Winning Space Race with Data Science

Shreyash Sabde
2/12/2021

Outline

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

Executive Summary

❑ SUMMARY OF METHODOLOGIES

1) DATA COLLECTION

2) DATA WRANGLING

3) EDA WITH DATA VISUALIZATION

4) EDA WITH SQL

5) BUILDING WITH AN INTERACTIVE MAP WITH FOLIUM

6) BUILDING DASHBOARD WITH PLOTLY DASH

7) PREDICTIVE ANALYSIS (CLASSIFICATION)

❑ SUMMARY OF ALL RESULTS

1) EXPLORATORY DATA ANALYSIS RESULTS

2) INTERACTIVE ANALYTICS DEMO IN SCREENSHOTS

3) PREDICTIVE ANALYSIS RESULTS

Introduction

- Project background and context

1. We predicted if the falcon 9 first stage will land successfully. SpaceX advertises falcon 9 rocket launches on its website, with the cost of 62 million dollars, other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. 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 alternate company wants to bid against SpaceX for a rocket launch.

- Problems you want to find answers

1. What influences if the rocket will land successfully?
2. The effect each relationship with certain rocket variables will impact in determining the success rate of a successful landing.
3. What conditions does SpaceX have to achieve to get the best results and ensure the best rocket success landing rate.

Section 1

Methodology

Methodology

- Data collection methodology:

  SpaceX API

  Web Scraping

- Perform data wrangling

  One hot encoding data fields and dropping irreverent columns

- Perform exploratory data analysis (EDA) using visualization and SQL

  Plotting – Scatter plots, Bar graphs to show relationships between variables to show patterns of data.

- Perform interactive visual analytics using Folium and Plotly Dash

- Perform predictive analysis using classification models

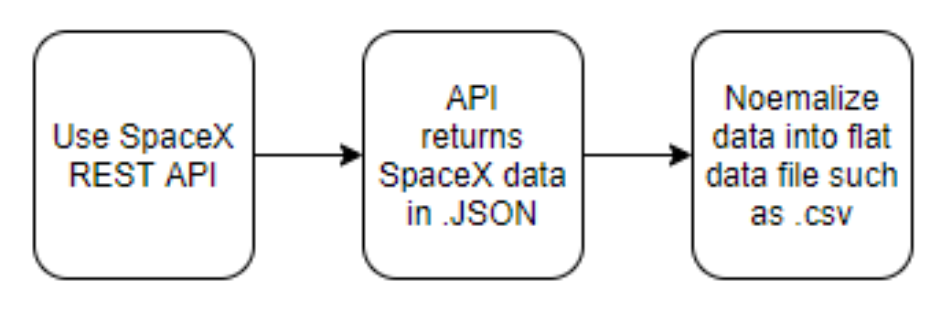
  How to build, tune, evaluate classification models

Data Collection

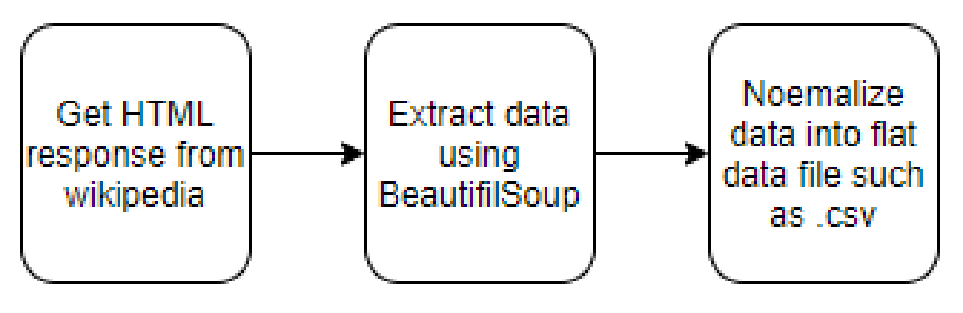
DESCRIBE HOW DATA SETS WERE COLLECTED.

- THE DATASETS ARE COLLECTED USING THE SPACEX REST API AND WEB SCRAPING WIKIPEDIA USING BEAUTIFULSOUP.
- USING THIS DATA WE WANT TO PREDICT WHETHER SPACEX WILL ATTEMPT TO LAND A ROCKET OR NOT.
- THE API GIVES THE INFORMATION ABOUT ROCKET USED, PAYLOAD DELIVERED, LAUNCH SPECIFICATIONS, LANDING SPECIFICATIONS AND LANDING OUTCOME.

Using the SpaceX REST API



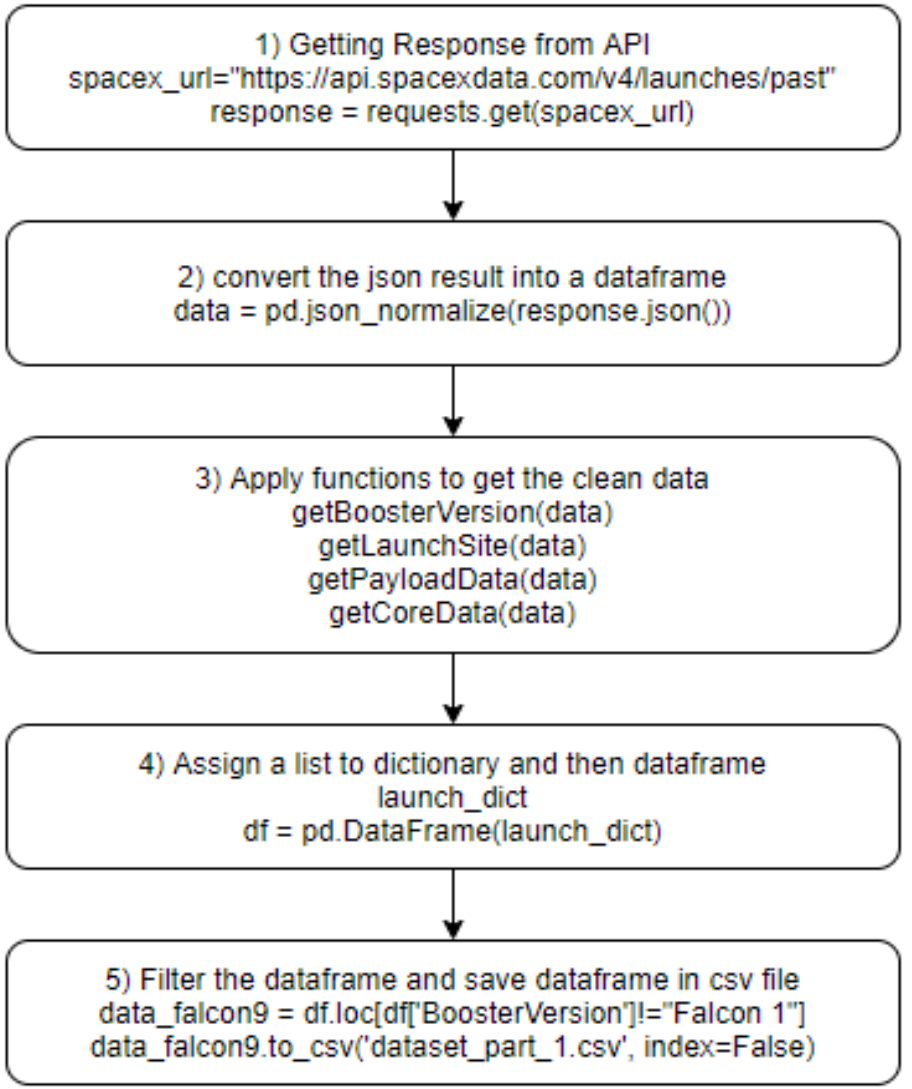
Web scraping using BeautifulSoup



Data Collection – SpaceX API

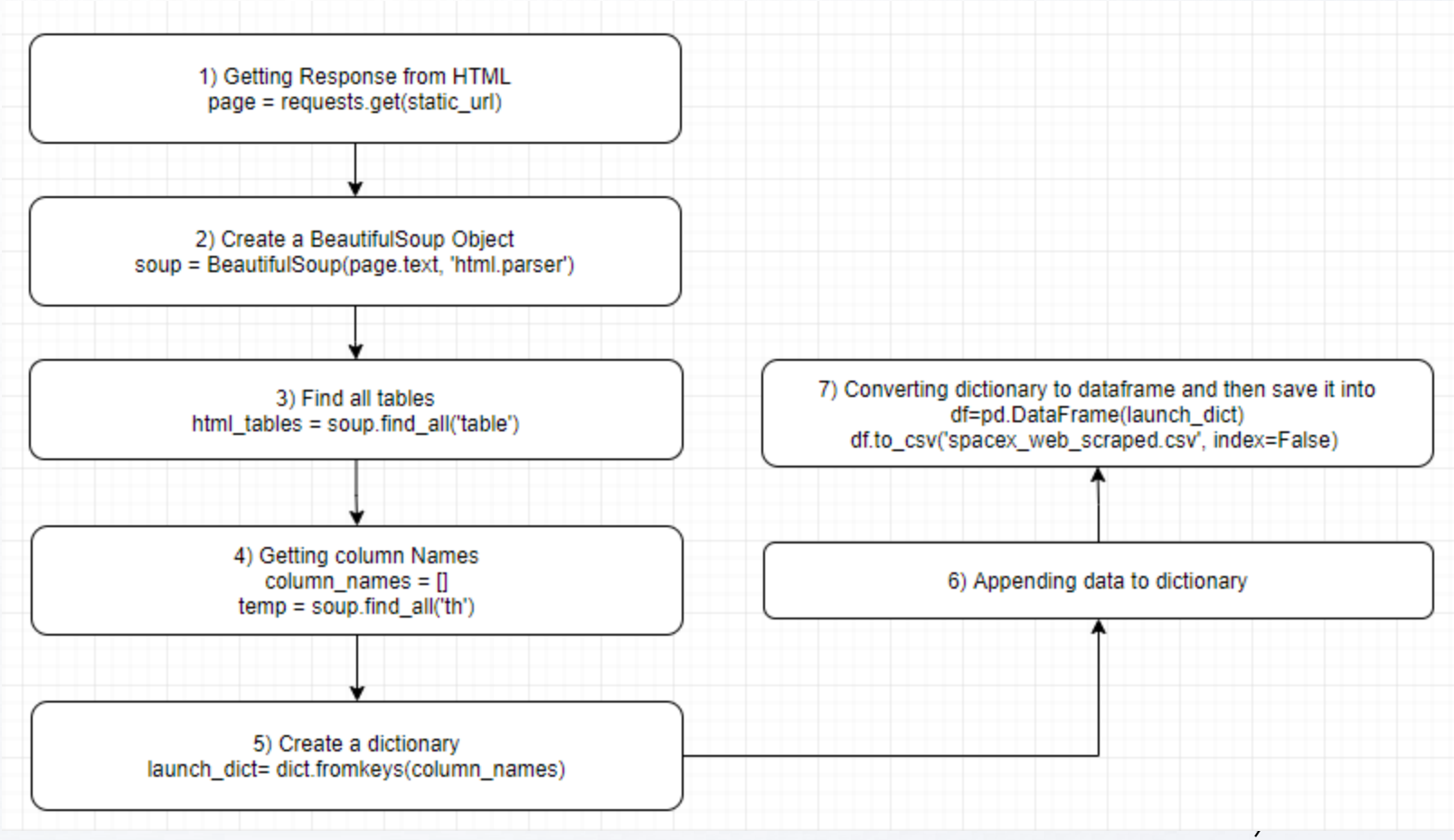
- DATA COLLECTION WITH SPACEX REST API

Data collection API



Data Collection - Scraping

- Data collection using web scraping method



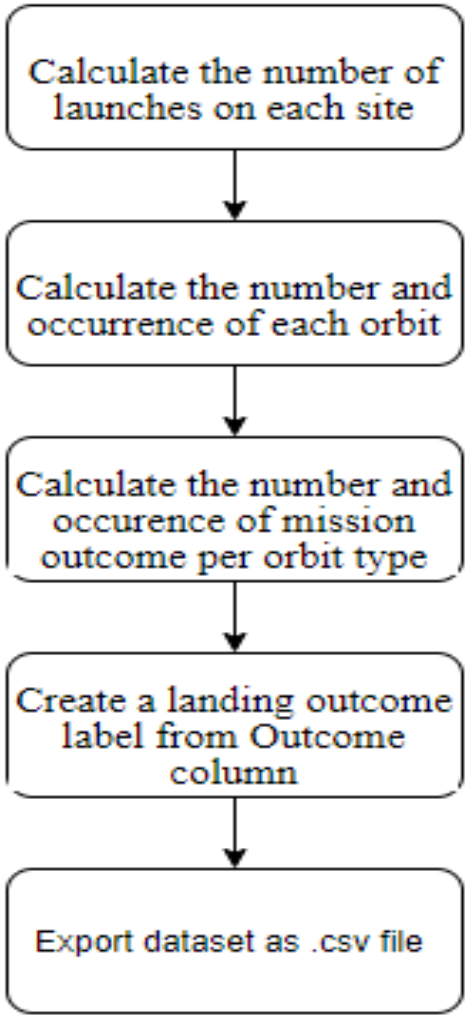
[Data collection Web Scraping]

Data Wrangling

- INTRODUCTION

1) There are several different cases where booster did not land successfully. sometimes, landing was attempted but failed due to accident.
2) Based on the different outcomes, we classify them as good or bad outcome and convert those outcomes into training labels with 1 as successfully landed, 0 means landing was unsuccessful.
3) Landing successful (outcomes) = true ASDS, true RTLS, true ocean
4) Landing unsuccessful (outcomes) = NONE NONE, false ASDS, false ocean, none ASDS, false RTLS

Data Wrangling



EDA with Data Visualization

- SUMMARIZE WHAT CHARTS WERE PLOTTED AND WHY YOU USED THOSE CHARTS
- SCATTER PLOTS

1) FLIGHT NUMBER VS PAYLOAD MASS
2) FLIGHT NUMBER VS LAUNCH SITE
3) PAYLOAD VS LAUNCH SITE
4) ORBIT VS FLIGHT NUMBER
5) PAYLOAD VS ORBIT TYPE
6) ORBIT VS PAYLOAD MASS

- scatter plots show how much one variable is affected by another. the relationship between two variables is called their correlation. scatter plots usually consists of a large body of data.

EDA with Data Visualization

- BAR GRAPH

1) SUCCESS RATE VS ORBIT

- a bar diagram makes it easy to compare sets of data between different groups at a glance. the goal is to show relationship between two axes. also shows big change in data over time.

- LINE GRAPH

1) SUCCESS RATE VS YEAR

- line graphs are useful in that they show data variables and trends very clearly and can help to make predictions about result of data not recorded yet.

EDA with SQL

PERFORMED SQL QUERIES TO GATHER INFORMATION ABOUT THE DATASET.

EDA with SQL

1) display the names of the unique launch sites in the space mission
2) display 5 records where launch sites begin with the string ‘CCA'
3) display the total payload mass carried by boosters launched by NASA (CRS)
4) display average payload mass carried by booster version f9 v1.1
5) list the date when the first successful landing outcome in ground pad was achieved.
6) list the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
7) list the total number of successful and failure mission outcomes.
8) list the names of the booster versions which have carried the maximum payload mass.
9) list the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
10) 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.

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map.
- 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 data frame launch outcomes (failure, successes) to classes 0, 1 with red and green markers on the map in a markercluster().
- Using haverine's formula we calculated the distance from the launch site to various landmarks to find various trends about what is around launch site to measure patterns. lines are drawn on map to measure distance to landmarks.
- Example of some trends in which launch site is situated in

1) are launch sites in close proximity to railways? no
2) are launch sites in close proximity to highways? no
3) are launch sites in close proximity to coastlines? no
4) do launch sites keep certain distance away from cities? yes

Interactive Map

Build a Dashboard with Plotly Dash

- The dashboard is built with dash web framework.
- Graphs

1) Pie chart
    1) shows the total launches by certain site
    2) display relative proportions of a multiple classes of data
    3) size of a circle can be made proportional to total quantity it represents
2) Scatter plots
    1) shows the relationship between outcome and payload mass for the different booster versions
    2) the range of data flow, i.e. maximum and minimum value, can be determined.
    3) observations and readings are straightforward.

Dashboard with Plotly

Predictive Analysis (Classification)

1) Building Model
- Load our datasets by collecting data
- Transform data
- Split our data into train and test datasets
- Check how many test samples we have
- Decide which type of machine learning algorithm we want to use
- Set our parameters and algorithms to GridSearchCV
- Fit our datasets into GridsearchCV objects and train our dataset

2) Evaluating Model
- Check accuracy fir each model
- Get tuned hyperparmeters for each type of algorithms
- Plot confusion matrix

3) Improving Model
- Feature engineering
- Algorithm tuning

4) Finding the best performing Classification Model
- The model with best accuracy score wins the best performing model.

- PREDICTION MODEL

Results

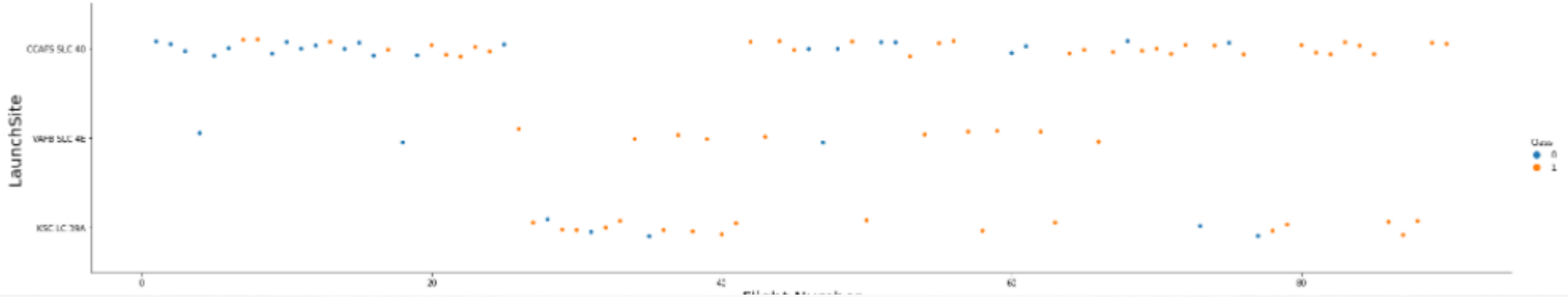
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

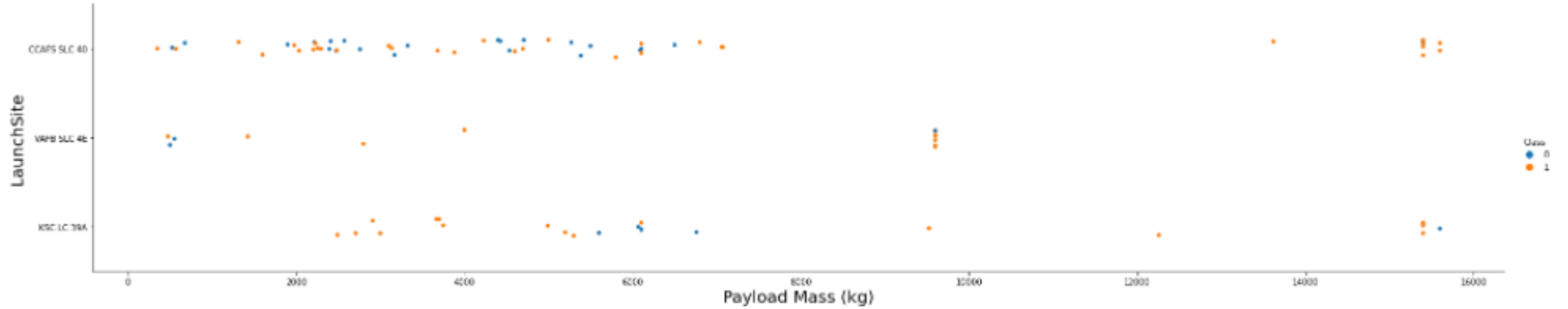
```
In [4]: # Plot a scatter point chart with x axis to be Flight Number and y axis to be the launch site, and hue to be the class value
sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("LaunchSite",fontsize=20)
plt.show()
```



Launch site CCAFS SLC 40 have more successful launches than KSC LA 39A and VAFB SLC 4E.

Payload vs. Launch Site

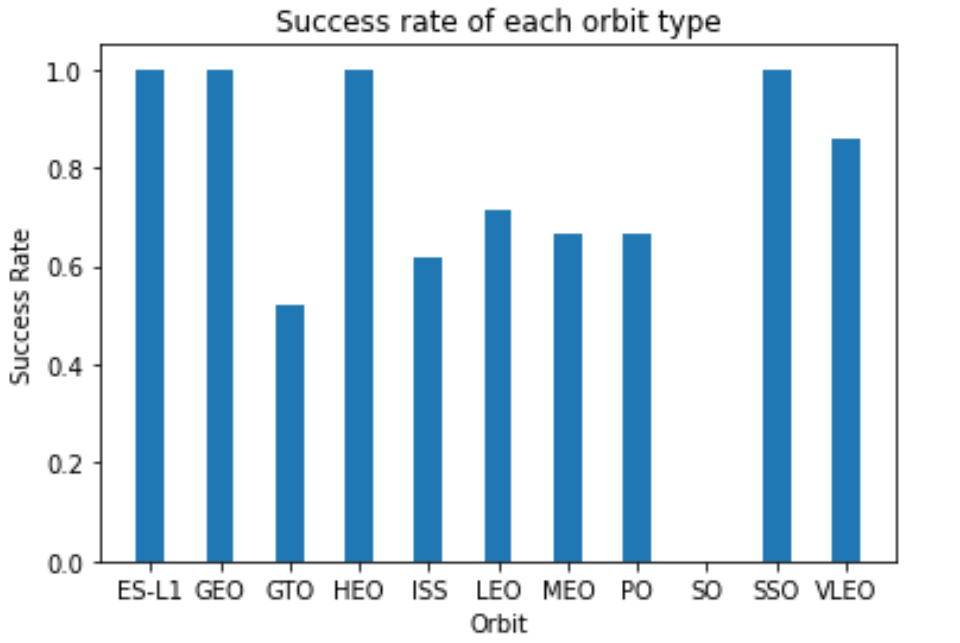
```
In [5]: # Plot a scatter point chart with x axis to be Pay Load Mass (kg) and y axis to be the launch site, and hue to be the class value
sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("Payload Mass (kg)",fontsize=20)
plt.ylabel("LaunchSite",fontsize=20)
plt.show()
```



- The greater the payload for launch site CCAFS SLC 40 the higher the success rate for the rocket. there is not quite a clear a pattern found using this visualization to make decision if launch site is dependent on pay load mass for successful launch.
- We can see that for the VAFB SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).

Success Rate vs. Orbit Type

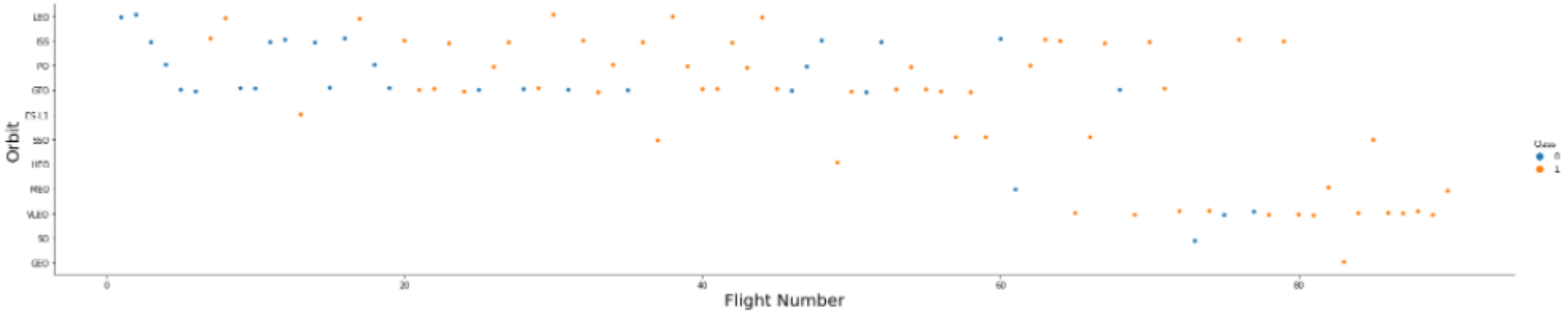
```
In [8]: plt.bar(df_1['Orbit'], df_1['Class'], width = 0.4)
        plt.xlabel("Orbit")
        plt.ylabel("Success Rate")
        plt.title("Success rate of each orbit type")
        plt.show()
```



ORBIT ES-L1, GEO, HEO, SSO has the best success rate.

Flight Number vs. Orbit Type

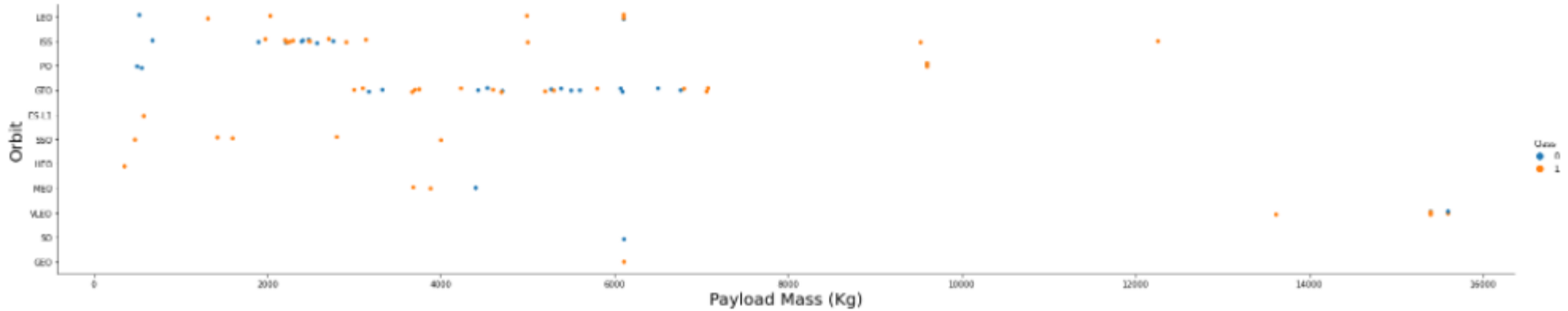
In [9]:
```
# Plot a scatter point chart with x axis to be FlightNumber and y axis to be the Orbit, and hue to be the class value
sns.catplot(y="Orbit", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```



- We can see that in 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

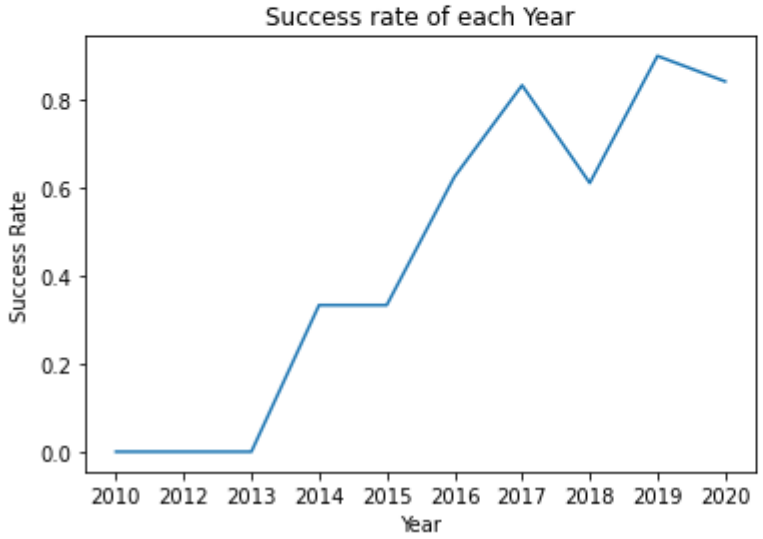
```
In [10]: # Plot a scatter point chart with x axis to be Payload and y axis to be the Orbit, and hue to be the class value
sns.catplot(y="Orbit", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("Payload Mass (Kg)",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```



- 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

```
In [16]: # Plot a line chart with x axis to be the extracted year and y axis to be the success rate
         # Plot a line chart with x axis to be the extracted year and y axis to be the success rate
         X = df_3['Year']
         Y = df_3['Class']
         plt.plot(X, Y)
         plt.xlabel("Year")
         plt.ylabel("Success Rate")
         plt.title("Success rate of each Year")
         plt.show()
```



We can observe that the success rate since 2013 shown great improvement and increases till 2020.

All Launch Site Names

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

```
In [12]: %sql select distinct(Launch_Site) from SPACEXTBL
```

```
 * ibm_db_sa://jtq88607:***@b1bc1829-6f45-4cd4-bef4-10cf0819(
Done.
```

Out[12]:

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

- Unique launch sites in the column are as shown in the outcome.
- ‘distinct()’ function is used to get the unique names.

Launch Site Names Begin with 'CCA'

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

In [13]:
```
%sql select * from SPACEXTBL where Launch_Site like 'CCA%' limit 5;
```

```
 * ibm_db_sa://jtq88607:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32304/BLUDB
Done.
```

Out[13]:

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

- 5 records where launch sites begin with `cca` are as shown in above outcome.

Total Payload Mass

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

```
In [14]: %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer = 'NASA (CRS)';
 * ibm_db_sa://jtq88607:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.c1ogj3sd0tgtu0lqde00.
Done.
Out[14]:
1
45596
```

- Total payload carried by boosters from NASA is 45596 kg.

Average Payload Mass by F9 v1.1

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

```
In [15]: %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version like 'F9 v1.1%';
          * ibm_db_sa://jtq88607:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.c1ogj3sd0tgtu0lqde00.databases.
         Done.
Out[15]:
```

1
2534

- Average payload mass carried by the booster version F9 v1.1 is 2534 kg.

First Successful Ground Landing Date

List the date when the first successful landing outcome in ground pad was acheived.

Hint:Use min function

In [16]:
```
%sql select Date from SPACEXTBL where landing__outcome = 'Success (ground pad)' limit 1;
```

```
 * ibm_db_sa://jtq88607:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.c1ogj3sd0tgtu0lqde00.dat
Done.
```

Out[16]:

DATE
2015-12-22

- 22-12-2015 is the first date when the successful landing outcome in ground pad was achieved.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

In [17]:
```
%sql select Booster_Version from SPACEXTBL \
where PAYLOAD_MASS__KG_ between 4000 and 6000 and landing__outcome = 'Success (drone ship)';
```

```
 * ibm_db_sa://jtq88607:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.c1ogj3sd0tgtu0lqde00.databases.appdomain.clou
Done.
```

Out[17]:

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

- 4 types of booster versions are present with payload between 4000 – 6000 kg which have the landing outcome as successful.

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
In [18]: %sql select Mission_Outcome, count(Mission_Outcome) AS Count from SPACEXTBL GROUP BY Mission_Outcome;
 * ibm_db_sa://jtq88607:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.c1ogj3sd0tgtu0lqde00.databases.appdoma
Done.
```

Out[18]:

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

- Queried the distinct mission outcome and then group them by their count, to get the idea about the success and failure count of mission.

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [19]:
```
%sql select Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTBL \
where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

```
 * ibm_db_sa://jtq88607:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.c1ogj3sd0tgtu0lqde00.databa
Done.
```

Out[19]:

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

- Booster version are queried by equating payload mass to maximum payload mass from the column.

2015 Launch Records

In [34]:
```
%sql select Booster_Version, Launch_Site from SPACEXTBL \
where YEAR(Date) = 2015 and landing__outcome = 'Failure (drone ship)';
```

```
 * ibm_db_sa://jtq88607:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32304/BLUDB
Done.
```

Out[34]:

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- Booster versions, and launch site of launch which failed to land in the year 2015.

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

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

In [74]:
```
%sql select landing__outcome, count(landing__outcome) AS Count from SPACEXTBL \
WHERE (Date >'2010-06-04') AND (Date < '2017-03-20') GROUP BY landing__outcome ORDER BY Count DESC;
```

```
 * ibm_db_sa://jtq88607:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32304/BLUDB
Done.
```

Out[74]:

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

- Between date 2010-06-04 and 2017-03-20, the landing outcomes are as shown in the above outcome.

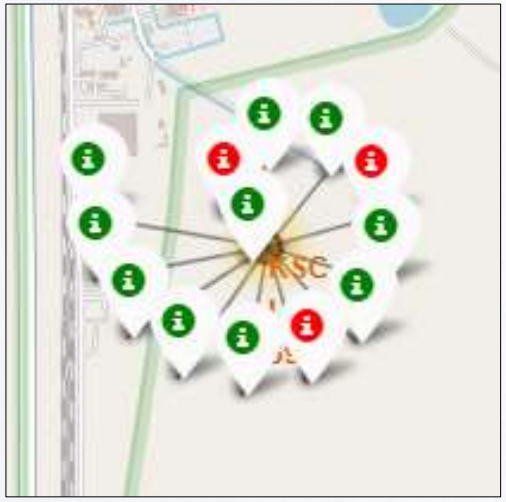
Section 4

Launch Sites Proximities Analysis

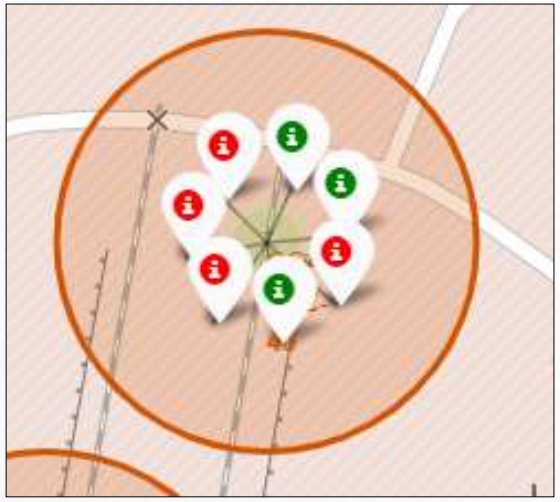
All launch Sites global map markers

- We can see that the SPACEX launch sites are in the united states of America coasts. Florida and California

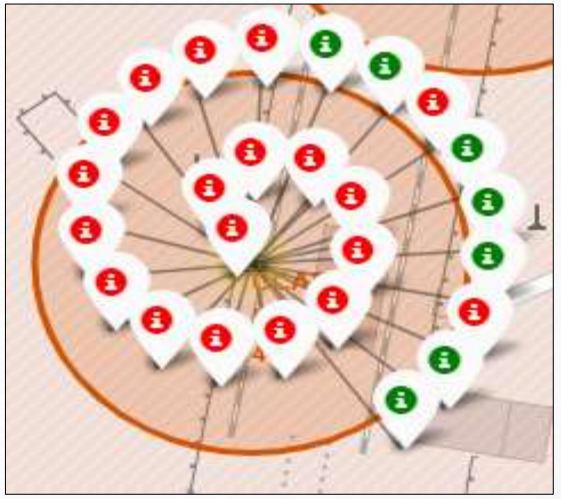
Color labelled Markers



KSC LC-39A

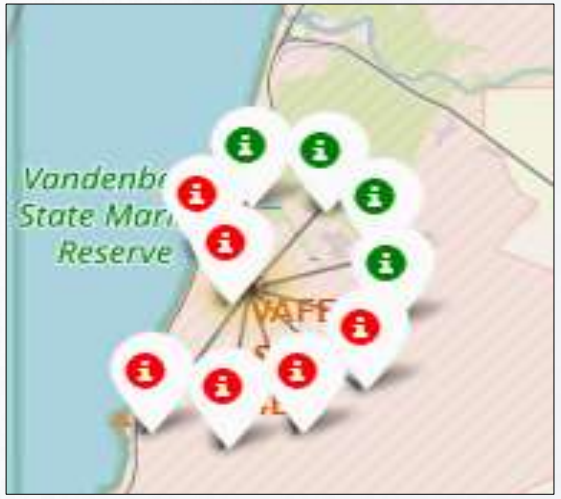


CCAFS SLC-40



CCAFS LC-40

Florida launch Sites

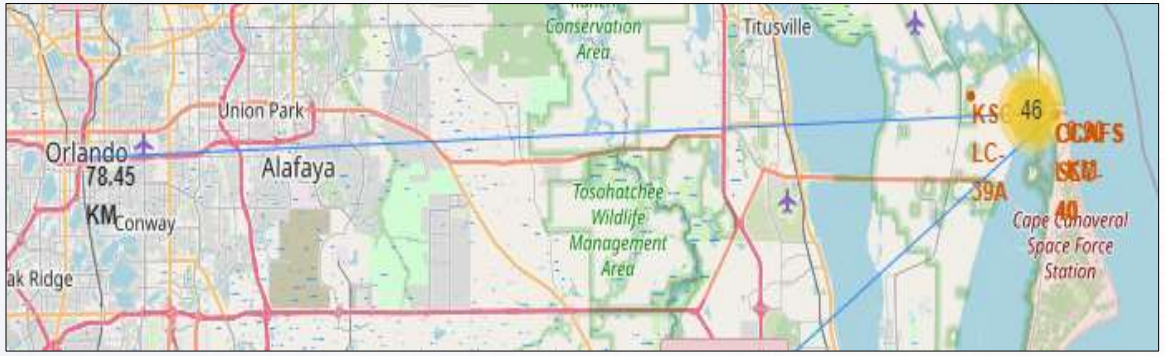


VAFB SLC-4E

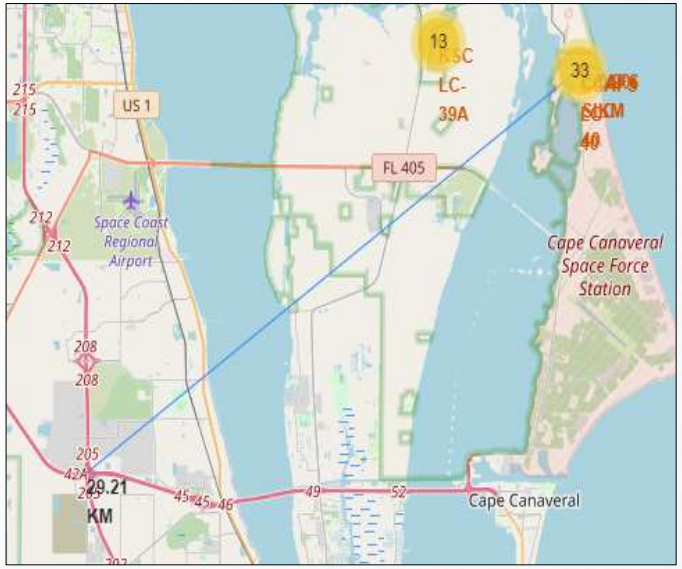
California Launch Site

- green marker shows successful launches and red marker shows failures

Distance from launch sites to landmarks



Distance to city



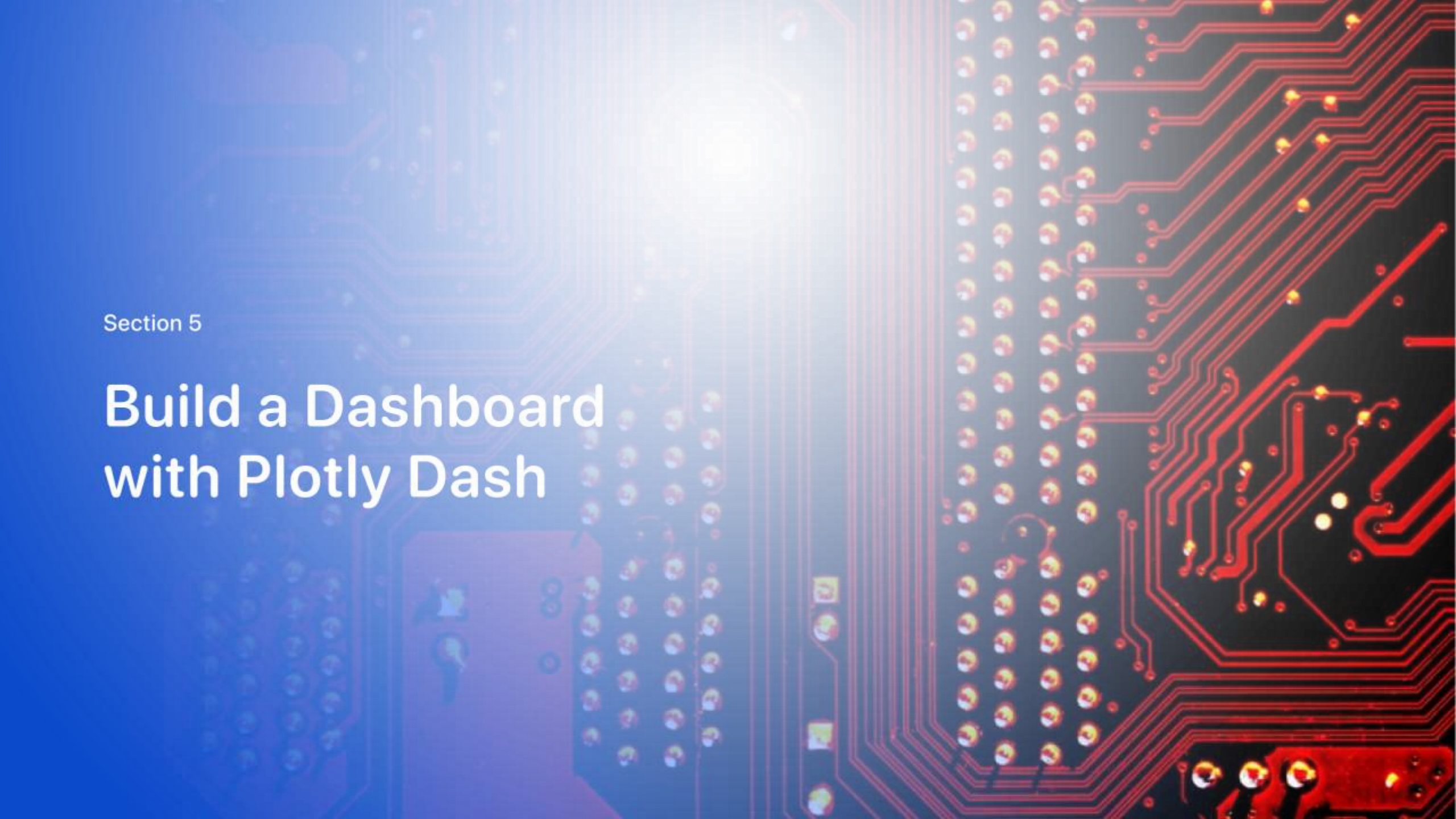
Distance to closest Highway

Distance to coastlines

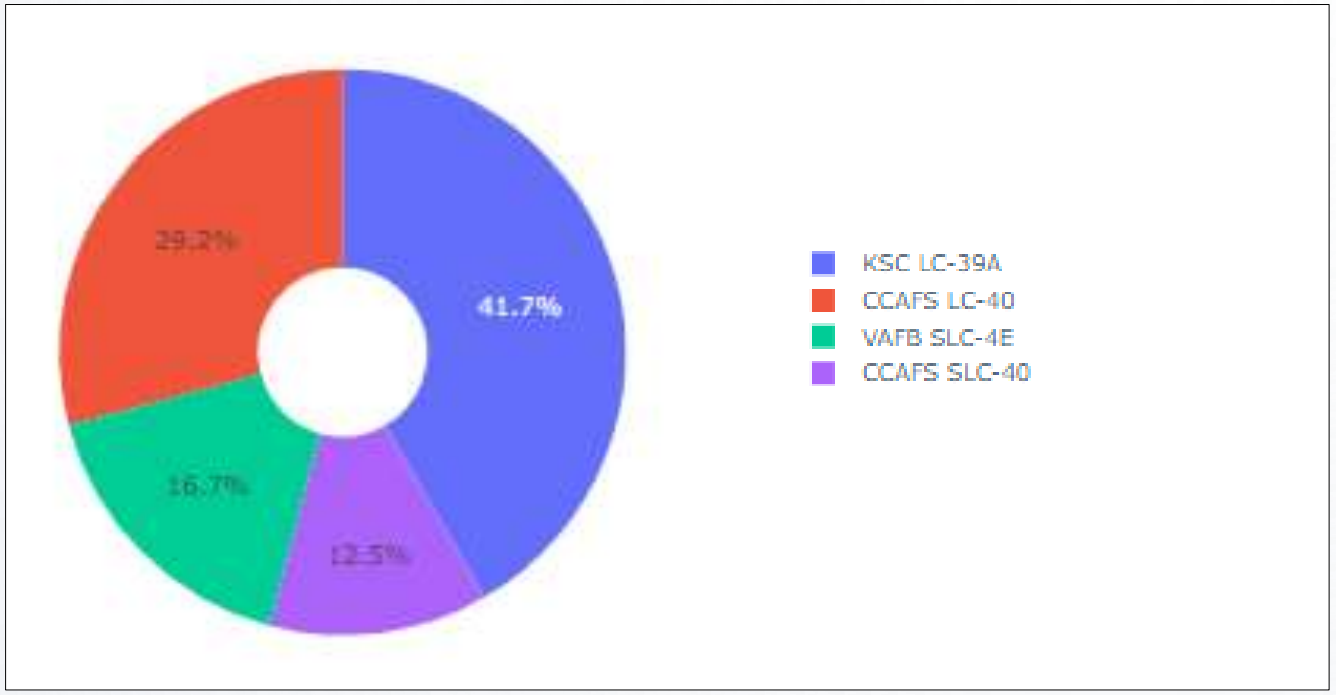
- Are launch sites in close proximity to railways? = No
- Are launch sites in close proximity to highways? = No
- Are launch sites in close proximity to coastline? = Yes
- Do launch sites keep certain distance away from cities? = Yes

Section 5

Build a Dashboard with Plotly Dash

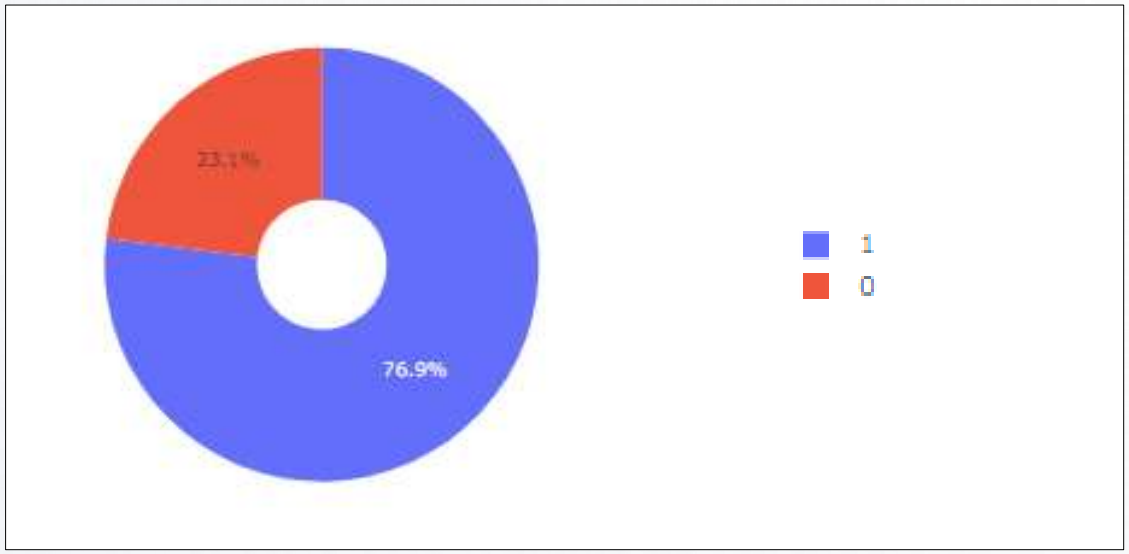


Success percentage achieved by each launch site



- We can see that KSC LC-39A had the most successful launches from all sites.

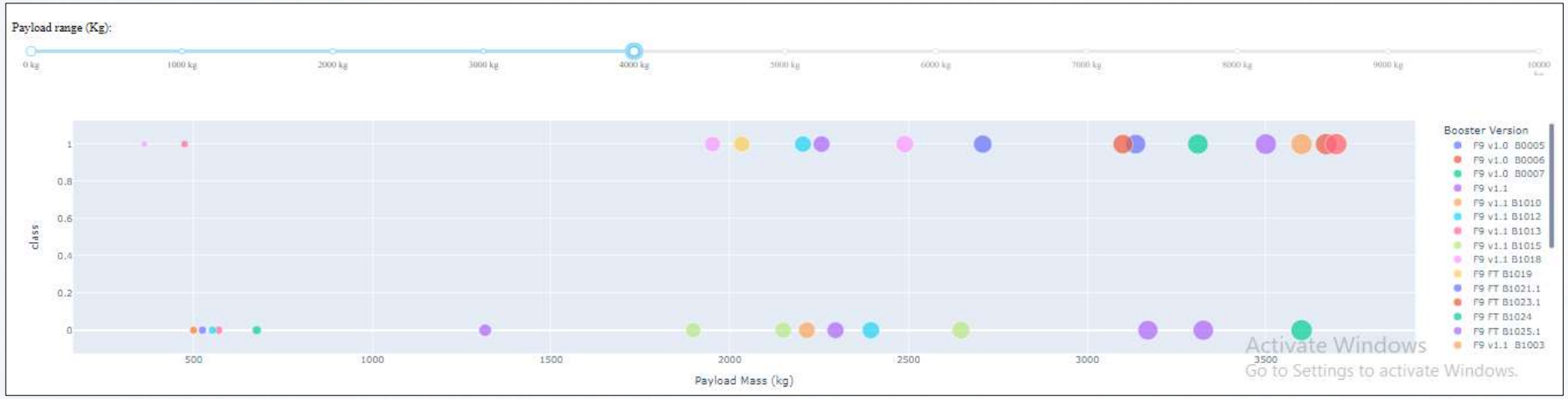
Pie chart for the launch site with highest launch success ratio



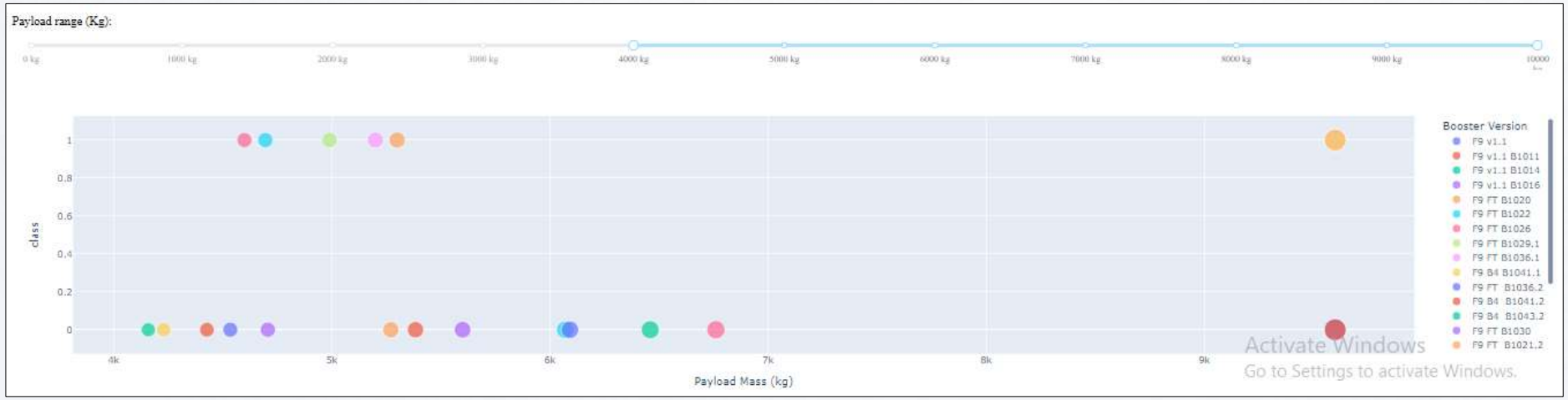
- Out of all launch sites the KSC LC-39A has most successful launches.
- KSC LC-39A achieved a 76.9% success rate while getting the 23.1% failure rate.

Payload Vs Launch Outcome

scatter plot for all sites with different payload selected in range slider



Low weighted payloads 0 – 4000 kg



High weighted payloads 4000 – 10000 kg

- We can see the success rates for low weighted payloads is higher than heavy weighted payloads.

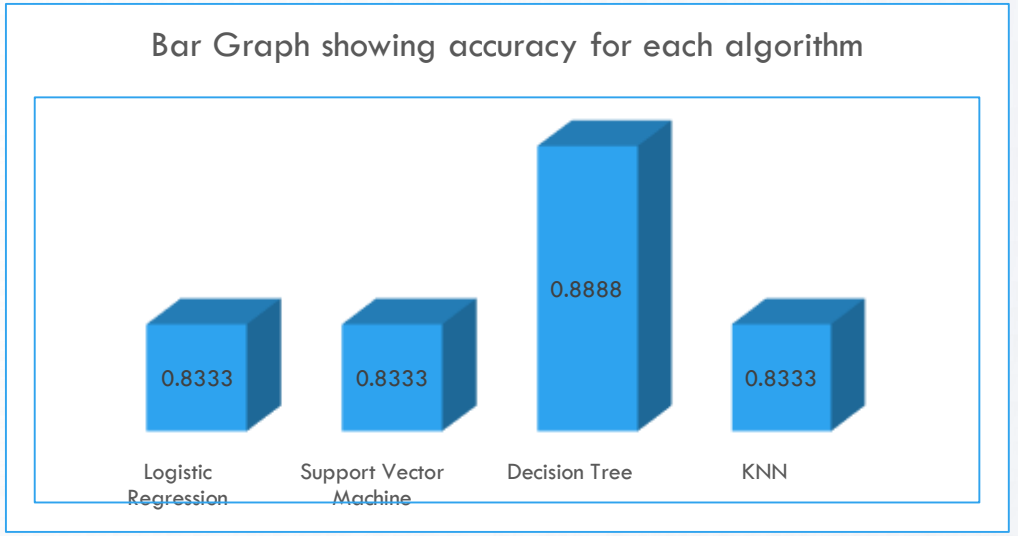
Section 6

Predictive Analysis (Classification)

Classification Accuracy

Using this function, we will find the best algorithm as per accuracy.

```
bestalgorithm = max(algorithms, key=algorithms.get)
```



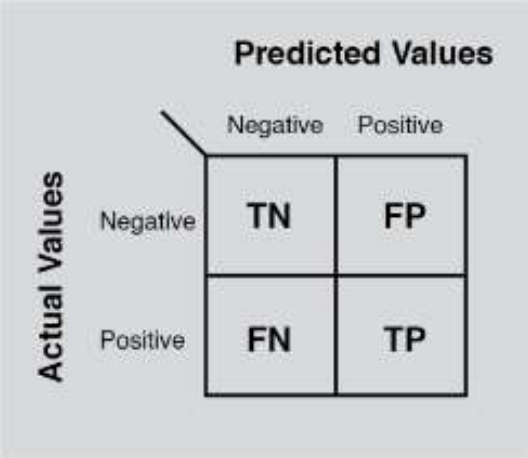
Tree algorithm wins

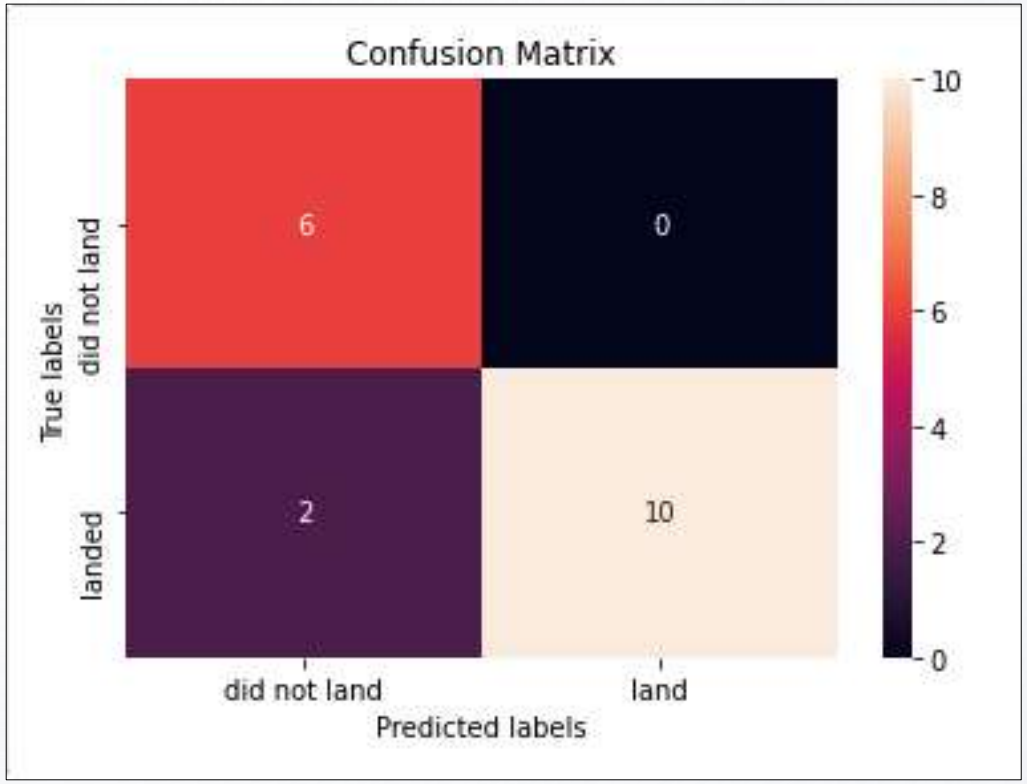
```
Best Algorithm is Tree with a score of 0.8607142857142858
Best Params is : {'criterion': 'gini', 'max_depth': 14, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 10,
'splitter': 'best'}
```

- As we can see decision tree gives the best accuracy from the available algorithms.
- After selecting this best parameters for decision tree classifier using validation data, we achieved 88.88% accuracy on test data

Confusion Matrix

- Examining confusion matrix, we see that tree can distinguish between the different classes.
- We have some false negatives in the classification result.





Confusion Matrix for Decision tree

Conclusions

- ✓ Decision tree algorithm is the best classifier for this dataset.
- ✓ Low weighted payloads perform better than heavier weighted payloads.
- ✓ The success rate for SPACEX launches improves with time. It is more than 80% as of now.
- ✓ We can see KSC LC-39A has the most successful launches from all sites.
- ✓ Orbit GEO, HEO, SSO, ES-11 has the best success rate.

Appendix

- Haversine formula
- Adggoogle maps
- Module SQL server

Thank you!