



IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Randall Hidalgo Sánchez

2025-07-14

<https://github.com/randall-hidalgosanchez/IBM-Capstone>



Outline

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

Executive Summary

- Data was obtained through SpaceX API and complemented with information scrapped from Wikipedia's article on Falcon9 launches. Data was cleaned and standardized as preparation for each of the following steps. Different visualizations were plotted and models tested.
- The model that better fitted the data was a Decision Tree, however, that was a model with the best parameters as other models behaved better when using basic arguments.

Introduction

- 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.
- We will predict if the Falcon 9 first stage will land successfully.

Section 1

Methodology

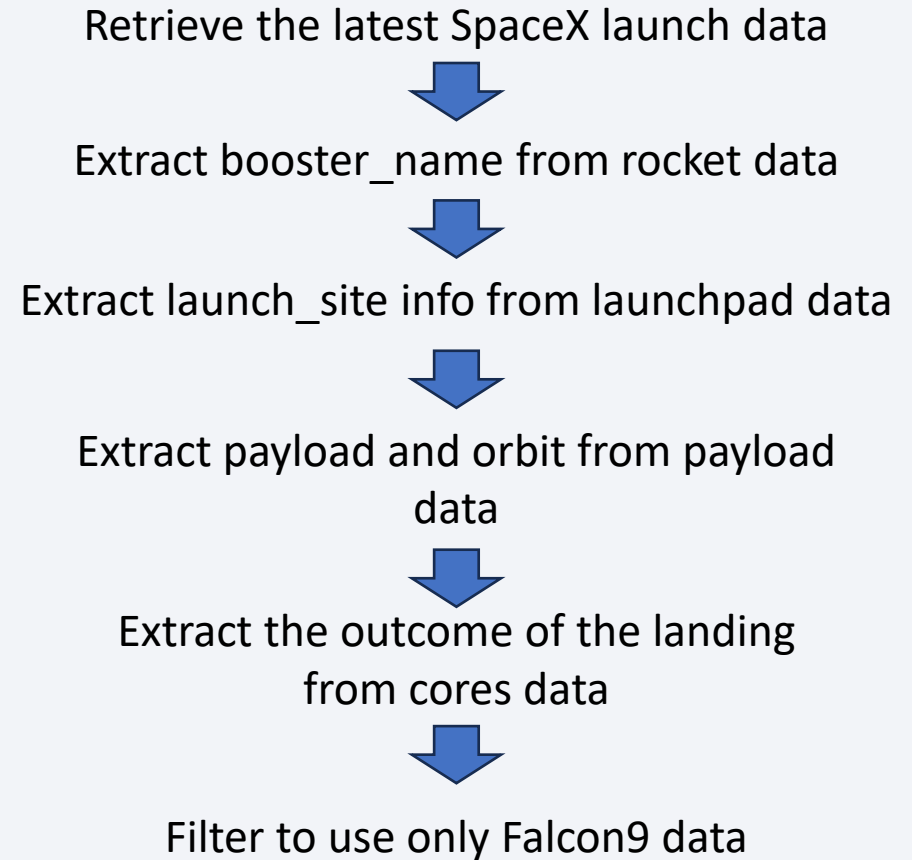
Methodology

Executive Summary

- Data collection methodology:
 - Data was obtained from SpaceX API and complemented with Falcon9 launches information scrapped from Wikipedia
- Perform data wrangling
 - Data was processed to obtain a variable “Outcome” with 1 meaning landing was successful and 0 meaning landing was not successful.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
 - Four models were tested: Logistic regression, SVM, Decision Trees and KNN.
 - Cross-validations was implemented to obtain the best parameters

Data Collection – SpaceX API

- Data was collected through SpaceX API
- <https://github.com/randall-hidalgosanchez/IBM-Capstone/blob/main/Scripts/01.%20jupyter-labs-spacex-data-collection-api.ipynb>



Data Collection - Scraping

- Get Falcon9 launch records from a HTML table in Wikipedia
- <https://github.com/randall-hidalgosanchez/IBM-Capstone/blob/main/Scripts/02.%20jupyter-labs-web scraping.ipynb>



Data Wrangling

- Data was first checked for null values
- Then, variables were checked to be the required type (numeric, categorical)
- Number of launches on each site were calculated
- Then, the number and occurrence of each orbit were also obtained
- Lastly, a new variable was created with the landing outcome labeled
- <https://github.com/randall-hidalgosanchez/IBM-Capstone/blob/main/Scripts/03.%20labs-jupyter-spacex-Data%20wrangling.ipynb>

EDA with Data Visualization

- The relation between flight number and payload mass was plotted using a scatterplot and colored with the outcome. This helps to see if the payload has any effect on the outcome.
- Then, flight numbers was plotted against launch site, again colored with launching outcome. This will help know if the site has any effect on the success of the landing.
- Launch site was plotted against payload mass. This helps to know if the mass in conjunction with site affects the outcome.
- Success rate was plotted for each orbit. This allows us to know if it is easier to land on certain orbits.
- Success rate was plotted through time to know how the success has changed along time.
- <https://github.com/randall-hidalgosanchez/IBM-Capstone/blob/main/Scripts/05.%20pandas-matplotlib-edadataviz.ipynb>

EDA with SQL

- A table was first created containing SpaceX info
- A query to obtain the name of each site was included
- Then, a snippet of the data was shown
- Total payload mass carried by boosters launched by NASA was calculated
- Also, average payload mass carried by booster version F9 v1.1
- Date of first successful landing was obtained
- Finally, the count of landing outcomes were listed and ranked
- https://github.com/randall-hidalgosanchez/IBM-Capstone/blob/main/Scripts/04.%20jupyter-labs-eda-sql-coursera_sqlite.ipynb

Build an Interactive Map with Folium

- A map of the NASA Johnson Space Center at Houston, Texas was plotted to give context on the location.
- Another map of the launching sites was added. This map also contains the outcome of each launch.
- The distance from those launching sites to certain sites of interest were calculated.
- https://github.com/randall-hidalgosanchez/IBM-Capstone/blob/main/Scripts/06.%20lab_jupyter_launch_site_location_interactive.ipynb

Build a Dashboard with Plotly Dash

- A dropdown menu for each site
- A pie chart showing the success rate
- A slider for the payload range
- A scatterplot showing the correlation between payload and launch success
- <https://github.com/randall-hidalgosanchez/IBM-Capstone/blob/main/Scripts/07.%20spacex-dash-app.py>

Predictive Analysis (Classification)

- Different models were tested to find which one fits better the data. Those models aim to predict successful launches.
- Those models are: logistic regression, support vector machine, decision tree, and K nearest neighbors
- A confusion matrix of the results was plotted for each model
- All models were compared utilizing their accuracy
- https://github.com/randall-hidalgosanchez/IBM-Capstone/blob/main/Scripts/08.%20SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

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

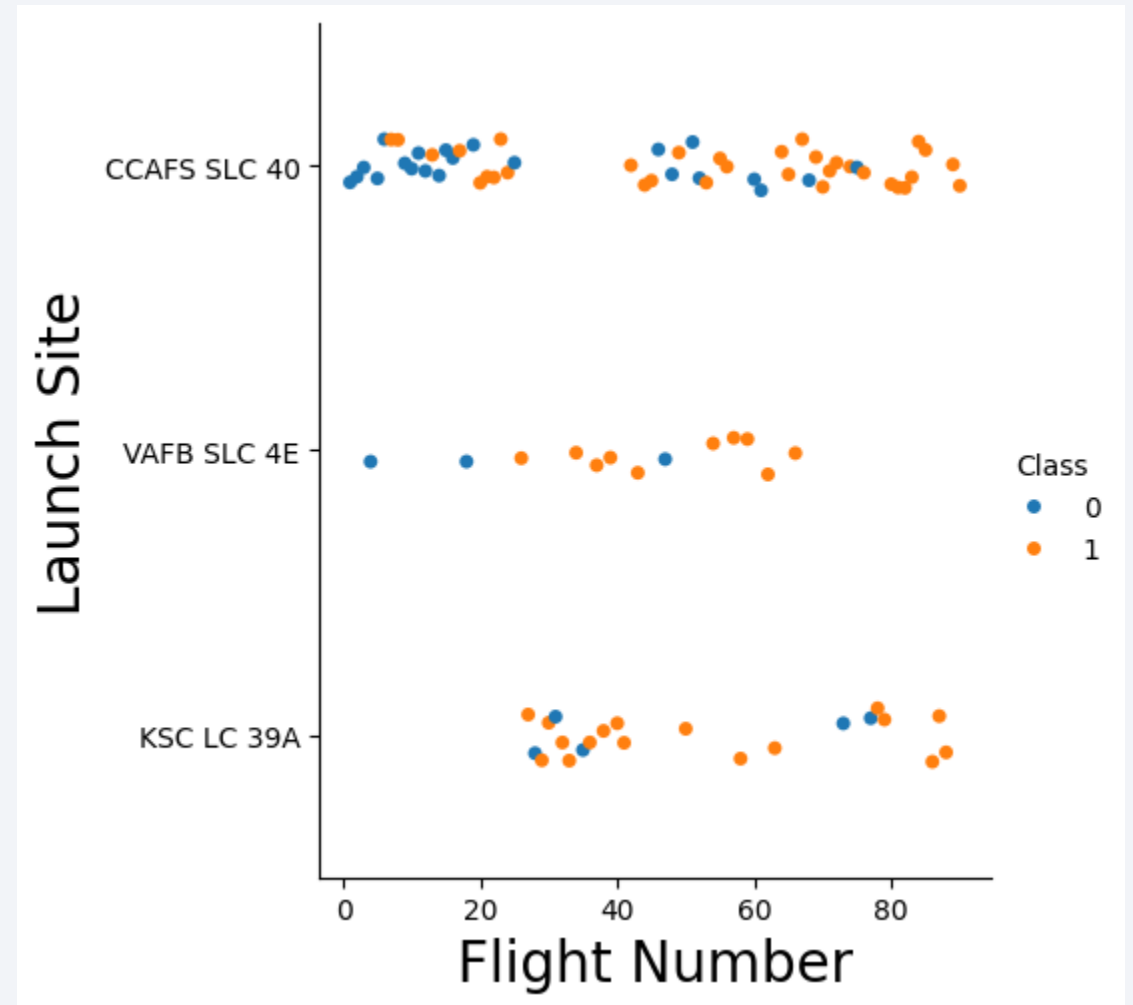
The background of the slide is an abstract composition. It features a dark blue base color. Overlaid on this are numerous diagonal streaks in shades of red and cyan. A faint, light blue grid pattern is also visible, particularly in the lower half of the image. The overall effect is dynamic and technological.

Section 2

Insights drawn from EDA

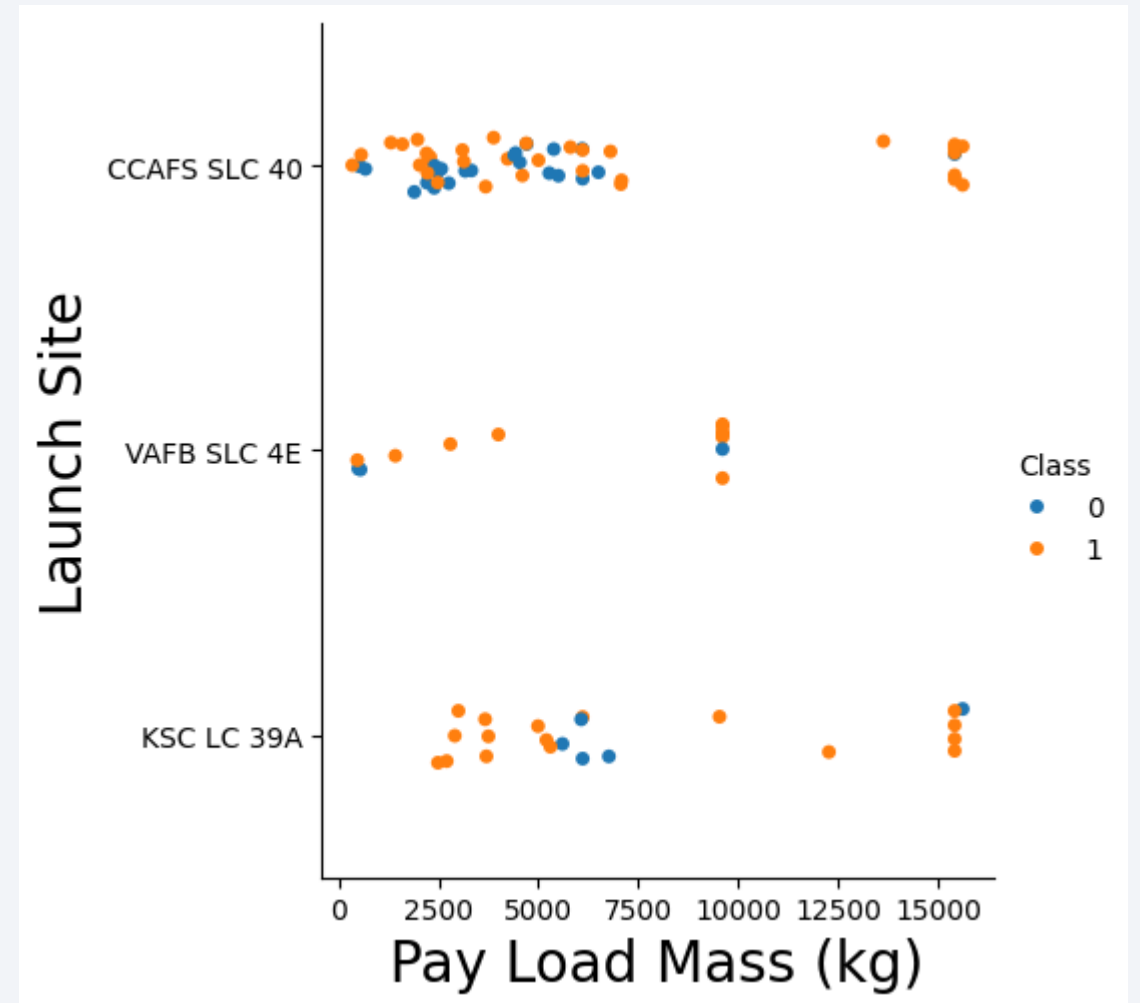
Flight Number vs. Launch Site

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



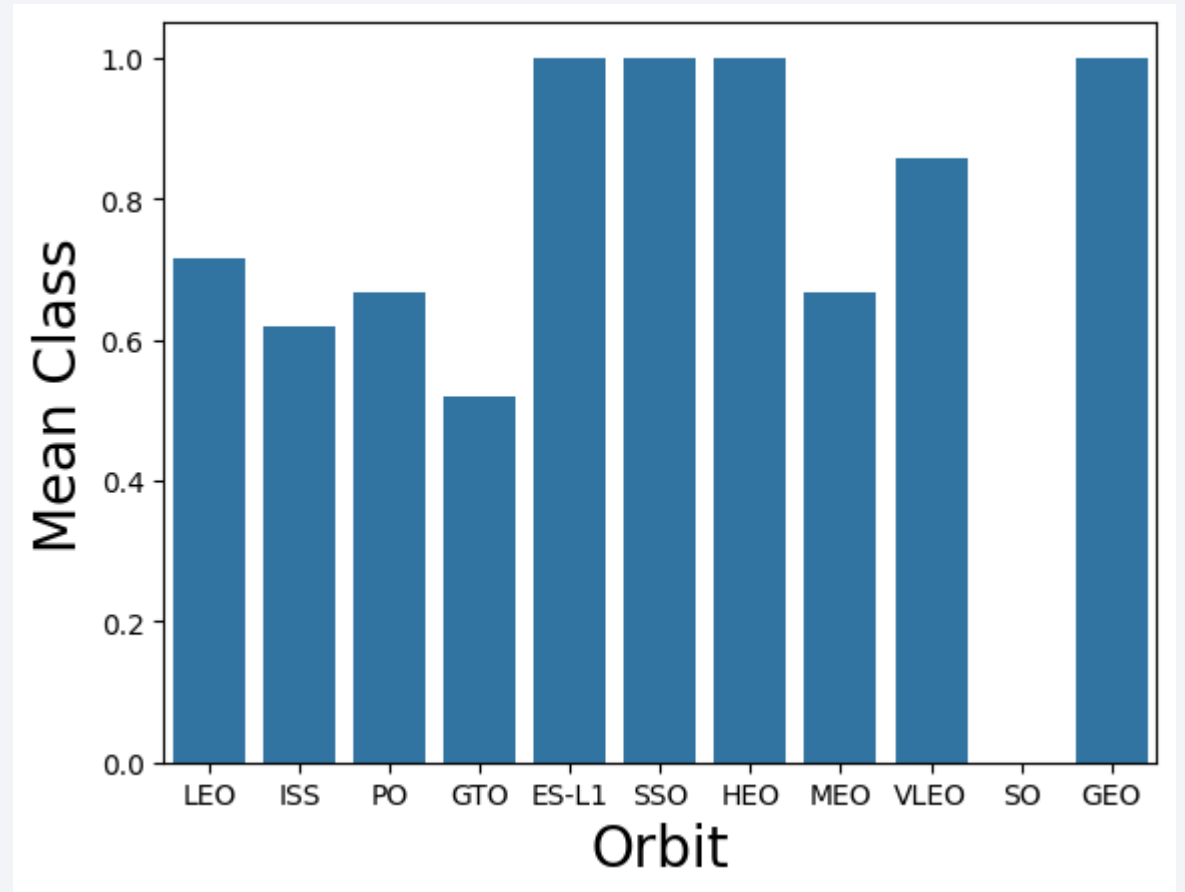
Payload vs. Launch Site

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



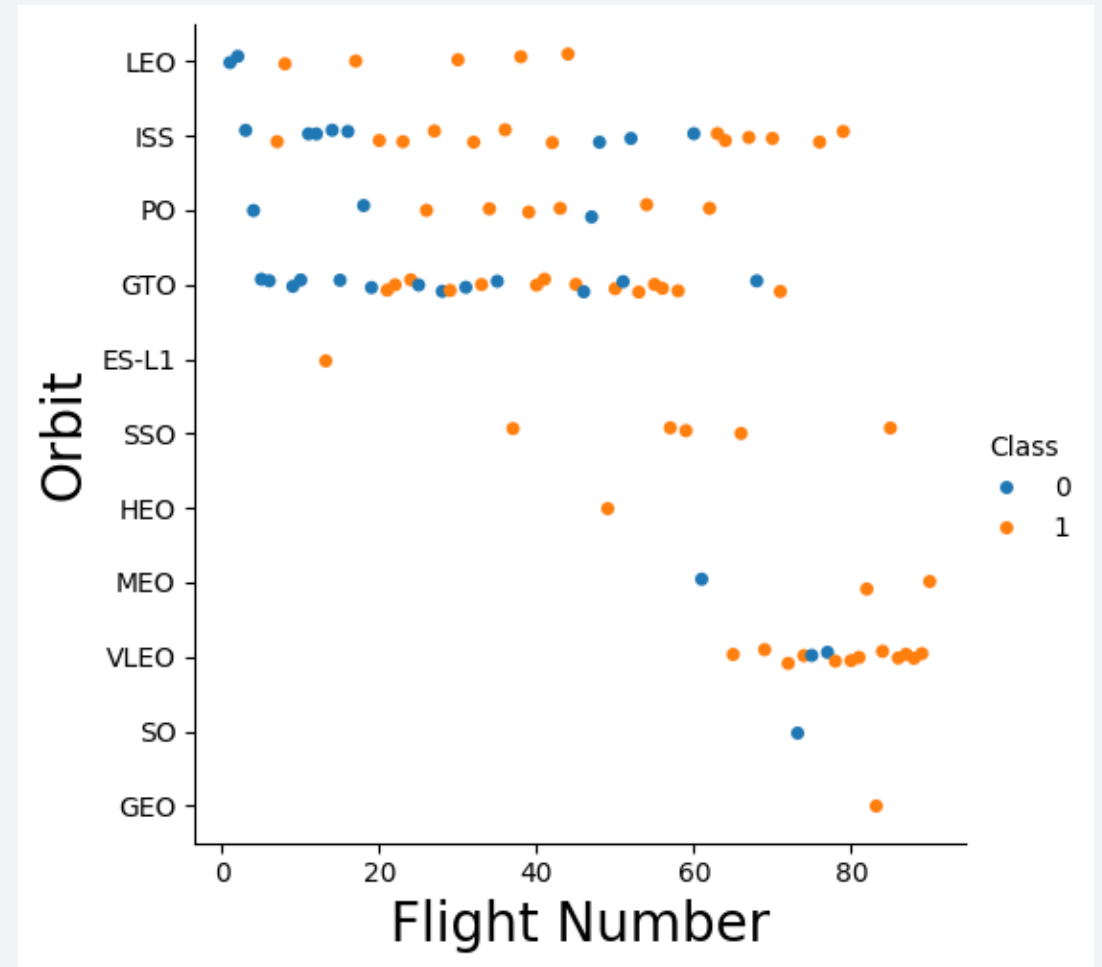
Success Rate vs. Orbit Type

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



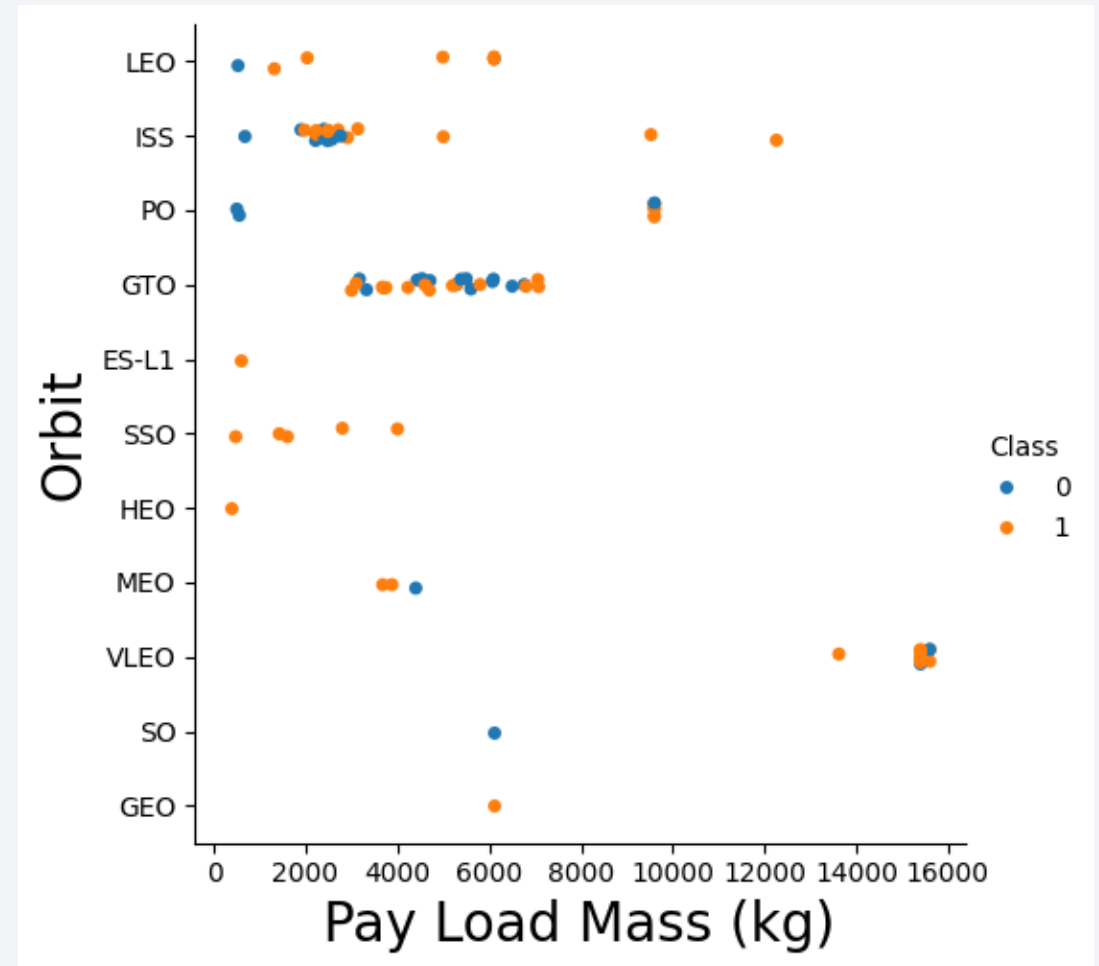
Flight Number vs. Orbit Type

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



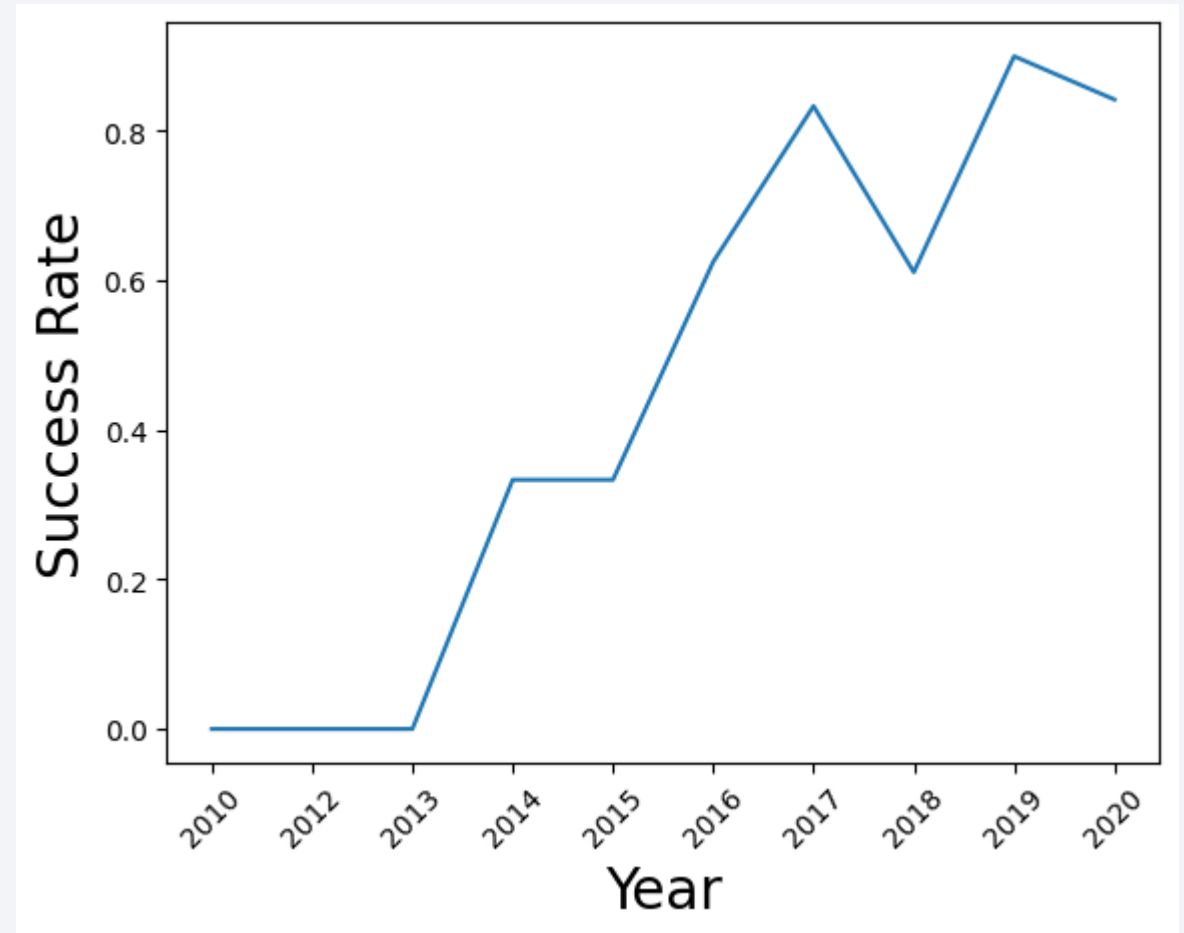
Payload vs. Orbit Type

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



Launch Success Yearly Trend

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



All Launch Site Names

- The following query takes the variable Launch_Site from the table SPACEXTABLE and shows the unique values.

```
%sql select distinct Launch_Site from SPACEXTABLE;

* 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'

- This code takes the variable Launch_Site from the table SPACEXTABLE and extracts 5 rows where Launch_Site is “CCAFS SLC-40”.

```
%sql select * from SPACEXTABLE where Launch_Site = 'CCAFS SLC-40' limit 5;
```

* sqlite:///my_data1.db
Done.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS_KG	Orbit	Customer	Mission_Outcome	Landing_Outcome
2017-12-15	15:36:00	F9 FT B1035.2	CCAFS SLC-40	SpaceX CRS-13	2205	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2018-01-08	1:00:00	F9 B4 B1043.1	CCAFS SLC-40	Zuma	5000	LEO	Northrop Grumman	Success (payload status unclear)	Success (ground pad)
2018-01-31	21:25:00	F9 FT B1032.2	CCAFS SLC-40	GovSat-1 / SES-16	4230	GTO	SES	Success	Controlled (ocean)
2018-03-06	5:33:00	F9 B4 B1044	CCAFS SLC-40	Hispasat 30W-6 PODSat	6092	GTO	Hispasat NovaWurks	Success	No attempt
2018-04-02	20:30:00	F9 B4 B1039.2	CCAFS SLC-40	SpaceX CRS-14	2647	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- The following query calculates the sum of PAYLOAD_MASS_KG for all rows from SPACEXTABLE table where Customer is “NASA (CRS)”

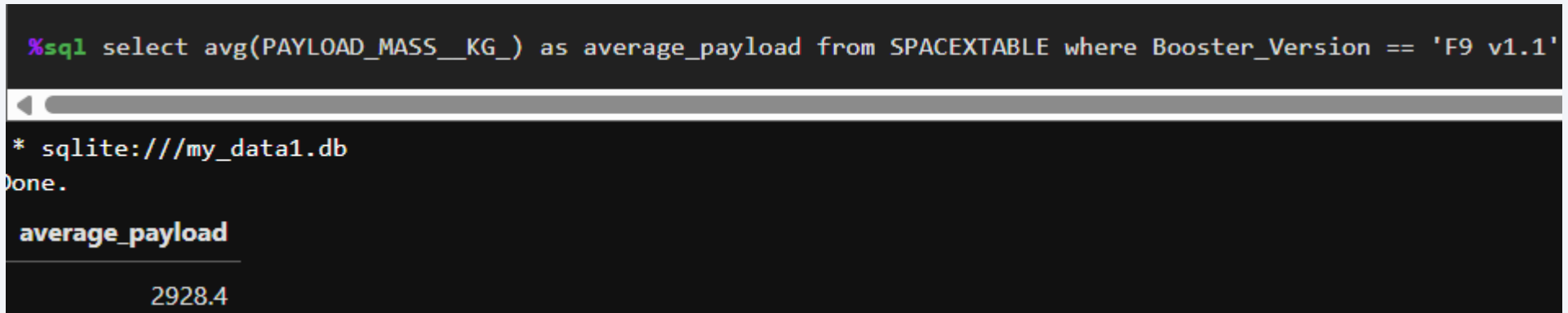
```
%sql select sum(PAYLOAD_MASS_KG_) as total_payload from SPACEXTABLE where Customer = 'NASA (CRS)'
```

```
* sqlite:///my_data1.db  
Done.  
  
total_payload  
-----  
45596
```

Average Payload Mass by F9 v1.1

- The following query calculates the average of PAYLOAD_MASS_KG for all rows from SPACEXTABLE table where Booster_Version is “F9 v1.1”

```
%sql select avg(PAYLOAD_MASS_KG_) as average_payload from SPACEXTABLE where Booster_Version == 'F9 v1.1'
```



The screenshot shows a SQLite terminal window with a dark background. The prompt is `* sqlite:///my_data1.db`. The user has entered the SQL query `select avg(PAYLOAD_MASS_KG_) as average_payload from SPACEXTABLE where Booster_Version == 'F9 v1.1'`. The terminal shows the output as a table with one column, `average_payload`, and one row with the value `2928.4`.

average_payload
2928.4

First Successful Ground Landing Date

- The following query finds the date of the first successful landing on the SPACEXTABLE dataframe
- %sql select min(Date) as first_successful_landing from SPACEXTABLE where Mission_Outcome = 'Success' and Landing_Outcome = 'Success (ground pad)' and Date is not null;
- The first successful landing was on 2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- %sql select distinct Booster_Version from SPACEXTABLE where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000;

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

Total Number of Successful and Failure Mission Outcomes

- %sql select Mission_Outcome, count(Mission_Outcome) as total from SPACEXTABLE group by Mission_Outcome;

Mission_Outcome	total
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

```
%sql select Booster_Version from SPACEXTABLE where PAYLOAD_MASS_KG_ = (select max(PAYLOAD_MASS_KG_) from SPACEXTABLE);
```

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

```
Done.
```

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

2015 Launch Records

- %sql select substr(Date, 6,2) as month, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTABLE where substr(Date,0,5)='2015' and Landing_Outcome = 'Failure (drone ship)' order by month;

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

- %sql select Landing_Outcome, count(Landing_Outcome) as total from SPACEXTABLE where Date between '2010-06-04' and '2017-03-20' group by Landing_Outcome order by total desc;

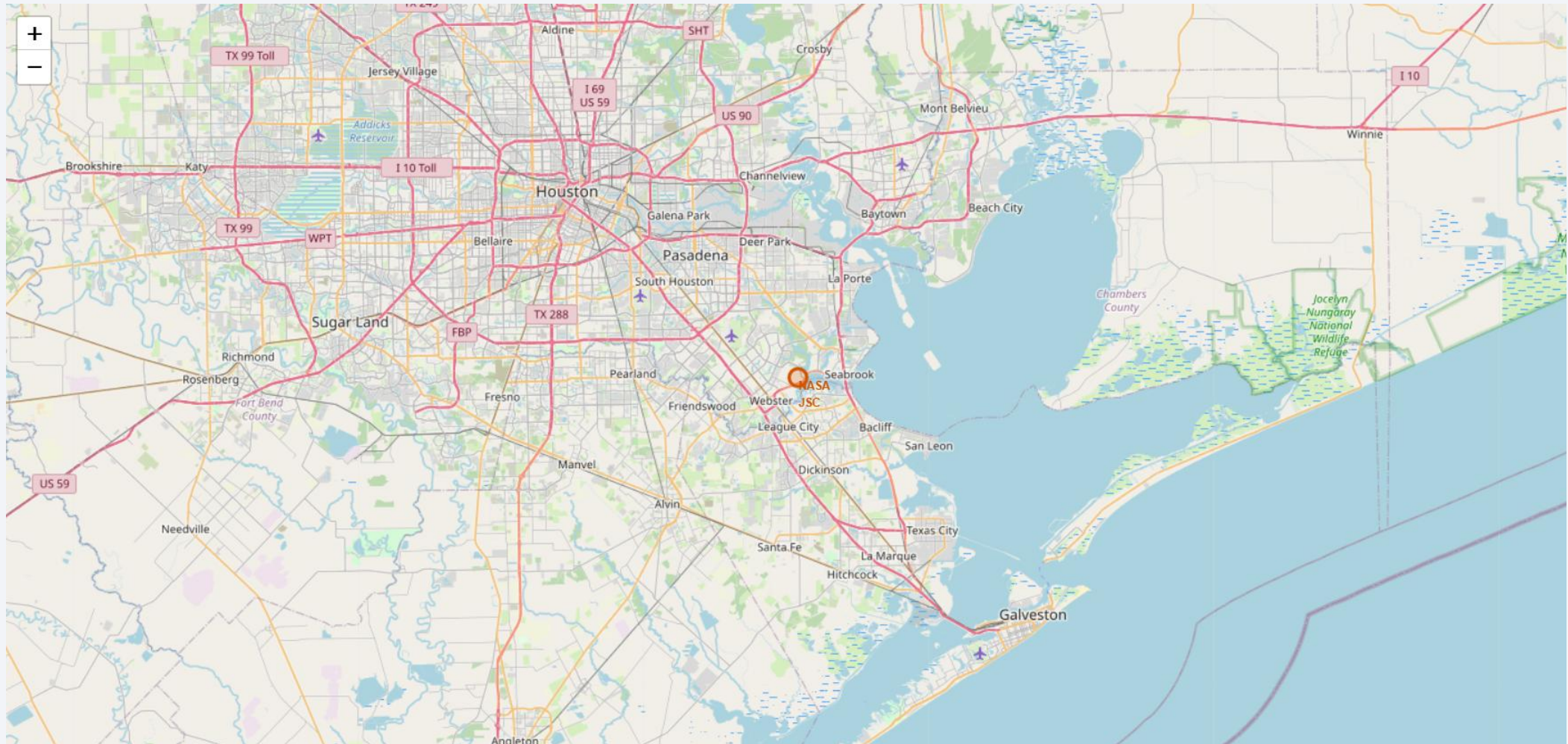
Landing_Outcome	total
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The image is a composite of a solid blue rectangle on the left and a satellite photograph of Earth on the right. The Earth's surface is dark, with numerous bright yellow and orange lights representing cities and urban areas. The horizon of the Earth is visible, separating the dark surface from the deep blue of the atmosphere and the blackness of space.

Section 3

Launch Sites Proximities Analysis

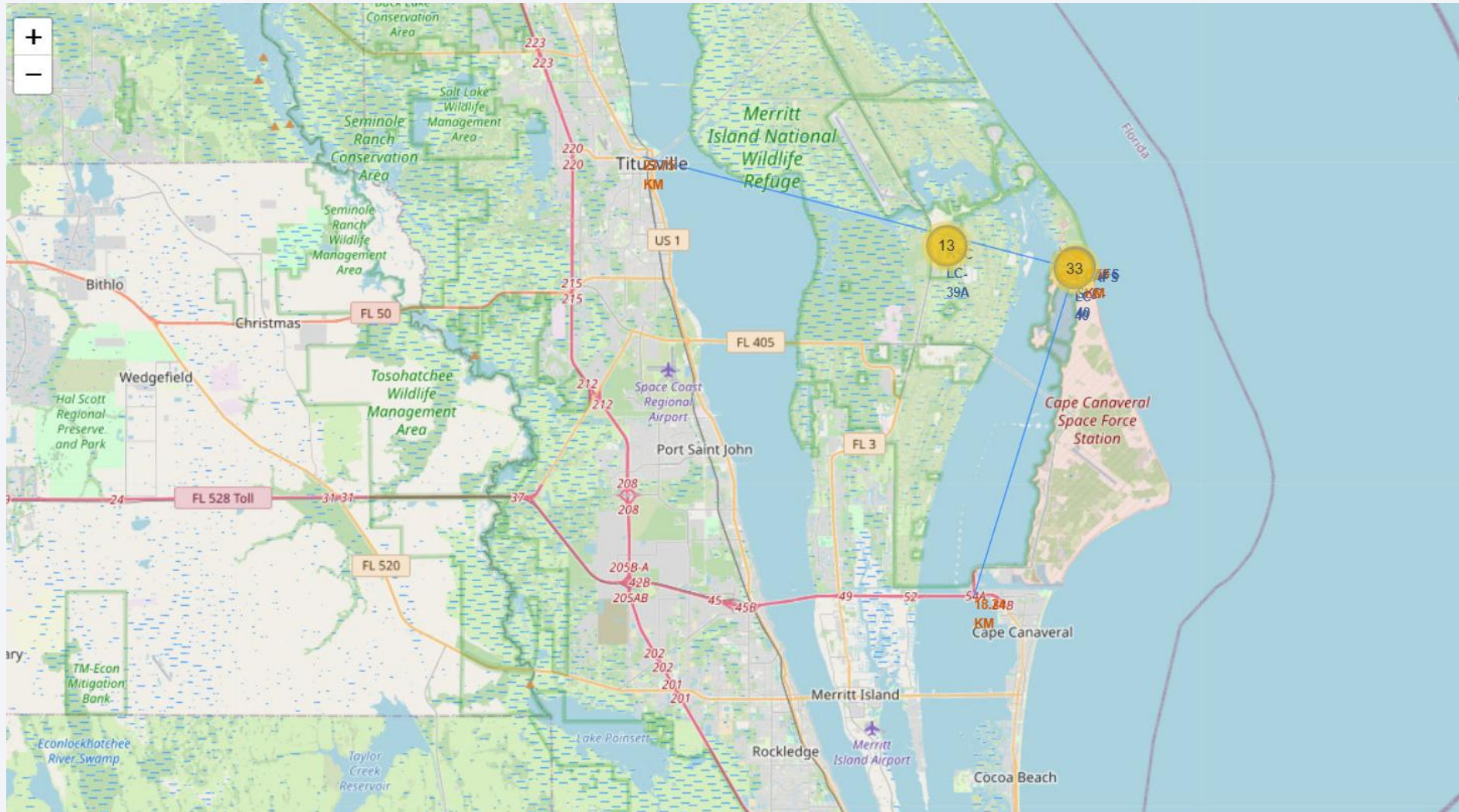
NASA Houston, Texax



Launching sites



Distance to sites of interest



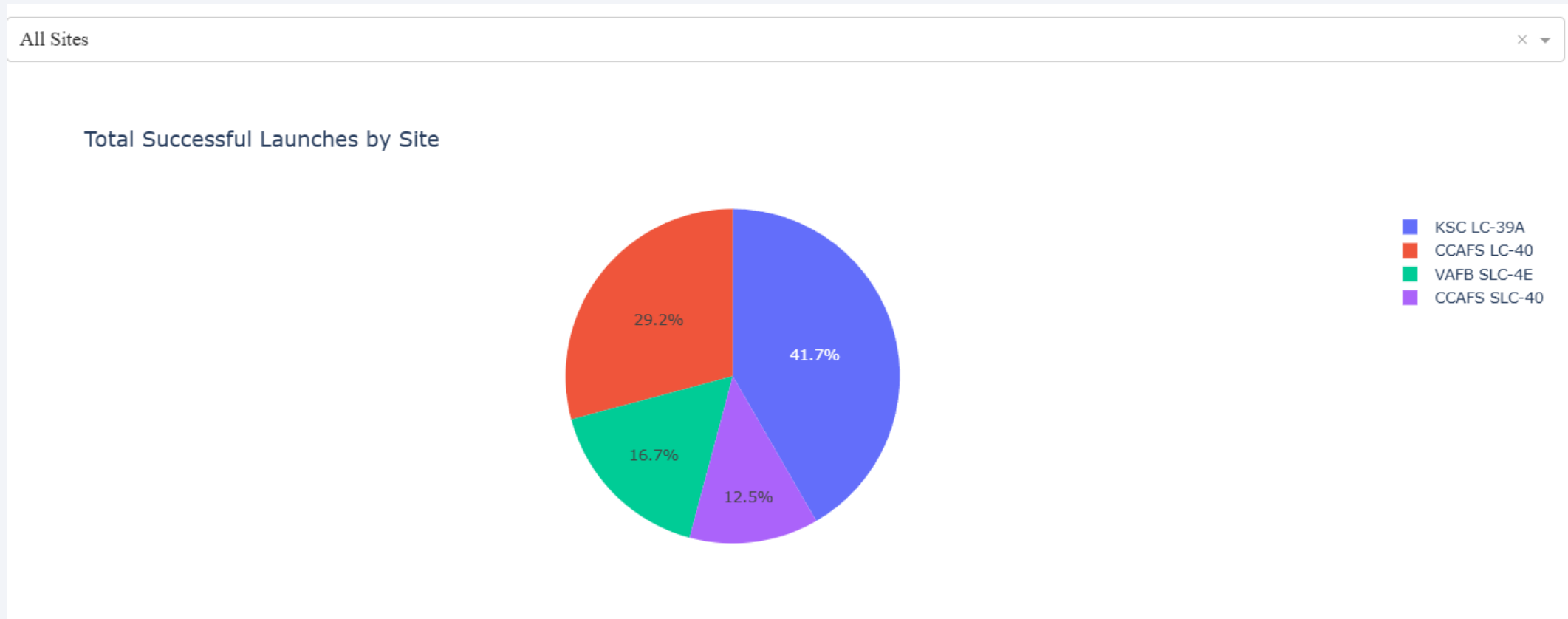


Section 4

Build a Dashboard with Plotly Dash

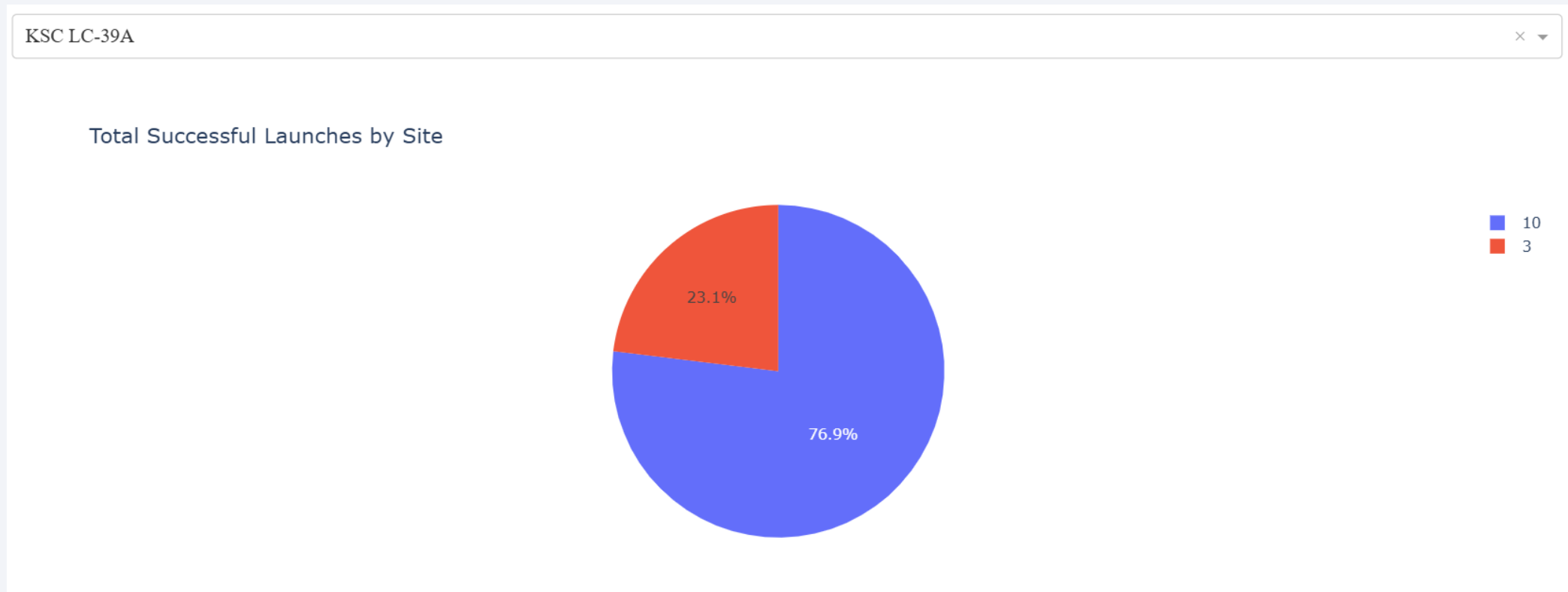
Successful Launches

- In the following plot we can observe the percentage of successful launches by launching site. The location with the highest success is KSC LC-39A



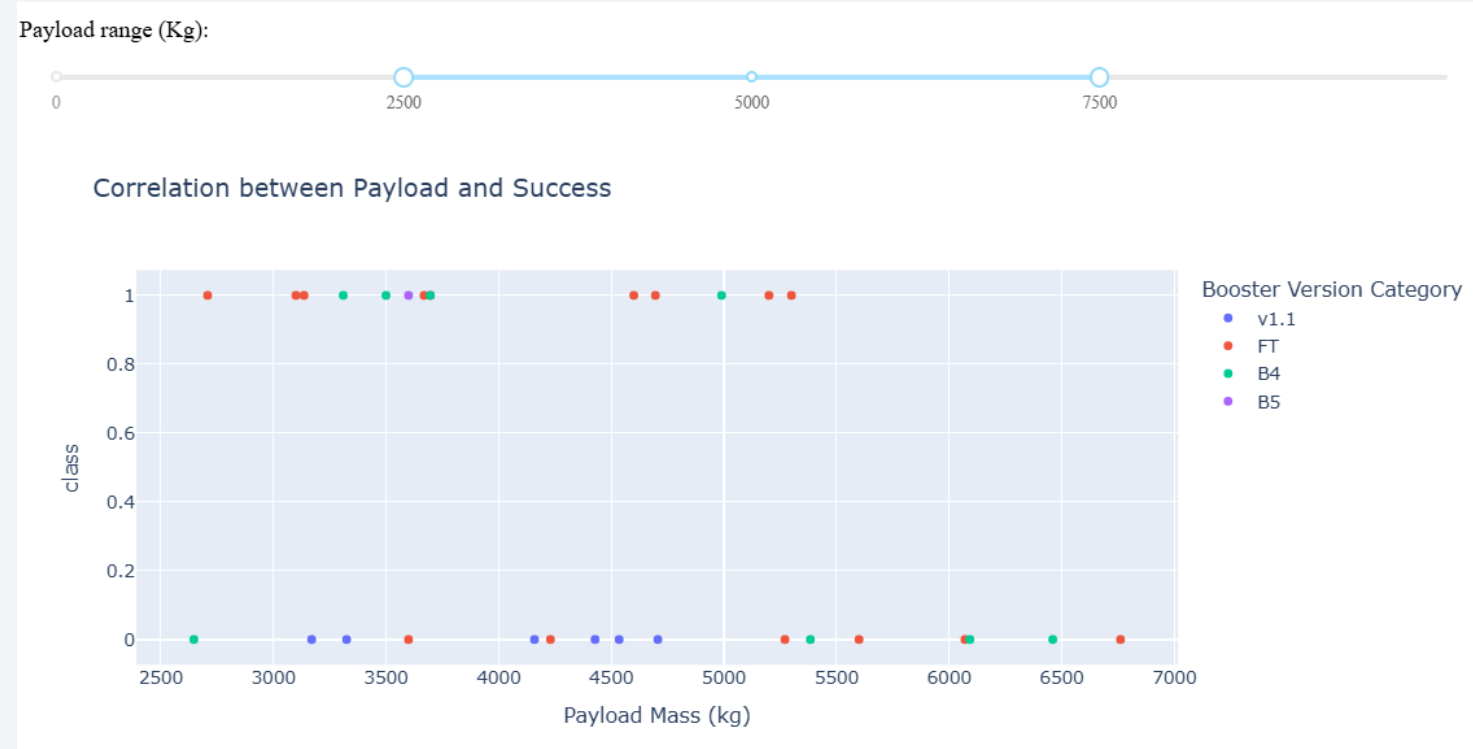
Successful Launches on KSC LC-39A

- Following the previous plot, here we observe the proportion of successful launches on KSC LC-39A. 76.9 % of the launches were successful.



Payload vs Success

- Here we observe the correlation between Payload and Success, using the payload range between 2500-7500 (Kg).



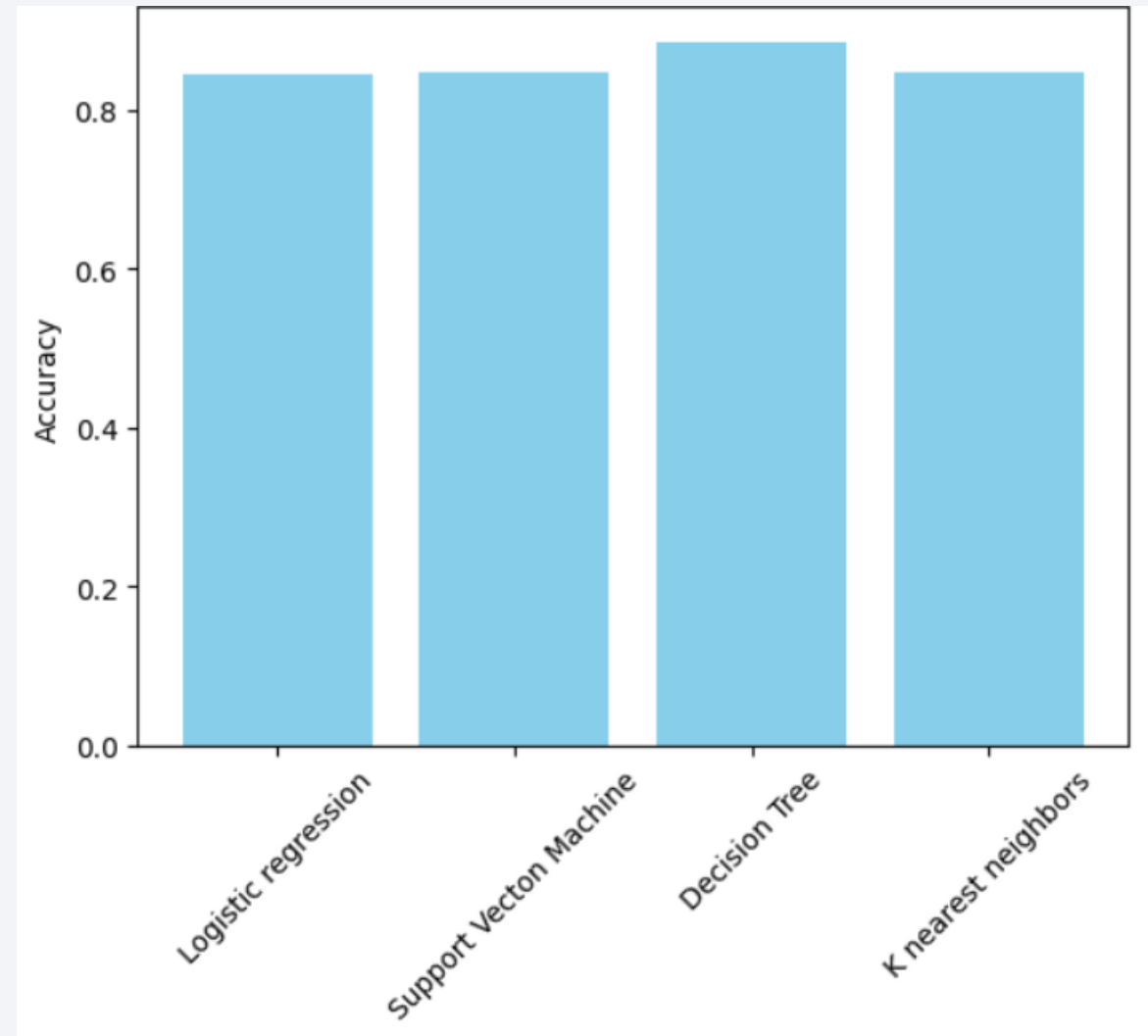


Section 5

Predictive Analysis (Classification)

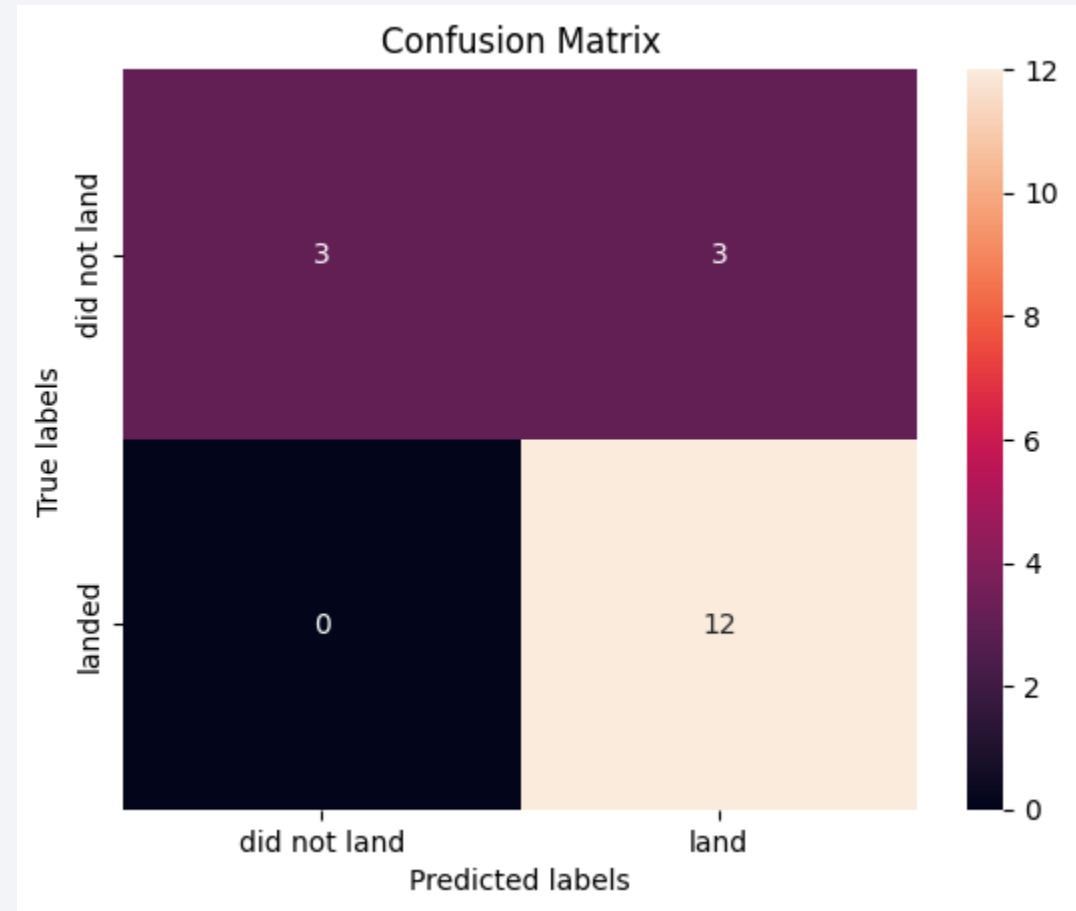
Classification Accuracy

- The plot compares the accuracy of all models. For this comparison the model used is the one with the best parameters. Based on this plot, the decision tree model has the higher accuracy.



Confusion Matrix

- The confusion matrix shows that the model predicted correctly 15 of the 18 samples in the test data.
- Only 3 samples that did not landed were predicted wrongly (as landed)



Conclusions

- SPACEX API and Falcon9 Wikipedia were useful for predicting successful landings
- Plotly dashboard is appropriate to show data
- Decision Tree (with adequate parameters) proved the best model

Appendix



Skills
Network

SpaceX Falcon 9 first stage Landing Prediction

Thank you!

