Winning Space Race with Data Science

Sandro Joel Hernández Goicochea
June 2nd, 2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Space X has found great success in the race for commercial space travel because they have managed to reuse the stage 1 of their rockets. For this reason it is very important that the stage 1 landing on earth is successful. This work performs a statistical analysis to determine relationships between the parameters of each flight and the success or failure of the stage 1 landing. The parameters considered for the analysis are, for example, the launch site, payload mass, orbit, booster version , etc. In this analysis, classification methods such as Logistic regression, SVM, Tree or KNN were used to predict the outcome of new releases. GridSearchCV was used to determine the best parameters for each method and finally, the accuracies of the four methods were compared using the test set. The final result shows that the Tree method obtained the lowest precision of 0.7222, while the Logistic Regression, SVM and KNN methods present the same precision of 0.8333.

Introduction

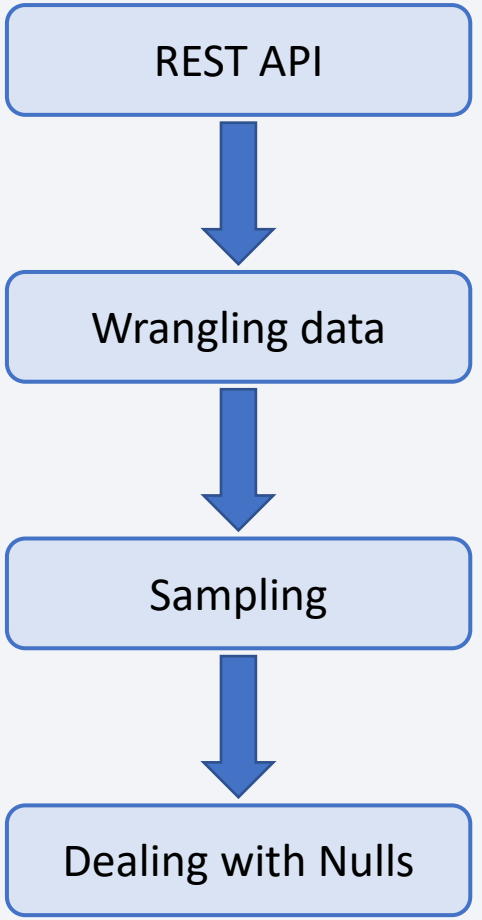
- We are in the commercial space age. Companies are making space travel affordable.
- Space X is maybe the most successful company in this area. They can do rocket launches relatively inexpensive.
- Much of the savings are because Space X can reuse the first stage of their rockets.
- However not in every launch the first stage land successfully. There are cases when the landing is failed.
- The problem that we want to solve is to predict if the landing of the first stage of rocket is successfully or not. We will determine it using the past rocket launch information such as payload, orbit and customer.

Section 1

Methodology

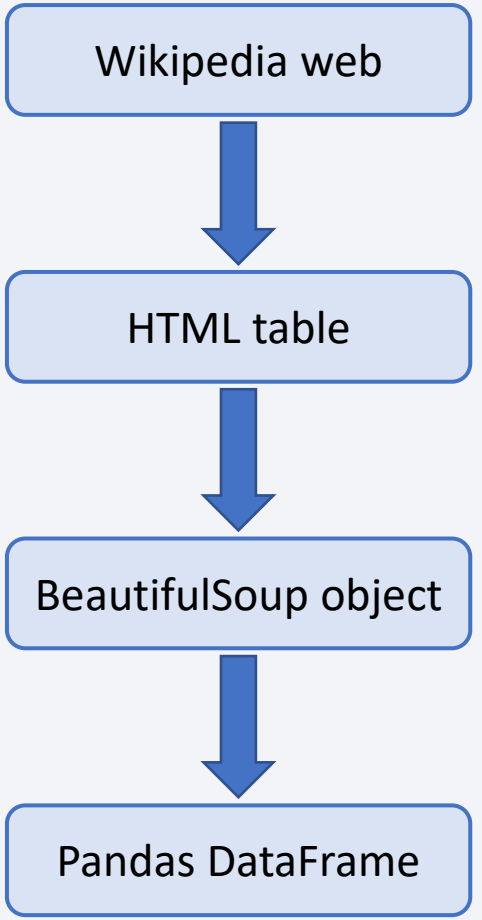
Data Collection – SpaceX API

- First json file is requested from REST API.
- Then pandas function json_normalize is used to get information as a dataframe.
- Finally, data is sampled and Null elements are removed.
- https://github.com/sandro-hernandez/IBM_Final_Capstone/blob/7871d9e5d124eabf082af678b3a749fe759f6923/Data%20Collection/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Using Requests and BeautifulSoup libraries Falcon 9 launch records HTML table is extracted from Wikipedia.
- Then the table is parsed and converted it to a dataframe.
- https://github.com/sandro-hernandez/IBM_Final_Capstone/blob/7871d9e5d124eabf082af678b3a749fe759f6923/Data%20Collection/jupyter-labs-webscraping.ipynb



Data Wrangling

- Data is wrangled in order to have a Pandas dataframe with just the relevant data and without null elements.
- Categorical data is converted in numerical through the use of dummy variables.
- https://github.com/sandro-hernandez/IBM_Final_Capstone/blob/7871d9e5d124eabf082af678b3a749fe759f6923/Data%20Wrangling/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- To determine the relationship between the different variables that characterize each flight, plots between some variables were made:
  1. Flight Number vs Launch Site
  2. Payload vs Launch Site
  3. Success rate vs Orbit type
  4. FlightNumber vs Orbit type
  5. Payload vs Orbit type
  6. Launch success yearly trent
- https://github.com/sandro-hernandez/IBM_Final_Capstone/blob/7871d9e5d124eabf082af678b3a749fe759f6923/Exploratory%20Data%20Analysis/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- SQL queries performed were made to find relevant information about the dataset. Queries carried out were about to find:
  - ❖ the names of the unique launch sites in the space mission
  - ❖ records where launch sites begin with the string 'CCA'
  - ❖ the total payload mass carried by boosters launched by NASA (CRS)
  - ❖ average payload mass carried by booster version F9 v1.1
  - ❖ the date when the first successful landing outcome in ground pad was achieved.
  - ❖ the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000 kilograms
  - ❖ the total number of successful and failure mission outcomes
  - ❖ the names of the booster_versions which have carried the maximum payload mass. Use a subquery
  - ❖ the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - ❖ the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order
- https://github.com/sandro-hernandez/IBM_Final_Capstone/blob/7871d9e5d124eabf082af678b3a749fe759f6923/Exploratory%20Data%20Analysis/jupyter-labs-eda-sql-coursera.ipynb

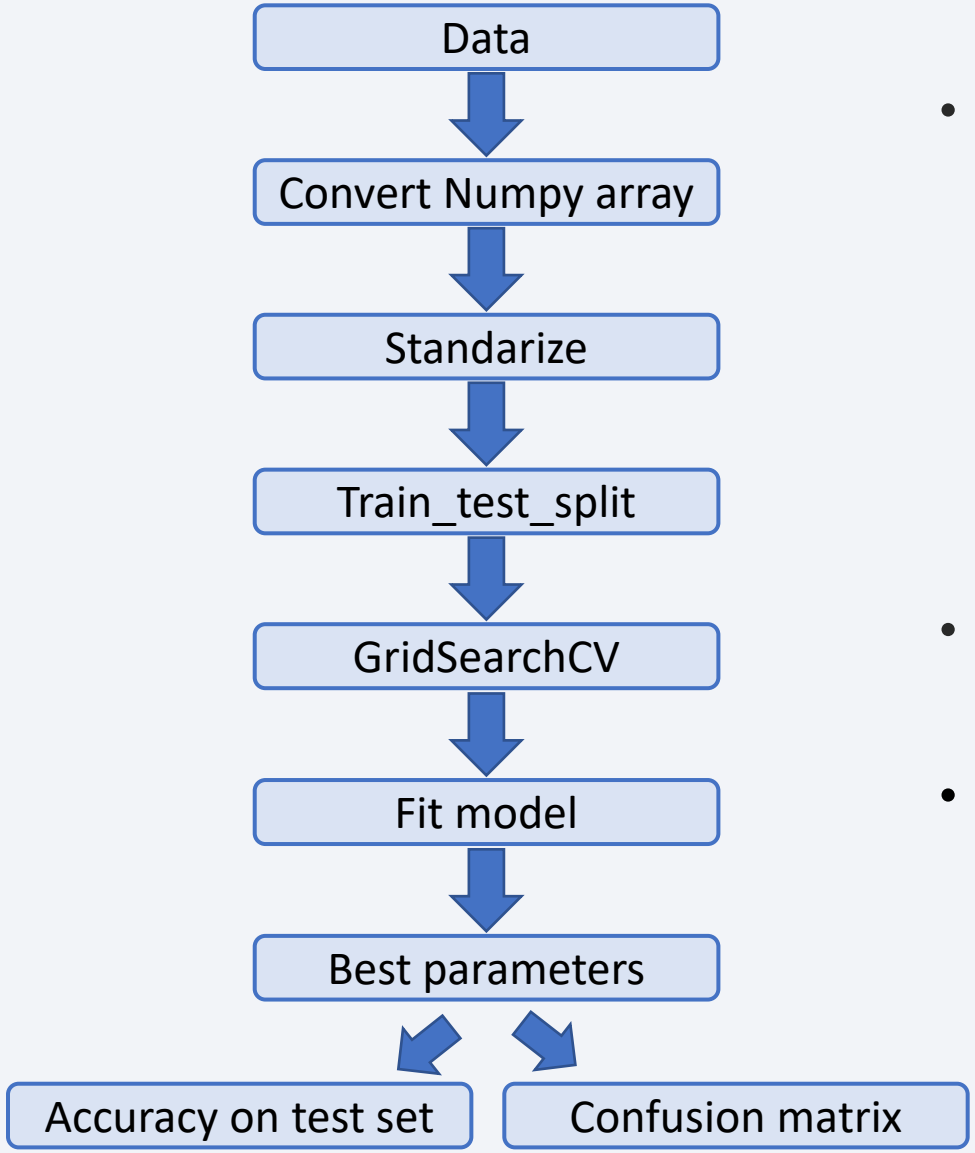
Build an Interactive Map with Folium

- As part of the analysis an interactive map was created with some markers:
    1. Markers and circles in Launch sites: To identify clearly were each Launch site is located in the map
    2. Create marker cluster: To visualize how many successes and failures correspond to each Launch site.
    3. Polyline and markers indicating distances to its proximities: to compare and derive geospatial location information.
- https://github.com/sandro-hernandez/IBM_Final_Capstone/blob/7871d9e5d124eabf082af678b3a749fe759f6923/Interactive%20Visual%20Analytics%20and%20Dashboard/lab_jupyter_launch_site_location.ipynb

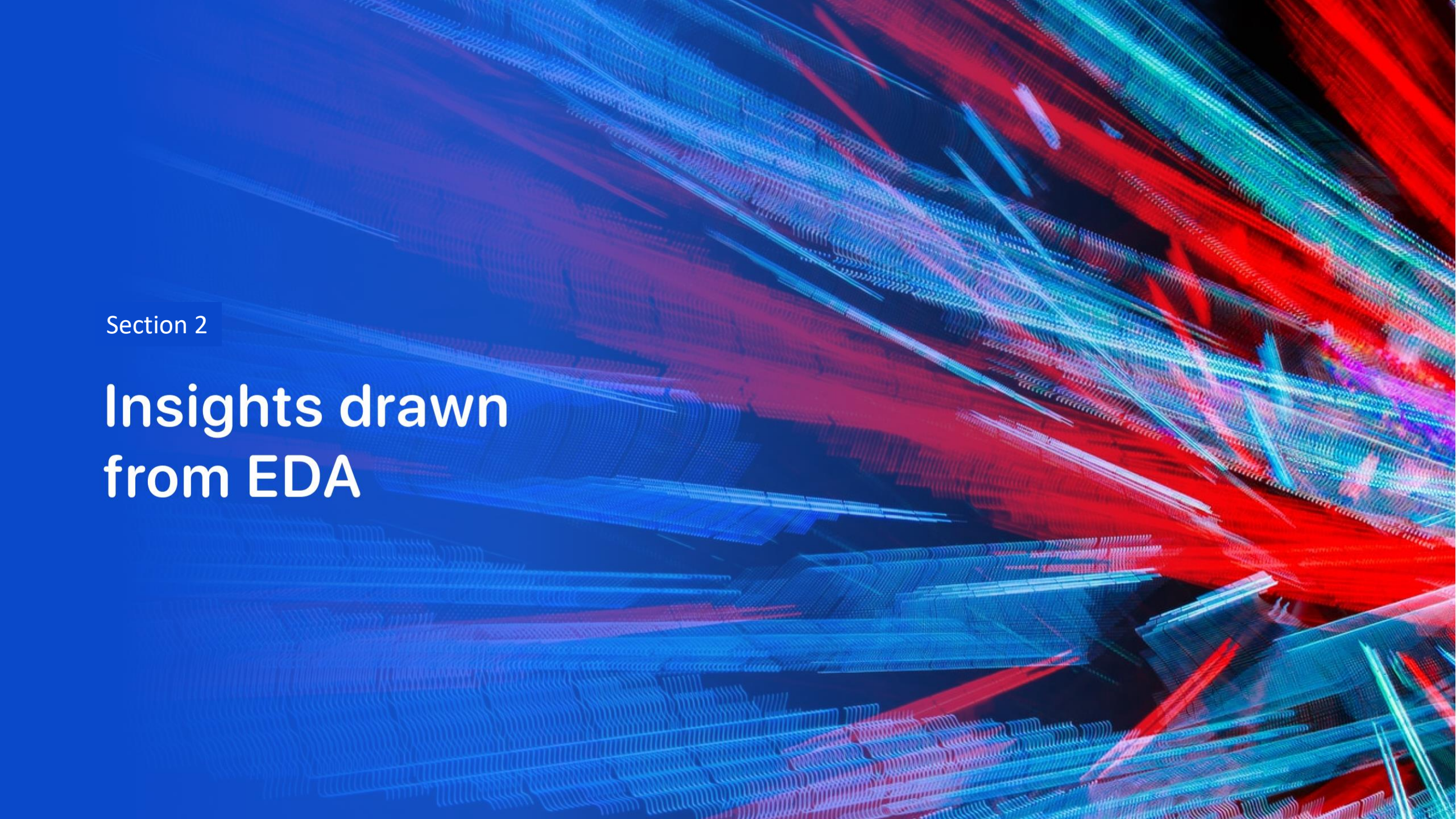
Build a Dashboard with Plotly Dash

- Dashboard shows important information visually. Graphs created were:
  - Pie chart of launch success count for each sites.
  - Points plot of Payload vs. Launch Outcome for all sites, with a slide to select different payload in the range of 0 and 10000 kg. Points colors represent the different Booster Version Category.
- https://github.com/sandro-hernandez/IBM_Final_Capstone/blob/7871d9e5d124eabf082af678b3a749fe759f6923/Interactive%20Visual%20Analytics%20and%20Dashboard/spacex_dash_app.py

Predictive Analysis (Classification)

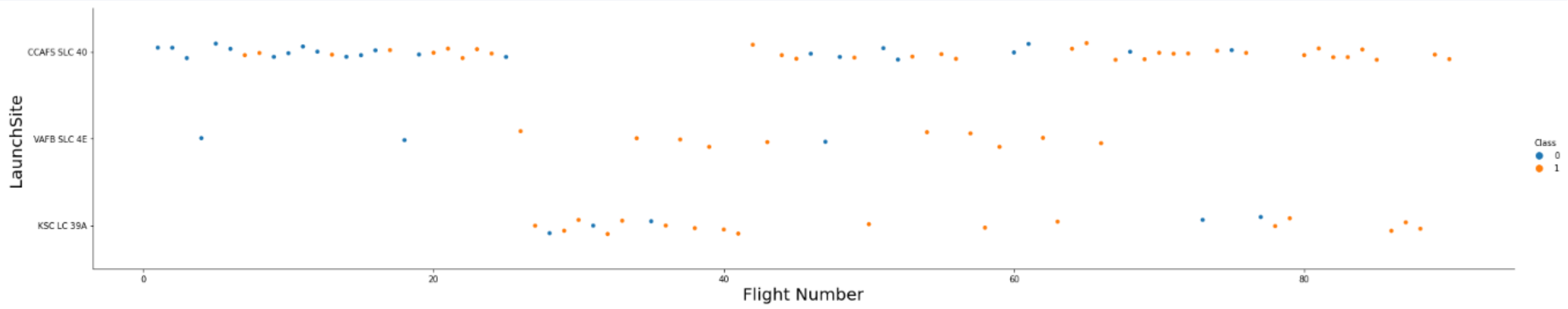


- The process was realized for the following classification methods:
  - logistic regression
  - support vector machine
  - decision tree classifier
  - k nearest neighbors
- From the four results, the method with the best accuracy was selected.
- https://github.com/sandro-hernandez/IBM_Final_Capstone/blob/7871d9e5d124eabf082af678b3a749fe759f6923/Predictive%20Analysis/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb


Section 2
Insights drawn from EDA

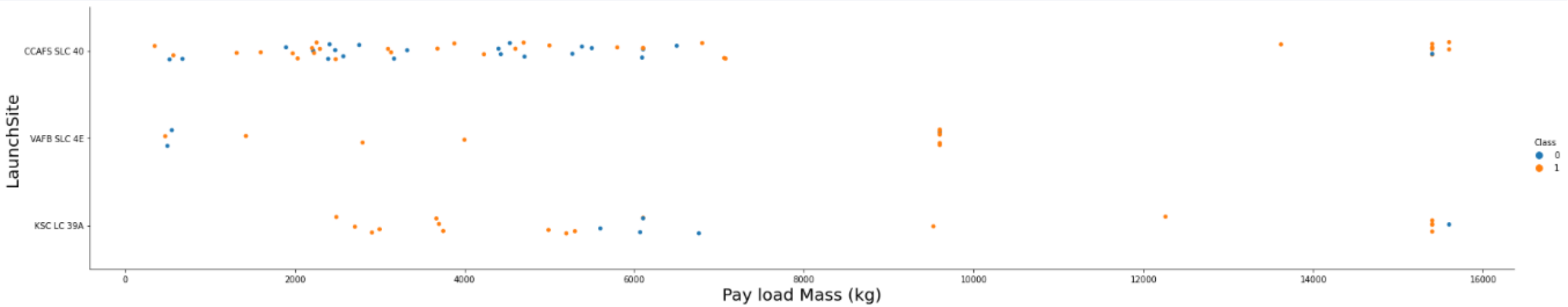
Flight Number vs. Launch Site

- Figure shows a scatter plot of Flight Number vs. Launch Sites
- The largest number of launches were made from CCAFS SLC-40.
- From launch number 80 onwards the launches were successful.



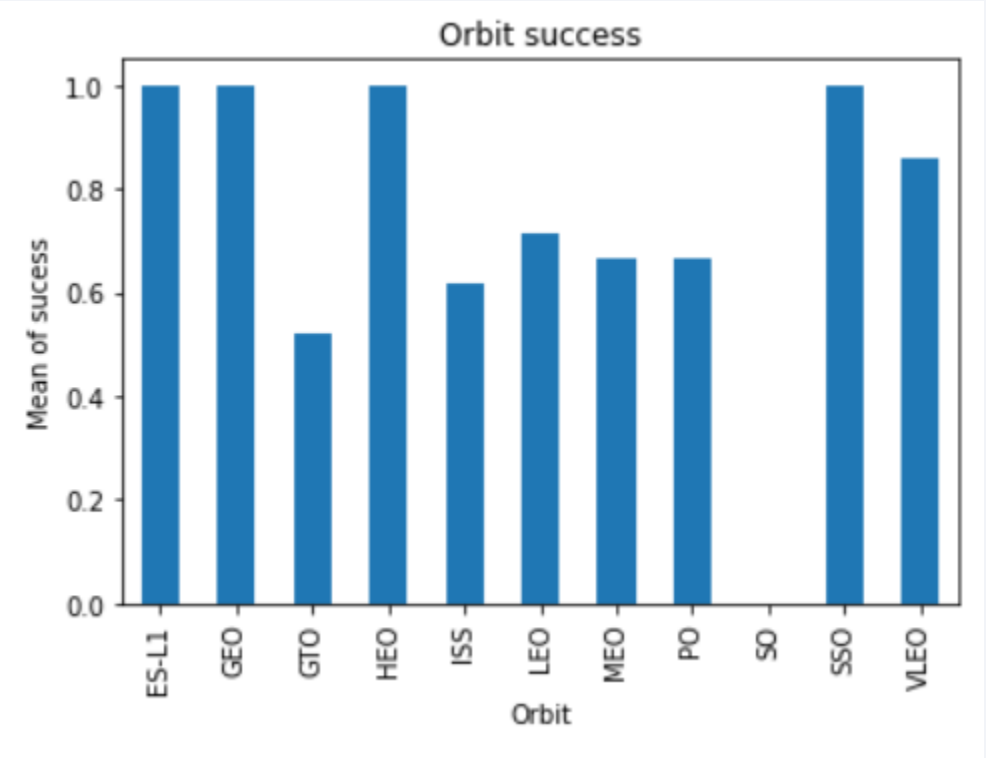
Payload vs. Launch Site

- Figure shows a scatter plot of Payload Mass in kilograms vs. Launch Sites.
- Launches made from VAFB SLC 4E with payloads greater than 1000 kilos were successful.
- Launches made from KSC LC 39A with payloads lower than 5000 kilos were successful.



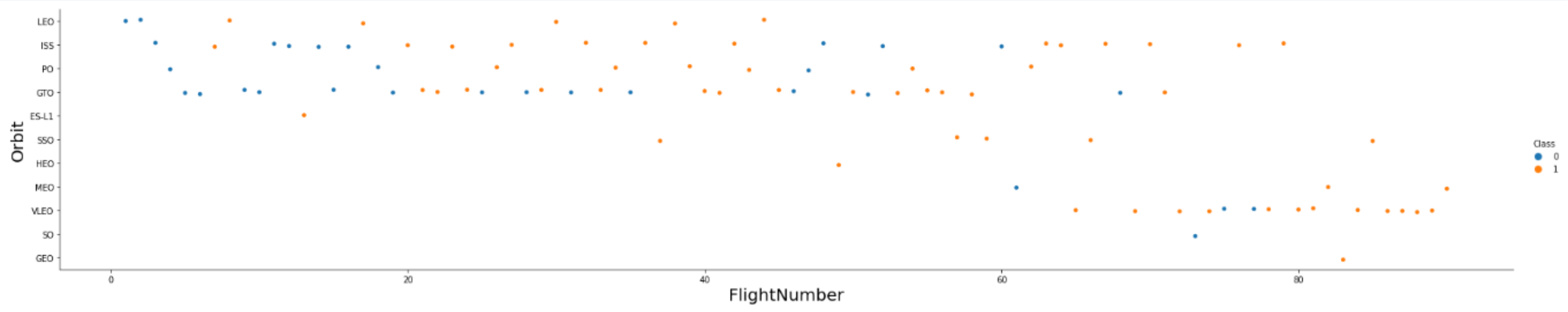
Success Rate vs. Orbit Type

- Figure shows a bar chart of the success rate of each orbit type.
- All the launches with orbits ES-L1, GEO, HEO or SSO were 100% successful.
- Launches with orbit type GTO have the lowest success rate



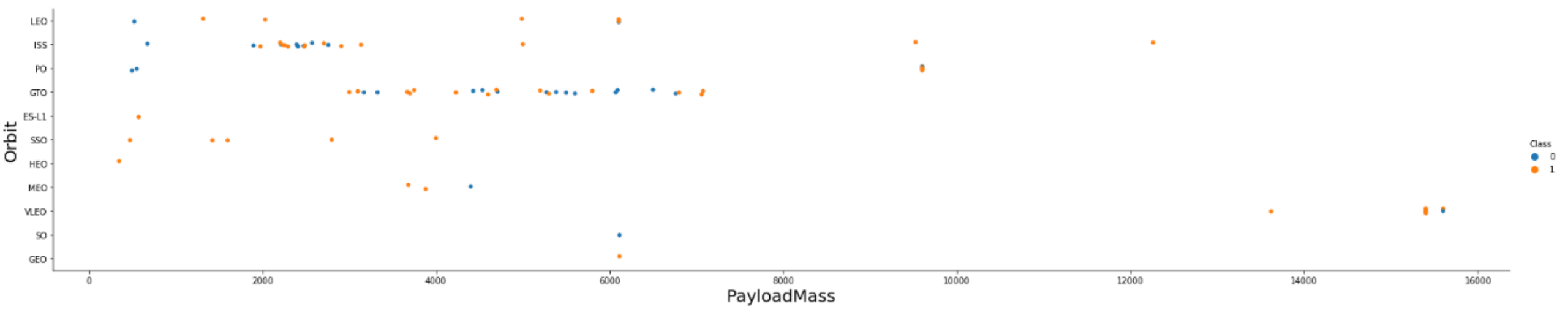
Flight Number vs. Orbit Type

- Figure shows a scatter plot of Flight number vs. Orbit type
- Launches with orbits ES-L1, GEO, HEO or SSO were 100% successful, but were few.
- Launches with orbit type LEO and FlightNumber above 5 were successful.
- The largest number of launches were made with orbit type ISS or GTO.



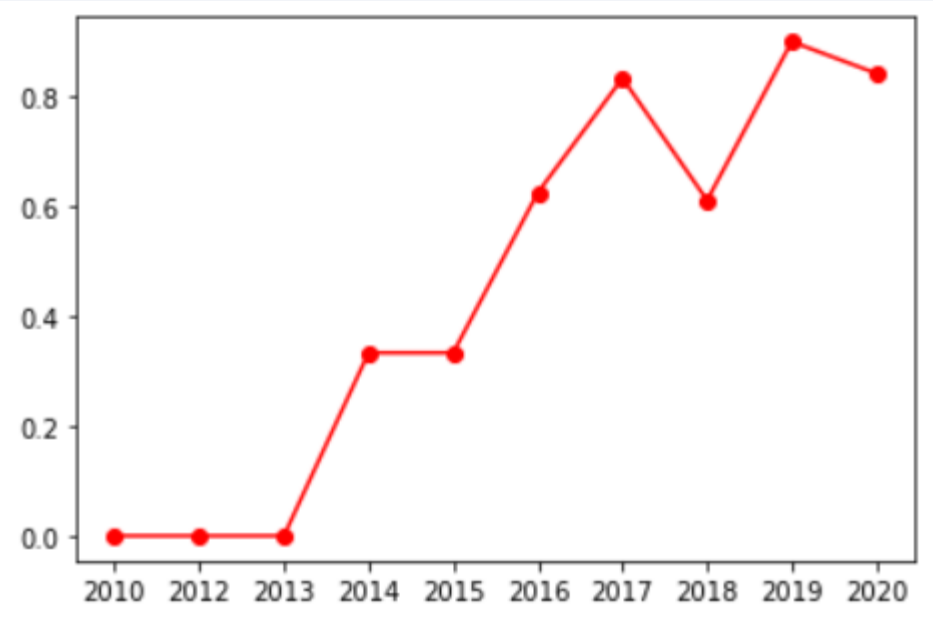
Payload vs. Orbit Type

- Figure shows a scatter plot of Payload vs. Orbit type
- Launches with orbit type ISS and Payload above 3000 kg were successful.
- Launches with Payload above 8000 kg were made with orbits ISS, PO or VLEO.



Launch Success Yearly Trend

- Figure shows a line chart of yearly average success rate.
- There is an upward trend. Success increases over time except in the year 2018.
- Launch success starts in 2013.
- In 2019 success rate almost reaches 100 %.



All Launch Site Names

- Find the names of the unique launch sites
- Present your query result with a short explanation here

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Present your query result with a short explanation here

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

- The total payload by boosters from NASA in kilograms is:

1
45596

Average Payload Mass by F9 v1.1

- The average payload mass in kilograms carried by booster version F9 v1.1 is:

1
2928

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad is:

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 kg but less than 6000 kg are:

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

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission from different outcomes:

landing_outcome	2
Controlled (ocean)	5
Failure	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	22
Precluded (drone ship)	1
Success	38
Success (drone ship)	14
Success (ground pad)	9
Uncontrolled (ocean)	2

Boosters Carried Maximum Payload

- Names of the boosters which have carried the maximum payload mass.

booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015:

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

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

- 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:

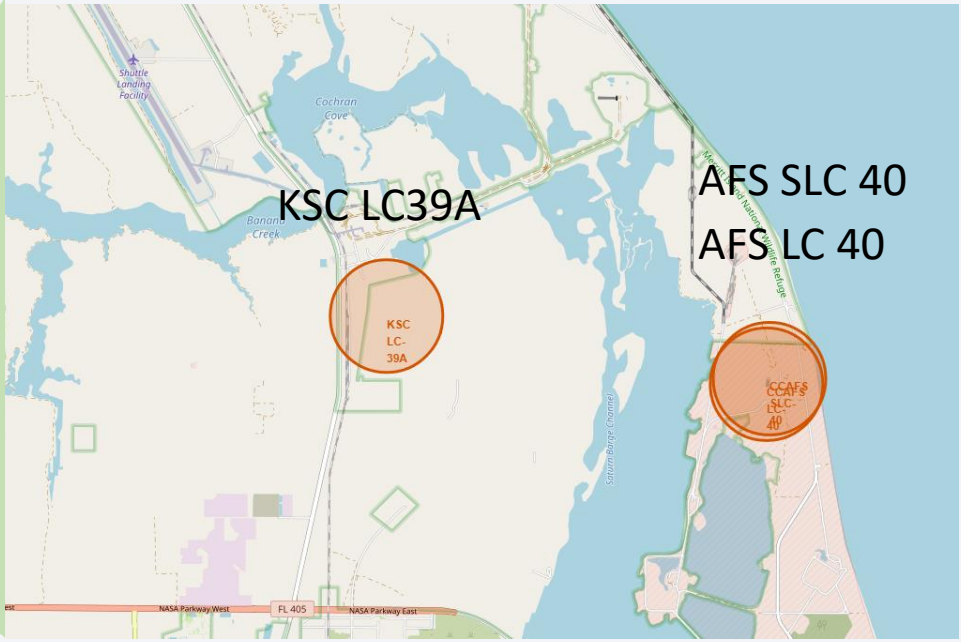
landing_outcome	2
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

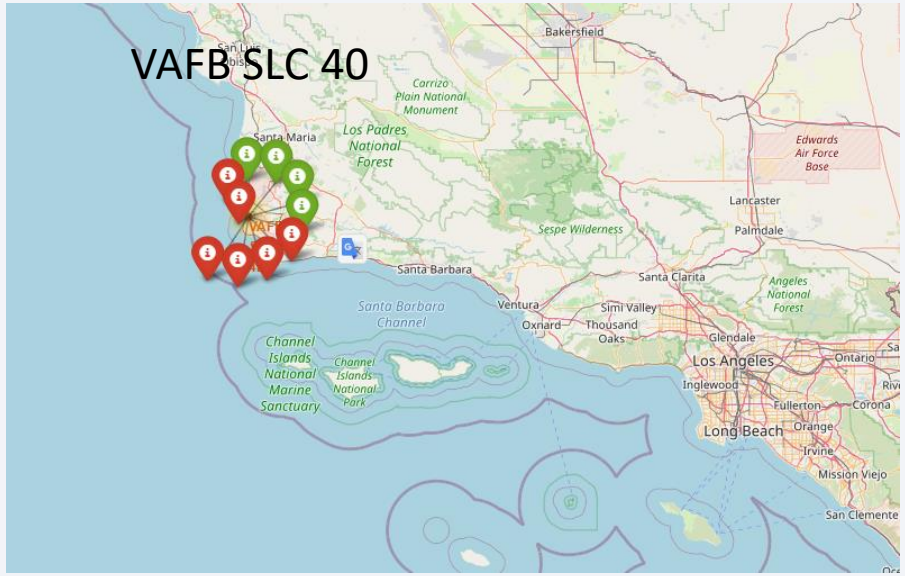
Geospatial location of launch sites

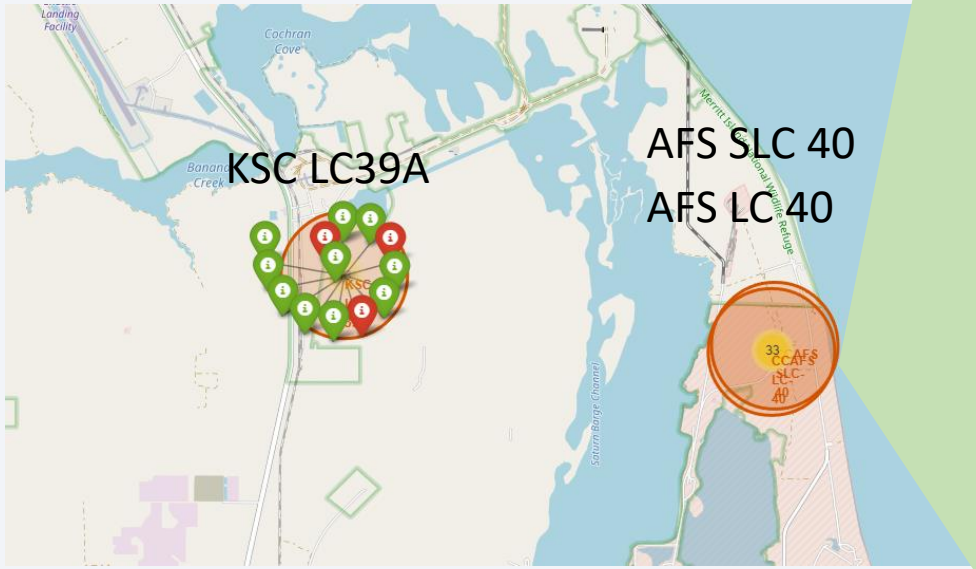
- The four launch sites are located in USA.
- VAFB SLC 40 is located in west coast and the rest (KSC LC39A, AFS SLC 40 and AFS LC 40) are located in the east coast.

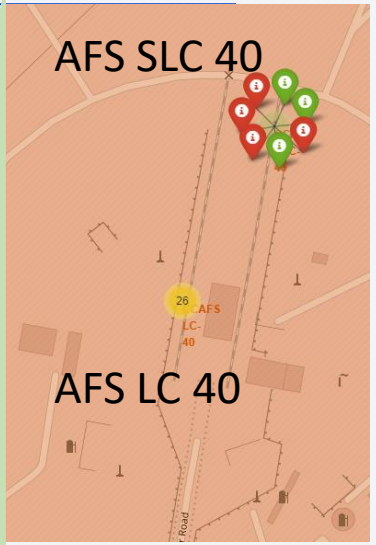


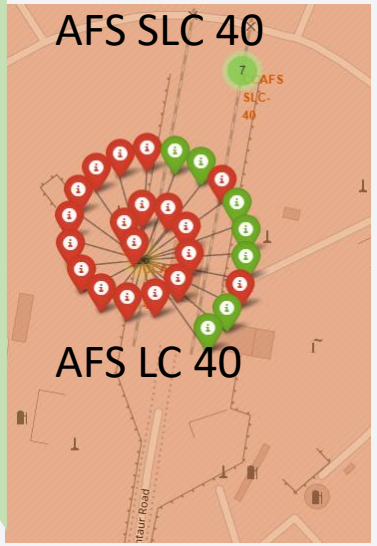
Launch outcomes on the map

- Red markers represents failures and green markers, successful launches.
- Launches were made mostly from AFS LC 40.
- The greatest success rate is from KSC LC 39A.



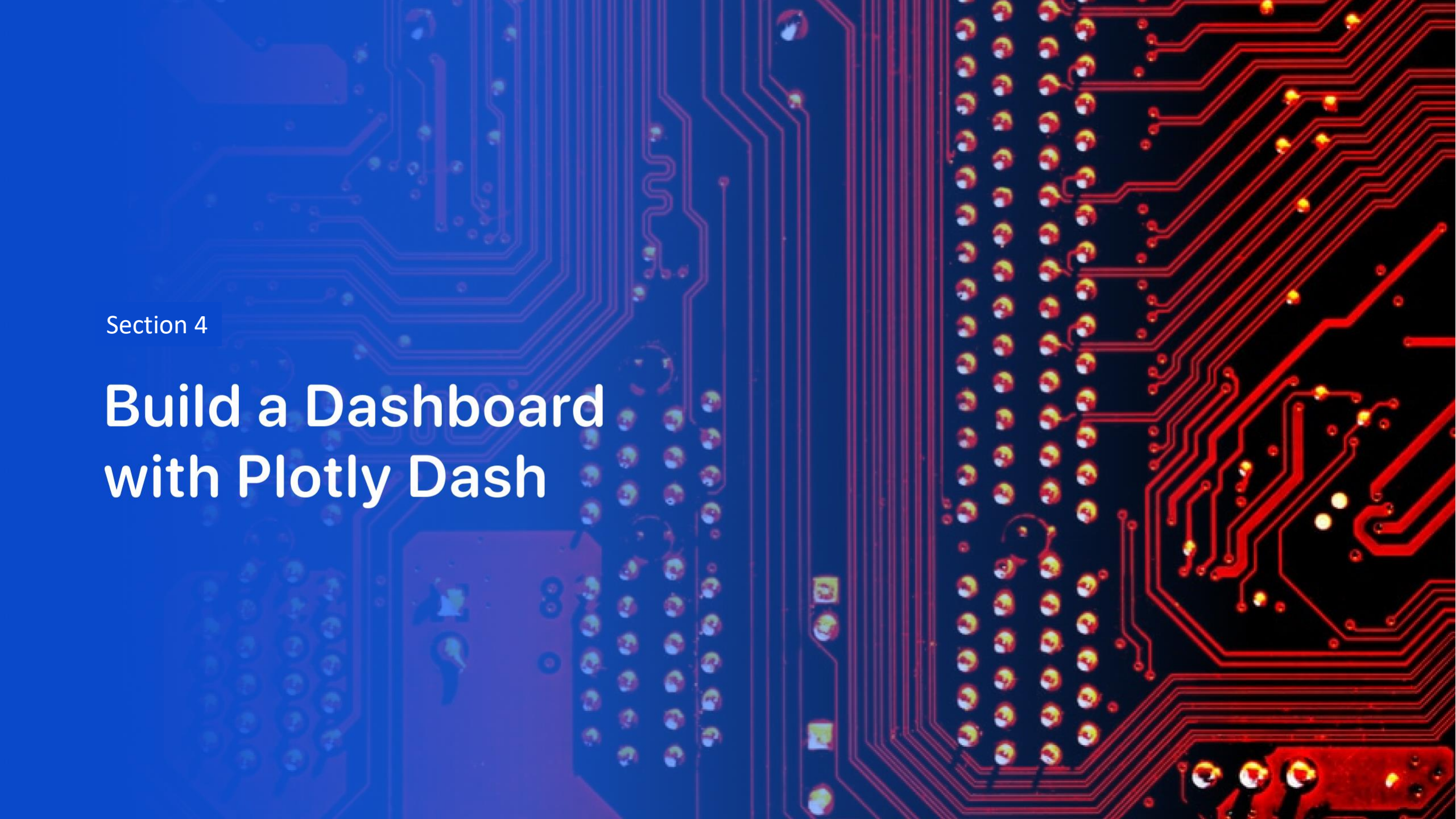






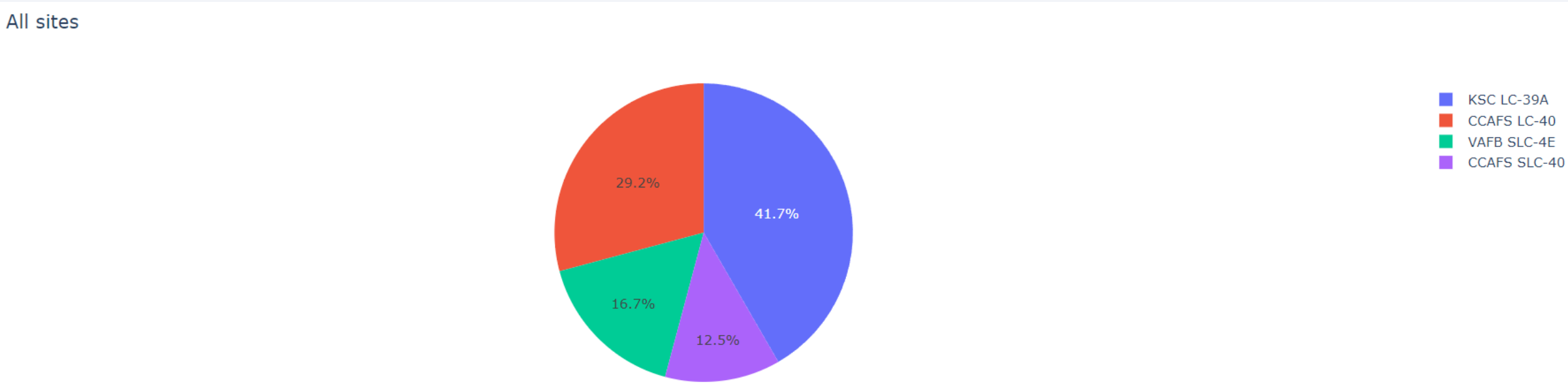
Distances to proximities for AFS SLC-40

- Figure shows distance from a selected launch site, AFS SLC 40, to its proximities such coastline, with distance calculated and displayed


Section 4
Build a Dashboard with Plotly Dash

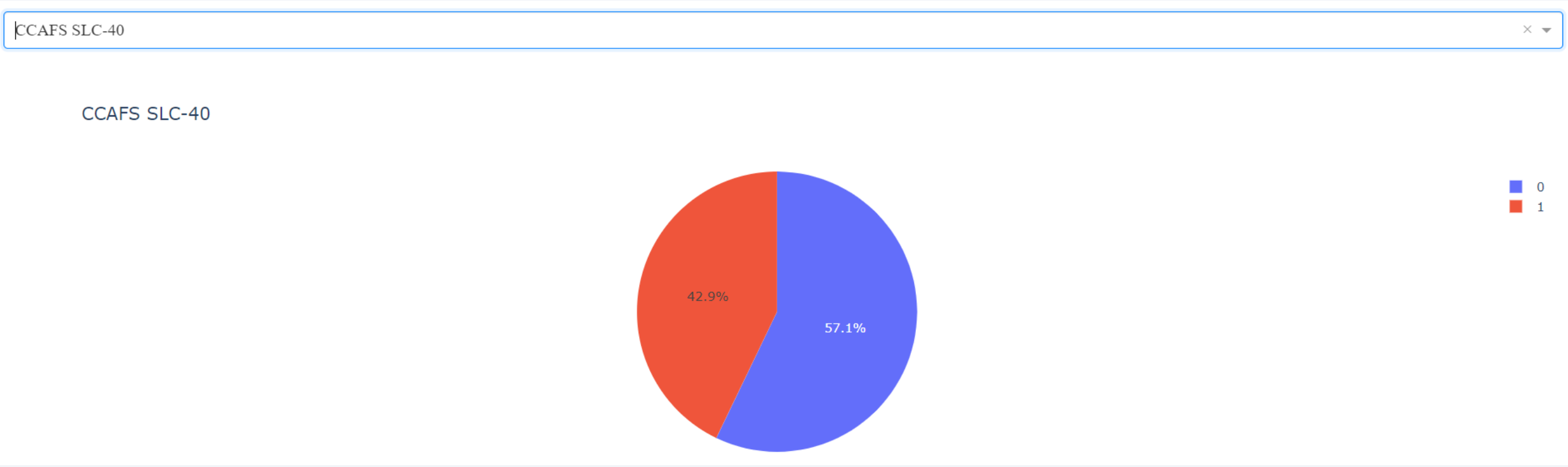
Launch sites success rate

- Figure shows the rate of launch success count for each sites.
- The most successful launches were from KSC LC-39A. It is almost half of all the successful launches.
- The lowest rate is from CCAFS SLC-40 12.5%.



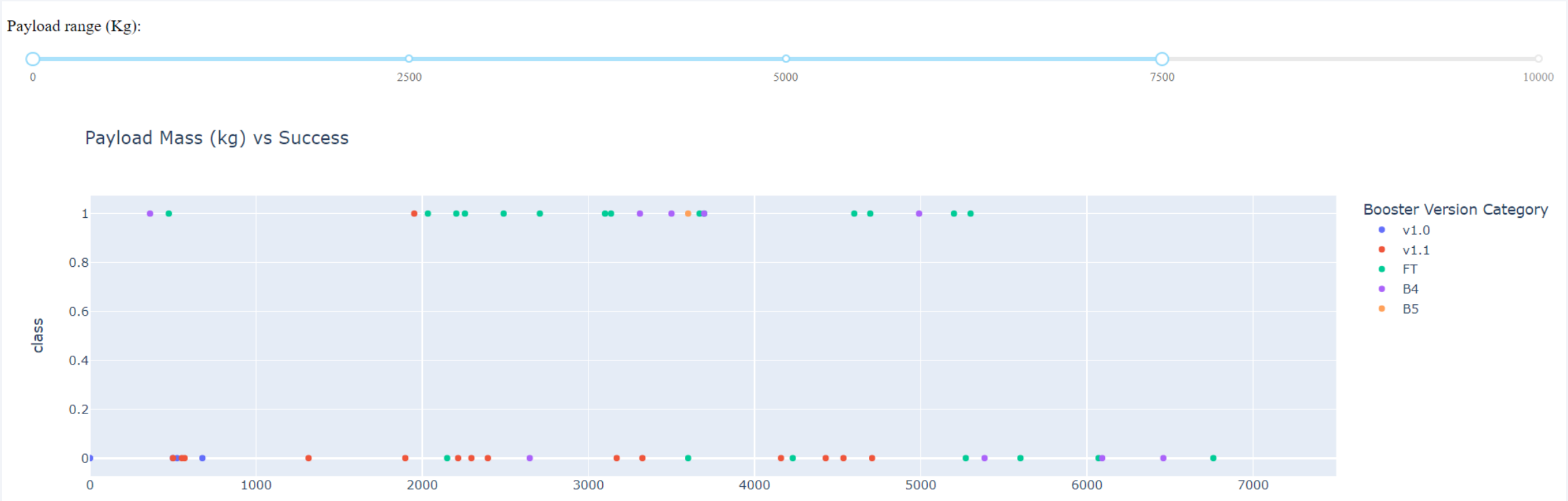
CCAFS SLC-40 success rate

- Figure shows the pie chart for the launch site with highest launch success ratio.
- CCAFS SLC has the least quantity of successful launches. However, it has the greatest success ratio (57.1%).



Payload vs Launch Outcome

- Figure shows a scatter plot of Payload vs. Launch Outcome for all sites, with different payload selected in the range of 0 and 7500 kg. Colors represent the different Booster Version Category.
- Above 5000 kg, Booster version B4 has 100% success ratio.
- Booster version v1.1 has just one successful launch.

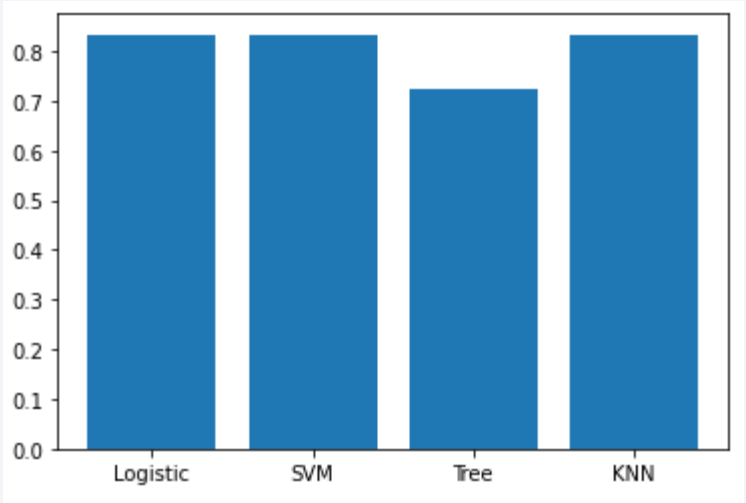


Section 5

Predictive Analysis (Classification)

Classification Accuracy

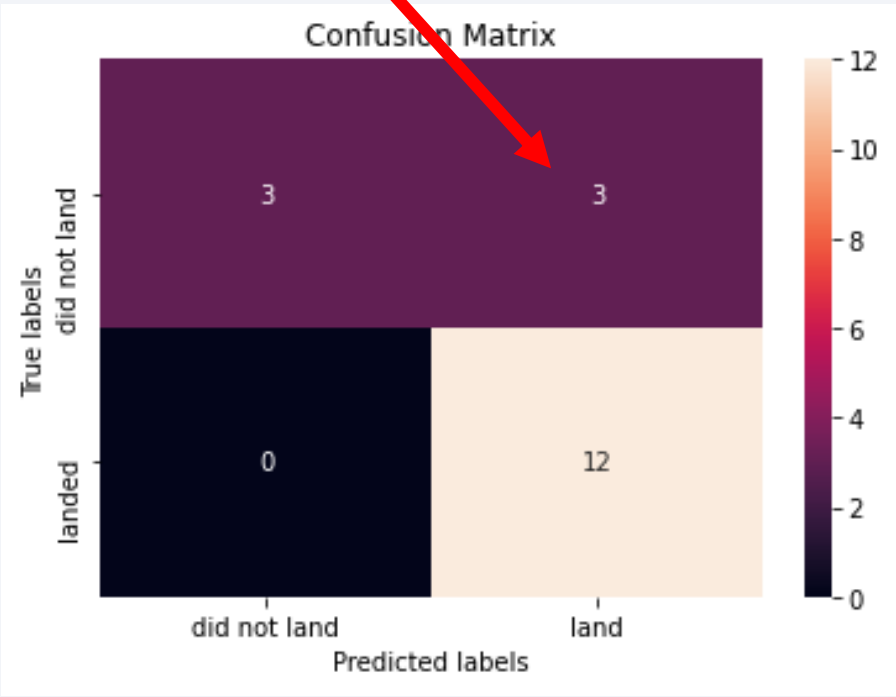
- Figure shows the built model accuracy for all built classification models, in a bar chart.
- The models with the highest classification accuracy are three: Logistic Regression, SVM and KNN algorithm. They have an accuracy of 0.8333.
- The parameters used to split data were test_size=0.2 and random_state=2.
- The best kernel for SVM was sigmoid.
- The number k of neighbors for KNN algorithm was 9.



	Method	Score
0	Logistic	0.833333
1	SVM	0.833333
2	Tree	0.722222
3	KNN	0.833333

Confusion Matrix

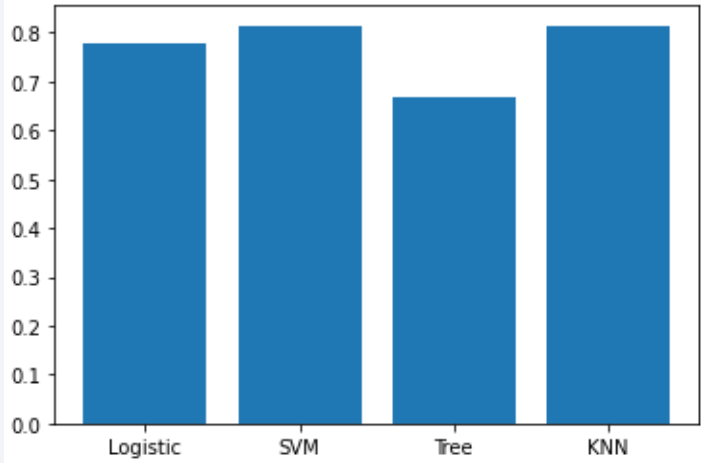
- Figure shows the confusion matrix of one of the best performing models, Logistic Regression.
- We find that there are 3 false positives and none false negatives.

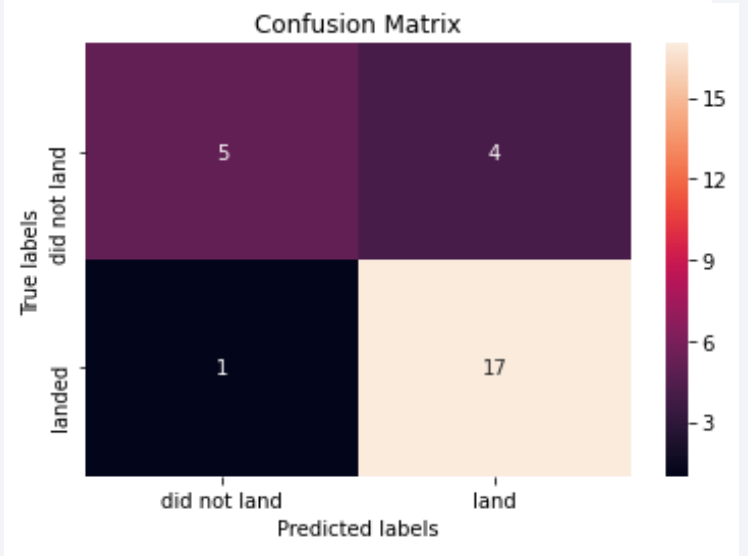


Classification Accuracy using test_size=0.3 and random_state=10

- Figure shows the built model accuracy for all built classification models, in a bar chart.
- The models with the highest classification accuracy are two: SVM and KNN algorithm. They have an accuracy of 0.8148.
- The parameters used to split data were test_size=0.3 and random_state=10.
- The best kernel for SVM was sigmoid.
- The number k of neighbors for KNN algorithm was 5.

	Method	Score
0	Logistic	0.777778
1	SVM	0.814815
2	Tree	0.666667
3	KNN	0.814815





Conclusions

- For payloads lower than 5000 kilos is convenient to use KSC LC 39A. Because all the launches in that range have been successful.
- Launches to orbit types ISS and GTO were more common than others. However, their success rates is relatively low .
- The method that should be used to make predictions has to be selected between Logistic Regression, SVM and KNN algorithm. They have the greatest accuracy, 8333, in the three methods.
- However, if we change the parameters to split the data in train and test sets, results vary and Logistic accuracy decreases below SVM and KNN algorithm accuracies.

Thank you!