DATA SCIENCE CAPSTONE FINAL ASSIGNMENT

Roberto Freitas Rodrigues

17/03/2022

OUTLINE

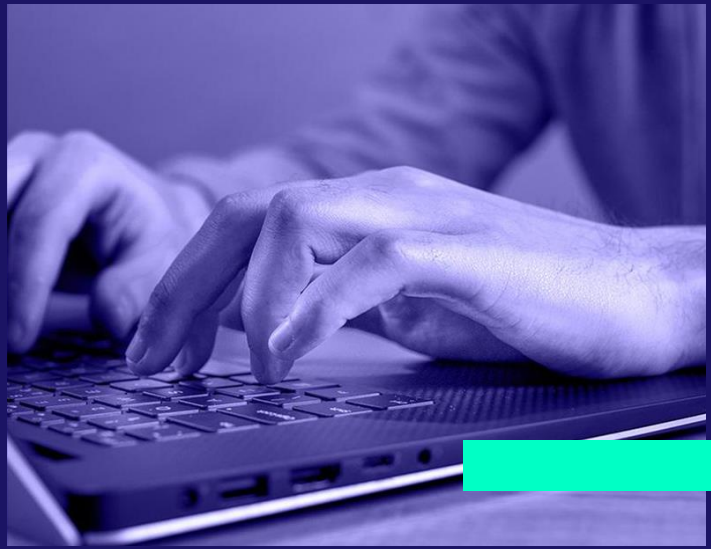
1) Executive Summary;
2) Introduction;
3) Methodology;
4) Results;
5) Conclusion;
6) Appendix

INTRODUCTION

SpaceX can send rockets to space without spending a lot of money. The company 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. In this project we will predict the successful landings of the first stage of Falcon 9 to evaluate the viability of a new company called SpaceY that wants to compete directly with SpaceX.

Therefore, we want to find some answers:

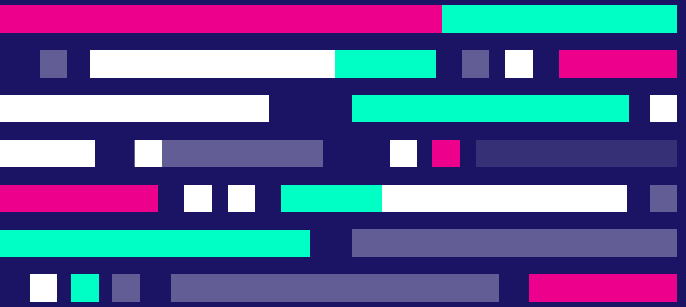
1) What are the factors to a successful landing?
2) What is the best location to make launches?



01

METHODOLOGY

METHODOLOGY



- Data colection methodology:
  - SpaceX Rest API
  - Web Scrapping from Wikipedia
- Perform data wrangling
  - Dropping irrelevante columns and NaN values
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Evaluation of the most accurate classification method

DATA COLLECTION

Data collection is the process of gathering, measuring, and analyzing accurate data from a variety of relevant sources to find answers to research problems, answer questions, evaluate outcomes, and forecast trends and probabilities.

Step 1

Acquire data from an API or a Web Page

Step 2

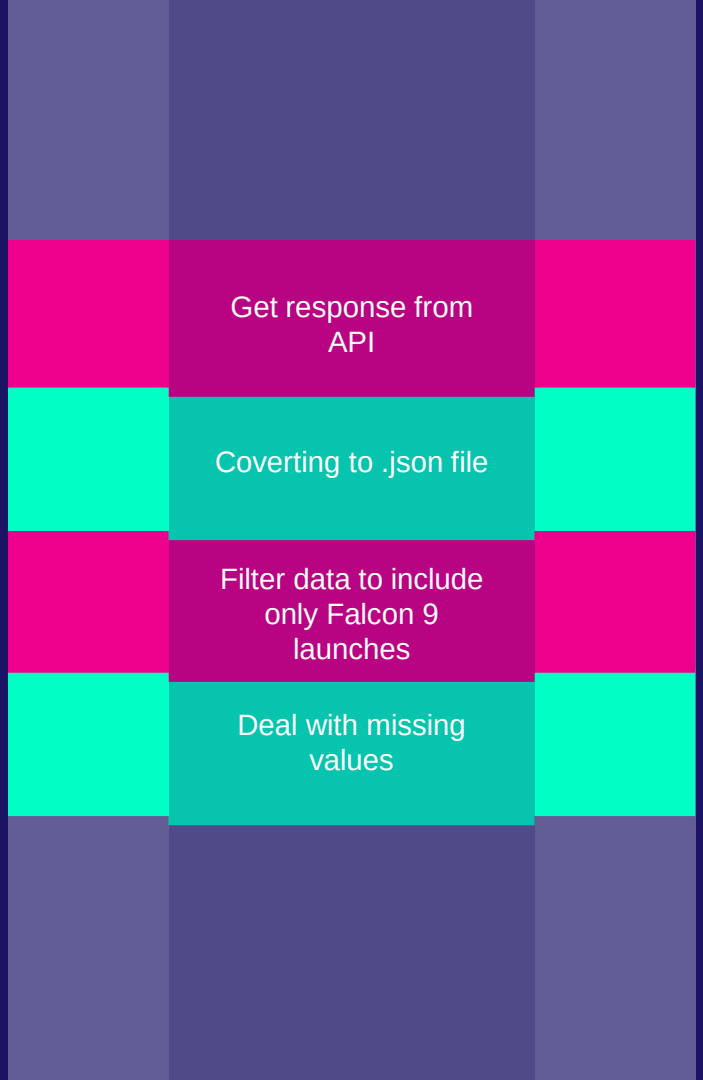
Transform data into a workable dataframe

Step 3

Eliminate undesirable values and columns

DATA COLLECTION – SpaceX API

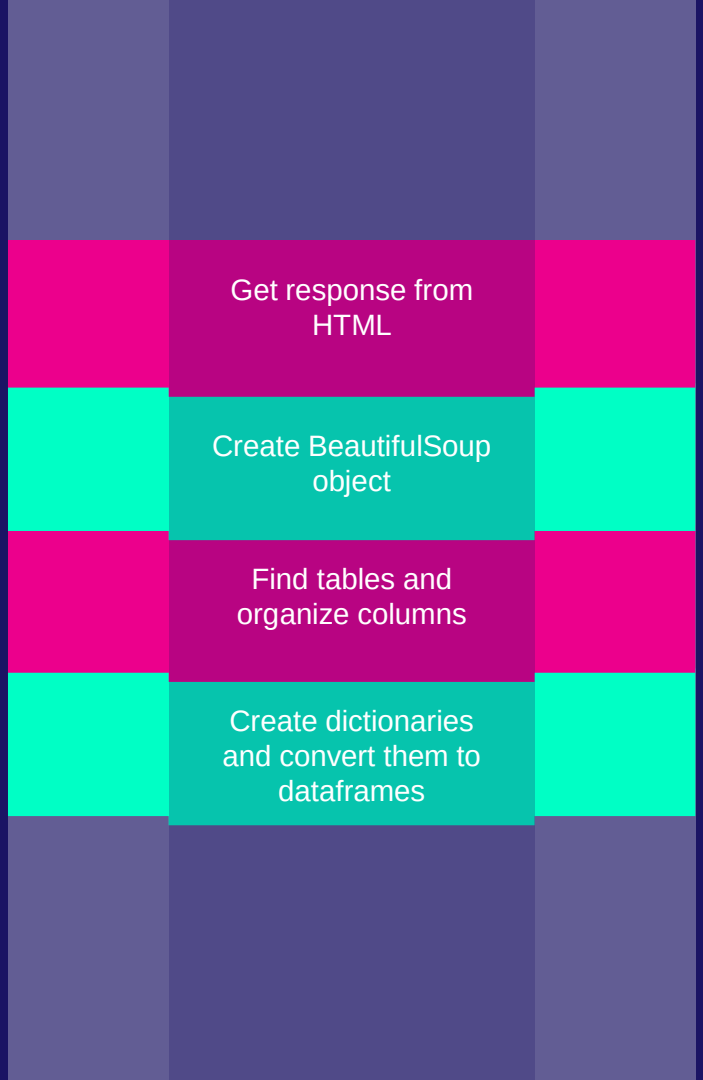
- Data can be collected via SpaceX public API;
- The flowchart indicates how the API was used to collect the data.



GitHub URL

DATA COLLECTION – WEB SCRAPING

- Data can be collected via Wikipedia;
- The flowchart indicates how the used to collect the data.

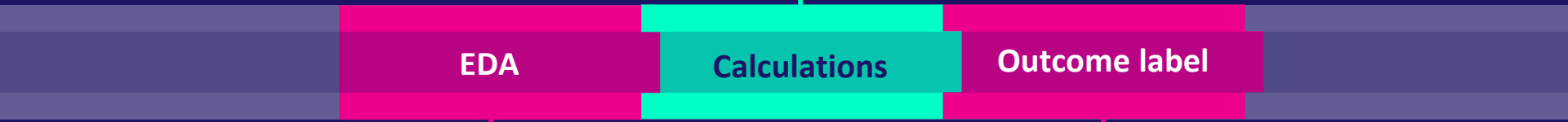


GitHub URL

DATA WRANGLING

Data wrangling is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics. The goal of data wrangling is to assure quality and useful data.

Were calculated launches per site, ocurrence of each orbit and ocurrences of missions outcome per orbit type



Was performed Exploratory Data Analysis (EDA) on the dataset

Finally, was created the “Landing Outcome” label from the outcome column.

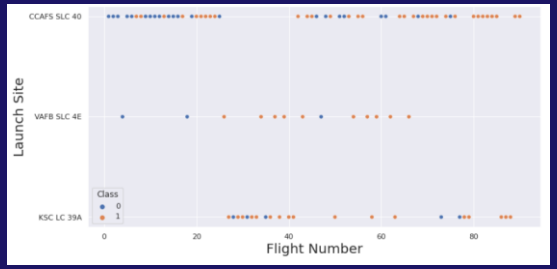
GitHub URL (Data wrangling)

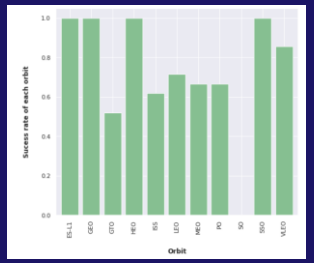
GitHub URL (EDA)

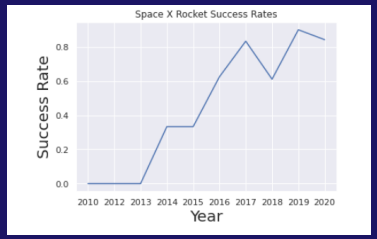
EDA WITH DATA VISUALIZATION

Data visualization

Were used scatterplots, barplots and lineplots to explore the data







GitHub URL

EDA WITH SQL

SQL stands for Structured Query Language which is basically a language used by databases. This language allows to handle the information using tables and shows a language to query these tables and other objects related (views, functions, procedures, etc.).

We performed the following SQL queries:

- 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 successful 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 failed landing_outcomes in drone ship, their booster versions, and launch site names for 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

GitHub URL

BUILD AN INTERACTIVE MAP WITH FOLIUM

Folium makes it easy to visualize data that's been manipulated in Python on an interactive leaflet map. It enables both the binding of data to a map for choropleth visualizations as well as passing rich vector/raster/HTML visualizations as markers on the map. Were used longitude and latitude coordinates of each launch sites with named labels of the sites. We also used markers to define success and failure launches. All the maps objects are defined below.

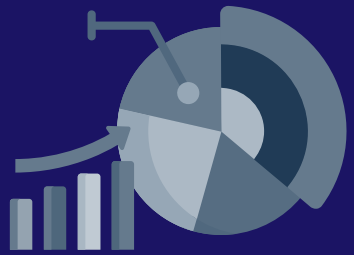
- Map Marker (folium.Marker()): Map object to make a mark on map
- Icon Marker (folium.Icon()): Create an icon on map
- Circle Marker (folium.Circle()): Create a circle where the Mark is being placed
- PolyLine(folium.PolyLine()): Create line between two or more points
- Marker Cluster Objetc(folium.MarkerCluster()): Simplify a map containing many markers with the same coordinate

GitHub URL

BUILD A DASHBOARD WITH PLOTLY DASH

Dash is a python framework created by plotly for creating interactive web applications. Dash is written on the top of Flask, Plotly. js and React. js. With Dash, you don't have to learn HTML, CSS and Javascript in order to create interactive dashboards, you only need python.

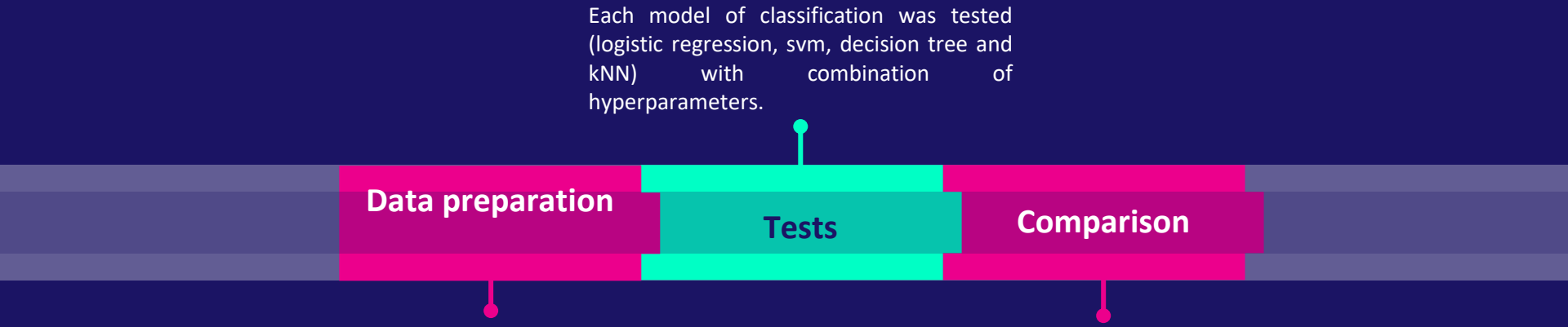
Were used graphs and plots to visualize percentage of launches per site and payload range. Therefore, we were able to analyze the relation between payloads and launch sites, identifying the best launch site.



GitHub URL

PREDICTIVE ANALYSIS (CLASSIFICATION)

Predictive analytics is the use of historical data, statistical algorithms, predictive modeling, and big data machine learning techniques to help organizations predict future outcomes more accurately, plan for unknown events, and discover opportunities in future activities. Were compared four classification models to find out whats was the best.



GitHub URL

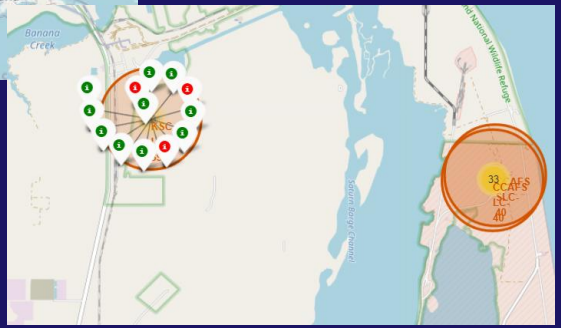
RESULTS

EDA results:

- SpaceX has 4 different launch sites;
- The first launches were also performed by NASA;
- The average payload of F9 v1.1 booster is 2.928kg;
- The first succesful landing ocurred in 2015
- Falcon 9 boosters version were successful at landing in drone ships, having payload above the average;
- Two booster version failed at landing in drone ships in 2015 (F9 v1.1 B1012 and F9 v1.1 B1015);
- The number of successful landings evolved better as the years passed;

RESULTS

Interactive analytics:



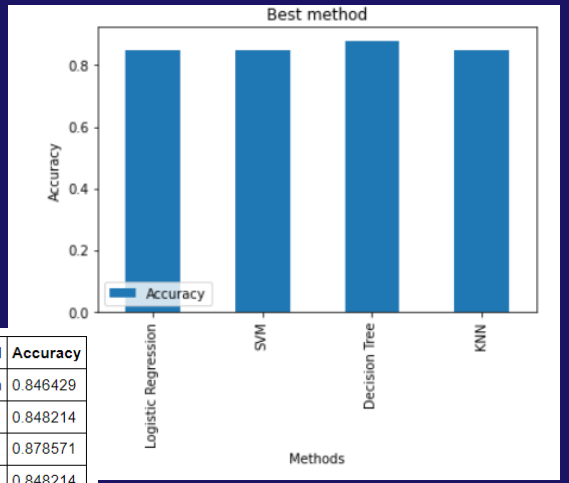
We were able to visualize and identify that the launch sites are commonly placed near the sea because of the safety. And were also capable of identify the successful and failure landings;

RESULTS

Predictive Analysis:

We were able to identify the best model to predict successful landings, having accuracy of 88%.

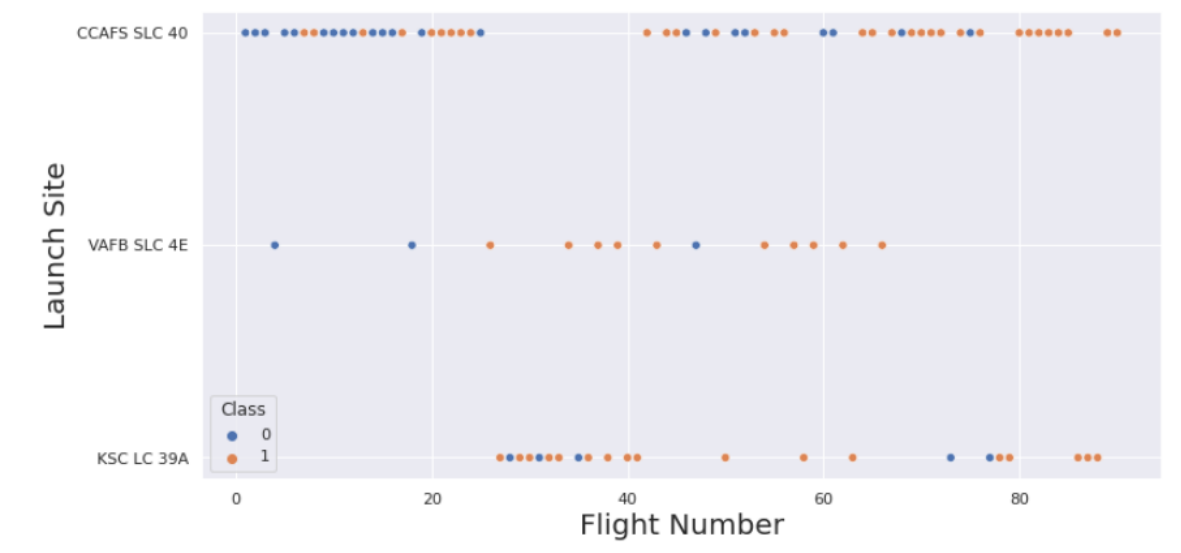
	Method	Accuracy
0	Logistic Regression	0.846429
1	SVM	0.848214
2	Decision Tree	0.878571
3	KNN	0.848214



02

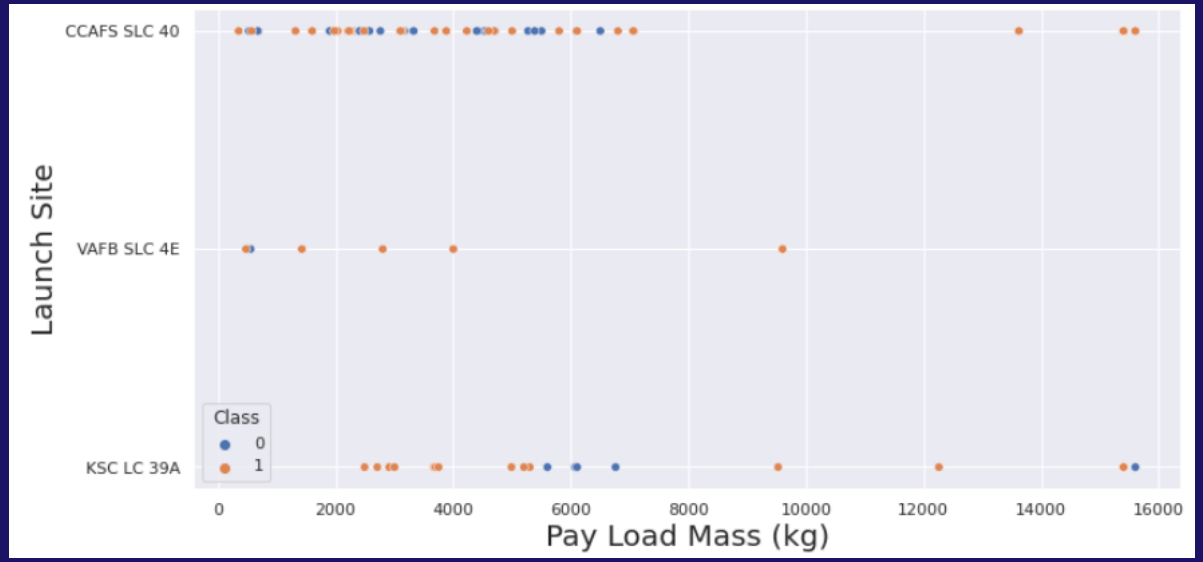
INSIGHTS DRAWN FROM EDA

Flight Number vs. Launch Site



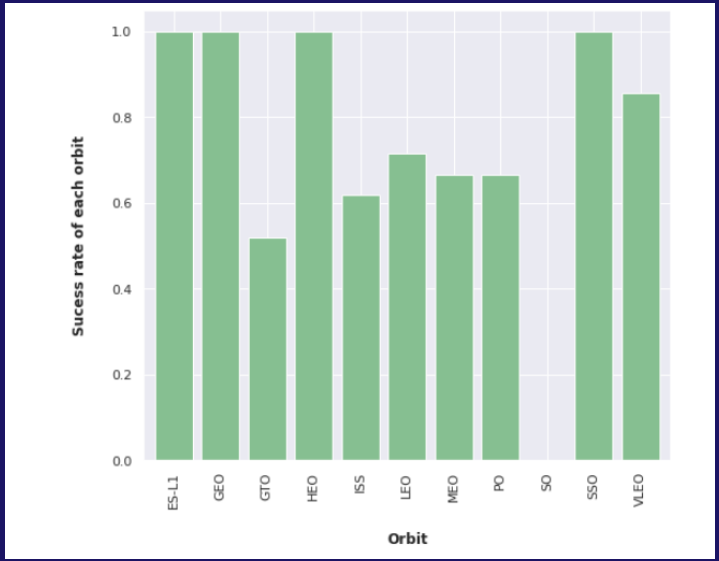
It's possible to identify that the success rate increase with more flight numbers with more flight numbers.

Payload vs. Launch Site



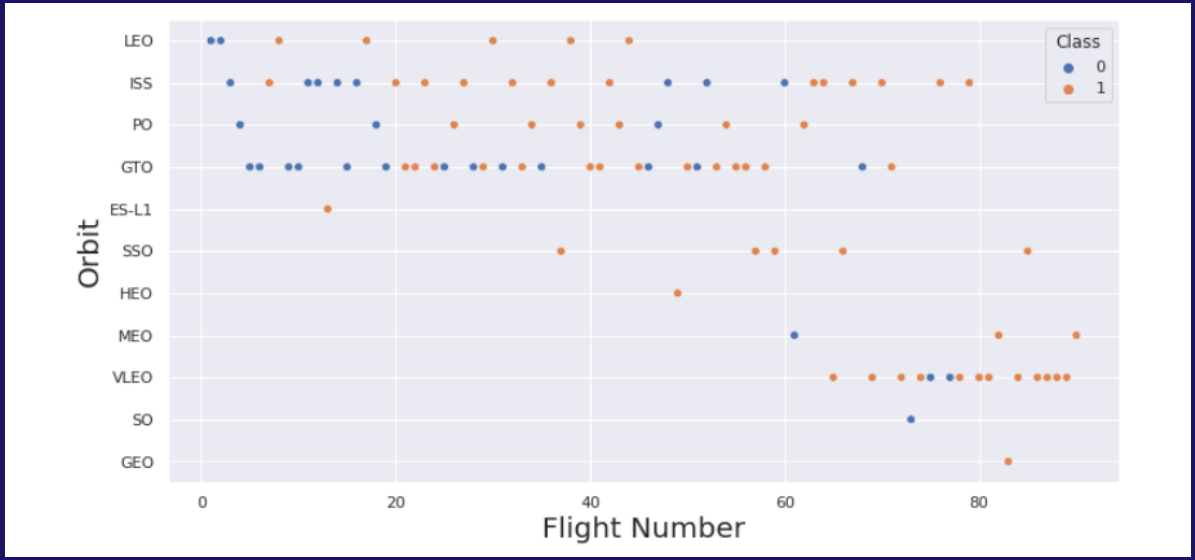
Payloads with mass greater than 8000 Kg have higher success rate.

Success Rate vs. Orbit Type



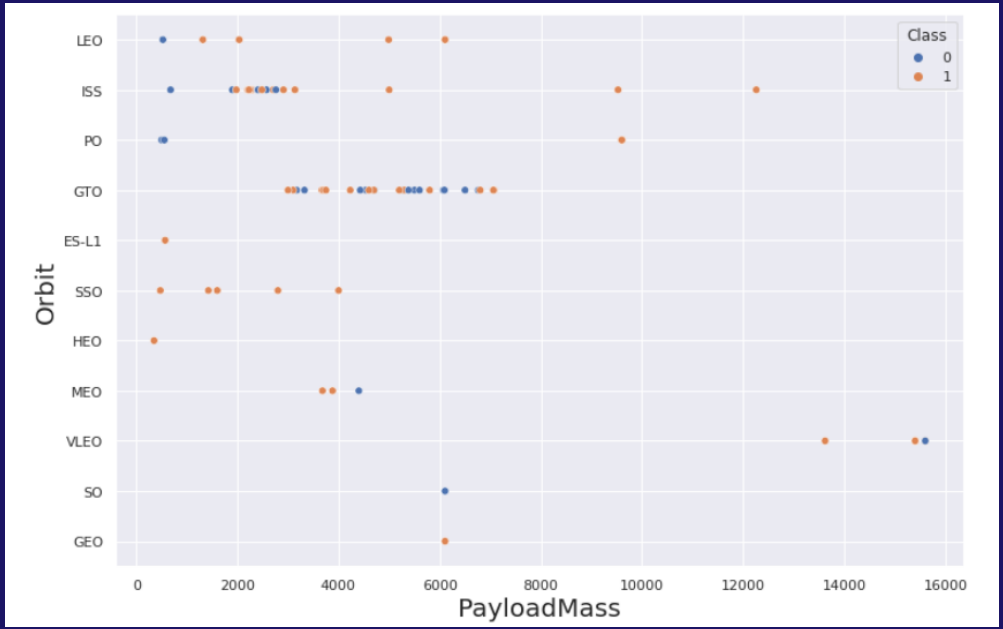
The orbits with highest success rate are: ES-L1, GEO, HEO and SSO.

Flight Number vs. Orbit type



The success rate improved over time.

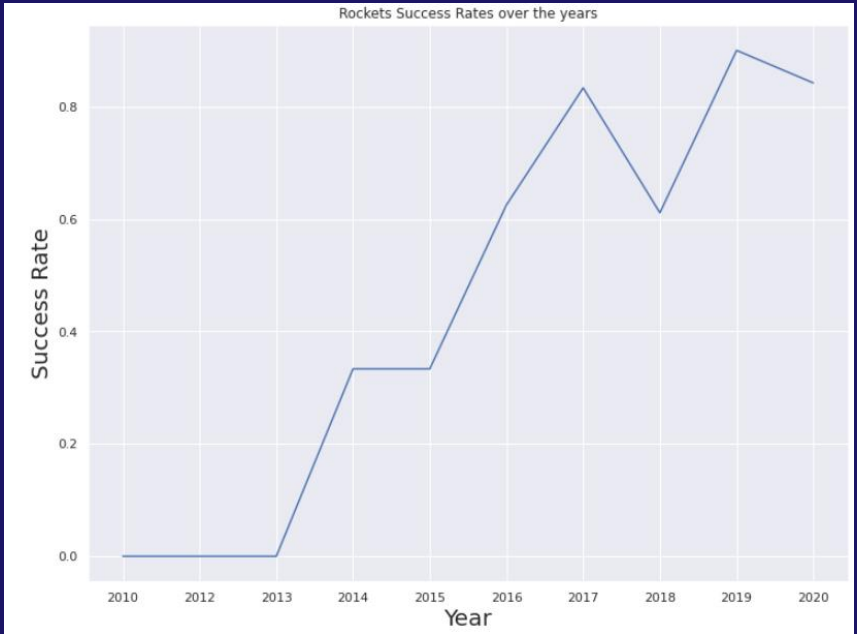
Payload vs. Orbit type



With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

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



Since 2013 the the success rate kept increasing until 2020.

03

EDA with SQL

All Launch Site Names

```
In [8]: %sql select distinct LAUNCH_SITE as "Launch_Sites" from SPACEXTBL;
```

In this query we pulled only non-repeating values for the “Launch Site” column using the command distinct.

Out[8]:

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

Launch Site Names Begin with 'CCA'

In [9]:
```
%sql select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' limit 5;
```

In this query we were capable to display only the records containing with “Launch Site” column containing ‘CCA’ and limited it to show only the first five records from the table “SPACEXTBL”.

Out[9]:

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

```
In [10]: %sql select SUM(PAYLOAD_MASS__KG_) as "Total Payload Mass by NASA (CRS)" from SPACEXTBL where CUSTOMER = 'NASA (CRS)';
```

In this query we used the “SUM” command to calculate the total Payload Mass (kg) fetching the costumer by name (NASA (CRS)).

Out[10]:

Total Payload Mass by NASA (CRS)
45596

Average Payload Mass by F9 v1.1

```
In [11]: %sql select AVG(PAYLOAD_MASS__KG_) as "Average Payload Mass by Booster Version F9 v1.1" from SPACEXTBL where BOOSTER_VERSION = 'F9 v1.1';
```

In this query we used the “AVG” command to calculate the average Payload Mass (kg) only for boosters “F9 v1.1”.

Out[11]:

Average Payload Mass by Booster Version F9 v1.1
2928

First Successful Ground Landing Date

```
In [12]: %sql select MIN(DATE) as "Successful Landing Outcome in Ground Pad" from SPACEXTBL where LANDING__OUTCOME = 'Success (ground pad)'
```

In this query we used the “MIN” command to show the minimun (or first) date where the “Landing Outcome” was a “Success (ground pad)”.

Out[12]:

Successful Landing Outcome in Ground Pad
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
In [13]: %sql select BOOSTER_VERSION from SPACEXTBL where LANDING__OUTCOME = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000
```

In this query we used two conditions to filter the results. The first was that the “Landing Outcome” should be a successful one and the Payload Mass (kg) should be between 4000 kg and 6000kg

Out[13]:

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

Total Number of Successful and Failure Mission Outcomes

In [21]:
```
%sql select COUNT(MISSION_OUTCOME) as "Successful and Failure Missions" from SPACEXTBL \
where MISSION_OUTCOME like 'Success%' or MISSION_OUTCOME like 'Failure%';
```

In this query we used calculated the total number of successful and failure missions using only the “Mission Outcome” with ‘Success’ and ‘Failure’ words, totalizing 101.

Out[21]:

Successful and Failure Missions
101

Boosters Carried Maximum Payload

In [22]:
```
%sql select distinct BOOSTER_VERSION as "booster versions which have carried the maximum payload mass" from SPACEXTBL \
where PAYLOAD_MASS__KG_ = (select MAX(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

In this query we used the DISTINCT command to find unique booster version which had the maximum “Payload Mass”

booster versions which have carried the maximum payload mass
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

In [23]:
```
%sql select MONTH(DATE) as "Month" , BOOSTER_VERSION, LAUNCH_SITE from SPACEXTBL \
where year(DATE)='2015' and LANDING__OUTCOME='Failure (drone ship)'
```

In this query we used the specified the month using the command MONTH() to show the 2015 Launch Records

Out[23]:

Month	booster_version	launch_site
1	F9 v1.1 B1012	CCAFS LC-40
4	F9 v1.1 B1015	CCAFS LC-40

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

In [24]:
```
%sql select LANDING__OUTCOME as "Landing Outcome", COUNT(LANDING__OUTCOME) as "Total" from SPACEXTBL \
where DATE between '2010-06-04' and '2017-03-20'\
group by LANDING__OUTCOME order by count(LANDING__OUTCOME) desc;
```

In this query we used the used only the dates between 2010-06-04 and 2017-03-20 and ranked it using the “desc” command.

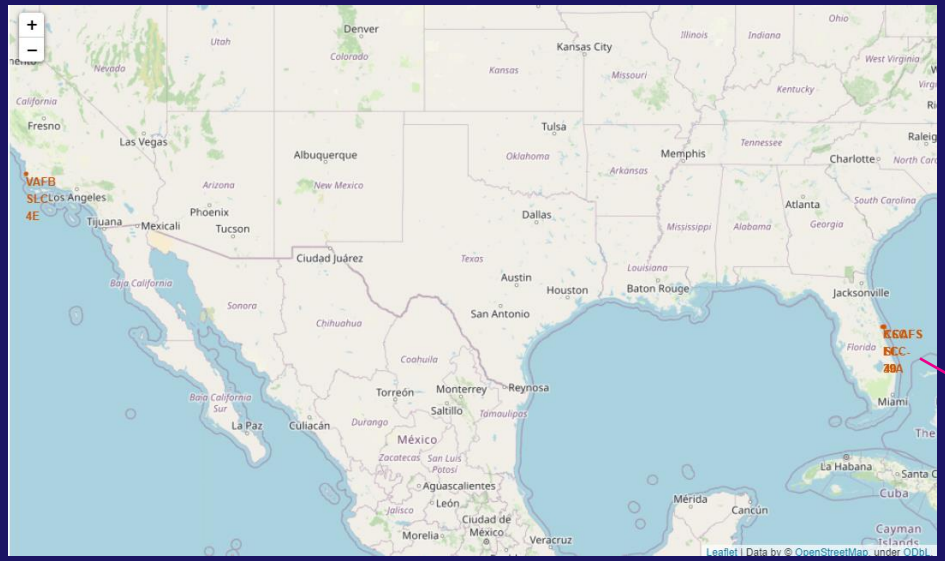
Out[24]:

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

04

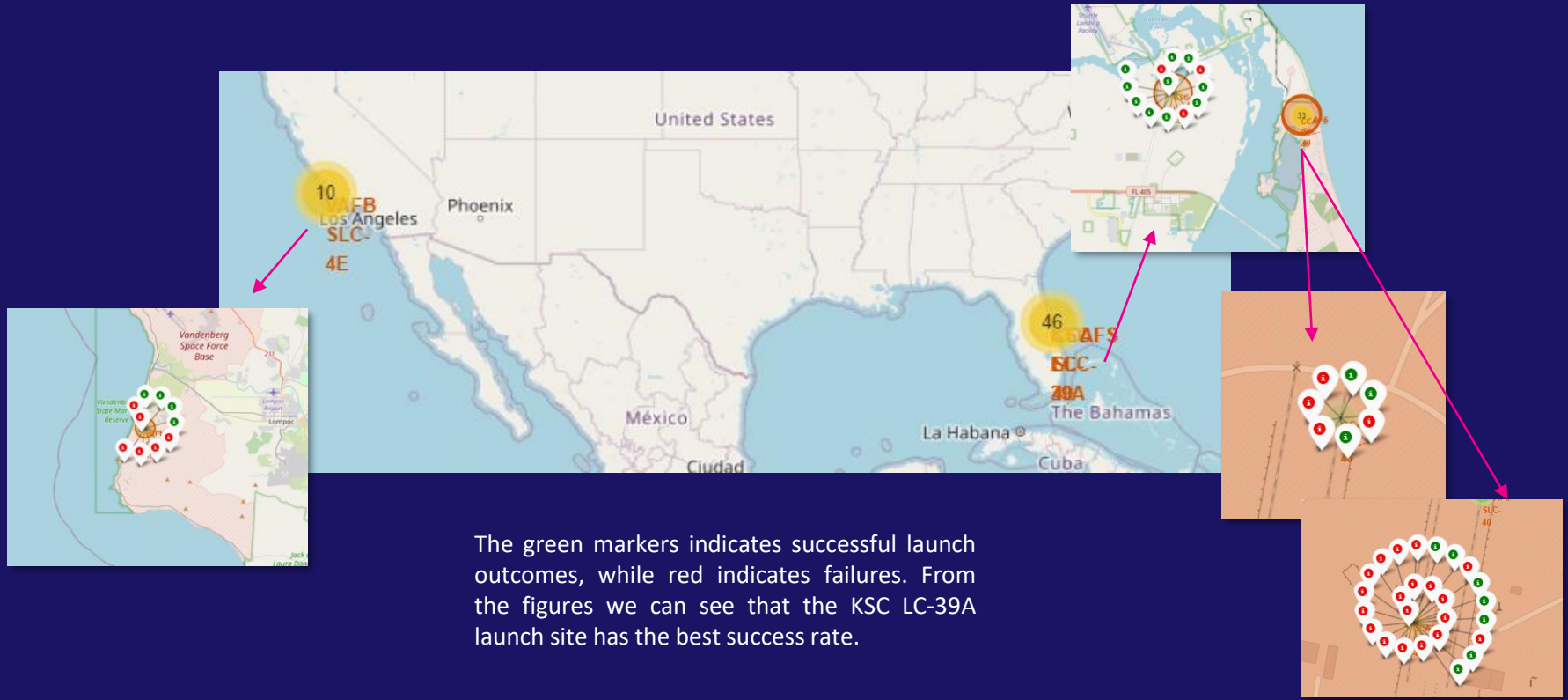
Launch Sites Proximities Analysis

Launch Sites



We can see the VAFB SLC-4E, KSC LC-39A and CCAFS SLC-40 launch sites, all above the Equator Line. We observe that they are all near the coast, maybe for safety.

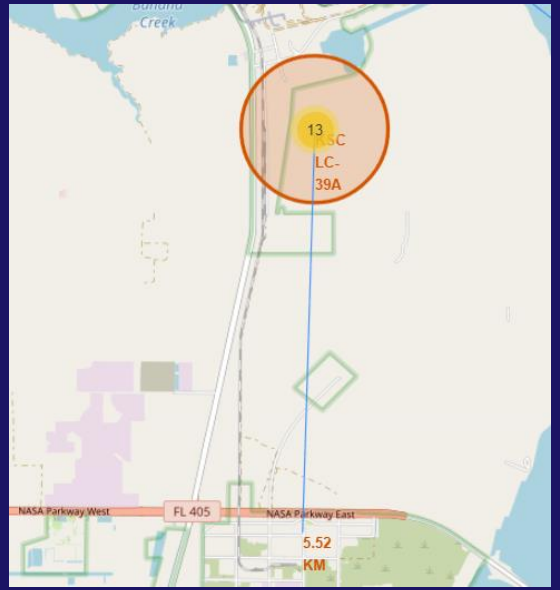
Launch Outcomes



The green markers indicates successful launch outcomes, while red indicates failures. From the figures we can see that the KSC LC-39A launch site has the best success rate.

Distance between launch site to its proximities

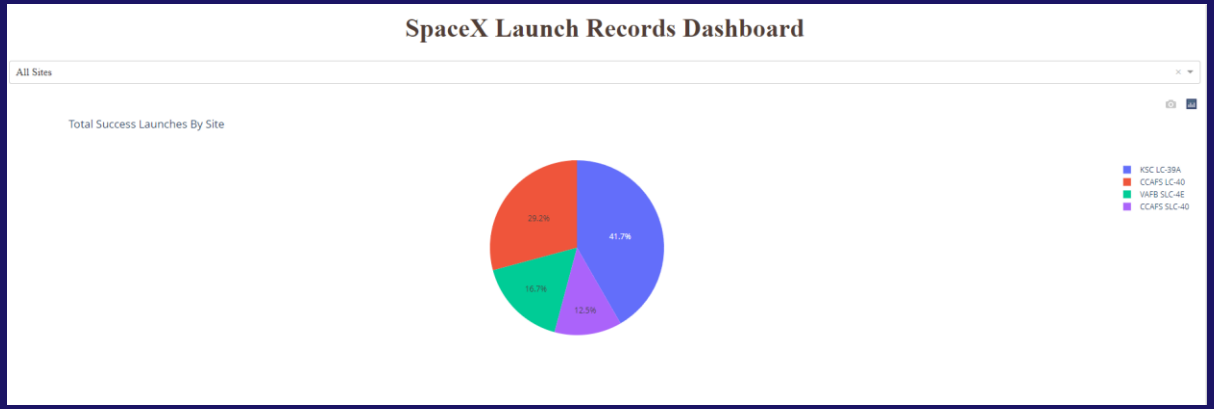
We have the example of the KSC LC-39A Launch Site, that is 5.52 km away of a road. It turns out to be a safe place to do rocket launches.



06

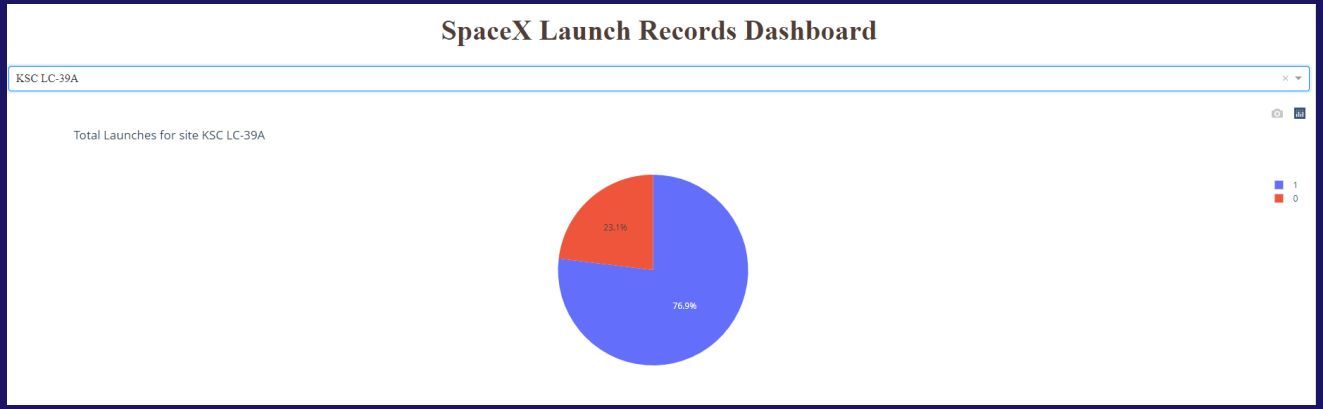
Build a Dashboard with Plotly Dash

Total Success Launches by Site



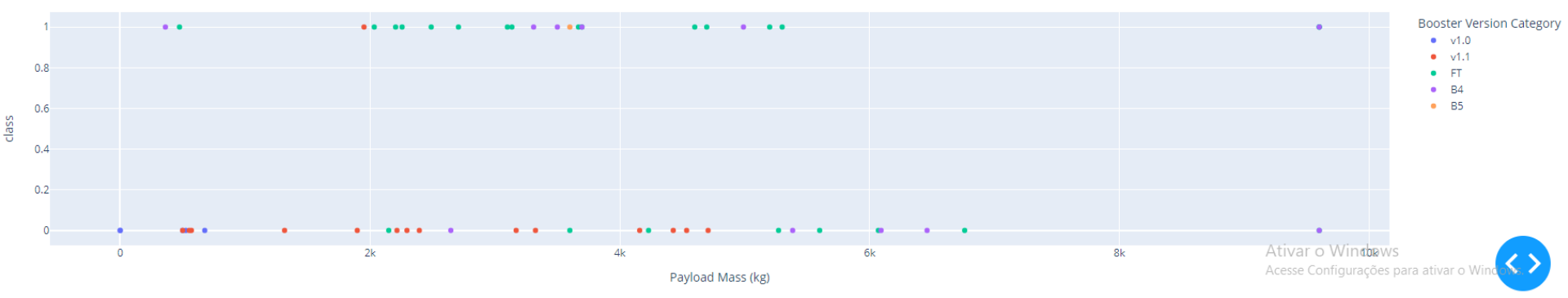
We see that the KSC LC-39A seems to be the most successful launch site. It seems that this launch site impacts on the success of the SpaceX missions.

Total Success Launches by Site



We confirm that the KSC LC-39A has the most successful missions, reaching more then 75%.

Total Success Launches by Site



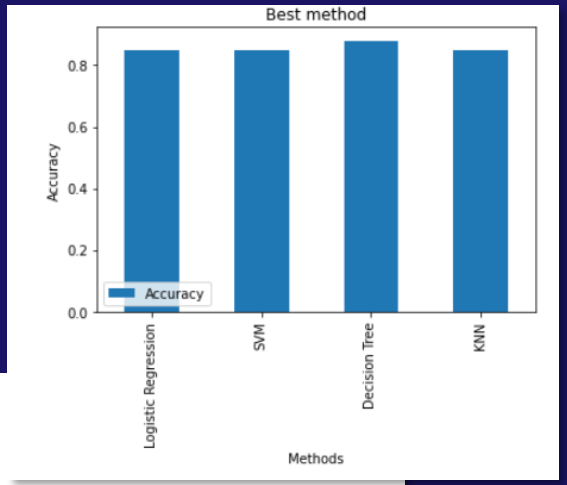
With the scatterplot we observe that the FT boosters are the most successful of all versions.

07

Predictive Analysis (Classification)

Classification Accuracy

- We utilized four different classification methods: Logistic Regression, SVM, Decision Tree, kNN.
- The method that showed the best accuracy was the Decision Tree method with accuracy of 88%.

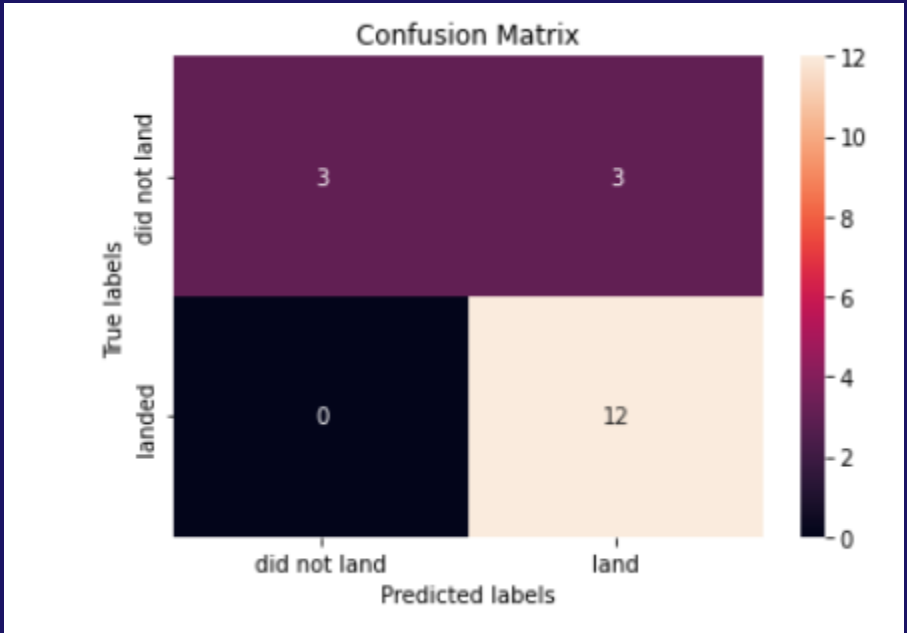



```
The Decision Tree method performs best with accuracy of 0.8785714285714284
```

Out[47]:

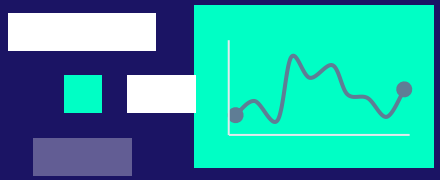
	Method	Accuracy
0	Logistic Regression	0.846429
1	SVM	0.848214
2	Decision Tree	0.878571
3	KNN	0.848214

Confusion Matrix of Decision Tree



Unfortunately, all confusion matrixes were the same because all accuracies are close to each other.

Conclusions

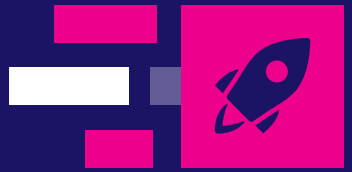


Launches increasing

The successful launches have increased with time since 2013, and can soon or later reach the required target;

Orbits

The orbits ES-L1, GEO, HEO and SSO have the highest success rates;

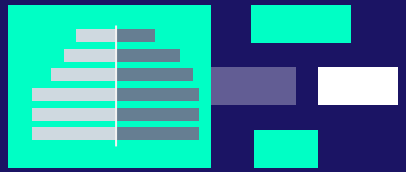


Best launch site

The best launch site KSC LC-39A;

Classification

The best classification method for this dataset is the Decision Tree Method.



Appendix

- The Jupyter notebooks from all the data shown here are on this GitHub: https://github.com/robfreitas96/Applied-Data-Science-Capstone;
- For security purposes the Db2 service credentials were omitted in the EDA with SQL laboratory.
- Plotly Dash code and some screenshots of the app are also in Github.

THANKS!

CREDITS: This presentation template was created by Slidesgo, incluiding icons by Flaticon, and infographics & images by Freepik.