IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Nikhil Mattu
9/14/24

Outline

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

Executive Summary

Methodologies:

- Data collection, web-scraping, and data wrangling
- Exploratory data visualization
- Exploratory data analysis using SQL
- Visualization with a Folium map
- Visualization with a Plotly Dashboard
- ML Prediction/Classification Testing

Results:

- Identified patterns in success rate based on orbit, payload mass, and launch site
- Found that the launch outcome success rate, on average, increased from 2013 to 2020
- Performed several SQL queries to receive information such as first successful launch, total payload, and average payload
- Found similarities between where launch sites were located and subjects that were (or were not) within their proximity
- Identified different counts of success and success rates for all four launch sites; also found how booster version and payload impacted success
- The best prediction/classification method for the SpaceX dataset used was decision tree classifier

Introduction

Overall goal: predict if the Falcon 9 first stage of SpaceX will land successfully

Context:
SpaceX advertises Falcon 9 rocket launches as having a far lower cost than some competitor companies, provided that it is able to reuse its first stage. Determining if the first stage will land will allow us to determine the launch cost.

Importance?:
One could leverage this information to their advantage if an alternate provider wanted to bid against SpaceX for a rocket launch.

Problems to solve:

1. Can we collect specific parts of data from the SpaceX dataset?
2. Can we notice patterns in the data between different features (e.g. how might payload mass) by graphing plots of these features?
3. Can we identify patterns based on location by developing a map of the launch sites with associated outcomes at each?
4. Can we find the best prediction model for classification so we can achieve our overall goal?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Retrieve Falcon9 data using the SpaceX API
  - Obtain data by web-scraping from a Wikipedia page for SpaceX Falcon9 launches
- Perform data wrangling
  - Using data collected with the SpaceX API, calculate number of launches on each site, calculate times that each orbit was used, and create landing outcome column containing values representing success or failure and create a launch outcome column
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Create test/train split of our data > with various models, fit them and use grid search to find the best parameters to use for each (with associated accuracies in finding them for training data)> identify accuracies of models on testing data

Data Collection

Methods:

1. Data collection using the SpaceX API

Request and parse SpaceX launch data with a GET request → Filter data frame with data to only include Falcon 9 launches → Deal with missing values

2. Data collection via web-scraping

Request the Falcon9 Launch Wikipedia page → Extract all column/variable names from the HTML table header → Create data frame by parsing the launch HTML tables

Data Collection – SpaceX API

Create functions that will return a desired portion of the data (based on function’s input of data) via requests.get()

1

Create a response (with requests.get()) and checked the content of launch data from the SpaceX API

2

Request and parse the SpaceX launch data; decode response content as a json (with .json()) and normalized into a pandas dataframe with .json_normalize()

3

Use the SpaceX API to get information about the launches for specific columns and set as initial ‘data’ dataframe variable

4

Retrieve data using step 1’s functions for desired columns, transferring the individual data to separate global variables / lists for each desired columns

5

Create a dictionary with these lists and turn dictionary into a new data frame

6

Filter new data frame to only have Falcon 9 launches

7

Deal with missing values (calculate the mean for Payload Mass with .mean() and replace nan values with .replace())

8

Export to CSV for data wrangling

https://github.com/MagmaLeo/Coursera-Applied-Data-Science-Capstone/blob/7e96c81d5de6be3c6372716c1e7db7685f087763/Module1/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

Request Falcon9 Launch HTML page with requests.get() as an HTTP response

Create a beautiful soup object from the response with BeautifulSoup() and html.parser type

Collect all relevant column names from the HTML table header by (find all tables on the page > identify the desired table > extract column names one by one from this table)

Create data frame by parsing the launch HTML tables (create dictionary with keys being the column names extracted > fill in dictionary values with data that we extract by row from the table for each column)

Convert dictionary to dataframe

https://github.com/MagmaLeo/Coursera-Applied-Data-Science-Capstone/blob/7e96c81d5de6be3c6372716c1e7db7685f087763/Module1/jupyter-labs-webscraping.ipynb

Data Wrangling

1. Load csv (as a dataframe) from data collection section (may not need to mention)
2. Calculate number of launches on each site (use value_counts() on LaunchSite column ot fin number of launches on each site)
3. Calculate number of occurrences of each orbit (^ same but for Orbit column)
4. Calculate number and occurrence of mission outcome of the orbits to determined number of landing_outcomes (^ same but Outcome column)
5. Create landing outcome label from Outcome column (label with 0 if bad outcome – based on a set of specific “failed” outcome values -, else 1) – i.e. turning categorical into numerical – if 0, failure, if 1, first stage landed successfully -> add these corresponding values to df as “Class” column
6. Export as csv for EDA with data visualization

https://github.com/MagmaLeo/Coursera-Applied-Data-Science-Capstone/blob/7e96c81d5de6be3c6372716c1e7db7685f087763/Module1/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

Load data wrangling csv as df, using this df as our ‘data’ for plotting

→ Scatter plot of flight number vs payload mass with class (failed or success outcome) as the hue

→ Scatter plot of flight number vs launch site with class as hue

→ Scatter plot of launch sites vs their payload mass with class as hue

→ Create bar chart for success rate of each orbit (grouping by orbit and giving mean ‘class’ value for each)

→ Scatter plot of flight number vs orbit with class as hue

→ Scatter plot of payload mass vs orbit with class as hue

→ Line chart of year vs average success rate (to get average launch success trend – achieved by grouping by data and giving mean ‘class’ value for each year)

→ Then select features that will be used in success prediction in the future module > create dummy variables for categorical columns/features (orbits, launchsite, landingpad, and serial) > cast all numeric columns/features to float64

https://github.com/MagmaLeo/Coursera-Applied-Data-Science-Capstone/blob/7e96c81d5de6be3c6372716c1e7db7685f087763/Module2/edadataviz.ipynb

EDA with SQL

<u>Queries performed:</u>

- Select unique (“distinct”) launch sites
- Display 5 records where launch sites began with string ‘CCA’
- Display total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List data when the first successful landing outcome in ground pad was achieved
- List names of the boosters that have success in drone ship and have payload mass >4000 but <6000
- List total number of successful and failure mission outcomes
- List names of the booster versions that carried the maximum payload mass using a subquery
- List records displayed month names, failure landing_outcomes in drone ship, booster versions, and 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

https://github.com/MagmaLeo/Coursera-Applied-Data-Science-Capstone/blob/7e96c81d5de6be3c6372716c1e7db7685f087763/Module2/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Building an Interactive Map with Folium

Map Objects:

- Map –overall map
- Circle – to add highlighted circle areas around NASA Johnson Space Center and launch sites
- Marker (with Icon) – to add a marker with an icon for each launch site at their coordinates; also added one for each launch outcome
- Marker cluster – to group launch outcomes of the same launch site
- MousePosition – to see where on the map (by coordinates) that your cursor is hovering over
- Polyline – to draw lines between a launch site and proximities such as a coastline and highway (to visualize distance between them)

https://github.com/MagmaLeo/Coursera-Applied-Data-Science-Capstone/blob/7e96c81d5de6be3c6372716c1e7db7685f087763/Module3/lab_jupyter_launch_site_location.ipynb

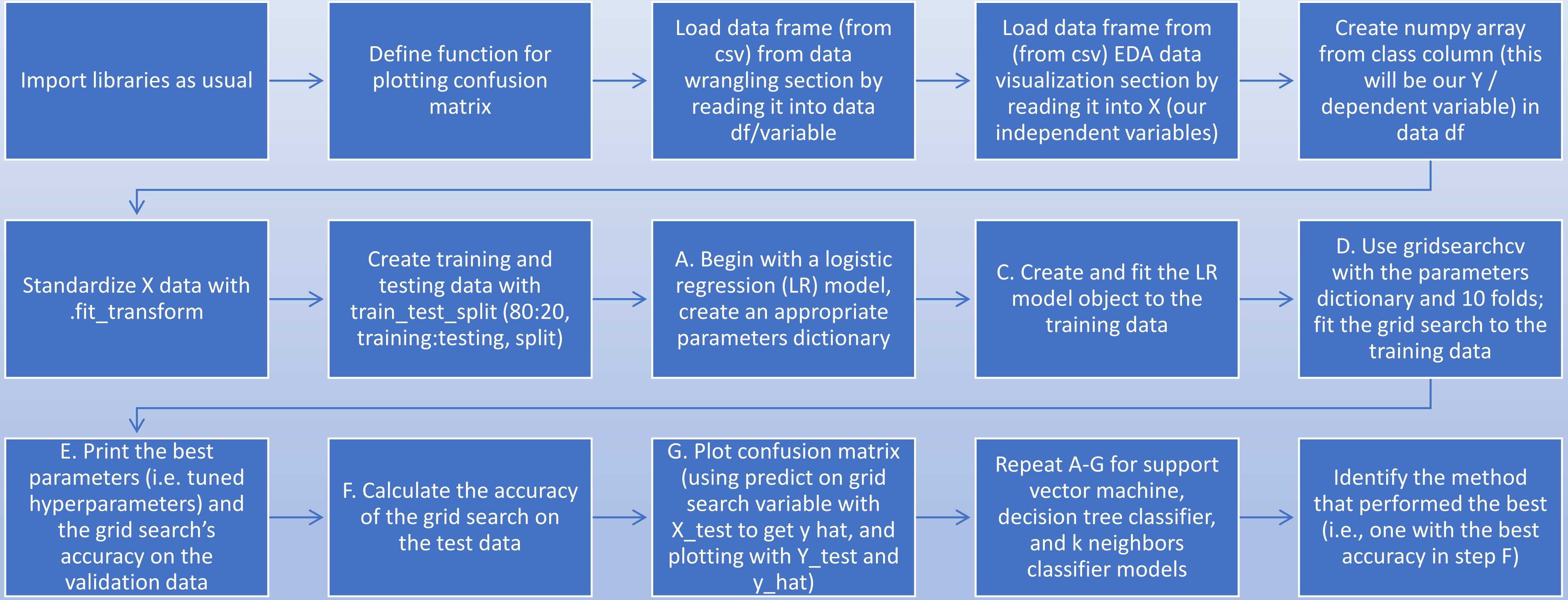
Building a Dashboard with Plotly Dash

Dashboard features:

- Launch site drop down menu
  - Allows user to select a given launch site (or ‘all sites’) to see the associated data
- Pie chart for success rate based on the selected launch site (including an ‘All sites’ option for proportions of counts of success for all launch sites)
  - For individual sites, this helps visualize rate of success, so we can see which sites had a higher or lower number of successful launch outcomes
  - For ‘All sites’, this helps as we can compare which sites had a greater or lesser total count of successes
- Scatter plot of payload mass (kg) vs launch outcome (class column) with color of each point indicating booster version + a range slider for setting minimum and maximum payload data
  - The scatter plot allows for visualization of the success rate across different payload masses for different booster versions (depending on the selected launch site from the dropdown menu); we can use this to see which booster versions had higher or lower launch success rates
  - The slider lets us focus in on specific ranges of payload mass for the scatter chart, allowing us to more easily see patterns of success (or failure) for or being unaffected by carrying certain payloads (i.e., we can use this to see which payload ranges had higher or lower success rates)

https://github.com/MagmaLeo/Coursera-Applied-Data-Science-Capstone/blob/7e96c81d5de6be3c6372716c1e7db7685f087763/Module3/Dashboard%20with%20Plotly%20Dash%20Lab.py

Predictive Analysis (Classification)



https://github.com/MagmaLeo/Coursera-Applied-Data-Science-Capstone/blob/7e96c81d5de6be3c6372716c1e7db7685f087763/Module4/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

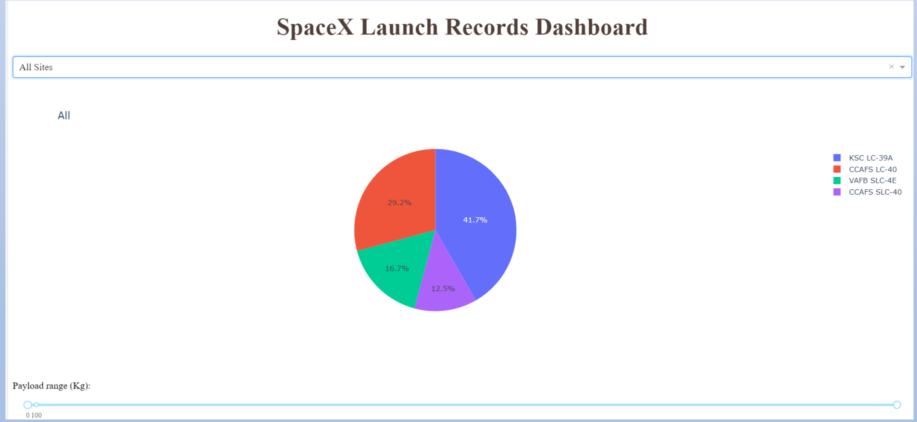
Exploratory Data Analysis:

- Greater success rate for later flight numbers
- No rockets from VAFB SLC 4E were launched with a payload > 10000KG
- Most rockets were launched with a payload < 8000kg
- ES-L1, GEO, HEO, and SSO orbits had the highest success rates (~100%)
- Success increased with the number of flights for the LEO orbit, but for others, such as GTO, there was no apparent relationship
- LEO, ISS, and PO orbits appeared had greater success with carrying heavier payload masses
- Success rate on average increased from 2013 to 2020
- Average payload mass ~3000kg
- First successful landing on Dec. 22, 2015

Predictive Analysis Results:

- Decision tree classifier model had the best prediction accuracy on test data (developed from a Space X dataset) compared to logistic regression, support vector machine, and K-neighbors models

Interactive Analytics:



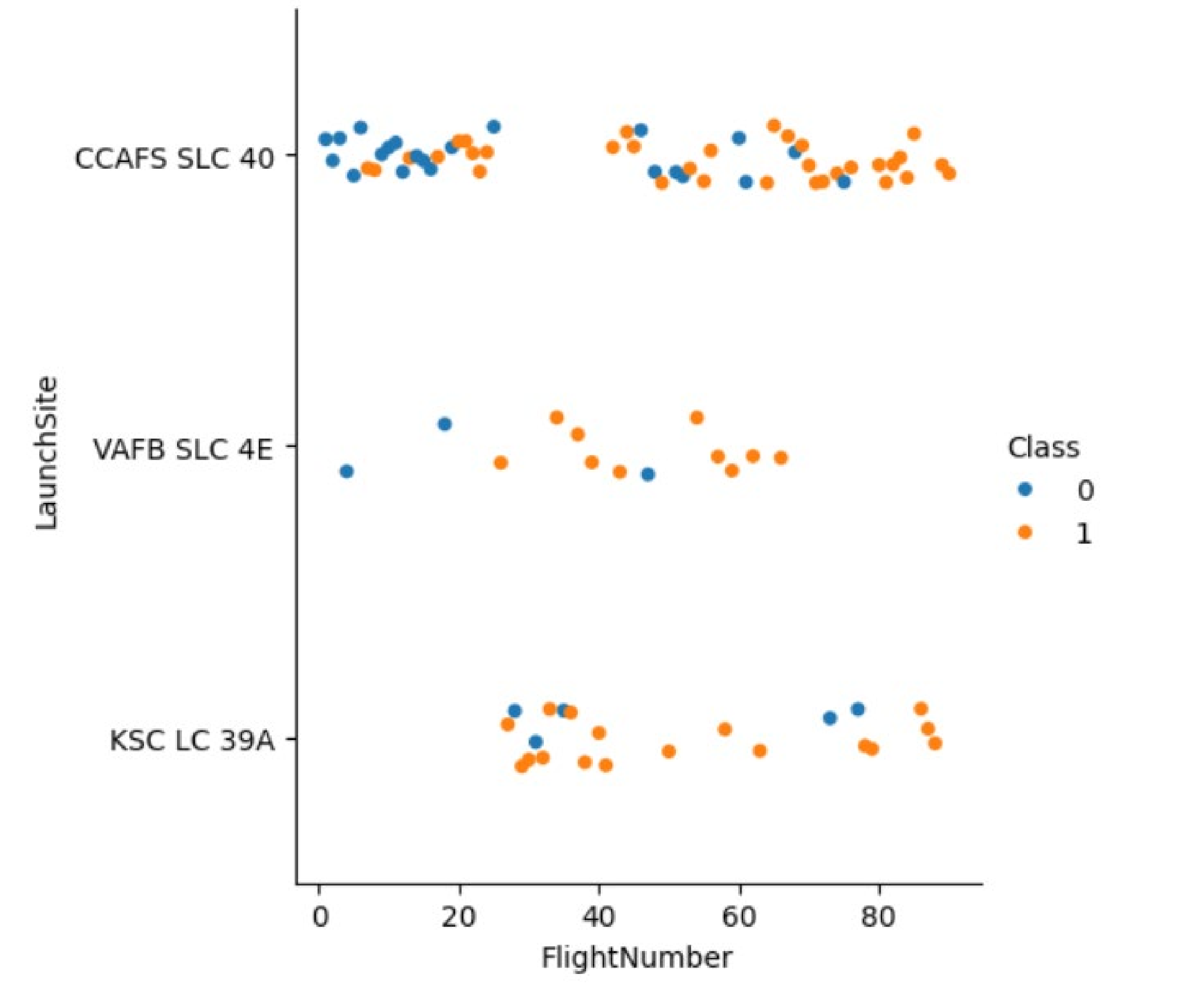
Examples: Folium map depicting distances between a launch site and proximities (coastline and a highway) (upper image); Plotly dashboard depicting proportions for number of successes of all launch sites

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

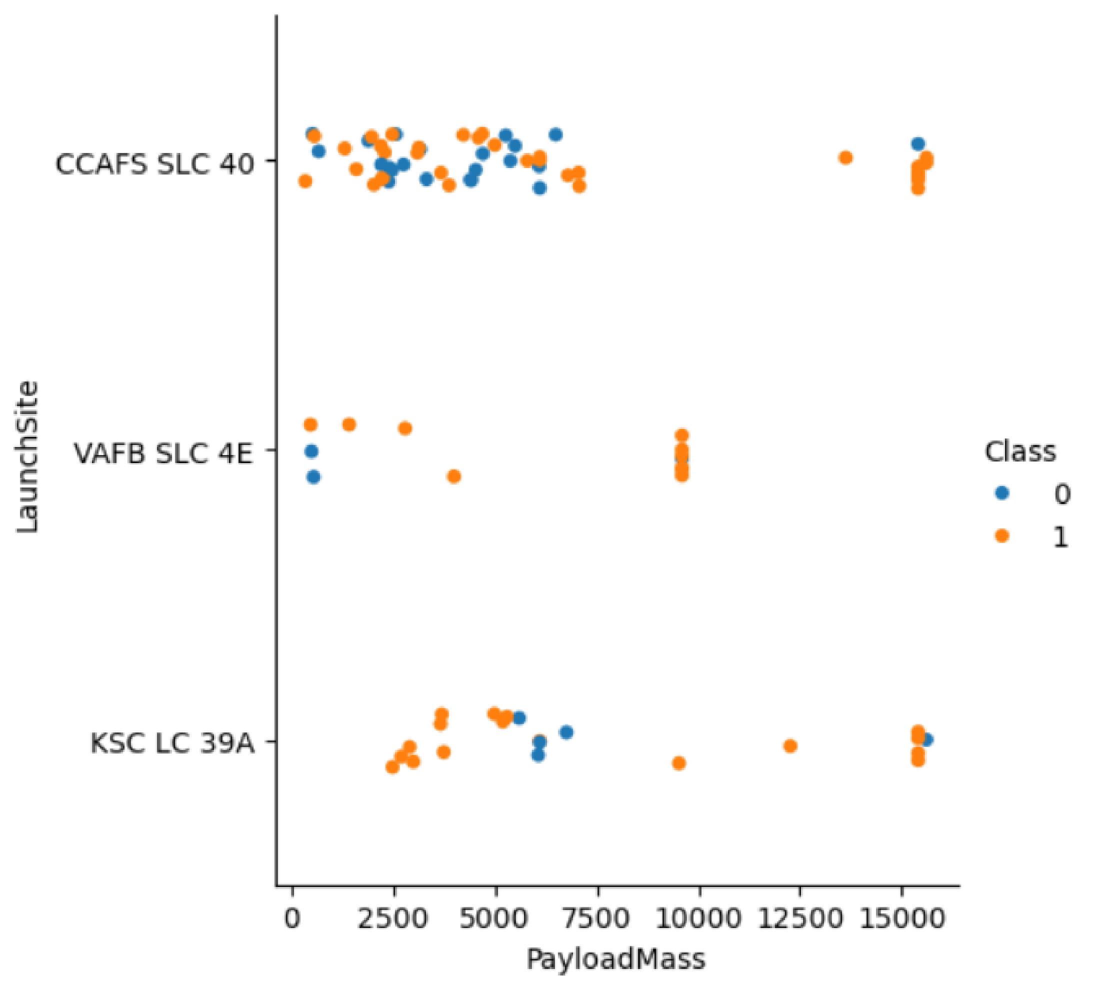
- Observed an overall greater rate of success for later flight numbers



Payload vs. Launch Site

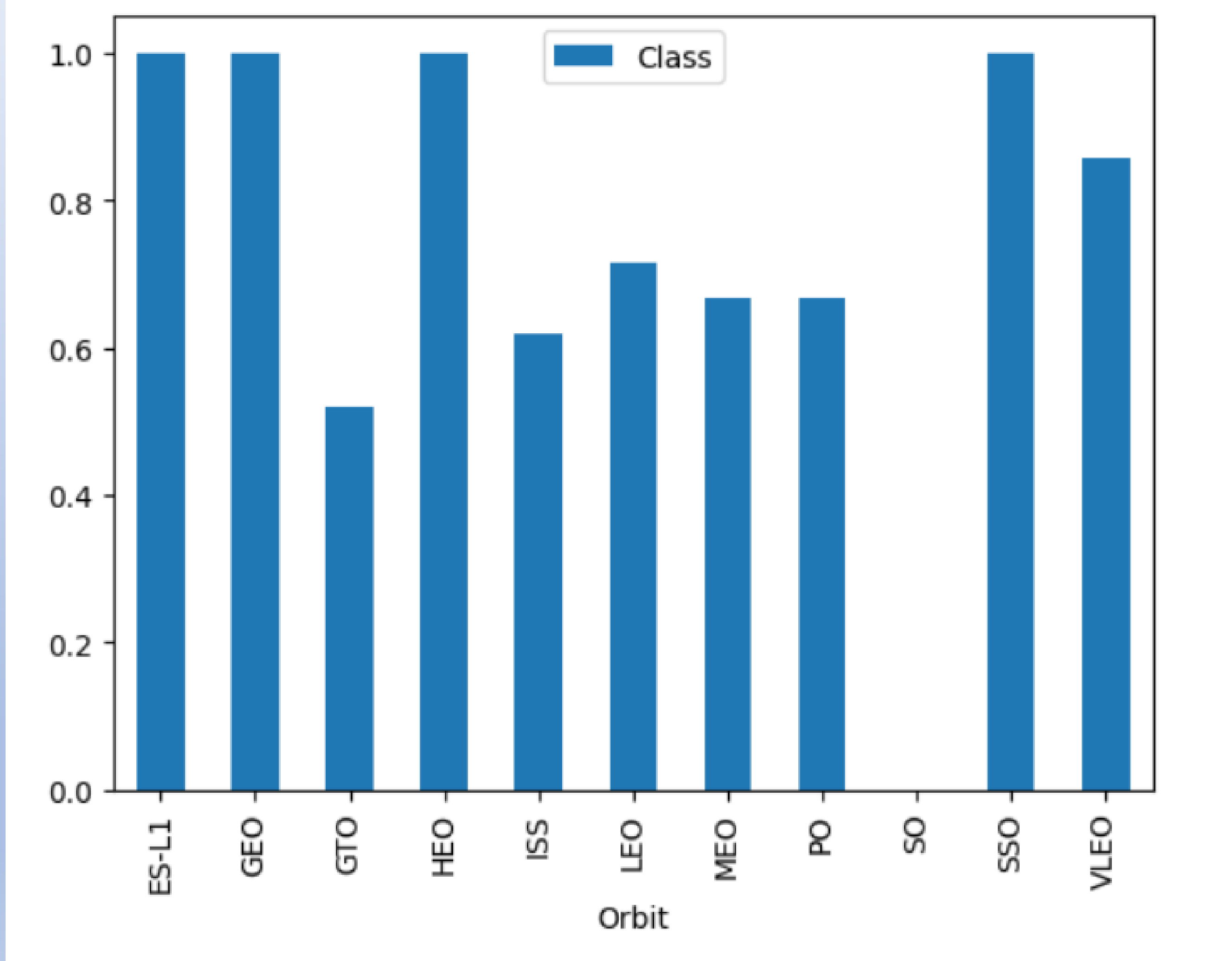
- No rockets were launched with payload > 10000KG for VAFB SLC 4E
- A majority of rockets were launched with a payload under 8000kg

<seaborn.axisgrid.FacetGrid at 0x64e5f60>



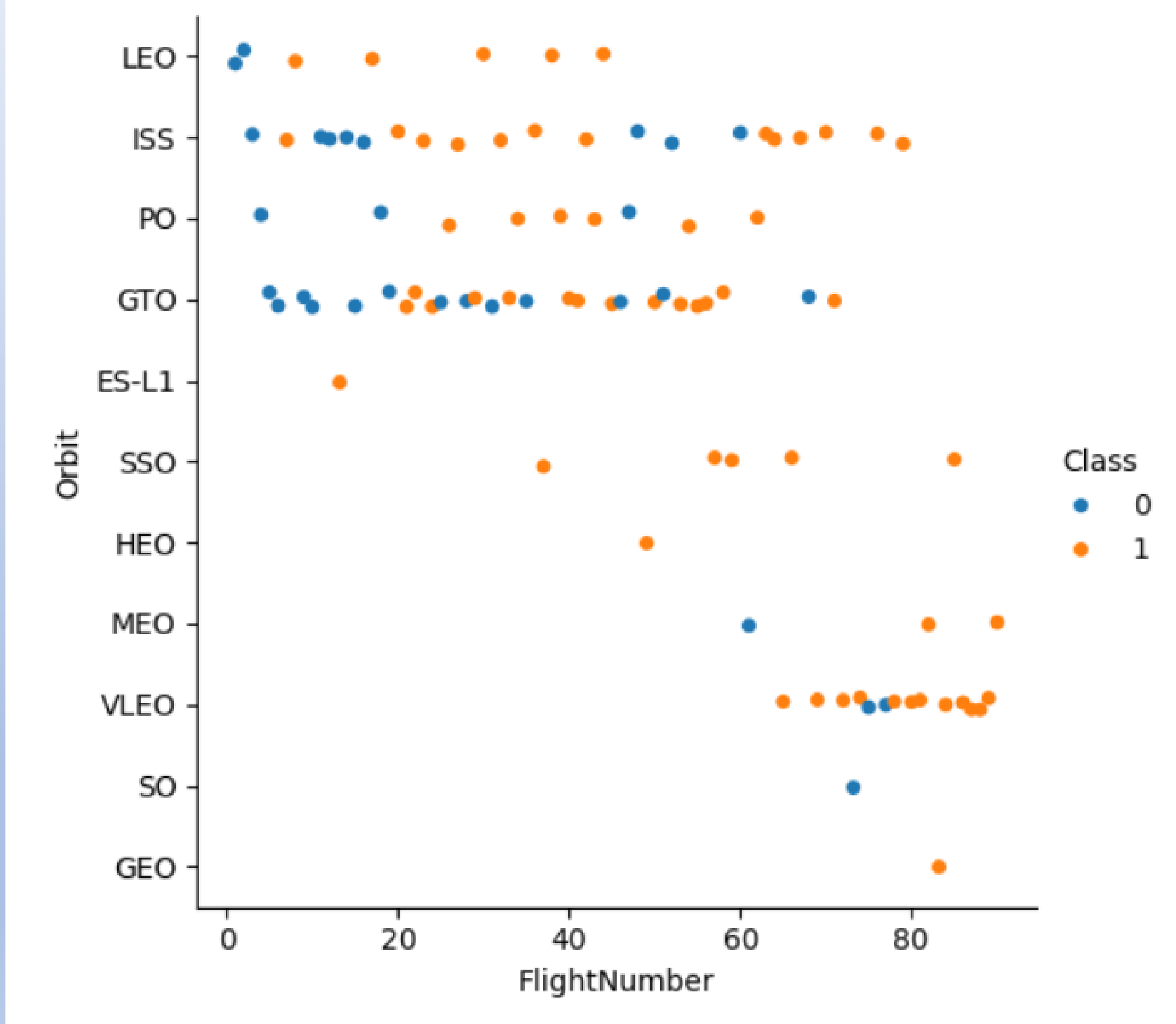
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO had the highest success rates (~100%)



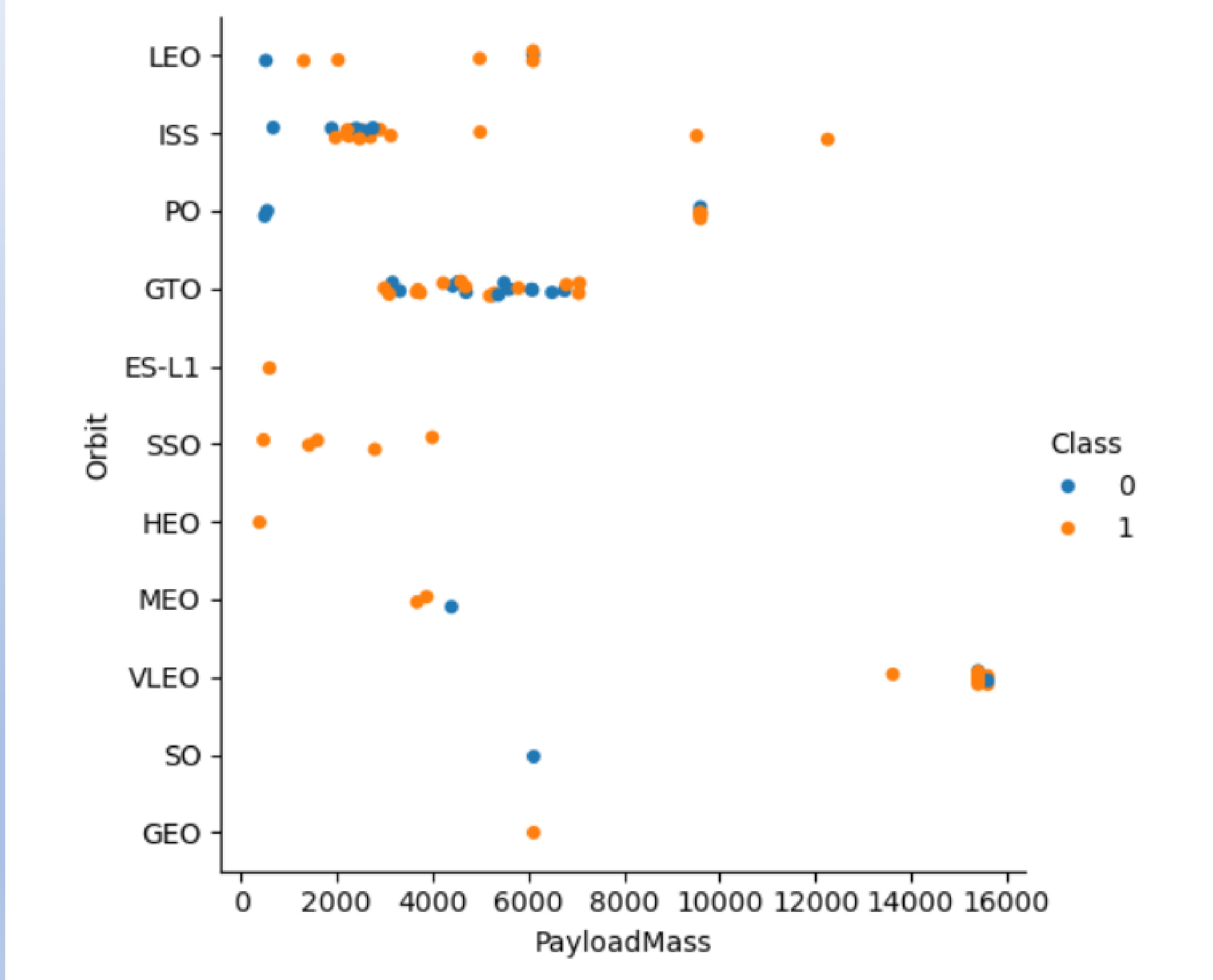
Flight Number vs. Orbit Type

- For LEO, success increased with number of flights, whereas others such as GTO appeared to have no relationship



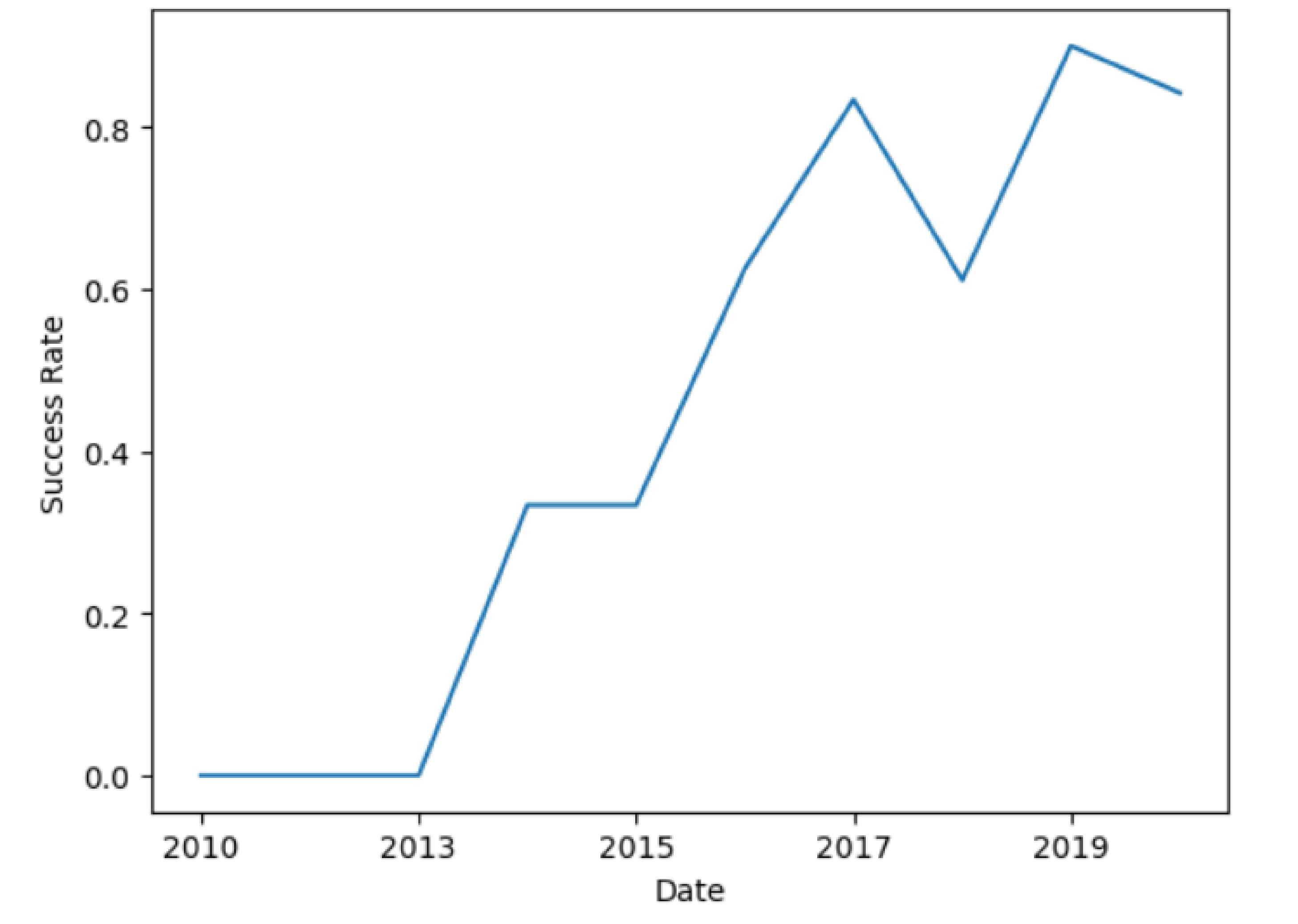
Payload vs. Orbit Type

- LEO, ISS, and PO appeared to have more frequent success with heavier payload masses



Launch Success Yearly Trend

- Success rate on average increased from 2013 to 2020



All Launch Site Names

- For our Space X table data, we are able to see the four possible launch sites used by rockets

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

Launch Site Names Begin with 'CCA'

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

- Above is a listing of five example data for launch sites beginning with CCA

(i.e., entries with launch site CCAFS LC-40 or CCAFS LC-40)

Total Payload Mass

- With this query, we can observe the total payload carried by boosters from NASA

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Below is the result of a query used to find the average payload mass carried by booster version F9 v1.1

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- The following the result of an SQL query used to obtain the data of the first successful ground landing (i.e., first successful landing outcome on the ‘ground pad’)

Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- With an SQL query, we are also able to obtain a list of boosters which have successfully landed on the drone ship while holding a payload mass greater than 4000 kg but less than 6000 kg

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

Total Number of Successful and Failure Mission Outcomes

- Another SQL query allows us to see the the total number of successful and failure mission outcomes

Success MOs	Failure MOs
100	1

Boosters Carried Maximum Payload

- Here, we have a list of the names of the boosters that had carried the maximum payload mass

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

- In terms of landing outcomes in 2015, we are able to obtain a list of failed ones for drone ship with associated booster and launch site information

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

- Finally, we were able to use an SQL query to rank the number per landing outcome type in descending order for dates between 2010-06-04 and 2017-03-20

Landing_Outcome	Occurrence
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

Section 3

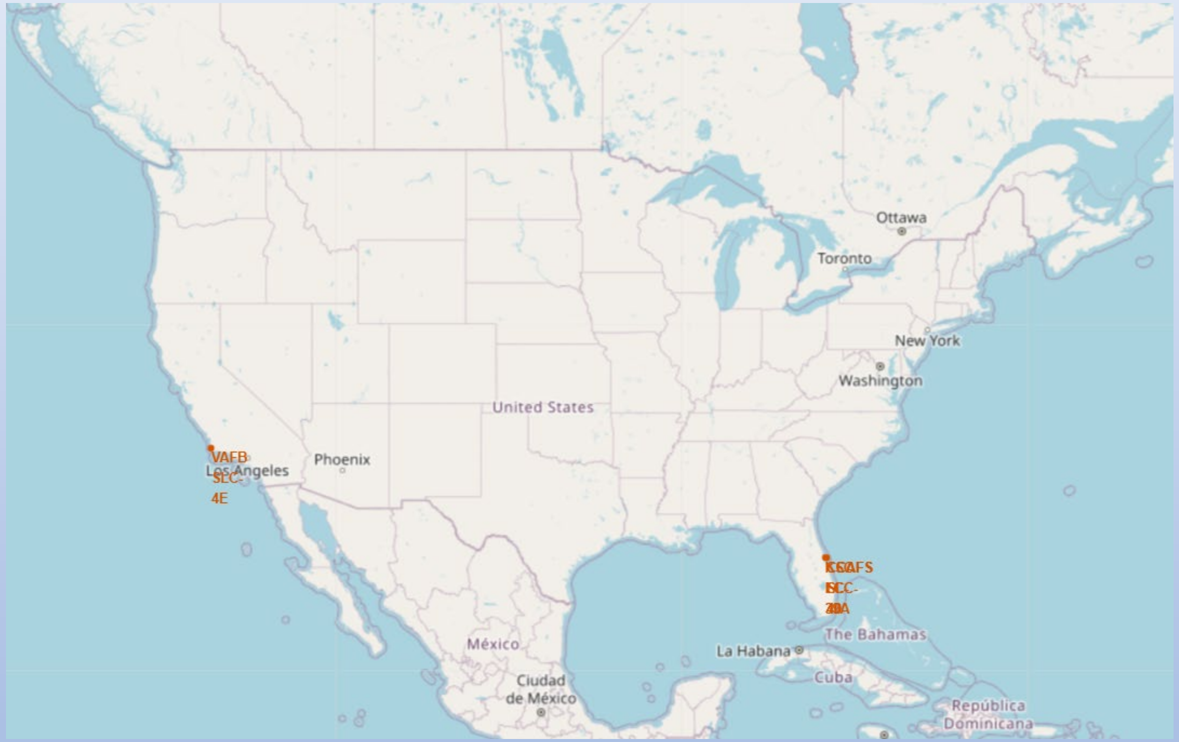
Launch Sites Proximities Analysis

Launch Sites Locations

- Displayed, we have our four launch sites, VAFB SLC-4E, KSC LC-39A, CCAFS LC-40, and CCAFS SLC-40 (latter three in similar region of on the right side of map)
- With our Folium map, we see that all launch sites are:

  1. further south in the US (closer to the equator)

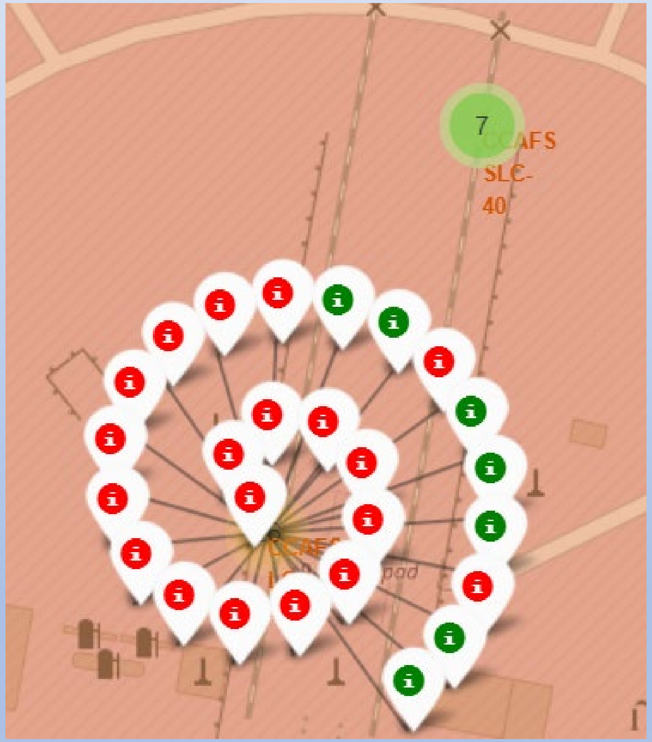
  2. close to the coastline



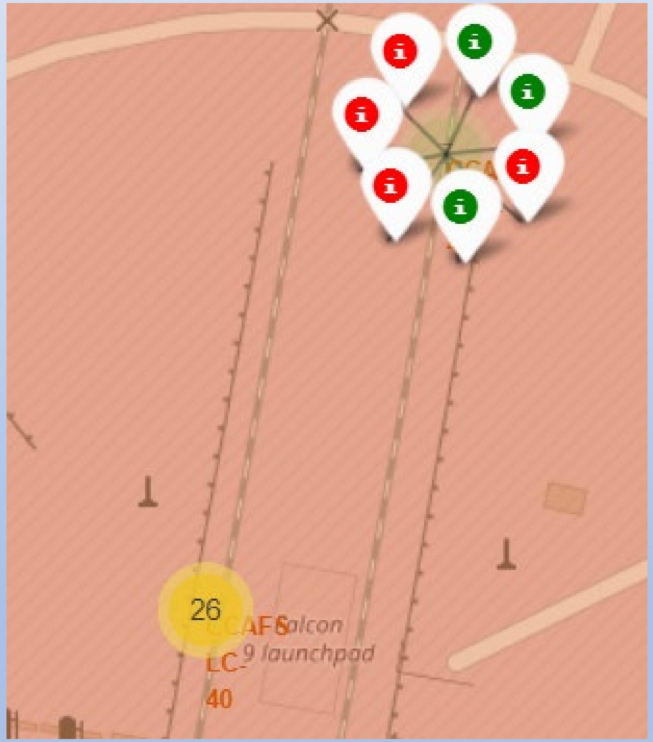
Launch Outcomes Per Site

- On our map, we have numbered clusters holding launch outcomes per site; here we are zoomed into the region of CCAFS LC-40, and CCAFS SLC-40
- We are able to see outcomes for each of the two clusters, where green markers indicate successful outcomes, while red indicate failures; such visualization allows us to identify sites with greater or lesser success rates
- Based on the depicted outcomes per cluster here, we are able to determine that the CCAFS SLC-40 site had a higher success rate than CCAFS LC-40

CCAFS LC-40 Outcomes (26)



CCAFS SLC-40 Outcomes (7)



Commonalities of Launch Site Proximities

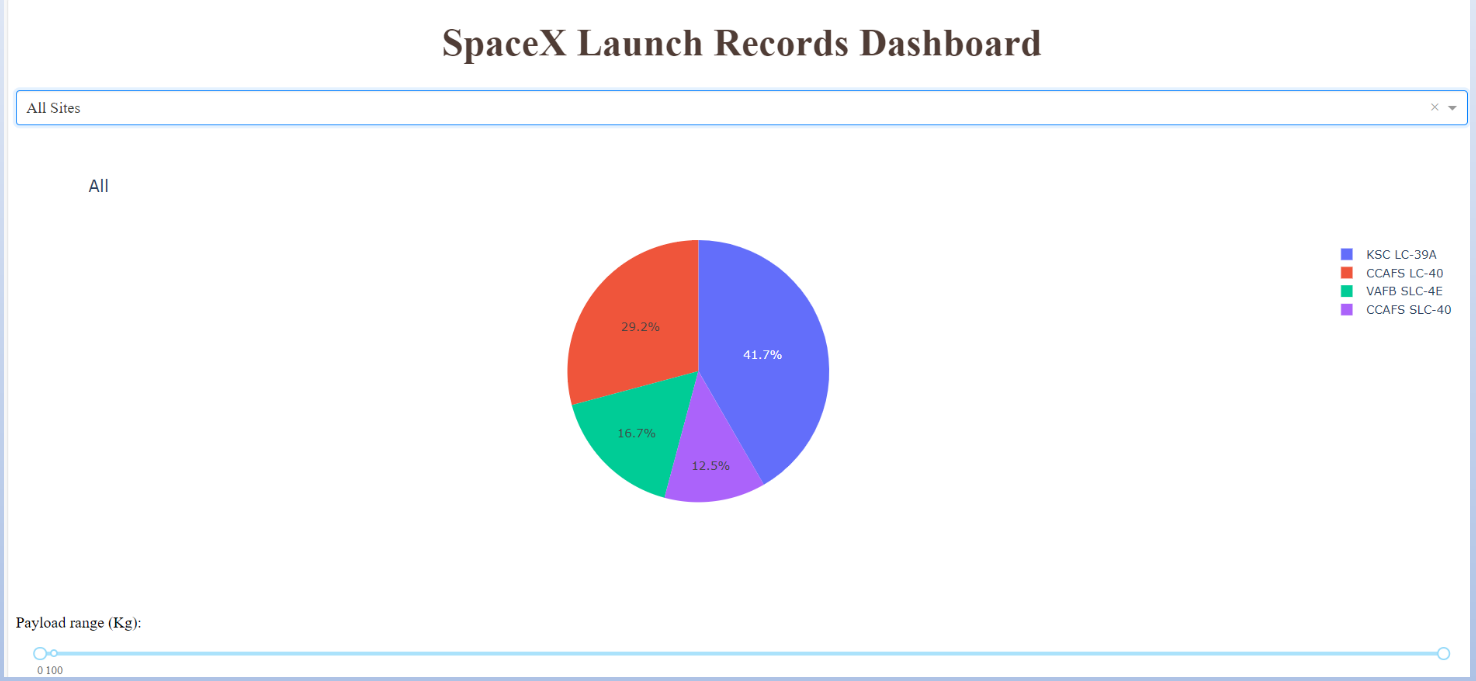
- In exploring our Folium map, we are able to identify commonalities between launch sites and their proximities (or lack thereof):
  - Sites tend to be near coastlines, highways, and railways
  - Sites tend to be further away from cities
- An example is shown here for the close proximity of the coastline and a highway to launch site CCAFS SLC-40

Section 4

Build a Dashboard with Plotly Dash

Successful Outcomes of All Launch Sites

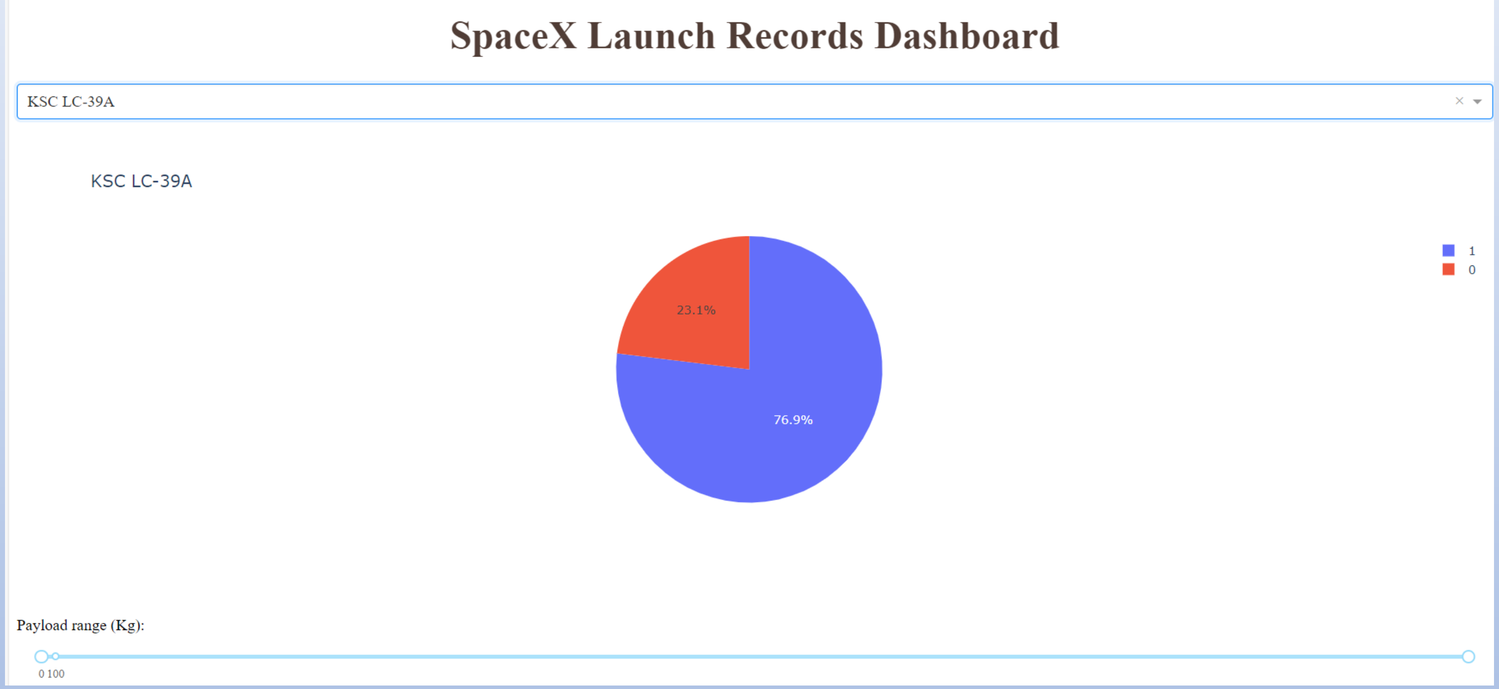
- Depicted are the proportions of total successful outcomes success each launch site had in relation to one another
- From this, we can observe that CCAFS SLC-40 had the fewest successes, while KSC LC-39A had the most
- This is not, however, indicative of the success rates, which we will show next…



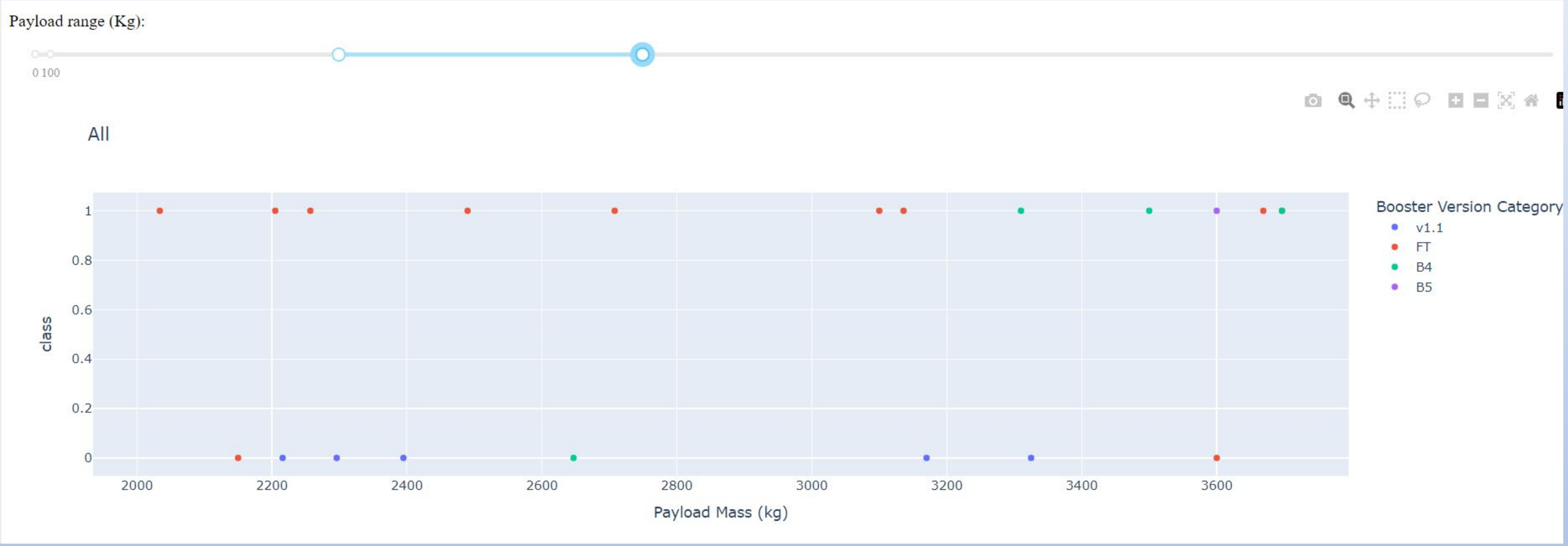
Launch Site with Highest Success Rate

- The launch site KSC LC-39A not only presented with the highest count of successful outcomes, but also had the best success rate out of the four sites (KSC LC-39A > CCAFS LC-40 > VAFB SLC-4E > CCAFS SLC-40)

- To be specific, we are able to see that more than ¾ of the launch outcomes were successful for this site



Payload Mass (Kg) vs Launch Outcome (Class) for All Sites



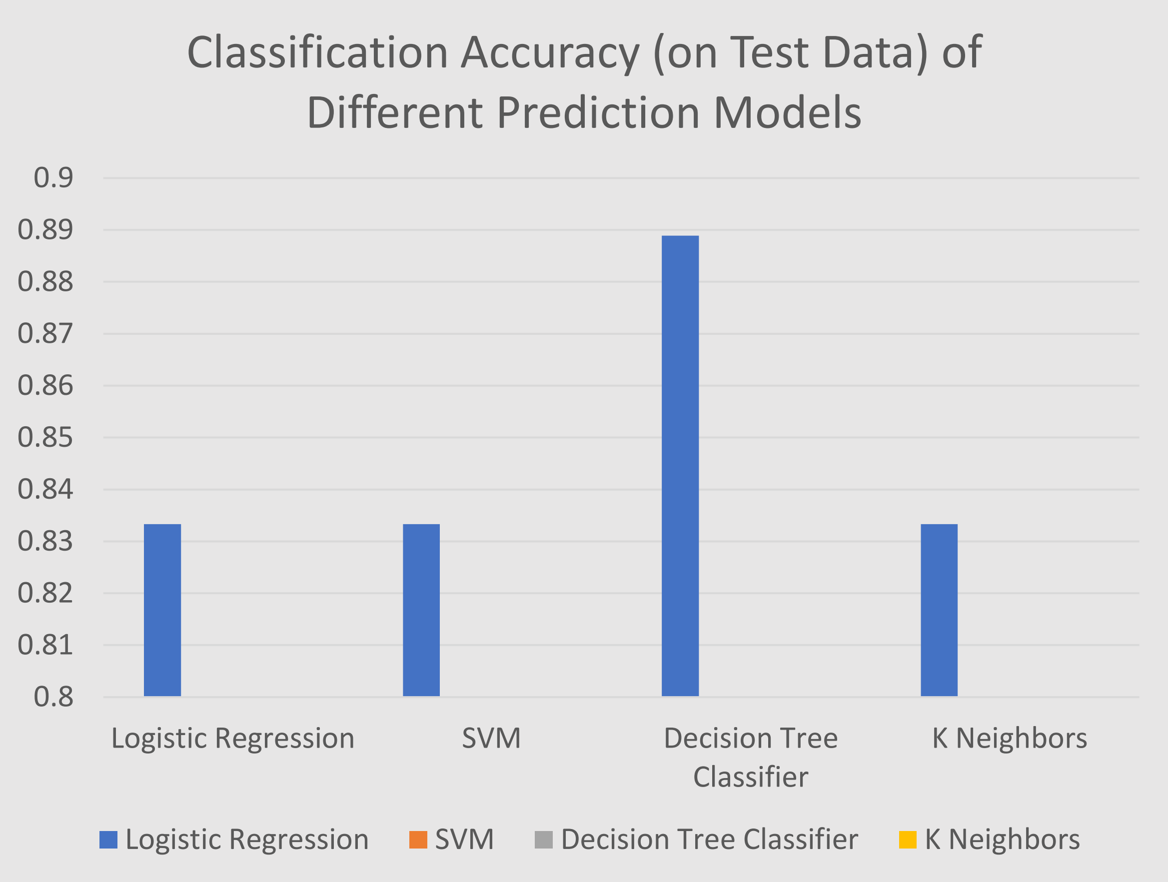
- With our dashboard, we can select a smaller payload range and view the outcomes (i.e., class; 0 for failure, 1 for success) for given booster versions (from any launch site) for said range
- Here we have a payload mass range of roughly 2000 to 3800 Kg
- By limiting the plot to this range, we see that for payloads between 2000 and 3800 Kg , FT and B4 had a greater rate of success, while v1.1 appeared with the opposite; additionally, we see that there were not many launches (only 1) with a payload mass in this range that had used the B5 booster
- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.

Section 5

Predictive Analysis (Classification)

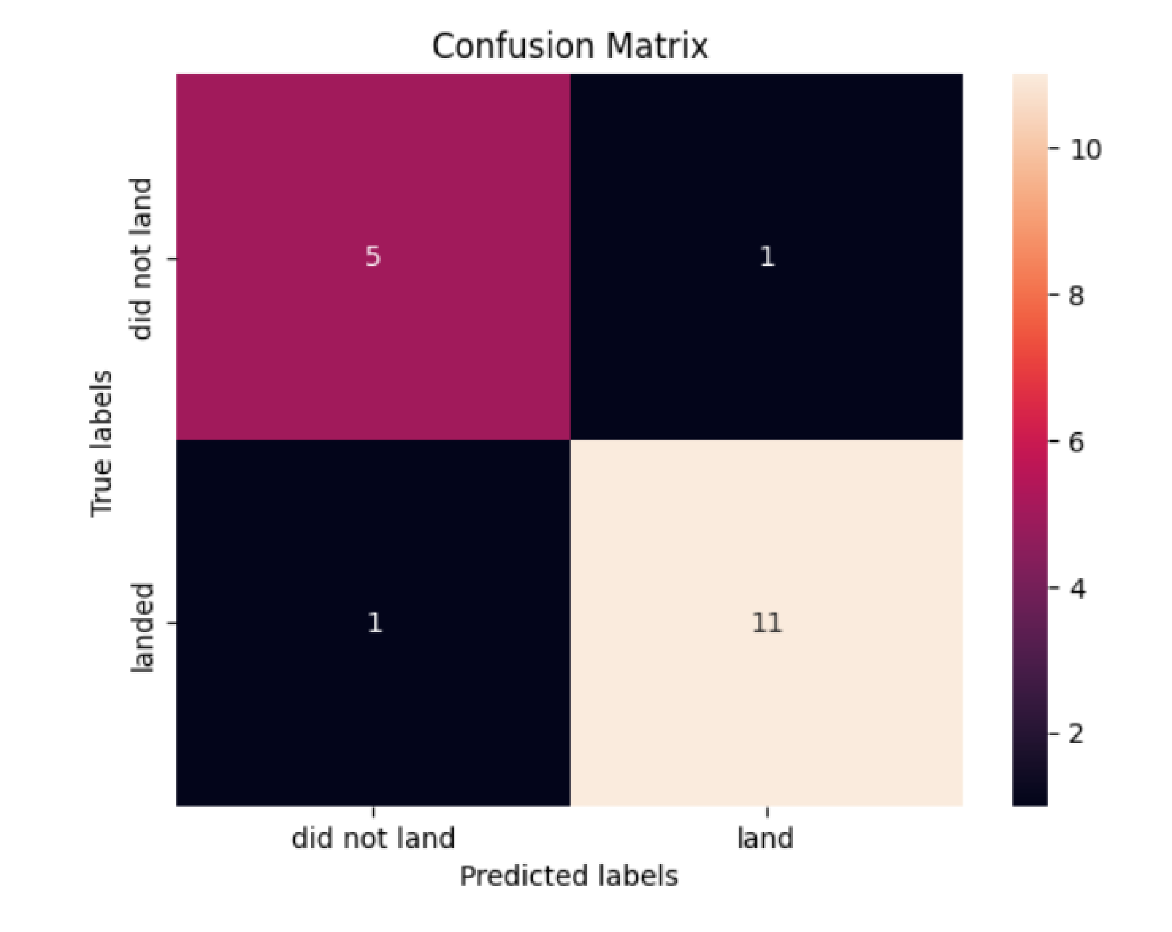
Classification Accuracy

- Using logistics regression, support vector machine, decision tree classifier, and K-neighbors classification models, we could predict Space X launch outcomes based on past data
- The accuracy of each model using test data is depicted in the graph, where decision tree classifier had the highest accuracy



Confusion Matrix (Decision Tree Classifier)

- The decision tree classifier, the best performing model, offers a confusion matrix (based on input data) as displayed
- From this matrix, we see a very low number (1) for each of our false positives (upper right) and false negatives (lower left), whereas we have much higher counts for our true positives (lower right; 11) and true negatives (upper left; 5) – this indicates that most of the model's predictions of known data matched with the known data itself, hence, it has a decent classification accuracy



Conclusions

Exploratory Data Analysis:

- Identified patterns between features such as success rate, flight number, orbit used, type of landing, payload mass, and launch site
- Observed that the launch success rate on average increased from 2013 to 2020
- Queried various other information such as when the first successful landing was, the total and average payload masses

Predictive Analysis Results:

- Decision tree classifier model had the best prediction accuracy on test data (developed from a Space X dataset) compared to logistic regression, support vector machine, and K-neighbors models

Interactive Analytics:

- Identified that launch sites tended to be close to coastlines, highways, and railways; launch sites also tended to be distant from cities
- Was able to identify trends in successes based on certain payloads and booster versions
- Saw what launch sites had higher or lower counts of successful outcomes, and what sites presented with higher or lower success rates overall

Appendix

Github link for all work:

https://github.com/MagmaLeo/Coursera-Applied-Data-Science-Capstone.git

Thank you!