Winning Space Race with Data Science

KAMESH KHANDEKAR

2022-03-29

Outline

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

Executive Summary

- Summary of methodologies :
  - Collect data from public SpaceX API and its Wikipedia page.
  - Created labels column “**class**”, which classifies the successful landings.
  - Explored and visualize data using SQL, visualization, folium maps, and dashboards.
  - Gathered relevant columns to be used as features.
  - Used one hot encoding to change all categorical variables to binary. Standardized data and used GridSearchCV technique to find the best parameter for machine learning models.
  - Visualize accuracy score of all models.
- Summary of all results
  - Four machine learning models were produced:
    - Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors.
  - All produced similar results with accuracy rate of about 83.33%. All models over predicted successful landings. It requires more data to make model better and for better model determination and accuracy.

Introduction

- Project background and context
  - SpaceX(Space Exploration Technologies Corp.) is developing a satellite internet constellation named Starlink to provide commercial internet service.
  - Founded in 2002 by Elon Musk.
  - Reducing space transportation cost with pricing from $165 million USD to $62 million USD.
  - It has best pricing compared to other space providers.
  - The project is for SpaceY which wants to compare with SpaceX.
- Problems you want to find answers
  - SpaceY is working to train a machine learning model to predict successful recovery of stage 1 from rockets.

Section 1

Methodology

Methodology

Executive Summary

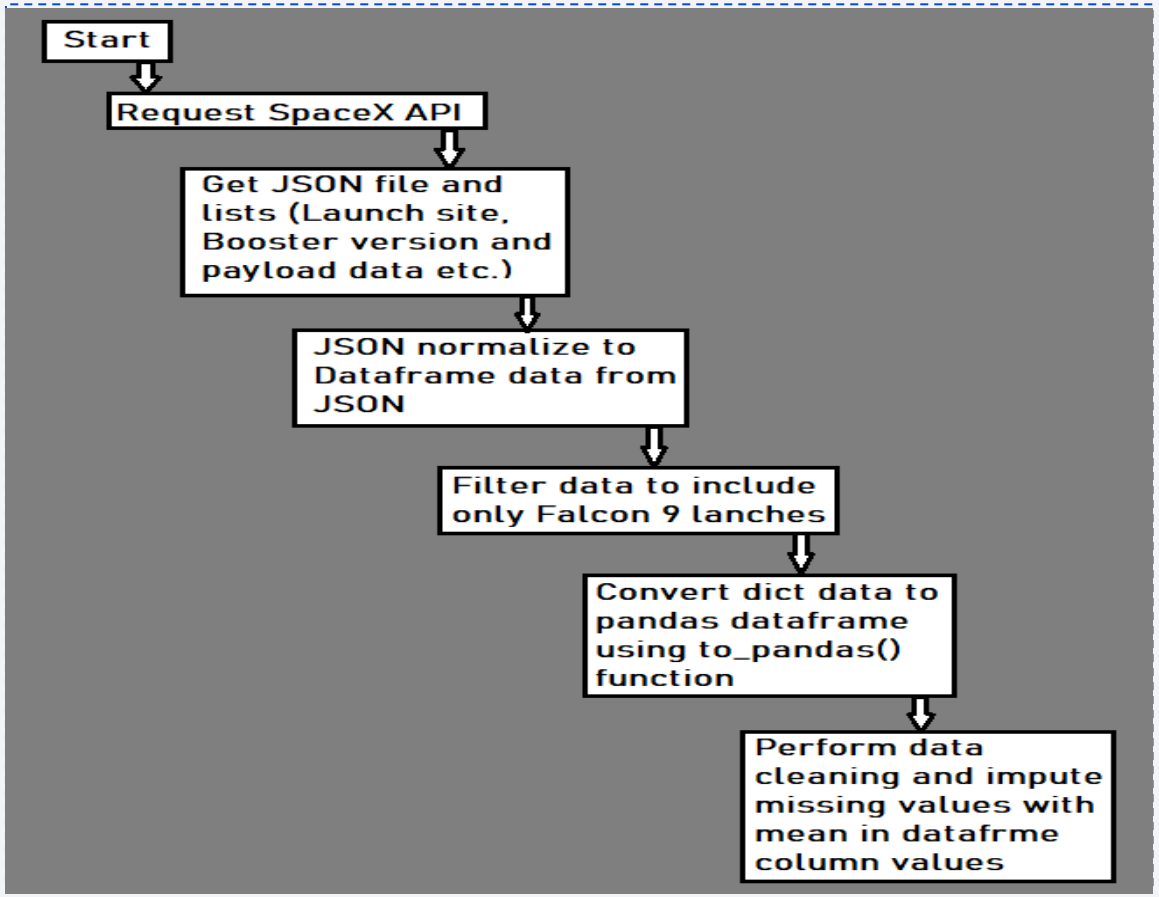
- Data collection methodology:
  - The combination of data from SpaceX API and the Wikipedia of SpaceX page.
- Perform data wrangling
  - Data is transformed and mapped from one raw data form into another form with intent of making it more appropriate and valuable for various tasks.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build : training data and Training labels, tune : GridSearchCV, evaluate classification models : cross-validation and feature extraction.

Data Collection

- Data collection process involved in a combination of SpaceX API and Wiki pages' data via web scrapping from tables in the wiki page.
- Then in next slide will show the flowchart of data collection from SpaceX API and after that it will show flowchart of data collection from web scrapping using Beautiful Soup.
- SpaceX Api columns after parsing the request data :
  - Serial, Longitude, Latitude, PayloadMass, Orbit, LaunchSite, Outcome, Flights, Legs, FlightNumber Date, BoosterVersion, LandingPad, Block, ResusedCount, GridFins, Resused
- SpaceX Api columns after parsing the request data :
  - Orbit, Customer, FlightNo, LaunchSite, Payload, BoosterLanding, Date, Time, PayloadMass, LaunchOutcome, VersionBooster

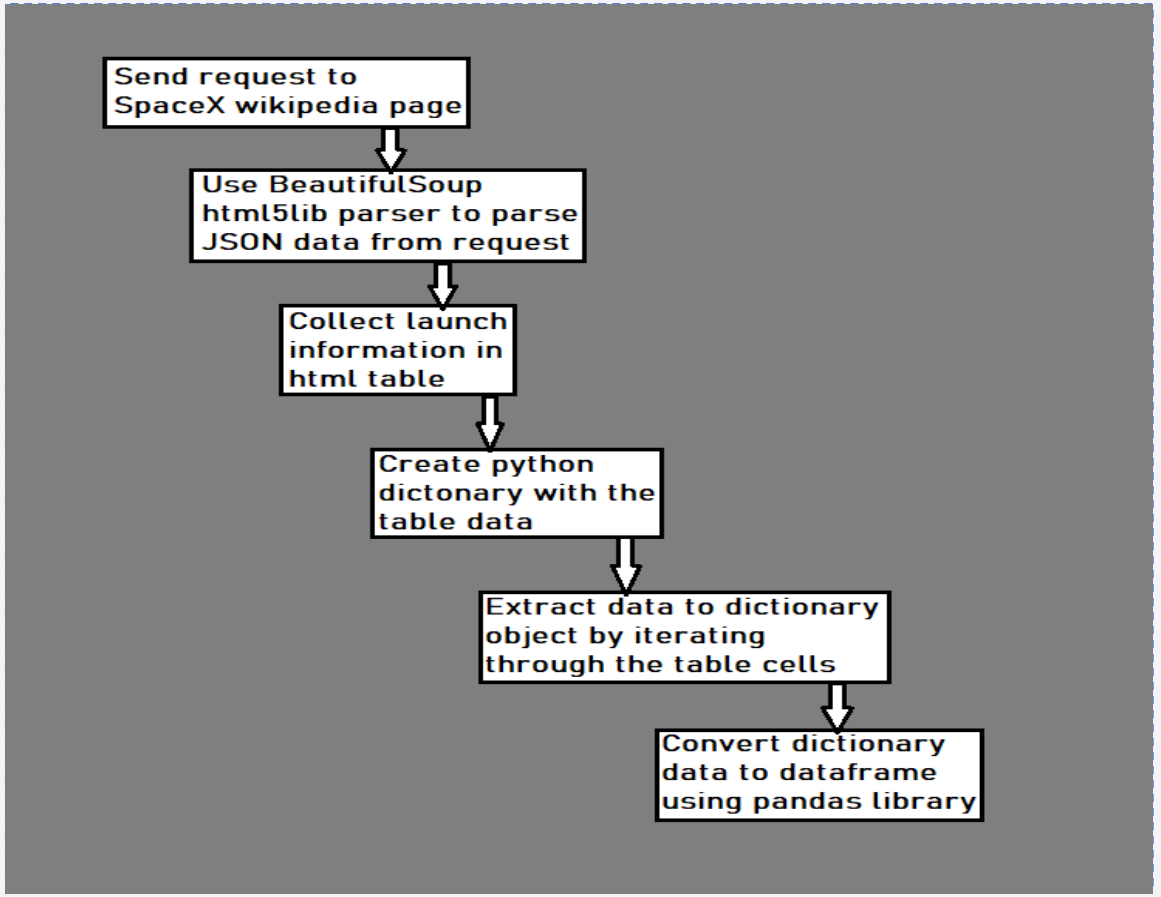
Data Collection – SpaceX API

- Presented data collection with SpaceX REST calls using key phrases and flowcharts :
  - Kindly, refer to figure shown on right side.
- GitHub URL of the completed SpaceX API calls notebook :
  - https://github.com/kamesh01/Applied-Data-Science---Capstone/blob/main/Week%201%20Introducation/Capstone_Data_Collection_API.ipynb



Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts :
  - Kindly, refer to figure shown on right side.
- GitHub URL of the completed web scraping notebook :
  - https://github.com/kamesh01/Applied-Data-Science---Capstone/blob/main/Week%201%20Introducation/Capstone_Data_collection_with_WebScrapping.ipynb



Data Wrangling

- First create training label with landing outcomes where successful =1 & failure = 0. Outcome column has two components Landing location and missing outcomes.
- Value mapping: new training label's column "Class" with a value of 1 if missing outcome is True else 0 for false value.
- Ex. – There are two situations :
  - True ASDS, True RTLS & True Ocean set to 1
  - False ASDS, None None, False RTLS set to 0
- GitHub URL of your completed data wrangling related notebooks :
  - https://github.com/kamesh01/Applied-Data-Science---Capstone/blob/main/Week%201%20Introducation/Capstone_Data_wrangling_EDA_.ipynb

EDA with Data Visualization

- **Exploratory data analysis(EDA)** is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.
- Plots used and why :
  - Scatter plots, line plots and bar plots are used *to explore the relationship between different variables to decide if a relation exists, so that it could be used in training the machine learning models and get idea about data using visualization.*
- GitHub URL of your completed EDA with data visualization notebook :
  - https://github.com/kamesh01/Applied-Data-Science---Capstone/blob/main/Week%202%20EDA/EDA_with_Data_Visualization.ipynb

EDA with SQL

- Summarizing the SQL queries are performed :
  - Created table and loaded data in DB2 database.
  - Used SQL python integration using notebook and SQL Alcamy.
  - Used where clauses to filter the records from dataset.
  - Retrieved launch site name, various payload sizes, booter version info and landing outcomes.
  - Used date range to get information(in some date range) of different parameters like outcomes, launch site name and others.
- GitHub URL of your completed EDA with SQL notebook :
  - https://github.com/kamesh01/Applied-Data-Science---Capstone/blob/main/Week%202%20EDA/EDA_with_SQL.ipynb

Build an Interactive Map with Folium

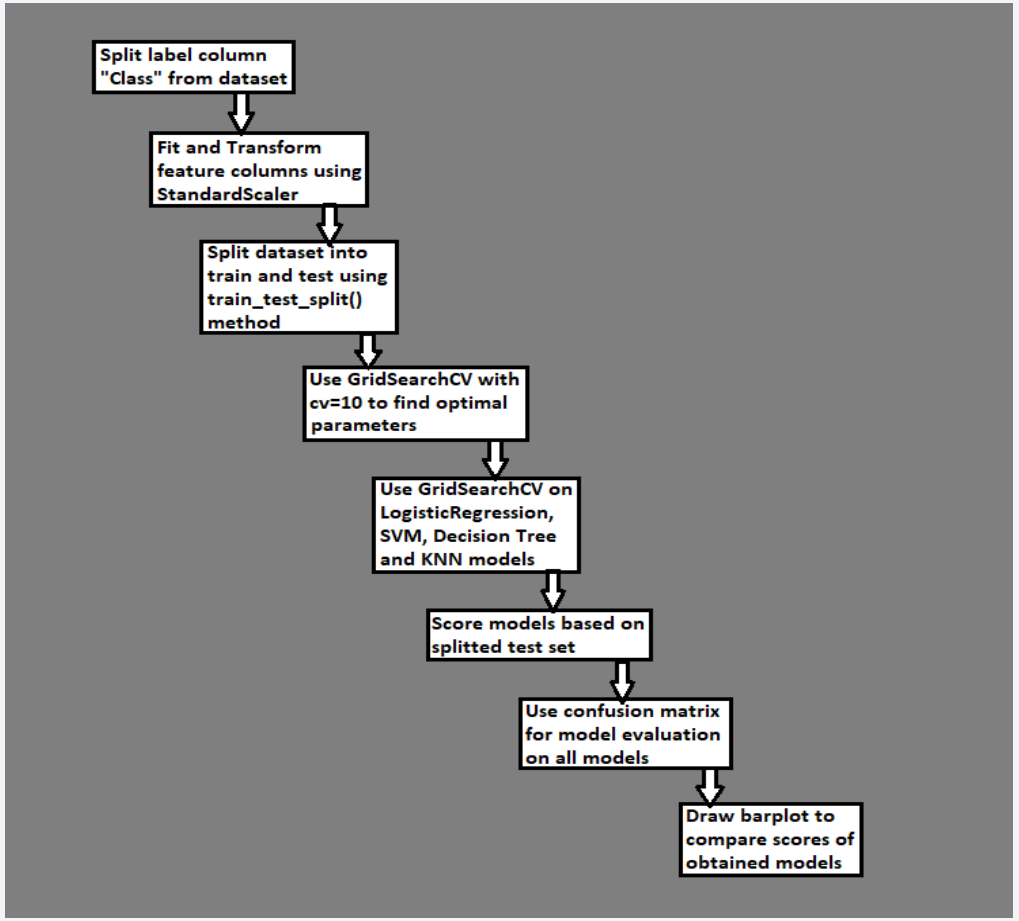
- Summary of what map objects such as markers, circles, lines, etc. that's been created and added to a folium map :
  - In folium maps, makers are used for Launch site, successful and unsuccessful landings location with proximity key locations can be any of Railway, Highway, Coast and city or near to it.
- These markers provides ease and allows us to understand better to which location is best fit to launch sites. Visualization helps in finding successful landing to its corresponding locations.
- GitHub URL of your completed interactive map with Folium map :
  - https://github.com/kamesh01/Applied-Data-Science---Capstone/blob/main/Week%203%20Interactive%20Visual%20Analytics%20and%20Dashboard/Interactive_Visual_Analytics_with_Folium_lab.ipynb

Build a Dashboard with Plotly Dash

- The dashboard includes a pie chart and a scatter plot.
- The Pie chart is useful to show distribution of successful landings across all launch sites and individual launch site success rates with distributed visual.
- The pie-chart is used to visualize launch site success rate.
- In Scatter plot it takes two input : All sites or individual site and payload_mg on a slider between 0 to 10000 kg.
- The scatter plot is used to see how success varies across launch sites, payload_mg, and booster version category.
- GitHub URL of your completed Plotly Dash lab :
  - https://github.com/kamesh01/Applied-Data-Science---Capstone/blob/main/Week%203%20Interactive%20Visual%20Analytics%20and%20Dashboard/spacex_dash_app.py

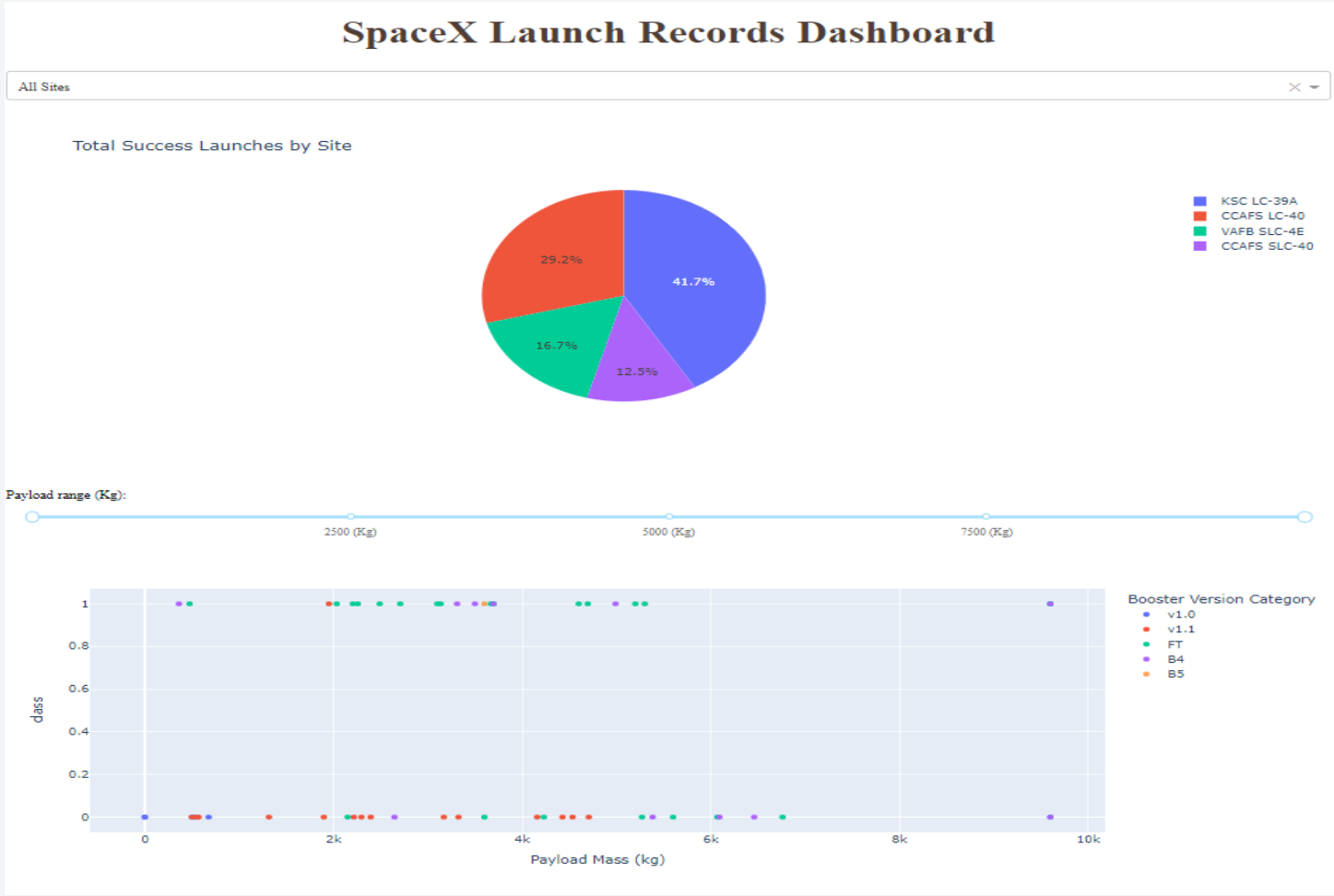
Predictive Analysis (Classification)

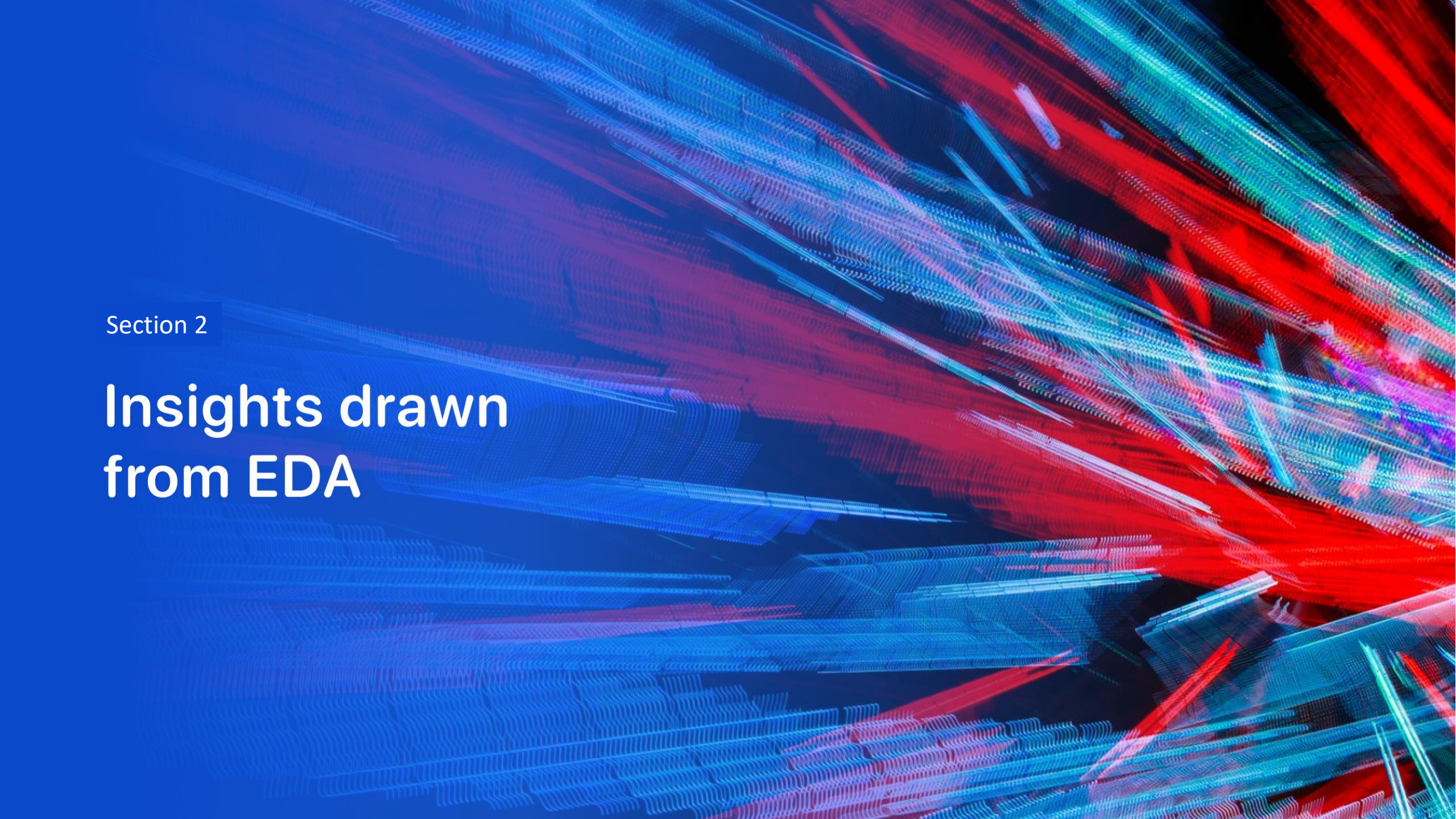
- Build model using different algorithms like SVM, KNN, Decision tree, Logistic regression.
- Use GridSearchCV with cv =10 parameter, to find the best hyperparameter to get the best parameter for model for highest accuracy.
- Use scoring method to score each built models based on the splitted test dataset.
- For model evaluation, use confusion matrix on each generated models.
- Draw barplot and visualize the score for better efficiency/accuracy of each model.
- For key phrases and flowchart refer figure to right side.
- GitHub URL of your completed predictive analysis lab :
  - https://github.com/kamesh01/Applied-Data-Science---Capstone/blob/main/Week%204%20Predicative%20Analysis%20-%20Classification/Machine_Learning_Prediction.ipynb



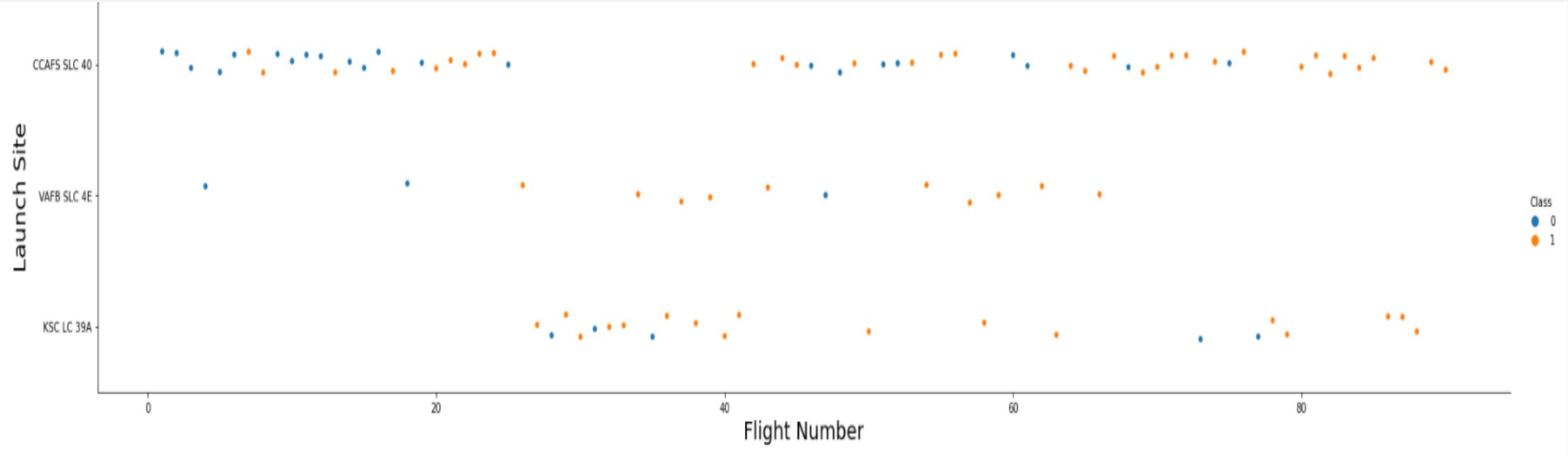
Results

- The figure shown on right side is the Plotly Dashboard with a Pie Chart and a Scatter Plot.
- In the coming slides, will see the results of EDA with visualization, EDA with SQL, Interactive Map with Folium and finally the results of our model with about 83% accuracy.




Section 2
Insights drawn from EDA

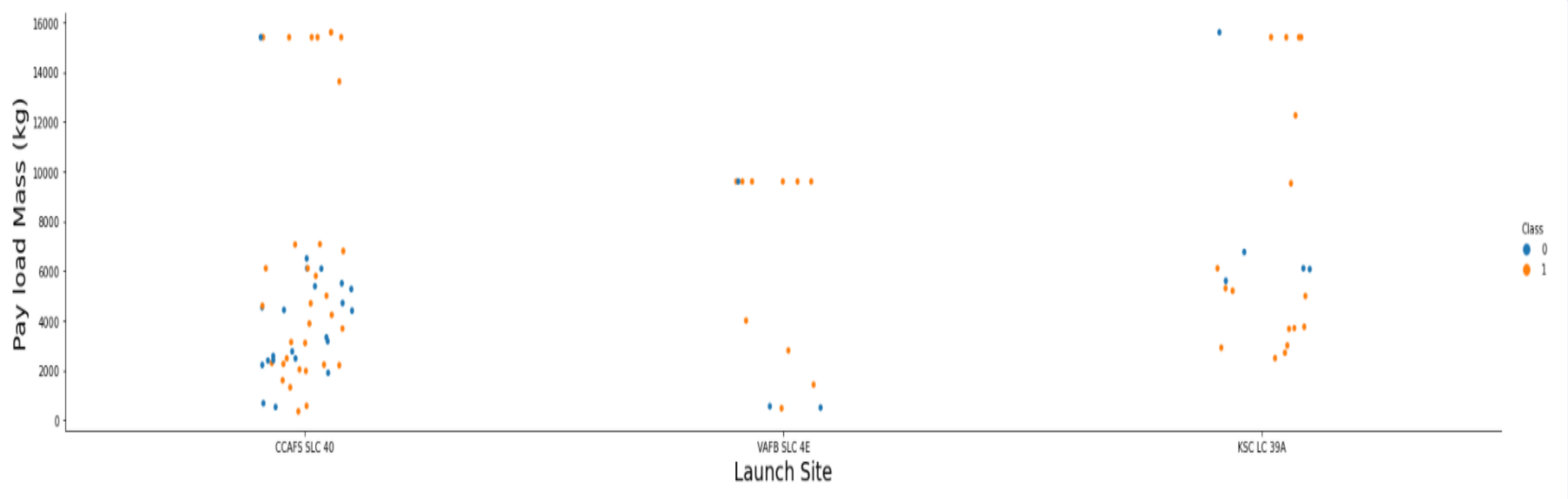
Flight Number vs. Launch Site



Orange color indicates : Successful launches
Blue color indicates : Unsuccessful launches

- The above graph suggests an increase in success over time.
- Likely a big breakdown around flight 20 which significantly increased success rate.
- CCFA appears to be the main launch site as it has the most volume in green color.

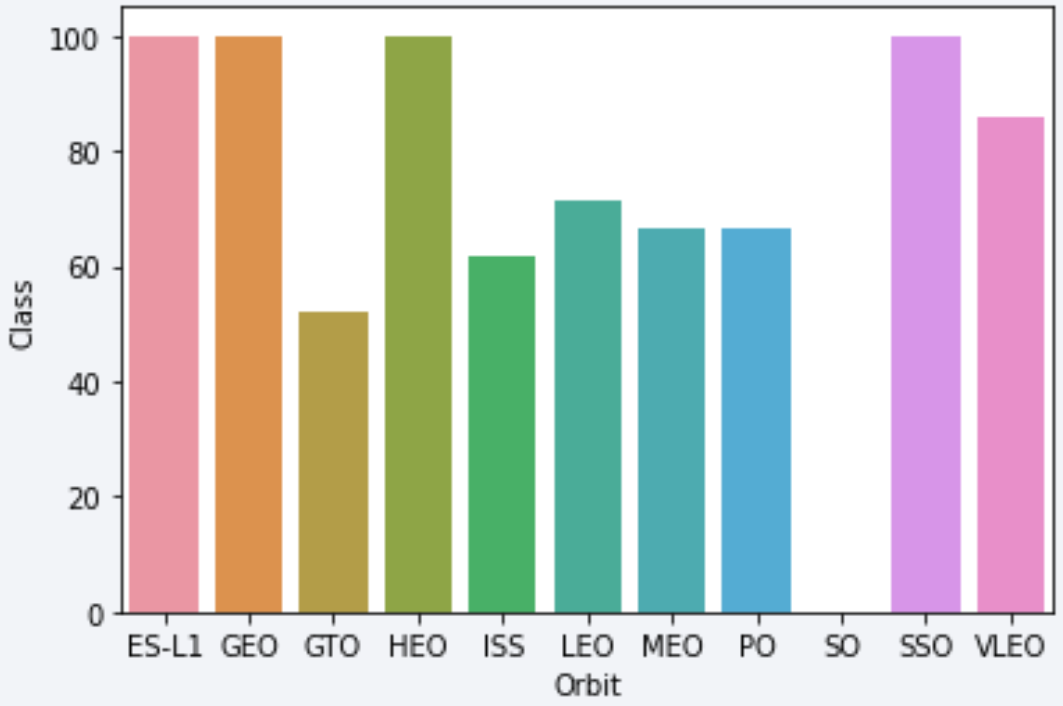
Payload vs. Launch Site



Orange color indicates : Successful launches
Blue color indicates : Unsuccessful launches

- By observing Payload Vs. Launch Site scatter point chart, will find for the VAFB-SLC LaunchSite there are no rockets launched for HeavyPayload_Mass(greater than 10000).

Success Rate vs. Orbit Type



Success rate scale with

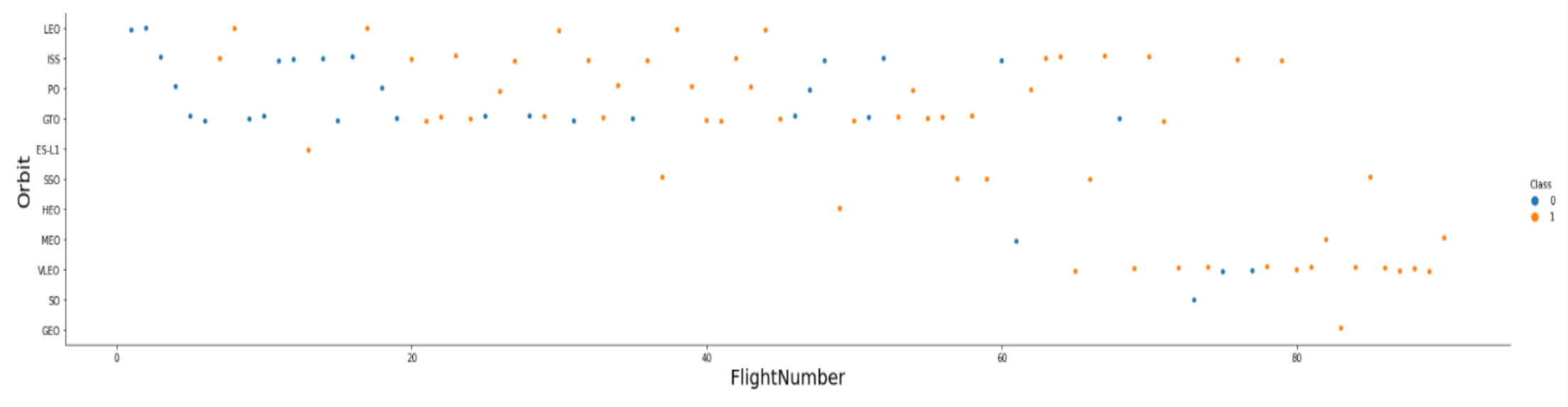
0%

60%

100%

Orbit types : ES-L1, GEO, HEO and SSO has 100% success rate, GTO has 50% success rate, LEO has 40% decent success rate, SO has 0(0% success rate)

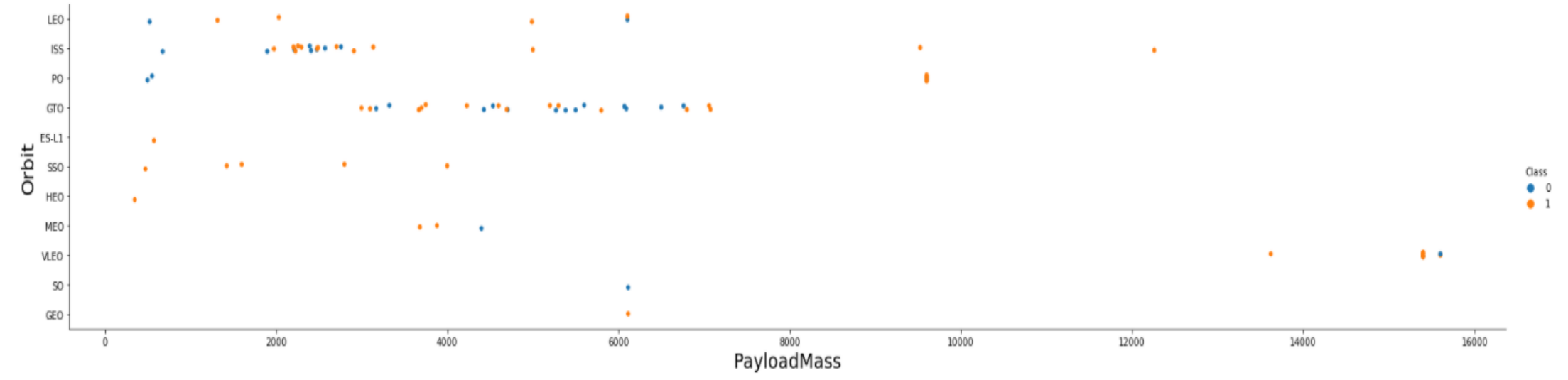
Flight Number vs. Orbit Type



Orange color indicates : Successful launches
Blue color indicates : Unsuccessful launches

- Launch orbit preferences changed over Flight Number.
- Launch outcomes seems to correlate with this preference.
- SpaceX started with LEO orbit which saw moderate success LEO.
- Hence, the graph shows that SpaceX performs better in lower orbits synchronous orbits.

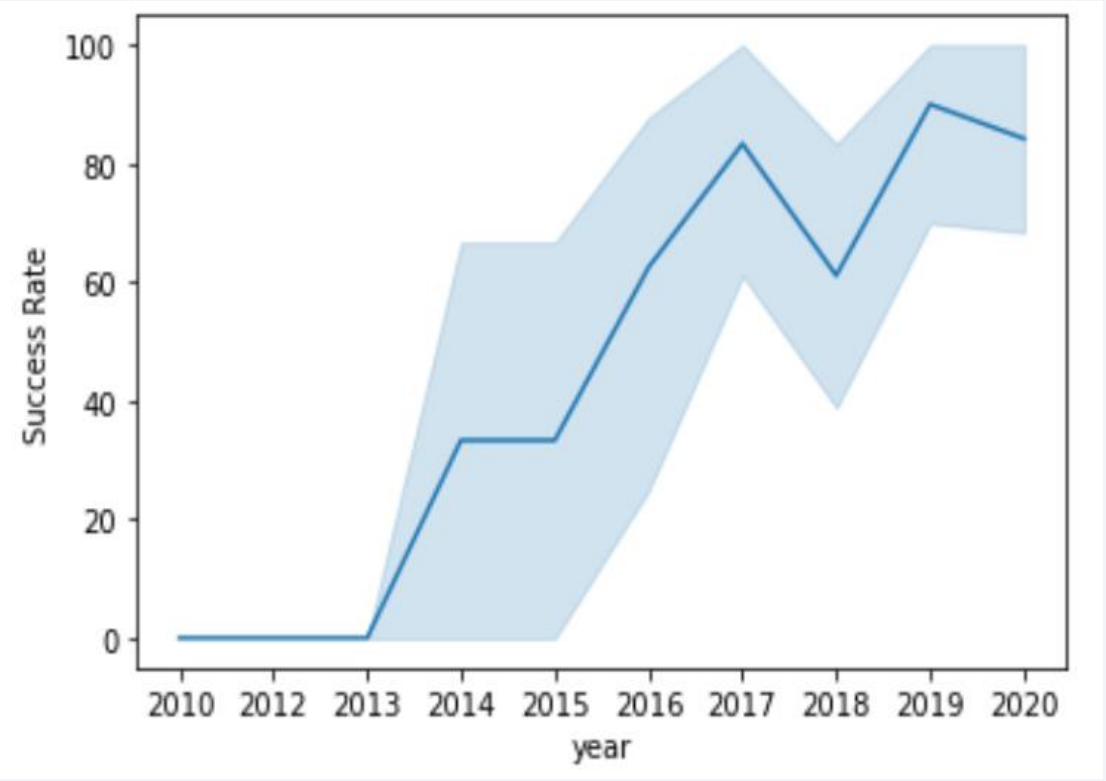
Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.
- Payload mass seems to correlate with orbit.
- SSO and LEO seems to have relatively low payload mass.

Launch Success Yearly Trend

- From the figure it can be observed that the success rate, since 2013 kept increasing till 2020.
- The success rate in year around 2019, it is highest that is 90%
- Hence, it can be concluded that success generally increases over time since 2013.



All Launch Site Names

Display the names of the unique launch sites in the space mission

```
%sql select DISTINCT LAUNCH_SITE from SPACEXDATA
```

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

- The shown query is giving unique launch site names from database.
- From looking at the figure, it is clear in understanding that there only 4 launch sites by SpaceX.
- But CCAFS LC-40 and CCAFS SLC-40 looks same, it might be possible that there might be entry error.
- So, if there is any entry error or typing mistake then, there are only 3 unique sites for launching SpaceX.

Launch Site Names Begin with 'CCA'

Display 5 records where launch sites begin with the string 'CCA'

```
%sql select * from SPACEXDATA where launch_site like 'CCA%' limit 5
```

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

- With the help of above query shown in figure, we can get only 5 records where launch sites begin with `CCA`.
- Hence, we can understand that there some records in database with starting letters as 'CCA'.

Total Payload Mass

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql select sum(payload_mass__kg_) as sum from SPACEXDATA where customer like 'NASA (CRS)'
```

SUM
45596

- The query adds all the payload mass kg where the customer was NASA.
- From the output we can understand that NASA has been carried 45,596 boosters till now, as per the current database information.

Average Payload Mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1

```
%sql select avg(payload_mass__kg_) as Average from SPACEXDATA where booster_version like 'F9 v1.1%'
```

average
2534

- The above query is calculating the average of payload mass kg of boosters with applying the filter with booster version which starts from ‘F9 v1.1’
- Hence, it can be seen that average payload mass of F9 v1.1 has 2534 kg, which is quite low in payload mass range as per the current database.

First Successful Ground Landing Date

List the date when the first successful landing outcome in ground pad was acheived.

```
%sql select min(date) as Date from SPACEXDATA where landing__outcome = 'Success (ground pad)'
```

DATE
2015-12-22

- The above query returns the first successful ground pad landing date.
- First ground pad landing wasn't until the end of 2015.
- Successful landings in general appear starting on ending of 2014.

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%sql select booster_version from SPACEXDATA \
where (payload_mass__kg_ BETWEEN 4001 AND 5999) \
AND (landing__outcome like 'Success (drone ship)')
```

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

- The query returns 4 successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- The query exclude payload mass which is 4000 or 6000 but returns all the records between these range.

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
%sql SELECT mission_outcome, count(*) as Count \
FROM SPACEXDATA \
GROUP by mission_outcome \
ORDER BY mission_outcome
```

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

- The above query returns count of each mission outcome.
- The result shows that SpaceX is having 99% success rate, it means that most of the landing failures are intended.
- From the result, one launch has an unclear payload status and unfortunately one failure in flight.

Boosters Carried Maximum Payload

- The shown query returns the booster versions that carried the highest payload mass of 15600 kg
- The booster F9 B5 B10 variety has similar booster versions.
- This shows that payload mass correlates with the booster version that is being used.
- It tells that F9 B5 B1… series has the highest payload mass capacity of boosters.

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%sql select booster_version, payload_mass__kg_ \
from SPACEXDATA \
where payload_mass__kg_=(select max(payload_mass__kg_) from SPACEXDATA)
```

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

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%sql select MONTHNAME(DATE) as Month, landing__outcome, booster_version, launch_site \
from SPACEXDATA where DATE like '2015%' AND landing__outcome like 'Failure (drone ship)'
```

MONTH	landing__outcome	booster_version	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- The query returns the month, landing outcome, booster version, launch site which have launched on date 2015.
- From the output it is clearly appeared that, in year 2015 stage 1 failed to land on a drone ship, one of those failure was in January, and one was in April.

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

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

```
%sql select landing__outcome, count(*) as count from SPACEXDATA where Date between '2010-06-04' AND '2017-03-20' \
GROUP by landing__outcome ORDER BY count Desc
```

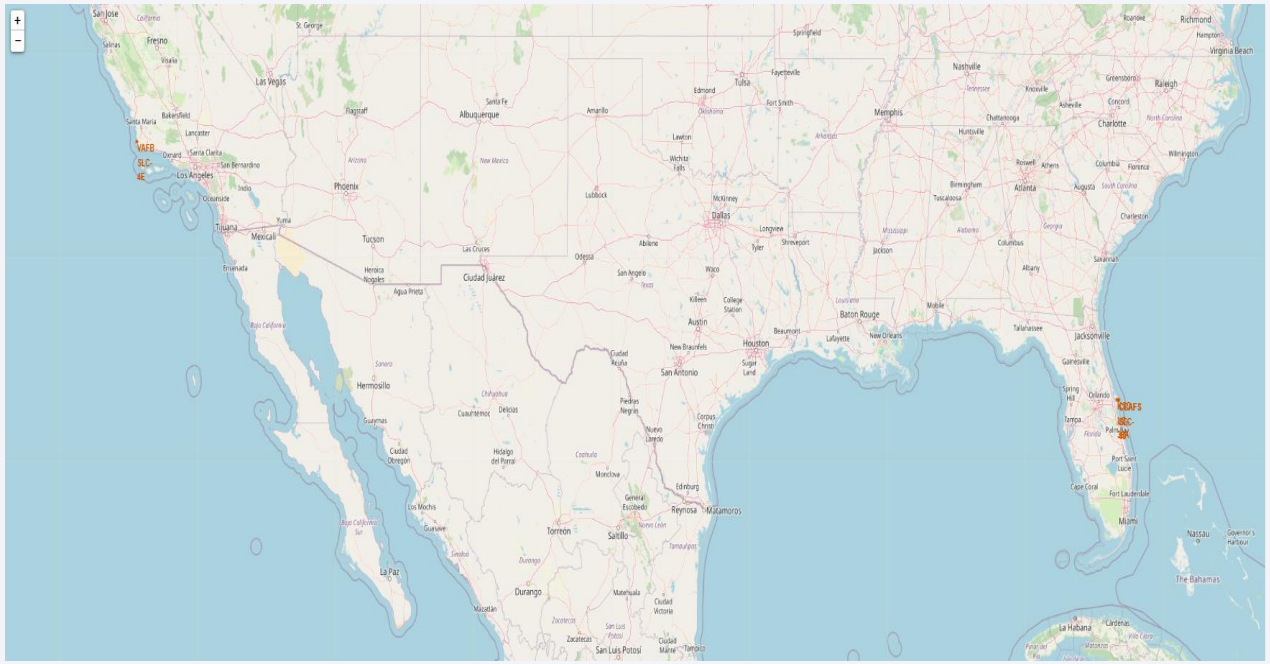
landing__outcome	COUNT
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

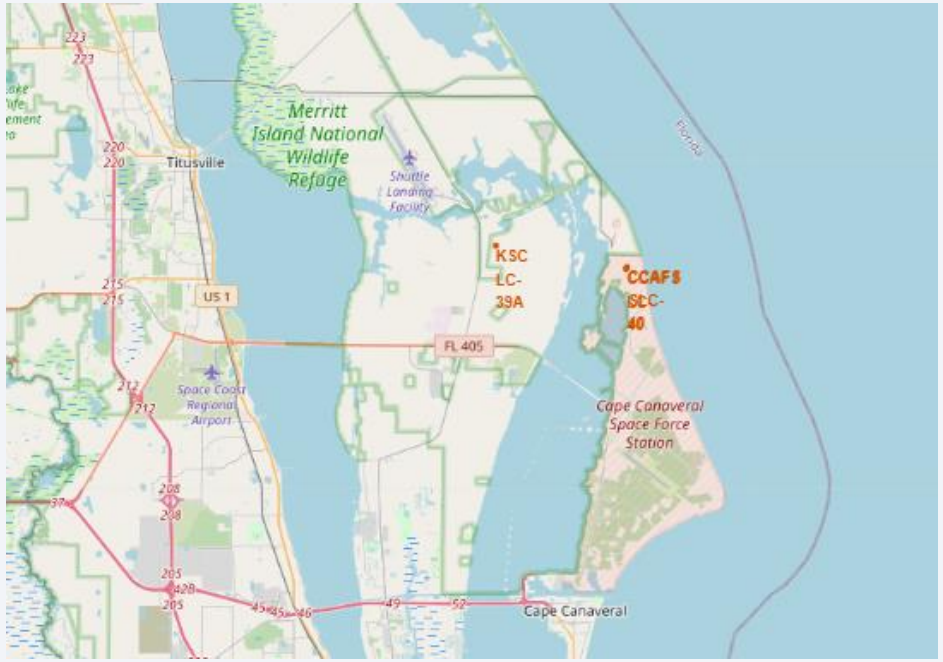
- The query returns a list of counts and of all the landing outcome.
- It includes success and failure and other type of landing outcome with a date range between 2010-06-04 and 2017-03-20.
- After getting the records between the date range, it is sorted in descending order of count of landing outcome.

Section 3

Launch Sites Proximities Analysis

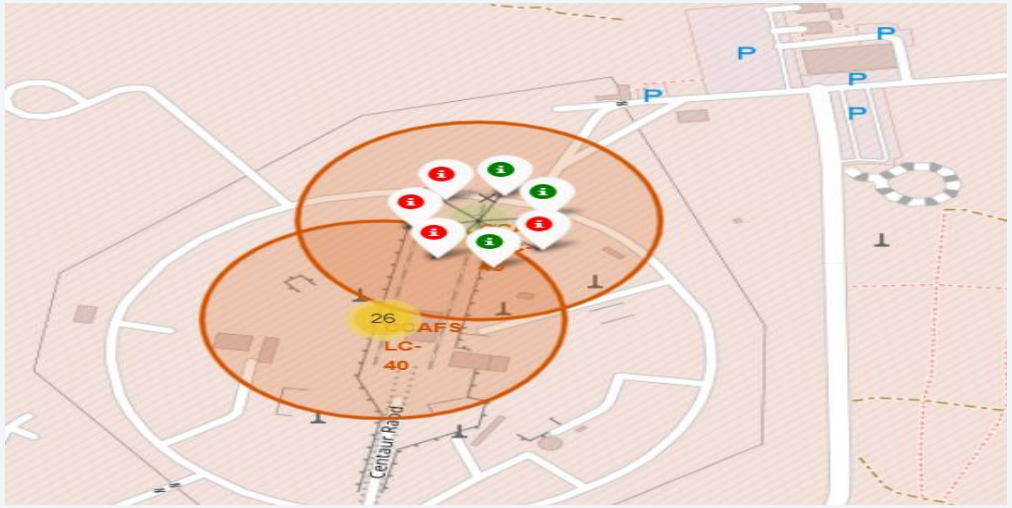
Launch Site Locations





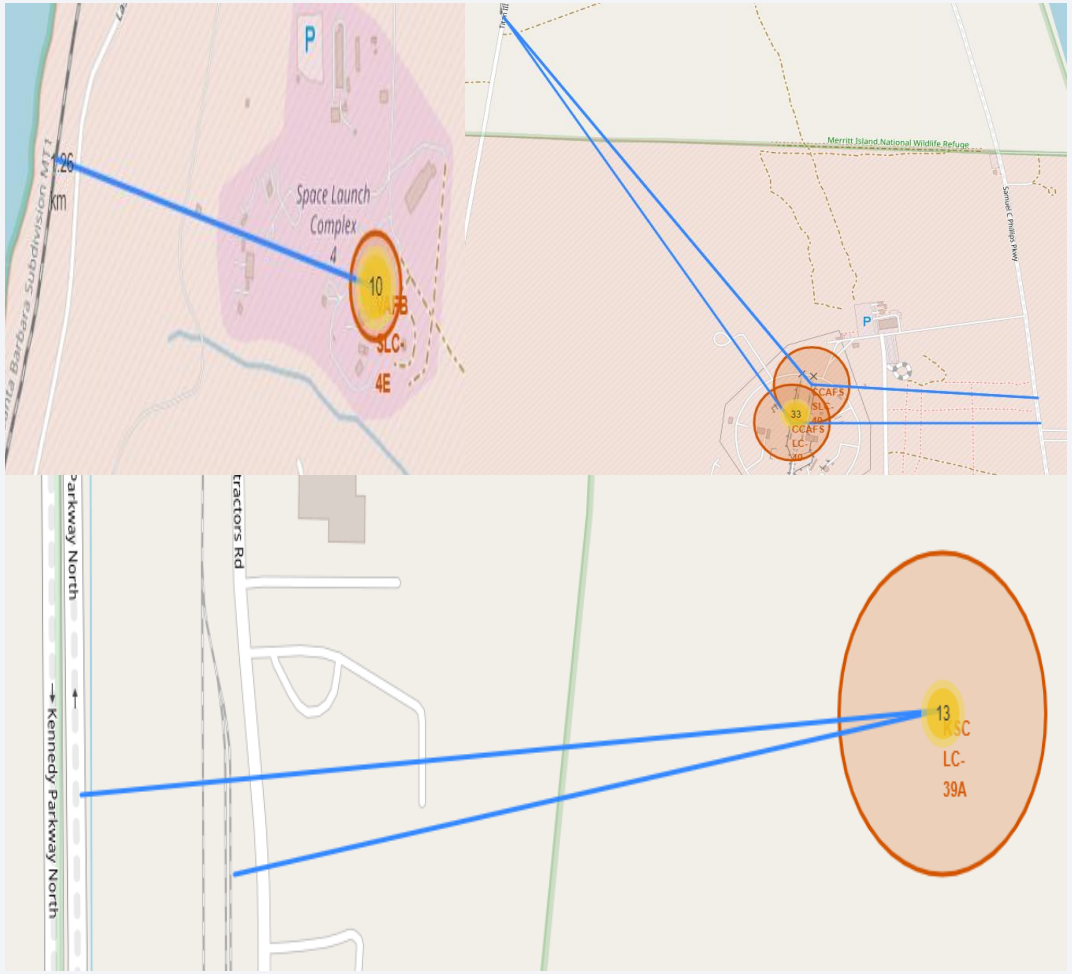
- The above map is showing all the launch site locations globally in US.
- The left map is showing all launch sites relative to US map.
- The right map is showing the two launch sites locations.
- From the above map, all sites are near to ocean.

Success and failure launch sites color-coded markers

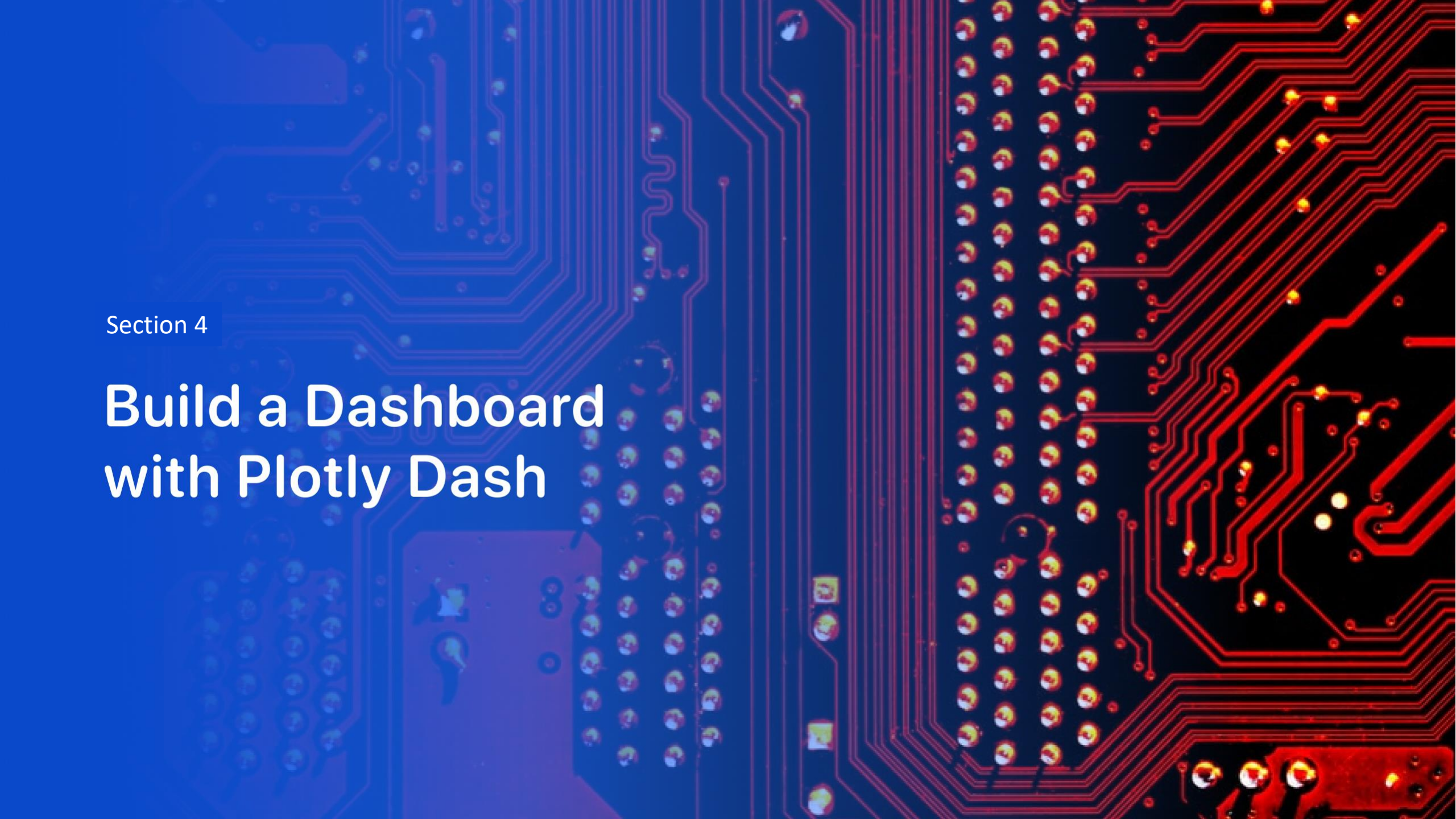


- From the color-labeled markers in marker clusters, you should be able to easily identify which launch sites have relatively high success rates.
- The clusters on folium map can be used to click on to display each successful(green color) and unsuccessful(red color) landing sites
- In this top screenshot shows CCAFS SLC-40 has 3 successful landings and 5 unsuccessful landings.
- In the bottom screenshot shows VAFB SLC-4E has 4 successful(green) and 6 unsuccessful landings(red).

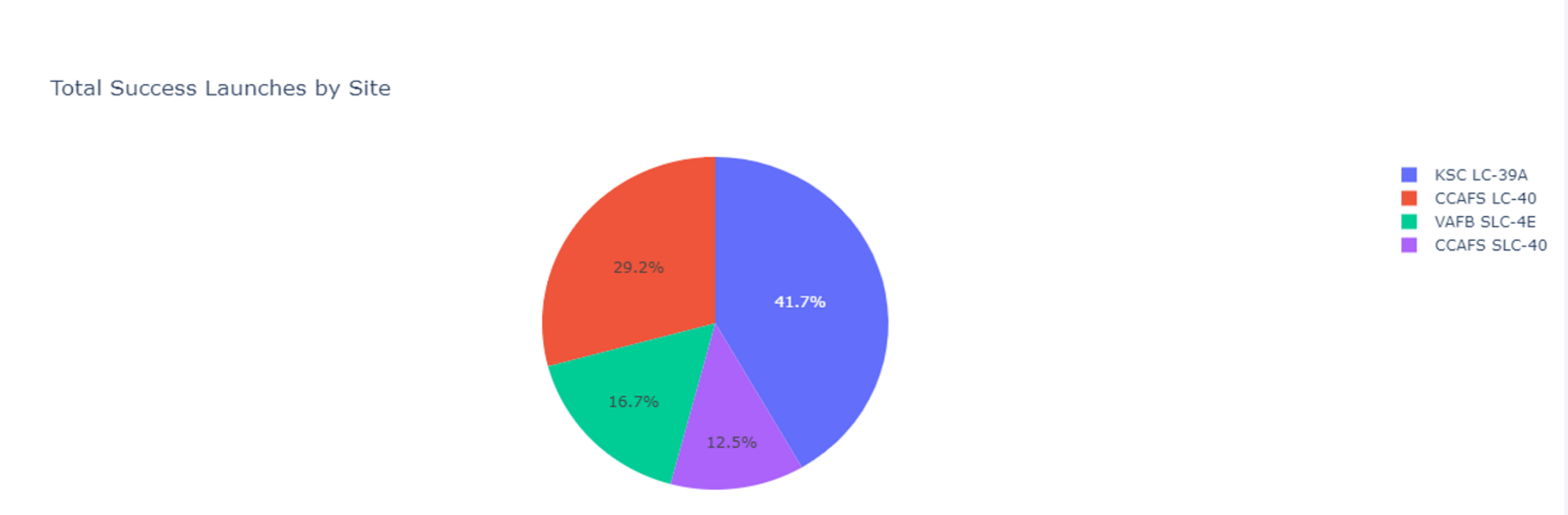
Distance between a launch site to its proximity locations



- With proximity we can calculate the distance between two points on the map based on their Lat and Long values.
- There are different launch sites by SpaceX, each launch site is showing distance from nearest railway, highway and city that is the major factor for transportation cost.
- Launch sites are also closest to some coasts but distance from cities, so launch failure can land in the sea to avoid rockets falling on populated area.

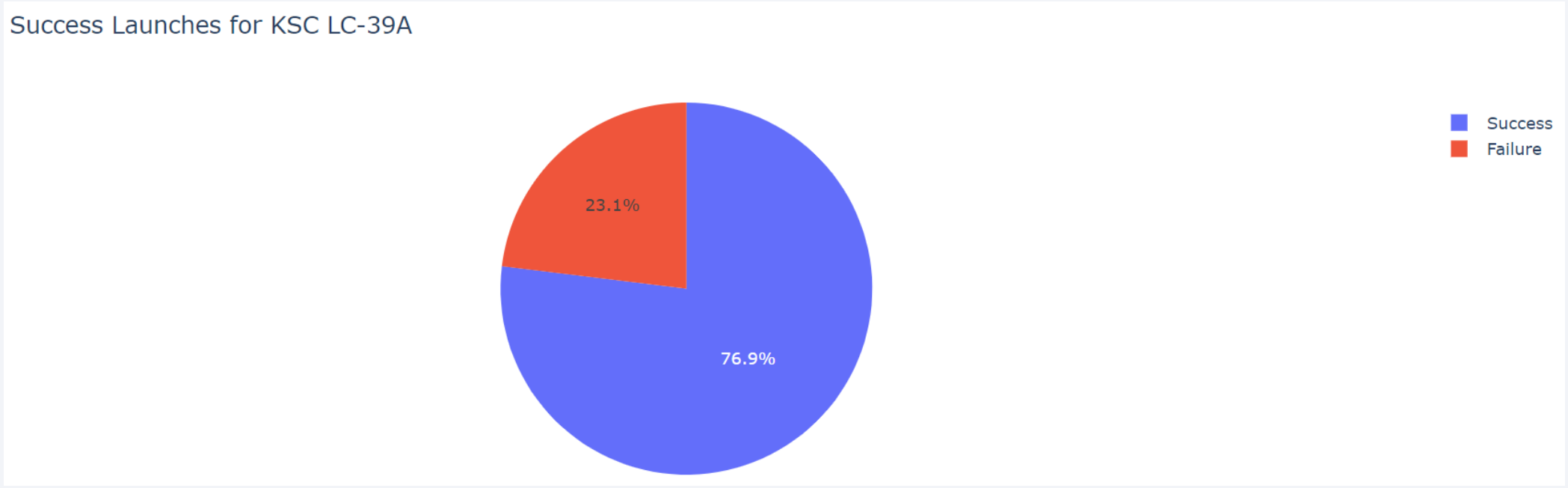

Section 4
Build a Dashboard with Plotly Dash

Successful launch sites – Across all launches



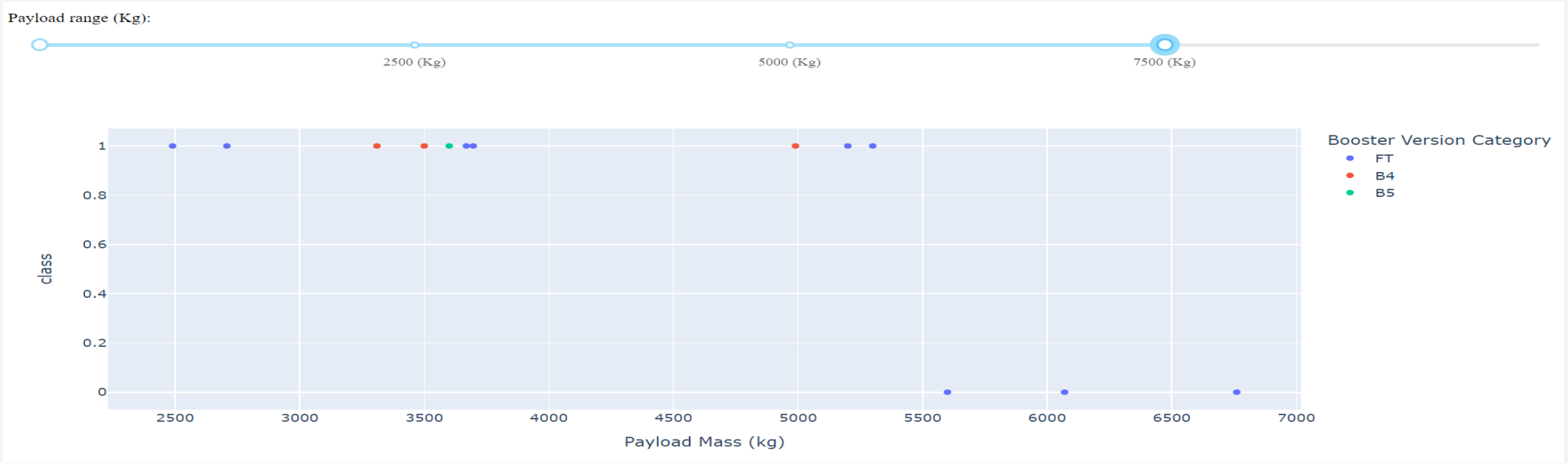
- The pie chart is showing distribution of successful landings across all launch sites.
- CCAFS LC-40 has most of the successful landings were performed.
- CCAFS SLC-40 is new name for launch site so it can be ignored as a smallest distributor in successful launches. After CCAFS SLC-40 is ignored, VAFB SLC-4E is the smallest distributor in successful launch sites.

Highest launch success ratio



- The pie chart shows that the launch site KCS LC-39A has the highest success rate with 76.9%
- Blue color is for success ratio and Orange is for failure ratio.

Payload vs. Launch Outcome with booster versions

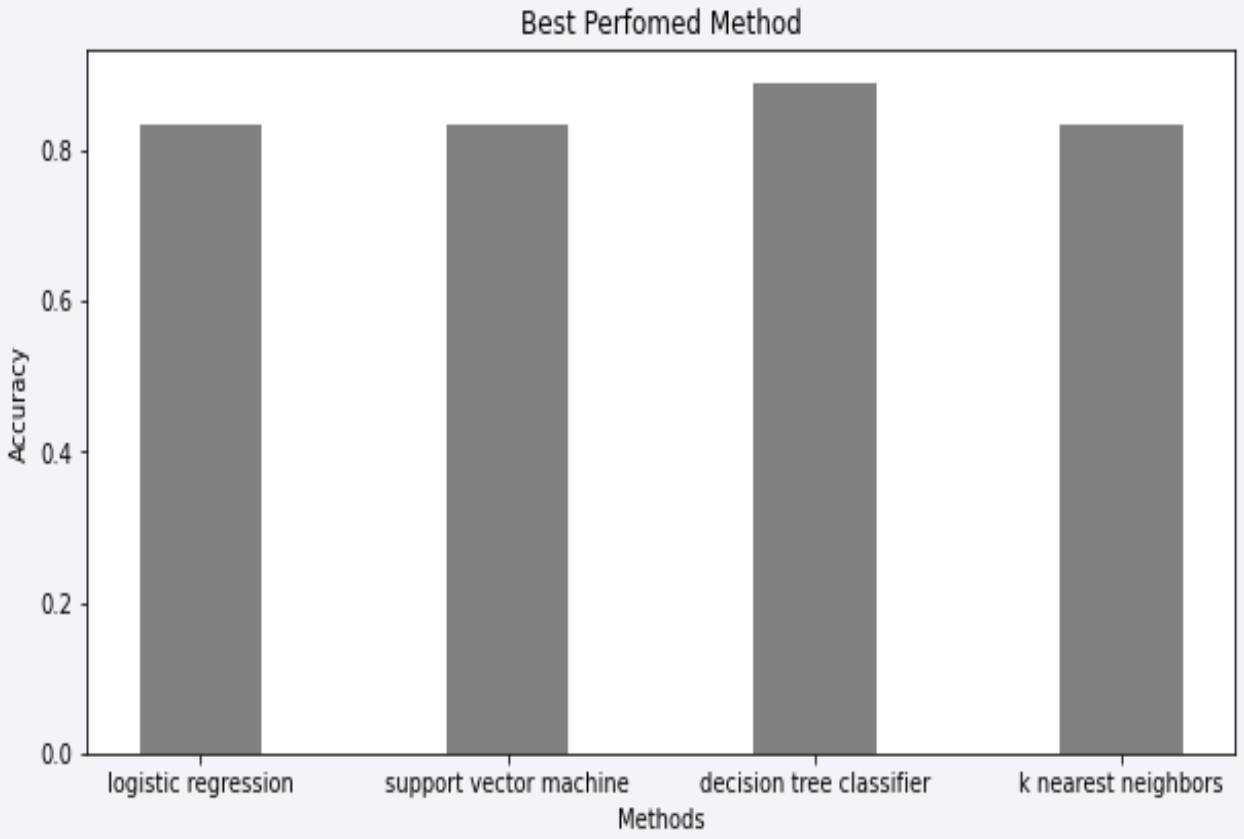


- Payload dashboard has a payload range selector. However, this is sent from 0-1 ratio.
- In the dashboard it has option to slide for different payload mass, in the above screenshot 0-7500kg payload is showing.
- There 3 booster versions(FT, B4 and B5) with the capacity of range between 0 – 7500 kg.

Section 5

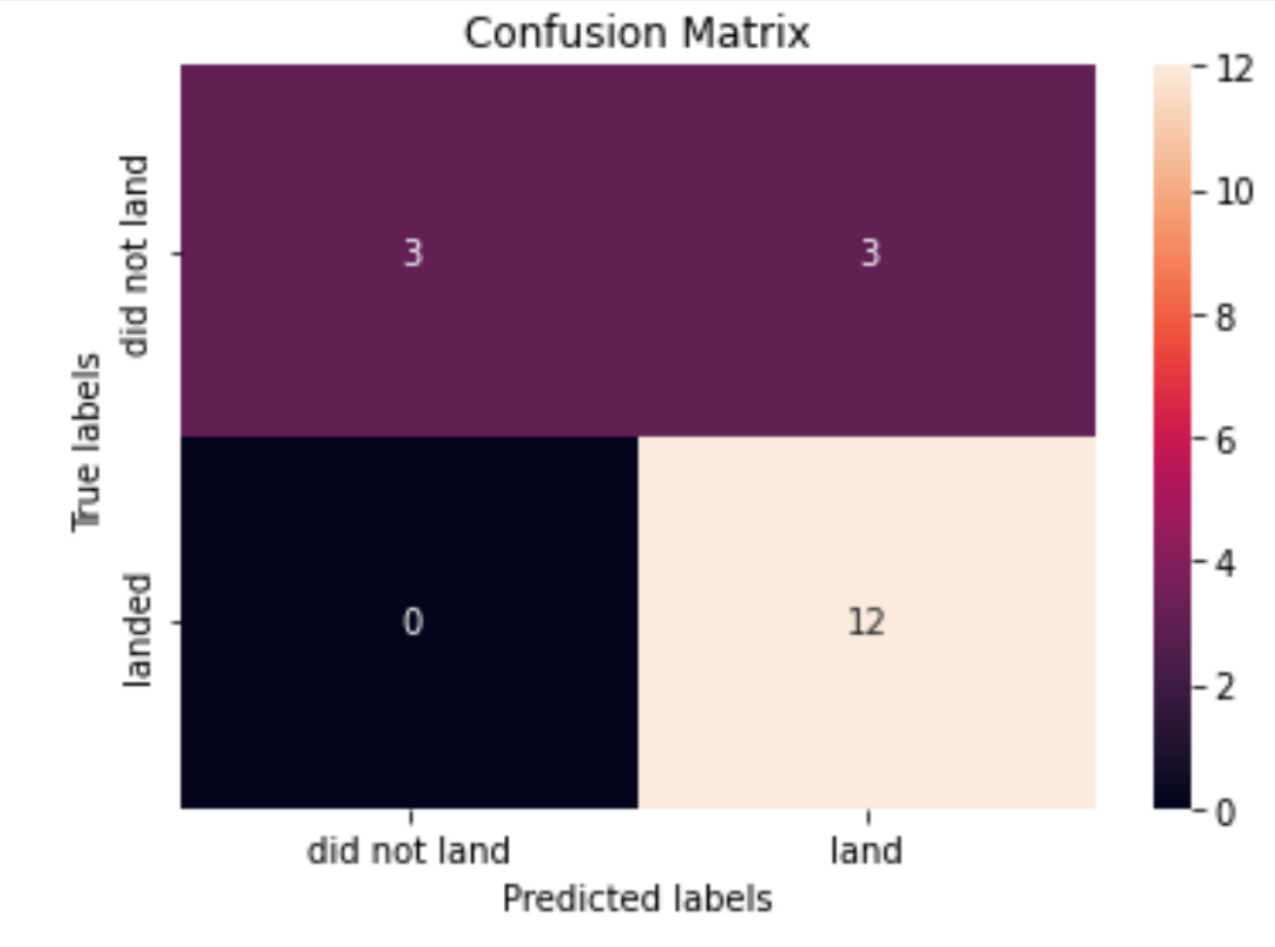
Predictive Analysis (Classification)

Classification Accuracy



- All models has approximately same appearance in the bar plot on the test dataset(small in size of 18 records) at 83.33% accuracy.
- This might cause large variance in accuracy in some ML models like **Decision tree classifier** while repetitive runs.
- Due to small size of data it is hard to take decision for good accuracy. Hence, we need more data to determine the best model.

Confusion Matrix



- To determine the correct predications, we should start from top left to bottom right boxes.
- From the matrix graph, it is visible that all models performed the same for test data but the confusion matrix is the same across all models.
- The model predicated 12 successful landings when the label was true for successful landings.
- The model predicated 3 unsuccessful landings when the true label was unsuccessful landings.
- For the **False-Positive** case - The model predicated 3 successful landings when the true label was unsuccessful landings.

Conclusions

- End Goal : Develop a machine learning model for SpaceY who wants to compete against SpaceX.
- The major goal was to build a predictive model that can correctly predict that when stage1 will land successfully is it worth to save $100 million USD.
- To make the predictive model, in this project SpaceX public API and SpaceX Wikipedia page are used.
- Data collection is done with SpaceX API and Wikipedia page by performing some methodologies, like data collection, web scrapping, data wrangling etc.
- Extracted, Transformed and loaded the gathered data into DB2 database.
- Created interactive dashboard for visualization using Plotly python library from the data sets from a csv file.
- Created machine learning model with an accuracy of 83.33%, which concludes that the landing of stage1 can help in reducing cost to SpaceY company.
- By visualization analysis we have found that best locations for launch site, capacity of different booter versions and what are the best landing locations across entire US.
- The accuracy could be more accurate if there is more data to build the machine learning model.
- However, Successfully generated a machine learning model that predict the successful landings or launch locations site.

Appendix

- Github URL for relevant assets that have been created during this project :
  - https://github.com/kamesh01/Applied-Data-Science---Capstone

- Speacilly thanks to all the instructors :
  - https://www.coursera.org/professional-certificates/ibm-data-science#instructors

Thank you!