IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Tatum Hackler
6/27/2024

Outline

Executive Summary

Summary of methodologies

- Data collection
- Data wrangling
- Exploratory Data Analysis with SQL
- Exploratory Data Analysis with visualizations
- Building interactive maps with Folium
- Building interactive dashboard with Plotly Dash
- Predictive analysis (classification)

Summary of all results

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

Introduction

Project background and context

SpaceX is widely known as a leader in the commercial space industry. A hypothetical rival company would like the compete with SpaceX. SpaceX advertises Falcon 9 rocket launches on its website with a cost of $62 million; other known providers cost upwards of $165 million. The savings can be primarily attributed to SpaceX's reuse of the first stage. Therefore, if we can determine whether the first stage will land, we can determine the cost of the launch.

Questions to be answered

- Which factors determine if the rocket will land successfully?
- What is the best machine learning model to predict the outcome of Falcon 9 landings?
- Will a future Falcon 9 landing be successful? What will be the cost of the launch given the predicted outcome?

Section 1

Methodology

Methodology

Executive Summary

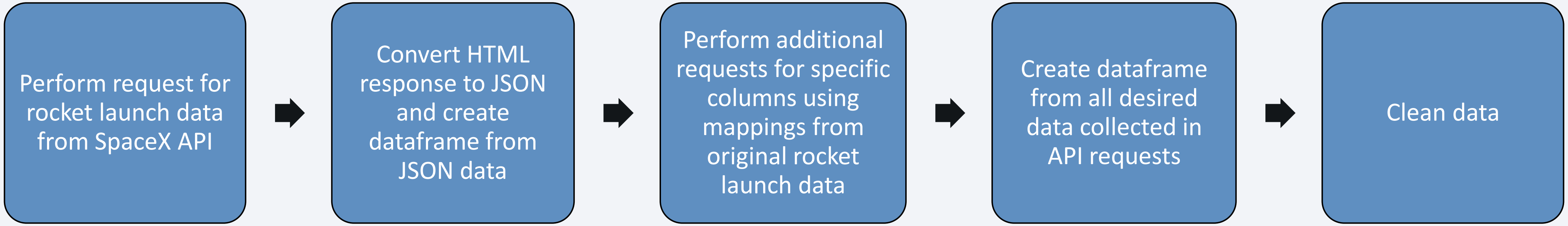
- Collect data from SpaceX via SpaceX REST API and Wikipedia via web scraping
- Clean and shape data for use in data exploration, visualizations, and applying machine learning models
- Perform exploratory data analysis (EDA) using visualization and SQL
- Create interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

- Data was collected both directly from SpaceX using SpaceX's REST API and from SpaceX Falcon 9 Launches Wikipedia page with tables containing launch records using web scraping
- Data columns obtained from SpaceX using SpaceX's REST API:
  - FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- Data columns obtained from Wikipedia table using web scraping:
  - Flight No., Launch site, Payload, Payload mass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API

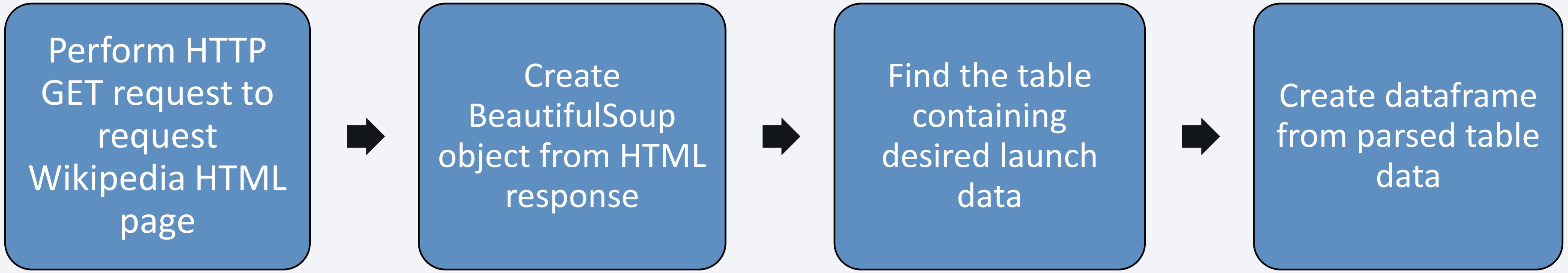
- Collected data directly from SpaceX using SpaceX's API
- Multiple API calls were required to obtain all desired columns by using mappings between composite launch data and data specifically for rockets, launchpads, payloads, and cores
- Performed data cleaning on collected data
  - Detailed on data wrangling slide (slide #10)



- GitHub URL: https://github.com/hacklert/spaceX/blob/master/data-collection-api.ipynb

Data Collection - Scraping

- Performed web scraping to extract data from Wikipedia page that has tables containing Falcon 9 launch data



- GitHub URL: https://github.com/hacklert/spaceX/blob/master/data-collection-webscraping.ipynb

Data Wrangling

- Cleaned data collected using SpaceX API
  - Filtered out all non-Falcon 9 launches
  - Handled missing values for payload mass by replacing them with the mean
- Performed EDA on collected data to see the number of launches for each launch site, orbit type, and landing outcome type
- Created new landing outcome column ‘Class’ from the Outcome column to be used as the target for classification
  - The Class column contains a 1 if the outcome was a success or a 0 if the outcome was a failure

Clean data collected via SpaceX API → Perform EDA on data to gain insight about distribution of launches for each site, orbit, and mission outcome → Create landing outcome label from ‘Outcome’ column that can be used for classification

- GitHub URL: https://github.com/hacklert/spaceX/blob/master/data-wrangling.ipynb

EDA with Data Visualization

- Visualizing Launch Site trends
  - Scatterplot of landing outcomes with respect to flight number and launch site
  - Scatterplot of landing outcomes with respect to payload mass and launch site
- Visualizing Orbit Type trends
  - Bar chart of landing outcomes with respect to orbit type
  - Scatterplot of landing outcomes with respect to flight number and orbit type
  - Scatterplot of landing outcomes with respect to payload mass and orbit type
- Visualizing time-based trends
  - Line plot of landing outcome success rate by year

- GitHub URL: https://github.com/hacklert/spaceX/blob/master/eda-visualizations.ipynb

EDA with SQL

- SQL queries performed:
  - Display names of unique launch sites
  - 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
  - Display the date when the first successful landing outcome in ground pad was achieved
  - Display the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000 kg
  - Display the total number of successful and failure mission outcomes
  - Display the names of the booster versions which have carried the maximum payload mass
  - Display the failure landing outcomes in drone ship, booster versions, launch site for the months in year 2015
  - Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20
- GitHub URL: https://github.com/hacklert/spaceX/blob/master/eda-sql.ipynb

Build an Interactive Map with Folium

- Created interactive maps with Folium to display:
  - Markers of all launch sites
    - Used marker with circle, popup label, and text label for each launch site to show their geographical locations and identify general proximity to coasts and the Equator
  - Colored markers of landing outcomes for each launch site
    - Added colored markers (green for success, red for failure) using marker clusters to represent landing outcomes to identify which launch sites have higher success rates
  - Distances between a launch site and its proximities
    - Used lines labeled with distances to show the distance between launch site VAFB SLC-4E to the nearest coastline, railway, highway, and city

- GitHub URL: https://github.com/hacklert/spaceX/blob/master/interactive_visualizations_folium.ipynb

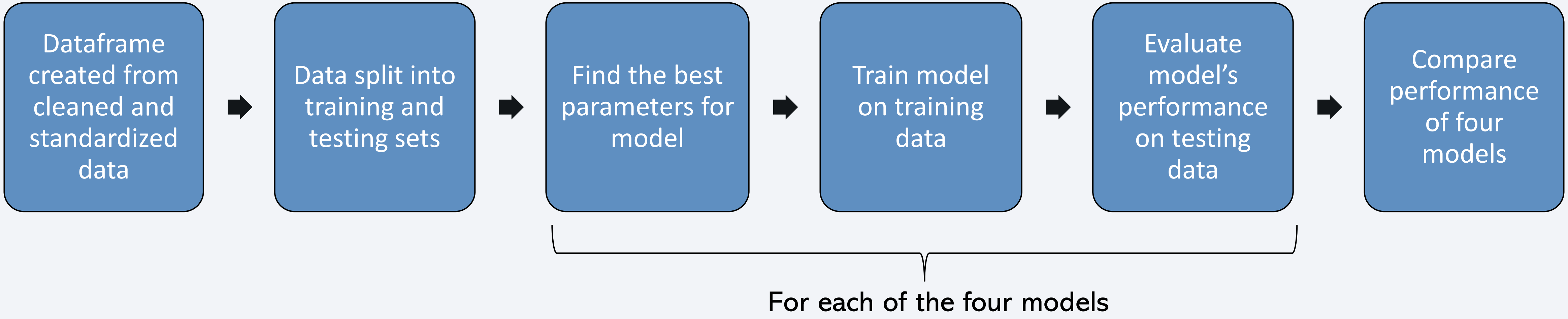
Build a Dashboard with Plotly Dash

- Created dashboard with Plotly Dash to display:
  - Pie chart showing distribution of successful landing outcomes for all launch sites
  - Pie chart showing success / failure landing outcome ratio for a particular launch site
  - Scatterplot showing payload mass vs. landing outcome for various booster versions for all launch sites
  - Scatterplot showing payload mass vs. landing outcome for various booster versions for a particular launch site
- Interactions available on dashboard:
  - A dropdown allows user to select a particular launch site to display graphs for or 'All Sites' to display graphs that cover all sites
  - A slider allows user to select a payload mass range to use for the scatterplot

- GitHub URL: https://github.com/hacklert/spaceX/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)

- Data was split into training set (80%) and testing set (20%)
- Four machine learning classification models were trained and evaluated: logistic regression, Support Vector Machine (SVM), decision tree, and k-nearest neighbors
- The best hyperparameters were selected for each respective model using grid search (GridSearchCV)
- Each model's performance was evaluated by calculating the accuracy score



- GitHub URL: https://github.com/hacklert/spaceX/blob/master/classification-predictive-analysis.ipynb

Results

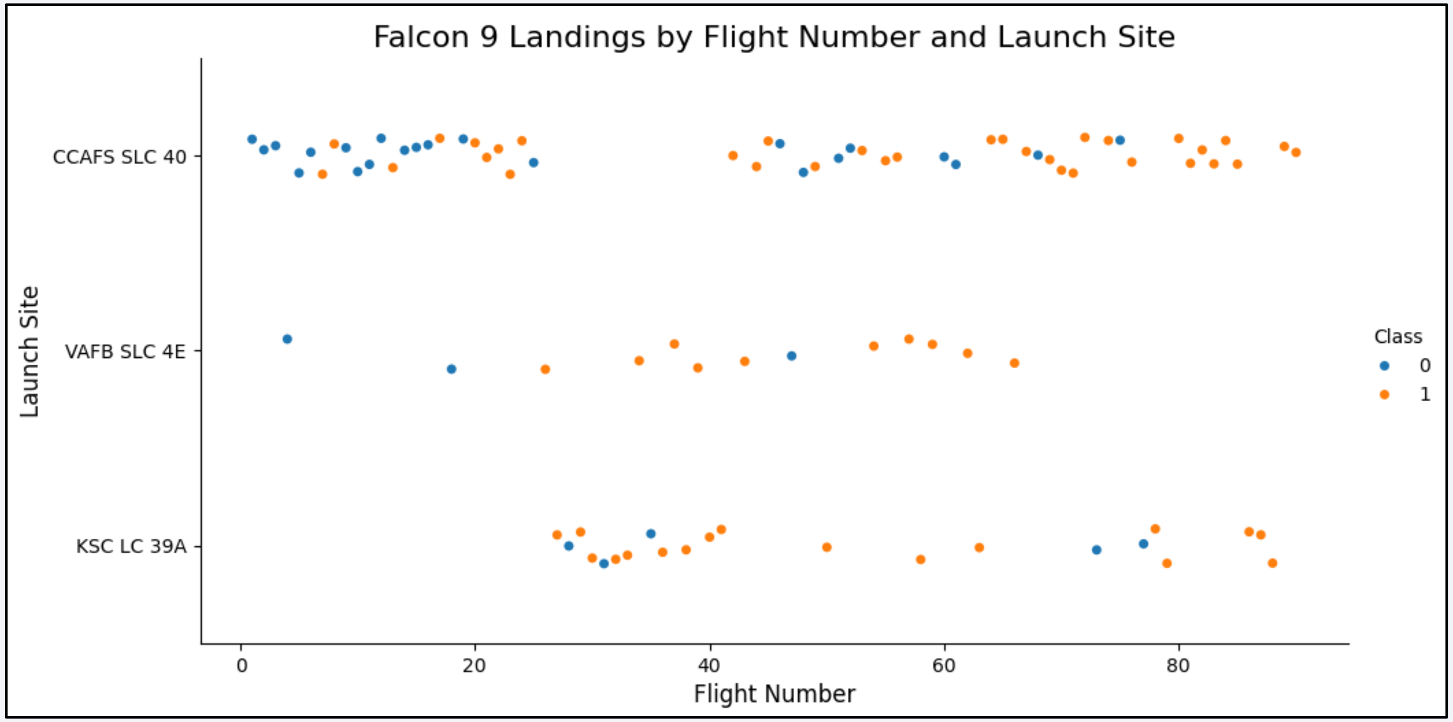
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

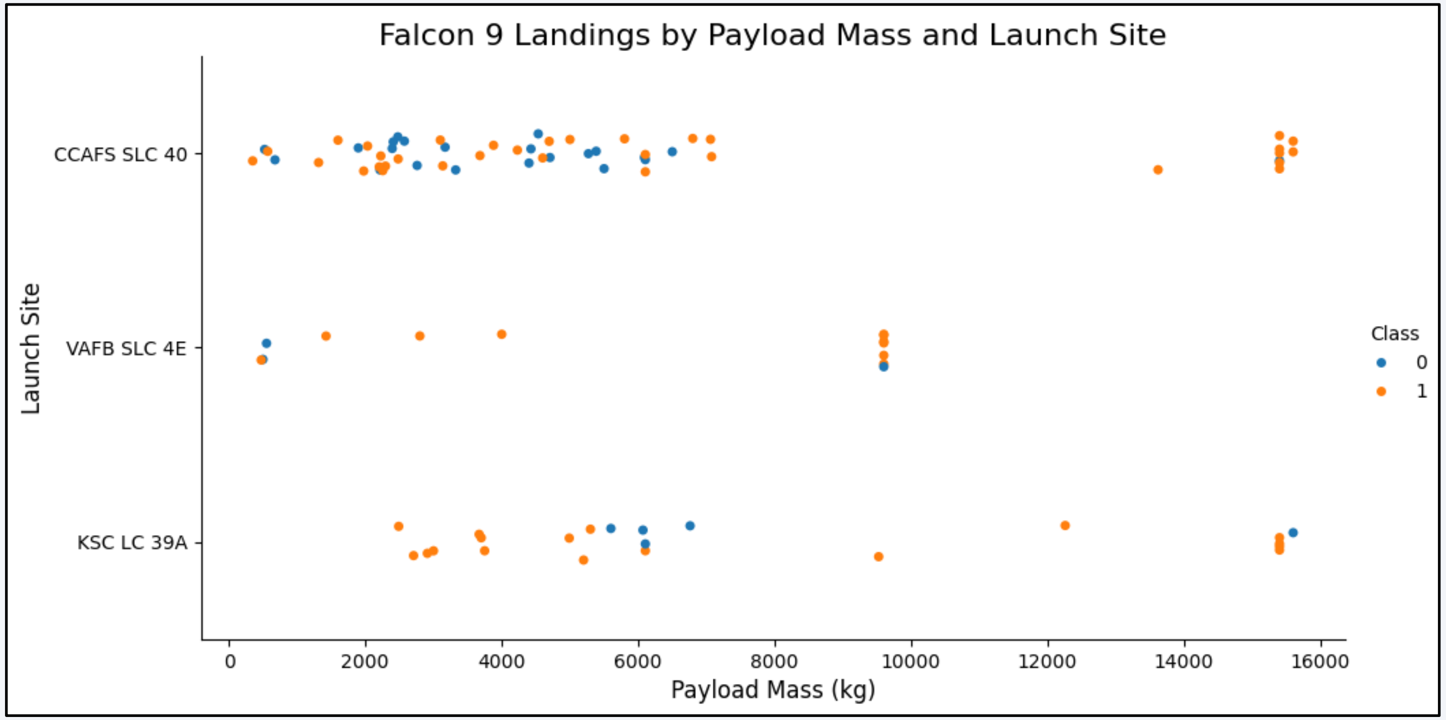
Flight Number vs. Launch Site

- Most of the launches occurred at CCAFS SLC 40
- Success rate appears to vary between the launch sites; VAFB SLC 4E and KSC LC 39A have a higher success rate than CCAFS SLC 40
- Success rate appears to increase across all launch sites as the flight number increases



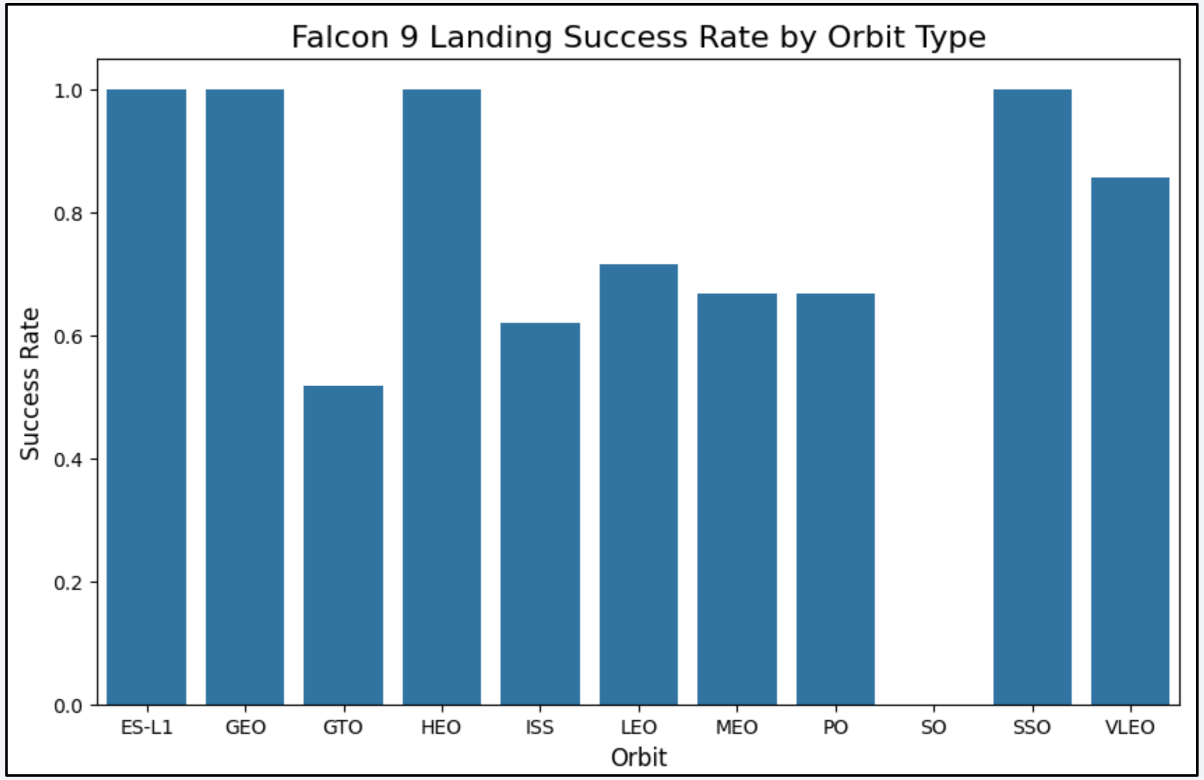
Payload vs. Launch Site

- Only a small number of launches had a payload mass over 10,000 kg, with 0 heavy payload mass launches out of VAFB SALC 4E
- At CCAFS SLC 40, there is no clear relationship between payload mass and success rate for payloads less than 8,000 kg
- Success rate appears to increase across all launch sites as the payload mass increases



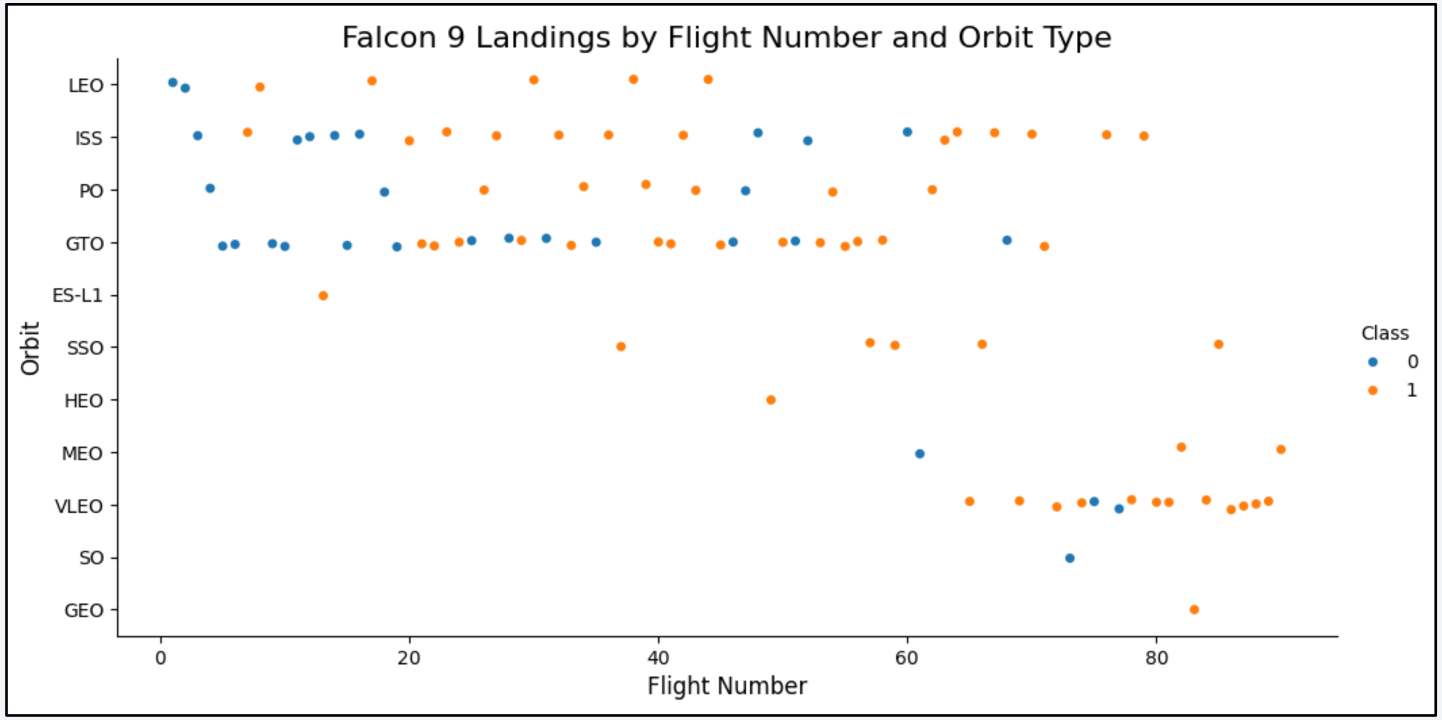
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO orbits have no failed landings
- SO orbit has no successful landings
- All other orbits have at least 50% success rate



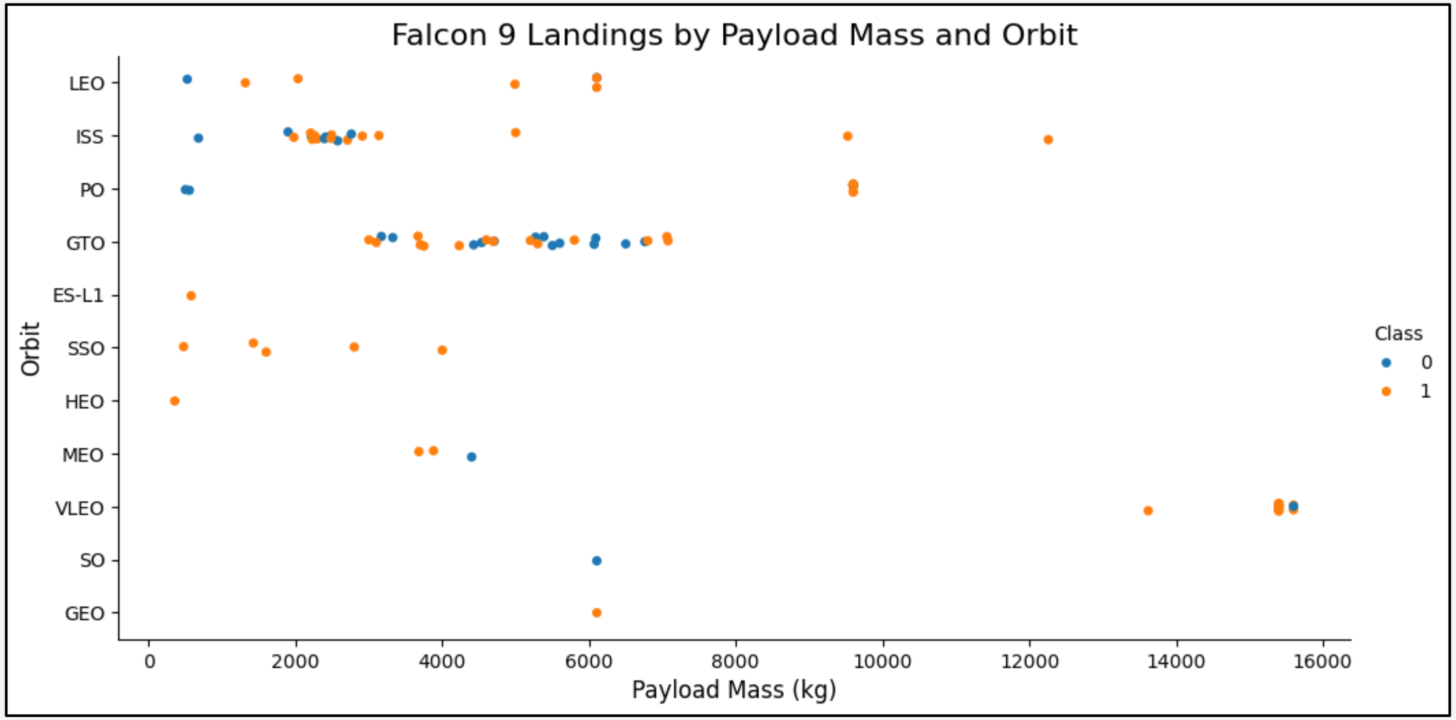
Flight Number vs. Orbit Type

- There is a relationship between flight number and success rate in LEO orbit; success rate appears to increase with flight number
- There is no clear relationship between flight number and success rate in GTO orbit



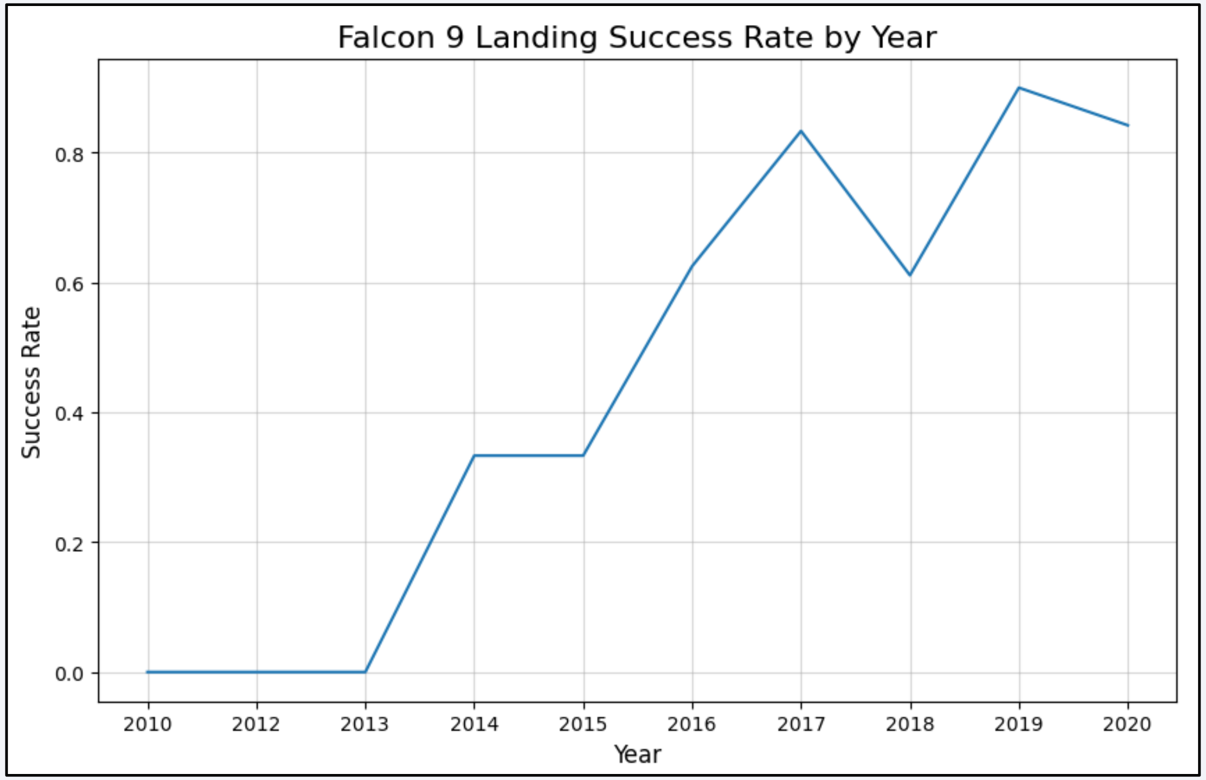
Payload vs. Orbit Type

- Success rate appears to increase with payload mass in LEO, ISS, and PO orbits
- There is no clear relationship between payload mass and success rate in GTO orbit



Launch Success Yearly Trend

- Success rate continued to increase between 2013 and 2020



All Launch Site Names

- Displaying names of each unique launch site
- Query: SELECT DISTINCT Launch_Site FROM SPACEXTABLE

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

Launch Site Names Begin with 'CCA'

- Get records for launch sites where name begins with ‘CCA’ and display the first five
- Query: SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5

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

- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Query: SELECT sum(payload_mass__kg_) AS Total_Payload_Mass_kg FROM SPACEXTABLE WHERE Customer='NASA (CRS)’

Total_Payload_Mass_kg
45596

- The total payload mass for boosters launched by NASA (CRS) is 45596 kg

Average Payload Mass by F9 v1.1

- Displaying the average payload mass carried by booster version F9 v1.1
- Query: SELECT avg(payload_mass__kg_) AS Avg_Payload_Mass_kg FROM SPACEXTABLE WHERE Booster_Version='F9 v1.1'

Avg_Payload_Mass_kg
2928.4

- The average payload mass for booster F9 v1.1 is 2928.4 kg

First Successful Ground Landing Date

- Getting the date when the first successful landing outcome in ground pad was achieved
- Query: SELECT min(Date) AS Date FROM SPACEXTABLE WHERE Landing_Outcome='Success (ground pad)'

Date
2015-12-22

- The first successful ground landing took place on December 12, 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

- Listing the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- Query: SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome='Success (drone ship)' AND payload_mass__kg_ BETWEEN 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

- Getting the total number of successful and failure mission outcomes
- Query: SELECT Mission_Outcome, count(*) AS ‘Count’ FROM SPACEXTABLE GROUP BY Mission_Outcome

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

- There are 100 successful missions and 1 failure in the dataset

Boosters Carried Maximum Payload

- Listing the names of the boosters which have carried the maximum payload mass
- Query: SELECT Booster_Version FROM SPACEXTABLE WHERE payload_mass__kg_ = (SELECT max(payload_mass__kg_) FROM SPACEXTABLE)

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

- Listing the failed landing outcomes in drone ship, their booster versions, and launch site names from 2015
- Query: SELECT substr(Date, 6, 2) AS Month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE Landing_Outcome='Failure (drone ship)' AND substr(Date, 0, 5)='2015'

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

- There were two failed landing outcomes in the drone ship in 2015: one in January and one in April

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

- Ranking the count of each unique landing outcome between 2010-06-04 and 2017-03-20, in descending order
- Query: SELECT Landing_Outcome, count(*) AS 'Count' FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY count(*) DESC

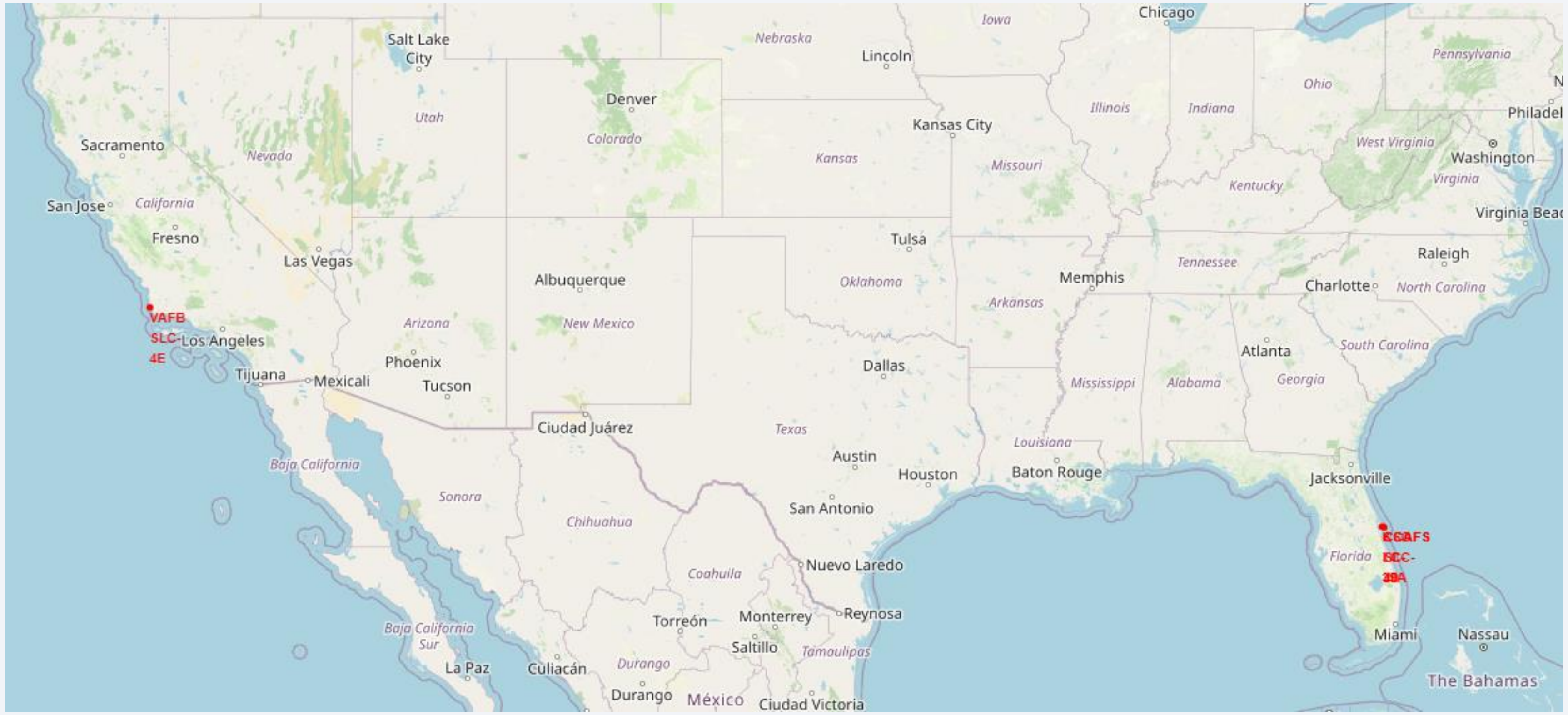
Landing_Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

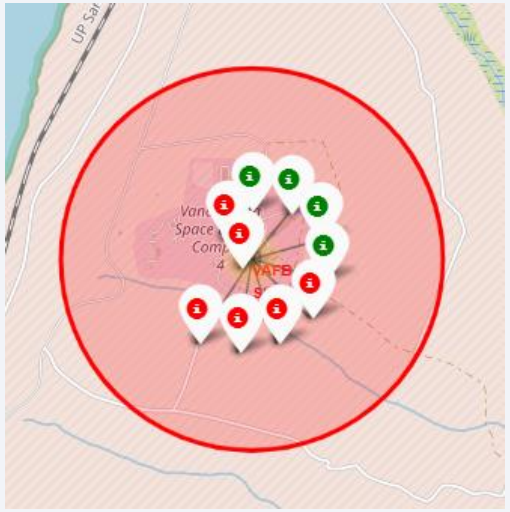
Falcon 9 Launch Sites

- Map with markers for all the Falcon 9 launch sites
- All launch sites are in proximity to the coast



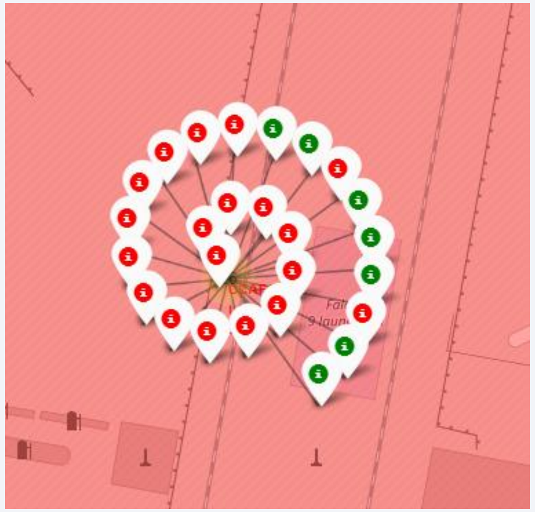
Map Markers for Success / Failure Landings

- The launch sites are displayed below with markers for each landing; green markers indicate successful landings and red markers indicate failed landings
- The markers make it easier to compare success rate between the launch sites
- KSC LC-39A has the highest overall success rate and CCAFS LC-40 has the lowest overall success rate

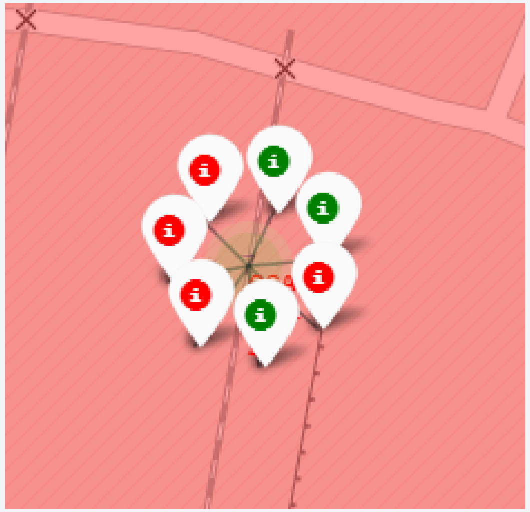


VAFB SLC-4E

KSC LC-39A



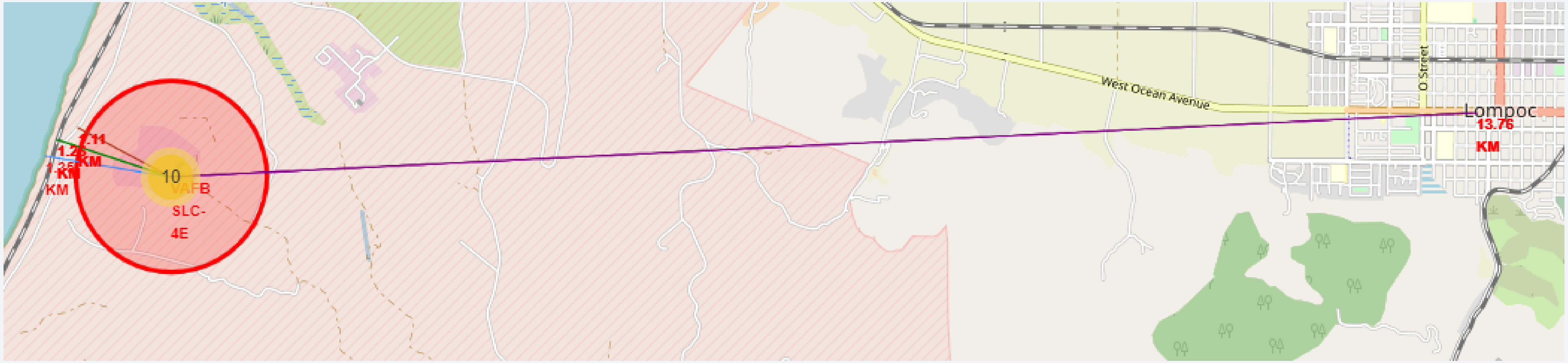
CCAFS LC-40



CCAFS SLC-40

Proximities for VAFB SLC-4E

- Launch site VAFB SLC-4E is shown in the screenshot with nearest coastline (blue line), railway (green line), highway (brown), and city (purple)
- Launch sites are generally in close proximity to a coastline, railway, and highway; this is likely to facilitate controlled water landings, transportation to / from the site for employees and raw material suppliers
- Launch sites are generally a distance away from the nearest city, with VAFB SLC-4E having the closest city across all launch sites at 13.76 km away; this is likely to avoid potential damage to the city or injuries in the event of rocket launch or landing failures

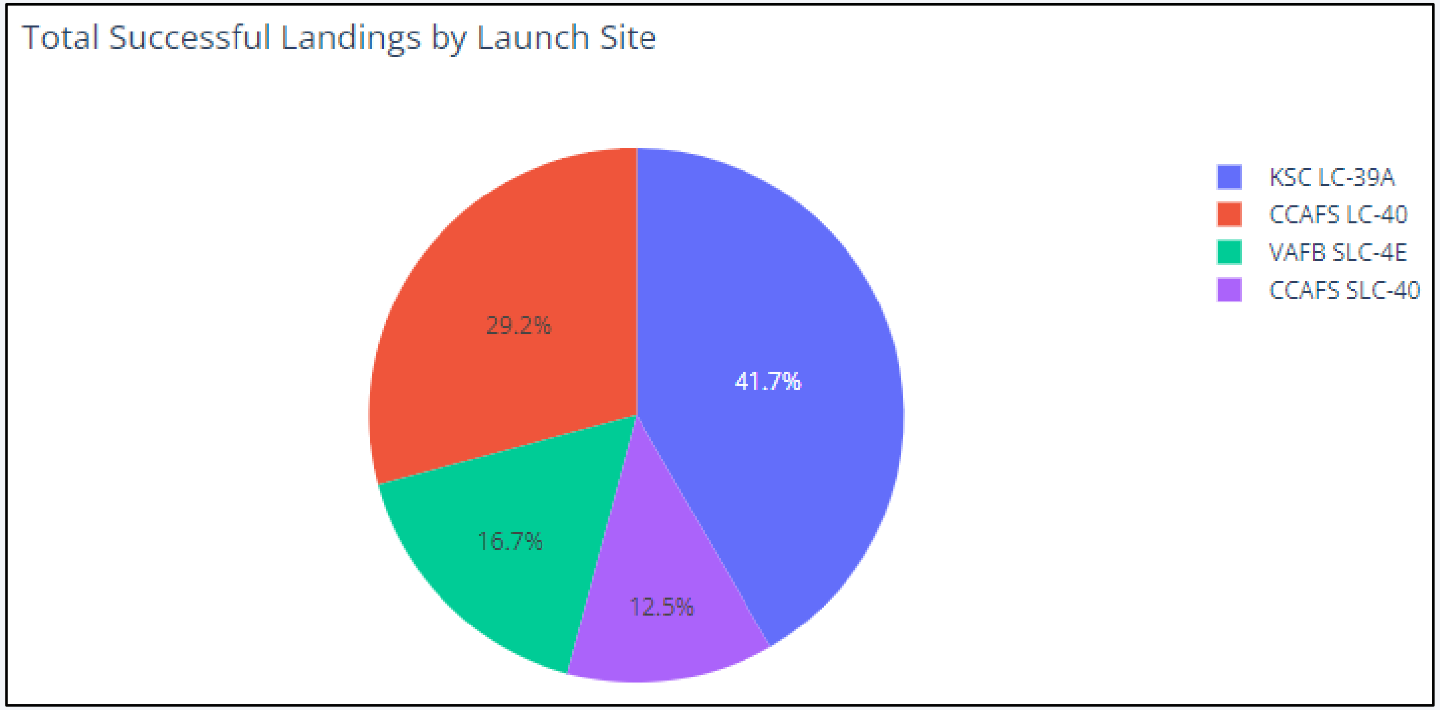


Section 4

Build a Dashboard with Plotly Dash

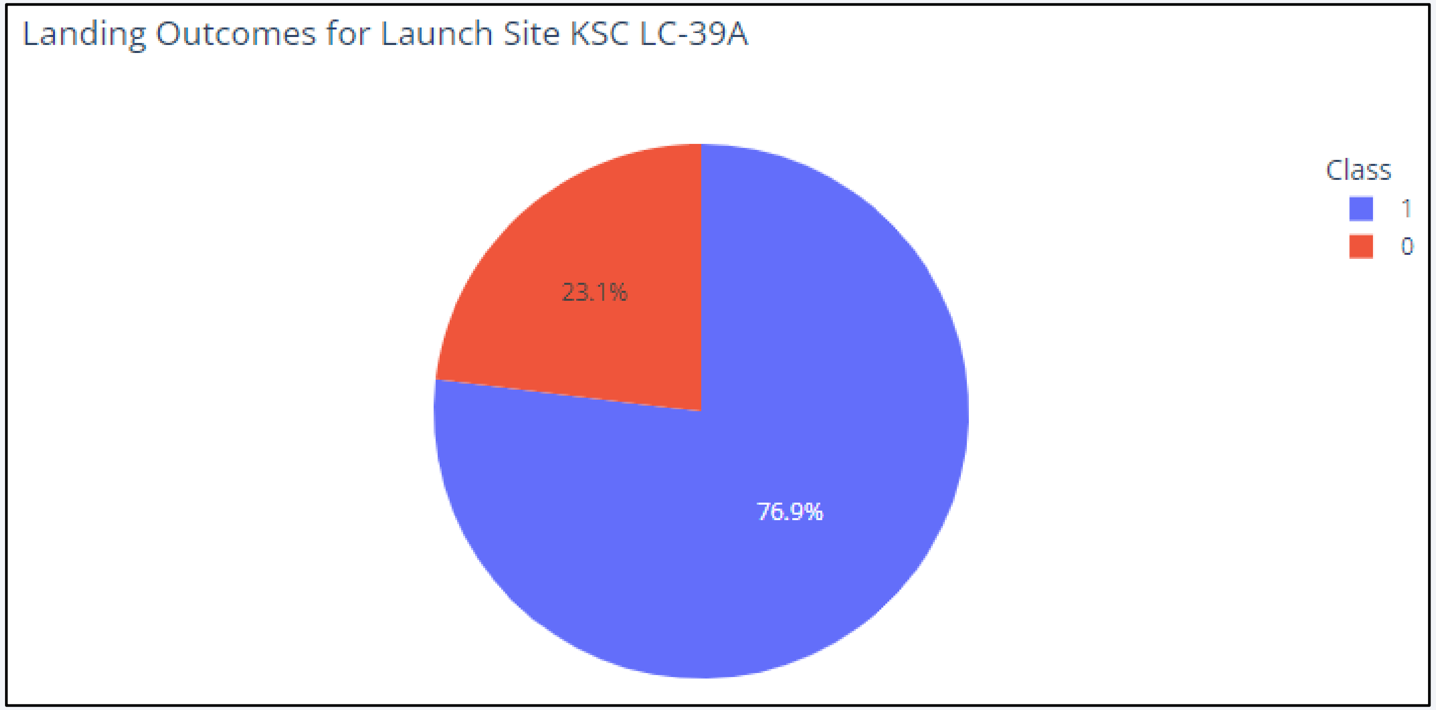
Landing Success Count for All Launch Sites

- The pie chart displays the distribution of successful landings between all the launch sites
- Most of the successful landings occurred at KSC LC-39A



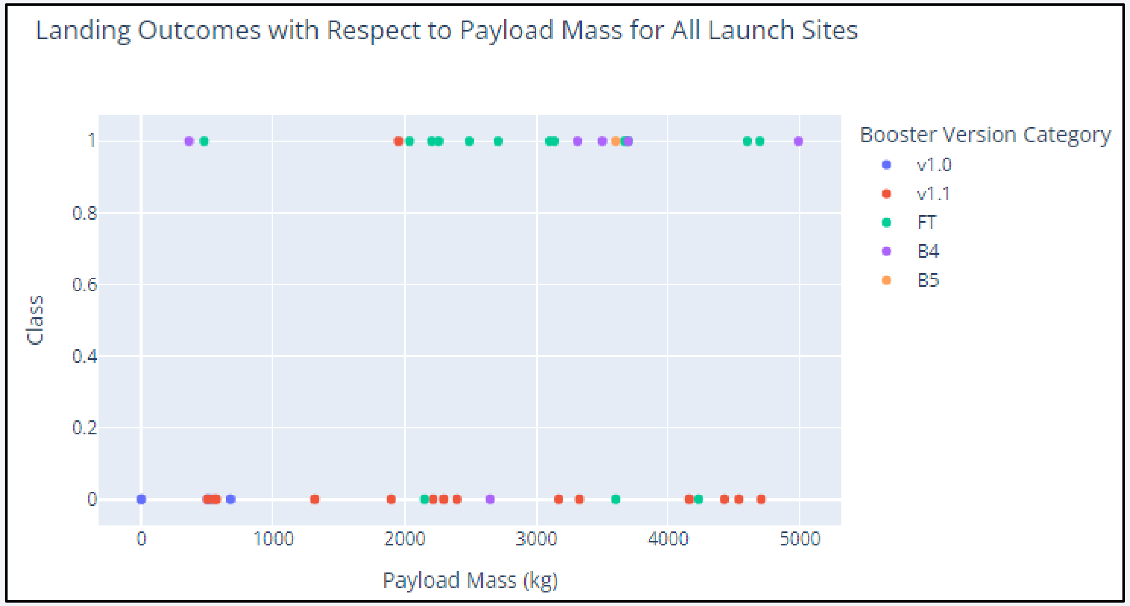
Ratio of Landing Outcomes at KSC LC-39A

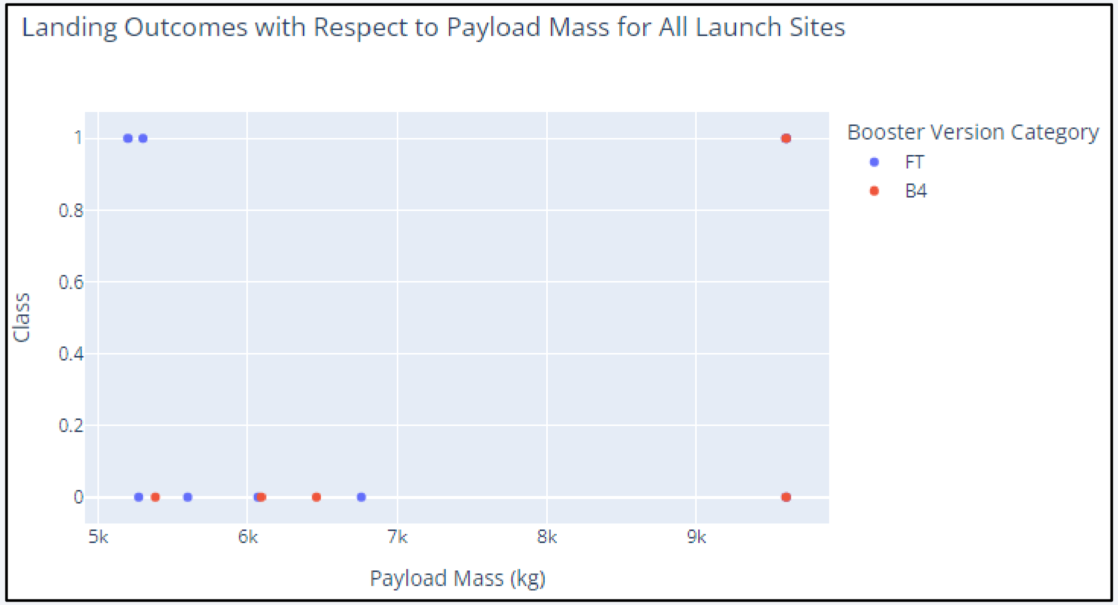
- The pie chart displays the ratio of landing outcomes at the launch site KSC LC-39A
- KSC LC-39A had the highest success ratio of all the launch sites, with a 76.9% success rate



Payload vs. Launch Outcome for All Launch Sites

- The scatterplots display payload mass vs. launch outcome for all launch sites, with a payload range of 0 to 5,000 kg on the left and 5,000 to 10,000 kg on the right
- Smaller payload launches appear to have a higher success rate
- Booster FT appears to have the highest success rate between booster types



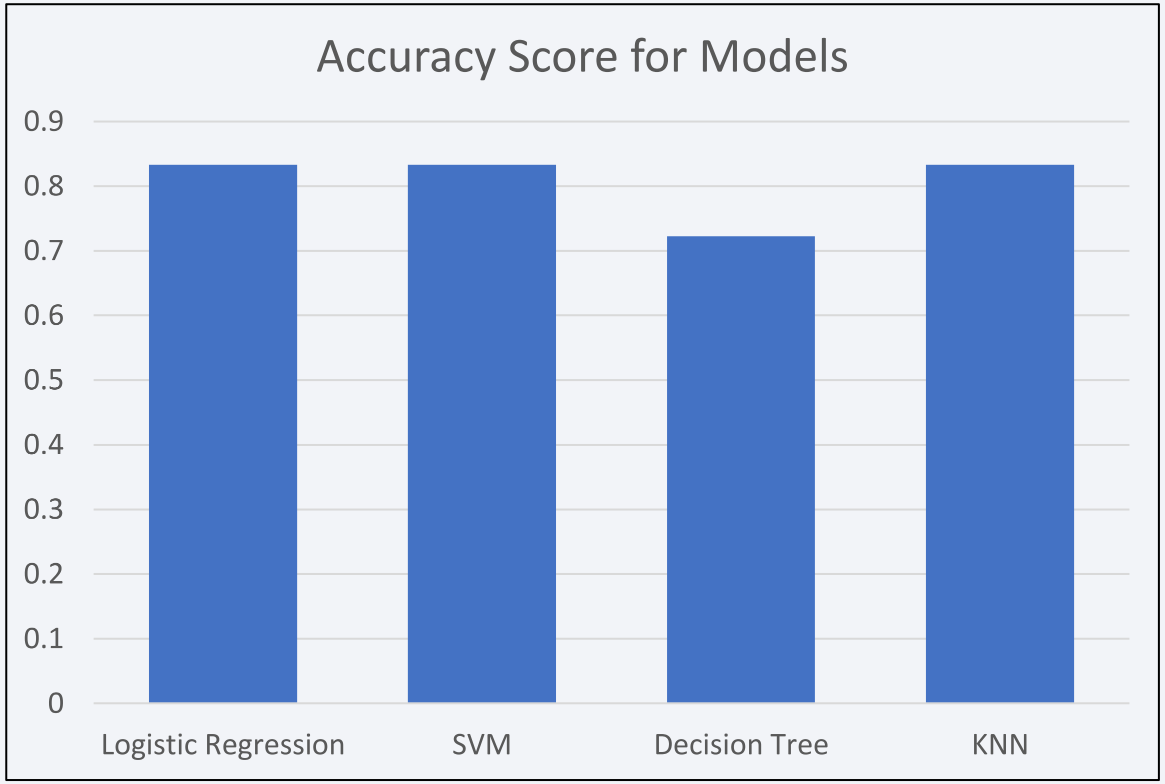


Section 5

Predictive Analysis (Classification)

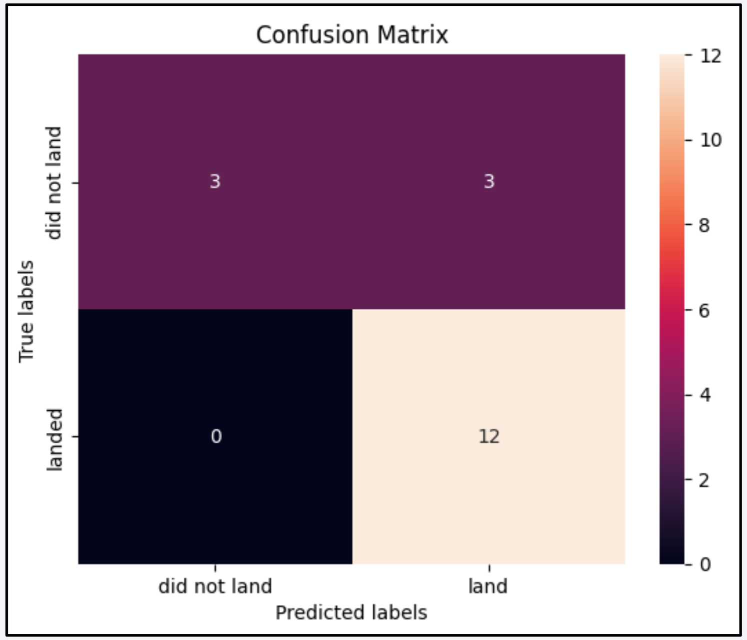
Classification Accuracy

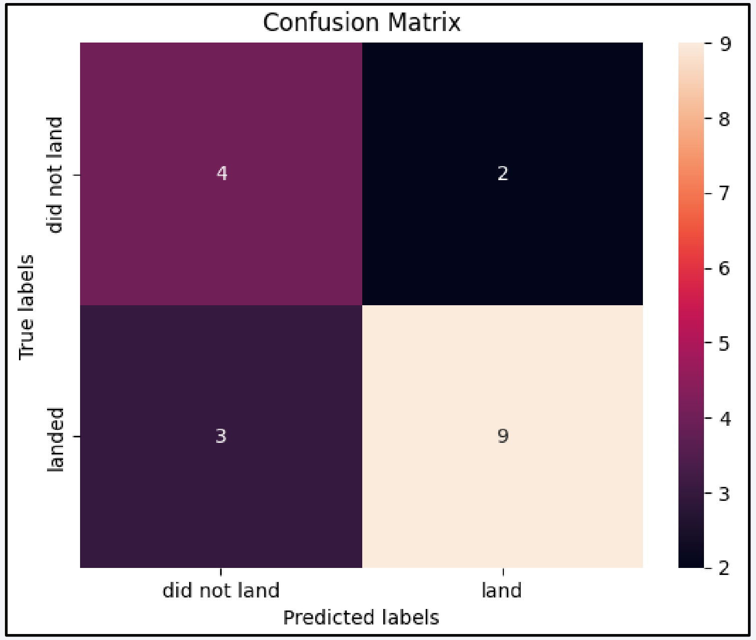
- The four models were evaluated using the accuracy score
- Logistic regression, SVM, and KNN models performed the same with an accuracy score of 0.83
- Decision tree performed slightly worse with an accuracy score of 0.72



Confusion Matrix

- The confusion matrix for the three models that performed the same is shown on the left
  - The weak point of these models is predicting false positives (predicting that landing is successful when it was not)
  - The models correctly predicted 100% of successful landings and only 50% of failed landings
- The confusion matrix for the decision tree model is shown on the right
  - The model struggled with both false positives and false negatives
  - The model was only able to correctly predict 75% of successful landings and 67% of failed landings





Conclusions

- Flight number, launch site, orbit type, and payload mass are significant factors in helping to predict the landing outcome of a Falcon 9 launch; insights gained through EDA suggest that:
  - Success rate increases with flight number (greater flight number indicates more launches that have taken place)
  - Success rate varies significantly between launch sites
  - Success rate varies significantly between orbit types
  - Success rate increases with payload mass
- Various classification machine learning models can be used to predict the landing outcome with high accuracy
  - With optimally tuned hyperparameters, logistic regression, SVM, and KNN models are equally effective at predicting landing outcomes (83.3% accurate), while a decision tree model performs slightly worse (72.2% accurate)

Appendix

- Provided source datasets:
  - SpaceX API static JSON: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json
  - Wikipedia page: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
  - Data used for EDA with SQL: https://github.com/hacklert/spaceX/blob/master/Spacex.csv
  - Geo data used for Folium maps: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/spacex_launch_geo.csv
  - Data used for Plotly Dash dashboard: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/spacex_launch_dash.csv
- Produced datasets:
  - SpaceX API data collection: https://github.com/hacklert/spaceX/blob/master/dataset_part_1.csv
  - Web scraping data collection: https://github.com/hacklert/spaceX/blob/master/spacex_web_scraped.csv
  - Data wrangling: https://github.com/hacklert/spaceX/blob/master/dataset_part_2.csv
  - EDA / feature engineering: https://github.com/hacklert/spaceX/blob/master/dataset_part__3.csv

Thank you!