IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Anna Maslowska-Gornicz
05.10.2021

Outline

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

Summary

- Public data of SpaceX were collected from a Wikipedia.
- Python and SQL were used to clean, convert into data frame, analyze and visualize.
- Four machine learning models (logistic regression, support vector analysis, decision tree and K nearest neighbors) were produced to predict if the first stage of Falcon 9 lands successfully.
- With three methods (logistic regression, support vector analysis and K nearest neighbors) the accuracy was ~80% and only decision tree accuracy was significant lower (~70%).

Introduction

- Project background and context:

Space Y would like to compete with SpaceX founded by Billionaire industrialist Allon Musk.

- Problems you want to find answers:

Determine the price of each launch (first stage) from SpaceX and create dashboards.

We would like to answer on question: Can we determine if SpaceX will reuse the first stage using the machine learning and public information?

Section 1

Methodology

Methodology

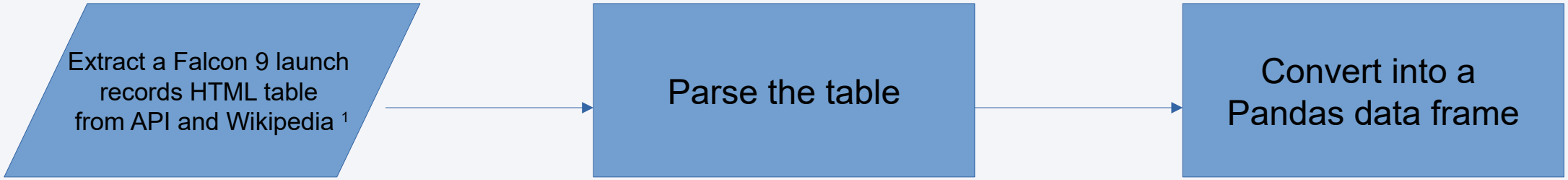
OUTLINE

Data collection methodology:

- Combined data about SpaceX from API and Wikipedia.

- Perform data wrangling
  - Reviewing some of the attributes: FlightNumber, Date, Booster version, Payload mass Orbit, Launch Site, Outcomes
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection



1. https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Data Collection – SpaceX API

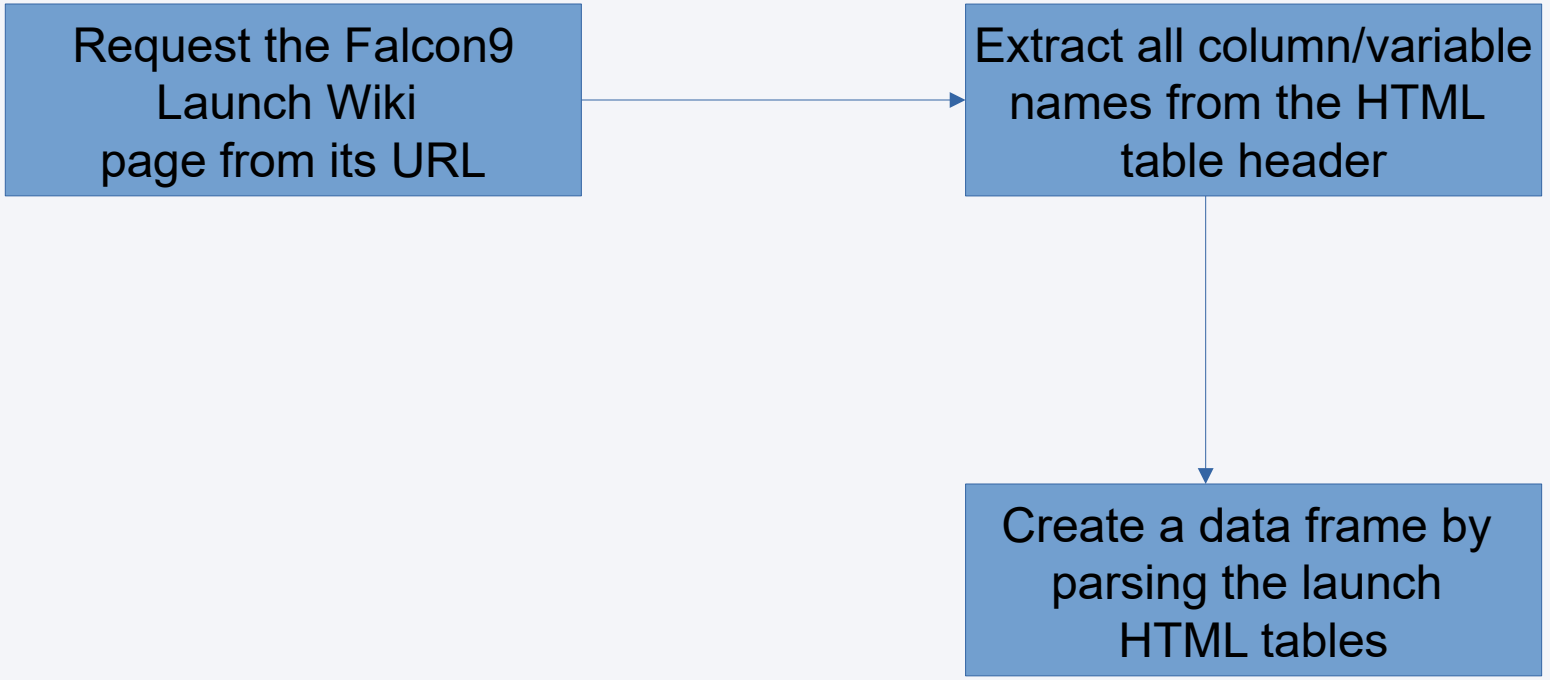
Request and parse the SpaceX launch data using the GET request

Filter the data frame to only include Falcon 9 launches

Dealing with missing Values

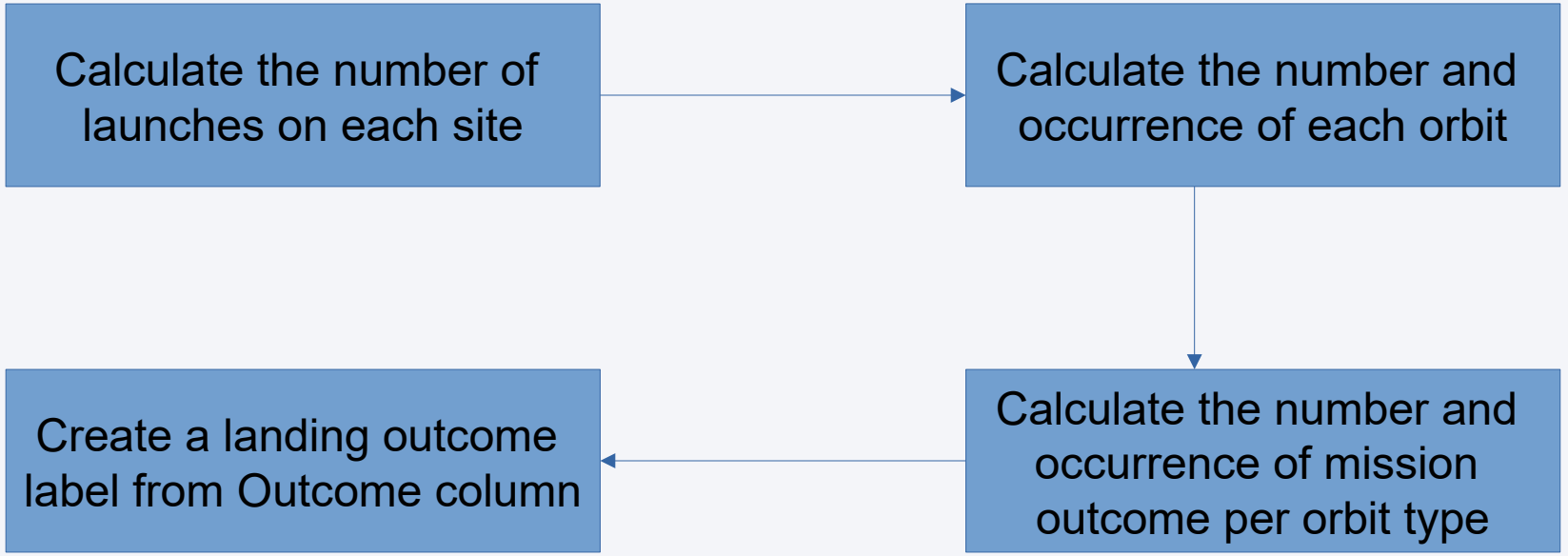
https://github.com/TheAnuska/testrepo1/blob/master/Complete%20the%20Data%20Collection%20API%20Lab.ipynb

Data Collection - Scraping



https://github.com/TheAnuska/testrepo1/blob/main/Complete%20the%20Data%20Collection%20with%20Web%20Scraping%20lab.ipynb

Data Wrangling



https://github.com/TheAnuska/testrepo1/blob/main/Complete%20the%20EDA%20lab.ipynb

EDA with Data Visualization

1) Data wrangling on selected features: Flight Number, Payload Mass, LaunchSite, Orbit type, Class and Year.

2) Skatter plots: FlightNumber vs LaunchSite, FlightNumber vs. PayloadMass, Payload Vs. LaunchSite, FlightNumber vs Orbit type, Payload vs Orbit type.

3) Bar chart: success rate of each orbit type.

4) Linear plot: to visualize the launch success yearly trend.

https://github.com/TheAnuska/testrepo1/blob/main/Complete%20the%20EDA%20with%20Visualization%20lab.ipynb

EDA with SQL

1) Loaded data into IBM DB2 database.

2) SQL used for queries in Jupyter notebook.

3) Queries made to better understanding datasets.

4) Queries such as: average payload mass carried by booster version F9 v1.1, successful landing outcome, list of the boosters, list of the launches with several conditions.

https://github.com/TheAnuska/testrepo1/blob/main/Complete%20the%20EDA%20with%20SQL%20lab.ipynb

Build an Interactive Map with Folium

1) Mark all launch sites on a map.
2) Markers to investigate the success/failed launches.
3) Circles to highlight area with a text label on a specific coordinate.
4) Draw a line between a launch site to the selected coastline point.

https://github.com/TheAnuska/testrepo1/blob/main/Complete%20the%20Interactive%20Visual%20Analytics%20with%20Fol.ipynb

Build a Dashboard with Plotly Dash

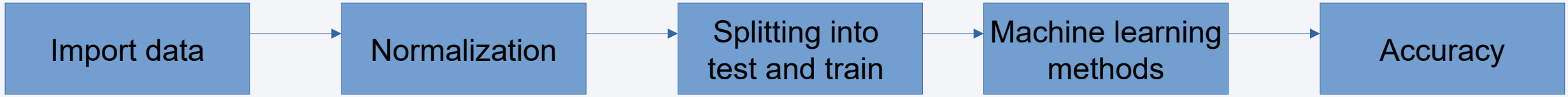
1) A pie chart visualizing total success landings across all launch sites or individual.

2) A scatter plot to see success count.

https://github.com/TheAnuska/testrepo1/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

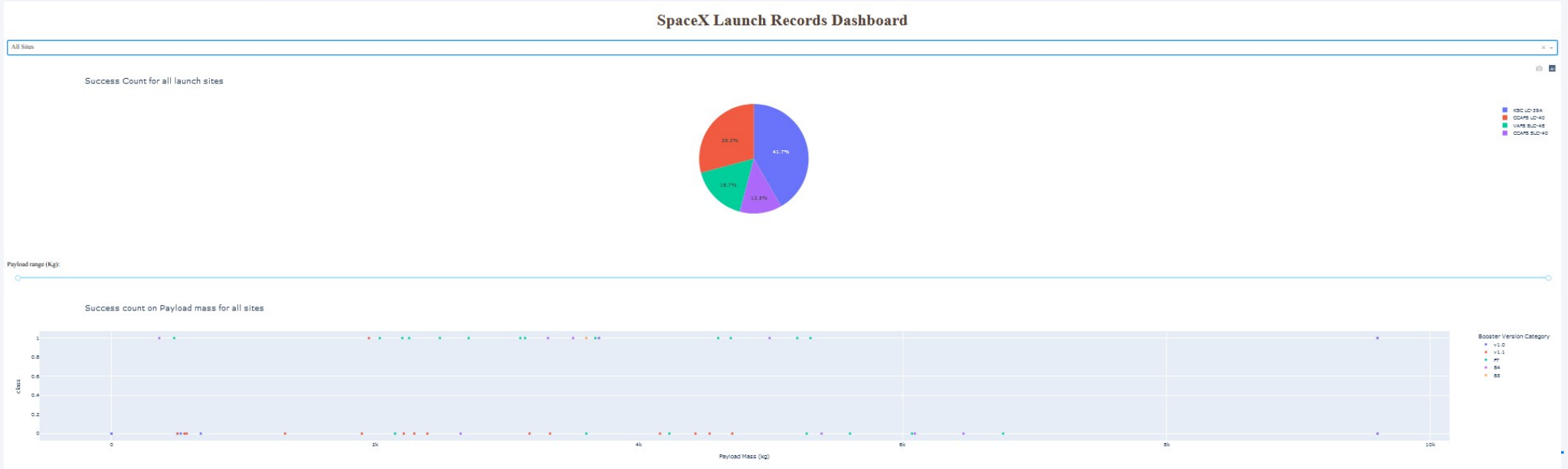
1. Import data frame from previous step.
2. Normalization data and splitting into train and test sets.
3. Use machine learning methods: logistic regression, support vector machine, decision tree and k nearest neighbors.
4. Calculate accuracy for each method using method score.

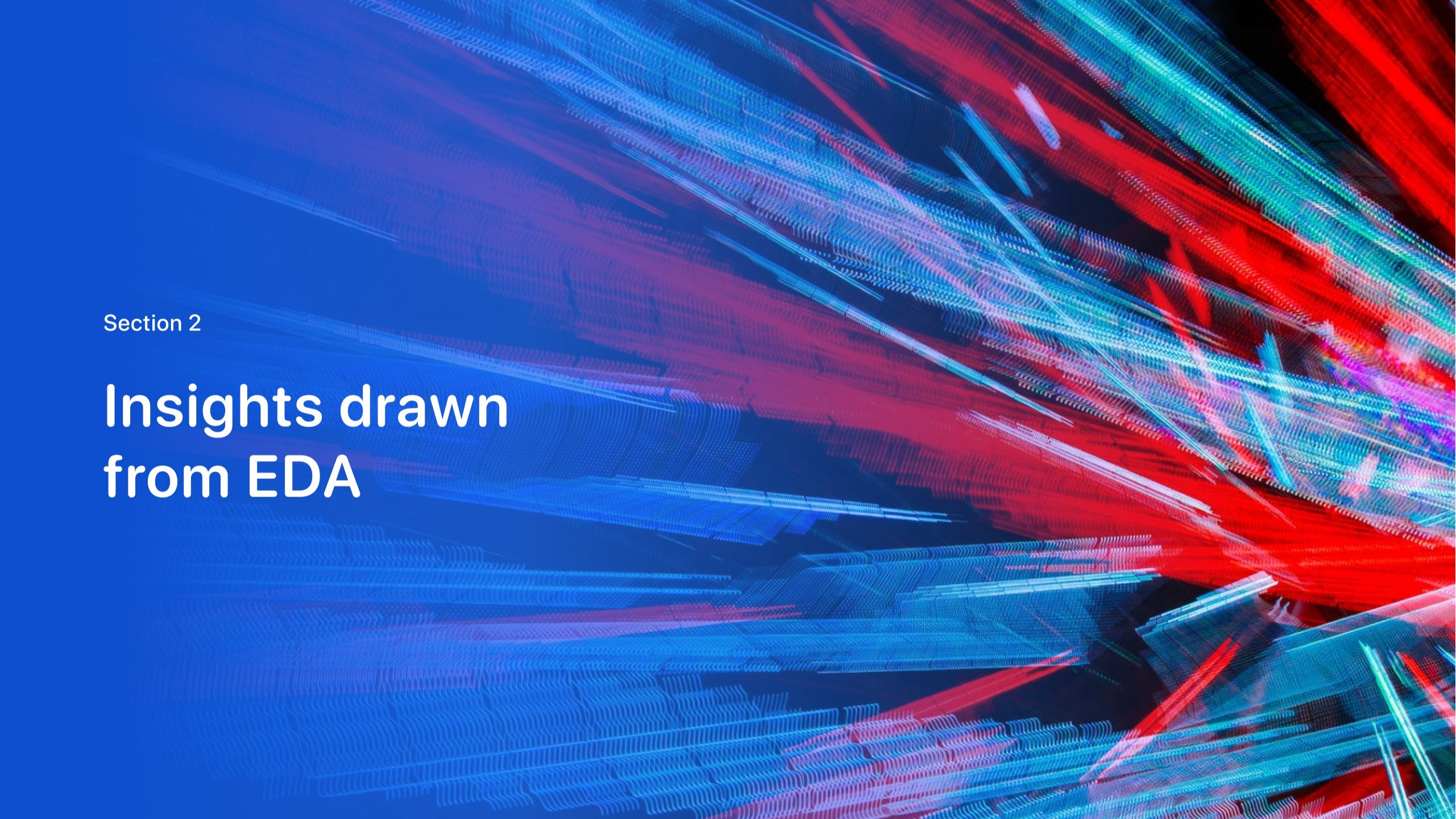


https://github.com/TheAnuska/testrepo1/blob/main/Complete%20the%20Machine%20Learning%20Prediction%20lab.ipynb

Results

This is a preview of the Plotly dashboard. The following sides will show the results of EDA with visualization, EDA with SQL, Interactive Map with Folium, and finally the results of our model with about 83% accuracy.

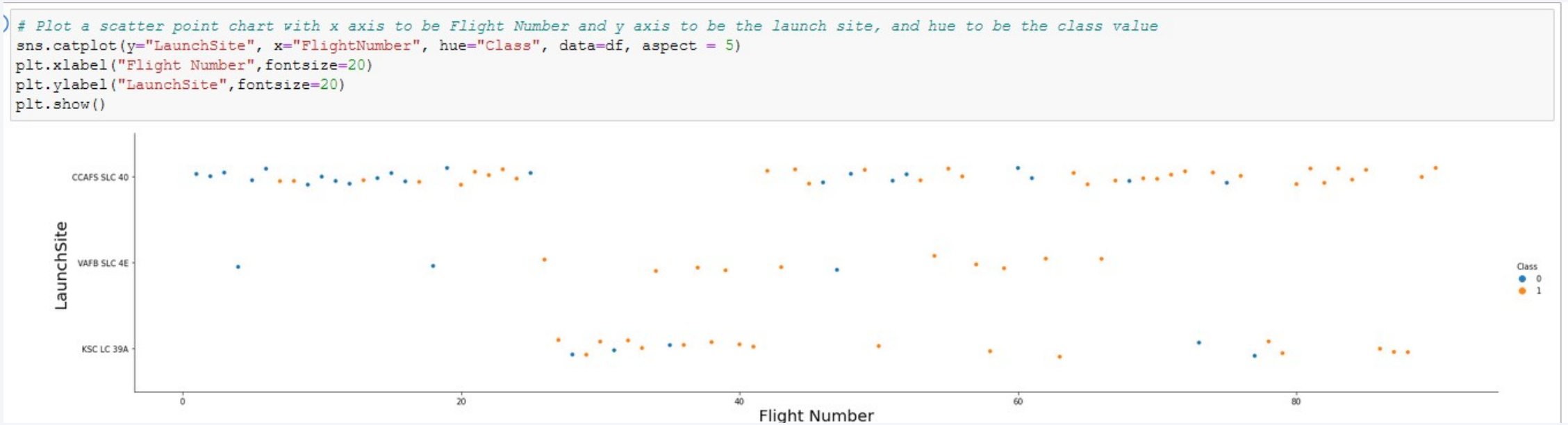



Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

- The Launch Site CCAFS SLC 40 was the most used.
- CCAFS SLC 40 – at the begging the outcome was negative (blue) and later was positive (red).

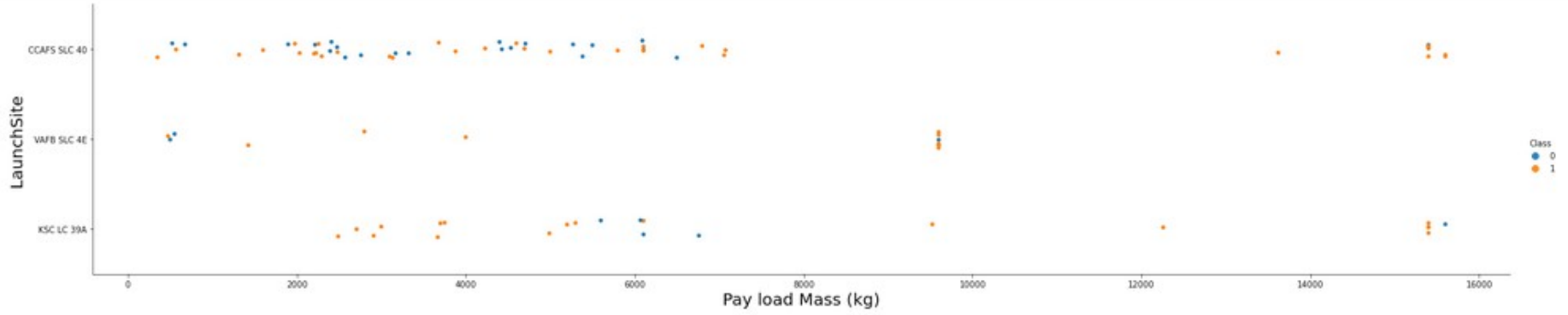
```
# Plot a scatter point chart with x axis to be Flight Number and y axis to be the launch site, and hue to be the class value
sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("LaunchSite",fontsize=20)
plt.show()
```



Payload vs. Launch Site

- We can see that the Pay load increase with the grow of the Launch site.
- Launch negative (blue), launch positive (red).

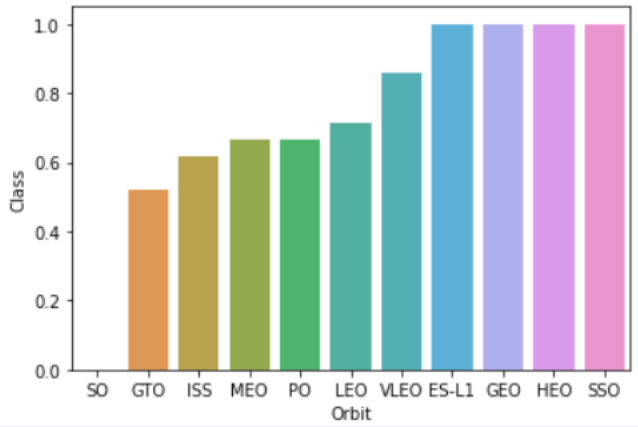
In [8]:
```
# Plot a scatter point chart with x axis to be Pay Load Mass (kg) and y axis to be the launch site, and hue to be the class v
alue
sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("Pay load Mass (kg)",fontsize=20)
plt.ylabel("LaunchSite",fontsize=20)
plt.show()
```



Success Rate vs. Orbit Type

- 4 (ES-L1, GEO, HEO, SSO) orbits out of the all have the mean=1
- On orbit GTO the mean value is the lowest, around 0.5.

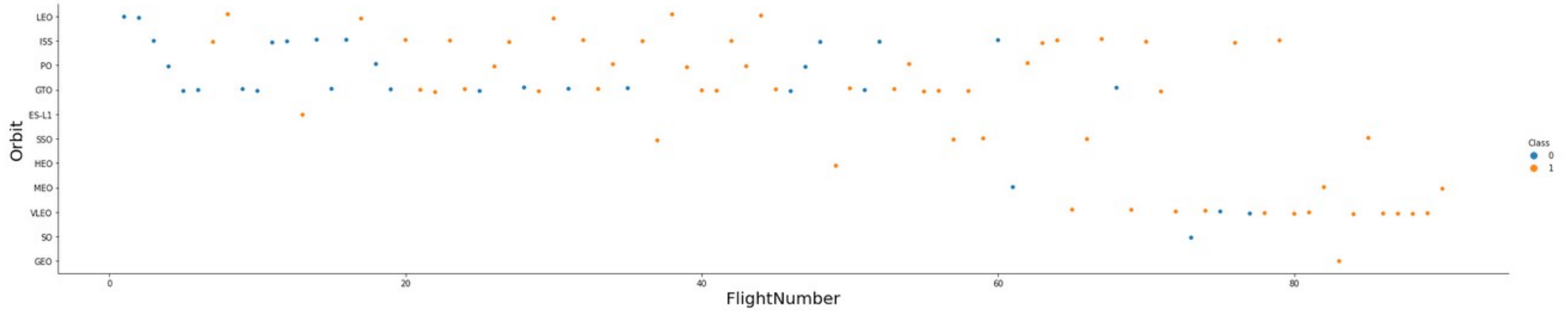
In [16]:
```
# HINT use groupby method on Orbit column and get the mean of Class column
df_barchart = df[['Orbit', 'Class']].groupby(['Orbit']).mean()
df_barchart = df_barchart.reset_index()
df_barchart = df_barchart.sort_values('Class')
sns.barplot(x='Orbit', y='Class', data=df_barchart)
plt.show()
```



Flight Number vs. Orbit Type

- In the LEO orbit the success appears related to the number of flights.
- There is no relation with the flight number in GTO orbit.
- In some cases we are not able to see any relation because of the not enough data.

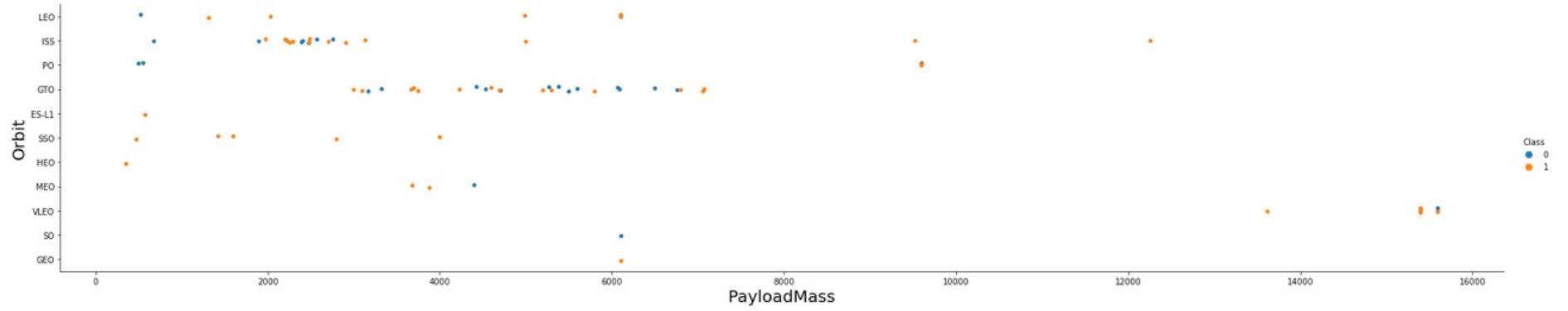
```
# Plot a scatter point chart with x axis to be FlightNumber and y axis to be the Orbit, and hue to be the class value
sns.catplot(y="Orbit", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("FlightNumber",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```



Payload vs. Orbit Type

Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

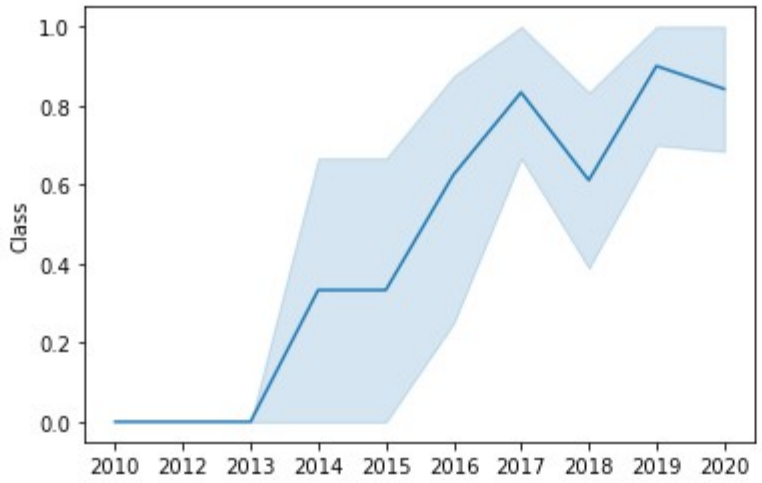
```
# Plot a scatter point chart with x axis to be Payload and y axis to be the Orbit, and hue to be the class value
sns.catplot(y="Orbit", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("PayloadMass",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```



Launch Success Yearly Trend

- Success rate since 2013 kept increasing till 2020
- 95% confidence interval

```
# Plot a line chart with x axis to be the extracted year and y axis to be the success rate
sns.lineplot(data=df, x=year, y='Class')
plt.show()
```



EDA WITH SQL

Exploratory data analysis with
SQL DB2 integrated
in Python with SQLALCHEMY

All Launch Site Names

QUERY: %sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL

Unique launch_site

CCAFS LC-40

CCAFS SLC-40

KSC LC-39A

VAFB SLC-4E

Launch Site Names Begin with 'CCA'

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

```
In [6]: %sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
 * ibm_db_sa://qqm87606:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

Out[6]:

DATE	TIME	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing_outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

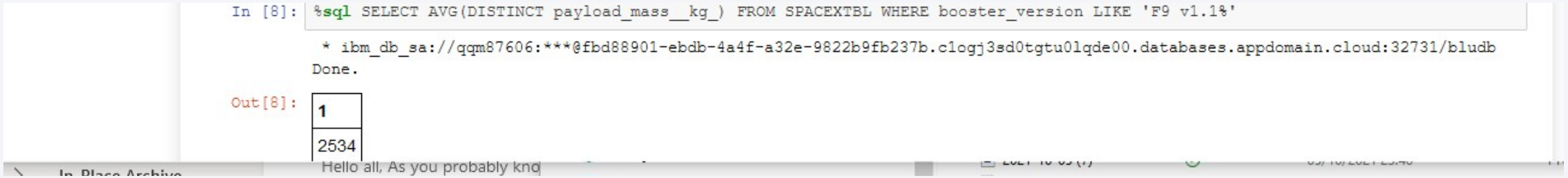
Total Payload Mass

```
In [7]: %sql SELECT SUM(DISTINCT payload_mass__kg_) FROM SPACEXTBL WHERE customer='NASA (CRS)'
 * ibm_db_sa://qqm87606:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
Out[7]:
| 1 |
|---|
| 45596 |
```

The total payload mass carried by boosters launched by NASA (CRS) is 45596.

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1 is 2534.



First Successful Ground Landing Date

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

```
In [ ]: # DATE  Time (UTC)      booster_version         launch_site     payload         payload_mass__kg_       orbit   customer
        mission_outcome         Landing _Outcome
```

```
In [16]: %sql SELECT * FROM SPACEXTBL WHERE landing_outcome like 'Success (ground pad)'
```

```
 * ibm_db_sa://qqm87606:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu01qde00.databases.appdomain.cloud:32731/bludb
Done.
```

Out[16]:

DATE	TIME	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing_outcome
2015-12-22	01:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)
2016-07-18	04:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2017-06-03	21:07:00	F9 FT B1035.1	KSC LC-39A	SpaceX CRS-11	2708	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-08-14	16:31:00	F9 B4 B1039.1	KSC LC-39A	SpaceX CRS-12	3310	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-09-07	14:00:00	F9 B4 B1040.1	KSC LC-39A	Boeing X-37B OTV-5	4990	LEO	U.S. Air Force	Success	Success (ground pad)
2017-12-15	15:36:00	F9 FT B1035.2	CCAFS SLC-40	SpaceX CRS-13	2205	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2018-01-08	01:00:00	F9 B4 B1043.1	CCAFS SLC-40	Zuma	5000	LEO	Northrop Grumman	Success (payload status unclear)	Success (ground pad)

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 [19]: %sql SELECT * FROM SPACEXTBL WHERE (payload_masa__kg_>4000 AND payload_masa__kg_<6000)
```

```
 * ibm_db_sa://qqm87606:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu01qde00.databases.appdomain.cloud:32731/bludb
Done.
```

Out[19]:

DATE	TIME	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing_outcome
2014-08-05	08:00:00	F9 v1.1	CCAFS LC-40	AsiaSat 8	4535	GTO	AsiaSat	Success	No attempt
2014-09-07	05:00:00	F9 v1.1 B1011	CCAFS LC-40	AsiaSat 6	4428	GTO	AsiaSat	Success	No attempt
2015-03-02	03:50:00	F9 v1.1 B1014	CCAFS LC-40	ABS-3A Eutelsat 115 West B	4159	GTO	ABS Eutelsat	Success	No attempt
2015-04-27	23:03:00	F9 v1.1 B1016	CCAFS LC-40	Turkmen 52 / MonacoSAT	4707	GTO	Turkmenistan National Space Agency	Success	No attempt
2016-03-04	23:35:00	F9 FT B1020	CCAFS LC-40	SES-9	5271	GTO	SES	Success	Failure (drone ship)
2016-05-06	05:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-08-14	05:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2017-03-16	06:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2017-09-07	14:00:00	F9 B4 B1040.1	KSC LC-39A	Boeing X-37B OTV-5	4990	LEO	U.S. Air Force	Success	Success (ground pad)
2017-10-11	22:53:00	F9 FT B1031.2	KSC LC-39A	SES-11 / EchoStar 105	5200	GTO	SES EchoStar	Success	Success (drone ship)
2018-01-08	01:00:00	F9 B4 B1043.1	CCAFS SLC-40	Zuma	5000	LEO	Northrop Grumman	Success (payload status unclear)	Success (ground pad)
2018-01-31	21:25:00	F9 FT B1032.2	CCAFS SLC-40	GovSat-1 / SES-16	4230	GTO	SES	Success	Controlled (ocean)
2018-06-04	04:45:00	F9 B4 B1040.2	CCAFS SLC-40	SES-12	5384	GTO	SES	Success	No attempt
2018-08-07	05:18:00	F9 B5 B1046.2	CCAFS SLC-40	Merah Putih	5800	GTO	Telkom Indonesia	Success	Success
2018-11-15	20:46:00	F9 B5 B1047.2	KSC LC-39A	Es hail 2	5300	GTO	Es hailSat	Success	Success
2018-12-23	13:51:00	F9 B5B1054	CCAFS SLC-40	GPS III-01	4400	MEO	USAF	Success	No attempt
2019-02-22	01:45:00	F9 B5 B1048.3	CCAFS SLC-40	Nusantara Satu, Beresheet Moon lander, S5	4850	GTO	PSN, SpaceIL / IAI	Success	Success
2019-05-12	14:17:00	F9 B5 B1051.2	VAFB SLC-4E	RADARSAT Constellation, SpaceX CRS-18	4200	SSO	Canadian Space Agency (CSA)	Success	Success
2020-06-30	20:10:46	F9 B5B1060.1	CCAFS SLC-40	GPS III-03, ANASIS-II	4311	MEO	U.S. Space Force	Success	Success
2020-07-20	21:30:00	F9 B5 B1058.2	CCAFS SLC-40	ANASIS-II, Starlink 9 v1.0	5500	GTO	Republic of Korea Army, Spaceflight Industries (BlackSky)	Success	Success
2020-11-05	23:24:23	F9 B5B1062.1	CCAFS SLC-40	GPS III-04 , Crew-1	4311	MEO	USSF	Success	Success

There were a number of boosters with the payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

Calculate the total number of successful and failure mission outcomes

List the total number of successful and failure mission outcomes

[24]:
```
%sql SELECT COUNT(mission_outcome), mission_outcome FROM SPACEXTBL GROUP BY mission_outcome
```

```
 * ibm_db_sa://qqm87606:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/b
Done.
```

t[24]:

1	mission_outcome
1	Failure (in flight)
99	Success
1	Success (payload status unclear)

Boosters Carried Maximum Payload

There were 12 boosters which have carried the maximum payload mass

```
In [34]: %sql SELECT booster_version FROM SPACEXTBL WHERE payload_mass__kg_=(SELECT max(payload_mass__kg_) FROM SPACEXTBL)
         # SELECT max(payload_mass__kg_) FROM SPACEXTBL
```

```
 * ibm_db_sa://qqm87606:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

Out[34]:

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

There were 2 failed landing_outcomes in drone ship in year 2015

```
In [39]: %sql SELECT landing_outcome, booster_version, launch_site FROM SPACEXTBL WHERE DATE LIKE '2015%' AND landing_outcome LIKE 'Failure%'

 * ibm_db_sa://qqm87606:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

Out[39]:

landing_outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

There were 5 Failure (drone ship) and 3 Success (ground pad) landing outcomes between the date 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

In [43]:
```
%sql SELECT COUNT(landing_outcome), landing_outcome FROM SPACEXTBL WHERE (DATE BETWEEN '2010-06-04' AND '2017-03-20') AND (la
nding_outcome like 'Failure (drone ship)' or landing_outcome like 'Success (ground pad)') GROUP BY landing_outcome
```

```
 * ibm_db_sa://qqm87606:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

Out[43]:

1	landing_outcome
5	Failure (drone ship)
3	Success (ground pad)

Reference Links

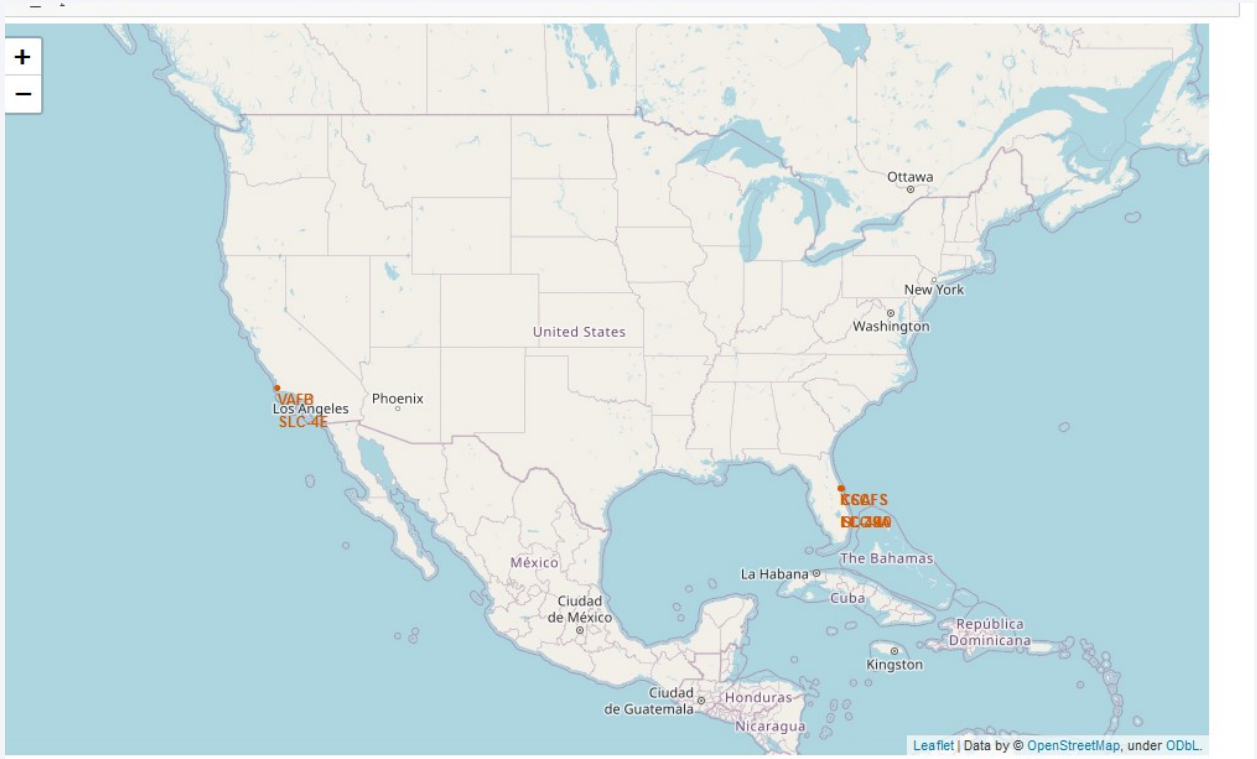
Section 4

Launch Sites Proximities Analysis

All launch sites on a map

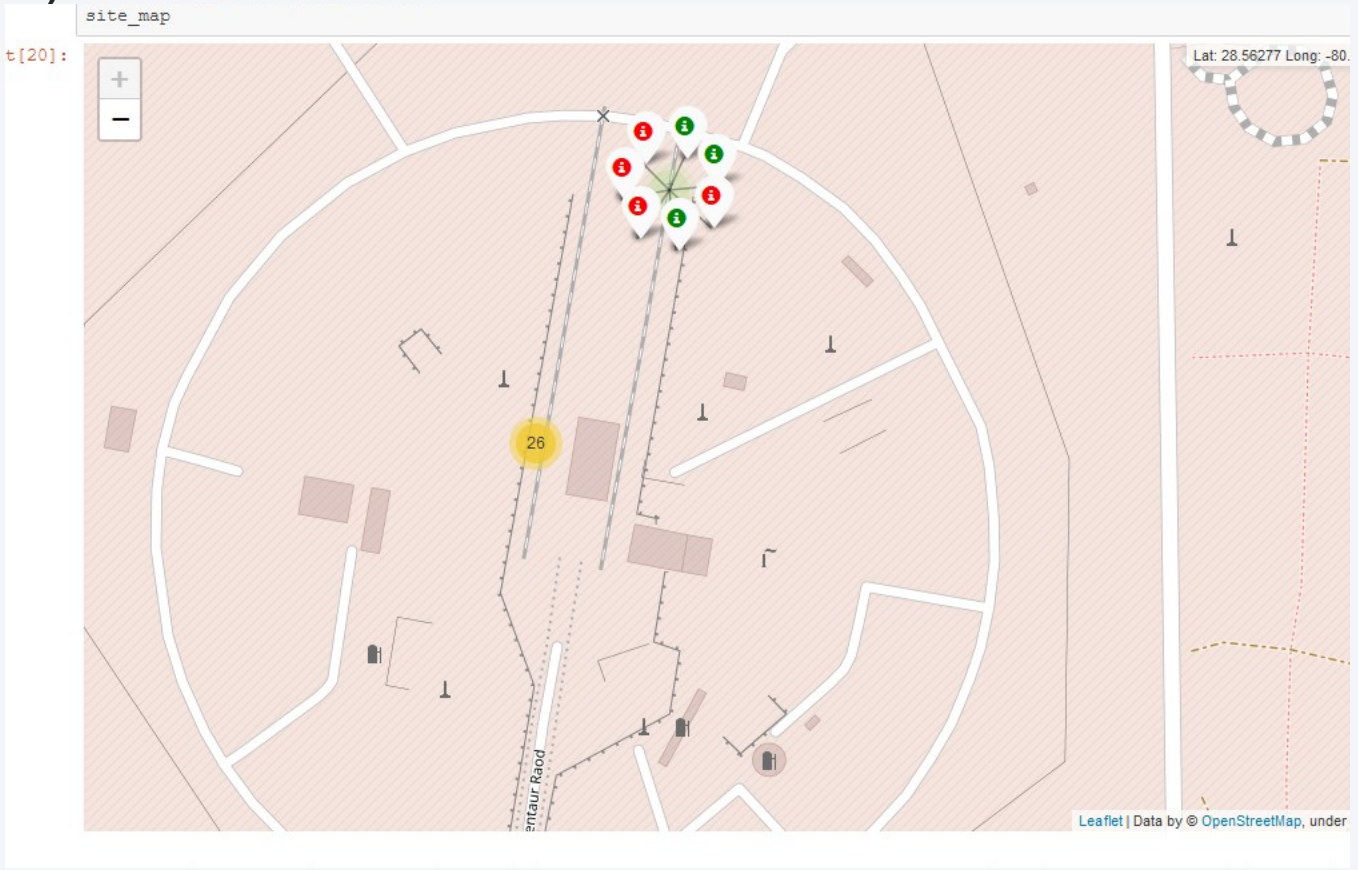
All launch sites are in proximity to the Equator line.

All launch sites are very close proximity to the coast.



The success/failed launches for each site on the map

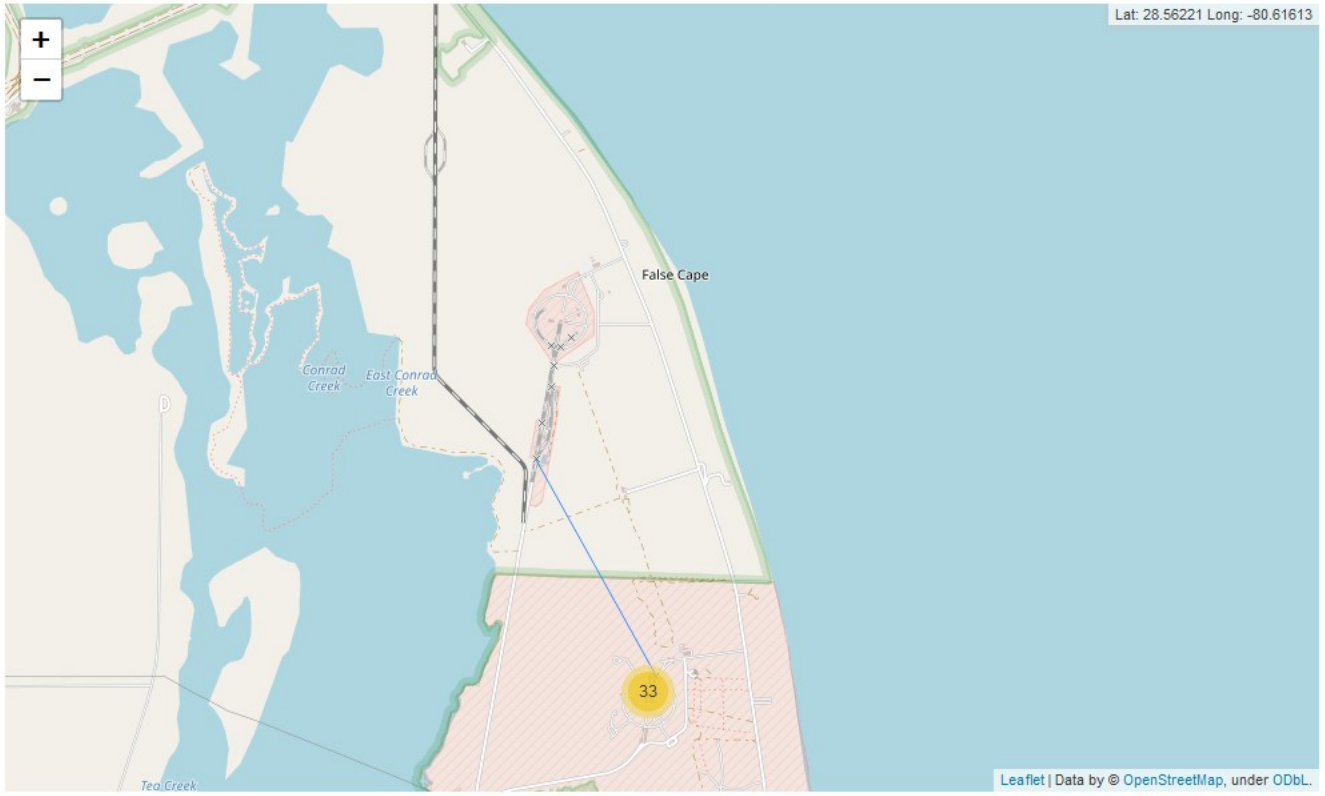
The color indicate which launch sites have relatively high success rate (green).

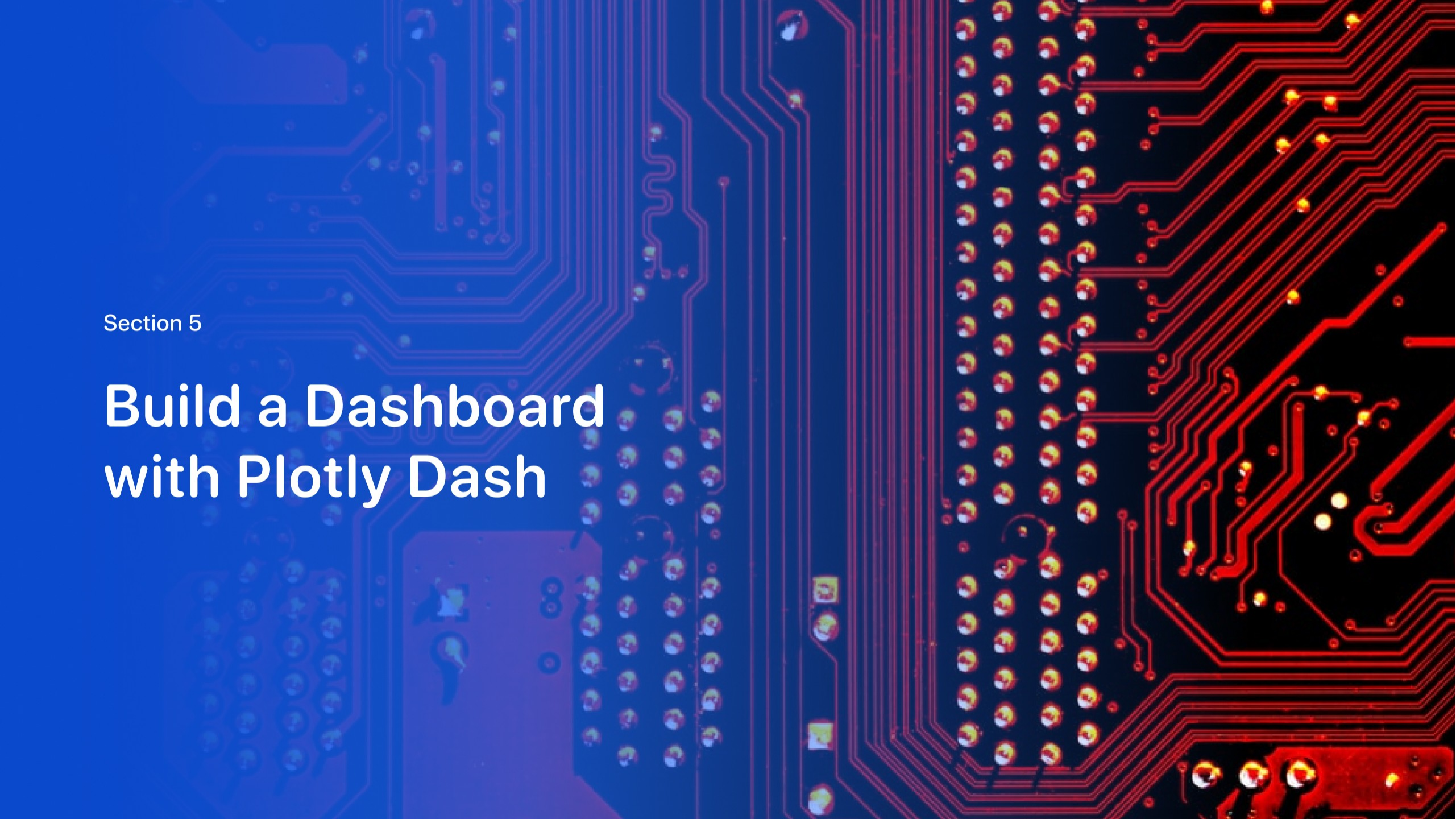


The distances between a launch site to its proximities

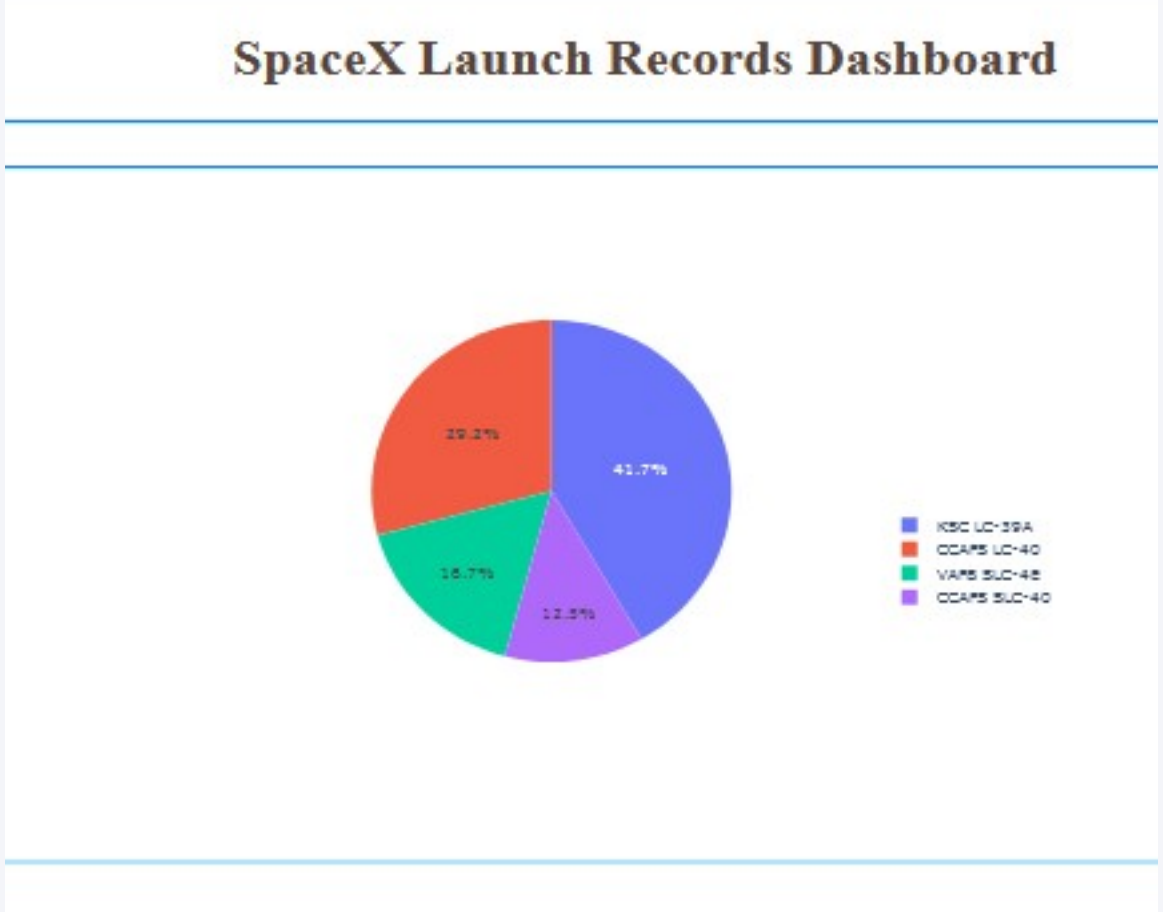
```
[30]: # Create a `folium.PolyLine` object using the coastline coordinates and launch site coordinate
      # lines=folium.PolyLine(locations=coordinates, weight=1)
      lines=folium.PolyLine(locations=coordinates, weight=1)
      site_map.add_child(lines)
```

t[30]:




Section 5
Build a Dashboard with Plotly Dash

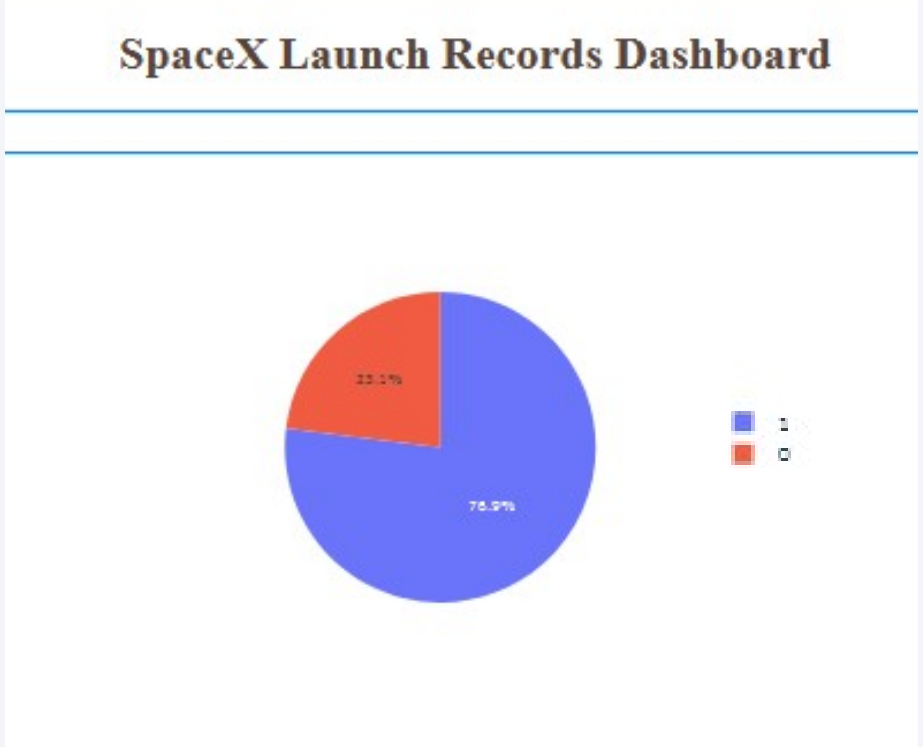
Success count for all launch sites



The distribution of successful landing across all launch sites.

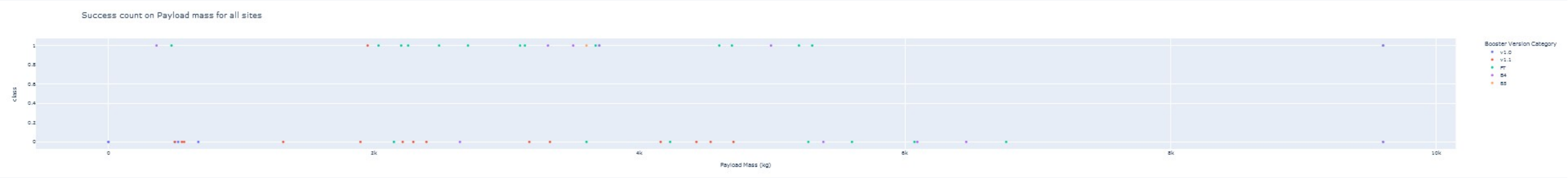
The launch site with highest lunch success ratio.

The lunch site KSC LC-39A is with the highest ratio.



Success count on Playload mass for all sites

- It is shown that most of the booster version with success have the weight between 2k and 4k.

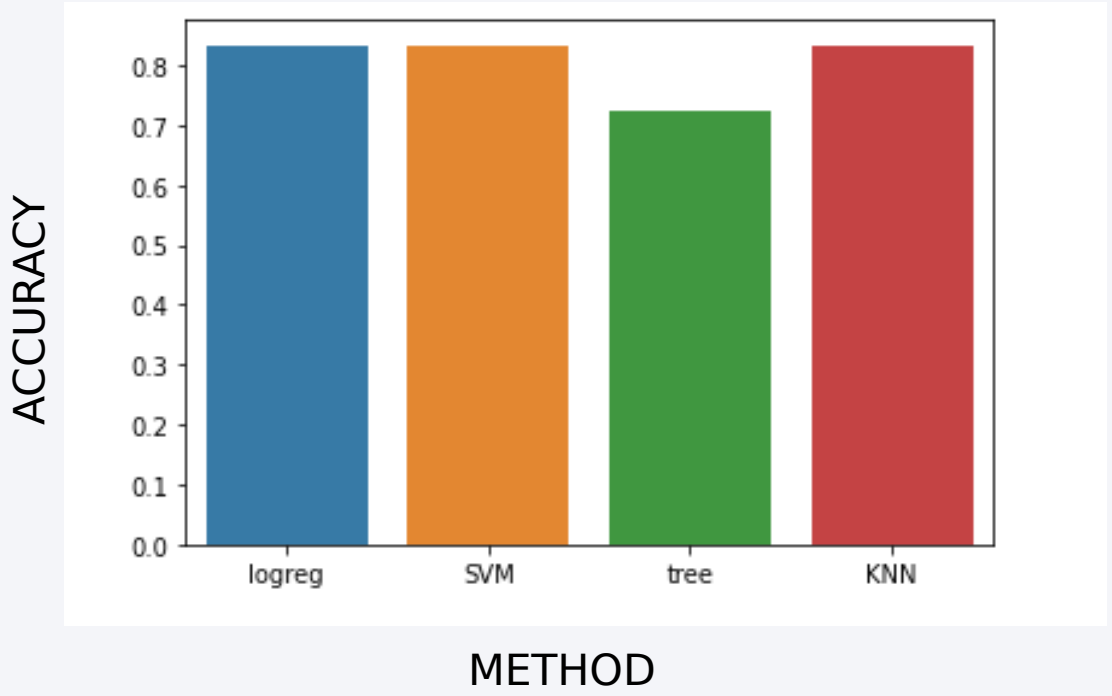


Section 6

Predictive Analysis (Classification)

Classification Accuracy

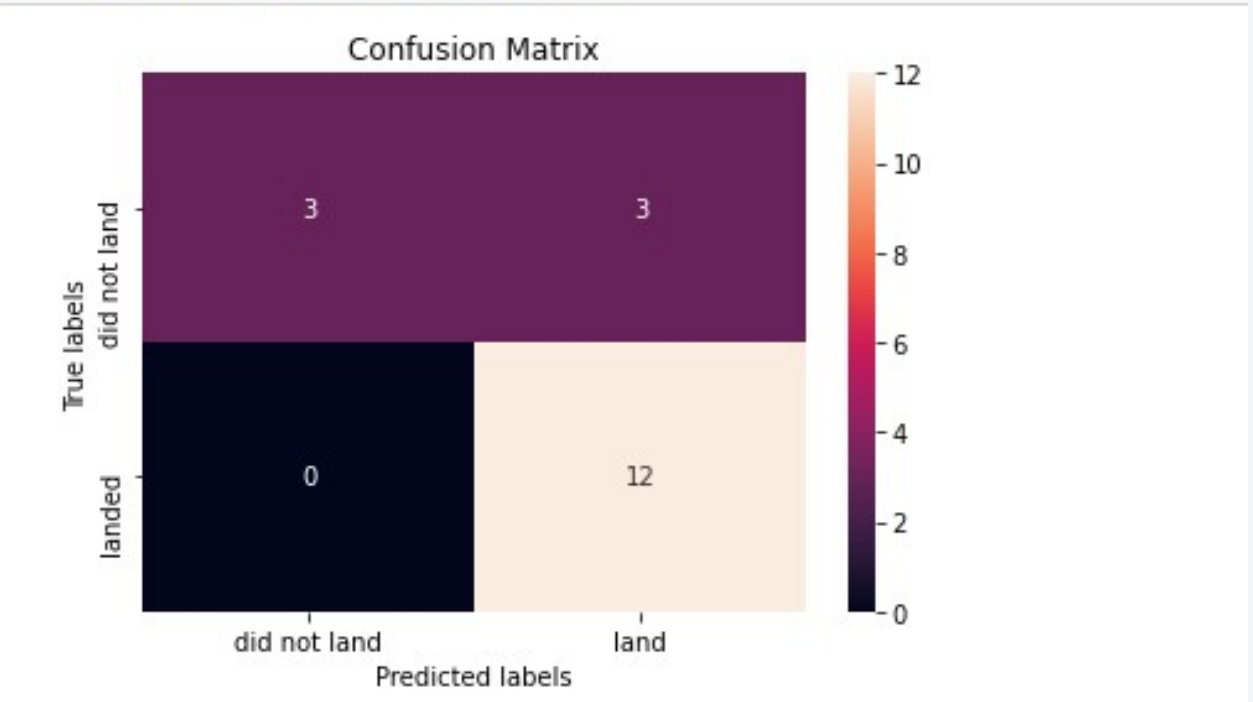
- We can see that accuracy in only tree model is lower than the rest.



Confusion Matrix

- The confusion matrix of the K nearest neighbor.

- Examining the confusion matrix on the right, we can see that our models can distinguish different classes.
- One common issue with all models is False Positives. A false positive is the misclassification of a successful landing as a failed to land.
- We have three landings that are false positives as seen in the top right cell.



Conclusions

- Task: develop machine learning model for SpaceY who wants to bid against SpaceX.
- The goal was to predict when Stage 1 will successfully land to save ~$100 minion USD.
- Data for the investigation were from API and Wikipedia.
- Investigations: data cleaning, sorting and selecting features, dashboard visualization
- Machine learning model with accuracy 80%.
- This model can be used to predict relatively high accuracy whether a lunch will have a successful stage 1 landing.
- To future investigations more data will be needed to determine the best machine learning model and improve accuracy.

Appendix

All used code in this project are available on GitHub.

Thank you!