Winning Space Race with Data Science

IBM Data Science Capstone project SpaceX

Dilapsky

2021/11/20

Outline

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

Executive Summary

Summary of methodologies

- Data Collection
- Data Wrangling
- EDA Data Visualization
- SQL Enquiry
- Create Visualize Map using folium
- Predictive Analysis (Clustering, KNN and etc)

Summary of all results

- Data is analyzed in the data collection
- Data is visualized and explored using EDA, SQL and map Folium
- An interactive Dash Board is created
- The predictive models performed very well in prediction

Introduction

Project background

- We are predicting whether the first stage of the SpaceX Falcon 9 rocket will land successfully. We will use the dataset from SpaceX API, and use our data science method to do the data analysis.

Project Mission

- Know How & Knowledge to adjust from lesson learn in case of improve the future success .
- Related key to the successful launches.
- Find out the variable that depend upon of outcome of launches.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
- Using REST API and web scraping methods
- Perform data wrangling
- Dealing with null and missing values, explore data types and standardization
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
- Linear Regression, SVM, Decision Tree and KNN models, model evaluation

Data Collection – SpaceX API

- For this project we will use Rest API to get data from defined URL
- This API get the data of rocket such as lunch time. Load and etc
- This data is used for predict the rocket attempt to land or not
- End point is https://api.spacexdata.com/v4/
- (Core, Rocket, Lunchpad and PayLoadmass)
- We also request information from Wikipedia via web wrapping by Beautifulsoup

Request Data
Space X API
Web Scraping

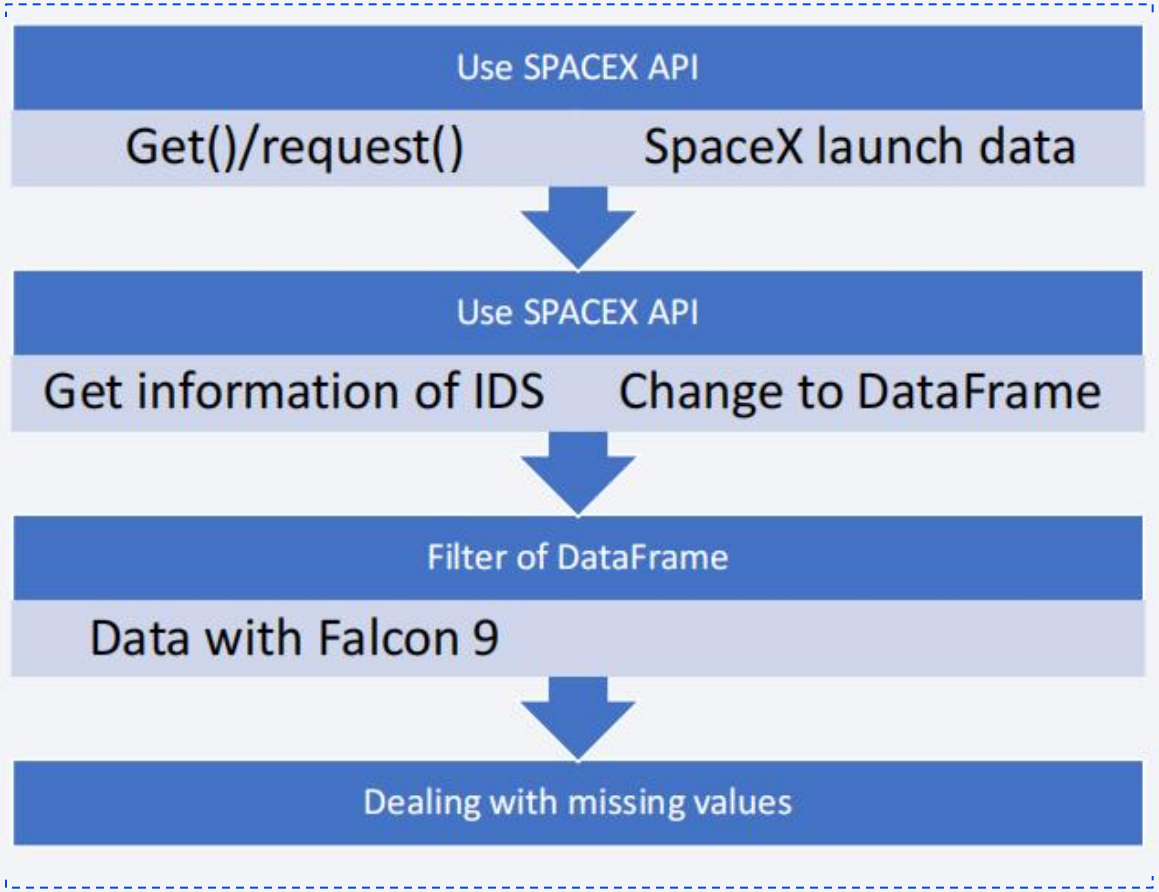
Store Data
Get/request function
JSON Objects

Normalization
Transform JSON to dataframe

Data Wrangling
Sampling data
Dealing with nulls

Data Collection – SpaceX API

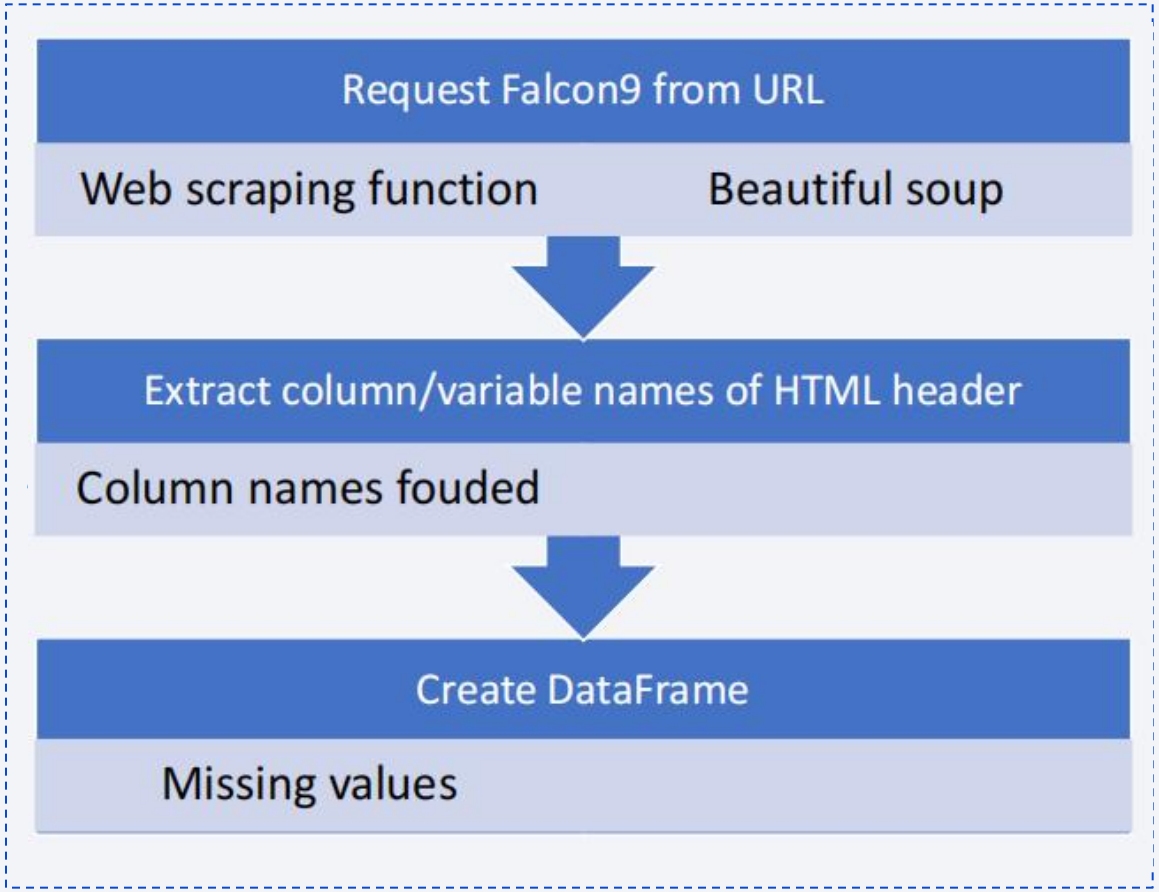
Flowchart of SpaceX API calls



Link: https://github.com/Dilapsky/IBM-Data-Science-Professional-Certificate/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

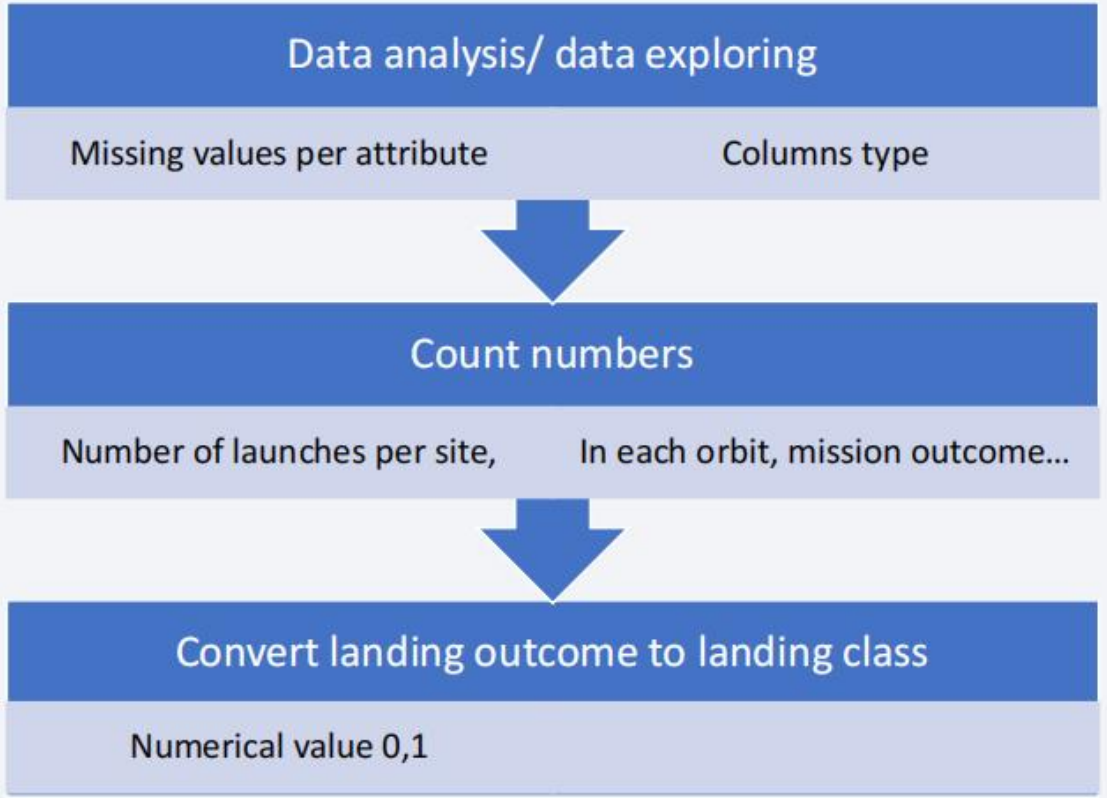
Flowcharts of web scraping process



Link: https://github.com/Dilapsky/IBM-Data-Science-Professional-Certificate/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

Flowcharts of data wrangling process



Link: https://github.com/Dilapsky/IBM-Data-Science-Professional-Certificate/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

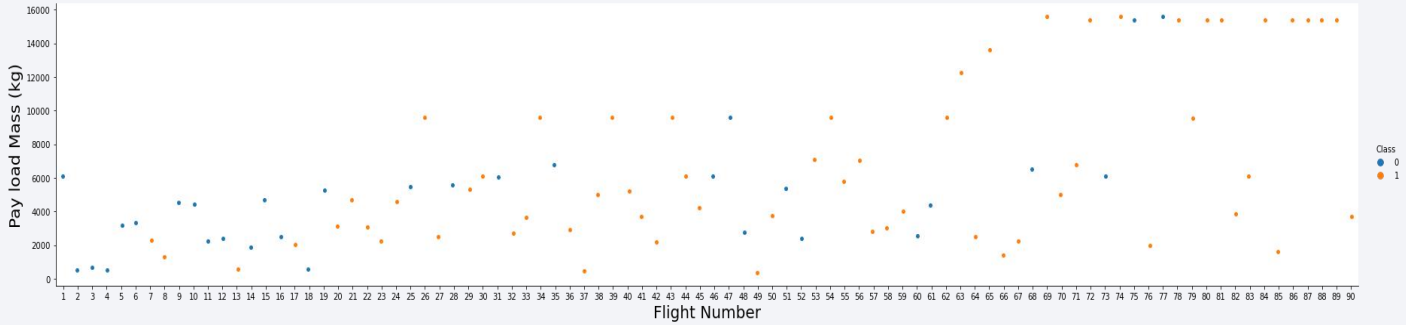
EDA with Data Visualization

- Scatter plot is firstly used to analyze the data, to get a quick image of the data, in followed pairs: FlightNumber vs. PayloadMass, FlightNumber vs. LaunchSite, PayloadMass vs. LaunchSite, lightNumber vs. Orbit type, Payload vs. Orbit type.
- Bar chart is used to show the relationship between success rate of each orbit type, line chart is used to show the launch success yearly trend, they are very obvious and easy to use.
- Using dummy variables to convert categorical columns to numerical columns, prepare for the machine learning model.

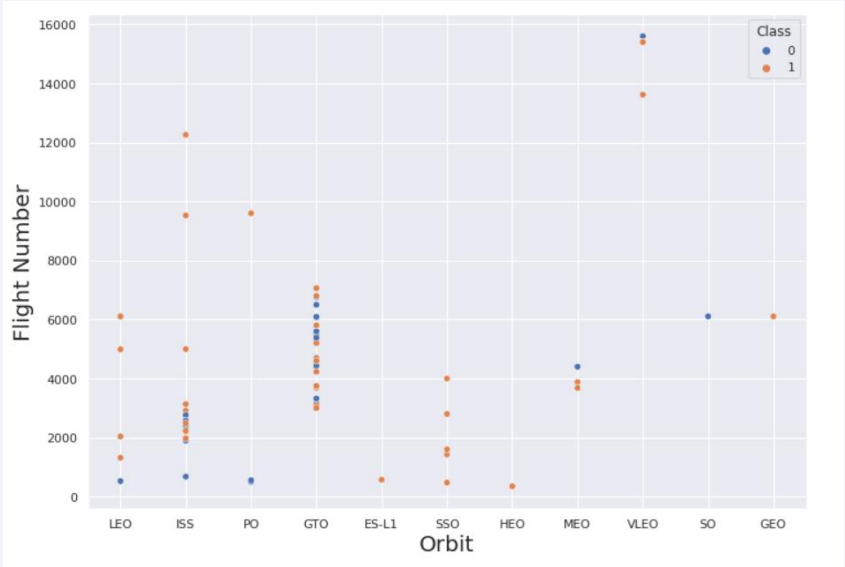
Link: https://github.com/Dilapsky/IBM-Data-Science-Professional-Certificate/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with Data Visualization

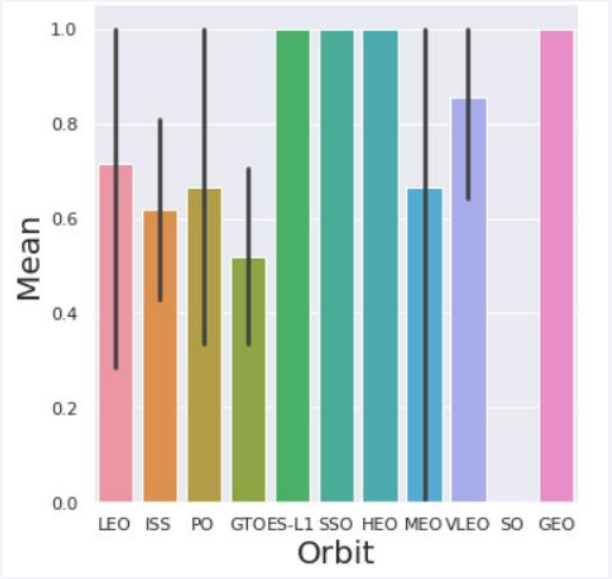
Categorical plot



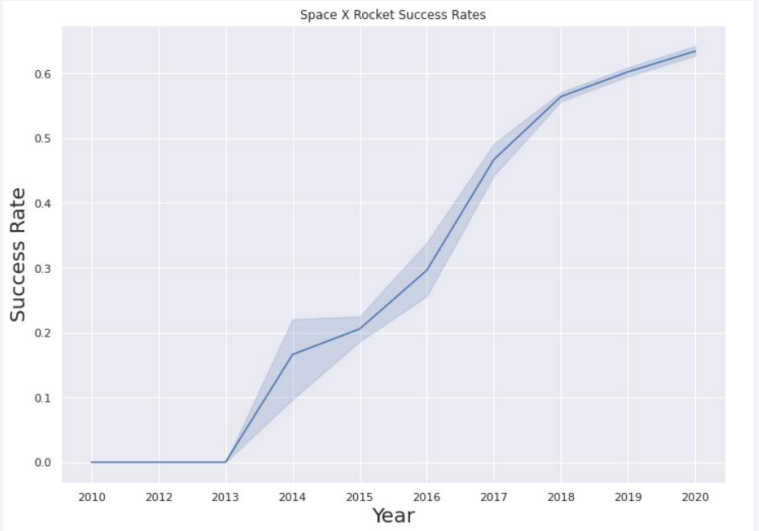
Scatter plot



Bar chart



Trend plot



EDA with SQL

- SQL queries like: SELECT, WHERE, LIKE, LIMIT, SUM, AVG, COUNT, MIN, MAX and DATE were performed to explore and analyze the data.
- Example:

```
"select DISTINCT Launch_Site from spacextbl "
```

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

Link: https://github.com/Dilapsky/IBM-Data-Science-Professional-Certificate/blob/main/jupyter-labs-eda-sql-coursera.ipynb

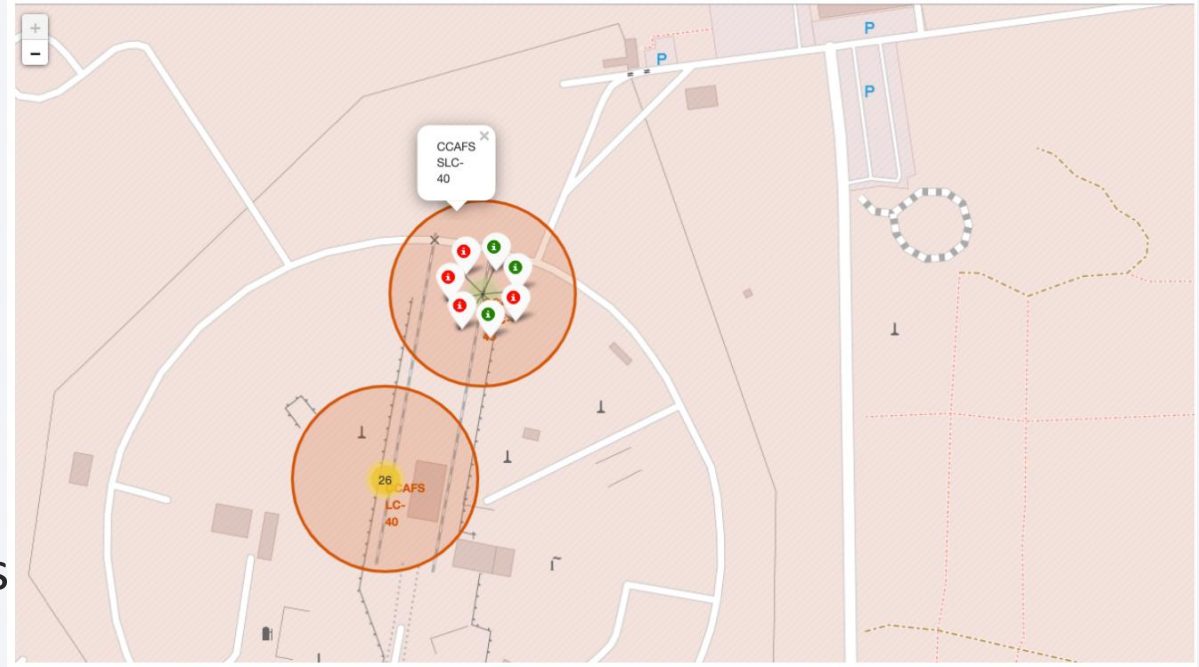
Build an Interactive Map with Folium

- The interactive map is developed in following steps:

1. marked the launch site with circle marker, obvious and direct on the map

2. mark the success/failed launches for each site on the map, get a quick image of success numbers and success rate

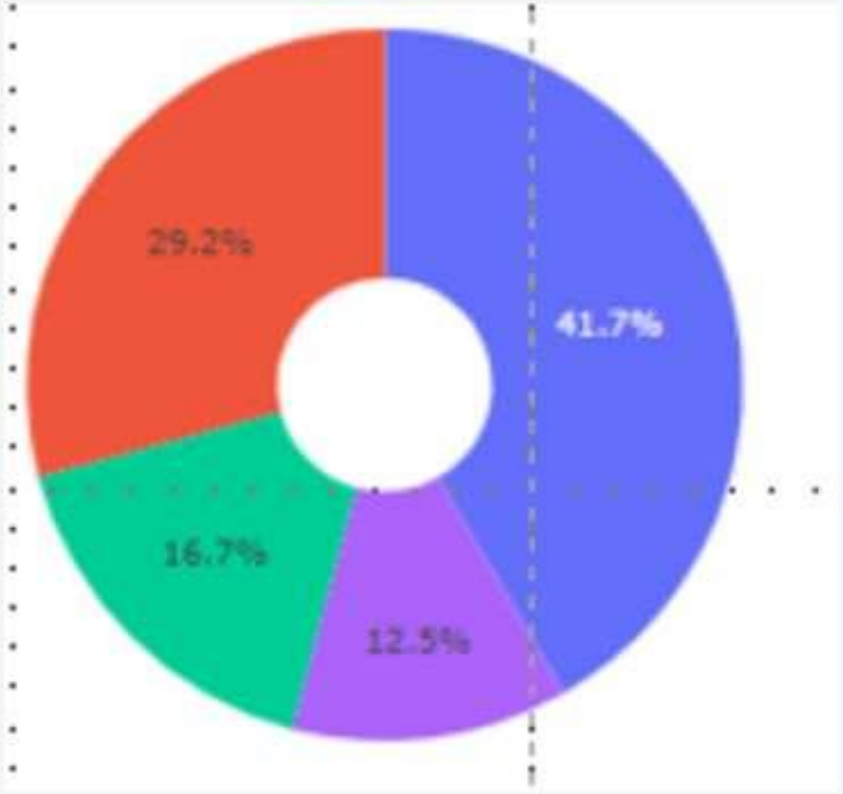
3. explore and analyze more details about proximities of launch sites, find out if the success launch have relationship with proximities conditions



Link: https://github.com/Dilapsky/IBM-Data-Science-Professional-Certificate/blob/main/lab_jupyter_launch_site_location.ipynb

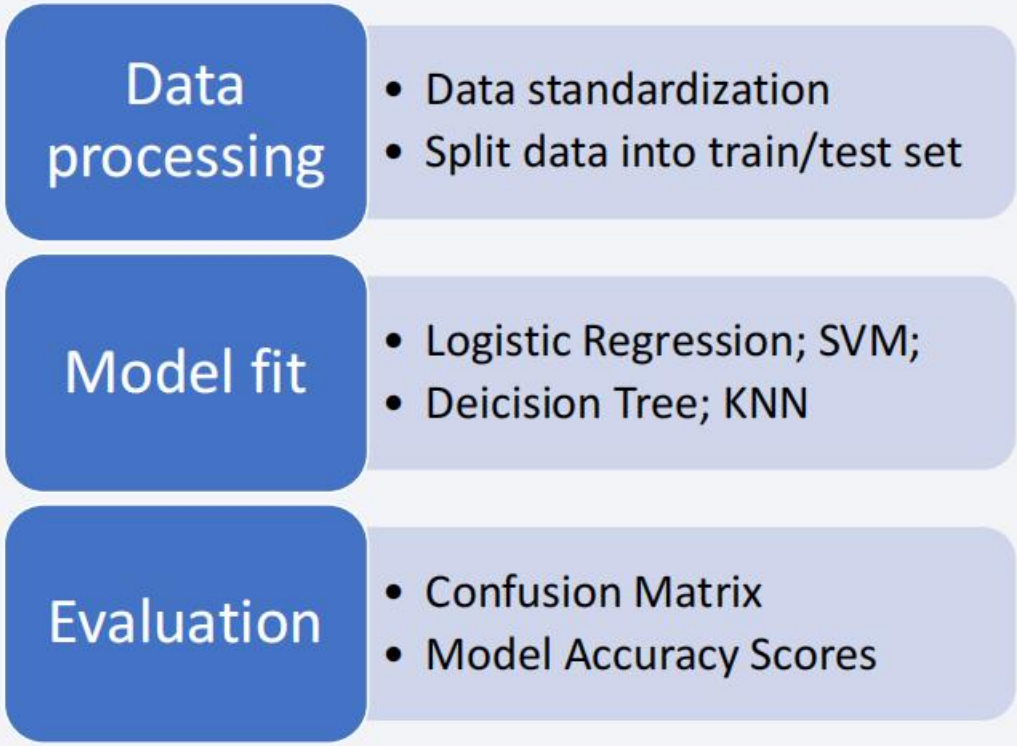
Build a Dashboard with Plotly Dash

- In the interactive dash board:
- A pie chart shows relationship of total success launches by site is created, it interactively shows the success number and successful rate of each site and in total
- A scatter plot shows the correlation between payload and success for all site is created, it includes a range slider, one can easily define the correlation of different payload range



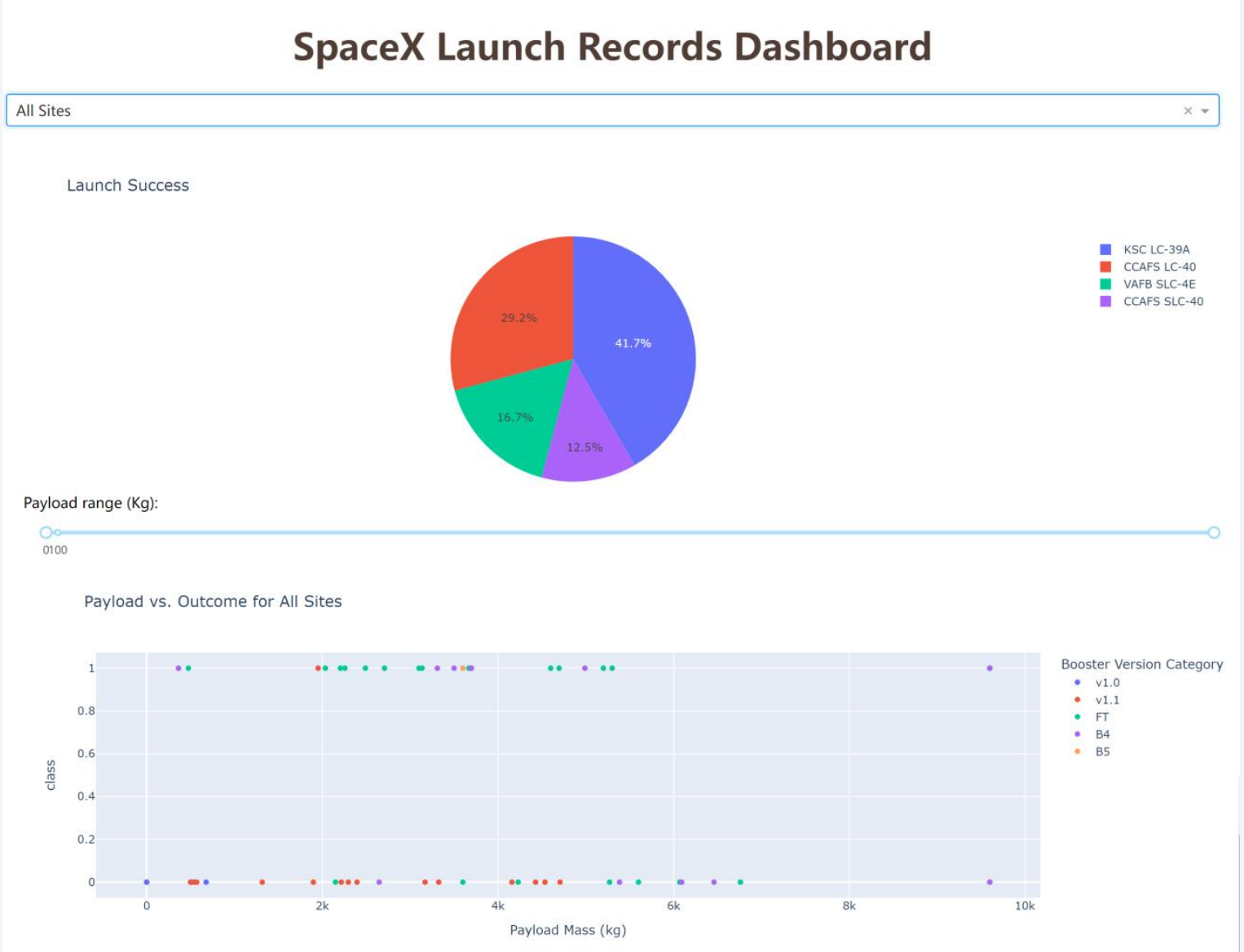
Predictive Analysis (Classification)

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. In this lab, I create a machine learning pipeline to predict.



Link: https://github.com/Dilapsky/IBM-Data-Science-Professional-Certificate/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results



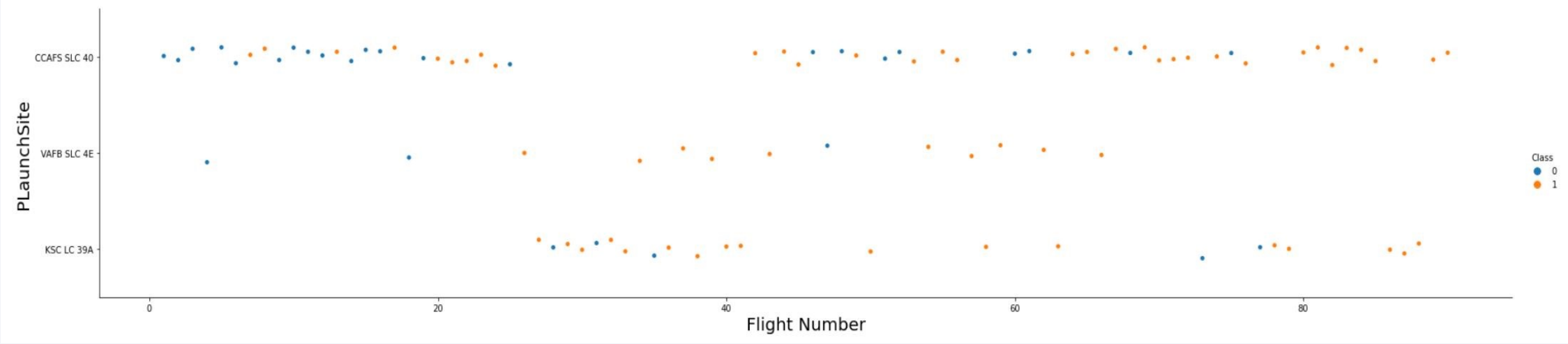
Exploratory data analysis founded good bases for predictive models, and reveals the correlations among data sets.

- Linear Regression, SVM, Decision Tree and K_x0002_Nearest-Neighbor model are applied for prediction and received relatively higher accuracy.

Section 2

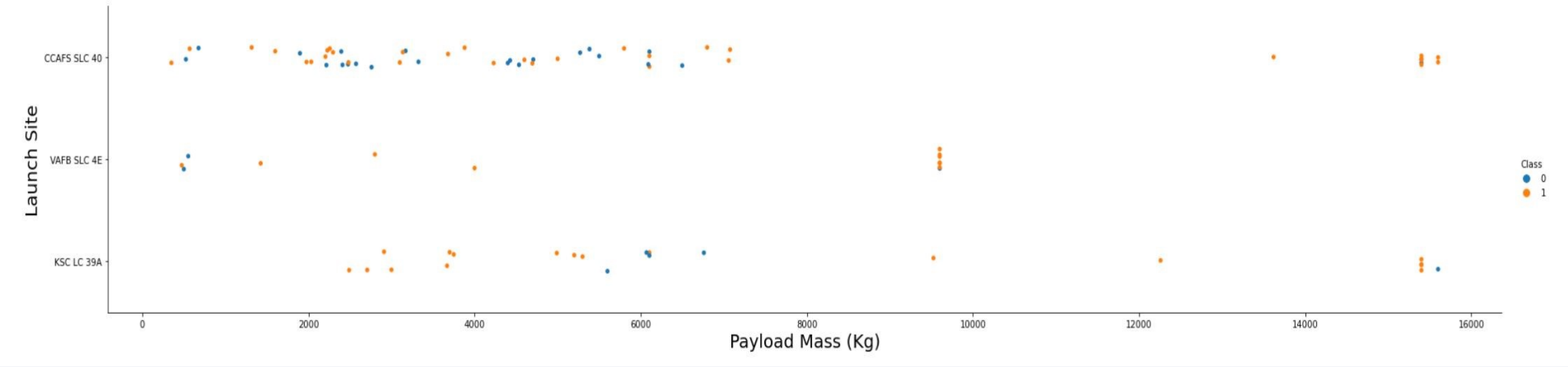
Insights drawn from EDA

Flight Number vs. Launch Site



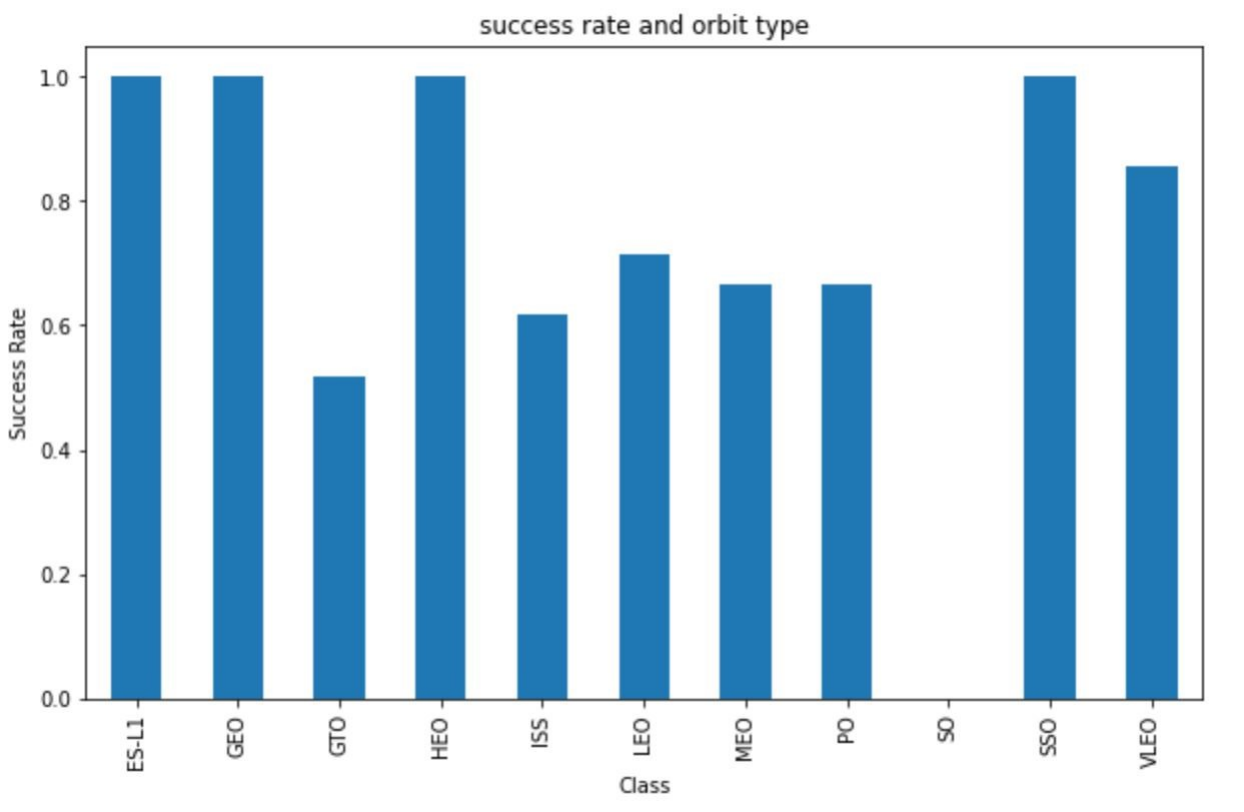
- the front flight number of site "CCAFS SLC 40" has a relative low success rate; the site "VAFB SLC 4E" has relative higher successful rate but with small sample set; the site "KSC LC 39A" is in the middle level between the other two classes.

Payload vs. Launch Site



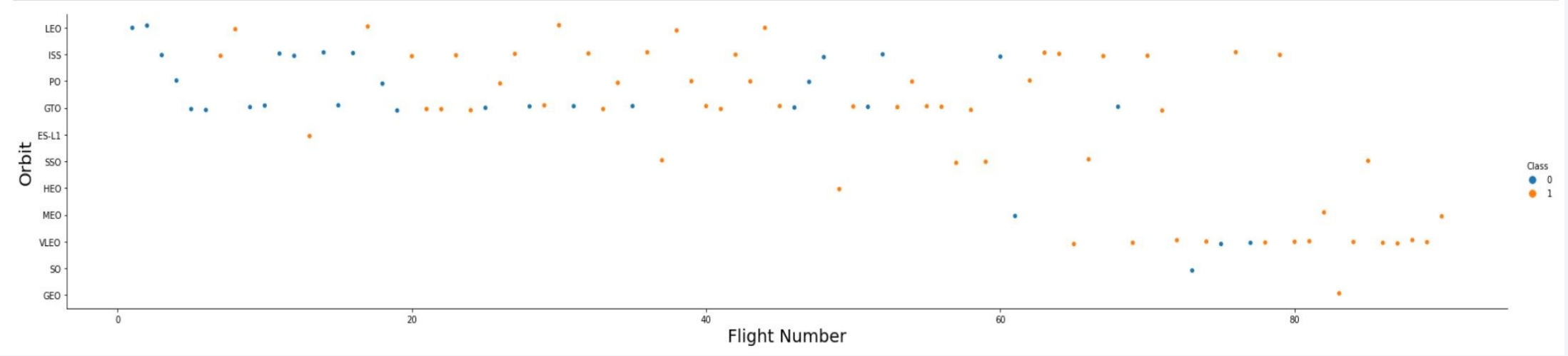
- When the payload mass is over 8000kg, the success rate of the three launch site increases significantly to nearly 100%; at site “VAFB SLC 4E”, the success rate performs good of a payload between 1000kg to 5000kg; at site “KSC LV 39A” it performs very well in the rage of payload under 7000kg.

Success Rate vs. Orbit Type



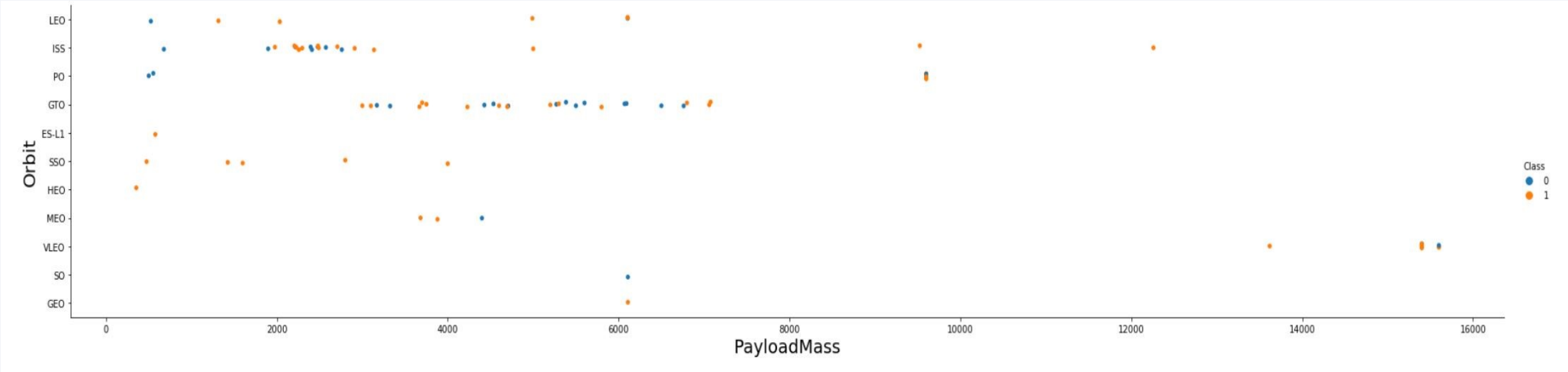
- Orbit type ES.L1, GEO, HEO, SSO have a success rate of 100%
- Orbit type SO has a success rate of 0%
- Orbit type ISS, LEO, MEO, PO, VLEO have also success rate over 50%
- The success rate has great correlation with orbit type

Flight Number vs. Orbit Type



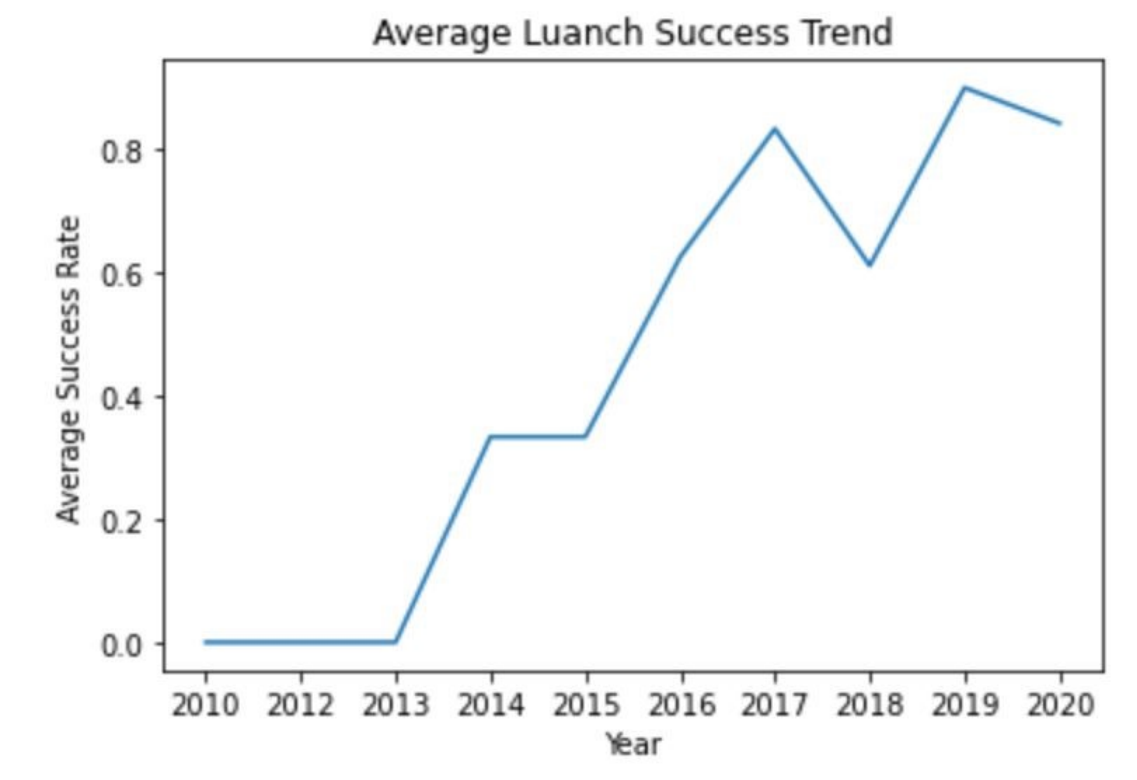
- Orbit type ES-L1, SSO, HEO, MEO, SO, GEO do not have enough number of flights. Therefore the success rate of those orbit types are lacking of samples.

Payload vs. Orbit Type



- For the orbit type ES-L1, SSO, HEO, MEO have only small payloads (under 4000kg), they have a good success rate
- For type LEO, ISS, VLEO, they have good success rate over 4000kg
- There are nosignificant correlations between orbit type and payloads in other types

Launch Success Yearly Trend



- The launch success rate is increasing from year 2010 to 2020.
- In year 2018 shows a small “valley”

All Launch Site Names

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

- Four unique launch site names are found in the data set
- CCAFS LC-40
- CCAFS SLC-40
- KSC LC-39A
- VAFB SLC-4E

Launch Site Names Begin with 'CCA'

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

- Five records are with site name begin with ‘CCA’ are shown in the table
- all of them are with orbit type LEO, launch site in CCAFS LC-40

Total Payload Mass

- the total payload carried by boosters from NASA is 45596kg

- NASA is the largest customer of SPACEX, so they have also a very high total payload

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2534 kg
- The booster version with extension v1.1 (B version) is also included

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad is 2015-12-22
- With the booster version F9 FT B1019, launch site at CCAFS LC-40, orbit LEO, payload mass 2034kg

Successful Drone Ship Landing with Payload between 4000 and 6000

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2016-05-06	05:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-08-14	05:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-10-11	22:53:00	F9 FT B1031.2	KSC LC-39A	SES-11 / EchoStar 105	5200	GTO	SES EchoStar	Success	Success (drone ship)

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

  F9 FT B1022, F9 FT B 1026, F9 FT B1021.2, F9 FT B1031.2

- They are all use the orbit GTO

Total Number of Successful and Failure Mission Outcomes

mission_outcome	COUNT
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- Only one mission outcome is failure
- They other 100 are success, include only one outcome with “payload status unclear”

Boosters Carried Maximum Payload

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

- The table lists the names of the booster which have carried the maximum payload mass
- All the booster is F9 B5 B series varieties.

2015 Launch Records

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- The failed landing_outcomes in drone ship, and their booster versions, and launch site names for in year 2015 are list in the table

- They are all with booster version F9 v1.1 B101x and launch site in CCAFS LC -40

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

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2016-06-15	14:29:00	F9 FT B1024	CCAFS LC-40	ABS-2A Eutelsat 117 West B	3600	GTO	ABS Eutelsat	Success	Failure (drone ship)
2016-03-04	23:35:00	F9 FT B1020	CCAFS LC-40	SES-9	5271	GTO	SES	Success	Failure (drone ship)
2016-01-17	18:42:00	F9 v1.1 B1017	VAFB SLC-4E	Jason-3	553	LEO	NASA (LSP) NOAA CNES	Success	Failure (drone ship)
2015-04-14	20:10:00	F9 v1.1 B1015	CCAFS LC-40	SpaceX CRS-6	1898	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)
2015-01-10	09:47:00	F9 v1.1 B1012	CCAFS LC-40	SpaceX CRS-5	2395	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)

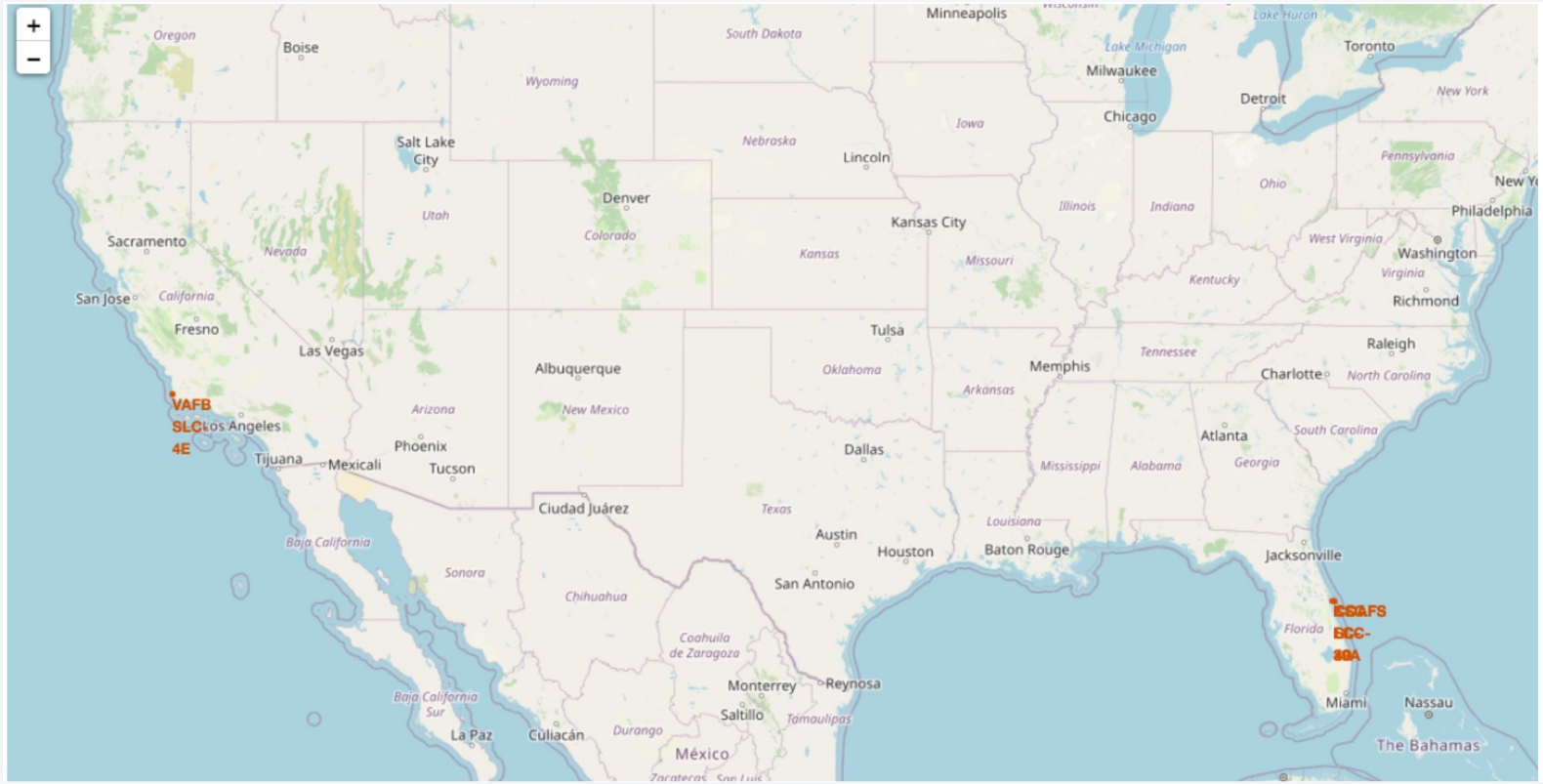
DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2016-07-18	04:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2015-12-22	01:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

- There are 5 landing outcomes Failure (drone ship) and 3 landing outcome Success (ground pad) between the date 2010-06-04 and 2017-03-20

Section 4

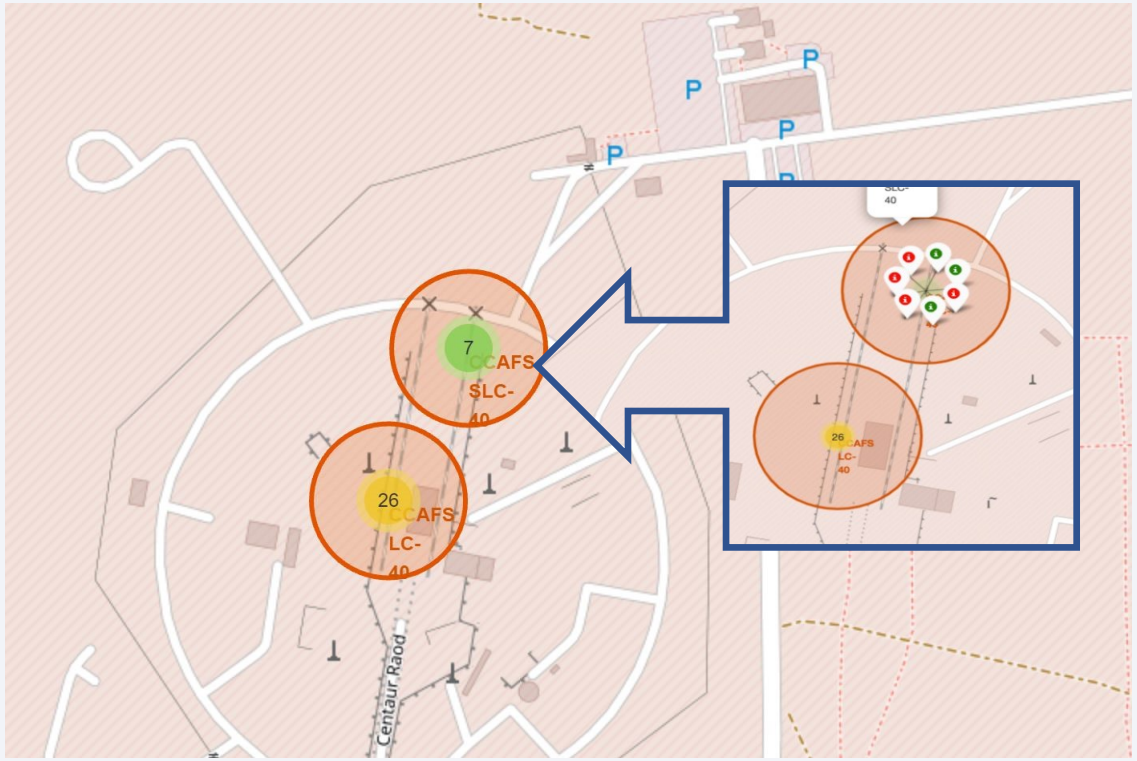
Launch Sites Proximities Analysis

Launch Site on the map



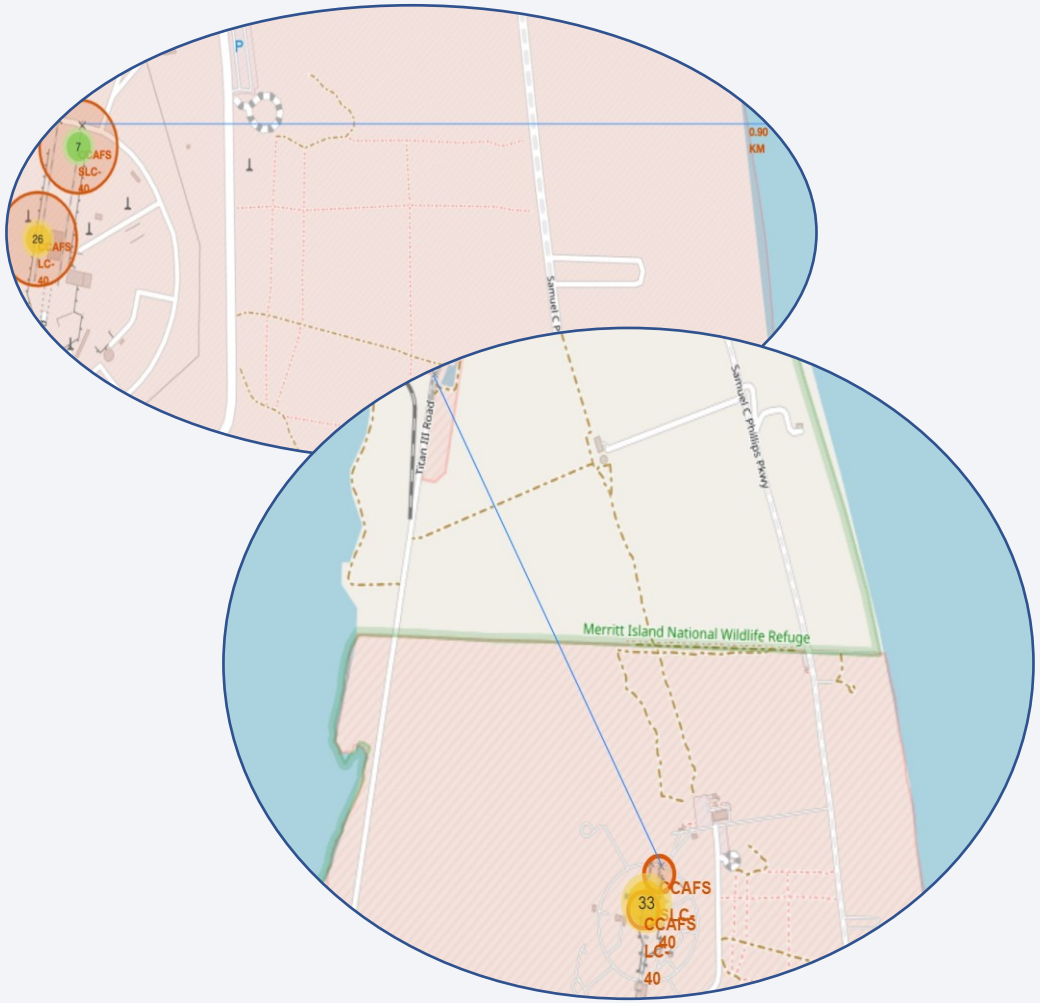
- The launch sites are all in the very southern area of US
- The launch sites are all close to coast

Map of success and failed launches for each site

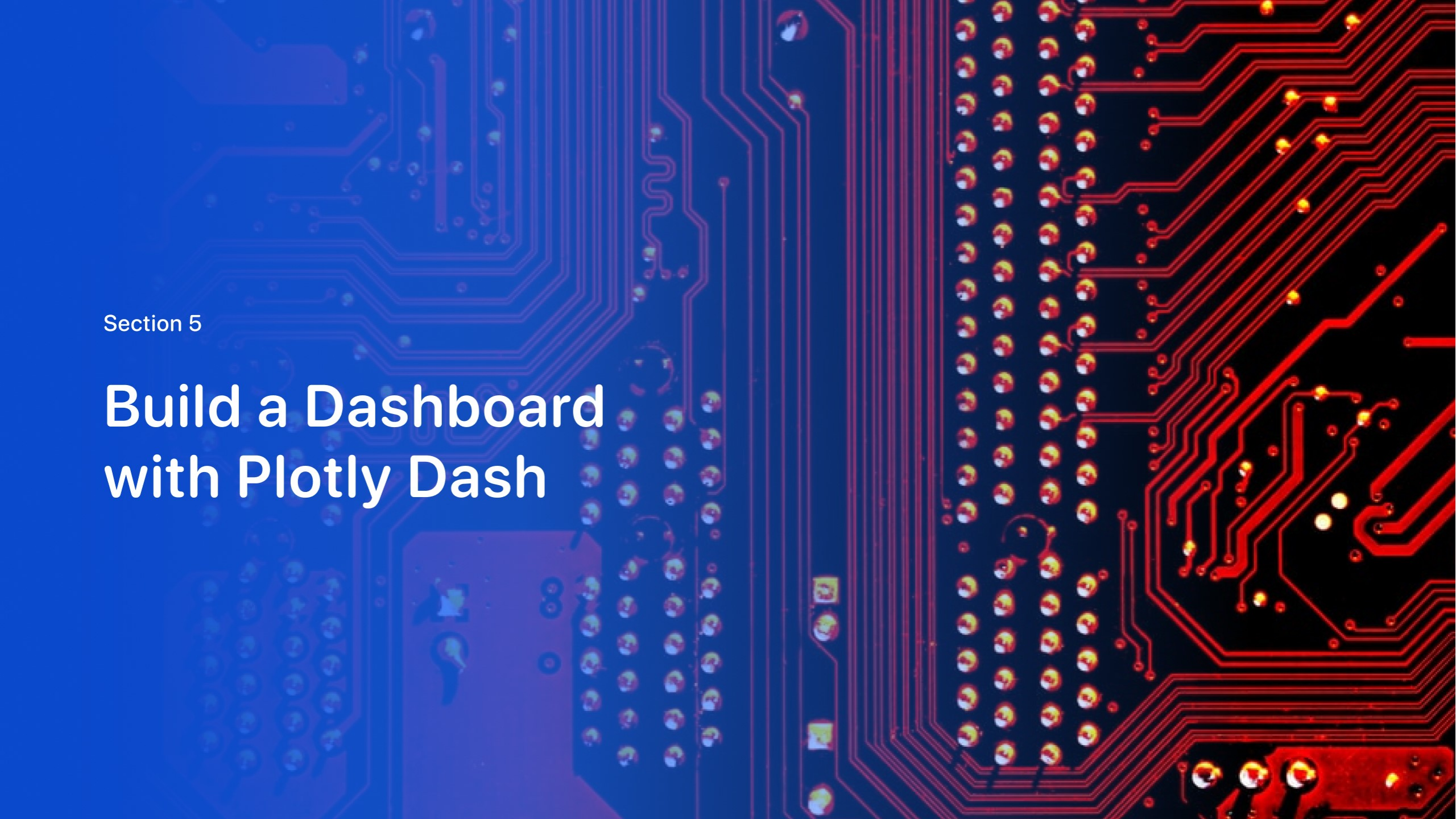


- From this map, it is easy to identify which launch sites have relatively high success rates, and which launch sites have more launches
- For example, at CCAFS SLC-40 site, 7 launches and the success rate is 57%

map of distances between a launch site to its proximities



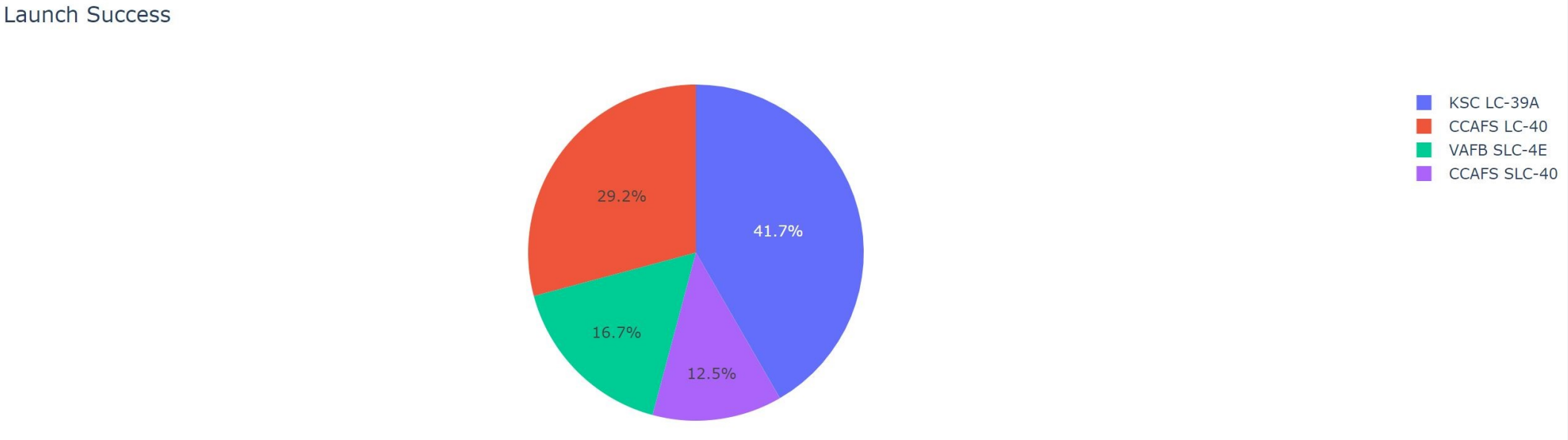
- The map clearly shows the distances between the launch sites to other facilities.
- For example, CCAFS SLC-40 is very close to a coast, less than 1km. It is also close to a rail terminal station for a distance about 1.8km. It is relatively close to highway, but away from large cities like Orlando.



Section 5

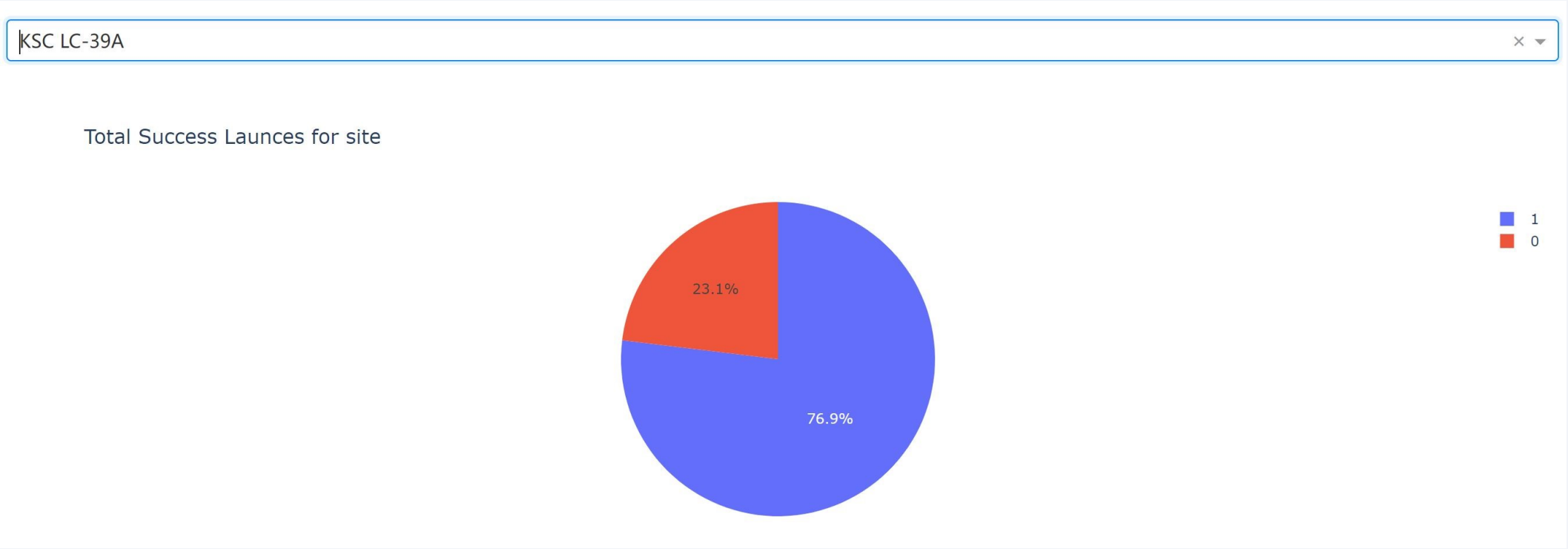
Build a Dashboard with Plotly Dash

Launch Success – All Sites



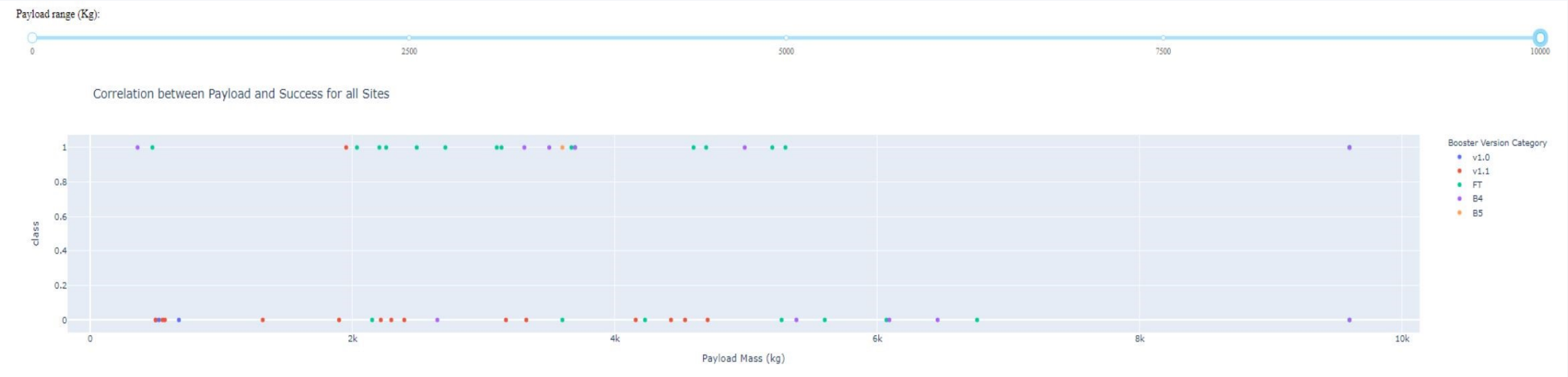
- KSC LC-39A contributes the highest portion of launch success; the second is CCAFS LC-40; they other two sites have smaller launch successes.

Launch Site with highest launch success ratio



- KSC LC-39A has the highest launch success ratio.

Payload vs. Launch Outcome

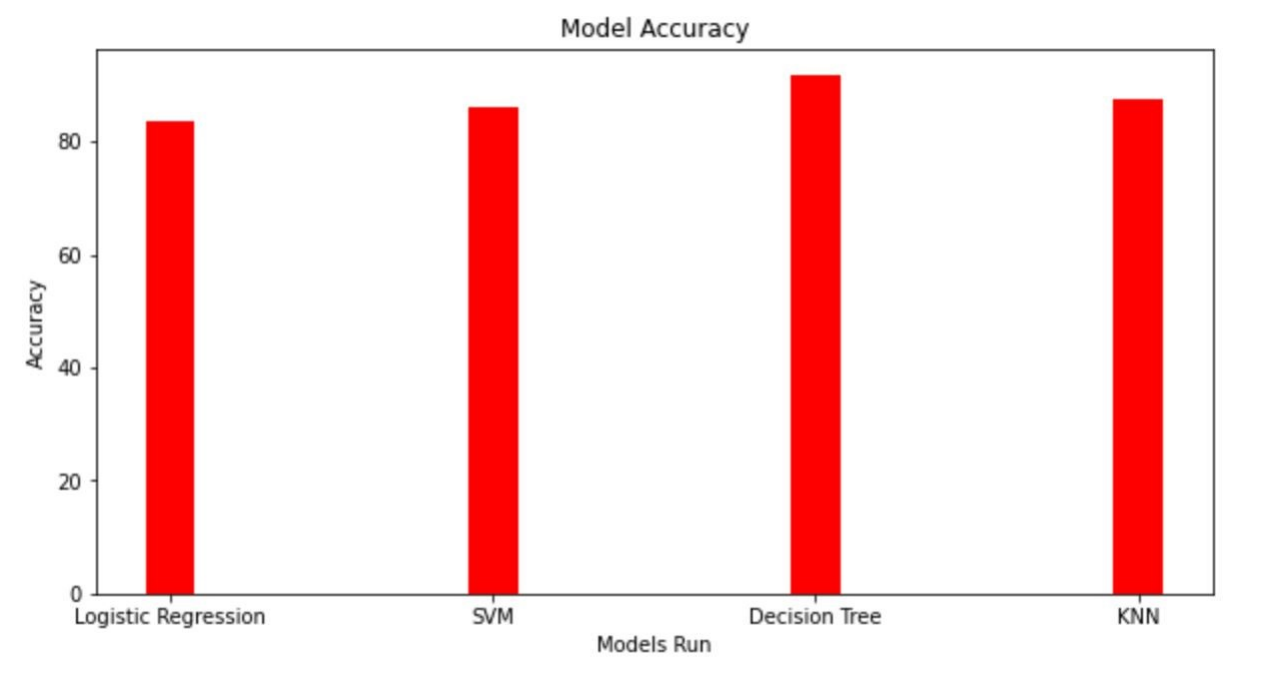


- With the payload in the range 2000kg-6000kg, it has the highest launch success rate; payload below 2000kg and over 6000kg, the success rate is low.
- The FT booster version perform very well between the payload of 2000kg to 6000kg

Section 6

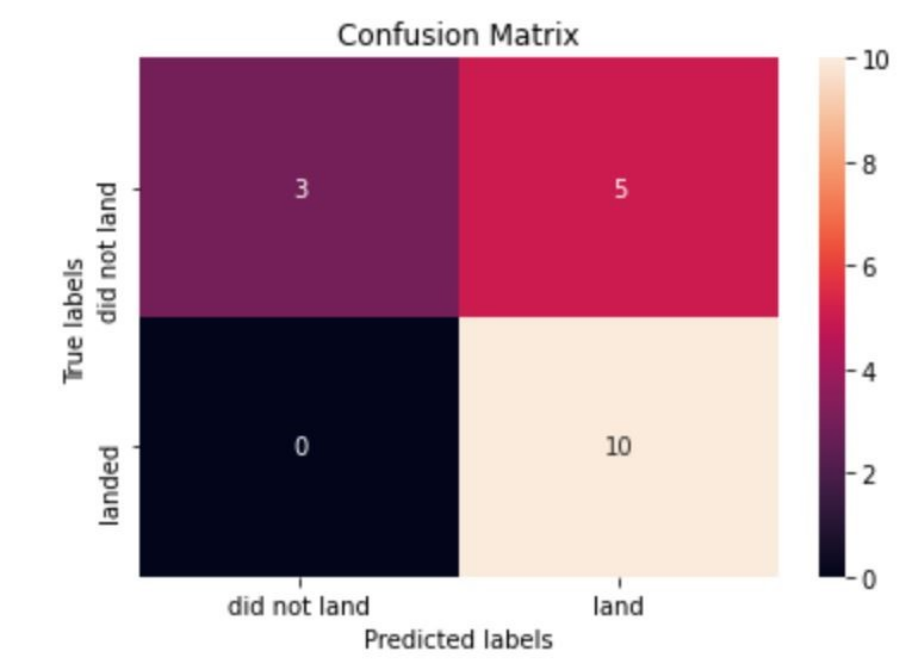
Predictive Analysis (Classification)

Classification Accuracy



- The prediction models showing very good accuracy scores.
- The best model is “Decision Tree”, with an accuracy score of 0.91.

Confusion Matrix



- The confusion matrix of Decision Tree, easy to figure out it performs good predictions
- No False Positive, then the TPR=100%
- False Negative is also very low

Conclusions

- Exploratory data analysis founded good bases for predictive models, and reveals the correlations among data sets, for example, the correlation between orbit type and success rate is significant
- Scatter plot shows that payload and launch site contribute much to the success rate
- SQL function helps easily to group the landing outcomes and sort by attributes
- Maps visualize directly the location of launch site, and show the distances between launch site and proximities, for example, launch sites are always located near cost
- From interactive dash board, one can directly view the launch success rate per site; and the effect of payload in different range
- The predict models selected in the task works very well on the prediction, the Decision tree model has the highest accuracy.

Appendix

- Help information on:
- Pandas: https://pandas.pydata.org/
- Dashboard Plotly: https://dash.plotly.com/
- GitHub Documentation: https://docs.github.com/
- IBM Cloud Park Docs: https://cloud.ibm.com/docs
- SQL Functions: https://www.w3schools.com/sql/sql_ref_sqlserver.asp
- Learning materials from IBM Data Science Course
- Wikipedia Page – SpaceX: https://en.wikipedia.org/wiki/SpaceX
- Wikipedia Page – Confusion Matrix: https://en.wikipedia.org/wiki/Confusion_matrix

Thank you!