IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Bob
05/01/22

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - Exploratory Data Analysis (Data Visualisation and SQL)
  - Interactive Map (Folium)
  - Dashboard (Ploty Dash)
  - Classification (Predictive Analysis)
- Summary of all 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 one can accurately predict the likelihood of the first stage rocket landing successfully, the cost of a launch can be determined.
- Problems you want to find answers
  - What factors influences the success rate of a rocket landing?
  - How does each factor affect the landing success rate for a certain rocket type?
  - What conditions are required to achieve the highest landing success rate?

Section 1

Methodology

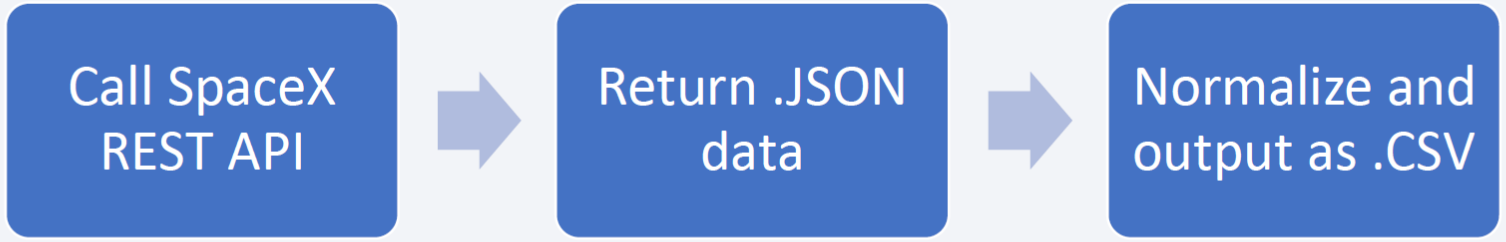
Methodology

Executive Summary

- Data collection methodology:
  - SpaceX REST API
  - Web Scraping (Falcon 9 historical launch records from a Wikipedia page)
- Perform data wrangling
  - Transform data collected based on landing outcomes
  - 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
  - Build, tune, and evaluate classification models

Data Collection

1. SpaceX REST API
   - A series of GET request was performing to return information from API in json format
   - Data is subsequently normalized using Pandas to output a CSV file



2. Wikipedia Falcon 9 Launch data Web Scraping
   - Extract Data using beautiful soup
   - Normalize and output as CSV file

Data Collection – SpaceX API

1. Get response from SpaceX API:

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

2. Convert response as a dataframe:

```
data1 = response.json()
data= pd.json_normalize(data1)
data
```

3. Use Helper function to extract relevant :

```
getBoosterVersion(data)
getLaunchSite(data)
getPayloadData(data)
getCoreData(data)
```

4. Construct dictionary:

```
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}
```

5. Clean up unwanted data:

```
# Hint data['BoosterVersion']!='Falcon 1'
data_falcon9 = data[data.BoosterVersion == 'Falcon 9']
data_falcon9.loc[:,'FlightNumber'] = list(range(1, data_falcon9.shape[0]+1))
```

6. Address missing values

```
# Calculate the mean value of PayloadMass column
Mean_PayloadMass = data_falcon9.PayloadMass.mean()
# Replace the np.nan values with its mean value
data_falcon9['PayloadMass'] = data_falcon9['PayloadMass'].replace(np.nan, Mean_PayloadMass)
```

7. Output flat file (.CSV)

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

https://github.com/CBX00/Data-Scientist-Capstone/blob/febc7b030671ff1c90e2c1d3df1f7eb3303f03b9/Data%20Collection%20API.ipynb

Data Collection - Scraping

1. Define HTML Source

```
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).text
```

2. Create Beautiful Soup Object

```
soup = BeautifulSoup(response, 'html.parser')
```

3. Find tables

```
html_tables = soup.find_all("table")
print(html_tables)
```

4. Extract Column Header

```
column_names = []

# Apply find_all() function with `th` element on first_launch_table
# Iterate each th element and apply the provided extract_column_from_header() to
get a column name
# Append the Non-empty column name (`if name is not None and len(name) > 0`) int
o a list called column_names
temp = soup.find_all('th')
for x in range(len(temp)):
    try:
     name = extract_column_from_header(temp[x])
     if (name is not None and len(name) > 0):
        column_names.append(name)
    except:
     pass
```

5. Create dictionary

```
launch_dict= dict.fromkeys(column_names)

# Remove an irrelevant column
del launch_dict['Date and time ( )']

# Let's initial the launch_dict with each value to be an empty list
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'] = []
# Added some new columns
launch_dict['Version Booster']=[]
launch_dict['Booster landing']=[]
launch_dict['Date']=[]
launch_dict['Time']=[]
```

6. Append data to Dictionary keys

7. Convert Dictionary into Dataframe

```
#df=pd.DataFrame(launch_dict)

headings = []
for key,values in dict(launch_dict).items():
    if key not in headings:
        headings.append(key)
    if values is None:
        del launch_dict[key]

def pad_dict_list(dict_list, padel):
    lmax = 0
    for lname in dict_list.keys():
        lmax = max(lmax, len(dict_list[lname]))
    for lname in dict_list.keys():
        ll = len(dict_list[lname])
        if ll < lmax:
            dict_list[lname] += [padel] * (lmax - ll)
    return dict_list

pad_dict_list(launch_dict,0)

df = pd.DataFrame.from_dict(launch_dict)
df.head()
```

8. Output flat file (.csv)

```
df.to_csv('spacex_web_scraped.csv', index=False)
```

https://github.com/CBX00/Data-Scientist-Capstone/blob/febc7b030671ff1c90e2c1d3df1f7eb3303f03b9/Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling

1. Perform exploratory Data Analysis and determine Training Labels
   - Exploratory Data Analysis
   - Determine Training Labels
2. Calculate the number of launches on each site
3. Calculate the number and occurrences of each orbit
4. Calculate the number and occurrence of mission outcome per orbit type

https://github.com/CBX00/Data-Scientist-Capstone/blob/febc7b030671ff1c90e2c1d3df1f7eb3303f03b9/EDA.ipynb

EDA with Data Visualization

- Scatter Plot of:
  - Flight Number vs Launch Site
  - Payload vs Launch Site
  - Flight Number vs Orbit type
  - Payload vs Orbit type
- Bar graph of success rate of each Orbit type
- Line graph of Yearly launch success trend

https://github.com/CBX00/Data-Scientist-Capstone/blob/febc7b030671ff1c90e2c1d3df1f7eb3303f03b9/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- Insights gathered via SQL:
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in ground pad was achieved.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
  - List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - Rank the count of successful landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

https://github.com/CBX00/Data-Scientist-Capstone/blob/febc7b030671ff1c90e2c1d3df1f7eb3303f03b9/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Using Launch_Site coordinates, launch data were marked on an interactive map to aid visualization
- Launch_Outcome can also be visualized based on the classes:
  - Successful launch (Class 1) in Green
  - Failed launch (Class 0) in Red
- Distances between an identified launch site (CCAFS SLC-40) to its proximities were determined to answer some trends related to launch site location:
  - Are launch sites in close proximity to railways?
  - Are launch sites in close proximity to highways?
  - Are launch sites in close proximity to coastline?
  - Do launch sites keep certain distance away from cities?

https://github.com/CBX00/Data-Scientist-Capstone/blob/febc7b030671ff1c90e2c1d3df1f7eb3303f03b9/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

- A Dashboard provide live trending of proportion of successful launches by varying launch site when SpaceX conducts new launches which will help to improve prediction accuracy overtime
- A Scatterplot of Launch Outcome vs Payload Mass (kg) for different booster versions provide an easily observable relationship between the variables

https://github.com/CBX00/Data-Scientist-Capstone/blob/febc7b030671ff1c90e2c1d3df1f7eb3303f03b9/spacex_dash_app.ipynb

Predictive Analysis (Classification)

- Building the classification model
  - Load dataset into NumPy and Pandas
  - Transformation: Standardize data and reassign to variable
  - Split data to Train and Test set
  - Determine test sample count
  - Perform logistic regression object and create GridSearchCV
  - Fit dataset into GridSearchCV and train with dataset
- Evaluating and Improving the classification model
  - Check accuracy of each model
  - Get best parameter (Tuned Hyperparameters)
  - Plot Confusion Matrix
- Finding the best performing classification model
  - Determining the model with the best accuracy score

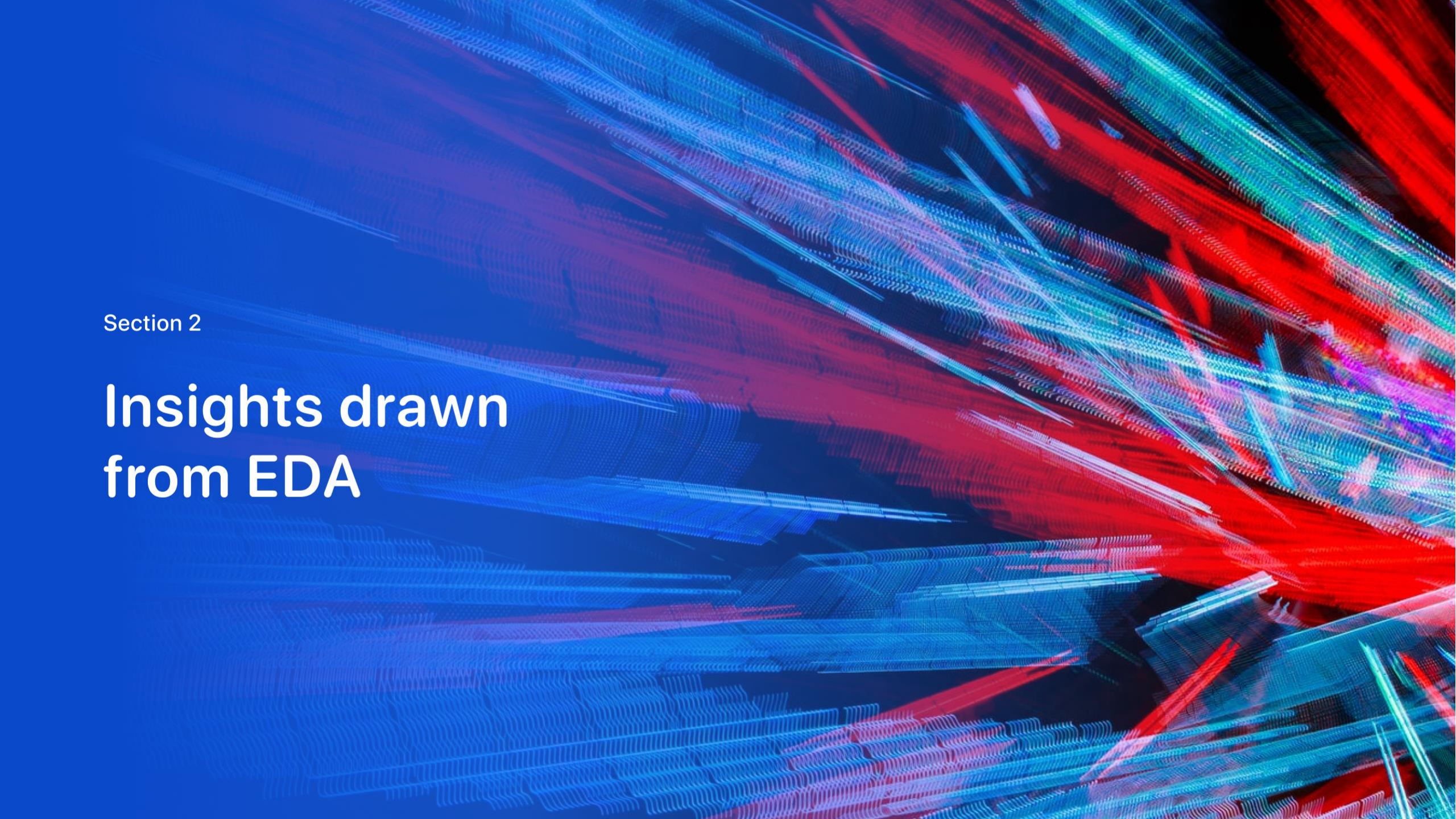
https://github.com/CBX00/Data-Scientist-Capstone/blob/febc7b030671ff1c90e2c1d3df1f7eb3303f03b9/Machine%20Learning%20Prediction.ipynb

Results

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

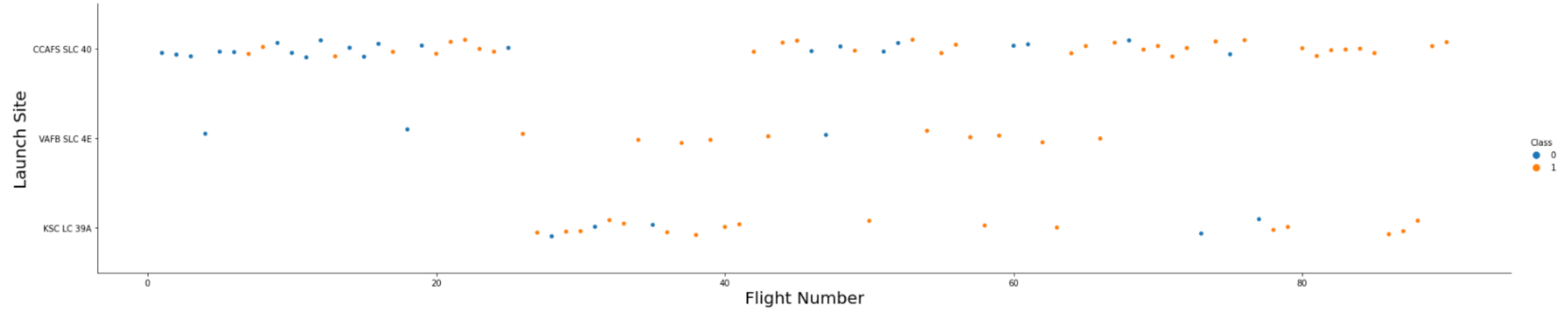
Section 2

Insights drawn from EDA



Flight Number vs. Launch Site

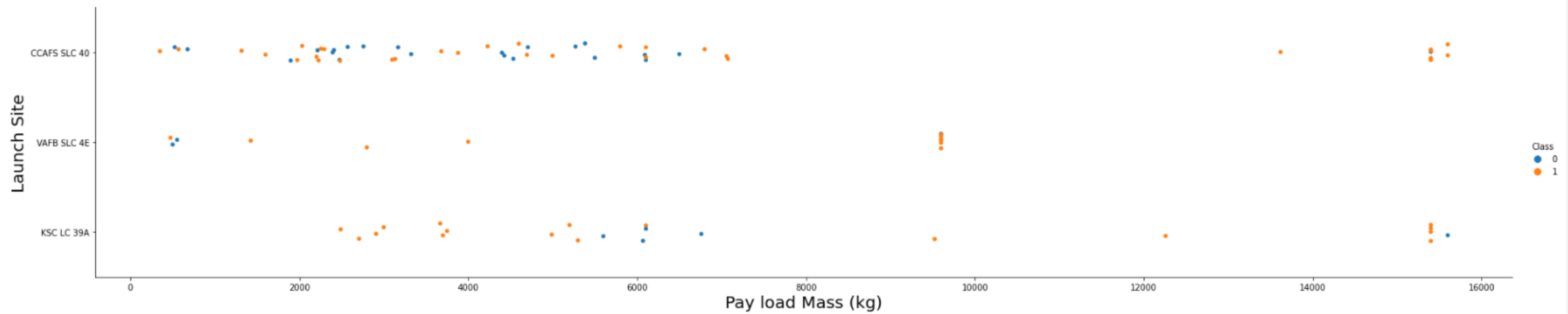
- Scatter plot of Flight Number vs. Launch Site



- General trend of higher success rate with higher flight number

Payload vs. Launch Site

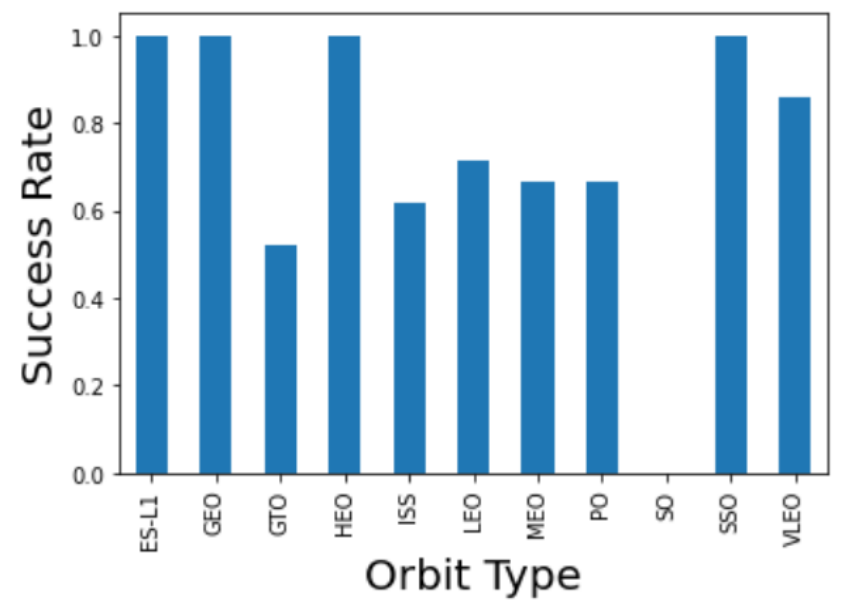
- Scatter plot of Payload vs. Launch Site



- Launches from CCAFS SLC 40 have a higher success rate at greater payload mass for
- No obvious trend if launch site is dependent on payload mass for successful launch

Success Rate vs. Orbit Type

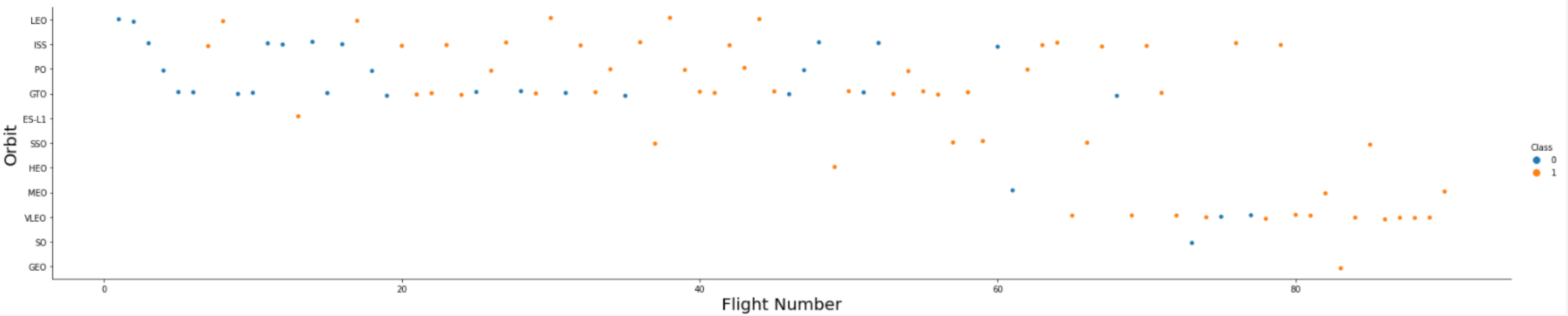
- Show a bar chart for the success rate of each orbit type



- Orbit type with the highest success rates are ES-L1, GEO, HEO, SSO

Flight Number vs. Orbit Type

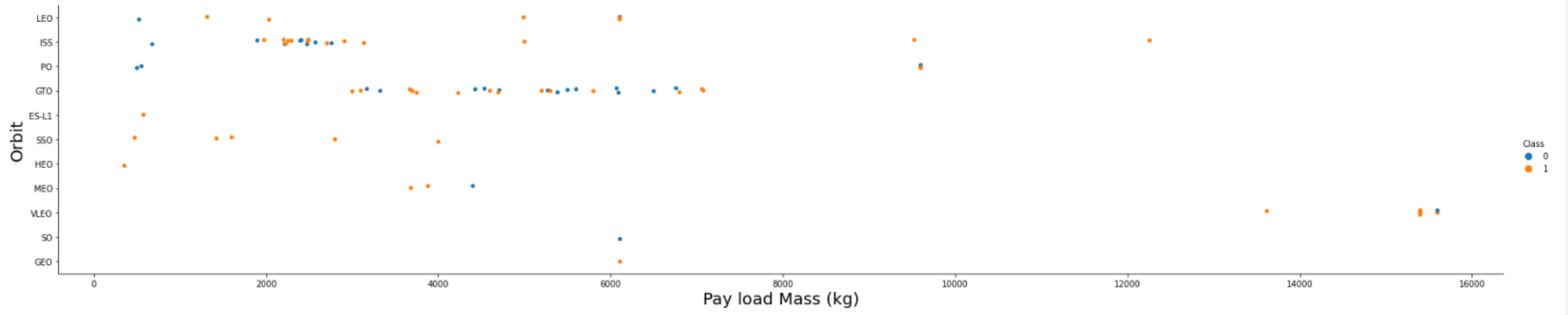
- Scatter point of Flight number vs. Orbit type



- Success appears related to the number of flights for LEO orbit
- No obvious relationship between flight number when in GTO orbit.

Payload vs. Orbit Type

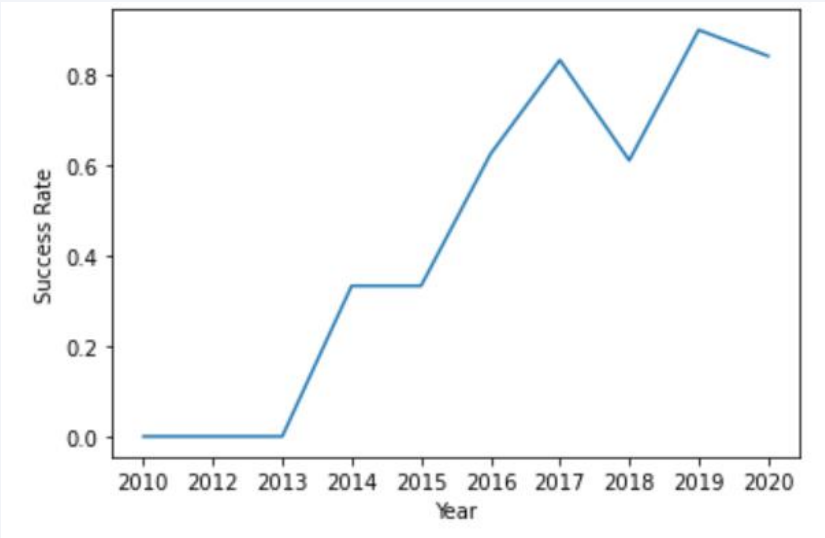
- Scatter point of payload vs. orbit type



- With heavy payloads - successful landing or positive landing rates are higher for Polar, LEO, and ISS orbit types.
- No obvious trend for GTO orbit

Launch Success Yearly Trend

- Line chart of yearly average success rate



- In general, success rate of launches have been increasing since 2010

All Launch Site Names

- Names of the unique launch sites

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

```
%sql select distinct(LAUNCH_SITE) from SPACEXDATASET
```

```
 * ibm_db_sa://fxn66192:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f
Done.
```

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`

```
%sql select * from SPACEXDATASET where LAUNCH_SITE like 'CCA%' limit 5
```

```
 * ibm_db_sa://fxn66192:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

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 was 45596 kg

```
%sql select SUM(PAYLOAD_MASS__KG_) from SPACEXDATASET where CUSTOMER = 'NASA (CRS)'
```

```
 * ibm_db_sa://fxn66192:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.c
Done.
```

1
45596

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1

```
%sql select AVG(PAYLOAD_MASS__KG_) from SPACEXDATASET where BOOSTER_VERSION = 'F9 v1.1'
```

```
 * ibm_db_sa://fxn66192:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.data
Done.
```

1
2928

First Successful Ground Landing Date

- Dates of the first successful landing outcome on ground pad

```
%sql select MIN(DATE) from SPACEXDATASET where LANDING__OUTCOME = 'Success (ground pad)'
```

```
 * ibm_db_sa://fxn66192:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.data
Done.
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
%sql select DISTINCT(BOOSTER_VERSION) from SPACEXDATASET where LANDING__OUTCOME = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS_
```

```
 * ibm_db_sa://fxn66192:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes

```
%sql select COUNT(MISSION_OUTCOME) from SPACEXDATASET
```

```
 * ibm_db_sa://fxn66192:***@ba99a9e6-d59e-4883-8fc0-d6a
Done.
```

1
101

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass

```
%sql select BOOSTER_VERSION from SPACEXDATASET where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXDATASET)
```

```
 * ibm_db_sa://fxn66192:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

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

```
%sql select Booster_Version, Launch_Site from SPACEXDATASET WHERE DATE like '2015%' AND Landing__Outcome = 'Failure (drone ship)'
```

```
 * ibm_db_sa://fxn66192:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

- Ranked 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

```
%sql select LANDING__OUTCOME, COUNT(LANDING__OUTCOME) \
from SPACEXDATASET \
WHERE DATE between '2010-06-04' AND '2017-03-20' \
GROUP BY LANDING__OUTCOME \
ORDER BY COUNT(LANDING__OUTCOME) DESC
```

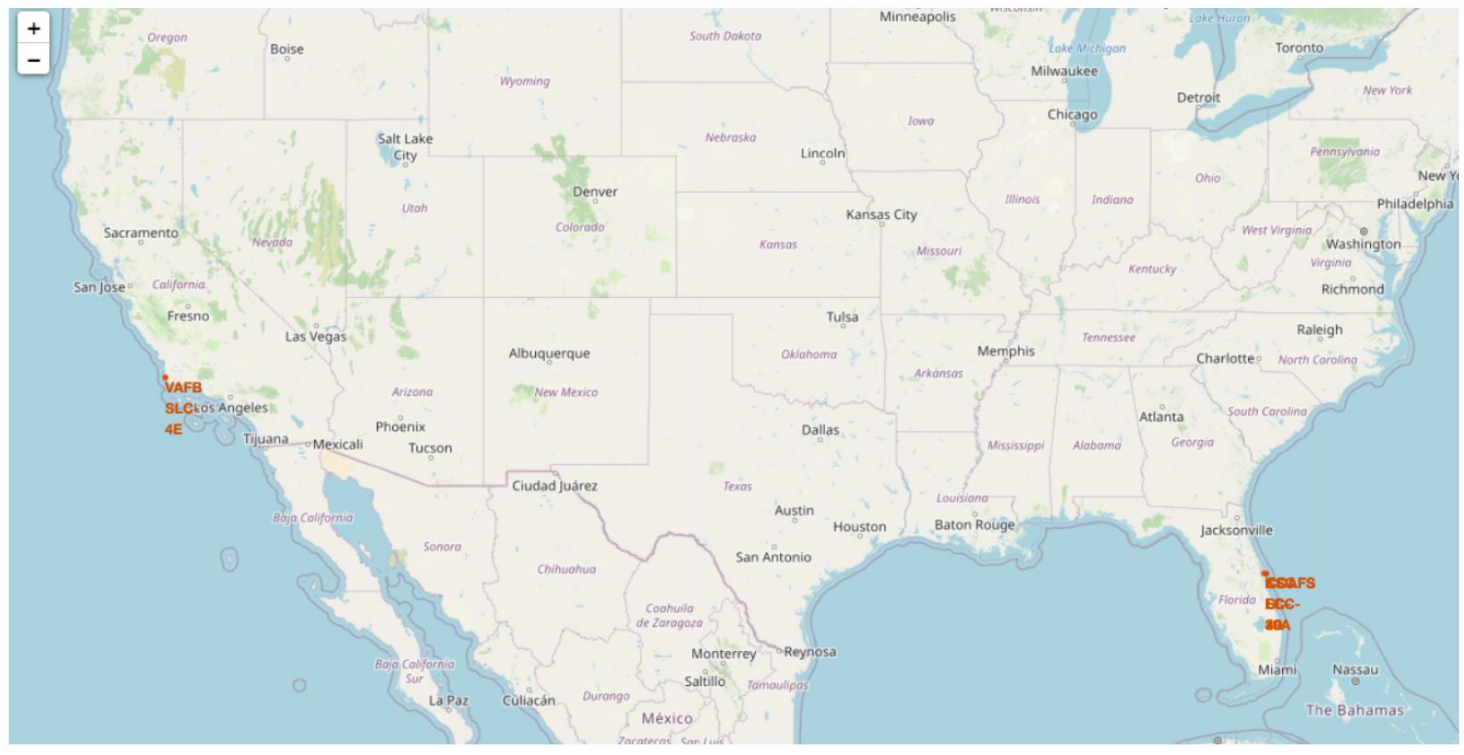
```
 * ibm_db_sa://fxn66192:***@ba99a9e6-d59e-4883-8fc0-d6a
Done.
```

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

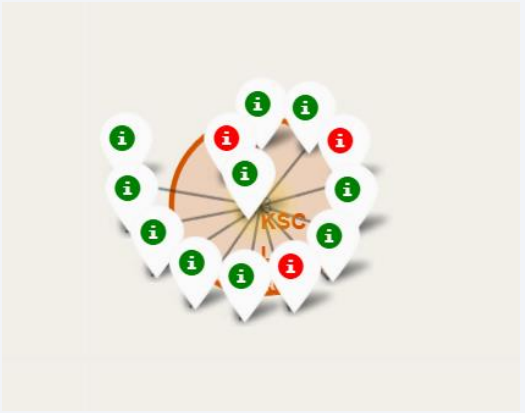
Launch Sites



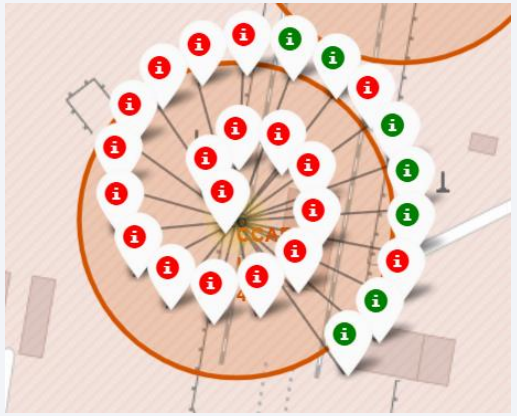
- Launch sites are located at coastlines and close to the equator

Launch outcomes at each launch site

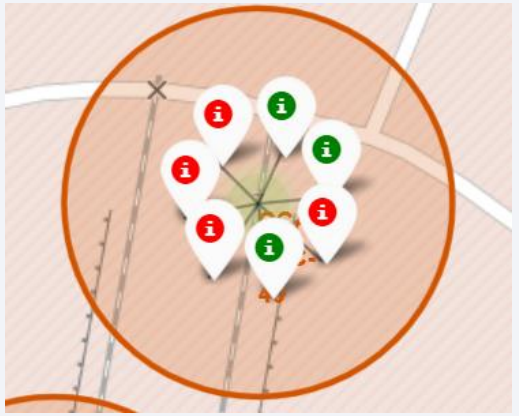
VAFB SLC-4E



KSC LC-39A



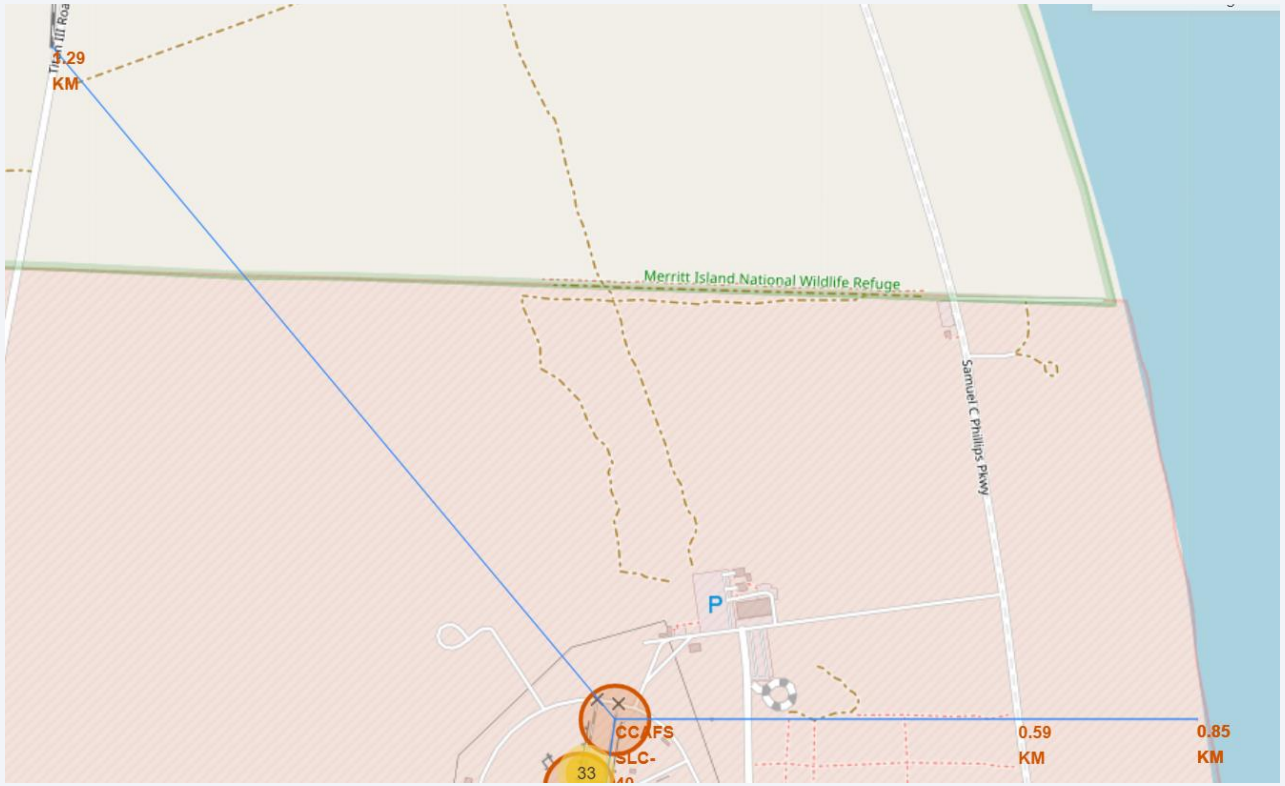
CCAFS LC-40



CCAFS SLC-40

- Green markers indicate successful launches
- Red markers indicate unsuccessful launches
- From this visual representation, KSC LC-39A has the highest proportion of successes, followed by CCAFS SLC-40, VAFB SLC-4E and CCFS LC-40

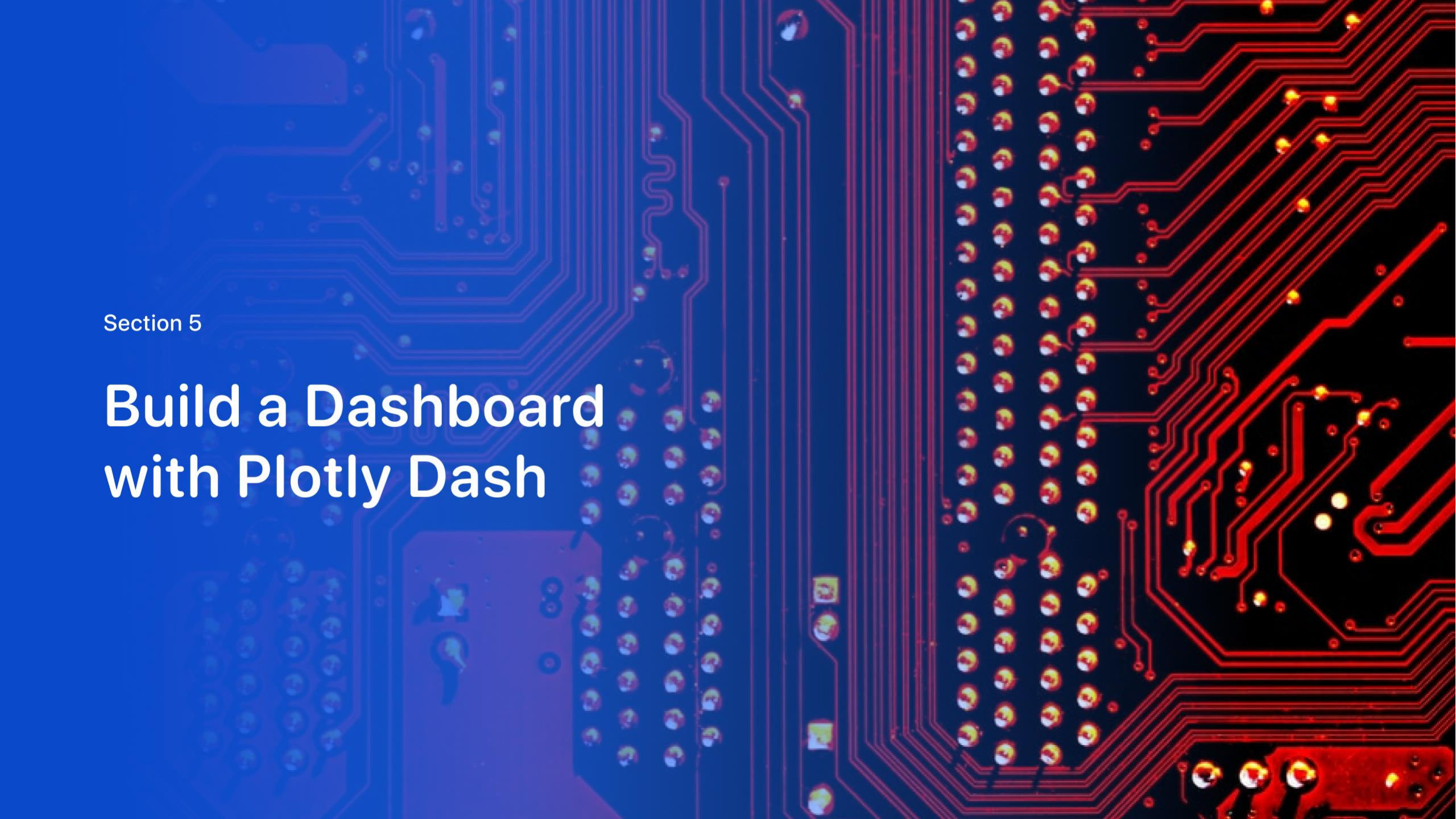
Proximity of launch site to various infrastructure



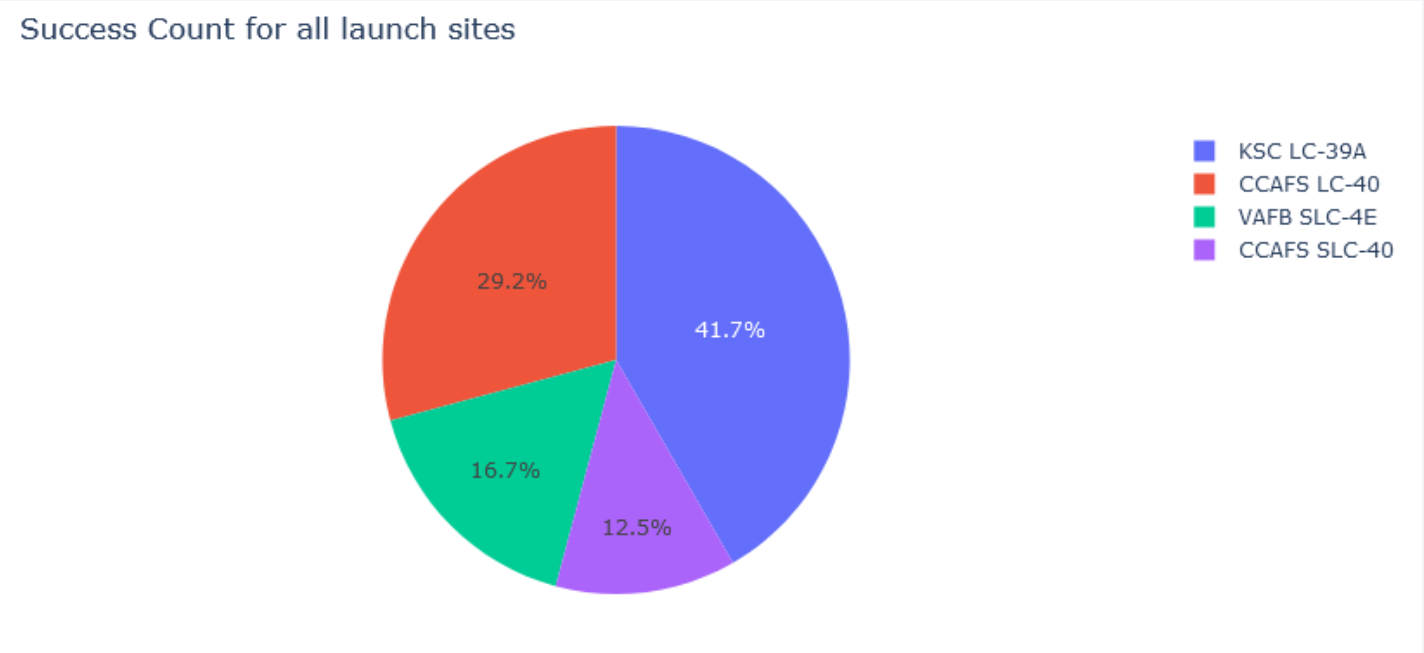
- Launch sites may be located close to transport infrastructure like railroads and highways, however they are located far away from cities

Section 5

Build a Dashboard with Plotly Dash

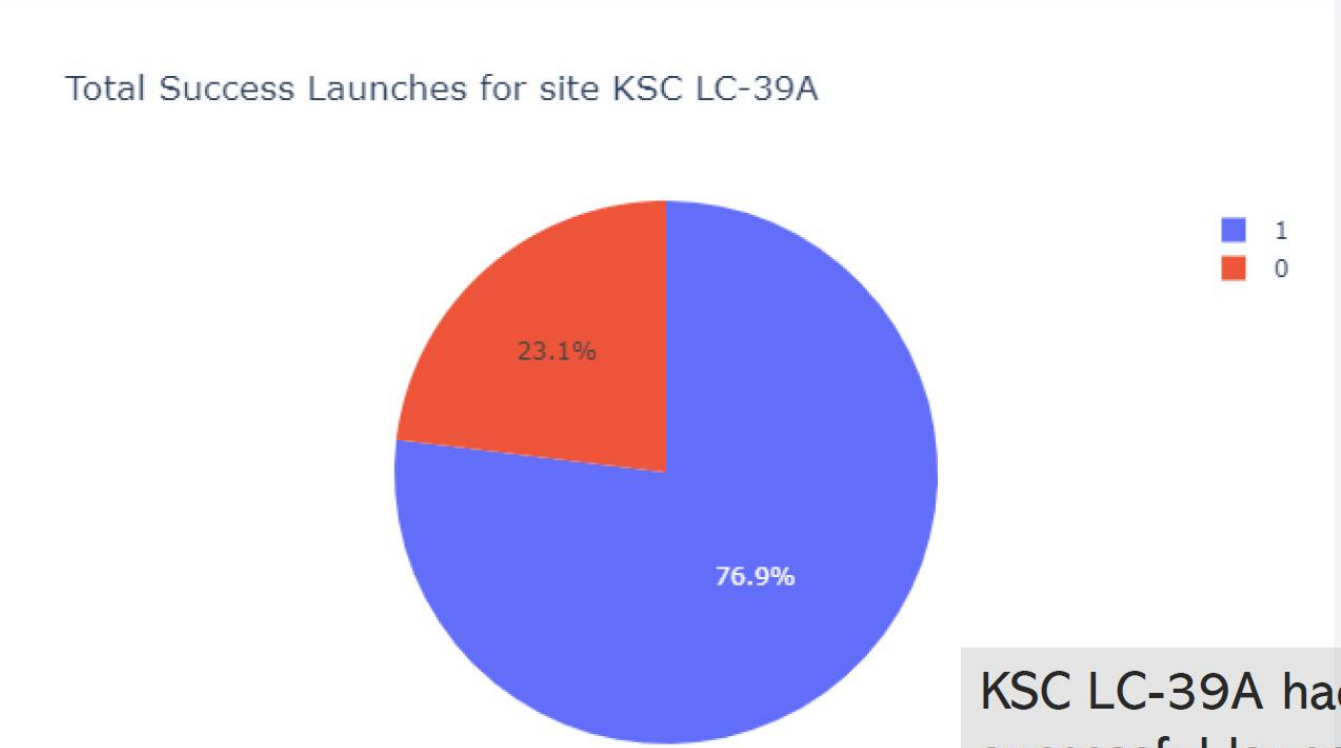


Success rate for all launch sites



- KSC LC-39A has the highest proportion of successes, followed by CCAFS LC-40, VAFB SLC-4E and CCAFS SLC-40

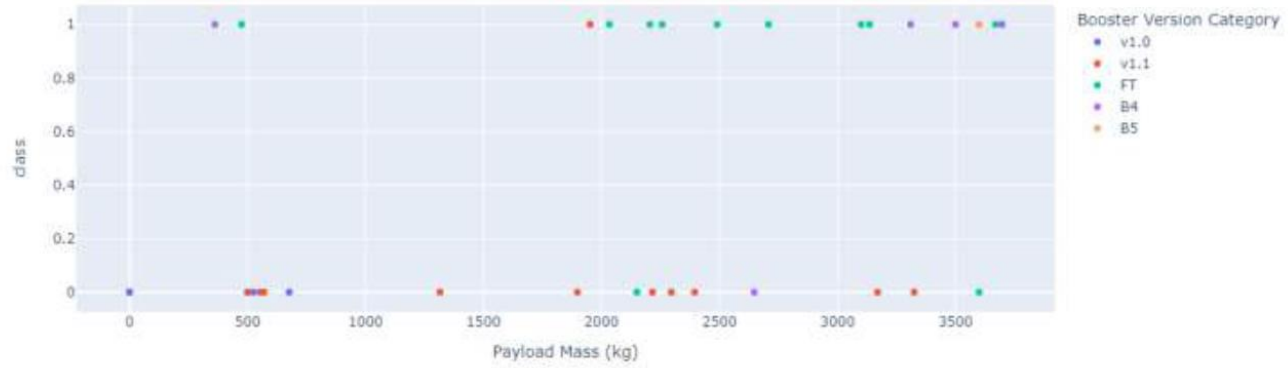
Launch site with highest launch success ratio



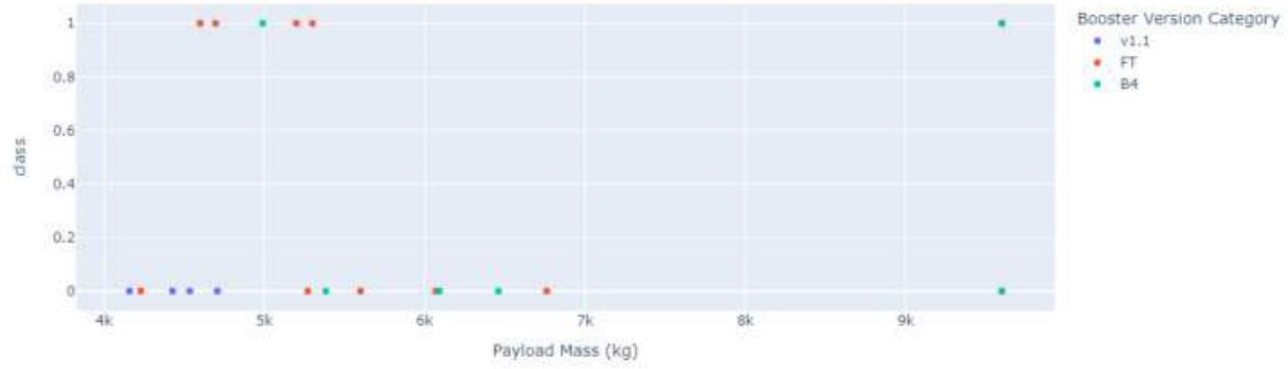
KSC LC-39A had a 76.9% successful launch rate

Payload vs Launch Outcome

Success count on Payload mass (0 to 4000kg) for all sites



Success count on Payload mass (4000kg and above) for all sites



1) Higher success rate at lower Payload mass
2) FT Booster Version seems to have a higher success rate

Section 6

Predictive Analysis (Classification)

Classification Accuracy

- Accuracy of various models as follows:
  - Logistic regression: 0.875
  - SVM: 0.889
  - Tree: 0.861
  - KNN: 0.861
- Hence SVM has the highest accuracy

Confusion Matrix

- The Matrix shows that SVM was able to predict accurately true positive of successful landing, but it still have several false positive for failure to land

Conclusions

- Orbit ES-L1, GEO, HEO and SSO have the best Success Rates
- Overall the success rates of launches increase over the years
- KSC LC-39A had the most successful launches from all launch sites
- Lower weighted payload performed better than heavier ones
- SVM Classifier Algorithm is best suited for Machine Learning of this data set

Thank you!