Winning Space Race with Data Science

Jiaxin Peng
2.2.2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies:
  - Data collection: SpaceX REST API and web scraping via Wikipedia.
  - Data Analysis: SQL and Data Visualization
  - Data prediction: Folium interactive map, Dashboard and Machin learning.
- Summary of all results:

1. SSO, HEO, ES-L1 and SSO have the highest success rate.
2. KSC LC-39A has the highest success launch rate than other launch sites.
3. FT booster version has the highest counts.
4. The Decision tree model has the highest accuracy in prediction.

Introduction

- Project background and context

  According to public information, the launch cost of SpaceX is relatively inexpensive. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. We help the new company to prepare the information for racket launch.

- Problems:
  - To what extent SpaceX can reuse the first stage?
  - How to determine the cost of the launch and the first stage?
  - What is the success rate?
  - Are there any restrictions, like payload mass or launch site selection?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Via SpaceX API
  - Webscraping from SpaceX Wikipedia page
- Perform data wrangling
  - Group information according to landing sites, orbits and mission outcomes.
- 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

- Data are collection in two different ways.
- 1. Collect from SpaceX REST API directly.
  - Import and connect with REST API → collect from JSON file → create dataframe → only include Falcon 9 → missing value detect and adjustment
- 2. Collect data from Wikipedia via webscraping.
  - Connect Wiki via requests → extract Falcon 9 table by BeautifulSoup → create dataframe

Data Collection – SpaceX API

- Via SpaceX API, I connect, collect, create and clean.


```
spacex_url="https://api.spacexdata.com/v4/launches/past"
response = requests.get(spacex_url)
print(response.content)
```

```
launch_dict = {'FlightNumber': list(data['flight_number']),
'Date': list(data['date']),
'BoosterVersion':BoosterVersion,
'PayloadMass':PayloadMass,
'Orbit':Orbit,
'LaunchSite':LaunchSite,
'Outcome':Outcome,
'Flights':Flights,
'GridFins':GridFins,
'Reused':Reused,
'Legs':Legs,
'LandingPad':LandingPad,
'Block':Block,
'ReusedCount':ReusedCount,
'Serial':Serial,
'Longitude': Longitude,
'Latitude': Latitude}
```

```
data_falcon9 = launch_df[launch_df['BoosterVersion']=='Falcon 9']
```

```
PayloadMass_mean = data_falcon9['PayloadMass'].mean()
# Replace the np.nan values with its mean value
data_falcon9['PayloadMass'].replace(np.nan, PayloadMass_mean, inplace=True)
data_falcon9.isnull().sum()
```

```
data_falcon9.to_csv('dataset_part_1.csv', index=False)
```

- GitHub URL

Data Collection - Scraping

- Scraping via Wikipedia SpaceX website

Request URL and collect information → Filter the Falcon9 table and collect column name → Create DataFrame

```
static_url =
"https://en.wikipedia.org/w/index.php?
title=List_of_Falcon_9_and_Falcon_Heavy_launches
&oldid=1027686922"
response = requests.get(static_url)
soup = BeautifulSoup(response.text,
'html.parser')
```

```
html_tables = soup.find_all('table')
first_launch_table = html_tables[2]
print(first_launch_table)
column_names = []
for th in first_launch_table.find_all('th'):
    name = extract_column_from_header(th)
    if name is not None and len(name) > 0:
        column_names.append(name)
```

```
launch_dict= dict.fromkeys(column_names)
# Remove an irrelvant column
del launch_dict['Date and time ( )']
launch_dict['Flight No.'] = []
launch_dict['Launch site'] = []
launch_dict['Payload'] = []
launch_dict['Payload mass'] = []
launch_dict['Orbit'] = []
launch_dict['Customer'] = []
launch_dict['Launch outcome'] = []
launch_dict['Version Booster']=[]
launch_dict['Booster landing']=[]
launch_dict['Date']=[]
launch_dict['Time']=[]
df=pd.DataFrame.from_dict(launch_dict,orient='in
dex')
df=df.transpose()
```

- GitHub URL

Data Wrangling

- The main purpose of this step is to create a column named landing class to distinguish the successful and failed cases with 1 and 0.
- There are several steps:

Calculate number of launches on each site

Calculate the number of occurrence of each orbit

Calculate counts of mission outcome per orbit type

Create landing outcome lable

Check the performance

```
# Apply value_counts() on column LaunchSite
df['LaunchSite'].value_counts()
```

```
# Apply value_counts on Orbit column
df['Orbit'].value_counts()
```

```
# landing_outcomes = values on Outcome column
landing_outcomes = df['Outcome'].value_counts()
for i,outcome in enumerate(landing_outcomes.keys()):
    print(i,outcome)
```

```
# landing_class = 0 if bad_outcome
# landing_class = 1 otherwise
landing_class = df['Outcome'].apply(lam
x: 0 if x in bad_outcomes else 1)
df['Class']=landing_class
df[['Class']].head(8)
```

```
df.head(5)
df["Class"].mean()
```

- [GitHub URL]

EDA with Data Visualization

- Scatter plot (easily investigate the correlation between two variables):
  - Flight number to Payload Mass
  - Flight number to Launch Sites
  - Payload Mass to Launch Site
  - Flight number to Orbit
  - Payload Mass to Orbit
- Bar chart (easily compare different values under different categorises): Orbit success rate
- Line plot (clearly show the trend between two variables): Year to Success rate
- GitHub URL

EDA with SQL

- SQL queries in this project:
  - the names of the unique launch sites in the space mission
  - 5 records where launch sites begin with the string 'CCA'
  - the total payload mass carried by boosters launched by NASA (CRS)
  - average payload mass carried by booster version F9 v1.1
  - the date when the first successful landing outcome in ground pad was achieved
  - the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - the total number of successful and failure mission outcomes
  - the names of the booster_versions which have carried the maximum payload mass. Use a subquery
  - the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - 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
- GitHub URL

Build an Interactive Map with Folium

- Mark all lunch sites on a map as the marker with circles (radius 1000, color red) and lines between the launch site and closest coastline, city, railway and highway (blue line).
- Add the information can help to investigate whether the success rate has relationship with the launch site location.
- Adding the objects can help to answer the question:
  - launch sites in close proximity to railways? (Yes)
  - Are launch sites in close proximity to highways? (Yes)
  - Are launch sites in close proximity to coastline? (Yes)
  - Do launch sites keep certain distance away from cities? (Yes)
- GitHub URL

Build a Dashboard with Plotly Dash

- Launch sites dropdown menu: to select specific launch site for further investigation
- Pie chart for the success rate,
- Payload range bar for filtering specific range of landing outcomes.
- scatter plot for all success and failure cases.
- Add the information can help to find whether the launch sites can affect the success rate. Also, investigate the correlation between payload mass and mission outcomes.

- GitHub URL

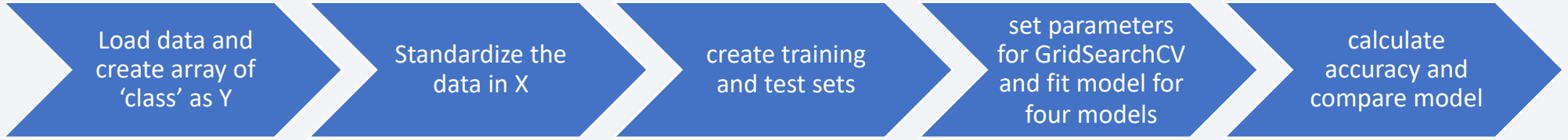
Predictive Analysis (Classification)

- Collect the adjusted dataframe from the previous section.
- Transfer dataframe “class” to NumPy array for machine learning as Y.
- Standardize the dependent variables X.
- Create training set and test set for both X and Y.
- Select four machine learning model: logistic regression, SVM, KNN and decision tree.
- Set the parameters and apply GridSearchCV with 10 folds.
- Fit the train set with four models.
- Calculate the accuracy to find the best performing classification model

- GitHub URL

Predictive Analysis (Classification)

- Flowchart:




```
data = pd.read_csv("https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/dataset_part_2.csv")
data.head()

X = pd.read_csv('https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/dataset_part_3.csv')
X.head(100)

Y = data['Class'].to_numpy()
```

```
transform = preprocessing.StandardScaler()
X = transform.fit_transform(X)
```

```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=2)

Y_test.shape
```

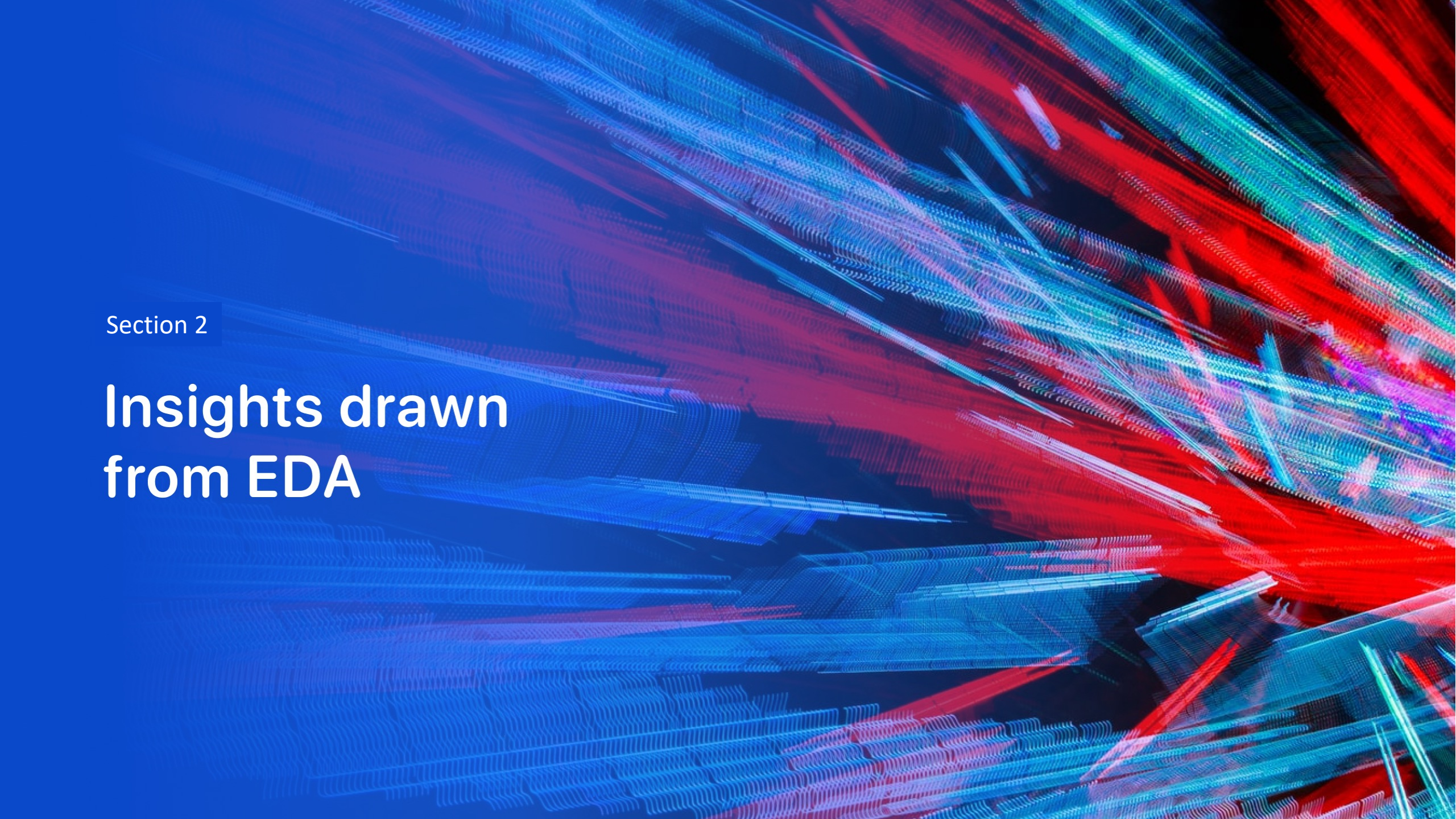
```
parameters ={"C":[0.01,0.1,1],'penalty':['l2'], 'solver':['lbfgs']}# l1 lasso l2 ridge
lr=LogisticRegression()
logreg_cv = GridSearchCV(lr,parameters,cv=10)
logreg_cv.fit(X_train,Y_train)
sorted(logreg_cv.cv_results_.keys())
```

```
print("tuned hpyerparameters :(best parameters) ",logreg_cv.best_params_)
print("accuracy :",logreg_cv.best_score_)
logreg_cv.score(X_test,Y_test)
```

- GitHub URL

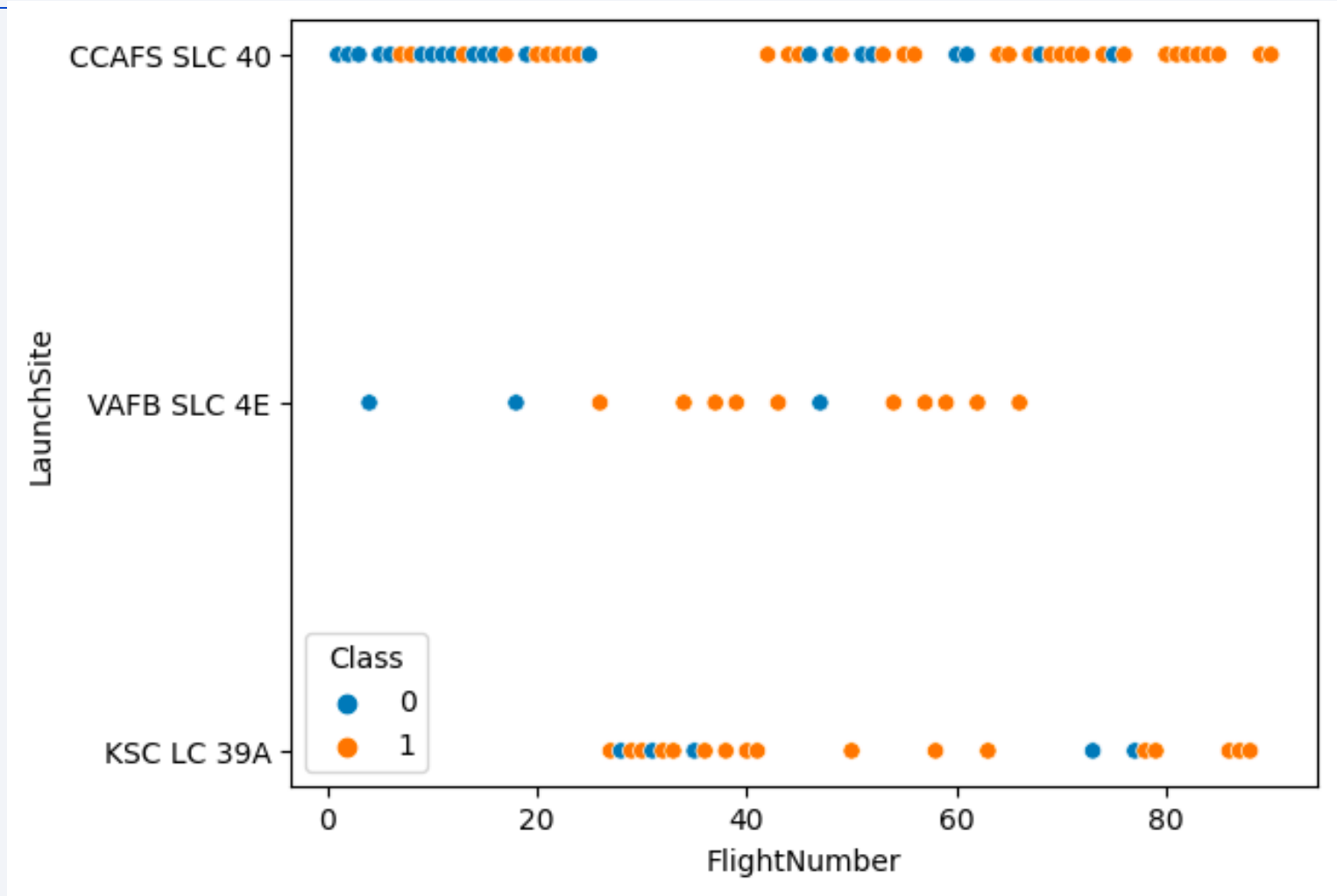
Results

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


Section 2
Insights drawn from EDA

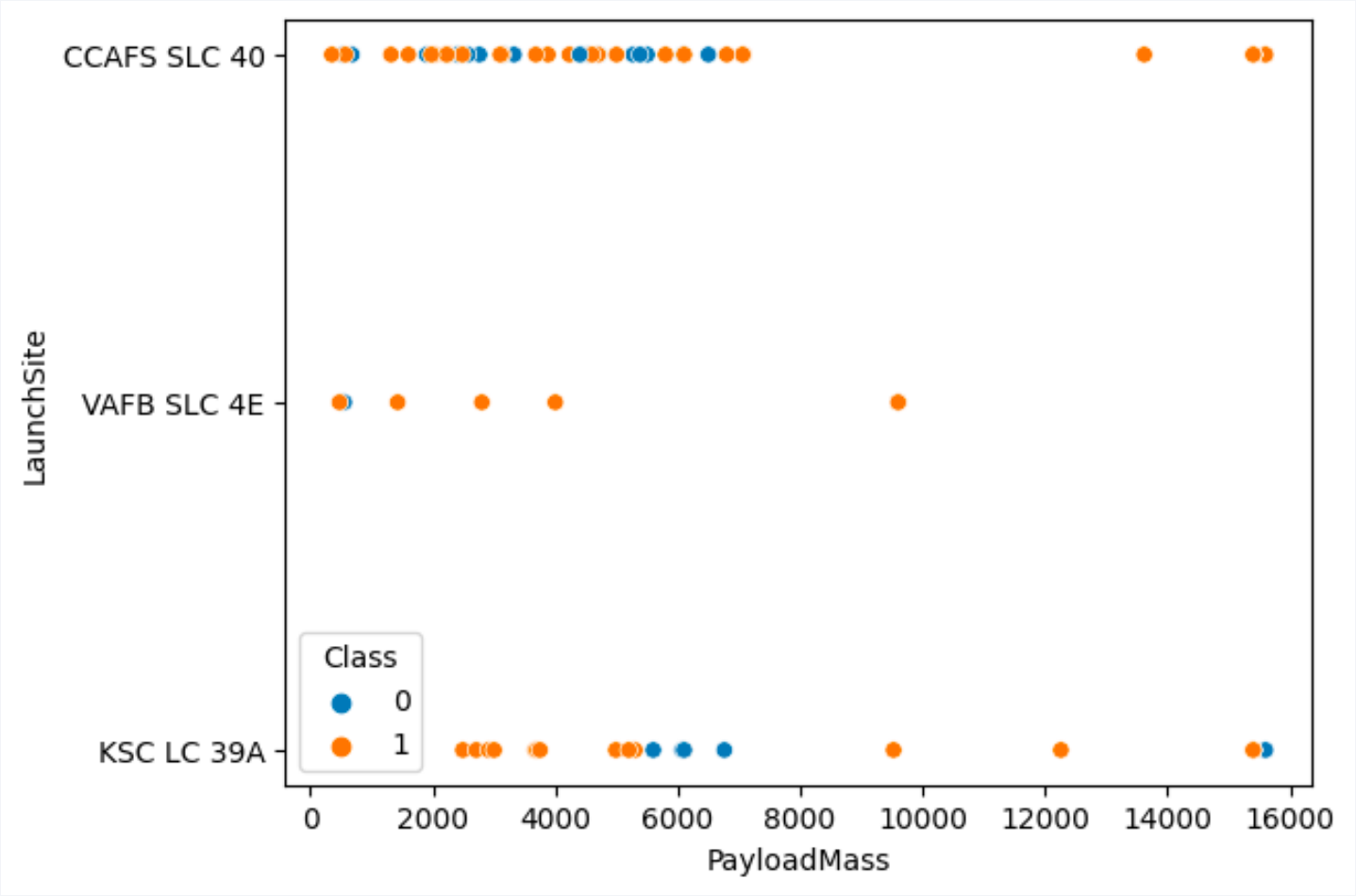
Flight Number vs. Launch Site

- Blue dot represents failure and orange dot means success.
- 1. At the beginning, test are had in CCAFS SLC 40. Then change to KSC LC 39A and final back to CCAFS SLC 40.
- 2. Most launches are at CCAFS SLC 40.



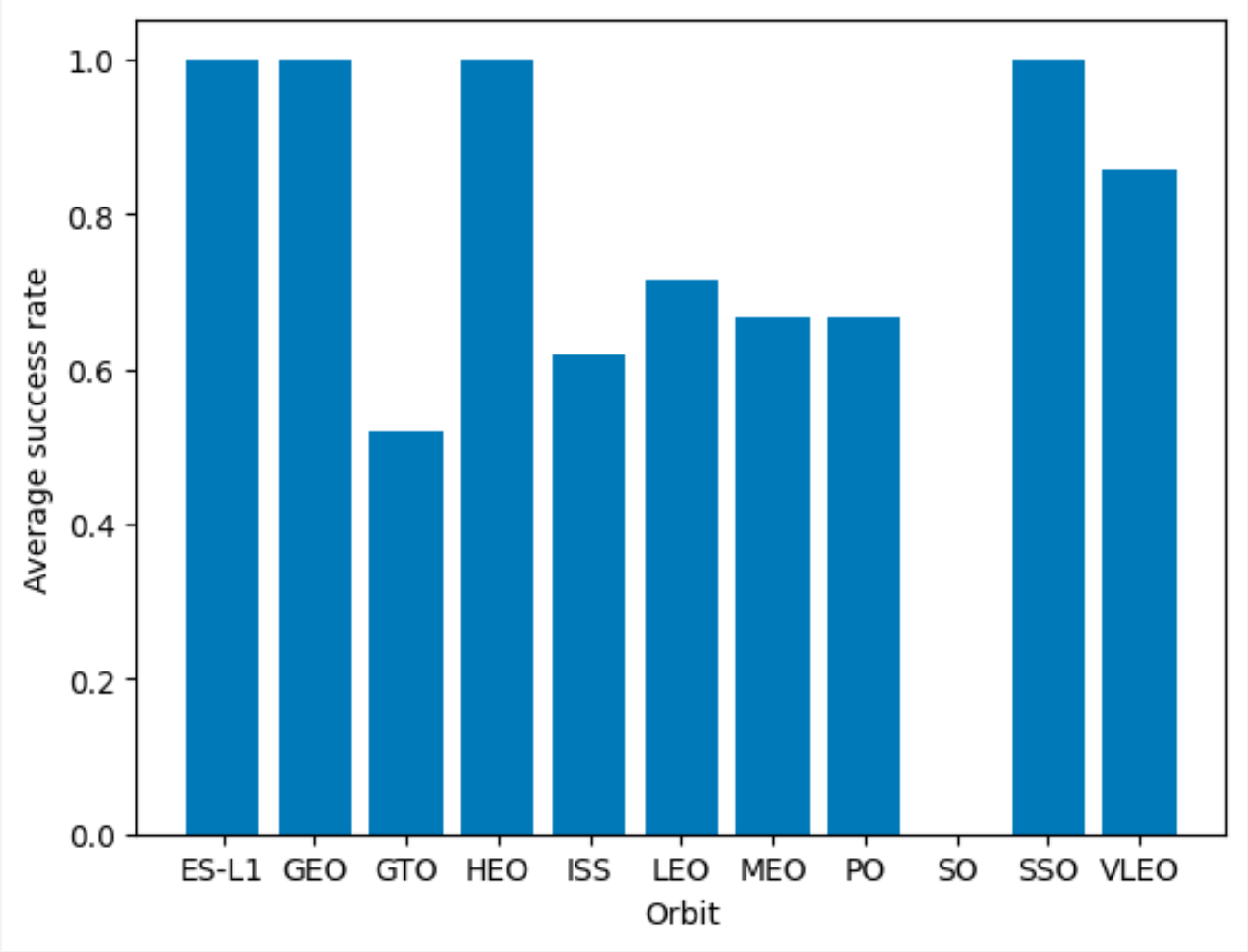
Payload vs. Launch Site

- This plot shows the relationship between payload mass and launch site.
- Higher payload mass with less failure rate.
- CCAFS and KSC take the majority proportions of small and medium flights.



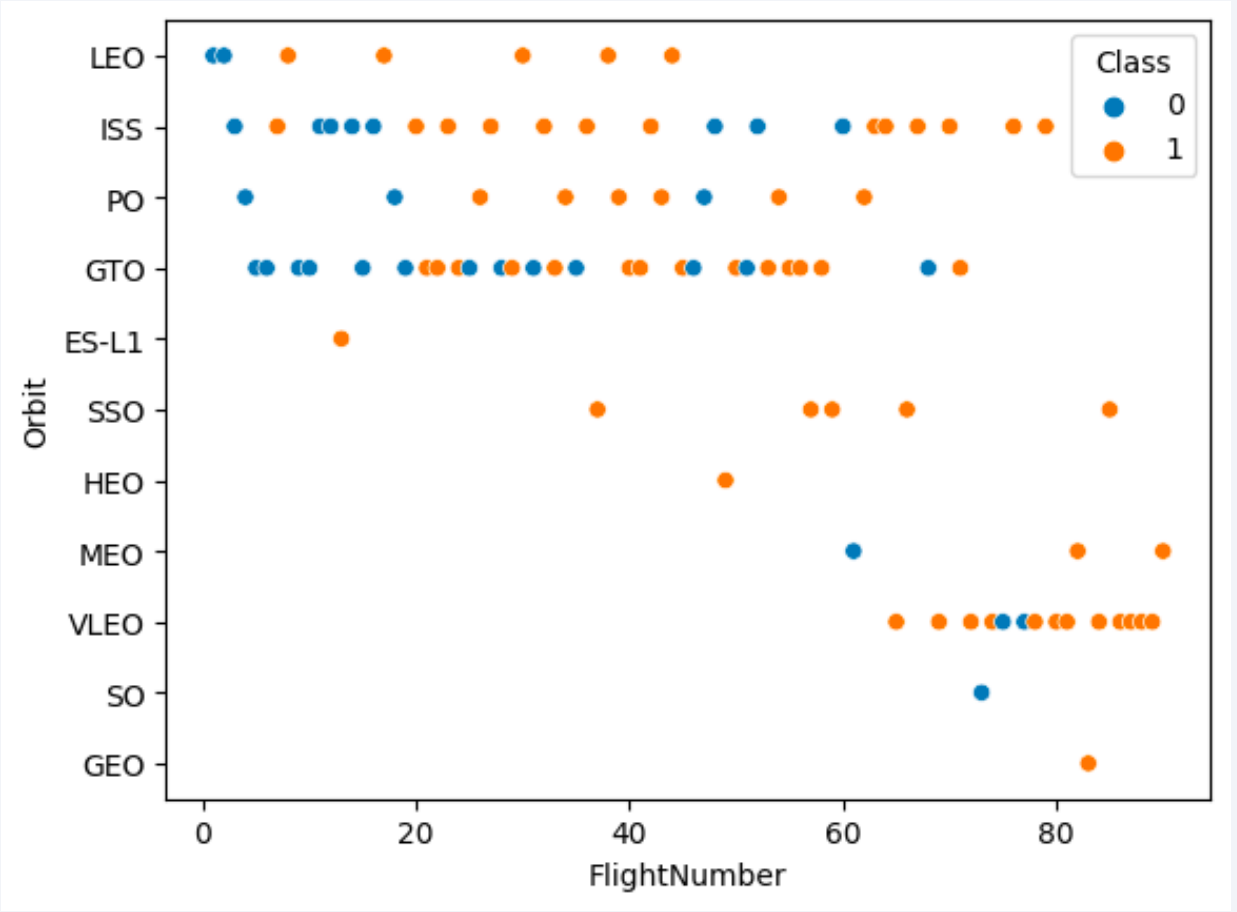
Success Rate vs. Orbit Type

- Bar chart shows the success rate for each Orbit.
- Four out of 11 have all missions launched successfully, which are ES-L1, GEO, HEO and SSO.
- All flights sent to SO were failed.



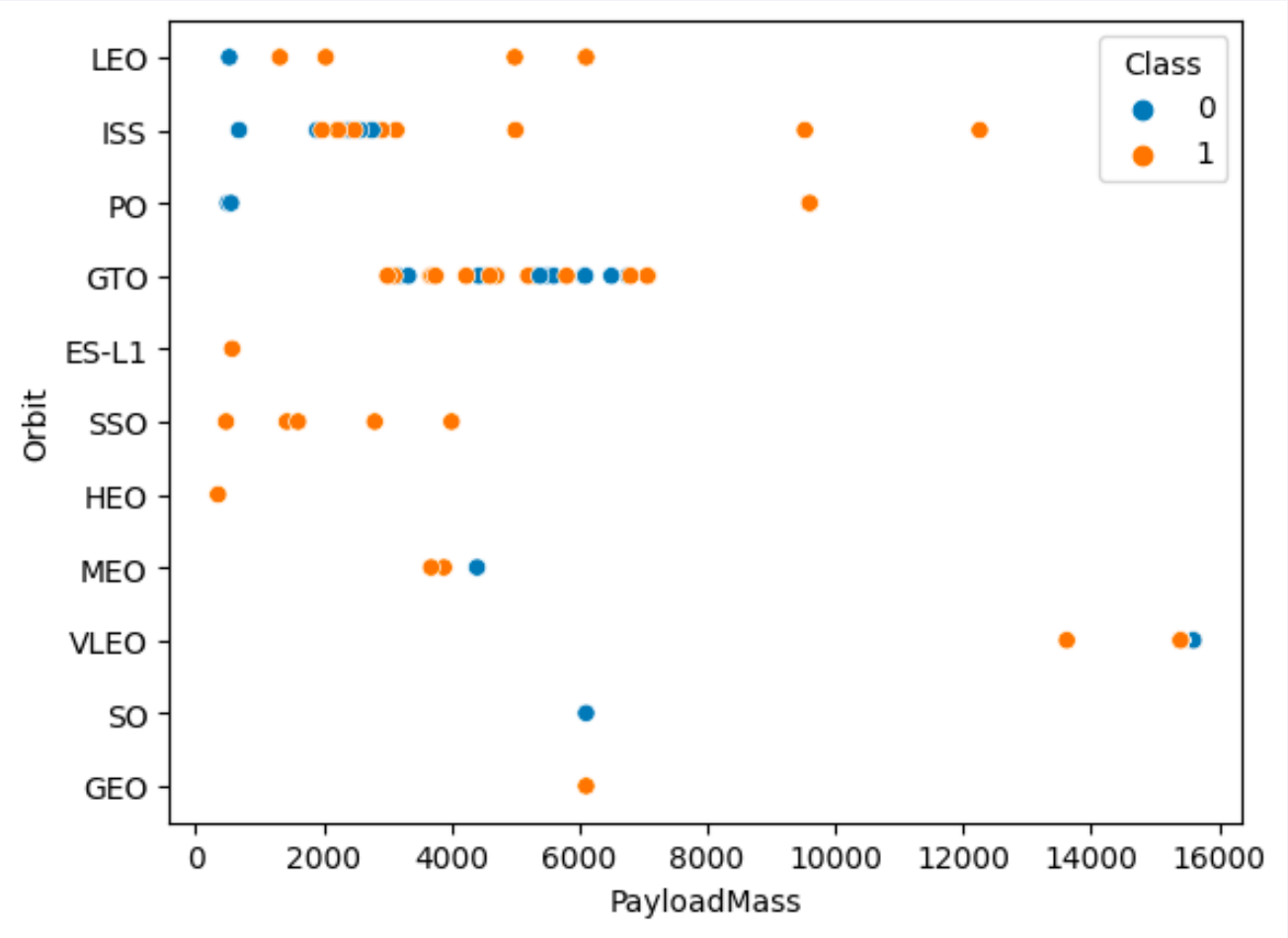
Flight Number vs. Orbit Type

- The scatter plot shows whether the success case has a correlation with orbit.
- SSO, HEO, ES-L1 and SSO have the highest success rate.
- With the increase in flights, the success rate of other orbits is also better.



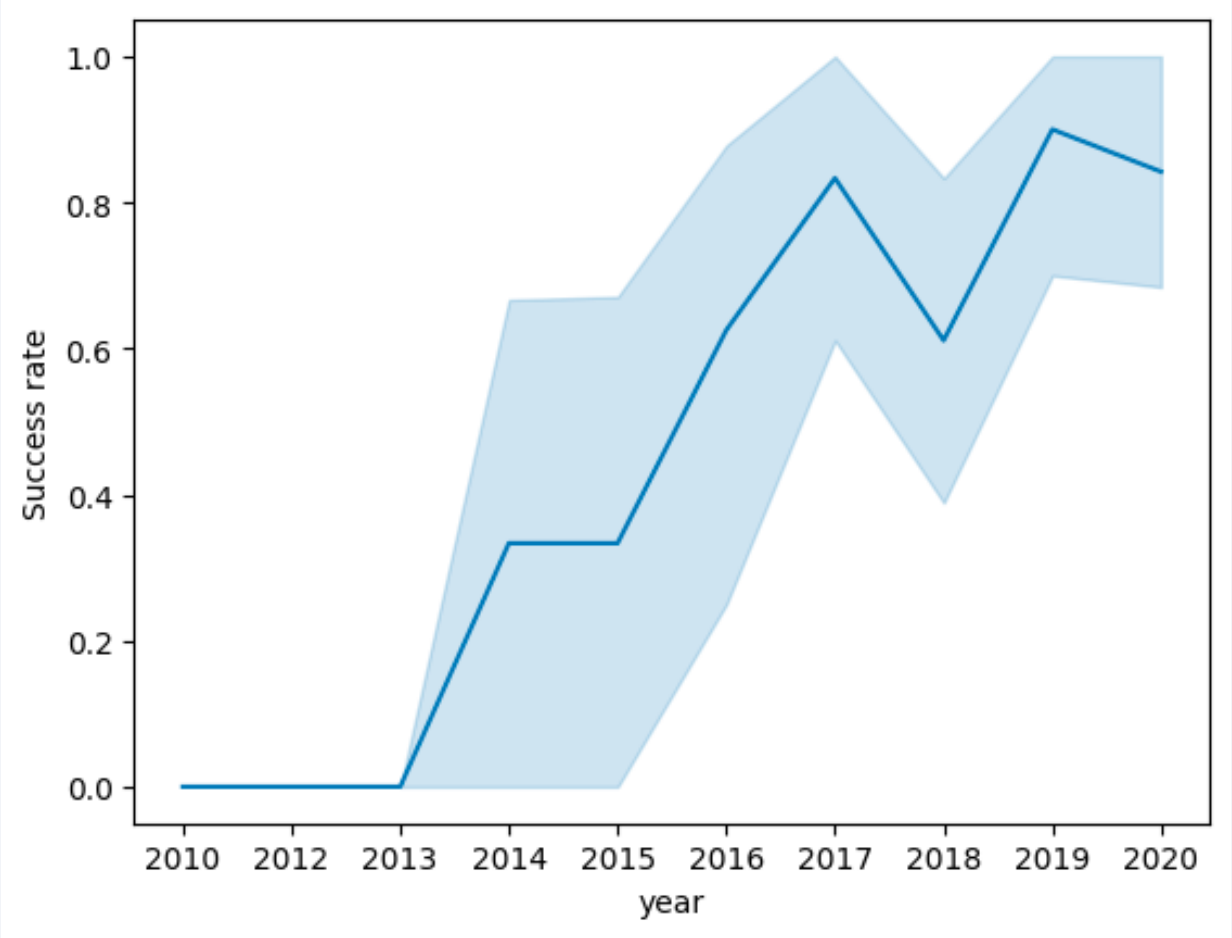
Payload vs. Orbit Type

- Scatter plot shows the relationship between payload and orbit.
- With heavy payloads the successful landing or positive landing rate are more for PO, LEO and ISS.
- ES, SSO and HEO have successful launches, however the data amount is small. It is hardly to judge.



Launch Success Yearly Trend

- The average success rate increases with from 2013 to 2020. But it drops in the year 2018.



All Launch Site Names

- SELECT distinct launch_site from spacex
- The list shows four launch sites for all SpaceX flights.

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

Launch Site Names Begin with 'CCA'

- select * from spacex where launch_site like '%CCA%' limit 5
- CCAFS start to send flights from 2010/6/4

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX
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
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)

Total Payload Mass

- select sum(payload_mass__kg_) as total_payload_mass from spacex where customer = 'NASA (CRS)'
- The number counts the total payload mass about 45000 kg.

total_payload_mass
45596

Average Payload Mass by F9 v1.1

- select avg(payload_mass__kg_) as Average_payload_mass from spacex where booster_version = 'F9 v1.1'
- The average payload mass of F9 v1.1 is about 2928, which is a small amount in the payload mass range.

average_payload_mass
2928

First Successful Ground Landing Date

- select min(date) as first_success_landing from spacex where landing__outcome = 'Success (ground pad)'
- The first success landing on ground pad is on 2015.12.22, 3 years after the first success case.

first_success_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- select booster_version from spacex where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000
- The success launches landed on drone ship is all having the booster F9 FT with version B10** with mid-range payload mass.

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

Total Number of Successful and Failure Mission Outcomes

- select substr(mission_outcome,1,7) as mission_outcome_type, count(mission_outcome) as counts from spacex group by substr(mission_outcome,1,7)

- The mission success rate is about 99.01%.

mission_outcome_type	counts
Failure	1
Success	100

Boosters Carried Maximum Payload

- select booster_version from spacex where payload_mass__kg_ = (select max(payload_mass__kg_) from spacex) order by booster_version
- The boosters can carry maximum payload (15600) with specific version after F9 B5 B1048.4.

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

2015 Launch Records

- select landing__outcome, booster_version, launch_site from spacex where landing__outcome = 'Failure (drone ship)' and year(date) = 2015
- Only two landing failure cases. Both are launched at CCAFS LC-40 and booster version is F9 v1.1 B1012 and 1015.

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

- select count(landing__outcome) as count_number from spacex where landing__outcome in ('Failure (drone ship)','Success (ground pad)') and date between '2010-06-04' and '2017-03-20' order by count(landing__outcome) desc
- 8 cases failed with above two situations between the specific time period.

count_number
8

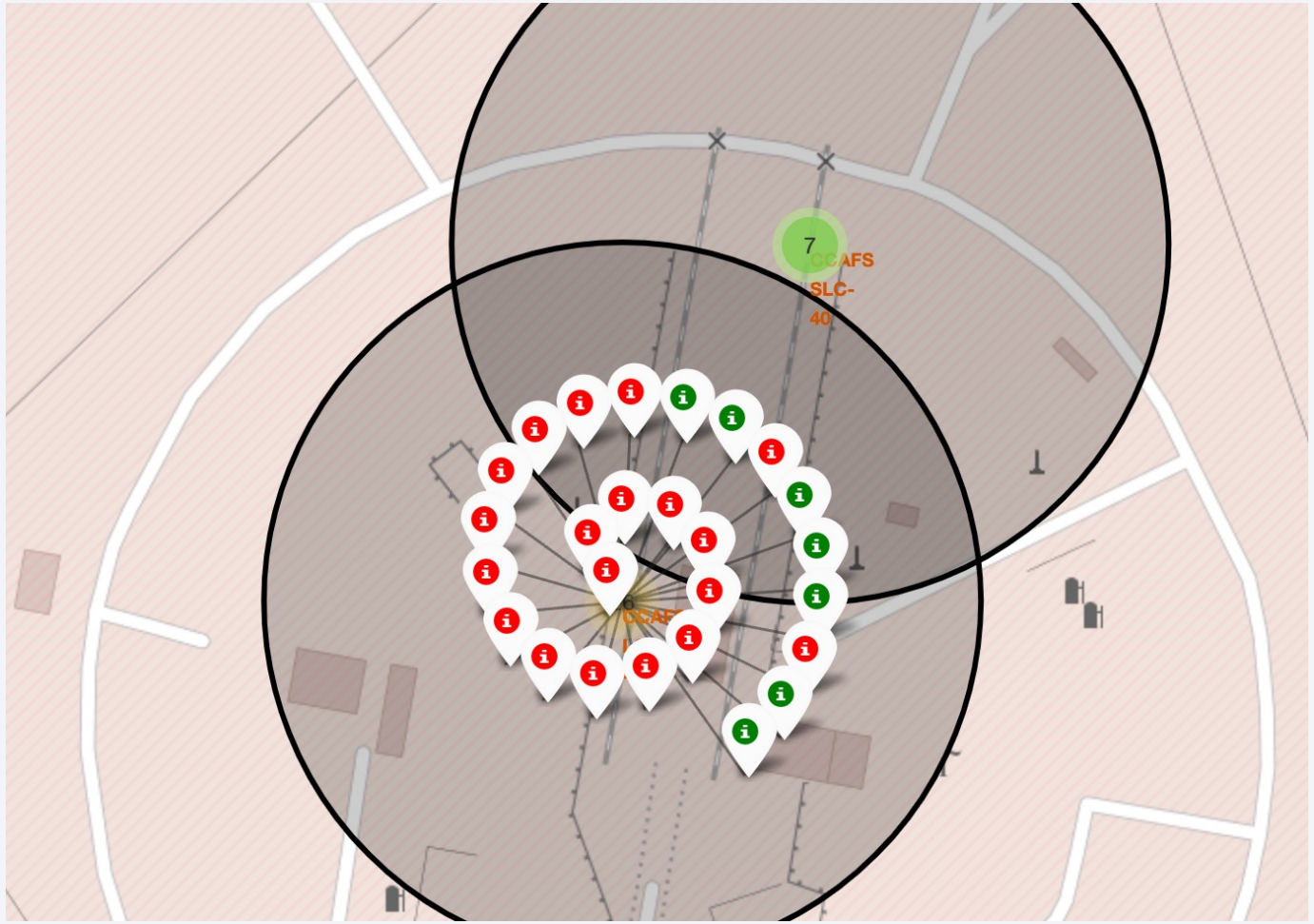

Section 3
Launch Sites
Proximities Analysis

All launch sites on world map

- All launch sites locate around the Tropic of Cancer (23°27’ N).

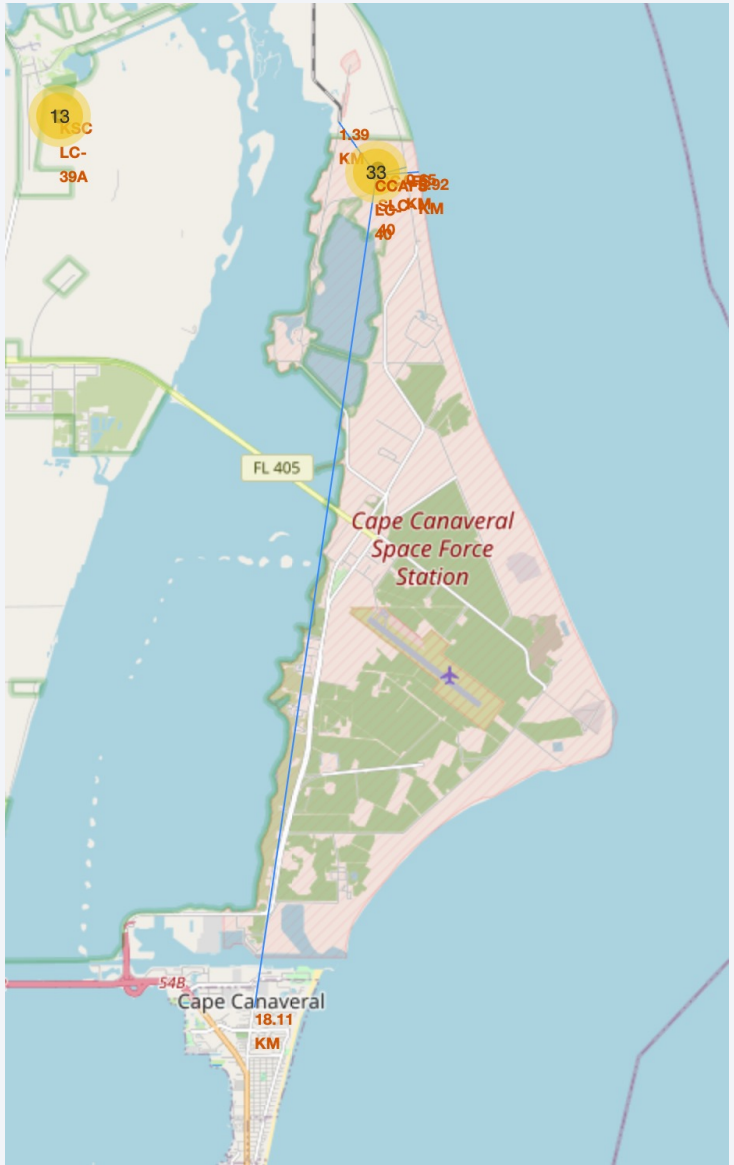
Example of colour-labeled launch outcomes

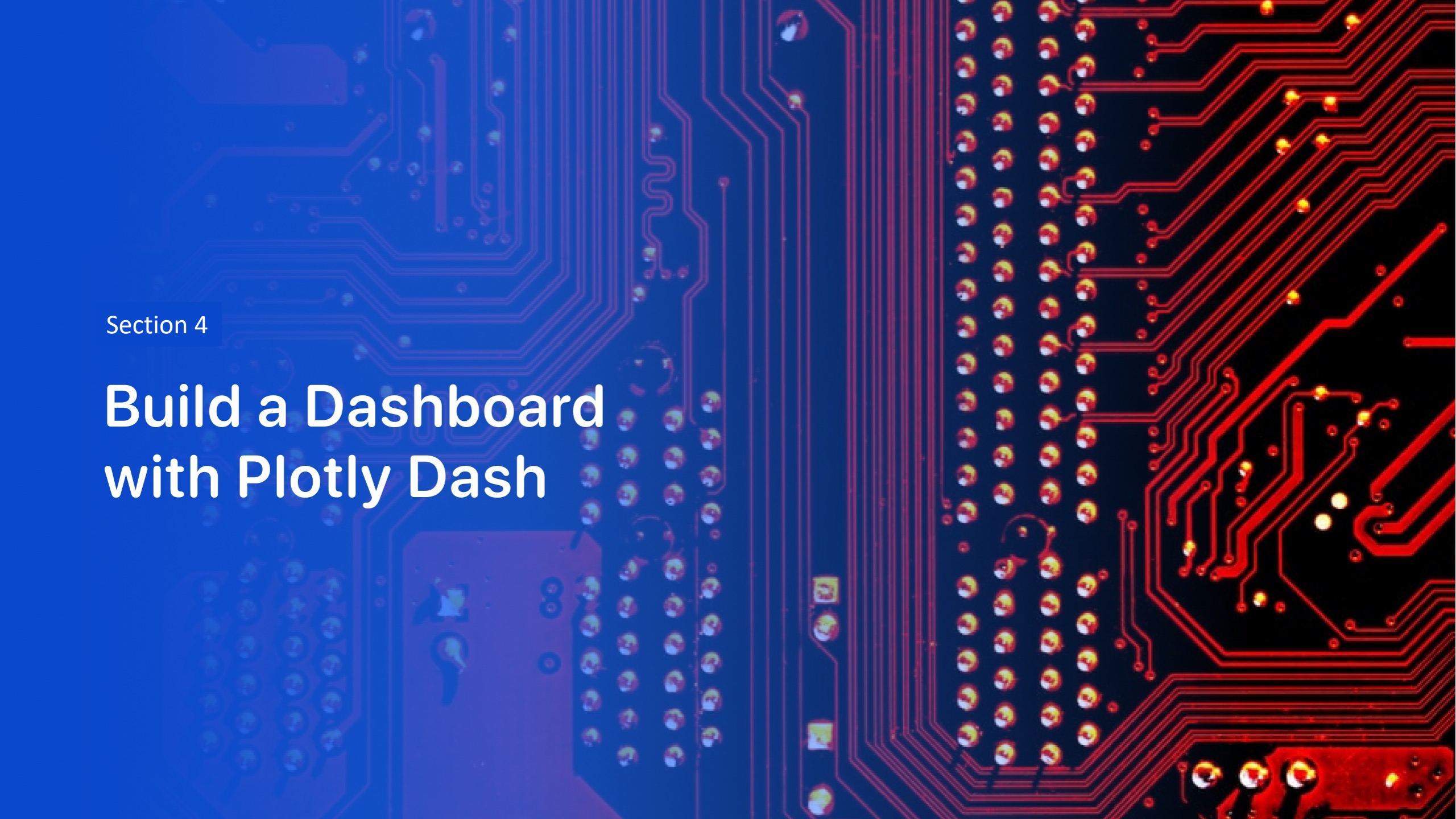
- Green (class = 1) are success and Red (class = 0) is failed.
- The display is for launch site, CCAFS LC-40



Distance between launch site and city, railway, etc.

- The graph shows that the launch site should be in close to the coastline, railways and highways.
- However, the launch site should be slight far from city than highways or railways.

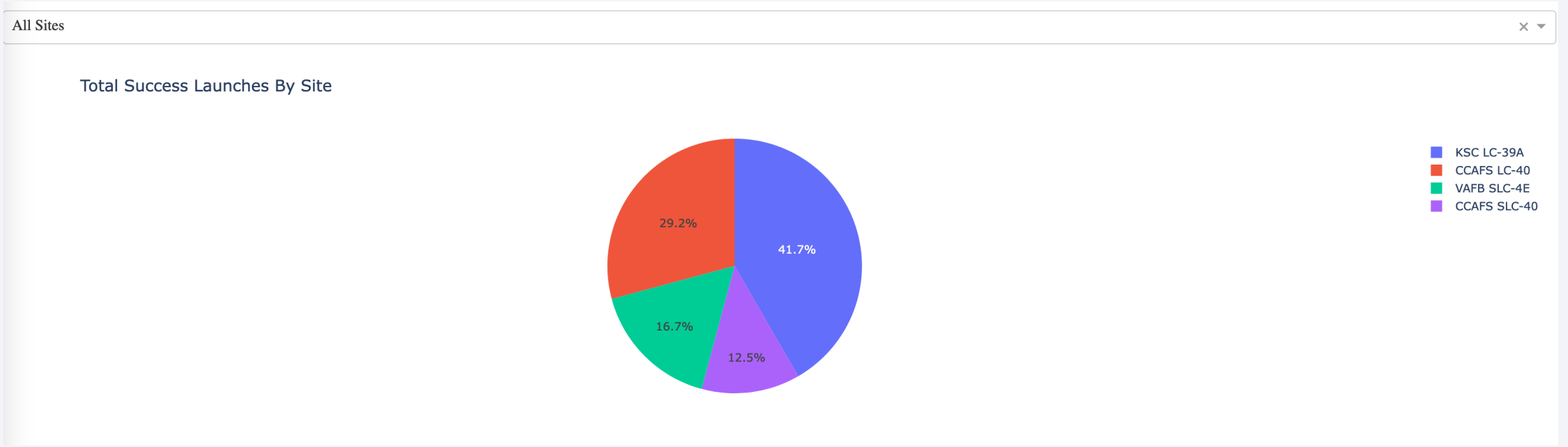




Section 4

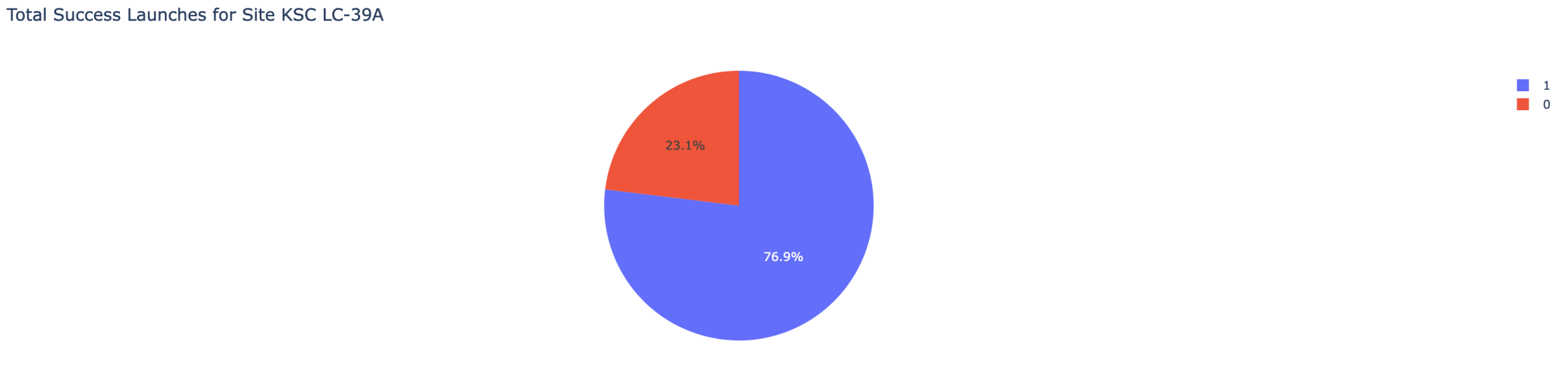
Build a Dashboard with Plotly Dash

Pie char for total success launches by site



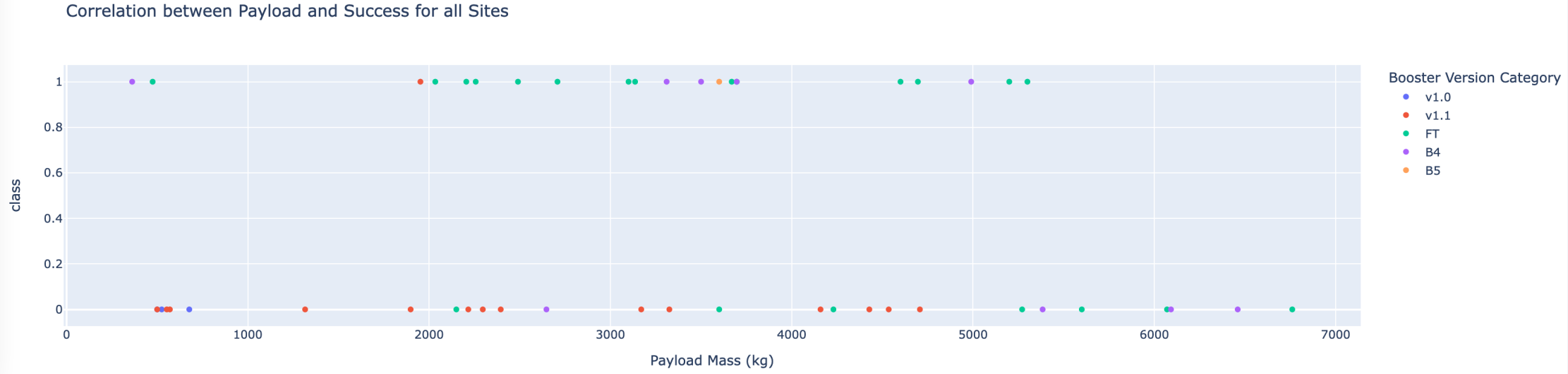
- Launch sites, KSC LC-39A and CCAFS LC-40 counts the majority of success launches.

Total success launches for site KSC LC-39A



- KSC LC-39A has the highest success launch rate than other launch sites.

Correlation between Payload and Success for all sites



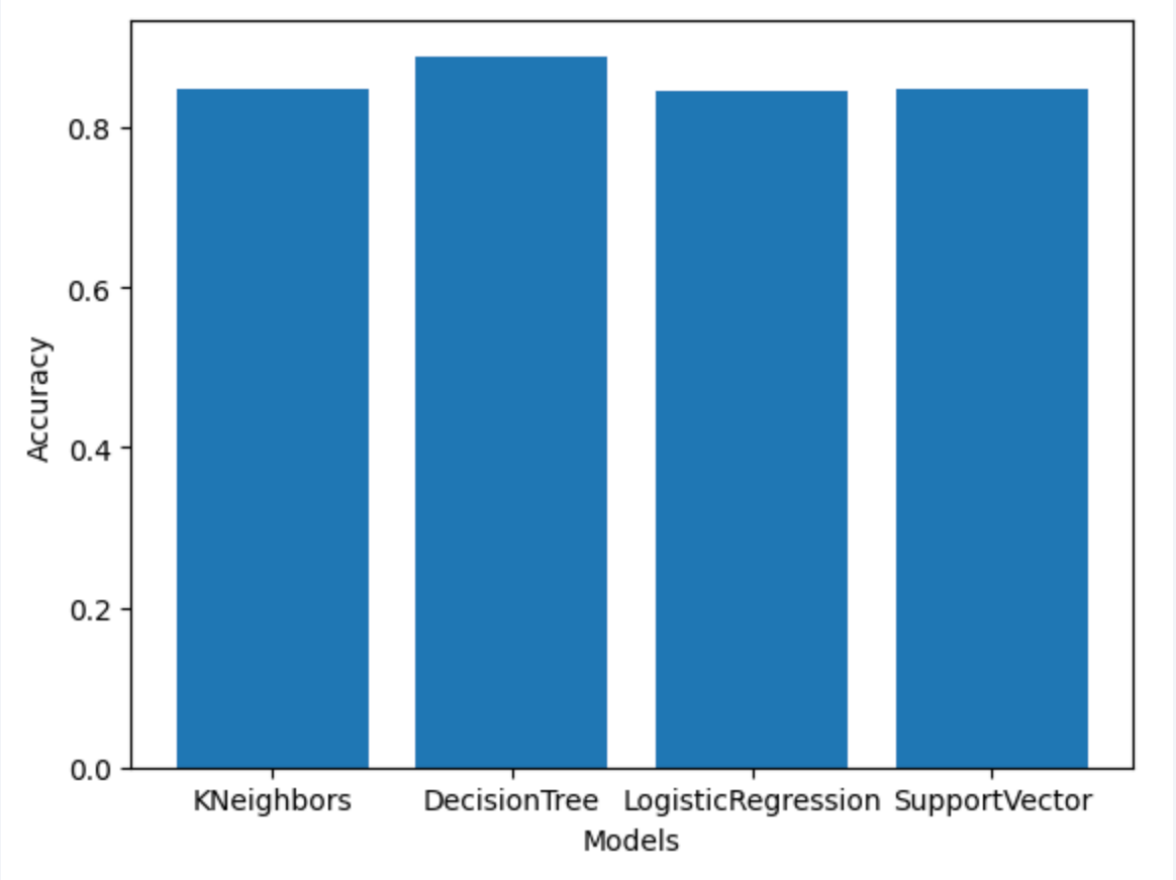
- Most success landing proportion are the flights with payload mass around 2000 to 5000. FT booster version has the highest counts.

Section 5

Predictive Analysis (Classification)

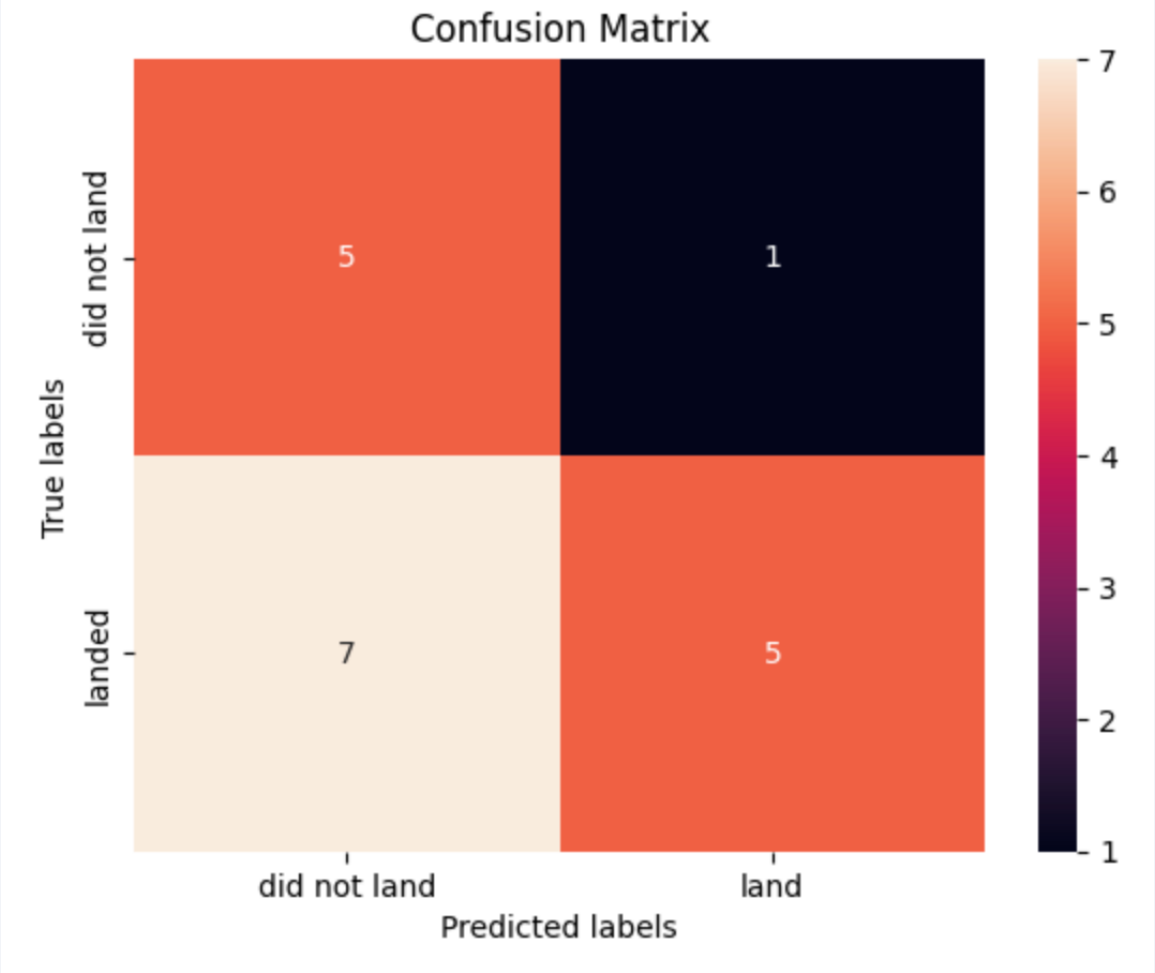
Classification Accuracy

- Best model is Decision Tree with a score of 0.889



Confusion Matrix

- The confusion matrix for the decision tree classifier shows that the classifier can distinguish between the different classes.
- The system can distinguish the “did not land” very well since only 1 case is predicted wrong. But the model cannot correctly predict the ”land” situation 7 cases are identified as “did not land”.



Conclusions

- SSO, HEO, ES-L1 and SSO have the highest success rate.
- The average success rate increases with from 2013 to 2020.
- KSC LC-39A has the highest success launch rate than other launch sites.
- Most success landing proportion are the flights with payload mass around 2000 to 5000.
- FT booster version has the highest counts.
- The Decision tree model has the highest accuracy in prediction.

Thank you!