IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Dongyue Li
09/14/2025

Outline

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

Executive Summary

- Summary of methodologies
  - Collecting, processing, cleaning, and visualizing Falcon 9 rockets and launch site data
  - Train predictive machine learning models to connect the rocket and launch site data with past launch outcomes
  - Predict the launching outcome based on the trained model and the input rocket and launch site data
- Summary of all results
  - All four trained classifiers can predict the launching outcome with >80% accuracy
  - All four classifiers can identify which parameter has the most predictive power for the predictions
  - Along the process, dashboards and maps were created to better understand the data

Introduction

- Project background and context
  - SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars, while other providers' cost upward of 165 million dollars each.
  - Much of SpaceX's savings is because SpaceX can reuse the first stage of rockets. 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.
- Problems you want to find answers
  - Can we predict if the Falcon 9 rocket's first stage will land successfully or not based on data of the rocket, orbit, launch site, and previous launch outcome using machine learning methods?

Section 1

Methodology

Methodology

Executive Summary

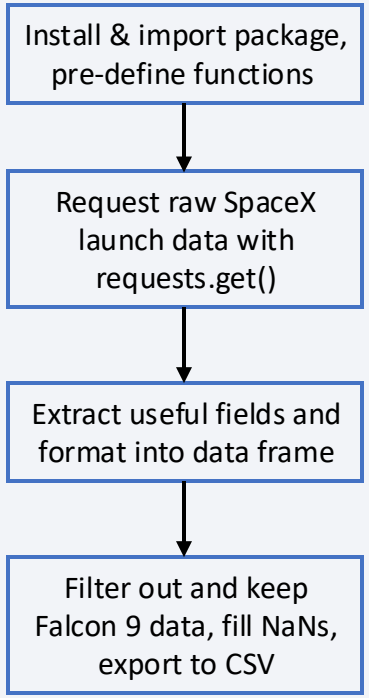
- Data collection methodology:
  - Collect data of past launches through the SpaceX API requests and web scraping of Wikipedia webpage
- Perform data wrangling
  - Data selection: keep only the data fields useful for launch outcome prediction
  - Data cleaning: fill NaNs and remove samples irrelevant or contain erroneous values
  - Data formatting and encoding: convert data to format that facilitate model prediction
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Using SQL and Python to collect, visualize, and map data from different sources
- Perform interactive visual analytics using Folium and Plotly Dash
  - Building dashboard to interactively visualize and explore the relationship among data fields
- Perform predictive analysis using classification models
  - Build, tune, and evaluate 4 classification models to predict launching outcome

Data Collection

- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts

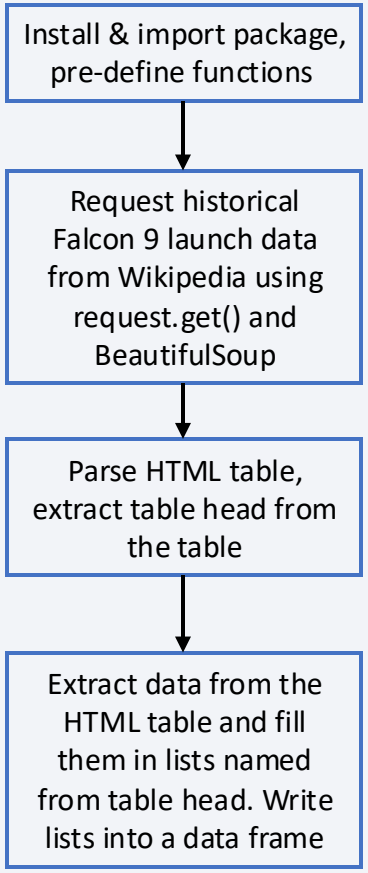
Data Collection – SpaceX API

- Collecting detailed rocket and launch data with SpaceX API
  - Using requests.get(url), status should be 200
  - Many data fields downloaded, kept only useful ones, including 'rocket', 'payloads', 'launchpad', 'cores', 'flight_number', 'date_utc', etc.
  - Apply pre-defined functions to format the raw data from the API to more readable format in a Pandas dataframe
  - Filter the data frame and only keep Falcon 9 launches
  - Data wrangling of the dataframe to fill NaN data
- https://github.com/dongyuegeo/ibm_data_science/blob/main/Capstone/Code/1-jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Falcon 9 historical launch records from a Wikipedia page scraping
  - Using BeautifulSoup to scrape a url response
  - Extract the data fields from the 3rd table on that page
  - Create a dataframe and parse the extracted table to fill the dataframe
- https://github.com/dongyuegeo/ibm_data_science/blob/main/Capstone/Code/2-jupyter-labs-webscraping.ipynb



Data Wrangling

- Data processing
  - To create a clean dataset with only relevant fields to facilitate insight uncovering
  - Data field selection:
    - Focus only on Falcon 9 launches as of June 2021
    - Only keep the data about rockets and launches, which are useful for launch outcome prediction
  - Data wrangling and cleaning:
    - Explore data fields in dataframe to get familiar with the data, including data type, size, and missing values
    - Properly fill Null values
  - Data formatting and encoding:
    - Use pre-defined functions to extract to extract data from the messy raw table directly scraped from Internet
    - Process the data into format that facilitate the outcome prediction, e.g. encode the launch outcome to 0 and 1
- https://github.com/dongyuegeo/ibm_data_science/blob/main/Capstone/Code/3-labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
  - A series of plots were made to (1) understand the distribution and value of the data, and (2) explore the correlation among two variables. These exploratory analysis help familiarize with the data and facilitate data feature engineering.
  - Plots in this lab include:
    - Scatterplot of FlightNumber vs. PayloadMass
    - Scatterplot of FlightNumber vs LaunchSite
    - Scatterplot of PayloadMassa vs LaunchSite
    - Barplot of Success Rate by Orbit Type
    - Scatterplot of FlightNumber vs Orbit type
    - Scatterplot of PayloadMass vs Orbit type
    - Yearly trend of success rate
- https://github.com/dongyuegeo/ibm_data_science/blob/main/Capstone/Code/5-eda%20data%20viz.ipynb

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
  - 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 all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function.
  - 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.
- https://github.com/dongyuegeo/ibm_data_science/blob/main/Capstone/Code/4-jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

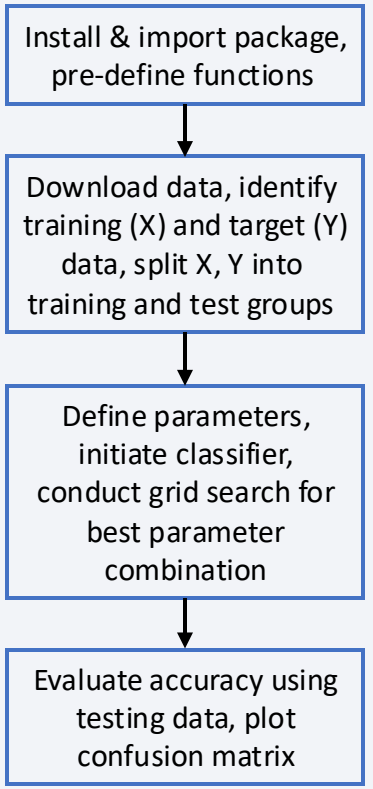
- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
  - The four launch sites were added as markers to the map
  - For each launch site, the launch outcomes were added as colored icon to show total number of launches and the number of successful/failed launches
  - At VAFB SLC-4E, circles were added to show number of launches and a line was added to show the distance between the launch site and a nearby rail way
- Explain why you added those objects
  - The folium map allows to interactively visualize the location of each launch site and figure out how to find the optimal sites
- https://github.com/dongyuegeo/ibm_data_science/blob/main/Capstone/Code/6-lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- The plots, graphs, and interactions added to the dashboard include:
  - A dropdown input allows to interactively select the success rate at which individual launch site or all sites to show
  - Pie charts showing the successful launch ratio at any individual site or all sites
  - A range slider to select the range of payload
  - A scatterplot showing the correlation between the selected payload and their launch outcome
- Explain why you added those plots and interactions
  - These plots and interactions allows to perform interactive visual analytics on SpaceX launch data in real-time
- https://github.com/dongyuegeo/ibm_data_science/blob/main/Capstone/Code/7-spacex-dash-app.py

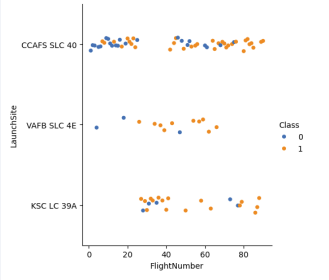
Predictive Analysis (Classification)

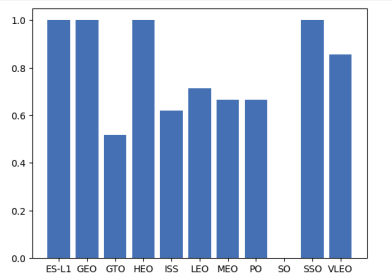
- Summarize how you built, evaluated, improved, and found the best performing classification model
  - Starting from loading necessary Python libraries
  - Download rocket and launch data, format the data and organize into feature data (X) and target vector (Y)
  - Perform train test split
  - For each classifier:
    - Define parameter set for grid search to choose from
    - Initiate the classifier
    - Define pipeline and grid search across all parameter combinations to choose the parameter combination with best accuracy
    - Test the trained classifier with the best performing parameters to predict the launching outcome using test data
    - Evaluate test data accuracy and generating confusion matrix
- https://github.com/dongyuegeo/ibm_data_science/blob/main/Capstone/Code/8_SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



Results (more in the next session)

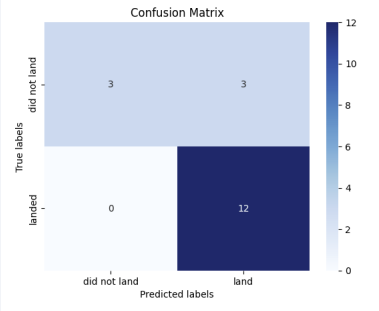
- Exploratory data analysis results

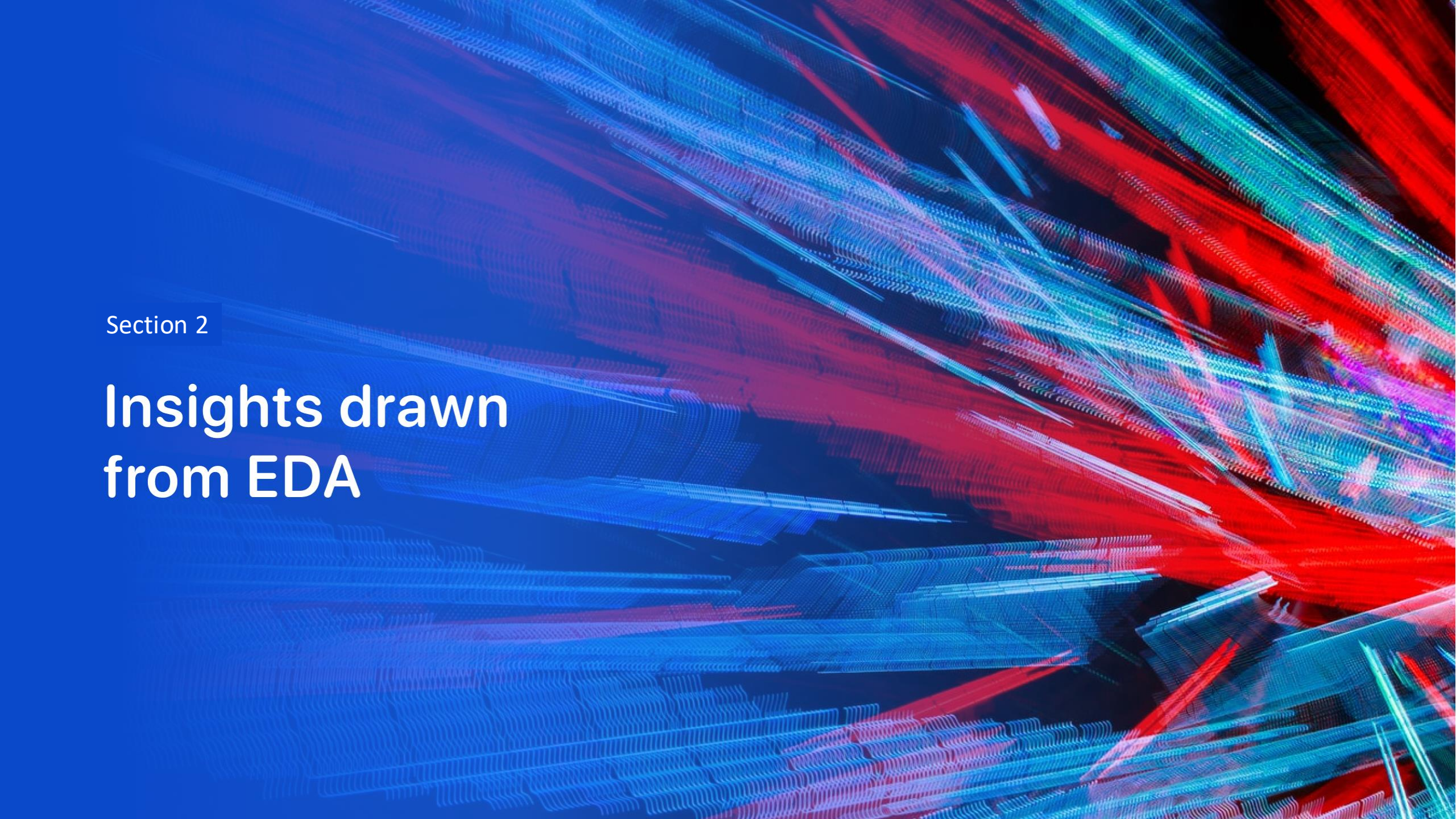




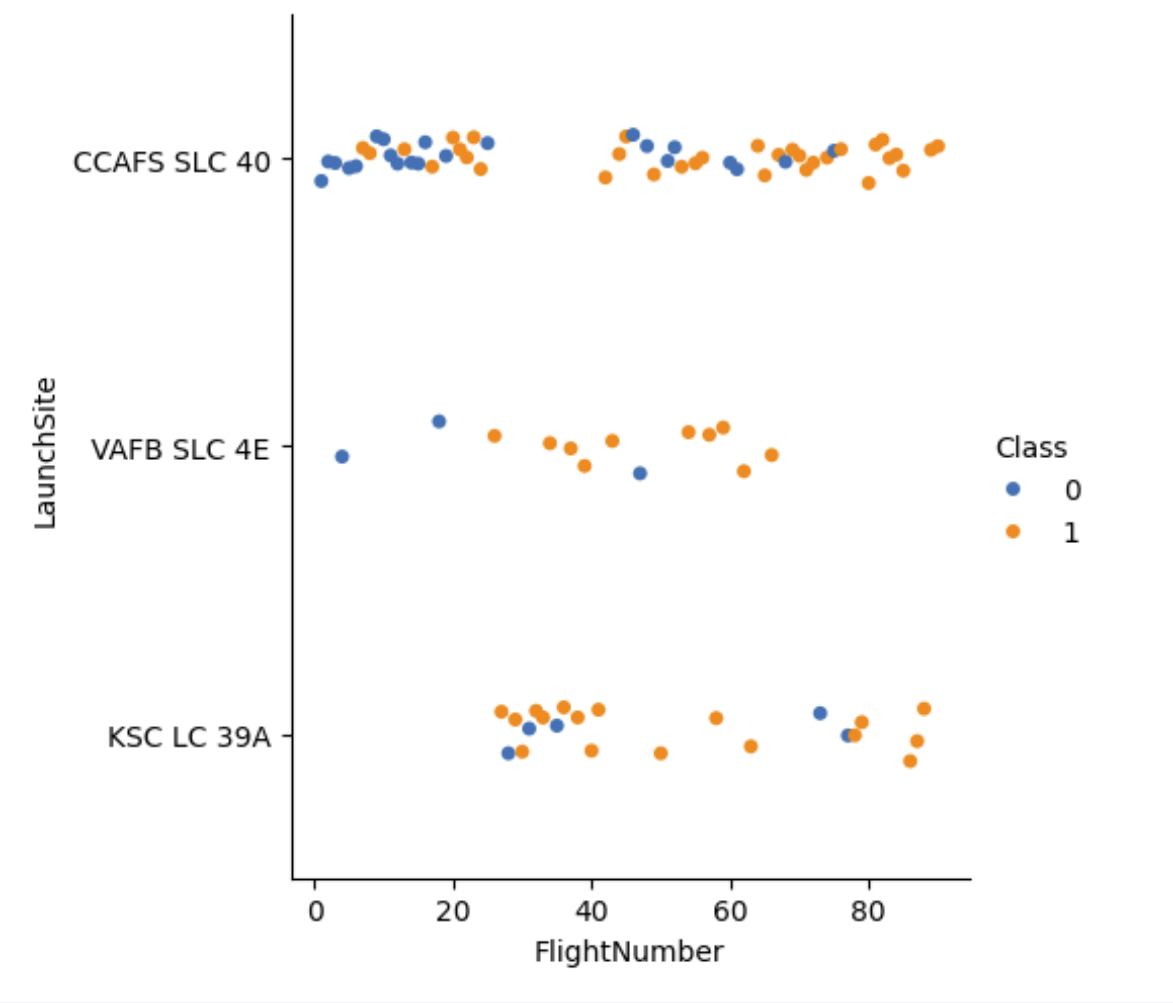
- Interactive analytics demo in screenshots

- Predictive analysis results



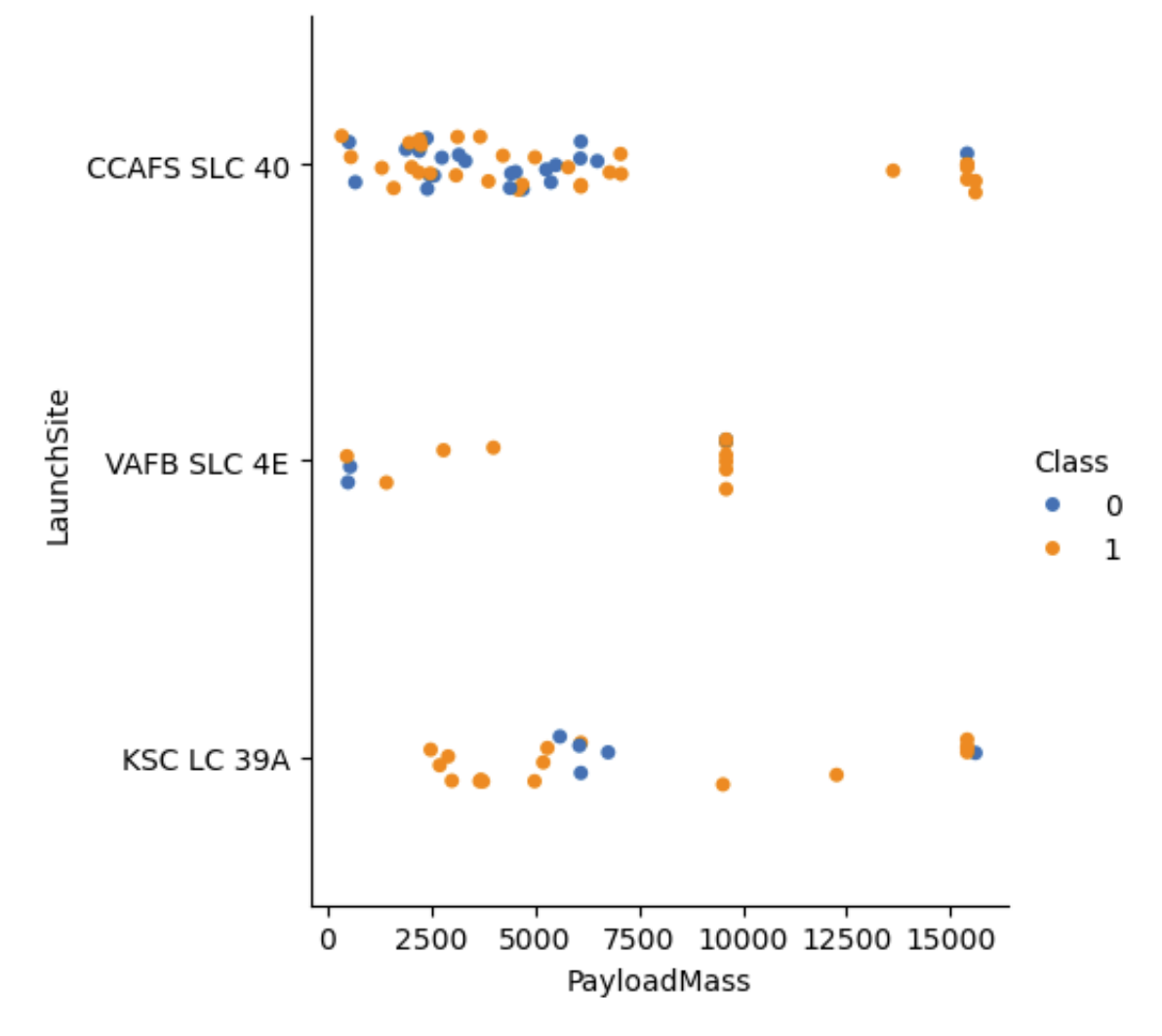

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



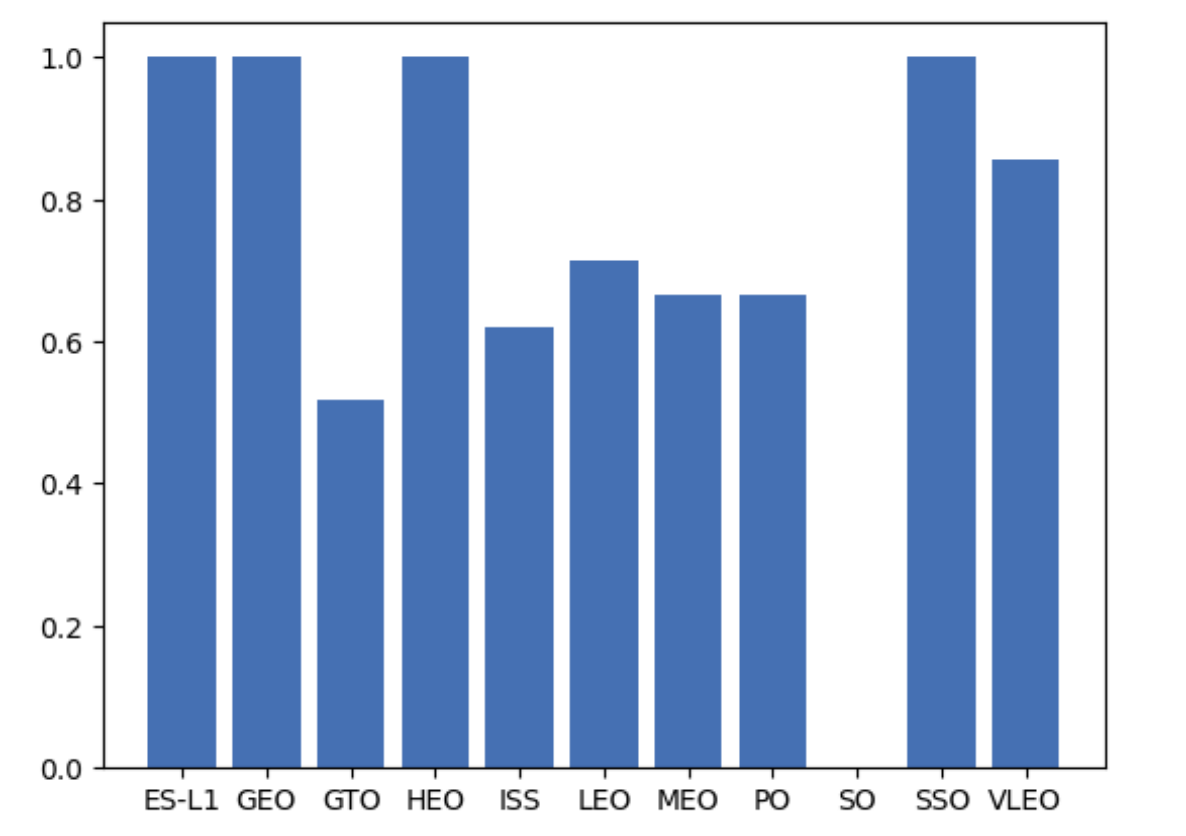
- In all 3 sites, early launches have more unsuccessful cases. Successful launch rate increase as more launches are conducted at the site
- CCAFS SLC 40 tends to be the primary site, as it has the most number of launches and the earliest launches
- VAFB SLC4 4E has the most significant improvement in success rate

Payload vs. Launch Site



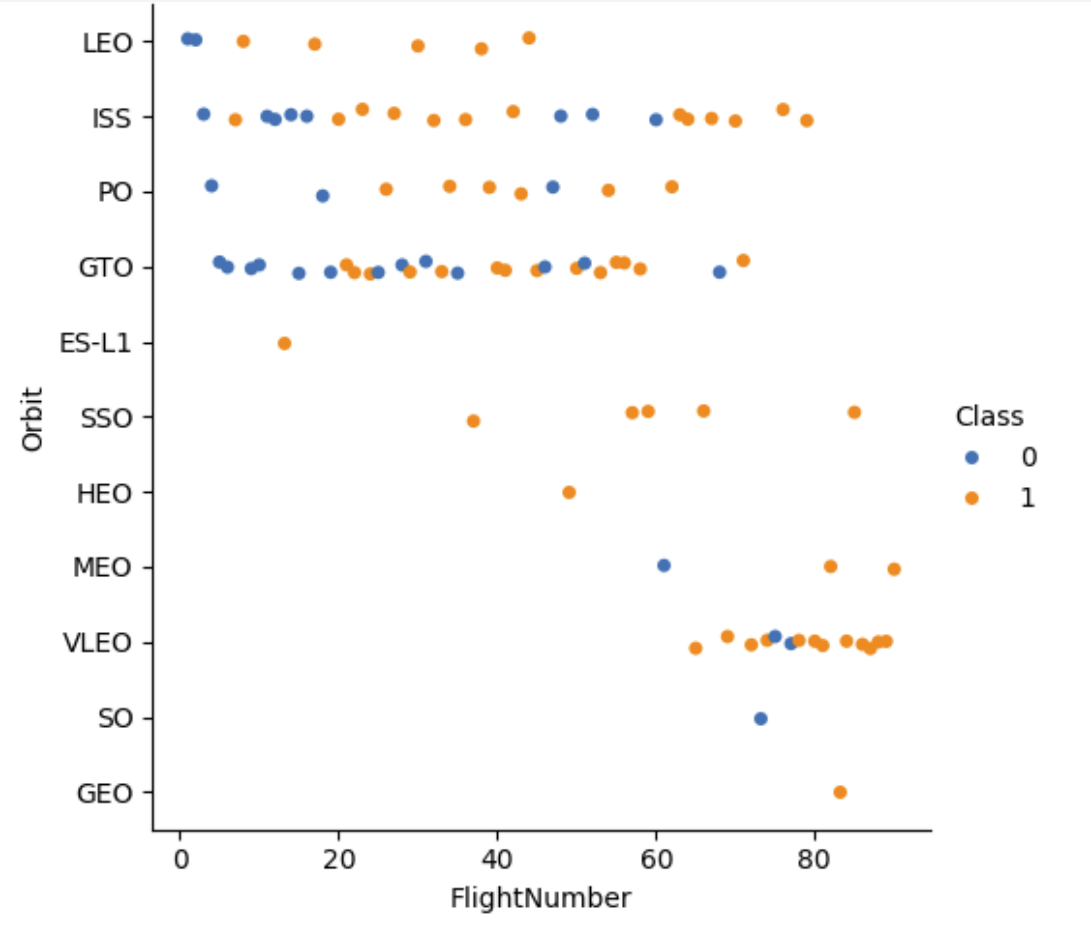
- Heavy payloads (>10,000) are launched in CCAFS SLC 40 and KSC LC 39A
- VAFB SLC 4E launches relatively small payloads
- CCAFS SLC 40 has higher success rates than KSC LC 39A for launching heavy payloads

Success Rate vs. Orbit Type



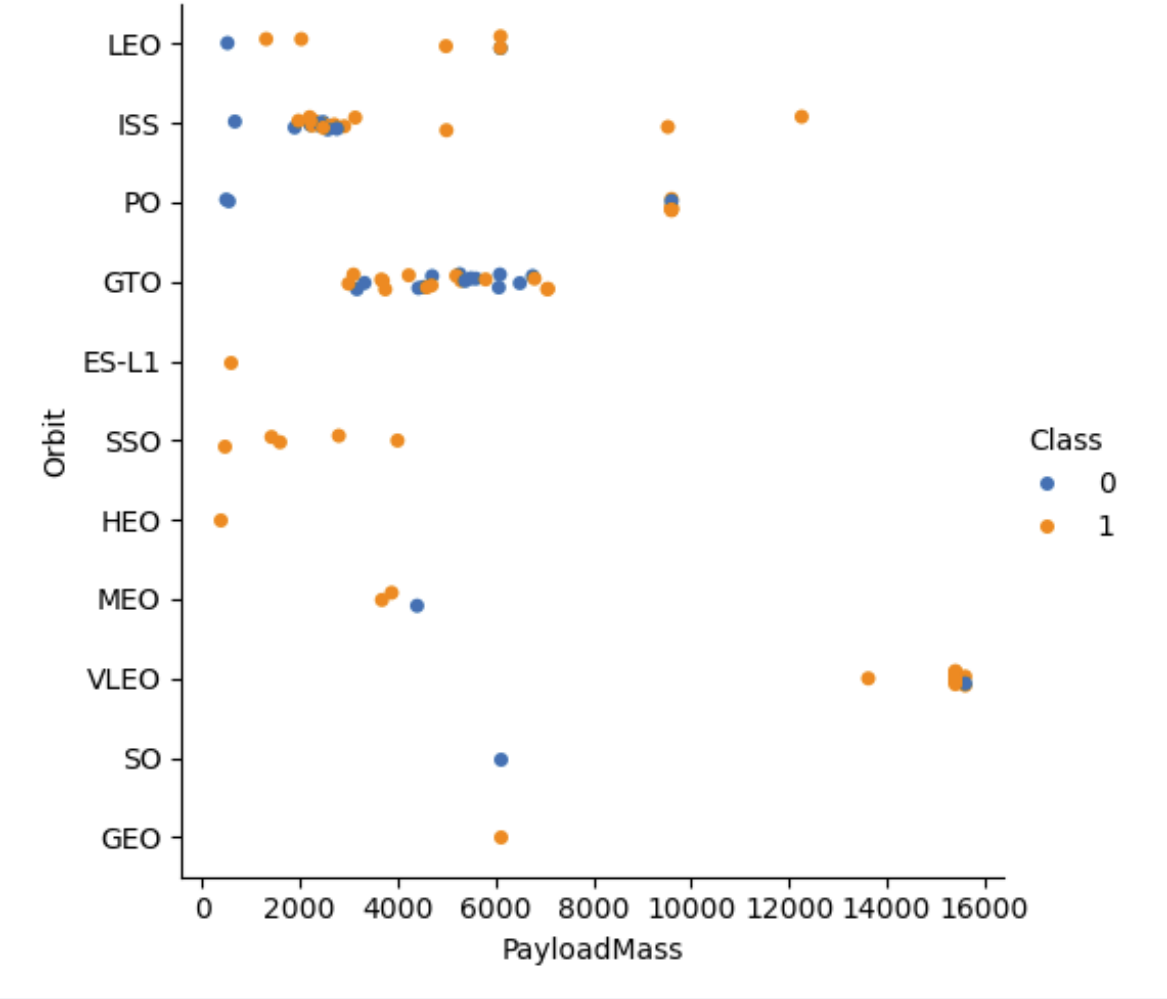
- Heavy payloads (>10,000) are launched in CCAFS SLC 40 and KSC LC 39A
- VAFB SLC 4E launches relatively small payloads
- CCAFS SLC 40 has higher success rates than KSC LC 39A for launching heavy payloads

Flight Number vs. Orbit Type



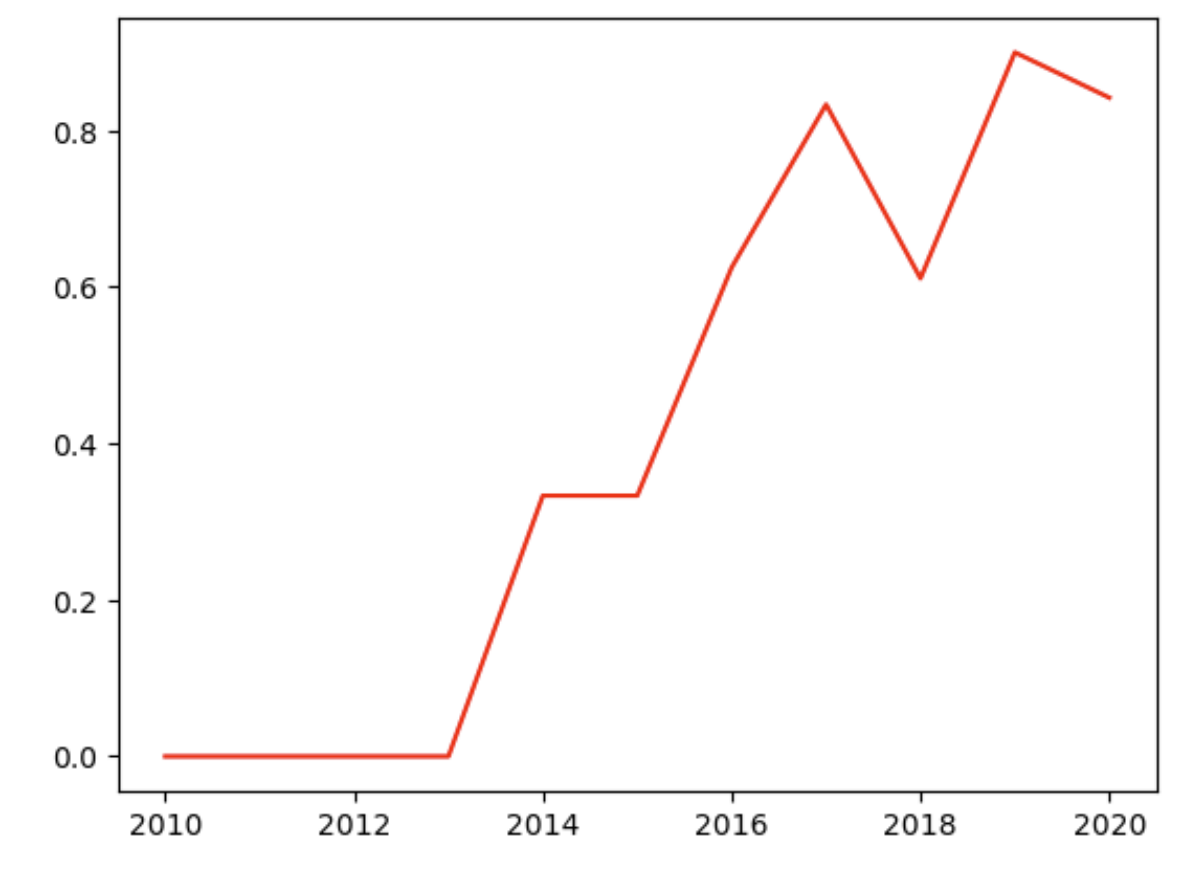
- Early launches (those with small flight numbers, <40) focus on near-surface orbit, as these orbits do not require large rockets and is comparatively easier than heavy rockets
- Early flights has more failures (0). Success rate increases after 20 launches.

Payload vs. Orbit Type



- Generally, small payloads are mounted at near-surface or low orbit satellite
- Large payloads could be at mid-orbit or high-orbit

Launch Success Yearly Trend



- Success rate is 0 ahead of 2013
- Success rate kept increasing from 2013 to 2020

All Launch Site Names

- Find the names of the unique launch sites
  - %sql select distinct "Launch_Site" from SPACEXTABLE

- Present your query result with a short explanation here
  - There are 4 unique launch sites for Falcon 9, as below

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
  - %sql select * from SPACEXTABLE where "launch_site" like "CCA%" limit 5
- Present your query result with a short explanation here
  - As attached below, the 5 records have 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

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
  - %sql select sum("PAYLOAD_MASS__KG_") as total_mass from SPACEXTABLE where customer like "NASA%"
- Present your query result with a short explanation here
  - Total payload carried by boosters from NASA is 99980 kg

total_mass
99980

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
  - %sql select avg("PAYLOAD_MASS__KG_") as ave_mass from SPACEXTABLE where "Booster_Version" like "F9 v1.1%"
- Present your query result with a short explanation here
  - average payload mass carried by booster version F9 v1.1 is 2535 kg

ave_mass
2534.6666666666665

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
  - %sql select min("Date") from SPACEXTABLE where "Landing_Outcome"="Success (ground pad)"
- Present your query result with a short explanation here
  - first successful landing outcome on ground pad was on 2015-12-22

min("Date")
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
  - %sql select "Payload" from SPACEXTABLE where ("Landing_Outcome"="Success (drone ship)") AND ("PAYLOAD_MASS__KG_">4000 and "PAYLOAD_MASS__KG_"<6000)
- Present your query result with a short explanation here
  - The following boosters have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Payload
JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
  - %sql select "Mission_Outcome", count(*) as count_out from SPACEXTABLE group by "Mission_Outcome"
- Present your query result with a short explanation here
  - The total number of successful and failure mission outcomes is as below:

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
  - %sql select "Booster_Version", "PAYLOAD_MASS__KG_" from SPACEXTABLE where "PAYLOAD_MASS__KG_" = (select max("PAYLOAD_MASS__KG_") from SPACEXTABLE)
- Present your query result with a short explanation here

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

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - %sql select substr(Date, 6,2) as month, "Booster_Version", "Launch_Site" from SPACEXTABLE where (substr(Date,0,5)='2015') AND ("Landing_Outcome"="Failure (drone ship)")
- Present your query result with a short explanation here

month	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

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

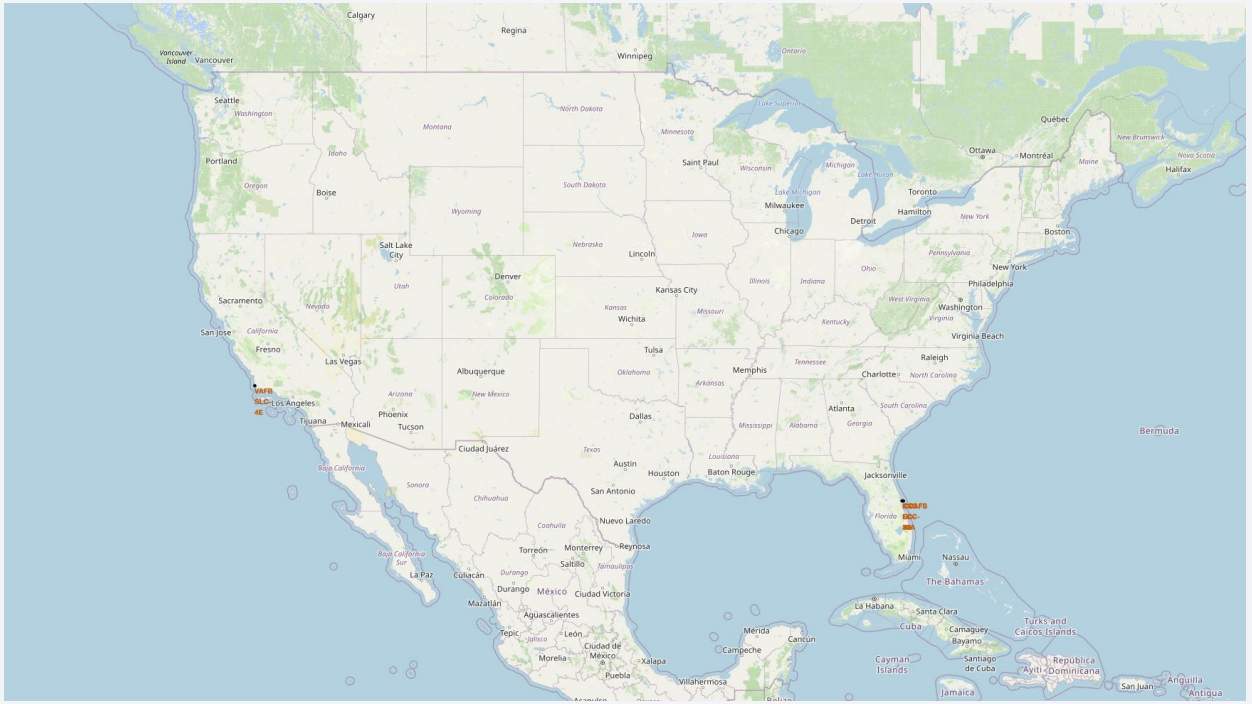
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order
  - %sql select "Landing_Outcome", count("Landing_Outcome") as count_out from SPACEXTABLE where ("Date">"2010-06-04" and "Date"<"2017-03-20") group by "Landing_Outcome" order by count_out desc
- Present your query result with a short explanation here

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

Section 3

Launch Sites Proximities Analysis

Location of Four Launching Sites



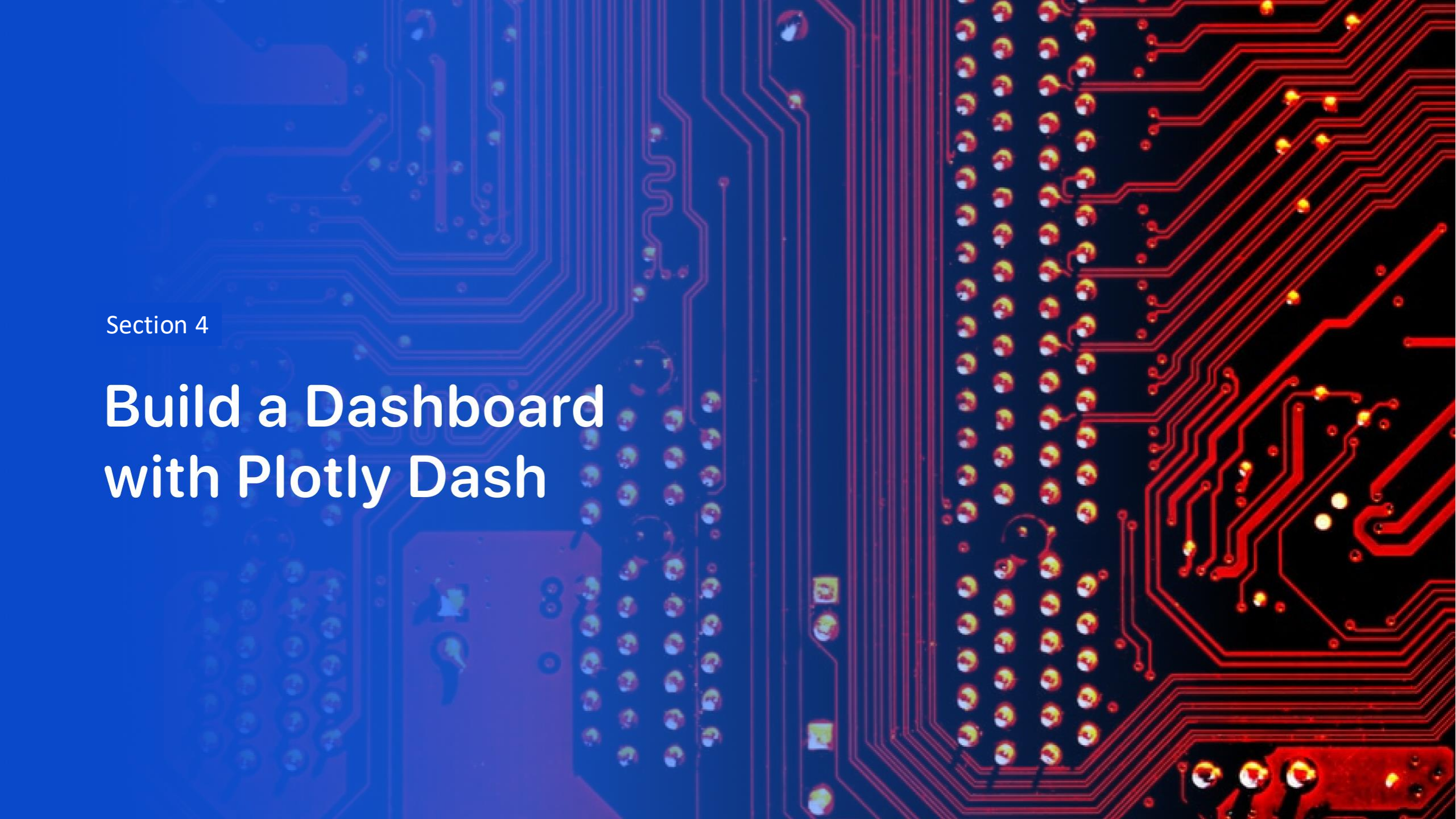
- Screenshot (left) shows the location of four launching sites
  - Circle marks represent all launching sites are in low-latitude area in the U.S.
  - All sites are close to the ocean

VAFB SLC-4E Launching Record

- The screenshot on the left shows the launching outcomes at this site
  - Big black circle shows the extent of the launch site, while marker cluster with colored icons represent past launches and outcomes
  - There were 10 total launches
  - Among the 10, 6 were successful (red) and 4 were failed (green)

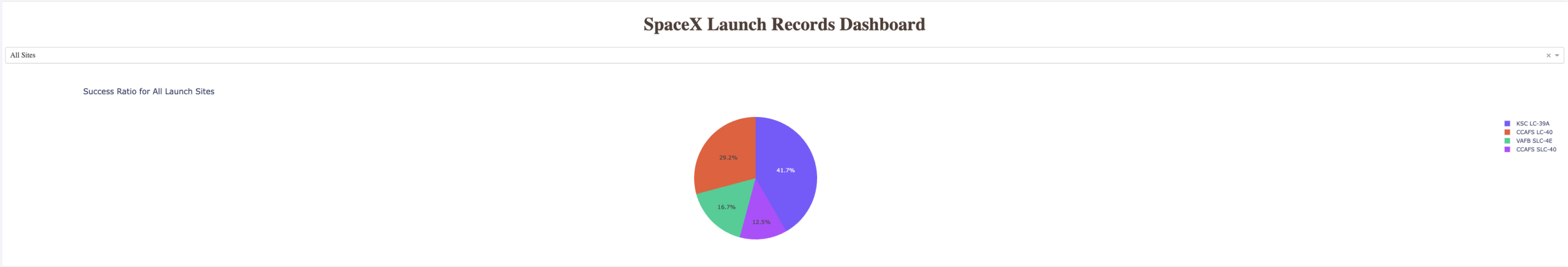
Distance between VAFB SLC-4E and Railway

- Screenshot shows the launch pad and nearby railways
- The shortest distance between the launch pad and the railway is about 0.015 km


Section 4
Build a Dashboard with Plotly Dash

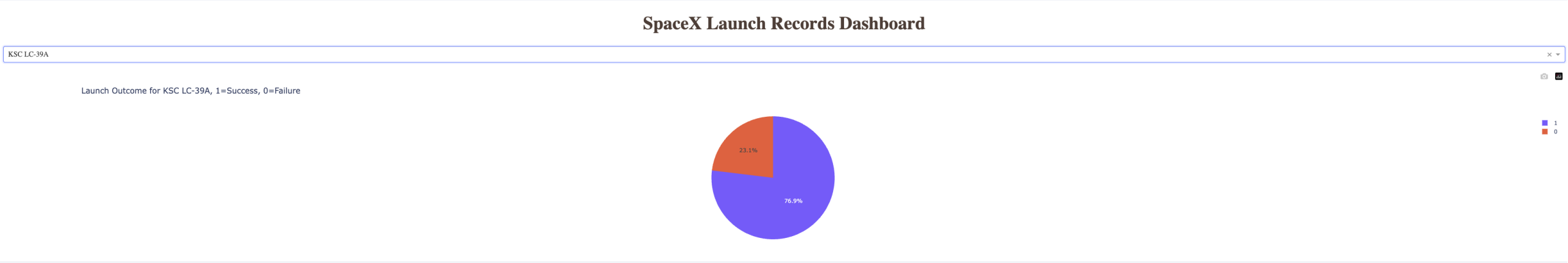
Pie chart for Success Launches by All Site

- Screenshot of pie chart for all sites below show the successful launches at each site to the total number of successful launches
- KSC LC-39A has the largest count of successful launches, while CCAFS SLC-40 has the least successful among the four sites



Launch Outcomes at Sites with Highest Success Ratio

- Pie chart shows the success and unsuccess launch ratio at Site KSC LC-39A, which is the one with the highest success ratio; 73.1% success vs 26.9% fail



Payload vs. Launch Outcome Scatter Plot for All Sites

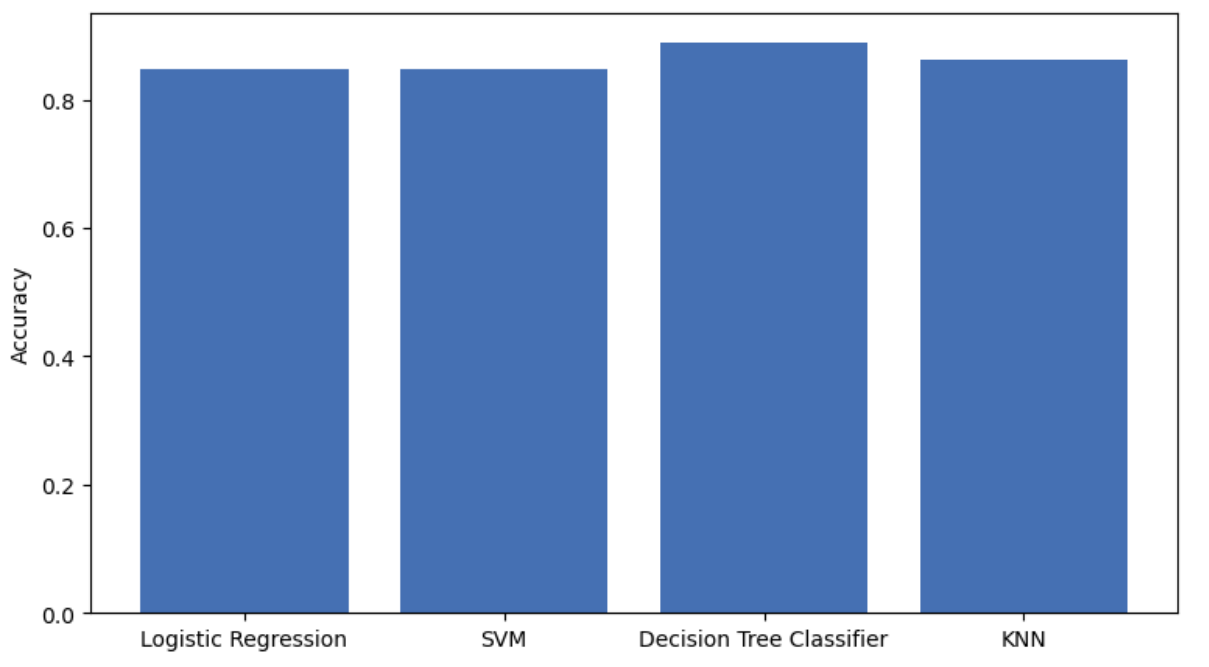
- Screenshots show Payload vs. Launch Outcome scatter plot for all sites with different payload selected in the range slider
- Payload range from 0 to 9600 Kg. Booster version FT has the largest amount of successful launches. There is no clear correlation between payload and outcome.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

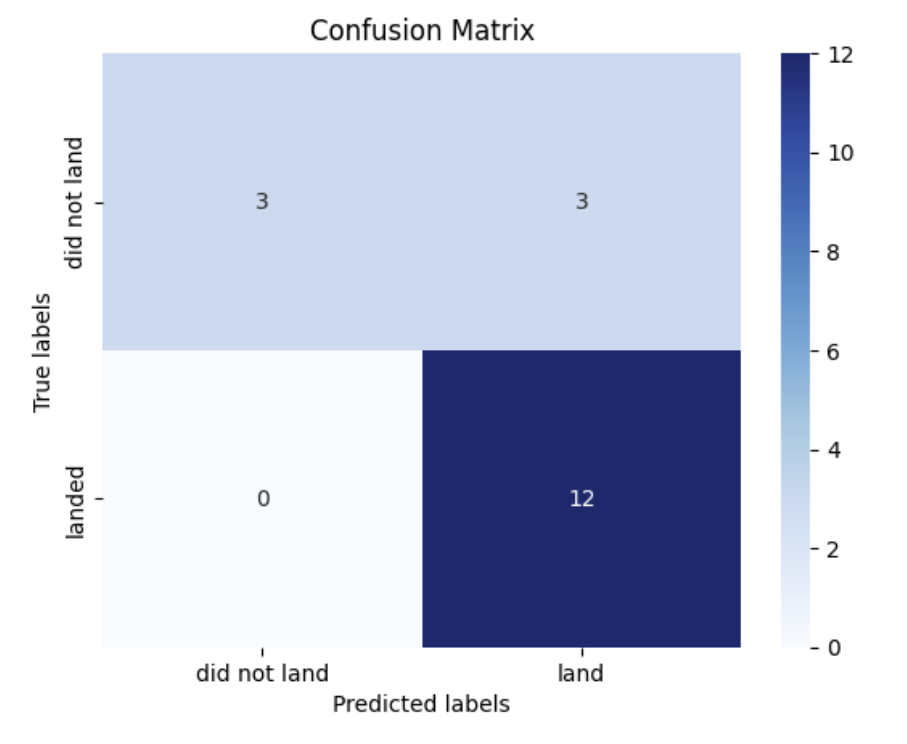
Figure 1. Accuracy of the four classification methods



- Conclusion: Decision Tree Classifier has the highest accuracy among the four methods

Confusion Matrix

Figure 1. confusion matrix of decision tree classifier



- The classifier successfully forecasts 15 out of the 18 landing outcome, including 12 successful landing and 3 unsuccessful landing
- The classifier has 3 false negative and 0 false positive forecasts

Conclusions

- All four methods successfully predicted over 80% of the landing outcomes
- Among the 4 prediction methods, decision tree classifier has the highest forecasting accuracy
- Grid Search is able to find the best combination of parameters for each method
- The more complex methods, i.e. those having more parameters or takes longer time to run, do not necessarily render more accurate predictions

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!