Winning Space Race with Data Science

Ramesh D Jadhav

OUTLINE

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

EXECUTIVE SUMMARY

Summary of Methodologies:

- We used a data driven approach to find a AI based solution to find , if we can predict whether a particular rocket (Falcon 9) will successfully re-land after its launch.
- The results seams promising. We used Different machine learning models (Logistic Regression, SVC, Decision Tree, KNN) and concluded that decision tree best suits for this application.

Summary of all results:

The Decision Tree gave training accuracy of 0.89 while giving a test accuracy of 0.88.

INTRODUCTION

Project Background and Context:

We all know that SpaceX is the first company to reuse the first stage(Main Part in a rocket) for future launches. So, by using the historical data of SpaceX‘s rocket launches we can find if a future rocket’s first stage can re-land safely using Artificial Intelligence. The main objective of the project is to come up with a model that takes in future rocket launches initial data and predicts if it lands safely back to earth. Using this model, we can decide how much we need to invest in future launches. This is of great advantage to business.

Problem Statement:

The problem is simple. We should develop a model that predicts the success of a future rockets re-landing .

Section 1

Methodology

METHODOLOGY

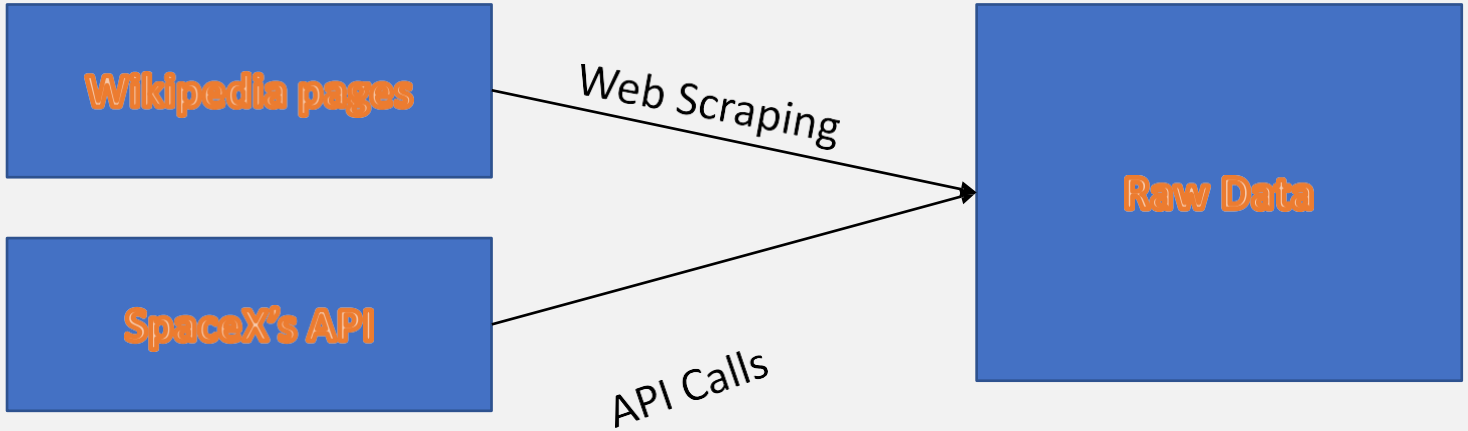
Executive Summary

- 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

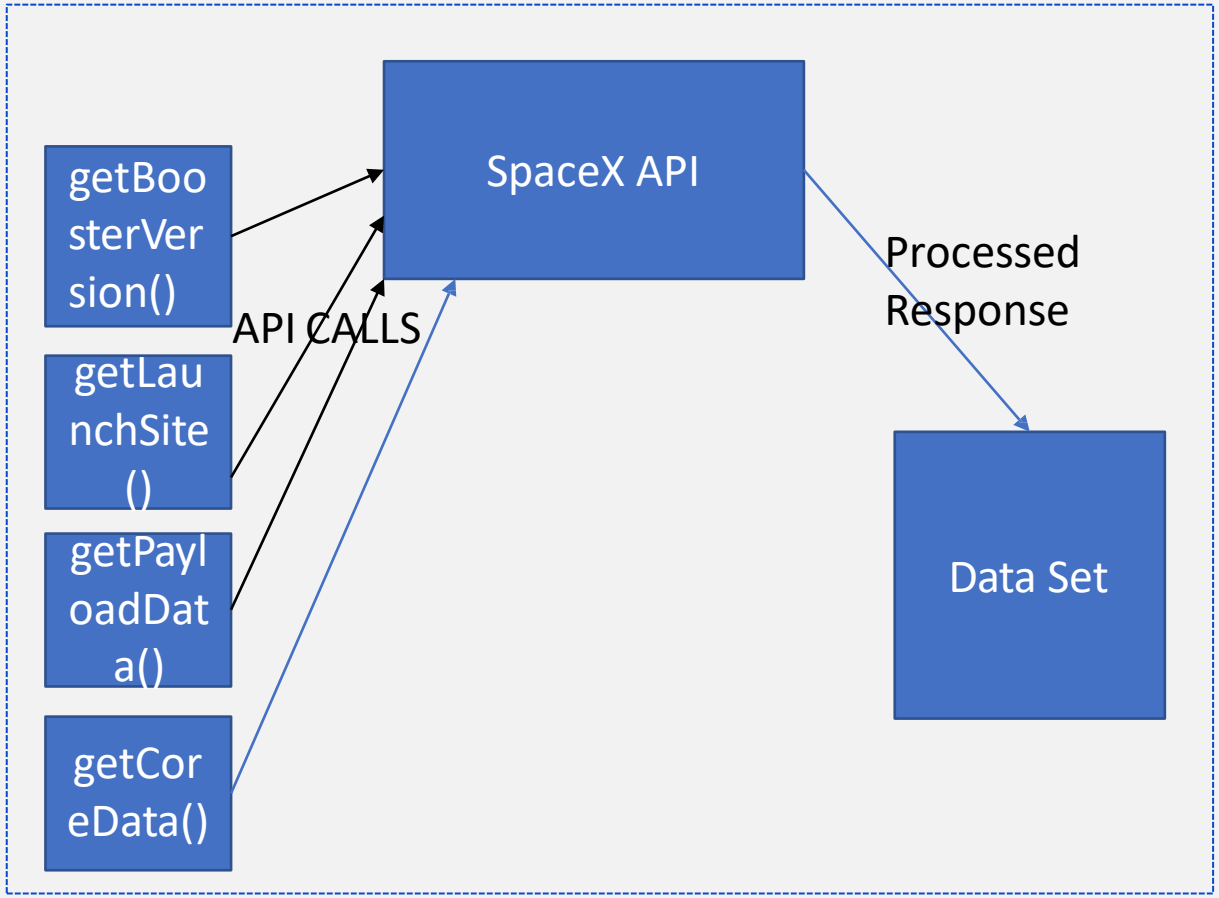
- This is the main phase in any Data Science task.
- We used two approaches for the data collection.

1. Using SpaceX's API
2. Web scraping of Wikipedia pages



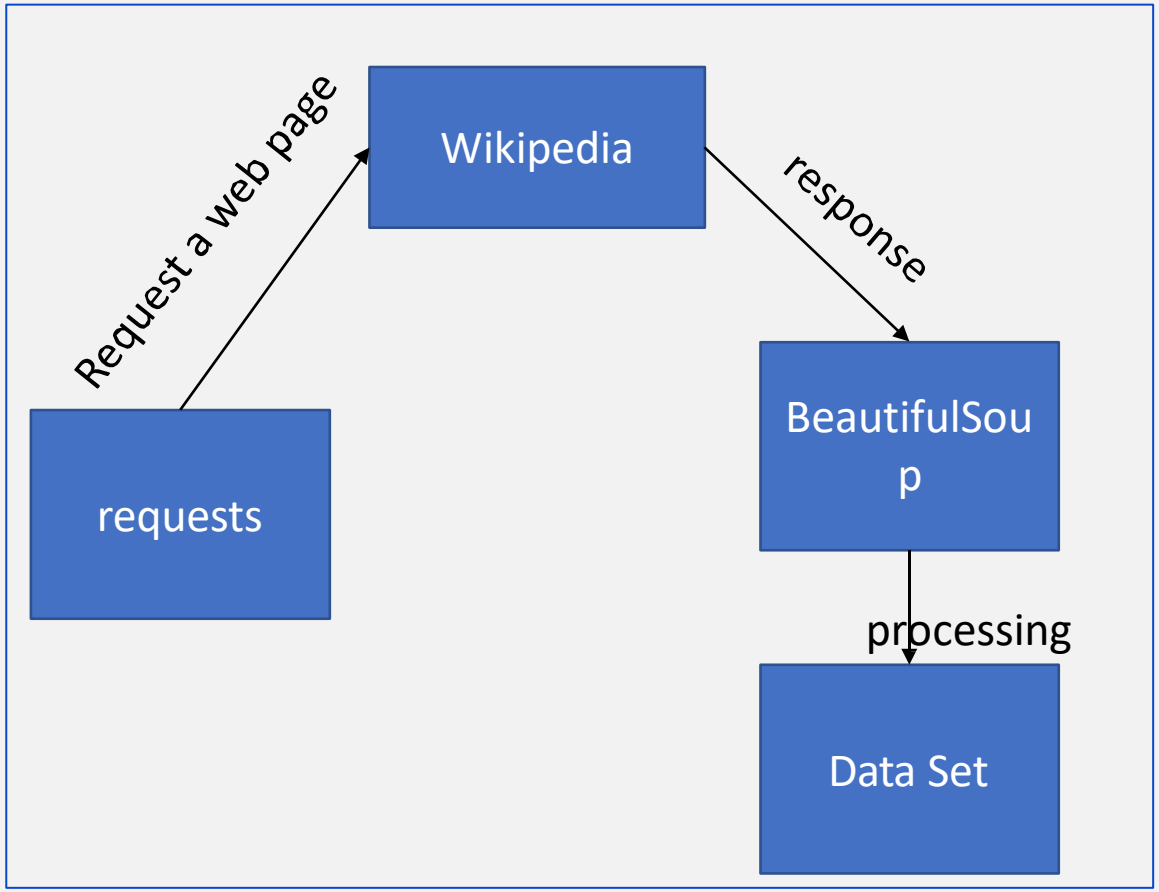
DATA COLLECTION – SPACEX API

- We used SpaceX's API to collect extra useful information like booster versions, launch site data, payload data, core data.
- We use rocket id, launch pad id, payload id's as parameters for our API calls.



DATA COLLECTION - SCRAPING

- We used Wikipedia tables to collect some crucial data about SpaceX's previous launches.
- We use BeautifulSoup, requests modules to do the scraping.



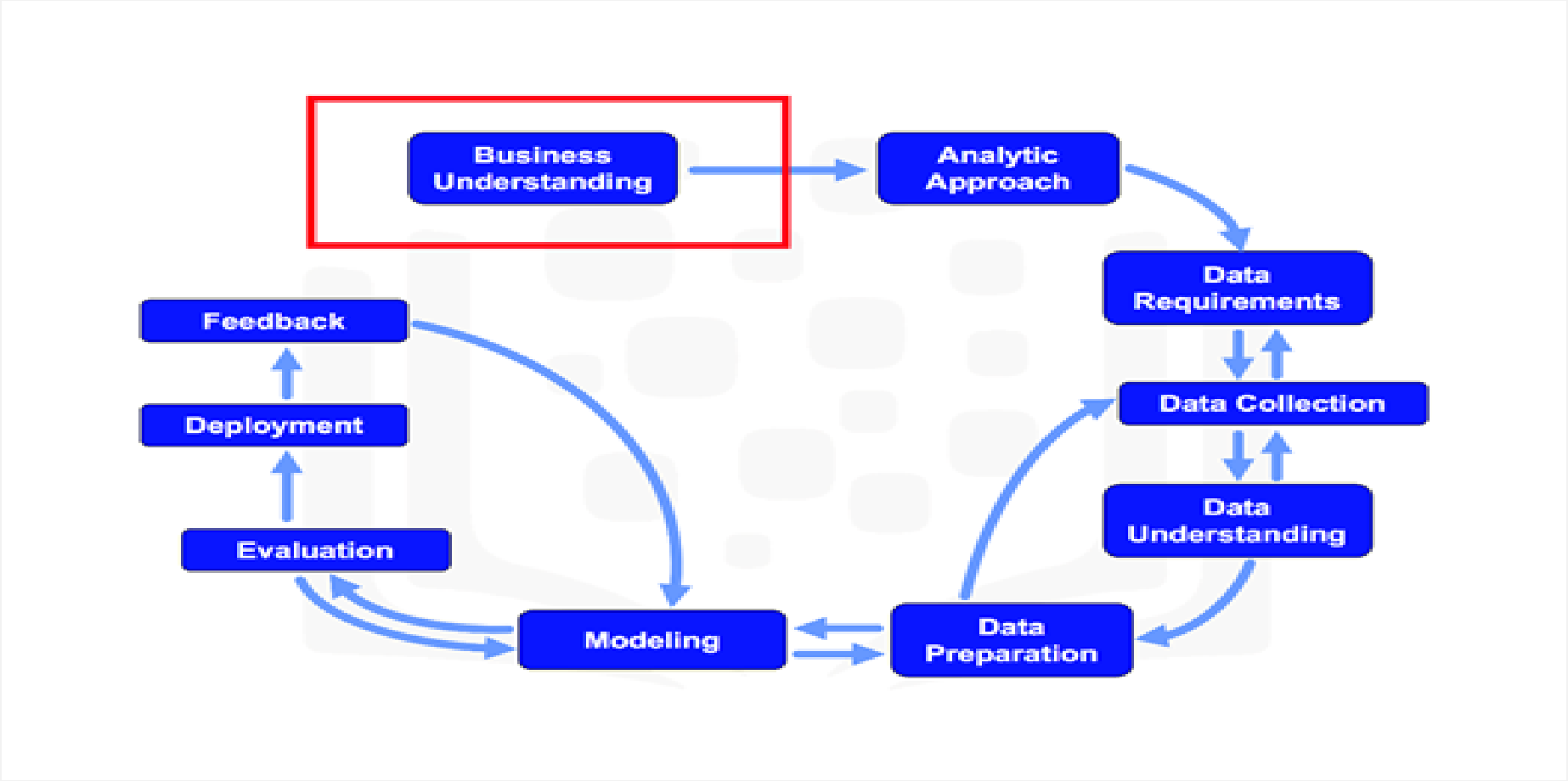
DATA WRANGLING

- The dataset obtained from data collection is unclean. We need to process it for further usage.
- Using Data Wrangling , we can make the data more machine readable and helpful.
- We found some missing values (Null/None) values in PayloadMass column. We replaced those missing values with the mean of the other PayloadMass entries.
- We replaced the false/none values of landing_outcomes column to 0 and others with 1.

DATA WRANGLING

- Data wrangling, a core data analysis technique is not done in one fell swoop–it's an iterative process that helps you get to the cleanest, most usable data possible prior to your analysis. Without data wrangling, the data set could be nearly impossible to sift through to find crucial insights. Each step in the data wrangling process exposes new potential ways that the data might be "re-wrangled," all driving towards the ultimate goal of generating the most robust data for final analysis.

METHODOLOGY

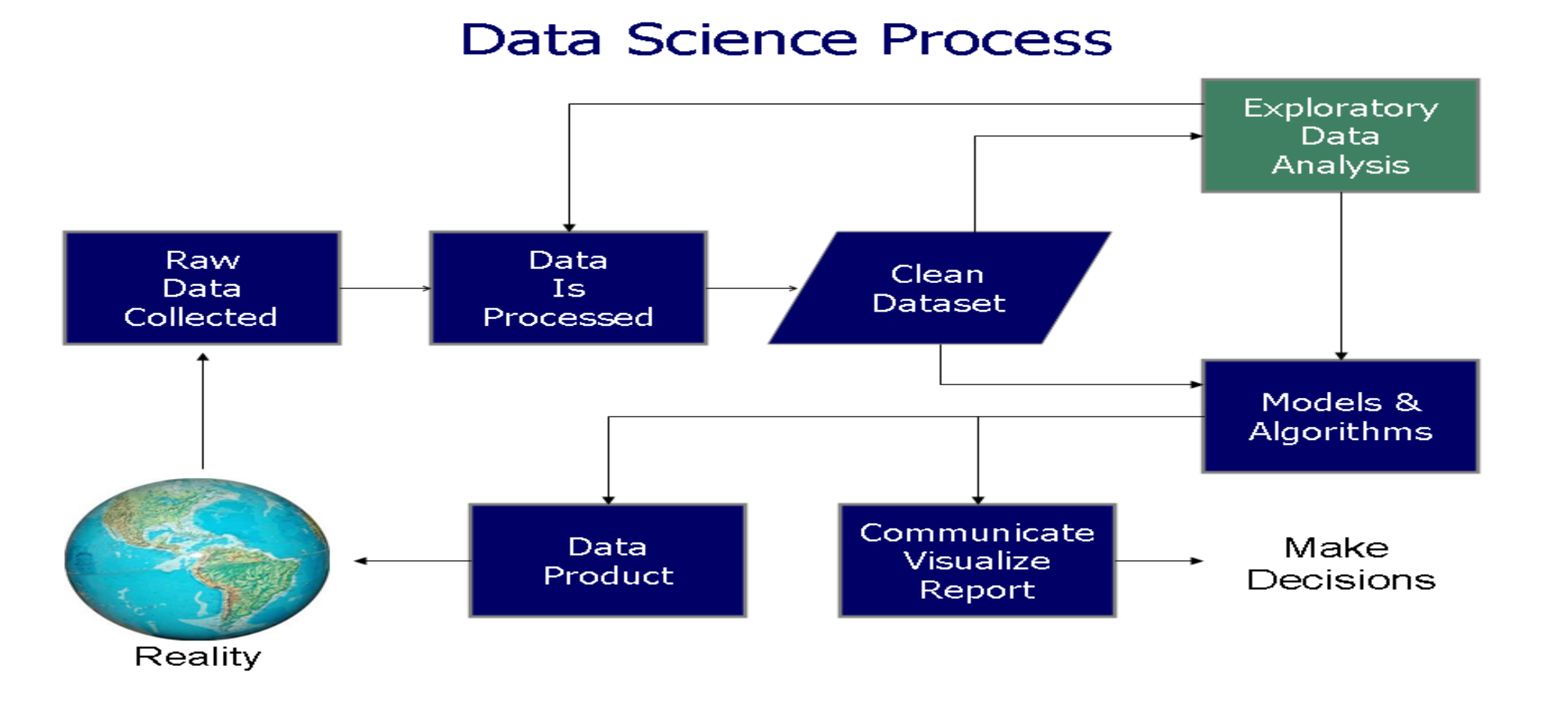


EDA WITH DATA VISUALIZATION

- During exploratory data analysis phase, we plotted different plots like

1. Cat plot of flight_numbers, payload mass to find some patterns.
2. Scatter plot of flight_numbers, launch_site to find any patterns.
3. Bar chart to find success rate of different orbits.
4. Scatter plot of payload and orbit.
5. Line plot for yearly success rate of total launches.

EDA AND INTERACTIVE VISUAL ANALYTICS METHODOLOGY



EDA WITH SQL

1. select UNIQUE(LAUNCH_SITE) from SPACEXDATASET -> find unique launch sites.
2. Select * from SPACEXDATASET WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5 -> to print first 5 records with launch site starting with 'CCA'
3. Select MAX(PAYLOAD_MASS__KG_) from SPACEXDATASET WHERE CUSTOMER='NASA (CRS)' -> to find max payload of nasa (crs).
4. Select AVG(PAYLOAD_MASS__KG_) from SPACEXDATASET WHERE BOOSTER_VERSION='F9 v1.1' -> find avg payload mass carried by booster F9 v1.1. Etc.

BUILD AN INTERACTIVE MAP WITH FOLIUM

- Initially , we plotted the 4 launch sites available in our dataset(CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E) based on their latitude and longitude.
- Next, we used markers_cluster to mark all success/failed launces in each site.
- Next, we tried to find the distance of these sites from cities, sea shores, railroads etc.
- This analysis can help in finding if a launch will be successful based on their proximities.

BUILD A DASHBOARD WITH PLOTLY DASH

- The dashboard is interactive based on launch site and payload mass.
- Based on launch site, we plotted a pie chart of their success rates.
- On scatter plot , we plotted how payload effects the success rate of a launch.
- These analysis helps to find the key features that can be used in model development in later phases.

PREDICTIVE ANALYSIS (CLASSIFICATION)

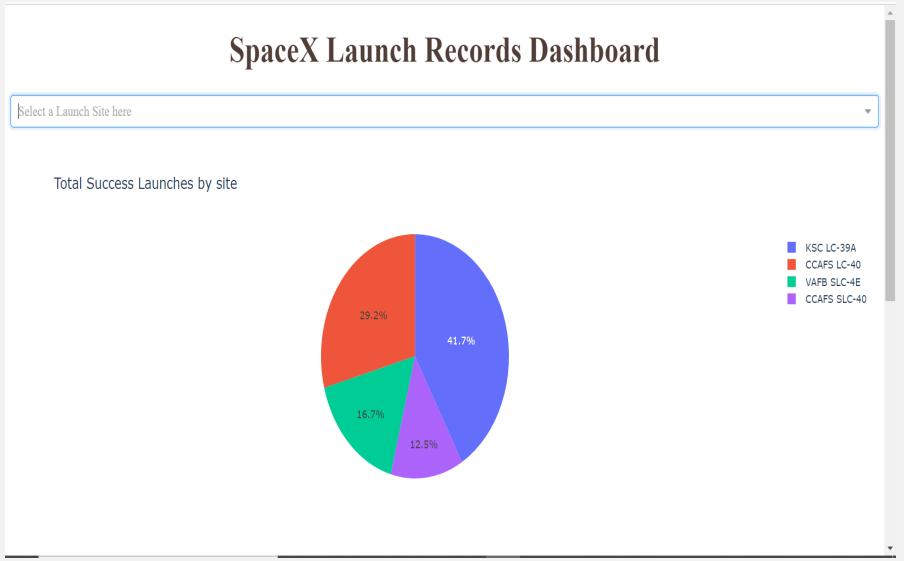
- We used logistic regression, SVC, Decision Tree, KNN as our models.
- First, we load the data, perform some standardization. Next we divide data in train and test sets.
- We use training set to iteratively find the best parameters for that data and finalized the parameters.
- Finally we find the best model of the four modes based on test set performance.

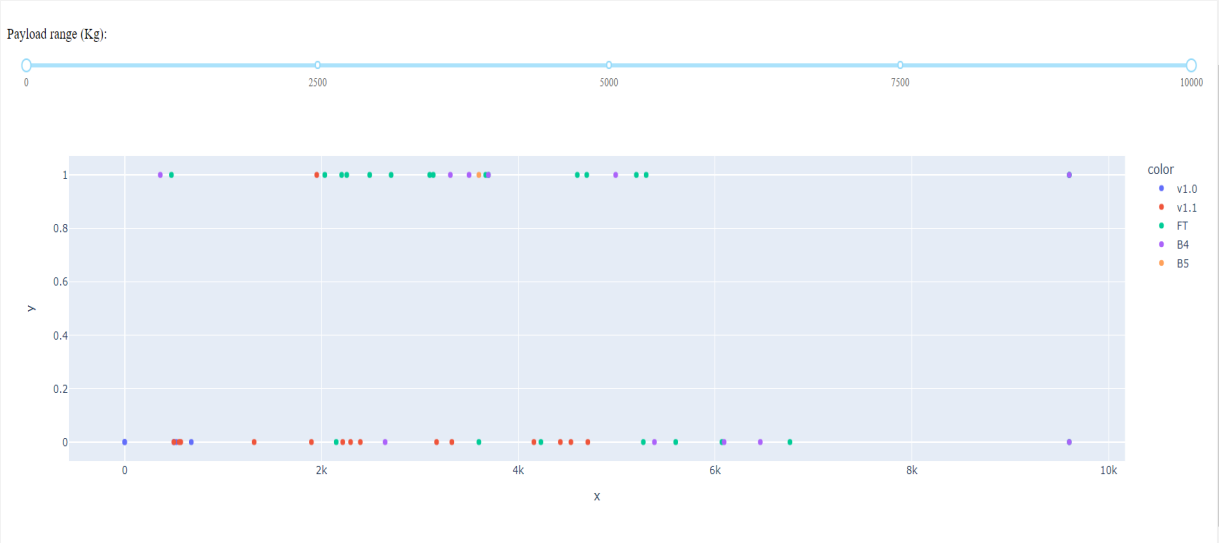
RESULTS

- Exploratory data analysis results

1. From 2013 there is significant increase in the success rate.
2. There is 99% of successful mission outcome (Not landing outcome)
3. KSC LC-39A Launch site has highest Successful landing rate.

- Interactive analytics demo in screenshots
- Predictive analysis results (test set)

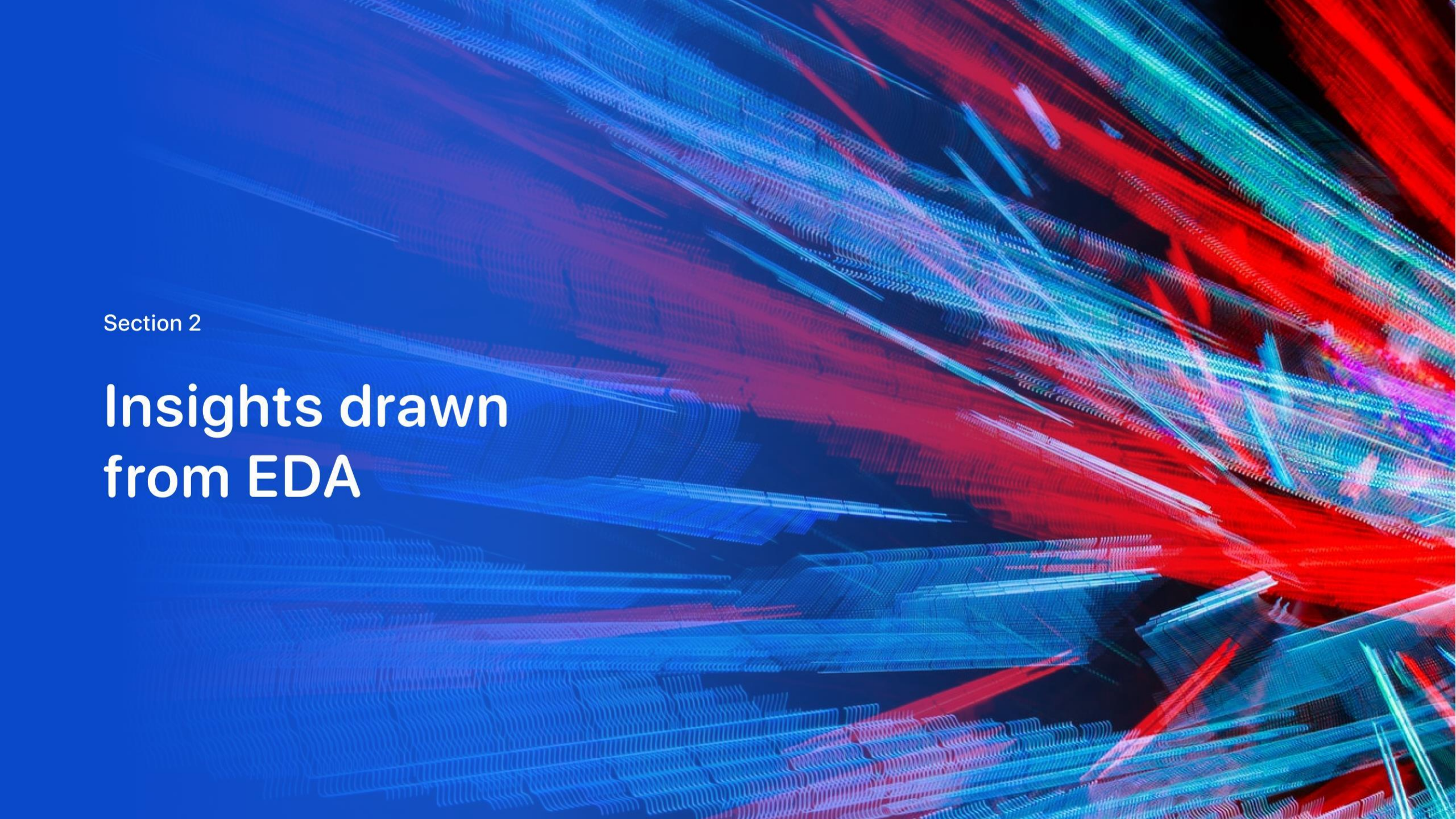




- Decision Tree 0.88
- Logistic regression 0.83
- SVM 0.83
- KNN 0.83

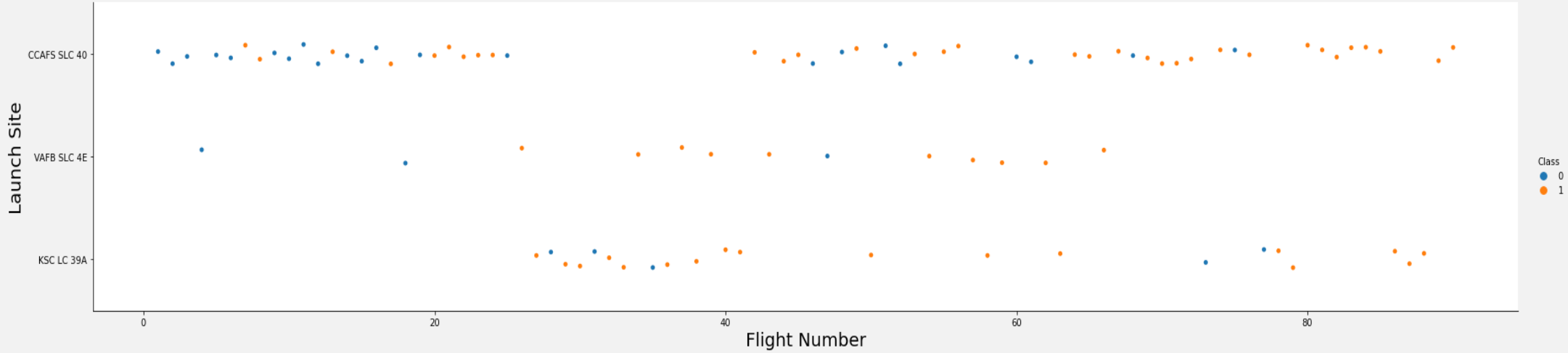
Section 2

Insights drawn from EDA



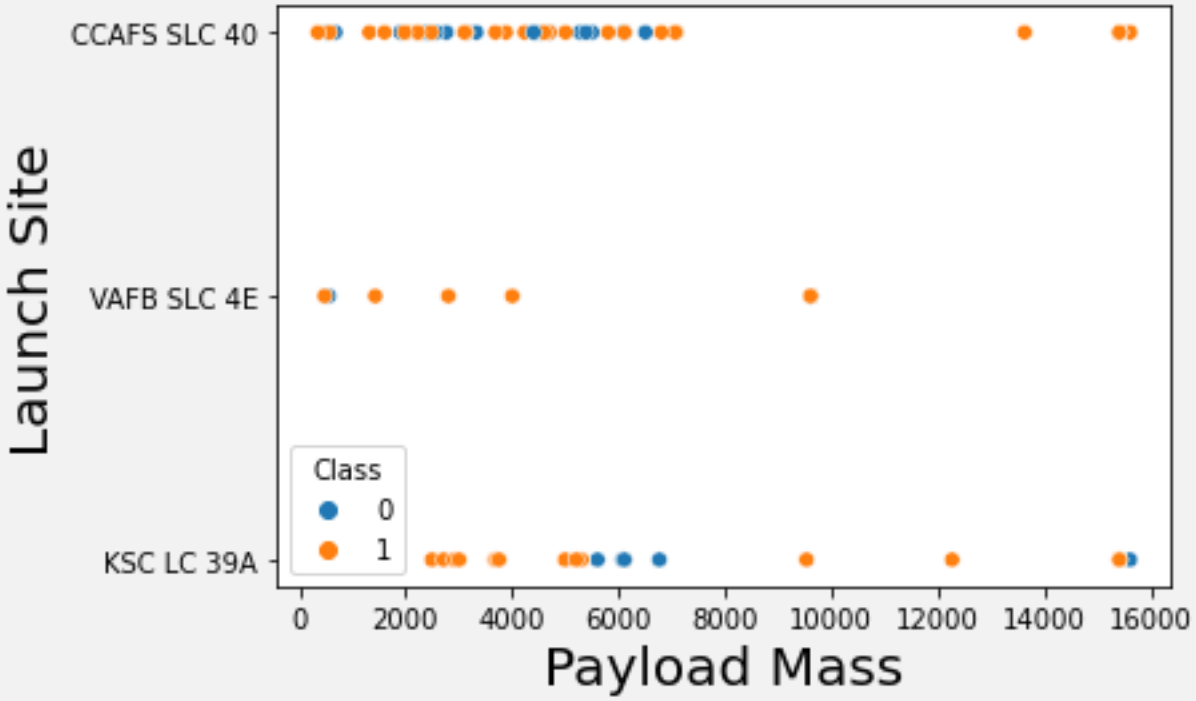
FLIGHT NUMBER VS. LAUNCH SITE

- We can see different launch site have different success rate.
- We can see that KSC LC 39A has 77% success rate.
- CCAFS LC-40 has 60 % success rate.
- VAFB SLC 4E has 76% success rate.



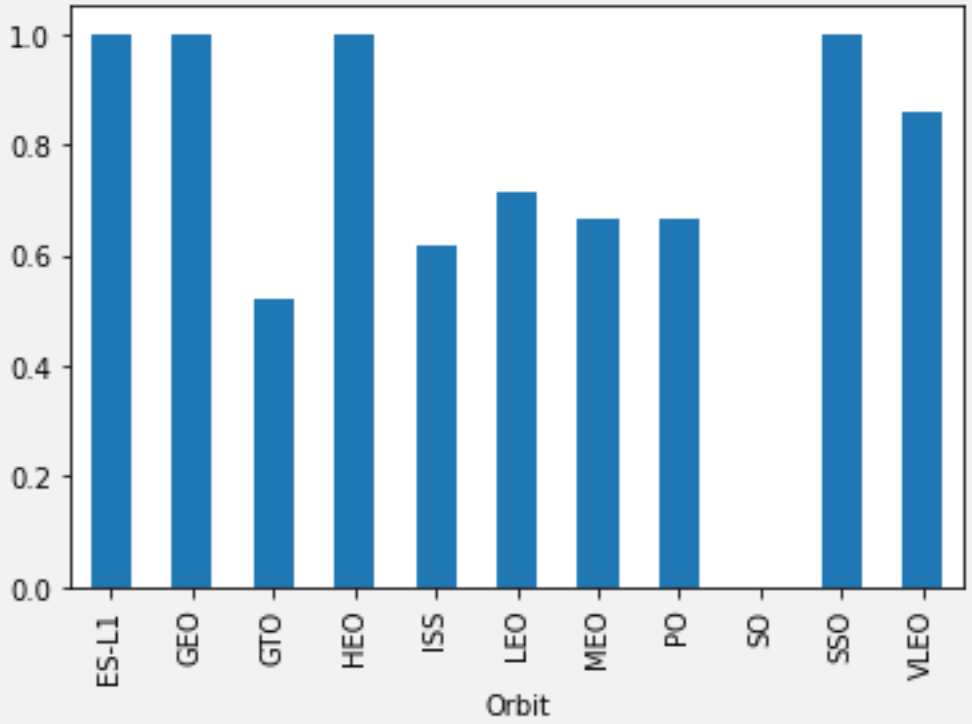
PAYLOAD VS. LAUNCH SITE

- We can see that payload greater that 8000kg has almost 100% success rate.
- While launches less than 8000kg payload has 50-60 % success rate.



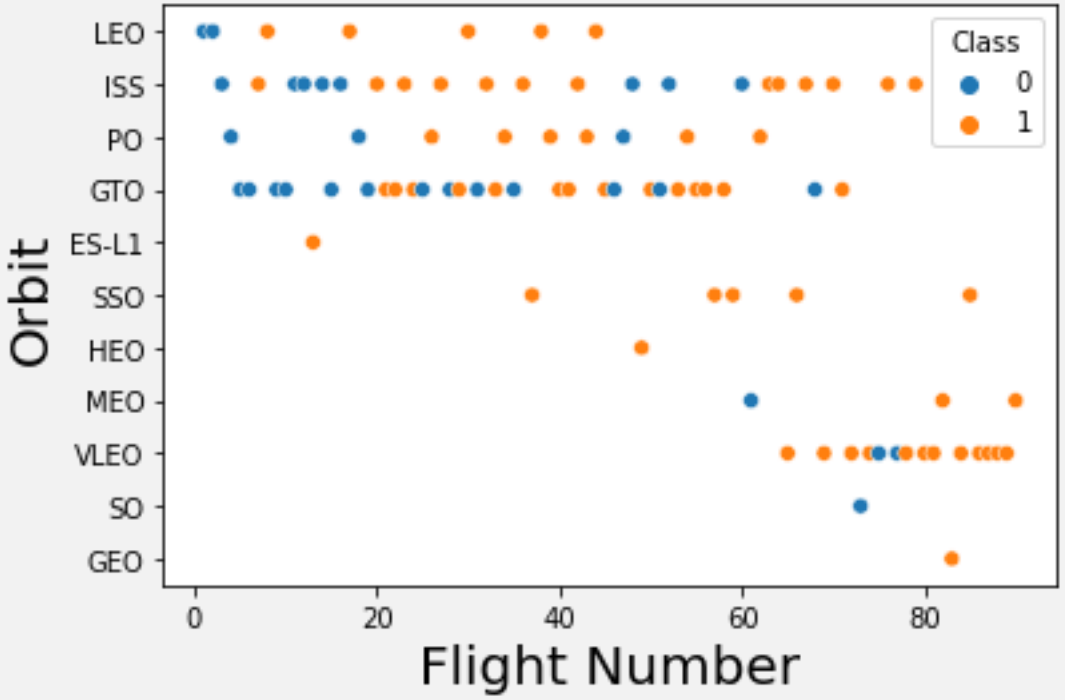
SUCCESS RATE VS. ORBIT TYPE

- We can see that ‘SO’ orbit has 0% success rate.
- While ES-L1, GEO, HEO, SSO have 100% success rate.



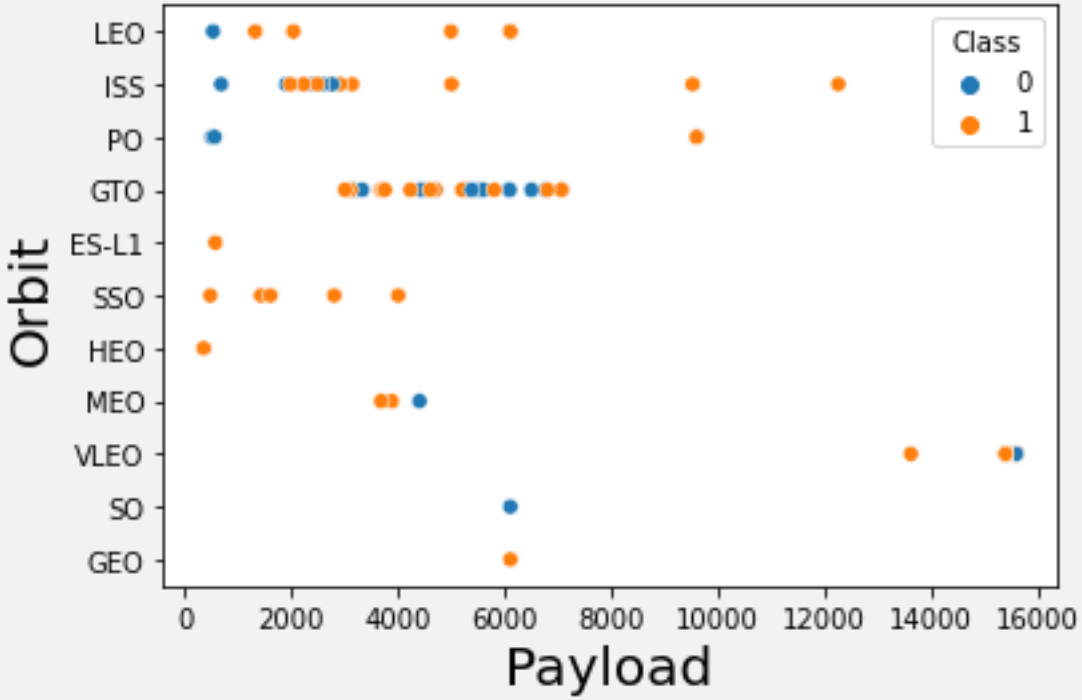
FLIGHT NUMBER VS. ORBIT TYPE

- We can see that flight numbers above 75 are all successful.



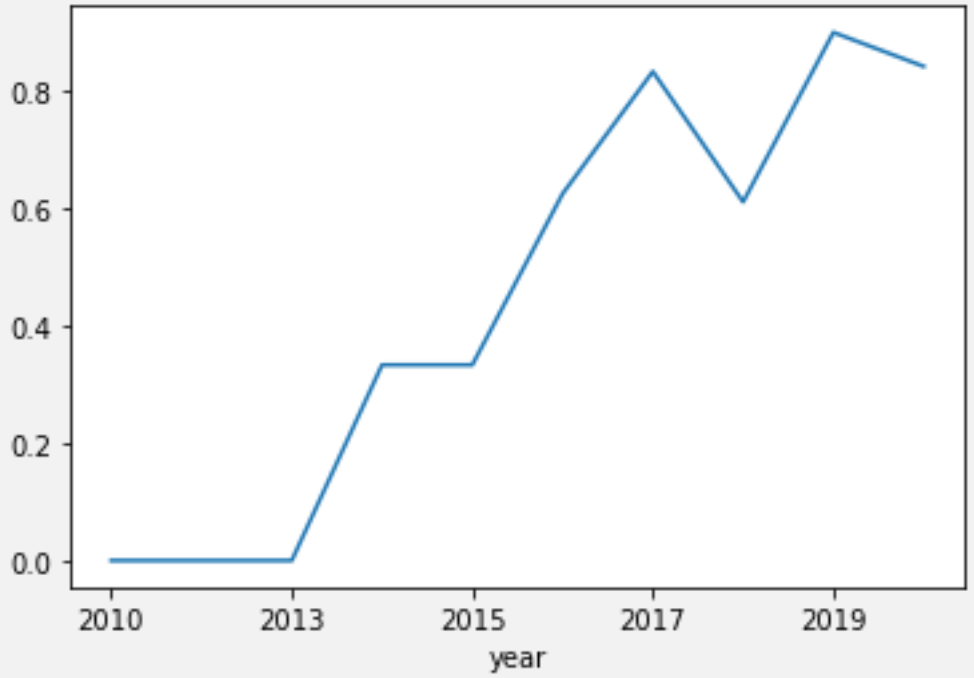
PAYLOAD VS. ORBIT TYPE

- We can see that orbits ISS, PO with payload greater than 8000kg are 100 successful.
- SSO is 100% successful with all payloads.
- HEO, GEO have only one launch as of now.



LAUNCH SUCCESS YEARLY TREND

- We can see that from 2013 , success rate has increased significantly.
- Highest success rate is in 2019.



ALL LAUNCH SITE NAMES

- Find the names of the unique launch sites

-> select UNIQUE(LAUNCH_SITE) from SPACEXDATASET

- There are four unique launch sites.

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

LAUNCH SITE NAMES BEGIN WITH 'CCA'

- Find 5 records where launch sites begin with `CCA`

-> Select * from SPACEXDATASET WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5

- 4/5 of these launches are for nasa.

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

TOTAL PAYLOAD MASS

- Calculate the total payload carried by boosters from NASA

-> Select MAX(PAYLOAD_MASS__KG_) from SPACEXDATASET WHERE CUSTOMER='NASA (CRS)

- MAX Payload carried by SPACEX rockets for nasa is 3310 kg.

AVERAGE PAYLOAD MASS BY F9 V1.1

- Calculate the average payload mass carried by booster version F9 v1.1

-> select AVG(PAYLOAD_MASS__KG_) from SPACEXDATASET WHERE BOOSTER_VERSION='F9 v1.1'

- Average payload carried by F9 v1.1 is 2928KG

FIRST SUCCESSFUL GROUND LANDING DATE

- Find the dates of the first successful landing outcome on ground pad

-> select MIN(DATE) from SPACEXDATASET WHERE LANDING__OUTCOME='Success (ground pad)'

- The successful landing on ground pad occurred on 2015- 12-22.

SUCCESSFUL DRONE SHIP LANDING WITH PAYLOAD BETWEEN 4000 AND 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

-> select BOOSTER_VERSION from SPACEXDATASET WHERE LANDING__OUTCOME = 'Success (drone ship)' and PAYLOAD_MASS__KG_>4000 and PAYLOAD_MASS__KG_<6000

- There are 4 Booster version

Out[24]:

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

TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSION OUTCOMES

- Calculate the total number of successful and failure mission outcomes
- select COUNT(*) from SPACEXDATASET WHERE MISSION_OUTCOME LIKE 'Success%'
- select COUNT(*) from SPACEXDATASET WHERE MISSION_OUTCOME LIKE 'Failure%’
- Almost 100% mission outcome are success.

BOOSTERS CARRIED MAXIMUM PAYLOAD

- List the names of the booster which have carried the maximum payload mass

-> select BOOSTER_VERSION from SPACEXDATASET where PAYLOAD_MASS__KG_=(select MAX(PAYLOAD_MASS__KG_) from SPACEXDATASET)

- There are 12 booster versions.

Out[28]:

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 failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

-> SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE from SPACEXDATASET WHERE LANDING__OUTCOME LIKE 'Failure (drone ship)' and DATE LIKE '2015%'

- There are two booster in ccafs lc-40 site.

Out[31]:

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
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 landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

-> select LANDING__OUTCOME,COUNT(LANDING__OUTCOME) as cou from SPACEXDATASET GROUP BY LANDING__OUTCOME ORDER BY cou DESC

- There are 10 types of landing outcomes.

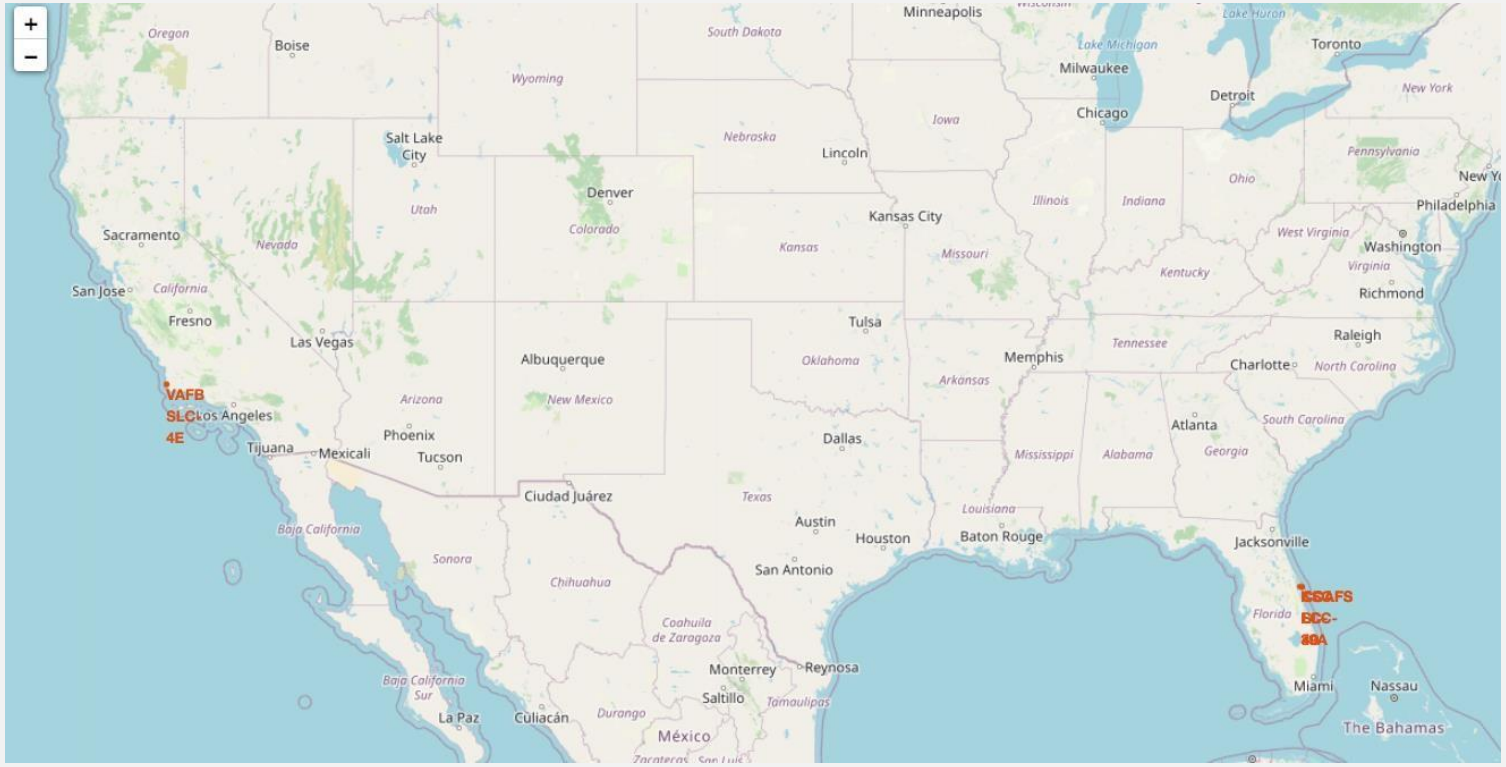
Out[39]:

landing_outcome	cou
Success	38
No attempt	22
Success (drone ship)	14
Success (ground pad)	9
Controlled (ocean)	5
Failure (drone ship)	5
Failure	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

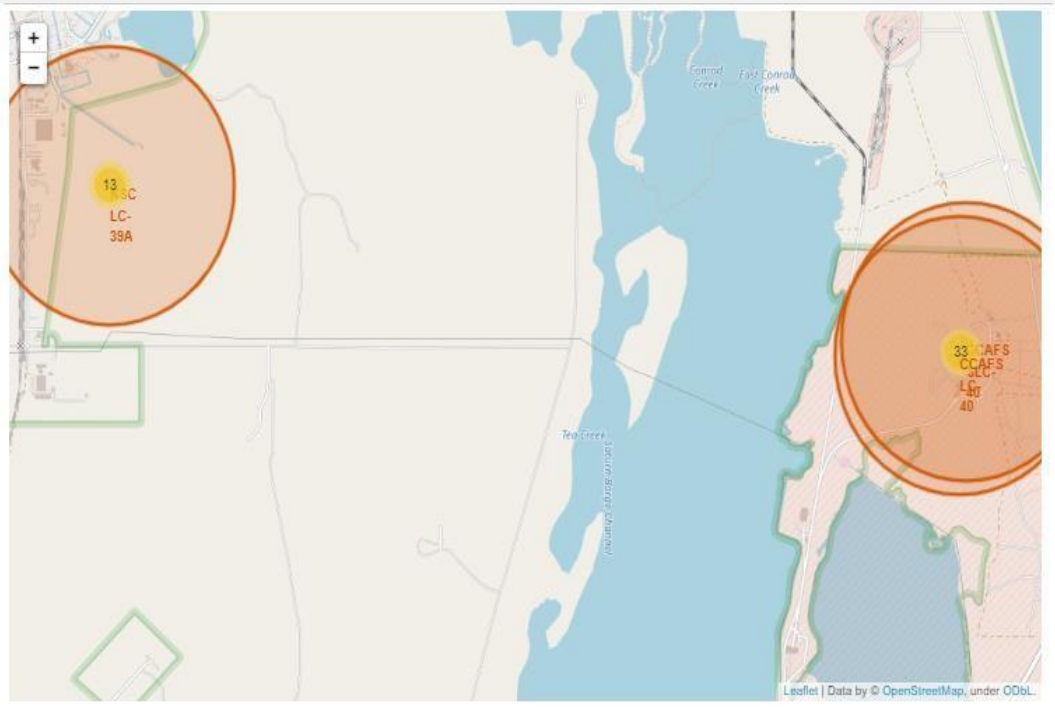

Section 4
Launch Sites
Proximities Analysis

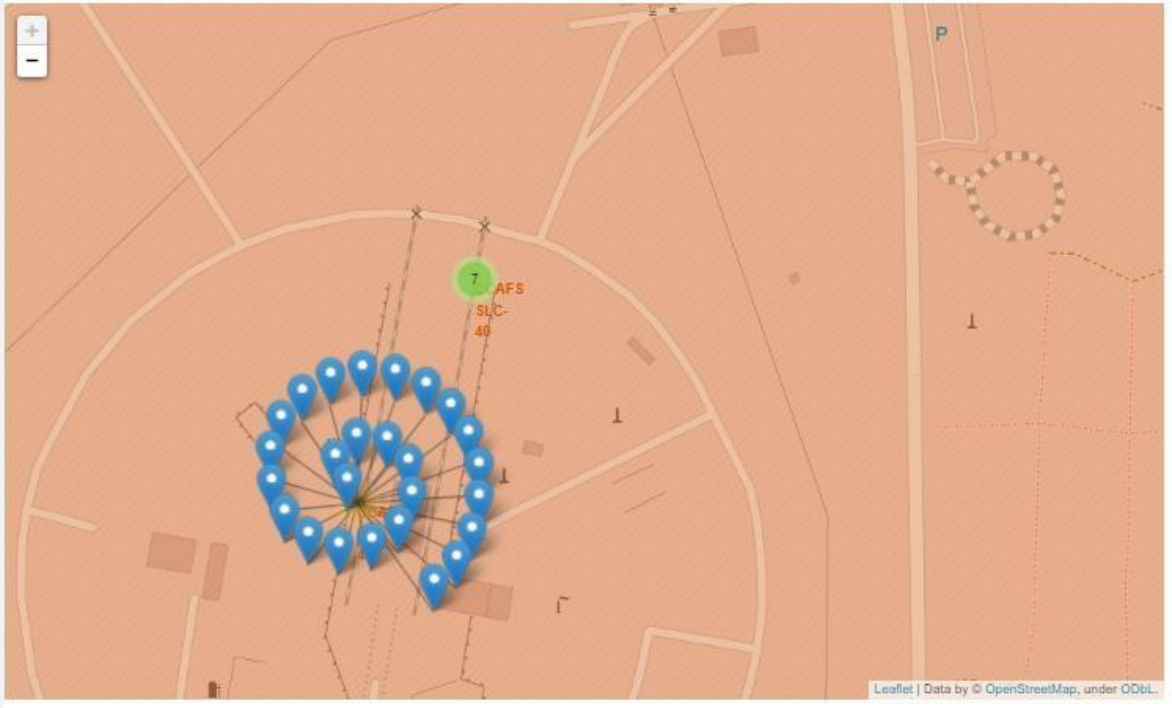
LAUNCH SITES OF SPACEX

- We can see that all the launch site are close to ocean and away from cities.



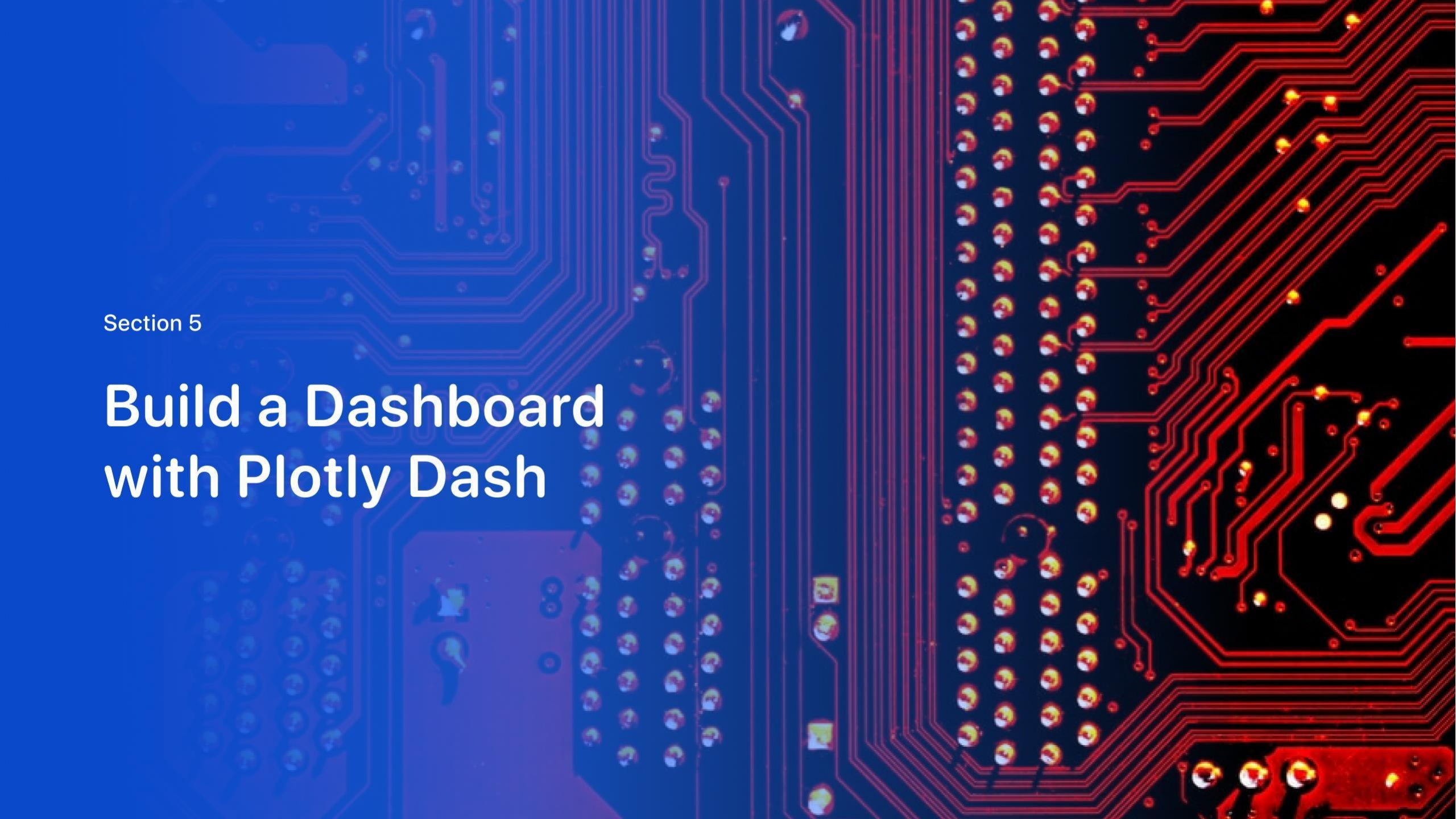
DIFFERENT LAUNCH LOCATIONS CLUSTERS.





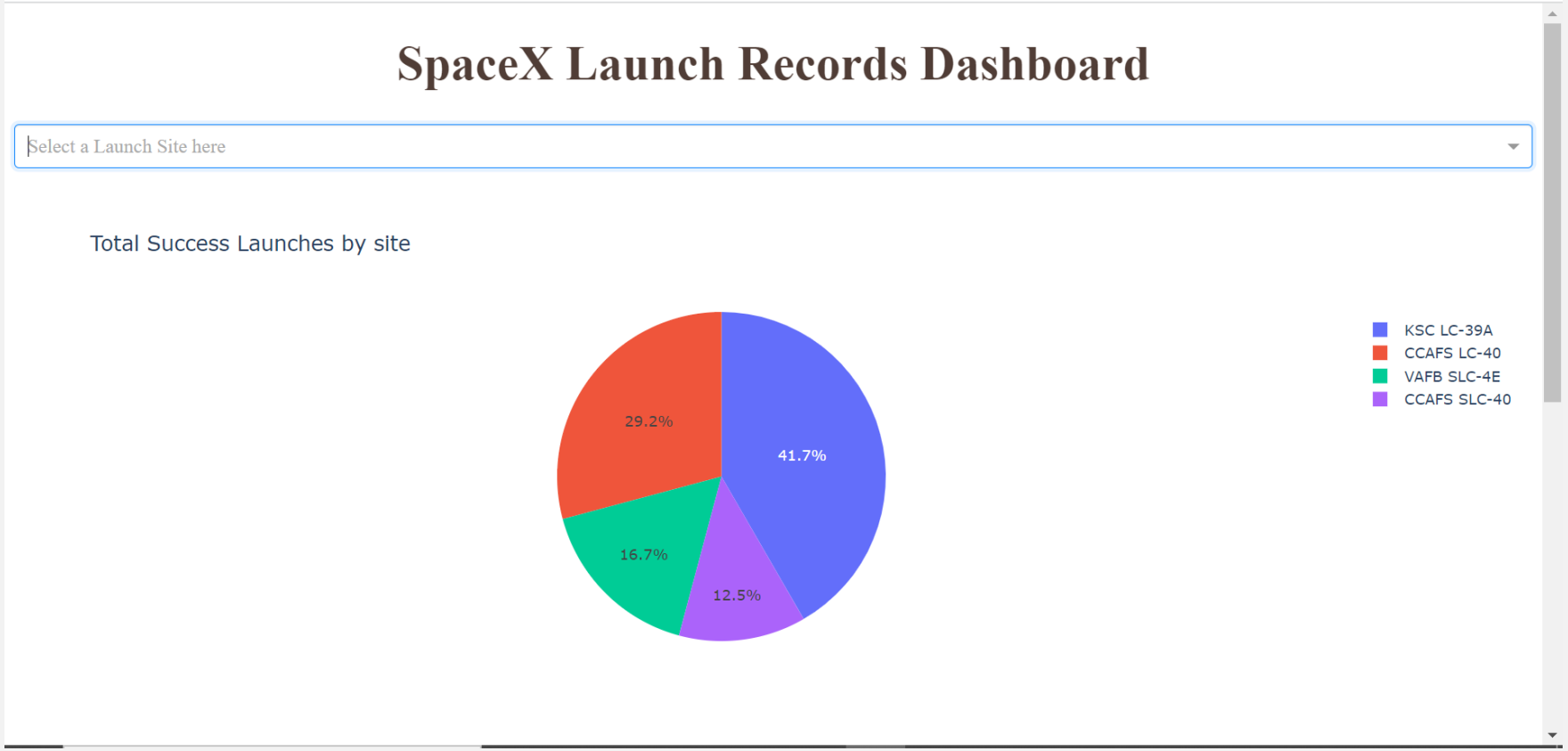
<Folium Map Screenshot 3>

Nearest highway is around 0.6 km away.


Section 5
Build a Dashboard with Plotly Dash

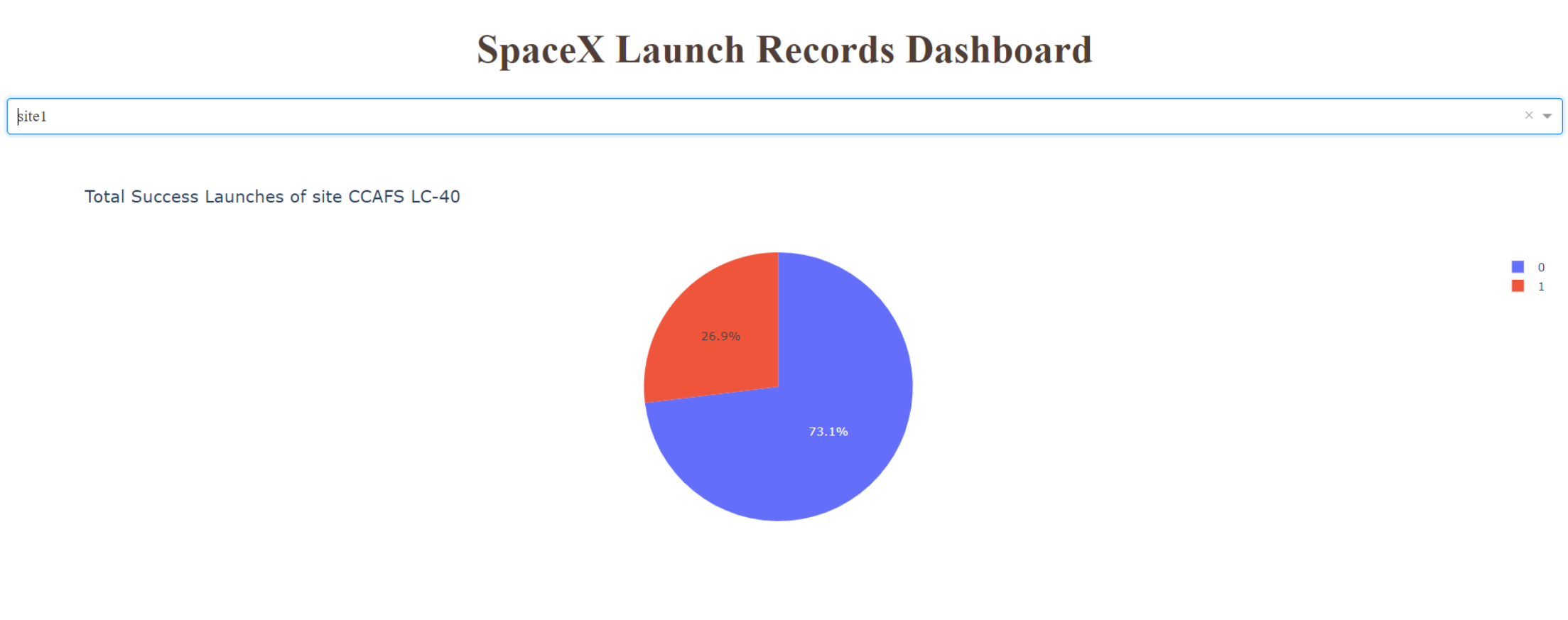
Pie chart for all launch sites.

KSC LC-39A has more number of successful launches.

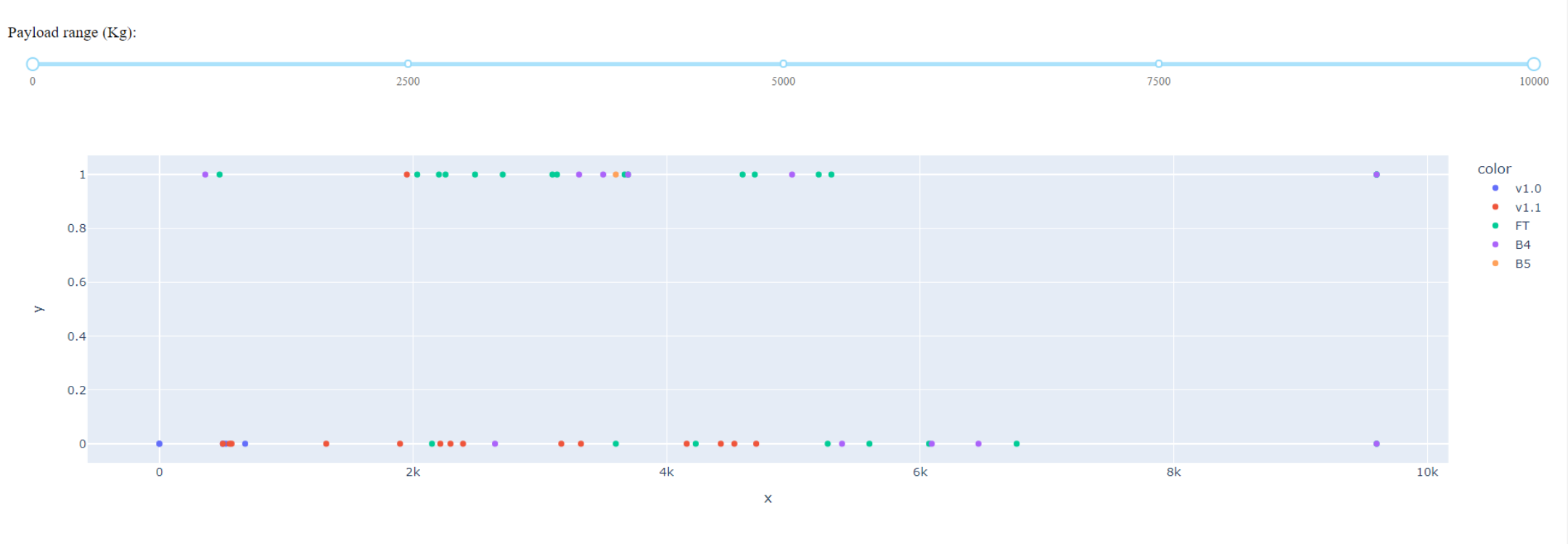


Pie chart for CCAFS LC-40 Launch site

73 % of launches are successful in this site.



PAYLOAD VS LANDING OUTCOME

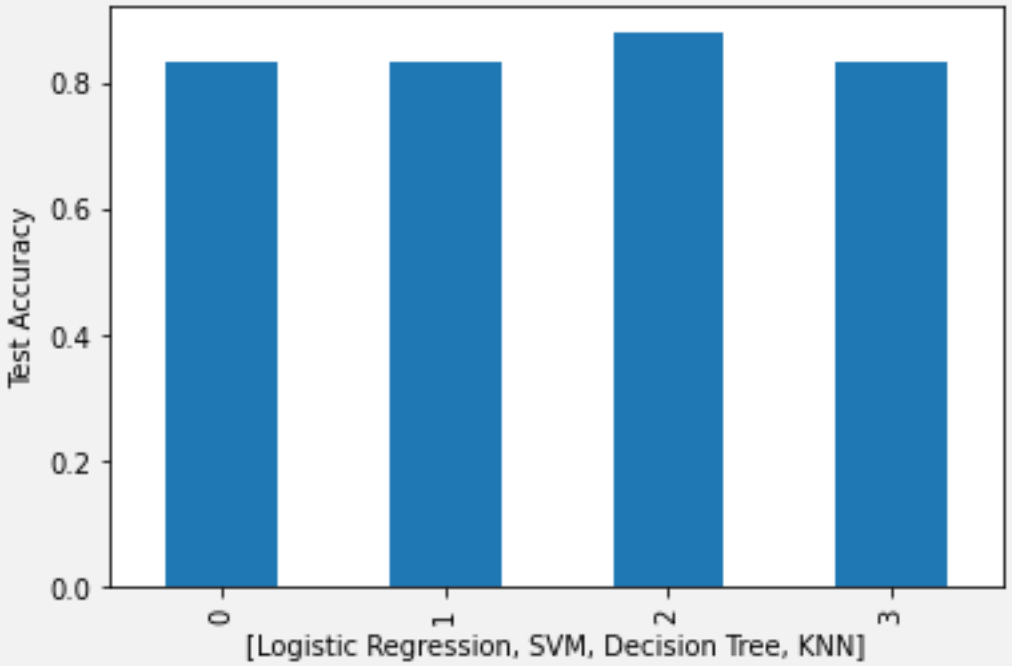


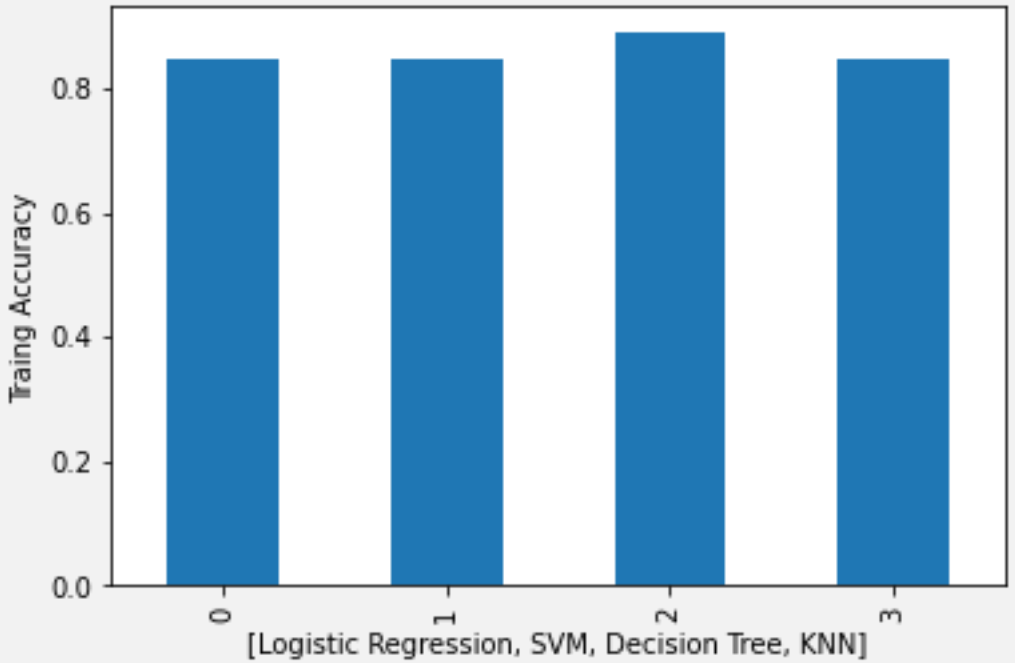
Section 6

Predictive Analysis (Classification)

CLASSIFICATION ACCURACY

- We can see that Decision Tree has the best training accuracy(0.89) and Test accuracy(0.88).





CONFUSION MATRIX

•We can see that 12/12 are correctly predicted for Positive class.

• 3/6 are correctly predicted for negative class.

CONCLUSIONS

- From this project, we conclude that Decision Tree with gini as criterion , max_depth of 8, Auto as max features, random splitter, min_samples_leaf =4 and min_samples_split=10 gives the best prediction for our application.
- With more data and good generalization of the model, we can improves the performance further.

APPENDIX

- Dataset:

Out[3]:

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	La
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.5
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.5
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.5
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610829	34.6
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28.5

Converting categorical columns to one hot vectors.

Out[43]:

	PayloadMass	Flights	Block	ReusedCount	ES-L1	GEO	GTO	HEO	ISS	LEO	...	B1048	B1049	B1050	B1051	B1054	B1056	B1058	B1059	B1060	B1062
0	6104.959412	1.0	1.0	0.0	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	525.000000	1.0	1.0	0.0	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	677.000000	1.0	1.0	0.0	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	500.000000	1.0	1.0	0.0	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	3170.000000	1.0	1.0	0.0	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

5 rows × 76 columns

- https://colab.research.google.com/drive/1f3eaq-sQTAcxGLtOu_O0c6pUyA-zBCwA
- https://courseraassessments.s3.amazonaws.com/assessments/1636885531447/6d05cd9b-7119-44c2-b2d5-af7dfbe21892/6th.pdf
- https://eu-de.dataplatform.cloud.ibm.com/analytics/notebooks/v2/f796ea32-abea-454e-a2ef-79bed6448d05/view?access_token=c96b9a11bee88b9a51dd2d4c7f70aed86a8463b95e87a0b3b2bfb94de808f20e
- https://colab.research.google.com/drive/1ZbS1YUOkr1-jm7-IUQYKHebYNeRS7vdF#scrollTo=053600XdCy8f