IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Haowei>
<2023-12-20>

Outline

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

Executive Summary

- The following methodologies were used to analyze data:
  - Data Collection (BeautifulSoup, Request, SpaceX Rest API)
  - Data Wrangling
  - Exploratory Data Analysis (Correlated Analysis, Data Visualization with Folium)
  - Interactive Data Dashboard (Dash, Ploty)
  - Predictive Analysis (SVM, KNN, LR etc.)
- Summary of all results
  - Exploratory Data Analsis Results
  - Interactive Dashboard Demo
  - Predictive Analysis Results

Introduction

- Project background and context
  - SpaceX, a trailblazer in commercial space exploration, has revolutionized affordability in space travel. Advertising Falcon 9 rocket launches at $62 million, a fraction of the competitors' $165 million, the key lies in their ability to reuse the first stage. Our aim is to predict the first stage's successful landing using public data and machine learning models, crucial in estimating launch costs.
- Problems you want to find answers
  - How do variables such as payload mass, launch site, number of lights, and orbits affect the success of the first stage landing?
  - Does the rate of successful landings increase over the years?
  - Are there any easy-interpreted machine learning models suitable for predicting unseen results?

Section 1

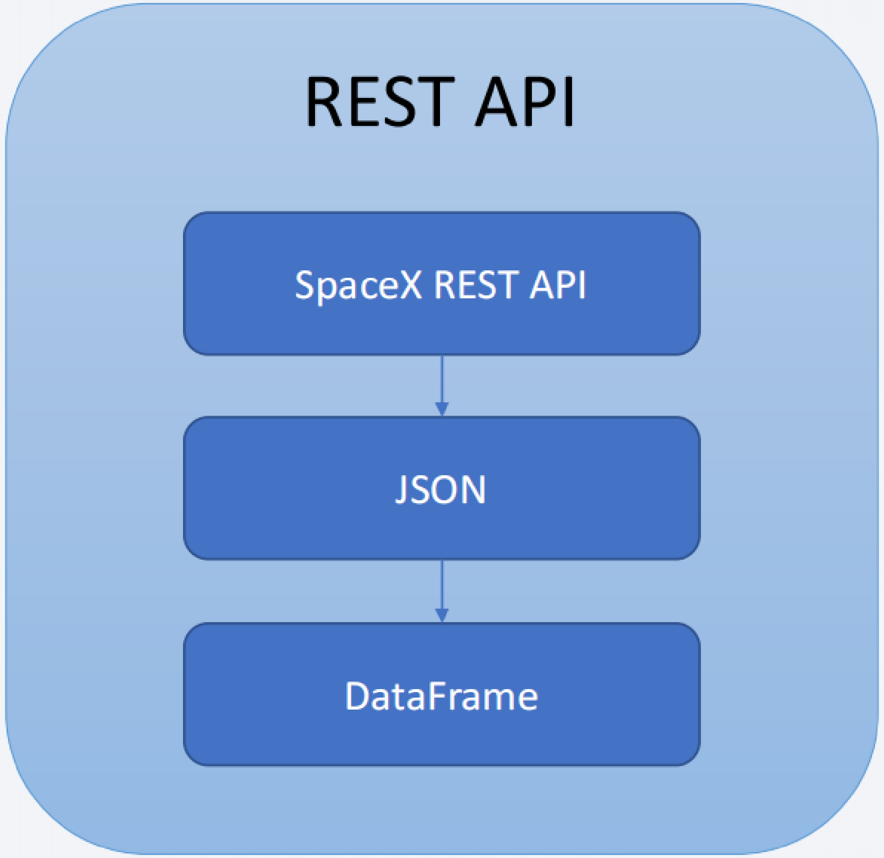
Methodology

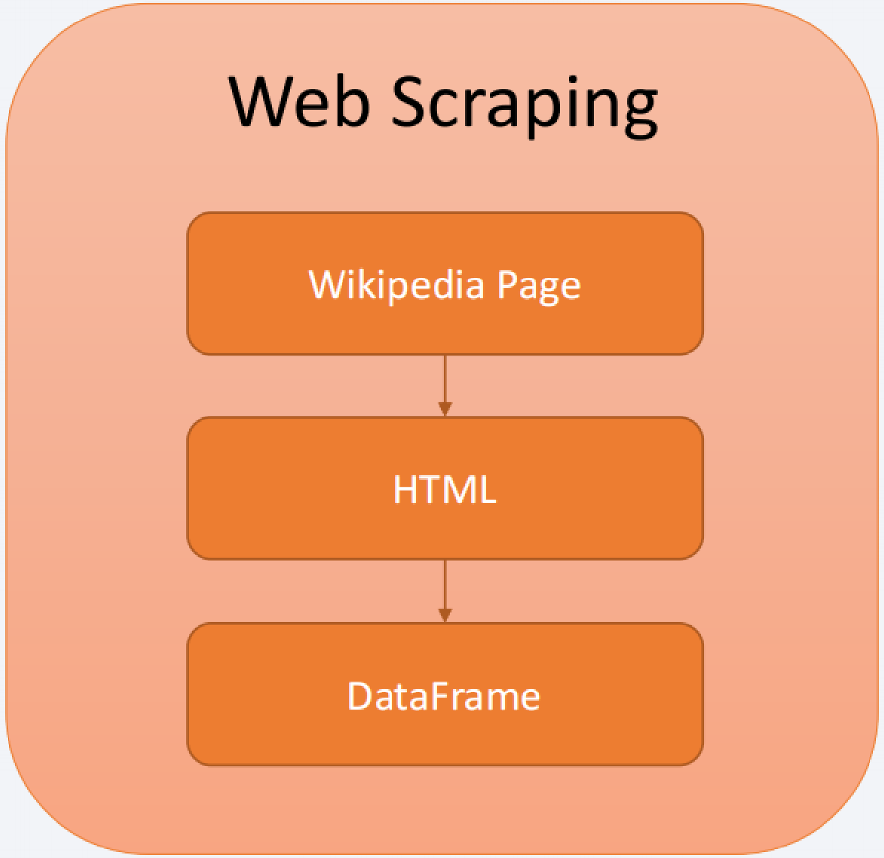
Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX Rest API (https://api.spacexdata.com/v4/rockets/)
  - Web Scraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- Perform data wrangling
  - Data Filtering, Dealing with Missing Value, One-Hot-Encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Well-built, fine-tuned classification models to seek best evaluation

Methodology





Data Collection

- Data collection process involved a combination of API requests from SpaceX REST API and web scraping data from a table in SpaceX's Wikipedia pages.
  - Data Columns are obtained by using SpaceX REST API:
    - FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
  - Data Columns are obtained by using Wikipedia Web Scraping:
    - Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch, outcome, Version Booster, Booster landing, Date, Time

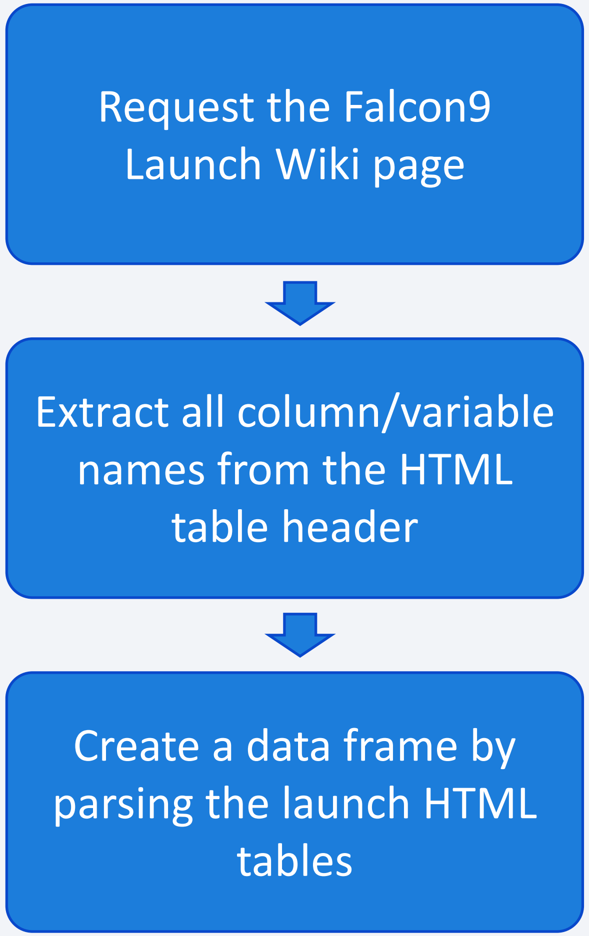
Data Collection – SpaceX API

- SpaceX offers a public API from where data can be obtained and then used
- This API was used according to the flowchart beside and then data is persisted.

- Source code: [Github-Data-Collection]

Data Collection - Scraping

- Data from SpaceX launches can also be obtained from Wikipedia
- Data are downloaded from Wikipedia according to the flowchart and then persisted



- Source code: [Github-Data-Collection]

Data Wrangling

- EDA was performed on the collected dataset
- Calculate the number of launches
- Calculate the number and occurrence of each orbit
- Calculate the number and occurrence of mission outcome per orbit type
- Create a landing outcome label from Outcome column

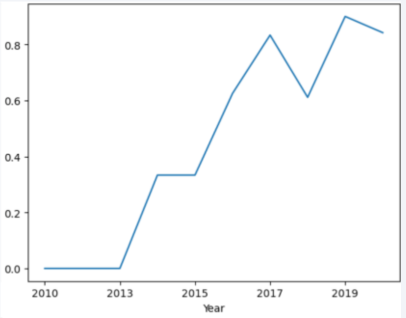
- Source code: Github-Data-Wrangling

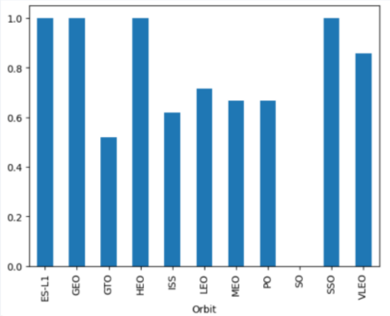
EDA with Data Visualization

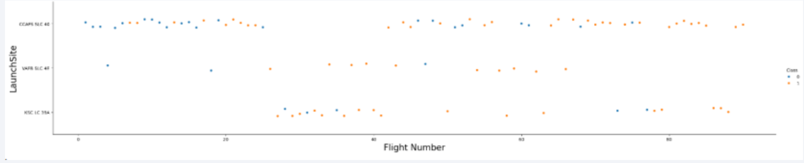
- Summarize what charts were plotted and why you used those charts

- Source code: [Github-Data-Viz]

Scatter Plot	To get relationship between variables, e.g.: • FlightNumber vs. Orbit type • Payload vs. Orbit type • FlightNumber vs. PayloadMass • FlightNumber vs. Launch Site
Bar Plot	To plot success rate of each orbit
Line Plot	To get the yearly average launch success trend







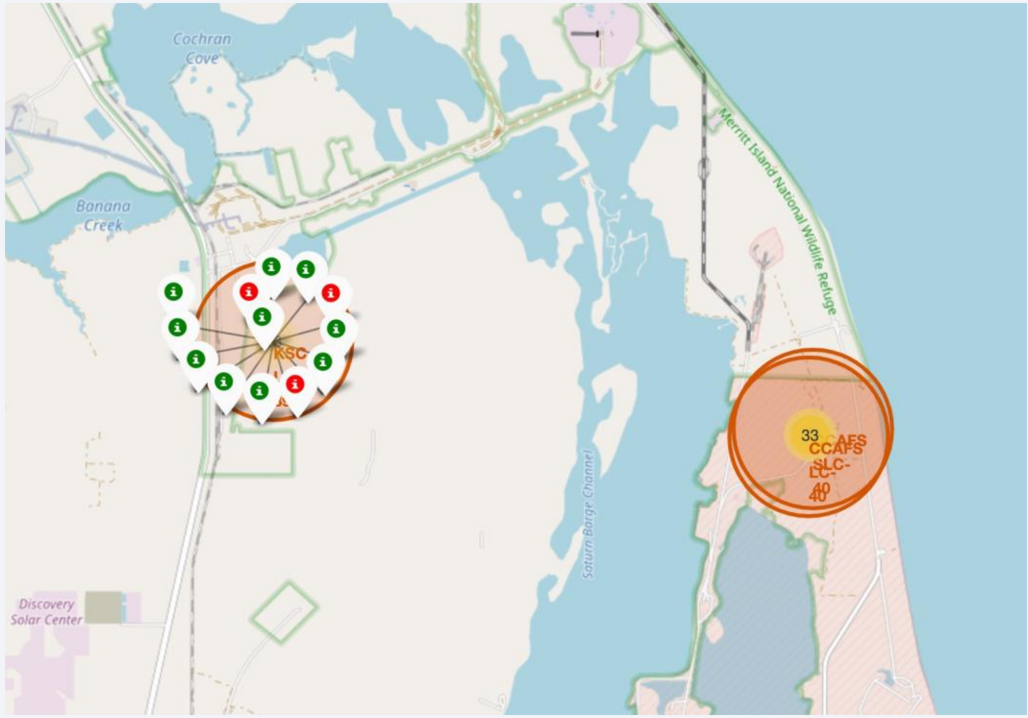
EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
  - Types of lauch sites in space mission
  - Top 5 launch sites belong to 'CCA’
  - Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg
  - Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20

  - Source code: Github-EDA-SQL

Build an Interactive Map with Folium

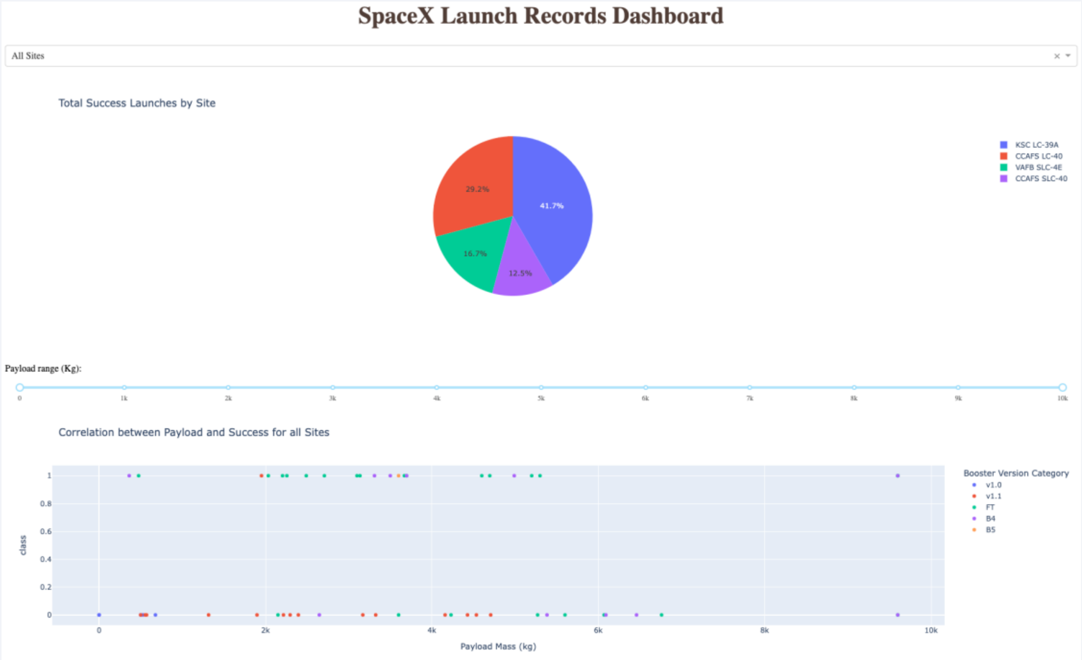
- Add Circles for Launch sites and Markers for labels
- Add MarkerCluster for successful and failed launches
- Add Lines for calculate distance between launch sites and their proximities



- Source code: [Github-Interactive-Viz]

Build a Dashboard with Plotly Dash

- With a Dropdown menu and a Pie Chart, we can get success launches distribution by launch site
- Additionally, with a Range Slider and a Scatter Plot, we can analyze the correlation between Payload and Success for different launch sites



- Source code: [Github-Interactive-Viz]

Predictive Analysis (Classification)

- Four classification models were utilized: Logistic Regression, Support Vector Machine, Decision Tree and K-Nearest Neighbors

- Source code: [Github-Models]

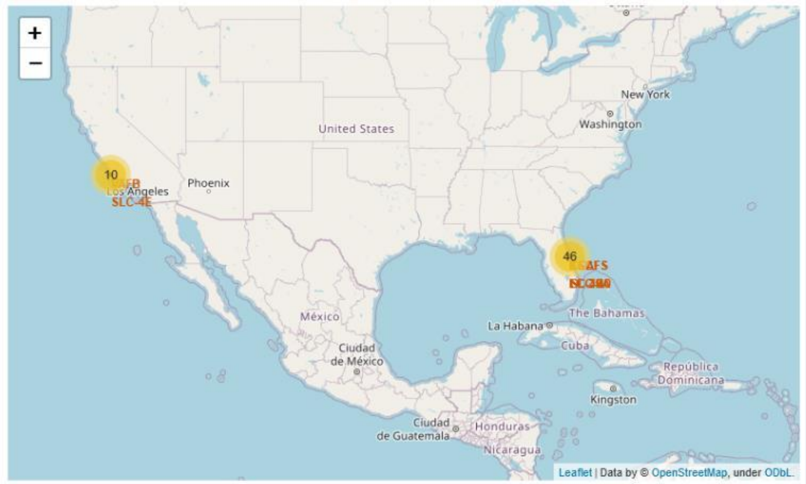
Data preparation and standardization → Hyperparameters fine-tuned by Grid Search method → Comparison of results

Results

- Exploratory data analysis results
  - Space X uses 4 different launch sites;
  - The first launches were done to Space X itself and NASA;
  - The average payload of F9 v1.1 booster is 2,928 kg;
  - The first success landing outcome happened in 2015 fiver year after the first launch;
  - Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average;
  - Almost 100% of mission outcomes were successful;
  - Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015;
  - The number of landing outcomes became as better as years passed

Results

- Interactive analytics demo in screenshots
  - Interactive analytics highlighted a pattern: launch sites are typically situated in secure coastal areas with strong logistical infrastructures
  - Most launches happens at east cost launch sites



Results

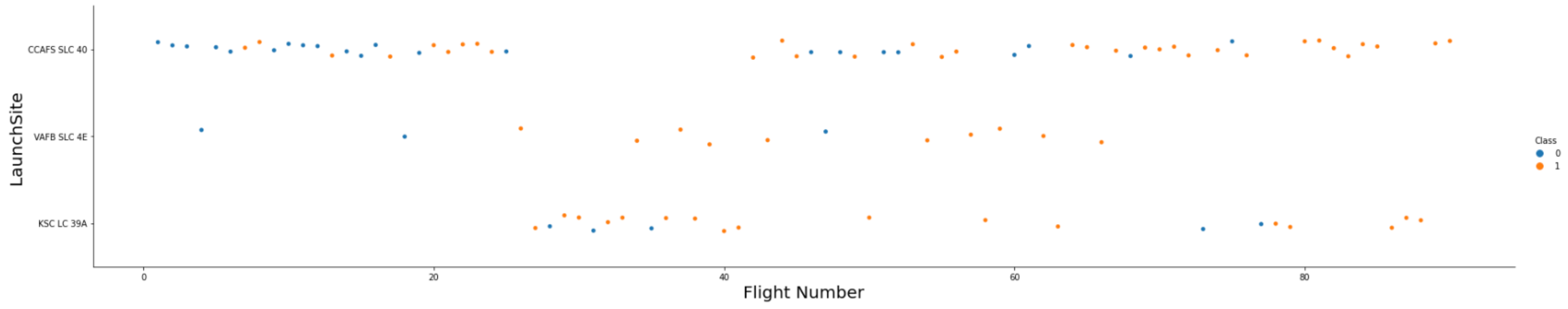
- Predictive analysis results
  - Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings with 83.33% test accuracy

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

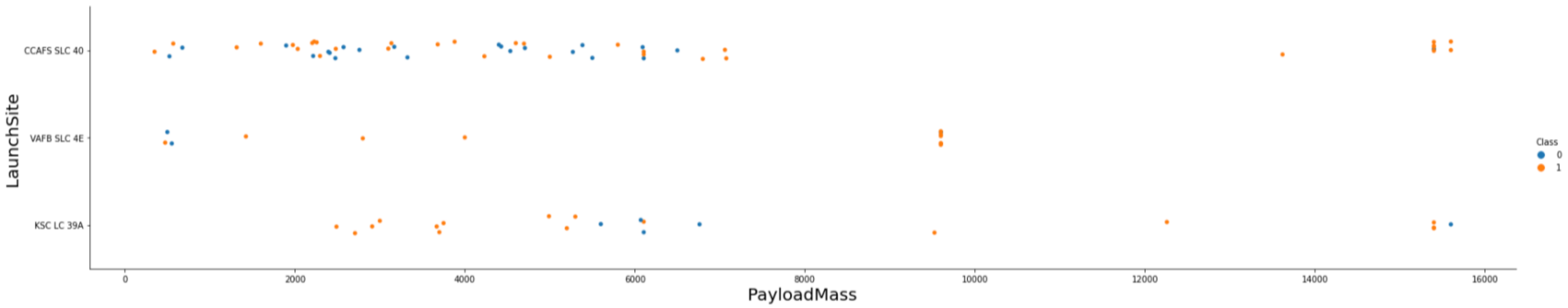
```
[4] # Plot a scatter point chart with x axis to be Flight Number and y axis to be the launch site, and hu
    sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)
    plt.xlabel("Flight Number",fontsize=20)
    plt.ylabel("LaunchSite",fontsize=20)
    plt.show()
```



Explanation: We can see from the scatter plot that as flight number increases, there are more successful first stage landing. With small flight numbers, launches happens more in the site CCAFS SLC 40 and with much lower success rate. Although there are less launches in VAFB SLC 4E and KSC LC 39A, higher success rate can be seen in these two sites.

Payload vs. Launch Site

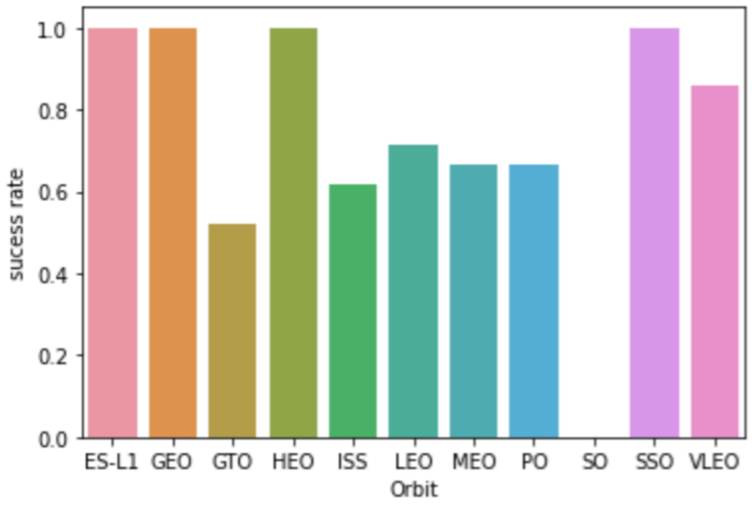
```
[5] # Plot a scatter point chart with x axis to be Pay Load Mass (kg) and y axis to be the launch site, a
    sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect = 5)
    plt.xlabel("PayloadMass",fontsize=20)
    plt.ylabel("LaunchSite",fontsize=20)
    plt.show()
```



Explanation: With higher Payload the success rate is much higher. And in KSC LC39A launchsite we can see much higher success rate with low Payload whereas this rate is mucher lower in CCAFS SLC 40 launchsite. Besides, there no rockets launched in VAFB-SLC for Payload greater than 10000. Furthermore, with Payload more than 9500, we can see very high success rate overall.

Success Rate vs. Orbit Type

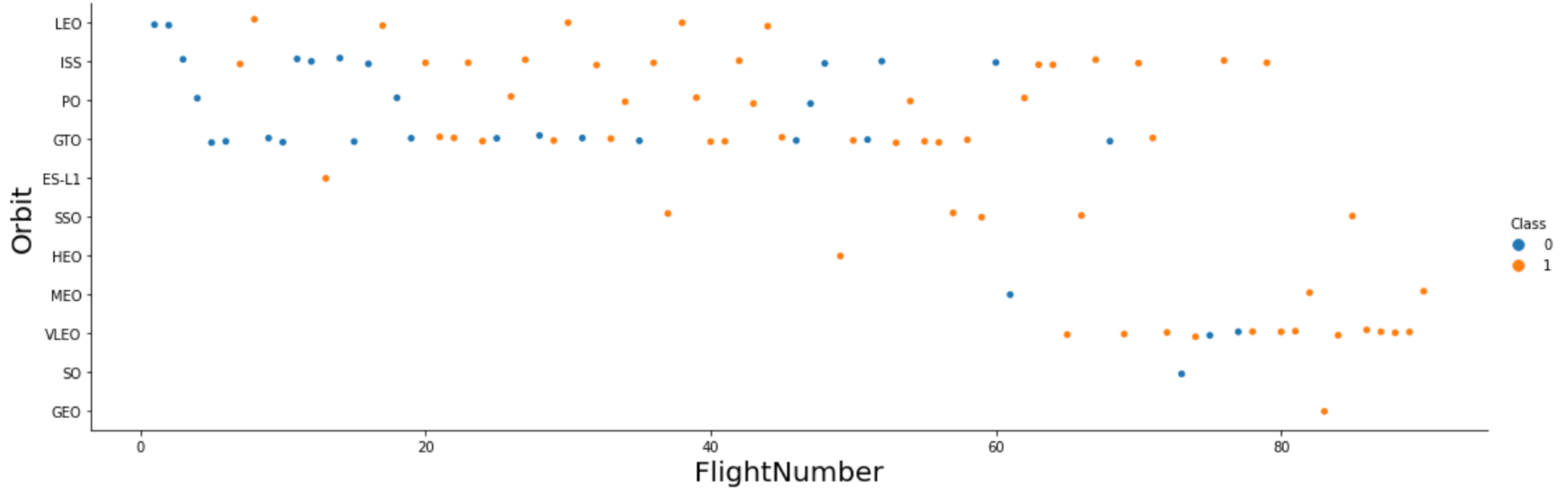
```
[ ] sns.barplot(y='Class', x='Orbit', data=df_success_rate)
    plt.xlabel("Orbit",fontsize=10)
    plt.ylabel("sucess rate",fontsize=10)
    plt.show()
```



Explanation: From the Bar Plot we can see for Orbit type ES-L1, GEO, HEO, and SSO have the highest success rate, which is 100%. And we also find in SO orbit, the rate is zero.

Flight Number vs. Orbit Type

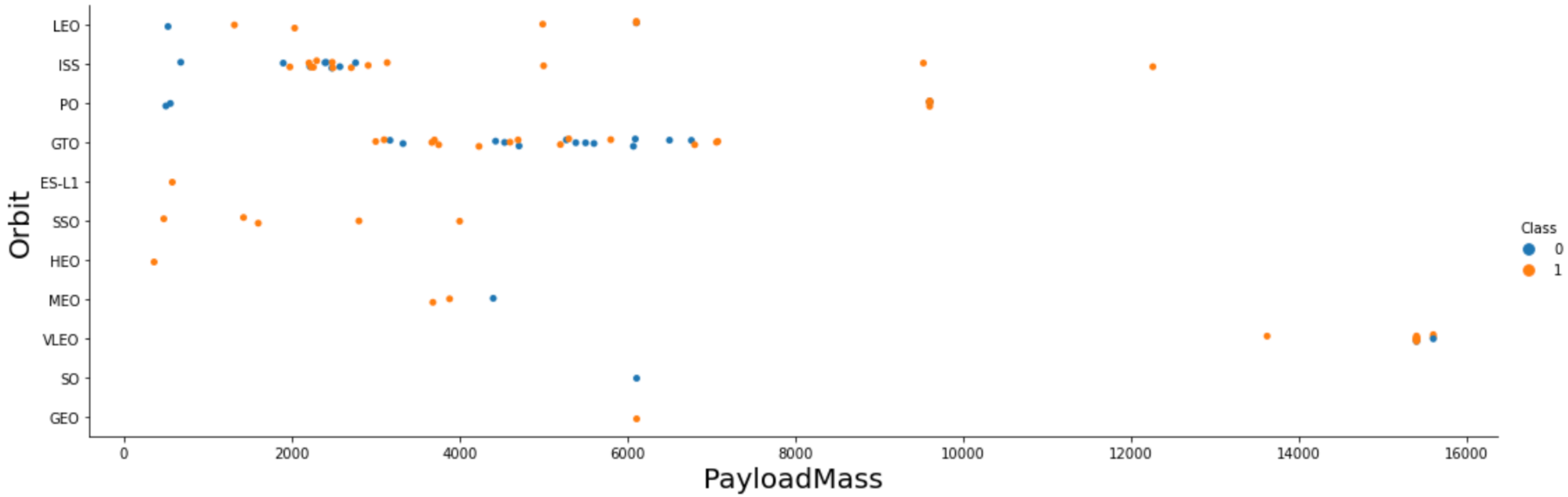
```
[9] # Plot a scatter point chart with x axis to be FlightNumber and y axis to be the Orbit, and hue to be
    sns.catplot(y="Orbit", x="FlightNumber", hue="Class", data=df, aspect = 3)
    plt.xlabel("FlightNumber",fontsize=20)
    plt.ylabel("Orbit",fontsize=20)
    plt.show()
```



Explanation: In ES-L1, GEO, HEO, and SSO orbits, all launches are successful. There is clear relationship between flight number and success rate in LEO orbit since as flightnumber increases, the success rate increases. In contrast, there is no such obvious relationship 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 v
    sns.catplot(y="Orbit", x="PayloadMass", hue="Class", data=df, aspect = 3)
    plt.xlabel("PayloadMass",fontsize=20)
    plt.ylabel("Orbit",fontsize=20)
    plt.show()
```

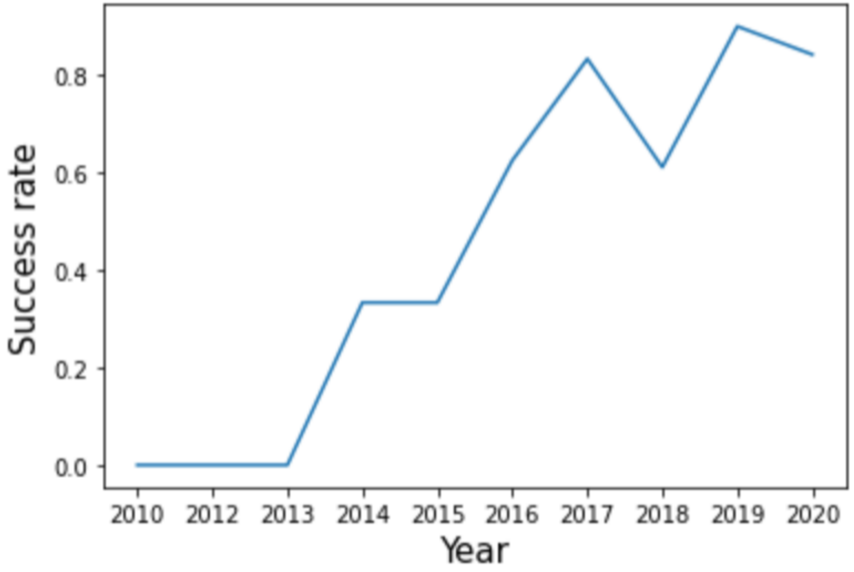


Explanation: With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

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

```
[14] sns.lineplot(y='Class', x='Year', data=df_year_success)
     plt.xlabel("Year",fontsize=15)
     plt.ylabel("Success rate",fontsize=15)
     plt.show()
```



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

All Launch Site Names

Four Launch Sites:

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

```
%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

1 in western coast

- VAFB SLC-4E

3 in eastern coast

- KSC LC-39A
- CCAFS SLC-40
- CCAFS LC-40

Launch Site Names Begin with 'CCA'

Display 5 records where launch sites begin with the string '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

5 launches happened in LEO orbit, and four of them were from NASA

Total Payload Mass

Display the total payload mass carried by boosters launched by NASA (CRS)

```
In [50]:  %sql select sum(PAYLOAD_MASS__KG_) as total_payload_mass from SPACEXTABLE where Customer = 'NASA (CRS)'

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

Out[50]:

total_payload_mass
45596

The total payload carried by boosters from NASA(CRS) is 45596

Average Payload Mass by F9 v1.1

In [53]:
```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version like "F9 v1.1%"
```

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

Out[53]:

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

The average payload mass carried by booster version F9 v1.1 is 2534.67

First Successful Ground Landing Date

In [74]:
```
%sql select min(Date) from SPACEXTABLE where Landing_Outcome like 'Success (ground pad)'
```

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

Out[74]:

min(Date)
2015-12-22

The first successful landing outcome on ground pad is 2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
In [81]: %sql select Booster_Version from SPACEXTABLE where Landing_Outcome = "Success (drone ship)" and Payload_MASS__KG_ between 4000 and 6000
```

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

Out[81]:

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

Successful Drone Ship Landing with Payload between 4000 and 6000:
F9 FT B1022; F9 FT B1026; F9 FT B1021.2; F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

In [85]:

```
%sql select Mission_Outcome, count(Mission_Outcome) from SPACEXTABLE group by Mission_Outcome
```

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

Out[85]:

Mission_Outcome	count(Mission_Outcome)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Generally, there are 100 successful outcomes, only 1 failed

Boosters Carried Maximum Payload

```
%%sql

select Booster_Version from SPACEXTBL
where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

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

```
In [95]: %sql select substr(Date, 6, 2) as monthnames, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTABLE \
         where substr(Date, 0, 5) = '2015' and Landing_Outcome like 'Failure%'

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

Out[95]:

monthnames	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

Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

```
%sql select count(Landing_Outcome) from SPACEXTABLE \
where Date between '2010-06-04' and '2017-03-20' \
group by Landing_Outcome
```

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

count(Landing_Outcome)
3
5
2
10
1
5
3
2

Section 3

Launch Sites Proximities Analysis

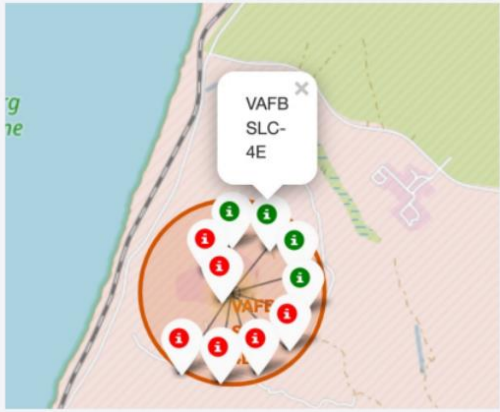
Locations of Launch Sites on Maps

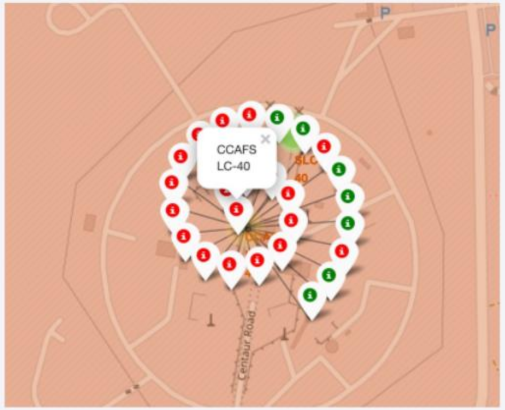
- VAFB SLC-4E in the west
- Other three in the east

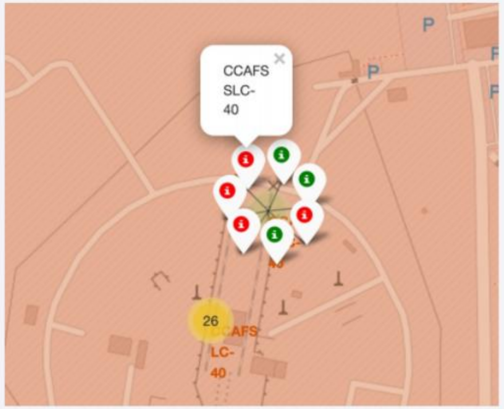
Launch Site	Lat	Long
CCAFS LC-40	28.56230197	-80.57735648
CCAFS SLC-40	28.56319718	-80.57682003
KSC LC-39A	28.57325457	-80.64689529
VAFB SLC-4E	34.63283416	-120.6107455

Display Launch Outcome

- From the color labels, we can easily see
  - KSC LC-39A has a rather higher success rate
  - Whereas CCAFS LC-40 and CCAFS SLC-40 have much lower rate

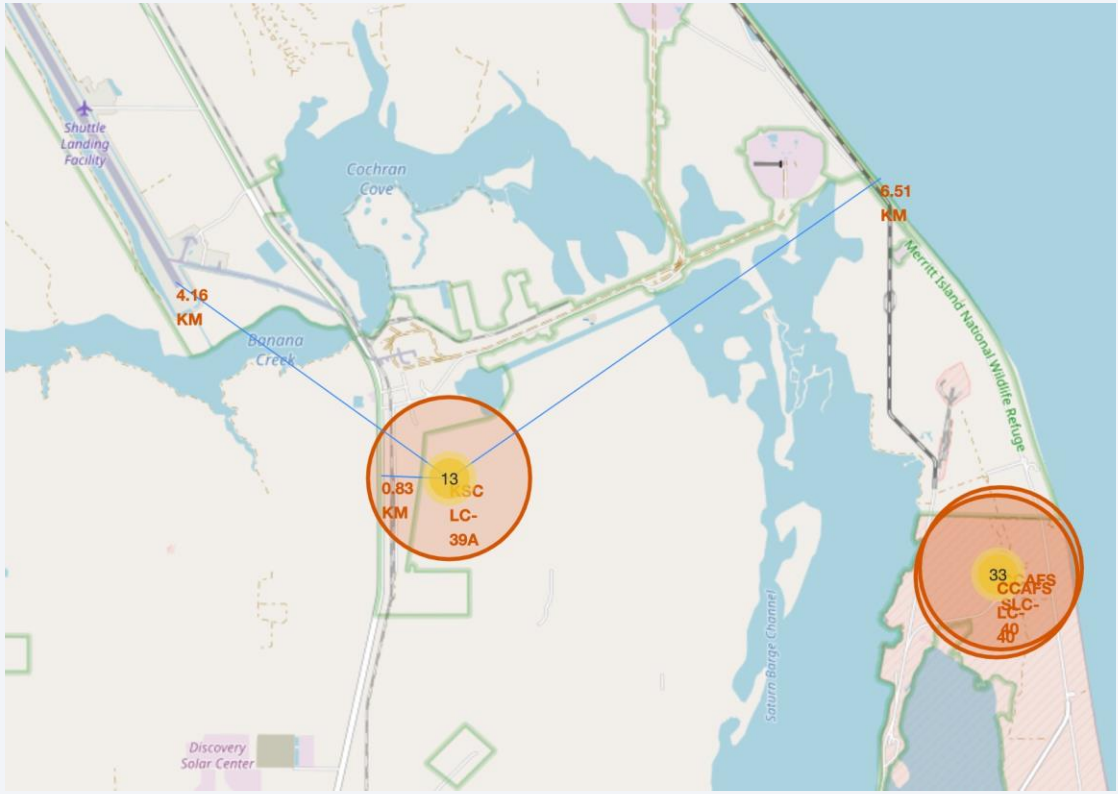






Distance to Proximities

- The distance from KSC LC-39A to the nearest shuttle landing facility is about 4.16 km
- The distance from KSC LC-39A to the nearest highway is less than 1 km
- The distance from KSC LC-39A to the coastline is around 6.5 km.

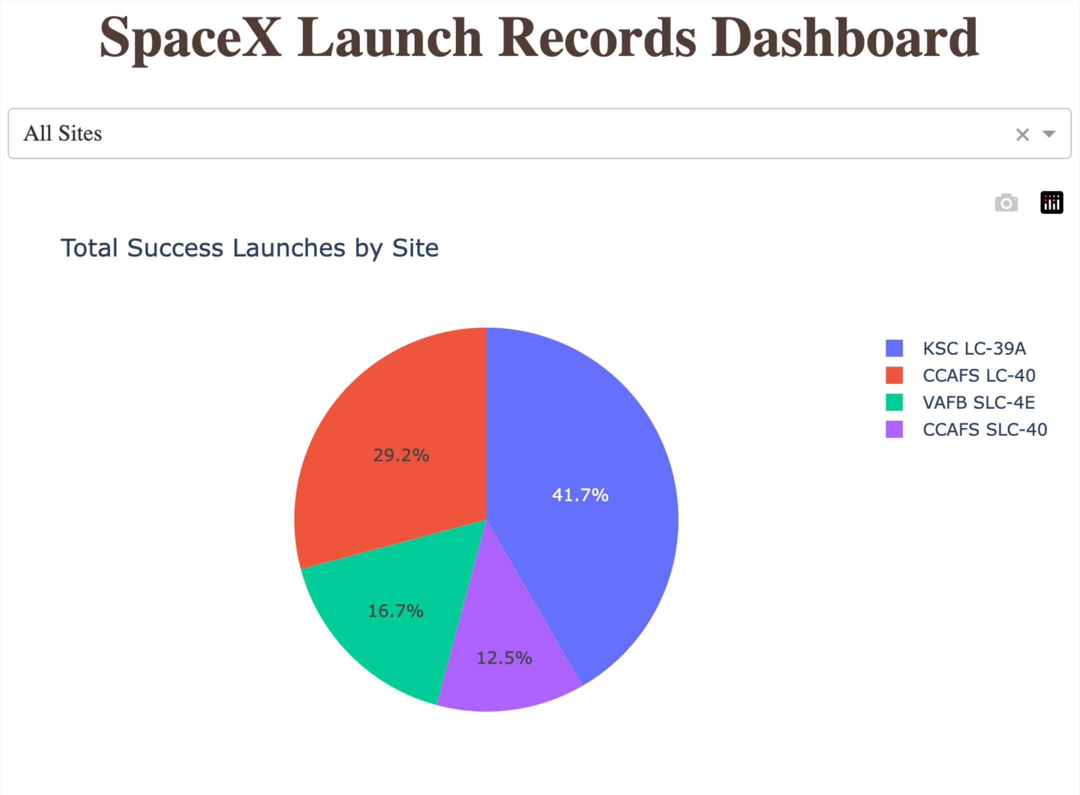


Section 4

Build a Dashboard with Plotly Dash

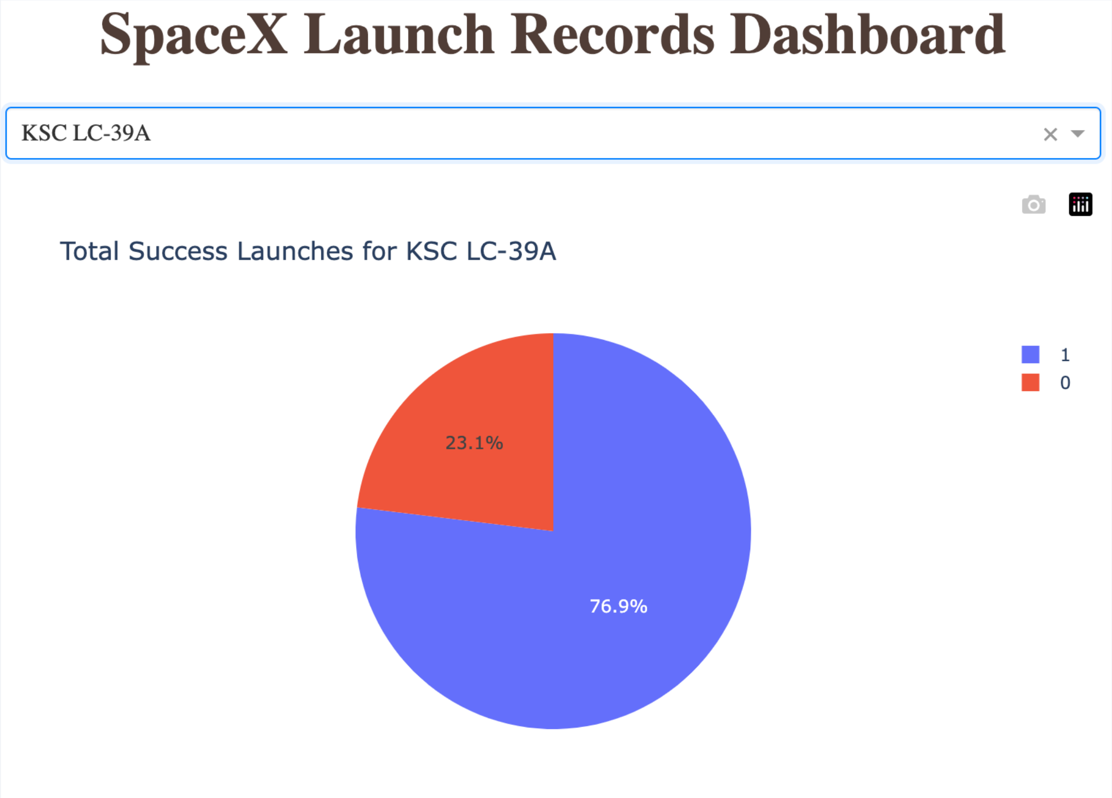
Total Success Launches for All Sites

- Total Success Launches for All Sites is
  - CCAFS LC-40: 29.2%
  - VAFB SLC-4E: 16.7%
  - KSC LC-39A: 41.7%
  - CCAFS SLC-40: 12.5%



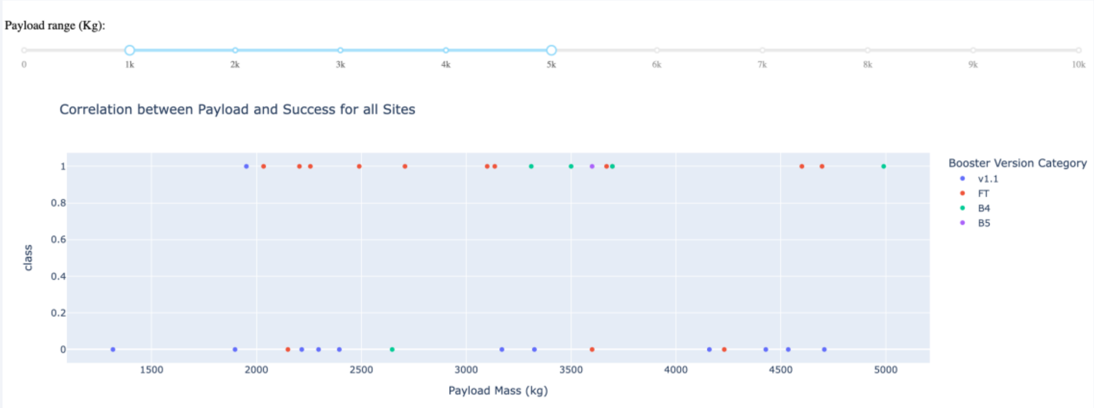
Success Ratio for KSC LC-39A

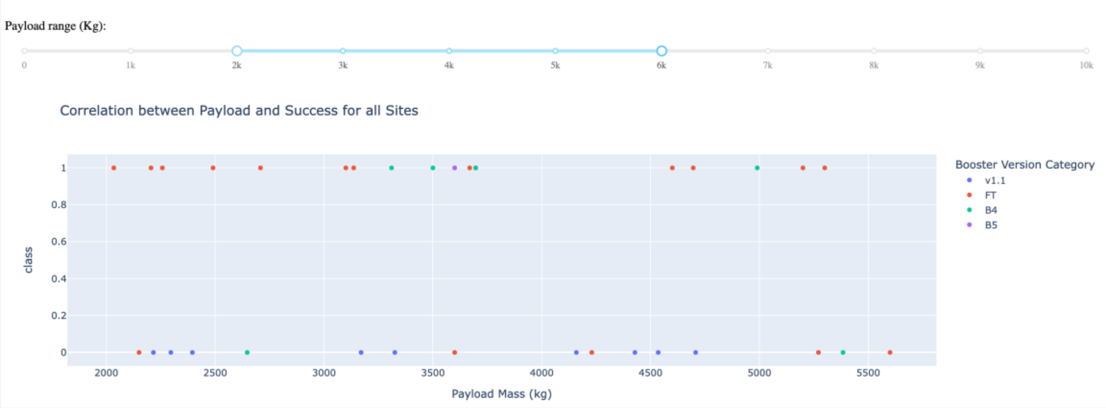
- The launch site with highest launch success ratio is KSC LC-39A with success rate of 76.9%.



Correlation Between Payload and Success

- Payload range in [3000, 4000] has the largest success rate
- Booster version of FT has the largest success rate



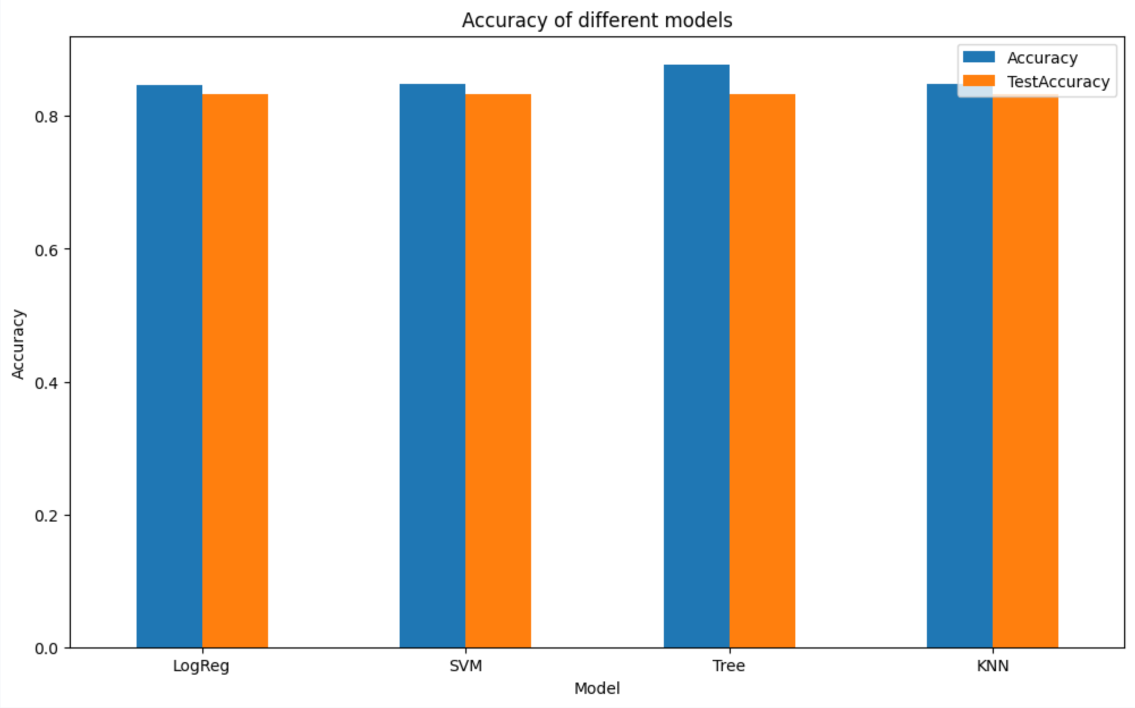


Section 5

Predictive Analysis (Classification)

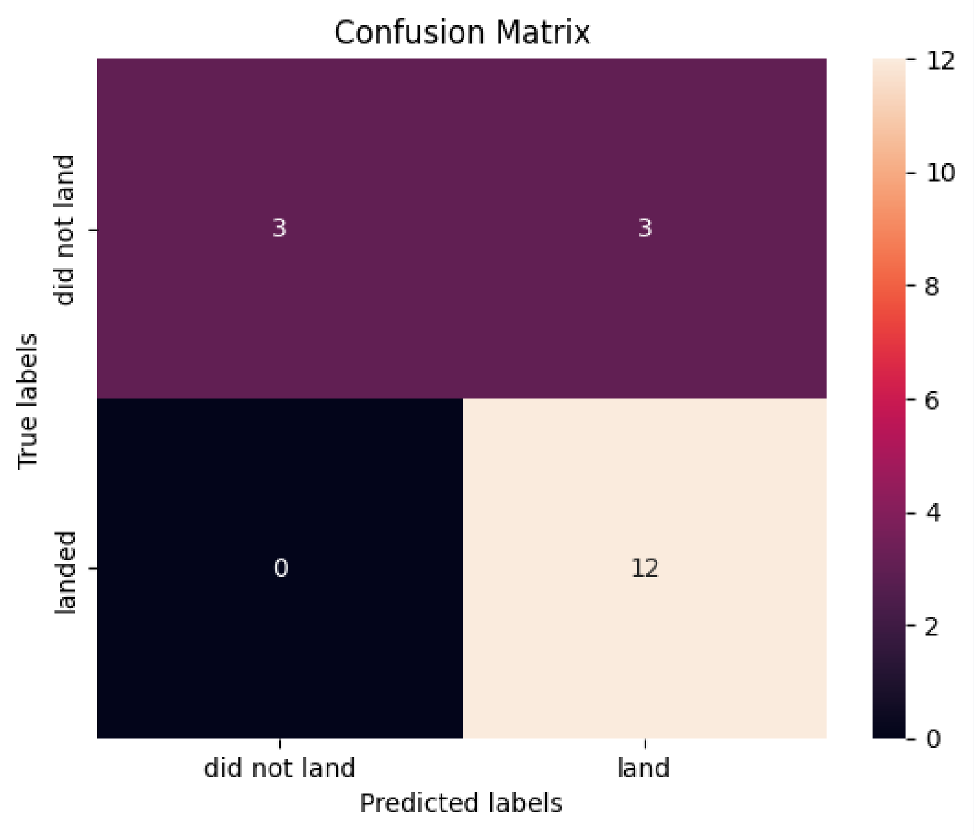
Classification Accuracy

- Decision Tree model has the highest classification accuracy (83.33%)



Confusion Matrix

- Given by the confusion matrix, there is not False Positive prediction



Conclusions

- The dataset has 90 rows of data, with 83 columns. With 80/20 split, we have 72 rows of training data and 18 rows of testing data.
- By GridSearchCV, we trained four models which have all best performance on test dataset
- By compared with other models, the Decision Tree outperform the others
- Generally, the outcome might fluctuate due to lack of amount of datasets

Thank you!