IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mahsa Ghavipanjehtorkamani

02/21/2023

Outline

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

Executive Summary

- In this project, we want to predict Success or Failure of SpaceX landing, So we will use classification models and we applies four classification models: Logistic Regression, SVM, Decision Tree, KNN.
- We applied all these models and we get SVM as the best model then we applied GridSearchCV to find the best parameters for our models, after applying best parameters to our models we can see all models have same accuracy.

Introduction

- SpaceX company advertises Falcon9 rocket with the cost of 62 million dollars while other companies provides the cost upward of 165 million dollars. Because they believe that if Falcon9 land successfully in the first stage they can reuse that.
- The aim of this project is to predict success or failure of Falcon9 landing and by using model we can improve the Falcon9 then we can increase accuracy of landing successfully.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- 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 Collected by two ways:
  - Rest API
  - Web Scraping.
- You need to present your data collection process use key phrases and flowcharts

Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

- Data Collection with API GitHub URL

1.get response from URL

2.normalize .JSON file of response

3.get intended columns

4.get intended values from intended columns

5.create a dictionary and add values in that

6.convert dataset to pandas data frame

7.save dataset as .csv file

Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts

- Data Collection with Web Scraping GitHub URL

1.get response from URL
2.create a soup by using Beautiful Soup
3.find and extract all tables in created soup
4.select intended table
5.get dataset column name from selected table
6.get values of column from selected table
7.assign column names and column values to a dictionary
8.convert dictionary to pandas data frame
9.save dataset as .csv file

Data Wrangling

- In this Project, for data wrangling, we handle the missing data . We define class column in our dataset to define a class for each record based on their outcomes. We have 2 different class 0 for failed outcomes and 1 for successful ones.
- Data Wrangling GitHub URL

EDA with Data Visualization

- In this Project, we used scatter plot for Flight Number vs. Launch Sites, Payload Mass vs. Launch Sites, Flight Number vs. Orbit Type ,Orbit Type vs. Payload Mass to see relation between them, line plot to see trends of success during years and bar plot to see success rate for each Orbit Type.
- Data Visualization GitHub URL

EDA with SQL

- SELECT query to get : unique names of launch sites, list all flights that has been in launch sites that starts with ‘CCA’ , total payload mass when customer is ‘NASA’, average payload mass for ‘F9 v1.1’, date of first successful landing, booster version when payload mass is between 4000 and 6000, total number of failure and success landings, booster version that is carried maximum payload mass, count successful landing between two specific date.
- SQL GitHub URL

Build an Interactive Map with Folium

- In folium map, we used circle and marker to display name of each launch sites and their name, we used cluster marker to demonstrate successful and failed landing in each launch site and actually we used lines to show distance between each launch site and other locations like coastline, railroad, highway, city and etc.
- Folium Map GitHub URL

Build a Dashboard with Plotly Dash

- In this project, we created a plotly dashboard to have an interactive plots. In this Dash, we used pie chart and scatter plot. Pie chart to see percentage of success in each launch site and percentage of failure and success in each launch site and in different range of payload mass. Scatter plot to show success and failure in different payload and launch site.
- Plotly Dash GitHub URL

Predictive Analysis (Classification)

- In this project, we want to predict success and failure, so we need use classification models. Logistic regression model, support vector machine, decision tree, K nearest neighbors have been deployed. After deploying these models, we used R-squared and matrix confusion to evaluate models and then by using grid search we improved our models.
- Modeling GitHub URL

Results

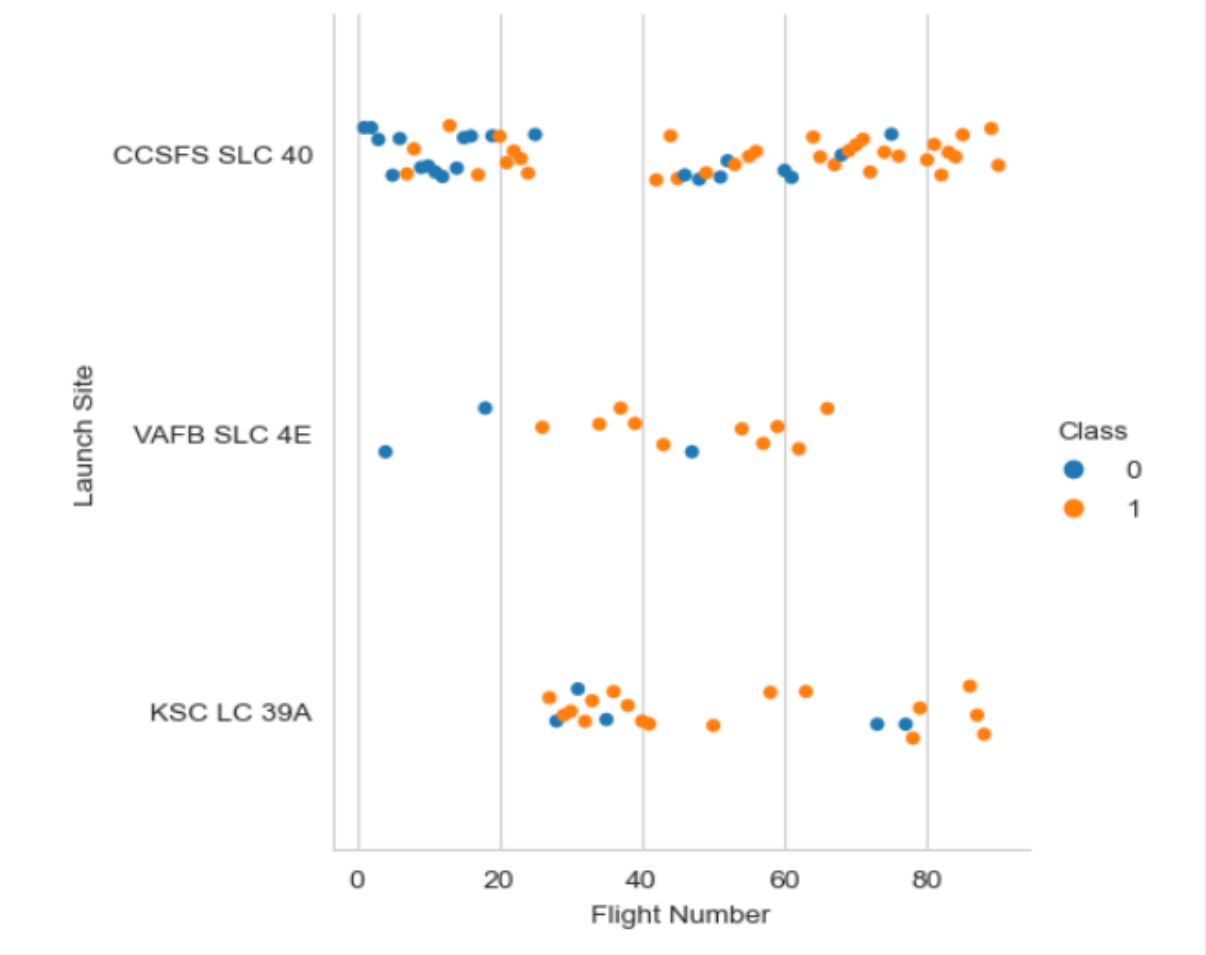
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

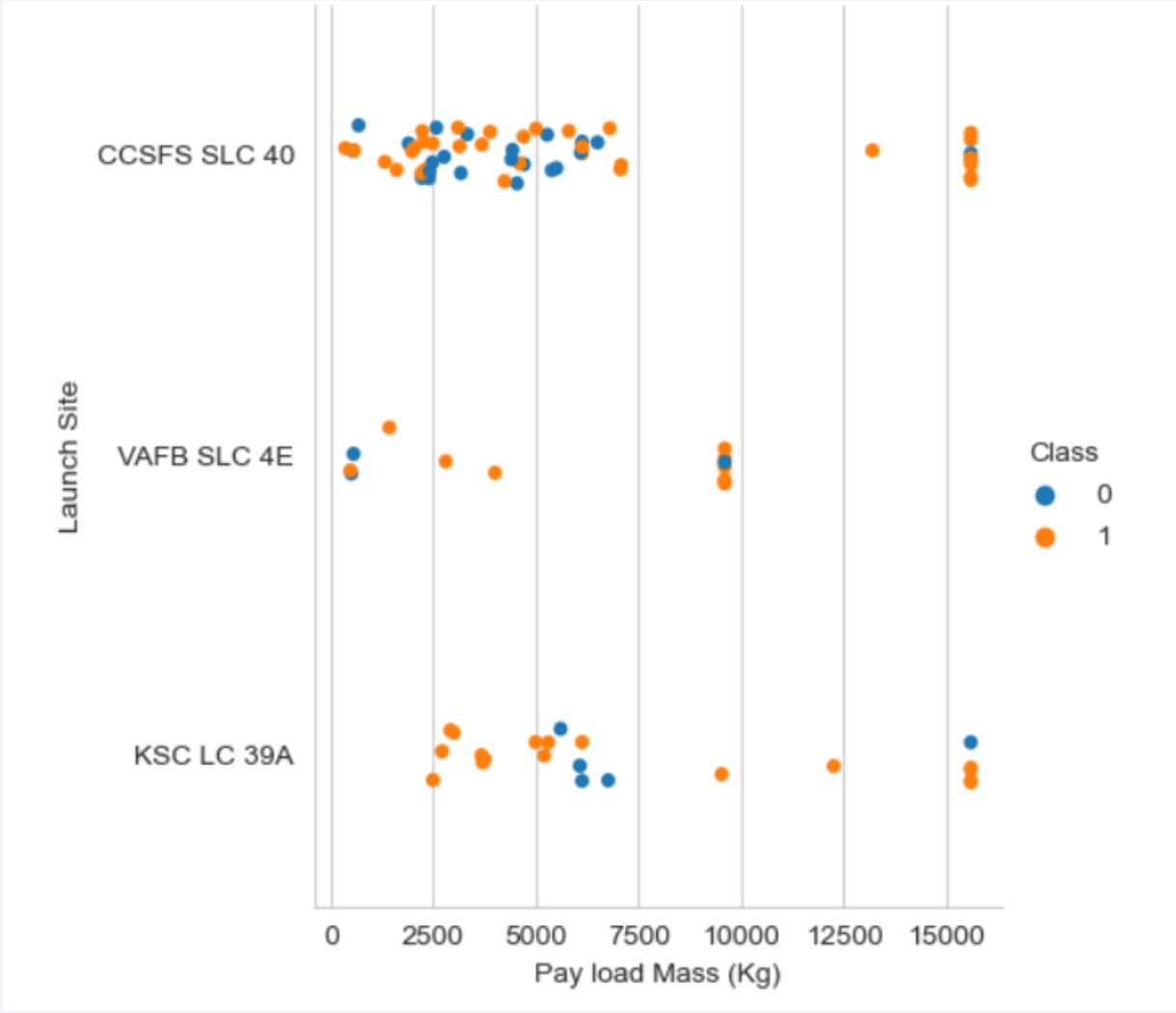
Flight Number vs. Launch Site

In this plot, we can see for different launch site with different flight numbers we have varies results. for launch site CCSFS SLC 40, we can see in the first flights number of failure is more than success but by increasing the flight number the number of success landing is increasing. in launch site VAFB SLC 4E,we don't have enough number of flights to predict success or failure of landing but most of them are successful for flight numbers greater than 20. for launch site KSC LC 39A,we don't have any flight numbers less than 20 but greater than 20,most of them are successful.



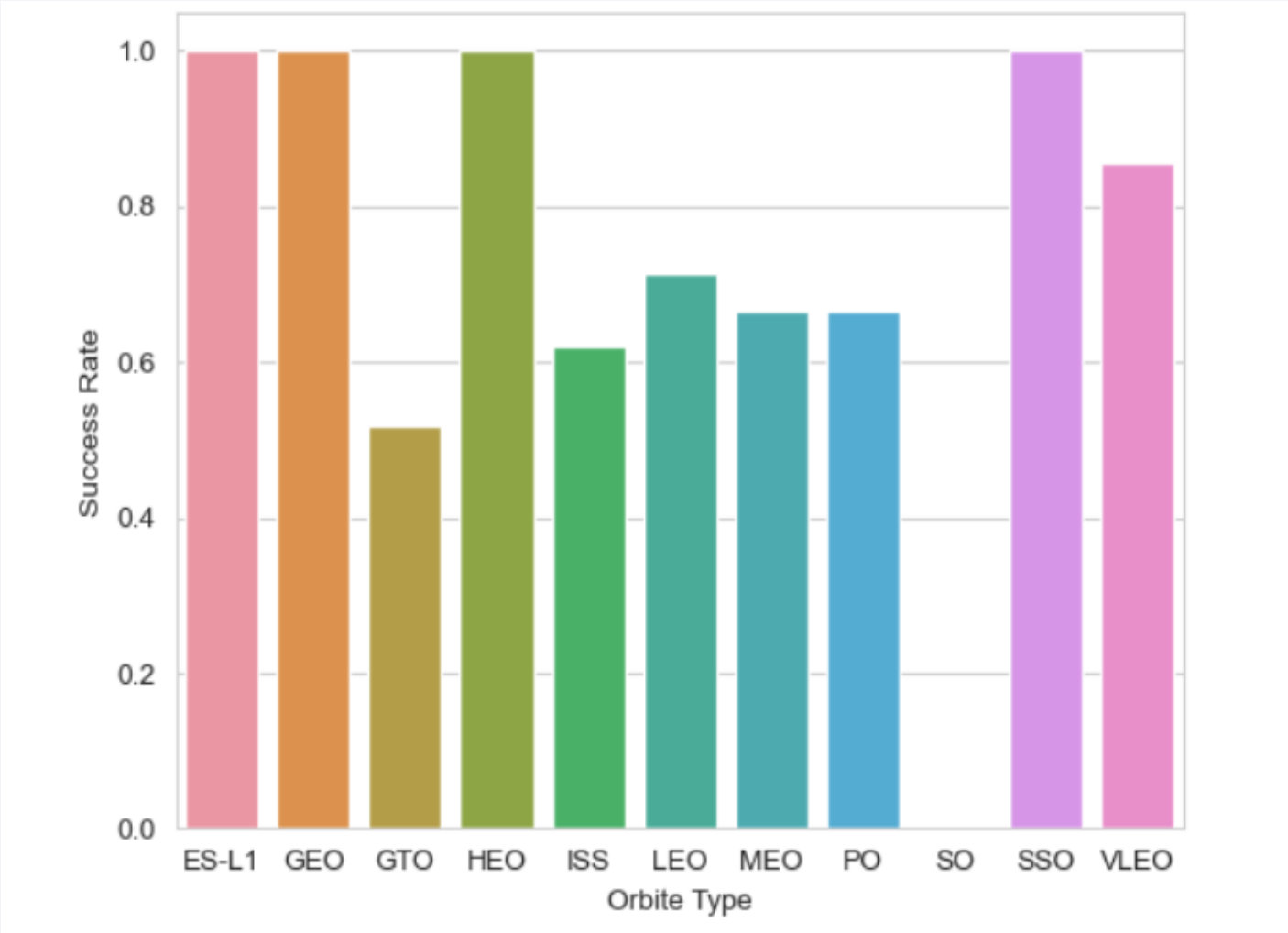
Payload vs. Launch Site

In this plot, we can see different amount of Payload mass in different Launch site and their landing result. For Launch site KSC LC 39A, they have pay load mass greater than 2500 and most of the are successful, only when payload mass is between 5000 and 7500 Kg we have more failure result. For launch site VAFB SLC 4E,we don't have enough data, most of the have payload mass between 0 and 5000 and we have few data when payload is near 10000 and failure and success result are almost same. For Launch site CCSFS SLC 40,most data have payload mass between 0 and 7500 and result almost same but payload mass greater than 12500 most of landing are successful.



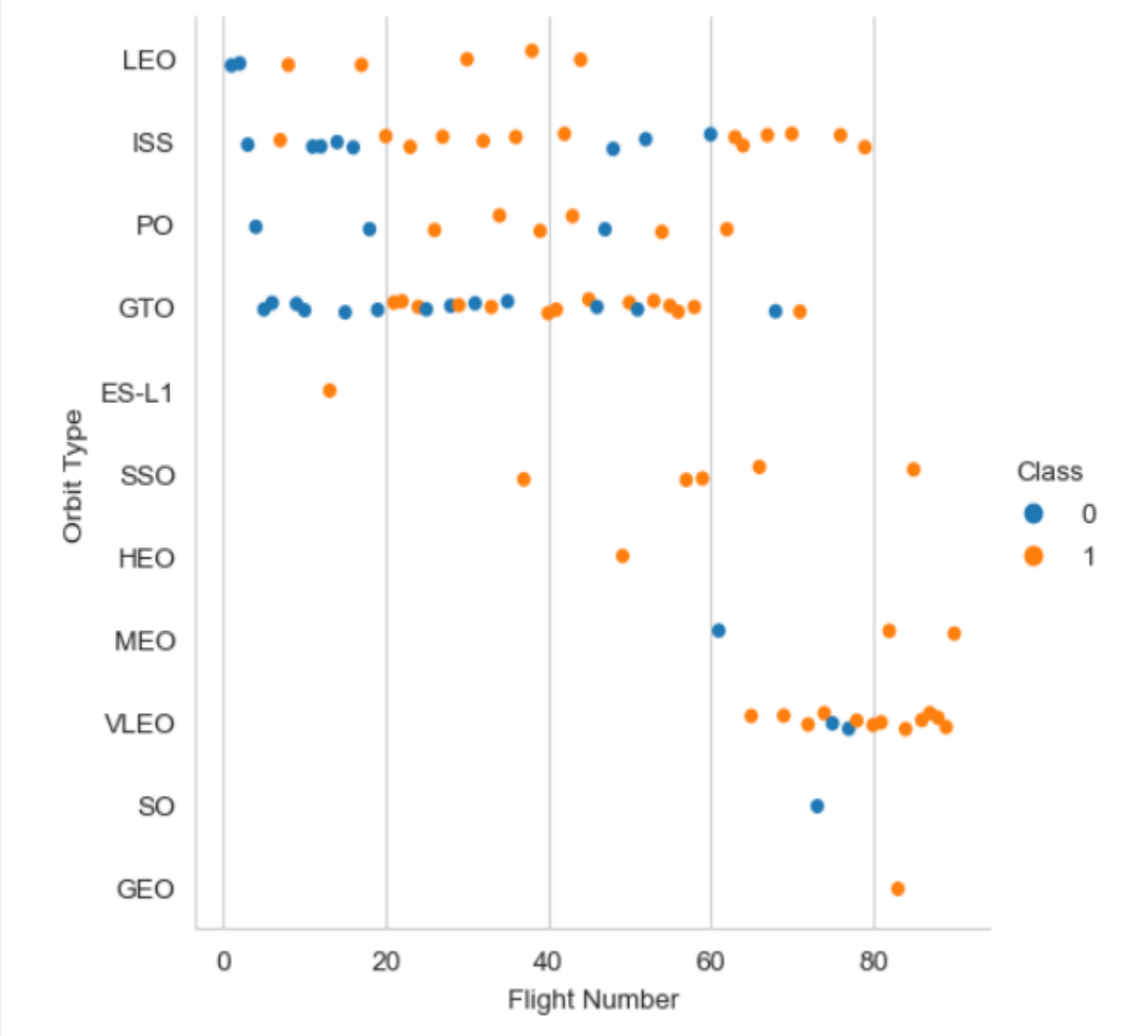
Success Rate vs. Orbit Type

We have 11 different types of Orbit. Based on this Bar plot,ES-L1,GEO,HEO and SSO are 100% successful. VLEO success rate is greater than 80%,ISS,LEO,MEO and PO success rate id between 60% and 80%,GTO success rate is between 40% and 60% and SO is 0% successful.



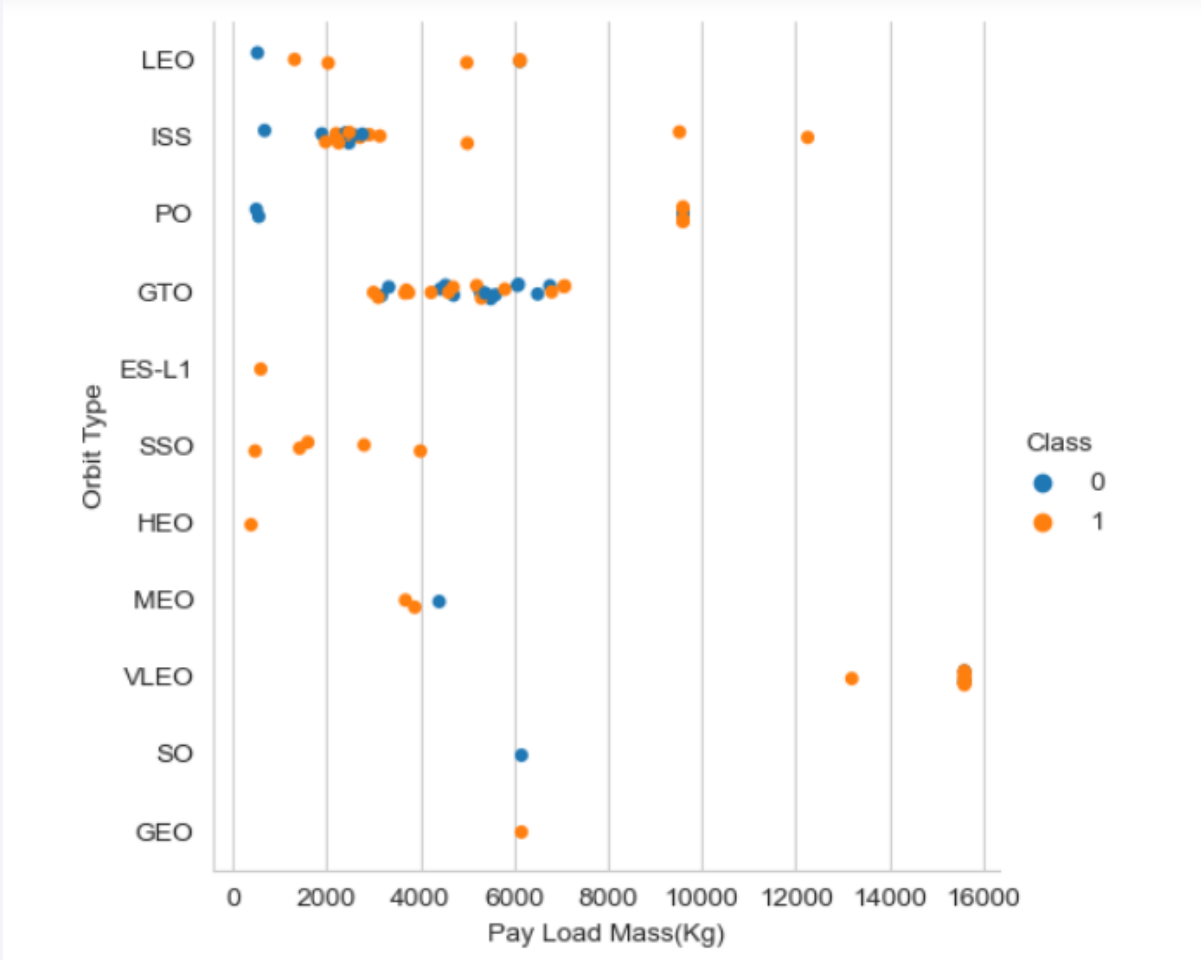
Flight Number vs. Orbit Type

In this Explanation We will use this Plot and Success Rate vs. Orbit Type plot, In GEO,HEO and ES-L1 we have only one flight and that is successful so rate of success is 100%.In orbit type SO we have only one flight and that is failed so success rate is 0%.for VLEO, LEO, ISS, MEO, PO, GTO and SSO we have more than 1 flights and they are arranged in order from the largest to the smallest rate: SSO,VLEO,LEO,PO,MEO,ISS,GTO.



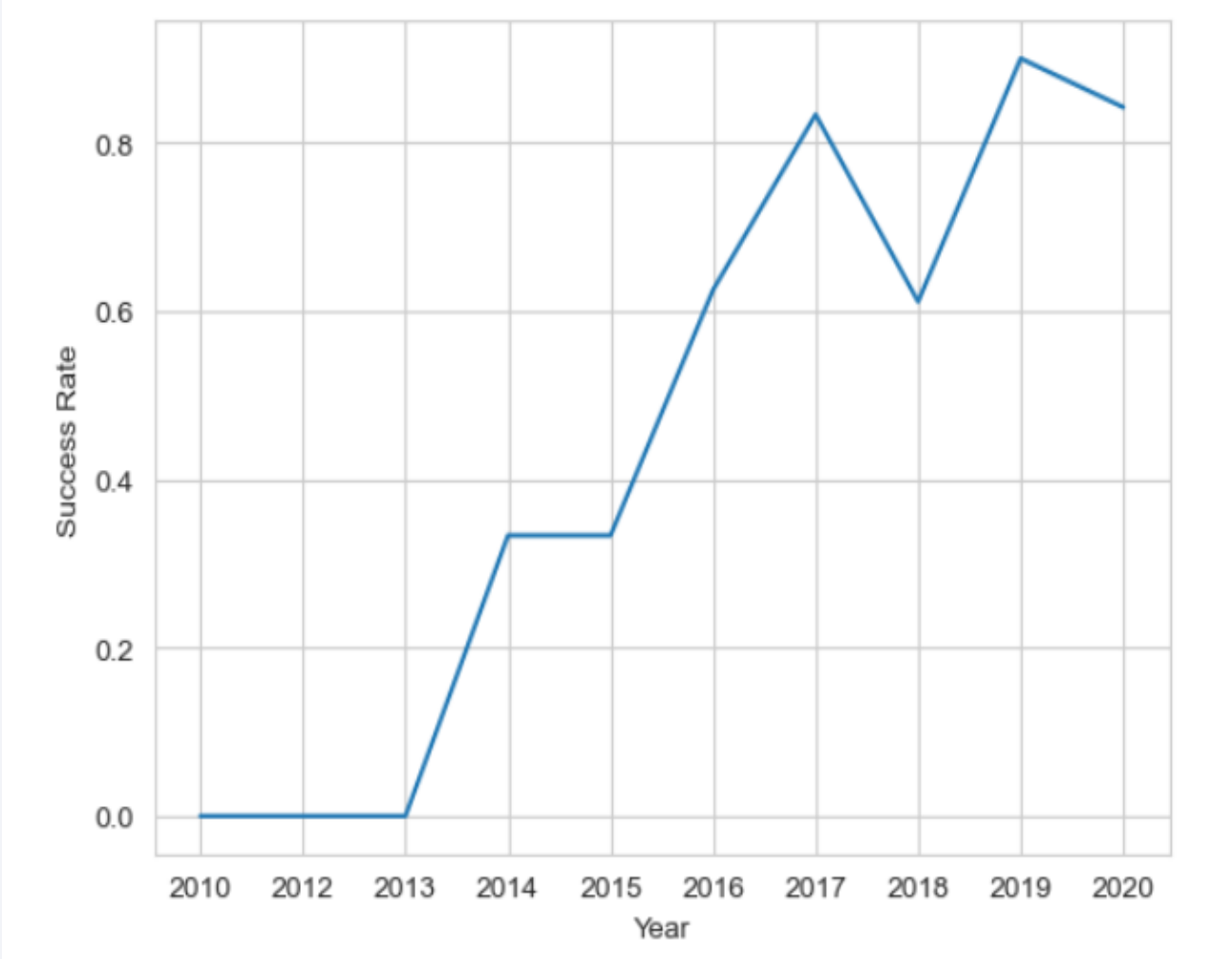
Payload vs. Orbit Type

In this plot, we can see success and failure of landing in different orbit types with different payload mass. in orbit type LEO, we have only flights with payload mass less than 6000 and most of them are successful and we don't have any idea about payloads greater than 6000.In ISS, density of payload is between 2000 and 4000 and for less than 2000 we have only on failed flight and for more than 6000 we have 3 successful flight. In orbit type PO , we have only two amount of payload mass, one is between 0 and 2000 and all are failed and another amount is almost 1000 and most of them are successful. For GTO payload mass is between 2000 and 8000 we don't have data for less than 2000 and more than 8000 for this orbit. For ES-L1 and HEO we have only one flight with payload mass less than 1000 and its successful. For SSO we have few flights that they have payload mass less than 4000 and all are successful. For MEO, we have few data when payload is almost 4000 and most of them are successful. For orbit VLEO, we have only few data that has payload mass greater than 12000 and all are successful. For Orbit SO and GEO , we have only one flight with payload mass almost 6000 and for SO is failed and GEO is successful.



Launch Success Yearly Trend

In this line plot, we can see the success rate of landing from 2010 to 2020. from 2010 to 2013 success rate was 0% . In 2013, that increased sharply. In 2014, that was steady. From 2015 to 2017,that raised aggressively then in 2017 that decreased harshly. In 2018,Success rate went up sharply and that was almost 95% then In 2019,that came down close to 85%.



All Launch Site Names

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

```
%%sql
SELECT DISTINCT(Launch_Site) from SPACEXTBL
```

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

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

Launch Site Names Begin with 'CCA'

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

```
%%sql
SELECT *
FROM SPACEXTBL
WHERE UPPER(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
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

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

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE upper(Customer) LIKE 'NASA%'
```

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

SUM(PAYLOAD_MASS__KG_)
99980

Average Payload Mass by F9 v1.1

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

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE Booster_Version LIKE 'F9 v1.1%'
```

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

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

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

Hint:Use min function

```
%%sql
SELECT min(Date)
FROM SPACEXTBL
WHERE upper(Mission_Outcome) LIKE 'SUCCESS'
```

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

min(Date)
01-03-2013

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

```
%%sql
SELECT Booster_Version
FROM SPACEXTBL
WHERE "Landing _Outcome" LIKE 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
```

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

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

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
%%sql
SELECT Distinct Mission_Outcome,Count(Mission_Outcome)
FROM SPACEXTBL
group by Mission_Outcome
```

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

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

Boosters Carried Maximum Payload

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

```
%%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

2015 Launch Records

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,Date,"Landing _Outcome",Booster_Version,Launch_Site
FROM SPACEXTBL
Where "Landing _Outcome"='Failure (drone ship)' AND substr(Date,7,4)='2015';
```

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

Month	Date	Landing _Outcome	Booster_Version	Launch_Site
01	10-01-2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	14-04-2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

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

```
%%sql
Select count(*) 'Number of success landing'
From SPACEXTBL
Where upper("Landing _Outcome") like 'S%' AND Date between '04-06-2010' AND '20-03-2017';
```

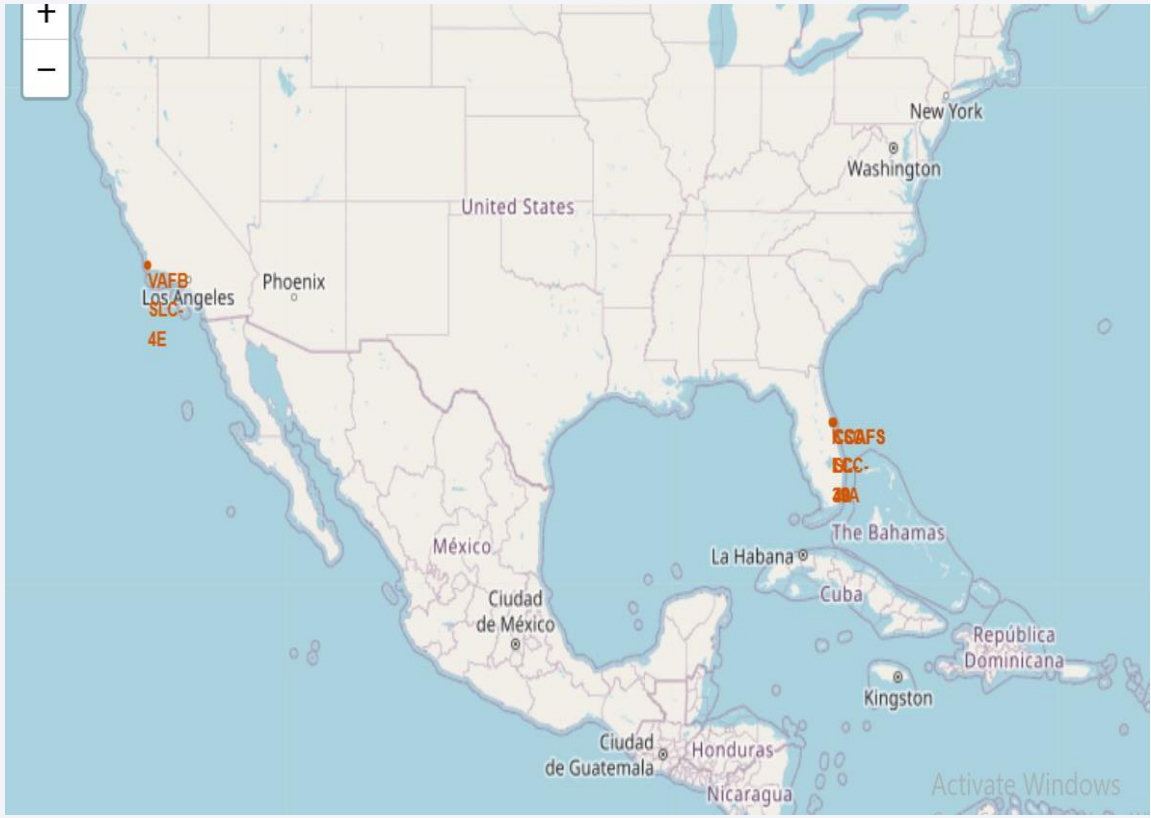
```
 * sqlite:///my_data1.db
Done.
```

Number of success landing
34


Section 3
Launch Sites
Proximities Analysis

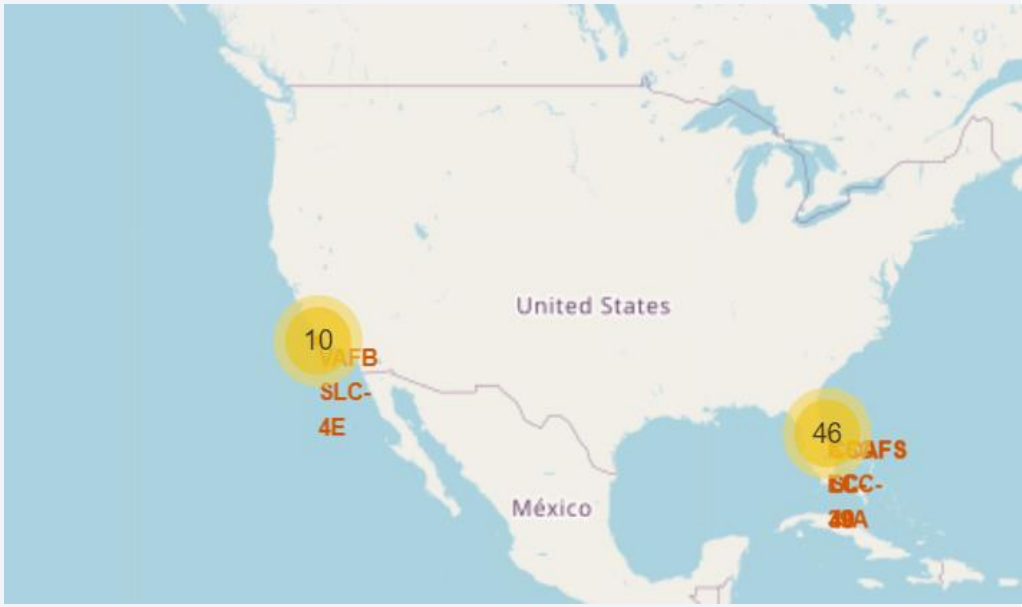
Launches Sites on Folium Map

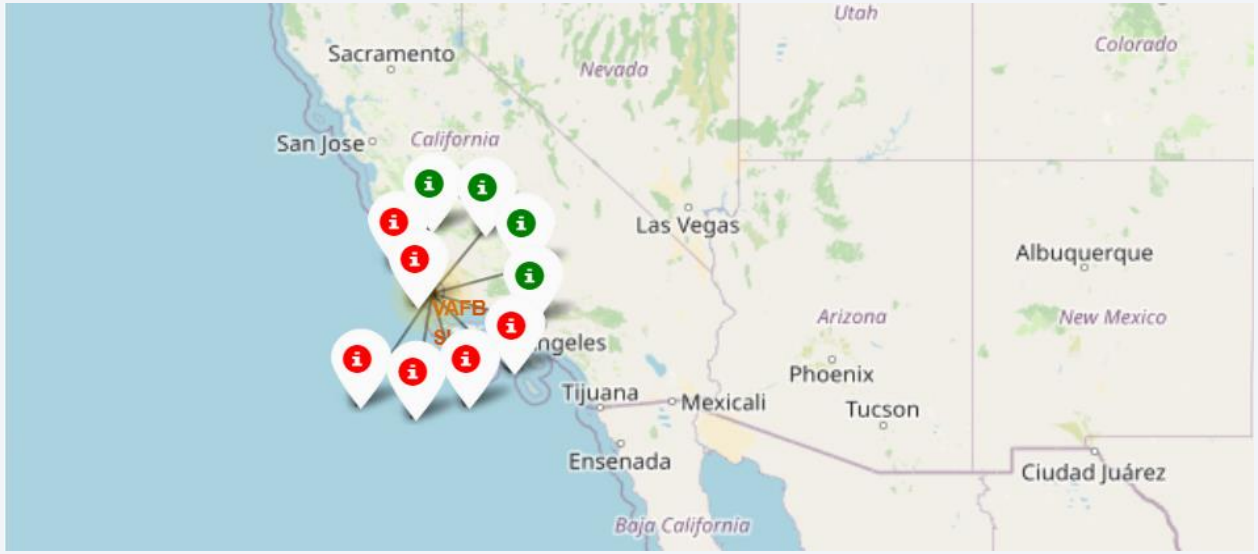
In our dataset, we have only 4 launch sites ,1 of them is in Los Angeles and the rest are located in Jacksonville.



Outcome of landing on Folium Map

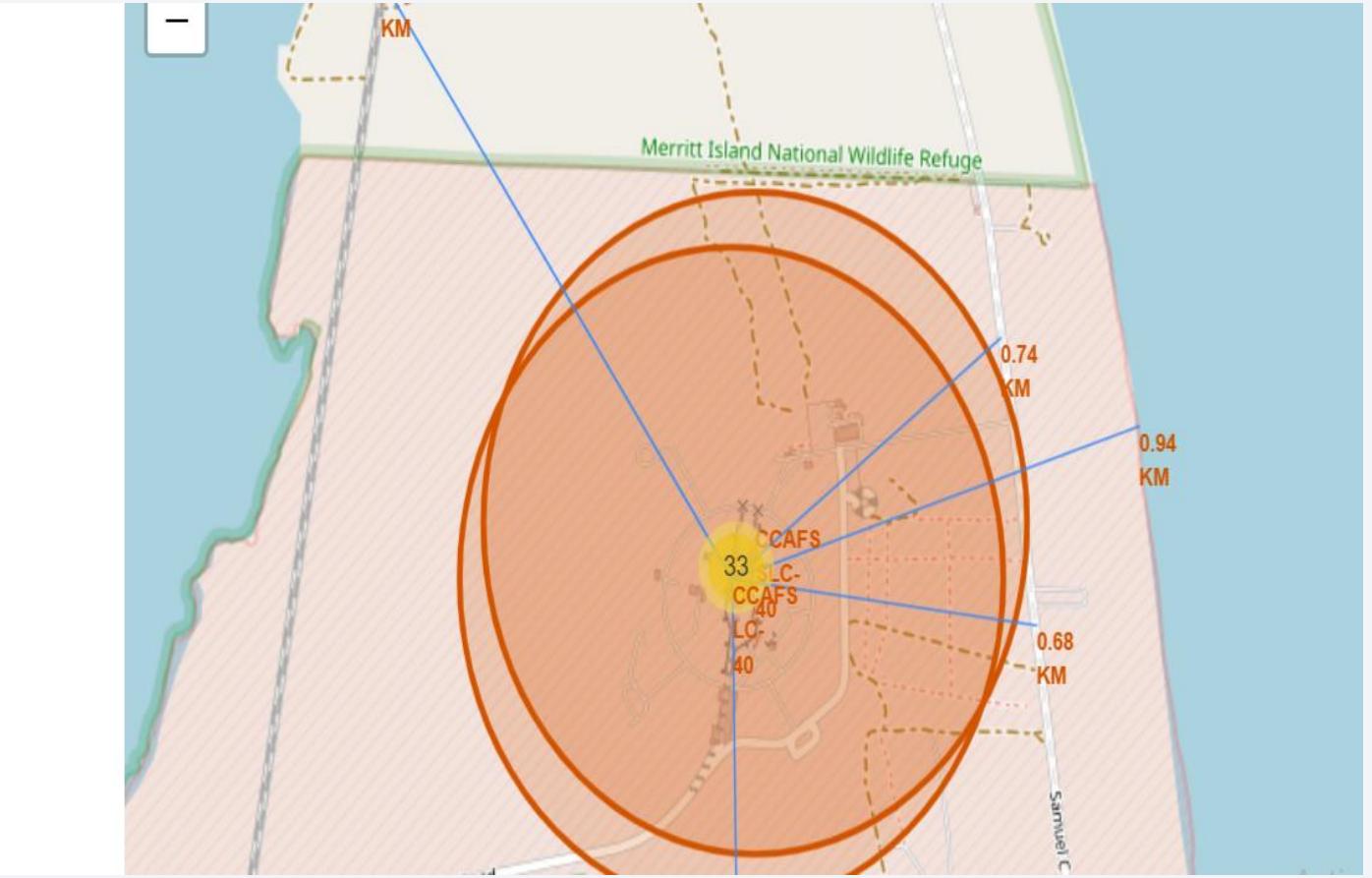
In first image, we can see we have 10 landing in launch site located in Los Angeles and 46 landing in the rest launch sites.4 out of 10 landing in Los Angeles were successful and 6 of them failed.

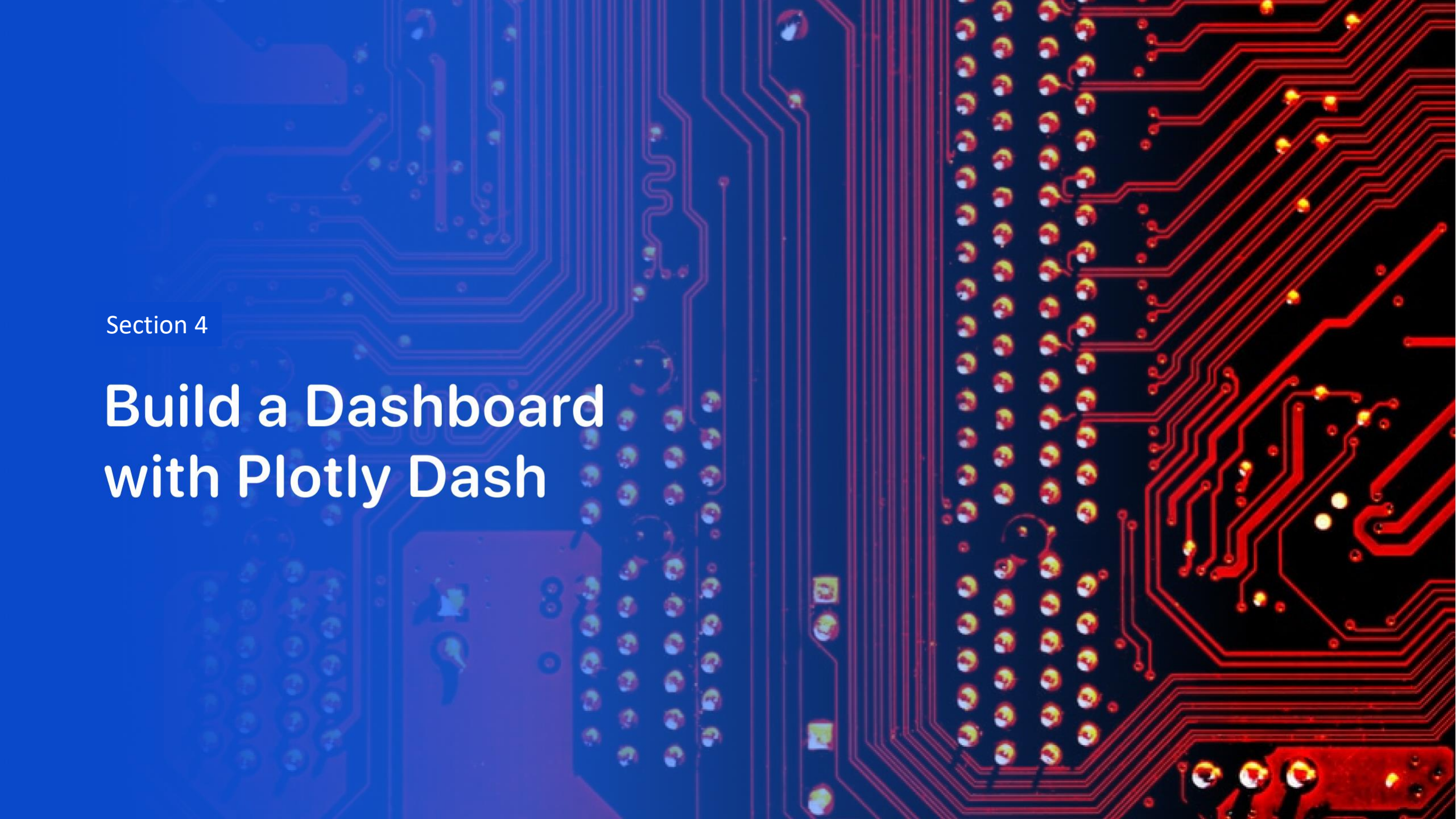




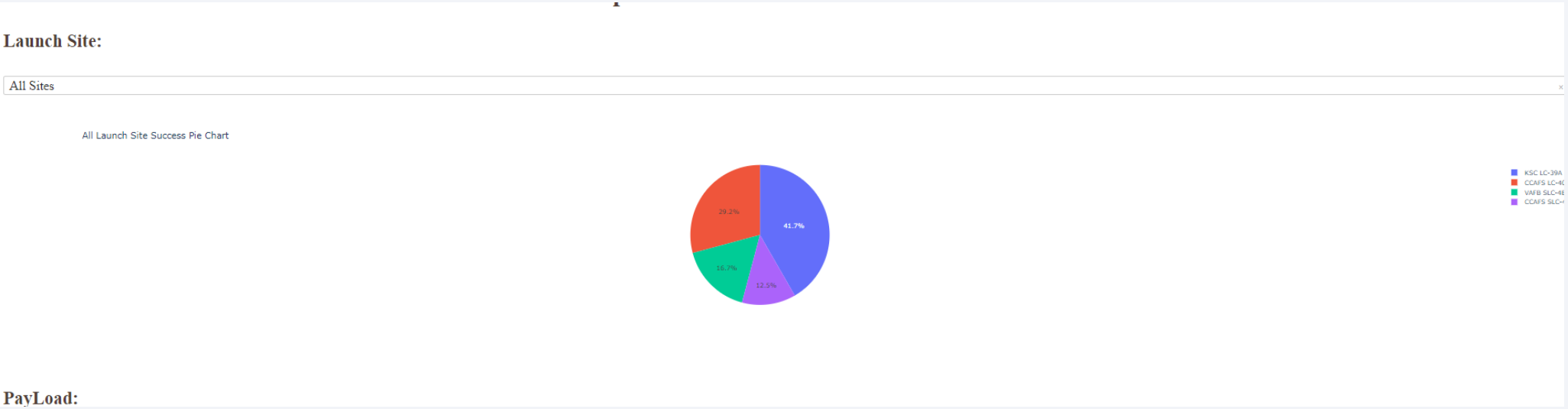
Distances from Launch site

- In this image, we can see distance between launch sites and some locations like highway, railroad, city and etc. to see are they appropriate to do landing.



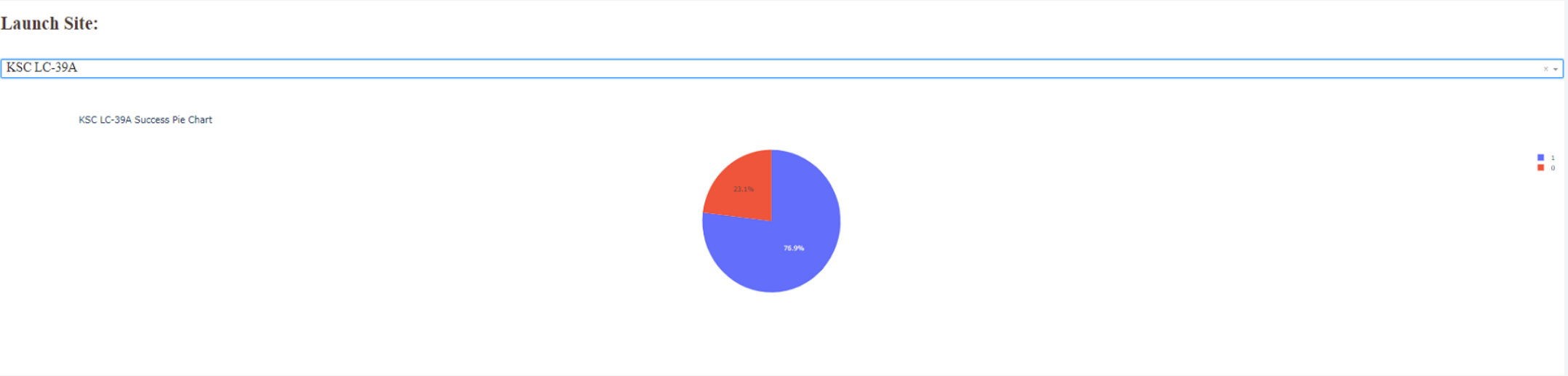

Section 4
Build a Dashboard with Plotly Dash

ALL Launch Site Success Pie Chart



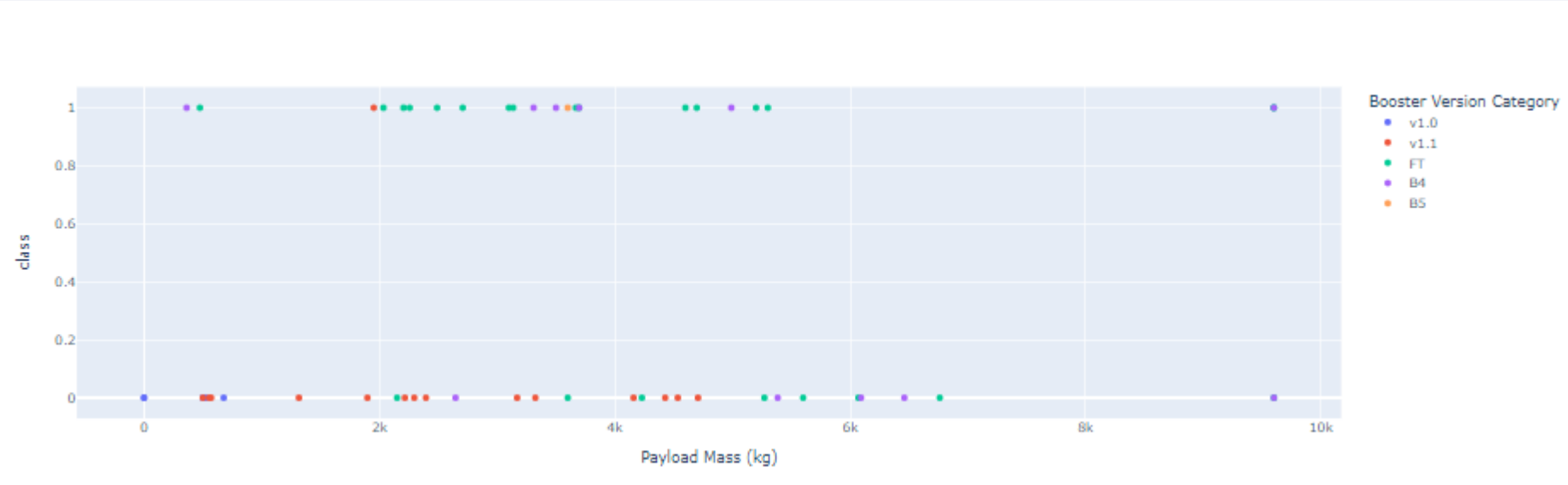
Based on this Pie Chart , we can see from all Launch Site Success 41.7% related to KSC LC-39A,29.2% is related to CCAFS LC-40,16.7% is related to VAFB SLC-4E and 12.5% is related CCSFS SLC-40

Highest Launch Success Ration Pie Chart



KSCLC-39A has the highest launch success Ration and from all of this launch site,76.9% was successful and 23.1% was failed.

Payload vs. Launch Site Scatter Plot



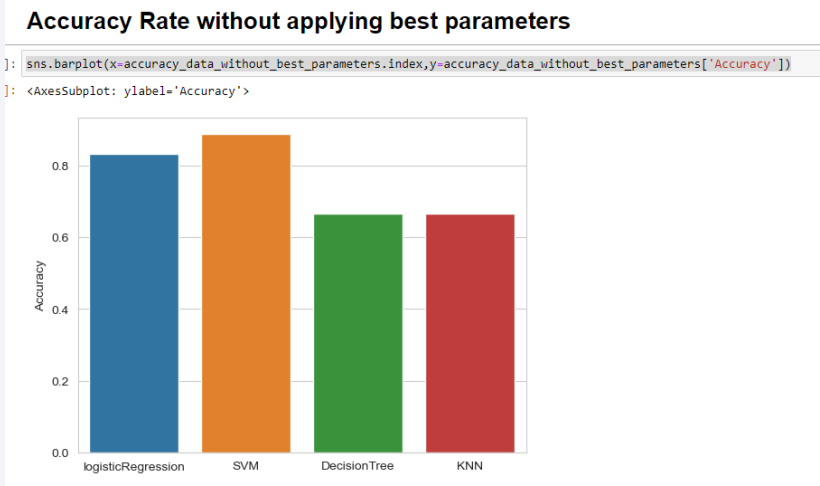
In this scatter plot, we can see in the payload mass from 0 to 10000,KSC LC-39A has the greatest success rate. KSC LC-39A has the greatest success rate in payload mass between 2000 and 6000.

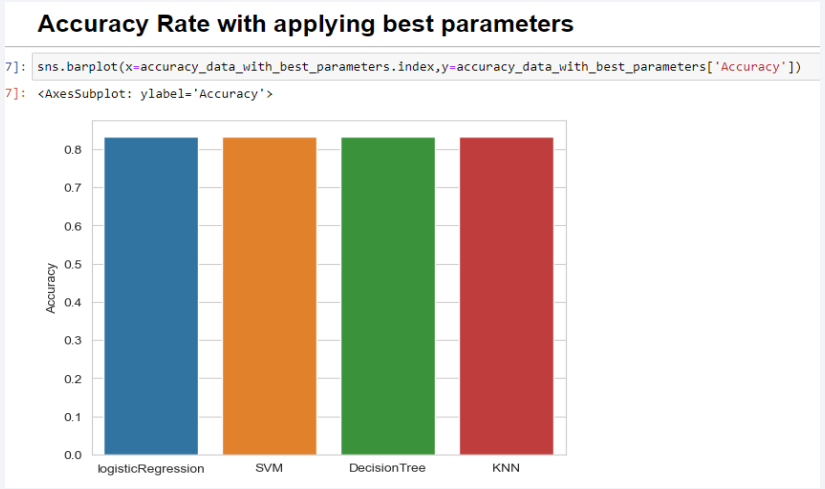
Section 5

Predictive Analysis (Classification)

Classification Accuracy

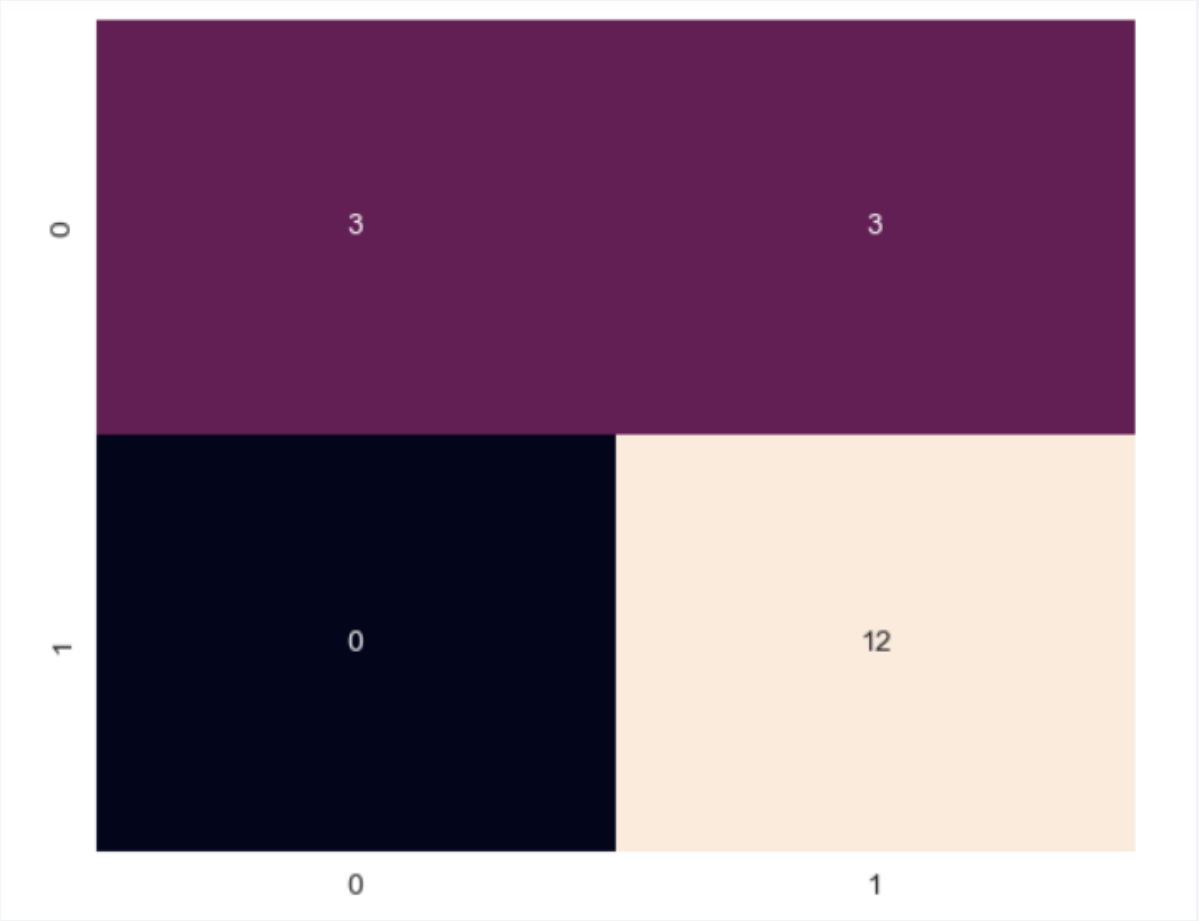
- First graph display accuracy in all 4 models without applying best parameters, we can see SVM is the best model. But after finding best parameters for our models we can see all models have same accuracy.





Confusion Matrix

based on this plot for confusion matrix, we have 18 data in our test set. 12 out of 18 are positive and our model can predict all positive data correctly. But 6 out of 18 are negative values and our model can predict only 3 of them correctly.



Conclusions

- Based on R-squared and confusion matrix of logistic regression, SVM, KNN and decision tree, we can see before applying the best parameters, SVM has the best R-squared.
- We improved our model, but we should try to reduce the number of flight that we predict them they will successful but they failed.
- To improve our model we can search about other factors that have effects on the out come of Falcon9 landing.

Thank you!