IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ivor Li
09/22/2022

Outline

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

Executive Summary

- **Summary of methodologies**
  - Data is collected through (1) SpaceX launch data collected from SpaceX REST API and (2) Web scrapping from Wikipedia for additional rocket and launch data
  - Data is transformed from the raw data into more readily used formats through various data wrangling steps
  - Exploratory Data Analysis is performed using SQL and Charts that provide better understanding of data and visualizing the relationships among variables
  - Folium is used to perform location and proximity analytics of all launch sites
  - A dashboard application was built with Plotly Dash to perform interactive visual analytics on SpaceX launch Data in real-time
  - Several Classification Machine Learning modes are built and evaluated for the best model to predict the outcome of landing
- **Summary of all results**
  - There are correlations between the launch success and the attributes given in the dataset (Launch Site, Orbit Type, Flight Number, Payload Mass
  - Low pay load weights seems to have higher success rates. Some launch sites performed better than the other (e.g. KSC LC-39A has the successful rate of 76.9%)
  - The success rates for SpaceX has been improving year on year.
  - The launch sites are located with close proximity with coastline, railways, highways but far away from cities
  - Decision Tree Is the best Machine Learning model to predict the success landing based on the data

Introduction

- Project background and context

  The project objective is to 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.

- Problems you want to find answers

  1. To find out those factors that may impact the landing outcome
  2. The relationship between those factors and the landing outcome
  3. The factors that may improve the chance of landing outcome

Section 1

Methodology

Methodology

Executive Summary

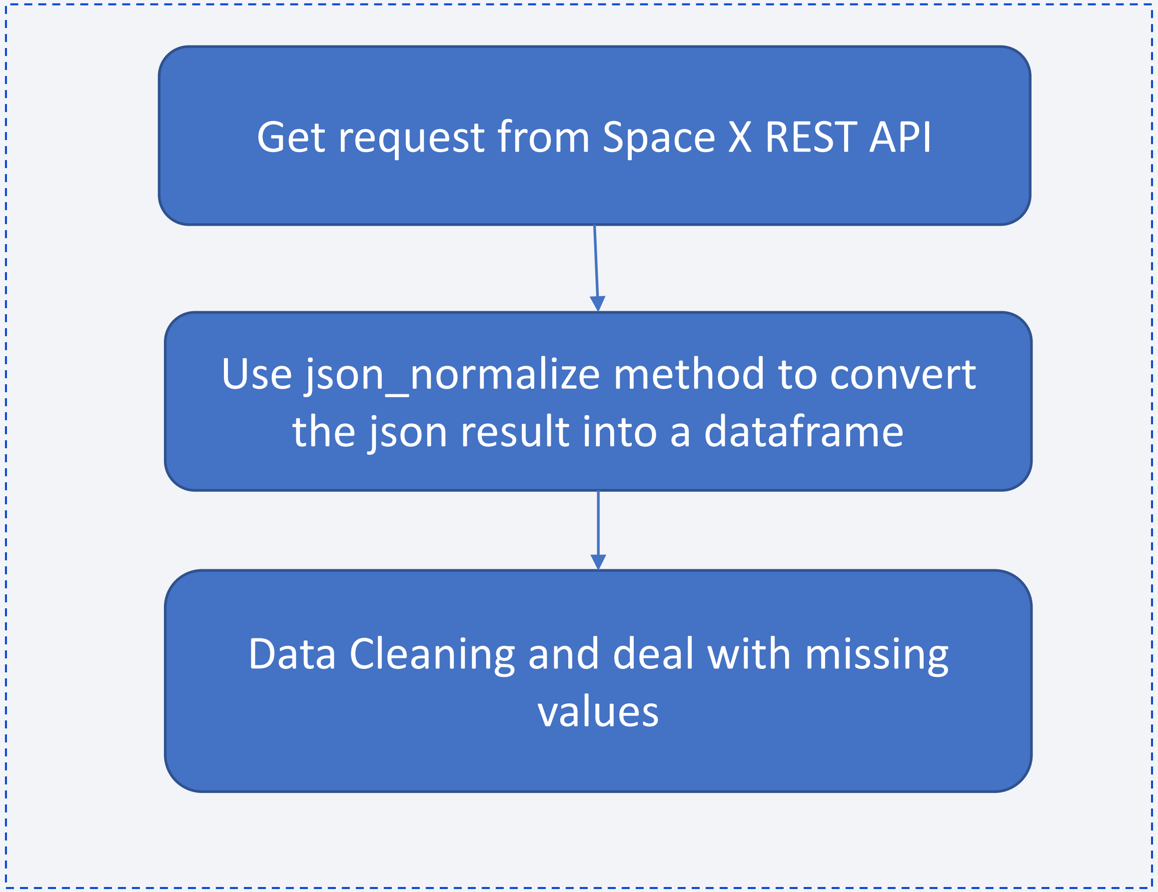
- Data collection methodology:
  1. SpaceX launch data collected from SpaceX REST API
  2. Web scrapping from Wikipedia for additional rocket and launch data
- Perform data wrangling
  - On-hot encoding for categorical features
  - Clean up all Null values
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistics Regression, KNN, SVM, Decision Tree models are used and evaluated

Data Collection – SpaceX API

- Getting SpaceX launch data from SpaceX REST API

- GitHub URL of the completed SpaceX API calls notebook

  https://github.com/ivorli/Falcon_9/blob/master/Falcon%209%20-%20Data%20Collection.ipynb



Data Collection - Scraping

- Web scrapping from Wikipedia for additional rocket and launch data

- GitHub URL of the completed notebook

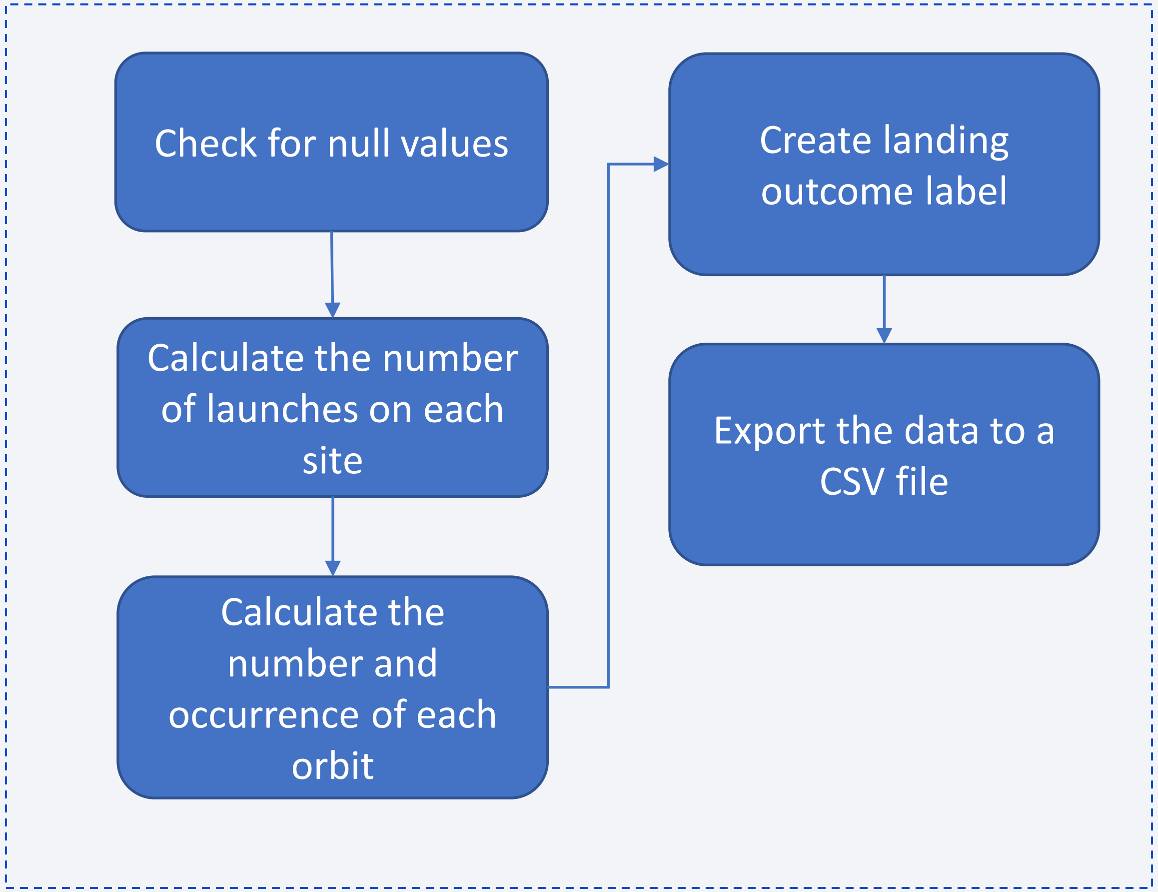
  https://github.com/ivorli/Falcon_9/blob/master/Falcon%209%20-%20Web%20Scrapping.ipynb

Data Wrangling

- transform the raw data into more readily used formats.

- GitHub URL of the completed notebook

  https://github.com/ivorli/Falcon_9/blob/master/Falcon%209%20-%20Data%20Wrangling.ipynb



EDA with Data Visualization

- Scatter plots, bar chart and line plots were used during the EDA process
- **Scatter plot** is a useful to visualize the relationship between attributes
  - Pay load Mass
  - Flight number
  - Launch Site
  - Orbit type
  - Success Rates
- **Bar chart** is useful for plotting the values of an attribute. It is used here to find the successful rate of each Orbit type.
- **Line chart** is excellent to visualize trend. It is used here to show if there is any trend of success rate year on year.

GitHub URL https://github.com/ivorli/Falcon_9/blob/master/Falcon%209%20-%20EDA%20with%20Visualization%20lab.ipynb

EDA with SQL

SQL queries performed to further understand the dataset:

- Display the names of the unique launch sites

  %sql select distinct(LAUNCH_SITE) from SPACEXTBL

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

  %sql SELECT * from SPACEXTBL where Launch_Site LIKE 'CCA%' limit 5

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

  %sql SELECT sum(PAYLOAD_MASS__KG_) FROM SPACEXTBL where Customer = 'NASA (CRS)'

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

  %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL where Booster_Version = 'F9 v1.1

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

  %sql select min(Date) from SPACEXTBL where [Landing _Outcome] = 'Success (ground pad)'

- 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 distinct(Booster_Version) from SPACEXTBL where [Landing _Outcome] = 'Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4000 and 600

- List the total number of successful and failure mission outcomes

  %sql select MISSION_OUTCOME, count(MISSION_OUTCOME) as Count from SPACEXTBL GROUP BY MISSION_OUTCOME;

- List the names of the booster_versions which have carried the maximum payload mass.

  %sql select BOOSTER_VERSION as Booster_Ver from SPACEXTBL where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTBL)

- 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.

  %sql SELECT substr(Date, 6, 2),MISSION_OUTCOME,BOOSTER_VERSION,LAUNCH_SITE FROM SPACEXTBL where substr(Date,1,4) ='2015'

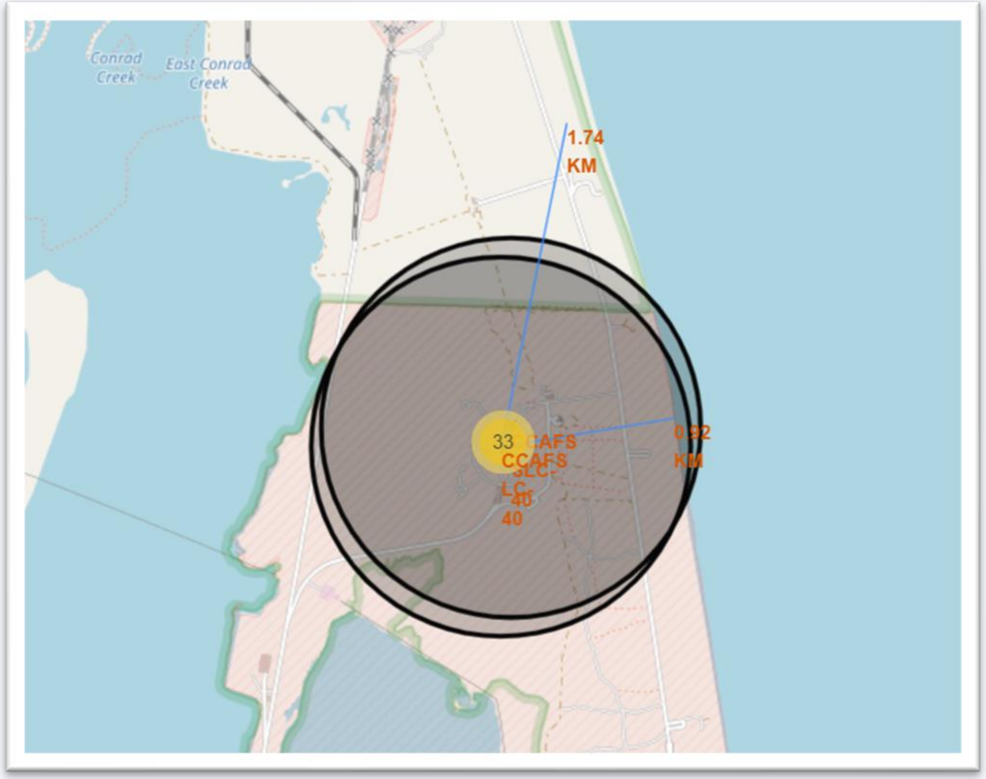
- Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

  %sql SELECT [Landing _Outcome], count([Landing _Outcome]) as Count from SPACEXTBL WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY [Landing _Ou

GitHub URL https://github.com/ivorli/Falcon_9/blob/master/Falcon%209%20-%20EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- The launch success rate may depend on the location and proximities of a launch site. Folium is used here to perform map analytics.
- Latitude and longitude coordinates of launch sites and point of interest are firstly identified from the dataset and the interactive map, following map objects are being added:
  - Markers
  - Circles
  - Lines
- Markers and Circles helped to visualize the locations on map. The outcomes are assigned to different class and marked with Red or Green to distinguish failure or success
- Lines object help to visualize the distance calculated between launch sites and nearest railways, highways, coastline and cities



GitHub URL https://github.com/ivorli/Falcon_9/blob/master/Falcon%209%20-%20Interactive%20Visual%20Analytics%20with%20Folium%20lab%20-%20site%20location.ipynb

Build a Dashboard with Plotly Dash

- A dashboard application was built with Plotly Dash to allow users to perform interactive visual analytics on SpaceX launch Data in real-time
- Below plots were built
    - A pie chart showing success launch rates by site and Booster version
    - A Scatter plot showing the relationship between payload mass, success class and booster version
- Below interactions were built
    - A drop-down input for sites
    - A range slider where users can choose the payload range
- The above setup would tell us success launch rates from different sites, payload mass range and booster version

GitHub URL https://github.com/ivorli/Falcon_9/blob/master/Falcon%209%20-%20spacex_dash_app.py

Predictive Analysis (Classification)

Machine learning pipeline to predict if the first stage will land

Model Building
- Load the dataframe
- Create class labels / Standardize data
- Split train and test data set
- Create model, fit the train data

Model Evaluation
- Calculate the accuracy
- Create confusion matrix

Model Improvement
- Parameters tuning using GridSearchCV
- Re-evaluate the model with tuned parameters

Best Model
- Compare the accuracy and prediction score of each model to find the best one

GitHub URL https://github.com/ivorli/Falcon_9/blob/master/Falcon%209%20-%20SpaceX_Machine%20Learning%20Prediction.ipynb

14

Results

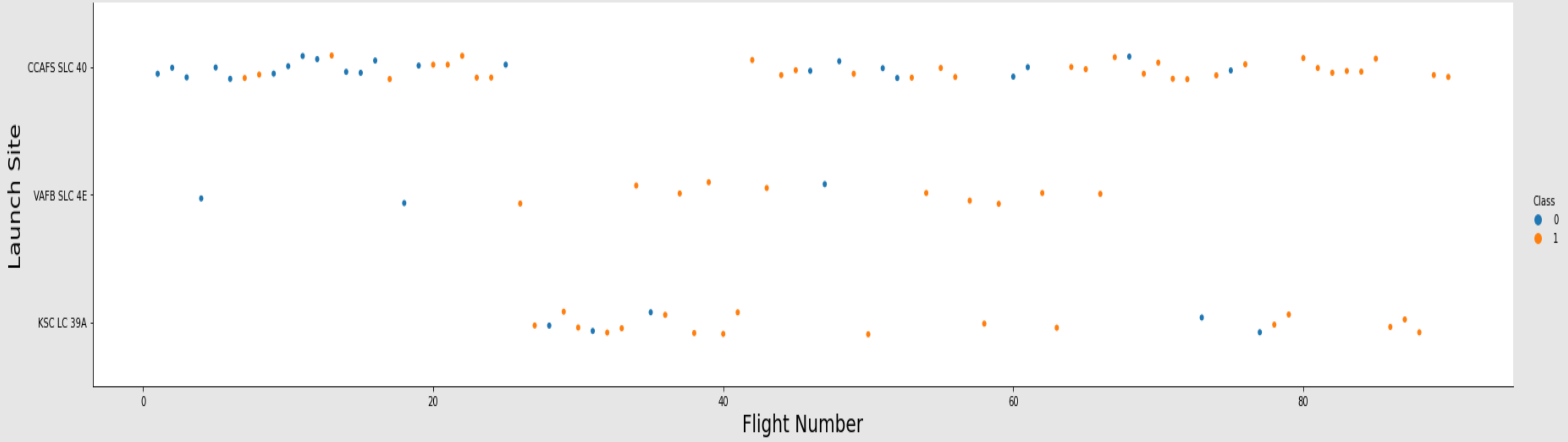
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

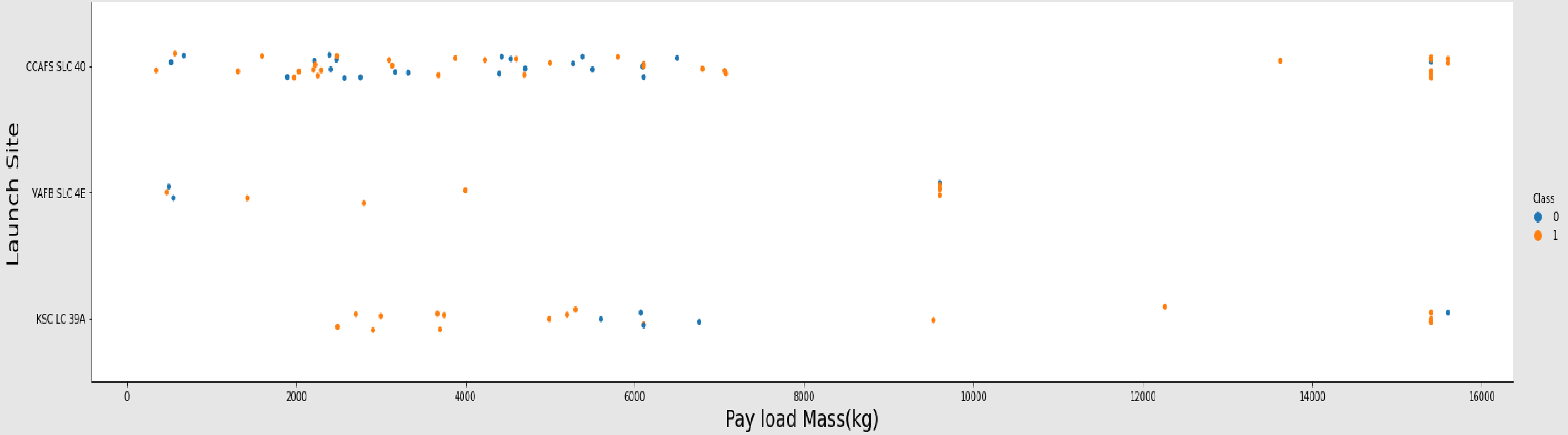
Flight Number vs. Launch Site

Launch Site KSC LC-39A and VAFB SLC 4E has a higher success rate when the flight number is larger



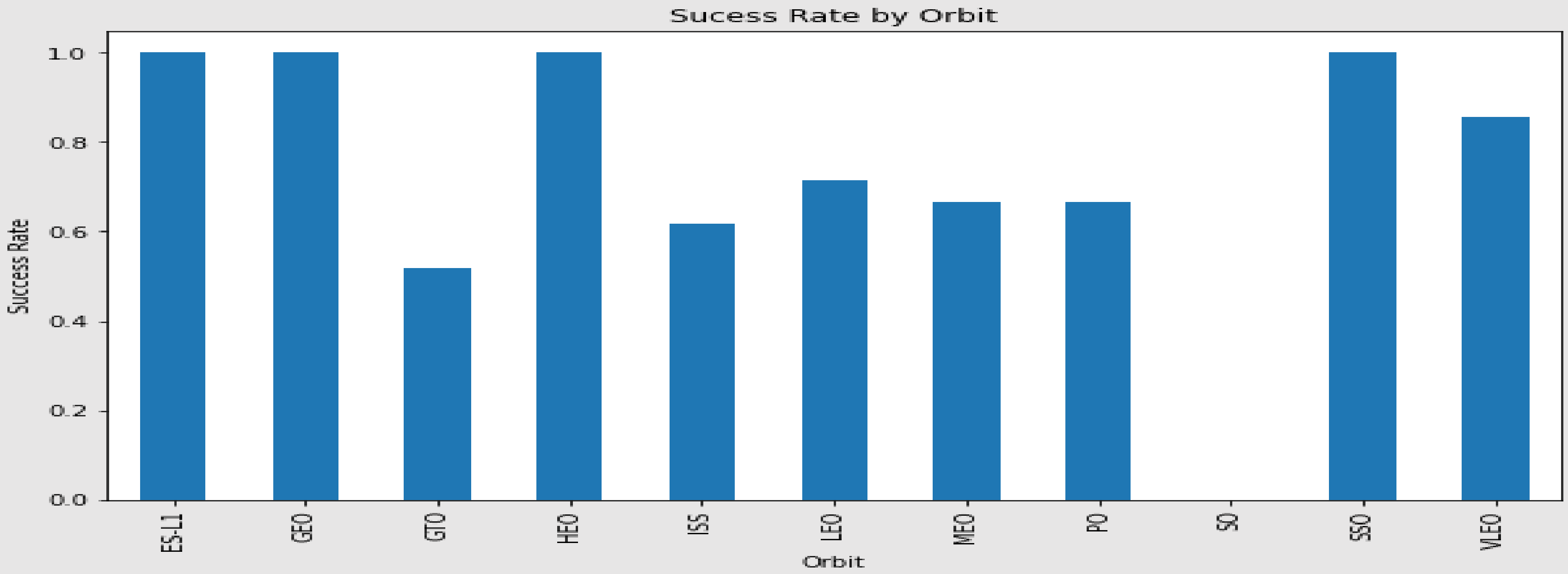
Payload vs. Launch Site

- Successful rate is much higher when the pay load mass is over 7000kg
- VAFB-SLC does not have heavy pay load mass



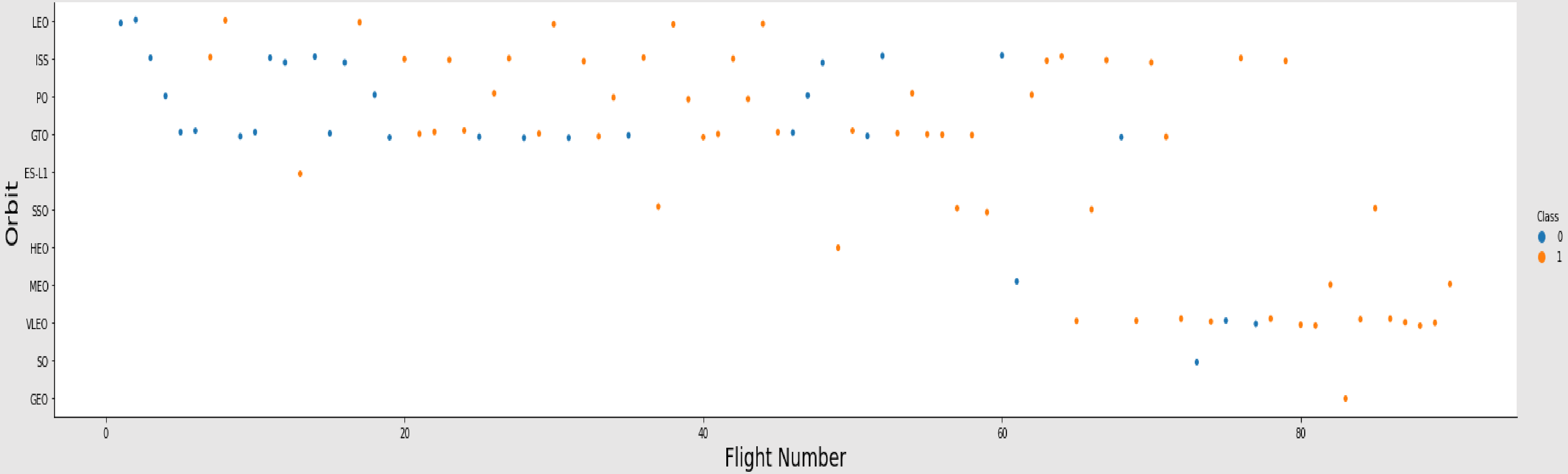
Success Rate vs. Orbit Type

There are strong relationship between success rate and orbit types. ESL1, GEO, JEO, SSO have 100% successful rates while other obit varies.



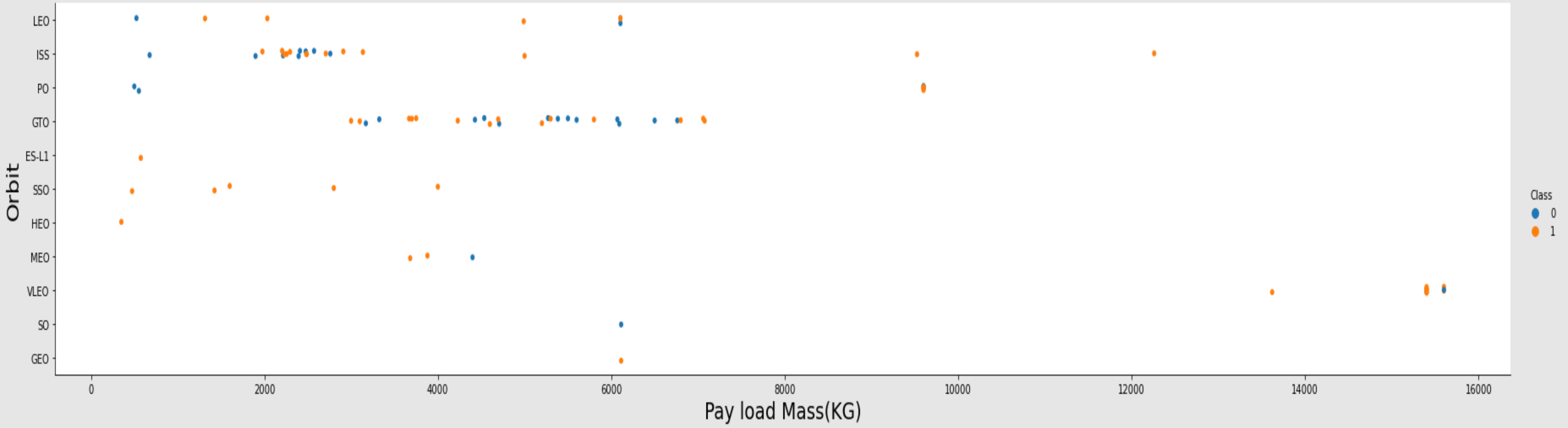
Flight Number vs. Orbit Type

In general, the larger the flight number, the higher the success rate. GTO does not show the same pattern.



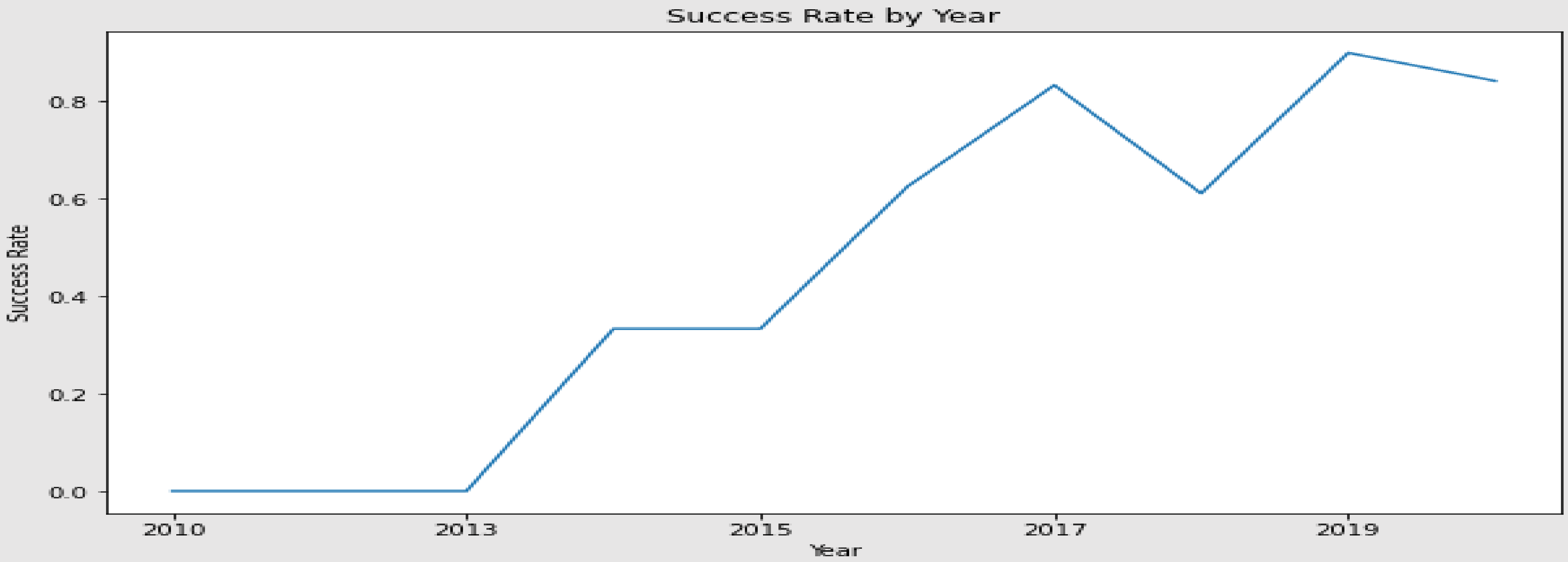
Payload vs. Orbit Type

- Heavier payload show higher success rate for LEO, ISS and PO orbit
- The other orbits do not show obvious patterns



Launch Success Yearly Trend

- Success rates have been improving year on year, with slight drop in 2018.



All Launch Site Names

- DISTINCT is used to select the unique name of the launch site. There are 4 in total:

```
%sql select distinct(LAUNCH_SITE) from SPACEXTBL
```

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

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

Launch Site Names Begin with 'CCA'

- Select launch site name only begin with “CCA”

```
%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
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- The total payload mass is 45596KG where customer is NASA(CRS)

```
%sql SELECT sum(PAYLOAD_MASS__KG_) FROM SPACEXTBL where Customer = 'NASA (CRS)'
```

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

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- The payload mass is 2928.4kg, for booster version F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL where Booster_Version = 'F9 v1.1'
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Min(Date) will give the “oldest” date result
- The date of the first successful landing outcome is 2015-12-22

```
%sql select min(Date) from SPACEXTBL where [Landing _Outcome] = 'Success (ground pad)'
```

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

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
%sql select distinct(Booster_Version) from SPACEXTBL where [Landing _Outcome] = 'Success (drone ship)' and PAYLOAD_MASS__KG_ BETW
```

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

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

Total Number of Successful and Failure Mission Outcomes

- There are total 100 success and 1 failure mission outcomes

```
%sql select MISSION_OUTCOME, count(MISSION_OUTCOME) as Count from SPACEXTBL GROUP BY MISSION_OUTCOME;
```

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

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

Boosters Carried Maximum Payload

- Booster which have carried the maximum payload mass

```
%sql select BOOSTER_VERSION as Booster_Ver from SPACEXTBL where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

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

Booster_Ver
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

```
%sql SELECT substr(Date, 6, 2),[LANDING _OUTCOME],BOOSTER_VERSION,LAUNCH_SITE FROM SPACEXTBL where substr(Date,1,4) ='2015' AND
```

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

substr(Date, 6, 2)	Landing _Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	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

```
%sql SELECT [Landing _Outcome], count([Landing _Outcome]) as Count from SPACEXTBL WHERE Date BETWEEN '2010-06-04' AND '2017-03-2(
```

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

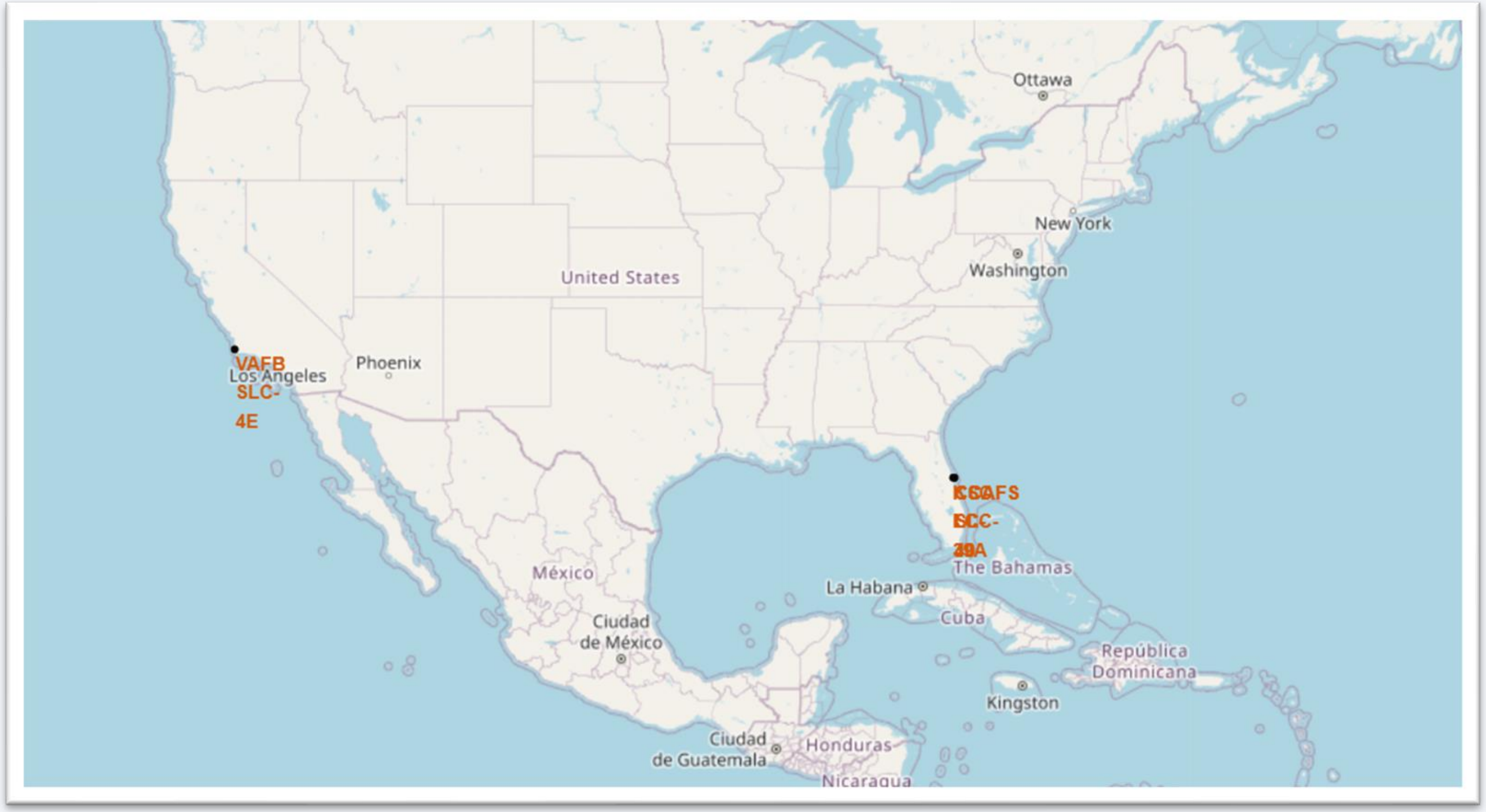
Landing _Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

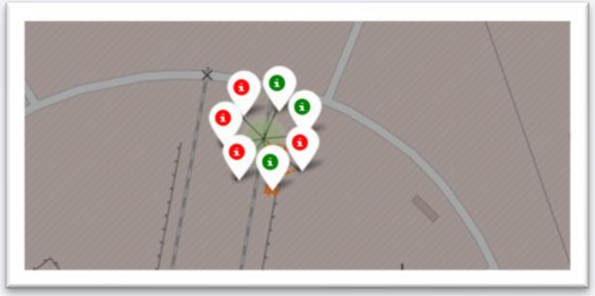
Location of all SpaceX Launch Sites

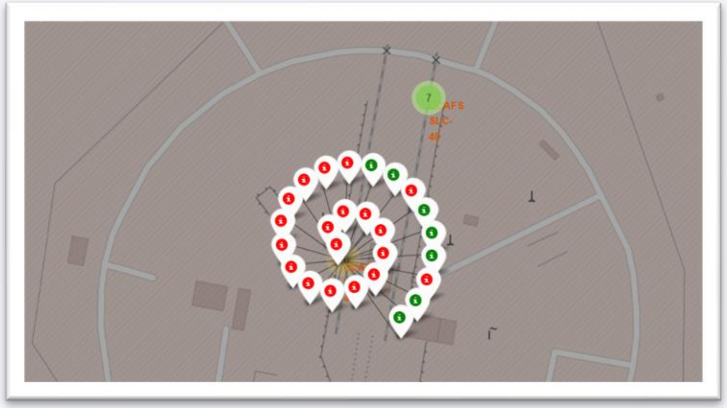
- Launch sites are either located in Eastern or Western along the coastline of the USA

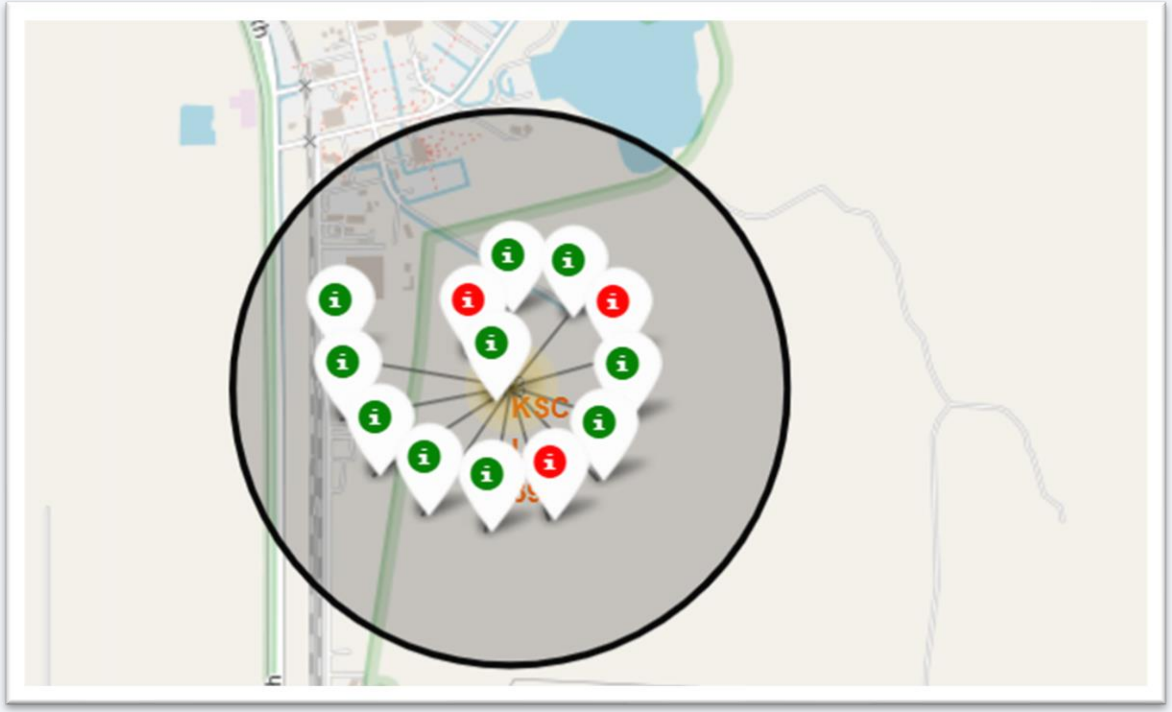


Launch Outcomes of Sites

- Green marker indicate launch success while red failure
- This provide visualized outcomes of each site





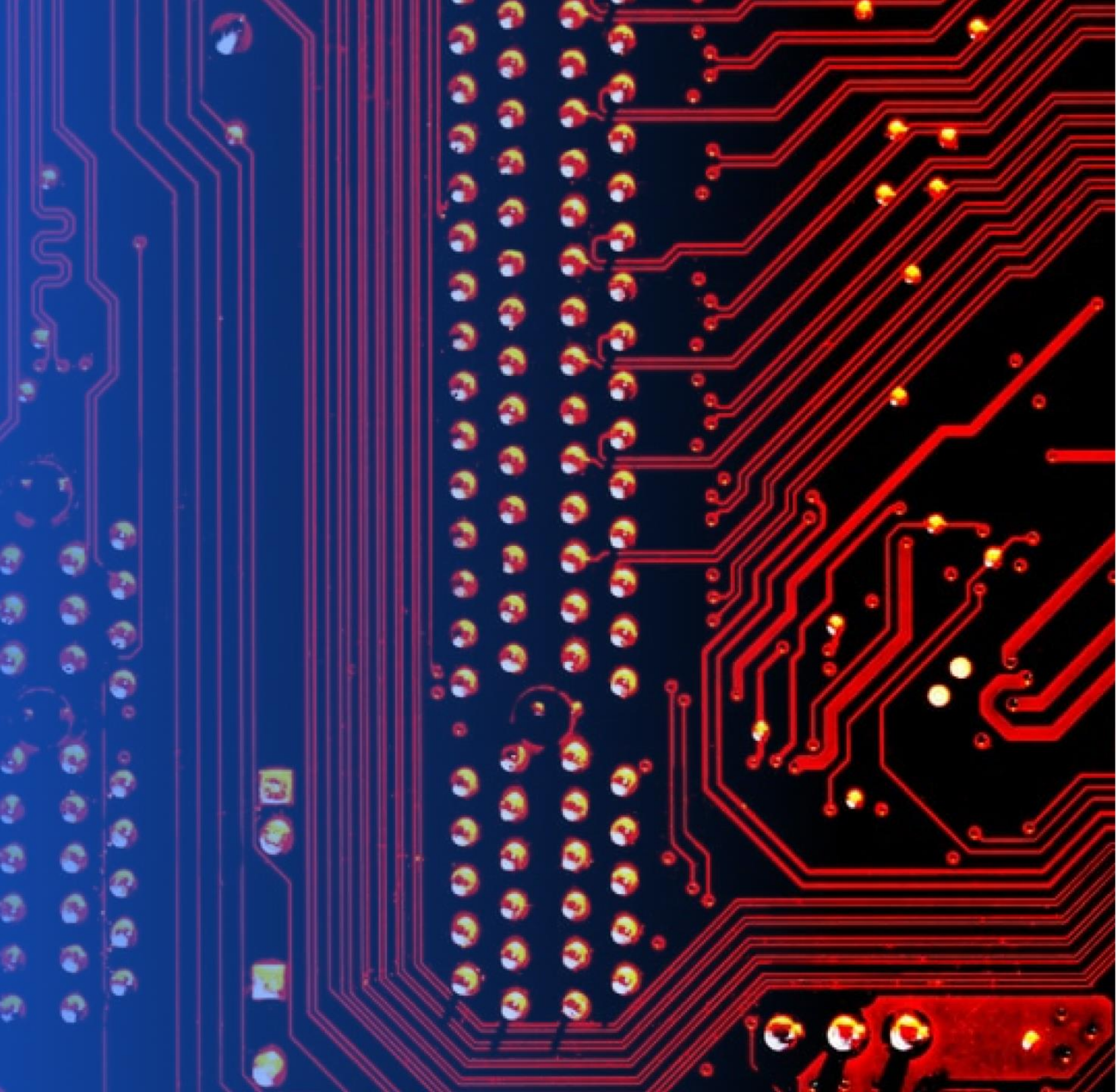


Launch Sites Proximities

- Concluded findings - Launch sites are in close proximity to railways, highways, coastline but far away from cities

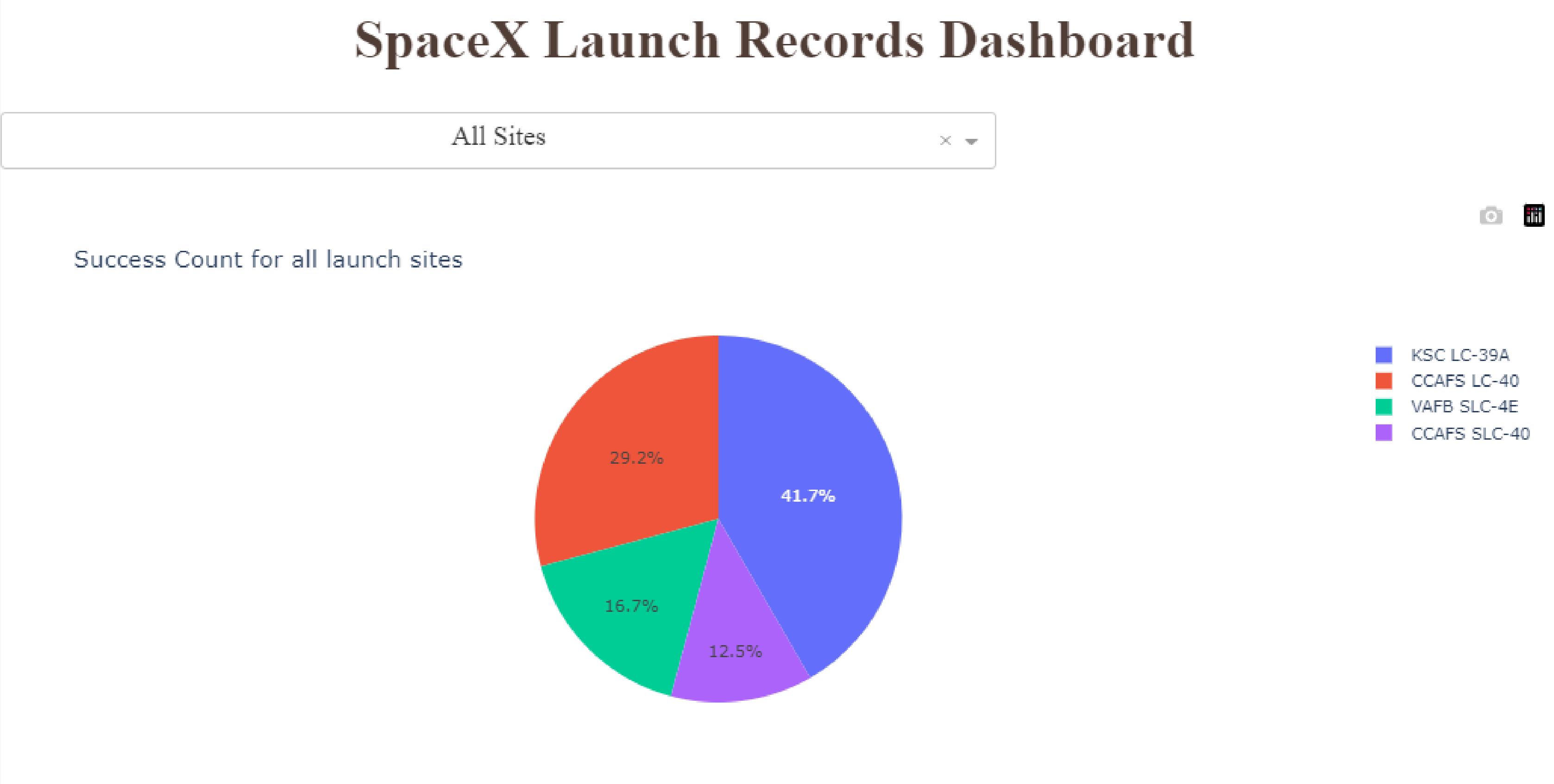
Section 4

Build a Dashboard with Plotly Dash



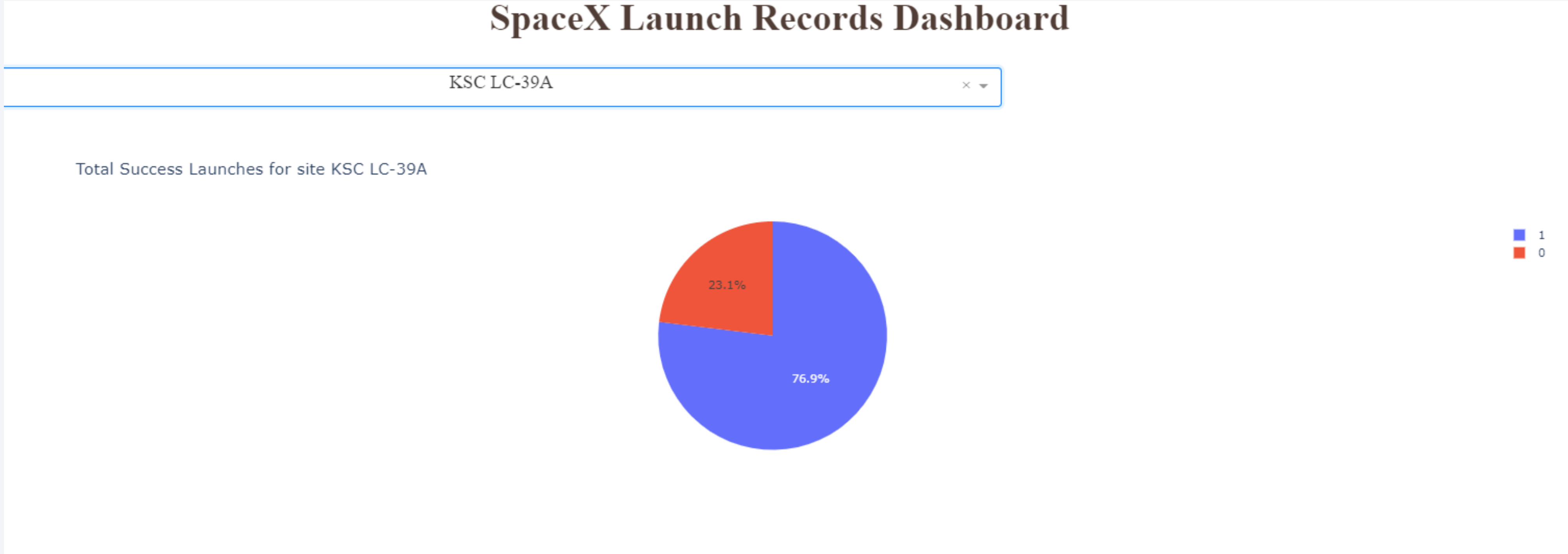
Launch Success Count of All Sites

- Below pie chart show the success launch count of each site. KSC LC-39A has most success launch count



The Highest Success Launch Rate Site

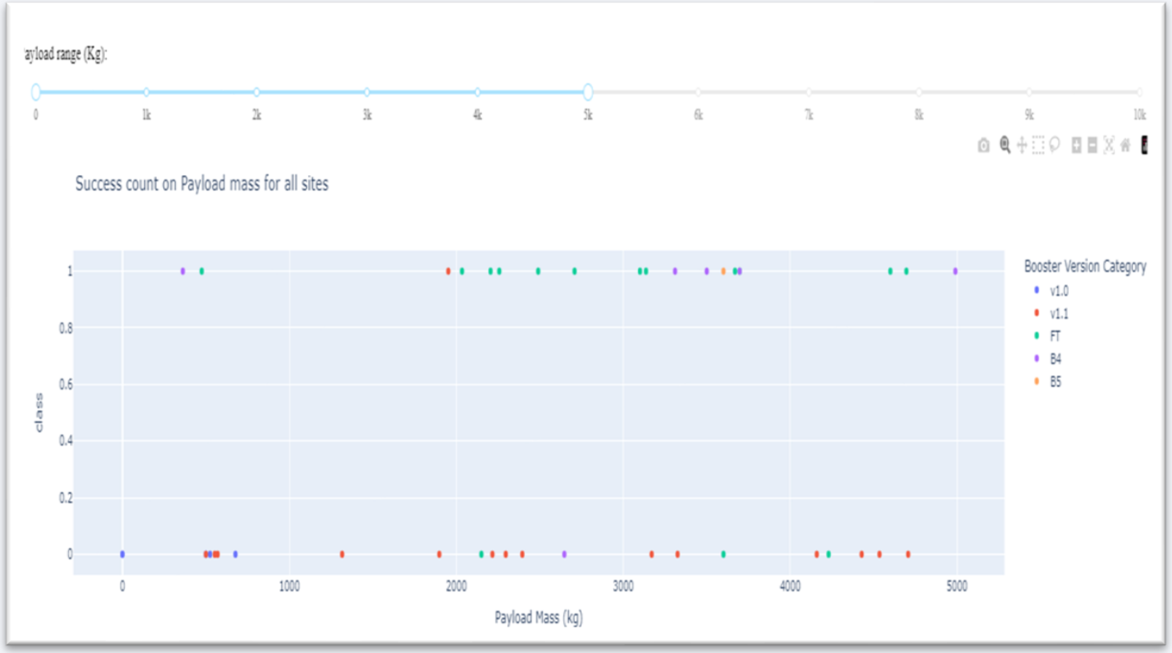
- KSC LC-39A has the highest 76.9 % launch success rate



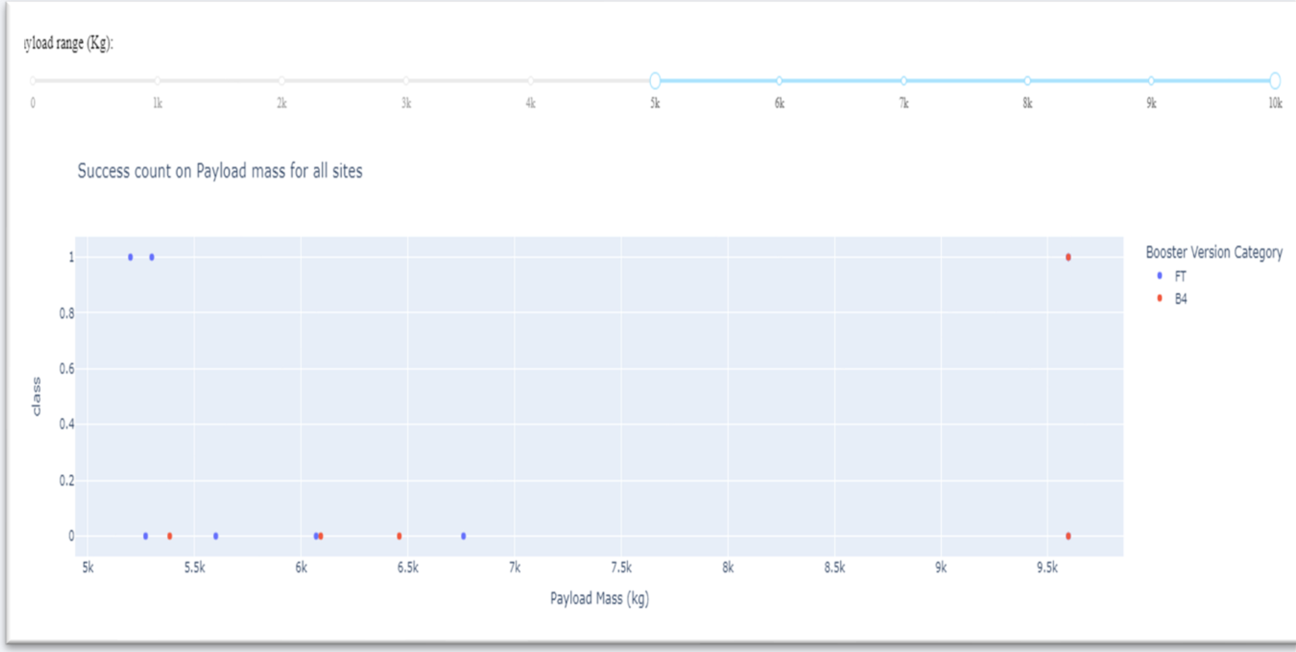
Payload Mass vs Launch Outcome

- It looks like lower payload mass has higher success rate

0 – 5000 kg



5000 – 10000 kg

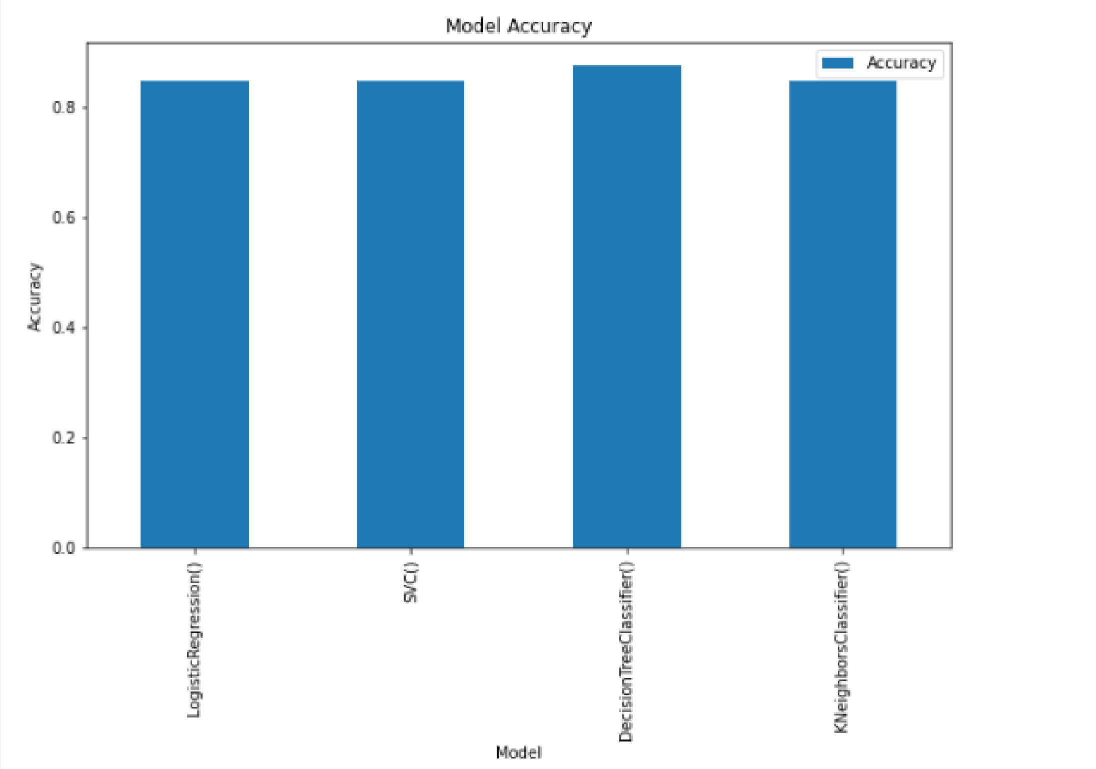


Section 5

Predictive Analysis (Classification)

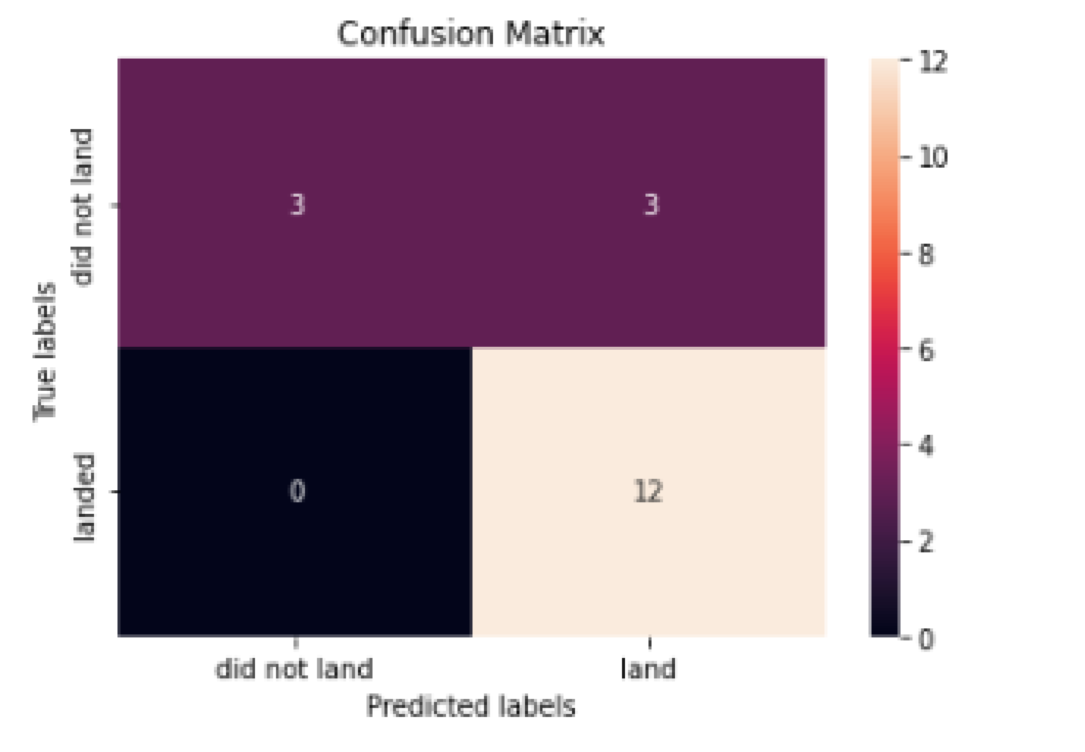
Classification Accuracy

- Decision Tree has the highest accuracy among all models used



Confusion Matrix

- Confusion Matrix of the Decision Model – the problem is the false positives (predict success while the true label is failure)



Conclusions

- There are correlations between the launch success and the attributes given in the dataset (Launch Site, Orbit Type, Flight Number, Payload Mass
- Low pay load weights seems to have higher success rates. Some launch sites performed better than the other (e.g. KSC LC-39A has the successful rate of 76.9%)
- The success rates for SpaceX has been improving year on year.
- The launch sites are located with close proximity with coastline, railways, highways but far away from cities
- Decision Tree Is the best Machine Learning model to predict the success landing based on the data

Thank you!