IBM Data Science Capstone

AUTHOR: SETTARA PRAMOD

Contents

Executive Summary

- Summary of Methodologies:
  - Data Collection
  - Data Wrangling
  - Exploratory Data Analysis with Visualization
  - Exploratory Data Analysis with SQL
  - Visual Analytics with Folium
  - Dashboard with Dash
  - Predictive Analysis working

Executive Summary

- Summary of Results:
  - EDA Findings
  - Prediction through Support vector machine, Classification Tree & Logistics Regression
  - For SVM sigmoid kernel provides better result on validation dataset
  - Hyper parameters for decision tree classifier through validation data

Introduction

The spaceX advertises the launch of Falcon 9 rocket through their website, it has a cost of 62 million USD, whereas the other providers cost are 165 million USD each, the main reason of the savings of spaceX is because that it can be reused the first stage We can determine the cost of the launch if we are determining that spaceX first stage will land The extracted information can be used for other companies who are interested to bid spaceX for launch of the rocket. We will be predicting that either spaceX first stage would be successfully be landed or not

Methodology

- Data collection methodology
- Data Wrangling
- Exploratory Data Analysis (EDA) using visualization and SQL
- Interactive Visual Analytics using Folium and Plotly
- Predictive Analysis using classification models

Methodology

- **Data collection methodology**
  - SpaceX Rest API
  - Web Scrapping Wikipedia
- **Data Wrangling**
  - Hot encoding data fields and dropping irrelevant columns

Methodology

- **Exploratory Data Analysis (EDA) using visualization and SQL**
  - Scatter/bar graphs
- **Predictive Analysis using classification models**
  - Classification models

Data Collection Methodology

- Data Collection Via Spacex API
  1. Get response from API
  2. Convert to json file
  3. Clean Data
  4. Convert to Dataframe
  5. Filter & export as file

Data Collection Methodology

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
0	1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin1A	167.743129	9.047721
1	2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2A	167.743129	9.047721
2	4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2C	167.743129	9.047721
3	5	2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin3C	167.743129	9.047721
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857

Data Collection Via Spacex API

Data Collection Methodology

- Data Collection Via Web Scraping
  1. Get response from HTML
  2. Create a Soup Object
  3. Find Tables
  4. Get columns names
  5. Create Dict & append data to key
  6. Convert Dict to data frame
  7. Export as CSV

Data Collection Methodology

2020 [edit]

In late 2019, Gwynne Shotwell stated that SpaceX hoped for as many as 24 launches for Starlink satellites in 2020,[490] in addition to 14 or 15 non-Starlink launches. At 26 launches, 13 of which for Starlink satellites, Falcon 9 had its most prolific year, and Falcon rockets were second most prolific rocket family of 2020, only behind China's Long March rocket family.[491]

[hide] Flight No.	Date and time (UTC)	Version, Booster[b]	Launch site	Payload[c]	Payload mass	Orbit	Customer	Launch outcome	Booster landing
78	7 January 2020, 02:19:21[492]	F9 B5 ♺ B1049.4	CCAFS, SLC-40	Starlink 2 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Success (drone ship)
	Third large batch and second operational flight of Starlink constellation. One of the 60 satellites included a test coating to make the satellite less reflective, and thus less likely to interfere with ground-based astronomical observations.[493]								
79	19 January 2020, 15:30[494]	F9 B5 ♺ B1046.4	KSC, LC-39A	Crew Dragon in-flight abort test[495] (Dragon C205.1)	12,050 kg (26,570 lb)	Sub-orbital[496]	NASA (CTS)[497]	Success	No attempt
	An atmospheric test of the Dragon 2 abort system after Max Q. The capsule fired its SuperDraco engines, reached an apogee of 40 km (25 mi), deployed parachutes after reentry, and splashed down in the ocean 31 km (19 mi) downrange from the launch site. The test was previously slated to be accomplished with the Crew Dragon Demo-1 capsule;[498] but that test article exploded during a ground test of SuperDraco engines on 20 April 2019.[419] The abort test used the capsule originally intended for the first crewed flight.[499] As expected, the booster was destroyed by aerodynamic forces after the capsule aborted.[500] First flight of a Falcon 9 with only one functional stage — the second stage had a mass simulator in place of its engine.								
80	29 January 2020, 14:07[501]	F9 B5 ♺ B1051.3	CCAFS, SLC-40	Starlink 3 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Success (drone ship)
	Third operational and fourth large batch of Starlink satellites, deployed in a circular 290 km (180 mi) orbit. One of the fairing halves was caught, while the other was fished out of the ocean.[502]								
81	17 February 2020, 15:05[503]	F9 B5 ♺ B1056.4	CCAFS, SLC-40	Starlink 4 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Failure (drone ship)
	Fourth operational and fifth large batch of Starlink satellites. Used a new flight profile which deployed into a 212 km × 386 km (132 mi × 240 mi) elliptical orbit instead of launching into a circular orbit and firing the second stage engine twice. The first stage booster failed to land on the drone ship[504] due to incorrect wind data.[505] This was the first time a flight proven booster failed to land.								
82	7 March 2020, 04:50[506]	F9 B5 ♺ B1059.2	CCAFS, SLC-40	SpaceX CRS-20 (Dragon C112.3 ♺)	1,977 kg (4,359 lb)[507]	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
	Last launch of phase 1 of the CRS contract. Carries *Bartolomeo*, an ESA platform for hosting external payloads onto ISS.[508] Originally scheduled to launch on 2 March 2020, the launch date was pushed back due to a second stage engine failure. SpaceX decided to swap out the second stage instead of replacing the faulty part.[509] It was SpaceX's 50th successful landing of a first stage booster, the third flight of the Dragon C112 and the last launch of the cargo Dragon spacecraft.								
83	18 March 2020, 12:16[510]	F9 B5 ♺ B1048.5	KSC, LC-39A	Starlink 5 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Failure (drone ship)
	Fifth operational launch of Starlink satellites. It was the first time a first stage booster flew for a fifth time and the second time the fairings were reused (Starlink flight in May 2019).[511] Towards the end of the first stage burn, the booster suffered premature shut down of an engine, the first of a Merlin 1D variant and first since the CRS-1 mission in October 2012. However, the payload still reached the targeted orbit.[512] This was the second Starlink launch booster landing failure in a row, later revealed to be caused by residual cleaning fluid trapped inside a sensor.[513]								
84	22 April 2020, 19:30[514]	F9 B5 ♺ B1051.4	KSC, LC-39A	Starlink 6 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Success (drone ship)

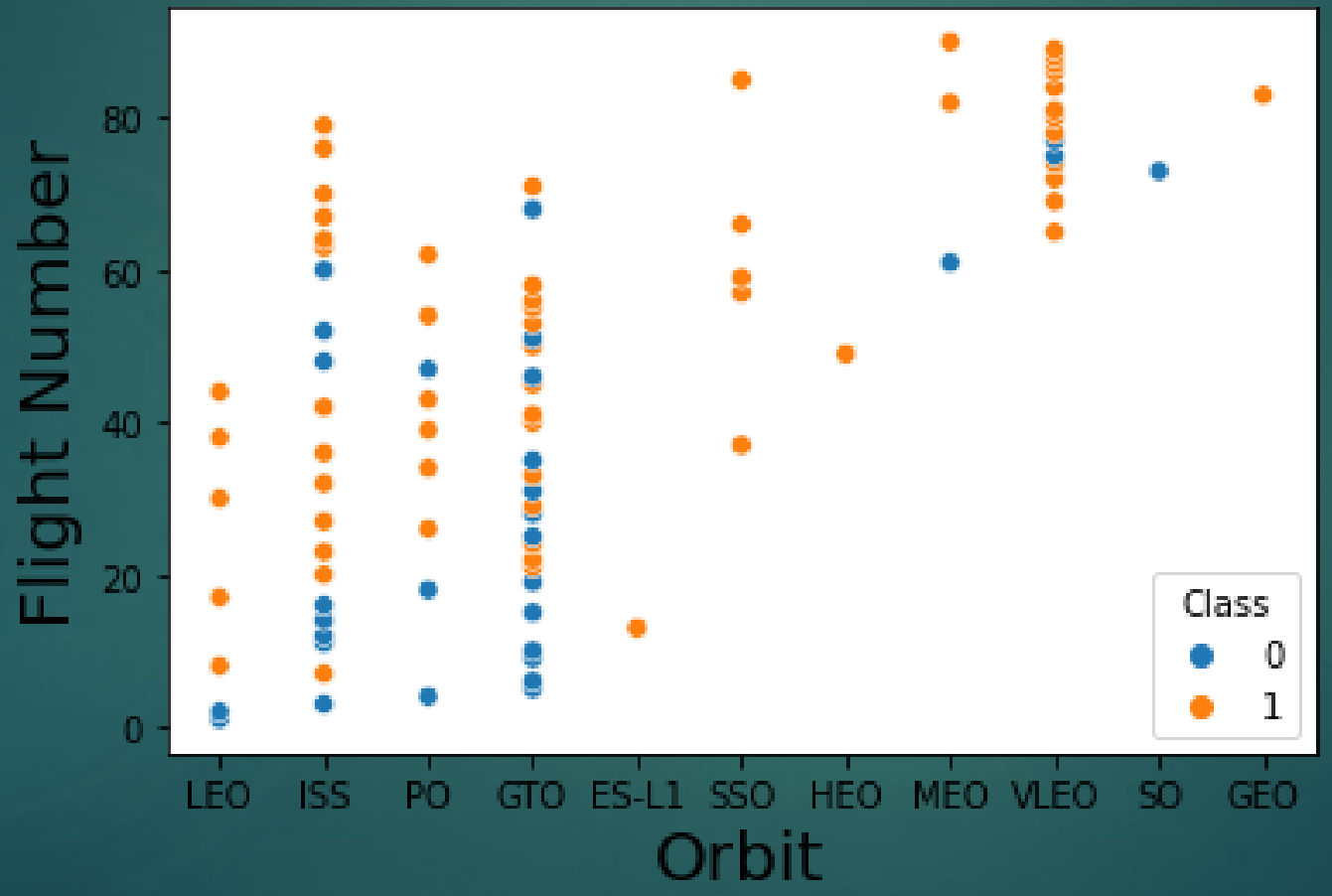
Data Collection Via Web Scraping

Data Wrangling

- Calculate number of launches at each site
- Calculate number and occurrence of each orbit
- Calculate number and occurrence of mission outcome per orbit type
- Create landing outcome label from Outcome column
- Create dictionary and append data to keys
- Export as .CSV file

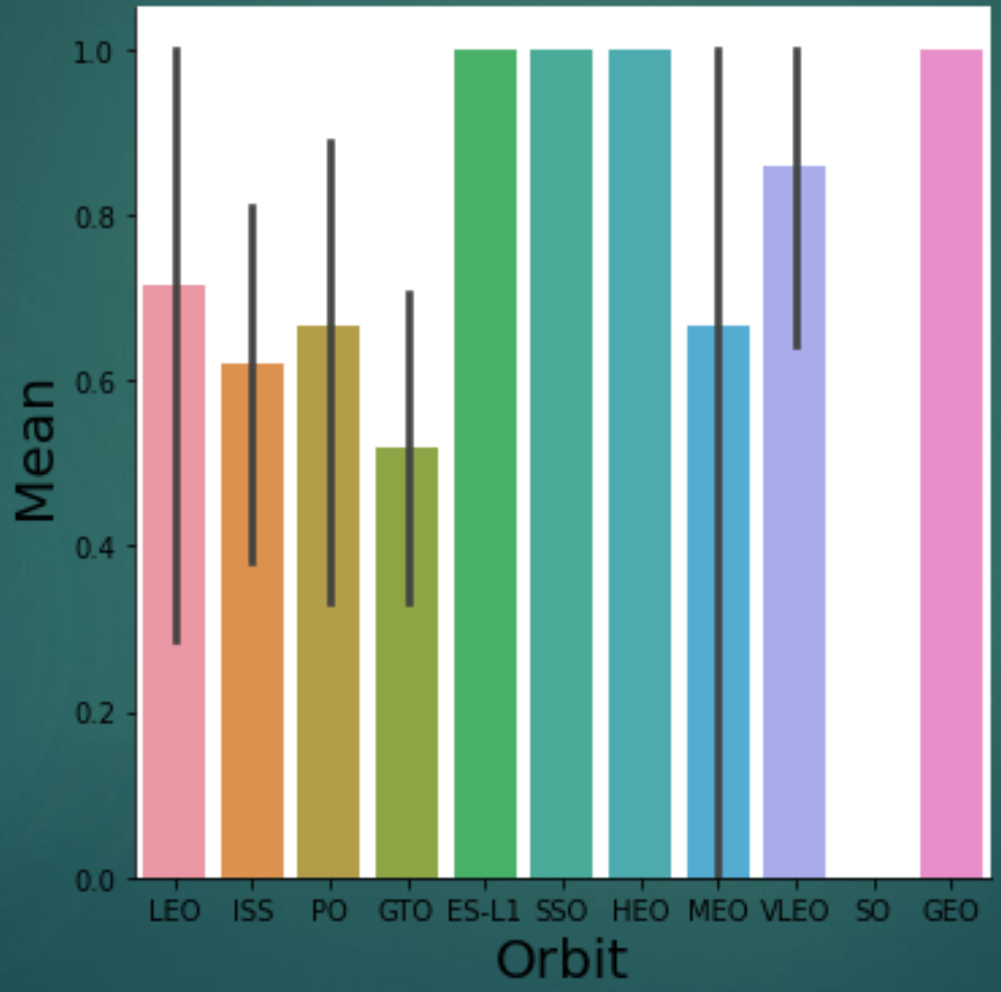
EDA Using Visualization And SQL

- **Scatter graph**: to determine whether there is a noticeable dependency between the attributes



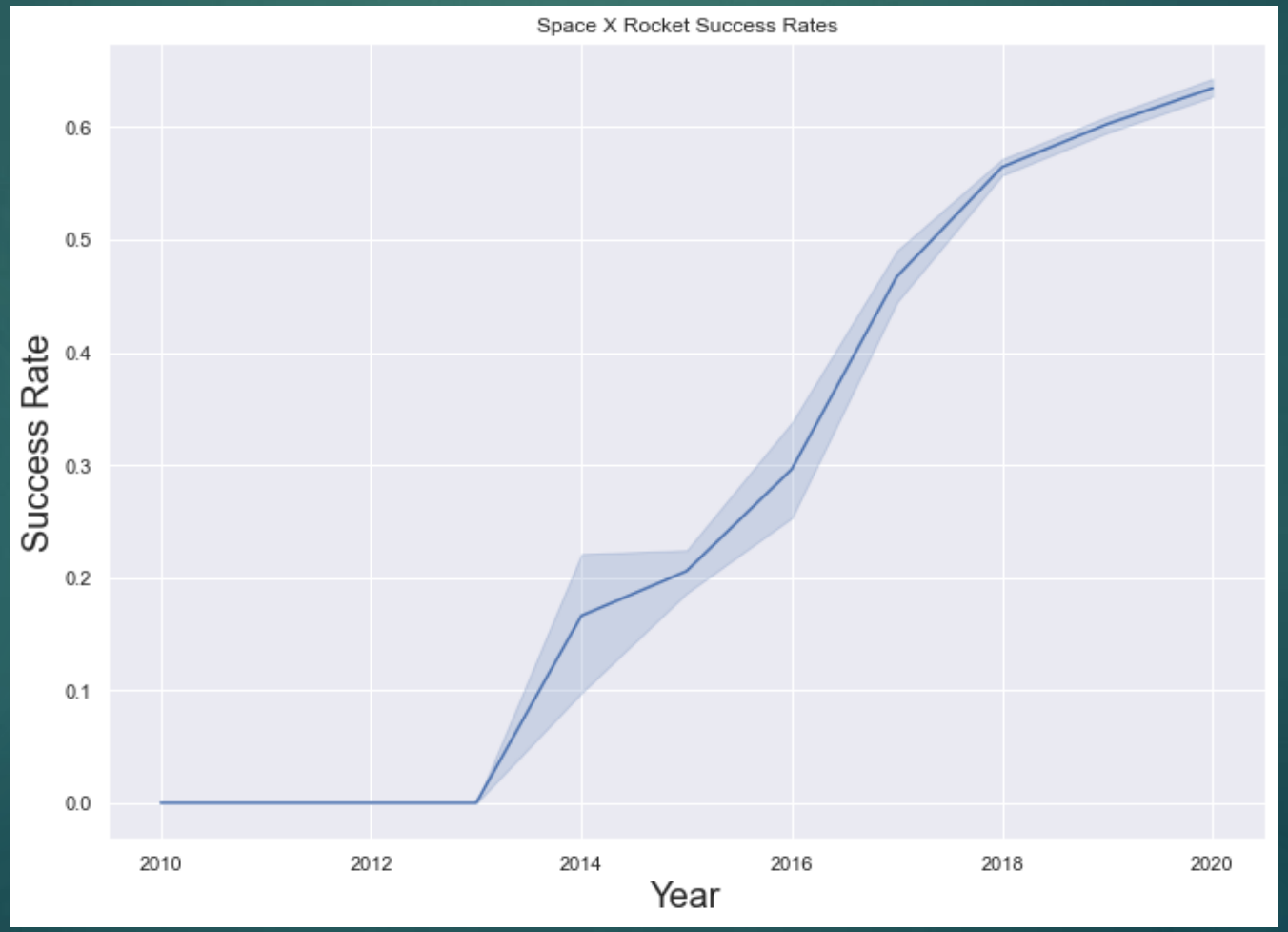
EDA Using Visualization And SQL

- Bar graph: to help identify any visual trends or relationships



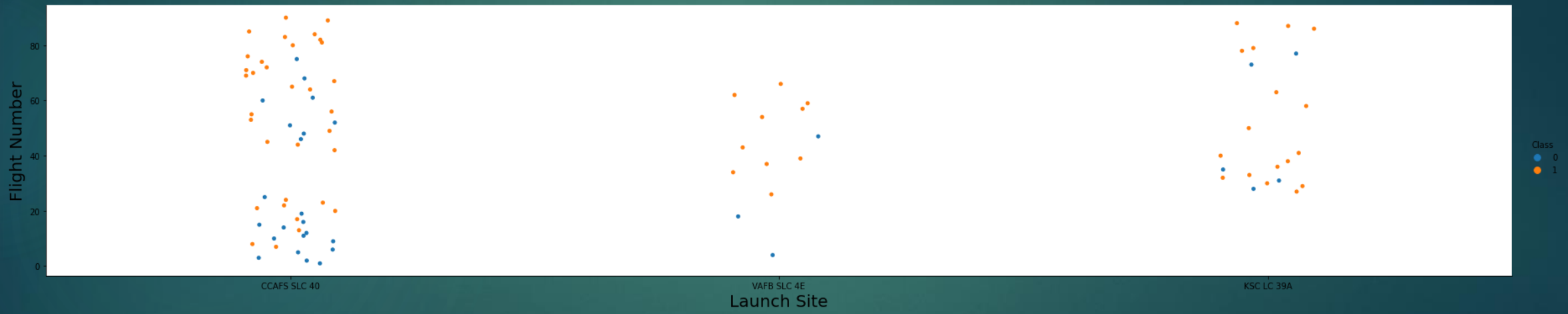
EDA Using Visualization And SQL

- Line graph: helps to track the direct relationship and pattern between the data point



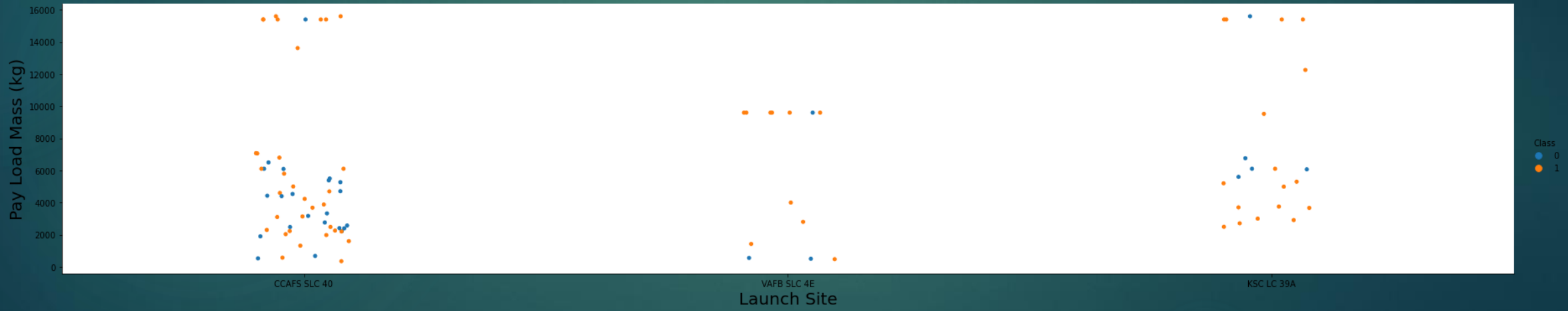
EDA Using Visualization And SQL

- Flight Number Vs. Launch Site
- With higher flight numbers (> 30), the success rate increases



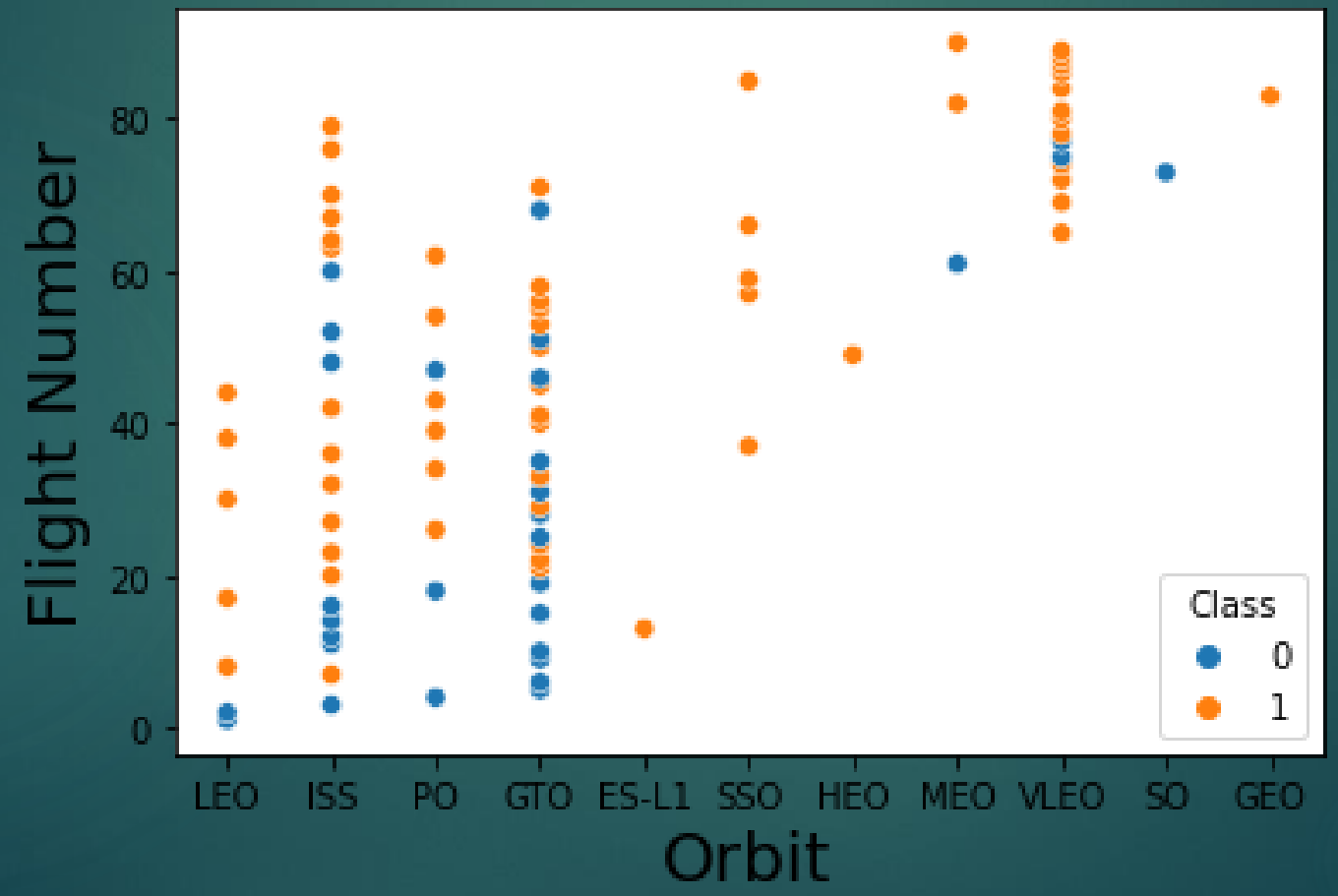
EDA Using Visualization And SQL

- Payload Vs Launch Site
- With greater payload mass (> 7000 KG), the higher the success rate for the rocket but payload mass and launch site are not directly correlated



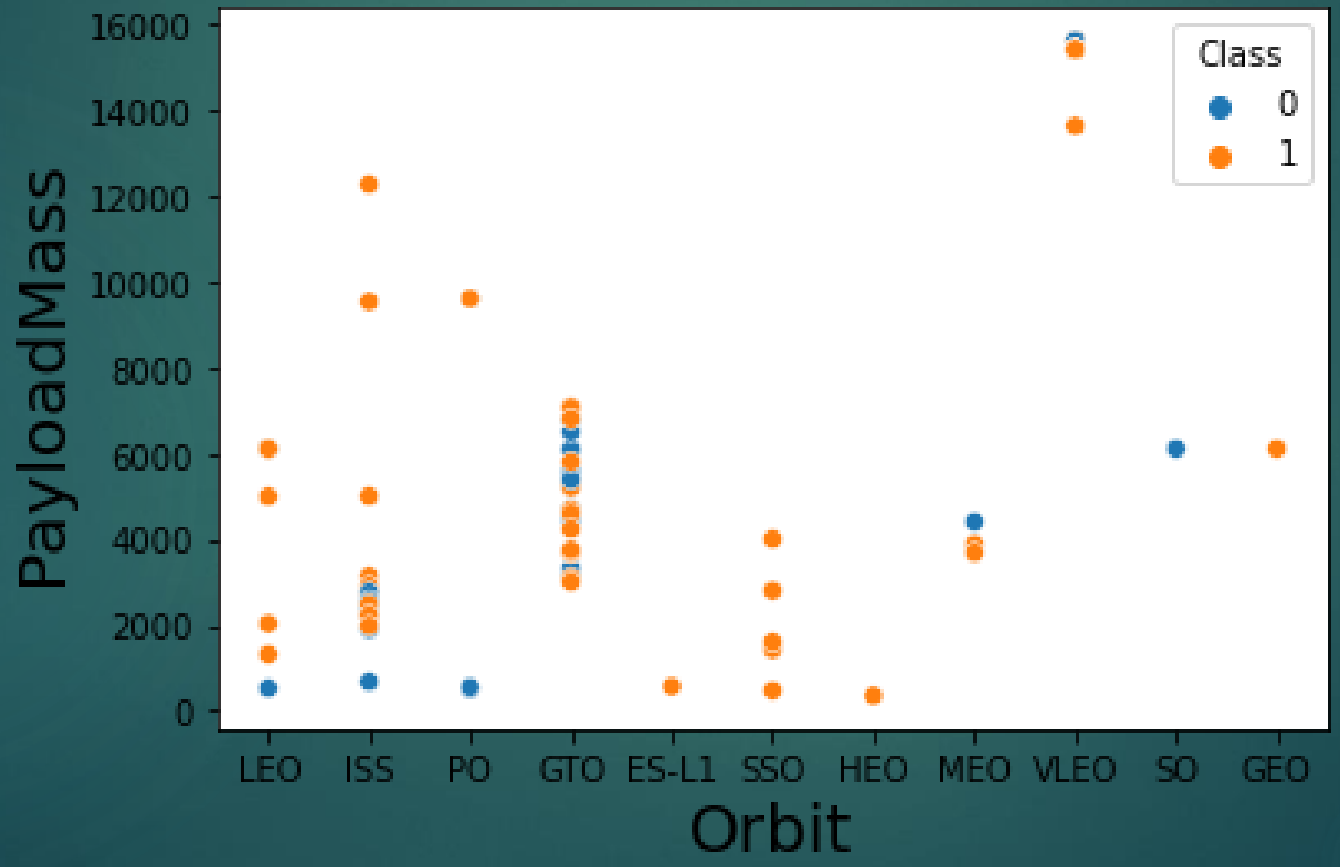
EDA Using Visualization And SQL

- Flight Number Vs Orbit Type
- The LEO orbit had the highest success rate with a higher number of flights



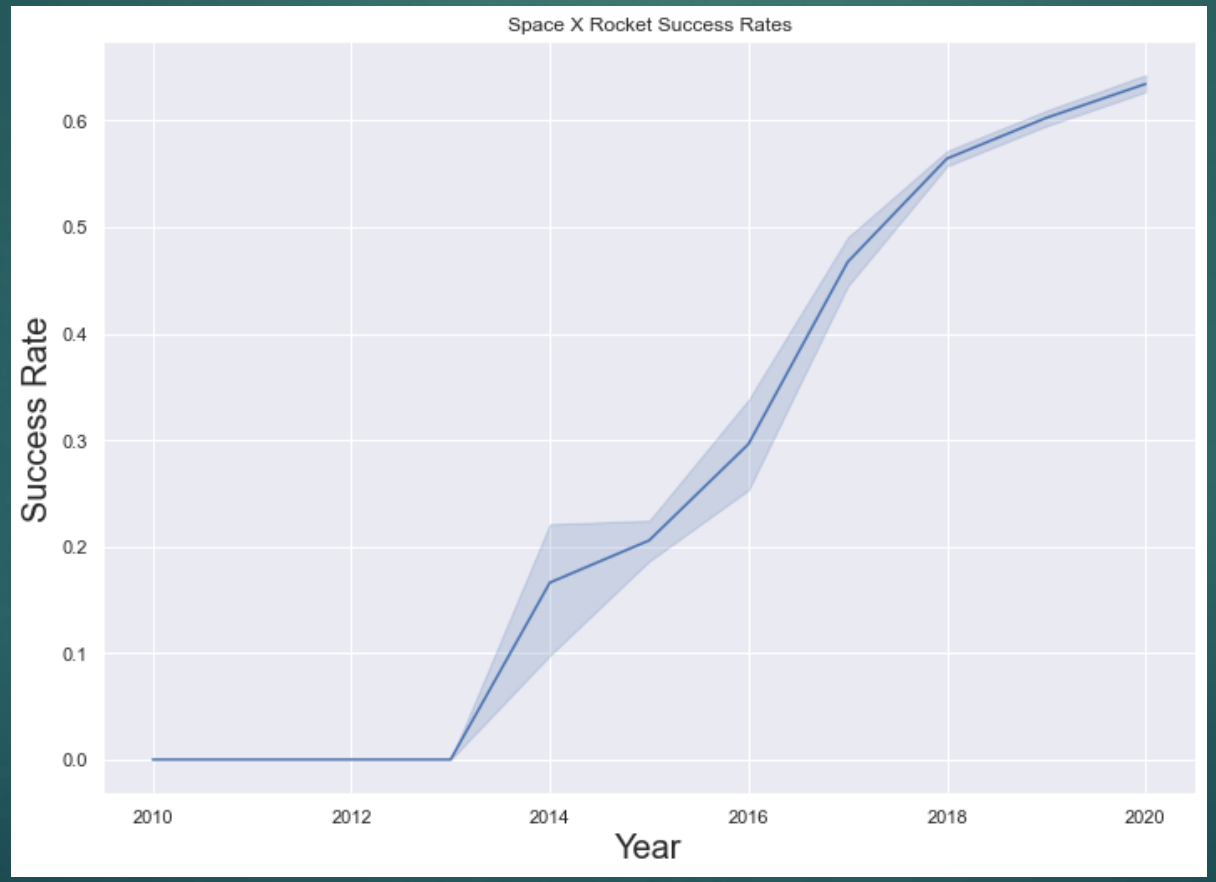
EDA Using Visualization And SQL

- Payload Vs. Orbit Type
- Higher payloads negatively impact the orbits



EDA Using Visualization And SQL

- Launch Success Yearly Trend
- Success rate since 2013 has increased consistently



EDA Using Visualization And SQL

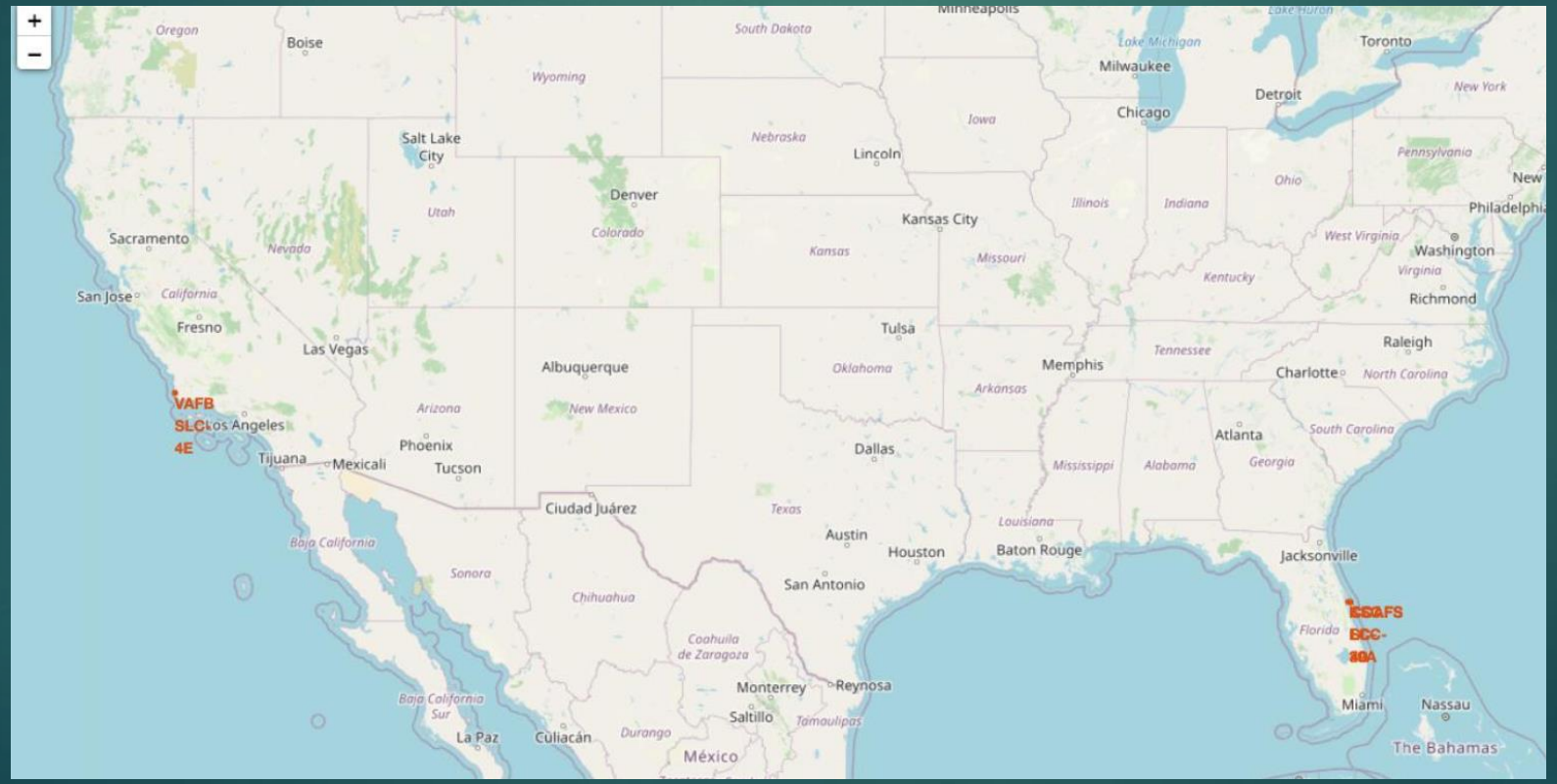
- Queries performed:
  - Display the names of the unique launch sites
  - Display 5 records where launch sites begin with the string ‘CCA’
  - Display total payload mass carried by boosters launched by NASA
  - Display average payload mass carried by booster version F9
  - List the date where the successful landing outcome in drone ship was achieved

EDA Using Visualization And SQL

- Queries performed:
  - List the names of the boosters which have success in ground pad and have a payload mass greater than 4000 and less than 6000
  - List the total number of successful and failed mission outcomes
  - List the names of the booster versions which have carried the maximum payload mass
  - List the failed landing outcomes in drone ship, booster versions and launch site names in 2015
  - Rank the count of landing outcomes between 2010 and 2017

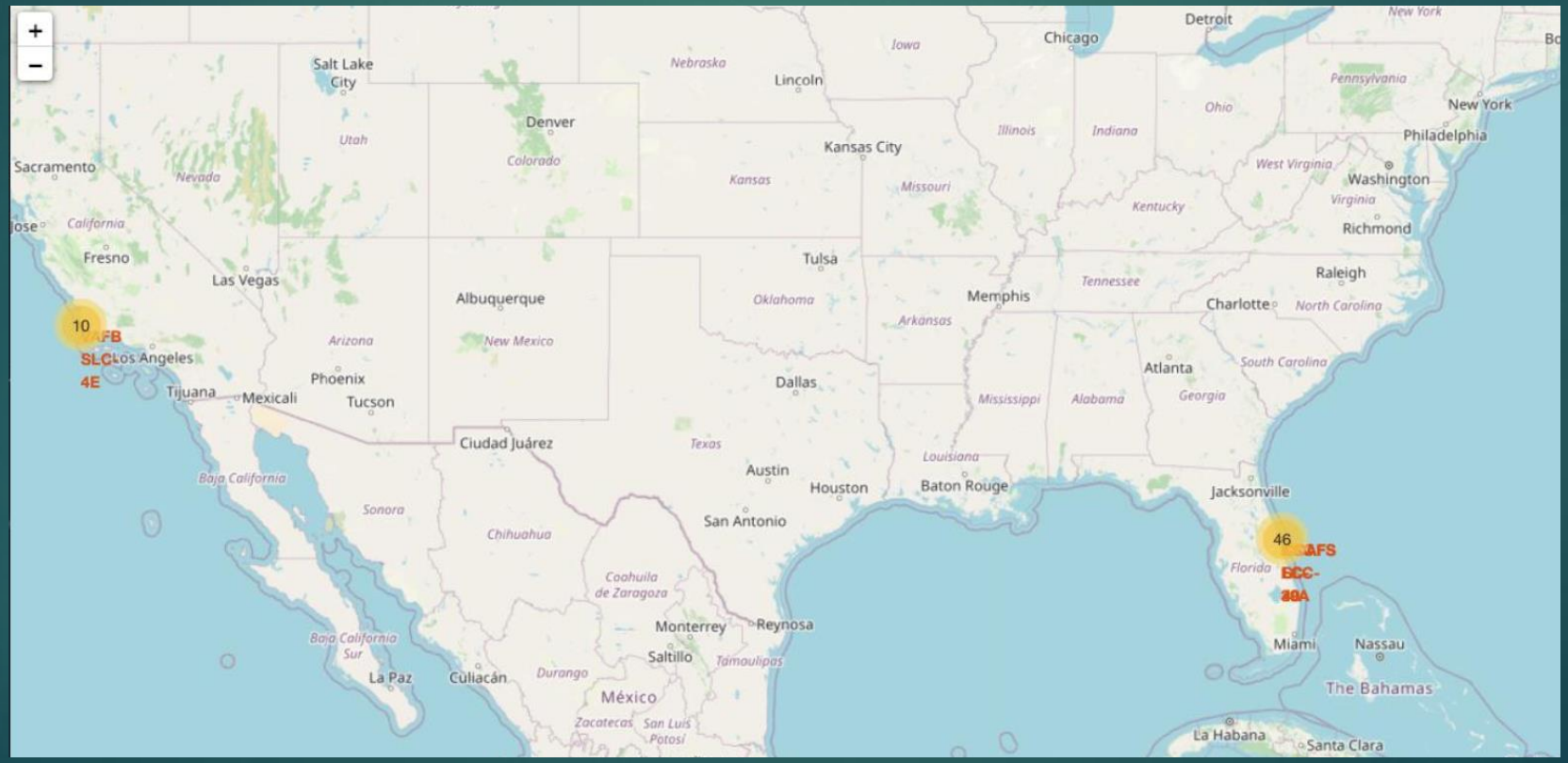
Interactive Visual Analytics using Folium and Plotly

- All Launch Sites' Location Markers On A Global Map



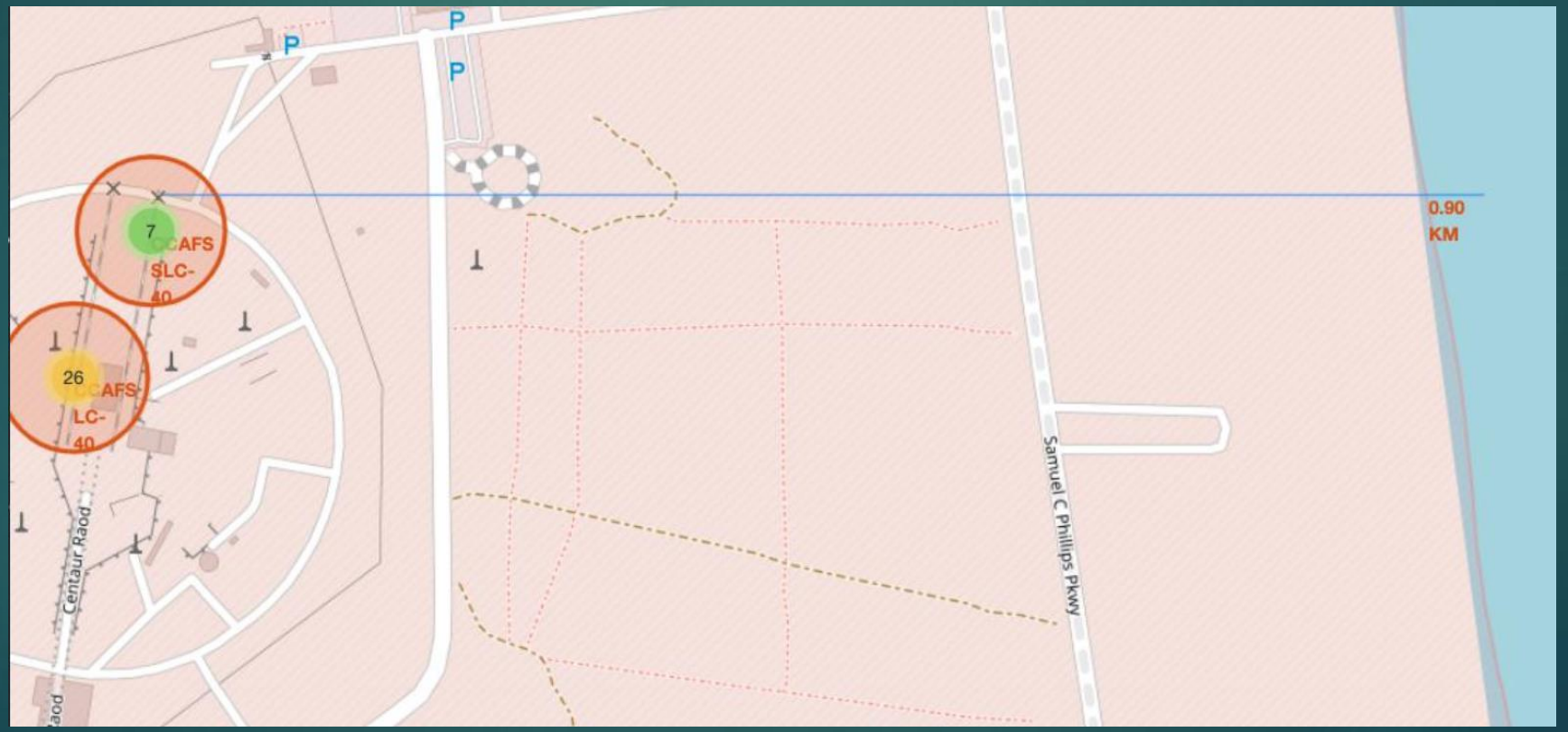
Interactive Visual Analytics using Folium and Plotly

- The Success/Failed Launches For Each Site On The Map



Interactive Visual Analytics using Folium and Plotly

- The Distances Between A Launch Site To Its Proximities



Predictive Analysis (Classification)

- **Built the model**
  - Load the engineered data into a dataframe
  - Transform and standardize the data using Numpy
  - Split the data into training and test data sets
  - Check how many samples were created and set our parameters/algorithms
  - Fit the datasets into the GridSearchCV objects to train the model
- **Evaluating the model**
  - Check the accuracy of the model and plot the Confusion Matrix
  - Finding the best performing classification model
  - Use the highest accuracy score

Results

KNN

```
parameters = {'n_neighbors': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],
              'algorithm': ['auto', 'ball_tree', 'kd_tree', 'brute'],
              'p': [1,2]}

KNN = KNeighborsClassifier()
```

```
gscv = GridSearchCV(KNN,parameters,scoring='accuracy',cv=10)
knn_cv = gscv.fit(X_train,Y_train)
```

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

```
tuned hpyerparameters :(best parameters)  {'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}
accuracy : 0.8482142857142858
```

TASK 11

Calculate the accuracy of tree_cv on the test data using the method score:

```
print("accuracy: ",knn_cv.score(X_test,Y_test))
```

```
accuracy:  0.8333333333333334
```

Results

Logistic Regression

```
parameters ={'C':[0.01,0.1,1],
             'penalty':['l2'],
             'solver':['lbfgs']}
```

```
parameters ={"C":[0.01,0.1,1],'penalty':['l2'], 'solver':['lbfgs']}# l1 lasso l2 ridge
lr=LogisticRegression()
```

```
gscv = GridSearchCV(lr,parameters,scoring='accuracy',cv=10)
logreg_cv = gscv.fit(X_train,Y_train)
```

We output the GridSearchCV object for logistic regression. We display the best parameters using the data attribute best_params_ and the accuracy on the validation data using the data attribute best_score_.

```
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
```

Results

SVM

```
parameters = {'kernel':('linear', 'rbf','poly','rbf', 'sigmoid'),
              'C': np.logspace(-3, 3, 5),
              'gamma':np.logspace(-3, 3, 5)}
svm = SVC()
```

```
gscv = GridSearchCV(svm,parameters,scoring='accuracy',cv=10)
svm_cv = gscv.fit(X_train,Y_train)
```

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

```
tuned hpyerparameters :(best parameters)  {'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}
accuracy : 0.8482142857142856
```

Results

Decision Tree

```
parameters = {'criterion': ['gini', 'entropy'],
     'splitter': ['best', 'random'],
     'max_depth': [2*n for n in range(1,10)],
     'max_features': ['auto', 'sqrt'],
     'min_samples_leaf': [1, 2, 4],
     'min_samples_split': [2, 5, 10]}

tree = DecisionTreeClassifier()
```

```
gscv = GridSearchCV(tree,parameters,scoring='accuracy',cv=10)
tree_cv = gscv.fit(X_train,Y_train)
```

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

```
tuned hpyerparameters :(best parameters)  {'criterion': 'entropy', 'max_depth': 18, 'max_features': 'auto', 'min_samples_leaf': 1, 'min_s
amples_split': 2, 'splitter': 'best'}
accuracy : 0.8785714285714287
```

TASK 9

Calculate the accuracy of tree_cv on the test data using the method score:

```
print("accuracy: ",tree_cv.score(X_test,Y_test))
```

```
accuracy:  0.9444444444444444
```

Accuracy

	Algorithm	Accuracy
0	Logistic Regression	0.846429
1	SVM	0.848214
2	KNN	0.848214
3	Decision Tree	0.862500

Conclusion

- Number of launched success have increases over the years
- ELS1, GEO, HEO, SSO have good success rate
- KSCLC39 have best success rate.
- Launch outcome impacted by the payload mass.
- Accuracy is same on the test model
- Decision tree classifier has highest accuracy