IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Alan Gordon Caswell
July 6, 2024

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

- This project is designed to utilize publicly-available data regarding SpaceX launches to assist Space Y calculate the cost of launches performed by SpaceX as it strives to compete with SpaceX.
- The primary problem that is attempting to be answered is what the cost is per launch by SpaceX.
- The secondary problem is predicting whether the first stage will be reused between launches
- These problems will both be solved through the utilization of a machine learning model provided with public data about historical launches by SpaceX.

Section 1

Methodology

Methodology

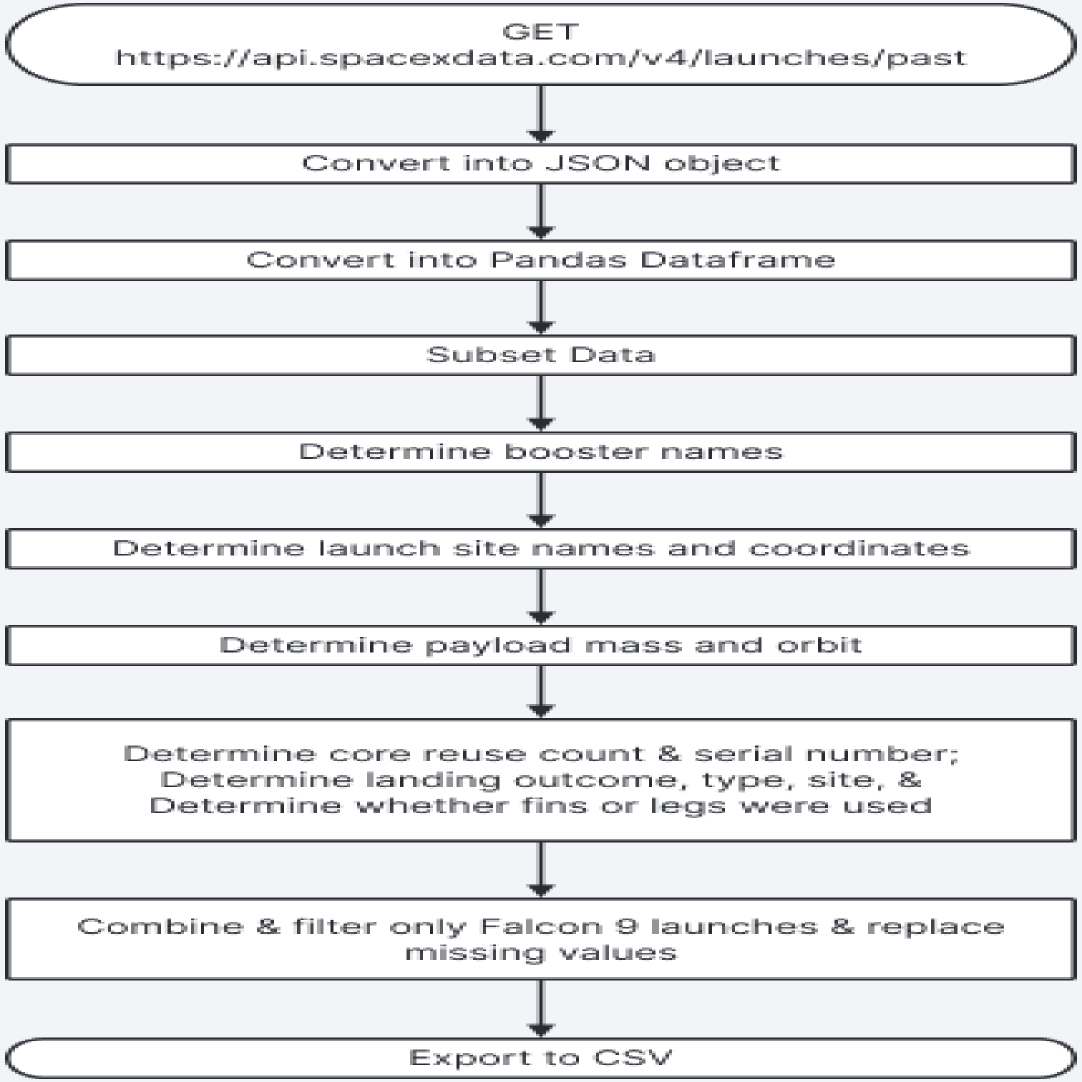
- Data collection using SpaceX's public REST API and through webscraping Wikipedia articles detailing SpaceX launch data
- Perform data wrangling using Python and Pandas
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - This included building, tuning, and evaluating classification models to determine the optimal model for use

Data Collection

- Data sets were collected using two complementary methods:
  - Requesting data from the SpaceX REST API
  - Webscraping data found on Wikipedia regarding SpaceX launches
- In both cases, the data received was parsed into Pandas dataframes for further analysis

Data Collection – SpaceX API

- REST API Get call placed to SpaceX API, seeking past launch data
- Received data converted into a JSON object, followed by converted into a Pandas Dataframe.
- Converted data subsetted to include only "rocket", "payloads", "launchpad", "cores", "flight number", and "date"
- Additional REST API call placed to determine booster names
- Additional REST API call placed to determine launch site names and coordinates
- Additional REST API call placed to determine payload mass and orbit
- Additional REST API call placed to determine a variety of information about the landing and cores
- All information obtained is combined and filtered to include only Falcon 9 launches, and missing values are replaced with applicable mean values
- https://github.com/GordCaswell/Applied_Data_Science_Specialization/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



GET https://api.spacexdata.com/v4/launches/past

Convert into JSON object

Convert into Pandas Dataframe

Subset Data

Determine booster names

Determine launch site names and coordinates

Determine payload mass and orbit

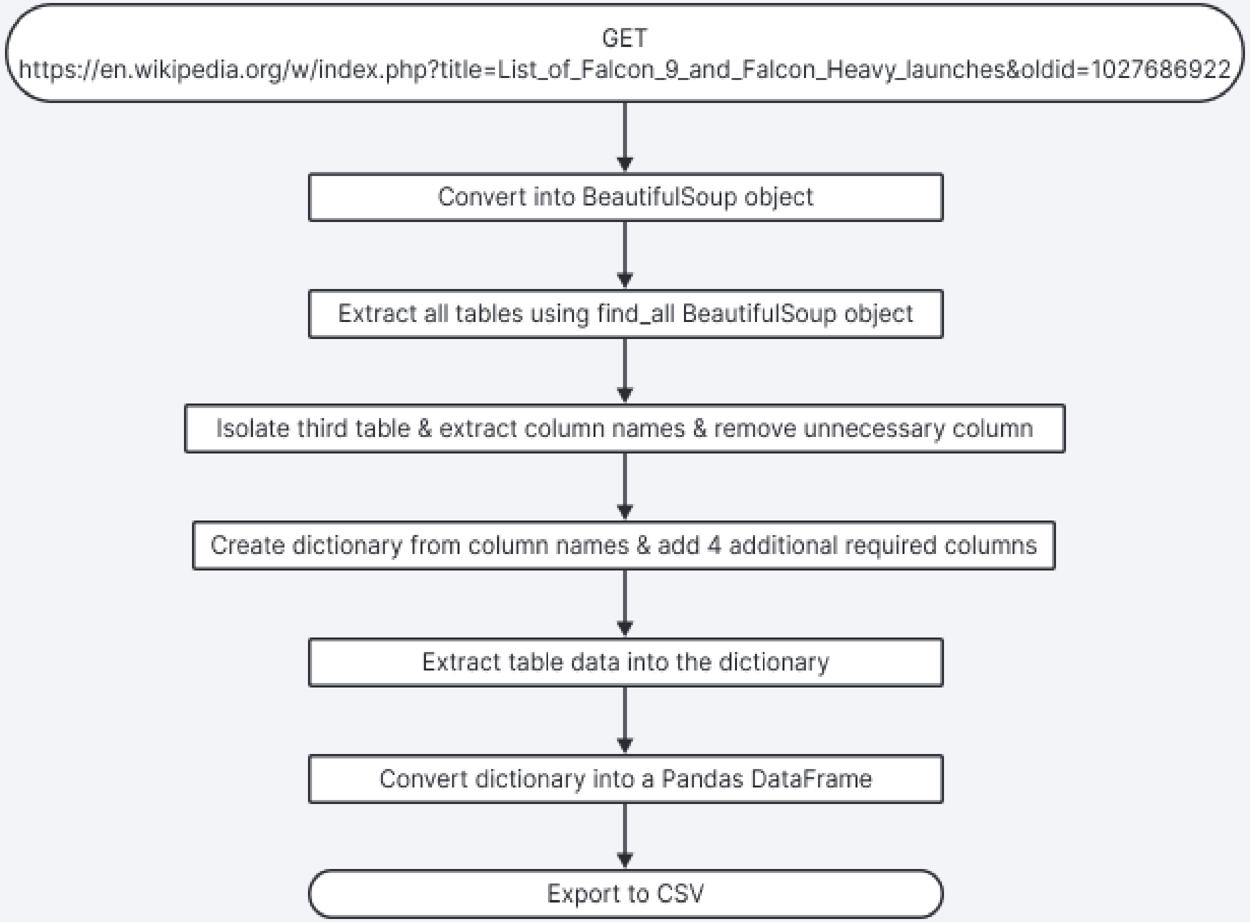
Determine core reuse count & serial number; Determine landing outcome, type, site, & Determine whether fins or legs were used

Combine & filter only Falcon 9 launches & replace missing values

Export to CSV

Data Collection - Scraping

- REST API Get call placed to "List of Falcon 9 and Falcon Heavy launches" Wikipedia page as it was displayed June 9, 2021.
- Received data converted into a BeautifulSoup object.
- Extracted all tables using the find_all BeautifulSoup object.
- Isolated the third table and extracted the column names from that table, removing one column.
- Created a dictionary from the column names, adding 4 additional required columns.
- Extracted table data into the dictionary.
- Converted the dictionary into a Pandas DataFrame.
- Exported to CSV.
- https://github.com/GordCaswell/Applied_Data_Science_Specialization/blob/main/notebooks/jupyter-labs-webscraping.ipynb



GET https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

Convert into BeautifulSoup object

Extract all tables using find_all BeautifulSoup object

Isolate third table & extract column names & remove unnecessary column

Create dictionary from column names & add 4 additional required columns

Extract table data into the dictionary

Convert dictionary into a Pandas DataFrame

Export to CSV

Data Wrangling

- Loaded DataFrame created from previously-scraped Wikipedia data.
- Calculated the number of launches from each launch site.
- Calculated the number and occurrence of each type of orbit.
- Calculated the number and occurrence of mission outcomes.
- Grouped mission outcomes based on failure or success of the mission.
- Added a column to the DataFrame to denote failure or success of the mission using a boolean variable assigned 0 or 1 respectively.
- Exported to CSV.
- https://github.com/GordCaswell/Applied_Data_Science_Specialization/blob/main/notebooks/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Several charts were utilized to examine the relationship between various factors that could affect mission success.
- These included examining mission success as applied to Flight Numbers versus Payload Mass, to Flight Numbers vs Launch Sites, to Launch Sites vs Payload Mass, to Payload Mass vs Orbit type, the success rate of each Orbit type, and the success rate by year.
- https://github.com/GordCaswell/Applied_Data_Science_Specialization/blob/main/notebooks/edadataviz.ipynb

EDA with SQL

- Previously-saved CSV file converted into SQL database
- Displayed each Launch Site's name
- Displayed 5 records launched from sites starting with "CCA"
- Displayed the total payload mass launched by NASA (CRS)
- Displayed the average payload mass launched by booster F9 v1.1
- Displayed the date of the first successful landing on a ground landing pad
- Displayed the boosters that have successfully landed on drone ships when launching payloads in the range 4000-6000KG
- Displayed the number of failed and successful missions
- Displayed the boosters that have launched the maximum payload
- Displayed records of failed drone landings in 2015 by month
- Ranked landing outcomes between June 4, 2010 and March 20, 2017
- https://github.com/GordCaswell/Applied_Data_Science_Specialization/blob/main/notebooks/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Launch Site Circles, Launch Coordinate Markers, Marker Clusters, & Distance Markers were plotted on a Folium map
- Launch Site circles were plotted on the map to denote the location and name of each launch site, with the exact coordinates of each launch being plotted as part of a marker cluster at each launch site. Additionally, distance markers were plotted on the map to denote the distance from coastlines, railroads, highways, and cities as relating to each launch site.
- https://github.com/GordCaswell/Applied_Data_Science_Specialization/blob/main/notebooks/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

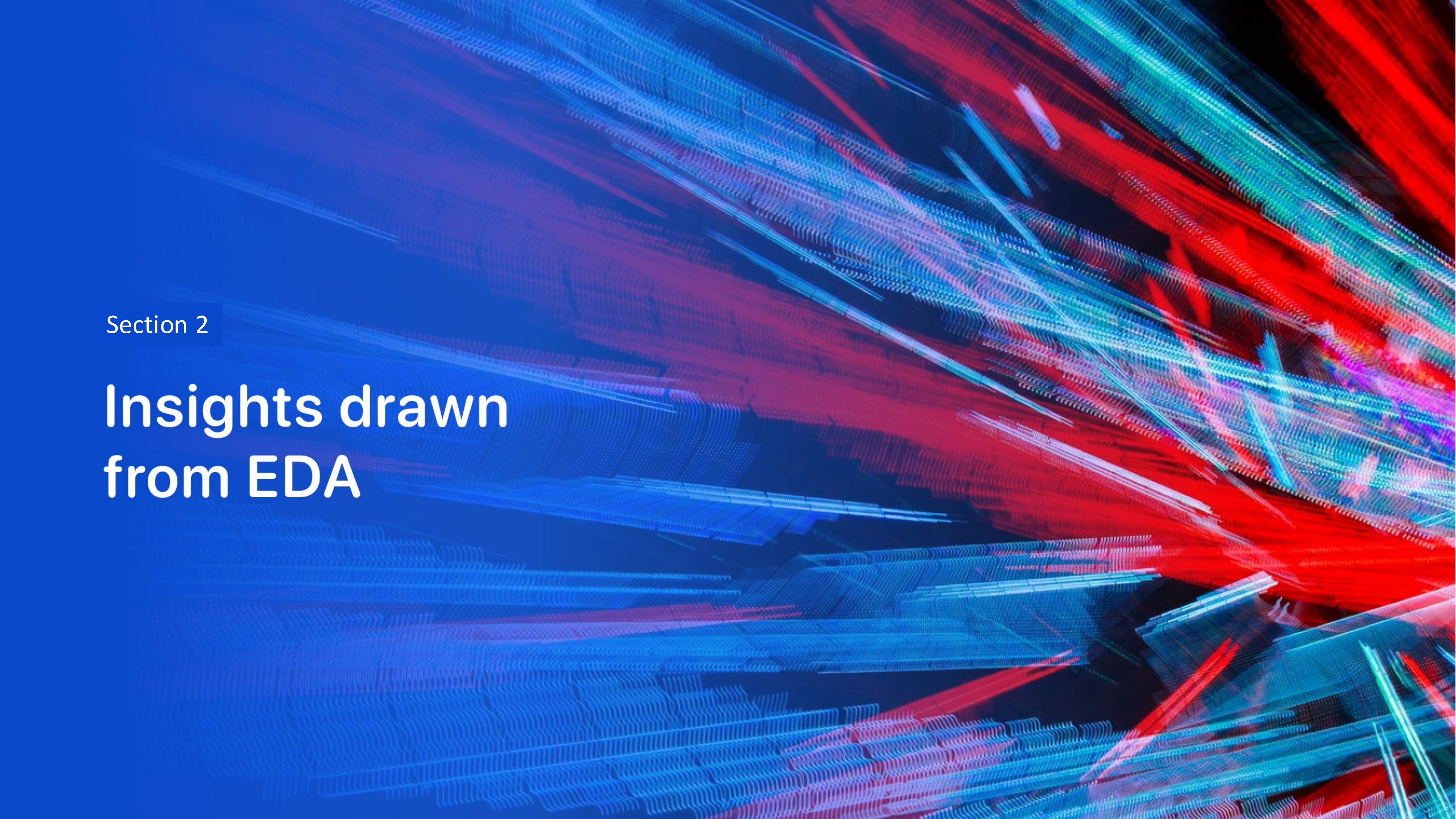
- There are two interactive selectors included in the Dashboard:
  - A dropdown allowing selection of by launch site, defaulted to include all launch sites
  - A range slider to select payload mass ranges, defaulted to show all payload masses
- These selectors drive the display of two plots:
  - A pie chart displaying the number and percentage of successful launches
    - Note that when considering all launch sites, this number and percentage is a total and percentage of those successful launches, rather than the number and percentage of successful launches against the total launches at each site when looking at individual sites
  - A scatter plot displaying the correlation between payload masses and mission success
- https://github.com/GordCaswell/Applied_Data_Science_Specialization/blob/main/code/spacex_dash_app.py

Predictive Analysis (Classification)

- Converted the mission success column from the existing Pandas DataFrame into a NumPy array.
- Standardized and Transformed existing DataFrame.
- Split DataFrame and NumPy into training and test data sets.
- Created and tested logistic regression, support vector machine, decision tree, and k next nearest objects, creating a Grid SearchCV obect for each.
- Fit each to the best parameters for each object.
- Tuned each object's hyperparameters, calculated the accuracy of each object, and plotted the confusion matrix for each.
- Lastly, listed the accuracy of each contrasted with the others.
- https://github.com/GordCaswell/Applied_Data_Science_Specialization/blob/main/notebooks/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

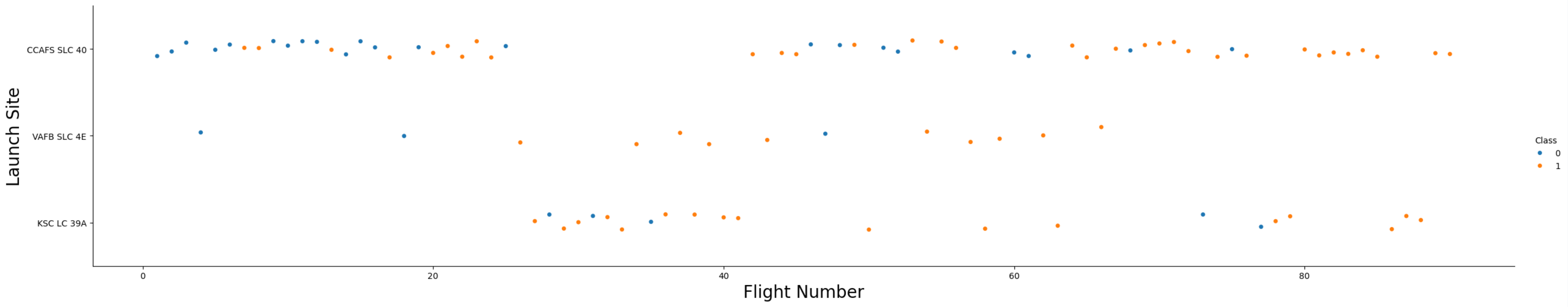
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



Section 2

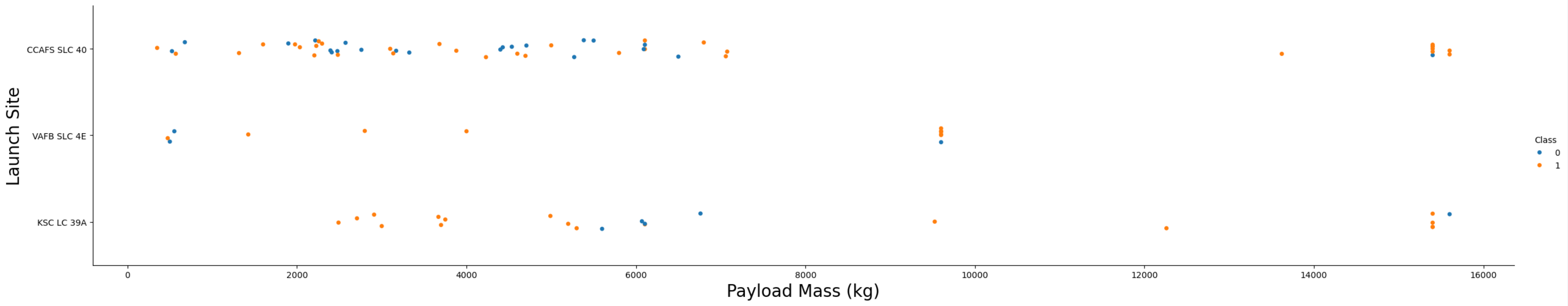
Insights drawn from EDA

Flight Number vs. Launch Site



As launches increased in number, a higher success rate was experienced for launches at all sites

Payload vs. Launch Site

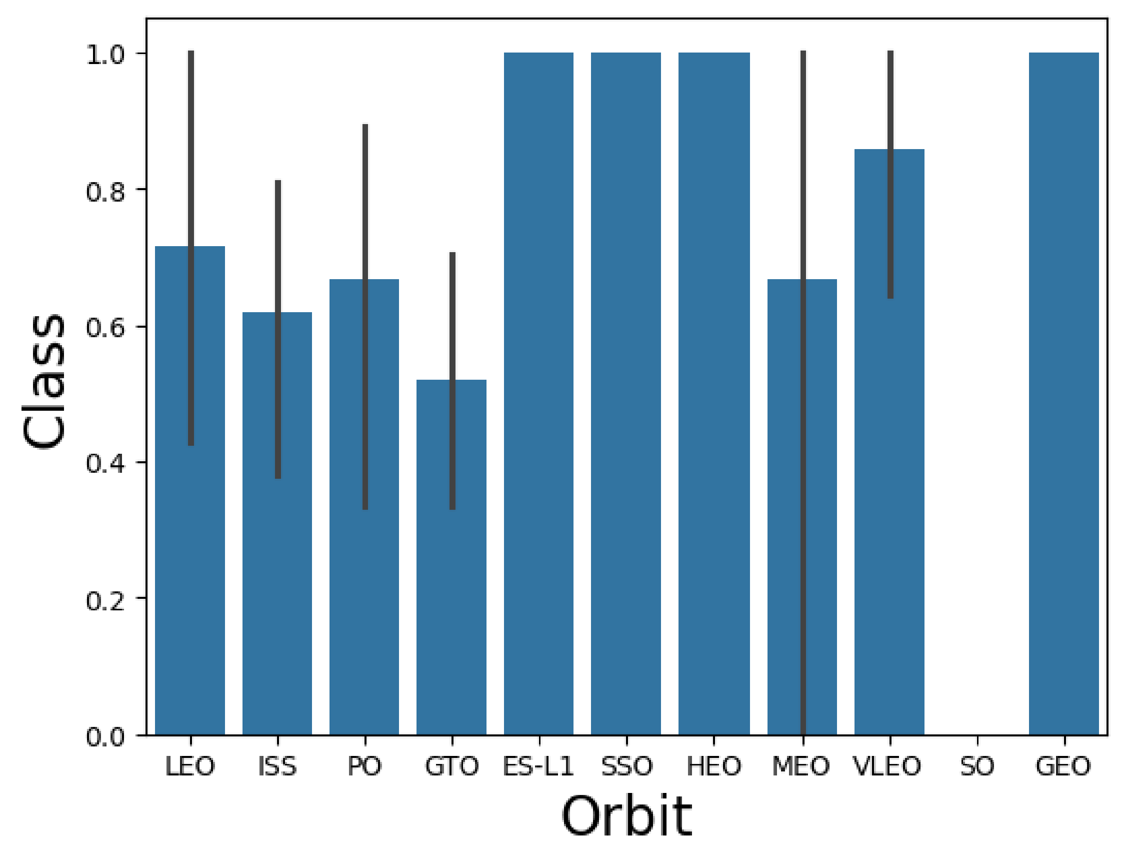


It is evident that the VAFB launch site launches only lighter-weight payloads, and the CCAFS launch site does not have any launches with payloads in the 8000-1000KG range

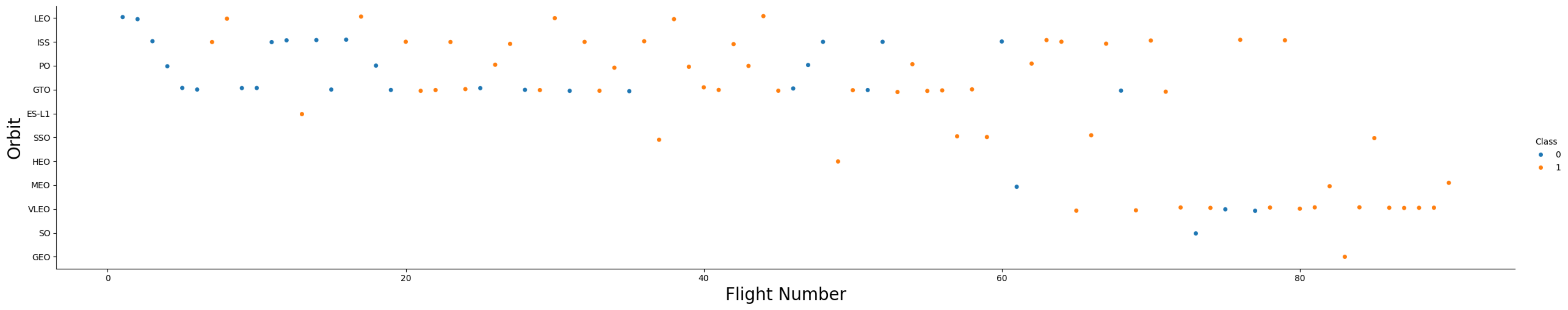
Success Rate vs. Orbit Type

The most successful orbit types are to:

- Lagrange Point 1 orbits
- Heliosynchronous orbits
- Highly elliptical orbits
- Geosynchronous orbits



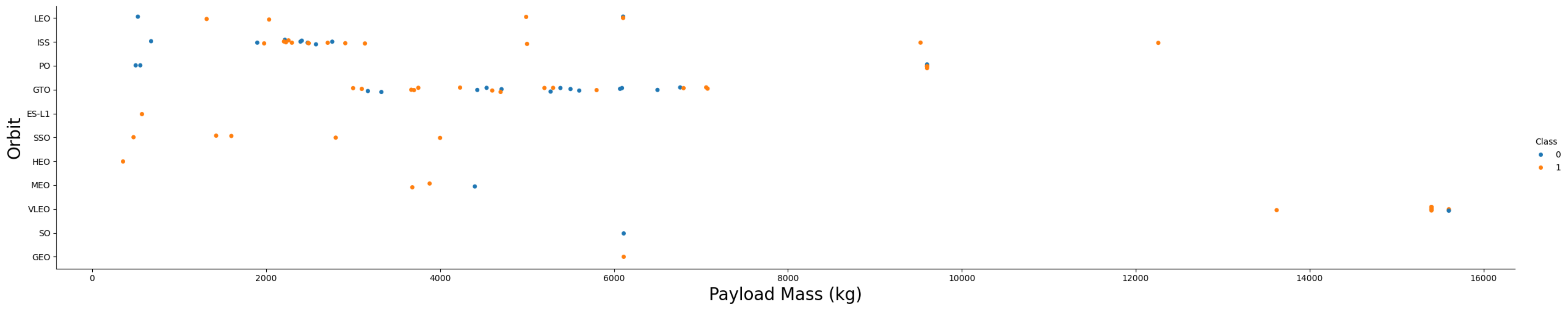
Flight Number vs. Orbit Type



This plot demonstrates that it is evident that sun-synchronous orbits have been consistently successful, but were not attempted initially.

Also, there are several orbit types that were not attempted until later on in the launch attempts.

Payload vs. Orbit Type



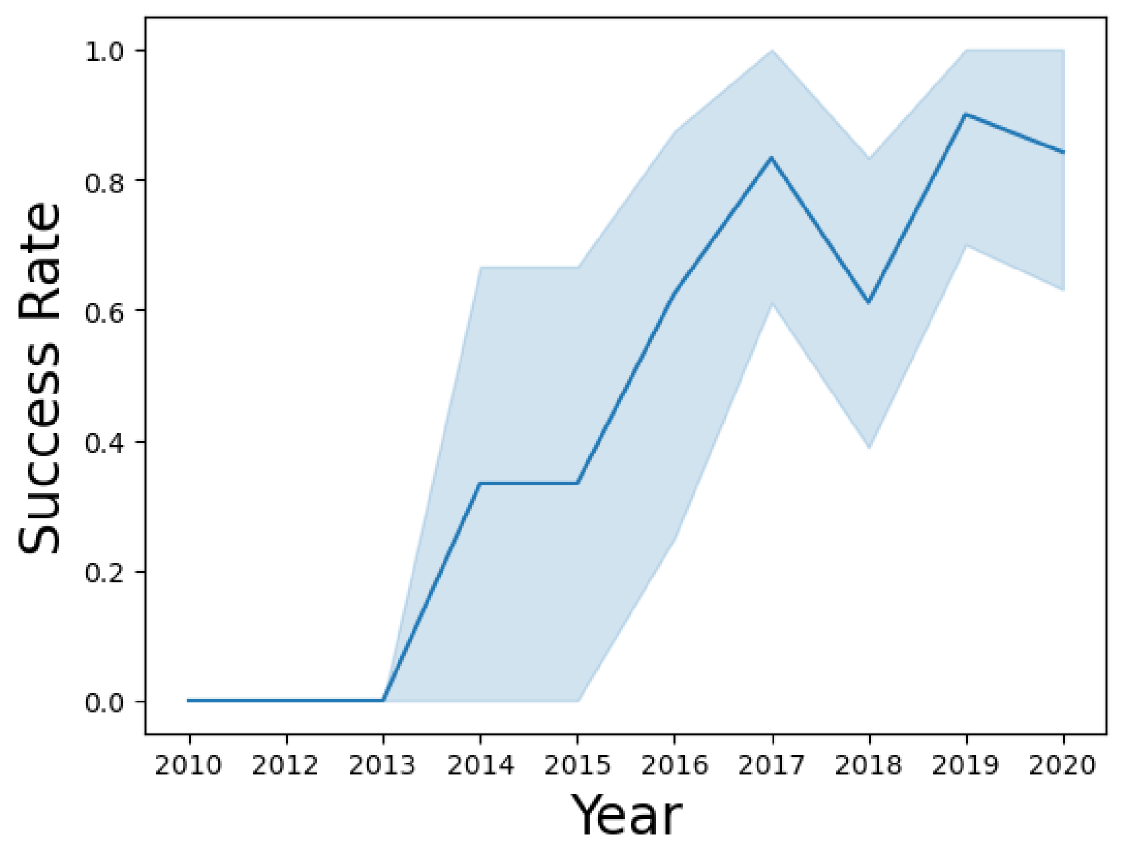
It is clear from this plot that heavy payloads are only launched to a limited number of orbit types.

Certain orbit types have a higher success rate with lighter payloads, while at the same time, others have more success with heavier payloads.

Launch Success Yearly Trend

As the years progress launch success steadily increases year-over-year.

An exception to this is observed in 2018, but success bounces back in 2019.



All Launch Site Names

- The SQL query %sql select DISTINCT "Launch_Site" from SPACEXTABLE returns a table as follows:

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

- The listed Launch Sites are the abbreviated notations used in the data representing the four launch sites used.

Launch Site Names Begin with 'CCA'

- The SQL query %sql select * from SPACEXTABLE where "Launch_Site" like "CCA%" limit 5 returns a table as follows:

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	Space X 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

- The listed records represent a selection of 5 records in the database with Launch Site beginning with CCA.

Total Payload Mass

- The SQL query %sql select sum(PAYLOAD_MASS_KG_) as "Total_Payload_Mass" from SPACEXTABLE where "Customer" like "NASA (CRS)" returns a table as follows:

Total_Payload_Mass
45596

- This shows that the total payload mass carried by boosters for the NASA (CRS) customer totals to 45596 KG.

Average Payload Mass by F9 v1.1

- **The SQL query** %sql select avg(PAYLOAD_MASS_KG_) as "Average_Payload_Mass" from SPACEXTABLE where "Booster_Version" like "F9 v1.1%" **returns a table as follows:**

Average_Payload_Mass
2534.6666666666665

- This shows that the average payload mass carried on boosters using booster version F9 v1.1 is 2354.66KG.

First Successful Ground Landing Date

- **The SQL query** %sql select min("Date") as "First_Successful_Outcome" from SPACEXTABLE where "Landing_Outcome" like "Success (ground pad)" **returns a table as follows:**

First_Successful_Outcome
2015-12-22

- This shows that the first successful landing on a ground landing pad occurred on December 22, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

- **The SQL query** %sql select "Booster_Version" from SPACEXTABLE where "Landing_Outcome" like "Success (drone ship)" and PAYLOAD_MASS__KG_ between 4000 and 6000 **returns a table as follows:**

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

- This shows that there are a total of 4 boosters that have successfully landed on a drone ship when launching payloads between 4000 and 6000KG.

Total Number of Successful and Failure Mission Outcomes

- **The SQL query** %sql select count(case when "Landing_Outcome" like "Success%" then 1 end) as "Successful_Outcome", count(case when "Landing_Outcome" like "Failure%" then 1 end) as "Failure_Outcome" from SPACEXTABLE **returns a table as follows:**

Successful_Outcome	Failure_Outcome
61	10

- This shows that there were a total of 61 successful landings in contrast to 10 failed landings.

Boosters Carried Maximum Payload

- **The SQL query** %sql select "Booster_Version" from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE) **returns a table as follows:**

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 1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

- This shows that there were 12 boosters that have launched the maximum payload.

2015 Launch Records

- **The SQL query** %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" like "Failure (drone ship)" **returns records for only two numbered months, but does not return the month names. As a result, the query can be modified to display the month names for these months:** %sql select (case when substr(Date,6,2) like "01" then "January" when substr(Date,6,2) like "04" then "April" end) as month, "Landing_Outcome", "Booster_Version", "Launch_Site" from SPACEXTABLE where substr(Date,0,5)='2015' and "Landing_Outcome" like "Failure (drone ship)" **and returns a table as follows:**

month	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- This shows that there were failed drone ship landings in both January and April 2015

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

- **The SQL query** %sql select "Landing_Outcome", count("Landing_Outcome") as "Count" from SPACEXTABLE group by "Landing_Outcome" order by count("Landing_Outcome") desc **returns a table as follows:**

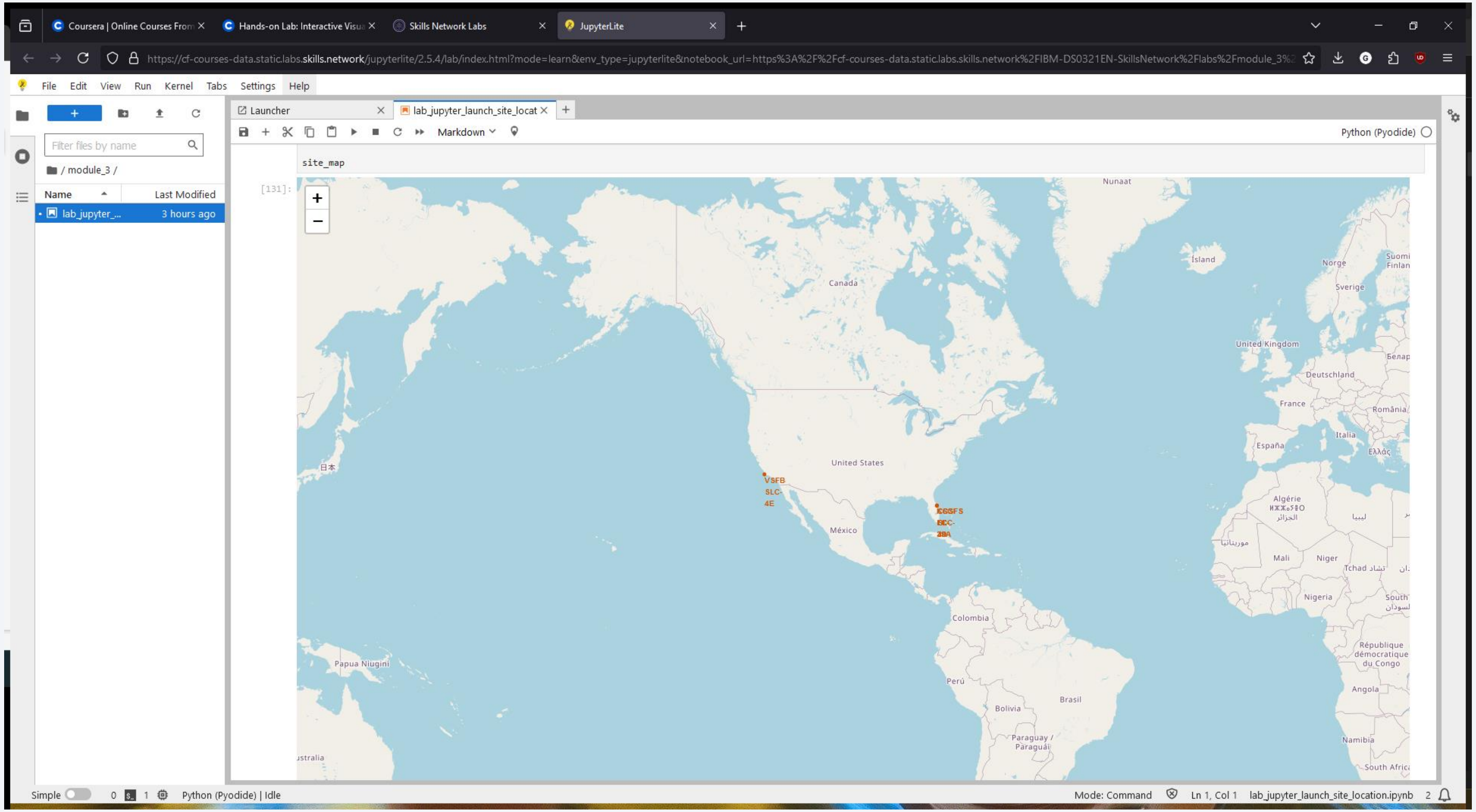
Landing_Outcome	Count
Success	38
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Failure	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1
No attempt	1

- This shows that there were more successful landings, by a significant number, than were failed landings or where no landing was attempted.

Section 3

Launch Sites Proximities Analysis

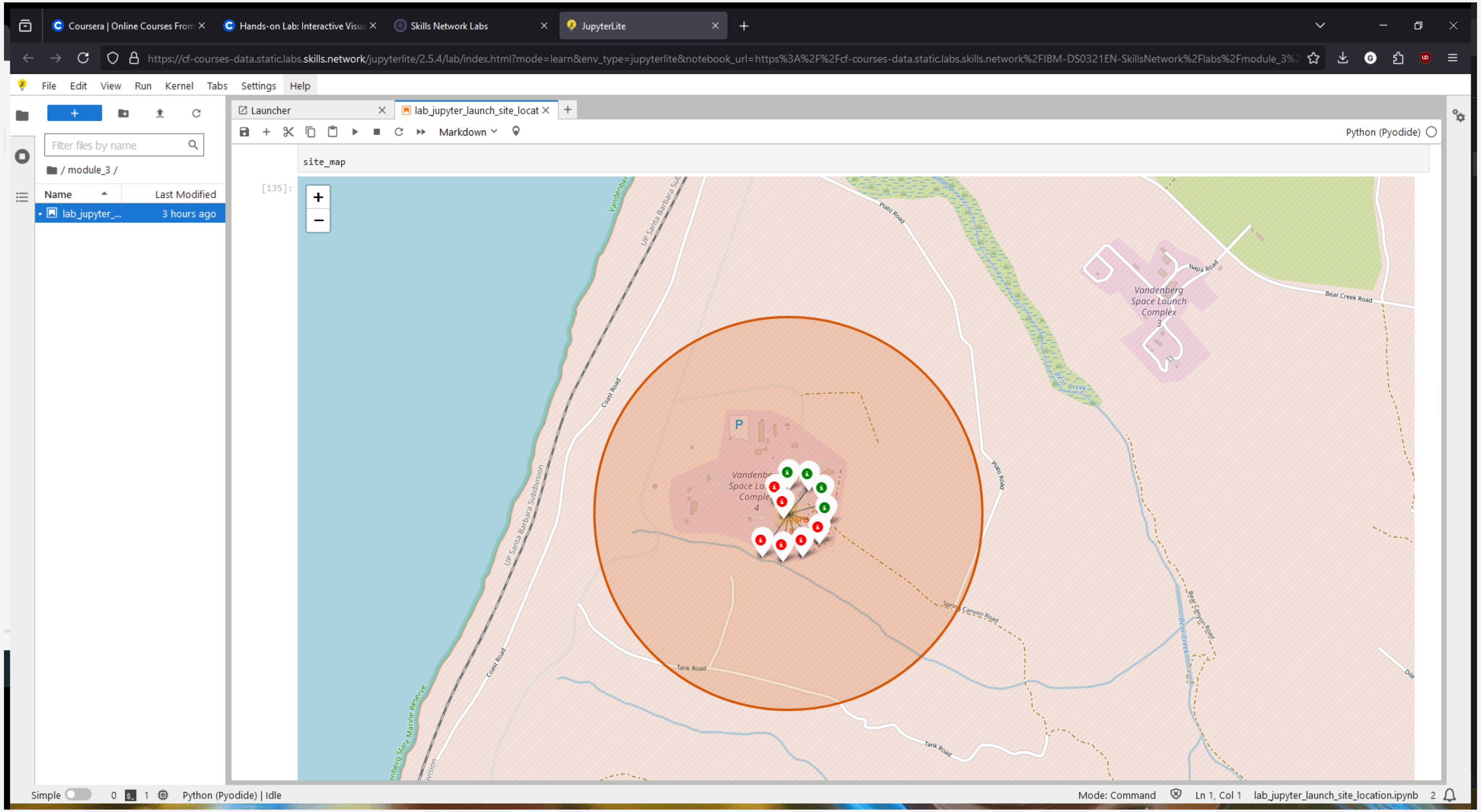
SpaceX Launch Site Locations



```
site_map
```

- This map shows the location of the launch sites on a global-scale map.
- The three Florida-based sites at this zoom level are layered on top of one another, whereas the California-based site is clearly visible on its own.

Launch Outcomes

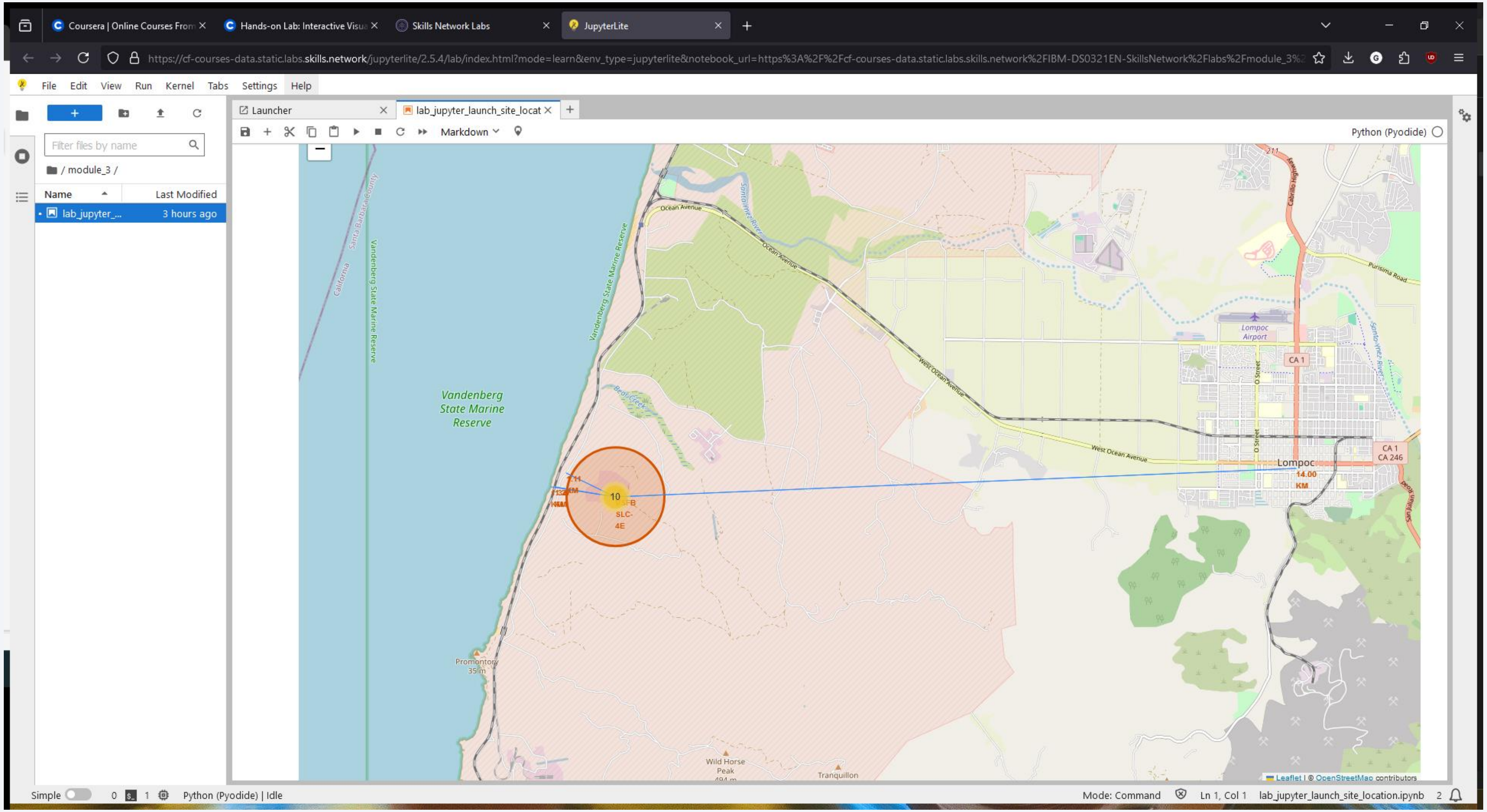


```
site_map
```

Coursera | Online Courses From
Hands-on Lab: Interactive Visu
Skills Network Labs
JupyterLite
https://cf-courses-data.static.labs.skills.network/jupyterlite/2.5.4/lab/index.html?mode=learn&env_type=jupyterlite&notebook_url=https%3A%2F%2Fcf-courses-data.static.labs.skills.network%2FIBM-DS0321EN-SkillsNetwork%2Flabs%2Fmodule_3%
File
Edit
View
Run
Kernel
Tabs
Settings
Help
Filter files by name
/ module_3 /
Name
Last Modified
lab_jupyter_...
3 hours ago
Launcher
lab_jupyter_launch_site_locat
Markdown
Python (Pyodide)
[135]:
Vandenberg Space Launch Complex 3
Vandenberg Space Launch Complex 4
Plato Road
Napa Road
Bear Creek Road
Coast Road
Tank Road
Spring Canyon Road
Bear Canyon Road
UP Santa Barbara Subdivision
Simple
Python (Pyodide) | Idle
Mode: Command
Ln 1, Col 1
lab_jupyter_launch_site_location.ipynb

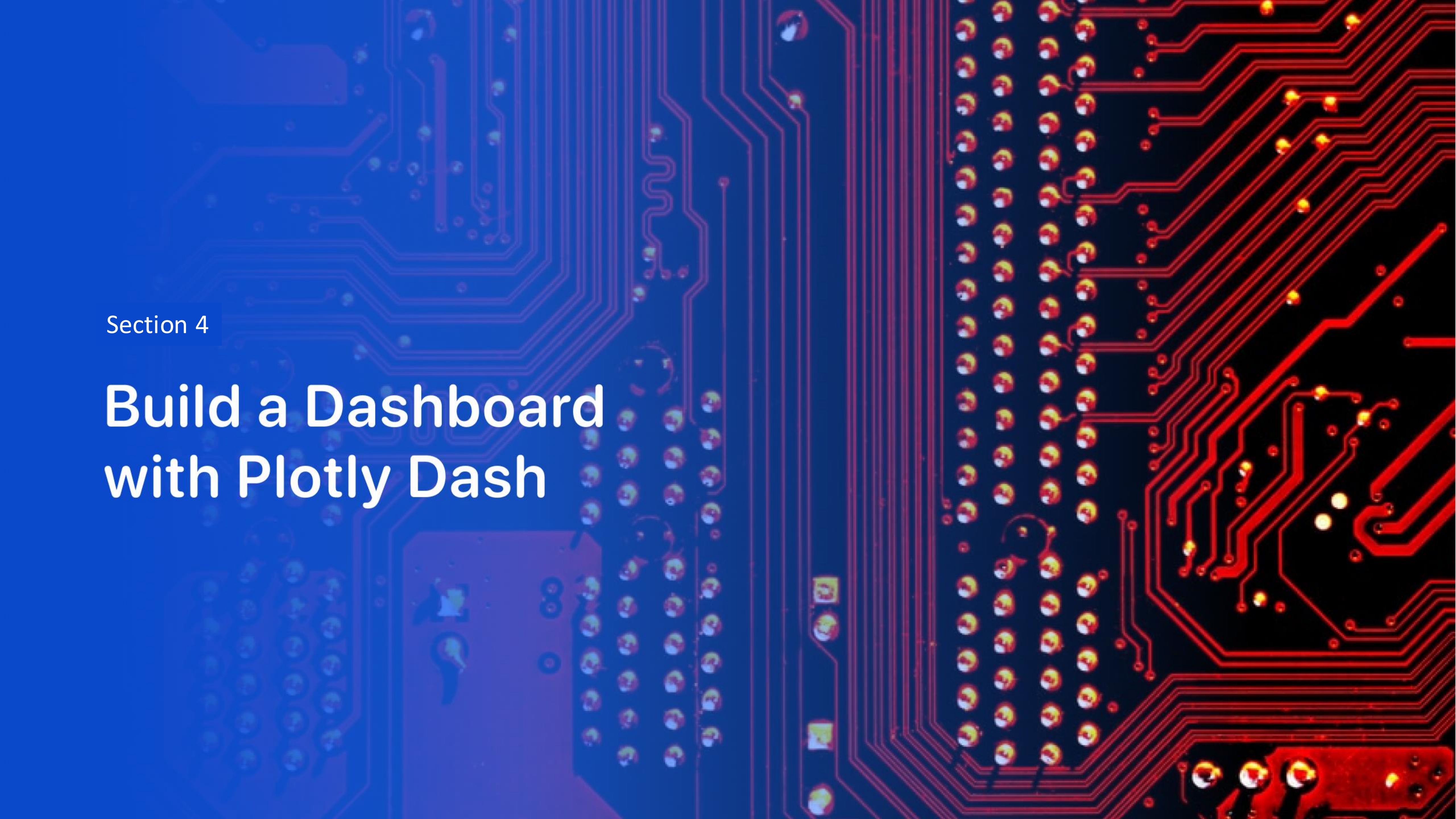
- This shows the launch outcomes of launches originating at the Vandenberg Launch Site.

Launch Site Proximities



https://cf-courses-data.static.labs.skills.network/jupyterlite/2.5.4/lab/index.html?mode=learn&env_type=jupyterlite&notebook_url=https%3A%2F%2Fcf-courses-data.static.labs.skills.network%2FIBM-DS0321EN-SkillsNetwork%2Flabs%2Fmodule_3%2

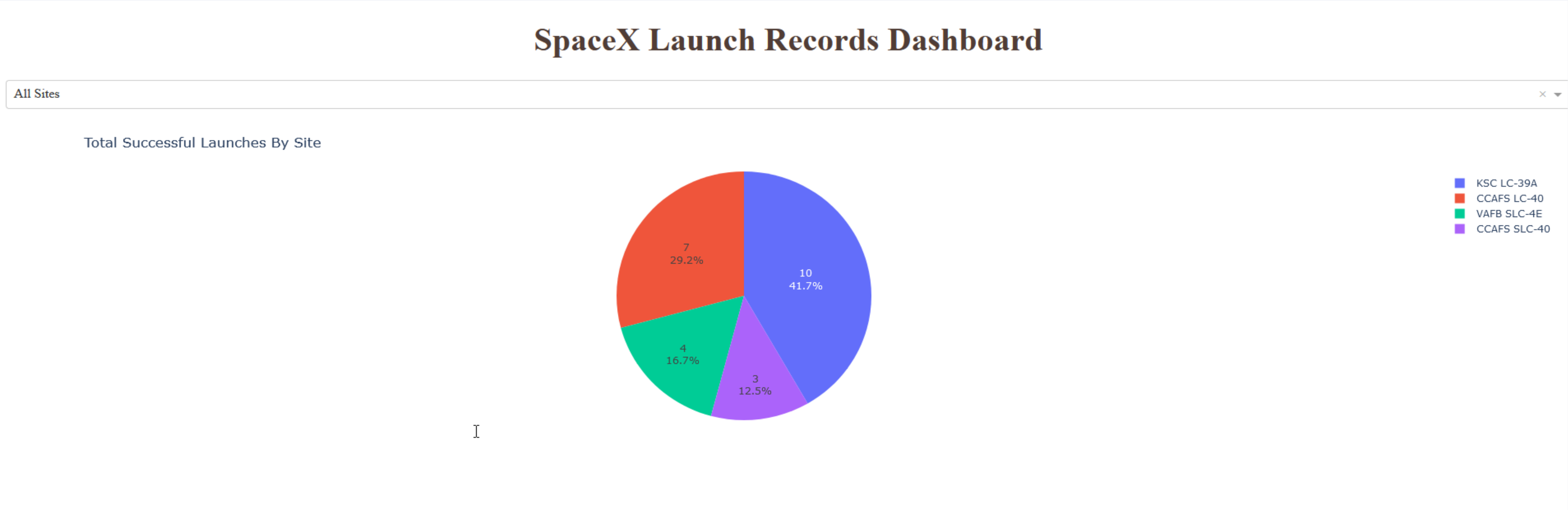
- This shows the proximity of the Vandenberg Launch Site to the nearest city, railroad, coastline, and highway.



Section 4

Build a Dashboard with Plotly Dash

Total Successful Launches Dashboard Screenshot



SpaceX Launch Records Dashboard

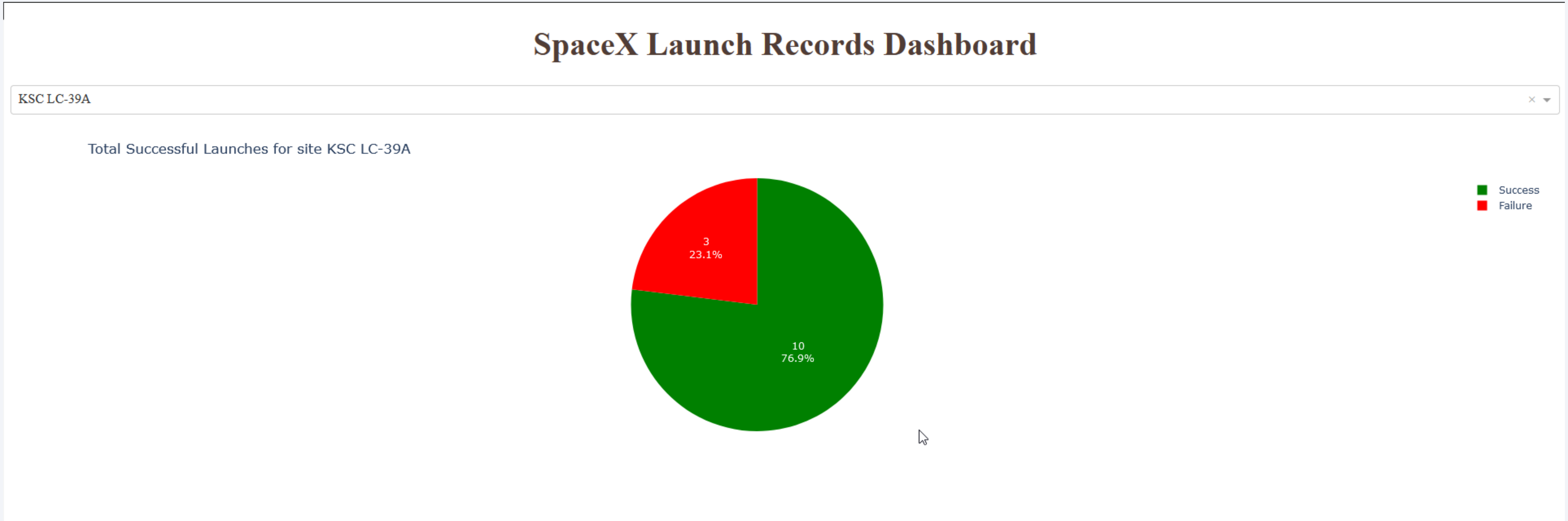
All Sites

- This dashboard screenshot shows that the selector indicates a selection for all launch sites, with the accompanying pie chart displaying the total successful launches per site.

- Note that the percentages here are calculated against the total number of successful launches combined, as opposed to the total number of launches inclusive of both successful and failed launches.

- It is evident from this screenshot that the KSC LC-39A launch site has had the greatest quantity of successful launches out of the four launch sites.

Total Successful Launches for KSC LC-39A

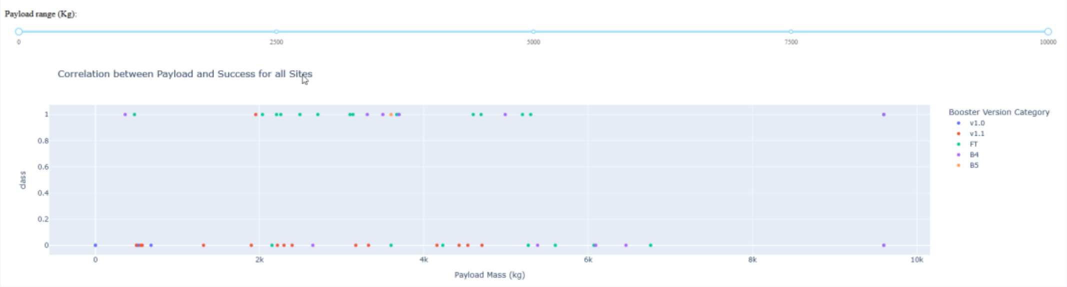

SpaceX Launch Records Dashboard

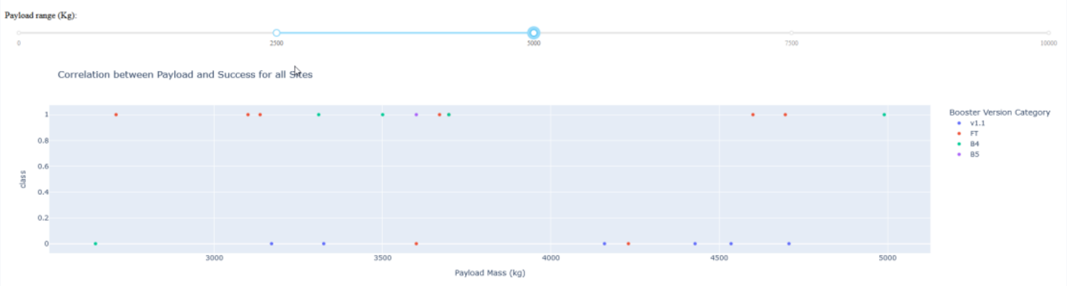
KSC LC-39A

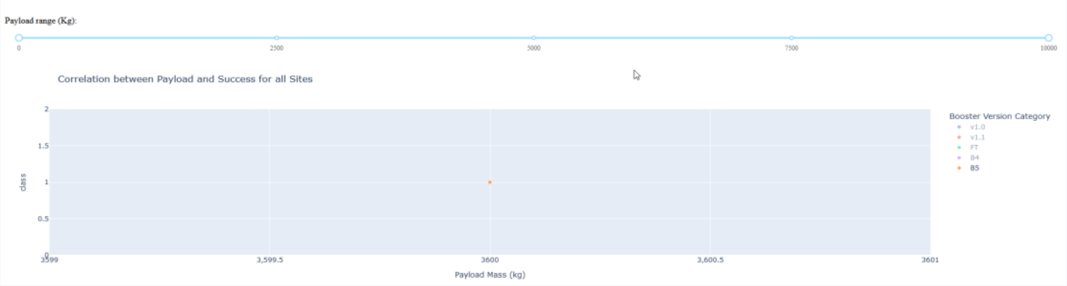
Total Successful Launches for site KSC LC-39A

- This screenshot displays the breakdown of launch success for the KSC LC-39A site.
- It can be observed that of the 13 launches that have occurred from this site, the vast majority of these have been successful, with a total of 10/13 ending up successful.

Varied Scatter Plots - Payload Mass vs Success – All Sites







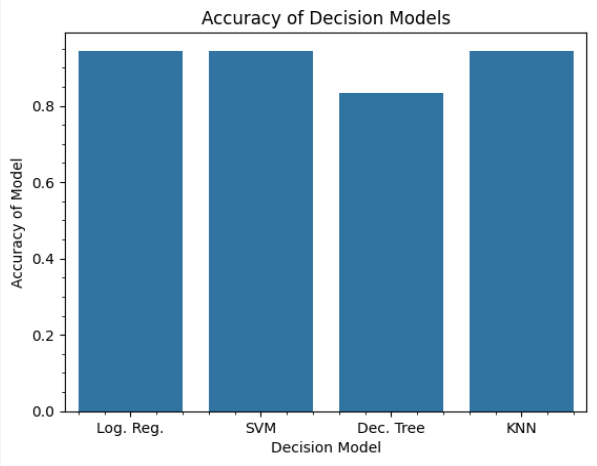
- The first screenshot displays the correlation between payload mass and success for all payload masses
- The second screenshot displays the correlation of payload vs success for the range of 2500 and 5000kg, showing that there is a high rate of success in this range of 60%
- Lastly, the third screenshot displays the correlation of payload mass to success for the B5 booster version, which has a 100% success rate for the single launch of this booster version.

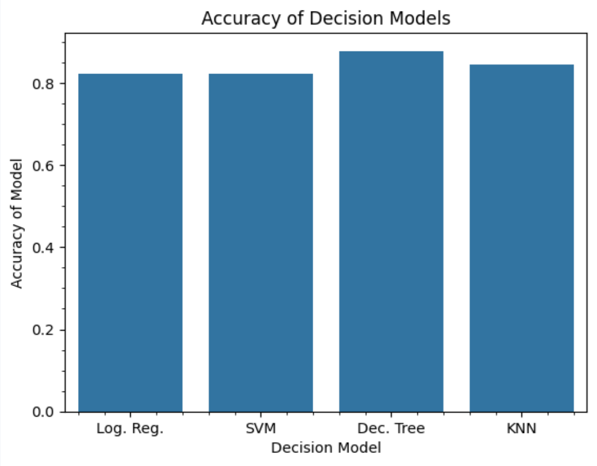
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- As can be observed in the first bar chart, interestingly all classification models are very similar in accuracy using test data.
- The only model showing a difference is the decision tree model.
- This is likely due to fit data failing to fit in the model as expected.
- Interestingly, when compared to the training data in the second bar chart, there is a far greater range of accuracy.

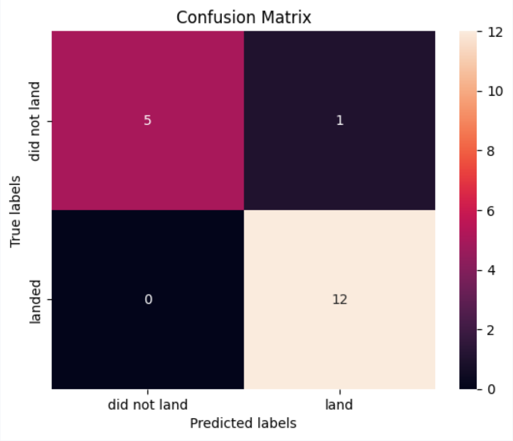




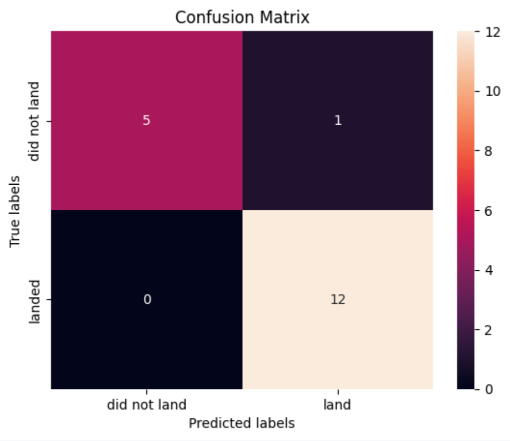
Confusion Matrix

As the test data shows, 3 out of the 4 models perform equally as well as the others. As such, the confusion matrix for theses will be similar as well, as can be observed in the three matrices displayed.

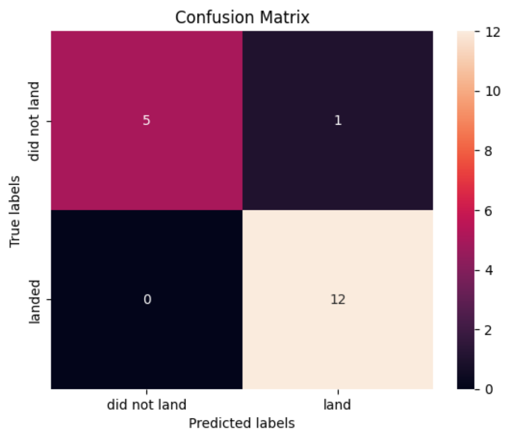
Logistics Regression Model



Support Vector Machine Model



K nearest neighbor Model



Conclusions

- Continued launches lead to greater success rates, both for ground and water-based landings.
- Provided that SpaceY is able to successfully leverage technology of a similar nature as SpaceX, it will be possible for SpaceY to become profitable for launches.
- Payloads with moderate payload mass, between the range of 2500 and 5000kg, are likely to be able to launched most successfully.
- SpaceY will likely experience the greatest success from a relatively remote location, close to the equator, a coast, but with both highway and railroad access.

Thank you!