IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Luiz Furtado
4/2023

Outline

- Executive Summary & Introduction
- Methodology
- Data Collection
- Exploratory Data Analysis (static and interactive)
- Launch Sites Proximity Analysis
- Interactive Dashboard
- Predictive Analysis
- Results
- Conclusion
- Appendix

Section 1

Executive Summary

Executive Summary

- Winning the Space Race with Data Science

Project for 'SpaceY'

The commercial space age is here, companies are making space travel affordable for everyone. Perhaps the most successful company nowdays is SpaceX. SpaceX's accomplishments include: Sending spacecraft to the International Space Station. Starlink, a satellite internet constellation providing satellite Internet access. Sending manned missions to Space. One reason SpaceX can do this is the rocket launches are relatively inexpensive. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.

Hypothesis: If we can determine if the first stage will land, we can determine the cost of a launch.

The payload is enclosed in the fairings. Stage two, or the second stage, helps bring the payload to orbit, but **most of the work is done by the first stage**. This stage does most of the work and is much larger than the second stage. This stage is quite large and expensive. Unlike other rocket providers, SpaceX's Falcon 9 can recover the first stage. Sometimes the first stage does not land. Sometimes it will crash as shown in this clip. Other times, Space X will sacrifice the first stage due to the mission parameters like payload, orbit, and customer.

photo source: Reddit.

Executive Summary

- Summary of methodologies

'**Space Y**' is a rocket launch company that would like to compete with SpaceX founded by Billionaire industrialist 'Allon Musk'.

Our job was:

✓ To determine the price of each launch.

- we did this by gathering information about Space X through their API and created dashboards for our team.

✓ We also determined if SpaceX will reuse the first stage.

- Instead of using rocket science to determine if the first stage will land successfully, we trained a machine learning model and,
- use public information to predict if SpaceX will reuse the first stage.

✓ Tools we used for this project:

- CRISP-DS/DM Framework (Cross Industry Standard Process for Data Mining) to go through all the steps;
- We used python inside a Jupiter Notebook to access SpaceX data through an API;
- We used Numpy and Pandas to collect, transform, select and clean the data as well to carry out the first Exploratory Data Analysis;
- Through SQL we did a more deep Exploratory Analysis;
- We used Matplotlib and Seaborn to be able to do static visual data analysis and Folium and Plotly Dash to perform an interactive visual data analysis;
- We used Sklearn to create and train different models to predict our goals mentioned above.

Executive Summary

- Summary of all results

FEATURES:

The period of analysis was between 2010 and 2020.

We had 17 features in the data base which 'payload mass', 'launch location' , 'booster version' and 'target orbit' were the principal ones to determine the success of a launch (successful landing of the first stage when it comes back).

LAUNCH SITES:

During this period were used 4 launch sites: 3 in Cape Canaveral: CCAFS SLC 40, CCAFS LC 40, KSC LC 39A and one in California: VAFB SLC 4E.

The last two (KSC and VAFB) have the highest successful landing rate: 77%.

There are no rockets launched with heavypayload mass (more than 10.000Kg) in the VAFB-SLC launch site.

All the sites are located less than 1Km from coast and near at least one highway and one railway. As well as far away at least 10 km from the nearest city.

PAYLOAD MASS:

The payload mass can be separated in 4 categories: 0 to 2.500Kg, 2.501 to 5.000Kg, 5.001 to 7.500Kg and 7.500 to more.

The average payload mass carried was: 2.535Kg.

And the maximum payload carried was 15.600Kg.

ORBIT:

The orbit that has been targeted the most times was: GTO (geosynchronous/geostationary) but with the least successful rate.

The most successful were: ES-L1, GEO, HEO, SSO and VLEO which means the low and medium earth orbit has greater successful rate.

CUSTOMER:

The principal customer of SpaceX was: NASA with 17% of the payload mass carried.

SUCCESSFUL RATE:

The success rate mean was 67% in more than 90 launch attempts in the period.

The most successful booster version was: the F9 v1.1

The first successful landing outcome in ground pad was achieved in 1-2017.

After 20 launches reached in 2013 the rate of successful landings increase quickly until reach 24 successful landings in 2020.

PREDICTION MODEL:

Our best result was with the building of a DecisionTree ML Model with accuracy 89% to predict when landing will be successful or not.

Introduction

- Project background and context

The commercial space age is here, companies are making space travel affordable for everyone. Perhaps the most successful company nowdays is SpaceX. SpaceX's accomplishments include: Sending spacecraft to the International Space Station. Starlink, a satellite internet constellation providing satellite Internet access. Sending manned missions to Space. One reason SpaceX can do this is the rocket launches are relatively inexpensive. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.

- Problems we want to find answers

Hypothesis: If we can determine if the first stage will land, we can determine the cost of a launch.

The payload is enclosed in the fairings. Stage two, or the second stage, helps bring the payload to orbit, but most of the work is done by the first stage. This stage does most of the work and is much larger than the second stage. This stage is quite large and expensive. Unlike other rocket providers, SpaceX's Falcon 9 can recover the first stage. Sometimes the first stage does not land. Sometimes it will crash as shown in this clip. Other times, Space X will sacrifice the first stage due to the mission parameters like payload, orbit, and customer.

Other questions:

- How much is the cost of a launch?
- What are the factors that affect the landing?
- Can we access and use the information that we need? How do we access? How this data come? Does it need to be crafted?
- Can we use this data to predict landing outcomes with good accuracy?

Section 2

Methodology

Methodology

- Data collection methodology:
  - We made a get request to the SpaceX API. You also did some basic data wrangling and formating. Source: https://api.spacexdata.com/
- Data wrangling
  - In the data set, there are several different cases where the booster did not land successfully. We mainly converted those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.
- Exploratory data analysis (EDA) using visualization tools and SQL
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classification models
  - We determined what would be the Training Labels than,
  - We found the best Hyperparameters for SVM, Classification Trees and Logistic Regression to determine the best Model to be adopt.

Section 3

Data Collection

Data Collection

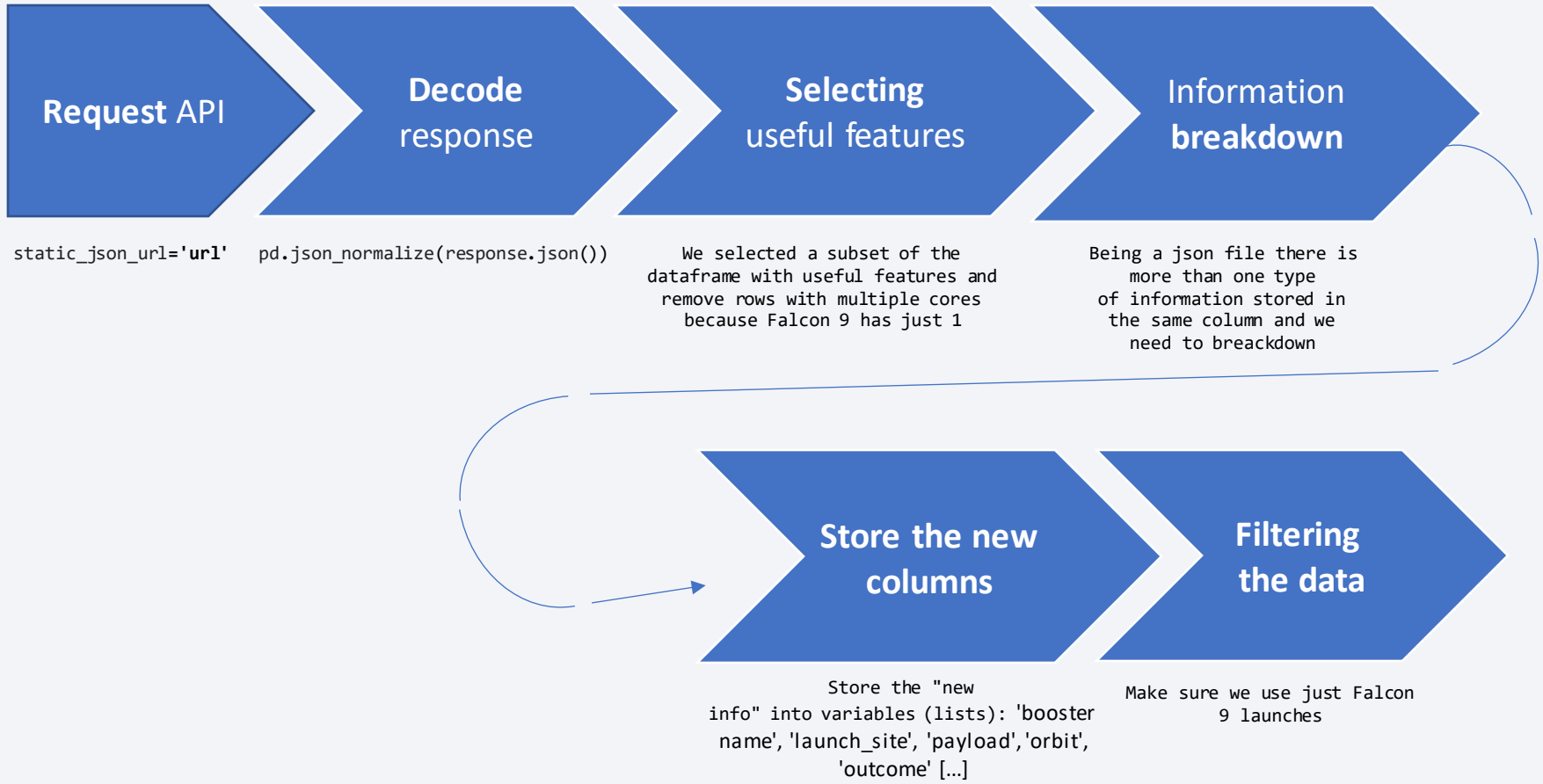
GitHub url:
Week 1_Lab 1_Data Collection.ipynb

Data Collection

- How data sets were collected
  - We requested to the SpaceX API first, than we had to clean the requested data. We used help functions to transform the .json files to readable files in form of data frames with the useful features. From each column on .json we could grab the information we needed i:
    - from 'rocket' we learnt the **booster name**;
    - from 'launchpad' we got the **names**, **latitude** and **longitude** of the sites;
    - from 'payload' we discover the **mass of the payload** and the **orbit** that it is going to;
    - from 'cores' to learn the **outcome** of the landing, the **type** of the landing, **number of flights with that core**, whether **gridfins** were used, **wheter the core is reused**, wheter **legs were used**, the **landing pad**, the **block of the core** which is a number used to seperate version of cores, the **number of times this specific core has been reused**, and the **serial** of the core.

Data Collection

- Data collection process



Data Wrangling

GitHub url:
Week 1_Lab 2_Data Wrangling.ipynb

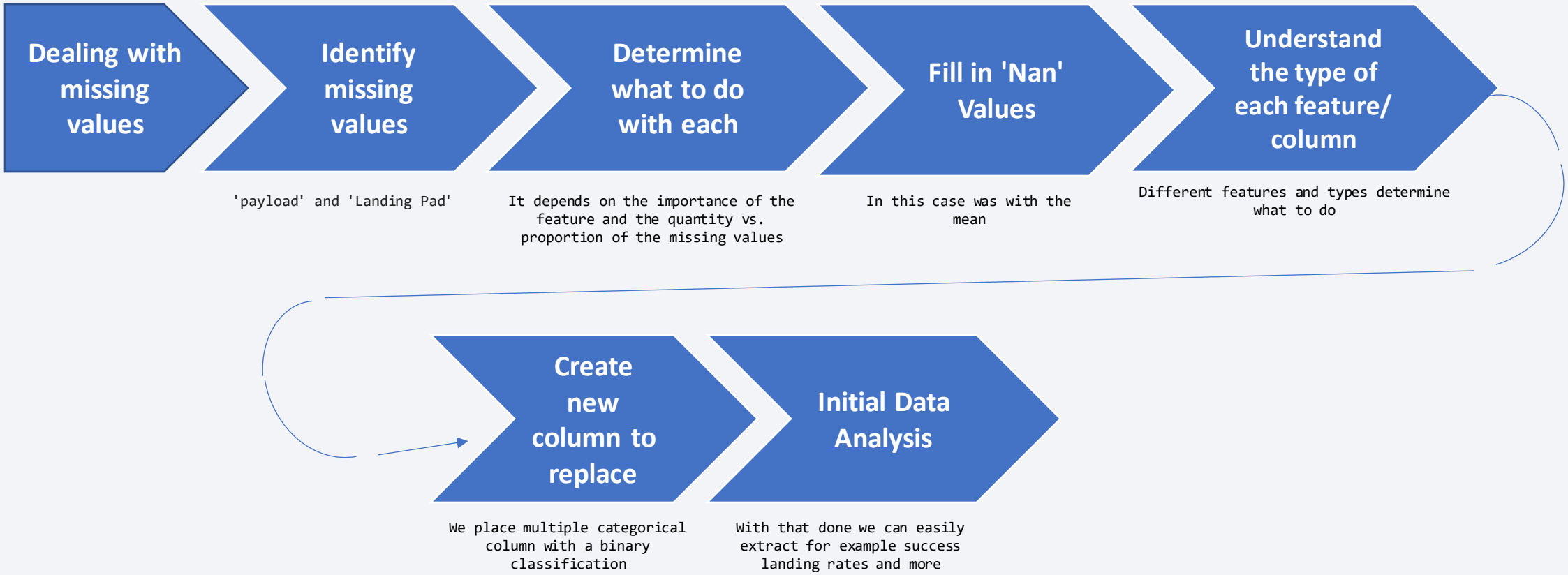
Data Wrangling

- How data were processed

  - First we dealt with **missing values** identifying it and calculating the impact of it using the percentage of the missing values in each attribute,
  - For the '**payload**' attribute we fill in the 'NaN' values with the mean of the payloads,
  - We have identified which columns are numerical and categorical and as the '**outcome**', out target, was a multiple categorical variable we need to transform that,
  - We created a new column '**Class**' to show the outcome with numbers where 1 meaning the first stage landed successfully and 0 the opposity,
  - Than we can go forward and start an **initial data analysis** extracting for example success landing rates and more.

Data Wrangling

- Data wrangling process



Section 4

Insights from Exploratory Data Analysis

EDA with Data Visualization

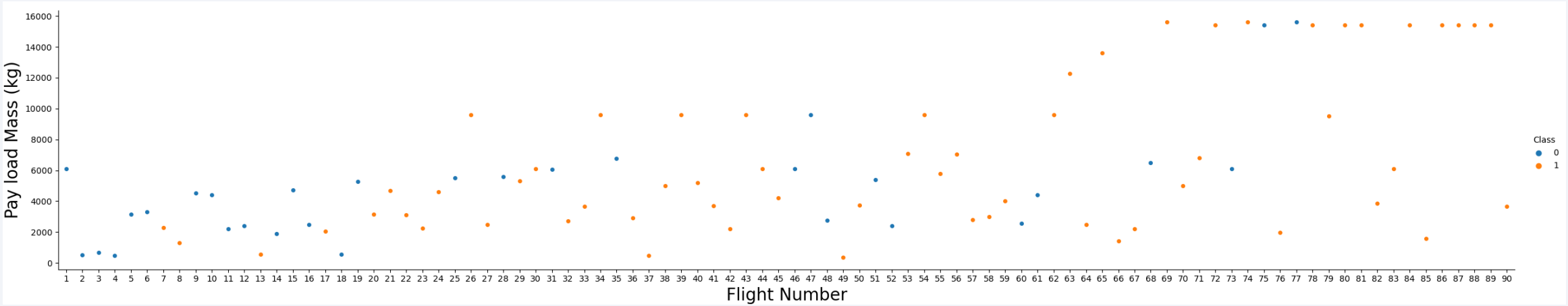
GitHub url:
Week 2_Lab 4_EDA with Visualization.ipynb

EDA with Data Visualization

- FlightNumber vs. PayloadMass

We see that as the flight number increases, the first stage is more likely to land successfully. Which means that there is an increase in successful rate as the company gets more experienced.

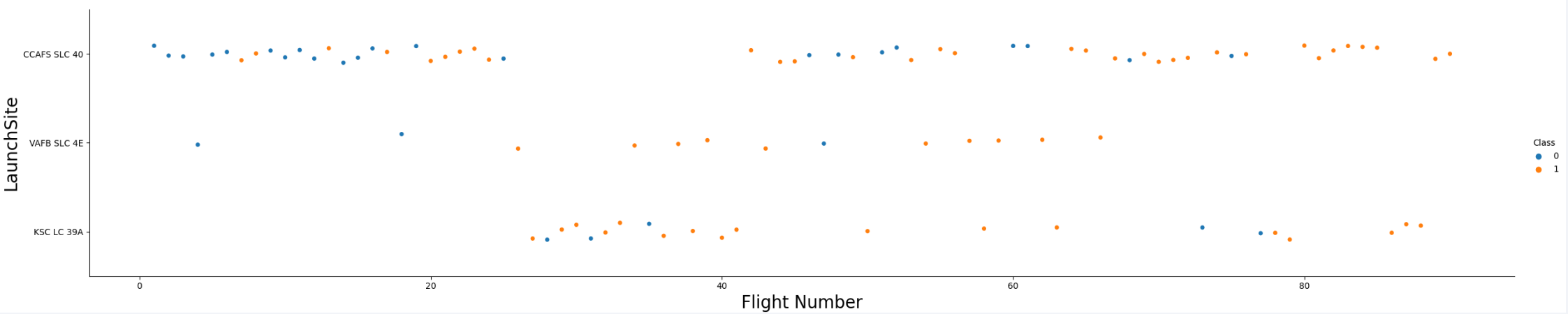
The payload mass is also important; it seems the more massive the payload, the less likely the first stage will return.



EDA with Data Visualization

- FlightNumber vs. LaunchSite

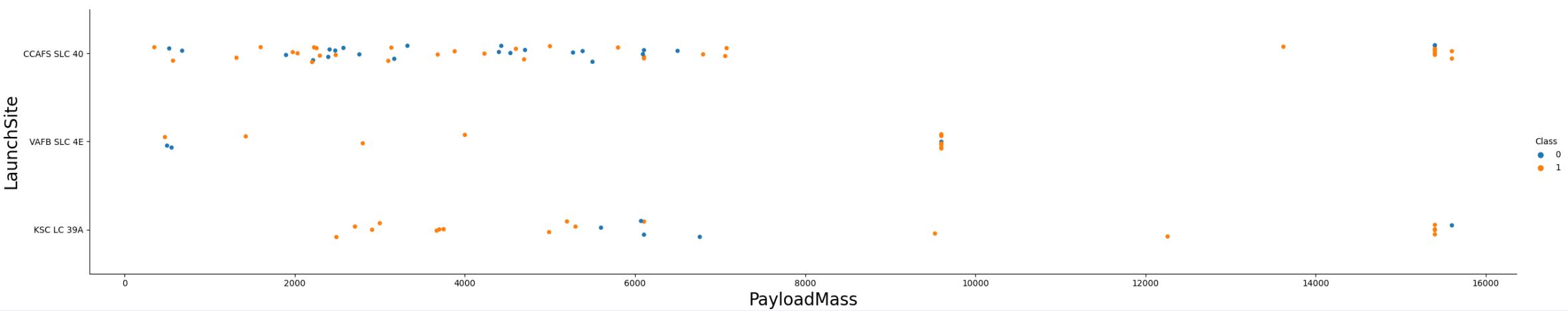
We see that different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%. One of the sites (CCA SLC-40) was much more used at the early flights, so it contains less successful landings.



EDA with Data Visualization

- Payload Mass vs. LaunchSite

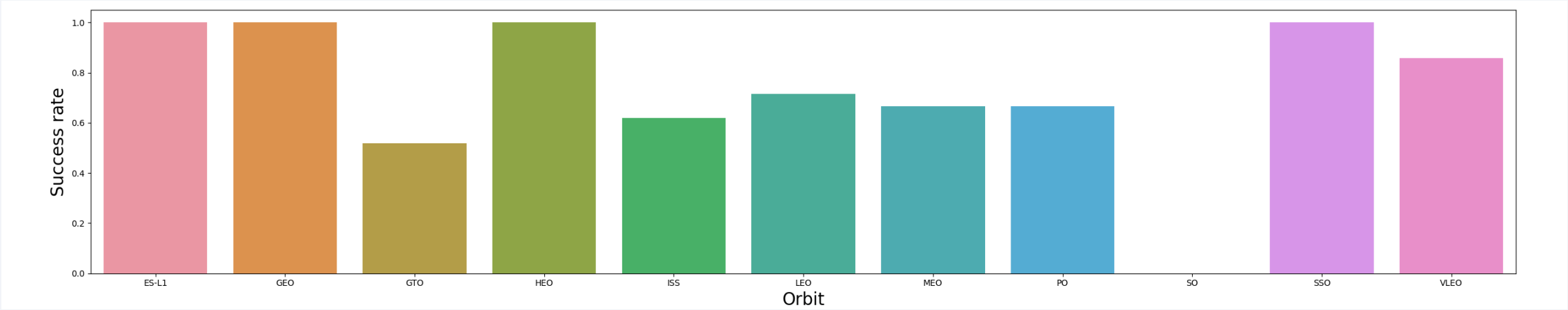
The VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).



EDA with Data Visualization

- Succes rate vs. Destination orbit

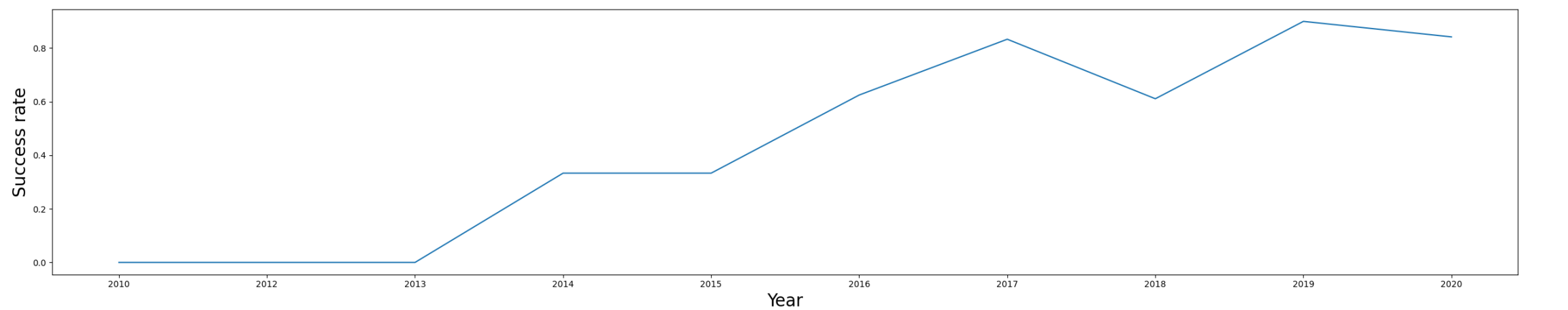
The most successful were: ES-L1, GEO, HEO, SSO and VLEO which means that low and medium earth orbit has greater successful rate.



EDA with Data Visualization

- Succes rate vs. Year

After 20 launches mark reached in 2013 the rate of successful landings increase quickly until reach 24 successful landings in 2020.



EDA with SQL

GitHub url:
Week 2_Lab 3_Complete the EDA with SQL.ipynb

EDA with SQL

- SQL queries performed:

Names of the unique launch sites:

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

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

EDA with SQL

Total payload mass carried by boosters launched by NASA (CRS):

Customer	Total_Payload_Carried_kg
NASA (CRS)	48213

Average payload mass carried by booster version F9 v1.1:

Booster_Version	Avg_Payload_Carried_kg
F9 v1.1 B1003	2534.67

When the first succesful landing outcome in ground pad was achieved:

Landing_Outcome	MIN(Date)
Success (ground pad)	01-05-2017

Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000 (shown with average payload):

Booster_Version	Payload_Mass_Kg
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

EDA with SQL

Total number of successful and failure mission outcomes:

Sucess_Landing	Fail_Landing
61	10

Names of the booster_versions which have carried the maximum payload mass:

Booster_Version	Payload_Mass_Kg
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

EDA with SQL

List of the records which are displaying the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015:

Months	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

Count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order:

Number_of_Landings
57


Section 5
Launch Sites
Proximities Analysis

Interactive Map with Folium

GitHub url:
Week 3_Lab 5_Interactive Visual Analytics with Folium.ipynb

Interactive Map with Folium

- Map objects:

We added **circles** for each launch site and a **markers** for each launch (flight number/row);

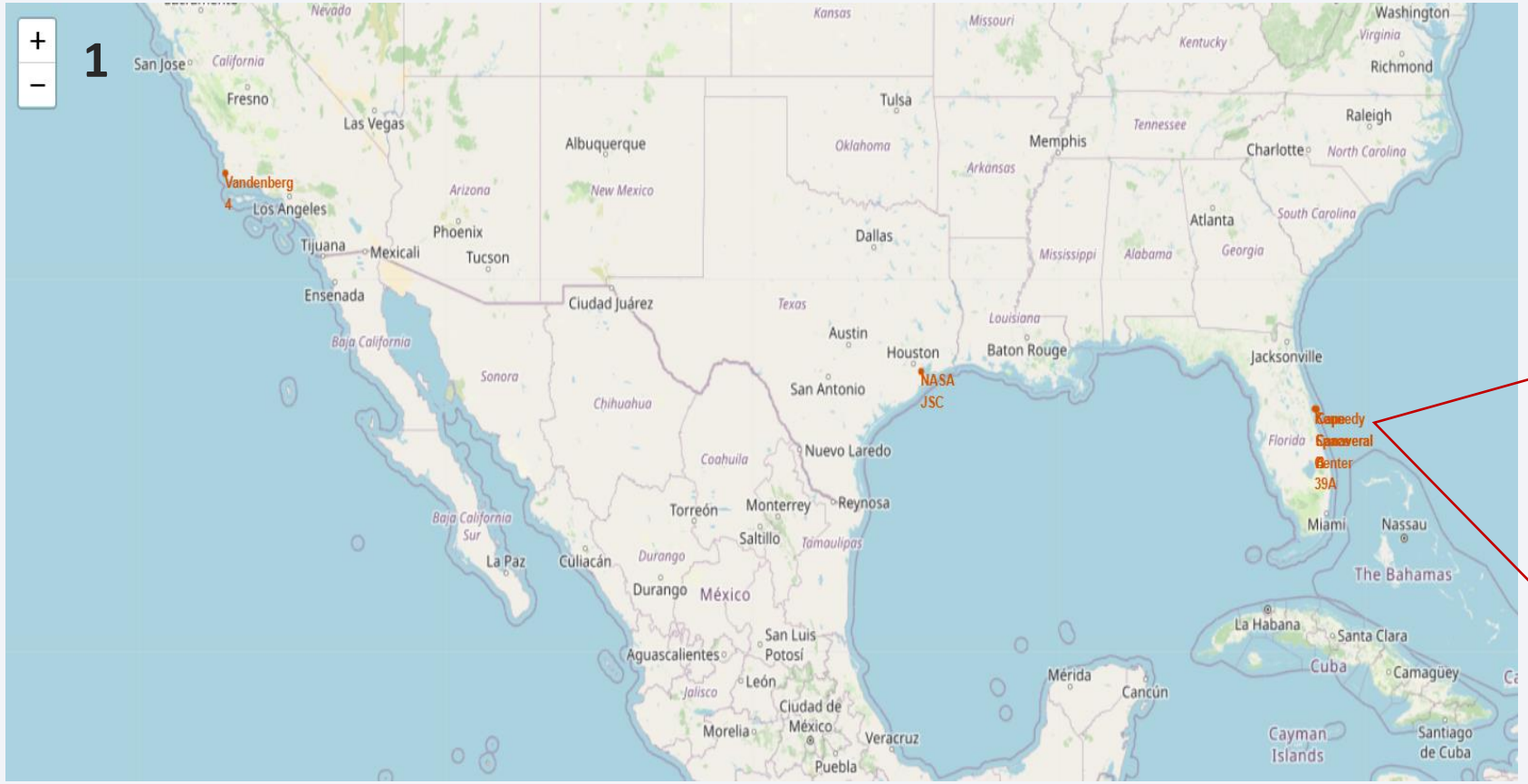
Inside the circles we can see the summary of quantity of launches in each one,

Also, when clicked, we can expand and see a **'pin'** for each launch. If the pin is **green** indicates de landing was successful, if it's **red**, unsuccessful.

We can see some **blue lines** that indicates the **distance** to the coast, to the nearest highway and to the nearest railway.

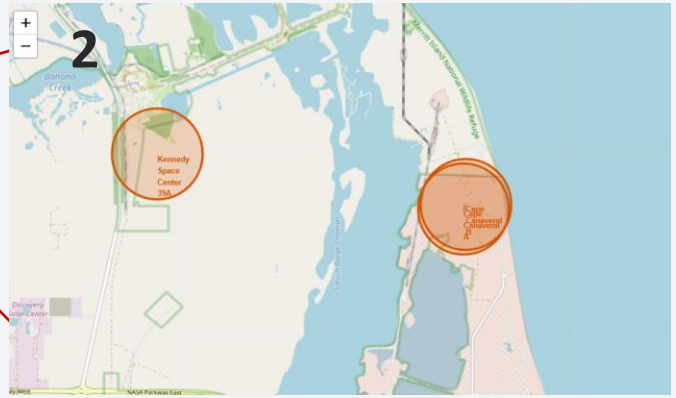
Interactive Map with Folium

- Launch Sites Map



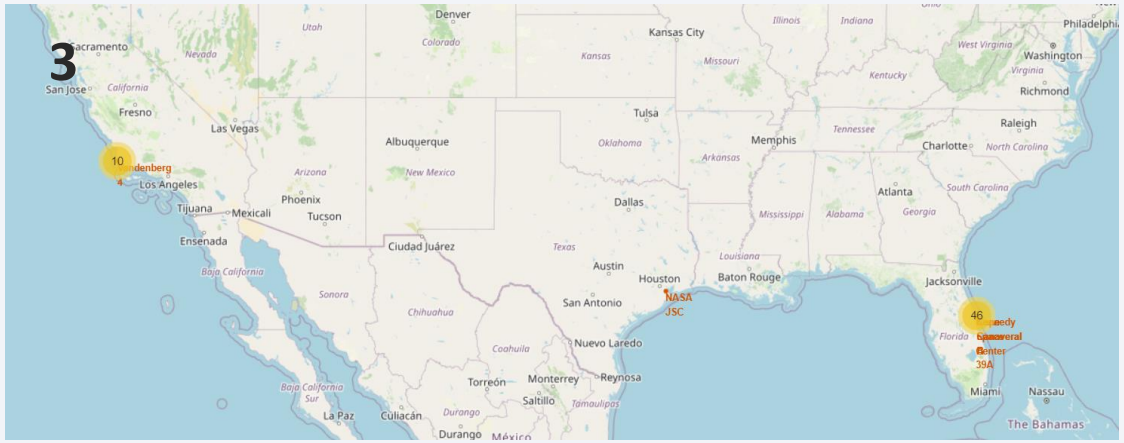
On the left: We can see a red marker with a circle around for each of the 4 launch sites plus NASA Space Center in Houston.

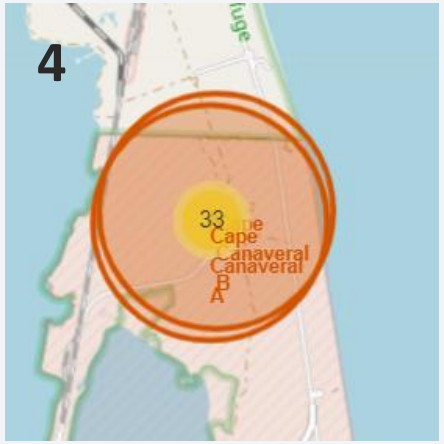
Below: a close look to KSC, CSAFS SLC and CCAFS LC.



Interactive Map with Folium

- Launches Outcomes





Images 3 and 4: We can see in yellow a cluster of all icon markers representing the launching outcomes for each launch.

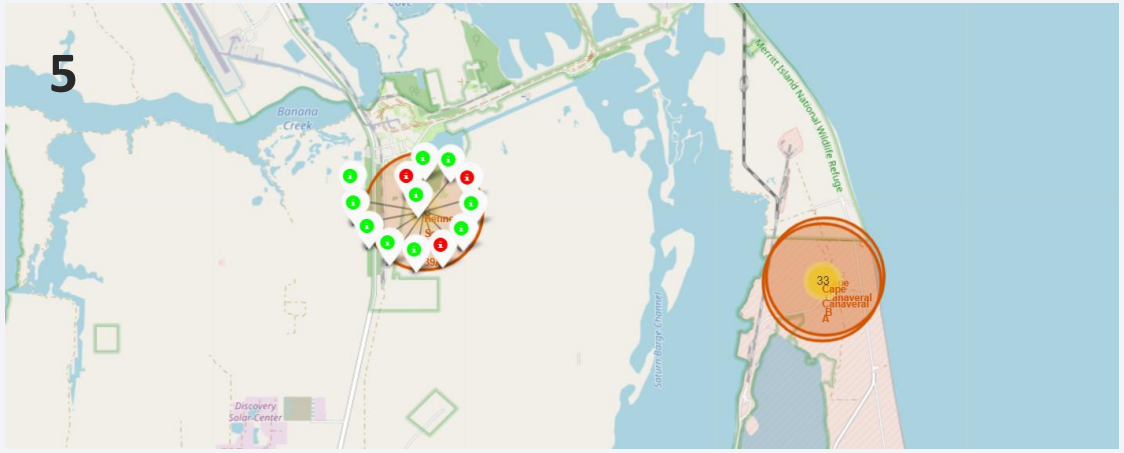


Image 5: By zooming and clicking on the cluster is possible to easily see the outcomes and have an idea of success rate. **Green** if successed and **red** if failed.

Interactive Map with Folium

- Launch Site and its proximities

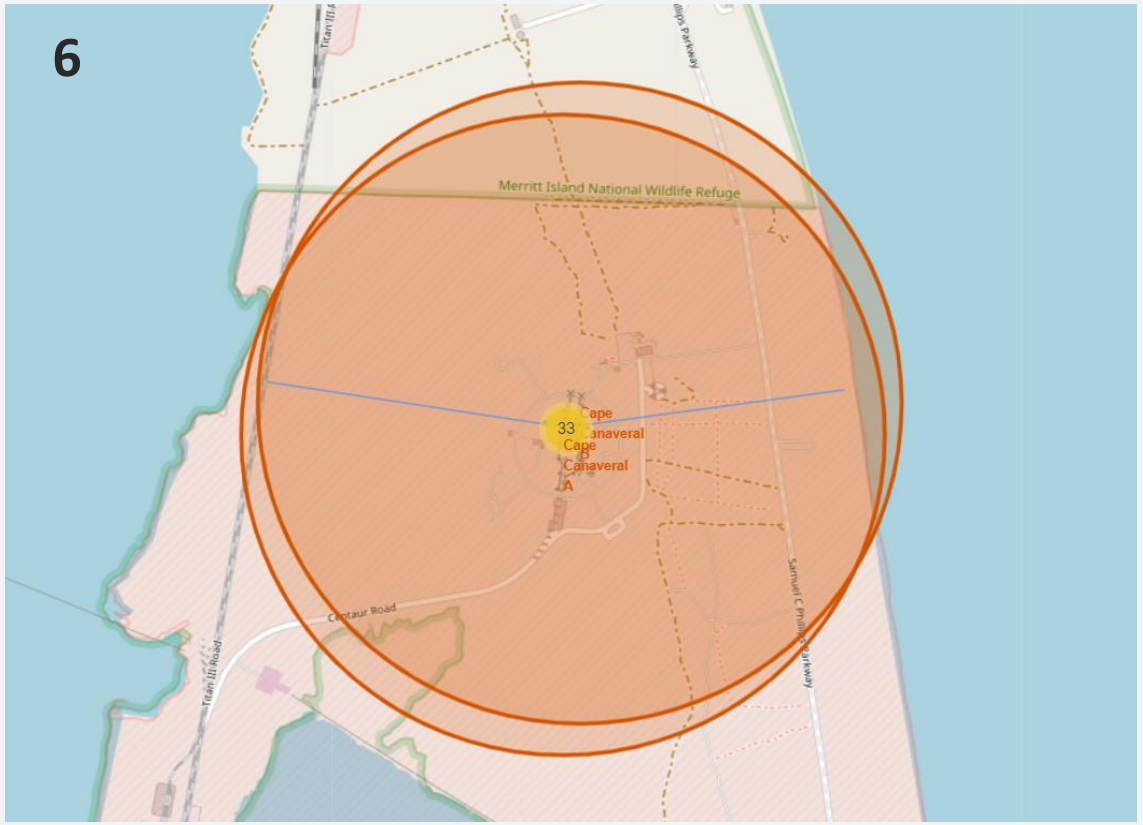
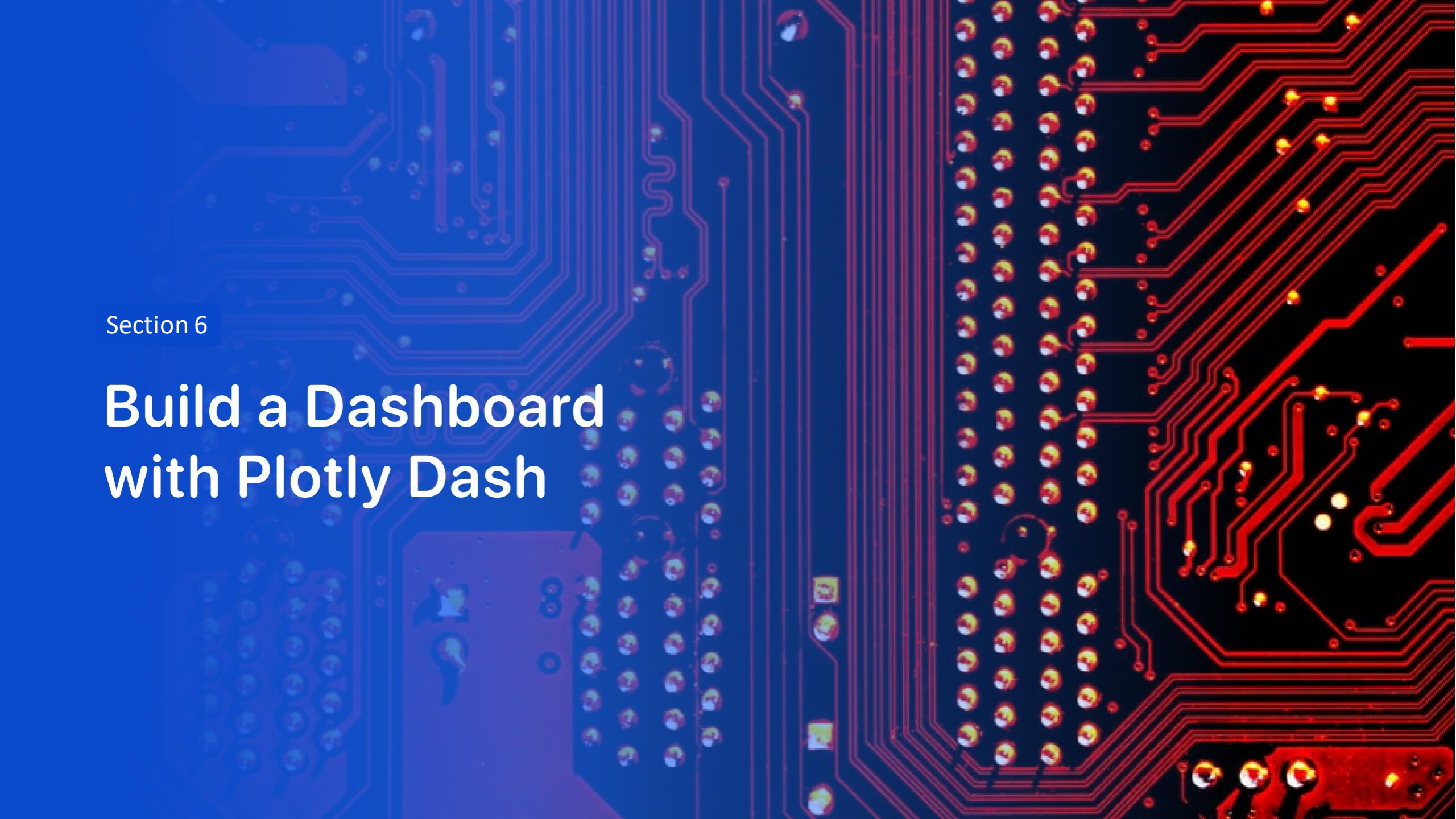


Image 6: Through the **blue line** we can observe the distance of Cape Canaveral Launch Sites to the railway on the left and highway and the coast on the right.


Section 6
Build a Dashboard with Plotly Dash

Dashboard with Plotly Dash

GitHub url:
Week 3_Lab 6_Interactive Dashboard with Ploty Dash.py

Dashboard with Plotly Dash

- We can see:

  Launch Success Rate and Count for all sites and each one of them,

  Payload vs. Launch Outcome for all sites, with different payload.

- With that we can:

  Find the launch site with highest launch success ratio,

  Compare different payloads ranges and discover the more successful.

Dashboard with Plotly Dash

- Launch Success % and Count for all sites

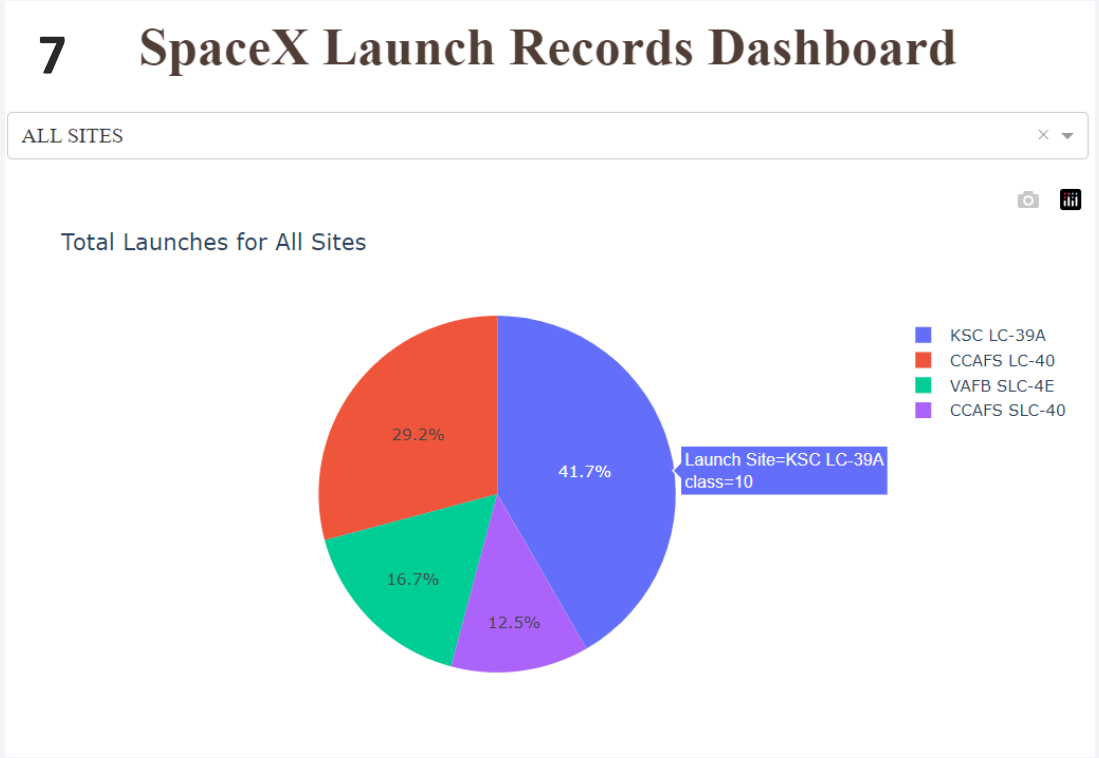


Image 7: We can see the numbers of all sites or select just one of them using the dropdown menu.

Also hovering the mouse over the pie chart we can see the count of success for that location. Ex.: KSC count = 10

Dashboard with Plotly Dash

- Launch site with highest launch success ratio

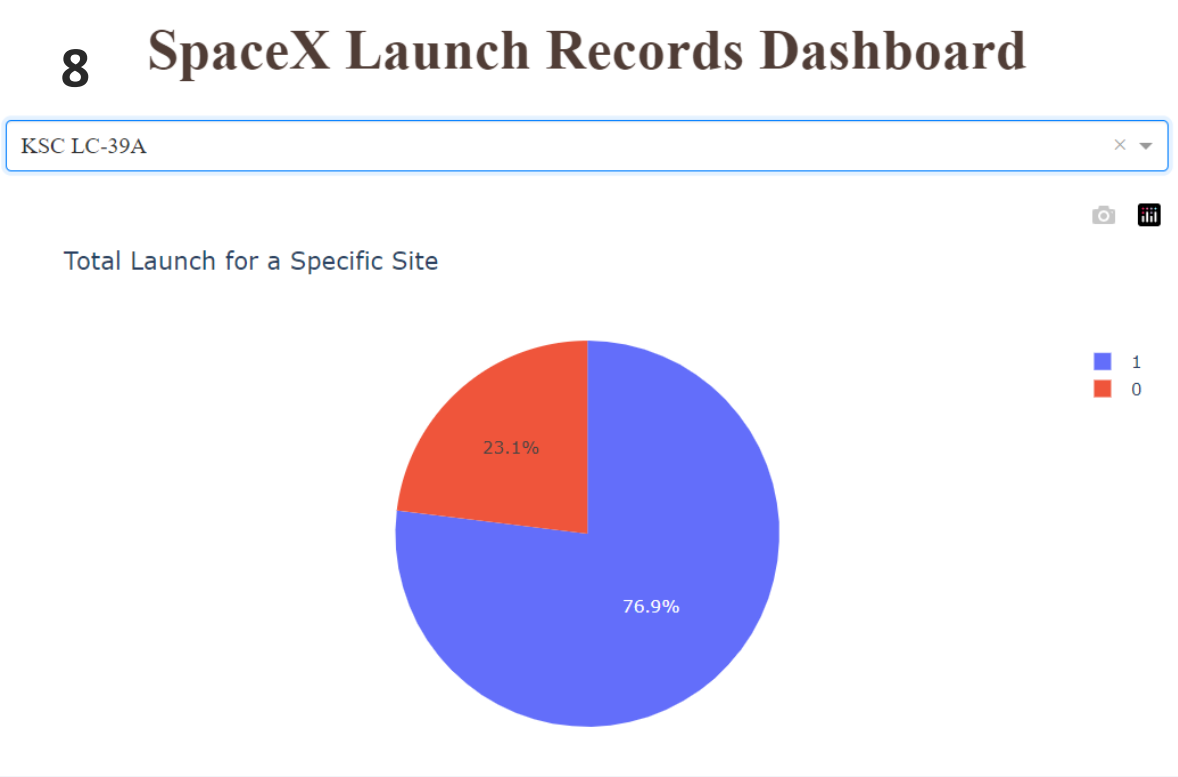
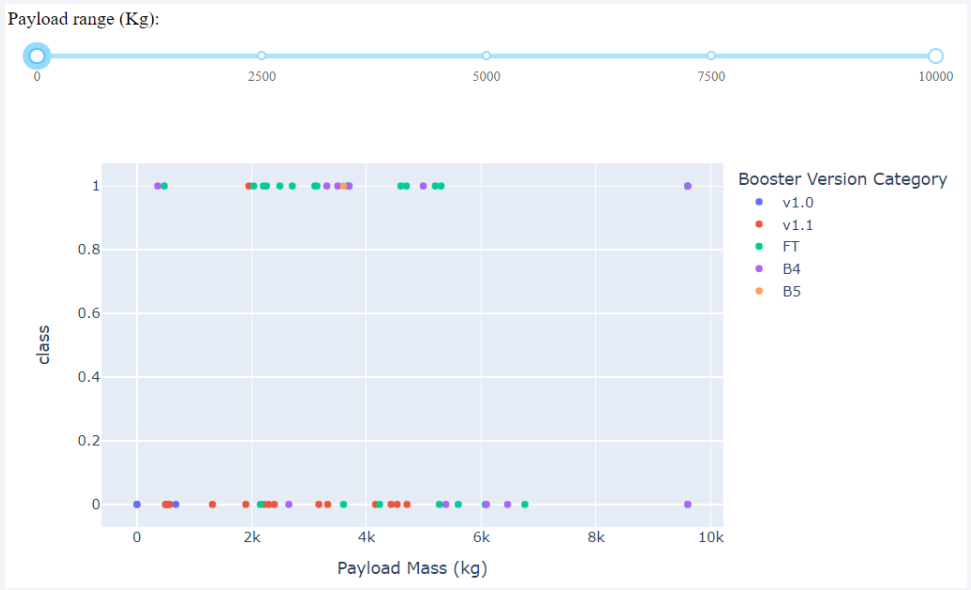


Image 8: Kennedy Space Center (KSC) has the highest launch success ratio and counting10 successful landings.

Dashboard with Plotly Dash

- Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider



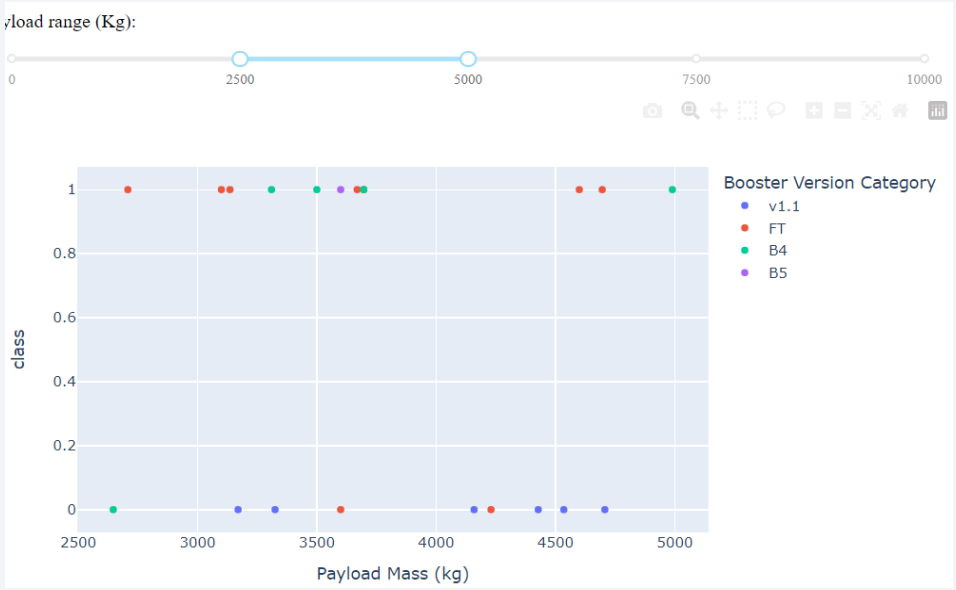


Image 9 and 10: payload between 2.500 and 5.000Kg has much better success rate than other ranges.

Section 7

Predictive Analysis (Classification)

Predictive Analysis (Classification)

GitHub url:
Week 4_Lab 7_Machine Learning Prediction.ipynb

Predictive Analysis (Classification)

- Building, evaluating, improving and finding the best performing classification model

  We **load** the data from IBM Cloud that was previous extracted from SpaceX public data through API,

  Perform some feature engineering: transform in numpy array and standarize de features,

  Train, test and split,

  Create models,

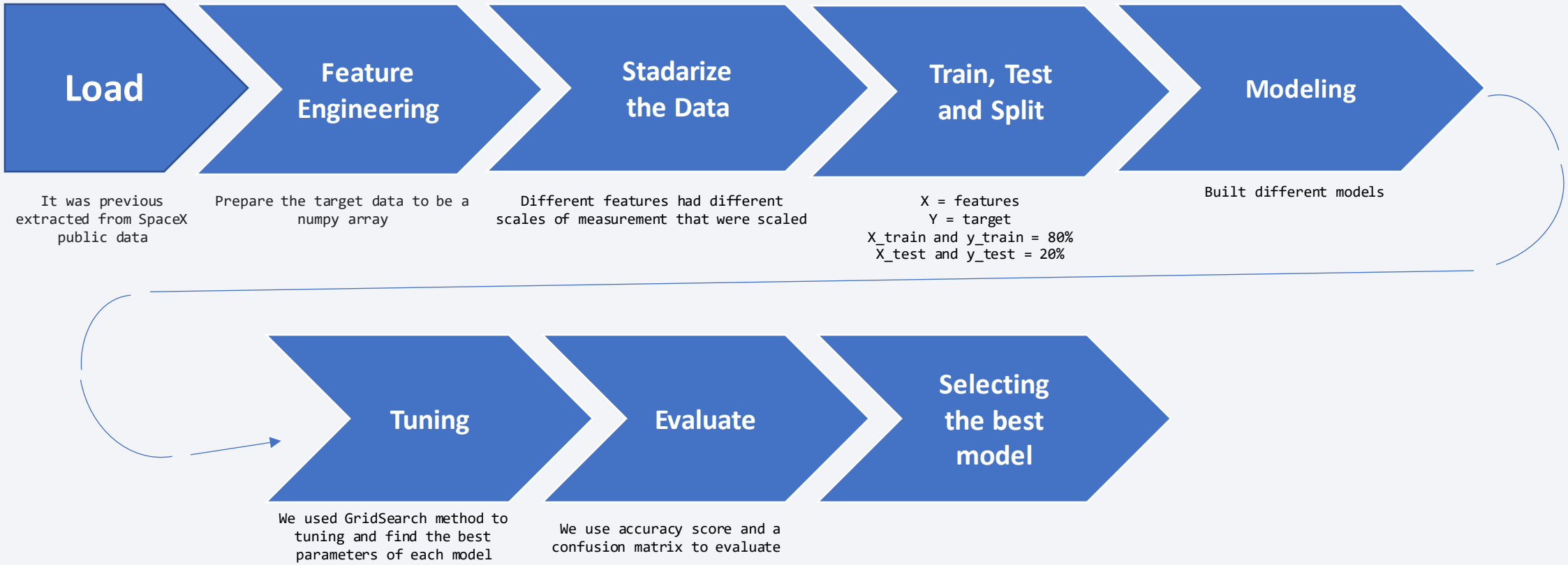
  Tuning with GridSearch to found best hyperparameters,

  Evaluate the models,

  Compere and select the best model.

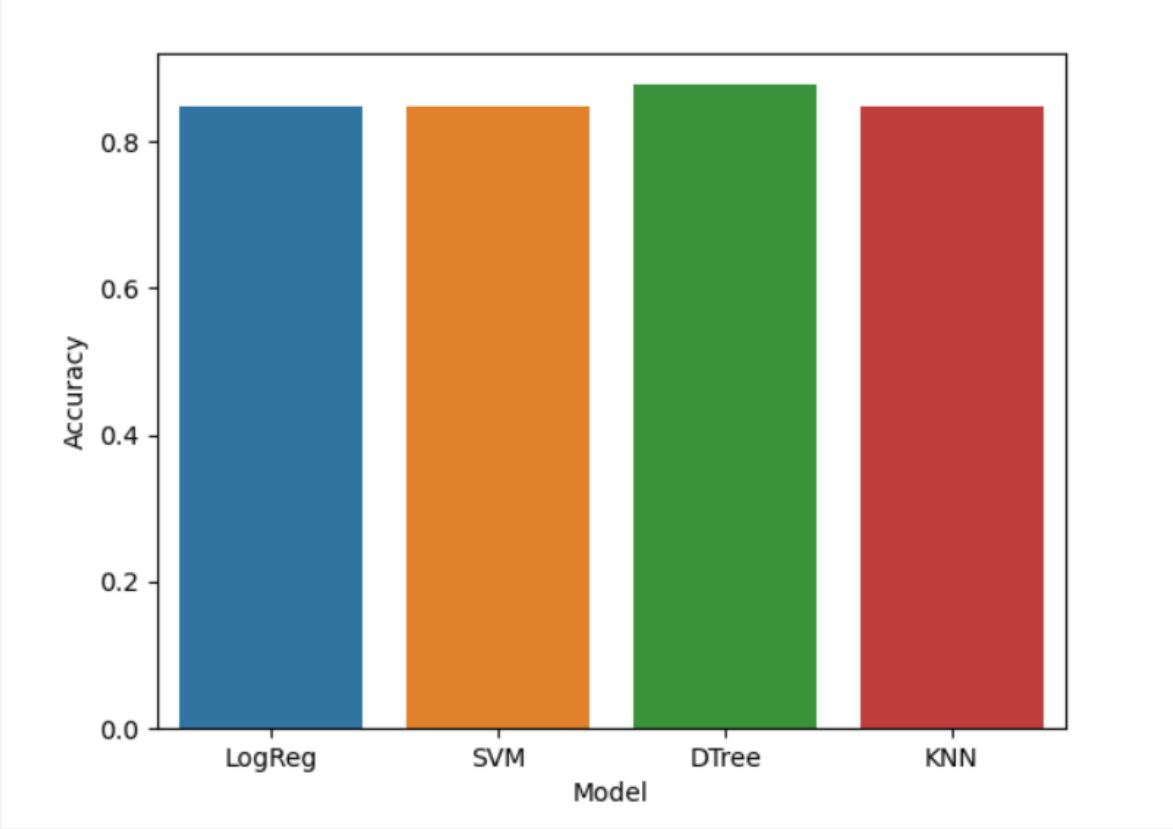
Predictive Analysis (Classification)

- Model development process



Classification Accuracy

- Model accuracy for all built classification models, in a bar chart



The best perform model was the Decision Tree Classifier with Accuracy of 0.8892857142857145

Which means it can predict the outcome correct 89% of times.

Confusion Matrix

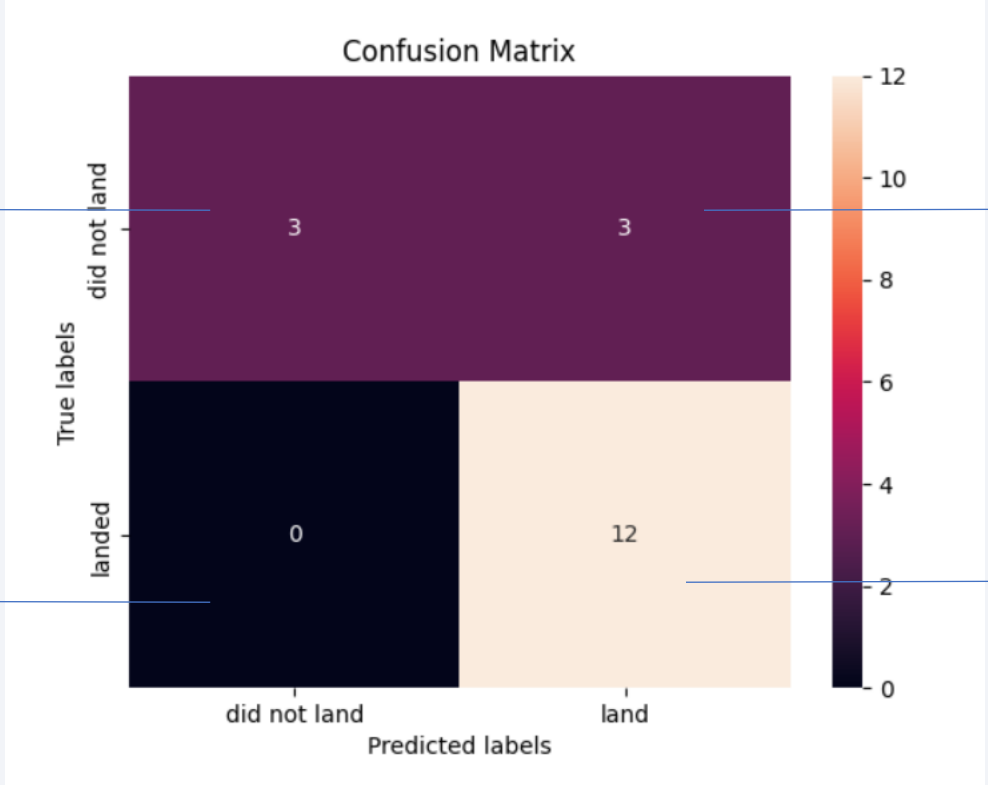
- Confusion matrix of the best performing model: Decision Tree Classifier

TN – True Negative

Predict that it would not land and actual it did not.

FN– False Negative

Predict that it would not land but it landed well.



FP– False Negative

Predict that it would land but it actual it did not.

TP– True Positive

Predict that it would land and it also actual landed well.

Section 8

Results

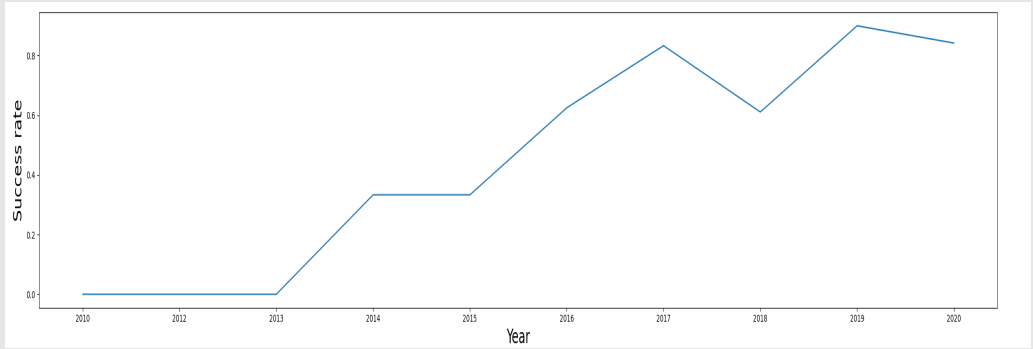
Exploratory Data Analysis Results

- Exploratory data analysis results

Company getting more experience

As INCREASES Flight Number than INCREASES Success rate

Success Rate Yearly Trend



20 launches reached in 2013

First successful landing outcome in ground pad was achieved in 2017

24 successful landings in 2020

Exploratory Data Analysis Results

Low Orbit Targets have greater success rate

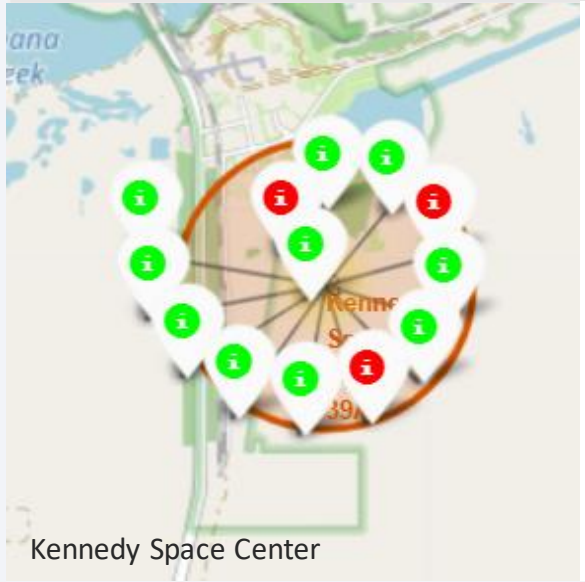
The principal customer of SpaceX was: **NASA with 17% of the payload mass carried**

The most successful booster version was: the **F9 v1.1**

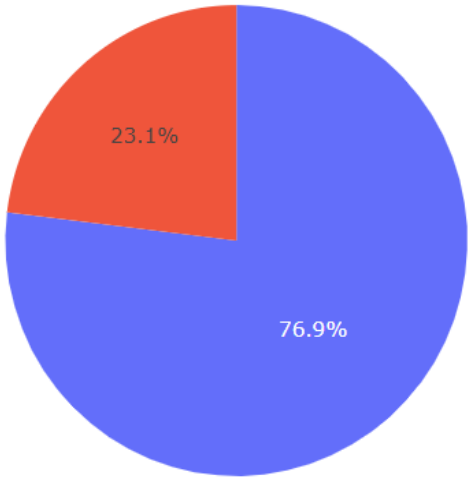
Payloads between 2.500 and 5.000Kg has much better success rate than other ranges.

Interactive Data Analysis Results

KSC LC-39A and VAFB SLC 4E
Launch Sites
have a highest success rate of 77%

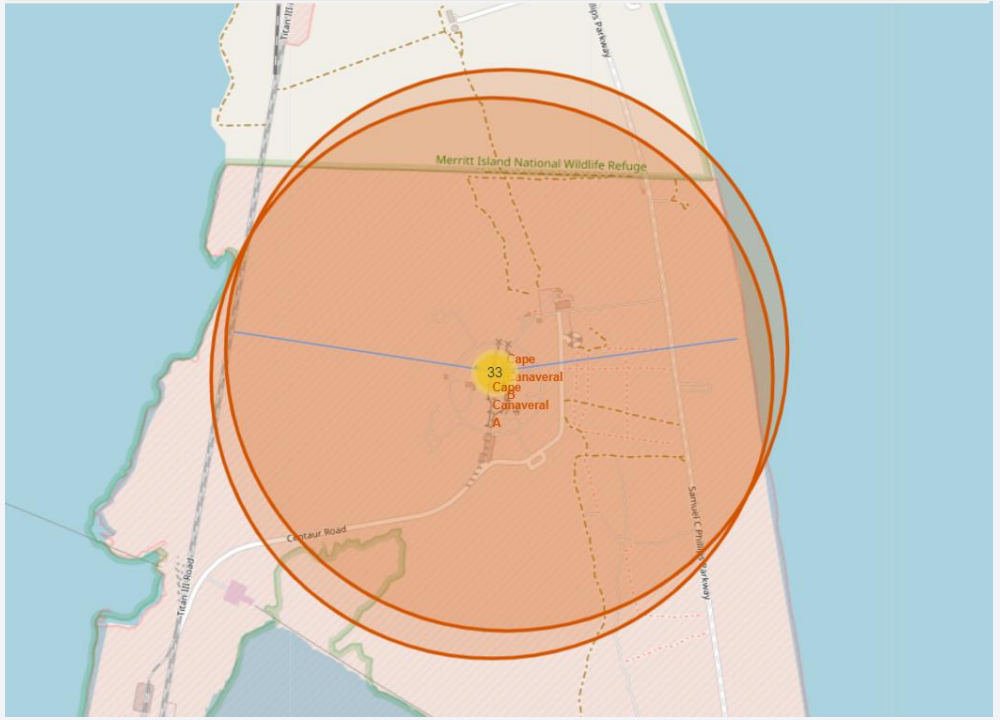


Kennedy Space Center

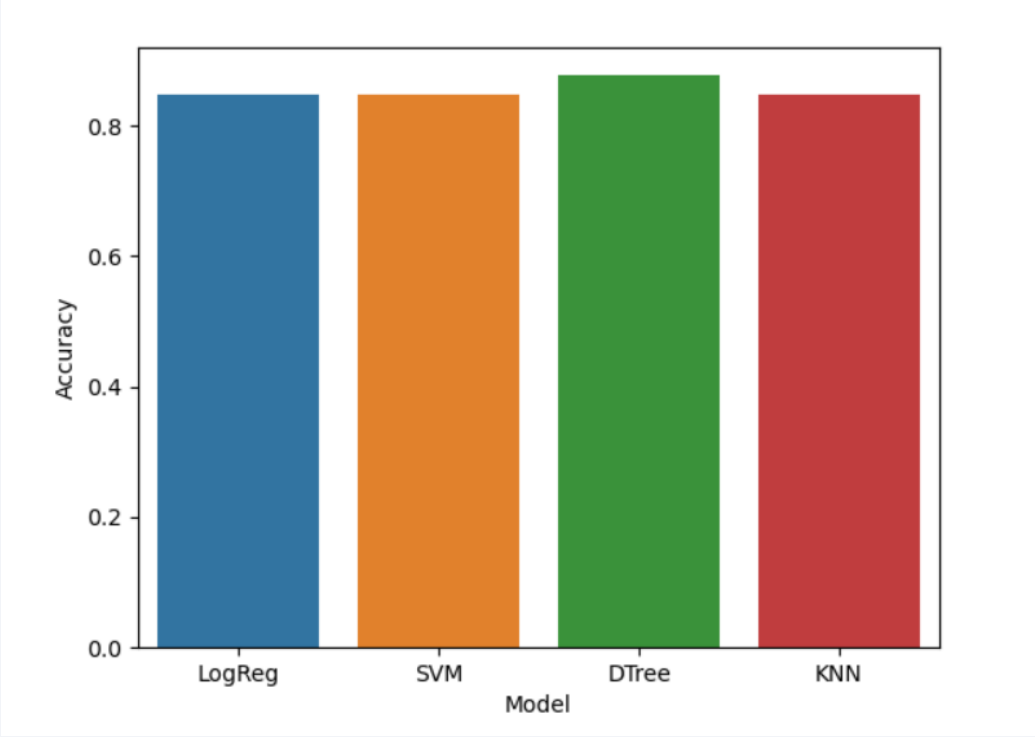


Kennedy Space Center

All launch sites are less than 1Km away from any railway, highway and the coast and at least 10Km away from nearest city.



Predictive Analysis Results



The best perform model was the Decision Tree Classifier with

Accuracy of 0.8892857142857145

Which means it can predict the outcome correct 89% of times.

Section 9

Conclusions

Conclusions

- We can **predict very well (89% accuracy) when a landing will be successful**,
- We can achieve that using the information about:
  - The Experience of the Company (flight number),
  - Characteristics of the Launch Sites,
  - Payload Mass,
  - Target Orbit and
  - Booster Version.
- We also can make recommendations to achieve greater successful rates as:
  - Use F9 v.11 booster version,
  - Launch from Kennedy Space Center of Vandenberg Launch Complex,
  - Target low orbit and,
  - Carry a Payload Mass between 2.500Kg and 5.000Kg,

Appendix

- Access to all relevant assets of this project:
  - Python code snippets,
  - SQL queries,
  - charts,
  - Notebook outputs,
  - data sets

GitHub url:
[My IBM-Applied-Data-Science-Capstone Repository]

Thank you!