Winning Space Race with Data Science

<Hua Yang>
<10/15/2022>

Outline

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

Executive Summary

- Methodologies: Use flowchart explain all progress
- Results: use screenshot shows all results

Introduction

- Background:
  - Space X 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 Space X 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 space X for a rocket launch.
- Context:
  - Use Python collect data, use SQL and Python do data wrangling, use Python do data dash, and use Python do data visualization. Finally, use classification models and Python program perform predictive analysis
- Solve Problems:
  - SpaceX still have invest value?
  - Do we can copy another SpaceX?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data collect come from two resources. One is use Python requests get request to the SpaceX API: https://api.spacexdata.com/v4/
  - Another is use Python BeautifulSoup do Webscraping from Wikipedia(List of Falcon 9 and Falcon Heavy launches): https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- Perform data wrangling
  - I use pandas and numpy exploratory data analysis(EDA) 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
  - I use pandas, numpy, matplotlib, seaborn, sklearn create a column for the class, standardize the data, split into training and test data, then find the method performs

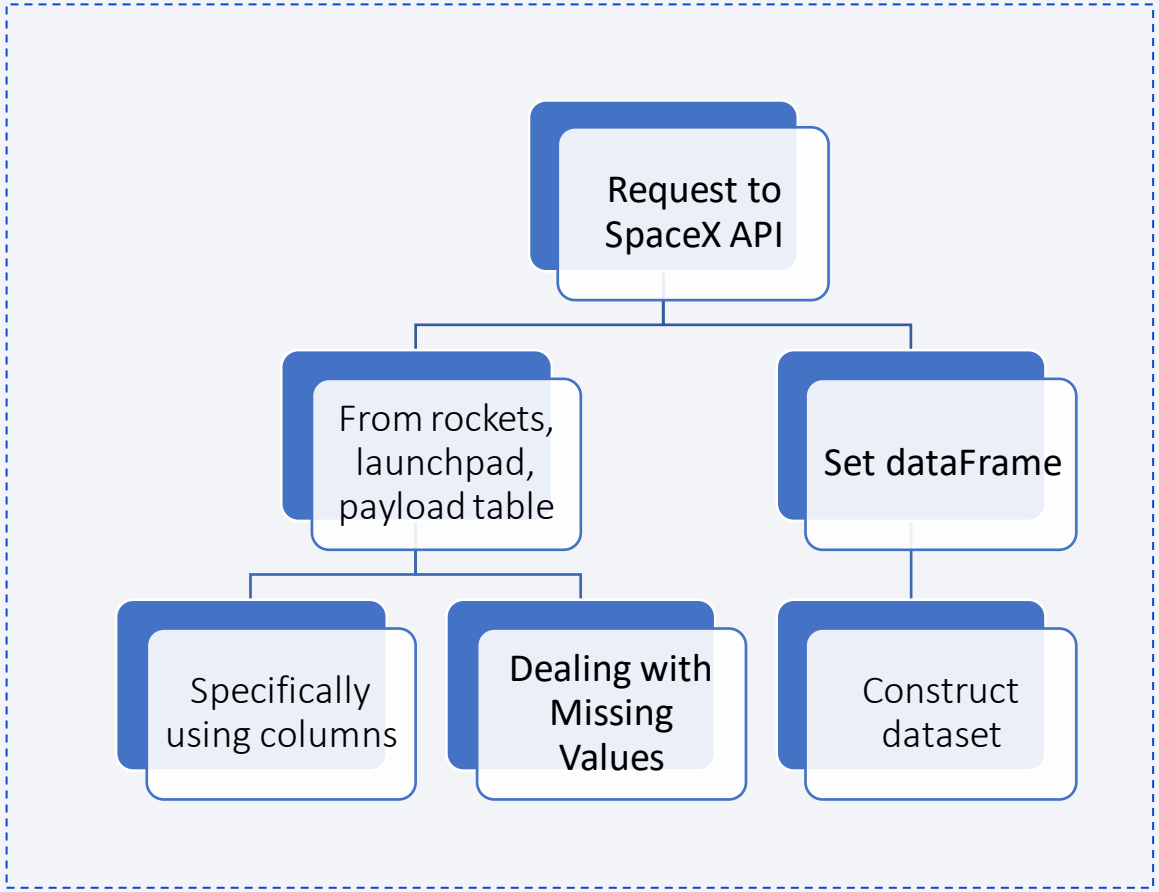
Data Collection

- There have two methods come from Data Collection:
  - Collect Data from SpaceX API
  - Collect Data from Wikipedia

Data Collection – SpaceX API

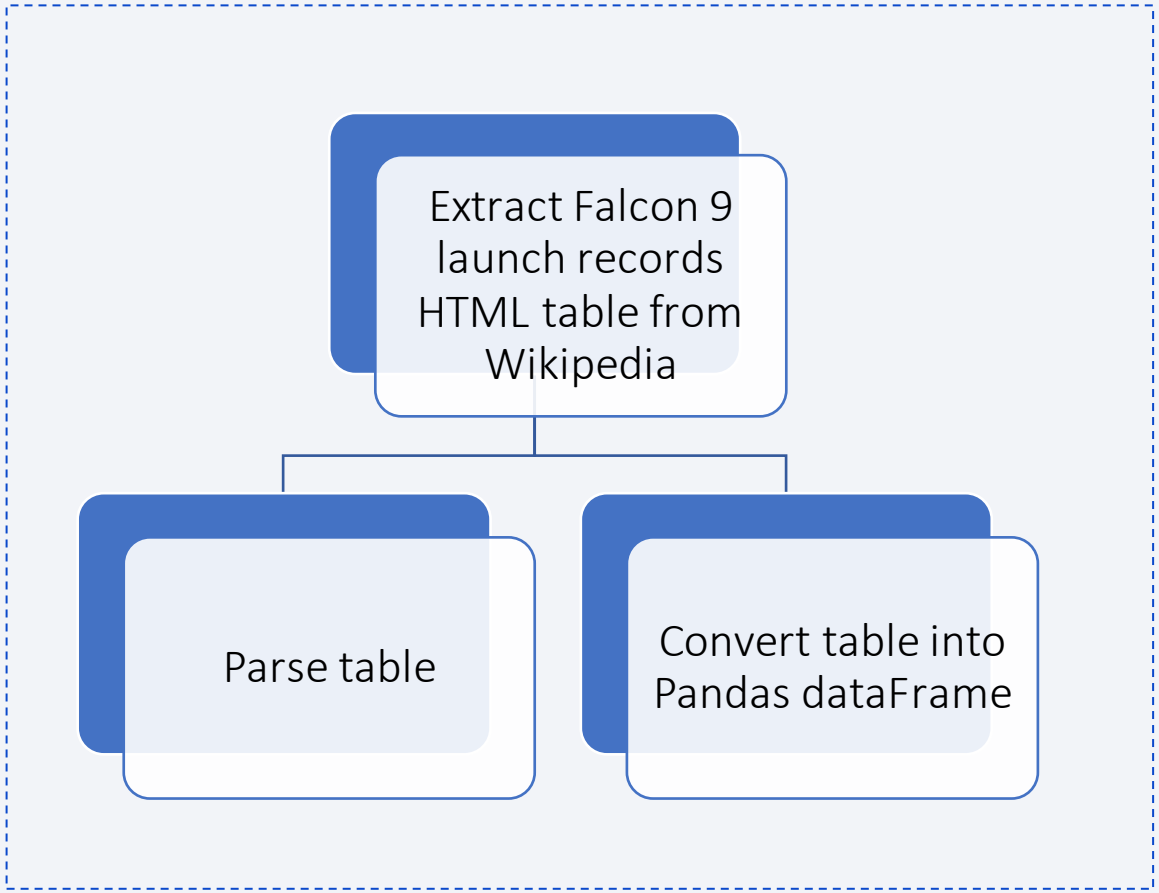
- Use requests request to the SpaceX API and use pandas, numpy and datetime Clean the requested data

- Hua Yang's GitHub: https://github.com/shasha920/SpaceXFalcon9firststageLandingPredictionPython/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

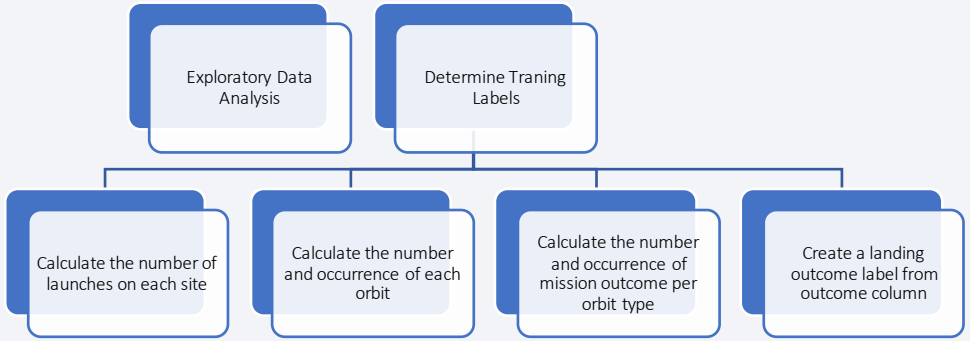
- Use Python BeautifulSoup do Webscraping from Wikipedia(List of Falcon 9 and Falcon Heavy launches): https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- Hua Yang's GitHub: https://github.com/shasha920/SpaceXFalcon9firststageLandingPredictionPython/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Use pandas and numpy do Data Wrangling

- Hua Yang's GitHub: https://github.com/shasha920/SpaceXFalcon9firststageLandingPredictionPython/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- There use 3 methods chart do Data Visualization, included scatter point chart(shows two different value's relationship), bar chart(shows success rate of each orbit), line chart(shows average success rate per year)
- Hua Yang's GitHub: https://github.com/shasha920/SpaceXFalcon9firststageLandingPredictionPython/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- Use SQL queries
  - Display the names of 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 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.
  - List the records which will display the month names, failure landing_outcomes in drone ship, booster versions, launch_site for the months in year 2015.
- Hua Yang's GitHub: https://github.com/shasha920/SpaceXFalcon9firststageLandingPrediction Python/blob/main/jupyter-labs-eda-sql_sqllite.ipynb

Build an Interactive Map with Folium

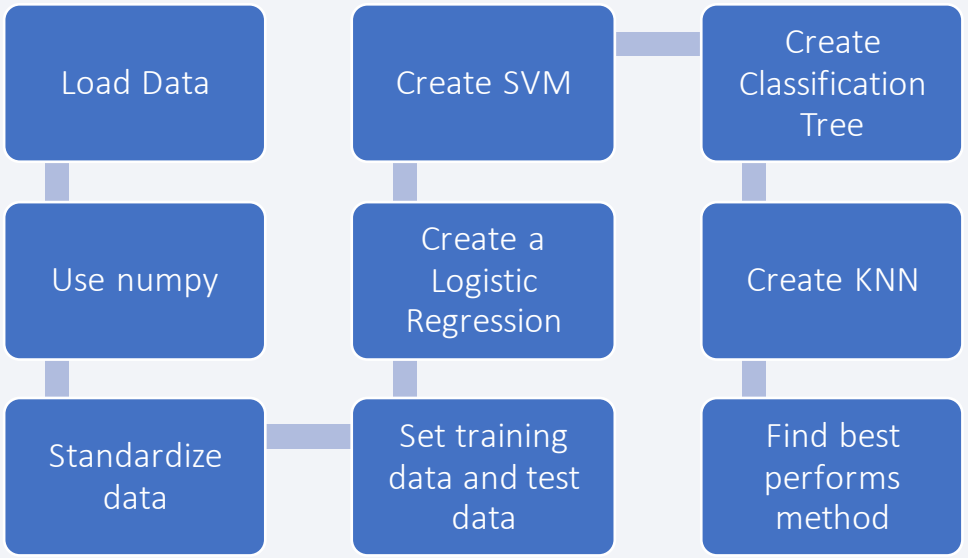
- Use folium.Map, folium.marker, folium.Circles, folium.PolyLines, etc, created and added to a folium map
- folium.Map: initial different Launch Site
- folium.marker: shows two Launch Site distance
- folium.Circles: highlight Launch Site
- folium.PolyLines: draw a Polyline between Launch Site with like railways, highways, coastline, cities
- Hua Yang's GitHub: https://github.com/shasha920/SpaceXFalcon9firststageLandingPredictionPython/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Use Pie chart plot and scatter point chart plot added to a dashboard
- Pie chart plot: when click site dropdown choose site, shows success rate
- Scatter point chart plot: when click site dropdown choose site, click payload_slider, shows success rate
- Hua Yang's GitHub: https://github.com/shasha920/SpaceXFalcon9firststageLandingPredictionPython/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

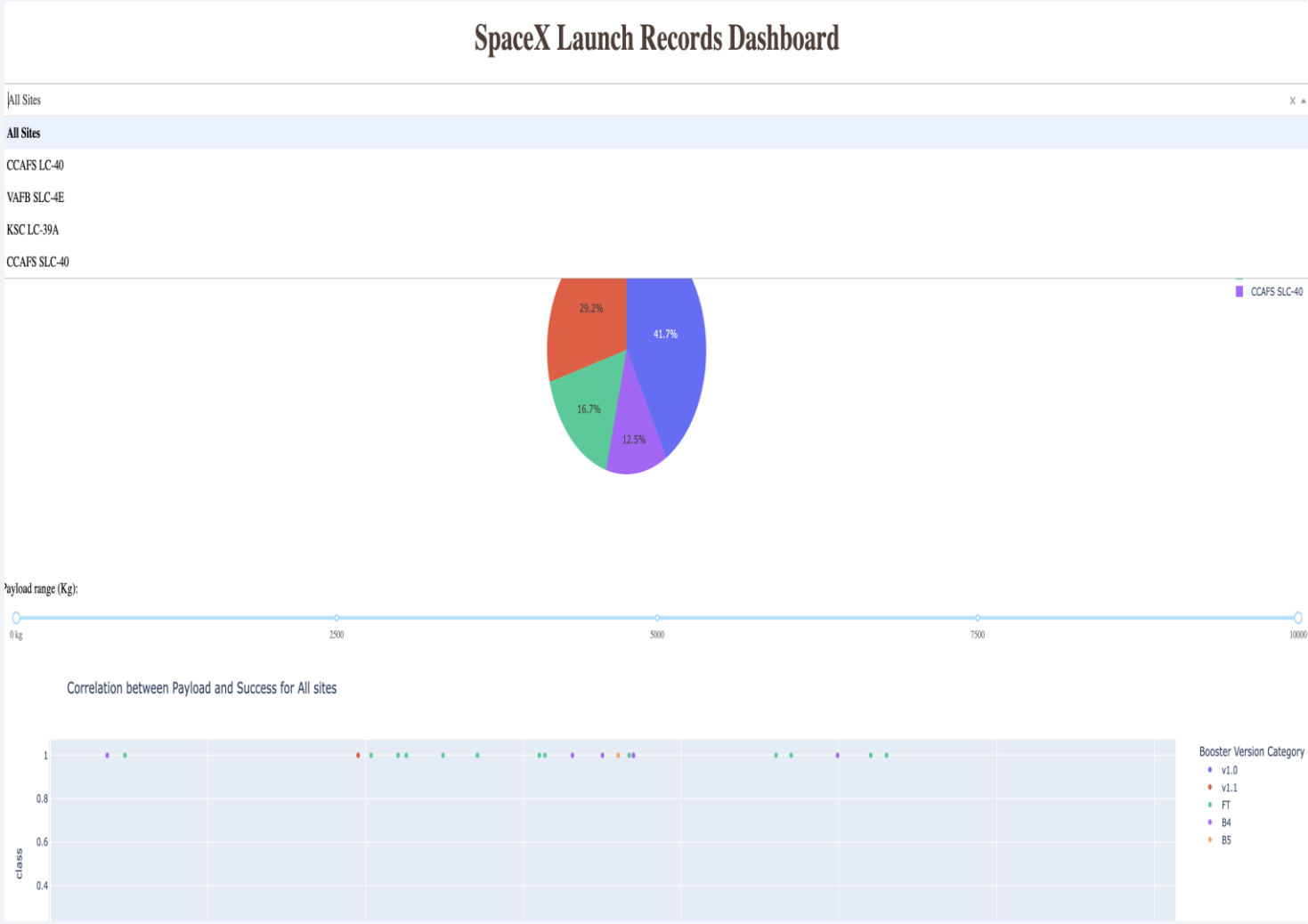
- Perform exploratory Data Analysis and determine Training Labels, then use SVM, Classification Trees, KNN and Logistic Regression to find best Hyperparameter
- You need present your model development process using key phrases and flowchart
- Hua Yang's GitHub: https://github.com/shasha920/SpaceXFalcon9firststageLandingPredictionPython/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



Results

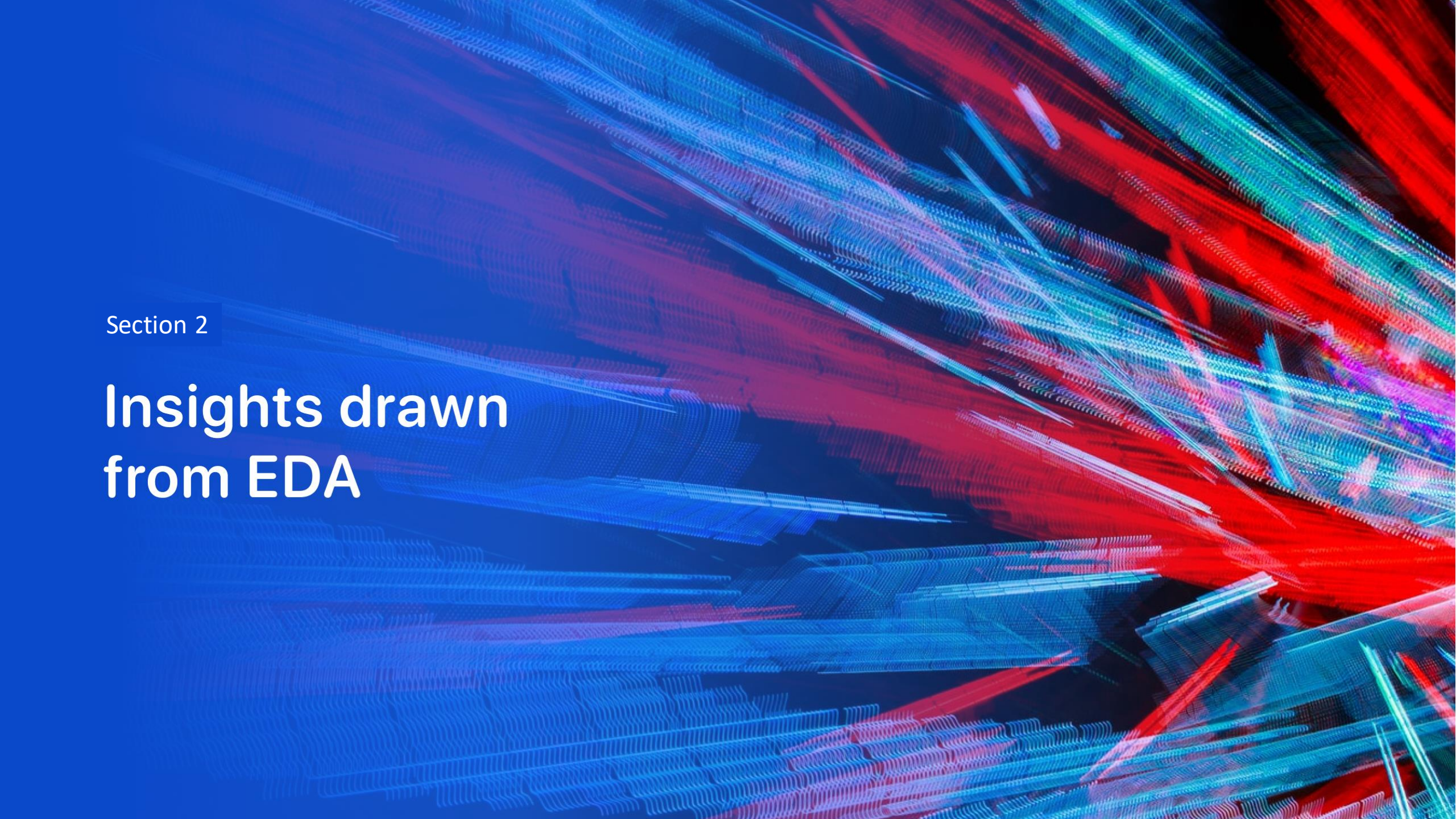
- As Data Analysis I found all launch sites close proximity to railways, highways, coastline, and keep certain distance away from cities

- As Machine Learning Prediction, shows Space X first stage will land score is 83.33%



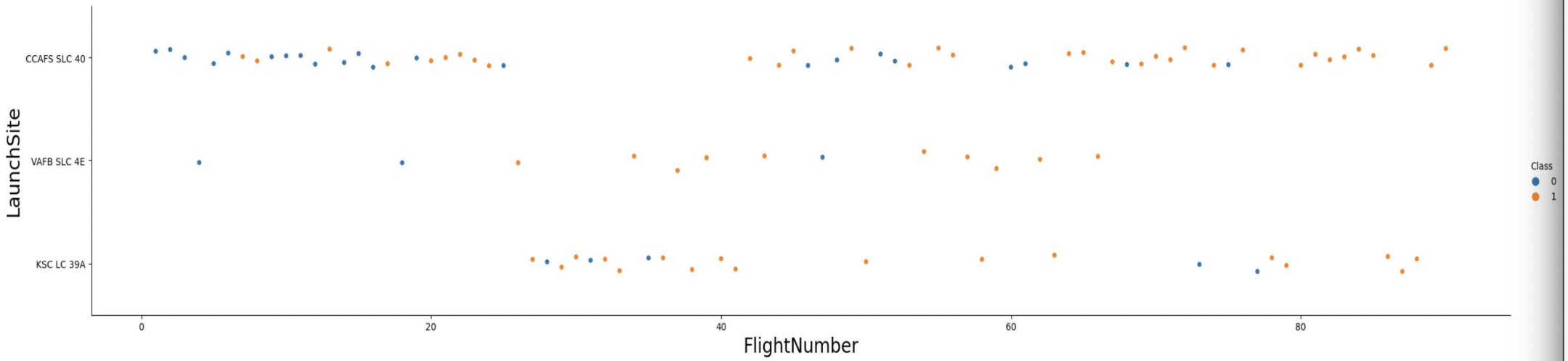
Section 2

Insights drawn from EDA



Flight Number vs. Launch Site

```
# Plot a scatter point chart with x axis to be Flight Number and y axis to be the launch site, and hue to be the class value
sns.catplot(x="FlightNumber",y="LaunchSite",hue="Class",data=df,aspect=5)
plt.xlabel("FlightNumber",fontsize=20)
plt.ylabel("LaunchSite",fontsize=20)
plt.show()
```

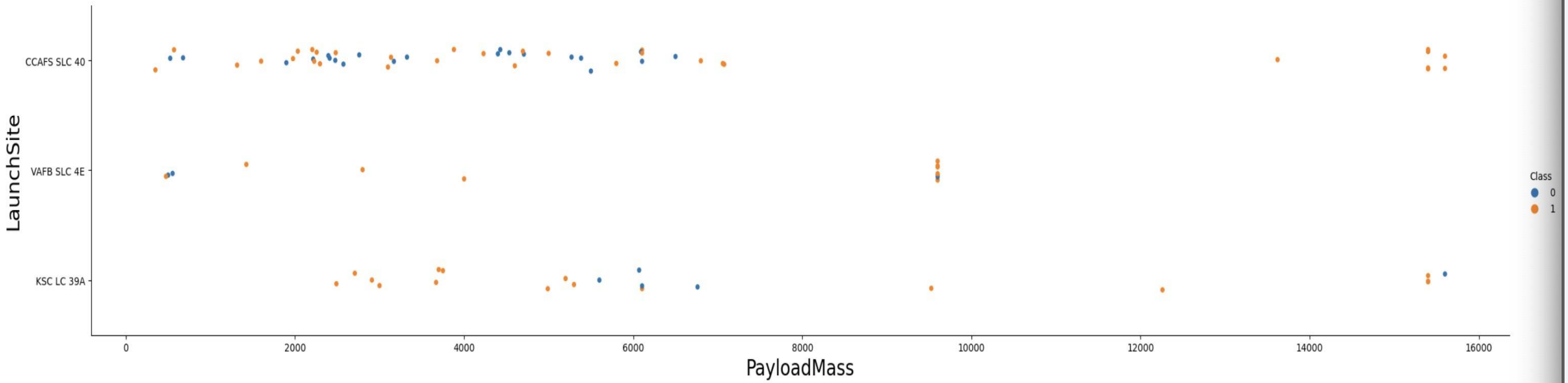


Now try to explain the patterns you found in the Flight Number vs. Launch Site scatter point plots.

Conclusion: Now, we can see CCDAFS SLC-40 have many times flight number, VAFB SLC 4E and KSC LC 39A have less flight number. Begin flight have more failure, but after have more successful flight

Payload vs. Launch Site

```
# Plot a scatter point chart with x axis to be Pay Load Mass (kg) and y axis to be the launch site, and hue to be the class va.
sns.catplot(x="PayloadMass",y="LaunchSite",hue="Class",data=df,aspect=5)
plt.xlabel("PayloadMass",fontsize=20)
plt.ylabel("LaunchSite",fontsize=20)
plt.show()
```

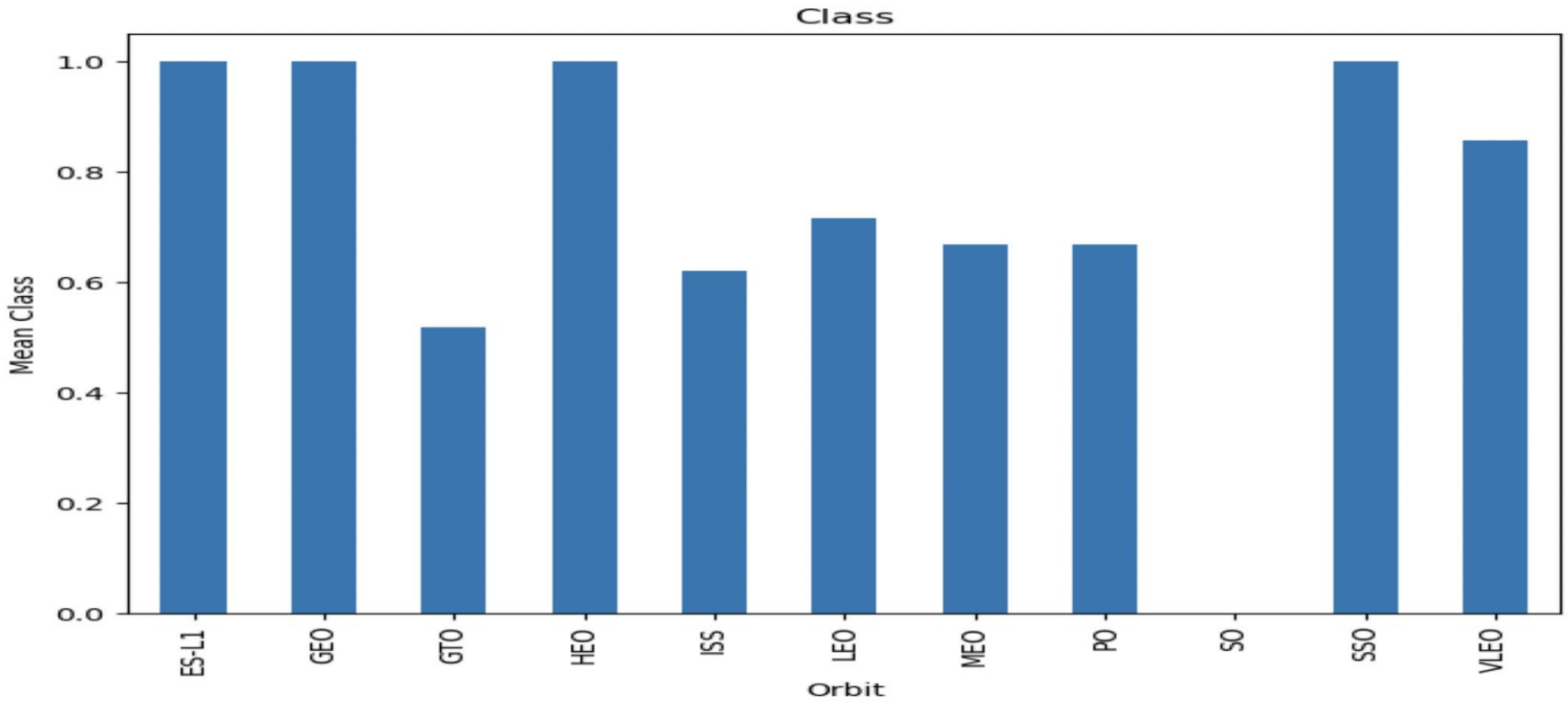


Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type

```
avg_class.plot(kind='bar',title='Class',ylabel='Mean Class', xlabel='Orbit',figsize=(8,6))
```

```
Out[6]: <AxesSubplot:title={'center':'Class'}, xlabel='Orbit', ylabel='Mean Class'>
```

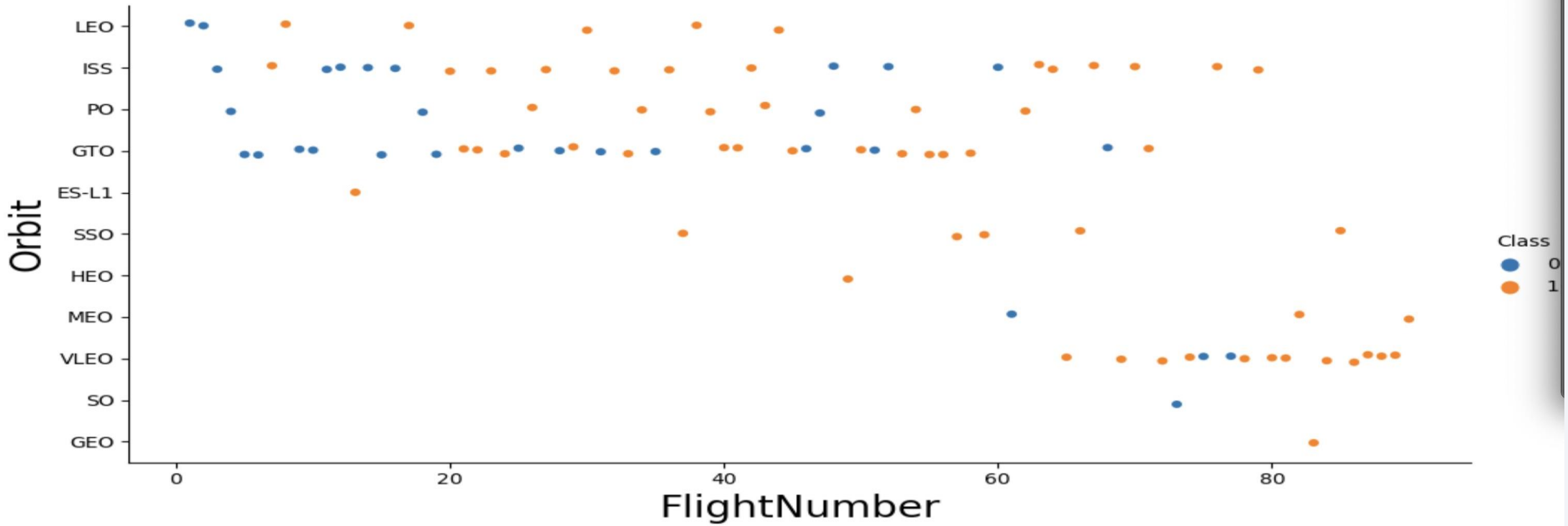


Analyze the ploted bar chart try to find which orbits have high sucess rate.

Conclusion: ES-L1,GEO,HEO,SSO have highest successful rate

Flight Number vs. Orbit Type

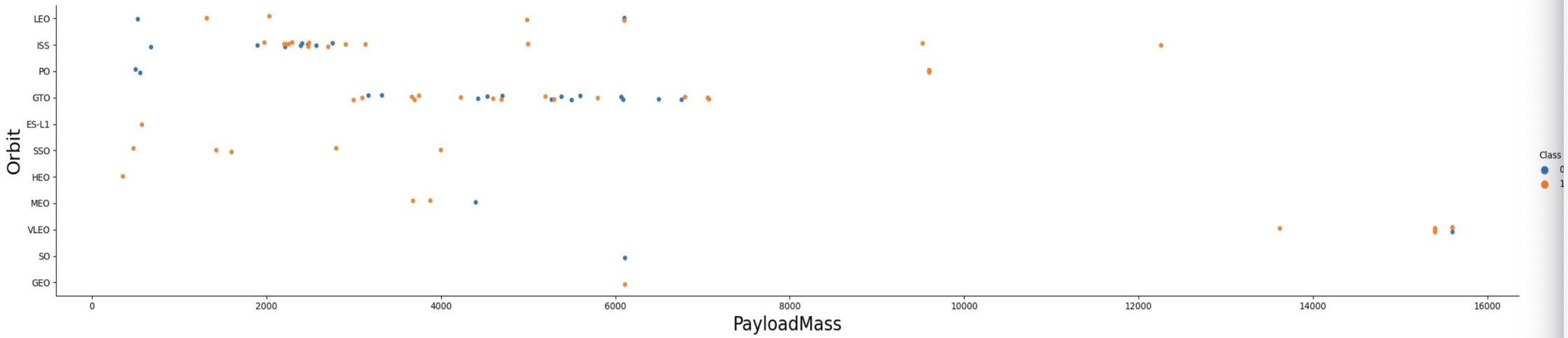
```
# Plot a scatter point chart with x axis to be FlightNumber and y axis to be the Orbit, and hue to be the class value
sns.catplot(x="FlightNumber",y="Orbit",hue="Class",data=df,aspect=2)
plt.xlabel("FlightNumber",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```



You should see that 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

```
# Plot a scatter point chart with x axis to be Payload and y axis to be the Orbit, and hue to be the class value
sns.catplot(x="PayloadMass",y="Orbit",hue="Class",data=df,aspect=5)
plt.xlabel("PayloadMass",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```

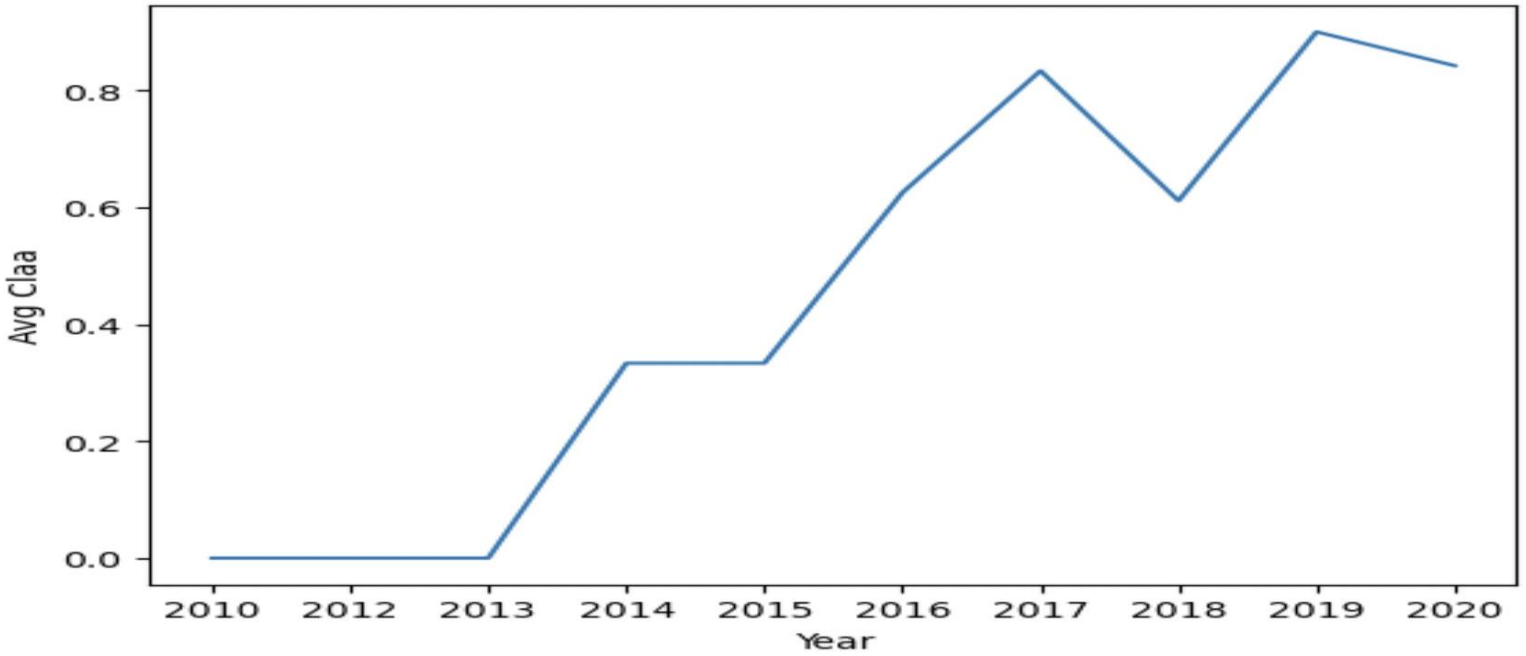


With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

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

```
# Plot a line chart with x axis to be the extracted year and y axis to be the success rate
list=Extract_year(df['Date'])
df1 = pd.DataFrame(list, columns =['year'])
df1['Class']=df['Class']
sns.lineplot(data=df1, x=np.unique(list) , y=df1.groupby(['year'])['Class'].mean())
plt.xlabel('Year')
plt.ylabel('Avg Claa')
plt.show()
```



you can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

Follow is Hua Yang use SQL find All Launch Site Names:

```
%sql select distinct "Launch_Site" from spacextbl
```

```
 * sqlite:///my_data1.db
Done.
```

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

Launch Site Names Begin with 'CCA'

Follow is Hua Yang use SQL find Launch Site Names Begin with 'CCA':

```
%sql select * from spacextbl where "Launch_Site" like "CCA%" limit 5
```

```
 * sqlite:///my_data1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

Follow is Hua Yang use SQL find Total Payload Mass:

```
%sql select SUM("PAYLOAD_MASS__KG_") from spacextbl WHERE "Customer"=="NASA (CRS)"
```

```
 * sqlite:///my_data1.db
Done.
```

SUM("PAYLOAD_MASS__KG_")
45596

Average Payload Mass by F9 v1.1

Follow is Hua Yang use SQL find Average Payload Mass by F9 v1.1:

```
%sql select AVG("PAYLOAD_MASS__KG_") from spacextbl WHERE "Booster_Version"=="F9 v1.1"
```

```
 * sqlite:///my_data1.db
Done.
```

AVG("PAYLOAD_MASS__KG_")
2928.4

First Successful Ground Landing Date

- Follow is Hua Yang use SQL find First Successful Ground Landing Date:

```
%sql select min("Date") from spacextbl where "Landing _Outcome"== "Success (ground pad)"
```

```
 * sqlite:///my_data1.db
Done.
```

min("Date")
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

- Follow is Hua Yang use SQL find Successful Drone Ship Landing with Payload between 4000 and 6000:

```
%sql select "Booster_Version" from spacextbl where "Landing _Outcome"== "Success (drone ship)" AND ("PAYLOAD_MASS__KG_" betwee
```

```
 * sqlite:///my_data1.db
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

- Follow is Hua Yang use SQL find Total Number of Successful and Failure Mission Outcomes:

```
%sql select COUNT("Mission_Outcome") from spacextbl
```

```
 * sqlite:///my_data1.db
Done.
```

COUNT("Mission_Outcome")
101

Boosters Carried Maximum Payload

- Follow is Hua Yang use SQL find Boosters Carried Maximum Payload:

```
%sql select distinct "Booster_Version", "PAYLOAD_MASS__KG_" from spacextbl where "PAYLOAD_MASS__KG_" = (select max("PAYLOAD_MAS
```

```
 * sqlite:///my_data1.db
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

- Follow is Hua Yang use SQL find 2015 Launch Records:

```
%sql select substr(Date,4,2) as month,"Landing _Outcome","Booster_Version","Launch_Site" from spacextbl where substr("Date",7,
```

```
 * sqlite:///my_data1.db
Done.
```

month	Landing _Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- Follow is Hua Yang use SQL find Rank Landing Outcomes Between 2010-06-04 and 2017-03-20:

```
%sql select "Date",count("Landing _Outcome")as countOfSuc from spacextbl WHERE ("Date" between  '04-06-2010' and '20-03-2017')
```

```
 * sqlite:///my_data1.db
Done.
```

Date	countOfSuc
07-08-2018	20
08-04-2016	8
18-07-2016	6

Section 3

Launch Sites Proximities Analysis

<Each Site's Location on a Map>

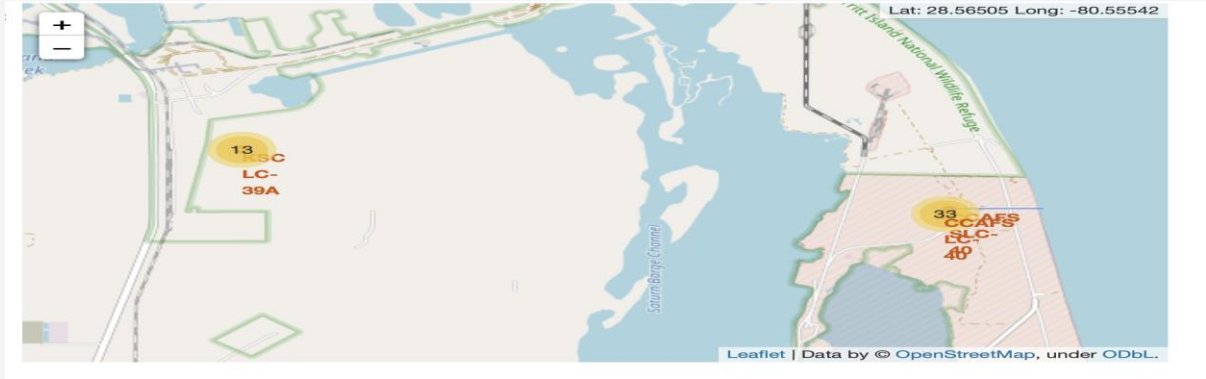
- Total 4 Launch Site location two main coast, one coast near CA, one coast near FL

<Count Successful rate on Map>

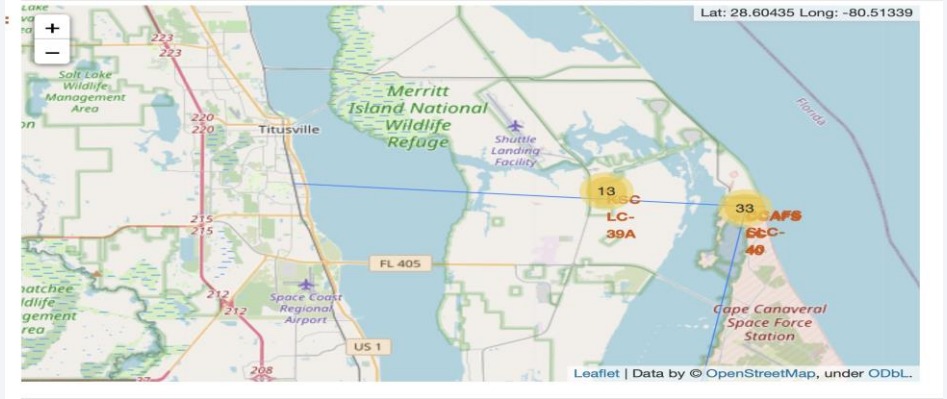
Set Successful as class 1, fail as 0. Then count Successful on Map

<Site's Location distance with Coastline, Railway, Cities, and Highway>

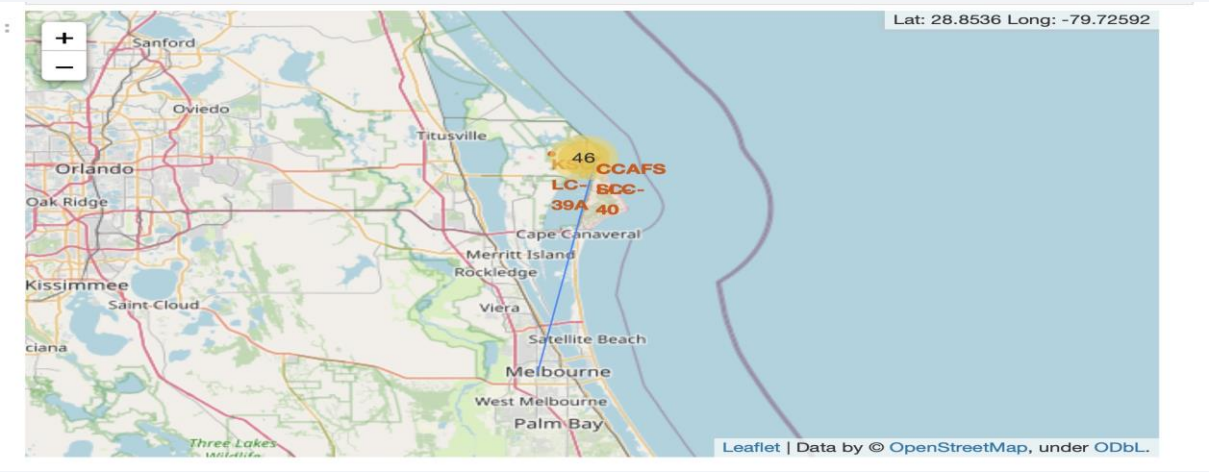
- Near Coastline



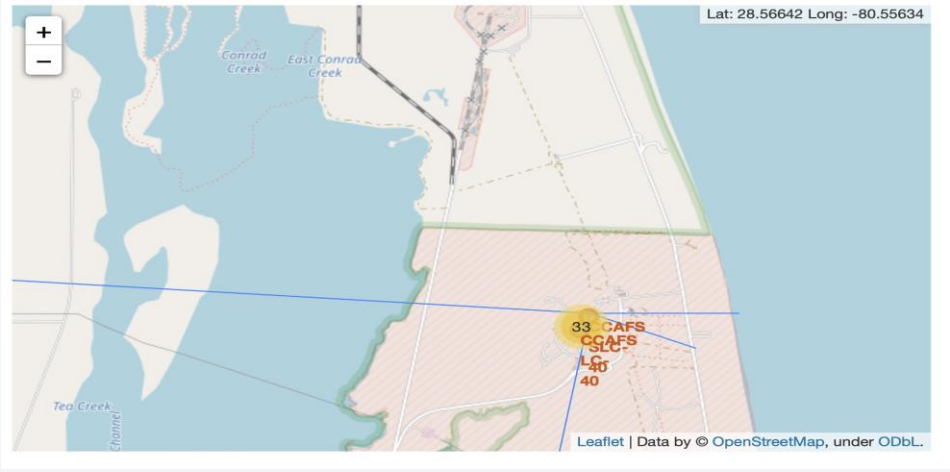
Near Railway

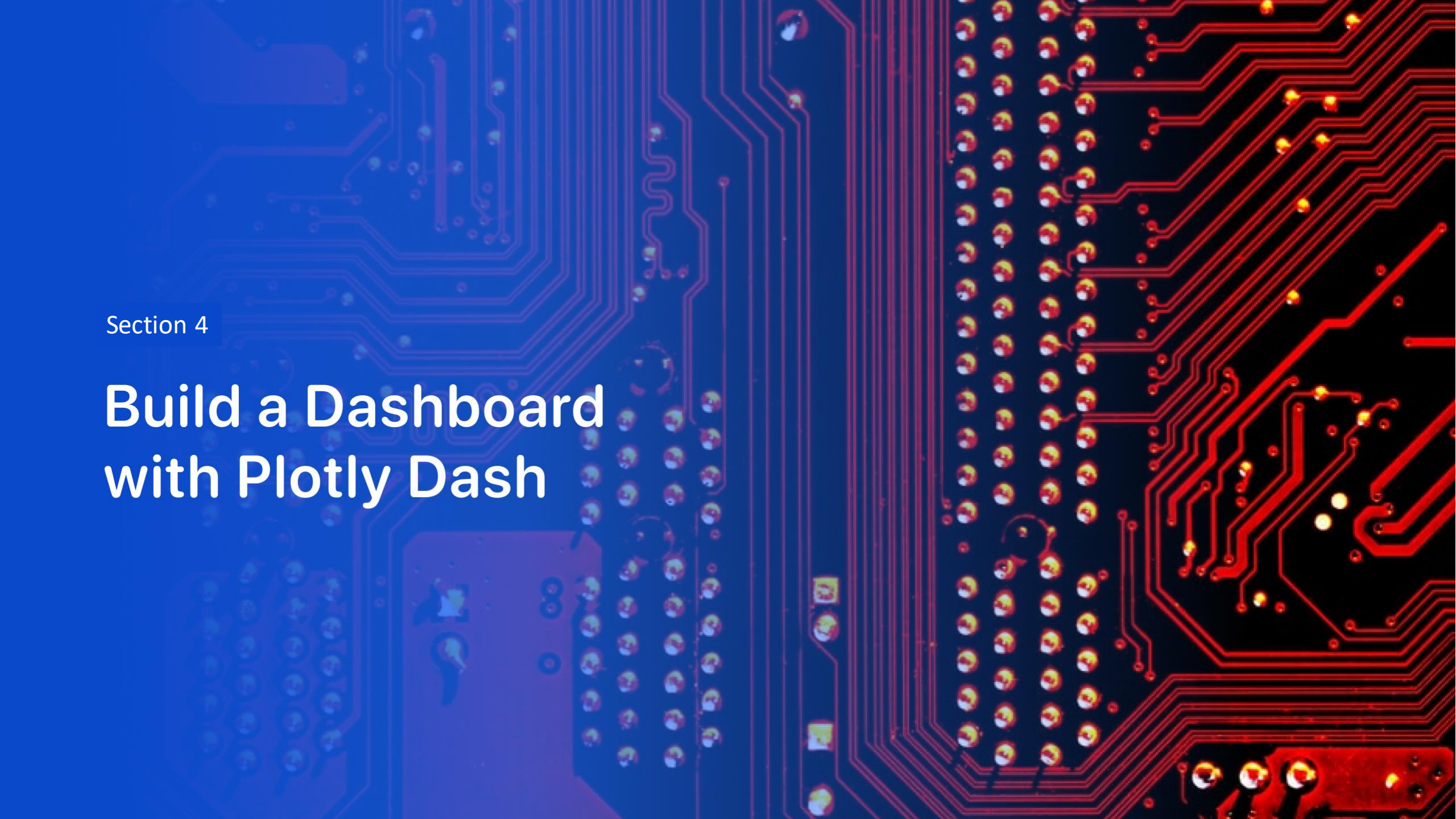


Far away Cities



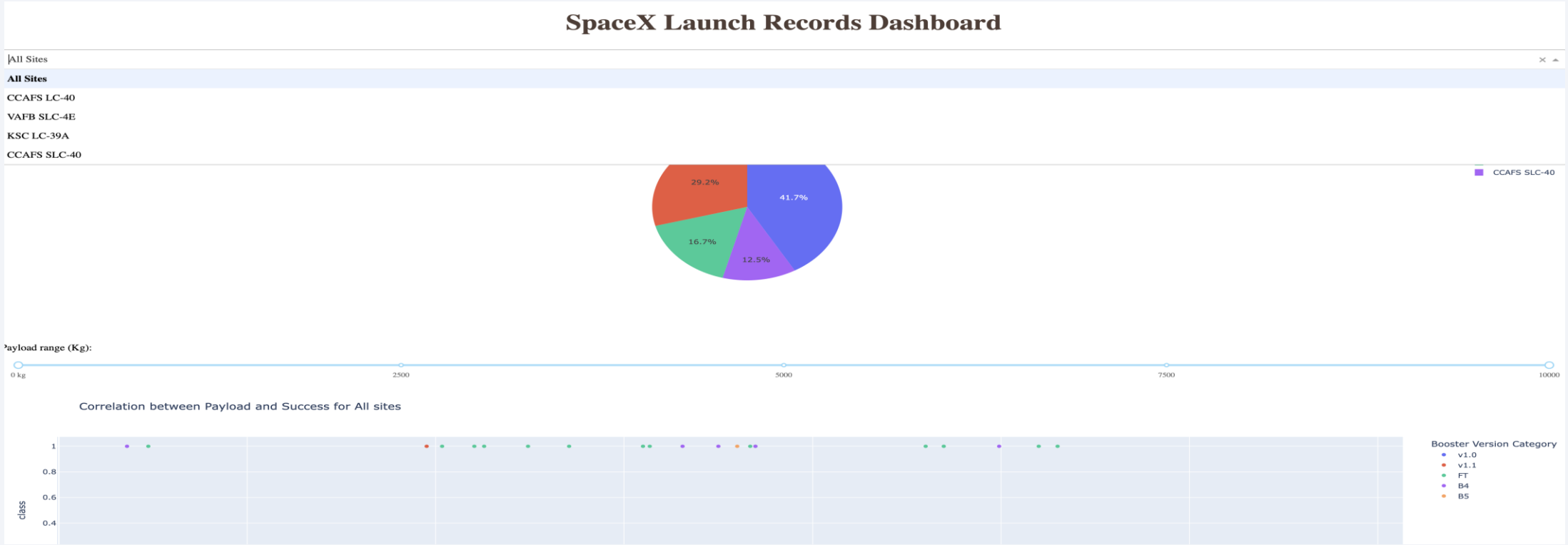
Near Highway




Section 4
Build a Dashboard with Plotly Dash

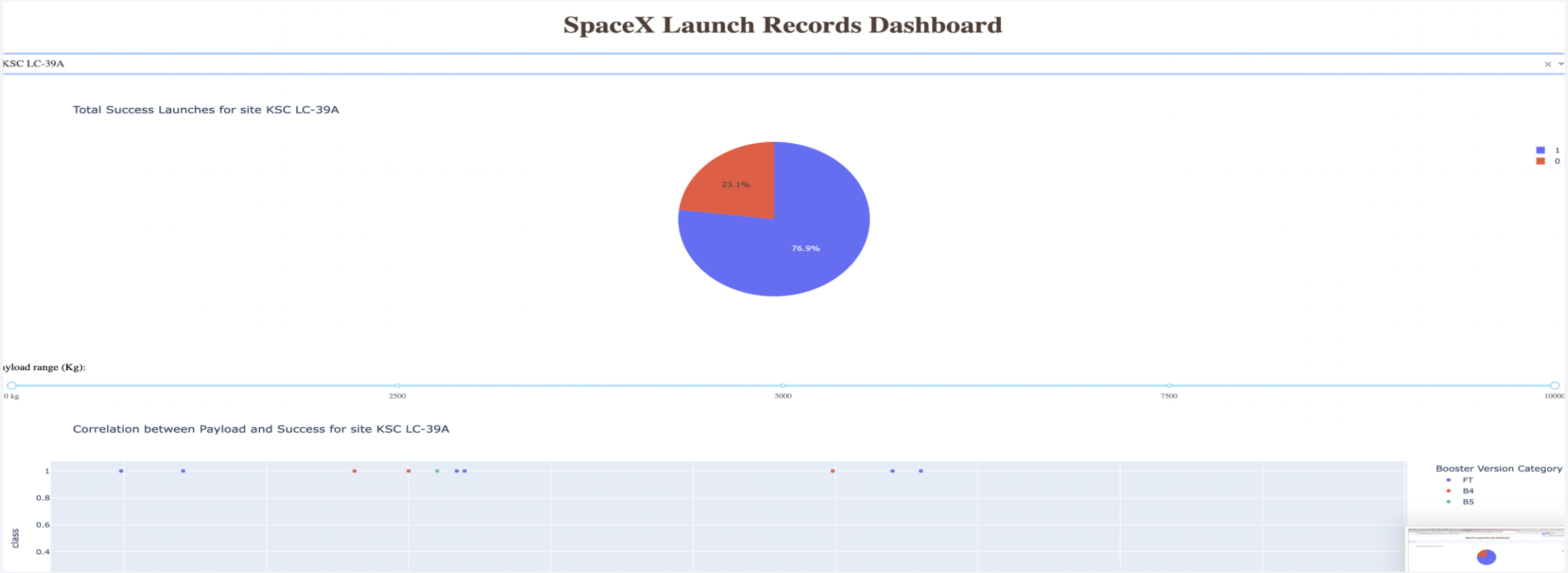
<The launch success count for all sites>

- **Follow shows the launch success count for all sites Pie Chart and Scatter point Chart, we can see KSC LC-39A has highest success count**



<KSC LC-39A distribution>

- Now we see **KSC LC-39A distribution**



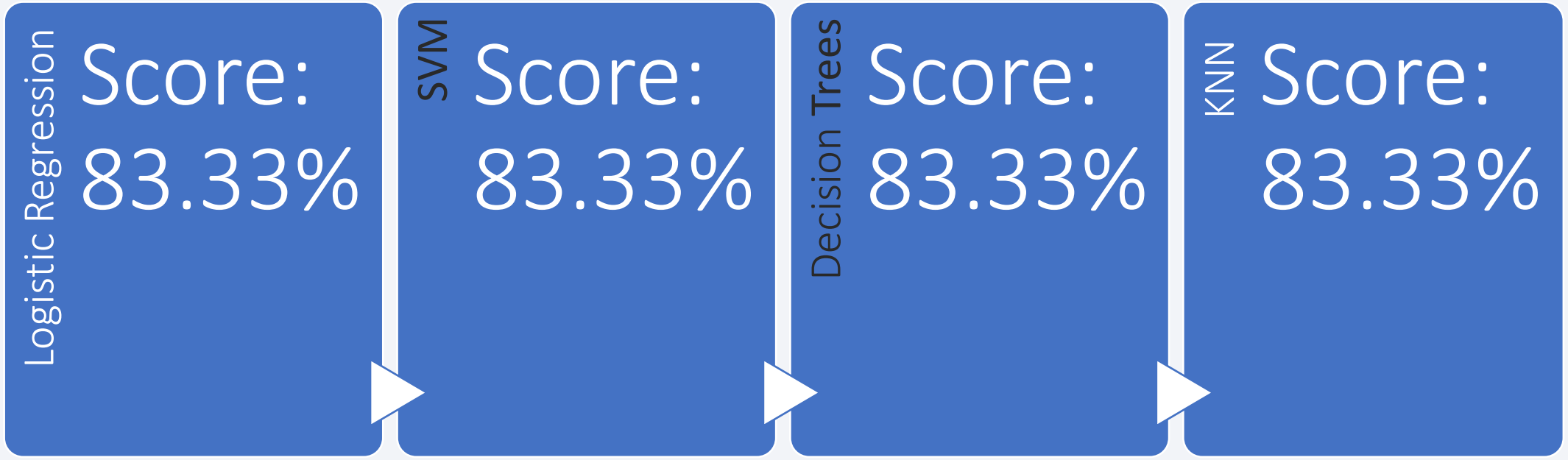
<Payload vs. Launch Outcome scatter plot>

Payload vs. Launch Outcome scatter plot for all sites, shows FT has the largest success rate, etc.

Section 5

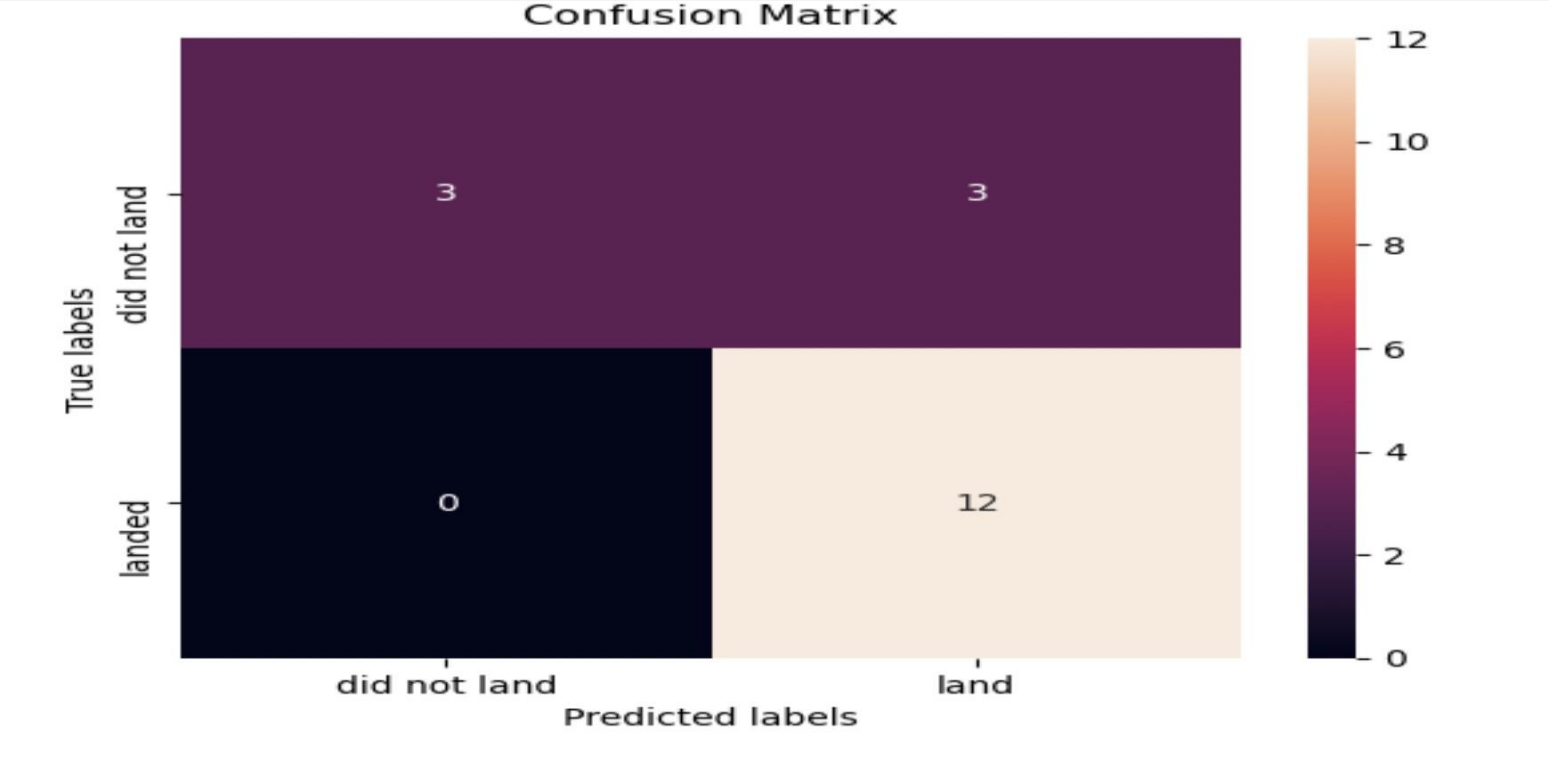
Predictive Analysis (Classification)

Classification Accuracy



- They almost all point to score as 83.33%, so I think they are all performs best.

Confusion Matrix



- Confusion Matrix pattern almost same

Conclusions

- Different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
- Sucess rate since 2013 kept increasing till 2020
- Launch sites close proximity to railways, highways, coastline, and keep certain distance away from cities
- Use SVM, Decision Trees, KNN and Logistic Regression to find best Hyperparameter score, found they almost 83.33%, any one could predict best Hyperparameter
- Use SVM, Decision Trees, KNN and Logistic Regression made Confusion Matrix, get same results, so anyone could do Confusion Matrix.
- Get results:
  - SpaceX have invest value
  - We can copy another SpaceX use same progress

Appendix

- All code see Hua Yang's GitHub:

https://github.com/shasha920/SpaceXFalcon9firststageLandingPredictionPython

Thank you!