IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Duy Hoang
Feb 28, 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection
  - Data wrangling
  - EDA with data visualization
  - EDA with SQL
  - Building an interactive map with Folium
  - Building a Dashboard with Plotly Dash
  - Predictive analysis (Classification)
- Summary of all results
  - Exploratory data analysis results
  - Interactive analytics
  - Predictive analysis

Introduction

- Project background and context
    - 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
- Problems you want to find answers
    - Predicting if the first stage of the SpaceX Falcon 9 rocket will land successfully

Section 1

Methodology

Methodology

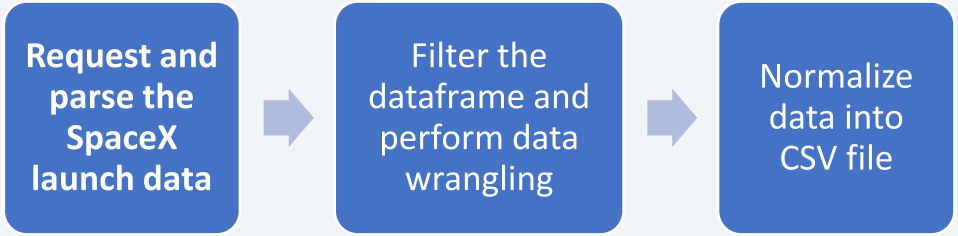
Executive Summary

- Data collection methodology:
    - Data Collected from SpaceX API and Wikipedia ([Falcon 9 and Falcon Heavy Launches](#))
- Perform data wrangling
    - The training labels are derived from the outcomes of the booster landing, indicating either success or failure.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Find best Hyperparameter for SVM, Classification Trees and Logistic Regression

Data Collection

- Data collection process includes a combination of API requests to SpaceX API and web scraping data "Falcon 9 and Falcon Heavy Launches Records" from Wikipedia
  - SpaceX API Data Columns: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
  - Wikipedia Data Columns: Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

SpaceX API

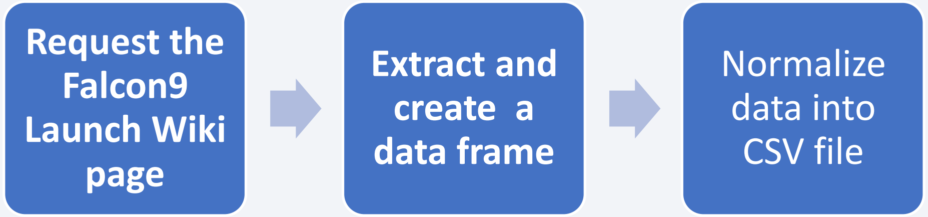


Request and parse the SpaceX launch data

Filter the dataframe and perform data wrangling

Normalize data into CSV file

Wikipedia



Request the Falcon9 Launch Wiki page

Extract and create a data frame

Normalize data into CSV file

Data Collection – SpaceX API

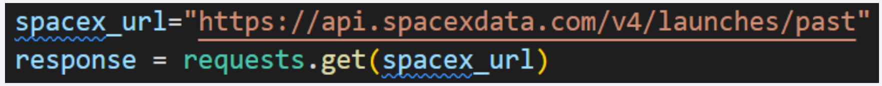
1. Request and parse the SpaceX launch data:

2. json result into a data frame

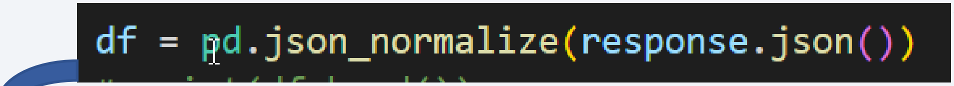
3. Using custom functions to clean data

4. Create data frame

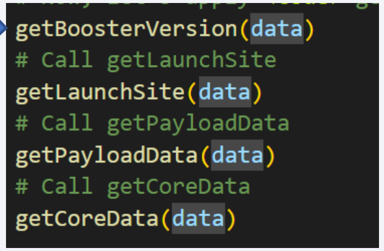
5. Filtering data frame and exporting to a CSV



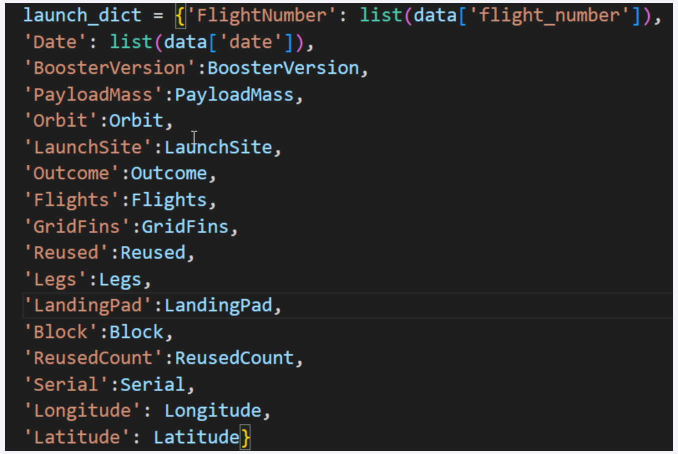
```
spacex_url="https://api.spacexdata.com/v4/launches/past"
response = requests.get(spacex_url)
```



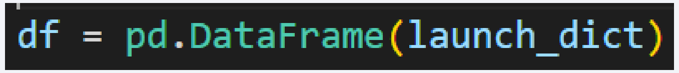
```
df = pd.json_normalize(response.json())
```



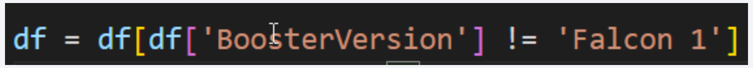
```
getBoosterVersion(data)
# Call getLaunchSite
getLaunchSite(data)
# Call getPayloadData
getPayloadData(data)
# Call getCoreData
getCoreData(data)
```



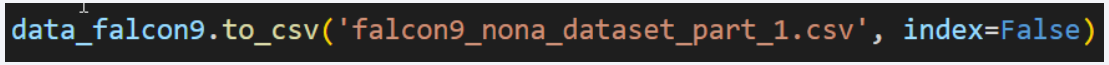
```
launch_dict = {'FlightNumber': list(data['flight_number']),
'Date': list(data['date']),
'BoosterVersion':BoosterVersion,
'PayloadMass':PayloadMass,
'Orbit':Orbit,
'LaunchSite':LaunchSite,
'Outcome':Outcome,
'Flights':Flights,
'GridFins':GridFins,
'Reused':Reused,
'Legs':Legs,
'LandingPad':LandingPad,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



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



```
df = df[df['BoosterVersion'] != 'Falcon 1']
```



```
data_falcon9.to_csv('falcon9_nona_dataset_part_1.csv', index=False)
```

GitHub URL: https://github.com/duy-genashtim/Data-Science-and-MachineLearning-Capstone/blob/main/01-spacex-data-collection-api.py

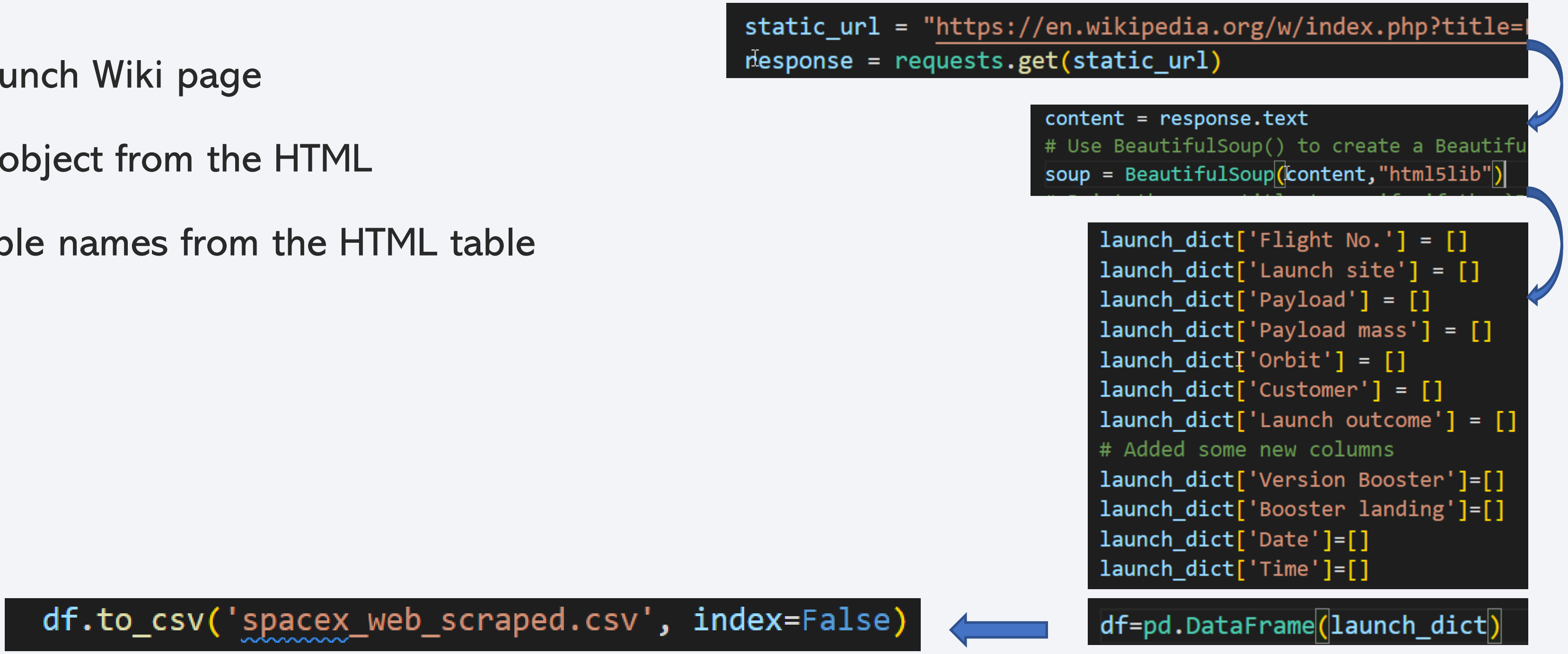
Data Collection - Scraping

1. Request the Falcon9 Launch Wiki page

2. Create a BeautifulSoup object from the HTML

3. Extract all column/variable names from the HTML table

4. exporting to a CSV



```
static_url = "https://en.wikipedia.org/w/index.php?title=
response = requests.get(static_url)
```

```
content = response.text
# Use BeautifulSoup() to create a Beautifu
soup = BeautifulSoup(content,"html5lib")
```

```
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']=[]
```

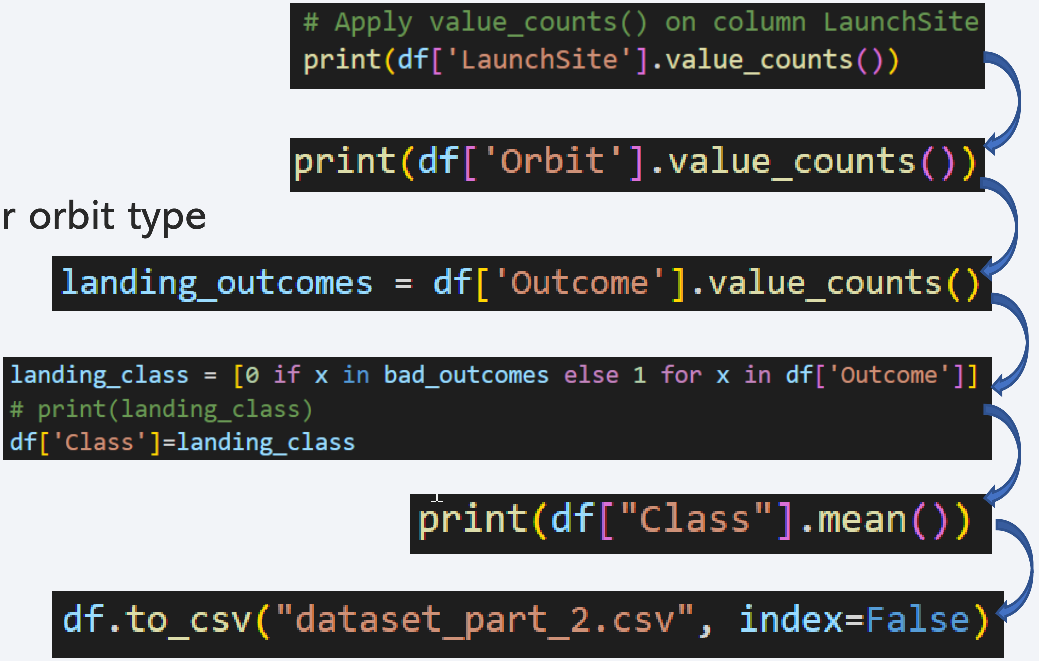
```
df=pd.DataFrame(launch_dict)
```

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

GitHub URL: https://github.com/duy-genashtim/Data-Science-and-MachineLearning-Capstone/blob/main/02-web_scraping.py

Data Wrangling

1. Calculate the number of launches on each site
2. Calculate the number and occurrence of each orbit
3. Calculate the number and occurence of mission outcome per orbit type
4. Create a landing outcome label from Outcome column
5. Determine the success rate
6. Exporting dataset to a CSV



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

```
print(df['Orbit'].value_counts())
```

```
landing_outcomes = df['Outcome'].value_counts()
```

```
landing_class = [0 if x in bad_outcomes else 1 for x in df['Outcome']]
# print(landing_class)
df['Class']=landing_class
```

```
print(df["Class"].mean())
```

```
df.to_csv("dataset_part_2.csv", index=False)
```

GitHub URL: https://github.com/duy-genashtim/Data-Science-and-MachineLearning-Capstone/blob/main/03-data_wrangling.py

EDA with Data Visualization

- Scatter chart:
  - Flight Number vs. Launch Site
  - Payload vs. Launch Site
  - Flight Number vs. Orbit Type
  - Payload vs. Orbit Type

  A scatter plot shows how much one variable is affected by another. The relationship between two variables is called a correlation.
- Bar Chart
  - Success Rate Vs. Orbit Type

  A bar chart is useful for comparing and displaying categorical data with their corresponding values, making it easy to identify patterns and trends in the data.
- Line Chart
  - Year Vs. Success Rate

  A line chart is useful for displaying trends over time or showing the relationship between two variables by representing them as continuous lines.

GitHub URL: https://github.com/duy-genashtim/Data-Science-and-MachineLearning-Capstone/blob/main/04-data_visualization.py

EDA with SQL

- SQL queries (Sqllite)
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'KSC'
  - 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 where the succesful landing outcome in drone ship was acheived
  - List the names of the boosters which have success in ground pad 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 records which will display the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017
  - Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

GitHub URL: https://github.com/duy-genashtim/Data-Science-and-MachineLearning-Capstone/blob/main/05-sqlite_queries.py

Build an Interactive Map with Folium

- Map objects added to folium
  - Circular markers were added to indicate launch sites
  - Red and green marker clusters were added to indicate successful/failed landings at each launch site
  - Line markers were added to indicate the distance of launch site from coastline, nearest city, highway, and railway
  - Text markers indicating the distance of site from said entities

GitHub URL: https://github.com/duy-genashtim/Data-Science-and-MachineLearning-Capstone/blob/main/06-launch_sites_locations_analysis.py

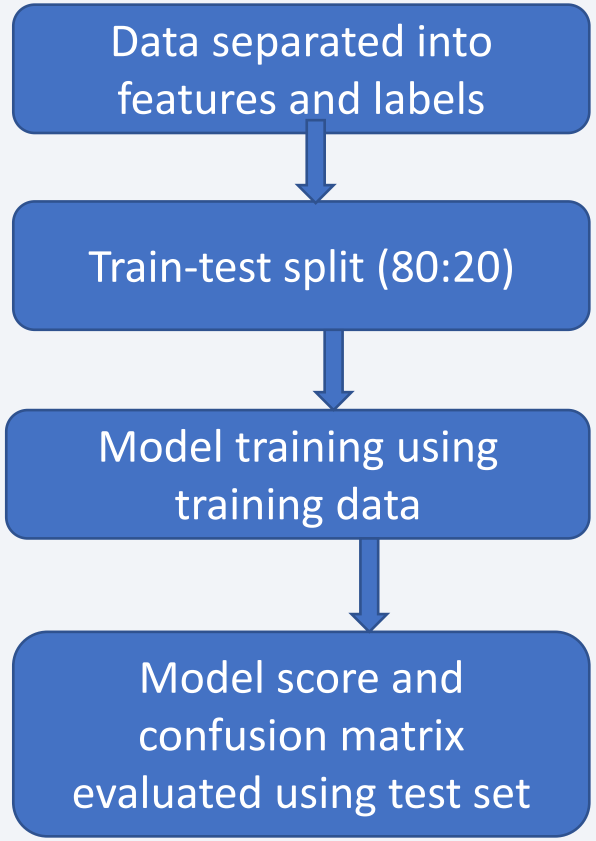
Build a Dashboard with Plotly Dash

- Pie Chart
  - Pie charts are useful for showing the proportion or percentage distribution of categorical data in a visually appealing way.
- Scatter Chart
  - A scatter chart is used to display the correlation between two variables, showing the relationship and pattern in the data.

GitHub URL: https://github.com/duy-genashtim/Data-Science-and-MachineLearning-Capstone/blob/main/07-spacex_dash_app.py

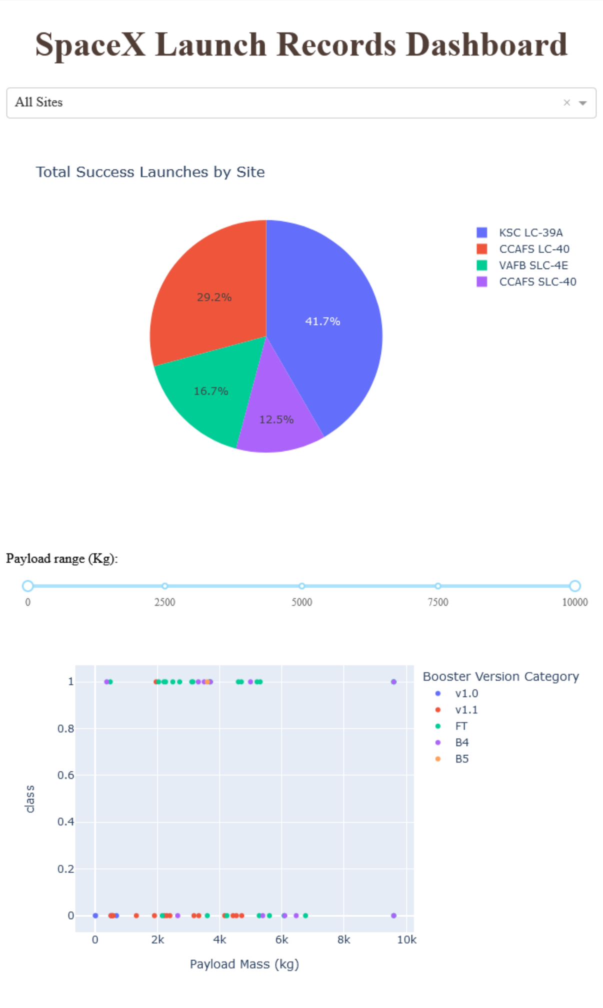
Predictive Analysis (Classification)

- Perform exploratory Data Analysis and determine Training Labels
  - Create a NumPy array from the column Class
  - Standardize the data
  - Split into training data and test data
- Find best Hyperparameter for Support Vector Machine, Decision Tree, K Nearest Neighbors, and Logistic Regression
- Calculate the accuracy for each and get the best one


Data separated into features and labels
Train-test split (80:20)
Model training using training data
Model score and confusion matrix evaluated using test set

GitHub URL: https://github.com/duy-genashtim/Data-Science-and-MachineLearning-Capstone/blob/main/08-machine_learning_prediction.py

Results



SpaceX Launch Records Dashboard

All Sites

Payload range (Kg):

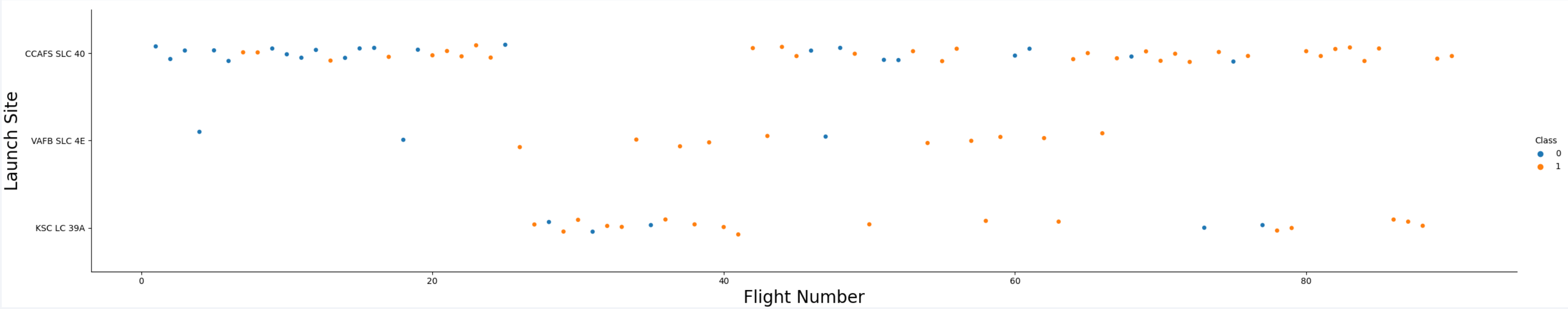
- The left screenshot is a preview of the Dashboard with Plotly Dash
- Exploring Launch sites with payload ranges

Section 2

Insights drawn from EDA

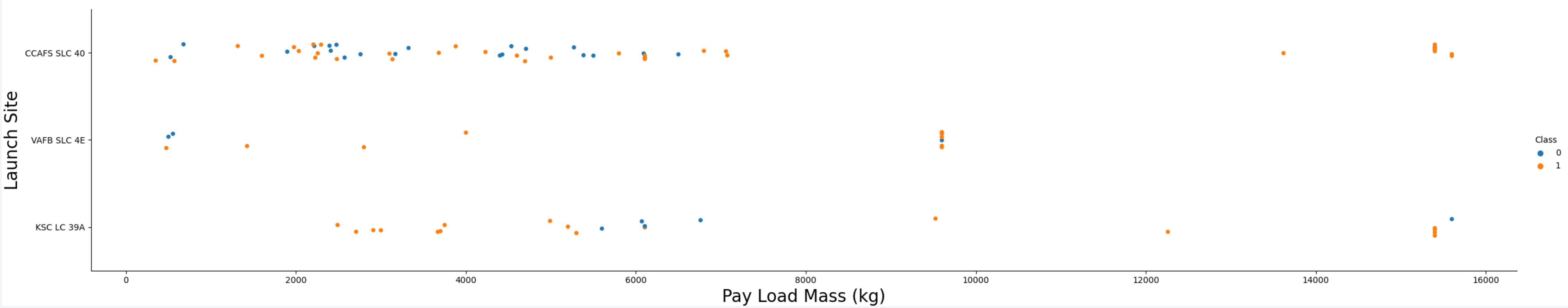
Flight Number vs. Launch Site

- Class 0 represents unsuccessful launch, and Class 1 represents successful launch
- The success rate increased as the number of flights increased



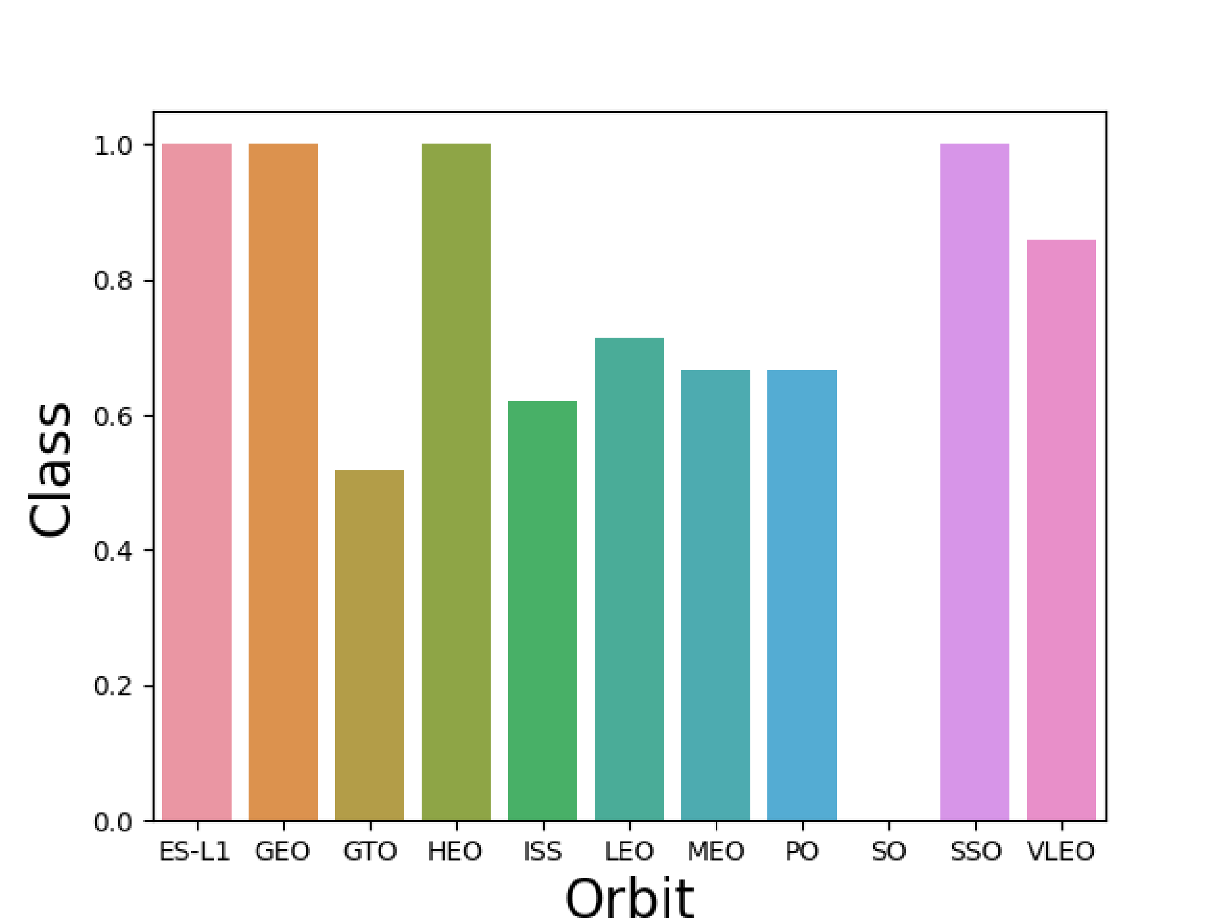
Payload vs. Launch Site

- Class 0 represents unsuccessful launch, and Class 1 represents successful launch
- The greater the payload mass for Launch Site CCAFS SLC 40 the higher the success rate for the Rocket. There is not quite a clear pattern to be found using this visualization to make a decision if the Launch Site is dependent on Pay Load Mass for a success launch.



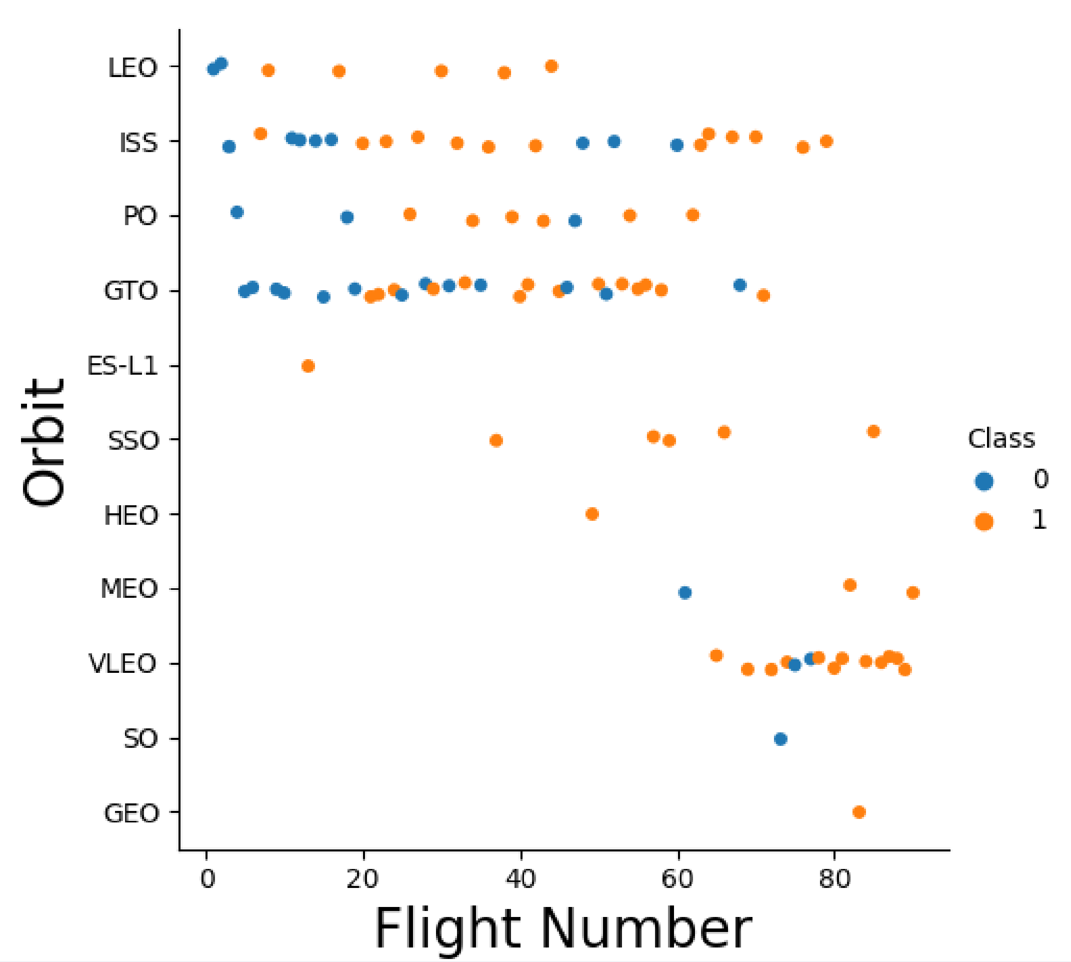
Success Rate vs. Orbit Type

- Orbit types SSO, HEO, GEO, and ES-L1 have the highest success rates (100%)
- Success rate for SO orbit is the lowest



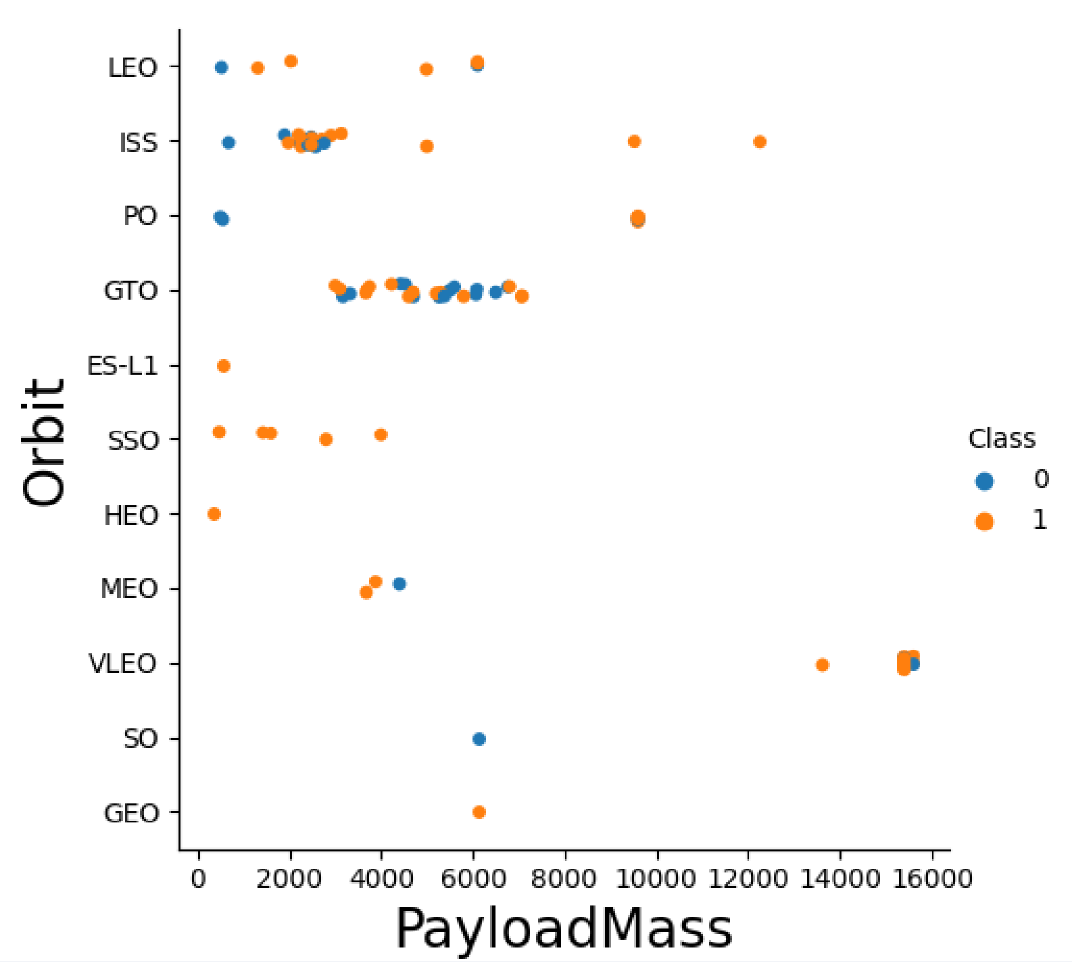
Flight Number vs. Orbit Type

- Class 0 represents unsuccessful launch, and Class 1 represents successful launch
- 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.



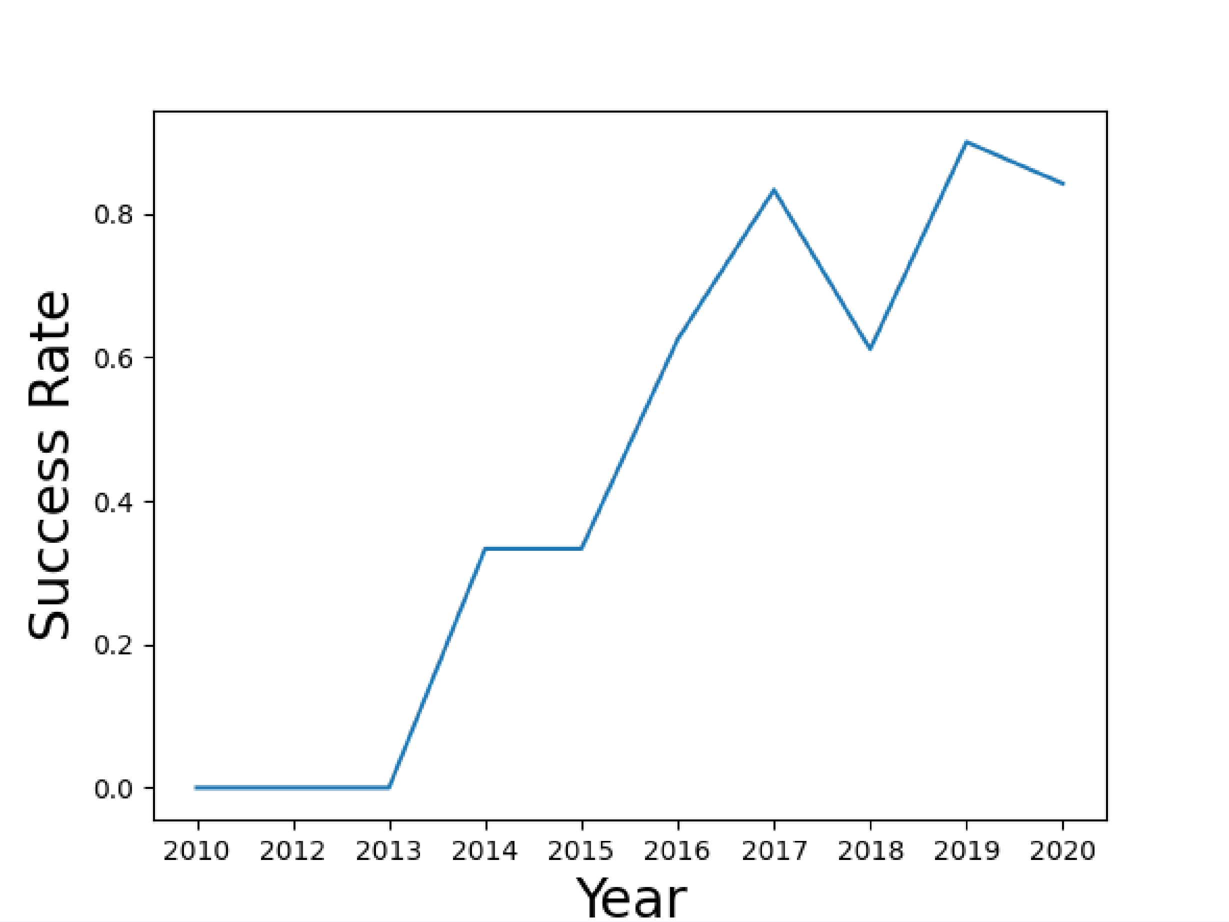
Payload vs. Orbit Type

- Class 0 represents unsuccessful launch, and Class 1 represents successful launch
- 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

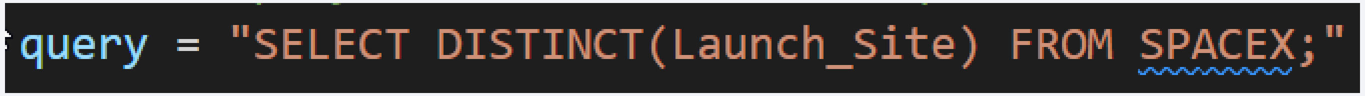
- The sucess rate since 2013 kept increasing till 2020
- The rate decreased slightly in 2018



All Launch Site Names

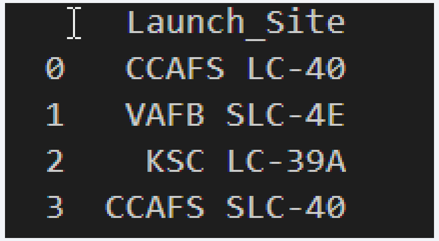
Find the names of the unique launch sites

- Query



```
query = "SELECT DISTINCT(Launch_Site) FROM SPACEX;"
```

- Result

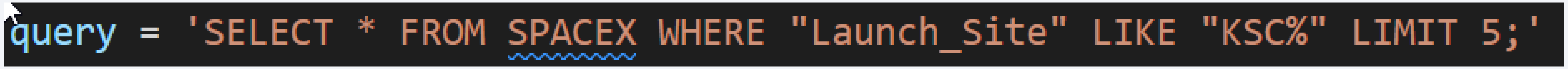


```
     Launch_Site
0    CCAFS LC-40
1    VAFB SLC-4E
2     KSC LC-39A
3   CCAFS SLC-40
```

Launch Site Names Begin with 'KSC'

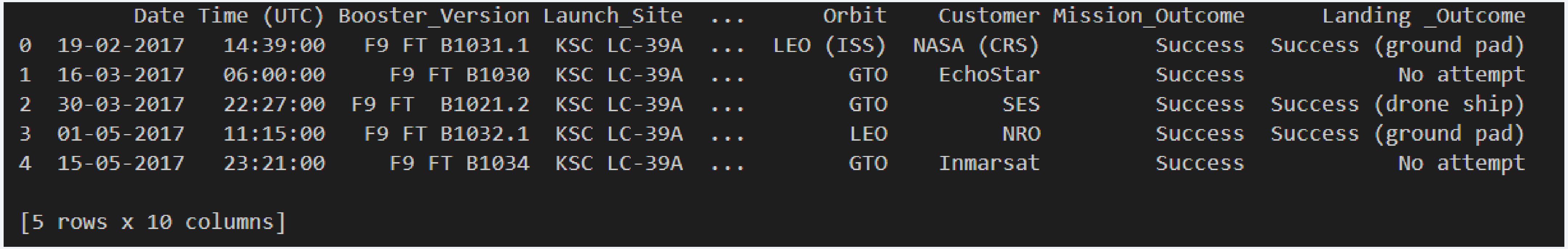
Find 5 records where launch sites’ names start with `KSC`

- Query



```
query = 'SELECT * FROM SPACEX WHERE "Launch_Site" LIKE "KSC%" LIMIT 5;'
```

- Result



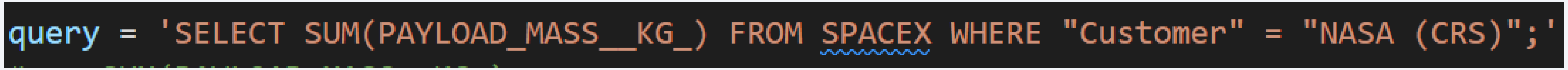
```
         Date Time (UTC) Booster_Version Launch_Site  ...      Orbit     Customer Mission_Outcome      Landing _Outcome
0  19-02-2017   14:39:00   F9 FT B1031.1  KSC LC-39A  ...  LEO (ISS)  NASA (CRS)         Success  Success (ground pad)
1  16-03-2017   06:00:00     F9 FT B1030  KSC LC-39A  ...        GTO    EchoStar         Success            No attempt
2  30-03-2017   22:27:00  F9 FT  B1021.2  KSC LC-39A  ...        GTO         SES         Success  Success (drone ship)
3  01-05-2017   11:15:00   F9 FT B1032.1  KSC LC-39A  ...        LEO         NRO         Success  Success (ground pad)
4  15-05-2017   23:21:00     F9 FT B1034  KSC LC-39A  ...        GTO    Inmarsat         Success            No attempt

[5 rows x 10 columns]
```

Total Payload Mass

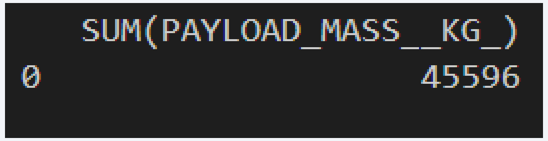
Calculate the total payload carried by boosters from NASA

- Query



```
query = 'SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEX WHERE "Customer" = "NASA (CRS)";'
```

- Result

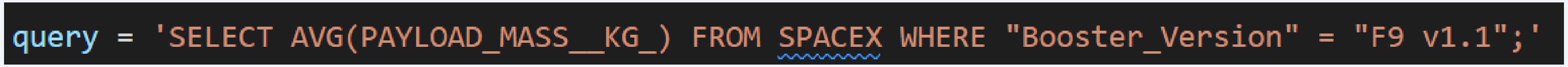


```
   SUM(PAYLOAD_MASS__KG_)
0                   45596
```

Average Payload Mass by F9 v1.1

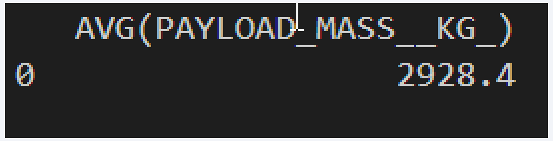
Calculate the average payload mass carried by booster version F9 v1.1

- Query



```
query = 'SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEX WHERE "Booster_Version" = "F9 v1.1";'
```

- Result

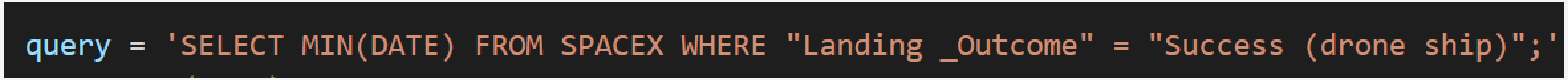


```
   AVG(PAYLOAD_MASS__KG_)
0                  2928.4
```

First Successful Ground Landing Date

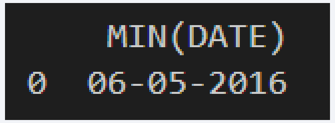
Find the dates of the first successful landing outcome on drone ship.

- Query



```
query = 'SELECT MIN(DATE) FROM SPACEX WHERE "Landing _Outcome" = "Success (drone ship)";'
```

- Result

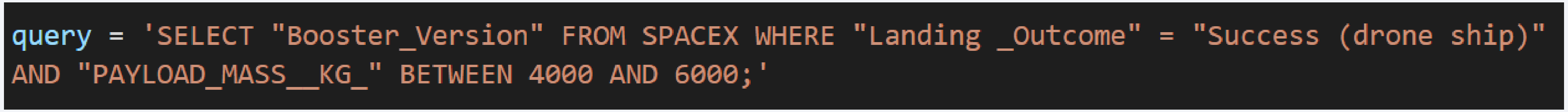


```
    MIN(DATE)
0  06-05-2016
```

Successful Drone Ship Landing with Payload between 4000 and 6000

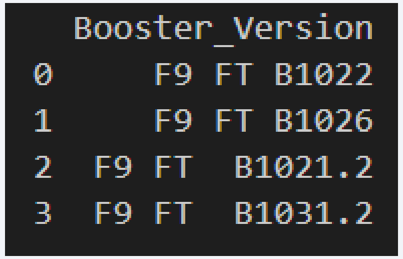
List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

- Query



```
query = 'SELECT "Booster_Version" FROM SPACEX WHERE "Landing _Outcome" = "Success (drone ship)"
AND "PAYLOAD_MASS__KG_" BETWEEN 4000 AND 6000;'
```

- Result

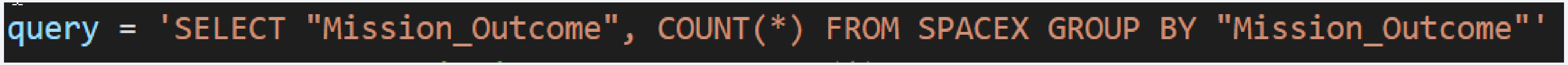


```
  Booster_Version
0     F9 FT B1022
1     F9 FT B1026
2  F9 FT  B1021.2
3  F9 FT  B1031.2
```

Total Number of Successful and Failure Mission Outcomes

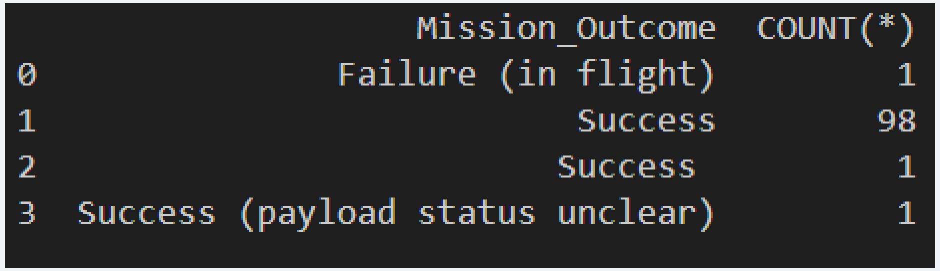
Calculate the total number of successful and failure mission outcomes

- Query



```
query = 'SELECT "Mission_Outcome", COUNT(*) FROM SPACEX GROUP BY "Mission_Outcome"'
```

- Result

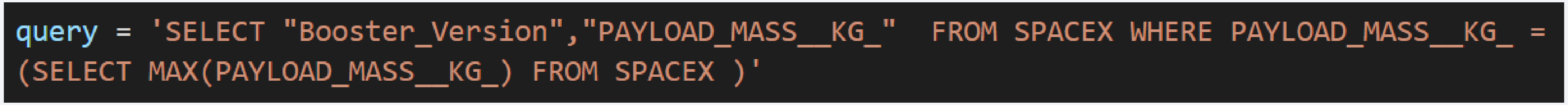


```
                    Mission_Outcome  COUNT(*)
0               Failure (in flight)         1
1                           Success        98
2                          Success          1
3  Success (payload status unclear)         1
```

Boosters Carried Maximum Payload

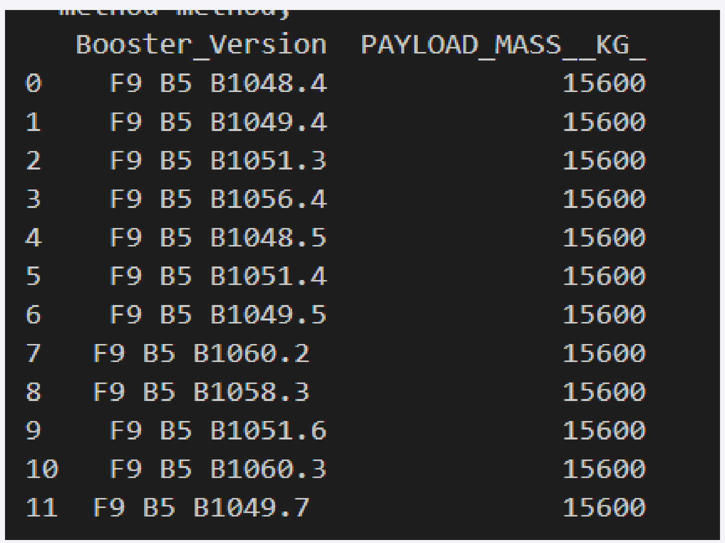
List the names of the booster which have carried the maximum payload mass

- Query



```
query = 'SELECT "Booster_Version","PAYLOAD_MASS__KG_"  FROM SPACEX WHERE PAYLOAD_MASS__KG_ =
(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEX )'
```

- Result

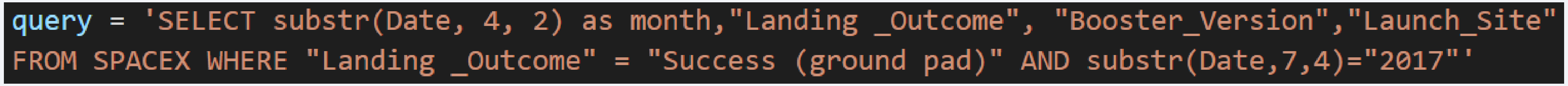


```
   method method)
   Booster_Version  PAYLOAD_MASS__KG_
0    F9 B5 B1048.4              15600
1    F9 B5 B1049.4              15600
2    F9 B5 B1051.3              15600
3    F9 B5 B1056.4              15600
4    F9 B5 B1048.5              15600
5    F9 B5 B1051.4              15600
6    F9 B5 B1049.5              15600
7   F9 B5 B1060.2               15600
8   F9 B5 B1058.3               15600
9    F9 B5 B1051.6              15600
10   F9 B5 B1060.3              15600
11  F9 B5 B1049.7               15600
```

2015 Launch Records

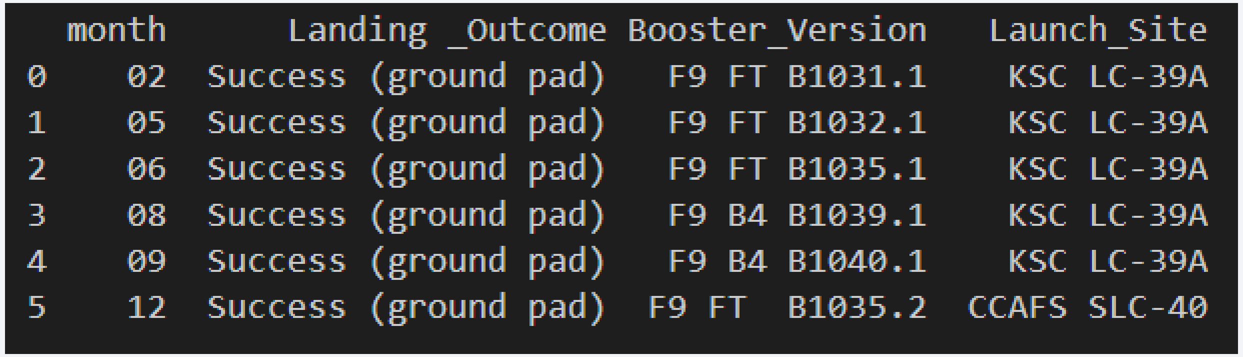
List the records which will display the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017

- Query



```
query = 'SELECT substr(Date, 4, 2) as month,"Landing _Outcome", "Booster_Version","Launch_Site"
FROM SPACEX WHERE "Landing _Outcome" = "Success (ground pad)" AND substr(Date,7,4)="2017"'
```

- Result

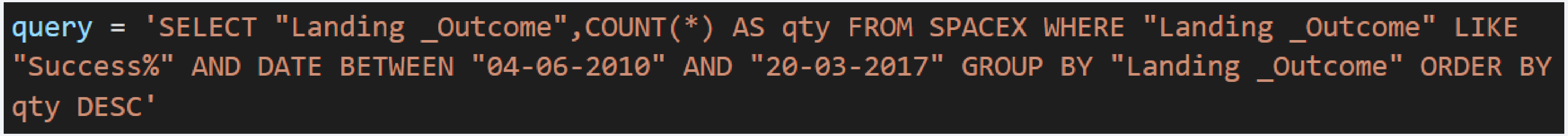


```
  month      Landing _Outcome Booster_Version   Launch_Site
0    02  Success (ground pad)   F9 FT B1031.1   KSC LC-39A
1    05  Success (ground pad)   F9 FT B1032.1   KSC LC-39A
2    06  Success (ground pad)   F9 FT B1035.1   KSC LC-39A
3    08  Success (ground pad)   F9 B4 B1039.1   KSC LC-39A
4    09  Success (ground pad)   F9 B4 B1040.1   KSC LC-39A
5    12  Success (ground pad)  F9 FT  B1035.2 CCAFS SLC-40
```

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

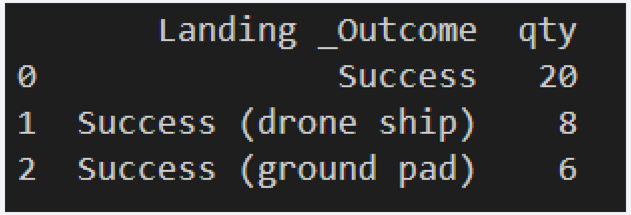
Rank the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order

- Query



```
query = 'SELECT "Landing _Outcome",COUNT(*) AS qty FROM SPACEX WHERE "Landing _Outcome" LIKE
"Success%" AND DATE BETWEEN "04-06-2010" AND "20-03-2017" GROUP BY "Landing _Outcome" ORDER BY
qty DESC'
```

- Result

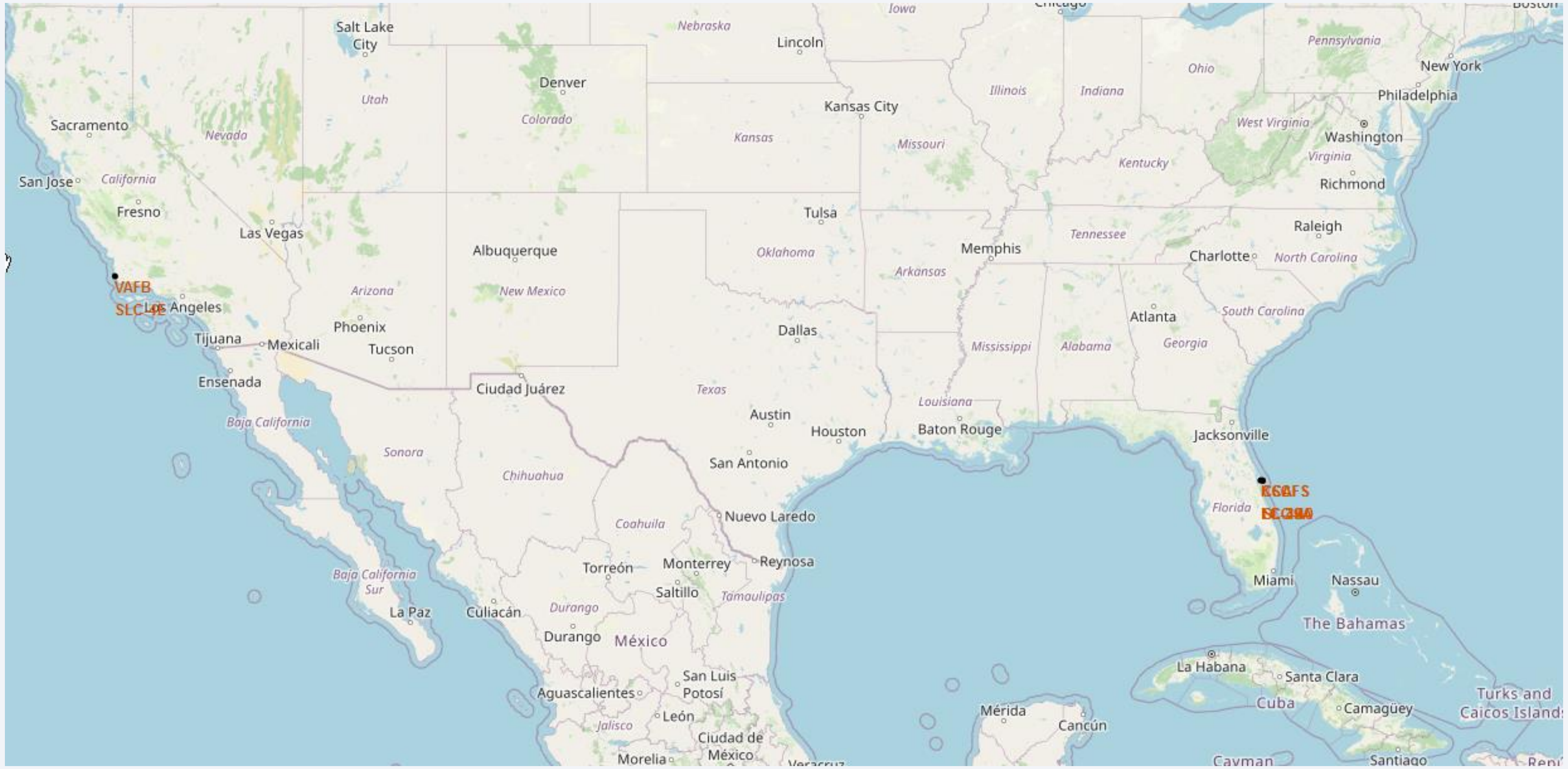


```
       Landing _Outcome  qty
0               Success   20
1  Success (drone ship)    8
2  Success (ground pad)    6
```

Section 3

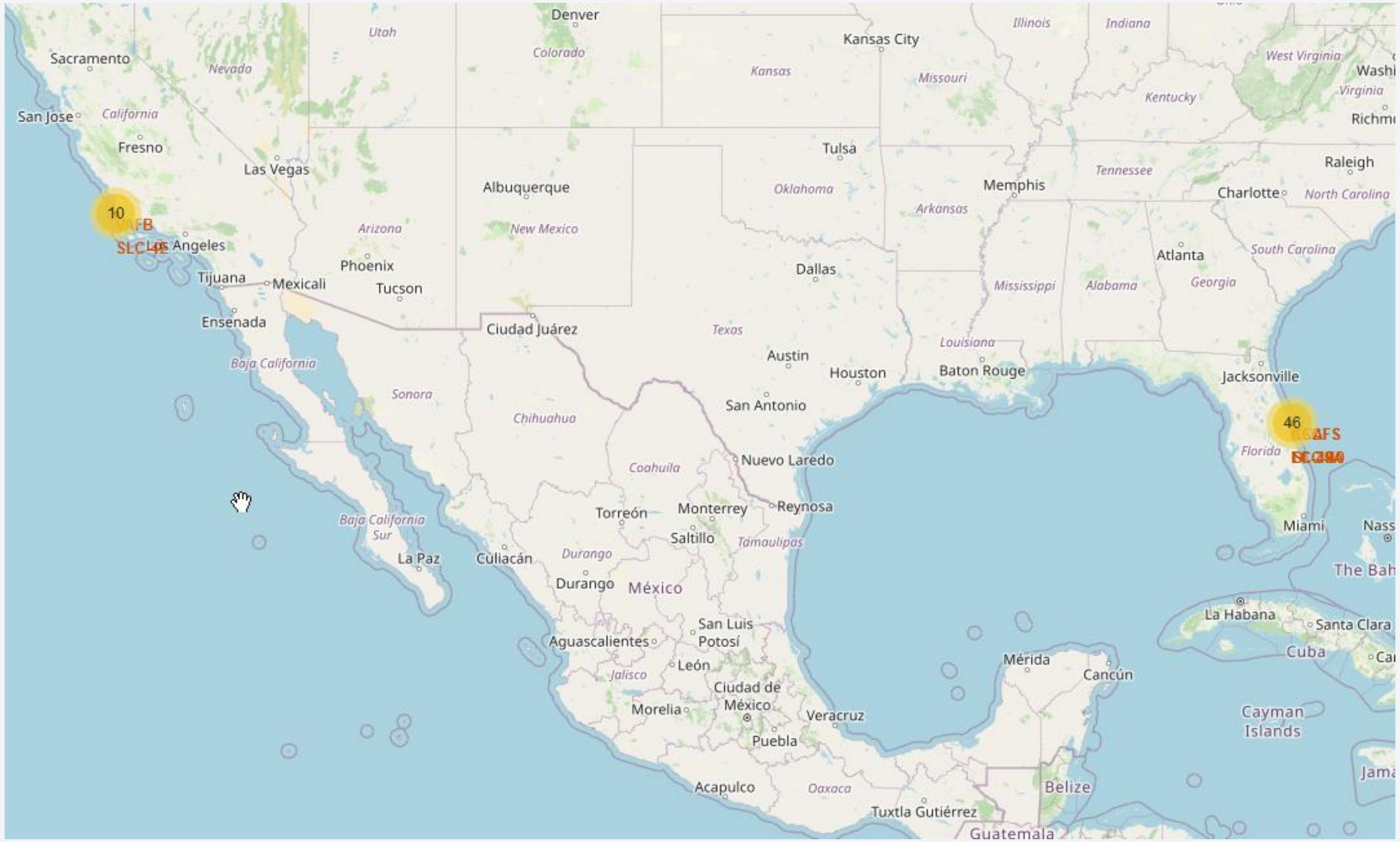
Launch Sites Proximities Analysis

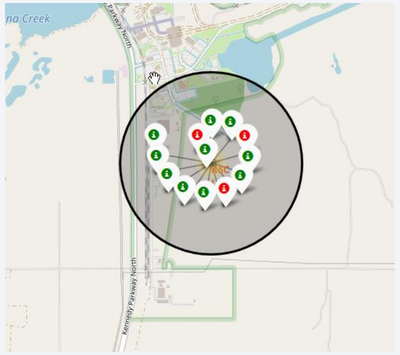
All Launch Sites



- We can see that the SpaceX launch sites are in the United States of America coasts. Florida and California

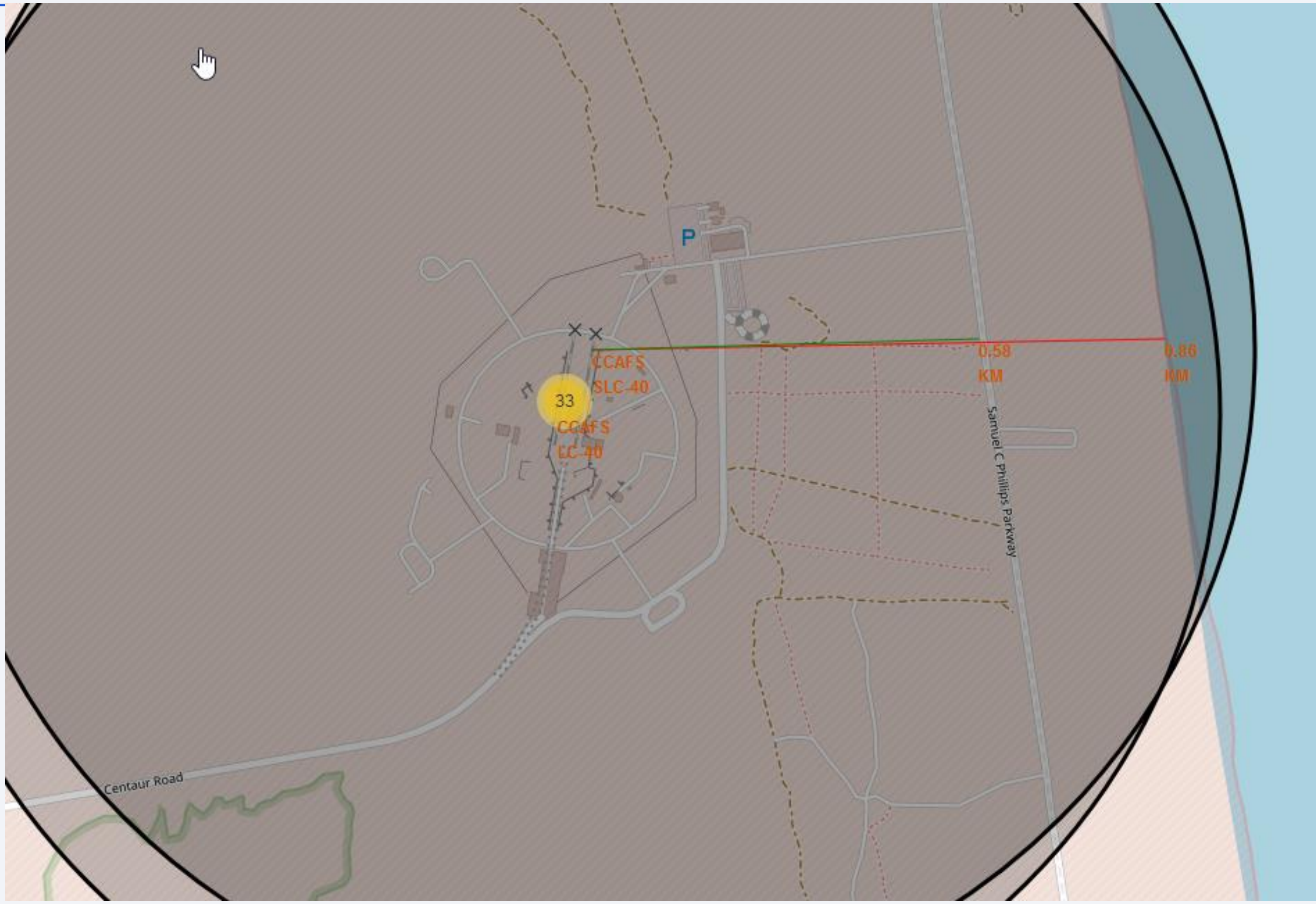
Success/Failed Launches for each site on the map





By clicking on the marker clusters, successful landing (green) or failed landing (red) are displayed.

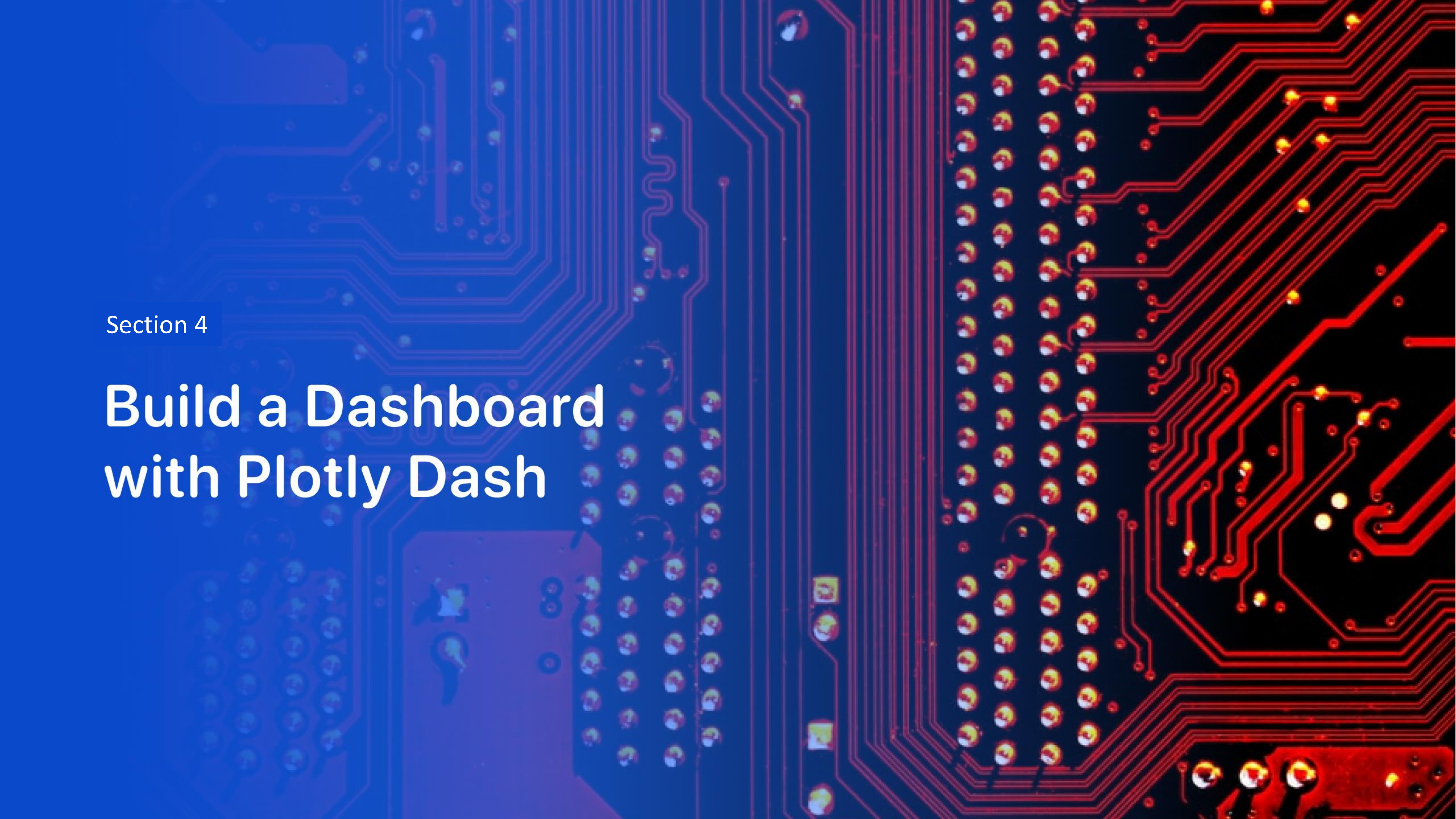
Proximities of Launch Sites



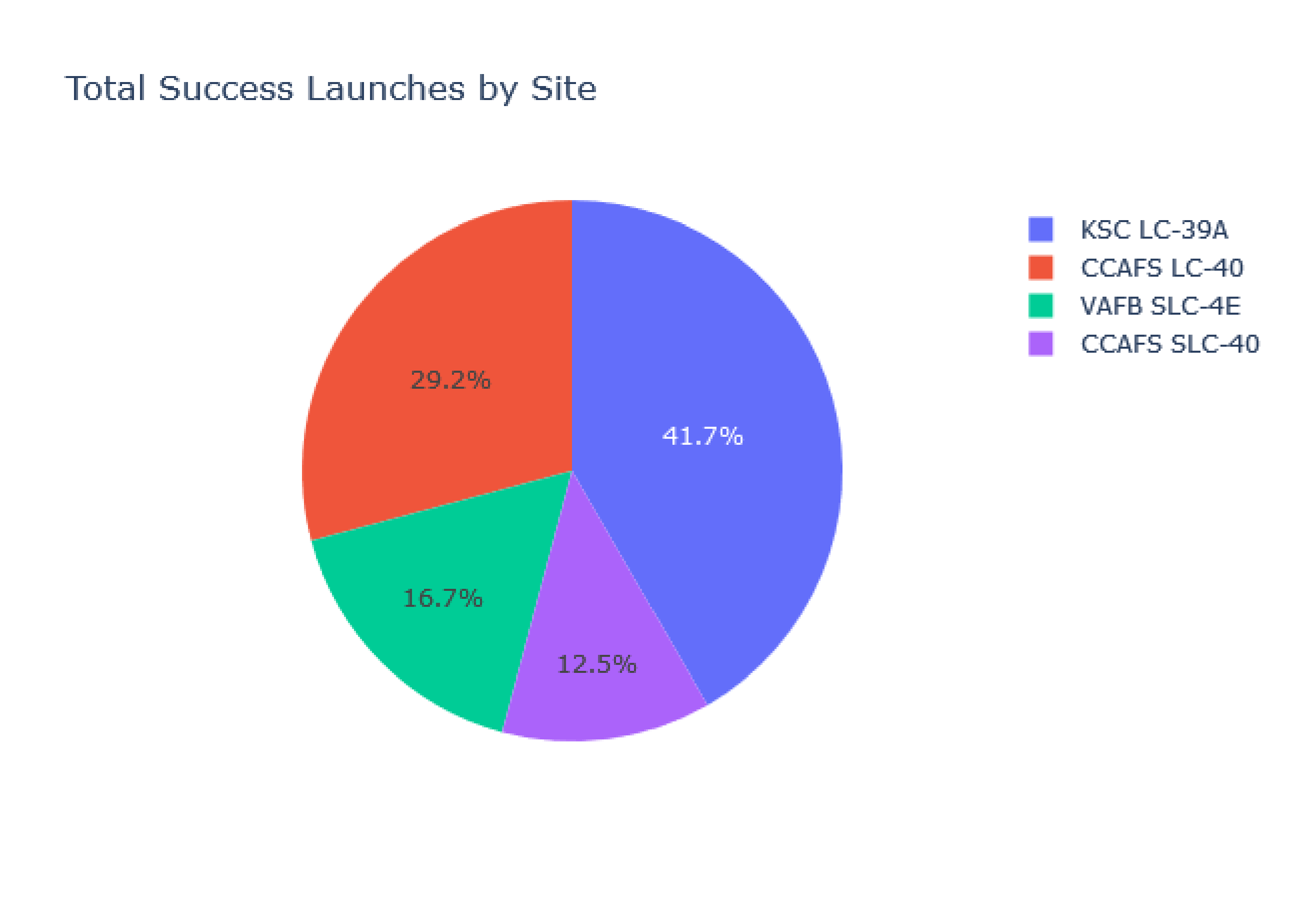
- We can see that the launch site is 0.86KM away from the coastline.

Section 4

Build a Dashboard with Plotly Dash

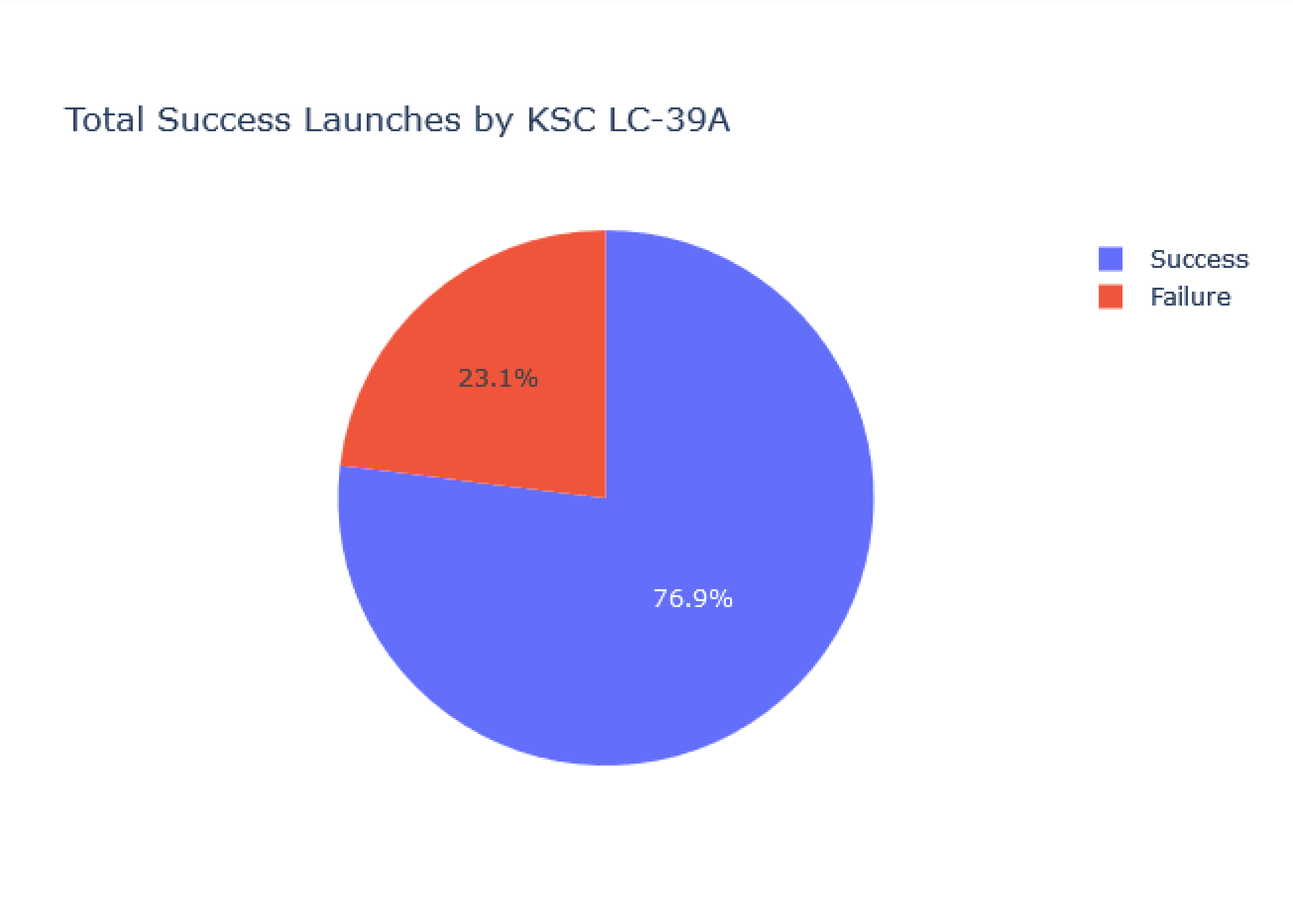


Total Success Launches by All Sites



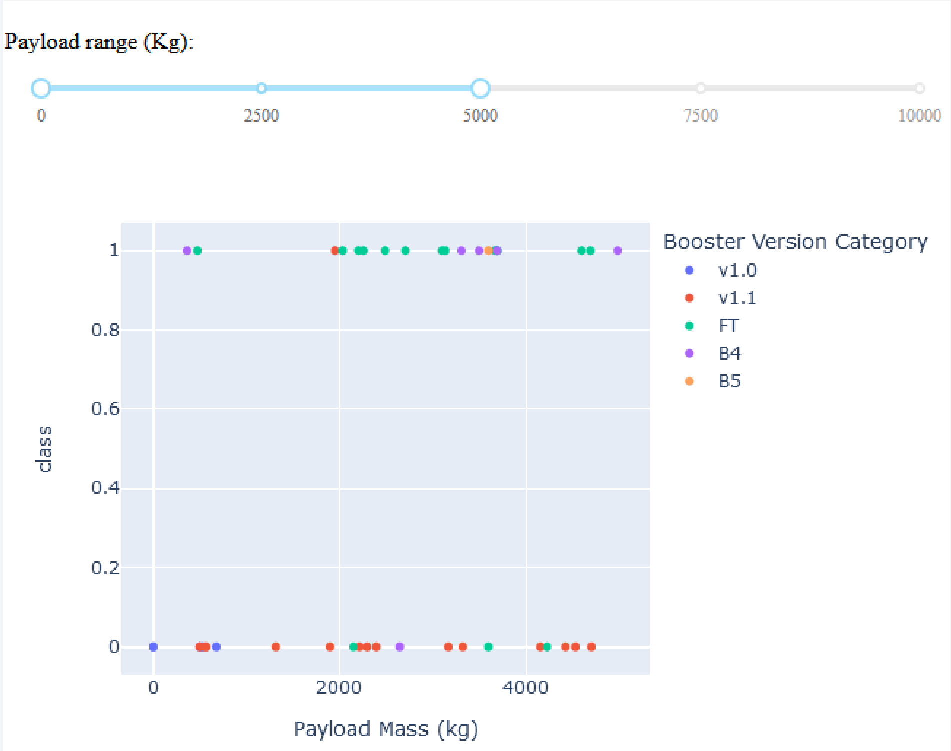
- We can see that KSC LC-39A had the most successful launches from all the sites

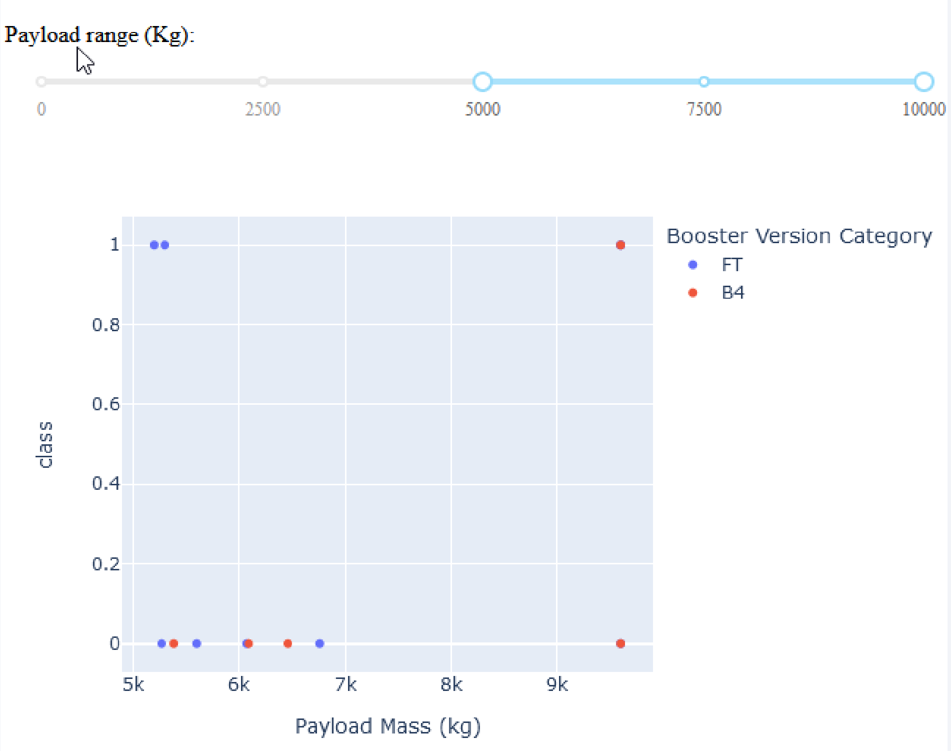
Launch Site with Highest Launch Success Ratio



KSLC-39A has the highest success rate with 10 landing successes (76.9%) and 3 landing failures (23.1%).

Payload vs. Launch Outcome Scatter Plot for All Sites



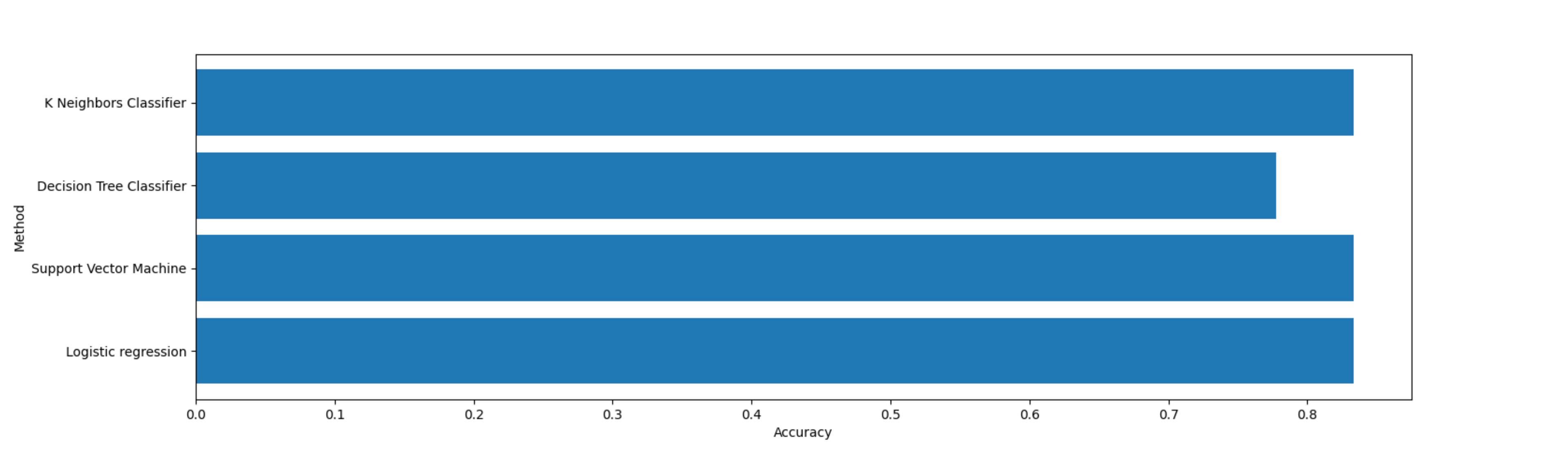


- These figures show that the launch success rate (class 1) for low weighted payloads(0-5000 kg) is higher than that of heavy-weighted payloads(5000-10000 kg).

Section 5

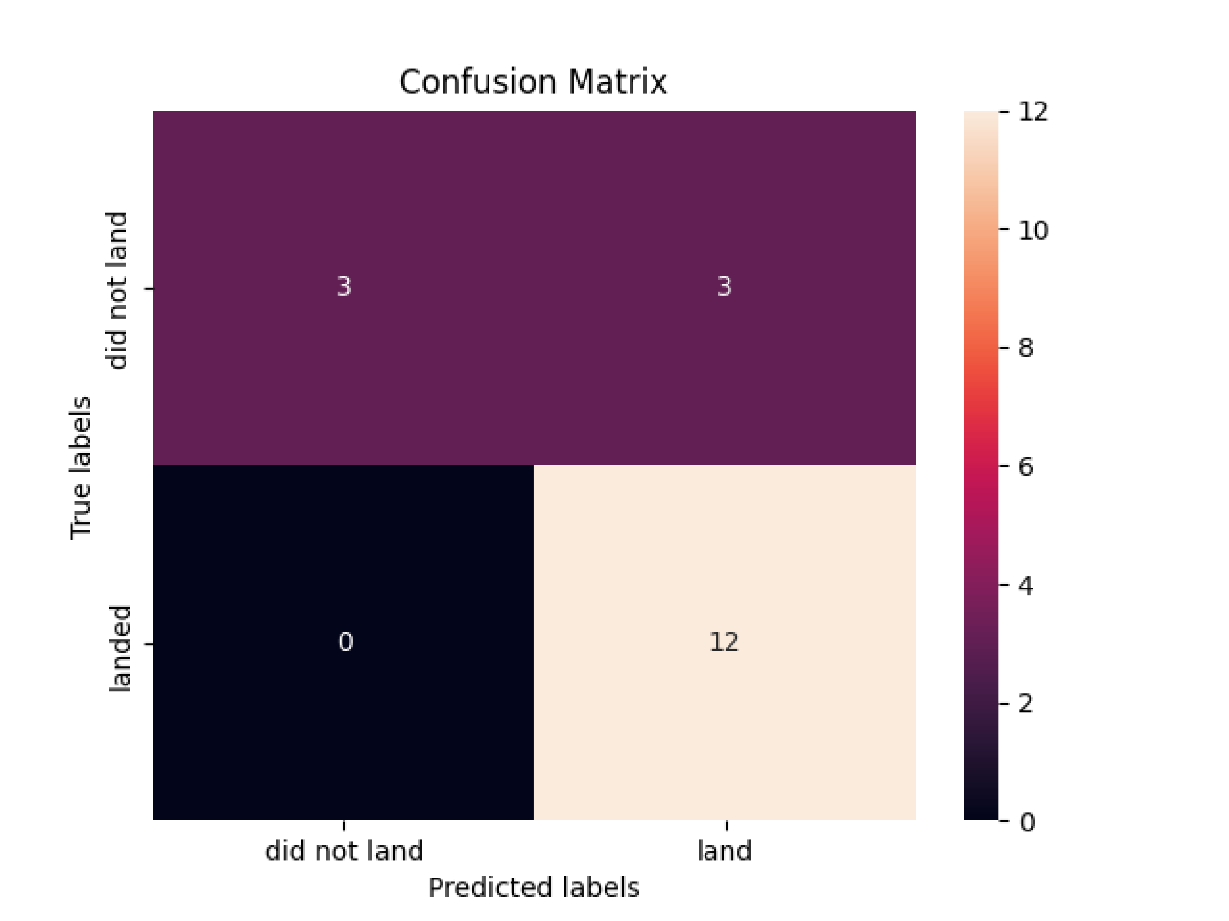
Predictive Analysis (Classification)

Classification Accuracy



- Logistic Regression, SVM, and KNN classifiers can be used since they all have the same accuracy

Confusion Matrix



- The models predicted 12 successful landings when the true label was successful and 3 failed landings when the true label was failure. But there were also 3 predictions that said successful landings when the true label was failure (false positive).
- Overall, these models predict successful landings.

Conclusions

- As the number of flights increased, the success rate increased, and recently it has exceeded 80%
- Orbit GEO, HEO, SSO, and ES-L1 have the highest success rate (100%)
- Ideal rocket launch sites are near the coast, railways, and highways, and far from urban areas.
- Payloads with lower weight have a higher launch success rate compared to those with heavier weight.
- The models exhibit similar accuracy (83.33%), but insufficient data size precludes determination of an optimal model.

Appendix

- GitHub Code: https://github.com/duy-genashtim/Data-Science-and-MachineLearning-Capstone
- Course URL: https://learning.edx.org/course/course-v1:IBM+DS0720EN+2T2021/home

Thank you!