IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

TSHIDAHO OSBORN MUKWEVHO
04.11.2024

Outline

EXECUTIVE SUMMARY

INTRODUCTION

METHODOLOGY

RESULTS

CONCLUSION

APPENDIX

Executive Summary

Summary Of Methodologies

- Data Collection
- Data Wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an Interactive Map with Folium
- Building a Dashboard with Plotly Dash
- Predictive Analysis(Classification)

Summary OF All Results

- Exploratory Data Analysis Results
- Interactive Analytics Demo in Screenshots
- Predictive Analytics Results

Introduction

- Project background and context
- SpaceX is the most successful company of the commercial space age, making space travel affordable. The company advertises Falcon9 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 we can determine if the first stage will land, we can determine the cost of a launch. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.
- Problems you want to find answers
- How variables such as payload mass, launch site, number of flights and orbits affect the success of the first stage landing?
- Does the rate of a successful landings increase over the years?
- What is the best algorithm that can be used for binary classification in this case?

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology:

Using SpaceX Rest API

Using Web Scrapping from Wikipedia

Perform data wrangling

Filtering data

Dealing with missing values

Using One Hot Encoding to prepare the data to a binary classification

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

Data collection process involved a combination of API request from SpaceX Rest API and Webscrapping data from a table in SpaceX's Wikipedia.

-Used both of data collection methods to get a complete information about the launches for a more detailed analysis.

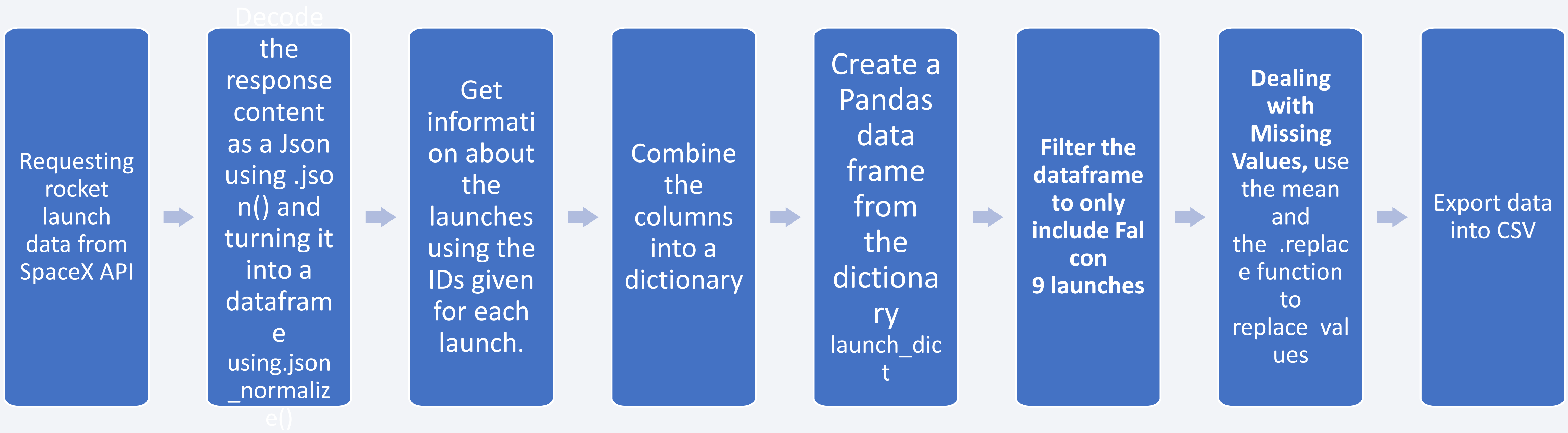
Data Columns are obtained by using SpaceX REST API

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, RescueCourt, Serrial, Longitude, Latitude.

Data columns are obtained by using Wikipedia Web Scrapping

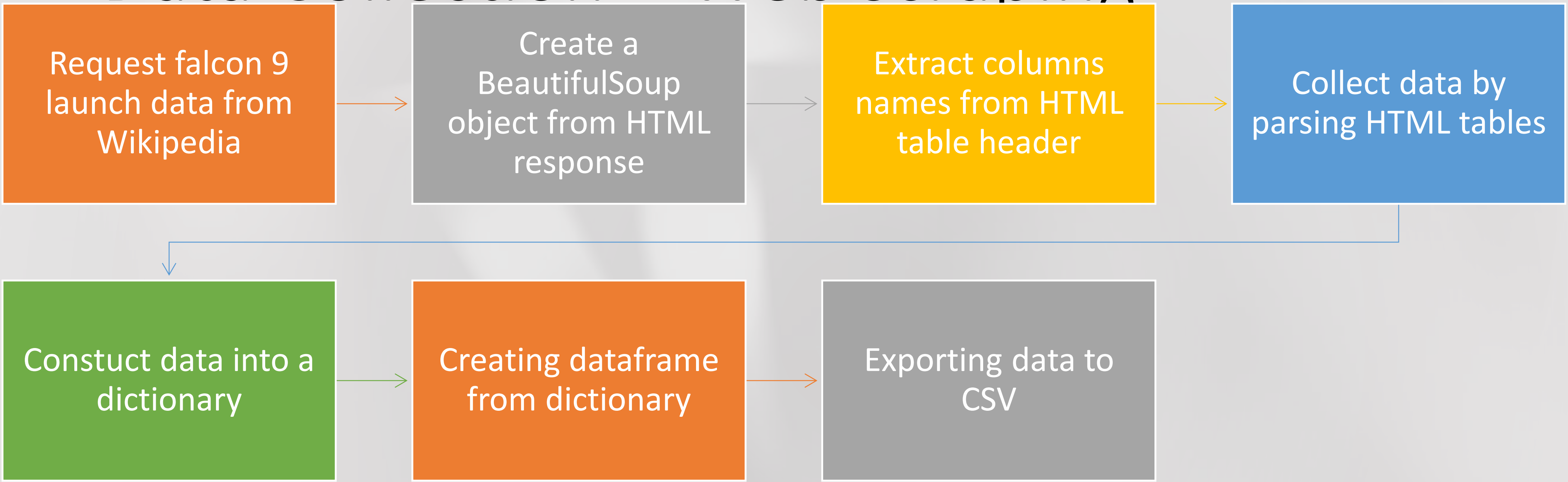
Flight NO., Launch site, Payload, PyaloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster Landing, date, Time

Data Collection – SpaceX API



[Data-science-capstone/jupyter-labs-spacex-data-collection-api.ipynb at main · osborn-engine/Data-science-capstone]

Data Collection – Web scraping



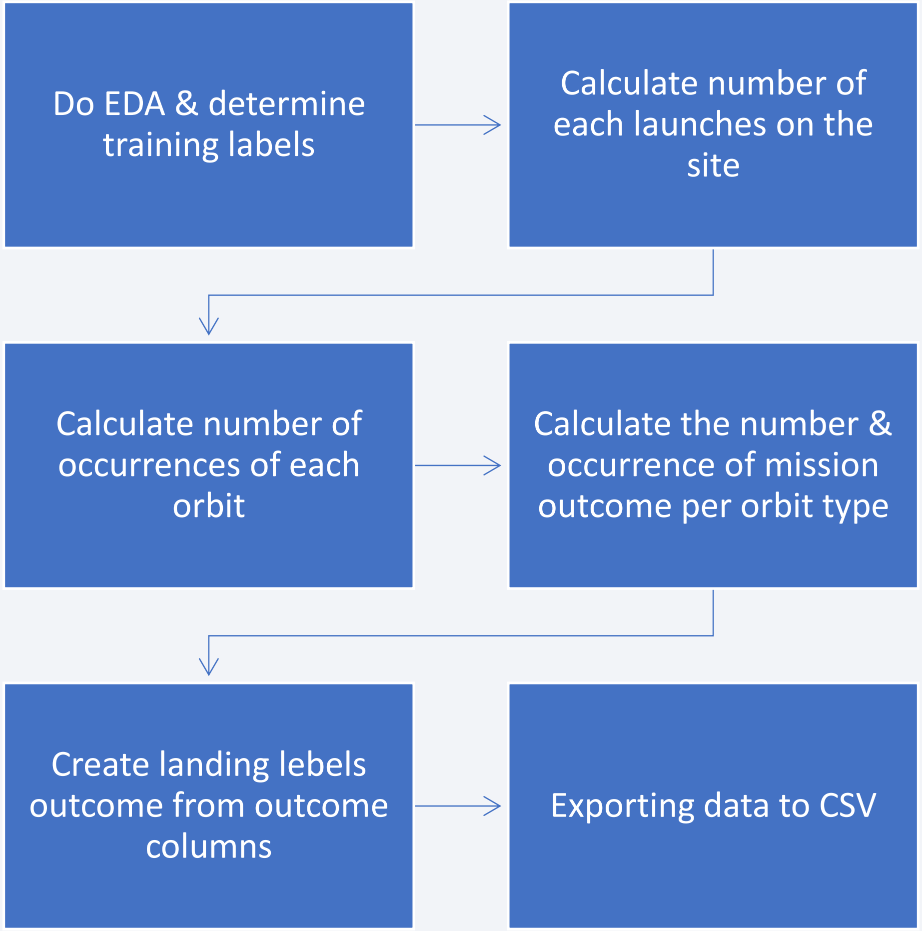
[Data-science-capstone/jupyter-labs-webscraping.ipynb at main · osborn-engine/Data-science-capstone]

Data Wrangling

In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad. True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship.

We will mainly convert those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.

Data-science-capstone/labs-jupyter-spacex-Data wrangling.ipynb at main · osborn-engine/Data-science-capstone



EDA with Data Visualization

- **FlightNumber vs. PayloadMass**, relationship between **Flight Number and Launch Site**, relationship between **Payload Mass and Launch Site**,

- relationship between success rate of each orbit type, relationship between **FlightNumber and Orbit type**, relationship between **Payload Mass and Orbit type,** Visualize the **launch success yearly trend.**

- Scatter plot shows the relationship between variables, if a relationship exists, they could be used in machine learning model

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

- Line charts show trends in data overtime(time series)

- Data-science-capstone/edadataviz (1).ipynb at main · osborn-engine/Data-science-capstone

EDA with SQL

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first succesful landing outcome in ground pad was acheived.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
- 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.
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

[Data-science-capstone/jupyter-labs-eda-sql-coursera_sqllite.ipynb at main · osborn-engine/Data-science-capstone]()

Build an Interactive Map with Folium

- *Mark all launch sites on a map*
- To add each site's location on a map using site's latitude and longitude coordinates
  - To add a highlighted circle area with a text label on a specific coordinate

- Mark the successful/failed launches for each site on the map
  - to enhance the map by adding the launch outcomes for each site, and see which sites have high success rate

- Calculate the distances between a launch site to its proximities
  - Add a mouse position on the map to get coordinates for a mouse over a point on the map

[Data-science-capstone/lab_jupyter_launch_site_location (1).ipynb at main · osborn-engine/Data-science-capstone]

Build a Dashboard with Plotly Dash

- Launch sites Dropdown list
  - Add a dropdown list to enable Launch Site selection

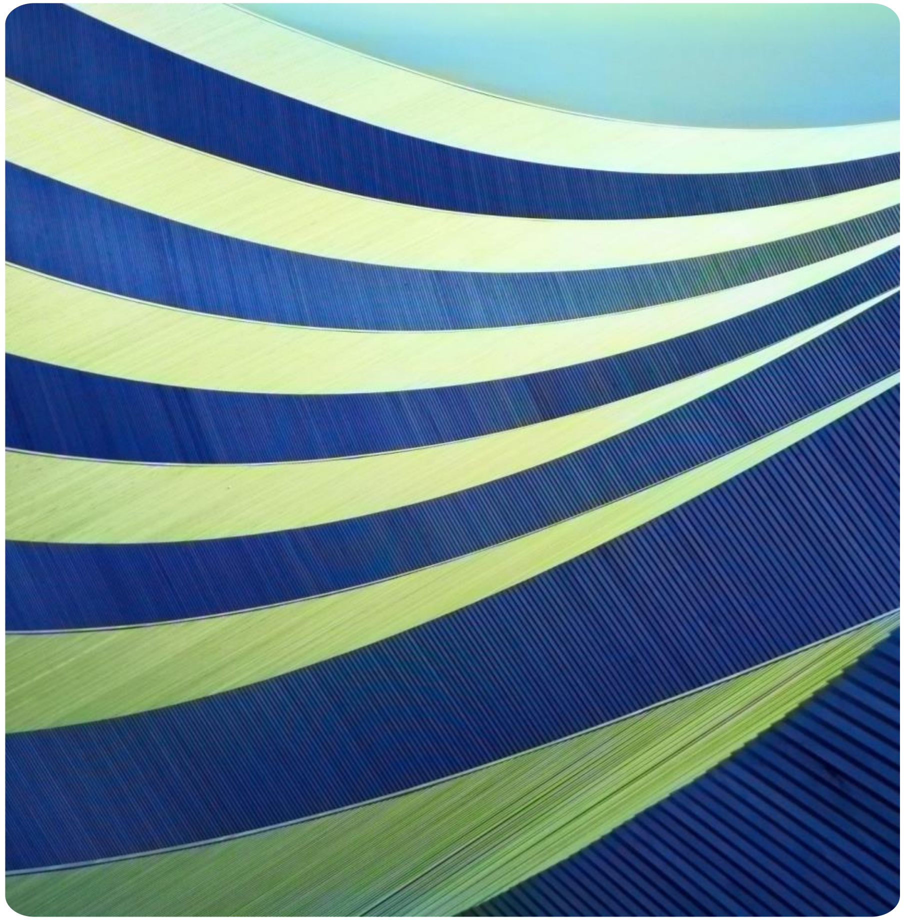
- Pie chart showing successful launches
- Add a pie chart to show the total successful launches count for all sites. If a specific launch site was selected, show the Success vs. Failed counts for the site

- Payload Range
  - Add a slider to select payload range

- Scatter Chart
- Add a callback function for `site-dropdown` and `payload-slider` as inputs, `success-payload-scatter-chart` as output

Data-science-capstone/spacex_dash_app.py at main · osborn-engine/Data-science-capstone

Predictive Analysis (Classification)



Data-science-capstone/SpaceX_Machine Learning Prediction_Part_5 (1).ipynb at main · osborn-engine/Data-science-capstone

Create a NumPy array from the column in data → Standardize the data in then reassign it to the variable X → Use the function train_test_split to split the data X and Y into training and test data.

→ Create a logistic regression object then create a GridSearchCV object with cv = 10. → Calculate the accuracy on the test data using the method → Finding best performing method for Jaccard and f1 metrics scores

→ Examining confusion metrics for all methods → Calculating the accuracy of the tests data using score() for all methods → Applying GridsearchCV for logreg, SVM,Desion trees and KNN models

Results

EXPLORATORY DATA ANALYSIS RESULTS

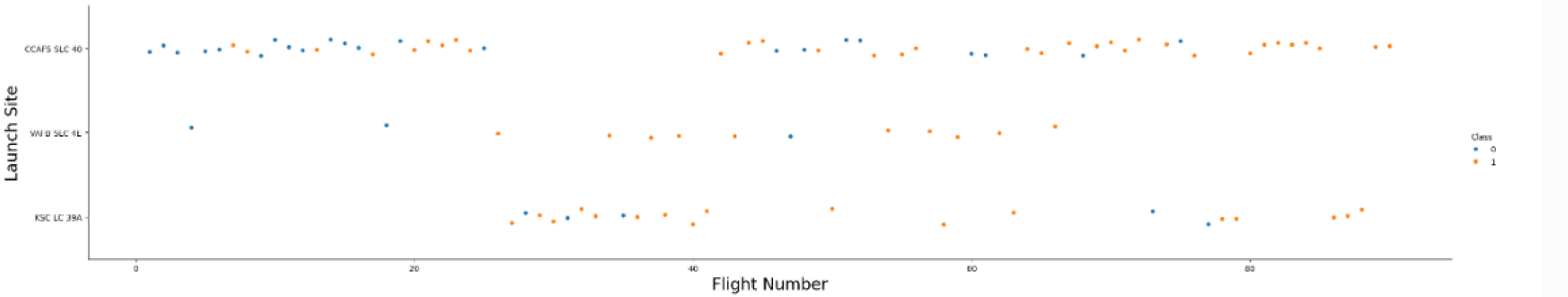
INTERACTIVE ANALYTICS DEMO IN SCREENSHOTS

PREDICTIVE ANALYSIS RESULTS

Section 2

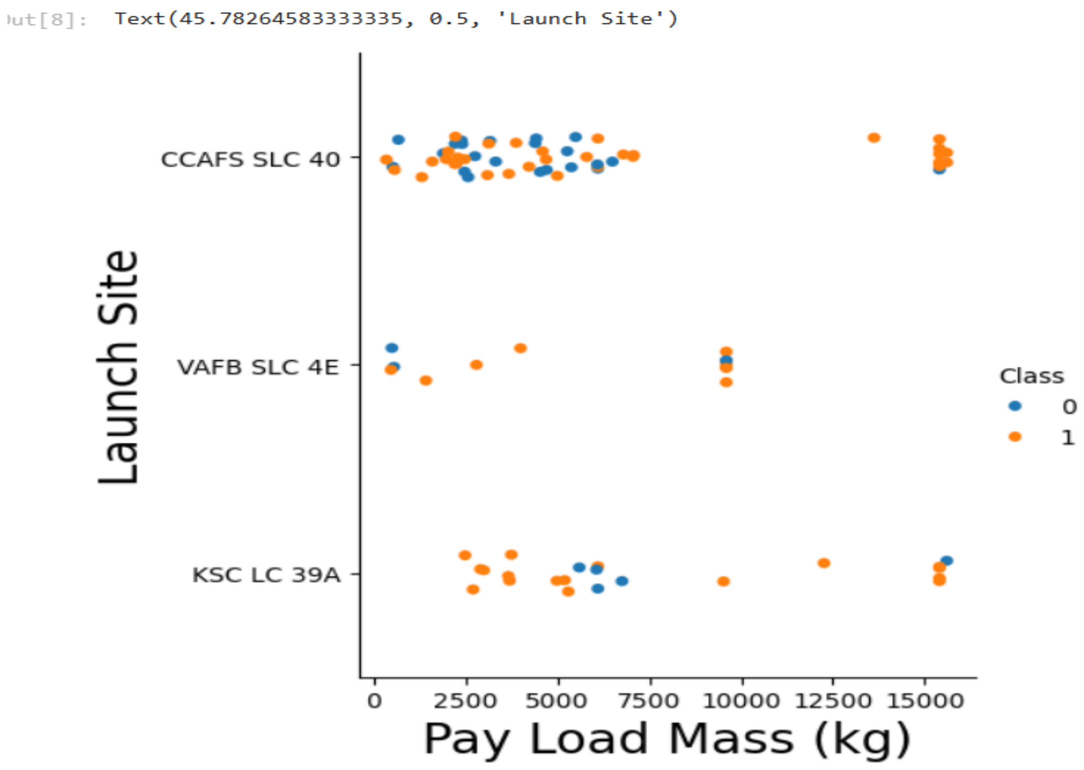
Insights drawn from EDA

Flight Number vs. Launch Site



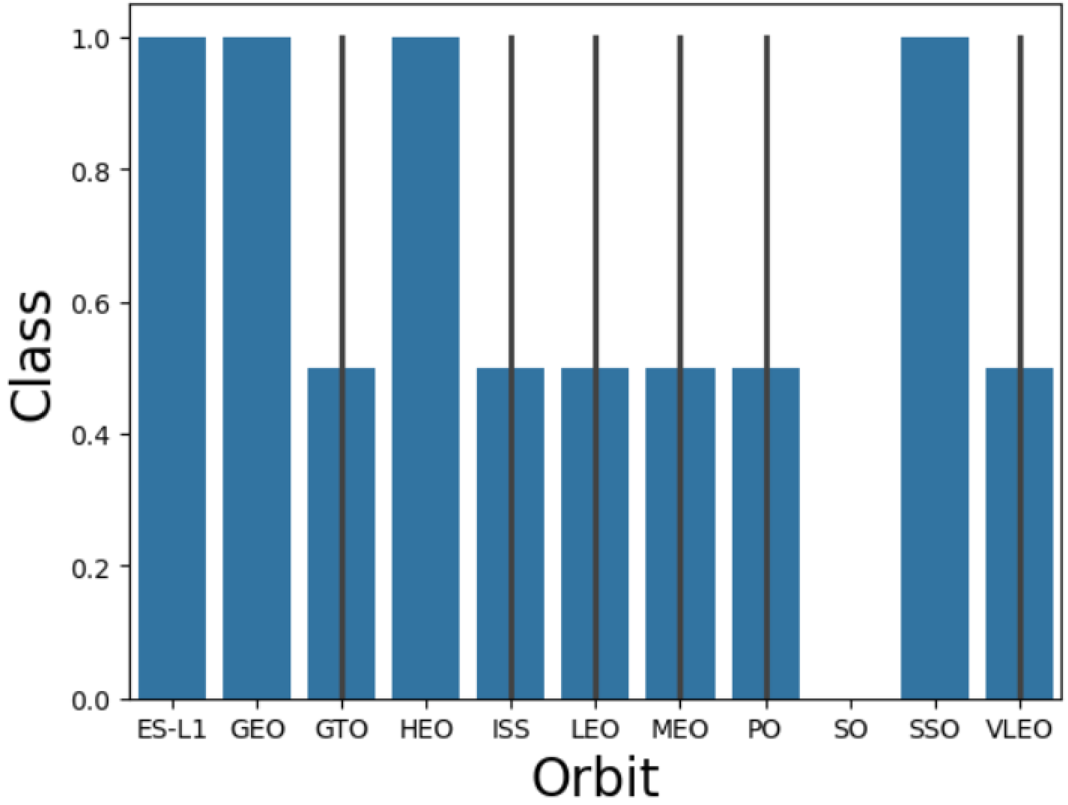
- The earliest flights all failed while the latest flights all succeeded
- The CCAFS SLC 40 launch site has about a half of all launches
- VAFB SLC 4E and KSC LC 39A have higher success rates
- It can be assumed that each new launch has a higher rate of success

Payload vs. Launch Site



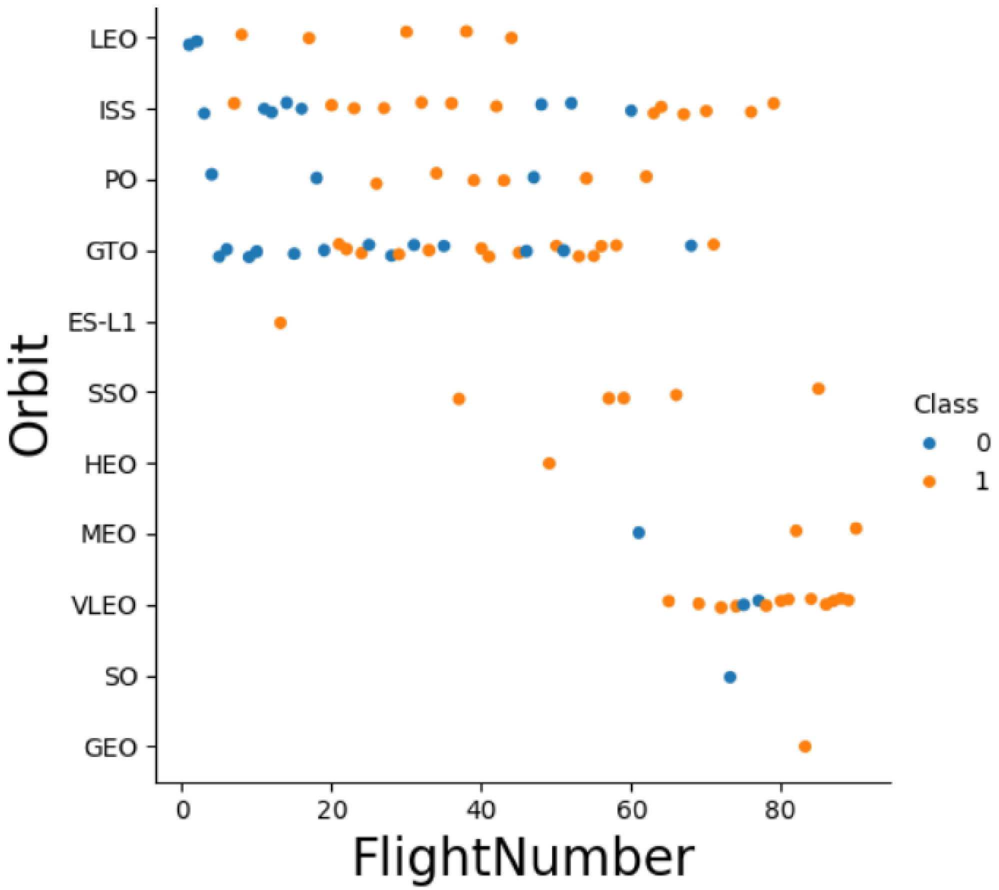
- For every launch site the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

Success Rate vs. Orbit Type



- Orbits with 100% success rate: - ES-L1, GEO, HEO, SSO
- Orbits with 0% success rate: - SO
- Orbits with success rate between 50% and 85%:

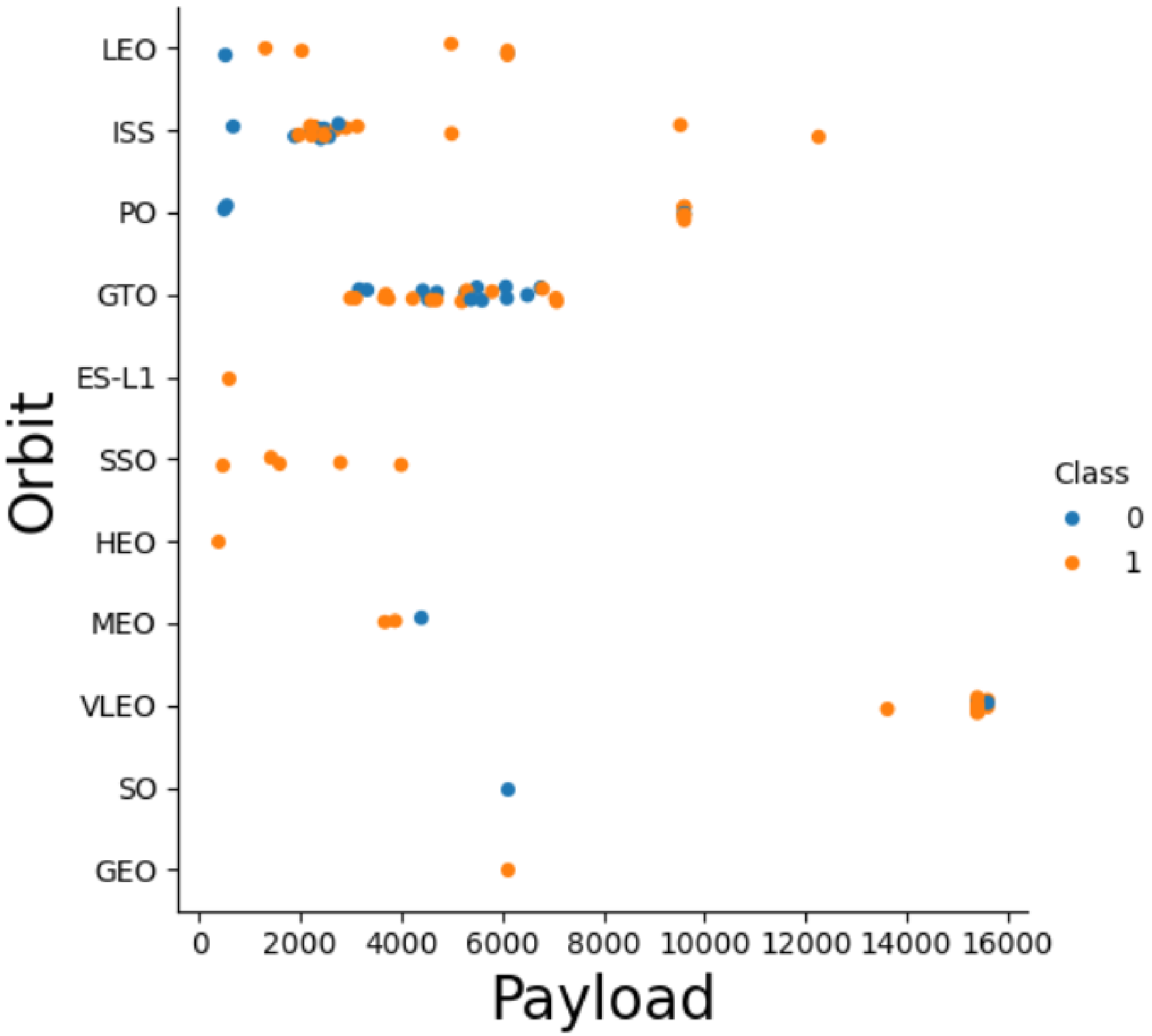
Flight Number vs. Orbit Type



- 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

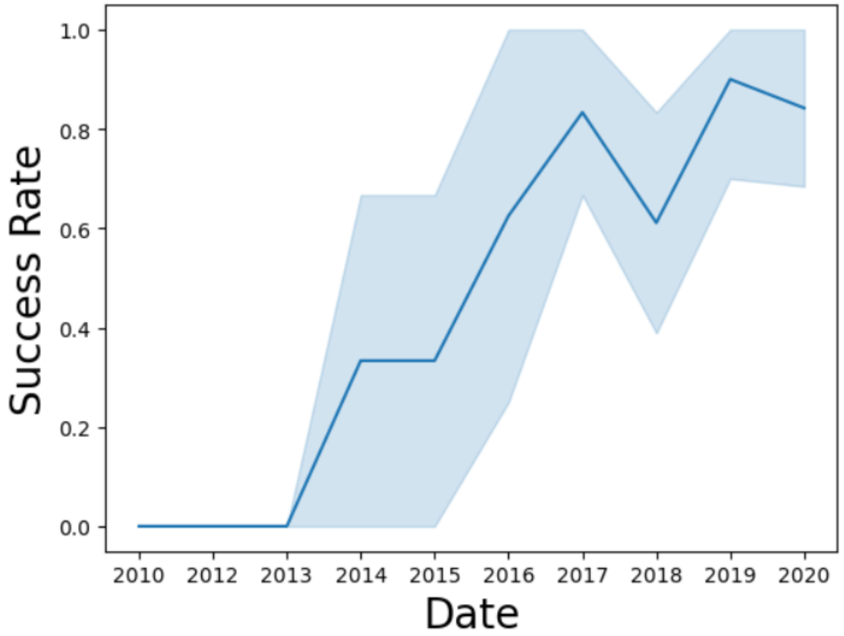
Payload vs. Orbit Type

- Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits



The success rate since 2013 kept increasing till 2020.

Launch Success Yearly Trend



Out[10]:

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

Displaying the names of the unique launch sites in the space mission.

All Launch Site Names

Launch Site Names Begin with 'CCA'

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

In [11]:
```
%sql SELECT* FROM SPACEXTABLE WHERE Launch_Site like 'CCA%' Limit 5;
```

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

Out[11]:

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

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

In [12]:
```
%sql SELECT AVG("PAYLOAD_MASS__KG_") AS TOTAL_PAYLOAD_MASS FROM SPACEXTABLE WHERE "Customer" ='NASA (CRS)';
```

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

Out[12]:

TOTAL_PAYLOAD_MASS
2279.8

Total Payload Mass

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

Task 4

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

```
In [13]: %sql SELECT AVG("PAYLOAD_MASS__KG_") AS AVERAGE_PAYLOAD FROM SPACEXTABLE WHERE "Booster_Version" = 'F9 v1.1';
```

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

Out[13]:

AVERAGE_PAYLOAD
2928.4

Average Payload Mass by F9 v1.1

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

In [14]:
```
%sql SELECT MIN("Date") AS First_Successful_Landing_Date FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (ground pad)'
```

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

Out[14]:

First_Successful_Landing_Date
2015-12-22

First Successful Ground Landing Date

Listing the date when the first successful landing outcome in ground pad was achieved

Successful Drone Ship Landing with Payload between 4000 and 6000

```
In [15]: %sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)' AND "PAYLOAD_MASS__KG_" > 40
```

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

Out[15]:

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

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

In [16]:
```
%sql SELECT "Mission_Outcome", COUNT(*) AS Total_Count FROM SPACEXTABLE GROUP BY "Mission_
```

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

Out[16]:

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

Total Number of Successful and Failure Mission Outcomes

Listing the total number of successful and failure mission outcomes

In [17]:
```
%sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX ("PAYLOAD_MASS__KG_") FROM SPACEXTABl
```

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

Out[17]:

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

Listing the names of the booster versions which have carried the maximum payload mass

Boosters Carried Maximum Payload

2015 Launch Records

Out[34]:

Month_Name	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40

Out[34]:

Month_Name	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015

Out[32]:

Landing_Outcome	Outcome_Count
Failure (drone ship)	5
Success (ground pad)	3

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

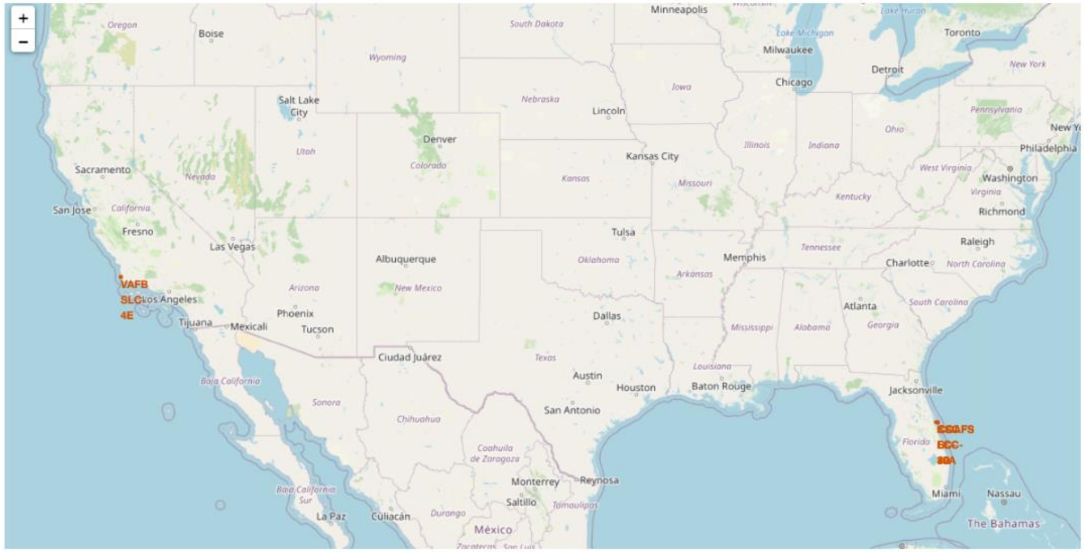
- Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20 in descending order

Section 3

Launch Sites Proximities Analysis

All Launch Sites Location Markers in a Global Map

• Most of Launch sites are in proximity to the Equator line. The land is moving faster at the equator than any other place on the surface of the Earth. Anything on the surface of the Earth at the equator is already moving at 1670 km/hour. If a ship is launched from the equator it goes up into space, and it is also moving around the Earth at the same speed it was moving before launching. This is because of inertia. This speed will help the spacecraft keep up a good enough speed to stay in orbit. • All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimizes the risk of having any debris dropping or exploding near people

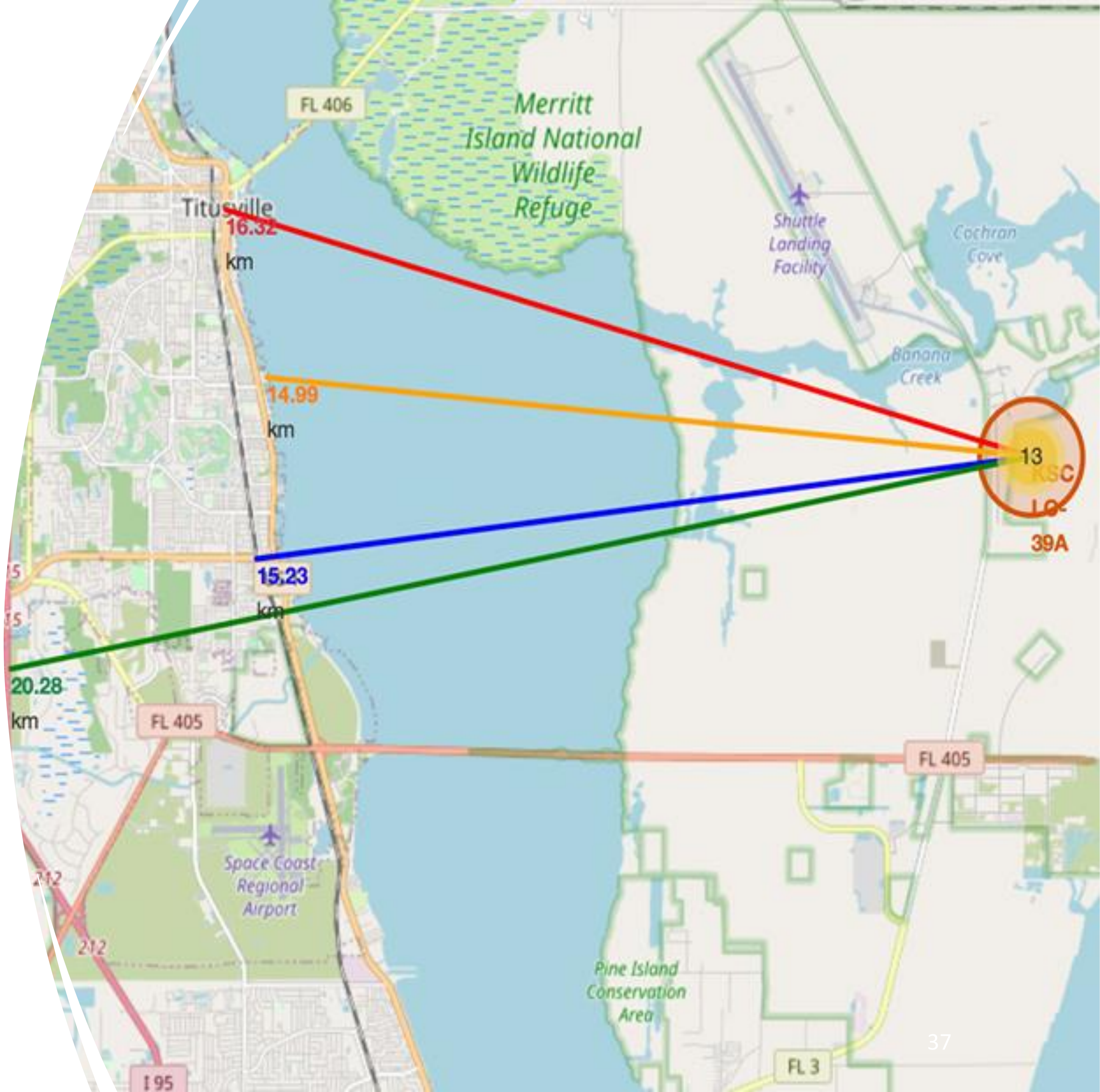


Colored Labeled Launch on the Map

- From the color-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 CCAFS SLC-40 has a high Success Rate.

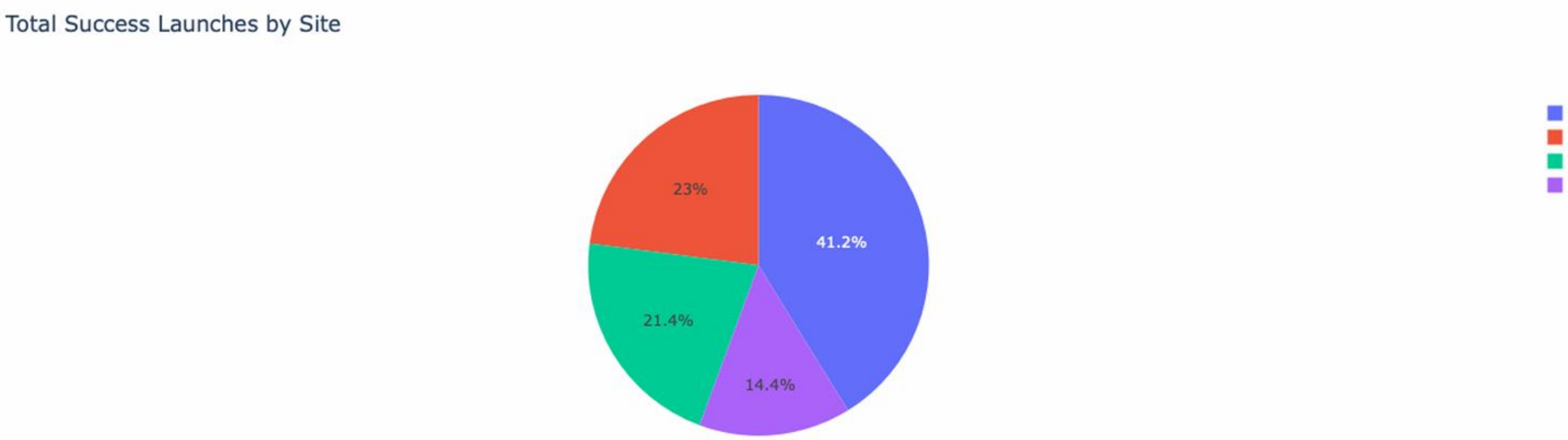
Distance between the launch site and its proximities

- From the visual analysis of the launch site KSC LC-39A we can clearly see that it is: - relative close to railway (15.23 km) - relative close to highway (20.28 km) - relative close to coastline (14.99 km) • Also the launch site KSC LC-39A is relative close to its closest city Titusville (16.32 km). • Failed rocket with its high speed can cover distances like 15-20 km in few seconds. It could be potentially dangerous to populated areas.



Section 4

Build a Dashboard with Plotly Dash

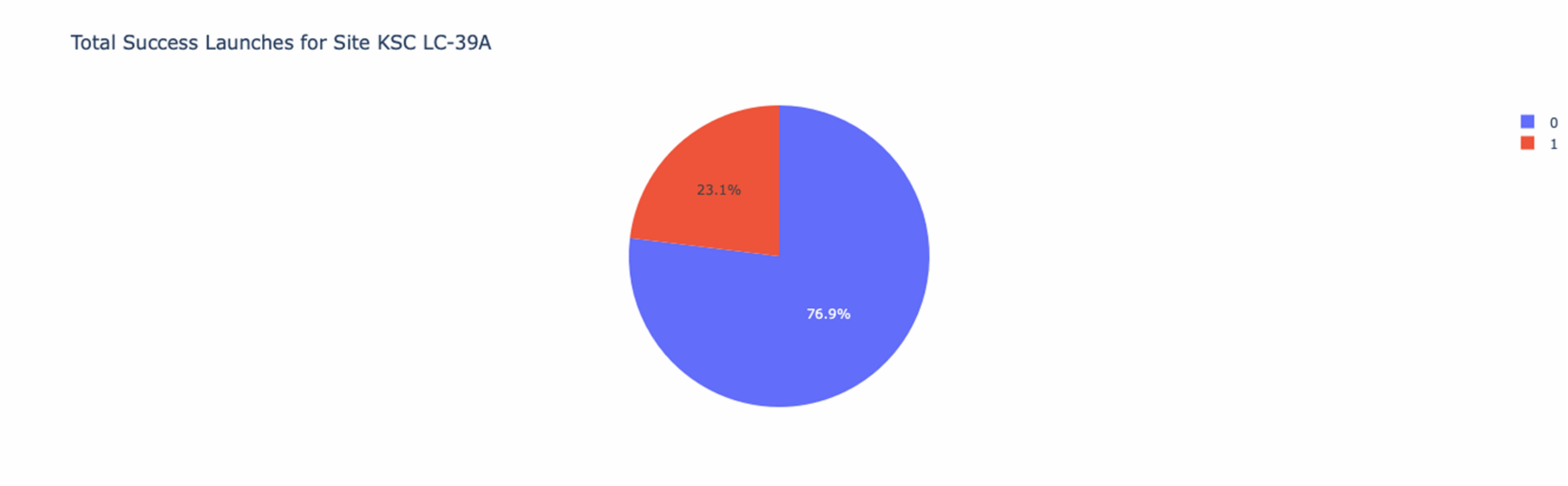


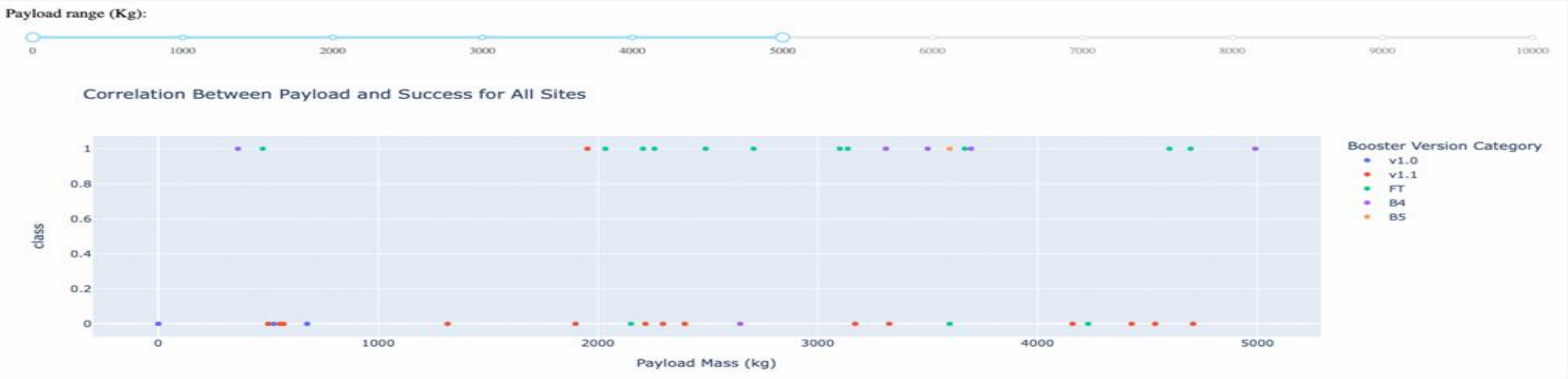
Launch Success Count For All Sites

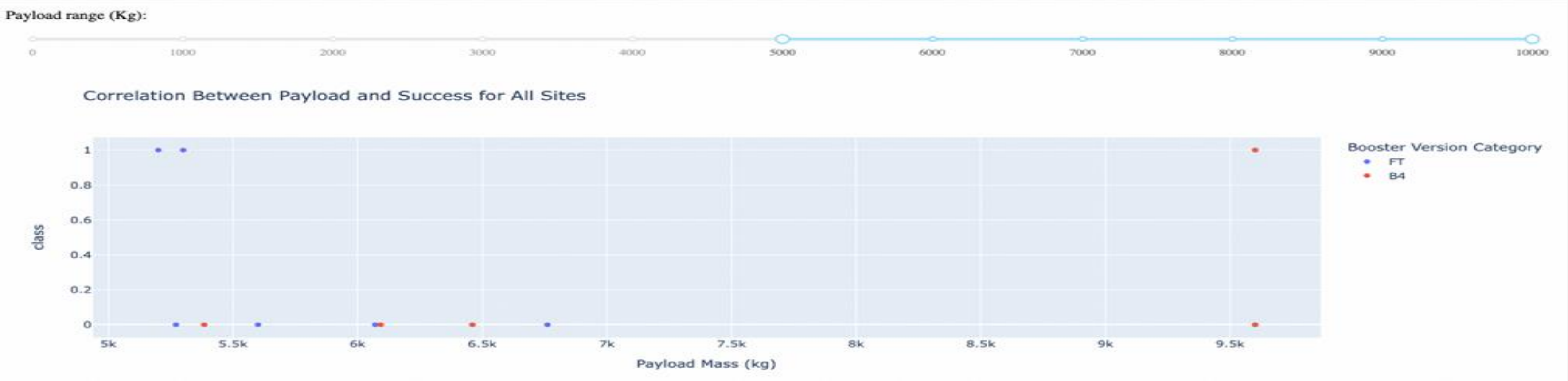
The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches

Highest Success Ratio Launch, by the Launch Site

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







PayLoad Mass vs. Launch Outcome For All Sites

The charts show that payloads between 2000 and 5500 kg have the highest success rate.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Scores and Accuracy of the Test Set

[57]:

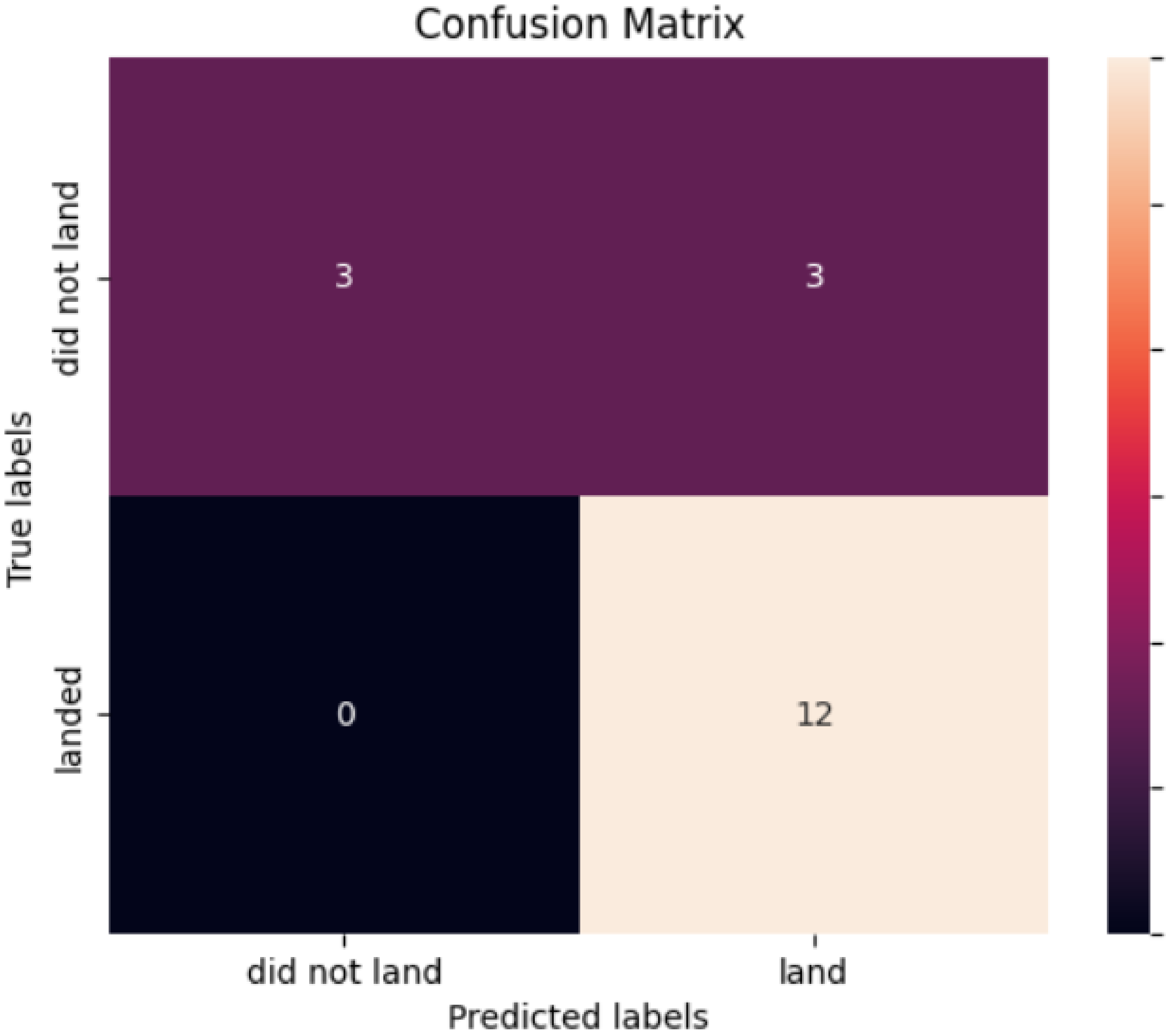
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

Scores and Accuracy of the Entire Data Set

[58]:

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.819444	0.819444
F1_Score	0.909091	0.916031	0.900763	0.900763
Accuracy	0.866667	0.877778	0.855556	0.855556

Based on the scores of the Test Set, we can not confirm which method performs best. • Same Test Set scores may be due to the small test sample size (18 samples). Therefore, we tested all methods based on the whole Dataset. • The scores of the whole Dataset confirm that the best model is the Decision Tree



Confusion Matrix

- Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives

Conclusions

ALGORITHM PERFORMANCE: THE DECISION TREE MODEL EMERGED AS THE BEST ALGORITHM FOR THIS DATASET.

PAYLOAD MASS: LAUNCHES WITH A LOWER PAYLOAD MASS TEND TO SHOW BETTER RESULTS COMPARED TO THOSE WITH A LARGER PAYLOAD MASS.

LAUNCH SITES: MOST LAUNCH SITES ARE LOCATED NEAR THE EQUATOR AND IN CLOSE PROXIMITY TO THE COAST.

SUCCESS RATE OVER TIME: THE SUCCESS RATE OF LAUNCHES HAS INCREASED OVER THE YEARS.

TOP LAUNCH SITE: KSC LC-39A BOASTS THE HIGHEST SUCCESS RATE AMONG ALL LAUNCH SITES.

ORBIT SUCCESS RATES: ORBITS ES-L1, GEO, HEO, AND SSO HAVE A 100% SUCCESS RATE.

Appendix

- **Special Thanks to:**
- Instructors
- Coursera
- IBM
- **Data Collection:**
- SpaceX REST API: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- Wikipedia Web Scraping: Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time
- **Data Wrangling:**
- Filtering the data
- Dealing with missing values
- Using One Hot Encoding to prepare the data for binary classification
- **Exploratory Data Analysis (EDA):**
- Visualization: Scatter plots, Bar charts, Line charts
- SQL Queries: Unique launch sites, Payload mass, Booster versions, Mission outcomes
- **Interactive Visual Analytics:**
- Folium: Markers of all Launch Sites, Coloured Markers of launch outcomes, Distances between Launch Sites and proximities
- Plotly Dash: Launch Sites Dropdown List, Pie Chart, Slider of Payload Mass Range, Scatter Chart
- **Predictive Analysis (Classification):**
- Models: Logistic Regression, SVM, Decision Tree, KNN
- Metrics: Jaccard_score, F1_score, Confusion Matrix
- Best Model: Decision Tree Model

Thank you!