Winning Space Race with Data Science

Marco Multari
March 02, 2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
    - through API
    - Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis
    - with SQL
    - with Data Visualization
  - Interactive Visual Analytics with Folium
  - Dashboard with Plotly Dash
  - Machine Learning Prediction (Classification)
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

Introduction

- Project background and context

SpaceX is a private aerospace company founded in 2002 by Elon Musk with the goal of creating the technologies to reduce the costs of accessing space and enable the colonization of Mars. One of the keys to saving is the use of partially reusable rockets. So, the ability to identify the components that positively influence the success of the first-stage rocket landing facilitates cost optimization.

The purpose of this project is to find the best classifier to predict if the first stage of Falcon9 rocket will land successfully or not which in-turn can be used to determine and minimize the cost by SpaceX.

- Problems you want to find answers
  - To identify, if any, the factors determining the success of the Falcon 9 first-stage landing.
  - Evaluate how various factors impact the success of the landing.
  - Conditions that allow achieving the best results.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX Rest API and Web Scraping from Wikipedia.
- Perform data wrangling
  - One-hot encoding was applied to categorical features
  - Dropping irrelevant columns
- 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

To collect the data, we used a combination of API requests from the SpaceX REST API and data extraction from a table in the SpaceX Wikipedia page with the Web Scraping methodology.
We had to use both data collection methods to obtain complete information on the launches for a more detailed analysis.

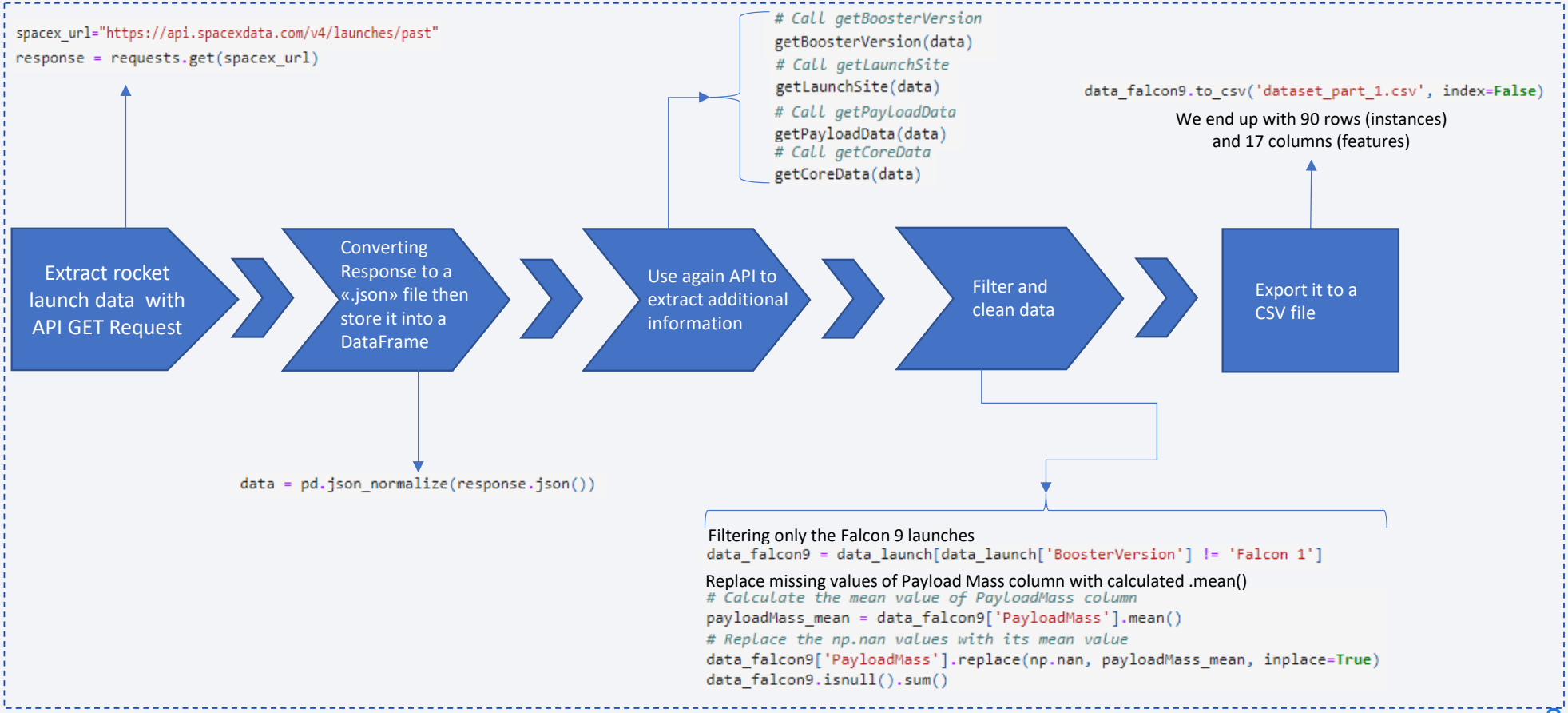
Data Columns are obtained by using SpaceX REST API:

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

Data Columns are obtained by using Wikipedia Web Scraping:

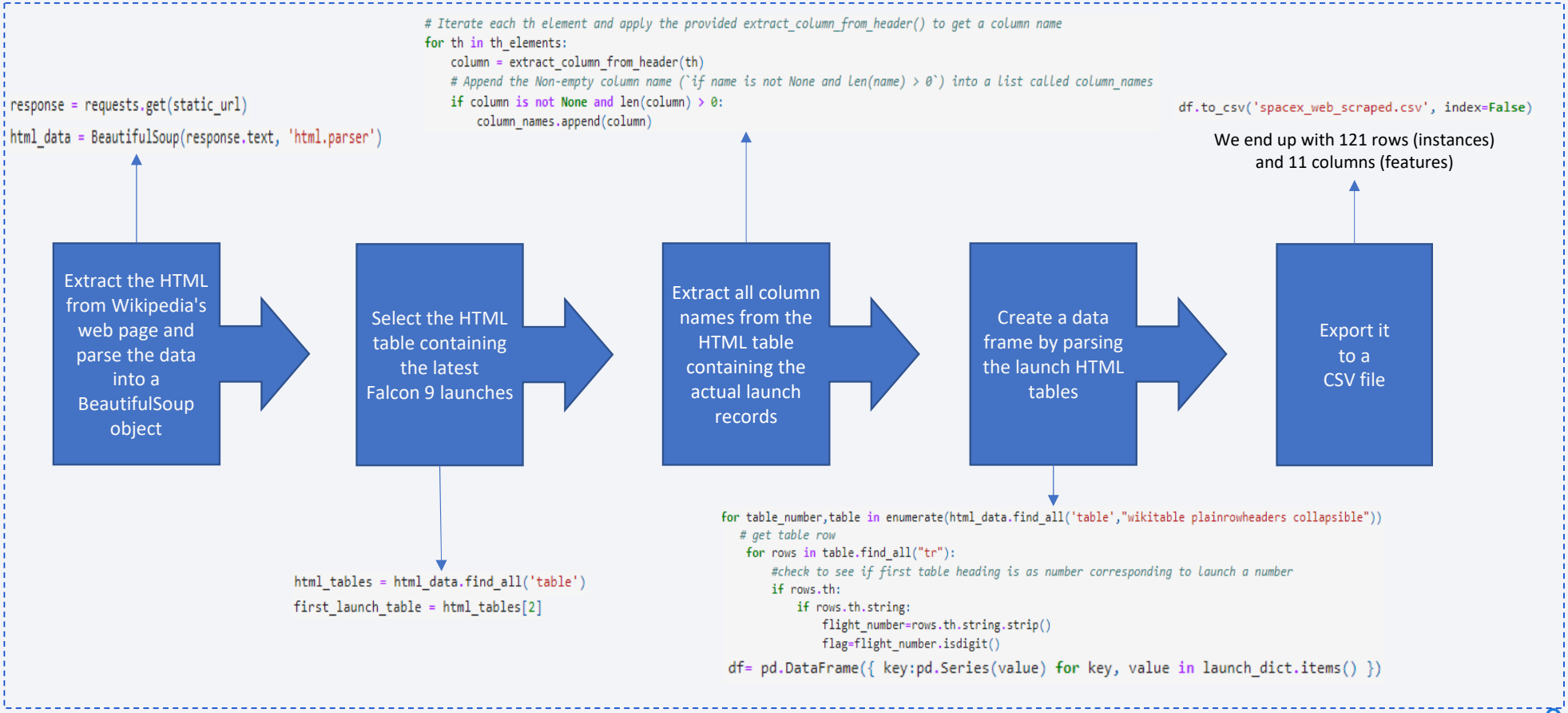
Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



GitHub Notebook: Data Collection - API

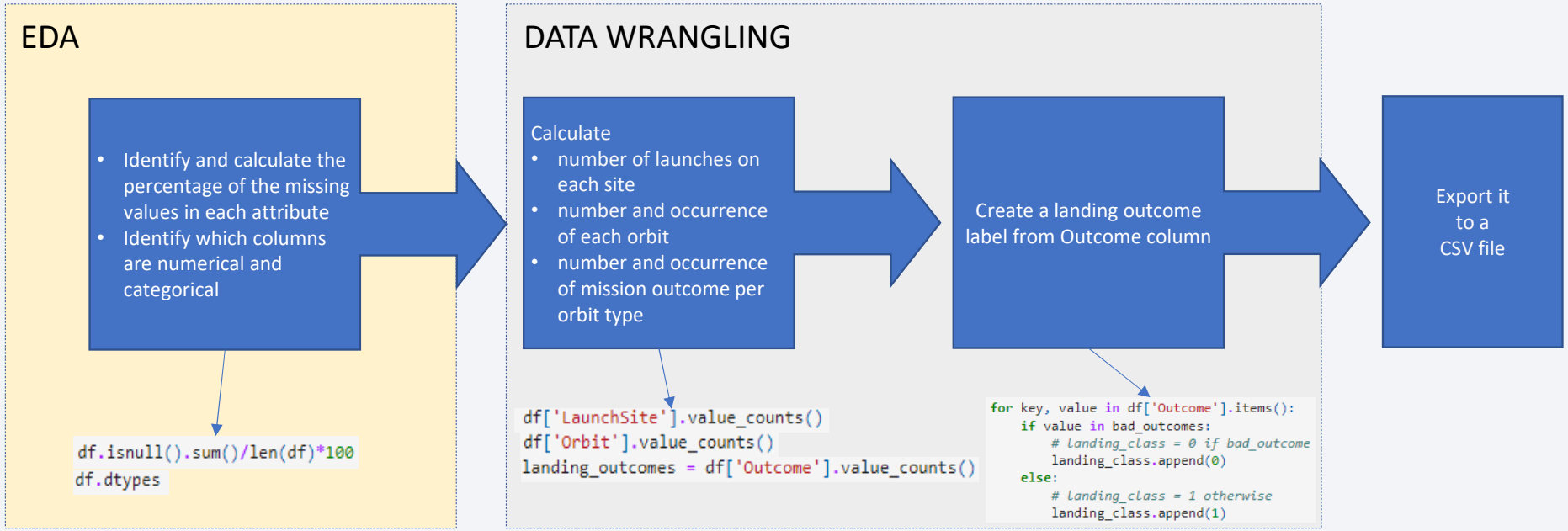
Data Collection - Scraping



[GitHub URL: Data Collection – Web Scraping]

Data Wrangling

Through an exploratory data analysis, the training labels were identified, and cleaning tasks to be applied to the data were defined. The data is subsequently processed so that there are no missing entries, and categorical features are encoded using one-hot encoding.



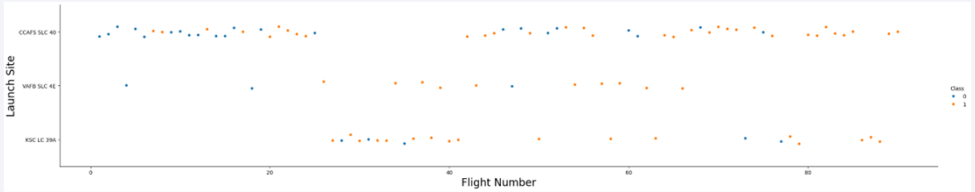
GitHub URL: Data Collection – Data Wrangling

EDA with Data Visualization (1/2)

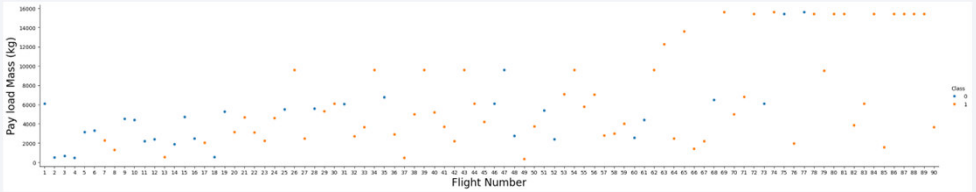
Scatter point Chart

Scatter plots show the relationship between variables. If a relationship exists, they could be used in machine learning model.

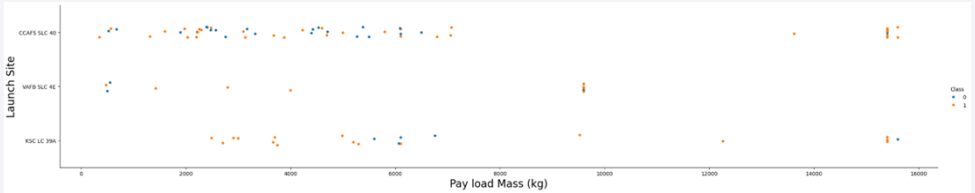
Visualize the relationship between Payload and Launch Site



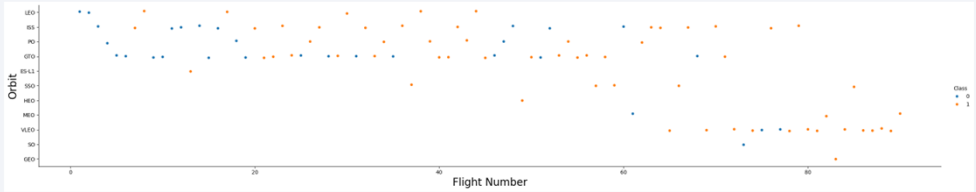
Visualize the relationship between Flight Number and Launch Site



Visualize the relationship between Payload and Orbit type



Visualize the relationship between Flight Number and Orbit type

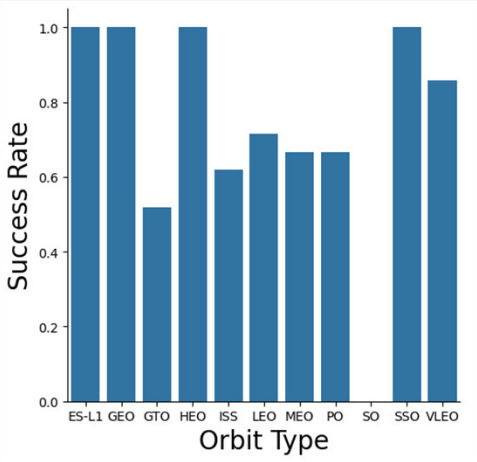


GitHub URL: Data Collection – EDA with Data Visualization

EDA with Data Visualization (2/2)

Bar Chart

Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value.

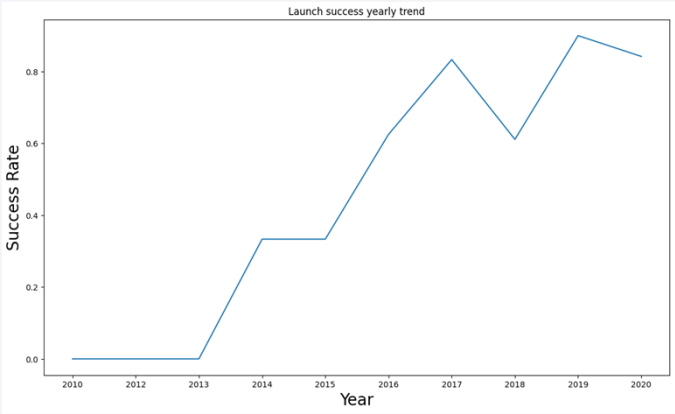


Visualize the relationship between success rate of each orbit type

Line Chart

Line charts show trends in data over time (time series).

Visualize the launch success yearly trend



GitHub URL: Data Collection – EDA with Data Visualization

EDA with SQL

Performed SQL queries:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string ‘CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Ranking 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: Data Collection – EDA with SQL

Build an Interactive Map with Folium

With the support of the Folium library, elements have been displayed on a map for the purpose of finding geographical patterns on launch sites useful for defining optimal locations to build new launch sites and perform an interactive visual analysis to discover relationships between launch sites and success rate.

Markers of all Launch Sites:

- Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations.

Coloured Markers of the launch outcomes for each Launch Site:

- Added coloured Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.

Distances between a Launch Site to its proximities:

- Added coloured Lines to show distances between the Launch Site VAFB SLC-4E and its proximities like Railway, Highway, Coastline and Closest City.

GitHub URL: Interactive MAP with Folium (through nbviewer.org)

Build a Dashboard with Plotly Dash

Using the Plotly Dash library, a web page has been developed with a dashboard for the interactive visualization of data related to launches and some features that could influence the outcome.

The dashboard is composed as follows:

- a dropdown list to enable Launch Site selection;
- a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected;
- a slider to select Payload range;
- a scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster version.

GitHub URL: Interactive Dashboard with Ploty Dash

Predictive Analysis (Classification)

Preliminary tasks

- Creating a NumPy array from the column “Class” in data
- Standardizing the data with StandardScaler, then fitting and transforming it

Splitting the data into training and testing sets with train_test_split function

Models evaluation

Classification Model	Selection best hyperparameters with GridSearchCV (cv = 10)	Calculating model accuracy with method .score()
Logistic Regression	{'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}	0.8333333333333334
SVM	{'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}	0.8333333333333334
Decision Tree	{'criterion': 'gini', 'max_depth': 6, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 5, 'splitter': 'random'}	0.833333
KNN	{'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}	0.833333

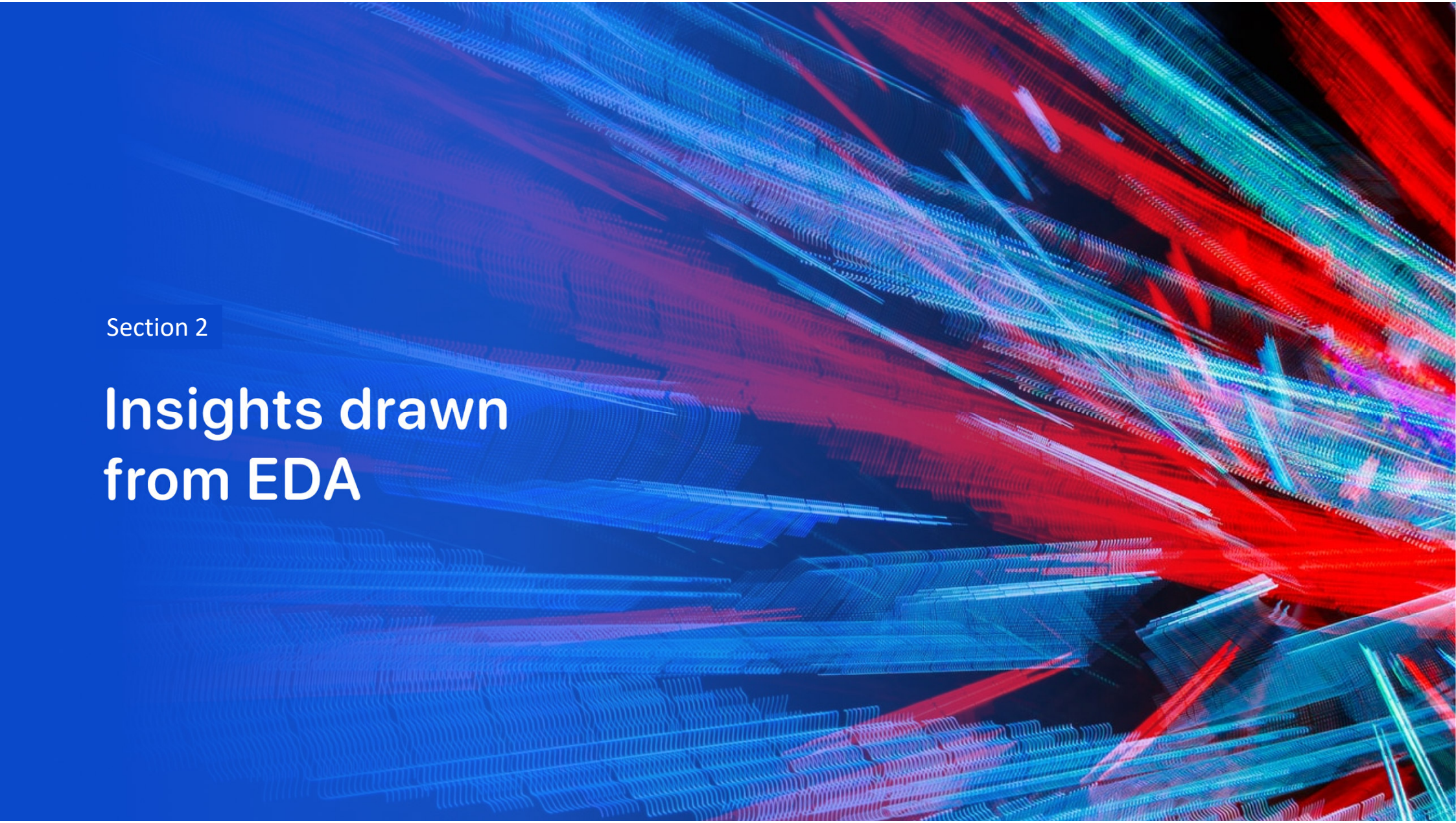
Methods comparison	
Logistic regresssion	0.846429
SVM	0.848214
Decision tree	0.873214
KNN	0.848214

Decision Tree

GitHub URL: Predictive Analysis

Results

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

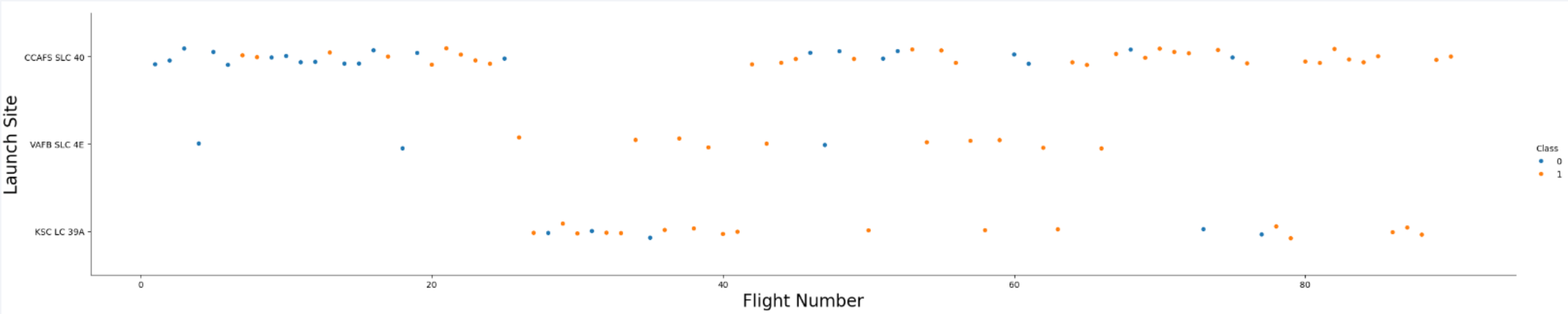


Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

Relationship between Flight Number and Launch Site

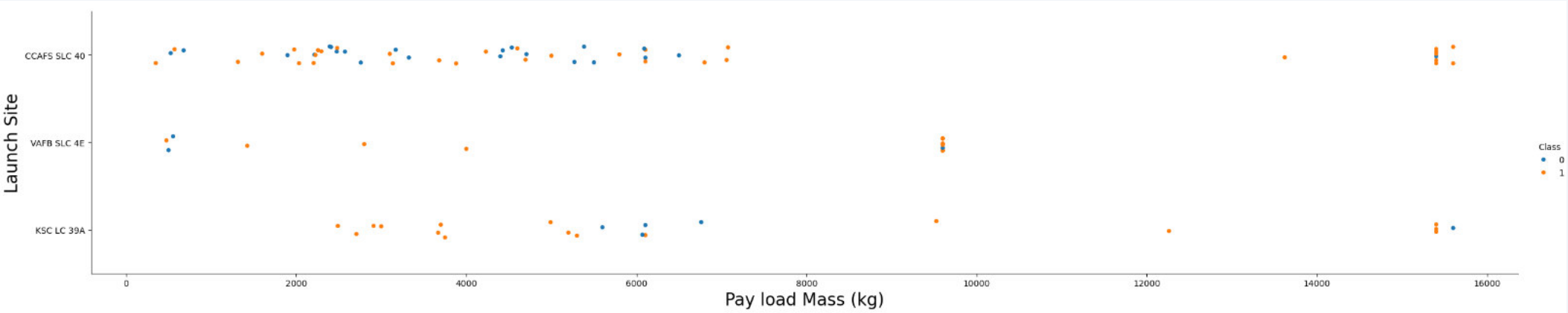


Explanation:

- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- Each new launch has a higher rate of success.

Payload vs. Launch Site

Relationship between Payload and Launch Site

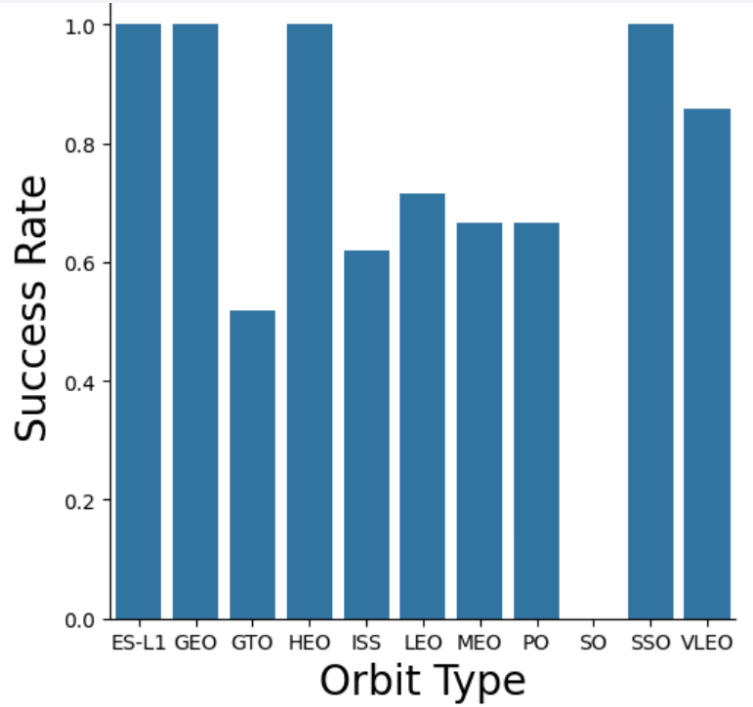


Explanation:

- CCAFS SLC 40 has higher success rate for payload mass over 7000 kg
- VAFB-SLC has no rockets launched for mass greater than 10000 kg.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg.

Success Rate vs. Orbit Type

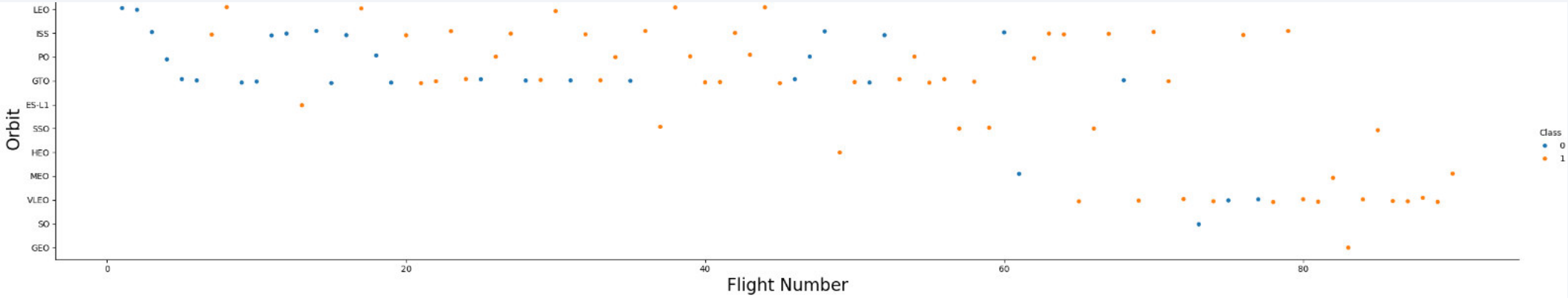
Relationship between success rate of each orbit type



Explanation:
ES-L1, GEO, HEO, SSO have the high success rate

Flight Number vs. Orbit Type

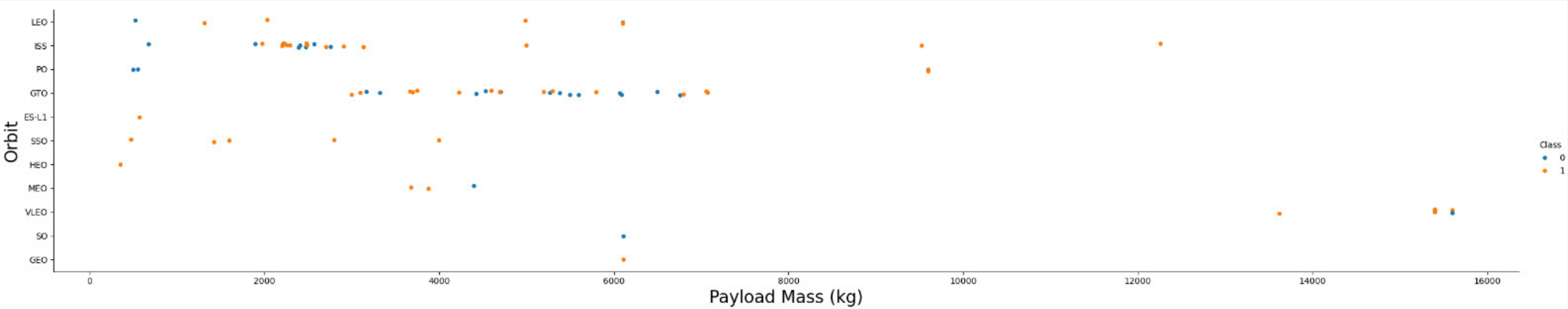
Relationship between FlightNumber and Orbit type



You should see that 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.

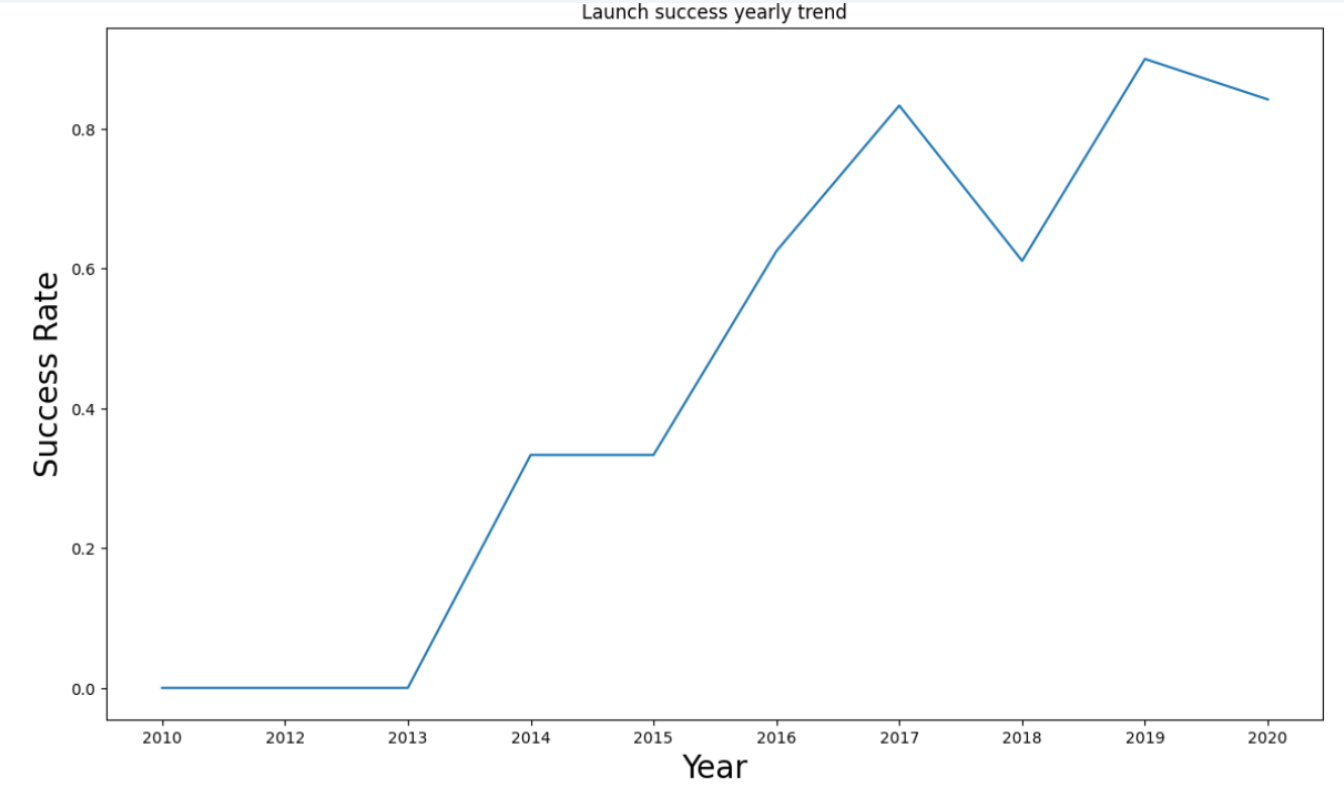
Payload vs. Orbit Type

Relationship between Payload and Orbit type



With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
However, for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend



The success rate since 2013 kept increasing till 2020.

All Launch Site Names

- Selection from the SpaceX Data Base of the launch sites names.
- The DISTINCT clause was used.

```
%sql select distinct launch_site from SPACEXTABLE;
```

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

Launch Site Names Begin with 'CCA'

- The select was used to list 5 records where launch sites begin with `CCA`

```
%sql select * from SPACEXTABLE where launch_site like 'CCA%' limit 5;
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
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 (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Displaying the total payload carried by boosters from NASA (CRS).
- The SUM() function was used.

```
%sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXTABLE where customer = 'NASA (CRS)';
```

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

total_payload_mass
45596

Average Payload Mass by F9 v1.1

- Displaying average payload mass carried by booster version F9 v1.1.
- The AVG() function was used.

```
%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXTABLE where booster_version like '%F9 v1.1%';
```

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

average_payload_mass
2534.6666666666665

First Successful Ground Landing Date

- List the date when the first succesful landing outcome in ground pad was acheived.
- The MIN() function was used.

```
%sql select min(date) as first_successful_landing from SPACEXTABLE where landing_outcome = 'Success (ground pad)';
```

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

first_successful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.
- The WHERE and BETWEEN clause was used.

```
%%sql select booster_version
        from SPACEXTABLE
       where landing_outcome = 'Success (drone ship)'
        and payload_mass__kg_ between 4000 and 6000;
```

```
 * 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

- Calculate the total number of successful and failure mission outcomes.
- The COUNT() function and GROUP BY clause was used.

```
%%sql select mission_outcome, count(*) as total_number
        from SPACEXTABLE
    group by mission_outcome;
```

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

Mission_Outcome	total_number
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass.
- A sub select was used.

```
%sql select booster_version from SPACEXTABLE where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTABLE);
```

```
 * 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 failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015.

```
%%sql select substr(Date, 6,2) as month, date, booster_version, launch_site, landing_outcome
        from SPACEXTABLE
       where landing_outcome = 'Failure (drone ship)'
         and substr(Date,0,5)= '2015';
```

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

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

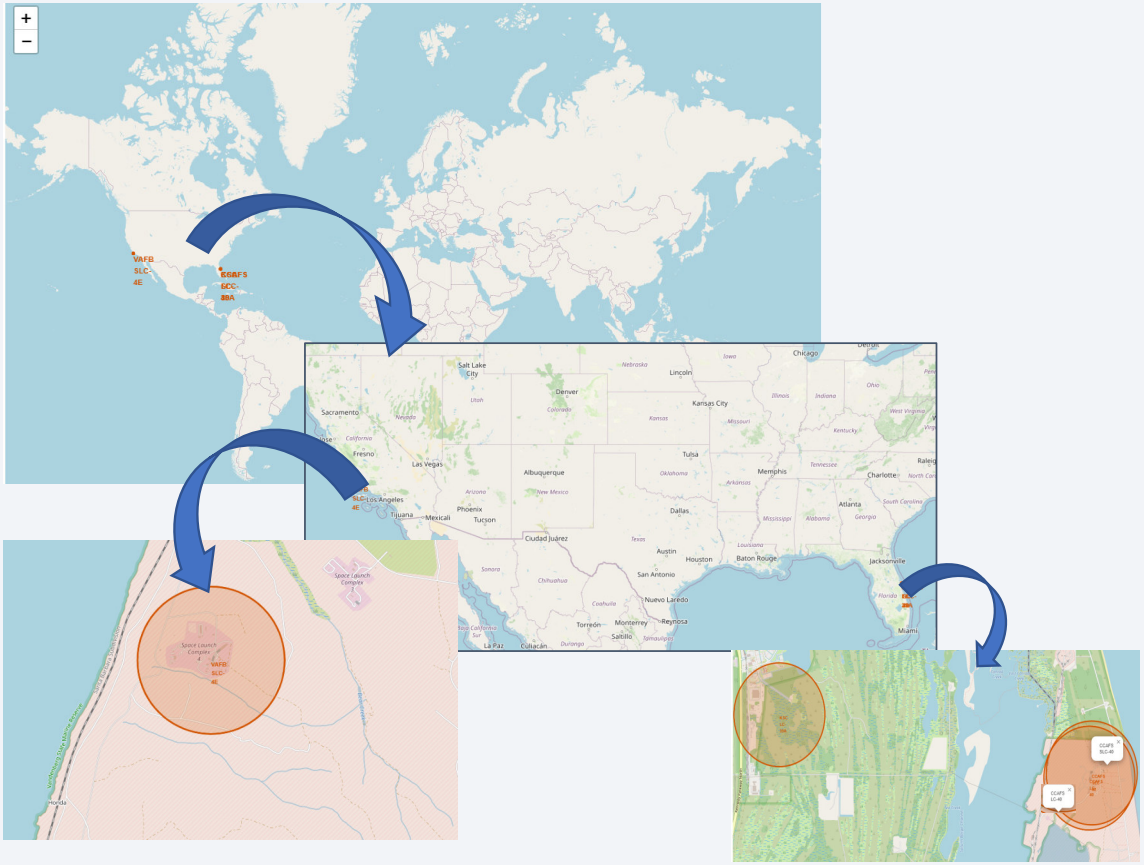
- Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.

```
%%sql select landing_outcome, count(*) as count_outcomes
        from SPACEXTABLE
       where date between '2010-06-04' and '2017-03-20'
    group by landing_outcome
    order by count_outcomes desc;
```

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


Section 3
Launch Sites
Proximities Analysis

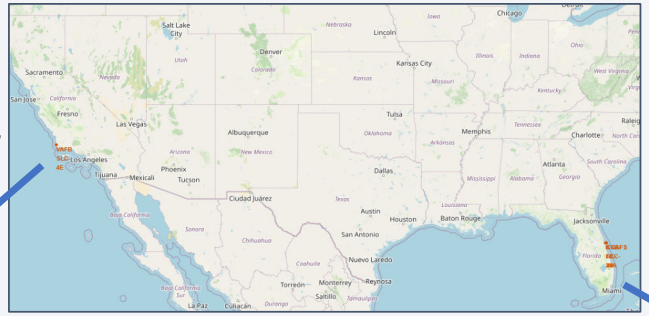
SpaceX launch sites on map

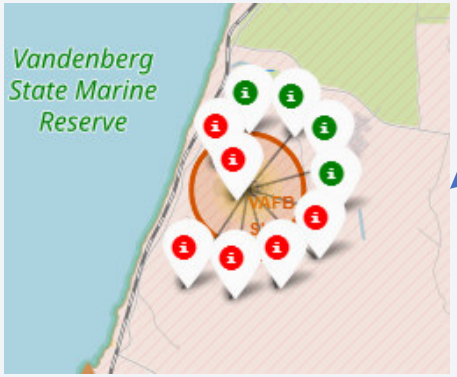


Findings:

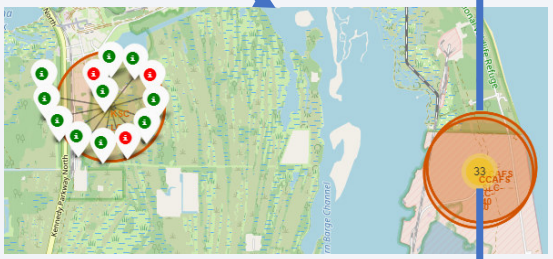
- SpaceX launch sites are in the USA (Florida and California)
- The launch sites are located near the coast to minimize the risk of debris and explosions hitting people in the event of an accident.
- The launch sites are in proximity to the Equator line where the land is moving faster than any other place on the surface of the Earth. This speed will help the spacecraft keep up a good enough speed to stay in orbit.

Success/failed launches for each site on the map

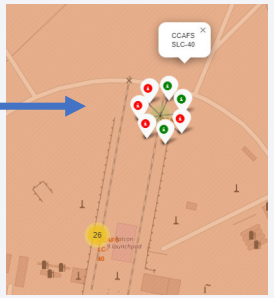




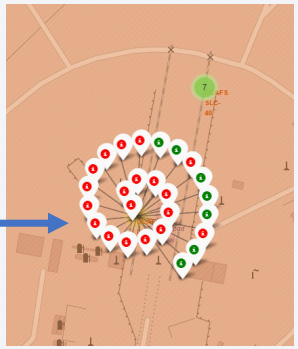
VAFB SLC-4E



KSC LC-39A



CCAFS SLC-40



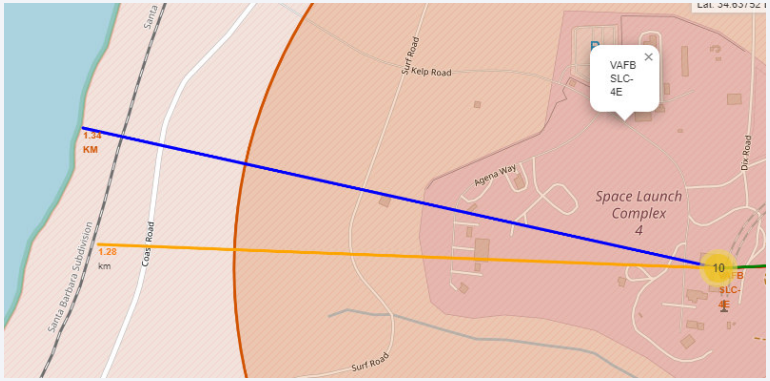
CCAFS LC-40

Explanation:

- From the colour-labeled markers we should be able to easily identify which launch sites have relatively high success rates.
  - Green Marker = Successful Launch
  - Red Marker = Failed Launch
- Launch Site KSC LC-39A has a very high Success Rate.

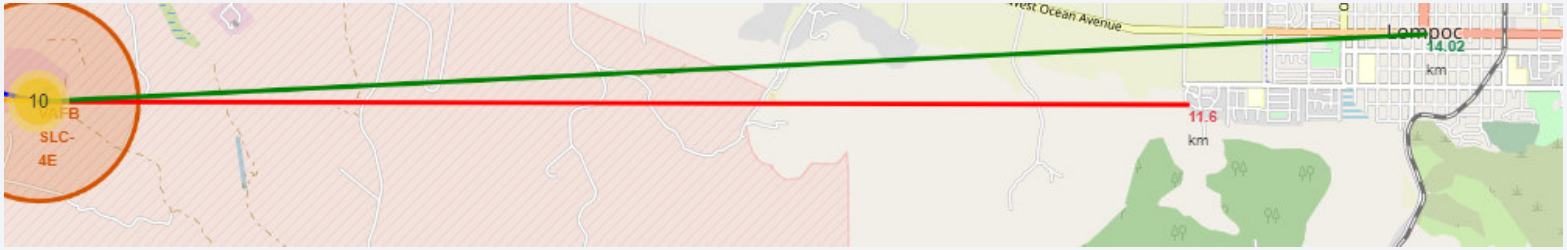
Distances between VAFB SLC-4E launch site to its proximities

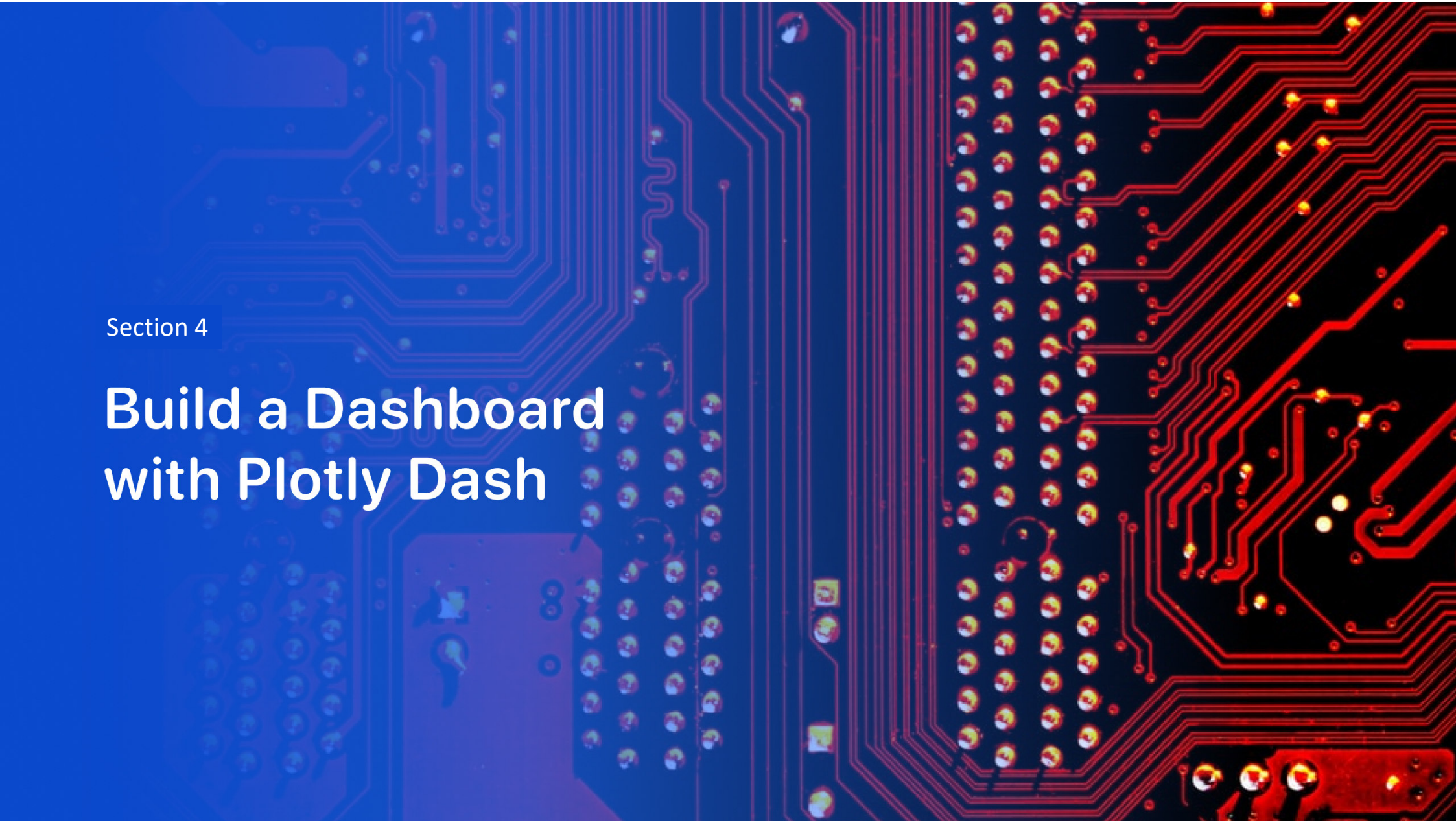
From the visual analysis of the launch site VAFB SLC-4E we can see that it is:



- close to coastline (1.34 Km)
- close to railway (1.28 Km)

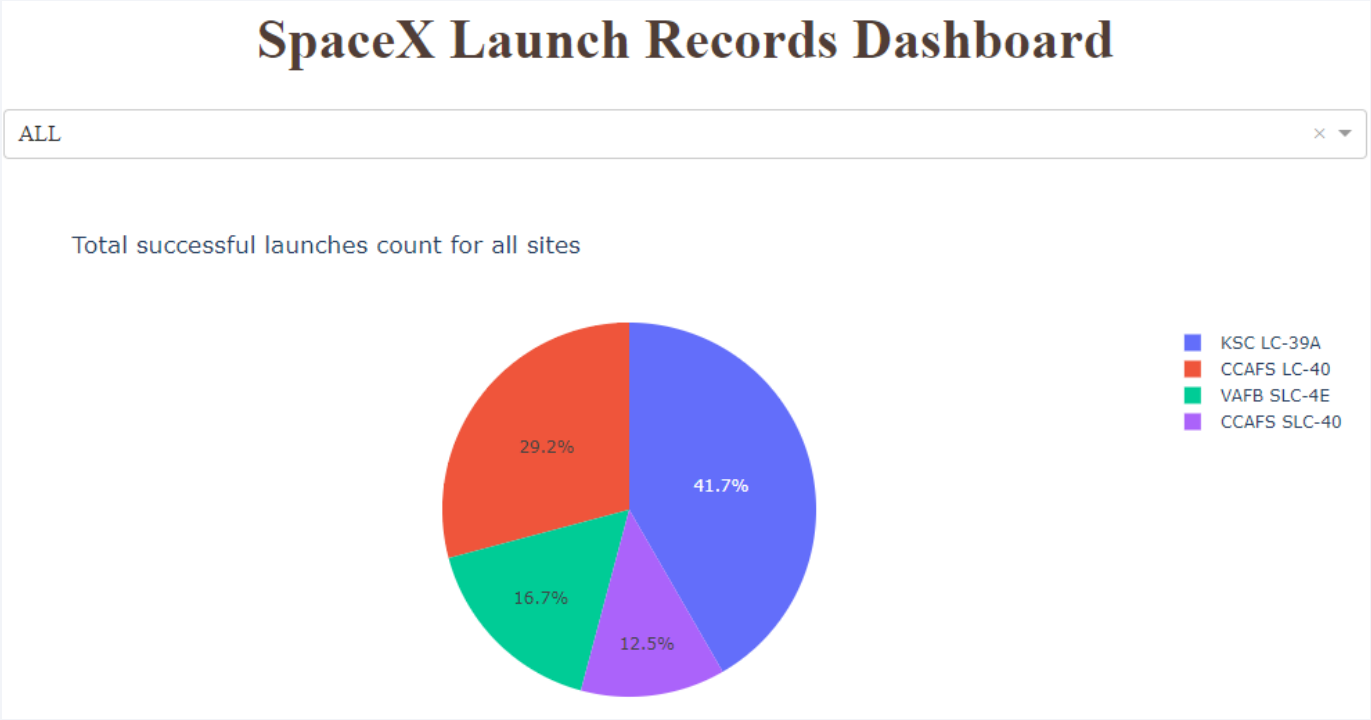
- relative close to its closest city Lompoc (11.6 km)
- relative close to highway (14.02 km)




Section 4
Build a Dashboard with Plotly Dash

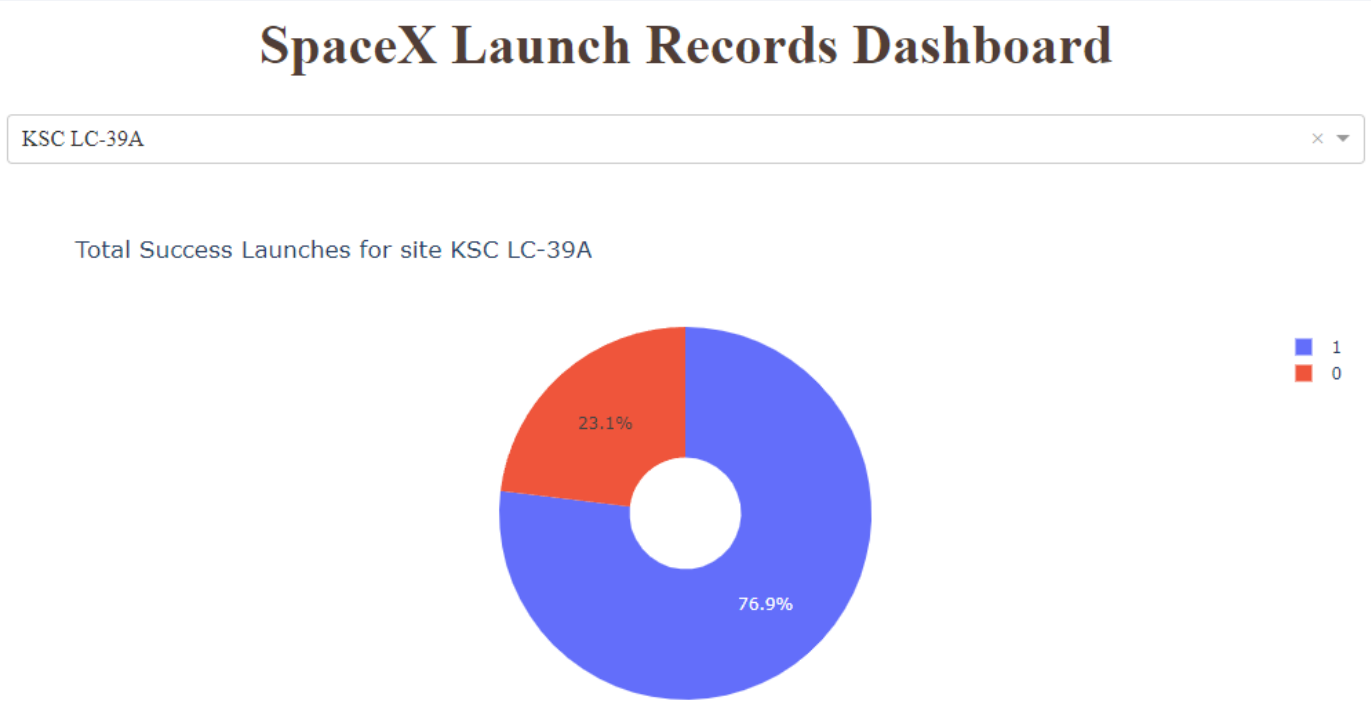
Launch success count for all sites

The chart shows the percentage of success from all the sites, and KSC LC-39A has the most successful launches.



Statistics for KSC LC-39A

Detail values for KSC LC-39A site.
KSC LC-39A has the highest launch success rate (76.9%).



Payload Mass vs. Launch Outcome for all sites

By selecting payload ranges, it is highlighted that within the range of 5000 to 10000 kg, the highest success rate of launches is recorded.

Class 0 represents failed launches while class 1 represents successful launches.

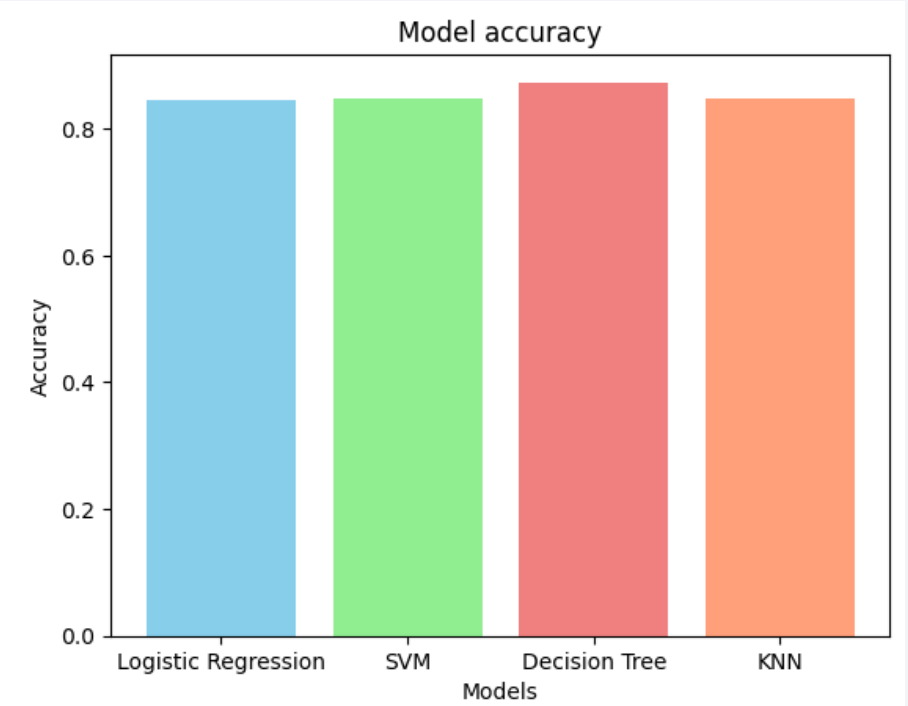
Section 5

Predictive Analysis (Classification)

Classification Accuracy

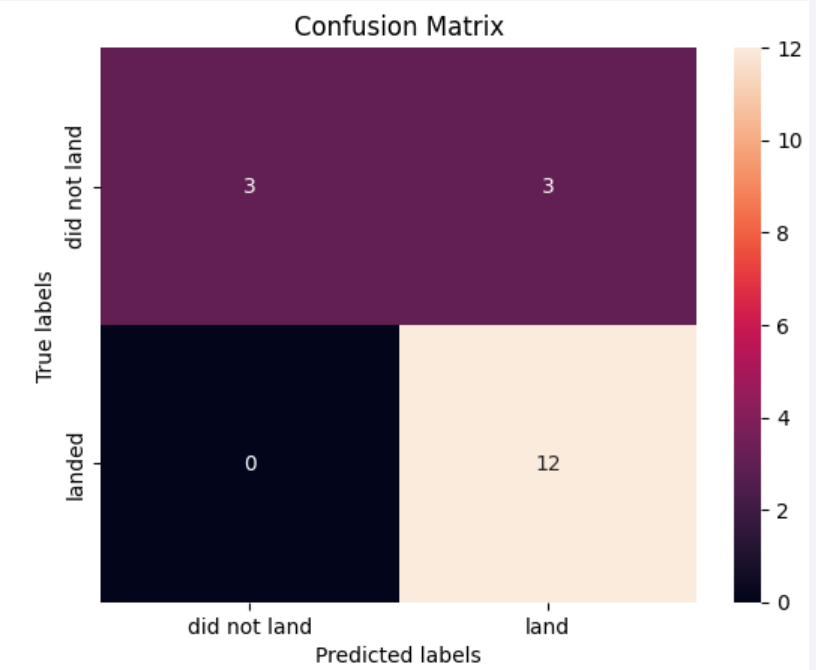
- All built classification models display a very similar level of accuracy.
- Trying the models on whole Dataset reveals that the decision tree achieves the highest accuracy.

	Scores
Logistic regresssion	0.846429
SVM	0.848214
Decision tree	0.873214
KNN	0.848214



Confusion Matrix

The confusion matrix for the decision tree classifier shows that the classifier can distinguish between the different classes. The major problem is the false positives.



Conclusions

In the exploratory data analysis phase, it was observed that certain features may correlate with the outcome of launches.
For example:

- The success rate of launches increases over the years. Launch success rate started to increase in 2013.
- KSC LC-39A has the highest launch success rate (76.9%).
- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.

Four different machine learning algorithms were tested to identify patterns from the launch data that allow us to predict the outcome of future launches. Among these, **the decision tree classification model emerged as the most effective**.

Appendix

IBM Applied Data Science Capstone

Project resources

SpaceX Falcon-9 page

Thank you!