IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Gianluca Giulini>
<13/02/2025>

Outline

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

Executive Summary

- The methodologies adopted in this project are data collection (from public SpaceX API and by scrapping SpaceX Wikipedia page), data wrangling, exploratory data analysis (EDA) using visualization and SQL, interactive data analytics using Folium and Plotly Dash, and predicative analysis (with machine learning algorithms) using classification models.

- The results include data analysis results, data visualization results, and predicative analysis results.

Introduction

- Project background and context
  - The commercial space age is here and so many companies are working to make space travel affordable for everyone. Likely the most successful is SpaceX that advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars whereas other providers cost more than 165 million dollars. Savings are mostly due to the fact that the first stage will be re-used after the launch Therefore, the goal of this project is to predict if the Falcon 9 first stage will land successfully, which can eventually determine the cost of a launch. Such information will be also useful for any company wishing to bid against SpaceX for a rocket launch investment.
- Problems to be answered
  - What attributes and training labels need to be used for the predicative model.
  - The effect of each feature on the outcome of the launch.

Section 1

Methodology

Methodology

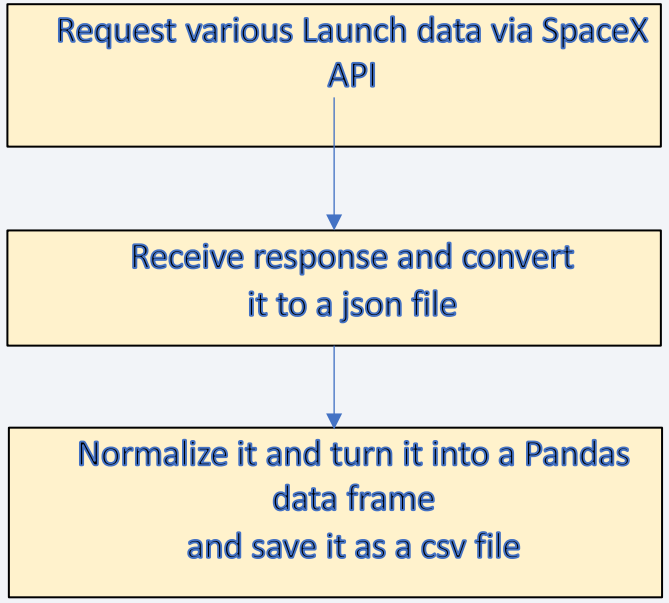
Executive Summary

- Data collection methodology:
  - Make a request to the SpaceX API
  - Web scraping from the Wikipedia page (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
- Perform data wrangling
  Convert landing outcomes to a binary format (like turning "successful" and "failed" into 1s and 0s) and create labels for outcomes
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  Initialize a set of classification models. Then, utilize Grid Search to determine the optimal hyperparameters for each model. Finally, compare the training and test accuracies of these models to identify the one with the highest performance.

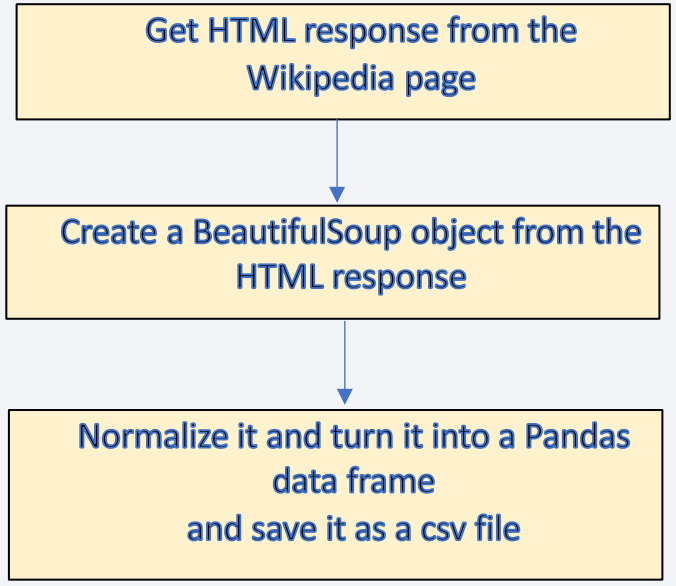
Data Collection

- Data collection process
  - For SpaceX API, we make a request to the SpaceX API and then go on to clean the requested data.
  - For web scraping, we extract a Falcon 9 launch records HTML table from Wikipedia and then parse the table and convert it into a Pandas data frame

SpaceX API

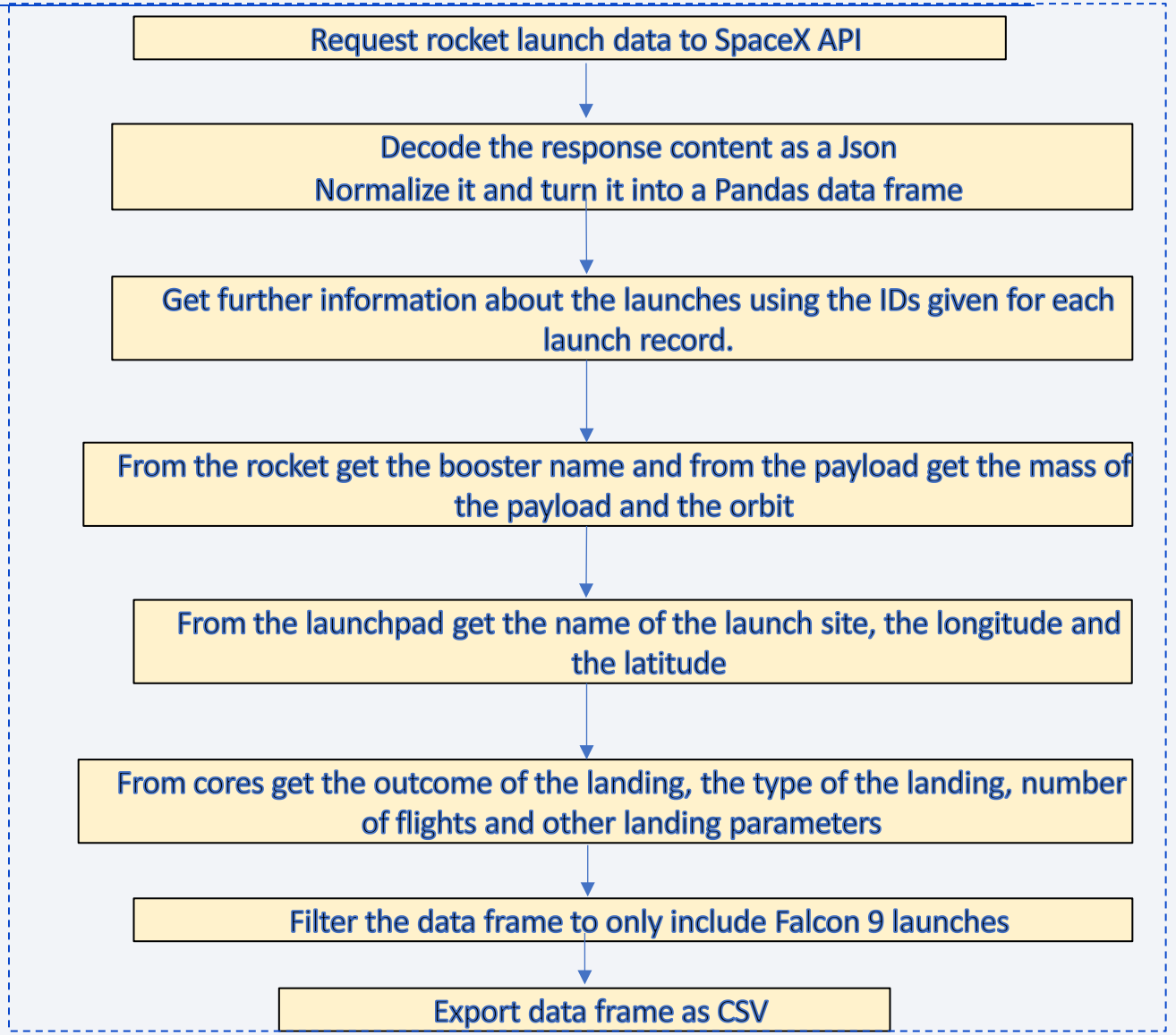


Web scraping



Data Collection – SpaceX API

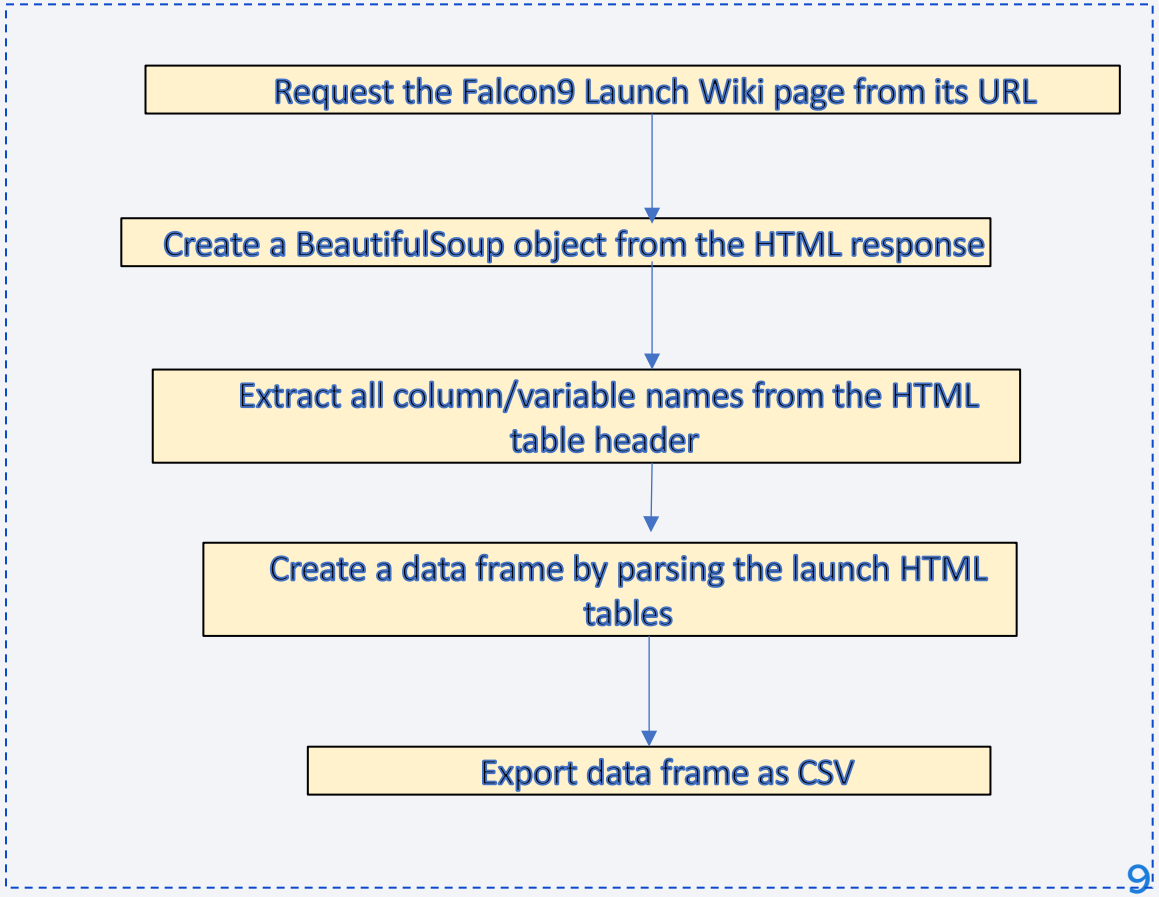
- Link to the SpaceX API calls on my notebook
- https://github.com/bejaflor66/applied-data-science-capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

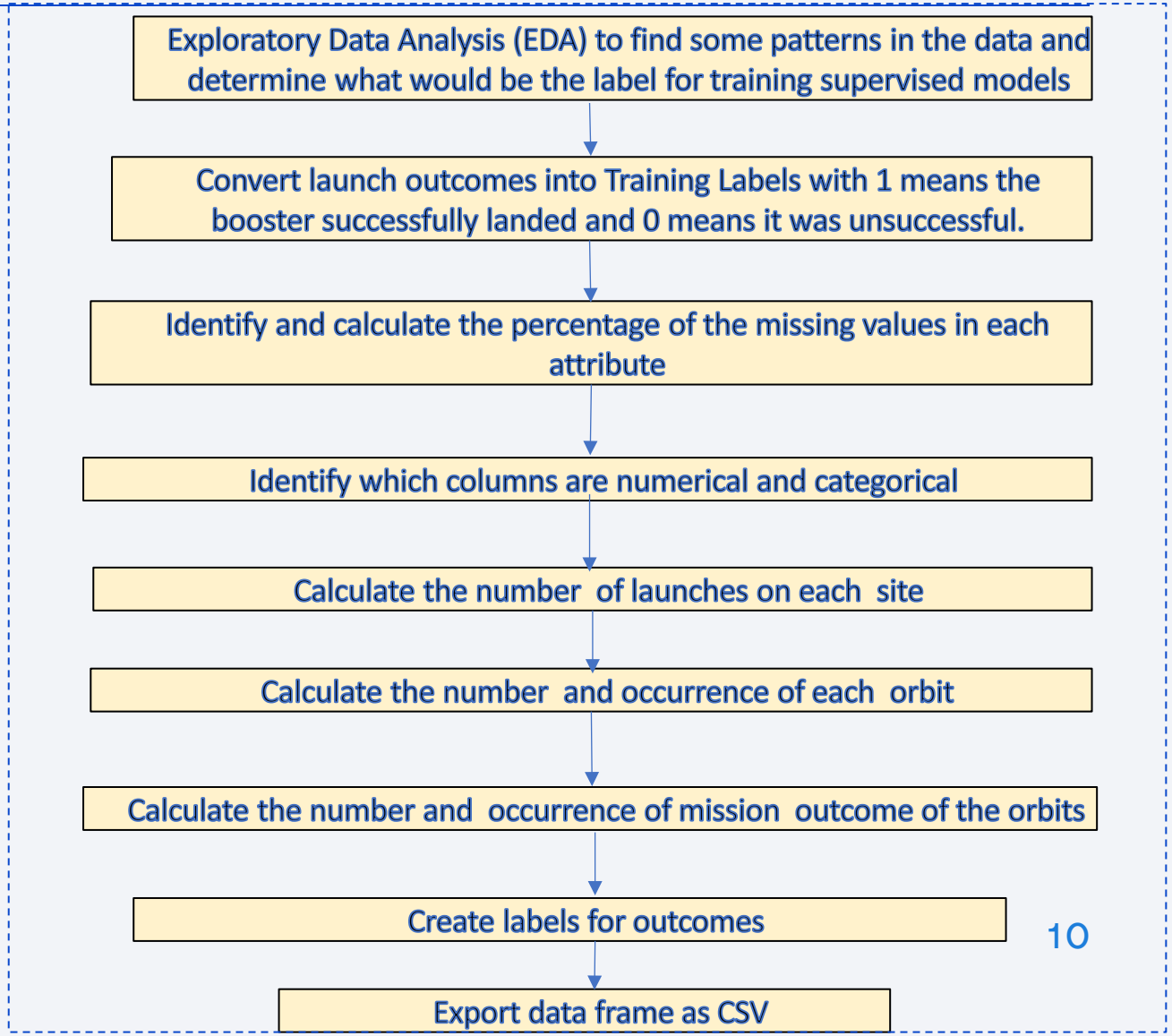
- Link to the Web scraping notebook

- https://github.com/bejaflor66/applied-data-science-capstone/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Link to the Data Wrangling notebook
- https://github.com/bejaflor66/applied-data-science-capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- Scatter plots to check how much the success of a landing is related to the following:
  - Flight Number vs. Payload Mass
  - Flight Number vs. Launch Site
  - Flight Number vs. Orbit Type
  - Pay Load Mass vs. Orbit
- Bar chart:
  - Success Rate of Each Orbit Type
- Line chart:
  - Year vs. Success Rate

Reference to the EDA with data visualization notebook:
https://github.com/bejaflor66/applied-data-science-capstone/blob/main/edadataviz.ipynb

EDA with SQL

- List of SQL query tasks performed for EDA
  - **Display the names of the unique launch sites in the space mission**
  - **Display 5 records where launch sites begin with the string '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 succesful landing outcome in ground pad was acheived.**
  - **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 will 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.**

Reference to the EDA with SQL notebook: https://github.com/bejaflor66/applied-data-science-capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- United States Folium Map was created to show interactively the Launch sites
- Circle markers were added to indicate the different launch sites. The color of the markers represent the launch outcomes, with green=success and red=failure
- Lines were added to represent the distance between launch sites and its proximities.

- Reference to the Interactive Map notebook: https://github.com/bejaflor66/applied-data-science-capstone/blob/main/lab_jupyter_launch_site_location.ipynb

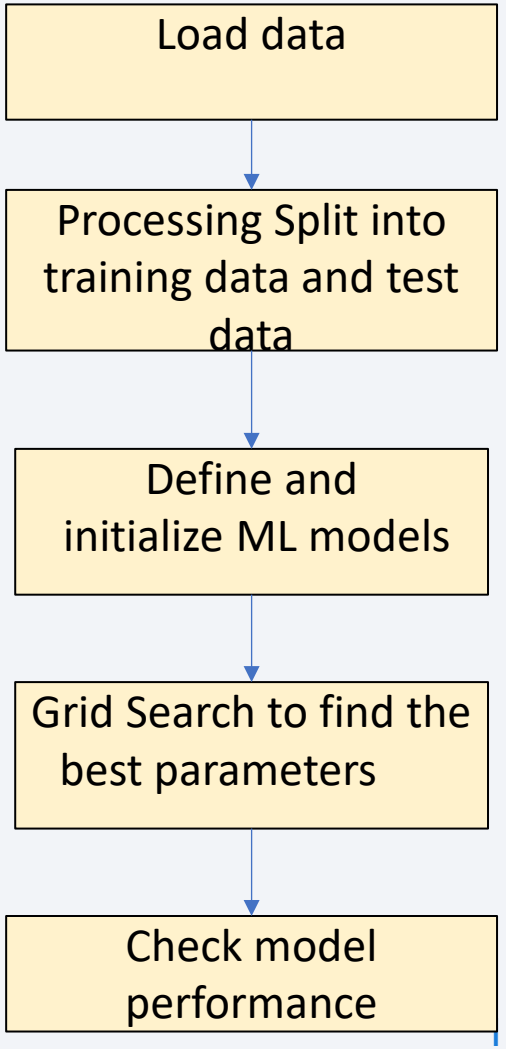
Build a Dashboard with Plotly Dash

- A Dashboard was developed to show:
  - A dropdown section which lists “All Sites” and individual launch sites to modify the results of the Pie chart
  - A Pie chart showing the total successful launches of all sites or a specific launch site
  - xA range slider which allows to select different Payload Mass ranges to filter the range of payload mass for the scatter plots.
  - Scatter plots showing the relationship between Outcome and Payload Mass for different Booster Versions.

Reference to the Dashboard Python project: https://github.com/bejaflor66/applied-data-science-capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

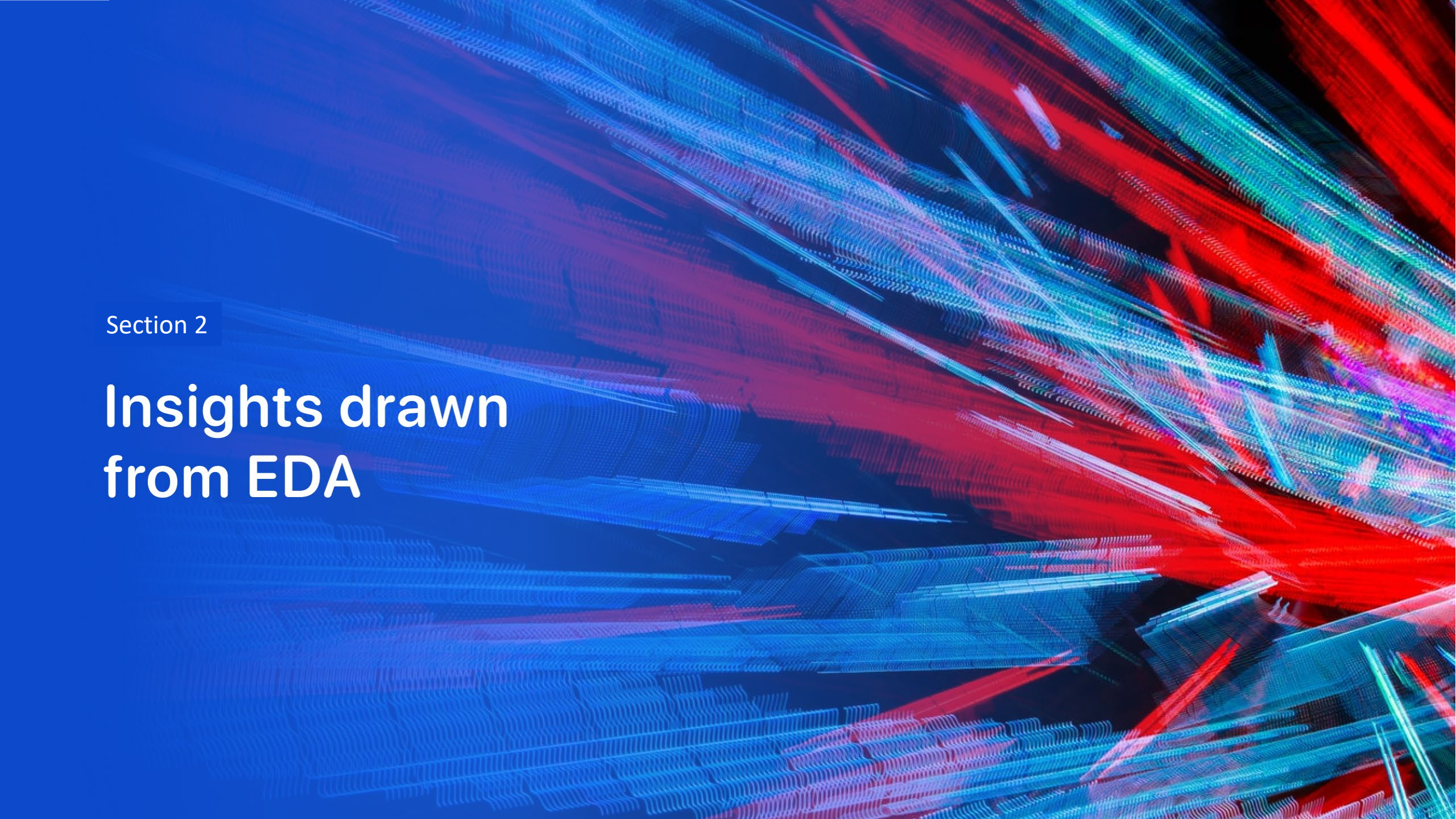
- Model Build-up
  - Load the dataset into a pandas dataframe
  - Data standardization and normalization
  - Split the data into training and testing data
  - Create a Logistic Regression, SVM, Decision Tree and KNN objects to initialize the ML model
  - Fit the objects to find the best parameters through GridSearch
- Model Evaluation
  - Check the accuracy on test data accuracy scores
  - See the confusion matrix
- Find the best performing classification model
  - Choose the model with the highest accuracy scores



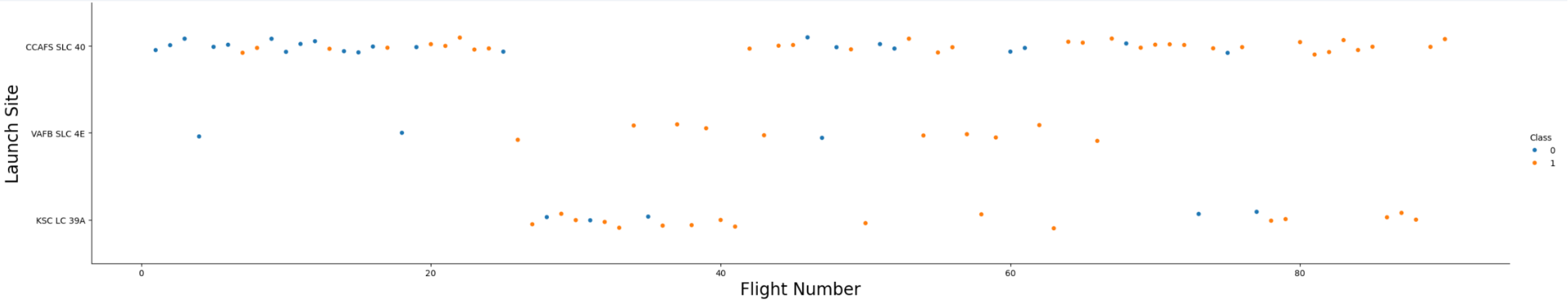
- Reference to the predictive analysis notebook:
- https://github.com/bejaflor66/applied-data-science-capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

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


Section 2
Insights drawn from EDA

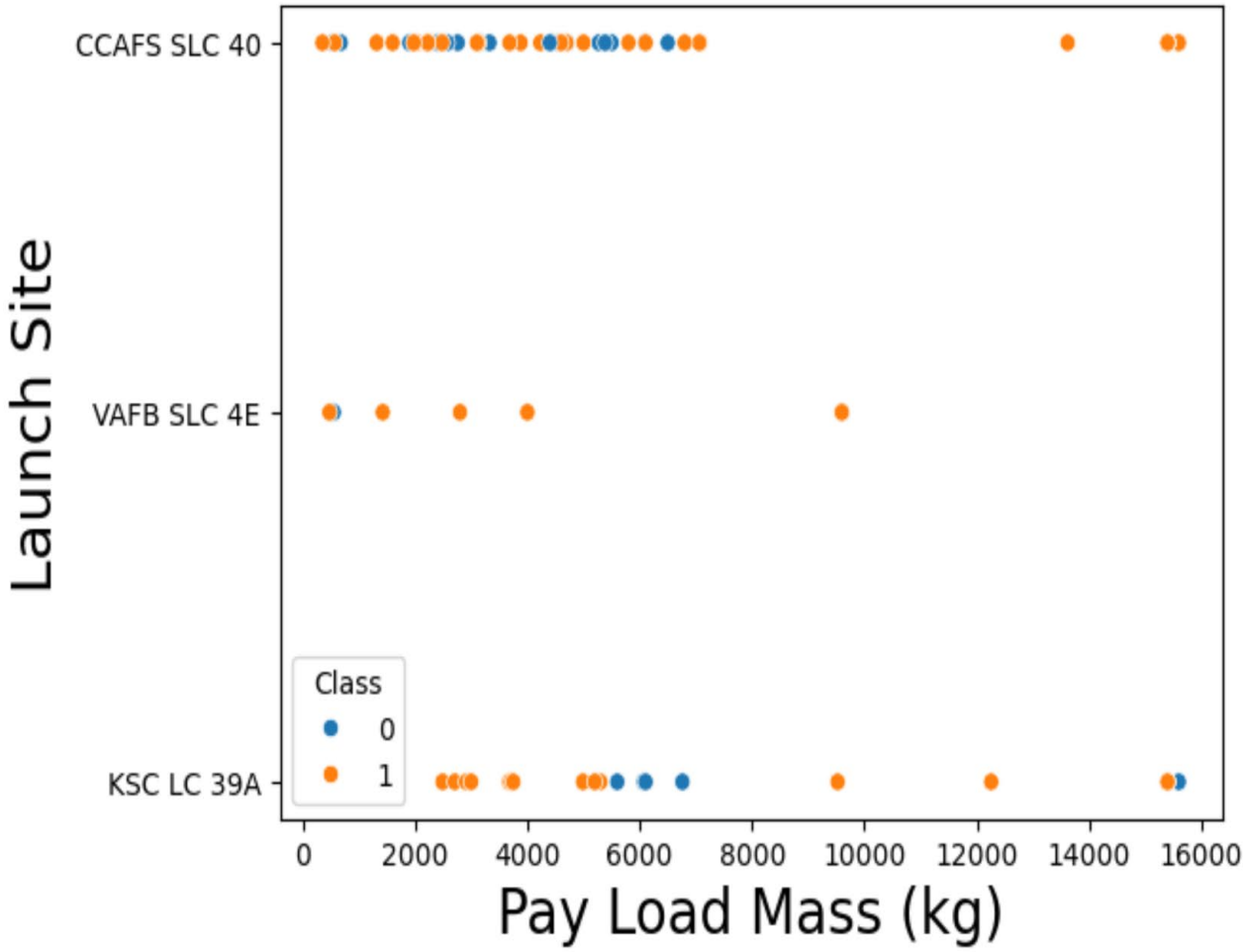
Flight Number vs. Launch Site




```
sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)
```

- As shown in the chart, as the flight number increases, the launch outcome is likely to be successful.

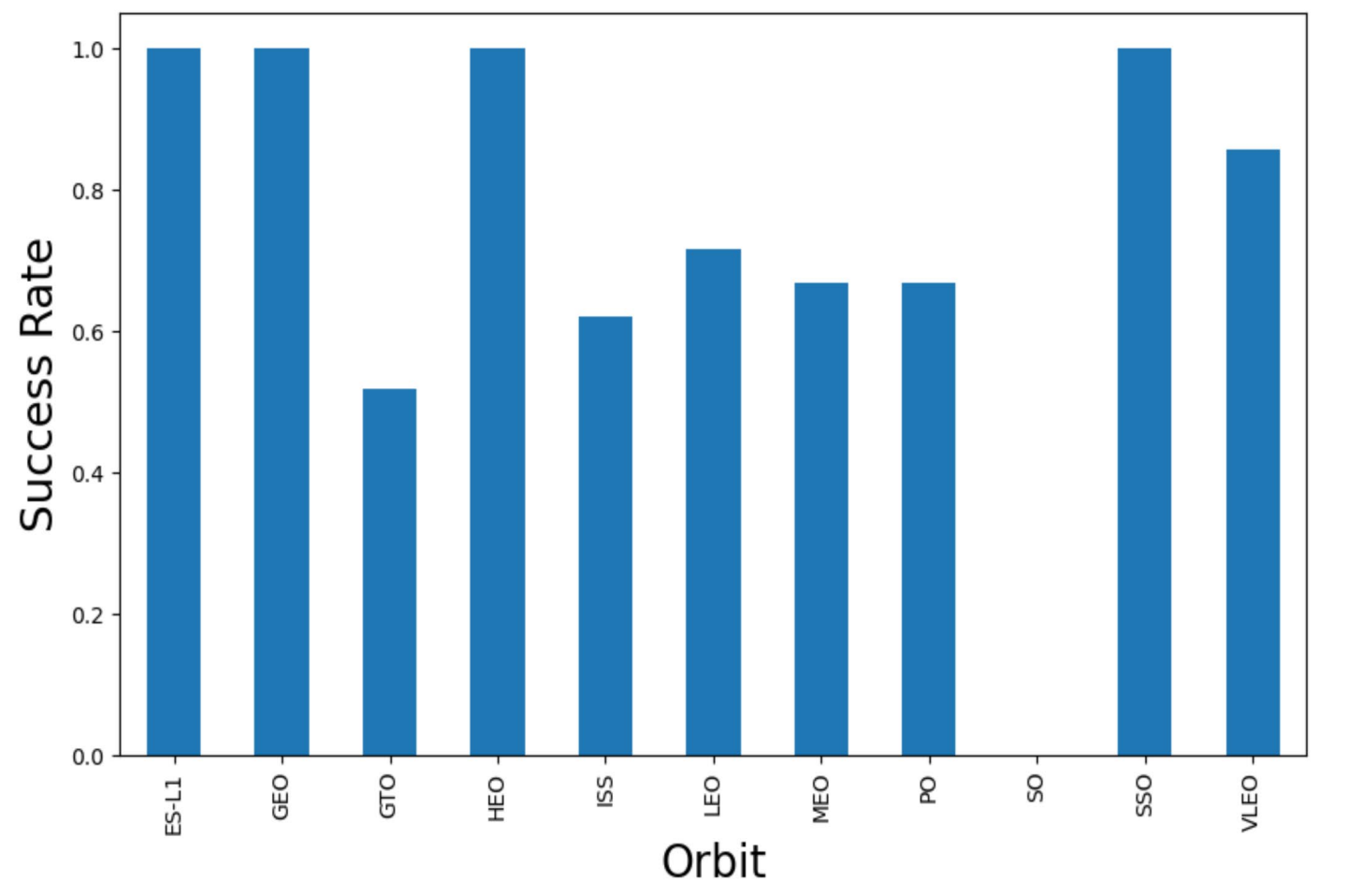
Payload vs. Launch Site




```
sns.scatterplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df)
```

- For CCAFS SLC-40, for small and medium payload mass the outcome does not seem to depend on it, but it is successful with big load mass;
- For VAFB SLC-4E, as payload mass increases, the launch outcome tends to be successful;
- For KSC LC-39A, with payload mass between 6000 and 8000 kg, the launch outcome failed.

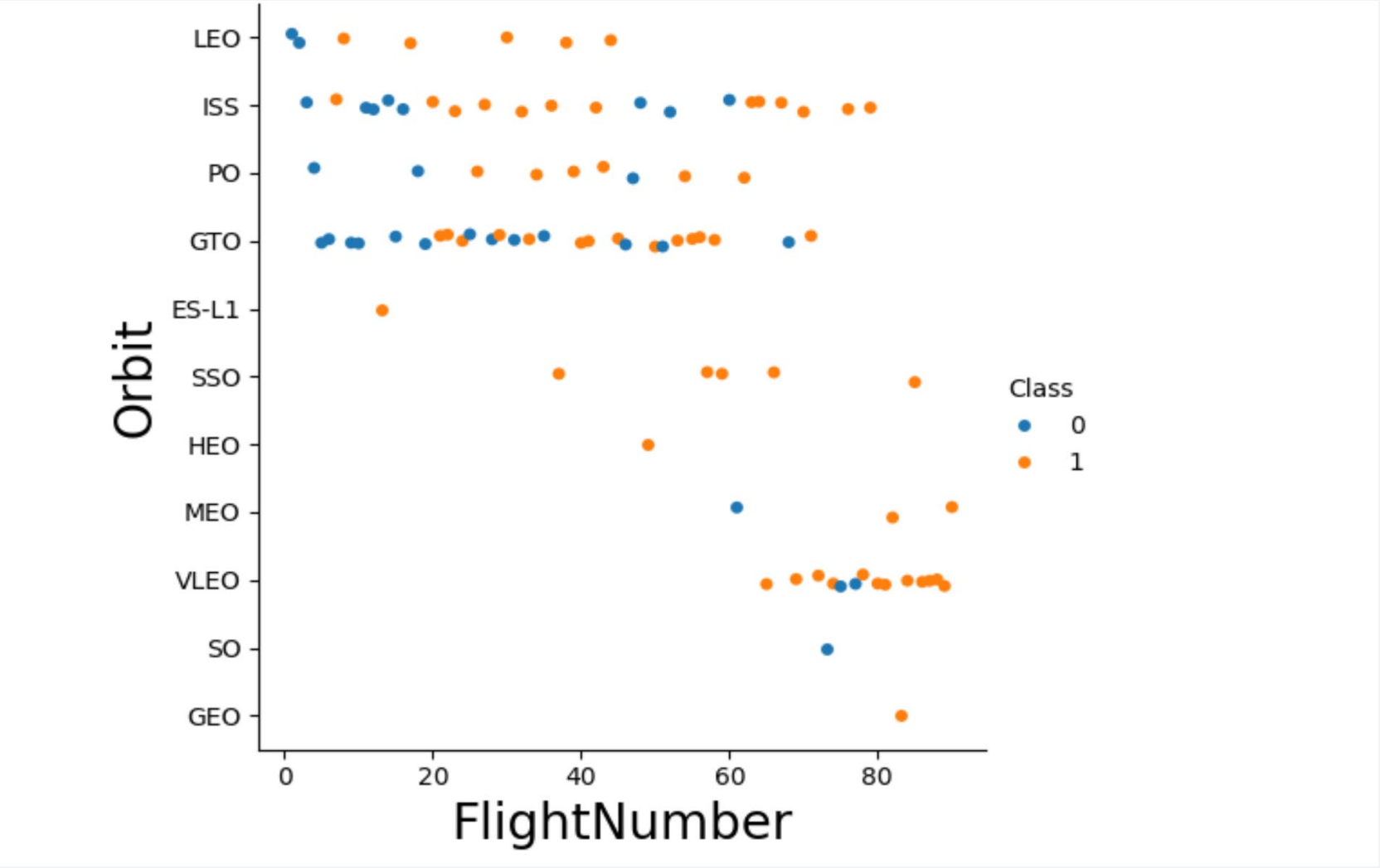
Success Rate vs. Orbit Type



- ES-L1, GEO, HEO and SSO have the highest success rate. Vleo has 85% success rate while the others have between 45% and 70 % success rate. SO has no successful launch based on data available


```
df.groupby("Orbit")['Class'].mean().plot(kind='bar',figsize=(10, 6))
```

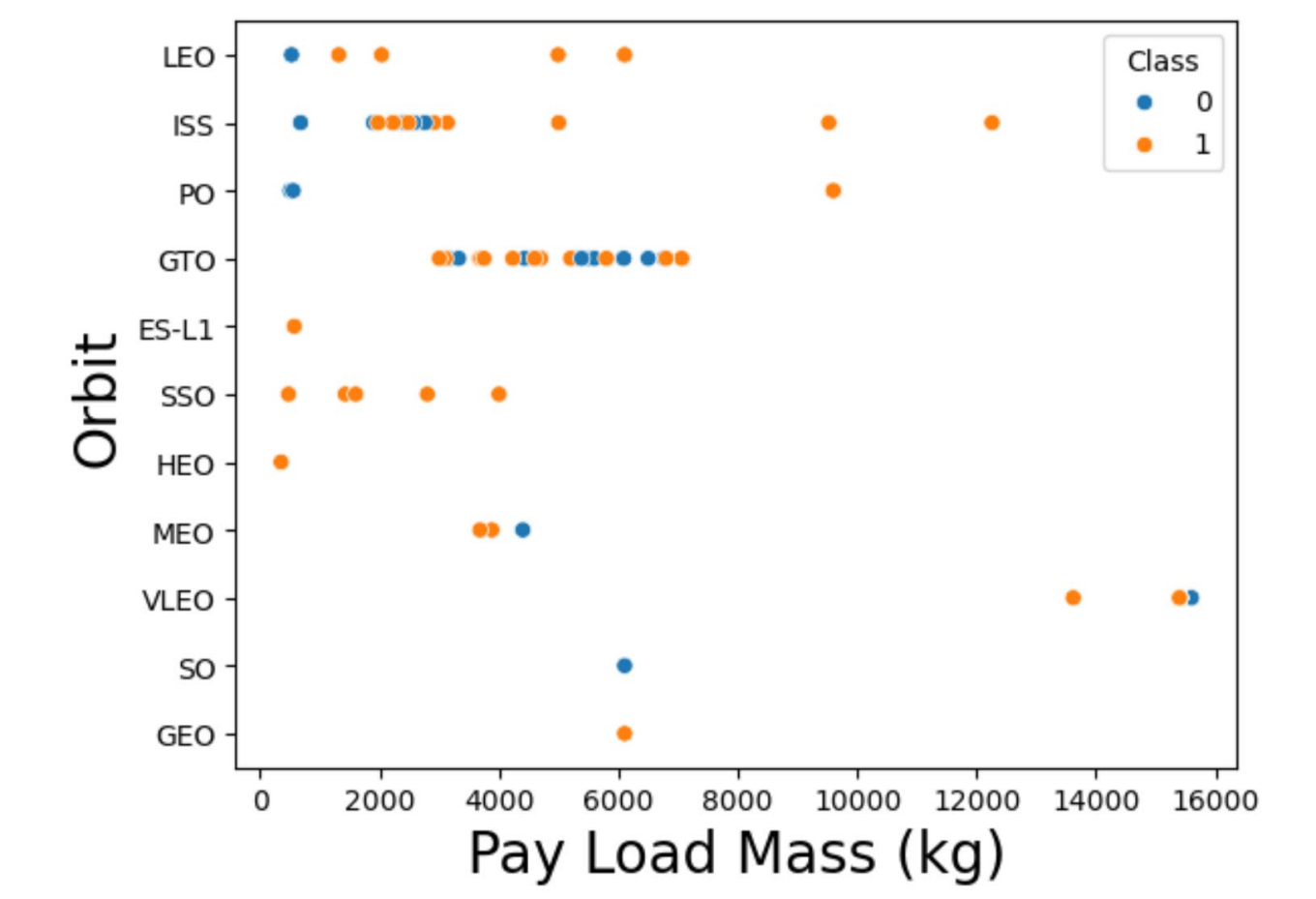
Flight Number vs. Orbit Type



- In the LEO orbit the landing tends to be successful with increasing the number of flight. Likewise VLEO while SSO always shows positive landing. Finally there seems to be no relationship between flight number and success rate in GTO orbit.


```
sns.catplot(y="Orbit", x="FlightNumber", hue="Class", data=df)
```

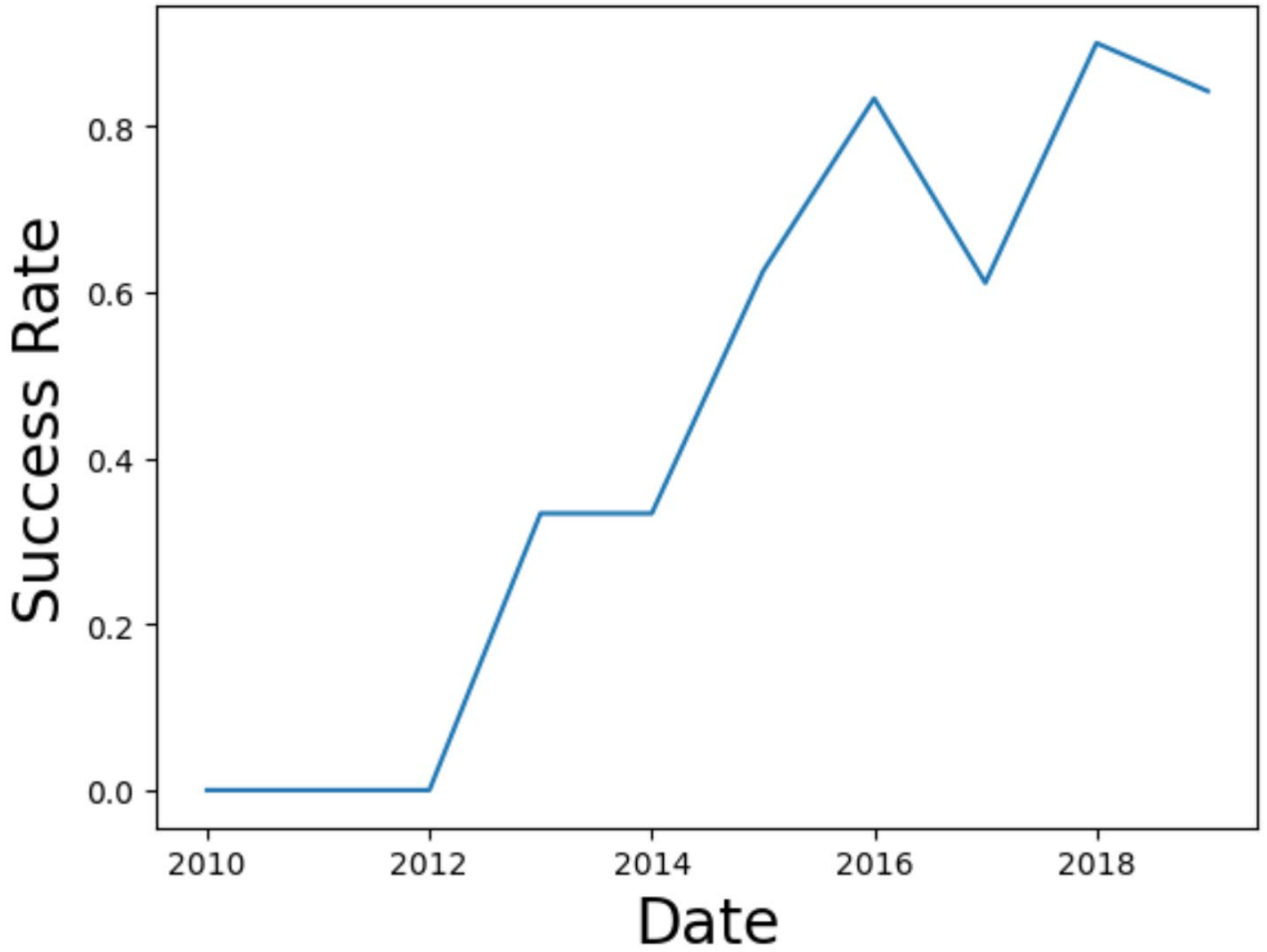
Payload vs. Orbit Type




```
sns.scatterplot(y="Orbit", x="PayloadMass", hue="Class", data=df)
```

- GTO suPolar, LEO and ISS show to have succesful landing with big payloads
- Success rate does not seem to be influenced by Pay Load Massive. SSO is always positive regardless of the Mass while for the others, we have too few records to identify a real trend.

Launch Success Yearly Trend



- As of 2013 it is noticed that the success rate kept increasing.

```
plt.plot(years, mean_yearly_success)
```

All Launch Site Names

- Query

```
%sql select DISTINCT "Launch_Site" from SPACEXTABLE order by 1
```

- Result

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

- **Explanation**

The query retrieves the names of the unique launch sites in the space mission thanks the use of DISTINCT which select unique values.

Launch Site Names Begin with 'CCA'

- Query

```
%sql select * from SPACEXTABLE where "Launch_Site" like 'CCA%' LIMIT 5
```

- Result

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

- Explanation

The query displays 5 records where launch sites begin with the string 'CCA'; LIMIT makes that only 5 sites are shown; the LIKE "CCA%" select all names beginning with "CCA"

Total Payload Mass

- Query

```
%sql select "Launch_Site", sum("PAYLOAD_MASS__KG_") from SPACEXTABLE where "Customer" like '%CRS%'
```

- Result

Launch_Site	SUM(PAYLOAD_MASS__KG_)
CCAFS LC-40	48213

- Explanation

Displays total payload mass carried by boosters launched by NASA (CRS) by applying the SUM to calculate the total value and LIKE to limit the choice to NASA (CRS)

Average Payload Mass by F9 v1.1

- Query

```
%sql select "Booster_Version",  AVG("PAYLOAD_MASS__KG_") from SPACEXTABLE where  "Booster_Version" = 'F9 v1.1'
```

- Result

Booster_Version	AVG("PAYLOAD_MASS__KG_")
F9 v1.1	2928.4

- Explanation

Display average payload mass carried by booster version F9 v1.1 with the use of AVG to calculate the average value and LIKE “F9 v1.1” to specify the booster version

First Successful Ground Landing Date

- Query

```
%sql select MIN("Date") from SPACEXTABLE where  "Landing_Outcome" = 'Success (ground pad)'
```

- Result

MIN("Date")
2015-12-22

- Explanation

List the date when the first successful landing outcome in ground pad was achieved using MIN to find the earliest date and WHERE clause to specify that the landing outcome is in ground pad

Successful Drone Ship Landing with Payload between 4000 and 6000

- Query

```
%sql select "Booster_Version", "PAYLOAD_MASS__KG_", "Landing_Outcome" from SPACEXTABLE where "Landing_Outcome" = 'Success (drone ship)' and
"PAYLOAD_MASS__KG_" > 4000 and "PAYLOAD_MASS__KG_" < 6000
```

Booster_Version	PAYLOAD_MASS__KG_	Landing_Outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

- Explanation

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000 thanks to where clause that limit Landing outcome to ‘Success(drone ship)’and Payload Mass in the mentioned interval

Total Number of Successful and Failure Mission Outcomes

- Query

```
%sql select Mission_Outcome, count(Mission_Outcome)  from SPACEXTABLE group by Mission_Outcome
```

- Result

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

- Explanation

List the total number of successful and failure mission outcomes using the COUNT to pick the total number of mission outcomes and GROUP BY clause to have the results differentiated by the mission outcomes

Boosters Carried Maximum Payload

- Query

```
%sql select DISTINCT "Booster_Version", "PAYLOAD_MASS__KG_" from SPACEXTABLE where  "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_")
from SPACEXTABLE) ORDER BY 1
```

- Result

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

- Explanation

List the names of the booster versions which have carried the maximum payload mass by using a subquery clause in the where clause to limit the choice to the maximum payload mass

2015 Launch Records

- Query

```
%sql SELECT substr(Date, 6,2) as month, substr(Date,0,5) as year, "Landing_Outcome","Booster_Version", "Launch_Site" from SPACEXTABLE
where substr(Date,0,5)='2015' and "Landing_Outcome" = 'Failure (drone ship)'
```

- Result

month	year	Landing_Outcome	Booster_Version	Launch_Site
01	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Explanation

The query returns the 2015 Launch Records by using in the where clause the date substring and limiting the results to the failure landing outcomes in drone ship.

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

- Query

```
%sql SELECT "Landing_Outcome",count("Landing_Outcome") as count, "Date" from SPACEXTABLE
 where date between '2010-06-04' and '2017-03-20' group by "Landing_Outcome" order by count DESC
```

- Result

Landing_Outcome	count	Date
No attempt	10	2012-05-22
Success (drone ship)	5	2016-04-08
Failure (drone ship)	5	2015-01-10
Success (ground pad)	3	2015-12-22
Controlled (ocean)	3	2014-04-18
Uncontrolled (ocean)	2	2013-09-29
Failure (parachute)	2	2010-06-04
Precluded (drone ship)	1	2015-06-28

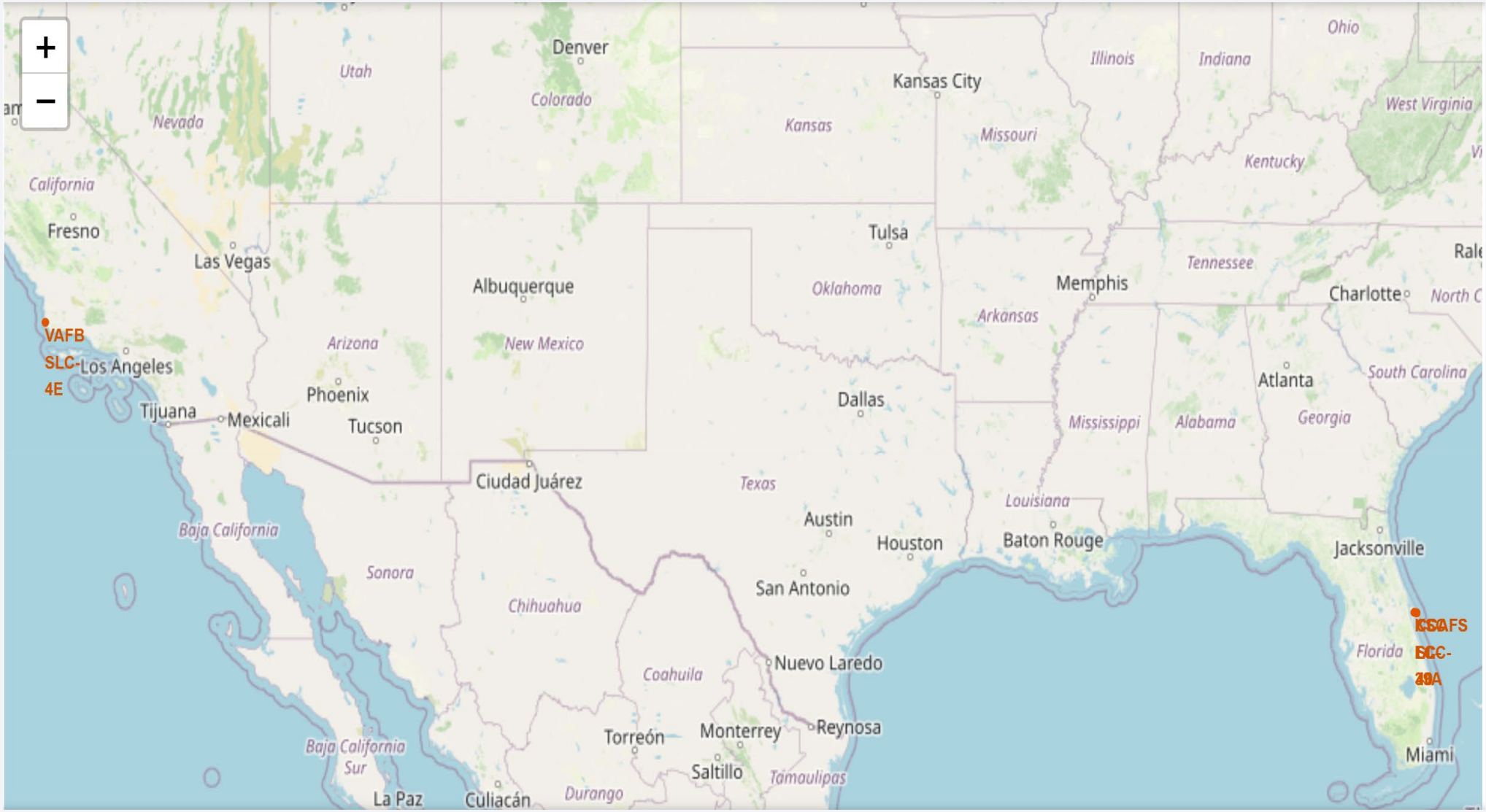
- Explanation

The query Ranks the count of landing between the date 2010-06-04 and 2017-03-20, in descending order by using the count and grouping the result by the Landing Outcome within the desired time frame.

Section 3

Launch Sites Proximities Analysis

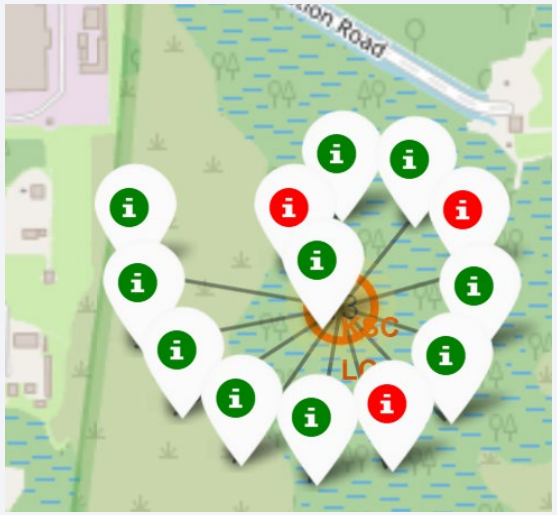
Folium Map of United States with all the Launch Sites marked



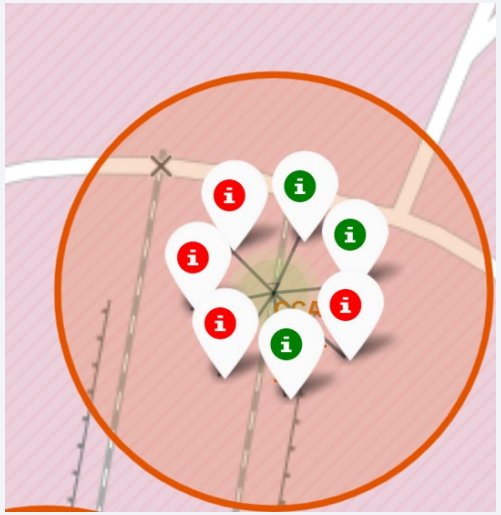
- Findings: All launch sites are not close to the Equator line and they are all in very close proximity to the coast to reduce the payload costs involved in first stage landing on land.

Mark Success/Failed Launches For Each Site

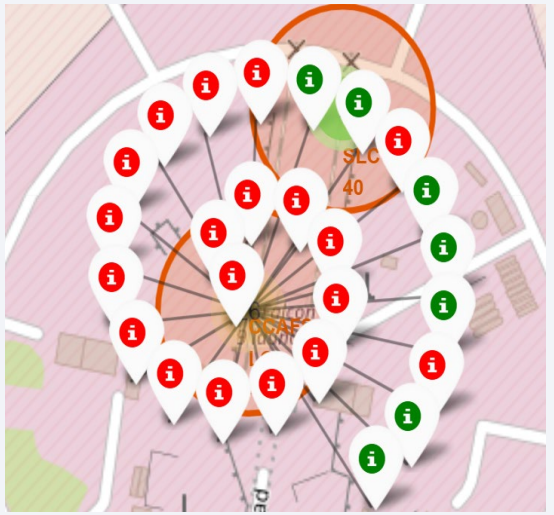
KSC LC-39A



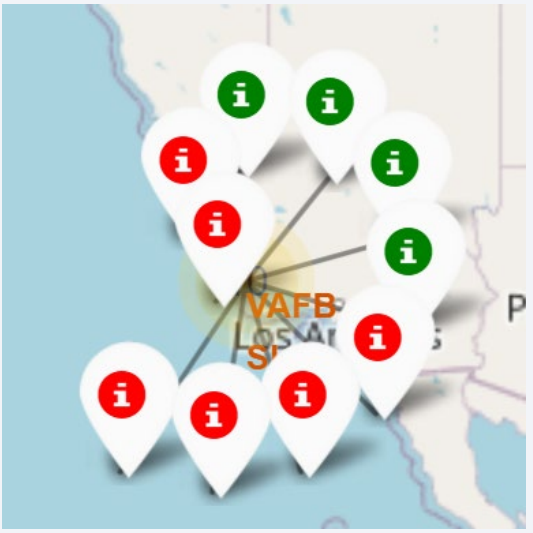
CCAFS SLC-40



CCAFS LC-40



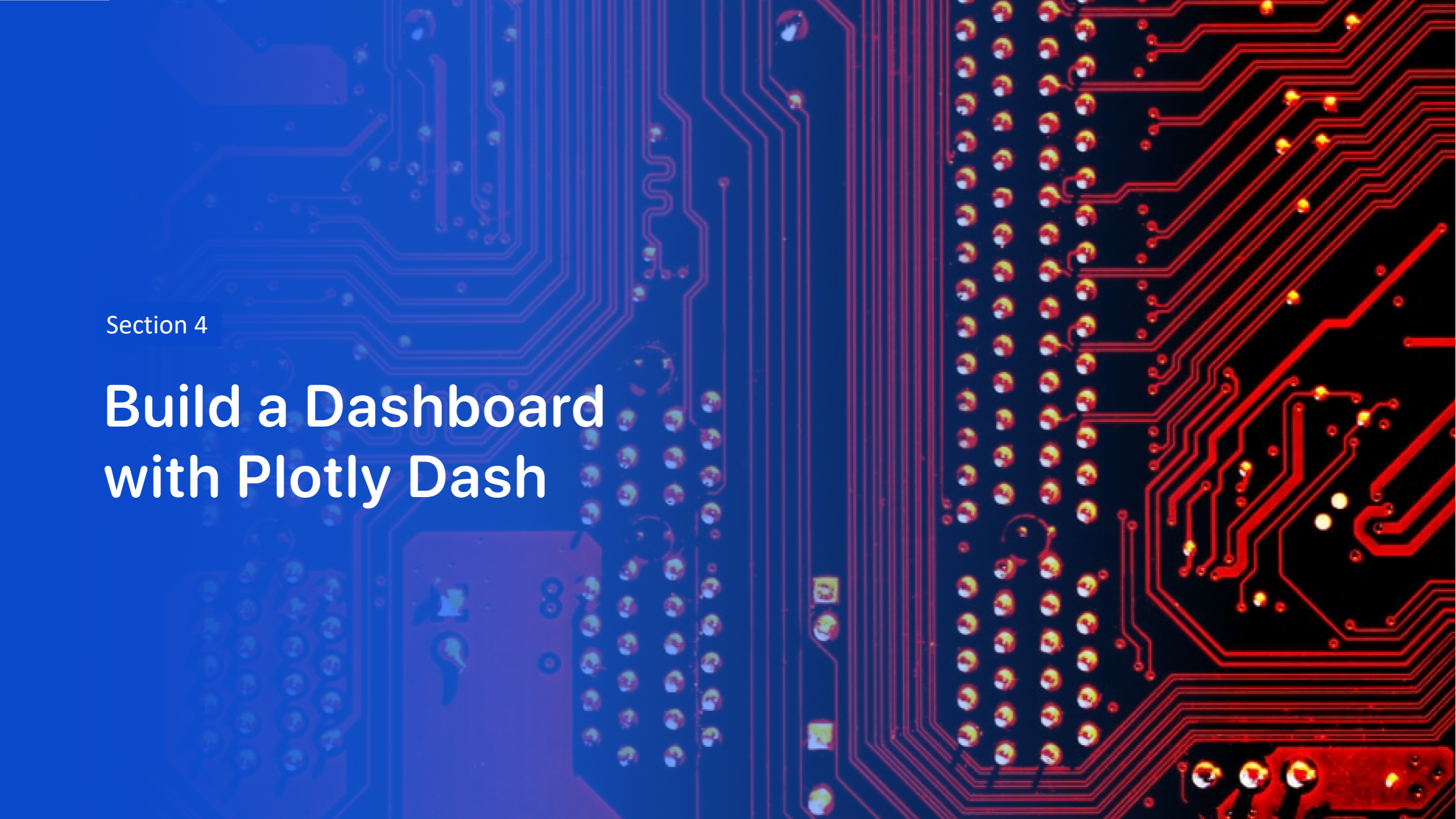
VAFB SLC-4E



Findings: KSC LC-39A has the highest success rate; the others have relatively low success rate.

Distances Between Launch Sites to Its Proximities

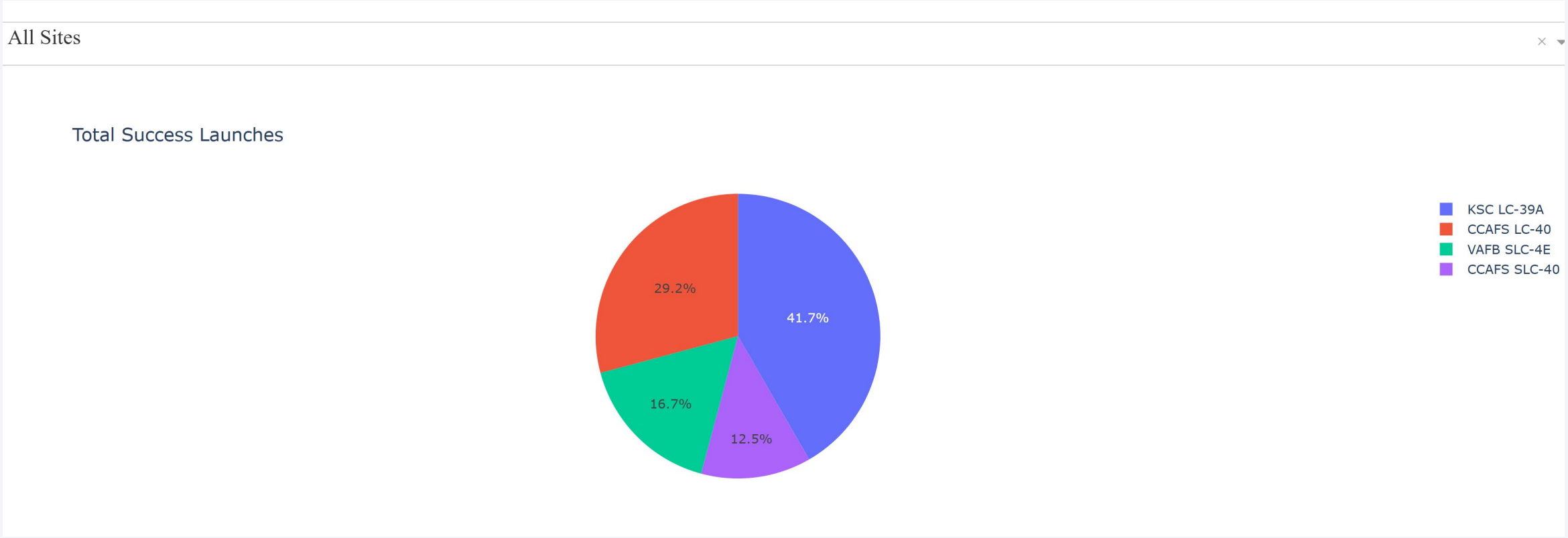
Findings: Launch sites are in close proximity to the coastline but not to railways nor highways. Additionally they keep certain distance away from cities.



Section 4

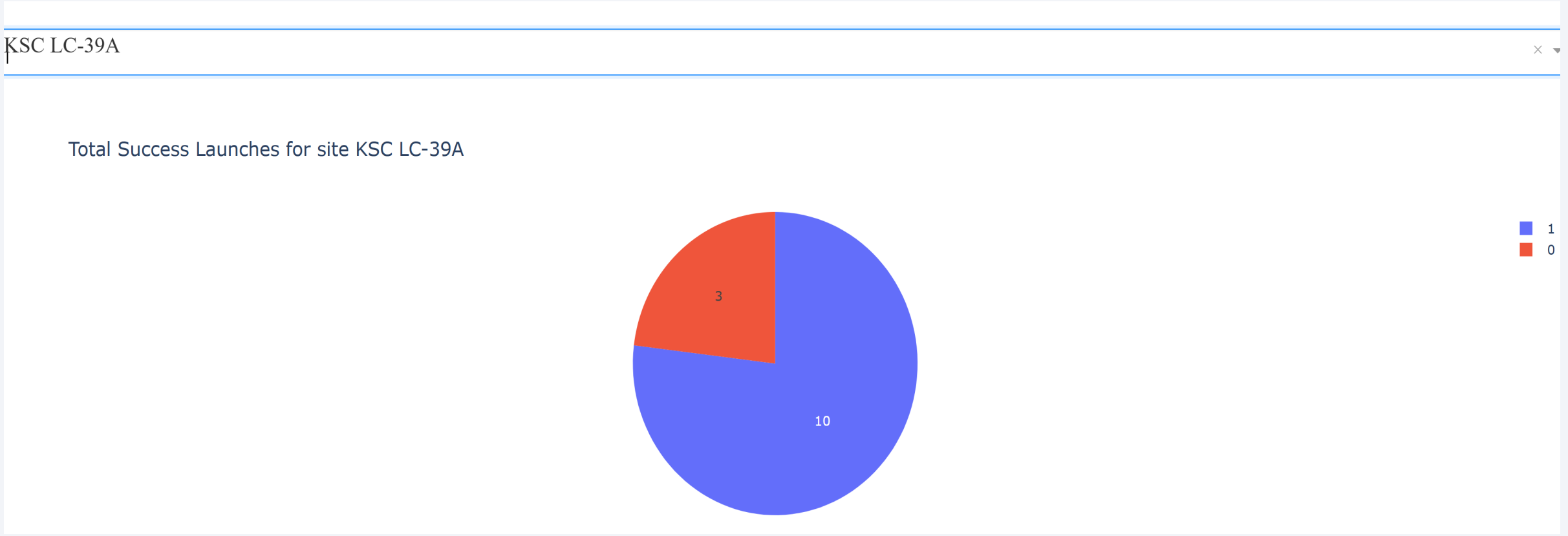
Build a Dashboard with Plotly Dash

Total Success Launches For All Sites



Findings: KSC LC-39A has both the highest launch success rate (41,7%) and the largest successful launches (10)

Launch Site with highest launch success ratio



Findings: KSC LC-39A had 10 successful launches ratio out of 13

Correlation Between Payload Mass and Success 1

Findings: FT and B5 booster version have the highest launch success rate even though there is only one B5 launch ever

Findings: The range with the highest launch success rate is the 2000 -:- 3800 kg of Payload Mass

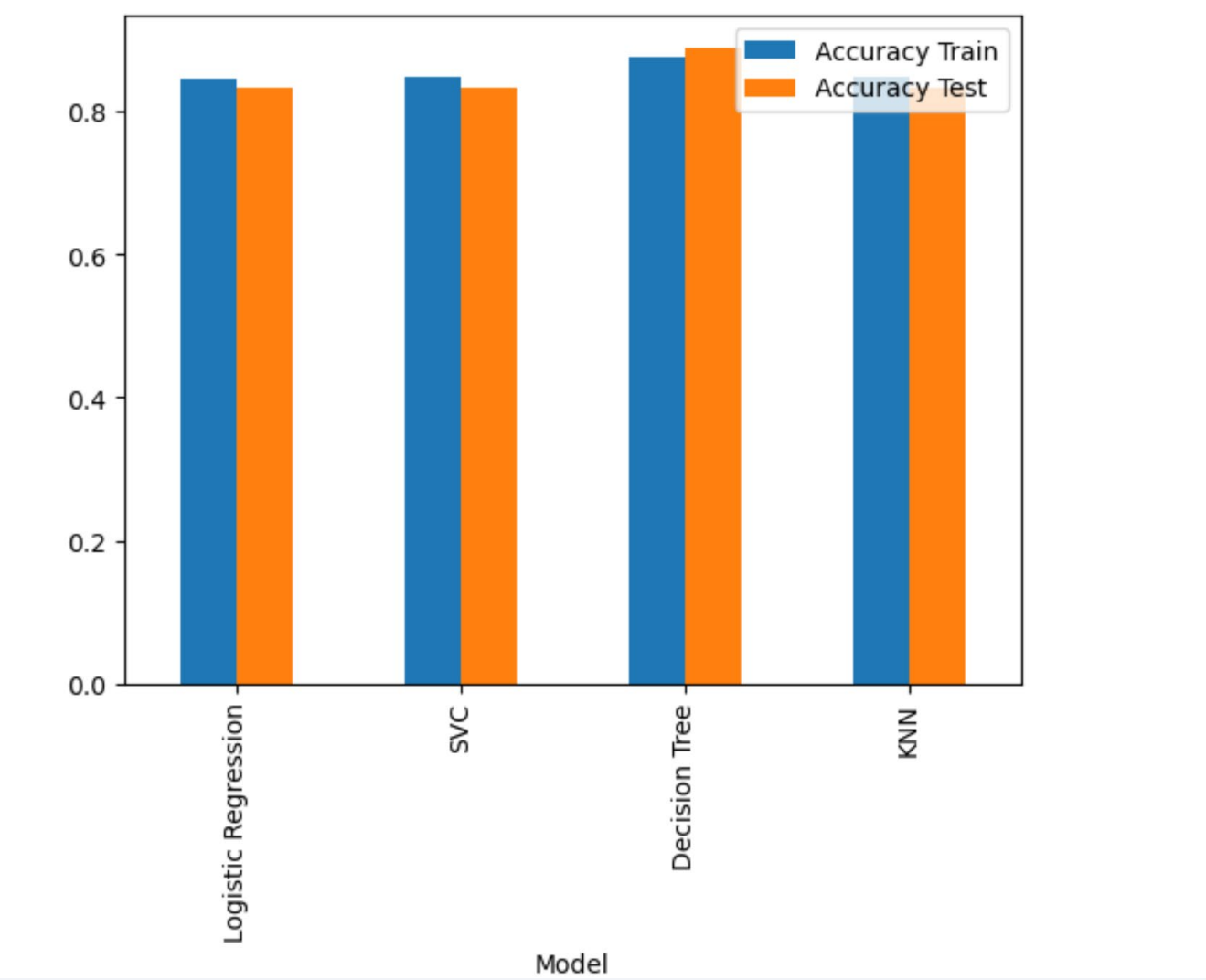
Correlation Between Payload Mass and Success 2

Findings: The range with the lowest launch success rate is any Payload Mass greater than 5.500 kg

Section 5

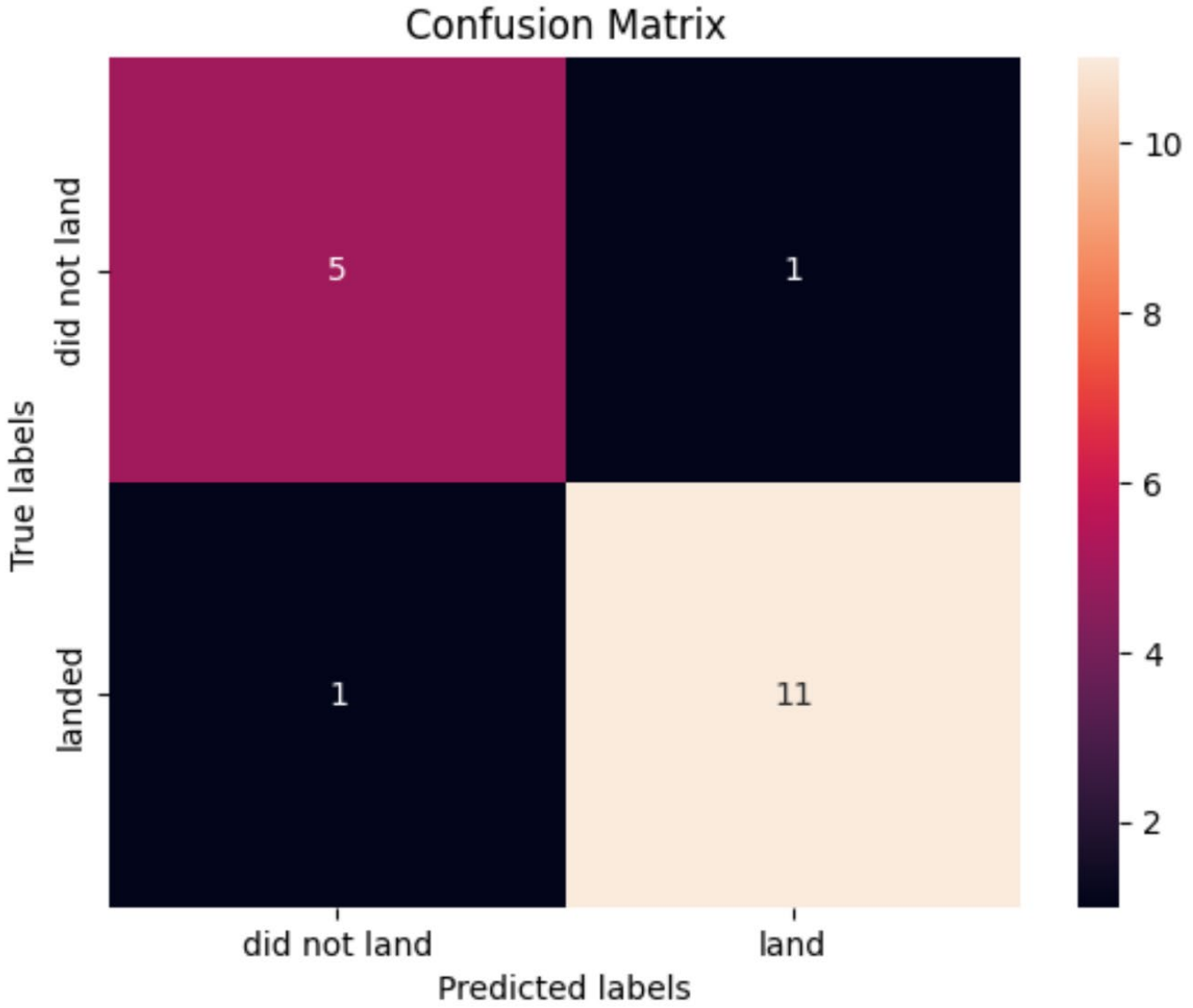
Predictive Analysis (Classification)

Classification Accuracy



- The bar chart shows that the decision tree is the best model because it has the highest training and test accuracy scores

Confusion Matrix



- For the Decision Tree Model, out of 18 predictions, we had 1 False Positive, 1 False Negative, 11 True Positive and 5 True Negative.
- The recall rate is quite high (recall=TP/(TP+FN)=11/12 = 0.916)

Conclusions

- SpaceX has steadily improved its launch success rate over the years.
- Launches to destinations like Earth-Sun L1, Geostationary Earth Orbit (GEO), High Earth Orbit (HEO), and Sun-Synchronous Orbit (SSO) tend to have the highest success rates.
- In general the most successful launches occur when the payload mass falls within the range of 2000 to 3800 kg. but launches to GTO, Sun-Polar, LEO, and ISS have a high success rate for landing even with substantial payloads
- Among SpaceX launch sites, KSC LC-39A boasts the most successful launches
- All launch sites are not close to the Equator line and they are all in very close proximity to the coast to reduce the payload costs involved in first stage landing on land
- Among the models tested, the tree classifier demonstrated the highest predictive accuracy for launch success.

Appendix

- A full list of datasets, with links to the respective notebooks used for this SpaceX project is available at this link:

https://github.com/bejaflor66/applied-data-science-capstone

Thank you!