IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Shrikrishna Bhagirath Rajule

7th April 2022

Outline

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

Executive Summary

- To predict the cost of SpaceX launch, we collected the past launch data from the SpaceX APIs and by web scrapping through the internet, performed data wrangling to improve the dataset. Performed EDA, using basic statistical analysis, data visualizations and SQL, to understand and get initial insights from the dataset. Further performed interactive visual analytics by making a dashboard using Plotly and Dash. In the end, performed predictive analysis using machine learning classification algorithms such as Logistic Regression, KNN, SVM and Decision Trees.
- EDA & interactive visual analytics gave us insights on basic structure, type, features and characteristics of the dataset and the relationship between features and features & success rate of SpaceX launches. Through predictive analysis we were able to build the best possible classification models for the engineered dataset. The best performing models turned out to be the SVM and KNN with accuracy of 83.3% each.

Introduction

- Space is becoming more accessible to humans than ever before. Commercial space flights are becoming affordable or at least cheaper than before. Thanks to Elon Musk's SpaceX company who is trying to innovate and make it possible to launch a rocket to space with cost of flight way much cheaper than the rocket launches provided by other space companies like Blue Origin, Virgin Galactic, etc. They are able to make such cheaper rockets due to the fact that 1st stage and other few part of the SpaceX's Falcon 9 rocket is recovered and reused.
- As a Data Scientist for the company SpaceY, I would like to predict if the future launches of SpaceX's Falcon 9 rockets successfully recovers the 1st stage and to see if the whole mission is successful and to know the cost of the launch. Hence , use this information to have some vantage point over SpaceX, to direct & attract customers to SpaceY.

Section 1

Methodology

Methodology

Executive Summary

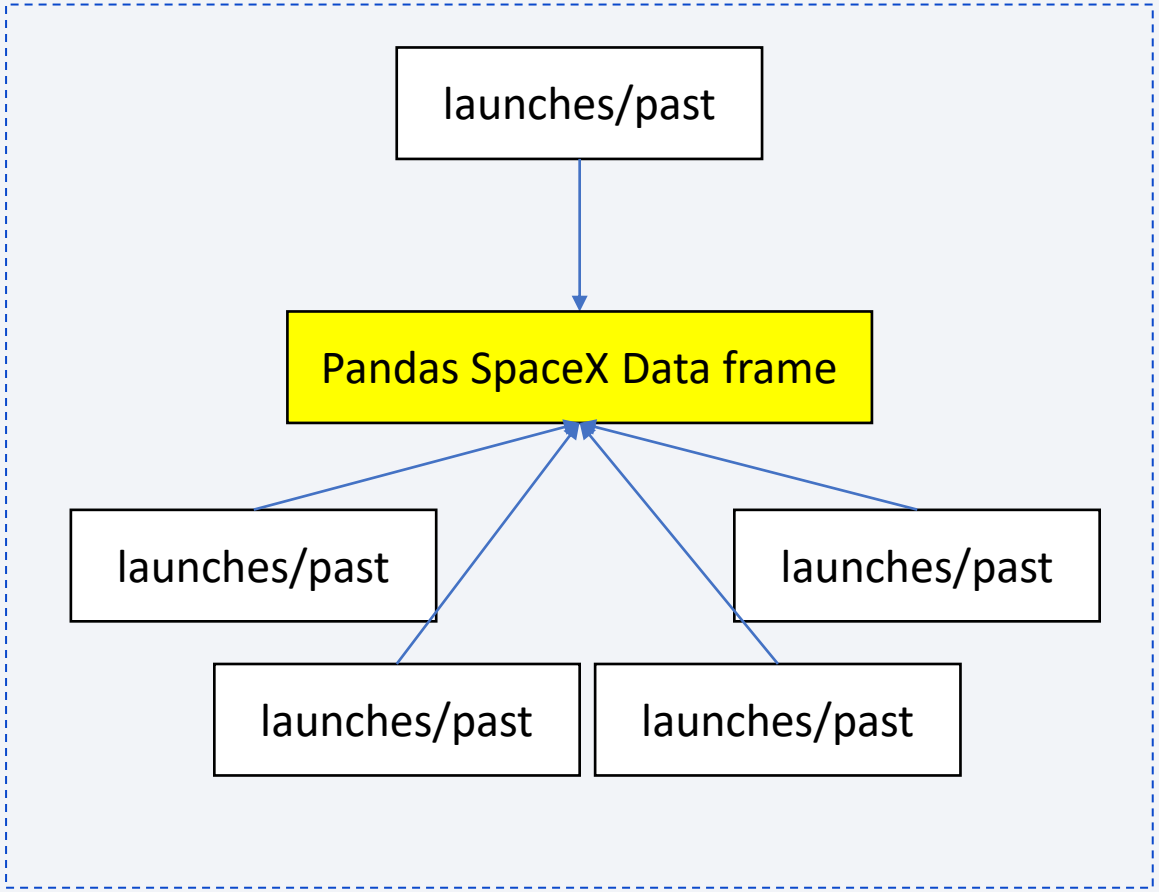
- Data collection methodology:
  - Using SpaceX API and web scrapping through the internet (Wikipedia).
- Perform data wrangling
  - Dealing with missing values. Replacing NAN values of the feature “PayloadMass” with mean.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Normalizing, splitting dataset into Training and Test, using Grid Search Cross Validation method to train, validate and find the best hyperparameters for the model, then using Test set to find the accuracy of the model.

Data Collection

- Collection of the dataset were mainly done using two sources, SpaceX APIs and Wikipedia by performing web scrapping.
- **SpaceX APIs** – Using “past launches” API we get the dataset and then using “payloads”, “rockets”, “launchpads” and “cores” API we substitute the IDs with names to “past launches” data frame.
- **Web Scrapping** – Using “beautifulsoup” python library to scrape data from web. Parsing the request response to get the data from tables of the Wikipedia page of SpaceX launches.

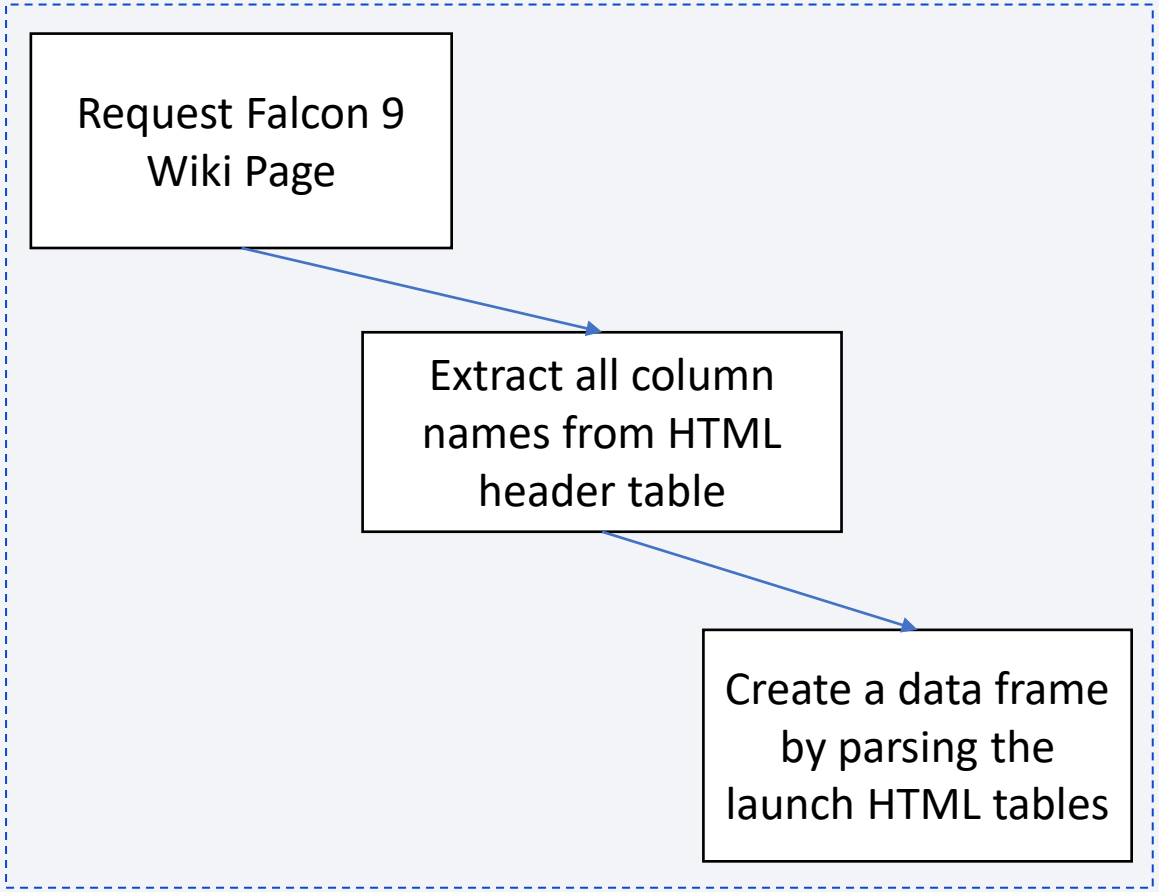
Data Collection – SpaceX API

- API - https://api.spacexdata.com/v4/
- Past launches API path - “launches/past”
- Auxiliary API paths - “rockets/”, “launchpads/”, “payloads/”, “cores/”
- Request to SpaceX APIs and clean the requested dataset.
- [Data Collection SpaceX API Github Link]



Data Collection - Scraping

- Extract a Falcon 9 launch records HTML table from Wikipedia
- Parse the table and convert it into a Pandas data frame
- Data Collection Web Scrapping Github Link



Data Wrangling

- Found the columns/features with null/missing values.
- “PayloadMass” – had 5 null entries, replaced the missing values with the mean of the whole column.
- “LandingPad” – had 26 null entries, the null entries in this column infers that landing pads were not used in the launch.
- Create a landing outcome label using Outcome column.
- Exported the dataset to CSV after the data wrangling process.
- Data Wrangling Github Link – 1
- Data Wrangling Github Link – 2

EDA with Data Visualization

- FlightNumer vs PayloadMass – Flight number increases, higher payload flight's 1st stage landing success rate also increases.
- FlightNumer vs LaunchSite – Flight number increases, success rate of flights also increases at all launch sites.
- PayloadMass vs LauchSite - VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000)
- Success rate for each orbit type – ES-L1, GEO, HEO and SSO have success rate of 100%
- PayloadMass vs OrbitType - heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- SpaceX launches yearly trend – Success rate from 2013 increases drastically until 2020.
- EDA Data Visualization Github Link

EDA with SQL

- Displayed the names of the unique launch sites in the space mission
- Displayed 5 records where launch sites begin with the string 'CCA'
- Displayed the total payload mass carried by boosters launched by NASA (CRS)
- Displayed average payload mass carried by booster version F9 v1.1
- Listed the date when the first successful landing outcome in ground pad was acheived.
- Listed the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listed the total number of successful and failure mission outcomes
- Listed the names of the booster_versions which have carried the maximum payload mass. Use a subquery
- Listed the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Ranked 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
- EDA SQL Github Link

Build an Interactive Map with Folium

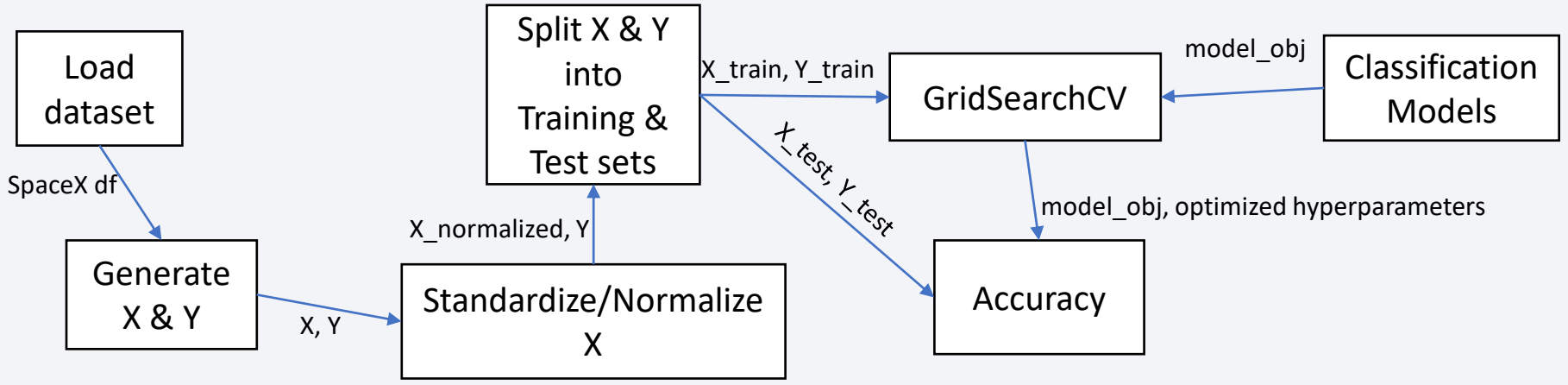
- folium.Map – to create to show the basic map with preferred coordinates and zoom level
- folium.Circle – to add highlighted circle area on specific coordinates
- folium.Marker – to dad text on specific coordinates
- folium.Popup – to add label/popup when clicked on the circle.
- folium.PolyLine – to draw a line between two specific points on the map
- Interactive Map with Folium Github Link

Build a Dashboard with Plotly Dash

- Pie Chart for successful launches from all sites together and individual – To understand which launch site has high success rate amongst each other and also with respect to itself.
- Scatter plot to check the correlation between PayloadMass and Success Rate – Custom range for PayloadMass gave us a lucid insight on how the success rate is affected due to the varying PayloadMass values. It also gave us an idea of which BoosterVersionCategory gave us the more successful launches.
- Dashboard with Plotly Dash Github Link

Predictive Analysis (Classification)

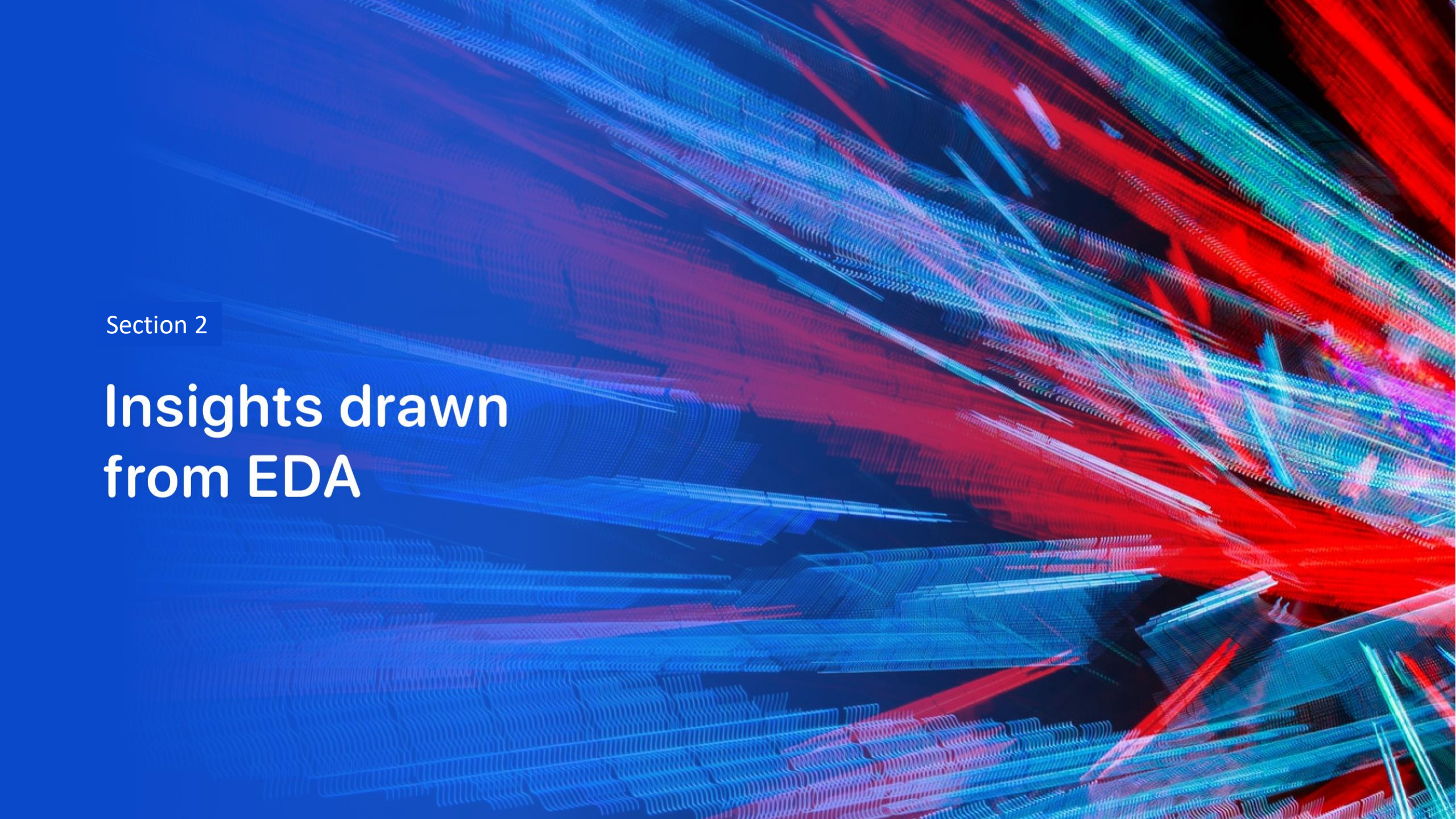
- Loaded the dataset into pandas data frame -> Created a separate numpy array of “Class” labels and assigned it to Y -> Standardized/Normalized the dataset and assigned it to X -> Split the dataset into Training (X_train, Y_train) and Test (X_test, Y_test) sets -> Used Cross Validation and Grid Search methodologies on Training dataset to train and find the best hyper parameters & weights for each of the classification model -> Calculated the score or the accuracy for each model using to Test dataset and using best generated hyperparameters & weights from GridSearchCV -> Compared the accuracy with each model and then selected the model with best accuracy.



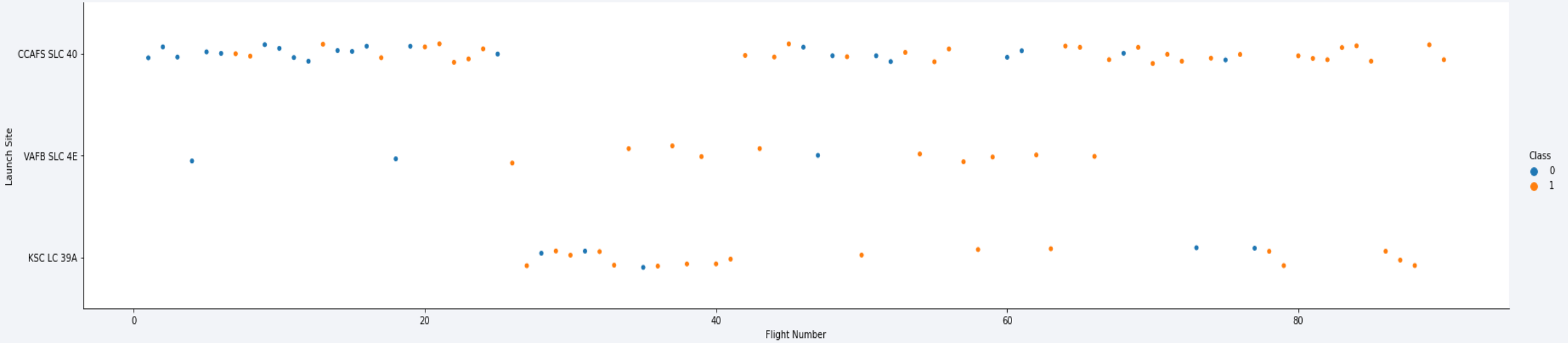
- [Predictive Analysis Classification Models Github Link]

Results

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

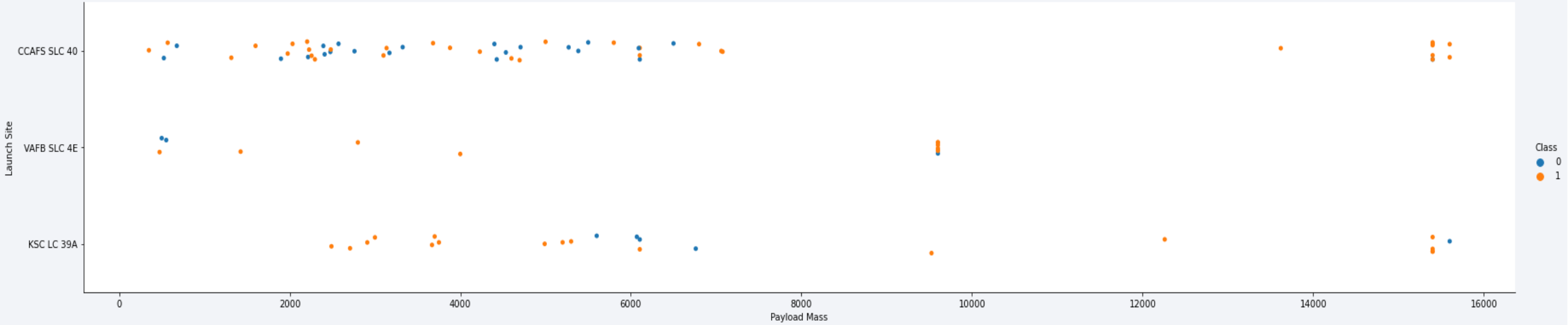

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



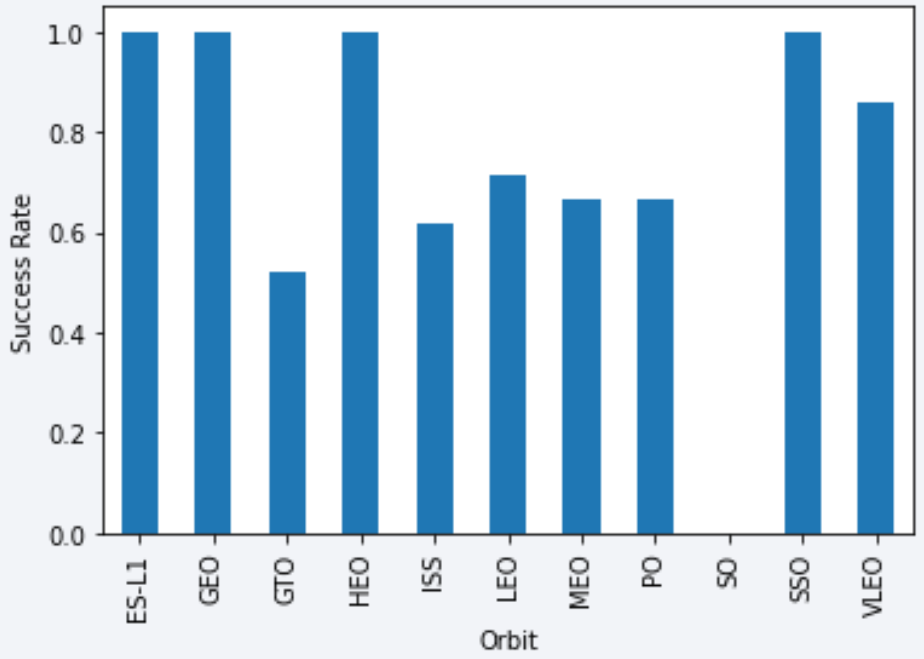
- Irrespective of launch site, we see that as the number of flights increases we see more successful flights and consistent performance.
- Most flights were launched from CCAFS SLC 40 and the least from VAFB SLC 4E

Payload vs. Launch Site



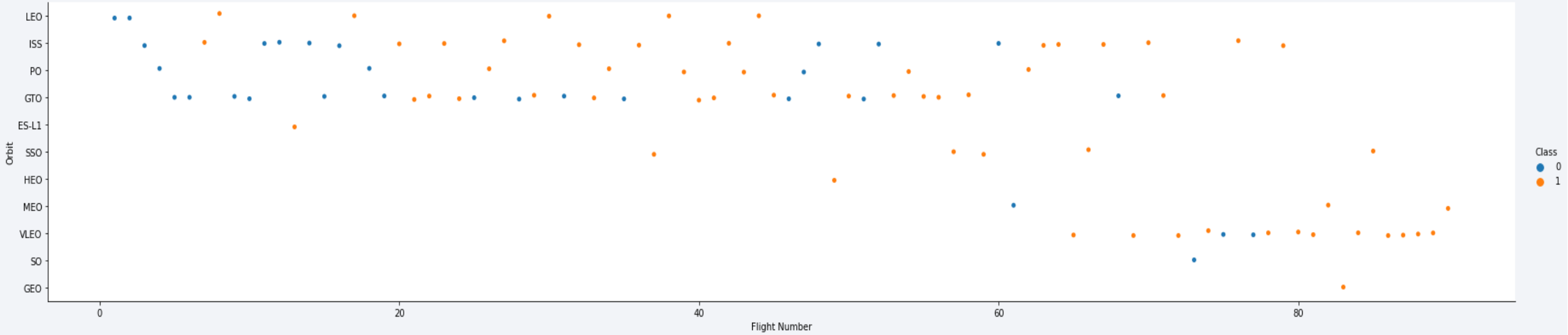
- For the VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000).

Success Rate vs. Orbit Type



- ES-L1, GEO, HEO & SSO have 100% success rate.
- While the SO have 0% success rate.

Flight Number vs. Orbit Type

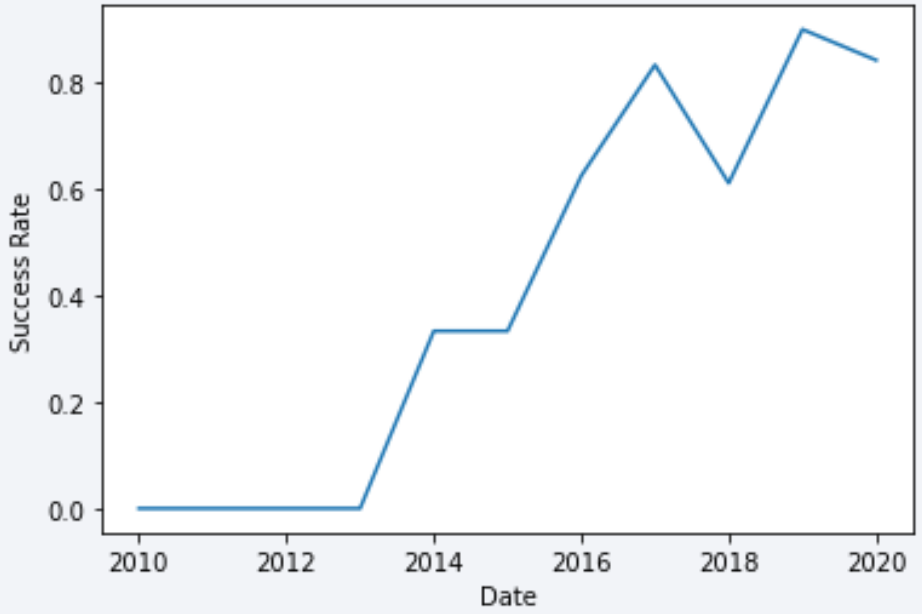


- The LEO orbit success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type

- There is no relationship for GTO and payload as there is no sign of positive or negative correlation.
- Whereas, other orbits certainly have some improvement in success rate with heavy payload deployment.

Launch Success Yearly Trend



- Success rate since 2013 kept increasing till 2020

All Launch Site Names

```
%%sql

SELECT DISTINCT(LAUNCH_SITE)
FROM SPACEXDATASET
```

```
 * ibm_db_sa://jgr76283:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/BLUDB
Done.
```

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

- To select the distinct launch sites from the database table “SPACEXDATASET”

Launch Site Names Begin with 'CCA'

```
%%sql

SELECT *
FROM SPACEXDATASET
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5
```

```
 * ibm_db_sa://jgr76283:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/BLUDB
Done.
```

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

- To select 5 rows/records where we can find launch site names beginning with “CCA”

Total Payload Mass

```
%%sql

SELECT SUM(PAYLOAD_MASS__KG_)
FROM SPACEXDATASET
WHERE CUSTOMER = 'NASA (CRS)'
```

```
 * ibm_db_sa://jgr76283:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/BLUDB
Done.
```

1
45596

- To select the total payload mass carried by boosters launched by NASA (CRS)

Average Payload Mass by F9 v1.1

```
%%sql

SELECT AVG(PAYLOAD_MASS__KG_)
FROM SPACEXDATASET
WHERE BOOSTER_VERSION = 'F9 v1.1'
```

```
 * ibm_db_sa://jgr76283:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/BLUDB
Done.
```

1
2928

- To select average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

```
%%sql

SELECT MIN(DATE)
FROM SPACEXDATASET
WHERE LANDING__OUTCOME LIKE 'Success%'
```

```
 * ibm_db_sa://jgr76283:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/BLUDB
Done.
```

1
2015-12-22

- To select the date when the first successful landing outcome in ground pad was achieved.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql

SELECT DISTINCT(BOOSTER_VERSION)
FROM SPACEXDATASET
WHERE LANDING__OUTCOME LIKE 'Success (drone%' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
```

```
 * ibm_db_sa://jgr76283:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/BLUDB
Done.
```

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

- To list 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

```
%%sql

SELECT MISSION_OUTCOME, COUNT(*)
FROM SPACEXDATASET
GROUP BY MISSION_OUTCOME
```

```
 * ibm_db_sa://jgr76283:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/BLUDB
Done.
```

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

- To calculate the total number of successful and failure mission outcome.

Boosters Carried Maximum Payload

```
%%sql

SELECT DISTINCT(BOOSTER_VERSION)
FROM SPACEXDATASET
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_)
                           FROM SPACEXDATASET)
```

```
 * ibm_db_sa://jgr76283:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/BLUDB
Done.
```

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

- To list the names of the booster which have carried the maximum payload mass

2015 Launch Records

```
%%sql

SELECT BOOSTER_VERSION, LAUNCH_SITE, LANDING__OUTCOME
FROM SPACEXDATASET
WHERE LANDING__OUTCOME LIKE 'Failure (drone%' AND YEAR(DATE) = 2015
```

```
 * ibm_db_sa://jgr76283:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/BLUDB
Done.
```

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

- To list 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

```
%%sql

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

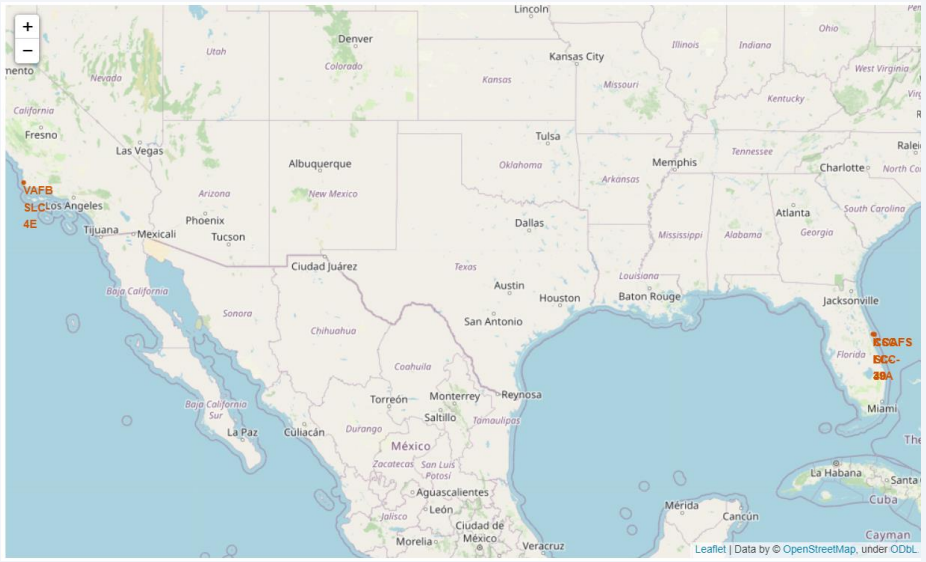
```
 * ibm_db_sa://jgr76283:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/BLUDB
Done.
```

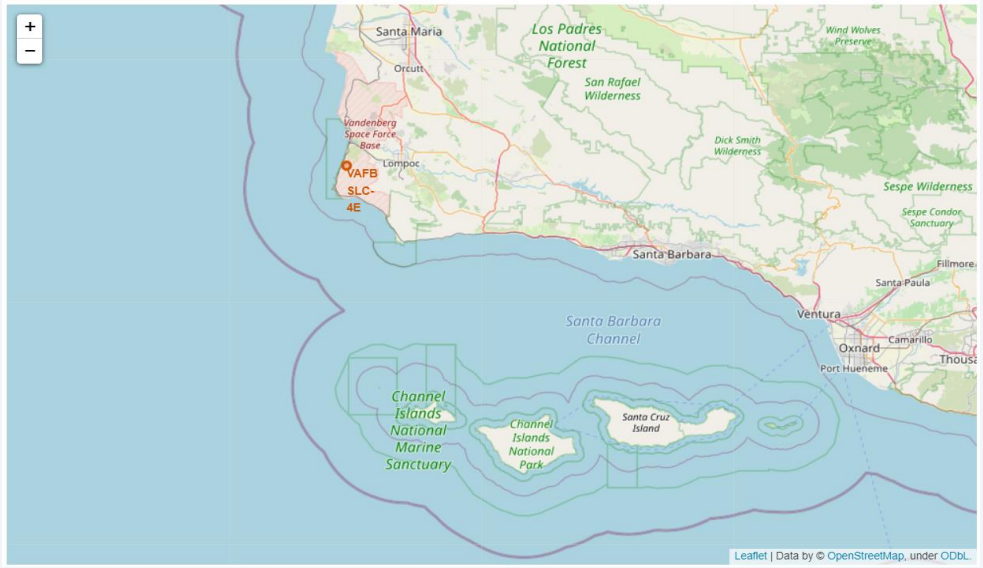
landing__outcome	total
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

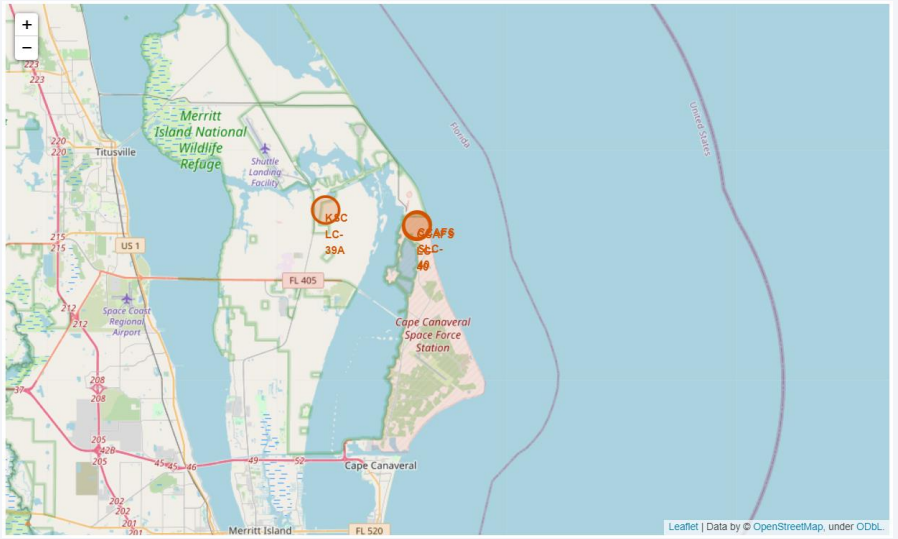
- To rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order


Section 3
Launch Sites
Proximities Analysis

Launch Sites on Map

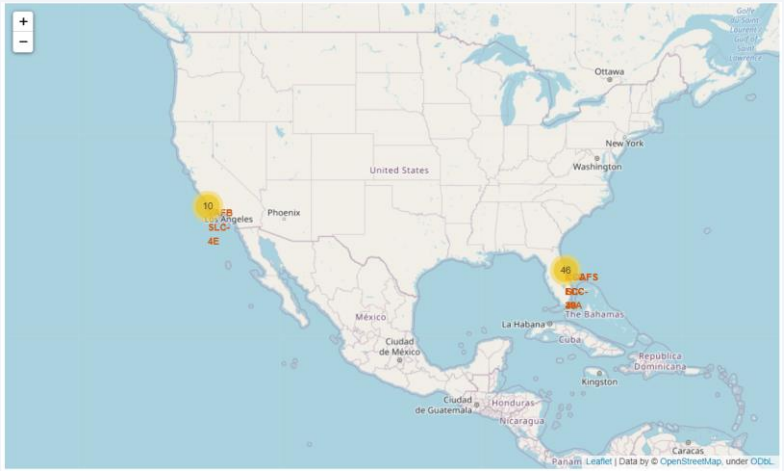


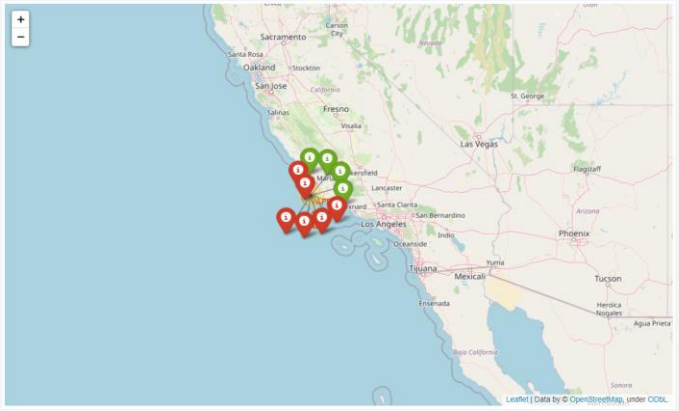


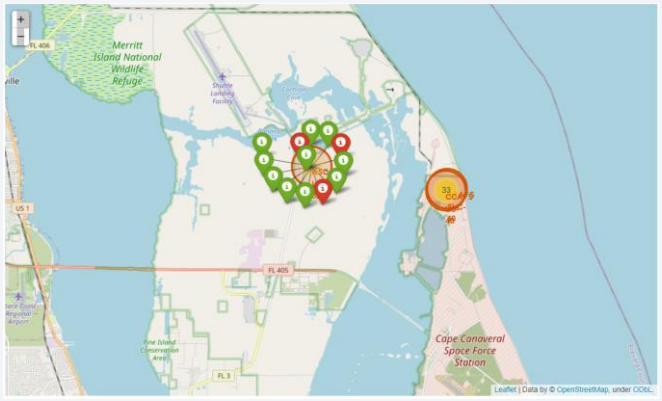


- The launch sites are near coast.
- The launch sites are also close to the equator line.

Flight Outcomes at each Launch Site

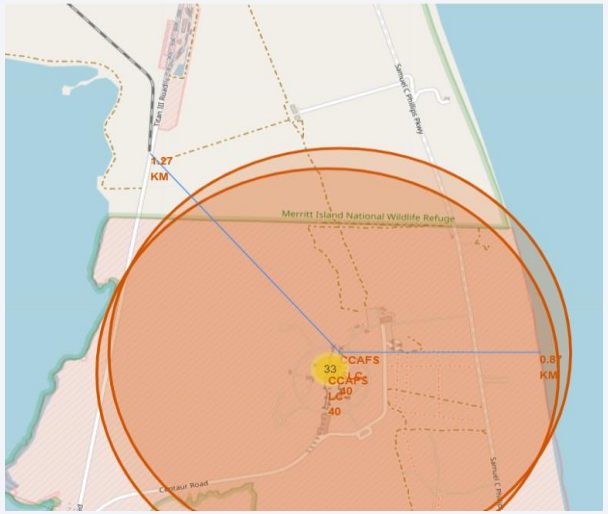


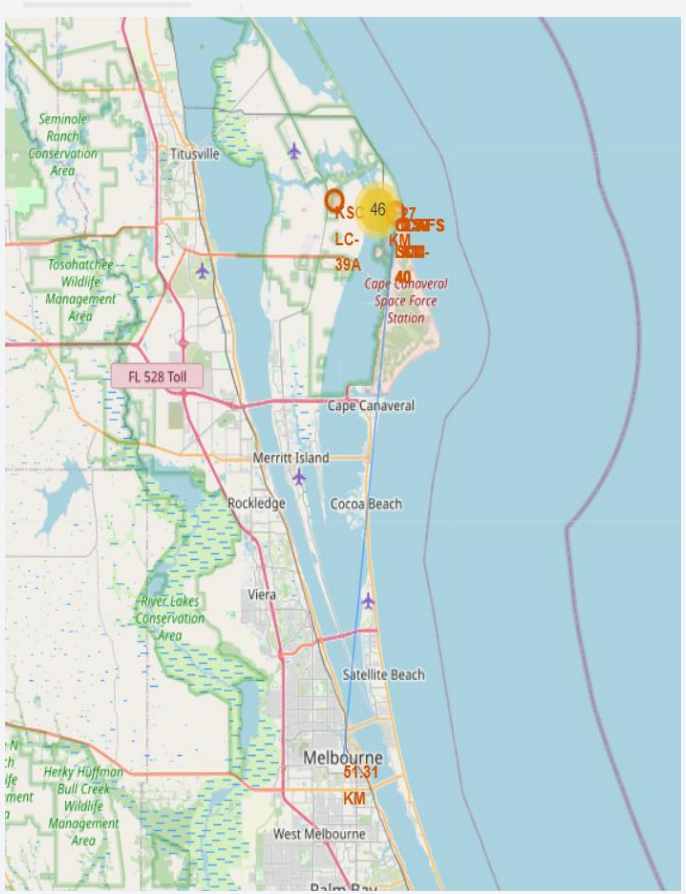




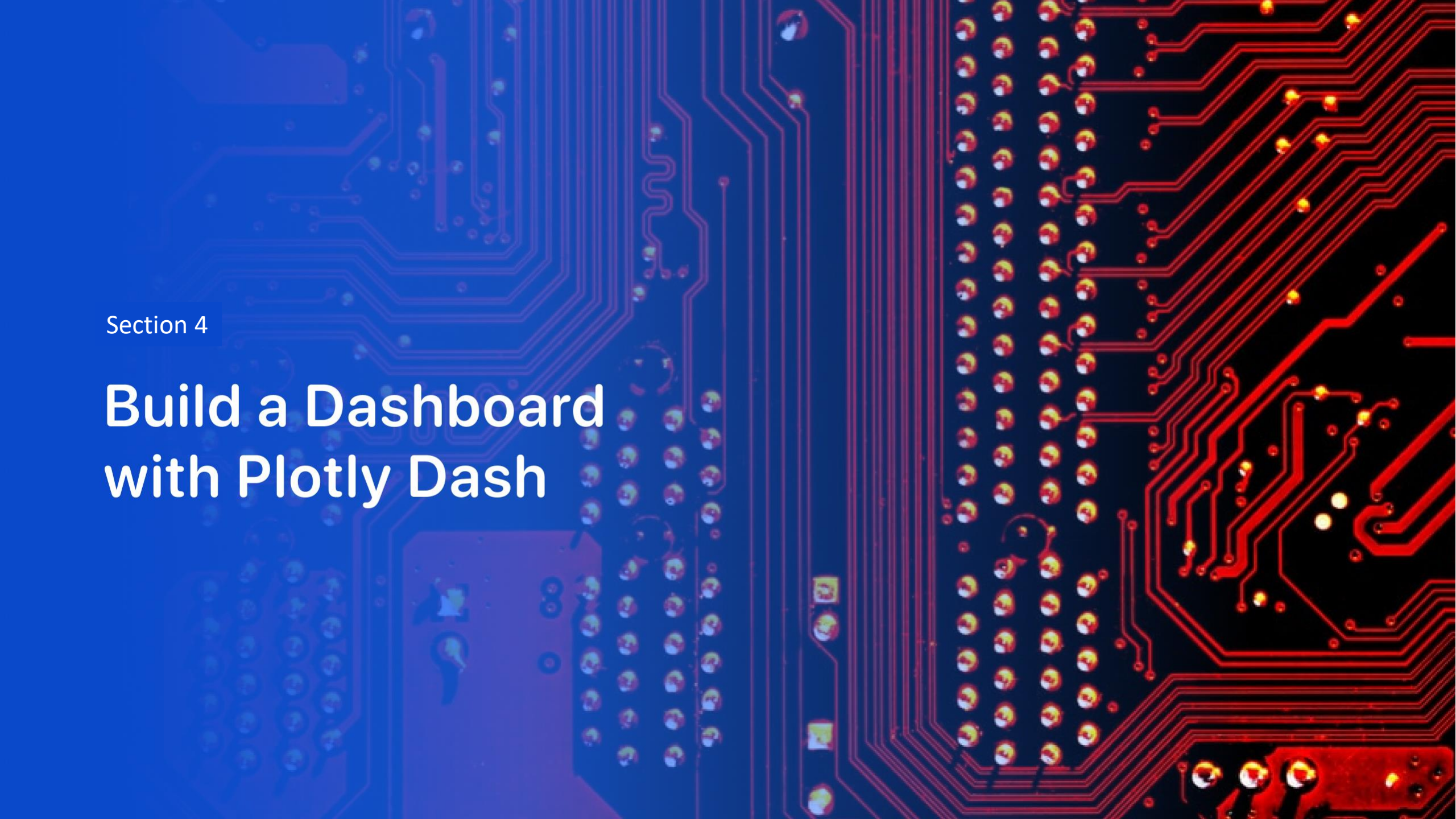
- From the map clusters, the launch site with high success rate is seen to be KSC LC - 39A

Launch Site Proximities

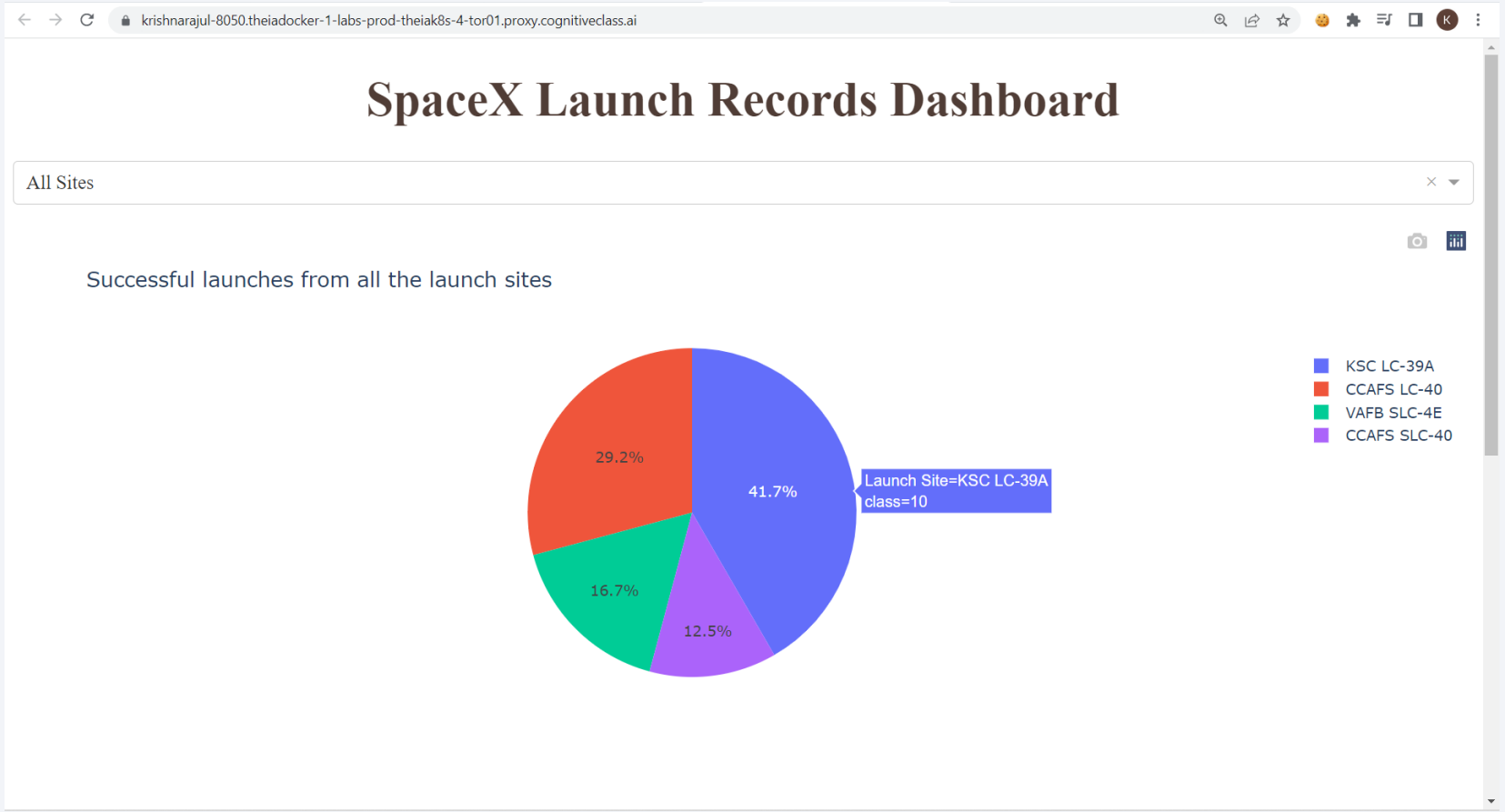




- Coastline, Railway and Highways are very near to the launch sites.
- Whereas, cities are much farther away from the launch sites.

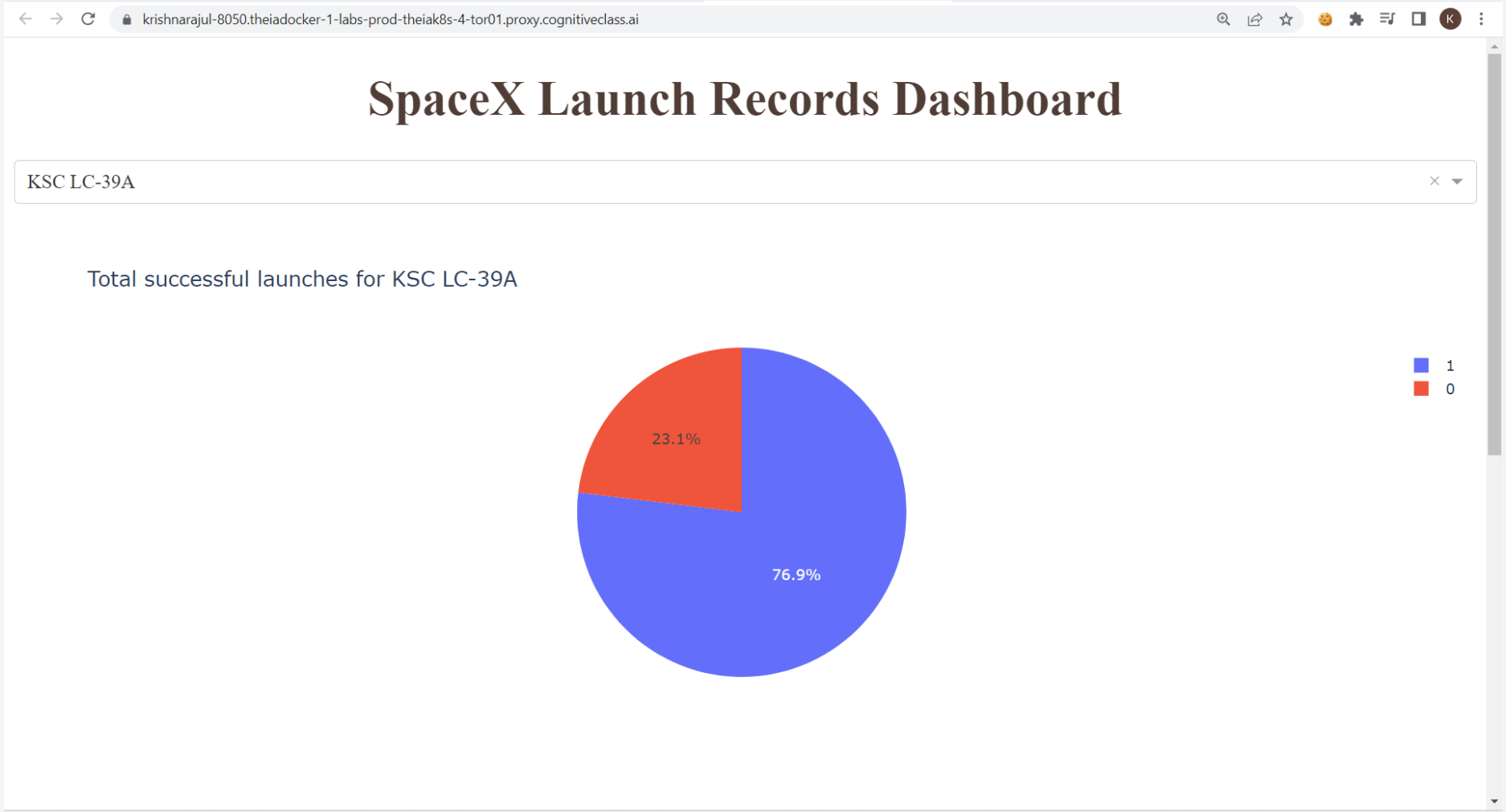

Section 4
Build a Dashboard with Plotly Dash

Launch Success Count for all Sites



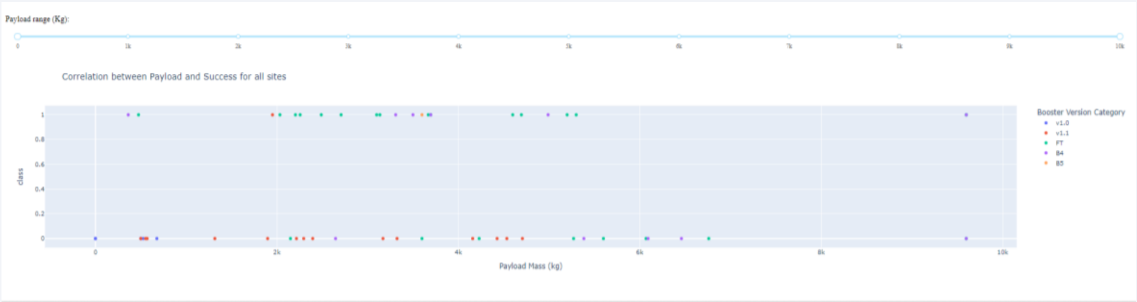
- KSC LC-39A has the most successful launches compared to other considered in our data.
- Around 41.7% of all successful launches, were taken place from the KSC LC-39A launch site.
- The least between all of them being from CCAFS SLC-40 at 12.5%

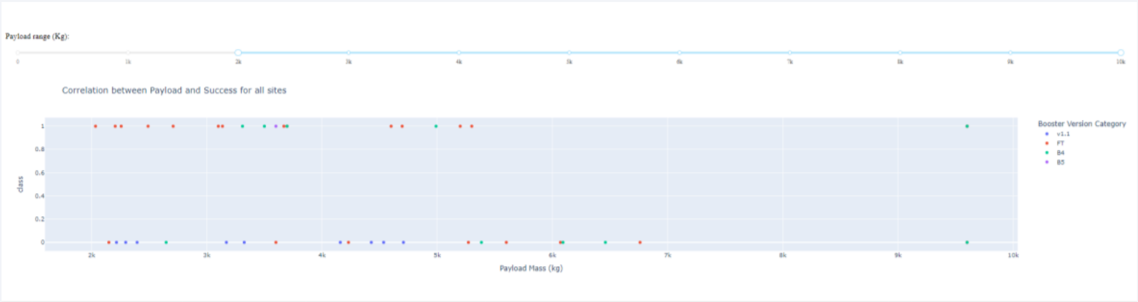
Highest Success Rate at a Launch Site

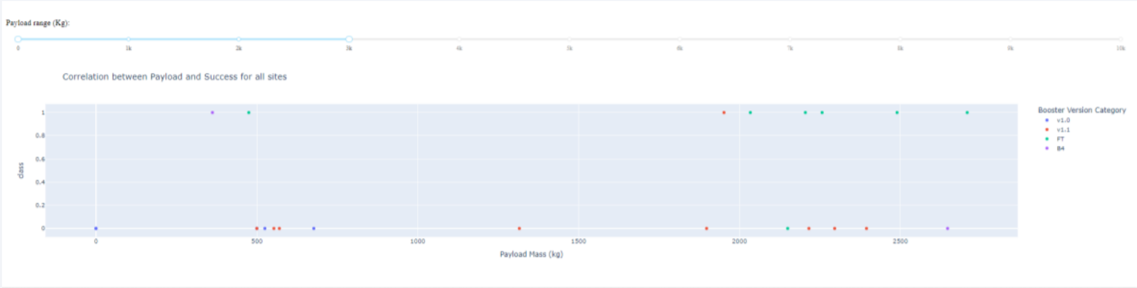


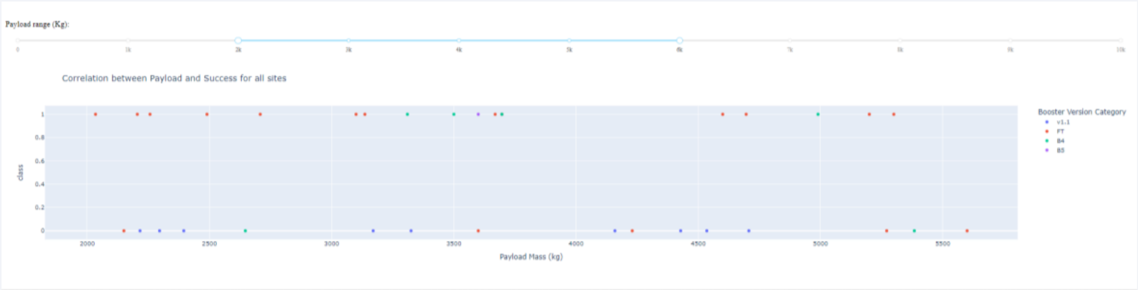
- KSC LC-39A has the highest success rate with 76.9% successful launches against 23.1% being unsuccessful.

Payload vs Booster Version vs Outcomes







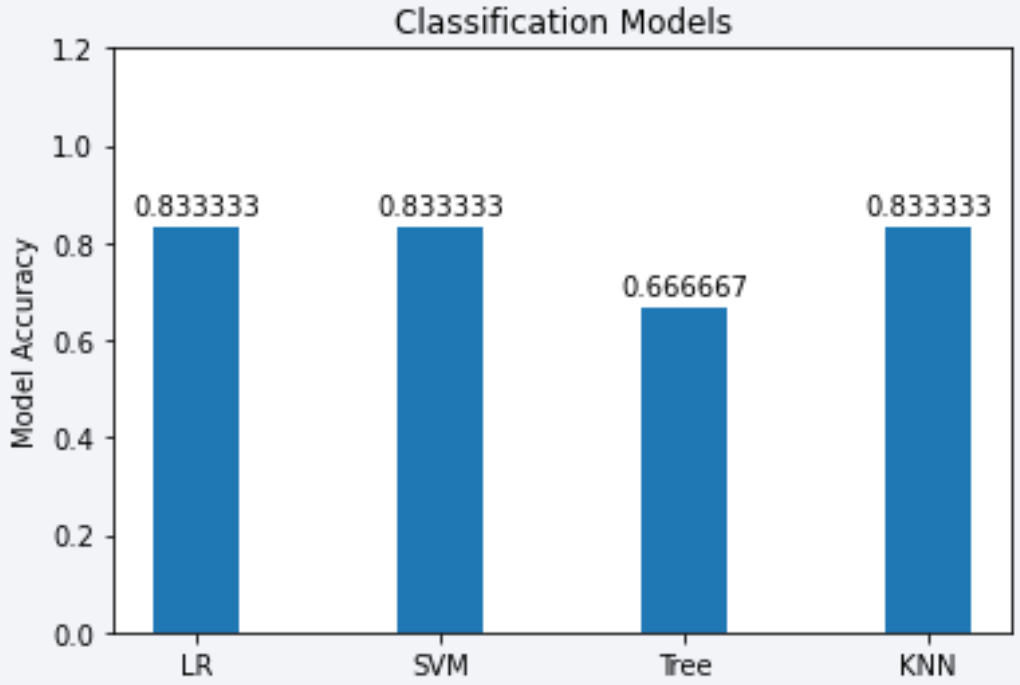


- Launches outcomes are positive or successful for payload range between 2k to 5.6k.
- The least successful launch outcomes are 0-2k and post 6k.
- The booster version category mostly used during successful launches is FT.

Section 5

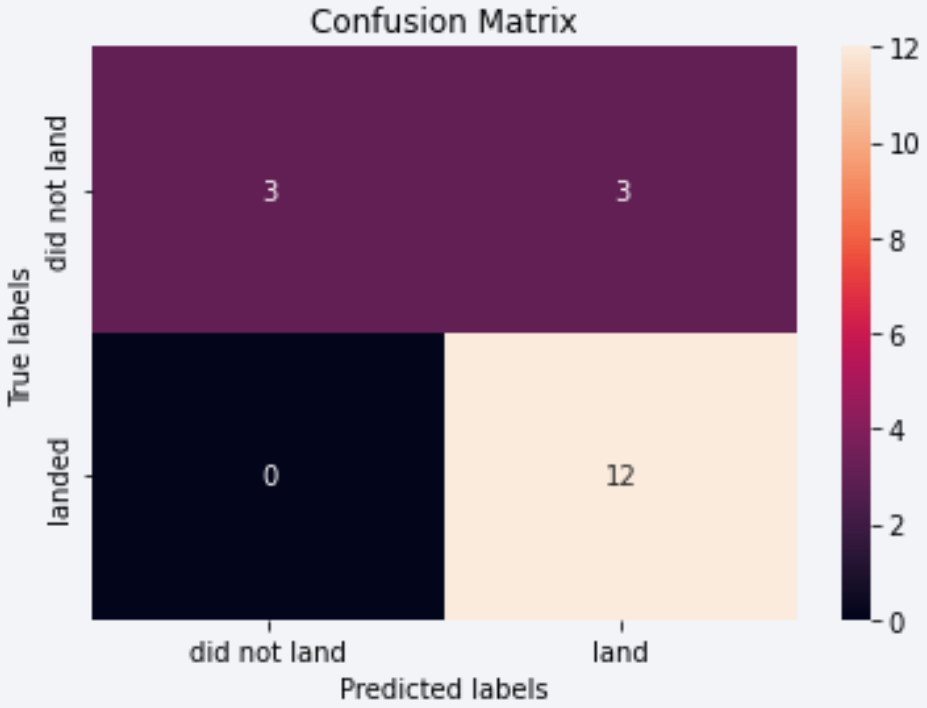
Predictive Analysis (Classification)

Classification Accuracy



- SVM, KNN and Logistic Regression models have accuracy of 83.3%.
- Therefore these models are the best to predict the outcome and to predict the cost of the launch.

Confusion Matrix



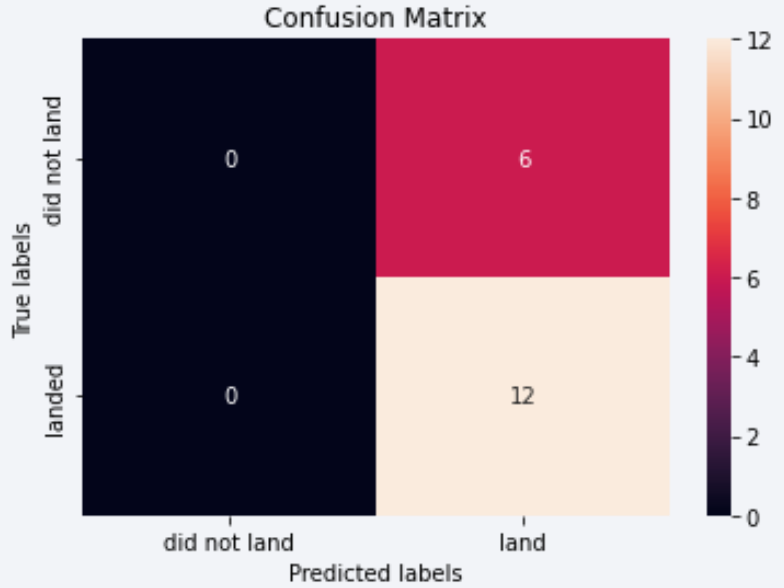
- All the 12 “landed” True labels are predicted correctly to be landed.
- But for the 6 “did not land” True label the model has predict 3 correctly to be “did not land” and 3 incorrectly to be “landed”.
- But this is the best possible accuracy and confusion matrix can be trained with the current dataset and model parameters.

Conclusions

- Data was collected from SpaceX APIs and Wikipedia
- Data Wrangling, EDA and Feature Extraction operations were performed.
- EDA with SQL and Visualization techniques showed us relationship or correlation between different features in the dataset.
- Using Plotly, Dash and Folium we got valuable insights of the dataset.
- Post Pre processing, we built classification models which gave us good accuracy to predict the outcome or cost of the SpaceX launch.

Appendix

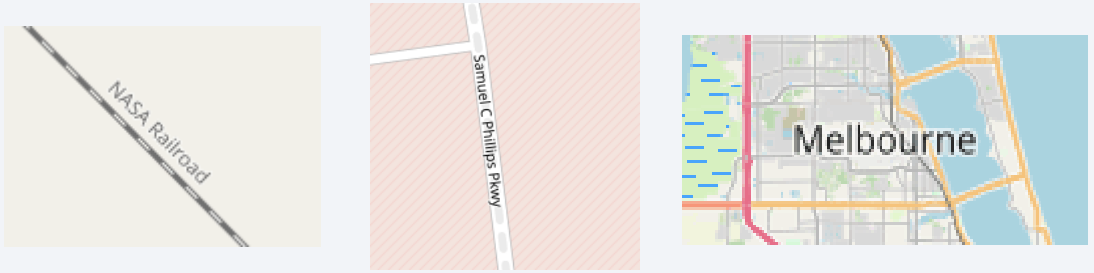
SpaceX Falcon9 Successful and Unsuccessful Landing



Decision Tree Confusion Matrix

	FlightNumber	PayloadMass	Flights	GridFins	Reused	Legs	Block	ReusedCount	Orbit_ES-L1	Orbit_GEO	...	Serial_B1048	Serial_B1049	Seri
0	1.0	6104.959412	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.0	0.0	0.0
1	2.0	525.000000	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.0	0.0	0.0
2	3.0	677.000000	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.0	0.0	0.0
3	4.0	500.000000	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.0	0.0	0.0
4	5.0	3170.000000	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
85	86.0	15400.000000	2.0	1.0	1.0	1.0	5.0	2.0	0.0	0.0	...	0.0	0.0	0.0
86	87.0	15400.000000	3.0	1.0	1.0	1.0	5.0	2.0	0.0	0.0	...	0.0	0.0	0.0
87	88.0	15400.000000	6.0	1.0	1.0	1.0	5.0	5.0	0.0	0.0	...	0.0	0.0	0.0
88	89.0	15400.000000	3.0	1.0	1.0	1.0	5.0	2.0	0.0	0.0	...	0.0	0.0	0.0
89	90.0	3681.000000	1.0	1.0	0.0	1.0	5.0	0.0	0.0	0.0	...	0.0	0.0	0.0

Process dataset – Post one hot encoding and all numeric values



Railway, Highway & City Symbols

Thank you!