



IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Richard Chandra>
<March 6th, 2022>



Outline

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

Executive Summary

- In this presentation, we, Space Y, try to compete with Space X in commercialize affordable space travel
- Space X could give a lower number because they land the first stage of their launched rocket
- We analyze Space X's launches to determine factors that contribute to the result for the first stage to land and try to use machine learning classification algorithm to predict the outcome
- The methodology for this analysis are:
 - Data collection (with Space X API, web scrapping, and wrangling data)
 - Exploratory Data Analysis (with SQL, Visualization, and Interactive Dashboard)
 - Predict the outcome with several classification algorithm (Logistic regression, Decision Tree, Support Vector Machine, k-Nearest Neighbour)

Executive Summary

- Result from analysis:
 - We can see that every year, the scope of orbit that being explored become more varied and also increase on successes
 - We find out that KSC LC 39-A launch site have the most highest success rate compared to other launch site
 - One of the most widely use drone ship to help landing the first stage is drone ship with booster version F4
- We can build a prediction for the landing outcome with several machine learning algorithm
 - The accuracy for several machine learning algorithm is 83.3% with type 1 error as the error of the prediction

Introduction - Background

- We are entering the commercial space age, companies are trying to make space travel affordable for everyone.
- One of the most successful company is Space X, which accomplishment include sending spacecraft to the International Space Station.
- Space X claims that their Falcon 9 rocket launches on its website with a cost of 62 million dollars, more cheaper than other providers which cost upwards of 165 million dollars each, because they can reuse the first stage by landing it.



Introduction - Problem

- Our company, Space Y, would like to compete with Space X.
- We want to know the price of each launch and predict the if Space X will reuse the first stage.



Section 1

Methodology

Methodology

Executive Summary

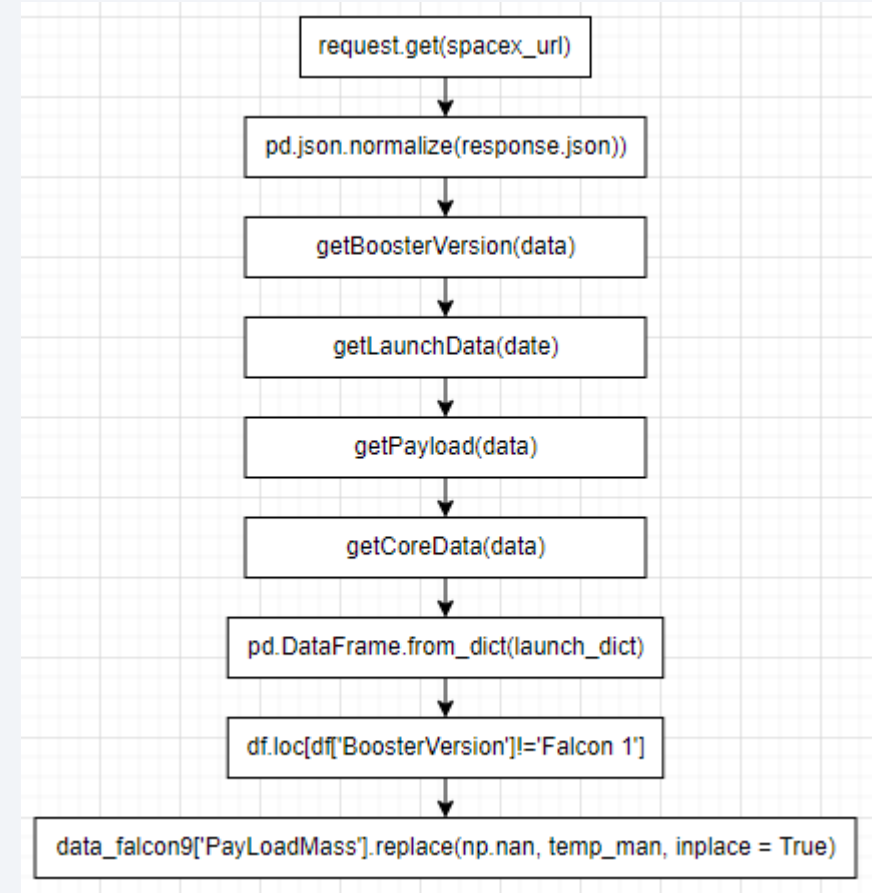
- Data collection methodology:
 - Describe how data was collected
- Perform data wrangling
 - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
 - How to build, tune, evaluate classification models

Data Collection

- There are two methods that we use to collect Falcon 9 data:
 - First, we collect the data by using SpaceX API. The collected data was also cleaned the data by replacing the missing data (NaN value) with the mean of data.
 - Second, we scrap Falcon 9 launch records with BeautifulSoup library. We extract Falcon 9 launch records HTML table from Wikipedia. The data was parsed the table and convert it into Pandas data frame.

Data Collection – SpaceX API

- We collect data from given SpaceX API
- The process are:
 - Request data from SpaceX API with the given URL
 - Turning the given data and turn it into Pandas dataframe using `.json_normalize()`
 - Use the API again to get information about the launches using the IDs given for each launch
 - From given function, we extract all of the data and turn it into dataframe
 - Filter the dataframe to only include Falcon 9 launches
 - Dealing with missing value

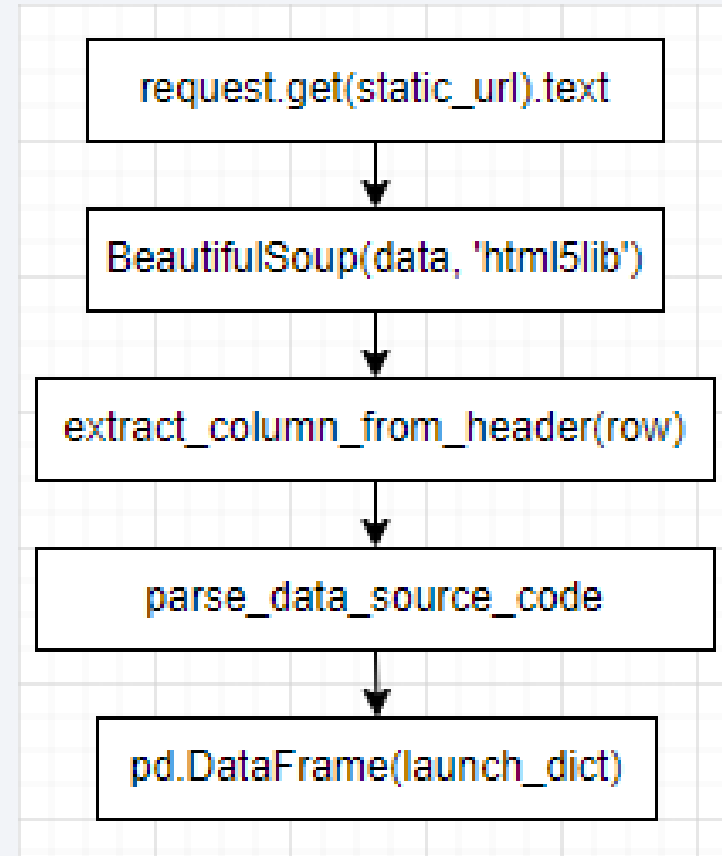


Data Collection – SpaceX API - Result

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Leg	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857
...
89	86	2020-09-03	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	2	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	8	B1060	-80.603956	28.608058
90	87	2020-10-06	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	3	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	8	B1058	-80.603956	28.608058
91	88	2020-10-18	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	6	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	10	B1051	-80.603956	28.608058
92	89	2020-10-24	Falcon 9	15600.0	VLEO	CCSFS SLC 40	True ASDS	3	True	True	True	5e9e3033383ecbb9e534e7cc	5.0	8	B1060	-80.577366	28.561857
93	90	2020-11-05	Falcon 9	3681.0	MEO	CCSFS SLC 40	True ASDS	1	True	False	True	5e9e3032383ecb6bb234e7ca	5.0	3	B1062	-80.577366	28.561857

Data Collection - Scraping

- In this process, we collect the data from Wikipedia page titled “List of Falcon 9 and Falcon Heavy Launches”
- The process are:
 - Request the Falcon9 Launch HTML Page
 - Create a BeautifulSoup object from the HTML response
 - Extract the information
 - Parse the data and turn it into Pandas dataframe



Data Wrangling

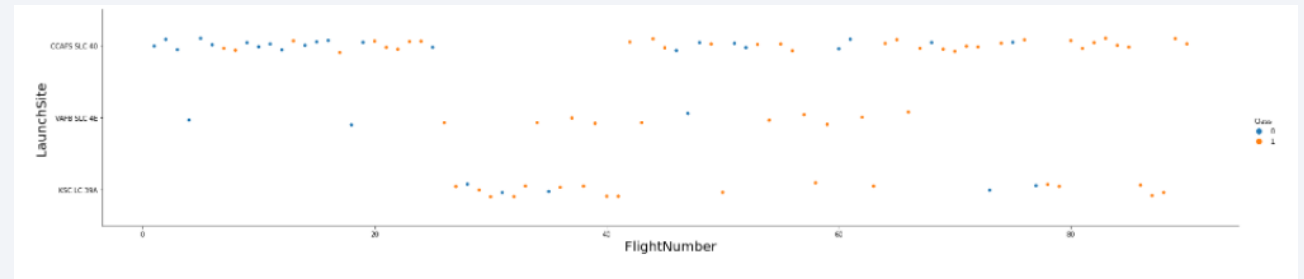
- In this part, we try to do Exploratory Data Analysis and determining training labels
- The process are:
 - We try to find the number of launch based on the Launch Site
 - The launches also analyzed from the orbit they were launched at
 - We calculate the number and occurrence of mission outcome per orbit type
 - Create the landing outcome label from Outcome column

```
df['Class']=landing_class  
df[['Class']].head(8)
```

	Class
0	0
1	0
2	0
3	0
4	0
5	0
6	1
7	1

EDA with Data Visualization

- We use Cat plot, Lineplots, and Bar plot to do the Exploratory Data Analysis and do Feature Engineering for predicting the future event
 - The cat plot used color to differ between success first stage landing and failed landing
- We use matplotlib and seaborn to help us visualize the data
- The details:
 - Catplots:
 - Flight Number vs Pay Load Mass
 - Pay Load Mass vs Launch Site
 - Flight Number vs Orbit Type
 - Pay Load Mass vs Orbit Type
 - Bar plot:
 - Orbit vs Success Rate (class)
 - Line Plot:
 - Launch Success Yearly Trend



Example of Cat Plot (Launch Site vs Flight Number) with Class (Success Rate) separate their color

EDA with SQL

- Another way to explore our data is by using SQL in Python with the help of db2 library
- In this assignment, there are several SQL query that the result are:
 - Display the name 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 successful landing outcome in ground pad was achieved

EDA with SQL

- 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
- List the names of the booster_versions which have carried the maximum payload mass
- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- 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

Build an Interactive Map with Folium

- In this section, I mark all launch sites on map, mark the success/failed launches for each site on the map, and calculate the distances between a launch site to its proximities
- We use Folium to show the real map in Python
- We added several mark to point the location on map so we have the knowledge about the launch sites position
 - We also try to add several line to pinpoint the distance between a launch site and the nearby infrastructure



Build a Dashboard with Plotly Dash

- In this section, there is two type of data that I have to show
 - Total success launches by site
 - We have four different launch sites and we would like to see which one has the largest success count
 - We show the data with pie chart
 - We can choose to show all sites or selecting specific site with drop bar
 - Correlation between payload and success for all sites
 - We wanted to find if variable payload is correlated to mission outcome
 - We show the data with scatter plot
 - We can adjust the minimum and maximum of payload (0-10000 kg)
- The dashboard was build with Plotly Dash

Predictive Analysis (Classification)

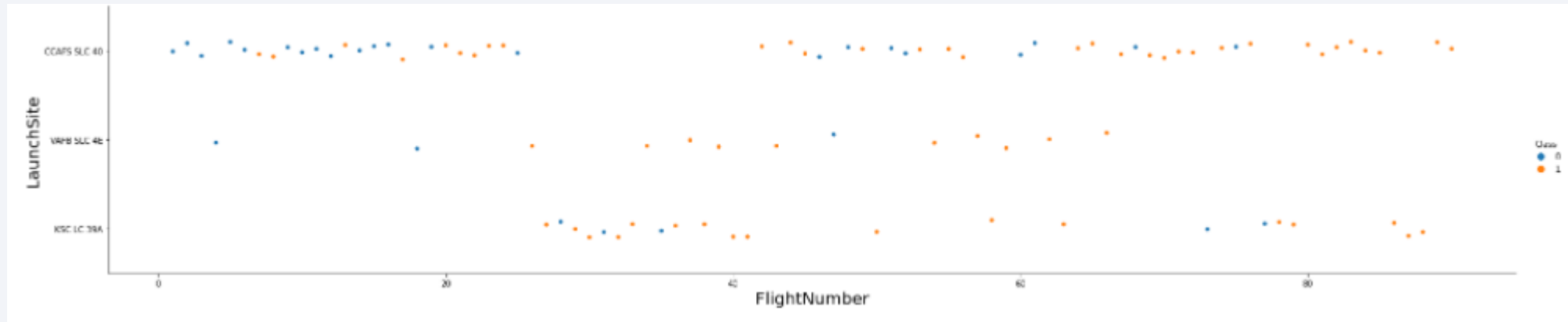
- In this section, we try to predict Space X Falcon 9 First Stage Landing Prediction
- Processes:
 - Import library and load database
 - Create NumPy array from column Class in data
 - Standardization of data
 - Split the data to train set and test set
 - Run a classification with: Logistic regression, Support Vector Machine, Decision Tree, and k Nearest Neighbors
 - We evaluate every model's score and compare each model to the others

The background of the slide is an abstract composition. It features a dark blue base color. Overlaid on this are numerous diagonal streaks in shades of red and cyan. A faint, light blue grid pattern is also visible, particularly in the lower half of the image. The overall effect is dynamic and technological.

Section 2

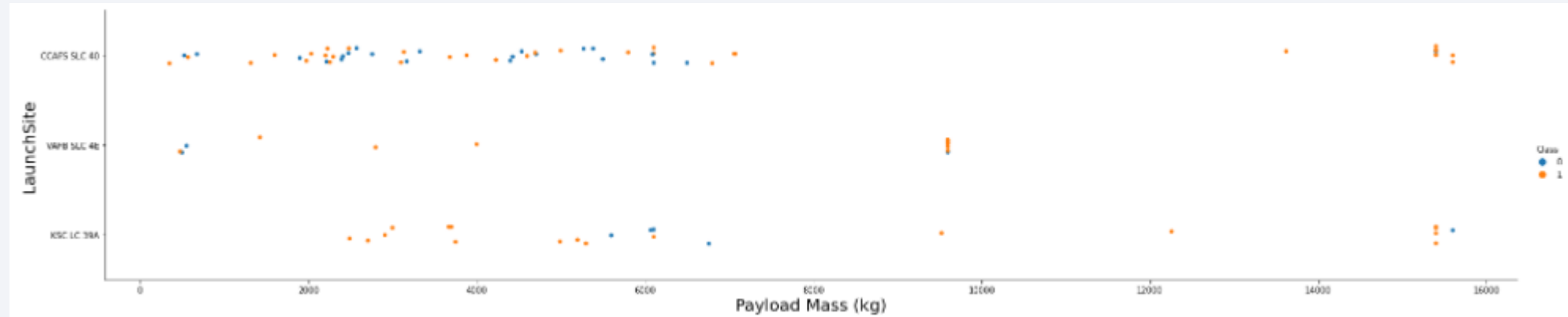
Insights drawn from EDA

Flight Number vs. Launch Site



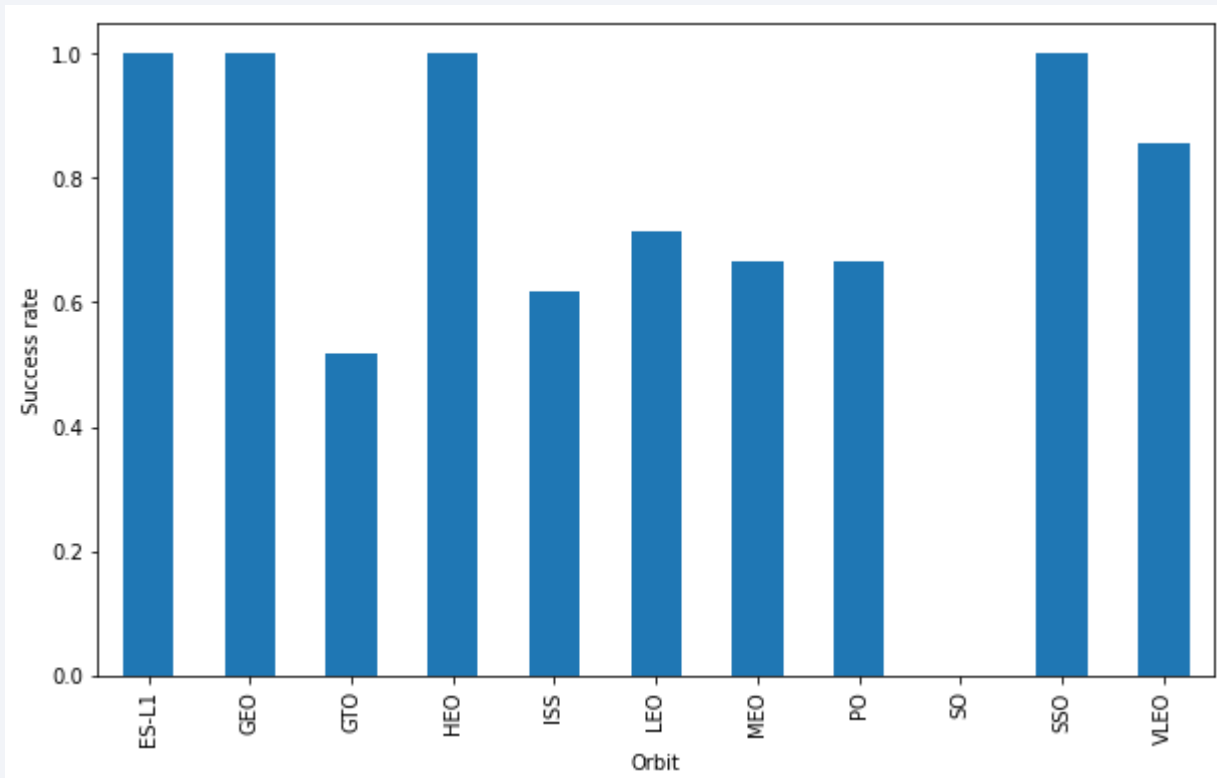
- The scatter plot shows us three launch sites: CCAFS SLC 40, VAFB SLC 4E, KSC LC 39A, with successful first stage landing noted with orange dot and failed landing with blue dot, and being compared to flight number
- Several things that we can learn from this scatter plot
 - CCAFS SLC 40 have the most failure on landing it's first stage
 - In CCAFS case, we see failure more frequently occur on the early flight number
 - VAFB SLC 4E send the least number of flight
 - Most of the later flight number have a success rather than the early flight number

Payload vs. Launch Site



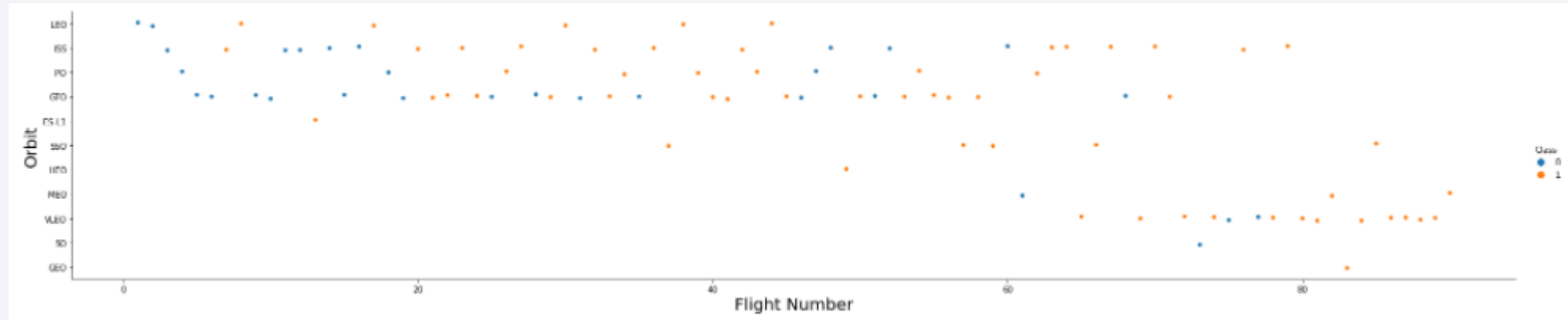
- The scatter plot shows us three launch sites: CCAFS SLC 40, VAFB SLC 4E, KSC LC 39A, with successful first stage landing noted with orange dot and failed landing with blue dot, and being compared to payload mass (kg)
- Several things that we can learn from this scatter plot
 - Most of the launch happened with payload mass below 8000 kg
 - CCAFS SLC 40 launch the most flight which have payload mass below 8000 kg
 - Launches that have payload mass higher than 8000 kg are more likely to succeed than below 8000 kg

Success Rate vs. Orbit Type



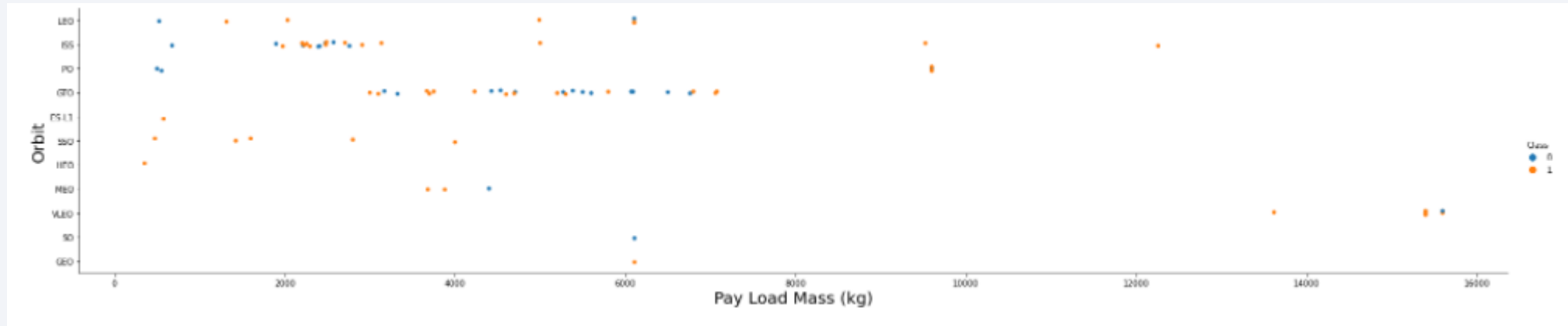
- We use bar chart to analyze between success rate and orbit type
- From the chart, we can see that
 - SO have the lowest success rate with score 0
 - ES-L1, GEO, HEO, and SSO have the maximal success rate
 - The second highest success rate was flight to the VLEO

Flight Number vs. Orbit Type



- The scatter plot shows us flight orbit and being compared to flight number, with orange colour indicate the first stage component successfully landing and the blue marks the failure
- Several things that we can learn from this scatter plot
 - Most of the earlier launch happened to occur around LEO, ISS, PO, and GEO
 - Earlier launch have more failure than the later launch
 - Flight number above 60 have more scope of orbit that being explored
 - All flight with flight number above 80 have successfully landed it's first stage

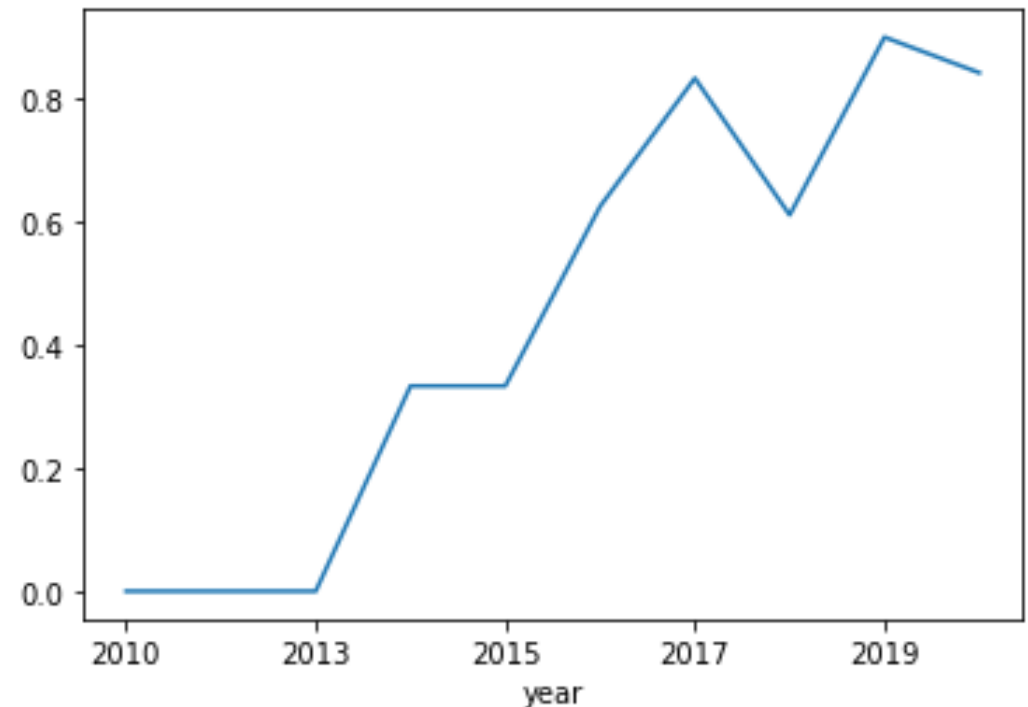
Payload vs. Orbit Type



- The scatter plot shows us flight orbit and being compared to payload mass, with orange colour indicate the first stage component successfully landing and the blue marks the failure
- Several things that we can learn from this scatter plot
 - GEO has the most variance on payload mass
 - Most of the flight occur with payload mass below 8000 kg
 - The heaviest flight was flight with VLEO as the orbit
 - Most ISS orbit flight have a payload around 2000 – 4000 kg

Launch Success Yearly Trend

- This is a line chart with x axis to be Year and y axis to be Average Success Rate
- From this data, we can conclude
 - Starting from 2013, the average success rate of first stage landing is increasing for each year, except for 2018
 - 2015 to 2017 is the highest increase on average success rate
 - Starting from 2017, the success rates are higher than 0.6



All Launch Site Names

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

- This is all launch site names that we observe the data from
- We can achieve by using SELECT DISTINCT to show use unique launch_site from database SPACEXTBL

```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL;
```

Launch Site Names Begin with 'CCA'

DATE	time__utc__	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

- This show 5 records that have launch site names begin with 'CCA'
- We can achieve this with this query by using LIKE to find CCA word and limit the result to 5 using LIMIT

```
%sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```


Total Payload Mass

- We can calculate the total payload mass with SQL query

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)';
```

We can use SUM to count the total payload mass and using WHERE clause to only process data that have customer value 'NASA (CRS)'

- The result of calculation is (in kg)

total_payload
45596

Average Payload Mass by F9 v1.1

- We can calculate the average payload mass of flight that have booster version F9 v1.1 with this query

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION LIKE 'F9 v1.1%';
```

We can achieve these result by using LIKE to find which data's booster_version is 'F9 v1.1' and use AVG to find the average number of payload mass

- The result is (in kg)

avg_payload
2534

First Successful Ground Landing Date

- We can find the first successful ground landing date with this query

```
%sql SELECT MIN(DATE) FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Success';
```

We can receive the outcome by comparing the searching in every value of landing_outcome, if there is any success word

- This is the date where for the first time, ground landing occurred successfully

1
2018-07-22

Successful Drone Ship Landing with Payload between 4000 and 6000

booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 B4 B1043.1
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1046.3
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

- This is the list of boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000 kg
- We can achieve these result by using SQL query

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE LANDING__OUTCOME  
LIKE '%Success%' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

We are using LIKE and BETWEEN to find which one have word “Success” in their landing_outcome column and in interval of 4000 to 6000 for pay load mass

Total Number of Successful and Failure Mission Outcomes

- This is the list of the total number of successful and failed mission outcomes

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

- We can achieve it with using SQL query

```
%sql SELECT MISSION_OUTCOME, COUNT(*) AS TOTAL_NUM FROM SPACEXTBL GROUP BY MISSION_OUTCOME;
```

We are using COUNT and aggregate it with GROUP BY on mission_outcome to gain count how many success and failed flight occurred

Boosters Carried Maximum Payload

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

- This the list of the names of the booster version which have carried the maximum payload
- We can achieve this result with using SQL query

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL  
WHERE PAYLOAD_MASS__KG_ =  
(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

We are using subquery to return a single value for maximum payload, and then compared it with all the data's payload mass

2015 Launch Records

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- This is the list of failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- We can achieve it using SQL query

```
%sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL
```

```
WHERE LANDING__OUTCOME LIKE '%Fail%' AND YEAR(DATE) = 2015;
```

We can use wildcards % with LIKE to find which landing outcome was failed, and combine it with extracting the YEAR from DATE to match our desired year

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

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

- This is the count rank of landing_outcomes between 2010-06-01 and 2017-03-20 in descending order
- We can achieve these result by using SQL query that use COUNT and GROUP BY to count the landing outcome, BETWEEN to keep the processed data form 2010-06-01 to 2017-03-20, and ORDER BY to ranking the data by landing_count descendingly

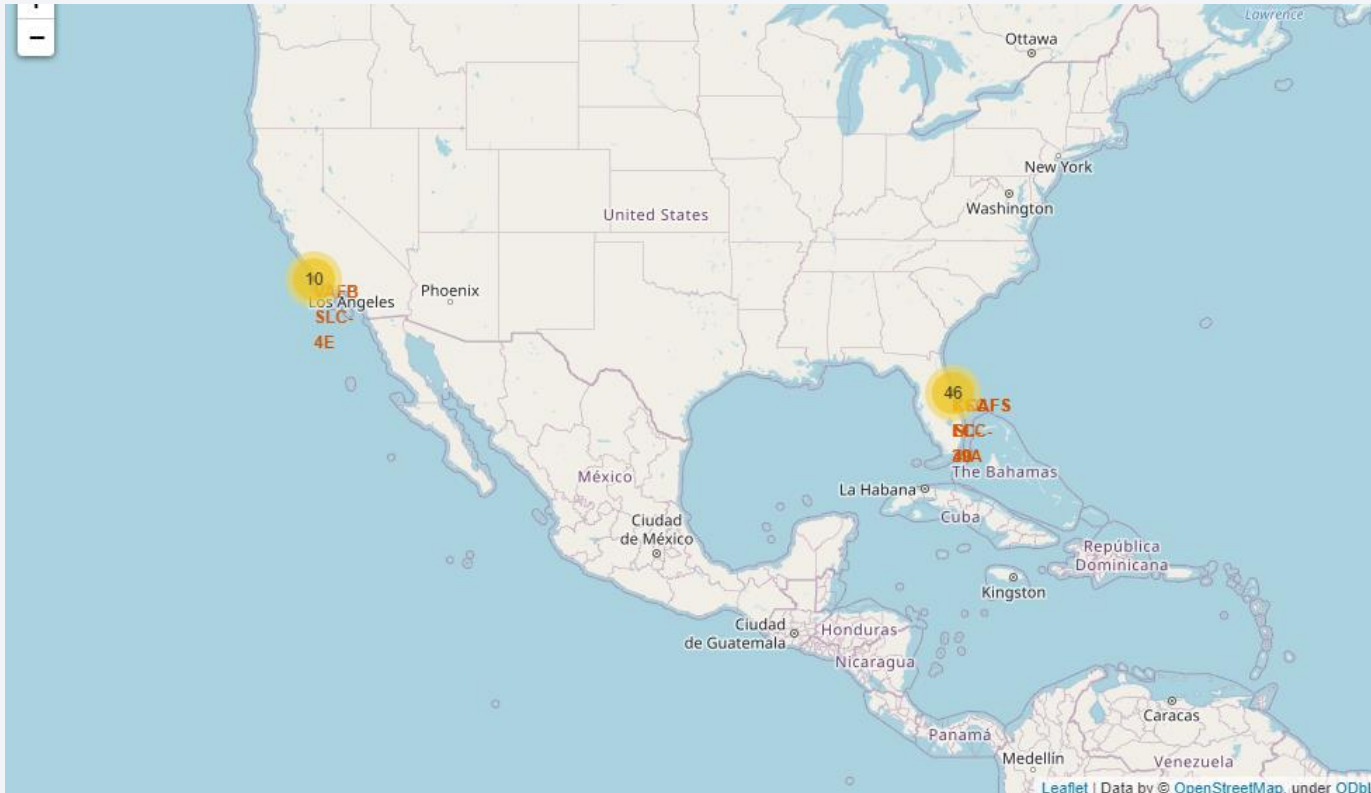
```
%sql SELECT LANDING__OUTCOME, COUNT(*) AS LANDING_COUNT FROM SPACEXTBL  
      WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY  
      LANDING__OUTCOME ORDER BY LANDING_COUNT DESC;
```

A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The background is a deep blue gradient.

Section 3

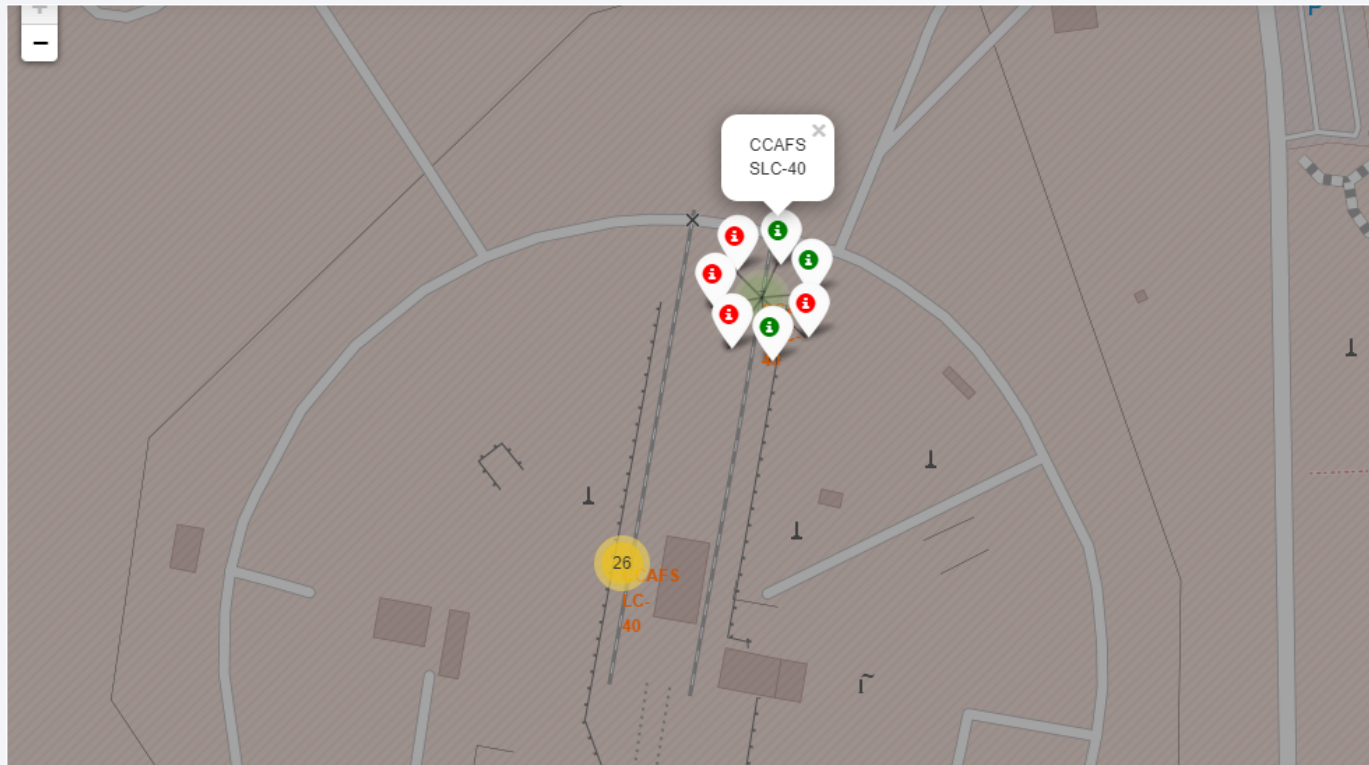
Launch Sites Proximities Analysis

All Launches Sites on Map



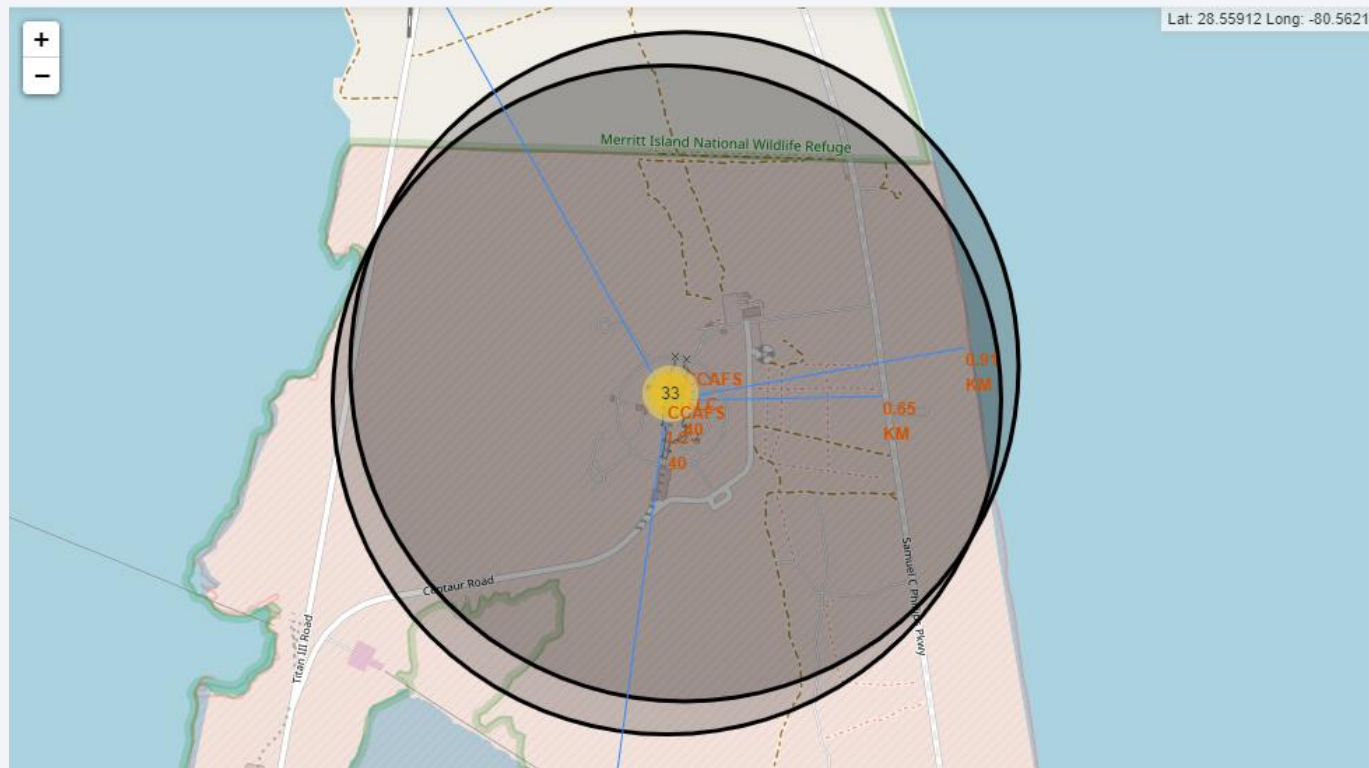
- This is all the location of launch sites on map
- With this map, we can visualize the location of launch sites. Two location that build to be launch site is California and Florida
- Most of the launch site build on Florida

Success and failed launches for each site on the map



- This map shows location of the launch site with the number of launches the launch site ever done
- We can click the circle that indicate the launch site to see the icon that indicate the result of the launch, green for successful launch, red for failed launch

Distances Between a Launch Site to Its Proximities



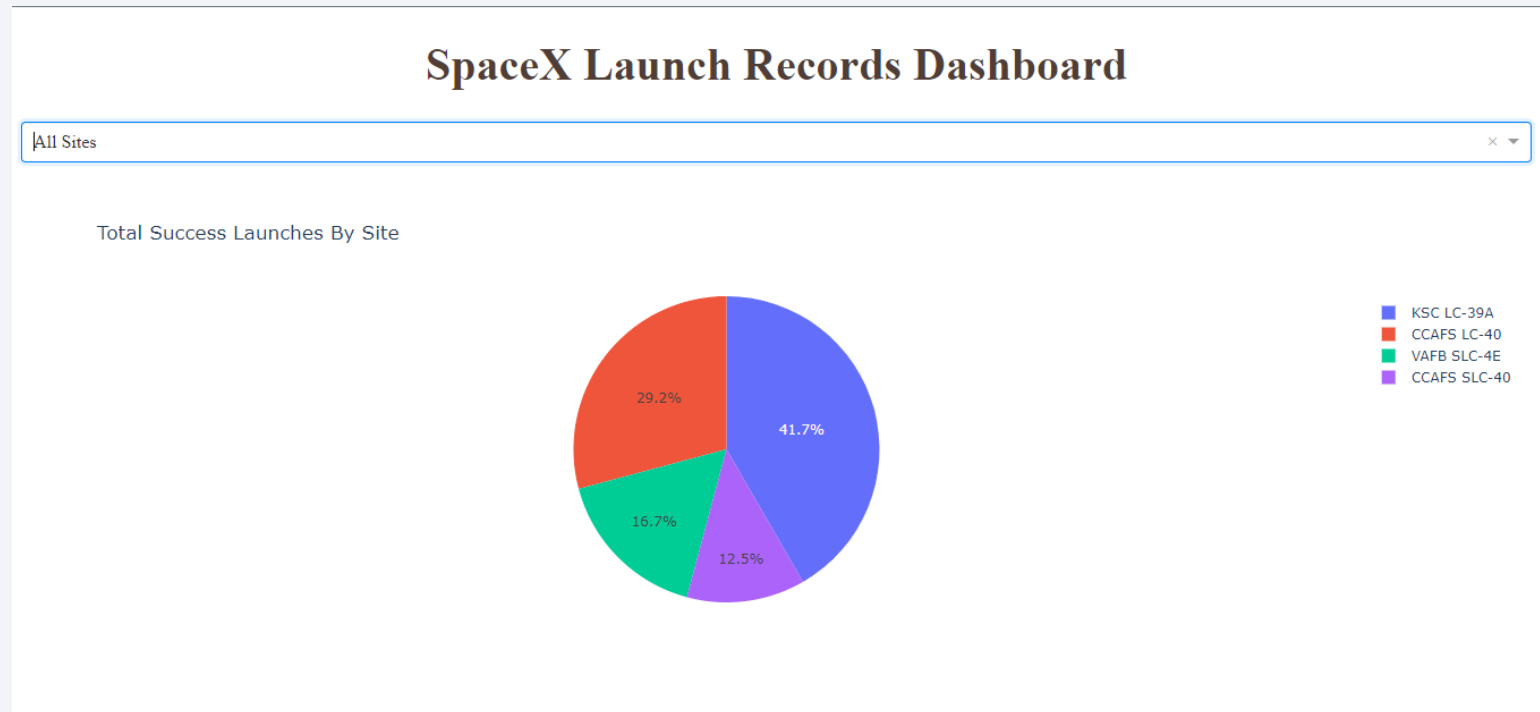
- This map shows location of the launch site with a line to indicate the distance between launch site and its proximities
- We can see that launch site is far from the city but close to coastline, highway, and railway
- This is intuitively correct as launch site can be dangerous if placed near the city



Section 4

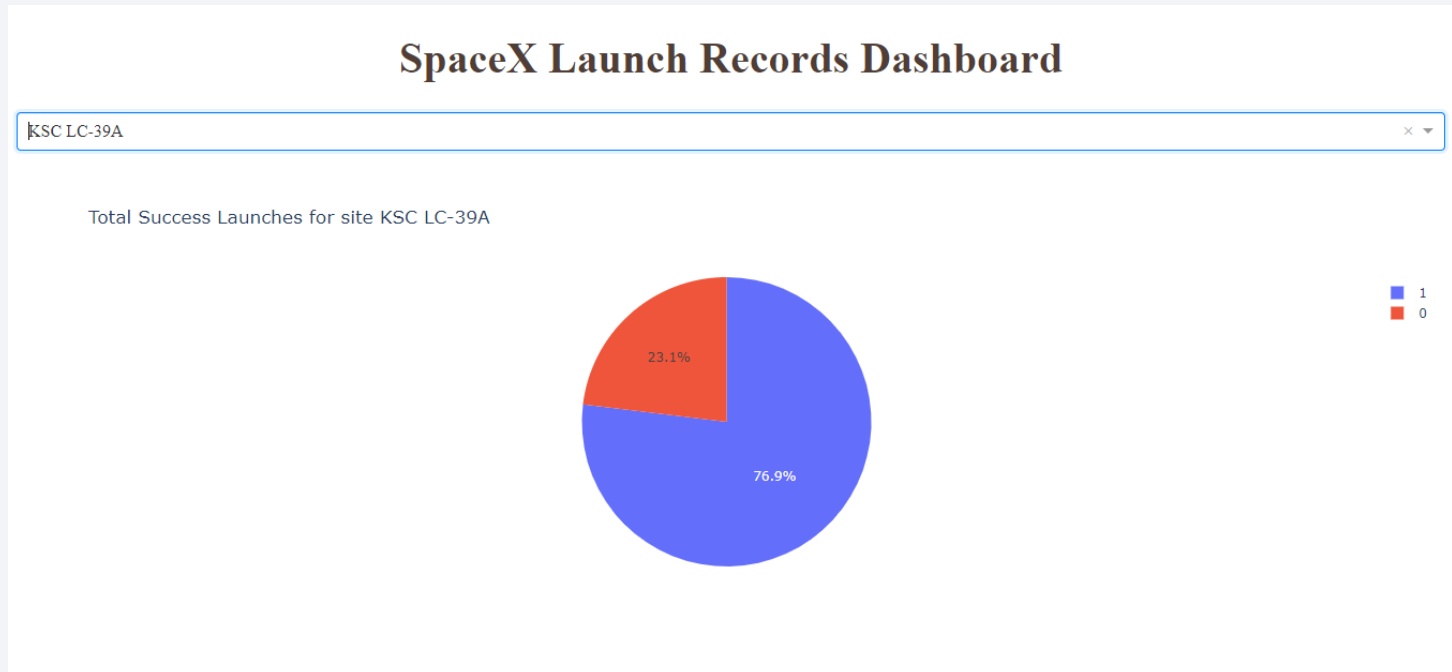
Build a Dashboard with Plotly Dash

Space X Launch Records Dashboard – Total Success Launches By Site



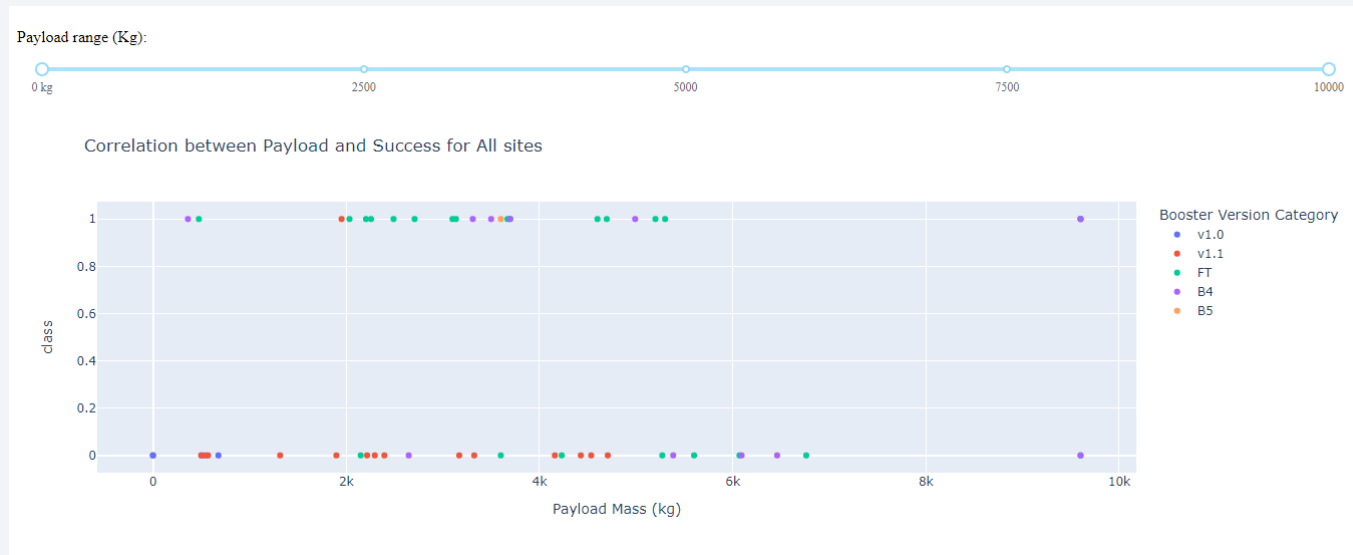
- This is the visualization for total success launches based on the launching site
- In this dashboard, we can select to show all sites or specific site by using the dropdown
- From the data, we can see that KSC LC-39A have the highest percentage of total success, compared to others
- CCAFS SLC-40 have the lowest percentage of total success

Space X Launch Records Dashboard – Pie Chart for Highest Launch Success Ratio



- KSC LC-39A got the highest success ratio compared to other launch site
- The success percentage is 76.9% and failure percentage is 22.1%

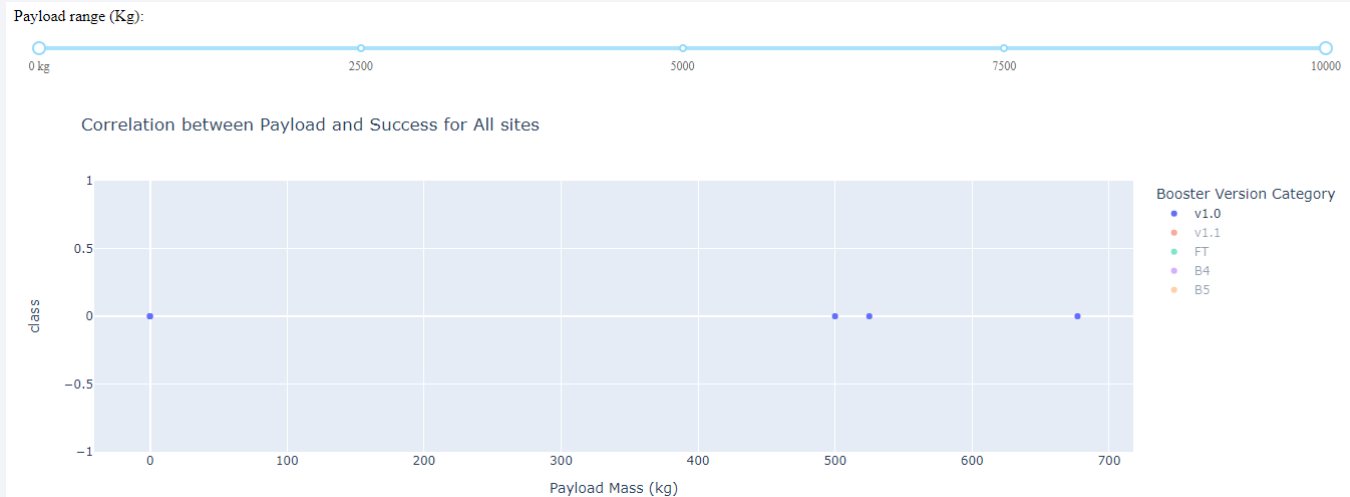
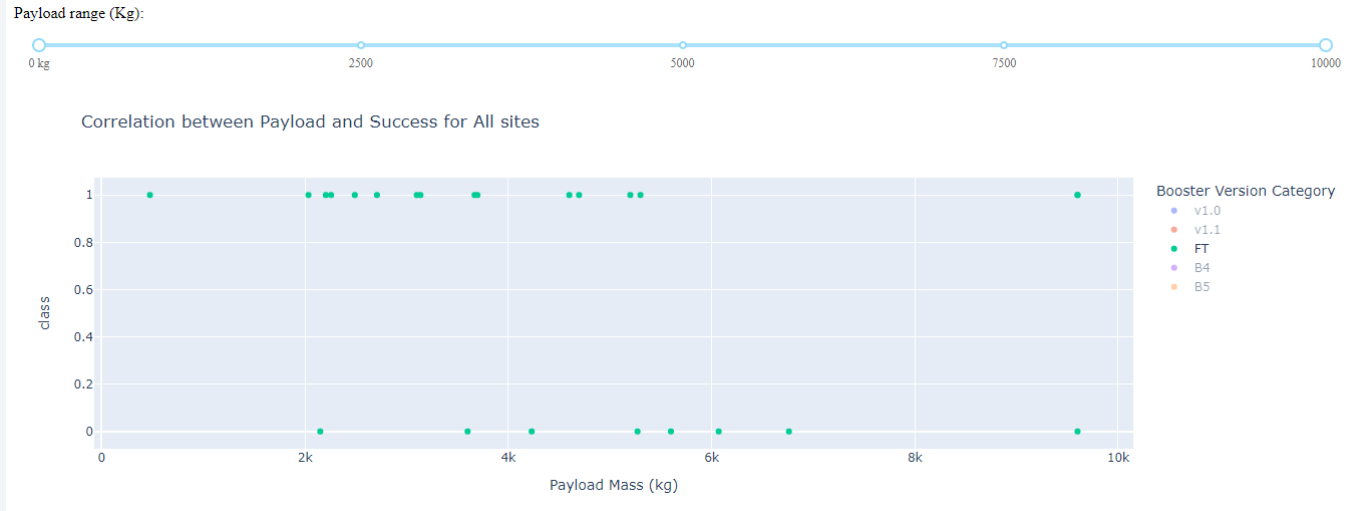
Space X Launch Records Dashboard – Payload vs Launch Outcome Scatter Plot



- This the scatter plot for finding correlation between payload and successes of launch for all class and have a color to distinct the booster version



Space X Launch Records Dashboard – Payload vs Launch Outcome Scatter Plot

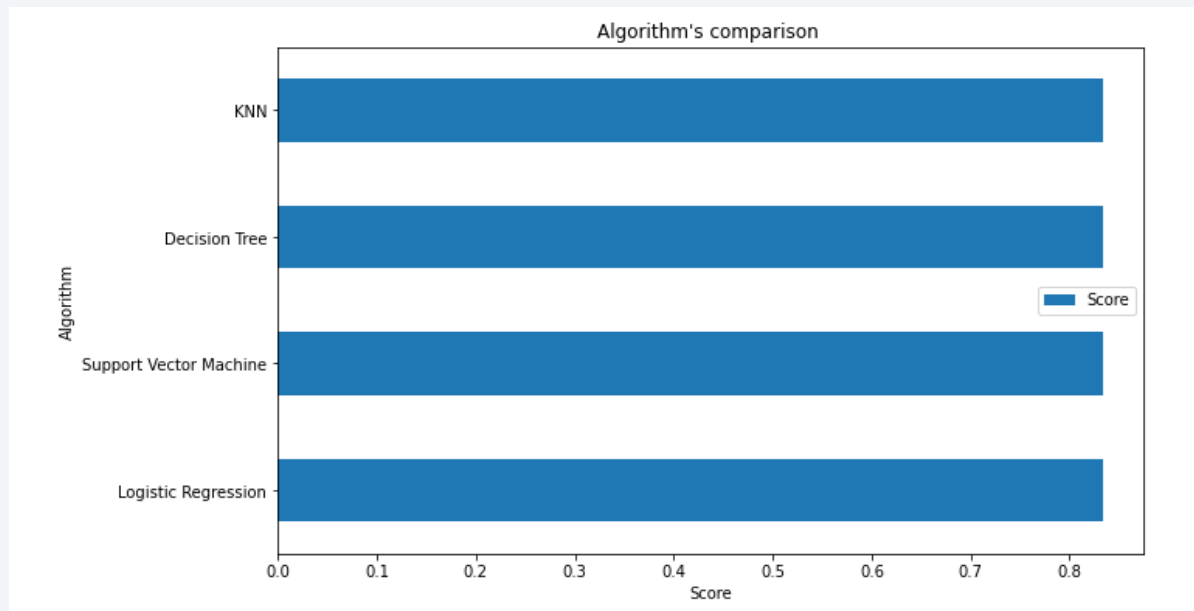


- As we can see, booster version category that have the most highest count and success rate is FT
- Booster version v1.0 never have a successful first stage landing according to the data
- From the success plot, we can't find the correlation between payload and output of launches
- Only booster version B4 have launched rocket with payload more than 8k kg

Section 5

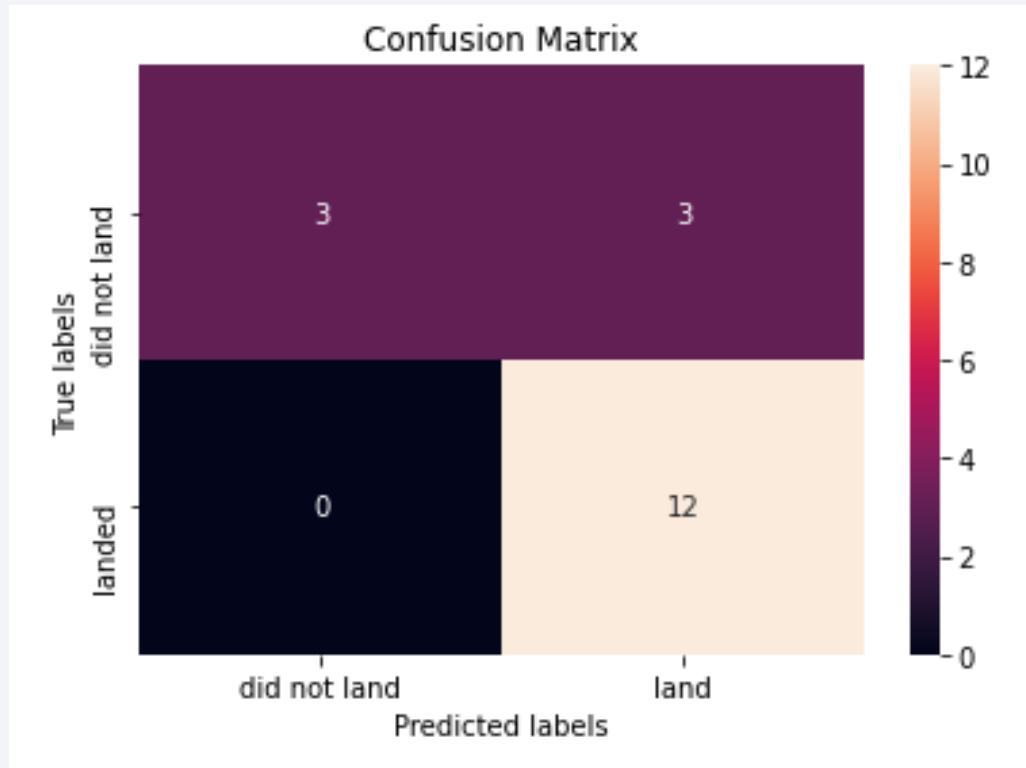
Predictive Analysis (Classification)

Classification Accuracy



- This is result of comparison between each classification algorithm for their result score
- In this attempt, all of the algorithm gain equal score

Confusion Matrix



- This is the confusion matrix for all the classification algorithm
- Errors come from false positive area, which means some prediction gives success for landing but what really happen was not

Conclusions

- We got five launch site to analysis: CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E
- Every year, the orbit for launch is becoming wider and success rate is increasing to about more than 80% success rate for first stage to land
- KSC LC-39A launch site have the most success rate for every launch
- We can see that most launch site in America placed near coastline and have nearby access for transportation
- From the test that we run for each classification algorithm, we found that we can predict the outcome of launch with 83.3% accuracy
- We can't find which one the best suited algorithm for this case, we need to take more test and train set and trying to change the hyper parameter to find the best solution

Appendix

Notebooks:

- https://jp-tok.dataplatform.cloud.ibm.com/analytics/notebooks/v2/5a671ad7-3728-45d8-9137-642aed78a1c2/view?access_token=d5db2da497353cb0bc1c5bf3228af43dc669c03d4bc2e0bc0d6c8d6f0533de0d
- https://jp-tok.dataplatform.cloud.ibm.com/analytics/notebooks/v2/8d20754a-cedb-43a7-ac85-ad9ed29d628c/view?access_token=d621611aa829870c6cf9afb43203fb8b012174c34531e7b07b9a1efdf4738b33
- https://jp-tok.dataplatform.cloud.ibm.com/analytics/notebooks/v2/daf6ac6b-8dbb-4865-8d78-6ab35cbb39fd/view?access_token=0d463faac509bf71c70c1c82a709d12881c06aeba942055263a1d54b4d08d596
- https://jp-tok.dataplatform.cloud.ibm.com/analytics/notebooks/v2/3e611002-f5af-46cb-ab68-862f66924461/view?access_token=343f8cb357e6bbc223f8a7fd7db32b9638acdae016a8c0b9d237bf667ba3c964
- https://jp-tok.dataplatform.cloud.ibm.com/analytics/notebooks/v2/51bf2c52-512f-40bf-9ef3-69c2861ecf11/view?access_token=d8b936d19f7e66a663e42039d8eb671c322ab5cf17a00d1ddb415aba5e27ca6e
- https://jp-tok.dataplatform.cloud.ibm.com/analytics/notebooks/v2/0900bf28-ce31-4629-8cd9-dc984098c7d3/view?access_token=d2bf6c65e4447c7318ed68aaf92cb0912b615b394d839610b67072186f125613
- https://jp-tok.dataplatform.cloud.ibm.com/analytics/notebooks/v2/5acc2aaa-a3e7-4a82-be6e-848482c14232/view?access_token=024e764711d147aedd9719dfd0c8d50701f294c5a4ccc48dca1c70b2fd7f2ada
- https://jp-tok.dataplatform.cloud.ibm.com/analytics/notebooks/v2/2ad70678-98fe-42a3-b39c-764e7cbc43ac/view?access_token=2061551cef65d6bc3f1087c24508e492093084a4bb206f9552386c40df32fcbc

Thank you!

