IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Douglas Chaves
12/05/2022

Outline

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

- GitHub: https://github.com/DouglasFr/Capstone

Executive Summary

- The focus was to find the best outcome results from the landing data acquired from SpaceX Rest API, comparing results from payload data, launch site, launch method, orbit, etc
- After the data was treated, we created Viz using both MatPlot and Seaborn library from Python data science tools.
- Several models (KNN, SVM, Log Reg, Decision Tree) was tested in order to find the most accurate predictor of a launch success.

Introduction

- In this Capstone Project we must predict the success rate of Falcon 9´s first stage. The value stipulated on the SpaceX website for the first stage is $62 Millions with additional costs reaching up to $165 Millions.
- If we can determine if the first stage will land successfully, we can also determine the cost of the launch. With this information in hand, we can simulate a rival company of SpaceX that wants to bid against SpaceX rocket launch costs.

Section 1

Methodology

Methodology

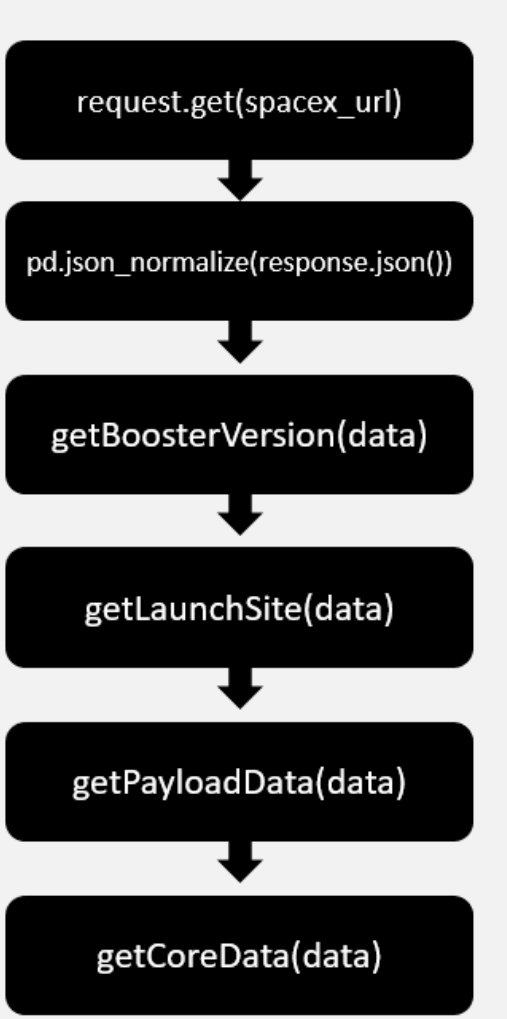
Executive Summary

- Data collection methodology:
  - SpaceX API and Wikipedia Web Scrapping
- Perform data wrangling
  - Treat missing values and create landing category variable.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Grid search to train, test and tune classification models.

Data Collection – SpaceX API

- Here is all SpaceX Rest API requests:
  - Past launch data: https://api.spacexdata.com/v4/launches/past
  - Booster data: https://api.spacexdata.com/v4/rockets/
  - Launch pads data: https://api.spacexdata.com/v4/launchpads/
  - Payload data: https://api.spacexdata.com/v4/payloads/
  - Outcomes of landing: https://api.spacexdata.com/v4/cores/

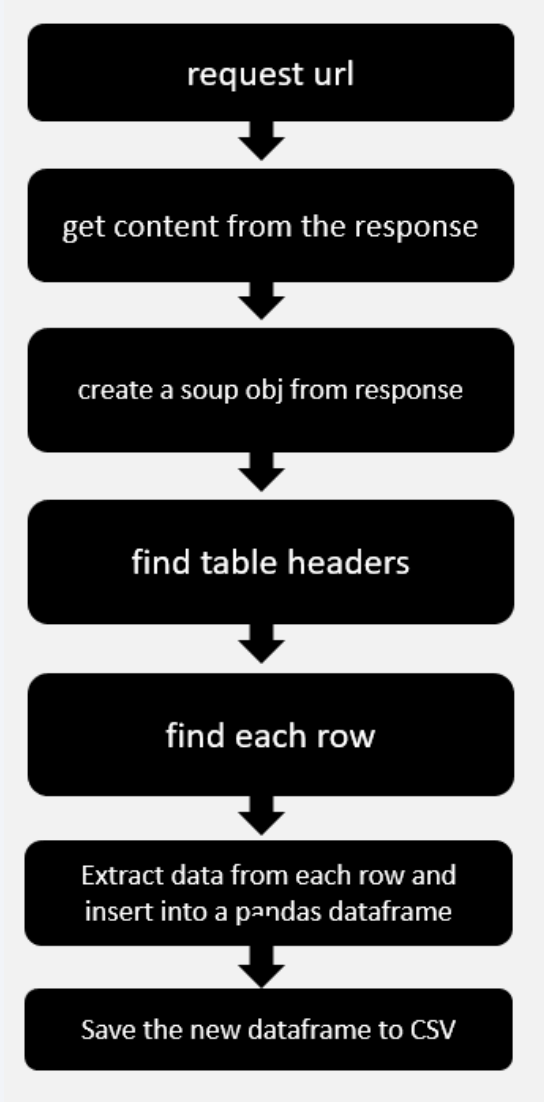
- https://github.com/DouglasFr/Capstone/blob/c4a516e642274882e216e1d6c6c4ab65ec5bbb9b/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- The ‘requests’ library was used to scrap a Wikipedia pages with the info needed: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- Function BeautifulSoap was used to parse the content returned in the response.
- The result was put into a pandas dataframe and saved as a CSV file.

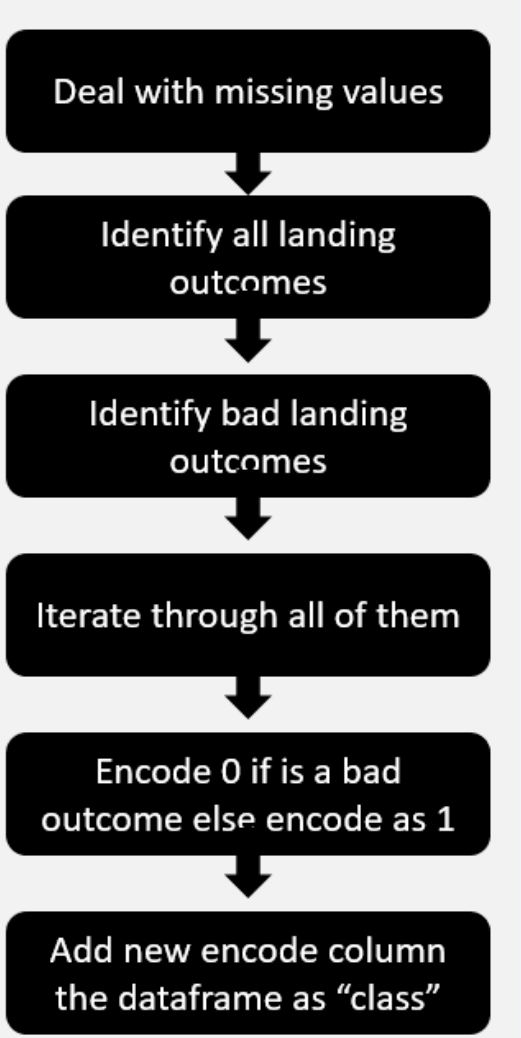
- https://github.com/DouglasFr/Capstone/blob/c4a516e642274882e216e1d6c6c4ab65ec5bbb9b/jupyter-labs-webscraping.ipynb



Data Wrangling

- A classification variable must be created to encode the results of the outcome of the landing as bad landing = 0 or good landing = 1, this variable name will be “class”.

- https://github.com/DouglasFr/Capstone/blob/c4a516e642274882e216e1d6c6c4ab65ec5bbb9b/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- All the Exploratory Data Analysis are done using either Matplotlib or Seaborn to create visualizations from the pandas dataframe
- The main Viz used was bar chart, scatter chart and line chart to show relationships between Flight Number, Launch Sites, Payload Mass and Orbit
- All Viz created was color coded using the “class” variable that as created to classify the landing outcome.

- https://github.com/DouglasFr/Capstone/blob/415917f311f0361599bed3df093163f31c5e36e6/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- All queries performed:
  - select distinct(LAUNCH_SITE) from SPACEX;
  - select * from SPACEX where LAUNCH_SITE like 'CCA%' limit 5;
  - select CUSTOMER, SUM(PAYLOAD_MASS__KG_) as TOTAL from SPACEX where CUSTOMER = 'NASA (CRS)' GROUP BY CUSTOMER;
  - select BOOSTER_VERSION, AVG(PAYLOAD_MASS__KG_) as AVERAGE from SPACEX where BOOSTER_VERSION = 'F9 v1.1' GROUP BY BOOSTER_VERSION;
  - select MIN(DATE) as FIRST_DATE from SPACEX where LANDING__OUTCOME = 'Success (ground pad)' ;
  - select BOOSTER_VERSION, PAYLOAD_MASS__KG_ from SPACEX where LANDING__OUTCOME = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000;
  - select MISSION_OUTCOME, count(*) as TOTAL FROM SPACEX GROUP BY MISSION_OUTCOME;
  - select BOOSTER_VERSION, PAYLOAD_MASS__KG_ from SPACEX where PAYLOAD_MASS__KG_ in (select max(PAYLOAD_MASS__KG_) from SPACEX);
  - select DATE,BOOSTER_VERSION, LAUNCH_SITE, LANDING__OUTCOME FROM SPACEX WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND DATE between to_date('01/01/2015','DD/MM/YYYY') and to_date('31/12/2015','DD/MM/YYYY');
  - select LANDING__OUTCOME, count(*) as TOTAL from SPACEX WHERE DATE between to_date('04/06/2010','DD/MM/YYYY') and to_date('20/03/2017','DD/MM/YYYY') group by LANDING__OUTCOME order by TOTAL DESC;
- https://github.com/DouglasFr/Capstone/blob/415917f311f0361599bed3df093163f31c5e36e6/jupyter-labs-eda-sql-coursera.ipynb

Build an Interactive Map with Folium

- In order to show the launch locations in the Folium map Circles and Markers was used.
- Each launch location as a MarkerCluster showing it´s outcome green mark (successful outcome) or red mark (failed outcome).
- Lines was drawn in the map to show distances between launch sites or the distance between the site and the coast and so on.

- https://github.com/DouglasFr/Capstone/blob/415917f311f0361599bed3df093163f31c5e36e6/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

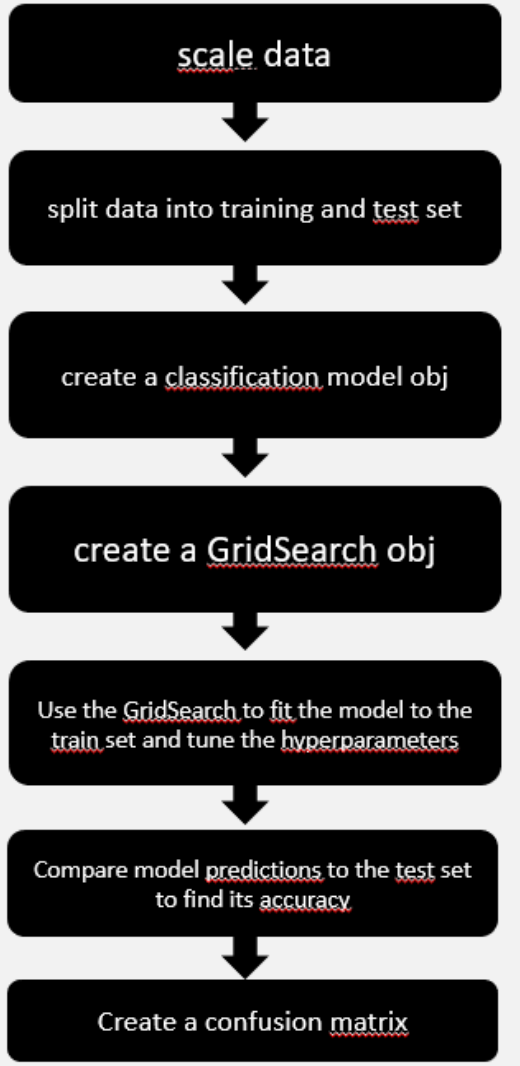
- An interactive dashboard allows users to analyze each launch site, its payload masses, booster type and the outcome (successful or failed).
- A pie chart shows the outcome for all launch sites
- A scatter plot shows what was the launch outcome accordingly to its payload mass and its data paints was colored by its booster type.

- https://github.com/DouglasFr/Capstone/blob/415917f311f0361599bed3df093163f31c5e36e6/Interactive%20Visuals%20Capstone.ipynb

Predictive Analysis (Classification)

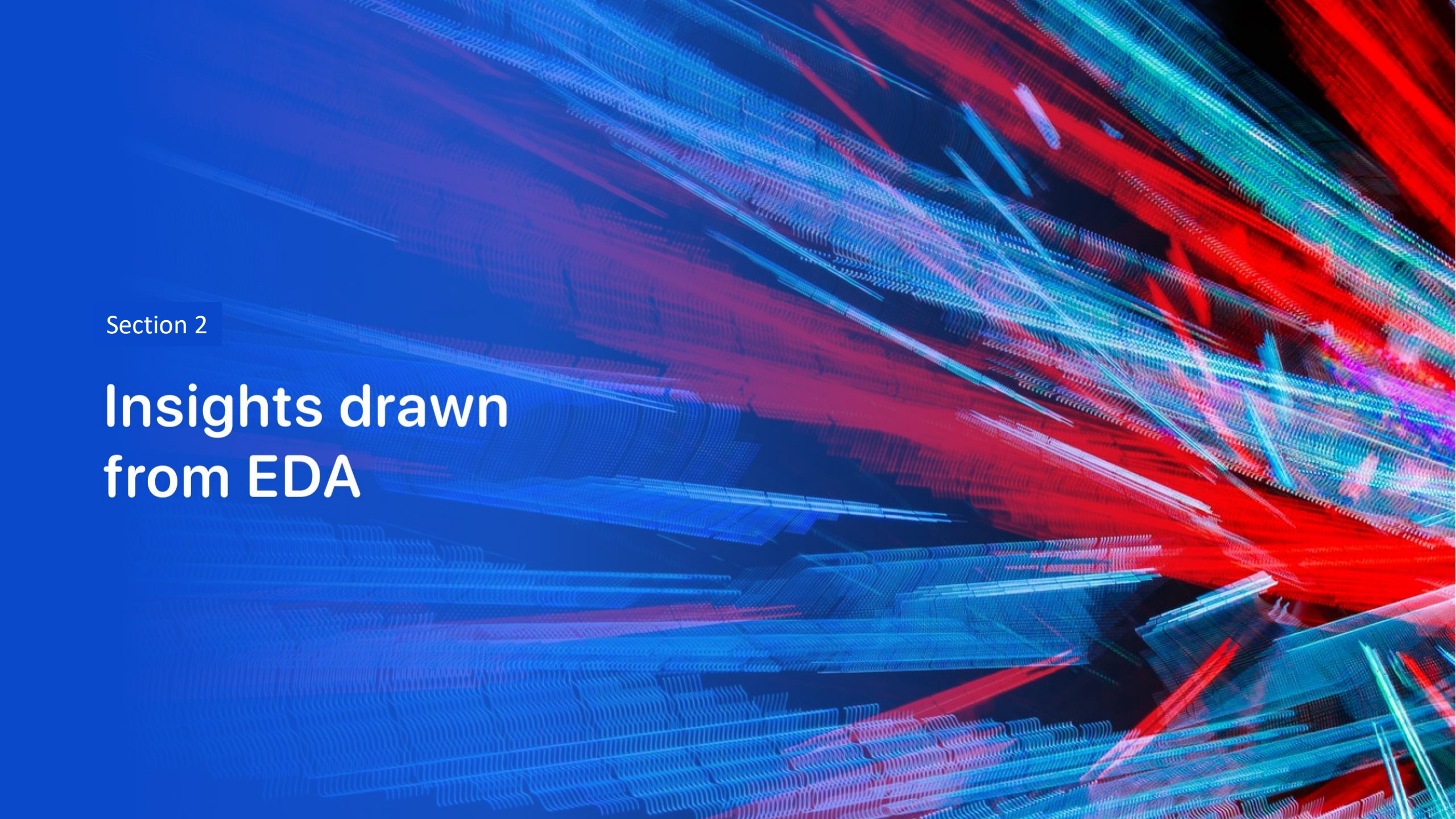
- Follow the instructions in the flowchart and use the classification algorithms bellow:
  - Logistic Regression
  - Support Vector Machine (SVM)
  - Decision Tree
  - K-nearest neighbors (KNN)

- https://github.com/DouglasFr/Capstone/blob/415917f311f0361599bed3df093163f31c5e36e6/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



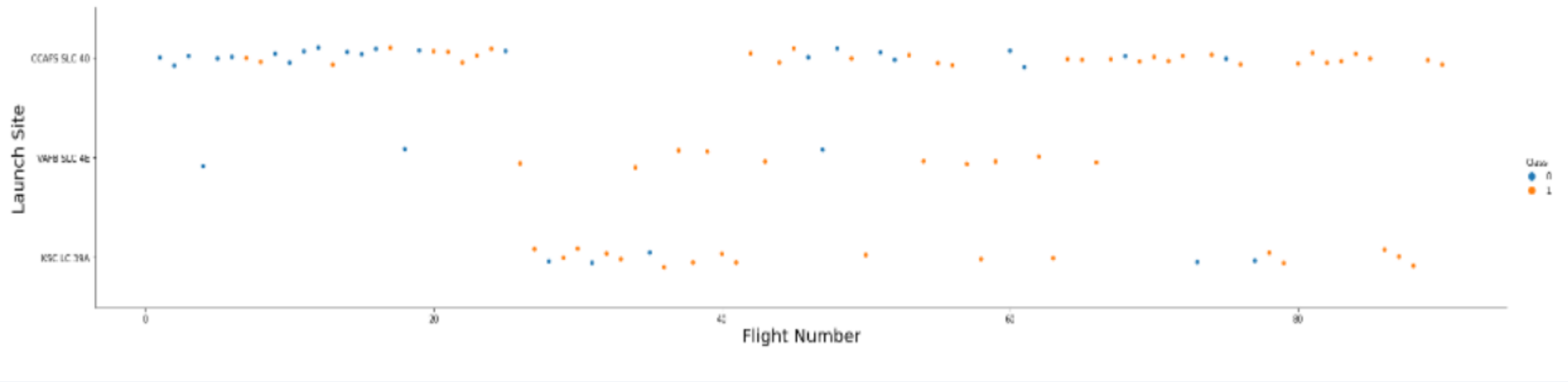
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn from EDA

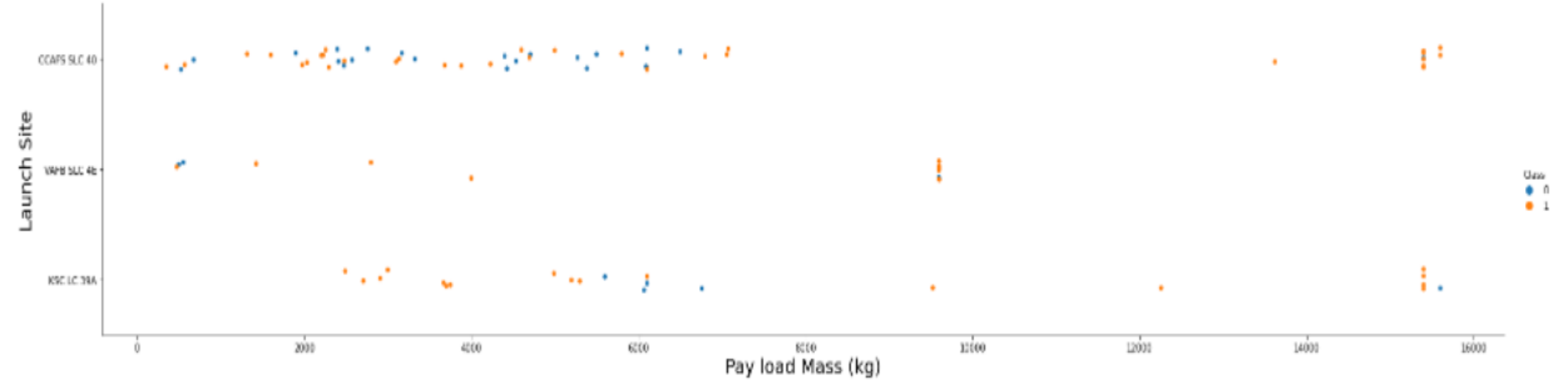
Flight Number vs. Launch Site

- Scatter plot made with Seaborn with the relation Flight Number x Launch Site, colored by outcome (blue = failed, orange = success)
- CCAFS represents the majority of the test's subjects, VAFB and KSC have the higher success rate.



Payload vs. Launch Site

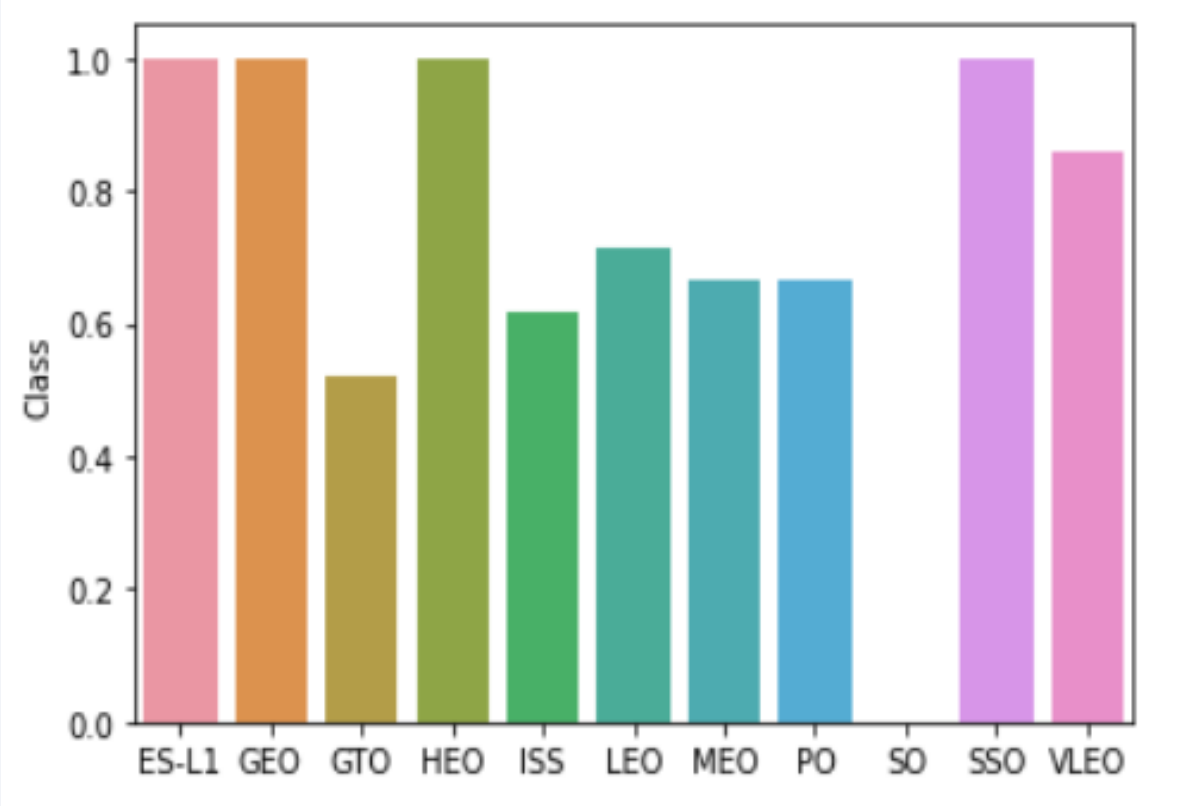
- Scatter plot made with Seaborn with the relation Payload x Launch Site, colored by outcome (blue = failed, orange = success)
- CCAFS had the highest number of launches with the mass between 0 and 8000 Kg.
- Launches with greater mass (16000 Kg) had a higher success rate.



Success Rate vs. Orbit Type

- Bar chart for the success rate of each orbit type

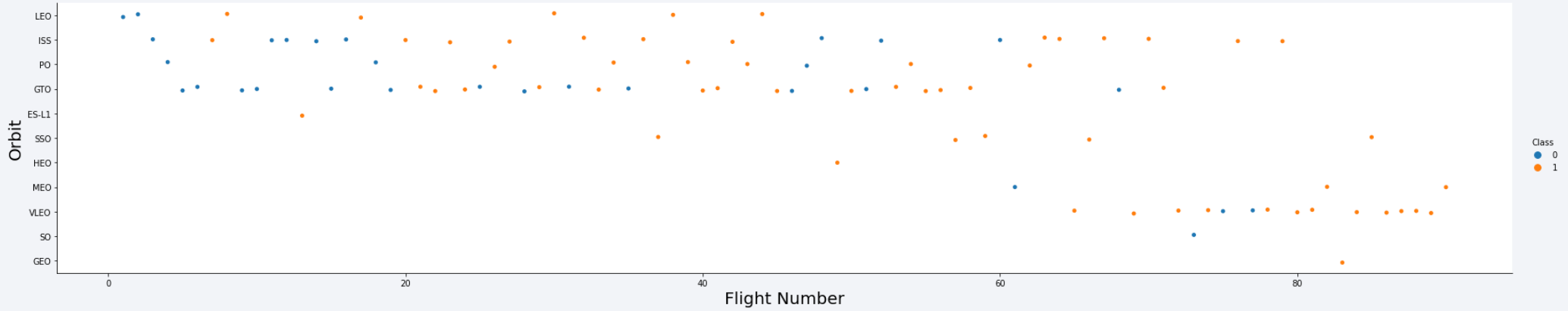
- If we analyze closely the number of samples, we could say that VLEO is the most successful of them, while GEO and HEO had a perfect 100%, they only had 1 sample of each, VLEO on the other hand had 14 samples and 80% successful landing outcomes



Flight Number vs. Orbit Type

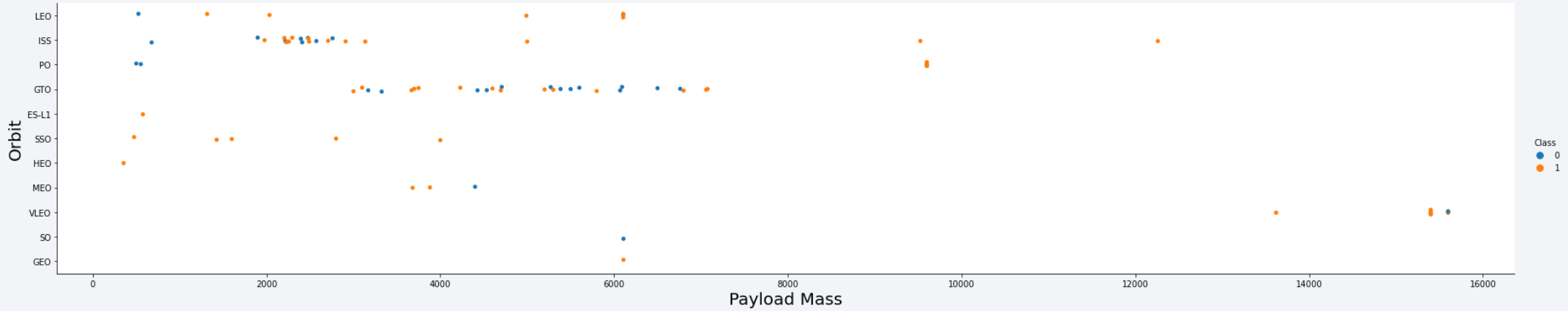
- Show a scatter point of Flight number vs. Orbit type, colored by outcome (blue = failed, orange = successful)

- All outcomes tend to improve with the numbers of flights.



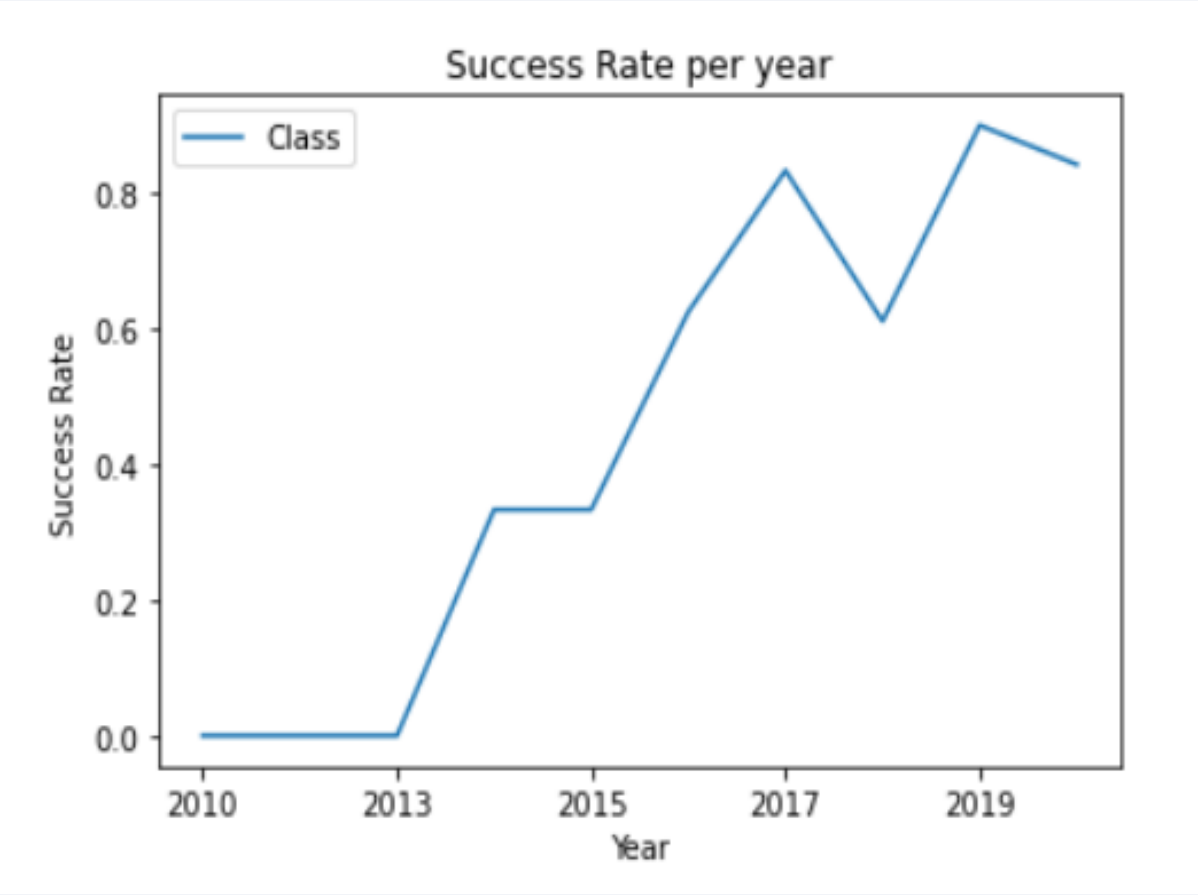
Payload vs. Orbit Type

- Show a scatter point of payload vs. orbit type, colored by outcome (blue = failed, orange = successful)
- We can see many samples located with payload between 0 to 8000 but no clear pattern among them.



Launch Success Yearly Trend

- A line chart of yearly average success rate
- We can see a steady increase of the success rate since 2013, there was a dip around 2018 but they went back on track soon after.



All Launch Site Names

- Query: select distinct(LAUNCH_SITE) from SPACEX
- The **distinct** clause select only the unique names for the Launch Site column.

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

Launch Site Names Begin with 'CCA'

- Query: select * from SPACEX where LAUNCH_SITE like 'CCA%' limit 5;
- The * select all values from the table; the Like 'CCA%' will filter strings that start with CCA and lastly limit 5 to only bring 5 rows of data.

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

- Query: select CUSTOMER, SUM(PAYLOAD_MASS__KG_) as TOTAL from SPACEX where CUSTOMER = 'NASA (CRS)' GROUP BY CUSTOMER;
- The main filter here is in the CUSTOMER column where only the 'NASA (CRS)' was selected, the group by was add in order to bring the Customer information in the initial select.

customer	total
NASA (CRS)	45596

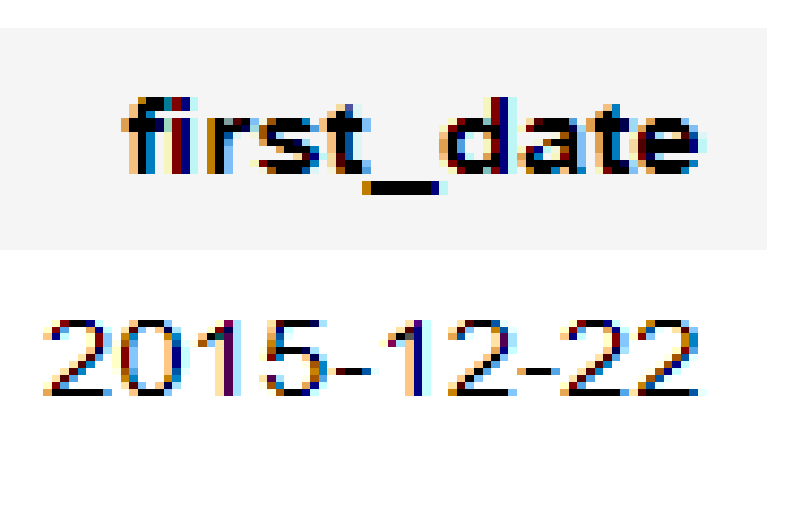
Average Payload Mass by F9 v1.1

- Query: select BOOSTER_VERSION, AVG(PAYLOAD_MASS__KG_) as AVERAGE from SPACEX where BOOSTER_VERSION = 'F9 v1.1' GROUP BY BOOSTER_VERSION;
- Similar situation from the previous query where a string filter was used, this time in the Booster Version column and using the AVG calculation instead of the SUM.

booster_version	average
F9 v1.1	2928

First Successful Ground Landing Date

- Query: select MIN(DATE) as FIRST_DATE from SPACEX where LANDING__OUTCOME = 'Success (ground pad)' ;
- This time the MIN function is called to get the first date in the dataset that have the condition of Success (ground pad) in the Landing Outcome;



Successful Drone Ship Landing with Payload between 4000 and 6000

- Query: select BOOSTER_VERSION, PAYLOAD_MASS__KG_ from SPACEX where LANDING__OUTCOME = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000;
- This time the between function was called in order to filter Payload Masses with value higher than 4000 and lower than 6000;

booster_version	payload_mass__kg_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

Total Number of Successful and Failure Mission Outcomes

- Select MISSION_OUTCOME, count(*) as TOTAL FROM SPACEX GROUP BY MISSION_OUTCOME;
- The group by function will divide the counted values using the filter Mission Outcome;

mission_outcome	total
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Query: select BOOSTER_VERSION, PAYLOAD_MASS__KG_ from SPACEX where PAYLOAD_MASS__KG_ in (select max(PAYLOAD_MASS__KG_) from SPACEX);
- A subquery was used to bring only the Booster Versions that have the MAX 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

2015 Launch Records

- Query: select DATE,BOOSTER_VERSION, LAUNCH_SITE, LANDING__OUTCOME FROM SPACEX WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND DATE between to_date('01/01/2015','DD/MM/YYYY') and to_date('31/12/2015','DD/MM/YYYY');
- A date formatting function 'To_Date' was used in order to make it easier to deal with the dates, filter the column landing outcome so that only the failure launches are selected and the between function in the dates of the year 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

- Query: select LANDING__OUTCOME, count(*) as TOTAL from SPACEX WHERE DATE between to_date('04/06/2010','DD/MM/YYYY') and to_date('20/03/2017','DD/MM/YYYY') group by LANDING__OUTCOME order by TOTAL DESC;

- A rank of all landing outcome with the dates between the year 2010 to 2017 is shown using the new calculated field Total as in the sort function DESC;

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


Section 3
Launch Sites
Proximities Analysis

All Launch Sites

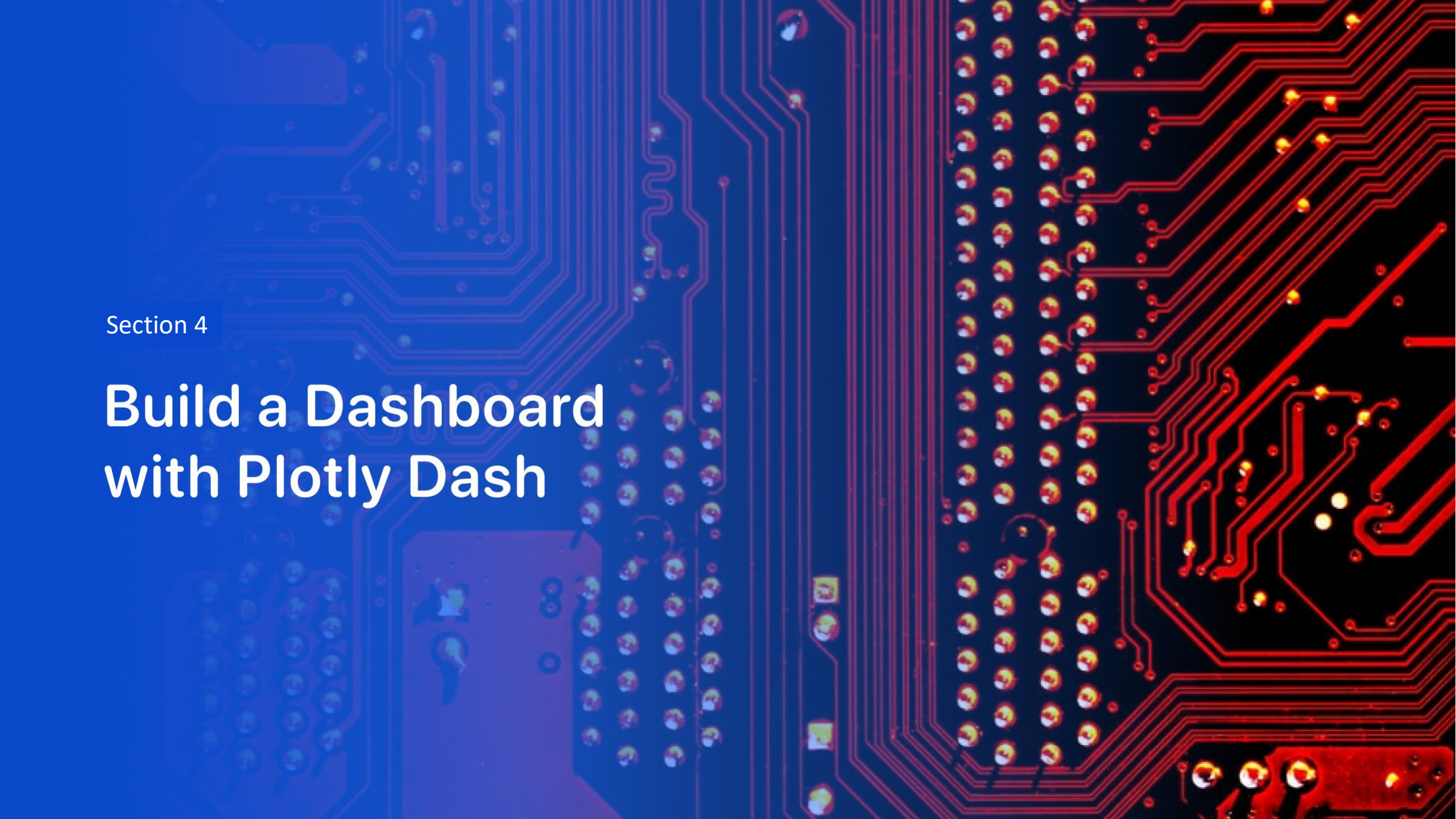
- All sites are next to the coast and major oceans.

Success/Failed launches

- An example of how the launch outcome is displayed in a folium map.

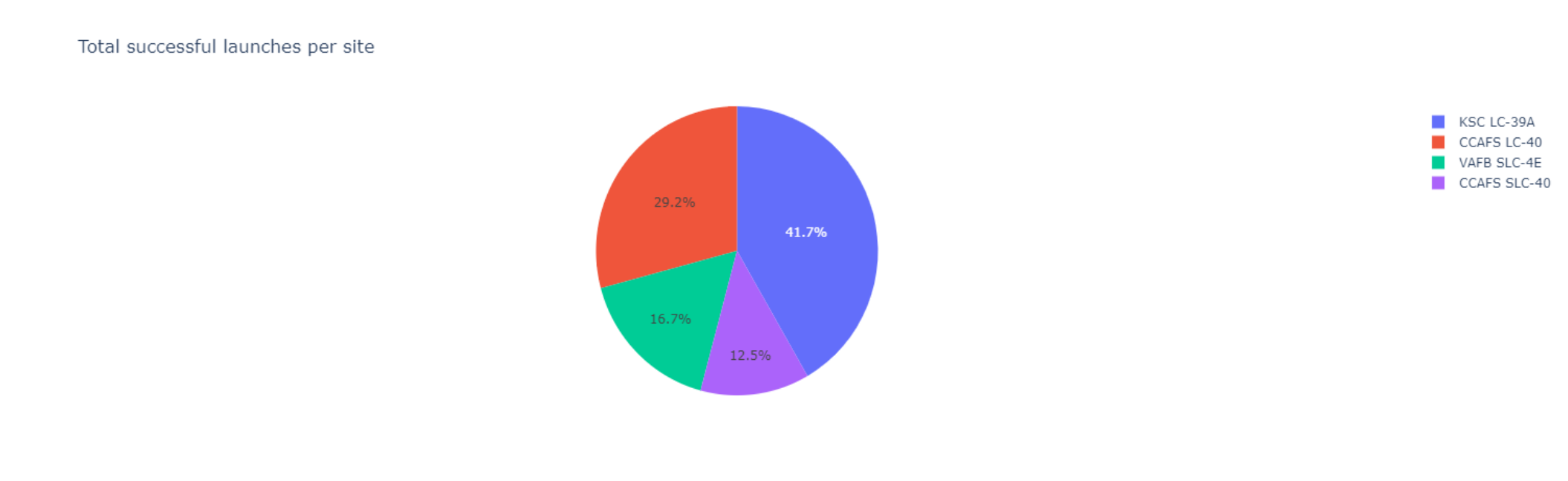
Distances between launch sites

- The map itself will show not only the distances between sites but also the distance from the coast.


Section 4
Build a Dashboard with Plotly Dash

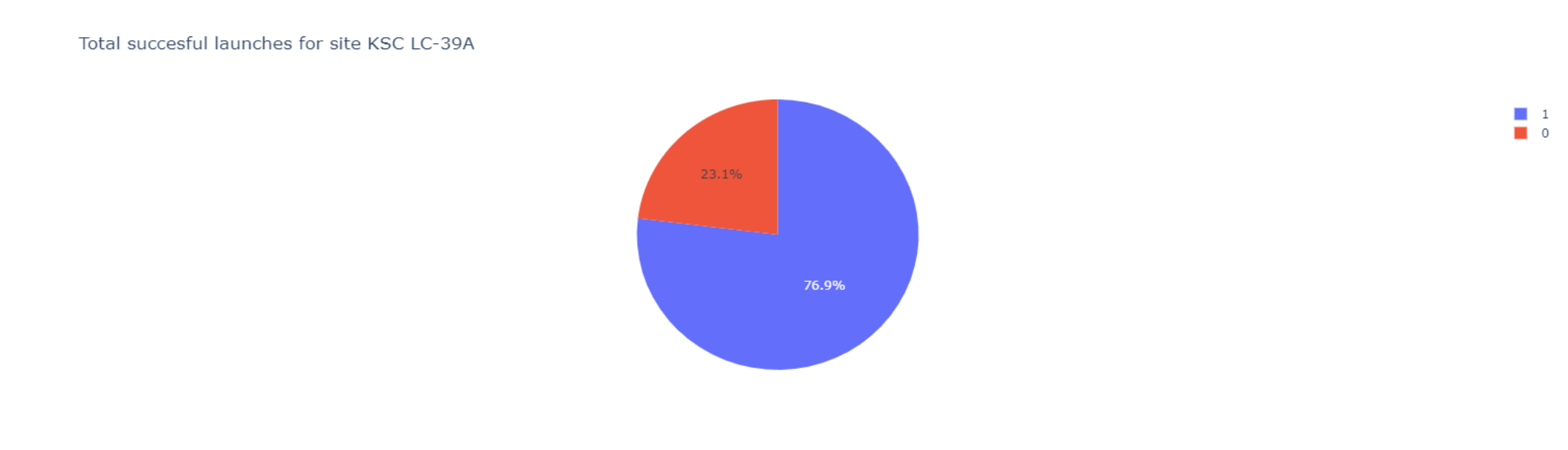
Successful launches across sites

- With a simple pie chart, we can see that most of the successful launches comes from KSC sites.



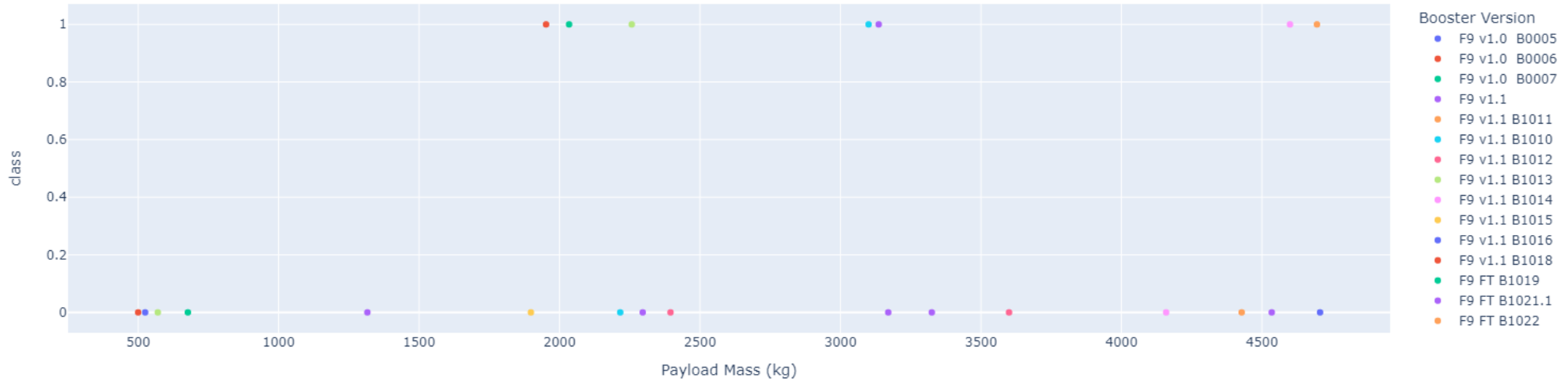
Site with highest success rate

- With a closer look to the data from KSC we can see a success rate of over 75%, the only other site that come close to this success is the CCAFS LC-40 with only 3% less success rate.



Payload (0 – 5000 KG) x Launch Outcome

- A simple scatter plot showing the relation between the launch outcome and payload mass (filtered to only show data from 0 to 5000 KG).

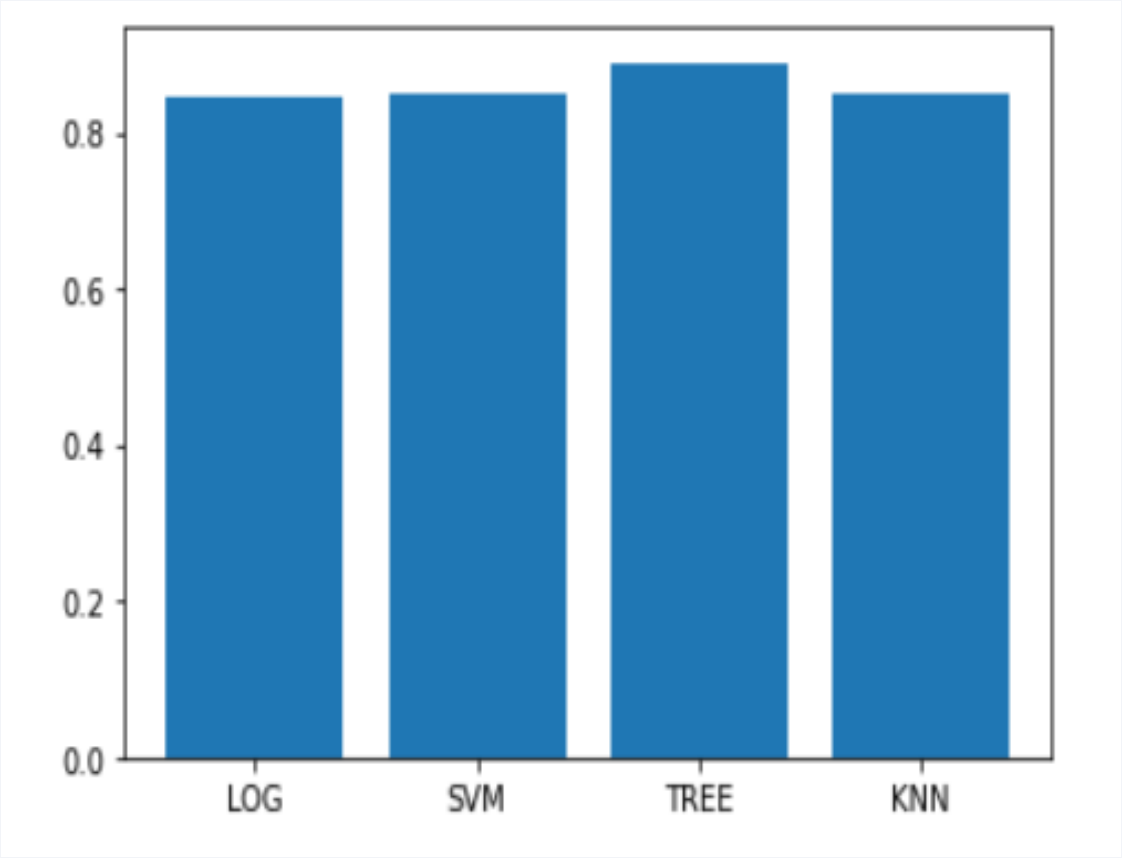


Section 5

Predictive Analysis (Classification)

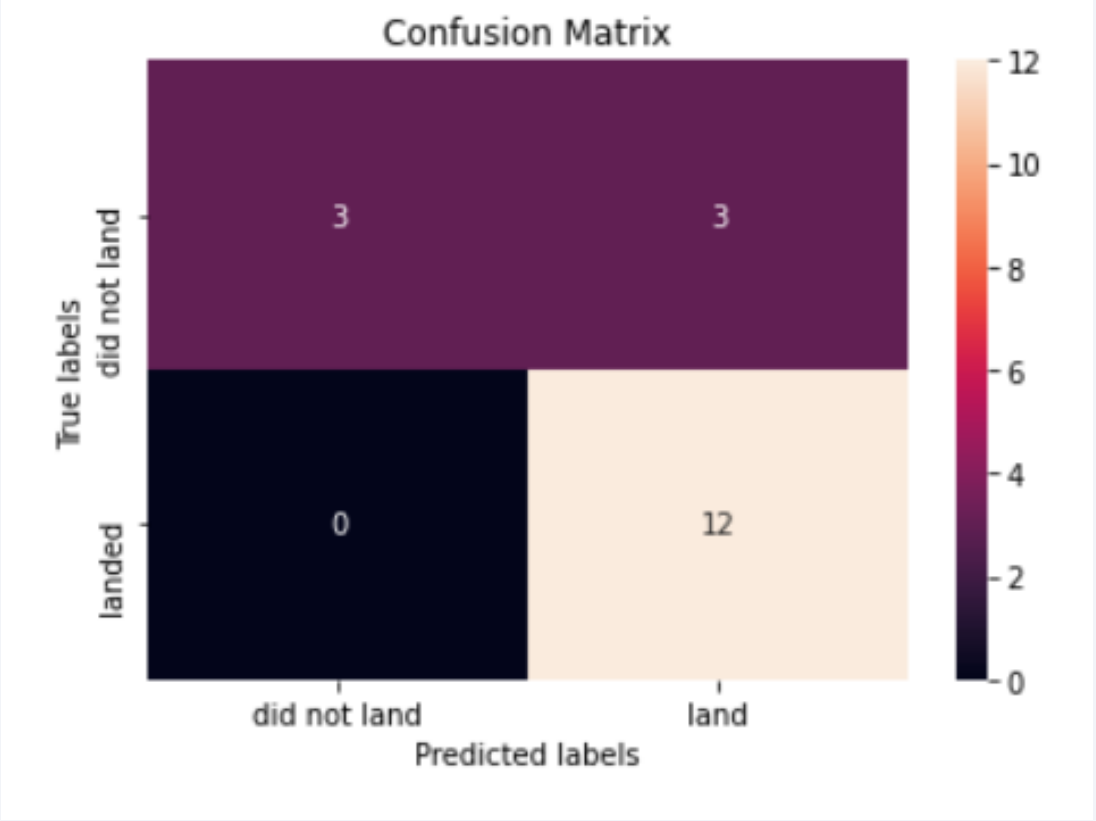
Classification Accuracy

- The best score from each model was used to create this Viz;
- The best model was the Decision Tree with an accuracy of 0.89.



Confusion Matrix

- The confusion matrix presented similar results for every model tested, with 3 TP, 3 FP, 0 FN and 12 TN;



Conclusions

- All launches success rates increased with time.
- The orbit type VLEO (very low earth orbit) had the best success rate taking in consideration the number of samples.
- The best landing outcomes were found in the KSC launch site, while the worse one was the CCAFS.
- By analyzing the “best_score_” from each model we conclude that Decision Tree was the best one for use.

Thank you!