IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Anurag Sharma

06-09-2021

Outline

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

Executive Summary

- For data collection we used Beautiful Soup and Python Requests libraries.
- For exploratory data analysis, the data was hosted on IBM Cloud DB2 instance and SQL was used for querying the data.
- The plots and charts were built in python using matplotlib and seaborn. The geographical data was visualized on folium maps.
- An interactive dashboard was also built using Plotly Dash framework.
- In the end we build several predictive models using Python’s machine learning framework called sklearn.
- The models were evaluated and compared and the model with the highest accuracy for finally selected.

Introduction

- Space X manages to perform space travel at much lower cost expense due to the recovery of their first stage rockets.
- There our organization, called Space Y, wants to predict whether after a launch Space X would recovery the first stage or not.
- If we can predict the likelihood of the first stage being recovered we can predict the overall expense of the launch being cheaper than other competitors.
- Therefore we intend to build a predict model to answer the same question and help us (Space Y) make a more competitive bid as a startup for rocket launches.

Section 1

Methodology

Methodology

Executive Summary

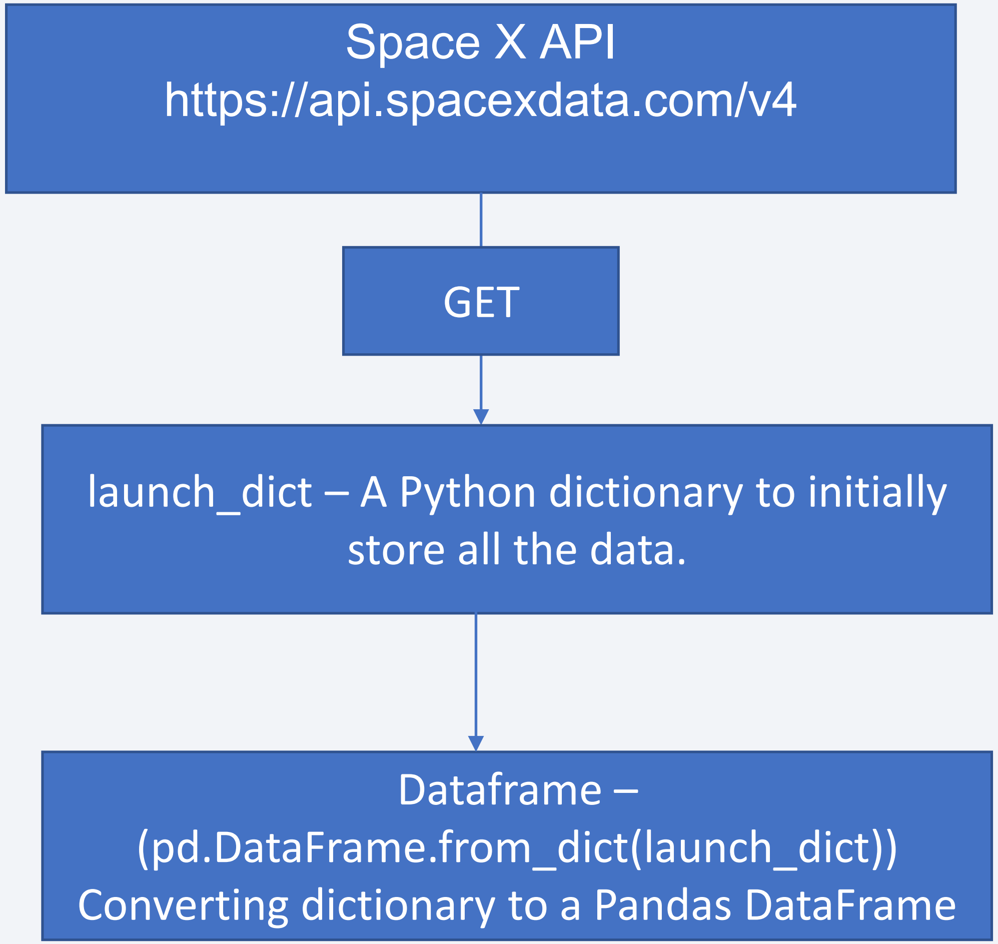
- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Data was collected from the publicly available API endpoints provided by Space X
- The data consisted of all the launch details related to Falcon rockets.
- https://api.spacexdata.com/v4
- Python’s requests library was used to make GET requests to the endpoints and download data in JSON format.
- The JSON was then converted to Pandas Data Frame.
- Also, historical records of Falcon 9 launches was collected by web scrapping from the wiki page.

Data Collection – SpaceX API

- Space X API was used to get the launch related data.
- Endpoints used were – payload, cores, launchpads, rockets
- https://github.com/anurag-ks/IBMDataScienceCapstoneProject/blob/main/lab_1.ipynb



Space X API
https://api.spacexdata.com/v4

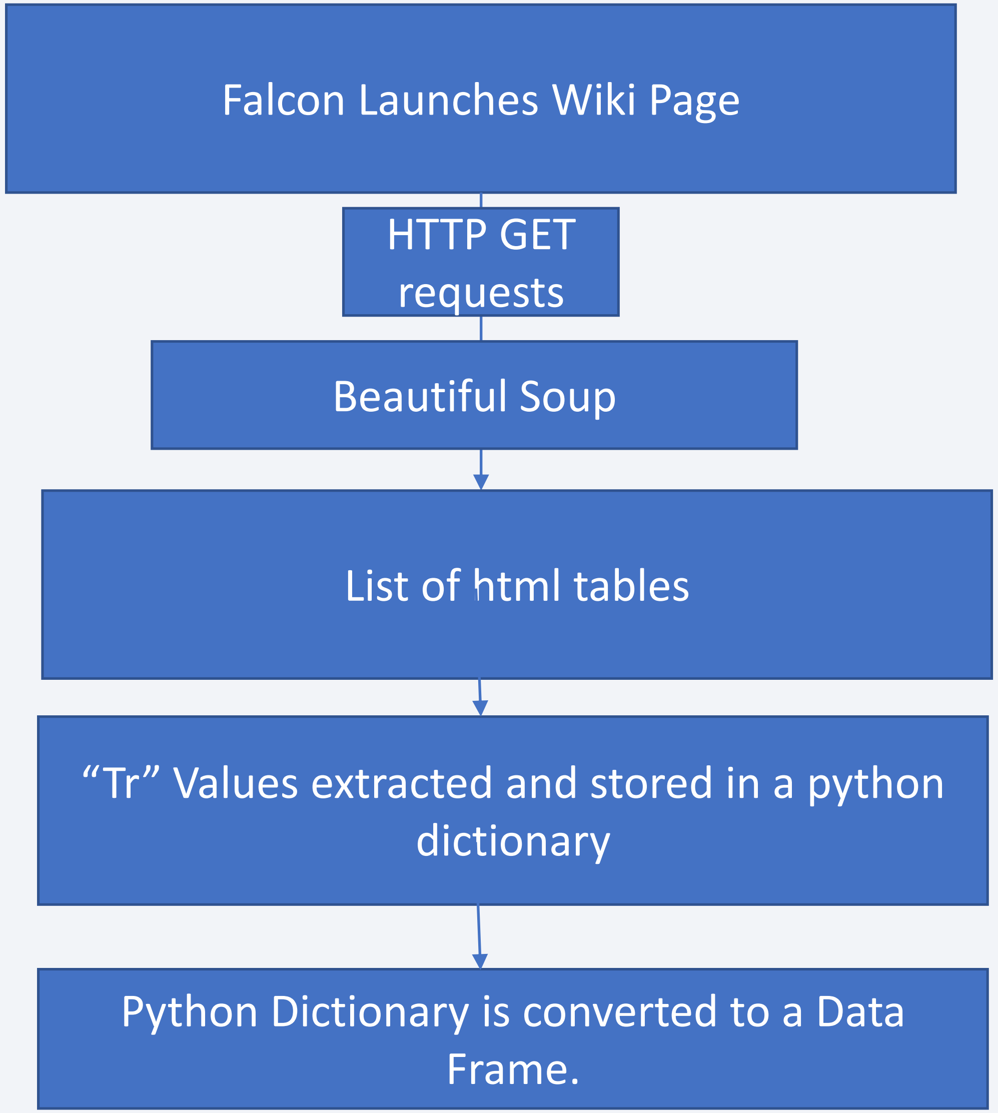
GET

launch_dict – A Python dictionary to initially store all the data.

Dataframe –
(pd.DataFrame.from_dict(launch_dict))
Converting dictionary to a Pandas DataFrame

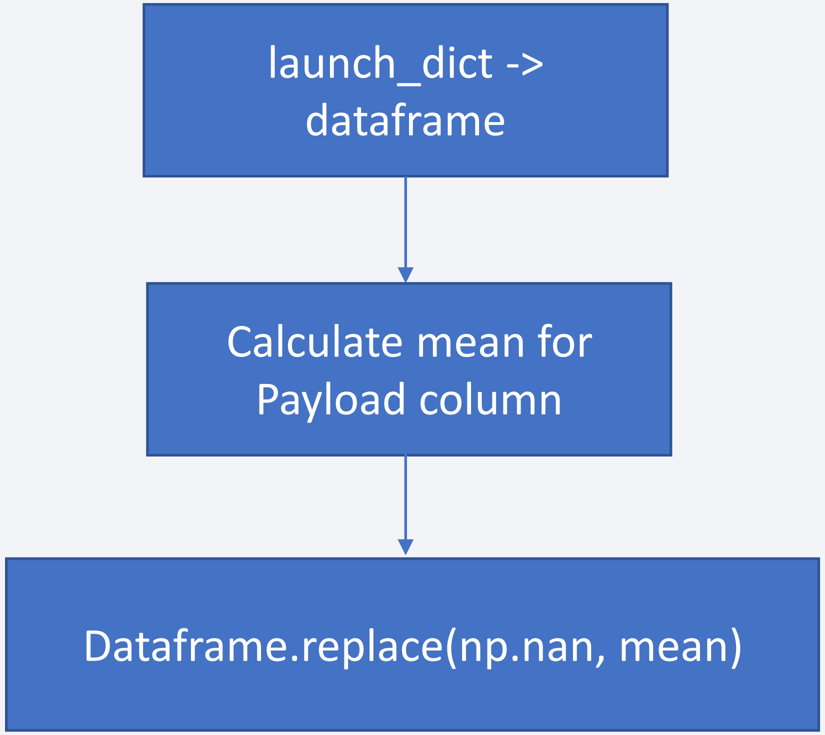
Data Collection - Scraping

- Falcon 9 historical launch data was collected from Wikipedia.
- We made used of Beautiful soup for web scrapping data from the html sites into a pandas data frame.
- https://github.com/anurag-ks/IBMDataScienceCapstoneProject/blob/main/Web%20Scrapping.ipynb


Falcon Launches Wiki Page
HTTP GET requests
Beautiful Soup
List of html tables
“Tr” Values extracted and stored in a python dictionary
Python Dictionary is converted to a Data Frame.

Data Wrangling

- From the main data frame we filtered out rows which contained information about Falcon 9 only.
- There were some missing values in Payload Mass column which were removed and replaced by the overall mean of the column.
- https://github.com/anurag-ks/IBMDataScienceCapstoneProject/blob/main/lab_1.ipynb


launch_dict -> dataframe

Calculate mean for Payload column

Dataframe.replace(np.nan, mean)

EDA with Data Visualization

- Several plots were made with the help of Python libraries – matplotlib, seaborn.
- The plots were made to visualize and understand different relationships the data features had with each other, for example a cat plot was made for “Flight No. vs Launch Site” relationship.
- Similarly, we had plotted scatter plot, bar charts and line plots for the same objective.
- The plots can be found by clicking on the link below.
- https://github.com/anurag-ks/IBMDataScienceCapstoneProject/blob/main/lab_4.ipynb

EDA with SQL

- SQL can be used to query and process data stored in a database.
- Space X launch data was stored in a DB2 instance on IBM Cloud.
- Using Python’s sqlalchemy, ibm_db_sa and ipython-sql we were able to connect to the IBM Cloud hosted database and perform SQL queries from Python.
- Several SQL queries were performed to explore the data set, for example –
    - SELECT MIN(DATE) FROM SPACEXTBL WHERE MISSION_OUTCOME='Success' AND LANDING__OUTCOME='Success (ground pad)';
- The above query was used to search for last date when the first successful landing outcome in ground pad was achieved.
- More queries can be found by clicking on the link below.
- https://github.com/anurag-ks/IBMDataScienceCapstoneProject/blob/main/lab_3.ipynb

Build an Interactive Map with Folium

- Folium which a python based library to generate map visualizations, was used to explore data on the map views.
- A successful launch can depend on the location of the launch site and it's proximity to near by locations such as sea coast, townships, rivers and etc.
- Hence it is important to find the optimal location for launch. For this reason we made use of Folium to generate interactive maps.
- https://github.com/anurag-ks/IBMDataScienceCapstoneProject/blob/main/lab_5.ipynb

Build a Dashboard with Plotly Dash

- Plotly Dash was used to build simple web based application (built on top of Flask micro framework).
- The web app is a simple dashboard which shows two plots, namely –
  - Pie Chart – Success Rate vs Launch Sites
  - Scatter Chart – Launch Outcome vs Payload Mass.
- We also added a dropdown from where the user can select which launch site they want see the data for. Also, we added a slider to select the range for payload mass for the second scatter plot.
- https://github.com/anurag-ks/IBMDataScienceCapstoneProject/blob/main/dash_app.py

Predictive Analysis (Classification)

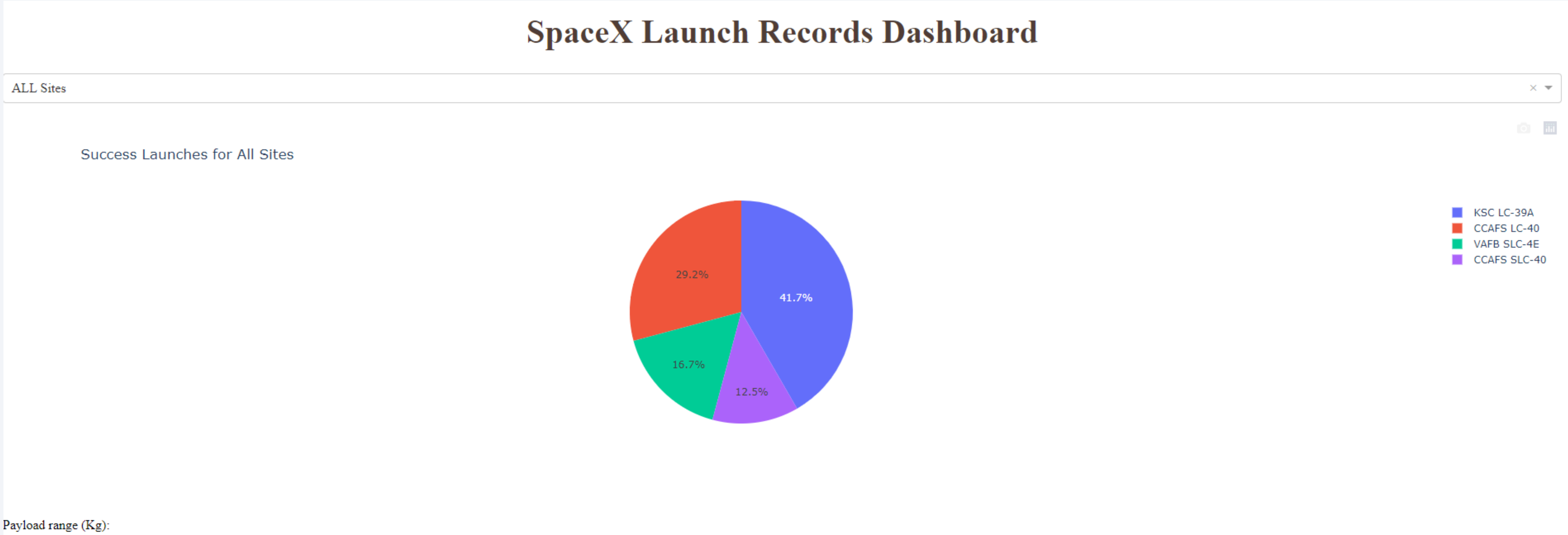
- Predictive models such as logistic regression, KNN, SVM and Decision Trees were used to make predictions on whether the launch will be successful or not based on available input features such as Payload Mass, Flight No, Orbit type and etc.
- The input data was split into training and test sets. The training set was initially used to build the models and the test sets were used to evaluate the accuracy of the same models.
- Based on the accuracy results we concluded which model performed best.
- The hyper parameters of each models were also figured out using the Grid Search approach.
- https://github.com/anurag-ks/IBMDataScienceCapstoneProject/blob/main/lab_6.ipynb

Results

- Our EDA included a wide range of data plots and map views.
- We were able visualize relationships between different input features and outcomes based on the same.
- Scatter Plots like that of Flight No. vs Outcome showed us that the outcome of a launch did not depend a lot on flight number, however the they were affected a lot by payload mass and the orbit selected for the launch.
- Using Folium we visualized locations where the successful launches had occurred and where failed launches had occurred. We also got to figure the distances of near by town ships or coasts from the launch zones.

Results

- Our Plotly Dashboard was used to create an interactive experience for the user to explore the data with dynamic inputs which then generated dynamic charts based on those given input data.



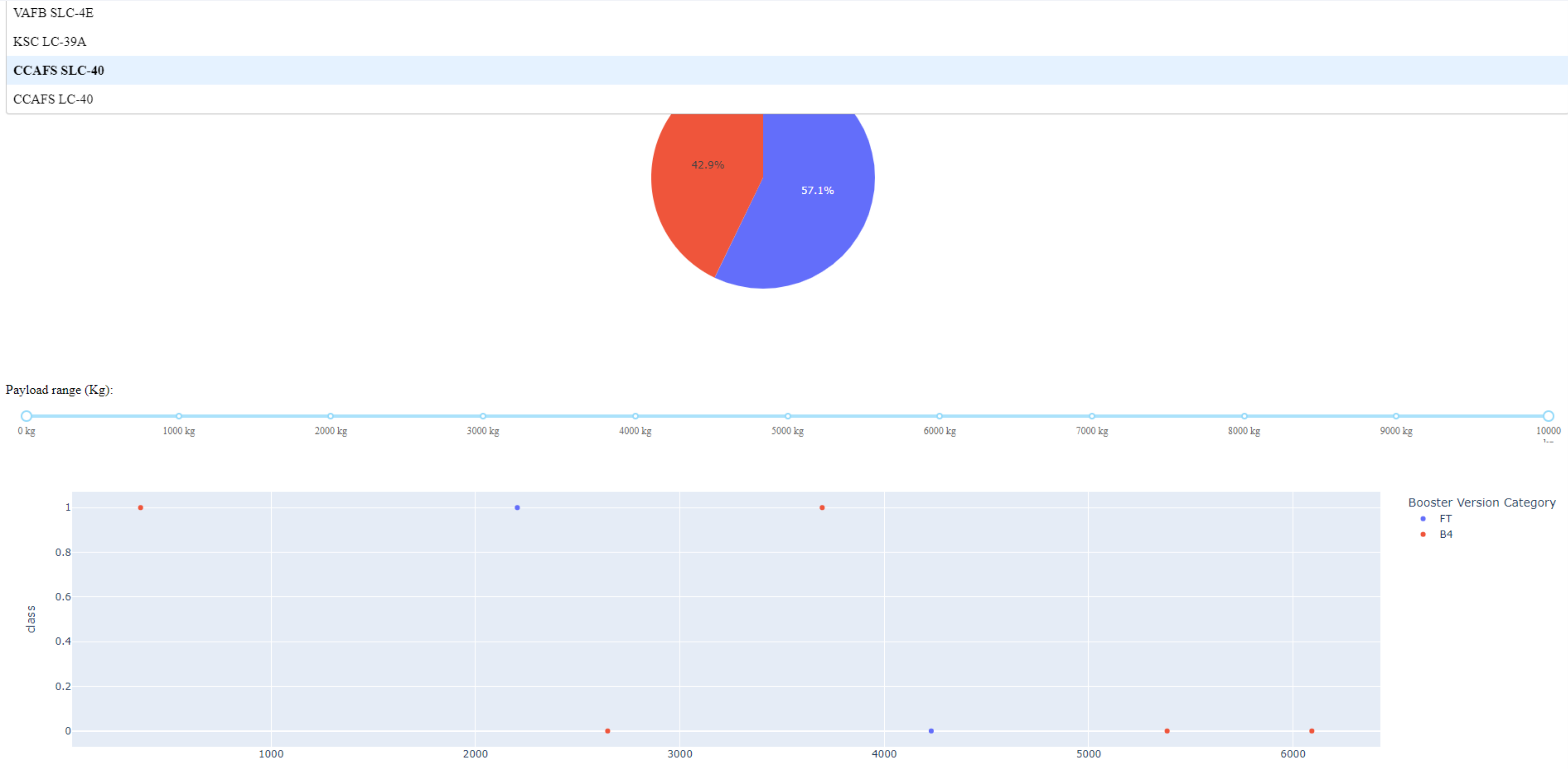
SpaceX Launch Records Dashboard

ALL Sites

Payload range (Kg):

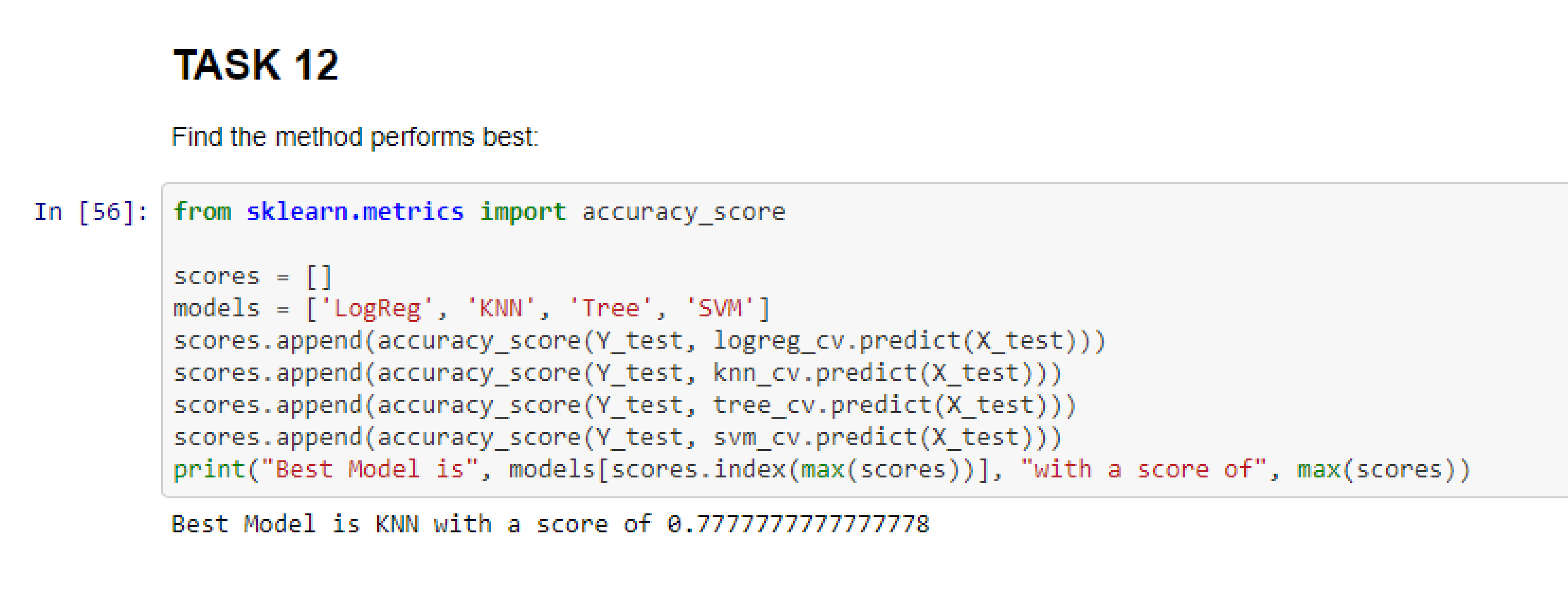
Results

More screenshots of the Plotly dashboard.



Results

- Also, based on the accuracy results of our predictive models we were able to figure out which model worked best for our data set.
- K-Nearest Neighbor model generally gave us the best accuracy score therefore we concluded that it should be the best model for predictive analysis of the given data.



TASK 12

Find the method performs best:

In [56]:

```
from sklearn.metrics import accuracy_score

scores = []
models = ['LogReg', 'KNN', 'Tree', 'SVM']
scores.append(accuracy_score(Y_test, logreg_cv.predict(X_test)))
scores.append(accuracy_score(Y_test, knn_cv.predict(X_test)))
scores.append(accuracy_score(Y_test, tree_cv.predict(X_test)))
scores.append(accuracy_score(Y_test, svm_cv.predict(X_test)))
print("Best Model is", models[scores.index(max(scores))], "with a score of", max(scores))
```

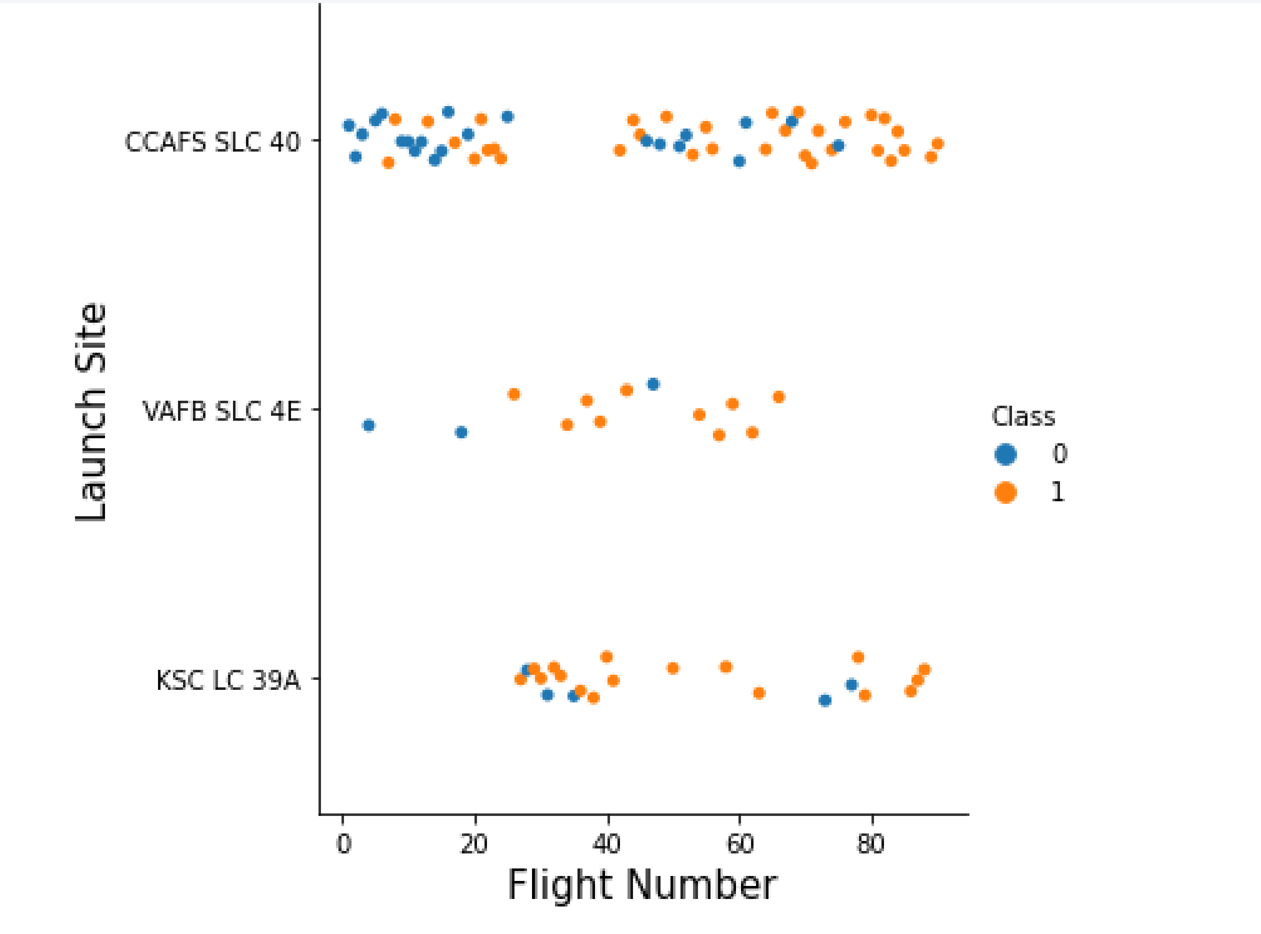
```
Best Model is KNN with a score of 0.7777777777777778
```

Section 2

Insights drawn from EDA

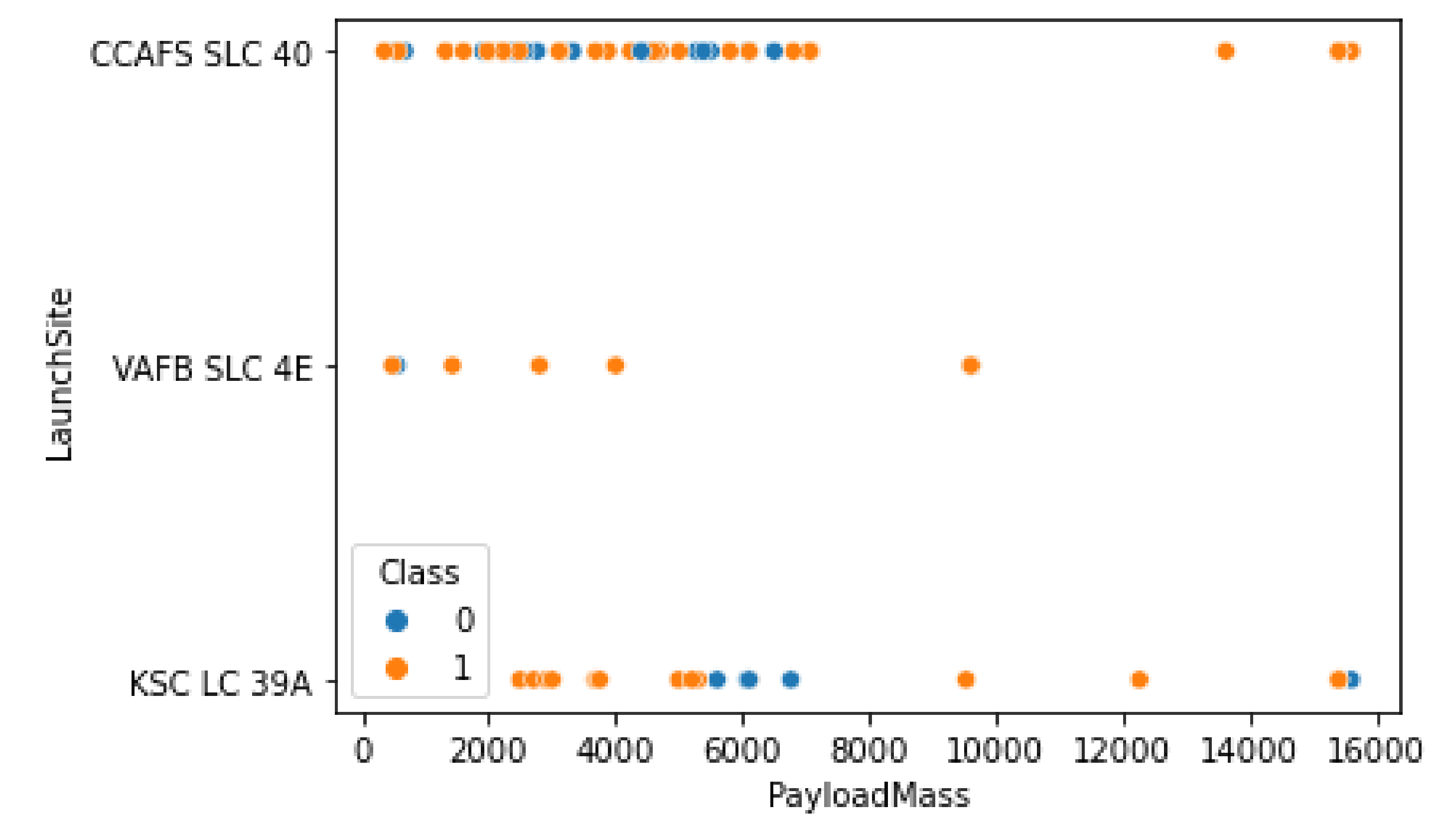
Flight Number vs. Launch Site

- Based on the graph, we can say that the launch site CCAFS SLC 40 has more launches than any other launch site.
- Also, the flight number doesn’t seem to have a distinct relationship to outcome of the flight.
- Same can be said for the launch site as well but we can see that KSC LC 39 A and VAFB SLC 4E has a relatively low failure rate as compared to CCAFS SLC 40.



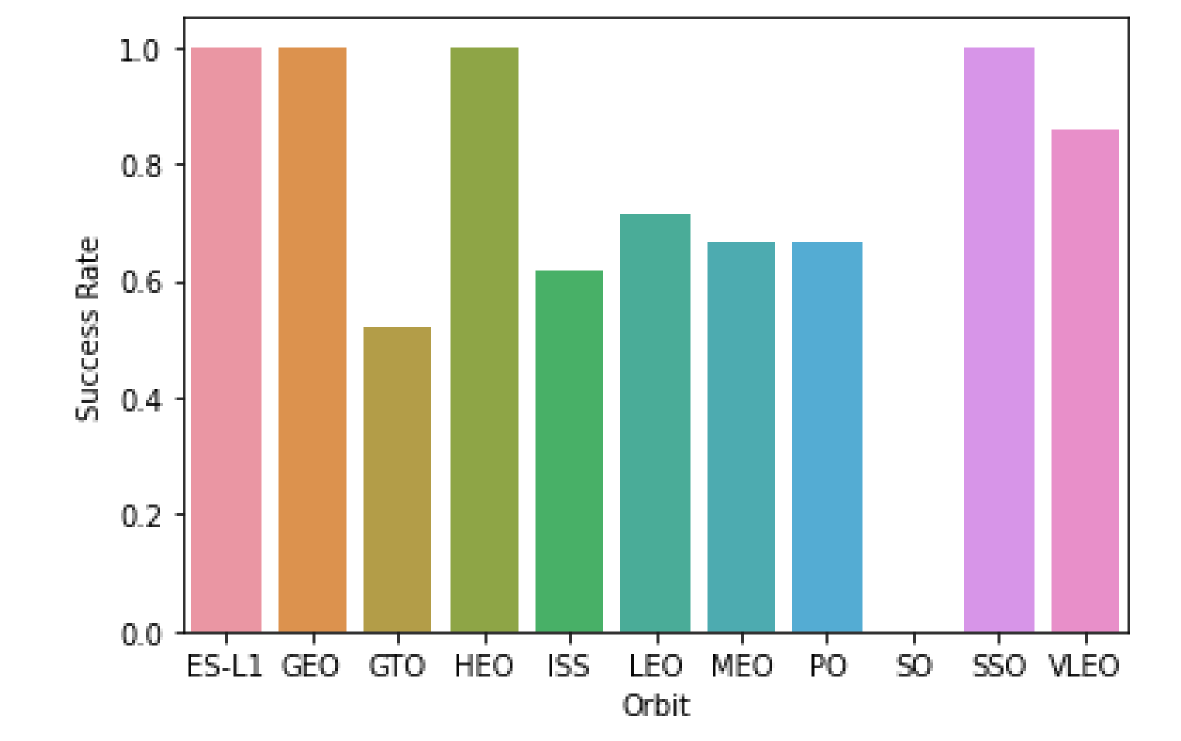
Payload vs. Launch Site

- Payload ranges from 0 kg to 16000 kg.
- The launches from KSC LC 39 A have a minimum payload of 2000 Kg
- Launches with Payload more than 8000 kg seems to have a successful outcome.
- CCAFS SLC 40 has the most launches with Payload ranging between 1000kg to 10000kg



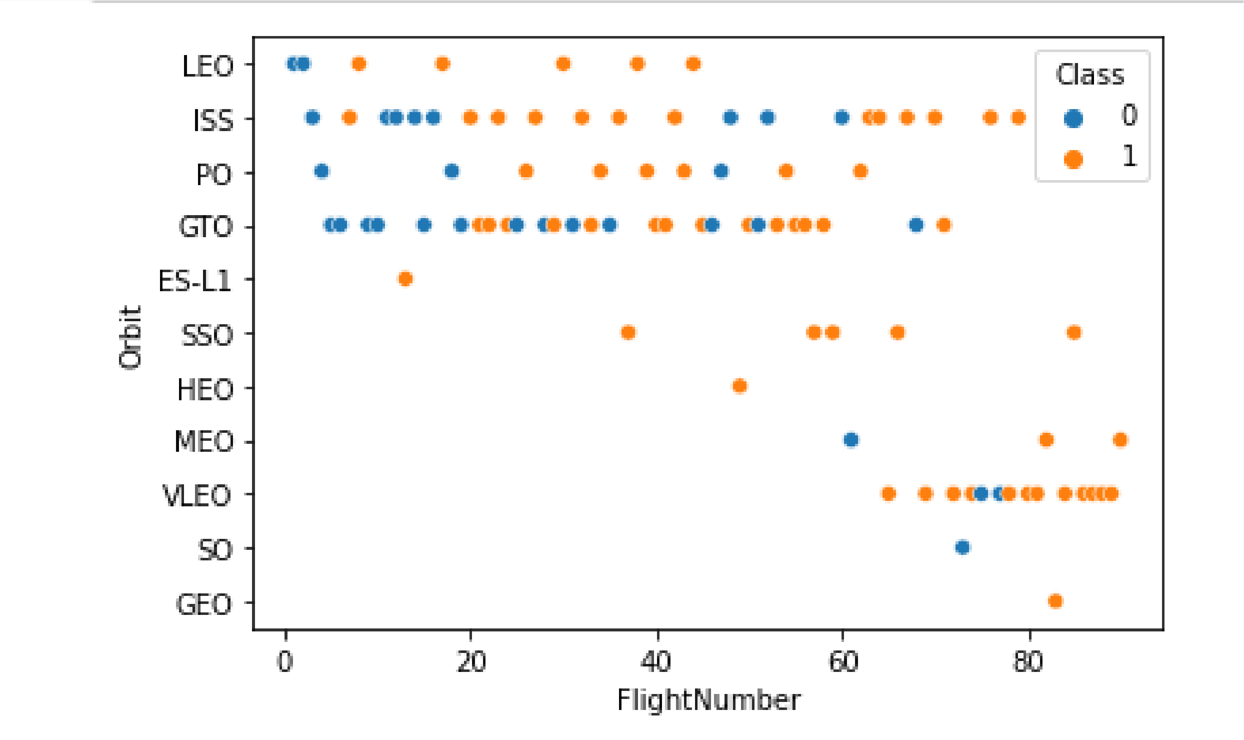
Success Rate vs. Orbit Type

- The following was the plot between Success Rate and the Orbit.
- As we can see GTO has the least amount of success.
- And, ES-L1, GEO, HEO and SSO have the highest success rate.



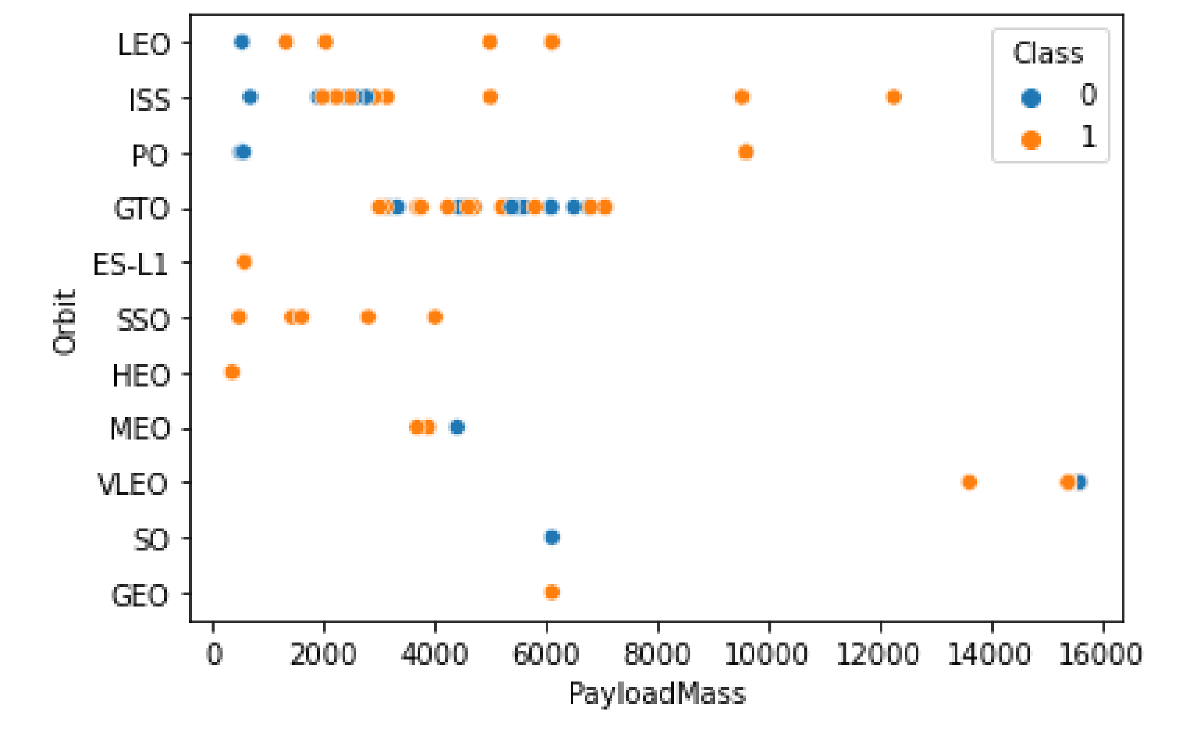
Flight Number vs. Orbit Type

- This is a scatter plot for Flight vs Orbit Type.
- You can 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

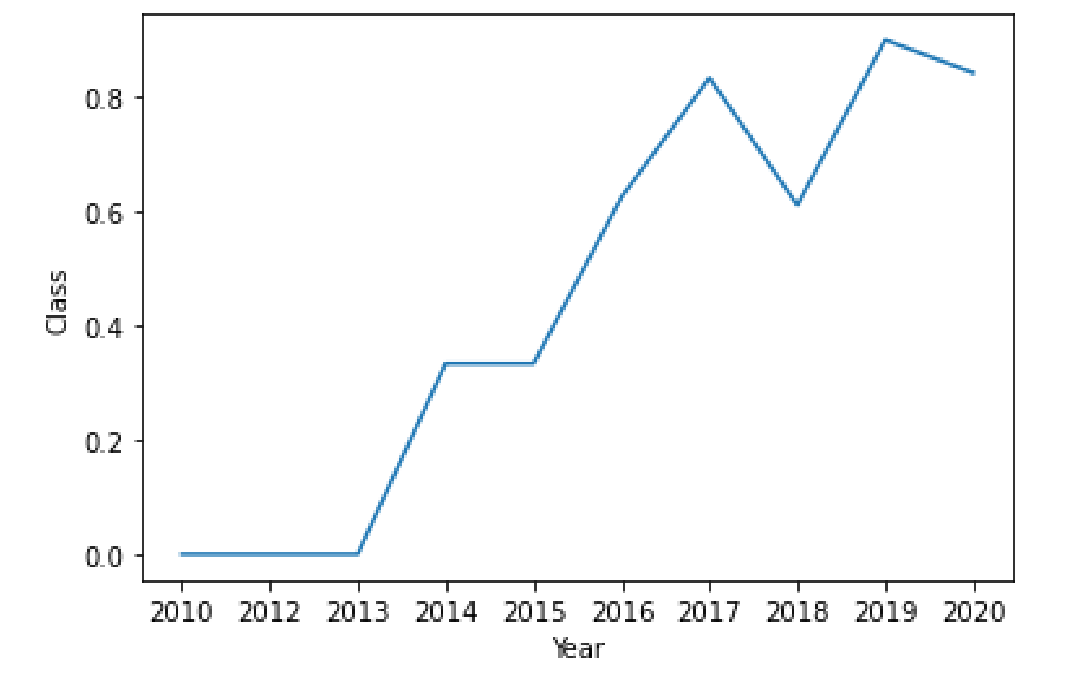
- The following is a scatter plot of payload mass vs orbit type.
- You should observe that Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.



Launch Success Yearly Trend

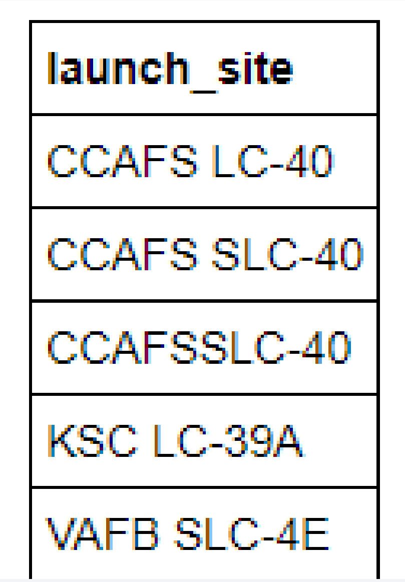
- The following is the trend of success vs year.

- We can observe that the rate has increased a lot after 2013 up to 2020.



All Launch Site Names

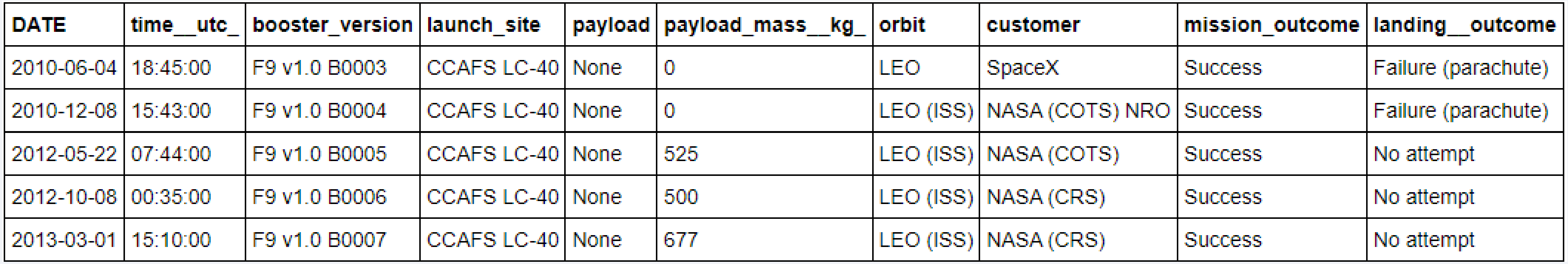
- All the unique launch sites are listed in the image provided.
- The query result was as follows -
- Query Used –
    - **select Unique(LAUNCH_SITE) from SPACEXTBL;**



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

Launch Site Names Begin with 'CCA'

- Query used –
  - select * from SPACEXTBL where LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
- Query Results -



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	None	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	None	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	None	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	None	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	None	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- The following query was used to calculate the total payload mass for the customer – “NASA (CRS)” in kilo grams.
- Query used –
  - SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER='NASA (CRS)‘
- Query Result – 45596 KG

Average Payload Mass by F9 v1.1

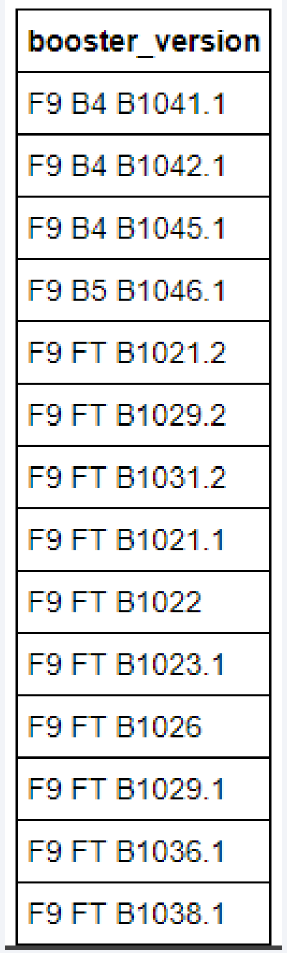
- Average payload mass carried by booster version F9 v1.1
- Query Used –
  - SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE BOOSTER_VERSION LIKE 'F9 v1.1%';
- Query Results –
  - 2534 kg

First Successful Ground Landing Date

- First successful landing outcome on ground pad
- Query Used –
  - SELECT MIN(DATE) FROM SPACEXTBL WHERE MISSION_OUTCOME='Success' AND LANDING__OUTCOME='Success (ground pad)';
- Query Results - 2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

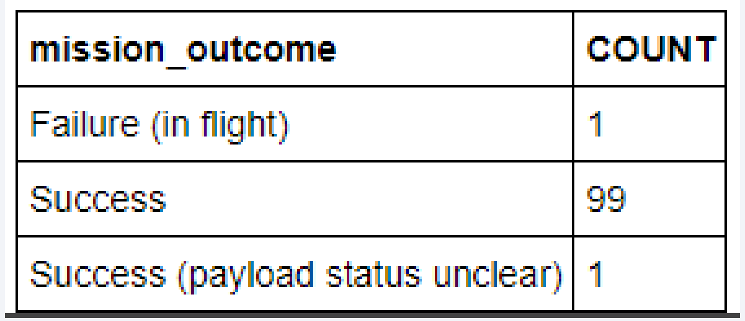
- Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- Query Used –
  - select distinct(booster_version) from SPACEXTBL where landing__outcome='Success (drone ship)‘ and PAYLOAD_MASS__KG_>=4000 and PAYLOAD_MASS__KG_<=6000;
- Query Results – (See the image on the side)



booster_version
F9 B4 B1041.1
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1
F9 FT B1021.2
F9 FT B1029.2
F9 FT B1031.2
F9 FT B1021.1
F9 FT B1022
F9 FT B1023.1
F9 FT B1026
F9 FT B1029.1
F9 FT B1036.1
F9 FT B1038.1

Total Number of Successful and Failure Mission Outcomes

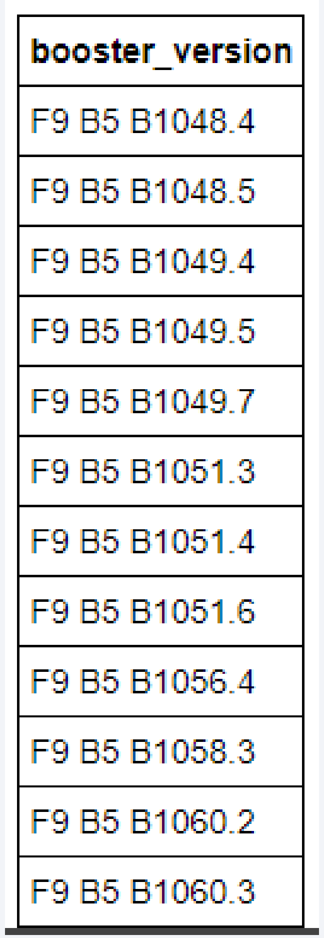
- Calculate the total number of successful and failure mission outcomes
- Query Used –
  - select mission_outcome, count(mission_outcome) as count from SPACEXTBL group by mission_outcome;
- Query Results –



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

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass
- Query Used –
  - select distinct(booster_version) from SPACEXTBL where payload_mass__kg_=(select max(payload_mass__kg_) from SPACEXTBL);
- Query Results –



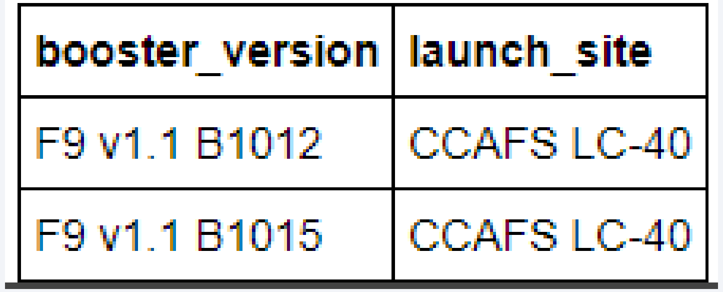
booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

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

- Query used –
  - select booster_version, launch_site from SPACEXTBL where landing__outcome='Failure (drone ship)' and YEAR(date)='2015';

- Query results –



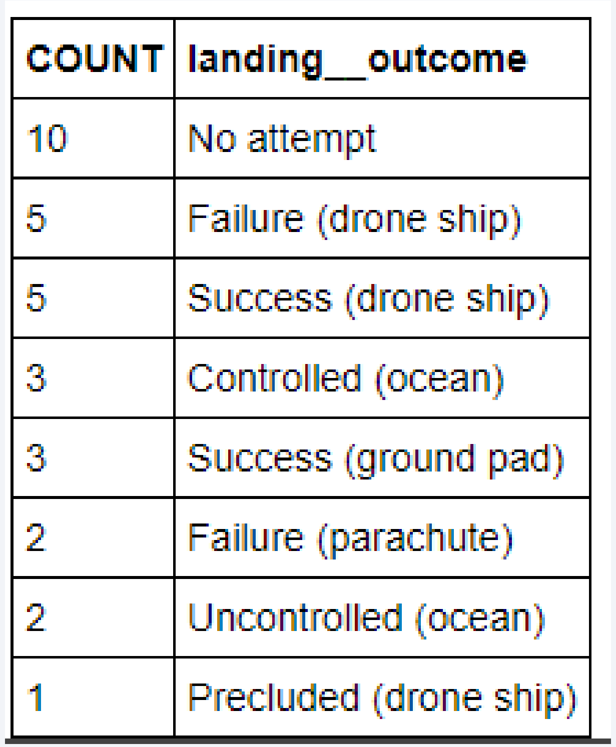
booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

- Query Used –
  - select count(landing__outcome) as count, landing__outcome from SPACEXTBL where date>='2010-06-04' and date<='2017-03-20' group by landing__outcome order by count(landing__outcome) desc;

- Query results – (see the side image)



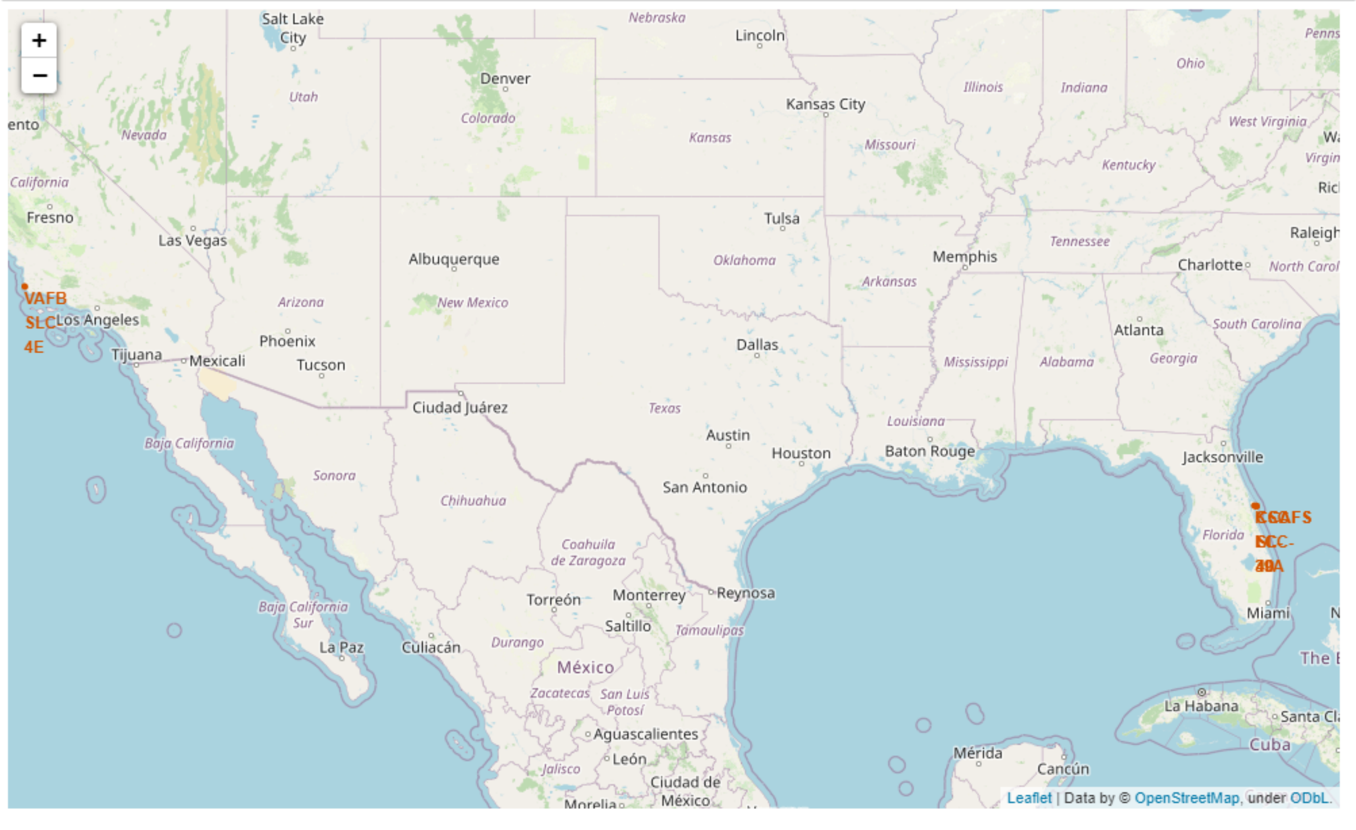
COUNT	landing__outcome
10	No attempt
5	Failure (drone ship)
5	Success (drone ship)
3	Controlled (ocean)
3	Success (ground pad)
2	Failure (parachute)
2	Uncontrolled (ocean)
1	Precluded (drone ship)

Section 4

Launch Sites Proximities Analysis

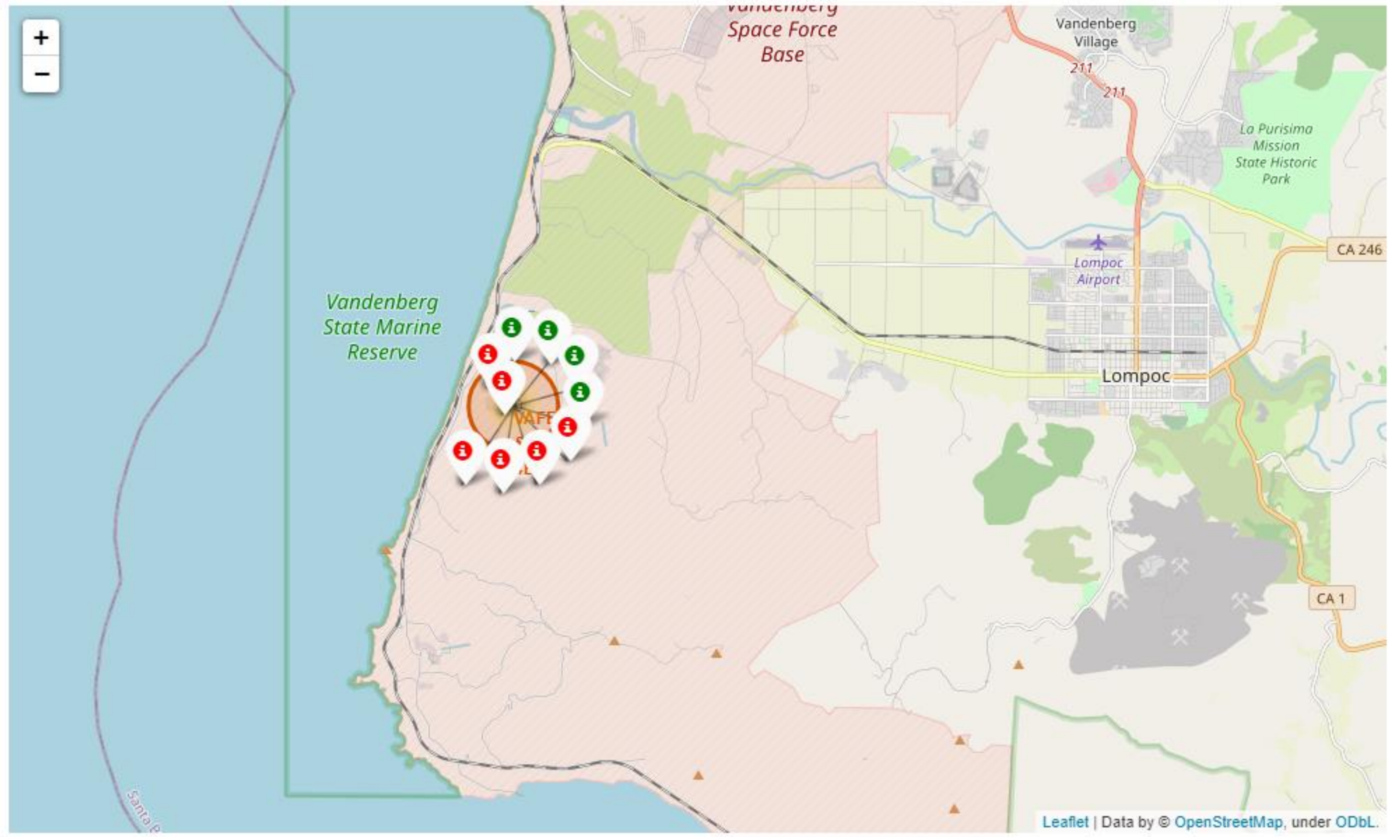
Overall Launch Sites locations

- The following map shows the different launch sites and the NASA JSC.
- They all have been marked on a standard map of the USA.
- We observe that most of the launch sites are located near to the sea shores.
- We have one near LA and the others are in Florida.



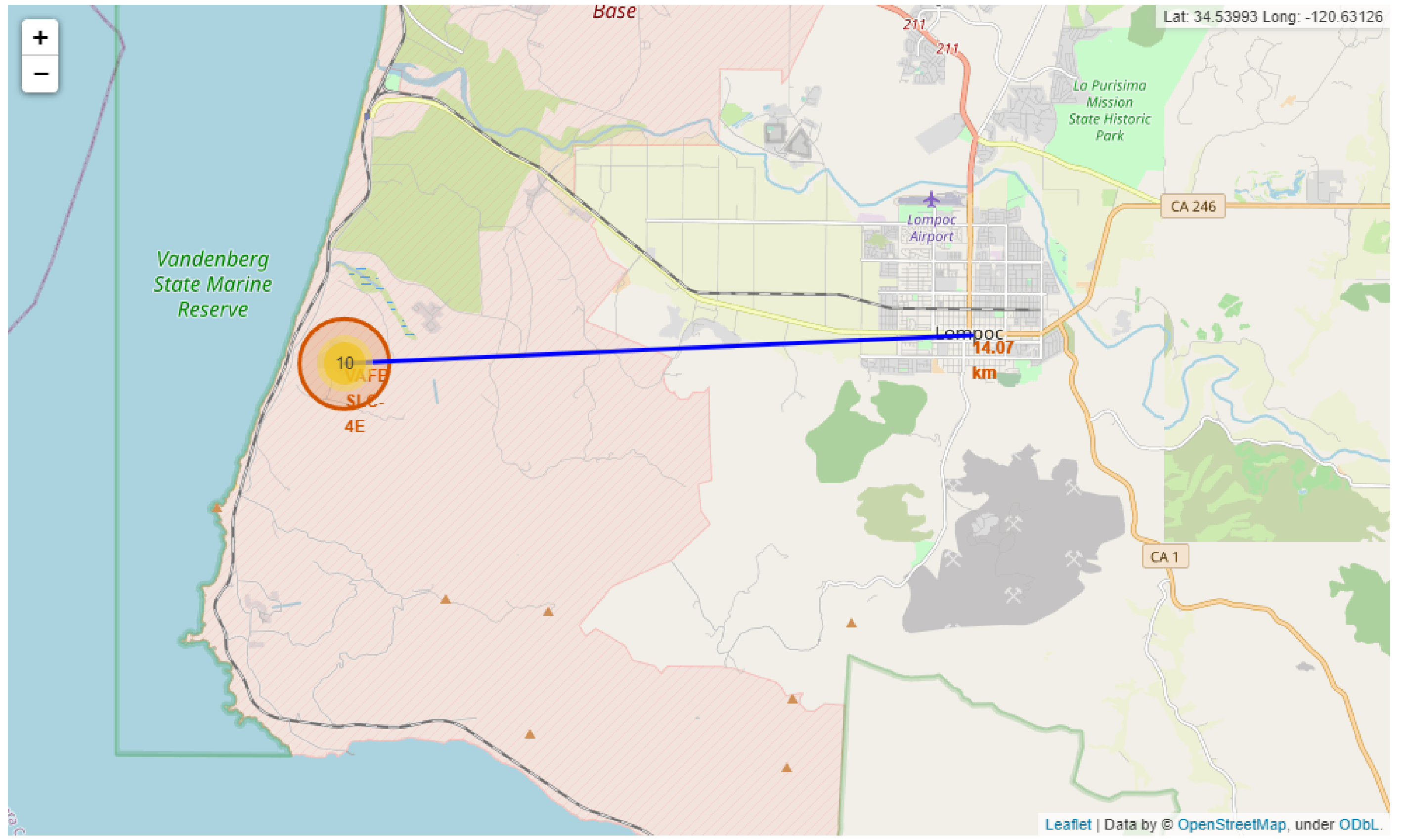
Success/Failed launches for each site

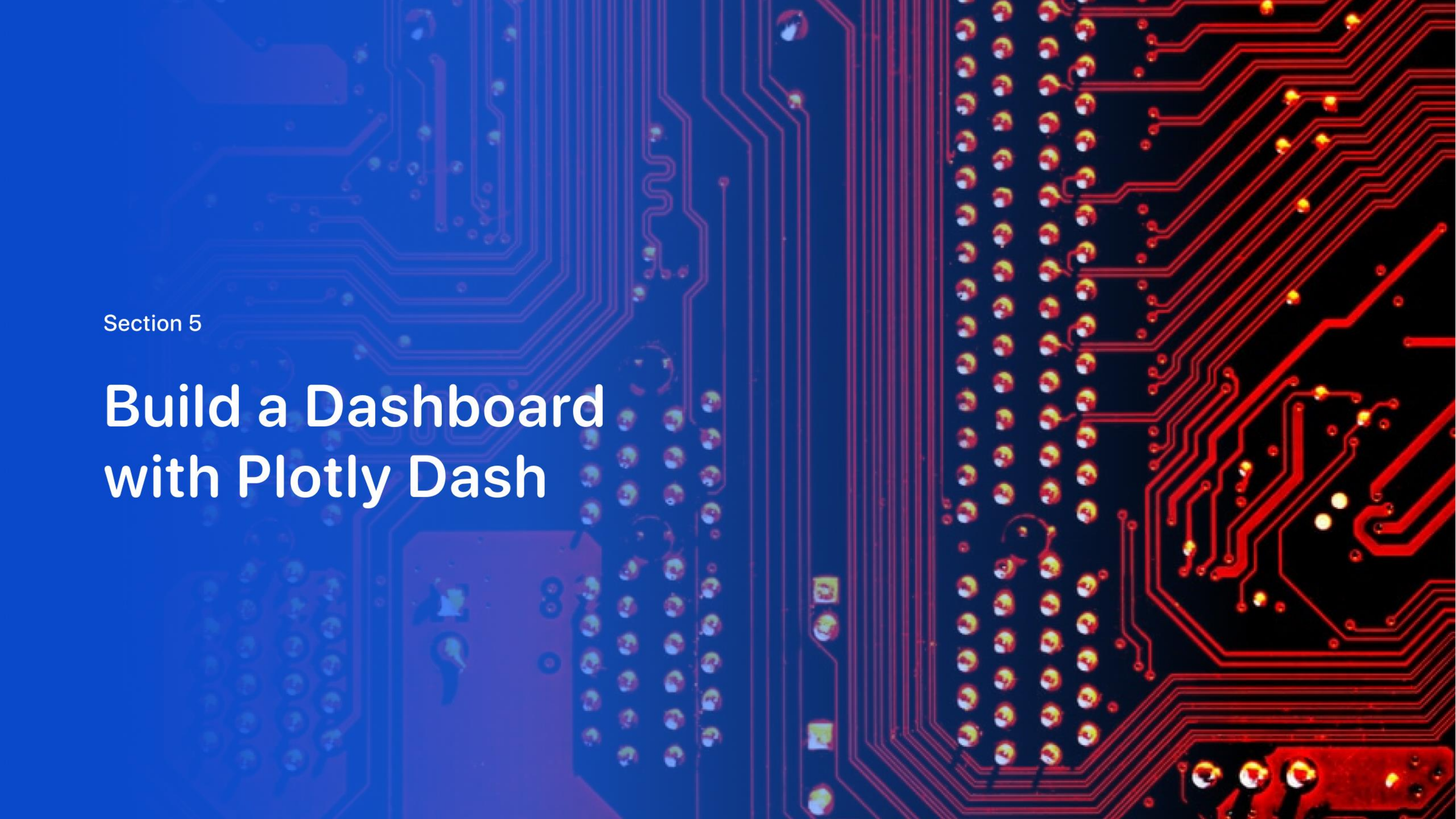
- In the following map we added markers to display the launch outcomes for each launch site.
- From the map we were able to visualize and locate which launch sites had the most successful launches.
- KSC LC 39 A


+
−
Vandenberg Space Force Base
Vandenberg Village
211
La Purisima Mission State Historic Park
Lompoc Airport
CA 246
Vandenberg State Marine Reserve
Lompoc
CA 1
Leaflet | Data by © OpenStreetMap, under ODbL.

Distance between launch sites & it’s proximities

- With folium we are also able to draw polygons, lines and other geometries on the map.
- Here, we calculated the distance between two points and then made a line in between them.
- As we can see, the following map shows us that there is airport (Lompoc Airport) within the 14.07km distance of the launch site – VAFE SLC 4E


+
−
Base
Lat: 34.53993 Long: -120.63126
211
211
La Purisima Mission State Historic Park
Lompoc Airport
CA 246
Vandenberg State Marine Reserve
10
VAFE SLC-4E
Lompoc
14.07 km
CA 1
Leaflet | Data by © OpenStreetMap, under ODbL.

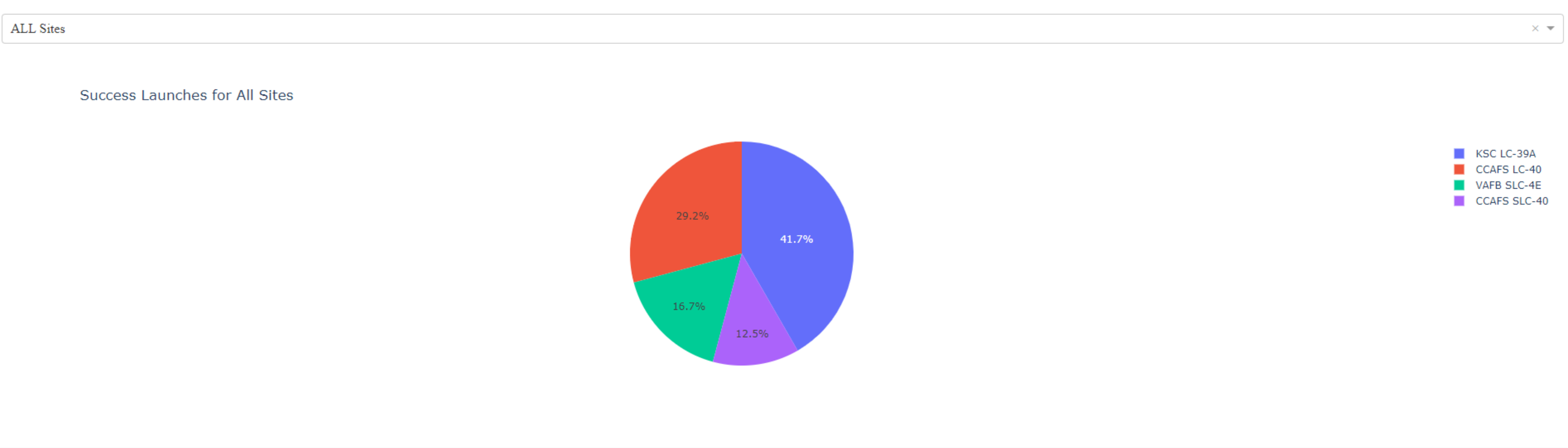


Section 5

Build a Dashboard with Plotly Dash

Success Rate for all Launch Sites

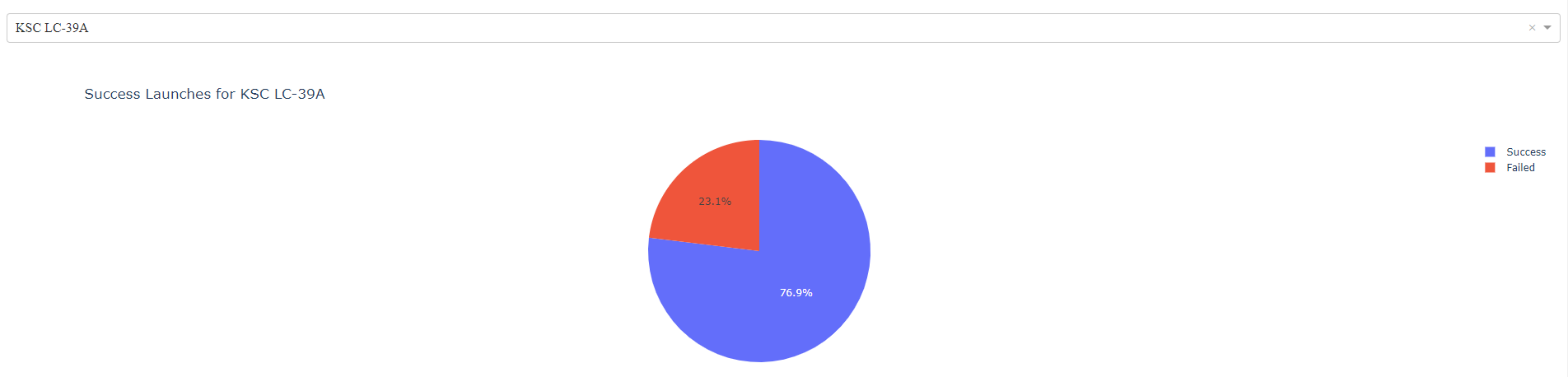
- The following is the pie chart showing the success ratio/percentage for all launch sites.
- We can see that Launch site KSC LC-39A has the highest success rate.
- Meanwhile CCAFS SLC 40 has the lowest success rate.



ALL Sites

Launch Site with Highest Success Ratio

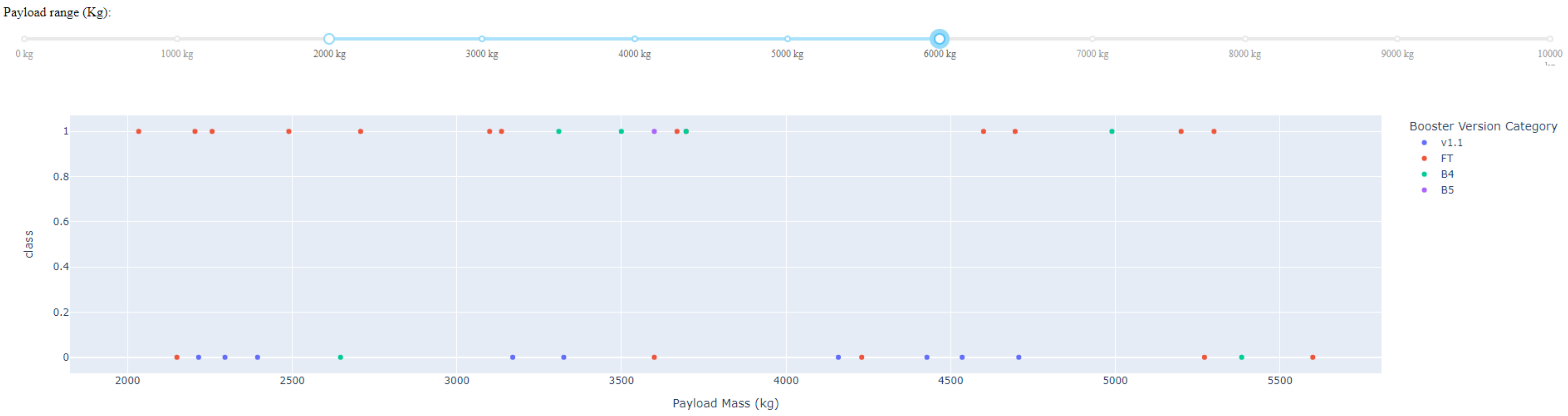
- Following is a pie chart for the launch site for highest success ratio built using the Plotly dashboard.
- The launch site is KSC LC 39-A which has a success rate of 76.9 % and a failure rate of 23.1 %.



KSC LC-39A

Payload vs Launch Outcomes

- The following scatter plot shows the relationship between launch outcomes and payload mass.
- We have payload mass range slider from which we can filter a range of payload mass for the scatter plot. Here we have selected a range between 2000kg and 6000kg
- The points are color coded based on the type of booster version used.
- We can see that many of v1.1 boosters have a failed launch outcome meanwhile FT boosters seems to be quite successful as compared to the others.

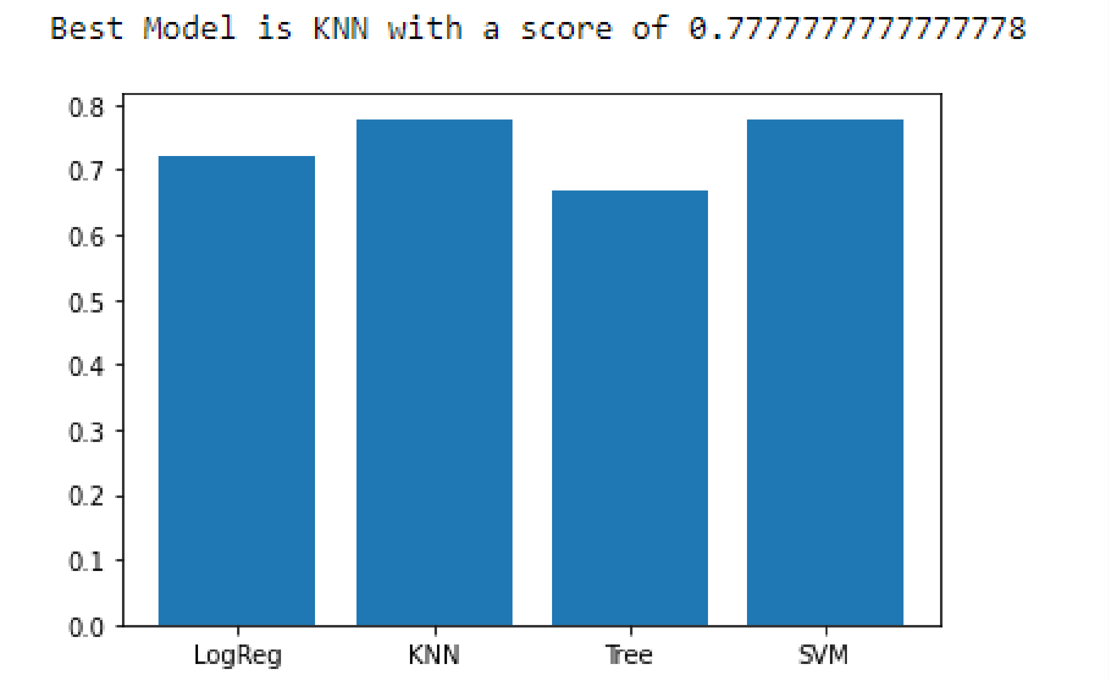


Section 6

Predictive Analysis (Classification)

Classification Accuracy

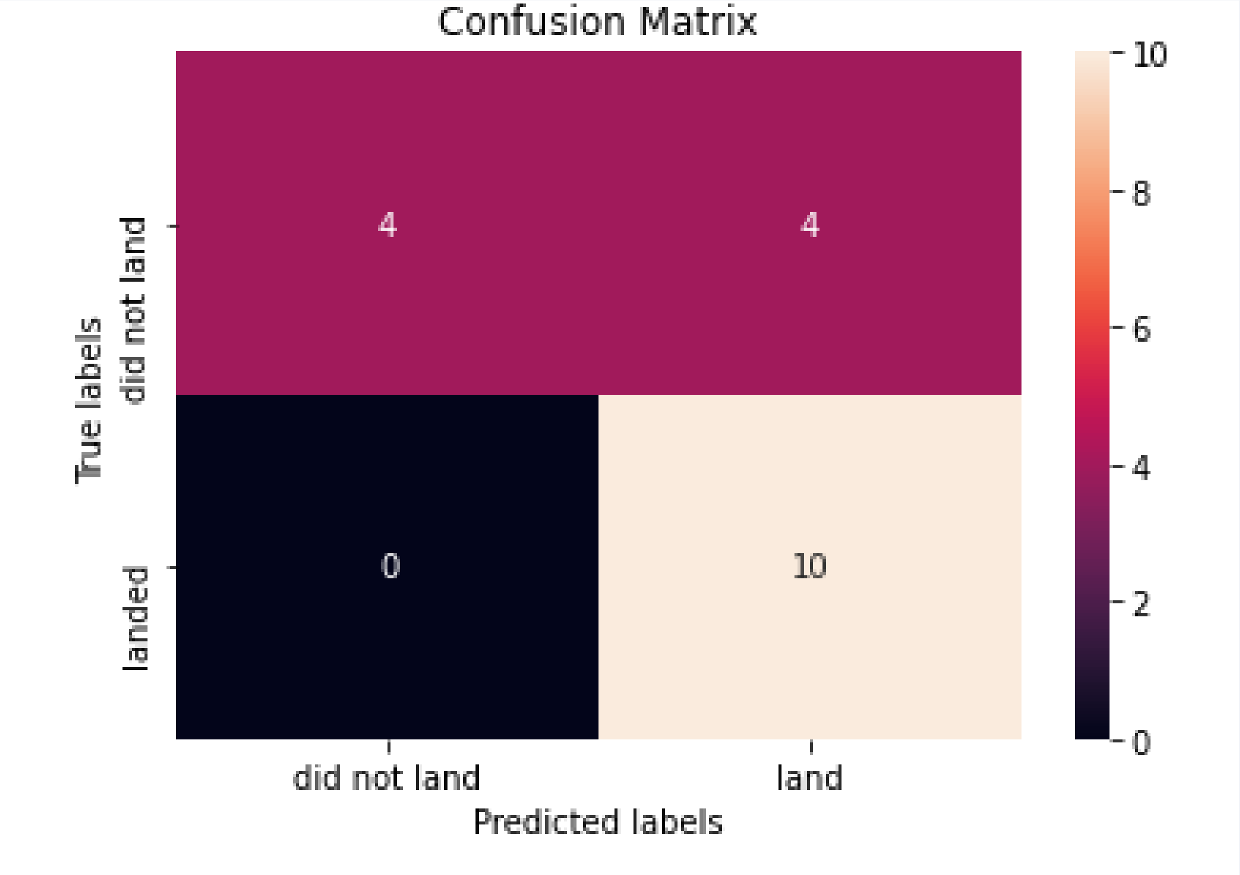
- All the models were trained and tested on the same dataset.
- As we can see the KNN model performed the best.
- The maximum accuracy we could get was 0.77 or 77%.



```
Best Model is KNN with a score of 0.7777777777777778
```

Confusion Matrix

- This is the confusion matrix produced for the KNN model.
- Out of the 10 flights that successfully landed the model was able to predict correctly all the 10.
- And, out of the 8 flights that did not land successfully, the mode was able predict only 4 correctly.



Conclusions

- We had aimed to build a predictive model with the help several input data and features from provided by Space X Falcon 9 launches.
- Upon comparison between different machine learning models, the KNN models was selected to be the best model for the given data set and input features.
- We can maybe improve the model accuracy by providing it with more data in future or maybe applying other machine learning techniques like neural networks.
- But in the end we were able to analyze and under stand the data and present it in a well structured manner with the help of dashboards and plots. These are essential for any data science related project before moving to building predictive models for the business.
- Hopefully our findings would be able to solve several business related questions for the organization and provide them with important insights as to how to proceed further.

Thank you!