Winning Space Race with Data Science

Sayer Niblett

August 1, 2024

Outline

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

Executive Summary

Summary of methodologies

- Data collection: API & web scraping.
- Data wrangling: replace missing values, code variables appropriately, create a landing_class variable.
- Exploratory data analysis: analyze outcomes by orbit type, payload mass, booster version; analyze outcomes using charts.
- Interactive data visualization: analyze outcomes and launch sites using Folium and Plotly Dash.
- Predictive analysis: using classification models—logistic regression, SVM, decision tree, and KNN.

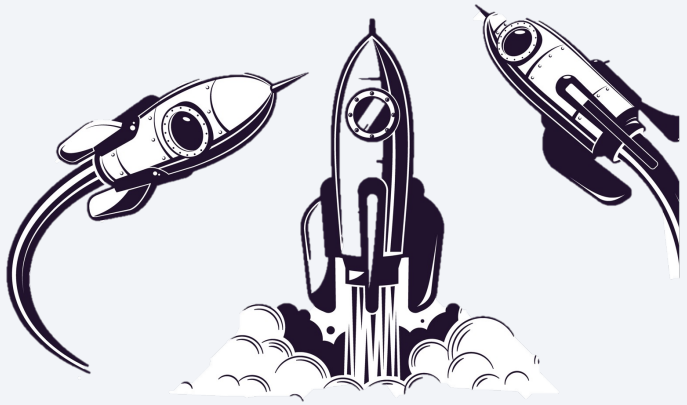
Summary of all results

- Exploratory data analysis:
  - Average landing success rate has increased over time.
  - Higher success at ES-L1, GEO, HEO, SSO orbits.
- Interactive analytics:
  - The launch site with the highest success rate is the Kennedy Space Center Launch Complex (KSC LC-39A). At this site, 76.9% of all launches were successful.
  - Highest launch success is between 2,000 and 6,000 kg payload mass.
  - At higher payload mass there are fewer successful launches.
- Predictive analysis:
  - The model with the most predictive power (highest accuracy) was Logistic Regression.

Introduction

- Project background and context
  - The commercial space age is here, companies are making space travel affordable for everyone.
  - Examples: Virgin Galactic (suborbital spaceflights), Rocket Lab (satellite provider), Blue Origin (sub-orbital and orbital reusable rockets), SpaceX.
  - SpaceX has sent manned missions to space and spacecraft to the International Space Station.
  - SpaceX's success? SpaceX's rocket launches are relatively inexpensive and much of that savings is because it can (sometimes) reuse the first stage of the launch.
  - Space Y, a company founded by Billionaire industrialist Allon Musk would like to compete with SpaceX.

- Problems you want to find answers
  - What will be the price of each launch?
  - Will the first stage land successfully?
  - Will SpaceX reuse the first stage of a given launch?



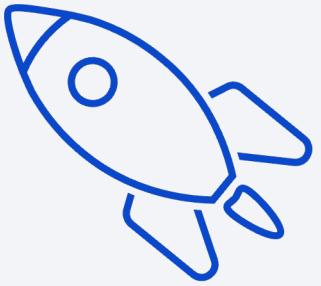
Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX-API, web scraping SpaceX Wikipedia page
- Perform data wrangling
  - Replace missing values, code variables appropriately, create a landing_class variable.
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Analyze outcomes by orbit type, payload mass, booster version; analyze outcomes using charts.
- Perform interactive visual analytics using Folium and Plotly Dash
  - Folium: visual analysis of launch sites on map.
  - Plotly Dash dashboard: visualize success rates by launch site, payload mass, and booster version.
- Perform predictive analysis using classification models
  - Classification models: logistic regression, SVM, decision tree, KNN.
  - Find the best hyperparameters using Grid Search
  - Select the best method based on the testing data



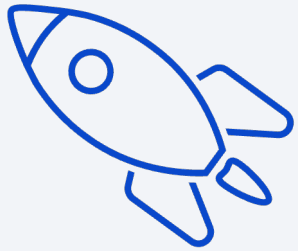
Data Collection

- The data sets were collected using an API and web scraping.
- SpaceX REST API: SpaceX launch data was gathered from the SpaceX REST API. This API provides data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome. That data was delivered in a JSON format and then converted to a data frame.
- Web scraping: Python's BeautifulSoup package was used to web scrape HTML tables containing Falcon 9 launch records. Data from the tables were parsed and then converted to a Pandas data frame.

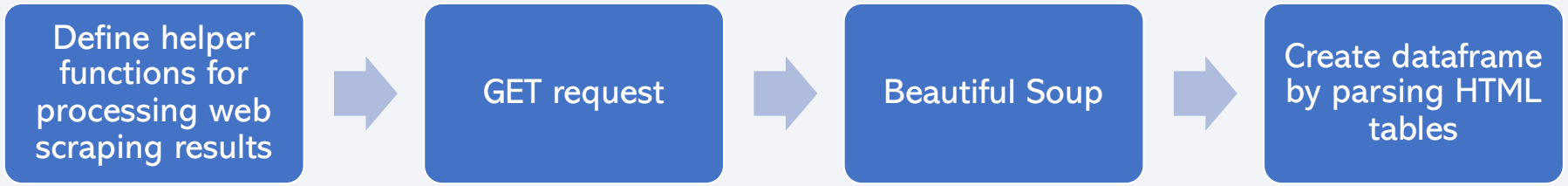
Data Collection – SpaceX API

Define helper functions for use with the API → Request and parse SpaceX launch data using GET request → Convert data to Pandas dataframe → Filter the dataframe to only include Falcon 9

- Define helper functions: getBoosterVersion, getLaunchSite, getPayloadData, getCoreData.
- Request data from SpaceX API with GET request.
- Convert data to a Pandas dataframe.
- Filter the data to include only Falcon 9 instances.
- GitHub URL to Jupyter Notebook (Lab 1: Collecting the data)

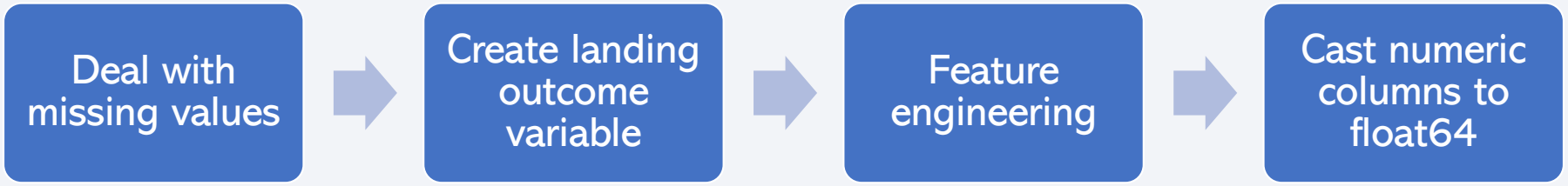


Data Collection - Scraping



- Define helper functions: date_time, booster_version, landing_status, get_mass, extract_column_from_header.
- Request Falcon9 Launch Wiki page from URL using GET request.
- Use BeautifulSoup to extract data from response.
- Create dataframe by parsing HTML tables.
- GitHub URL to Jupyter Notebook (Lab 1: Web scraping Falcon 9 and Falcon Heavy Launches Records from Wikipedia)

Data Wrangling



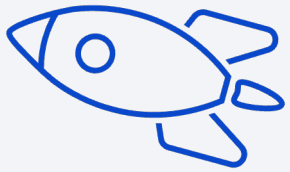
- Deal with missing values: Identify and calculate the percentage of the missing values in each attribute; replace missing values or drop observations.
- Create landing outcome variable: Create a list where the element is zero if the corresponding row in Outcome is in the set bad_outcome; otherwise, it's one; assign it to the variable landing_class.
- Feature engineering: create dummy variables for categorical columns.
- Cast the entire dataframe to variable type float64.
- GitHub URL to Jupyter Notebook (Lab 2: Data Wrangling)

EDA with Data Visualization: Charts Plotted

- Payload Mass vs. Flight Number: shows that as the flight number increases, the first stage is more likely to land successfully and that the more massive the payload, the less likely the first stage will return.
- Launch Site vs. Flight Number: shows that the most launches have occurred at CCAFS SLC 40 and that the launch site VAFB SLC 4E has the fewest failed launches.
- Launch Site vs. Payload Mass: shows that for the VAFB-SLC launch site there are no rockets launched for heavypayload mass (greater than 10000).
- Success Rate vs. Orbit Type: shows that the highest success rates are in ES-L1, GEO, HEO, SSO, and VLEO.
- Orbit Type vs. Flight Number: shows that in LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.
- Orbit Type vs. Payload Mass: shows that with heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- Launch Success Yearly Trend: shows that the success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing.
- GitHub URL to Jupyter Notebook (Lab 3: Exploring and Preparing Data)

EDA with SQL

- Exploratory data analysis was performed using SQL queries (see Appendix for code):
  - List of distinct launch sites
  - List of 5 records where launch sites begin with the string 'CCA'
  - Total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - Date when the first successful landing outcome in ground pad was achieved
  - Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - Total number of successful and failure mission outcomes
  - Names of the booster_versions which have carried the maximum payload mass
  - 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
- GitHub URL to Jupyter Notebook (Lab 3: SQL)

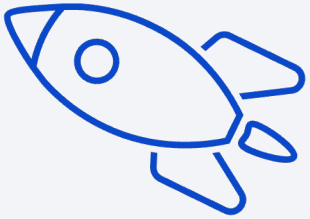


Build an Interactive Map with Folium

- Map objects were added to a Folium map, including:
  - Markers: to identify each unique launch site and their proximity to certain geographic features such as the equator and coasts.
  - Circles around the launch sites with radius of 1,000m.
  - Marker cluster: to show the launch outcomes (success/failure) for each launch site.
  - Lines: to analyze the distance from a launch site to the nearest coast or city.
- GitHub URL to Jupyter Notebook (Lab 4: Launch Sites Locations Analysis with Folium)

Build a Dashboard with Plotly Dash

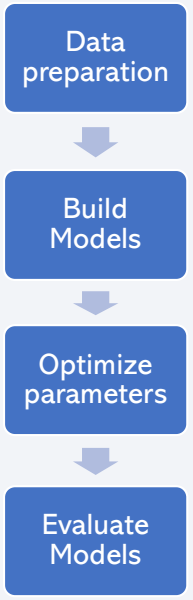
- The following plots/interactions were added to the dashboard:
  - Pie chart showing the success/failure rate for each launch site (and all launch sites).
  - Scatterplot showing the success/failure rate by payload mass and booster version (for all launch sites or a single selected launch site).
- These plots and interactions were added in order to allow the user to visually analyze the success/failure outcomes at individual launch sites, at various payload masses, and for different booster versions.
- This allows for analysis of questions like:
  - Which site has the largest successful launches?
  - Which site has the highest launch success rate?
  - Which payload range(s) has the highest launch success rate?
  - Which payload range(s) has the lowest launch success rate?
  - Which F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) has the highest launch success rate?



- GitHub URL to Dash app code

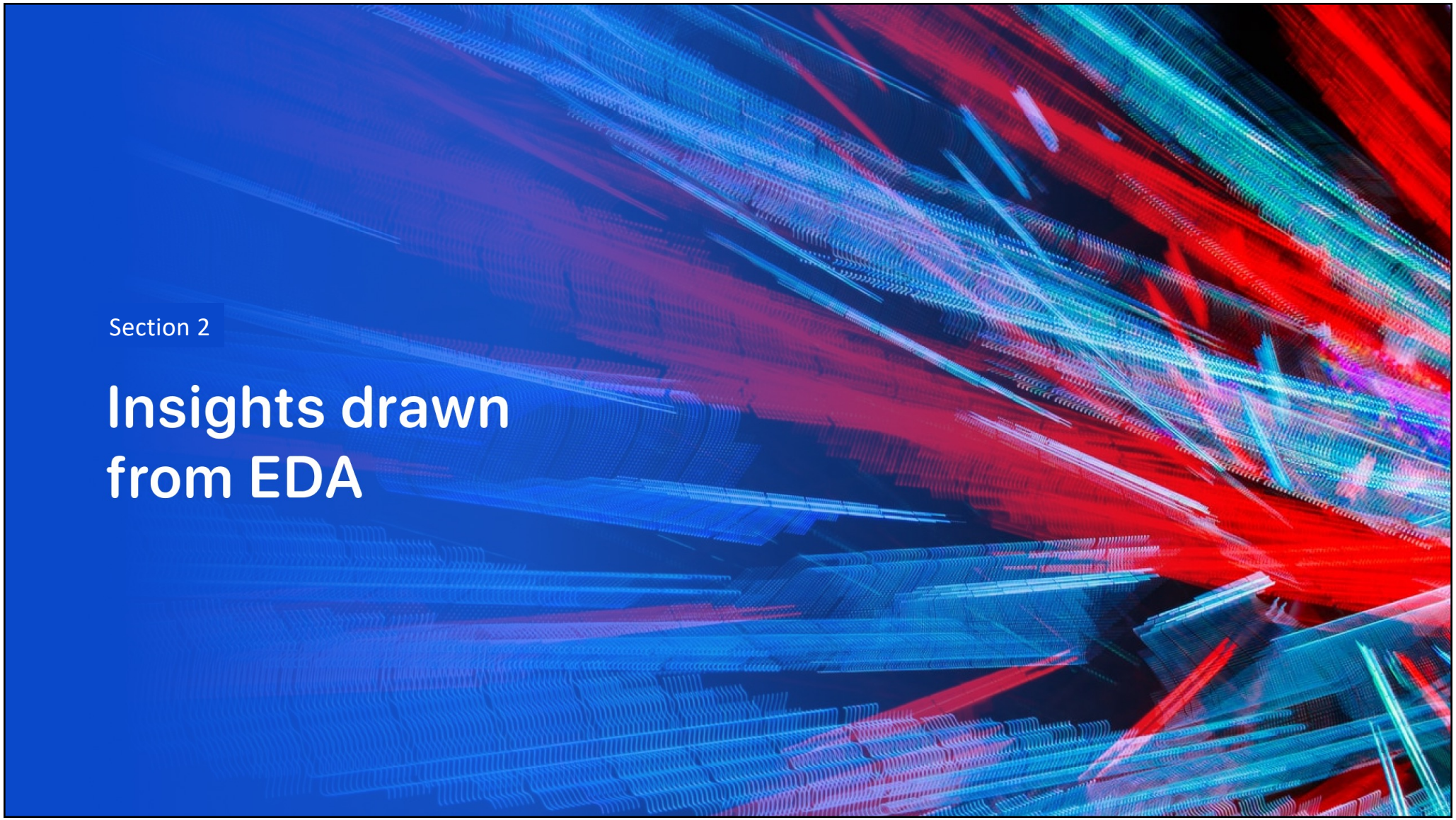
Predictive Analysis (Classification)

- Preparing data:
  - Create a column for the "Class" outcome (the target variable)
  - Standardize the data
  - Split the data X (predictor variables) and Y (Class) into training and testing data sets
- Model building:
  - Logistic Regression
  - Support Vector Machine (SVM)
  - Decision Tree
  - K-Nearest Neighbor (KNN)
- Model optimization
  - Use Grid Search to identify the best parameters for each model
- Model evaluation:
  - Confusion matrix
  - Compare the accuracy score of the testing data for each model
- GitHub URL to Jupyter Notebook (Lab 5: Machine Learning Prediction)

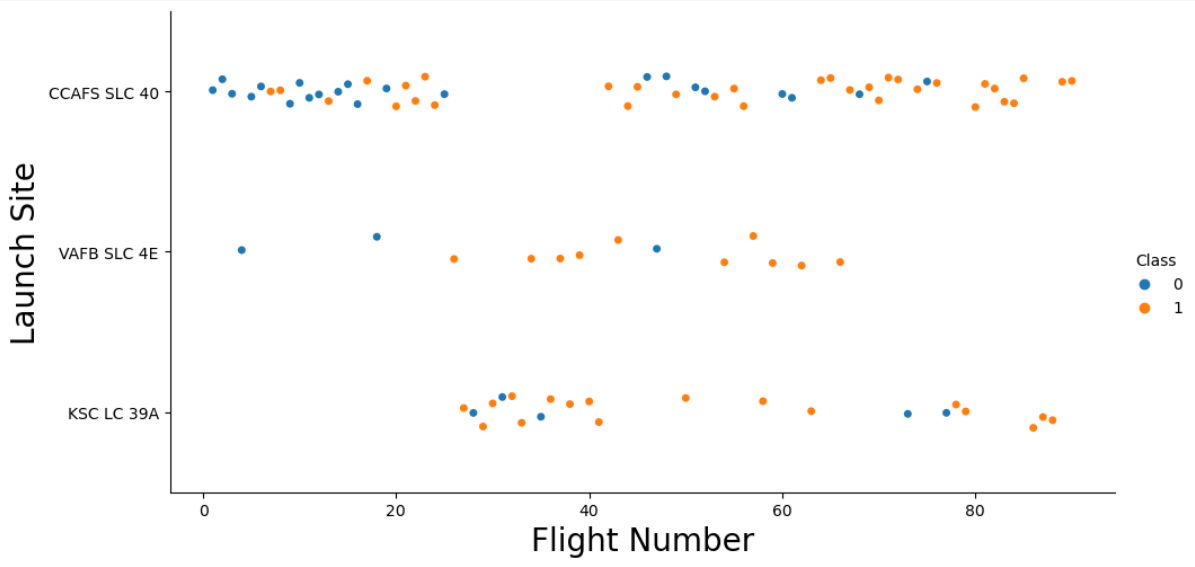


Results

- Exploratory data analysis results
  - Average landing success rate has increased over time.
  - Higher success at ES-L1, GEO, HEO, SSO orbits.
- Interactive analytics demo in screenshots
  - The launch site with the highest success rate is the Kennedy Space Center Launch Complex (KSC LC-39A).
  - At this site, 76.9% of all launches were successful.
  - Highest launch success between 2,000 and 6,000 kg payload mass.
- Predictive analysis results
  - The model with the most predictive power (highest accuracy) was Logistic Regression.

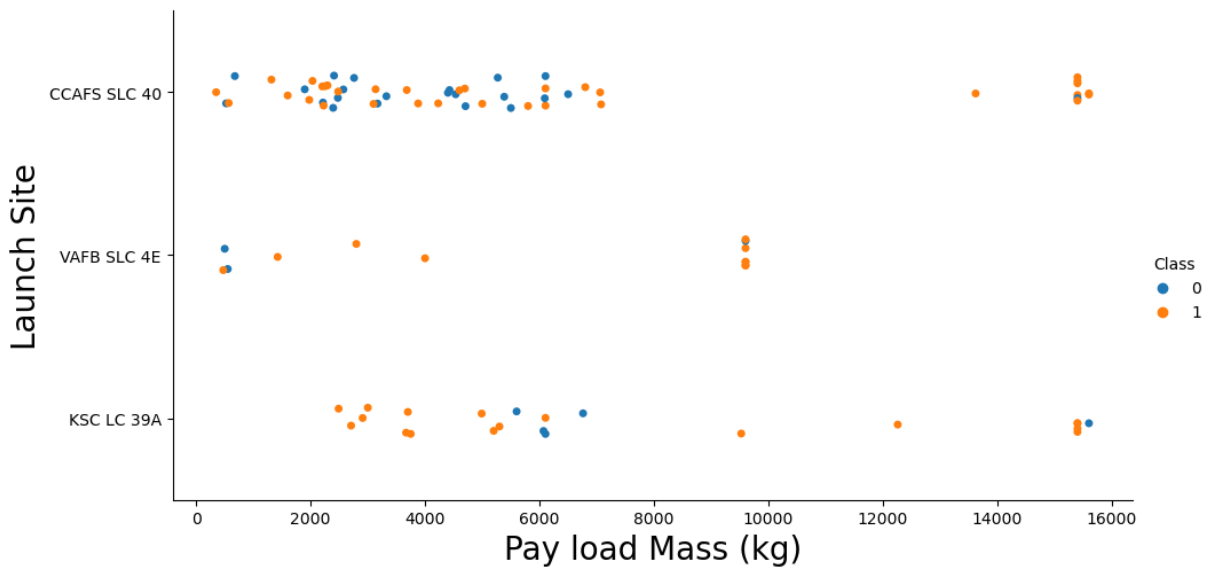

Section 2
Insights drawn
from EDA

Flight Number vs. Launch Site



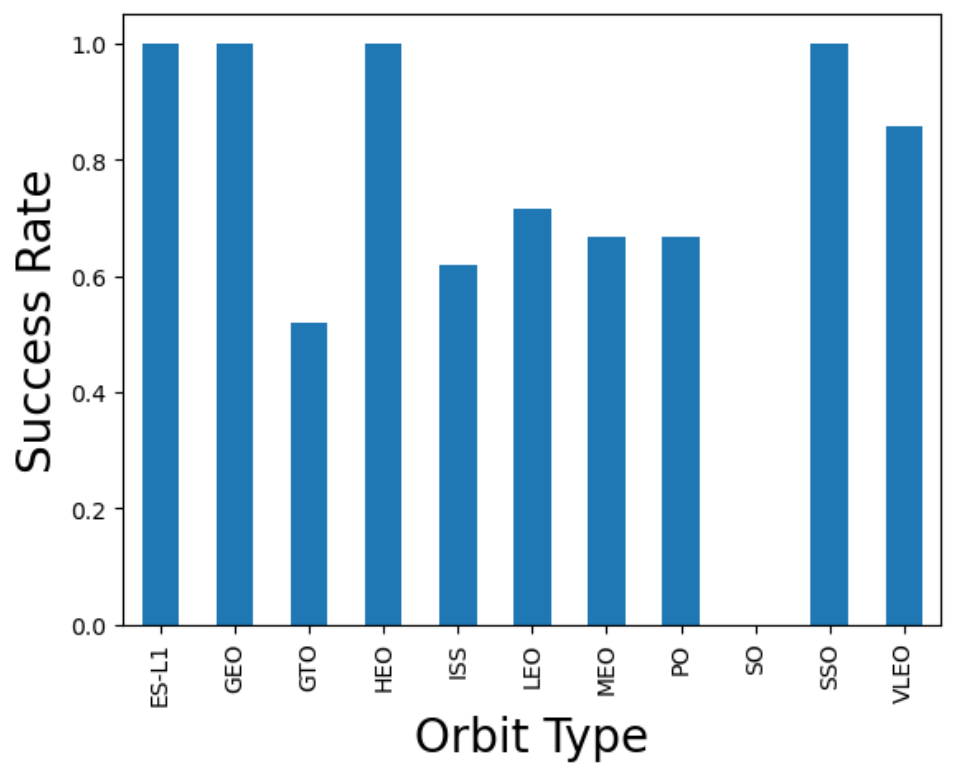
- It appears that the most launches have occurred at CCAFS SLC 40.
- Launch site VAFB SLC 4E has the fewest failed launches.

Payload vs. Launch Site



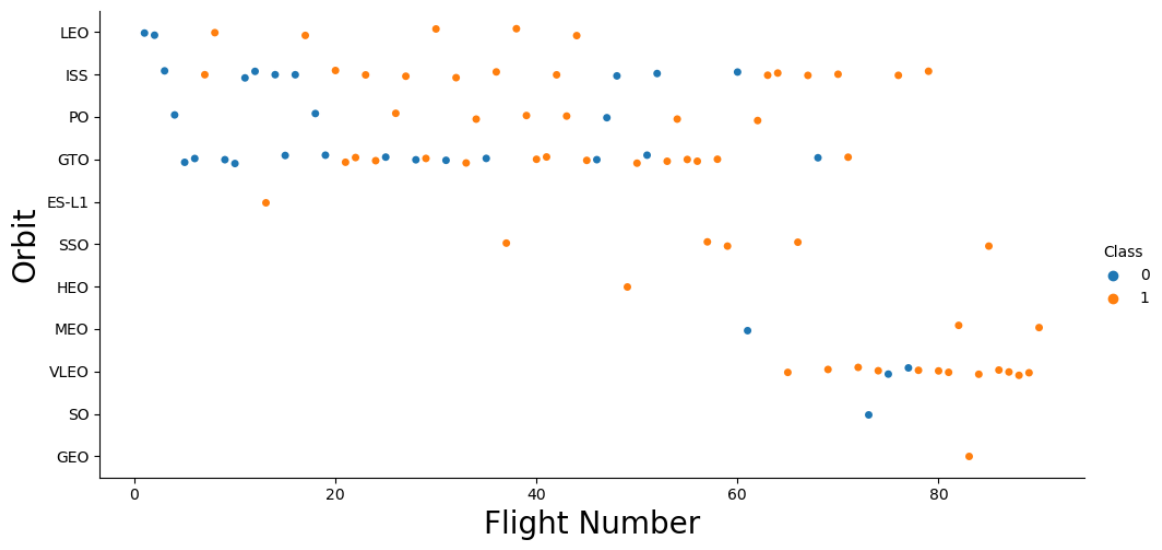
- For the VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000).

Success Rate vs. Orbit Type



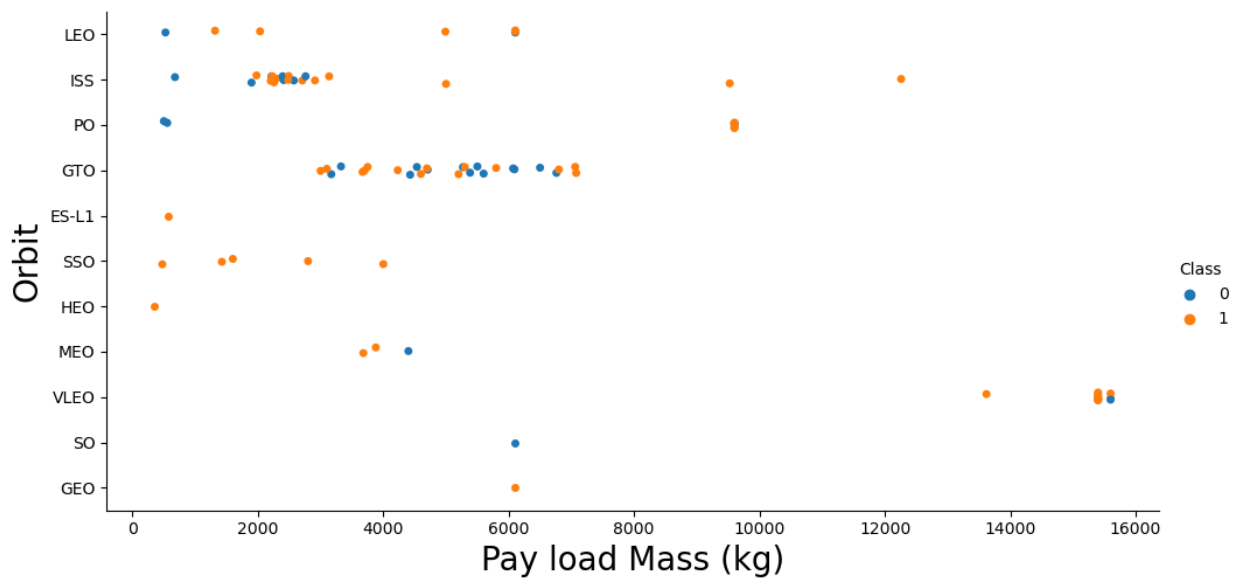
- The orbit types with the highest success rates are ES-L1, GEO, HEO, and SSO.

Flight Number vs. Orbit Type



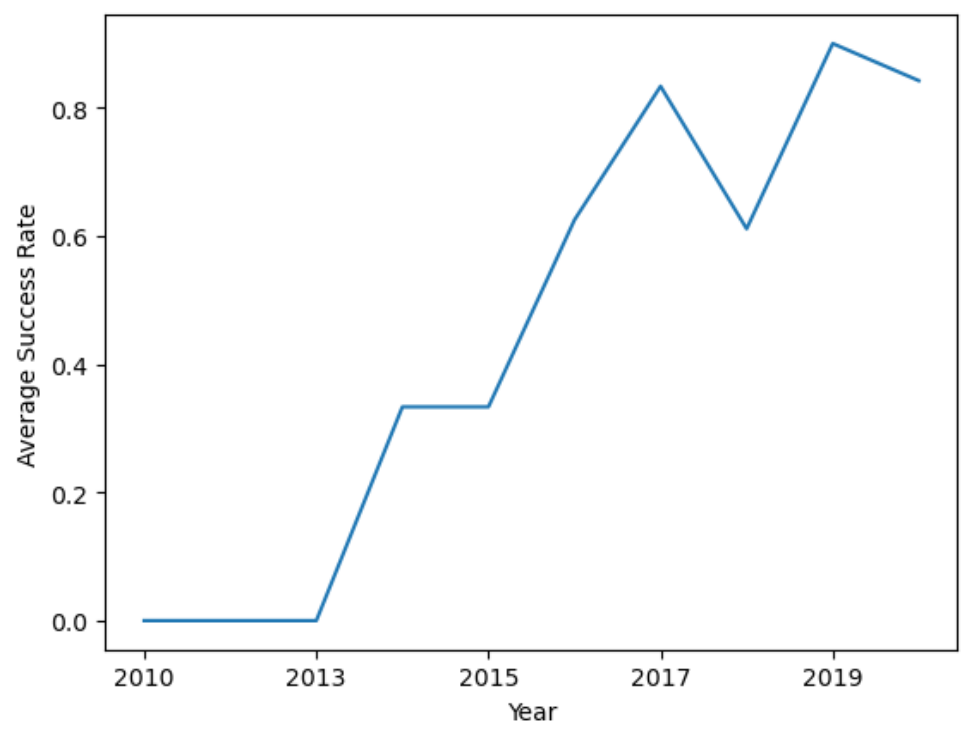
- In the LEO orbit the Success appears related to the number of flights;
- There seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



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

Launch Success Yearly Trend



- The success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing.

All Launch Site Names

- There are 4 unique launch sites (listed in the table to the right).
- Code: %sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTABLE.

Unique Launch Sites
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

- The table shows 5 records where launch sites begin with ‘CCA’
- Code: %sql SELECT * FROM SPACEXTABLE WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5

Total Payload Mass

- The total payload carried by boosters from NASA in our data set is 45,596 kg
- Code: %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE CUSTOMER == "NASA (CRS)"

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2,928.4 kg.
- Code: %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version == "F9 v1.1"

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad is December 22, 2015.
- Code: %sql SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome == "Success (ground pad)"

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- There are 4 boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 (listed in the table below).
- Code: %sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome == "Success (drone ship)" and PAYLOAD_MASS__KG_ BETWEEN 4001 and 5999.

Boosters which have successfully landed on drone ship and had payload mass > 4000 and < 6000 (Booster Version)
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- In our data set there were 100 successful missions and 1 unsuccessful mission.
- Code: %sql SELECT Mission_Outcome, COUNT(*) as NumOutcomes \ FROM SPACEXTABLE \ GROUP BY Mission_Outcome.

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

Boosters Carried Maximum Payload

Boosters Which Have 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

- The names of the 12 boosters which have carried the maximum payload mass are listed in the table to the left.
- Code: %sql SELECT Booster_Version \ FROM SPACEXTABLE \ WHERE PAYLOAD_MASS__KG_ == (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE).

2015 Launch Records

- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 are listed in the table below
- Code: %sql SELECT substr(Date, 6,2) as month, DATE, Landing_Outcome, Booster_Version, Launch_Site \ FROM SPACEXTABLE \ WHERE Landing_Outcome == "Failure (drone ship)" and substr(Date,0,5)='2015';

Failed Drone Ship Launches in 2015				
Month	**Date**	**Landing Outcome**	**Booster Version**	**Launch Site**
01	2015-01-10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015-04-14	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

Landing Outcome	Count
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

- The ranked count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order, is presented in the table to the left.
- The most common outcome was “No attempt”, followed by Success (drone ship) and failure (drone ship).
- Code: %sql SELECT Landing_Outcome, COUNT(Landing_Outcome) \ FROM SPACEXTABLE \ WHERE Date BETWEEN '2010-06-04' and '2017-03-20’ \ GROUP BY Landing_Outcome \ ORDER BY COUNT(Landing_Outcome) DESC.

Section 3

Launch Sites Proximities Analysis

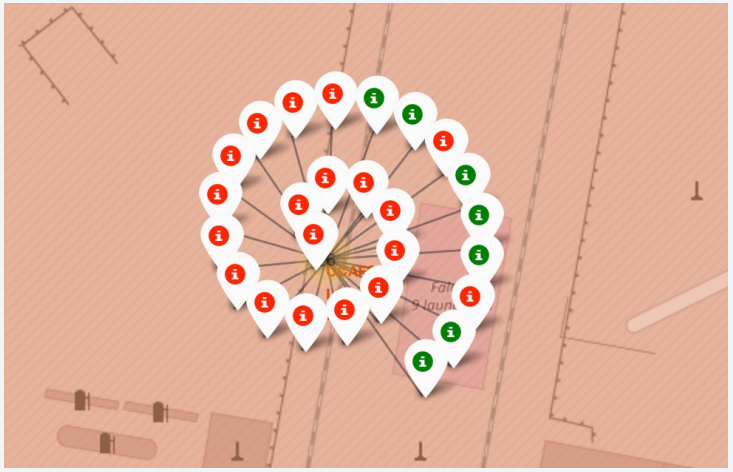
Folium Map: Launch Sites

- Launch sites are on the coasts in Florida and California.
- Launch sites are in proximity to the equator line.

Folium Map: Launch Outcomes

- Launch outcomes for each site were marked on the map.
- Successful launches are indicated by green markers.
- Unsuccessful launches are indicated by red markers.

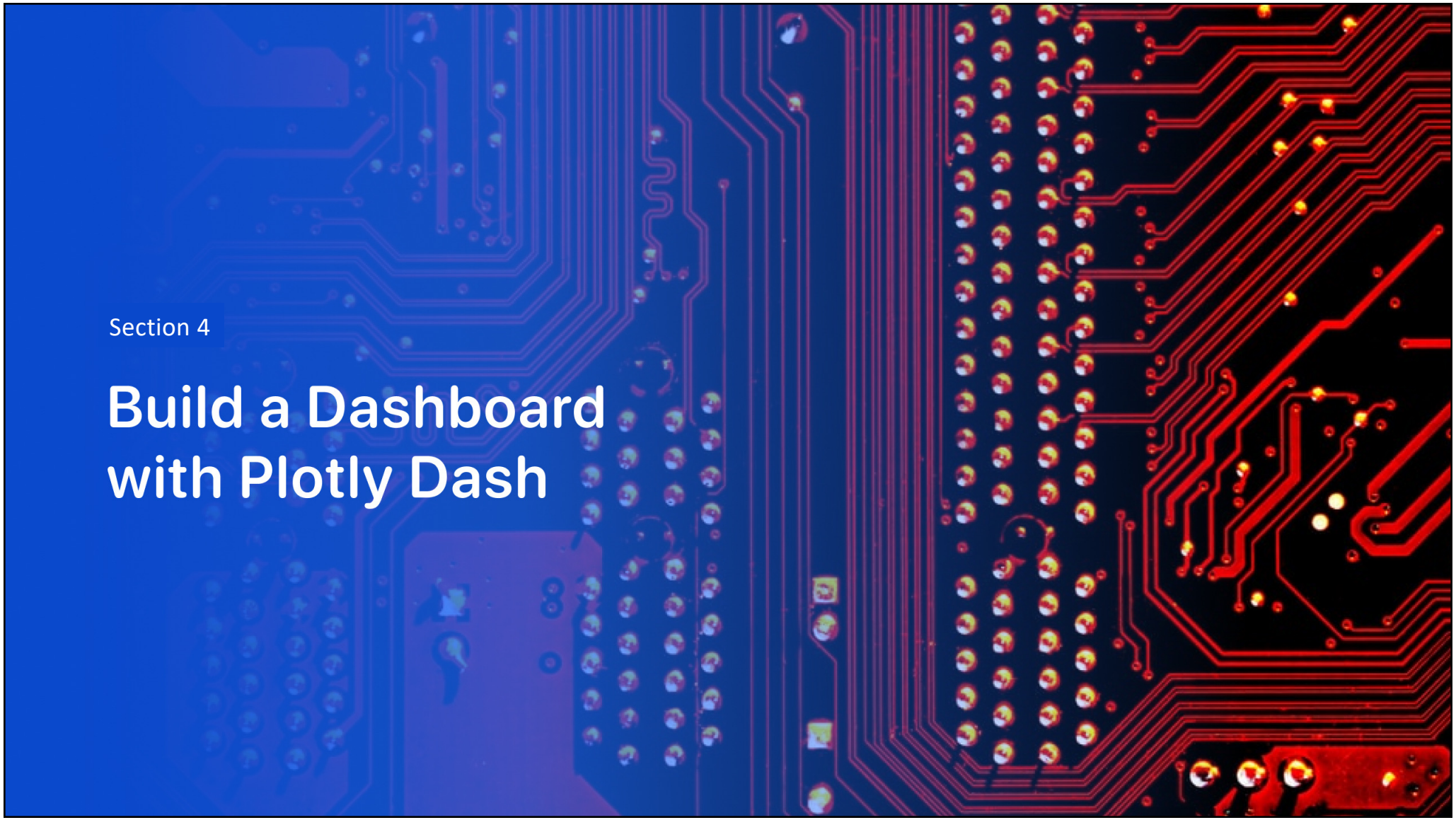
Figure: Launch outcomes at CCAFS LC-40



Folium Map: Proximity to Coast

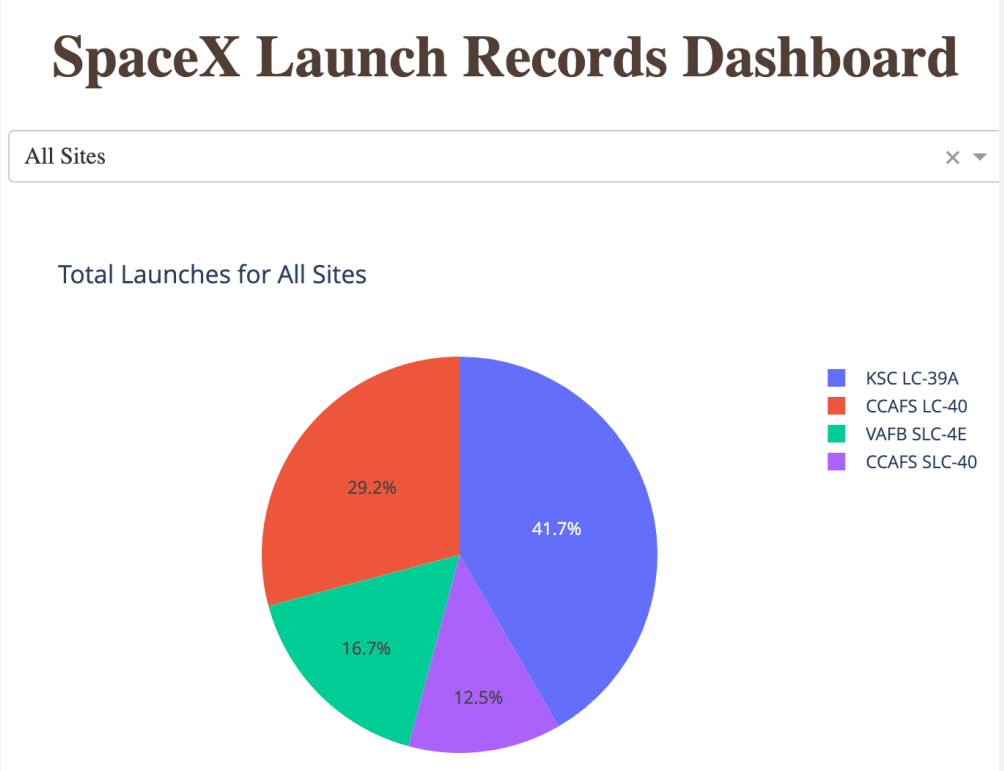
- The map to the right shows the distance from the CCAFS SLC-40 launch site to the coastline.
- Distance: 0.86km.

Figure: Distance from CCAFS SLC-40 to the Coast


Section 4
Build a Dashboard with Plotly Dash

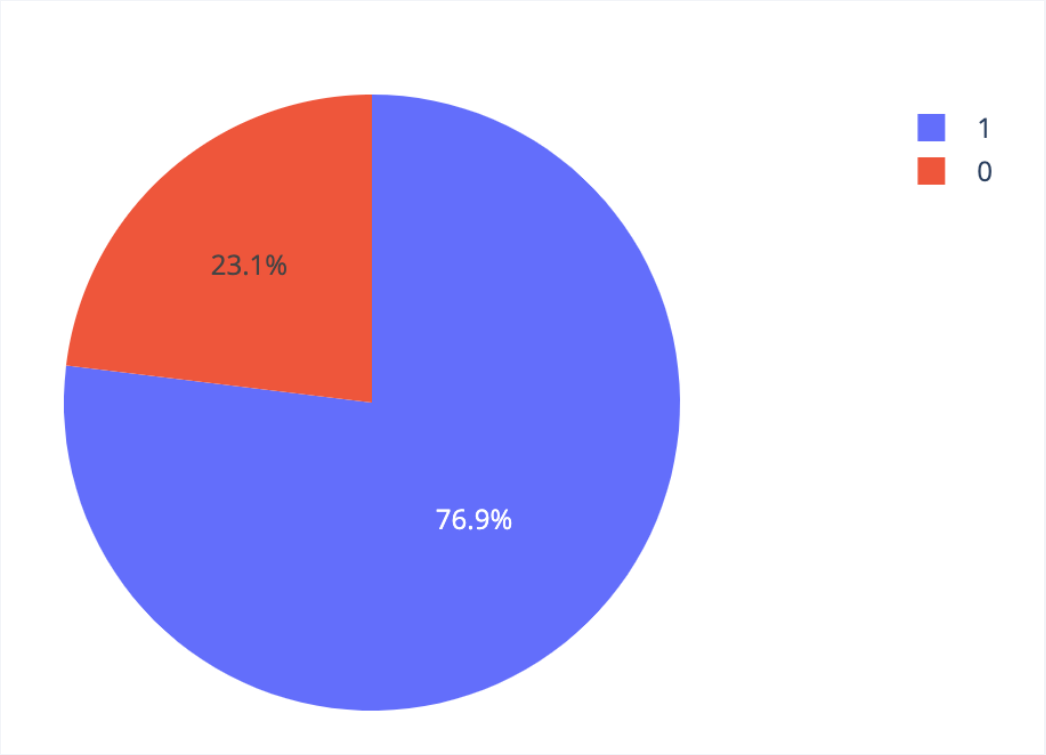
SpaceX Launch Records Dashboard: Launches by Site

- The Kennedy Space Center Launch Complex (KSC LC-39A) launch site accounts for the largest share of total launches.



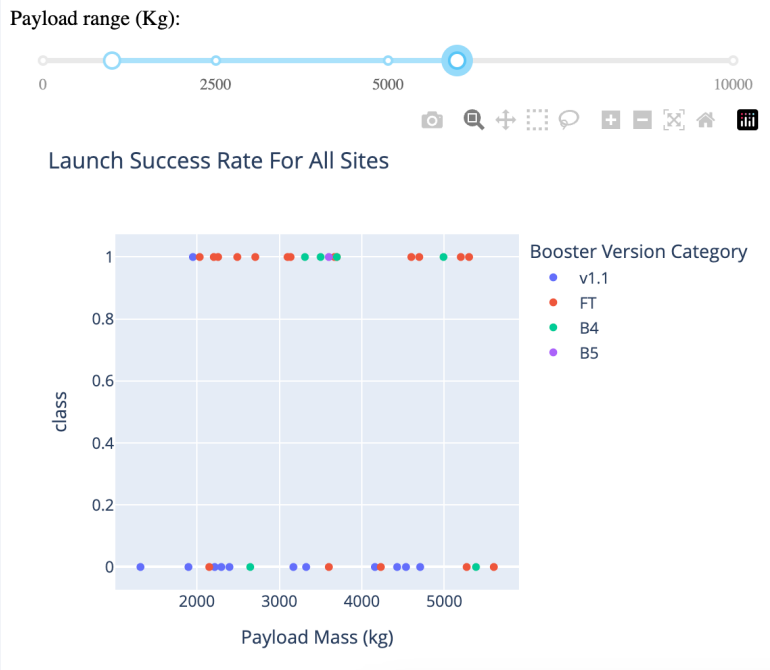
Highest Launch Success Rate: Kennedy Space Center

- The launch site with the highest success rate is the Kennedy Space Center Launch Complex (KSC LC-39A).
- At this site, 76.9% of all launches were successful.

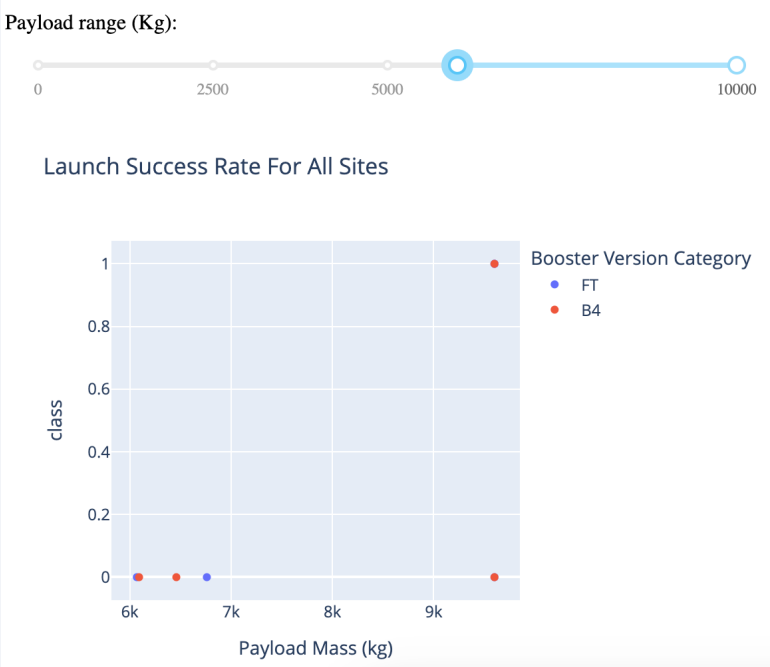


Launch Success Rates & Payload Mass

- The payload range with the highest success rate is between 2,000 and 6,000 Kg.

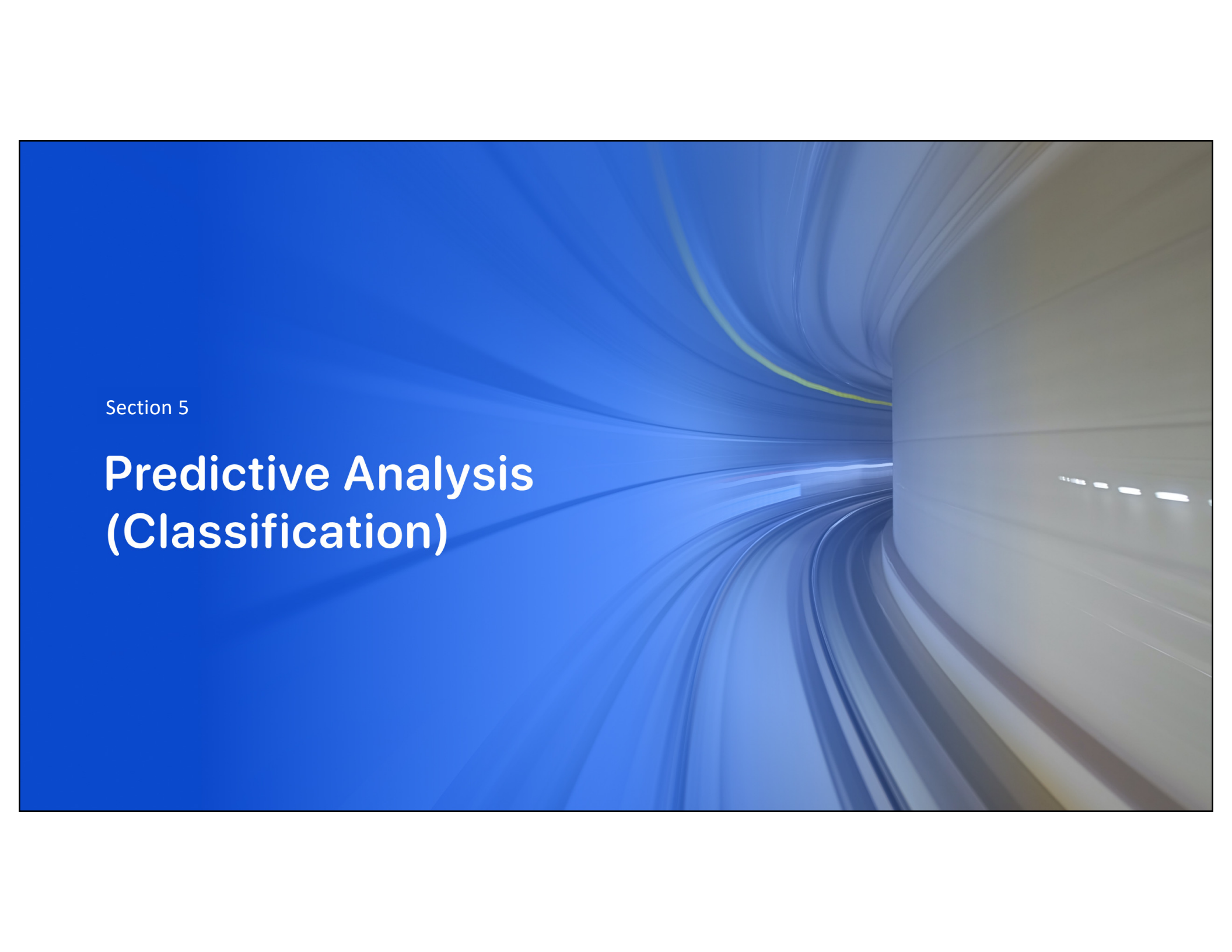


- At higher payload mass there are fewer successful launches.



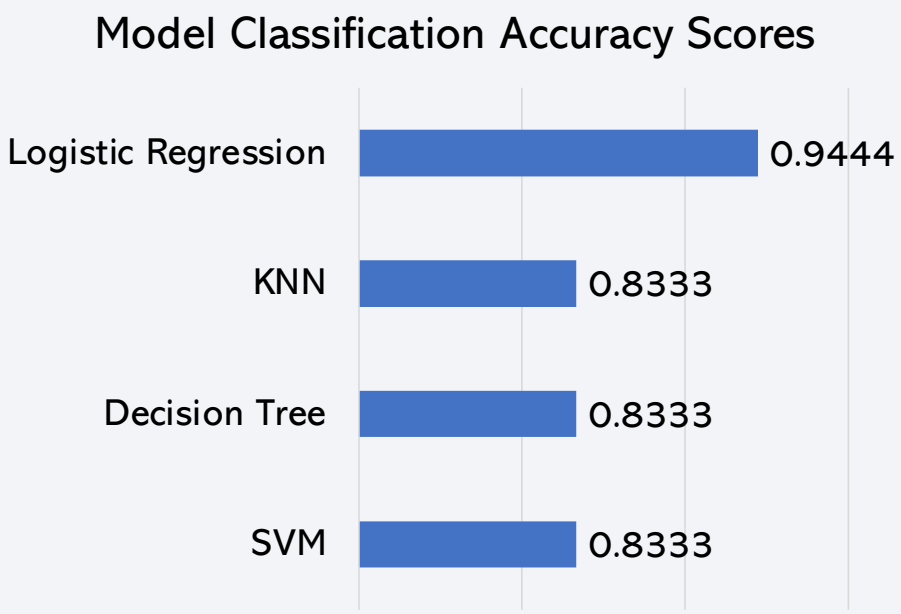
Section 5

Predictive Analysis (Classification)



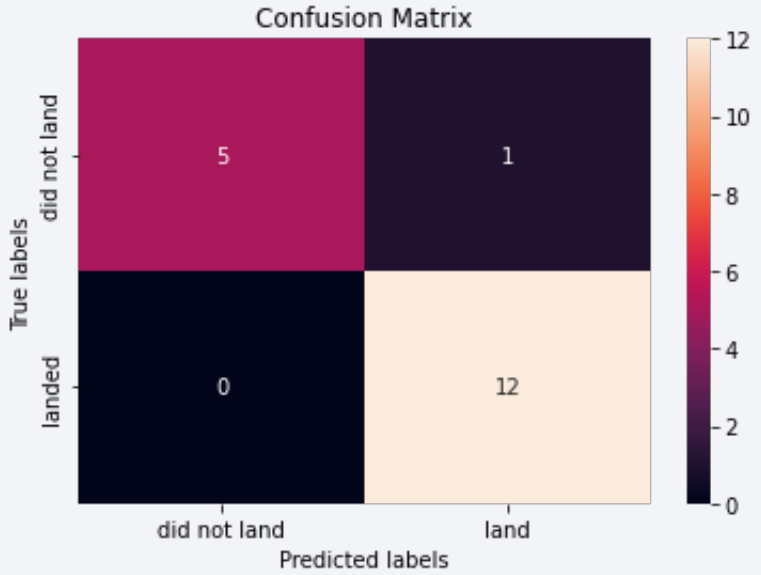
Classification Accuracy

- The model with the highest classification accuracy is Logistic Regression with an accuracy score of 0.9444.



Confusion Matrix: Logistic Regression

- Examining the confusion matrix, we see that logistic regression can distinguish between the different classes.
- True positives: 12
- True negatives: 5
- False positives: 1
- False negatives: 0



Conclusions

- The SVM, Decision Tree, and KNN models perform quite similarly:
  - All have an accuracy score of 0.8333.
  - All result in 3 false positives.

- The model with the highest classification accuracy is Logistic Regression:
  - Logistic Regression has an accuracy score of 0.9444.
  - Logistic Regression results in 0 false negatives and 1 false positive.

Appendix

EDA with SQL: Query Codes

Query	SQL Code
aExtract a list of distinct launch sites	%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTABLE
Display 5 records where launch sites begin with the string 'CCA'	%sql SELECT * FROM SPACEXTABLE WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
Display 5 records where launch sites begin with the string 'CCA'	%sql SELECT * FROM SPACEXTABLE WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
Display the total payload mass carried by boosters launched by NASA (CRS)	%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE CUSTOMER == "NASA (CRS)"
Display the average payload mass carried by booster version F9 v1.1	%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version == "F9 v1.1"
List the date when the first successful landing outcome in ground pad was achieved	%sql SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome == "Success (ground pad)"
List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000	%sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome == "Success (drone ship)" and PAYLOAD_MASS__KG_ BETWEEN 4001 and 5999;
List the total number of successful and failure mission outcomes	%sql SELECT Mission_Outcome, COUNT(*) as NumOutcomes \ FROM SPACEXTABLE \ GROUP BY Mission_Outcome;
List the names of the booster_versions which have carried the maximum payload mass	%sql SELECT Booster_Version \ FROM SPACEXTABLE \ WHERE PAYLOAD_MASS__KG_ == (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
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	%sql SELECT substr(Date, 6,2) as month, DATE, Landing_Outcome, Booster_Version, Launch_Site \ FROM SPACEXTABLE \ WHERE Landing_Outcome == "Failure (drone ship)" and substr(Date,0,5)='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	%sql SELECT Landing_Outcome, COUNT(Landing_Outcome) \ FROM SPACEXTABLE \ WHERE Date BETWEEN '2010-06-04' and '2017-03-20' \ GROUP BY Landing_Outcome \ ORDER BY COUNT(Landing_Outcome) DESC;

Thank you!