IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Shin Le>

<Sep 12th, 2023>

Outline

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

Executive Summary

- Methodologies:
  - Data Collection, Data Wrangling, Web Scraping, SQL, Graph, Dashboard, Classification Models
- Results:
  - The success rate kept increasing positively in years.
  - All launch sites were located near coastlines.
  - KSC LC-39A has the highest success rate
  - Decision Tree could be the best model for our project with high accuracy. So, Space X Falcon 9 can be confidently landed at the first stage.

Introduction

- Project background and context:
  - SpaceX has gained worldwide attention for a series of historic milestones.
  - It is the only private company ever to return a spacecraft from low-earth orbit, which it first accomplished in December 2010. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars wheras 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 SpaceX for a rocket launch.
  - This dataset includes a record for each payload carried during a SpaceX mission into outer space.

- Problems you want to find answers:
  - Which booster version has the highest success rate?
  - Which machine learning model has the highest accuracy?
  - Which region is concentration of rockets development?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - collect and make sure the data is in the correct format from the SpaceX API
  - clean the requested data
- Perform data wrangling
  - Determine the Null/ missing values, then replace them with appropriate values such mean in our data set. In some case, we will create a dummy variable to solve this.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Perform exploratory Data Analysis and determine Training Labels
  - Find best Hyperparameter for SVM, Classification Trees and Logistic Regression

Data Collection

- Collect and make sure the data is in the correct format from the SpaceX API
  - Make a get request to the SpaceX API
  - Do some basic data wrangling and formating.
    - Request to the SpaceX API
    - Clean the requested data

Data Collection – SpaceX API

Now let's start requesting rocket launch data from SpaceX API with the following URL:

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

```
response = requests.get(spacex_url)
```

```
print(response.content)
```


```
b'[{"fairings":{"reused":false,"recovery_attempt":false,"recovered":false,"ships":[]},"links":{"patch":
{"small":"https://images2.imgbox.com/94/f2/NN6Ph45r_o.png","large":"https://images2.imgbox.com/5b/02/QcxH
Ub5V_o.png"},"reddit":{"campaign":null,"launch":null,"media":null,"recovery":null},"flickr":{"small":
[],"original":[]},"presskit":null,"webcast":"https://www.youtube.com/watch?v=0a_00nJ_Y88","youtube_id":"0
a_00nJ_Y88","article":"https://www.space.com/2196-spacex-inaugural-falcon-1-rocket-lost-launch.html","wik
ipedia":"https://en.wikipedia.org/wiki/DemoSat"},"static_fire_date_utc":"2006-03-17T00:00:00.000Z","stati
c_fire_date_unix":1142553600,"net":false,"window":0,"rocket":"5e9d0d95eda69955f709d1eb","success":fals
e,"failures":[{"time":33,"altitude":null,"reason":"merlin engine failure"}],"details":"Engine failure at
33 seconds and loss of vehicle","crew":[],"ships":[],"capsules":[],"payloads":["5eb0e4b5b6c3bb0006eeb1e
1"],"launchpad":"5e9e4502f5090995de566f86","flight_number":1,"name":"FalconSat","date_utc":"2006-03-24T2
2:30:00.000Z","date_unix":1143239400,"date_local":"2006-03-25T10:30:00+12:00","date_precision":"hour","up
coming":false,"cores":[{"core":"5e9e289df35918033d3b2623","flight":1,"gridfins":false,"legs":false,"reuse
d":false,"landing_attempt":false,"landing_success":null,"landing_type":null,"landpad":null}],"auto_updat
e":true,"tbd":false,"launch_library_id":null,"id":"5eb87cd9ffd86e000604b32a"},{"fairings":{"reused":fals
e,"recovery_attempt":false,"recovered":false,"ships":[]},"links":{"patch":{"small":"https://images2.imgbo
x.com/f9/4a/ZboXReNb_o.png","large":"https://images2.imgbox.com/80/a2/bkWotCIS_o.png"},"reddit":{"campaig
n":null,"launch":null,"media":null,"recovery":null},"flickr":{"small":[],"original":[]},"presskit":nul
l,"webcast":"https://www.youtube.com/watch?v=Lk4zQ2wP-Nc","youtube_id":"Lk4zQ2wP-Nc","article":"https://w
ww.space.com/3590-spacex-falcon-1-rocket-fails-reach-orbit.html","wikipedia":"https://en.wikipedia.org/wi
ki/DemoSat"},"static_fire_date_utc":null,"static_fire_date_unix":null,"net":false,"window":0,"rocket":"5e
9d0d95eda69955f709d1eb","success":false,"failures":[{"time":301,"altitude":289,"reason":"harmonic oscilla
tion leading to premature engine shutdown"}],"details":"Successful first stage burn and transition to sec
ond stage, maximum altitude 289 km, Premature engine shutdown at T+7 min 30 s, Failed to reach orbit, Fai
led to recover first stage","crew":[],"ships":[],"capsules":[],"payloads":["5eb0e4b6b6c3bb0006eeb1e2"],"l
aunchpad":"5e9e4502f5090995de566f86","flight_number":2,"name":"DemoSat","date_utc":"2007-03-21T01:10:00.0
00Z","date_unix":1174439400,"date_local":"2007-03-21T13:10:00+12:00","date_precision":"hour","upcoming":f
alse,"cores":[{"core":"5e9e289ef35918416a3b2624","flight":1,"gridfins":false,"legs":false,"reused":fals
e,"landing_attempt":false,"landing_success":null,"landing_type":null,"landpad":null}],"auto_update":tru
e,"tbd":false,"launch_library_id":null,"id":"5eb87cdaffd86e000604b32b"},{"fairings":{"reused":false,"reco
very_attempt":false,"recovered":false,"ships":[]},"links":{"patch":{"small":"https://images2.imgbox.com/6
```


- GitHub URL:

[SpaceX API]

Data Collection - Scraping

- static_json_url='https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-Skills

```
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}
```

```
# Show the head of the dataframe
data.head()
```

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Re
0	1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None	1	False	
1	2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None None	1	False	
2	4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll	None None	1	False	
3	5	2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll	None None	1	False	
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	

- GitHub URL: SpaceX API

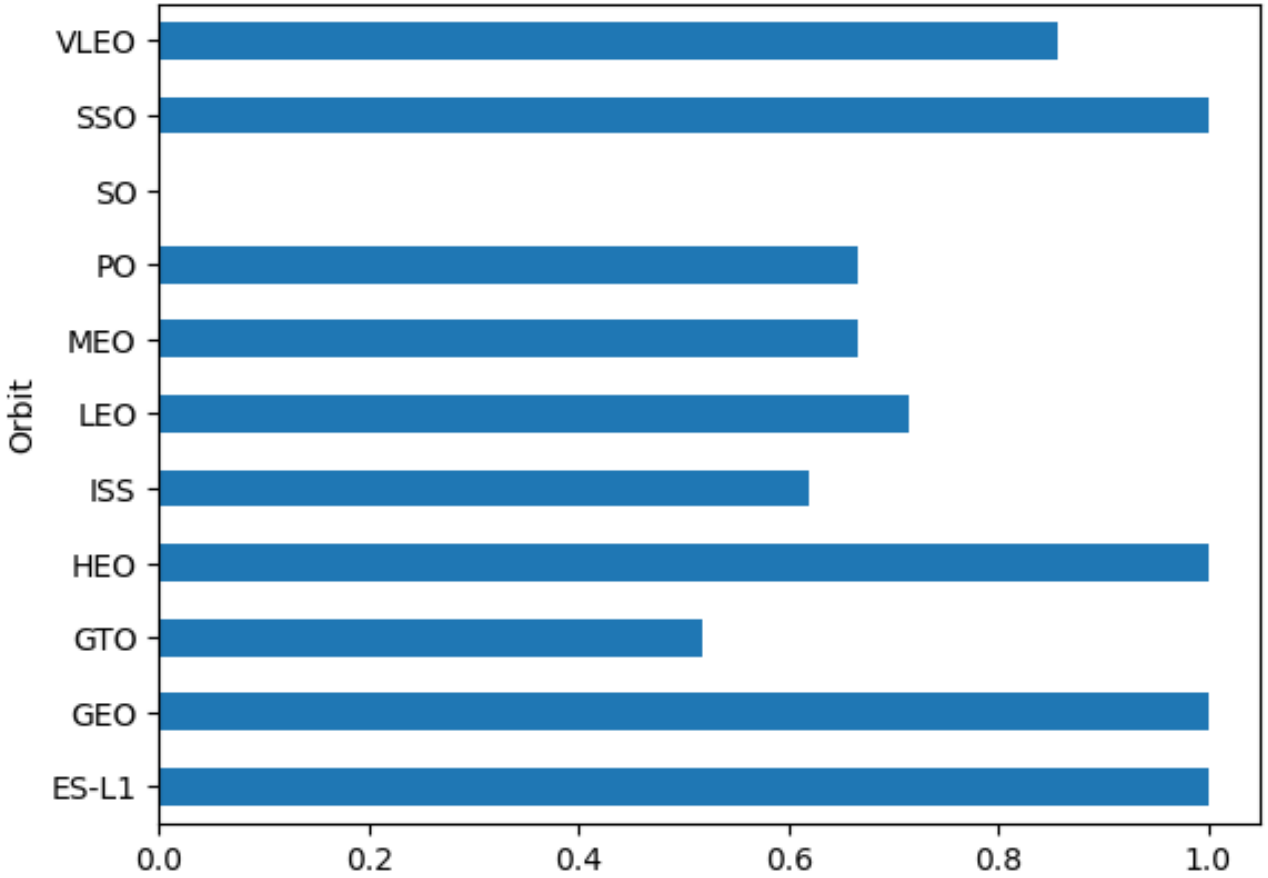
Data Wrangling

- Data processing:
  - Finding and replace the rows that are missing values in our dataset:
    - Using the mean() for the “PayloadMass” in replace missing elements in that columns
    - We also have missing values on columns “LandingPad,” but we will create dummy variables to categorical columns. Using ‘BeautifulSoup’ to extract Falson 9 launch record, then parse the table and convert it into a pandas data frame **from the HTML table**

- GitHub URL: SpaceX API Web Scraping Data Wrangling Dummy Variables

EDA with Data Visualization

- We will perform some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models
    - We attempted to convert the ‘Outcomes” into Training Labels (1 vs 0 as successful vs unsuccessful landing). Then calculating the success rate.
    - Creating dummy variables to categorical columns, then cast all numeric columns to ‘float64’
- GitHub URL: Data Wrangling Exploring and Preparing Data



The relationship between success rate of each orbit type

EDA with SQL

- Using SQL queries to understanding the SpaceX dataset:
  - **Display the names of the unique launch sites in the space mission**
  - **Display 5 records where launch sites begin with the string 'CCA'**
  - **Display the total payload mass carried by boosters launched by NASA (CRS)**
  - **Display average payload mass carried by booster version F9 v1.1**
  - List the date when the first succesful 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**
  - **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 URL: EDA with SQL

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map:
  - Mark all launch sites on a map
    - Create and add folium.Circle and folium.Marker for each launch site on the site map
  - Mark the success/failed launches for each site on the map
    - Marker clusters shows many markers having the same coordinate
  - Calculate the distances between a launch site to its proximities:
    - Find any railway, highway, coastline, etc.
    - Mark down their coordinates in order to the distance to the launch site
      - Ex: Mark down a point on the closest coastline using MousePosition and calculate the distance between the coastline point and the launch site.
        - Coordinates= [[28.56367, -80.57163], [28.56216 , -80.56778]]

- We can see that most of launch sites are very close to coastline and far away from cites
- GitHub URL: Locations Analysis with Folium

Build a Dashboard with Plotly Dash

- Dashboard:
  - We create a dashboard to show the interactions and comparison among four different launch sites based on the success rate
    - Add a Launch Site Drop-down Input Component
    - Add a callback function to render 'success-pie-chart' based on selected site dropdown
    - Add a Range Slider to Select Payload
    - Add a callback function to render the success-payload-scatter-chart scatter plot

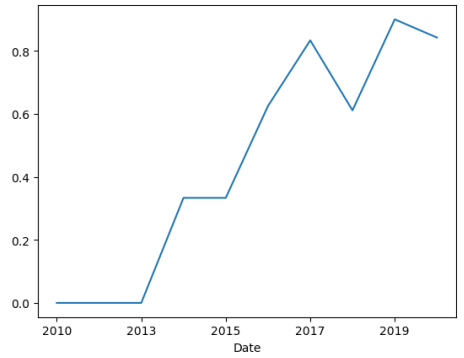
- GitHub URL: Dashboard with Ploty dash (code) Dashboard Dataset

Predictive Analysis (Classification)

- Perform exploratory Data Analysis and determine Training Labels
  - Create a column for the class
  - Standardize the data
  - Split into training data and test data
- Find best Hyperparameter for SVM, Classification Trees and Logistic Regression
  - Find the method performs best using test data
- GitHub URL: Machine Learning Prediction

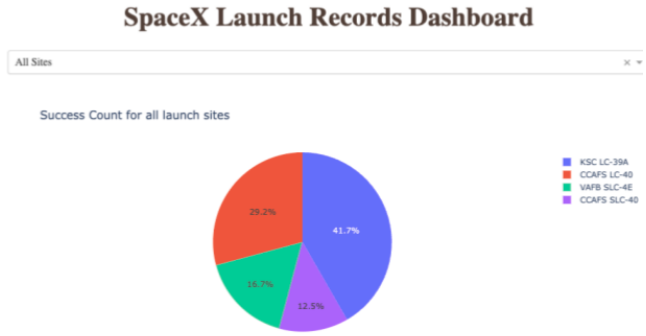
Results

- Exploratory data analysis results



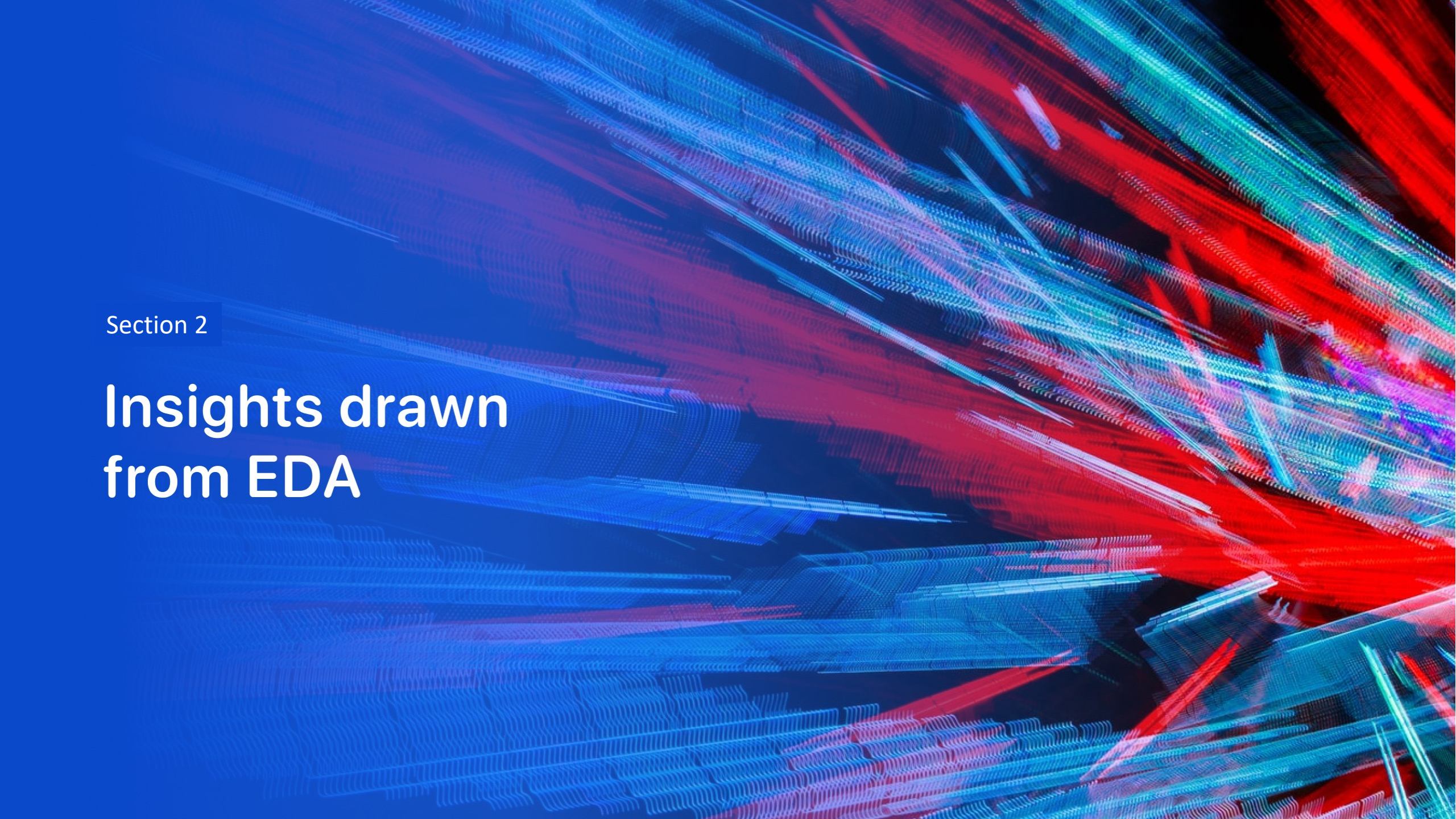
you can observe that the sucess rate since 2013 kept increasing till 2020

- Interactive analytics demo in screenshots



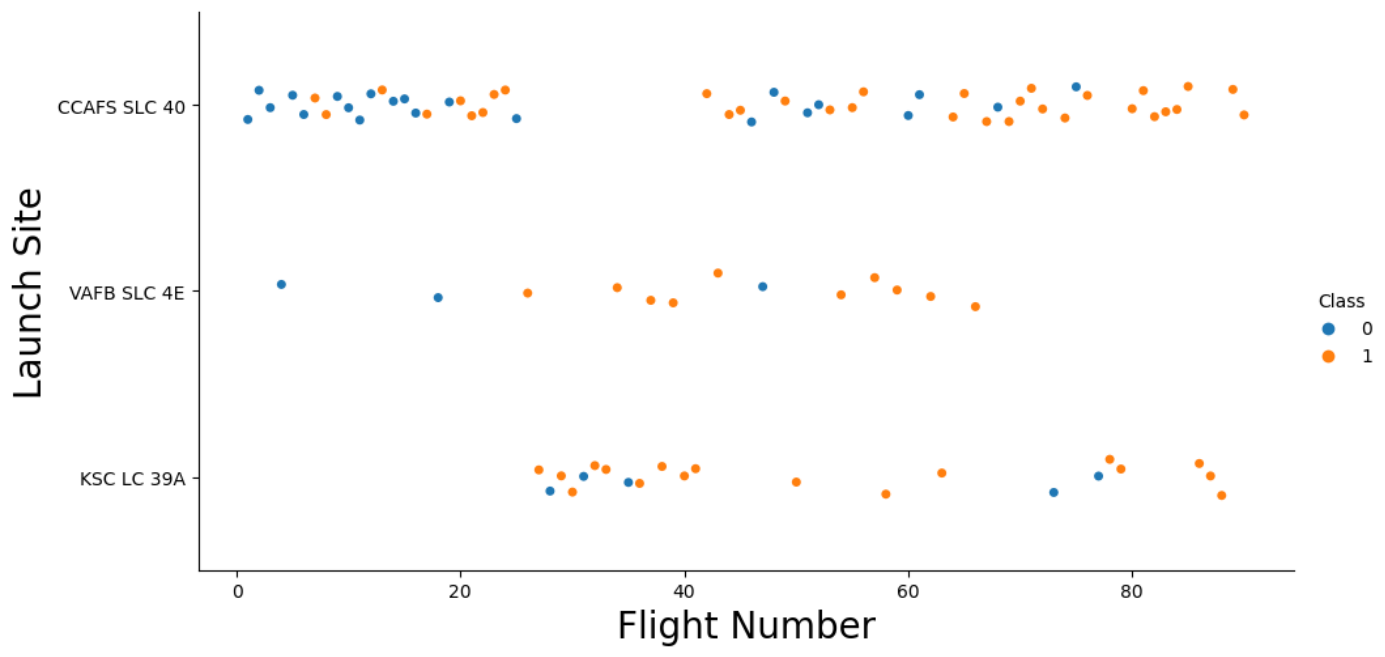
- Predictive analysis results

	Classification Method	Accuracy_Training Data	Accuracy_Test Data
0	Logistic Regression	0.846429	0.833333
1	SVM	0.848214	0.833333
2	Decision Tree	0.875000	0.833333
3	KNN	0.848214	0.833333


Section 2
Insights drawn from EDA

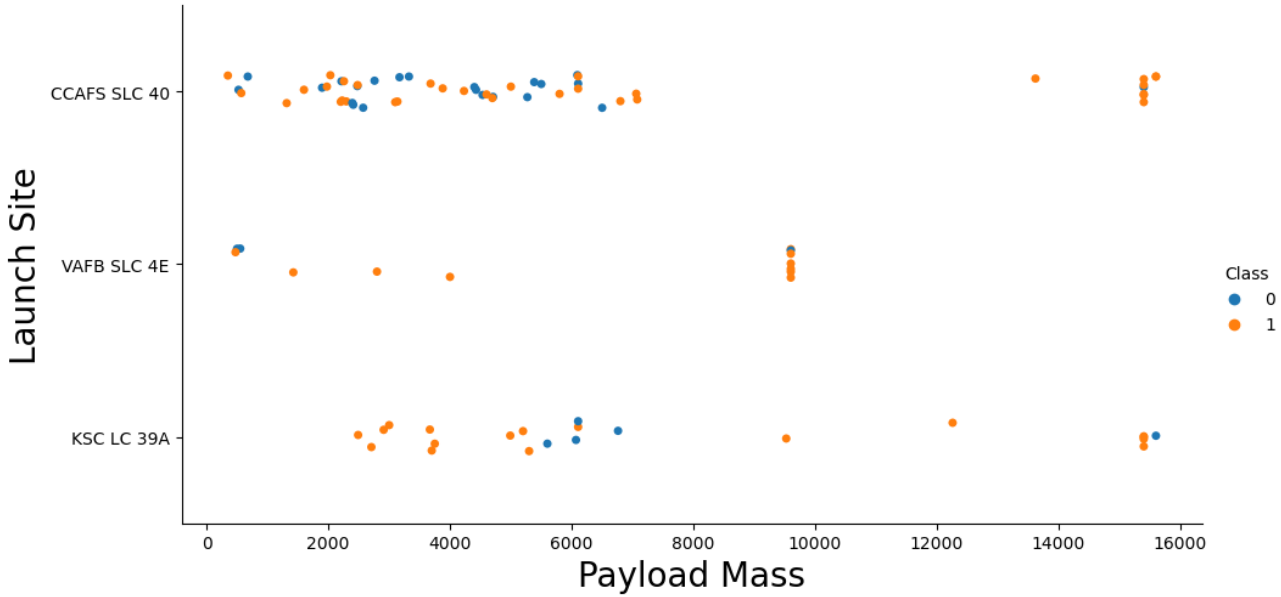
Flight Number vs. Launch Site

- We can see that CCAFS SLC40 is the Launch Site which has the highest number of success flights. It is also the one that has the the highest number of fail flights
- The flight number is increasing, and the probability of success is also increasing



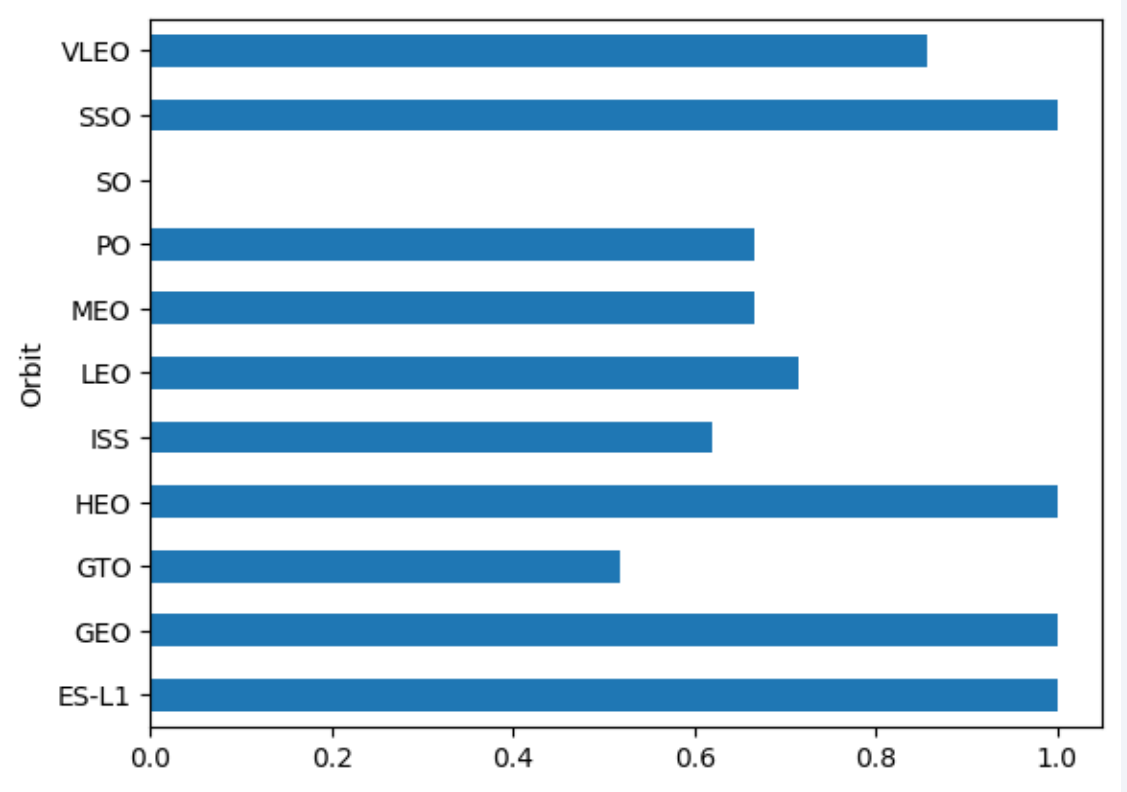
Payload vs. Launch Site

- We can see that SpaceX focuses on developing the Launch Site in weight from 2000 - 8000 kg. These sites have higher success rate.
- There are many massive rockets with heavy payloads (10000-16000 kg)that were successfully launched



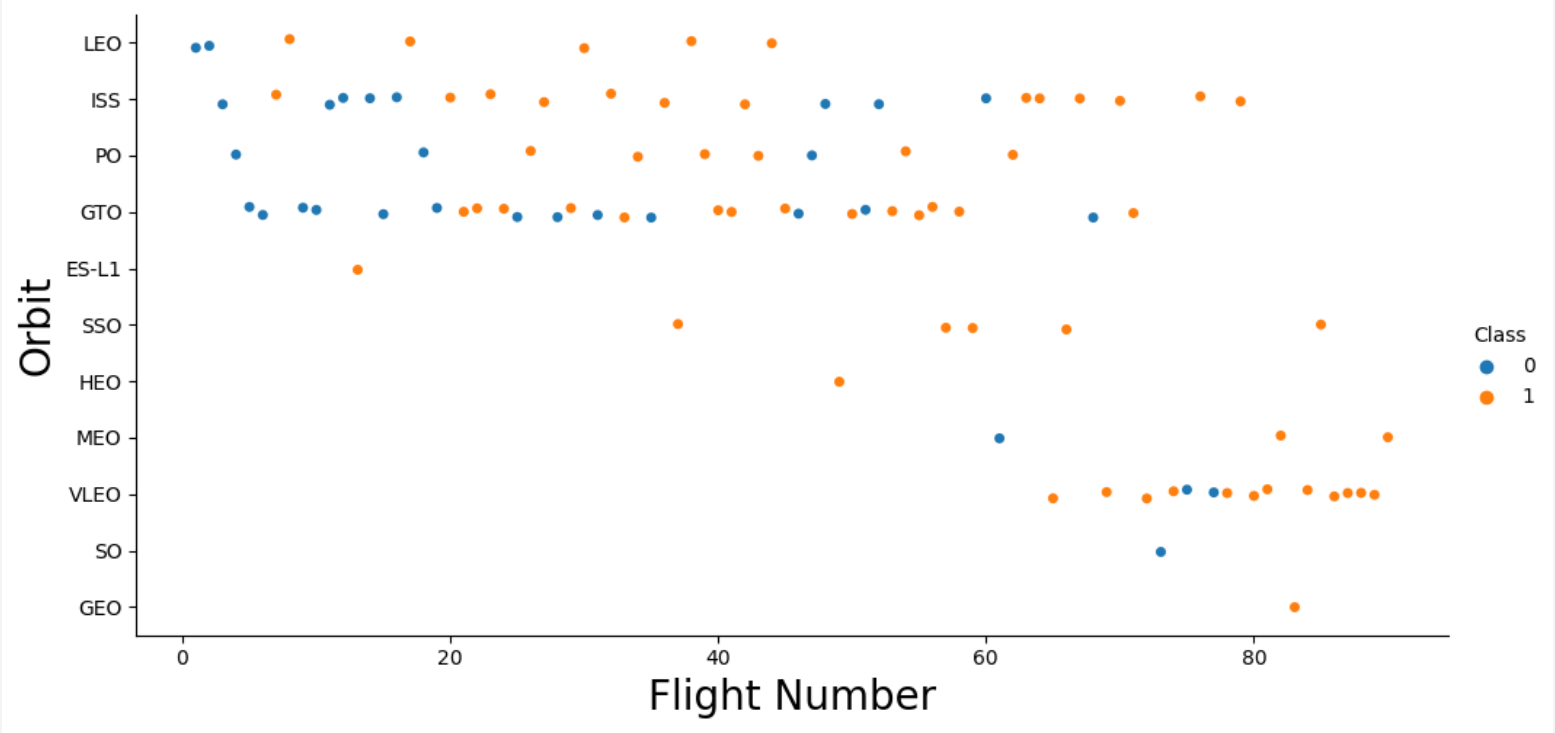
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO are the orbits have the highest success rate of 1.00. That means, rockets were completely and successfully launched along these orbits.



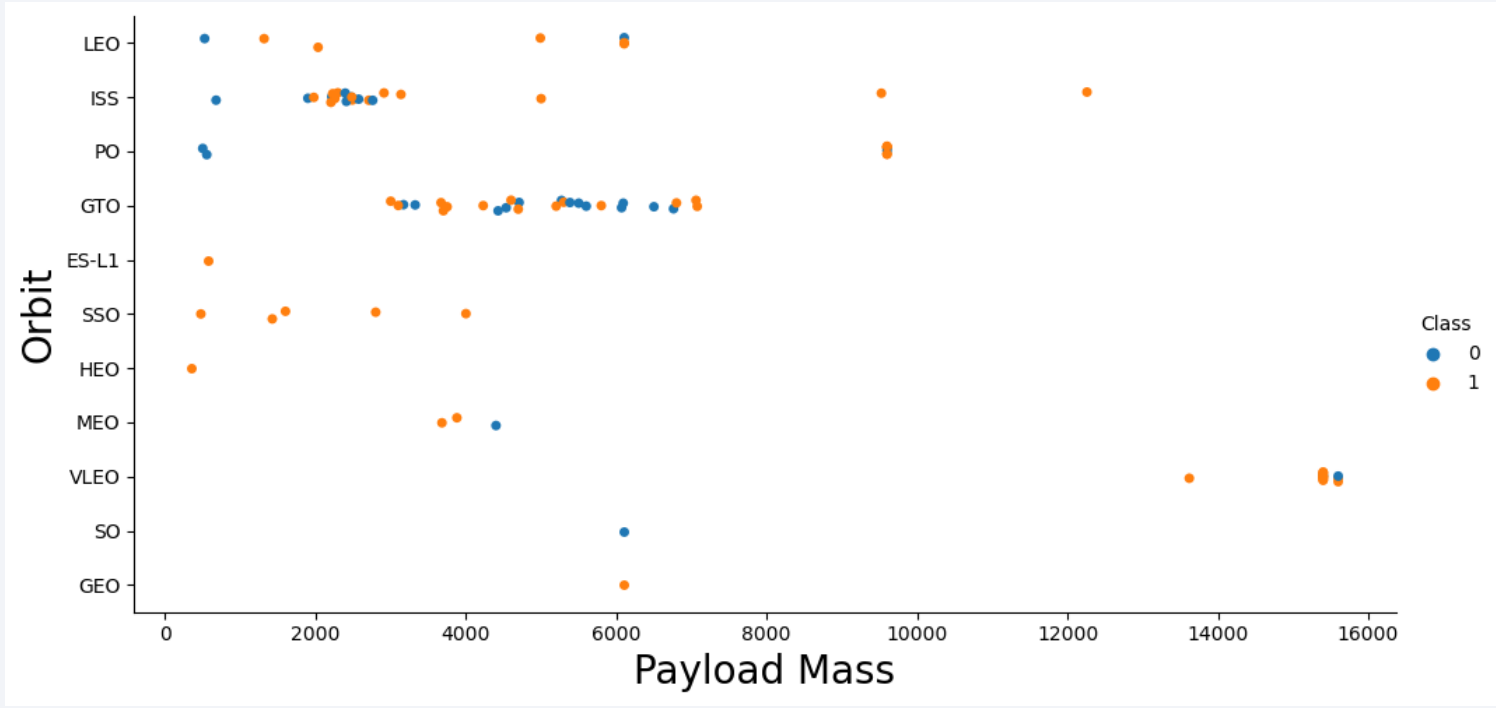
Flight Number vs. Orbit Type

- The LEO orbit is 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.



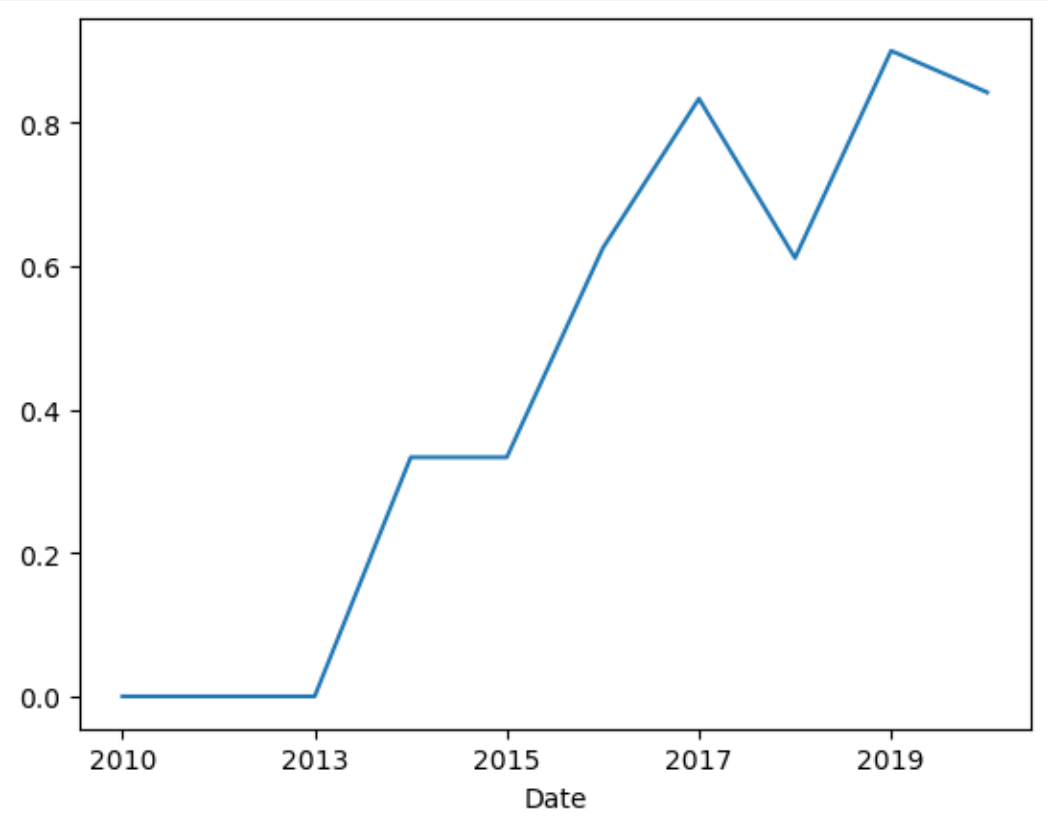
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- 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 success rate since 2013 kept increasing till 2020



All Launch Site Names

- There are four distinct launch sites

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

```
%sql SELECT DISTINCT(Launch_Site) FROM SPACEXTABLE
```

```
 * 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'

Display 5 records where launch sites begin with the string 'CCA'

```
%sql SELECT * FROM SPACEXTABLE 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
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success
2010-08-12	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
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success

Total Payload Mass

- The total payload carried by boosters from NASA

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

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS "The total payload mass" FROM SPACEXTABLE WHERE Customer ="NASA (CRS)"
```

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

The total payload mass
45596

Average Payload Mass by F9 v1.1

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

Display average payload mass carried by booster version F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS "The average payload mass" FROM SPACEXTABLE\
WHERE Booster_Version LIKE "F9 v1.1%"
```

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

The average payload mass
2534.6666666666665

First Successful Ground Landing Date

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

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%sql SELECT MIN(Date), Landing_Outcome FROM SPACEXTABLE \
WHERE Landing_Outcome LIKE "%success%"
```

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

MIN(Date)	Landing_Outcome
2015-12-22	Success (ground pad)

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, Landing_Outcome FROM SPACEXTABLE \
WHERE Landing_Outcome LIKE "%success (drone%"\
AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
```

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

Booster_Version	Landing_Outcome
F9 FT B1022	Success (drone ship)
F9 FT B1026	Success (drone ship)
F9 FT B1021.2	Success (drone ship)
F9 FT B1031.2	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
%sql SELECT Mission_Outcome, COUNT(Mission_Outcome) From SPACEXTABLE GROUP BY Mission_Outcome
```

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

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

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%sql SELECT Booster_version, PAYLOAD_MASS__KG_ FROM SPACEXTABLE \
WHERE PAYLOAD_MASS__KG_ =(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

```
 * 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

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

```
%sql SELECT   Date AS month, landing_outcome, booster_version, launch_site FROM SPACEXTABLE\
WHERE landing_outcome LIKE "failure (drone%" AND Date LIKE '%2015%'
```

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

month	Landing_Outcome	Booster_Version	Launch_Site
2015-10-01	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

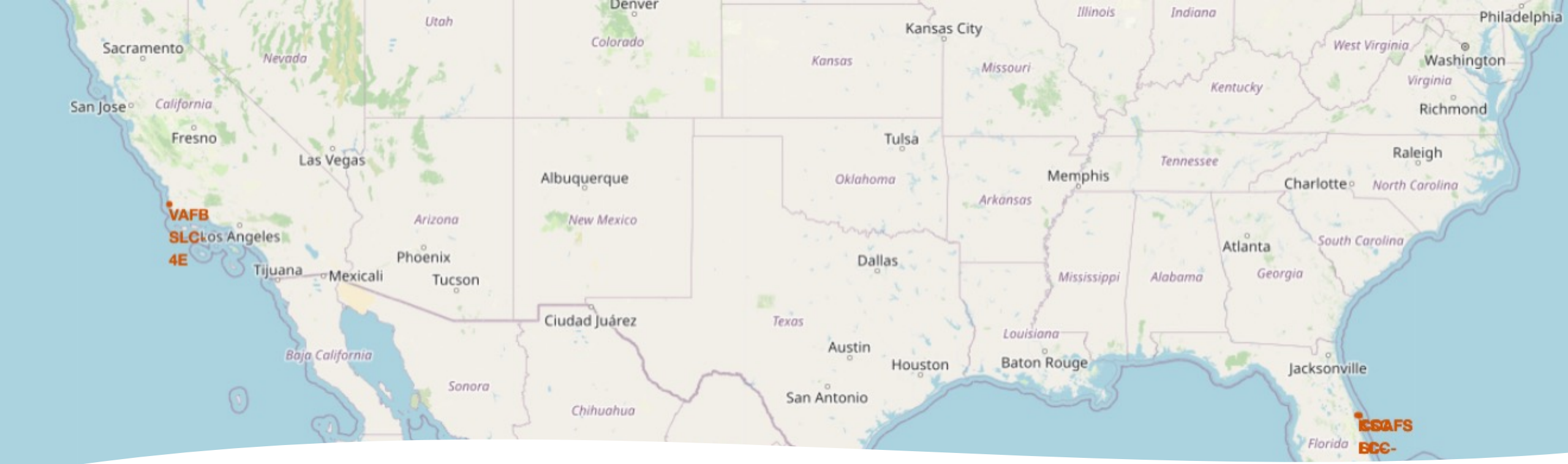
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

```
%sql SELECT landing_outcome, COUNT(landing_outcome) FROM SPACEXTABLE WHERE Date Between '2010-06-04' and '2017-03
```

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

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


Section 3
Launch Sites
Proximities Analysis

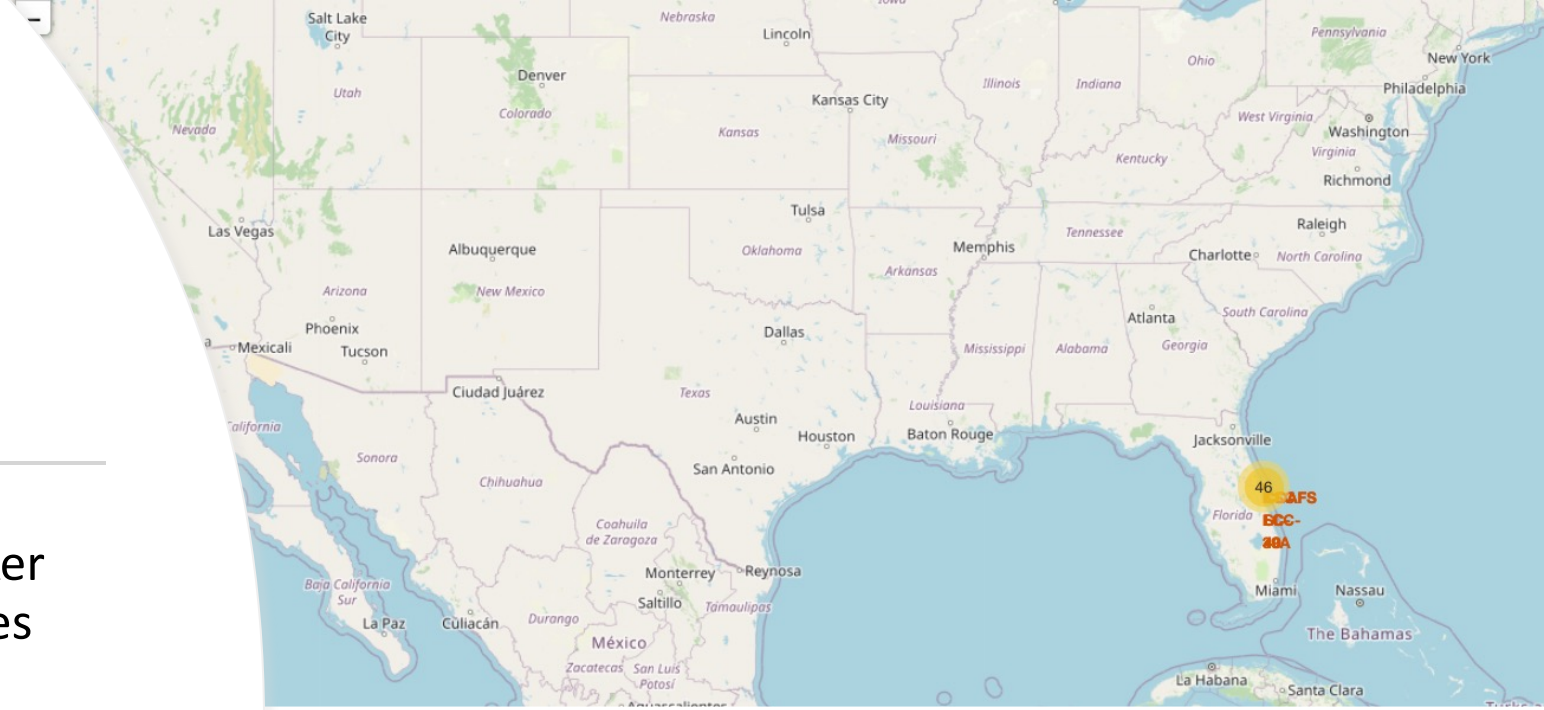


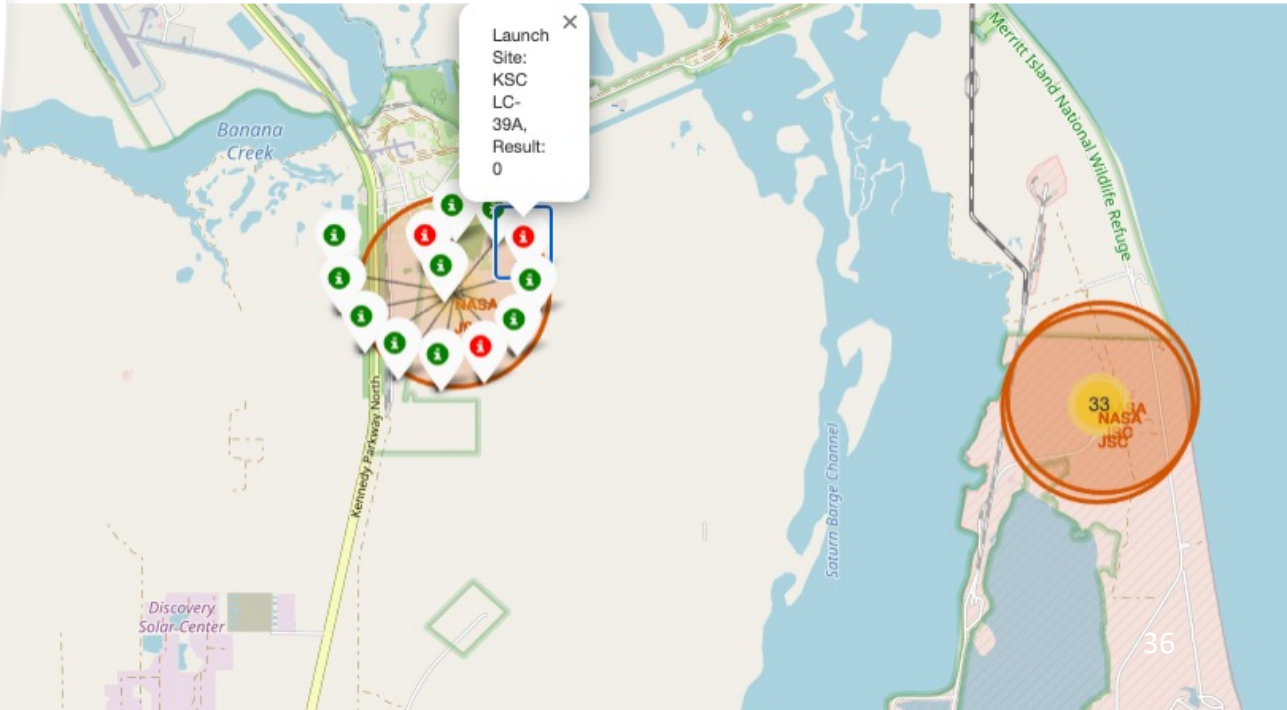
Folium Map : Mark all launch sites on a map

- All launch sites were located near coastlines. Florida is the state where rockets are concentrated and developed

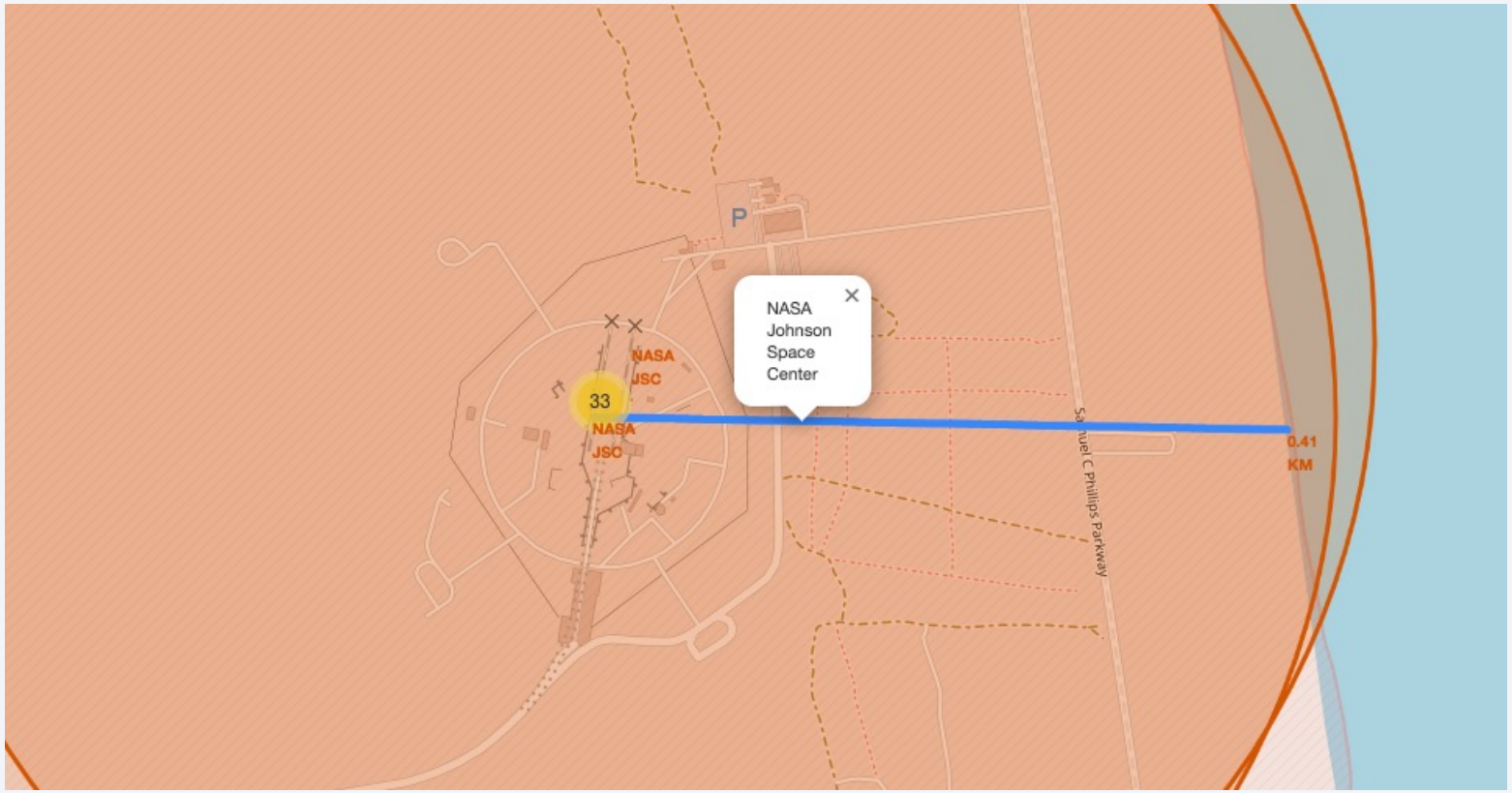
Folium Map: Mark the success/failed launches for each site

- From the color-labeled markers in marker clusters, we easily identify that launch sites KSC LC-39A have relatively the highest success rates.

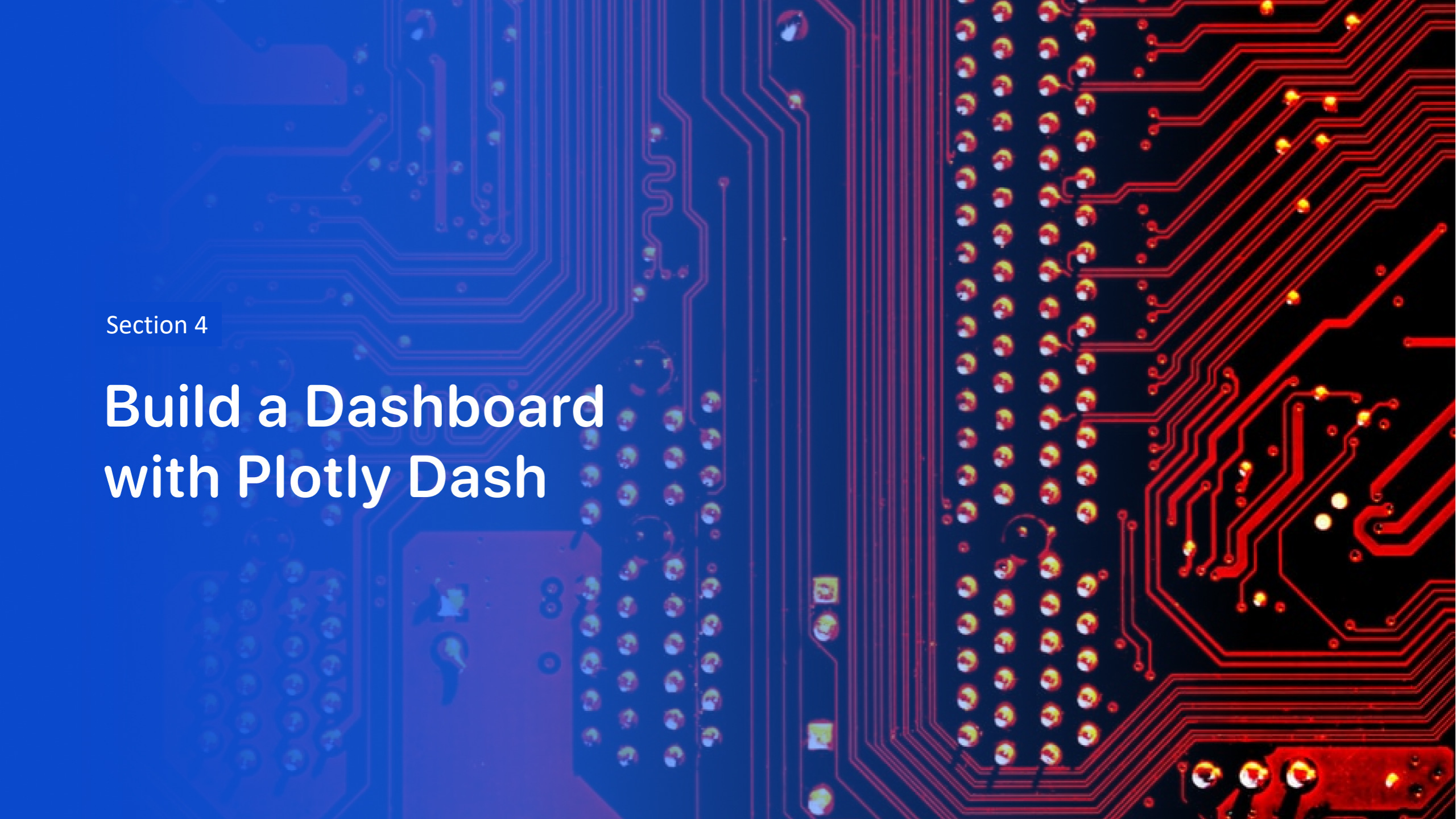




Folium Map: The distances between a launch site to its proximities

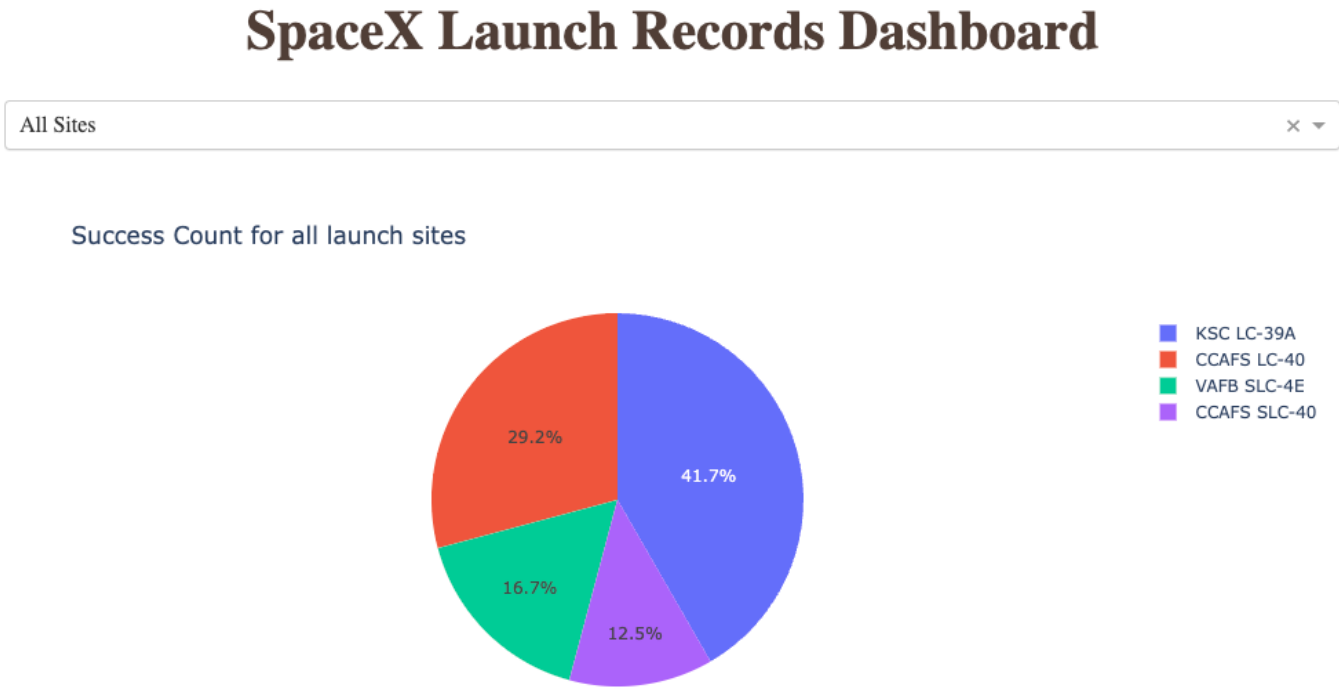


NASA JSC is located just 0.41 kilometers east of the coastline.


Section 4
Build a Dashboard with Plotly Dash

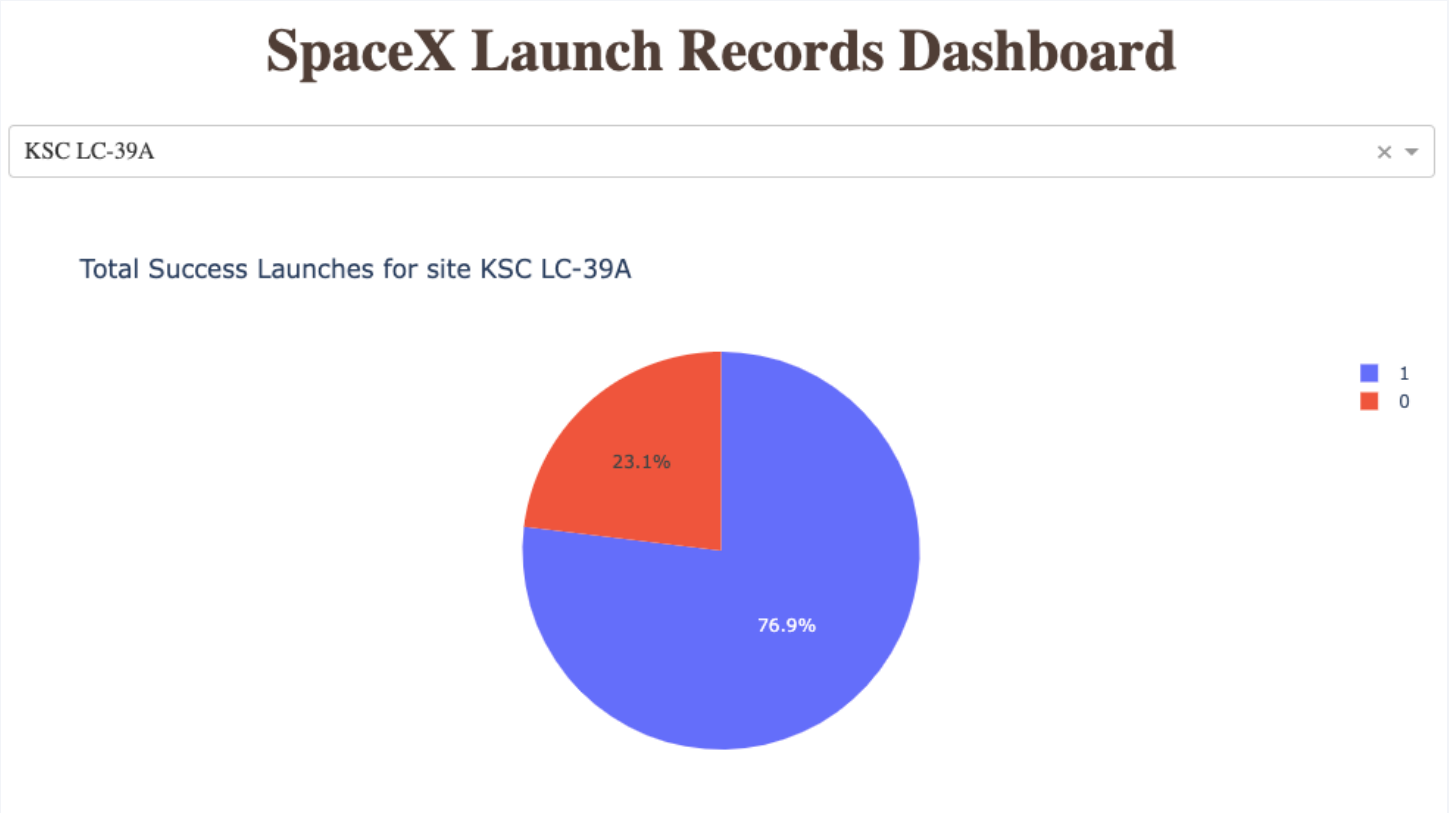
Dashboard: Success Count for all launch sites

- Obviously, KSC LC-39A has the highest success rate when it comes to rocket launches.



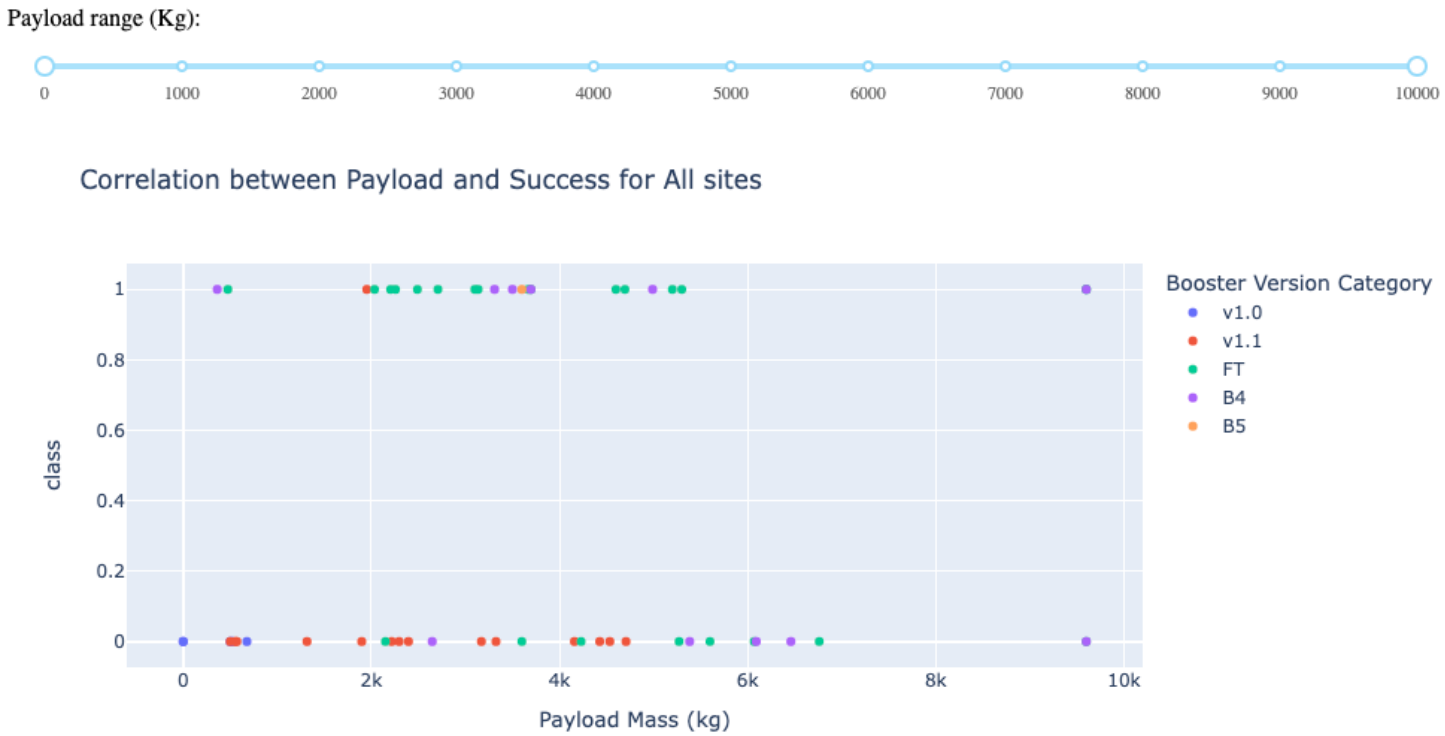
Dashboard: The pie chart for the launch site with highest launch success ratio

The success rate of KSC LC-39A is 76.9%



Dashboard: Payload vs. Launch Outcome scatter plot for all sites

- FT is the booster version have the largest success rate with payload range from 2000 to 6000 kg.



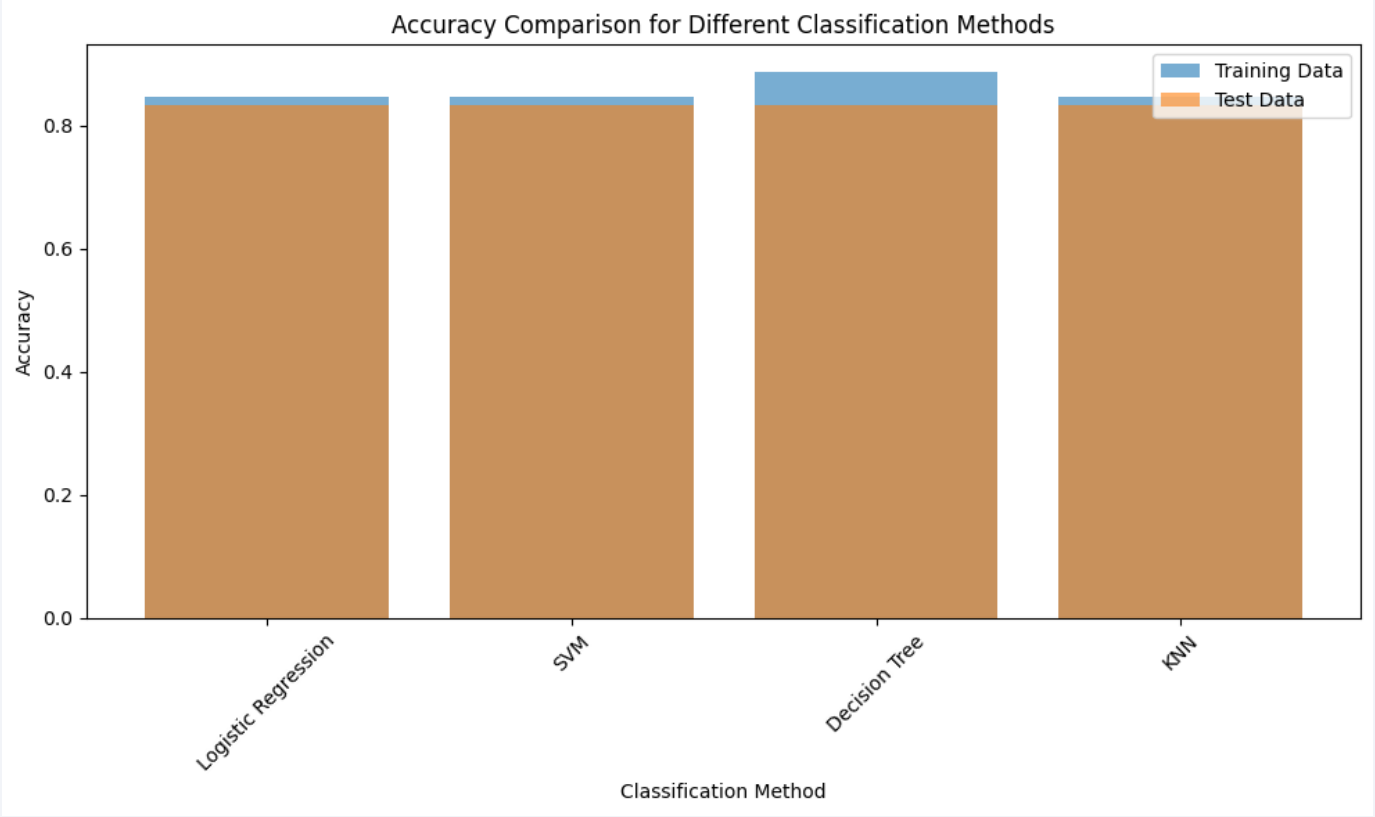
Section 5

Predictive Analysis (Classification)

Classification Accuracy

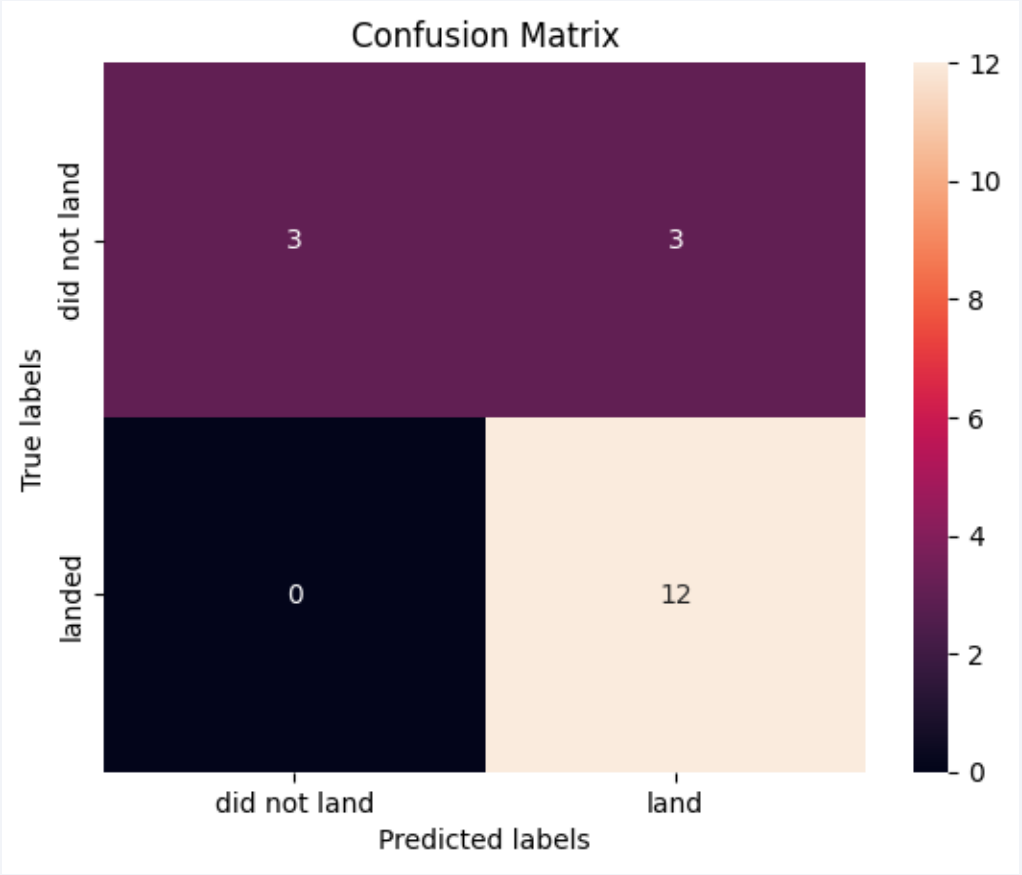
Decision Tree model has the highest classification accuracy. However, the other model have high accuracies.

	Classification Method	Accuracy_Training Data	Accuracy_Test Data
0	Logistic Regression	0.846429	0.833333
1	SVM	0.848214	0.833333
2	Decision Tree	0.887500	0.833333
3	KNN	0.848214	0.833333



Confusion Matrix

- We can see the true label (landed) and predicted label (land) completely matches in Confusion Matrix (it shows as an element of 12).



Conclusions

- The success rate since 2013 kept increasing till 2020.
- All launch sites were located near coastlines. Florida is the state where rockets are concentrated and developed
- KSC LC-39A has the highest success rate when it comes to rocket launches.
- Decision Tree is the best model for our project with accuracy of 88%.
- So, Space X Falcon 9 can be confidently landed at the first stage based on the positive result and a high accuracy of model in Machine Learning prediction.

Appendix

- Creating a table and bar chart to summarize and compare our models in machine learning prediction in Task 12 in Machine Learning Prediction

Thank you!