IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

João Maçãs
December 2021

# Outline

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

# Executive Summary

- Summary of methodologies
  - In this project it were applied methodologies covering: a) Data collect from different sources and Data wrangling; b) Exploratory Data analysis using visualization techniques and SQL; c) Analysis via interactive visual analytics using Folium (Proximity Analysis) and Plotly Dash; d) Predictive analysis using Classification models
- Summary of all results
  - From the analysis to the data obtained we concluded that as launches attempts were increasing so the success rate was increasing. The site KSC LC-39A (John F. Kennedy Space Center, Florida - Launch Complex 39A) is the one with more successful launches. The classification model to predict the success of a launch present an accuracy of ~83,33%.

# Introduction

- Project background and context
  - I've been hired, as Data Scientist, by a private rocket company named Space Y. Space Y is interested in competing with SpaceX. One information my company want to find is the price of each launch. One approach for this target is to evaluate the data SpaceX announces on the internet. There one can find that SpaceX advertises Falcon 9 rocket launches have a cost of 62 million dollars. When looking for other providers advertisements the cost goes upward of 165 million dollars each. Much of the savings in SpaceX is because they can reuse the first stage of their rockets.
- Problems you want to find answers
  - The problem statement in this project is if we can determine if the first stage from Falcon 9 will land, as this way we would be able to determine the cost of a launch.
  - We'll use machine learning model and public information from SpaceX site to determine if Falcon 9 stage one will land, i.e., to predict the rate of first stage reuse.

Section 1

# Methodology

# Methodology

## Executive Summary

- Data collection methodology
  - Data collected via SpaceX REST API and via Web Scraping related Wiki pages
- Perform data wrangling
  - Sampling, dealing with Nulls and set Labels for training models
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Classification models built and tuned using GridSearchCV. Evaluation done with Score and Confusion Matrix.

# Data Collection

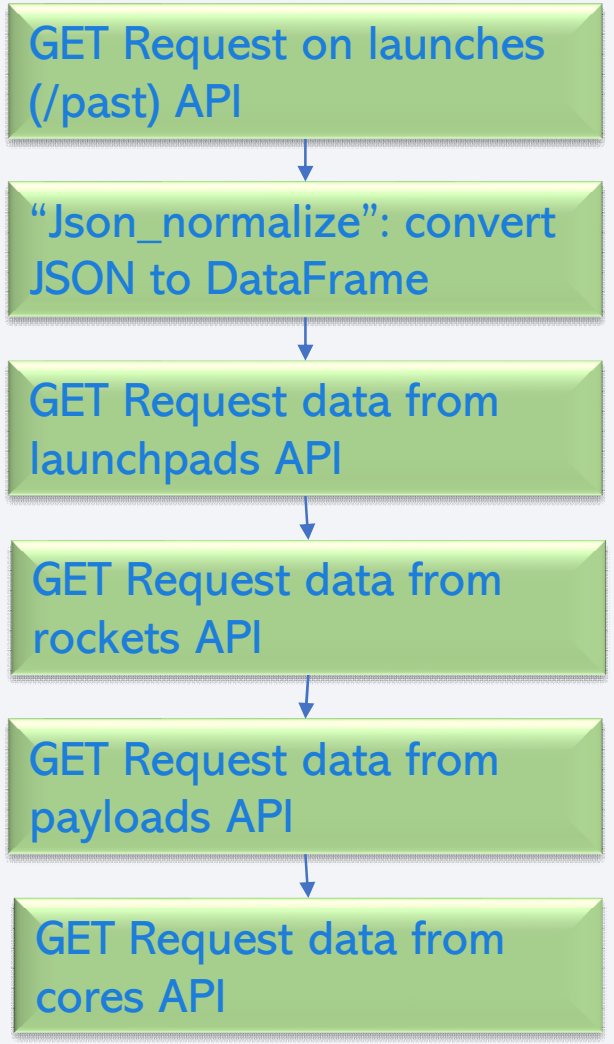
- In this project 2 approaches were used to collect data sets
  - Via an API put available by SpaceX, namely the **SpaceX REST API**. API offers different endpoints and for this project, the main information come from “endpoint api.spacexdata.com/v4/launches/past”, which provides info on past launches. Other endpoints were used to extract info on rockets, launchpad (to learn which sites are used and their geo position), payload (to learn the mass of the payload and the orbit that it is going to) and cores (to learn the outcome of the landing, the type of the landing, number of flights with that core, among many other data on cores)
  - By **Web Scraping** related Wiki pages. To do that it was used the Python BeautifulSoup package, which allow to web scrape some HTML tables that contain Falcon 9 launch records.

# Data Collection – SpaceX API

- SpaceX launch data is obtained by sending a “get request” message to API, using “requests” library. The response was in the form of a JSON, specifically a list of JSON objects (each object represent a launch). To convert the structured JSON data result to a DataFrame, it was used the json_normalize function to “normalize” the structured data into a flat table.
- REST API callings done:

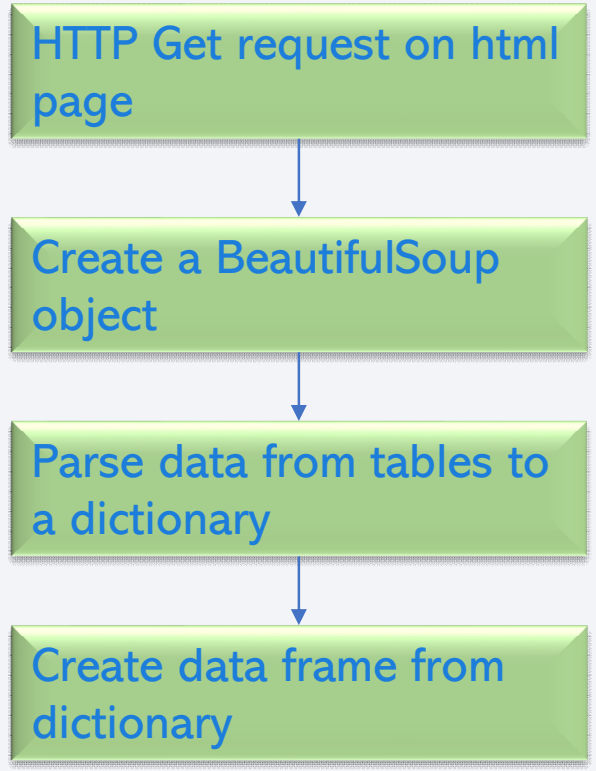
  Perform a GET request to the launches API
  Perform a GET request to the rockets API
  Perform a GET request to the launchpads API
  Perform a GET request to the payloads API
  Perform a GET request to the cores API

- GitHub URL of the completed SpaceX API calls notebook:
  Coursera_Capstone/1.1 - Data Collection API.ipynb at master · JoaoMacas/Coursera_Capstone · GitHub



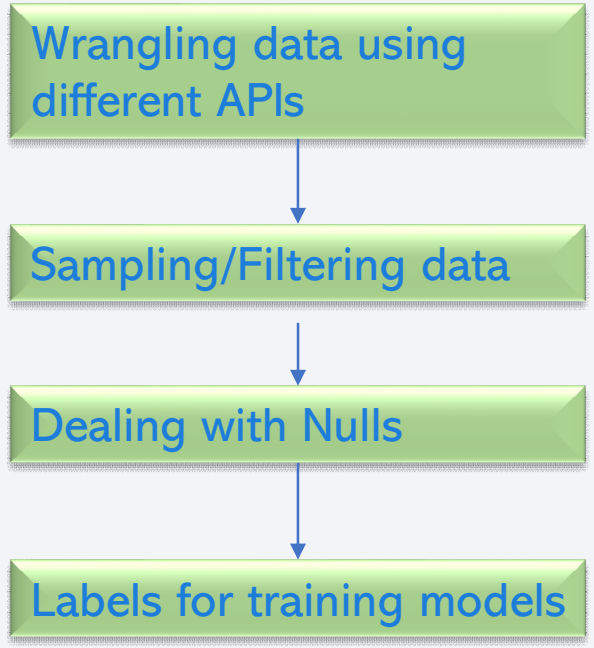
# Data Collection - Scraping

- Web scraping data process and respective data processing consisted on:
  - Perform an HTTP GET method to request the Falcon9 Launch HTML page, as an HTTP response.
  - Create a BeautifulSoup object from the HTML response
  - Parse the data from tables and convert them into a dictionary
  - Create a data frame from dictionary, for further visualization and analysis.
- GitHub URL of the completed web scraping notebook:
  Coursera_Capstone/1.2 - Data Collection with Web Scraping.ipynb at master · JoaoMacas/Coursera_Capstone · GitHub



# Data Wrangling

- Some processing was done on raw data to transform it into a clean and more meaningful data. It was applied:
  - Wrangling Data using several different API
    - As the raw data obtained was showing ID for different rocket features it was used another API endpoint (see previous slide to have list of other endpoint API accessed) to get specific data associated with each ID. All the processed data was stored in lists and was then used to create dataset used.
  - Sampling Data
    - For the project, only the data on Falcon 9 is needed, and because raw data has other falcons included one have applied filtering/sampling actions to have only Falcon 9 data.
  - Dealing with Nulls
    - Some data contains NULL values. In order to make the dataset viable for analysis we have dealt with the null values inside the PayloadMass. Null values have been replaced by the the mean of the PayloadMass data.
  - Labels for training models
    - The data set obtained had several different cases where the booster did not land successfully. We wanted to mark all those cases in a single manner, i.e., mark the outcome for that launch as failed (0 value). In the same way all successful launches are marked with 1.
- GitHub URL of completed data wrangling related notebooks: Coursera_Capstone/1.1 - Data Collection API.ipynb at master · JoaoMacas/Coursera_Capstone · GitHub and Coursera_Capstone/1.3 - EDA - Data Wrangling.ipynb at master · JoaoMacas/Coursera_Capstone · GitHub



# EDA with Data Visualization

- A set of scatter points charts were plotted (see list below), to check the contribution from each variable (feature) for the success of the launch. Goal is to select the features that will be used in success prediction in the future module.
  - a) the FlightNumber vs. PayloadMass; b) the FlightNumber vs LaunchSite; c) the Payload vs Launch Site; d) the FlightNumber vs Orbit and e) the Payload vs. Orbit
  - All these charts were done with Class overlay.
- We also plot a bar chart for the success rate of each orbit. Goal is to verify if there are any relationship between success rate and orbit type.
- Finally, we also plot a line chart showing the average success rate over the Years. Goal is to see the average launch success trend overs the years, and, in fact, we see that the success rate since 2013 kept increasing till 2020.
- GitHub URL completed EDA with data visualization notebook:
  Coursera_Capstone/2.2 - EDA with Data Visualization.ipynb at master · JoaoMacas/Coursera_Capstone · GitHub

# EDA with SQL

- Some SQL queries were performed for exploratory data analysis:
  - 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 achieved
  - 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.
  - 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 of completed EDA with SQL notebook: Coursera_Capstone/2.1 - EDA with SQL.ipynb at master · JoaoMacas/Coursera_Capstone · GitHub

# Build an Interactive Map with Folium

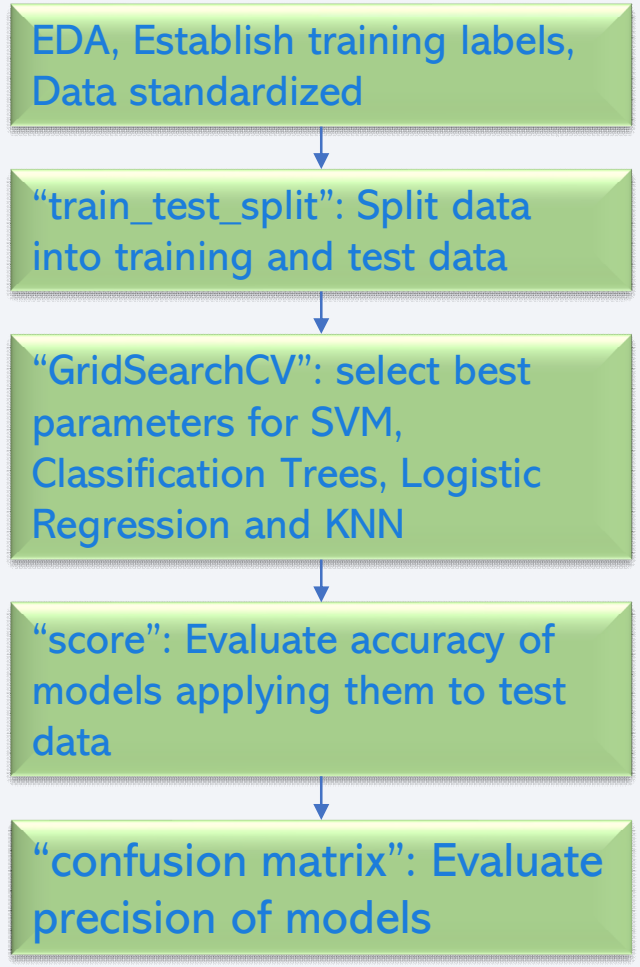
- To explore if launch success rate was dependent on the location of the launch, and its proximities, interactive Map with Folium were built.
- It was built a map of the United States centred on the location to the NASA Johnson Space Centre at Houston, Texas.
- In the maps it was marked, using Folium map markers and circles, a) all launch sites and b) the success/failed launches for each site on the map.
- In the maps it were marked lines, together with the distances, between a launch site and its relevant buildings/infrastructures in the proximities.
- These objects were added to maps to understand the locations and their surroundings, in an attempt to identify location success factors for launching space rockets
- GitHub URL of completed interactive map with Folium map: Coursera_Capstone/3.1 - Interactive Visual Analytics with Folium.ipynb at master · JoaoMacas/Coursera_Capstone · GitHub

# Build a Dashboard with Plotly Dash

- Using a Plotly Dash it was built a dashboard with a pie chart presenting the success rate of launches in all and in each site. Values are seen using a dropdown menu (interaction item) to select a site/all sites. In same dashboard it was available a scatter chart to see the correlation between Payload and Success in each site. Selection of payload range values is done using a range slider (interaction item).
- With the dashboard created one can get info on which site has the largest successful launches, which site has the highest launch success rate, which payload has the highest/lowest launch success rate, and which f9 booster version has highest launch success rate.
- GitHub URL of completed Plotly Dash lab: Coursera_Capstone/3.2 - Spacex Plotly Dashsboard.ipynb at master · JoaoMacas/Coursera_Capstone (github.com)

# Predictive Analysis (Classification)

- A machine learning pipeline to predict if the first stage will land was created. These predictions were based on a classification model.
- As first step to evaluate the classification models some preparations were done, namely: a) performed an Exploratory Data Analysis; b) establish the training labels by identify the feature used for data classification (namely the Class column: 1 – successful launch; 0 – unsuccessful launch); c) data was standardized using "Transform" and "Fit" of StandardScaler ; d) split the data into training data and test data, using the function "train_test_split".
- Then to evaluate and select the best parameters and accuracy for the different classification models it was applied hyperparameters (using the function GridSearchCV) for SVM, for Classification Trees, for Logistic Regression and for KNN.
- Finally, the models were evaluated (i.e., checked their accuracy) to find the best one by applying them to the test data and using "score" function, and to check their precision by plotting and observing the confusion matrix.
- GitHub URL of completed predictive analysis lab: Coursera_Capstone/4.1 - Machine Learning Prediction.ipynb at master · JoaoMacas/Coursera_Capstone (github.com)

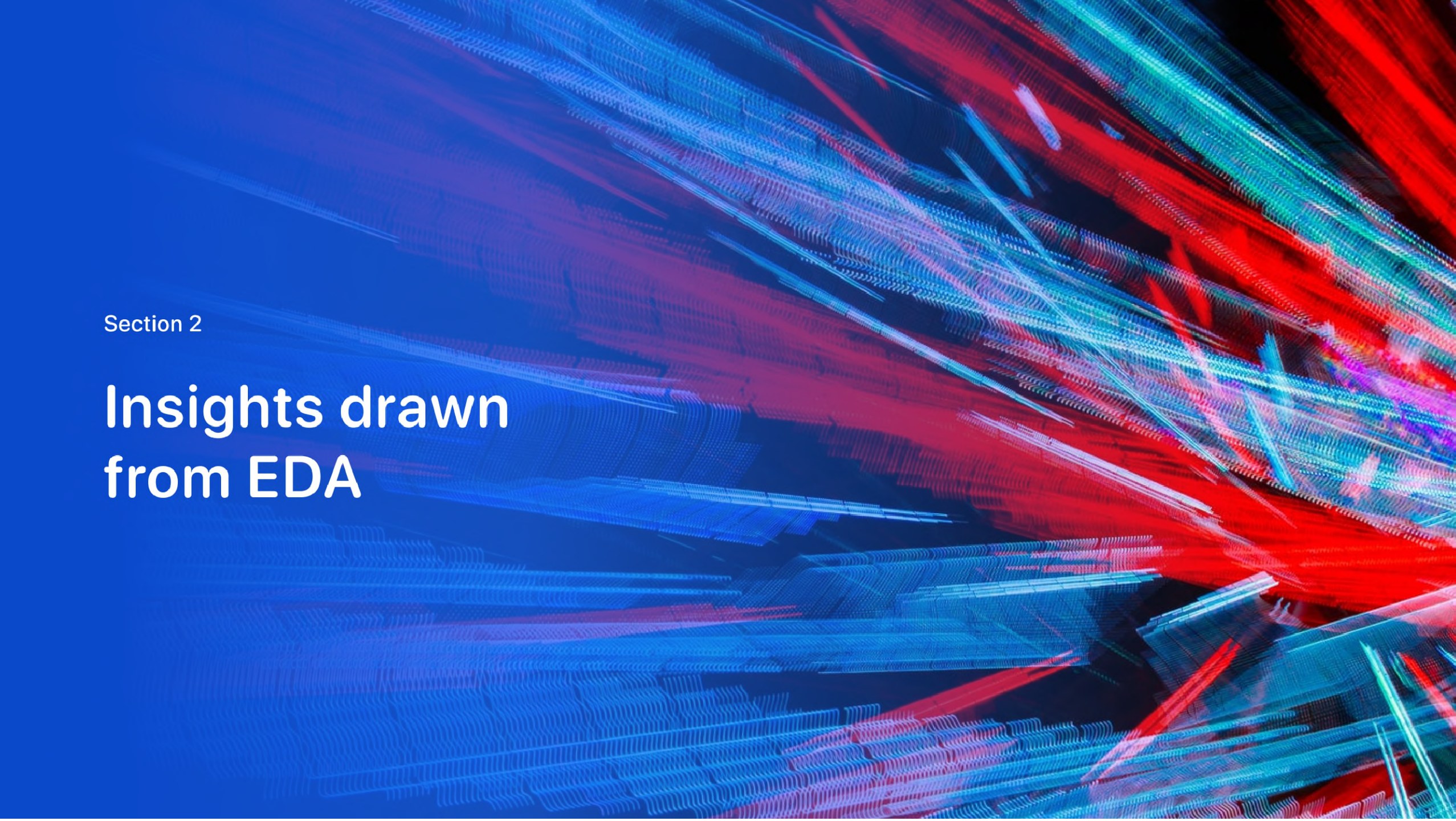


# Results

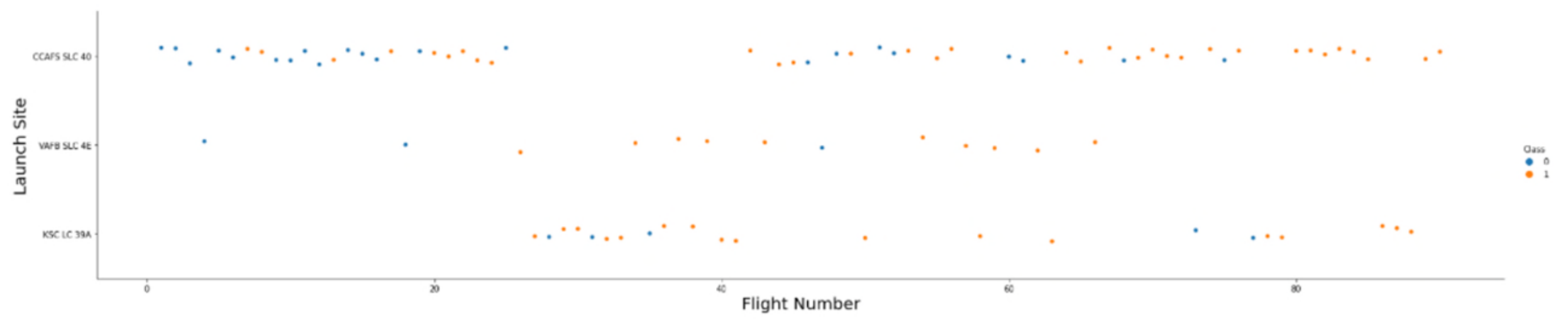
- One of the overall conclusions from Exploratory Data Analysis is that as launches attempts were increasing so the success rate was increasing. In fact, we see that the success rate since 2013 kept increasing till 2017, then presented a decline (2018), but recover in the following 2 years. When looking to orbit selected and how that correlates with success launch there are 4 orbit type that stood out as a very relevant for launch success: ES-L1; GEO; HEO and SSO. When considering also the rocket payload mass, one can see that with heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS orbits. Two interesting numbers I would like to highlight are: average Payload Mass by F9 v1.1 was 2534 Kg, and that first Successful Ground Landing occurred on 22-12-2015. More results and detailed info can be seen in section “Insights drawn from EDA”.
- From the launch sites proximity analysis, it was confirmed the closeness of launch sites to costs/sea and in, as much as possible, isolated places from crowdy places (in case of any bad takeoff the rocket falls in water away from population). Besides that, the sites are as well close to railway lines, which is convenient to transport parts of the rocket.
- From other interactive analysis we conclude site KSC LC-39A (John F. Kennedy Space Center, Florida - Launch Complex 39A) is the site with more successful launches and site CCAFS SLC-40 (Cape Canaveral Space Force Station, Florida – Space Launch Complex 40) is the site with less successful launches rate. First sire mentioned is also the site that presents the with highest launch success ratio, which permits to conclude this site presents goods characteristics for launching rockets. More info and screenshots can be seen in section “Launch Sites Proximities Analysis” and “Build a Dashboard with Plotly Dash”.
- The classification models built present the same performance accuracy of ~83,33%. More info can be seen in section “Predictive Analysis (Classification)”.

Section 2

# Insights drawn from EDA

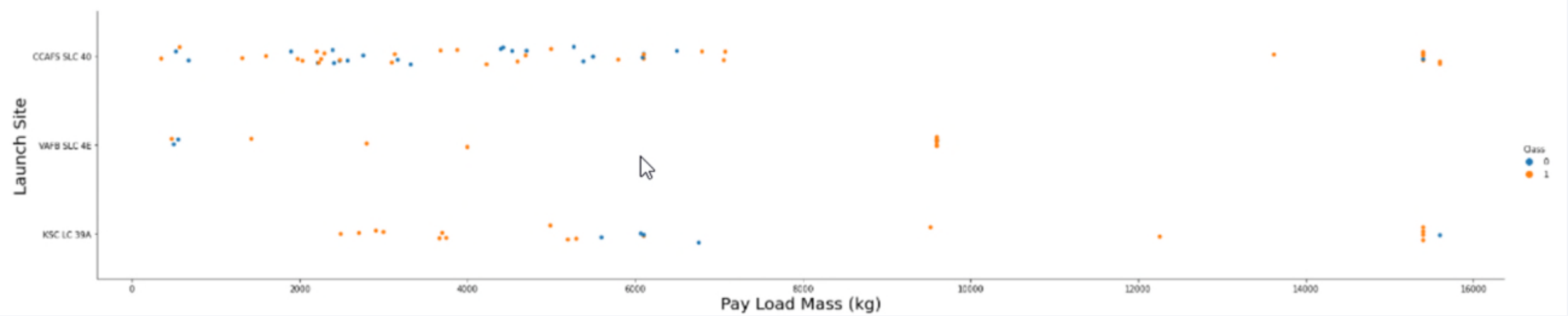


# Flight Number vs. Launch Site



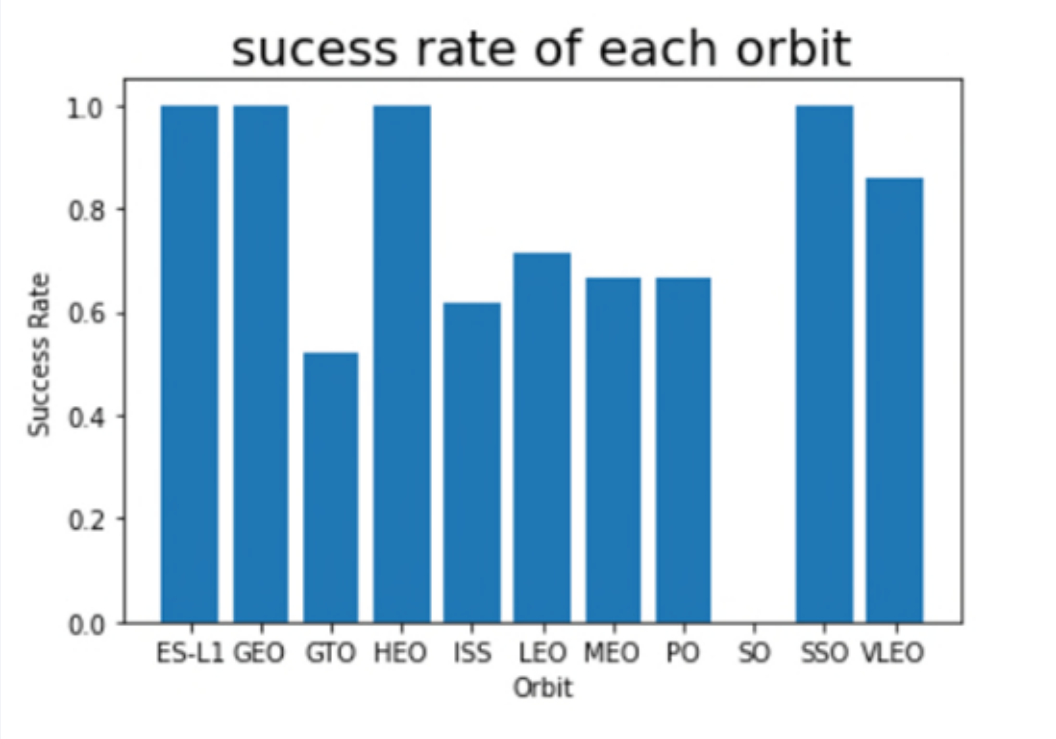

- Here is a scatter plot of Flight Number vs. Launch Site
- One can see that as launches attempts were increasing the success rate was increasing as well. Success rate in first launches was very low.

# Payload vs. Launch Site



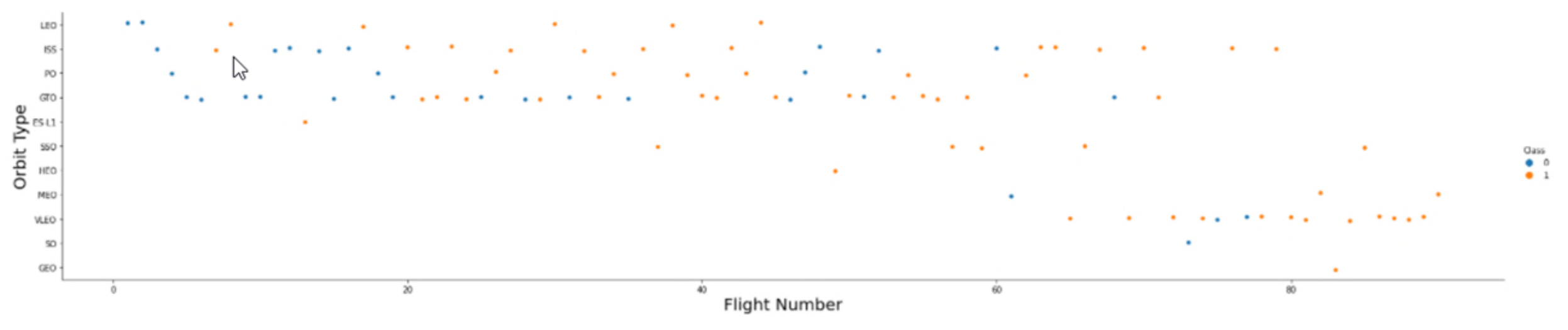

- Here is a scatter plot of Payload vs. Launch Site
- Observing the results one can see that for the VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000).

# Success Rate vs. Orbit Type



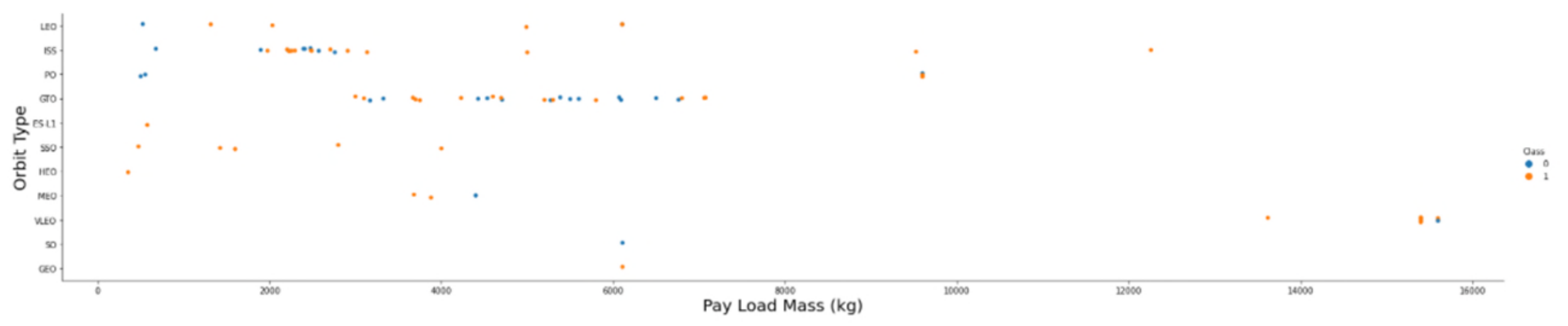

- Here is a bar chart for the success rate of each orbit type. In this chart one can see the relationship between success rate and orbit type.
- 4 orbit type stood out as a very relevant for launch success: ES-L1; GEO; HEO and SSO.

# Flight Number vs. Orbit Type




- Here is a scatter point of Flight number vs. Orbit type
- One can see that in the LEO orbit the high success appears related to the (low) number of flights and in the opposite spectrum of success we see there seems to be no relationship between flight number when in GTO orbit.

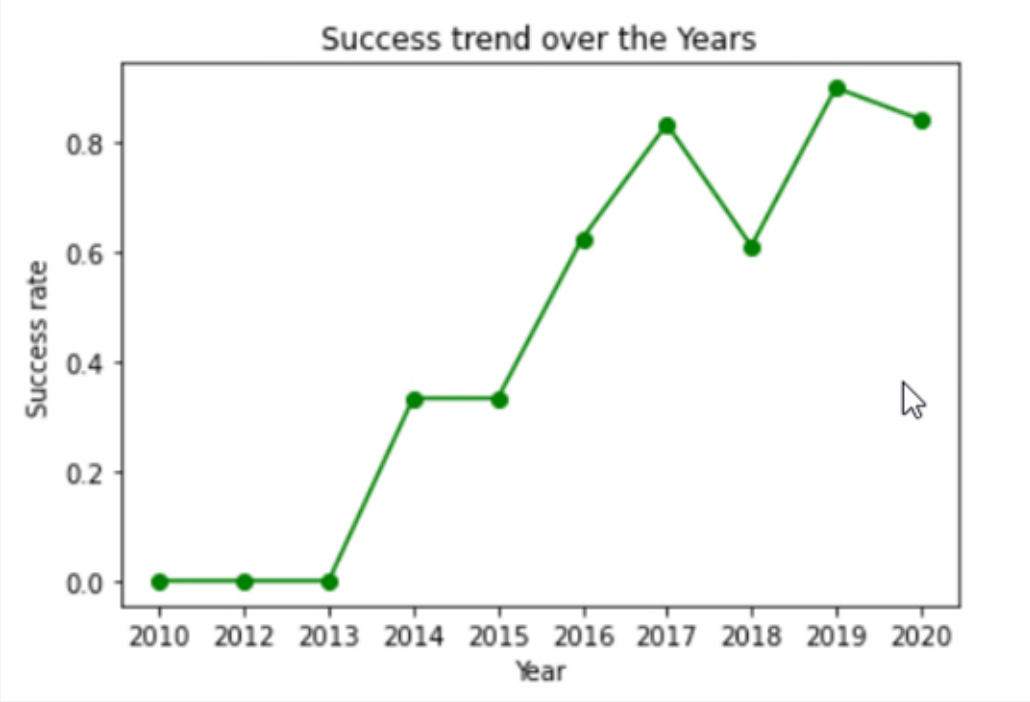
# Payload vs. Orbit Type




- Here is a scatter point of Payload vs. Orbit type
- One can see that with heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS orbits. However, for GTO we cannot distinguish this well as both positive landing rate and negative landing (unsuccessful mission) are both there.

# Launch Success Yearly Trend

- Here is a line chart of yearly average success rate
- In this graph one can see the average launch success trend. In fact, we see that the success rate since 2013 kept increasing till 2017, then presented a decline (2018), but recover in the following 2 years.



# All Launch Site Names

- An SQL query was issue to find all the names of the unique launch sites.
- Result obtained with a query to select “DISTINCT LAUNCH_SITE” values:

| launch_site |
| --- |
| CCAFS LC-40 |
| CCAFS SLC-40 |
| KSC LC-39A |
| VAFB SLC-4E |

# Launch Site Names Begin with 'CCA'

- An SQL query was issue to find 5 records where launch sites begin with `CCA`.
- Result obtained with a query to select the data “WHERE LAUNCH_SITE LIKE ‘CCA’ ” and limited to “LIMIT 5” entries:

| 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

- An SQL query was issue to calculate the total payload mass carried by boosters from NASA (CRS).
- Result obtained with a query with “SUM(PAYLOAD_MASS__KG_)” and “WHERE CUSTOMER LIKE 'NASA (CRS)’ ”:

# Average Payload Mass by F9 v1.1

- An SQL query was issue to calculate the average payload mass carried by booster version F9 v1.1.
- Result obtained with a query with “AVG(PAYLOAD_MASS__KG_)” and “WHERE BOOSTER_VERSION LIKE ‘F9 v1.1%’ ”:

# First Successful Ground Landing Date

- An SQL query was issue to find the dates of the first successful landing outcome on ground pad
- Result obtained with a query with “MIN(DATE)” and “WHERE LANDING__OUTCOME LIKE ‘Success (ground pad)’ ”:

# Successful Drone Ship Landing with Payload between 4000 and 6000

- An SQL query was issue to list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- Result obtained with a query “WHERE LANDING__OUTCOME LIKE 'Success (drone ship)’ ” and “PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000”:

| booster_version | payload_mass__kg_ | landing__outcome |
|---|---|---|
| F9 FT B1022 | 4696 | Success (drone ship) |
| F9 FT B1026 | 4600 | Success (drone ship) |
| F9 FT B1021.2 | 5300 | Success (drone ship) |
| F9 FT B1031.2 | 5200 | Success (drone ship) |

# Total Number of Successful and Failure Mission Outcomes

- An SQL query was issue to calculate the total number of successful and failure mission outcomes
- Success rate result obtained with a query with “COUNT(LANDING__OUTCOME)” and “WHERE LANDING__OUTCOME LIKE 'Success%’ ”:

```
Number Successful outcome:
| 1  |
+----+
| 61 |
+----+
```

- Failure rate result obtained with a query with “COUNT(LANDING__OUTCOME)” and “WHERE LANDING__OUTCOME LIKE ‘Failure%' ”:

```
Number Unsuccessful outcome:
| 1  |
+----+
| 10 |
+----+
```

# Boosters Carried Maximum Payload

- An SQL query was issue to list the names of the booster which have carried the maximum payload mass
- Result obtained using a subquery, in a query with “WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) from SPACEXDATASET)”:

| 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 |

# 2015 Launch Records

- An SQL query was issue to list the failed ”landing_outcomes” in drone ship, their booster versions, and launch site names for in year 2015
- Result obtained with a query with “WHERE LANDING__OUTCOME LIKE 'Failure (drone ship)' AND YEAR(DATE) = 2015”:

| DATE | booster_version | launch_site | landing__outcome |
|---|---|---|---|
| 2015-01-10 | F9 v1.1 B1012 | CCAFS LC-40 | Failure (drone ship) |
| 2015-04-14 | F9 v1.1 B1015 | CCAFS LC-40 | Failure (drone ship) |

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

- An SQL query was issue to 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
- Result obtained with a query selecting “LANDING__OUTCOME, COUNT (*) as "Count of outcomes" ”WHERE DATE >= '2010-06-04' AND DATE <= '2017-03-20’ “GROUP BY LANDING__OUTCOME” and “ORDER BY "Count of outcomes" DESC”:

| landing__outcome | Count of outcomes |
|---|---|
| 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 |

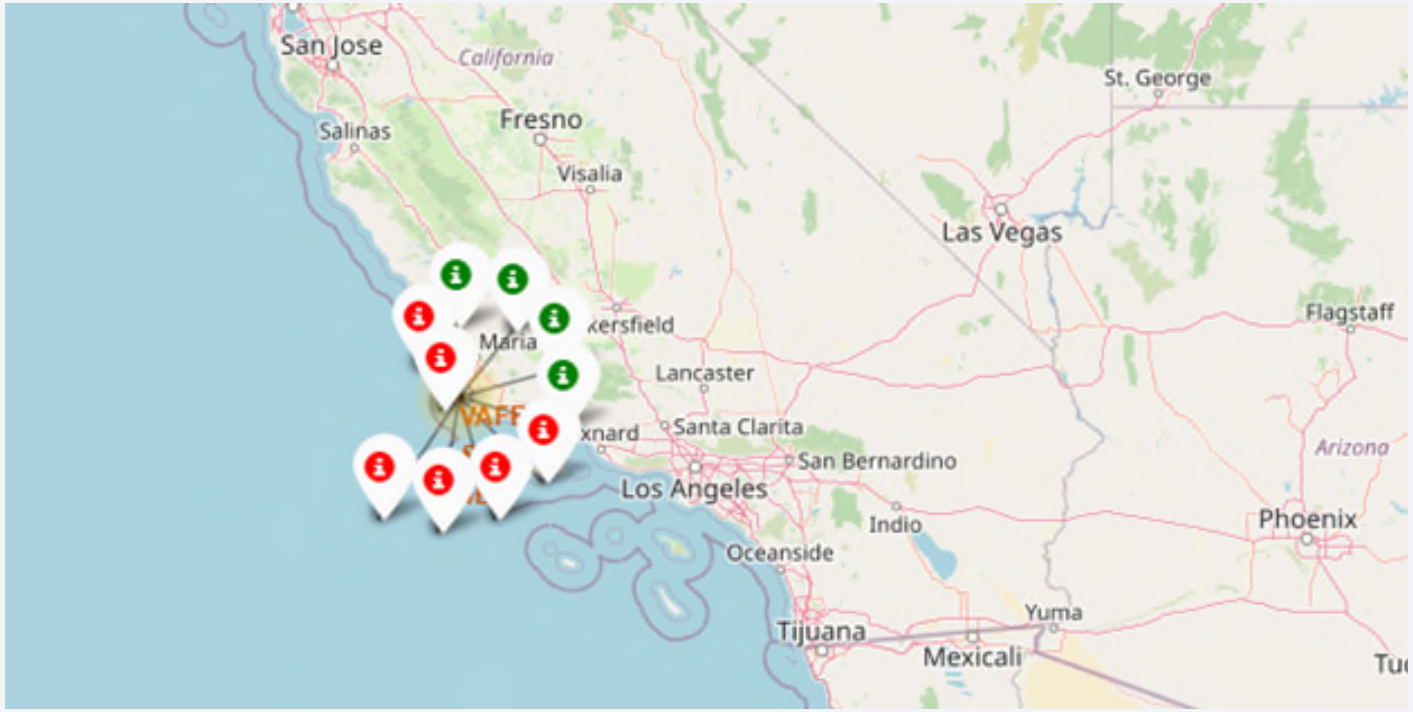
Section 4

# Launch Sites Proximities Analysis

# All launch sites on a USA map

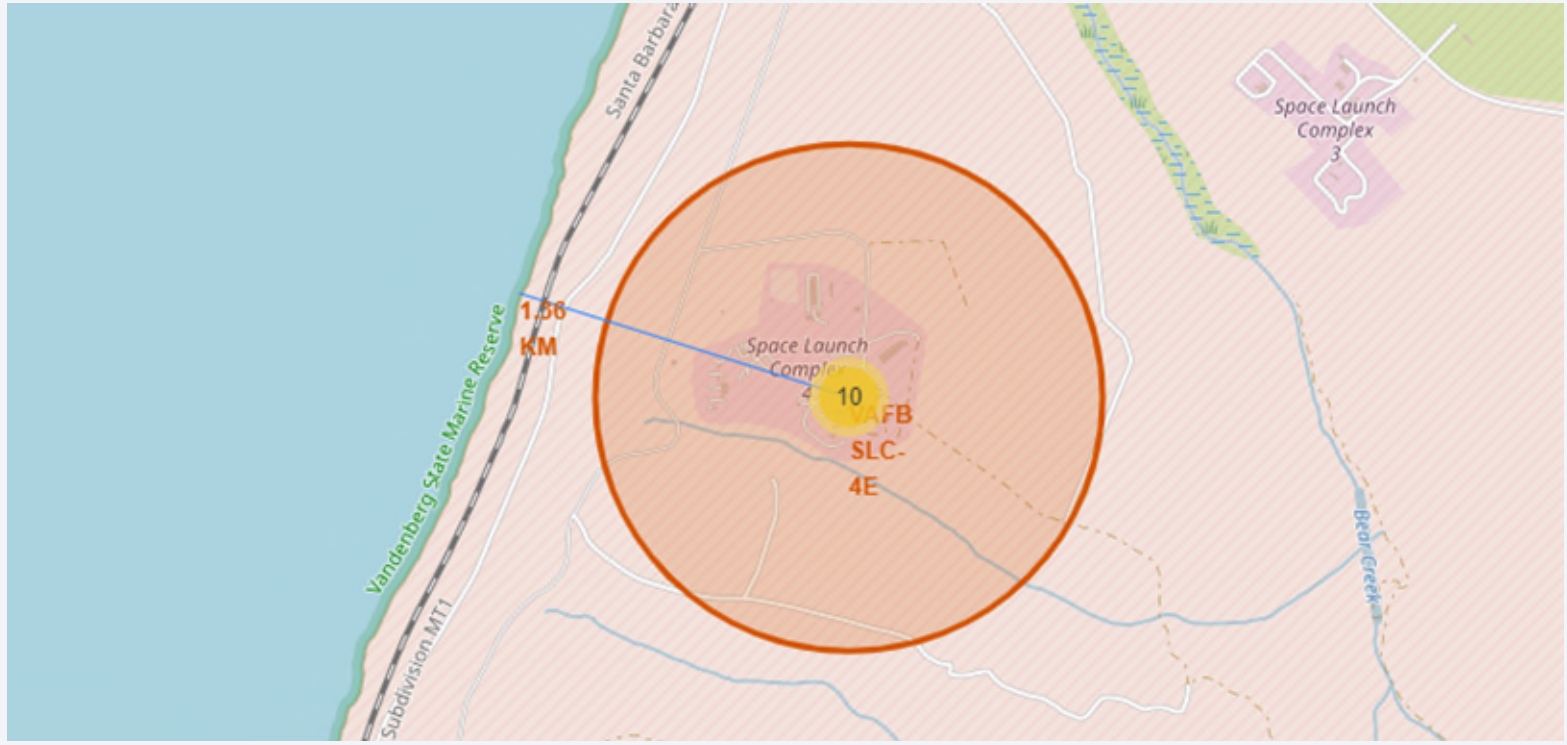

- Launch sites are marked with circles and have a text reference for them. For the launch sites in southeast coast, since they are very close the references are overlapping.
- Launch sites are close to costs/sea and in, as much as possible, isolated places from crowdy places.

# Success/failed launches for site VAFB-SLC4

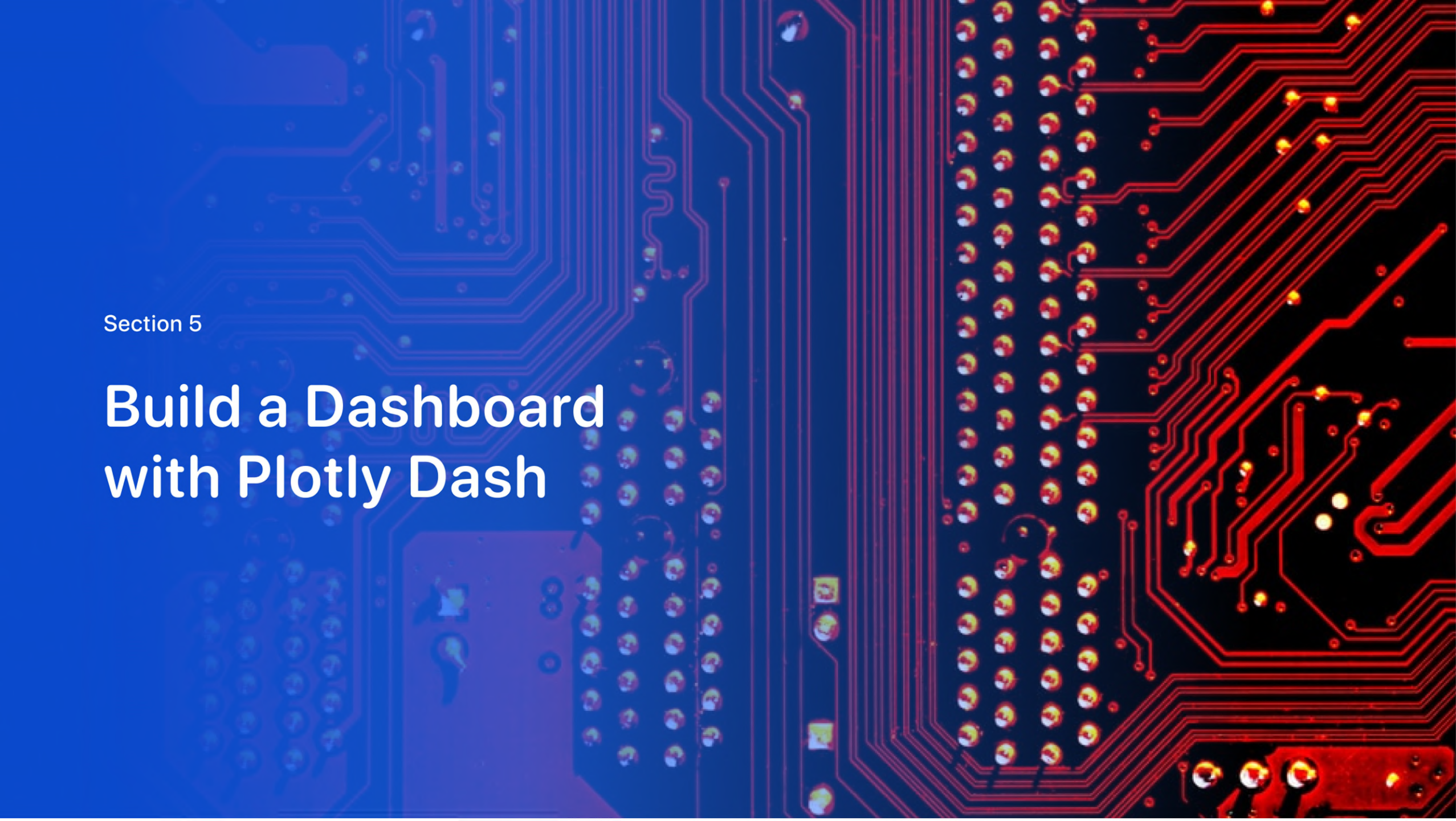



- In this map are marked the success (in green circle) and failed (in red circle) launches in Space Launch Complex 4 (SLC-4) at Vandenberg Air Force Base (VFAB), California, U.S.
- With these distinct marks for success and failure one can have an immediate understanding on success rate in this site: 4 out of 10.

# Launch site VAFB-SLC4 distance to coast



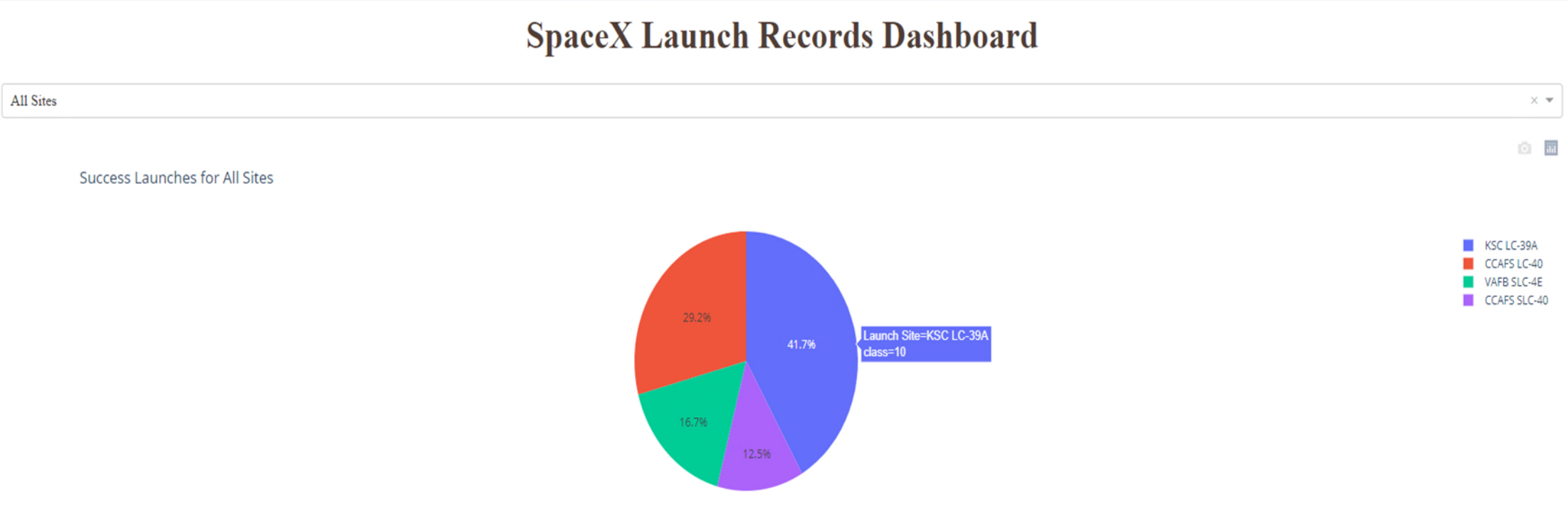

- In this map is marked with a line the distance (which is also displayed) of Space Launch Complex 4 (SLC-4) at Vandenberg Air Force Base (VFAB), California, U.S, to the coast (Vandenberg State Marine Reserve).
- We can see the site is not only very close to see (in case of any bad takeoff the rocket falls in water away from population), and as well close to a railway line, which is convenient to transport parts of the rocket.



Section 5

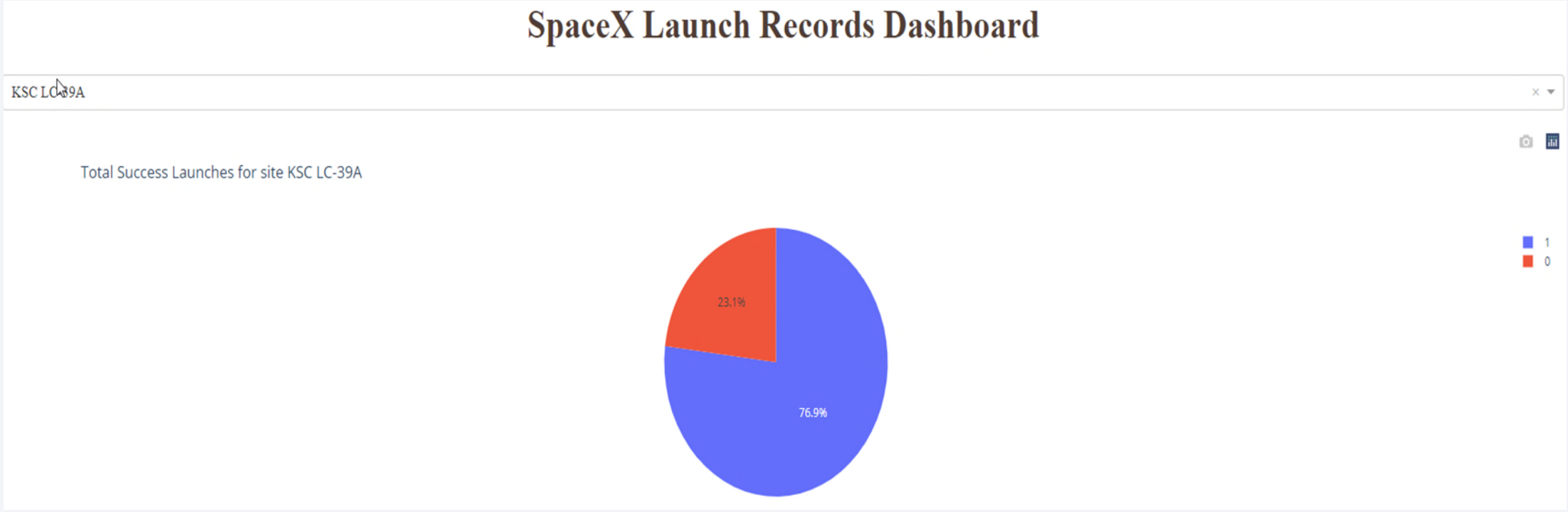
# Build a Dashboard with Plotly Dash

# Launch success count for all sites



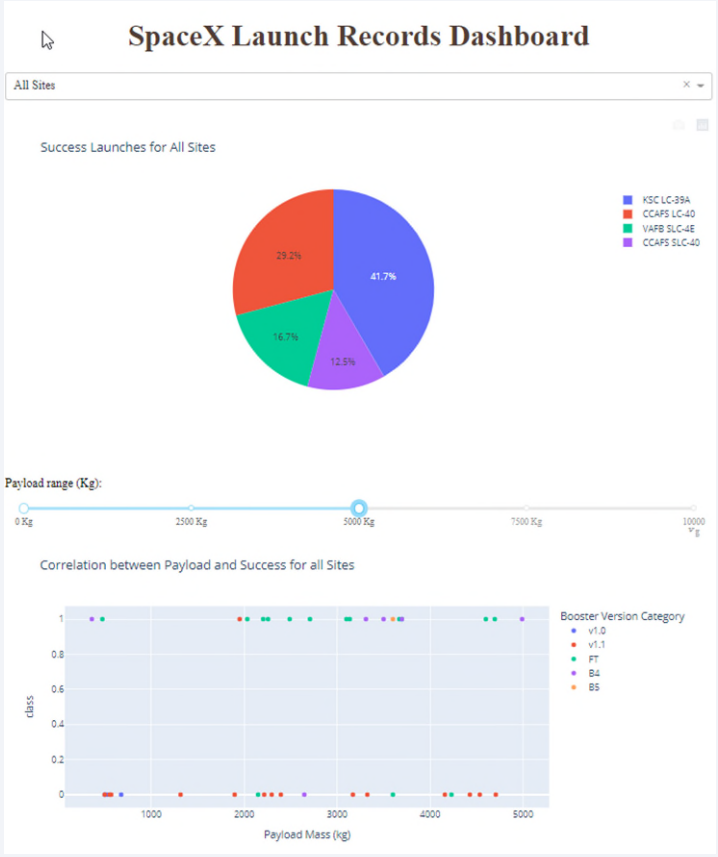

- In this pie chart one can compare the success counts for each site. We see site KSC LC-39A (John F. Kennedy Space Center, Florida - Launch Complex 39A) is the site with more successful launches and site CCAFS SLC-40 (Cape Canaveral Space Force Station, Florida – Space Launch Complex 40) is the site with less successful launches rate.

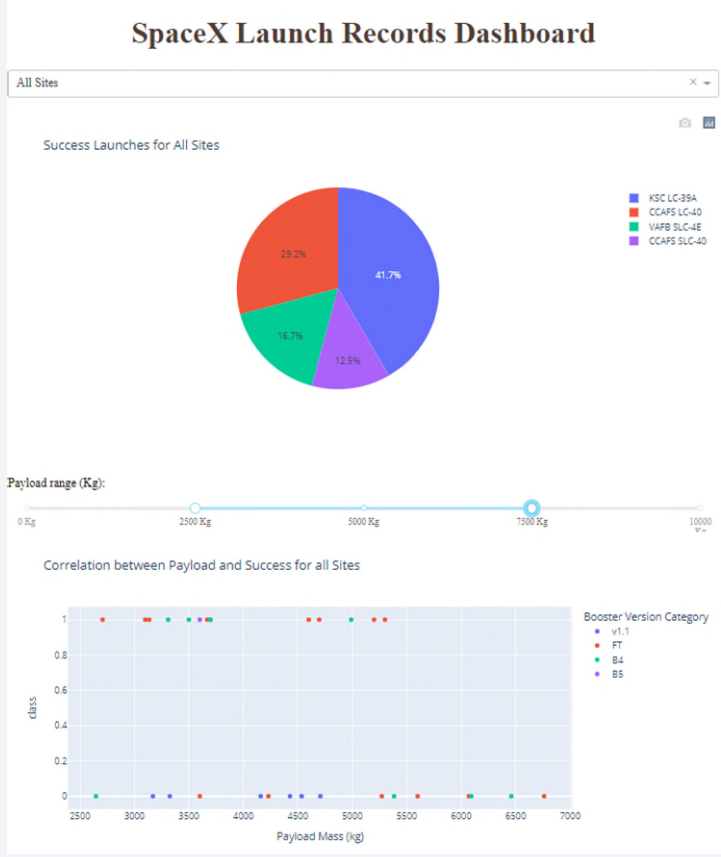
# Launch site with highest launch success ratio

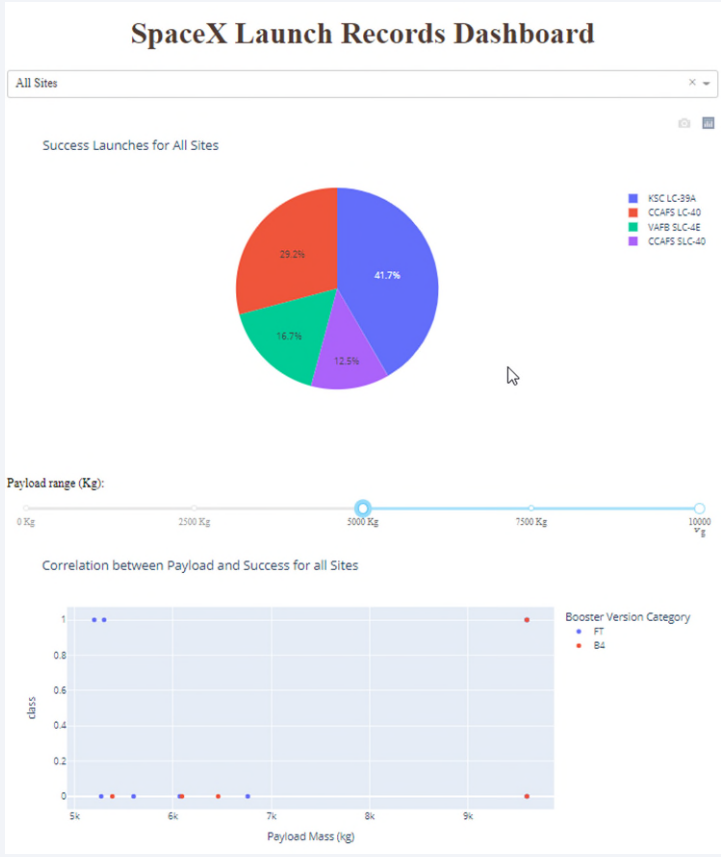



- Site KSC LC-39A (John F. Kennedy Space Center, Florida - Launch Complex 39A) is the site that presents the with highest launch success ratio. Previously (previous slide) we saw it was the site with more successful launches. This gives us the indication this site presents goods characteristics for launching rockets.

# Payload vs. Launch Outcome scatter plot for all sites




In these graphs one can see that success rate is higher for payload range between 2K Kg and 5K Kg. For payload above 6k Kg and below 9K Kg the success rate is 0!

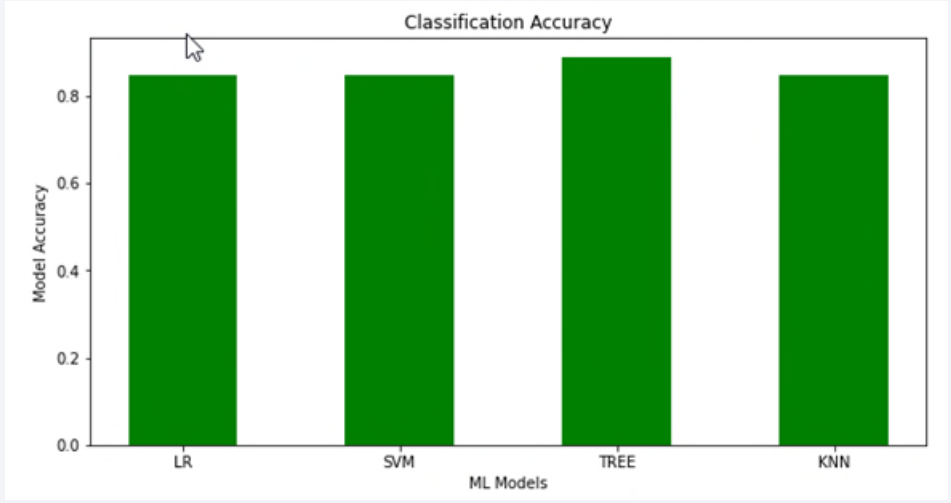
Looking to the booster version, the ones presenting largest success rate is FT and B4.

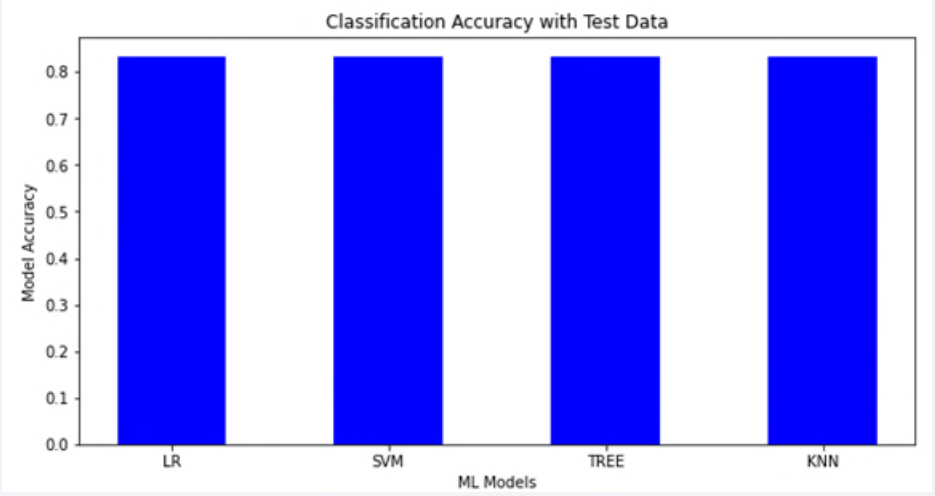
Section 6

# Predictive Analysis (Classification)

# Classification Accuracy

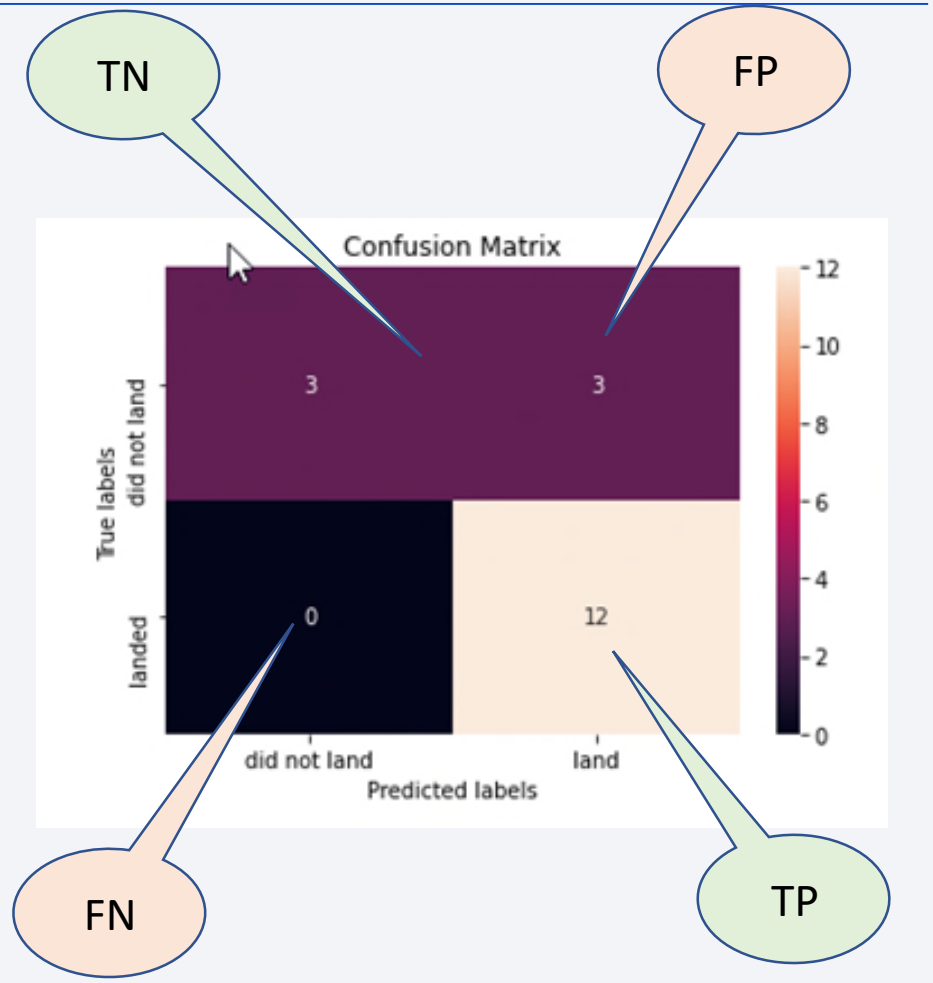
- For all classification models considered and where the “GridSearchCV” was applied it was observed the model accuracy:
  - on the validation data via the attribute “best_score_”. The values are presented in the bar chart on the right with green bars. One can see the Classification Tree model is the one with highest classification accuracy of ~87,32%.
  - On test data via the function “score”. The values are presented in the bar chart on the right with blue bars. One can see the all models present the same performance accuracy of 83,33%.
- In overall the models predict a successful landing with an accuracy of 83,33%.



# Confusion Matrix

- On the right side is the confusion matrix obtained for Classification Tree model. Note however that all models present the same values in confusion matrix.
- Examining the confusion matrix, we see that model can distinguish between the different Class values (success vs failure). Still the problem we see is the false positives, i.e., classifying as positive (success) when it should have classified as failure.
- The model presents
  - A precision of 0.80 ( TP / (TP+FP) = 12 / (12+3) )
  - A sensitivity (recall) of 1 (TP / (TP+FN) =12 / (12+0) ).



# Conclusions

- For the target of building a predictive model of successful launch data was obtained from several sources.
- Several methodologies were used to work the raw data obtained and identify which features (like launch site, rocket payload mass, …) could be used to build the predictive model, i.e., which feature have some significance degree correlated with the launch success.
- A classification model to predict if a launch is successful was built and present an accuracy level of ~83.33%.
- In order to improve the model accuracy and eventually distinct between the classification different models tested more data would be required.

# Appendix

- All assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project are available in Coursera Capstone project in GitHub project.

# Thank you!