IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Andy Tsai
2022/07/13

Outline

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

Executive Summary

- Summary of methodologies
    - Data are collected from SpaceX API and web scraping from SpaceX Wikipedia page
    - By using data visualization and create a dashboard, some insight can be found by figure.

- Summary of all results
    - Top majority:
      to develop a model to predict successful Stage 1 recovery in order to lower the cost.
    - The machine learning model with best accuracy is KNN classification, with accuracy of 90.3%

Introduction

- Project background and context
  - The commercial space age is coming. Offering space travel affordable for everyone may be feasible in recent future.
  - Some company has already launch their plan, such as Virgin Galactic and SpaceX
  - SpaceX is the benchmark in this field owing to lower rocket price

- Problems you want to find answers
  - To built a company(SpaceY) which can compete with SpaceX, top majority is to train a model to predict successful Stage 1 recovery .

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data from SpaceX public API and Wikipedia page of SpaceX will be used in this project
- Perform data wrangling
  - Convert landing outcomes(successful/ unsuccessful) into (1/0) class
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Machine learning methodologies, including Decision Tree Classifier, Logistic regression, KNN and SVM will be used.
  - GridSearchCV will be used to tuning model.

Data Collection

- Describe:
  - Data were collected by two ways: SpaceX public API and Wikipedia page.
  - Selecting useful data and dealing with missing value will be down in this step.
  - The flow of collecting data from both ways are shown in following slides.

Data Collection – SpaceX API

- The flow chart of collecting data from SpaceX API show on right hand side.
- The step can conclude with 2 points:
  - Query and select data
  - Clean data

- Notebook: 1-jupyter-labs-spacex-data-collection-api.ipynb

Request data from SpaceX API

↓

Decode request data into .json

↓

Normalize request(.json) into dataframe

↓

Select specific column fron dataframe ('rocket', 'payloads', 'launchpad', 'cores' …)

↓

Rescreen each column to extract useful data for analysis

↓

Filter dataframe to only include Falcon 9 launches

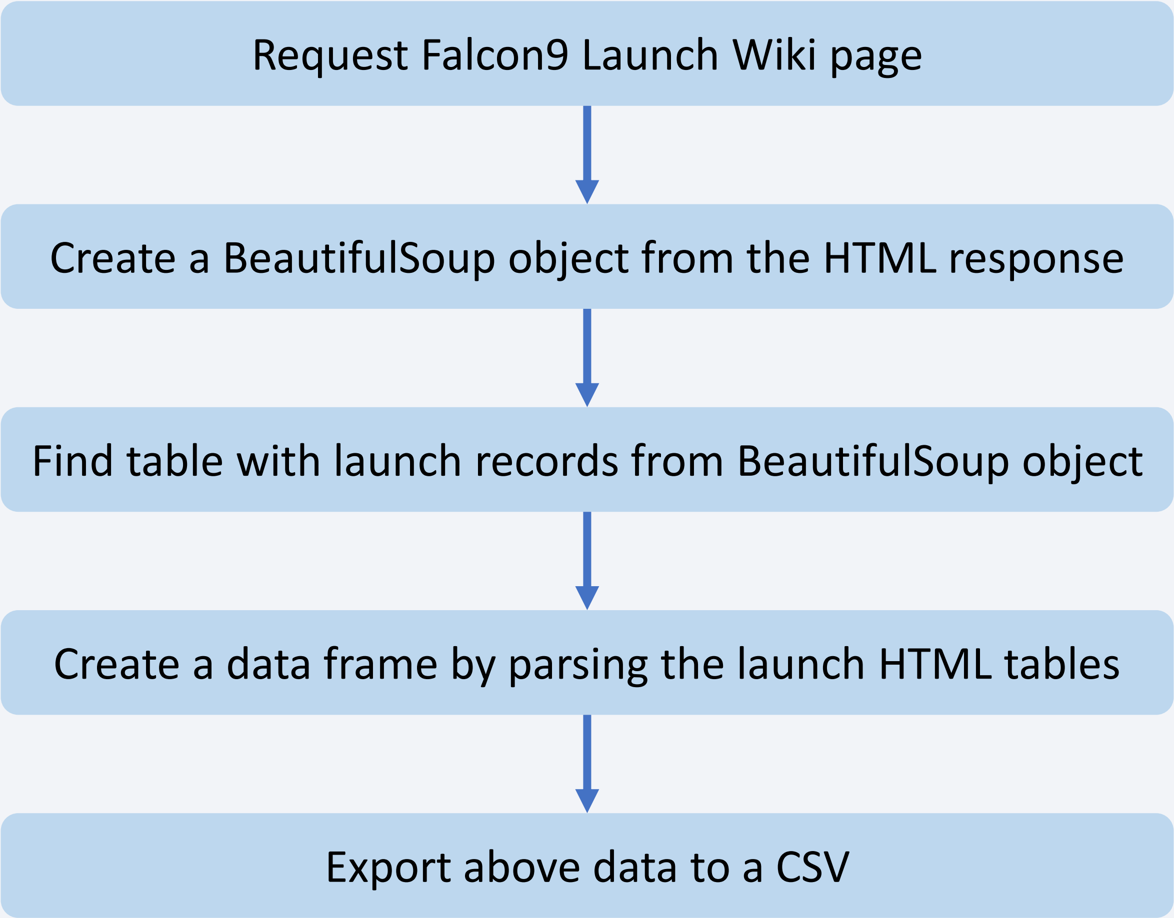
↓

Using mean value to deal with missing values

Data Collection - Scraping

- The flow chart of collecting data from Wikipedia page.
- Data collect from wiki and SpaceX API will be used in later analysis.

- Notebook: 2-jupyter-labs-webscraping.ipynb



Data Wrangling

- How data were processed:
  1. Import collected data

  2. Outcome column expose two message: ‘Mission Outcome’ and ‘Landing Location’

  3. We try to label the outcomes.
     - If mission success, then lable will be ‘1’,iIf mission fail, then lable will be ‘0’,

  4. Therefore, [True ASDS, True RTLS, True Ocean] will be labeled as ‘1’, [None None, False ASDS, False Ocean, None ASDS, False RTLS] will be labeled as ‘0’.

- Notebook: 3-labs-jupyter-spacex-Data wrangling.ipynb

EDA with Data Visualization

- Summarize:
  - By using data visualization, the relation of successful rate with other properties (and other relationship) can be found.
    - Success rate generally increases over years.
    - Success rate is relative to orbit type.
    - Launch Site has its payload mass limit.
    - Payload mass is relative to orbit type.

- Notebook: 4-jupyter-labs-eda-sql-coursera_sqllite.ipynb

EDA with SQL

- Summary of SQL queries
    - Loaded data set from .db file
    - Find the names of the unique launch sites
    - Find 5 records where launch sites begin with `CCA`
    - Calculate the total payload carried by boosters from NASA
    - Calculate the average payload mass carried by booster version F9 v1.1
    - Find the dates of the first successful landing outcome on ground pad
    - Select boosters name with specific payload mass .

- Notebook: 4-jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Summarize & explanation
  - Folium maps mark the important property on the map, such as Launch Sites, successful and unsuccessful landings, and a proximity distance to railway, highway, coast, and city.
  - Interactive map allows us to understand launch sites' relative relationship in real world.

- Notebook: 5-lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Summarize & explanation
  - Dashboard includes two parts: a pie chart and a scatter plot.
  - A Range_Slider and a Dropdown items are added to select specific data.
  - The pie chart is used to visualize launch site success rate.
  - The scatter is used to see the relationship between launch sites, payload mass, and booster version category

- Python script: 6-spacex_dash_app.py

Predictive Analysis (Classification)

- The flow chart of predictive analysis are shown on the right hand.

- The KNN classification model (with best parameter found by GridSearchCV) come out with best accuracy.

- Notebook: 7-SpaceX_Machine Learning Prediction_Part_5.ipynb

Data Standardize

↓

Split the data into train set and test set

↓

Testing different classification model for prediction

↓

Using GridSearchCV to find the best model parameter

↓

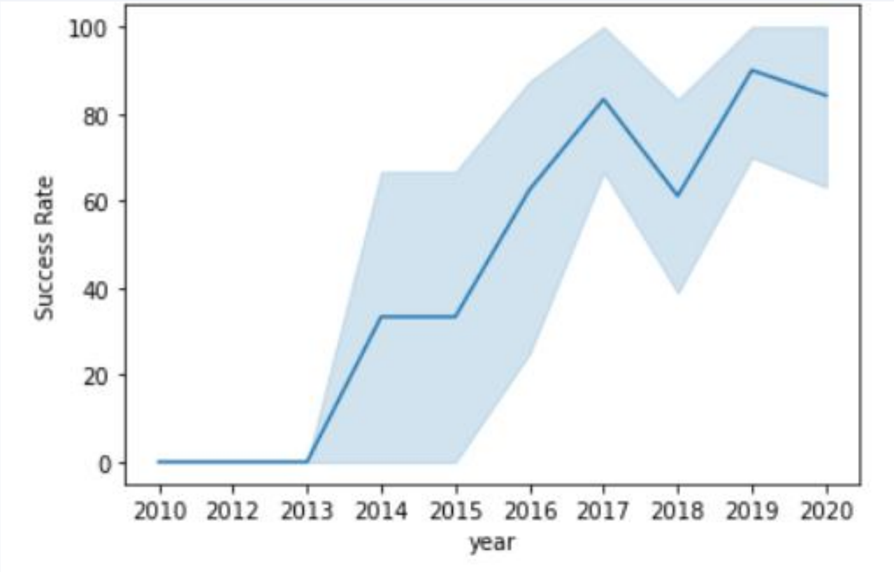
Check R square and confusion matrix evaluate result

↓

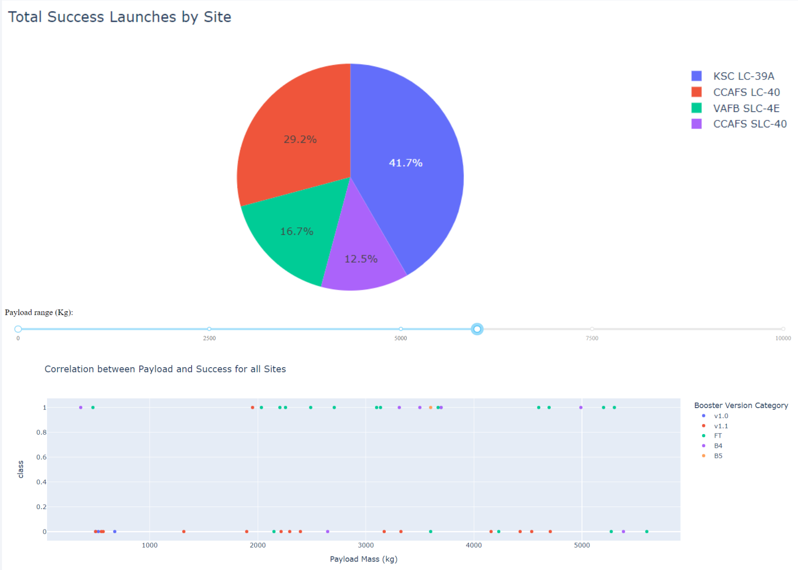
Using bar chart to comparing different model

Results

- Exploratory data analysis results



- Interactive analytics demo in screenshots



- Predictive analysis results

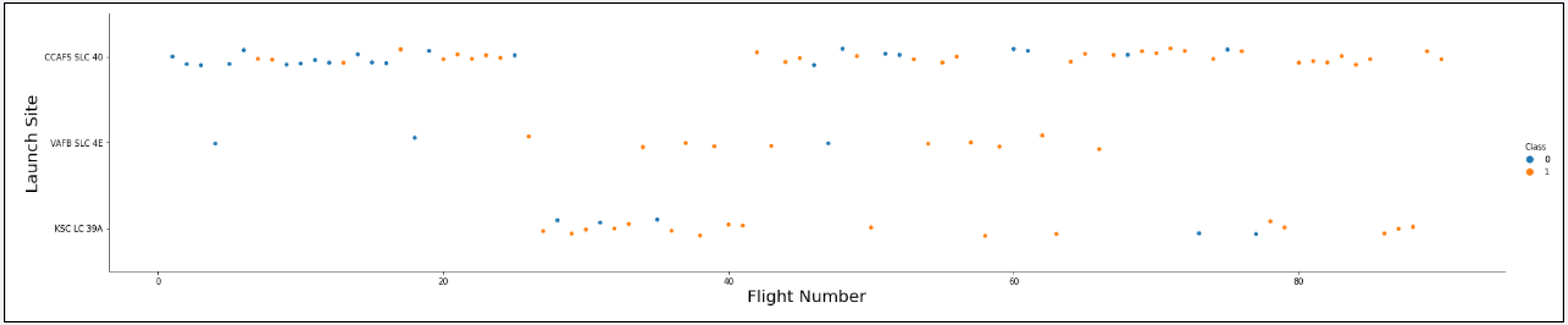
	Algorithm	Accuracy Score	Best Score
0	Logistic Regression	0.833333	0.847222
1	Support Vector Machine	0.833333	0.847222
2	Decision Tree	0.777778	0.888889
3	K Nearest Neighbours	0.833333	0.902778

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

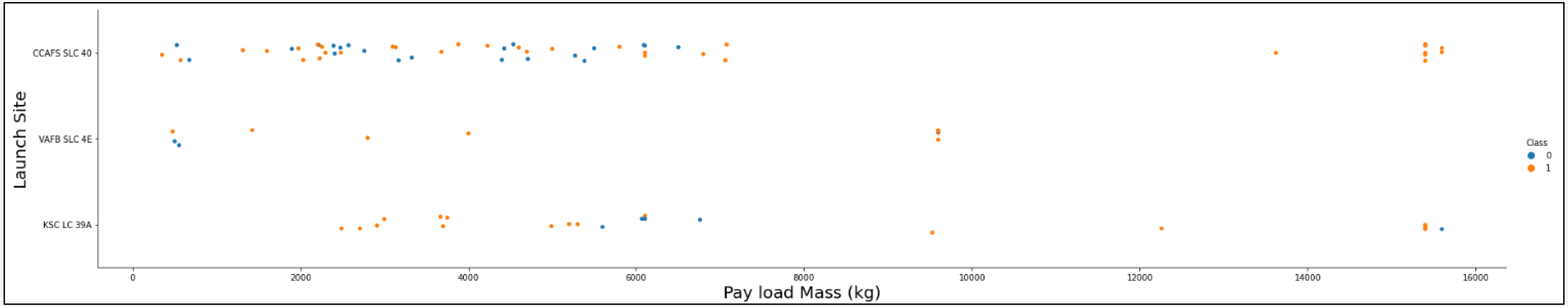
- Scatter plot of Flight Number vs. Launch Site



- Explanations
  - If flight number above 20, then successful rate rapidly increase.
  - Different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
  - CCAFS is the main launch site which has the highest flight number.

Payload vs. Launch Site

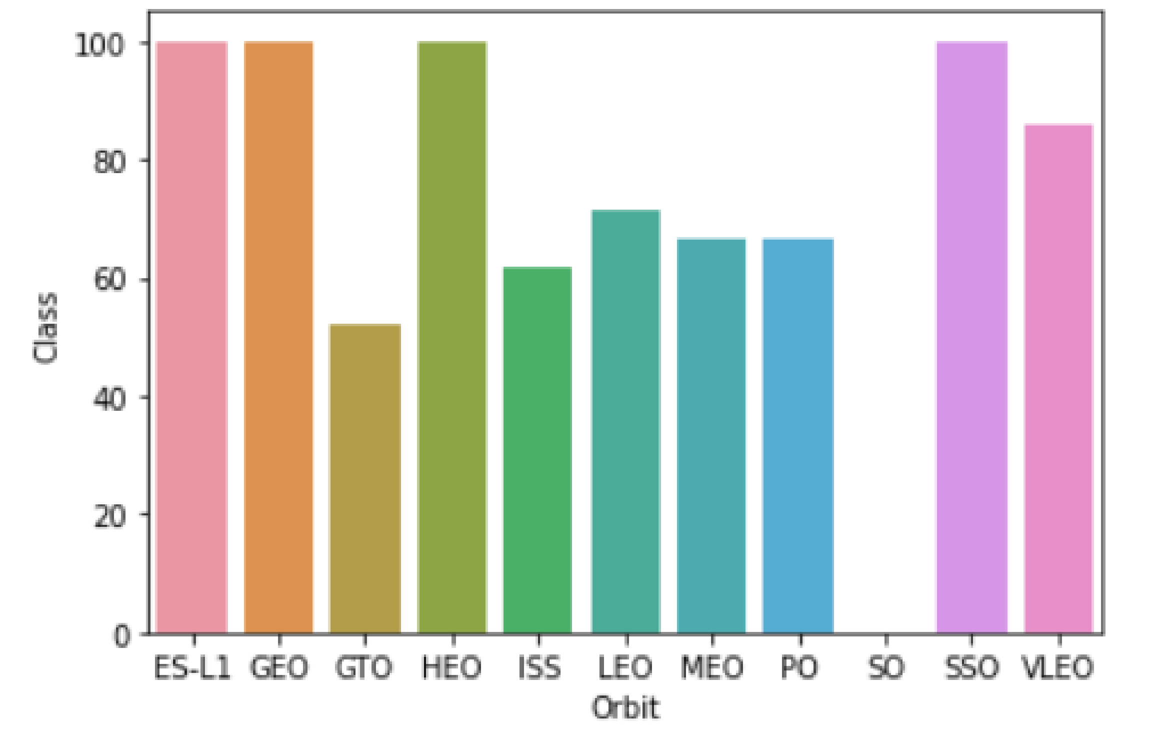
- Scatter plot of Payload vs. Launch Site



- Explanations
  - Launch sites VAFB SLC 4E's max pay load mass is below 10000.
  - Most payload mass in a range of 0-6000 kg.

Success Rate vs. Orbit Type

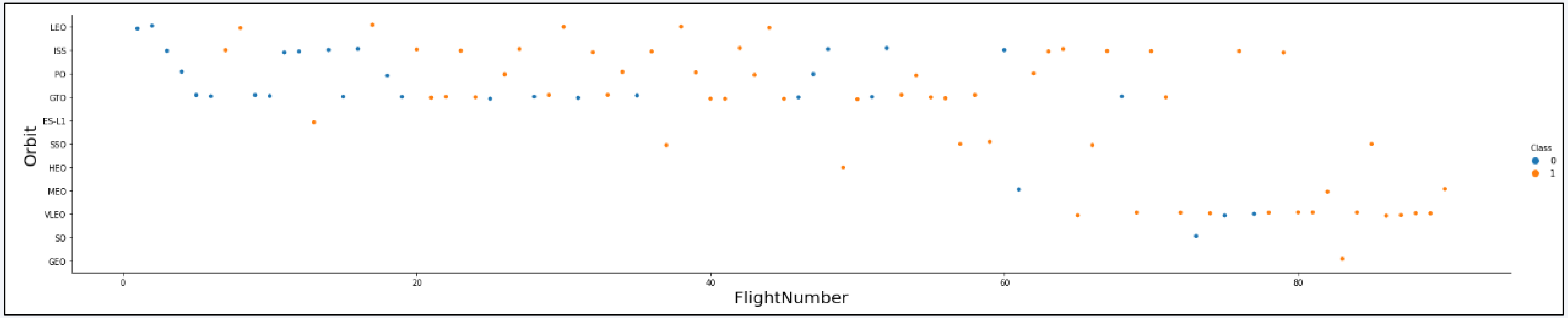
- Bar chart for the success rate of each orbit type



- Explanations
  - SSO has 100% success rate with 5 sample.
  - ES-L1 , GEO , HEO have 100% success rate with only 1 sample
  - SO (1) has 0% success rate
  - GTO (27) has the around 50% success rate but largest sample

Flight Number vs. Orbit Type

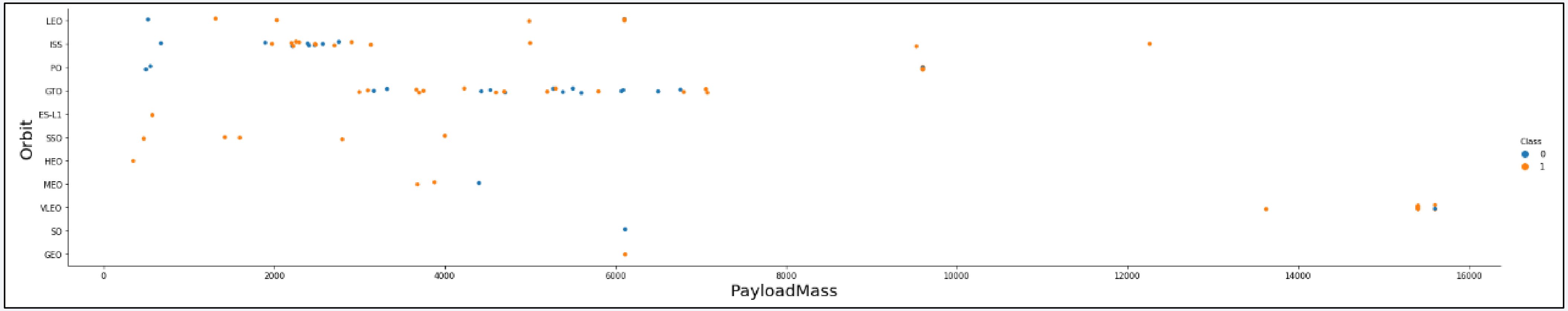
- Scatter point of Flight number vs. Orbit type



- Explanations
  - With flight number become bigger, the success rate increase.
  - The later setting Orbit preferences(with bigger flight number) perforce better.
  - SpaceX appears to perform better in LEO or SO

Payload vs. Orbit Type

- Scatter point of payload vs. orbit type

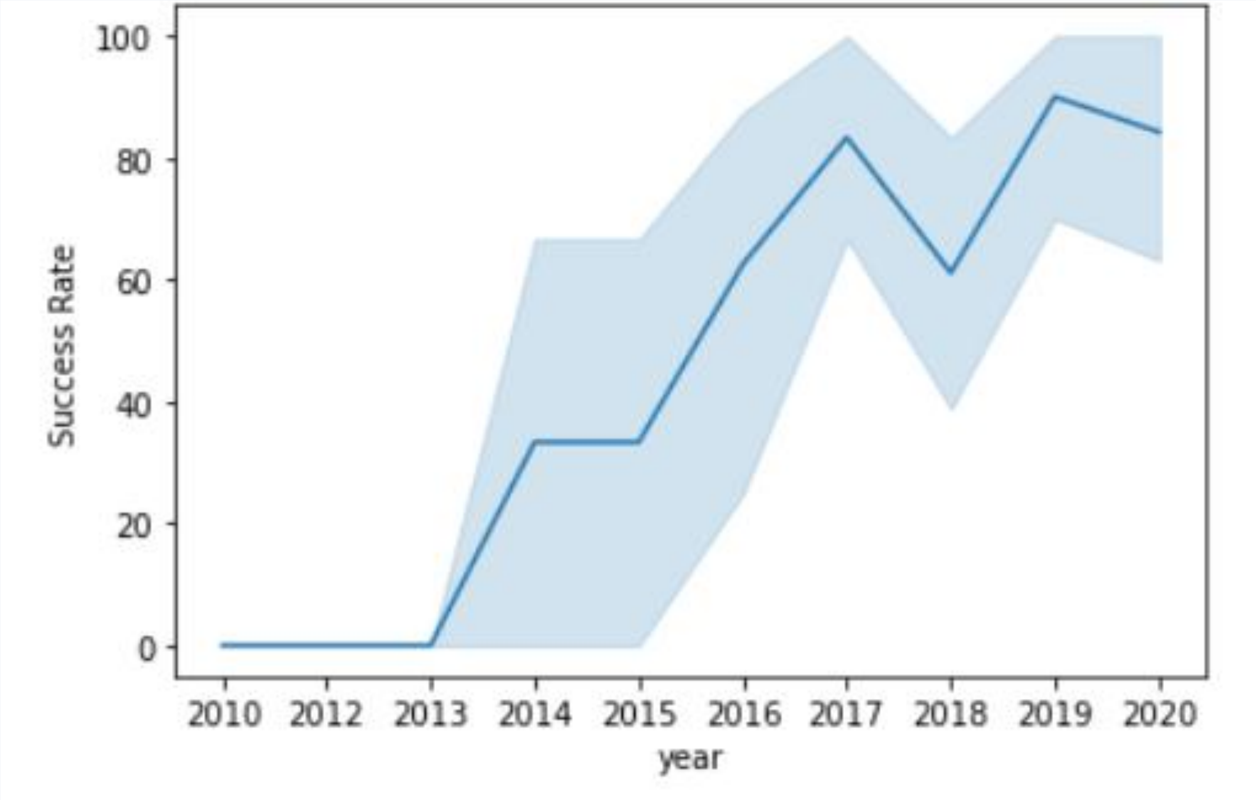


- Explanations
  - LEO and SSO seem to have relatively low payload mass
  - VLEO always com with largely higher payload then other orbit.

Launch Success Yearly Trend

- Line chart of yearly average success rate

- Explanations
  - Success rate generally increases over years.
  - Success rate only slight dip in 2018.
  - Success is above 80% in recent years.



All Launch Site Names

- Find the names of the unique launch sites
- Present your query result with a short explanation here
    - 4 launch site are listed in the right fig. They are:
        - CCAFS LC-40
        - VAFB SLC-4E
        - KSC LC-39A
        - CCAFS SLC-40

```
%sql select DISTINCT LAUNCH_SITE from SPACEXTBL
```

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

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Present your query result with a short explanation here

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

 * sqlite:///my_data1.db
Done.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	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
  - The total payload carried by boosters from NASA is 45996 kg.

```
%sql select sum(payload_mass__kg_) as sum from SPACEXTBL where customer like 'NASA (CRS)'
```

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

sum
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
  - The average payload mass carried by booster version F9 v1.1 is about 2543.67 kg.

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

```
%sql select avg(payload_mass__kg_) as Average from SPACEXTBL where booster_version like 'F9 v1.1%'
```

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

Average
2534.6666666666665

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
  - The first successful landing outcome on ground pad is 2017/1/5

```
%sql select min(date) as Date from SPACEXTBL where "Landing _Outcome" like 'Success (ground pad)'
```

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

Date
01-05-2017

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
  - 4 boosters name retrieve which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
%%sql
select booster_version from SPACEXTBL
where (mission_outcome like 'Success')
AND (PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000)
AND ("Landing _Outcome" like 'Success (drone ship)')
```

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

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 successful outcomes is 100 times while there were 1 time payload status is unclear.

```
%sql SELECT mission_outcome, count(*) as no_Count FROM SPACEXTBL GROUP by mission_outcome
```

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

Mission_Outcome	no_Count
Failure (in flight)	1
Success	98
Success	1
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
  - Table on the right hand shows the max payload mass is 15600.
  - These booster versions are all F9 B5 B10xx.x variety.

```
%%sql
select booster_version, PAYLOAD_MASS__KG_ from SPACEXTBL
where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Present your query result with a short explanation here
  - 2 failed landing_outcomes were queried in above condition.

```
%%sql
select substr(Date, 4, 2) as Month, "Landing _Outcome", booster_version, launch_site from SPACEXTBL
where substr(Date,7,4)='2015'
AND "Landing _Outcome" like 'Failure (drone ship)'
```

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

Month	Landing _Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order
- Present your query result with a short explanation here
  - This query returns a list of landings outcomes between 2010-06-04 and 2017-03-20.
  - There are total 8 types of result and can category into 3 types:
    - Success: including [Success, Success (drone ship), Success (ground pad)]
    - Failure: including [Failure, Failure (drone ship), Failure (parachute)]
    - Other: including [No attempt, Controlled (ocean)]

```
%%sql
select "Landing _Outcome", count(*) as no_count from SPACEXTBL
where Date >= '04-06-2010' AND Date <= '20-03-2017'
GROUP by "Landing _Outcome" ORDER BY no_count Desc
```

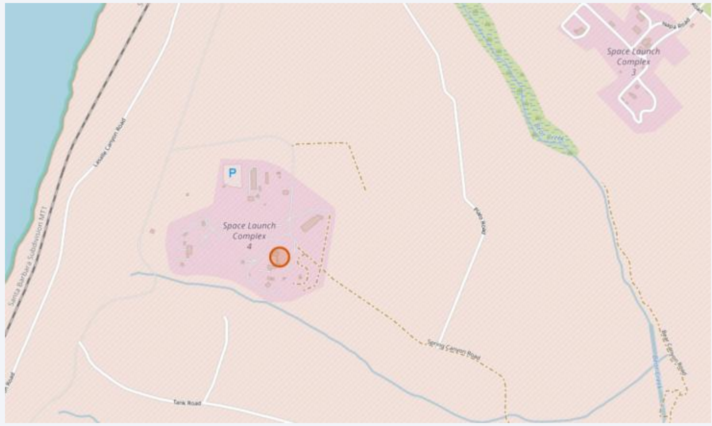
```
 * sqlite:///my_data1.db
Done.
```

Landing _Outcome	no_count
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

Section 3

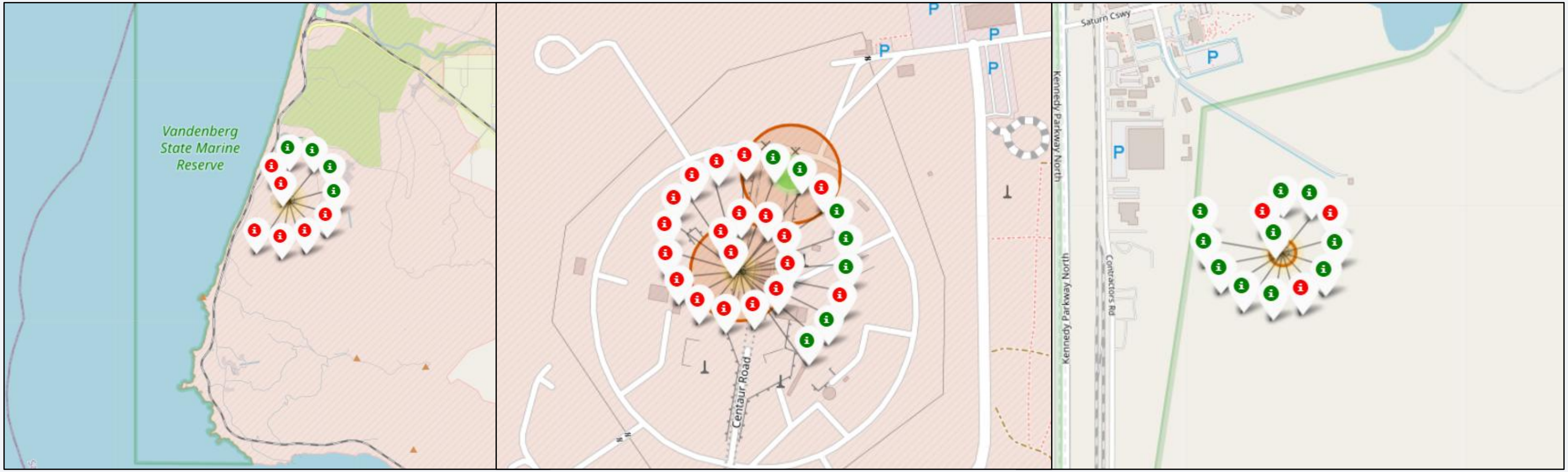
Launch Sites Proximities Analysis

All launch sites' location on map



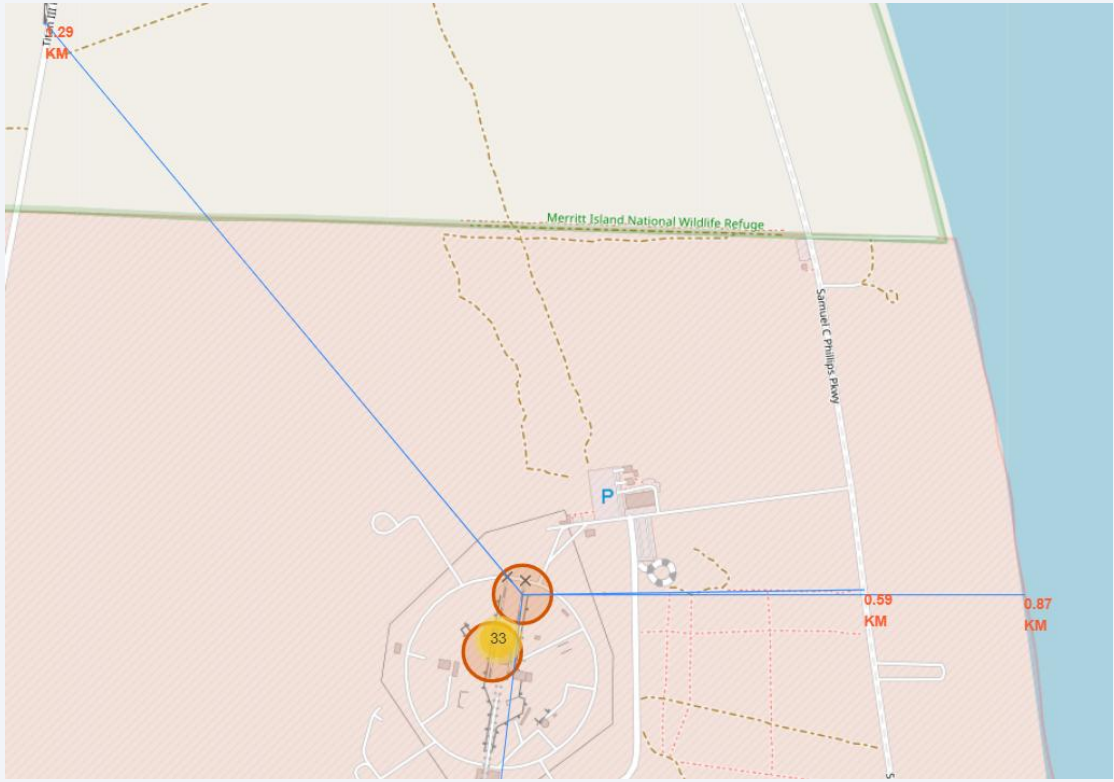
- Important elements and findings:
  - Only 1 launch site on west coast, most launch site on east coast.
  - All launch site near the sea.

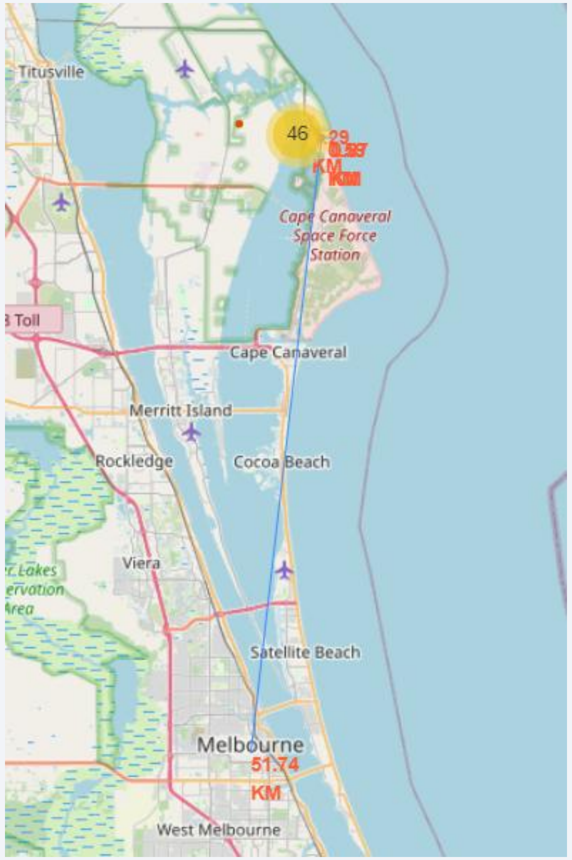
Map of color-labeled launch outcomes



- Important elements and findings:
  - By using clusters, launch outcome of each site can be visualize
  - In the figure above, green icon means successful landing while red icon means failed landing.

Distance to key location



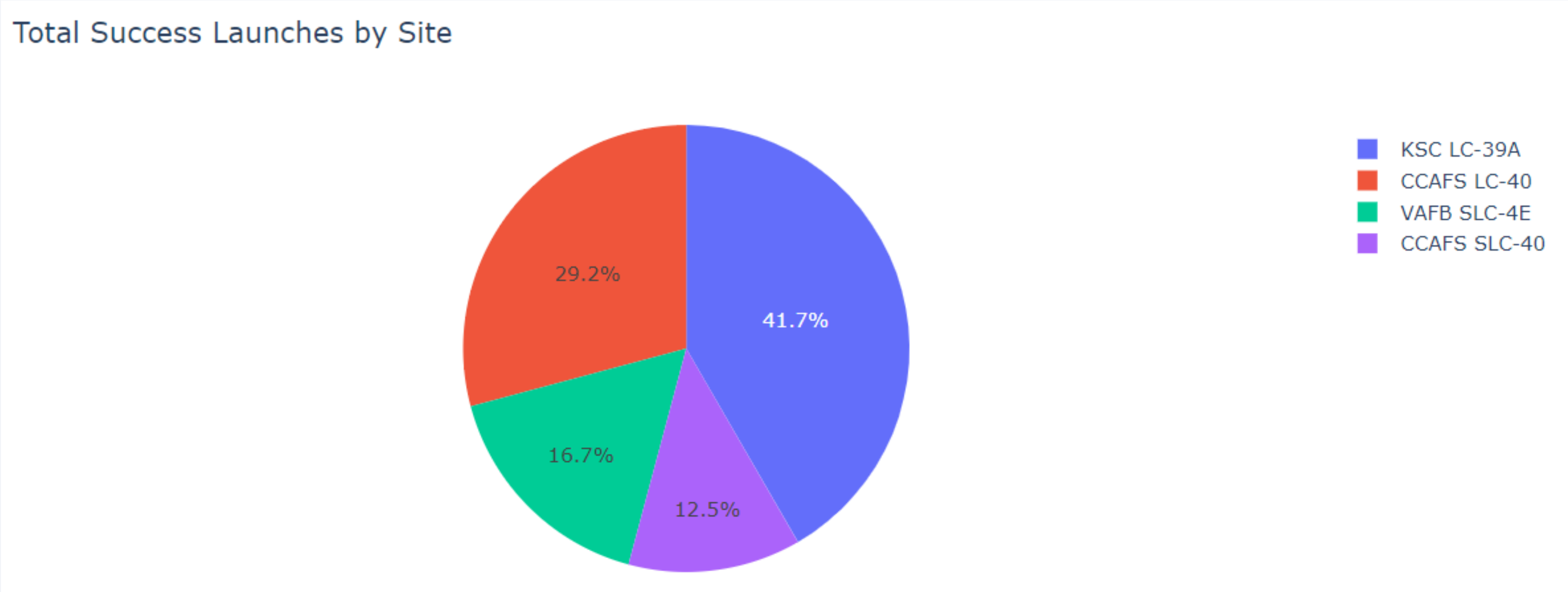


- Important elements and findings, take CCAFS SLC-40 as example:
  - Launch sites are close to railways, highways.
  - Launch sites are close to coasts.
  - Launch sites are relative far from populated areas.

Section 4

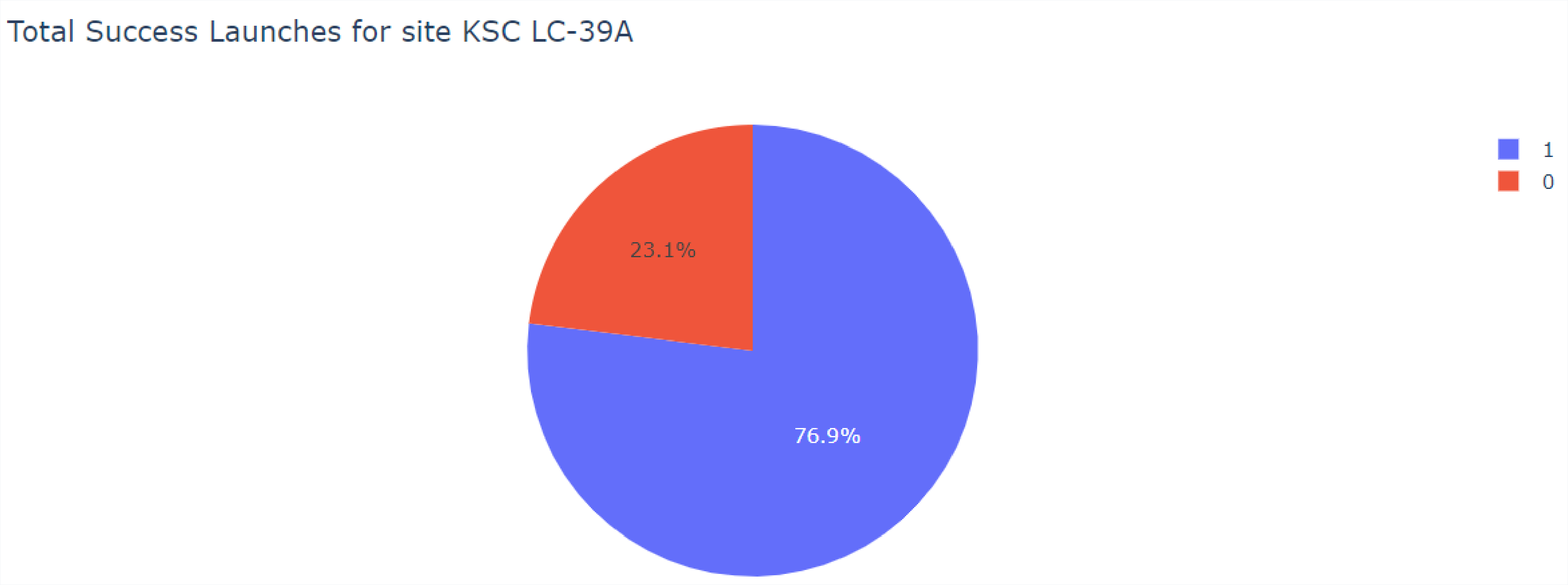
Build a Dashboard with Plotly Dash

Pie chart of total success by launches site



- Important elements and findings:
  - KSC LC-39A account for the highest success amount (41.7%) in all launches site.
  - CCAFS SLC-40 account for the lowest success amount(12.5) in all launches site.

Pie chart for launch site with highest launch success ratio



- Important elements and findings:
  - KSC LC-39A has the highest success rate(76.9%)

Payload vs. Launch Outcome scatter plot for all sites

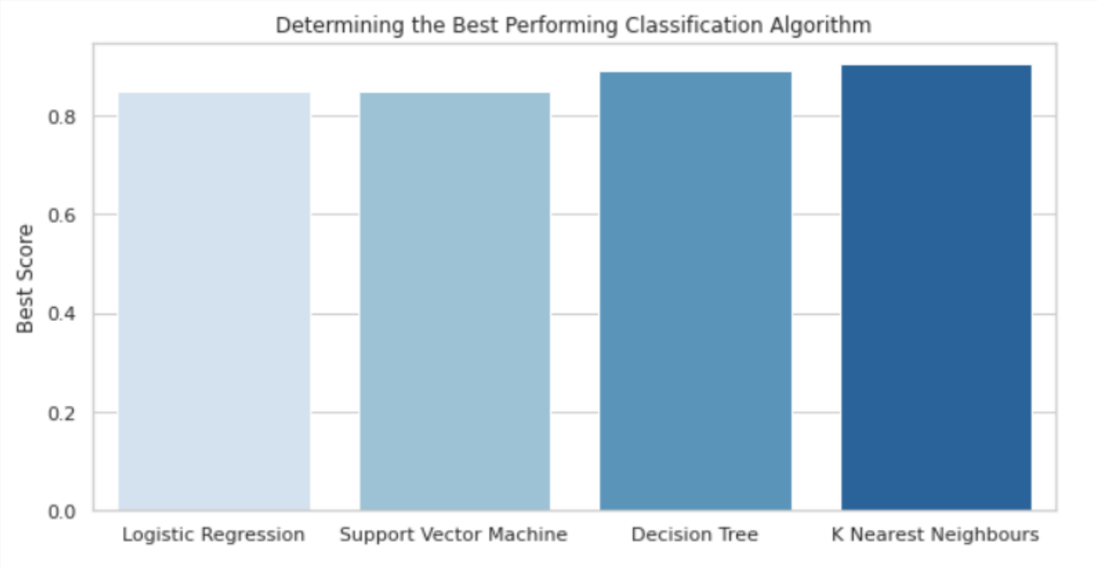
- Important elements and findings:
  - In payload range [6000~10000 kg], almost all the launch outcome is false
  - On the other hands, in payload range [0~6000 kg], the launch outcome become successful more often.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

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

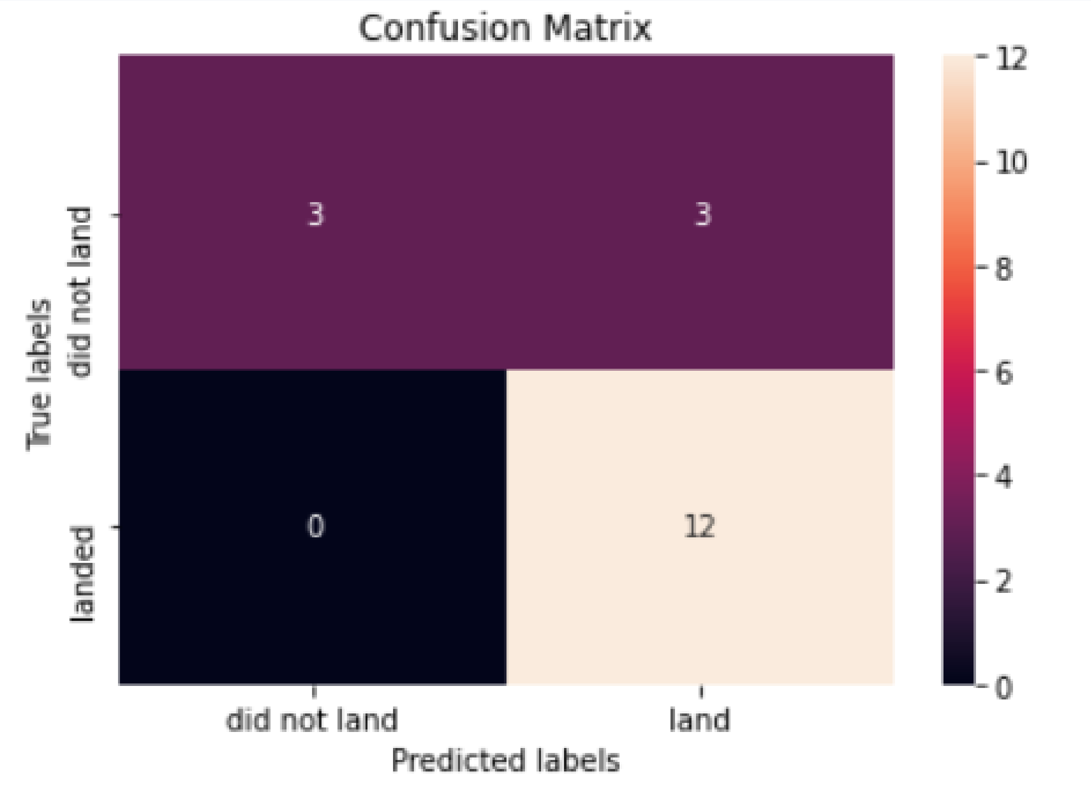


	Algorithm	Accuracy Score	Best Score
0	Logistic Regression	0.833333	0.847222
1	Support Vector Machine	0.833333	0.847222
2	Decision Tree	0.777778	0.888889
3	K Nearest Neighbours	0.833333	0.902778

- Find which model has the highest classification accuracy
  - At the condition of [test_size=0.2, random_state=2] and using gridsearch_cv to find the best score, KNN has the highest classification accuracy
  - The best score is 0.903

Confusion Matrix

- Show the confusion matrix of the best performing model with an explanation
  - KNN models predicted 12 successful landings when the true label was successful landing.
  - KNN models predicted 3 unsuccessful landings when the true label was unsuccessful landing.
  - KNN models predicted 3 successful landings when the true label was unsuccessful landings (false positives).



Conclusions

- Our task: To built a company(SpaceY) which can compete with SpaceX
- Top majority:
  to develop a model to predict successful Stage 1 recovery in order to lower the cost.
- Data are collected from SpaceX API and web scraping from SpaceX Wikipedia page
- By using data visualization and create a dashboard, some insight can be found by figure.
- The machine learning model with best accuracy is KNN classification, with accuracy of 90.3%
- If more data or features can be collected or consider, the machine learning model may improve its accuracy.

Appendix

- Author's GitHub: https://github.com/CKMaxwell

Thank you!