IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ulisses J. Jacobo
2022/12/23

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies
  - The objective of this project is to determine if the first stage will land successfully. Drawing from techniques involving data collection, data wrangling, interactive visual analytics, exploratory data analysis, and predictive analysis. The resulting information would prove useful against other companies that would want to contest SpaceX contract in the future.
- Summary of all results
  - EDA Results
  - Interactive Map Results
  - Dashboard Results
  - Predictive Analysis Results

Introduction

- Project background and context
  - In 2002, SpaceX became the first commercial spaceflight company to successfully launch and return a spacecraft from earth orbit and the first to launch a manned flight and dock with the international space station(ISS). Its most notorious success is decreasing the cost of space flight compared to its competitors.
  - SpaceX advertises their Falcon 9 rocket with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, and much of the savings is because SpaceX can reuse the first stage.
- Problems you want to find answers
  - The objective of this project is to determine if the Falcon 9 first stage will land successfully.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data Collection through API
  - Data Collection with Web Scraping
- Perform data wrangling
  - Formatting Data and Training Labels
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistic Regression, Classification Trees, Support Vector Machine (SVM), K-Nearest Neighbors (KNN)

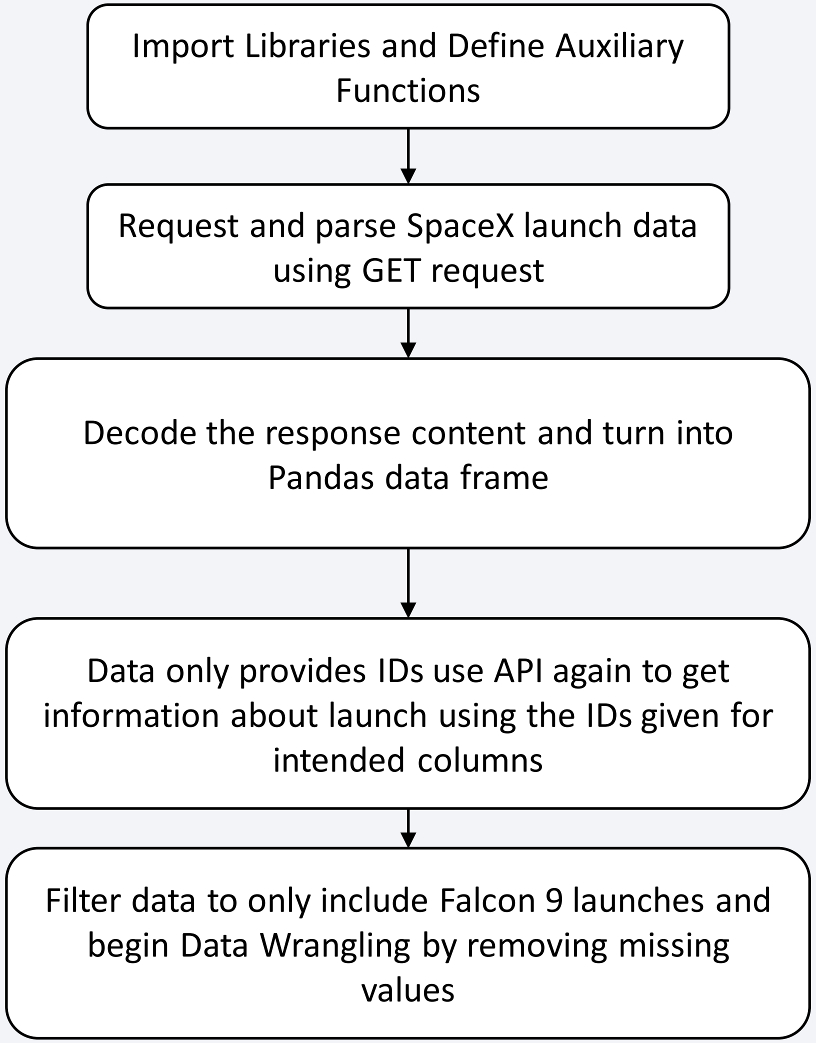
Data Collection

- Data Collection utilizing SpaceX API
  - Import Libraries and Define Auxiliary Functions
  - Request rocket launch data from SpaceX REST API URL
  - Parse and Clean data
- Data Collection with Web Scraping
  - Extract a Falcon 9 launch records HTML table from Wikipedia
  - Parse the table and convert it into a Pandas data frame

Data Collection – SpaceX API

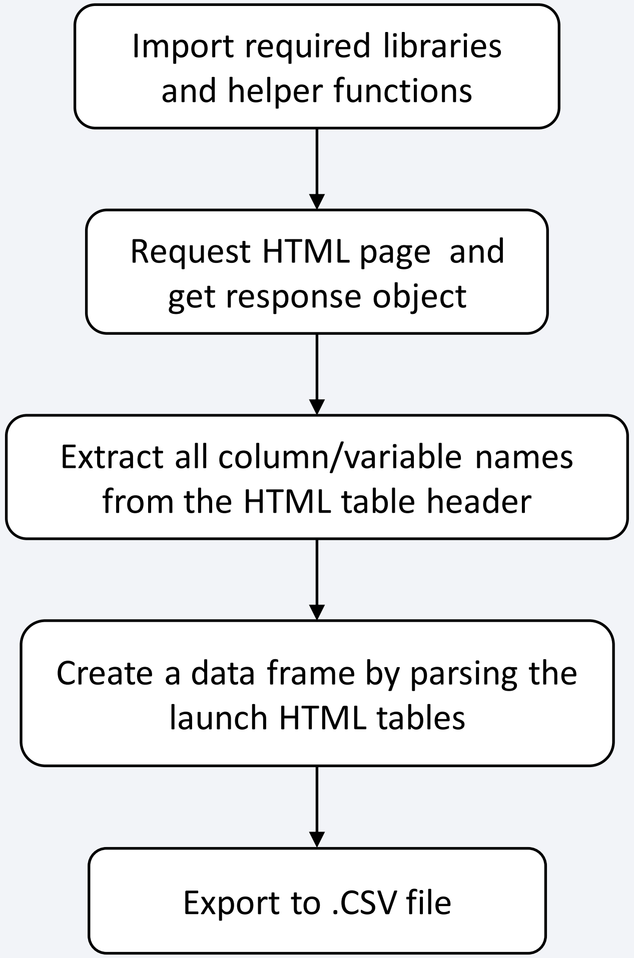
- Flowchart depicting data collection utilizing SpaceX REST API

- GitHub Link: https://github.com/cobo35/Data-Science-and-Machine-Learning-/blob/30495399a9873dd97e024c546ef81e481ca092a2/Data_Collection_API.ipynb



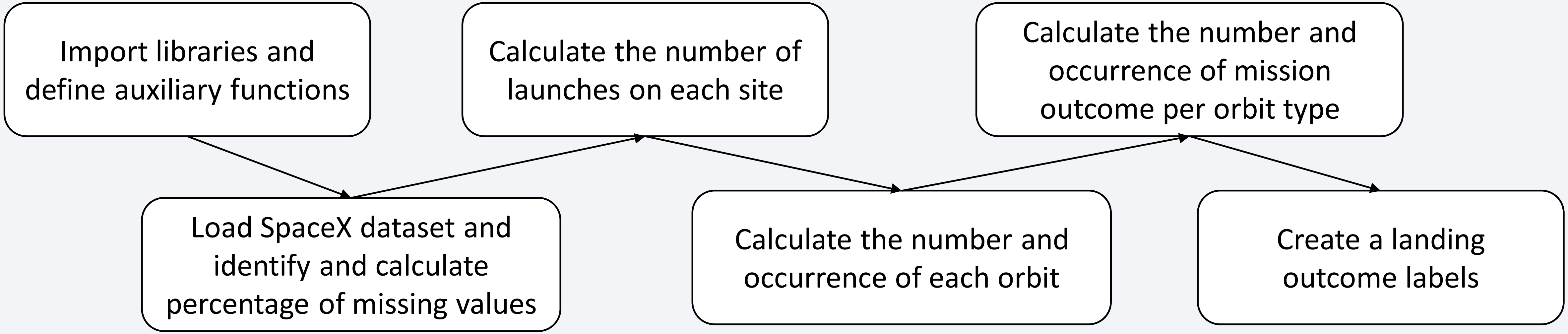
Data Collection – Web Scraping

- Flowchart depicting data collection Web scraping Falcon 9 and Falcon Heavy Launches Records from Wikipedia.
- Using Python Beautiful Soup library
- GitHub Link: https://github.com/cobo35/Data-Science-and-Machine-Learning-/blob/30495399a9873dd97e024c546ef81e481ca092a2/Web_Scraping.ipynb



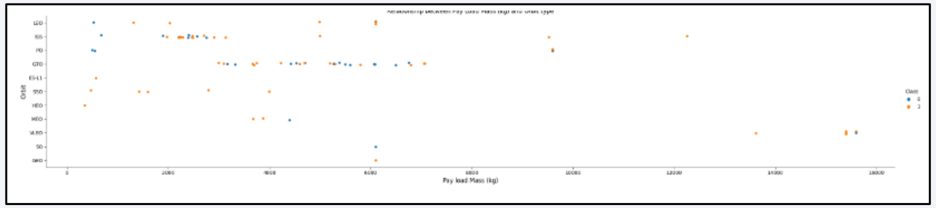
Data Wrangling

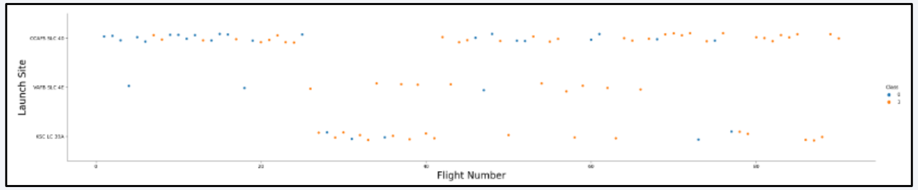
- Performing Exploratory Data Analysis (EDA) to find patterns in the data and determine what would be the label for training supervised models. The data set has several different cases where the booster did or did not land successfully. With data wrangling we will convert the different cases into training labels to identify successful and unsuccessful landings.
- GitHub Link: https://github.com/cobo35/Data-Science-and-Machine-Learning-/blob/30495399a9873dd97e024c546ef81e481ca092a2/Data_Wrangling.ipynb

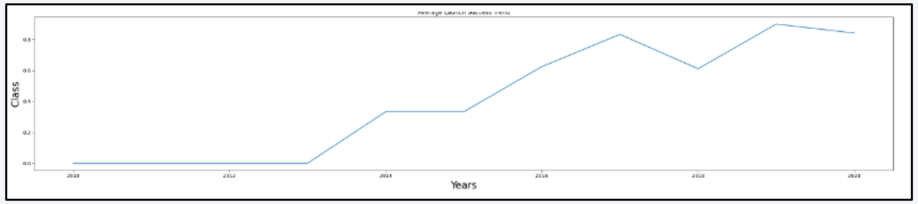


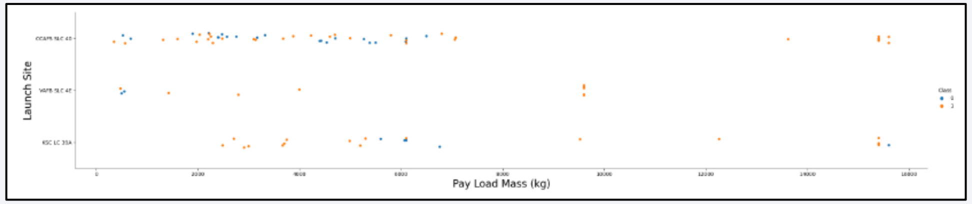
EDA with Visualization

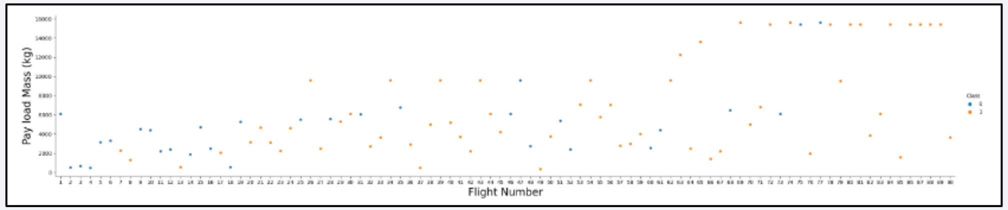
- Utilized scatter plots to show relationship between two variables and bar chart to compare metric values across different subgroups.
- GitHub Link: https://github.com/cobo35/Data-Science-and-Machine-Learning-/blob/30495399a9873dd97e024c546ef81e481ca092a2/EDA_with_Visualization.ipynb

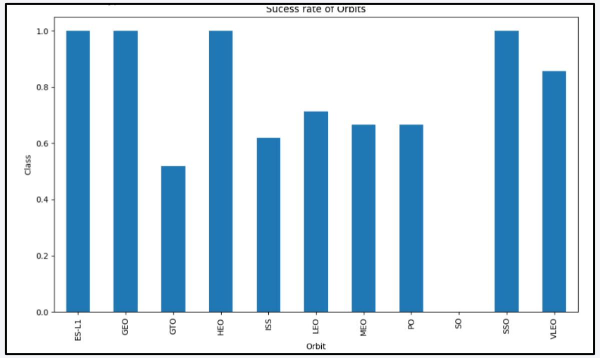








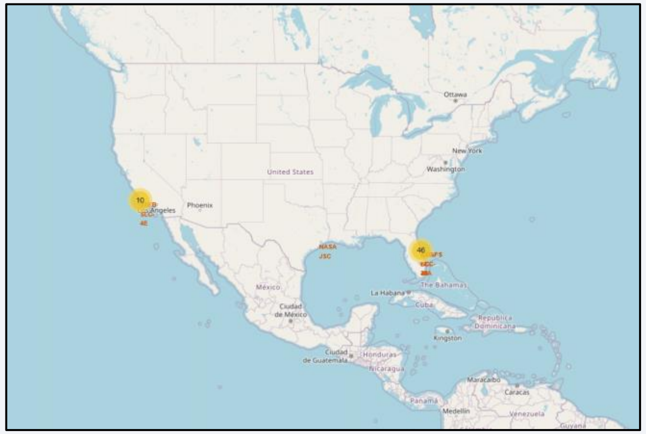




EDA with SQL

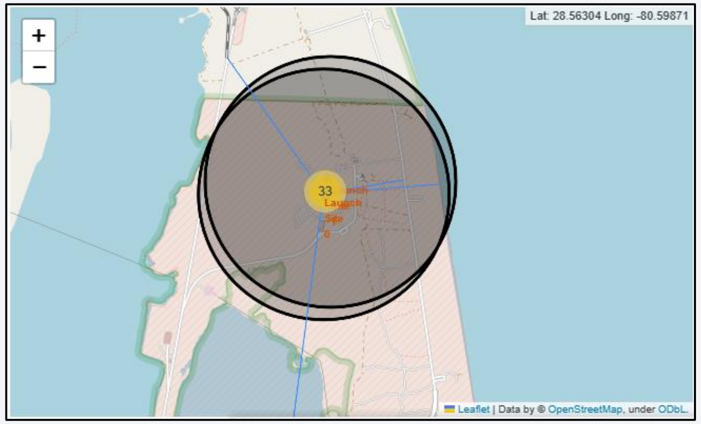
- Connect to the Database
- Display the names of the unique launch sites in the space mission
- Displayed specific records of the data set for review for clearer understanding of the dataset.
  - Displayed the total payload mass carried by boosters launched by NASA (CRS), average payload mass carried by booster version
- Listed the date where the successful landing outcome in drone ship was achieved and the names of the boosters which had success in ground pad and had payload mass greater between 4000 – 6000 kg
- Listed the total number of successful and failure mission outcomes
- Ranked the count of successful landing outcomes
- GitHub Link: https://github.com/cobo35/Data-Science-and-Machine-Learning-/blob/30495399a9873dd97e024c546ef81e481ca092a2/EDA_with_SQL.ipynb

Interactive Map with Folium



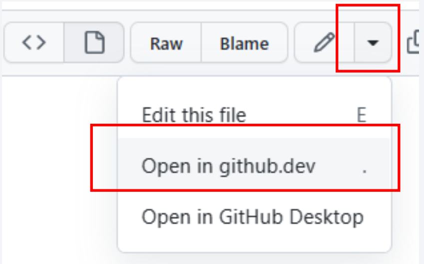
Depicted all launch sites on a map.

Marked success and failed launches to proper label



Added Vector Line to railway, highway, and city.

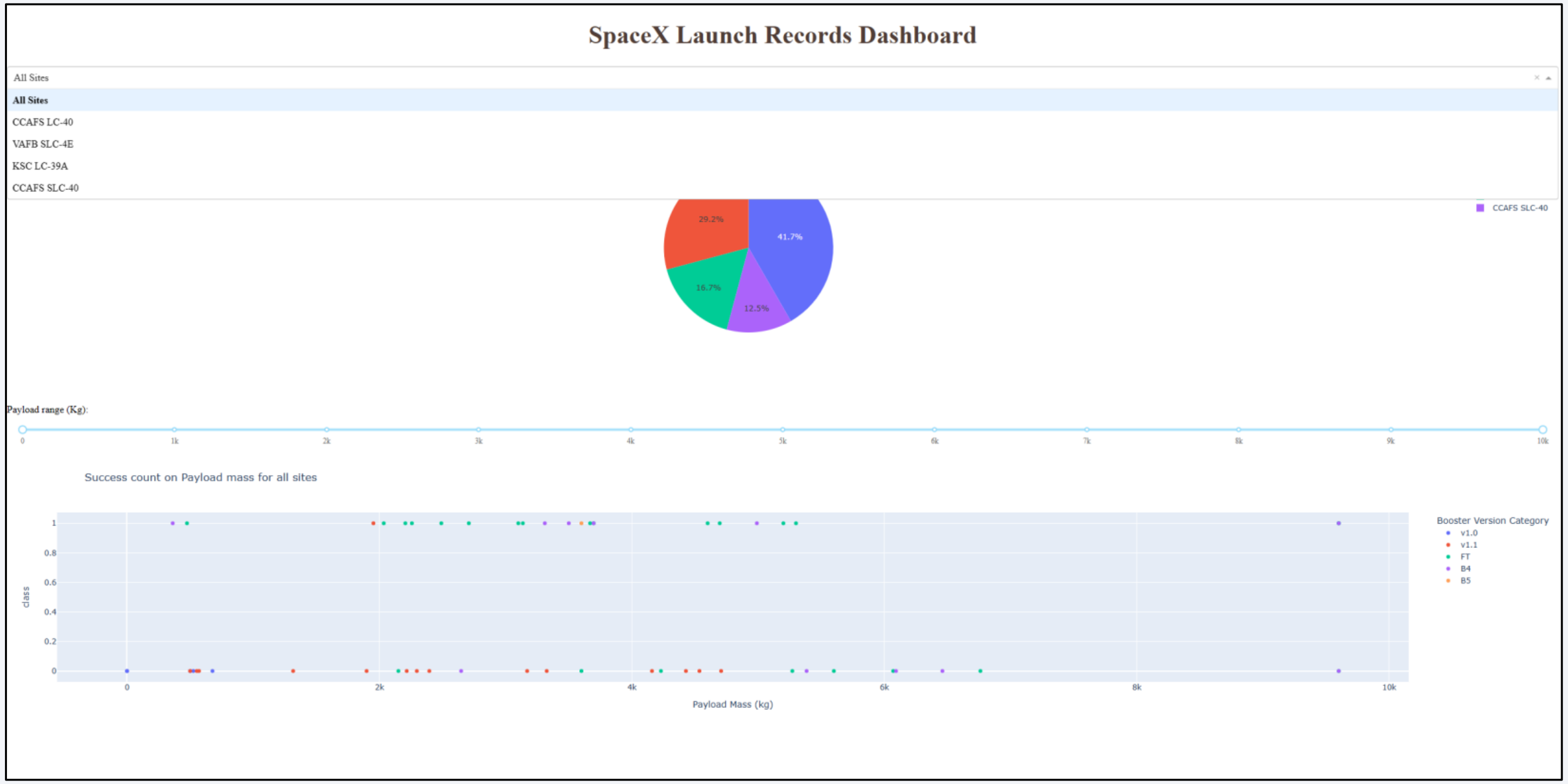
------- !Issue displaying full code in GitHub!------ to view select Dropdown and open in github.dev--------------------------------------→



- GitHub Link: https://github.com/cobo35/Data-Science-and-Machine-Learning-/blob/30495399a9873dd97e024c546ef81e481ca092a2/Interactive_Visual_Analytics_with_Folium.ipynb

Build a Dashboard with Plotly Dash

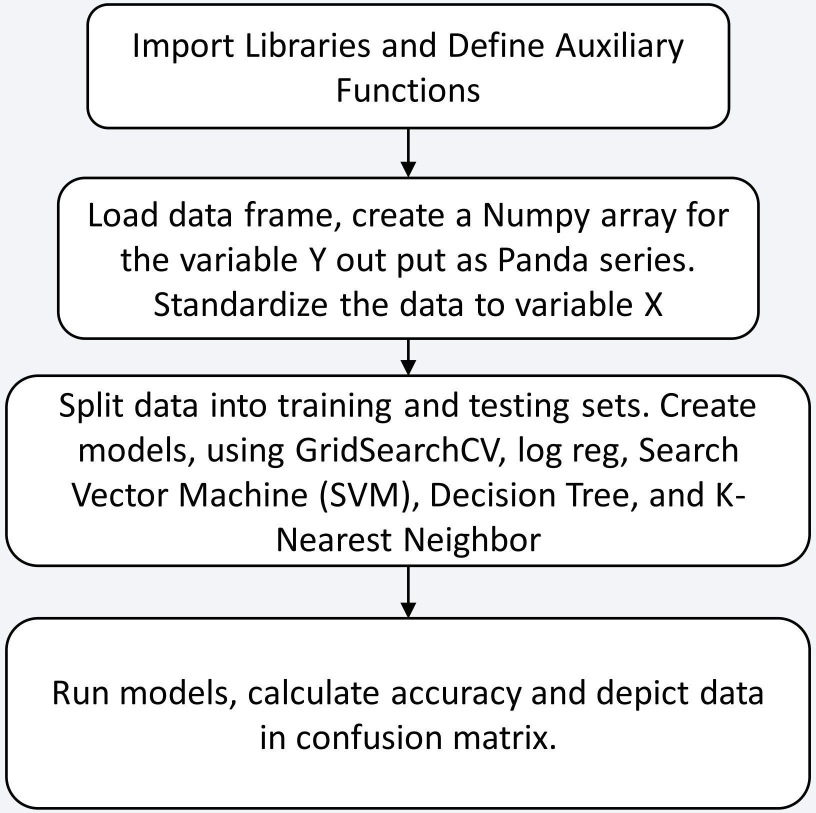
- Added dropdown list for launch site selection.
- Added pie chart breaking down successful launch sites and classes.
- Added a scatter plot to depict success count of payload mass for all sites.



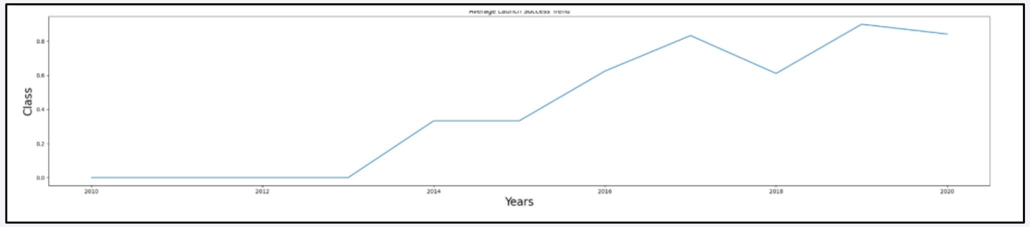
- GitHub Link: https://github.com/cobo35/Data-Science-and-Machine-Learning-/blob/30495399a9873dd97e024c546ef81e481ca092a2/DashPlottly.ipynb

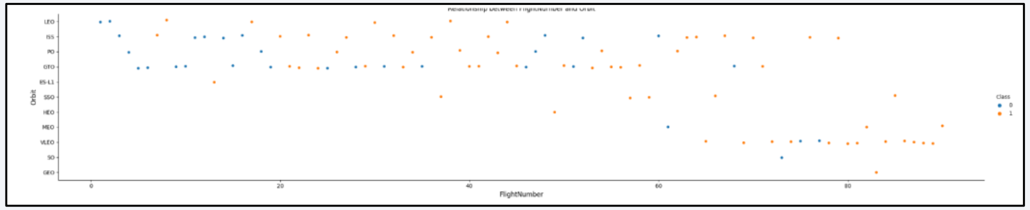
Predictive Analysis (Classification)

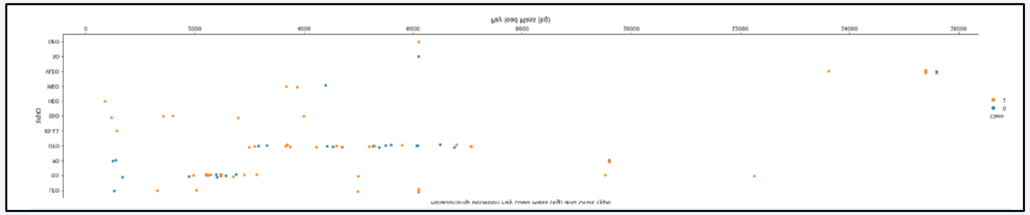
- Machine Learning Prediction Flow Chart
- GitHub Link: https://github.com/cobo35/Data-Science-and-Machine-Learning-/blob/7dd2c29d6e6e3708e1f358762e7ad7af6ec58cbe/Machine%20Learning%20Predictions.ipynb

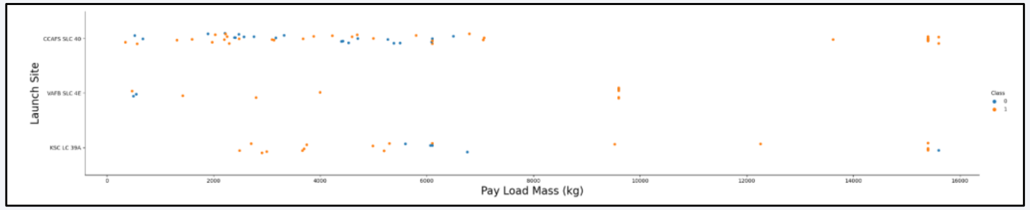


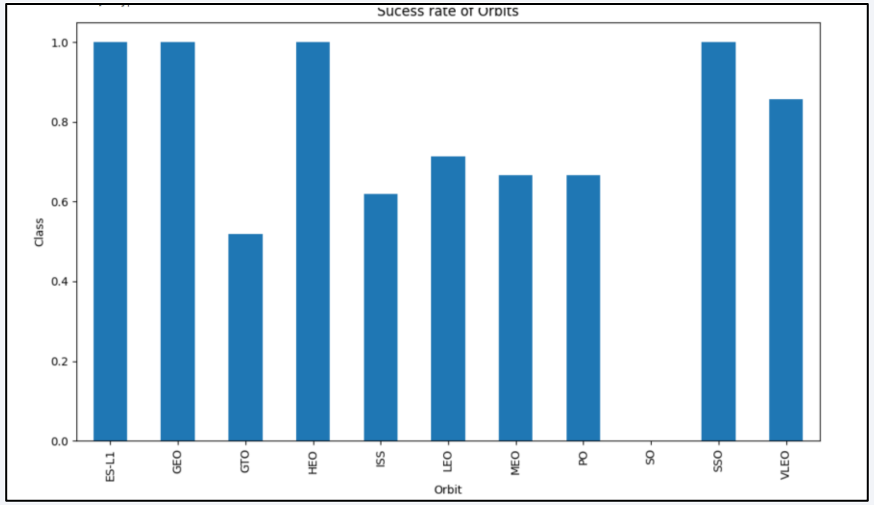
Exploratory Data Analysis Results in Screenshots

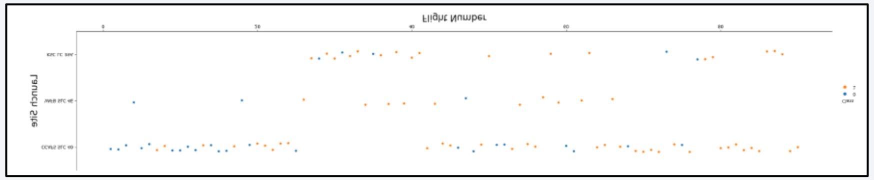


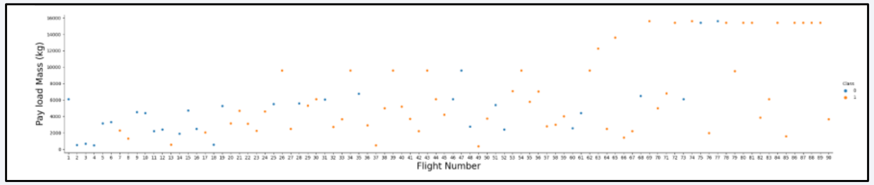




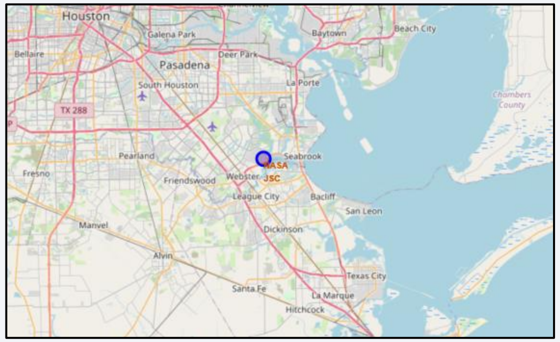


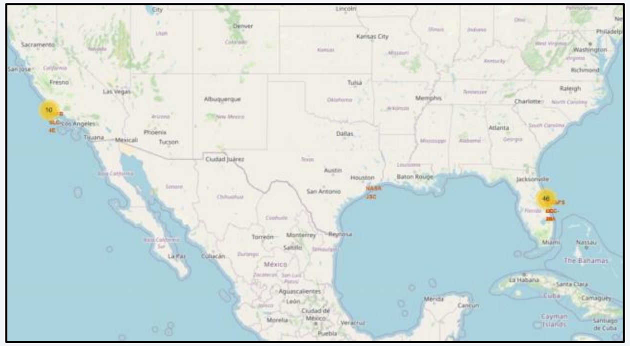


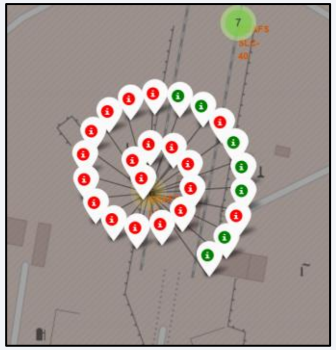


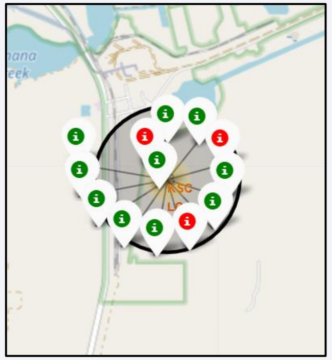


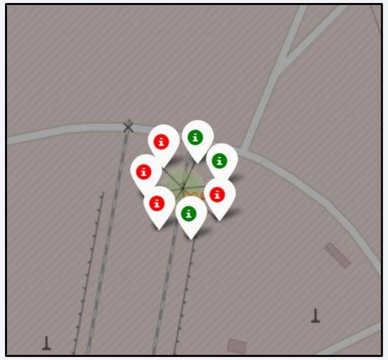
Interactive Analytics Demo in Screenshots

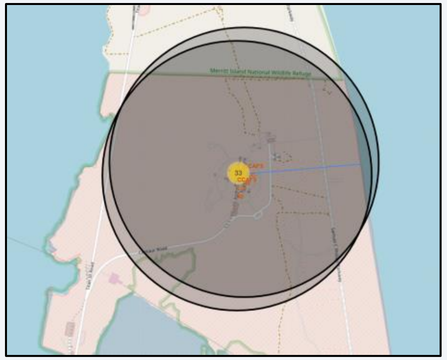


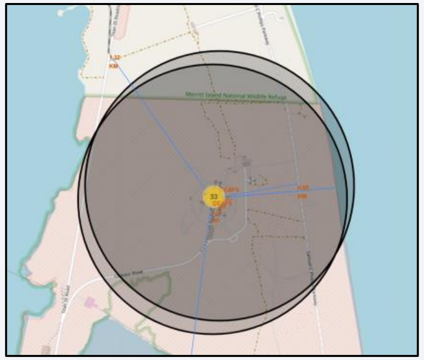


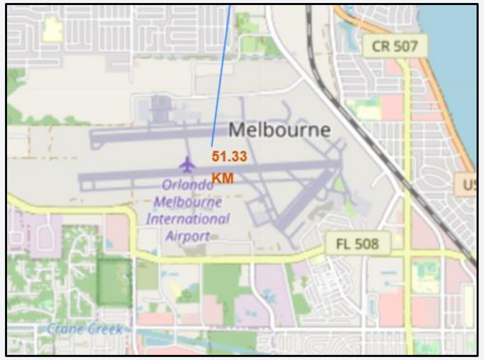






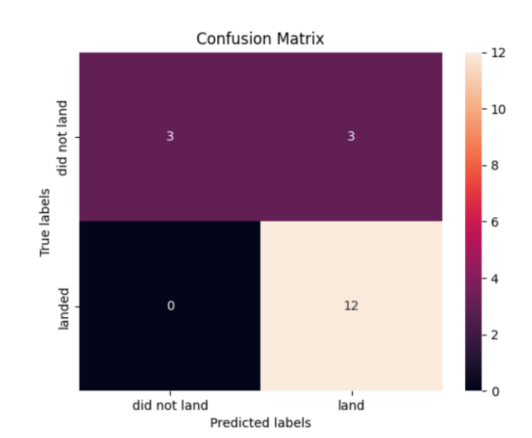




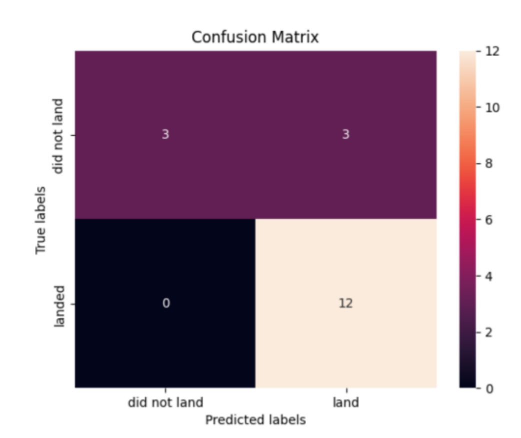


Predictive Analysis Results

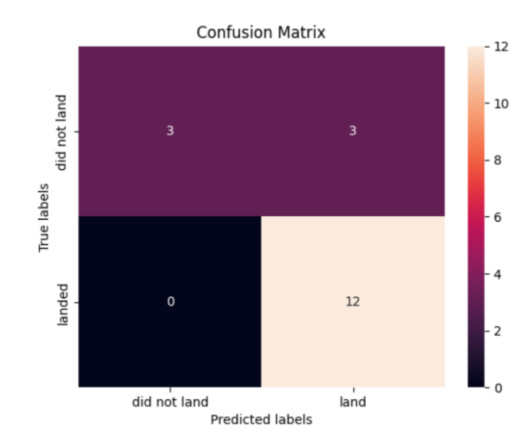
```
yhat = knn_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```




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

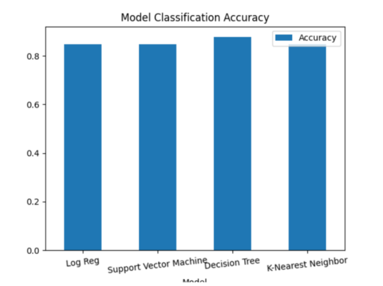



```
yhat=logreg_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```




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

	Accuracy
Log Reg	0.833333
Support Vector Machine	0.833333
Decision Tree	0.833333
K-Nearest Neighbor	0.833333




```
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
```

```
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
```

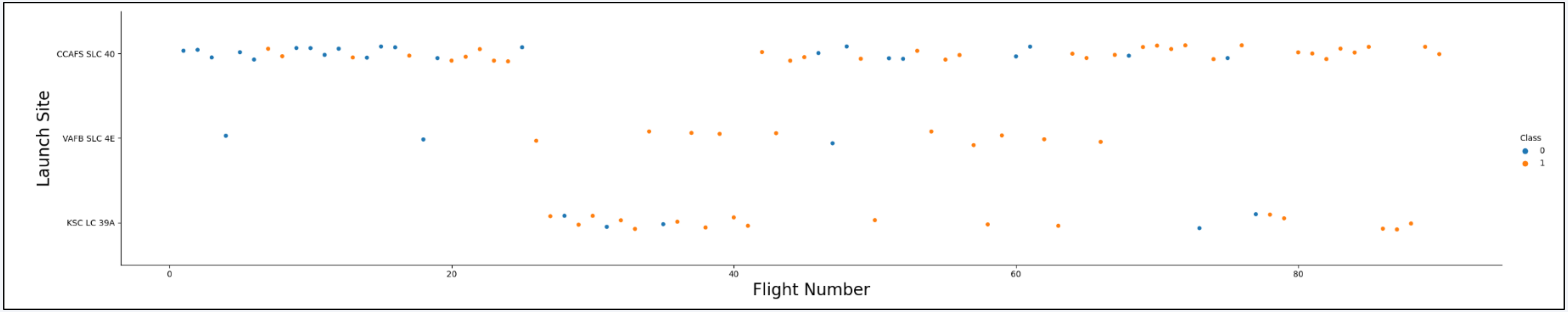
```
print("tuned hpyerparameters :(best parameters) ",svm_cv.best_params_)
print("accuracy :",svm_cv.best_score_)

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

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

tuned hpyerparameters :(best parameters)  {'criterion': 'gini', 'max_depth': 6, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 5, 'splitter': 'bes
t'}
accuracy : 0.8767857142857143
```

Section 2

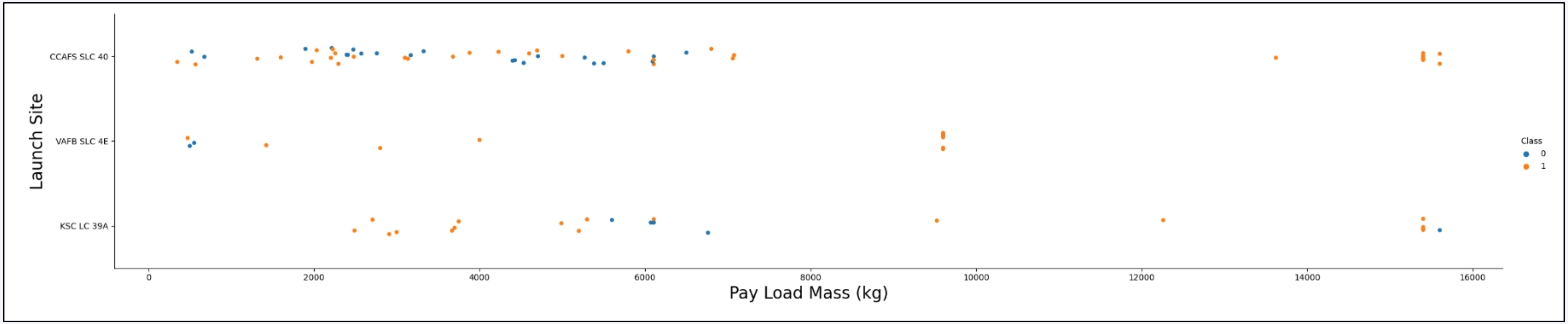
Insights drawn from EDA

Flight Number vs. Launch Site



- The scatter plot depicts that site CCAFS SLC 40 has the most consistency in usage and most flight number. While site VAFB SLC 4E has least usage and flight number. Site KSC LC 39A usage seems to algin with site CCAFS SLC 40 non-use.

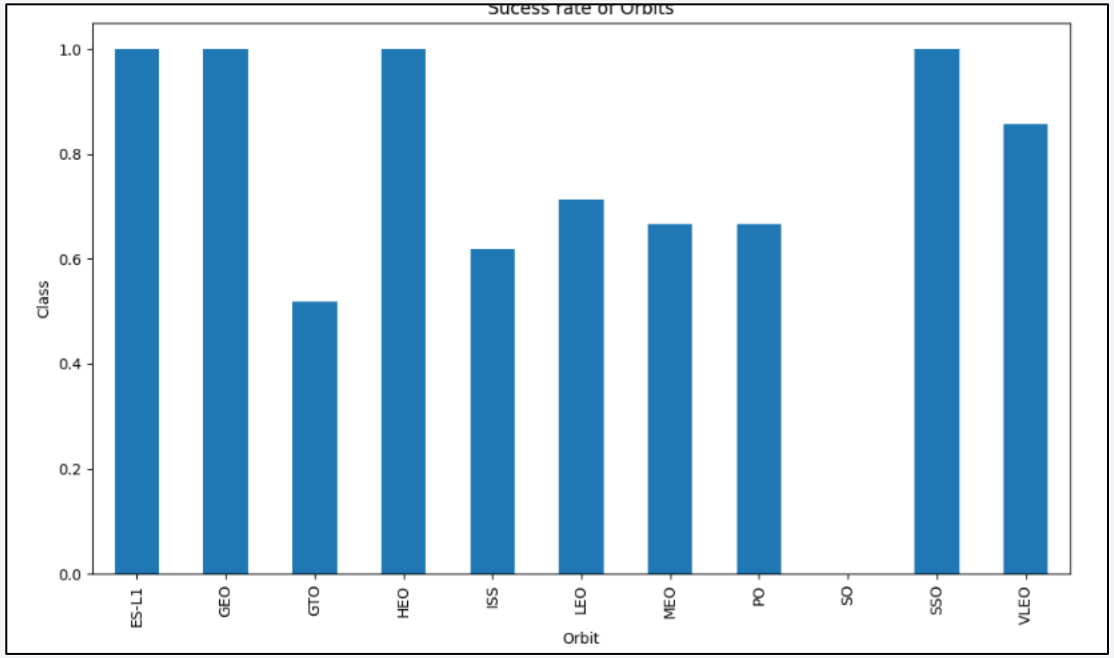
Payload vs. Launch Site



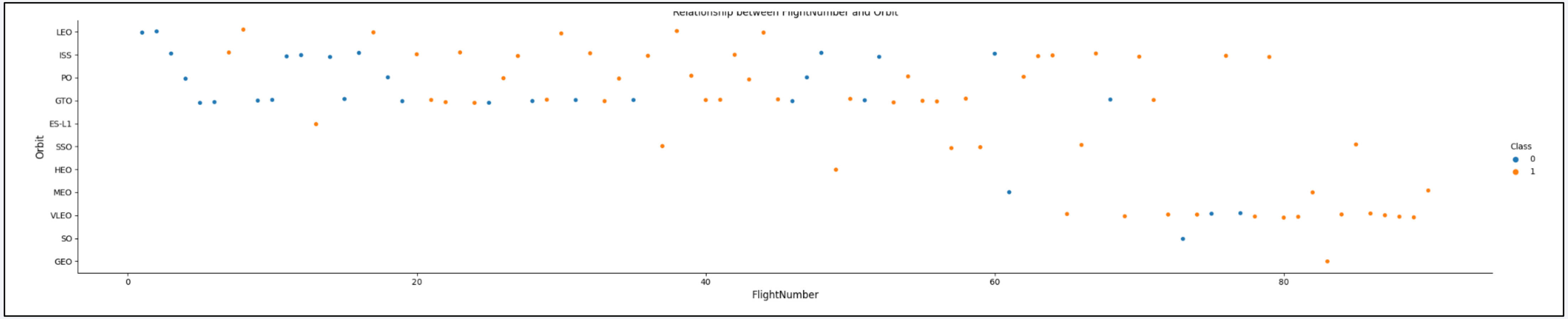
- At launch site VAFB SLC 4E there are no rockets launched for heavy pay load mass greater than 10000 kg.
- Launch site CCAFS SLC 40 is the most used with the site KSC LC 39A being the second most used.

Success Rate vs. Orbit Type

- Show a bar chart for the success rate of each orbit type
- Show the screenshot of the scatter plot with explanations

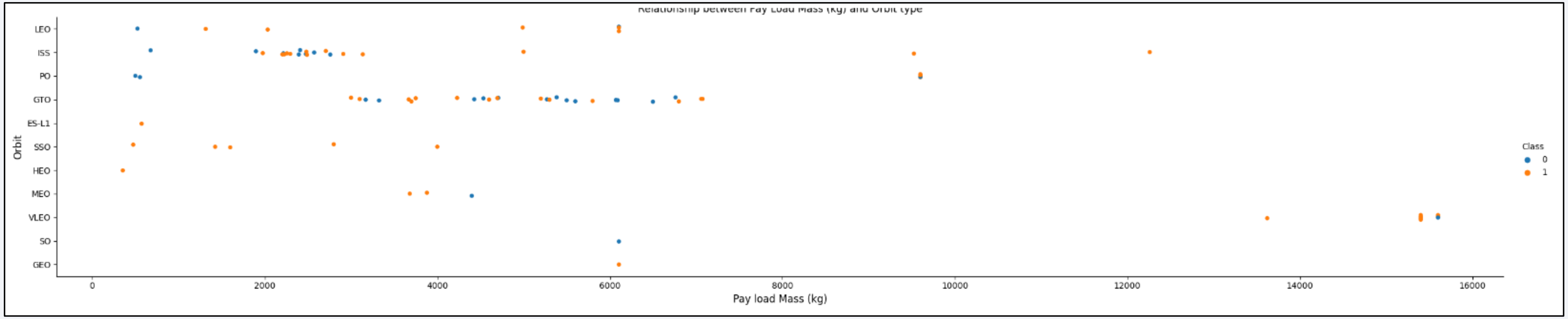


Flight Number vs. Orbit Type



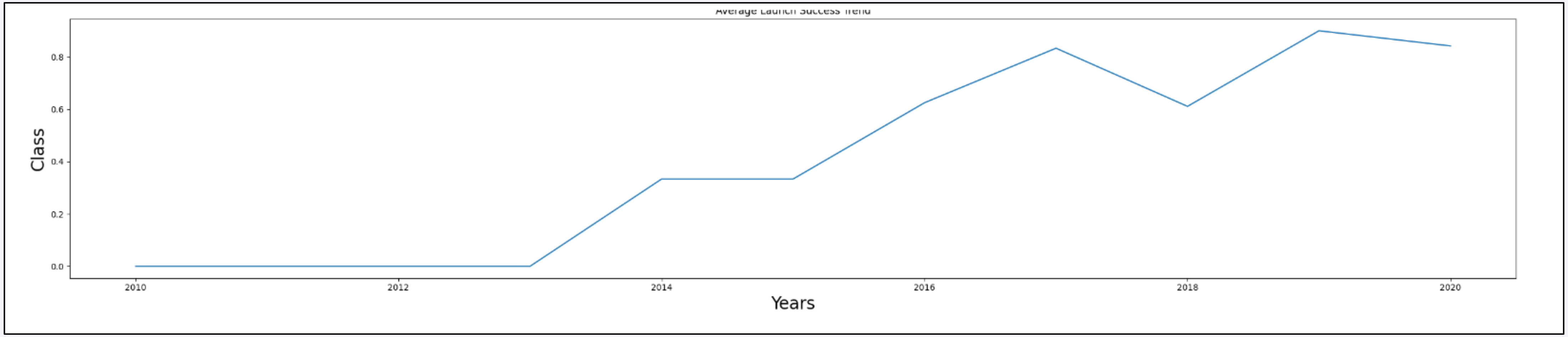
- The LEO orbit the success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

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



- It can be observed that overall since 2013 the success rate has increased till 2020 with minor dips seen in 2018.

All Launch Site Names

```
In [117…  %sql select distinct "Launch_Site" from "SPACEXTBL"
```

- Utilizing query displayed above. I displayed the names of the unique launch sites in the space mission.

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

Launch Site Names Begin with 'KSC'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
19-02-2017	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
16-03-2017	06:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
30-03-2017	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
01-05-2017	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
15-05-2017	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

```
%sql select * from "SPACEXTBL" where "Launch_Site" like 'KSC%' limit 5
```

- With the query above I displayed five records of the launch sites that begin with ‘KSC’.

Total Payload Mass

```
%sql select sum("PAYLOAD_MASS__KG_") as sum from "SPACEXTBL" where "Customer" like 'NASA (CRS)'
```

```
 * sqlite:///my_data1.db
Done.
```

sum
45596

- With the query above I calculated the total payload mass carried by boosters launched by NASA (CRS).

Average Payload Mass by F9 v1.1

```
%sql select avg("PAYLOAD_MASS__KG_") as Average from "SPACEXTBL" where "Booster_Version" like 'F9 v1.1'
```

```
 * sqlite:///my_data1.db
Done.
```

Average
2928.4

- With the query above I calculated the average payload mass carried by the booster version F9 v1.1.

First Successful Drone Ship Landing Date

List the date where the succesful landing outcome in drone ship was acheived.

Hint:Use min function

```
print('Real min date is 08-04-2016. It should be reading as DD-MM-YYYY')
%sql select min("Date") from "SPACEXTBL" where "Landing _Outcome" like 'Success (drone ship)'
```

```
Real min date is 08-04-2016. It should be reading as DD-MM-YYYY
 * sqlite:///my_data1.db
Done.
```

min("Date")
06-05-2016

- With the query above I observed the date when the first successful landing outcome in drone ship was achieved.

Successful Ground Pad Landing with Payload between 4000 and 6000

List the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000

```
%sql select "Booster_Version" from "SPACEXTBL" where ("Landing _Outcome" like 'Success (ground pad)') and ("PAYLOAD_MASS__KG_" between 4000 and 6000)
```

```
 * sqlite:///my_data1.db
Done.
```

Booster_Version
F9 FT B1032.1
F9 B4 B1040.1
F9 B4 B1043.1

- The query above presents the names of booster which have successfully landed on ground pad and had a pay load mass greater than 4000 but less than 6000.

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
%sql select "Mission_Outcome", count(*) from "SPACEXTBL" group by "Mission_Outcome"
```

```
 * sqlite:///my_data1.db
Done.
```

Mission_Outcome	count(*)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- With the query above the total number of successful and failure mission outcomes is presented. 100 successful and 1 failure mission outcomes.

Boosters Carried Maximum Payload

```
%sql select "Booster_Version" from "SPACEXTBL" where "PAYLOAD_MASS__KG_" = (select MAX("PAYLOAD_MASS__KG_") from "SPACEXTBL")
```

```
 * sqlite:///my_data1.db
Done.
```

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 query above lists the names of the booster version that have carried the maximum payload mass.

2017 Launch Records

```
%sql select "Date", "Landing _Outcome", "Booster_version", "Launch_Site" from "SPACEXTBL" where
("Landing _Outcome" like 'Success (ground pad)') and ("Date" like '%%-%%-2017')
```

```
 * sqlite:///my_data1.db
Done.
```

Date	Landing _Outcome	Booster_Version	Launch_Site
19-02-2017	Success (ground pad)	F9 FT B1031.1	KSC LC-39A
01-05-2017	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
03-06-2017	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
14-08-2017	Success (ground pad)	F9 B4 B1039.1	KSC LC-39A
07-09-2017	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A
15-12-2017	Success (ground pad)	F9 FT B1035.2	CCAFS SLC-40

- With the query above, I’m displaying successful landing outcomes in ground pad for 2017 along with Dates, booster version, and launch site.

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

```
%sql select "Landing _Outcome", count(*) as count from "SPACEXTBL" where (date >= '04-06-2010' and date <= '20-03-2017') and
("Landing _Outcome" like 'Success%') group by "Landing _Outcome" order by "Landing _Outcome" DESC
```

```
 * sqlite:///my_data1.db
Done.
```

Landing _Outcome	count
Success (ground pad)	6
Success (drone ship)	8
Success	20

- Using the query above, I ranked the count of successful landing outcomes between the date 2010-06-04 and 2017-03-20.

Section 3

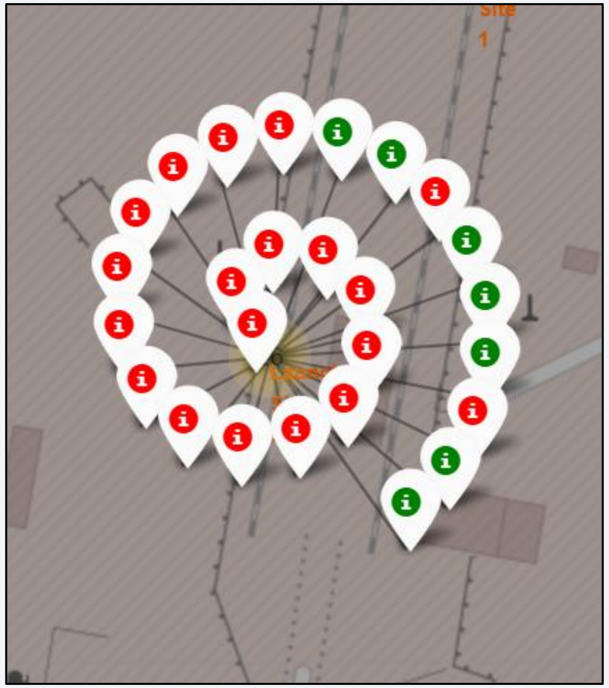
Launch Sites Proximities Analysis

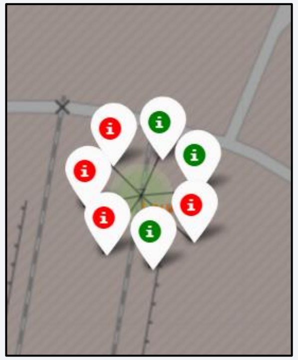
All Launch Sites on Global Map

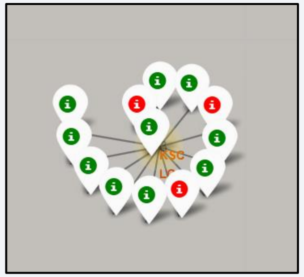
- All launch sites are located within the continental United States specifically in the states of Florida and California.

Launch Sites

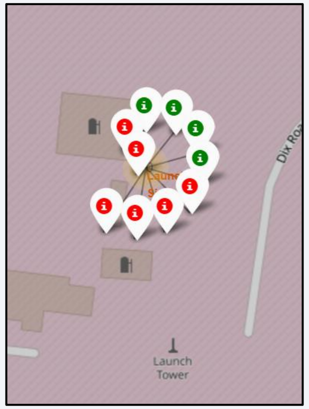
Florida







California



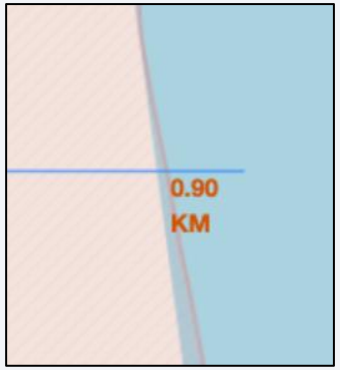
- **Red markers** indicate failed launches and **Green markers** indicate successful launches.

Launch Site Distance to Landmarks

Distance to City

Distance to railway

Distance to Highway



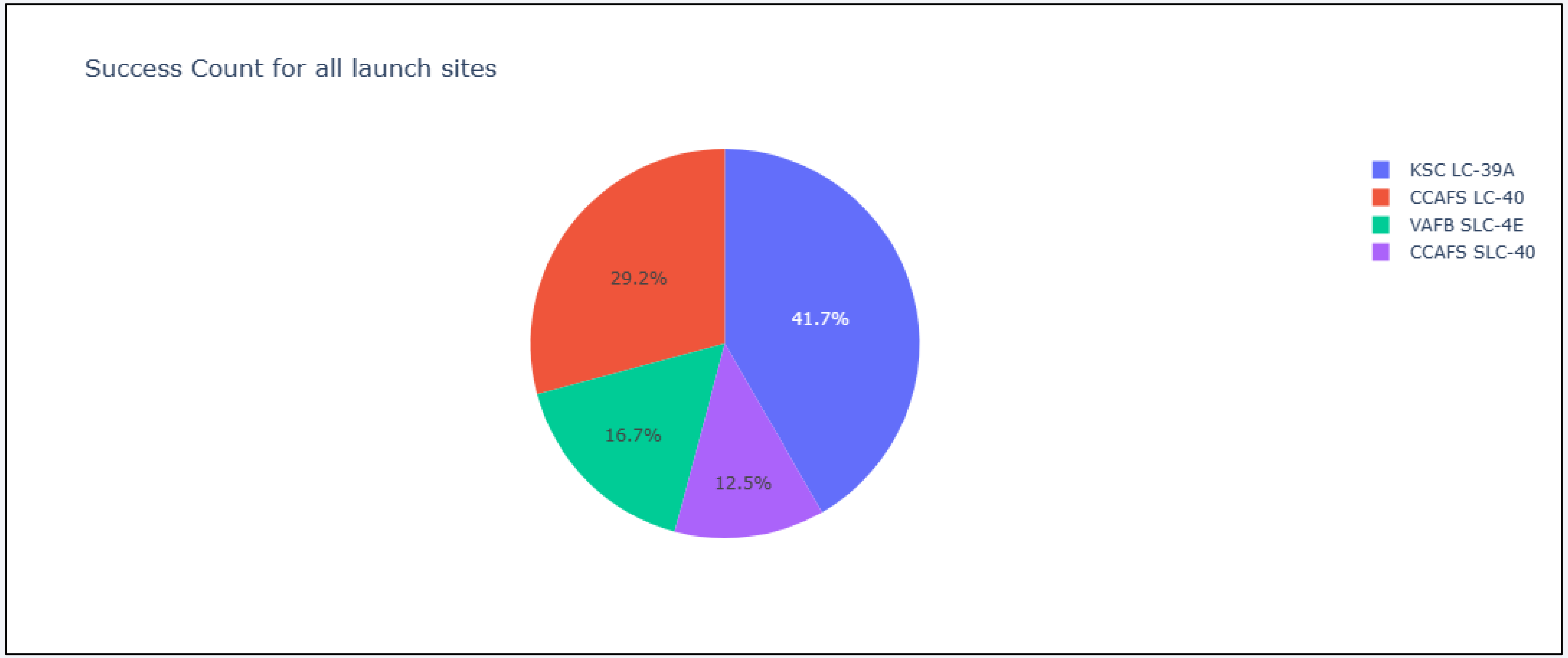
Distance to Coast

- Are launch sites in close proximity to railways? Yes
- Are launch sites in close proximity to highways? Yes
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes

Section 4

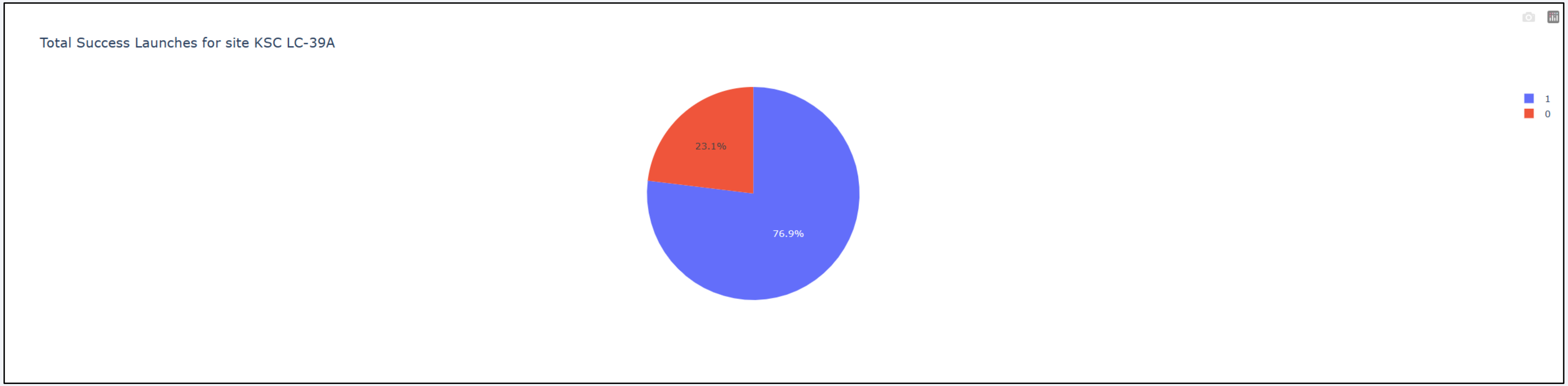
Build a Dashboard with Plotly Dash

Pie Chart Launch Success Count



- The chart above depicts that launch site KSC LC-39A had the most successful launches from all other sites. Followed then by CCAFS LC-40, VAFB SLC-4E and lasty by CCAFS SLC-40

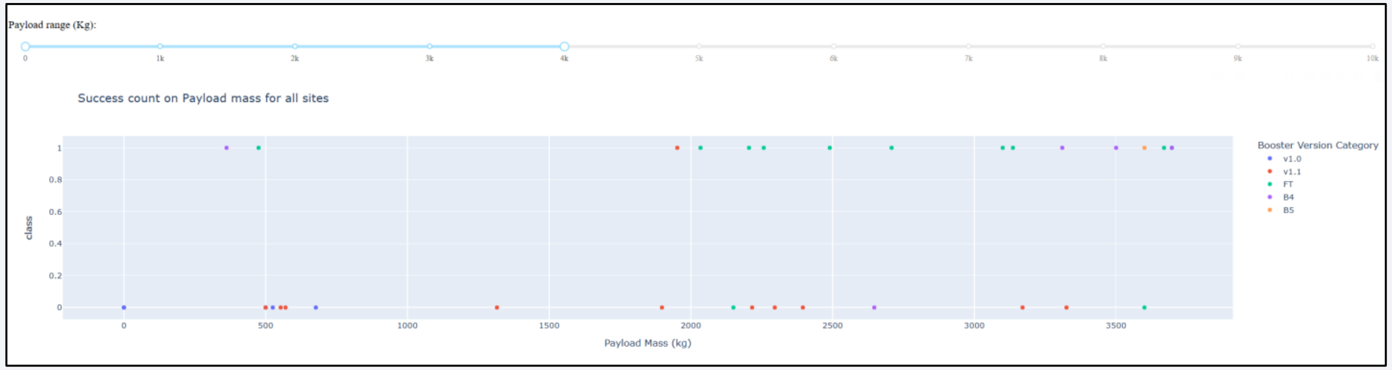
Pie Chart Depicting Highest Launch Site Success Ratio



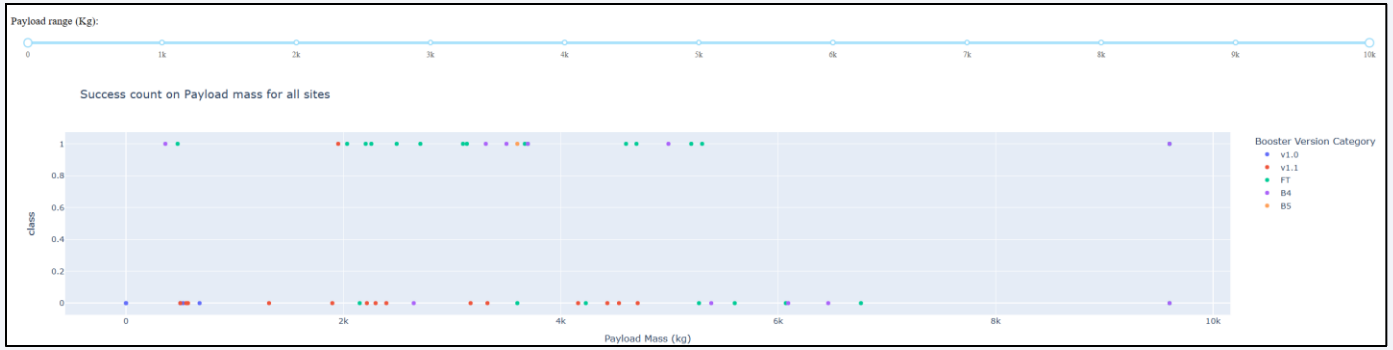
- Launch site KSC LC-39A had the higest success rate of 76.9% and the lowest failure rate of 23.1%

Payload vs. launch outcome scatter plot with different weight selected

Payload range (Kg): 0 - 4000



Payload range (Kg): 0 - 10000



- The scatter plot depicts that success rates are higher in the low weighted range between 2000 – 5000 rather than the heavy weighted ranges.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

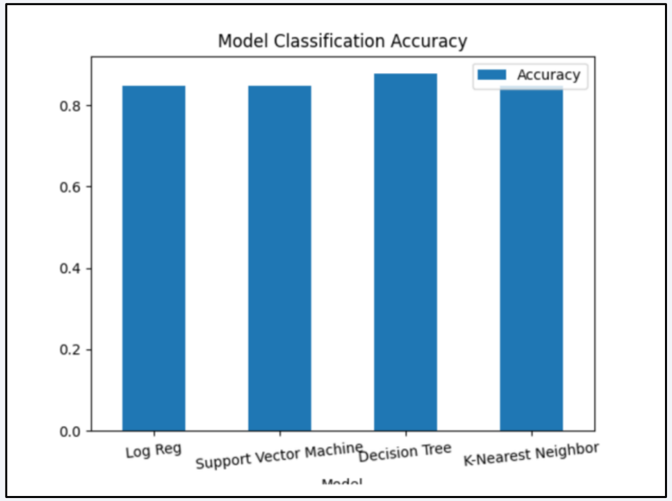
```
parameters = {'criterion': ['gini', 'entropy'],
     'splitter': ['best', 'random'],
     'max_depth': [2*n for n in range(1,10)],
     'max_features': ['auto', 'sqrt'],
     'min_samples_leaf': [1, 2, 4],
     'min_samples_split': [2, 5, 10]}

tree = DecisionTreeClassifier()
```

```
tree_cv = GridSearchCV(tree, parameters, cv=10)
tree_cv.fit(X_train, Y_train)
```

```
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']})
```

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

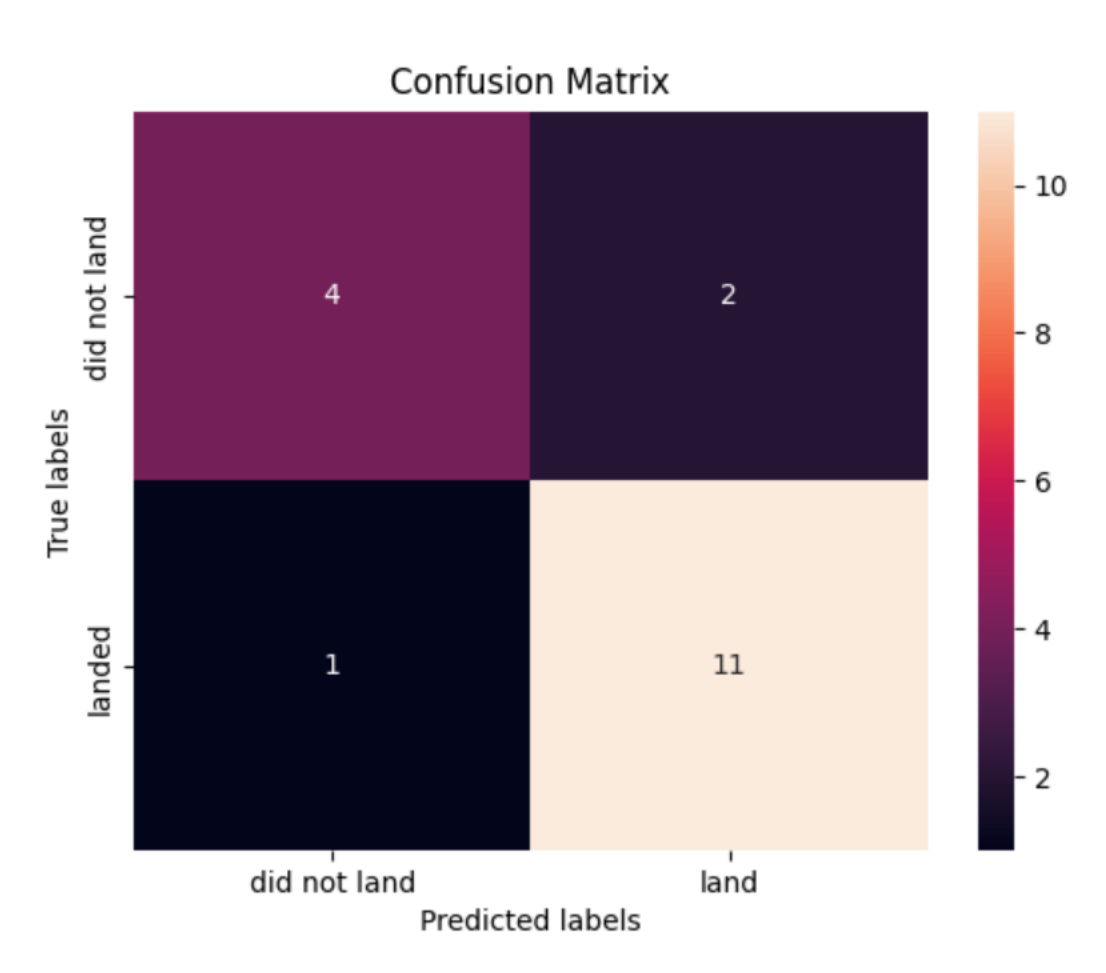
```
tuned hpyerparameters :(best parameters)  {'criterion': 'gini', 'max_depth': 6, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 5, 'splitter': 'bes
t'}
accuracy : 0.8767857142857143
```



- The Decision Tree model has the highest classification accuracy.

Confusion Matrix

- The decision tree model had the best accuracy. Distinguishing between different classes the matrix to the side presents 1 = FN, 11= TN, 2=FP, 4=TP.



Conclusions

- Launch site KSC LC-39A had the most successful launches.
- The Decision Tree model presents the most accurate results to be utilized in a machine learning model to predict if the first stage will land.
- There is a correlation between successful launches and payload mass (kg) between the ran ge of 2000 – 5000 kg
- Launch success rate have increased since 2013

Thank you!