WINNING SPACE RACE WITH DATA SCIENCE

RAHUL KARMAKAR

19-02-2023

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Summary
  - Visualization – Maps
  - Dashboards
  - Predictive Analysis
- Conclusion
- Appendix

EXECUTIVE SUMMARY

- The methodologies used for this project are as follows:
  - Data Collection using web scraping and SpaceX API;
  - Data Wrangling;
  - Exploratory Data Analysis with SQL and Visualization;
  - Interactive visual analysis with Dashboards;
  - Machine Learning for future prediction.

- Summary:
  - Machine Learning was able to find the best way to approach the competitive problem along with a higher accurate model.

INTRODUCTION

- Our objective is to find the best possible way for company Space Y to compete against the best version of Space X's Falcon 9 conforming to economy and efficiency.
- Economy and efficiency are dependent on reusability of the First Stage of the Falcon 9 rocket as well as the successful launches and landings from certain launch locations throughout USA.

METHODOLOGY

Data Collection

Data for Space X were obtained via Space X API and Web Scraping.

Data Wrangling

Helps in finding patterns and creating label for further training supervised model.

Exploratory Data Analysis using Visualization and SQL queries

Interactive visual analytics using Folium and Plotly Dash

Predictive analysis using classification models

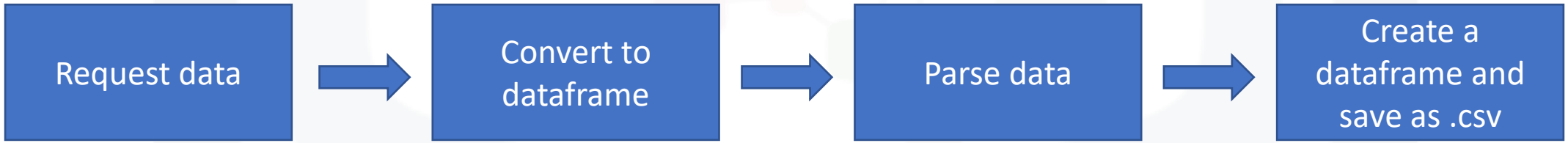
Use Logistic Regression, Support Vector Machine, Tree Decision Classifier and K-Nearest Neighbour for best fitting model.

DATA COLLECTION – SpaceX API

- Data was collected from public SpaceX API (https://api.spacexdata.com/v4/rockets/).
- The .json result was converted into a Pandas data frame.
- The data was then parsed and cleaned to include only Falcon 9 launches.
- The resulting data was exported to a .csv file for further actions.
- Link: Data Collection notebook.

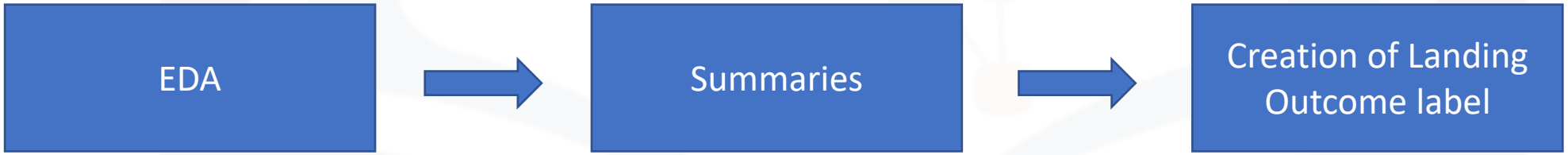
DATA COLLECTION – WEB SCRAPING

- The historical launch records of Falcon 9 rocket are available in Wikipedia.
- Data was scraped from a snapshot of the page `List of Falcon 9 and Falcon Heavy launches` updated on 9th June 2021 using the URL link "List of Falcon 9 and Falcon Heavy Launches".
- Link: Web Scraping notebook



DATA WRANGLING

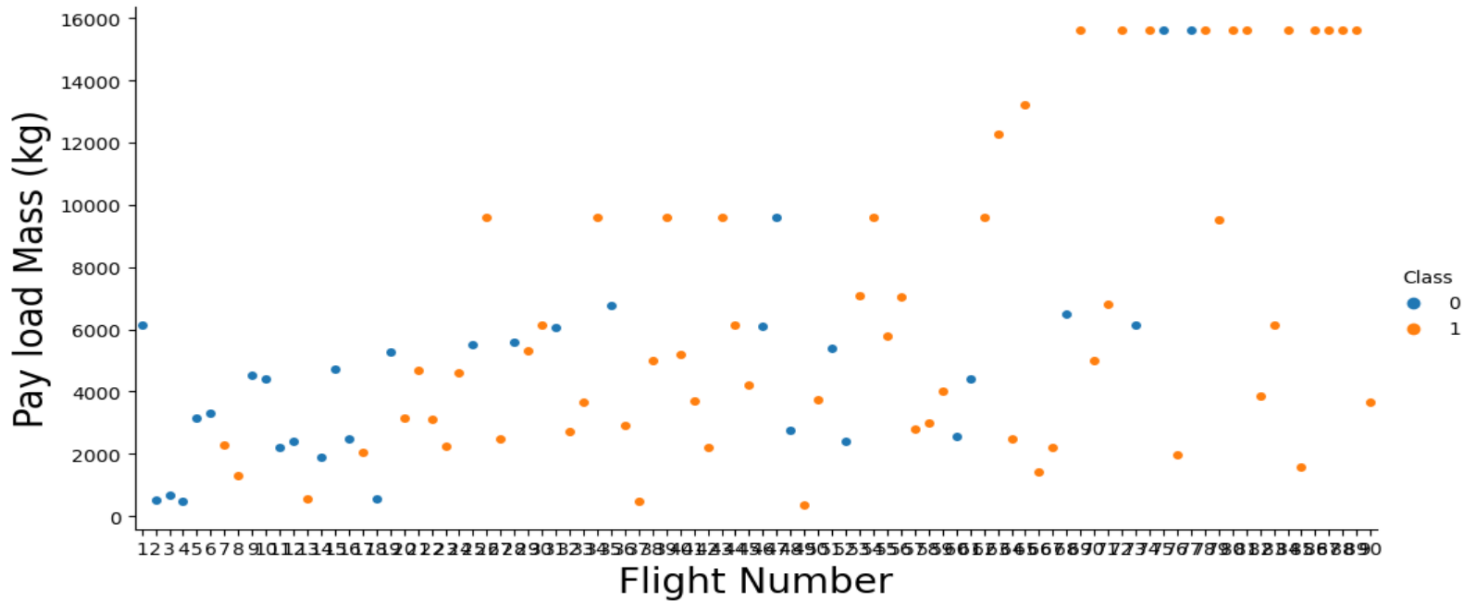
- Initial Exploratory Data Analysis (EDA) is performed using different visualization plots.
- Next, the number of launches on each site, number and occurrence of each orbit and mission outcome per orbit type were calculated.
- Later, a landing outcome label from Outcome column was created.
- Link: Data Wrangling notebook.



EDA WITH DATA VISUALIZATIONS

- The various relationships between different features were visualized using different plots. They were as follows:
  - Payload Mass vs Flight Number – catplot;

```
sns.catplot(y="PayloadMass", x="FlightNumber", hue="Class", data=df, aspect = 2)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Pay load Mass (kg)",fontsize=20)
plt.show()
```



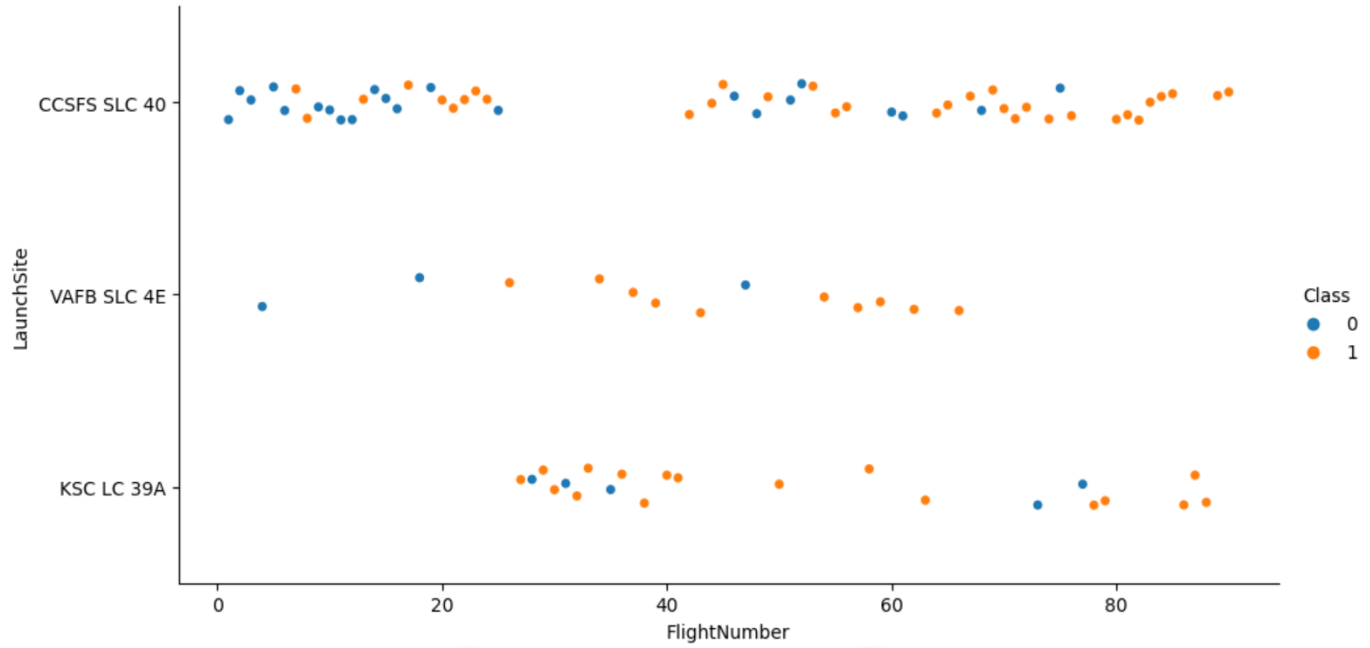
We see that different launch sites have different success rates. `CCAFS LC-40`, has a success rate of 60 %, while `KSC LC-39A` and `VAFB SLC 4E` has a success rate of 77%.

EDA WITH DATA VISUALIZATIONS

- The various relationships between different features were visualized using different plots. They were as follows:
  - Flight Number vs Launch Site – catplot:- Shows that KSC LC 39A is the most successful.

```
# 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=2)
```

```
<seaborn.axisgrid.FacetGrid at 0x122f3bf9d30>
```

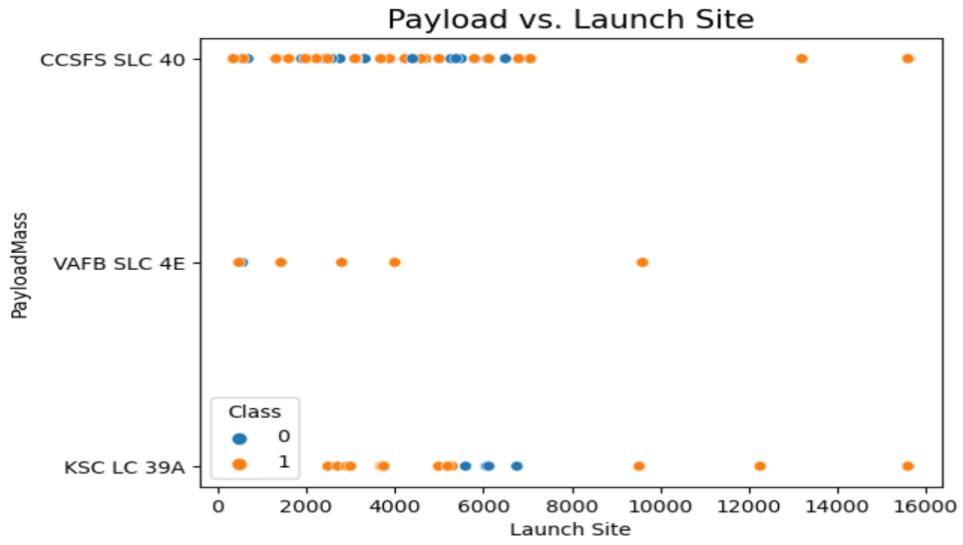


EDA WITH DATA VISUALIZATIONS

- The various relationships between different features were visualized using different plots. They were as follows:
  - Payload Mass and Launch Site – scatterplot;

```
# 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.scatterplot(data=df, x='PayloadMass', y='LaunchSite', hue='Class')
#sns.catplot(x='LaunchSite',y='PayloadMass', hue = 'Class', data=df,aspect =2, alpha = 0.6)
plt.title('Payload vs. Launch Site', fontsize = 15)
plt.xlabel('Launch Site')
plt.ylabel('PayloadMass')
plt.show
```

```
<function matplotlib.pyplot.show(close=None, block=None)>
```



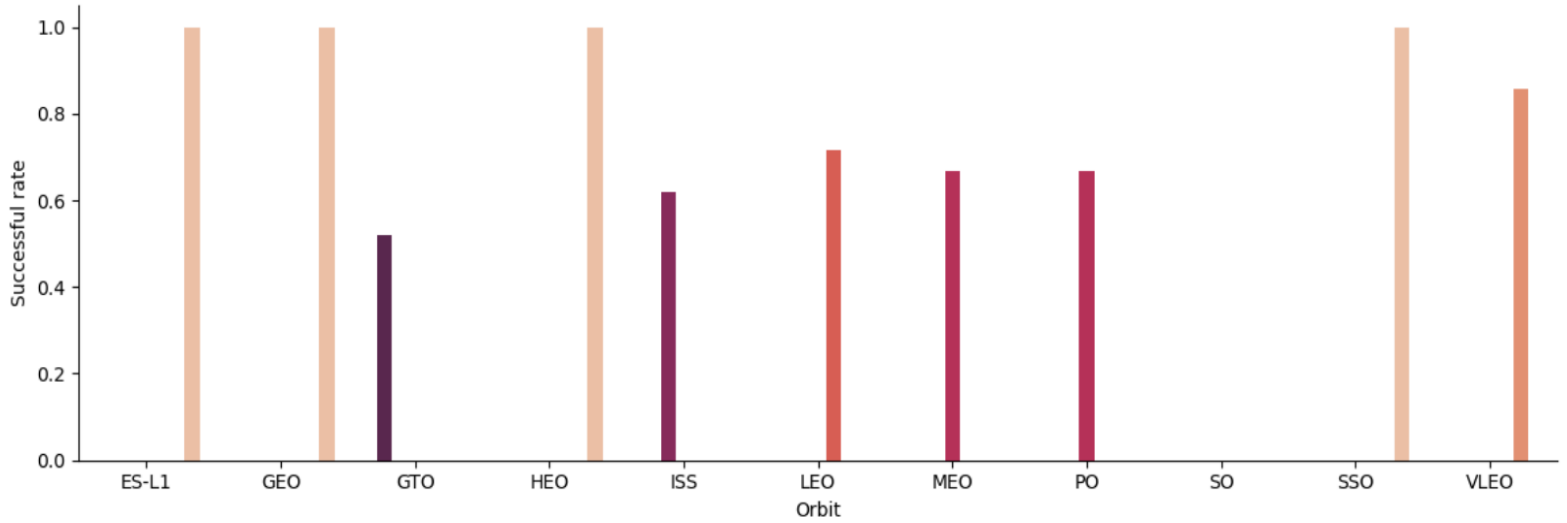
Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

EDA WITH DATA VISUALIZATIONS

- The various relationships between different features were visualized using different plots. They were as follows:
  - Orbit vs Successful rate – bar chart via catplot:- SSO, ES-L1, GEO and HEO are the most successful orbits to get launched to.

```
# HINT use groupby method on Orbit column and get the mean of Class column
df2 = df.groupby(['Orbit'])['Class'].mean().reset_index()
sns.catplot(data=df2, x='Orbit', y='Class', kind='bar',palette = 'rocket', hue = 'Class', height =4, aspect =3, legend = None)
plt.ylabel('Successful rate')
plt.show
```

```
<function matplotlib.pyplot.show(close=None, block=None)>
```

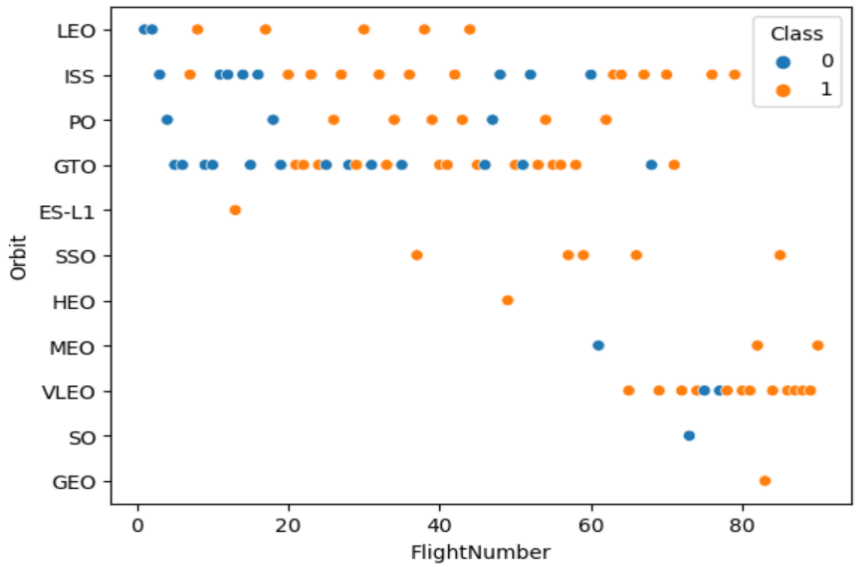


EDA WITH DATA VISUALIZATIONS

- The various relationships between different features were visualized using different plots. They were as follows:
  - Flight Number vs Orbit – scatterplot;

```
# 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.scatterplot(y='Orbit', x='FlightNumber', data=df, hue='Class')
plt.show
```

```
<function matplotlib.pyplot.show(close=None, block=None)>
```



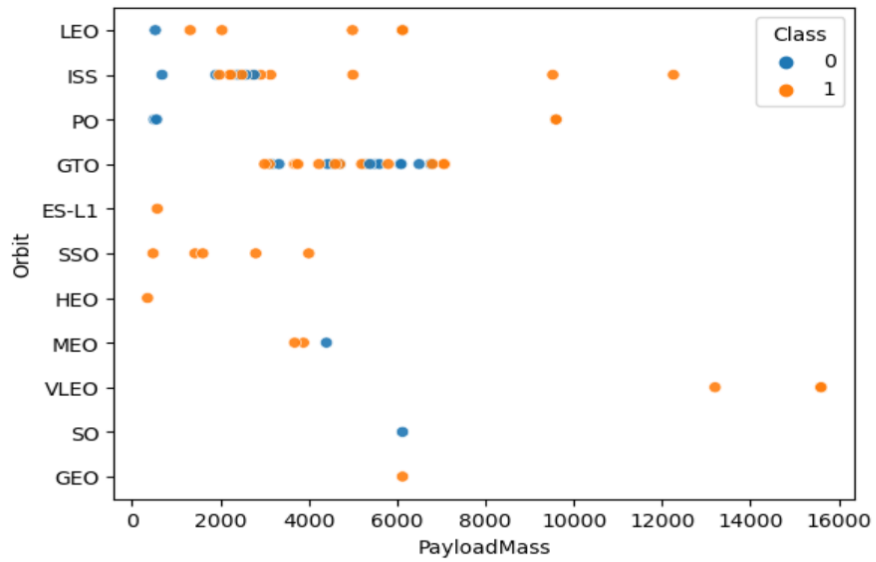
You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

EDA WITH DATA VISUALIZATIONS

- The various relationships between different features were visualized using different plots. They were as follows:
  - Payload Mass vs Orbit – scatterplot;

```
# 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.scatterplot(x='PayloadMass',y='Orbit',data=df, hue='Class',alpha=0.9)
plt.show
```

```
<function matplotlib.pyplot.show(close=None, block=None)>
```



With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

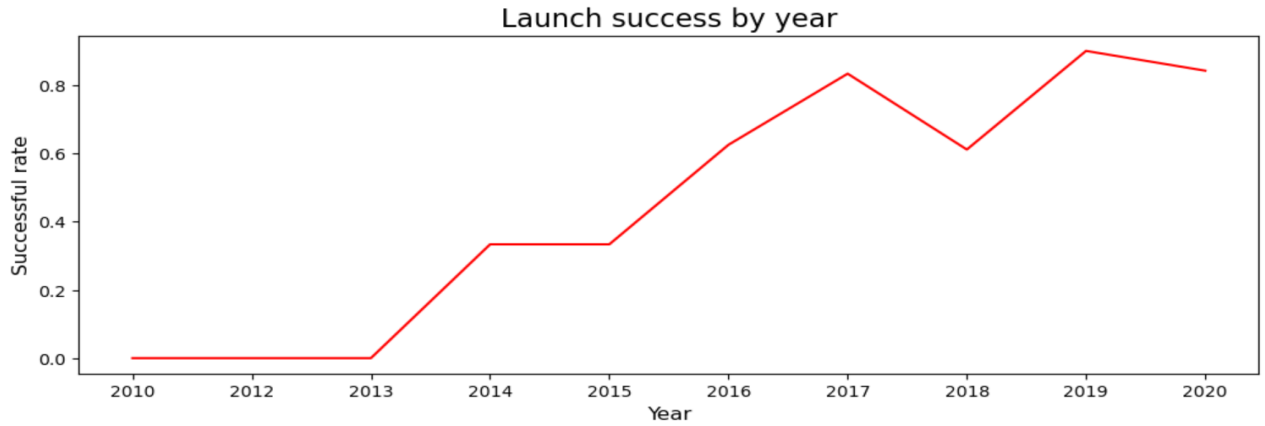
However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

EDA WITH DATA VISUALIZATIONS

- The various relationships between different features were visualized using different plots. They were as follows:
  - Launch success by year vs Successful rate – lineplot

```
# Plot a line chart with x axis to be the extracted year and y axis to be the success rate
plt.figure(figsize=(12,4))
sns.lineplot(x='Date',y='Class',data=df3, color ='red')
plt.xlabel('Year',fontsize=12)
plt.ylabel('Successful rate', fontsize=12)
plt.title('Launch success by year',fontsize=16)
plt.show
```

```
<function matplotlib.pyplot.show(close=None, block=None)>
```



you can observe that the sucess rate since 2013 kept increasing till 2020

- Link: Data Visualization notebook.

EDA WITH SQL

- The following 10 queries were performed:-
  - Display the names of the unique launch sites in the space mission;

Task 1

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

```
%sql select distinct Launch_Site from SPXTABLE;
```

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

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

EDA WITH SQL

- The following 10 queries were performed:-
  - Display 5 records where launch sites begin with the string 'CCA';

Task 2

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

```
%sql select Launch_Site from SPXTABLE where Launch_Site like 'CCA%' limit 5
```

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

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

EDA WITH SQL

- The following 10 queries were performed:-
  - Display the total payload mass carried by boosters launched by NASA (CRS);

Task 3

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

```
%sql select sum(PAYLOAD_MASS__KG_) as total_payload from SPXTABLE where Customer like 'NASA (CRS)';
```

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

total_payload
45596

EDA WITH SQL

- The following 10 queries were performed:-
  - Display average payload mass carried by booster version F9 v1.1;

Task 4

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

```
%sql select avg(PAYLOAD_MASS__KG_) as Avg_Payload from SPXTABLE where Booster_version like 'F9 v1.1';
```

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

Avg_Payload
2928.4

EDA WITH SQL

- The following 10 queries were performed:-
  - List the date when the first successful landing outcome in ground pad was achieved;

Task 5

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

Hint:Use min function

```
%sql select min(date) As Early_Date from SPXTABLE where Landing_Outcome like 'Success (ground pad)';
```

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

Early_Date
01-05-2017

EDA WITH SQL

- The following 10 queries were performed:-
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000;

Task 6

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

```
%sql select distinct Customer, Landing_Outcome, PAYLOAD_MASS__KG_ from SPXTABLE where Landing_Outcome ='Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000;
```

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

Customer	Landing_Outcome	PAYLOAD_MASS__KG_
SKY Perfect JSAT Group	Success (drone ship)	4696
SKY Perfect JSAT Group	Success (drone ship)	4600
SES	Success (drone ship)	5300
SES EchoStar	Success (drone ship)	5200

EDA WITH SQL

- The following 10 queries were performed:-
  - List the total number of successful and failure mission outcomes;

Task 7

List the total number of successful and failure mission outcomes

```
%sql select Mission_Outcome, Count(*) as Numbers from SPXTABLE group by Mission_Outcome;
```

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

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

EDA WITH SQL

- The following 10 queries were performed:-
  - List the names of the booster_versions which have carried the maximum payload mass. Use a subquery (showing only some outputs as rows are more in number);

Task 8

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

```
%sql select Booster_Version, Max_Payload from (select Booster_Version, max(PAYLOAD_MASS__KG_) as Max_Payload from SPXTABLE group by Booster_Version) as Sub;
```

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

Booster_Version	Max_Payload
F9 B4 B1039.2	2647
F9 B4 B1040.2	5384
F9 B4 B1041.2	9600
F9 B4 B1043.2	6460
F9 B4 B1039.1	3310
F9 B4 B1040.1	4990
F9 B4 B1041.1	9600
F9 B4 B1042.1	3500
F9 B4 B1043.1	5000
F9 B4 B1044	6092
F9 B4 B1045.1	362
F9 B4 B1045.2	2697
F9 B5 B1046.1	3600
F9 B5 B1046.2	5800

EDA WITH SQL

- The following 10 queries were performed:-
  - List the records which will display the month names, failure landing_outcomes in drone ship, booster versions, launch_site for the months in year 2015;

Task 9

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

```
%sql select substr(Date,4,2) as Month, Booster_Version, Launch_site from SPXTABLE where Landing_Outcome like 'Failure%drone%' AND substr(Date,7,4) = '2015';
```

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

Month	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

EDA WITH SQL

- The following 10 queries were performed:-
  - Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

Task 10

Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

```
%sql select Landing_Outcome, COUNT(*) as Numbers from SPXTABLE where Landing_Outcome like 'Success%' and Date between '04-06-2010' and '20-03-2017' group by Landing_Outcome order by Numbers desc;
```

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

Landing_Outcome	Numbers
Success	20
Success (drone ship)	8
Success (ground pad)	6

- Link: EDA with SQL notebook.

INTERACTIVE VISUAL ANALYTICS USING FOLIUM

- Markers, circles, lines and marker clusters were used in Folium maps.
  - Markers indicate points like launch sites;
  - Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Centre;
  - Marker clusters indicates groups of events in each coordinate, like launches in a launch site;
  - Lines are used to indicate distances between two coordinates.

- Link: Interactive visual analytics using Folium notebook.

INTERACTIVE VISUAL ANALYTICS USING PLOTLY DASH

- The following dashboard components were created using Plotly Dash:-
  - A dropdown list to enable Launch Site selection along with a call-back function for `site-dropdown` as input, `success-pie-chart` as output and a call-back function for `site-dropdown` and `payload-slider` as inputs, `success-payload-scatter-chart` as output;
  - A pie chart to show the total successful launches count for all sites;
  - A slider to select payload range;
  - A scatter chart to show the correlation between payload and launch success.

- Link: Plotly dash Python file.

PREDICTIVE ANALYSIS USING CLASSIFICATION MODELS

- The following tasks were performed in this part:-
  - Creating a numpy array and assigning it to a variable with output as Pandas series;
  - Standardizing a variable by transforming it;
  - Splitting both variables' data into test and training sets;
  - Predictive model evaluation with combinations of hyperparameters using the 4 models, i.e., Logistic Regression, Support vector machine, Decision tree classifier and K-nearest neighbour, were performed.
  - Their results were compared with each other to find the best fit model.

- Link: Machine Learning prediction notebook.

RESULTS

- From EDA analysis, we can infer the following:-
  - We see that different launch sites have different success rates, like CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%;
  - For the VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000);
  - ES-L1, GEO, HEO and SSO orbits seem to have the highest success rates;
  - In the LEO orbit, the Success appears related to the number of flights;
  - 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.
  - The success rate since 2013 kept increasing till 2020.

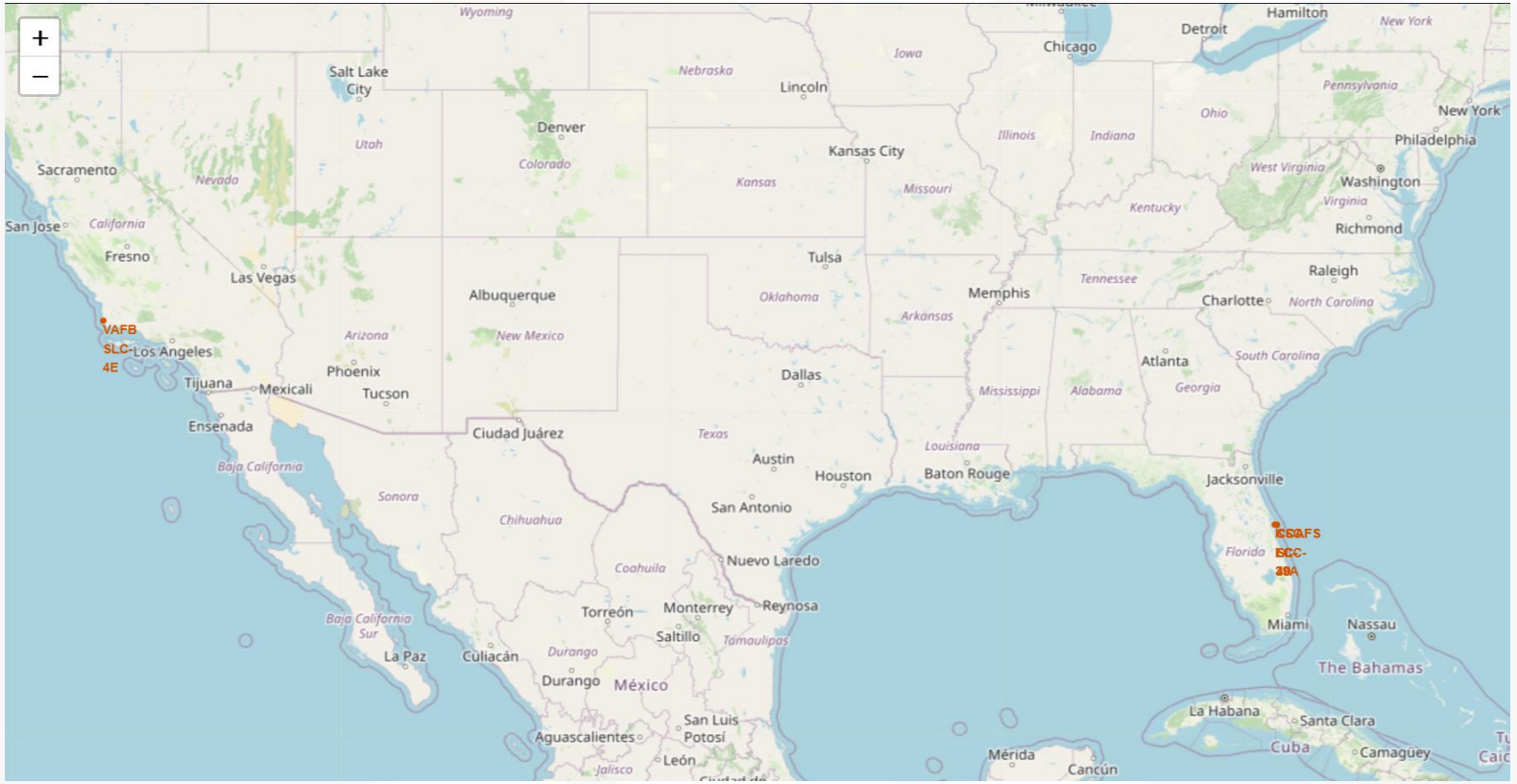
RESULTS

- From SQL query analysis, we found the following:-
  - The 4 launch sites are CCAFS LC-40, VAFB SLC-4E, KSC LC-39A and CCAFS SLC-40;
  - The total payload mass carried by boosters launched by NASA (CRS) is 45596 KG.
  - The average payload mass carried by booster version F9 v1.1 is 2928.4 KG;
  - The first successful landing outcome in ground pad was achieved on 01-05-2017;
  - Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are F9 FT B1021.2, F9 FT B1031.2, F9 FT B1022 and F9 FT B1026;
  - The count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order is shown as

Landing_Outcome	Numbers
Success	20
Success (drone ship)	8
Success (ground pad)	6

RESULTS – VISUALIZATION (MAPS)

- Launch site proximity analysis:



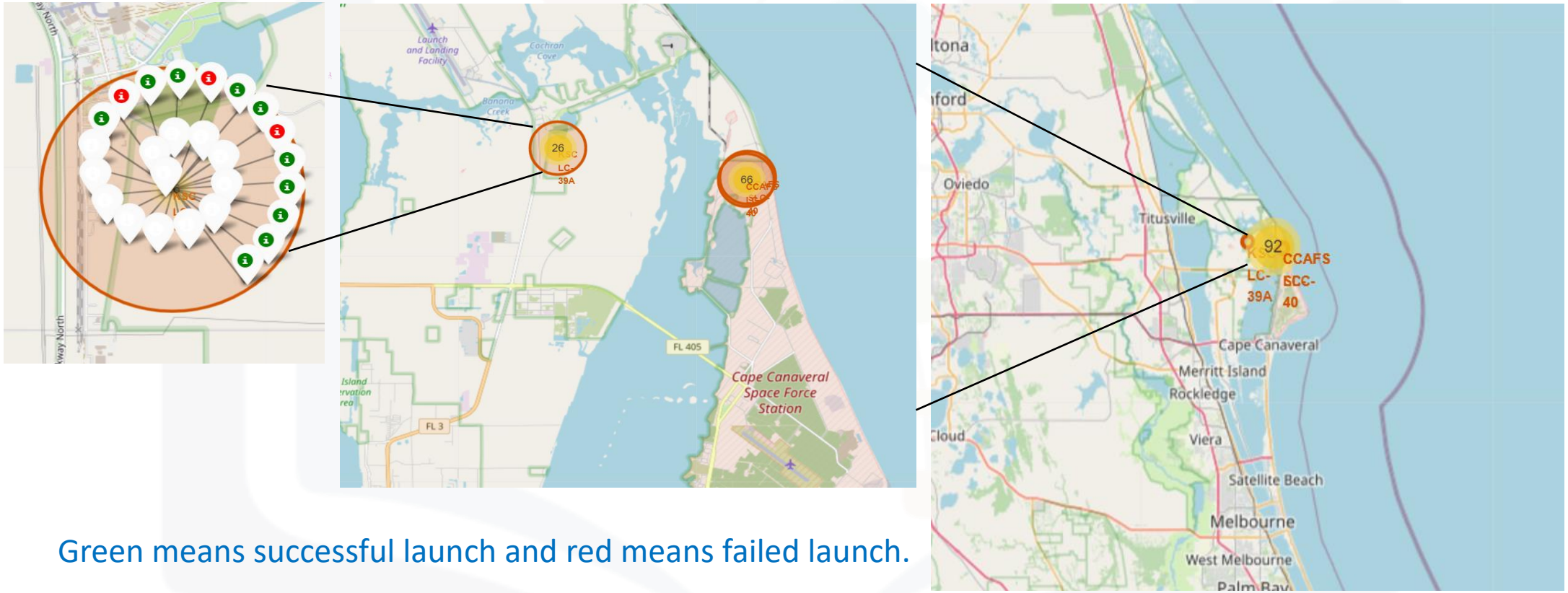
All launch sites are in places close to sea/ocean with railroad and road connectivity.

RESULTS – VISUALIZATION (MAPS)

- Success/failed launches for each site:

RESULTS – VISUALIZATION (MAPS)

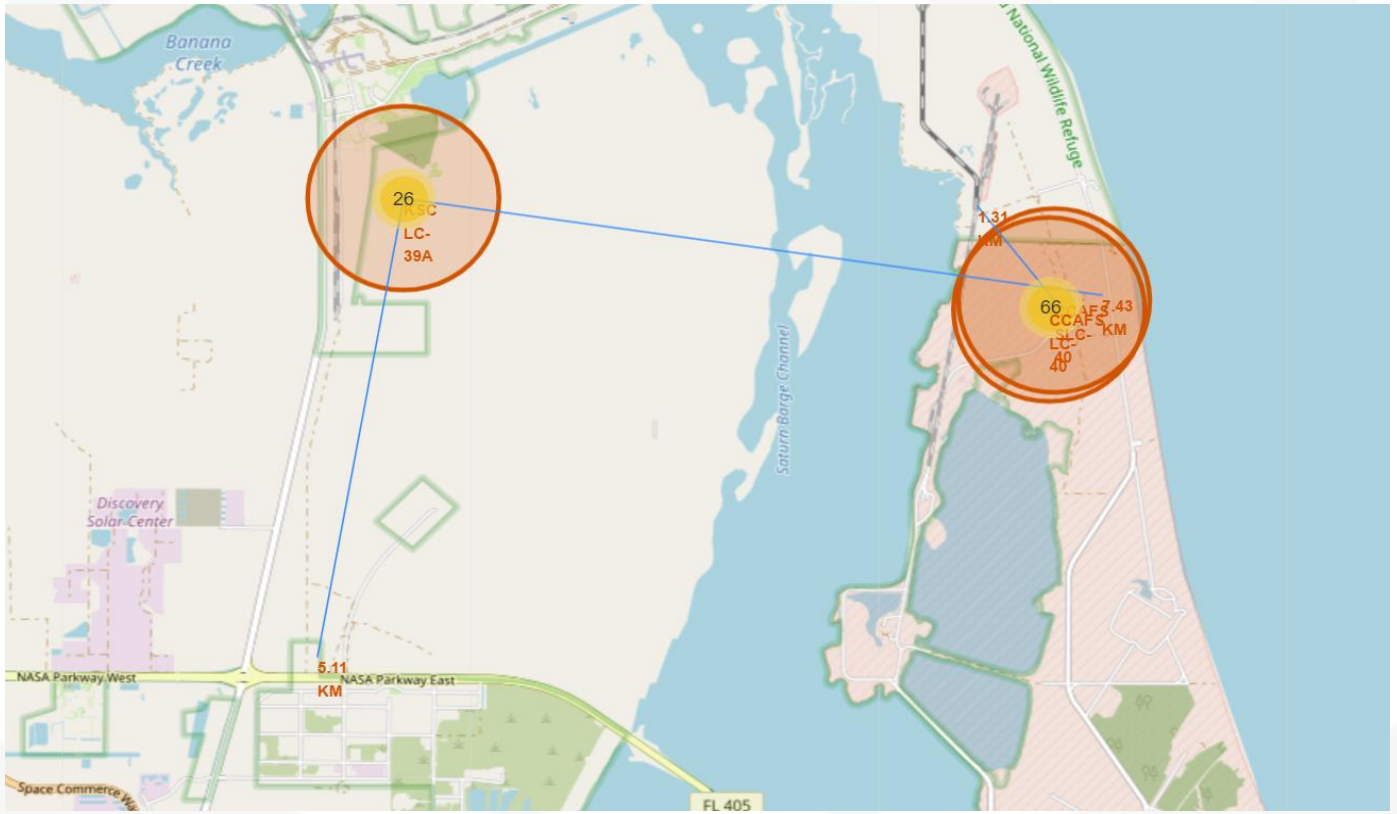
- Example of KSC LC-39A launch site launch outcomes:



Green means successful launch and red means failed launch.

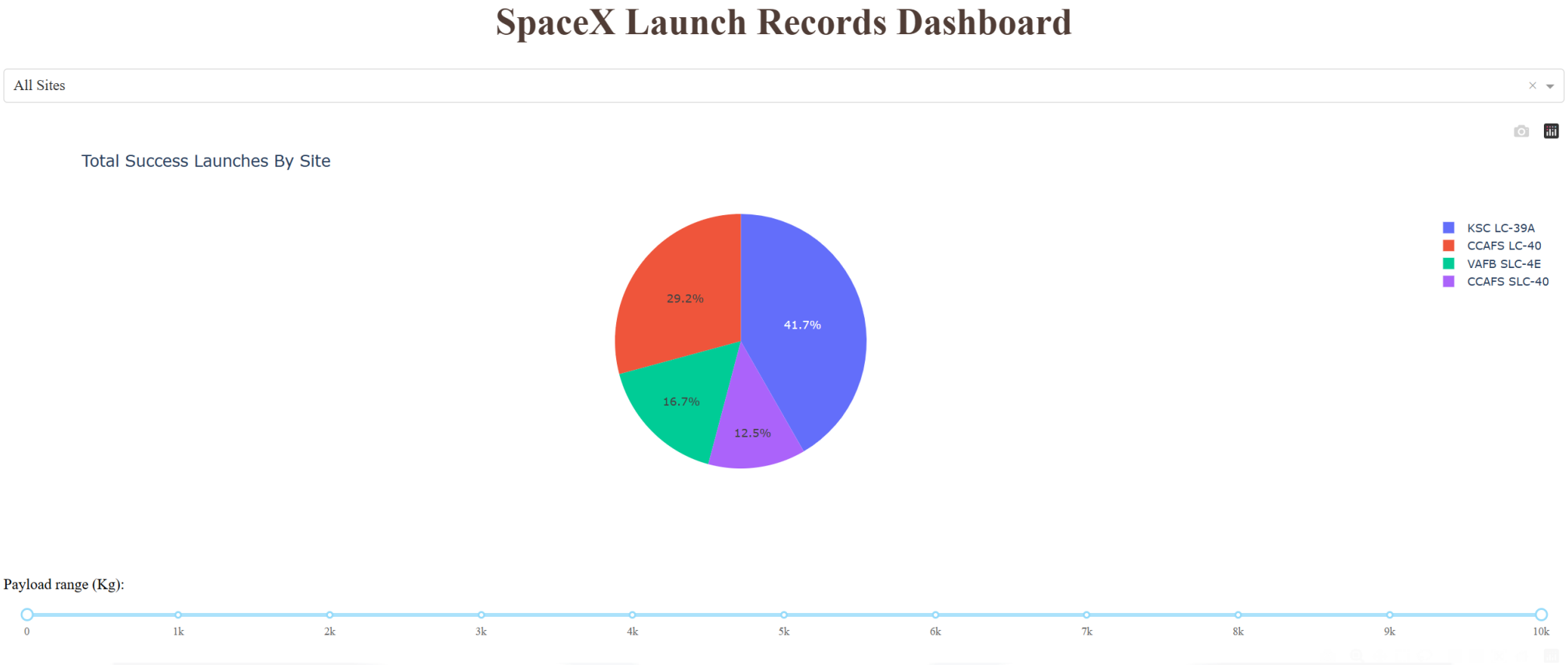
RESULTS – VISUALIZATION (MAPS)

- The sites are far away from residential areas due to safety.



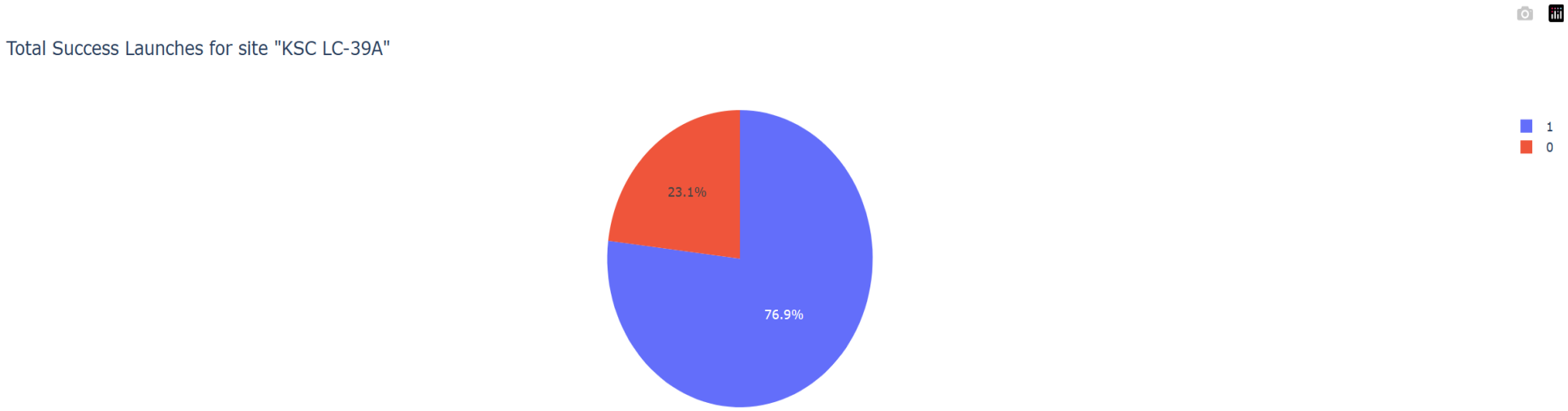
This picture shows the distances of sites with each other as well as nearest railroad and road for KSC LC-39A.

RESULTS – DASHBOARDS



This picture shows the launch success % by each site, with KSC LC-39A at 41.7%.

RESULTS – DASHBOARDS



This picture shows the launch (76.9%) and failure (23.1%) rates of KSC LC-39A.

RESULTS – DASHBOARDS

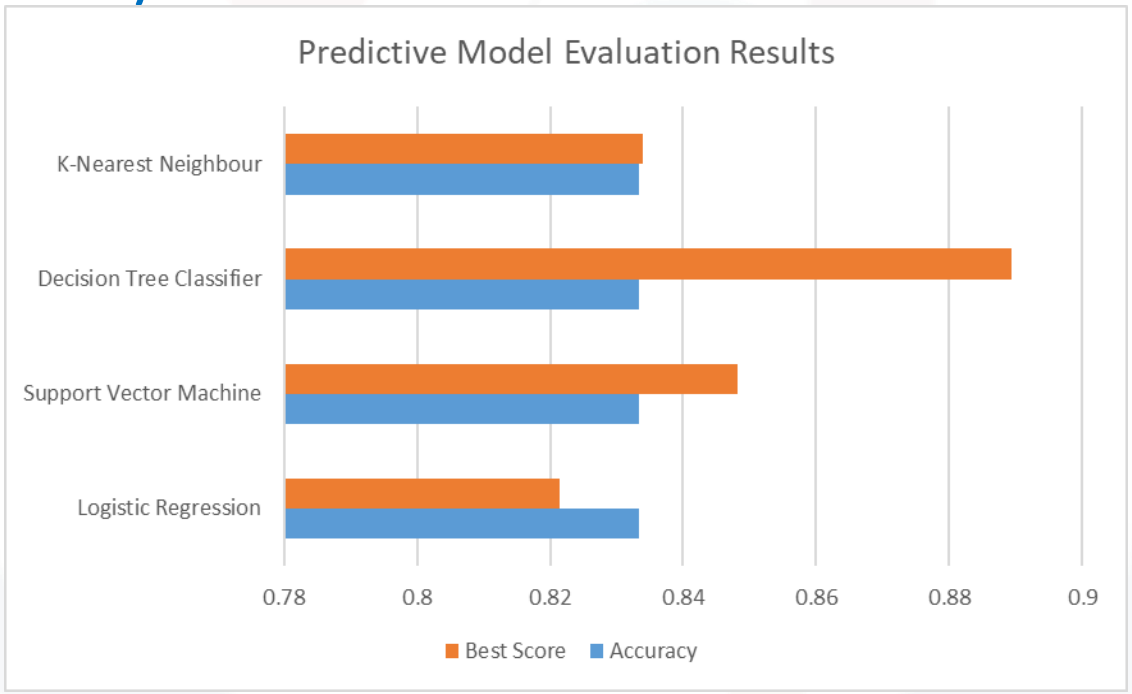
This picture shows correlation between Payload and Success for all sites along with the booster version category.

RESULTS – DASHBOARDS

- Major insights drawn from the Payload vs Success dashboard are as follows:
  - V1.0 booster version has the lowest success rate among all the booster versions.
  - FT booster version has the highest success rate among all the booster versions.
  - Most of the successful launches by all the booster versions happened in the range of 2000 KG to 6000 KG Payload Mass.
  - Most of the failed launch attempts by all the booster versions happened in the range of 0 KG to 7000 KG Payload Mass.
  - Only B4 and FT booster versions can carry a maximum Payload Mass of 9600 KG.
  - B4 booster version is versatile as it can carry different capacities.
- Link: Python Dashboard

RESULTS – PREDICTIVE ANALYSIS

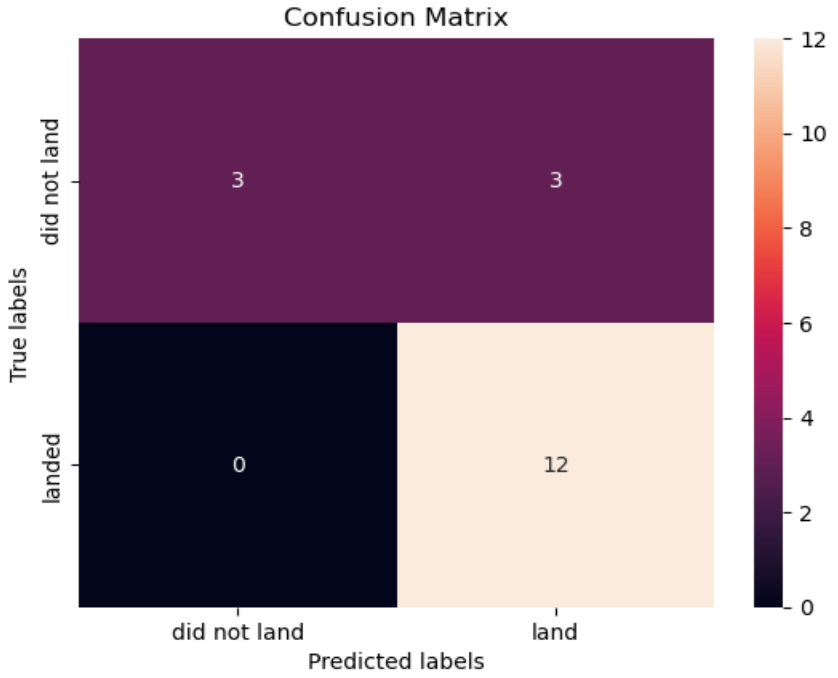
- Classification accuracy:



Even though all the models we have used have same accuracy, Decision Tree Classifier has the highest “Best Score”. Hence, we can say it is the best suited model for our training and test split data sets.

RESULTS – PREDICTIVE ANALYSIS

- Confusion Matrix of Decision Tree Classifier:



Confusion matrix of Decision Tree Classifier proves its accuracy by showing the big numbers of true positive and true negative compared to the false ones.

CONCLUSION

- Launches below 7000 KG are riskier.
- Decision Tree Classifier can be a good model for predicting successful launches with best parameters.
- There has been an increase in the success rate since 2013 which kept increasing till 2020.
- The best launch site is KSC LC 39A.
- FT booster version has the highest success rate and is versatile in carrying different payload capacities.

APPENDIX

- IBM Capstone Project
- List of Falcon 9 and Falcon Heavy launches

THANK YOU