IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Denis Kytschakov

2023-11-10

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection (official Spacex Data and Web scraping)
  - Data cleaning
  - Data exploration using visualization of feature correlation
  - Feature extraction
  - One-Hot-Coding
  - Split data into train and test sets
  - Fit train data and make predictions
- Summary of all results
  - It could be shown, that the total number successful landings has increased over the years
  - A relationship between Payload, Launch Site and targeted Orbit which affects successful landing
  - Using this data a successful landing of Flacon 9 first stage could be predicted with 83.33 % accuracy

Introduction

- **Project background and context**:
  - 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.
  - The problem to solve was a creation of data set and applying machine learning algorithms to estimate the probability of a successful landing

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from official SpaceX source and web scraping from wikipedia
- Perform data wrangling
  - Data has been examined to find some pattern in it what could be label for training of supervised models. We labeled different outcomes with just 0 or 1 and created an extra column in the data set called “class”.
- Perform exploratory data analysis (EDA) using visualization and SQL
  - We used SQL queries, map visualization and interactive graphs to explore data to see patterns and determine dependencies.

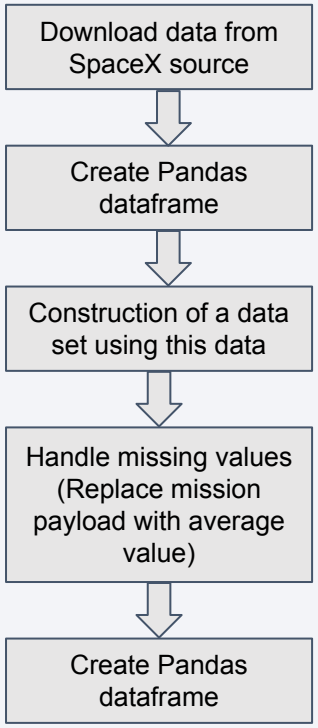
Methodology

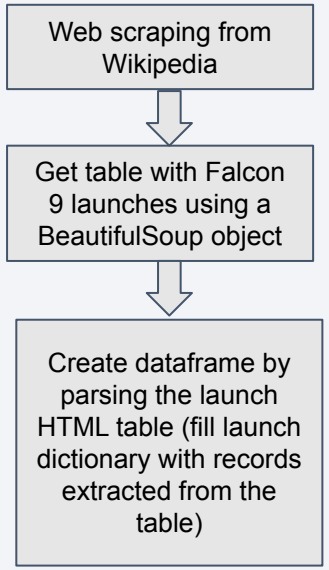
Executive Summary

- Perform interactive visual analytics using Folium and Plotly Dash
  - For the visualization we used Folium for creating a map with markers indicating launch sites. The markers had different colors depending on landing success.
  - Using Plotly we then created an interactive pie plot showing the landing success for different launch sites which could be chosen from a dropdown menu. We also created a slide bar to choose payload range for each launch site and show it as a scatter plot.
- Perform predictive analysis using classification models
  - We then used a pipeline to train different types of models: 1. Logistic Regression,

    2. Support-Vector-Machine, 3. k-Nearest-Neighbors, 4. Decision tree. We splittet the dataset into training and test parts, created grid search to find best parameters for our models, performed training and validated on test data. As a result we were able to predict the landing outcome with a 83.33 % accuracy

Data Collection



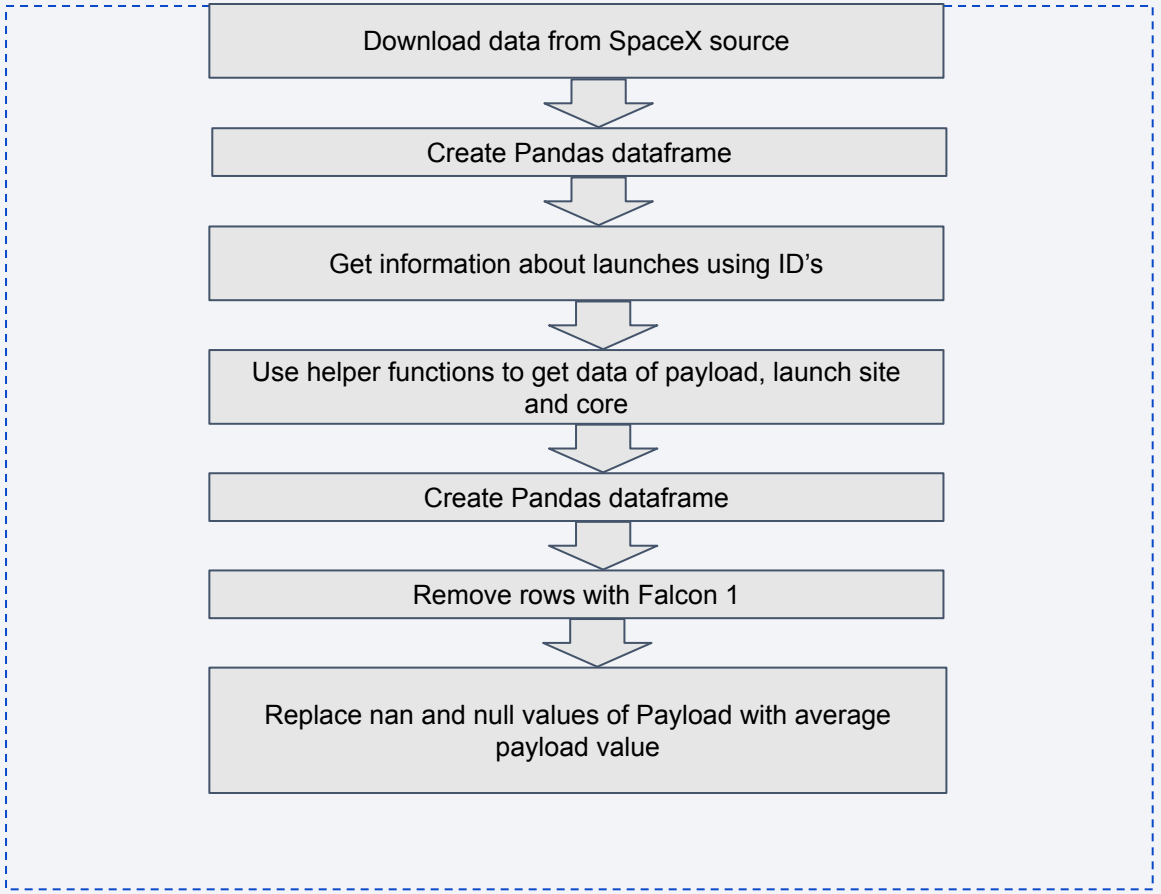


This two methods were used to get data related to launches of Falcon 9 rockets.

The obtained data has been processed to a data set that can be used for further examination.

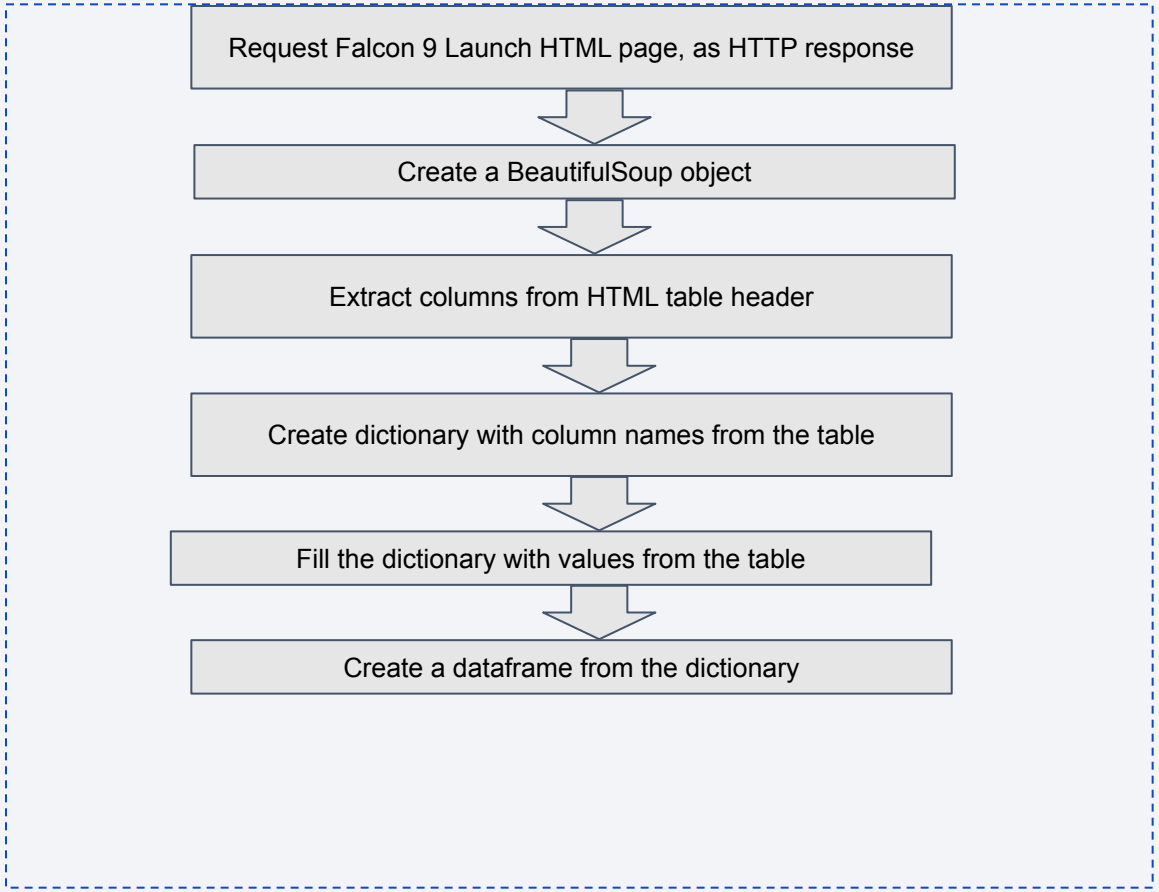
Data Collection – SpaceX API

- GitHub URL of the completed SpaceX API calls notebook:
  - Capstone_Project/LAB1_Filter_Data_Create_Dataset.ipynb at main · Denis19068811/Capstone_Project (github.com)



Data Collection - Scraping

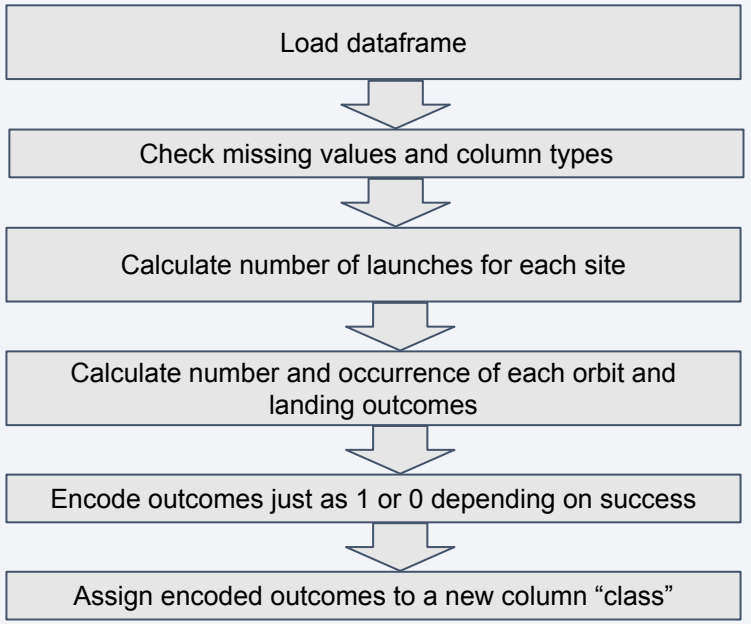
- BautifulSoup object was used for web scraping.
- A table with Falcon 9 launches was extracted from Wikipeida
- Dictionary was created with column names from the table
- Dictionary was filled with data from the table
- A dataframe was created from the dictionary



- GitHub URL of the completed web scraping notebook: Capstone_Project/LAB1.2_Web_Scraping.ipynb at main · Denis19068811/Capstone_Project (github.com)

Data Wrangling

- Previously created dataframe was used to calculate number of launches on each site.
- Then number and occurrence of each orbit were calculated
- Then number and occurrence of mission outcome were calculated
- Different outcomes were coded with 0 or 1 and assigned to a new column “class”



- GitHub URL of completed data wrangling:

  Capstone_Project/LAB2_Data_Wrangling.ipynb at main · Denis19068811/Capstone_Project (github.com)

EDA with Data Visualization

- A scatter plot was performed to display flight number against payload
- Relationship between Lauch Site and Flight Number was plotted using bar plot
- A scatter plot was created to show relationship between Launch Site and Flight Number. A distinction was made between successful and unsuccessful landings
- Another scatter plot shows relationship between Launch Site and Payload and how these parameters affect landing success
- A bar plot shows relationship between targeted orbit and success rate
- Another scatter plot shows relationship between Orbit and Flight Number colored depending on landing being successful or not
- There is another similar plot but with Orbit against Payload
- The next plot is a line plot indicating the trend of successful landings from 2013 until 2020
- After doing all these steps to examine data, the dataset was one-hot-coded to eliminate all categorical data and make it suitable for machine learning models

- GitHub URL of your completed EDA with data visualization notebook:
  Capstone_Project/LAB3_Explore_and_Preparing_Feature_Engineering.ipynb at main · Denis19068811/Capstone_Project (github.com)

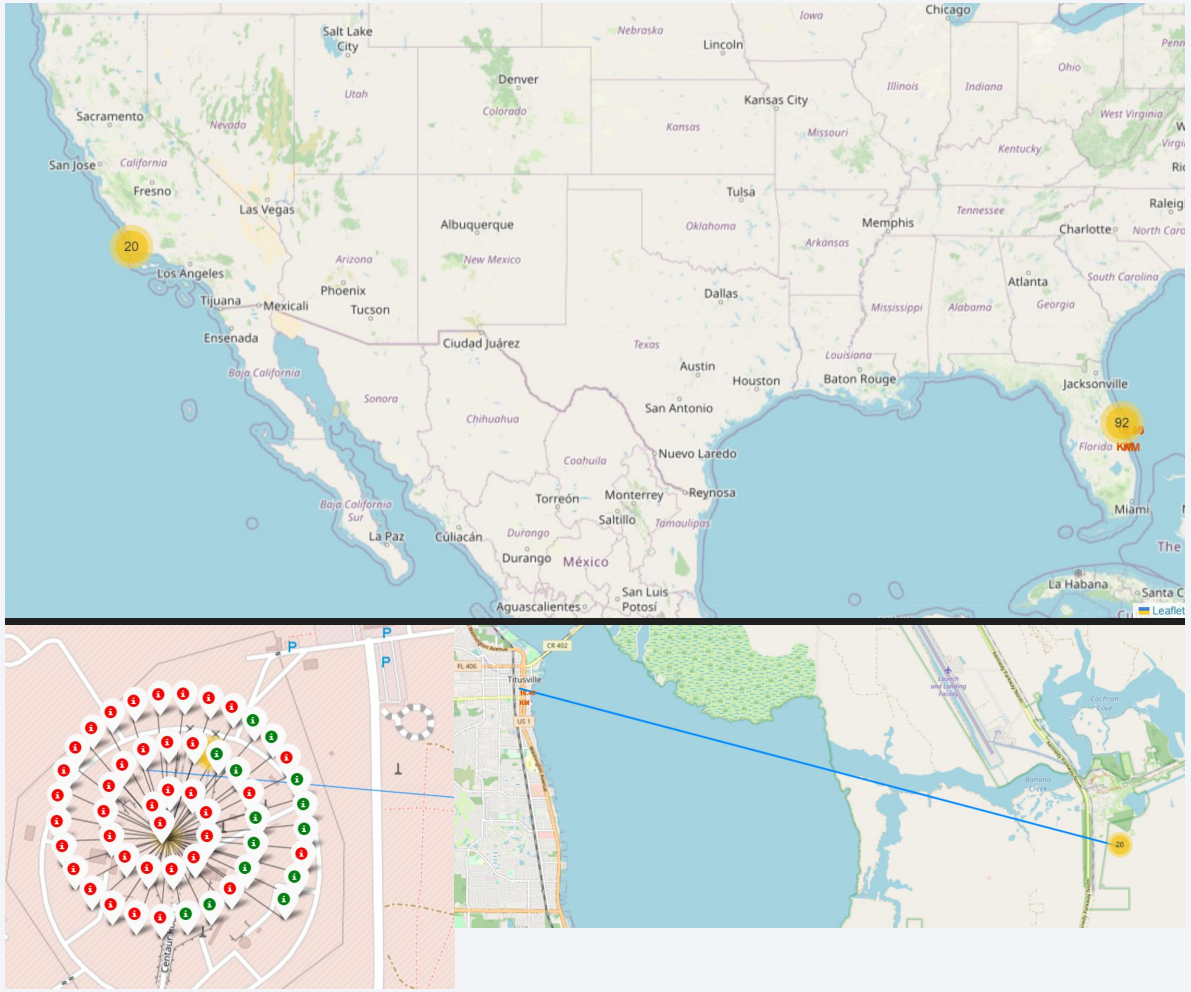
EDA with SQL

- Show the header of the table
- Show 5 records where launch site begin with “CCA”
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass
- List the records which display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months 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

- GitHub URL of your completed EDA with SQL notebook:
  Capstone_Project/LAB2_Data_Exploration_SQL.ipynb at main · Denis19068811/Capstone_Project (github.com)

Build an Interactive Map with Folium

- A map plot was created using folium object and coordinates of launch sites
- Different marker types were created and grouped into clusters that indicate successful and unsuccessful landings for each launch site.
- The distance from launch site to the nearest coastline, city etc. was calculated
- Doing all these steps helped to visualize and analyze the relationship between launch position and successful landings since trajectory resulting from objects near launch position may affect the success rate.
- GitHub URL of your completed interactive map with Folium map:
  Capstone_Project/LAB4_Interactive_Plot.ipynb at main · Denis19068811/Capstone_Project (github.com)



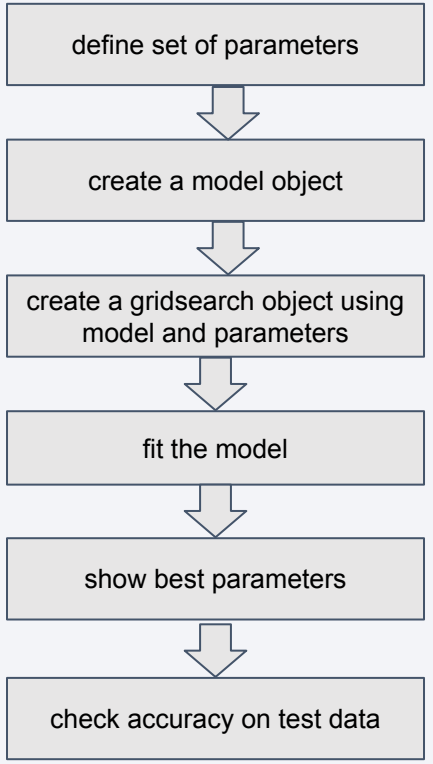
Build a Dashboard with Plotly Dash

- A dropdown menu was added to the dashboard to provide selection between launch site. It was also possible to choose all of them
- A pie chart was added, which displayed successful landings for chosen launch site
- A slide bar was added for selecting the range of payload
- Scatter plot was added to show landing outcomes depending on launch site and payload range

- These plots enable the used to review data interactively and choose values and areas of interest only
- GitHub URL of your completed Plotly Dash lab: Capstone_Project/spacex_dash_app.py at main · Denis19068811/Capstone_Project (github.com)

Predictive Analysis (Classification)

- Dataset was divided into X and Y (contains classes 0 or 1)
- As preprocessing step StandardScaler was applied to X
- Next the data was splitted into train and test sets (X_train, Y_train, X_test, Y_test)
- 4 models were trained: Logistic Regression, Support Vector Machine, Decision Tree, k-Nearest-Neighbours.
- For each model a set of parameters was created and GridSearch was used to find optimal ones
- Finally the models were fit using Data (X), Labels (Y) and GridSearch object. The output was a set of optimal parameters
- To show the results score and hitmap were displayed
- Scores of all 4 model were printed to determine the best performing model

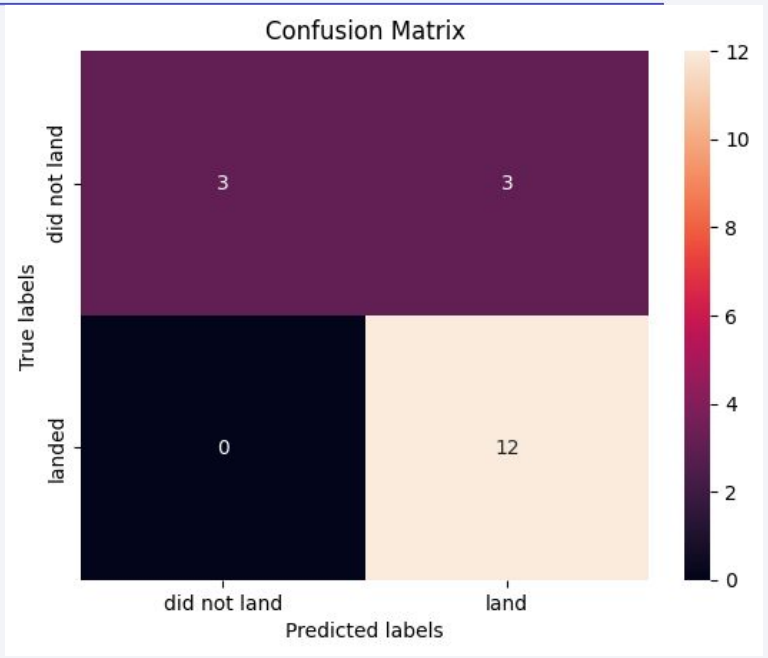


Model developing process

- GitHub URL of your completed predictive analysis lab: Capstone_Project/LAB5_Pipeline_for_Machinelearning.ipynb at main · Denis19068811/Capstone_Project (github.com)

Results

- Exploratory data analysis results:
  - Successful landings increase over time
  - There is a relationship between Orbit, Payload, Launch Site and successful landings



- Predictive analysis results
  - Accuracy of 83.33% could be achieved
  - The main problem is the prediction of false positives

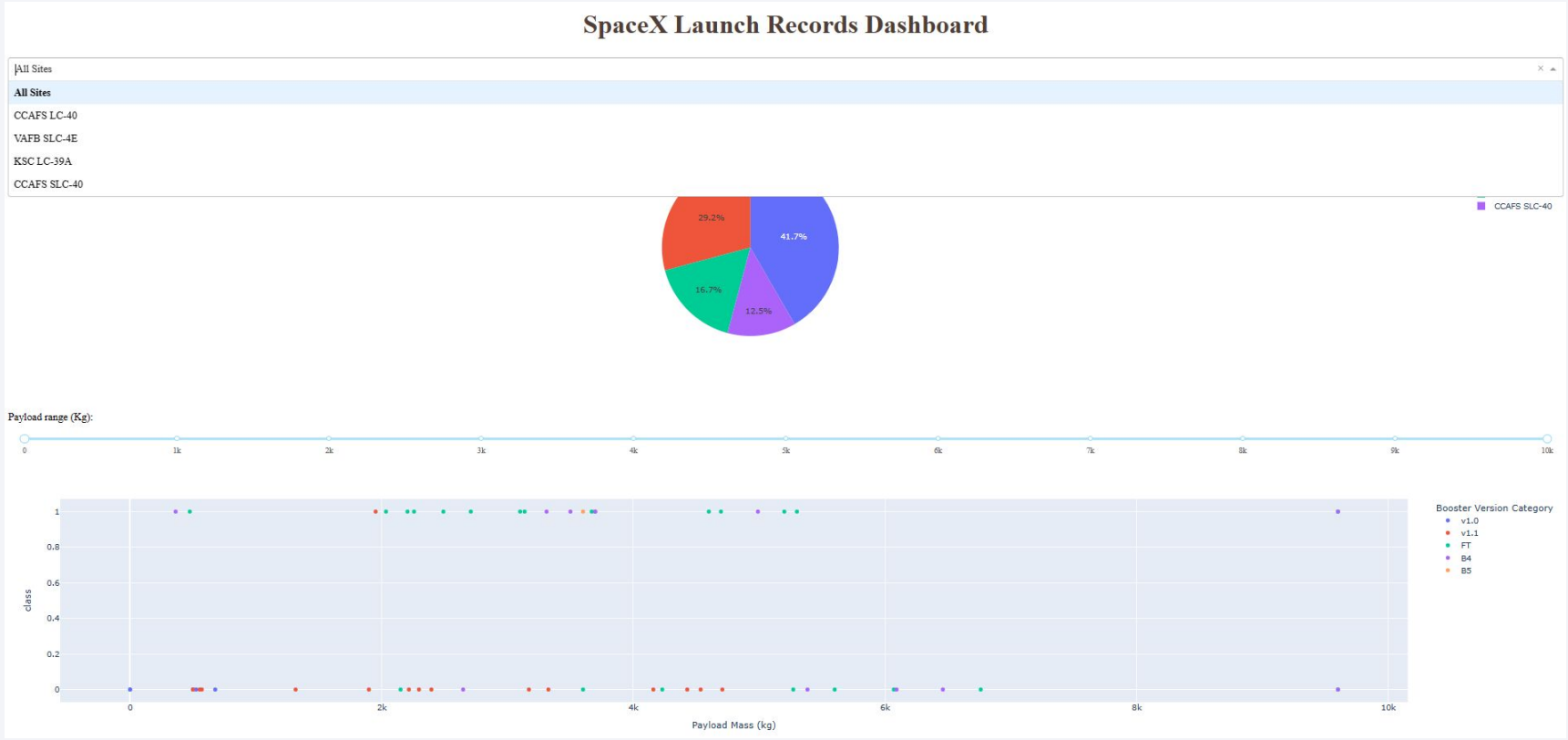
```
print(f'Logistic Regression Score:      {logreg_cv.score(X_test, Y_test)}')
print(f'Support Vector Machine Score:   {svm_cv.score(X_test, Y_test)}')
print(f'Decision Tree Classifier Score: {tree_cv.score(X_test, Y_test)}')
print(f'KNN Score:                      {knn_cv.score(X_test, Y_test)}')
```

✓ 0.0s

```
Logistic Regression Score:      0.8333333333333334
Support Vector Machine Score:   0.8333333333333334
Decision Tree Classifier Score: 0.8333333333333334
KNN Score:                      0.8333333333333334
```

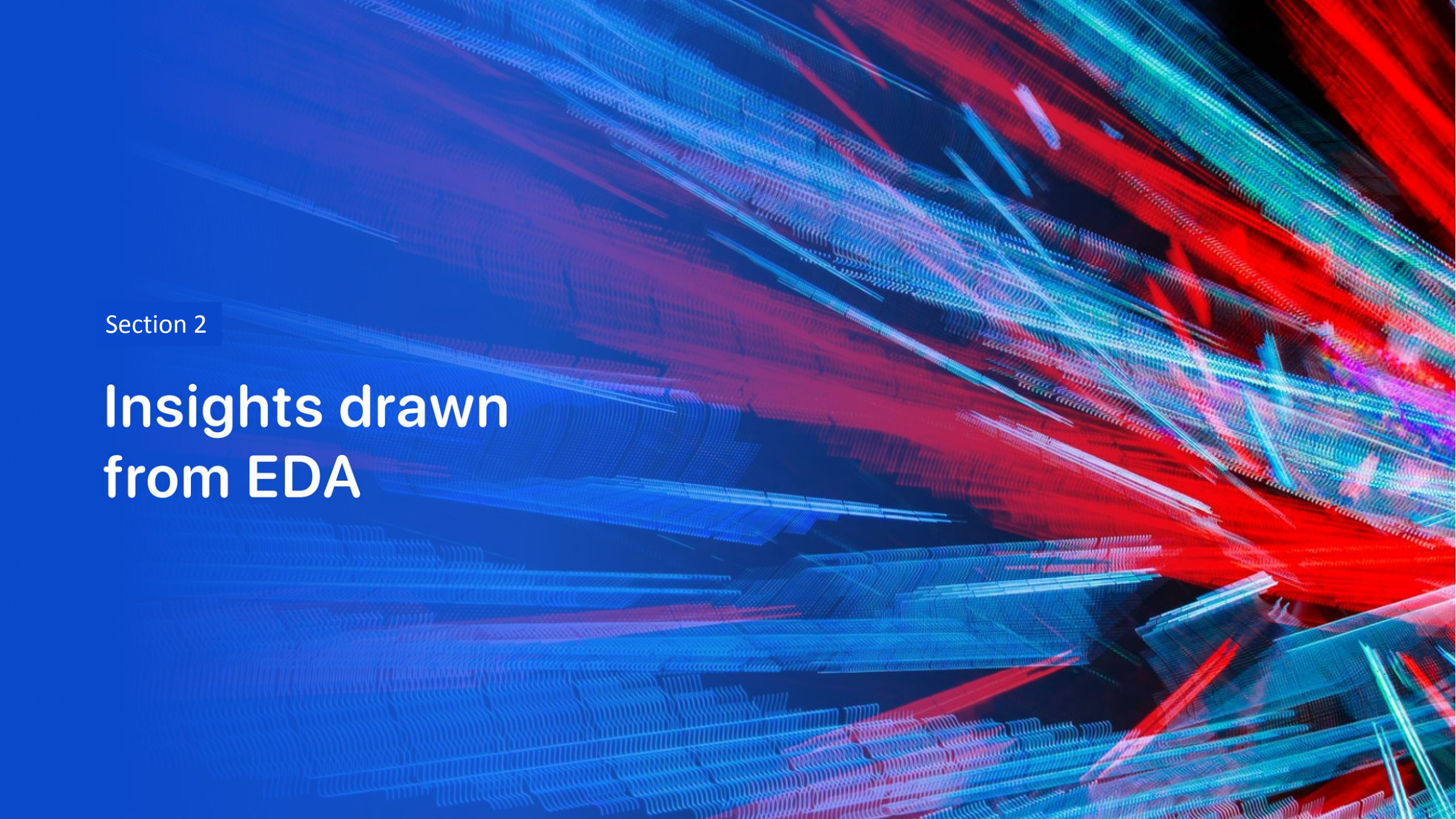
Results

- Interactive analytics demo in screenshots

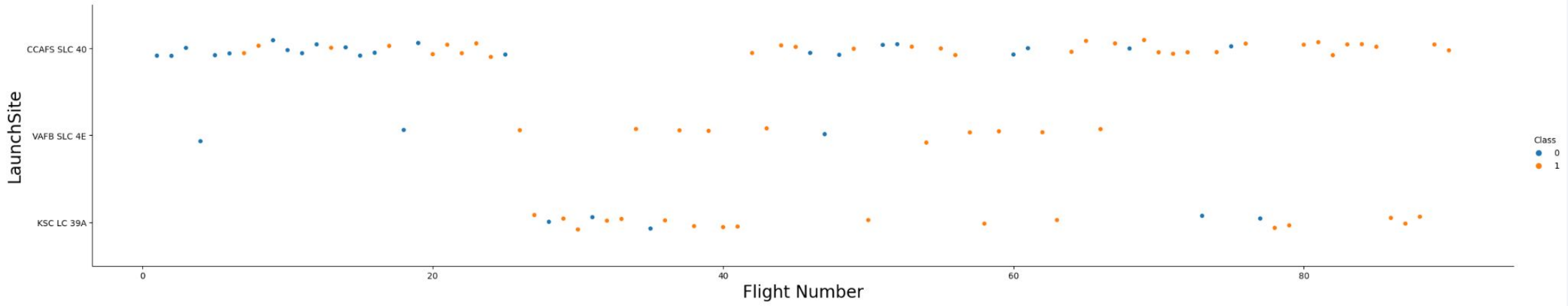


Section 2

Insights drawn from EDA

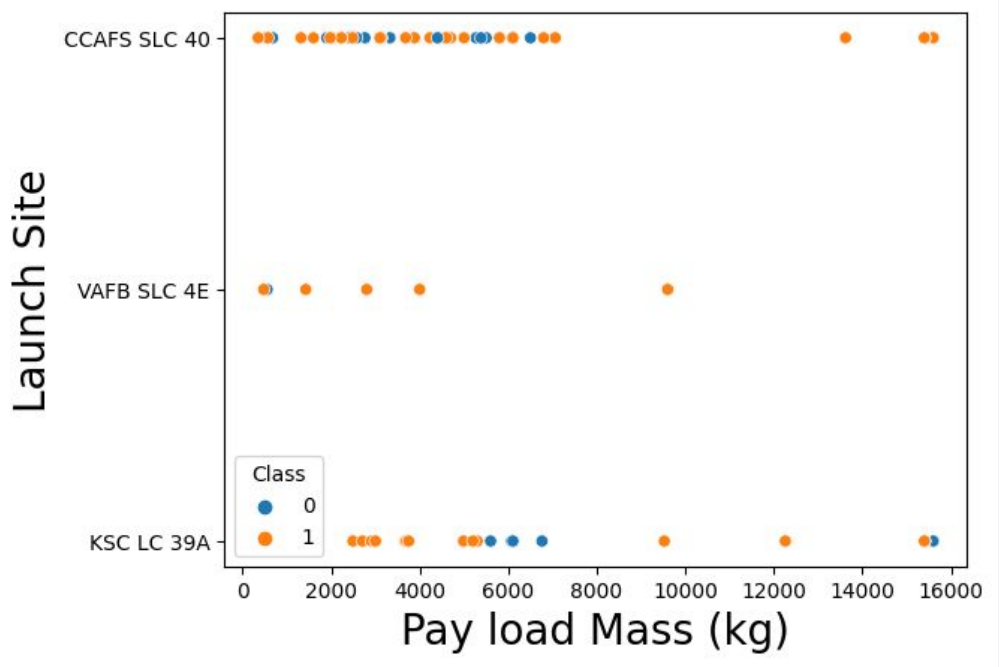


Flight Number vs. Launch Site



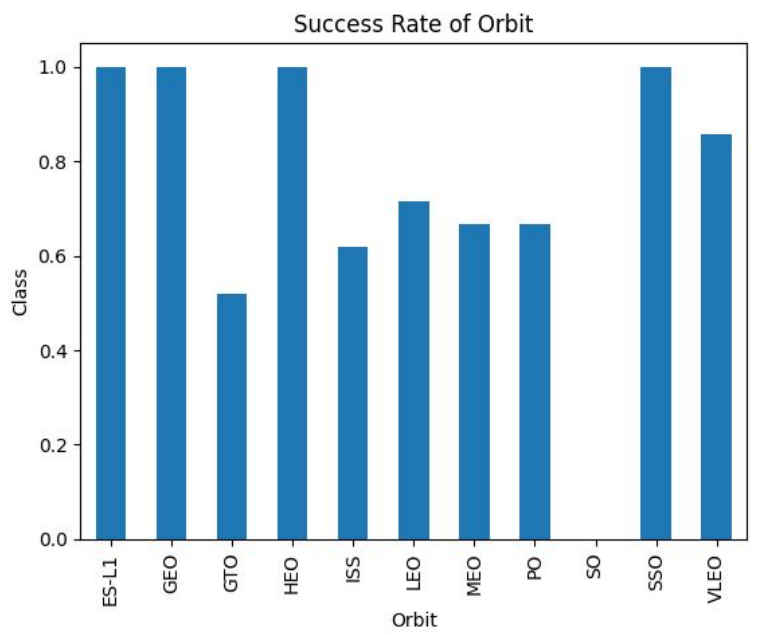
- Here we can see the launch site for each flight number and how successful was the outcome of each flight.

Payload vs. Launch Site



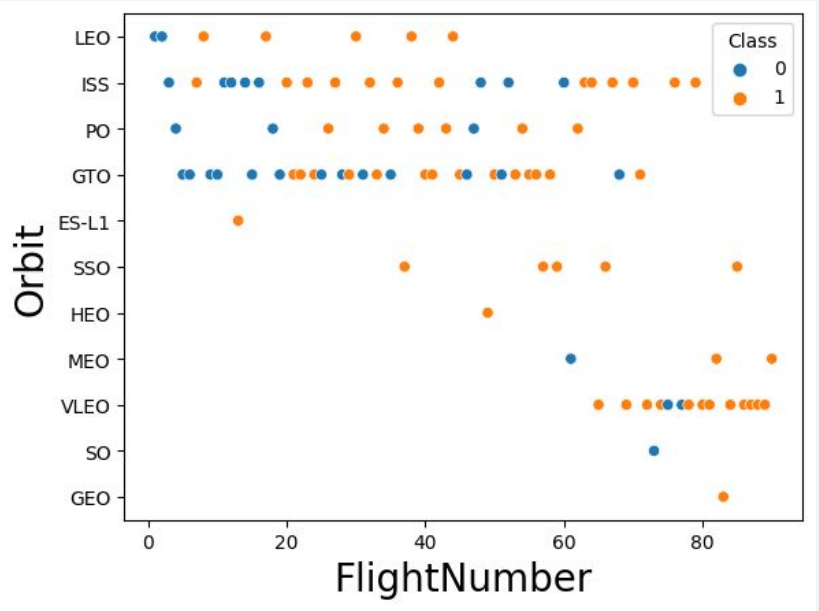
- In this plot we can see for KSC LC 39A that havy payload is more likely to lead to an unsuccessful landing.
- Missions starting from VAFB SLC 4E didn’t carry havy payload and were quite successful
- It is difficult to find a pattern for CCAFS SLC 40 because the Outcomes are mixed

Success Rate vs. Orbit Type



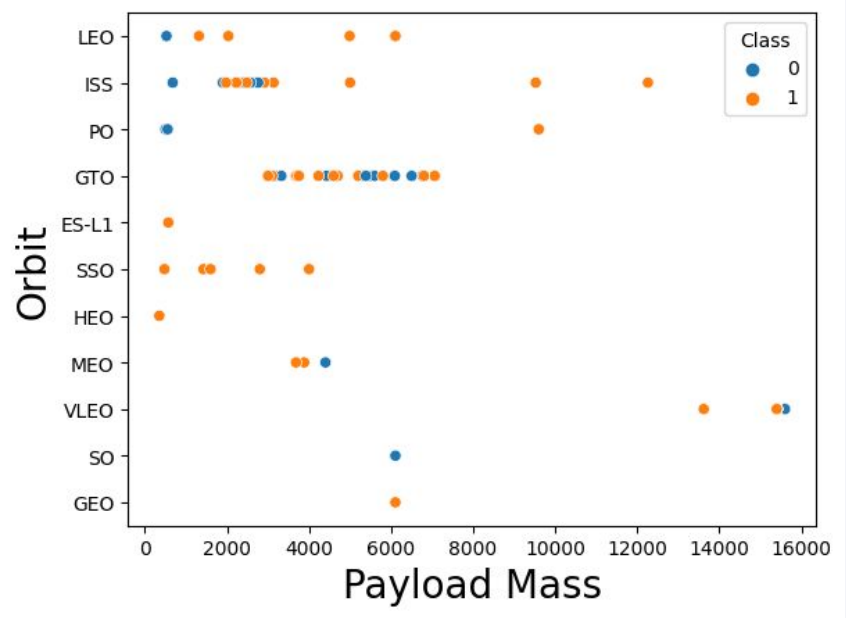
- In this plot we can see success rate of each targeted orbit. Flights to some orbits are more often successful then to others. This data can therefore be used in the dataset, because the outcome depends on orbit type.

Flight Number vs. Orbit Type



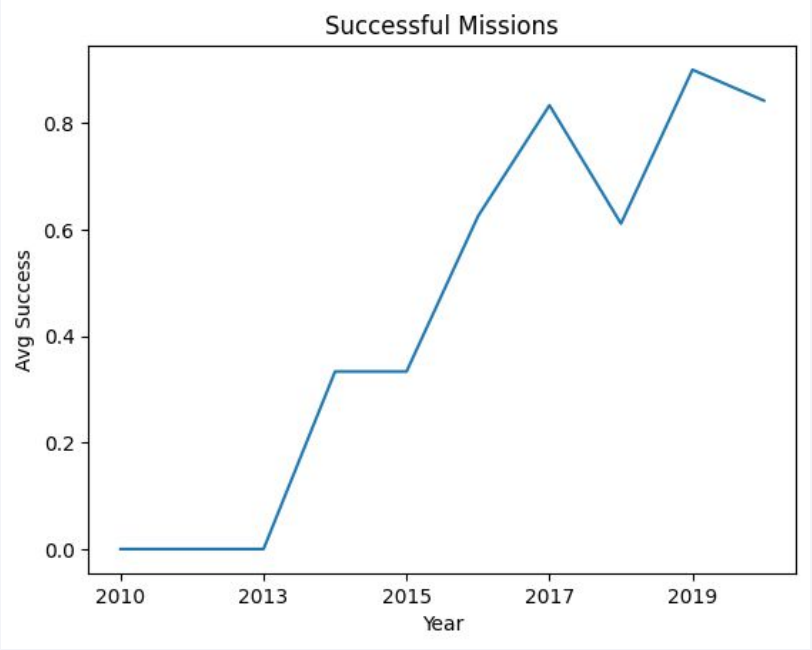
- Here we can see that number of successful landings increases with higher flight number

Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend



- In this chart we can see, that number of successful landings has increased over the years

All Launch Site Names

- The output of SQL query regarding unique Launch Site Names was:
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SCL-40

- The corresponding query code can be seen in the screenshot

```
%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'

```
%sql SELECT Launch_Site FROM SPACEXTABLE WHERE Launch_Site like 'CCA%' LIMIT 5;
```

[12]

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

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

- The result of the query can be seen in the screenshot above. The query returned 5 launch site names starting with CCA from spacex table

Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) as TOTAL_PAYLOAD FROM  (SELECT PAYLOAD_MASS__KG_ FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)');
```

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

TOTAL_PAYLOAD
45596

- The total payload carried by boosters from NASA is 45596 kg. This is the result of the SQL query shown in the screenshot above.

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) as AVG_PAYLOAD FROM (SELECT PAYLOAD_MASS__KG_ FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1');

 * sqlite:///my_data1.db
Done.

AVG_PAYLOAD
2928.4
```

- Average payload carried by booster F9 v1.1 is 2928.4 according to SQL query as shown in the screenshot above

First Successful Ground Landing Date

```
%sql SELECT MIN(Date) FROM (SELECT Date, Mission_Outcome FROM SPACEXTABLE WHERE Mission_Outcome = 'Success')
```

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

MIN(Date)
2010-04-06

- According to SQL query in the screenshot above the first successful landing was carried out on 2010-04-06

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
- The corresponding SQL query code can be found in the screenshot

```
%sql SELECT DISTINCT Booster_Version FROM (SELECT Booster_Version, PAYLOAD_MASS__KG_ FROM SPACEXTABLE WHERE Mission_Outcome = 'Success') \
WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

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

Booster_Version
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 FT B1032.2
F9 B4 B1040.2
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1046.3
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT Mission_Outcome, COUNT(Mission_Outcome) FROM SPACEXTABLE GROUP BY Mission_Outcome LIKE 'Success%'
```

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

Mission_Outcome	COUNT(Mission_Outcome)
Failure (in flight)	1
Success	100

- The query returned 100 successful missions and 1 failure.

Boosters Carried Maximum Payload

```
%sql SELECT DISTINCT Booster_Version FROM (SELECT Booster_Version, MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

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

Booster_Version
F9 B5 B1048.4

- The booster which has carried the maximum payload mass is F9 B5 B1048.4 according to query in the screenshot above

2015 Launch Records

```
%sql SELECT substr(Date, 6,2) as MONTH, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE Landing_Outcome = 'Failure (drone ship)' AND substr(Date,0,5)='2015';
```

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

MONTH	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- According to outcome of the query in the screenshot above there were two landing failures in 2015 which occurred in october and april.

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

```
%sql SELECT * FROM \
(SELECT Landing_Outcome, COUNT(Landing_Outcome) as total_outcome FROM SPACEXTABLE \
 WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome) ORDER BY total_outcome DESC;
```

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

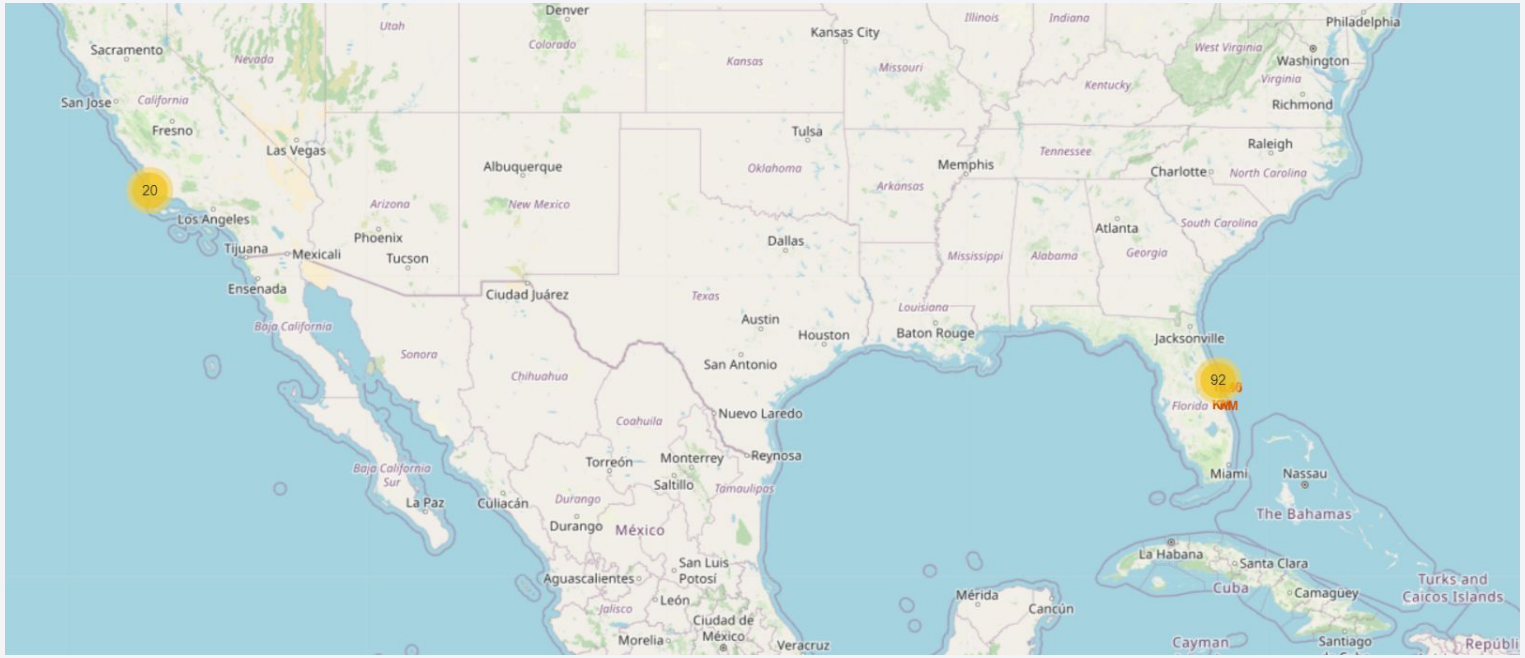
Landing_Outcome	total_outcome
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Success (ground pad)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	1
Precluded (drone ship)	1

- Screenshot above shows the rank of landing outcomes between 2010-06-04 and 2017-03-20 in descending order


Section 3
Launch Sites
Proximities Analysis

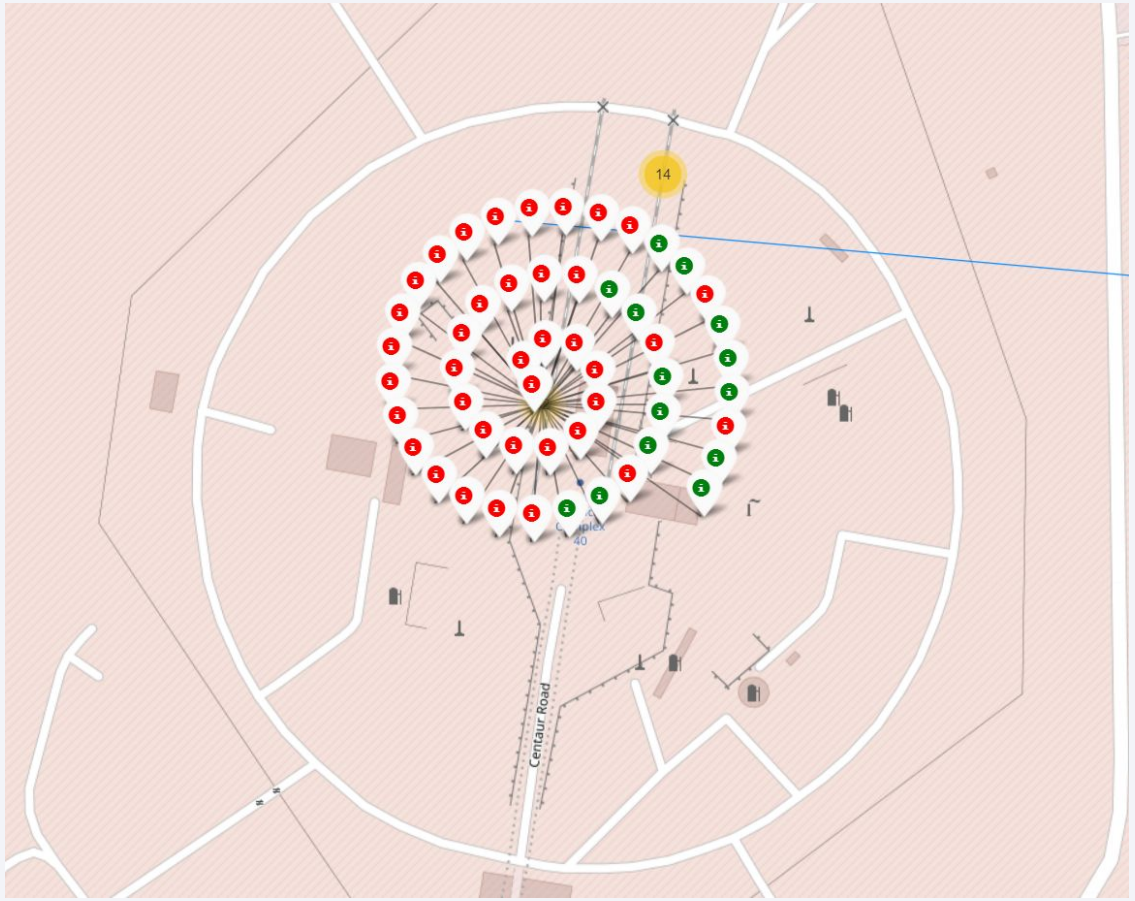
Launch Sites on global Map

- The map on the screenshot shows two sites. We can see that 20 launches were carried out near Los Angeles and 92 from Cape Canaveral



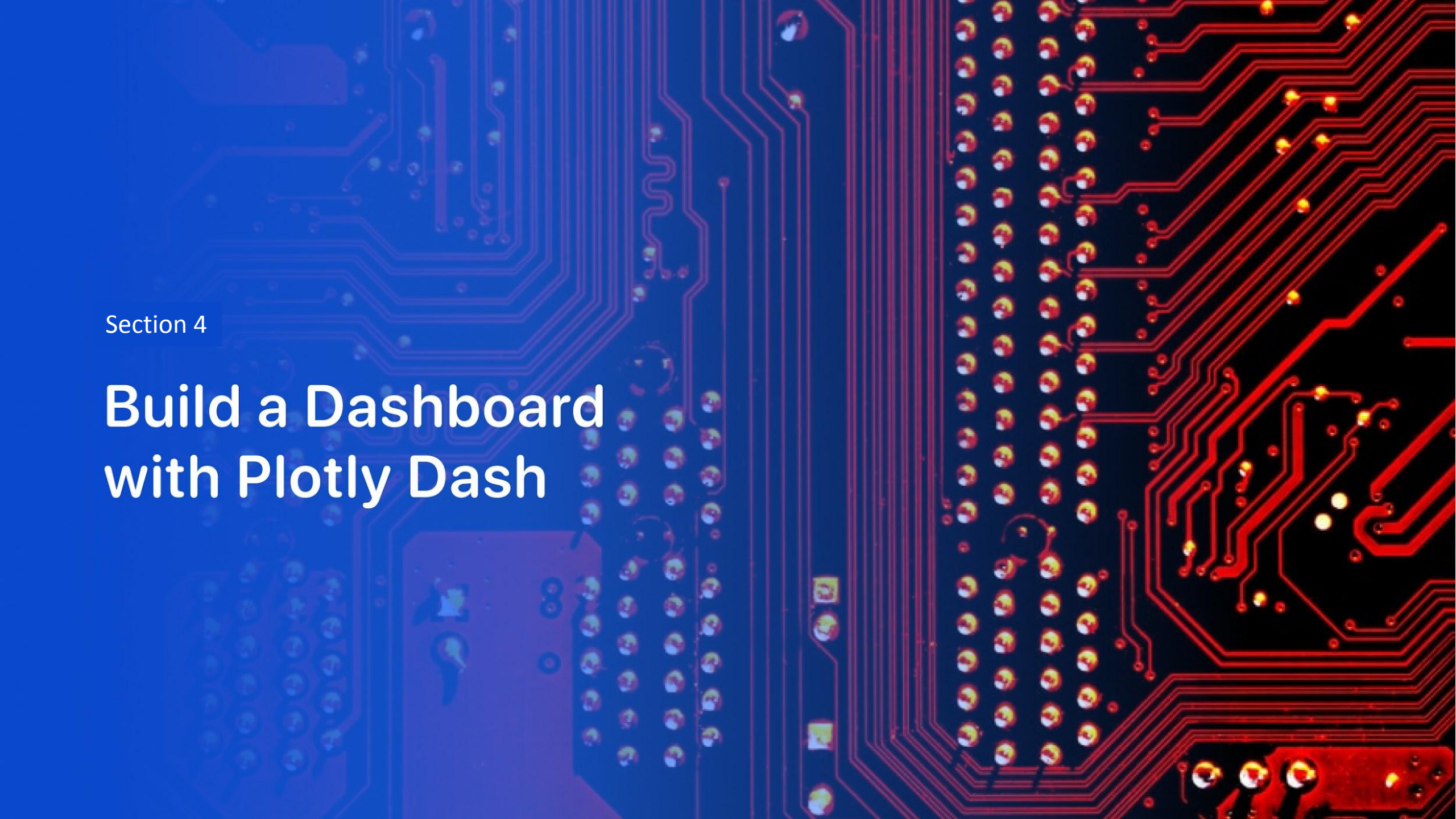
Launch Outcome

- Launch outcomes are labeled by makers. The color indicates success or failure.
- Markers lying next to each other are grouped into clusters

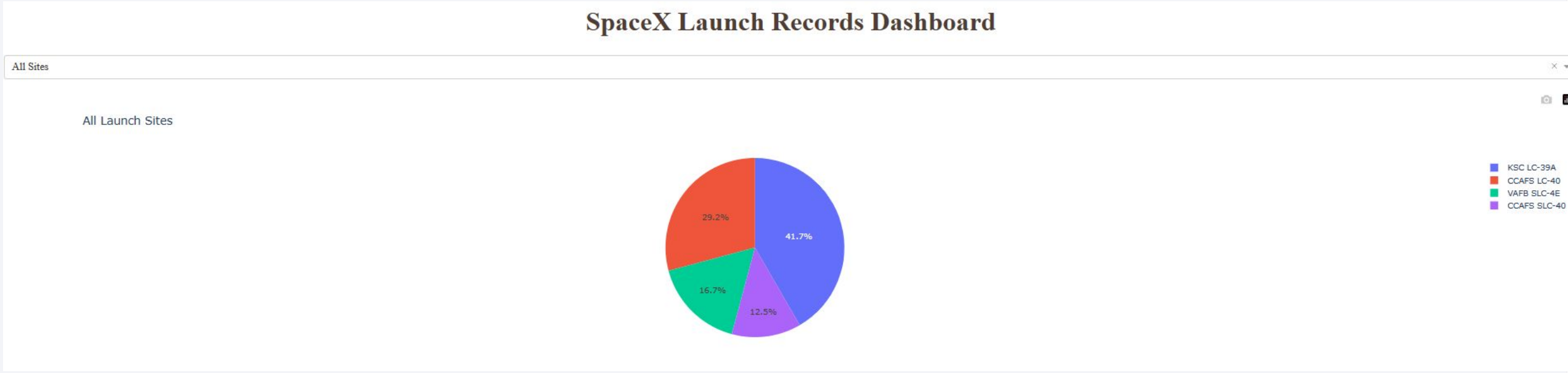


Distance to Proximites

- Explain the important elements and findings on the screenshot
- On the screenshot we can observe the distance of 16.40 km from launch site to the nearest city
- This is an example, how distances to proximities are calculated and depicted on a map plot.

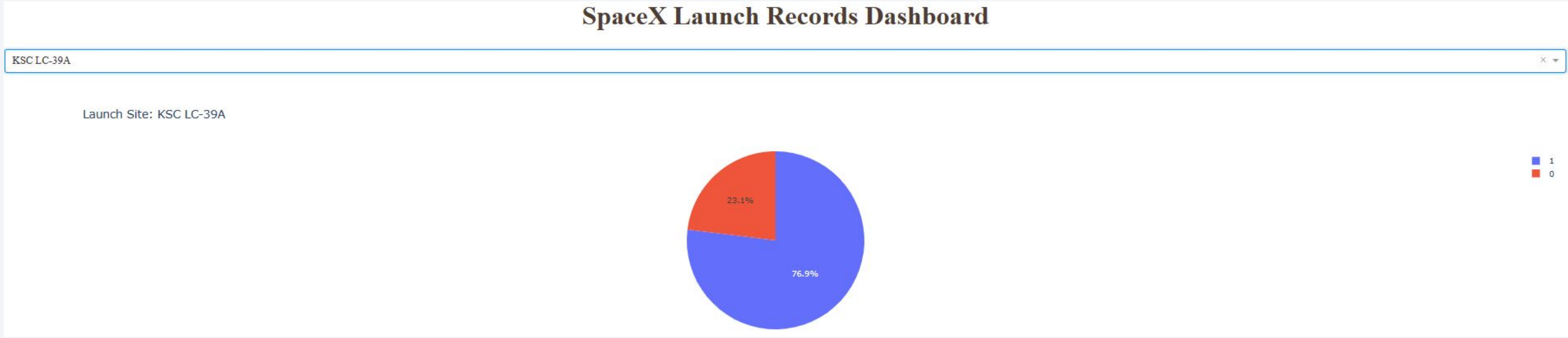

Section 4
Build a Dashboard with Plotly Dash

Pie Chart of All Launch Sites



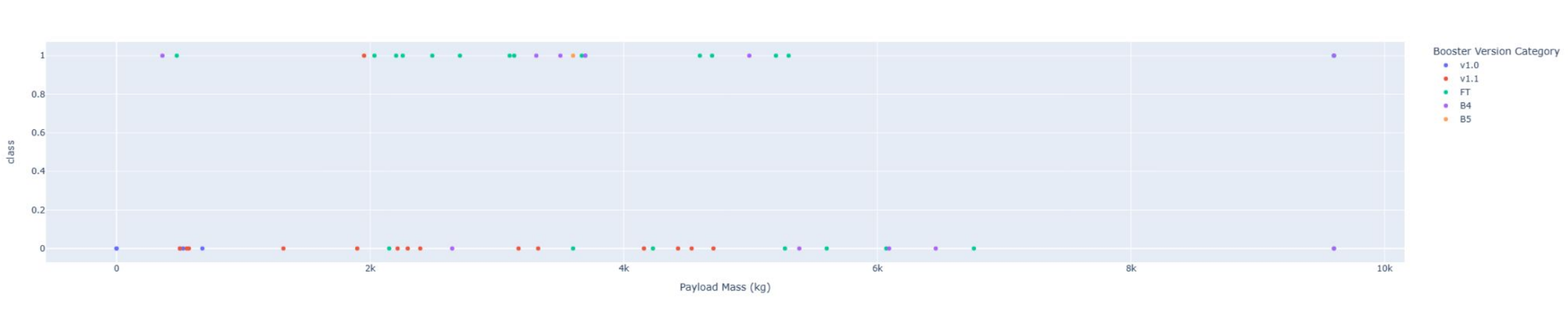
- In this screenshot a share of successful landings can be seen for each launch site

Highest launch success ratio



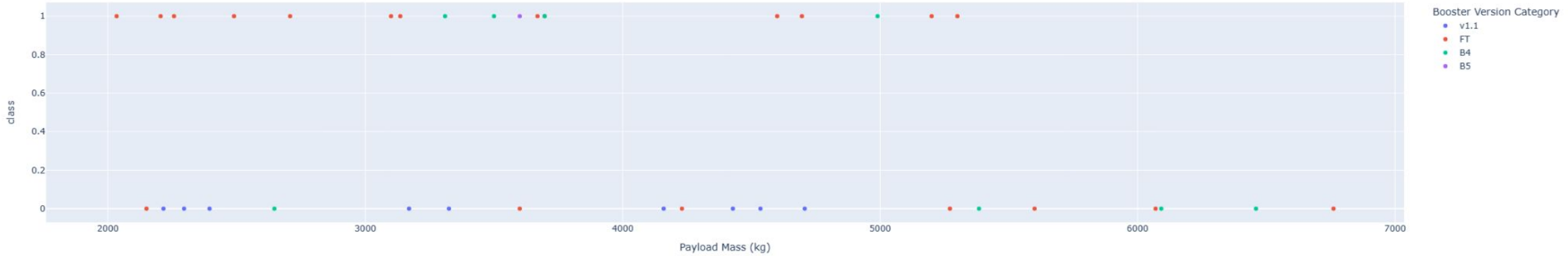
- From this screenshot can be taken, that missions starting from KSC LC-39A site have the highest launch to success ratio regarding landing performance

Payload vs. Launch Outcome



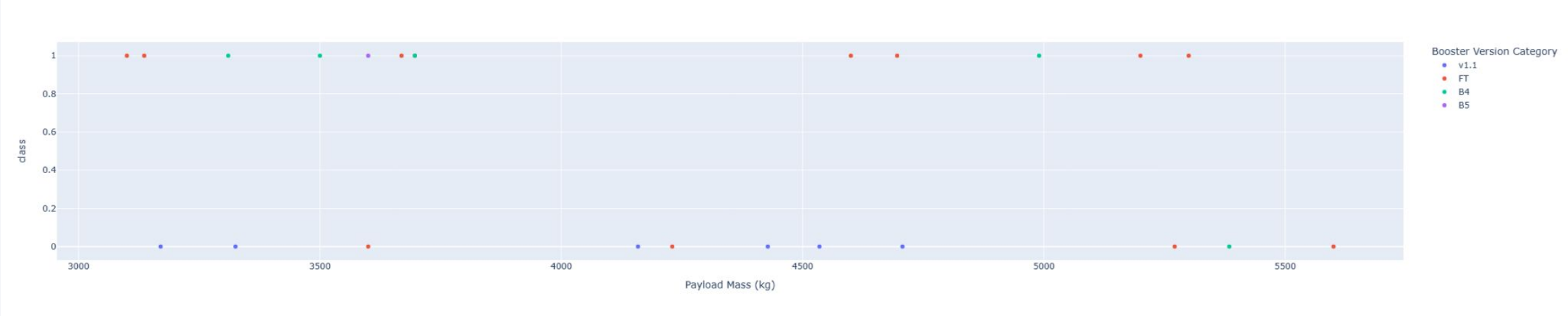
- Payload vs. Class plot for all launch sites. Payload range 0 to 10000 kg

Payload vs. Launch Outcome



- Payload vs. Class plot for all launch sites. Payload range 2000 to 8000 kg

Payload vs. Launch Outcome



- Payload vs. Class plot for all launch sites. Payload range 3000 to 6000 kg

Payload vs. Launch Outcome

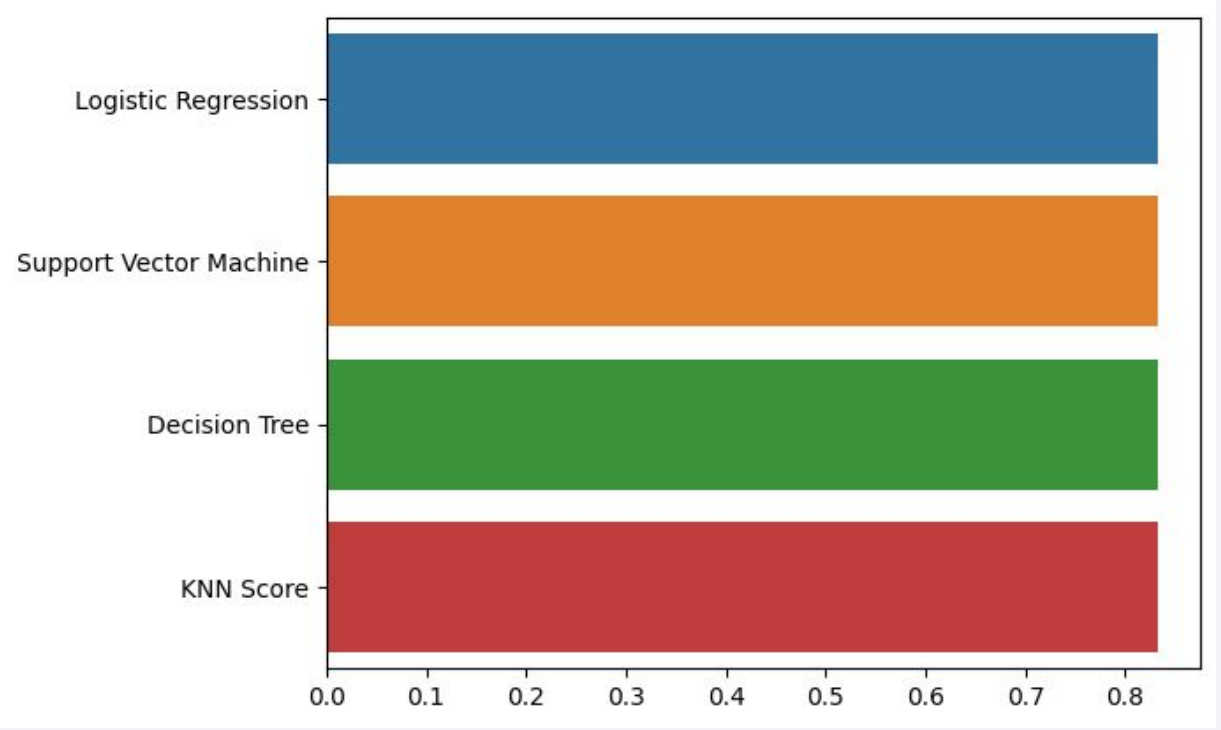
- Especially screenshots with wide range between 2000 and 8000 kg and 0 to 10000 kg demonstrate that the number of failed landings increase with higher payloads.

Section 5

Predictive Analysis (Classification)

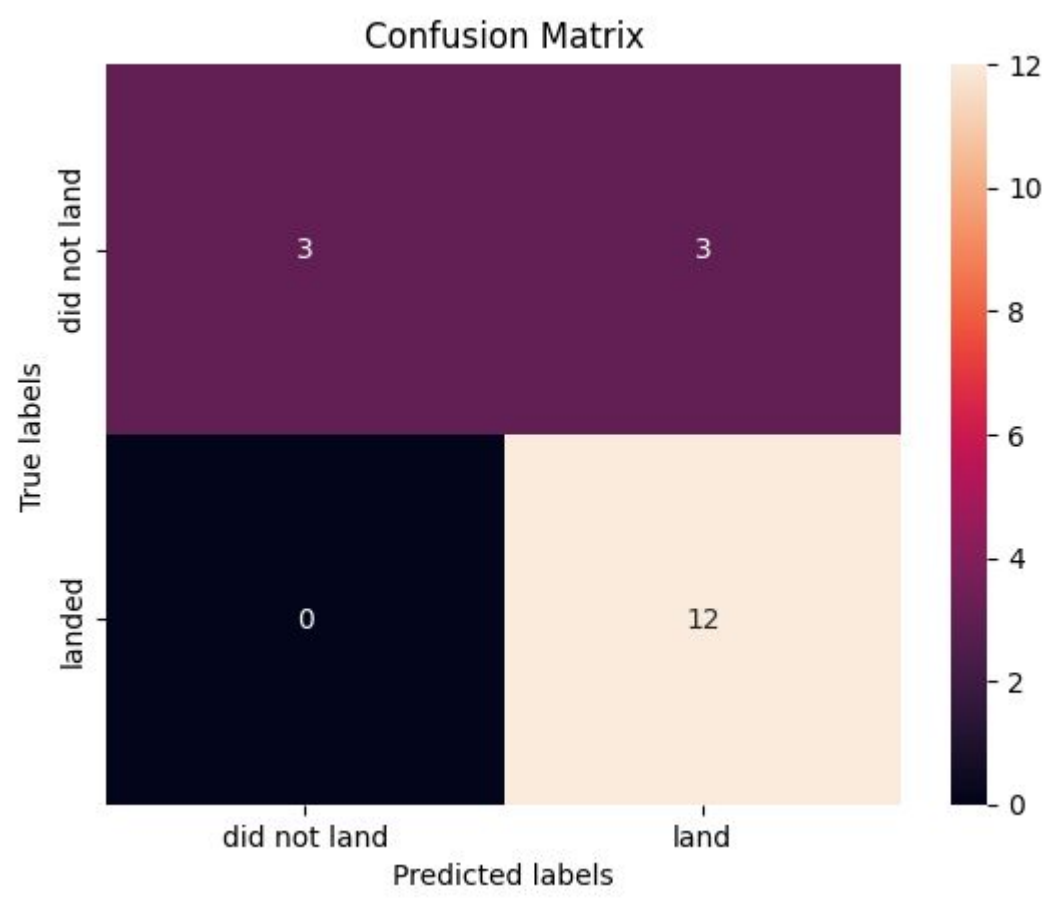
Classification Accuracy

- All 4 models seem to show same performance



Confusion Matrix

- As presented on the previous slide, all 4 models have performed with equal result of 83.33%.
- This confusion matrix shows that the main problem of all 4 models is prediction of false positives.



Conclusions

- Data was gathered from from SpaceX official source and from Wikipedia via web scraping.
- Analysis of this data revealed some relationships between launch circumstances and landing success.
- These discovered relationships were used to create a dataset to predict an outcome for future flights.
- Machine learning models were successfully trained on this data and achieved 83.33 % accuracy predicting successful landing.
- The problem is that the models tend to false positive predictions

Appendix

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude	Class
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.561857	0
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.561857	0
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.561857	0
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610829	34.632093	0
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28.561857	0

- Structure of the data set

Thank you!