Winning Space Race with Data Science

Manpreet Saluja
11-03-2022

Outline

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

Executive Summary

- Data was collected from public SpaceX API and SpaceX Wikipedia page using webscraping.
- Explored data using SQL, visualization , folium maps , and dashboards. Gathered relevant columns to be as useful features.
- Changed all categorical variables to binary using one hot encoding.
- Satandardized data and used GridSearchCV to find the best parameters for machine learning models.
- Visualized accuracy score of all models.
- Four machine learning models were produced: Logistic Regression,Support Vector Machine,Decision Tree Classifier and K nearest neighbors.
- All ML models produced similar results with accuracy rate of about 83%.
- More data is needed for better model determination and accuracy.

Introduction

BACKGROUND:

- Space X (Falcon 9) has best pricing ($62 million vs upwards $165 million USD)
- Largely due to ability to recover part of rocket (stage1)
- Space Y wants to compete with Space X.

Challenge:

- Determine the price of each launch
- Gather public information about Sapce X and create dashboards for the team.
- Train a machine learning model to predict successful stage 1 recovery.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Combined data from SpaceX public API and Space X Wikipedia page.
- Perform data wrangling
  - Classifying true landings as successful and unsuccessful otherwise.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Tuned models using GridSearchCV

Data Collection

The data collection process involved a combination of API requests from SPACEX REST API and Web Scraping data from a table in Wikipedia entry.

Both of the collection methods were used in order to get complete information about the launches for a more detailed analysis.

Data columns obtained by SpaceX REST API:

Flightnumber,Date,BoosterVersion,PayloadMass,Orbit,LaunchSite,Outcome,Flights,GridFins,Reused,Legs,LandingPad,Block,ReusedCount,Serial,Longitude,Latitude.

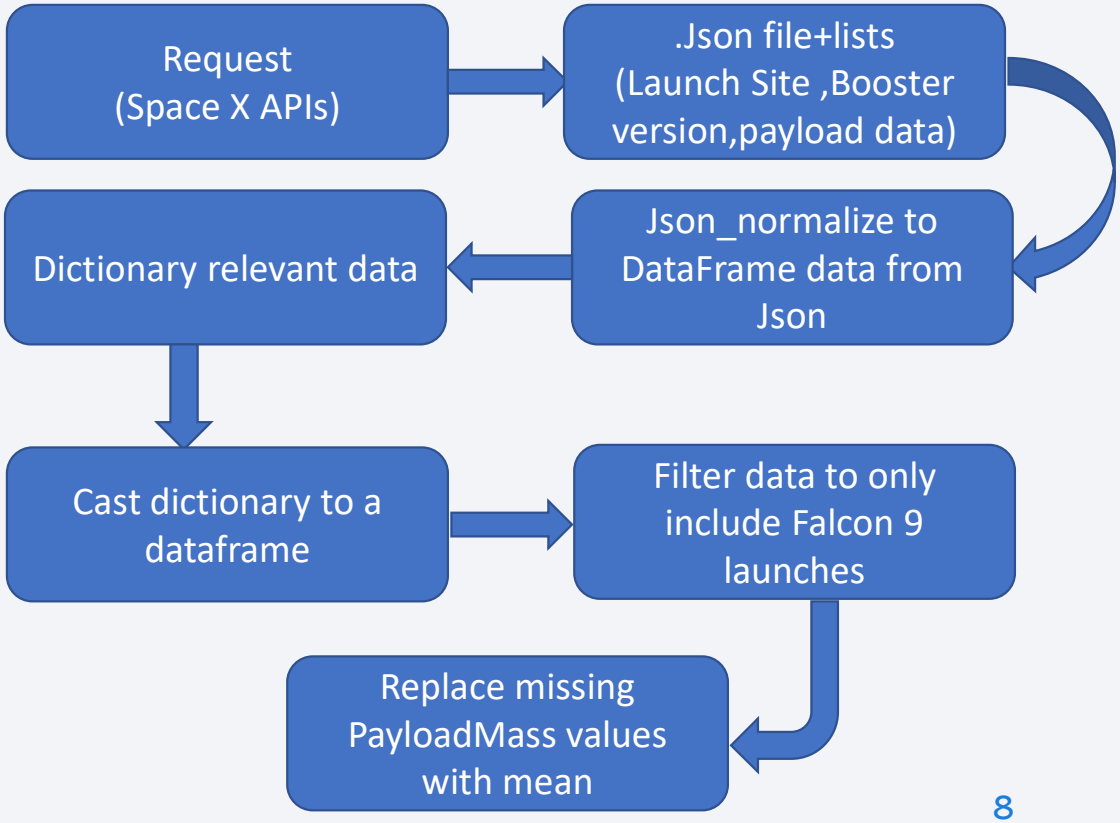
Data columns obtained by Web Scraping:

Launch Site,Payload,Payload Mass,Orbit,Customer,Launch Outcome,Version Booster,Booster Landing,Date,Time.

Data Collection – SpaceX API

URL:

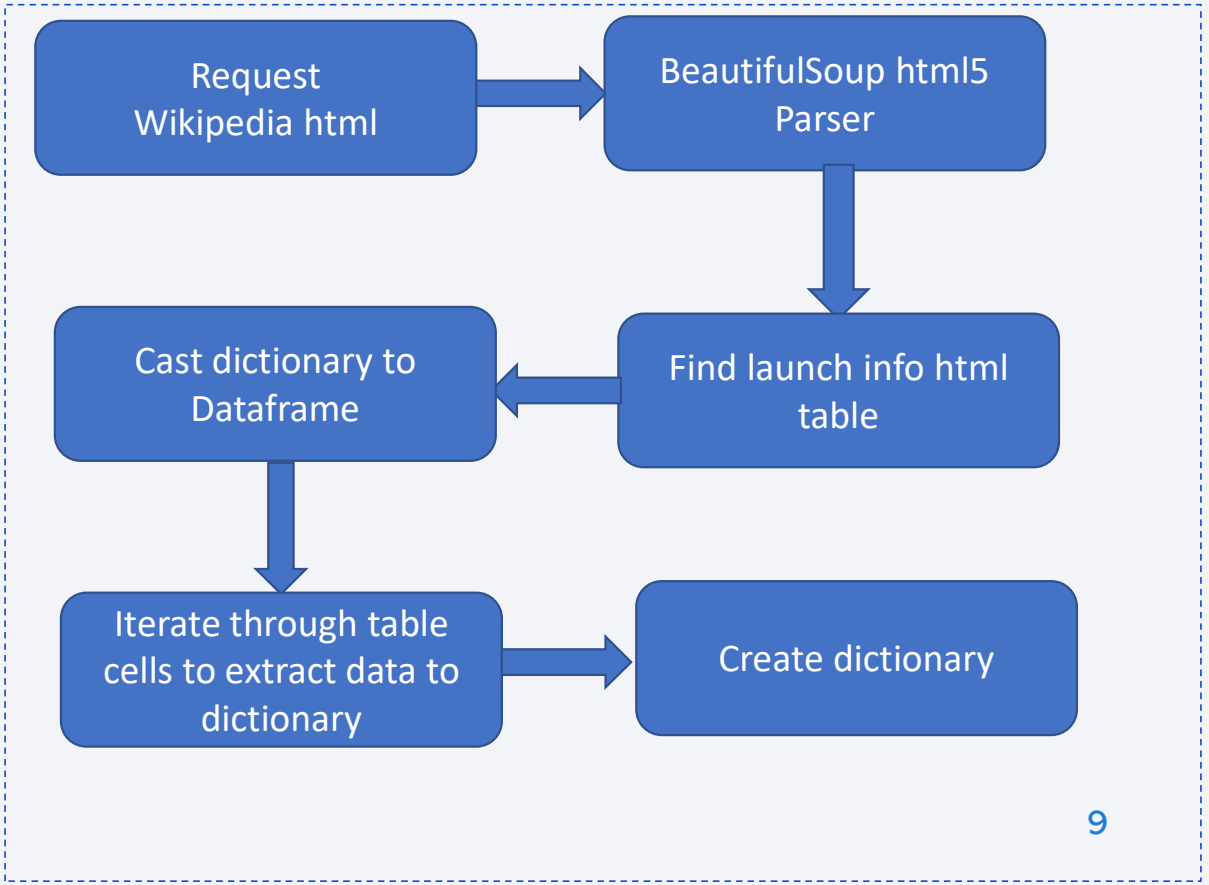
https://github.com/shibal-7/Applied-Data-Science-Capstone/blob/main/Data%20Collection%20API%20Lab.ipynb



Data Collection - Scraping

URL:

https://github.com/shibal-7/Applied-Data-Science-Capstone/blob/main/Data%20Collection%20with%20web%20scraping.ipynb



Data Wrangling

URL:

https://github.com/shibal-7/Applied-Data-Science-Capstone/blob/main/Data%20Wrangling.ipynb

- Created a training label with landing outcomes where successful=1 and failure=0.
- Outcome column has two components :'Mission Ouctome','Landing Outcome'.
- New training label column 'class' with a value of 1 if mission outcome is True and 0 otherwise.

Value mappings

- True ASDS ,True RTLS and True Ocean – set to 1
- None None , False ASDS, None ASDS , False Ocean , False RTLS –set to 0.

EDA with Data Visualization

URL: https://github.com/shibal-7/Applied-Data-Science-Capstone/blob/main/EDA%20with%20Data%20Visualisation.ipynb

EDA performed on variables Flight Number,Payload Mass,Launch Siter,Orbit,Class and Year.

Plots used:

- Flight Number vs Payload mass ,Flight Number vs Launch Site, Payload Mass vs Launch Site , Orbit vs Success rate, Flight number vs Orbit,Payload vs Orbit and Success Yearly trend.
- Scatter plots,line charts and bar plots were used to compare relationships between variables to decide if a relationship exists so that they could be used in training the machine learning model.

EDA with SQL

URL: https://github.com/shibal-7/Applied-Data-Science-Capstone/blob/main/EDA.ipynb

- Loaded data set into IBM DB2 Database
- Queried using SQ Python Integration
- Queries were made to get a better understanding of the dataset
- Queried information about launch site names,mission outcomes,various payload sizes of all customers and booster versions and landing outcomes.

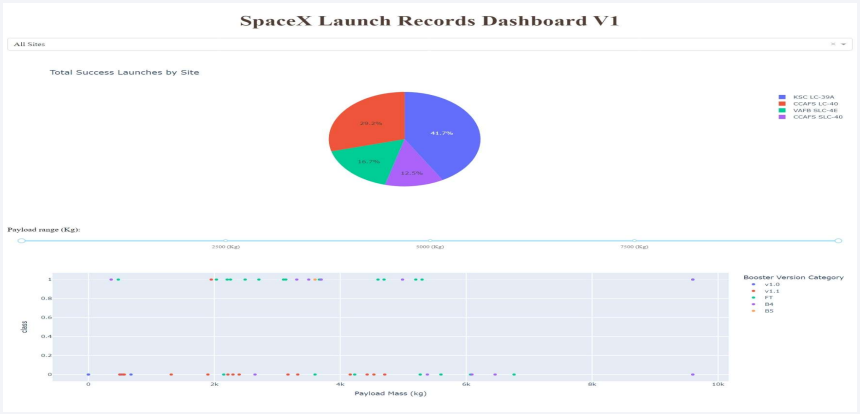
Build an Interactive Map with Folium

URL: https://github.com/shibal-7/Applied-Data-Science-Capstone/blob/main/Interactive%20Visual%20Analytics%20and%20dashboard.ipynb

Launch Sites Locations analysis with Folium

- Folium maps mark Launch Sites ,successful and unsuccessful landings, and a proximity example to key locations:Railway ,Highway ,Coast and City.
- This allows us to understand why launch sites may be located where they are .Also visualizes successful landings relative to location.

Build a Dashboard with Plotly Dash



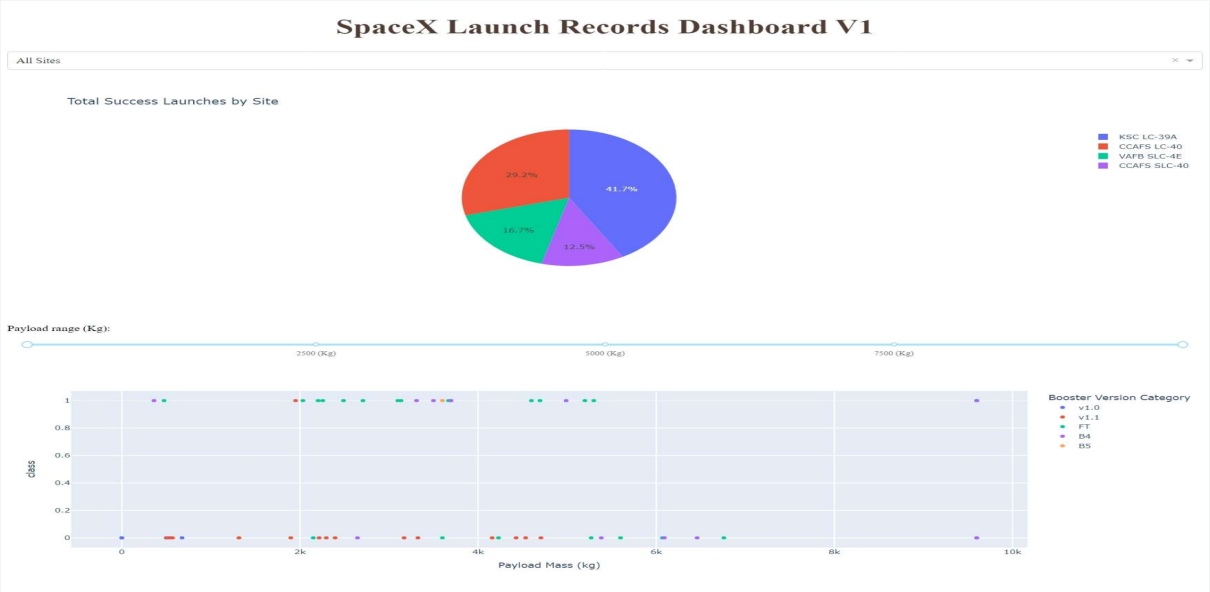
- Added a launch Site dropdown input component.
- The Dashboard contained acallback function to render pie chart displaying the total success launches by site. KSC LC-39A has the highest success rate of 41.7%.
- Added a callback function to render the success-payload-scatter-chart scatter plot.
- Added a Range Slider to Select Payload.

Predictive Analysis (Classification)

URL: https://github.com/shibal-7/Applied-Data-Science-Capstone/blob/main/Machine%20Learning.ipynb

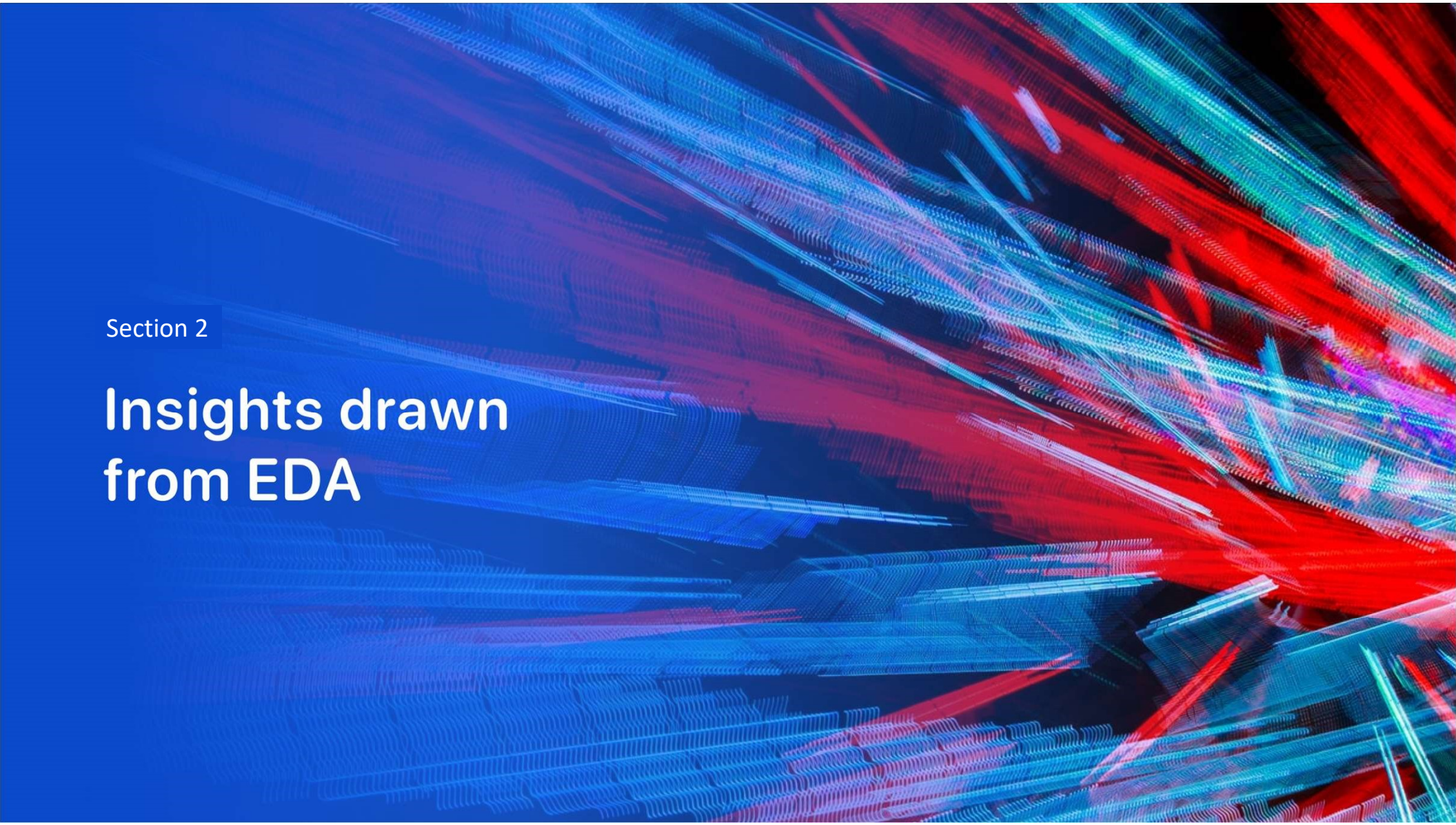
- Split label column / Class from dataset.
- Fit and transform features using standard scaler.
- Train,Test and Split.
- GridSearchCV(cv=10) to find optimal parameters.
- Confusion matrix for all models.
- Barplot to compare scores of models.

Results



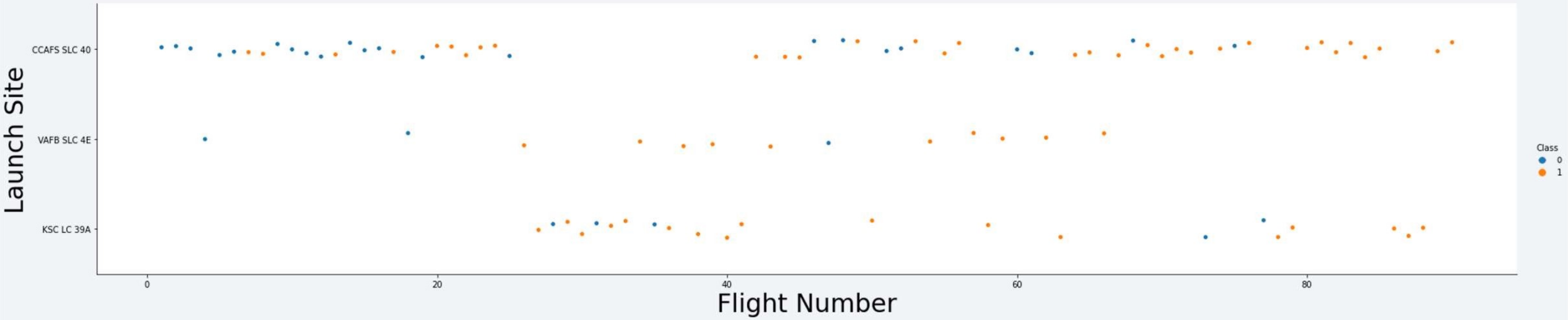
Preview of the Plotly Dashboard .

The following slides will show the results of EDA with visualization,EDA with SQL ,Interactive Map with folium and finally the results of our model with 83% accuracy.


Section 2
Insights drawn
from EDA

Flight Number vs. Launch Site

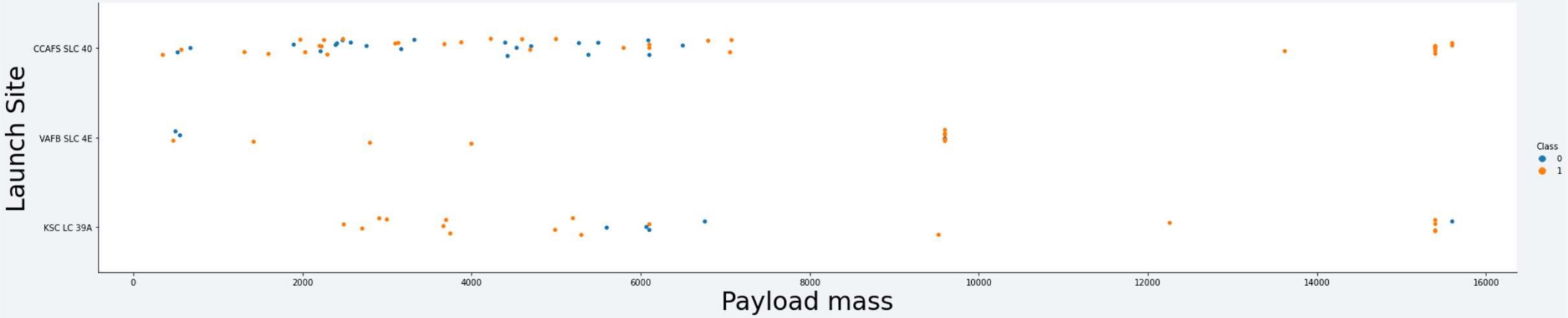
- Scatter plot of Flight Number vs. Launch Site



From the above Launch Site vs Flight Number scatter plot ,we see that as the flight number increases, the first stage is more likely to land successfully. Also the launch site VAFB SLC 4E and KSC LC 39A has a higher success rate than CCAFS SLC 40.

Payload vs. Launch Site

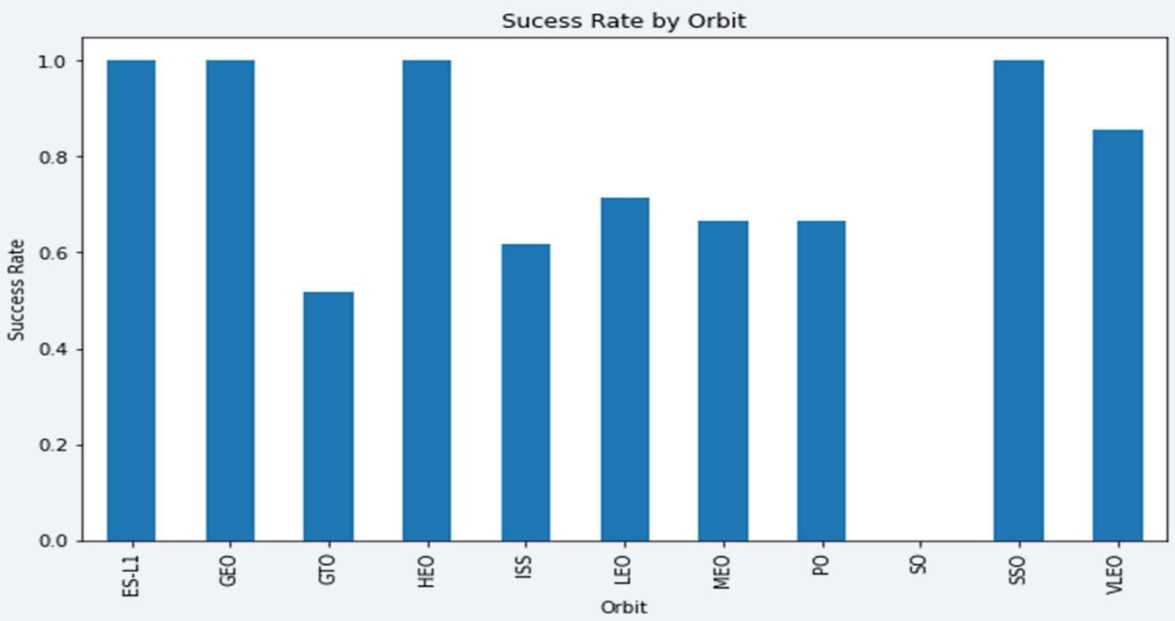
- Scatter plot of Payload vs. Launch Site



From the above Payload vs Launch Site scatter point chart, we observe that for the VAFB-SLC launchsite there are no rockets launched for heavy spayload mass(greater than 10000).

Success Rate vs. Orbit Type

- Bar chart for the success rate of each orbit type



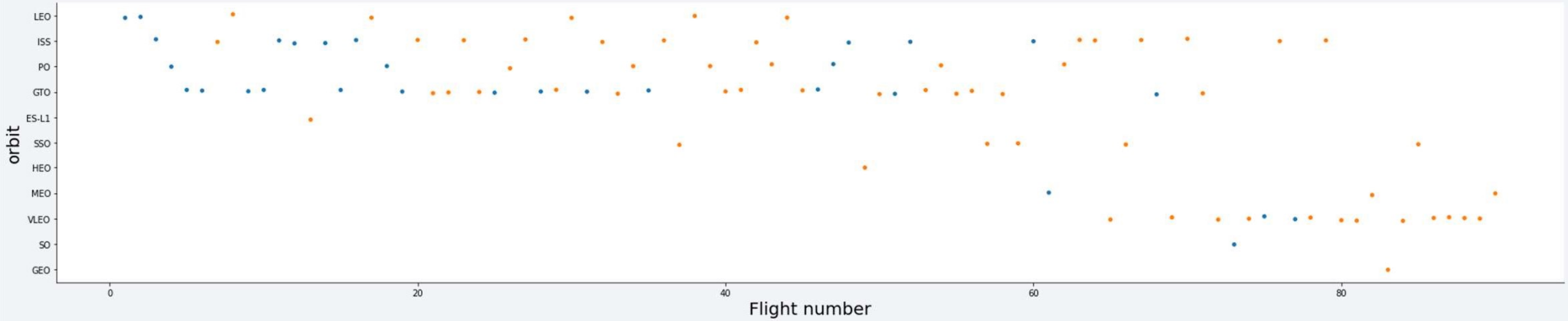
The orbit types ES-L1,GEO,HEO,SSO,VLEO have 100% success rates

The orbit type SO has zero success rate.

Orbits with success rates between 50% to 85% are GTO ,ISS,LEO,MEO,PO

Flight Number vs. Orbit Type

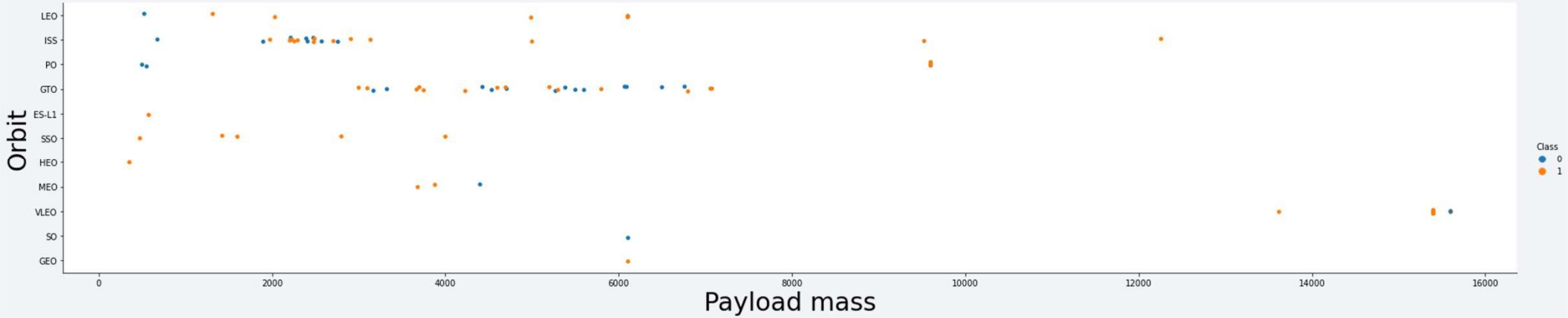
- Scatter plot of Flight number vs. Orbit type



From the given Flight number vs orbit type scatter plot ,we observe that for LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type

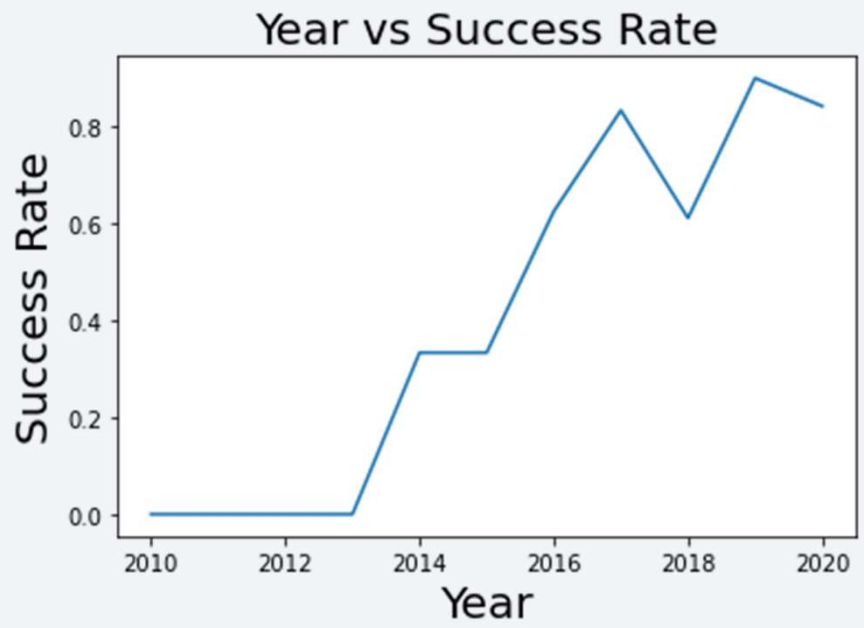
- Scatter point of payload vs. orbit type



- From the Payload vs Orbit type scatter plot,we observe that ,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.

Launch Success Yearly Trend

- Line chart of yearly average success rate



We observe that that the sucess rate since 2013 kept increasing till 2020.

All Launch Site Names

```
In [14]: %sql select DISTINCT(launch_site) from SPACEXTBL
 * ibm_db_sa://vpq97161:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

Out[14]:

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

Displaying the names of unique launch sites in the space mission.

Launch Site Names Begin with 'CCA'

```
In [22]: %sql SELECT * FROM SPACEXTBL WHERE launch_site LIKE 'CCA%' LIMIT 5
```

```
 * ibm_db_sa://vpq97161:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

Out[22]:

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

Displaying 5 launch site names which begin with CCA

Total Payload Mass

```
In [6]: %sql  SELECT sum(payload_mass__kg_) as total_payload_mass FROM SPACEXTBL where customer = 'NASA (CRS)'
 * ibm_db_sa://vpq97161:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

Out[6]:

total_payload_mass
45596

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

Average Payload Mass by F9 v1.1

```
In [11]: %sql SELECT AVG(payload_mass__kg_) as avg_payload_mass FROM SPACEXTBL where booster_version = 'F9 v1.1'
 * ibm_db_sa://vpq97161:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

Out[11]:

avg_payload_mass
2928

Displaying the average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

```
In [20]: %sql select min(DATE)  FROM SPACEXTBL where "Landing _Outcome" = 'Success (ground pad)' LIMIT 1
          * ibm_db_sa://vpq97161:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
         Done.
Out[20]:
         1
         2015-12-22
```

Displaying the first successful Ground Landing Date

Successful Drone Ship Landing with Payload between 4000 and 6000

In [22]:
```
%sql select booster_version,payload_mass__kg_ from SPACEXTBL WHERE "Landing _Outcome" = 'Success (drone ship)' and payload_mass__kg_ BETWEEN 4000 AND 6000
```

```
 * ibm_db_sa://vpq97161:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

Out[22]:

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

The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

```
In [21]: %sql SELECT mission_outcome,count(mission_outcome) from spacextbl group by mission_outcome  order by count(mission_outcome) DESC
 * ibm_db_sa://vpq97161:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

Out[21]:

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

The total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

```
In [23]: #%sql SELECT (SELECT MAX(payload_mass__kg_) from spacextbl), booster_version from spacextbl where payload_mass__kg_=(SELECT MAX(payload_mass__kg_) from spacextbl)  limit 1;
         %sql SELECT booster_version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) as max_payload FROM SPACEXTBL)
          * ibm_db_sa://vpq97161:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
        Done.
```

Out[23]:

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

The names of the booster which have carried the maximum payload mass

2015 Launch Records

```
In [27]: %%sql
SELECT DATE, "Landing _Outcome", Booster_Version, Launch_Site
FROM SPACEXTBL WHERE Year(Date) = '2015' AND "Landing _Outcome" LIKE '%drone ship%';
```

```
 * ibm_db_sa://vpq97161:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

Out[27]:

DATE	Landing _Outcome	booster_version	launch_site
2015-01-10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
2015-06-28	Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40

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

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

```
In [28]: %%sql
SELECT count(*), "Landing _Outcome" FROM SPACEXTBL
WHERE
Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY "Landing _Outcome" ORDER BY 1 DESC;
```

```
 * ibm_db_sa://vpq97161:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

Out[28]:

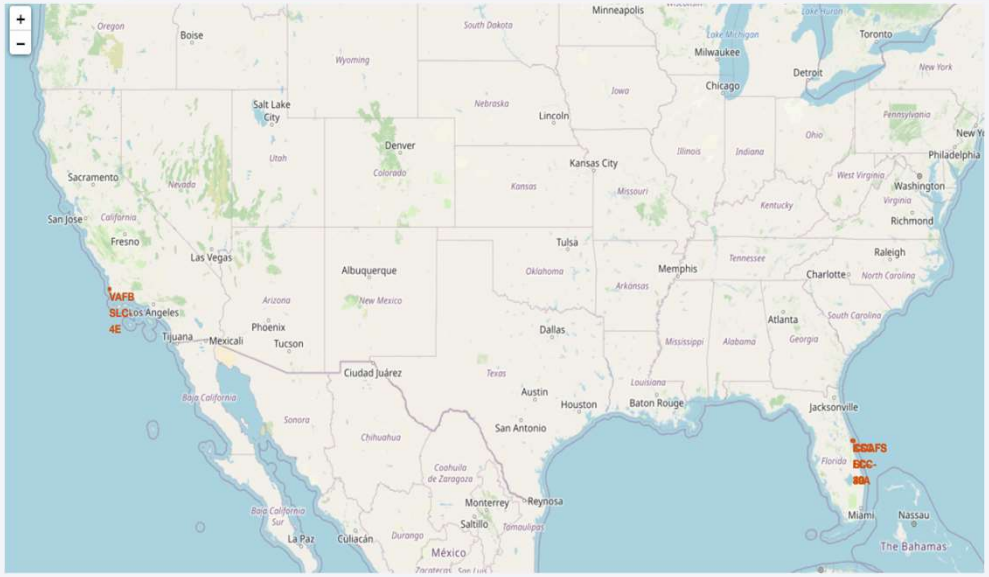
1	Landing _Outcome
10	No attempt
5	Failure (drone ship)
5	Success (drone ship)
3	Controlled (ocean)
3	Success (ground pad)
2	Failure (parachute)
2	Uncontrolled (ocean)
1	Precluded (drone ship)

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

Section 3

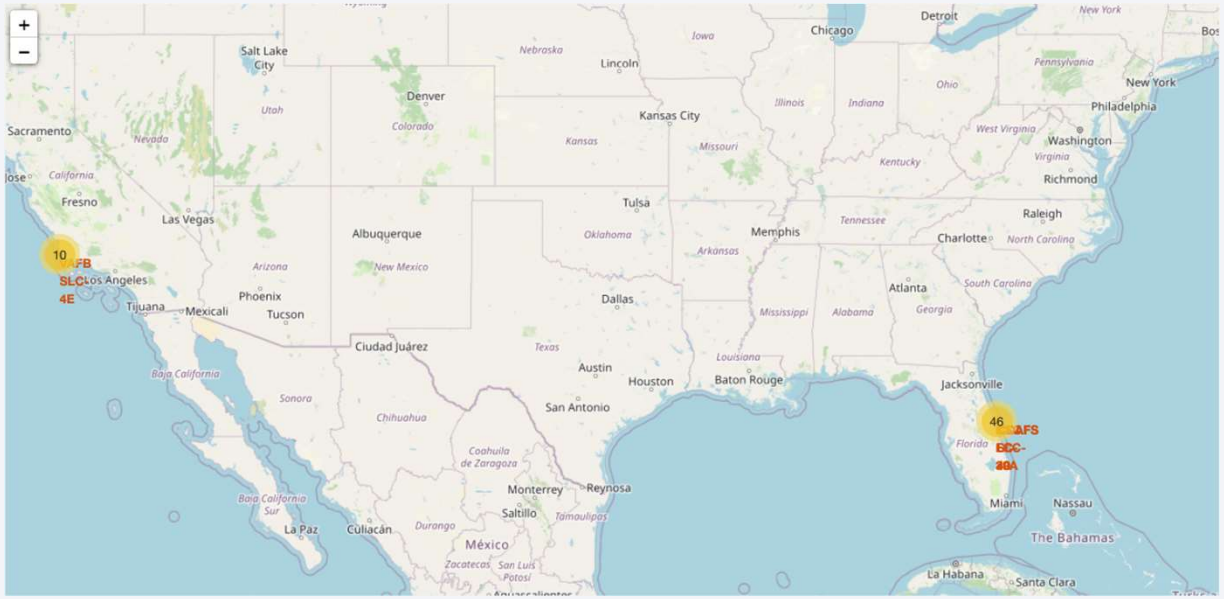
Launch Sites Proximities Analysis

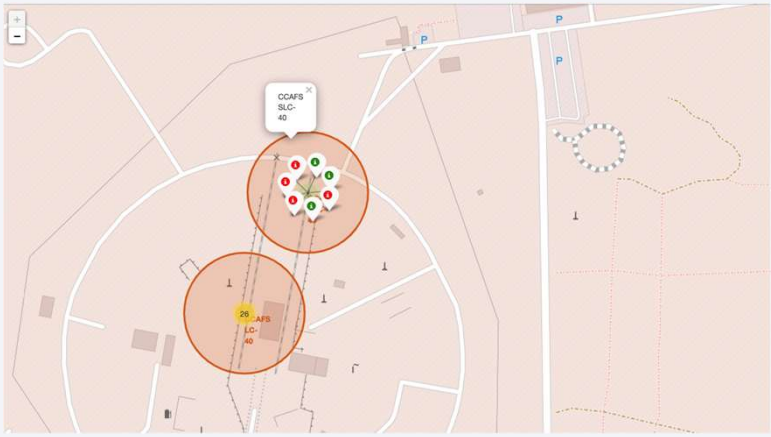
LAUNCH SITE LOCATIONS



The left map shows all launch sites relative US map.The right map shows the Florida launch sites since they are very close to each other. All launch sites are near the ocean.

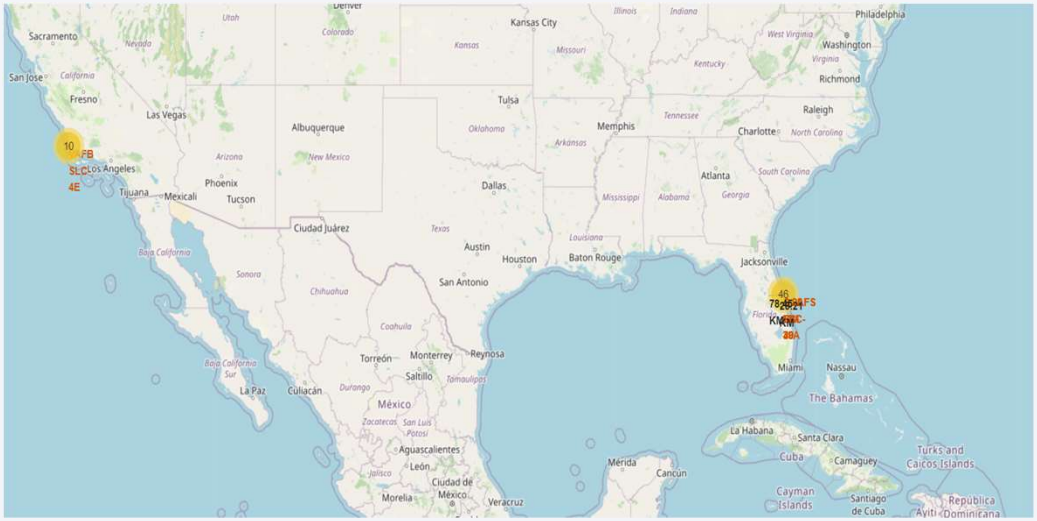
Color-Coded Launch Markers



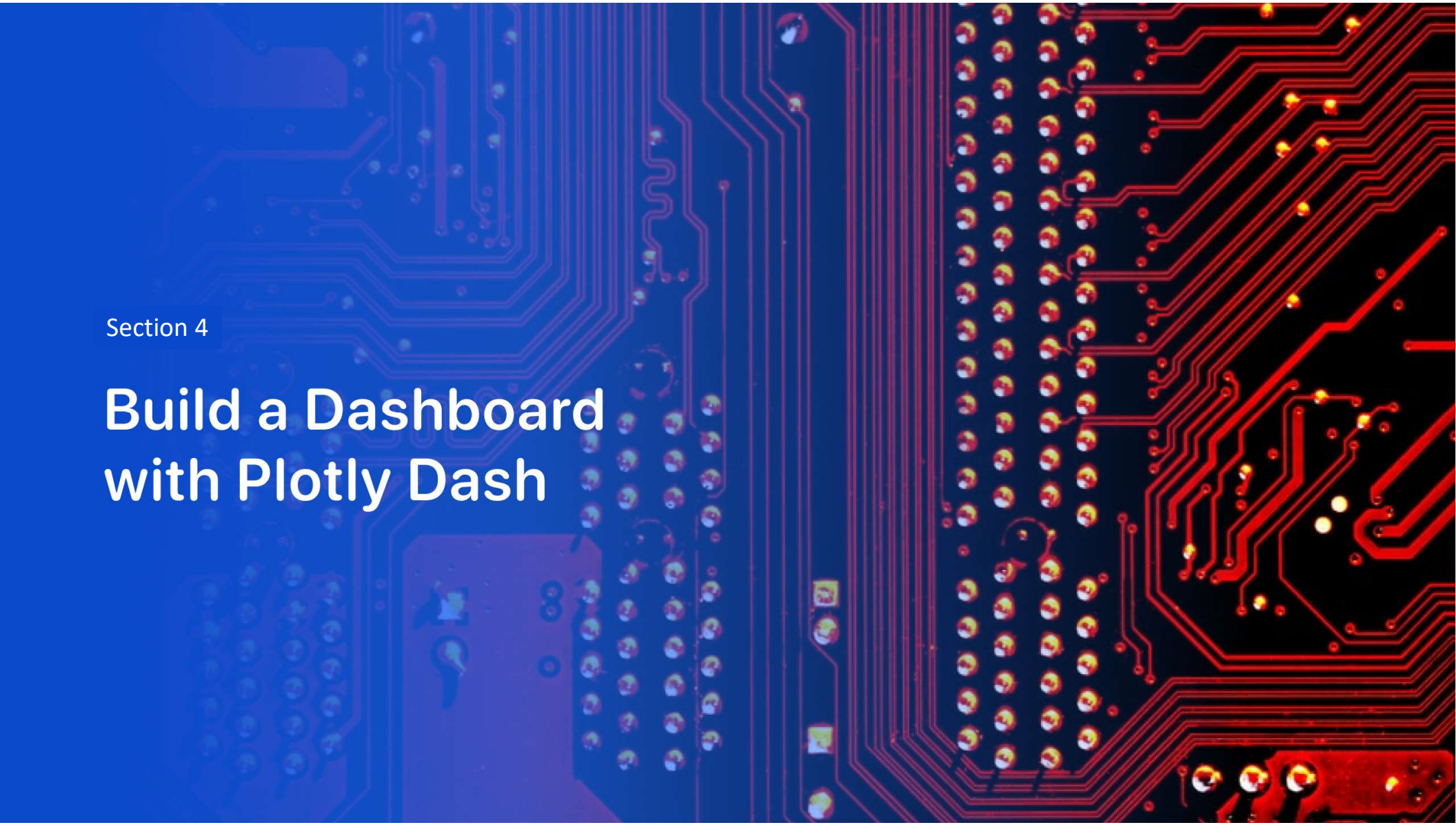


Clusters on Folium map can be clicked on to display each successful landing (green icon) and landing icon(red icon).In this example CCAFS SLC-40 shows 4 unsuccessful and 3 successful landings.

Key Location Proximities



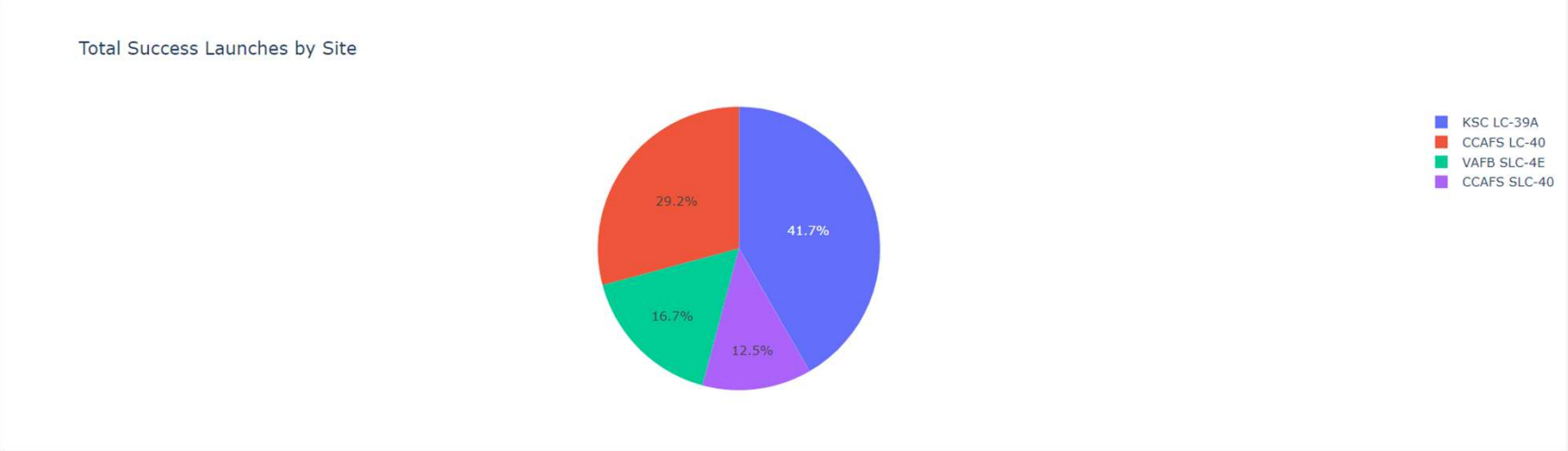
Using KSC LC-39A as an example ,launch sites are very close to railways for large part and supply transportation.Launch sites are close to highways for human and supply transport.Launch sites are closer to coasts and relatively far from cities so that launch failures can land in the sea to avoid rockets falling in densely populated areas.



Section 4

Build a Dashboard with Plotly Dash

Successful Launches across Launch Sites

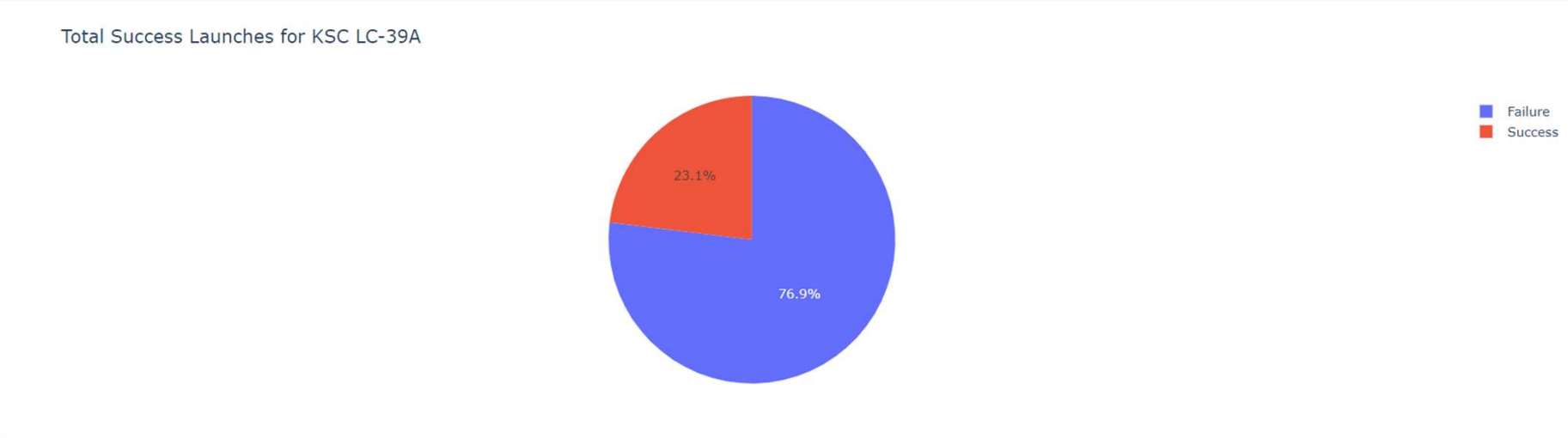


This is the distribution of successful landings across all launch sites.

VAFB has the smallest share of successful landings which may be due to smaller sample and increase in difficulty of launching in the west coast.

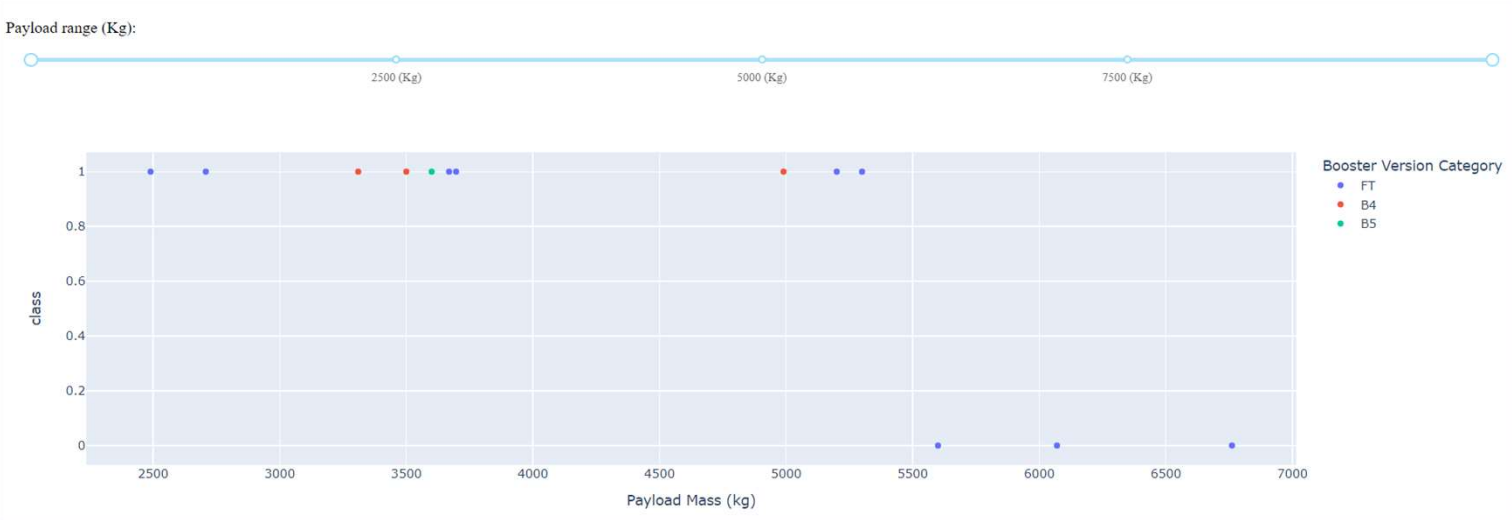
KSC LC-39A has the most number of successful landings while CCAFS LC-40 and VAFB SLC-4F have almost the same amount of successful landings.

Highest Success rate launch site



KSC LC-39A has the most number of successful landings.

Payload Mass vs Success rate vs Booster version Category



The Payload range selector is set from 0-10000 instead of max payload of 15600.

Class indicates 1 for successful landing and 0 for failure.

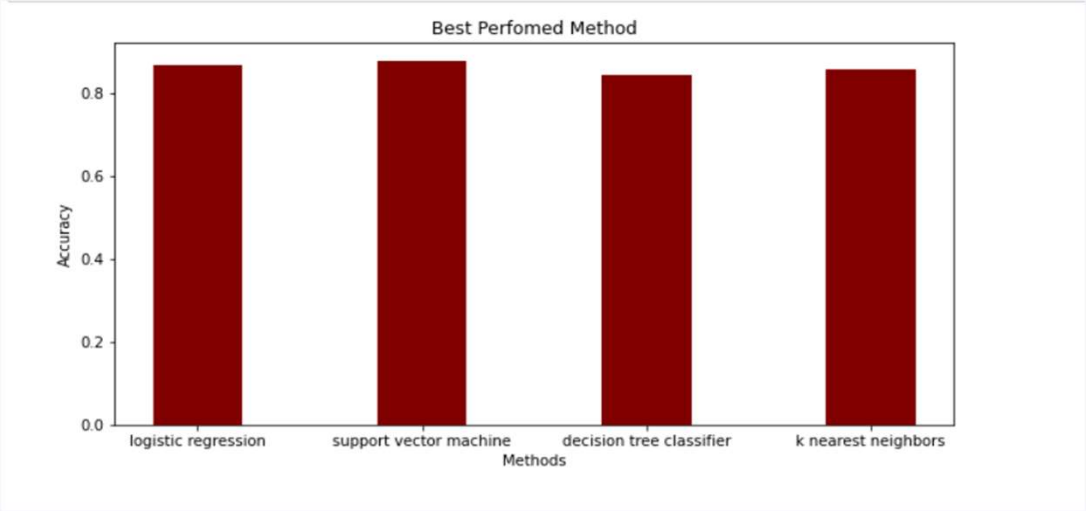
Scatter plot accounts for booster version category in color and number of launches in point size.

In the range 0-7500,there are two failed landings with payloads of zero kg.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

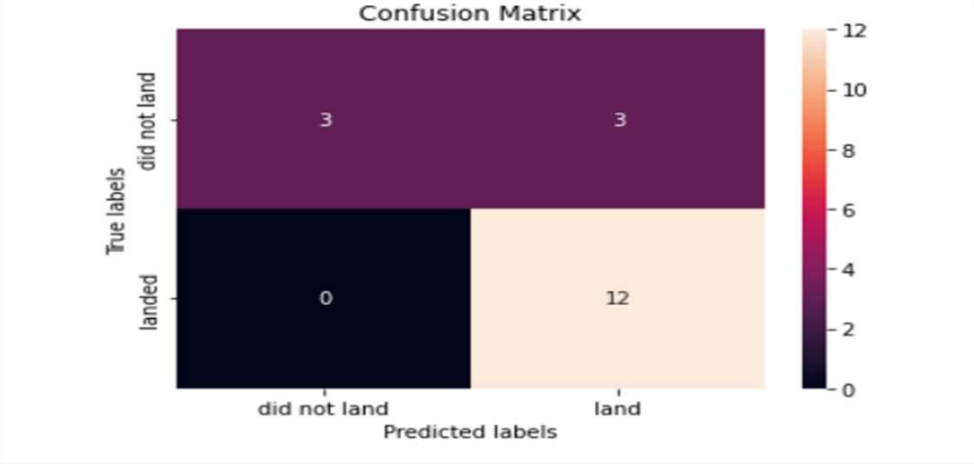


The models had virtually similar accuracy on the test set at 83.33% accuracy.

It should be noted that the test size is small at sample size of 18 only.

This can cause variance in accuracy results.Hence, we need more data to determine the best model.

Confusion Matrix



Since all models performed the same for the test set, the confusion matrix is the same across all models.

The models predicted 12 successful landings when the true label was successful landing.

The models predicted 3 unsuccessful landings when the true label was unsuccessful landing.

The models predicted 12 successful landings when the true label was unsuccessful landing.

(false positives)

Our models over predicted successful landings.

Conclusions

- Our task was to build a machine learning model for Space Y who wants to bid against Space X.
- The goal was to predict when Stage 1 will successfully land to save ~$100 million USD.
- Used data from a public SpaceX API and web scraping SpaceX Wikipedia page.
- Created data labels and stored data into a DB2 SQL database.
- Created a dashboard for visualization.
- Created a machine learning model with accuracy 83%
- Allon Mask of Space Y can use this model to predict with relatively high accuracy whether a launch will have a successful stage 1 landing before launch to determine whether the launch should be made or not.
- More data should be collected to better determine the best machine learning model and improve the accuracy.

Appendix

Github repository url:

https://github.com/shibal-7/Applied-Data-Science-Capstone

SpaceX url:

`https://api.spacexdata.com/v4/launches/past`

SpaceX data (Wikipedia):

```
"https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_
and_Falcon_Heavy_launches&oldid=1027686922"
```

Thank you!