IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Marco Morettini
07.11.2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion
- Appendix

Executive Summary

- **Summary of methodologies**
  - Data collection.
  - Data wrangling.
  - Exploratory Data Analysis with Data.

- **Visualization**
  - Exploratory Data Analysis with SQL.
  - Building an interactive map with Folium.
  - Building a Dashboard with Plotly Dash.
  - Predictive analysis. (Classification)

- **Summary of all Results**
  - Exploratory Data Analysis results.
  - Interactive analytics demo in screenshots.
  - Predictive analysis results.

Introduction

Project Background and context

In this capstone, we will predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website.

Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

Questions to be answered

- How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?

- Does the rate of successful landings increase over the years?

- What is the best algorithm that can be used for binary classification in this case?

Section 1

Methodology

Methodology

- Data collection methodology:
- Describe how data was collected.
- Perform data wrangling.
- Describe how data was processed.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models.
- How to build, tune, evaluate classification models

Data Collection

Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry.

We had to use both of these data collection methods in order to get complete information about the launches detailed analysis.

Data Columns are obtained by using SpaceX REST API:

- Flight Number, Date, BoosterVersion, Payload Mass, Orbit, LaunchSite, Outcome, Flights, Grid Fins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude.

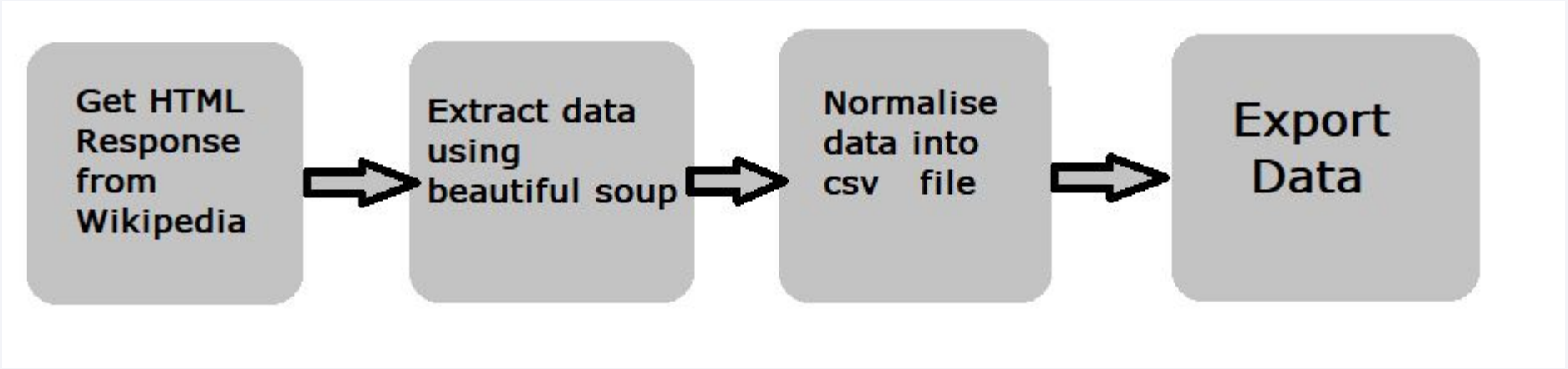
Data Columns are obtained by using Wikipedia Web Scraping:

- Flight Number, Launch site, Payload, Payload Mass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time.

Data Collection – SpaceX API

Datasets are collected from Rest SpaceX API

- The information obtained by the API are rocket, launches, payload information.
    - The Space X REST API URL is api.spacexdata.com/v4/

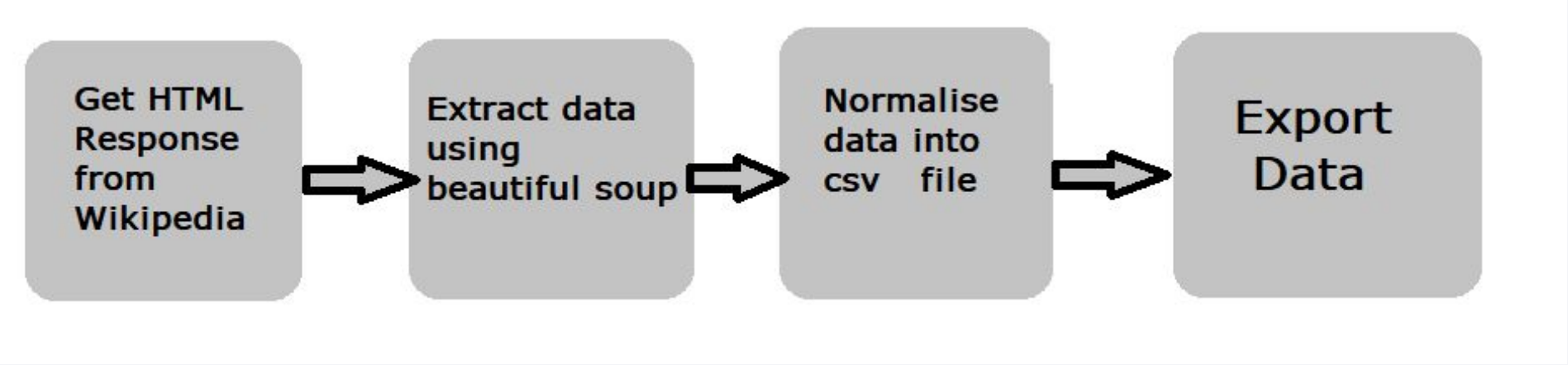


Link github Data Collection API

Data Collection - Scraping

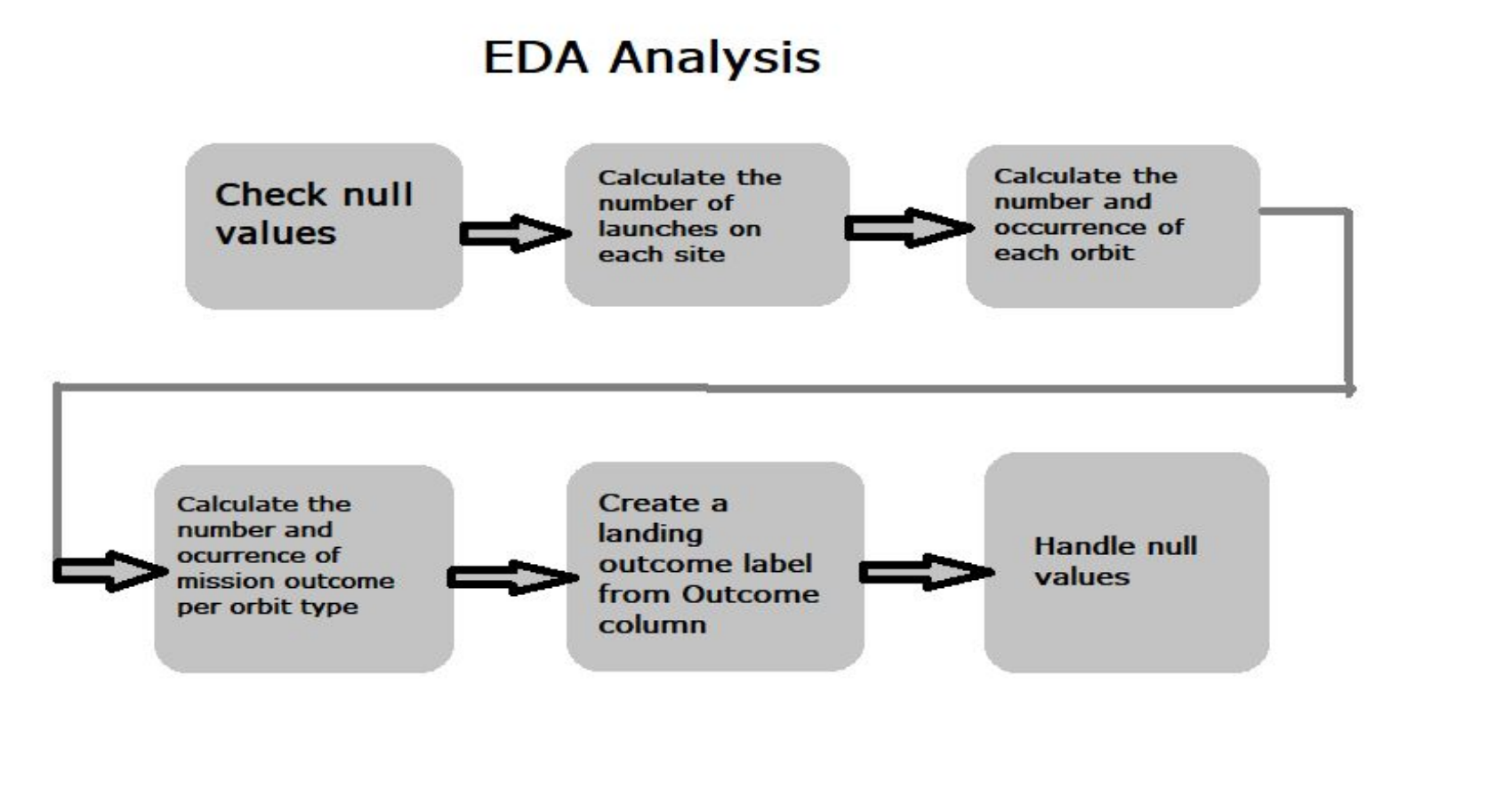
The information obtained by the web scraping of Wikipedia are launches, landing, payload information.

- URL is https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922



link github WebScraping

Data Wrangling



LInk github Data Wrangling

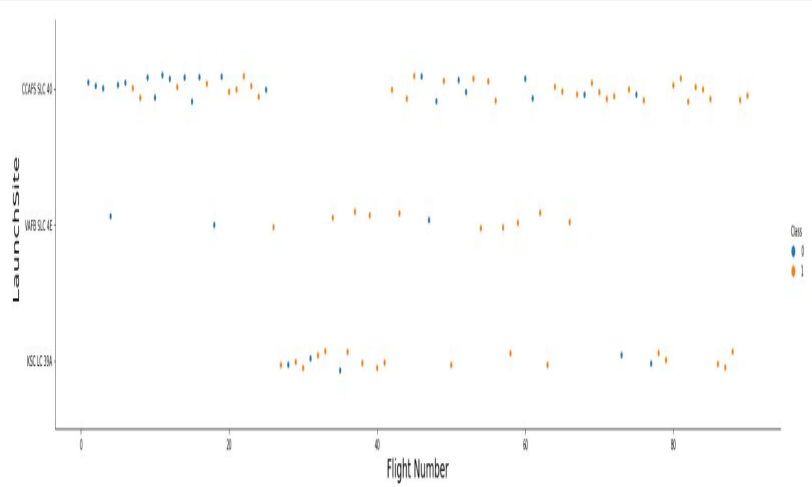
Data Wrangling

- In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean.

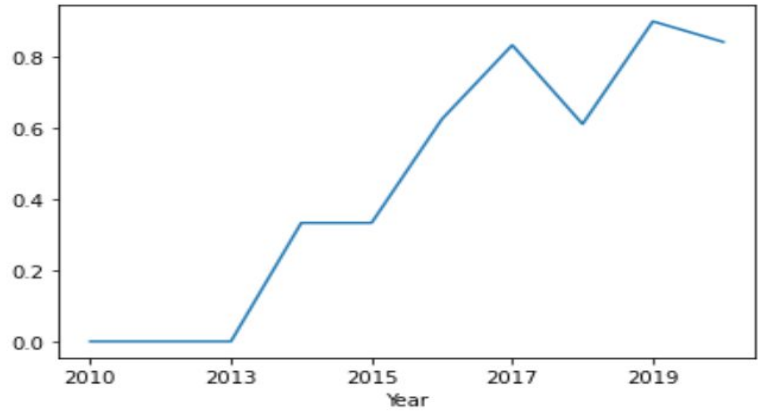
- True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad.True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship.

- We mainly convert those outcomes into Training Labels with “1” means the booster successfully landed, “0” means it was unsuccessful

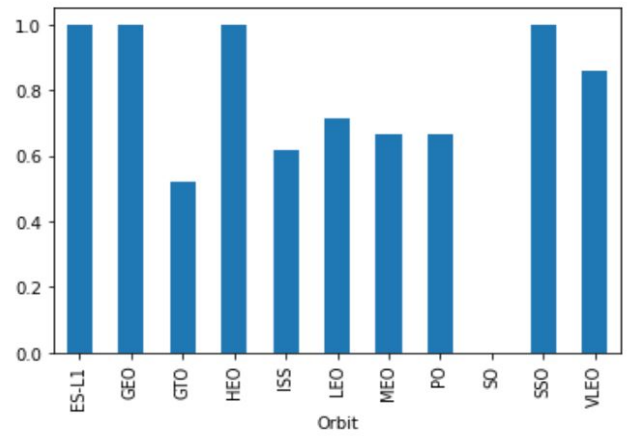
EDA with Data Visualization



Out[10]:



Out[6]:



- Scatter plots show relationship between variables. This relationship is called the correlation.
- Line graphs show data variables and their trends. Line graphs can help to show global behavior and make prediction for unseen data.
- Bar graphs show the relationship between numeric and categoric variables.

Link github EDA with Data visualization

EDA with Data Visualization

- Charts were plotted:
  -Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type and Success Rate Yearly Trend.
- Scatter plots show the relationship between variables. If a relationship exists, they could be used in machine learning model.
- Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value.
- Line charts show trends in data over time (time series)

EDA with SQL

Performed SQL queries:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string ‘CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes • Listing the names of the booster versions which have carried the maximum payload mass • Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015 • Ranking 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

Launch site names begin with `CCA`

In [19]:
```
%sql SELECT * FROM SPACEX WHERE "LAUNCH_SITE" LIKE '%CCA%' LIMIT 5
```

```
 * ibm_db_sa://pvy92609:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

Out[19]:

DATE	time_utc_	booster_version	launch_site	payload	payload_mass_kg_	orbit	customer	mission_outcome	landing_outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt
03-12-2013	22:41:00	F9 v1.1	CCAFS LC-40	SES-8	3170	GTO	SES	Success	No attempt

link github EDA with SQL

Build an Interactive Map with Folium

- Markers of all Launch Sites:

  - Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.

  - Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.

- Coloured Markers of the launch outcomes for each Launch Site:

  - Added coloured Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.

- Distances between a Launch Site to its proximities:

  - Added coloured Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City

link github Site Location with Folium

Build a Dashboard with Plotly Dash

Launch Sites Dropdown List:

- Added a dropdown list to enable Launch Site selection.

Pie Chart showing Success Launches (All Sites/Certain Site):

- Added a pie chart to show the total successful launches count for all sites and the

Success vs. Failed counts for the site, if a specific Launch Site was selected.

Slider of Payload Mass Range:

- Added a slider to select Payload range.

Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:

- Added a scatter chart to show the correlation between Payload and Launch Success.

link github Dashboard with Plotly Dash

Predictive Analysis (Classification)

- **Data preparation**
  - Load dataset
  - Normalize data
  - Split data into training and test sets.
- **Model preparation**
  - Selection of machine learning algorithms.
  - Set parameters for each algorithm to GridSearchCV
  - Training GridSearchModel models with training dataset
- **Model evaluation**
  - Get best hyperparameters for each type of model
  - Compute accuracy for each model with test dataset
  - Plot Confusion Matrix
- **Model comparison**
  - Comparison of models according to their accuracy
  - The model with the best accuracy will be chosen (see Notebook for result)

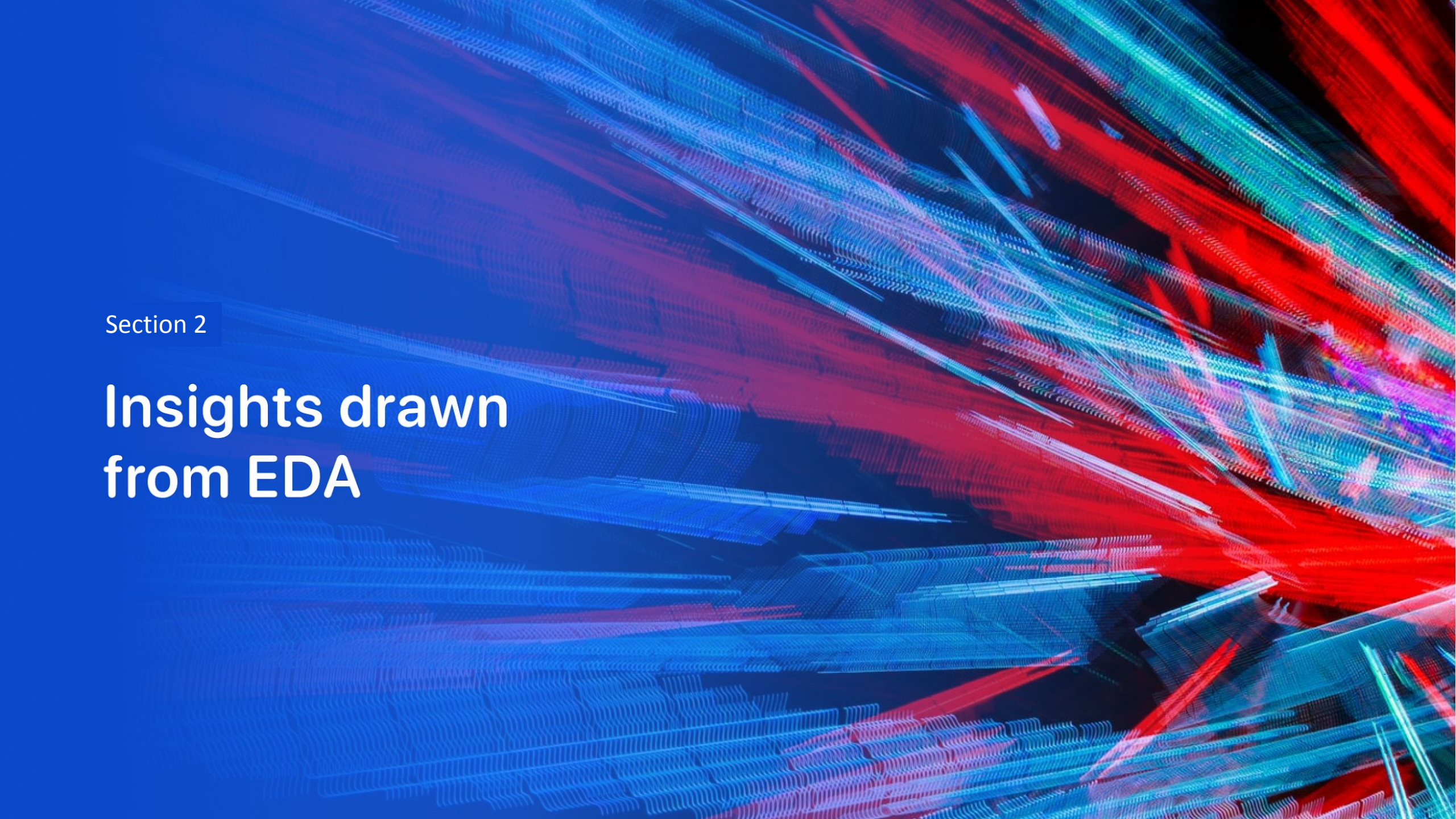
link github Machine Learning Prediction

Results

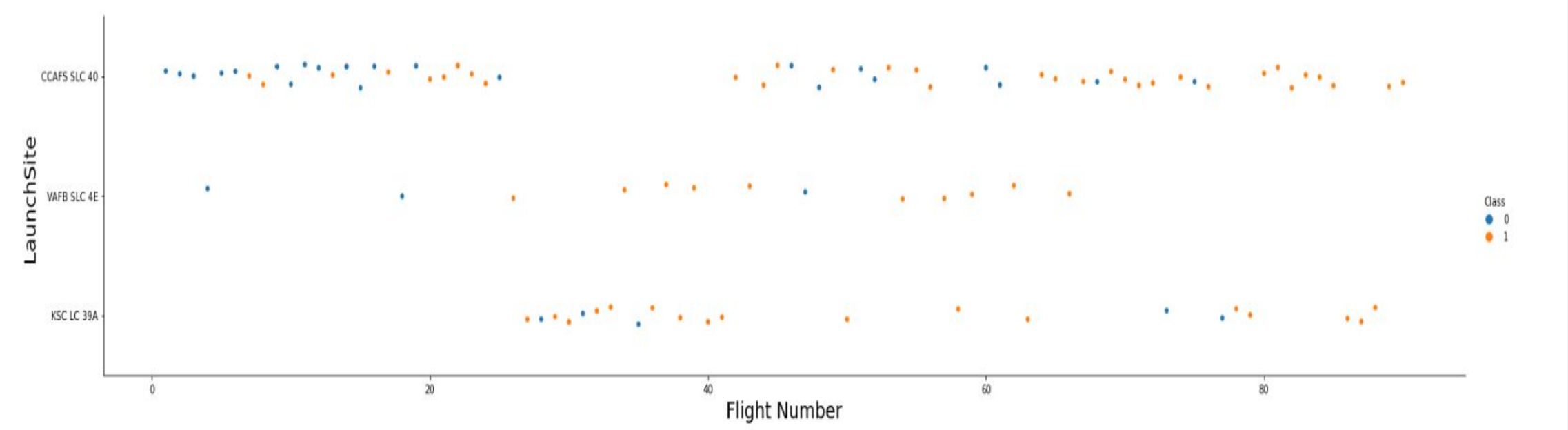
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

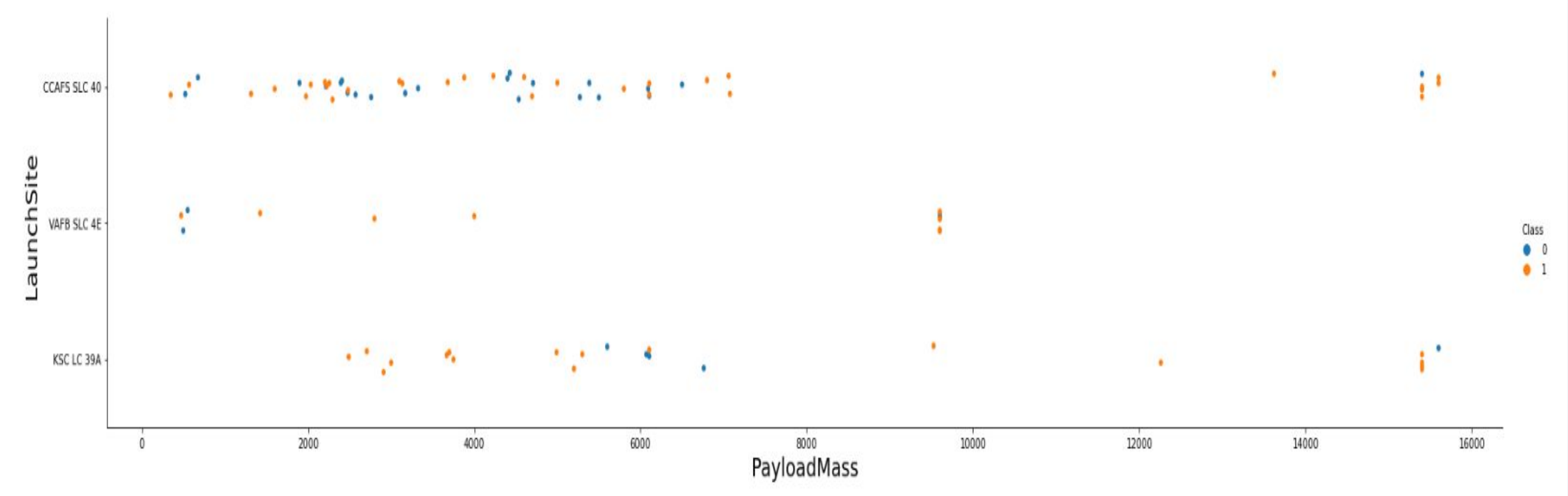


Flight Number vs. Launch Site



- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- For each site, the success rate is increasing.

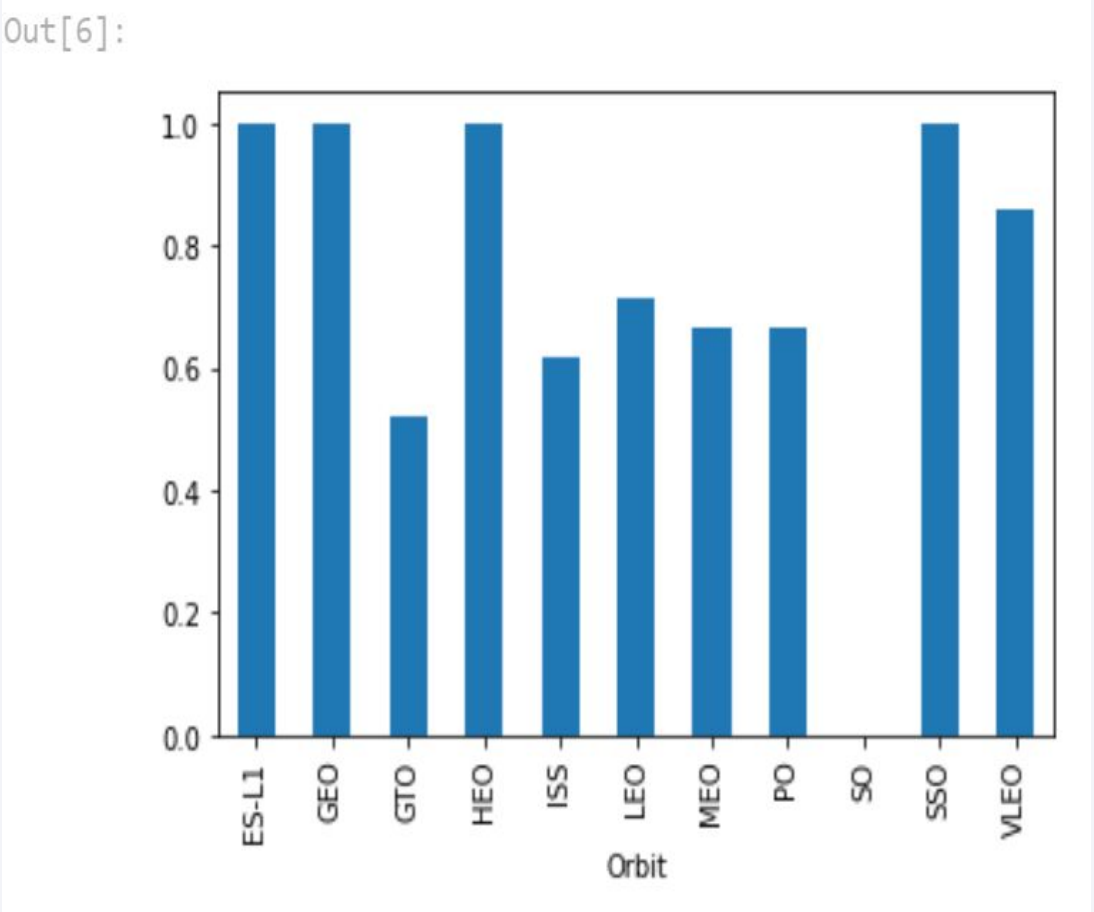
Payload vs. Launch Site



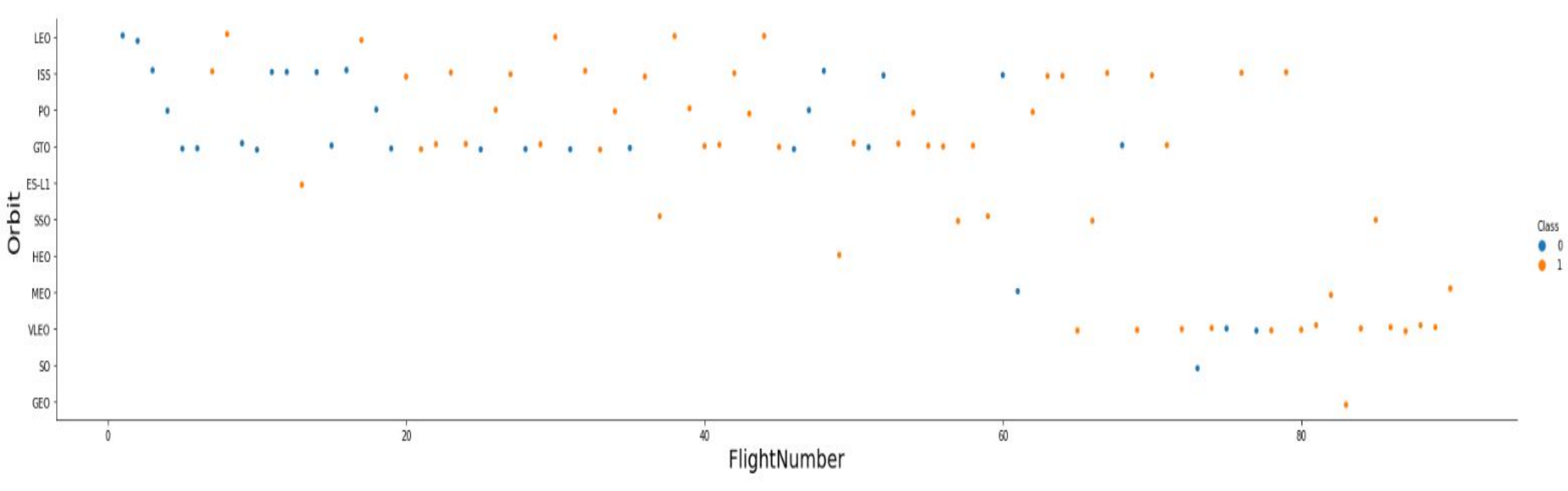
- KSC LC 39A has a 100% success rate for payload mass under 5500.
- Most of the launches with a load mass of more than 7000 kg were successful.
- For each launch site, the greater the mass of the payload, the greater the success installments.

Success Rate vs. Orbit Type

- Orbits with 100% success rate:
  - ES-L1, GEO, HEO, SSO
- Orbits with 0% success rate:
  - SO
- Orbits with success rate between 50% and 85%:
  - GTO, ISS, LEO, MEO, PO

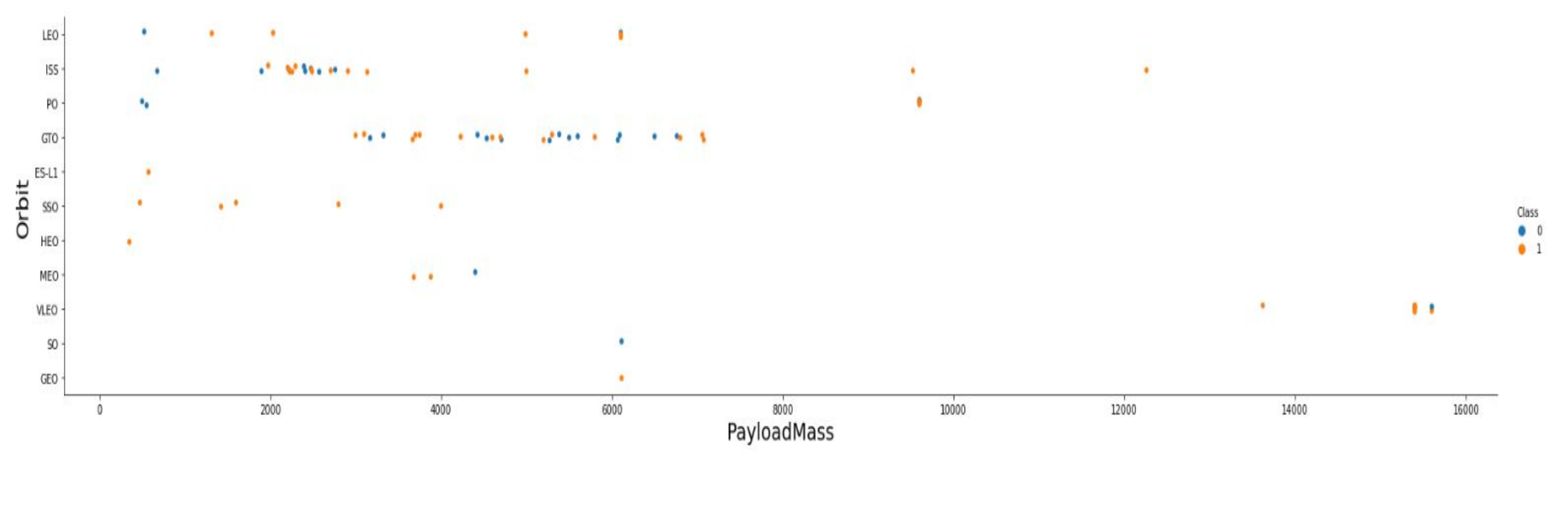


Flight Number vs. Orbit Type



- The success rate increases with the number of flights for the LEO orbit.
- No relationship between flight number when in GTO orbit.

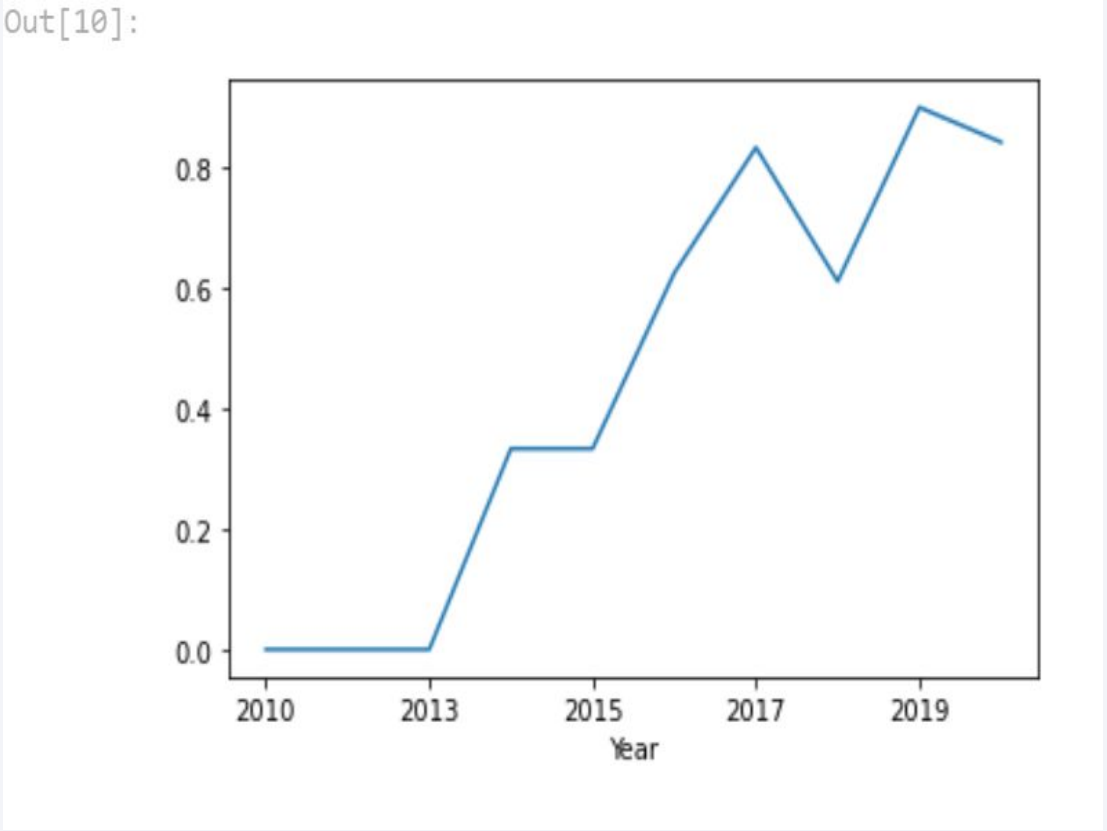
Payload vs. Orbit Type



Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

Launch Success Yearly Trend

Out[10]:



The success rate since 2013 kept increasing till 2020

All Launch Site Names

In [5]:
```
%sql SELECT DISTINCT LAUNCH_SITE as "Launch_Sites" FROM SPACEX;
```

```
 * ibm_db_sa://pvy92609:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

Out[5]:

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

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

Launch Site Names Begin with 'CCA'

In [19]:
```
%sql SELECT * FROM SPACEX WHERE "LAUNCH_SITE" LIKE '%CCA%' LIMIT 5
```

```
 * ibm_db_sa://pvy92609:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

Out[19]:

DATE	time_utc_	booster_version	launch_site	payload	payload_mass_kg_	orbit	customer	mission_outcome	landing_outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt
03-12-2013	22:41:00	F9 v1.1	CCAFS LC-40	SES-8	3170	GTO	SES	Success	No attempt

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

Total Payload Mass

```
In [18]: %sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEX WHERE "CUSTOMER" = 'NASA (CRS)'
 * ibm_db_sa://pvy92609:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
Out[18]: 1
38856
```

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

Average Payload Mass by F9 v1.1

```
In [20]: %sql SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEX WHERE "BOOSTER_VERSION" LIKE '%F9 v1.1%'
 * ibm_db_sa://pvy92609:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
Out[20]: 1
2534
```

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

First Successful Ground Landing Date

In [24]:

```
%%sql
SELECT MIN(Date)
FROM SPACEX
WHERE Landing__Outcome = 'Success (ground pad)';
```

```
 * ibm_db_sa://pvy92609:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

Out[24]:

1
01-05-2017

Listing the date when the first successful landing outcome in ground pad was achieved

Successful Drone Ship Landing with Payload between 4000 and 6000

In [26]:
```
%%sql
SELECT BOOSTER_VERSION
FROM SPACEX
WHERE LANDING__OUTCOME = 'Success (drone ship)'
    AND 4000 < PAYLOAD_MASS__KG_ < 6000;
```

```
 * ibm_db_sa://pvy92609:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

Out[26]:

booster_version
F9 FT B1021.1
F9 FT B1023.1
F9 FT B1029.2
F9 FT B1038.1
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

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

Total Number of Successful and Failure Mission Outcomes

In [27]:
```
%%sql
SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER
FROM SPACEX
GROUP BY MISSION_OUTCOME;
```

```
 * ibm_db_sa://pvy92609:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

Out[27]:

mission_outcome	total_number
Failure (in flight)	1
Success	64
Success (payload status unclear)	1
None	0

Listing the total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

In [29]:

```
%%sql
SELECT DISTINCT BOOSTER_VERSION
FROM SPACEX
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_)
    FROM SPACEX);
```

```
 * ibm_db_sa://pvy92609:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

Out[29]:

booster_version
F9 B4 B1041.2
F9 B4 B1041.1
F9 B5 B1049.2
F9 B5B1048.1
F9 FT B1036.2
F9 FT B1029.1
F9 FT B1036.1

Listing the names of the booster versions which have carried the maximum payload mass

2015 Launch Records

In [30]:
```
%%sql
SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE
FROM SPACEX
WHERE Landing__Outcome = 'Failure (drone ship)'
    AND YEAR(DATE) = 2015;
```

```
 * ibm_db_sa://pvy92609:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

Out[30]:

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

Listing the failed landing outcomes in drone ship in year 2015.

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

In [38]:

```
%%sql
SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS TOTAL_NUMBER
FROM SPACEX
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING__OUTCOME
ORDER BY TOTAL_NUMBER DESC
```

```
 * ibm_db_sa://pvy92609:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

Out[38]:

landing__outcome	total_number

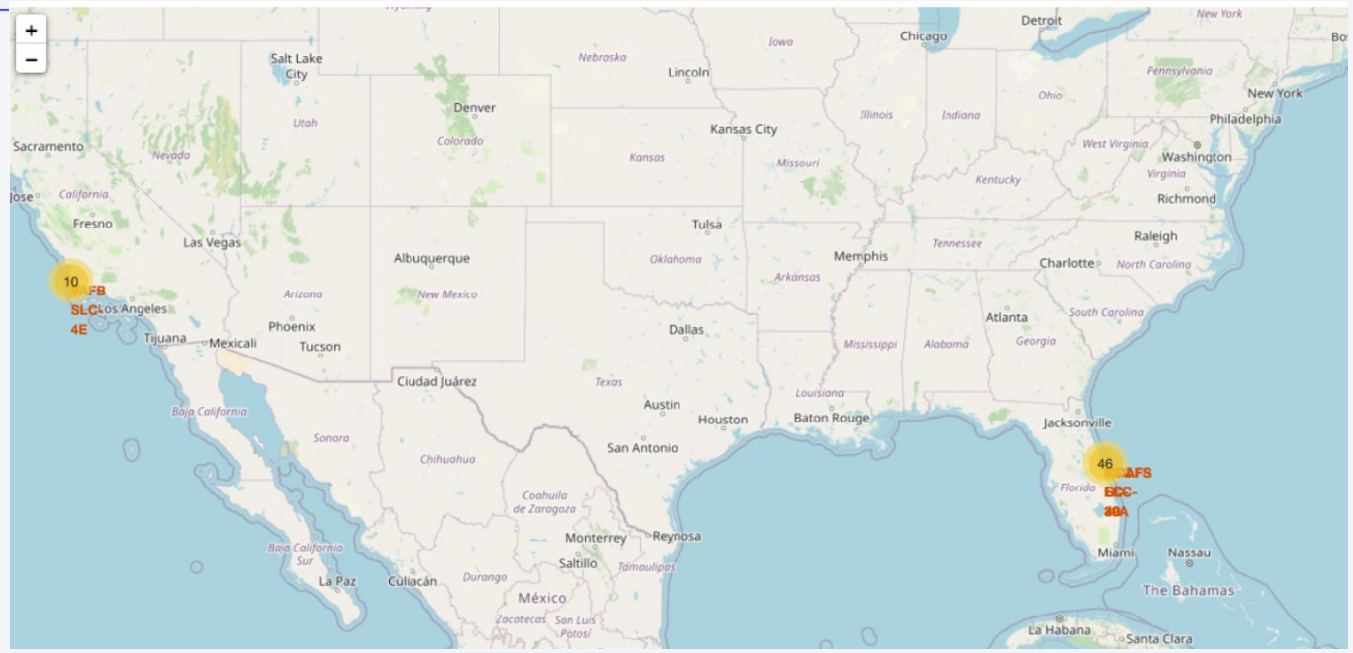
Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.


Section 3
Launch Sites
Proximities Analysis

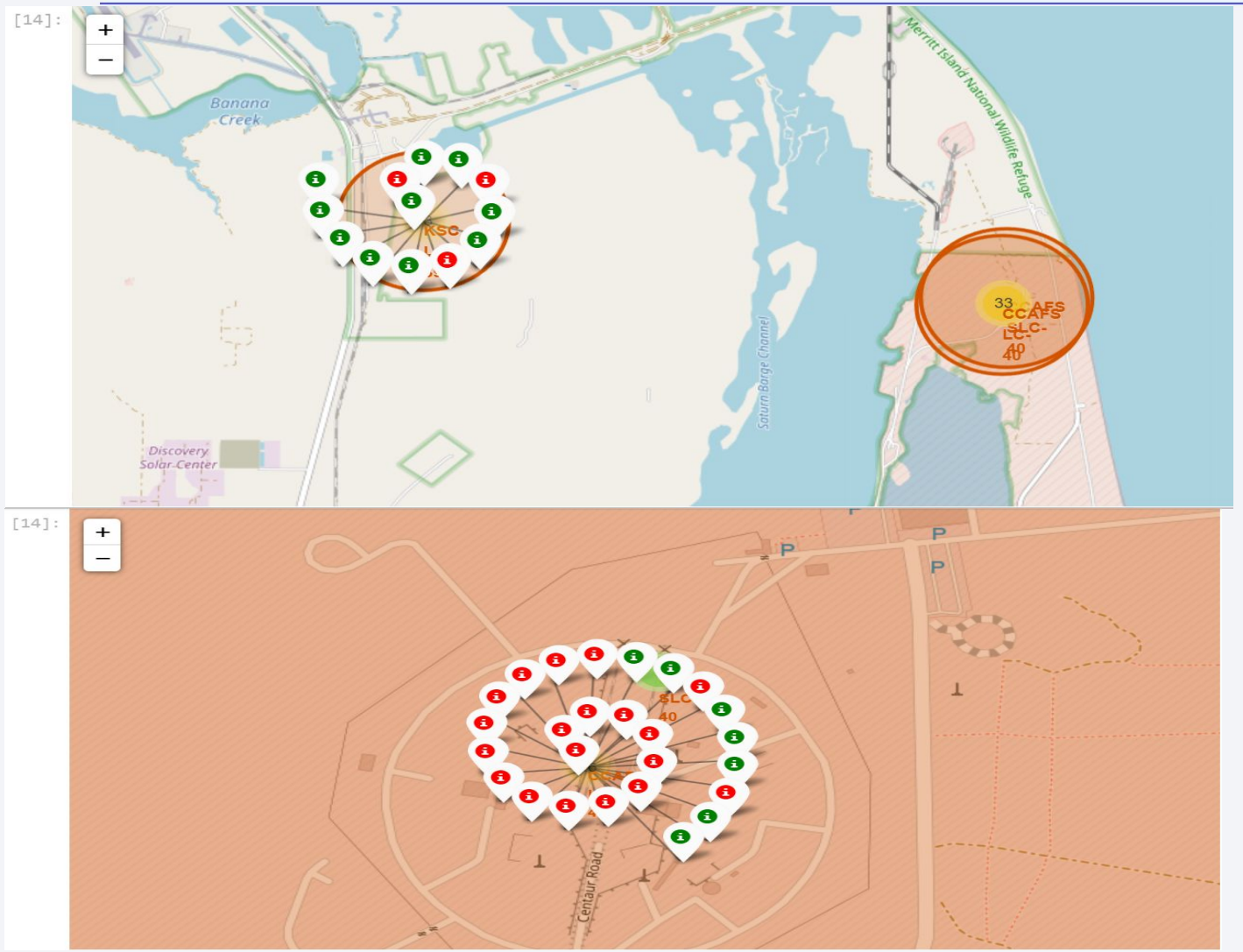
All launch sites' location markers on a global map

- Most of Launch sites are in proximity to the Equator line. The land is moving faster at the equator than any other place on the surface of the Earth. Anything on the surface of the Earth at the equator is already moving at 1670 km/hour. If a ship is launched from the equator it goes up into space, and it is also moving around the Earth at the same speed it was moving before launching. This is because of inertia. This speed will help the spacecraft keep up a good enough speed to stay in orbit.

- All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimises the risk of having any debris dropping or exploding near people



Colour-labeled launch records on the map



- From the colour-labeled markers we should be able to easily identify which launch sites have relatively high success rates. - Green Marker = Successful Launch - Red Marker = Failed Launch
- Launch Site KSC LC-39A has a very high Success Rate

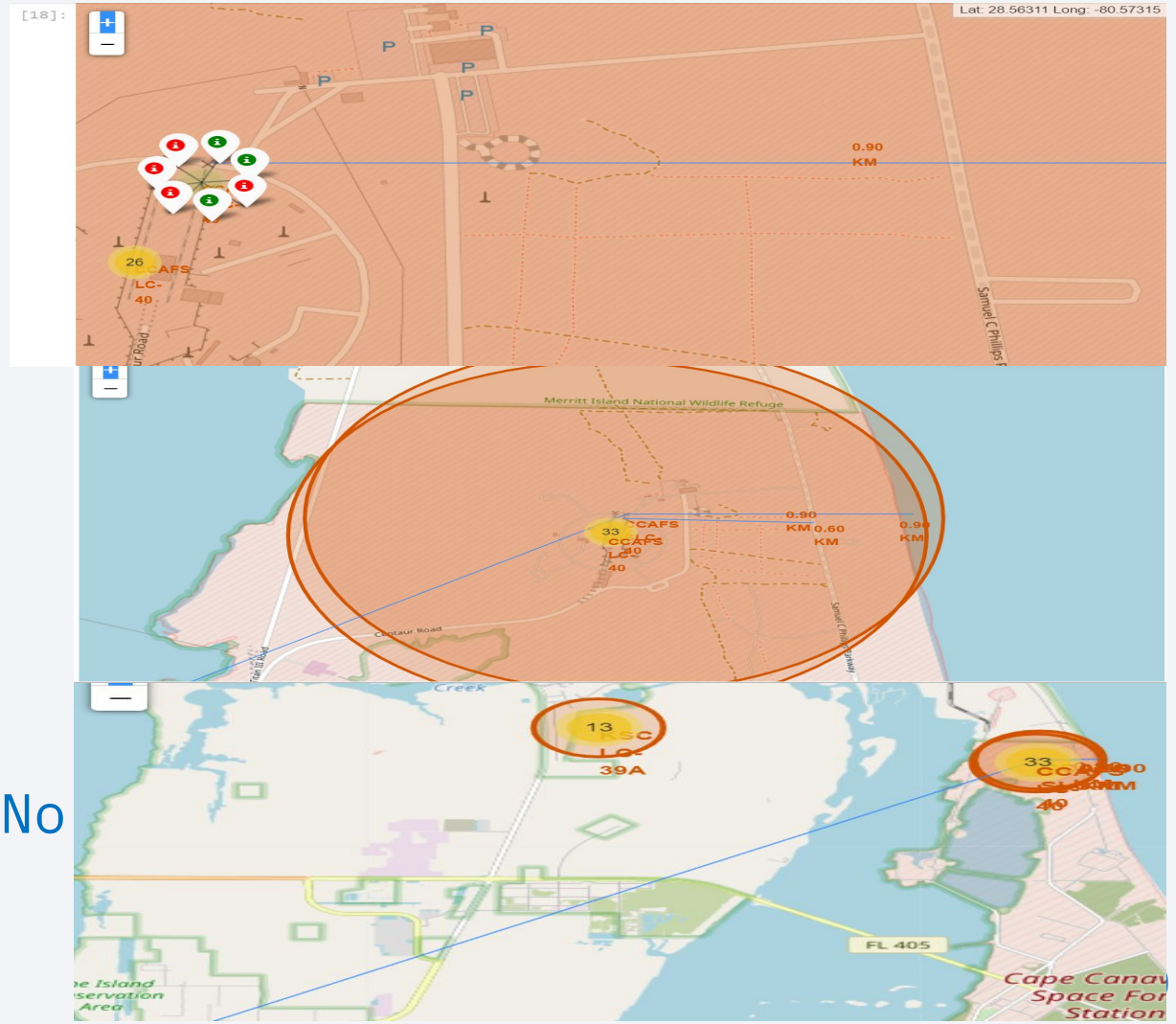
Folium Map
Distances between CCAFS SLC 40 and its proximities

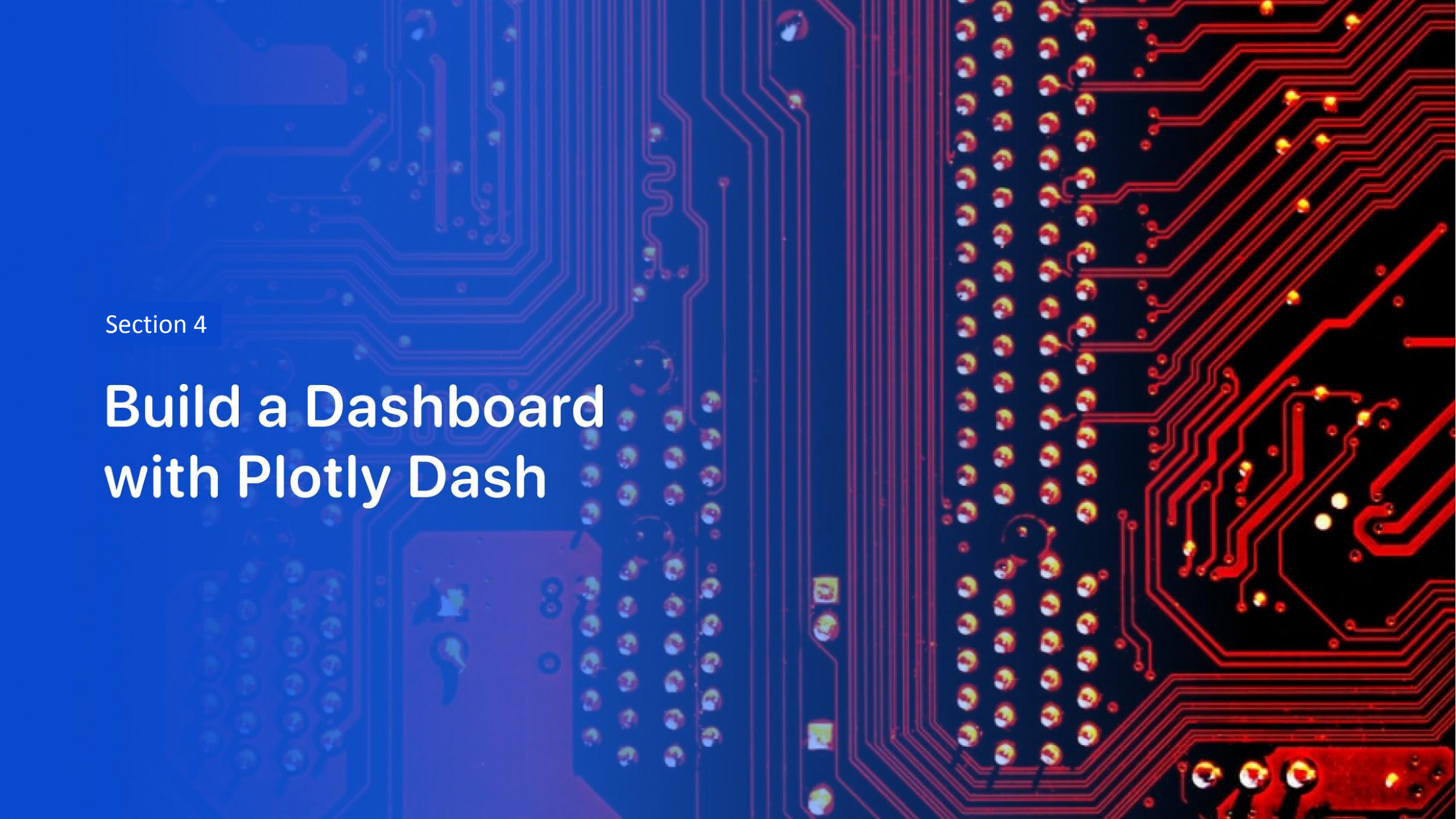
Is CCAFS SLC 40 in close proximity to railways ? Yes

Is CCAFS SLC 40 in close proximity to highways ? Yes

Is CCAFS SLC 40 in close proximity to coastline ? Yes

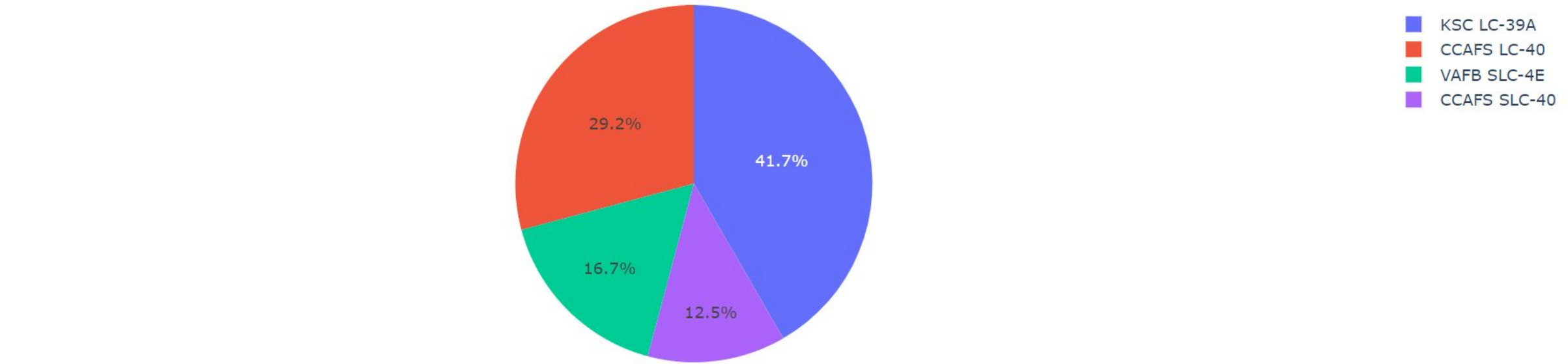
Do CCAFS SLC 40 keeps certain distance away from cities ? No




Section 4
Build a Dashboard with Plotly Dash

SpaceX Launch Records Dashboard

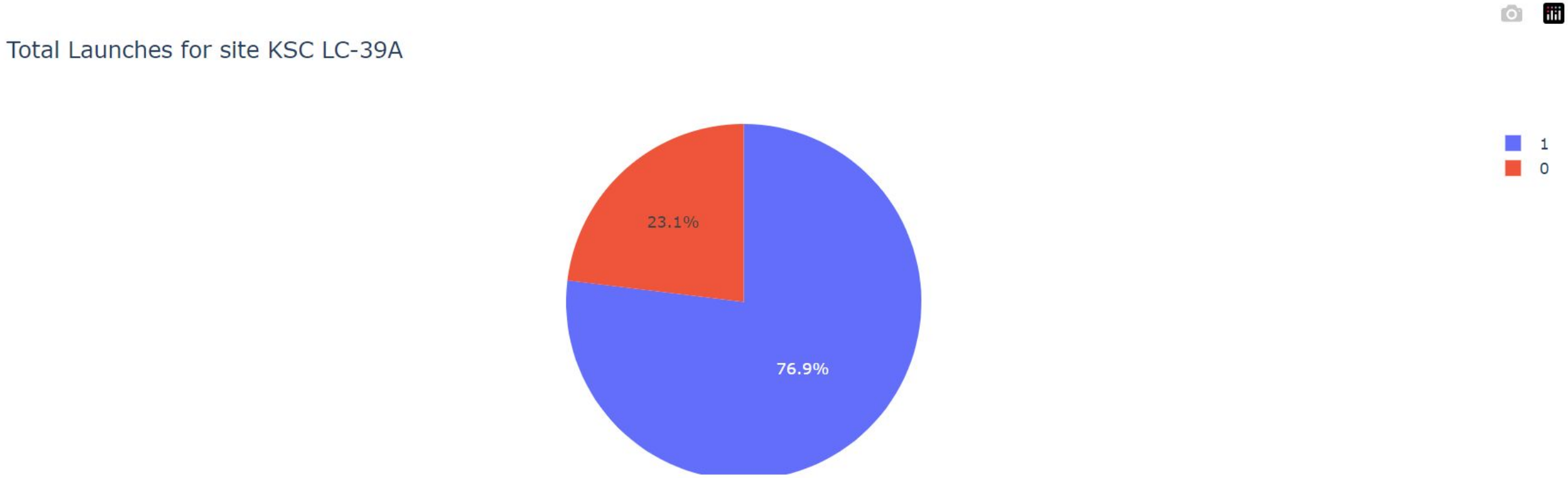
All Sites



The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches

SpaceX Launch Records Dashboard

KSC LC-39A



KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings

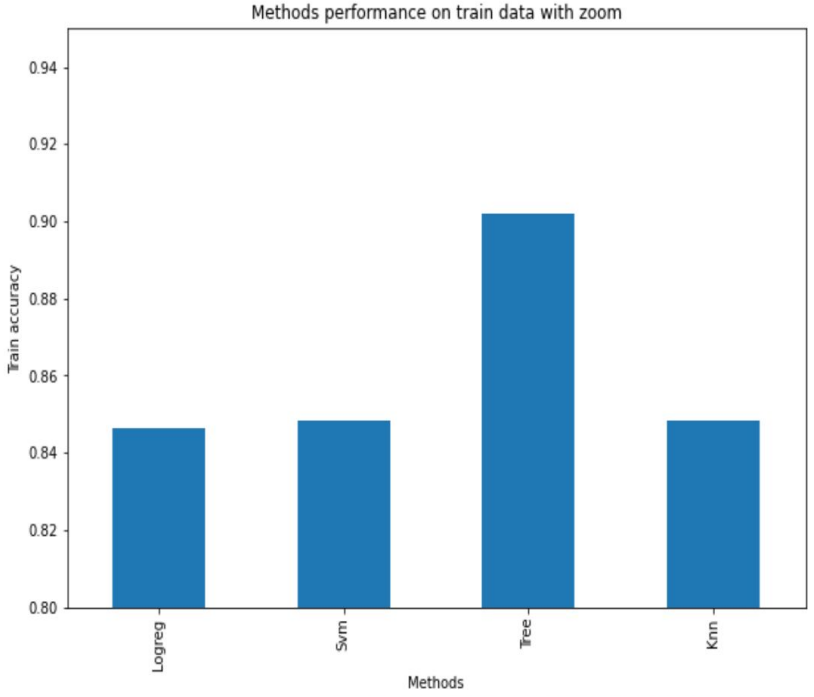
<Dashboard Screenshot 3>

The charts show that payloads between 2000 and 5500 kg have the highest success rate

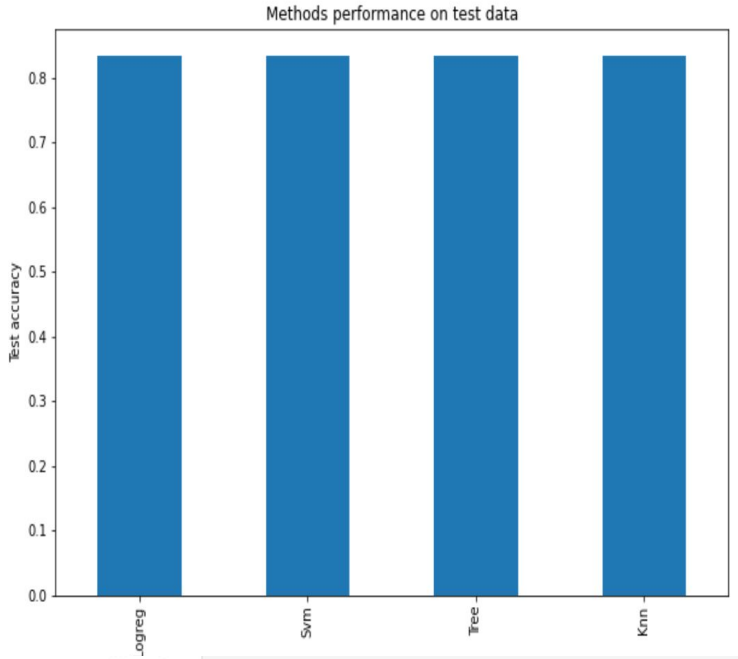
Section 5

Predictive Analysis (Classification)

Classification Accuracy



: Text(0.5, 1.0, 'Methods performance on test data')



Tree	0.901786	0.833333
Knn	0.848214	0.833333
Svm	0.848214	0.833333
Logreg	0.846429	0.833333

In [12]:

```
print("tuned hpyerparameters :(best parameters) ",logreg_cv.best_params_)
print("accuracy :",logreg_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}
accuracy : 0.8464285714285713
```

The scores of the whole Dataset

confirm that the best model is the

Decision Tree Model. This model

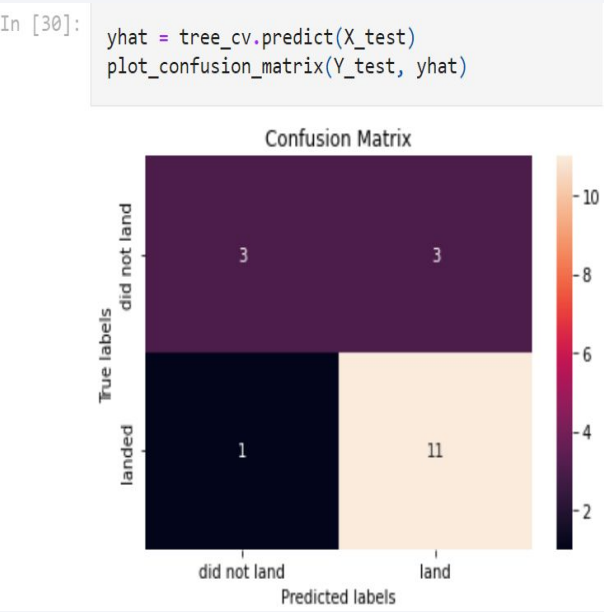
has not only higher scores, but also

the highest accuracy.

Confusion Matrix

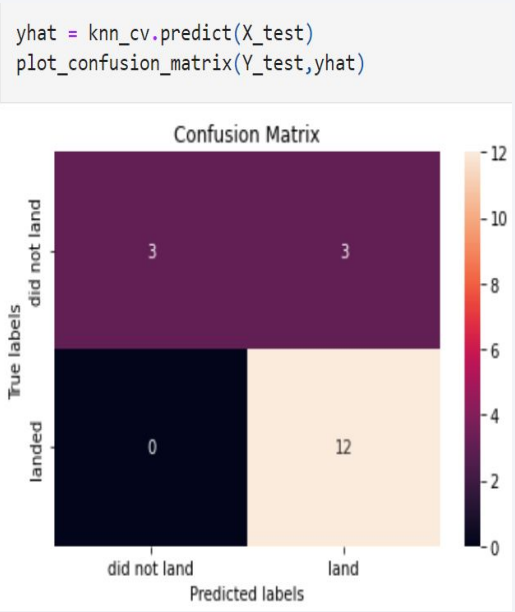
Decision Tree

```
In [30]: yhat = tree_cv.predict(X_test)
         plot_confusion_matrix(Y_test, yhat)
```



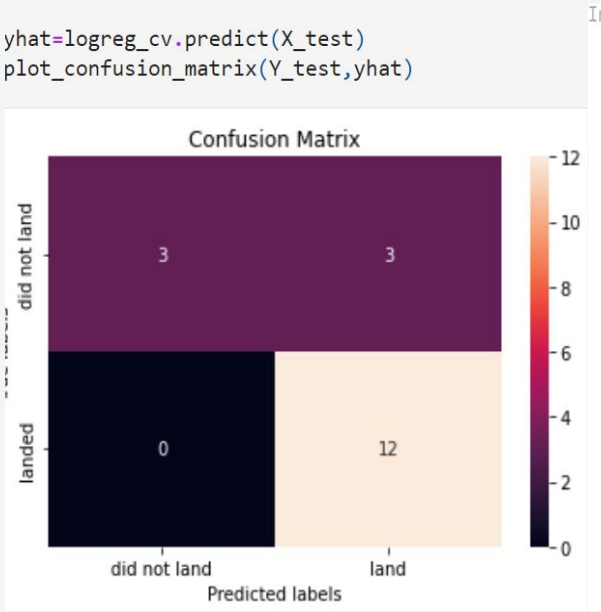
KNN

```
In [35]: yhat = knn_cv.predict(X_test)
         plot_confusion_matrix(Y_test,yhat)
```



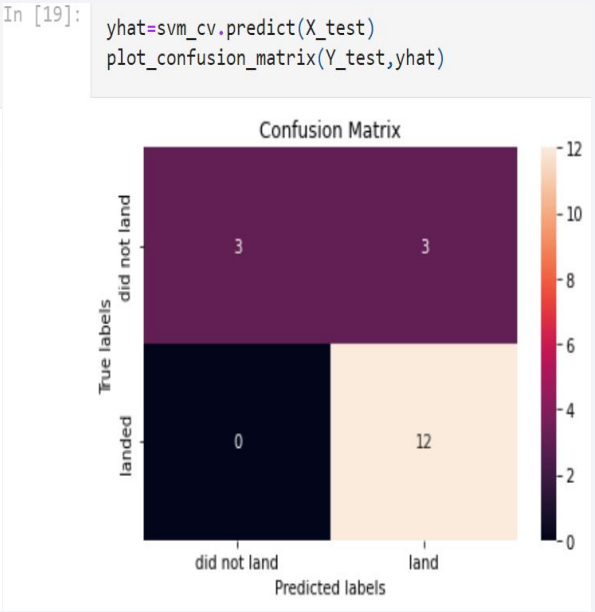
Logistic Regression

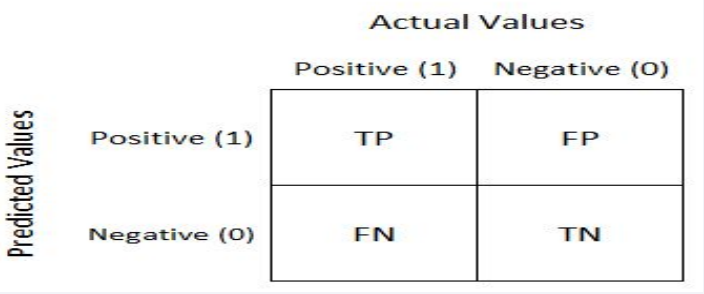
```
yhat=logreg_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



SVM

```
In [19]: yhat=svm_cv.predict(X_test)
         plot_confusion_matrix(Y_test,yhat)
```





As the Accuracy are all equal, the Confusion Matrix are identical.We see that the major problem is false positives.

Conclusions

- Decision Tree Model is the best algorithm for this dataset.
- Launches with a low payload mass show better results than launches with a larger payload mass.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- The success rate of launches increases over the years.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate

Thank you!