IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Rainer Aue
18. Dec. 2024

Outline

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

R. Aue

Executive Summary

- This study aims for determing the most relevant factors for successful SpaceX rocket launches, which have impact on SpaceX's reaching its ecomical goals.
- Data on SpaceX Falcon9 launches was gathered, cleaned and analyzed in order to identify what factors relate to successful mission outcomes as ....
  - Identifing key factors: payload mass, orbit, launch site, flight number, and booster version and
  - Finding patterns based on creation of interactive maps and dashboards.
  - Different modeling methods were tested by applying a variety of parameters to identify the best suited model and parameter configuration i.e. the most accurate results:
  - Logistic Regression, Support Vector Machine, Decision Tree and K-Nearest-Neighbour
- Decision Tree Classification produced the most accurate model (accuracy score ≈ 94%)
- Analysis code, datasets used and documentation can be found on GITHUB.

Introduction

- SpaceX provides historical data about Falcon 9 rocket launches starting from the year 2014 [1].
- SpaceX is able to reduce the cost of a rocket launch by landing the first stage of their Falcon 9 rocket so it can be re-deployed in future missions.
- By applying data science methodologies, launch data can be analyzed and predictions be made whether the first stage can be recovered.
  - Data can be analyzed using the flexible and broadly available programming *Python* platform, which has many built in data science tools and can interface with many other software platforms and languages.
- Main deliverable of this project is a comprehensive predictive analysis using supervised learning models that can predict whether the first stage of a Falcon 9 rocket will land given the various launch conditions.

Section 1

Methodology

Methodology

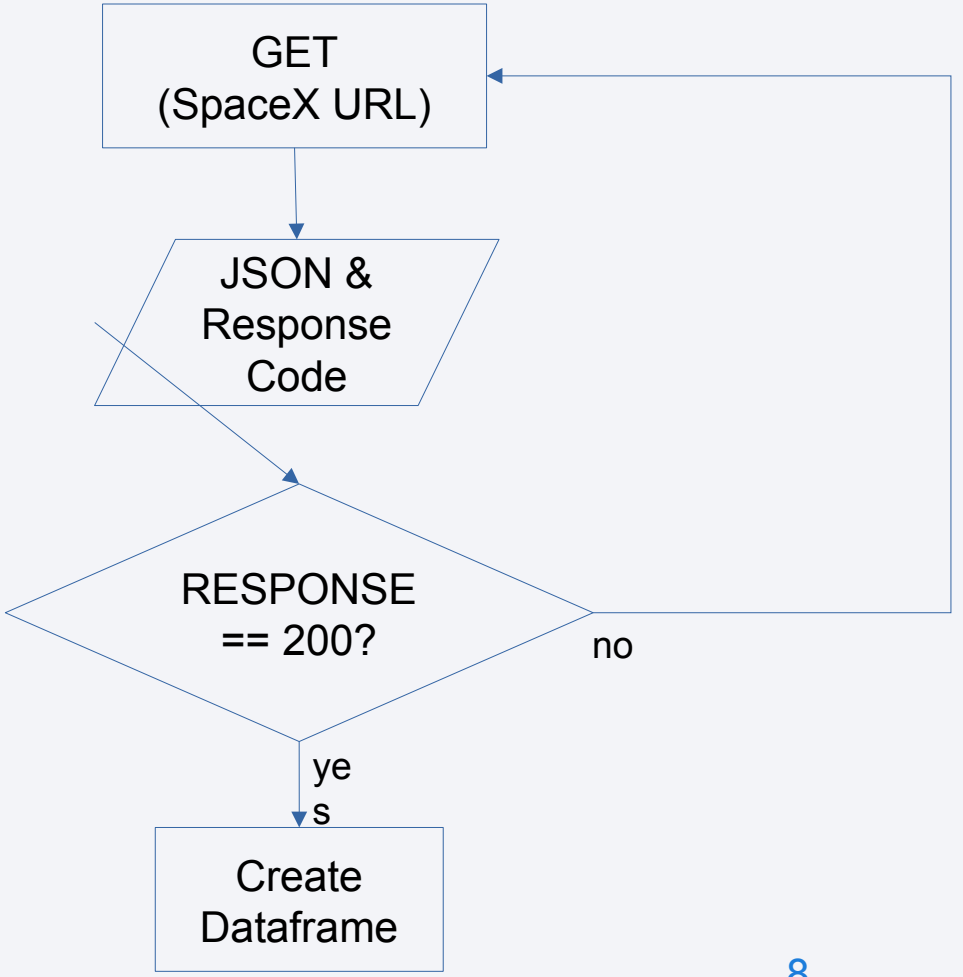
Executive Summary

- Data collection methodology:
  - SpaceX REST API (Datasets describing: Capsules, Cores, Launches, …)
  - Web-Scraping (i.e. collect datasets from web pages)
- Perform data wrangling
  - Data Discovery → Structuring → Cleaning → Enriching → Validating → Publishing
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build, tune, evaluate classification models (Logistic Regression, Decision Tree, K-Nearest-Neighbour, Support Vector Machine)

R. Aue

Data Collection – SpaceX API

- Collect data via *get(url) REST* call
- Create *PANDAS* data frame by using ther received *JSON* data



R. Aue

Reference:

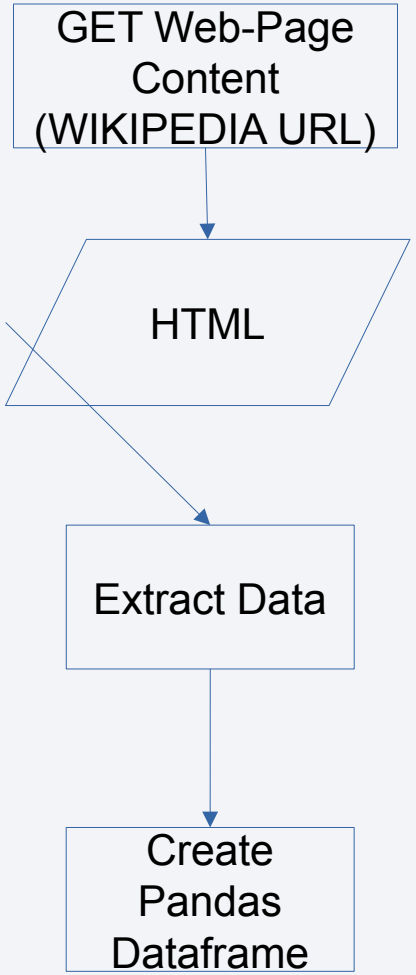
- SpaceX API calls Jupyter Notebook

Data Collection – Web Scraping

Motivation: Use if API is not available

General steps:

- Collect Falcon 9 launch data HTML data from Wikipedia
- Extract launch records *HTML* table
- Parse the table and convert it into a *PANDAS* data frame



R. Aue

Reference:

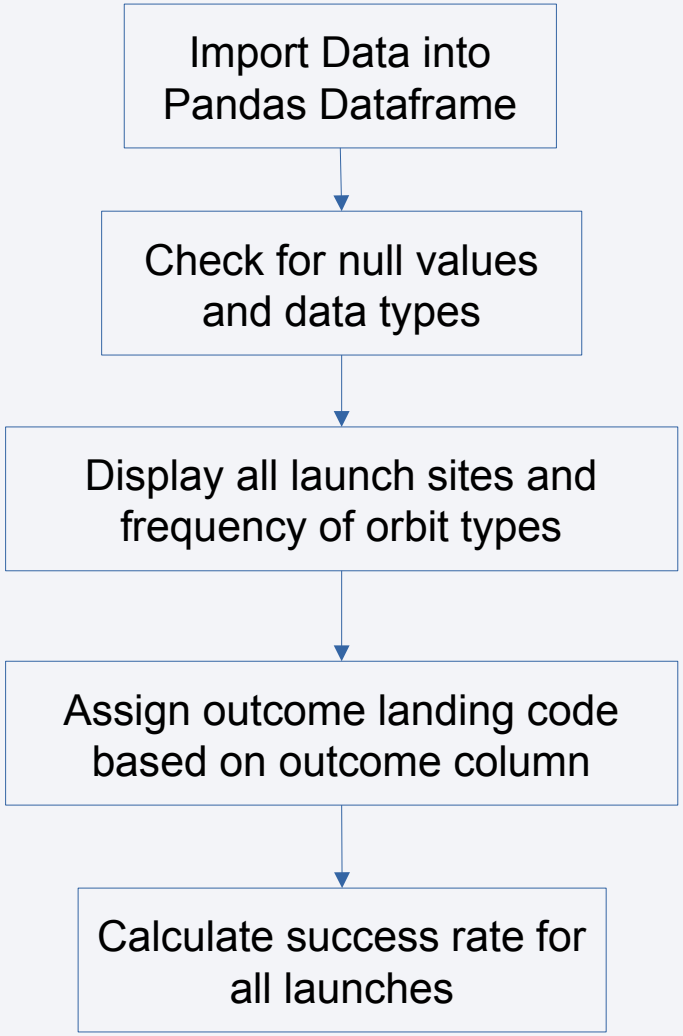
- SpaceX Web-Scraping Jupyter Notebook

Data Wrangling

- Identifying key data components
  - Missing data
  - Data types
- Breakdown of the data
  - Lauchesbysite/orbit
  - Outcome of each mission
- Feature engineering
  - Identify *good* vs. *bad* outcomes
  - Apply a class variable to outcome types
  - Calculate success rate from class frequency

Reference:

- SpaceX Data Wrangling Jupyter Notebook



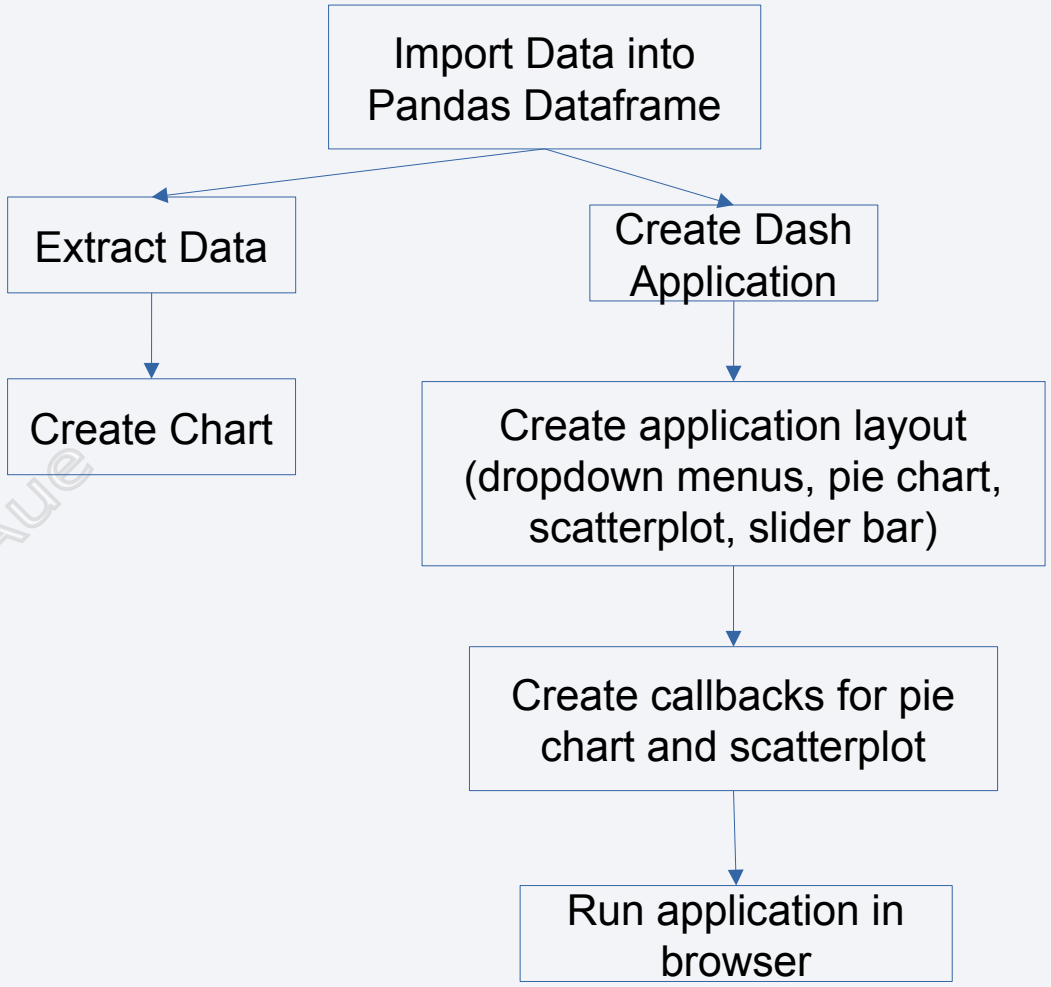
R. Aue

EDA with Data Visualization

Charts
- FlightNumbervs.Payload
- FlightNumbervs.LaunchSite
- Payload Mass (kg) vs. Launch Site
- Payload Mass (kg) vs. Orbit type

Interactiv Charts (Plotly/Dash)
- Success Rate vs. Launch Site
- Success Rate vs. Payload Mass

Analysis
- View relationship by using scatter plots. The variables could be useful for machine learning if a relationship exists
- Show comparisons among discrete categories with bar charts. Bar charts show the relationships among the categories and a measured value.



R. Aue

Reference:
- SpaceX Data Visualization Notebook & Python Dash Application

EDA with SQL

EDA was performed using Structured Query Language ("SQL") queries

- The following were examined to search for patterns in the data:
  - 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 names of 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 names of the booster_versions which have carried the maximum payload mass. Use a subquery.
  - 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.

R. Ale

Reference:

- SpaceX SQL Notebook

Build an Interactive Map with Folium

Using the Python library Folium, a geographic analysis of the launch sites was conducted

- Marked all launch sites using Folium Circles on a map of the United States, centered on NASA's Johnson Space Center in Houston
- Marked all the successful and failed launches for each site on the map using color-indexed Folium markers
- Calculated the distances between a launch site to important proximities like railways, highways, and coastlines – and marked an example with a line

R. Aue

Visual markers ease analyzing the relationship between launch sites and success rates, as well was what factors might influence launch site selection.

Reference:

- SpaceX Data Visualization Notebook using interactive Folium Map

Build a Dashboard with Plotly Dash

Dropdown List with Launch Sites

- Allow user to select all launch sites or a certain launchsite

Pie Chart Showing Successful Launches

- Allow user to see successful and unsuccessful launches as a percent of the total

Slider of Payload Mass Range

- Allow user to select payload mass range

Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version

- Allow user to see the correlation between Payload and Launch Success

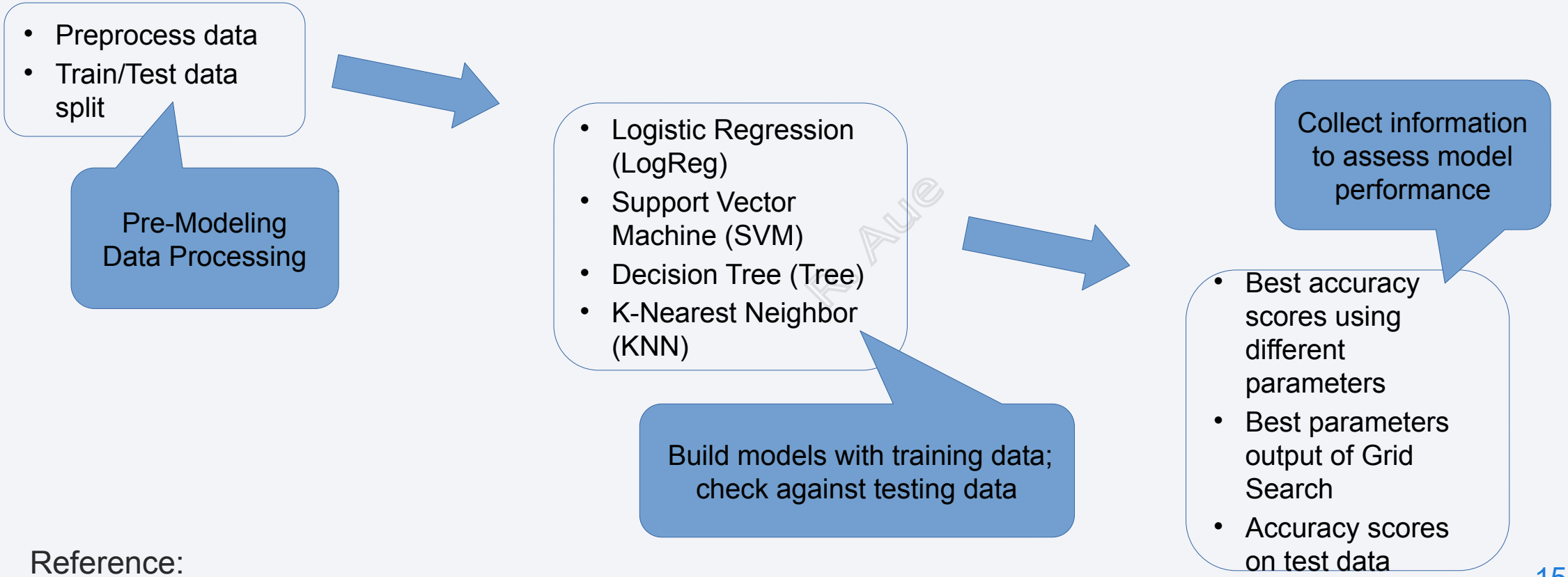
R. Aue

Reference:

- SpaceX Data Visualization using Plotly Dash

Predictive Analysis (Classification)

Multiple modeling techniques and strategies were explored to determine the best predictor of launch success or failure. The process followed a standard flow:



Reference:

- SpaceX Prdeictive Analysis Jupyter Notebook

Results

Exploratory Data Analysis

- Launch success has improved over time
- KSC LC-39A has the highest success rate among landing sites
- Orbits ES-L1, GEO, HEO and SSO have a 100% success rate

Visual Analytics

- Most launch sites are near the equator, and all are close to the coast
- Launch sites are far enough away from anything a failed launch can damage (city, highway, railway), while still close enough to bring people and material to support launch activities

Predictive Analytics

- Decision Tree model is the best predictive model for the given dataset (accuracy ≈ 94%)

Section 2

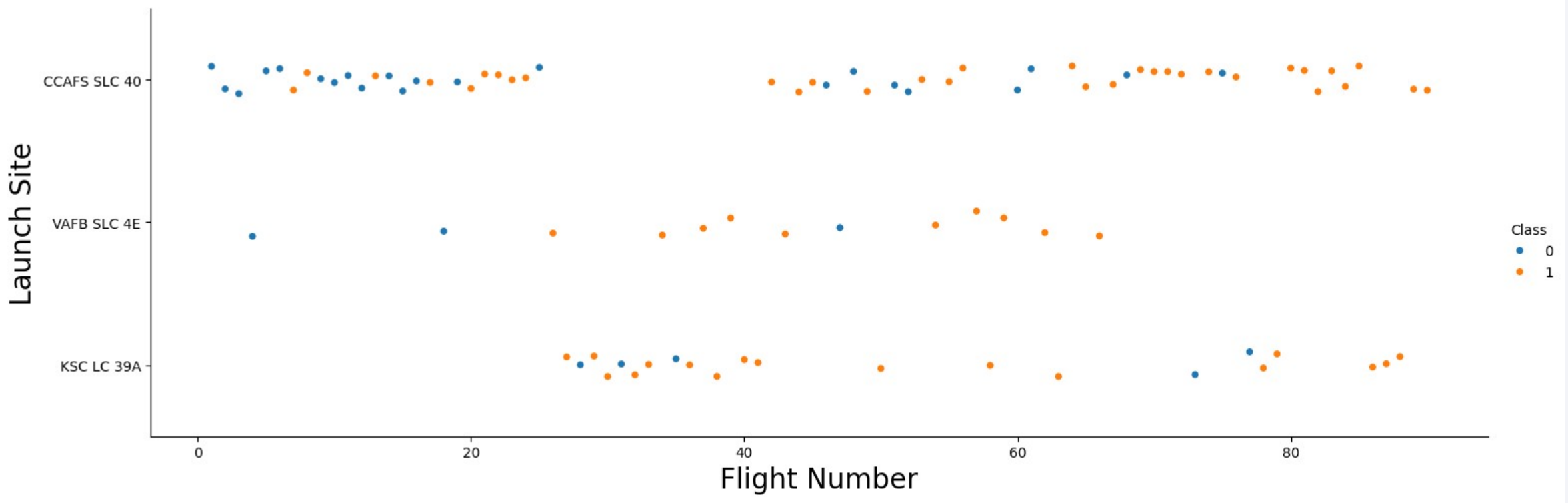
Insights drawn from EDA

R. Aue

Flight Number vs. Launch Site

Exploratory Data Analysis

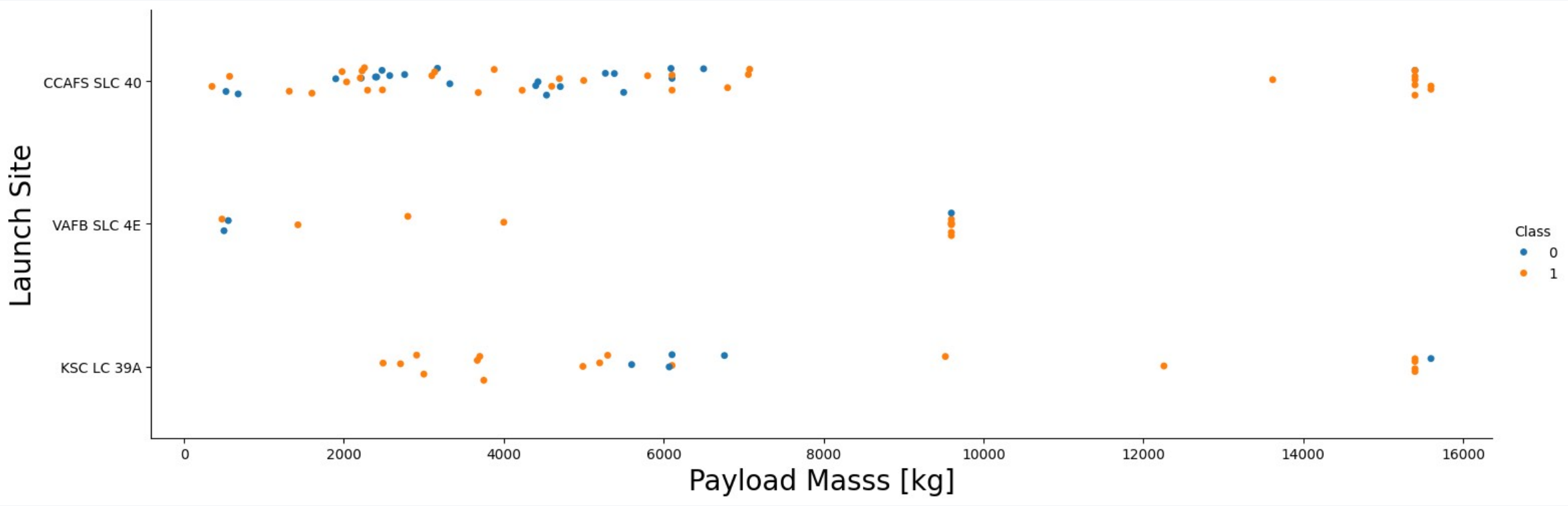
- Earlier flights had lower success rate (blue = fail)
- Later flights had higher success rate (orange = success)
- Approx. half of launches were from CCAFS SLC 40 launch site
- VAFB SLC 4E and KSC LC 39A have higher success rates
- New launches have higher tendency to succed



Payload vs. Launch Site

Exploratory Data Analysis

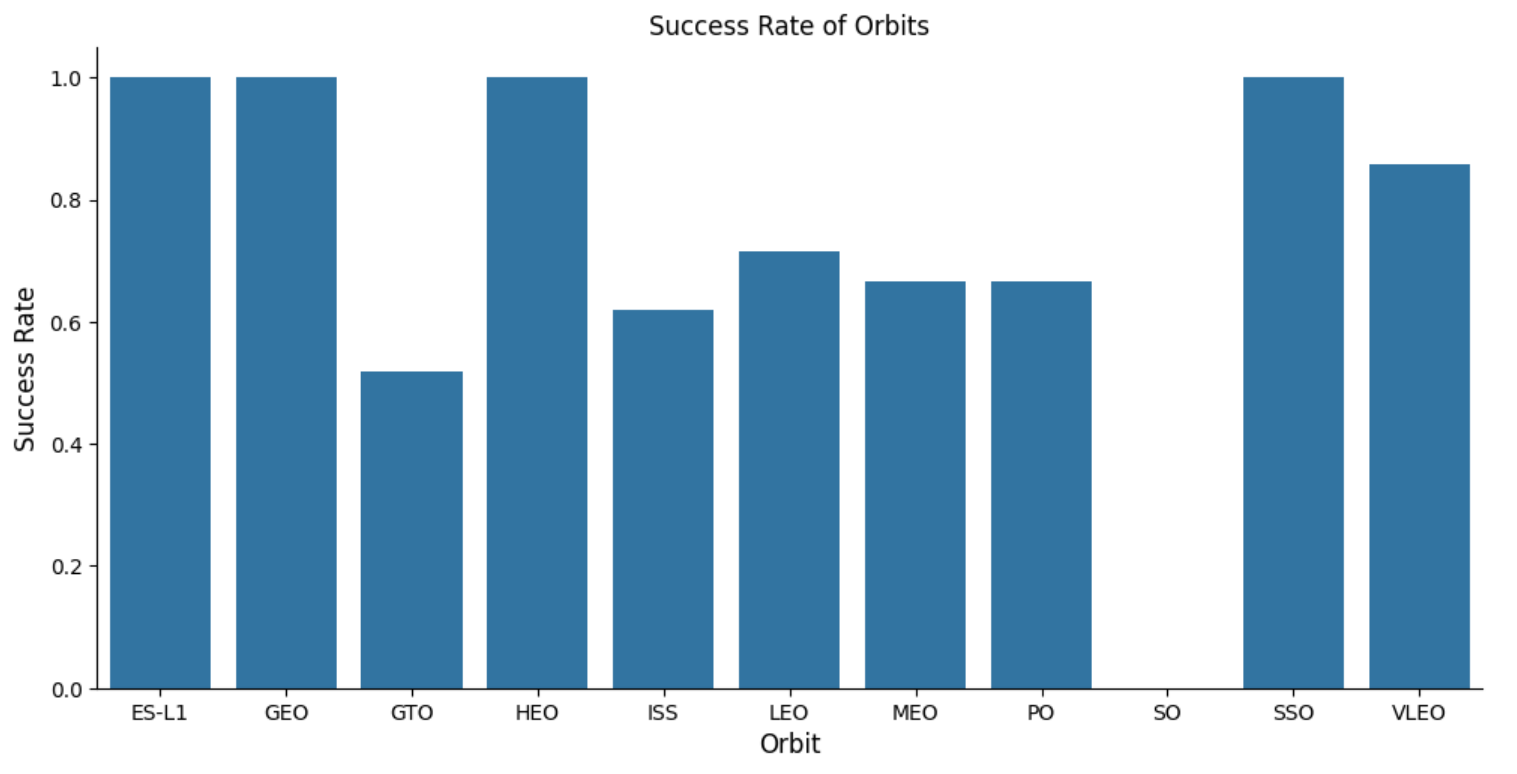
- Typically, the higher the payload mass, the higher the success rate
- Most launces with payload greater than 7,000 kg were successful
- KSC LC 39A has 100% success rate for launches less than 5,500 kg
- VAFB SKC 4E has not launched anything greater than ~10,000 kg



Success Rate vs. Orbit Type

Exploratory Data Analysis

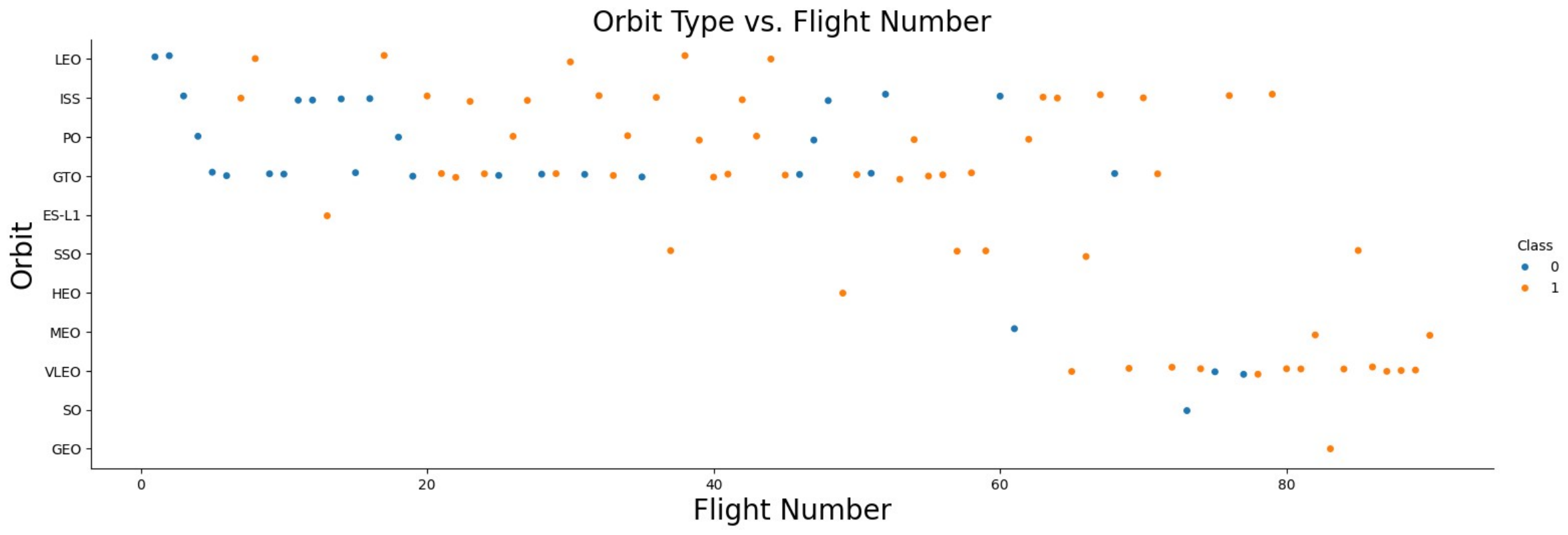
- 100% Success Rate: ES-L1, GEO, HEO and SSO
- 50% - 80% Success Rate: GTO, ISS, LEO, MEO, PO
- 0% Success Rate: SO



Flight Number vs. Orbit Type

Exploratory Data Analysis

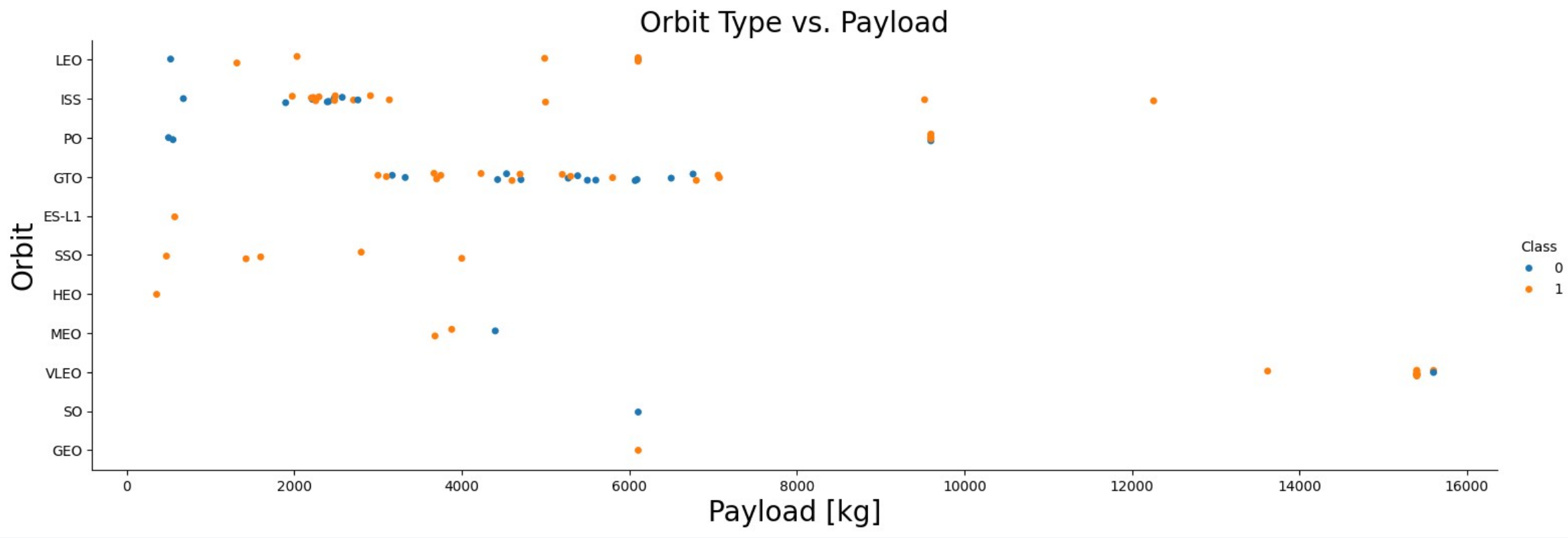
- The success rate typically increases with the number of flights for each orbit
- This relationship is highly apparent for the LEO orbit
  - The GTO orbit, however, does not follow this trend



Payload vs. Orbit Type

Exploratory Data Analysis

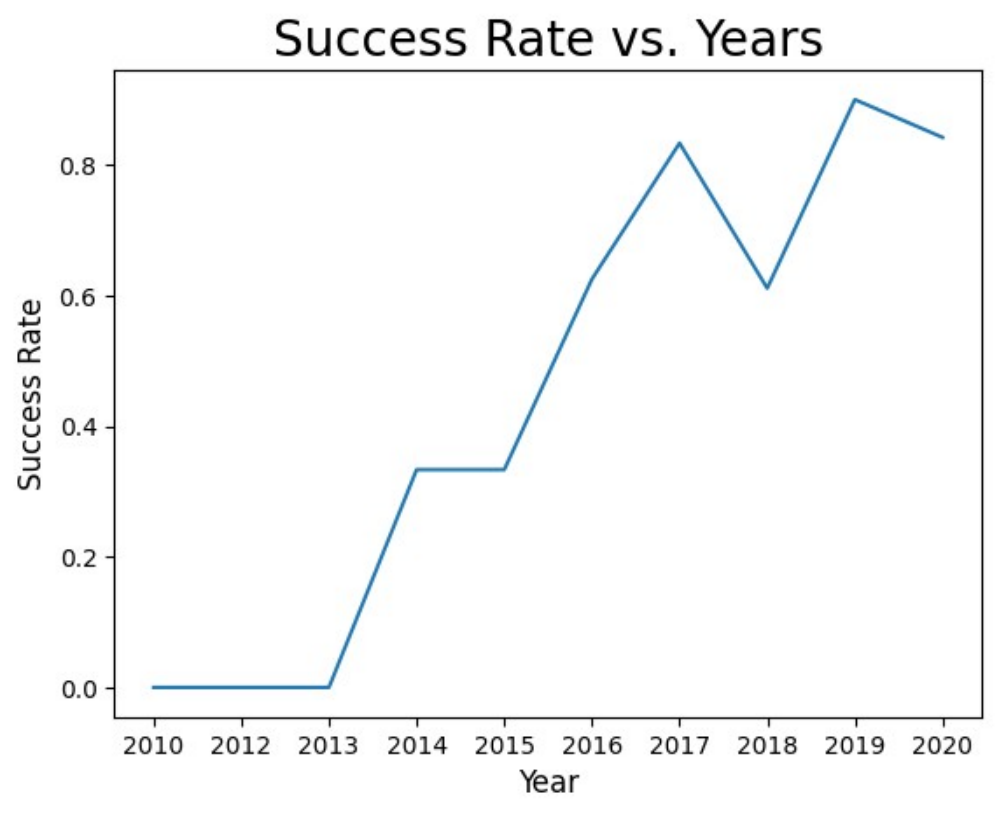
- Heavy payloads perform better in LEO, ISS and PO orbits
- The GTO orbit has mixed success with heavier payloads



Launch Success Yearly Trend

SpaceX has improved its launch capabilities over time, resulting in higher success rates year over year

- From 2010 to 2013, all landings were unsuccessful
- Starting in 2014, SpaceX began successfully landing Falcon 9 rockets
- Over time, SpaceX has improved its annual success rate to about 80%



All Launch Site Names

SQL Query:
%sql select distinct Launch_site from SPACEXTABLE;

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

R. Aue

- This SQL query extracts a list of unique launch site values
- Note: Cape Canaveral is represented by two values: CCAFS LC-40” and “CCAFS SLC-40”

Launch Site Names Begin with 'CCA'

SQL Query:
*%sql select * from SPACEXTABLE where Launch_site like 'CCA%' limit 5;*

Result:

- This query identifies the first 5 launches that took place in Cape Canaveral

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

SQL Query:

%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = 'NASA (CRS)';

Result:

sum(PAYLOAD_MASS__KG_)

45596

R. Aue

This query calculates the total payload amount launched by NASA (CRS) over all launches.

Average Payload Mass by F9 v1.1

SQL Query:

%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version = 'F9 v1.1';

Result:

avg(PAYLOAD_MASS__KG_)

2928.4

R. Aue

This query calculates the average payload for a Falvon 9 rocket that's version 1.1

1st Successful Ground Landing Date

SQL Query:
%sql select Date from SPACEXTABLE where Landing_outcome = 'Success (ground pad)' order by Date asc limit 1;

Result:
Date
2015-12-22

R. Aue

This query returns the date of the first successful landing on a ground pad.

Successful Drone Ship Landing with Payload between 4000 and 6000 kg

SQL Query:

%sql select distinct Booster_version from SPACEXTABLE where PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000 and Landing_Outcome = 'Success (drone ship)';

Result:

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

R. Aue

This query returns a list of versions of the Falcon 9 rocket that have successfully landed on a drone ship after carrying payloads greater than 4000 kg and less than 6000 kg.

Total Number of Successful and Failure Mission Outcomes

SQL Query:
%sql select count(Mission_Outcome) from SPACEXTABLE;

Result:
count(Mission_Outcome)
101

R. Aue

This query returns a total count of both successful and failed mission outcomes.

Boosters Carried Maximum Payload

SQL Query:

%sql select Booster_Version,PAYLOAD_MASS__KG_ from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE);

Result:

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

R. Aue

This query returns a list of the Falcon 9 versions that have carried ever carried the maximum payload.

2015 Launch Records

SQL Query:

*%sql select substr ("--JanFebMarAprMayJunJulAugSepOctNovDec", strftime ("%m", Date) * 3, 3) as Month, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTABLE where substr(Date,0,5) = '2015' and Landing_Outcome = 'Failure (drone ship)';*

Result:

Month	Landing_Outcome	Booster_Version	Launch_Site
Jan	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Apr	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

R. Aue

This query returns a list of all 2015 launch records with unsuccessful landings on a drone ship, including the launch site and version of the Falcon 9 booster used.

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

SQL Query:

%sql select Landing_Outcome, count(Landing_Outcome) from SPACEXTABLE where Date > '2010-06-04' and Date < '2017-03-20' group by Landing_Outcome order by count(Landing_Outcome) desc;

Result:

Landing_Outcome	count(Landing_Outcome)
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

R. Aue

This query returns a list of launch sites and a count of the launches/missions conducted at each site, in descending order by mission count, for launches that occurred between June 4, 2010 and March 20, 2017

Section 3

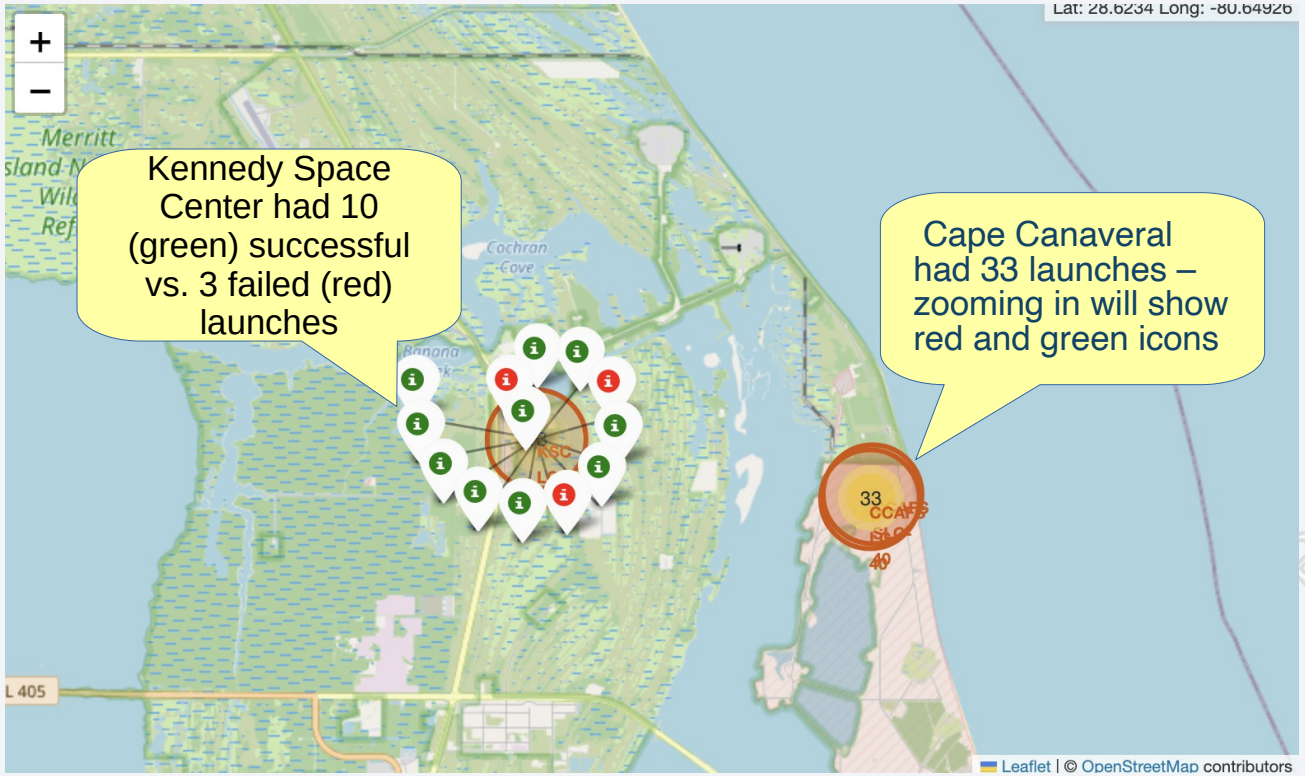
Launch Sites Proximities Analysis

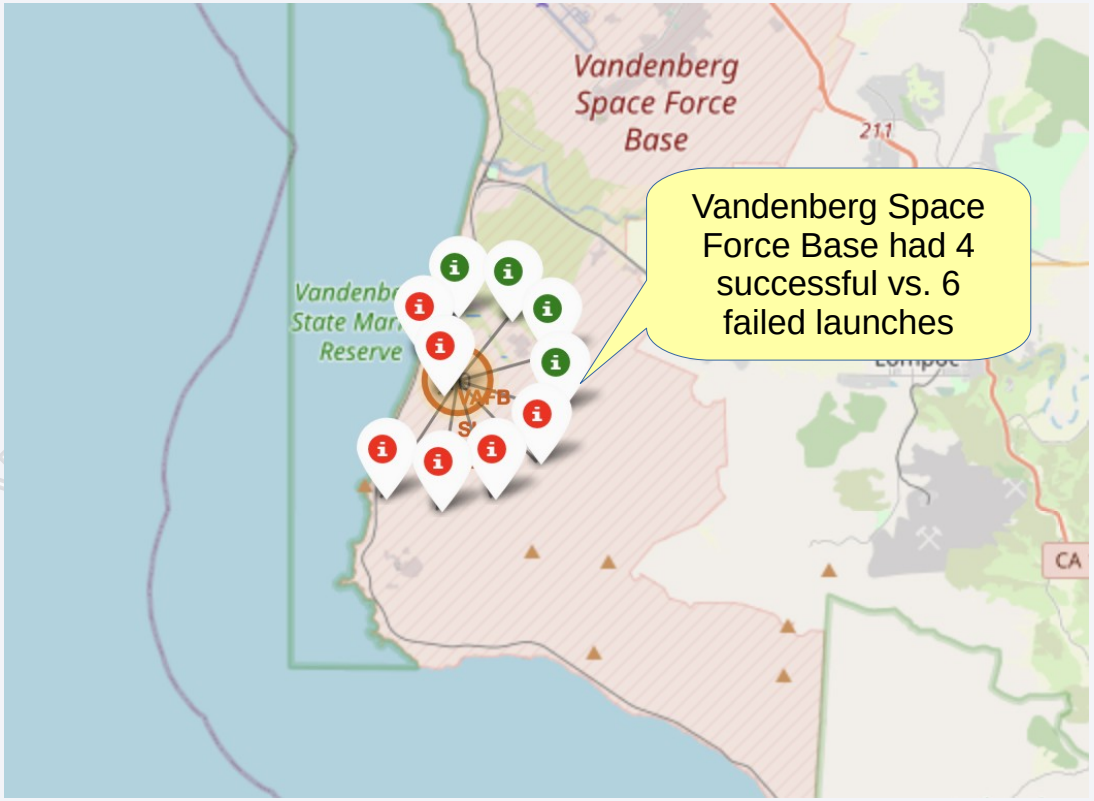
R. Aue

Launch Sites

- All launch sites located in the southern United States
  - the closer the launch site to the equator, the easier it is to launch to equatorial orbit;
  - Central Florida has a higher risk of severe weather conditions, but the lower latitude position overbalances that;

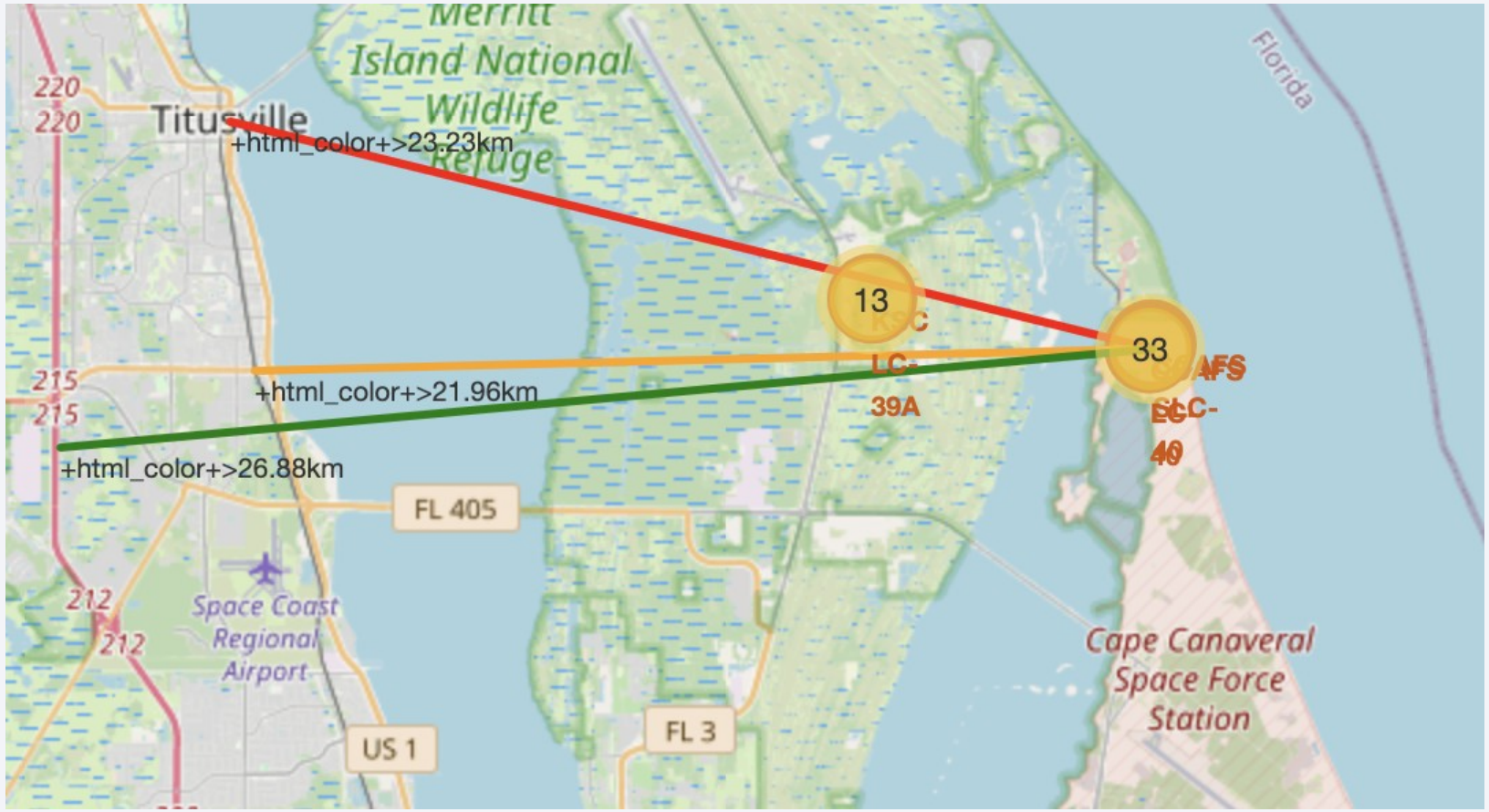
Launch Outcomes





- The folium maps were modified to have markers for each launch, color-coded to visually represent success or failure
- As a user zooms in (or out), they see more (or grouped) markers.

Distance to Proximities

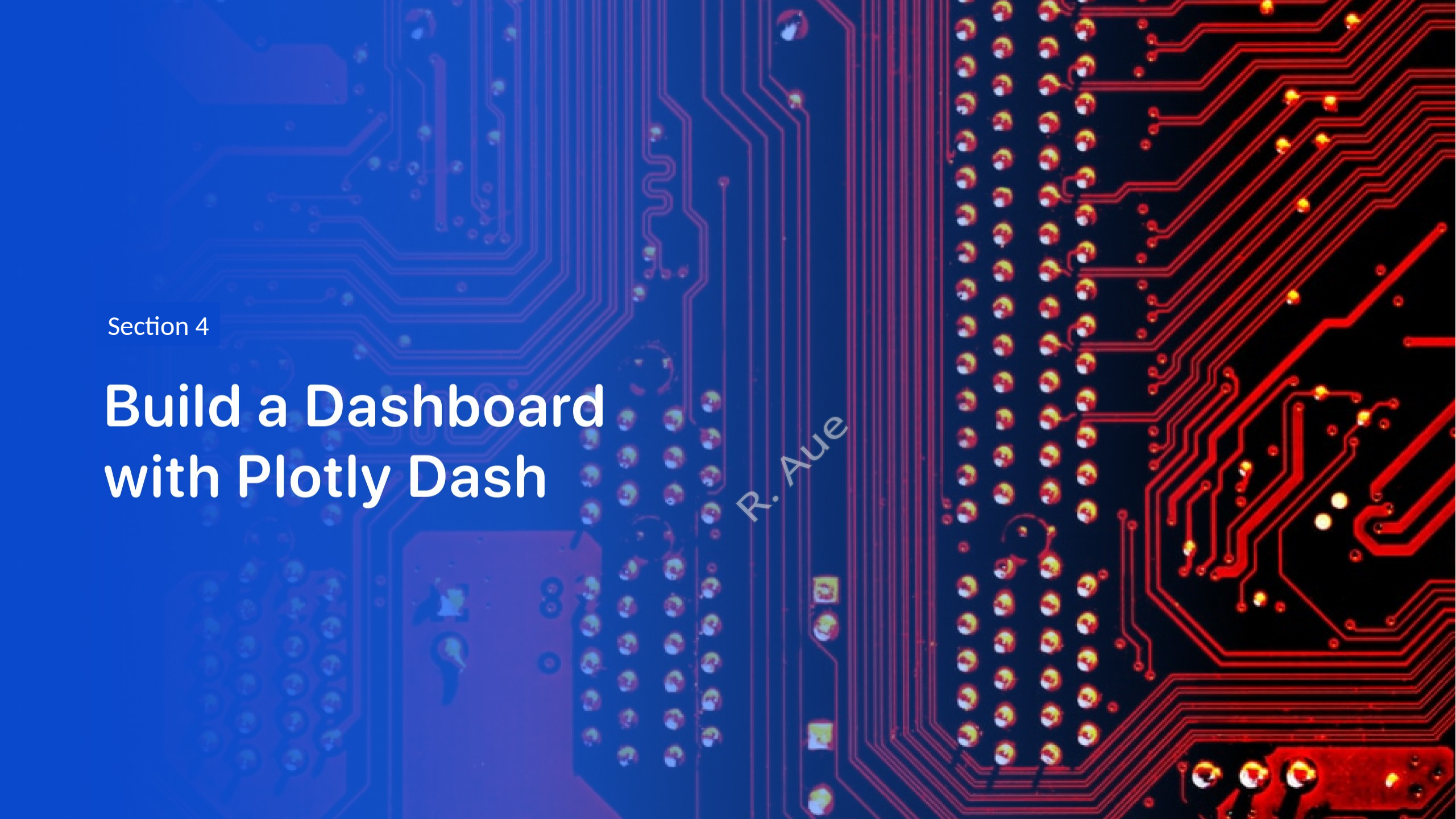


All 3 Launch Sites are …

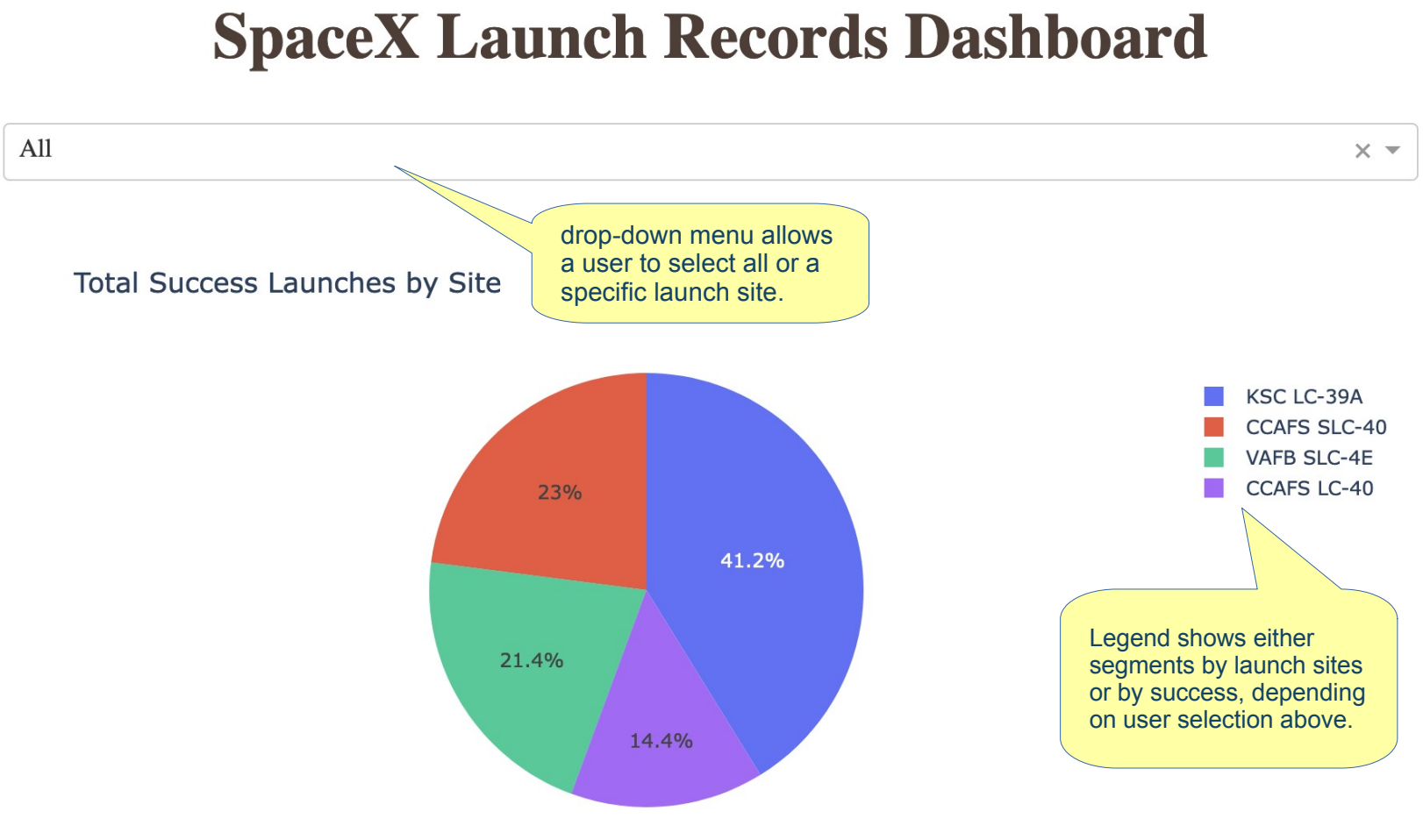
- close to water or coastlines
- close to highways and railways
- in certain distance to cities in case of failures.

Cape Caneveral CCAFS SLC-40 Site:

- 0.86 km from nearest coastline
- 21.96 km from nearest railway
- 23.23 km from nearest city
- 26.88 km from nearest highway

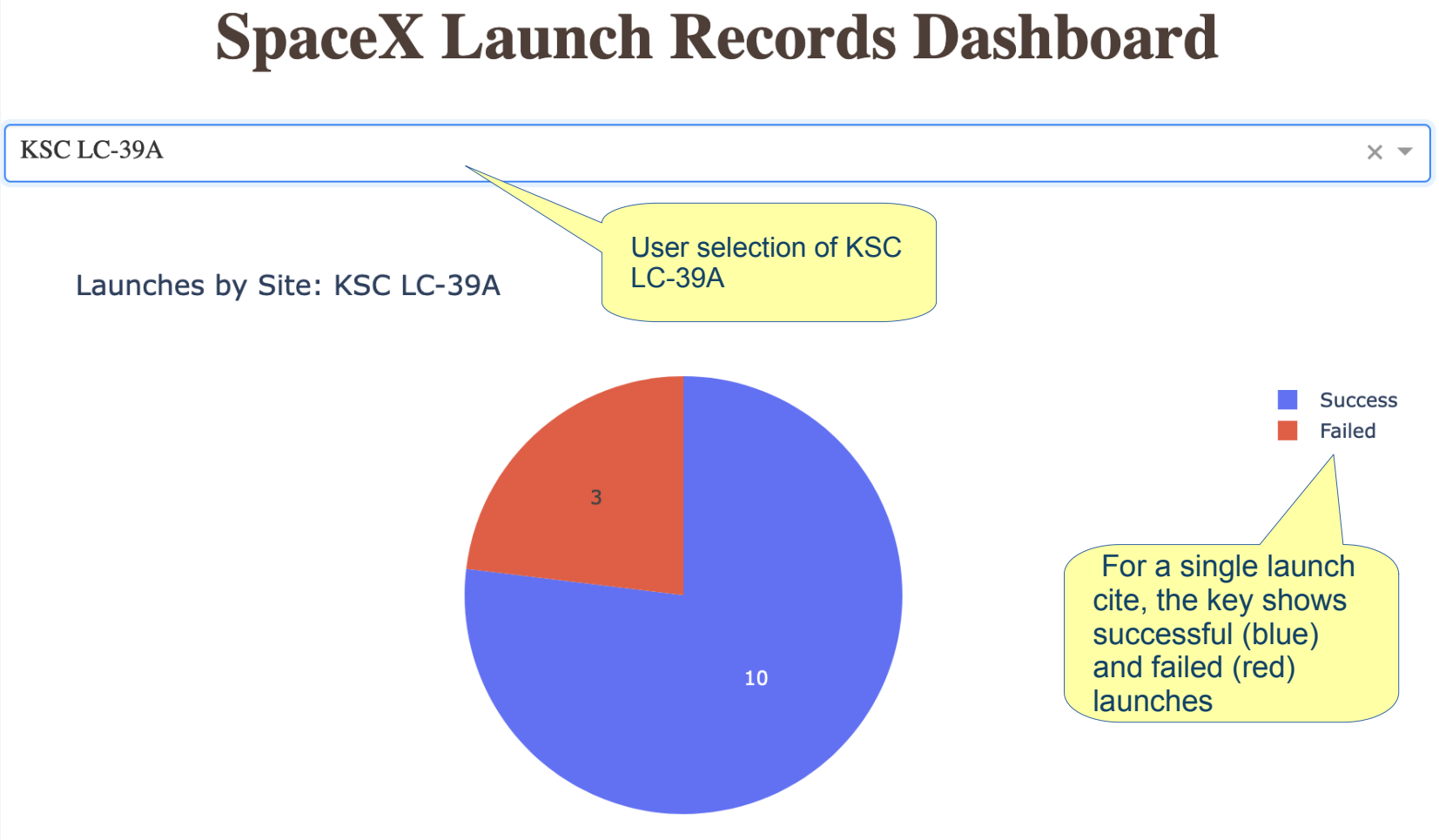

Section 4
Build a Dashboard with Plotly Dash
R. Aue

Launch Success by Site



- The majority of successful launches have occurred in Central Florida, either at Cape Canaveral (~41.6%) or Kennedy Space Center (~41.7%).
- Vandenberg Space Force Base is responsible for ~16.7% of successful launches.

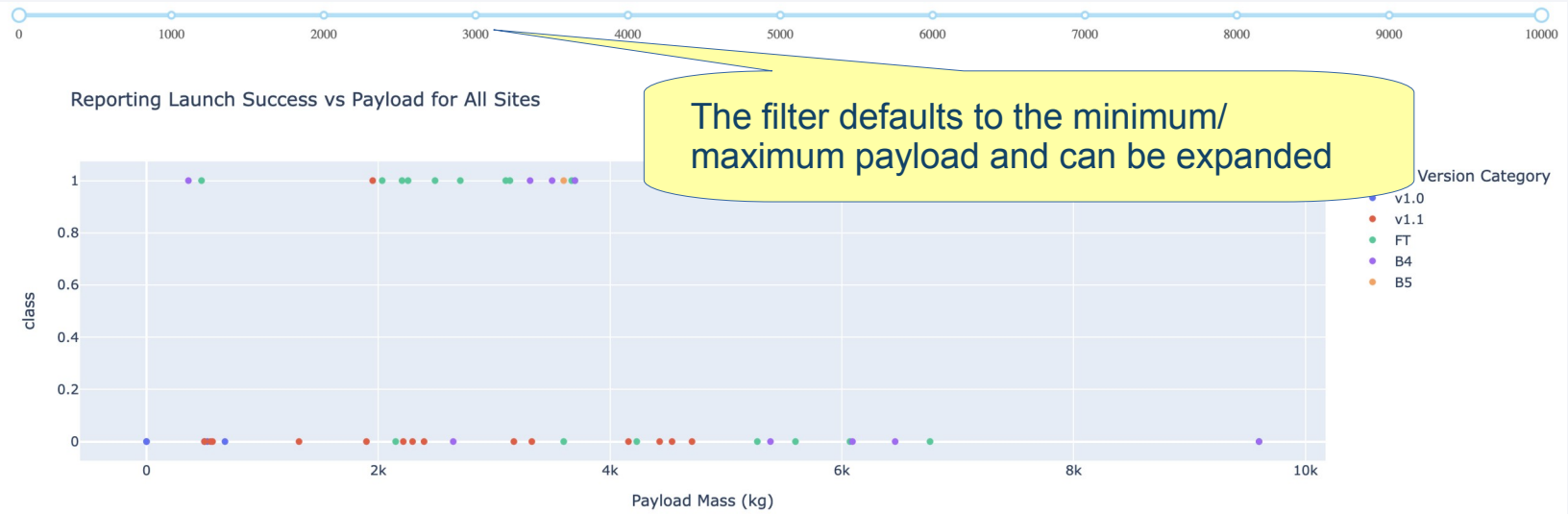
Launch Success KSC LC-39A

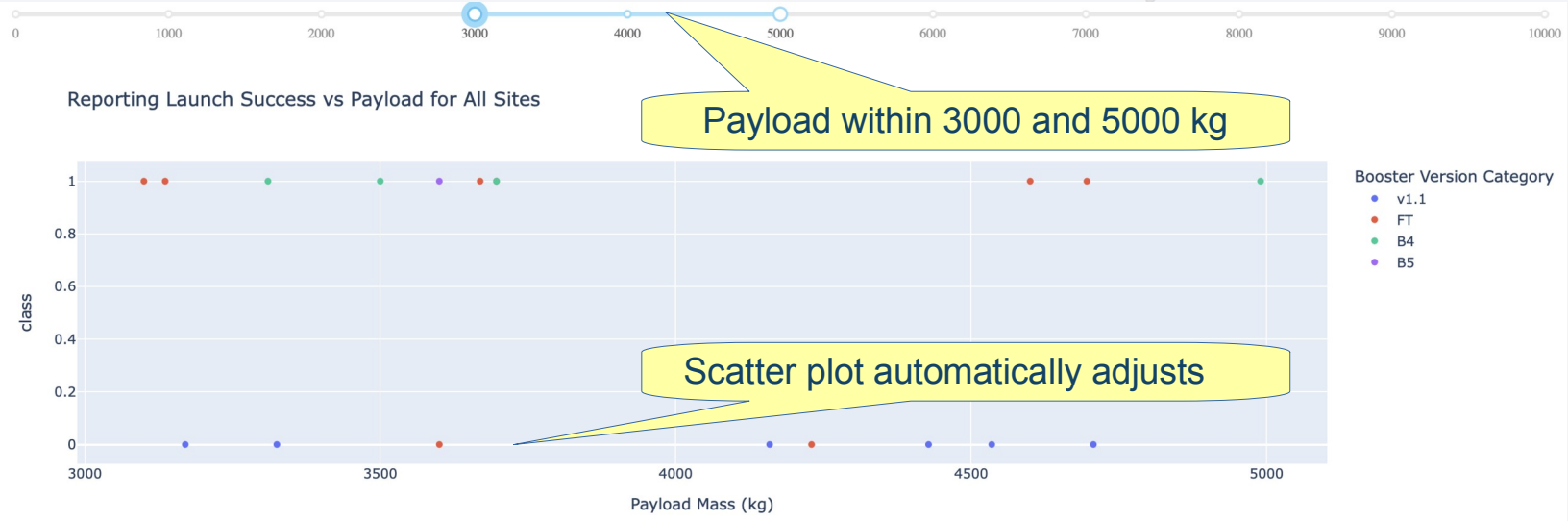


Success vs. Failed Launches

- KSC LC-39A had 10 successful launches and 3 failed launches
  - which the highest success rate amongst launch sites (76.9%)ailed

Payload Mass and Success



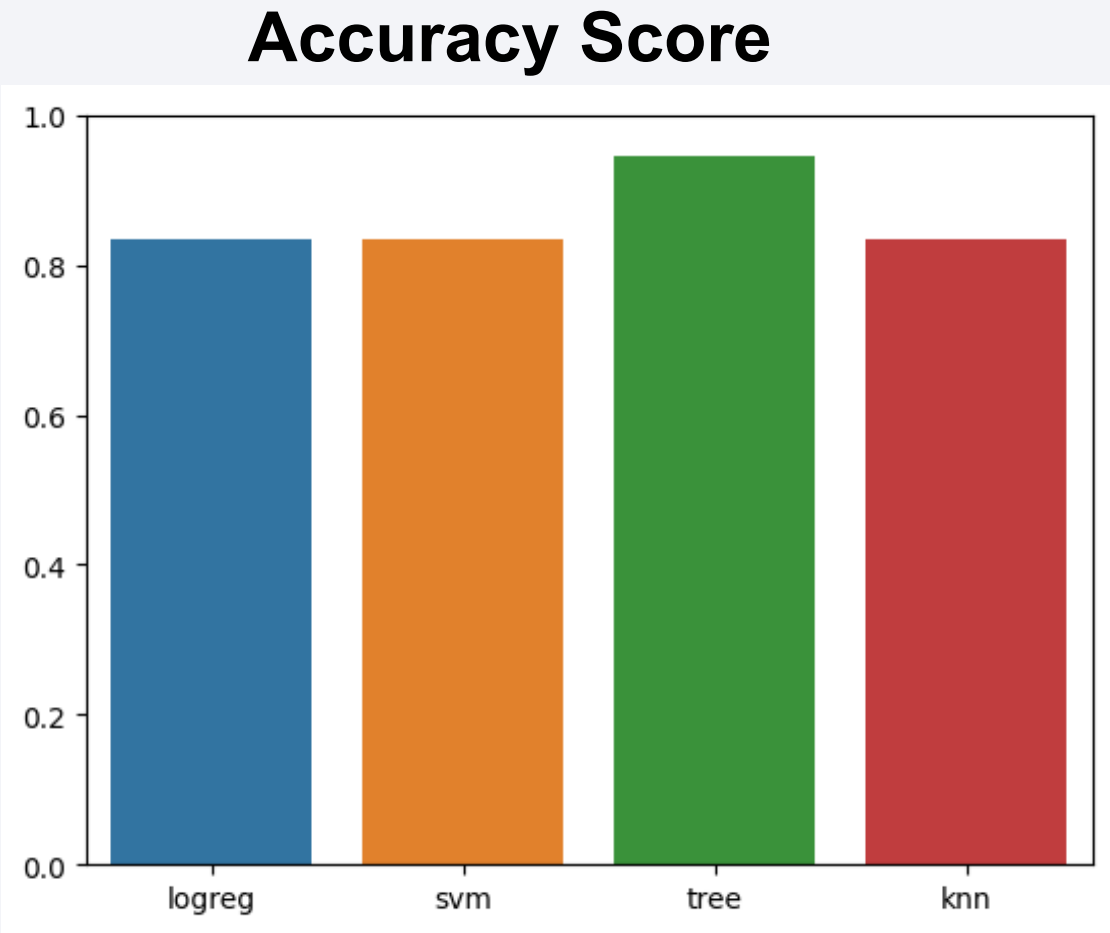


- This graph shows more information and can be adjusted by a slider
- Color-coded dots indicate that the “FT” Booster version has had the most successful launches of all versions, while “v1.1” has had the least
- Over time, launches have shifted away from “v1” and “v1.1” and presumably towards newer versions

Section 5

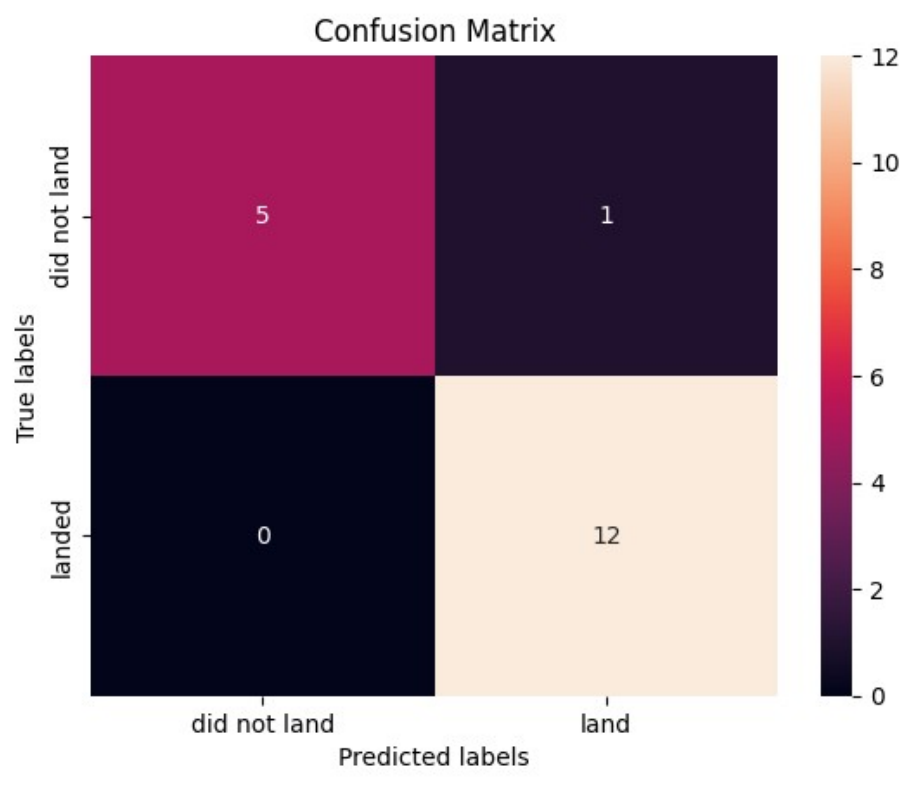
Predictive Analysis (Classification)

Classification Accuracy



- Accuracy scores were calculated for each of the 4 model types:
  - Logistical Regression
  - Support Vector Machine
  - Decision Trees
  - K-Nearest Neighbor
- Additionally, for each of the *Grid Searches* conducted to identify the strongest version of a specific model type, we identified:
  - The optimal parameters
  - That version's accuracy score
- The *Decision Tree* modeling technique had the highest *Accuracy Score*, at 0.944

Confusion Matrix *Decision Tree*



- The *Decision Tree* technique produced the most accurate model, with a test accuracy score of 0.944
  - The best parameters, per the grid search, were:
    - criterion: 'gini'
    - max_depth: 12
    - max_features: 'sqrt'
    - min_samples_leaf: 4
    - min_samples_split: 5
    - splitter: 'best'
  - The best accuracy score was 0.9018
- As shown in the confusion matrix, the model had only one false positive
  - 5 landing failures were predicted correctly
  - 12 landing successes were predicted correctly

Conclusions

- There is a relationship between the success of a launch and its payload, orbit, and launch site.
  - The flight number (as steadily incremented number) is also correlated to success – with larger flight numbers having higher success rates
  - Booster version is also correlated with success, though there is a correlation between booster version and experience (flight number)
- Launch sites are always located near major waterways (i.e. oceans)
  - Easy access to railways and highways is also prioritized
  - Launch cites in Florida have higher success rates
- The model technique with the most accurate predictions is Decision Tree
- Using a Decision Tree Model, we can predict the success or failure of a Falcon9 landing based on a number of key inputs – which will help determine the overall cost of a launch

Appendix

References:

[1] Falcon 9 and Falcon Heavy Launches 2010-2019, WIKIPEDIA

Assets:

- Main assets (i.e data files, Jupyter Notebooks, Python scripts) are stored in GITHUB
  - Other assets (e.g. charts, query results, score data) can be derived/generated by main assets.

R.Aue

Thank you!