Winning Space Race with Data Science

Liangqu Chen
10-22-2021

Outline

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

Executive Summary

- Summary of methodologies
  - ✓ HTTP Requests
  - ✓ Pandas and numpy for data manipulation and analysis
  - ✓ Beautifulsoup for data scraping
  - ✓ SQL queries
  - ✓ Exploratory data analysis
  - ✓ Data visualization with matplotlib, seaborn, Folium
  - ✓ Machine learning classification technics like SVM, classification trees, logistic regression
- Summary of all results
  - Accurate score of prediction for falcon 9 launch outcome is 0.833
  - Explored and show some correlations between some variables and the launch outcome by data visualization

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

  1. To predict SpaceX Falcon 9 launch outcomes

  2. To explore correlations between variables and launching outcomes

Section 1

Methodology

Methodology

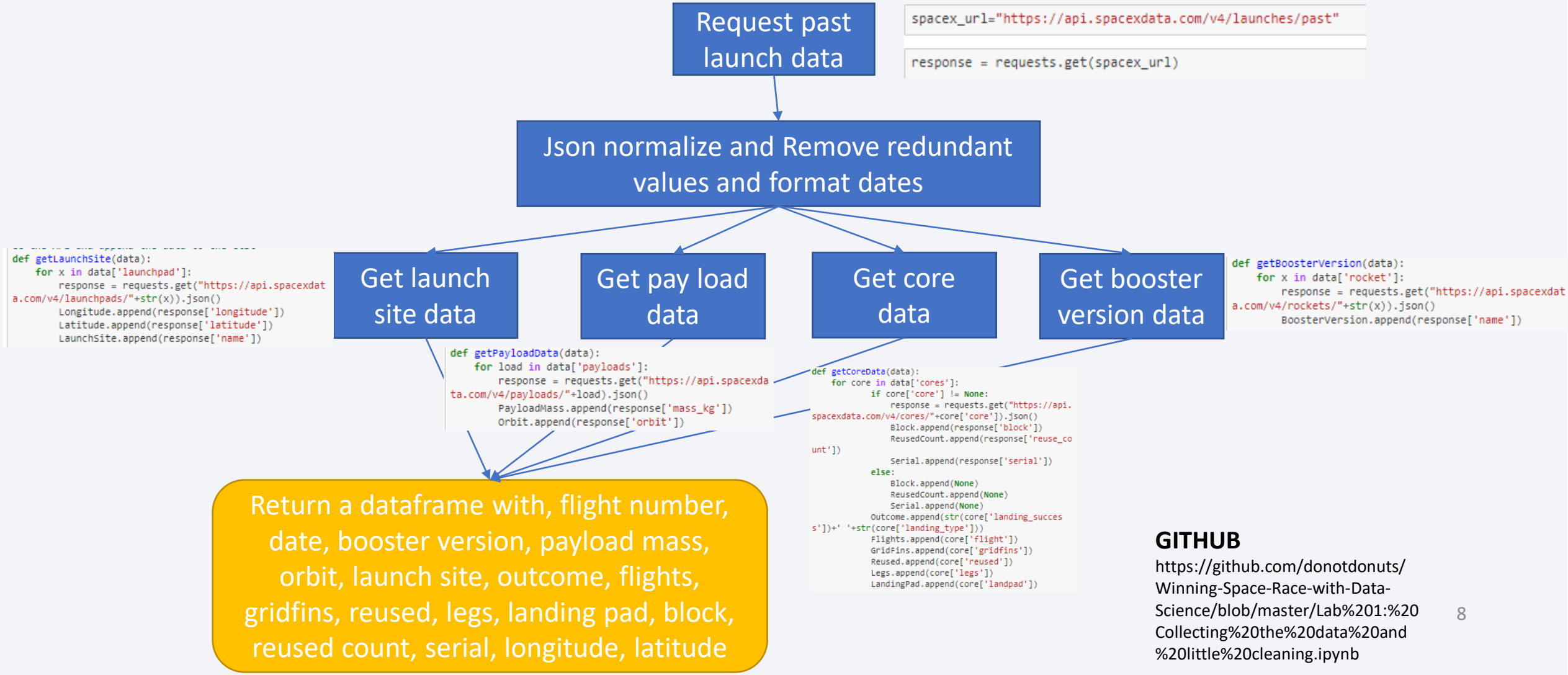
Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- 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

I collected data by requesting SpaceX data from SpaceX API and I also scraped Falcon 9 launch records in HTML table from Wikipedia and parsed the table by Beautifulsoup and convert it into Pandas data frames

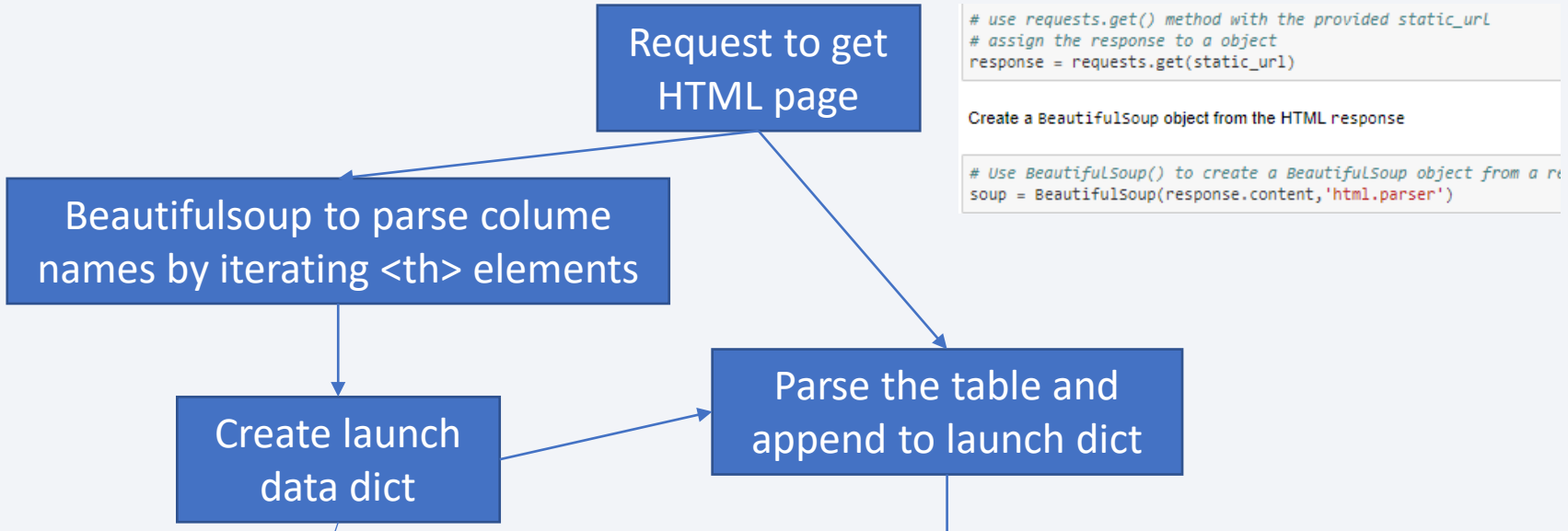
Data Collection – SpaceX API



GITHUB

https://github.com/donotdonuts/Winning-Space-Race-with-Data-Science/blob/master/Lab%201:%20Collecting%20the%20data%20and%20little%20cleaning.ipynb

Data Collection - Scraping




```
# use requests.get() method with the provided static_url
# assign the response to a object
response = requests.get(static_url)
```

Create a BeautifulSoup object from the HTML response

```
# Use BeautifulSoup() to create a BeautifulSoup object from a re
soup = BeautifulSoup(response.content,'html.parser')
```

```
launch_dict= dict.fromkeys(column_names)

# Remove an irrelvant 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']=[]
```

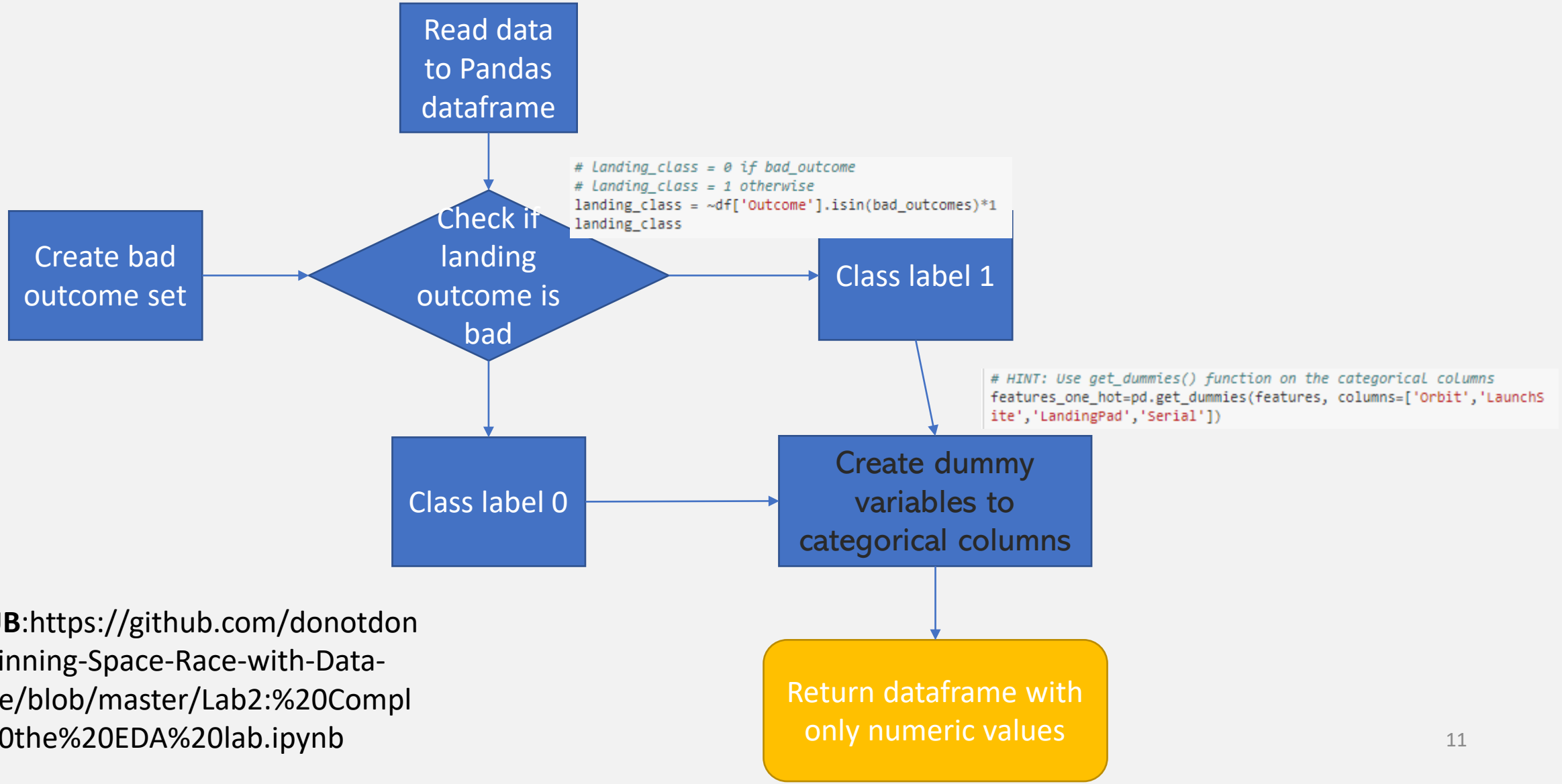
GITHUB:
https://github.com/donotdonuts/Winning-Space-Race-with-Data-Science/blob/master/Lab2:%20the%20Data%20Collection%20with%20Web%20Scraping.ipynb

```
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()
```

Data Wrangling

- In order to obtain the right independent variables, I cleaned up the attributes like cores and payloads as well as date format, while dropping nulls and replacing missing values with averages. Since the project is focused on Falcon 9, I removed any other launches other than Falcon 9.
- Label good outcomes to 1 and bad outcomes like False ASDS, False Ocean etc to 0. Create dummy variables to categorical columns so that is friendly for later machine learning process

Data Wrangling



GITHUB:https://github.com/donotdonuts/Winning-Space-Race-with-Data-Science/blob/master/Lab2:%20Complete%20the%20EDA%20lab.ipynb

EDA with Data Visualization

- **Scatter chart, Launch site vs flight number is plotted with outcomes marked by different colors**. It shows more launches have been done in CCAFS SLC 40 and it seems that VAFB SLC 4E has done the least launch but it has high successful rate as well as KSC LC 39 A.
- **Scatter chart, Pay load mass vs Launch site plotted with outcomes marked by different colors**. It shows more launches have been done in CCAFS SLC 40 and more launches have been done under 8000 KG in general, and it seems that higher the payload mass, high the success rate.
- **Bar chart, Orbit VS Outcome.** ES-L, GEO, HEO, SSO have 100% successful rate and GTO is the lowest.
- **Scatter chart, Orbit VS Flight Number plotted with outcomes in two colors.** In 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.
- **Scatter, Payloadmass vs Orbits with outcomes in two colors.** With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS
- **Line chart, Year vs outcomes.** Success rate raises from year 2013 to 2020

GITHUB: https://github.com/donotdonuts/Winning-Space-Race-with-Data-Science/blob/master/Lab4:%20Complete%20the%20EDA%20with%20Visualization%20lab.ipynb

EDA with SQL

- Display the names of the unique launch sites:
- Display 5 records begin with CCA
- Display the total payload mass carried by boosters
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was acheived.
- 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 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: https://github.com/donotdonuts/Winning-Space-Race-with-Data-Science/blob/master/Lab%203:%20Complete%20the%20EDA%20with%20SQL%20lab.ipynb

Build an Interactive Map with Folium

- Added a marker and a circle object to show NASA Johnson Space Center at Houston, Texas
- Added marker and circle objects to show three launch sites. One in California, two in Florida.
- Use marker to mark each launch location with green or red color to show launch outcome and use marker cluster to group markers
- Added a polyline object and marker to show distance line and distance in KM

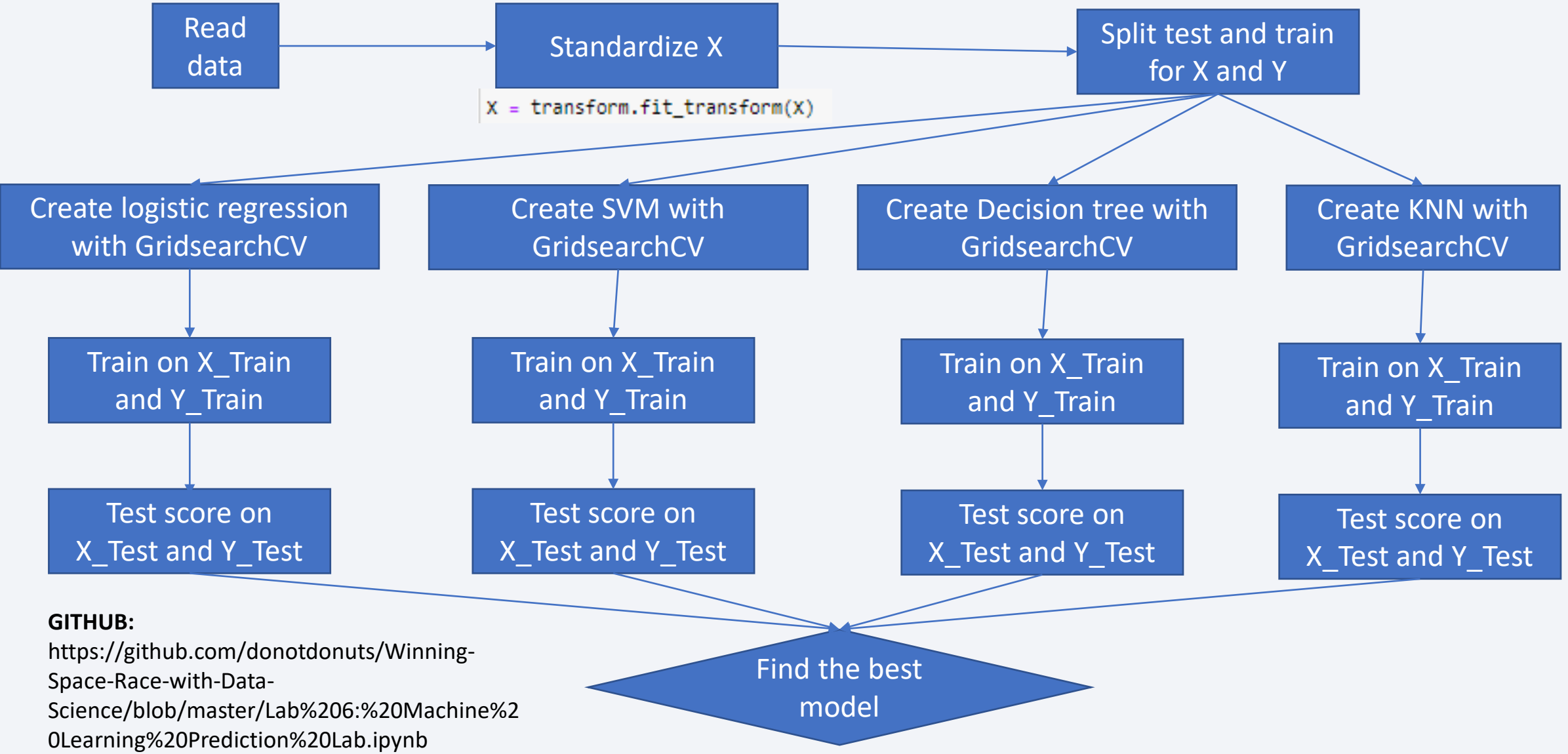
GITHUB: https://github.com/donotdonuts/Winning-Space-Race-with-Data-Science/blob/master/Lab5:%20the%20Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- Added success launch pie chart. Upon the choice of which sites or all sites, it shows the success rate for each site or all sites in a pie chart.
- Also added a scatter chart, Payload Mass vs outcome with different colors showing for different booster version category and this chart can be interacted by changed the slide bar for payload range.

GITHUB: https://github.com/donotdonuts/Winning-Space-Race-with-Data-Science/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)



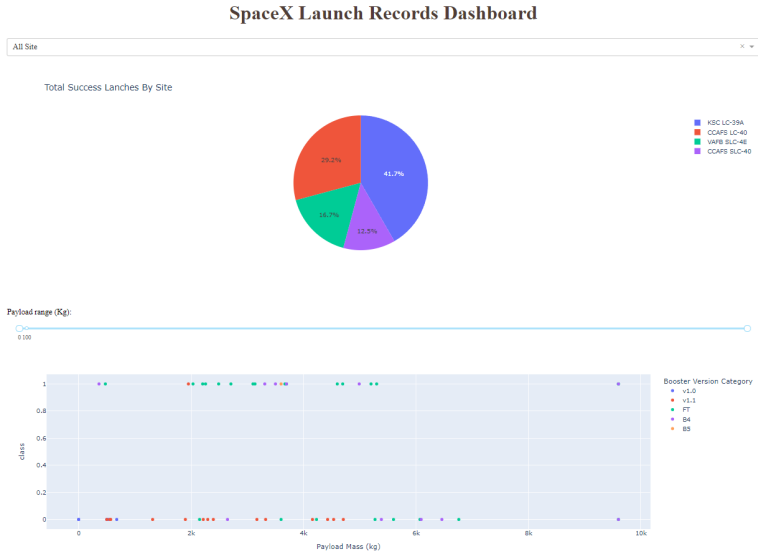
GITHUB:
https://github.com/donotdonuts/Winning-Space-Race-with-Data-Science/blob/master/Lab%206:%20Machine%20Learning%20Prediction%20Lab.ipynb

Results

- Exploratory data analysis results
  - It shows more launches have been done in CCAFS SLC 40 and it seems that VAFB SLC 4E has done the least launch but it has high successful rate as well as KSC LC 39 A.
  - It shows more launches have been done in CCAFS SLC 40 and more launches have been done under 8000 KG in general, and it seems that higher the payload mass, high the success rate.
  - ES-L, GEO, HEO, SSO have 100% successful rate and GTO is the lowest.
  - In the LEO orbit the Success appears related to the number of flights
  - With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS
  - Success rate raises from year 2013 to 2020
  - KSC LC-39A has the highest success counts and CCAFS SLC-40 has the lowest
  - It looks like 2k-4k payload mass ranges has the largest success rate
  - SpaceX usually carrys payload mass less than 6K

Results

- Interactive analytics demo in screenshots

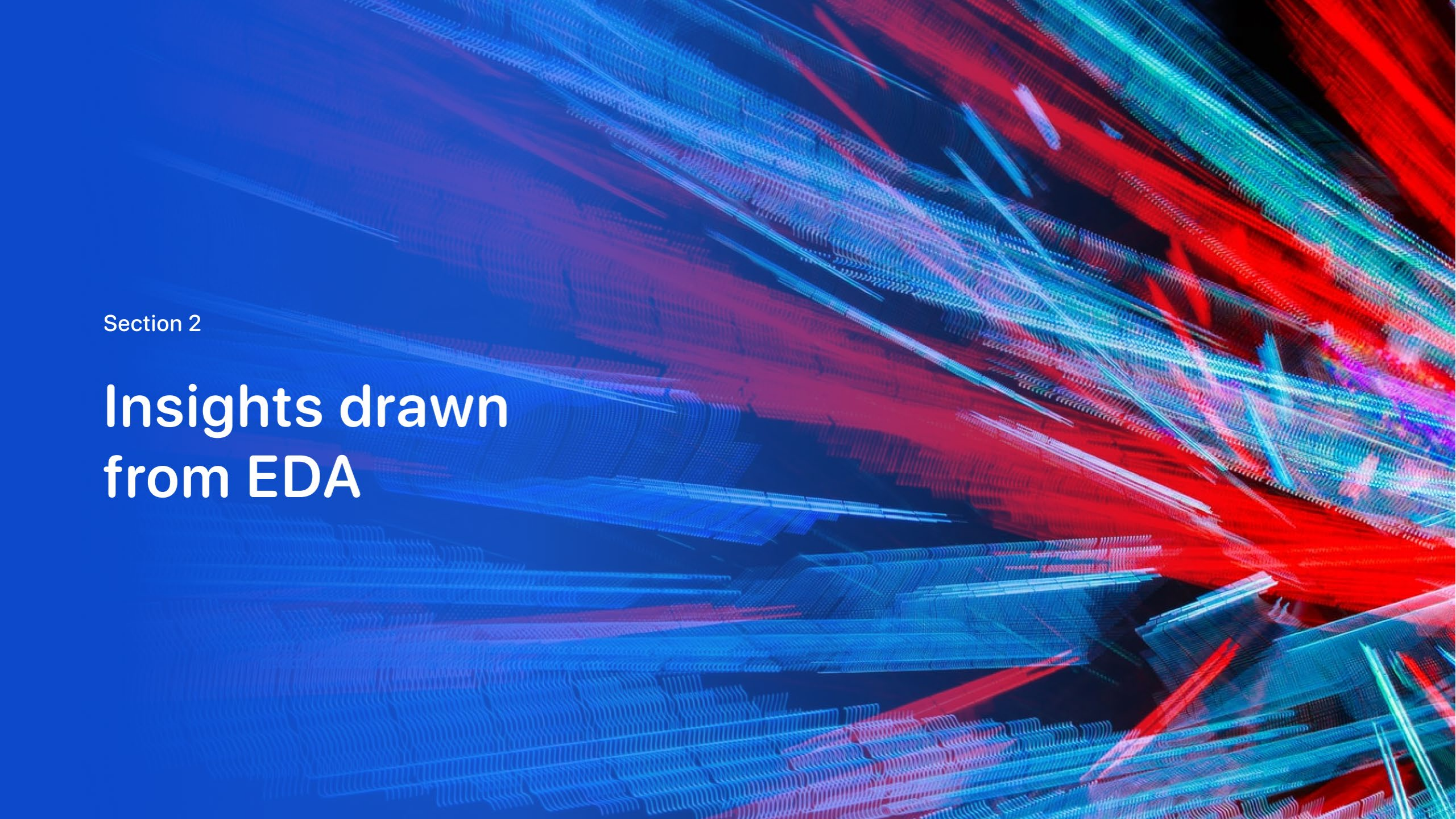


- Predictive analysis results

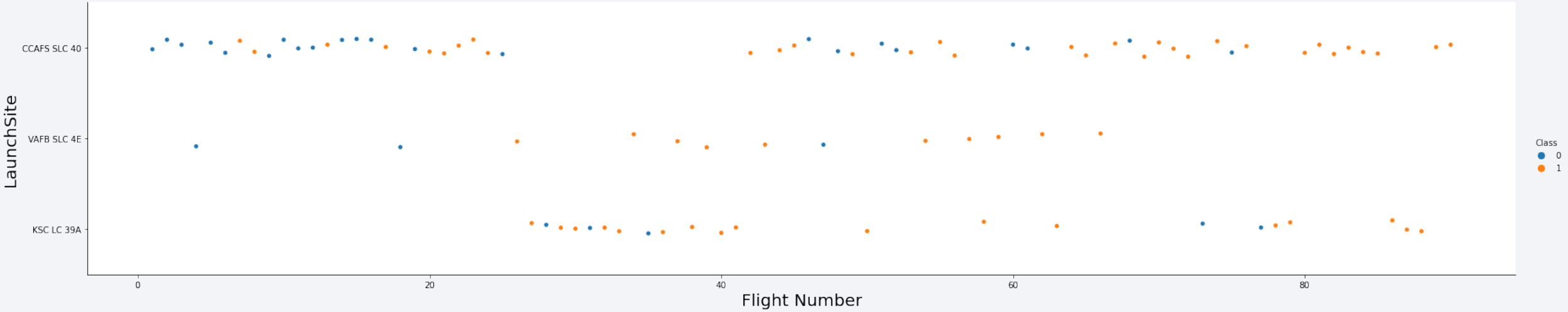
Only problem is false positve

```
print("logistic regression:",logreg_cv.score(X_test, Y_test))
print("SVM regression:",svm_cv.score(X_test, Y_test))
print('Decision tree:', tree_cv.score(X_test,Y_test))
print('knn:', knn_cv.score(X_test, Y_test))
```

```
logistic regression: 0.8333333333333334
SVM regression: 0.8333333333333334
Decision tree: 0.7777777777777778
knn: 0.8333333333333334
```

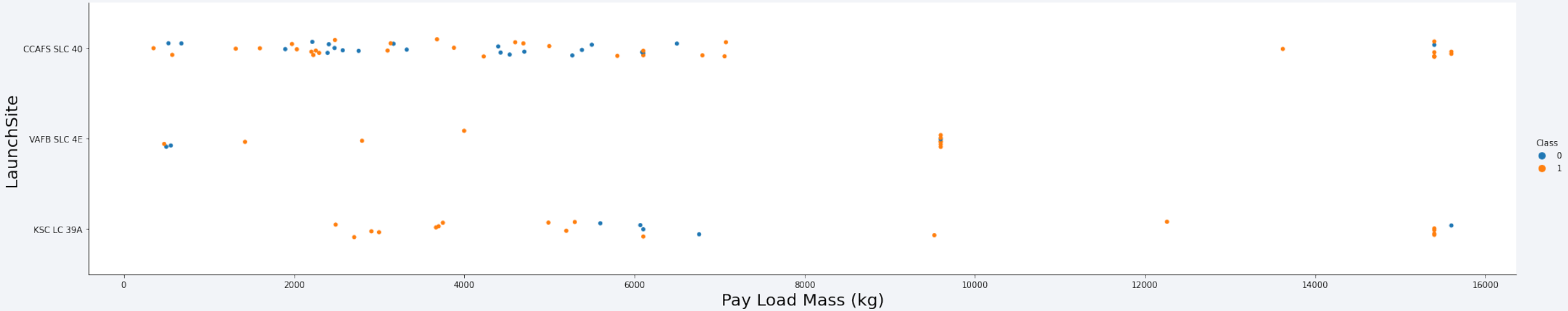

Section 2
Insights drawn
from EDA

Flight Number vs. Launch Site



- It shows more launches have been done in CCAFS SLC 40 and it seems that VAFB SLC 4E has done the least launch but it has high successful rate as well as KSC LC 39 A.

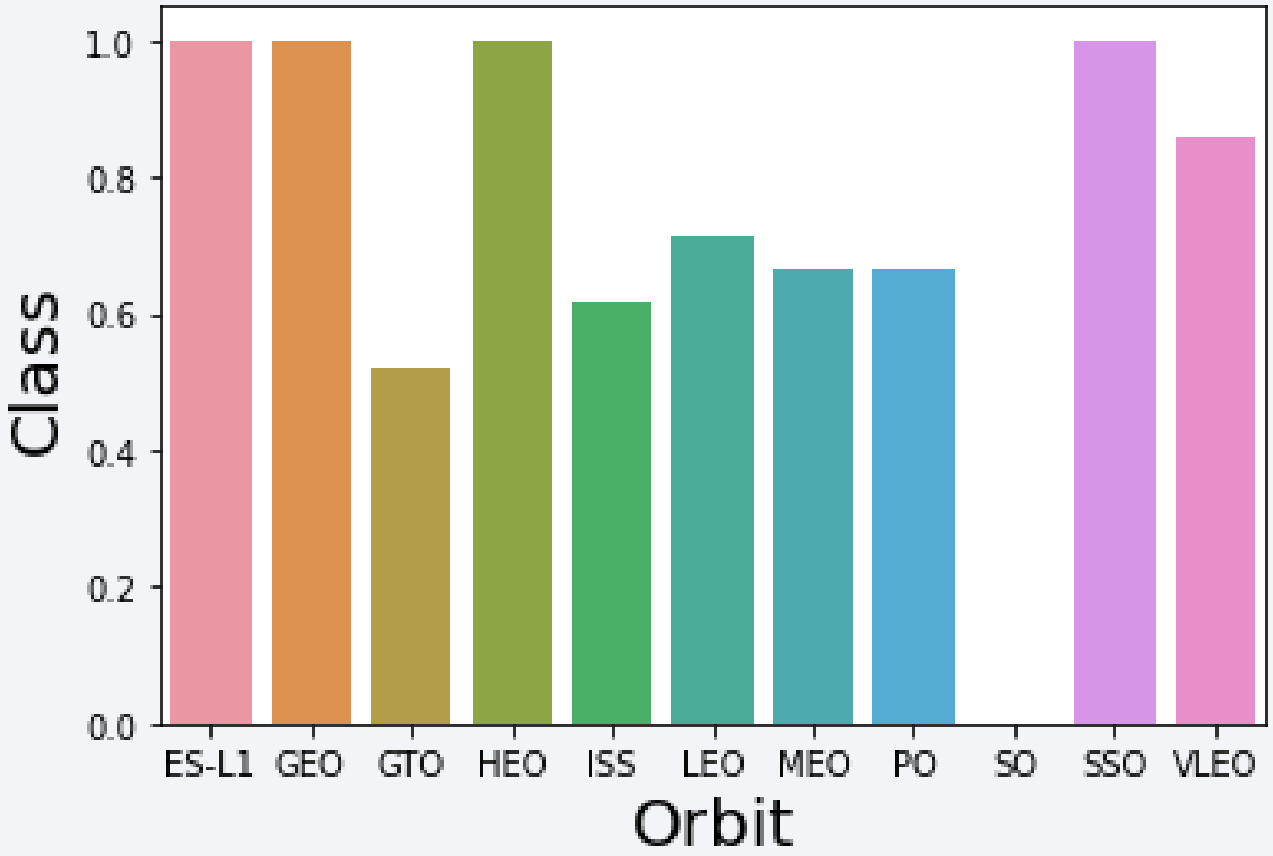
Payload vs. Launch Site



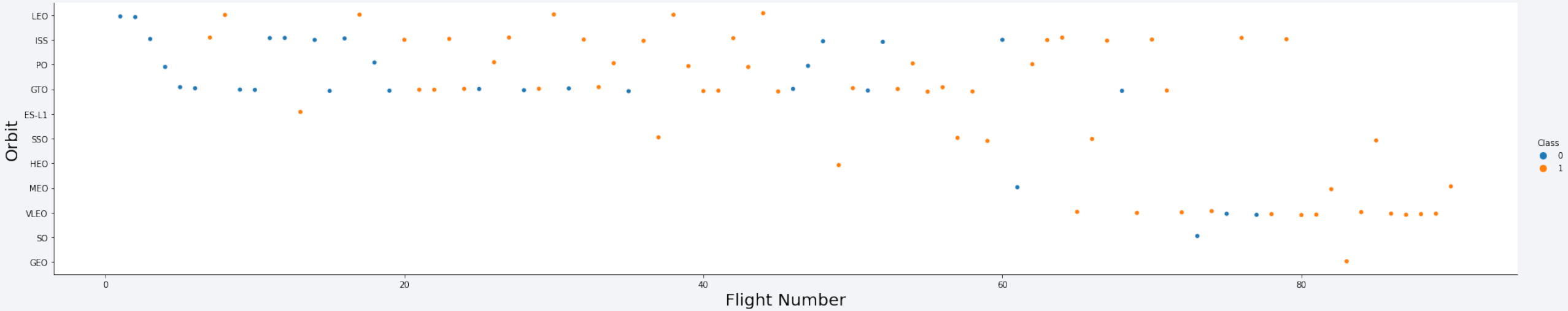
It shows more launches have been done in CCAFS SLC 40 and more launches have been done under 8000 KG in general, and it seems that higher the payload mass, high the success rate.

Success Rate vs. Orbit Type

- ES-L, GEO, HEO, SSO have 100% successful rate and GTO is the lowest.

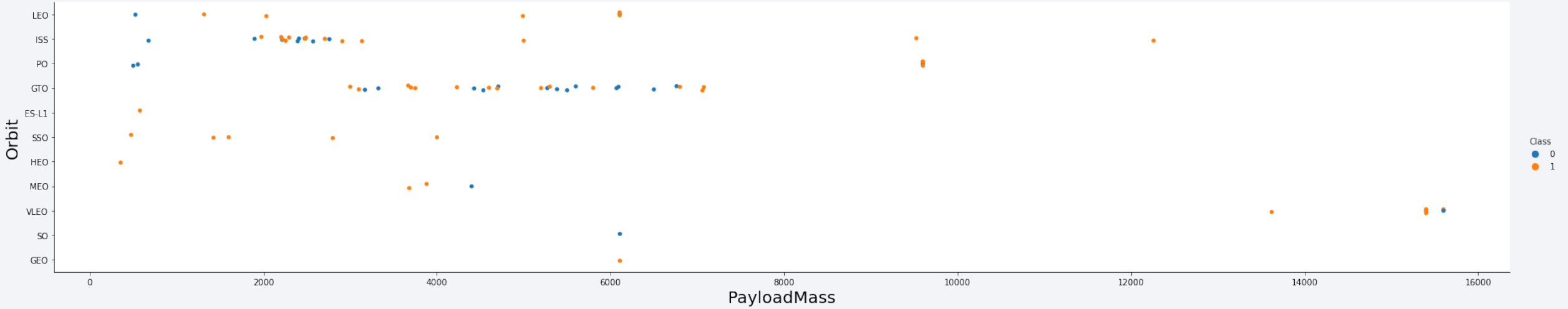


Flight Number vs. Orbit Type



In 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.

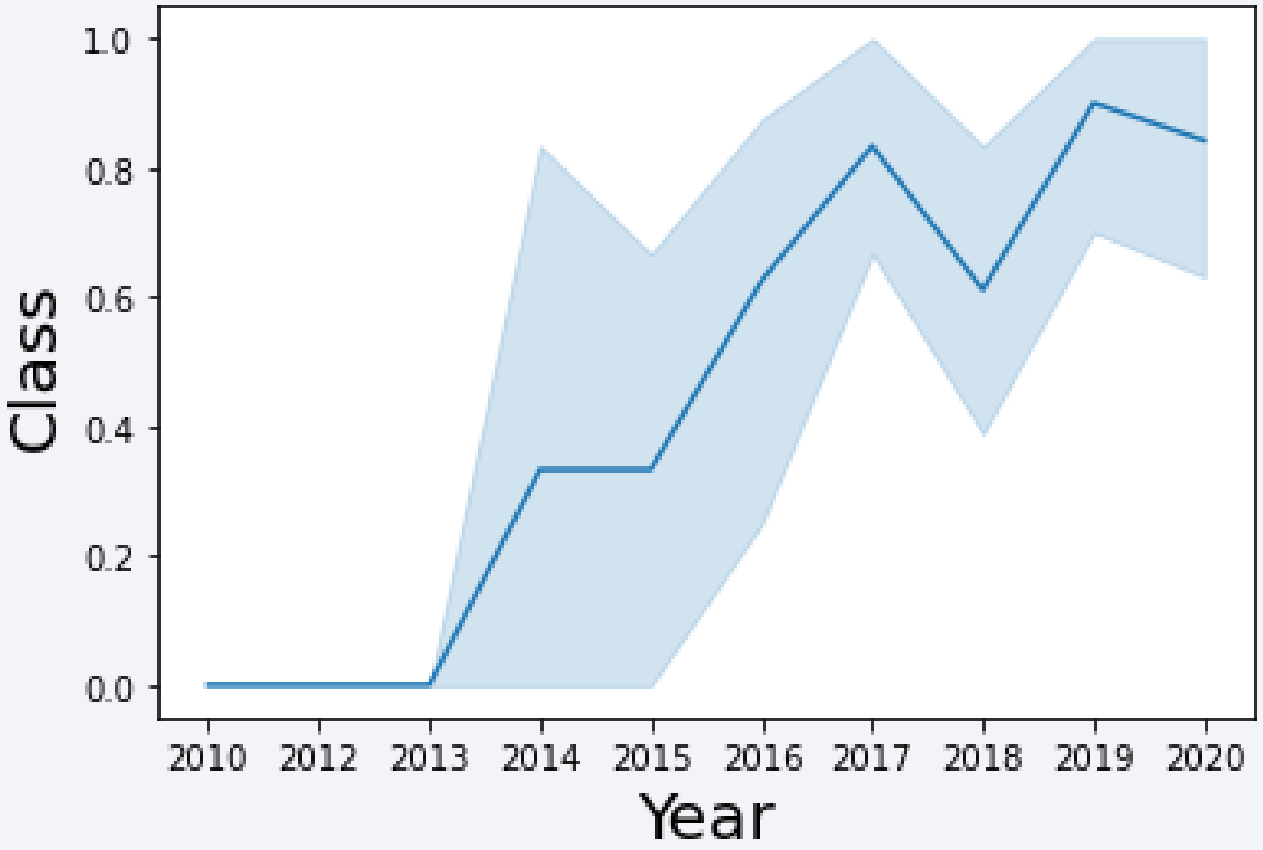
Payload vs. Orbit Type



With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS

Launch Success Yearly Trend

- Success rate raises from year 2013 to 2020



All Launch Site Names

- the names of the unique launch sites
- By using Distinct function, it shows one unique ones

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 SPAXEXTBL\
where LAUNCH_SITE like 'CCA%'\
fetch first 5 rows only;
```

```
 * ibm_db_sa://dgm77831:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/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

- Calculate the total payload carried by boosters from NASA

```
%sql Select sum(payload_mass__kg_) from SPAXEXTBL\
     where customer = 'NASA (CRS)'
```

```
 * ibm_db_sa://dgm77831:***@fbd88901-ebdb-4a4f-a32e-
Done.
```

1
45596

Average Payload Mass by F9 v1.1

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

```
%sql select avg(payload_mass__kg_) from SPAXEXTBL\
     where booster_version like 'F9 v1.1%'
```

```
 * ibm_db_sa://dgm77831:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237
Done.
```

1
2534

First Successful Ground Landing Date

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

```
%sql select min(Date) from SPAXEXTBL \
     where landing__outcome = 'Success (ground pad)'
```

```
 * ibm_db_sa://dgm77831:***@fbd88901-ebdb-4a4f-a32e-9822
Done.
```

1
2015-12-22

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 from SPACEXTBL \
     where landing__outcome = 'Success (drone ship)' \
     and (payload_mass__kg_ between 4000 and 6000)
```

```
 * ibm_db_sa://dgm77831:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.
Done.
```

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

```
%sql select mission_outcome, count(mission_outcome) from SPAXEXTBL \
     group by mission_outcome
```

```
 * ibm_db_sa://dgm77831:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0t
Done.
```

mission_outcome	2
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

```
%sql select Booster_Version, payload_mass__kg_ from SPAXEXTBL \
     where payload_mass__kg_ = (select max(payload_mass__kg_) from SPAXEXTBL)
```

```
 * ibm_db_sa://dgm77831:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0l
Done.
```

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

2015 Launch Records

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

```
%sql select Date, landing__outcome, Booster_Version, Launch_Site from SPAXEXTBL \
     where landing__outcome = 'Failure (drone ship)' and year(Date) = '2015'
```

```
 * ibm_db_sa://dgm77831:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.da
Done.
```

DATE	landing__outcome	booster_version	launch_site
2015-01-10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	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

```
%sql select landing__outcome, count(landing__outcome) from SPAXEXTBL \
     where Date between '2010-06-04' and '2017-03-20' \
     group by landing__outcome \
     order by count(landing__outcome) desc
```

```
 * ibm_db_sa://dgm77831:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0t
Done.
```

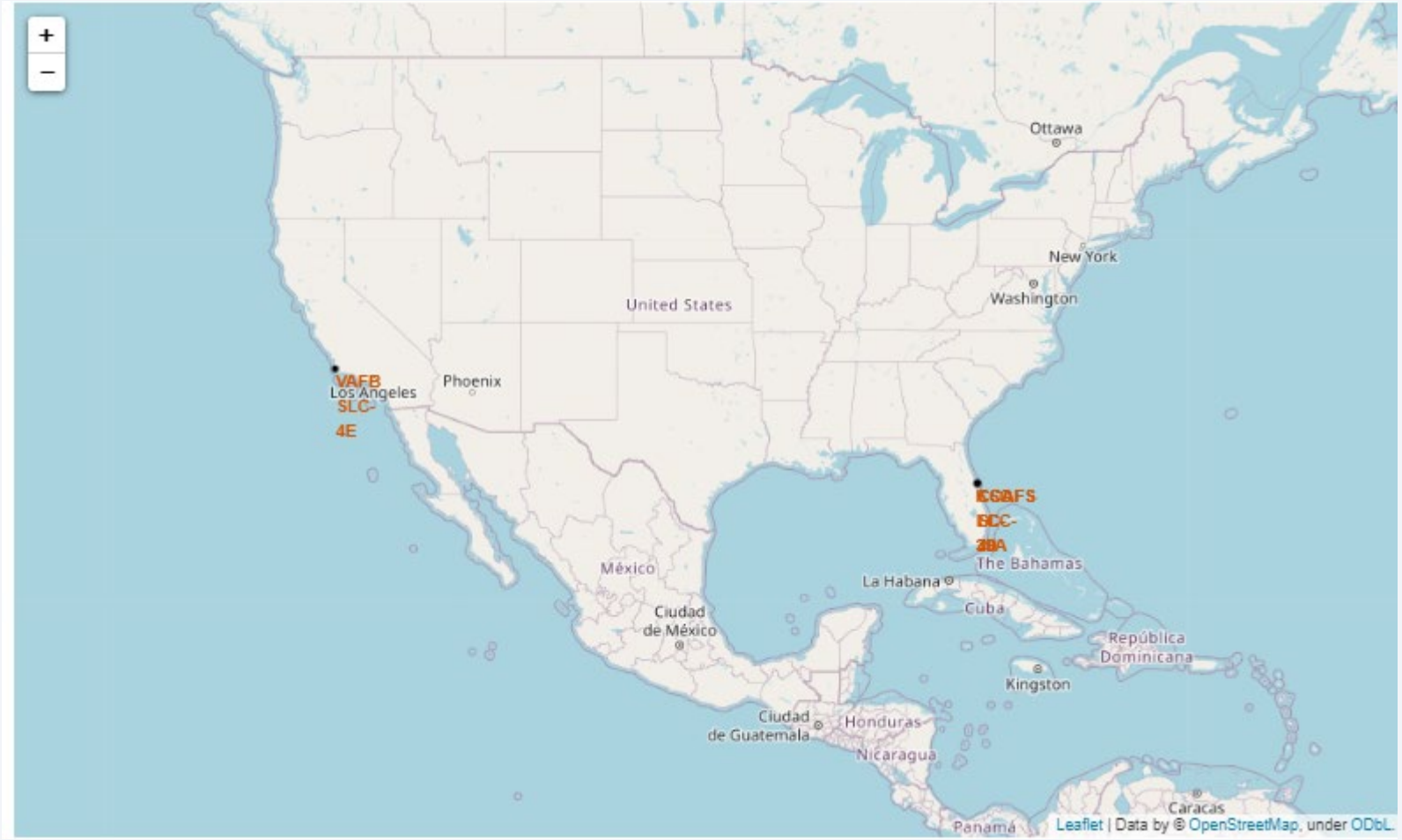
landing__outcome	2
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 4

Launch Sites Proximities Analysis

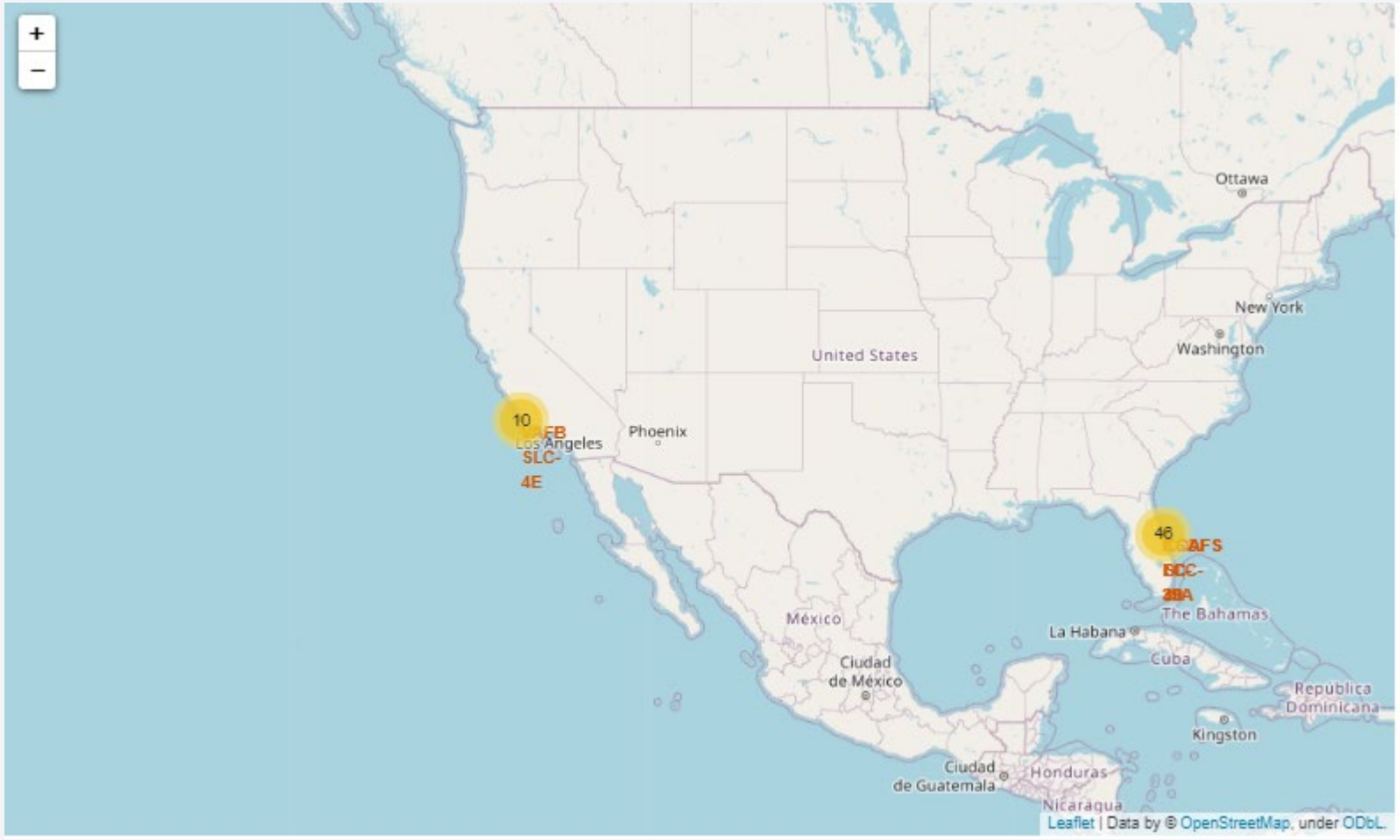
Launch sites on map

- Added marker and circle objects to show three launch sites. One in California, two in Florida.



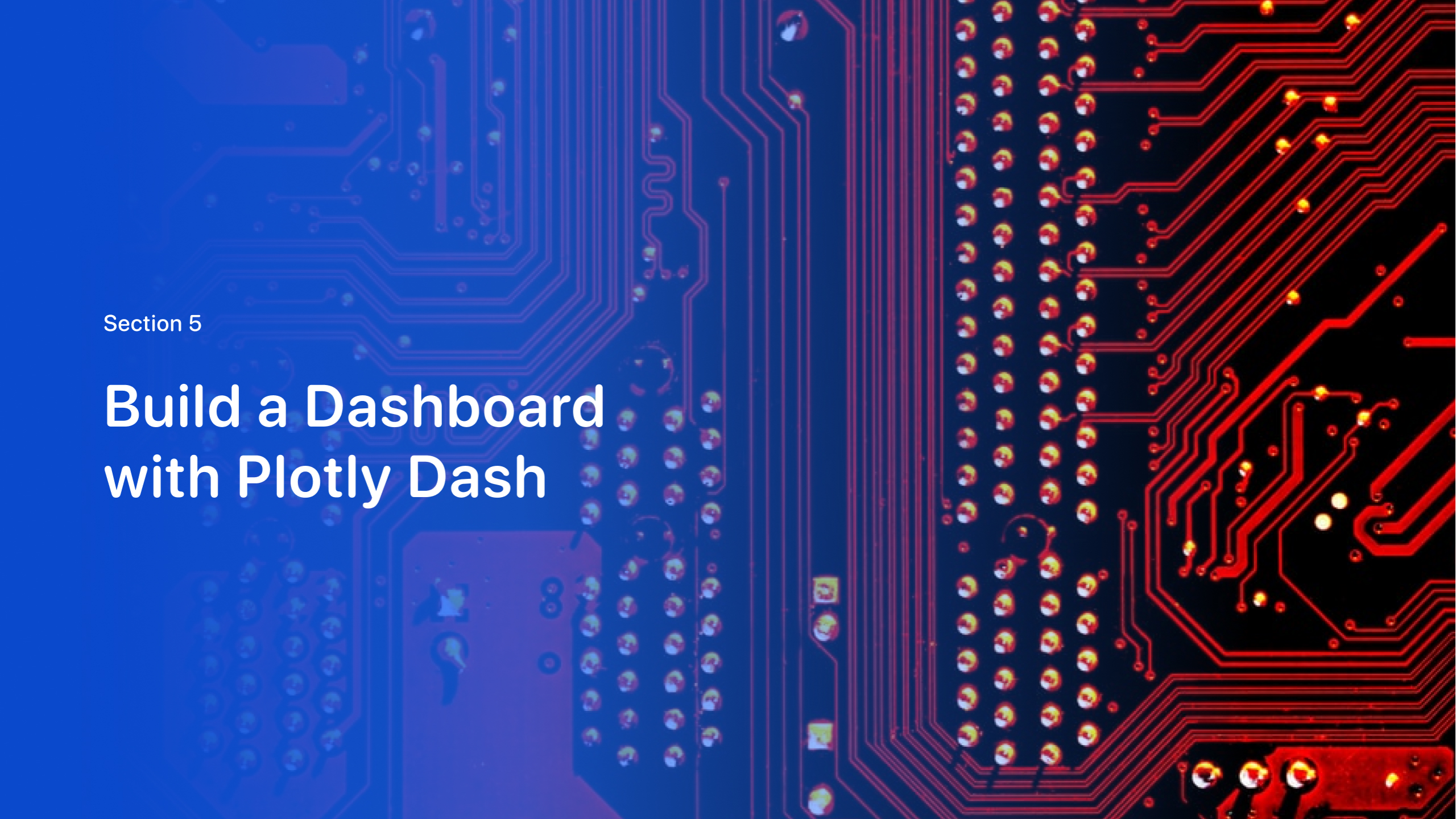
Mark success and fail launches for each site

- Use marker to mark each launch location with green or red color to show launch outcome and use marker cluster to group markers



Distance line to coastline

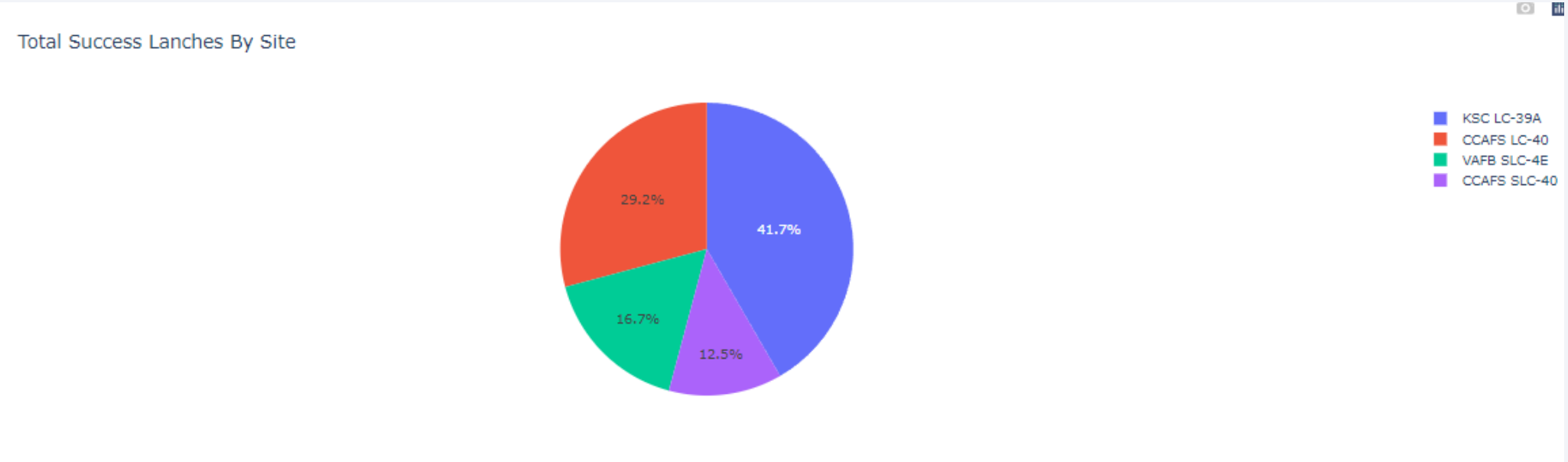
- Added a polyline object and marker to show distance line and distance in KM
- It shows distance between coastline and a launch center



Section 5

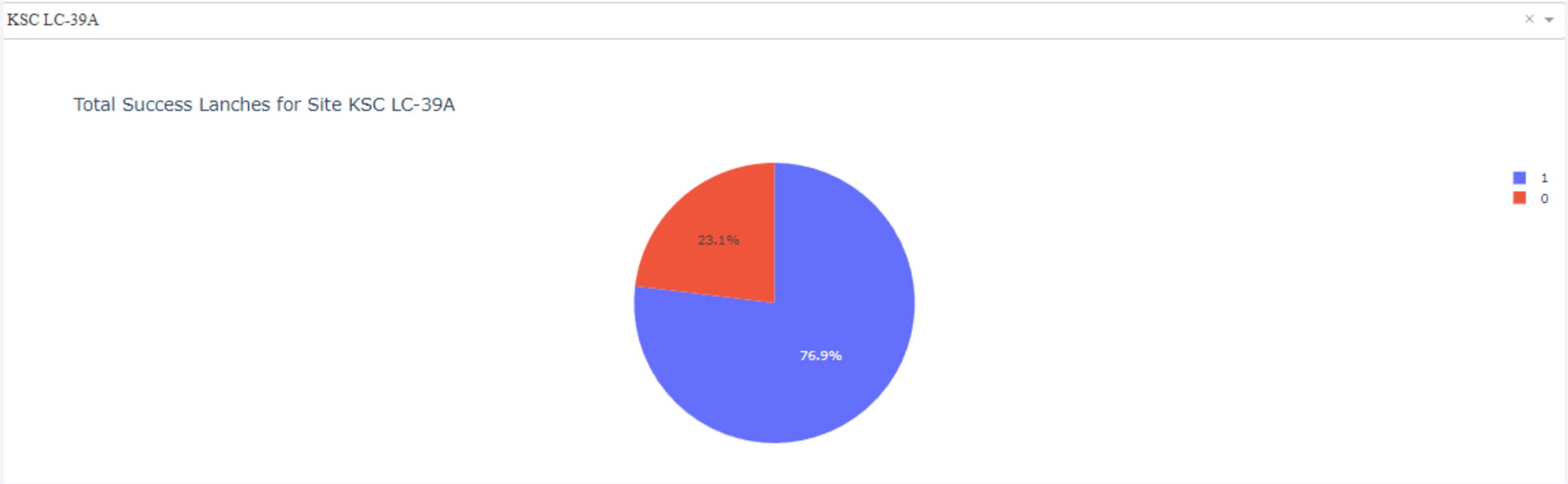
Build a Dashboard with Plotly Dash

Launch success count for all sites



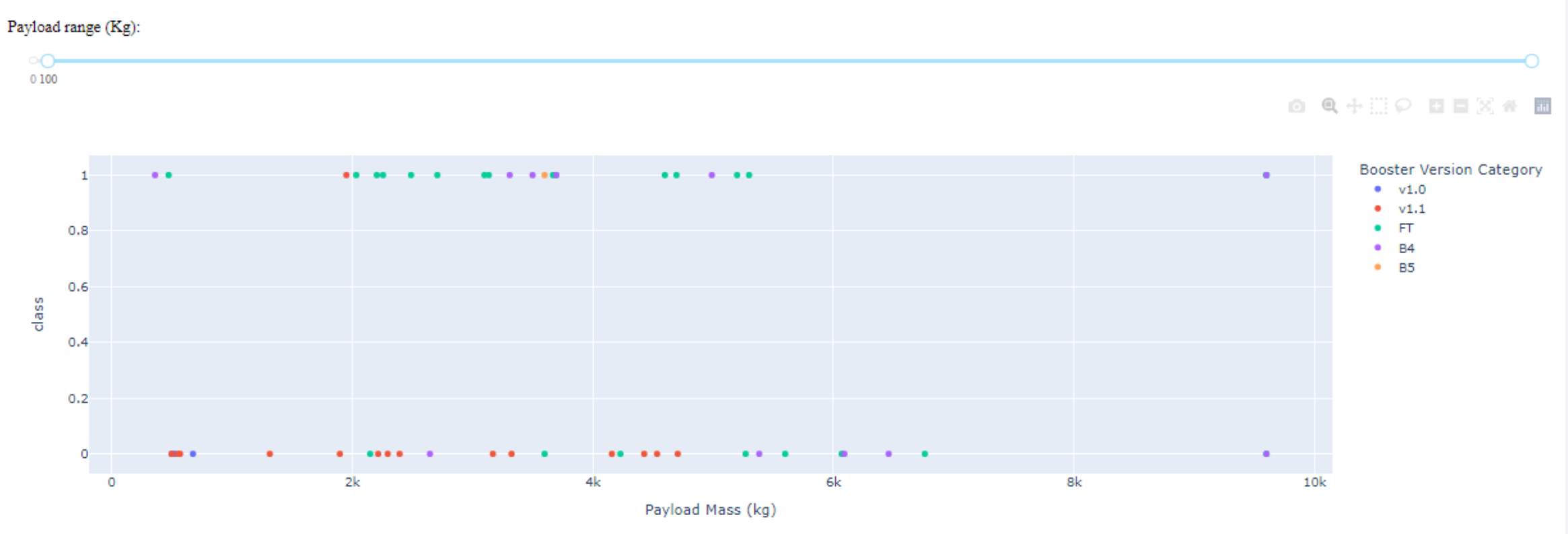
- KSC LC-39A has the highest success counts and CCAFS SLC-40 has the lowest

Launch site with highest launch success ratio



- KSC LC -39A has the highest launch success ratio—76.9%

Payload vs. Launch Outcome scatter plot for all sites



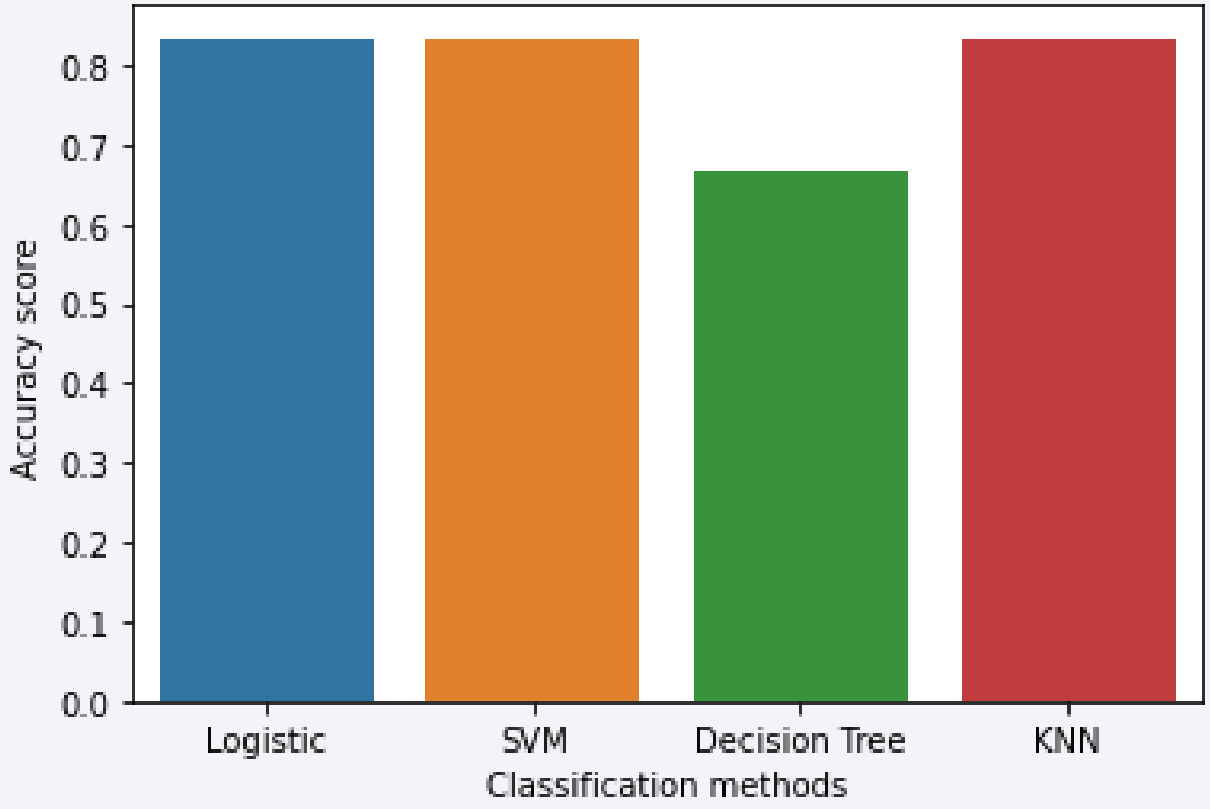
- It looks like 2k-4k payload mass ranges has the largest success rate
- SpaceX usually carrys payload mass less than 6K

Section 6

Predictive Analysis (Classification)

Classification Accuracy

- Decision tree is the lowest
- All others has accuracy score of 0.8333



Confusion Matrix

- Only concern in the matrix is the 3 predicted landed label but actually not landed

Conclusions

- It shows more launches have been done in CCAFS SLC 40 and it seems that VAFB SLC 4E has done the least launch but it has high successful rate as well as KSC LC 39 A.
- It shows more launches have been done in CCAFS SLC 40 and more launches have been done under 8000 KG in general, and it seems that higher the payload mass, high the success rate.
- ES-L, GEO, HEO, SSO have 100% successful rate and GTO is the lowest.
- In the LEO orbit the Success appears related to the number of flights
- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS
- Success rate raises from year 2013 to 2020
- KSC LC-39A has the highest success counts and CCAFS SLC-40 has the lowest
- It looks like 2k-4k payload mass ranges has the largest success rate
- SpaceX usually carrys payload mass less than 6K
- Decision tree model shows very high accuracy in training set but the lowest in test. It might be overfit. SVM, Logistic and KNN model predict better.

Appendix

- All codes and data set can be found in github link below:

  https://github.com/donotdonuts/Winning-Space-Race-with-Data-Science/tree/master

Thank you!