IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

NAEL SA'DI AQEL

01 – 12 – 2021

Outline

Executive Summary

- In our journey with this project, we will apply all knowledge that learnt through the courses in IBM certificate, it included (Data Collection, Data Wrangling, Exploratory Data Analysis using Visualization and SQL, Building an Interactive Visual Analytics using Folium and Plotly Dash, and finally, Building Predictive Classification Model using Machine Learning)
- After understanding and analyzing our data using the previous methodologies, we successfully built a predictive model with 85% accuracy, as what we will see in the next slides

Introduction

- In this project, 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.
- 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 SpaceX for a rocket launch.

Section 1

Methodology

Methodology

Executive Summary

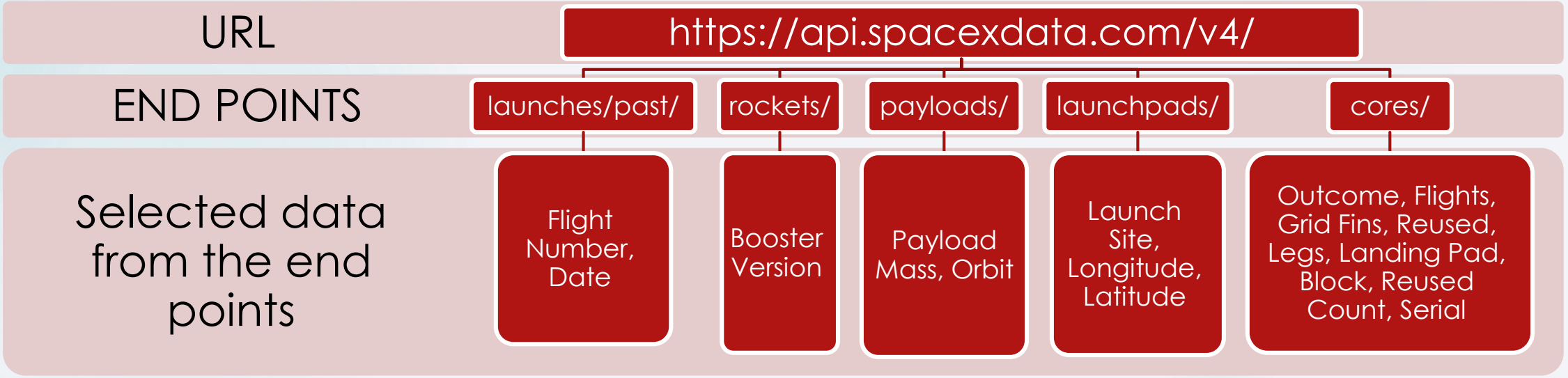
- Data Collection Methodology
  - Data collected from SpaceX (API) and Wikipedia (Web Scraping)
- Data Wrangling
- Exploratory Data Analysis (EDA) using Visualization
- Exploratory Data Analysis (EDA) using SQL
- Interactive Visual Analytics using Folium
- Interactive Visual Analytics using Plotly Dash
- Predictive Analysis using (Classification) Models

Data Collection

- We need any available data that contains the history and information of Falcon 9 trips and first stage landing which can be used to train our model, the data founded in two resources:
  - SpaceX website in a form of JSON which we can access by APIs provided from them
  - Wikipedia where it is available in form of HTML tables
- Our goal in this stage is to collect the data and format it in a data frames to start our process on it using the data science methods

Data Collection – SpaceX API

- Data collected using GET request of (URL+END POINT)
- From each response, a specific data was selected then saved to a data frame
- We filtered the data frame to the only included **"FALCON 9"** Launches
- The completed Data Collection API notebook is here

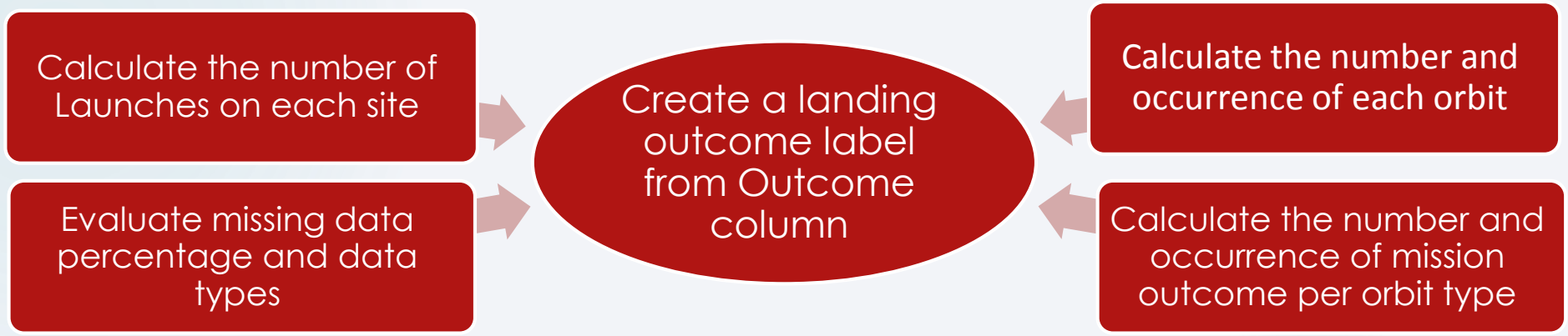


Data Collection - Scraping

- The historical **"FALCON 9"** Launch records had been collected from Wikipedia page titled: List of Falcon 9 and Falcon Heavy launches
- Below the web scraping process
- The completed Data Collection with Web Scraping notebook is here

Data Wrangling

- After required data were collected and formatted in one data frame, some analysis was performed to find patterns that would be the label for training the supervised models, below the process done
- At the end, an output column **"Class"** was created for the status of first stage landing (1: Success, 0: Fail)
- The completed Data Wrangling notebook is here



EDA with Data Visualization

- Using Matplotlib, we visualize the relationship between different items to prepare data feature engineering to use in our training models, the plots are:
  - Flight Number vs Launch Site
  - Payload Mass vs Launch Site
  - Success Rate vs Orbit Type
  - Flight Number vs Orbit Type
  - Payload Mass vs Orbit Type
  - Launch Success Yearly Trend
- The completed EDA with Visualization notebook is here

EDA with SQL

- Using SQL, we evaluated:
  - All Launch Site Names
  - Launch Site Names Begin with 'CAA'
  - Total Payload Mass Carried from NASA
  - Average Payload Mass by F9 v1.1
  - First Successful Ground Landing Date
  - Successful Drone Ship Landing with Payload between 4000 and 6000
  - Total Number of Successful and Failure Mission Outcomes
  - Boosters Carried Maximum Payload
  - Failure Drone Ship Landing in 2015
  - Rank Landing between 2010-06-04 and 2017-03-20
- The completed EDA with SQL notebook is here

Build an Interactive Map with Folium

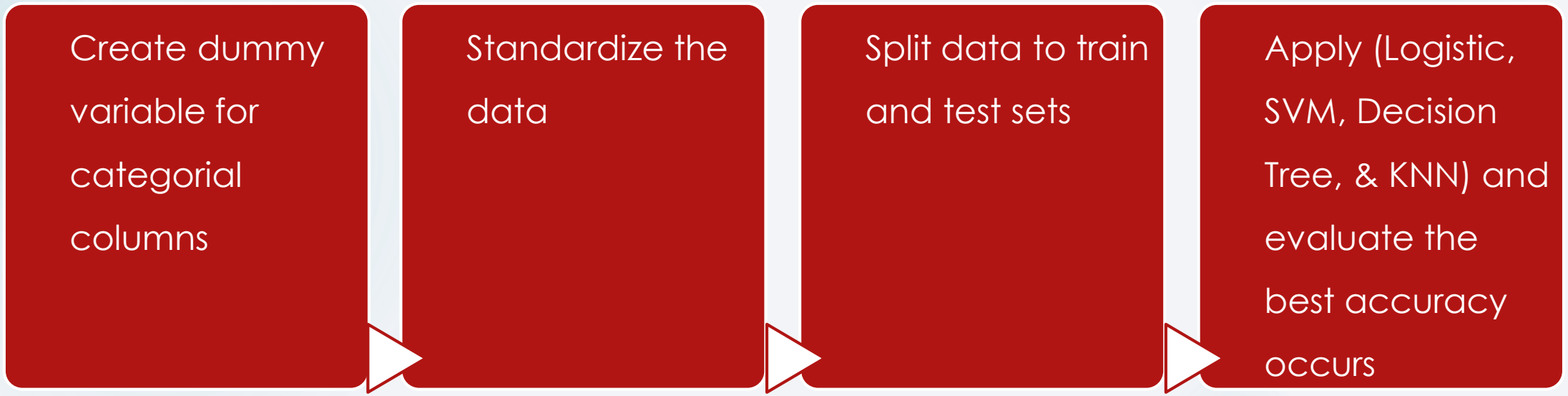
- All launches sites had been marked on the Map
- Then, for each site we made colored marked as per launch status (success - green / failure - red) for each launch happened
- And we marked and calculated the distance between each site and its proximities
- The purpose of this process is to make our data visualized into the map for better understanding
- The completed Data Visualization with Folium notebook is here

Build a Dashboard with Plotly Dash

- An interactive dashboard was created, it contains:
  - Dropdown List where you can choose the results all Launch Sites or a specific Launch Site
  - Pie Chart for the Total Success Launches as per your dropdown list selection
  - Range Slider to choose the range of payload mass you need to see in the Payload vs. Launch Outcome Plot, the available range is 0 to 10,000 kg with 1000 kg step
  - Scatter Chart of the Payload vs. Launch Outcome for the site selected in the dropdown list in the payload range selected by the range slider
- Some screenshots from the dashboard below:
  - Launch Success Pie Chart for All Sites
  - The Highest Launch Success Site Pie Chart
  - Payload vs. Launch Outcome for All Sites
- The completed Dashboard Application with Plotly Dash python code is here

Predictive Analysis (Classification)

- Below the predictive analysis process
- The completed Machine Learning Prediction notebook is here



Results

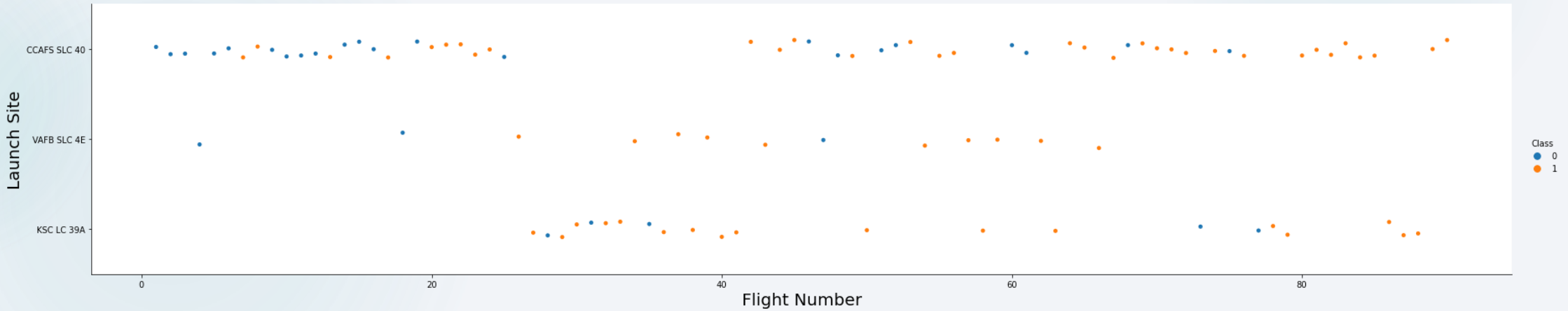
- Exploratory Data Analysis Results
- Interactive Analytics Demo in Screenshots – Maps
- Interactive Analytics Demo in Screenshots – Dashboard
- Predictive Analysis Results

Section 2

Insights drawn from EDA

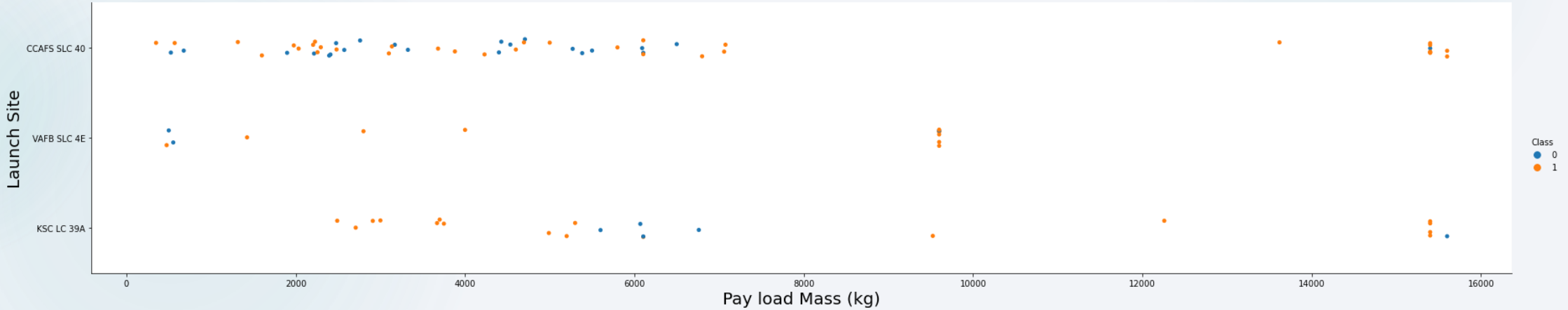
Flight Number vs. Launch Site

- We can notice that most KSC LC 39A & VAFB SLC 4E sites had a success landing, while most success landing on site CCAFS SLC 40 are the ones who had high flight numbers



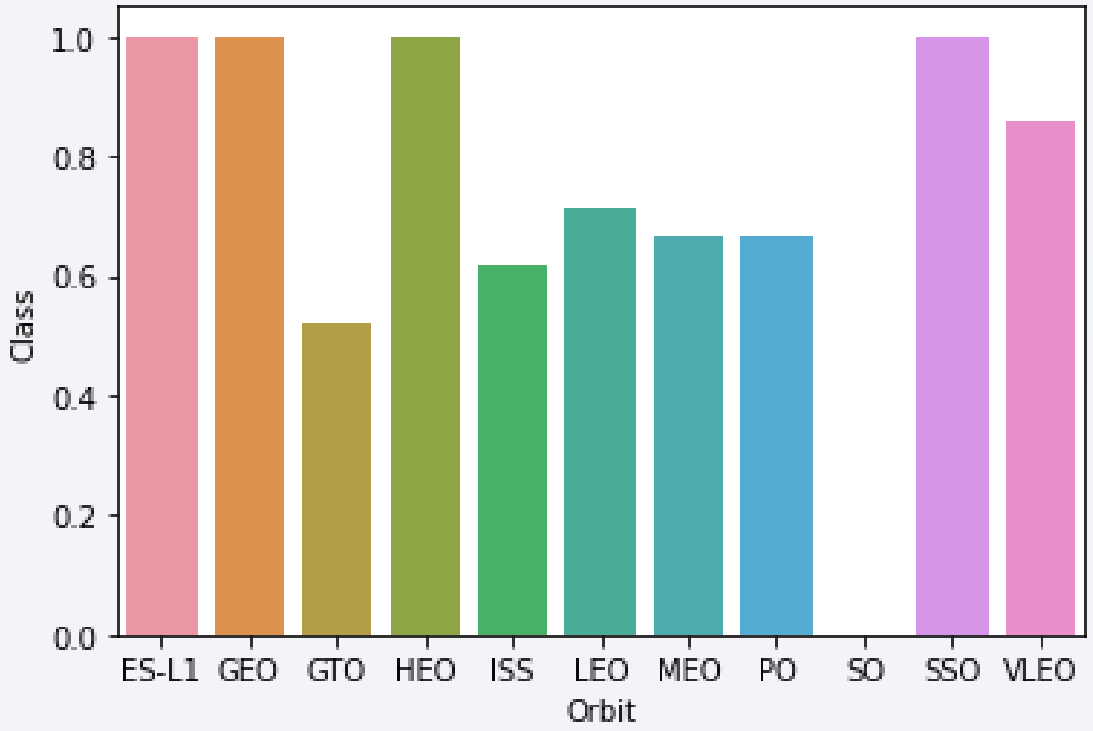
Payload vs. Launch Site

- We can notice that VAFB SLC 4E site has no heavy rockets and most of its landing are success, while most landing on KSC LC 39A site are success, but for CCAFS SLC 40 site the relation between payload and success landing is not that related



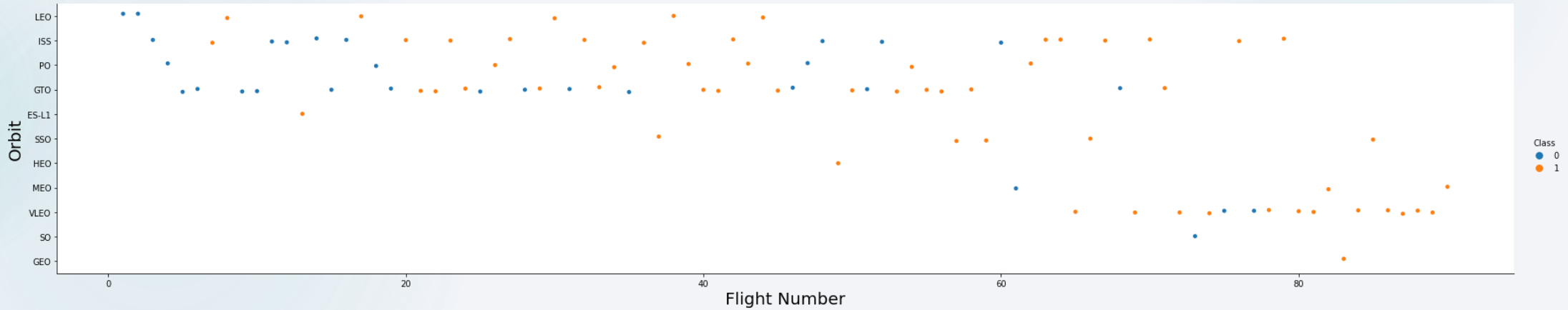
Success Rate vs. Orbit Type

- We can notice that ES-L1, GEO, HEO and SSO Orbits has 100% success rate, and the lowest success rate is for GTO Orbit with 50% rate



Flight Number vs. Orbit Type

- We can notice that the success rate increases as the flight number increase

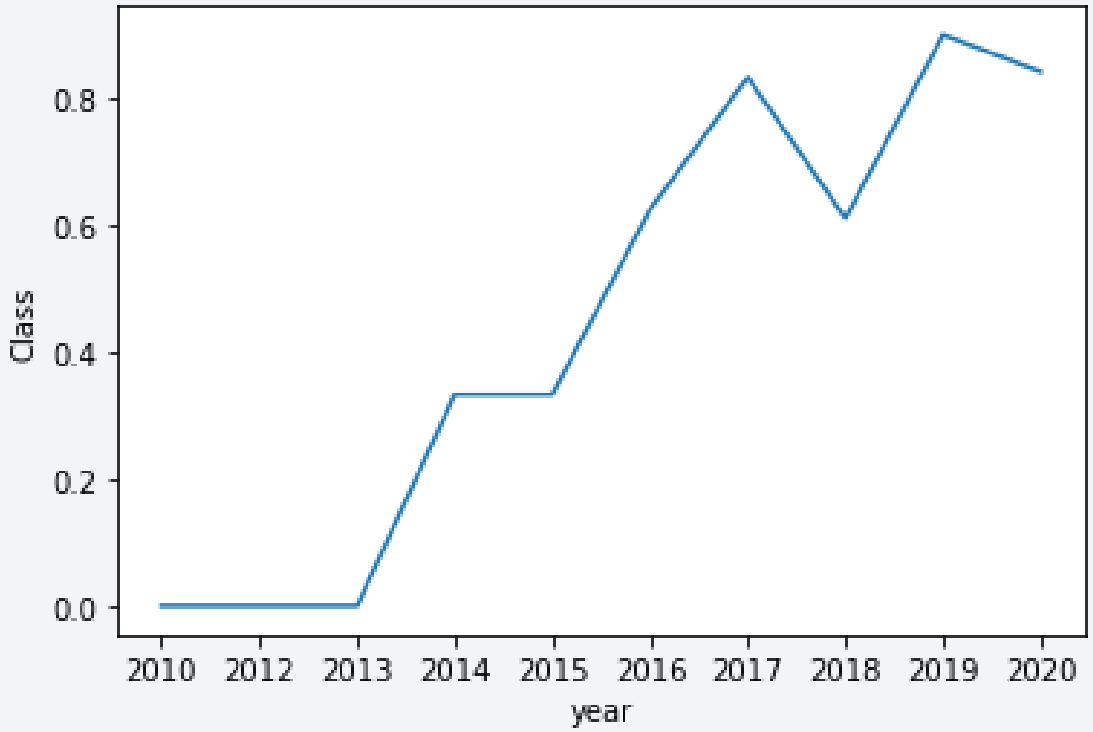


Payload vs. Orbit Type

- We can notice that the success rate increases as Payload Mass increases except for GTO Orbit

Launch Success Yearly Trend

- We can notice that generally the success rate is increased with years



All Launch Site Names

- The Launch Site names was founded using “UNIQUE” query in SQL, the result shows that we have 4 Launch Sites

```
In [6]: %sql SELECT UNIQUE(LAUNCH_SITE) FROM SPACEXTBL
 * ibm_db_sa://tlp33288:***@8e359033-a1c9-4643-8
Done.
Out[6]:
```

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

Launch Site Names Begin with 'CCA'

- The Launch Site names begin with ‘CCA’ was founded using “LIKE” query in SQL, “LIMIT” is used to show only first 5 rows

In [7]:
```
%sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
```

```
 * ibm_db_sa://tlp33288:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
Done.
```

Out[7]:

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 Carried from NASA

- The total payload mass from NASA was founded using “SUM” with “WHERE” as shown below

```
In [8]: %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)'
 * ibm_db_sa://tlp33288:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1o
Done.
Out[8]: 1
45596
```

Average Payload Mass by F9 v1.1

- The average payload mass by F9 v1.1 was founded using “AVG” with “WHERE” as shown below

```
In [9]: %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1'
 * ibm_db_sa://tlp33288:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lc
Done.
Out[9]: 1
2928
```

First Successful Ground Landing Date

- The first successful ground landing was founded using “MIN” with “WHERE” as shown below

```
In [10]: %sql SELECT MIN(DATE) FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Success (ground pad)'
 * ibm_db_sa://tlp33288:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.
Done.
Out[10]:          1
          2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- The booster version for success drone ship landing with payload between 4000 and 6000 was founded using “WHERE”, “AND” and “BETWEEN” queries as shown below

```
In [11]: %sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Success (ground pad)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
 * ibm_db_sa://tlp33288:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
Done.
```

Out[11]:

booster_version
F9 FT B1032.1
F9 B4 B1040.1
F9 B4 B1043.1

Total Number of Successful and Failure Mission Outcomes

- The total number of success and failure mission was founded using “COUNT” and “GROUP BY” queries as shown below

```
In [12]: %sql SELECT LANDING__OUTCOME,COUNT(*) AS TOTAL FROM SPACEXTBL GROUP BY LANDING__OUTCOME
 * ibm_db_sa://tlp33288:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.data
Done.
```

Out[12]:

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

Boosters Carried Maximum Payload

- The booster carried maximum payload was founded using SUBQUERY with “MAX” query as shown below

```
In [13]: %sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
 * ibm_db_sa://tlp33288:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120
Done.
```

Out[13]:

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

Failure Drone Ship Landing in 2015

- The booster version and launch site for failure drone ship landing in 2015 was founded using “WHERE”, “AND” and “YEAR” queries as shown below

```
In [14]: %sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND YEAR(DATE) = 2015

 * ibm_db_sa://tlp33288:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
Done.
```

Out[14]:

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

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

- The Rank landing outcome between mentioned dates was founded using “COUNT”, “WHERE”, “BETWEEN”, “GROUP BY” and “ORDER BY” as shown below

```
In [15]: %sql SELECT LANDING__OUTCOME,COUNT(*) AS TOTAL FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING__OUTCOME ORDER BY TOT
 * ibm_db_sa://tlp33288:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
Done.
```

Out[15]:

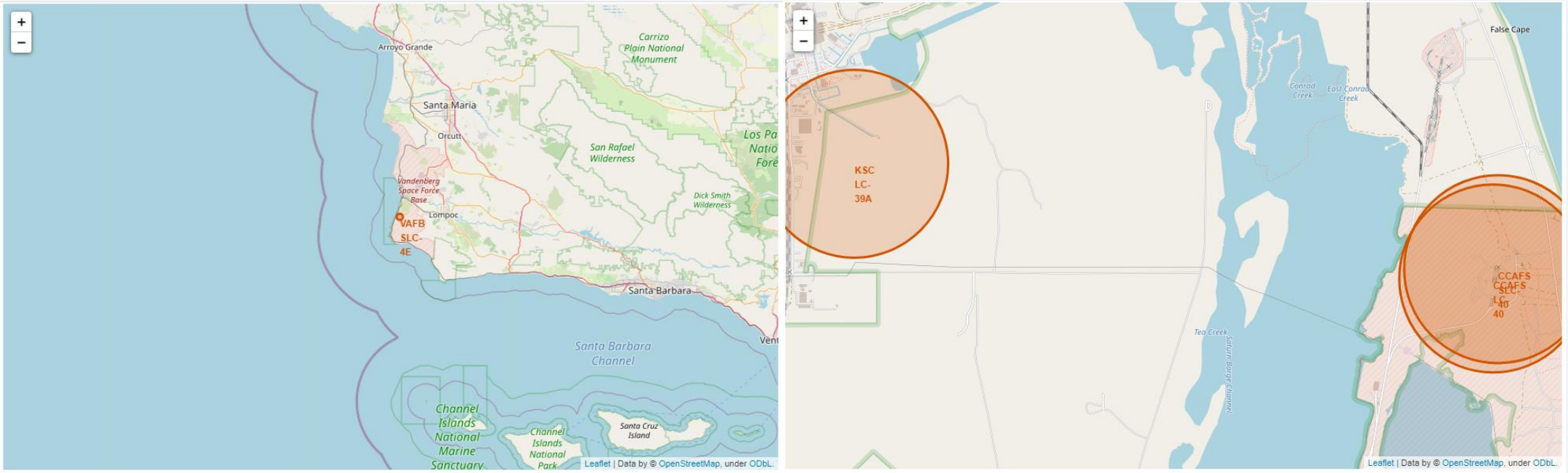
landing__outcome	total
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 4

Launch Sites Proximities Analysis

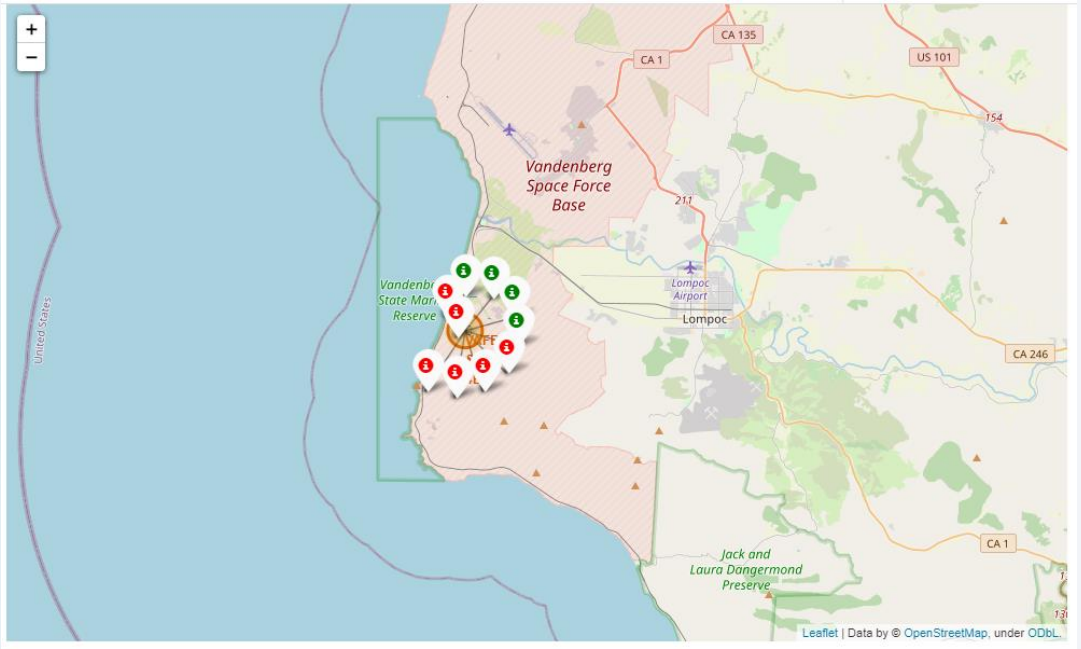
All Launch Sites' Marker on Map

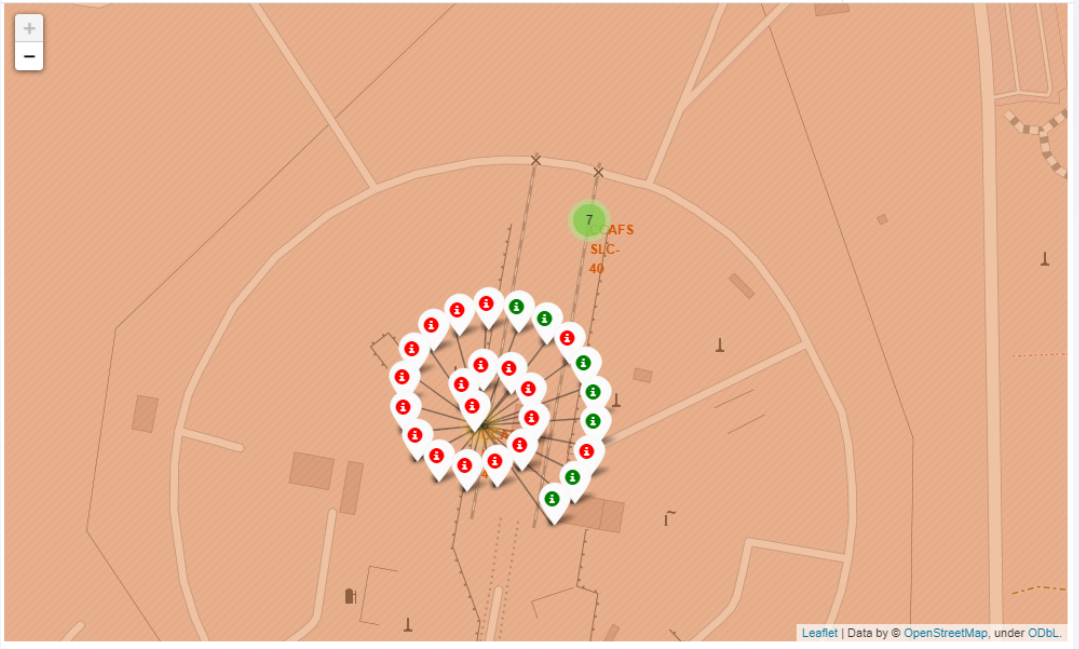
- To see the maps more clearly, the map separated into two parts (West and East)



Color-labeled Launch Outcome on Map

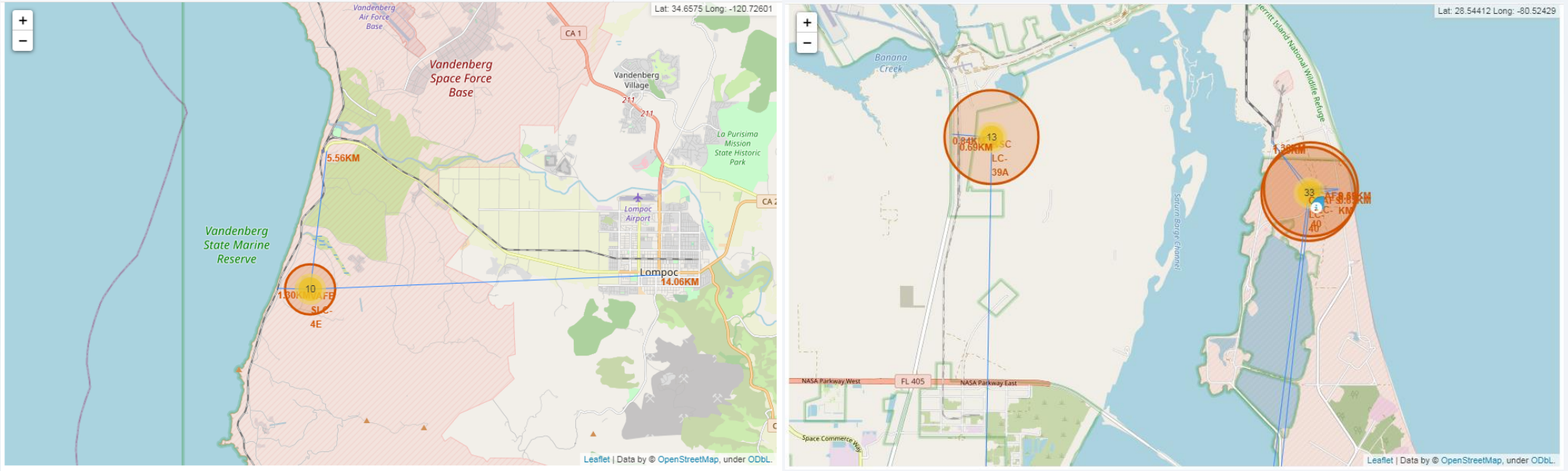
- To see the maps more clearly, the map separated into two parts (West and East)

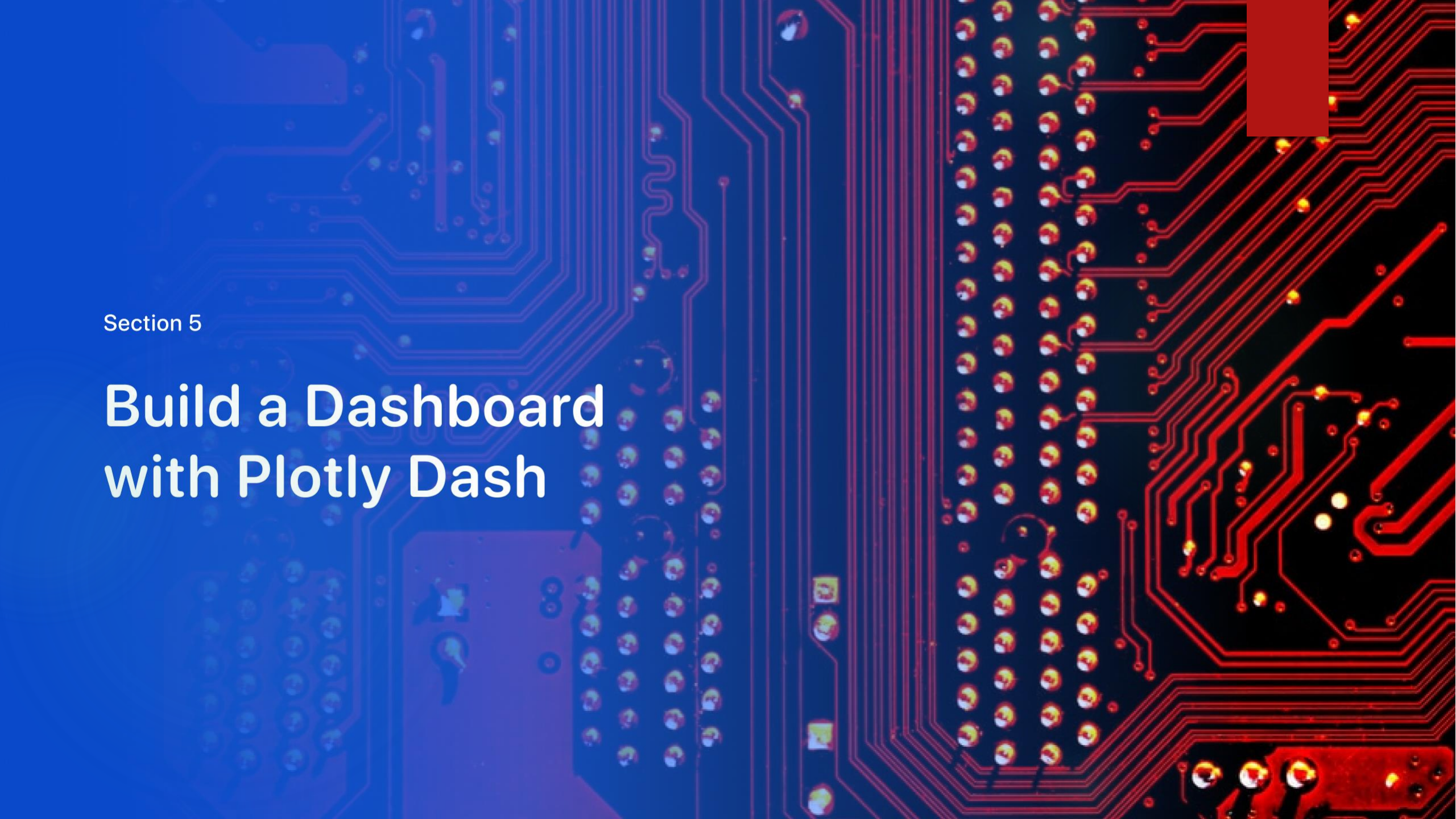




Sites Proximities on Map

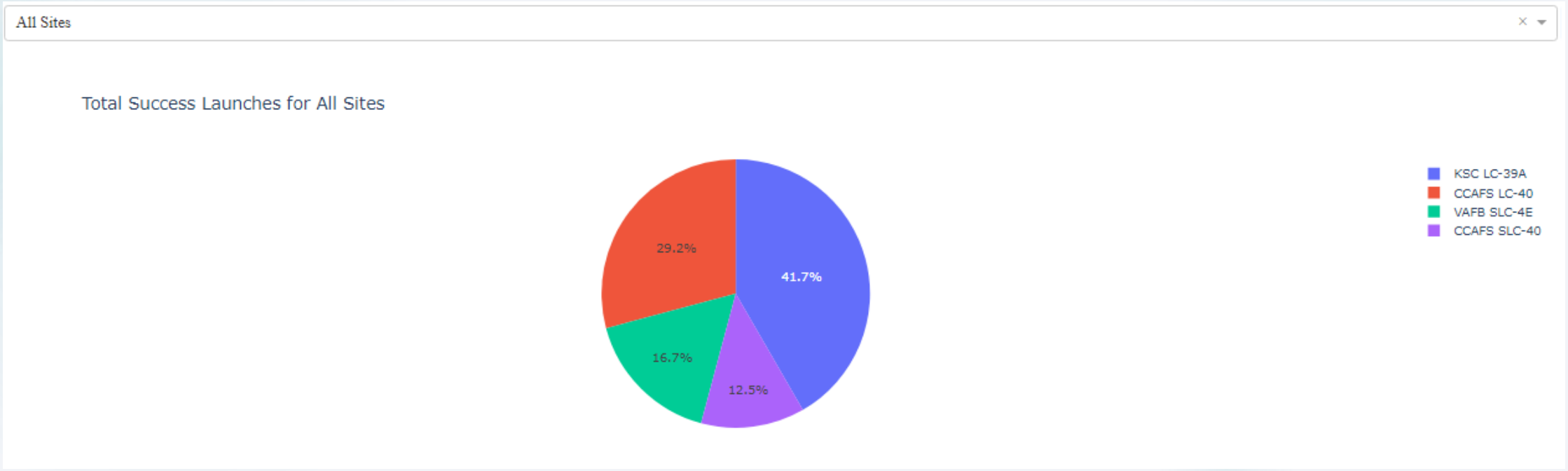
- To see the maps more clearly, the map separated into two parts (West and East)




Section 5
Build a Dashboard with Plotly Dash

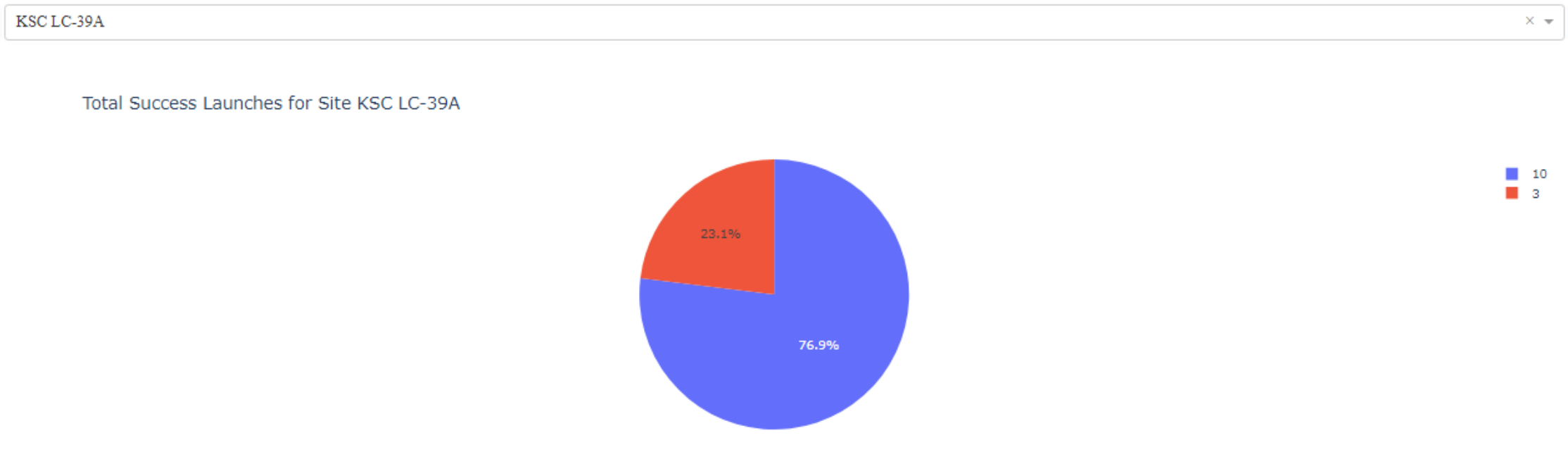
Launch Success Pie Chart for All Sites

- The highest success lunches was in KSC LC-39A Site, and the lowest success launches was in CCAFS CLS-40 Site



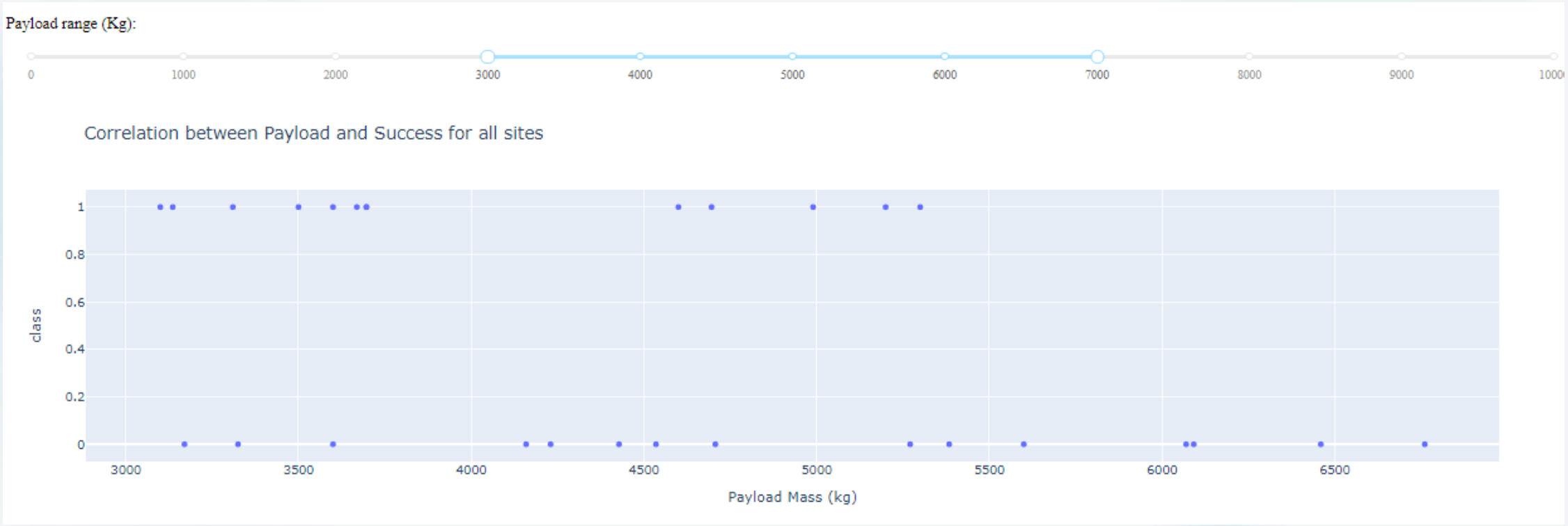
The Highest Launch Success Site Pie Chart

- The highest success lunches was in KSC LC-39A Site



Payload vs. Launch Outcome for All Sites

- The Payload vs. Launch Outcome for all sites in payload range between 3000 to 7000 kg. In this range, the most success launches was between 3000 to 4000 kg



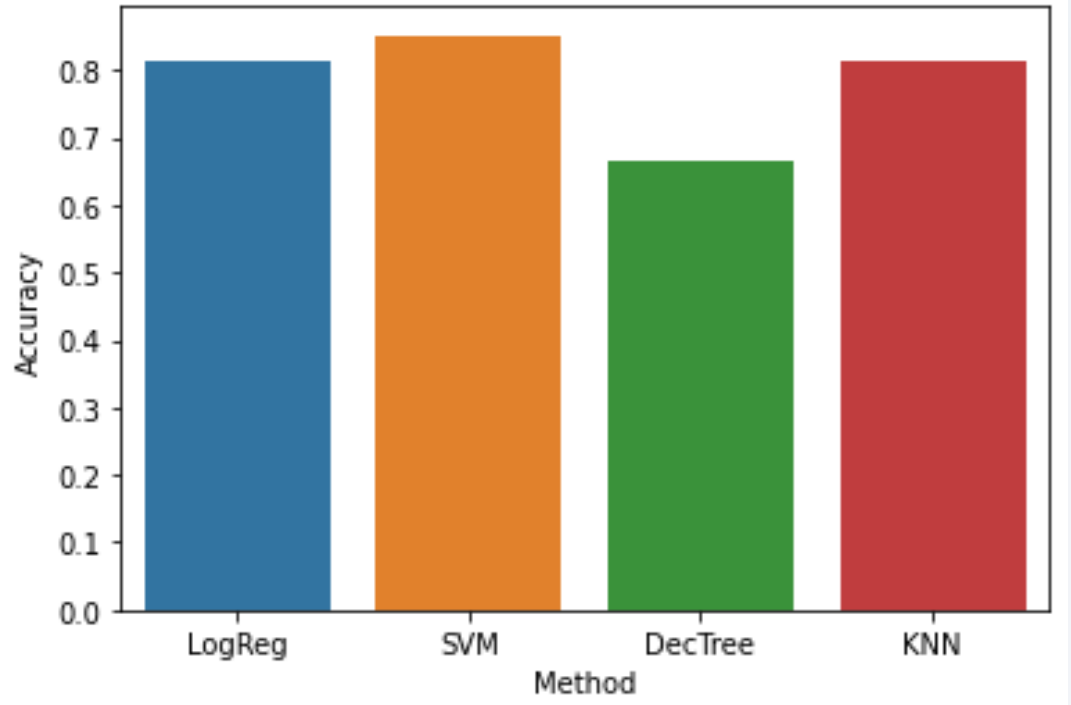
Section 6

Predictive Analysis (Classification)

Classification Accuracy

- We can see that the best accuracy is with SVM model with rate of 85%

	Method	Accuracy
0	LogReg	0.814815
1	SVM	0.851852
2	DecTree	0.666667
3	KNN	0.814815



Confusion Matrix of SVM Model

- We can see that the true landed values was 100% predicted from our model while not landed values was 60% correctly predicted, and as overall, our model is 85% accurate which is a good value

Conclusions

- Our model is acceptable to predict the results of launching based on the history of Falcon 9 rockets from SpaceX
- SpaceY company now have a good model where they can use to have a good revenue and can predict how the rockets will launch

Appendix

- The full GitHub respiratory that contains all work done in this project is here
- For more details regarding IBM Data Science Certificate, visit here

Thank you!