IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Amit Sehrawat
June – 7, 2022

Outline

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

Executive Summary (Methodology)

- Gathered SpaceX launch data using two different methods.
- Gathered data is stored using pandas data frame data structure and csv file.
- Data contains information about flight number, data, payload, launch outcome, launch sites, and many more such info.
- Performed EDA on collected using python and SQL.
- Performed interactive visualization using plotly and folium on the gathered data.
- Built 4 type of classifiers to predict the launch outcome of mission.

Executive Summary (Results)

- There are four launch sites, and all are located near coastline.
- There are clear relations between different variables like flight no. vs launch outcome at different locations and many such, results of which are presented in section . Performed interactive visualization using plotly and folium on the gathered data.
- Success rate increased significantly onwards 2013.
- KSC LC-39A has the highest success rate of 41.7 % among all launch sites. CCAFS SLC-40 has the lowest success rate of 12.5 % among all launch sites.
- All the four classifiers based on LR, SVM, DT, and KNN performed similar on test set with an accuracy of around 83.3 %.

Introduction

Project background and context

- “Space Exploration Technologies Corp. is an American spacecraft manufacturer, space launch provider, and a satellite communications corporation headquartered in Hawthorne, California.[1]”
- The company was founded with focus on reducing the space transportation costs to make space exploration more accessible.
- In this project, we performed the detailed analysis on the history of rocket launched by SpaceX and their success and failure outcome. We have looked in different factors affecting the outcome of a launch.

Introduction

Problems you want to find answers

- We want to look into how different launch variables like launch site, payload mass, booster versions, and many more factors influence a launch's outcome.
- This project aims to create a machine learning model to predict if the first stage will land given the data from the prior launches.
- The project's ultimate goal is to build a predictive model to evaluate the future outcome with acceptable accuracy to determine the cost of the successful mission.
- This model can then be used by SpaceX itself or another entity that wants to compete with SpaceX in space exploration.

Section 1

Methodology

Methodology

Data collection methodology:

- Using SpaceX public API.
- Using web scrapping of the wiki page of SpaceX launch.

Perform data wrangling

- The collected data is then cleaned using pandas and NumPy packages.
- We removed null value using some meaningful values like mean or other significant values.
- Created landing outcome label e.g., ”1” for mission success and “0” for failure.

Methodology

- Perform exploratory data analysis (EDA) using visualization and SQL.
- Perform interactive visual analytics using Folium and Plotly Dash.
- Perform predictive analysis using classification models
    - We used four different classification models and optimized hyperparameters to find the best model and parameters set.

Data Collection

- Data was collected using two different methods.

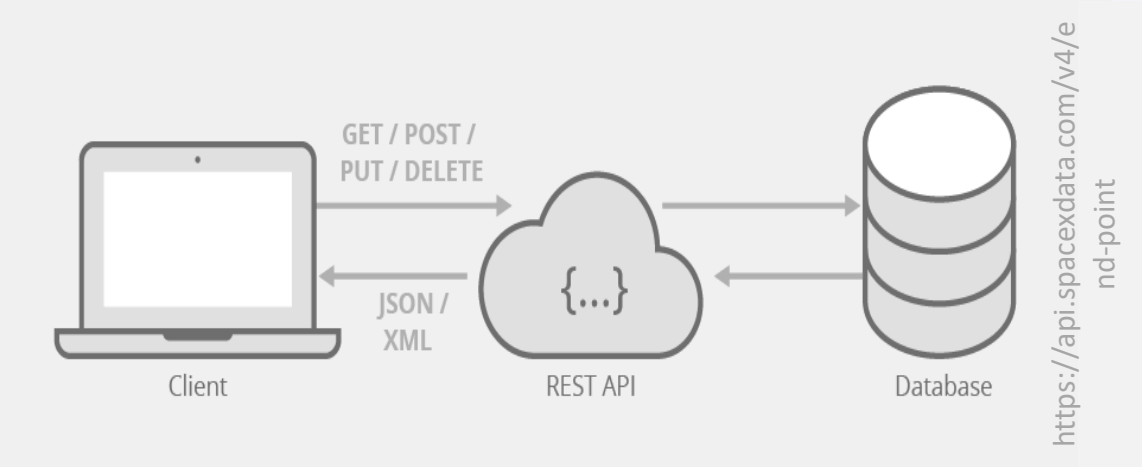
    1. Web scraping by using python *requests* and *Beautiful Soup* libraries.
    2. Using get request to the SpaceX API.

Data Collection – SpaceX API

- Make a get request to the SpaceX API using different resources like launch data, payload data, and booster version.

(refer: week1_spacex_data_collection_api.ipynb)

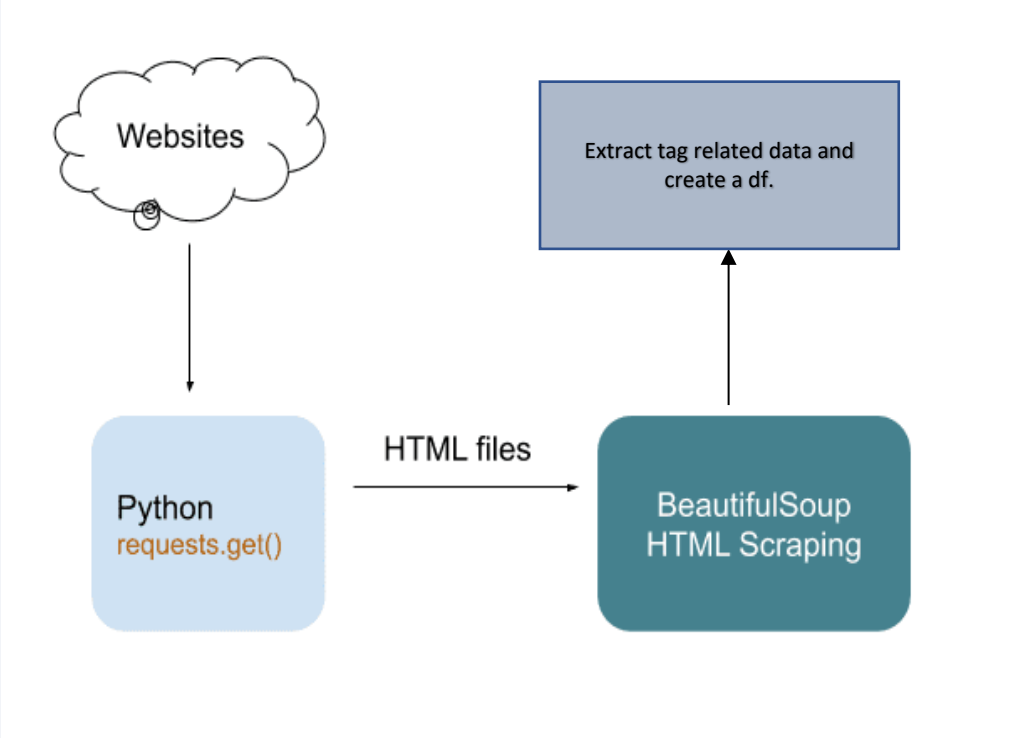
https://github.com/shera-amit/Coursera-capstone-final.git



Ref: https://www.sqlshack.com/create-rest-apis-in-python-using-flask/

Data Collection - Scraping

- Extract a Falcon 9 launch records HTML table from Wikipedia
- Parse the table and convert it into a Pandas data frame



(refer: week1_Data_Collection_with_Web_Scraping.ipynb)

https://github.com/shera-amit/Coursera-capstone-final.git

Data Wrangling

- Data frame (df) created using above mentioned methods like web scrapping and via API need to be cleaned to use in modelling or analytics.

- First, we calculated missing values in the df and then replaced those values with appropriate significant values for analysis purpose.

- We replaced missing payload mass values with mean value of payload mass column.

(refer: week1_spacex_data_collection_api.ipynb)

https://github.com/shera-amit/Coursera-capstone-final.git

EDA with Data Visualization

- We created categorical plot for:

  FlightNumber vs PayloadMass, FlightNumber vs LaunchSite, PayloadMass vs LaunchSite, FlightNumber vs Orbit.

- For EDA, we created categorical plots to evaluate the relationship between different variable (columns) in the dataset. Using categorical plots, we can infer the relationship between two plotted variables visually.
- We then created a bar plot showing success rate for different orbits.
- We also created line plot to show year wise success rate for launches.

(refer: week2_EDA_with_Data_Visualization.ipynb)

https://github.com/shera-amit/Coursera-capstone-final.git

EDA with SQL

- Unique launch sites in the space mission.
- 5 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 total number of successful and failure mission outcomes.

EDA with SQL

- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- 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 between the date 2010-06-04 and 2017-03-20, in descending order

(refer:week2_EDA_with_SQL.ipynb) https://github.com/shera-amit/Coursera-capstone-final.git

Build an Interactive Map with Folium

1. Markers for launch site on world map.

- Markers for launch site make it easy to visualize the geography of location.
- Markers helps to understand the relationship between location geographical like

    Are all launch sites in proximity to the Equator line?

    Are all launch sites in very close proximity to the coast?

Build an Interactive Map with Folium

2. Marker cluster for launch site on world map.

- We also wanted to see the success and failures for each launch location on the map.
- Marker clusters can be a good way to simplify a map containing many markers having the same coordinate.

Build an Interactive Map with Folium

3. Add lines to analyze and explore the close proximity to the launch location site.

- By drawing lines and labelling the distances, it is easy to infer the nearby location proximity to the marked site.
- Here, we calculated the distance to near coastal line, highway and railway.

(refer: week3_Interactive_Visual_Analytics_with_Folium_lab.ipynb)
https://github.com/shera-amit/Coursera-capstone-final.git

Build a Dashboard with Plotly Dash

- Added pie chart to show the success rate corresponding to different sites.
- Added categorical plot to show the outcome of the launch as a function of payload mass (Kg).
- Added Dropdown menu to select a particular launch site or all sites.

  This feature helps to visualize the results for a particular launch site.
- Created a Rangeslider to select the payload mass range.

  This feature make it easy to visualize the outcome for a particular payload mass range.

(refer: week3spacex_dash_app.py)
https://github.com/shera-amit/Coursera-capstone-final.git

Predictive Analysis (Classification Models)

In this exercise, we built four different classifications models.

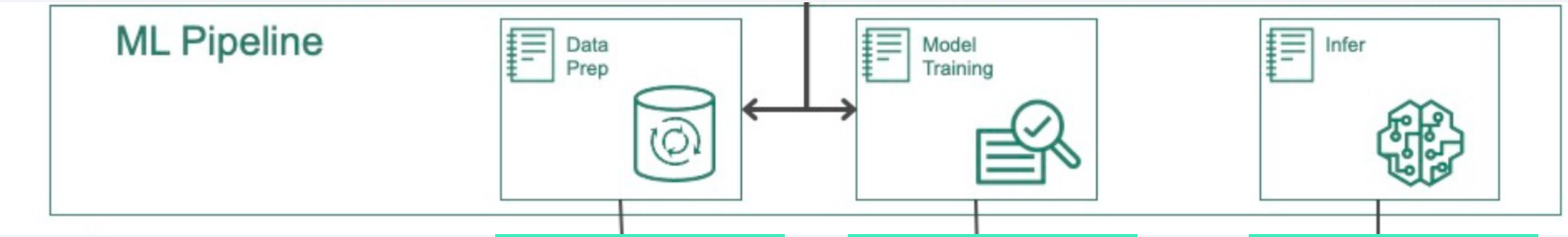
- Logistic regression
- Decision tree
- Support vector machine (SVM)
- K nearest neighbor (KNN)

(refer: week4_SpaceX_Machine Learning Prediction_Part_5.ipynb)
https://github.com/shera-amit/Coursera-capstone-final.git

Predictive Analysis (Model building)

- We loaded the clean data gathered from previous exercises.
- Extracted feature set (predictor variables) and outcome class (predicted variable).
- Splitted the data in training and test dataset using sklearn library for training and test purpose.
- Using sklearn grid search functionality, we iterate the model using multiple hyper-parameter values and found the best set of hyperparameters.
- Model evaluation was performed using accuracy score and confusion matrix metrics.
- We also evaluated the test set accuracy to find the best performing model.

Predictive Analysis (Flow chart)



1. Collected and wrangled using web scrapping, API.

2. Feature engineering.

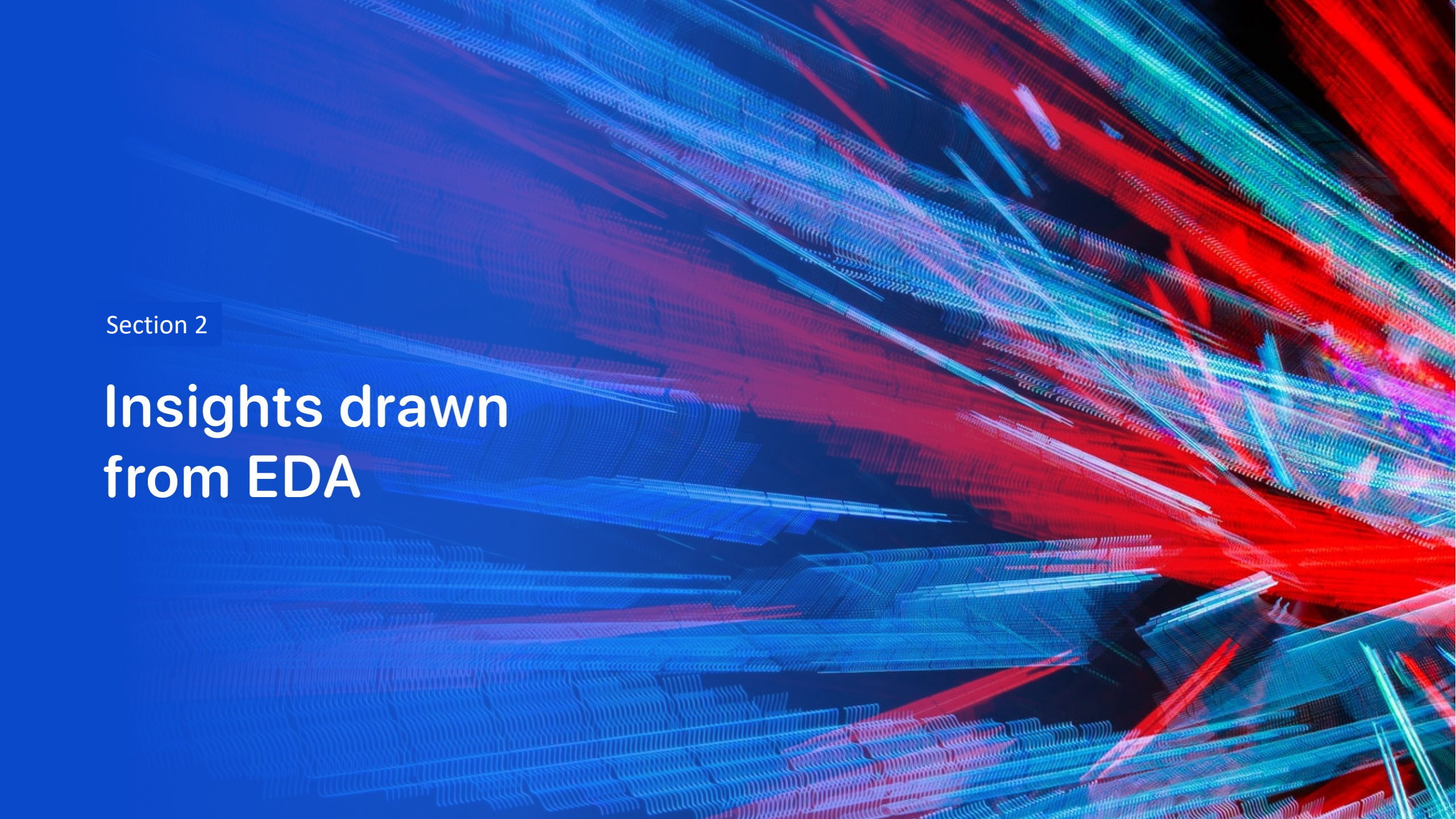
1. Built 4 different classifiers and evaluation of model performance on training data and test set.

1. Use the model for predictions.

Ref: https://portworx.com/blog/how-to-build-a-machine-learning-pipeline-using-kubeflow-and-portworx/

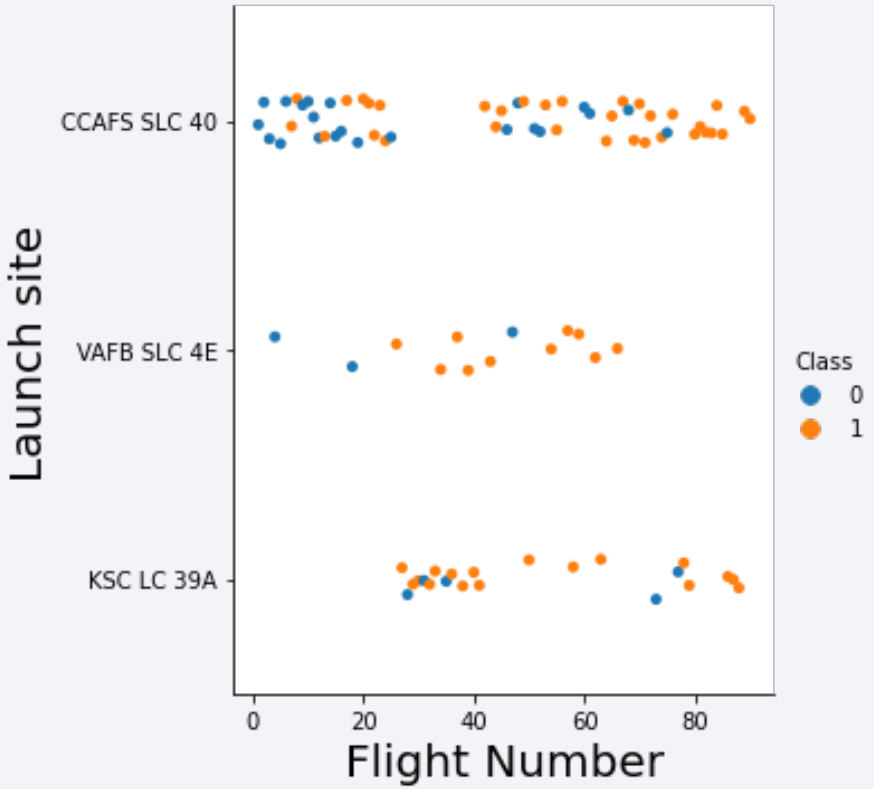
Results

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


Section 2
Insights drawn from EDA

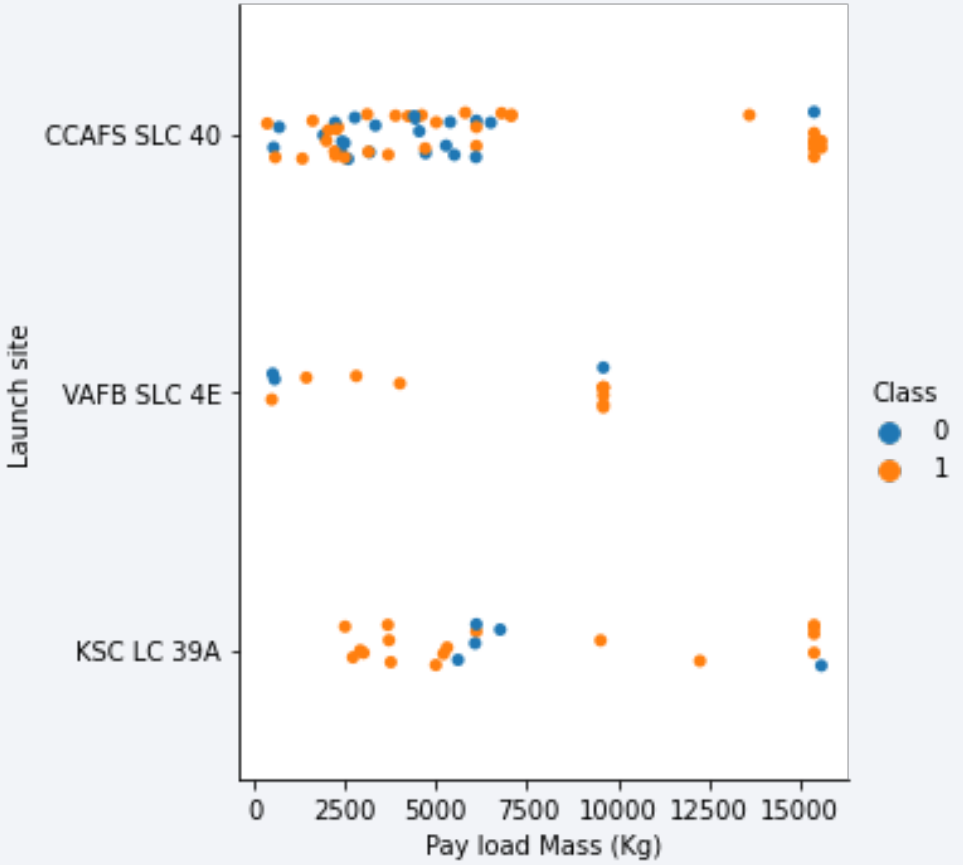
Flight Number vs. Launch Site

- For site CCAFS SLC 40, the success rate increased with number of flights.
- For site CCAFS SLC 40, at very low flight numbers success rate is very low.
- For other two launch sites VAFB SLC 4E and KSC LC 39-A success rate is very high compared to CCASFS SLC 40 launch site.
- There are a greater number of flights from CCAFS SLC 40 as compared to other two launch sites.



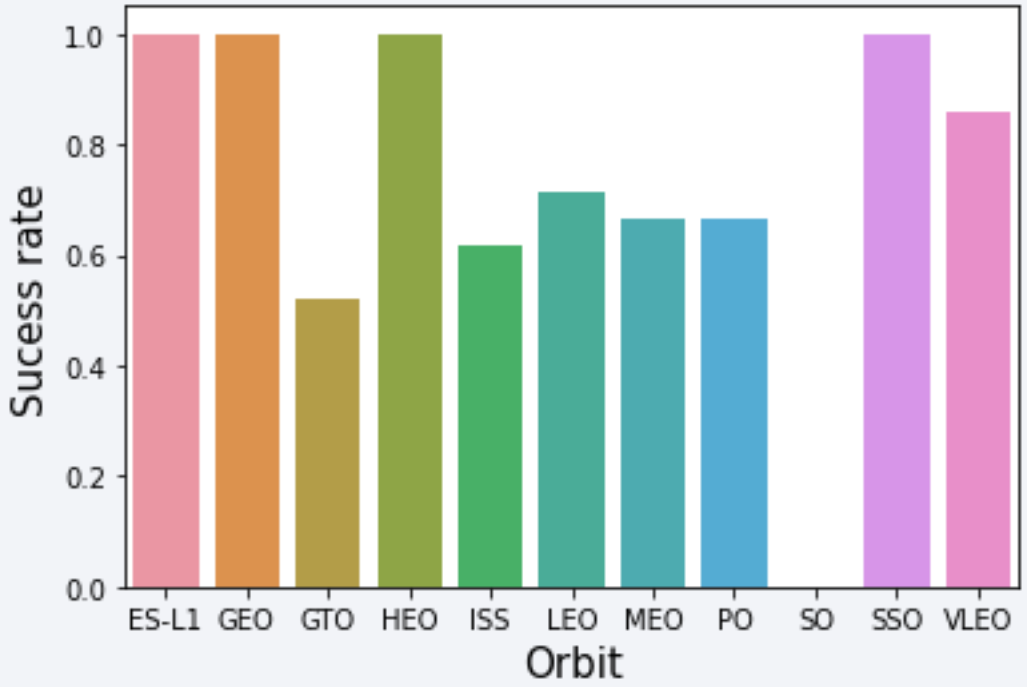
Payload vs. Launch Site

- For site CCAFS SLC 40, the success rate for very high payload (above 10k kg) is very high.
- For site VAFB SLC 4E , success rate is high for all ranges for payload but there are no very heavy rockets (> 10k kg).
- For launch site KSC LC 39A all rockers below 5k load are a success and between around 6k and 7.5k load there are mostly failures.



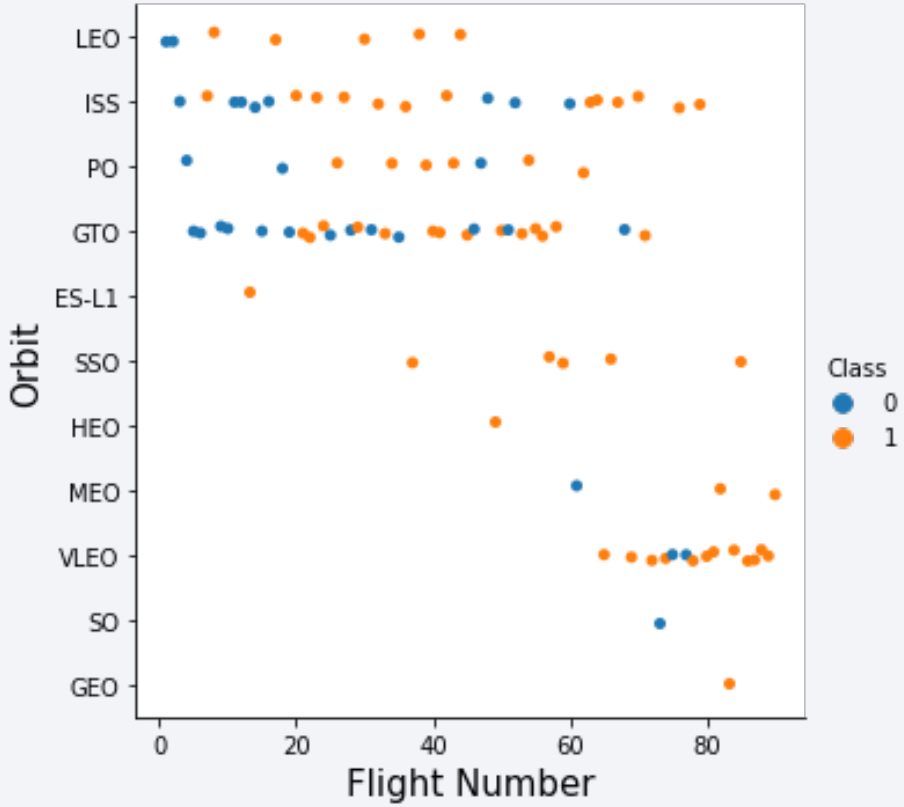
Success Rate vs. Orbit Type

- For orbit ES-L1, GEO, HEO, and SSO success rate is 1.
- Success rate for SO orbit is zero.
- For all orbits except SO success rate is above 50 %.



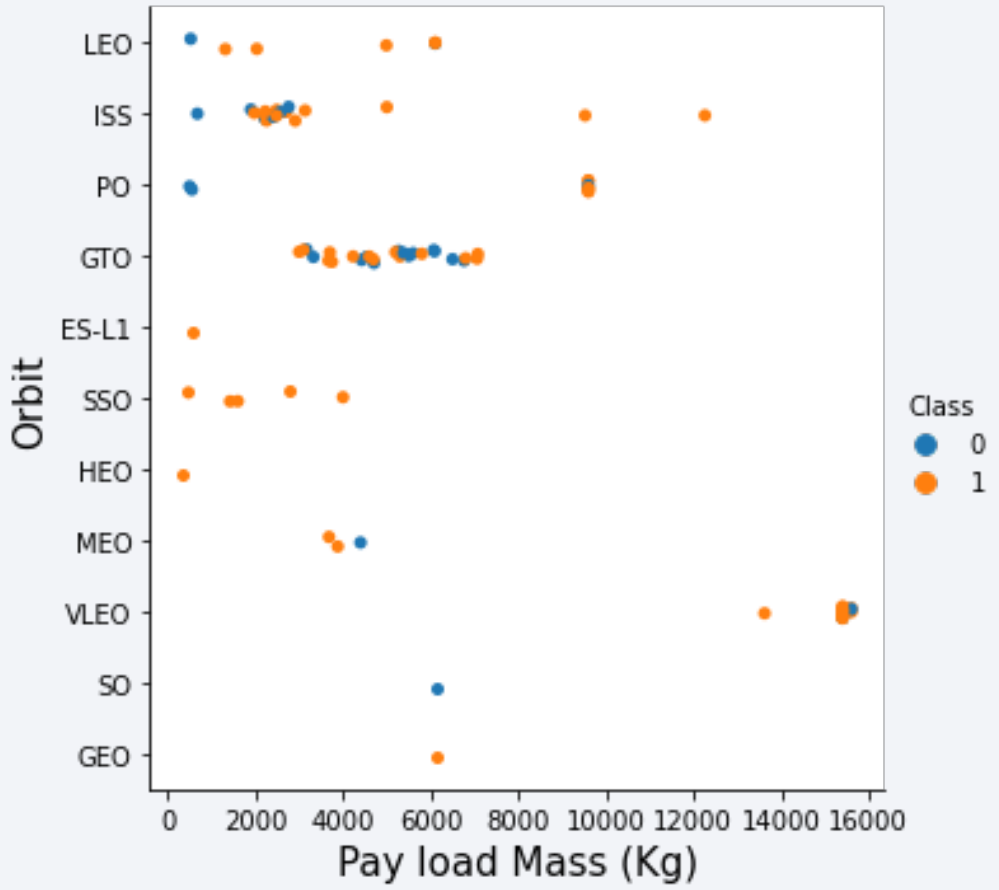
Flight Number vs. Orbit Type

- For LEO, ISS, GTO, and PO success of launch increases with flight number.
- For orbits GEO, SO, VLEO, MEO, HEO and SSO there are no record for the low flight numbers meaning there were launches to those orbit initially.
- For SSO orbit all launches are success.



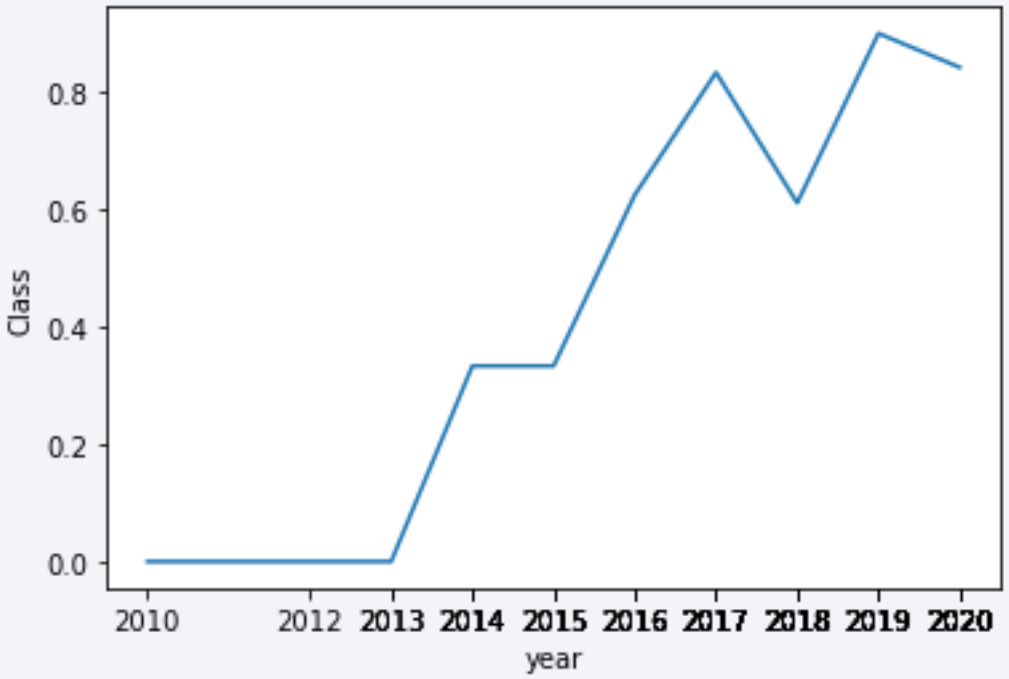
Payload vs. Orbit Type

- For LEO, ISS, and PO orbits success rate of launch increases with increase in payload mass.
- For GTO, there is no relation of success and failure of mission given the payload mass.
- Very heavy payload rockets are launched to the VLEO orbit and resulted in very high success rate.
- For SSO, HEO, MEO, SO, and GEO orbits pay load mass is less than 6k kg.



Launch Success Yearly Trend

- Success rate increased significantly onwards 2013.
- For the first time after 2013, success rate dipped compared to previous year in 2018.
- In recent year, success rate is nearly 90 %.



All Launch Site Names

```
select unique(launch_site) from SPACXTBL;
```

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

- We query the unique value using launch_site field from SPACXTBL.
- There are 4 unique values in the launch_site field.

Launch Site Names Begin with 'CCA'

```
select * from SPACXTBL
where launch_site LIKE 'CCA%'
LIMIT 5;
```

TE	TIME__UTC_	BOOSTER_VERSION	LAUNCH_SITE	PAY
-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Drago
-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Drago
-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Drago
-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	Space
-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	Space

F9 v1.0 B0007

- We filtered the query on launch_site field using LIKE operator “CCA%”.
- In next step, we limited our results using LIMIT keyword to first 5 results.

Total Payload Mass

```
select customer, sum(payload_mass__kg_) as sum_payload from spacxtbl
where customer = 'NASA (CRS)'
GROUP BY customer;
```

CUSTOMER	SUM_PAYLOAD
NASA (CRS)	45596

- We first group by our results using customer field and filtered the results where customer is “NASA (CRS)” .
- We then used the aggregation function sum() to calculate the total payload for NASA.

Average Payload Mass by F9 v1.1

```
select booster_version, AVG(payload_mass__kg_) as avg_payload_mass
from spacxtbl
where booster_version = 'F9 v1.1'
GROUP BY booster_version;
```

BOOSTER_VERSION	AVG_PAYLOAD_MASS
F9 v1.1	2928

- We first group by our results using booster_version field and filtered the results where booster_version is “F9 v1.1” .
- We then used the aggregation function AVG() to calculate the average payload for F9 v1.1 booster.

First Successful Ground Landing Date

```
select min(DATE) as first_sgl from spacxtbl
where landing__outcome = 'Success (ground pad)';
```

FIRST_SGL
2015-12-22

- We calculated the minimum of the date using min function on DATE field and results were filtered based on the condition of ‘Success (ground pad)’ landing.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
SELECT BOOSTER_VERSION
  FROM SPACXTBL
  WHERE LANDING__OUTCOME = 'Success (drone ship)' AND (
PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000 );
```

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

- For this query we filtered the results on two conditions using where clause and AND operator
    1. where landing outcome is ‘Success (drone ship)’
    2. Payload mass is in range 4000 and 6000 kg.

Total Number of Successful and Failure Mission Outcomes

```
SELECT MISSION_OUTCOME, COUNT(*)
  FROM SPACXTBL
  GROUP BY MISSION_OUTCOME;
```

MISSION_OUTCOME	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- We group by our results using MISSION_OUTCOME field and then used the COUNT function to calculate the total count for each mission outcome.

Boosters Carried Maximum Payload

```
SELECT BOOSTER_VERSION
  FROM SPACXTBL
  WHERE PAYLOAD_MASS__KG_ = (
  SELECT
  MAX(PAYLOAD_MASS__KG_)
  FROM SPACXTBL);
```

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

- To find the booster version where payload is maximum, first we used a subquery to find the max payload from PAYLOAD_MASS__KG_ field and used the result of this query to filter out the booster version having maximum payload.

2015 Launch Records

```
SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE
  FROM SPACXTBL
  WHERE YEAR(DATE) = 2015 AND LANDING__OUTCOME = 'Failure (drone
ship)';
```

LANDING__OUTCOME	BOOSTER_VERSION	LAUNCH_SITE
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- To find the booster version and launch site for the failed landing outcome in drone ship in the year of 2015. We used multiple condition with AND operator Where year = 2015 and landing outcome is failure on drone ship.

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

```
SELECT LANDING__OUTCOME, COUNT(*)
  FROM SPACXTBL
  WHERE DATE BETWEEN '2010-06-04' AND
'2017-03-20'
  GROUP BY LANDING__OUTCOME
  ORDER BY COUNT(*) DESC;
```

- 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

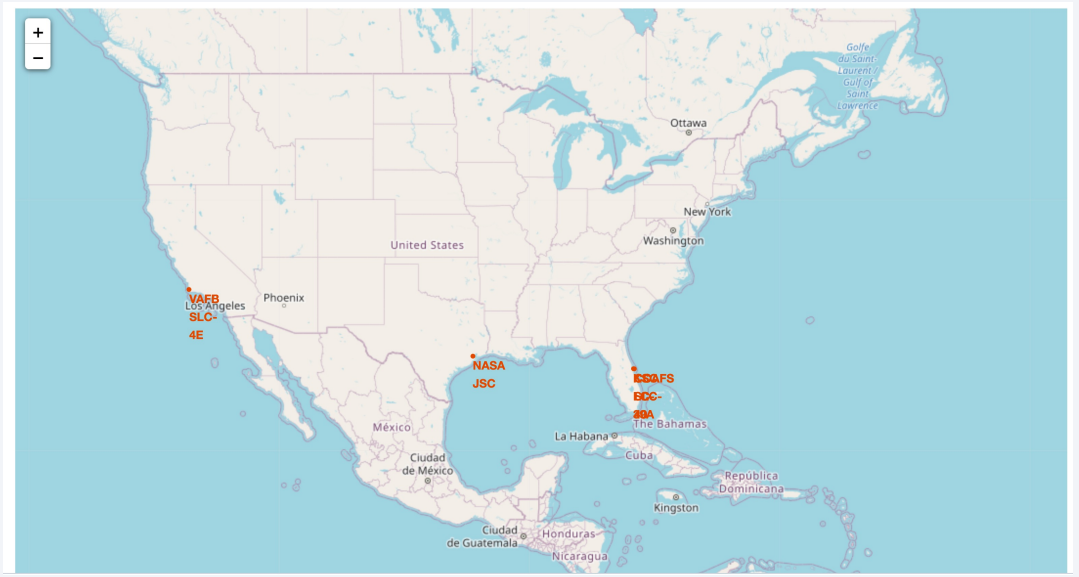
- First, we group by our data by landing outcome and then using filtering for given date range we count the total occurrence for each landing outcome and finally display our finding by descending order.

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

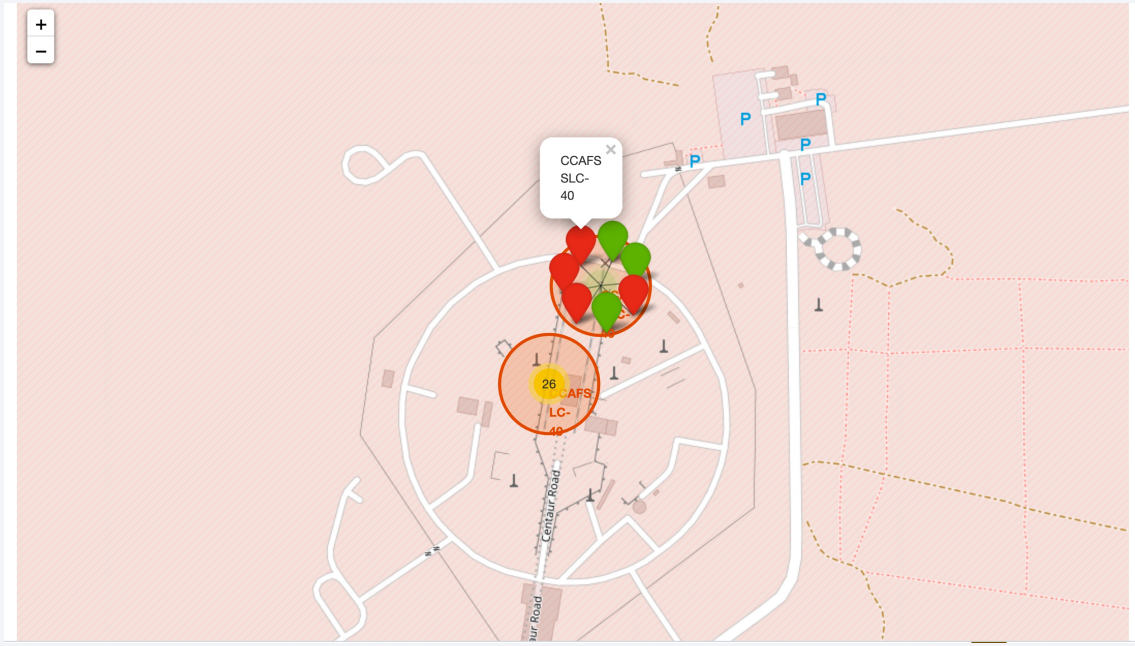
Location Markers Launch sites

- 4 location marked as launch sites with red labels on world map.
- Here, we used *marker* and *circle* to show the locations on the map.
- All launch sites are near to the coastal lines.
- We also marked the NASA JSC coordinates for the reference purpose.



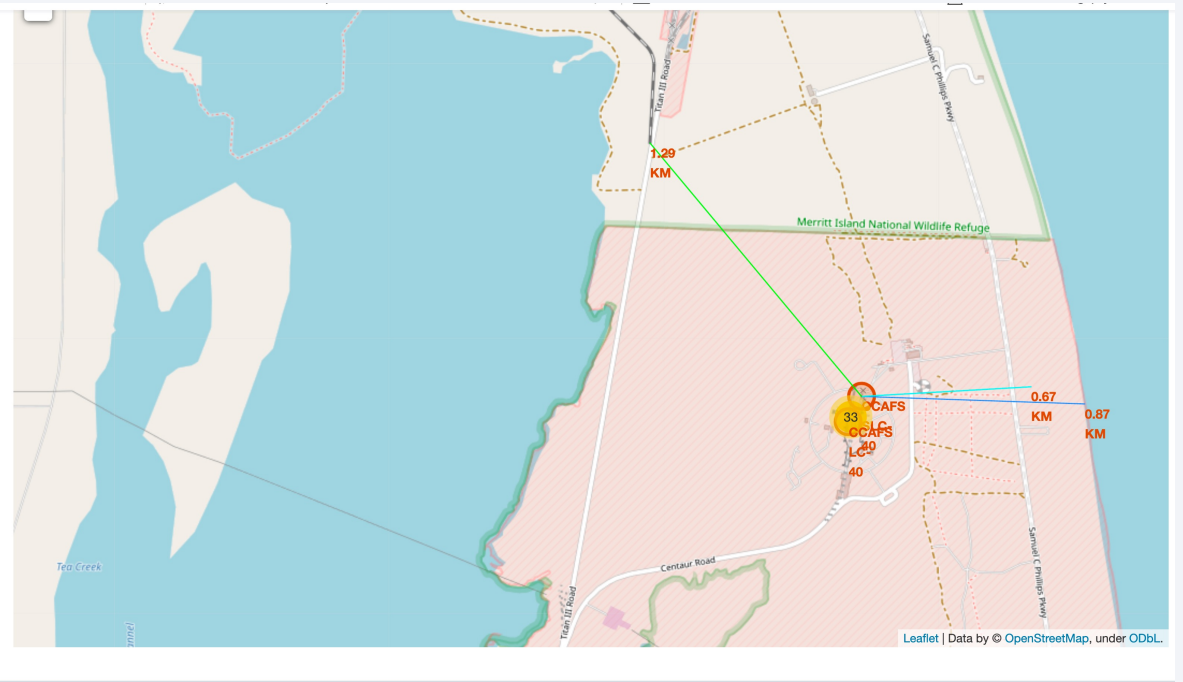
Launch outcomes with Marker clusters

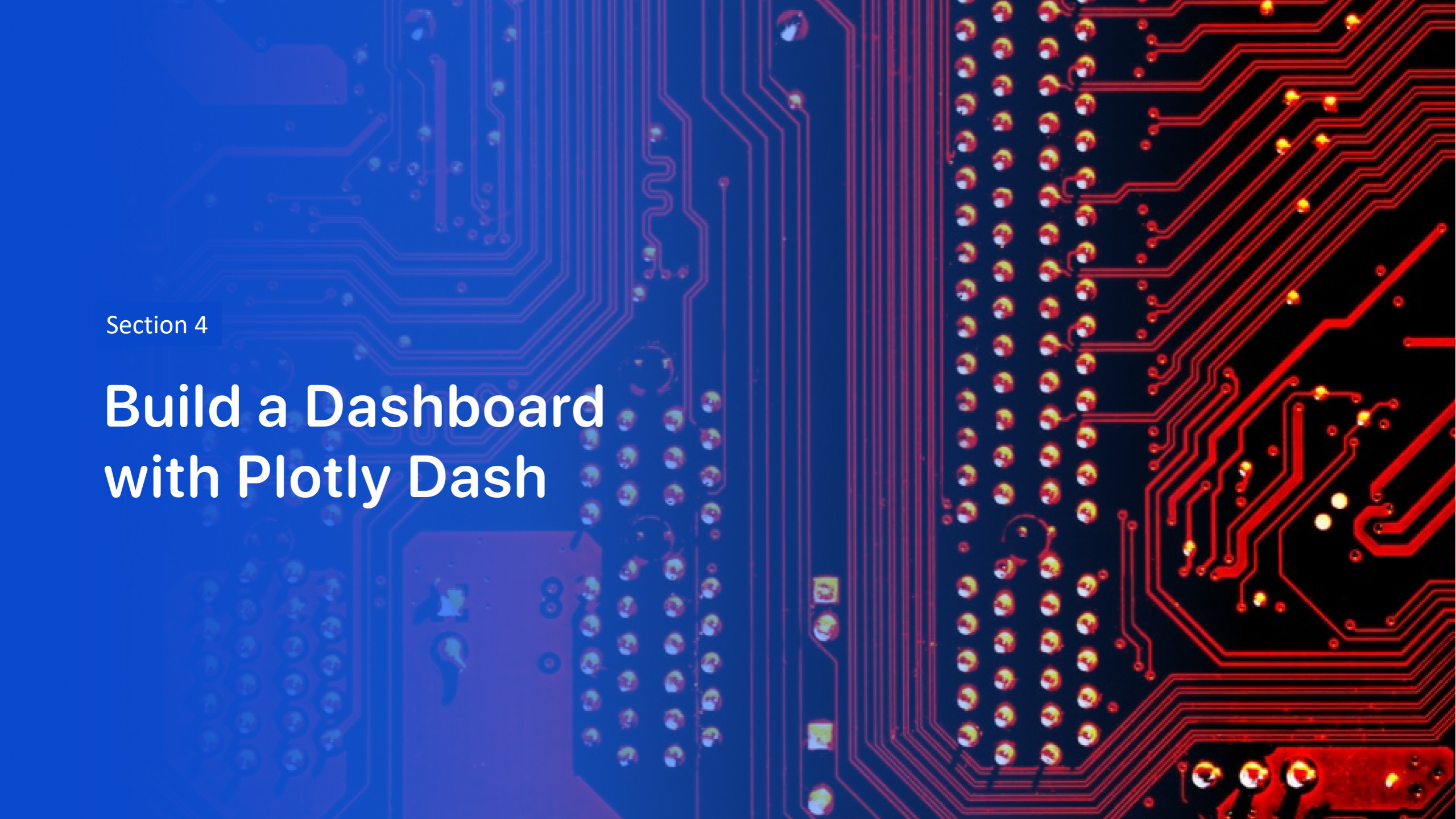
- Using Marker clusters in folium, we have shown launch outcome using two different color for each location site.
- For CCAFS SLC-40, we have 7 launches and out of those 3 (green) are success and 4 (red are failure).
- Marker cluster are great way to show lot of markers at single location using clustering.



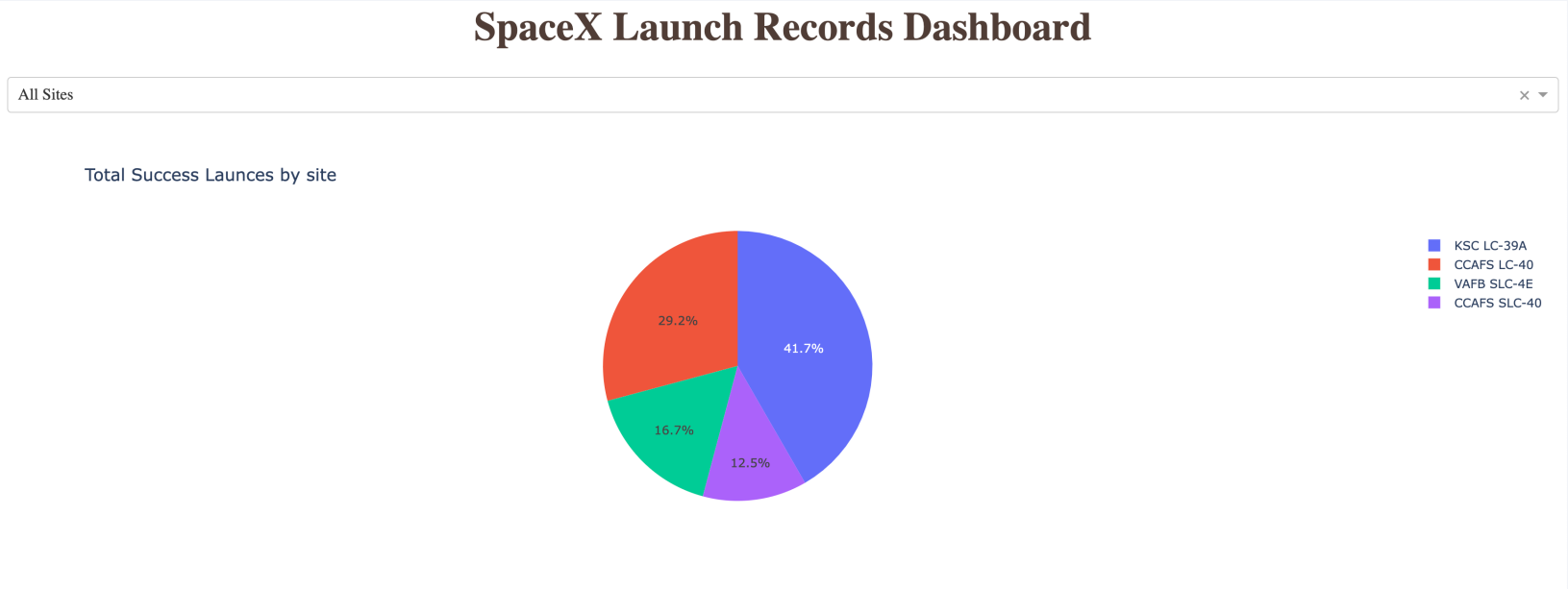
Launch site and proximity

- Using Polyline feature of folium , we have shown close proximities to CCAFS SLC-40 launch site with labels showing the distances.
- Nearest coastline (blue line) is approx. 0.87 km, highway (green line) is approx. 0.67 km and nearest railway track is located at 1.29 km from CCAFS SLC-40 launch site.
- Similarly, we can also show the proximities to other locations.



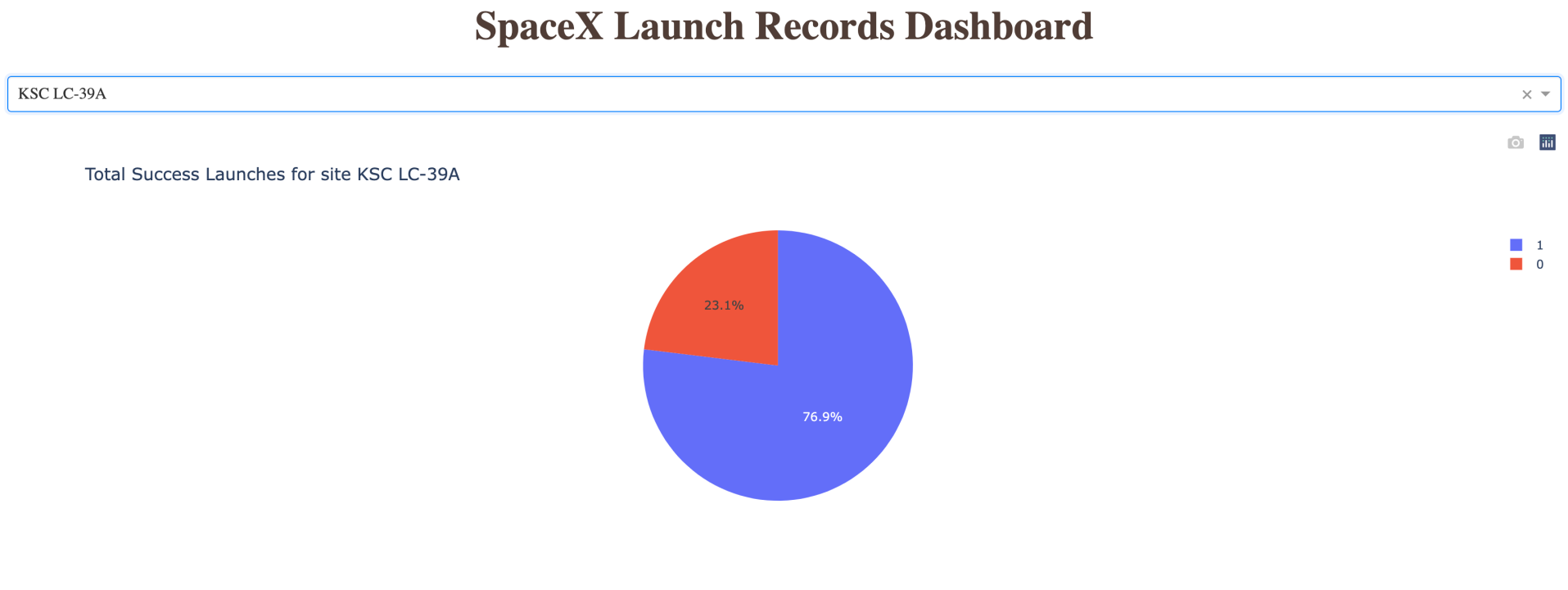

Section 4
Build a Dashboard with Plotly Dash

SpaceX Launch Records Dashboard (All sites)



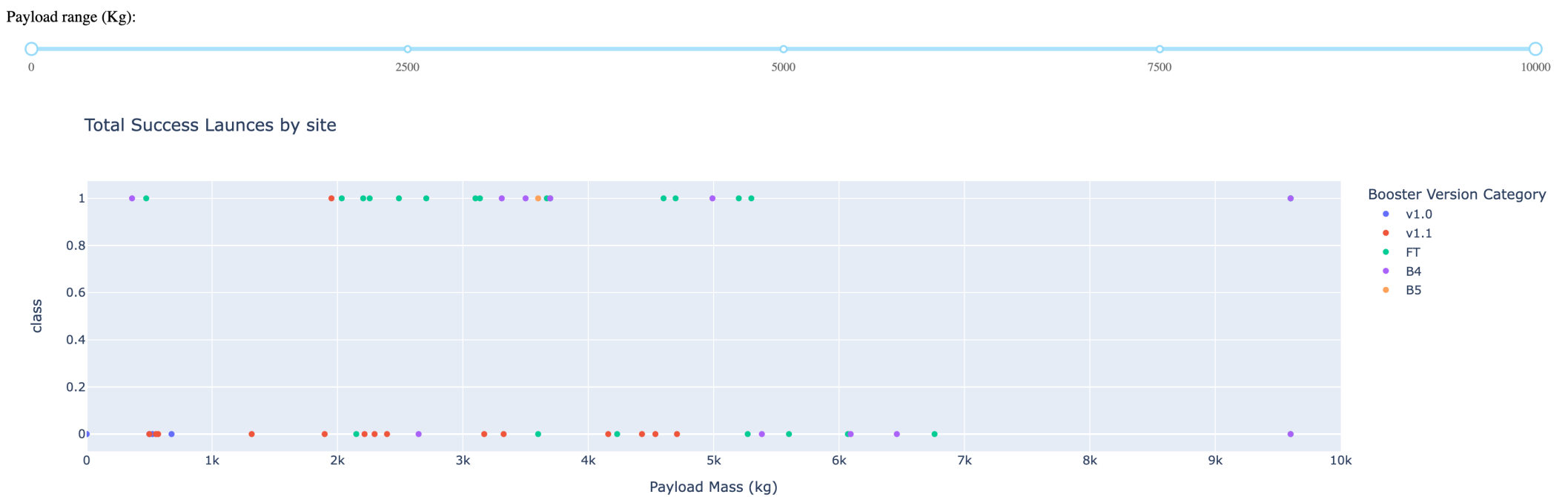
- KSC LC-39A has the highest success rate of 41.7 % among all launch sites.
- CCAFS SLC-40 has the lowest success rate of 12.5 % among all launch sites.

SpaceX Launch Records Dashboard (KSC LC-39A)



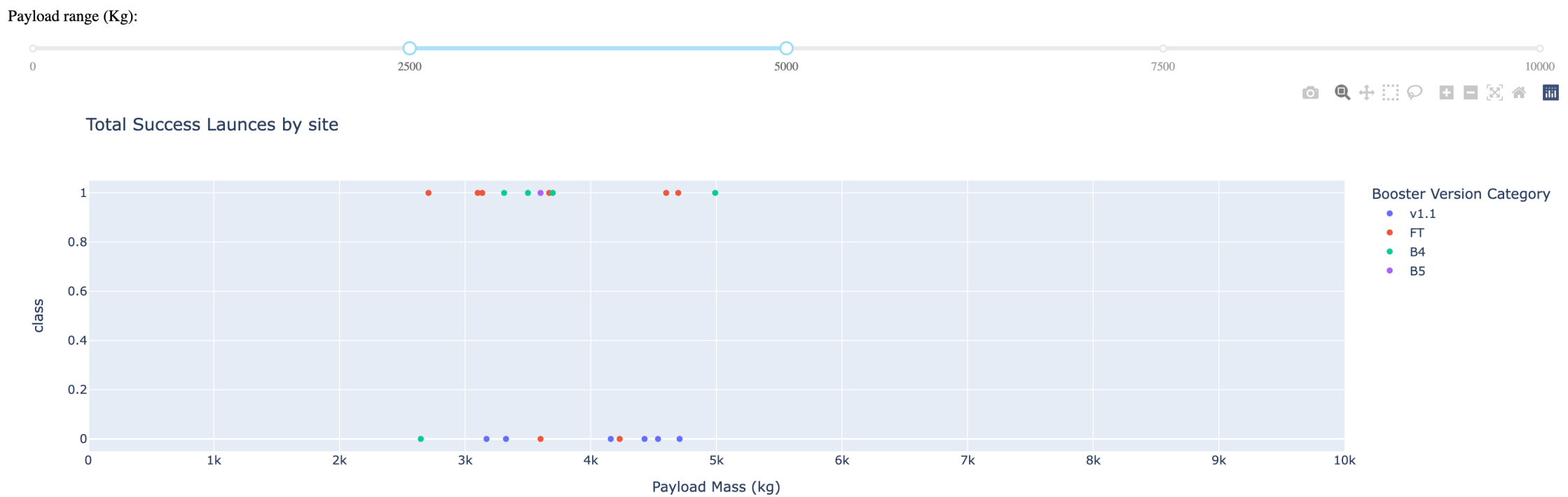
- For KSC LC-39A launch site 76.9 % launches were successful.

Payload vs Outcome for all site



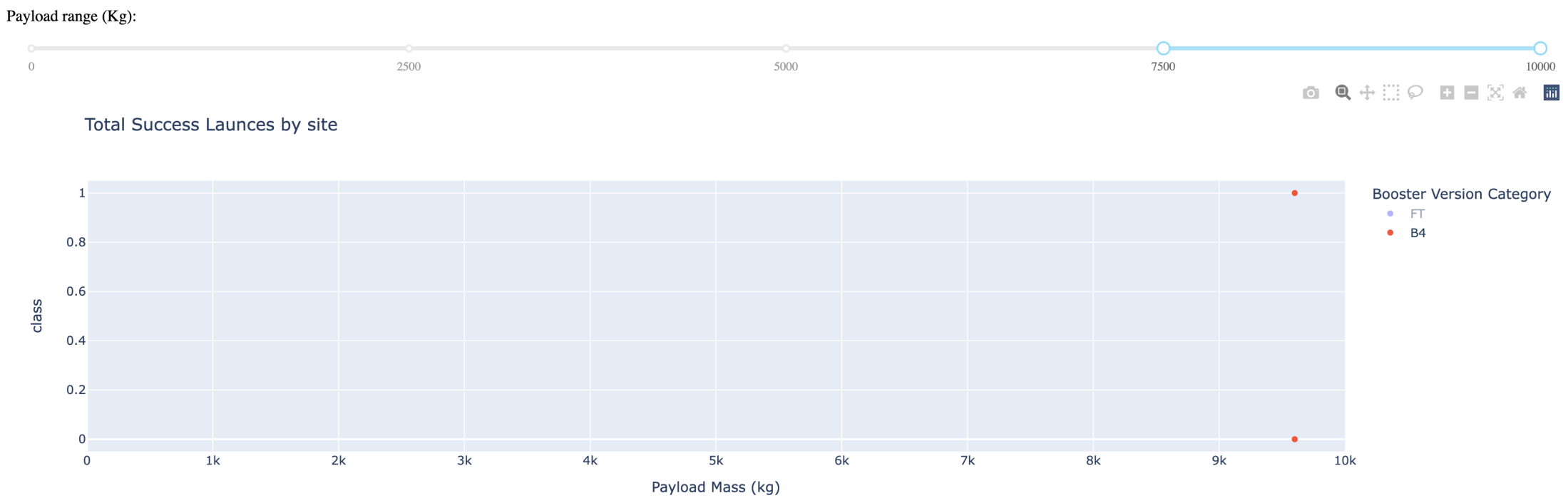
- Booster version v1.1 has very low success rate for all possible range of payload mass.
- There is no success launch for booster version v1.0.
- Booster FT and v1.1 has wide range of payload mass, but booster FT has good success rate as compared to v1.1.

Payload vs Outcome for all site



- For medium payload (2.5k to 5k range) B4 booster have high success rate.
- For medium payload (2.5k to 5k range) there is no success launch for booster v1.1.

Payload vs Outcome for all site

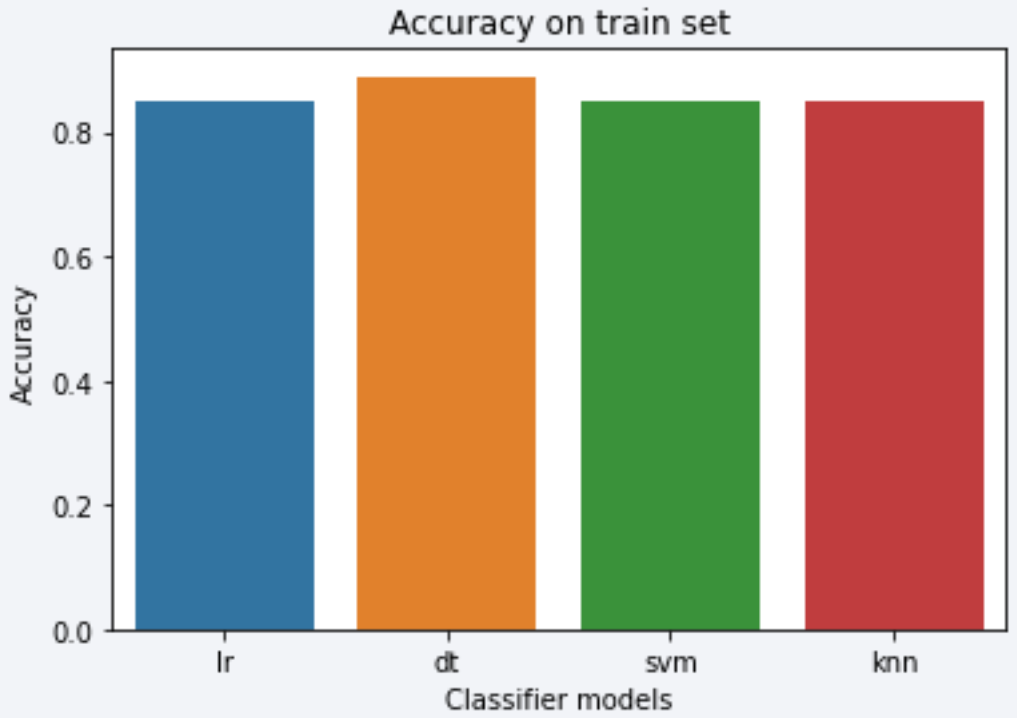


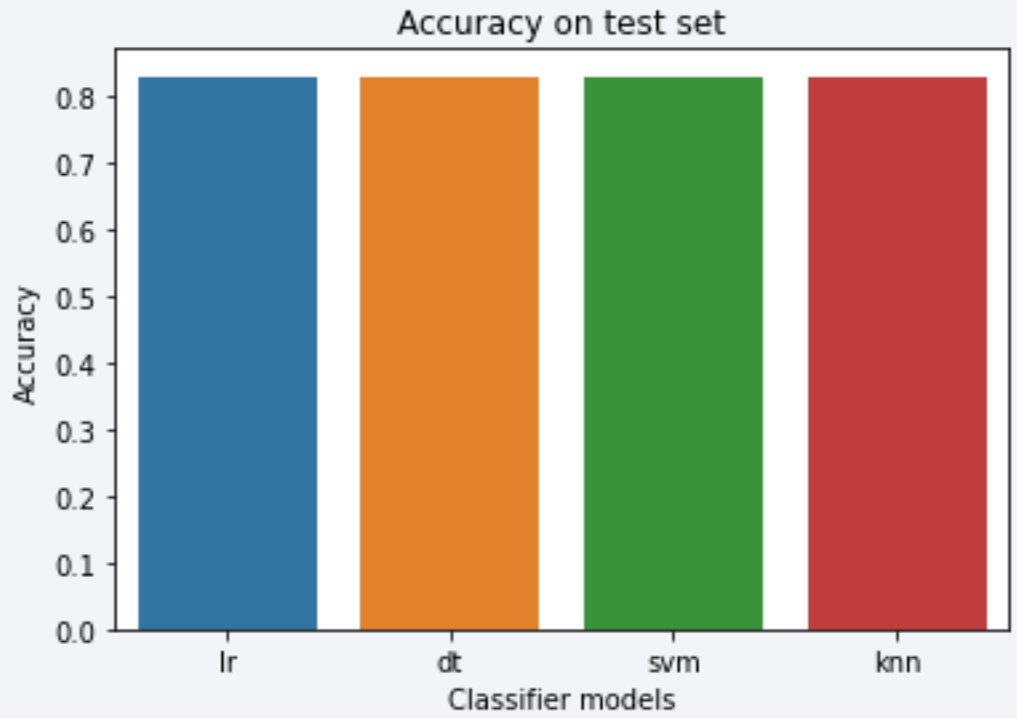
- For high payload mass (above 9k kg) there are only two booster version FT and B4 both with 50 % success rate..

Section 5

Predictive Analysis (Classification)

Classification Accuracy





- Practically all classifiers results in the same accuracy for test set (83.3 %).
- There is slight variation in accuracy for the train set for different classifiers. While all classifiers performs same on train set (85 % accuracy). Decision tree has an accuracy of 89 % on the train set.

Confusion Matrix

- All classifiers show the similar accuracy and similar confusion matrix. Here, we show the confusion matrix for the Logistic regression model.
- Here 3 prediction which did not land successfully are shown landed successfully by the model. (upper right quadrant of cf matrix.)

Conclusions

- we build four classifiers to predict the launch outcome for a SpaceX launch.
- All the four classifiers based on LR, SVM, DT, and KNN performed similar on test set
- With an accuracy of around 83.3 %.
- Model predicted few false positive case, meaning showing success landing where it failed.
- We can any of the four models to predict the launch outcome for the future landing with ~ 83 % accuracy.
- We can also build a model based on Random Forest algorithm to see if there is any further improvements.

Appendix

- GitHub repo containing all code notebook and relevant links and dataset.

  https://github.com/shera-amit/Coursera-capstone-final.git

Thank you!