITE,"Landing _Outcome" from spacex
where "Landing _Outcome"='Success (ground pad)' and YEAR(DATE)=2017;
```

Month Name	booster_version	launch_site	Landing _Outcome
February	F9 FT B1031.1	KSC LC-39A	Success (ground pad)
May	F9 FT B1032.1	KSC LC-39A	Success (ground pad)
June	F9 FT B1035.1	KSC LC-39A	Success (ground pad)
August	F9 B4 B1039.1	KSC LC-39A	Success (ground pad)
September	F9 B4 B1040.1	KSC LC-39A	Success (ground pad)
December	F9 FT B1035.2	CCAFS SLC-40	Success (ground pad)

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

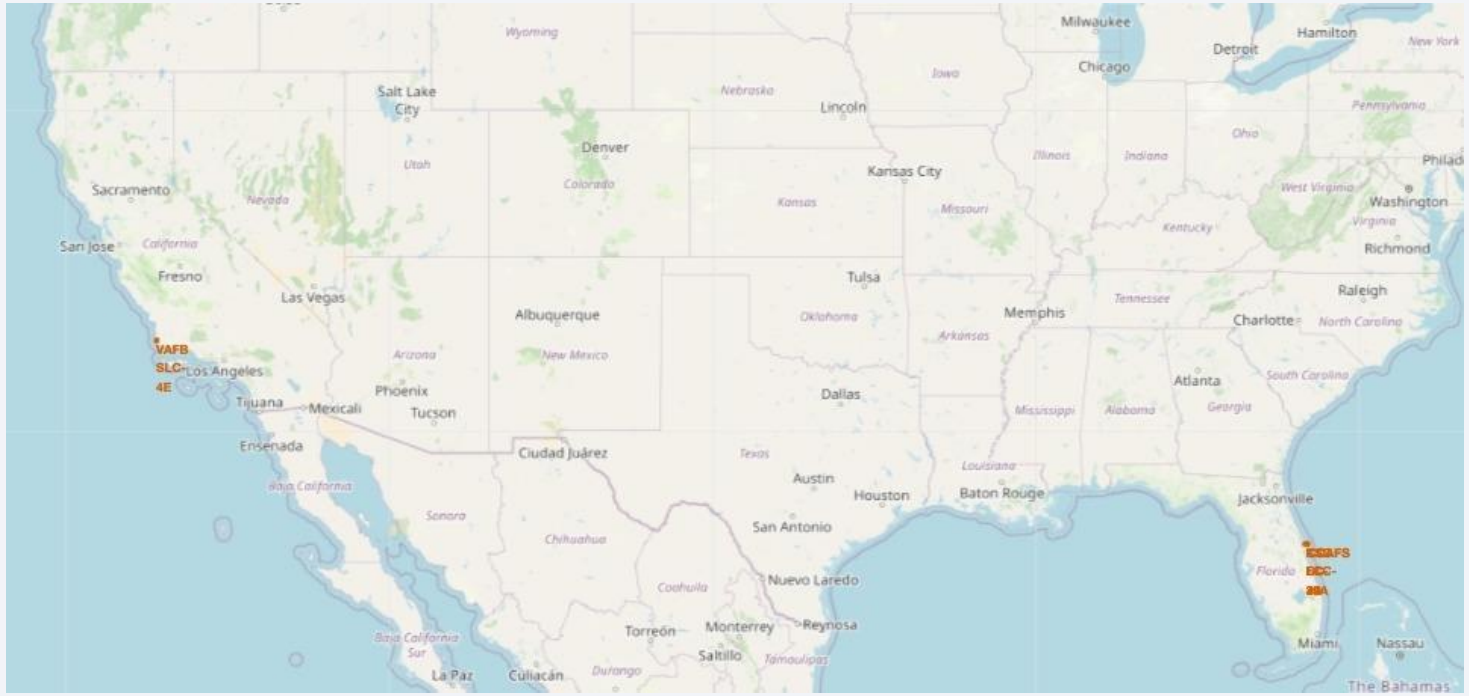
To rank the counts of possible successful landing outcomes between 2010-06-04 and 2017-03-20 the following query was executed

```
select "Landing _Outcome", count(*) as "Count" from spacex where DATE between '2010-06-04' and
'2017-03-20' group by "Landing _Outcome" having "Landing _Outcome" like 'Success%' order by 2;
```

Landing _Outcome	Count
Success (ground pad)	3
Success (drone ship)	5


Section 3
Launch Sites
Proximities Analysis

Location of Launch Sites



The positions of the launch sites are indicated in this screenshot. The Space Launch Complex in the Vandenberg Space Force Base (formerly the Air Force Base) is clearly seen on the western coast of America.

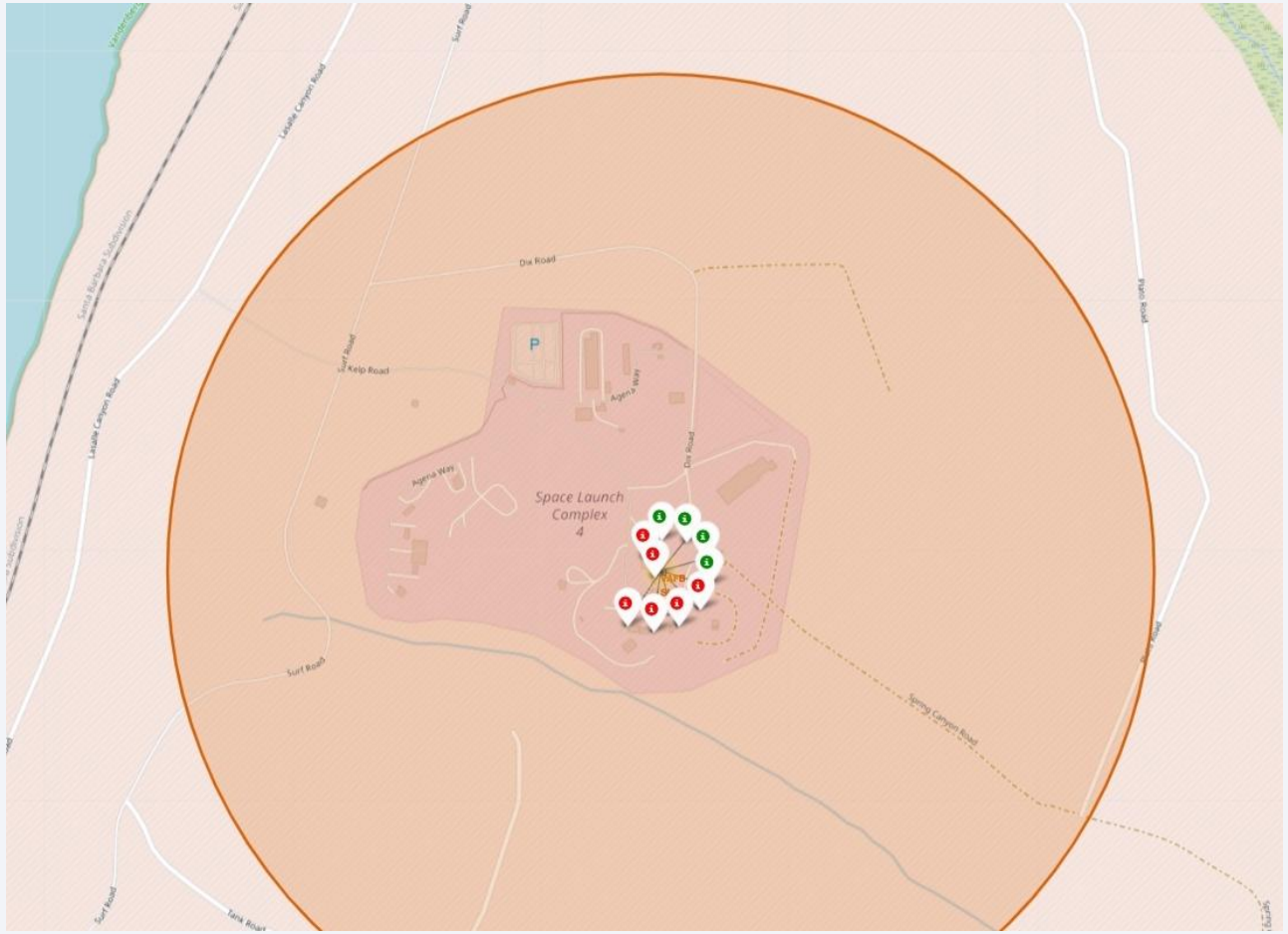
There are three more sites in close proximity in Florida. These can be seen more clearly in the next slide

Location of Launch Sites

Here we focus on those launch sites located on the east coast of Florida. The Kennedy Space Center Launch Complex can be easily seen on the left.

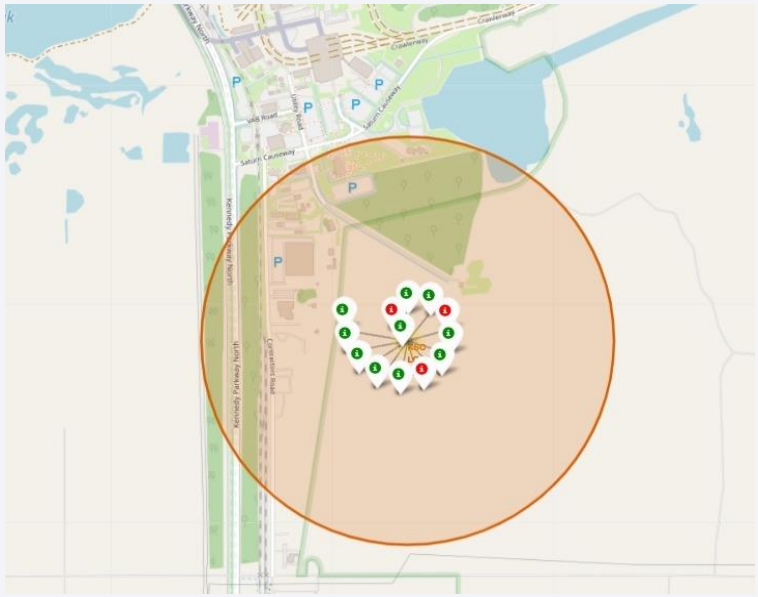
On the right, the Cape Canaveral Space Launch Complex (CCAFS SLC-40) can be seen overlapping with the Cape Canaveral Launch Complex (CCAFS LC-40). They refer to the same physical location with the latter being the name prior to a recent change.

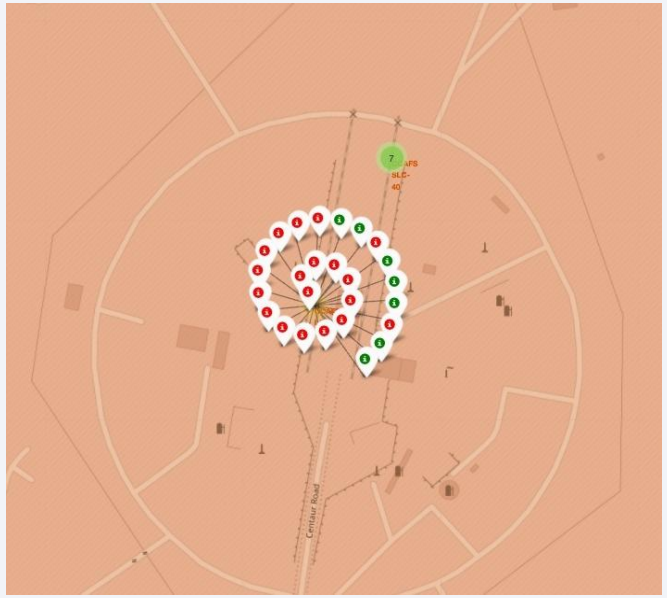
Representations of Launch Outcomes



In this map, focused on VAFB SLC-4E, we can see mission outcomes, with green markers indicating success and red indicating failure. The spiral shape is used to ensure each marker is clearly visible. Starting from the middle, the markers are placed in chronological order allowing us to see trends.

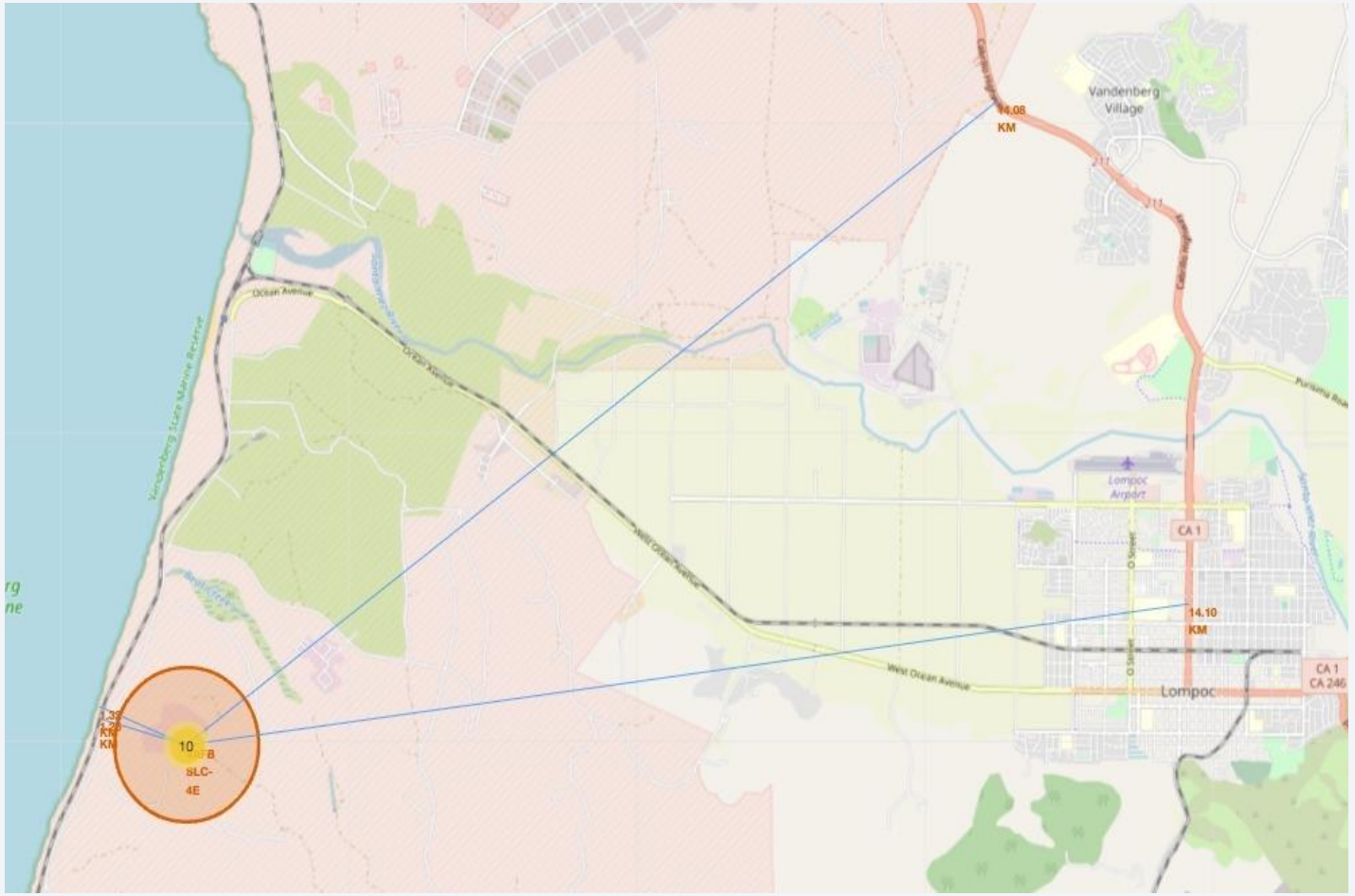
Representations of Launch Outcomes



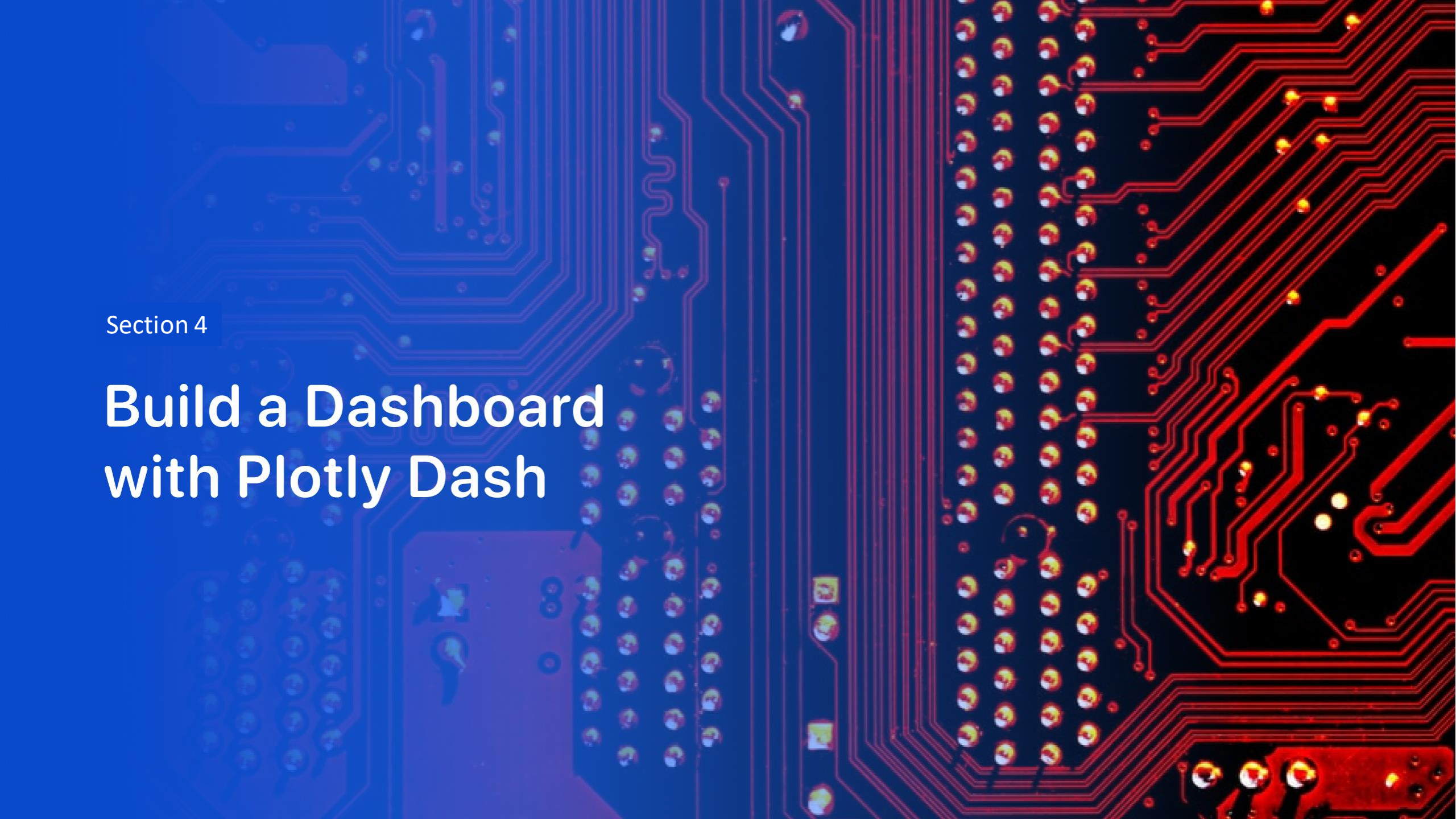


Similarly, we have representations of the outcomes for launches from KSC LC-39A, CCAFS SLC-40 and CCAFS LC-40, in order from left to right.

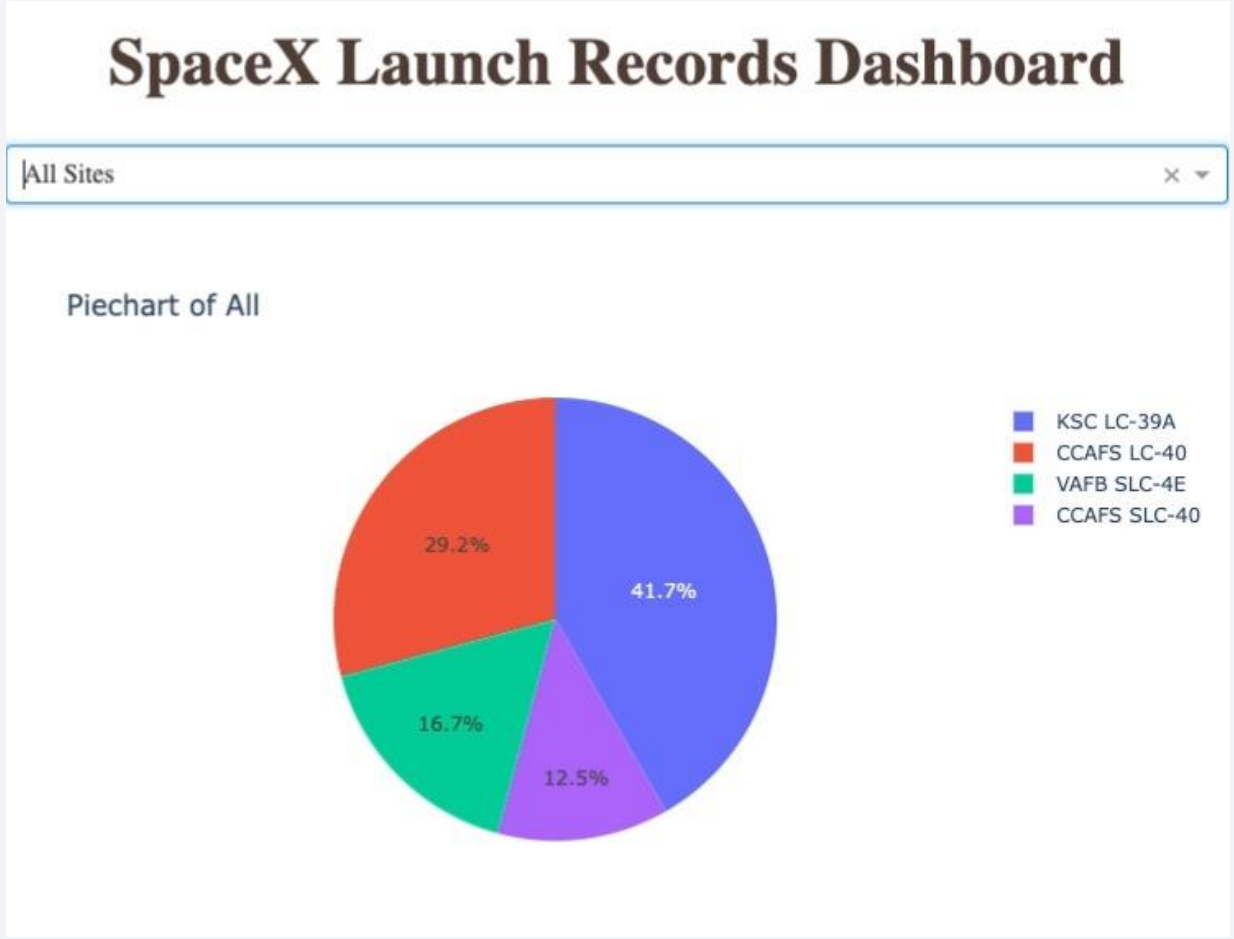
Proximity of Launch Sites to Important Features



This map illustrates the location of the Vandenberg Launch site in relation to the city of Lompoc (14.1 km away), the Cabrillo highway also known as Highway 1 or the Pacific Coast Highway (14.08 km at the point selected), also the coastline (1.32km) and the Santa Barbara Subdivision train line (1.26km).
This pattern is repeated with the other sites, that is, close to the ocean and railways and a greater distance from major highways and population centers. Conservation areas are also nearby, presumably to act as buffers in case of catastrophic accidents.

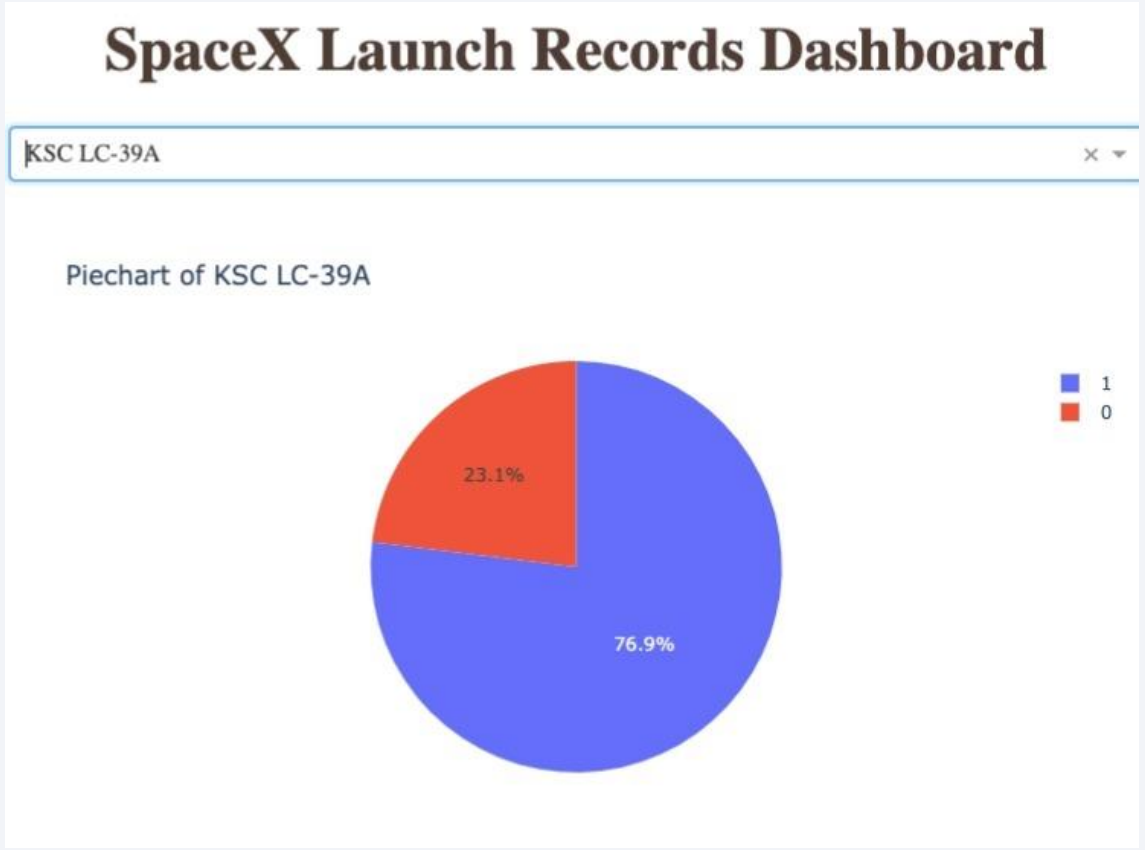

Section 4
Build a Dashboard with Plotly Dash

Successful Outcomes by Location



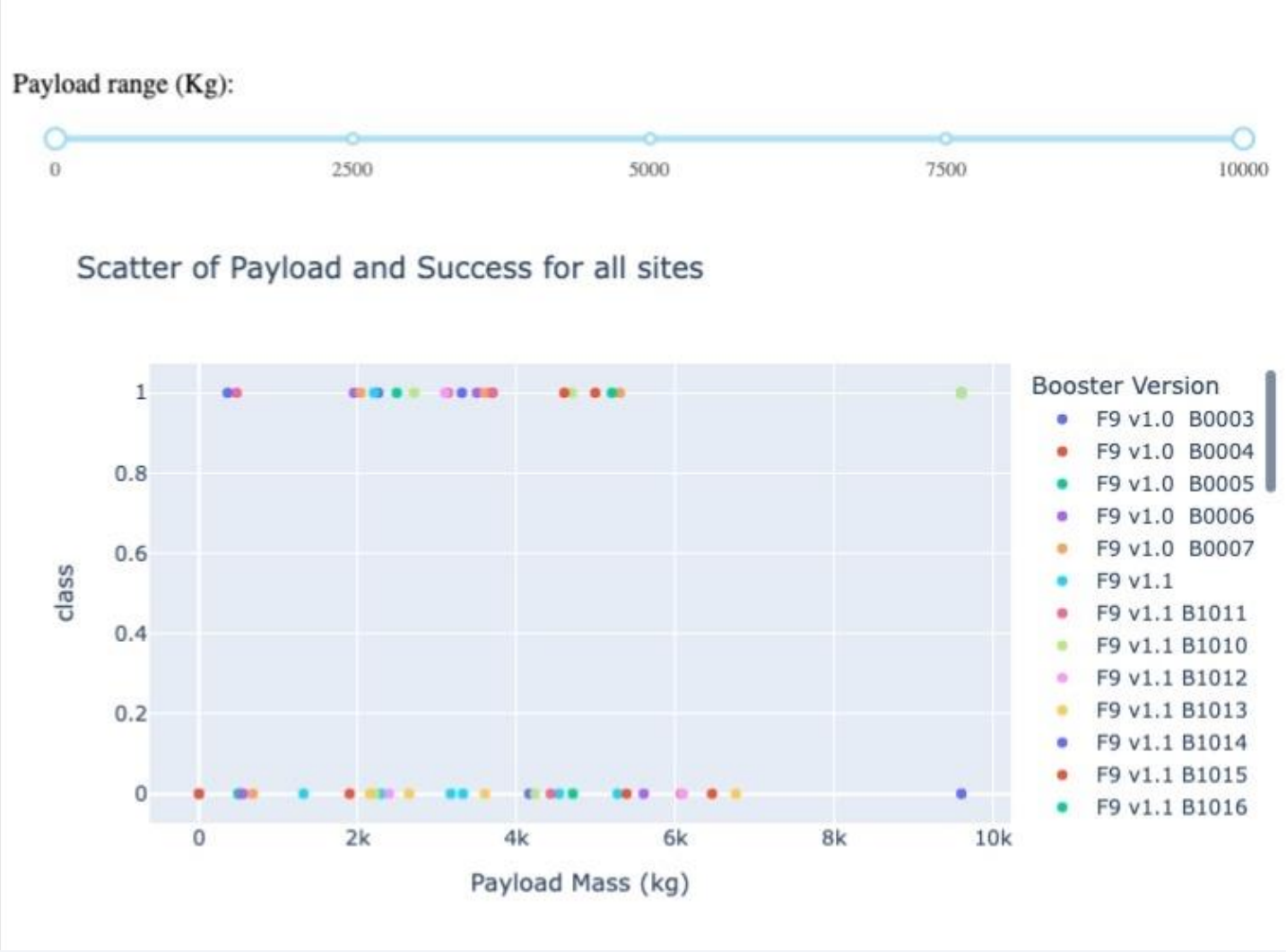
With the drop down list set to "All" we have a pie chart showing the distribution of successful missions. In this we can see the Kennedy Space Center had the most successful launches by a significant margin.

Kennedy Space Center Mission Breakdown



By selecting a launch site from the drop down list, we can see proportion of successful missions versus failures for that site. Doing this revealed that the Kennedy Space Center also had the highest proportion of successful launches.
As can be seen in the graph, 76.9% of launches were successful from this location.

Payload vs Launch Outcomes



Another feature of the dashboard allows us to display mission outcomes for specific ranges of payload mass. This screenshot shows the total distribution with booster version being indicated and selectable.
In this context "1" indicates success and "0" failure.

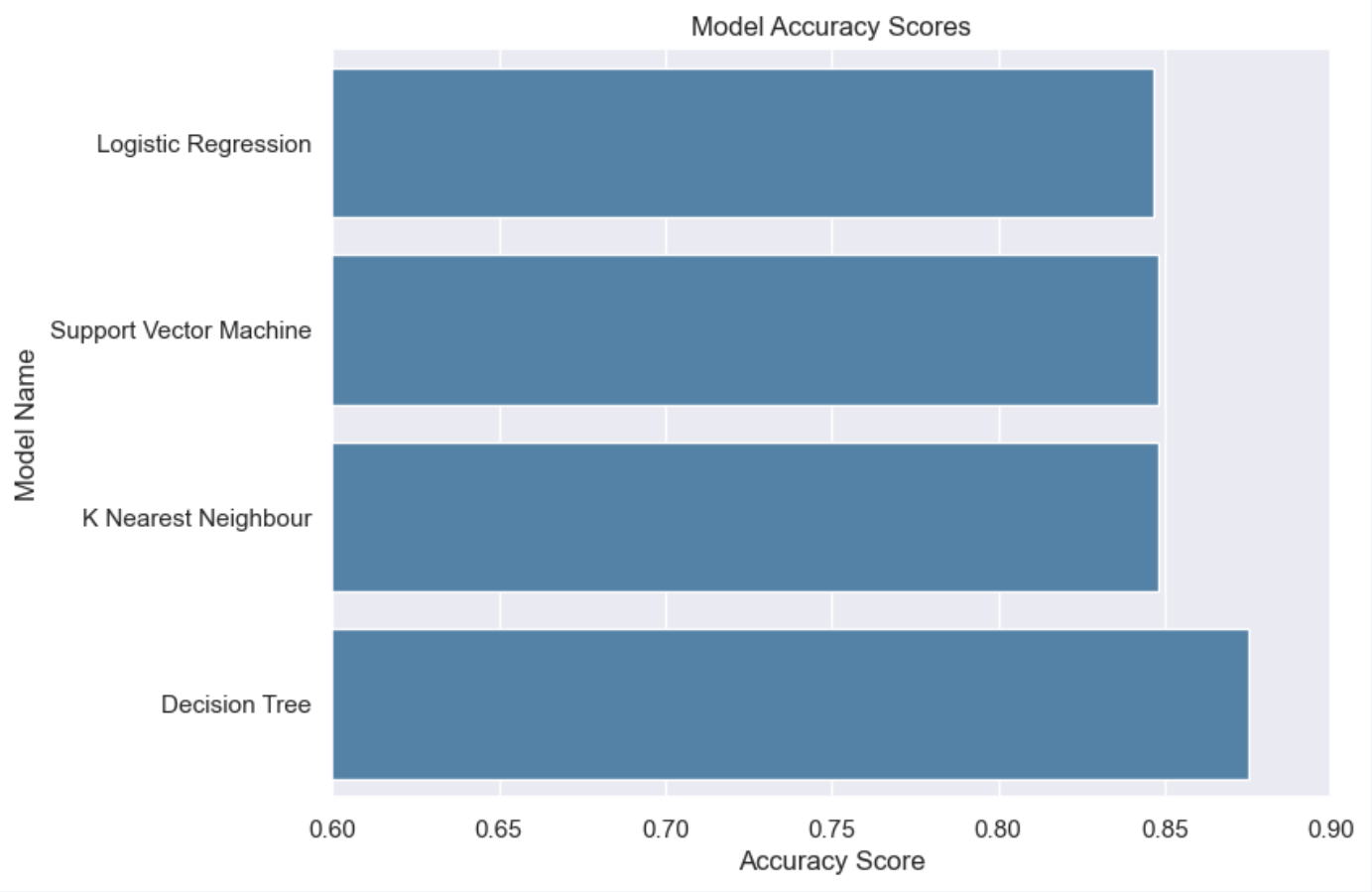
Payload vs Launch Outcomes

Here we have focused on payloads between 2500kg and 5000kg, which visually had the highest rate of success. There is no obvious indication of a superior booster version.

Section 5

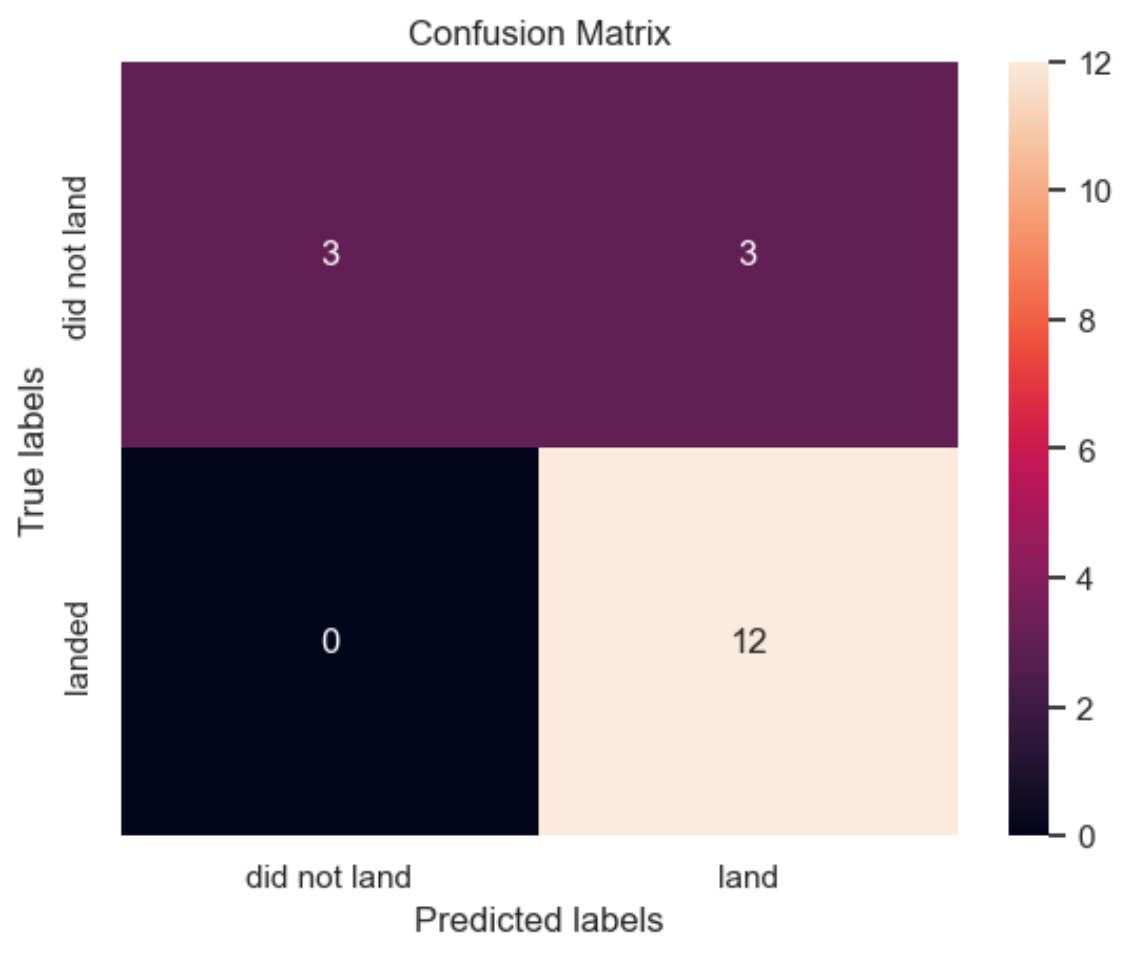
Predictive Analysis (Classification)

Classification Accuracy



As seen in the chart, all of the models performed well and with almost identical accuracy scores.
However, the model based on the Decision Tree algorithm was seen to perform the best with an accuracy score of 0.875

Confusion Matrix



Here we have the confusion matrix produced when the decision tree model was applied to test data.

This shows that the model has a bias towards predicting a successful landing. Of the missions that did end successfully, the model was able to predict this outcome 100% of the time. However in cases when the mission did not end well, the model was only 50% effective.

This behavior was also observed in all of the model predictions and suggests a need for further refinement of the data pipeline. In addition, expert knowledge on the field of launches may help produce more robust models.

Refinements

It is possible to determine what certain models view as the most significant feature in making their final prediction. For models such as the Decision Tree and Logistic Regression models, this can easily be done (see appendix for code fragments). Due to the non-linear nature of k-NN and SVM (when the sigmoid kernel is used), this is much harder for these cases.

For the first two models, it can be seen that 'Legs_True' (and equivalently 'Legs_False') are highly significant. These features refer to the presence of Landing Legs being present on the booster for this mission. In practice, the absence of Landing Legs may preclude the possibility of a successful landing. As such, it may be wise to remove records where legs are not present from the data set to allow other features to be more significant.

In addition, features referring to specific boosters (like Serial_B1050, Serial_B1058 and others) appear to be highly significant. It is possible that these boosters are now obsolete, being used by SpaceX for testing and development. If the intent of these models is to provide predictions to influence pricing, then it may be wise to do specific domain research to see if these boosters are still in use or if they should be removed as well.

Conclusions

In conclusion, this project undertook a comprehensive analysis to predict the likelihood of mission success for SpaceX launches in the aerospace industry. Through the utilization of various data visualization techniques, SQL queries, folium maps, and interactive visualizations using Plotly Dash, valuable insights were gained.

Exploratory data analysis, facilitated by SQL queries, provided a deeper understanding of the dataset. Visualizations such as scatter diagrams, pie charts, and bar charts aided in uncovering patterns and relationships between variables. Notably, a bar chart displayed the accuracy scores of different machine learning models, with the Decision Tree model emerging as the top performer by a narrow margin.

Additionally, interactive maps created with folium showcased the geographical distribution of launch sites, highlighted mission outcomes, and illustrated the proximity of essential infrastructures such as coastlines, railway tracks, highways, and cities to the Vandenberg Space Launch complex.

The Plotly Dash interactive dashboard allowed for dynamic exploration of the dataset. It featured a dropdown menu to select launch sites, a pie chart depicting success/failure outcomes, a range selector for payload mass, and a scatter plot visualizing the relationship between payload mass and mission success. These interactive visualizations provided a comprehensive overview and allowed users to gain insights by adjusting the parameters.

Conclusions

Throughout the project, advanced machine learning models such as decision trees, logistic regression, support vector machines, and k-nearest neighbors were employed to predict mission outcomes. The models were fine-tuned using GridSearchCV to optimize their performance.

Refinements were suggested based on feature importance analysis from the Decision Tree and logistic regression models. Recommendations included removing cases where the booster did not include landing legs, aggregating or removing individual booster types/serial numbers, and further domain-specific research to determine the relevance of certain booster types.

By integrating data visualization techniques, SQL queries, folium maps, and interactive visualizations using Plotly Dash, this project provided valuable insights into predicting mission success in the aerospace industry. These findings can inform decision-making processes, enhance risk assessment, and drive improvements in the competitive landscape of the aerospace market. Continued monitoring, data updates, and refinement of the models will further enhance their accuracy and reliability.

Appendix

- All Jupyter Notebooks can be found at https://github.com/rplmath/edXMLCapstone/tree/main
- The code to rank features is shown below

```
from sklearn.tree import export_text
best_tree = tree_cv.best_estimator_
tree_text_representation = export_text(best_tree, feature_names=X.columns.tolist())
print(tree_text_representation)
#this outputs the decision rules determined by the model. Features appearing earlier in the tree are more significant

log_reg=logreg_cv.best_estimator_
importance = log_reg.coef_[0]

features = sorted(zip(X.columns, importance), key=lambda x: abs(x[1]), reverse=True)

# Display the top 10 features
for feature, coef in features[:10]:
    print(f"Feature: {feature}, Score: {coef:.5f}")
```

✓ 0.6s

Thank you!