IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Vitali Babich
2022-06-03

Outline

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

Executive Summary

- Used methodologies
  - Data collected by API and web-scrapping
  - Data visualization by line, bar, scatter charts
  - Spatial data visualization
  - Used Logistic Regression, SVM, Decision Tree, KNN methods
- Summary of all results
  - SpaceX are increasing success rate of the launching
  - Part of the orbits are problem for SpaceX

Introduction

- This report is a part of the Applied Data Science Capstone
- Problems I want to find answers
  - Is the launching success rate increasing or decreasing?
  - Which criteria influence to success rate and which don't?
  - Can we predict success of the landing

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from coursera tasks, web scaping, using REST API
- Perform data wrangling
  - Data was processed using pandas functions
- 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 data to train and test sets, standardize data
  - Build models (LogisticRegression, Decision Tree, SVM, KNN)
  - Choose the best model

Replace np.nan values in the data with the mean

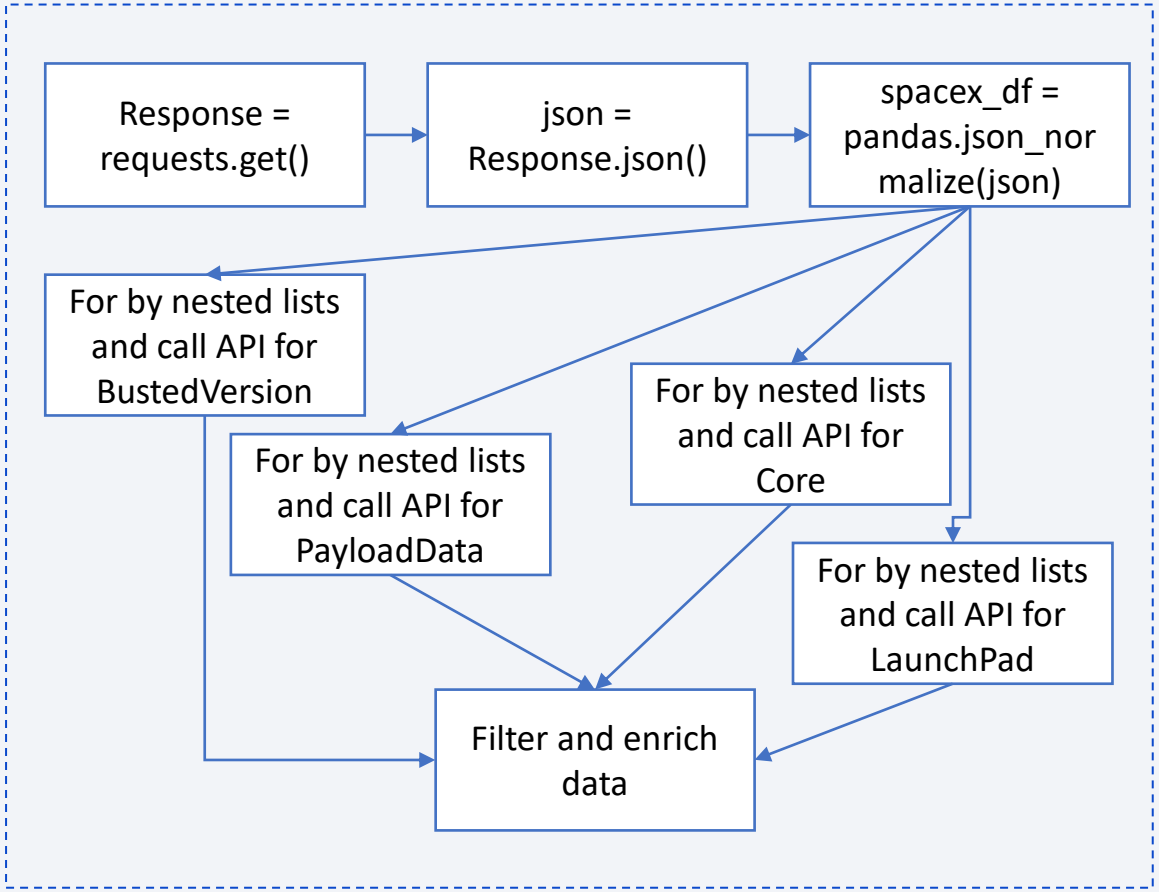
Data Collection

- REST API
  - Request to the SpaceX API
  - Clean the requested data
  - Replace missing values of PayloadMass with mean value
- Web scrapping
  - Get data from html page from Wikipedia (https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922)
  - Get tables and convert them to pandas DataFrame

Data Collection – SpaceX API

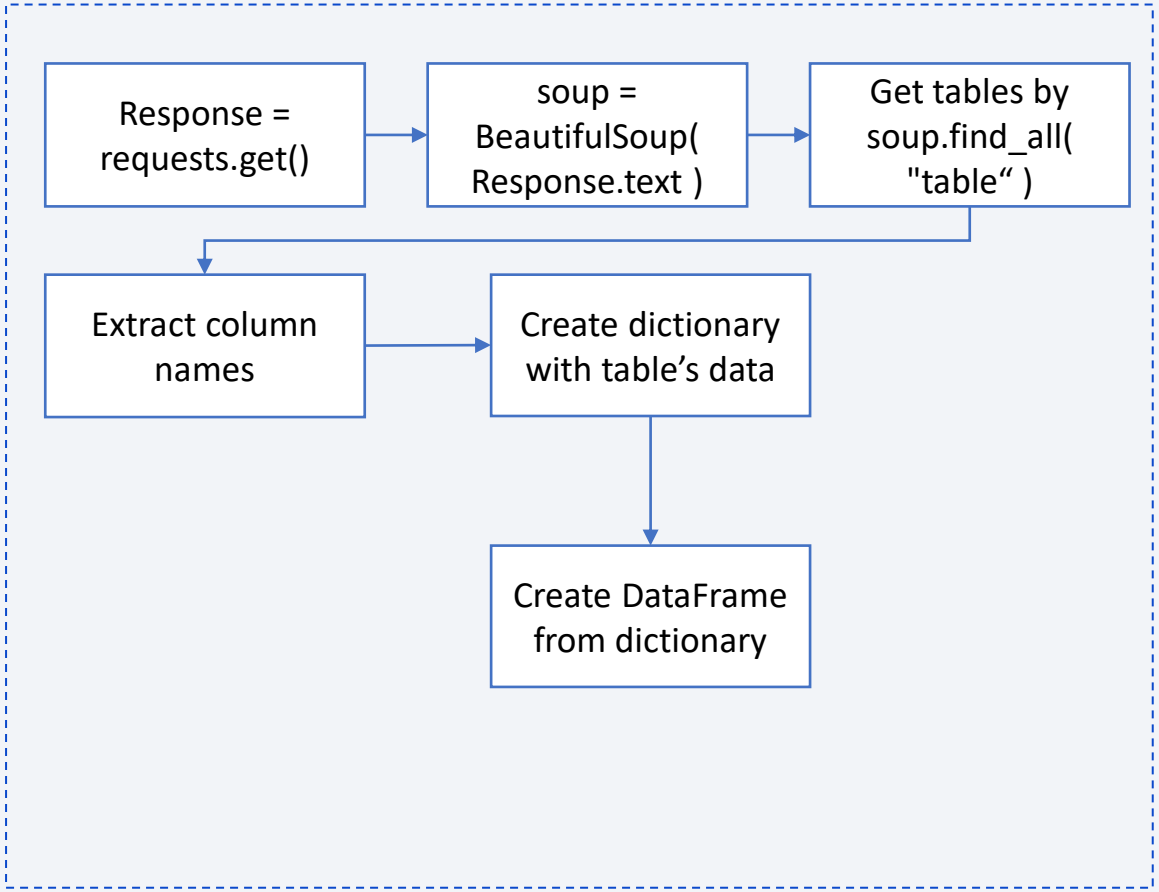
- Data collection by using functions get of the requests module.

- https://github.com/vitalyvb1974/ds/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



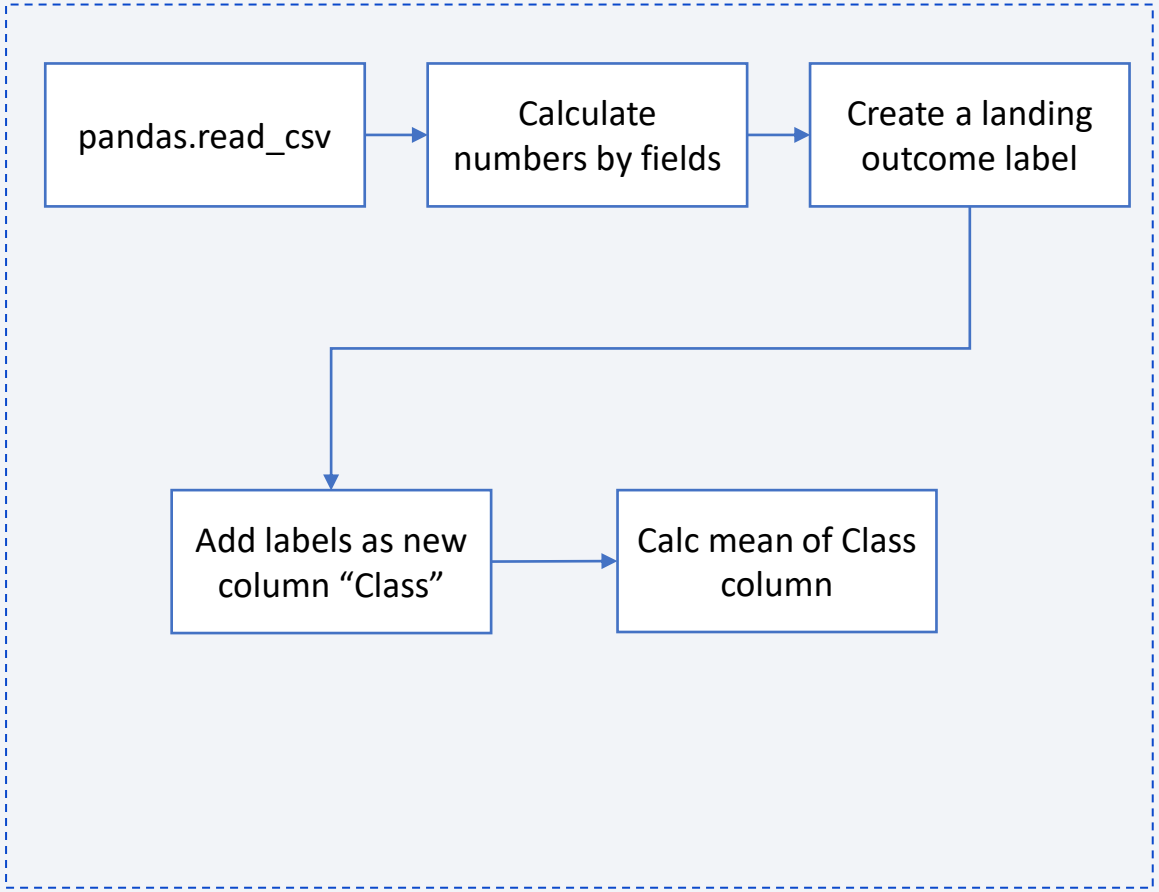
Data Collection - Scraping

- Scraping process includes receiving html page by function get of the requests module, parse it by methods of soup object (BeautifulSoup module), create dataframe from dict
- https://github.com/vitalyvb1974/ds/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Determine the number and occurrence for fields LaunchSite, Orbit, Outcome by value_counts()
- Create landing outcome labels (by comprehensive list) and add them to new column into dataframe
- https://github.com/vitalyvb1974/ds/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- Scatter plot to visualize the relationship between parameters (Payload vs. Orbit type, FlightNumber vs. Orbit type, Payload vs. Launch Site, Flight Number vs. Launch Site, FlightNumber vs. PayloadMass)
- Bar chart to visualize the mean value of the parameters (success rate of each orbit type)
- Line chart to visualize the mean the trend of the parameter (launch success yearly trend)
- https://github.com/vitalyvb1974/ds/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- I used SQLite database
- SQL queries are used to get distinct launch sites, 5 launch sites by pattern name, aggregate sum of the payload mass (in kg), average payload mass for distinct booster version, date of the first successful landing outcome, booster version with range of payload mass, total count of success and failed outcomes, booster version with max payload mass by sql subquery, list of failed landing_outcomes in drone ship and Rank the count of landing outcomes
- https://github.com/vitalyvb1974/ds/blob/main/jupyter-labs-eda-sql-coursera.ipynb

Build an Interactive Map with Folium

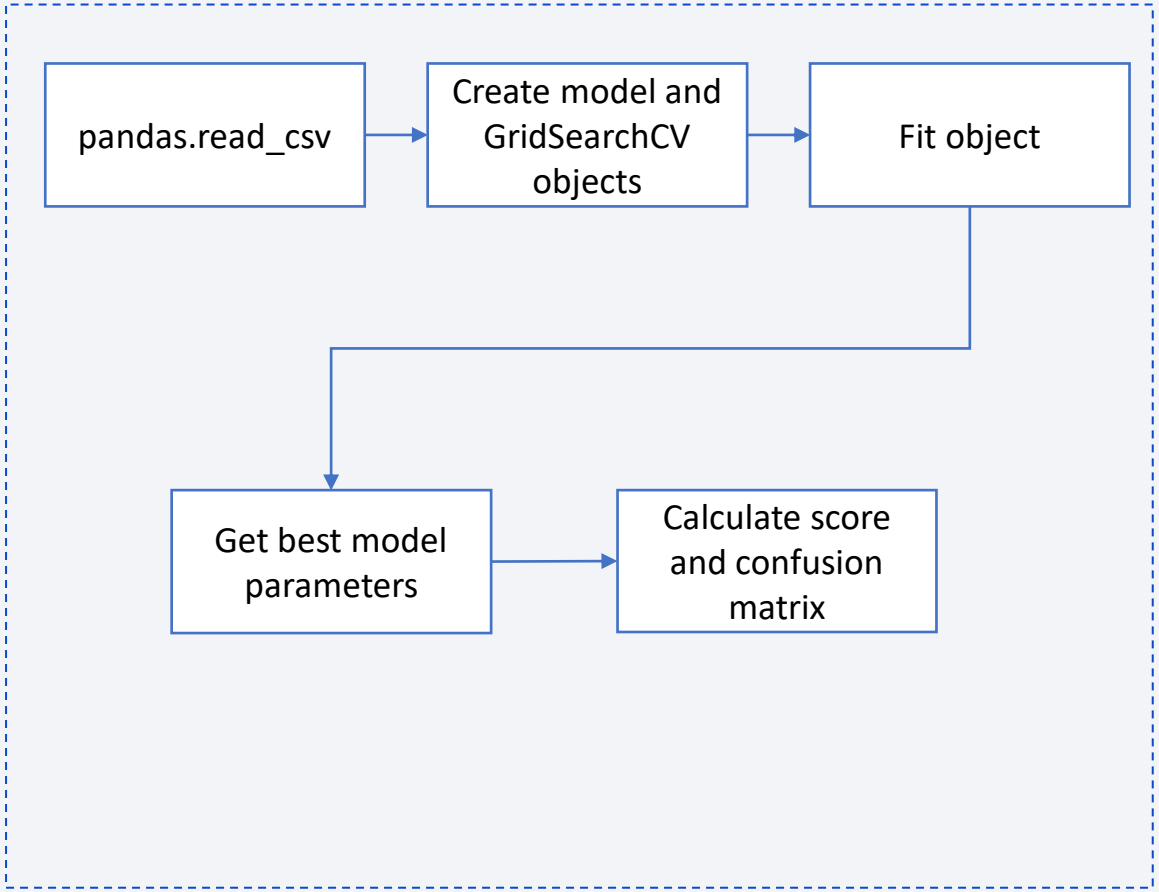
- I used folium.map object, and folium.Circle, folium.map.Marker, folium.plugins.MarkerCluster, folium.plugins.MousePosition, folium.features.DivIcon, folium.PolyLine, folium.Icon
- These objects are used to add a highlighted circle area to map (text and with a icon showing its name) (Circle and Marker), to simplify a map containing many markers having the same coordinate (MarketCluster), to specify icon (DivIcon and Icon), to add line (PolyLine), to get coordinate for a mouse over a point on the map (MousePosition)
- https://github.com/vitalyvb1974/ds/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- I have added:
  - Dropdown (to enable Launch Site selection) and rangeslider (to select payload range) interactions
  - Pie chart (to show the total successful launches count for all sites) and scatter plot (to visualize relationship between parameters)
- I select these visualizations because there accordance to goal like site selection, range selection, showing relation, showing total values
- https://github.com/vitalyvb1974/ds/blob/main/dashboard01.py

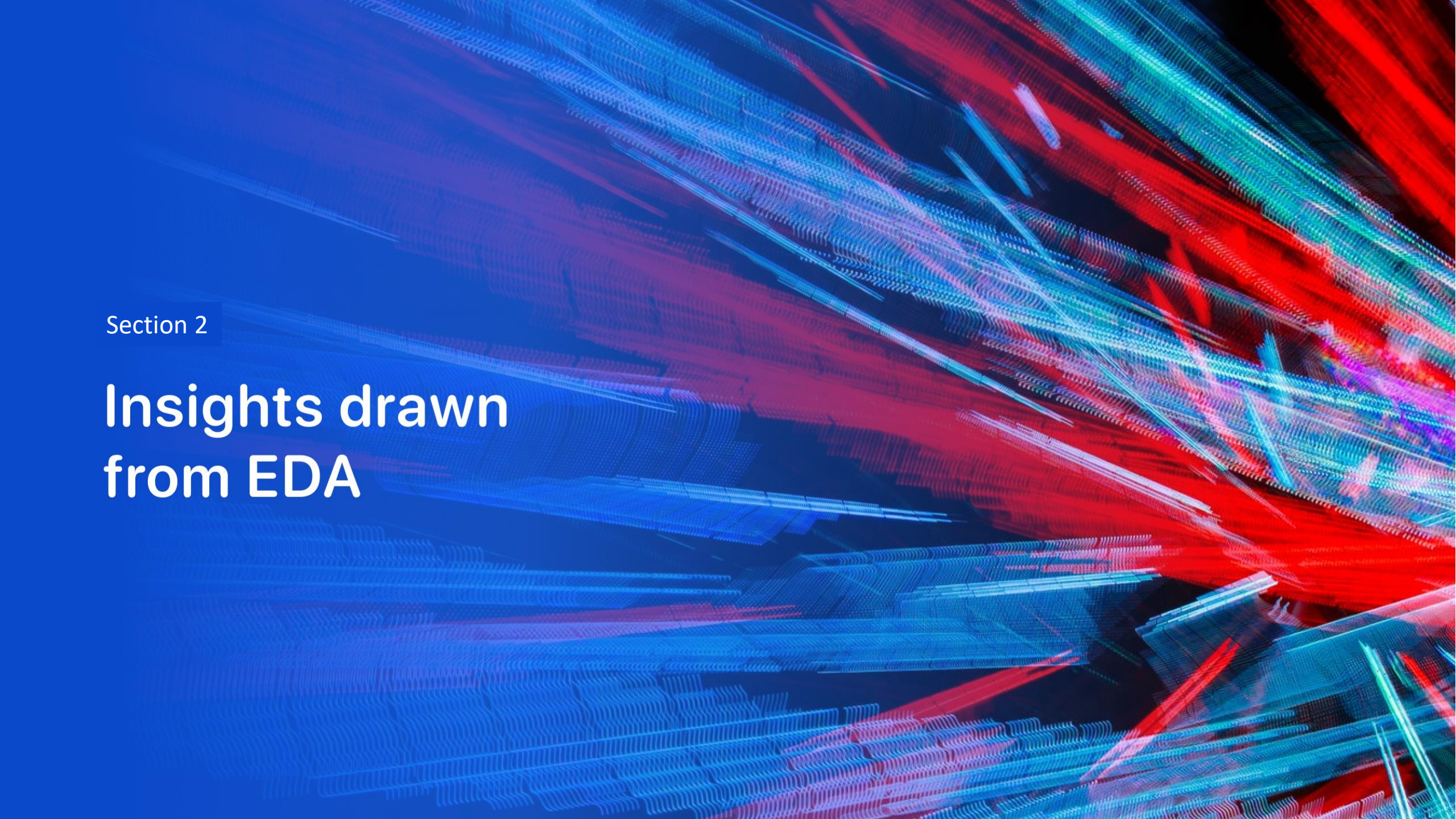
Predictive Analysis (Classification)

- I create Logistic Regression (LR), SVM, DecisionTree, KNN objects. Use GridSearchCV object with different parameters to find the best model (the best combination of parameters). Then models have been evaluated by test set.
- Process consists of loading data set, slitting it to train and test parameters, create objects and fit them with train data. Get score by test data.
- https://github.com/vitalyvb1974/ds/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



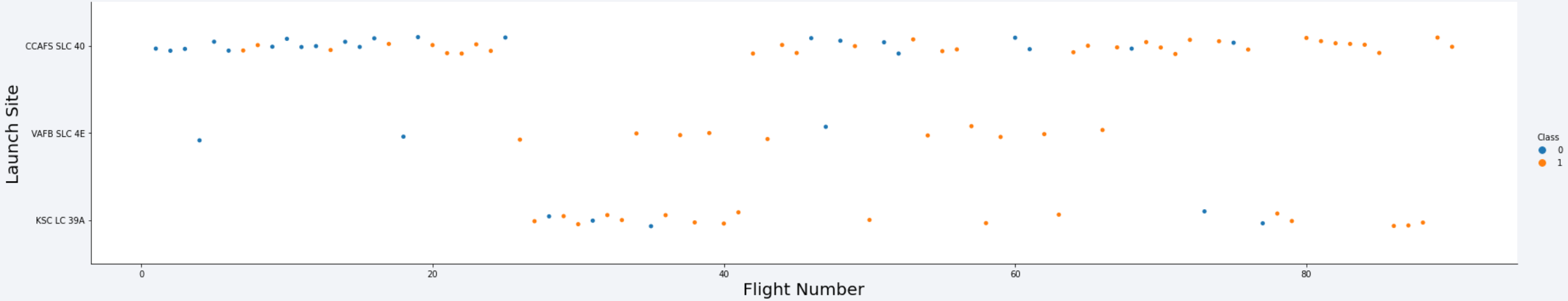
Results

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


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

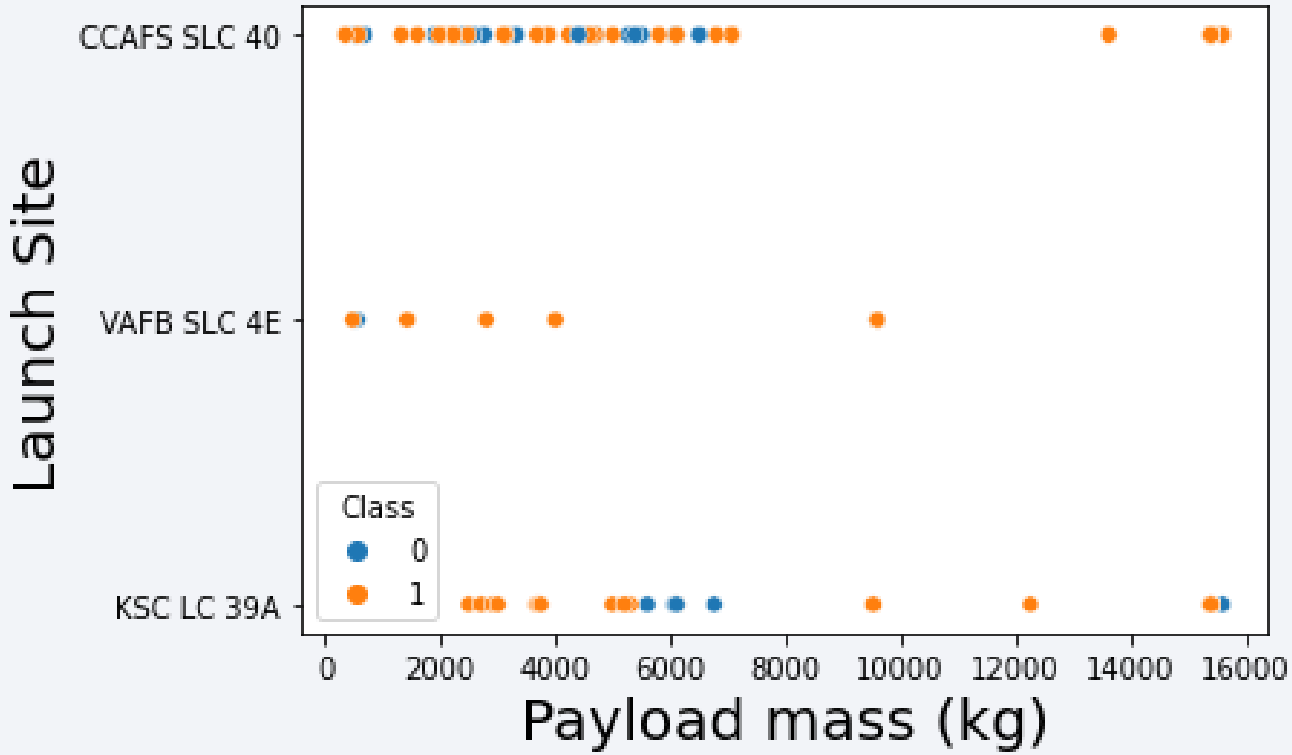
Flight Number vs. Launch Site scatter plot



Most of the landing on the CCAFS SLC 40 at the first stage was failed. Success rate increased after about 20 flight number.

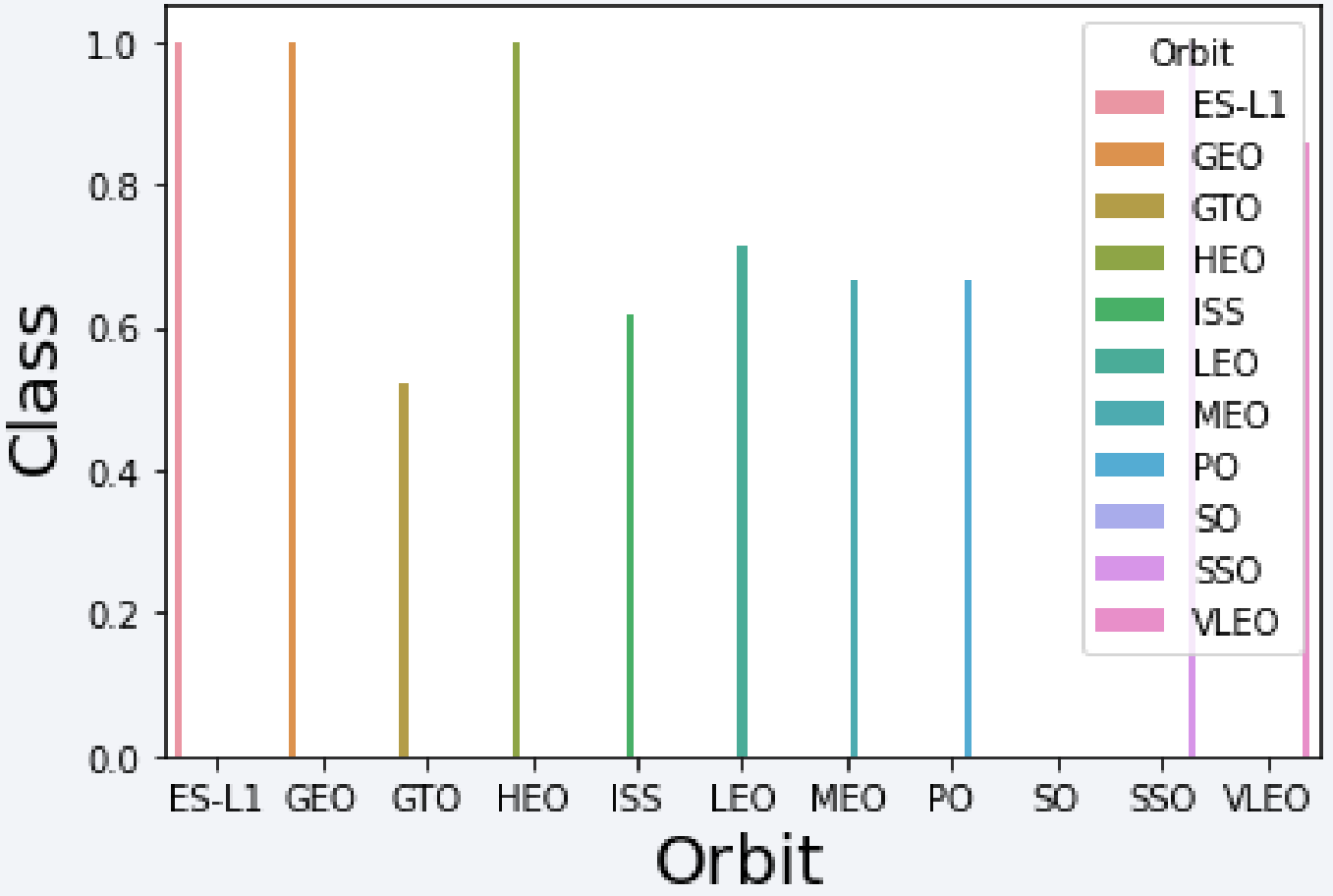
Payload vs. Launch Site

- A scatter plot of Payload vs. Launch Site
- CCAFS SLC 40 more suitable for heavy payload
- There are most of the problems with launches with payload between 4500 and 6500 kg



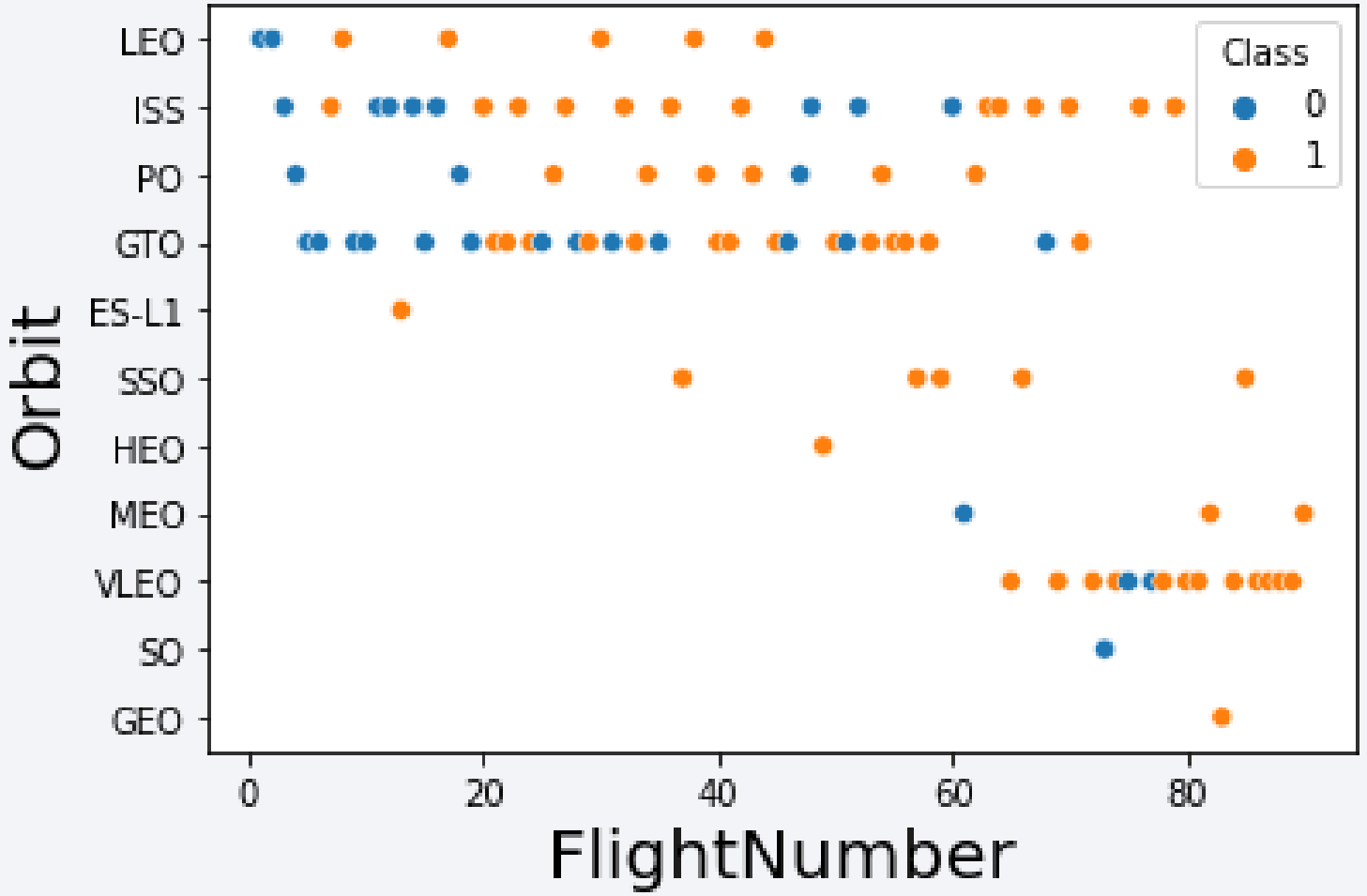
Success Rate vs. Orbit Type

- A bar chart for the success rate of each orbit type
- The least success rate is on the GTO orbit. So SpaceX cannot be used to launch geosynchronous spot for monitoring weather, communications and surveillance. The operator has to concentrate on increasing success rate on GTO orbit.
- The best rate in ES-L1, GEO, HEO and SSO orbits.



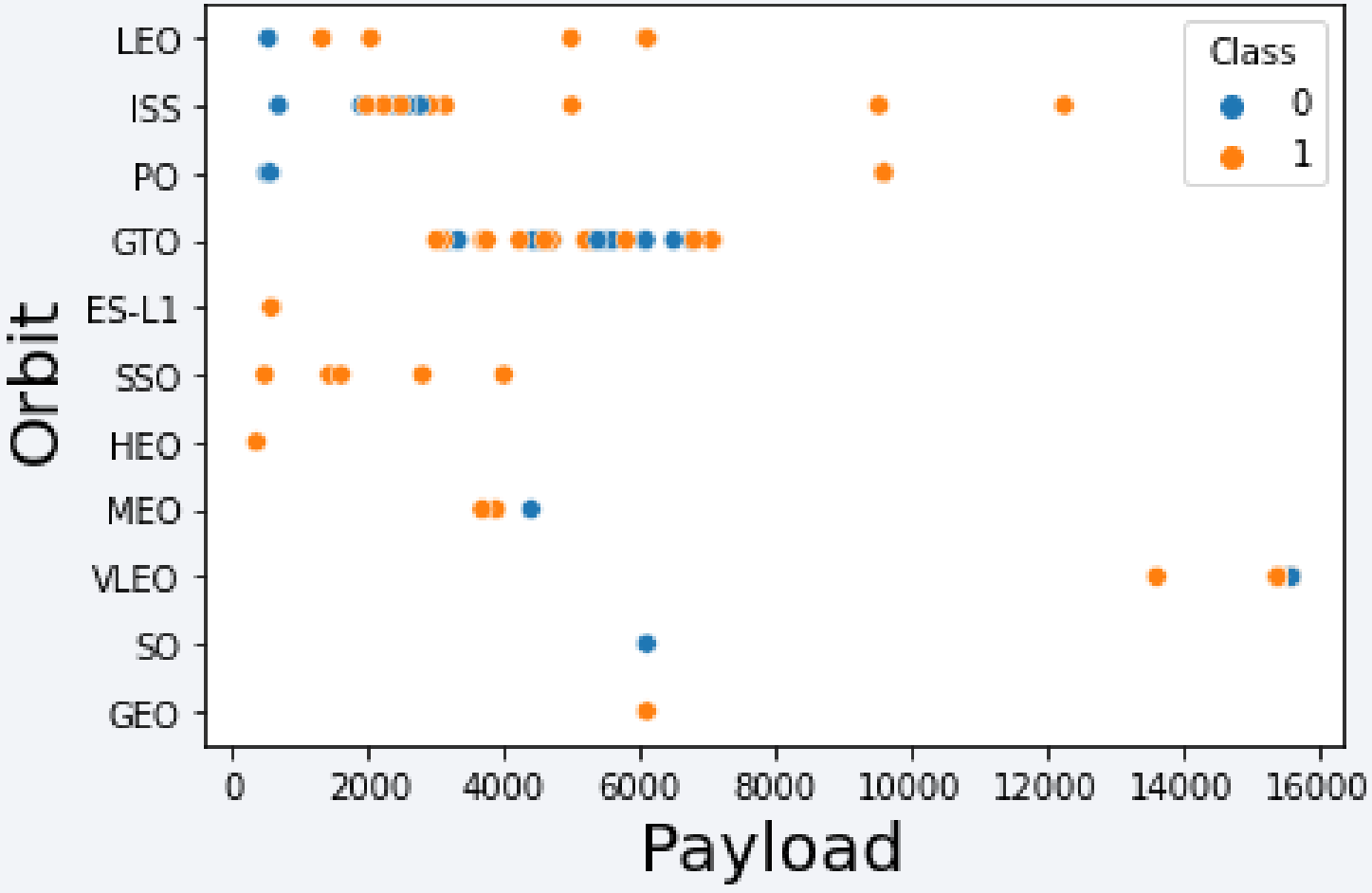
Flight Number vs. Orbit Type

- A scatter point of Flight number vs. Orbit type
- The Success appears related to the number of flights in the LEO orbit. But there is no relationship between flight number in GTO orbit.



Payload vs. Orbit Type

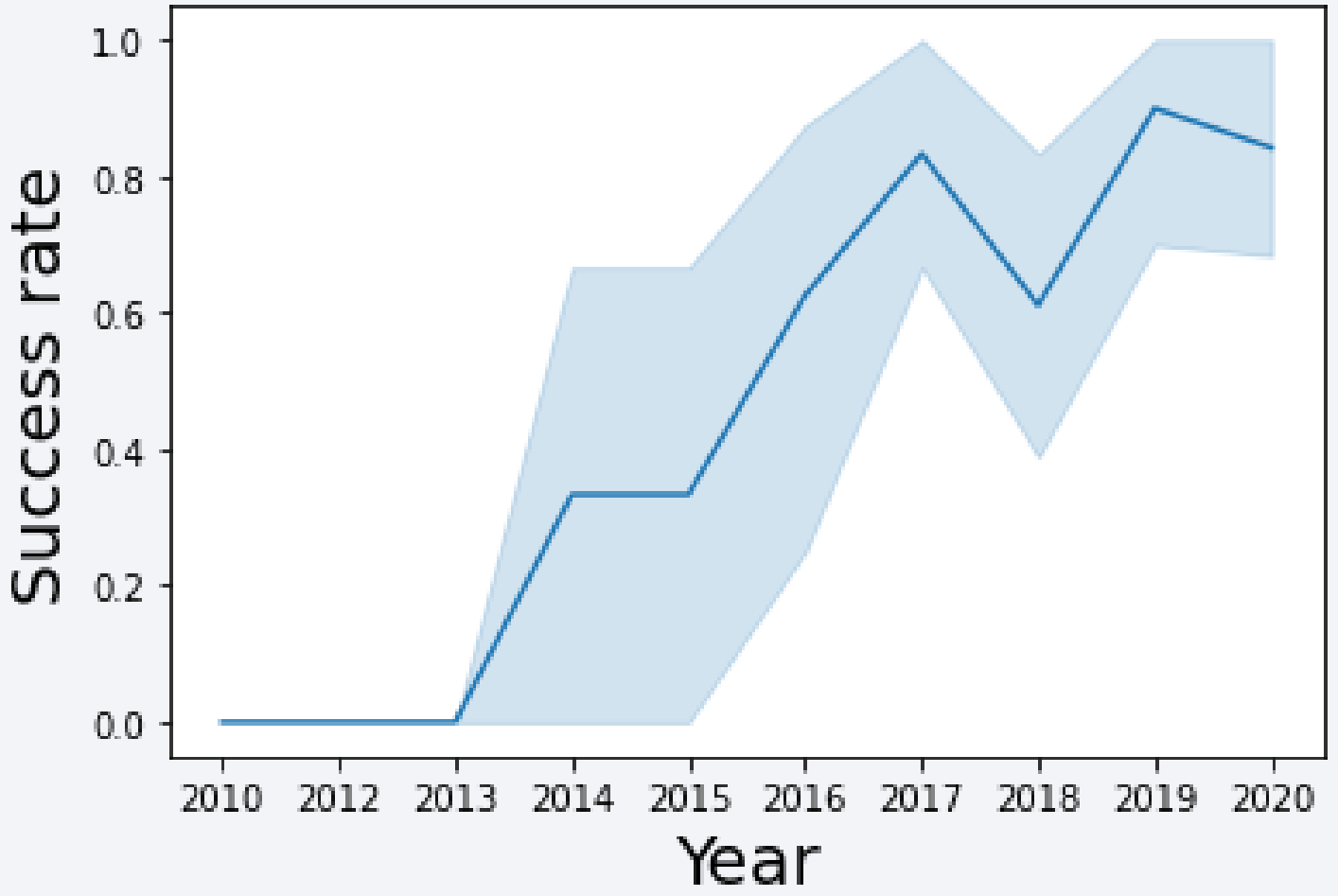
- A scatter point of payload vs. orbit type
- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- It seems, GTO success rate doesn't relate with orbit.



Launch Success Yearly Trend

- A line chart of yearly average success rate

- We can observe that the success rate since 2013 kept increasing till 2020 except 2018.



All Launch Site Names

- Find the names of the unique launch sites

- Present your query result with a short explanation here

select distinct launch_site from spacex;

SQL query select unique values of column.

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

Date	Time	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS_ _KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04 18:45:00	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	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	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	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	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

```
select * from spacex where launch_site like 'CCA%' limit 5;
```

Get values by pattern and limit result by 5 records.

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- Total payload 45,596 kg
- Payload by booster in the table
- SQL query: select sum(payload_mass__kg_) from spacex where customer = 'NASA (CRS)'
- Result:

sum(payload_mass__kg_)
45596

Booster_Version	sum(payload_mass__kg_)
F9 B4 B1039.2	2647
F9 B4 B1039.1	3310
F9 B4 B1045.2	2697
F9 B5 B1056.2	2268
F9 B5 B1058.4	2972
F9 B5 B1059.2	1977
F9 B5B1050	2500
F9 B5B1056.1	2495
F9 FT B1035.2	2205
F9 FT B1021.1	3136
F9 FT B1025.1	2257
F9 FT B1031.1	2490
F9 FT B1035.1	2708
F9 v1.0 B0006	500
F9 v1.0 B0007	677
F9 v1.1	2296
F9 v1.1 B1010	2216
F9 v1.1 B1012	2395
F9 v1.1 B1015	1898
F9 v1.1 B1018	1952

Average Payload Mass by F9 v1.1

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

2928.4 kg

- Present your query result with a short explanation here

SQL Query: select Booster_Version, avg(payload_mass__kg_) from spacex where Booster_version = 'F9 v1.1'

Result:

Booster_Version	avg(payload_mass__kg_)
F9 v1.1	2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

2015-12-22

- Present your query result with a short explanation here

SQL: select Landing_Outcome, min(Date) from spacex group by Landing_Outcome

Look at the row with landing_outcome = 'Success (ground pad)'

Landing_Outcome	min(Date)
Controlled (ocean)	2014-04-18 19:25:00
Failure	2018-12-05 18:16:00
Failure (drone ship)	2015-01-10 09:47:00
Failure (parachute)	2010-06-04 18:45:00
No attempt	2012-05-22 07:44:00
No attempt	2019-08-06 23:23:00
Precluded (drone ship)	2015-06-28 14:21:00
Success	2018-07-22 05:50:00
Success (drone ship)	2016-04-08 20:43:00
Success (ground pad)	2015-12-22 01:29:00
Uncontrolled (ocean)	2013-09-29 16:00:00

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- SQL: select booster_version, payload_mass__kg_ from spacex where landing_outcome = 'Success (drone ship)' and payload_mass__kg_ > 4000 and payload_mass__kg_ < 6000;

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

success	failure
61	10

- select sum(iif(landing_Outcome like '%Success%',1,0)) as success, sum(iif(lower(landing_Outcome) like '%failure%',1,0)) as failure from spacex;

Success landing more than failure.

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Sql query with subquery:

```
select distinct booster_version from spacex where payload_mass__kg_ = (select max(payload_mass__kg_) from spacex);
```

- A lot of type of booster have carried the maximum payload mass

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

- he failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Both of the launch site is CCAFS LC-40.

Landing_Outcome	Booster_Version	Launch_Site	year	Date
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015	2015-01-10 09:47:00
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015	2015-04-14 20:10:00

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

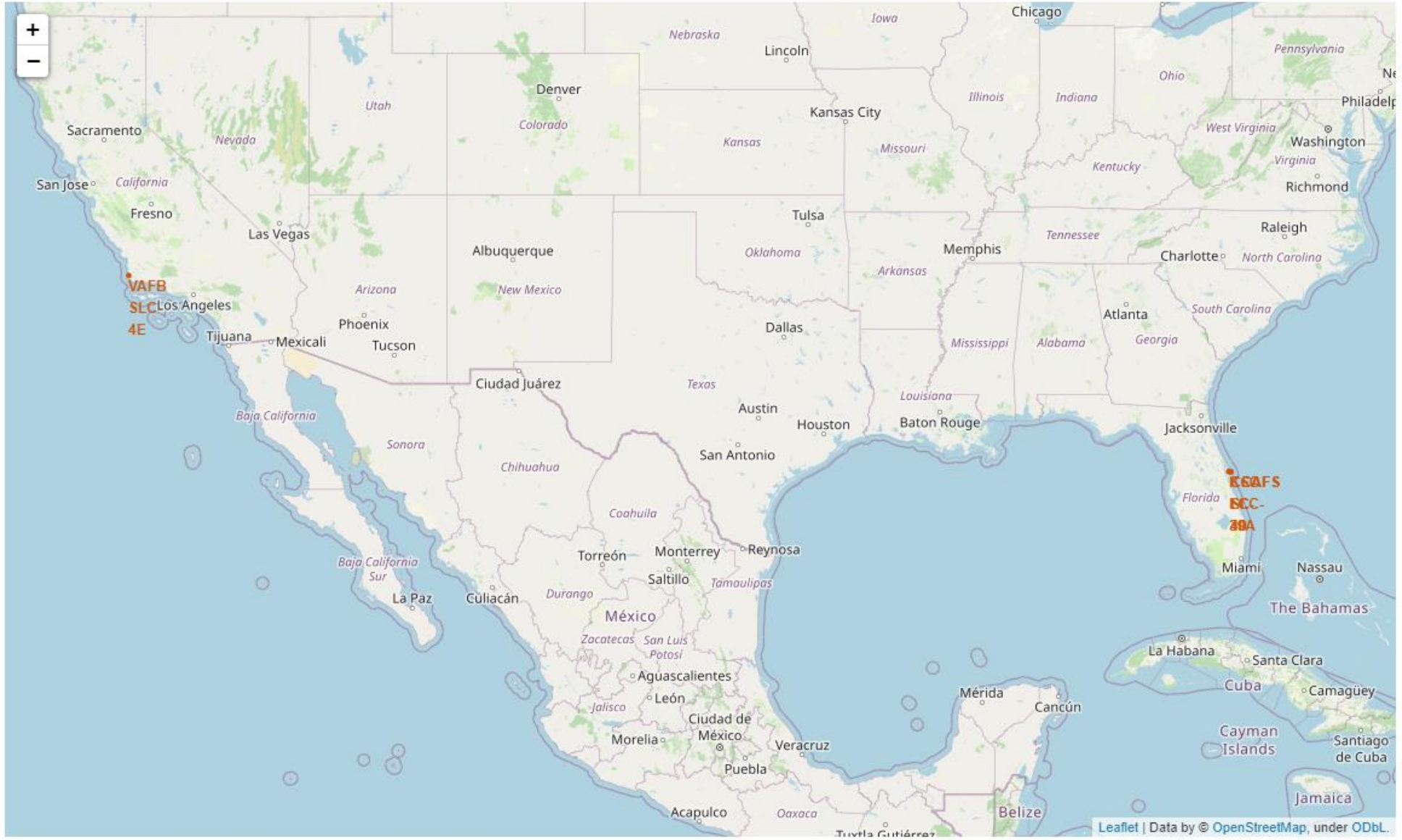
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order
- Result of the query ordered by count of landing outcome. The most occurred is “No attempt”

Landing_Outcome	count(landing_outcome)
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

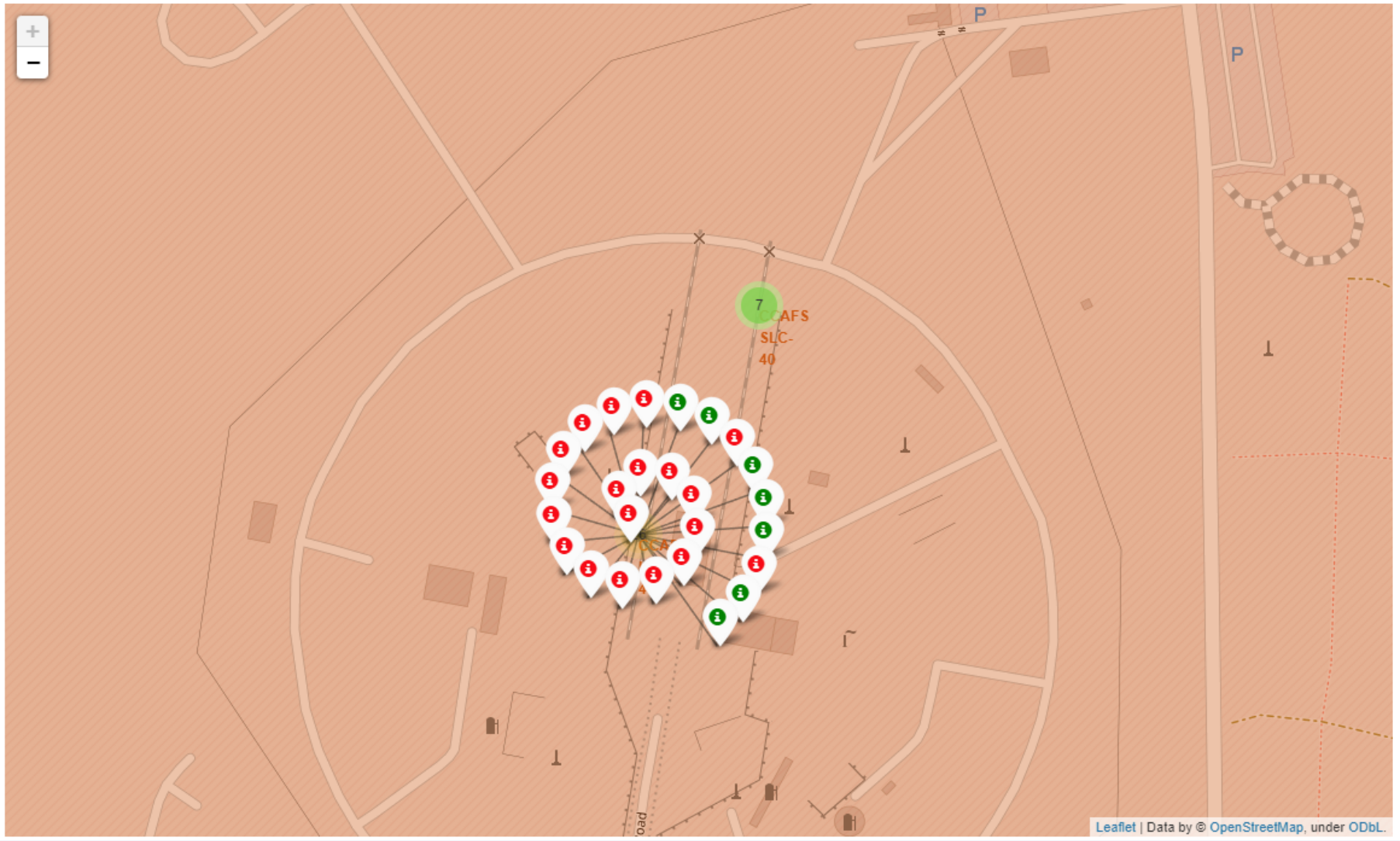
Section 3

Launch Sites Proximities Analysis

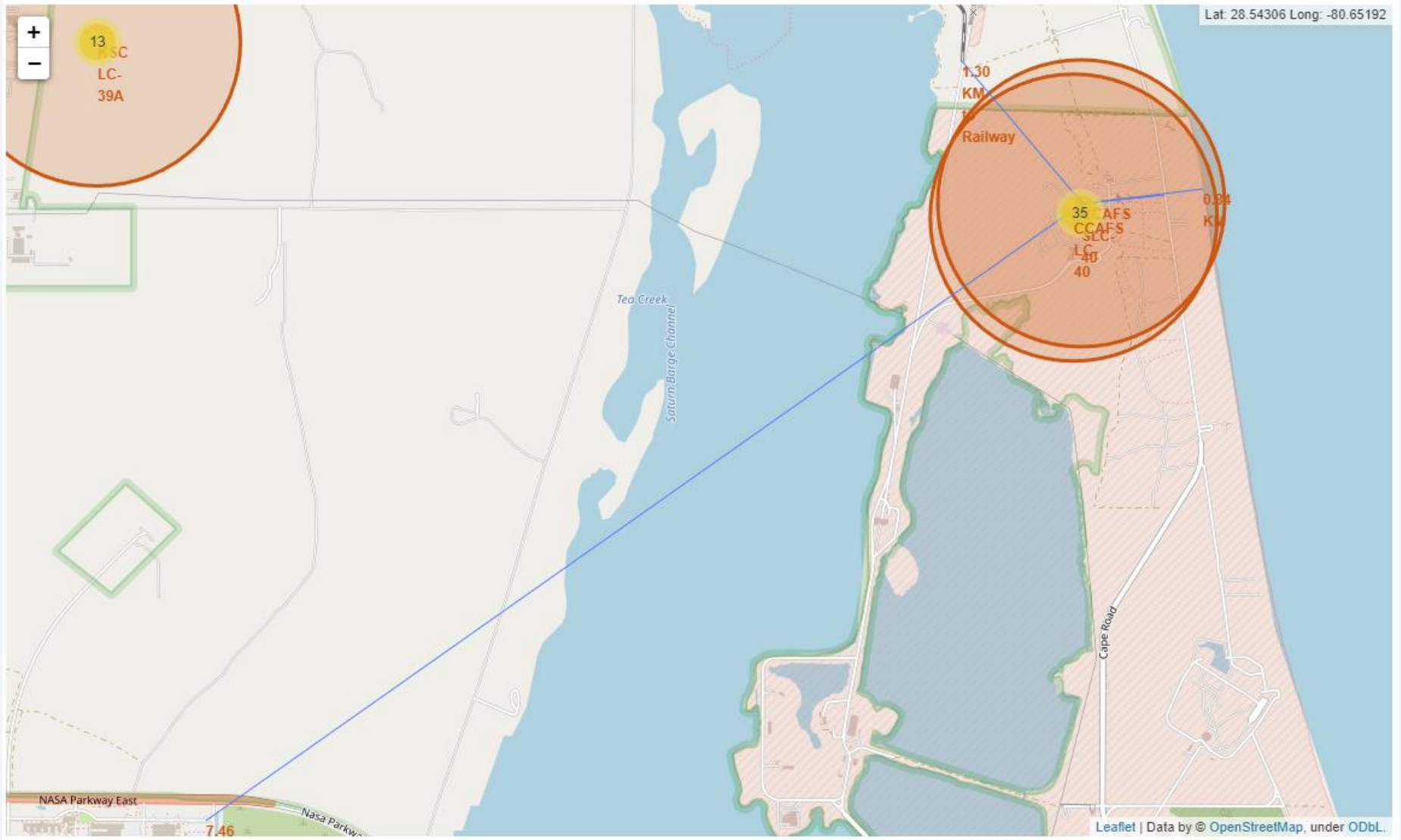
Launch sites

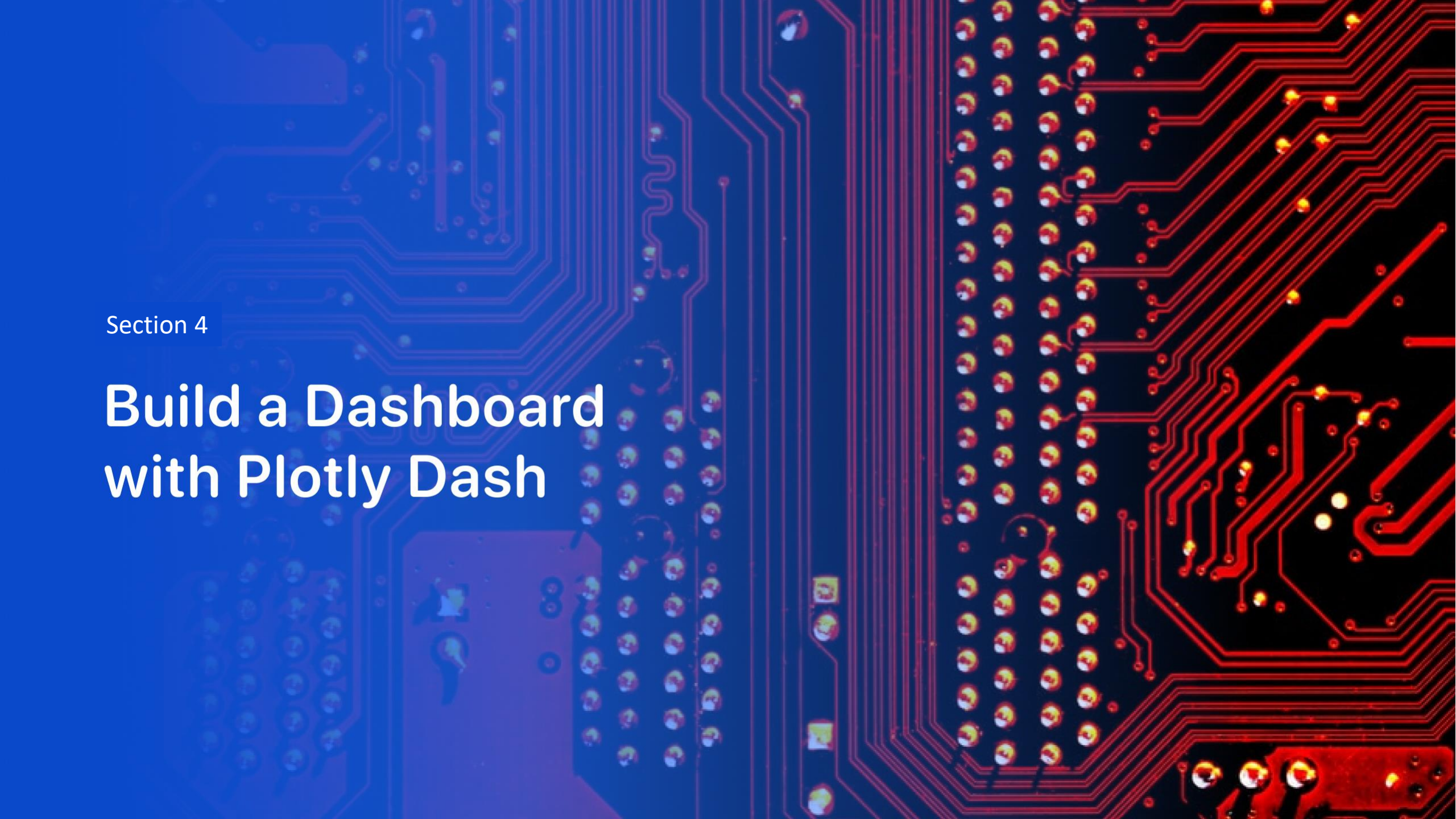


Launch outcomes



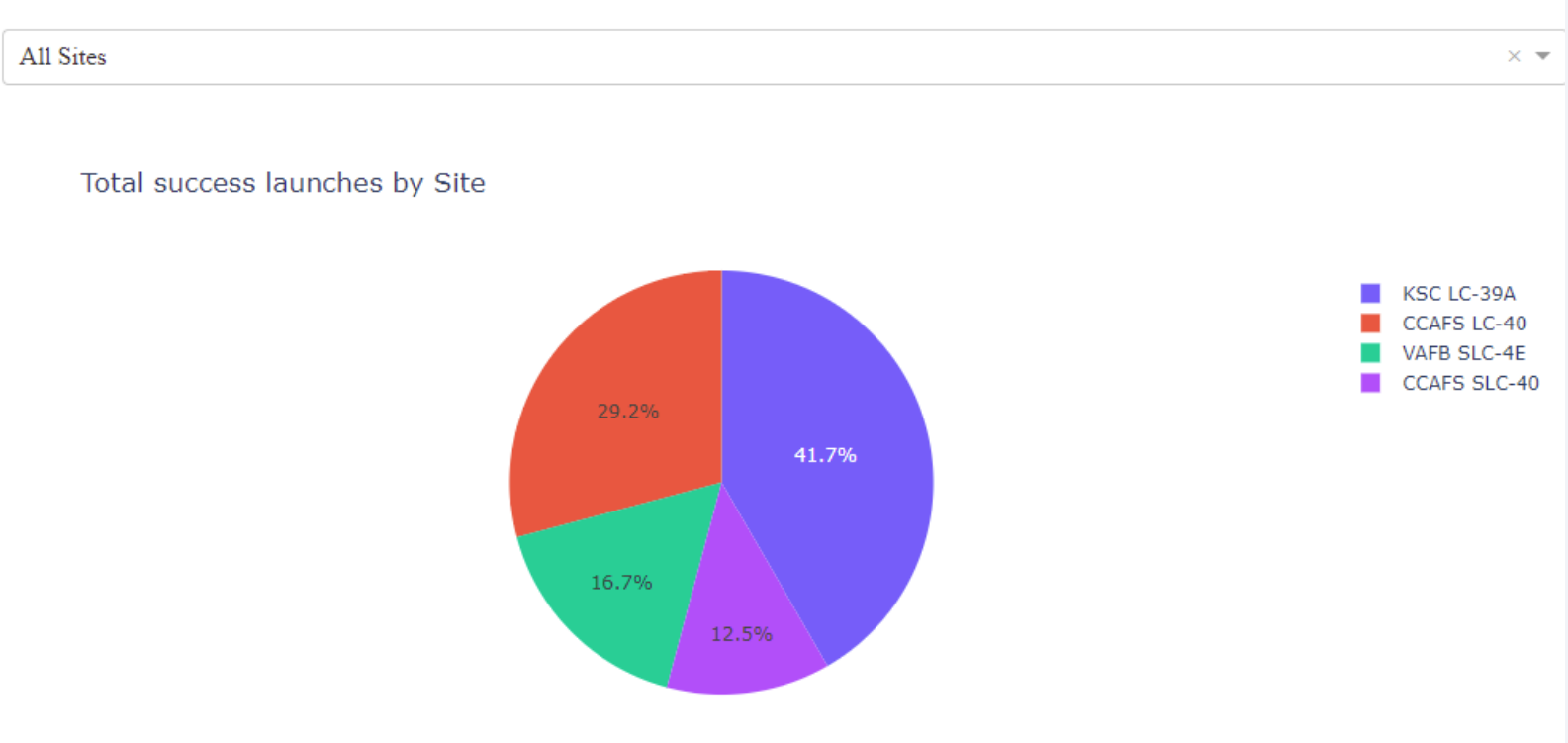
Distance to objects



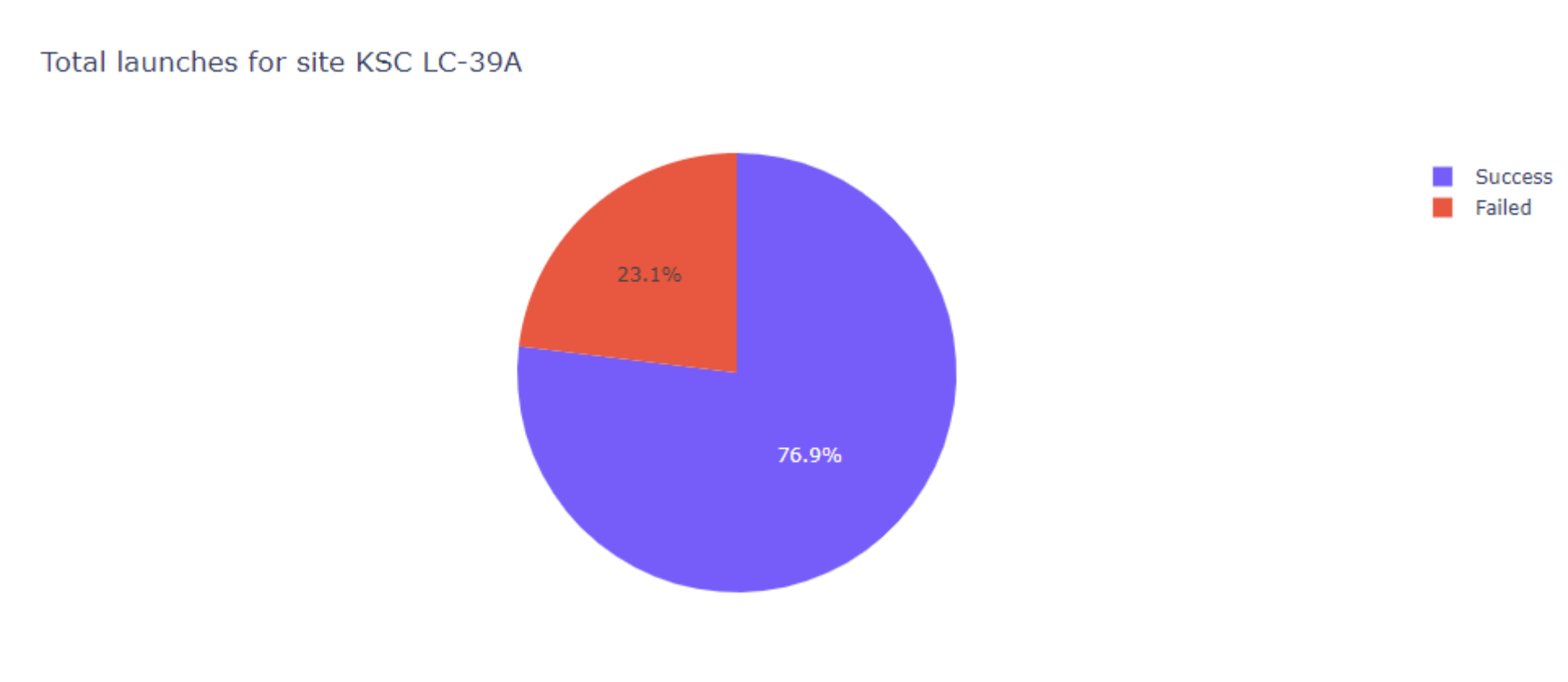

Section 4
Build a Dashboard with Plotly Dash

Total success launches by Site

The most success launches reached by KSC LC-39A site.



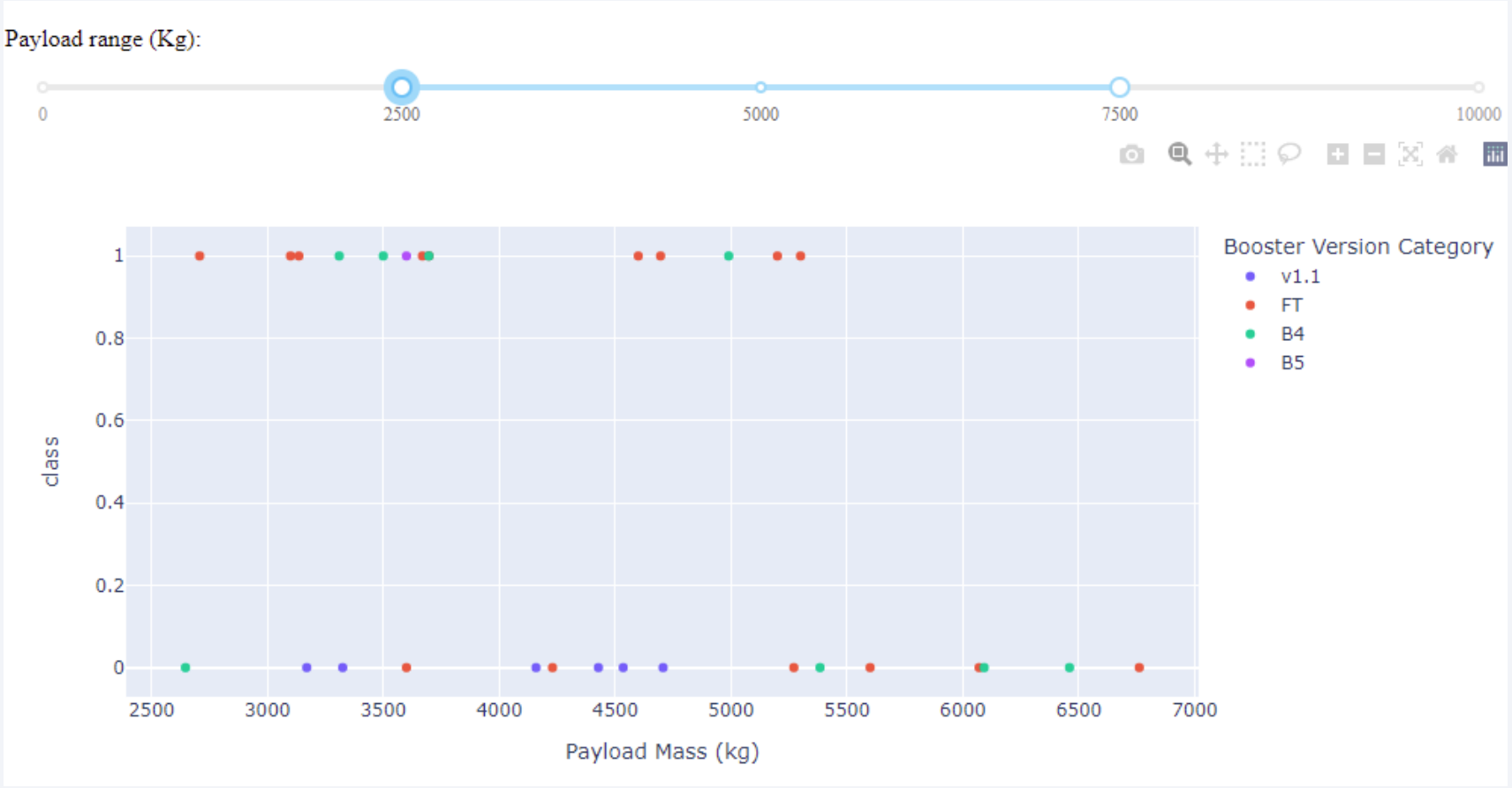
Success and failed launches by KSC LC-39A



Success rate is more than 76%

Payload vs. Launch Outcome in range 2500 to 7500 kg payload

- FT and B4 have better success rate in compare with v1.1 and B5.



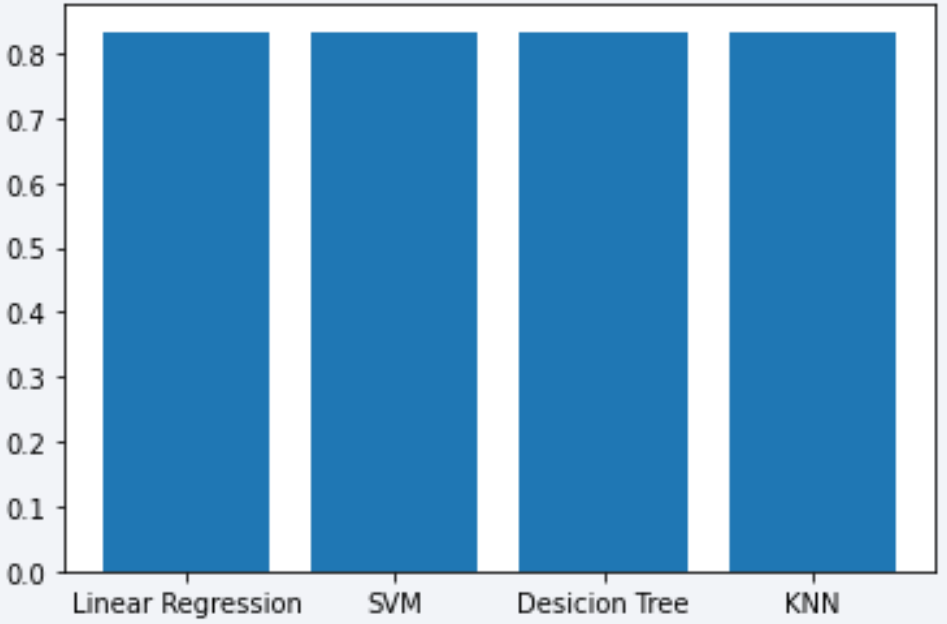
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Model accuracy for all built classification models, in a bar chart

- All model show same accuracy on the test set



Confusion Matrix

- All matrix have same view

Conclusions

- SpaceX success rate has been increasing during the research period
- GTO orbit is a problem orbit for SpaceX
- All tasks of the final assignment have done
- Different tools and methods of data analysis are used during final assignment
- Type of visualization depends on goal of visualization

Appendix

- Jupyter notebooks
  - https://github.com/vitalyvb1974/ds/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb
  - https://github.com/vitalyvb1974/ds/blob/main/jupyter-labs-eda-dataviz.ipynb
  - https://github.com/vitalyvb1974/ds/blob/main/jupyter-labs-eda-sql-coursera.ipynb
  - https://github.com/vitalyvb1974/ds/blob/main/jupyter-labs-spacex-data-collection-api.ipynb
  - https://github.com/vitalyvb1974/ds/blob/main/jupyter-labs-webscraping.ipynb
  - https://github.com/vitalyvb1974/ds/blob/main/lab_jupyter_launch_site_location.ipynb
  - https://github.com/vitalyvb1974/ds/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb
- Python code (dashboard)
  - https://github.com/vitalyvb1974/ds/blob/main/dashboard01.py
- SQLite database file (used instead of IBM DB2)
  - https://github.com/vitalyvb1974/ds/blob/main/spacex_sqlite.db

Thank you!