WINNING SPACE RACE WITH DATA SCIENCE

NGUYEN TRUNG TIN

10/21/2021

Outline

Executive Summary

Introduction

Methodology

Results

Conclusion

Appendix

EXECUTIVE SUMMARY

METHODOLOGIES

Data Collection And Wrangling

EDA with Visualization and SQL

Building Maps with Folium

Building a Dashboard with PlotlyDash

Predictive analytics

INTRODUCTION

ABOUT PROJECT

Background:

Now is the time of Comercial Space Age, but the cost for a rocket launch is huge. According Space X website, Falcon 9 rocket launch with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can recover first stage of launch rocket, the money will definitely decrease. And Space Y is a company wants to compete with Space X. We must help Space Y to do it.

Problem:

If we want to do it, we must know the outcome variables depend on what. How to successful for a landing.

METHODOLOGIES

Data Collection And Wrangling

- Using SpaceX Rest API
- Web Scrapping from Wikipedia

EDA with Visualization and SQL

- Data was clearned and transformed
- Use Line, Bar, Scatter graphs to show relation of data

Building Maps with Folium

- Using Folium to build interactive map

Building a Dashboard with PlotlyDash

- Using Plotly Dash to show interactive Dashboard

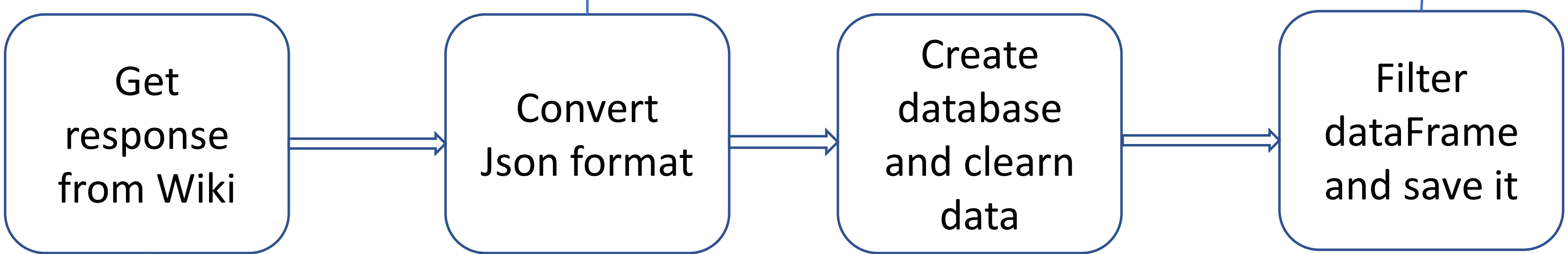
Predictive analytics

- Using Machine Learning to build model and predictive the successful rate.

DATA COLLECTION - SPACEX API

```
# Use json_normalize meethod to convert the json
response=requests.get(static_json_url)
data=pd.json_normalize( response.json())
```

```
# Hint data['BoosterVersion']!='Falcon 1'
data_falcon9= df[df['BoosterVersion']!='Falcon 1']
data_falcon9
```

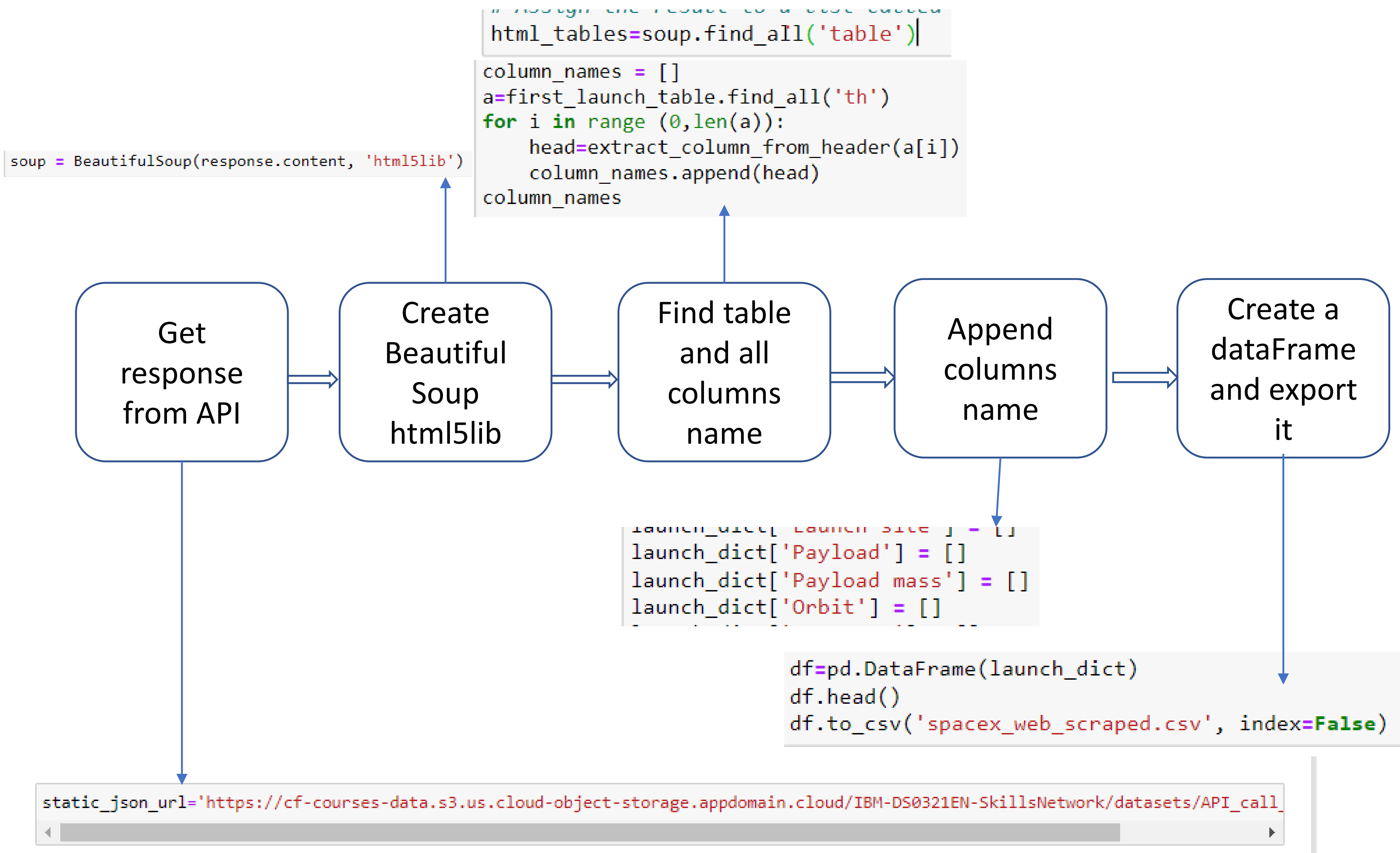



```
# Show the head of the dataframe
df=pd.DataFrame(launch_dict)
df.head()
```

	FlightNumber	Date	BoosterVersion	PayloadMass
0	1	2006-03-24	Falcon 1	20.0

```
static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"
response=requests.get(static_url)
```

DATA COLLECTION - SCRAPING



DATA WRANGLING

Calculate the num of launches on each site

```
df['LaunchSite'].value_counts()
```

```
CCAFS SLC 40    55
KSC LC 39A      22
VAFB SLC 4E     13
Name: LaunchSite, dtype: int64
```

Calculate the num and occurrence of each orbit

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

```
GTO     27
ISS     21
VLEO    14
PO       9
LEO      7
```

Calculate the number and occurence of mission outcome per orbit type

```
# landing_outcomes = values on Outcome column
landing_outcomes=df['Outcome'].value_counts()
landing_outcomes
```

```
True ASDS     41
None None     19
True RTLS     14
False ASDS     6
```

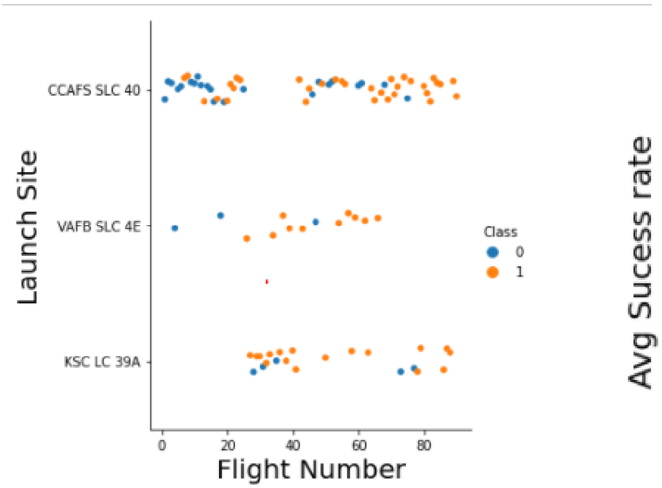
Create a landing outcome label from Outcome column

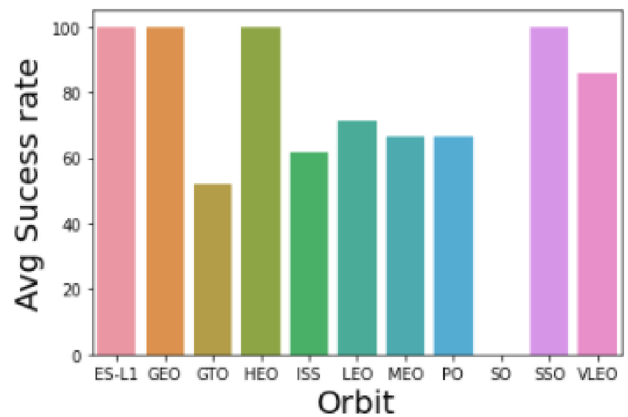
```
landing_class=[]
for i in range(0,len('Outcome')):
    p = df['Outcome'][i]
    for n in bad_outcomes:
        if p == n:
            landing_class.append(0)
        else:
            landing_class.append(1)
landing_class=pd.DataFrame(landing_class)
```

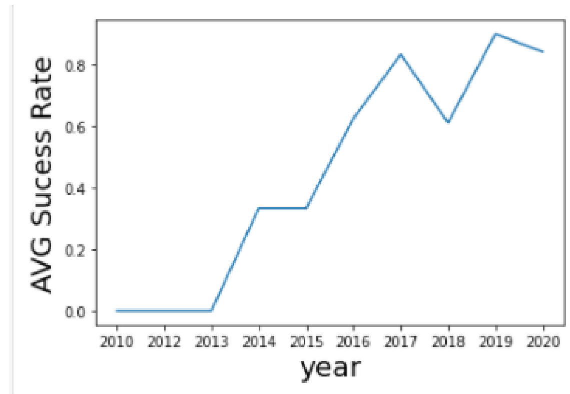
Create a dataFrame and export it

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

EDA With Visualization







Scatter Plot show correlation between:

- Flight Number vs. Payload Mass,
- Flight Number vs. Launch Site,
- Payload Mass vs. Launch Site,
- Flight Number vs Orbit Type,
- Payload Mass vs Orbit Type.

Scatter Plot show correlation between:

Success Rate vs Orbit type

Scatter Plot show correlation between:

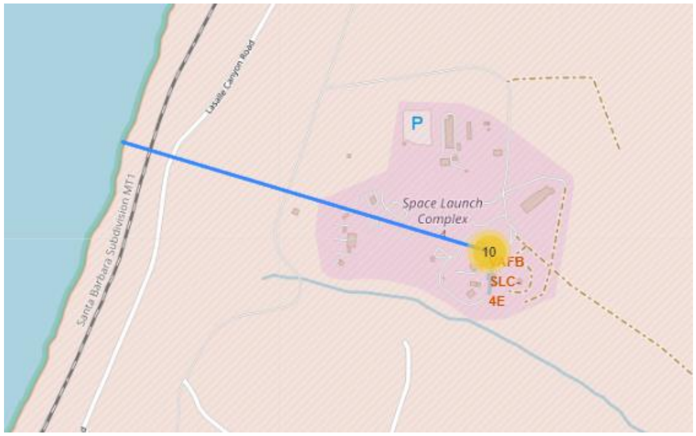
Success Rate Through Years

EDA With SQL

The information we get from SQL queries includes:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Ranking 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



Adding Circle Marker around each launch site with a label of the name of the launch site

Using Green marker to show successful launch and a red marker when a launch was failed.

By Using lines to indicate the distances between launch and its nearest such as Coastline .

Build a Dashboard with PlotlyDash

Dashboard includes a pie chart and a scatter plot:

- Pie Chart:

  Pie chart can be selected to show distribution of successful landings across all launch sites and can be selected to show individual launch site success rates.

  Scatter plot takes two inputs: All sites or individual site and payload mass on a slider between 0 and 10000 kg. The pie chart is used to visualize launch site success rate.

  The scatter plot can help us see how success varies across launch sites, payload mass, and booster version category.

PREDICTIVE ANALYSIS

Predictive analysis includes:

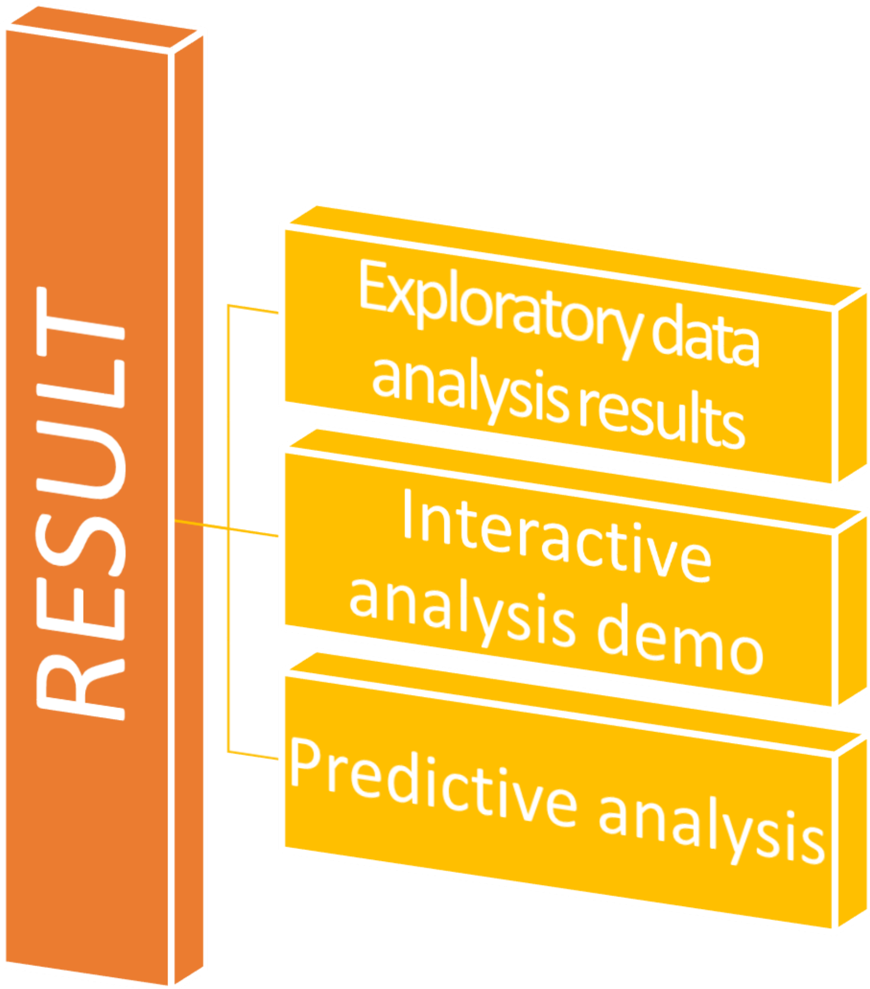
Building Model:

- Preprocessing and normalization data,
- Splitting data into train_test and size_test,
- Optimizing parameters by using Grid Search,
- Trainning models

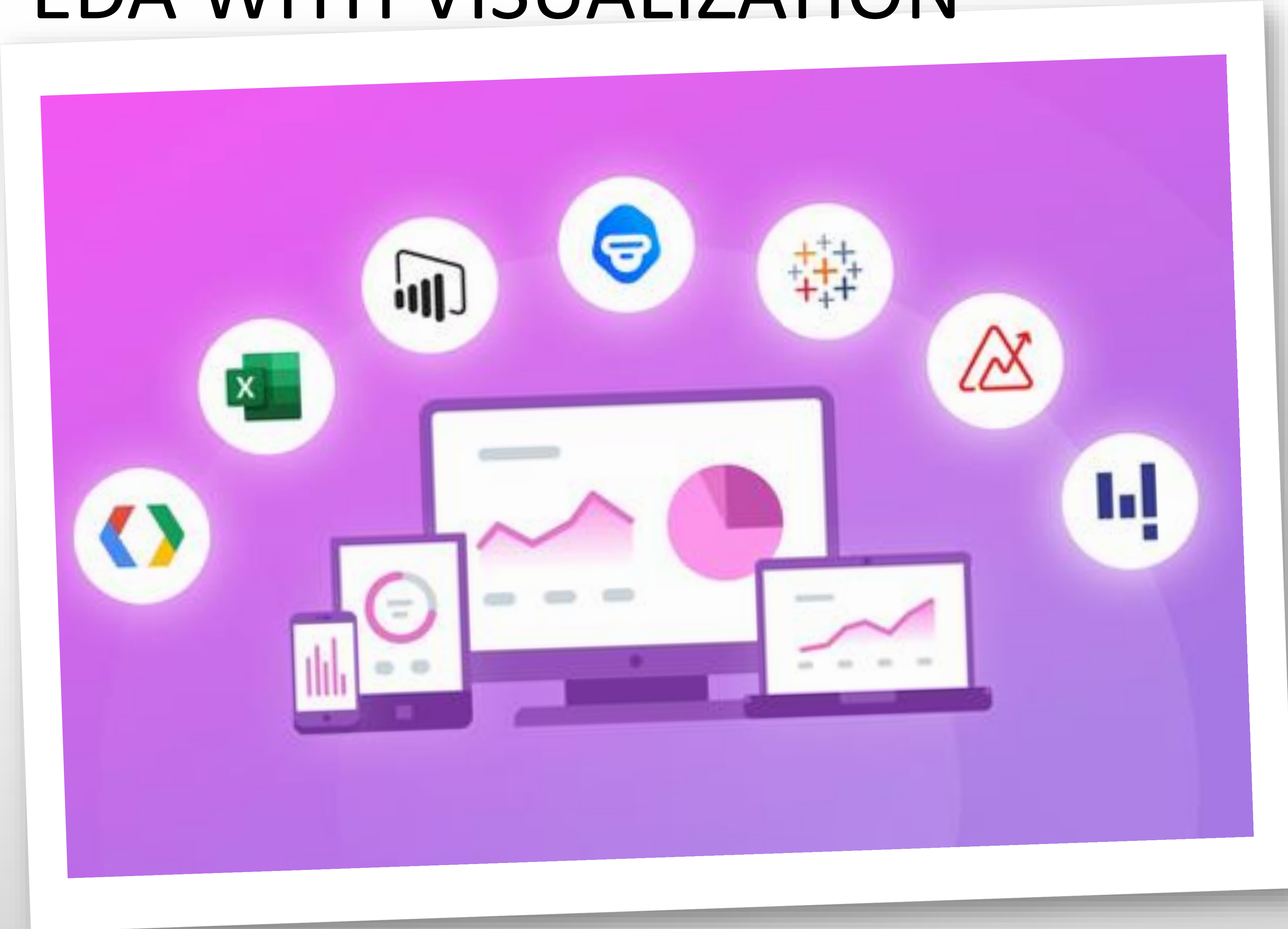
Evaluating Model:

- Checking accuracy for models,
- Plotting confusion matrix for models,
- Finding the best model with the highest accuracy

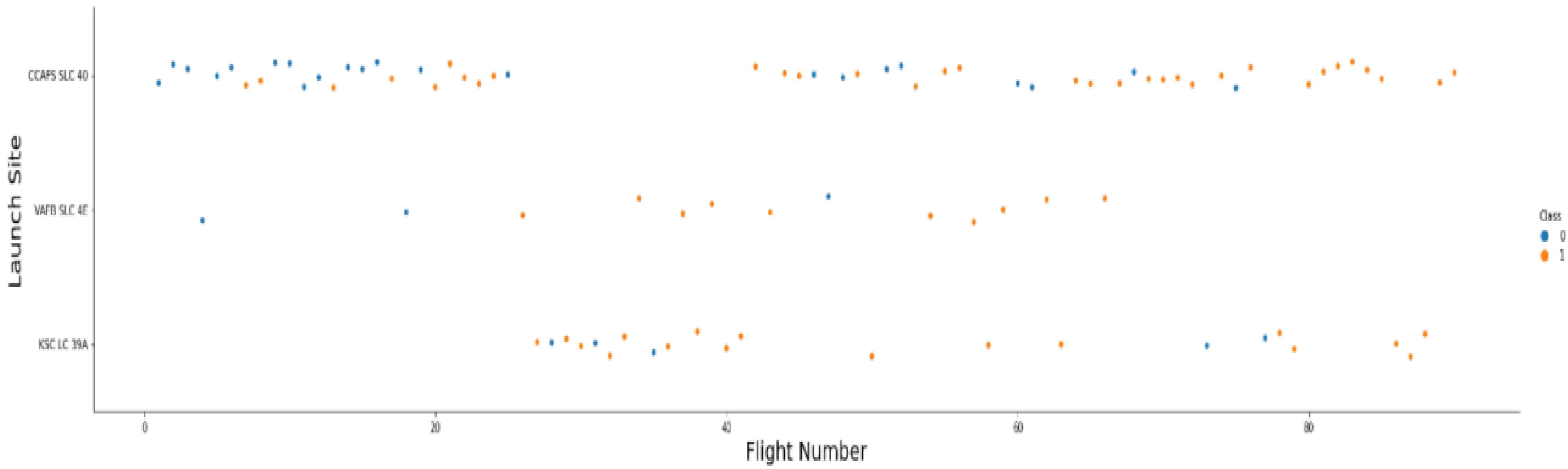
Result


RESULT
Exploratory data analysis results
Interactive analysis demo
Predictive analysis

EDA WITH VISUALIZATION

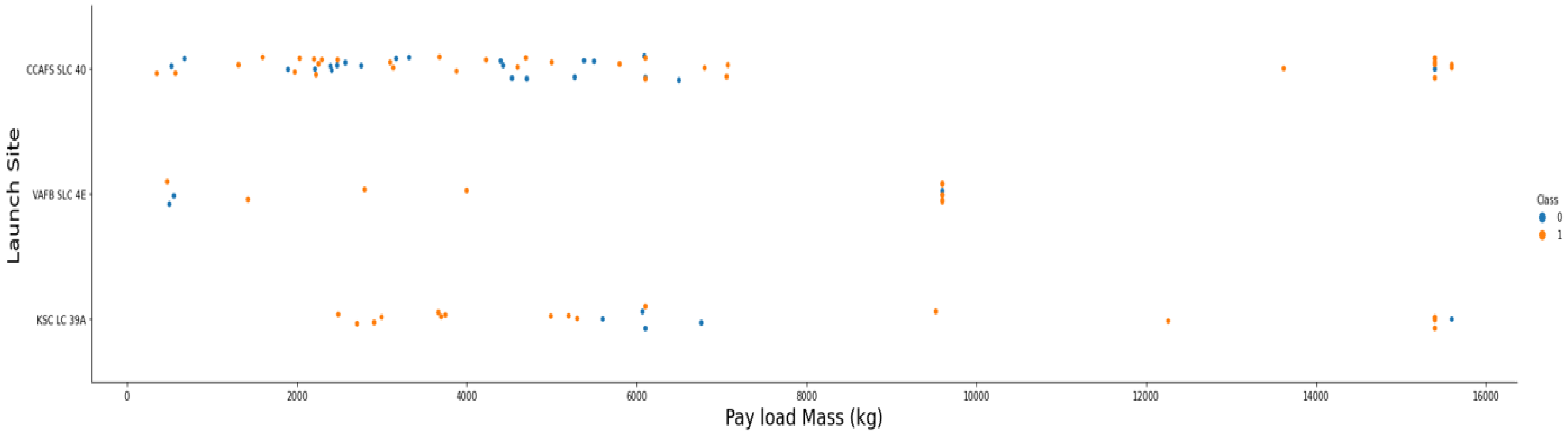


FLIGHT NUMBER VS. LAUNCH SITE



- The recent rocket launch have a higher success rate
- SLC 40 is the first and main launch site

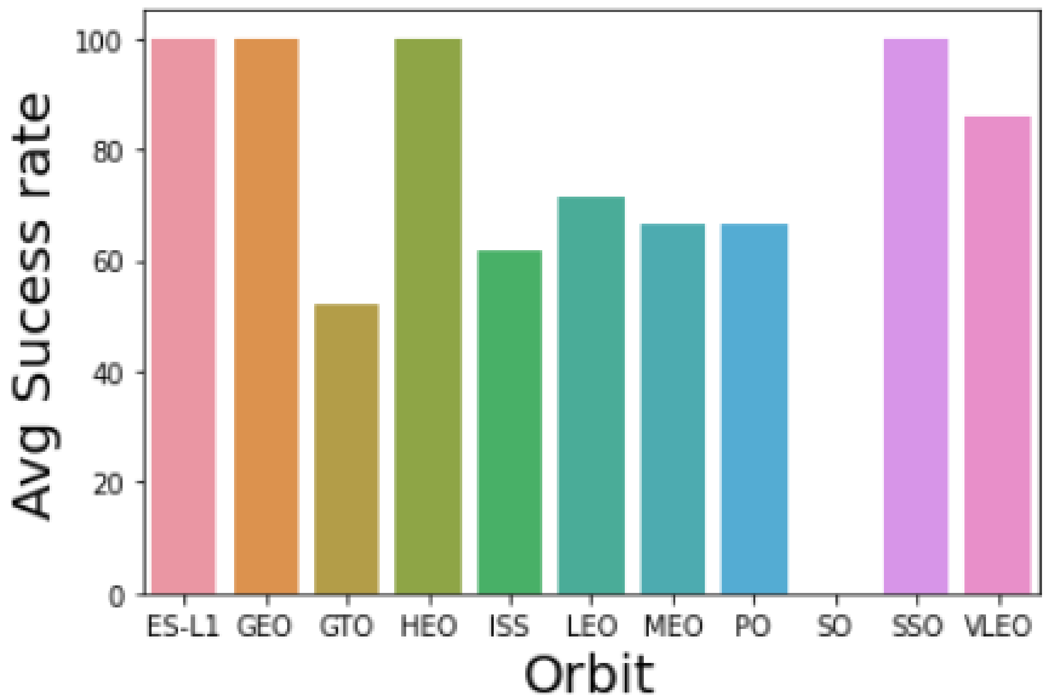
PAYLOAD VS. LAUNCH SITE



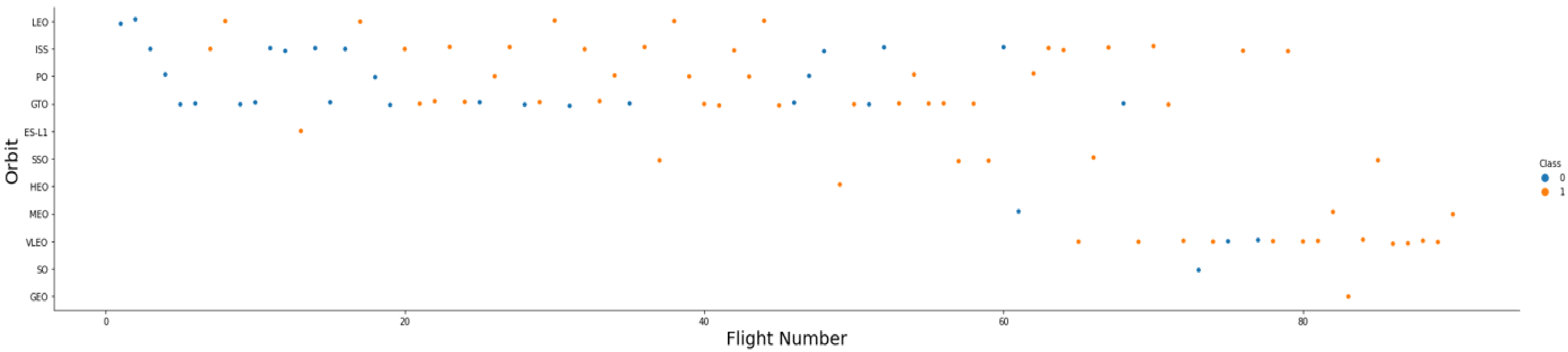
- The high Pay load Mass have a success rate higher.
- SLC 40 is the place has success rate of Small Pay load mass very low.

SUCCESS RATE VS. ORBIT TYPE

- EL-L1,GEO,HEO,SSO orbit all have successful
- It has never been successfully landed from SO orbit.
- The others orbit have above 50% success rate

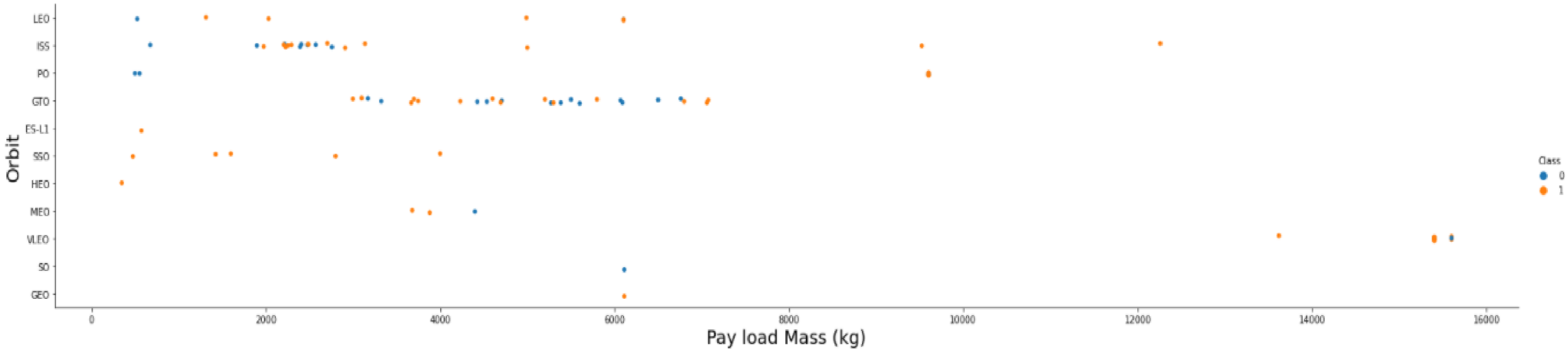


FLIGHT NUMBER VS. ORBIT TYPE



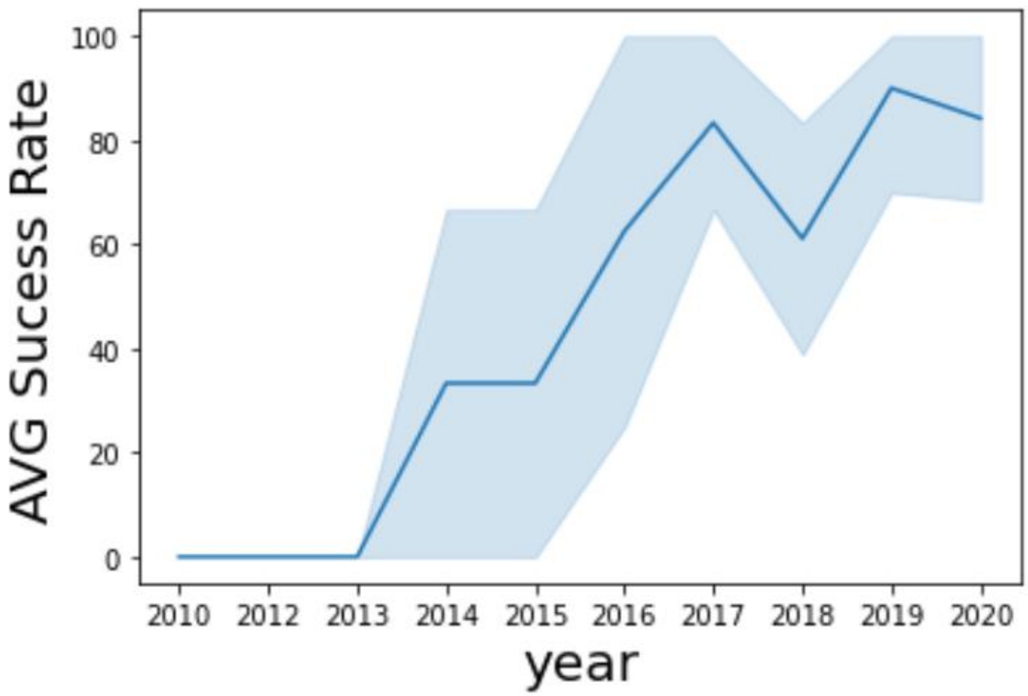
- The high Pay load Mass have a success rate higher.
- SLC 40 is the place has success rate of Small Pay load mass very low.
- The recent launch on VLEO have high success rate

PAYLOAD VS. ORBIT TYPE



- Heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here

PAYLOAD VS. ORBIT TYPE



- Success generally increases over time since 2013 with a slight dip in 2018 .
- Success in recent years at around 80%


EDA with
SQL
SELECT * FROM employees;
SELECT * FROM employees
WHERE emp_id = 2;
SELECT * FROM employees
ORDER BY name;

ALL LAUNCH SITE NAMES

```
%sql select distinct launch_site from spacextbl
```

```
 * ibm_db_sa://zvg47118:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

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

Explanation:

- Using select distinct statement to retrieve the unique launch site names.

LAUNCH SITE NAMES BEGIN WITH 'CCA'

```
%sql select * from spacextbl where launch_site like 'CCA%' limit 5
```

```
 * ibm_db_sa://zvg47118:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

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

Explanation:

- Using like to search a specific pattern in launch site column.

- Using limit clause to retrieve 5 records only

TOTAL PAYLOAD MASS

```
%%sql select sum(payload_mass__kg_) total_payload_mass
from spacextbl where customer = 'NASA (CRS)'
```

```
 * ibm_db_sa://zvg47118:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

total_payload_mass
45596

Explanation:

- Using sum() to calculate the total payload mass carried
- Using the condition customer = 'NASA (CRS)' to filter by customer

AVERAGE PAYLOAD MASS BY F9 V1.1

```
%%sql select avg(payload_mass__kg_) avg_payload_mass
from spacextbl where booster_version = 'F9 v1.1'
```

```
 * ibm_db_sa://zvg47118:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

avg_payload_mass
2928

Explanation:

- Using avg() to calculate average of payload mass carried

- Using the condition **booster_version ='F9 v1.1'** to filter by booster_version

FIRST SUCCESSFUL GROUND LANDING DATE

```
%%sql select min(date) min_date
from spacextbl where landing__outcome = 'Success (ground pad)'
```

```
 * ibm_db_sa://zvg47118:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

min_date
2015-12-22

Explanation:

- Using min() to return the smallest date
- Using the condition landing _outcome = 'Success (ground pad)' to filter by landing outcome.

SUCCESSFUL DRONE SHIP LANDING WITH PAYLOAD BETWEEN 4000 AND 6000

```
%%sql select booster_version from spacextbl
where landing__outcome = 'Success (drone ship)'
and payload_mass__kg_ > 4000 and payload_mass__kg_ < 6000
```

```
 * ibm_db_sa://zvg47118:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

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

Explanation:

- Using condition landing _outcome = 'Success (drone ship)' to filter by landing outcome.
- Using the condition payload_mass _kg_ > 4000 and payload_mass _kg_ < 6000 to filter by payload mass.

TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSION OUTCOMES

```
%%sql select
(case when mission_outcome like '%Success%' then 'Success' else 'Failure' end) mission_outcomes,
count(*) qty
from spacextbl group by (case when mission_outcome like '%Success%' then 'Success' else 'Failure' end)
```

```
 * ibm_db_sa://zvg47118:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

mission_outcomes	qty
Failure	1
Success	100

Explanation:

- Using count(*) to retrieve the number of records.
- Using group by to group by mission outcomes.

BOOSTERS CARRIED MAXIMUM PAYLOAD

```
%%sql select booster_version from spacextbl
where payload_mass__kg_ = (select max(payload_mass__kg_) from spacextbl)
```

```
 * ibm_db_sa://zvg47118:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

Explanation:

- Using the subquery to get maximum payload mass.
- Using max() to return the largest payload mass.

2015 FAILED DRONE SHIP LANDING RECORDS

```
%%sql select booster_version,launch_site from spacextbl
where landing__outcome = 'Failure (drone ship)' and year(date) = 2015
```

```
 * ibm_db_sa://zvg47118:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde
Done.
```

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

Explanation:

- Using year() to return the year part of a date.

RANKING COUNTS OF SUCCESSFUL LANDINGS BETWEEN 2010-06-04 AND 2017-03-20

```
%%sql select landing__outcome,count(landing__outcome) qty from spacextbl
where (date between '2010-06-04' and '2017-03-20')
group by landing__outcome order by 2 desc
```

```
 * ibm_db_sa://zvg47118:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lq
Done.
```

landing__outcome	qty
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

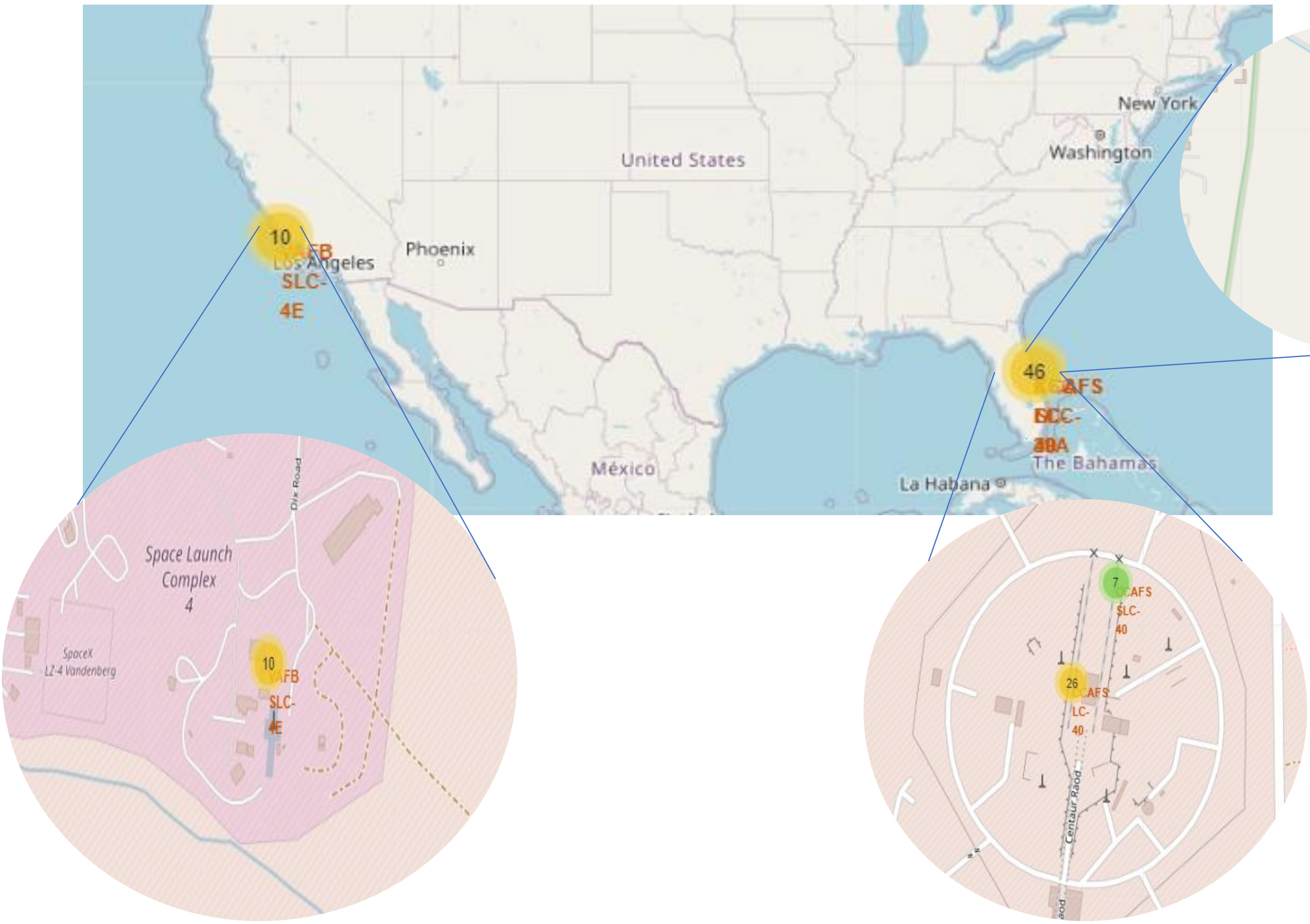
Explanation:

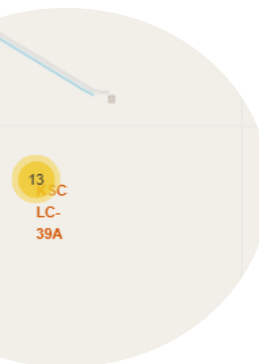
- Using group by to group by landing outcomes.
- Using order by 2 desc keyword to sort by 2nd column in descending order

Interactive Map with

Folium

SPACEX LAUNCH SITES

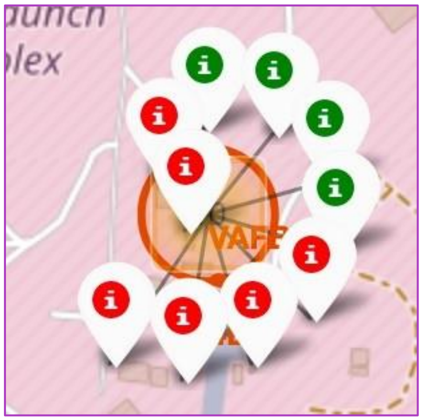




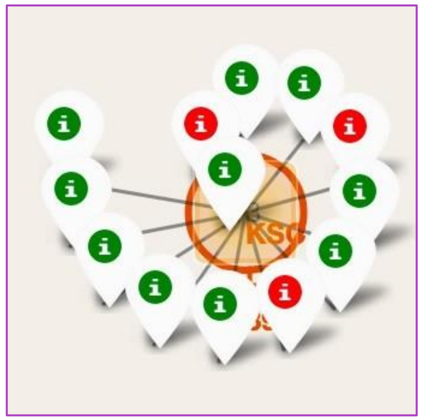
We can see all **4 launch site** are in very close to the ocean in order to minimize the risks in the event of an incident.

SPACEX LAUNCH SITE SUCCESS RATE

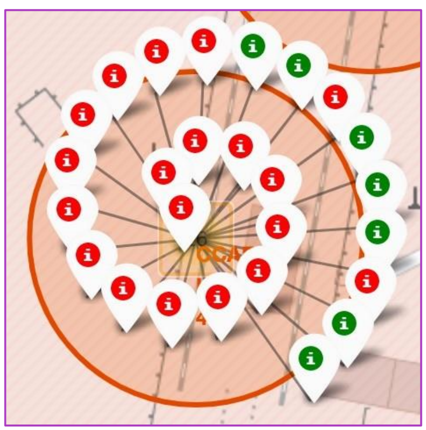
VAFB SLC-4E



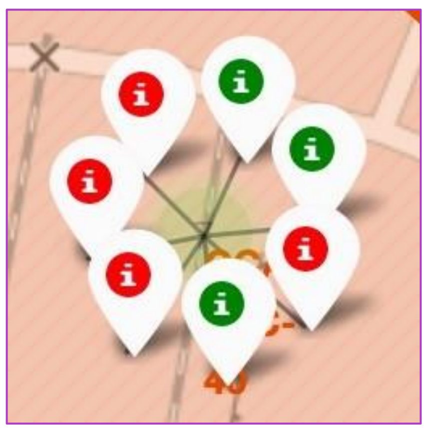
KSC LC-39A



CCAFS LC-40

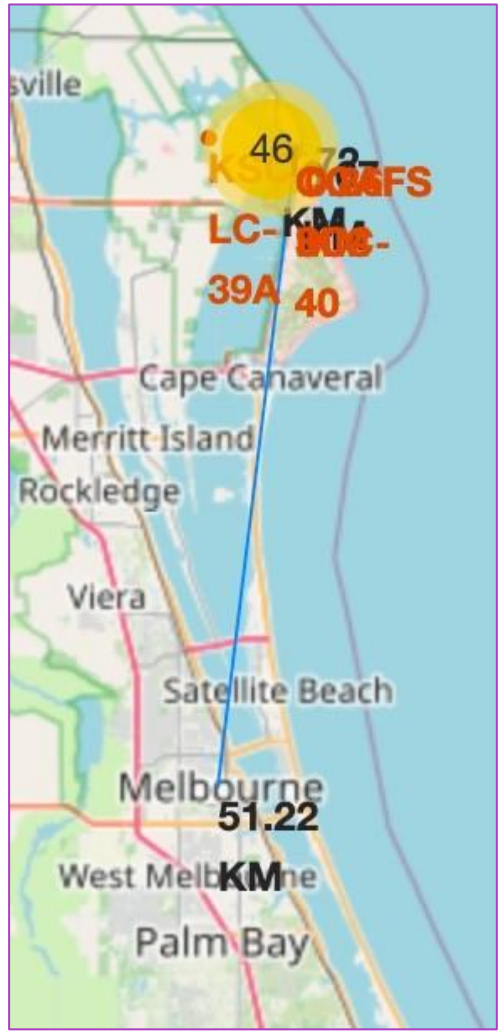


CCAFS SLC-40



- Green marker: successful landings
- Red marker: failed landings
- As screenshots shown, KSC LC-39A has the highest success rate and CCAFS LC-40 has the opposite result.

SPACEX LAUNCH SITE SUCCESS RATE

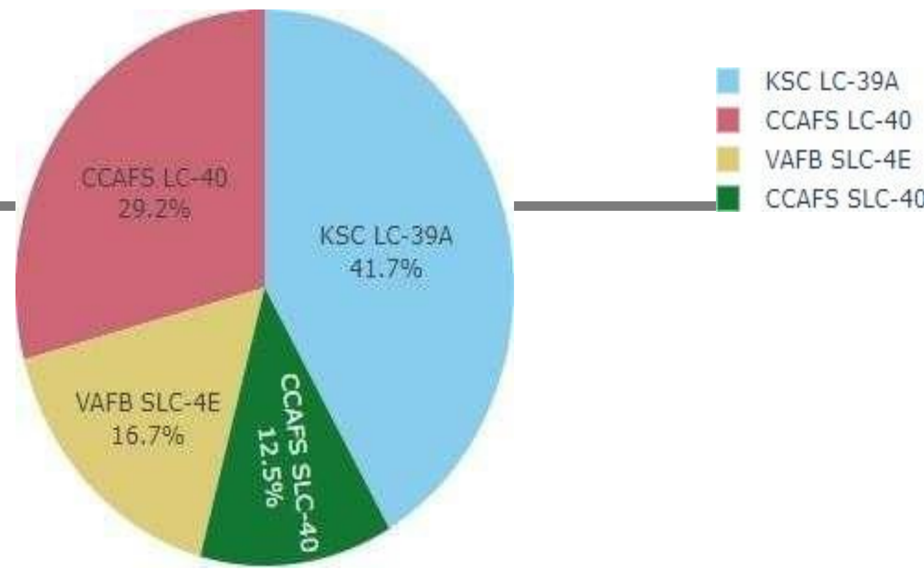


- Using KSC LC-39A as an example, launch sites are very close to railways for large part and supply transportation.
- Launch sites are close to highways for human and supply transport. Launch sites are also close to coasts and relatively far from cities so that launch failures can land in the sea in order to minimize the risks in the event of an incident.

Build a Dashboard with

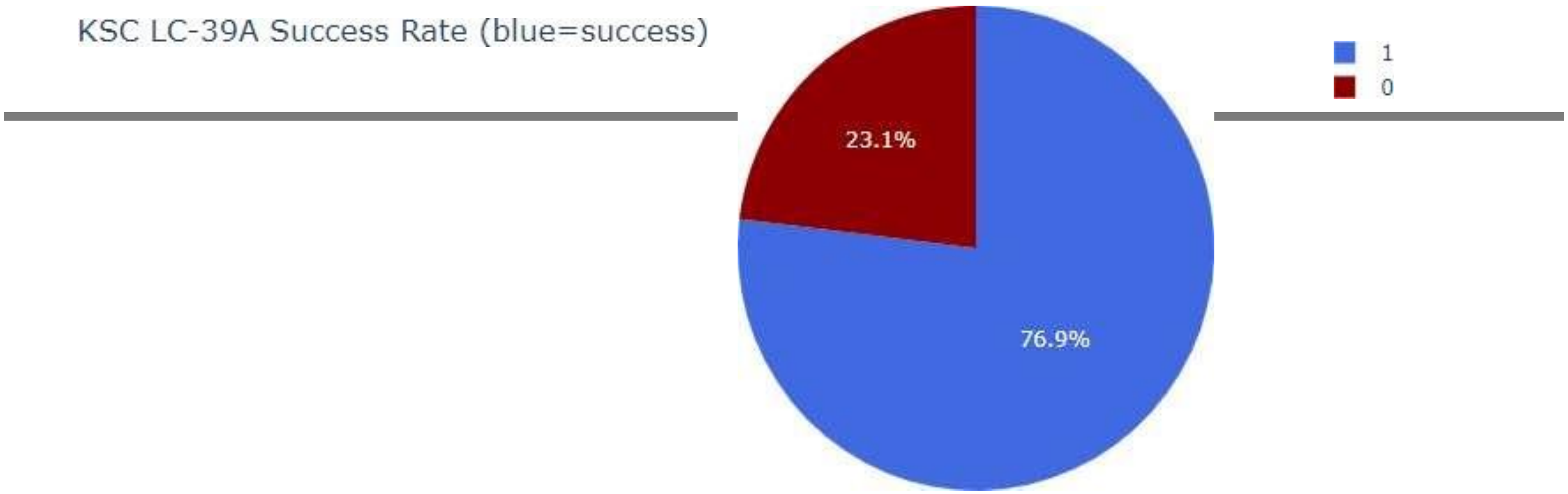
plotly | Dash

DASHBOARD
LAUNCH SUCCESS COUNT FOR ALL SITES



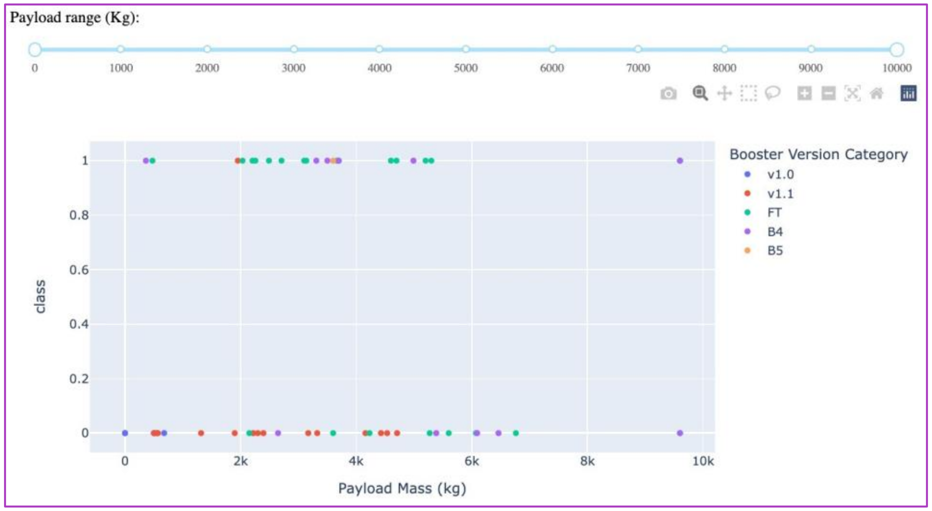
- The pie Chart show total success launch across all site.
- KSC LC-39A has the most success launch. Beside,the total of CCAFS SLC-40 and CCAFS LC- 40 has the same success with KSC LC-39A

DASHBOARD
LAUNCH SUCCESS COUNT FOR ALL SITES

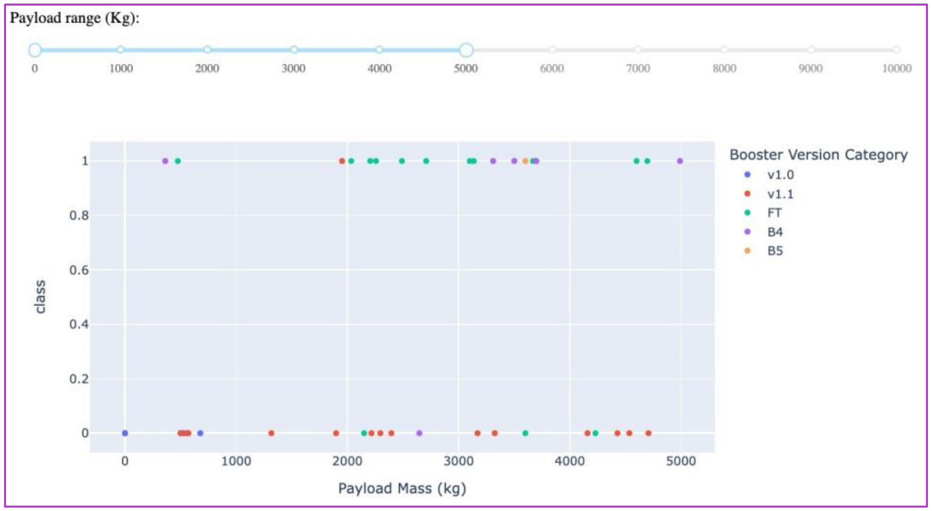


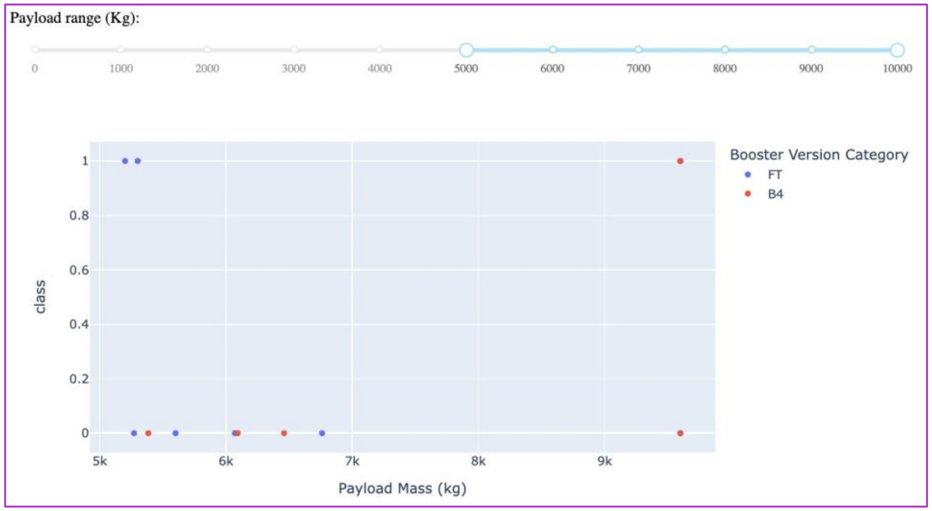
- KSC LC-39A has 76.9% success launch. That is so astonished.

DASHBOARD
LAUNCH SUCCESS COUNT FOR ALL SITES



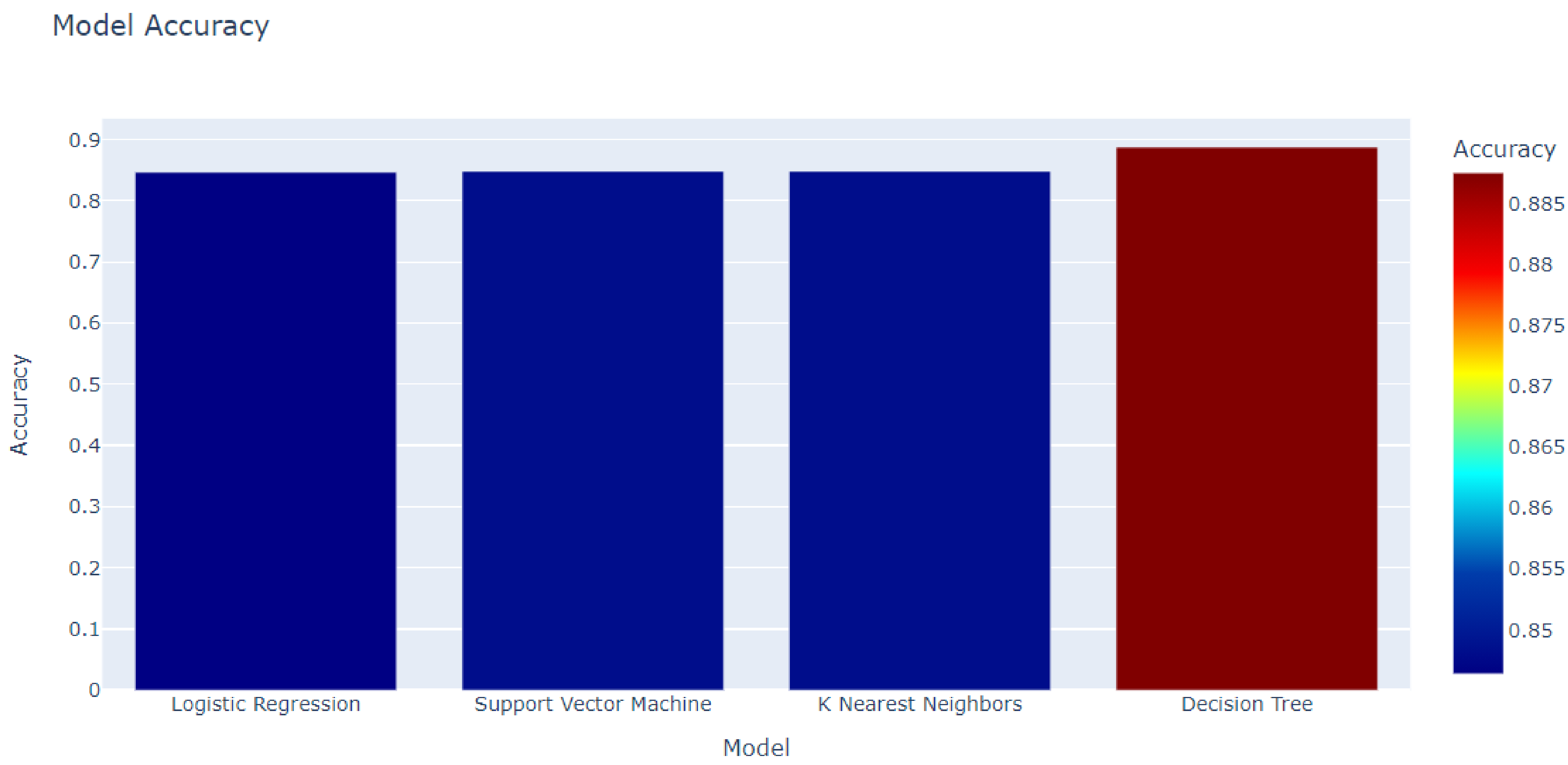
- The scatter plot shows the correlation between launch outcomes and payload mass by booster version.
- this is set from 0-10000 instead of the max Payload of 15600.
- Class indicates 1 for successful landing and 0 for failure.





PREDICTIVE ANALYSIS

Predictive Analysis



All models have a same best score but Decision Tree has a little better perform at 0.8875. All models have the same accuracy rate on the test are 0.8333.

CONCLUSIONS

- Our task: to develop a machine learning model for Space Y who wants to bid against SpaceX
- The goal of model is to predict when Stage 1 will successfully land to save much money as can.
- Using data from a public SpaceX API and web scraping SpaceX Wikipediapage
- Creating data labels and stored data into a DB2 SQL database
- Creating a dashboard for visualization
- Creating Machine Learning models to predictive analysis
- Decision Tree has a bit better performance that is a good choice for Space Y to use