Winning Space Race with Data Science

Grzegorz Kosek
4.12.2021

Outline

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

Executive Summary

- Summary of methodologies
  - The aim of this analysy was to find if Falcon 9 will land sucesfully. If it's possible than there is a chance to save money due to the possibility of re-using rocket.
  - In order to predict the sucesfull land I analysed data about previous launches.
  - I collected the data, performed its wrangling, used multiple visualization tools and created predictive analysy in order to establish what conditions would be best for a sucessful rocket landign.
- Summary of all results
  - Thanks to this analysis I found out which site should the company use to launch the rocket, what payload is the best and which booster is most reliable.

Introduction

- Project background and context
  - The purpuse of this project is to calculate the cost of rocket launch based on predictions whether the first stage will land sucesfully or not. We, as the Space Y, want to create a decent offer and be competitive against Space X.
  - If the first stage will land sucesfully, we will be able to re-use in in the next launch and, therefore, save big amount of money.

- Problems you want to find answers
  - What are the best conditions and features to perform a sucesfull landing of rocket's first stage?

Section 1

Methodology

Methodology

Executive Summary

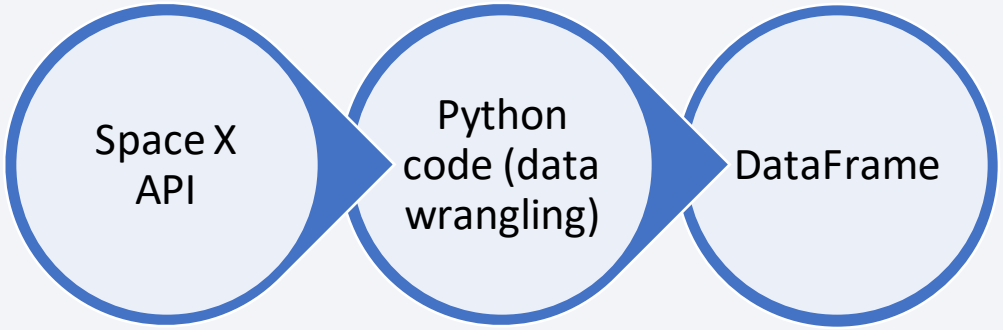
- Data collection methodology:
  - I used webscrapping method togheter with Space X API (https://api.spacexdata.com/)
- Perform data wrangling
  - Data was processed mostly by using libraries like Pandas and NumPy
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - I've created four classification models – Logistic Regression, Support Vector Machine, Classification Tree and K-Nearest Neighbors. I have also used Scikit-learn's models GridSearchCV in order to chose the best parameters for above models. All models have reached a quite sufficient accuracy.

Data Collection

- Describe how data sets were collected.

- Data were collected using Space X API 9 (https://api.spacexdata.com/) and defined functions which create the final dataframe.

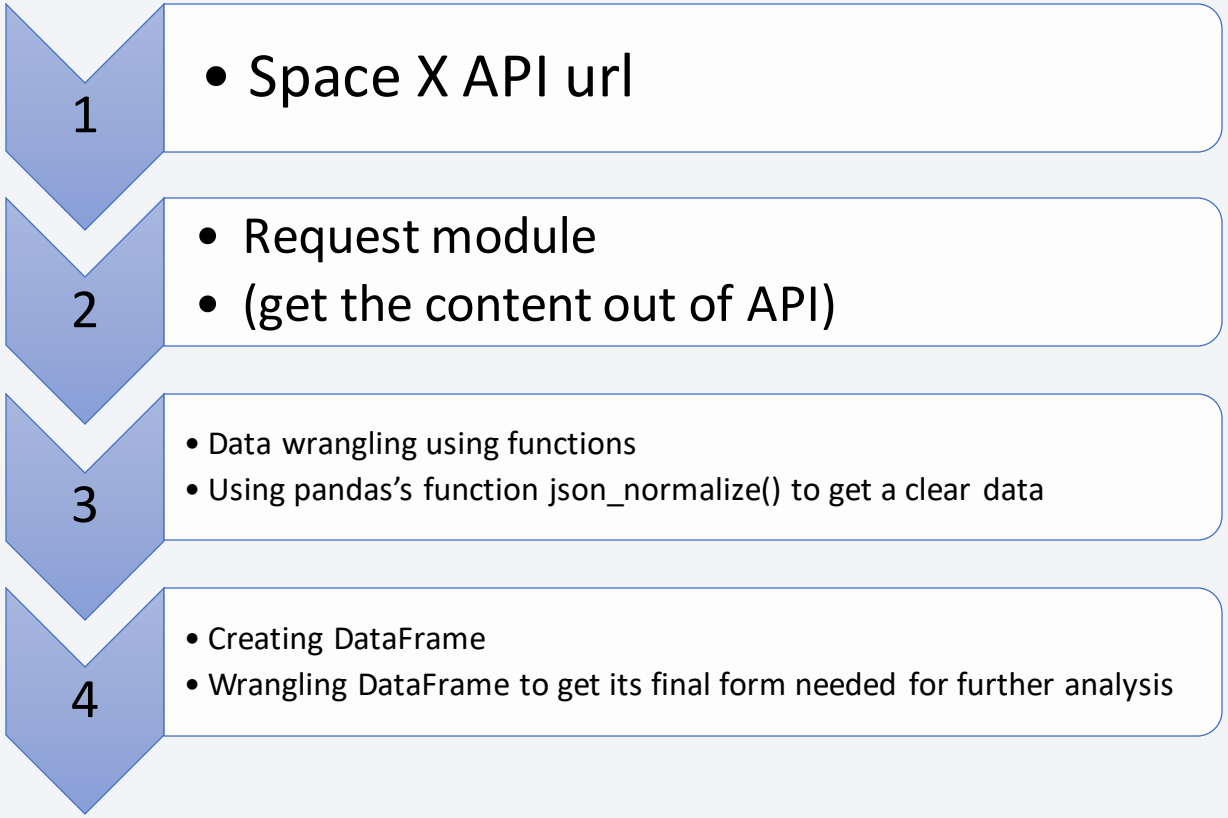
- You need to present your data collection process use key phrases and flowcharts



Data Collection – SpaceX API

Present your data collection with SpaceX REST calls using key phrases and flowcharts

- GitHub link to my notebook:
  - https://github.com/kosek-g/IBM_Data_Science_Final_Project/blob/main/Notebooks/w1-jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts

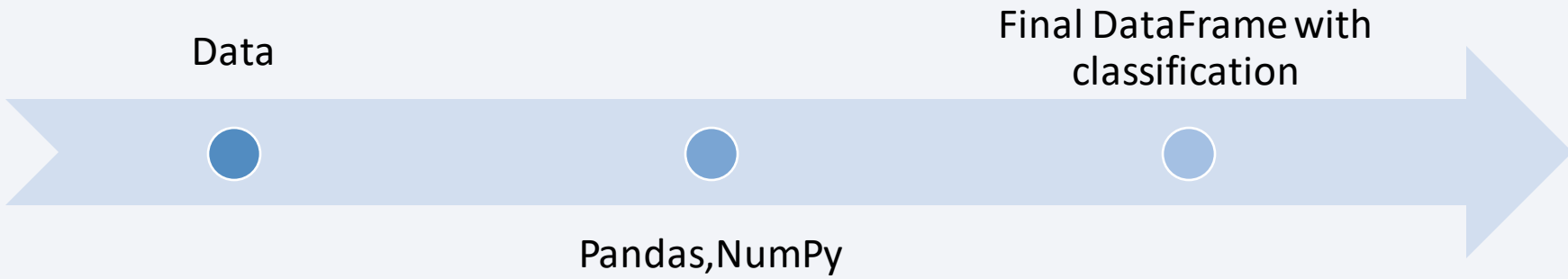
- GitHub link to my notebook:
  - https://github.com/kosek-g/IBM_Data_Science_Final_Project/blob/main/Notebooks/w1-jupyter-labs-webscraping.ipynb

1. Wikipedia url
2. Request module
   (get the content out of wiki)
3. Extracting Falcon 9's launches information using BeautifulSoup library
4. Creating DataFrame out of scraped information

Data Wrangling

- Describe how data were processed

- Data has been wrangled using Pandas and NumPy. Main goal was to create „Class” column showhing which landing was a success. They were clasified using „0” and „1”.
- GitHub link to my notebook:
  - https://github.com/kosek-g/IBM_Data_Science_Final_Project/blob/main/Notebooks/w1-labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts

- In that part of analysis, I have used few kinds of charts to describe data we are dealing with.
  - I have used scatter plots to show relations betwen:
    - Flight number and Palyload mass
    - Flight number and Launch site
    - Payload mass and Launch site
    - Flight number and Orbit type
    - Payload and Orbit type
  - Bar chart to show success rate for every orbit
  - Line plot to observe success rate over the years

GitHub link to my notebook: https://github.com/kosek-g/IBM_Data_Science_Final_Project/blob/main/Notebooks/w2-jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed

- Performed SQL queries
  - Unique launch sites names
  - 5 launch sites begining with name ‚CCA'
  - Total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - First successful landing
  - List of boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - Total number of successful and failure missions
  - List of booster versions which have carried the maximum payload mass
  - List of failed landing in drone ship with their booster versions, and launch site names in year 2015
  - Count of landing outcomes between 2010-06-04 and 2017-03-20 (desc order)

GitHub link to my notebook: https://github.com/kosek-g/IBM_Data_Science_Final_Project/blob/main/Notebooks/w2-jupyter-labs-eda-sql-coursera.ipynb

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map. Explain why you added those objects

- To the Folium Map I've added objects like:
  - folium.Circle – to highligh area with where the sites are. This feature make them visible on the whole world map
  - folium.Marker with folium.Cluster - to group and show all the launch places togheter with theri count and wheater they succed or not.
  - folium.MousePosition – to get coordinates for other objects on the map
  - folium.PolyLine – to draw lines between launch site and coasline, rails, roads and other cities

GitHub link to my notebook: https://github.com/kosek-g/IBM_Data_Science_Final_Project/blob/main/Notebooks/w3-lab_jupyter_launch_sites_loc.ipynb

(unfortunately, notebook on GitHub doesn't want to display maps generated by my code. In the repo notebook folder, there is also a html file with the same name which, after downloading, should display everythign in the correct way)

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard. Explain why you added those plots and interactions

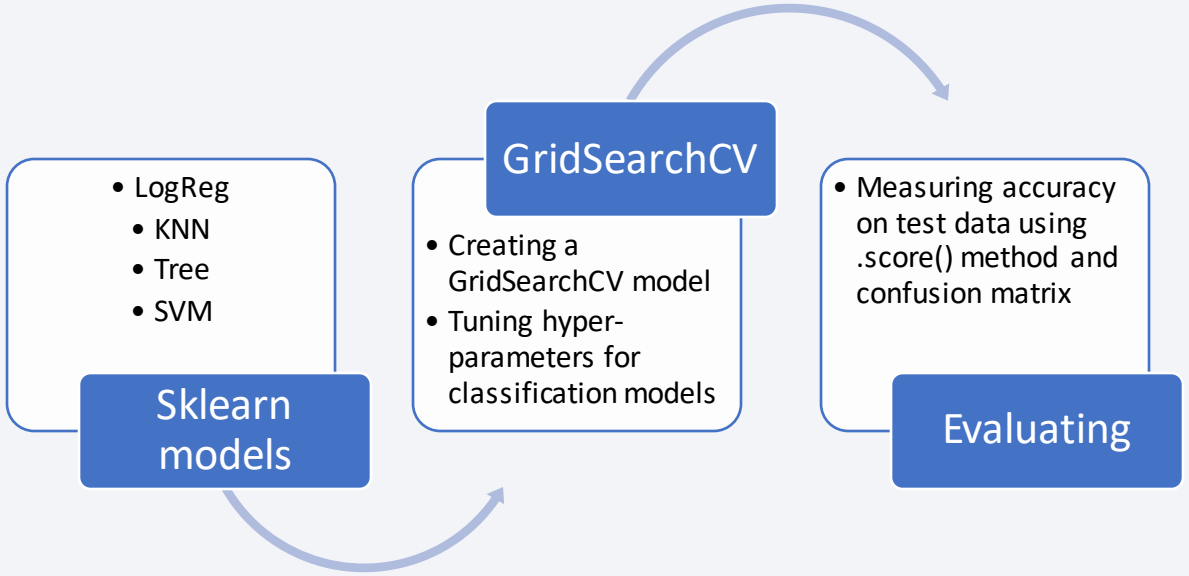
- Dashboard shows two interactive graphs. The first one is a pie chart showing the successful landing rate for sites. The second one is a scatter plot, and it shows relations between payload mass and class (0 – rocket did not land, 1 – successful landing). The booster version is also included (as a color on the plot) and of course everything can be filtered by single site using dropdown menu on the top. I've added this two graphs as they are important feauters when it comes to prediction. Payload mass might be crucial factor which can be seen on the plot.

- GitHub link to my notebook: https://github.com/kosek-g/IBM_Data_Science_Final_Project/blob/main/Notebooks/w3-plotly-dash.ipynb

Predictive Analysis (Classification)

Summarize how you built, evaluated, improved, and found the best performing classification model. You need present your model development process using key phrases and flowchart

I've created four classification models – Logistic Regression, Support Vector Machine, Classification Tree and K-Nearest Neighbors. I've builded them using models from Sci-Kit Learn. I have also used GridSearchCV in order to chose the best parameters for above models. To evaluate them, I've used .score() method on the test dataset. The result was indentical for all of them – 0,8333 even though, on train data they've reached different numbers.

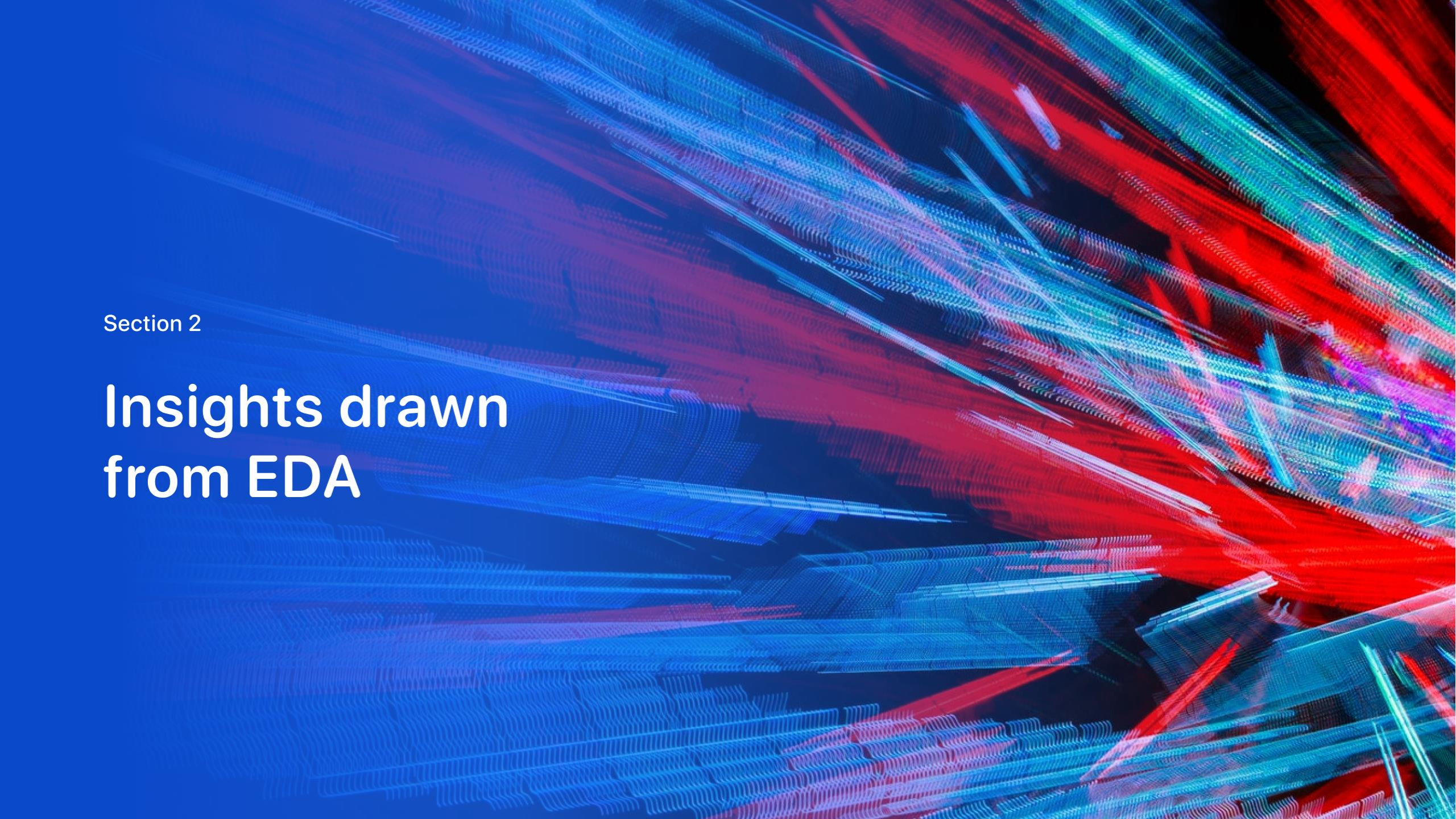


- GitHub link to my notebook: https://github.com/kosek-g/IBM_Data_Science_Final_Project/blob/main/Notebooks/w4-SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

- Everythign will be concluded in the next sections of this presentation.

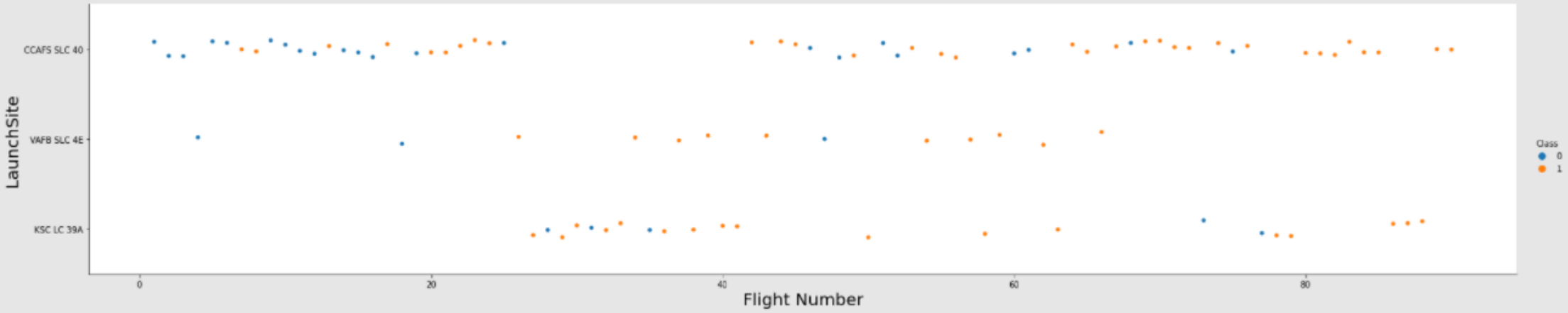


Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

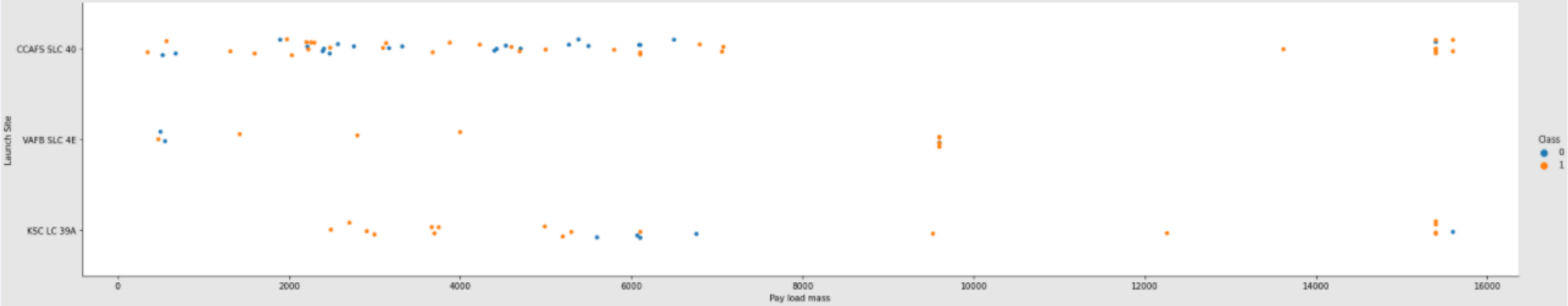
Show a scatter plot of Flight Number vs. Launch Site. Show the screenshot of the scatter plot with explanations



Scatter plot shows distribution of single launch with their locations and whether the landing was successful or not (marked with colors = blue for fail, orange for success).

Payload vs. Launch Site

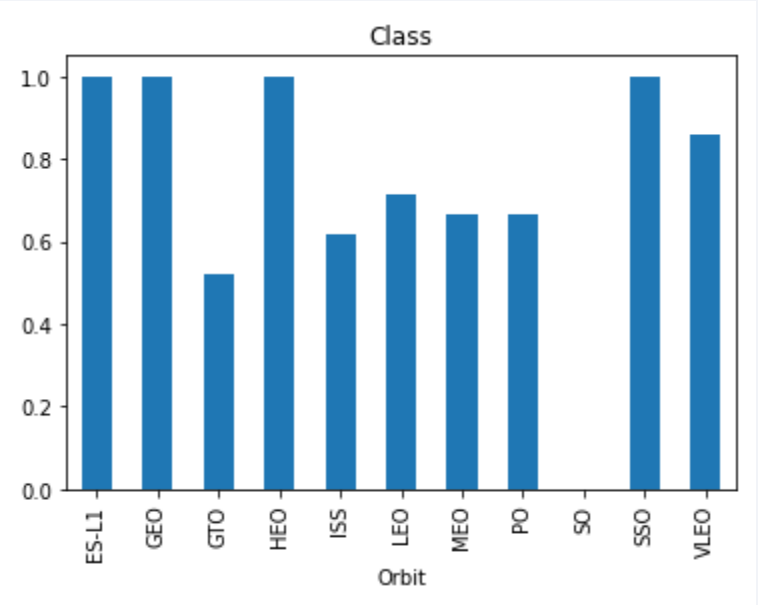
Show a scatter plot of Payload vs. Launch Site. Show the screenshot of the scatter plot with explanations



Scatter plot shows distribution of payload values with their locations and whether the landing was successful or not (marked with colors = blue for fail, orange for success).

Success Rate vs. Orbit Type

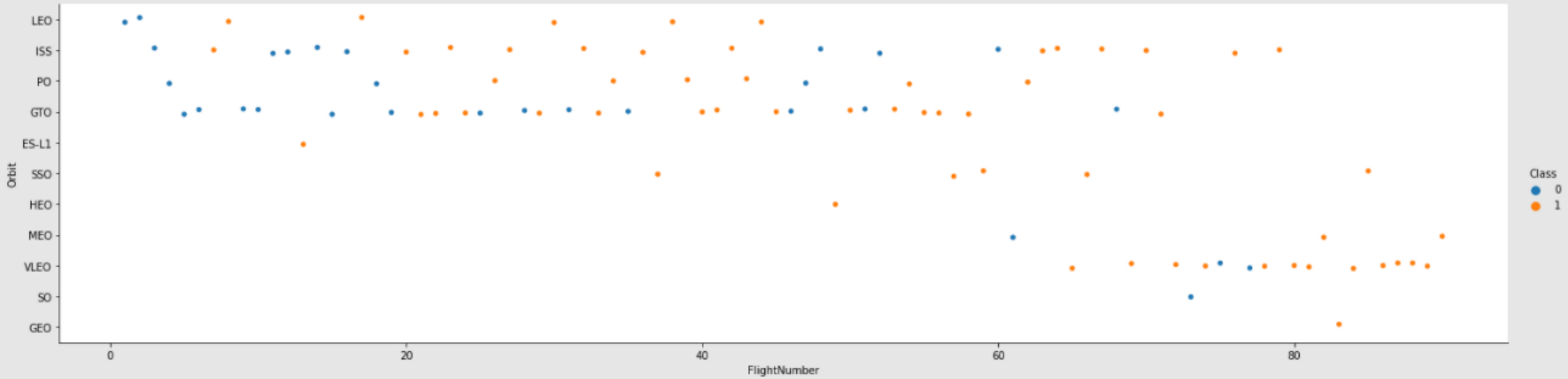
Show a bar chart for the success rate of each orbit type. Show the screenshot of the scatter plot with explanations



Bar chart shows the mean value for successful landing rate. As we can see, four orbits stand out with their means equal to 1. That means that all the launches sent to them had a successful landing.

Flight Number vs. Orbit Type

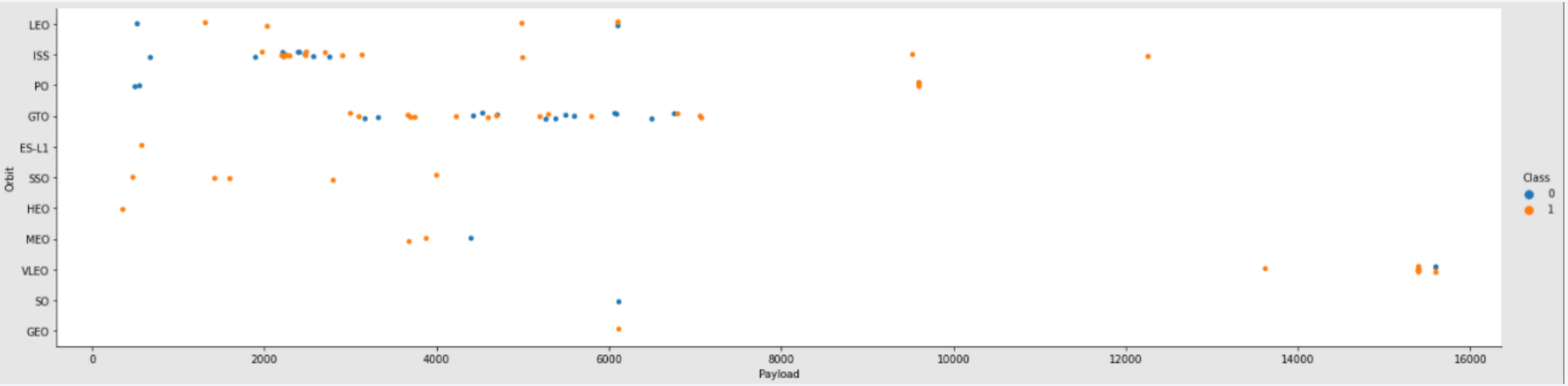
Show a scatter point of Flight number vs. Orbit type. Show the screenshot of the scatter plot with explanations



Scatter plot shows distribution of flight number with their orbits and whether the landing was successful or not (marked with colors = blue for fail, orange for success).

Payload vs. Orbit Type

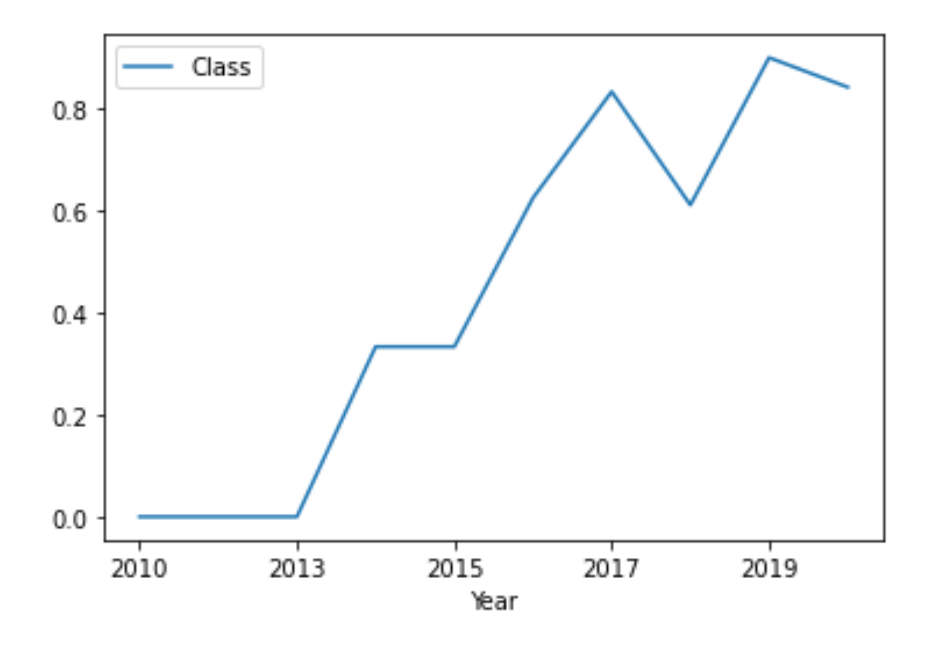
Show a scatter point of payload vs. orbit type. Show the screenshot of the scatter plot with explanations



Scatter plot shows distribution of payload values with their orbits and whether the landing was successful or not (marked with colors = blue for fail, orange for success).

Launch Success Yearly Trend

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



Line chart shows how the success rate increses throughout the last years.

All Launch Site Names

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

In [4]:
```
%%sql
SELECT DISTINCT(LAUNCH_SITE) FROM spacex;
```

```
 * ibm_db_sa://cgn84712:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/BLUDB
Done.
```

Out[4]:

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

There are four different sites. Their names can be seen above.

Launch Site Names Begin with 'CCA'

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

In [6]:
```
%%sql
select * from spacex where launch_site like '%CCA%' limit 5;
```

```
 * ibm_db_sa://cgn84712:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/BLUDB
Done.
```

Out[6]:

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

To show only 5 records I had to use the SQL command ‚LIMIT'

Total Payload Mass

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

```
In [9]: %%sql
        select sum(payload_mass__kg_) as sum_of_mass from spacex where customer = 'NASA (CRS)';
```

```
 * ibm_db_sa://cgn84712:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/BLUDB
Done.
```

Out[9]:

sum_of_mass
45596

Total payload mass carried by NASA's boosters were 45 596 kg.

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

```
In [10]: %%sql
         select avg(payload_mass__kg_) from spacex where booster_version = 'F9 v1.1';

          * ibm_db_sa://cgn84712:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/BLUDB
         Done.

Out[10]:    1
            2928
```

The average payload mass caried by booster F9 v1.1 was 2928 kg.

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

In [14]:
```
%%sql
select min(date) as min from spacex where mission_outcome = 'Success';
```

```
 * ibm_db_sa://cgn84712:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/BLUDB
Done.
```

Out[14]:

MIN
2010-06-04

The first successful landing took place on 4th of April 2010.

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

In [22]:
```
%%sql
select booster_version from spacex where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000;
```

```
 * ibm_db_sa://cgn84712:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/BLUDB
Done.
```

Out[22]:

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

As we can see, there are four boosters that fulfill given criteria.

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

In [24]:
```
%%sql
select mission_outcome, count(mission_outcome) as sum
from spacex
group by mission_outcome;
```

```
 * ibm_db_sa://cgn84712:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/BLUDB
Done.
```

Out[24]:

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

As we can see, there was only one failure. Rest of the missions were successful.

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

In [26]:
```
%%sql
select booster_version
from spacex
where payload_mass__kg_ = (select max(payload_mass__kg_) from spacex);
```

```
 * ibm_db_sa://cgn84712:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/BLUDB
Done.
```

Out[26]:

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

Above boosters can carry the maxium payload mass.

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

In [28]:
```
%%sql
select booster_version, launch_site, date, landing__outcome
from spacex
where
landing__outcome ='Failure (drone ship)' and year(date) = 2015;
```

```
 * ibm_db_sa://cgn84712:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/BLUDB
Done.
```

Out[28]:

booster_version	launch_site	DATE	landing__outcome
F9 v1.1 B1012	CCAFS LC-40	2015-01-10	Failure (drone ship)
F9 v1.1 B1015	CCAFS LC-40	2015-04-14	Failure (drone ship)

Above we can see which boosters were used in the failed landings in 2015. Locations are also provided.

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

In [31]:
```
%%sql
select landing__outcome, count(landing__outcome) as count
from spacex
group by landing__outcome
order by count desc;
```

```
 * ibm_db_sa://cgn84712:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/BLUDB
Done.
```

Out[31]:

landing__outcome	COUNT
Success	38
No attempt	22
Success (drone ship)	14
Success (ground pad)	9
Controlled (ocean)	5
Failure (drone ship)	5
Failure	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

As we can see, most of the landings in the given years were a success.

Section 4

Launch Sites Proximities Analysis

Site locations

Replace <Folium map screenshot 1> title with an appropriate title. Explore the generated folium map and make a proper screenshot to include all launch sites' location markers on a global map. Explain the important elements and findings on the screenshot

On the above map there are four locations. One on the west coast and three on the east coast in Florida. What they all have in common is the very small distance to the coast.

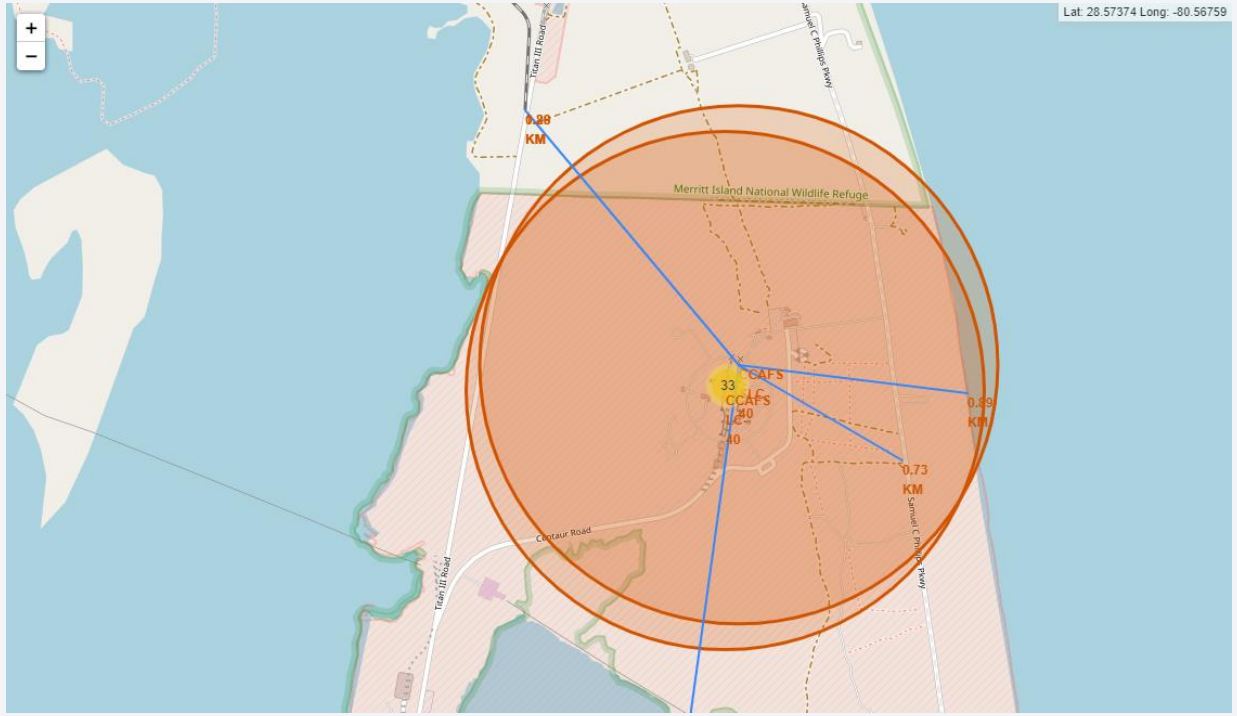
CCAFS LC-40 launches

Replace <Folium map screenshot 2> title with an appropriate title. Explore the folium map and make a proper screenshot to show the color-labeled launch outcomes on the map. Explain the important elements and findings on the screenshot

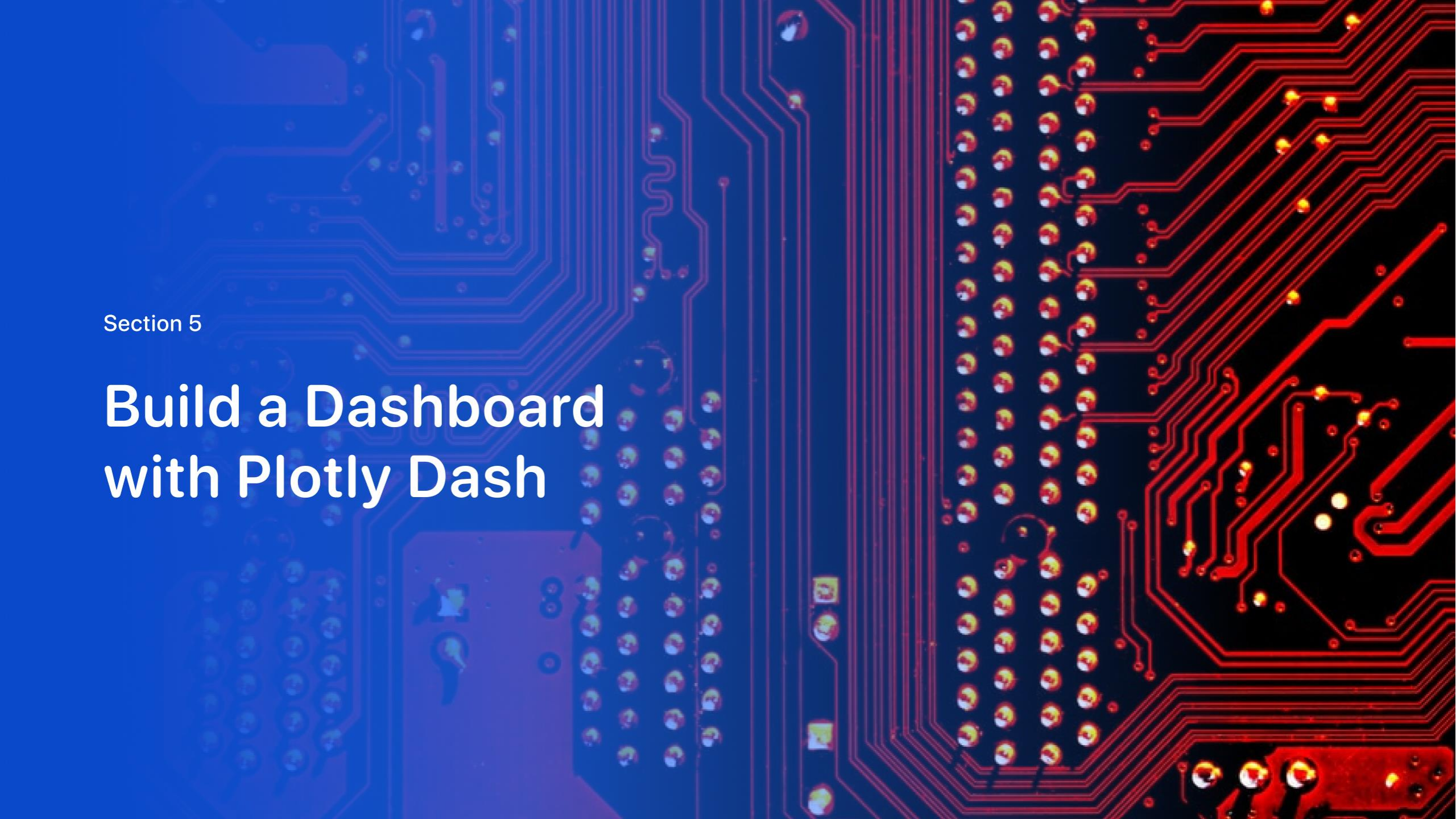
Screenshot on the left is a close-up to location CCAFS SLC-40 where we can see all the single landings. If there were successful, they are marked green. Othervise it's the color red. Above that location, we can see another's site name – CCAFS SLC – 40 with number 7. It means from that location there were seven launches performed. After close-up, we will be able to see similar view as for CCAFS LC-40 .

CCAFS SLC-40 distance to different objects

Replace <Folium map screenshot 3> title with an appropriate title. Explore the generated folium map and show the screenshot of a selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed. Explain the important elements and findings on the screenshot



Here we can see what distance if between site CCAFS SLC – 40 and closest objects like coast, rail, road and city. The blue lines go directly to that points. The closest city is Cape Canaveral and in straight line there is around 18km to it from our launch site.

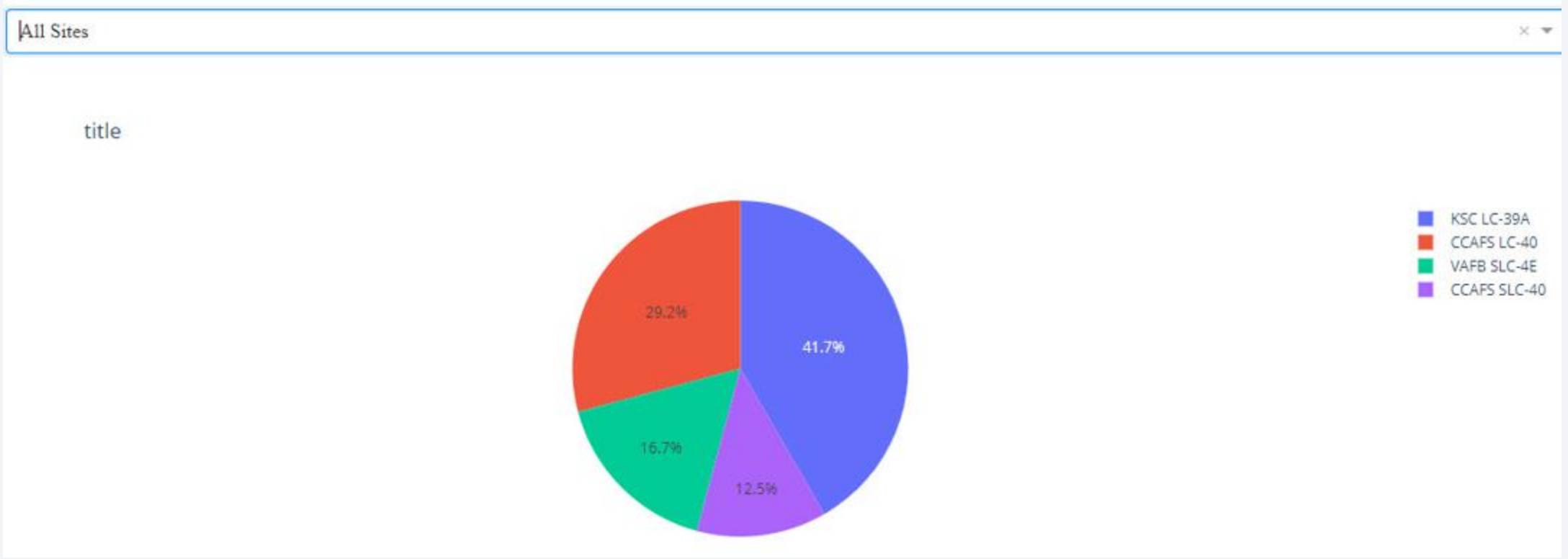


Section 5

Build a Dashboard with Plotly Dash

Space X Launch Records Dashboard

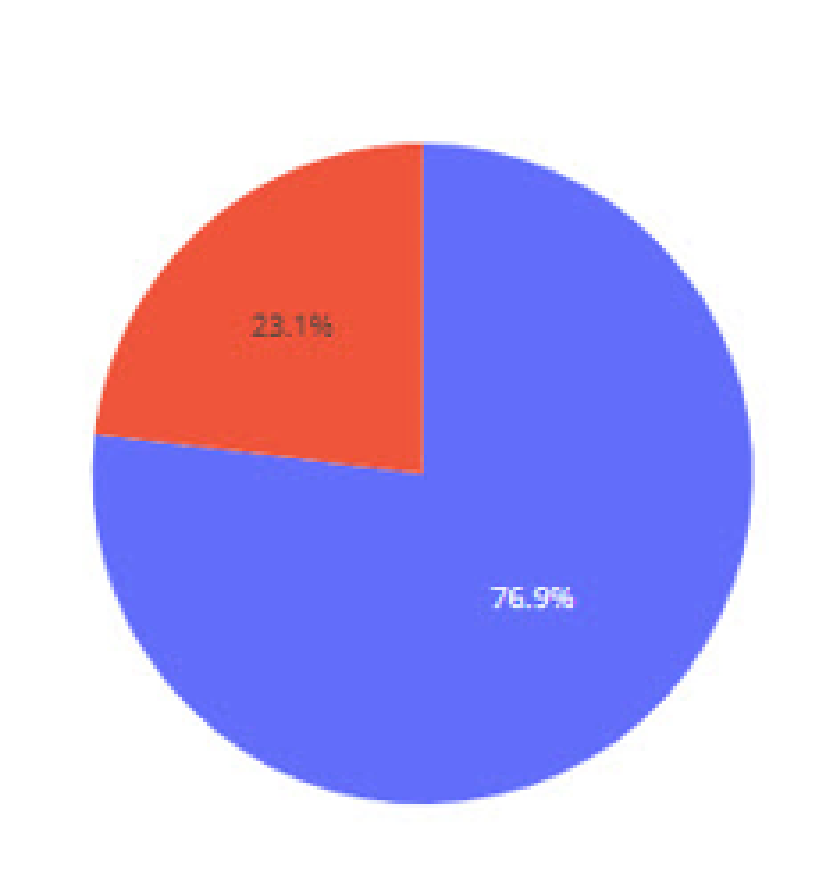
Replace <Dashboard screenshot 1> title with an appropriate title. Show the screenshot of launch success count for all sites, in a piechart. Explain the important elements and findings on the screenshot



The above pie chart shows us what was the success ratio for all sites. It clearly says that KSC LC-39A had the biggest number of successful mission.

Total success launches for KSC LC-39A

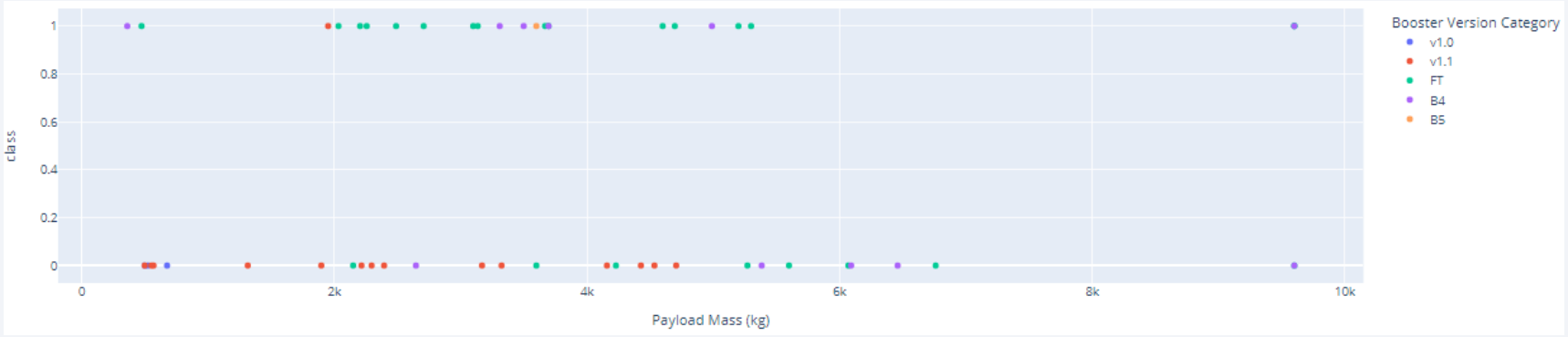
Replace <Dashboard screenshot 2> title with an appropriate title. Show the screenshot of the pie chart for the launch site with highest launch success ratio. Explain the important elements and findings on the screenshot

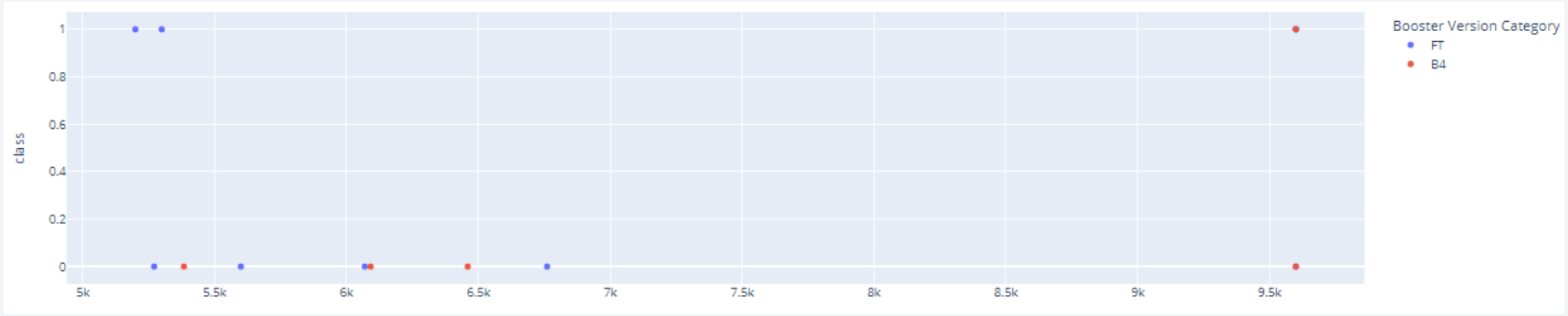


Blue color is as indicator for successful launches.
As we can see, KSC LC 39-A has a quite big score for it. 3 out of 4 launches there went well.

Correlation between payload and success for all site

Replace <Dashboard screenshot 3> title with an appropriate title. Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider. Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.





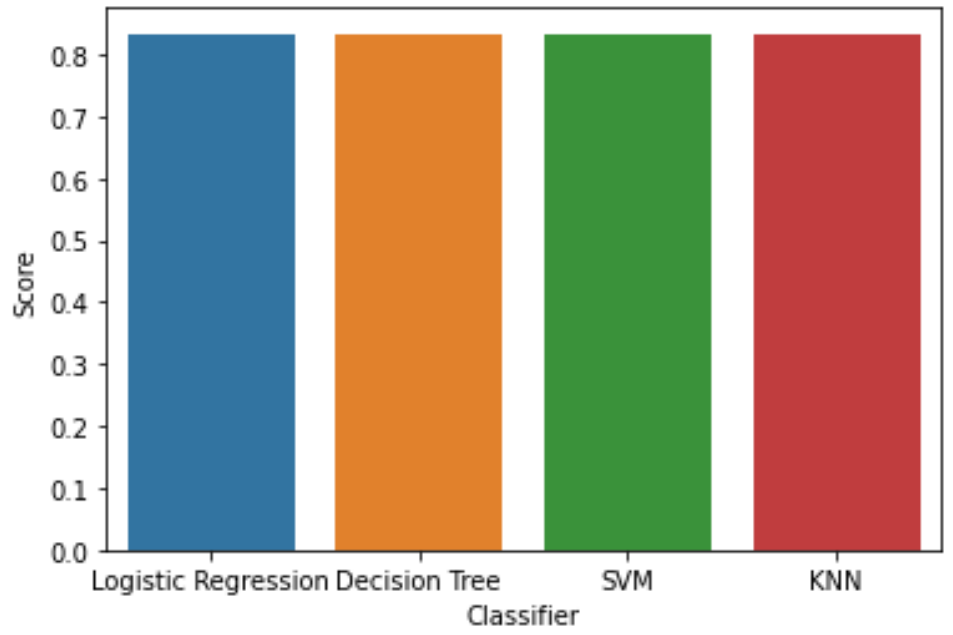
These scatter plots show correlation between payload and success launches. The first one focuses on whole range while the second one only on these payloads above 5000 kg. Thanks to that observation we know that bigger chance to perform a successful landing has a rocket with payload less than 6000kg. The perfect range would be 2k – 4k kg.

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



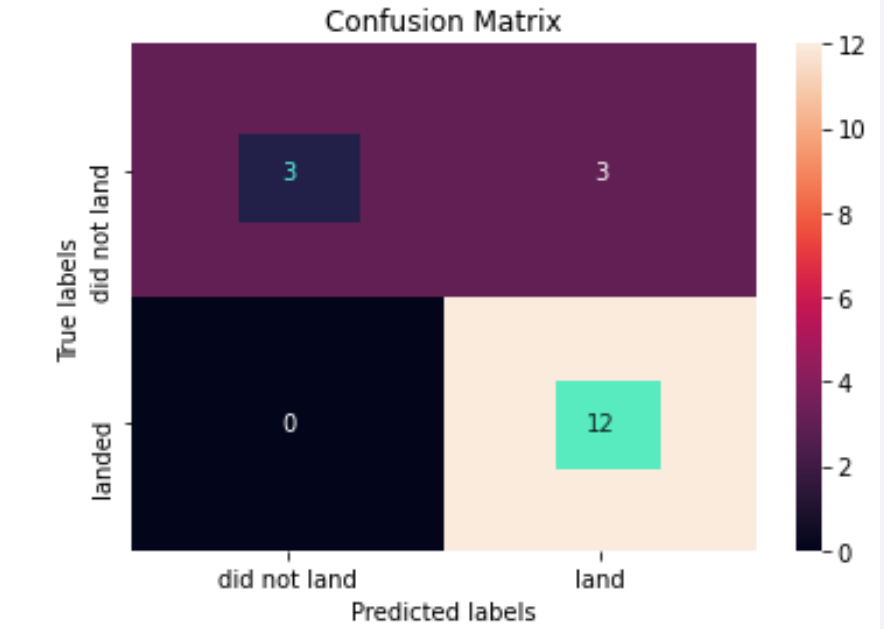
	Classifier	Score
0	Logistic Regression	0.833333
1	Decision Tree	0.833333
2	SVM	0.833333
3	KNN	0.833333

All models have reached the same score while evaluating test datasets with .score() method.

Confusion Matrix

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

```
In [36]: yhat_knn = knn_cv.predict(X_test)
         plot_confusion_matrix(Y_test,yhat_knn)
```



Because all models reached the same score, I will show only confusion matrix for KNN model. The others look excatly the same.

What we can see in that graph is that out of 18 records in the test datasets, 15 have been predicted correctly. 12 true positive (land-land) and 3 true negative (did not land – did not land). Our ML model was wrong about 3 records. In reality, 12 landing were successful and 6 not. Our model predicted that only 3 did not land.

Conclusions

Based on the performed analysys I can highlight few important points.

- The best site to perform a launch would be KSC LC-39A as it has the highest success rate
- Payload mass should close in the range of 4000 kg – 6000 kg. If it's below or above, the chance for failure grows
- All the missions sent to below orbit have a 100% success rate. It's worth to take it under consideration.
- Models I've built have more than 80% of accuracy which is consider as a good score. Thank's to them we can test different features and find the best parameters to prepare Space Y's successful launch and landing.

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

All the additional files like datasets and screenshots from Plotly dashboard can be found in my GitHub repo's folder „Additional". Python and SQL code have been executed within notebooks that can be find in the „notebook" folder. Also, in the firsts slides there are links to them. I haven't created any additional notebooks or files as there was no need for that. Sometimes I've created some extra code but that is included in the mentioned notebooks.

Thank you!