IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Meysam Modiryan
1 June 2022

Outline

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

EXECUTIVE SUMMARY

- ***The Methodologies to Analyze Data***

❑ Collecting the data with web scraping by Request to the SpaceX API & Clean the requested data

❑ Exploratory Data Analysis (EDA) by using SQL and VISUALIZATION

❑ Interactive Visual Analytics and Dashboard

❑ Machine Learning Prediction

- ***Summary of all results***

❑ Collection data was possible from public source

❑ By EDA we were able to figure out features' dependency and identify the best features for prediction

❑ And we understood which classification algorithm is better than the others to predict in machine learning prediction part.

Introduction

- ***Project background and context***

We predicted if the Falcon 9 first stage will land successfully. 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. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch Problems you want to find answers

- ***Desirable answers:***

  * The best way to estimate the total cost for launches, by predicting successful landings of the first stage of rockets;

  * Where is the best place to make launches.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from two sources
    - ❑Data collection API, actually Space X API (link)
    - ❑Web Scraping (link)
- Perform data wrangling
  - ❑we performed some Exploratory Data Analysis (EDA) to find some patterns in the data and determined what would be the label for training supervised models.

Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - ❑We Performed exploratory Data Analysis and determined Training Labels
        - ❖create a column for the class
        - ❖Standardize the data
        - ❖Split into training data and test data
    - ❑We found best Hyperparameter for SVM, Classification Trees and Logistic Regression and finally we found the method performs best using test data

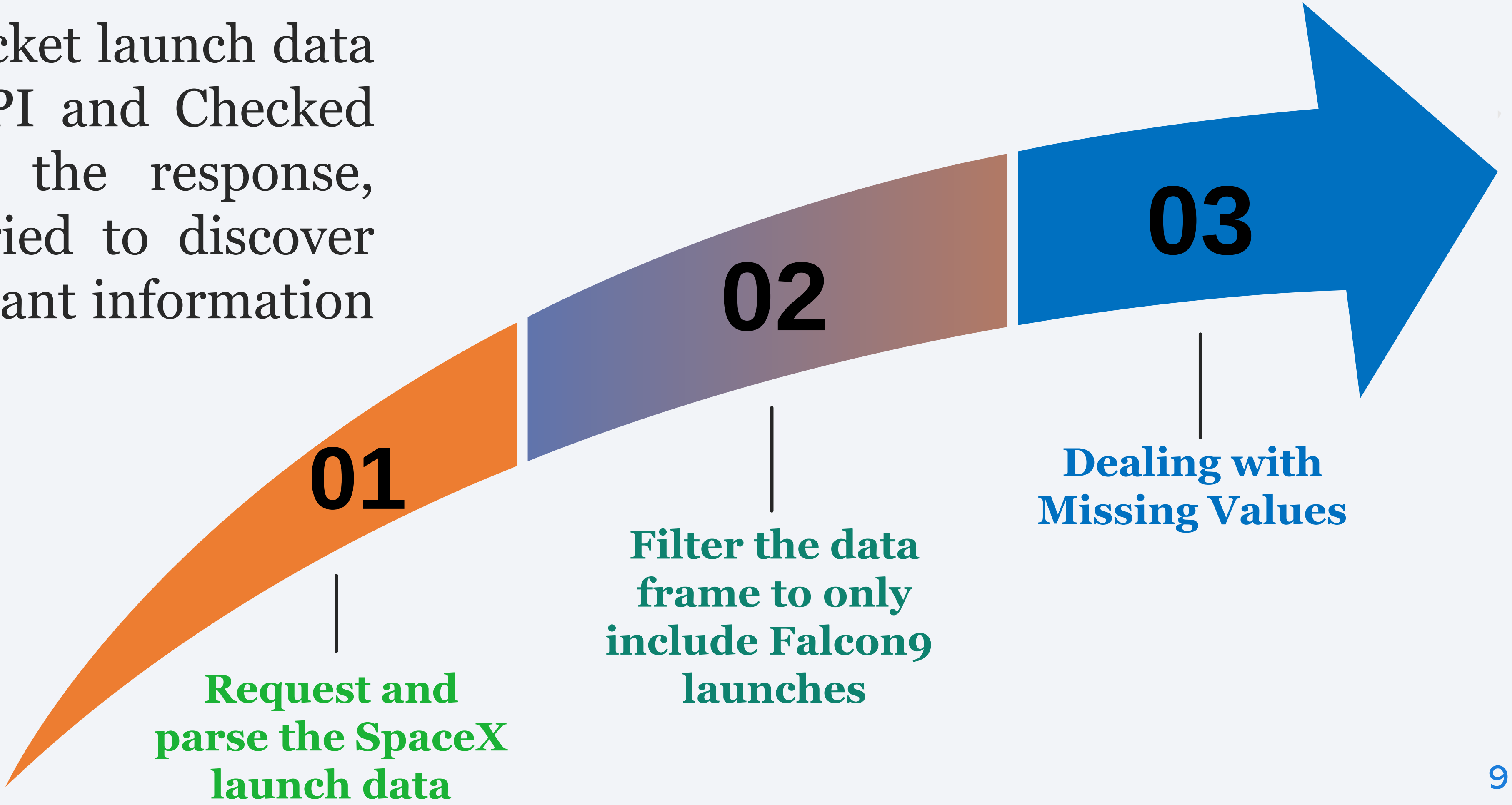
Data Collection

- Data sets were collected from:
  - Space X API
  - Wikipedia
- The links of both are provided in previous slides

In the following, we will talk about these in detail

Data Collection – SpaceX API

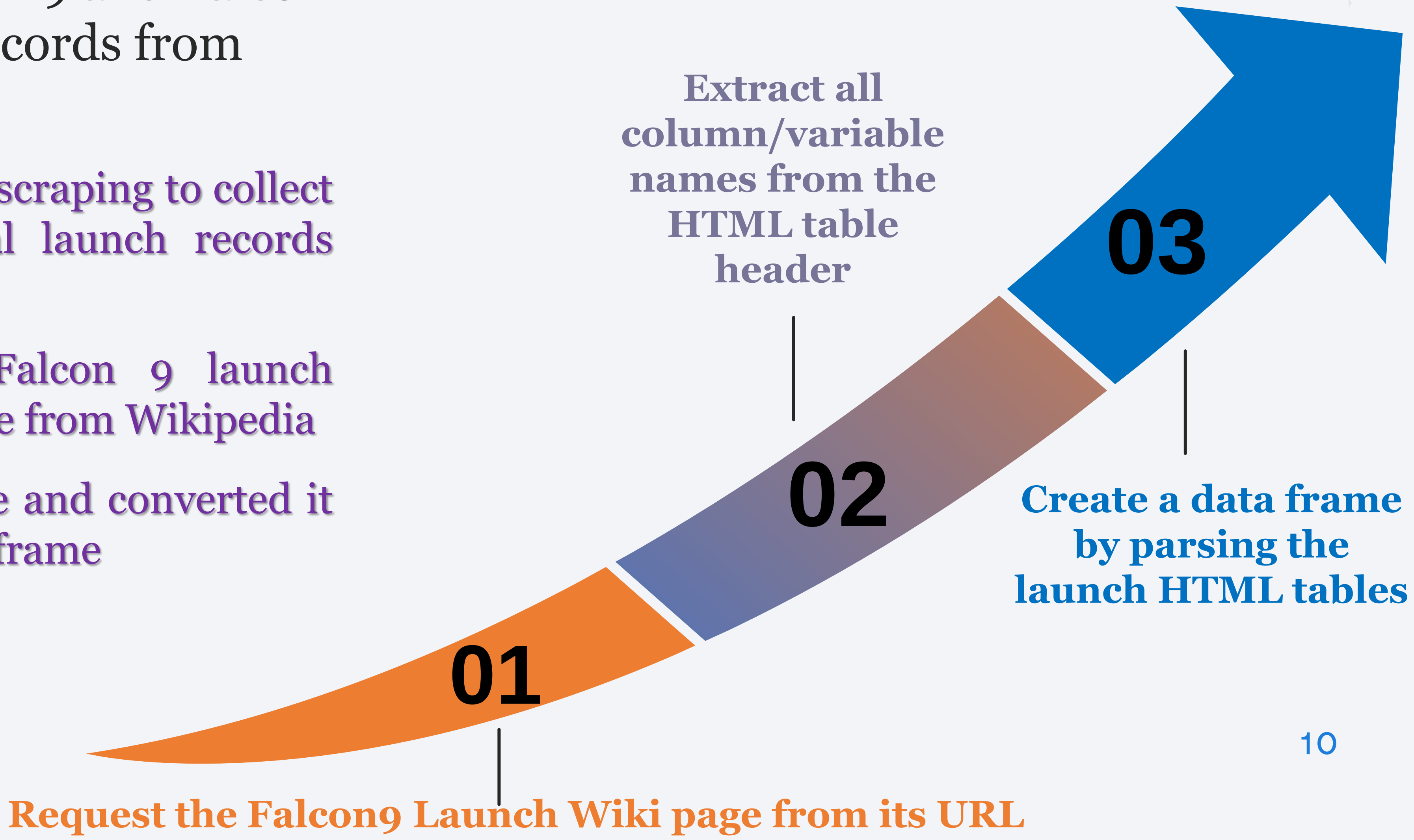
- we requested rocket launch data from SpaceX API and Checked the content of the response, after that we tried to discover some more relevant information for this project.



Source Code
Click here

Data Collection - Scraping

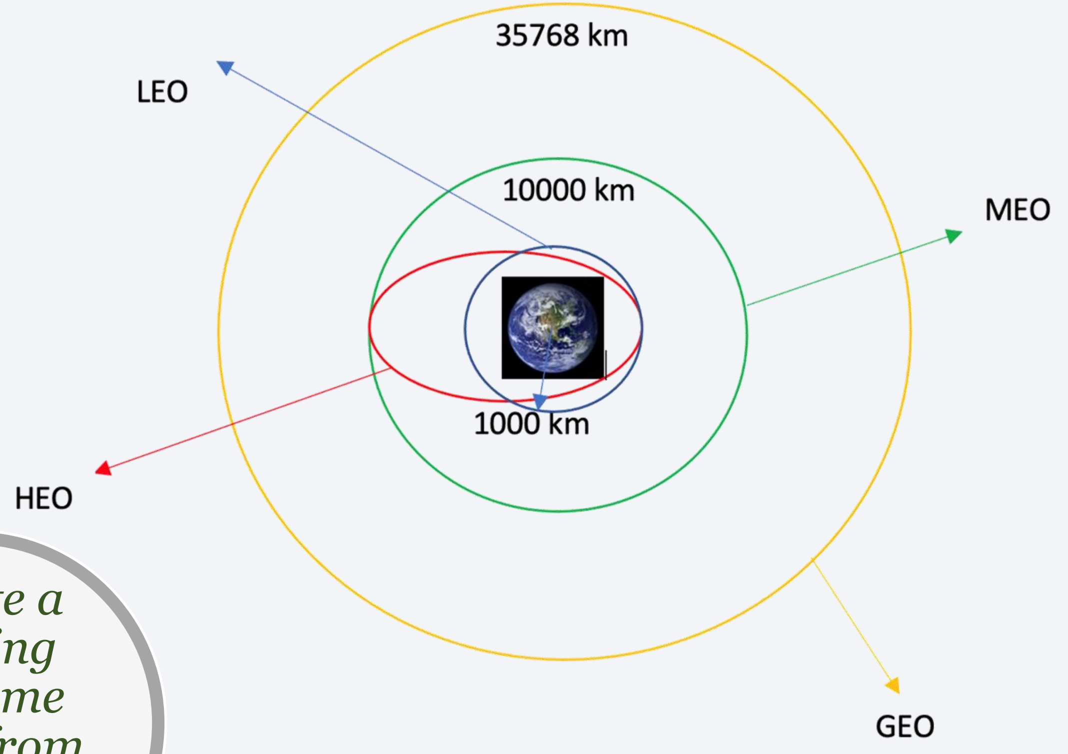
- Web scraping Falcon 9 and Falcon Heavy Launches Records from Wikipedia
  - ✓ we performed web scraping to collect Falcon 9 historical launch records from a Wikipedia
  - ✓ we Extracted a Falcon 9 launch records HTML table from Wikipedia
  - ✓ we parsed the table and converted it into a Pandas data frame



Click to Source code

Data Wrangling

some common orbit types are shown in the following plot



In this part, we performed some Exploratory Data Analysis (EDA) to find some patterns in the data and determined what would be the label for training supervised models.

1. *Calculate the number of launches on each site*
2. *Calculate the number and occurrence of each orbit*
3. *Calculate the number and occurrence of mission outcome per orbit type*
4. *Create a landing outcome label from Outcome column*

Source code: https://github.com/MeysamModiryan/Applied-Data-Science-Capstione/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

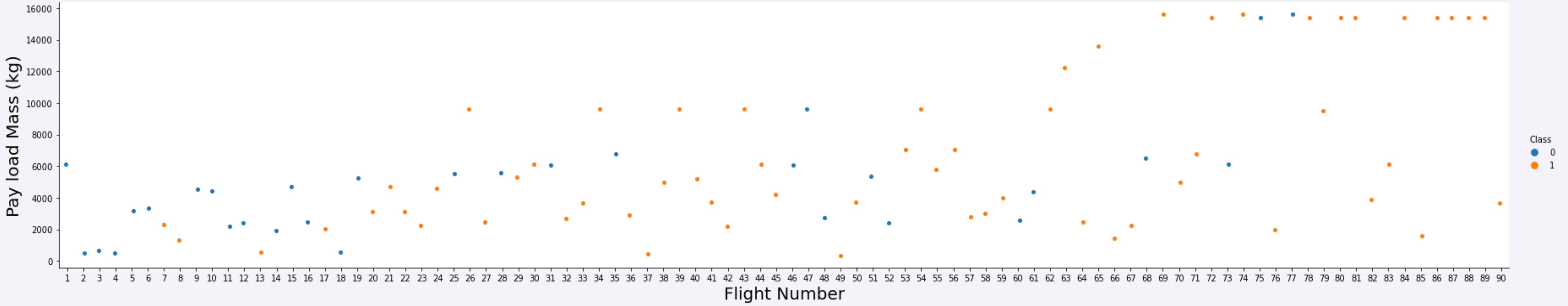
EDA with Data Visualization

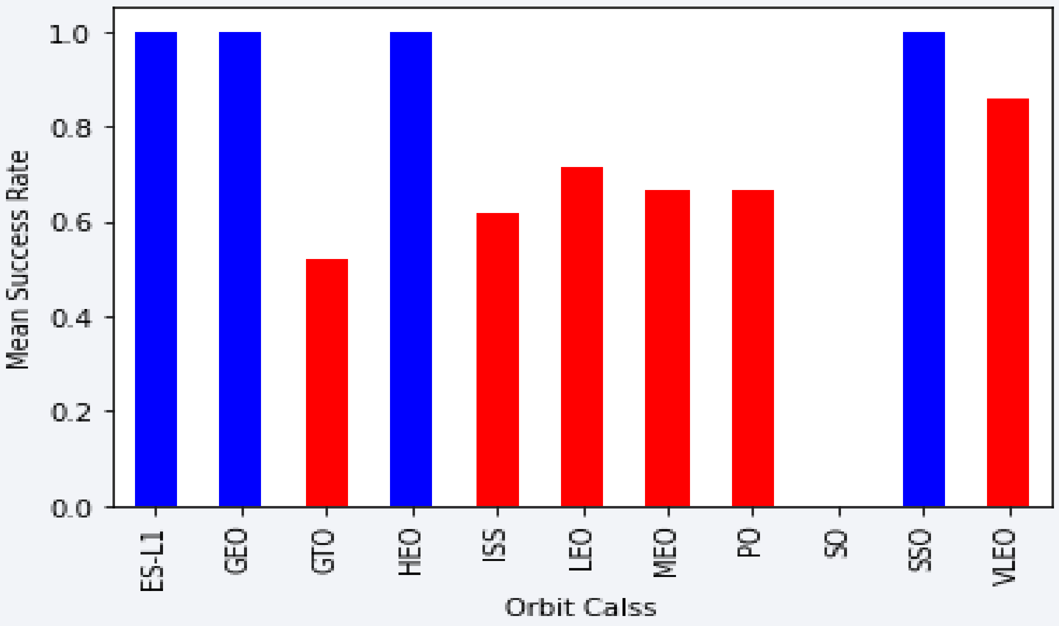
- We applied the catplot() function to some features to figure out the relation between them, Also we used bar chart and line chart to understand some relations between some features. The features that we visualized, are comen in the following:
  - ❑ Flight Number and Launch Site
  - ❑ Payload and Launch Site
  - ❑ the relationship between success rate of each orbit type
  - ❑ Flight Number and Orbit type
  - ❑ Payload and Orbit type
  - ❑ the launch success yearly trend

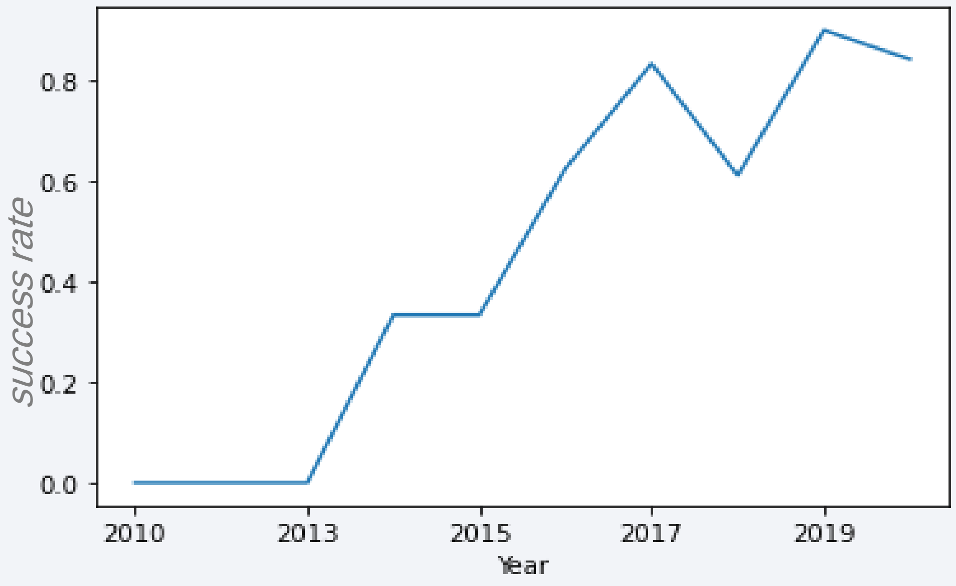
Some charts are provided in the next slide

Source code: https://github.com/MeysamModiryan/Applied-Data-Science-Capstione/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with Data Visualization







EDA with SQL

summarize the SQL queries we performed

- ***The names of the unique launch sites in the space mission***
- ***Top 5 records where launch sites begin with the string 'CCA'***
- ***The total payload mass carried by boosters launched by NASA (CRS)***
- ***Average payload mass carried by booster version F9 v1.1***
- ***the date when the first successful landing outcome in ground pad was achieved.***
- ***the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000***
- ***the total number of successful and failure mission outcomes***
- ***the names of the booster versions which have carried the maximum payload mass. Use a subquery***
- ***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***

Source code: https://github.com/MeysamModiryan/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera.ipynb

Build an Interactive Map with Folium

- In this part, we have performed more interactive visual analytics using Folium to complete the below tasks. By these, we should be able to find some geographical patterns about launch sites.
    - Mark all launch sites on a map
    - Mark the success/failed launches for each site on the map
    - Calculate the distances between a launch site to its proximities

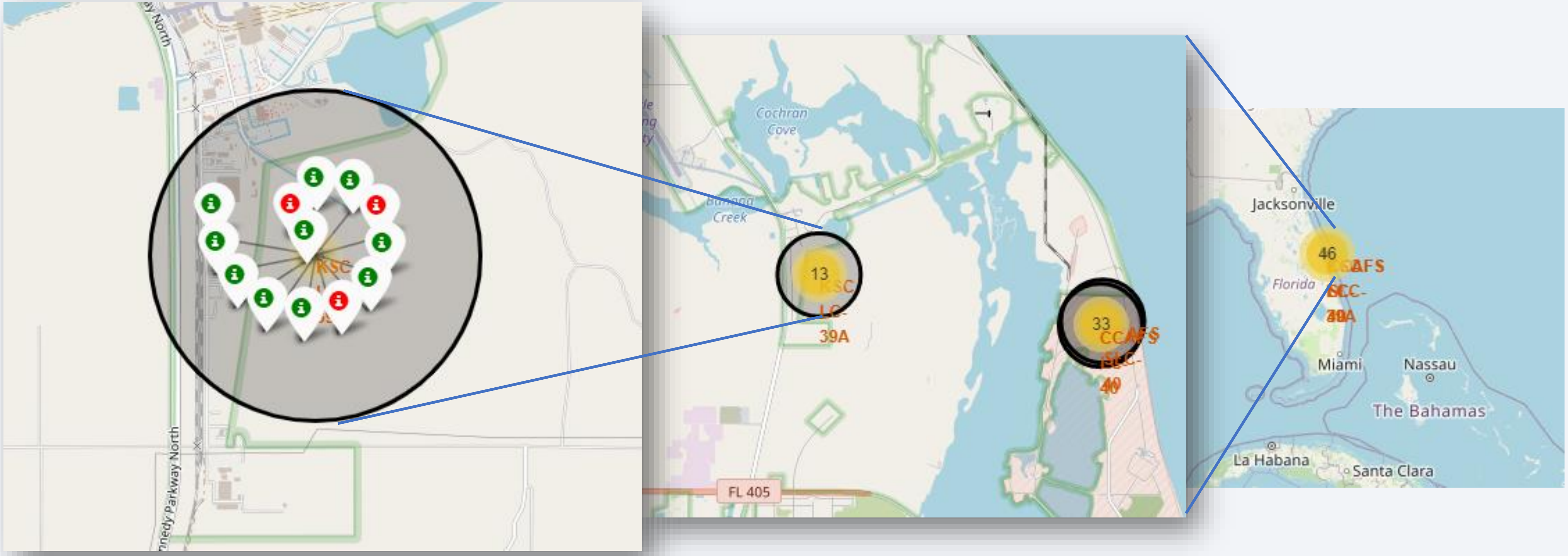
we used some map objects in order to above tasks that they were:

- ❑ Folium.Map(): To initial center location to be NASA Johnson Space Center at Houston, Texas.
- ❑ Folium.Circle(): To add a highlighted circle area with a text label on a specific coordinate
- ❑ Folium.map.Marker(): To Create a simple stock Leaflet marker on the map
- ❑ MarkerCluster(): to simplify a map containing many markers having the same coordinate.
- ❑ Folium.Marker(): To add objects to marker cluster. Also we used this to show the distance.
- ❑ Folium.PolyLine(): To draw a line between a launch site to the selected coastline point.
- ❑ Folium.Popup(): To Create a Popup instance that can be linked to a Layer.
- ❑ And also we used the add_child() function to add objects that were created to the map.

Source code: https://github.com/MeysamModiryan/Applied-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

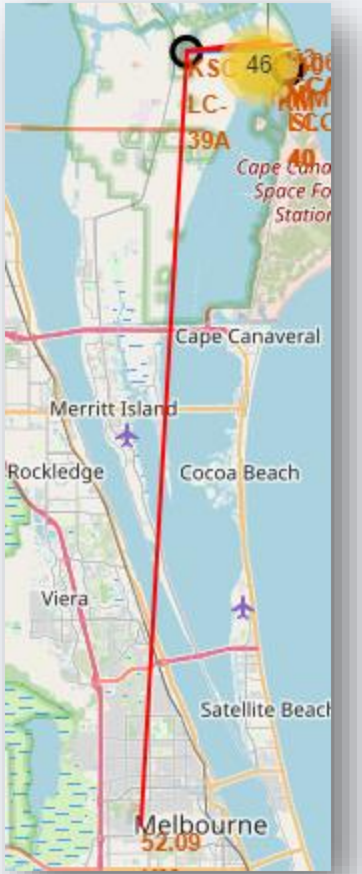
Build an Interactive Map with Folium

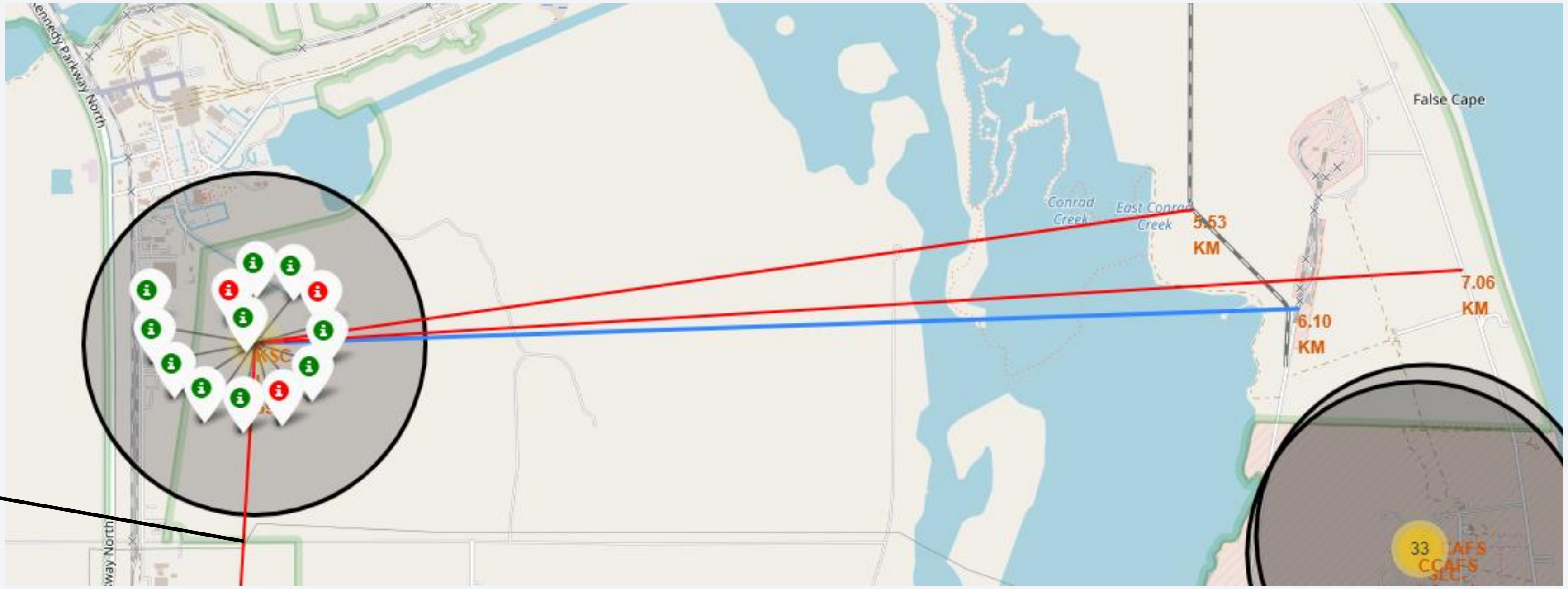
- Some of the plots that we created are provided the below:



Build an Interactive Map with Folium

- Some of the plots that we created are provided the below:





Line and distance between a launch site to its closest city, railway, highway, etc.

Build a Dashboard with Plotly Dash

✓ We used:

- ❑ Pie chart: *To show the total percentage success launches for sites*
- ❑ Scatter Plot: *To show the relation between Payload mass vs class*

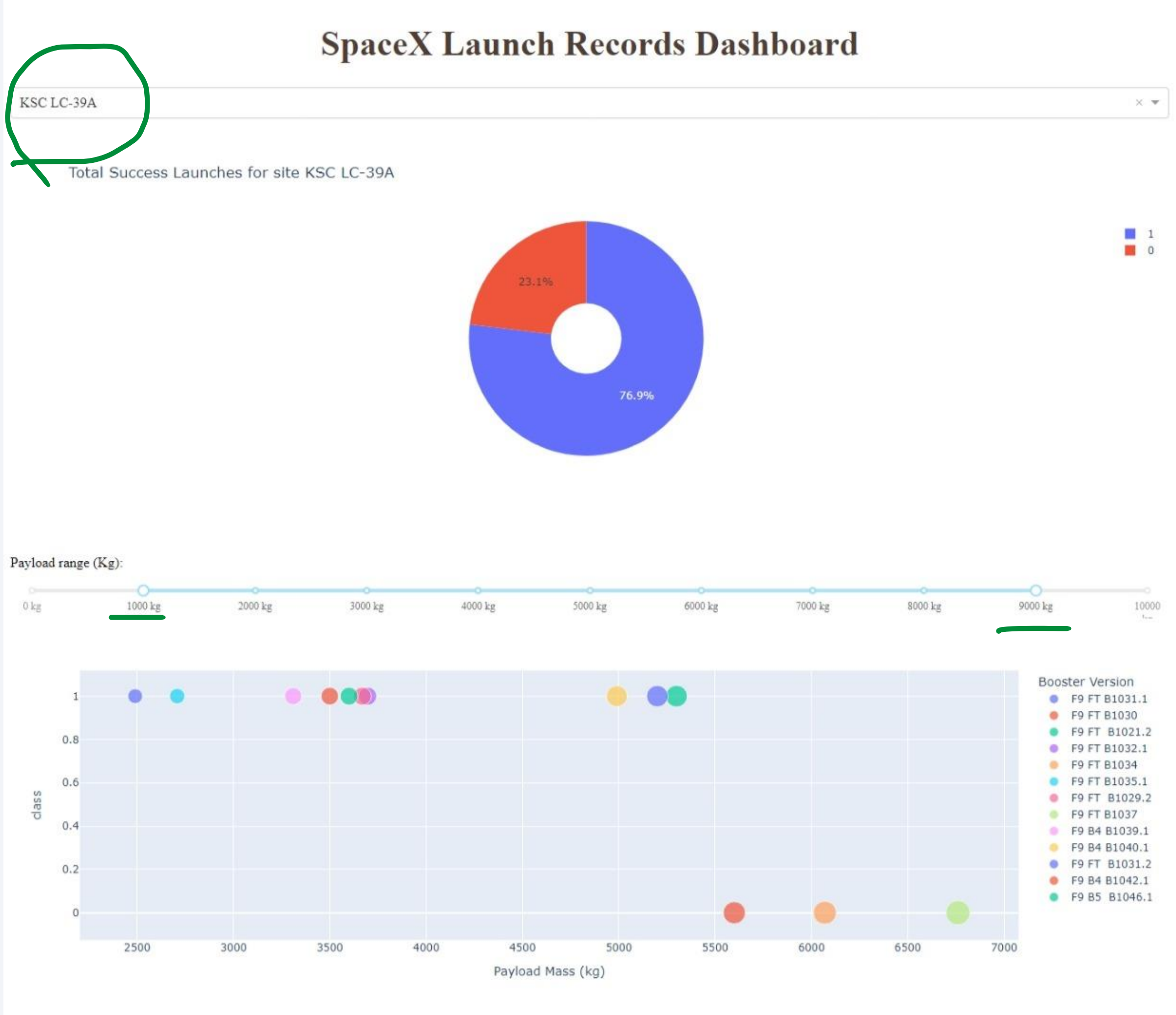
✓ Also we used:

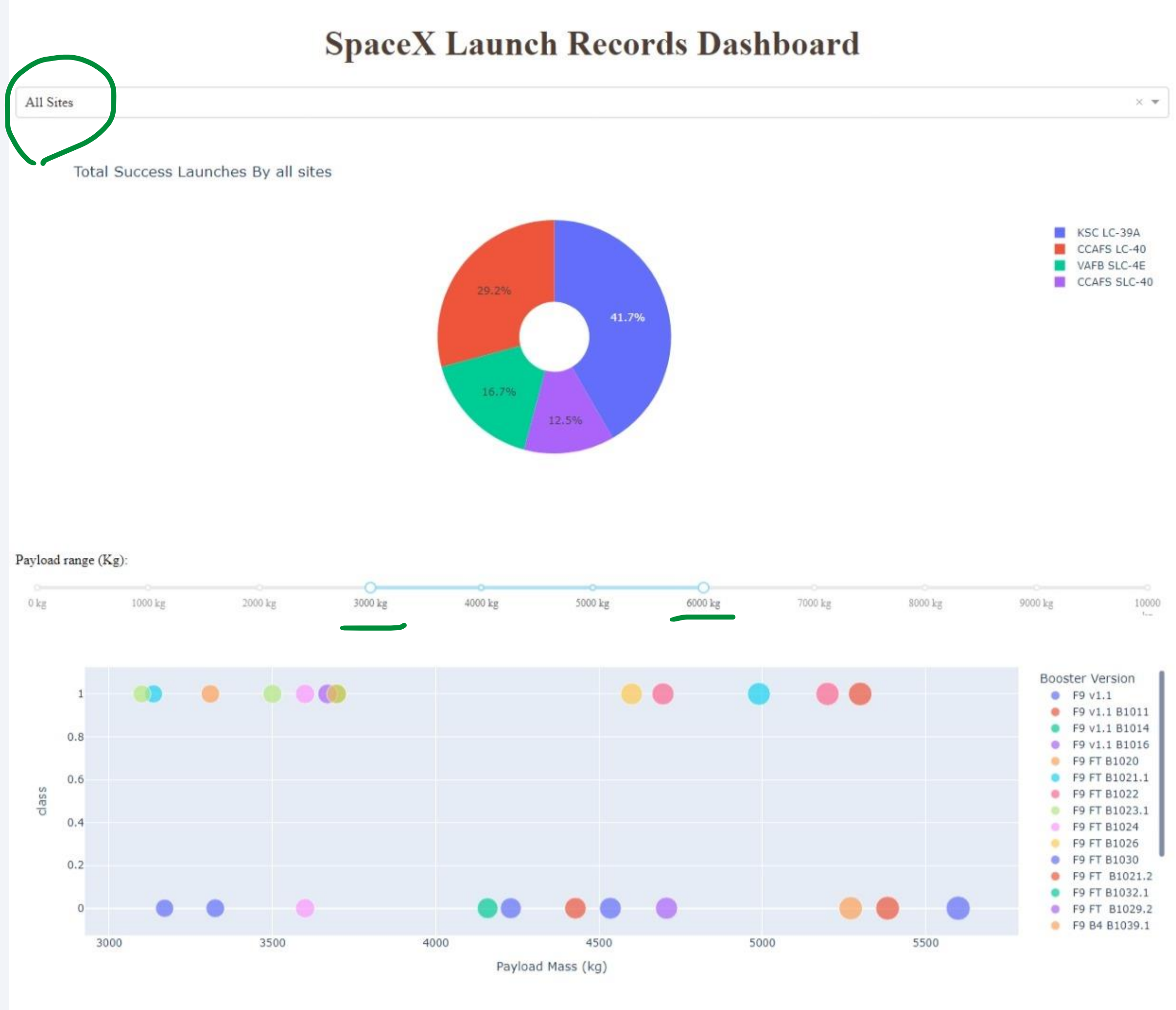
- ❑ Dropdown: *To filter or select a launch site on pie chart.*
- ❑ Rangeslider: *To filter and change the range of Payload on Scatter plot.*

This combination allowed to quickly analyze the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads.

Source code: *https://github.com/MeysamModiryan/Applied-Data-Science-Capstione/blob/main/spacex_dash_app1.py*

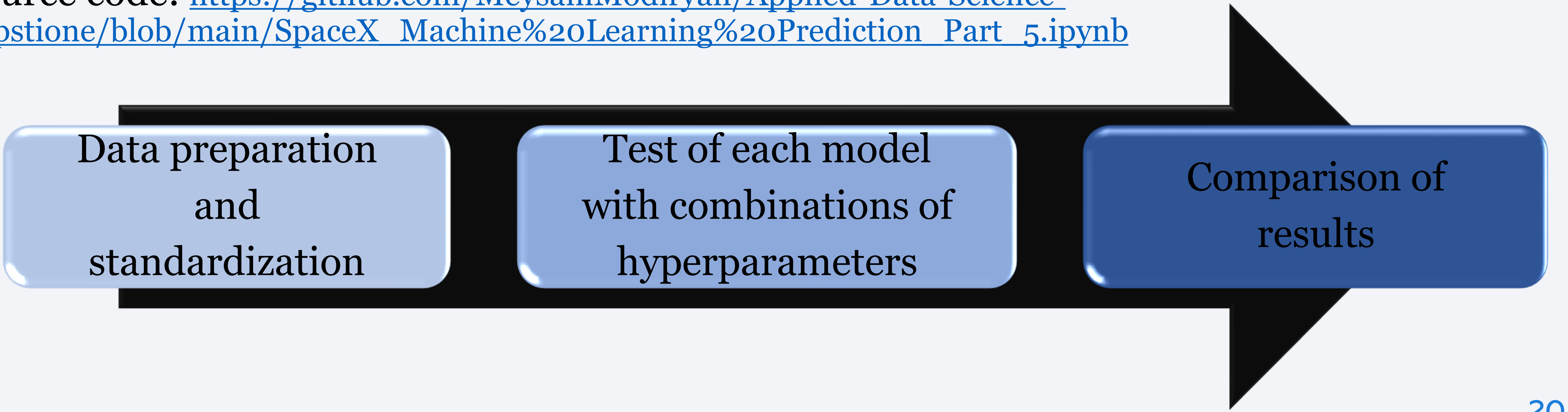
Build a Dashboard with Plotly Dash





Predictive Analysis (Classification)

- After loading the dataframe, we created a NumPy array from the column class as the label and standardized X (features that were selected from previous) and split them into training and test data. We trained some classification models like **LOGISTIC REGRESSION, SUPPORT VECTOR MACHINE, DECISION TREE**, and **K_NEAREST NEIGHBORS**, and also we used **GridysearchCV** in order to find the best model parameters. after calculating the accuracy of each model (train and test set accuracy) we found the method performed best.
- Source code: https://github.com/MeysamModiryan/Applied-Data-Science-Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

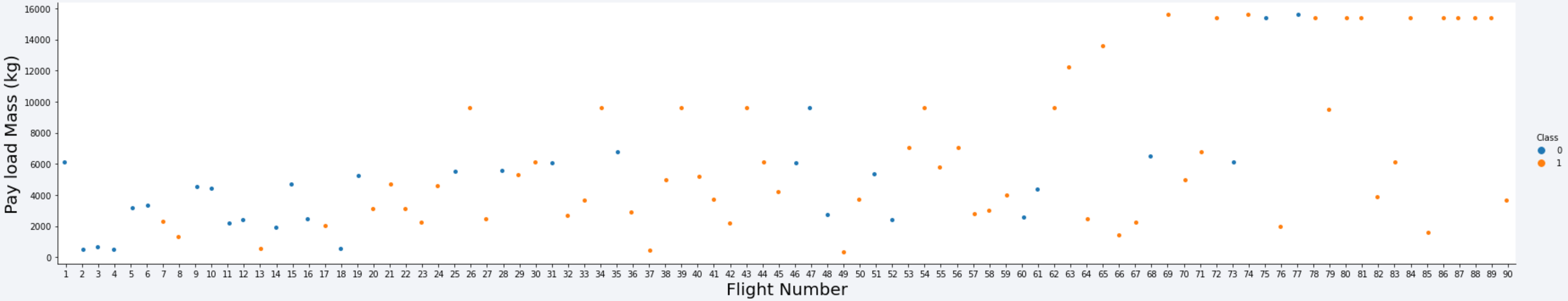


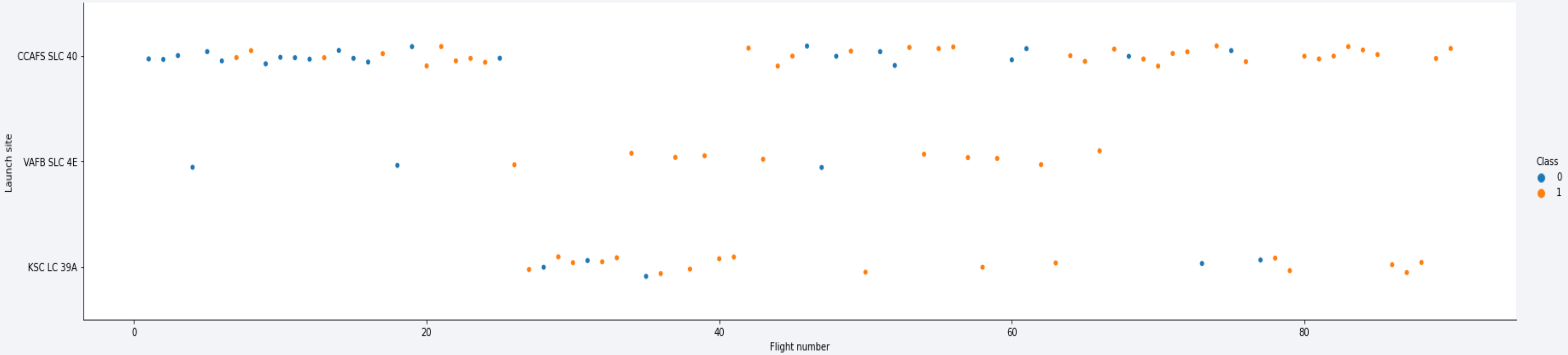
Results

- Exploratory data analysis results
  - Space X uses 4 different launch sites; (CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E)
  - The first launches were done to Space X itself and NASA;
  - average payload mass carried by booster version F9 v1.1 is 2928 Kg.
  - 2015-12-22 was the first successful landing outcome in-ground pad was achieved
  - Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average;
  - Almost 100% of mission outcomes were successful;
  - Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015;
  - The number of landing outcomes became as better as years passed.

Results

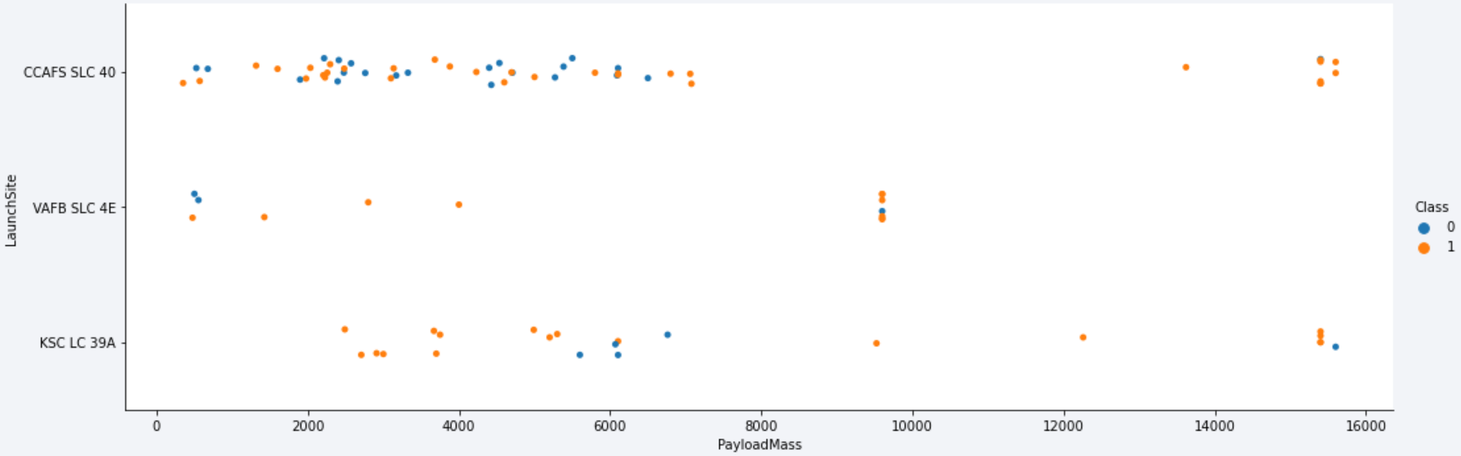
- Interactive analytics demo in screenshots (EDA with visualization)

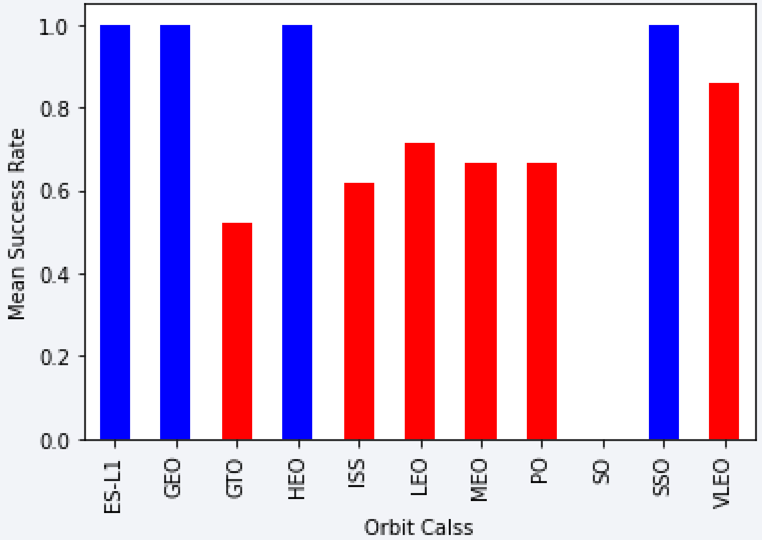


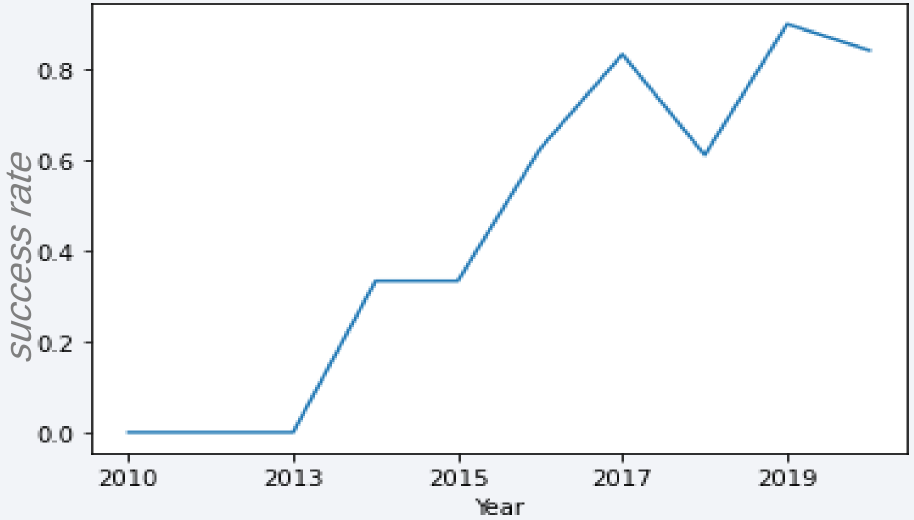


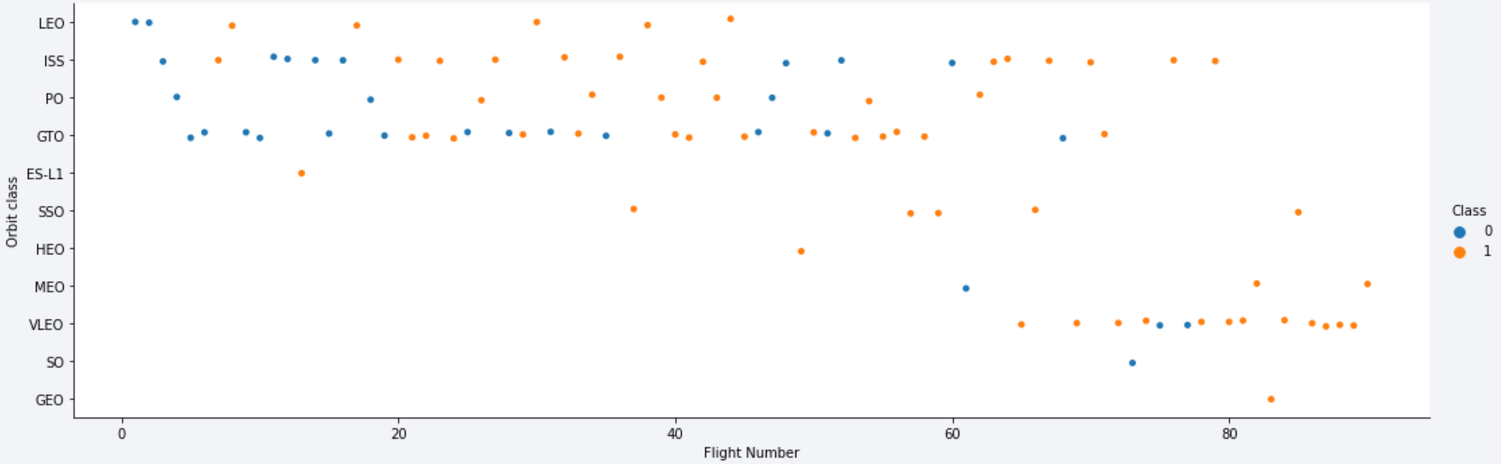
Results

- Interactive analytics demo in screenshots (EDA with visualization)









Results

- Interactive analytics demo in screenshots (EDA with SQL)

the names of the booster versions which have carried the maximum payload mass

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

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

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

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

the total payload mass carried by boosters launched by NASA (CRS)

total_payload
111268

average payload mass carried by booster version F9 v1.1

average_payload
2928

the date when the first successful landing outcome in ground pad was achieved.

first_successful
2015-12-22

the total number of successful and failure mission outcomes

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

the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

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

the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

Results

- Predictive analysis results

The Decision Tree classifier was the best model to predict a successful landing and had 89 percent accuracy for the training set and 84 percent for the test set.

```
SCORES = {'LOGISTIC REGRESSION MODEL':logreg_cv.best_score_,
          'SUPPORT VECTORE MACHINE MODEL': svm_cv.best_score_,
          'DECISION TREE MODEL':tree_cv.best_score_,
          'K_NEAREST NEIGHBORS MODEL':knn_cv.best_score_}
Best_Modle = max(SCORES, key=SCORES.get)
Best_Score = round(max(SCORES.values()), 2)*100
print('Best Model is: {} with %{} accuray for train model'.format(Best_Modle, Best_Score))
```

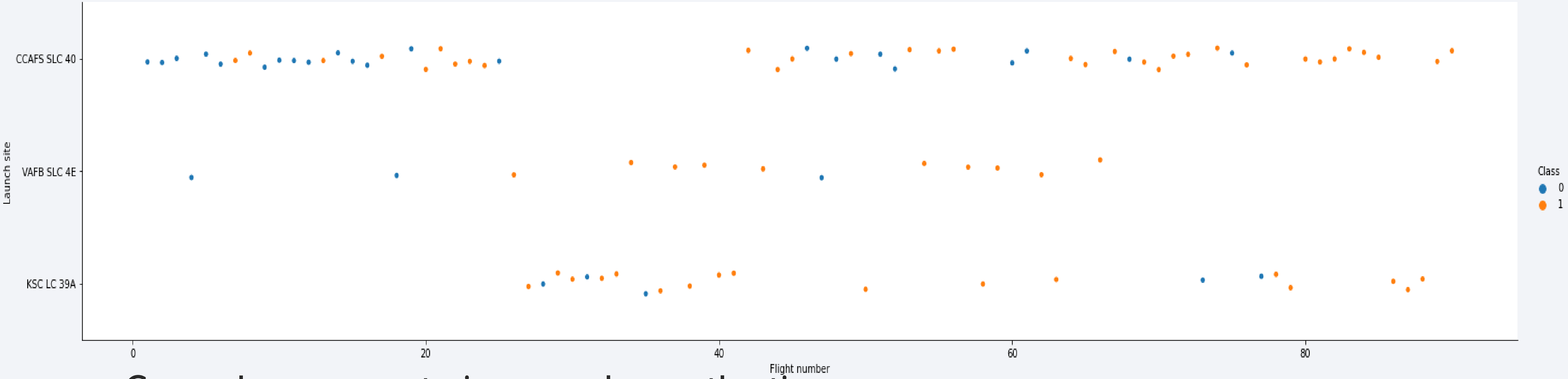
```
Best Model is: DECISION TREE MODEL with %89.0 accuray for train model
```

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

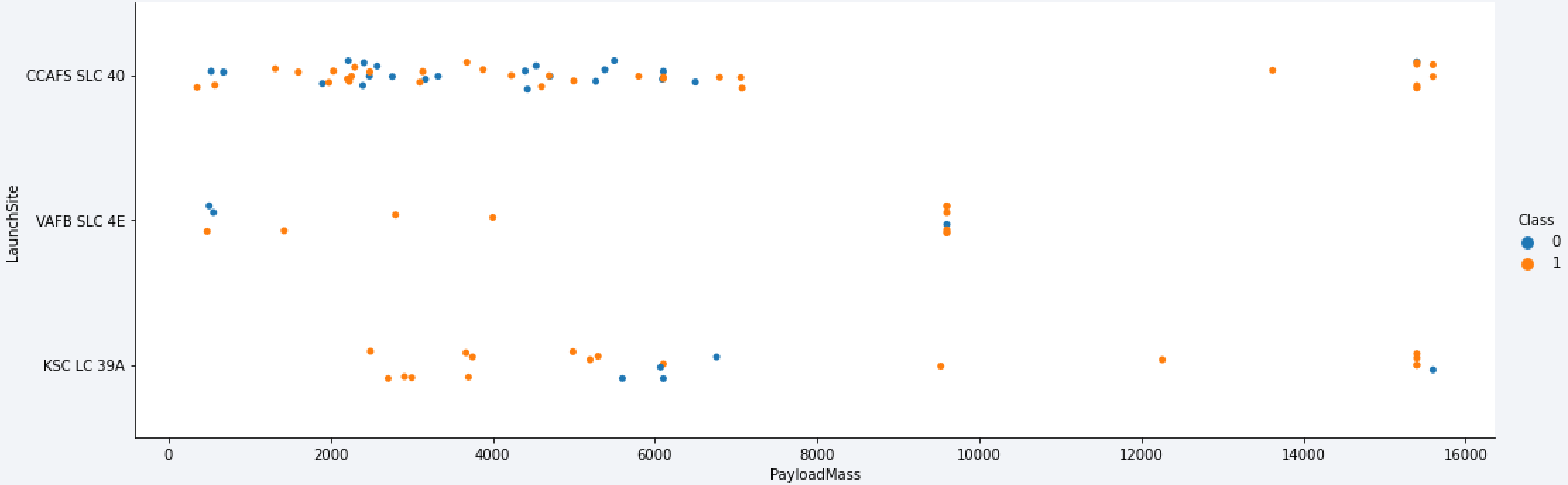
- scatter plot of Flight Number vs. Launch Site



- General success rate improved over the time;
- According to the plot, the best launch site is CCAF5-SLC40 for nowadays;
- Also, KSC LC 39A has a good success.

Payload vs. Launch Site

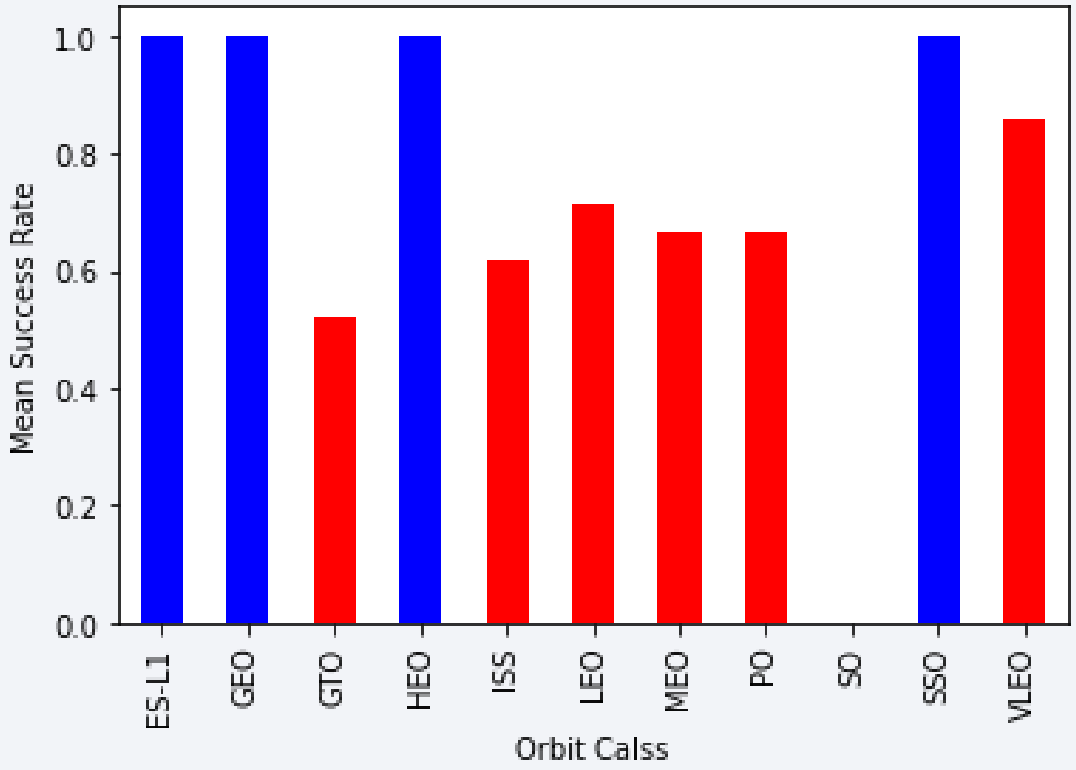
- *Scatter plot of Payload vs. Launch Site*



- *Payloads over 8000kg have a excellent success rate;*
- *KSC-LC39A site for payloads less than 5500kg has a good success rate;*
- *it seems for WAFB-SLC4E site, payloads over 10000 are impossible.*

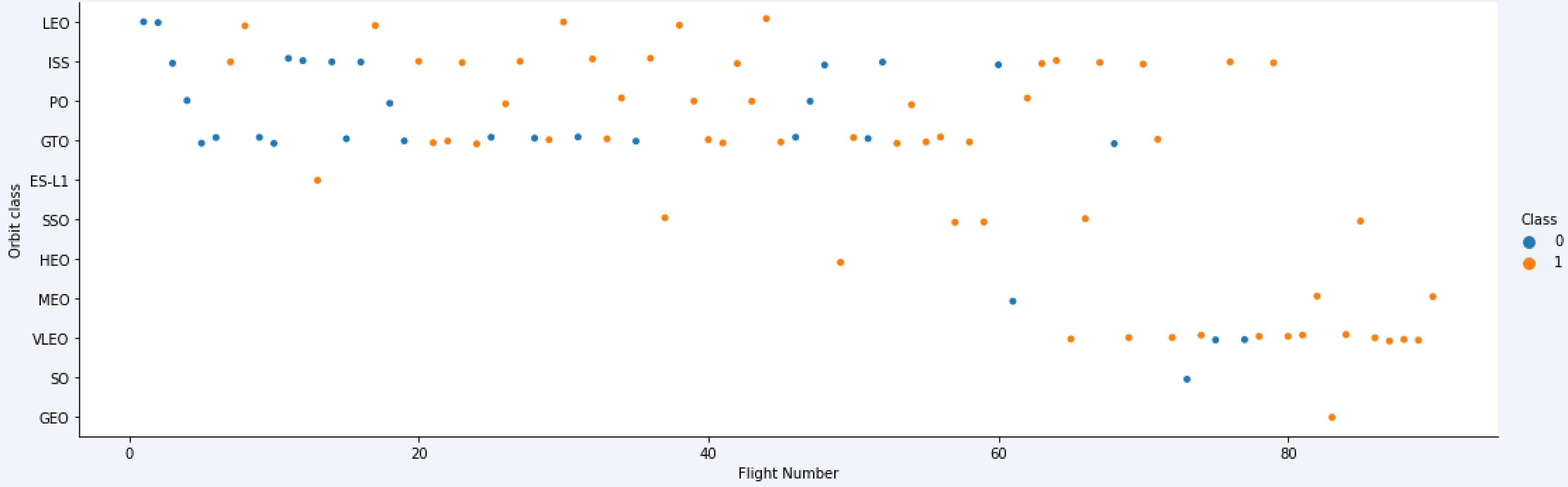
Success Rate vs. Orbit Type

- Bar chart for the success rate of each orbit type
- ES-L1, GEO, HEO, AND SSO have the highest mean success rate and they are shown as blue bar charts
- GTO has a lowest mean success rate between all orbits calls, about 50 percent.
- ISS, LEO, MEO, AND PO have almost the same function. About 60 percent mean success rate;
- VLEO class is not bad and its mean success rate is about 80 percent.
- the classes with the lowest success rates are shown in red color in the chart.



Flight Number vs. Orbit Type

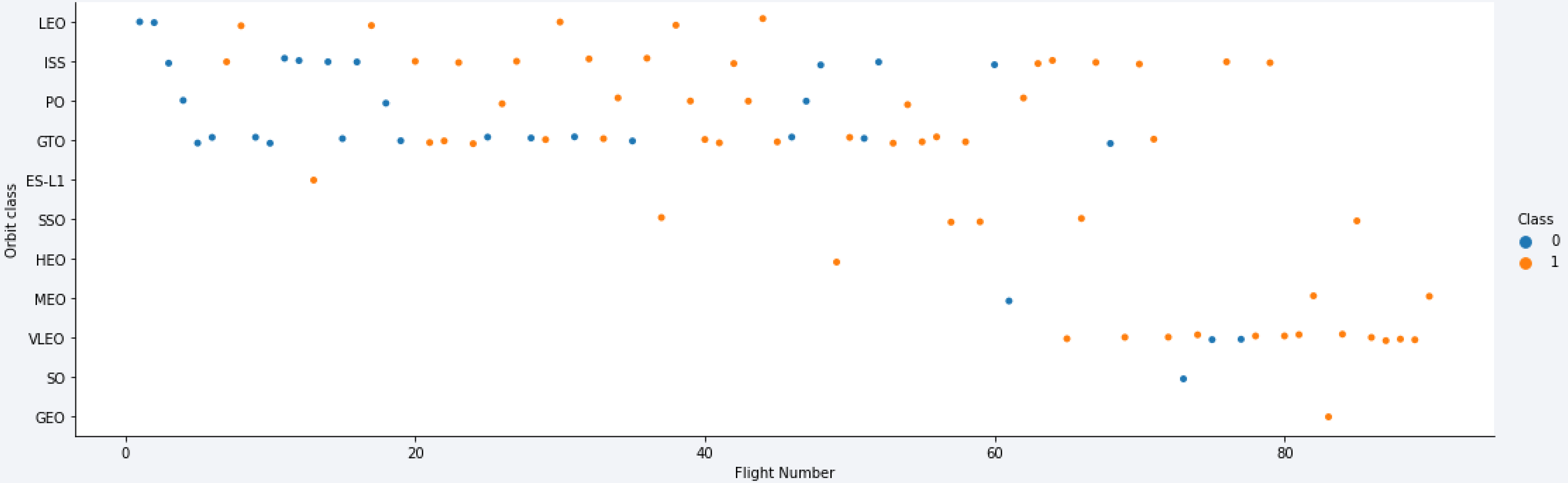
- Scatter point of Flight number vs. Orbit type



- ❑ **Apparently, success rate improved over time to all orbits;**
- ❑ **VLEO orbit seems a new business opportunity, due to recent increase of its frequency.**

Payload vs. Orbit Type

Scatter point of payload vs. orbit type

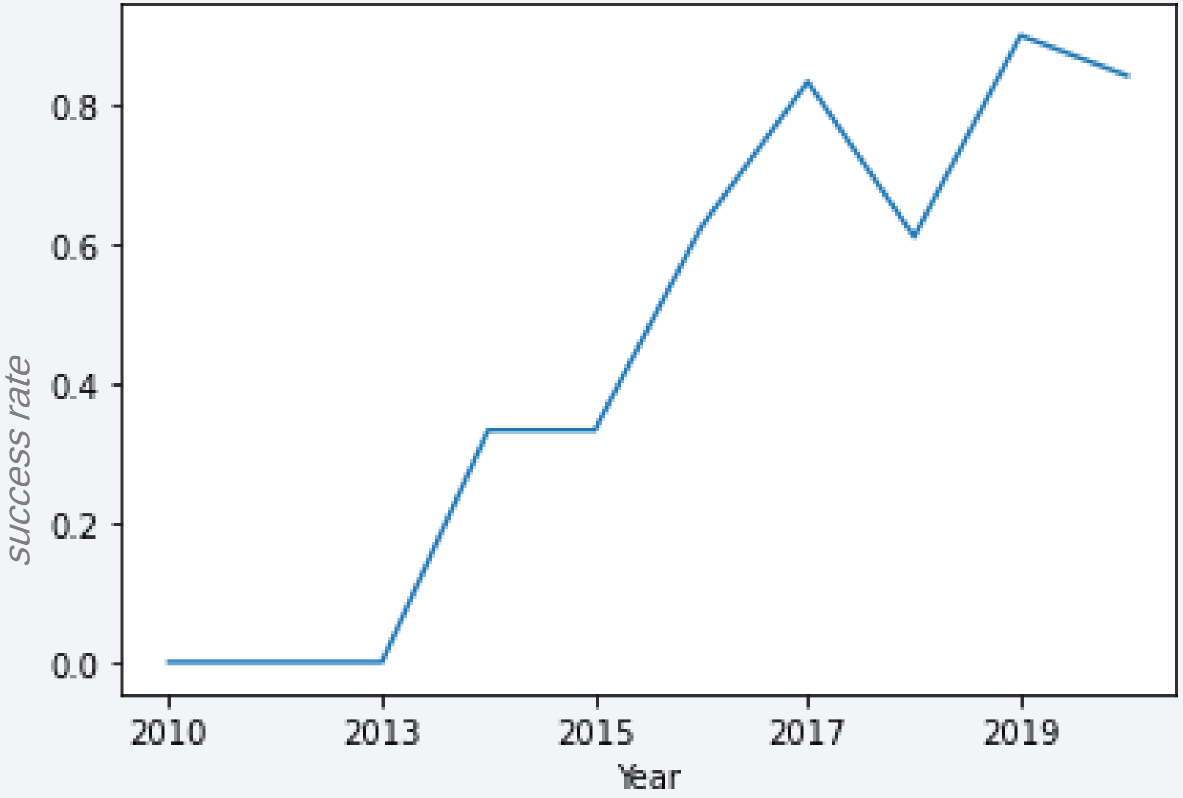


- it seems that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight numbers when in GTO orbit.
- There are few launches to the orbits SO and GEO.

Launch Success Yearly Trend

line chart of yearly average success rate

- we can observe that the success rate since 2013 kept increasing till 2020, although it had a decreasing trend between 2017 and 2018, generally, it was ascending



All Launch Site Names

According to data, there are four launch sites

They are obtained by selecting unique occurrences of "launch_site" values from the dataset.

Launch site
CAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Launch Site Names Begin with 'CCA'

- The 5 records where launch sites begin with `CCA`
- Here we can see five samples of Cape Canaveral launches.

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

- the total payload carried by boosters from NASA

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE PAYLOAD LIKE '%CRS%';
```

```
 * ibm_db_sa://fvj37188:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

total_payload
111268

Total payload calculated above, by summing all payloads whose codes contain 'CRS', which corresponds to NASA.

Average Payload Mass by F9 v1.1

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

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

```
 * ibm_db_sa://fvj37188:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

average_payload
2928

Filtering data by the booster version above and calculating the average payload mass we obtained the value of 2,928 kg.

First Successful Ground Landing Date

- the dates of the first successful landing outcome on ground pad

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

```
 * ibm_db_sa://fvj37188:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

first_successful
2015-12-22

2015-12-22 was the first successful landing outcome in-ground pad was achieved

Successful Drone Ship Landing with Payload between 4000 and 6000

The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
%%sql
SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE LANDING__OUTCOME= 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

```
 * ibm_db_sa://fvj37188:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

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

Selecting distinct booster versions according to the filters above, these 4 are the result.

Total Number of Successful and Failure Mission Outcomes

The total number of successful and failure mission outcomes

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

```
 * ibm_db_sa://fvj37188:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

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

Grouping mission outcomes and counting records for each group led us to the summary above.

Boosters Carried Maximum Payload

The names of the booster which have carried the maximum payload mass

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

```
 * ibm_db_sa://fvj37188:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

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

These are the boosters which have carried the maximum payload mass registered in the dataset.

2015 Launch Records

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

```
%%sql
SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Failure (drone ship)'
AND DATE_PART('YEAR', DATE) = 2015
```

```
 * ibm_db_sa://fvj37188:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c
Done.
```

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

The list has the only two occurrences.

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

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

```
 * ibm_db_sa://fvj37188:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.
Done.
```

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

This view of data alerts us that "No attempt" must be taken in account.

Section 3

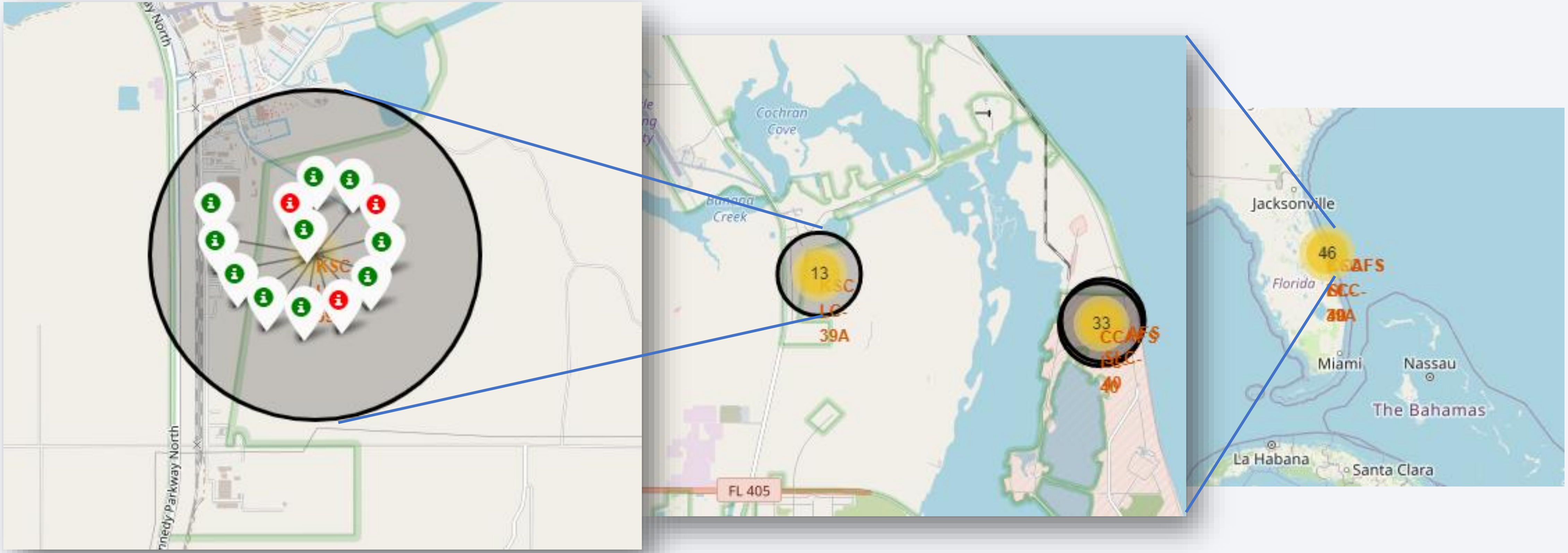
Launch Sites Proximities Analysis

< All launch sites on a map >

we can see the locations of each site on the map and also they are so near to the sea and it seems for human safety located there.

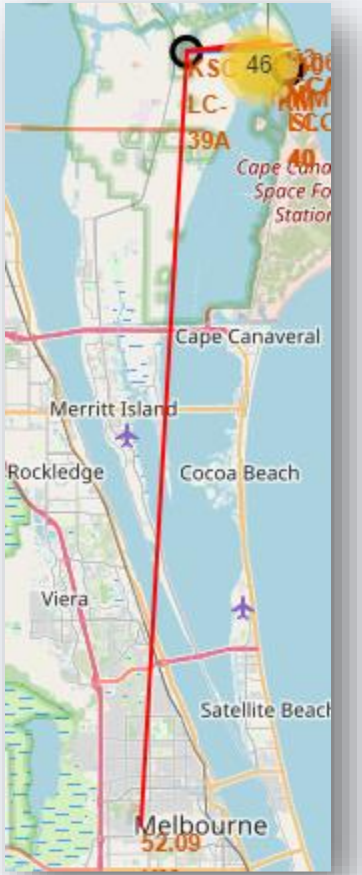
< Launch Outcomes by Site >

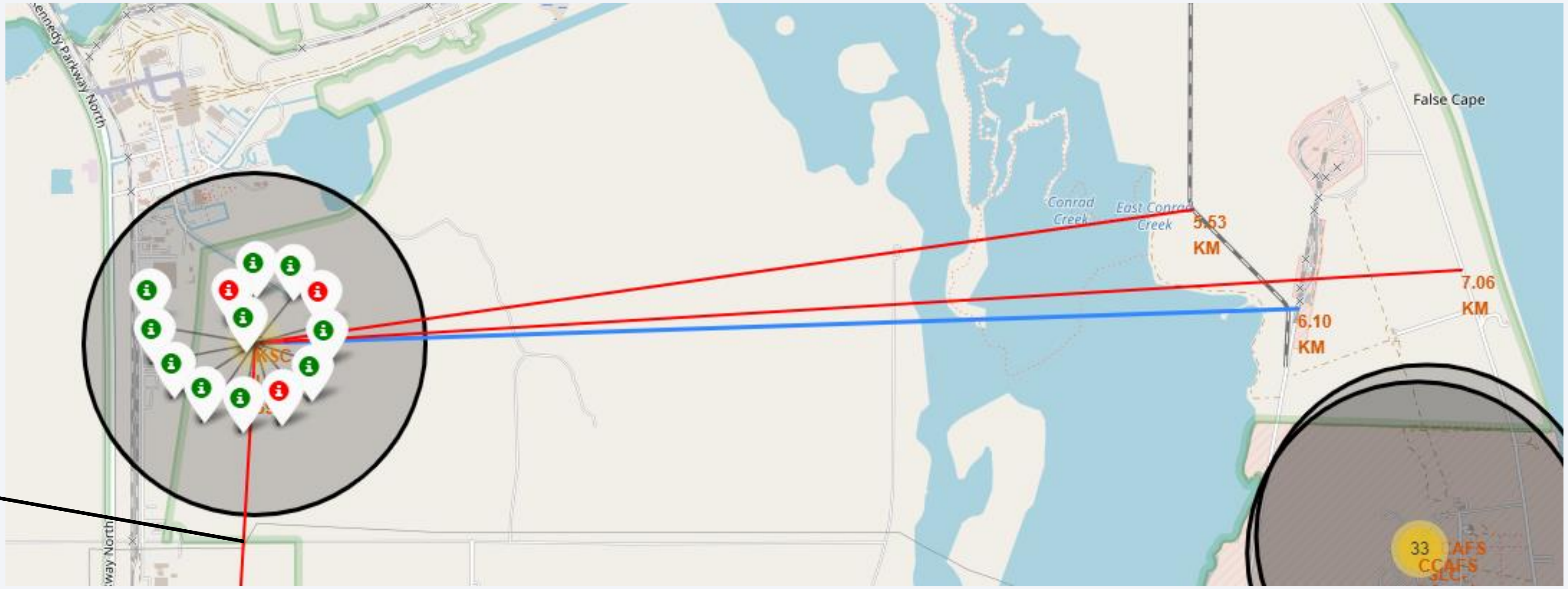
Example of KSC LC-39A launch site launch outcomes



<Logistics and Safety>

Launch site KSC LC-39A has good logistics aspects, being near railroad and road and relatively far from inhabited areas.





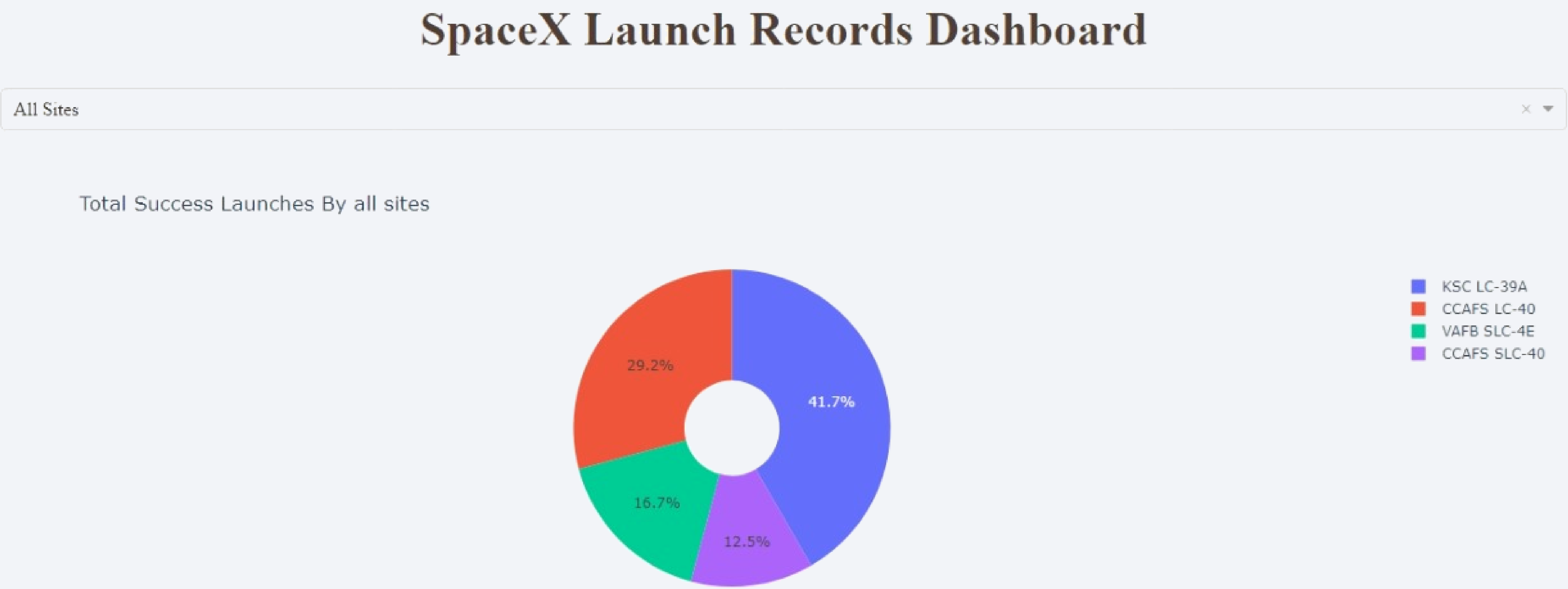
Line and distance between a launch site to its closest city, railway, highway, etc.

Section 4

Build a Dashboard with Plotly Dash

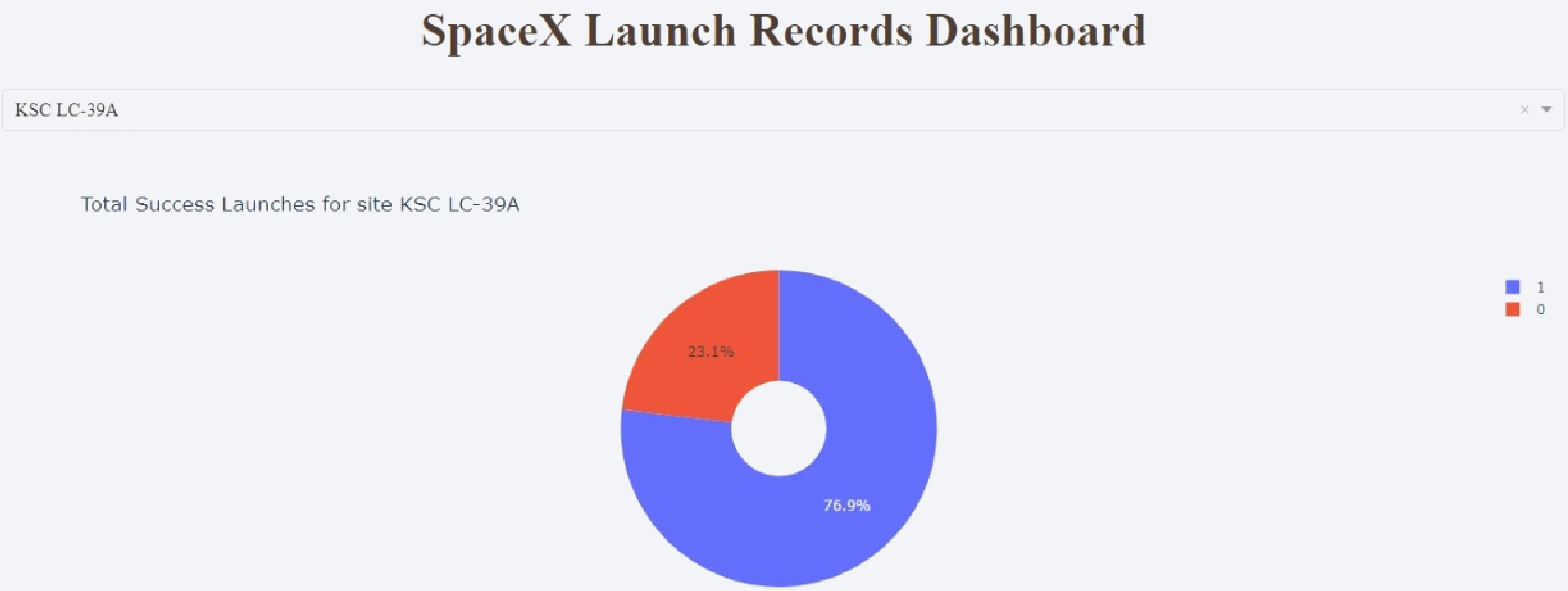
<Successful Launches by Site>

The place from where launches are done seems to be a very important factor of success of missions.



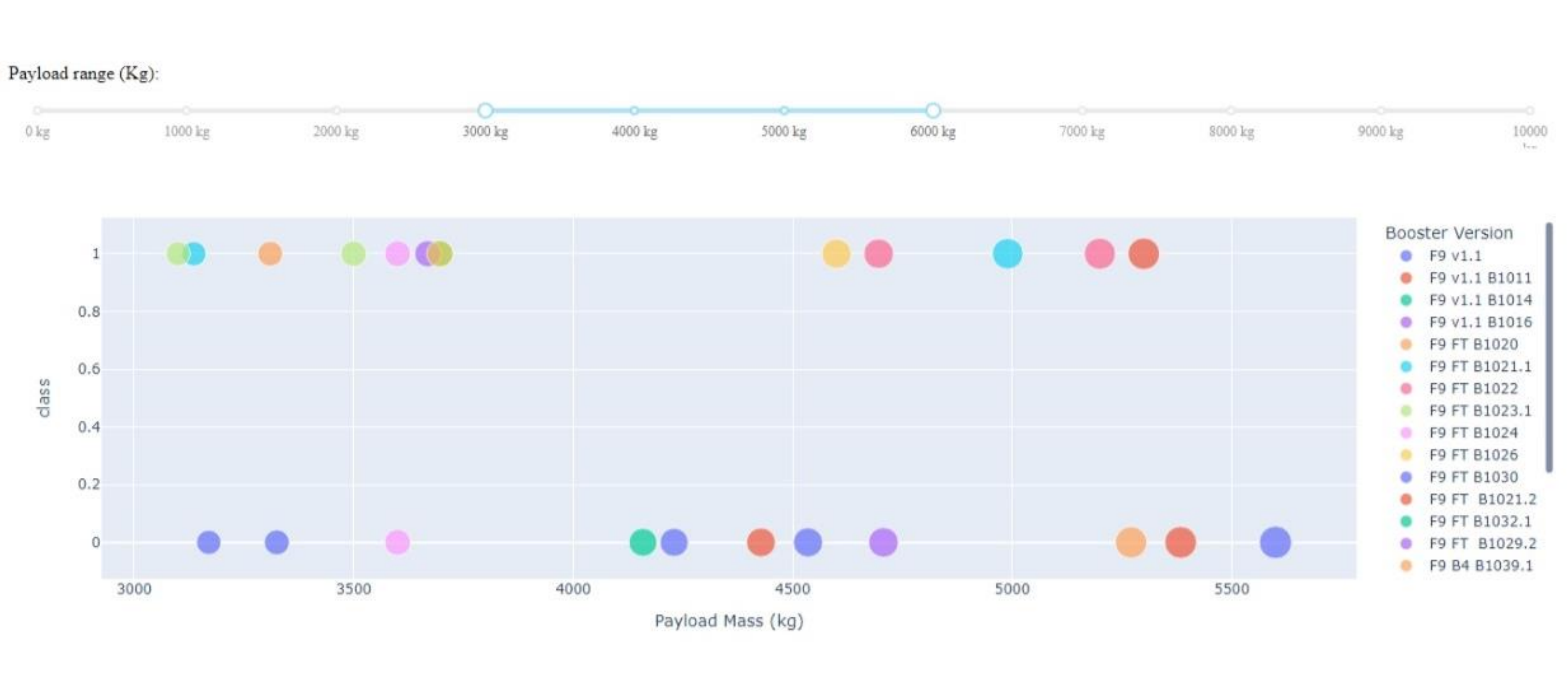
<The launch site with highest launch success ratio>

KSC-LC-39A has a highest success rate (about %77) between all launch sites.



<Payload vs. Launch Outcome scatter plot for all sites>

Payloads under 5500Kg and FT boosters are the most successful combination.

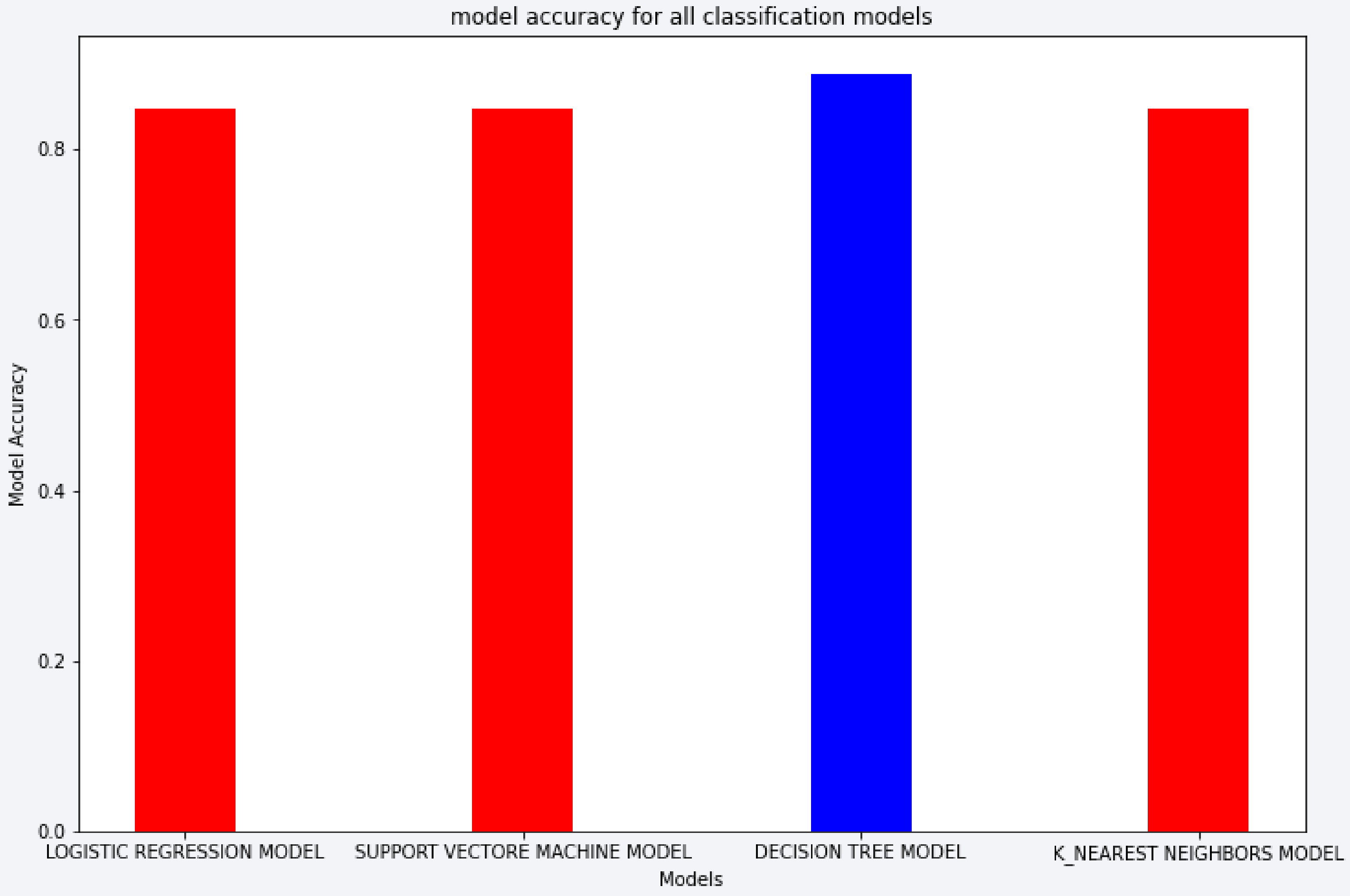


Section 5

Predictive Analysis (Classification)

Classification Accuracy

- All Classification models were trained and tested and their accuracy are plotted;
- The model with the highest classification accuracy is Decision Tree Classifier, which has accuracies over than 87%.



Confusion Matrix

- **Decision tree model has a good accuracy on test model and it has a excellent precision and recall.**
- **Precision** = 5 / (5+0) = 1
- **Recall** = 5 / (5+1) = 0/833
- **F1Score** = 0/907

Conclusions

- ❑Different data sources were analyzed, refining conclusions along the process;
- ❑We can see that KSC LC-39A had the most successful launches from all the sites
- ❑Orbit GEO, HEO, SSO, and ES-L1 have the best Success Rate
- ❑The success rates for SpaceX launches is directly proportional time in years they will eventually perfect the launches
- ❑For this data set Decision-Tree model has a better accuracy than other models

Appendix

- ***All Appendix that we used in this project are provided on GitHub, and its link is provided below. Unfortunately, Folium didn’t show maps on GitHub and also Machine Learning Prediction isn’t run on GitHub too.***
- ***https://github.com/MeysamModiryan/Applied-Data-Science-Capstione.git***

Thank you!