IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Thaidan Hoang
06/10/22

Outline

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

Executive Summary

- **Summary of methodologies**
  - Data collection
  - Data wrangling
  - EDA with data visualization
  - EDA with SQL
  - Building an interactive map with Folium
  - Building a Dashboard with Plotly Dash
  - Predictive analysis
- **Summary of all results**
  - EDA results
  - Interactive analytics
  - Predictive analysis

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.**

  The main research goal is to predict if the Falcon 9 first stage will land successful.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
    - Data Collection API
    - Data Collection with Web Scraping
- Perform data wrangling
    - Formating and determine Training Labels
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Logistic regression, SVM, Classification Trees and KNN.

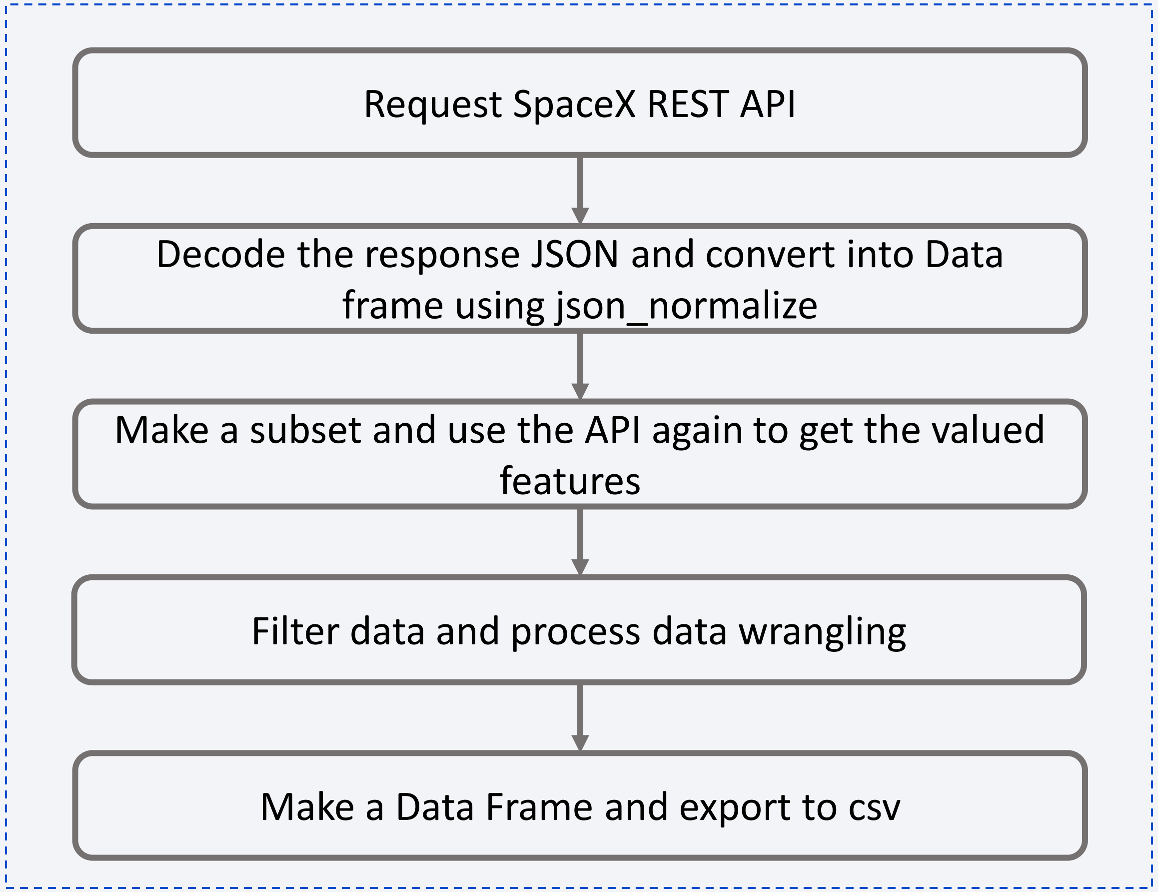
Data Collection

- **Data Collection API.**
  - Make a get request to the SpaceX API.
  - Clean the requested data.
- **Data Collection with Web Scraping.**
  - Extract a Falcon 9 launch records HTML table from Wikipedia.
  - Parse the table and convert it into a Pandas data frame.

Data Collection – SpaceX API

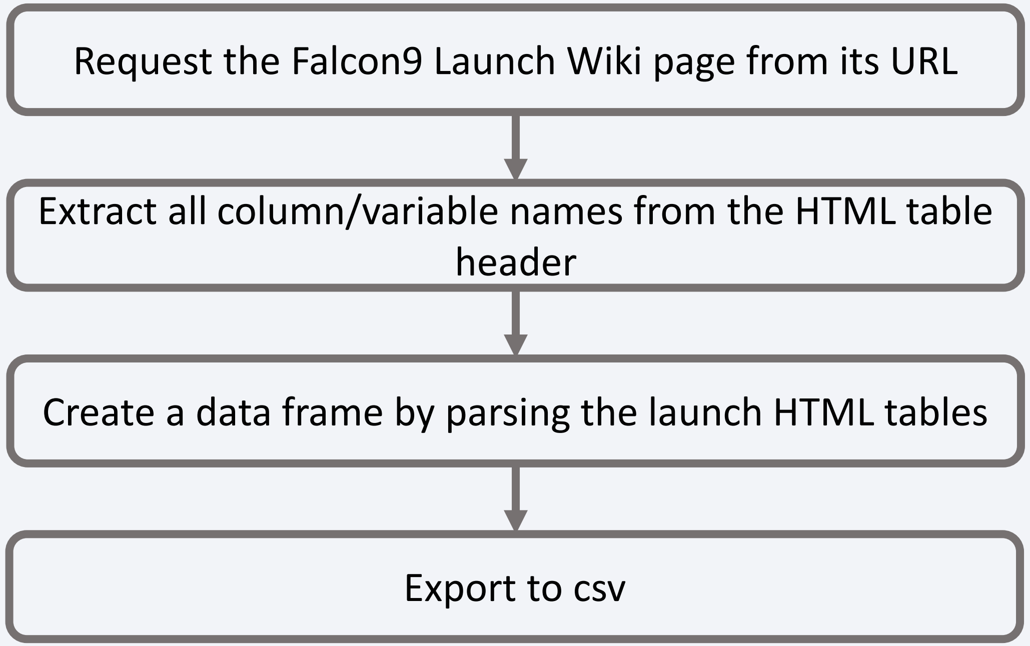
- **Flowcharts of data collection using SpaceX REST API**

- **GitHub Url:** https://github.com/atndan/Applied-Data-Science-Capstone/blob/main/data-collection-API.ipynb



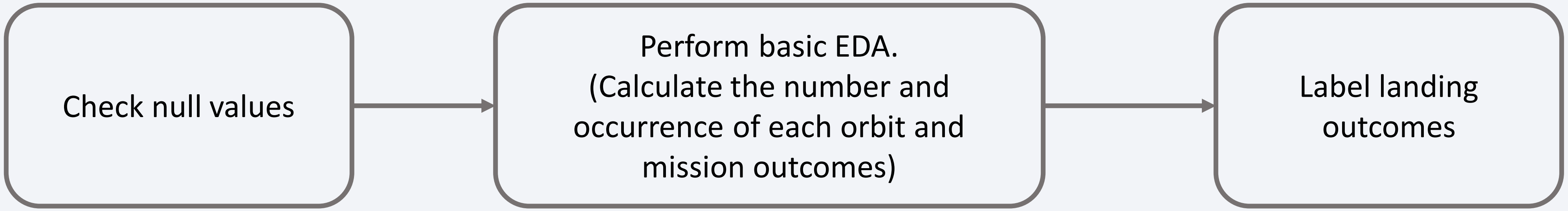
Data Collection - Scraping

- **Flowcharts of data collection with webscraping**
- **GitHub Url:** https://github.com/atndan/Applied-Data-Science-Capstone/blob/main/data-collection-webscraping.ipynb

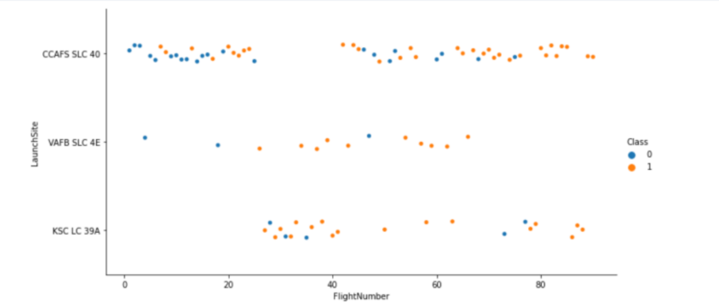


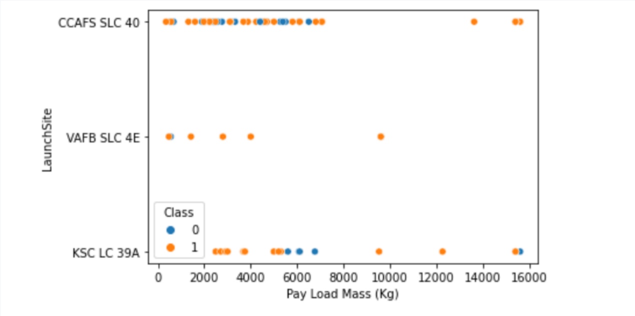
Data Wrangling

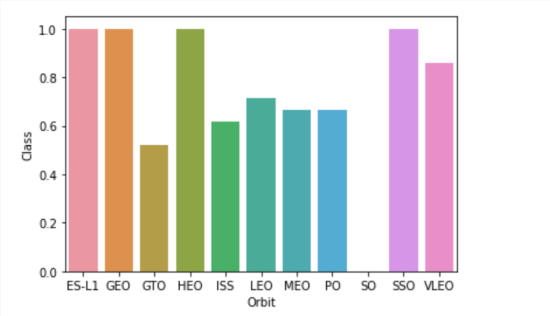
- Perform basic Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.
- GitHub Url: https://github.com/atndan/Applied-Data-Science-Capstone/blob/main/data-wrangling.ipynb

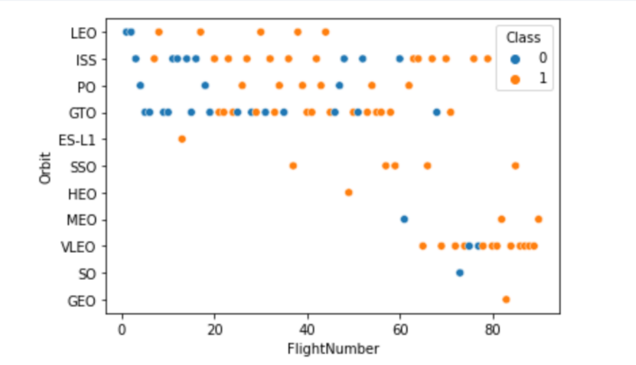


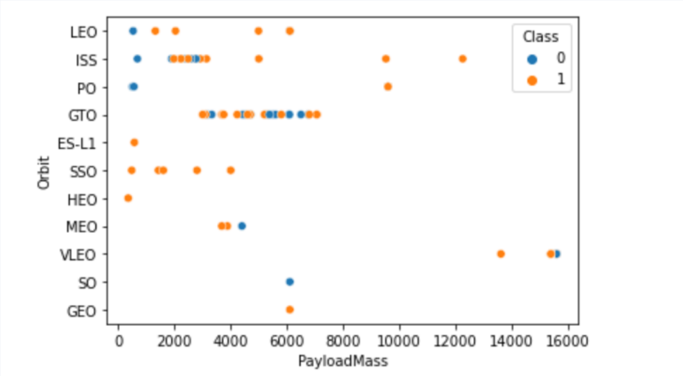
EDA with Data Visualization

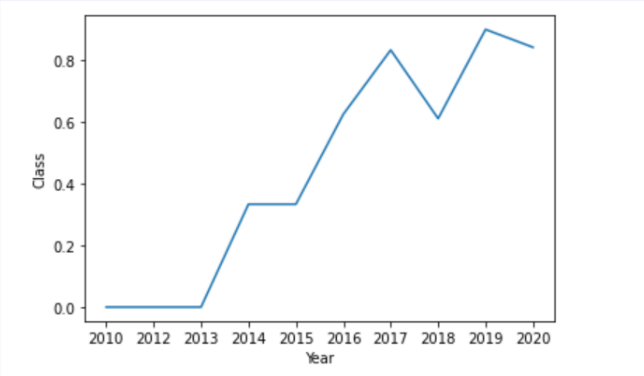












- We use scatter plot for one numeric attribute against another numeric attribute.
  And bar plot for one numeric attribute against category

- Github Url: https://github.com/atndan/Applied-Data-Science-Capstone/blob/main/eda-data-visualize.ipynb

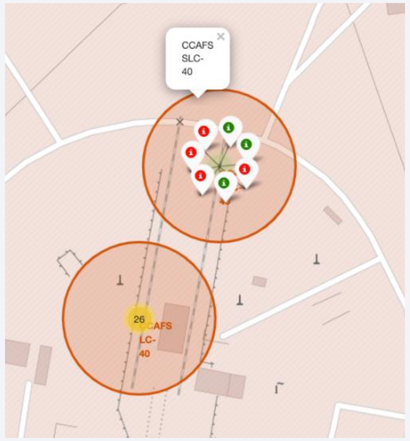
EDA with SQL

- Display the names of the unique launch sites in the space mission
- Display some records of launch sites to check how data look like
- Calculate total and average payload mass of some launch sites and booster version.
- List some interesting records, such as date of the first successful landing outcome in ground pad was archived, or the names of the boosters which have success in drone ship.
- List the total number of successful and failure mission outcomes
- Rank the count of landing outcomes

GitHub Url: https://github.com/atndan/Applied-Data-Science-Capstone/blob/main/eda-sql.ipynb

Build an Interactive Map with Folium

Mark all launch sites on a map



Mark the success/failed

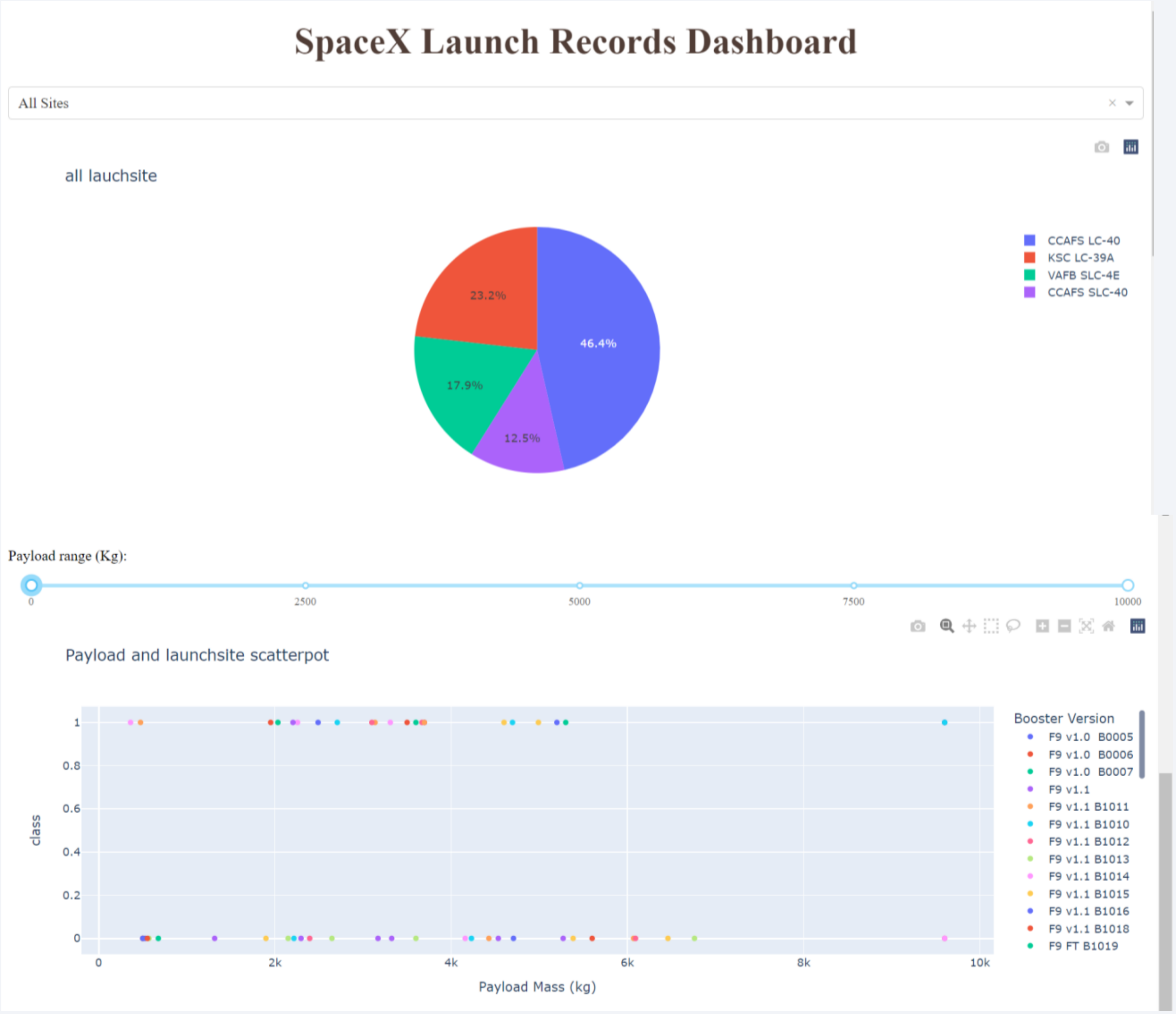
Add MousePosition, PolyLine to calculate some distances

- GitHub Url: https://github.com/atndan/Applied-Data-Science-Capstone/blob/main/map-analysis-Folium.ipynb

Build a Dashboard with Plotly Dash

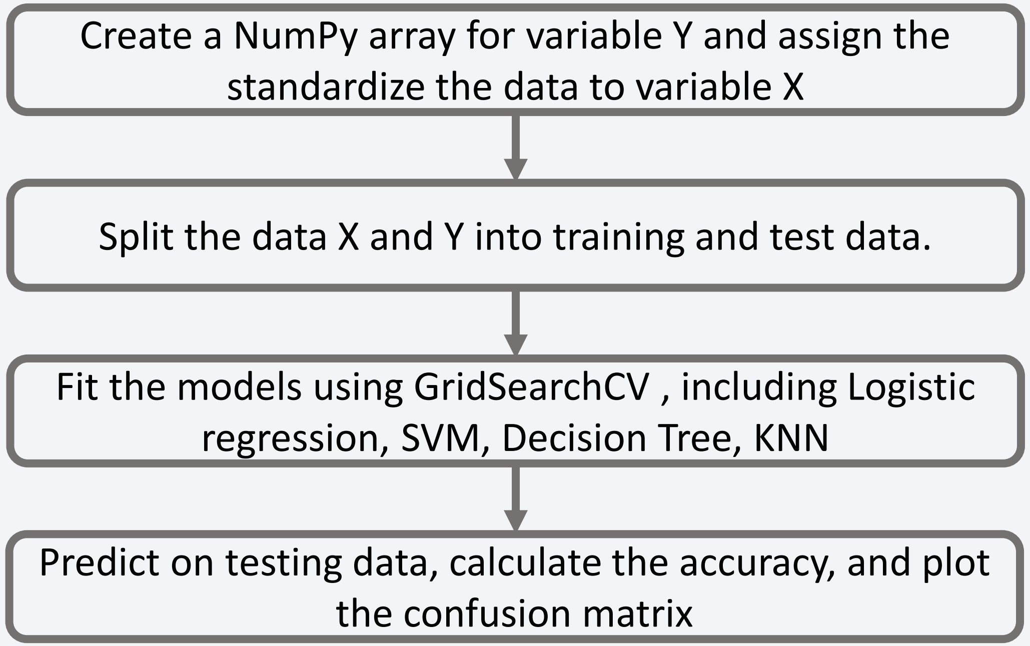
- Add a dropdown list to enable Launch Site selection
- Add a pie chart to show the total successful launches
- Add a scatter chart to show the correlation between payload and launch success

https://github.com/atndan/Applied-Data-Science-Capstone/blob/main/build-interactive-Dashboard-Ploty-Dash.ipynb



Predictive Analysis (Classification)

- Flowchart of Machine Learning Prediction
- GitHub URL: https://github.com/atndan/Applied-Data-Science-Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



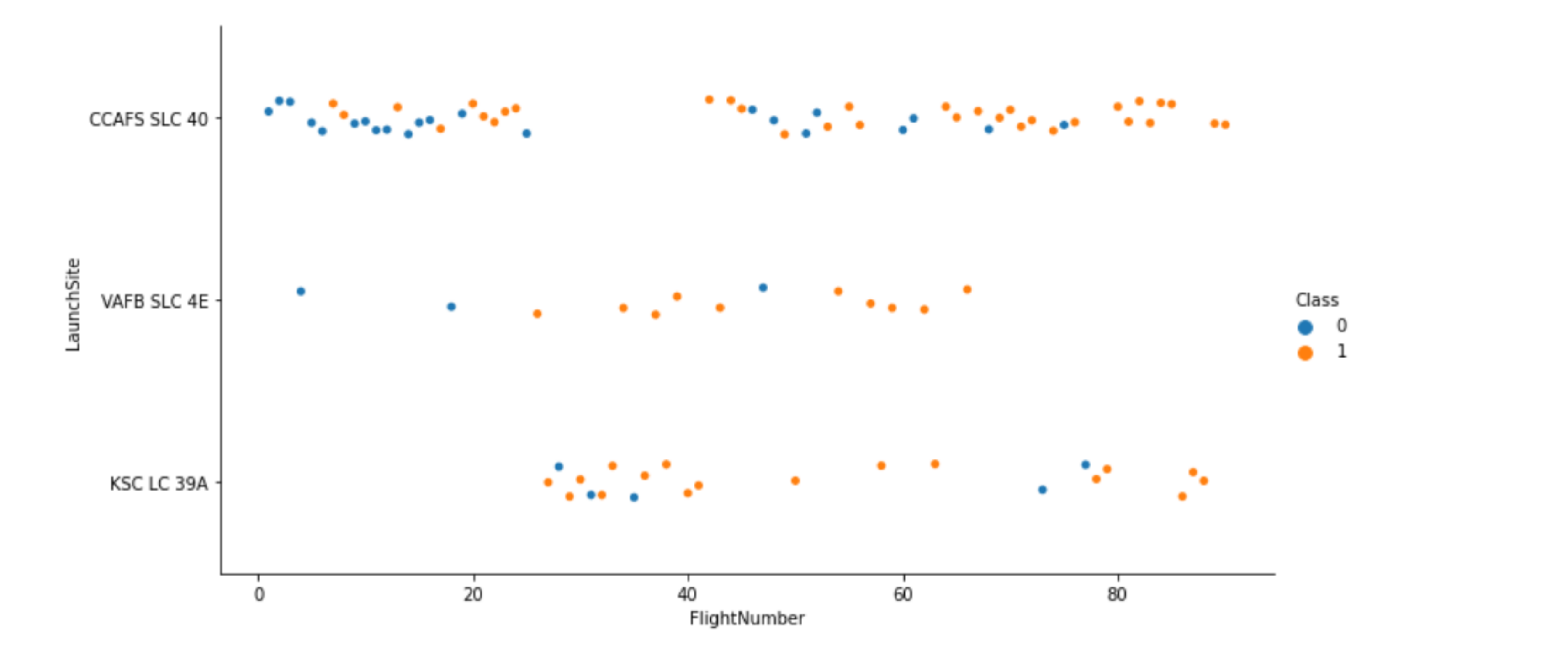
Results

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

Section 2

Insights drawn from EDA

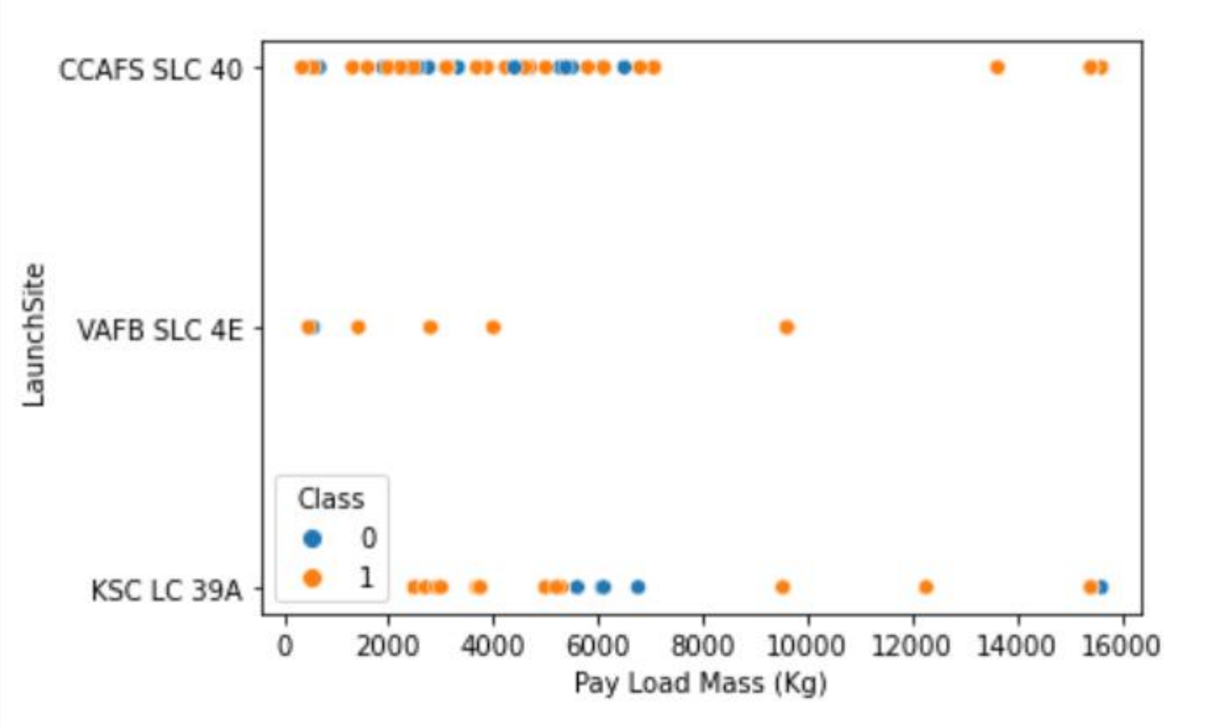
Flight Number vs. Launch Site



From the scatter plot of Flight Number vs. Launch Site, we can observe the launchsite CCAFS SLC 40 is the most used, and with the bigger flight number, the higher success rate is.

Payload vs. Launch Site

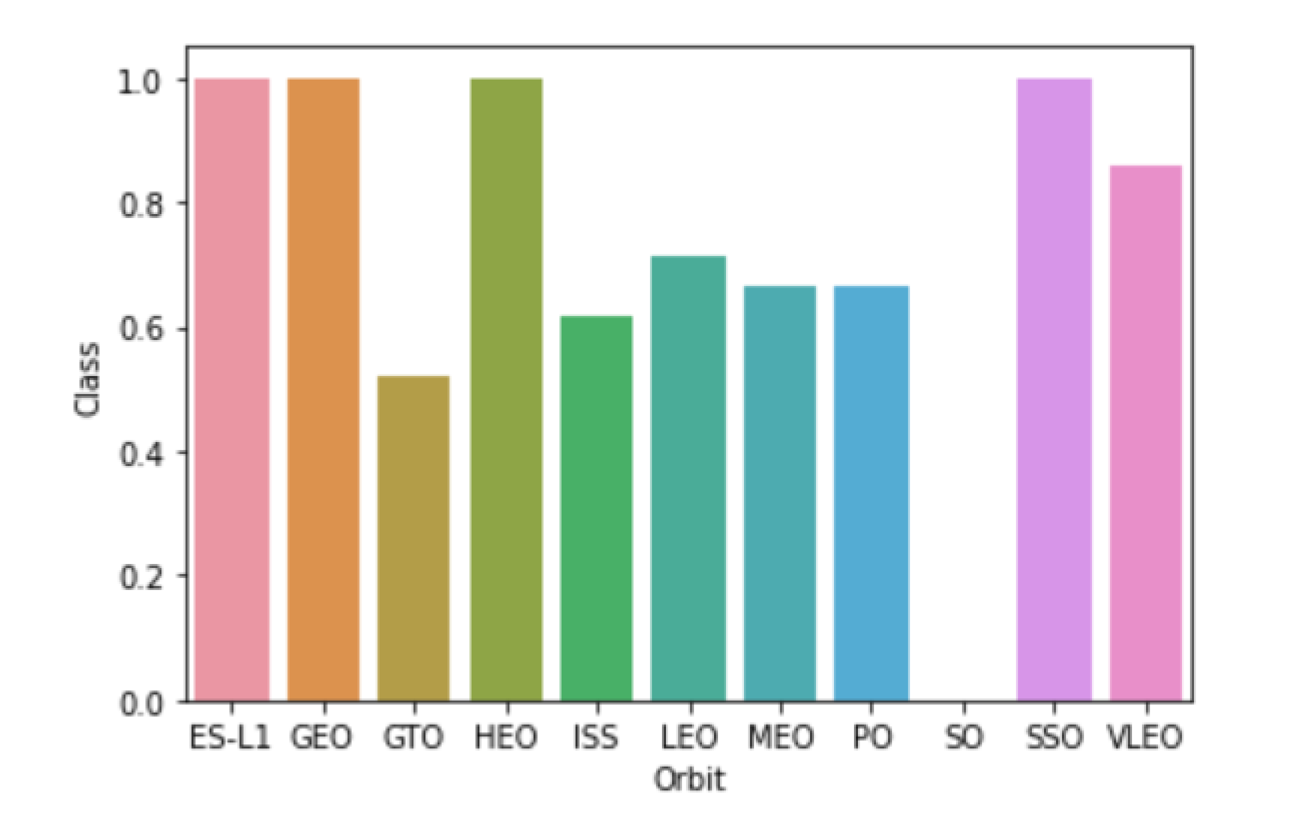
- We can observe the VAFB-SLC launch site, there are no rockets launched for heavy payload mass (greater than 10000).
- CCAFS SLC 40 is most used with low payload mass



Scatter plot of Payload vs. Launch Site

Success Rate vs. Orbit Type

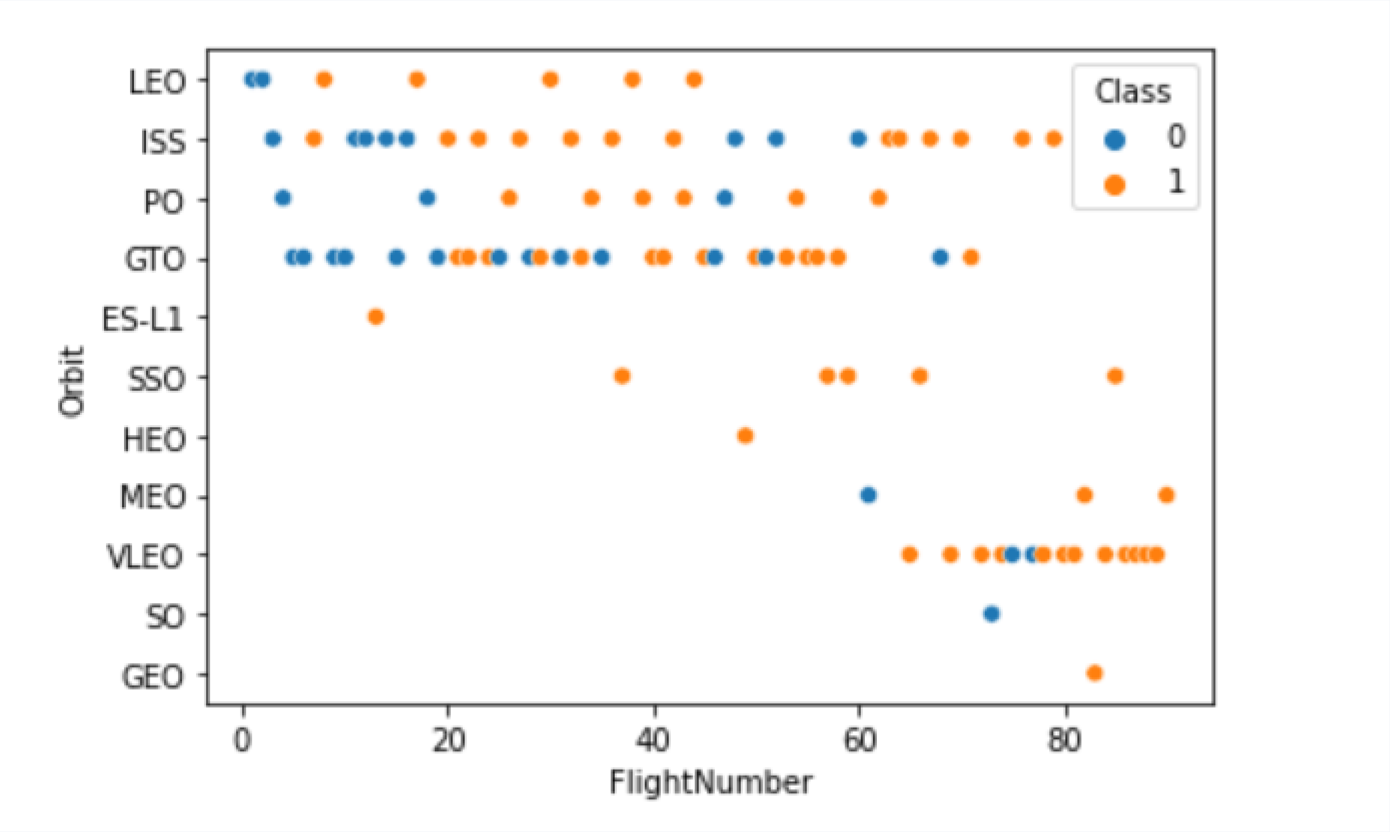
- The Orbit ES-L1, GEO, HEO, SSO have the highest success rate
- The orbit GTO has the lowest success rate



Flight Number vs. Orbit Type

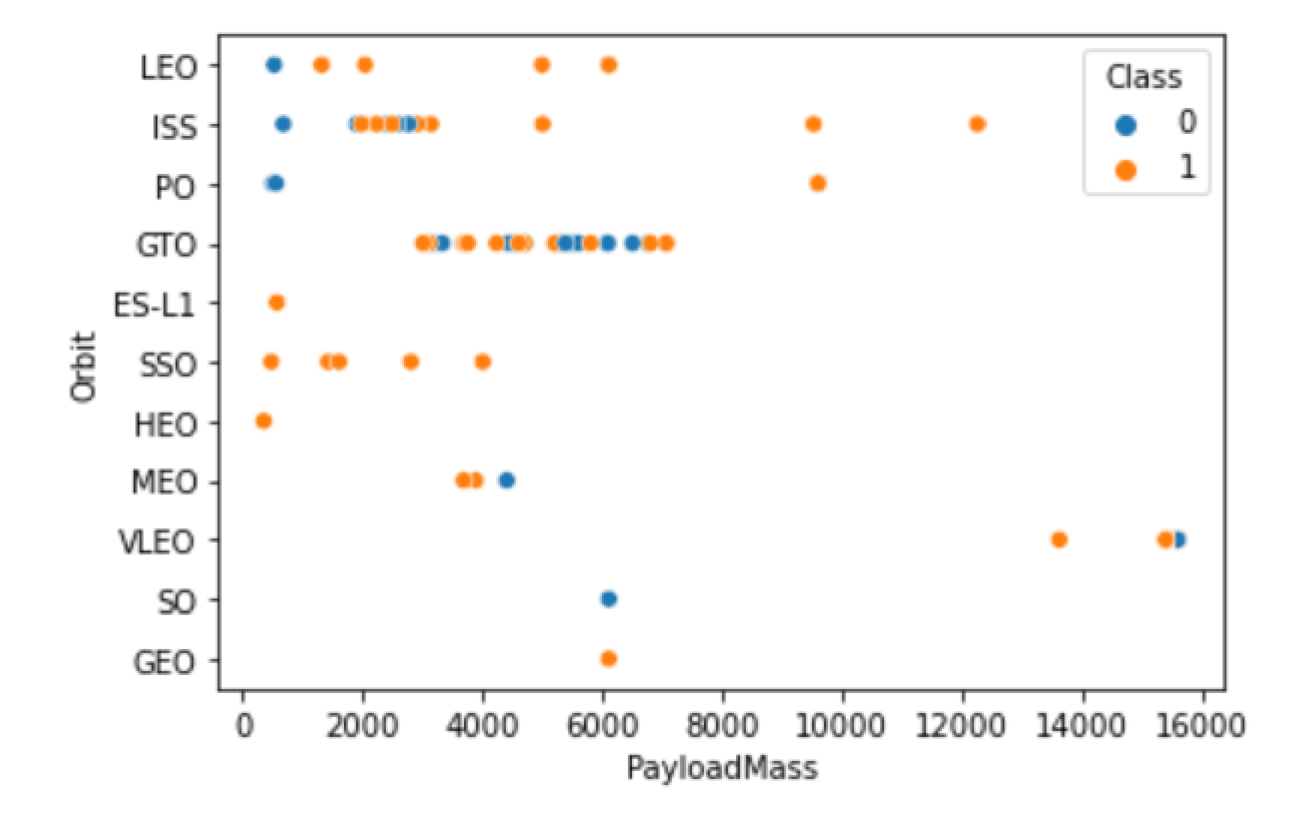
- The LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

- VLEO, MEO, SO, GEO are launched in recent year.



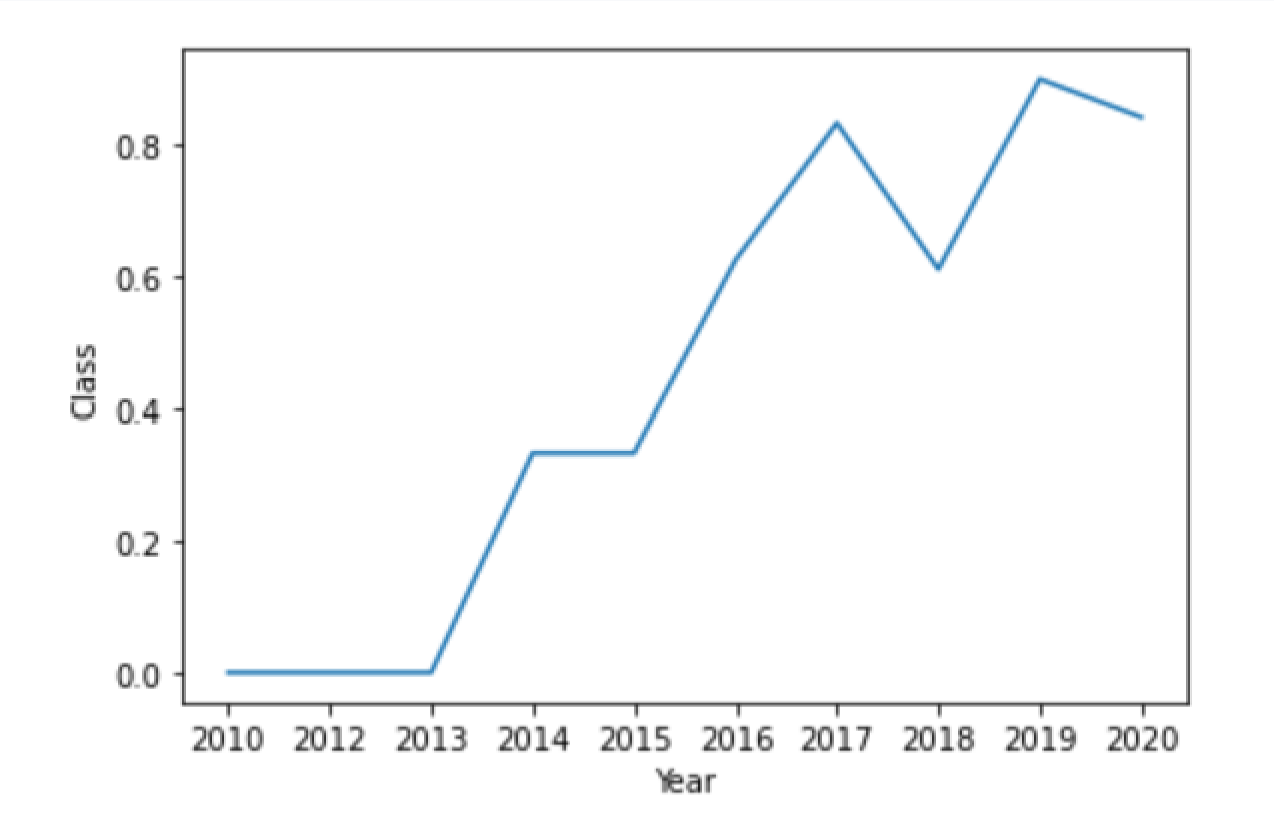
Payload vs. Orbit Type

- 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

- A clear trend can be observed that the success rate since 2013 kept increasing till 2020



All Launch Site Names

- query display the names of the unique launch sites in the space mission

```
%sql select DISTINCT LAUNCH_SITE from YKX98701.SPACEXTBL
```

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`

In [58]:
```
a2 = %sql select * from YKX98701.A where LAUNCH_SITE like 'CCA%' LIMIT 5
a2
```

```
 * ibm_db_sa://ykx98701:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[58]:

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

- Total payload carried by boosters from NASA

```
In [66]: %sql select sum(payload_mass__kg_) as sum from YKX98701.SPACEXTBL where customer like 'NASA (CRS)'

 * ibm_db_sa://ykx98701:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.

Out[66]:
```

SUM
45596

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.

```
In [46]: %sql select avg(payload_mass__kg_) as average from YKX98701.SPACEXTBL where booster_version like 'F9 v1.1'

 * ibm_db_sa://ykx98701:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[46]:

average
2928

First Successful Ground Landing Date

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

```
In [51]: %sql select min(DATE) as first_landing from YKX98701.A where landing__outcome like 'Success (ground pad)'

 * ibm_db_sa://ykx98701:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[51]:

first_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
In [52]: %sql select booster_version from YKX98701.A where landing__outcome like 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000

 * ibm_db_sa://ykx98701:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[52]:

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

In [54]:
```
%%sql select mission_outcome, count(*) as total
from YKX98701.A
group by  mission_outcome
having mission_outcome like 'Success%' or mission_outcome like 'Failure%'
```

```
 * ibm_db_sa://ykx98701:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[54]:

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

```
In [56]: %%sql select booster_version from YKX98701.A
         where payload_mass__kg_ =  (select max(payload_mass__kg_) from YKX98701.A)

 * ibm_db_sa://ykx98701:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[56]:

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

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

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

In [61]:

```
%%sql select landing__outcome, booster_version, launch_site
from YKX98701.A
where landing__outcome like 'Failure (drone ship)'
and DATE like '%2015%'
```

```
 * ibm_db_sa://ykx98701:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[61]:

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

In [65]:
```
%%sql
select  landing__outcome, count(*) as total

from(select landing__outcome, DATE
from YKX98701.A
where DATE between '2010-06-04' and '2017-03-20')

group by landing__outcome
order by total DESC
```

Out[65]:

landing__outcome	total
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

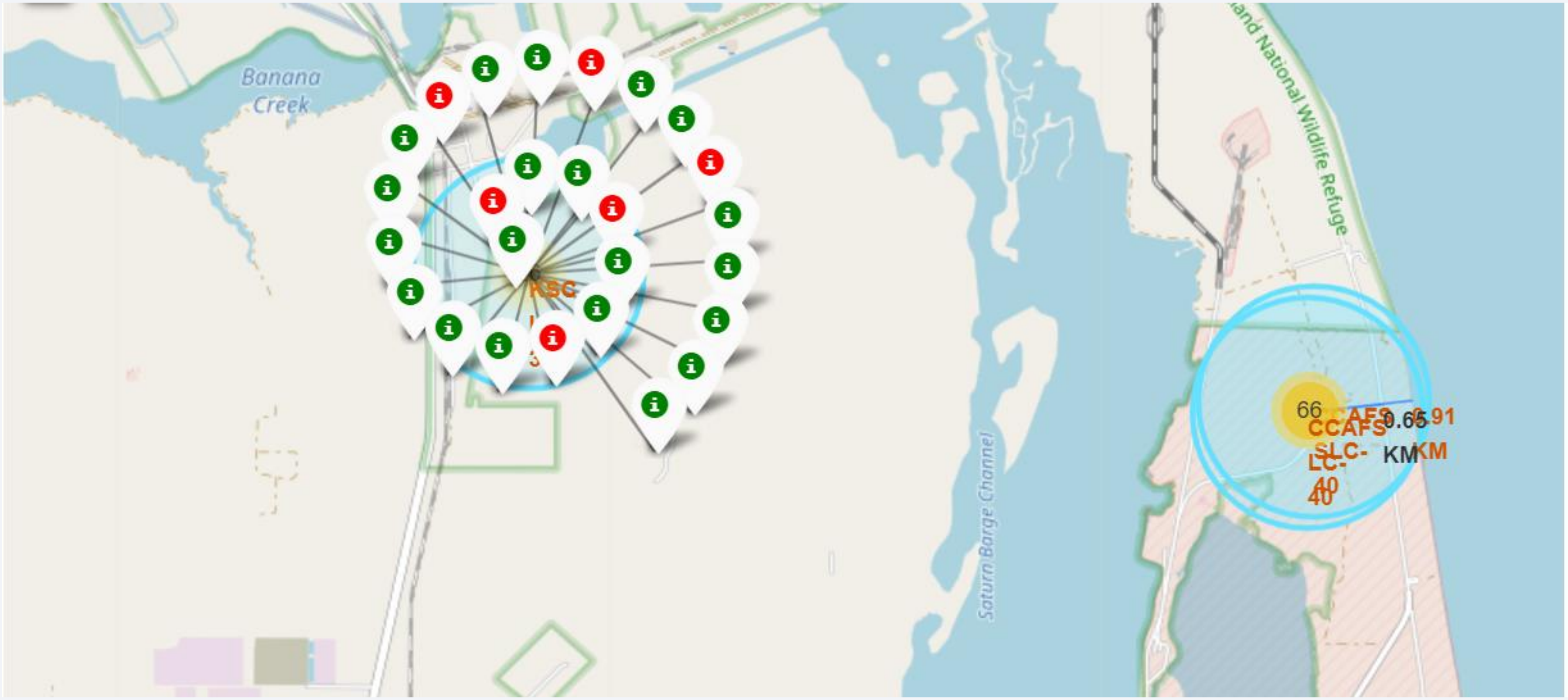
Section 3

Launch Sites Proximities Analysis

All launch sites are marked on the map

- All launch sites in proximity to the Equator line. Since rockets launched near the Equator get an additional natural boost that helps save the cost of putting in extra fuel and boosters.
- All launch sites in very close proximity to the coast. This could be probably not to damage the civil area.

Mark the success/failed launches



- KSC LC-39A can be observed to have a relatively high success rates.

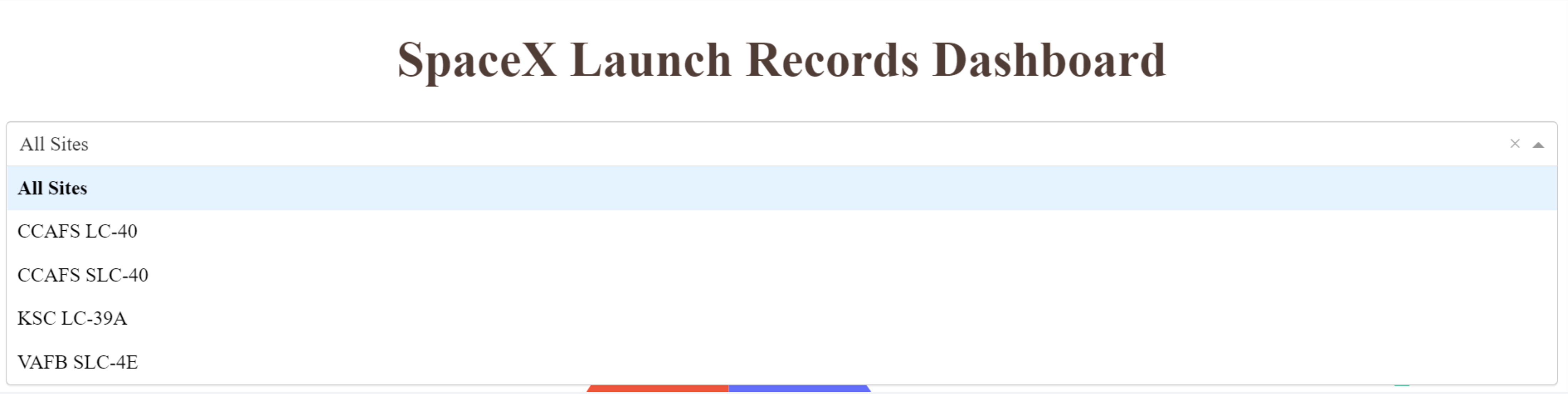
Distances between a launch site to its proximities

- Launch sites are in close proximity to coastline, railways, highway.
- Launch sites are kept a certain distance away from city.

Section 4

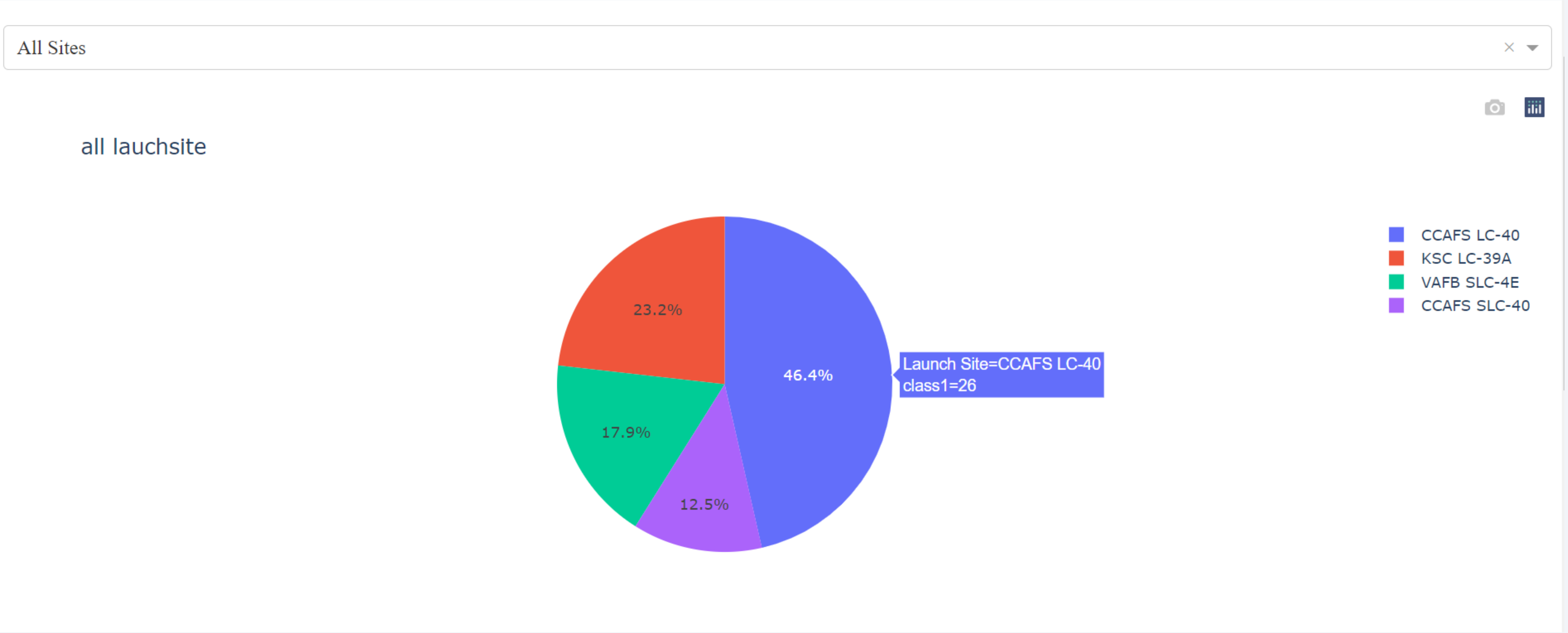
Build a Dashboard with Plotly Dash

Dropdown list enable Launch Site selection



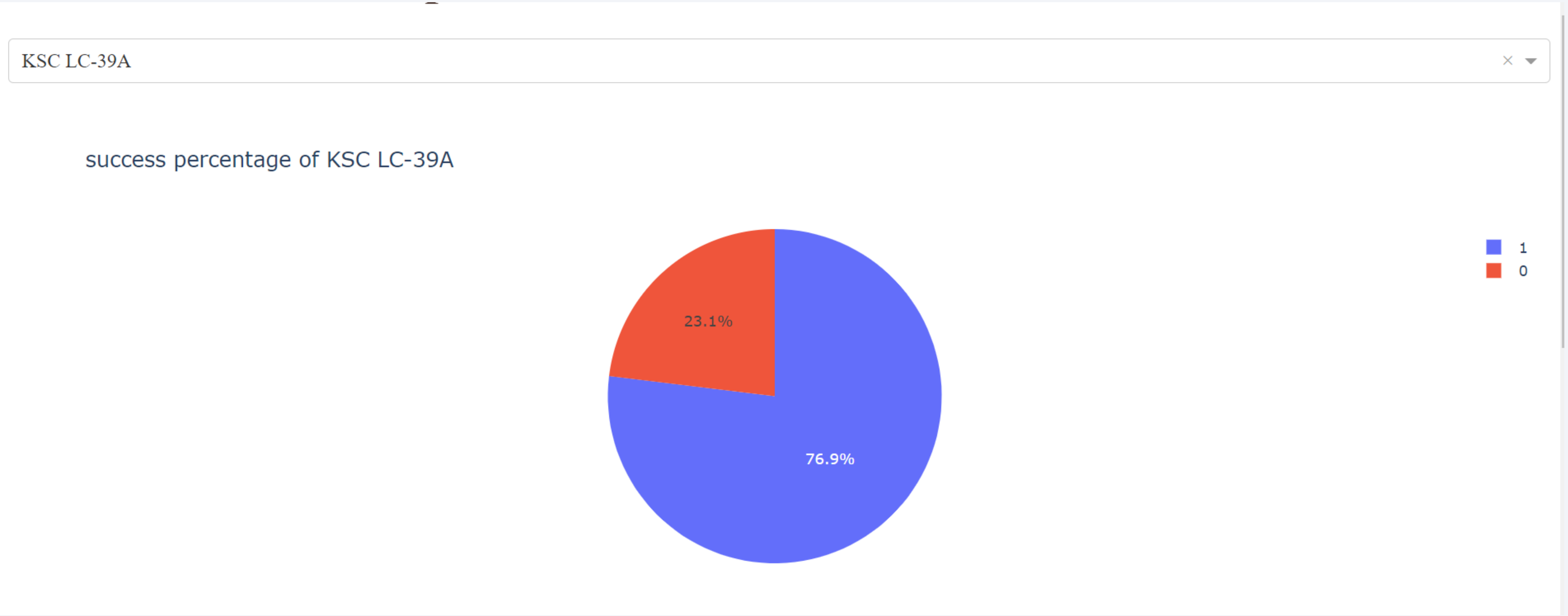
- Add a dropdown list to enable Launch Site selection, The default select value is for ALL sites

Pie chart of the total successful launches count



- Launch site CCAFS LC-40 is the most used launch site.

Pie chart of the total successful launches count



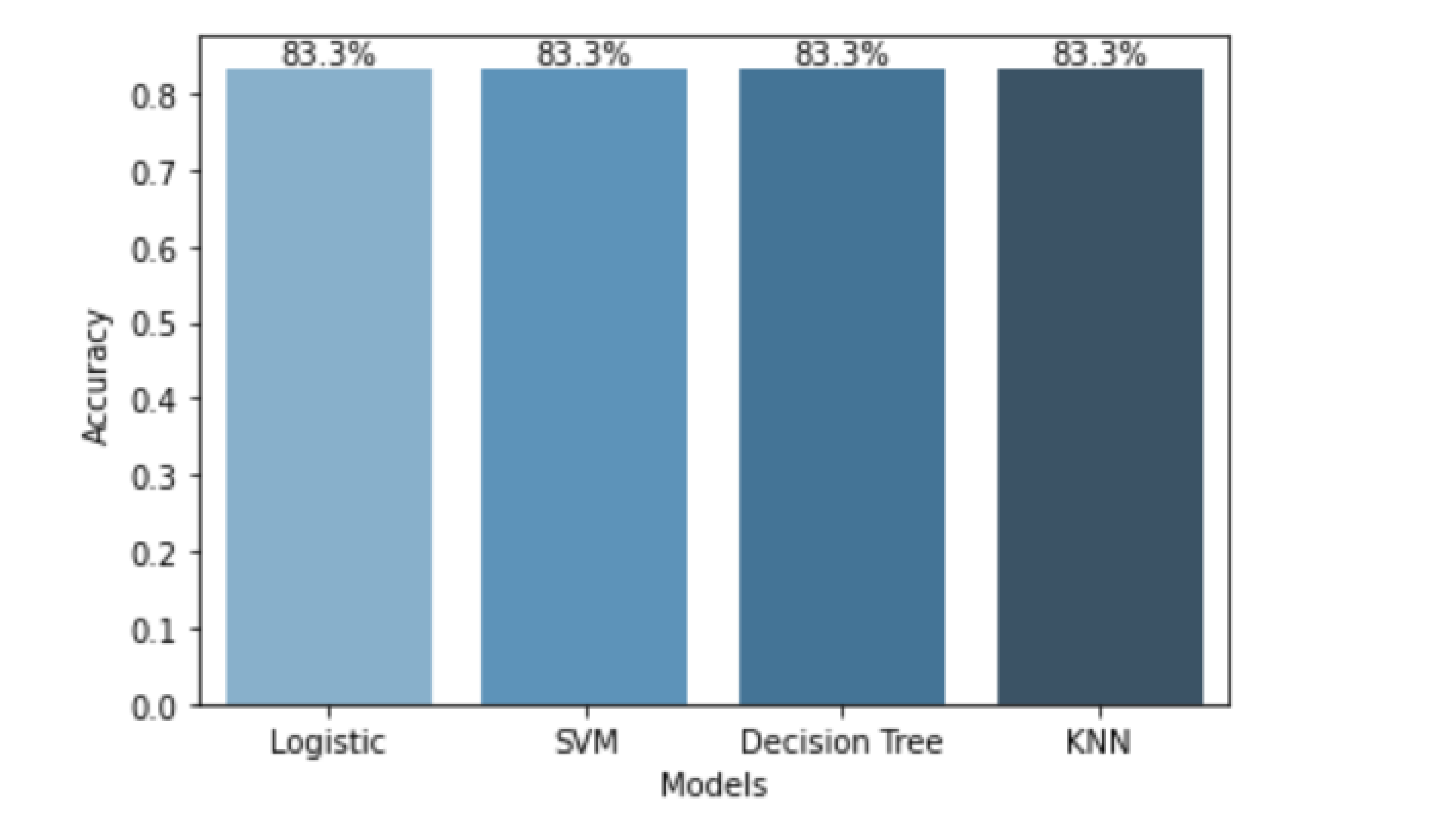
- Launch site KSC- LC-39A is has the highest success rate.

Payload vs launchsite scatter plot

Section 5

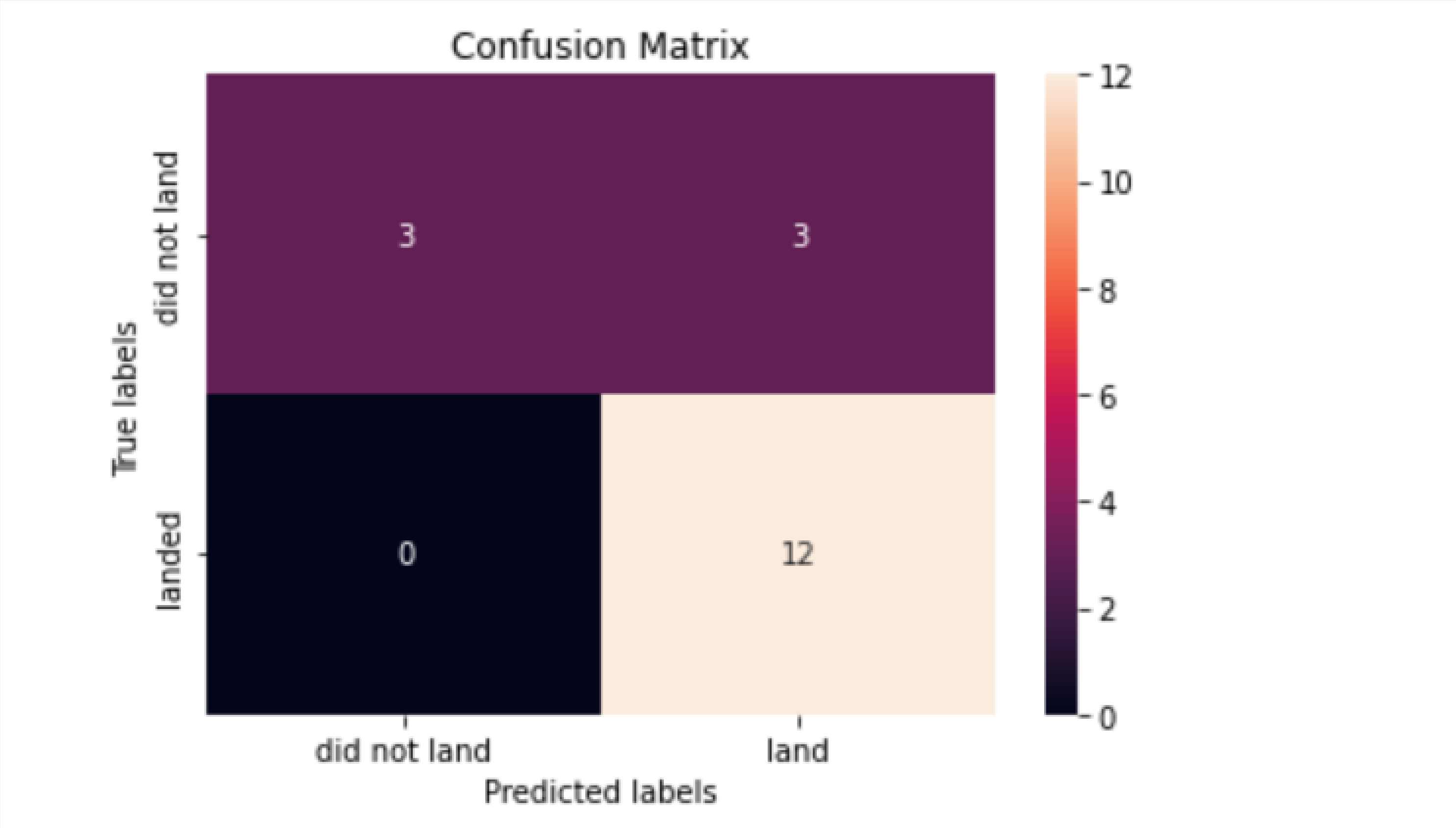
Predictive Analysis (Classification)

Classification Accuracy



- All models has equal accuracy, due to small training and testing data set

Confusion Matrix



- All models result in the same confusion matrix, and the major problem can be observed is the false positives.

Conclusions

- Flight Number, Launch Site, Payload Mass, Orbit Type, Launch Year, Booster Version are significant variable impact the landing outcome.
- A clear trend can be observed that the success rate since 2013 kept increasing.
- The Orbit ES-L1, GEO, HEO, SSO have the highest success rate.
- All launch sites in proximity to the Equator line. Launch sites are in close proximity to coastline, railways, highway. Launch sites are kept a certain distance away from city.
- All model prediction has same accuracy of 83% which is relatively good, significant enough to be useful in predicting if the first stage will land given the data from the preceding labs.

Thank you!