Winning Space Race with Data Science

Tadashi Ishikawa
01/03/2022

Outline

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

Executive Summary

- **Summary of methodologies**
  - Data collection
  - Data wrangling
  - Exploratory Analysis using SQL
  - Exploratory Analysis using Pandas and Matplotlib
  - Interactive Visual Analytics and Dashboard (Folium, Plotly Dash)
  - Predictive analysis (Classification)
- **Summary of all results**
  - Exploratory data analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Introduction

- Project background and context

  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.

- Problems you want to find answers

  To make competitive bid by the startup company, we would create predictive model which provides success rate of first stage landing of SpaceX.

Section 1

Methodology

Methodology

Executive Summary

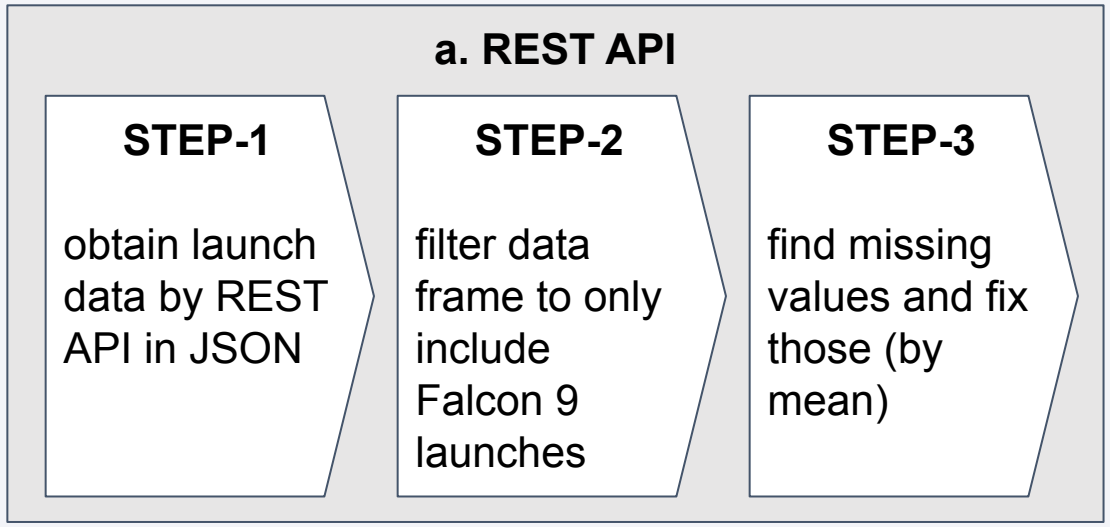
- Data collection methodology:
  - by SpaceX REST API
  - web scraping of SpaceX information on Wikipedia
- Perform data wrangling
  - fix missing values
  - encode categorical fields to One-hot vector and drop irrelevant columns
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - applied 4 classification algorithms, 1.Logistic Regression, 2.SVM, 3.Decision Tree, and 4.KNN to find the best option by using grid search method to tune hyper-parameters

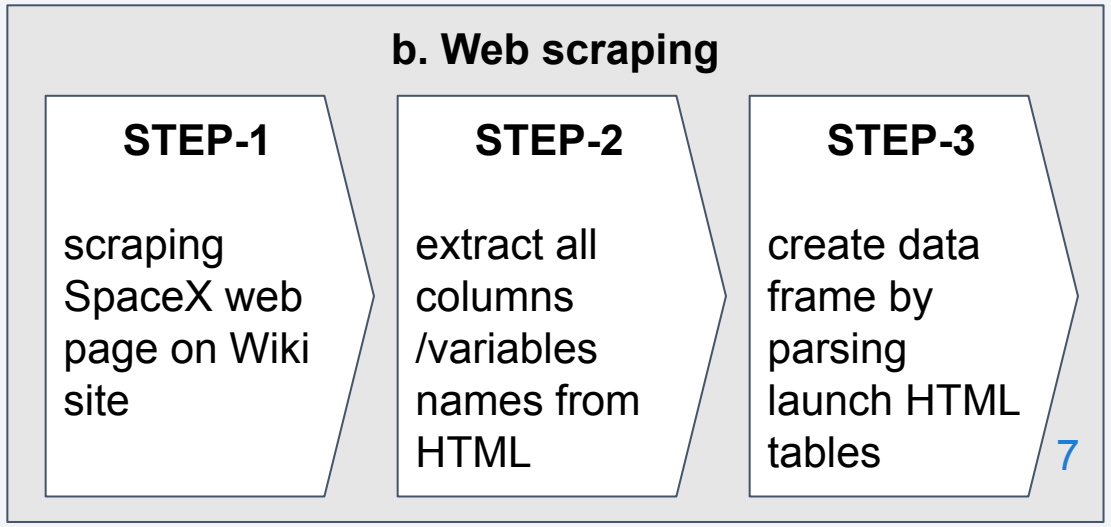
Data Collection

Data sets of Falcon 9 were collected by two approaches as follows;

a. SpaceX REST API
   obtain data of past launches about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome

b. Web scraping SpaceX pages on Wikipedia
   obtain data of past launches about launch site, payload, payload mass, orbit, launch outcome, and so on





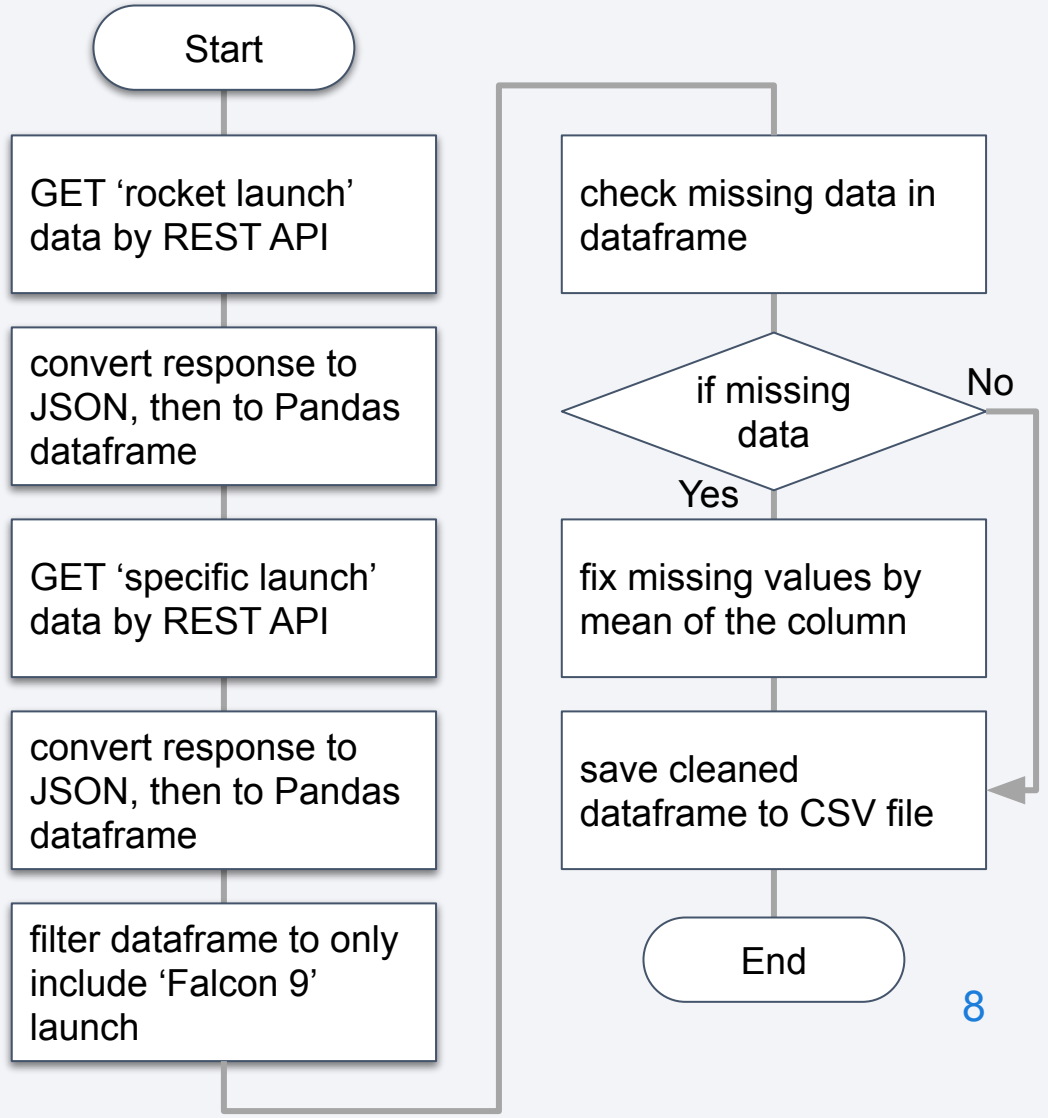
Data Collection – SpaceX API

Make a GET request to the SpaceX API. Then conduct some basic data wrangling and formatting:

- Request to the SpaceX API
  - Rocket launch data (summary)
  - Each launch data (specific)
- Clean the requested data

GitHub URL

https://github.com/kuma34989/testrepo/blob/master/Data%20Collection%20API.ipynb



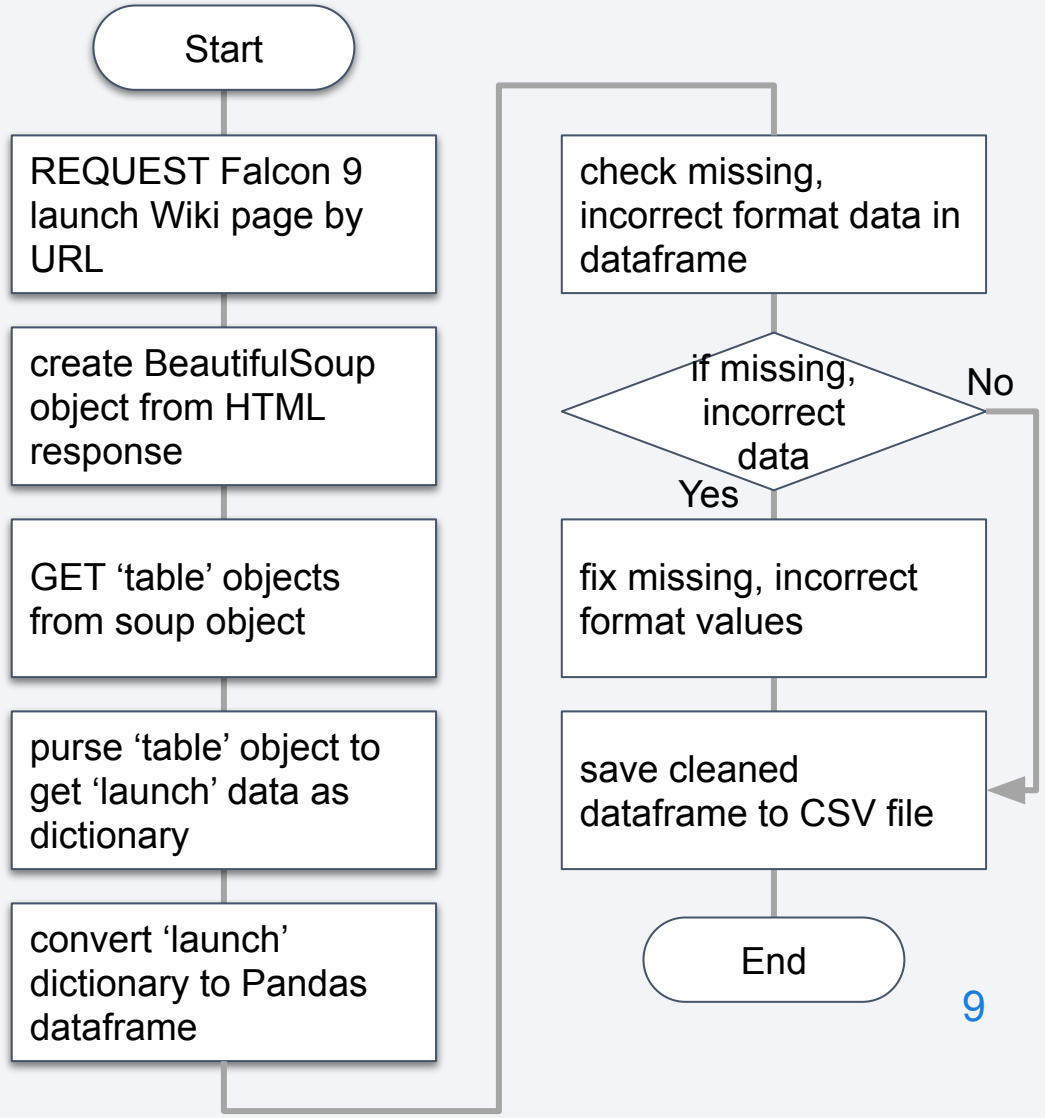
Data Collection - Scraping

Web scrap Falcon 9 launch records with BeautifulSoup:

- Extract a Falcon 9 launch records HTML table from Wikipedia
- Parse the table and convert it into a Pandas dataframe

GitHub URL

https://github.com/kuma34989/testrepo/blob/master/Data%20Collection%20with%20Web%20Scraping.ipynb



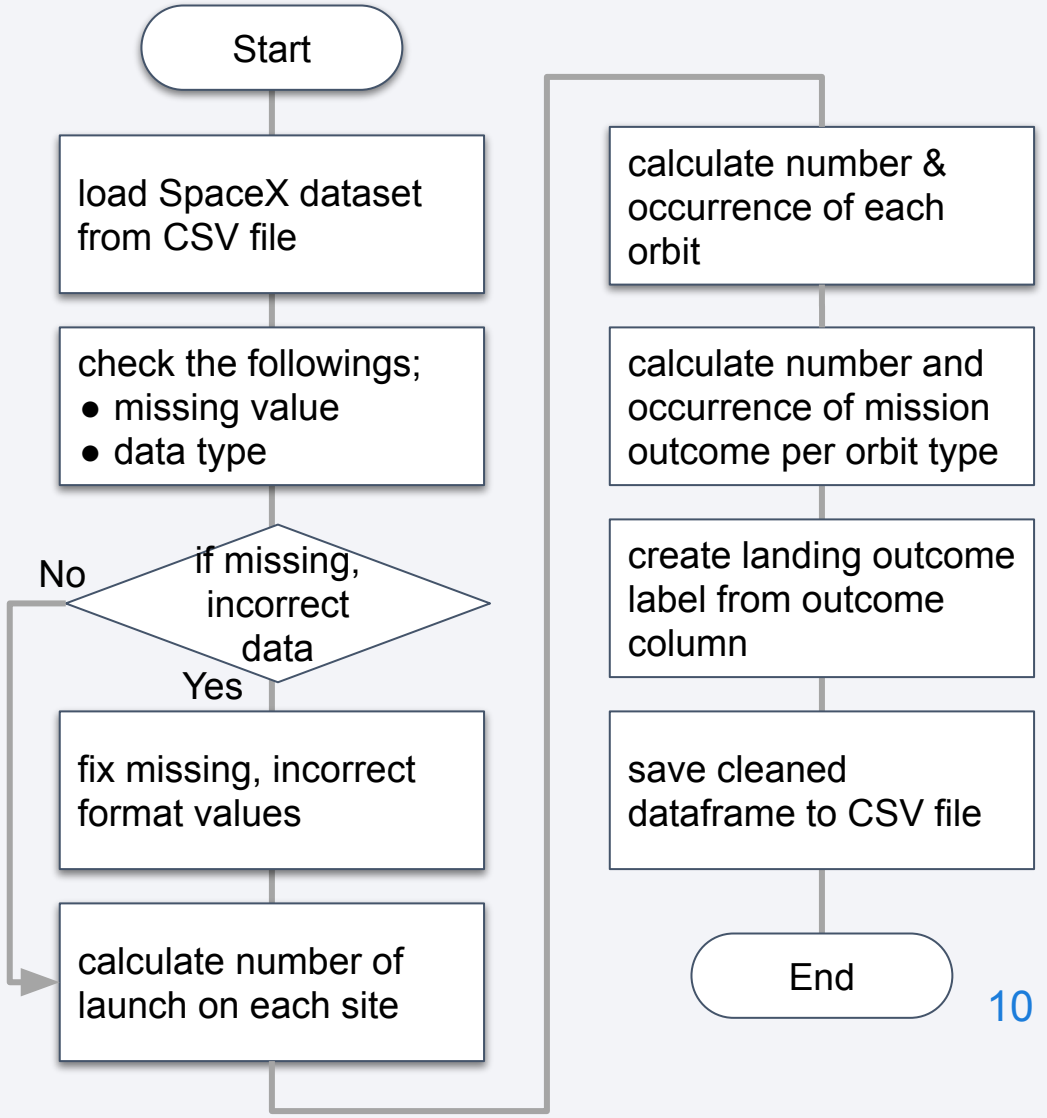
Data Wrangling

Perform exploratory Data Analysis and determine Training Labels

- Exploratory Data Analysis
- Determine Training Labels

GitHub URL

https://github.com/kuma34989/testrepo/blob/master/EDA.ipynb



EDA with Data Visualization

Charts plotted while EDA were;

Scatter chart:

- Flight Number vs. Payload Mass
- Flight Number vs. Launch Site
- Payload vs. Launch Site
- Flight Number vs. Orbit type
- Payload vs. Orbit type

(to understand relationship of each variable. Scatter plots show how much one variable is affected by another)

Bar chart:

- Launch success rate of each Orbit type

(to understand relationship between success rate vs. Orbit type. Is there any dependency ?)

Line chart:

- Launch success rate yearly trend

(to understand success rate over time. Has it been improved ?)

GitHub URL
https://github.com/kuma34989/testrepo/blob/master/EDA%20with%20Visualization.ipynb

EDA with SQL

SQL used in Ten tasks were;

1. select DISTINCT LAUNCH_SITE from spacextbl
2. select LAUNCH_SITE from spacextbl where LAUNCH_SITE like 'CCA%' limit 5
3. select sum(PAYLOAD_MASS__KG_) as total_payload_mass from spacextbl where CUSTOMER = 'NASA (CRS)'
4. select avg(PAYLOAD_MASS__KG_) as avg_payload_mass from spacextbl where BOOSTER_VERSION = 'F9 v1.1'
5. select DATE from spacextbl where lower(LANDING__OUTCOME) = 'success (ground pad)' order by DATE limit 1
6. select distinct BOOSTER_VERSION from spacextbl where lower(MISSION_OUTCOME) = 'success' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000
7. select MISSION_OUTCOME, count(*) as total_number from spacextbl group by MISSION_OUTCOME
8. select BOOSTER_VERSION, PAYLOAD_MASS__KG_ from spacextbl where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from spacextbl)
9. select BOOSTER_VERSION, LAUNCH_SITE, LANDING__OUTCOME, DATE from spacextbl where lower(LANDING__OUTCOME) not like 'success%' and DATE like '2015%'
10. select LANDING__OUTCOME, count(*) from spacextbl where DATE >= '2010-06-04' and DATE <= '2017-03-20' group by LANDING__OUTCOME order by count(*) DESC
    a. SELECT max(Date) from SPACEXTBL
    b. select min(payload_mass__kg_) from SPACEXTBL
    c. SELECT *, DAYNAME(DATE) FROM SPACEXTBL where DAYNAME(DATE)='Friday' LIMIT 5

GitHub URL

https://github.com/kuma34989/testrepo/blob/master/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

Created map objects to a folium map are as follows;

- **Circles:**
  to show four Space center locations
- **Markers:**
  to show individual Rocket launches with success and failed notion by color
- **Lines:**
  to show distance between a launch site to its proximities such as Coastline, City, Railway, and Highway

GitHub URL
https://github.com/kuma34989/testrepo/blob/master/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

Plots/graphs and interactions added to a dashboard are as follows;

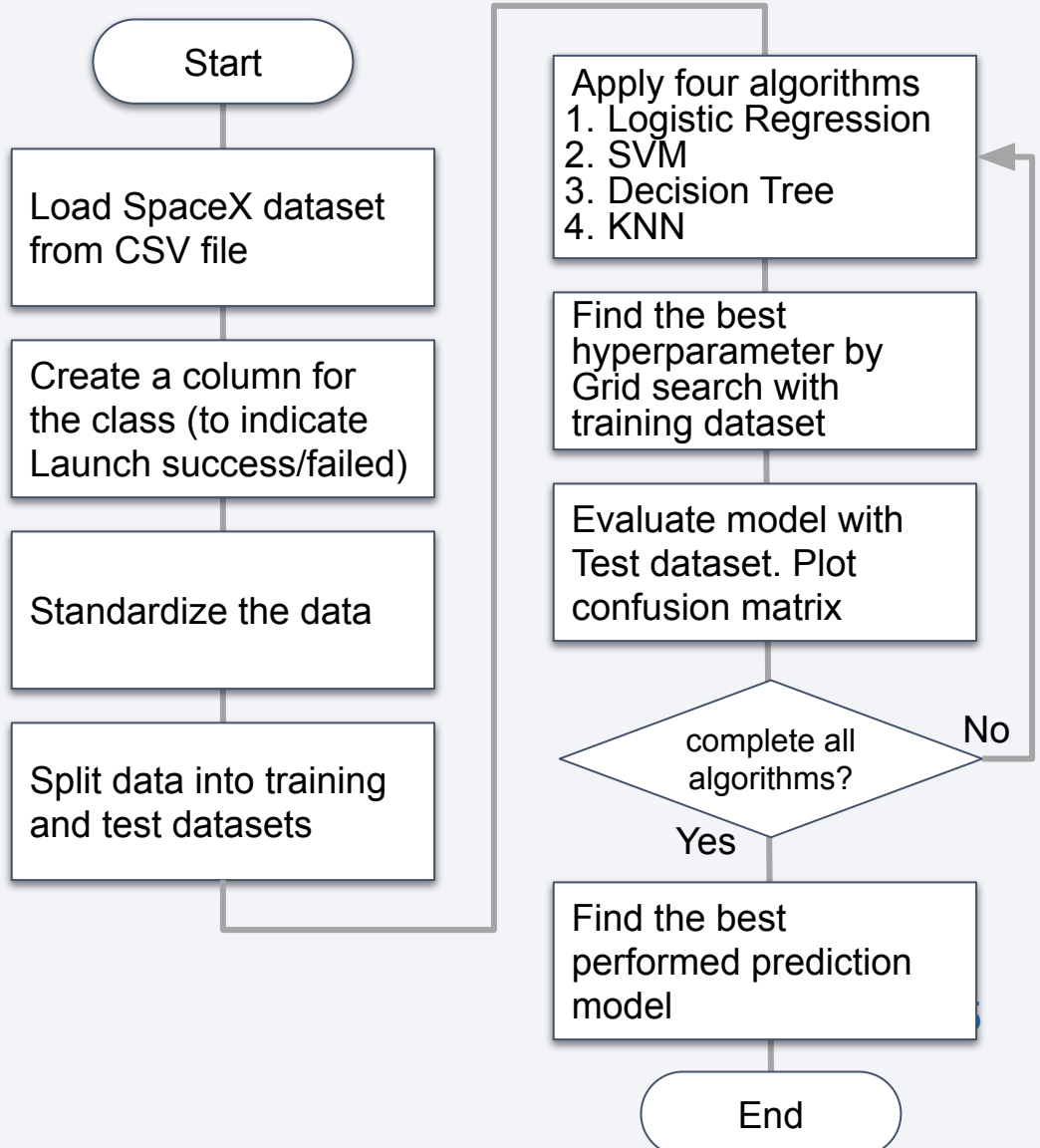
- **Drop Down list:**
  to enable Launch Site selection
- **Pie chart:**
  to show total successful launches count for all sites
- **Slider:**
  to select payload range for scatter chart
- **Scatter chart:**
  to show correlation between payload and launch success

GitHub URL
https://github.com/kuma34989/testrepo/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)

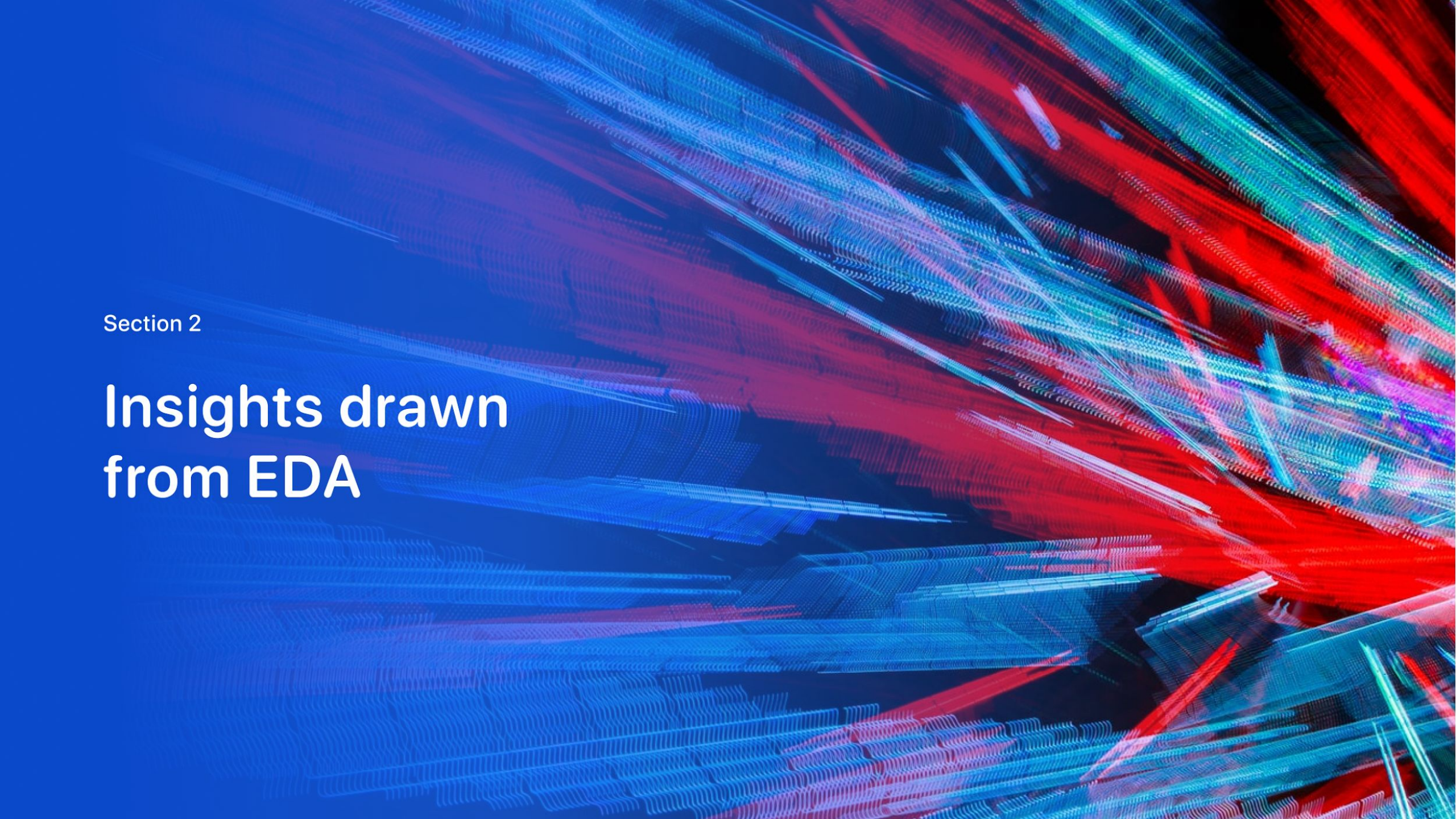
1. Perform exploratory Data Analysis and determine Training Labels
2. Establish classification model by four algorithms such as Logistic Regression, SVM, Decision Tree, and KNN
   a. Create model and find best hyperparameter by Grid search
   b. evaluate score and plot confusion matrix
3. Find the best performed prediction model among the above

GitHub URL
https://github.com/kuma34989/testrepo/blob/master/Machine%20Learning%20Prediction.ipynb



Results

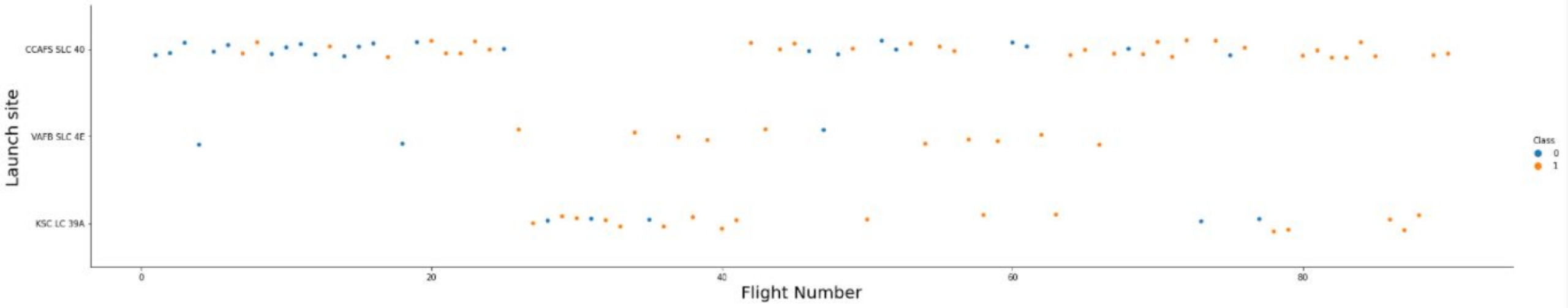
- **Exploratory data analysis results**
- **Interactive analytics demo in screenshots**
- **Predictive analysis results**


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

Findings from the chart.

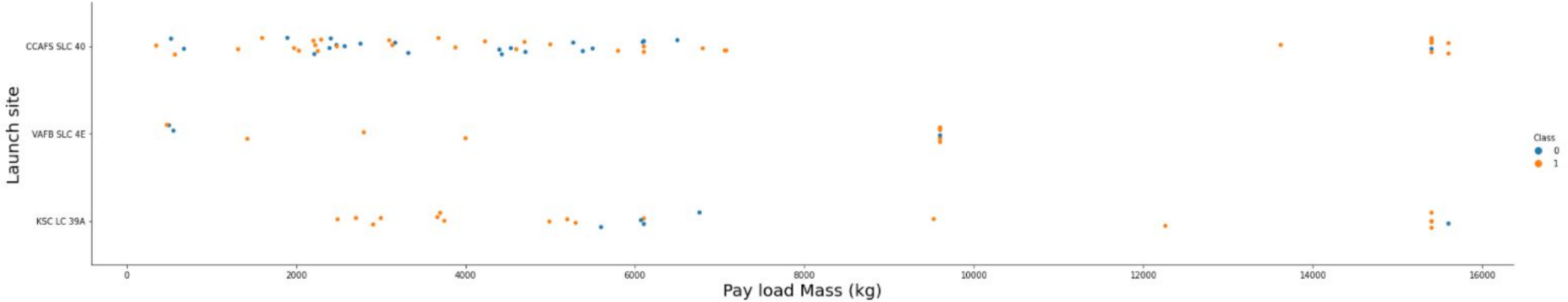
- At CCAFS SLC 40 site, success ratio of launch has been improved over-time. Up to Flight no.20 around, there were lots failed launches
- At VAFB SLC 4E site, there is NO rocket launch after Flight no.66
- KSC LC 39A seems to be a backup site of CCAFS SLC 40



Payload vs. Launch Site

Findings from the chart.

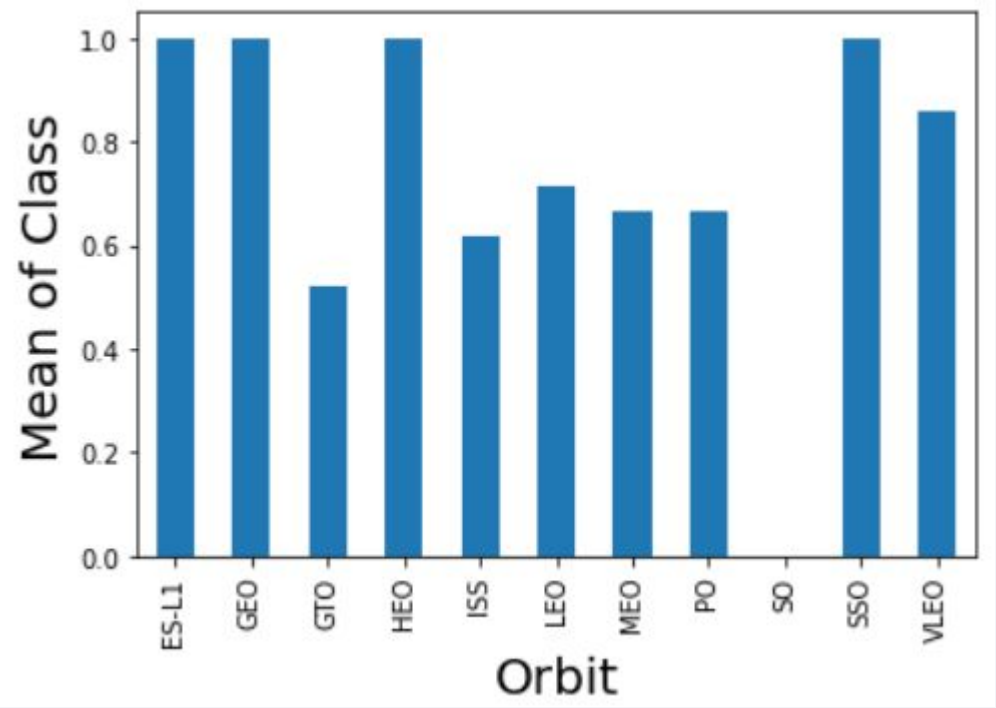
- At VAFB-SLC site, there is no rocket launched for heavy payload mass (greater than 10,000 kg)
- KSC LC 39A looks better than CCAFS SLC 40 in terms of success rate of light to medium payload mass (< 6,000 kg)



Success Rate vs. Orbit Type

Findings from the chart.

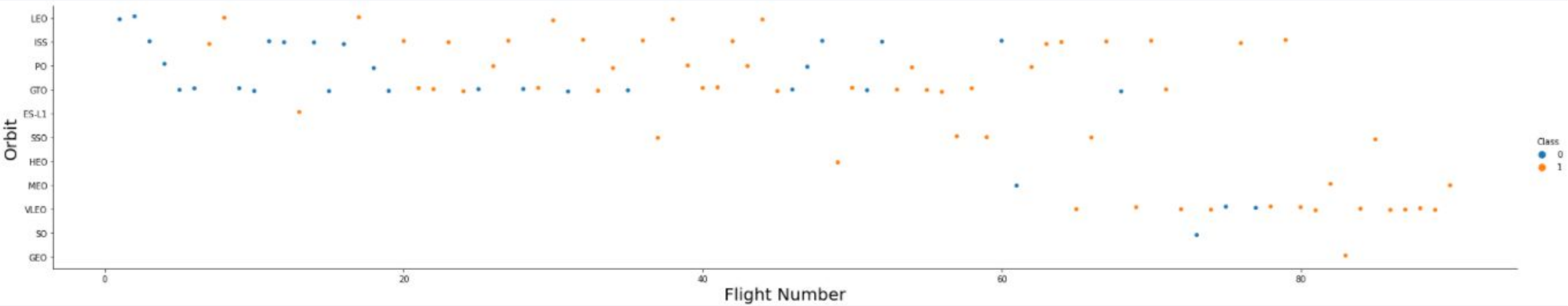
- Orbit ES-L1, GEO, HEO, SSO, and VLEO are higher in success ratio than others (Though ES-L1, GEO, HEO were only once respectively in records. Demand of VLEO(Very Low Earth Orbit) has been increased recently. Higher success ratio of it would be appreciated)



Flight Number vs. Orbit Type

Findings from the chart.

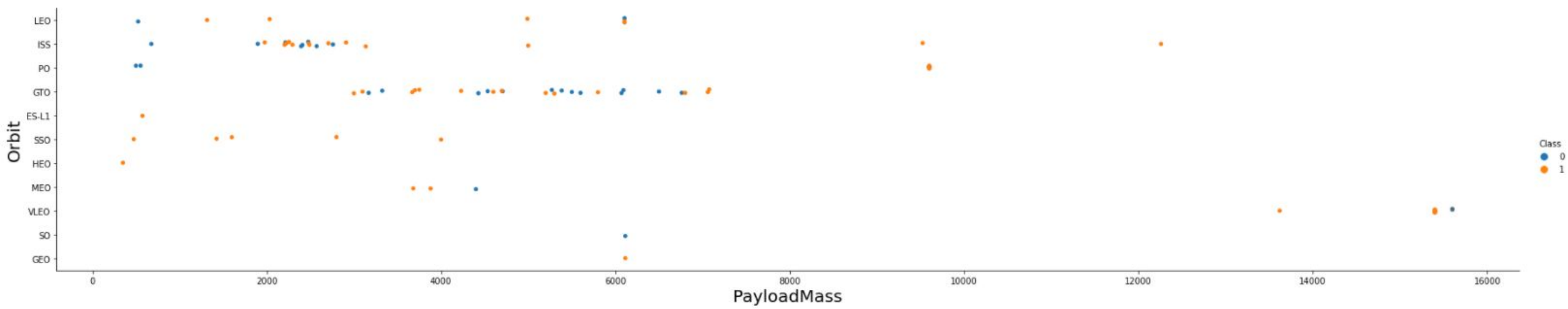
- GTO(Geostationary Transfer Orbit), ISS(International Space Station) are most popular options for satellite orbit
- Recently demand of VLEO (Very Low Earth Orbit) has been increased after Flight no. 60
- ES-L1, GEO, HEO were only once respectively in records



Payload vs. Orbit Type

Findings from the chart.

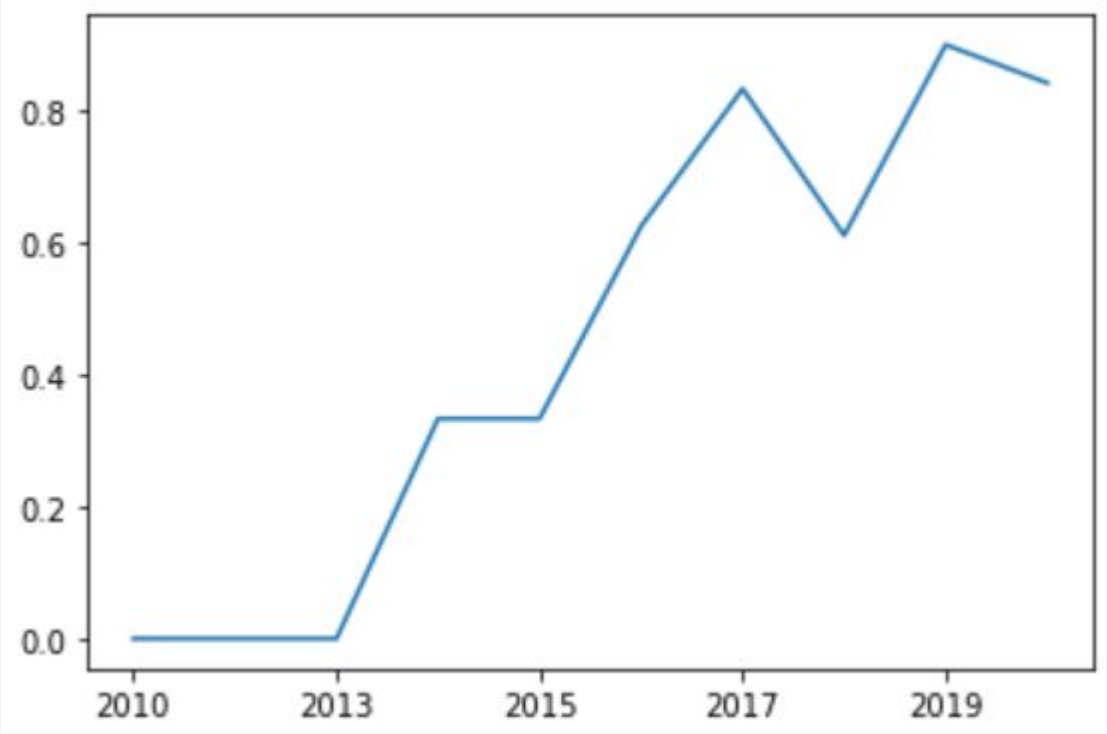
- With heavy payloads, success rate is more for Polar, LEO and ISS orbits
- For LSS, there are certain failed cases in light to medium payload (< 7,000 kg)
- For GTO, we cannot distinguish this well as both positive landing rate and negative landing



Launch Success Yearly Trend

Findings from the chart.

- Success rate of launch has been improved over time since 2013 as turning point until 2020



All Launch Site Names

List the names of the unique launch sites in the space mission

ANS: CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB-4E

TIPS:

Use keyword DISTINCT to show only unique values in LAUNCH_SITE column

```
In [11]: %sql select DISTINCT LAUNCH_SITE from spacextbl

 * ibm_db_sa://jjt62418:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:32286/BLUDB
Done.
```

Out[11]:

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

Launch Site Names Begin with 'CCA'

Query 5 records where launch sites begin with the string 'CCA'

ANS: as follows.

TIPS:

Use keyword LIKE with ‘CCA%’ to obtain LAUNCH_SITE start with word ‘CCA’

In [13]:
```
%sql select LAUNCH_SITE from spacextbl where LAUNCH_SITE like 'CCA%' limit 5
```

```
 * ibm_db_sa://jjt62418:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:32286/BLUDB
Done.
```

Out[13]:

launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

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

ANS: 45,596 kg

TIPS:

Use function SUM to obtain total of ‘PAYLOAD_MASS__KG_’ after filtering ‘CUSTOMER’ only as ‘NASA (CRS)’

```
In [20]: %sql select sum(PAYLOAD_MASS__KG_) as total_payload_mass from spacextbl where CUSTOMER = 'NASA (CRS)'
 * ibm_db_sa://jjt62418:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:32286/BLUDB
Done.
Out[20]: total_payload_mass
         45596
```

Average Payload Mass by F9 v1.1

Query average payload mass carried by booster version F9 v1.1

ANS: 2,928 kg

TIPS:

Use function ‘AVG’ to obtain average of ‘PAYLOAD_MASS__KG_’ after filtering ‘BOOSTER_VERSION’ only as ‘F9 v1.1’

```
In [22]: %sql select avg(PAYLOAD_MASS__KG_) as avg_payload_mass from spacextbl where BOOSTER_VERSION = 'F9 v1.1'
 * ibm_db_sa://jjt62418:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:32286/BLUDB
Done.
Out[22]: avg_payload_mass
         2928
```

First Successful Ground Landing Date

Query the date when the first successful landing outcome in ground pad was achieved.

ANS: 2015-12-22

TIPS:

Use keyword ‘ORDER BY’ and ‘LIMIT’ to obtain ‘DATE’ of the first successful landing

```
In [31]: %sql select DATE from spacextbl where lower(LANDING__OUTCOME) = 'success (ground pad)' order by DATE limit 1
 * ibm_db_sa://jjt62418:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:32286/BLUDB
Done.
Out[31]:
DATE
2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

Query the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

ANS: as shown right

```
In [32]: %sql select distinct BOOSTER_VERSION from spacextbl where lower(MISSION_OUTCOME) = 'success' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000
 * ibm_db_sa://jjt62418:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:32286/BLUDB
Done.
```

Out[32]:

booster_version
F9 B4 B1040.2
F9 B4 B1040.1
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5 B1058.2
F9 B5B1054
F9 B5B1060.1
F9 B5B1062.1
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1032.2
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1032.1
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016

TIPS:

Use operators ‘>’ and ‘<’ to specify condition of ‘PAYLOAD_MASS__KG_’.

Also use keyword ‘DISTINCT’ to get unique ‘BOOSTER_VERSION’

Total Number of Successful and Failure Mission Outcomes

Query the total number of successful and failure mission outcomes.

ANS: Success: 99, Failure: 1

TIPS:

Use keyword ‘GROUP BY’ to consolidate ‘MISSION_OUTCOME’ either as success or failed

In [35]:
```
%sql select MISSION_OUTCOME, count(*) as total_number from spacextbl group by MISSION_OUTCOME
```

```
 * ibm_db_sa://jjt62418:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:32286/BLUDB
Done.
```

Out[35]:

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

Boosters Carried Maximum Payload

Query the names of the booster_versions which have carried the maximum payload mass. Use a subquery.

```
In [37]: %sql select BOOSTER_VERSION, PAYLOAD_MASS__KG_ from spacextbl where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from spacextbl)
 * ibm_db_sa://jjt62418:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:32286/BLUDB
Done.
```

Out[37]:

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

ANS: as shown left

TIPS:

Use sub-query to obtain maximum value of ‘PAYLOAD_MASS__KG_’ then filter records by using the value

2015 Launch Records

Query the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015.

ANS: as shown below

TIPS:

Use keyword ‘LIKE’ and wildcard ’%’ to filter ‘LANDING_OUTCOME’ and ‘DATE’ as required

In [43]:
```
%sql select BOOSTER_VERSION, LAUNCH_SITE, LANDING__OUTCOME, DATE from spacextbl where lower(LANDING__OUTCOME) not like 'success%' and DATE like '2015%'
```

```
 * ibm_db_sa://jjt62418:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:32286/BLUDB
Done.
```

Out[43]:

booster_version	launch_site	landing__outcome	DATE
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)	2015-01-10
F9 v1.1 B1013	CCAFS LC-40	Controlled (ocean)	2015-02-11
F9 v1.1 B1014	CCAFS LC-40	No attempt	2015-03-02
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)	2015-04-14
F9 v1.1 B1016	CCAFS LC-40	No attempt	2015-04-27
F9 v1.1 B1018	CCAFS LC-40	Precluded (drone ship)	2015-06-28

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.

ANS: as shown below

TIPS:

Use ‘GROUP BY’ and ‘ORDER BY’ to filter and sort landing outcomes

In [46]:
```
%sql select LANDING__OUTCOME, count(*) from spacextbl where DATE >= '2010-06-04' and DATE <= '2017-03-20' group by LANDING__OUTCOME order by count(*) DESC
```

```
 * ibm_db_sa://jjt62418:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:32286/BLUDB
Done.
```

Out[46]:

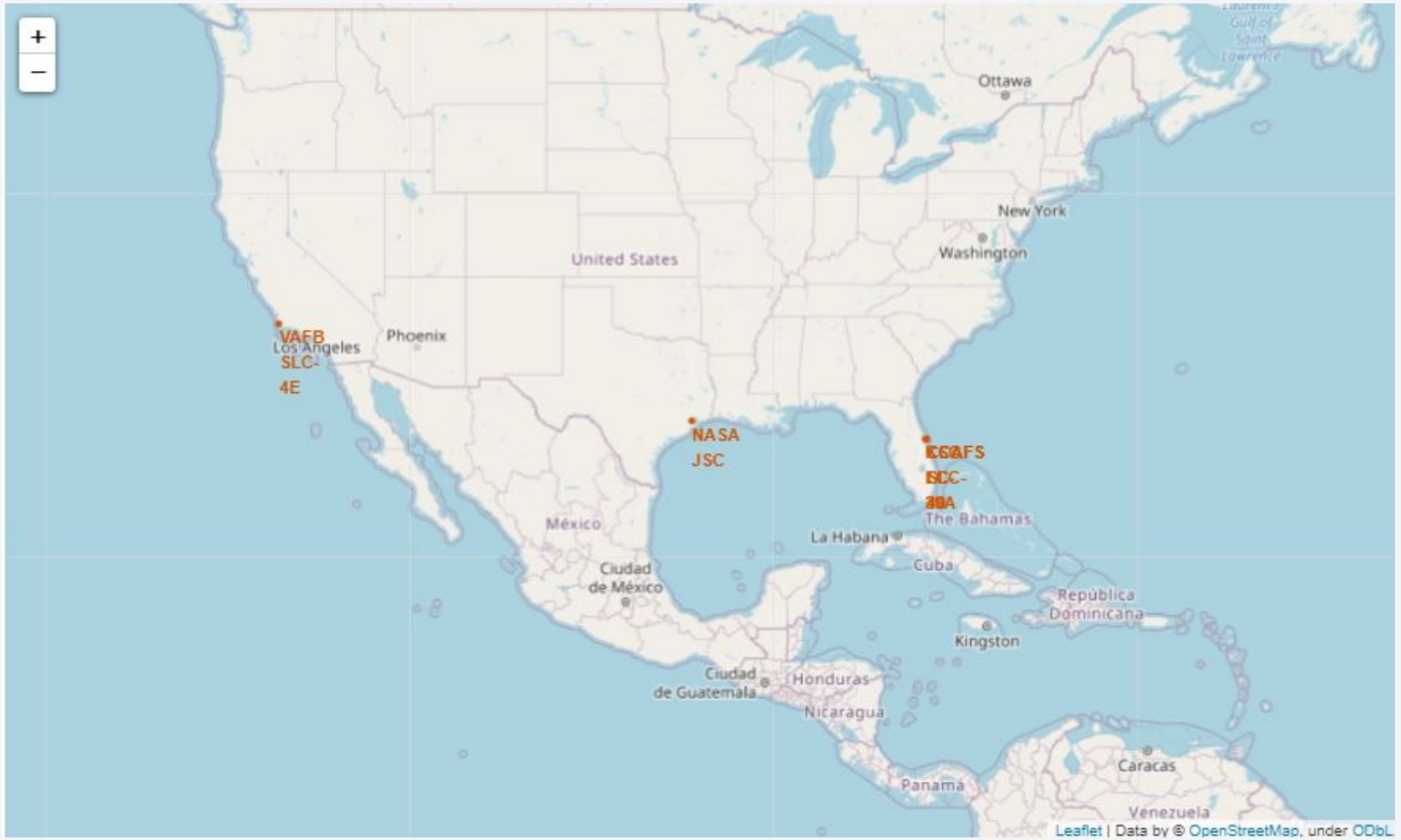
landing__outcome	2
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 4

Launch Sites Proximities Analysis

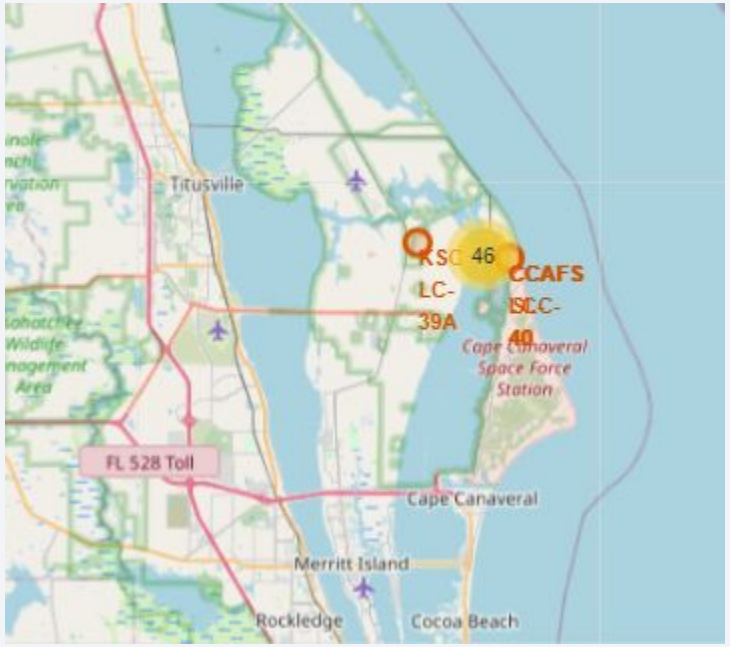
Rocket Launch sites in United States

Four launch sites in California and Florida are displayed together with NASA Johnson Space Center, Texas.



Rocket Launches per Success or Failed

By using MarkerCluster object of Folium, add set of markers to the map.
For instance, at Cape Canaveral Space Force Station (CCSFS), FL,
there were 46 rocket launches and each of launch in respective sites are displayed with Success/Failed notion by color (Green/Red)

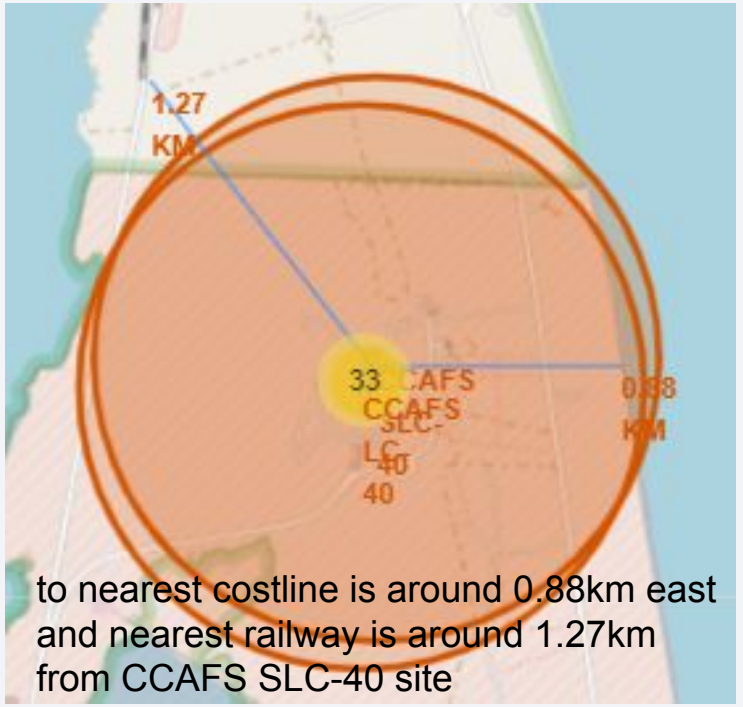


Distance from Launch site to major Landmarks

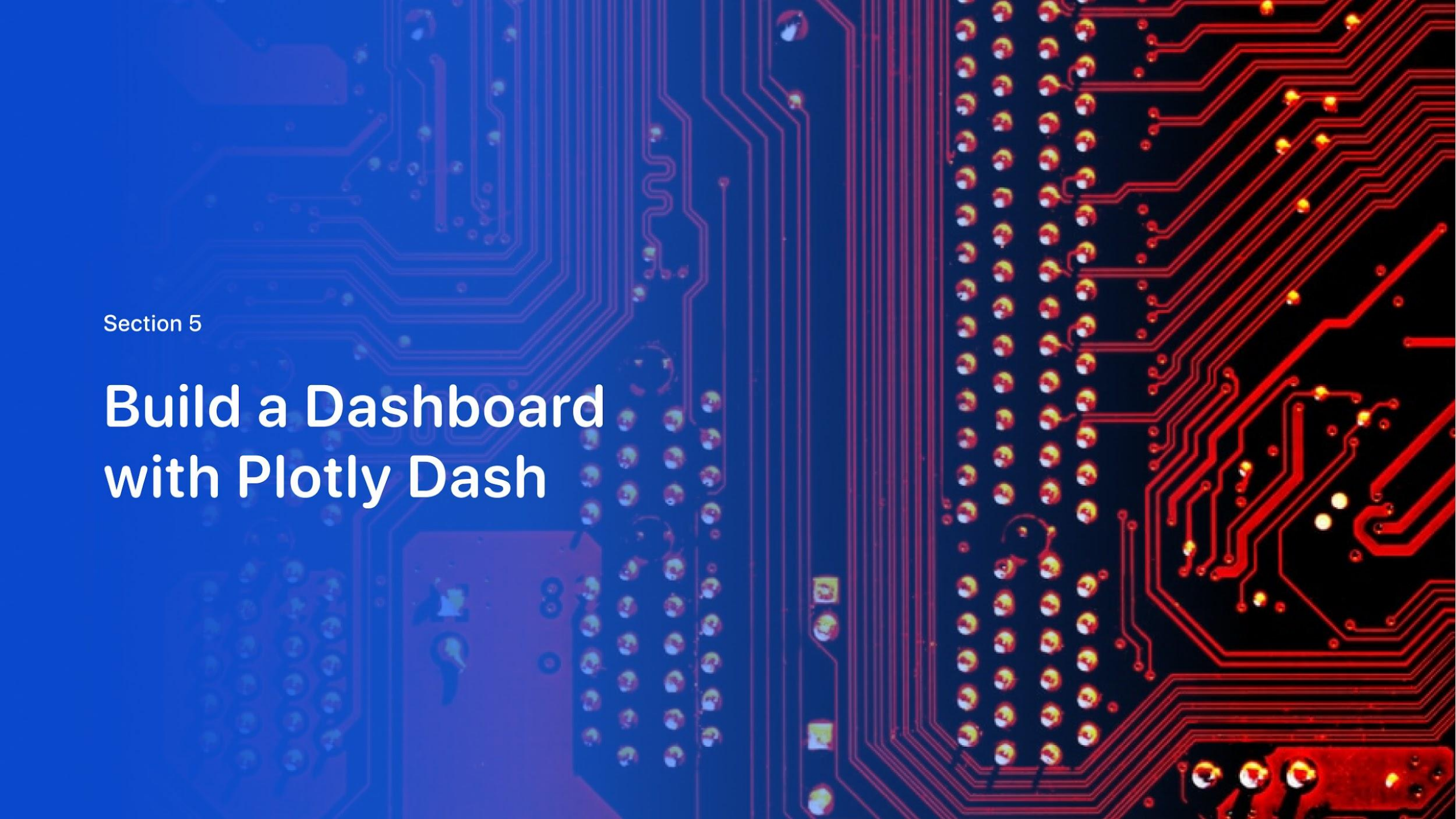
By using custom function to compute Distance between two locations, I measured it then plot line and marker in to the map.

For instance, from CCAFS SLC-40 site, distance to Nearest coastline was about 0.88km east and to nearest railway was about 1.27km respectively.

to nearest city, Titusville is located around 16.34km west
and nearest highway, US-1 is around 14.75km from KSC LC-39 site

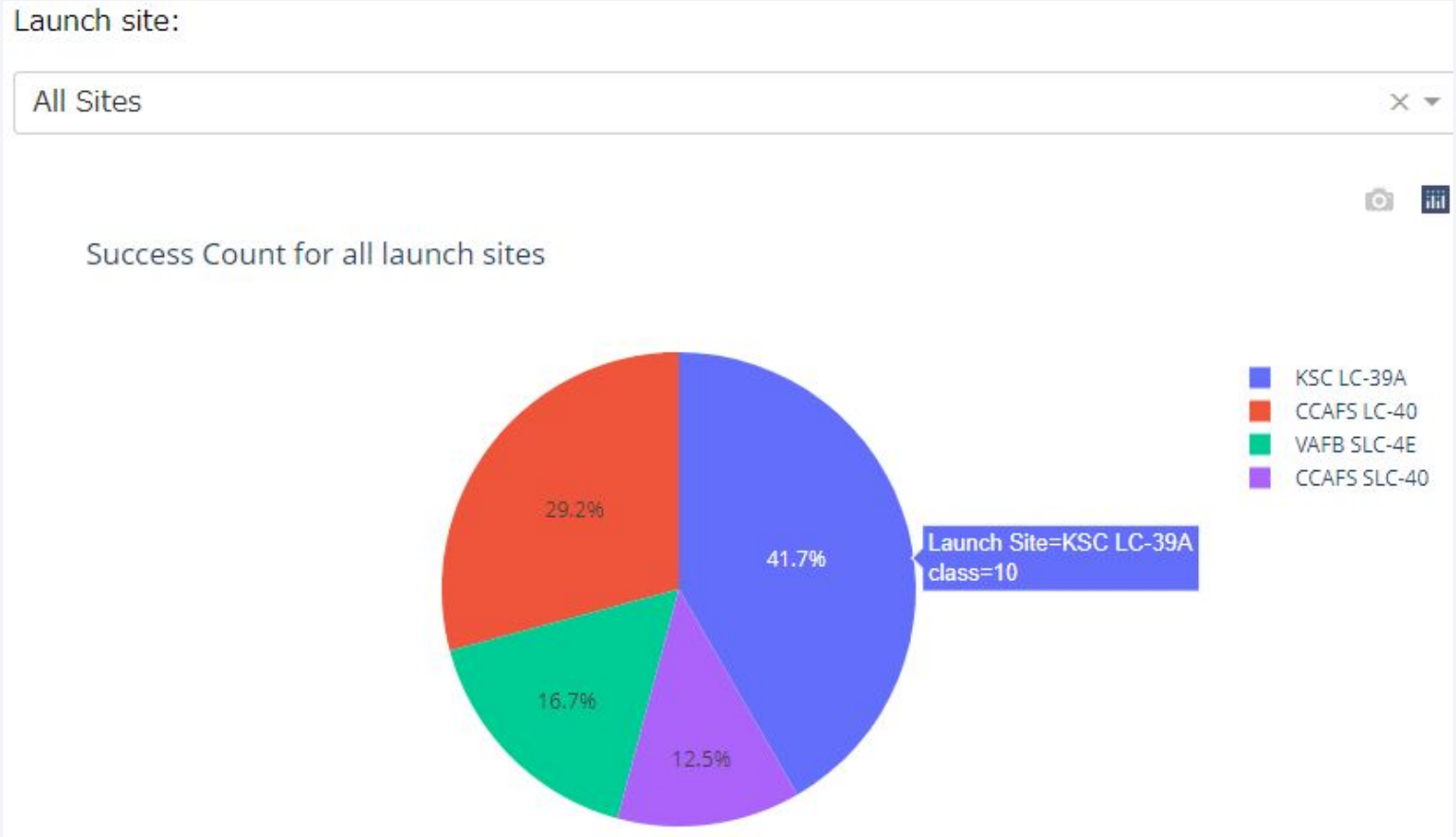


to nearest costline is around 0.88km east
and nearest railway is around 1.27km
from CCAFS SLC-40 site


Section 5
Build a Dashboard with Plotly Dash

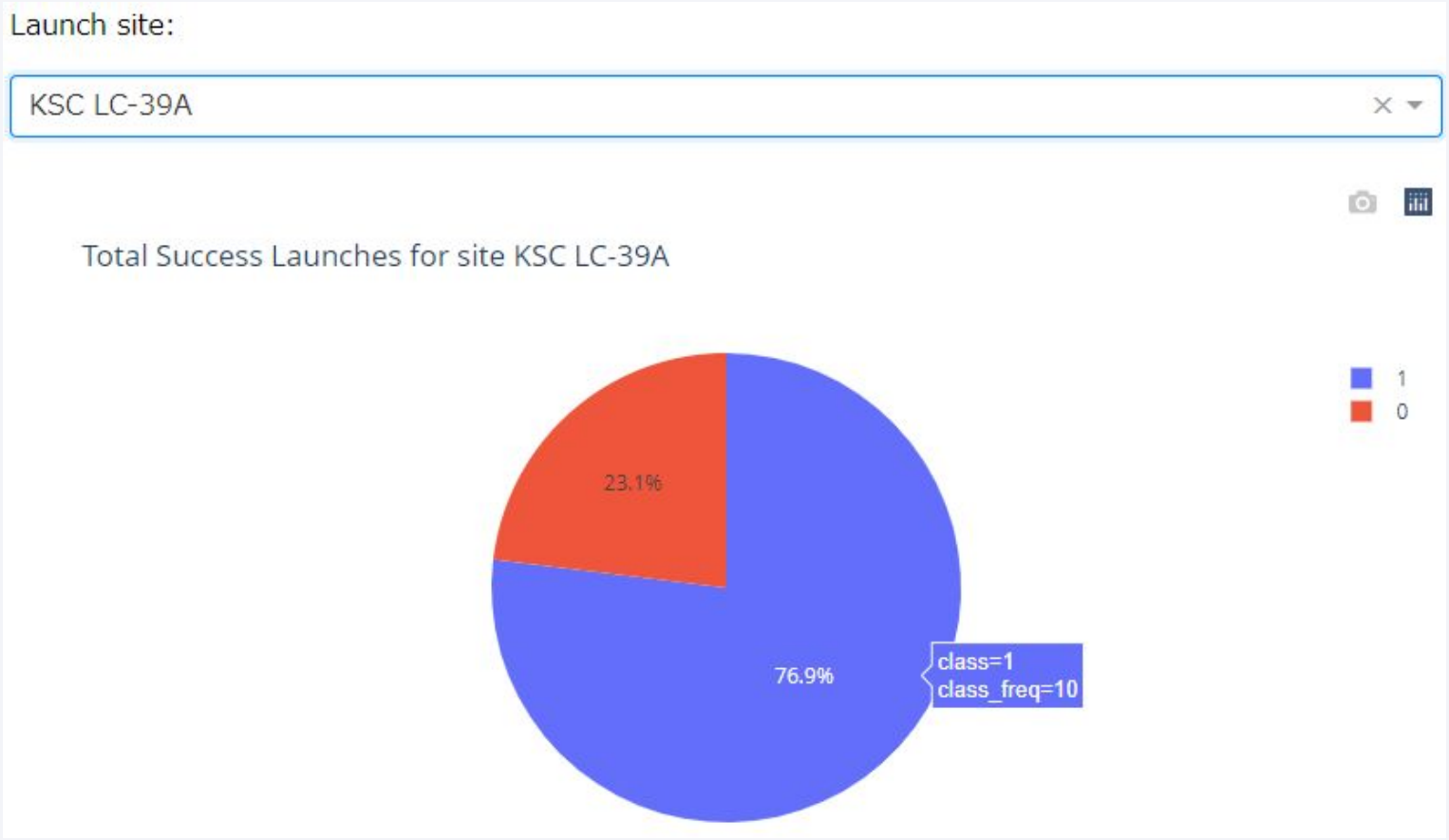
Success ratio of SpaceX Launch per site

Among four launch sites, KSC LC-39A located at Cape Canaveral Space Force Station (CCSFS), FL is the most number of successful launches (10 launches, 41.7% share) so far.



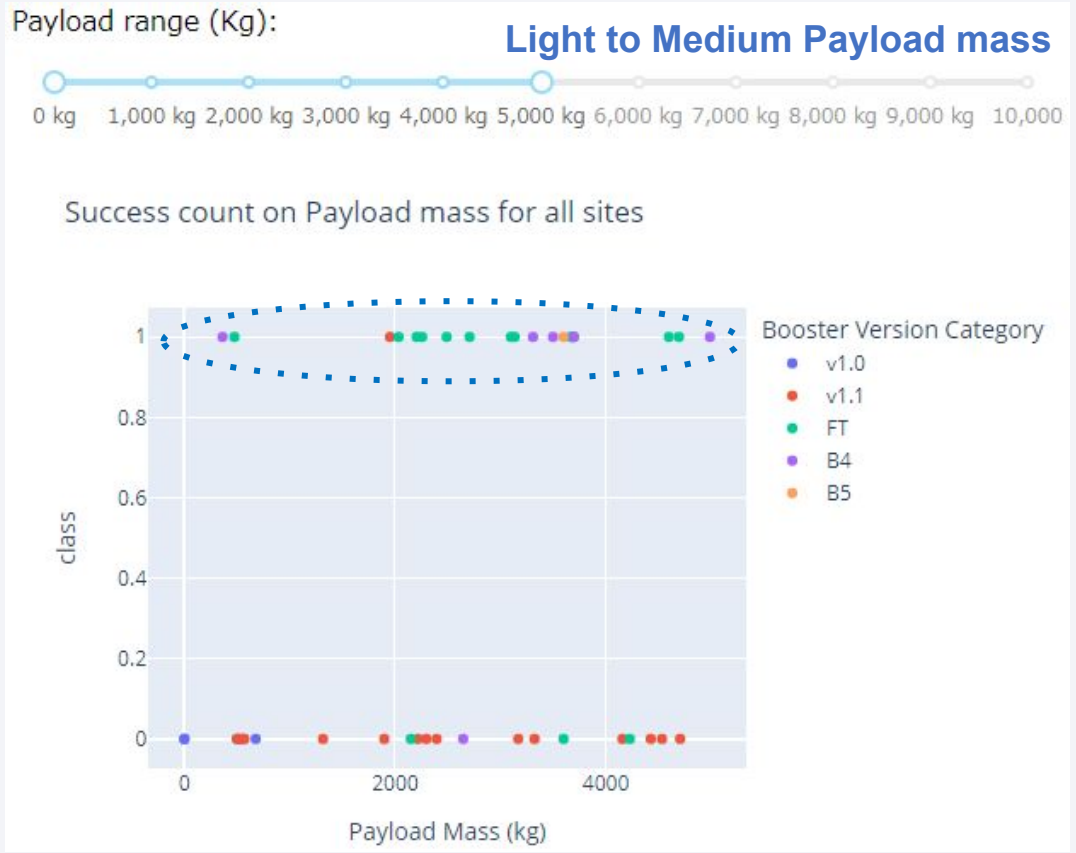
Highest success ratio at KSC LC-39A, CCFSF

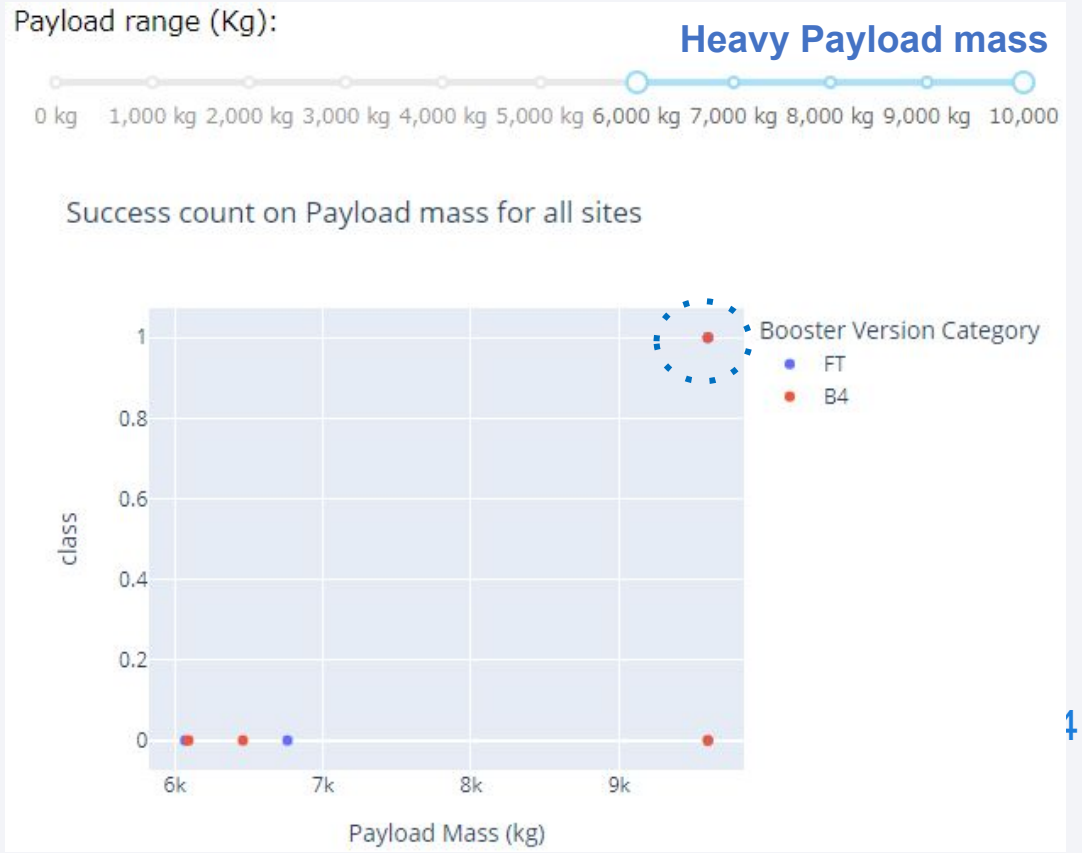
At KSC LC-39A, total of 13 launches, 10 were succeed and 3 were failed. Success ratio at the site is 76.9%, and this is the highest in four sites in terms of success ratio.



Payload mass, Booster dependency to success

- with heavy payload mass (>= 6,000kg), most of Booster version used were B4 and only 1 success so far.
- with light to medium payload mass (=< 5,000kg), Booster version FT and B4 showed higher success ratio than others



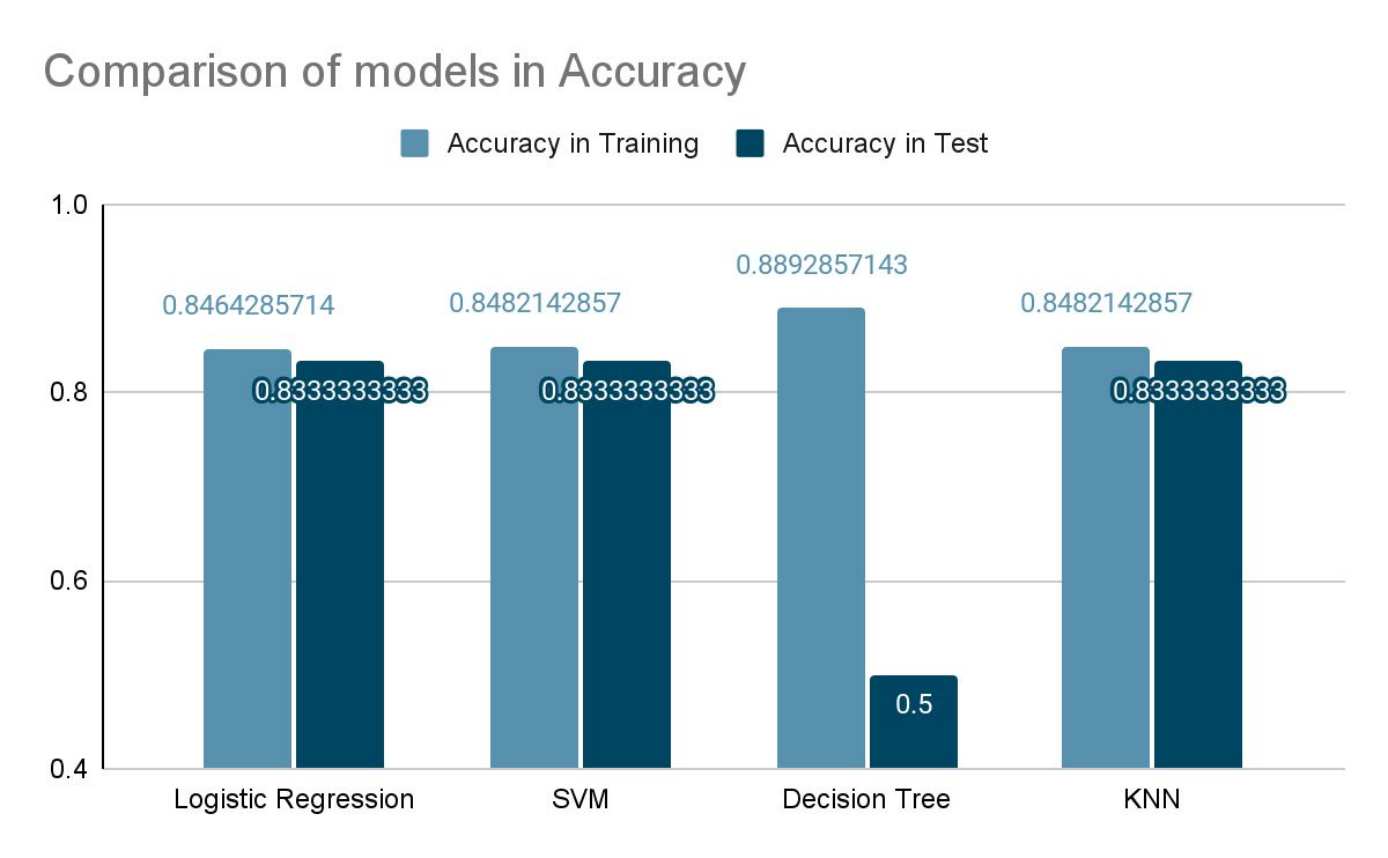


Section 6

Predictive Analysis (Classification)

Classification Accuracy

With Training dataset, Decision Tree showed the best performance, 0.8893.
With Test dataset, however, three algorithms showed the same best performance as around 0.8333 except Decision tree.

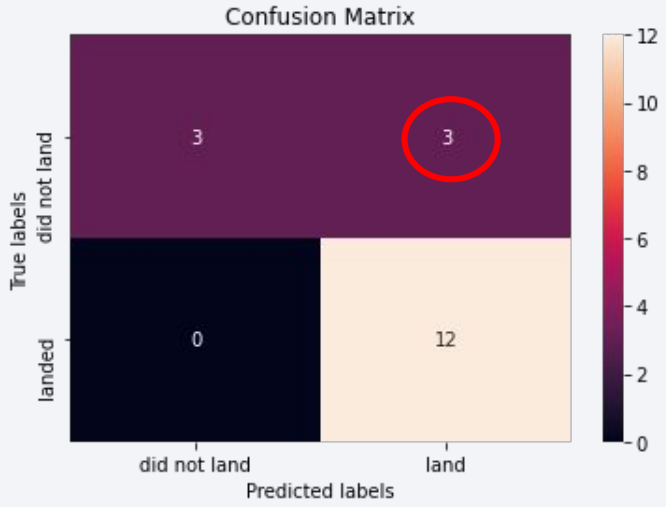


Confusion Matrix

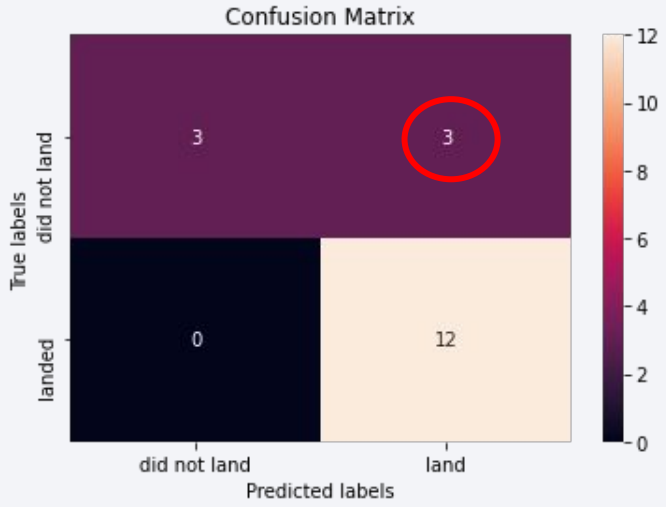
With Test dataset, three algorithms, Logistic regression, SVM, and KNN showed the same Accuracy as around 0.8333 as the best score.

According to Confusion matrix, all three algorithms missed Ground-truce ‘did not land’ as ‘land’ (False Negative) for three instances out of 18 cases.

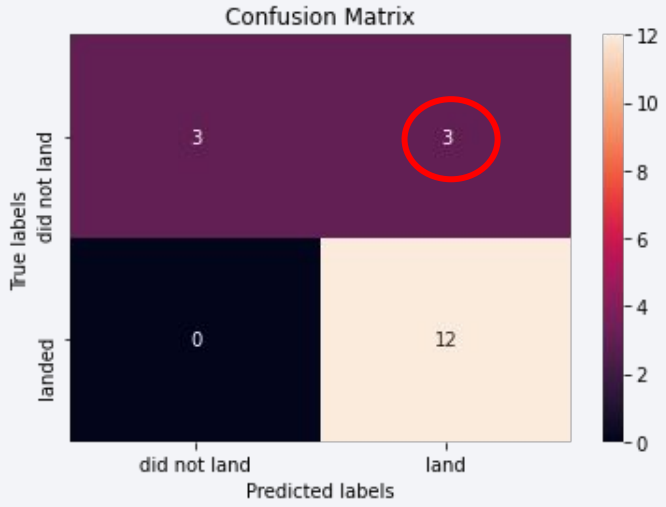
Logistic Regression



SVM



KNN



Conclusions

Key findings from series of data analysis are;

- Success rate of launch has been improved over time since 2013 as turning point
- Orbit ES-L1, GEO, HEO, SSO, and VLEO are higher in success.
  Demand of VLEO(Very Low Earth Orbit) has been increased recently. Higher success ratio of it would be appreciated by customers.
- Light and medium payload mass cases(=< 5,000kg) perform better than the heavier ones(>= 6,000kg). Heavier cases seems to be difficult
- Among four launch sites, KSC LC-39A located at Cape Canaveral Space Force Station (CCSFS), FL is the most successful site (10 launches, 41.7% share)
- Three classification algorithms, Logistic regression, SVM, and KNN, showed the best accuracy as 0.8333 to predict success/failed launches

Thank you!

Appendix