Winning Space Race with Data Science

Roxana Justiniano
December 21st, 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection from SpaceX API and Web scraping (Wiki)
  - Data Wrangling and Preprocessing
  - Exploratory Data Analysis
    - SQL
    - Data Visualization
  - Interactive Visual Analysis
    - Geographical Data (Folium)
    - Statistic Data (Dashboard)
  - Machine Learning Prediction

- Summary of all results
  - Exploratory Data Analysis
  - Interacting with the Data
  - Machine Learning Predictions

Introduction

Today we are in the middle of the Commercial Space era. The space companies are working on making Space travel more affordable. Reusing the First Stage of a rocket means great savings and it is a significant factor of the cost of a rocket launch.

Being able to understand the historical data and predict the probability of a successful recovery of the First Stage of a rocket will provide a better estimate of the cost of a launch.

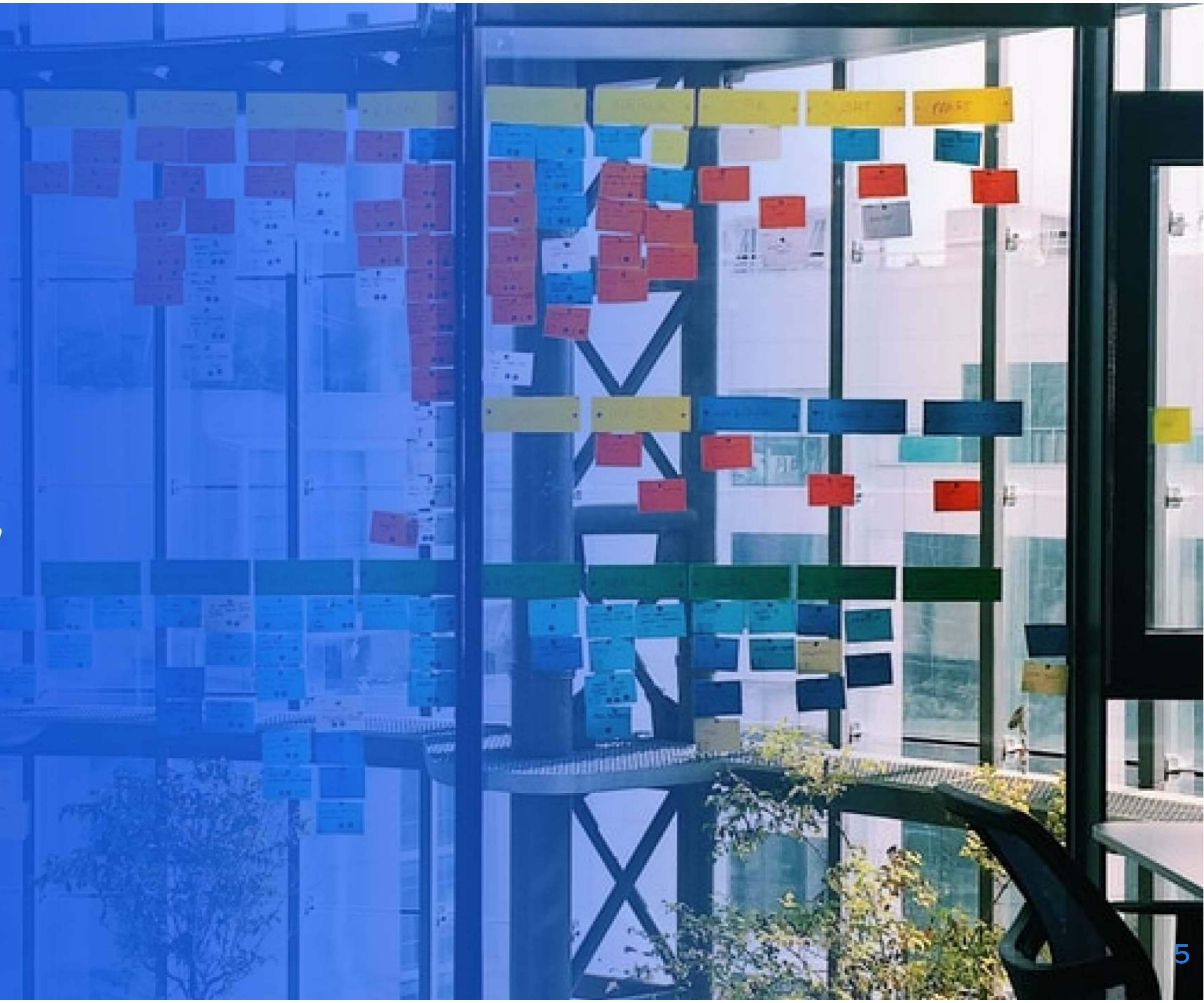
It could also point to other information like:

- What model of rocket has more recovery successes?
- Is there better locations for launching that improve the probability of recovery?

Let's see what story the data is telling us!!

Section 1

Methodology



Methodology

Executive Summary

- Data collection methodology:
  - Data Collection Stage 1: SpaceX API - https://api.spacexdata.com/v4/
  - Data Collection Stage 2: Web Scraping from Wiki - https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- Perform data wrangling
  - NULL values were replaced for Numeric fields (PayloadMass) using the PayloadMass mean
  - One-hot encoding was applied to Categorical Features.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistic Regression
  - SVC
  - Decision Tree
  - KNN

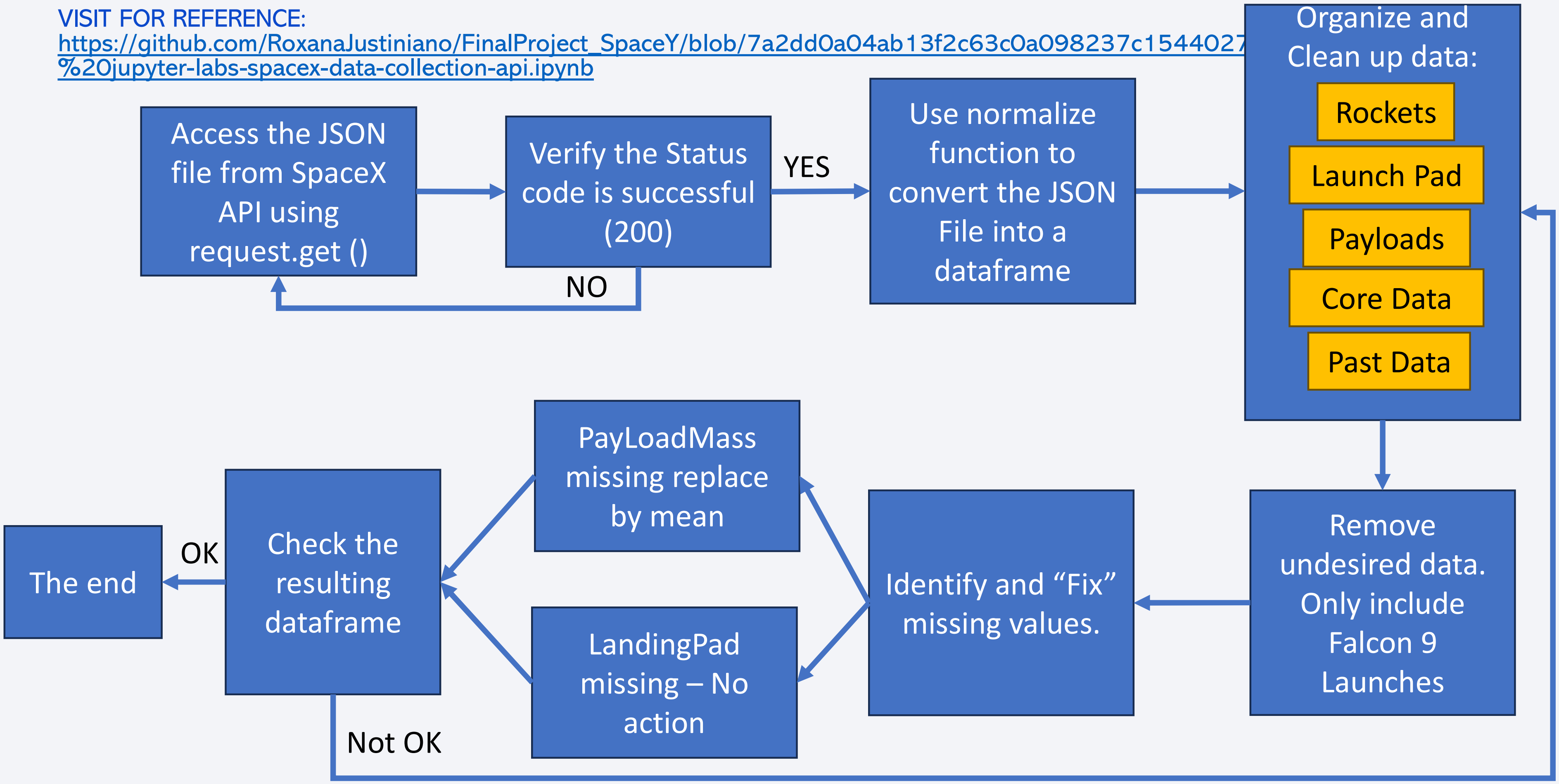
Data Collection

Data was collected from 2 Sources:

- SpaceX API
  - Rocket Data
  - Launch Pads Data
  - Payloads Data
  - Cores Data
  - Historical Data
- Web scraping Falcon 9 and Falcon Heavy Launches Records from Wikipedia.

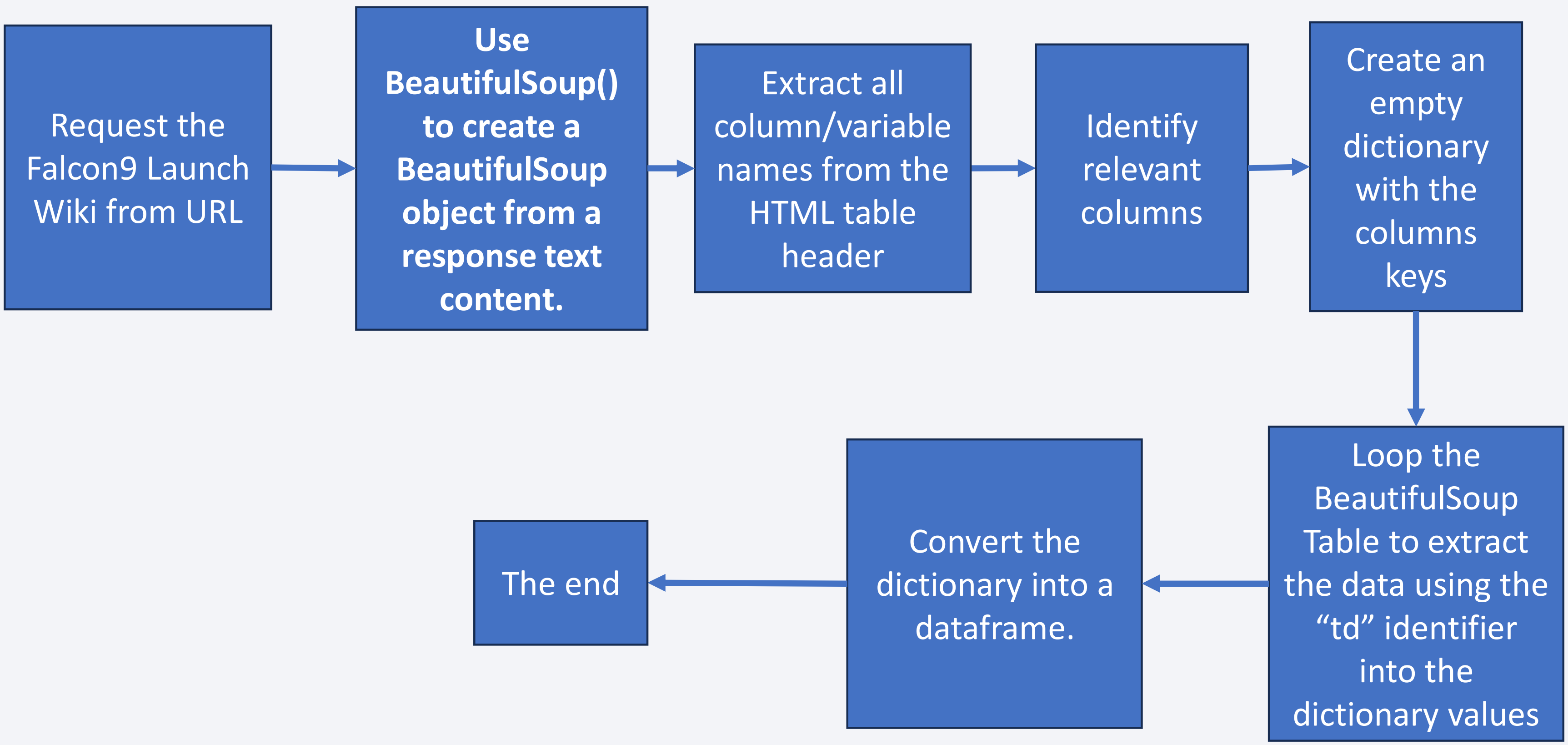
Data Collection – SpaceX API

VISIT FOR REFERENCE:
https://github.com/RoxanaJustiniano/FinalProject_SpaceY/blob/7a2dd0a04ab13f2c63c0a098237c1544027
%20jupyter-labs-spacex-data-collection-api.ipynb



Data Collection – Scraping

VISIT FOR REFERENCE: https://github.com/RoxanaJustiniano/FinalProject_SpaceY/blob/7a2dd0a04ab13f2c63c0a098237c154402799939/02%20-%20jupyter-labs-webscraping.ipynb



Data Wrangling

VISIT FOR REFERENCE: https://github.com/RoxanaJustiniano/FinalProject_SpaceY/blob/7a2dd0a04ab13f2c63c0a098237c154402799939/03%20-%20labs-jupyter-spacex-Data%20wrangling.ipynb

- Now we need to convert the data in a format it is meaningful and provides us useful information.
- This involves classification of the data based on different characteristics or column keys.
- We reviewed the following attributes:

Flight Number	Date	Booster Version	Payload Mass	Launch Site	Outcome(0/1)
Grid Fins	Legs	Block	Reused count	Serial	Longitude / Latitude

- Ordering and grouping the data based on these characteristics will allow us to identify relationship that leads us to decision making strategies in order to increment the possibilities of a successful first stage recovery and reuse in future launches.

EDA with Data Visualization

VISIT FOR REFERENCE:
https://github.com/RoxanaJustiniano/FinalProject_SpaceY/blob/7a2dd0a04ab13f2c63c0a098237c154402799939/05%20-%20jupyter-labs-eda-dataviz.ipynb.ipynb

- Catplot: Categorize the Launches by Launch Site versus Payload Mass
- Scatterplot:
  - Compared PayloadMass against Launch Site. A different view.
  - FlightNumber versus the Orbit, having different colors according the Outcome.
  - Compared PayloadMass against Orbit, again using different colors depending on the Outcome.
- Barplot: To have a visual of the average successful rate classified by Orbit.
- Lineplot: Map the success/failure rate through the years 2010-2020.
- Piechart: Show the % of success/failure of launches per Launch Site

EDA with SQL (part 1)

VISIT FOR REFERENCE: https://github.com/RoxanaJustiniano/FinalProject_SpaceY/blob/7a2dd0a04ab13f2c63c0a098237c154402799939/04%20-%20jupyter-labs-eda-sql-coursera_sqllite.ipynb

- Display the names of the unique launch sites in the space mission
  %sql select distinct Launch_Site from SPACEXTABLE

- Display 5 records where launch sites begin with the string 'CCA'
  %sql select * from SPACEXTABLE where Launch_Site like'CCA%' limit 5

- Display the total payload mass carried by boosters launched by NASA (CRS)
  %sql select SUM(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer like 'NASA (CRS)'

- Display average payload mass carried by booster version F9 v1.1
  %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version like 'F9 v1.1%'

- List the date when the first succesful landing outcome in ground pad was acheived.
  SOLUTION 1: %sql select * from SPACEXTABLE where Landing_Outcome like 'Success%ground pad%' order by Date asc limit 1
  SOLUTION 2: %sql select min(Date) from SPACEXTABLE where Landing_Outcome like 'Success%ground pad%'

- 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 SPACEXTABLE where Mission_Outcome='Success' and Landing_Outcome like '%drone%' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000

EDA with SQL (part 2)

VISIT FOR REFERENCE: https://github.com/RoxanaJustiniano/FinalProject_SpaceY/blob/49e5dcdc4a339e90a1cde994caef4bc3c8641051/04%20-%20jupyter-labs-eda-sql-coursera_sqllite.ipynb

- List the total number of successful and failure mission outcomes

```
%sql select Mission_Outcome,count(*) from SPACEXTABLE group by Mission_Outcome
```

- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery:

```
%sql select Booster_Version from SPACEXTABLE
        where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
        order by Date
```

- 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.

```
%sql select substr(Date,6,2) as Month, Landing_Outcome, Booster_Version, Launch_Site
        from SPACEXTABLE where Landing_Outcome like 'Failure%' and Date like '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.

```
%sql select Landing_Outcome, count(*) as Count_launches from SPACEXTBL
        where Date between '2010-06-04' and '2017-03-20’
        group by Landing_Outcome
        order by Count_launches desc
```

Build an Interactive Map with Folium

VISIT FOR REFERENCE: https://github.com/RoxanaJustiniano/FinalProject_SpaceY/blob/7a2dd0a04ab13f2c63c0a098237c154402799939/06%20-%20lab_jupyter_launch_site_location.ipynb

- Space Y – Visualization using Maps
  - Mark all launch sites on a map using Circular markers and Labels.
  - Mark the success/failed launches for each site on the map using markers clusters.
  - Calculate the distances between a launch site to its proximities using the calculate distance function and the Latitude and longitude information

- Visual information in the map helps us to identify important information like:
  - Is launching site a determining factor in the success of a launch?
  - Does the distance between launching site and a specific geographical site has an impact on the successful recovery rate? For example, the distance to the Equator or to the Coast Line?

Build a Dashboard with Plotly Dash

VISIT for REFERENCE: https://github.com/RoxanaJustiniano/FinalProject_SpaceY/blob/7a2dd0a04ab13f2c63c0a098237c154402799939/07%20-%20spacex_dash_app.py

- Dashboard Structure:
  - SECTION 1: Pie Chart to visualize the Successful Launch rate per Launching Site, where the user can dynamically select a summary of ALL Launching sites or a detailed info per each Launching site.
  - SECTION 2: Scatter Plot to visualize the Success / Failure occurrence versus the Payload Mass and classified per Booster Version (Hue). This will show the result according to the option selected in SECTION 1 (ALL Launching Sites or a specific one). The user can filter per Payload Mass range using the selection bar on top of the plot.
- These plots will let us identify how the launching site can impact the successful rate and how this relates with the Booster Version and Payload Mass of the rocket.

Predictive Analysis (Classification)

VISIT for REFERENCE: https://github.com/RoxanaJustiniano/FinalProject_SpaceY/blob/49e5dcdc4a339e90a1cde994caef4bc3c8641051/08%20-%20SpaceX_Machine_Learning_Prediction_Part_5.ipynb

- Methodology:

1. Using NumPy – create an array from the Class column (Success/Failure)
2. Standardize the data using the Scaler to fit and transform the information.
3. Split the data in Training Set and Testing Set.
4. Apply the Machine Learning algorithms: Logistic regresson, Support Vector Machine, Decision Tree, and K-Nearest Neighbor.
5. Calculate each model Accuracy using the function *score.*
6. Assess the confusion matrix for each of the models.
7. Evaluate and compare the models using Jaccard_Score, F1_Score and Accuracy to determine which is the best model.

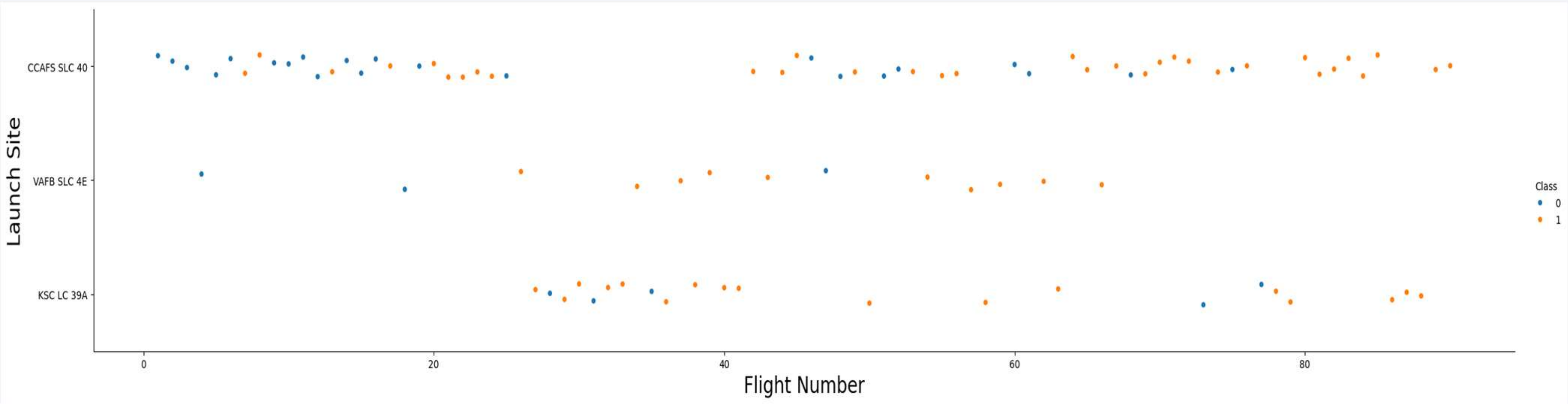
Results Summary

- Launch success has improved over time. We observed big steps one in 2014 and the most important in 2017.
- The launching site with the highest success rate is KSC LC-39A.
- Orbits ES-L1, GEO, HEO, and SSO have the highest success rate.
- When Payload Mass is too high (>6K) the successful rate drops to zero.
- Payload mass range with the highest success rate is between 2K and 5K
- Booster version with the highest rat of success is FT.
- We can observe 2 cluster of Launch Sites, both are close to the coast line.
- Both Launch Site cluster are relatively close to the Equator, but the one tat is closer to the Equator has a higher success rate.
- Launch sites are located away from cities, highways, railways or other locations that could be affected by a failure.
- Decision Tree Model is the most accurate predictive model.

Section 2

Insights drawn from EDA

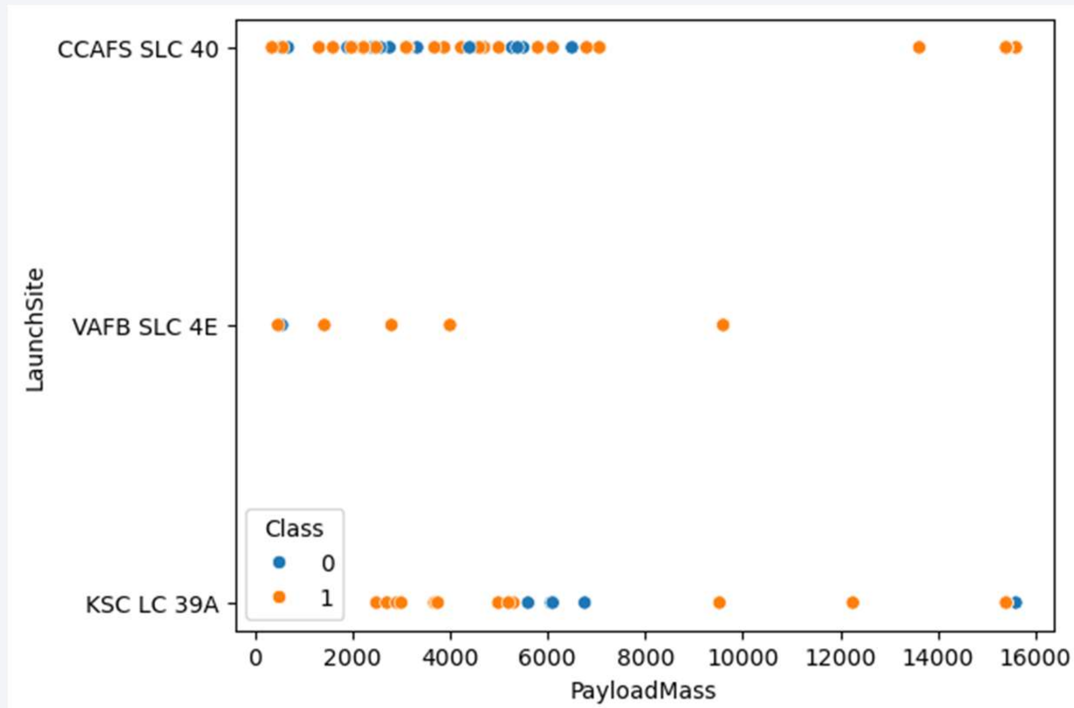
Flight Number vs. Launch Site



- Success rate improves in time. More recent flights show higher success rate.
- The site CCAFS SLC 40 has the higher number of launches (close to 50%)
- The site VAFB SLC 4E has not been used recently
- The Success rate of most recent flights in the locations CCAFS SLC 40 and KSC LC 39A is almost 100%

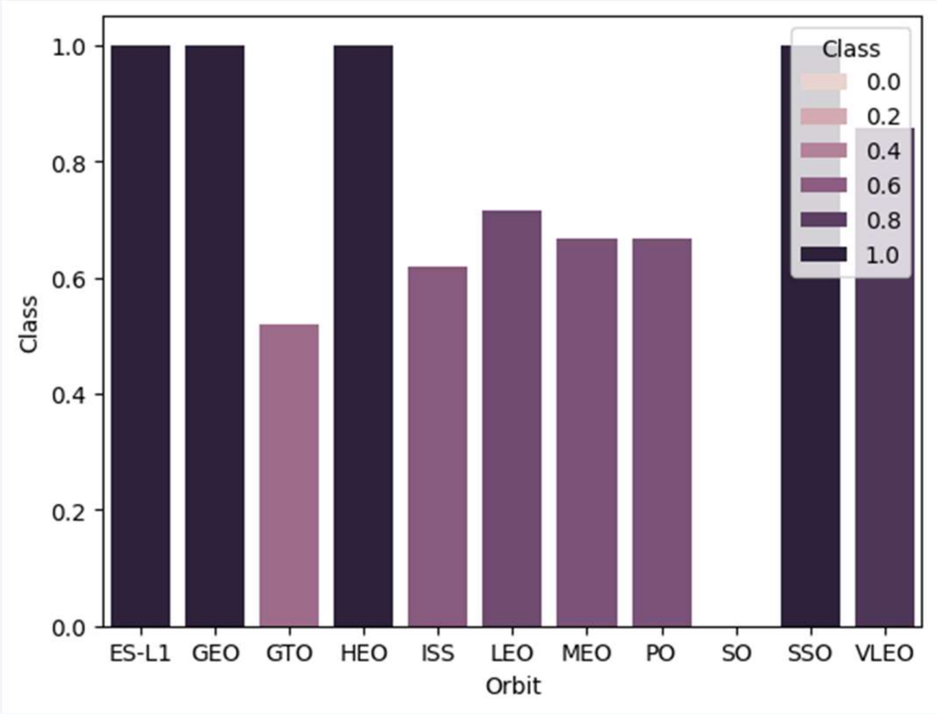
Payload vs. Launch Site

- Most of the launches are for a Payload Mass below 7,000 Kg
- The Location CCAFS SLC 40 has mixed results for Payload Mass less than 8,000 Kg, but a 100% success rate for Payload Mass higher than 12,000 Kg
- The location KSC LC 39A has a 100% success with rockets which Payload Mass is below 5,500 Kg



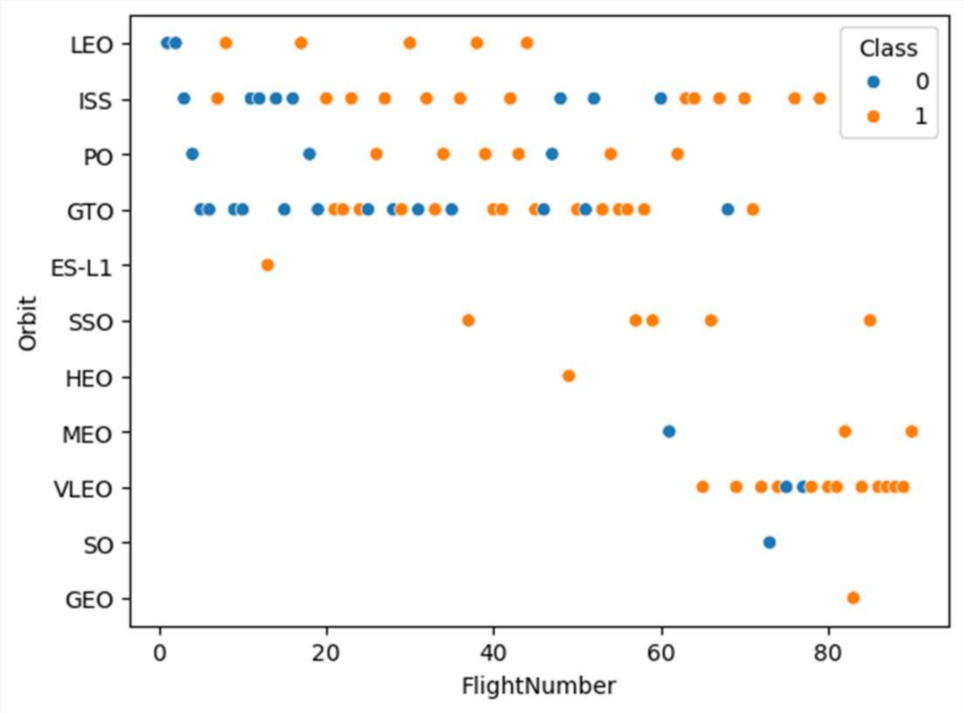
Success Rate vs. Orbit Type

- 100% Success Rate
  - ES-L1
  - GEO
  - HEO
  - SSO
- ~80% Success Rate
  - VLEO
- 50 – 79% Success Rate
  - GTO
  - ISS
  - LEO
  - MEO
  - PO



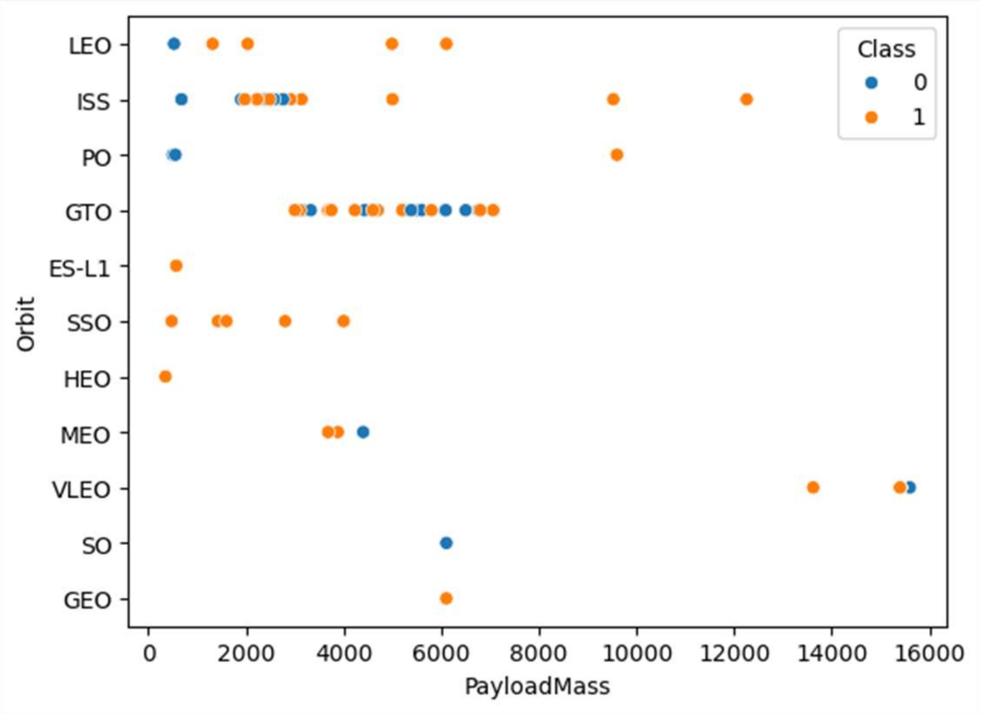
Flight Number vs. Orbit Type

- Success rate has improved over time. More recent launches have been more successful.
- There are orbits which show 100% success, but the sample is too small to be relevant.



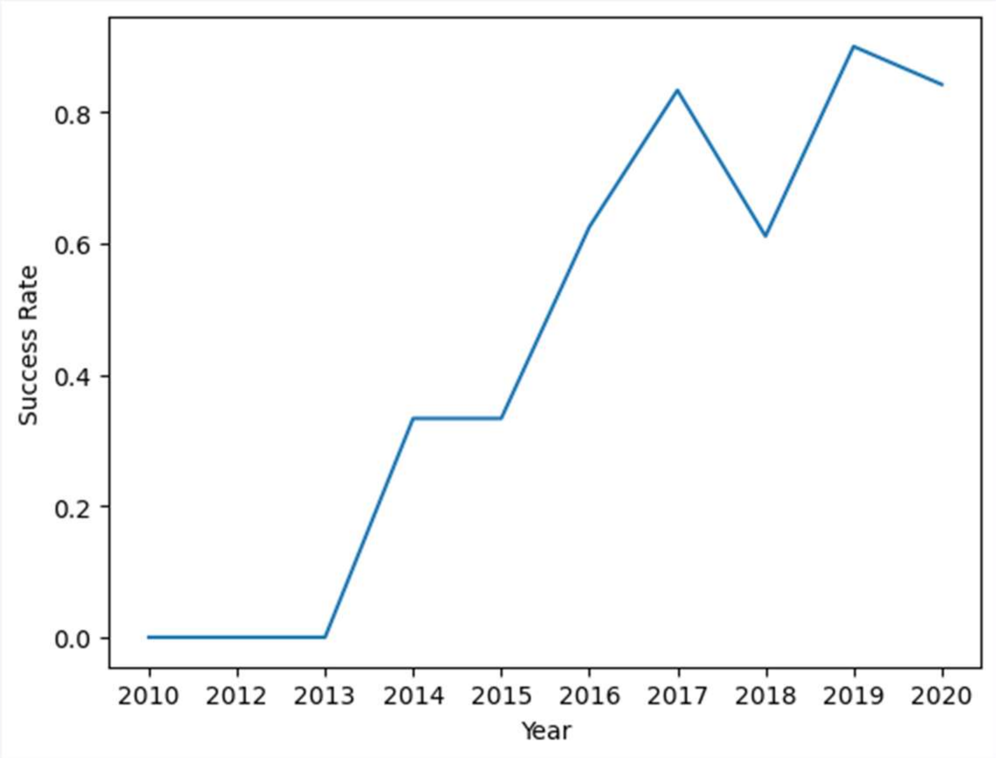
Payload vs. Orbit Type

- It seems higher Payload Mass rockets have a better success rate for certain orbits libe ISS, PO and VLEO. But still the sample is too small to reach a conclusive recommendation.
- For the orbits ES-L1, SSO, HEO, LEO, MEO the results show a higher success with lower Payload Mass.
- GTO Orbit shows mixed results for similar ranges of the Payload Mass, which probably indicates that the results are not correlated to the Orbit.



Launch Success Yearly Trend

- The success rate has significatively improves since 2014. Two big jumps occurred in 2014 and 2017.



All Launch Site Names

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

%sql select distinct Launch_Site from SPACEXTABLE

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

(*) using the function DISTINCT we remove all the duplicate values, getting the list of unique Launching Sites

Launch Site Names Begin with 'CCA'

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

- Using the function “like” and the ‘CCA%’ at the end of the selection condition we can get the entries which Launch_Site begin with “CCA”
- Using the LIMIT 5, we limit the number of entries retrieved to 5.

Total Payload Mass

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

```
%sql select SUM(PAYLOAD_MASS__KG_) from SPACEXTABLE
        where Customer like 'NASA (CRS)'
```

SUM(PAYLOAD_MASS__KG_)
45596

- Using the function SUM(Column_Name) we get the total for that column
- Using like command we filter the customer which contains NASA (CRS)

Average Payload Mass by F9 v1.1

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

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE
        where Booster_Version like 'F9 v1.1%’
```

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

- We get the average of the specified column using the command *avg()*

First Successful Ground Landing Date

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

```
%sql select min(Date) from SPACEXTABLE
        where Landing_Outcome like 'Success%ground pad%'
```

min(Date)
2015-12-22

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 SPACEXTABLE
        where Mission_Outcome='Success' and
        Landing_Outcome like '%drone%’
        PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000
```

Booster_Version
F9 FT B1020
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 Mission_Outcome,count(Mission_Outcome) from SPACEXTABLE
        group by Mission_Outcome
```

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

Boosters Carried Maximum Payload

- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%sql select Booster_Version from SPACEXTABLE
        where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
        order by Date
```

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.

```
%sql select substr(Date,6,2) as Month, Landing_Outcome, Booster_Version, Launch_Site
    from SPACEXTABLE
    where Landing_Outcome like 'Failure%' and Date like '2015%'
```

Month	Landing_Outcome	Booster_Version	Launch_Site
01	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 landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

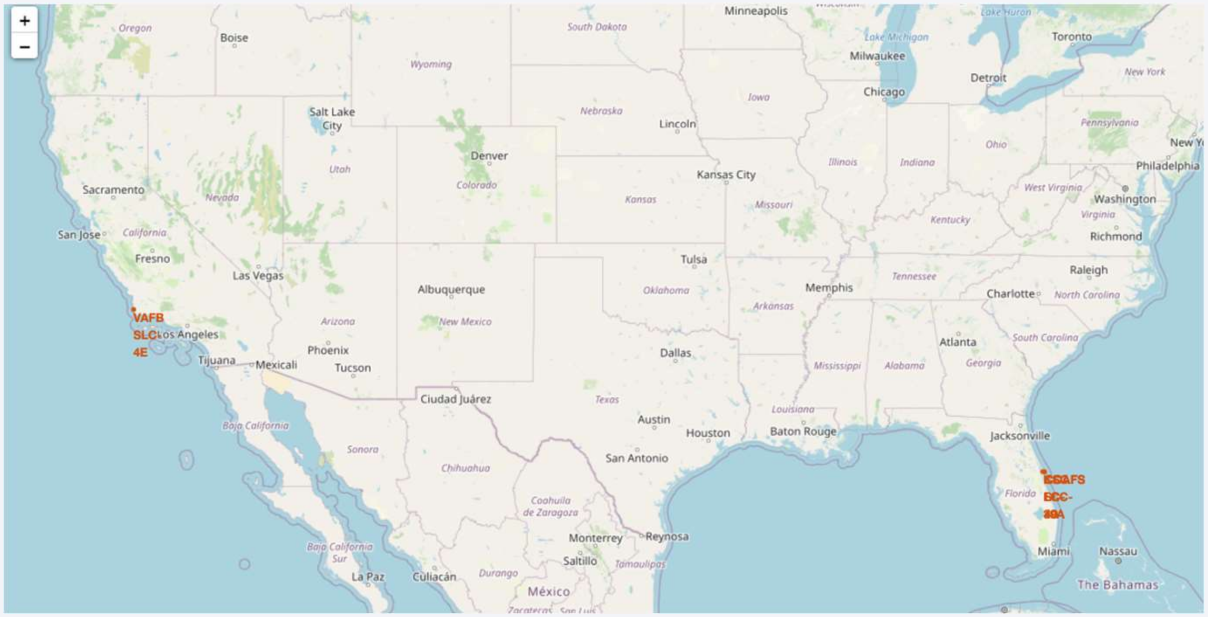
```
%sql select Landing_Outcome, count(*) as Count_launches from SPACEXTBL
    where Date between '2010-06-04' and '2017-03-20'
    group by Landing_Outcome order by Count_launches desc
```

Landing_Outcome	Count_launches
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

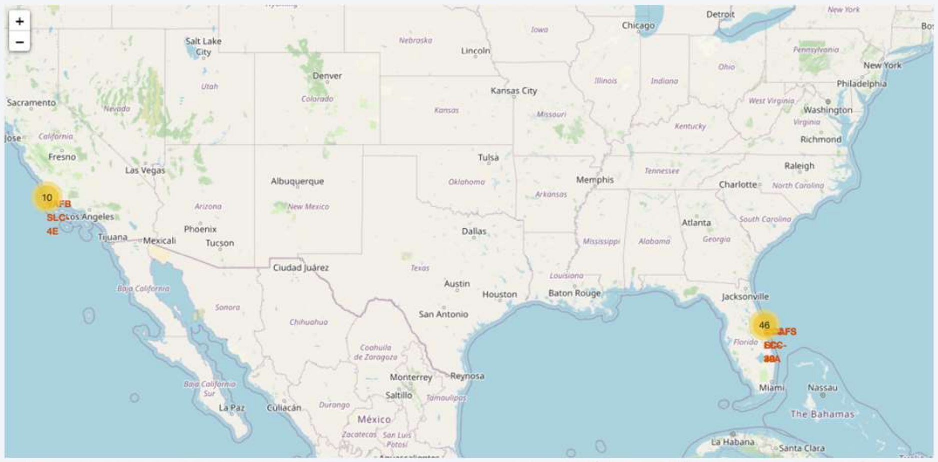
Launch Sites Proximities Analysis

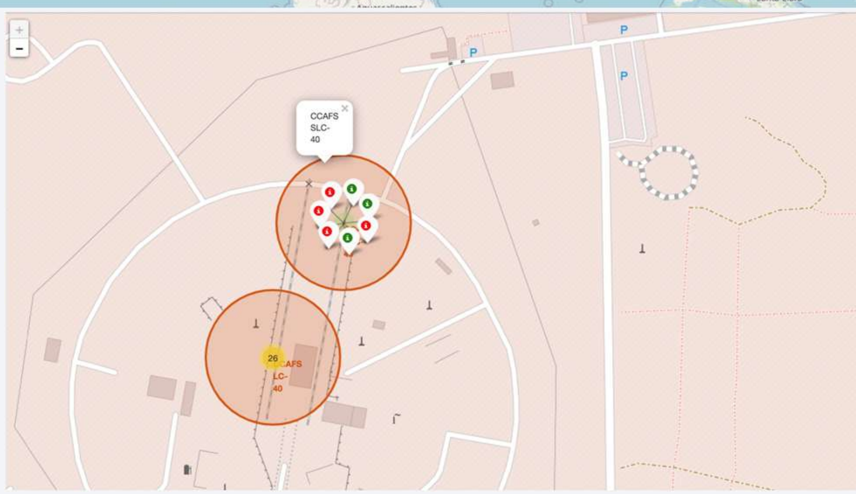
Space X - Launch Sites



- All Launch Sites are located in the South of US, closer to the Equator.
- All Launch Sites are close to the coastline.

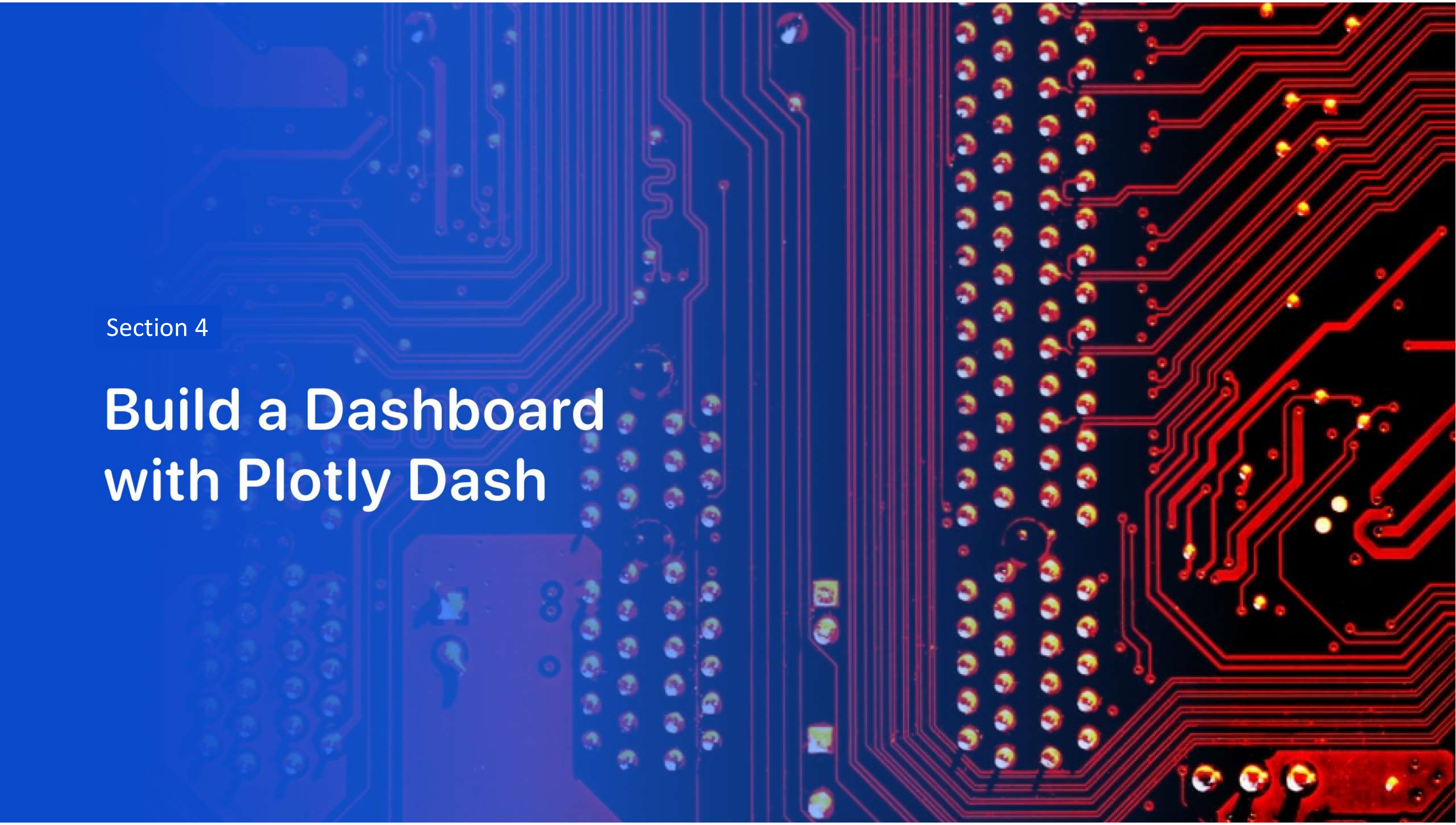
Launches Count per Launch Site and Status





- Most of the Launches took place in the Florida Location, probably because it is closer to the Equator.
- We found more successful launches in the Florida Location than in the California Location.

Launch Site distance to the Coastline



Section 4

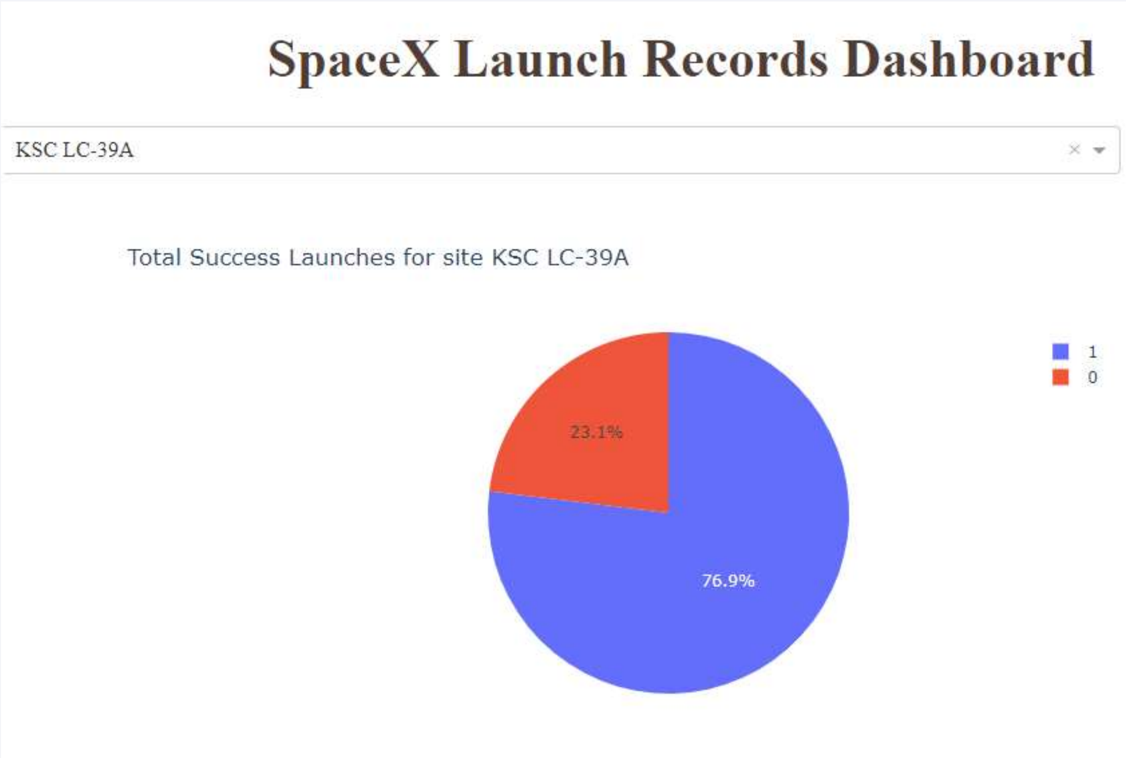
Build a Dashboard with Plotly Dash

Interactive Dashboard Results (ALL Launch Sites)

- The Launch Site with highest success rate is KSC LC-39A

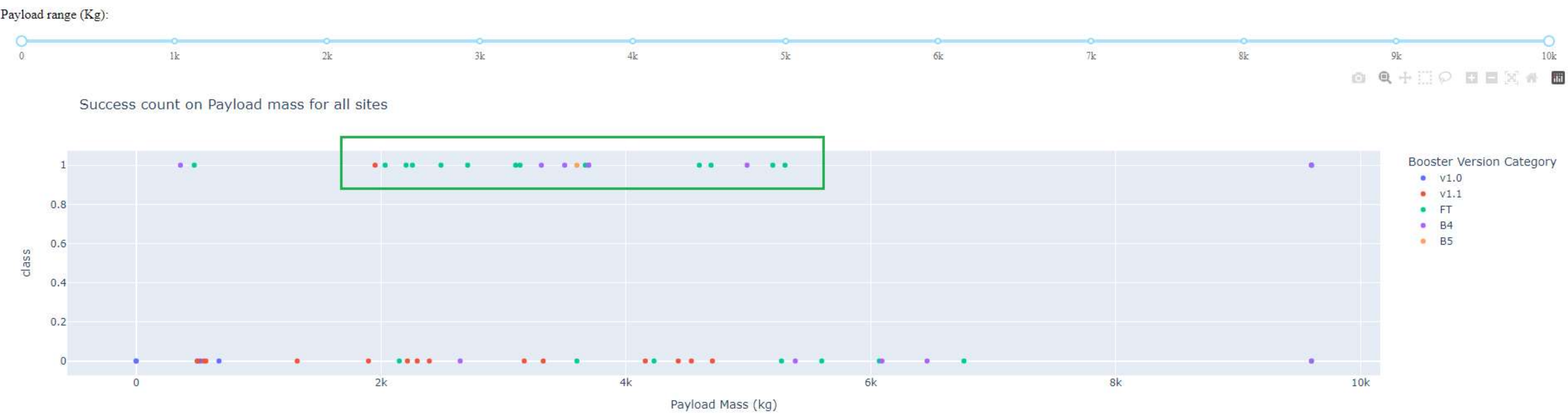
Interactive Dashboard Results (Top Launch Site)

- With a Success rate of 76.9%



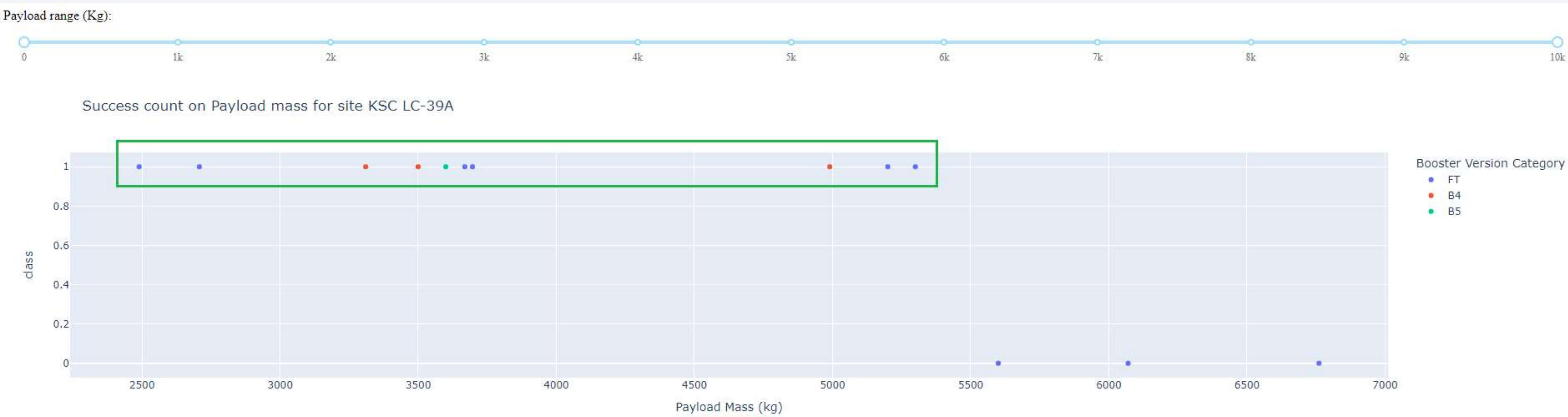
Interactive Dashboard Results (Payoad Mass & Version)

- Among ALL Sites, the rockets with better chance of success are the Booster version FT with a Payload Mass bigger than 2k and less than 5.5k.



Interactive Dashboard Results (Payload Mass & Version)

- Same result is observed in the Top Site (KSC LC-39A), the rockets with better chance of success are the Booster version FT with a Payload Mass bigger than 2k and less than 5.5k.



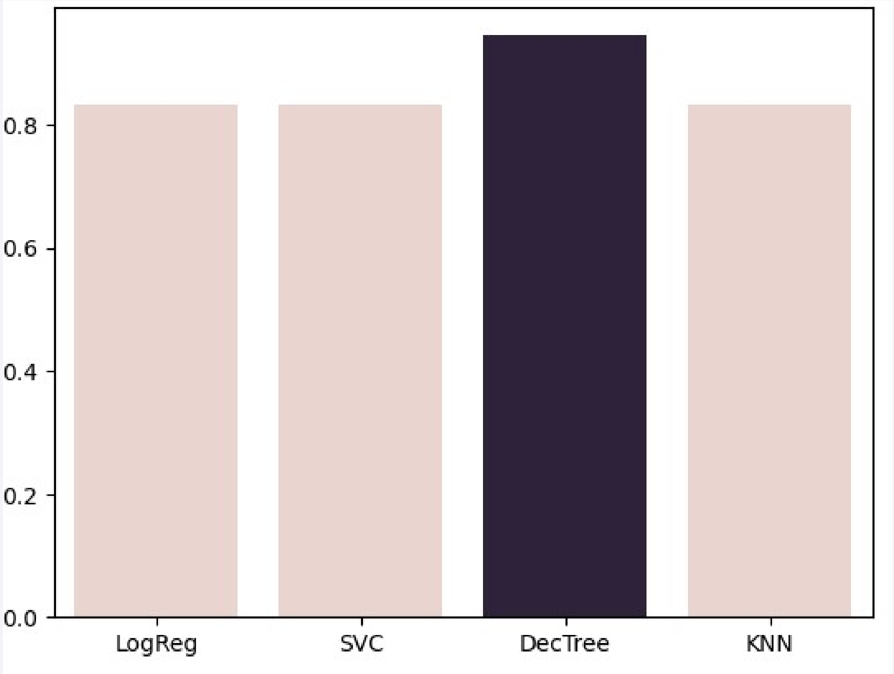
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Decision Tree is the most accurate model for this study case.

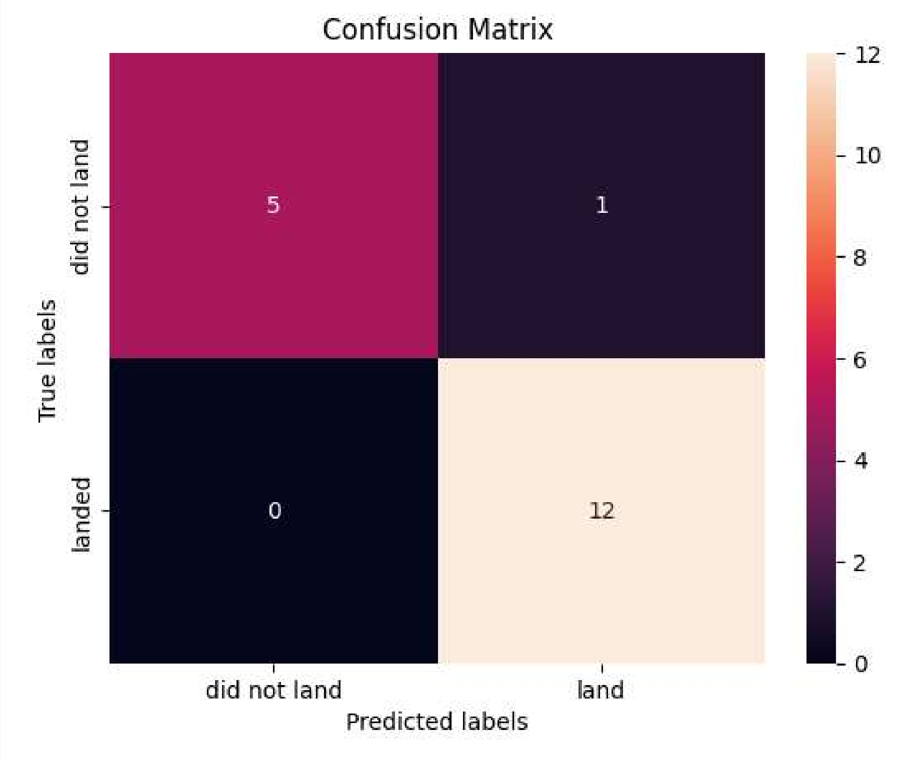
Model	Accuracy
Logaritmic Regression	0.833333
SVC	0.833333
Decision Tree	0.944444
KNN	0.833333



Confusion Matrix

- The confusion matrix is a summary of the performance of the model.
- Here we ilustrate the Decision Tree Confusion Matrix
- This model performs beter than the other models because is the one that resulted in less false positives (1 versus 3)
- Outputs:
    - 12 True positive
    - 5 True negative
    - 1 False positive
    - 0 False negative
- Precision = TP/(TP + FP) = 12 / 15 = 0.8
- Recall = TP / (TP + FN) = 12 / 12 = 1
- F1 Score = 2x(Precision x Recall) / (Precision + Recall)

  =2 x 0.8 / 1.8 = 0.89



Conclusions

- Launch success has improved over time. We observed big steps one in 2014 and the most important in 2017.
- The launching site with the highest success rate is KSC LC-39A.
- Orbits ES-L1, GEO, HEO, and SSO have the highest success rate.
- When Payload Mass is too high (>6K) the successful rate drops to zero.
- Payload mass range with the highest success rate is between 2K and 5K
- Booster version with the highest rat of success is FT.
- We can observe 2 cluster of Launch Sites, both are close to the coastline.
- Both Launch Site cluster are relatively close to the Equator, but the one tat is closer to the Equator has a higher success rate.
- Launch sites are located away from cities, highways, railways or other locations that could be affected by a failure.
- Decision Tree Model is the most accurate predictive model.

Appendix

- GITHUB LINK:

https://github.com/RoxanaJustiniano/FinalProject_SpaceY.git


Thank you!