IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Harshith Sakala Santhosh>
<25/02/2024>

Outline

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

Executive Summary

- In this capstone project, we have performed a predictive analytics on spacex data based on falcon 9 rocket launches to assess, analyse and predict the launch outcome. In addition, the data was used for spaceY to analyse and calculate price of launch in the next bid against SpaceX.
- After analysing falcon 9 launches, following trends were observed:

1) Flight number is directly proportional to launch success.
2) The rocket launches are done amongst 4 launch sites out of which ‘KSCLC39-A’ has the highest launch success rate and ‘CCAFSLC-40’ has the lowest success rate.
3) The launches to orbit VLEO has the highest success rate compared to other orbits followed by LEO,ISS and GTO has least success rate.
4) Payload mass is directly proportional to distance of orbit from earth. The more payload mass for more distant orbit the more probability of success.

Introduction

Reusable rocket launches are a significant advancement in space technology. A reusable launch vehicle has parts that can be recovered and re flown, carrying payloads from payloads from surface to other space. The most common parts or stage-1 parts are aimed for reuse which in return helps in reducing the price of launch by more than 50%.

The benefits of reusable rockets are significant, including reduced launch costs and increased frequency of launches. However, these benefits are offset by the costs of recovery and refurbishment. Despite these challenges, the development of reusable rockets continues to be a focus for many companies and is unexpected to play a crucial role in the future of space exploration.

Project Problem:

We will predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward 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 an alternate company wants to bid against SpaceX for a rocket launch?.

Major Problems:

1) What are the key factors that impact rocket launch?
2) Does launch location affects the success of launch?
3) Is there any relationship between the orbit altitude distance and payload mass and how it affects launch success?
4) Does booster has an impact on launch success?

Section 1

Methodology

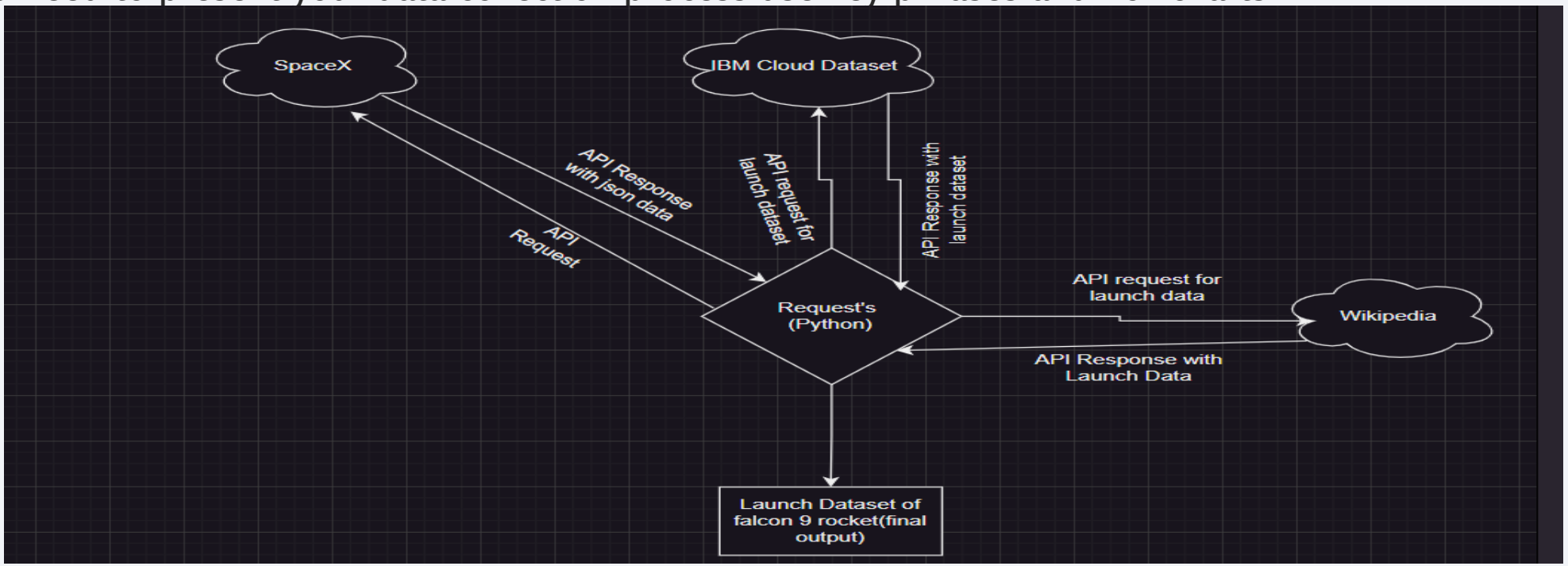
Methodology

Executive Summary

- Data collection methodology
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Data Collection:
  - The SpaceX data regarding falcon 9 was collected from SpaceX official website and Wikipedia using 5 different api's through python's request library. Out of these 5 API's, 3 of them are of SpaceX Api's, 1 from IBM Dataset and last one is from Wikipedia's page regarding falcon 9 launches. Though all API's had provided same launch data, but SpaceX and Wikipedia involved a lot of data processing to extract data.
- You need to present your data collection process use key phrases and flowcharts:



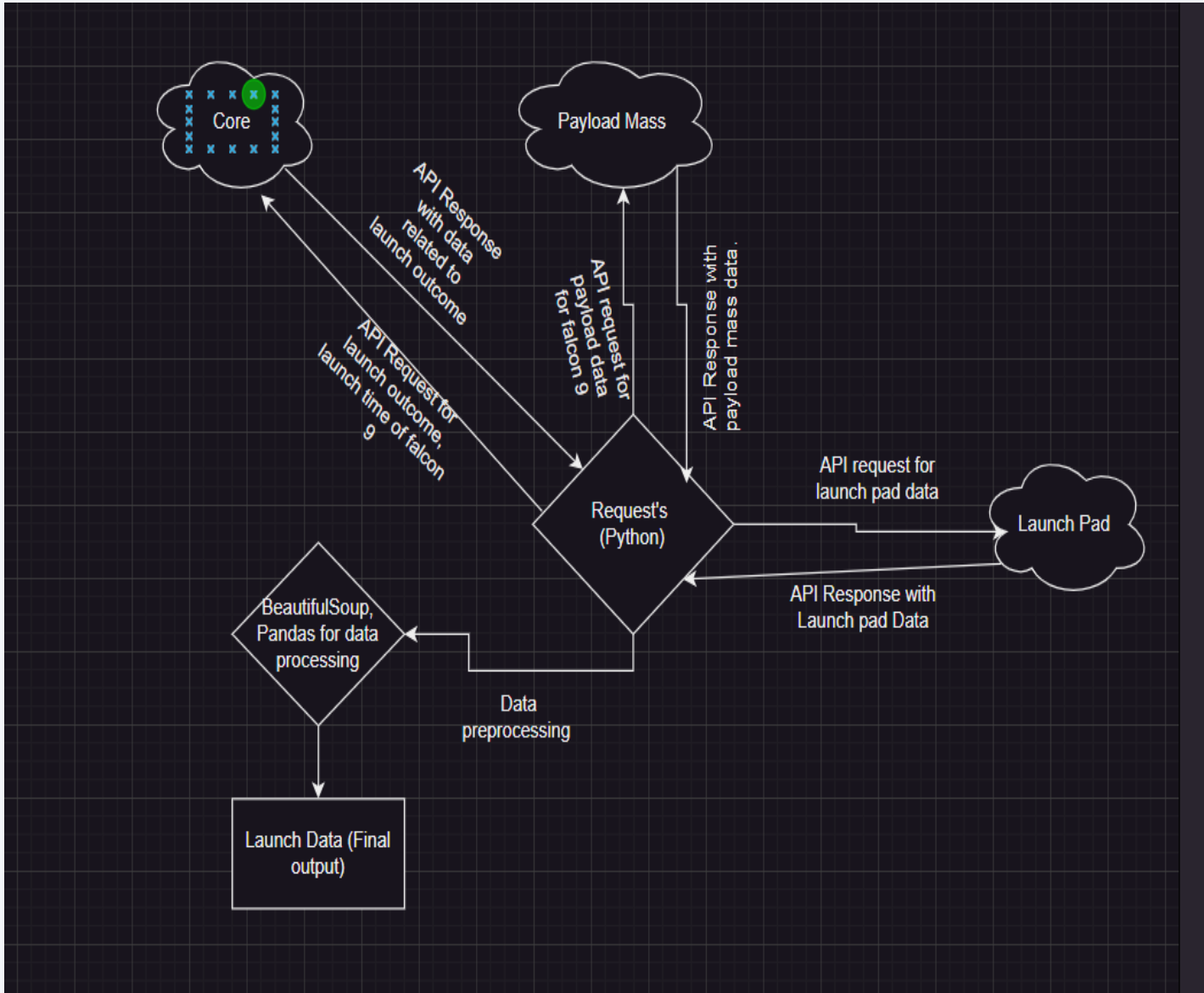
Data Collection – SpaceX API

- As mentioned in the previous slide, Three different APIs were used to collect data regarding launch data of falcon 9 rockets.
- 1) SpaceX's Core API is used to collect outcome of rocket launch and other data related to launch outcome.

2) SpaceX's payload is used to collect data about mass of payload and destined orbit.

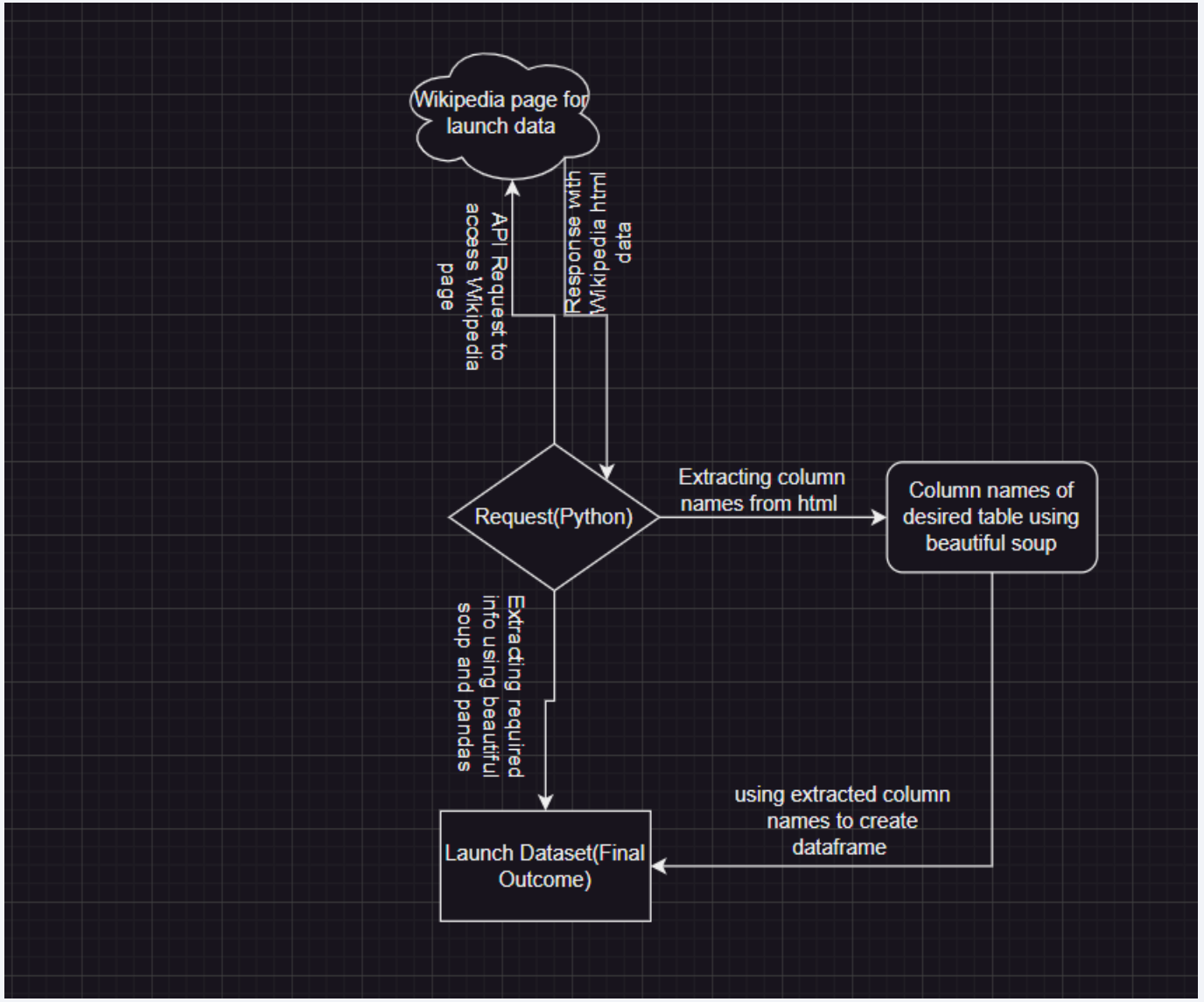
3) SpaceX's launchpad is used to collect data like name of launch pad, latitude and longitude of launchpad.

- GitHub (Navigate to jupyter-labs-spacex-data-collection-api.ipynb)



Data Collection - Scraping

- As mentioned earlier, Wikipedia page had a lot of tables regarding falcon launches. Using beautiful soup and pandas, required data is extracted and created a launch dataset.
- Github Link (Navigate to Jupyter _labs _web scraping.ipynb)



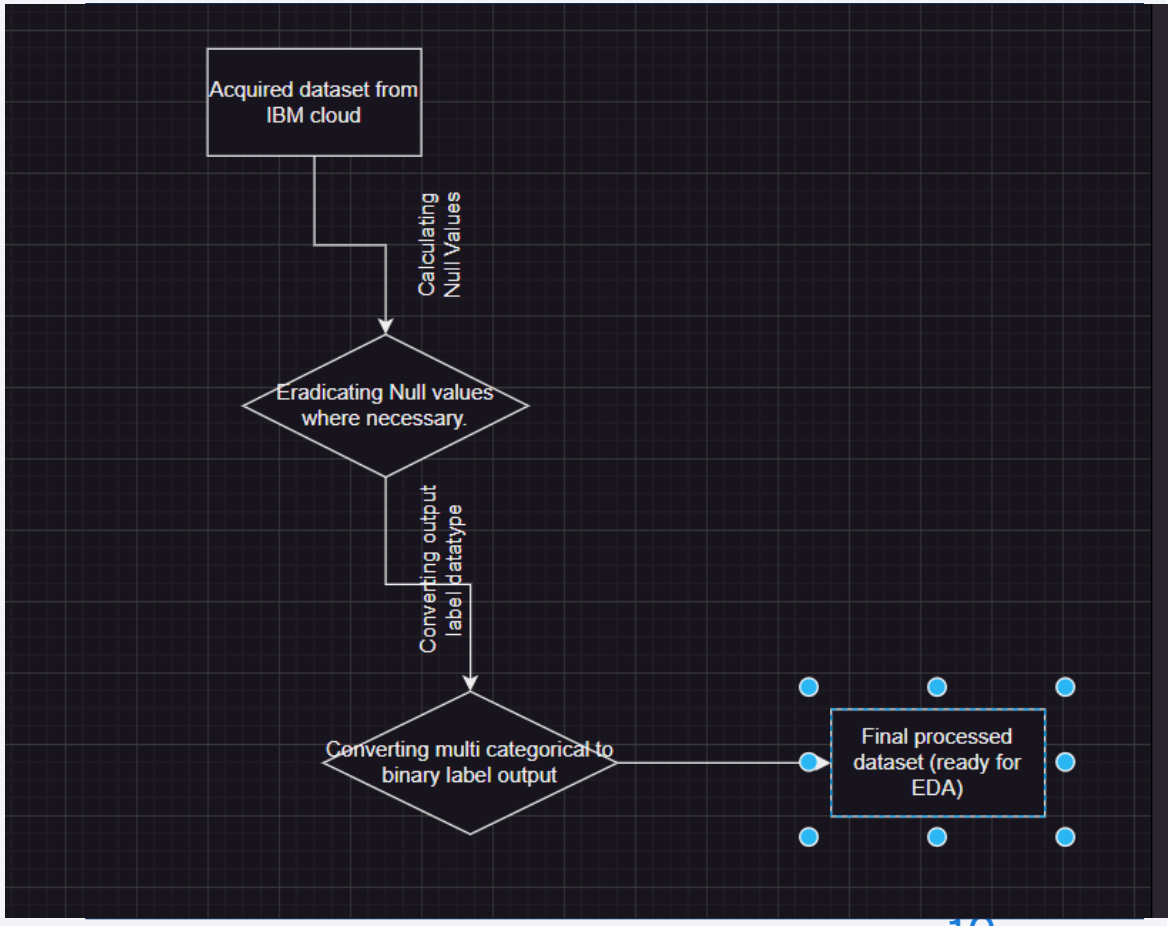
Data Wrangling

- The data that was used to make predictive analysis was from IBM cloud dataset.
- The Data Wrangling steps as follows:

1) Checking Null values in the data and taking actions accordingly.

2) Since the output label is categorical in nature and has different types of outcome based on launch type and launch success. The output label was converted into binary classifier label having ‘0’ for failed landing and ‘1’ for successful landing of first stage rocket.

- Github Link (Navigate to labs-jupyter-spacex-wrangling)

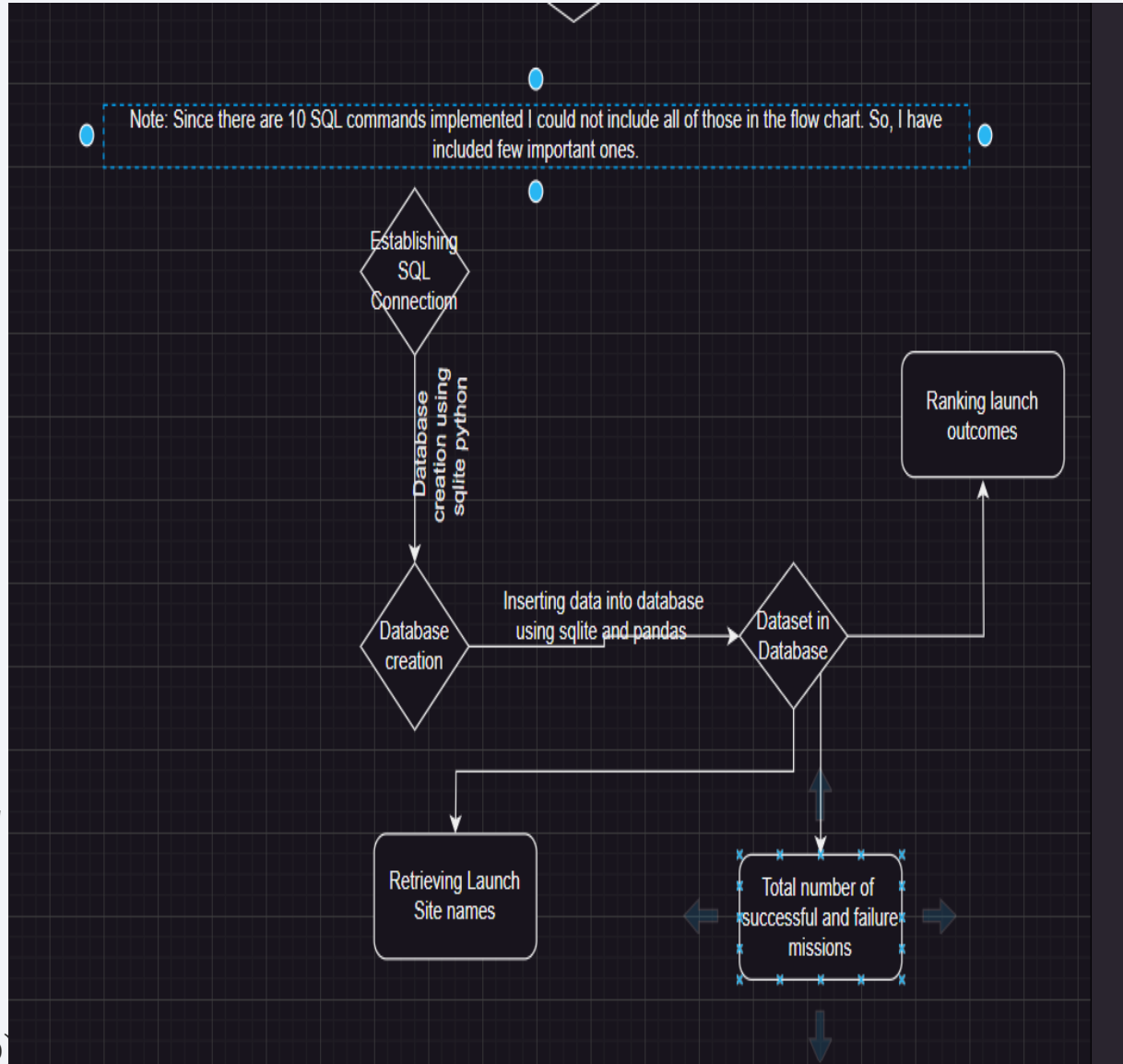


EDA with Data Visualization

- During EDA, To understand the impact of launch success by factors like launch site, payload mass etc… we have plotted graphs to understand relationship between various factors and how it is impacting success.
- We have used scatterplots, bar graphs, line plot for EDA.
- Scatter plot- To find relationship between Flight Number vs Payload Mass, Flight Number vs Launch Site etc….
- Bar Graphs- Used to measure success rate of each orbit, line plot- It is used to analyse timely success rate for each year.
- Github Link

EDA with SQL

- SQL Queries that were used during EDA, as follows:
- Launch site names: %sql select DISTINCT("Launch_Site") from SPACEXTABLE1;
- 5 records start with ‘CCA’ in launch site: %sql select * from SPACEXTABLE1 where "LAUNCH_Site" Like "CCA%" Limit 5;
- Total payload mass by NASA CRS Boosters: %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE1 where Customer is "NASA (CRS)" ;
- Average payload mass of booster f9 v1.1: %sql select AVG(PAYLOAD_MASS__KG_) from SPACEXTABLE1 where Booster_Version is "F9 v1.1" ;
- Date of first successful landing of landing pad: %sql select min(Date) as "first successful landing" from SPACEXTABLE1 where "Landing_Outcome" like "%(ground pad)%" and "Mission_Outcome" is "Success";
- booster version having success in landing pad and has payload mass between 4000 and 6000: %sql select "Booster_Version" from SPACEXTABLE1 where "Landing_Outcome" is "Success (drone ship)



EDA with SQL

- Booster version having success in landing pad and has payload mass between 4000 and 6000: %sql select "Booster_Version" from SPACEXTABLE1 where "Landing_Outcome" is "Success (drone ship)" and "PAYLOAD_MASS__KG_" between 4000 and 6000;

- Sum of success and failure missions: %sql select "Mission_Outcome",count("Mission_Outcome") As total_number from SPACEXTABLE1 group by "Mission_Outcome" ;

- Booster versions with highest payload mass: %sql select "Booster_version", "PAYLOAD_MASS__KG_" as maximum_payload_mass from SPACEXTABLE1 where "PAYLOAD_MASS__KG_" is (select max("PAYLOAD_MASS__KG_") from SPACEXTABLE1);

- Failure landing outcome, launch pad name, booster version for the year 2015: %sql select substr(Date,6,2) as month_name,substr(Date,0,5), "Landing_Outcome", "Booster_version","Launch_Site" from SPACEXTABLE1 where "Landing_Outcome" Like "%Failure%" and substr(Date,0,5)='2015';

- Ranking the landing outcomes based on count of outcomes: %sql select Landing Outcome,count("Landing_Outcome") from SPACEXTABLE1 where Date between "2010-06-04" and "2017-03-20" group by "Landing_Outcome" order by count(Landing_Outcome) DESC;

- Github Link (Navigate to jupyter_labs_sql_lite_eda)

Build an Interactive Map with Folium

- Based on requirements, Marker, Circle, Icon, Polyline, Marker Cluster, Mouse Position were used to pinpoint launch sites and calculated distance between launch sites and city, railways and Highway.
- In a nutshell,

1) Marker, Circle and Icon are basically used to pinpoint a particular location in this case it is launch site, railways so on…

2) Marker Cluster- It is used to avoid clumsiness while adding more markers in close proximity as this helps to form cluster among markers of same coordinates.

3) Polyline- As the name suggests, This is a line drawn between two or more locations.

4) Mouse position- Mouse position is basically useful in finding out the coordinates as soon as a cursor is placed on a random location in map.

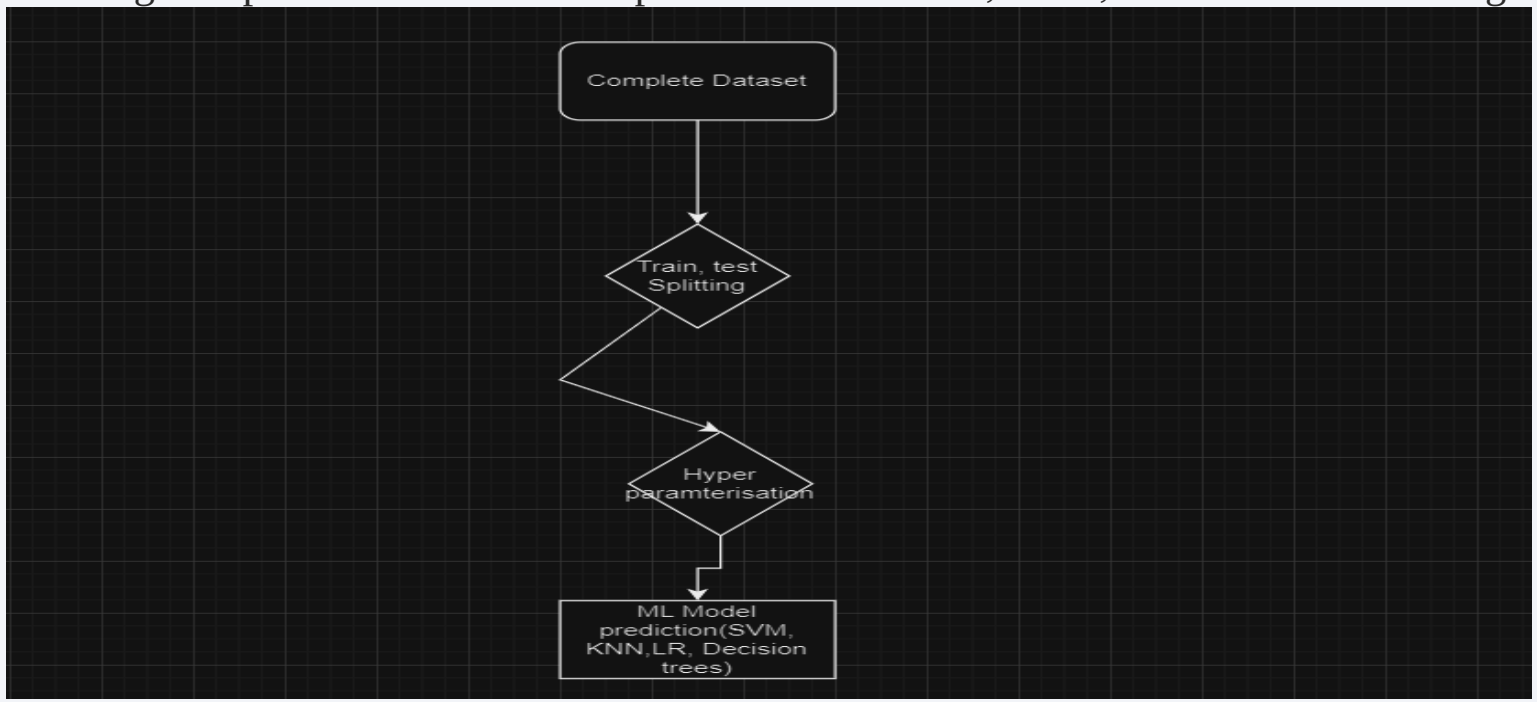
Github Link (Navigate to Jupyter_labs_launch_site_location_jupyterlite.ipynb)

Build a Dashboard with Plotly Dash

- Based on project requirements, we have used basic html tags like Div, Header tags and dash core components like dropdown, graph, range slider, pie chart, scatter plot.
- Since from EDA we understood that payload mass, launch site, booster version impacts successful landing of stage 1 of falcon 9.
- Drop down- This was used to select launch site to understand relation between that particular launch site with payload mass, booster version and its impact on success probability of stage 1 landing.
- Range slider- This range slider is used to select range of payload mass to understand its relation with booster version, launch site and success of launch.
- Pie chart- This chart shows the proportion success to failure of launches in a particular launch site.
- Scatter plot- This was plotted against payload mass and launch outcome to understand the impact of payload mass on launch success. In addition, it also tells relationship among booster version, payload mass and class for that particular launch site.

Predictive Analysis (Classification)

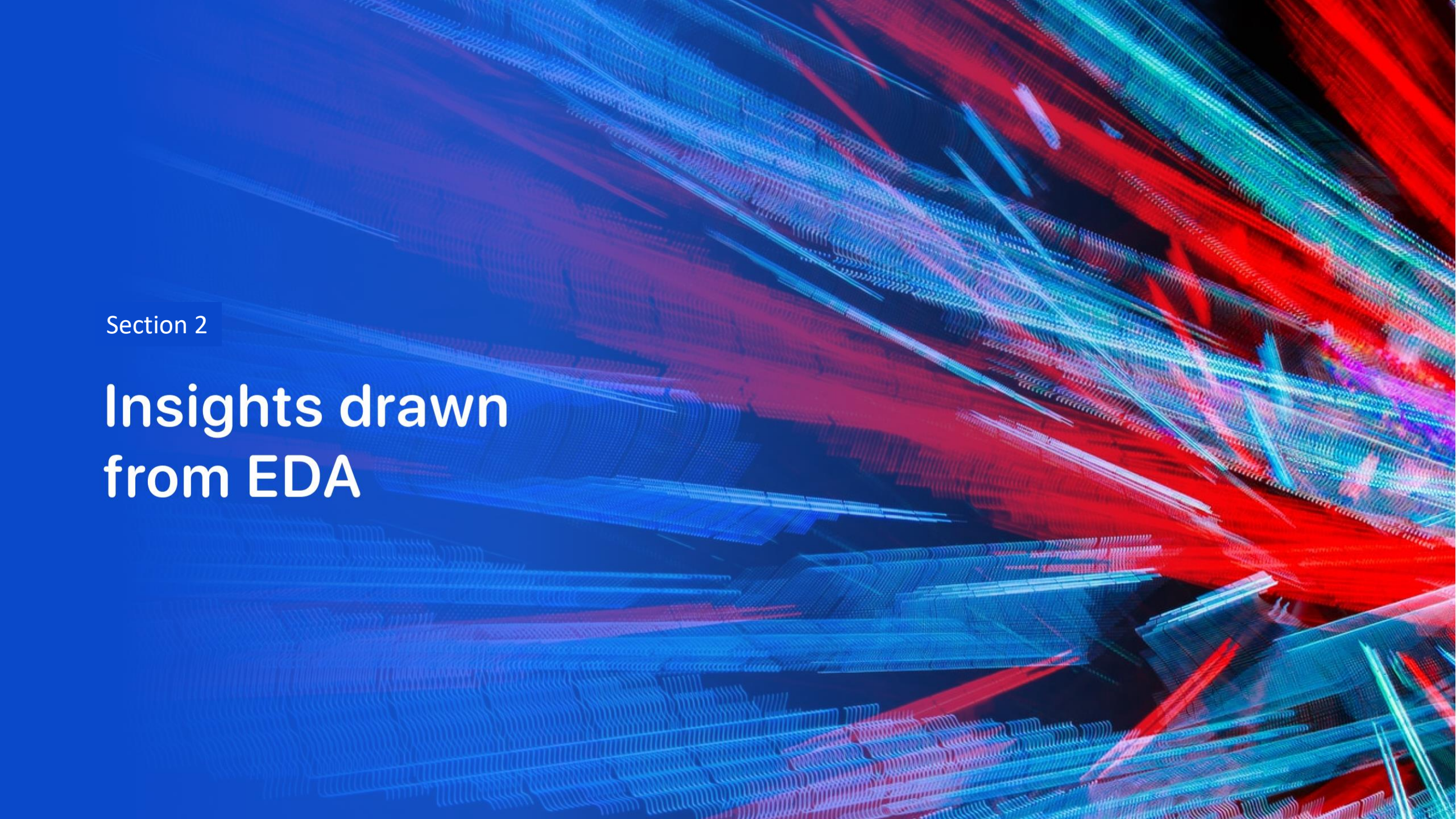
After conducting EDA, we understood that features like flight number, booster version, payload mass, launch location etc… were key impacting factors for predicting launch outcome. After having split data to train and test set hyperparameter tuning was performed to find best parameters for SVM, KNN, decision trees and Logistic regression.



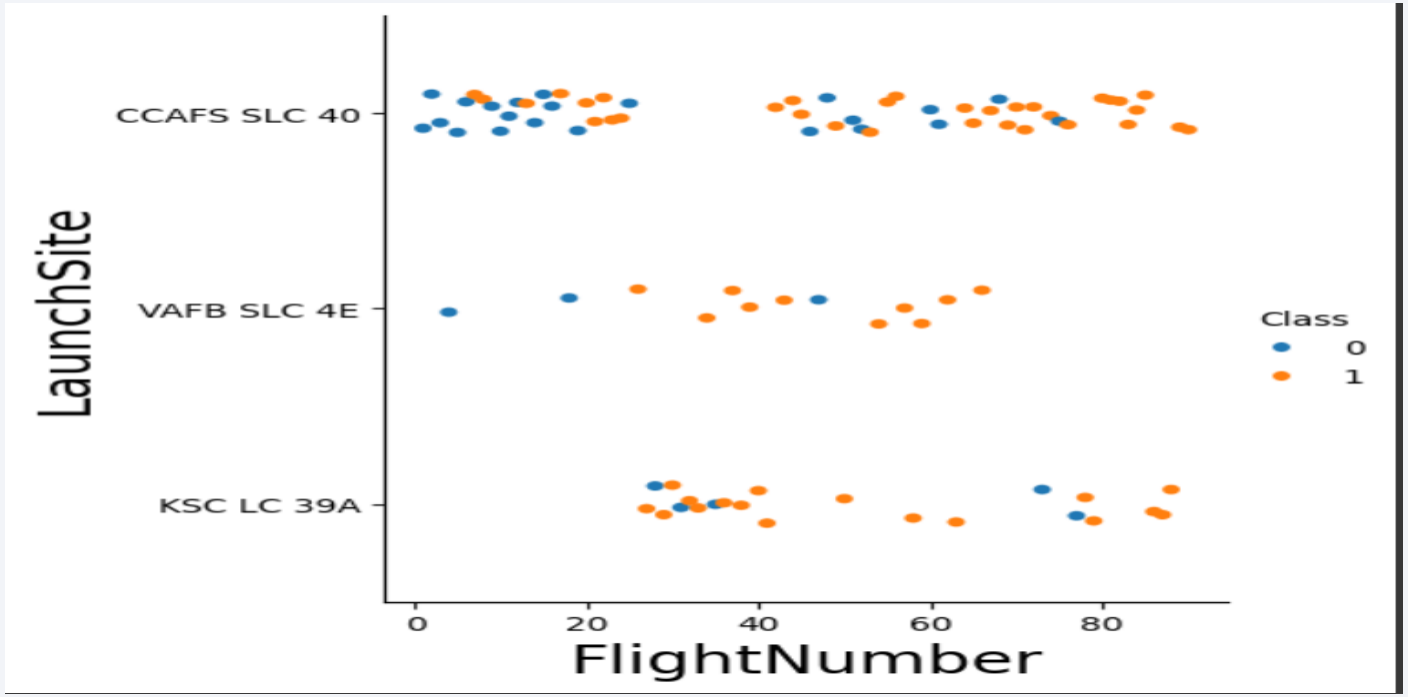
- Github Link

Results

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

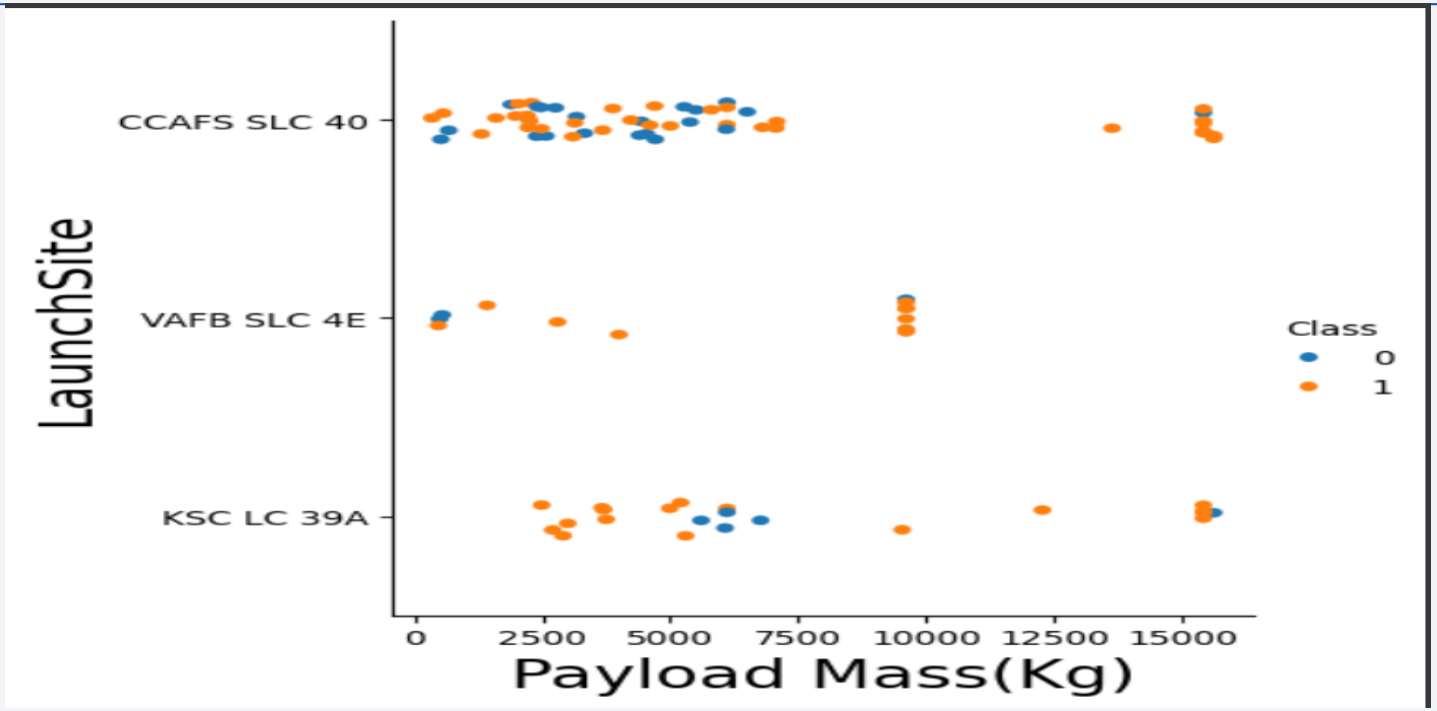

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



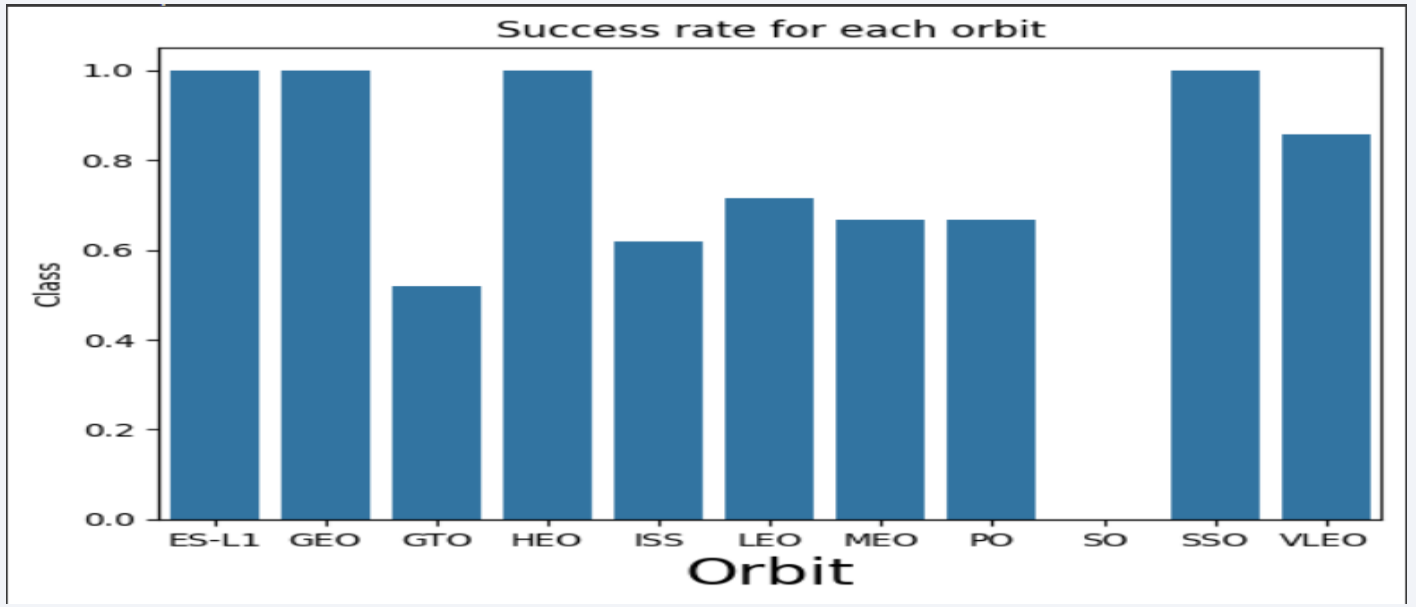
- It is clear that number of launches are unevenly distributed. The Flight number is directly proportional to launch success.
- KSCLC39A with significant number of launches has highest launch success rate amongst all. Apart from KSCLC39A, Lower flight number has more unsuccessful attempts in all launch sites irrespective of number of launches in each site.
- (Optional Suggestion)It is advisable to launch flight numbers less than ’20’ in KSCLC39A due to its high success rate and also it is surrounded with forest providing controlled air flow rather than having uncontrollable air flow in CCAFSSLC40 because of its close proximity with ocean and surrounded with barren land.

Payload vs. Launch Site



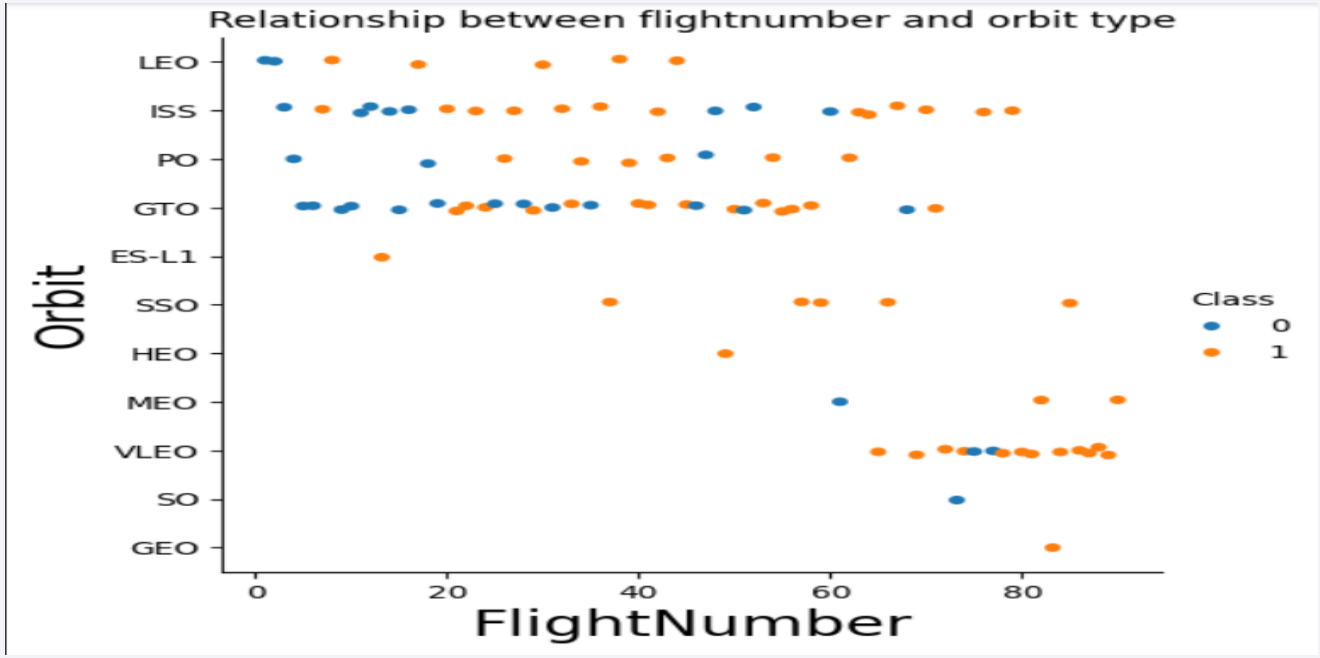
- It is clearly visible that payload mass has inverse relation with launch success up to 7500kg payload mass. In addition, most number of launches are concentrated with maximum payload of 7500kg.
- The payload mass more than 7500 kg has more success irrespective of launch sites which is contrary to our previous observation. KSCLC39A holds more success rate between 2500 to 7500kg. Having observed this trend, it is advisable to have more launches having payload more than 2500kg in launch site KSCLC39A. In case of VAFBSLC4E, we need to launch ones with more 10000kg mass payload to exactly understand how effective is this launch site in comparison to others.

Success Rate vs. Orbit Type



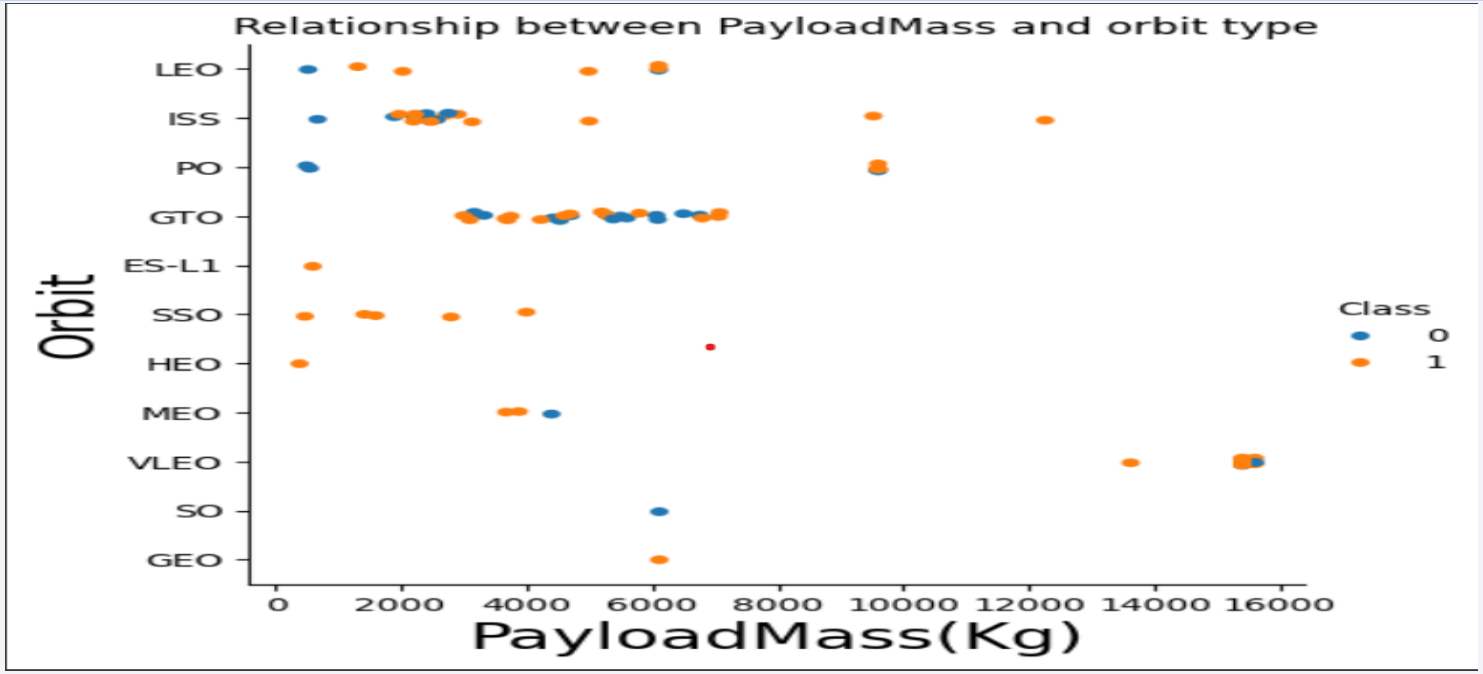
- The graph depicts that ES-L1, GEO, HEO has 100% success rate, but each have less than 5 launches. VLEO has 80% success rate with 12 launches.
- The orbits with max altitude of 20000km like ISS,LEO,VLEO,PO has seen more success than others. The average success probability is around 70% after combining all these orbits. The lowest success was observed for GTO with altitude of 35,000 km with more launches compared to other individual orbits.
- **It is clear from all these observations that more the altitude of orbit the lesser the chances of successful launches.**

Flight Number vs. Orbit Type



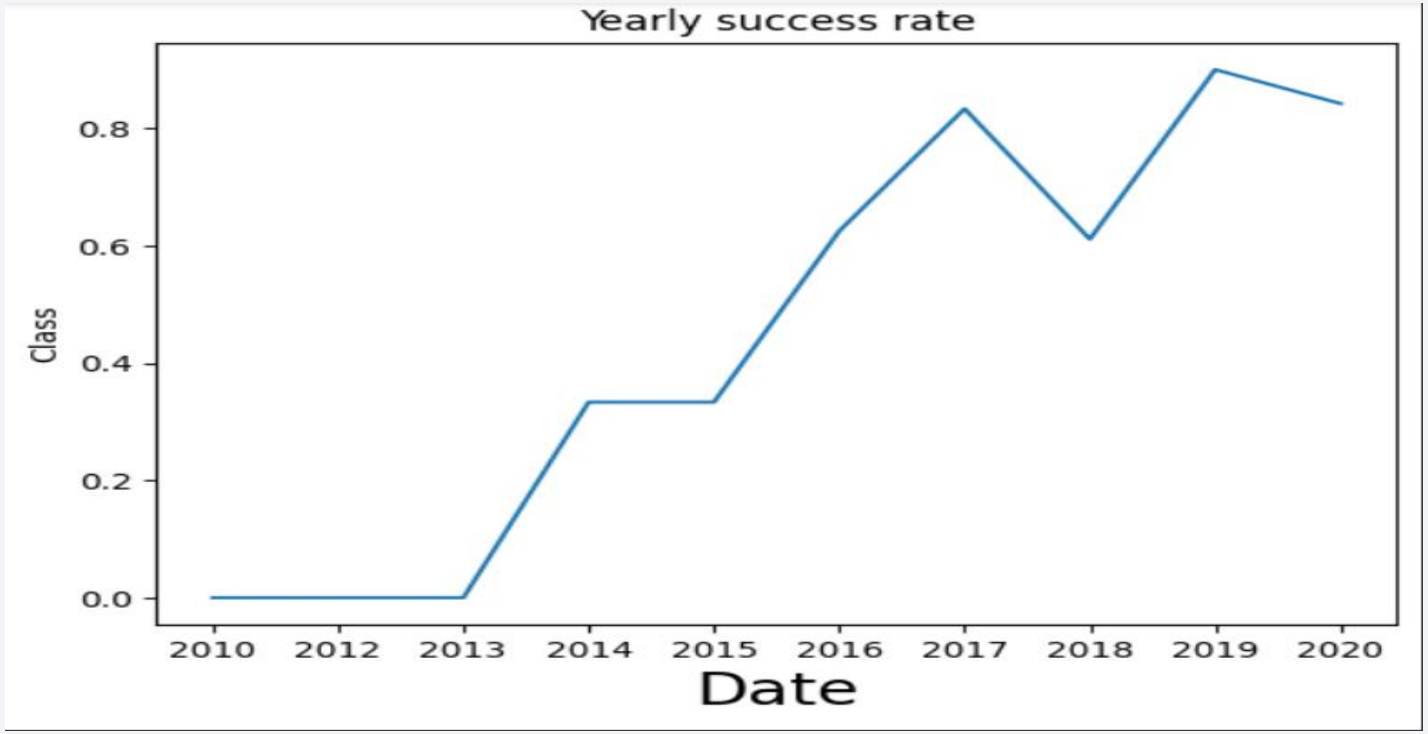
- It is clearly visible from graph that VLEO has highest success with success rate of 82% followed by LEO,ISS and GTO.
- The main reason behind high success rate for VLEO is its utilization of high flight number and the reason behind the high failure rate of GTO is around the utilization of 10 to 40 flight number which is very less compared to the altitude that GTO is above the earth. It is better to use high flight number for orbit that has more altitude.
- It is better to use low flight number for VLEO given that the altitude is around 450 km above earth which is very less compared to the altitude of GTO.

Payload vs. Orbit Type



- According to graph, VLEO has better success rate with heaviest payloads than other orbits. Though, SSO has 100% success rate it has fewer launches i.e., 5 which is very less compared to VLEO.
- Until below the altitude of 2000km the success rate of each orbit seems to be better than orbits like GTO,MEO etc… The possible reason is the payload as we can see that probability of successful launches increases for LEO,ISS after 2000 kgs but for GTO even after having a minimum payload of 4000 and maximum of 7000kg around has the lowest success rate amongst all.
- By these observations, we can say that as the altitude of the orbit increases the amount of payload need to increase gradually to increase the probability of launch success.

Launch Success Yearly Trend



- The graph clearly depicts that yearly success rate is increasing exponentially till 2020 despite of having drop between 2017 to 2018. The possible reason could be improvement in technology, introducing latest booster versions and more funding.

All Launch Site Names

```
%sql select DISTINCT("Launch_Site") from SPACEXTABLE1;
```

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

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

- The above SQLITE query provides unique launch sites in the dataset.

Launch Site Names Begin with 'CCA'

Task 2

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

```
%sql select * from SPACEXTABLE1 where "LAUNCH_Site" Like "CCA%" Limit 5;
```

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

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

- The above SQLITE query displays top 5 rows that has launch site name starting with ‘CCA’.

Total Payload Mass

Task 3

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

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE1 where Customer is "NASA (CRS)" ;
```

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

sum(PAYLOAD_MASS__KG_)
45596

- The above SQL query provides you the total payload mass carried by all NASA CRS rockets i.e., 45596.

Average Payload Mass by F9 v1.1

Task 4

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

```
%sql select AVG(PAYLOAD_MASS__KG_) from SPACEXTABLE1 where Booster_Version is "F9 v1.1" ;
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

- The above query provides the average payload mass carried by F9 V1.1 booster version i.e., 2928.4kg.

First Successful Ground Landing Date

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 "first successful landing" from SPACEXTABLE1 where "Landing_Outcome" like "%(ground pad)%" and "Mis
```

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

first successful landing
2015-12-22

- The above sql query presents the firs successful landing on ground pad i.e., 22-12-2015.

Successful Drone Ship Landing with Payload between 4000 and 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 "Booster_Version"  from SPACEXTABLE1 where "Landing_Outcome" is "Success (drone ship)" and "PAYLOAD_MASS__KG_" b
```

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

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

- The query above present the names of successful booster versions in drone ship which has payload mass between 4000 to 6000

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("Mission_Outcome") As total_number from SPACEXTABLE1 group by "Mission_Outcome" ;
```

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

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

- The above query provides the total number of successful and failure missions.

Boosters Carried Maximum Payload

Task 8

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

```
32]: %sql select "Booster_version", "PAYLOAD_MASS__KG_" as maximum_payload_mass from SPACEXTABLE1 where "PAYLOAD_MASS__KG_" is (s
 * sqlite:///my_data1.db
Done.
```

32]:

Booster_Version	maximum_payload_mass
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

- The above query presents booster_version names that carried maximum payload.

2015 Launch Records

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, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%sql select substr(Date,6,2) as month_name,substr(Date,0,5), "Landing_Outcome", "Booster_version","Launch_Site" from SPACEX
```

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

month_name	substr(Date,0,5)	Landing_Outcome	Booster_Version	Launch_Site
01	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- The above query displays month names, landing_outcomes, booster versions which had failure landing in the year 2015.

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

Task 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.

```
%sql select Landing_Outcome,count("Landing_Outcome") from SPACEXTABLE1 where Date between "2010-06-04" and "2017-03-20" grou
```

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

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

Reference Links

- The above query presents count of ten different landing outcomes.

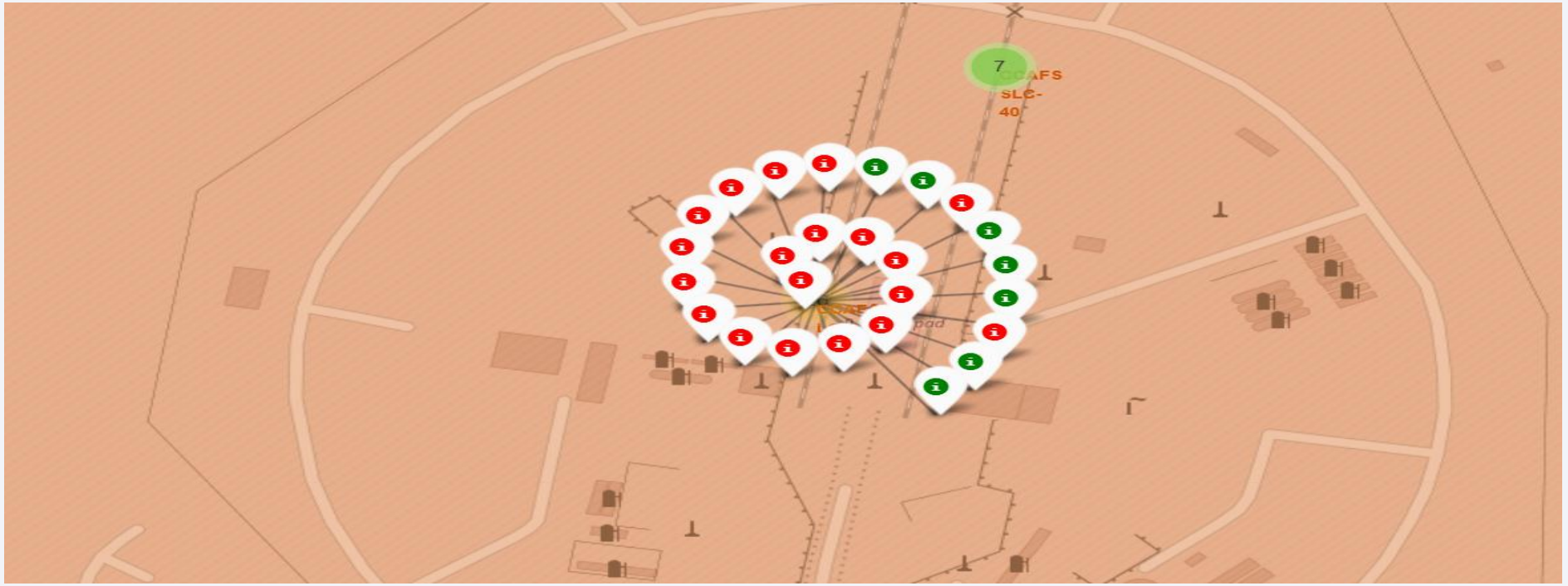
Section 3

Launch Sites Proximities Analysis

LAUNCH SITES ON FOLIUM

- Totally, there are 4 different launch sites. Each launch site is very close to coastal line and also connected with highways, railways and all these locations are little far away from cities.
- Out of these launch sites, VAFBSLC4E is very close to coastal line and KSCLC 39A is surrounded with forest and has the highest success rate out of all launch sites.

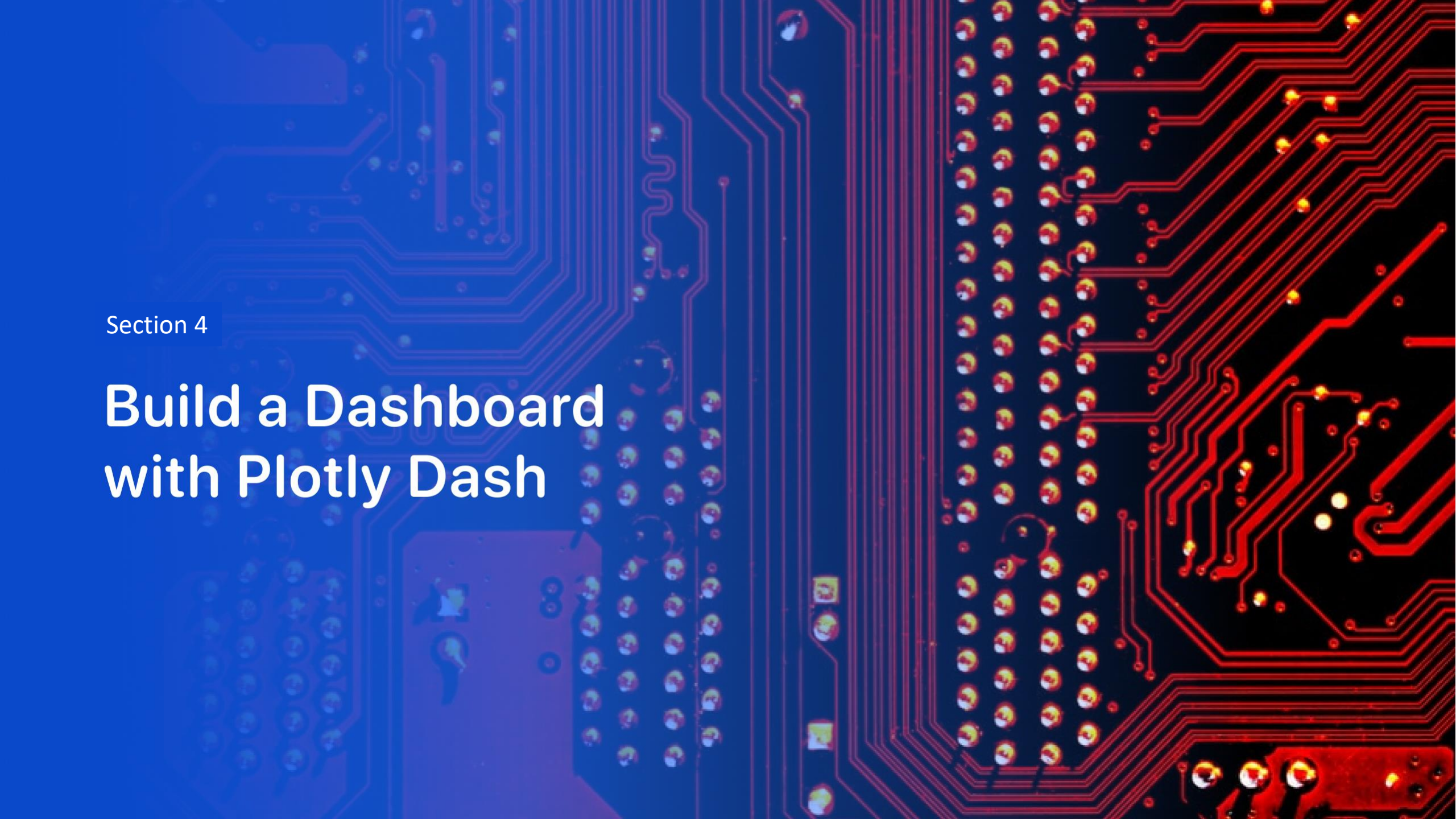
LAUNCH OUTCOMES ON FOLIUM MAP



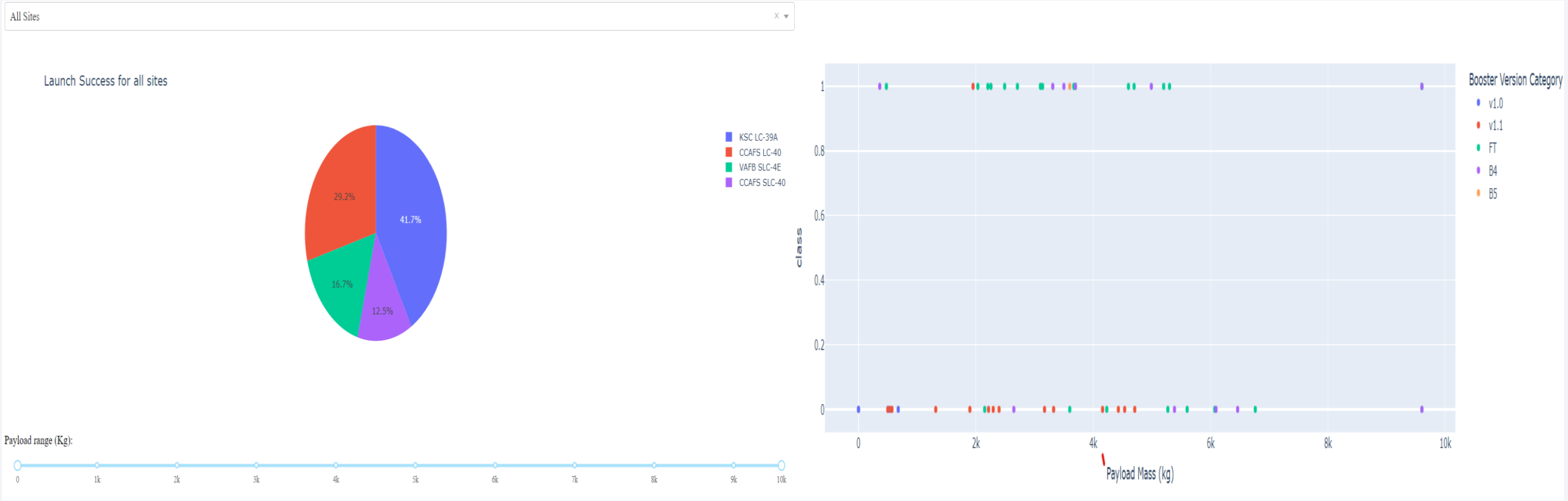
- The above figure depicts the launch outcomes of CCAFSLC-40 launch site which has the highest number of failure launches

DISPLAYING CLOSE PROXIMITIES TO A LAUNCH SITE

- The above picture display's close proximity to railways, highways and nearest cities to KSCLC39A launch site which has the highest launch success rate and the possible reason could be the controlled weather conditions like wind speed as it is surrounded by forests in three directions but the other sites are in barren lands and are very close to coastalline.

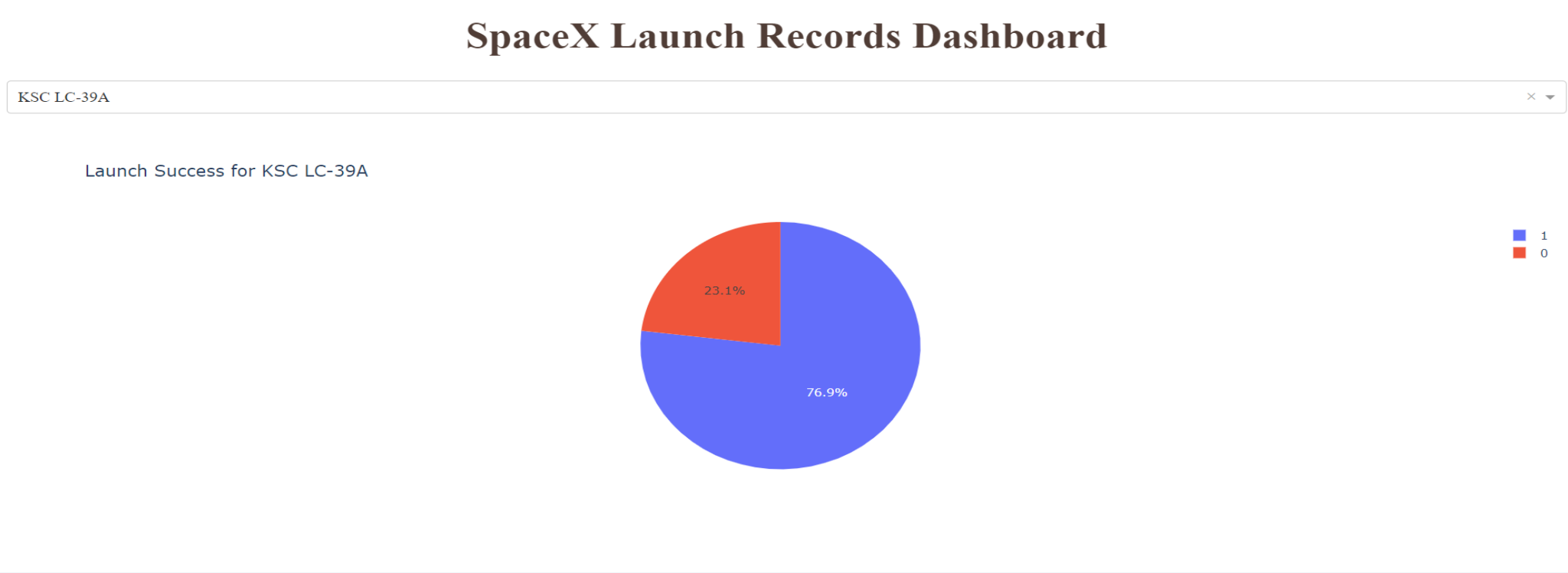

Section 4
Build a Dashboard with Plotly Dash

LAUNCH SUCCESS FOR ALL SITES



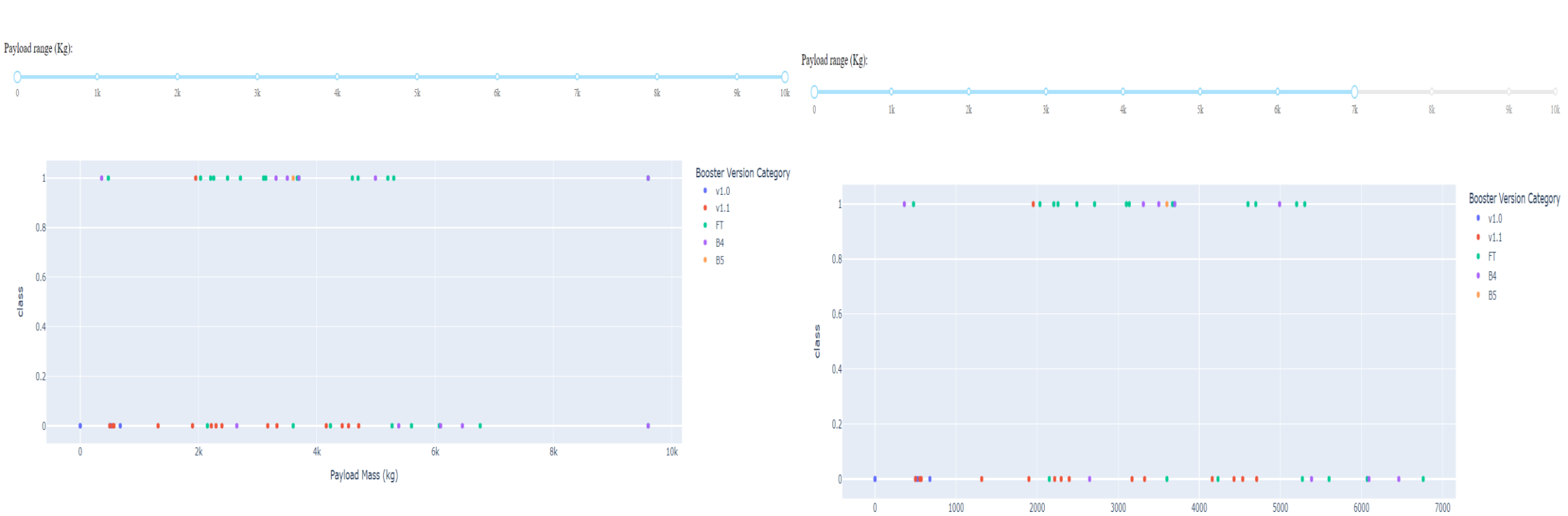
- From the graph, It is clear that we have got 21 success launches and out of this it is clear that more number of successful launches are in between 2000 kg to 4000 kg.

HIGHEST SUCCESS RATE AMONG ALL LAUNCH SITES



- It is clearly visible that KSCLC39A has the success rate among all the launch sites. Apart from KSCLC39A, all other sites are very close to coastal line and are present barren lands but this site is surrounded with forest which may help to maintain controlled wind speed and possibly the reason behind the highest launch success.

PAYLOAD VS LAUNCH OUTCOME FOR ALL SITES

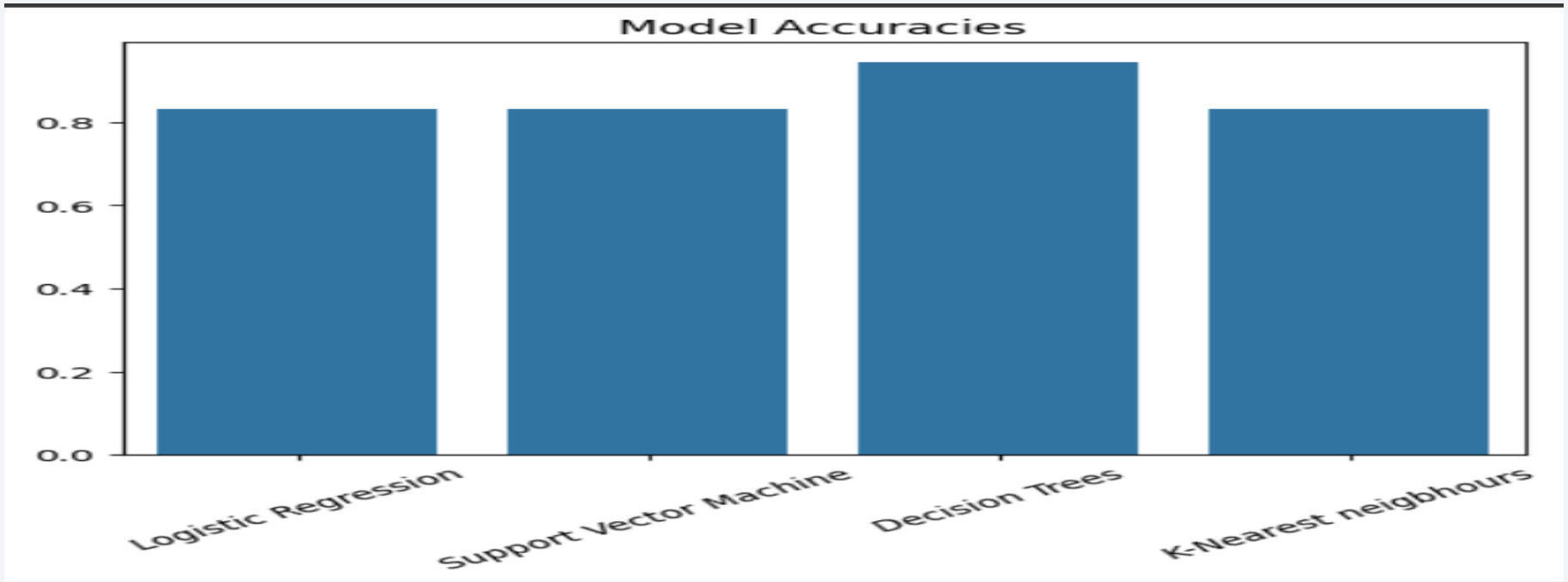


- From the graph it is clearly visible that booster version FT has the success rate and V1.1 has the lowest success rate. In addition, It is also observed that most of successful launches were between 2000kg to 6000 kg payload mass.

Section 5

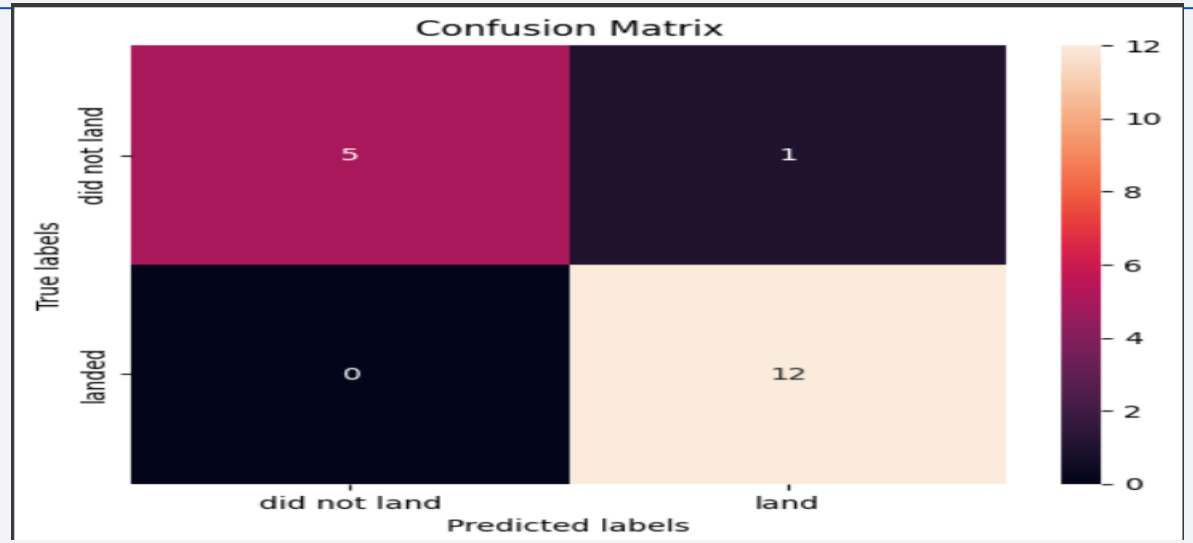
Predictive Analysis (Classification)

Classification Accuracy



- It is clear that decision trees is performing better than other models with accuracy of 94.4%.

Confusion Matrix

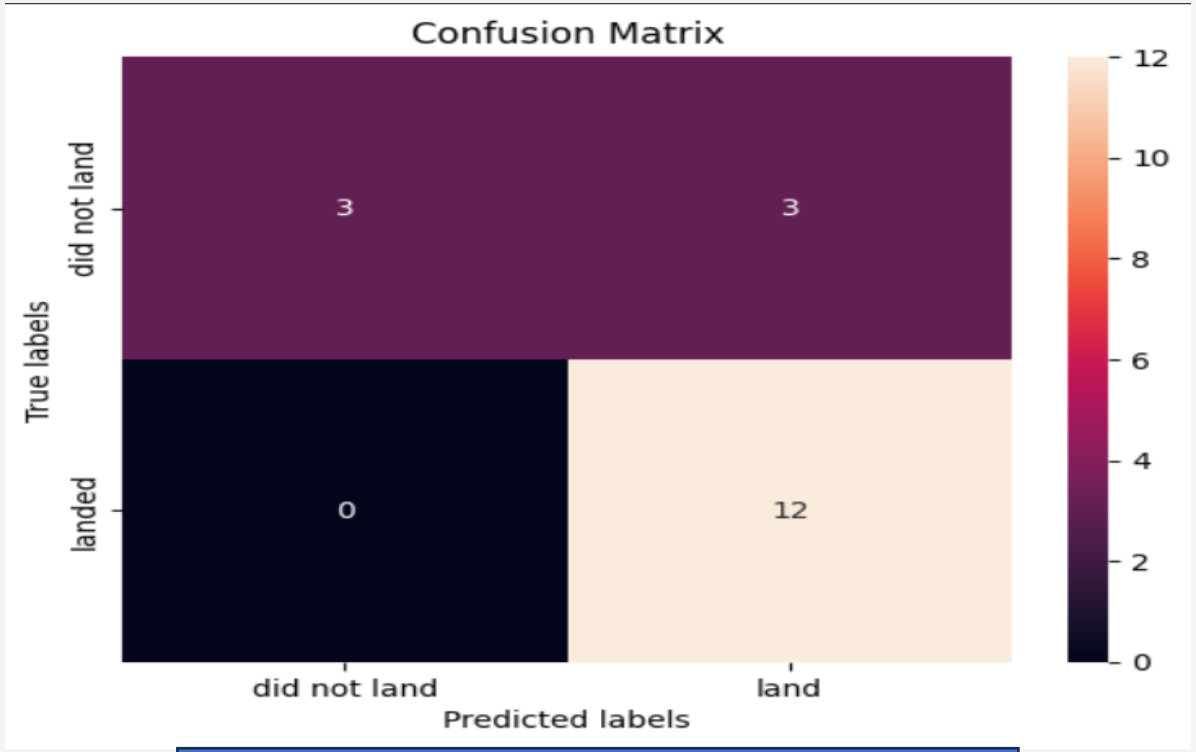


- As mentioned before, the decision trees is performing better than others with an accuracy of 94.4%. The other models are having accuracy of 83.33% and does have problem with having more false negatives than compared to decision trees.

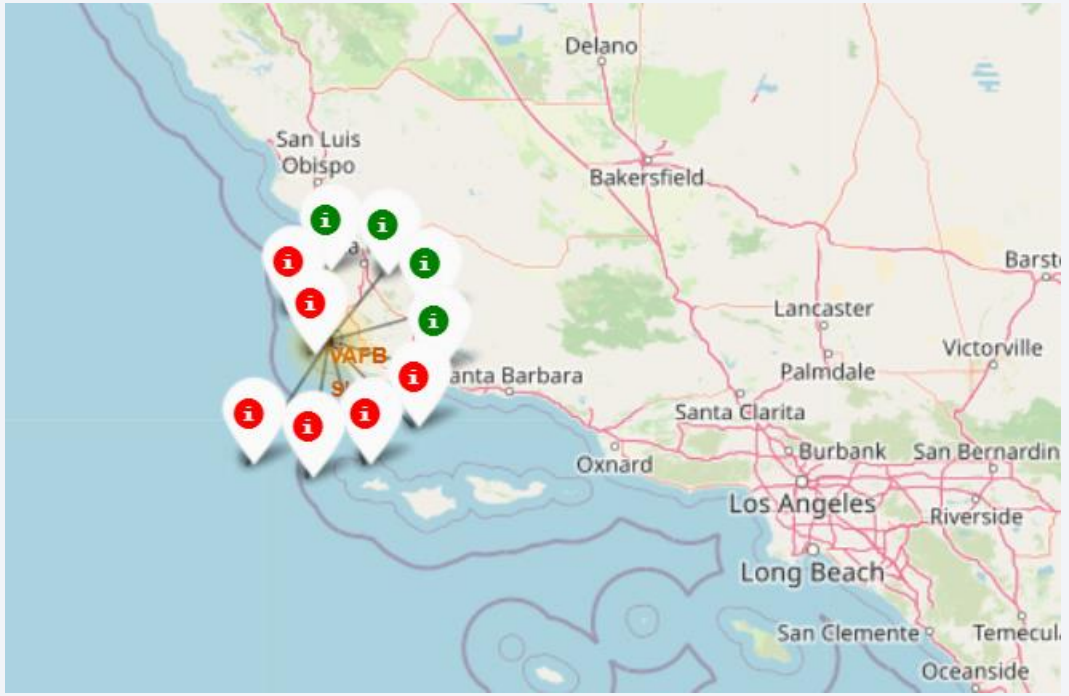
Conclusions

- From above observations, All launches are done amongst 4 launch sites. KSCLC39-A having highest success rate of around 80% amongst all.
- Most launches are done to these three orbits which are GTO, ISS,VLEO. Out of all VLEO has the highest success rate of 82% followed by LEO. Though we have orbits with 100% success but they are not considered as number of launches in those orbits are fewer than 5. The reason behind successful launches behind LEO,VLEO are because of having less altitude distance than other orbits i.e., 2000 km max.
- From previous observation, it was clear that flight number is directly proportional to launch success.
- It was observed that payload mass for orbits VLEO is very high and whereas for long distance orbits like GTO it was around 4000-6000 which is very less than the former orbits. Due to having used more payload mass for smaller orbits the success rate for these orbits are higher and comparatively, using very less for long distant orbits like GTO leading to less success rate.
- Payload mass is directly proportional to distance of orbit from earth. The more payload mass for more distant orbit the more probability of success.
- It is clearly visible that yearly launch success is increasing until 2020 despite of having dip between 2017 to 2018. The possible reason could be using of latest technology over time.

Appendix



Confusion matrix of SVM, KNN, Logistic Regression



Location of VAFBSLC4E Launch Site on map

Thank you!