IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ryo Taono
06/05/2023

Outline

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

Executive Summary

- **Summary of methodologies**
  - **Data Collection** – Using SpaceX API and web scraping techniques
  - **Data Wrangling** – Conducting Exploratory Data Analysis by manipulating data in Pandas data frame
  - **Data Visualization** – Using "Folium, Dashboard, Plots" to extract meaningful patten to guide the modeling process and analyze the launch site proximity
  - **Machine Learning Prediction** – Find best hyperparameter for Support Vector Machine, Classification Trees and Logistics Regression and build a predictive model
- **Summary of all results** (Booser Version - Falcon9)
  - Along with a coastline
  - All launch sites have higher success rate with higher payload mass (kg)
  - Success rate increased since 2013 until 2020

  - Success rate varies by launch sites – VAFB SLC 4E 70%

Introduction

Backgound

- 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 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.

- We would like to find the key factors of successful landings

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using web scraping techniques and SpaceX API
- Perform data wrangling
  - Preprocess the collected data by SQL and Pandas
- 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 - SpaceX API

- Import requests library to make HTTP requests which we will use to get data from API
- Create functions to extract the booster names, launch site' information , the payload mass, the orbit, out outcome of the launch.
- Request and parse the SpaceX launch data using the GET request
- Use json normalize method to convert the json result into a dataframe
- Combine the columns into the dictionary
- Create and clean the dataframe
- Export as a csv file

Data Collection – Web Scraping

- Request the Falcon9 launch Wiki page from its URL
- Create a BeautifulSoup object from the HTML response
- Extract all column/variable names from the HTML table header
- Create a dataframe by parsing the launch HTML tables
- Fill up the dictionary with the extracted records
- Export as a csv file

Data Wrangling

- Load and clean SpaceX dataset from last section
- Calculate # of launches on each site, of each orbit, and of mission outcome per orbit type
- Create a landing outcome label from Outcome column
- Export as a csv file

- https://github.com/17yo17/spacex/blob/master/data-wrangling.ipynb

EDA with Data Visualization

Charts

- Flight # vs. Payload Mass (kg)
- Flight # vs. Launch Site
- Payload Mass (kg) vs. Launch Site
- Payload Mass (kg) vs. Orbit Type
- https://github.com/17yo17/spacex/blob/master/eda-dataviz.ipynb

EDA with SQL

Display

- Names of the launch sites
- 5 records of the launch site begins with “CCA”
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1.

List

- Date of first successful landing outcome in ground pad was achieved
- Names of the boosters which have success in drone ship and payload mass greater than 4000 but less than 6000
- Total # of successful and failure mission outcomes
- Names of the booster versions which have carried the maximum payload mass
- Rank the count of successful landing outcomes in descending order

https://github.com/17yo17/spacex/blob/master/eda-sql.ipynb

Build an Interactive Map with Folium

- Circle all launch sites with their names and mark the success/failed launches for each site. Finally, calculate the distances between a launch site to its proximities along with a line and the calculated distance.
- These objects facilitates us to observe common features of a launch site as well as success rate of each site
- https://github.com/17yo17/spacex/blob/master/folium_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Dropdown list with launch sites
- Pie chart which displays successful rate in percentage of the total
- Scatter plot which displays ”Class vs. Payload Mass (kg)” along with range slider for Payload Mass (kg)

https://github.com/17yo17/spacex/blob/master/dashboard-spacex-launch-site.ipynb

Predictive Analysis (Classification)

Conduct EDA and Determine Training Labels

- Create a column for the class
- Standardize the data
- Split into training data and test data

Find Best Hyperparameter for SVM, Classification Trees, and Logistic Regression

- Passing GridSearchCV C-value: [0.01, 0.1, 1], penalty: l2, solver: lbfgs, estimator, cv: 10

Calculate the accuracy for each model on the test data

- Get the score with a confusion matrix to see that the model can distinguish between the different classes

Determine the best model

- Used Jaccard, F1, and the result from above

https://github.com/17yo17/spacex/blob/master/Machine_Learning_Prediction.ipynb

Results

Exploratory data analysis results

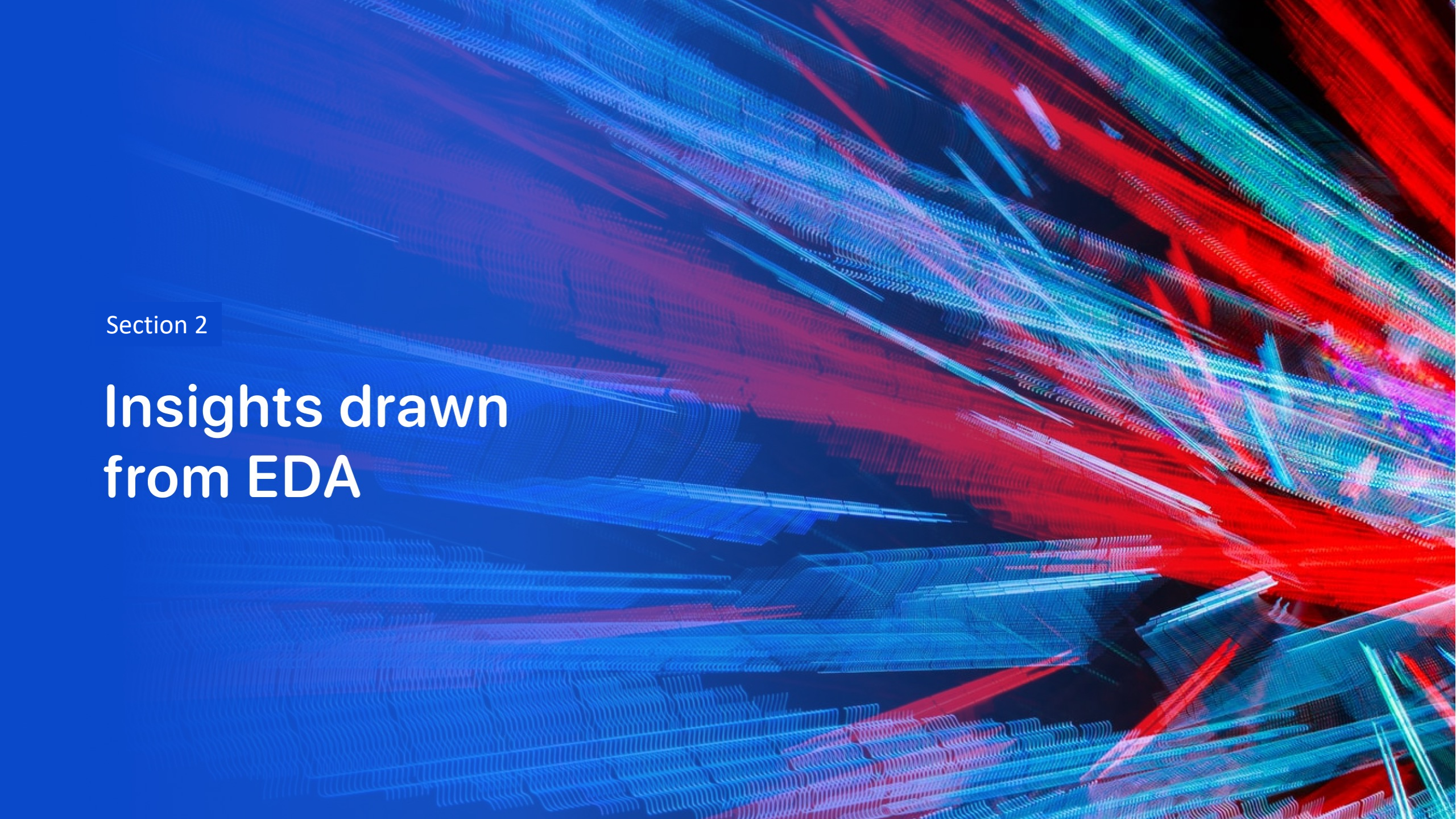
- Launch success has improved since 2013
- KSC LC-39A has the highest success rate
- Orbit type, ES-L1, GEO, HEO, and SSO have a 100% success rate whereas SO has a 0% success rate

Interactive analytics demo in screenshots

- All the sites are close to the coast, highway, railroad but away from cities

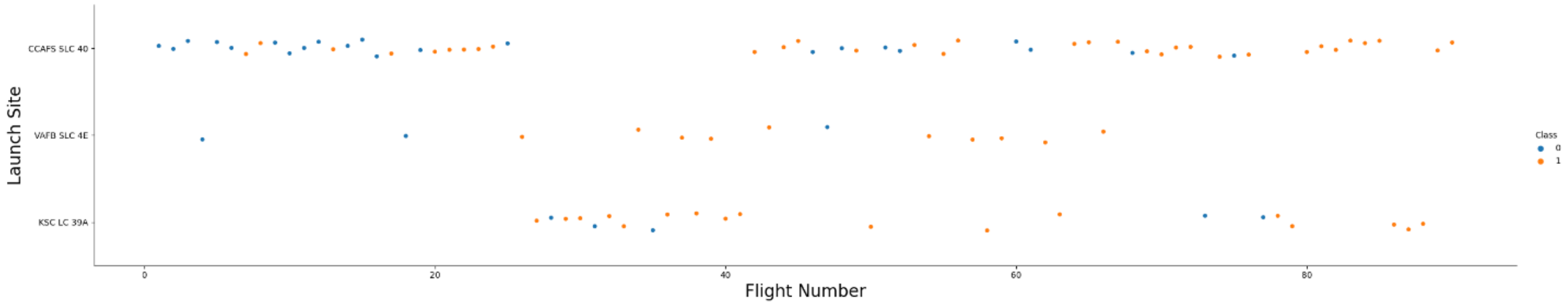
Predictive analysis results

- Decision tree model is the best predictive model for the dataset


Section 2
Insights drawn from EDA

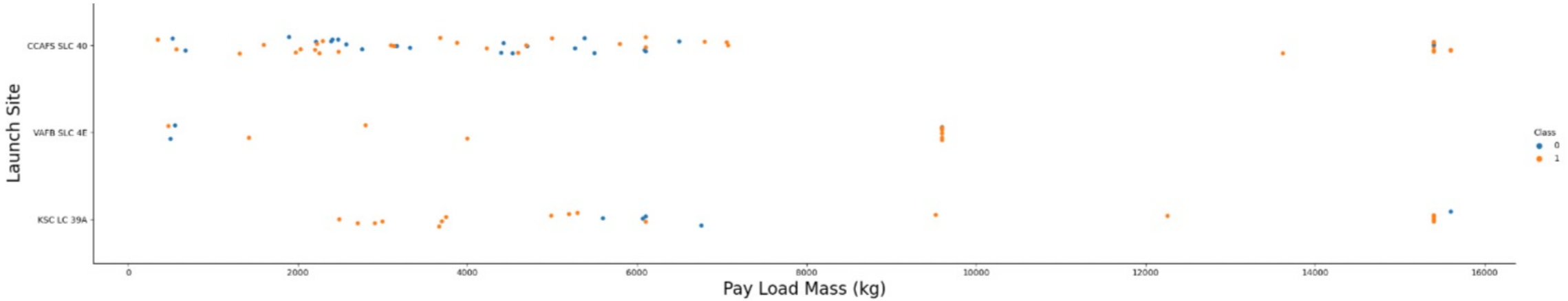
Flight Number vs. Launch Site

- Blue dots: unsuccessful landing, Orange dots: successful landing
- The higher successful landing with higher flight number over all, vice versa
- KSC LC 39A has no record flight # less than 25



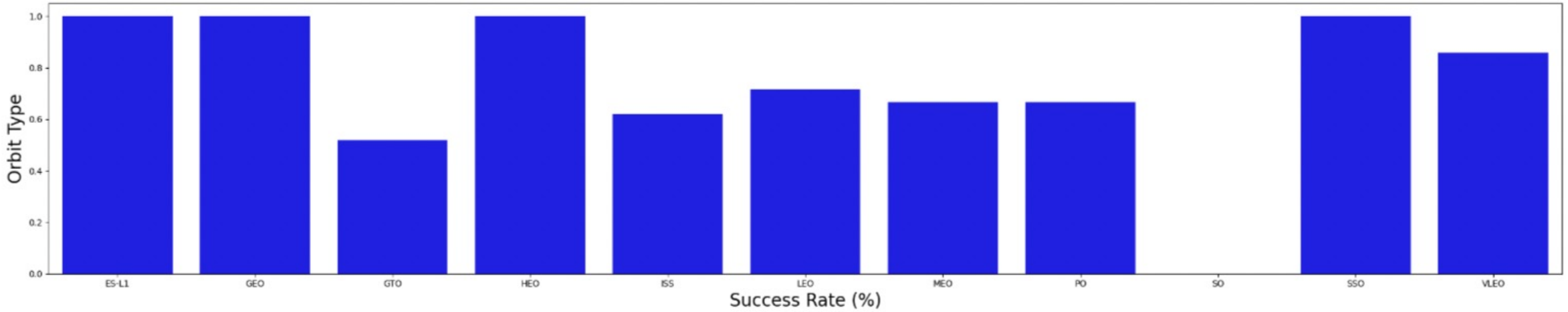
Payload vs. Launch Site

- There is no rockets launched for more than 10000 kg payload mass for the VAFB-SLC launch site
- Higher unsuccessful landing rate between 5000 ~ 7000 kg
- Higher successful rate over 7000 kg



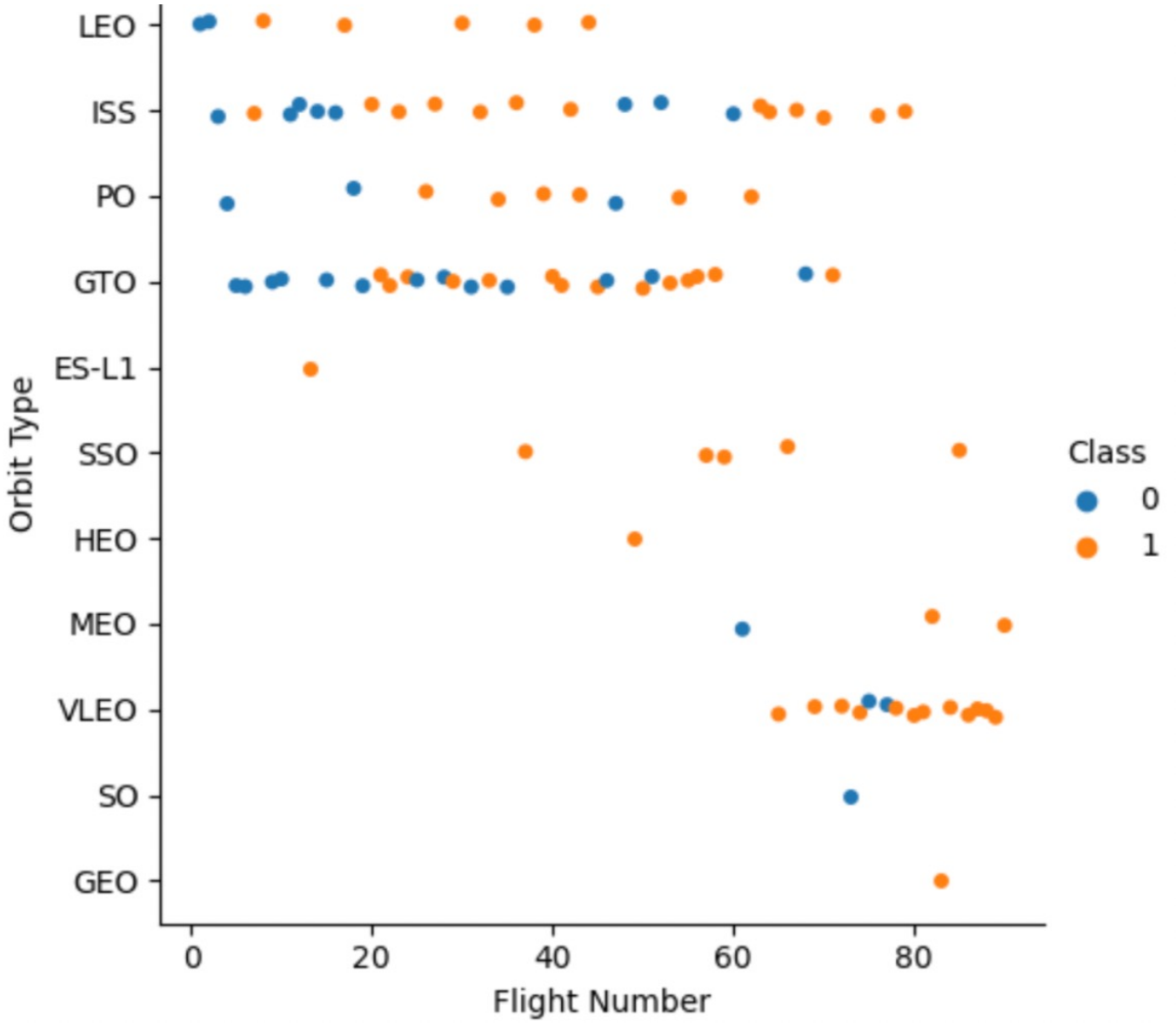
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, SSO with 100% success rate
- SO with 0% success rate



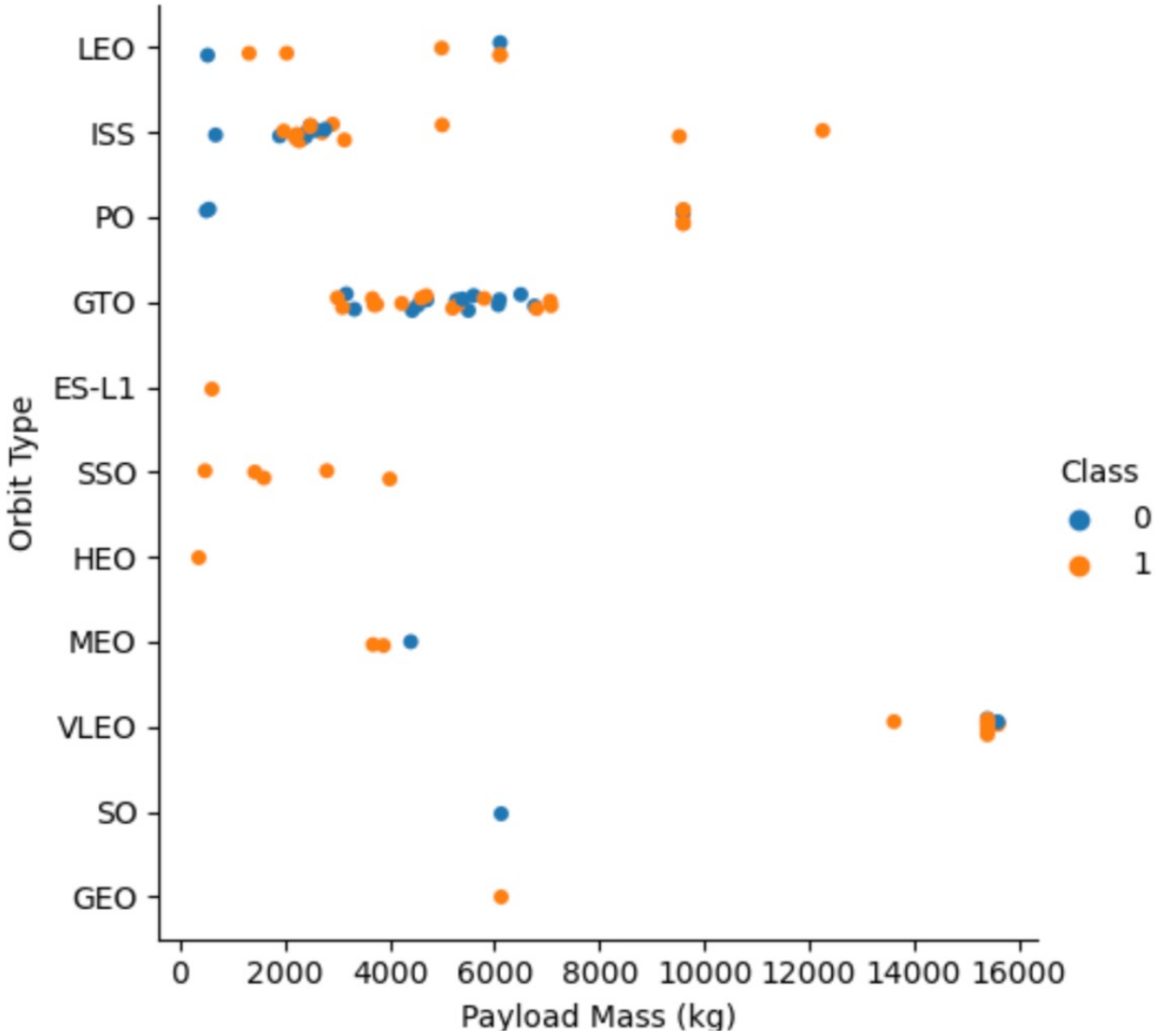
Flight Number vs. Orbit Type

- The Higher unsuccessful rate when the flight # is less than 20
- For LEO orbit, the rate related to the flight #
- For GTO orbit, there is no correlation



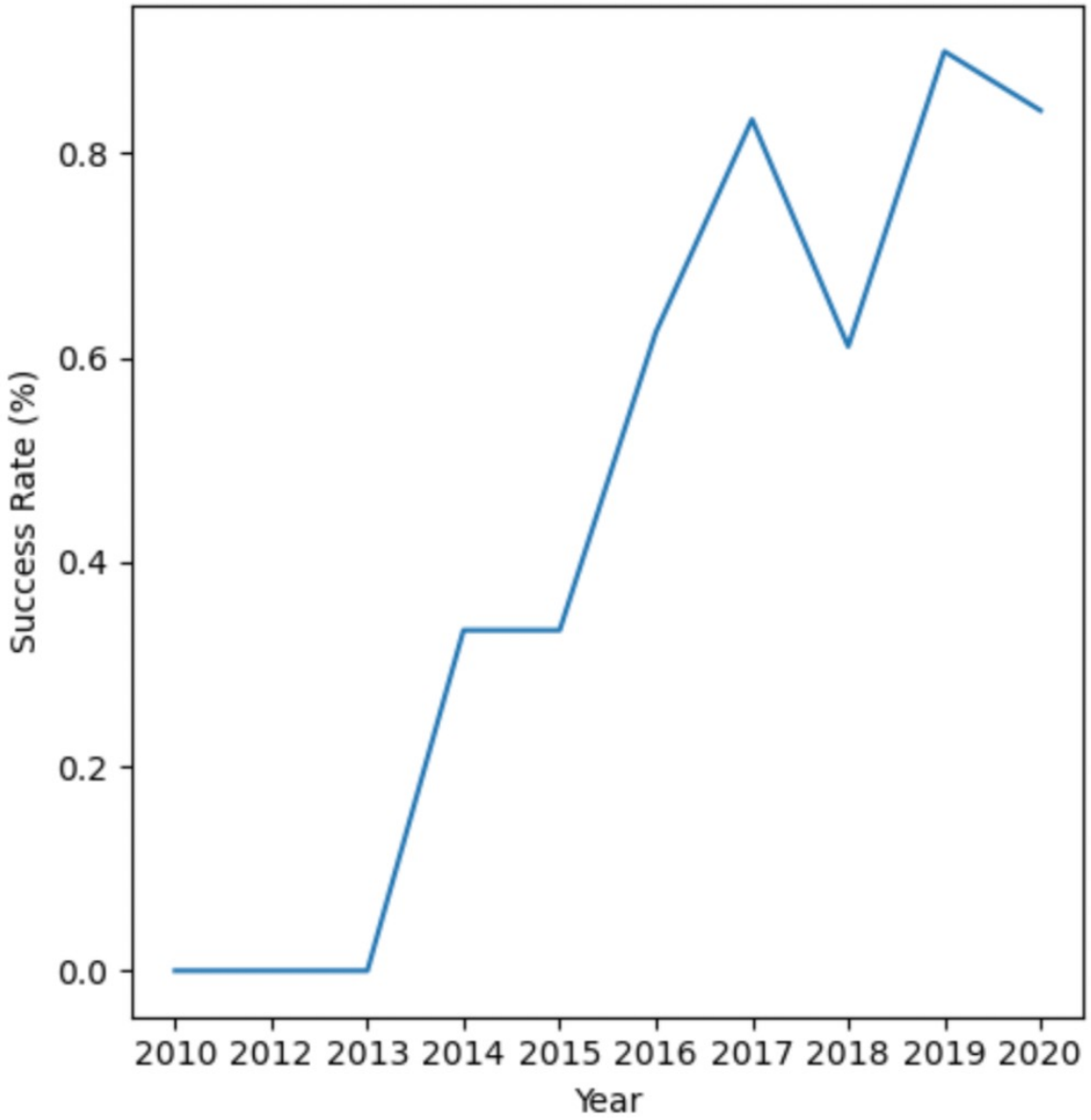
Payload vs. Orbit Type

- With heavy oayloads the successful landing rate are more for PO, LEO, and ISS
- For GTO, we can’t determine



Launch Success Yearly Trend

- The successful landing rate kept increasing since 2013 until 2020



All Launch Site Names

```
%sql select distinct Launch_Site from SPACEXTBL
```

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

- We can see that there are records with no launch site information

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXTBL where Launch_Site like 'CCA%' limit 5
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	NASA (CRS)	Success	No attempt

Total Payload Mass

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer = 'NASA (CRS)'
```

sum(PAYLOAD_MASS__KG_)
45596.0

- The total payload carried by boosters from NASA

Average Payload Mass by F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version = 'F9 v1.1'
```

avg(PAYLOAD_MASS__KG_)
2928.4

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

First Successful Ground Landing Date

```
%sql select * from SPACEXTBL where Landing_Outcome = 'Success (ground pad)' and Date = (select min(Date) \
from SPACEXTBL where Landing_Outcome = 'Success (ground pad)')
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome
01/08/2018	1:00:00	F9 B4 B1043.1	CCAFS SLC-40	Zuma	5000.0	LEO	Northrop Grumman	Success (payload status unclear)

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select distinct Booster_Version from SPACEXTBL \
where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000
```

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

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

Total Number of Successful and Failure Mission Outcomes

```
%sql select Mission_Outcome, count(*) from SPACEXTBL group by Mission_Outcome
```

Mission_Outcome	count(*)
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- Successful most of the times

Boosters Carried Maximum Payload

```
%sql select Booster_Version from SPACEXTBL \
where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

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

- Names of the booster_versions which have carried the maximum payload mass.

2015 Launch Records

```
%sql select substr(Date,4,2) as month, Landing_Outcome, Booster_Version, Launch_Site \
from SPACEXTBL \
where substr(Date,7,4) = '2015' and Landing_Outcome = 'Failure (drone ship)'
```

month	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- List of the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

```
%sql select Landing_Outcome, count(*) as count_success from SPACEXTBL \
where Landing_Outcome like 'Success%' and Date between '04-06-2010' and '20-03-2017' \
group by Landing_Outcome order by count_success desc
```

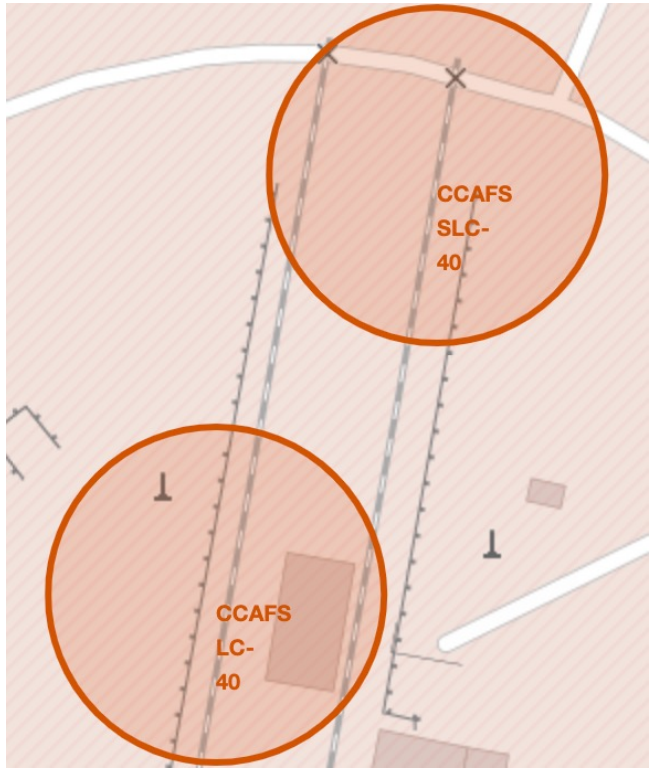
Landing_Outcome	count_success
Success	20
Success (drone ship)	8
Success (ground pad)	7

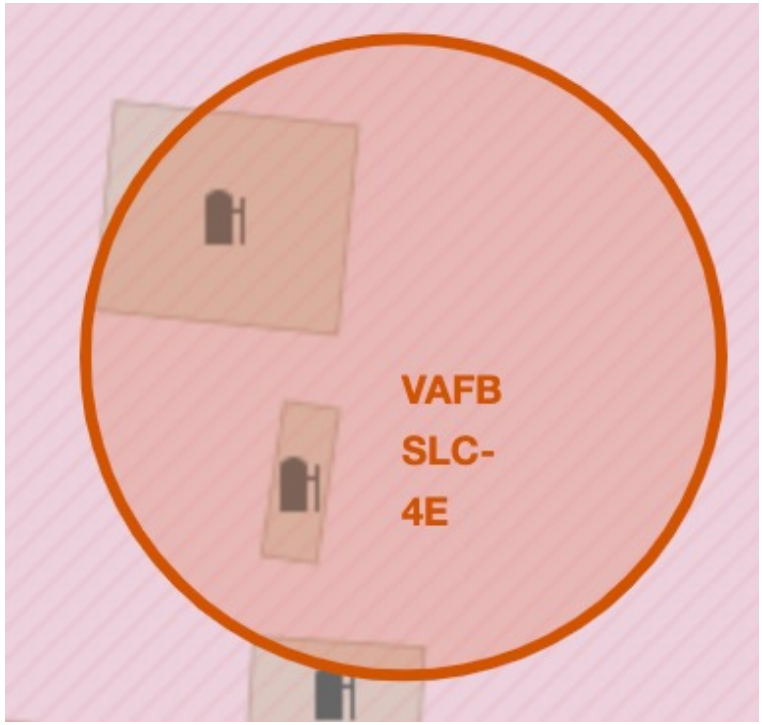
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

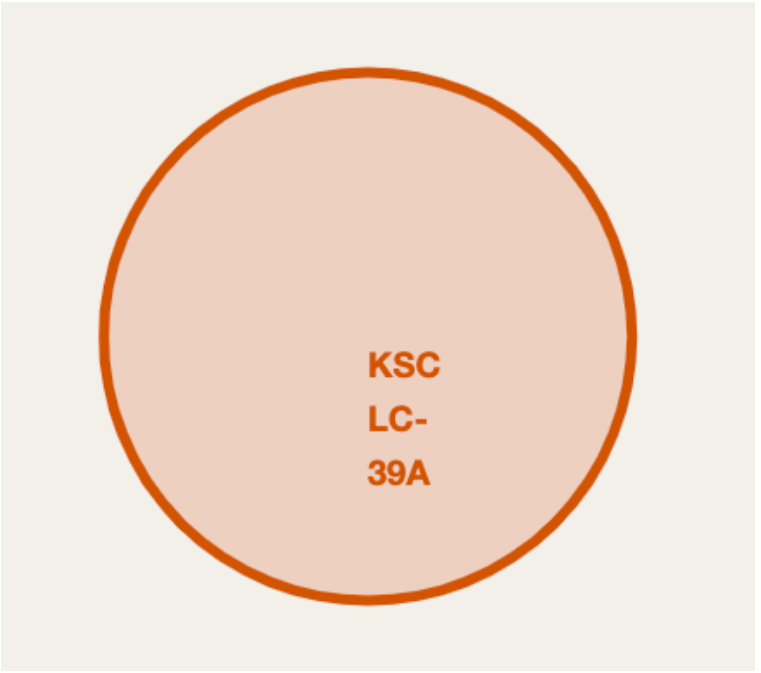

Section 3
Launch Sites
Proximities Analysis

Name Each Site with a Circle

- Tag a name each site with a circle on the map

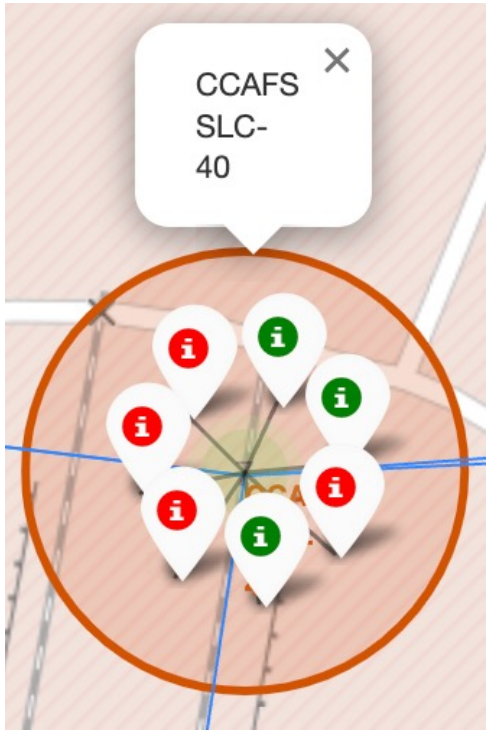


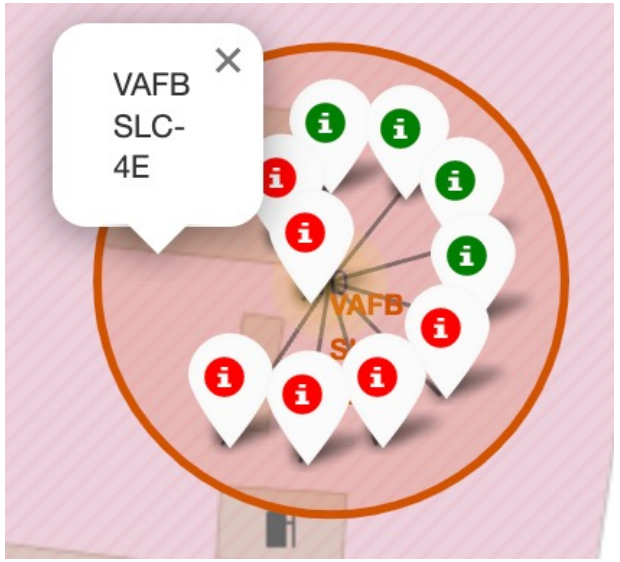


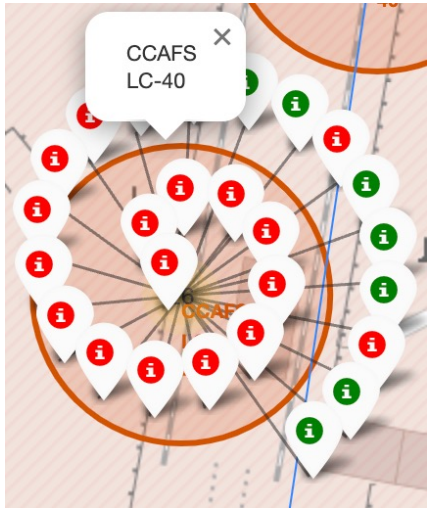


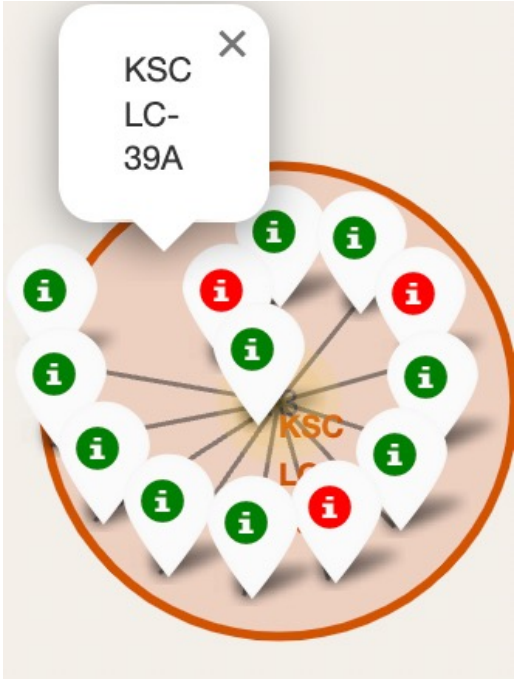
Mark Success/Failed Launches for Each Site

- Green tag as a success and Red tag as a failure
- You can see that KSC LC-39A has a higher success launch rate



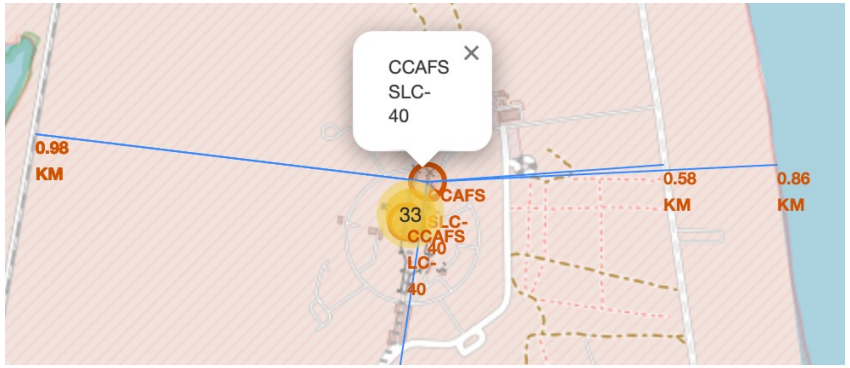


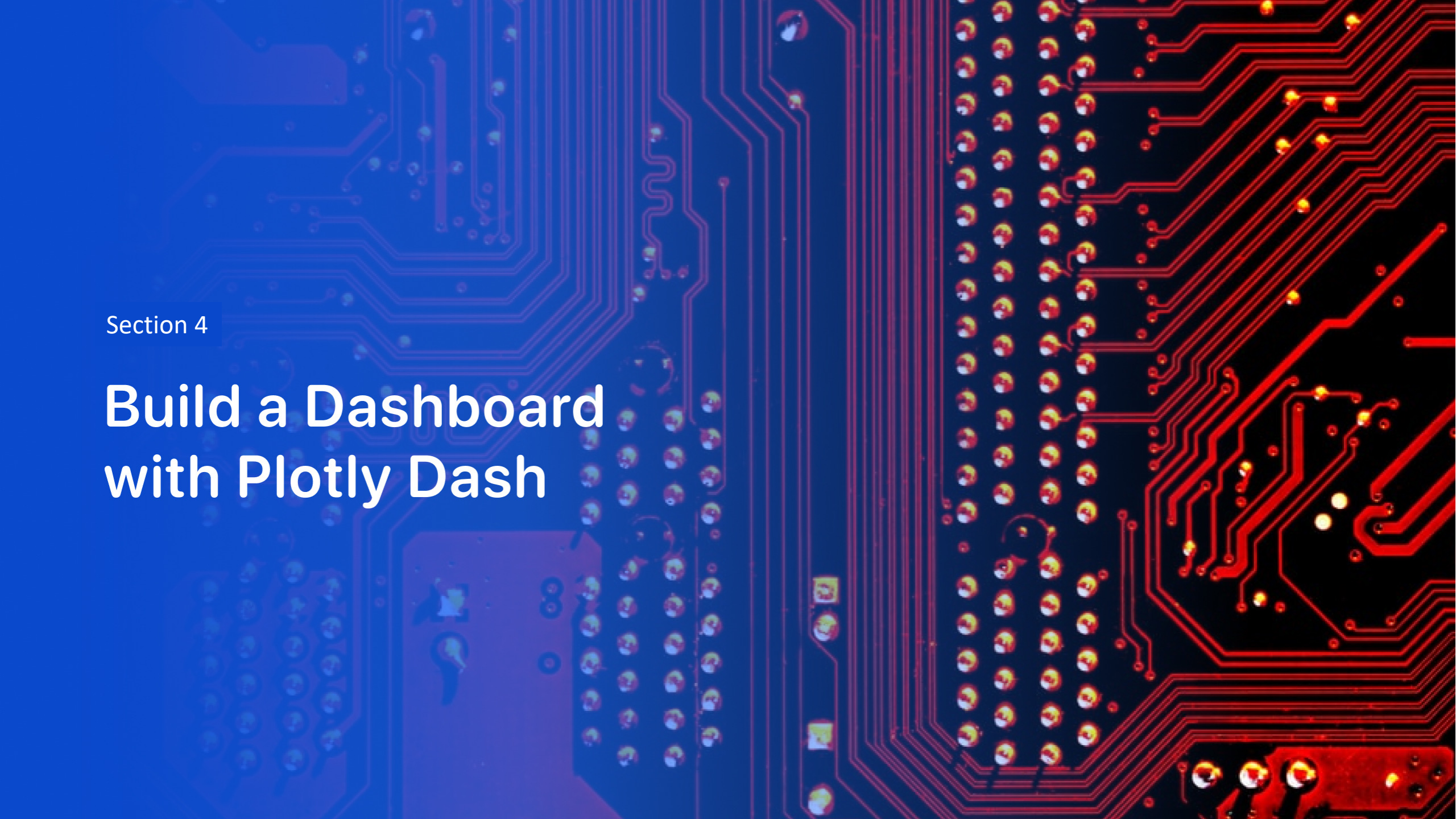




Distance to Proximities

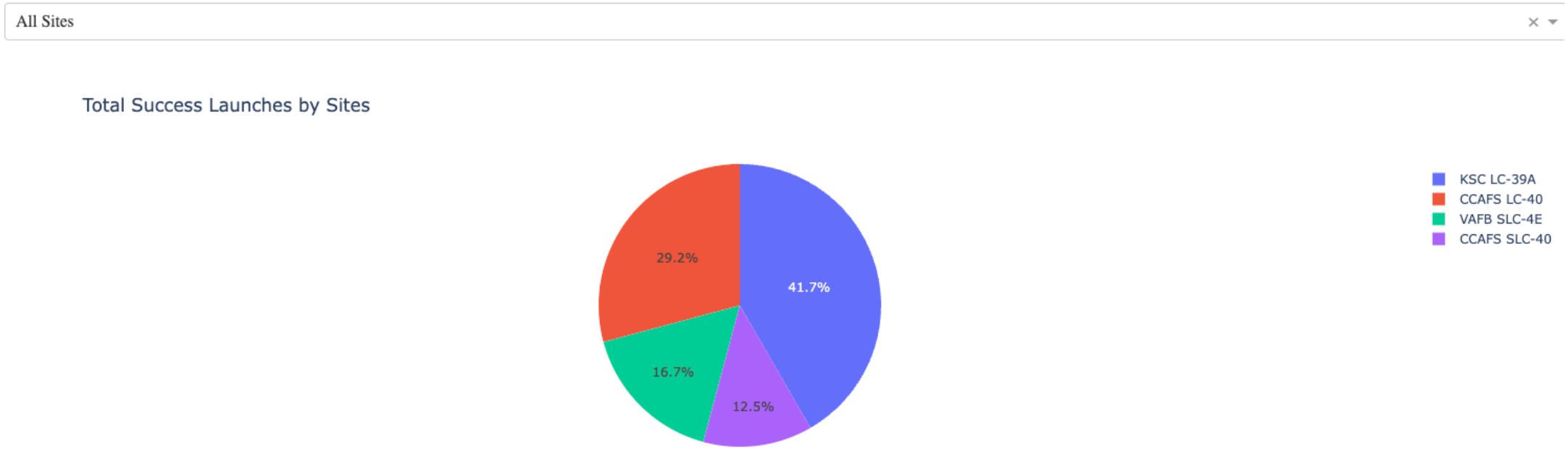
- Calculate the distance to proximities (such as city, railway, highway)
- Tag the distance and draw a line from the site, CCAFS SLC-40, to each proximity
- City: 18.19 km
- Railway: 0.96 km
- Highway: 0.58 km
- Coastline: 0.86 km




Section 4
Build a Dashboard with Plotly Dash

Successful Launches by Sites

- Display percentages of successful launches of each site
- KSC LC-39A has the highest percentage with 41.2%



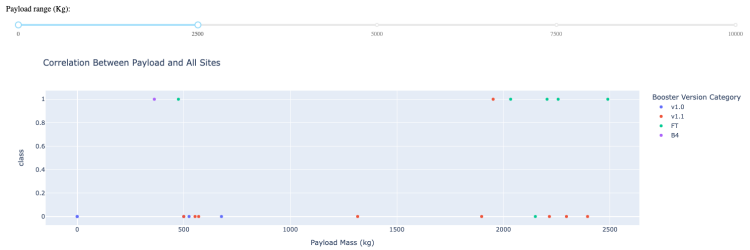
Site with Highest Successful Launch Rate

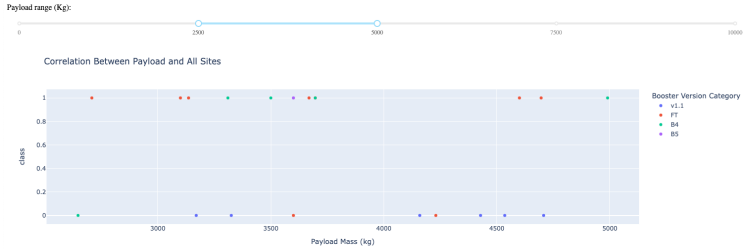
- KSC LC-39A has 76.9% success launch rate

Total Success Launches for site KSC LC-39A

Payload Mass vs. Successful Outcome

- Categorized by booster version
- Success as 1 and unsuccess with 0
- Payload Mass between 2000 to 5000 kg has the highest success rate



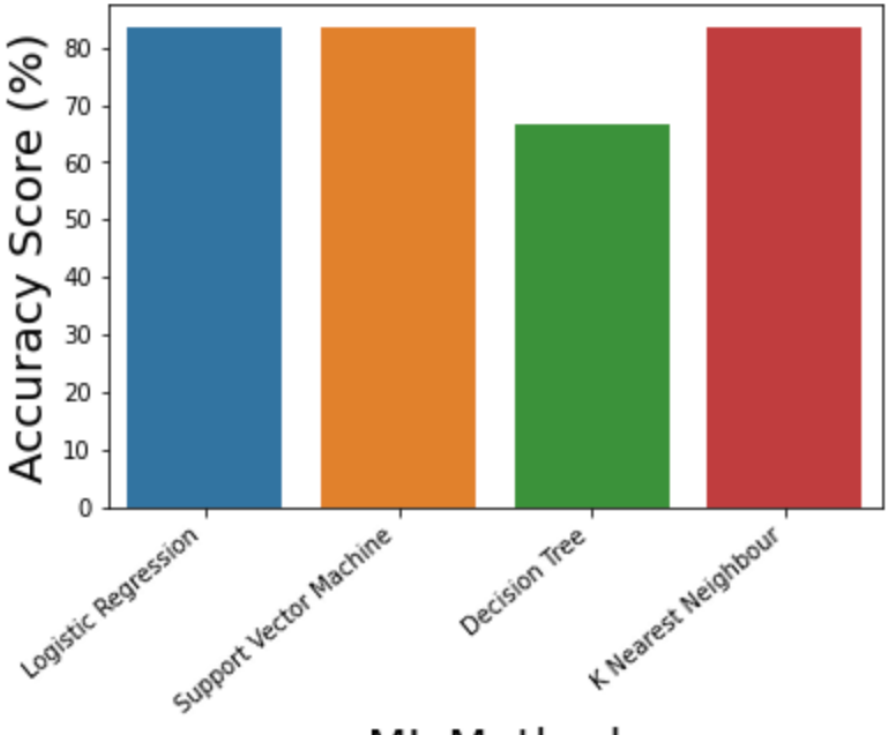


Section 5

Predictive Analysis (Classification)

Classification Accuracy

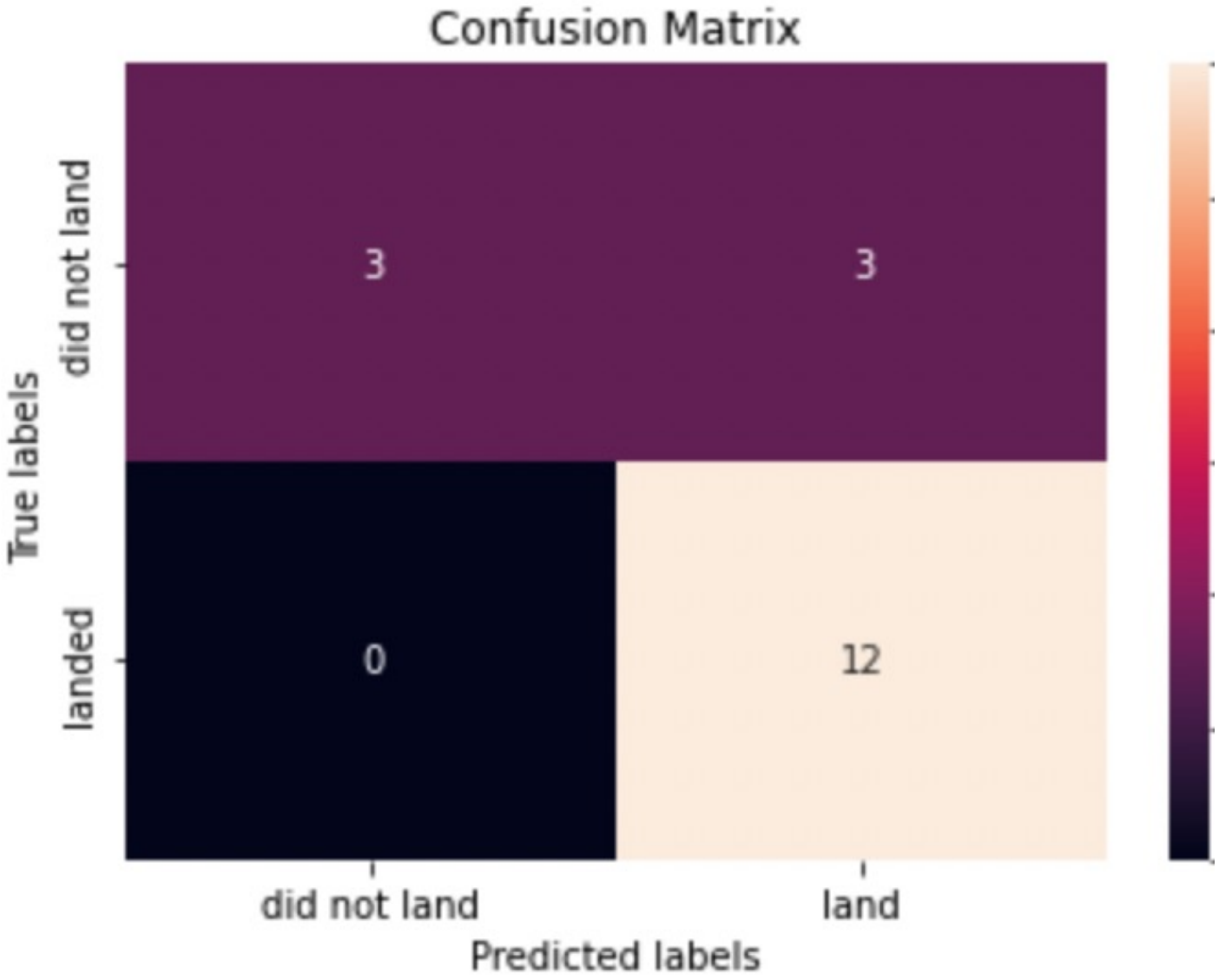
- Other than Decision Tree model, they all have accuracy with 83.33%



ML Method	Accuracy Score (%)
Logistic Regression	83.333333
Support Vector Machine	83.333333
Decision Tree	66.666667
K Nearest Neighbour	83.333333

Confusion Matrix

- Logistic regression, K nearest neighbors, and SVM had the same confusion matrix as in the figure.
- False positive seems to be erroneous
- Precision: TP/ (TP + FP)
  = .80
- Recall: TP/ (TP / FN)
  = 1
- F1 score: 2 * (Precision * Recall) / (Precision + Recall)
  = .89
- Accuracy: (TP + TN) / (TP + TN + FP + FN)
  = .833



Conclusions

- Key components for successful launches with Falcon9
  - Flight # - Higher is better
  - Payload Mass (kg) – 2000 to 5000 kg
  - Orbit type – SO type with 0%
  - Launch site - KSC LC-39A
  - Proximities – Coastline, Railway, Highway, and City
- Logistic regression, k nearest neighbors, and SVM ML models resulted in the identical accuracy on test set of 83.3%

Thank you!