IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Yulong WANG>
<2021-09-01>

Outline

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

Executive Summary

- Summary of methodologies

  In order to predict if the first stage of the SpaceX Falcon 9 rocket will land successfully, the methods of data collection, data wrangling, exploratory data analysis, data visualization, model development, model evaluation are applied.

- Summary of all results

  Launch site has a strong relation with success rate.

  CCAFS SLC-40 and KSC LC-39A are most feasible sites with highest success rate and amount of launch.

  The ocean & RTLS outcome have succussed on small and medium mass launches, and there is more chances for improvement.

Introduction

- Project background and context

  SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.

- Problems you want to find answers
  - Sources of data
  - Data managing
  - Data processing
  - Data analysis
  - Data understanding

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - API and web scraping mechanics are applied to collect the data.
- Perform data wrangling
  - Preprocess includes filling NaN, discard outlier, generating secondary data, ect.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Several classification models are built after exploratory, for example data standardize, training and test data split, parameters are tuned by using the grid searching method, finally, the best performance model is decided by score evaluation.

Data Collection

- We can say that there is no data there is no scientific data. Therefore, the fastest way to collect big data is through the Internet.
- API is the acronym for Application Programming Interface, which is a software intermediary that allows two applications to talk to each other. Web scraping is the process of using bots to extract content and data from a website.

Data Collection – SpaceX API

- A get request to the SpaceX Application Programming Interface (API), also the JSON to extract data about boosters.

- IBM Watson Studio JupyterNotebook Link

- Github JupyterNotebook Link

Data Collection - Scraping

- A get request to the SpaceX Application Programming Interface (API), also the JSON to extract data about boosters.

- IBM Watson Studio JupyterNotebook Link

- Github JupyterNotebook Link

Data Wrangling

- Transforming and mapping data from raw data, which including transforming data to dataframe, discarding the null values, discarding the outliers, normalizing the values, and some basic data exploring.

- IBM Watson Studio JupyterNotebook Link
- Github JupyterNotebook Link

- IBM Watson Studio JupyterNotebook Link
- Github JupyterNotebook Link

EDA with Data Visualization

- Scatter plot, Bar plot, and Line plot are used to visualize the data.

- IBM Watson Studio JupyterNotebook Link

- Github JupyterNotebook Link

- Scatter plot
  - Variable relationship
  - Easy to handle
- Bar plot
  - Categorical relationship
  - Impact visualization
- Line plot
  - Timeseries tendency

EDA with SQL

- IBM Watson Studio JupyterNotebook Link
- Github JupyterNotebook Link

- %load_ext sql
  - Load the sql and build a connection
- %%sql
  - Magic command for sql in python
- Where: extract or filter conditions
- Like: character string match detection
- Group by: summary rows
- Between: select values within a range
- Order: sort the result in order

Build an Interactive Map with Folium

- IBM Watson Studio JupyterNotebook Link
- Github JupyterNotebook Link

- Summary of map objects:
  - Markers: show a geo location from lat and lon data
  - Cluster: show a group of markers
  - Circles: show a single location
  - Lines: show distance between two markers

Build a Dashboard with Plotly Dash

- IBM Watson Studio JupyterNotebook Link
- Github JupyterNotebook Link

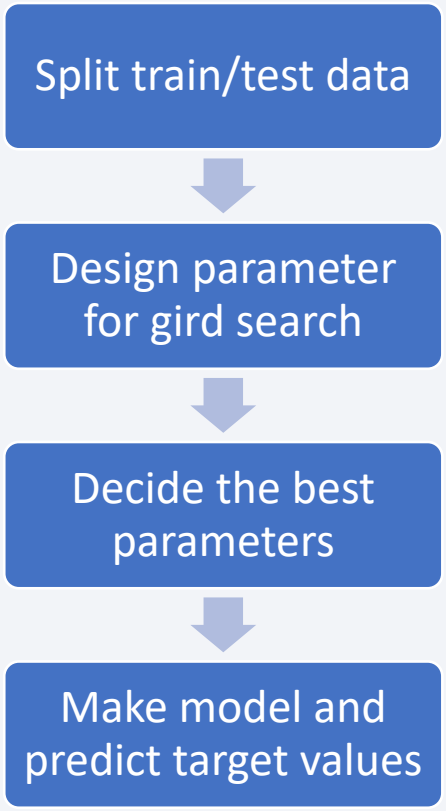
- Summary of plots:
  - Bar: shows categories differences
  - Line: reports timeseries changes
  - Pie: shows the percentage of events
  - Tree: shows complex relationship of variables in interactive way
  - Map: shows variables of states on a map

Predictive Analysis (Classification)

- The models are trained by the spited data sets, the optimal parameters are decided by applying grid searching method, and evaluated by scores.

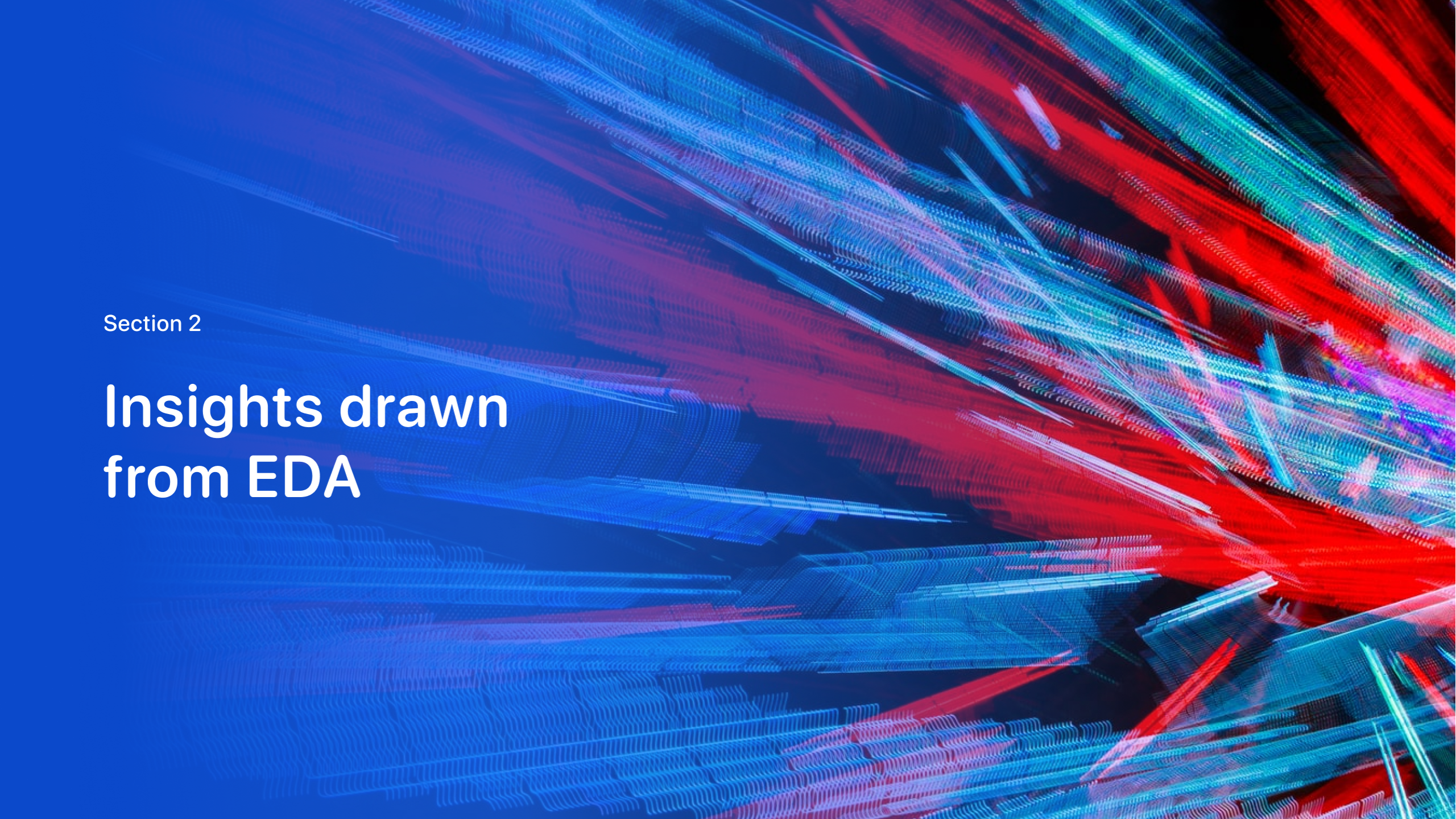
- IBM Watson Studio JupyterNotebook Link

- Github JupyterNotebook Link



Results

- Exploratory data analysis results
  - Both API and web crawling are capable to collect Xspace data
- Interactive analytics demo in screenshots
  - EDA with sql is effective for data filtering
  - EDA with interactive visualization provides informative information
  - Plotly Dash is powerful to show instant data change
- Predictive analysis results
  - The logistical regression model has the best accuracy of predicting.

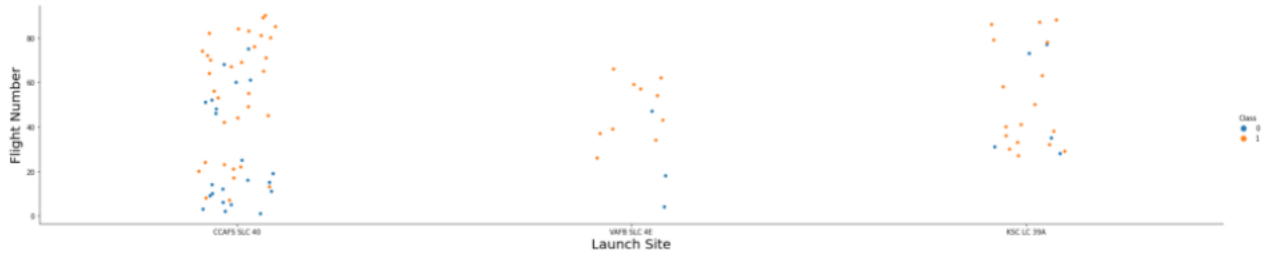


Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- Show a scatter plot of Flight Number vs. Launch Site
- Show the screenshot of the scatter plot with explanations



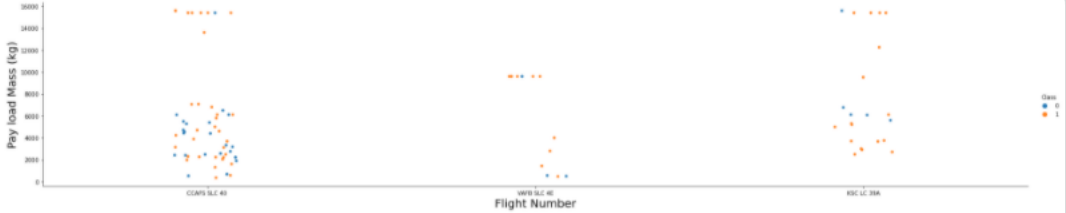
Now try to explain the patterns you found in the Flight Number vs. Launch Site scatter point plots.

- CCAFS SLC 40 seems to be the oldest launch platform and also lasted for a long time to support the new launches. It is the main launch site, and the failure rate decrease gradually.
- Launches at WAFB SLC 40 failed at the beginning.
- KSC has a low failure rate.

Payload vs. Launch Site

- Show a scatter plot of Payload vs. Launch Site

- Show the screenshot of the scatter plot with explanations



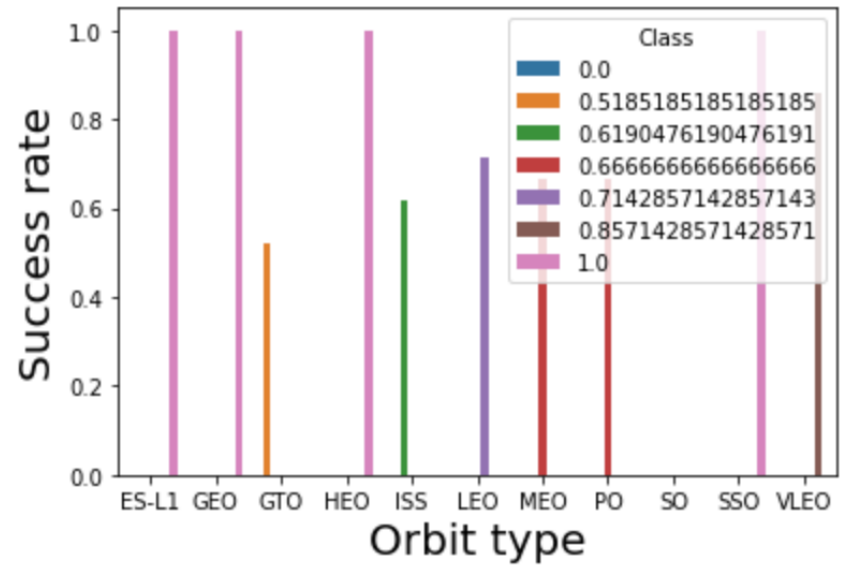
Now try to explain any patterns you found in the Payload Vs. Launch Site scatter point chart.

- CCAFS SLC 40 seems to be a predominant launch site, most small-medium mass level load mass launches are located here.
- WAFB SLC 40 seem to ba small mass launch platform, but quite important as it provides medium mass launches, the failure rate here is relatively low.

Success Rate vs. Orbit Type

- Show a bar chart for the success rate of each orbit type

- Show the screenshot of the scatter plot with explanations



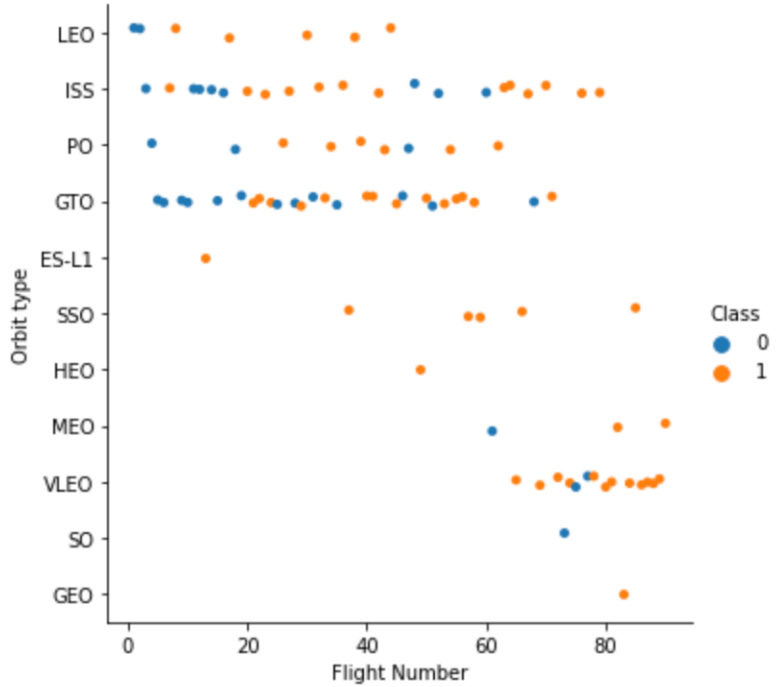
Analyze the ploted bar chart try to find which orbits have high sucess rate.

- Low success rate: GTO, ISS, MEO, PO and LEO
- High success rate: ES-L1, GEO, HEO, SSO

Flight Number vs. Orbit Type

- Show a scatter point of Flight number vs. Orbit type

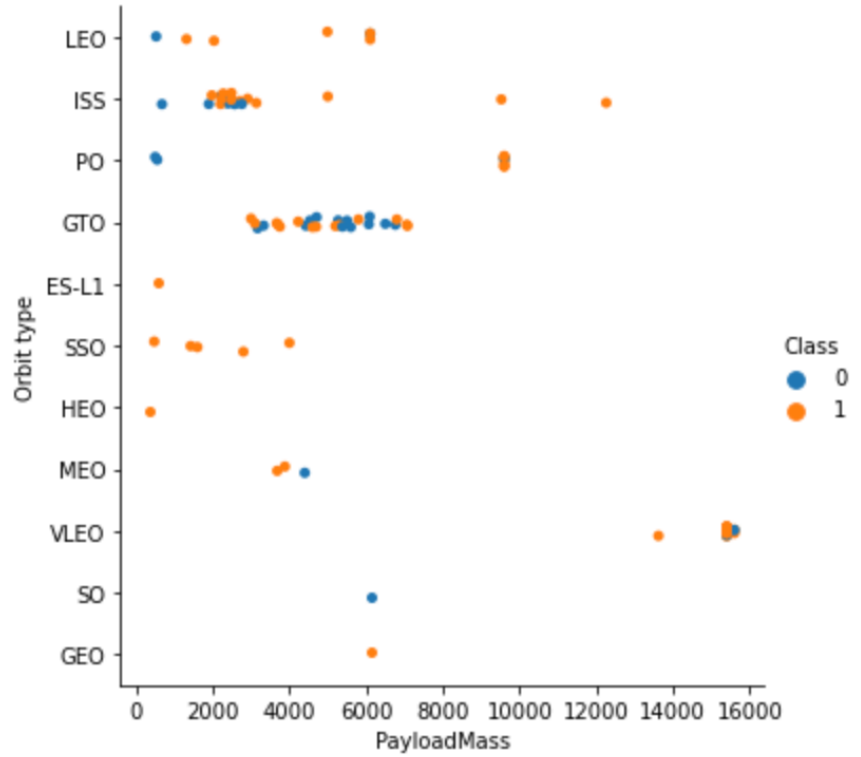
- Show the screenshot of the scatter plot with explanations



You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type

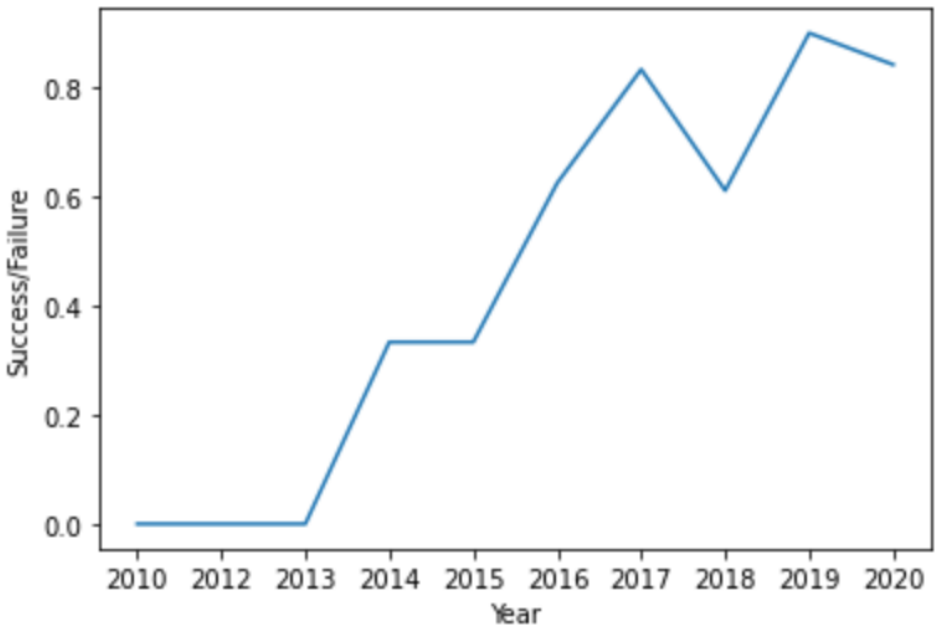
- Show a scatter point of payload vs. orbit type
- Show the screenshot of the scatter plot with explanations



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

Launch Success Yearly Trend

- Show a line chart of yearly average success rate
- Show the screenshot of the scatter plot with explanations



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

All Launch Site Names

- Find the names of the unique launch sites
- Present your query result with a short explanation here

  - There are four launch sites, CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, and VAFB SLC-4E.

```
%%sql
select distinct(LAUNCH_SITE)
from SPACEXTBL
```

```
 * ibm_db_sa://zgr48431:***@0c77d6f2
-5da9-48a9-81f8-86b520b87518.bs2io90
l08kqb1od8lcg.databases.appdomain.cl
oud:31198/bludb
Done.
```

Out[17]:

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Present your query result with a short explanation here
  - They are all CCAFS LC-40.

In [23]:
```
%%sql
select *
from SPACEXTBL
where LAUNCH_SITE like 'CCA%'
limit 5;
```

```
 * ibm_db_sa://zgr48431:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[23]:

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

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- Present your query result with a short explanation here

  - Well, um… no idea about it…

In [8]:

```
%%sql
select sum(PAYLOAD_MASS__KG_)
from SPACEXTBL
where CUSTOMER = 'NASA (CRS)'
;
```

```
 * ibm_db_sa://zgr48431:***@0c77d6f2-5
da9-48a9-81f8-86b520b87518.bs2io90l08k
qb1od8lcg.databases.appdomain.cloud:31
198/bludb
Done.
```

Out[8]:

1
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Present your query result with a short explanation here

  - Average of 3000 kg, quite small mass.

```
%%sql
select avg(PAYLOAD_MASS__KG_)
from SPACEXTBL
where BOOSTER_VERSION = 'F9 v1.1'
;
```

```
 * ibm_db_sa://zgr48431:***@0c77d6f2-5
da9-48a9-81f8-86b520b87518.bs2io90l08k
qb1od8lcg.databases.appdomain.cloud:31
198/bludb
Done.
```

Out[28]:

1
2928

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Present your query result with a short explanation here

  - First successful landing after 5 years attempts, LOL!!

```
In [37]:
```

```
%%sql
select min(DATE),
from SPACEXTBL
where landing__outcome like 'Success (g%';
```

```
 * ibm_db_sa://zgr48431:***@0c77d6f2-5
da9-48a9-81f8-86b520b87518.bs2io90l08k
qb1od8lcg.databases.appdomain.cloud:31
198/bludb
Done.
```

```
Out[37]:
```

1
2015-12-22

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

- Present your query result with a short explanation here

    - The version code seem to be continued

In [39]:

```
%%sql
select booster_version
from SPACEXTBL
where landing__outcome like 'Success (d%'
and payload_mass__kg_ between 4000 and 6000;
```

```
 * ibm_db_sa://zgr48431:***@0c77d6f2-5
da9-48a9-81f8-86b520b87518.bs2io90l08k
qb1od8lcg.databases.appdomain.cloud:31
198/bludb
Done.
```

Out[39]:

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Present your query result with a short explanation here
  - Mission outcomes are preforming well

In [49]:

```
%%sql
select mission_outcome, count(mission_outcome)
from SPACEXTBL
group by mission_outcome;
```

```
 * ibm_db_sa://zgr48431:***@0c77d6f2-5
da9-48a9-81f8-86b520b87518.bs2io90l08k
qb1od8lcg.databases.appdomain.cloud:31
198/bludb
Done.
```

Out[49]:

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Present your query result with a short explanation here

  - The version code seem to be continued

In [56]:

```
%%sql
select booster_version
from SPACEXTBL
where payload_mass__kg_=(select max(paylo
```

```
 * ibm_db_sa://zgr48431:***@0c77d6f2-5
da9-48a9-81f8-86b520b87518.bs2io90108k
qb1od8lcg.databases.appdomain.cloud:31
198/bludb
Done.
```

Out[56]:

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

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

- Present your query result with a short explanation here

    - Maybe drone ship is problem before improvement

```
In [60]:
```

```
%%sql
SELECT BOOSTER_VERSION,LAUNCH_SITE,landir
FROM SPACEXTBL
where EXTRACT(YEAR FROM DATE)='2015' and
```

```
 * ibm_db_sa://zgr48431:***@0c77d6f2-5
da9-48a9-81f8-86b520b87518.bs2io90l08k
qb1od8lcg.databases.appdomain.cloud:31
198/bludb
Done.
```

Out[60]:

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

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

- Present your query result with a short explanation here

  - Most failure cases are on drone ship recently.

```
In [68]:
```

```
%%sql
select landing__outcome, date
from spacextbl
where date between '2010-06-04' and '2017
order by date desc;
```

```
 * ibm_db_sa://zgr48431:***@0c77d6f2-5
da9-48a9-81f8-86b520b87518.bs2io90l08k
qb1od8lcg.databases.appdomain.cloud:31
198/bludb
Done.
```

```
Out[68]:
```

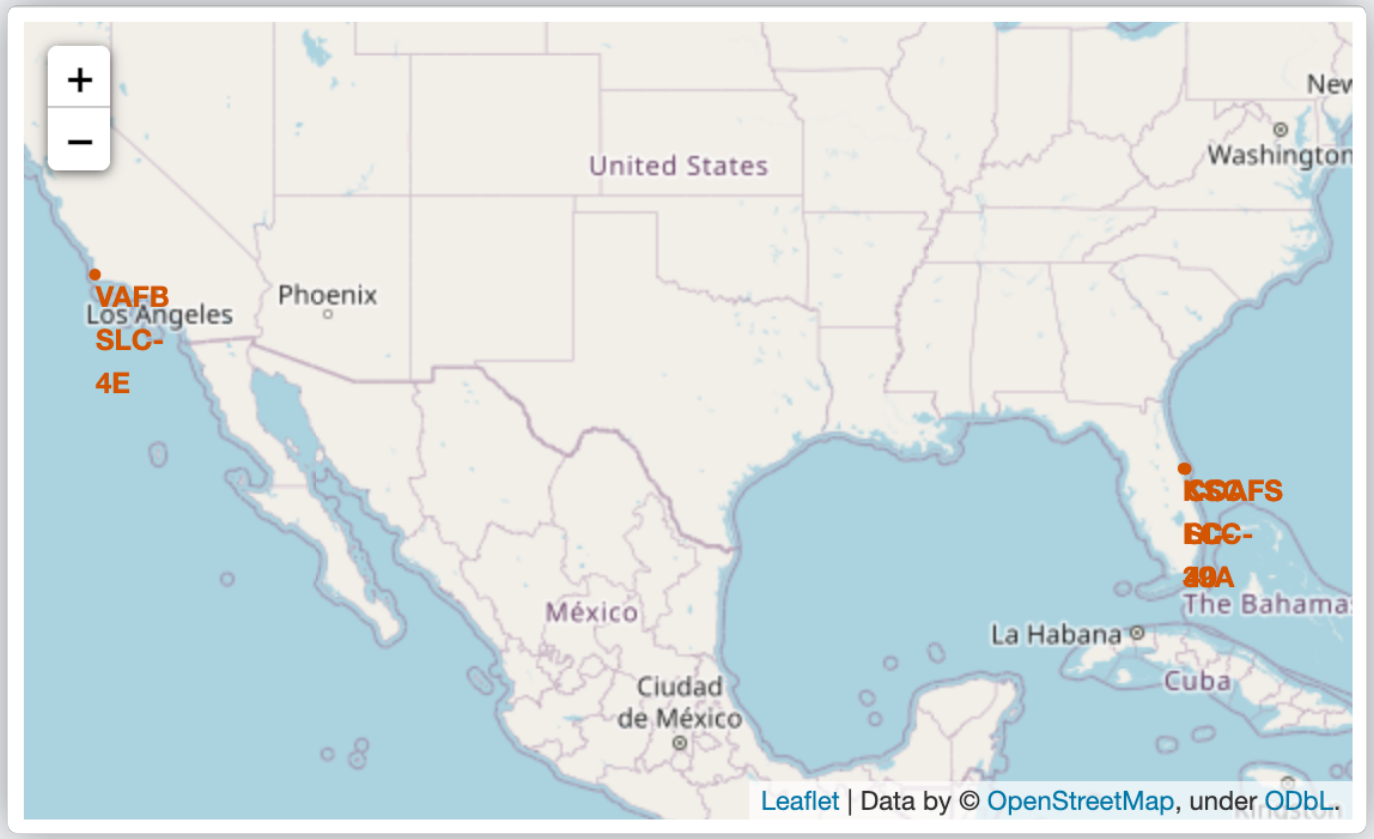
landing__outcome	DATE
No attempt	2017-03-16
Success (ground pad)	2017-02-19
Success (drone ship)	2017-01-14
Success (drone ship)	2016-08-14
Success (ground pad)	2016-07-18
Failure (drone ship)	2016-06-15
Success (drone ship)	2016-05-27
Success (drone ship)	2016-05-06
Success (drone ship)	2016-04-08
Failure (drone ship)	2016-03-04
Failure (drone ship)	2016-01-17
Success (ground pad)	2015-12-22

Section 4

Launch Sites Proximities Analysis

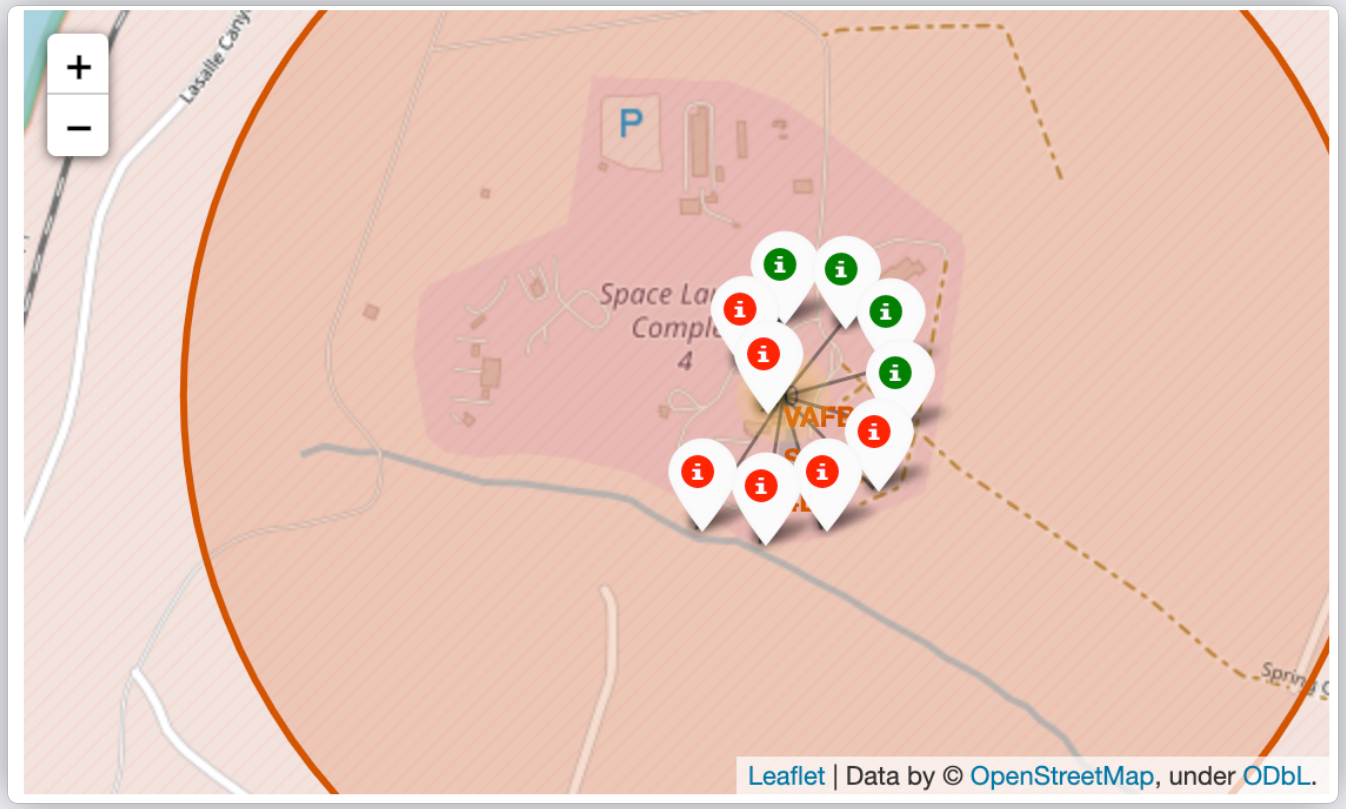
Best launch sites

- Explain the important elements and findings on the screenshot
  - Both ground and sea surface sites are necessary, also, south areas maybe a proper areas.
  - The transportation base chosen maybe important



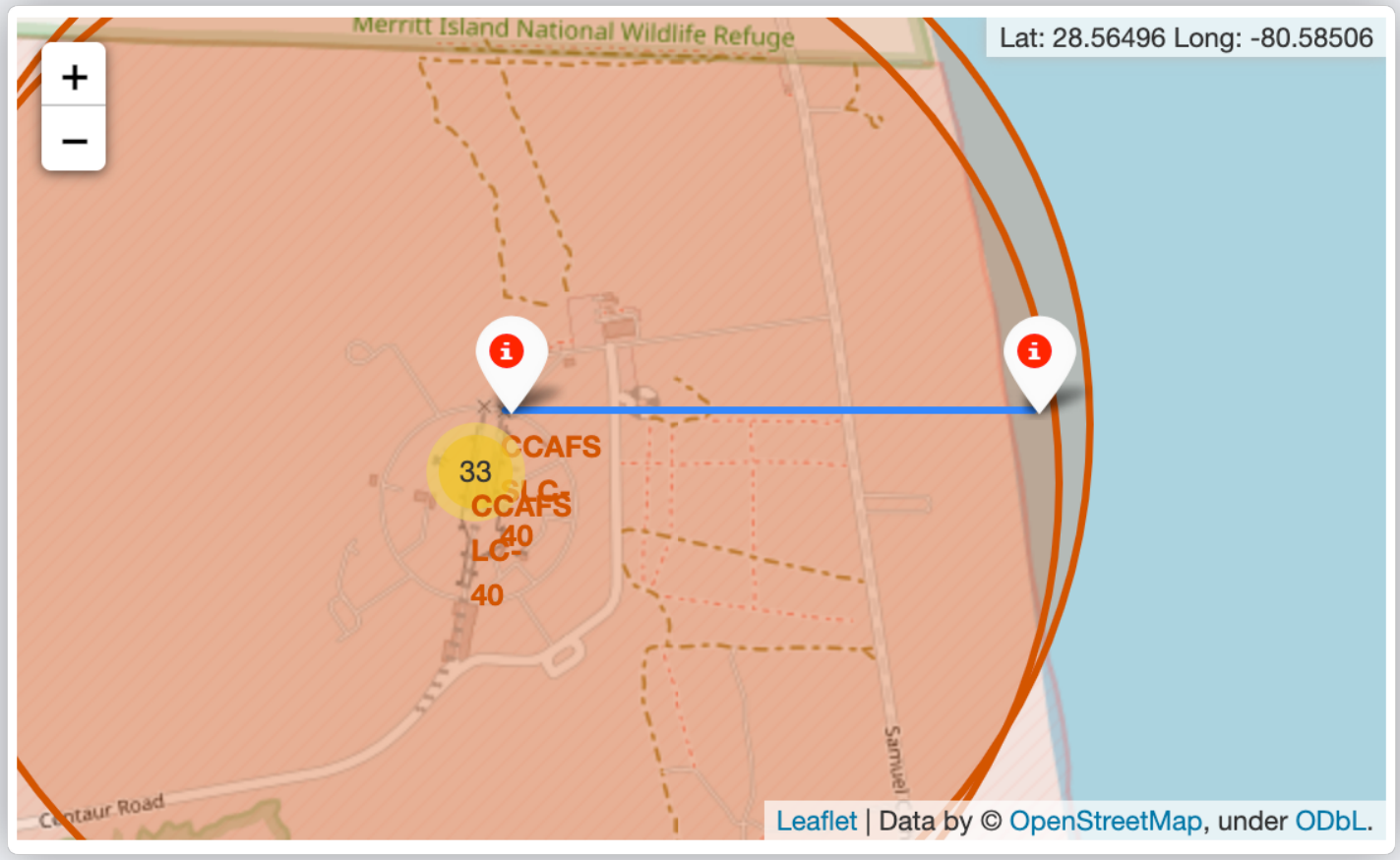
Red vs. Green

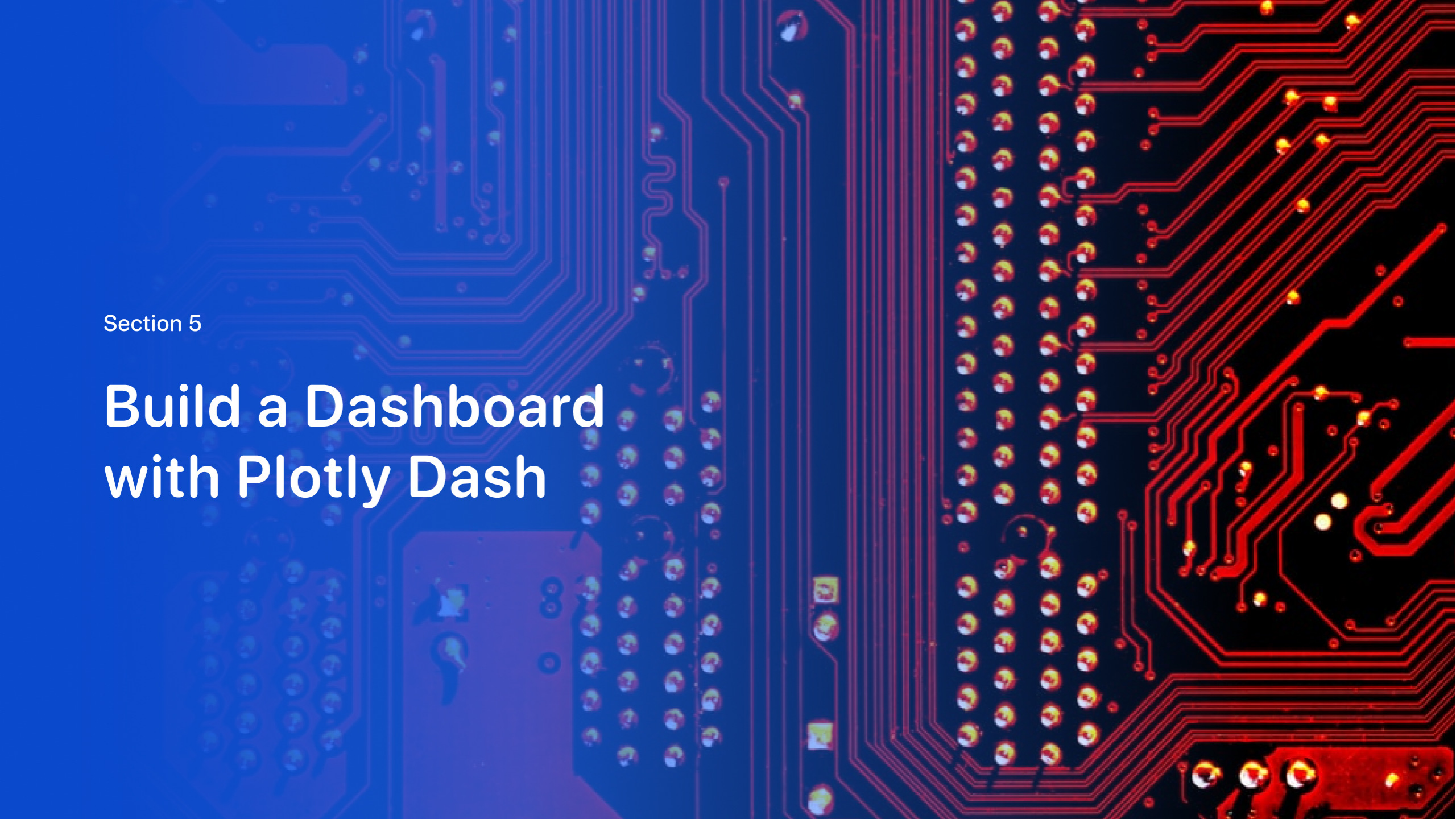
- Explain the important elements and findings on the screenshot
  - Color icons is impressive way to show the rate of success, in VABF, failure rate is high.



Transportation

- Explain the important elements and findings on the screenshot
  - Visualization shows importance of transportation, for example the railway, coast or port.



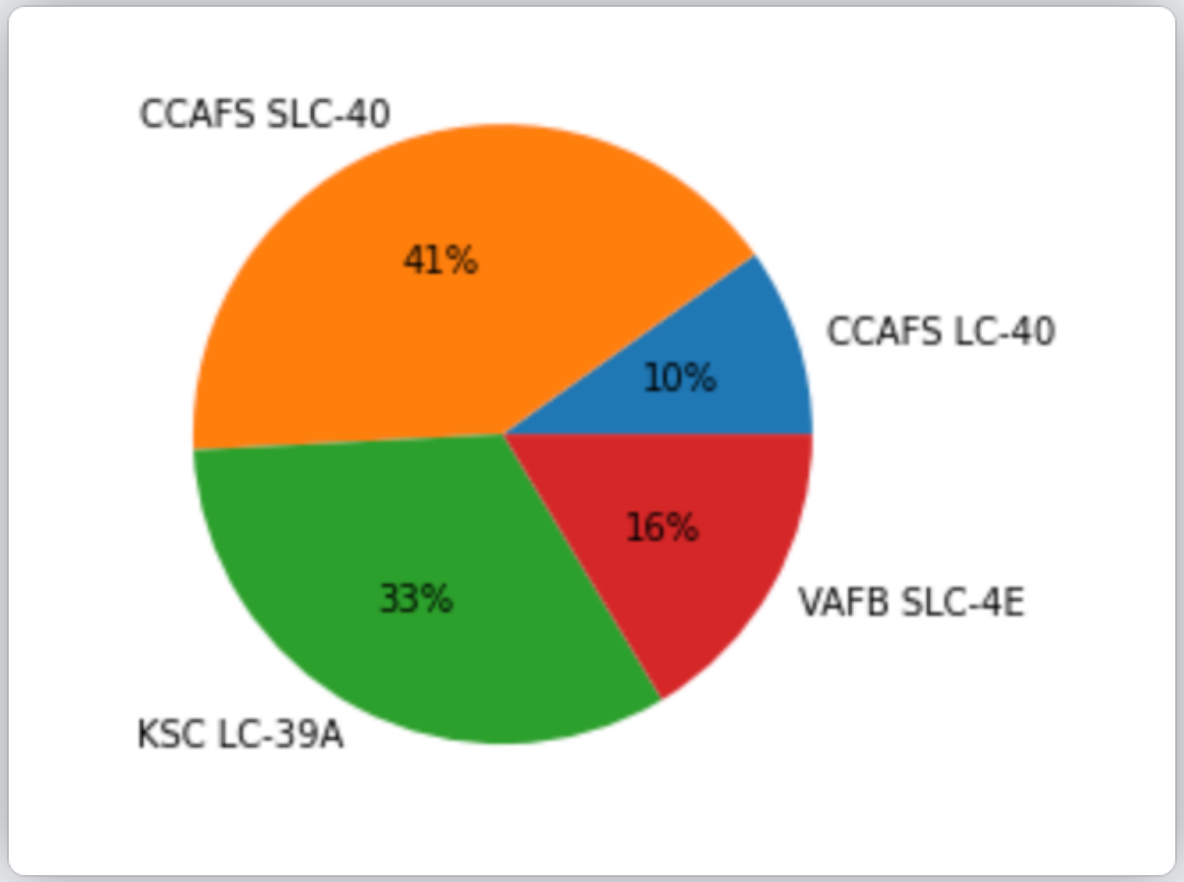


Section 5

Build a Dashboard with Plotly Dash

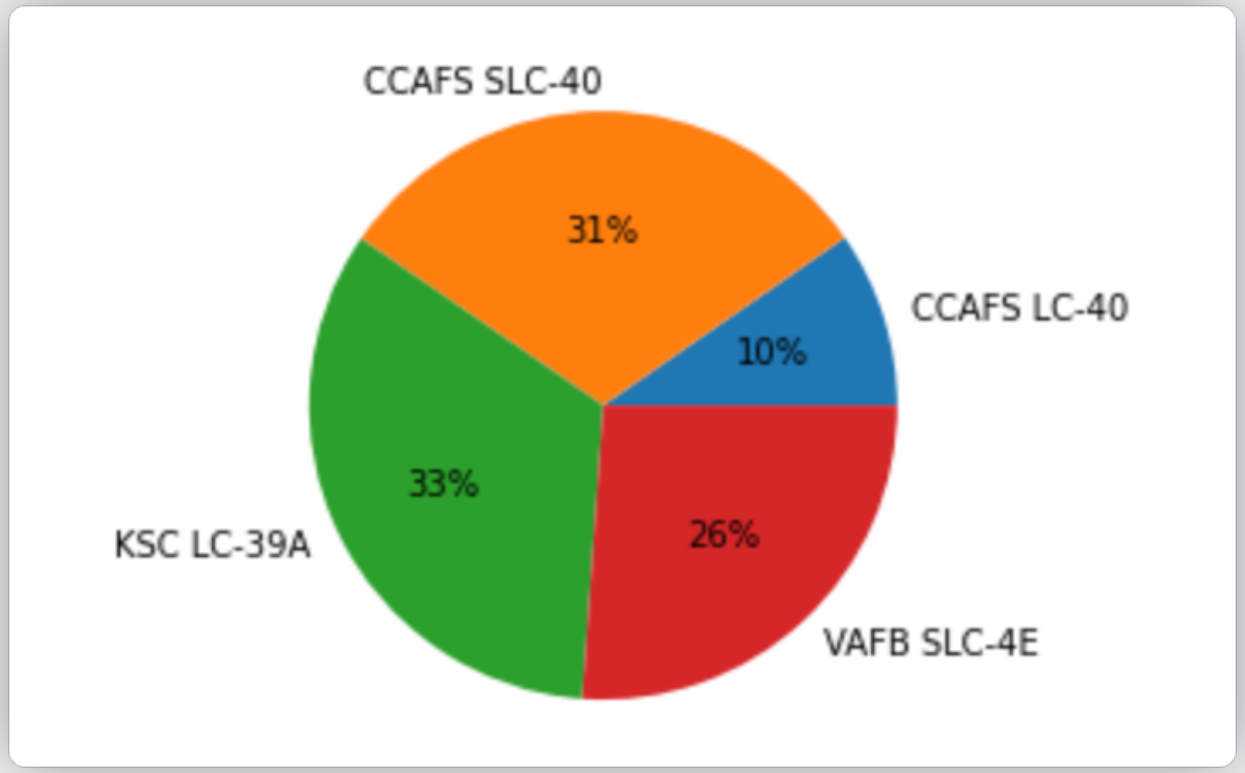
The most successful launch site

- Explain the important elements and findings on the screenshot
  - Key findings: The most successful launch sites are CCAFS SLC-40 and KSC LC-39A



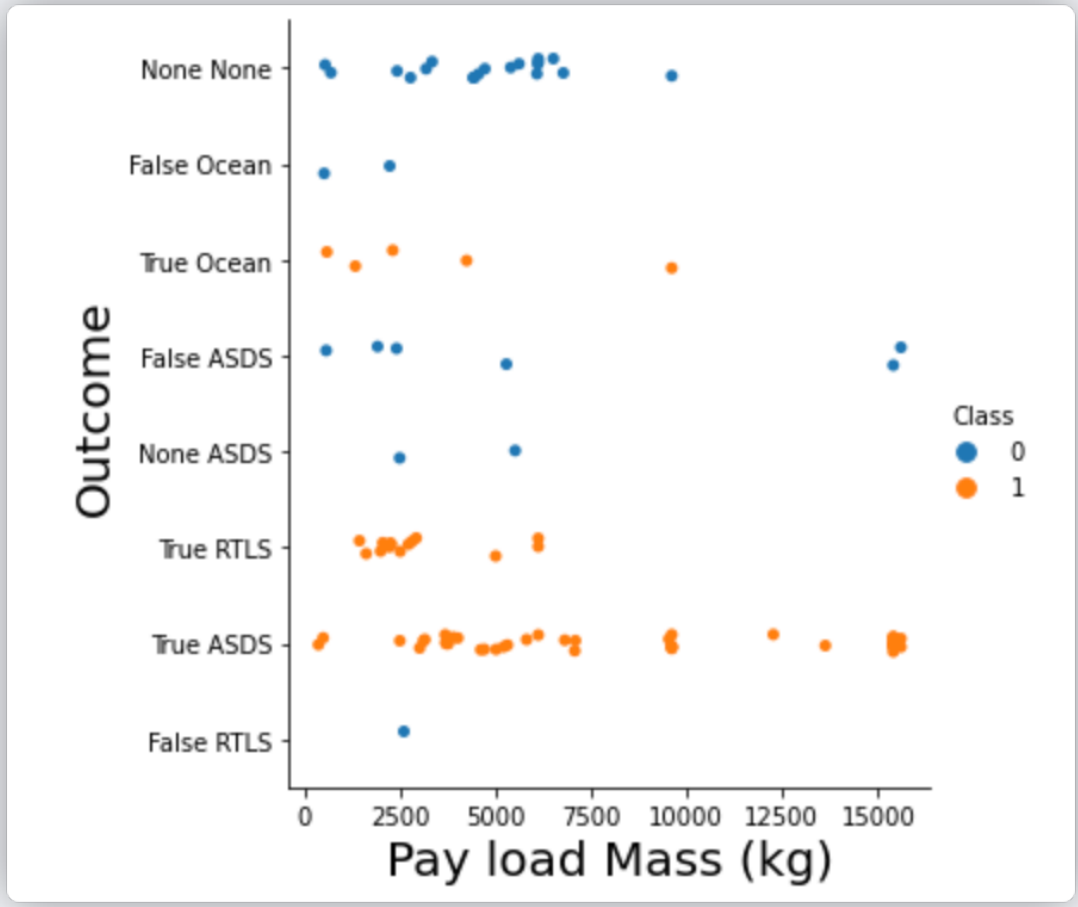
The most feasible launch site

- Explain the important elements and findings on the screenshot
  - KSC LC-39A has the highest successful rate.
  - The CCAFS SLC-40 follows.



Promising launch sites

- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.
  - One key point:
    - The ocean & RTLS outcome has high successful rate for the small and medium mass launch



Section 6

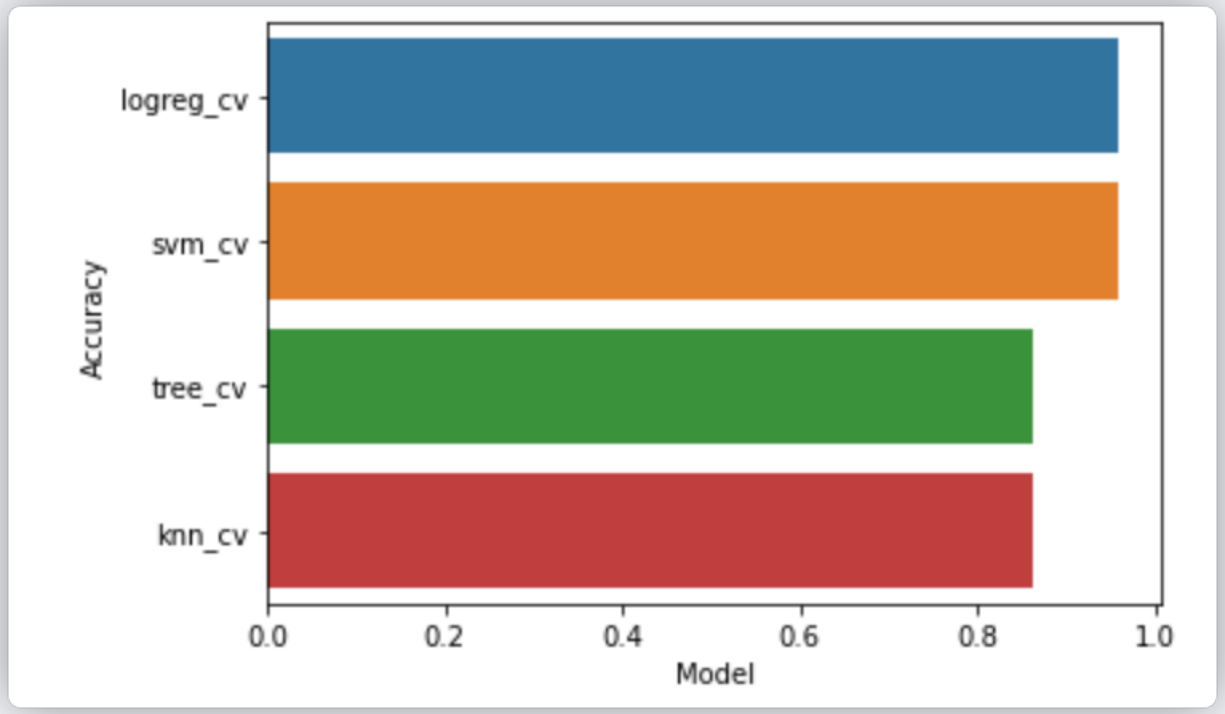
Predictive Analysis (Classification)

Classification Accuracy

- Visualize the built model accuracy for all built classification models, in a bar chart

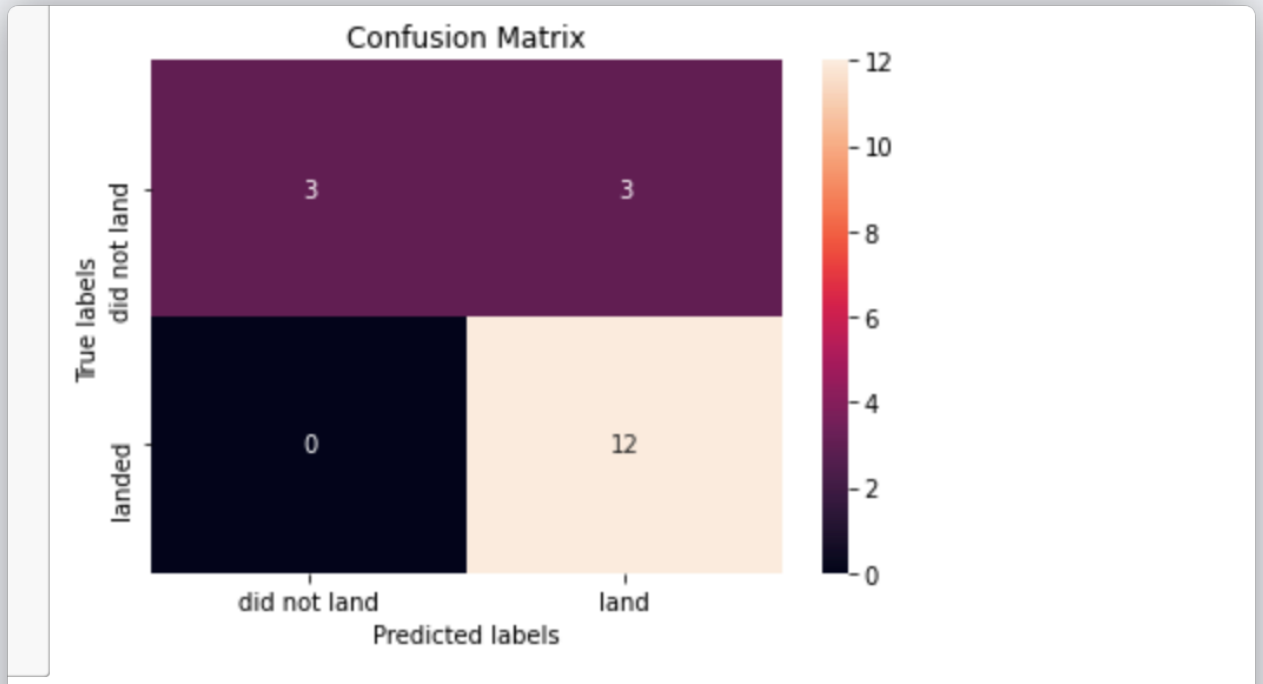
- Find which model has the highest classification accuracy

  - logistic regression model



Confusion Matrix

- Show the confusion matrix of the best performing model with an explanation



Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.

Conclusions

- Point 1: Launch site has a strong relation with success rate.
- Point 2: CCAFS SLC-40 and KSC LC-39A are most feasible sites with highest success rate and amount of launch.
- Point 3: The ocean & RTLS outcome have succussed on small and medium mass launches, and there is more chances for improvement.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

- SpaceX Falcon 9 first stage Landing Prediction
- Space X Falcon 9 First Stage Landing Prediction
- Complete the EDA with SQL lab
- Complete the EDA with Visualization lab
- Launch Sites Locations Analysis with Folium
- Space X Falcon 9 First Stage Landing Prediction

Thank you!