IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Daniel Santiago Sanabria Dueñas
25/10/2024

Outline

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

Executive Summary

- Summary of methodologies

- Data Collection SpaceX API
- Web scraping Falcon 9 and Falcon Heavy Launches Records from Wikipedia
- Exploratory Data Analysis (EDA) using SQL
- Perform interactive visual analytics using Folium and PlotlyDash
- Perform predictive analysis using classification models

- Summary of all results

- EDA results
- •Visual analytics and dashboards
- •Predictive analysis (classification model)
- •Confusion matrix
- •Conclusions

Introduction

- **Project background and context**

In this capstone, 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.

- **Problems you want to find answers**

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.

Section 1

Methodology

Methodology

Executive Summary

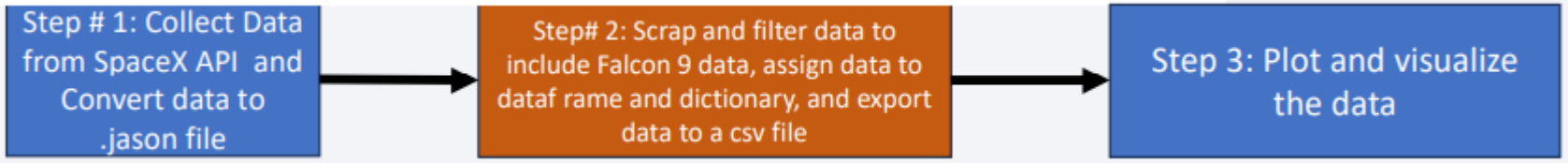
- Data collection methodology:
  - SpaceX API get request
- Perform data wrangling
  - Filter information from Falcon 9 only, change empty data form mass with mass mean
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

- **Describe how data sets were collected.**

Data collection with SpaceX REST calls using key phrases and flowcharts SpaceX API calls notebook

- **My GitHub**
- https://github.com/ZxThiagoZx17/Ciencia_datos_aplicada_capstone



Data Collection – SpaceX API

Present your dataWeb scraping Falcon 9 and Falcon Heavy Launches Records from Wikipedia. Web scrap Falcon 9 launch records with BeautifulSoup. Extract a Falcon 9 launch records HTML table from Wikipedia

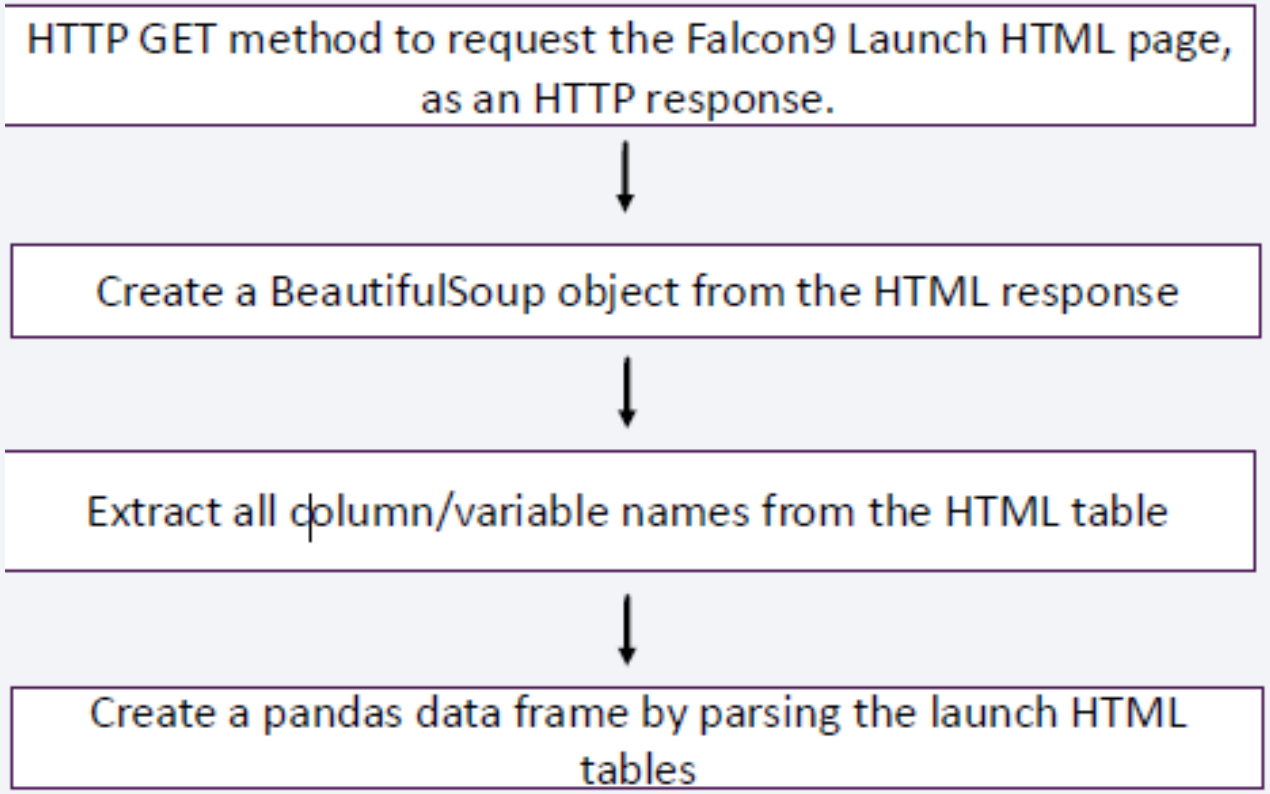
	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	Reus
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
89	86	2020-09-03	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	2	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	
90	87	2020-10-06	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	3	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	
91	88	2020-10-18	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	6	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	
92	89	2020-10-24	Falcon 9	15600.0	VLEO	CCSFS SLC 40	True ASDS	3	True	True	True	5e9e3033383ecbb9e534e7cc	5.0	
93	90	2020-11-05	Falcon 9	3681.0	MEO	CCSFS SLC 40	True ASDS	1	True	False	True	5e9e3032383ecb6bb234e7ca	5.0	

90 rows × 17 columns

https://github.com/ZxThiagoZx17/Ciencia_datos_aplicada_capstone

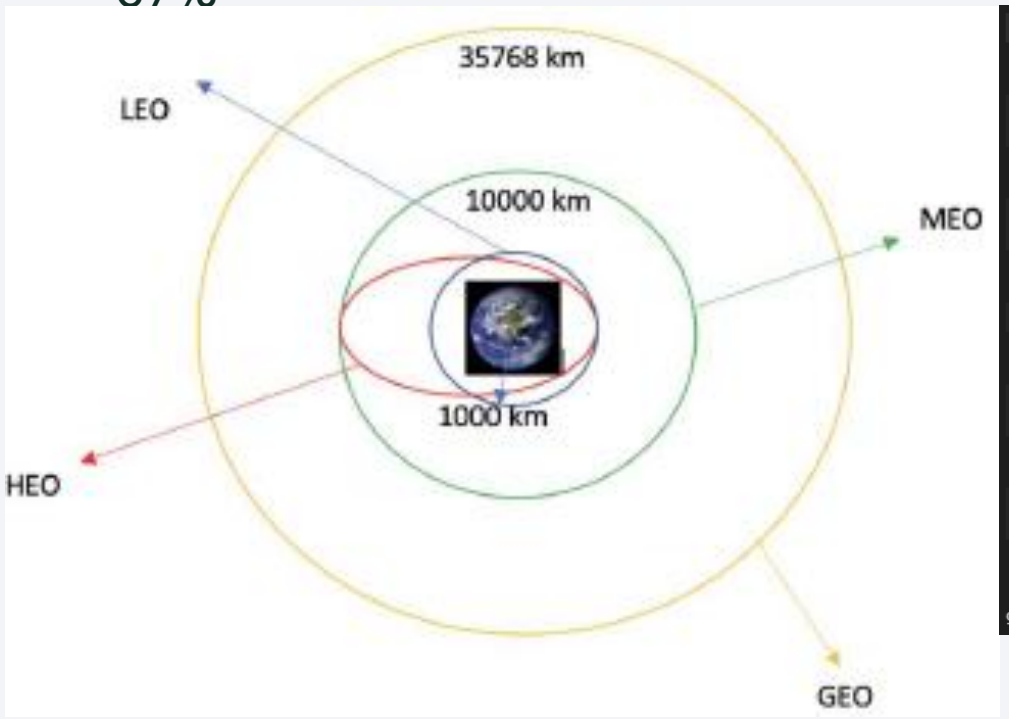
Data Collection - Scraping

- Web scraping Falcon 9 and Falcon Heavy Launches Records from Wikipedia. Web scrap Falcon 9 launch records with BeautifulSoup. Extract a Falcon 9 launch records HTML table from Wikipedia
- https://github.com/ZxThiagoZx17/Ciencia_datos_aplicada_capstone



Data Wrangling

From SpaceX dataset data collection. Using the method value_counts()on the column Launch Site to determine the number of launches on each site, Calculate the number and occurrence of each orbit, Calculate the number and occurence of mission outcome per orbit type. Create a landing outcome label from Outcome column. We can use the landing outcome label mean to determine the success rate = 67%



	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	Reus
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
89	86	2020-09-03	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	2	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	
90	87	2020-10-06	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	3	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	
91	88	2020-10-18	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	6	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	
92	89	2020-10-24	Falcon 9	15600.0	VLEO	CCSFS SLC 40	True ASDS	3	True	True	True	5e9e3033383ecbb9e534e7cc	5.0	
93	90	2020-11-05	Falcon 9	3681.0	MEO	CCSFS SLC 40	True ASDS	1	True	False	True	5e9e3032383ecb6bb234e7ca	5.0	

90 rows × 17 columns

https://github.com/ZxThiagoZx17/Ciencia_datos_aplicada_capstone

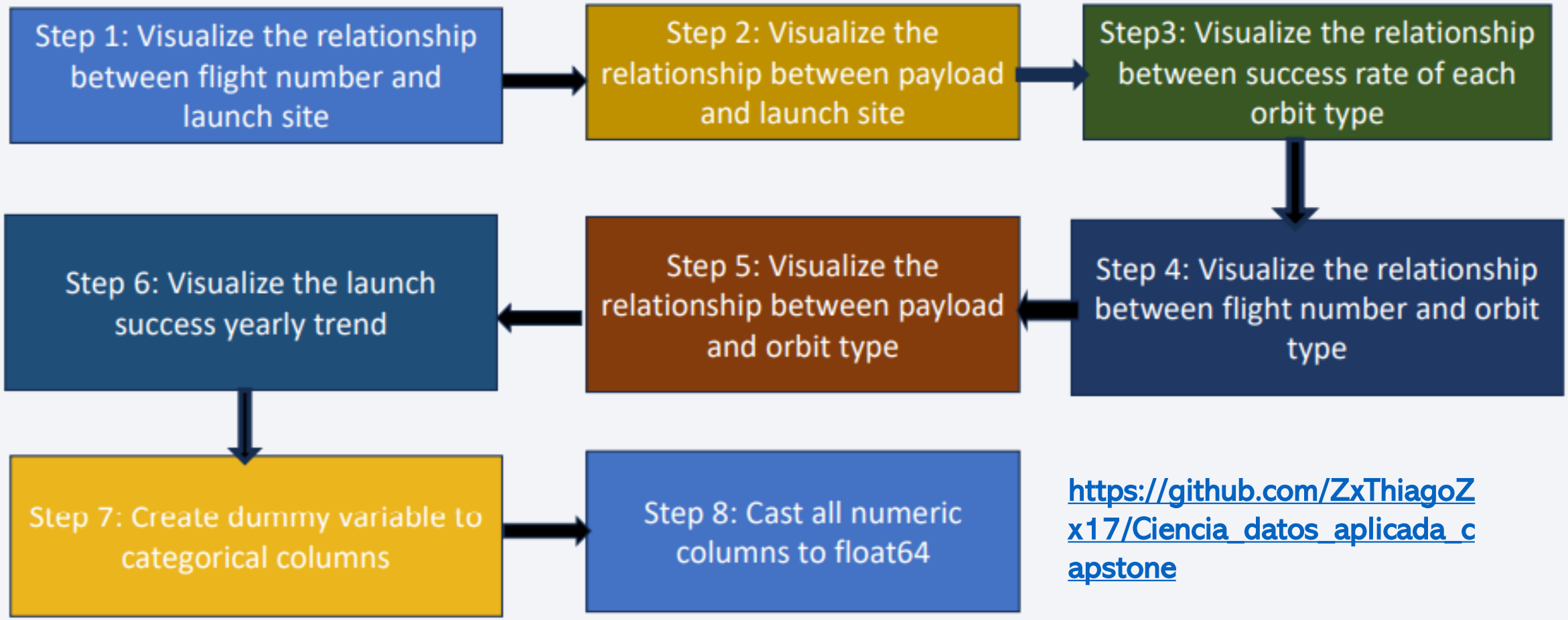
EDA with SQL

Names of the unique launch sites in the space mission:

- %sql SELECT DISTINCT LAUNCH_SITE as "Launch_Sites" FROM SPACEXTBL;
- Total payload mass carried by boosters launched by NASA (CRS):
- %sql SELECT SUM(PAYLOAD_MASS__KG_) as "Total Payload Mass(Kgs)", Customer FROM 'SPACEXTBL' WHERE Customer = 'NASA (CRS)';
- Average payload mass carried by booster version F9 v1.1:
- %sql SELECT AVG(PAYLOAD_MASS__KG_) as "Payload Mass Kgs", Customer, Booster_Version FROM 'SPACEXTBL' WHERE Booster_Version LIKE 'F9 v1.1%';
- Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.
- %sql SELECT * FROM SPACEXTBL WHERE "Landing _Outcome" LIKE 'Success%' AND (Date BETWEEN '04-06-2010' AND '20-03-2017') ORDER BY Date DESC;

https://github.com/ZxThiagoZx17/Ciencia_datos_aplicada_capstone

EDA and data visualization



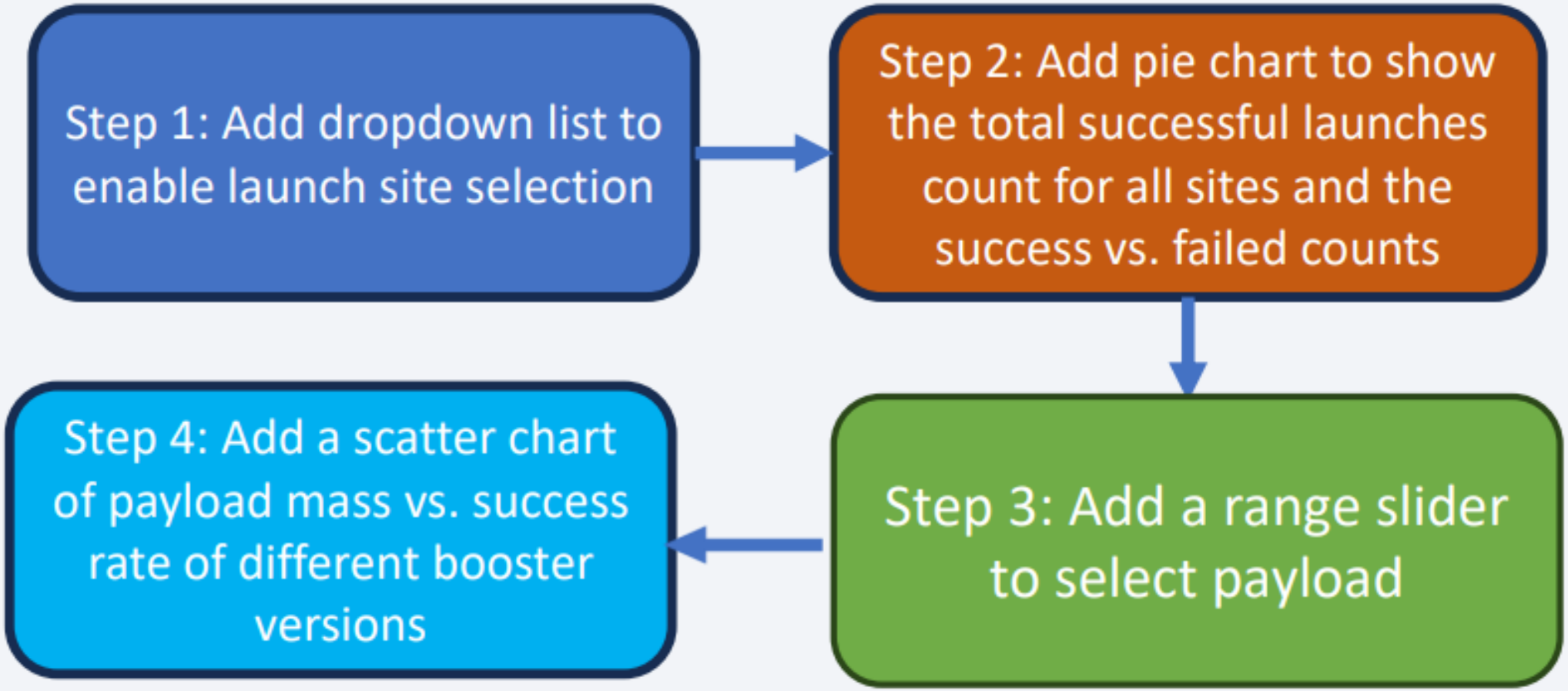
https://github.com/ZxThiagoZx17/Ciencia_datos_aplicada_capstone

Build an Interactive Map with Folium

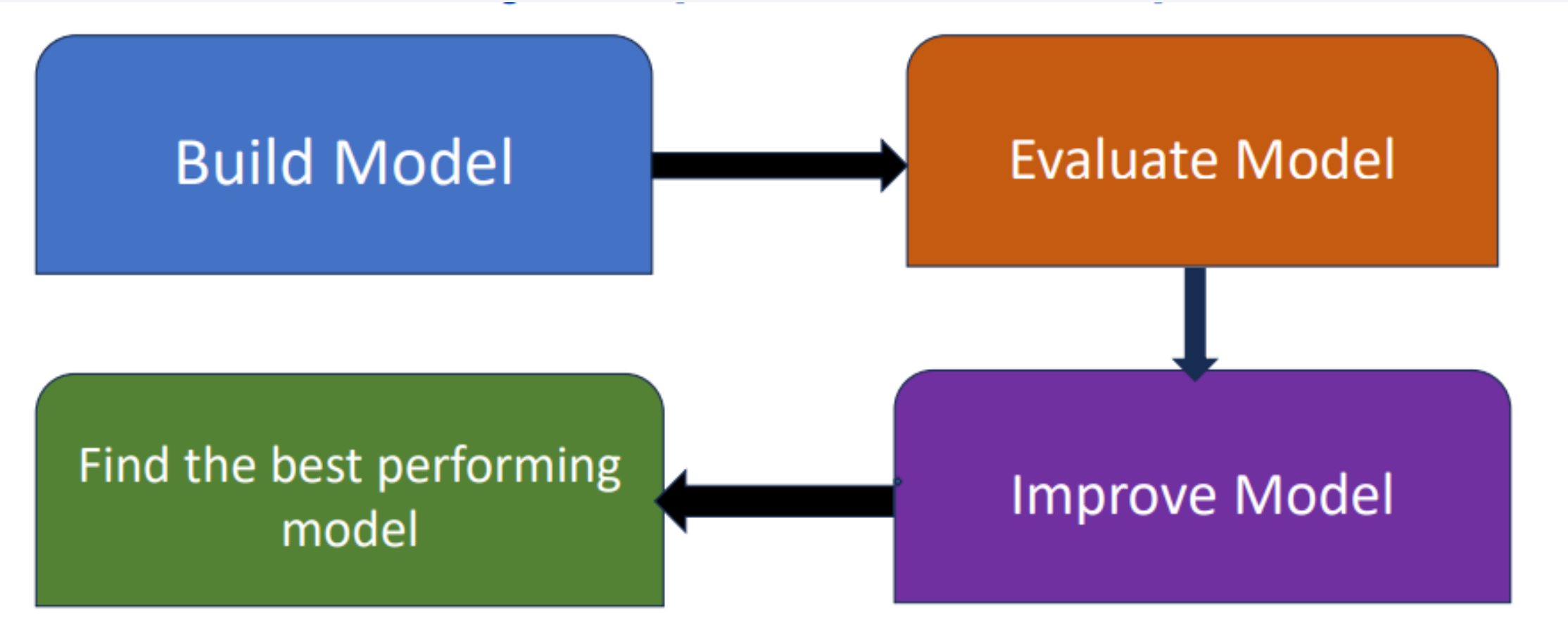
- On the Map it were included a circle mark for the launch sites.
- A color mark for the success rate in each site.
- Also it was calculated the distance to strategic locations such as railways, highways, coastline and cities. It can be observed as a line in the map

https://github.com/ZxThiagoZx17/Ciencia_datos_aplicada_capstone

Build a Dashboard with Plotly Dash



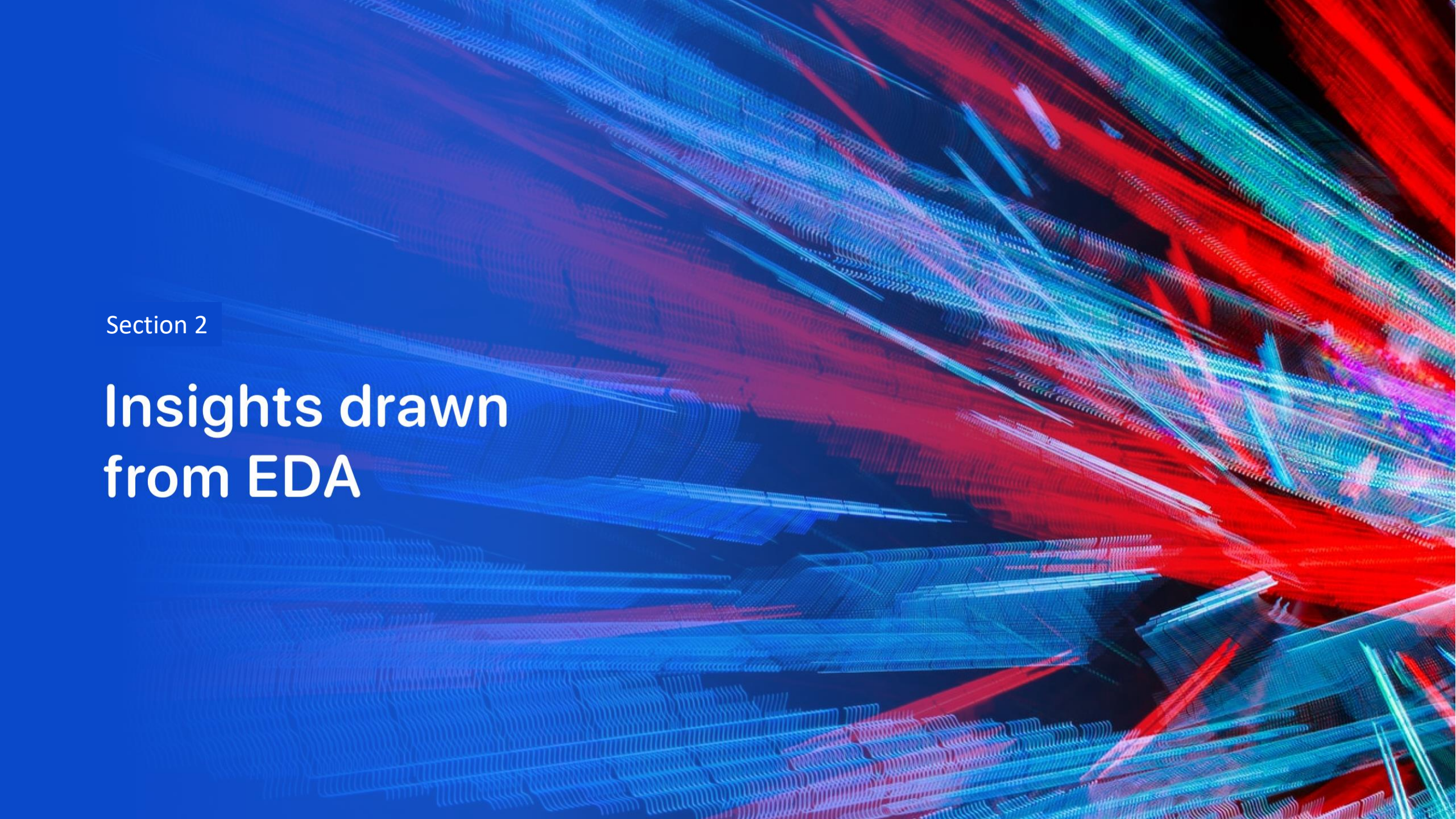
Predictive Analysis (Classification)



https://github.com/ZxThiagoZx17/Ciencia_datos_aplicada_capstone

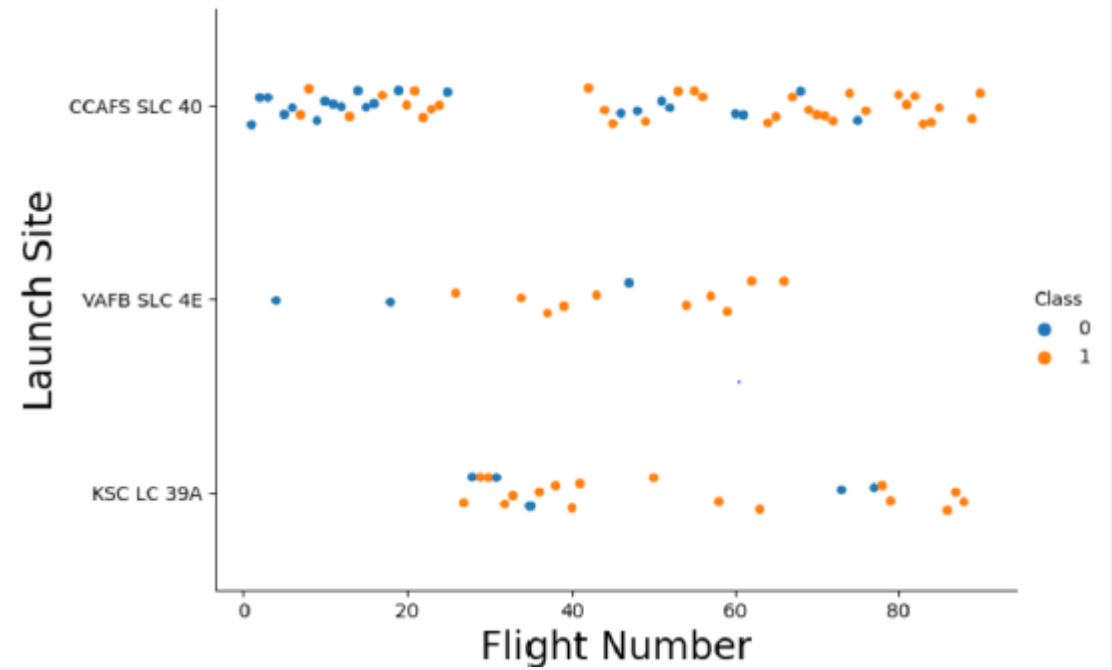
Results

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


Section 2
Insights drawn from EDA

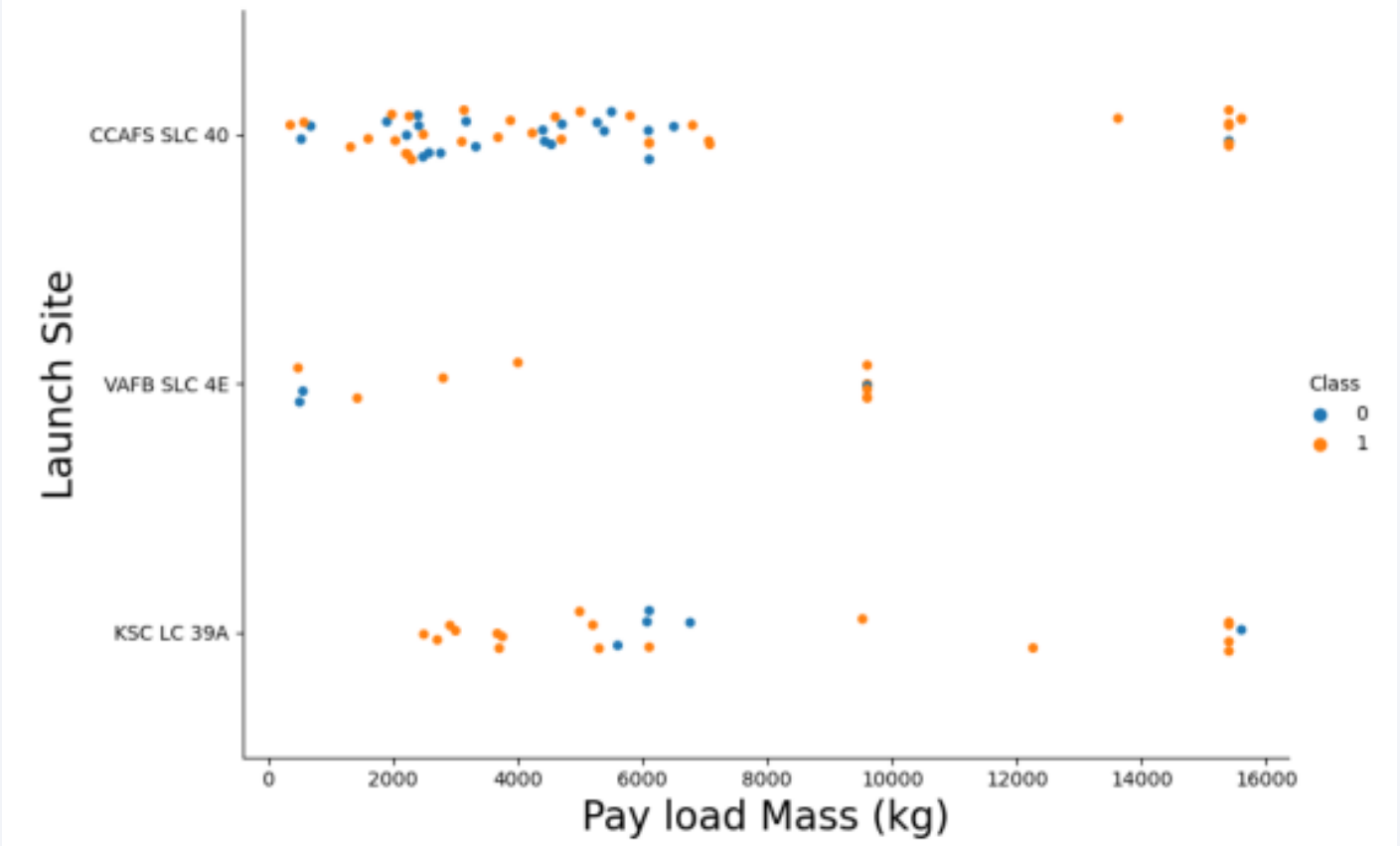
Flight Number vs. Launch Site

- The majority of the flights were launched from the CCAFS SLC 40 sites.
- The VAFB SLC 4E and KSC LC 39A sites have higher success rates than other sites.
- Newer flights have higher success rates than older flights.



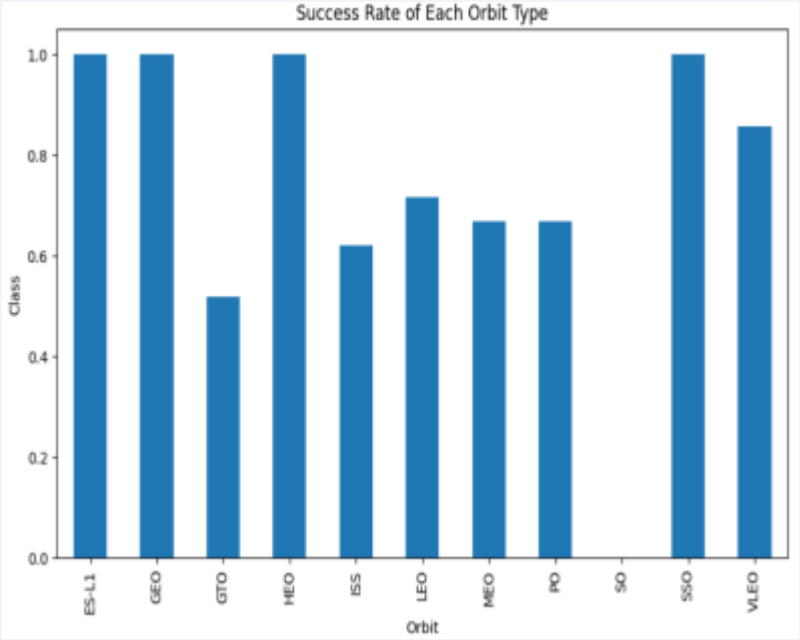
Payload vs. Launch Site

- If you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).
- KSCLC39A show a 100% success rate for payload below (5000)



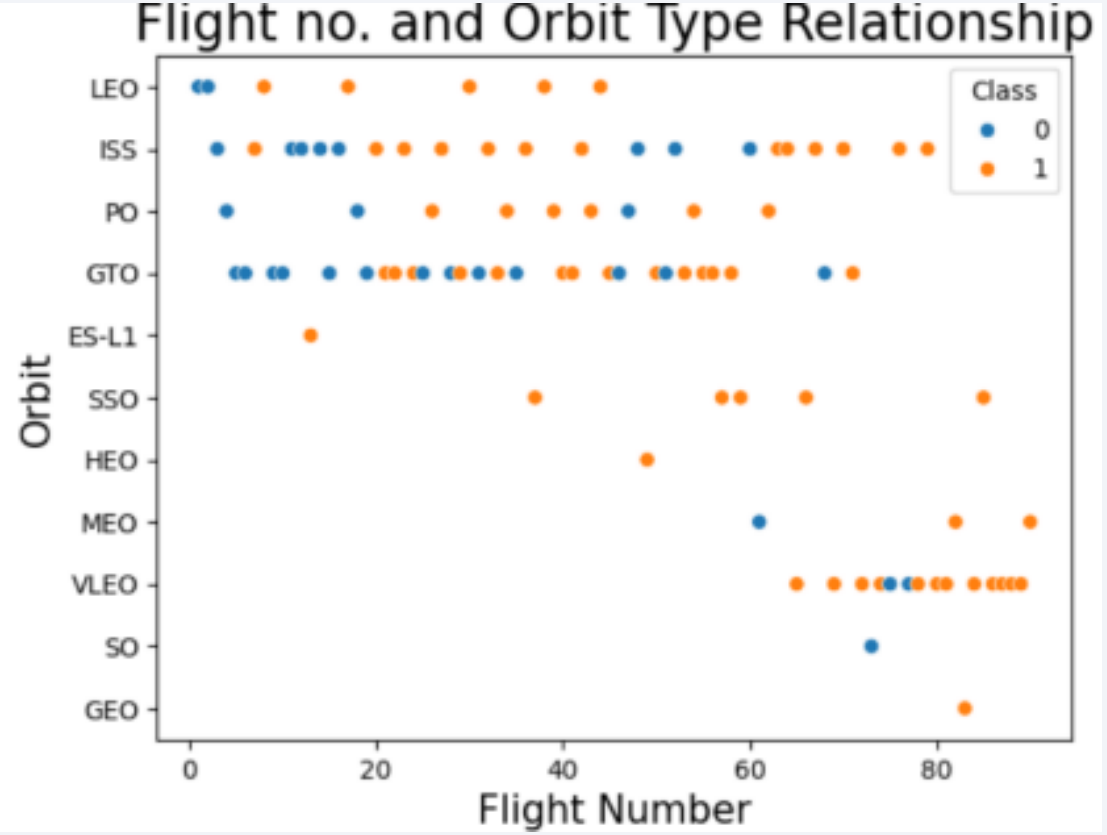
Success Rate vs. Orbit Type

- From this bar chart we can compare the success rate for each orbit type.
- Orbits ES-L1, GEO, HEO & SSO have the highest success rates at 100%, with GTO orbit having almost the lowest success rate at 50%. Orbit SO has 0% success rate.



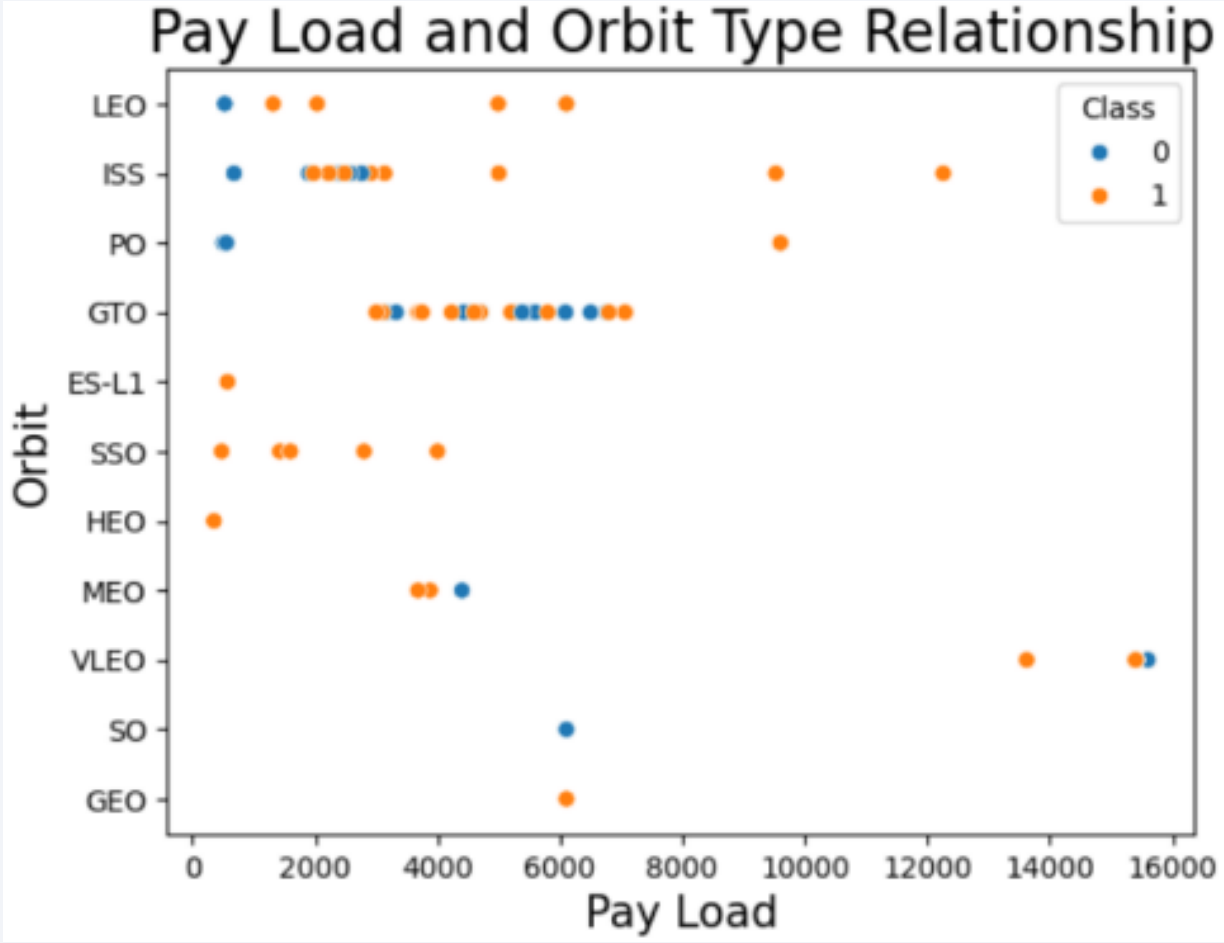
Flight Number vs. Orbit Type

- The majority of the flights were launches to the ISS and GTO orbits.
- The data suggests that there is no relationship between the flight number and the orbit type.



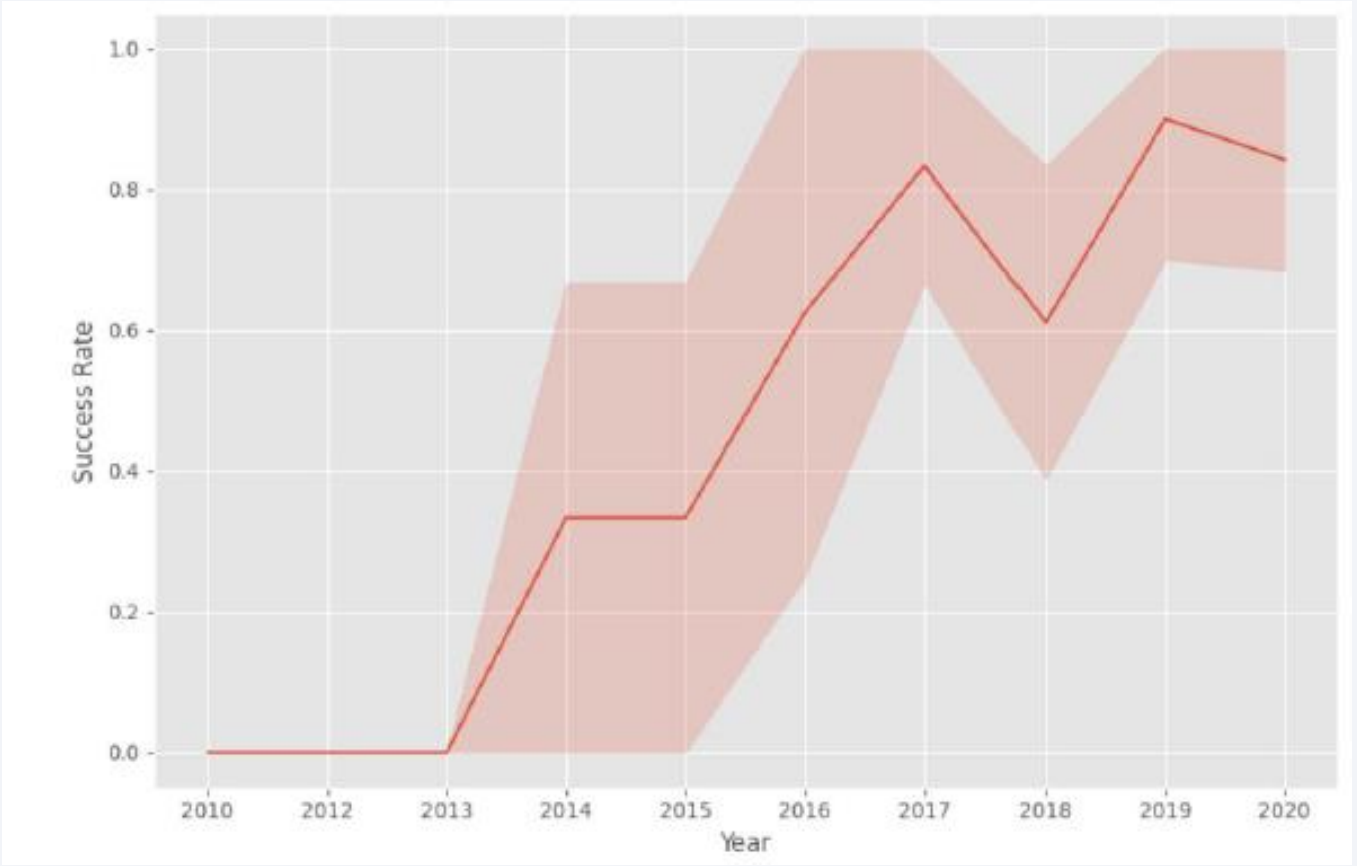
Payload vs. Orbit Type

- Payload masses above 10000 Kg were placed in PO, ISS and LEO orbits.
- Payload masses above 4000 and less than 8000 Kg were placed in the GTO orbit.



Launch Success Yearly Trend

- We can see a line chart of yearly average success rate
- You can observe that the success rate since 2013 kept increasing till 2020



All Launch Site Names

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

Launch Site Names Begin with 'CCA'

- 5 records for launch sites begin with the string 'CCA' and the query used for obtaining the information is shown below

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	[[SpaceX], \n]	Success\n	F9 v1.07B0003.18	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	[[.mw-parser-output .plainlist ol,.mw-parser-o...	Success	F9 v1.07B0004.18	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	[[NASA], (, [COTS],)\n]	Success	F9 v1.07B0005.18	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	[[NASA], (, [CRS],)\n]	Success\n	F9 v1.07B0006.18	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	[[NASA], (, [CRS],)\n]	Success\n	F9 v1.07B0007.18	No attempt\n	1 March 2013	15:10

Total Payload Mass

- %sql SELECT SUM(PAYLOAD_MASS__KG_) as "Total Payload Mass(Kgs)", Customer FROM 'SPACEXTBL' WHERE Customer = 'NASA (CRS)';
- The calculated total payload mass carried by boosters from NASA site =45596 Kg.

Total Payload Mass(Kgs)	Customer
45596	NASA (CRS)

Average Payload Mass by F9 v1.1

- %sql SELECT AVG(PAYLOAD_MASS__KG_) as "Payload Mass Kgs", Customer, Booster_Version FROM 'SPACEXTBL' WHERE Booster_Version LIKE 'F9 v1.1%';
- The average payload mass carried by booster version F9 v1.1=2534.7 Kg.

Payload Mass Kgs	Customer	Booster_Version
2534.6666666666665	MDA	F9 v1.1 B1003

First Successful Ground Landing Date

- %sql SELECT MIN(DATE) FROM 'SPACEXTBL' WHERE "Landing _Outcome" = "Success (ground pad)";
- The first successful landing outcome on a ground pad was in 2017-01-05.

Successful Drone Ship Landing with Payload between 4000 and 6000

- %sql SELECT DISTINCT Booster_Version, Payload FROM SPACEXTBL WHERE "Landing _Outcome" = "Success (drone ship)" AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

Booster_Version	Payload
F9 FT B1022	JCSAT-14
F9 FT B1026	JCSAT-16
F9 FT B1021.2	SES-10
F9 FT B1031.2	SES-11 / EchoStar 105

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes
- %sql SELECT "Mission_Outcome", COUNT("Mission_Outcome") as Total FROM SPACEXTBL GROUP BY "Mission_Outcome";

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- %sql SELECT "Booster_Version",Payload, "PAYLOAD_MASS__KG_" FROM SPACEXTBL WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTBL);

Booster_Version	Payload	PAYLOAD_MASS__KG_
F9 B5 B1048.4	Starlink 1 v1.0, SpaceX CRS-19	15600
F9 B5 B1049.4	Starlink 2 v1.0, Crew Dragon in-flight abort test	15600
F9 B5 B1051.3	Starlink 3 v1.0, Starlink 4 v1.0	15600
F9 B5 B1056.4	Starlink 4 v1.0, SpaceX CRS-20	15600
F9 B5 B1048.5	Starlink 5 v1.0, Starlink 6 v1.0	15600
F9 B5 B1051.4	Starlink 6 v1.0, Crew Dragon Demo-2	15600
F9 B5 B1049.5	Starlink 7 v1.0, Starlink 8 v1.0	15600
F9 B5 B1060.2	Starlink 11 v1.0, Starlink 12 v1.0	15600
F9 B5 B1058.3	Starlink 12 v1.0, Starlink 13 v1.0	15600
F9 B5 B1051.6	Starlink 13 v1.0, Starlink 14 v1.0	15600
F9 B5 B1060.3	Starlink 14 v1.0, GPS III-04	15600
F9 B5 B1049.7	Starlink 15 v1.0, SpaceX CRS-21	15600

2015 Launch Records

- List the failedlanding_outcomes in drone ship, their booster versions, andlaunch site names for in year 2015
- %sql SELECT substr(Date,7,4), substr(Date, 4, 2),"Booster_Version", "Launch_Site", Payload, "PAYLOAD_MASS__KG_", "Mission_Outcome", "Landing _Outcome" FROM SPACEXTBL WHERE substr(Date,7,4)='2015' AND "Landing _Outcome" = 'Failure (drone ship)';

substr(Date,7,4)	substr(Date, 4, 2)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Mission_Outcome	Landing _Outcome
2015	01	F9 v1.1 B1012	CCAFS LC-40	SpaceX CRS-5	2395	Success	Failure (drone ship)
2015	04	F9 v1.1 B1015	CCAFS LC-40	SpaceX CRS-6	1898	Success	Failure (drone ship)

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

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order
- %sql SELECT * FROM SPACEXTBL WHERE "Landing _Outcome" LIKE 'Success%' AND (Date BETWEEN '04-06-2010' AND '20-03-2017') ORDER BY Date DESC;

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
19-02-2017	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
18-10-2020	12:25:57	F9 B5 B1051.6	KSC LC-39A	Starlink 13 v1.0, Starlink 14 v1.0	15600	LEO	SpaceX	Success	Success
18-08-2020	14:31:00	F9 B5 B1049.6	CCAFS SLC-40	Starlink 10 v1.0, SkySat-19, -20, -21, SAOCOM 1B	15440	LEO	SpaceX, Planet Labs, PlanetIQ	Success	Success
18-07-2016	04:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
18-04-2018	22:51:00	F9 B4 B1045.1	CCAFS SLC-40	Transiting Exoplanet Survey Satellite (TESS)	362	HEO	NASA (LSP)	Success	Success (drone ship)
17-12-2019	00:10:00	F9 B5 B1056.3	CCAFS SLC-40	JCSat-18 / Kacific 1, Starlink 2 v1.0	6956	GTO	Sky Perfect JSAT, Kacific 1	Success	Success
16-11-2020	00:27:00	F9 B5B1061.1	KSC LC-39A	Crew-1, Sentinel-6 Michael Freilich	12500	LEO (ISS)	NASA (CCP)	Success	Success
15-12-2017	15:36:00	F9 FT B1035.2	CCAFS SLC-40	SpaceX CRS-13	2205	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)

Section 3

Launch Sites Proximities Analysis

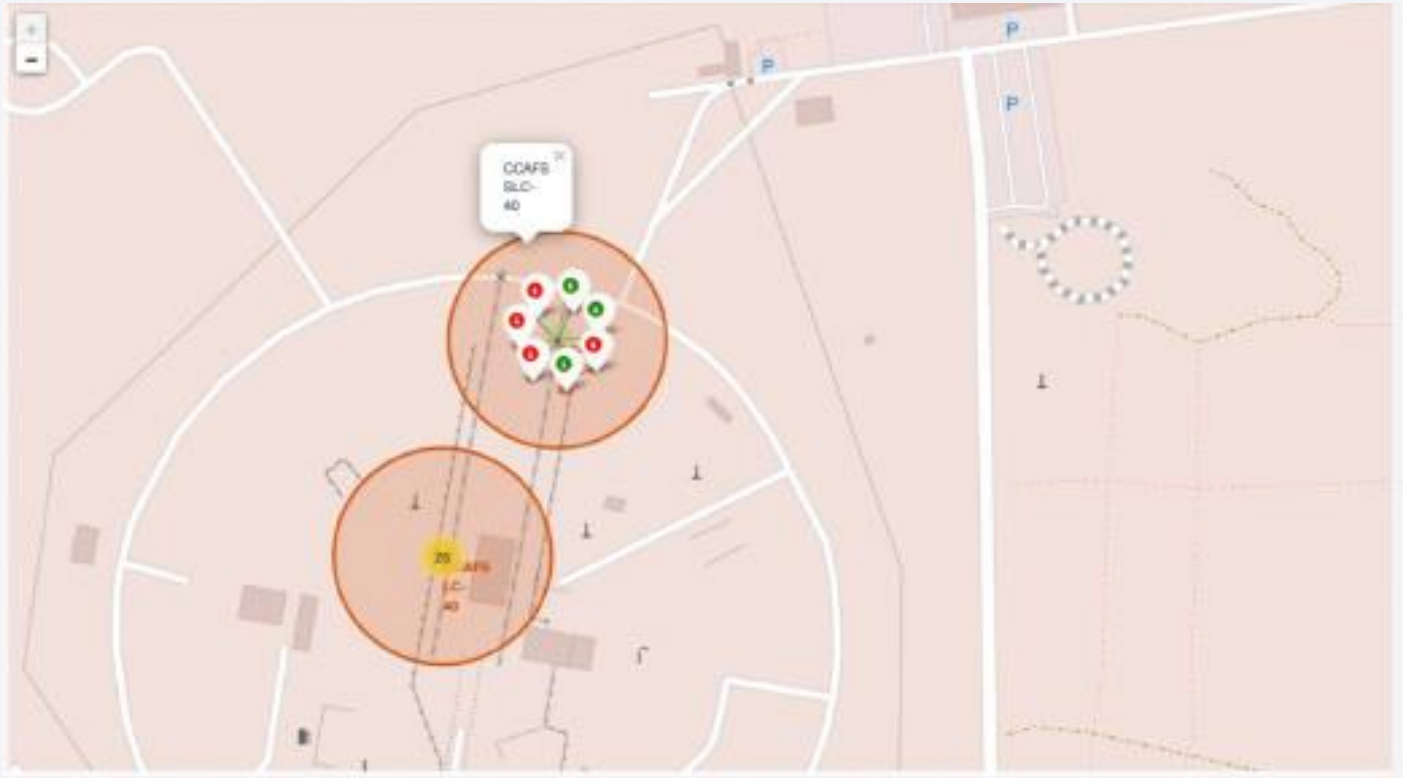
USA Launch Sites in California and Florida

• Most of Launch sites 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.

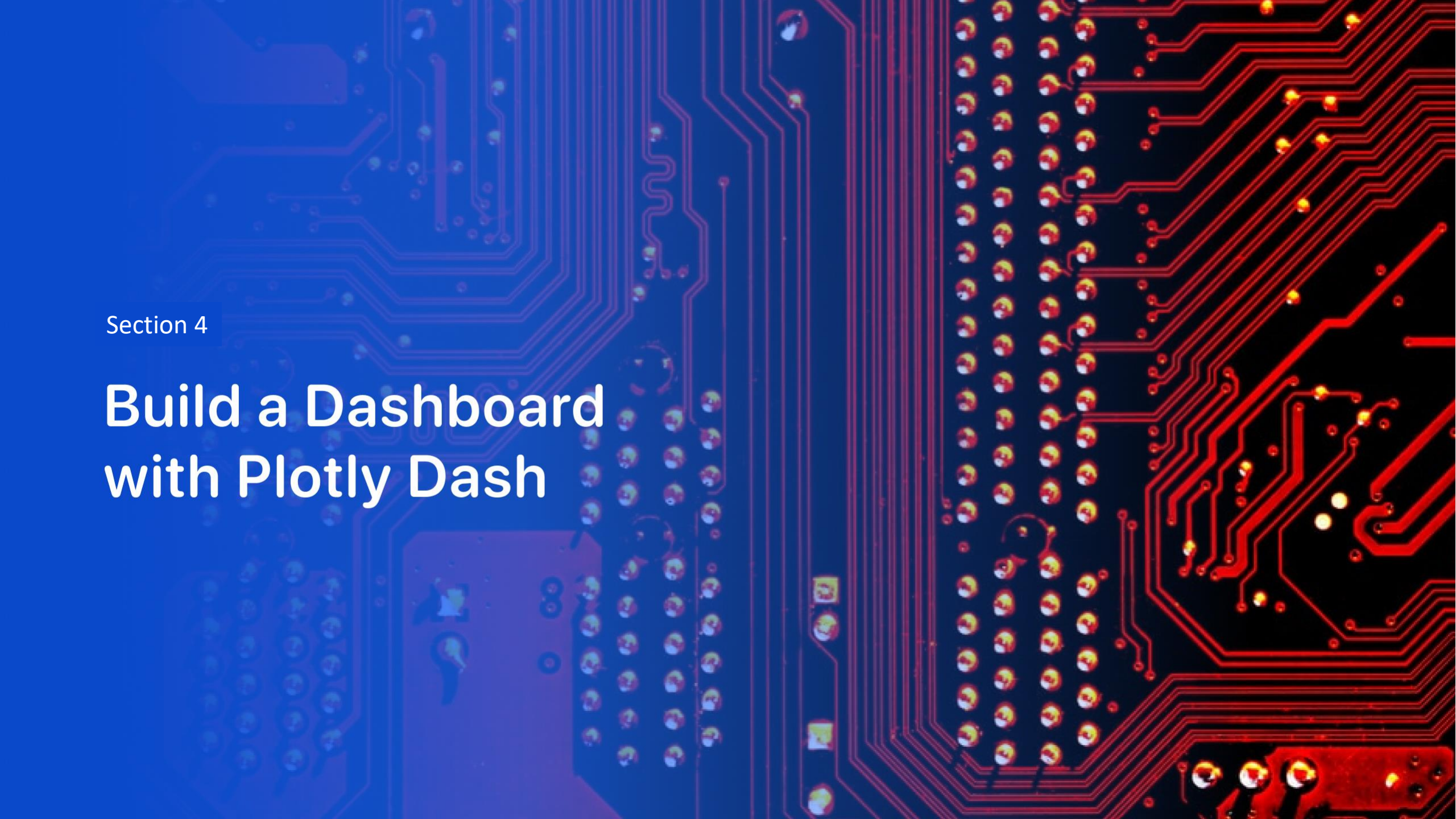
Color marked success/failed launches for each site on the map

From the color-labeled markers in marker clusters,we should be able to easily identify which launch sites have relatively high success rates.



Safe Distance to Launch Site

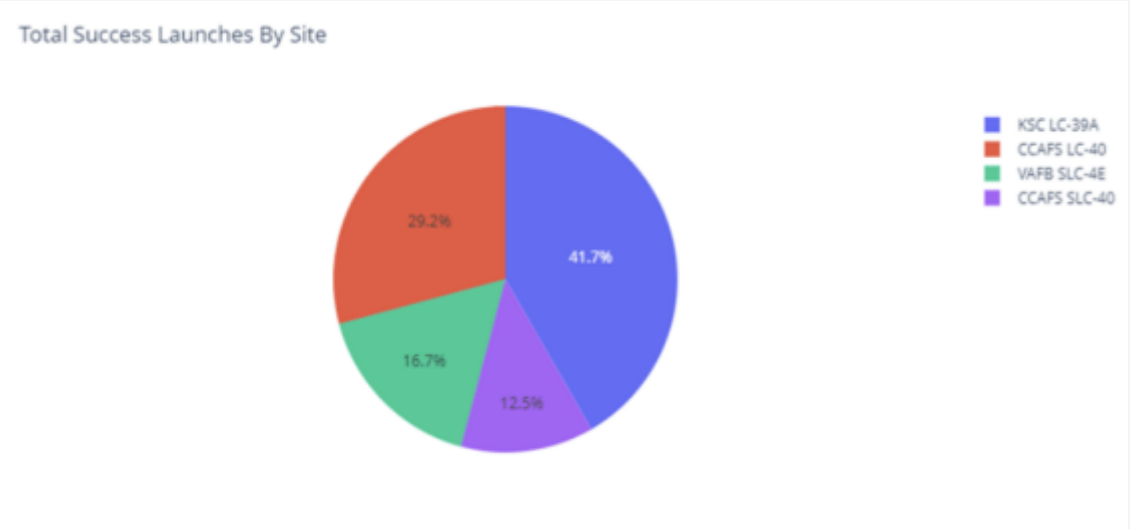
- The obtained results indicate that all launch sites are at safe distance from railway lines and cities.


Section 4
Build a Dashboard with Plotly Dash

Total Launch Success for All Sites

- The highest success launch rates were recorded at these sites :

1. KSC LC-39A (41.7%)
2. CCAFS LC-40 (29.2%)

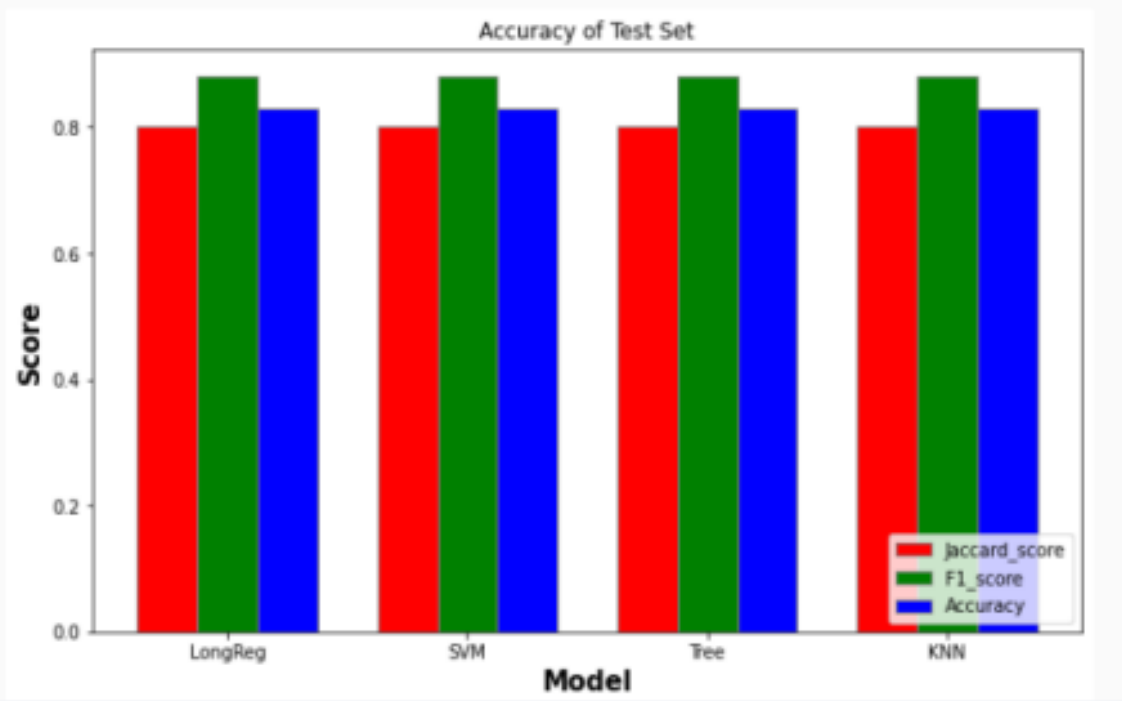


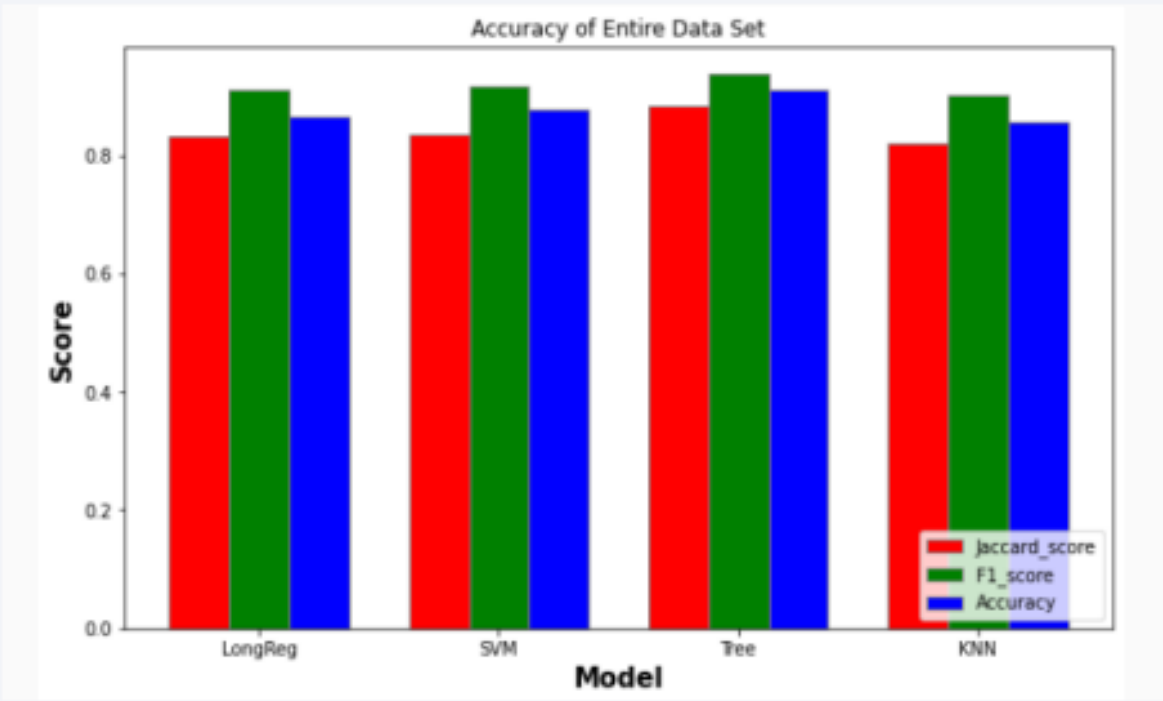
Section 5

Predictive Analysis (Classification)

Classification Accuracy

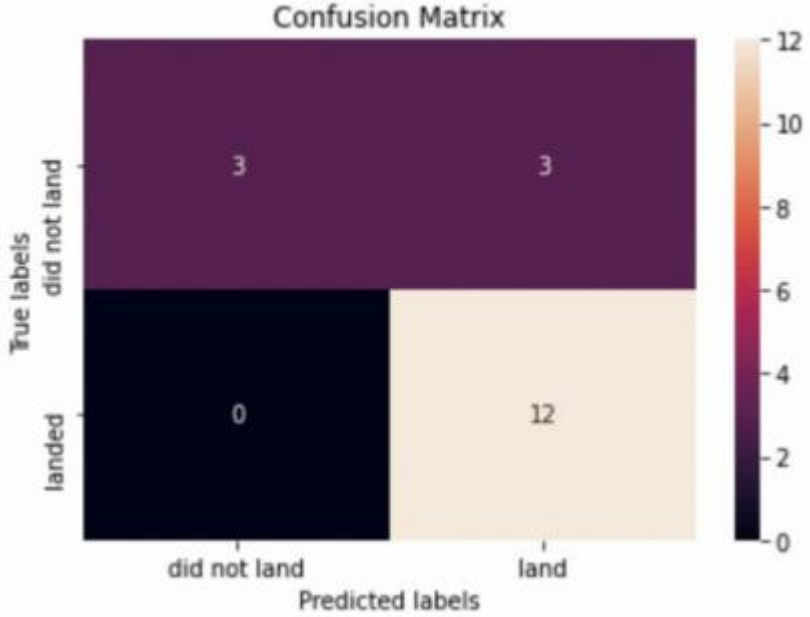
- Using the test set the same accuracy results were obtained from the four models.
- The Tree Model provided the best accuracy results for the entire data set.





Confusion Matrix

- The confusion matrix analysis suggests that the best performing model is the Logistic Regression model.
- The confusion matrix predicts 13 true positives, 3 false positives, 3 true positive, and 0 false negative.



Conclusions

- Point 1: On the initial analysis it can be observed a general success rate of 67% from the real data.
- Point 2: 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%
- Point 3: There is a clear relation between the outcome and variables like: flight number, payload mass, launch site and orbit. It can be observed on the exploratory data analysis using SQL and the visual analytics and dashboards presented on this project.
- Point 4: On the predictive analysis It was discovered that all 4 models (Logistic regression, SVM, decision tree, and KNN had the same accuracy on the test data = 83% also the same confusion matrix.
- For now we can use any of the 4 models to predict the outcome of a next mission with variables related in this project. It is recommend to repeat the 4 models evaluation by the time there are new data to compare.

Thank you!