Winning Space Race with Data Science

Parth Mehta

10-12-2021

Outline

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

Executive Summary

- Summary of methodologies

1. Data Collection – SpaceX API & Web Scraping
2. Data Wrangling
3. EDA with Data Visualization & SQL
4. Interactive Map with folium
5. Plotly Dash Board
6. Machine Learning Models- LogReg, SVM, KNN, Decision Tree

- Summary of all results

Complete Data Frame of Falcon 9,Success Rate, Launch Site's Location and surroundings, Charts and Plots, Accuracies Of the Models

Introduction

MANUFACTURING COST

SPACEX
$62 Millions

Other Providers
Upwards of $165 millions

HOW??

By Reusing The First Stage !

In this Project ,We Will find the key of a Success landing

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX API and Web Scraping
- Perform data wrangling
  - Filtering, Transforming, Changing and adding columns like df['Class']
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Split, Fit, Train, Test, Best Parameters, Best Score and Score method used

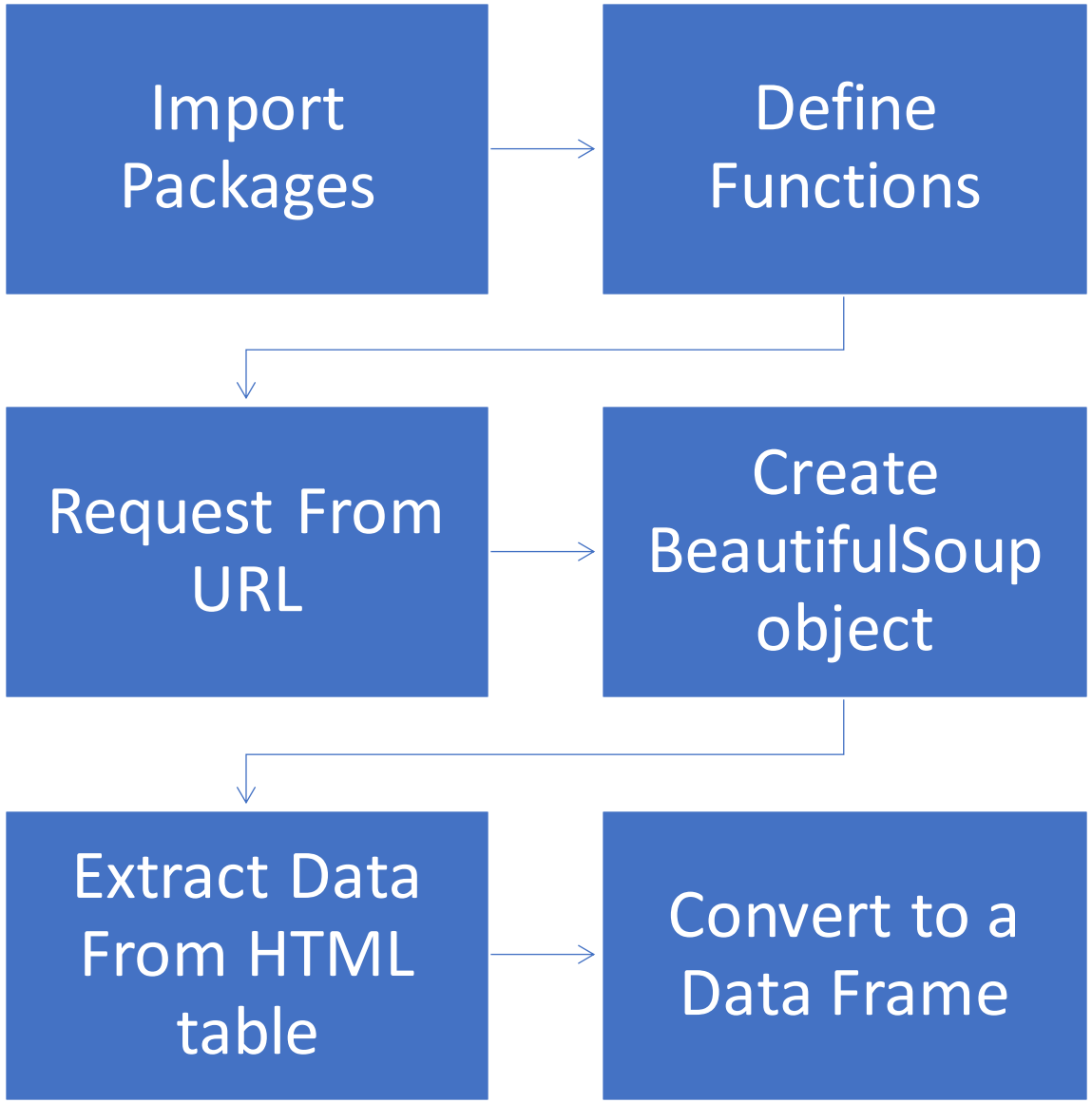
Data Collection – SpaceX API

- [ ] Import Libraries
- [ ] Define Auxillary Functions
- [ ] Get Request
- [ ] Normalize
- [ ] Cleaning Requested Data
- [ ] Filtering Falcon 9 Data
- [ ] Reser Flight Number

API GITHUB

Data Collection - Scraping

WEB SCRAPING GITHUB



Data Wrangling

- Missing Values
- Column Type
- No. of Launches per orbit , No. Of launches per site
- Landing Outcomes – TRUE =1 , False=0
- Add Class Column
- Success rate

EDA github

EDA with Data Visualization

- Flight Number Vs Payload Mass -

↑ Flight Number = ↑ Payload Mass & ↑ Success Rate

- Flight Number Vs launch Site – TOP = CCAFS SLC 40
- Payload vs launch site – Most Heavy Launches in CCAFS SLC 40
- Success rate vs Orbit – Higher in ES-L1, GEO , HEO , SSO

  Lower in GTO & ISS

- Flight Number VS orbit - ↑ Flight Number = ↑ VLEO Orbit
- Payload Mass vs Orbit – Highest Payload = VLEO Orbit
- Success Rate Vs Year - ↑ 2013-17 ↓ 2017-18 ↑ 2018-19 ↓ 2019-20
- EDA with data visualization github

EDA with SQL

- DISTINCT Statement
- WHERE + LIKE "%NAME%' - LIMIT
- SUM() + WHERE
- AVG() + WHERE
- MIN() + WHERE
- COUNT() + GROUPBY
- DISTINCT + WHERE (SUBQUERY MAX()]
- substr(Date, 4, 2) as month + substr(Date,7,4)='2015'
- WHERE + BETWEEN AND + LIKE "%NAME%' - ORDER BY Desc
- EDA with sql github

Build an Interactive Map with Folium

- CIRCLES – To Mark Launch Site
- CLUSTER – To show in a simple manner the success/failure of each launch at a launch site.
- LINES- Illustrating the distances between key places and the launch site.
- DISTANCE KM – Expressing a numerical fact

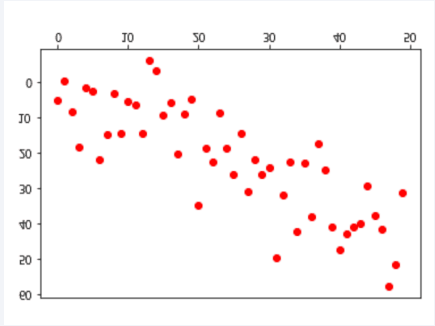
Interactive Visual Analytics With Folium Lab.ipynb

Build a Dashboard with Plotly Dash

DROPDOWN – For filtering Launch Sites

PIE CHART – Success rate vs Launch Site

SLIDER – Filtering Payload Mass



SCATTER CHART – Payload Mass vs Success Rate

[Dashboard Github]

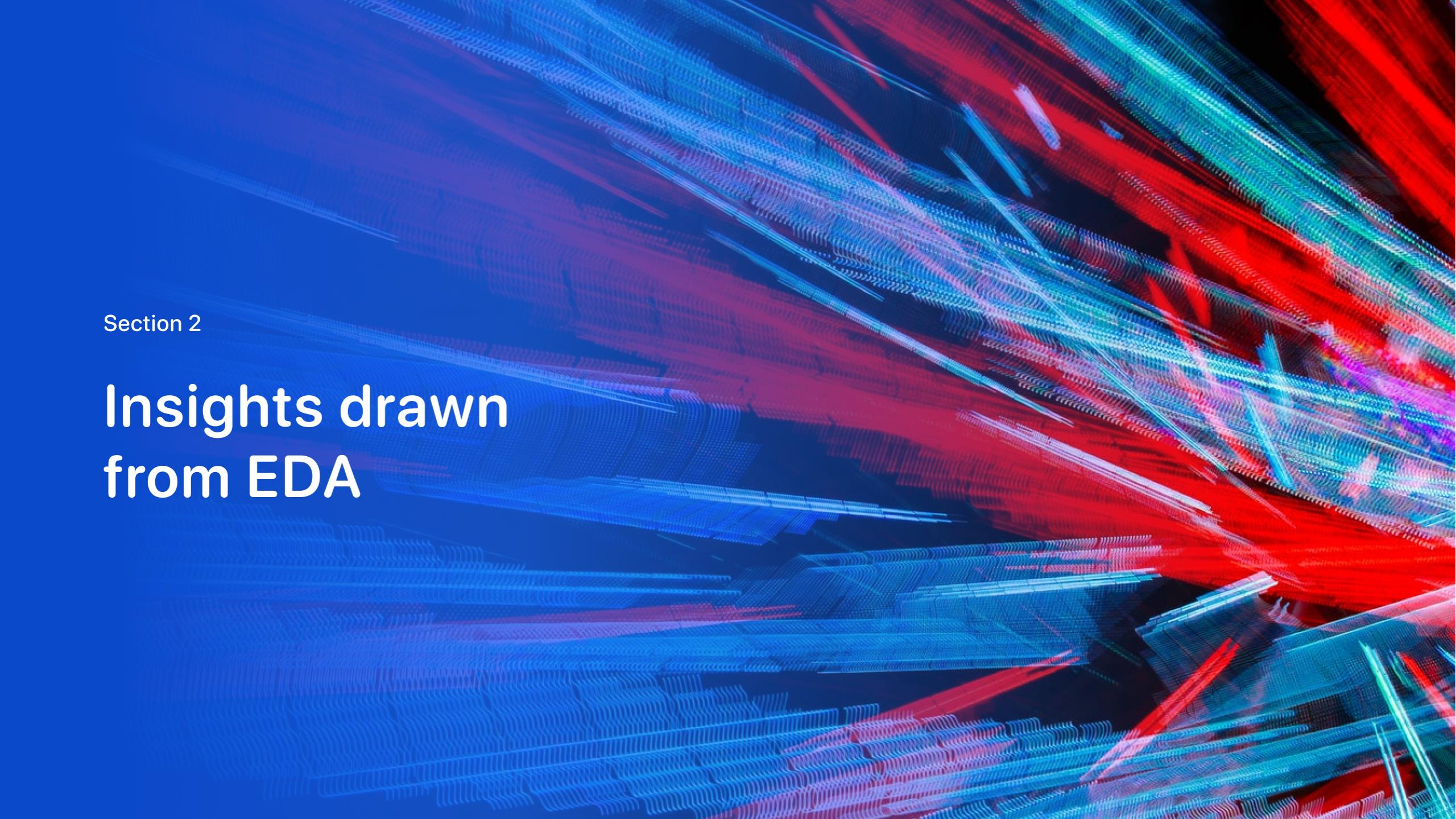
Predictive Analysis (Classification)

1. Define Y
2. Normalize and define X
3. Split the data : Train vs Test set
4. Run GridSearchCV objects with different classification techniques
5. Calculate their accuracies
6. Show each Confusion Matrix
7. Compare Accuracies

Machine Learning Prediction Github

Results

- Exploratory data analysis results
- Different Launch Sites: CCAFS LC-40, KSC LC-39A, VAFB SLC 4E, CCAFS SLC-40
- Most No. Launches: CCAFS SLC-40
- 100% Success Orbits: ES-L1,GEO,HEO, SSO
- Success Rate: 0.6667
- Total Payload Mass: 45,596KG
- First Successful Landing: 01-05-17
- Interactive analytics demo in screenshots
- Launch Site Latitude and Longitude
- Launch Site Plotting
- Marker color
- Cluster
- Closest Distances: Coast :0.90 km, City : 18.21 km , Highway : 0.58 km , Railway : 1.28 km
- Predictive analysis results
- 4 models: Logarithmic Regression, SVM, KNN Neighbors, Tree
- Best parameters
- Accuracy
- Score

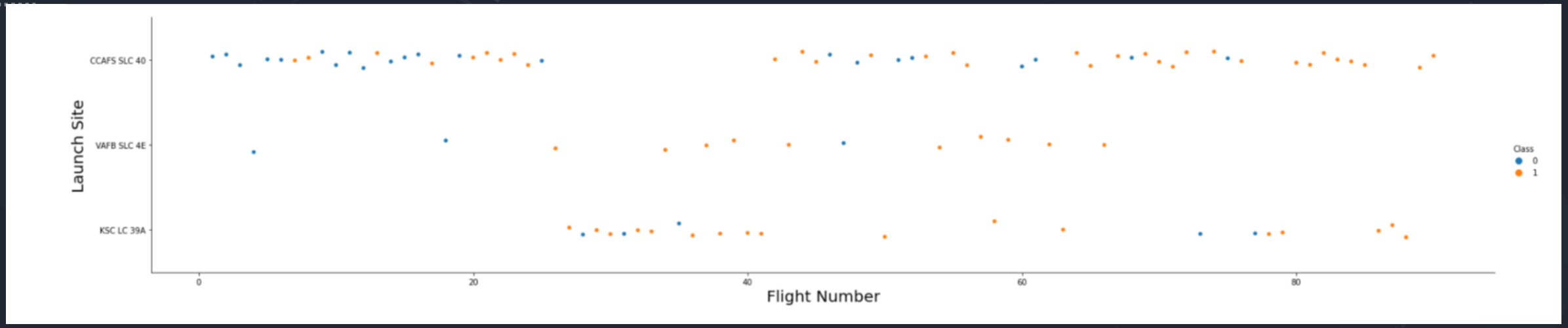


Section 2

Insights drawn from EDA

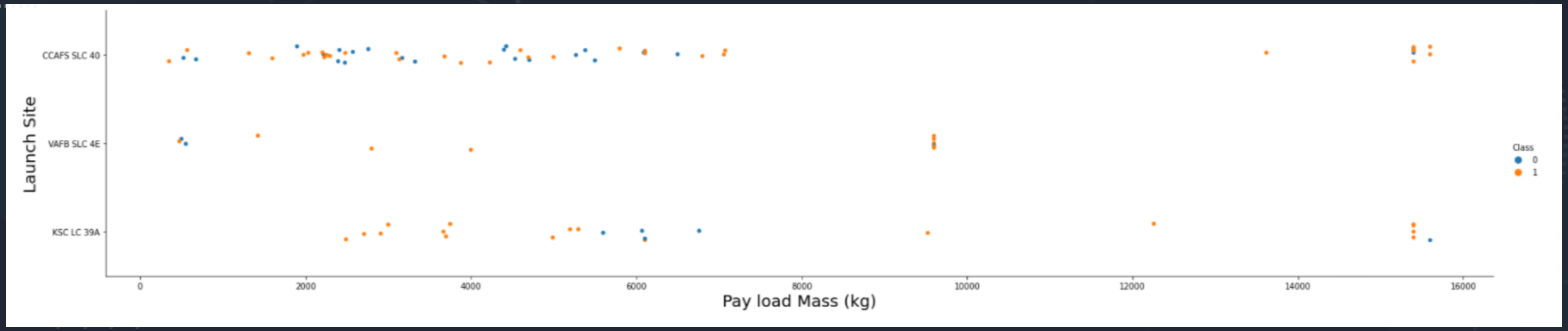
Flight Number vs. Launch Site

- Most launches at CCAFS SLC 40 (the more n° launches, the more successful)
- At VAFB SLC 4E, few n° of launches. No one for some launches.
- Late starting at KSC LC 39A, however, good successful rate and still working there.



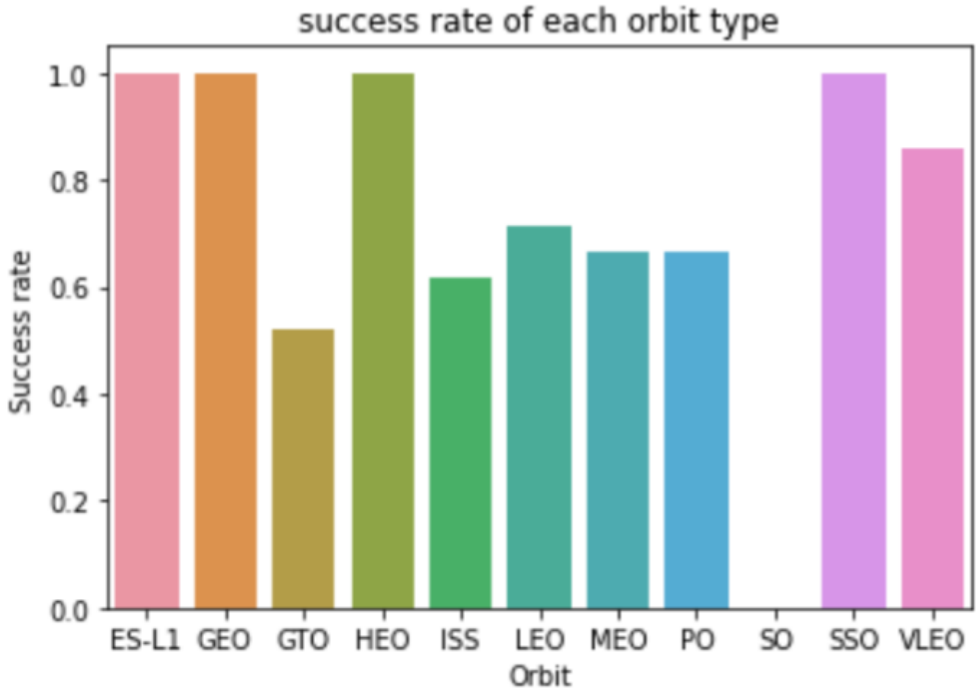
Payload vs. Launch Site

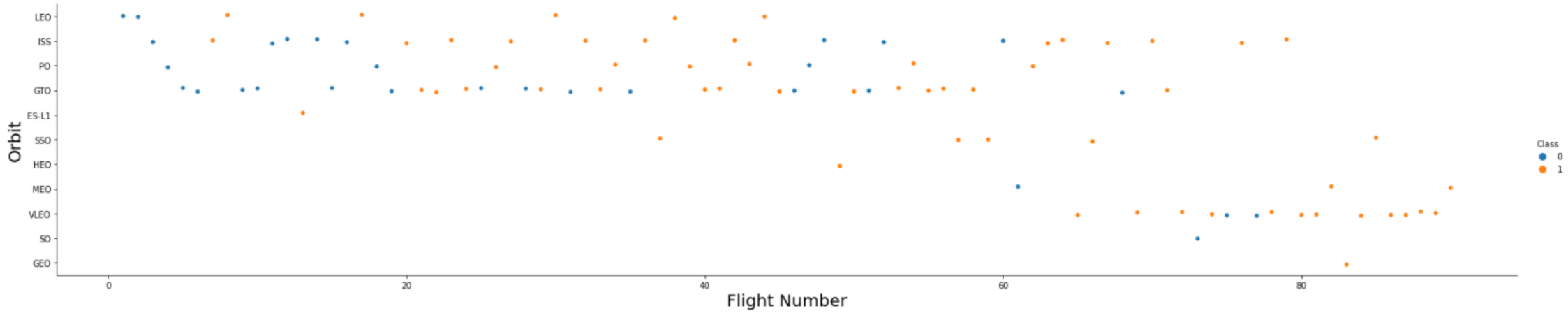
- For the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).



Success Rate vs. Orbit Type

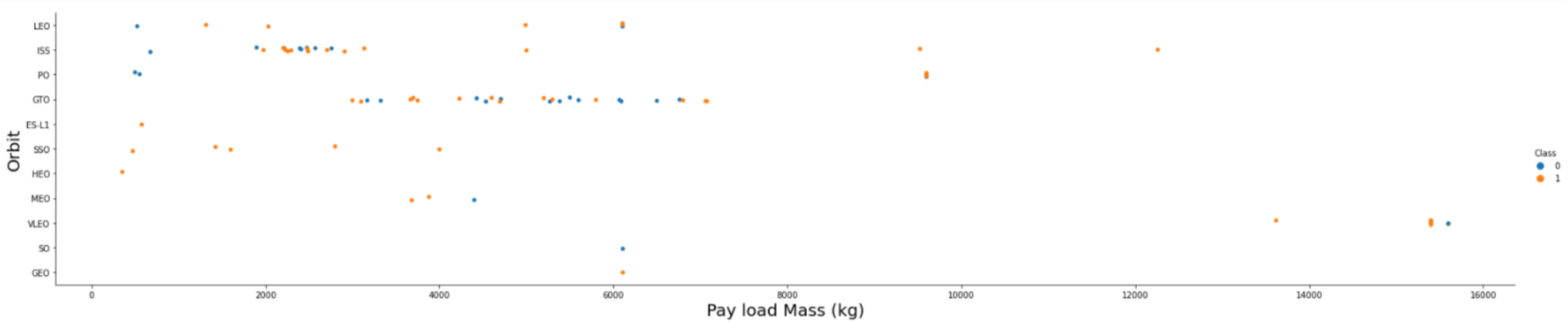
- 100 % success- ES-L1,GEO,HEO,SSO
- Failure- GTO,ISS,SO.





Flight Number vs. Orbit Type

- In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

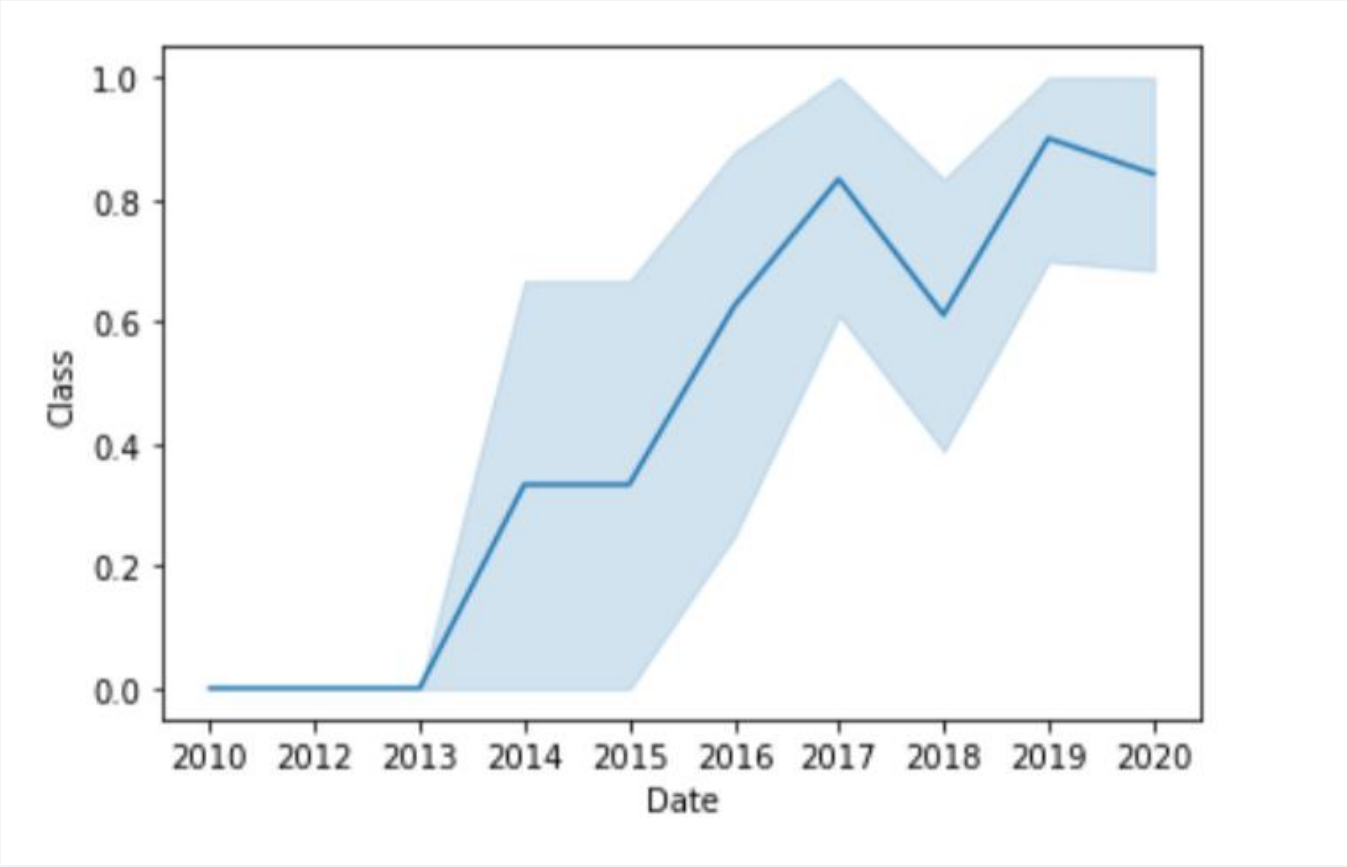


Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend

- you can observe that the sucess rate since 2013 kept increasing till 2020



All Launch Site Names

- **%sql** SELECT UNIQUE(Launch_Site) FROM SPACEXTBL

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

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	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

- **%%sql**
- **SELECT** **sum**(payload_mass__kg_),customer **from** SPACEXTBL **GROUP BY** customer **HAVING** customer**=**'NASA (CRS)'

1	customer
45596	NASA (CRS)

Average Payload Mass by F9 v1.1

- **%sql** SELECT avg(payload_mass__kg_),booster_version from SPACEXTBL GROUP BY booster_version HAVING booster_version ='F9 v1.1'

1	booster_version
2928	F9 v1.1

First Successful Ground Landing Date

- **%sql** SELECT min(date),landing__outcome from SPACEXTBL GROUP BY landing__outcome HAVING landing__outcome='Success (ground pad)'

1	landing__outcome
2015-12-22	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

- **%sql** SELECT booster_version FROM SPACEXTBL WHERE landing__outcome='Success (drone ship)' and (payload_mass__kg_ between 4000 and 6000)

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

Total Number of Successful and Failure Mission Outcomes

- `%sql SELECT COUNT(*),mission_outcome FROM SPACEXTBL GROUP BY mission_outcome`

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

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

2015 Launch Records

- **%sql** SELECT booster_version,launch_site,landing__outcome FROM SPACEXTBL WHERE landing__outcome = 'Failure (drone ship)' and YEAR(DATE)='2015'

booster_version	launch_site	landing__outcome
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

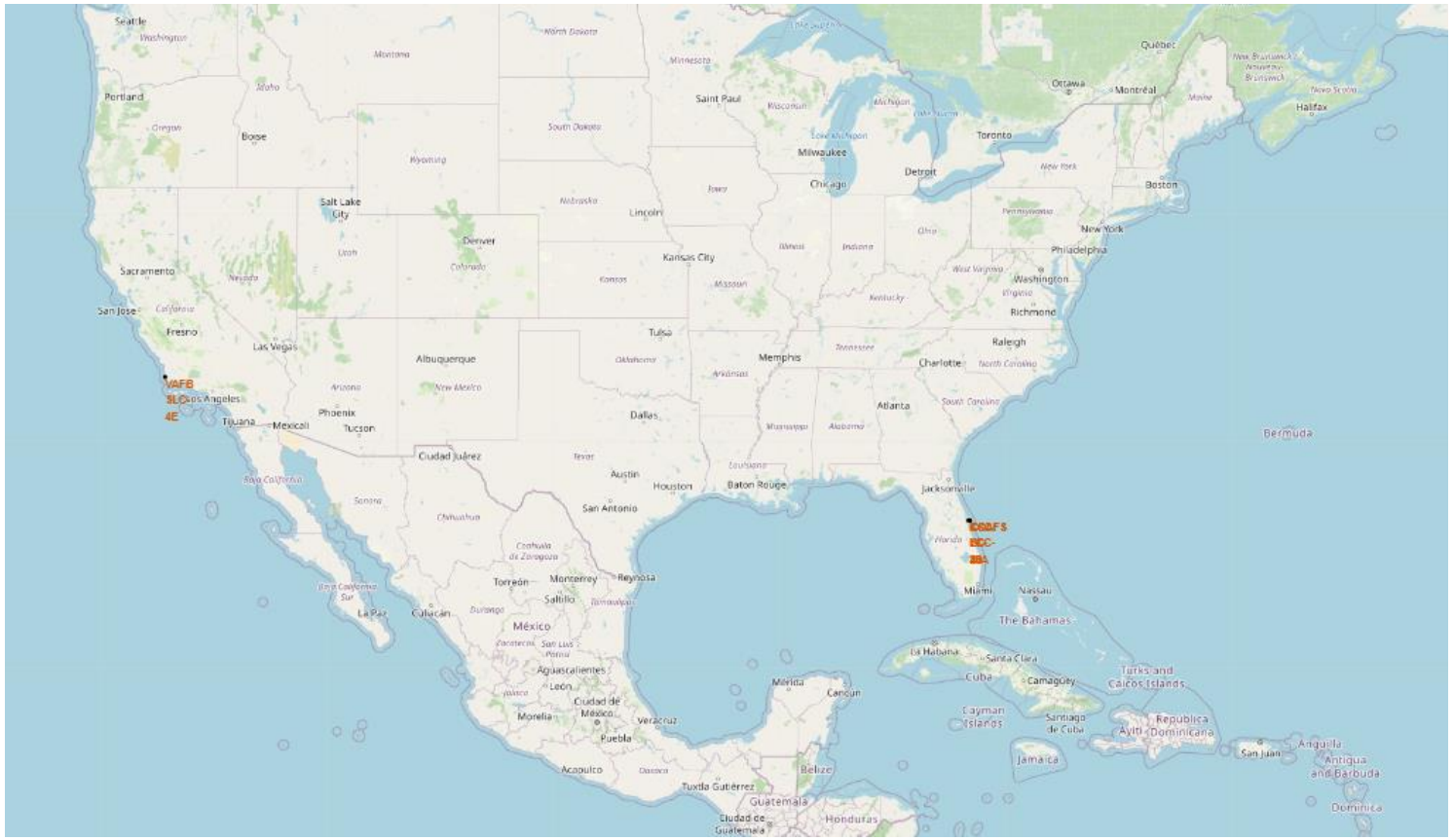
COUNT	landing__outcome
38	Success
22	No attempt
14	Success (drone ship)
9	Success (ground pad)
5	Controlled (ocean)
5	Failure (drone ship)
3	Failure
2	Failure (parachute)
2	Uncontrolled (ocean)
1	Precluded (drone ship)

- **%sql** SELECT COUNT(*) as count,landing__outcome from SPACEXTBL GROUP BY landing__outcome ORDER BY count DESC

Section 4

Launch Sites Proximities Analysis

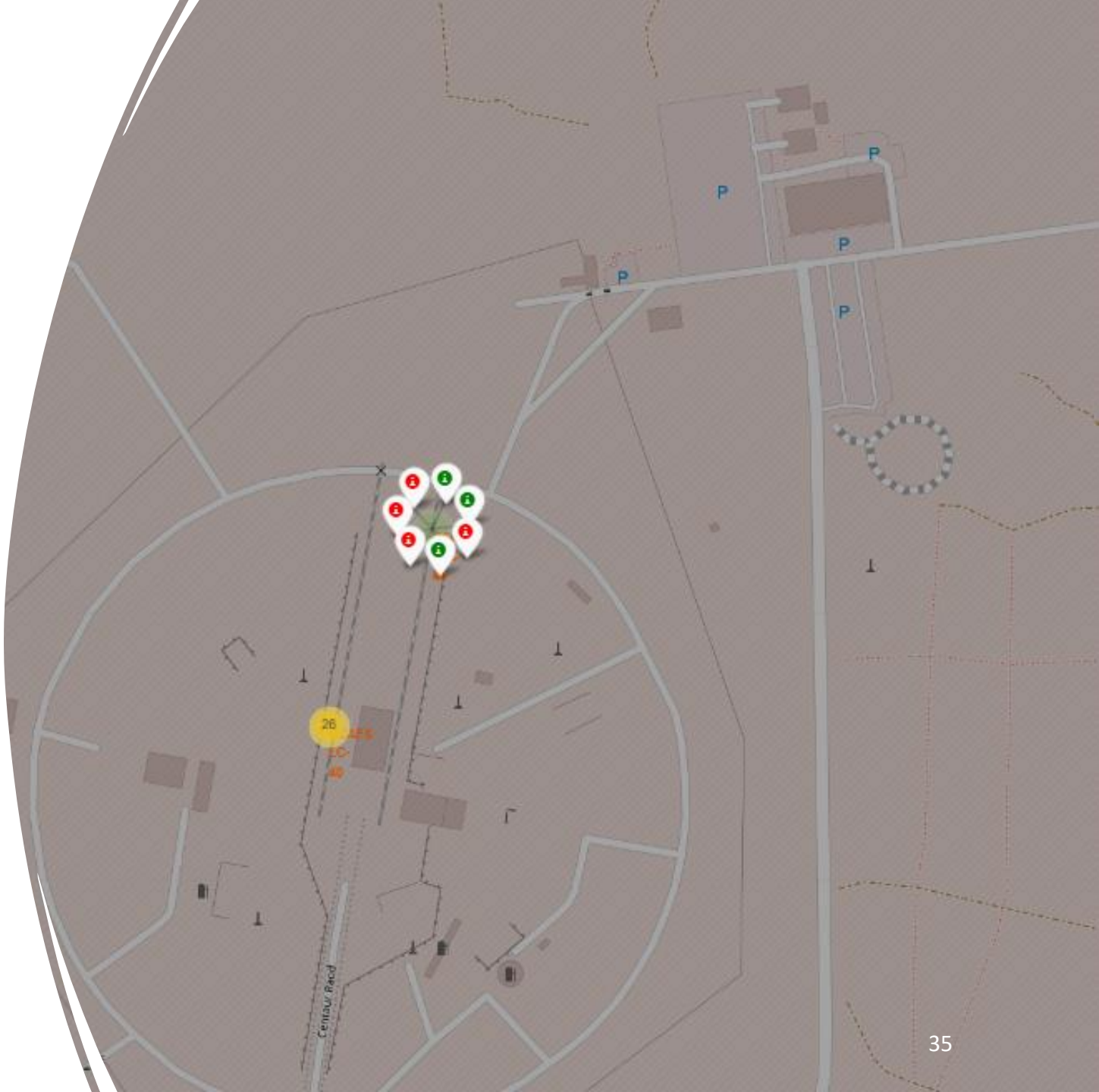
Launch Site Locations

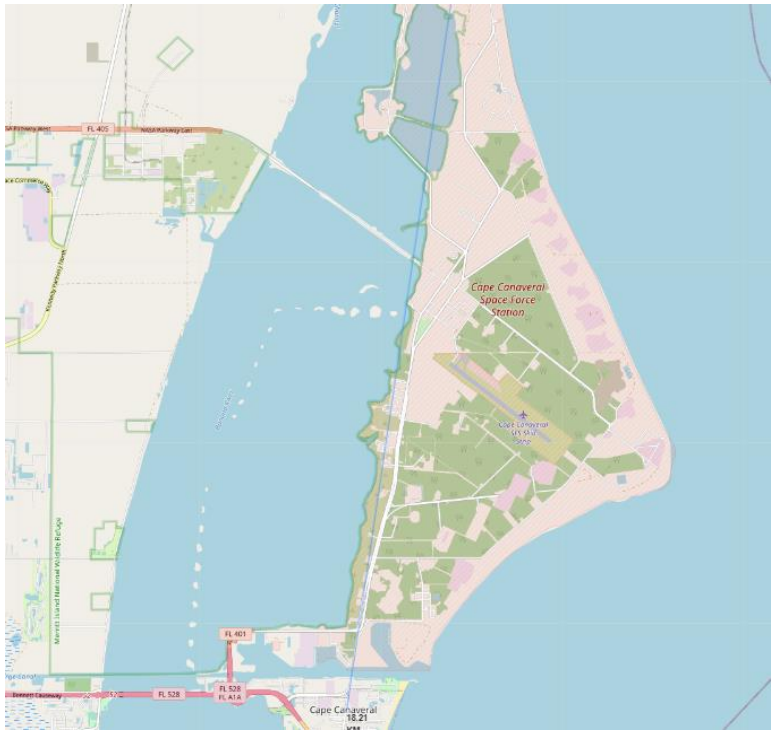


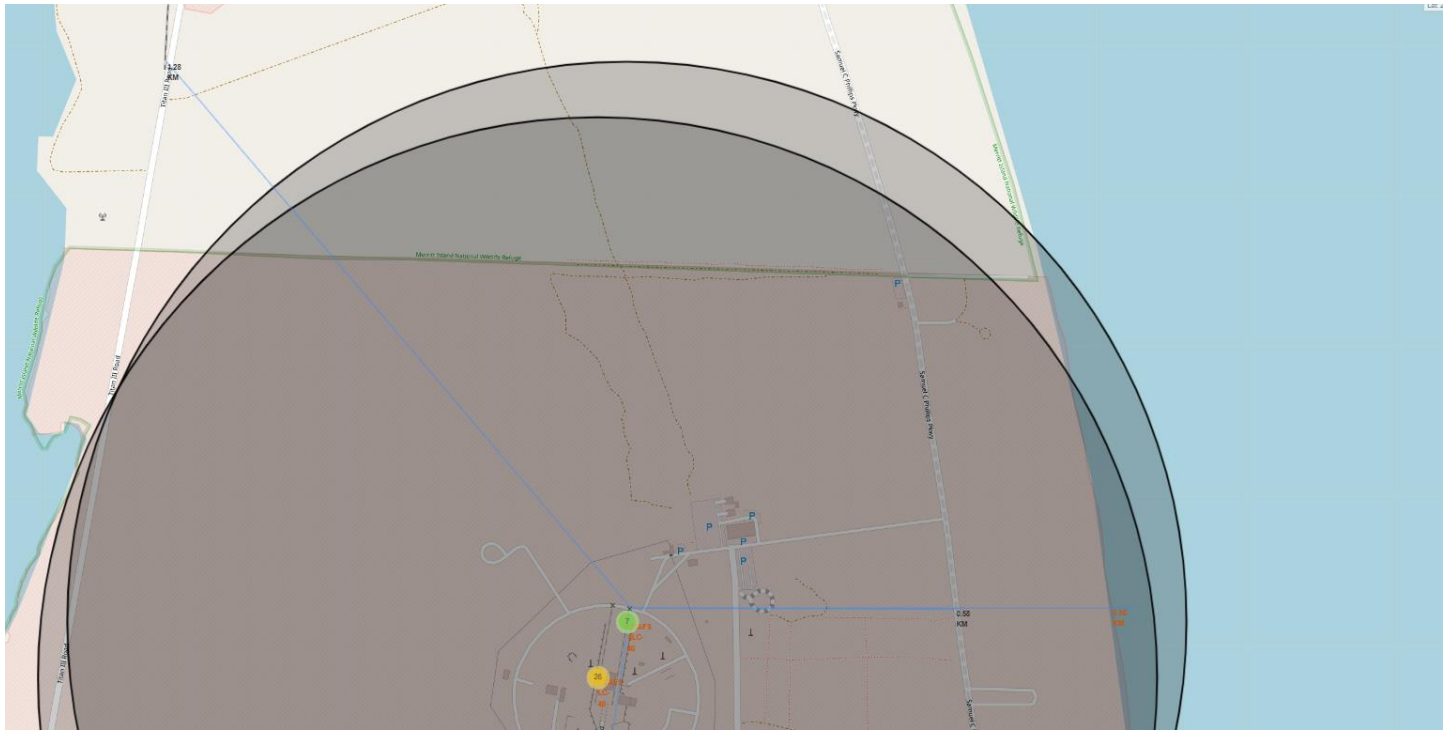
- All of them on the coast
- 3 out of 4 in the east coast
- Those on the east coast are pretty close

Launch Site Cluster

- Easy and accessible way to group and show the data (each launch site with its success/failure launch)

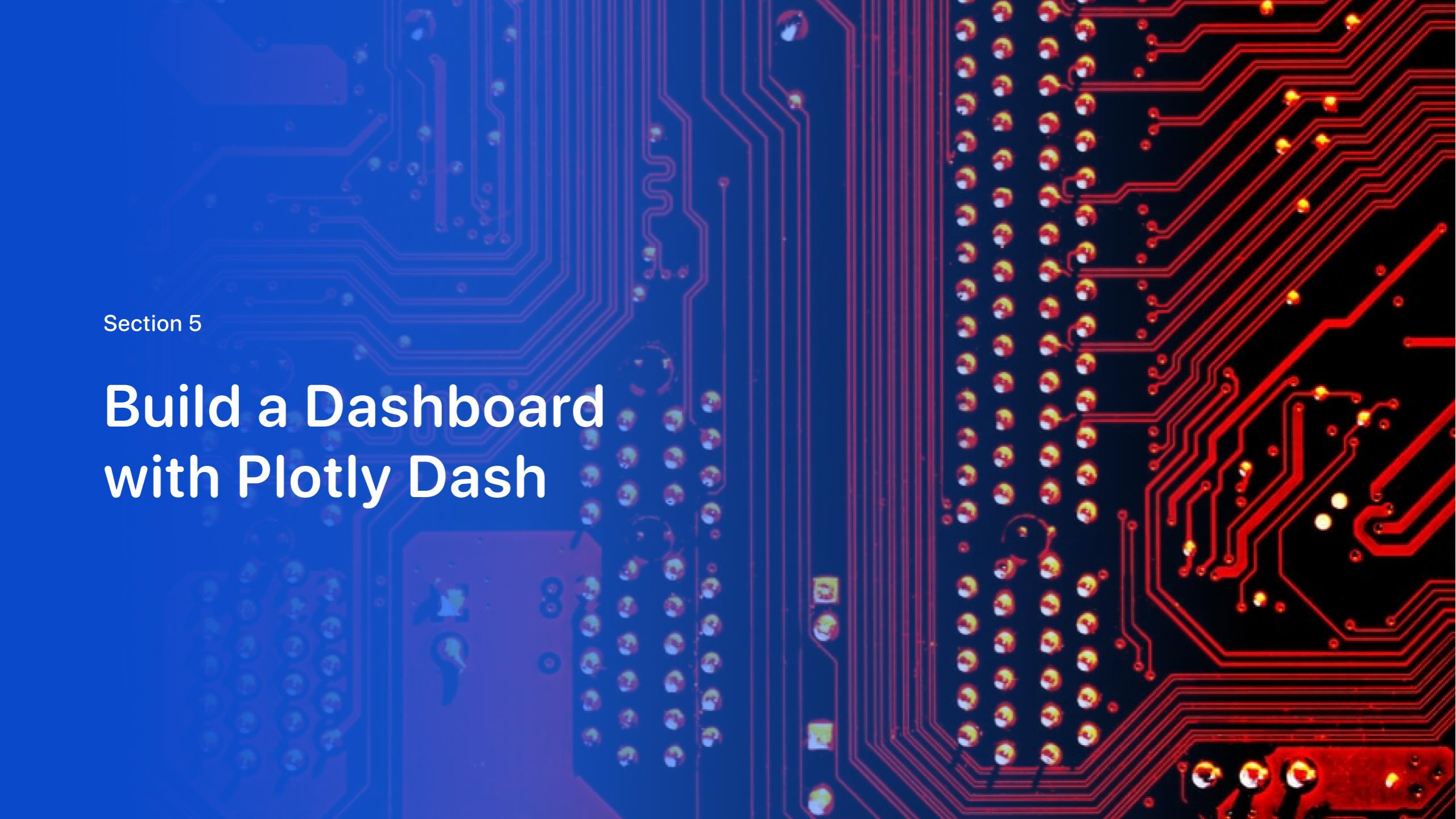






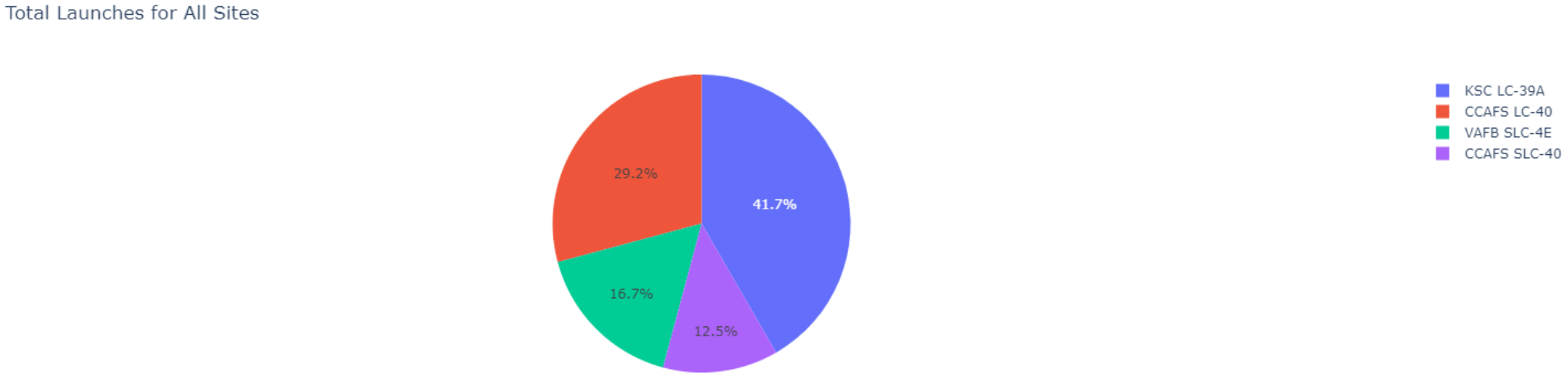
Exploring the surroundings of a launch site

- Closest Distances: Coast :0.90 km, City : 18.21 km , Highway : 0.58 km , Railway : 1.28 km



Section 5

Build a Dashboard with Plotly Dash

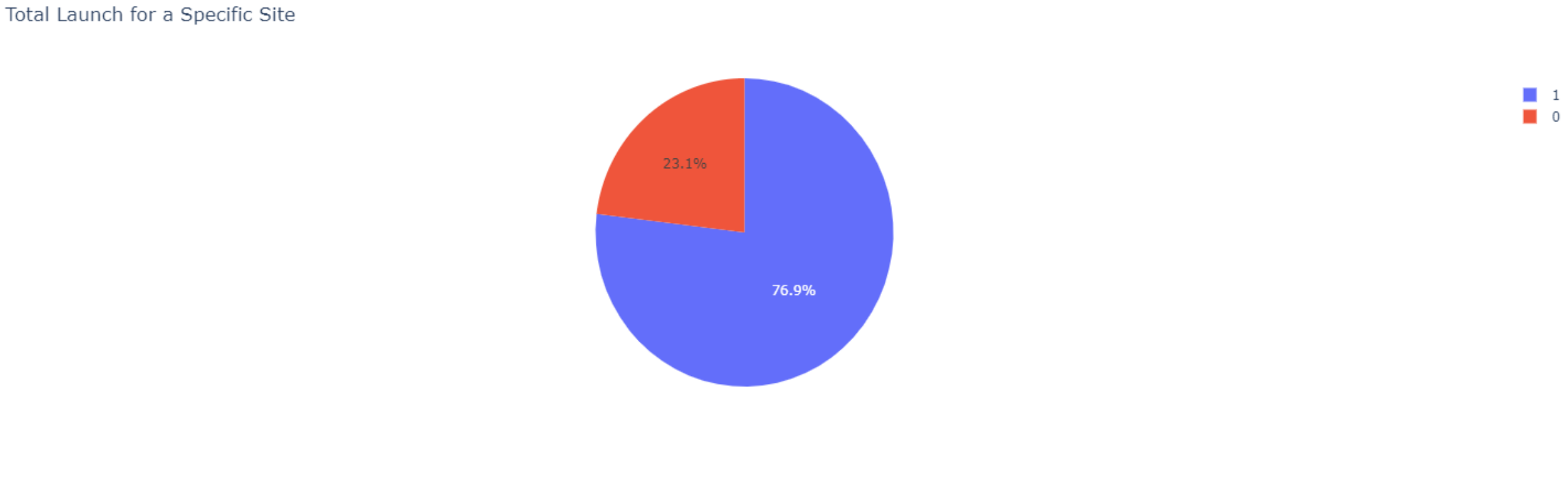


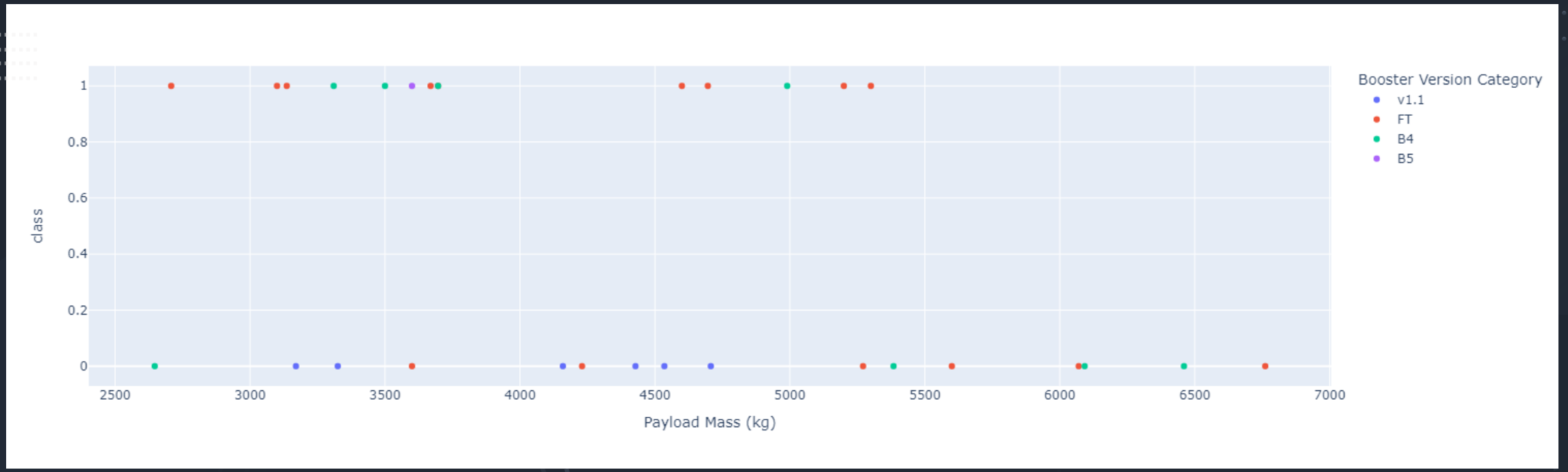
Launch Success Vs success rate

- KSC LC –39A Has the highest success rate

KSC LC-39A : THE BEST SITE

- KSC LC-39A: Although being the most successful one, nearly 25% of the launches fail!





Success Rate per Payload Mass for each Booster Version (2,500-7,500 kg)

- The FT Booster Version has the largest success rate

Section 6

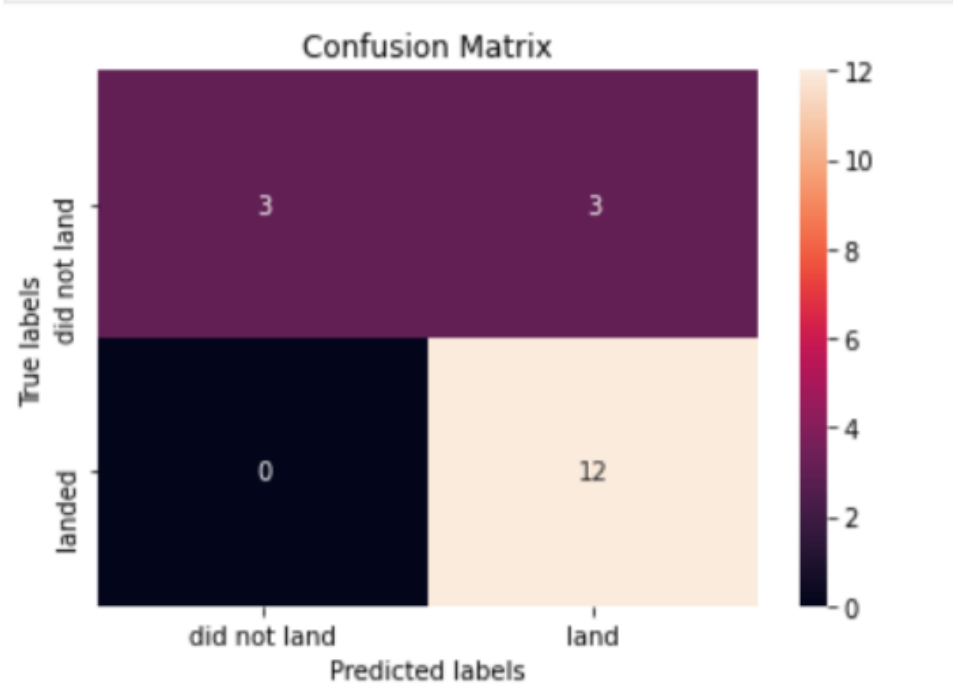
Predictive Analysis (Classification)

Classification Accuracy

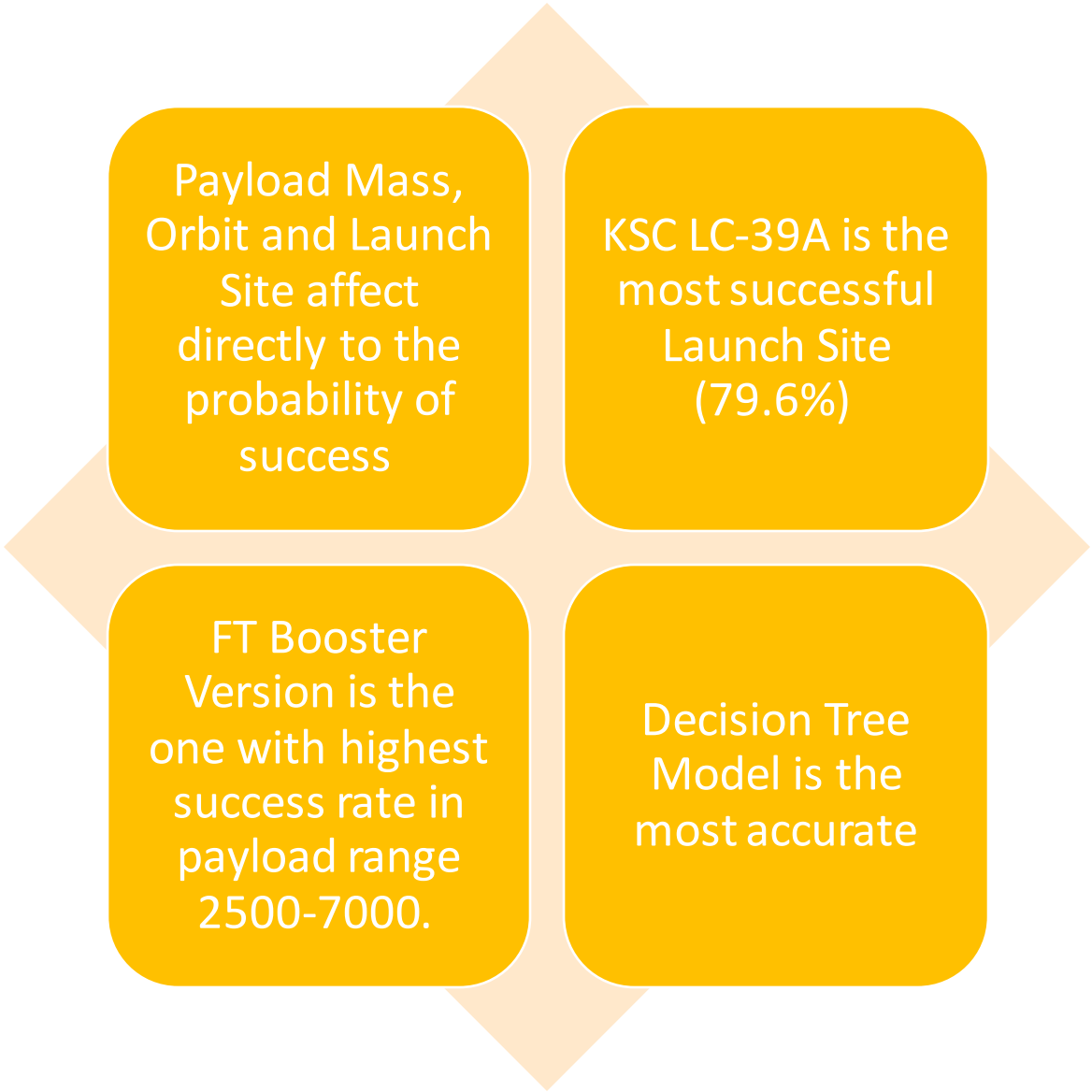
Decision Tree has the highest classification accuracy

Confusion Matrix

- The model predicted: 3 launches as not landed and 15 as landed
- Whereas true labels shows: 6 of them unsuccessfully landed and 12 did it with success.
- PRECISION: 0.5
- RECALL: 1
- ACURRANCY: 0.8334



Conclusion


Payload Mass, Orbit and Launch Site affect directly to the probability of success
KSC LC-39A is the most successful Launch Site (79.6%)
FT Booster Version is the one with highest success rate in payload range 2500-7000.
Decision Tree Model is the most accurate

Appendix

Model	Best Parameters	Score	Accuracy
Log Reg	{'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}	0.83	0.846
SVM	{'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}	0.83	0.848
DECISION TREE	{'criterion': 'gini', 'max_depth': 4, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 2, 'splitter': 'random'}	0.83	0.889
KNN	{'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}	0.83	0.848

Thank you!