Data Science Capstone Project

John Giblin

10/03/2025

IBM Developer

SKILLS NETWORK

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion
- GitHub Link
- Appendix

EXECUTIVE SUMMARY

- Summary of methodologies
  - Data collection
  - Data wrangling
  - Exploratory Data Analysis with Data Visualization
  - Exploratory Data Analysis with SQL
  - Building an interactive map with Folium
  - Building a Dashboard with Plotly Dash
  - Feature Engineering for predictive Analysis
  - Predictive analysis (Classification) Sub Point 3
- Summary of results
  - Exploratory Data Analysis results
  - Interactive analytics demo (charts, maps and dashboards)
  - Predictive analysis results

INTRODUCTION

Project Background/Context

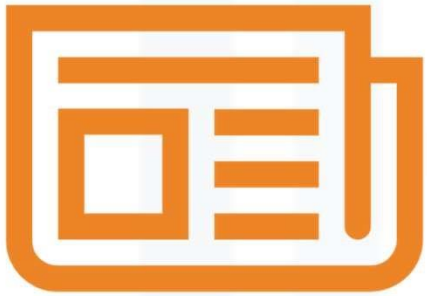
SpaceX has gained worldwide attention for a series of historic milestones. It is the only private company ever to return a spacecraft from low-earth orbit, which it first accomplished in December 2010.

Space X 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 Space X 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 space X for a rocket launch . Based on historical data of space X launches, we are going to predict if SpaceX will reuse the first stage, in other words if the first stage will land successfully. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch.

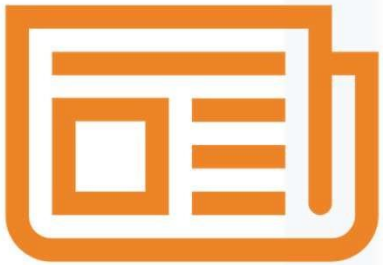
Questions to be answered

- Are there factors that affect the success of the first stage landing?
- Does the rate of successful landings increase over the years?
- What is the best algorithm that can be used to predict if first stage will land?

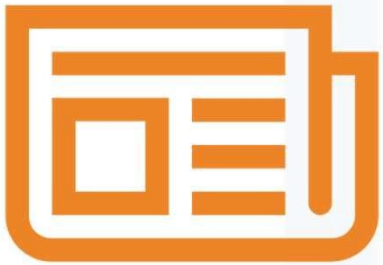
METHODOLOGY



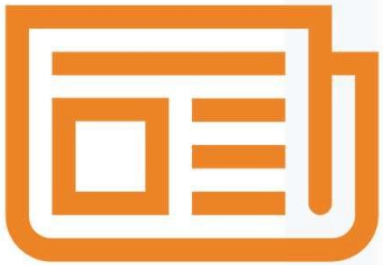
- Data collection
  - Making a get request to collect data from SpaceX Rest API
  - Filtering the data to only include Falcon 9 launches
  - Web Scraping Falcon 9 historical Launches Records from Wikipedia using Beautiful soup
- Data wrangling
  - Dealing with missing values
  - Data Analysis on the data columns (variables)
  - Filtering the data column named 'Outcome' to create an extra column named 'Class' for our Classification Model.
    - If the value of the column 'Class' is zero, the first stage did not land successfully;
    - one means the first stage landed Successfully



- Exploratory Data Analysis with Data Visualization
  - Relationship Between PayloadMass And LaunchSite
  - Relationship between FlightNumber and LaunchSite
  - Relationship between Success Rate(class) And Orbit Type
  - Relationship between FlightNumber And Orbit Type
  - Relationship Between PayloadMass And Orbit Type
  - Visualizing the Launch success yearly trend
- Exploratory Data Analysis with SQL
  - SpaceX Launch Sites Names
  - SpaceX Launch Site: CCAFS SLC-40
  - Nasa (CRS) Total Payload Mass Carried by Boosters
  - Average Payload Mass Carried by Booster Version F9 V1.1
  - First Successful Landing Outcome in Ground Pad (Date)
  - Total Number of Successful and Failure Mission Outcomes
  - Boosters with Success in Drone Ship & Payload Mass Greater Than 4000 But Less Than 6000
  - BoosterVersion Which Have Carried the Maximum Payload Mass
  - Failed Landing Outcomes in Drone Ship, Booster Versions for LaunchSite For the Months in Year 2015.
  - Landing Outcomes Between 2010-06-04 And 2017-03-20 In Descending Order.



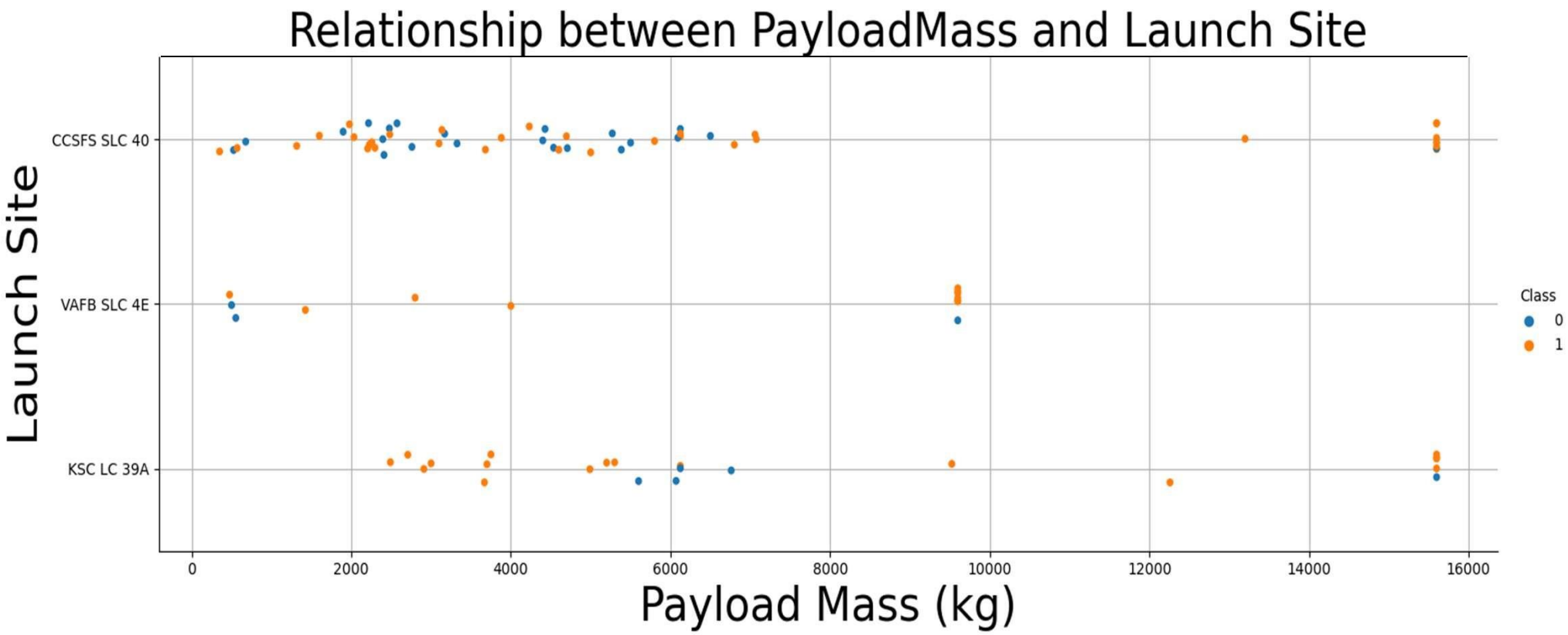
- Interactive visual analytics using Folium and Plotly-Dash
  - Interactive Map with Folium
    - Marking All Launch Sites Location on A Map
    - Marking the Success/Failed Launches for Each Site on The Map
    - Marking the Distances Between A Launch Site to Its Proximities
  - Building A Dashboard with Plotly Dash
    - Launch Success Rate(Average) For All Sites
    - Launch Success/Failure Count for Each Site
    - Payload Mass Vs Launch Success/Failure Count for Each Sites Varying Booster Version Categories
    - Payload Mass Vs Launch Success/Failure Count for VAFB SLC-4E
    - Payload Mass and Launch Success for Site CCAFS SLC-40
    - Payload Mass Vs Launch Success/Failure Count for KSC LC-39A



- Feature Engineering for predictive Modelling
  - Feature selection for a successful prediction
  - Creating dummy variables from the selected categorical columns
  - Casting dummy variables into numerical Variables
- Predictive analysis using classification models
  - Building, tuning and evaluation of classification models to ensure the best results
    - Accuracy score for the selected classification models
      - Accuracy score for the test set
      - Accuracy score for the entire data set
      - Confusion matrix: test dataset
      - Confusion matrix: entire dataset

EDA WITH VISUALIZATION RESULTS

RELATIONSHIP BETWEEN PAYLOADMASS AND LAUNCHSITES



DID NOT LAND SUCCESSFULLY : 0
LANDED SUCCESSFULLY : 1

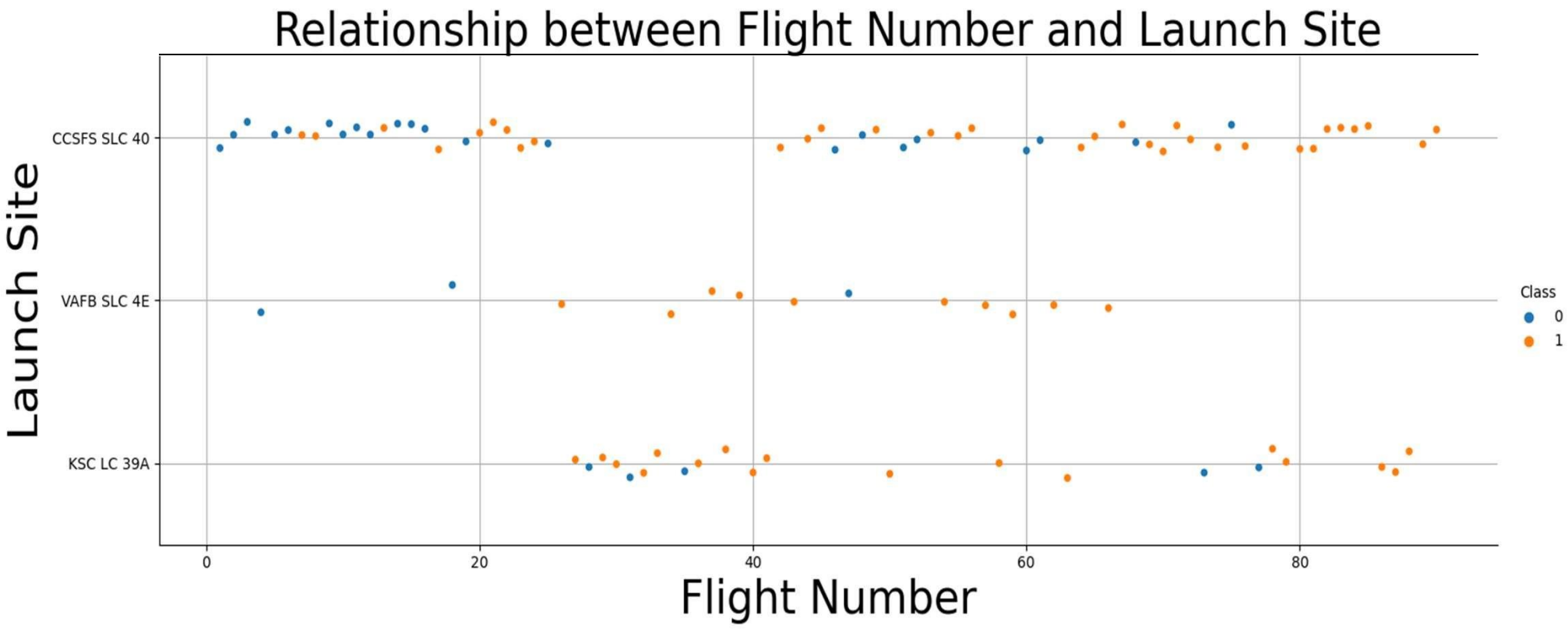
Fig. 1

RELATIONSHIP BETWEEN PAYLOADMASS AND LAUNCHSITES

Insight:

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

RELATIONSHIP BETWEEN FLIGHT NUMBER AND LAUNCHSITES



DID NOT LAND SUCCESSFULLY : 0
LANDED SUCCESSFULLY : 1

Fig. 2

RELATIONSHIP BETWEEN FLIGHT NUMBER AND LAUNCHSITES

Insight:

- FlightNumber indicates the continuous launch attempts.
- Majority of the earliest flights (lower flight number) failed while most of the latest flights(higher flight number) landed successfully (first stage landing)
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC has the lowest launch.
- It can be assumed that each new launch(latest) has a higher rate of success.
- KLC SC39A has no launching for flight number below 23 while VAFB SLC has no launching for flight number starting from 65 upward.

RELATIONSHIP BETWEEN SUCCESS RATE(CLASS) AND ORBIT TYPE

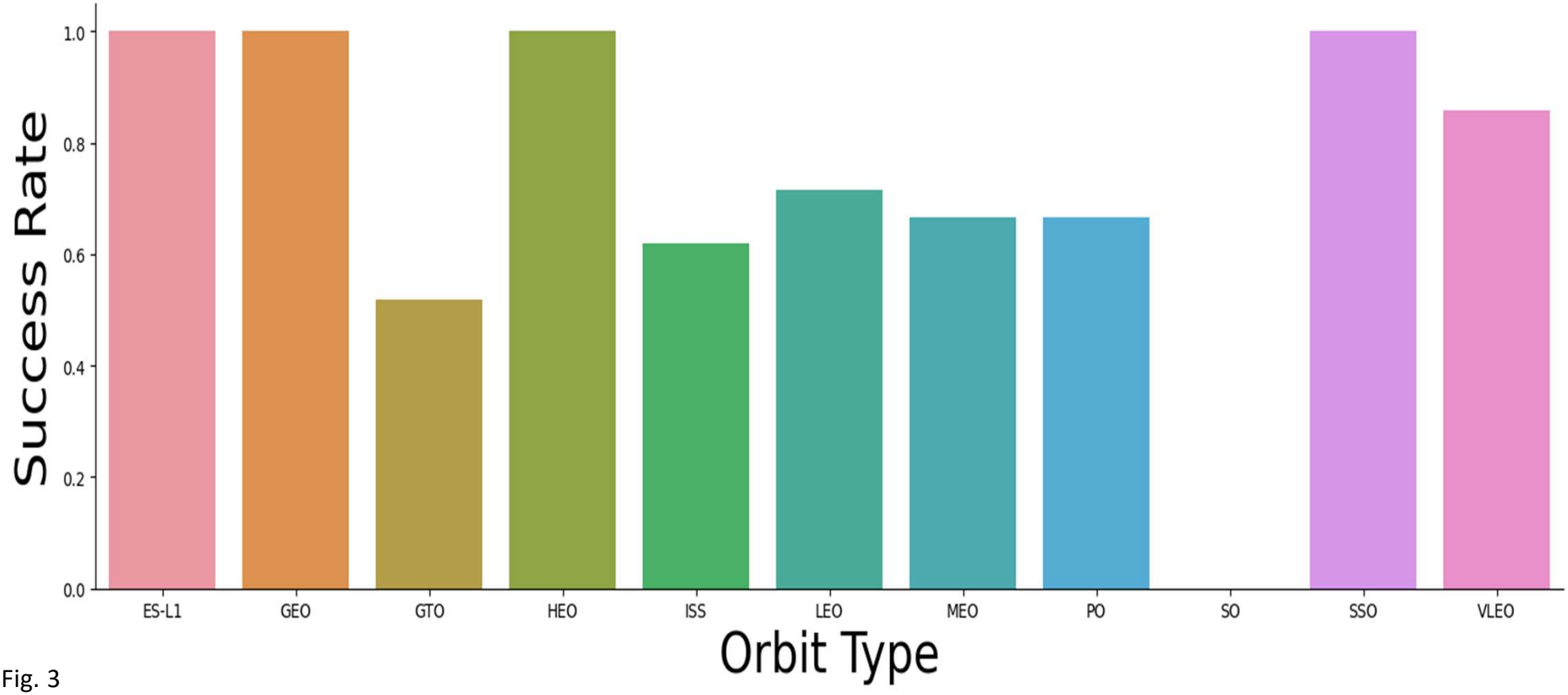


Fig. 3

RELATIONSHIP BETWEEN SUCCESS RATE(CLASS) AND ORBIT TYPE

Insight:

Which orbits have high success rate in the above bar chart?

- **ORBITS WITH 100% SUCCESS RATE ARE:**

  ES-L1, GEO, HEO, SSO

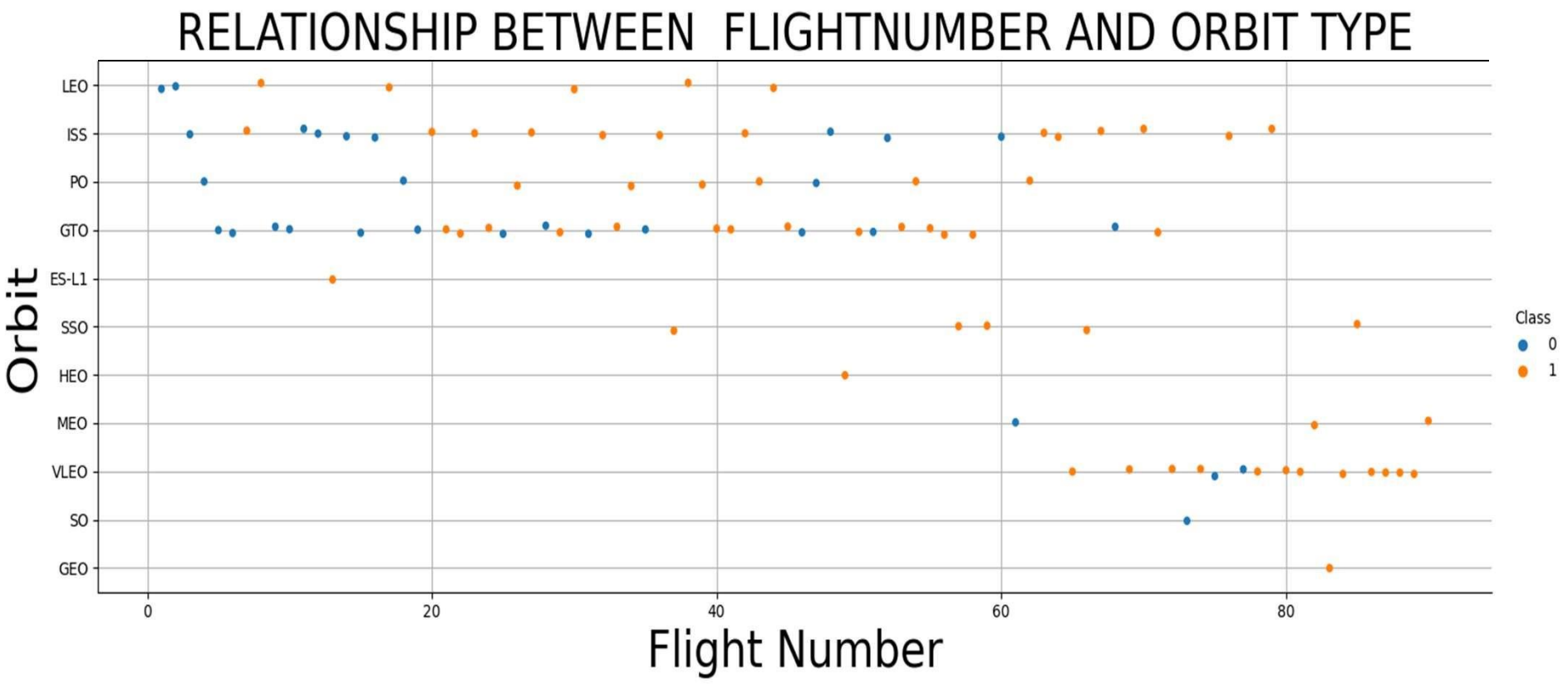
- **ORBITS WITH SUCCESS RATE BETWEEN 50% AND 85%:**

  GTO, ISS, VLEO, MEO, PO

- **ORBITS WITH 0% SUCCESS RATE IS:**

  SO

RELATIONSHIP BETWEEN FLIGHTNUMBER AND ORBIT TYPE



DID NOT LAND SUCCESSFULLY : 0
LANDED SUCCESSFULLY : 1

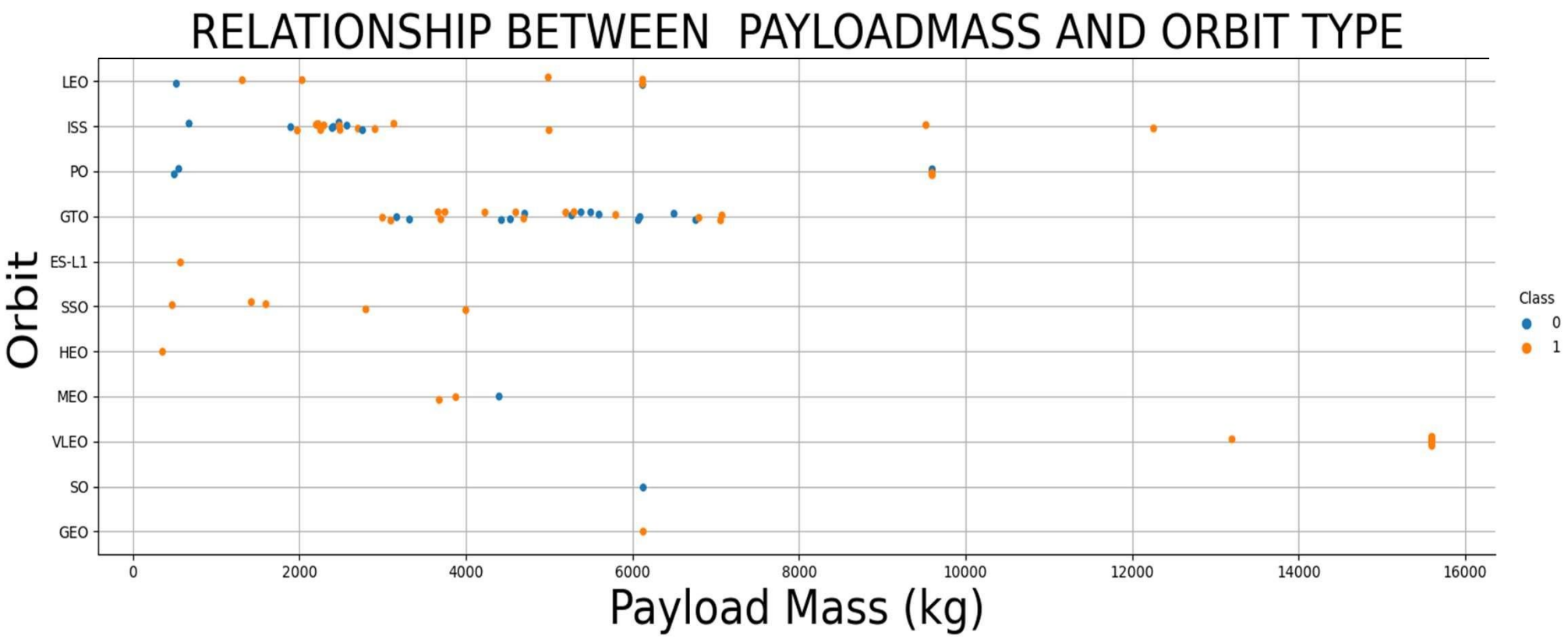
Fig. 4

RELATIONSHIP BETWEEN FLIGHTNUMBER AND ORBIT TYPE

Insight:

- 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 and success when in
- GEO orbit.
- Also, SSO had no unsuccessful landing while there were no launching for HEO,MEO, VLEO, SO and GEO with FlightNumber lower than 40.

RELATIONSHIP BETWEEN PAYLOADMASS AND ORBIT TYPE



DID NOT LAND SUCCESSFULLY : 0
LANDED SUCCESSFULLY : 1

Fig. 5

RELATIONSHIP BETWEEN PAYLOADMASS AND ORBIT TYPE

Insight:

- With heavy payloads the successful landing or positive landing rate are more for POLAR (PO), 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.

VISUALIZING THE LAUNCH SUCCESS YEARLY TREND

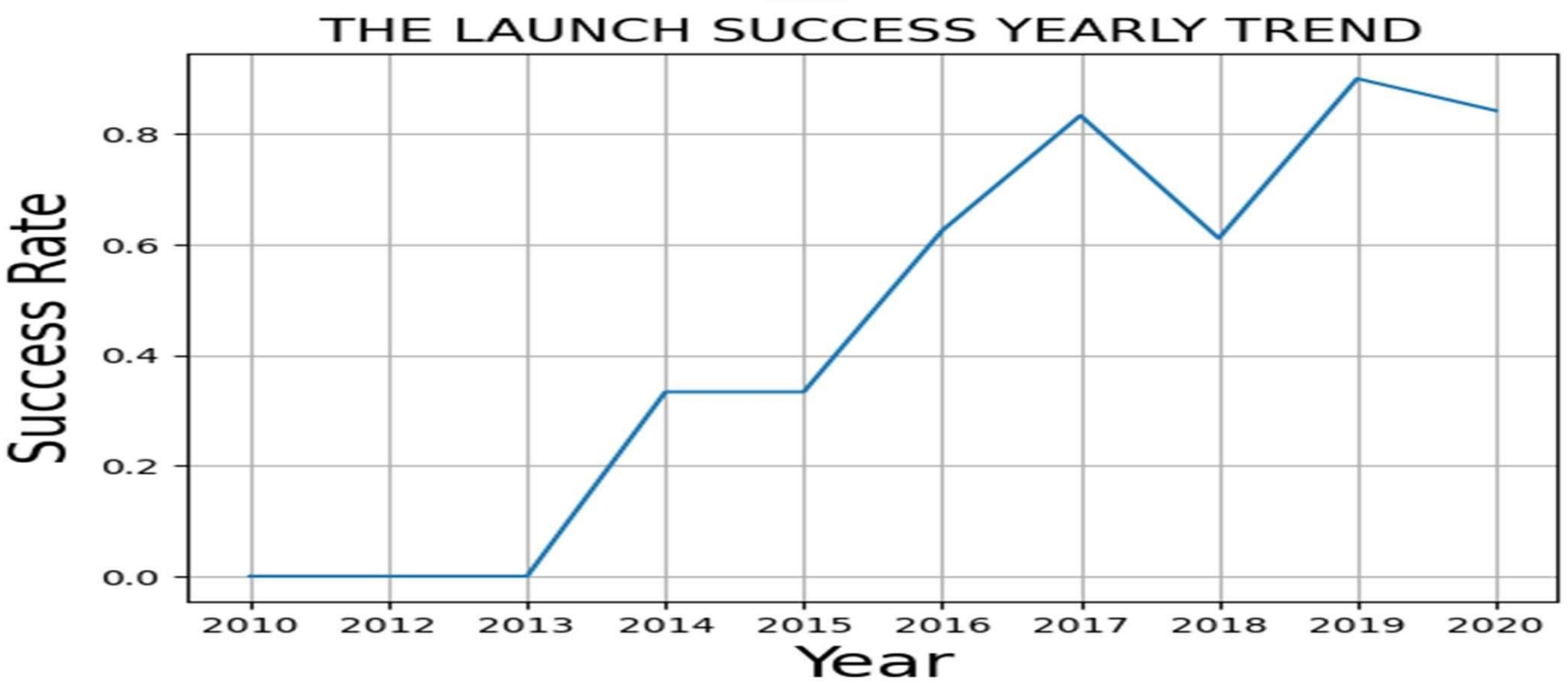


Fig. 6

VISUALIZING THE LAUNCH SUCCESS YEARLY TREND

Insight:

- You can observe that the success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing.

.

EXPLORATORY DATA ANALYSIS WITH SQL

SPACEX LAUNCH SITES NAMES

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

```
1 %sql select distinct launch_site from SPACEXTBL;
```

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

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

- CCAFS SLC-40: Cape Canaveral Space Launch Complex 40 (SLC-40)
- VAFB SLC-4E: Vanguard Space Base Space Launch Complex 4E (SLC-4E)
- KSCL LC-39A: Kennedy Space Center Launch Complex 39A (LC-39A)

Fig. 7

SPACEX LAUNCH SITE: CCAFS SLC-40

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

```
1 %sql select * from SPACEXTBL 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 SLC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS SLC-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 SLC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS SLC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS SLC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Fig. 8

NASA (CRS) TOTAL PAYLOAD MASS CARRIED BY BOOSTERS

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

```
1 %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXTBL where customer = 'NASA (CRS)';
```

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

total_payload_mass
45596

Fig. 9

AVERAGE PAYLOAD MASS CARRIED BY BOOSTER VERSION F9 V1.1

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

```
1 %sql select round(avg(payload_mass__kg_),2) as Avg_payload_mass from SPACEXTBL where booster_version like '%F9 v1.1%';
```

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

Avg_payload_mass
2534.67

Fig. 10

FIRST SUCCESFUL LANDING OUTCOME IN GROUND PAD (DATE)

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

Hint:Use min function

```
]:  1 %sql select min(Date) as first_successful_landing from SPACEXTBL where Landing_Outcome = 'Success (ground pad)';
 * sqlite:///my_data1.db
Done.
```

]:

first_successful_landing
2015-12-22

Fig. 11

TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSION OUTCOMES

List the total number of successful and failure mission outcomes

```
1 %sql select Mission_Outcome, count(*) as Total_number from SPACEXTBL 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

- Successful Mission Outcomes: **99**
- Failure in Flight: **1**
- Successful with payload status unclear: **1**

Fig. 12

BOOSTERS WITH SUCCESS IN DRONE SHIP & PAYLOAD MASS GREATER THAN 4000 BUT LESS THAN 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 = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000
```

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

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

Fig. 13

BOOSTER_VERSIONS WHICH HAVE CARRIED THE MAXIMUM PAYLOAD MASS

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

```
1 %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

Fig. 14

FAILED LANDING_OUTCOMES IN DRONE SHIP ,BOOSTER VERSIONS FOR LAUNCH_SITES FOR THE MONTHS IN YEAR 2015.

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,Date, Booster_Version, launch_site, Landing_Outcome from SPACEXTBL
where Landing_Outcome = 'Failure (drone ship)' and substr(Date,0,5)='2015';
```

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

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

Fig. 15

LANDING OUTCOMES BETWEEN *2010-06-04* AND *2017-03-20* IN DESCENDING ORDER.

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(*) as count_outcomes from SPACEXTBL
where Date between '2010-06-04' and '2017-03-20'
group by Landing_Outcome
order by count_outcomes desc;
```

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

Landing_Outcome	count_outcomes
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

Fig. 16

INTERACTIVE MAP WITH FOLIUM

NOTE:

- The launch success rate may also depend on the location and proximities of a launch site, i.e., the initial position of rocket trajectories.
- Finding an optimal location for building a launch site certainly involves many factors and hopefully we could discover some of the factors by analyzing the existing launch site locations.

MARKING ALL LAUNCH SITES LOCATION ON A MAP

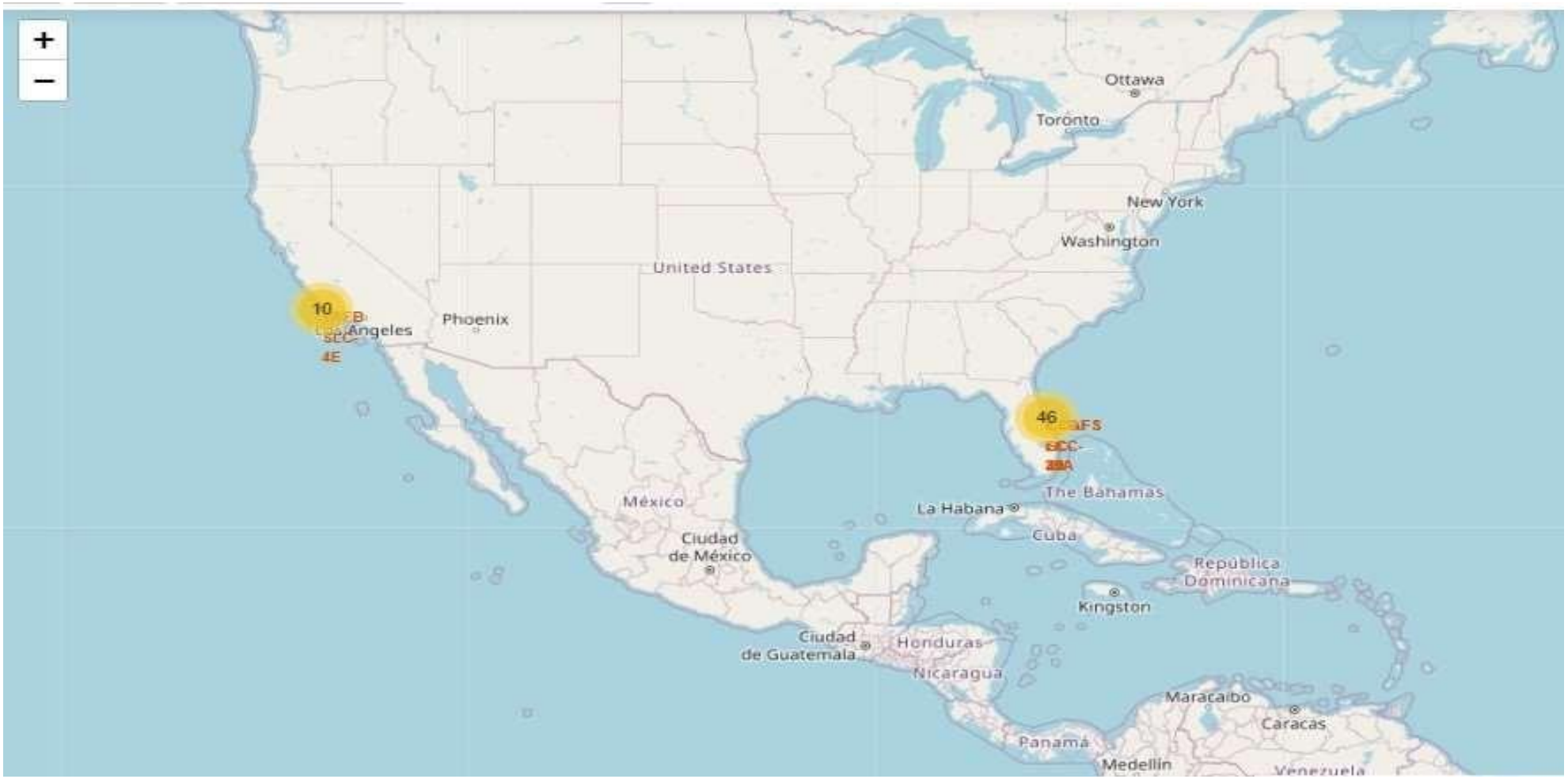
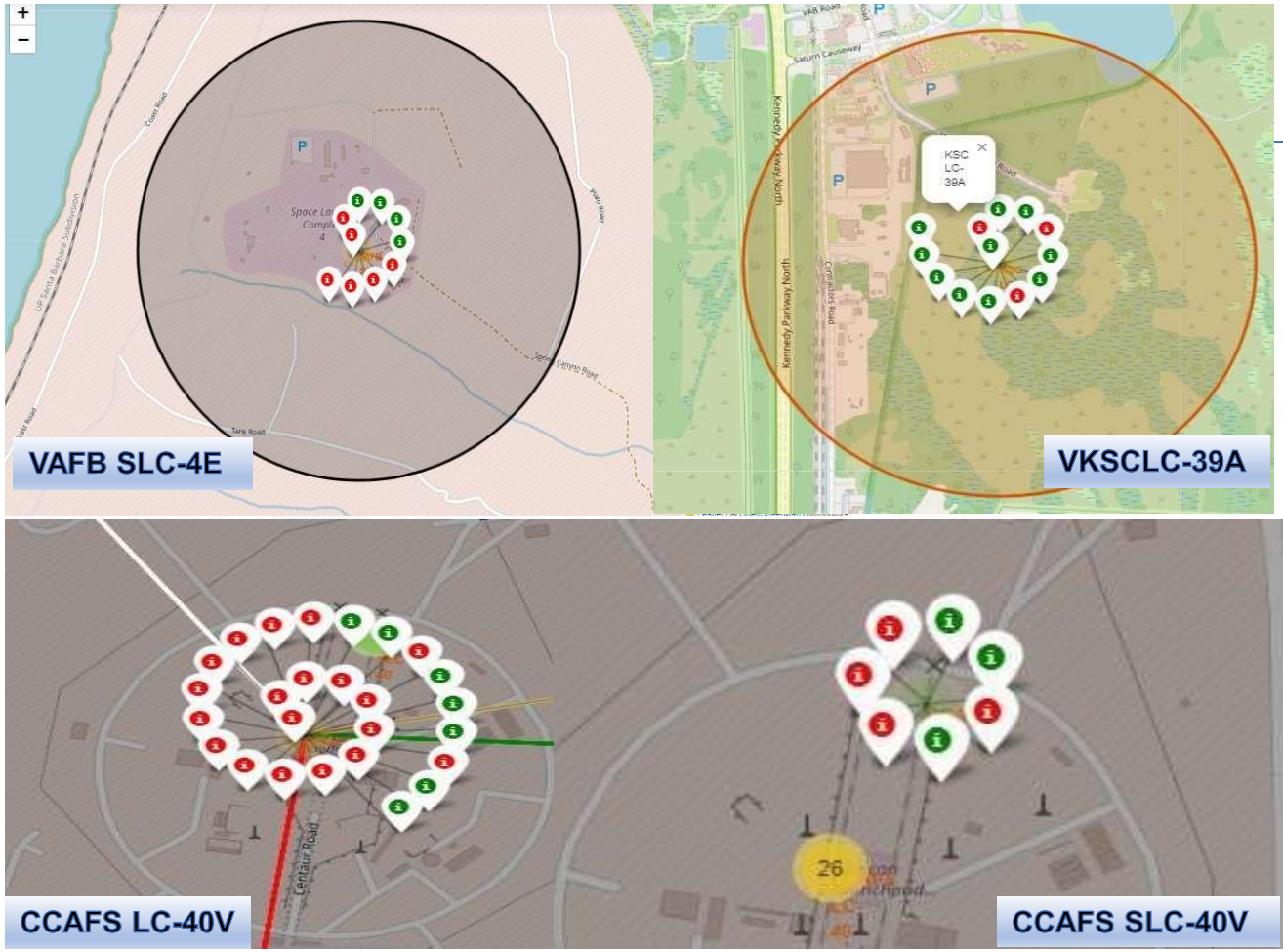


Fig. 17

MARKING THE SUCCESS/FAILED LAUNCHES FOR EACH SITE ON THE MAP



- Green Marker = Successful Launch
- Red Marker = Failed Launch

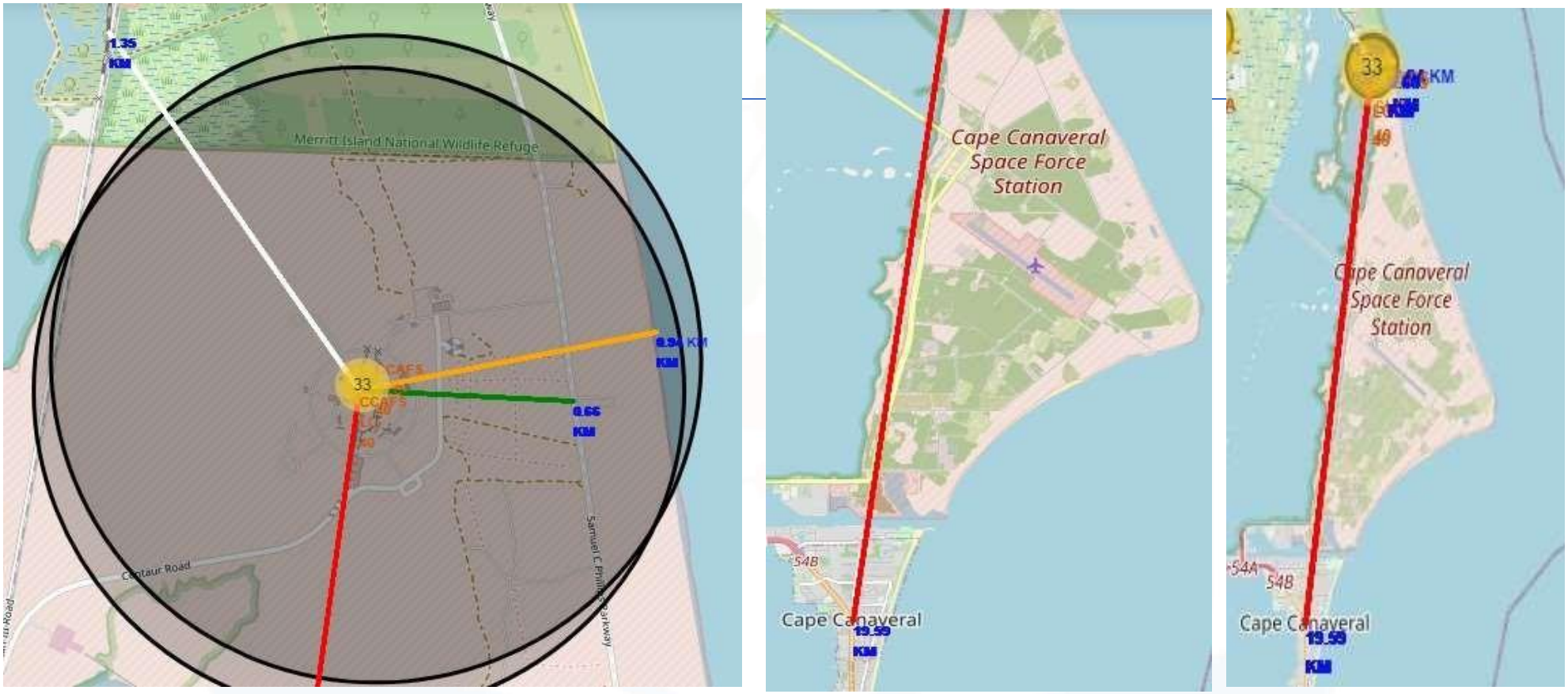
Fig. 18

MARKING THE SUCCESS/FAILED LAUNCHES FOR EACH SITE ON THE MAP

Insight:

- From the color-labeled markers (Red/Green) in marker clusters above, you should be able to easily identify which launch sites have relatively high success rates and lowest.

MARKING THE DISTANCES BETWEEN A LAUNCH SITE TO ITS PROXIMITIES



Railway:	1.35km
Coast line:	0.94 km
Highway:	0.66 km
City:	19.59 km

Fig. 19

MARKING THE DISTANCES BETWEEN A LAUNCH SITE TO ITS PROXIMITIES

Insight:

- The Launch sites (VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40) considered in this project are in proximity to the Equator line. Launch sites are made at the closest point possible to Equator line, because anything on the surface of the Earth at the equator is already moving at the maximum speed (1670 kilometers per hour). For example launching from the equator makes the spacecraft move almost 500 km/hour faster once it is launched compared half way to north pole.
- All launch sites considered in this project are in very close proximity to the coast.
- While starting rockets towards the ocean, we minimize the risk of having any debris dropping or exploding near people.
- From the visual analysis of the launch site **CCAFS SLC-40,** we can clearly see that it is:
  - relatively close to railway (**1.35 km**)
  - relatively close to highway (**0.66 km**)
  - relatively close to coast line (**0.94 km**)
- Also the launch site **CCAFS SLC-40** keeps certain distance away from its nearest city **Cape Canaveral** (**19.59 km**). This is so because failed rocket with its high speed can cover distances like **15-20 km** in few seconds. It could be potentially dangerous to populated areas.

Building a Dashboard with Plotly Dash

IBM Developer

SKILLS NETWORK

LAUNCH SUCCESS RATE(AVERAGE) FOR ALL SITES

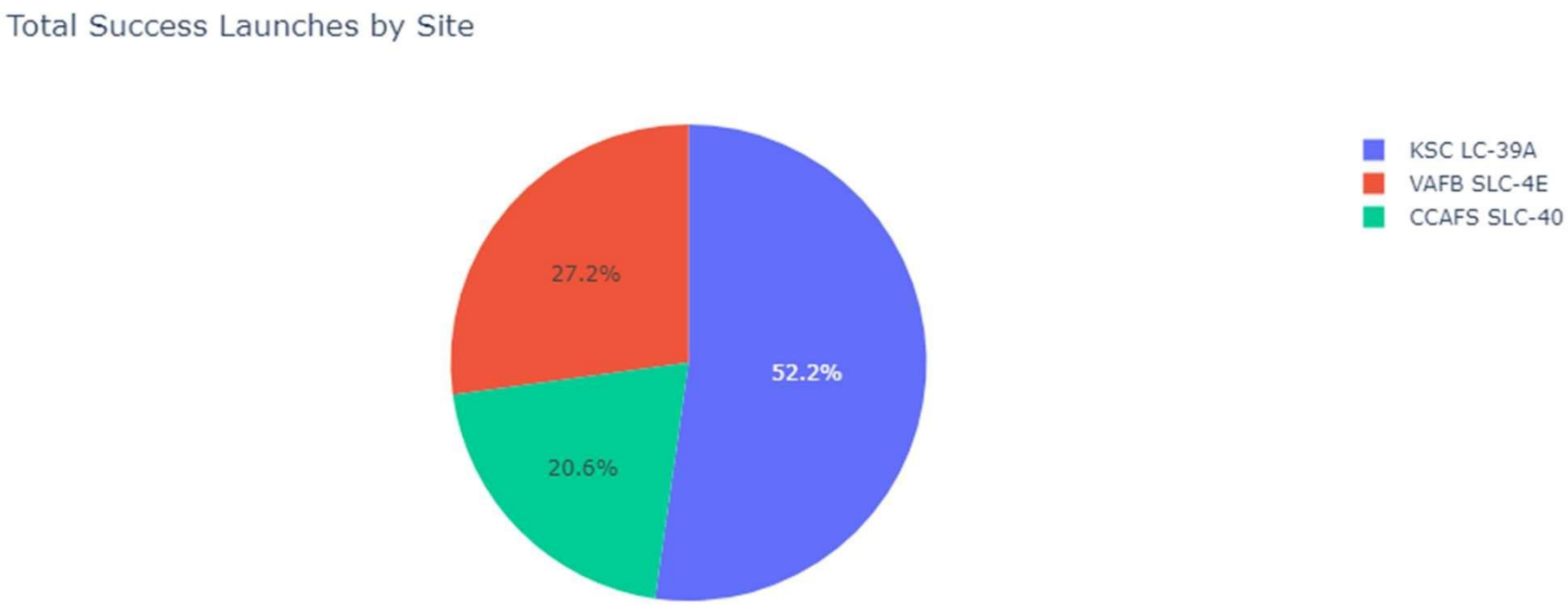


Fig. 20

LAUNCH SUCCESS RATE(AVERAGE) FOR ALL SITES

Insight:

The chart clearly shows that from all the launch sites, **KSC LC-39A** at **52.2%** has the overall average for successful launches while **CCAFS LC-40** has the least at **20.6%**.

LAUNCH SUCCESS/FAILURE COUNT FOR EACH SITES

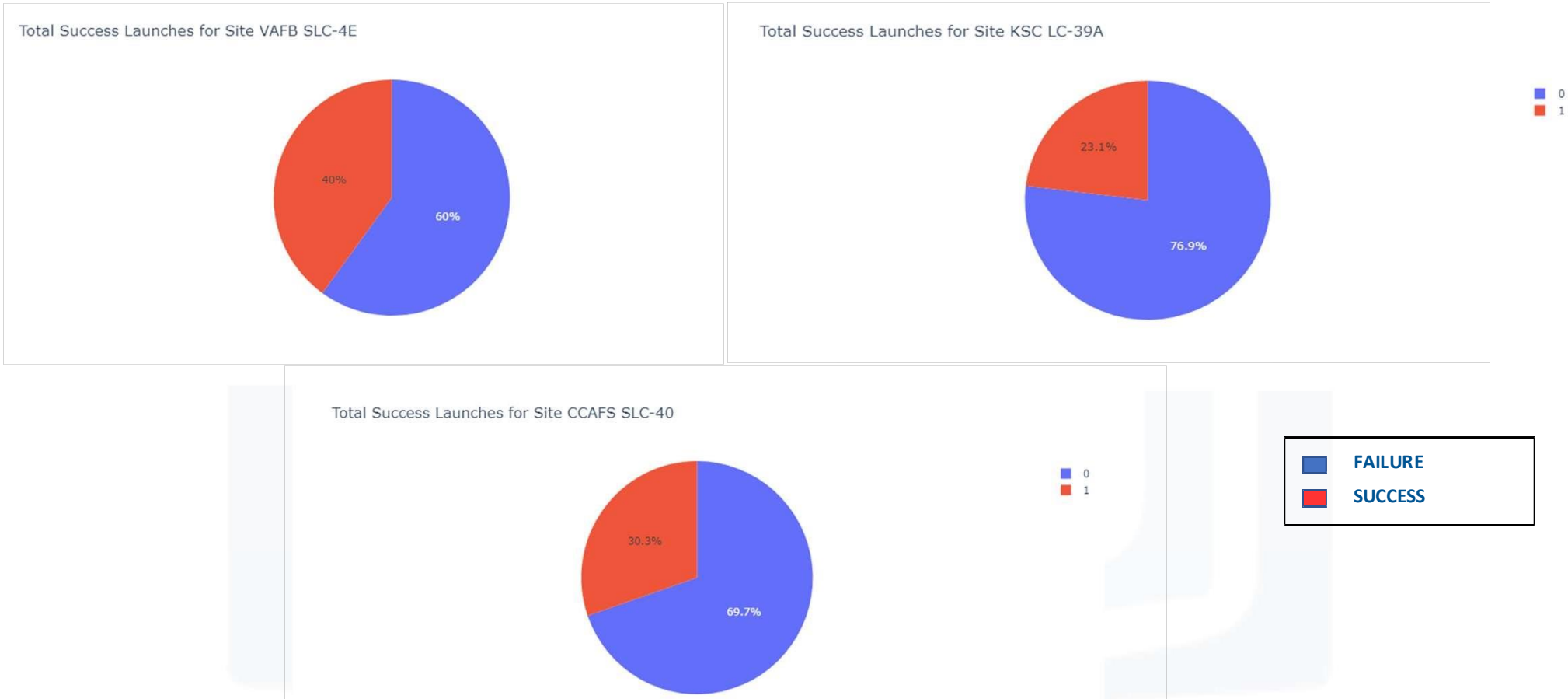


Fig. 21

LAUNCH SUCCESS/FAILURE COUNT FOR EACH SITES

Insight:

The chart clearly shows that from all the launch sites in our sampled dataset of 57 rows, **VAFB SLC-4E** at **40%** has the overall success rate count while, **KSC LC-39A** has the least success rate at **23.1%**.

PAYLOAD MASS VS LAUNCH SUCCESS/FAILURE COUNT FOR EACH SITE VARYING BOOSTER VERSION CATEGORIES

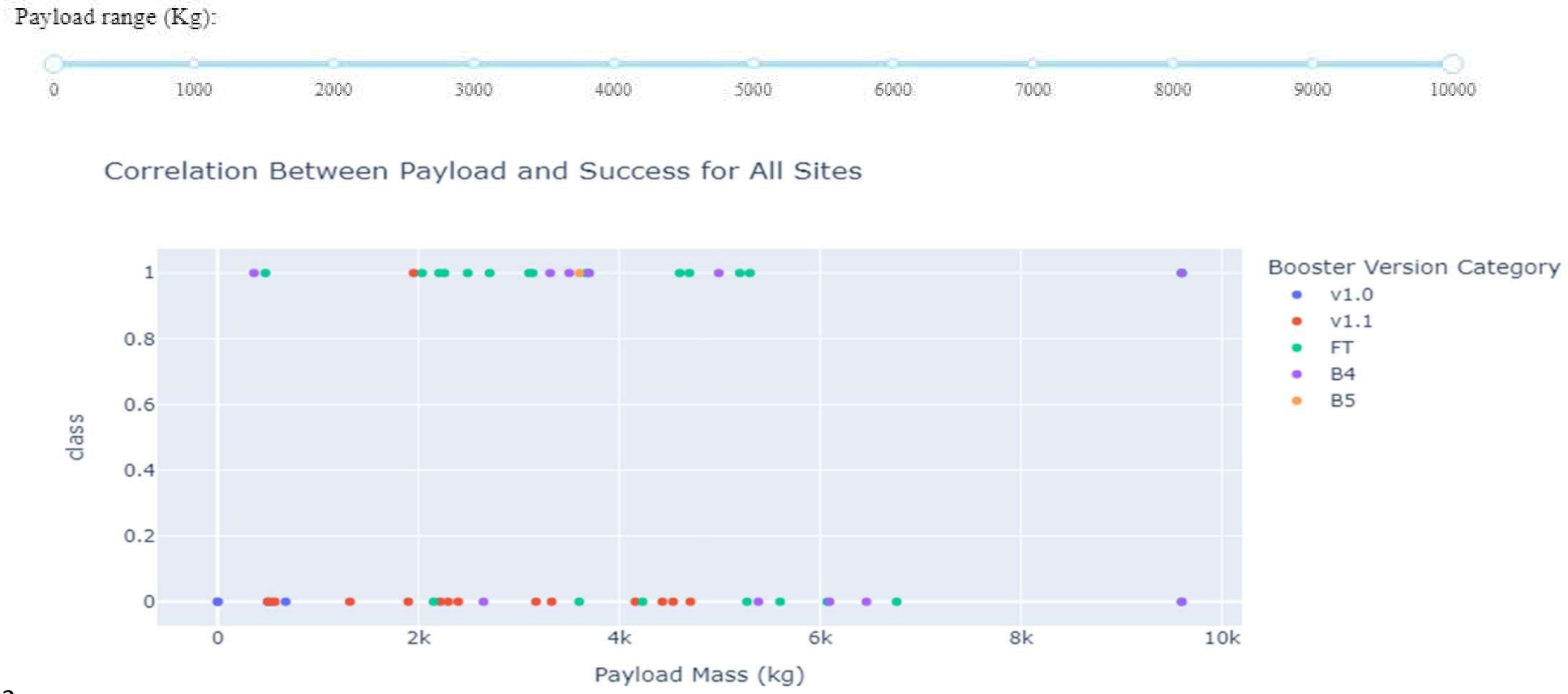


Fig. 22

PAYLOAD MASS VS LAUNCH SUCCESS/FAILURE COUNT FOR VAFB SLC-4E

Fig. 23

PAYLOAD MASS AND LAUNCH SUCCESS FOR SITE CCAFS SLC-40

Fig. 24

PAYLOAD MASS VS LAUNCH SUCCESS/FAILURE COUNT FOR KSC LC-39A

Fig. 25

Insight:

- In the payload mass vs launch success/failure the charts clearly shows that from all the launch sites in our sampled dataset of 57 rows, **booster version FT** has the highest success rate while **booster version v1.1** has the highest failure rate.
- The payload mass size ranges between **0 kg and 53000 kg** except an exception (outlier) with the booster version **B4** at **9600kg.**
- Payloads between **2000 kg** and **5700 kg** have the highest success rate.

FEATURE ENGINEERING FOR PREDICTIVE ANALYSIS

By now, we should have obtain some preliminary Insights about how each important variable would affect the success rate. From our dataset, here we select the features that will be used in our prediction module.

FEATURE SELECTION FOR A SUCCESSFUL PREDICTION

```
features = df[['FlightNumber', 'PayloadMass', 'Orbit', 'Flights',
               'GridFins', 'Reused', 'Legs', 'LandingPad', 'Block', 'ReusedCount', 'Serial']]

features.head()
```

	FlightNumber	PayloadMass	Orbit	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial
0	1	6123.55	LEO	1	False	False	False	NaN	1.0	0	B0003
1	2	525.00	LEO	1	False	False	False	NaN	1.0	0	B0005
2	3	677.00	ISS	1	False	False	False	NaN	1.0	0	B0007
3	4	500.00	PO	1	False	False	False	NaN	1.0	0	B1003
4	5	3170.00	GTO	1	False	False	False	NaN	1.0	0	B1004

Fig. 26

CREATING DUMMY VARIABLES FROM THE SELECTED CATEGORICAL COLUMNS

```
# HINT: Use get_dummies() function on the categorical columns
features_one_hot_ = pd.get_dummies(features[['FlightNumber', 'PayloadMass', 'Orbit', 'Flights',
                'GridFins', 'Reused', 'Legs', 'LandingPad', 'Block', 'ReusedCount', 'Serial']])

features_one_hot_.head()
```

	FlightNumber	PayloadMass	Flights	GridFins	Reused	Legs	Block	ReusedCount	Orbit_ES-L1	Orbit_GEO	...	Serial_B1048	Serial_B1049	Serial_B1050	Ser
0	1	6123.55	1	False	False	False	1.0	0	False	False	...	False	False	False	
1	2	525.00	1	False	False	False	1.0	0	False	False	...	False	False	False	
2	3	677.00	1	False	False	False	1.0	0	False	False	...	False	False	False	
3	4	500.00	1	False	False	False	1.0	0	False	False	...	False	False	False	
4	5	3170.00	1	False	False	False	1.0	0	False	False	...	False	False	False	

Fig. 27

CASTING DUMMY VARIABLES INTO NUMERICAL VARIABLES

```
# HINT: use astype function
features_one_hot_.astype('float64')
```

	FlightNumber	PayloadMass	Flights	GridFins	Reused	Legs	Block	ReusedCount	Orbit_ES-L1	Orbit_GEO	...	Serial_B1048	Serial_B1049	Serial_B1050	Se
0	1.0	6123.55	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.0	0.0	0.0	
1	2.0	525.00	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.0	0.0	0.0	
2	3.0	677.00	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.0	0.0	0.0	
3	4.0	500.00	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.0	0.0	0.0	
4	5.0	3170.00	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.0	0.0	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
85	86.0	15600.00	2.0	1.0	1.0	1.0	5.0	12.0	0.0	0.0	...	0.0	0.0	0.0	
86	87.0	15600.00	3.0	1.0	1.0	1.0	5.0	13.0	0.0	0.0	...	0.0	0.0	0.0	
87	88.0	15600.00	6.0	1.0	1.0	1.0	5.0	12.0	0.0	0.0	...	0.0	0.0	0.0	
88	89.0	15600.00	3.0	1.0	1.0	1.0	5.0	12.0	0.0	0.0	...	0.0	0.0	0.0	
89	90.0	3681.00	1.0	1.0	0.0	1.0	5.0	8.0	0.0	0.0	...	0.0	0.0	0.0	

Fig. 28

PREDICTIVE ANALYSIS USING CLASSIFICATION MODELS

IBM Developer

SKILLS NETWORK

ACCURACY SCORE FOR THE SELECTED CLASSIFICATION MODELS

ACCURACY SCORE FOR THE TEST SET

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

Fig. 29

ACCURACY SCORE FOR THE SELECTED CLASSIFICATION MODELS

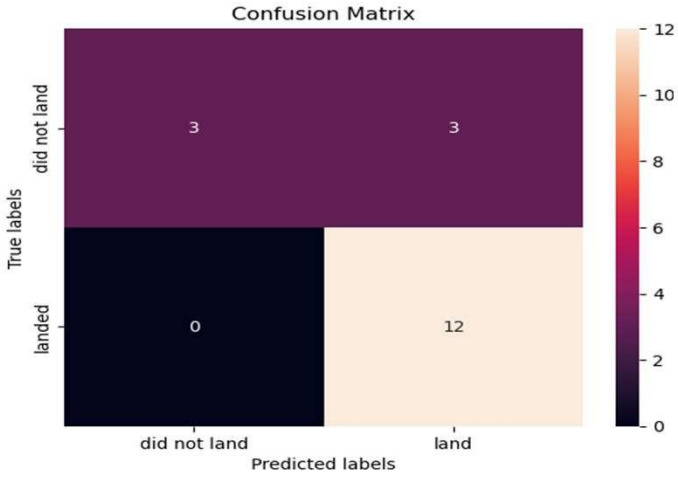
ACCURACY SCORE FOR THE ENTIRE DATA SET

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.819444	0.819444
F1_Score	0.909091	0.916031	0.900763	0.900763
Accuracy	0.866667	0.877778	0.855556	0.855556

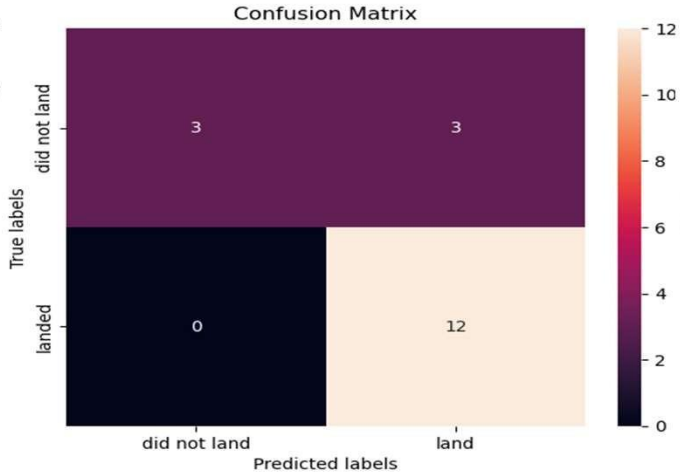
Fig. 30

CONFUSION MATRIX: TEST DATASET

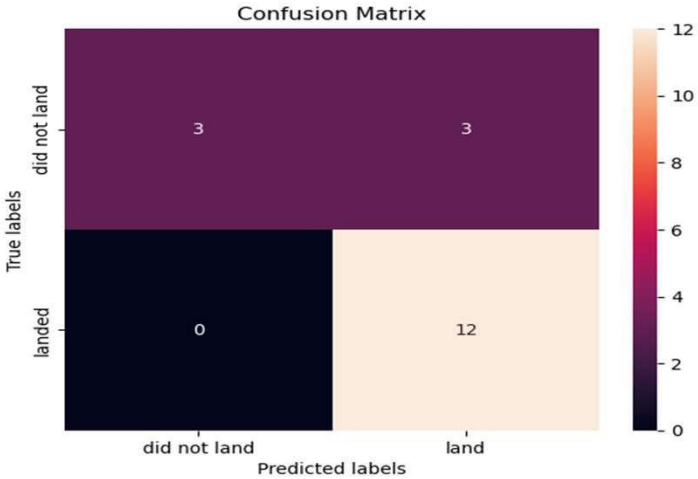
LOGREG



SVM



TREE



KNN

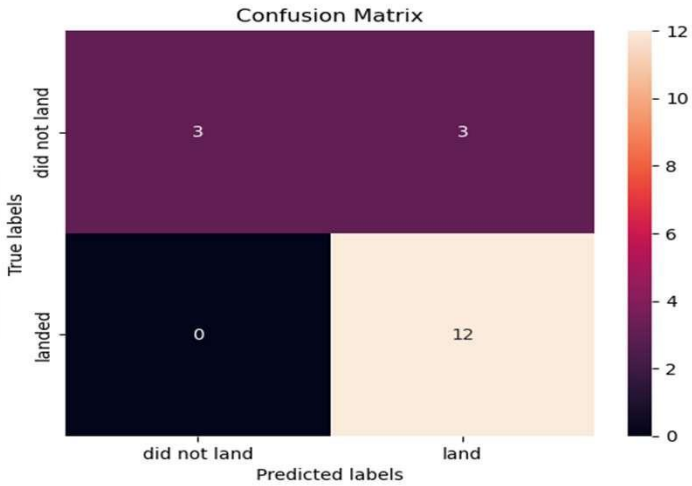
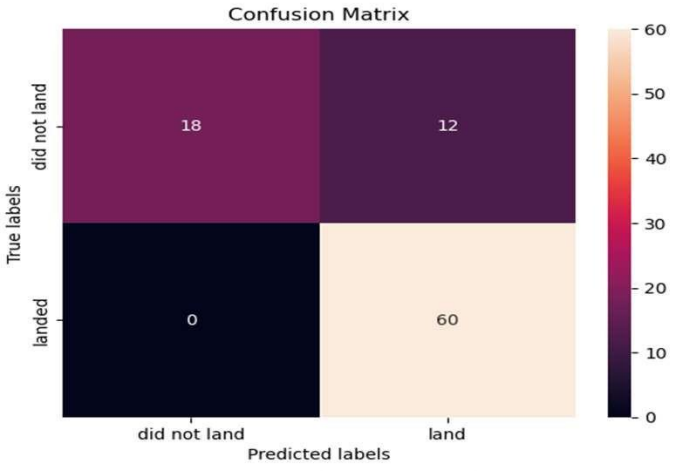


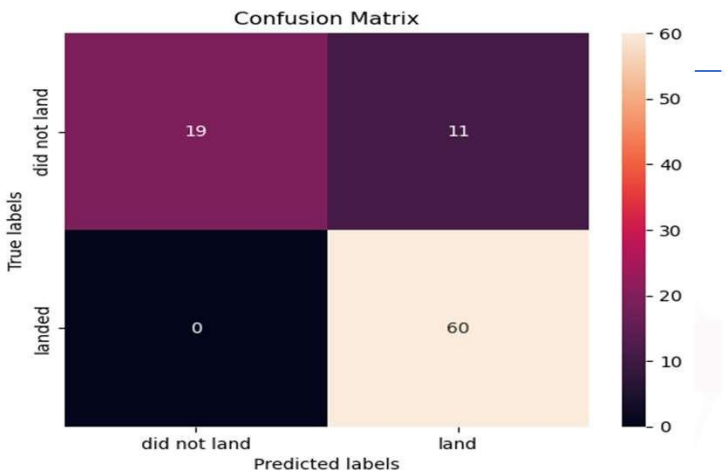
Fig. 31

CONFUSION MATRIX: ENTIRE DATASET

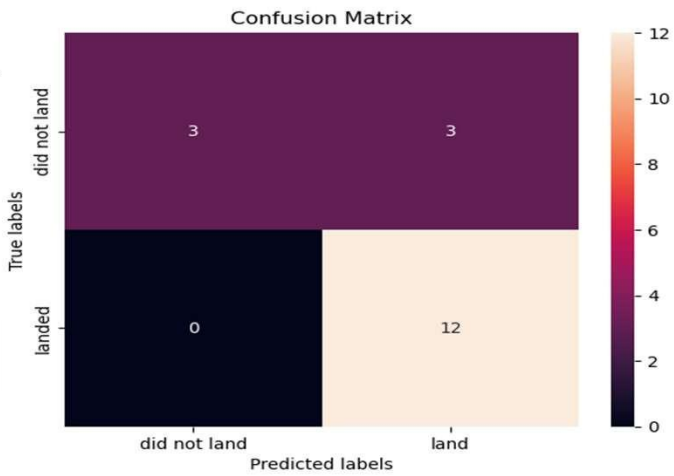
LOGREG



SVM



KNN



TREE

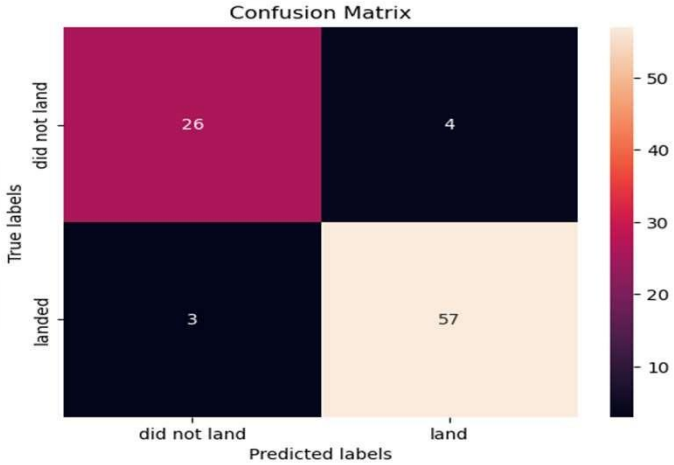


Fig. 32

PREDICTIVE ANALYSIS USING CLASSIFICATION MODELS

Insight:

- Based on the scores of the Test Set, we could not confirm which method performs best since all the models scores were same. Same Test Set scores and confusion matrix(Test Score) may be due to the small test sample size (18 samples).
- Therefore, we tested all methods based on the whole dataset and this gave a better result. The scores of the whole Dataset confirm that the best model is the **Decision Tree Model (TREE).**
- This model has not only higher scores, but also the highest accuracy as seen in the confusion matrix where out of 29 rows for ‘did not land’ we got 26 correctly while, out of 61 rows for ‘land’, we got 67 correctly.

CONCLUSION

According to our analysis, there are important factors affecting the success of the first stage landing :

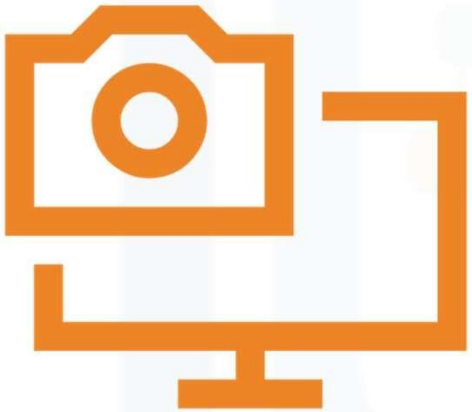
- The **CCAFS SLC 40** launch site has about a half of all launches. This means it is the most used with the highest success count, however, on an average success rating, **KSC LC-39A** has the highest average success rate.
- Launches with a **low payload mass** show better results than launches with a **larger payload mass**.
- **Orbits ES-L1, GEO, HEO and SSO** have 100% success rate
- Payloads between **2000 kg** and **5700 kg** have the highest success rate.
- **Payload Mass** with **Booster version FT** has the highest success rate while **Booster version v1.1** has the highest failure rate.
- The launch success rate depend also on the location and proximities of a launch site, i.e., the initial position of rocket trajectories.
- Most of launch sites are in proximity to the **Equator line** and all the sites are in very close proximity to **the coast**.
- The success rate of launches increases over the years. Since 2013 it kept increasing till 2017 (stable in 2014) and after 2015 it started increasing
- **Decision Tree Model** is the best algorithm for this dataset. Based on the scores of the Test Set, we could not confirm which method performs best since all the models scores were same. Same Test Set scores may be due to the small test sample size (18 samples).
- The scores of the whole Dataset confirm that the best model is the **Decision Tree Model (TREE), h**owever, it is recommended to obtain more data to better improve our models accuracy.

GitHub Link - https://github.com/chill-est/IBM-Data-Science---FINAL-PROJECT

IBM Developer

SKILLS NETWORK

APPENDIX



Special Thanks to:

Instructors

Coursera

IBM