IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Kushagra Singh
02/10/2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - EDA With Visualization
  - EDA With SQL
  - Interactive map with Folium
  - Dashboard with Plotly
  - Predictive Analysis
- Summary of all results
  - EDA
  - Analytic Dash and map demo screenshots
  - Analysis Results

Introduction

- SpaceX advertises that they can provide launch facilities at a cost of 62 Million Dollars, other companies provide the same service at a cost of upwards of 165 million Dollars, major factor in such high difference in cost is due to the ability of SpaceX to reuse the first stage. We need to predict the cost of rocket launch for a new company SpaceY. If we can determine whether the first stage can be reused using the data of SpaceX, Then we can use the resulted model in SpaceY and compete with SpaceX
- We need to find:
  - Factors that lead to successful landing of the first stage.
  - How much each factor affects the landing of the first stage.

Section 1

Methodology

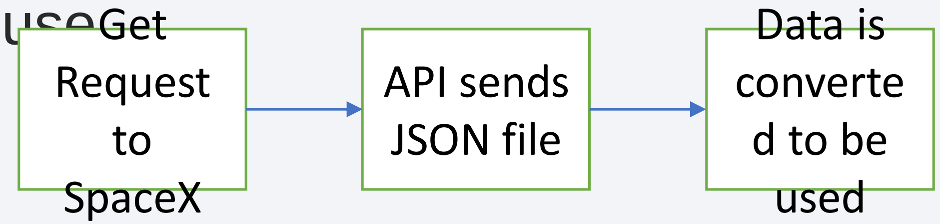
Methodology

- Data collection methodology:
  - Using SpaceX API
  - By Web Scraping Through [Wikipedia]
- Perform data wrangling
  - Preprocessing was done to deal with missing values and encoding was done to replace dummy variables
- EDA Was performed using Graphs and SQL
- Interactive Visual Analytics was done using Folium maps and Plotly Dash
- Predictive Analysis was performed using Classification Models
  - How to build, tune, evaluate classification models

Data Collection

- Data was collected from SpaceX about their rocket launches, our interest was the data about Falcon9 launch data.
    - We used SpaceX Rest API to gather Data like information about rocket used, payload, launch specifications, landing outcome, etc.

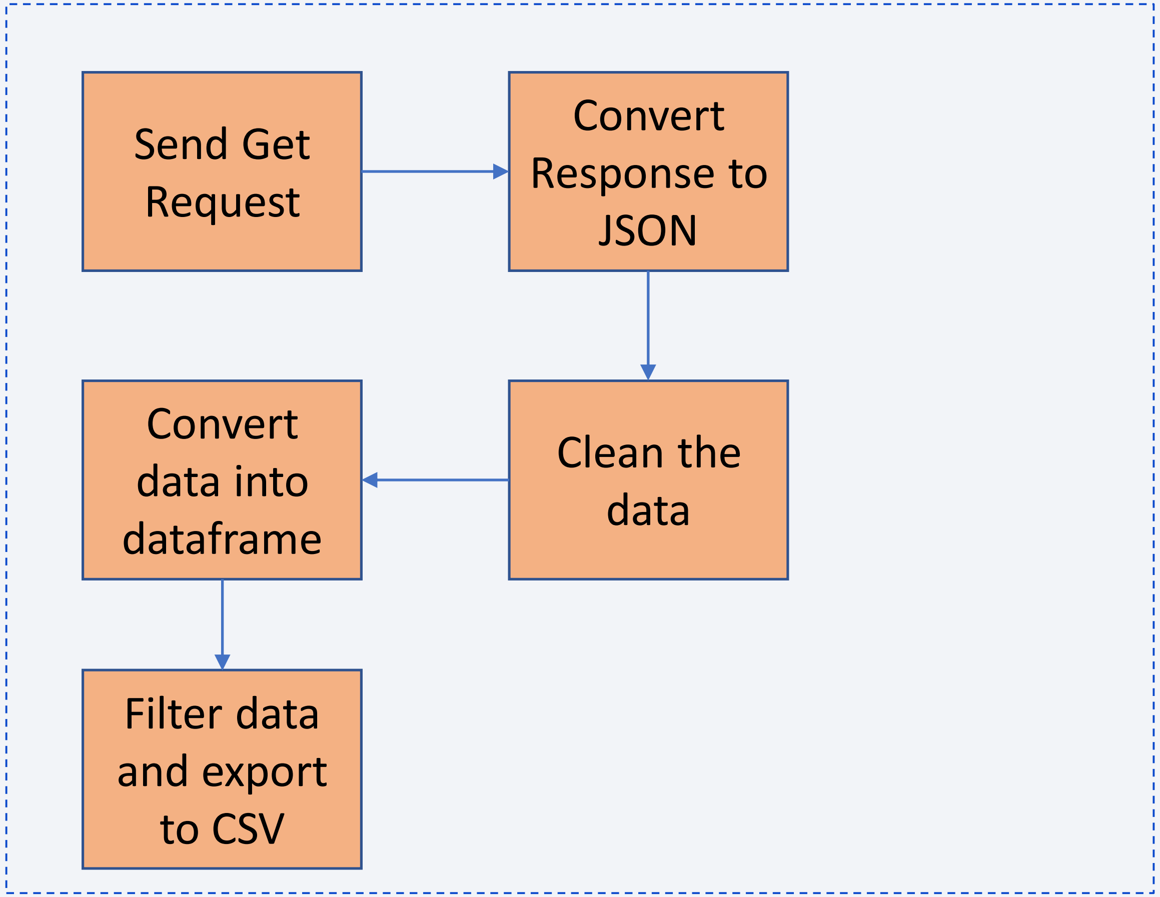
- The SpaceX Rest API endpoint “api.spacexdata.com/v4/” is used to send a get request to get data from the server, the SpaceX server sends a response packet to us in the form of a JSON file that we convert to normal readable format for our use



Data Collection – SpaceX API

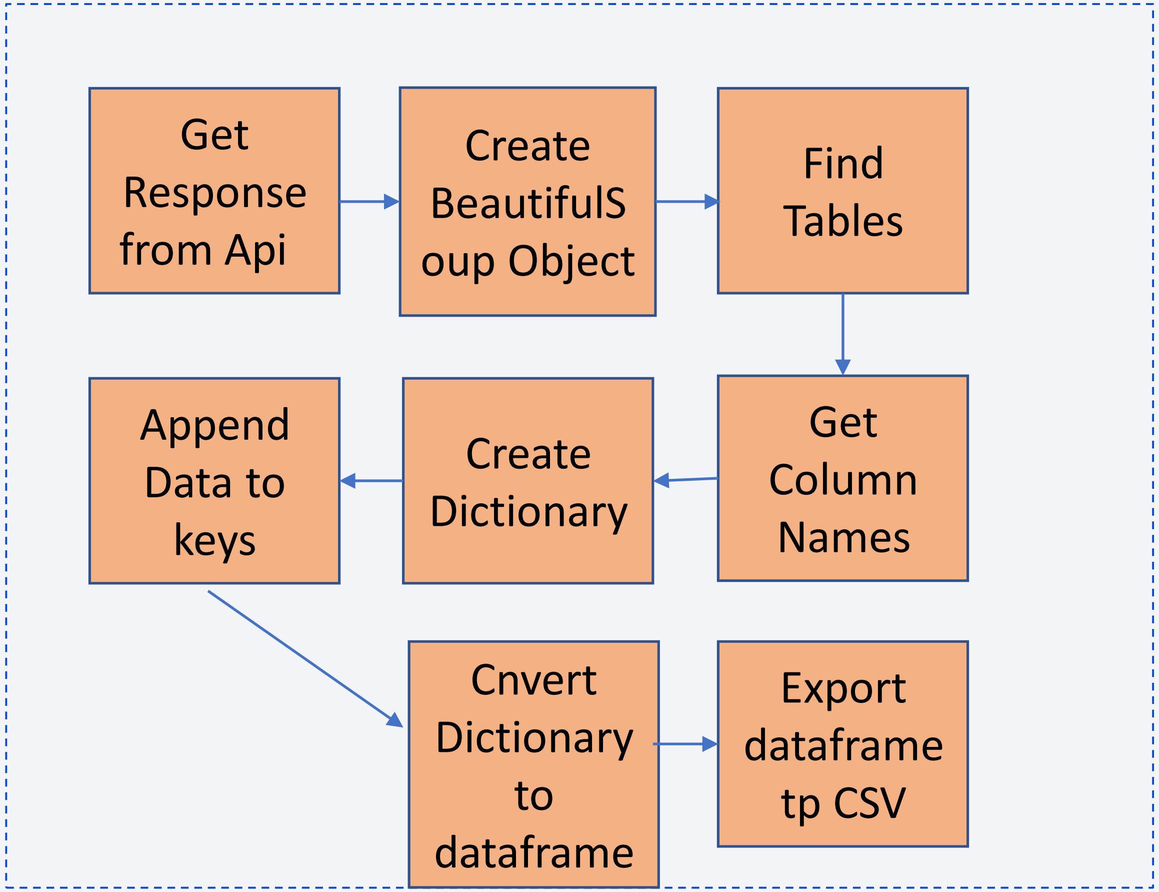
- We send get request to SpaceX API, convert the response to usable format and filter data for use.

- https://github.com/kushagraSh02/Coursera-Capstone/blob/main/Data%20Collection.ipynb



Data Collection - Scraping

- We get response from API in form of JSON, we extract tables using BeautifulSoup Object, we create DataFrame from the tables and export the DataFrame to a CSV file.



- https://github.com/kushagraSh02/Coursera-Capstone/blob/main/Data%20Collection.ipynb

Data Wrangling

- In the dataset there are many different outcomes like the first stage not landing successfully, landing failed due to an accident, whether landing happened on particular section of the ocean etc. True RTLS means the mission was successfully landed to ground, False RTLS means mission was not successfully landed on the ground pad. True ASDS means stage was successfully landed on a drone ship while False ASDS means that the stage was not successfully landed on the drone ship.
- In data Wrangling we convert all the successful outcomes to 1 and all the unsuccessful outcomes to 0.

- https://github.com/kushagraSh02/Coursera-Capstone/blob/main/Data%20Wragling.ipynb

EDA with Data Visualization

- Scatter Plots were drawn to show correlation between the various features.
- Bar Graphs were made to find relation between different values of payload and orbits
- Finally, line graph was drawn to show the trend of success rate with years.

- https://github.com/kushagraSh02/Coursera-Capstone/blob/main/EDA_Visualizations.ipynb

EDA with SQL

- Performed SQL queries on the dataset were:
- Displaying names of Unique launch sites in missions.
- Display 5 records where launch sites begin with the string 'CCA’.
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was acheived.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
- And more.

Build an Interactive Map with Folium

- **To visualize the Launch Data into an interactive map. We took the Latitude and Longitude Coordinates at each launch site and added a Circle Marker around each launch site with a label of the name of the launch site.**

- We assigned the dataframe launch_outcomes(failures, successes) to classes 0 and 1 with Green and Red markers on the map in a MarkerCluster().
- Using Haversine's formula we calculated the distance from the Launch Site to various landmarks to find various trends about what is around the Launch Site to measure patterns. Lines are drawn on the map to measure distance to landmarks

- https://github.com/kushagraSh02/Coursera-Capstone/blob/main/Folium%20Map.ipynb

Build a Dashboard with Plotly Dash

- Pie charts and scatter graphs were used in the Dashboard created using Plotly dash and flask.
- Pie chart was used to show the difference in the launches of the data, scatter plots were used to show correlation between the different payloads and outcomes which was a non-linear relationship.

Predictive Analysis (Classification)

- We imported our data into dataframe, transform and split the data in test and training sets, then we decide the learning model we need to implement by analyzing the data and problem in hand. After selecting the model, we apply GridSearchCV to select the best hyperparameters for the model and train the model on those parameters.
- Build the model => Evaluate the model => Optimize the model => Find best model.
- https://github.com/kushagraSh02/Coursera-Capstone/blob/main/Predictive%20Analysis.ipynb

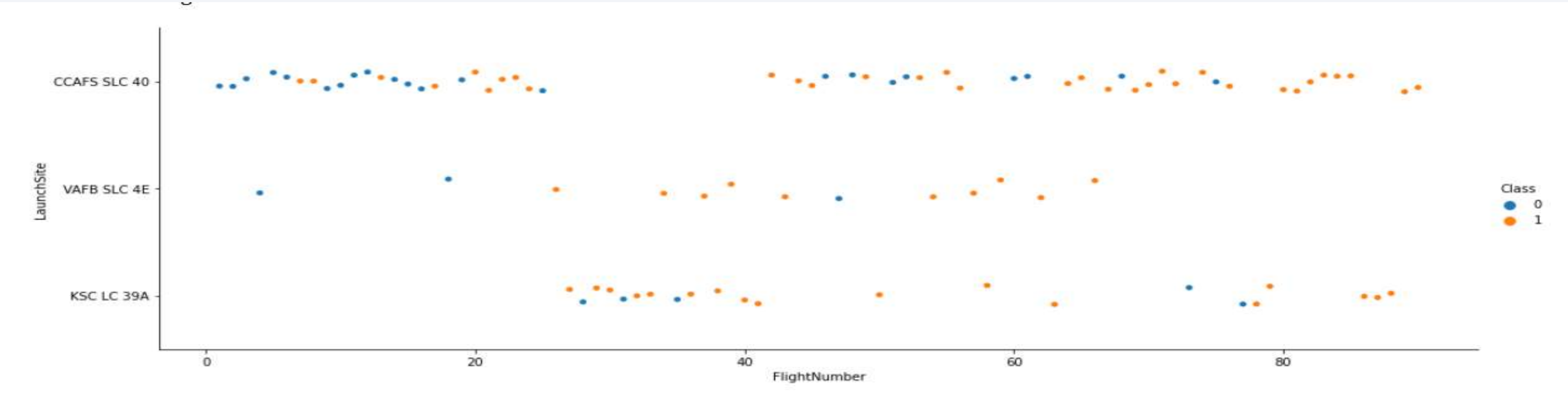
Results

- We can determine whether a Falcon 9 rocket launch would result in a successful landing of the first stage with an accuracy of greater than 83%.
- K Nearest Neighbor, logistic regression, and support vector machine models all performed similarly on test data with approximately 83% accuracy, while the Decision Tree model performed worst with a 78% accuracy score.

Section 2

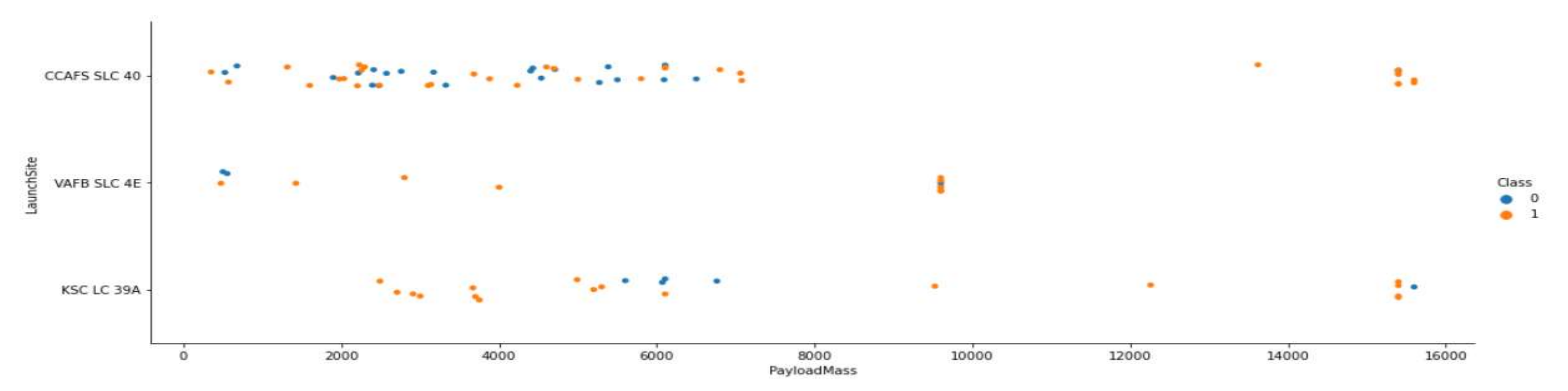
Insights drawn from EDA

Flight Number vs. Launch Site



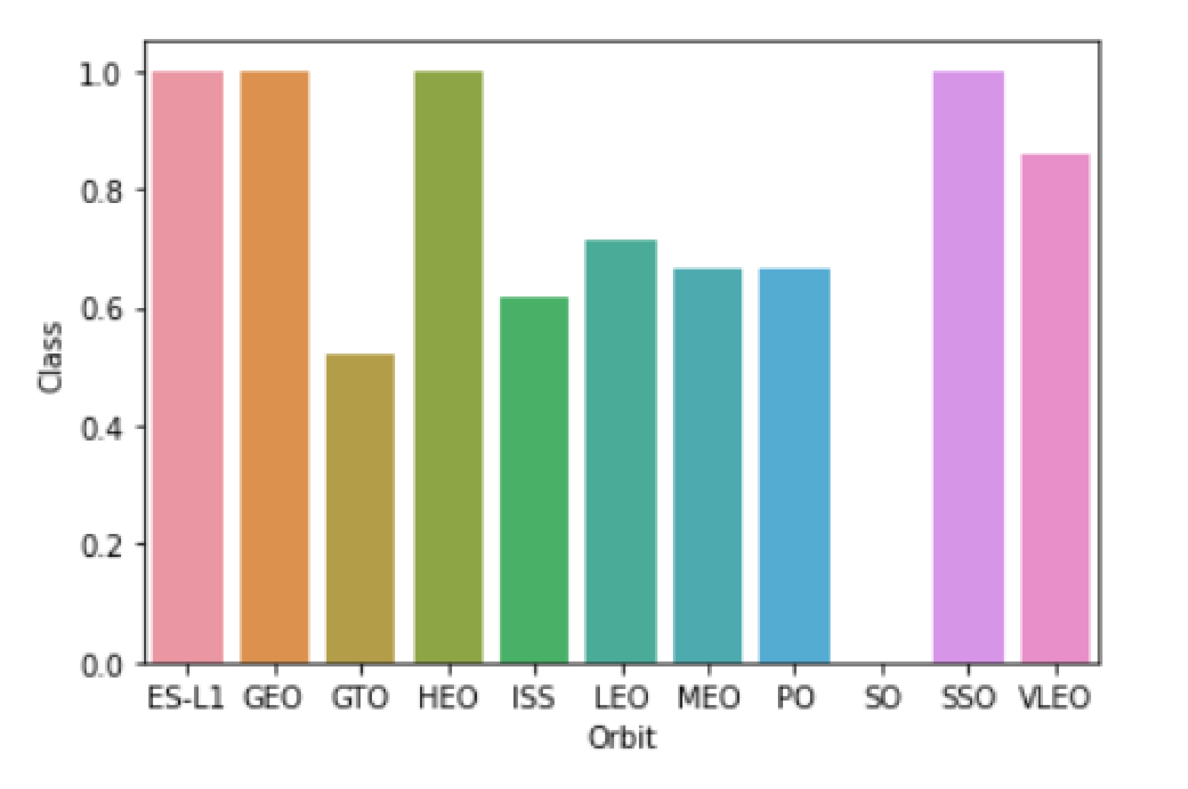
- The greater number of flights at a launch site the greater the success rate at a launch site

Payload vs. Launch Site



- The greater the payload mass for Launch Site CCAFS SLC 40 the higher the success rate for the Rocket. There is not quite a clear pattern to be found using this visualization to decide if the Launch Site is dependent on Pay Load Mass for a success launch.

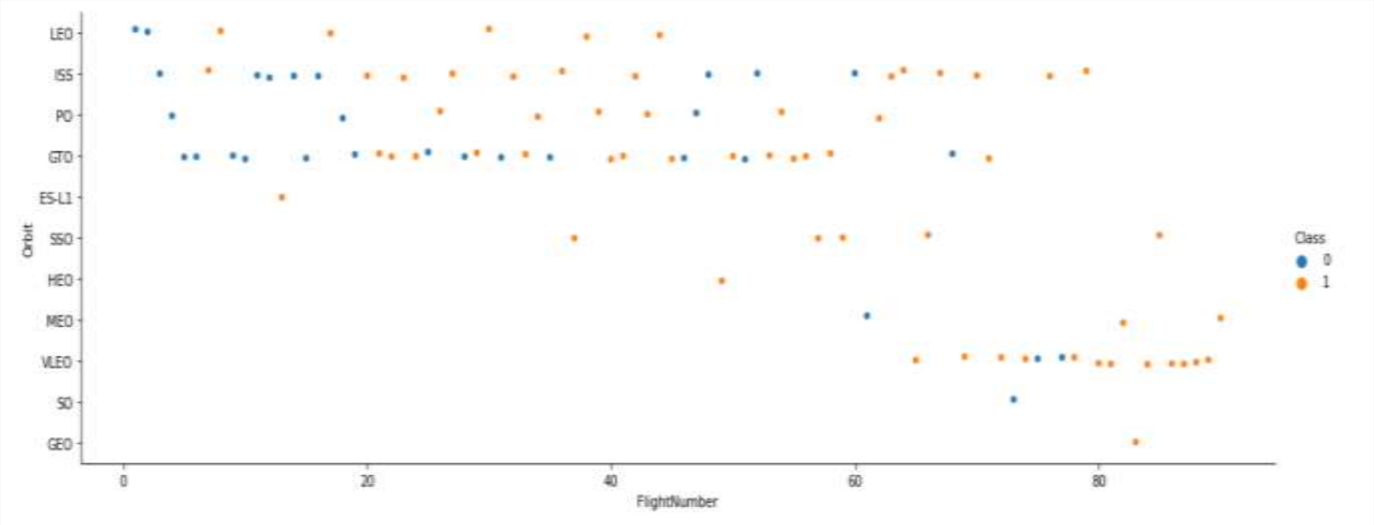
Success Rate vs. Orbit Type



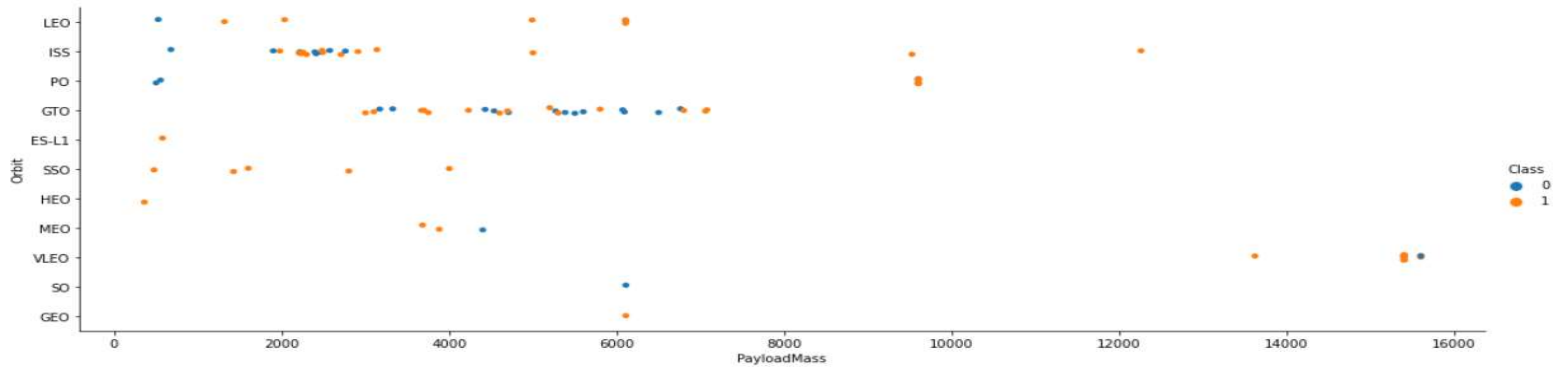
- Orbit GEO,HEO,SSO,ES-L1 have the best Success Rates.

Flight Number vs. Orbit Type

- In 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.



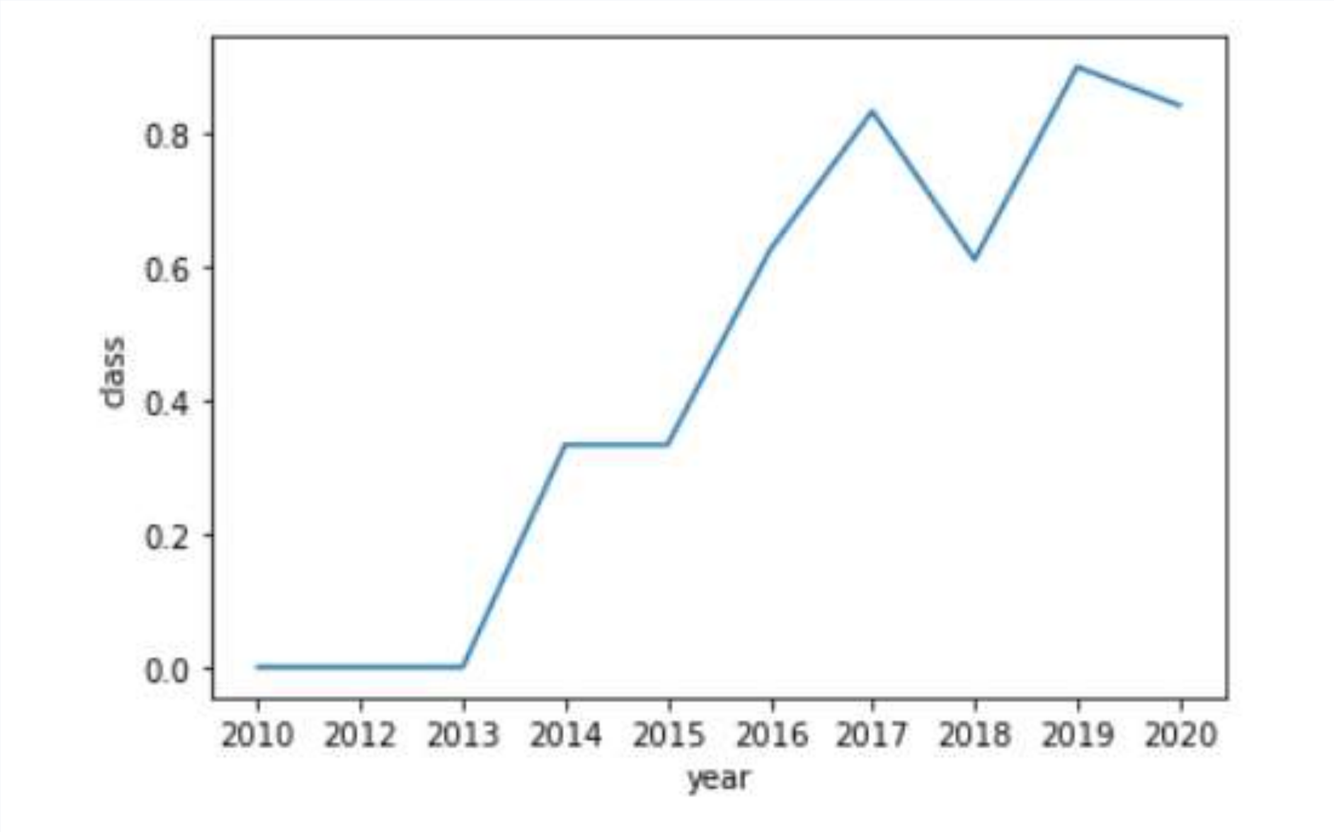
Payload vs. Orbit Type



- 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

- We can see that from 2013, the success rate started to increase and generally increased except in 2018.



All Launch Site Names

select unique(LAUNCH_SITE) from SPACEXDATASET

Out[9]:

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

Launch Site Names Begin with 'CCA'

- SELECT * FROM spacexdataset WHERE launch_site LIKE 'CCA%' limit 5

Out[26]:

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

Total Payload Mass

- SELECT sum(payload_mass__kg_) FROM spacexdataset WHERE customer = ‘NASA (CRS)’

1
45596

Average Payload Mass by F9 v1.1

- SELECT avg(payload_mass__kg_) FROM spacexdataset WHERE booster_version = 'F9 v1.1'

1
2928

First Successful Ground Landing Date

- SELECT date FROM spacexdataset WHERE landing__outcome = ‘Success (ground pad)’ ORDER BY date ASC LIMIT 1

DATE
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- SELECT booster_version, payload_mass__kg_ FROM spacexdataset WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ > 4000 AND payload_mass__kg_ < 6000

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

Total Number of Successful and Failure Mission Outcomes

- SELECT mission_outcome, count(*) as count FROM spacexdataset GROUP BY mission_outcome

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

Boosters Carried Maximum Payload

- SELECT distinct(booster_version) FROM spacexdataset WHERE payload_mass__kg_ = (SELECT max(payload_mass__kg_) FROM spacexdataset)

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

2015 Launch Records

- SELECT landing__outcome, booster_version, launch_site FROM spacexdataset WHERE landing__outcome = 'Failure (drone ship)' and year(date) = '2015'

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

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

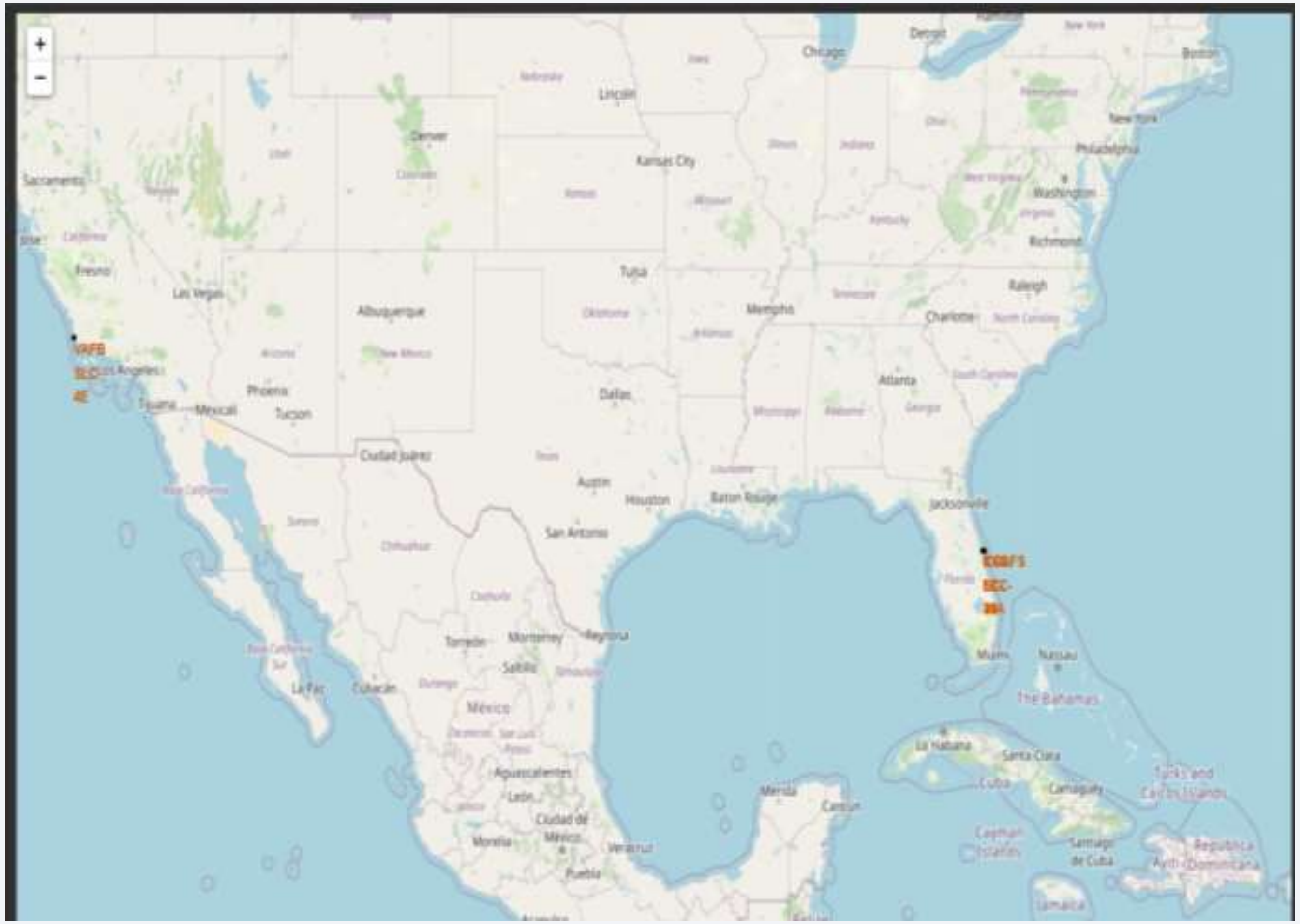
- SELECT landing__outcome, count(*) as count FROM spacexdataset WHERE date > '2010-06-04' and date < '2017-03-20' GROUP BY landing__outcome ORDER BY count desc

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

Section 3

Launch Sites Proximities Analysis

Falcon 9 Launch Sites.



- This map shows the locations of the different sites for Falcon 9 launches

Clusters of Launch Sites with Success

- This map shows a close up of a launch site as well as color coded labels indicating the landing success of failure of individual launches

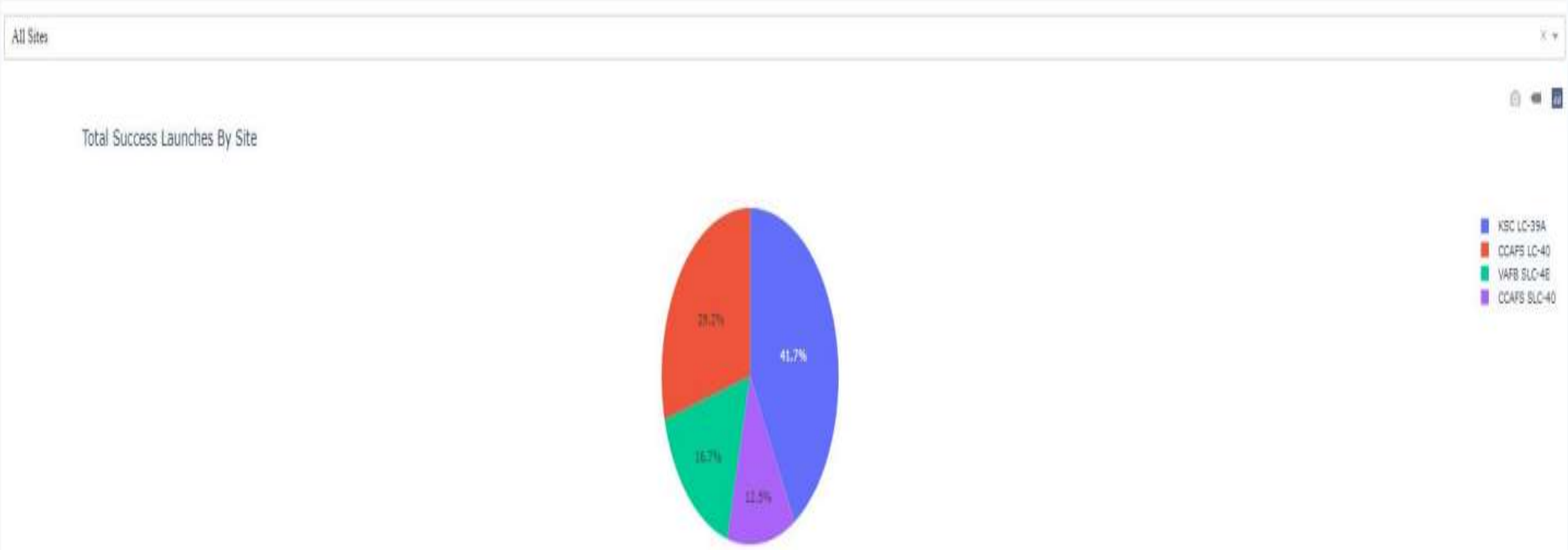
Locations nearby to Launch Sites

- The locations of Ocean nearby to launch sites.

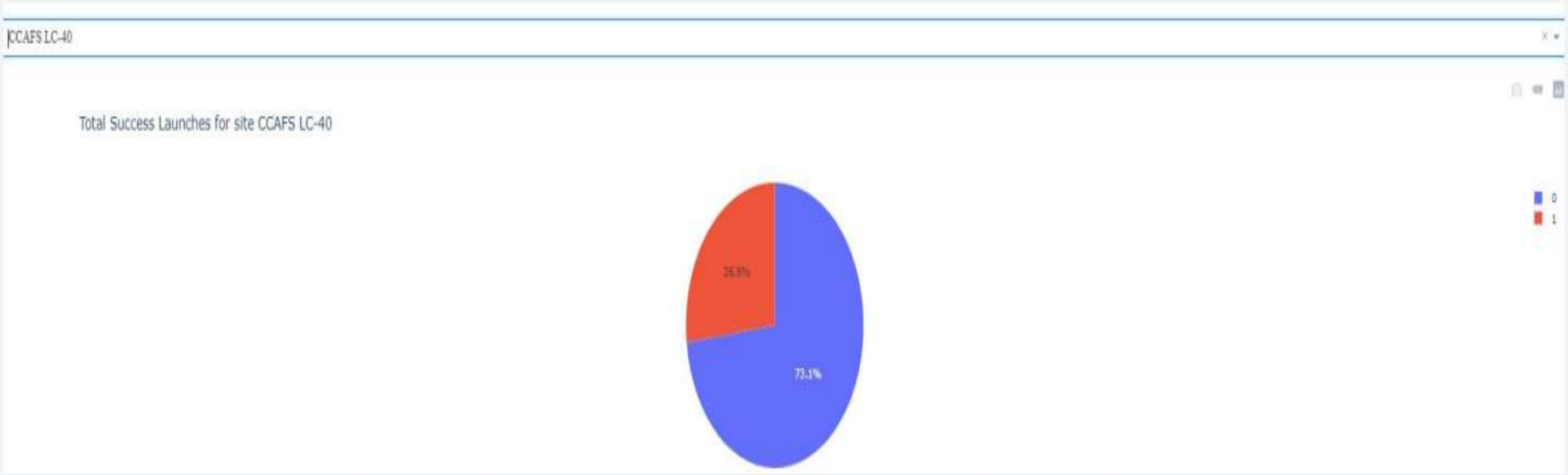
Section 4

Build a Dashboard with Plotly Dash

Success of each Launch Site



Success Rate of Individual Site



Payload V/S Success Rate

Section 5

Predictive Analysis (Classification)

Classification Accuracy

```
GridSearchCV(cv=10, estimator=KNeighborsClassifier(),
             param_grid={'algorithm': ['auto', 'ball_tree', 'kd_tree', 'brute'],
                         'n_neighbors': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],
                         'p': [1, 2]}): 0.8333333333333334
GridSearchCV(cv=10, estimator=SVC(),
             param_grid={'C': array([1.00000000e-03, 3.16227766e-02, 1.00000000e+00, 3.16227766e+01,
       1.00000000e+03]),
                         'gamma': array([1.00000000e-03, 3.16227766e-02, 1.00000000e+00, 3.16227766e+01,
       1.00000000e+03]),
                         'kernel': ('linear', 'rbf', 'poly', 'rbf', 'sigmoid')}): 0.8333333333333334
GridSearchCV(cv=10, estimator=LogisticRegression(),
             param_grid={'C': [0.01, 0.1, 1], 'penalty': ['l2'],
                         'solver': ['lbfgs']}): 0.8333333333333334
GridSearchCV(cv=10, estimator=DecisionTreeClassifier(),
             param_grid={'criterion': ['gini', 'entropy'],
                         'max_depth': [2, 4, 6, 8, 10, 12, 14, 16, 18],
                         'max_features': ['auto', 'sqrt'],
                         'min_samples_leaf': [1, 2, 4],
                         'min_samples_split': [2, 5, 10],
                         'splitter': ['best', 'random']}): 0.7777777777777778
```

Confusion Matrix

Conclusions

- Low weighted payloads perform better than the heavier payloads.
- It is possible to predict whether or not a Falcon 9 rocket will be landed successfully with greater than 83% accuracy.
- Launch variables can be used by competing companies to provide evidence of higher than estimated costs if it is predicted that there will not be a successful landing.
- Orbit GEO,HEO,SSO,ES-L1 has the best Success Rate

Thank you!