IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Emily F.
2022.07.09

Outline

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

Executive Summary

- Summary of methodologies
    - Data Collection using web scraping and SpaceX API;
    - Exploratory Data Analysis (EDA), including data wrangling, data visualization and interactive visual analytics;
    - Machine Learning Prediction.

- Summary of all results
    - Able to collected valuable data from public sources;
    - EDA allowed to identify which features are the best to predict success of launchings;
    - Machine Learning Prediction showed the best model to predict which characteristics drive this opportunity by the best way, using all collected data.

Introduction

- Project background and context

Cost of launch continues to remain a key barrier for new competitors in private space travel. SpaceX with its first stage reuse capabilities offers a key advantage against its competitors. Each SpaceX launch costs around $62M and SpaceX can reuse stage 1 for future launches. This provides SpaceX a unique advantage where other competitors are spending around 165 mission plus for each launch.

- Problems you want to find answers.

    - The best way to estimate the total cost for launches, by predicting successful landings of the first stage of rockets;

    - Where is the best place to make launches.

Section 1

Methodology

Methodology

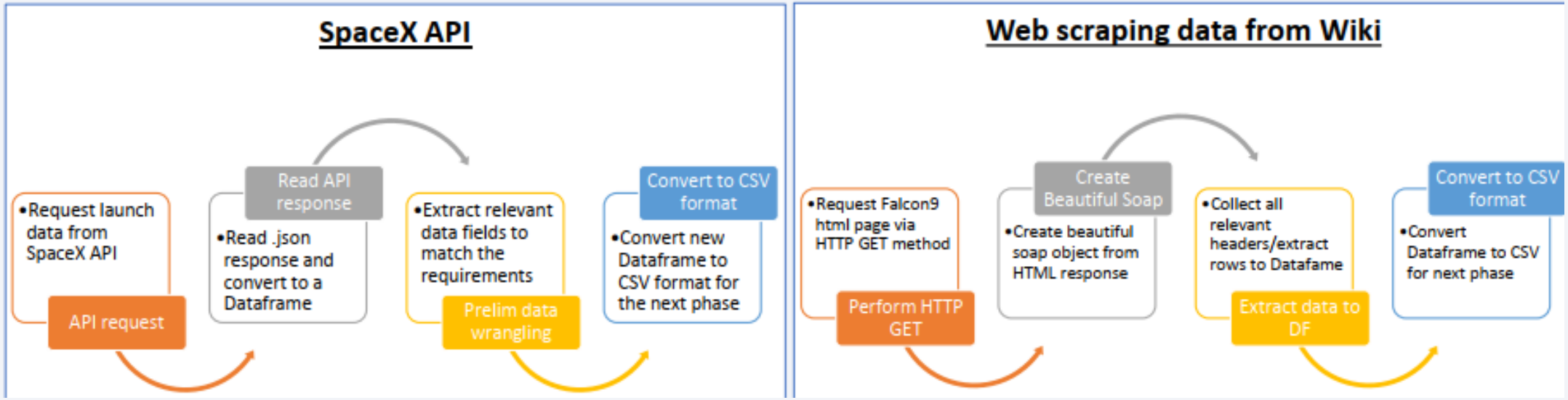
Executive Summary

- Data collection methodology:
  - SpaceX API
  - Web scrap Falcon 9 and Falcon Heavy launch records from Wikipedia
- Perform data wrangling
  - Determined labels for training the supervised models by converting mission outcomes in to training labels (0-unsuccessful, 1-successful)
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Created a column for ‘class’; standardized and transformed data; train/test split data; find best classification algorithm (Logistic regression, SVM, decision tree, & KNN) using test data

Data Collection

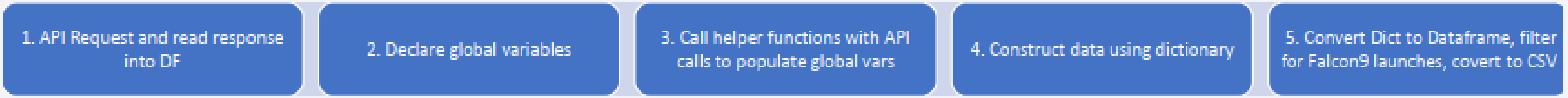
- Data sets were collected from Space X API (https://api.spacexdata.com/v4/rockets/) and from Wikipedia using web scraping techniques.

- Data collection process use key phrases and flowcharts



Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

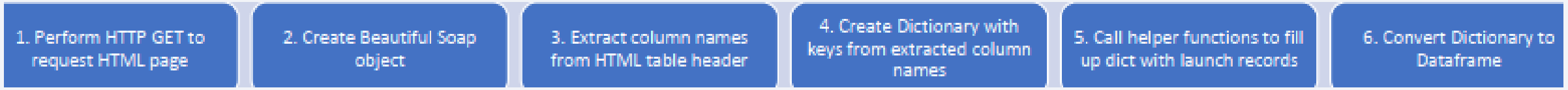


- GitHub URL of the completed SpaceX API calls notebook

https://github.com/Tech-for-Fun/DS-Project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts



- GitHub URL of the completed web scraping notebook

https://github.com/Tech-for-Fun/DS-Project/blob/main/jupyter-labs-webscraping.ipynb

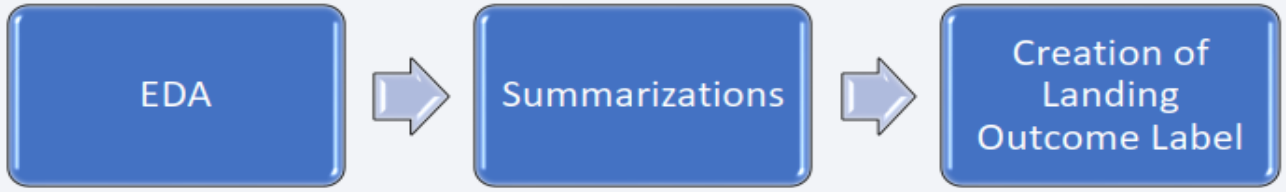
Data Wrangling

- Describe how data were processed
  - Explored data to determine the label for training supervised models

- Calculated the number of launches on each site, each orbit, mission outcome per orbit type

  - Encode a landing outcome from 'Outcome' column, when booster land successfully class =1, otherwise 0.

- data wrangling process using key phrases and flowcharts



- GitHub URL of your completed data wrangling related notebooks

https://github.com/Tech-for-Fun/DS-Project/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts

1. Scatter plot: Shows relationship or correlation between two variables making patterns easy to observe .
2. 2. Bar Chart: Commonly used to compare the values of a variable at a given point in time. Bar charts makes it easy to see which groups are highest/common and how other groups compare against each other.
3. 3. Line Chart: Commonly used to track changes over a period of time. It helps depict trends over time.

- GitHub URL of your completed EDA with data visualization notebook

https://github.com/Tech-for-Fun/DS-Project/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
  - Names of the unique launch sites in the space mission;
  - Top 5 launch sites whose name 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 between 4000 ~ 6000 kg;
  - Total number of successful and failure mission outcomes;
  - Names of the booster versions which have carried the maximum payload mass;
  - Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015; and

- GitHub URL of your completed EDA with SQL notebook

https://github.com/Tech-for-Fun/DS-Project/blob/main/jupyter-labs-eda-sql-coursera.ipynb

Build an Interactive Map with Folium

- Summarize what map objects created and added to a folium map and explain why.

Folium interactive map helps analyze geospatial data to perform more interactive visual analytics and better understand factors such location and proximity of launch sites that impact launch success rate.

Map objects created and added to the map:
- Mark all launch sites on the map. This allowed to visually see the launch sites on the map.
- Added 'folium.circle' and 'folium.marker' to highlight circle area with a text label over each launch site.
- Added a 'MarkerCluster()' to show launch success (green) and failure (red) markers for each launch site.
- Calculated distances between a launch site to its proximities (e.g., coastline, railroad, highway, city)

- GitHub URL of your completed interactive map with Folium map

https://github.com/Tech-for-Fun/DS-Project/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard and explain why.

- Added Launch Site Drop-down, Pie Chart, Payload range slide, and a Scatter chart to the Dashboard.

1. Launch Site Drop-down Input component to the dashboard to provide an ability to filter Dashboard visual by all launch sites or a particular launch site
2. Pie Chart to the Dashboard to show total success launches when 'All Sites' is selected and show success and failed counts when a particular site is selected
3. Payload range slider to the Dashboard to easily select different payload ranges to identify visual patterns
4. Scatter chart to observe how payload may be correlated with mission outcomes for selected site(s). The color-label Booster version on each scatter point provided missions outcomes with different boosters

- GitHub URL of your completed Plotly Dash lab

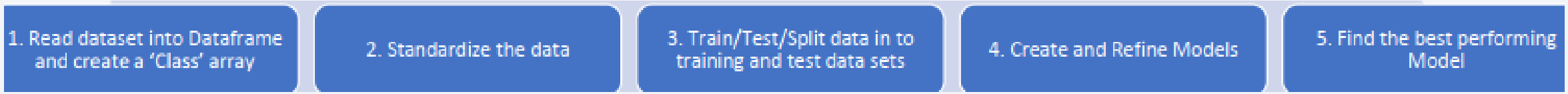
https://github.com/Tech-for-Fun/DS-Project/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model

- 4 classification models were compared: logistic regression, support vector machine, decision tree and k nearest neighbors.

- You need present your model development process using key phrases and flowchart



- GitHub URL of your completed predictive analysis lab: https://github.com/Tech-for-Fun/DS-Project/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

- Exploratory data analysis results

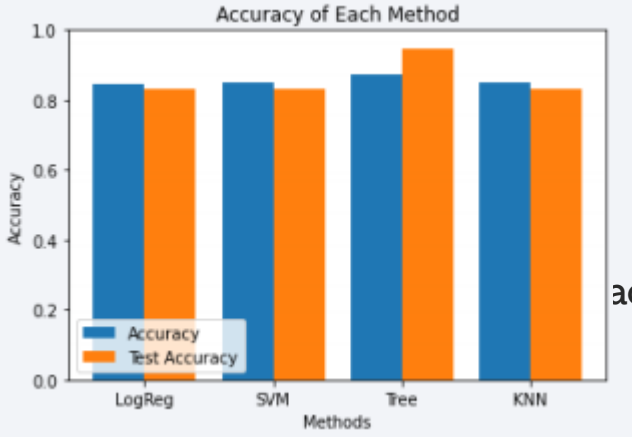
- Space X uses 4 different launch sites. The 1st launches happened in 2010 (Space X and NASA), the 1st success landing outcome was in 2015.
- Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average;
- Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015. The number of landing outcomes became as better as years passed.

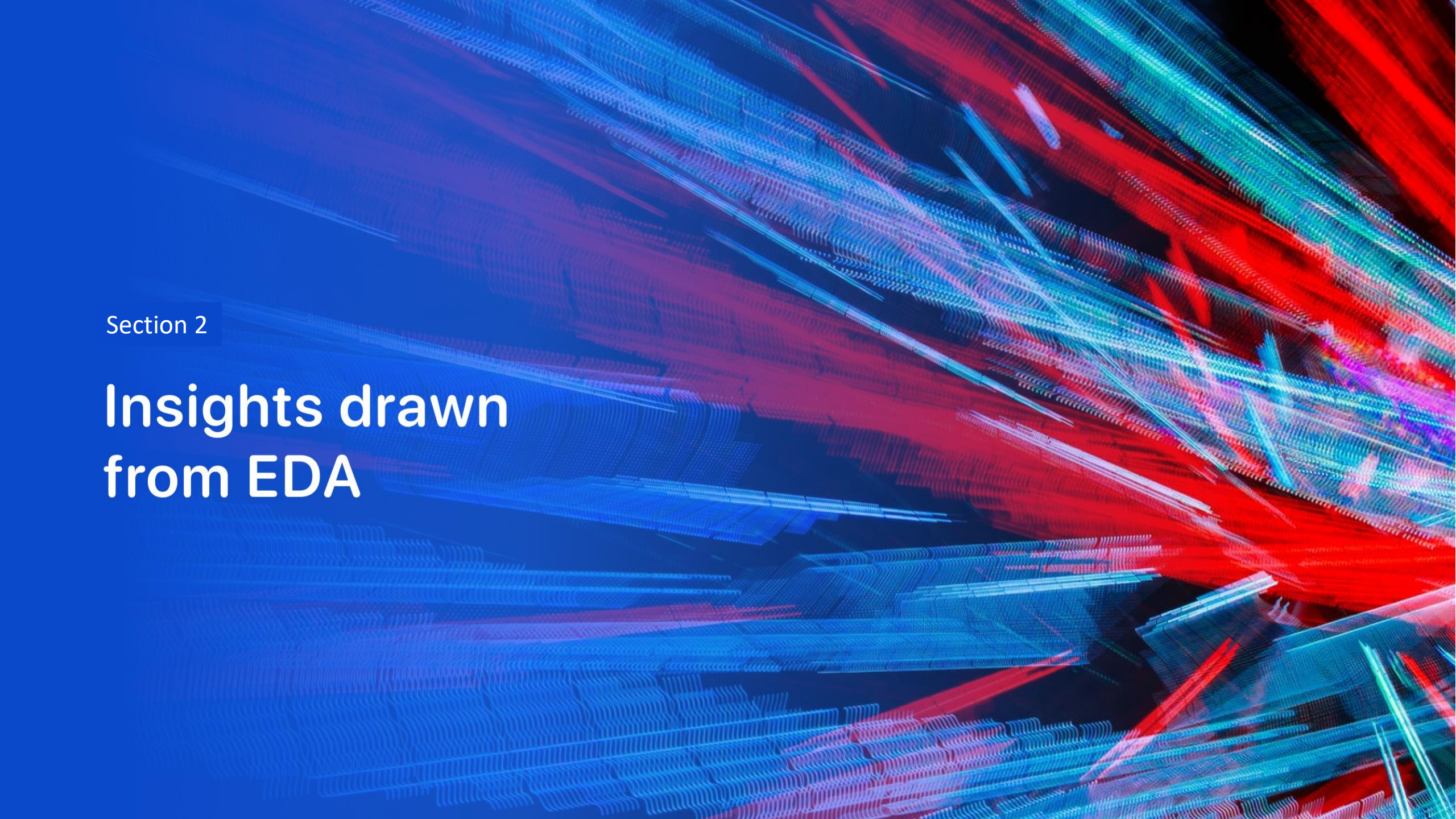
- Interactive analytics demo in screenshots

- Using interactive analytics was possible to identify that launch sites use to be in safety places, near sea, for example and have a good logistic infrastructure around. Most launches happens at east cost launch sites.

- Predictive analysis results

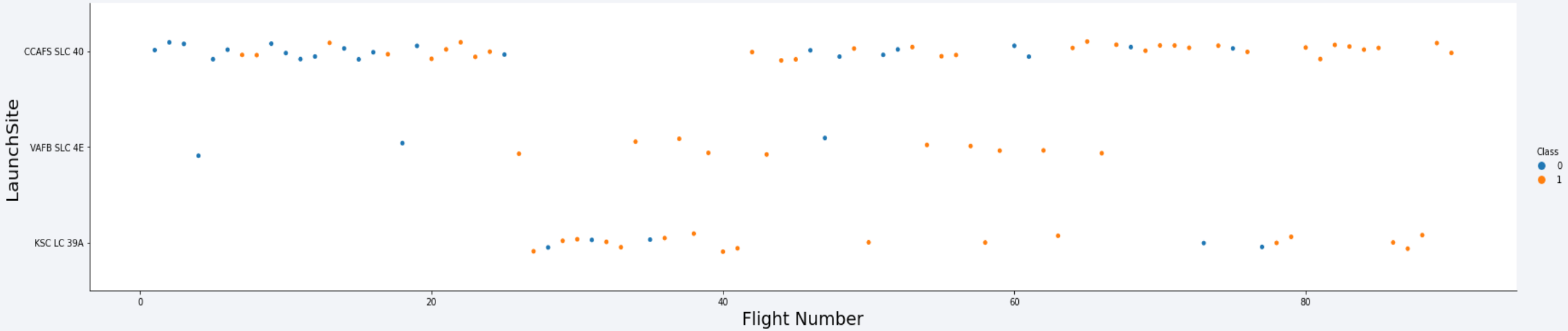
- Decision Tree Classifier is the best model to predict successful landings, over 87% and accuracy for test data over 94%.




Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

- Show a scatter plot of Flight Number vs. Launch Site



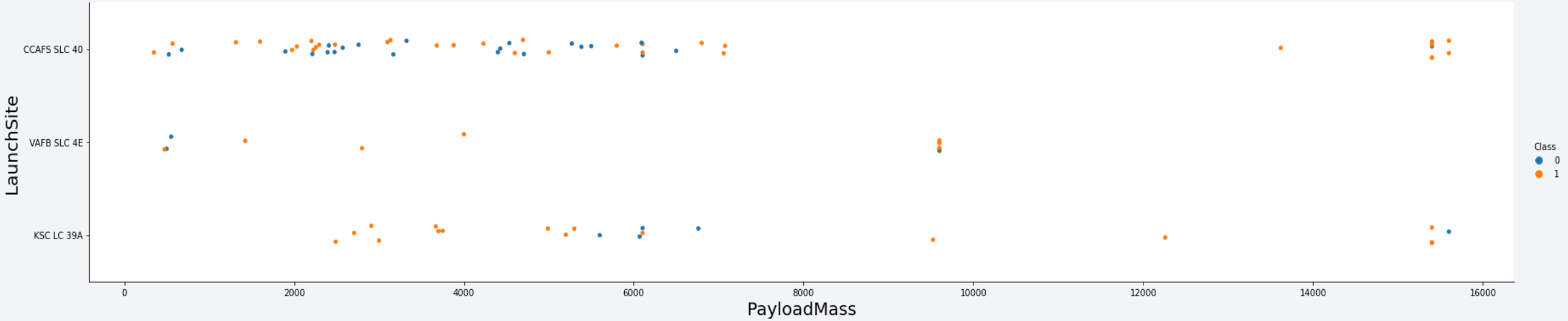
- Show the screenshot of the scatter plot with explanations

- It's possible to verify that the best launch site nowadays is CCAF5 SLC 40, second place VAFB SLC 4E and third place KSC LC 39A;

- The general success rate improved over time.

Payload vs. Launch Site

- Show a scatter plot of Payload vs. Launch Site

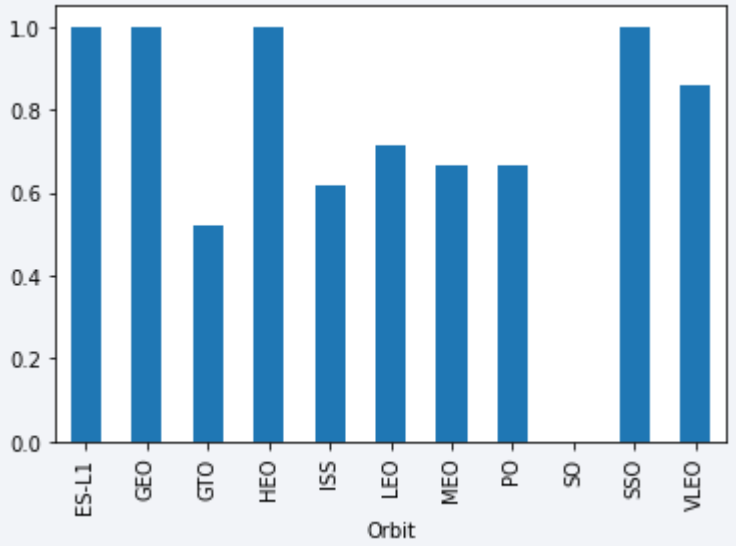


- Show the screenshot of the scatter plot with explanations

- Payloads over 9,000kg (about the weight of a school bus) have excellent success rate;
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.

Success Rate vs. Orbit Type

- Show a bar chart for the success rate of each orbit type

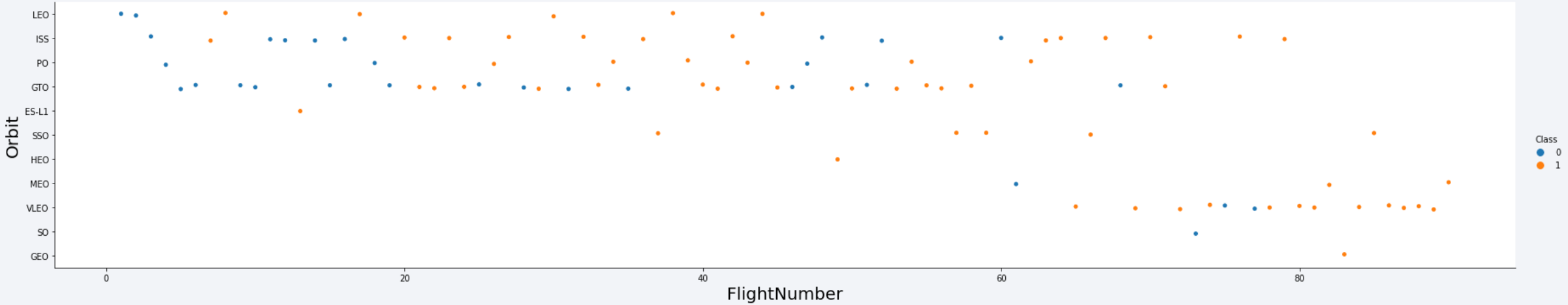


- Show the screenshot of the scatter plot with explanations

- The biggest success rates happens to orbits ranking:
- ES-L1;
- GEO;
- HEO;
- SSO;
- VLEO (above 80%);
- LFO (above 70%).

Flight Number vs. Orbit Type

- Show a scatter point of Flight number vs. Orbit type

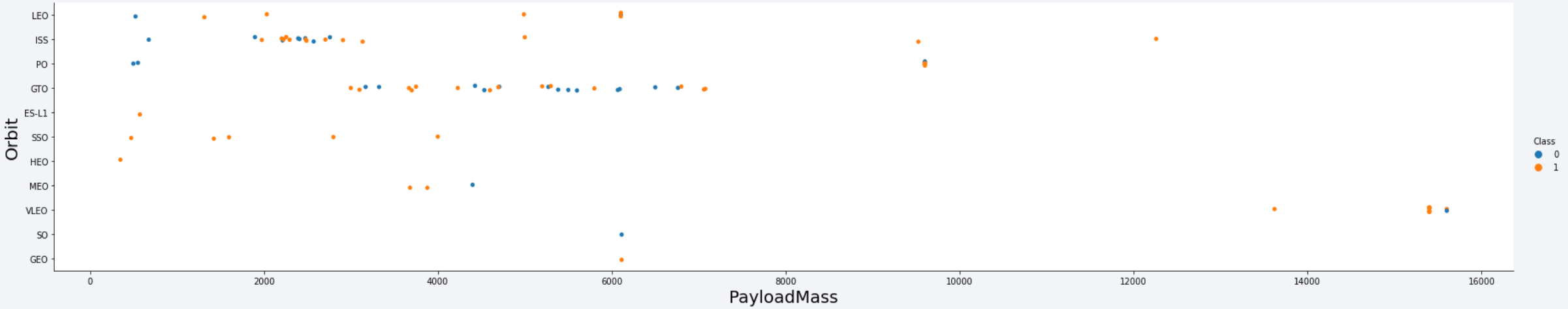


- Show the screenshot of the scatter plot with explanations

- Success rate improved over time to all orbits.
- VLEO orbit seems a new business opportunity, due to recent increase of its frequency
- For most orbits (LEO, ISS, PO, SSO, MEO, VLEO) successful landing rates appear to increase with flight numbers
- There is no relationship between flight number and orbit for GTO

Payload vs. Orbit Type

- Show a scatter point of payload vs. orbit type

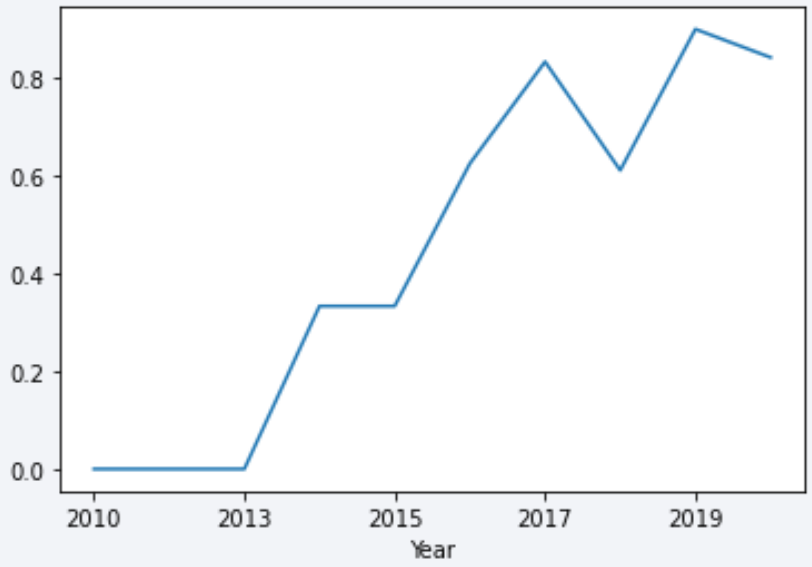


- Show the screenshot of the scatter plot with explanations

- There is no relation between payload and success rate to orbit GTO
- ISS orbit has the widest range of payload and a good rate of success
- There are few launches to the orbits SO and GEO.

Launch Success Yearly Trend

- Show a line chart of yearly average success rate



- Show the screenshot of the scatter plot with explanations

- Success rate (Class=1) increased by about 80% between 2013 and 2020
- Success rates remained the same between 2010 and 2013 and between 2014 and 2015
- Success rates decreased between 2017 and 2018 and between 2019 and 2020

All Launch Site Names

- Find the names of the unique launch sites

	Launch Site	Lat	Long
0	CCAFS LC-40	28.562302	-80.577356
1	CCAFS SLC-40	28.563197	-80.576820
2	KSC LC-39A	28.573255	-80.646895
3	VAFB SLC-4E	34.632834	-120.610746

- Present your query result with a short explanation here
- here is no relation between payload and success rate to orbit GTO;
  • ISS orbit has the widest range of payload and a good rate of success;
  • There are few launches to the orbits SO and GEO.

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Present your query result with a short explanation here

Display 5 records where launch sites begin with the string 'CCA'

```
sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

```
 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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
- Present your query result with a short explanation here

Display the total payload mass carried by boosters launched by NASA (CRS)

```
sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE PAYLOAD  LIKE '%CRS%';
```

```
 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases
Done.
```

total_payload
111268

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Present your query result with a short explanation here

Display average payload mass carried by booster version F9 v1.1

```
sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
```

```
 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.app
Done.
```

avg_payload
2928

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Present your query result with a short explanation here

List the date when the first successful landing outcome in ground pad was acheived.

Hint:Use min function

```
sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Success (ground pad)';
```

```
 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain
Done.
```

first_success_gp
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
- Present your query result with a short explanation here

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 DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND LANDING__OUTCOME = 'Success (drone ship)';
```

```
 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Present your query result with a short explanation here

List the total number of successful and failure mission outcomes

```
sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;
```

```
 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.clou
Done.
```

mission_outcome	qty
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL) ORDER BY BOOSTER_VERSION;
```

```
 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Present your query result with a short explanation here

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND DATE_PART('YEAR', DATE) = 2015;
```

```
 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
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

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(*) AS QTY FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING__OUTCOME ORDER BY QTY DE
```

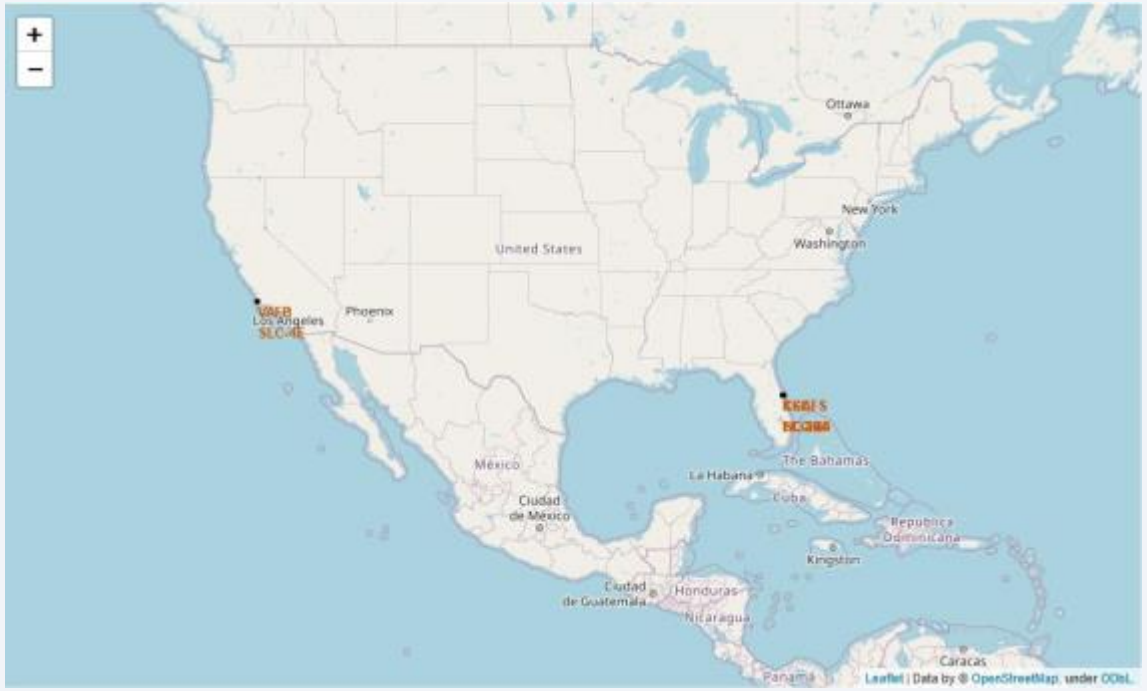
```
 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

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

Section 3

Launch Sites Proximities Analysis

SpaceX Falcon9 - Launch Sites Map



- Explore the generated folium map and make a proper screenshot to include all launch sites' location markers on a global map

- Each launch site contains a circle, label, and a popup to highlight the location and the name of the launch site.

- all launch sites are near the coast.

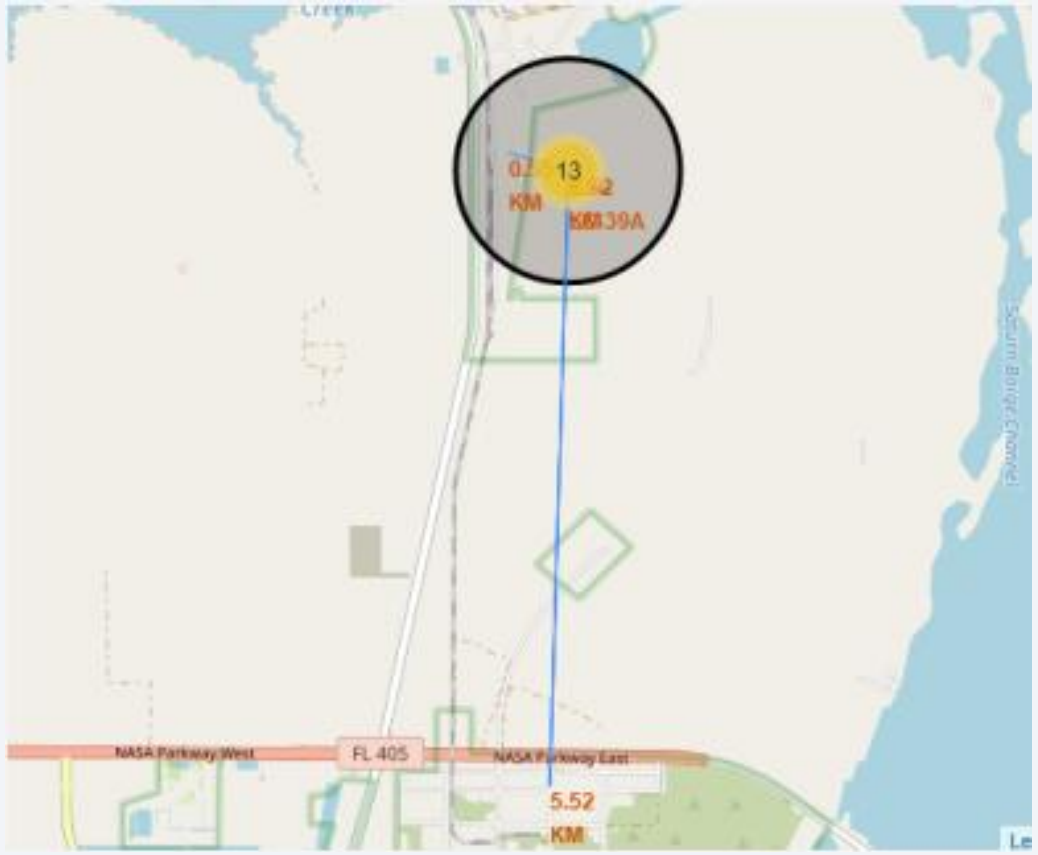
Launch Outcomes by Site

- Explain the important elements and findings on the screenshot

- The US map with all the Launch Sites. The numbers on each site depict the total number of successful and failed launches

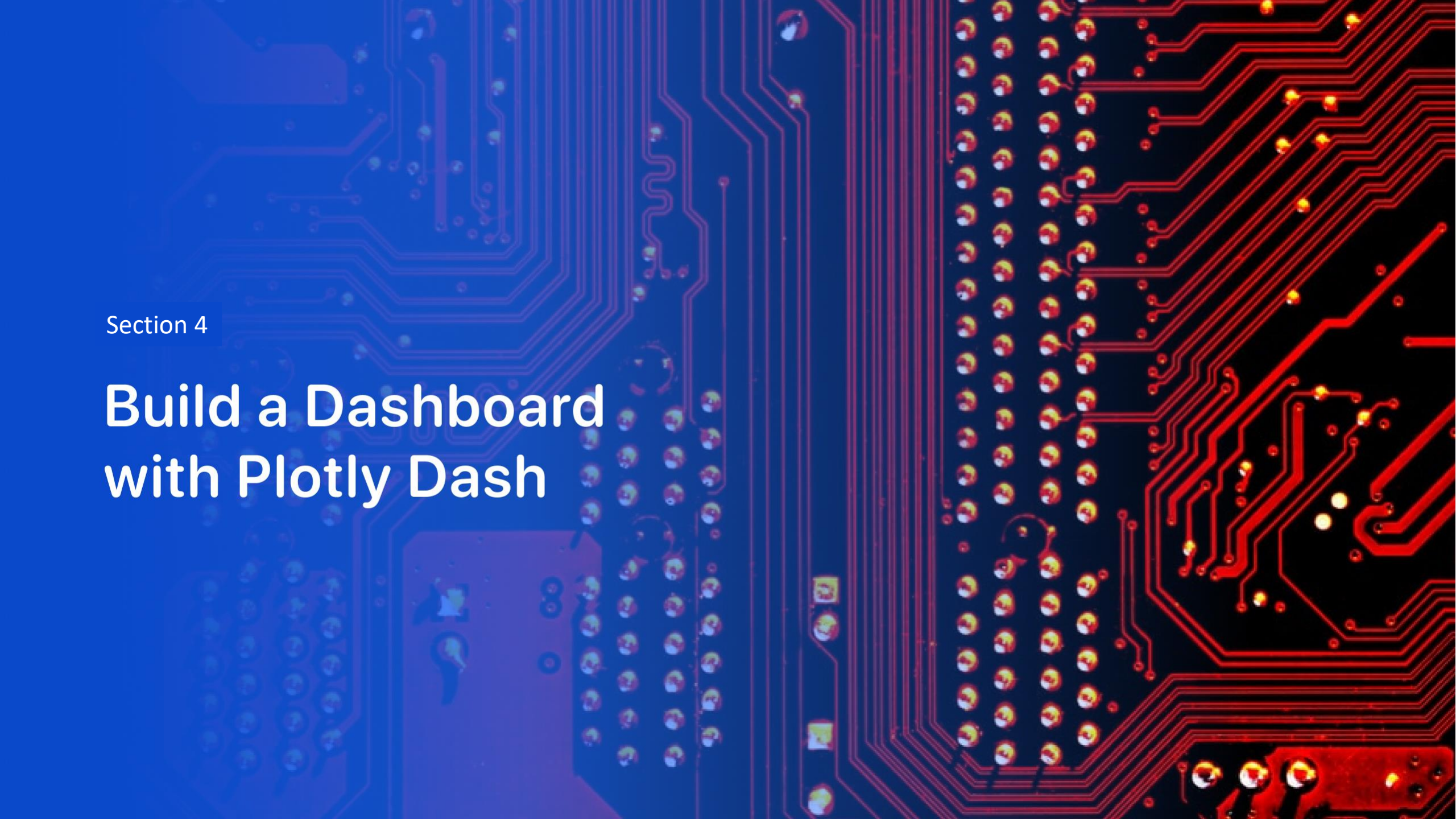
- By looking at each site map, KSC LC-39A Launch Site has the greatest number of successful launches

Logistics and Safety

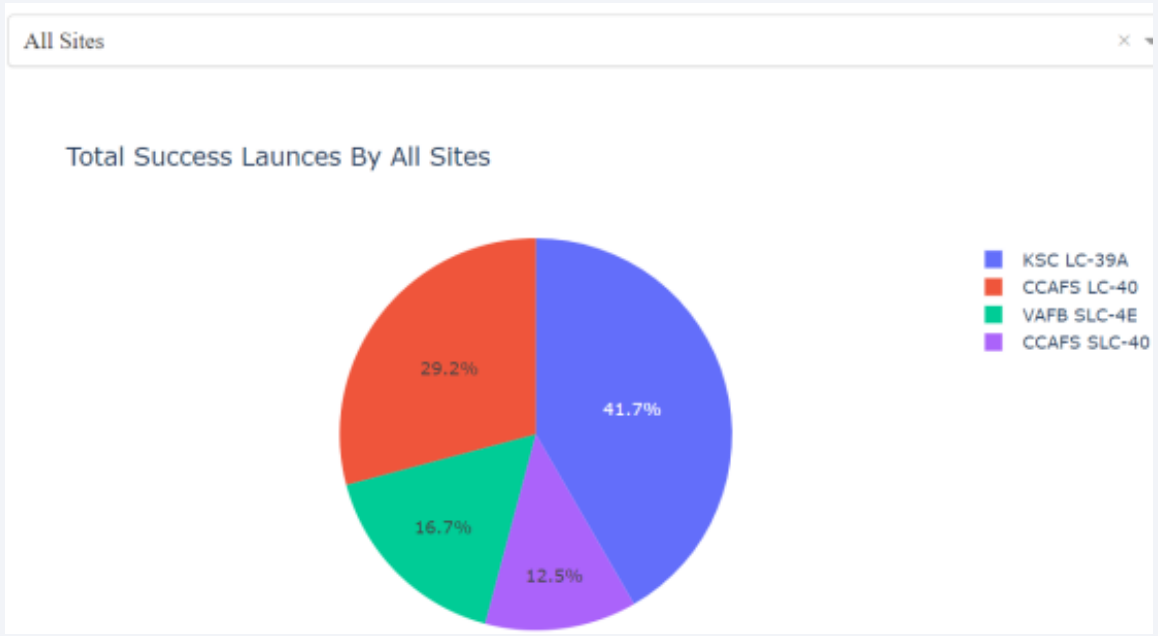


- Explain the important elements and findings on the screenshot

- Launch site KSC LC-39A has good logistics aspects, being near railroad and road and relatively far from inhabited areas


Section 4
Build a Dashboard with Plotly Dash

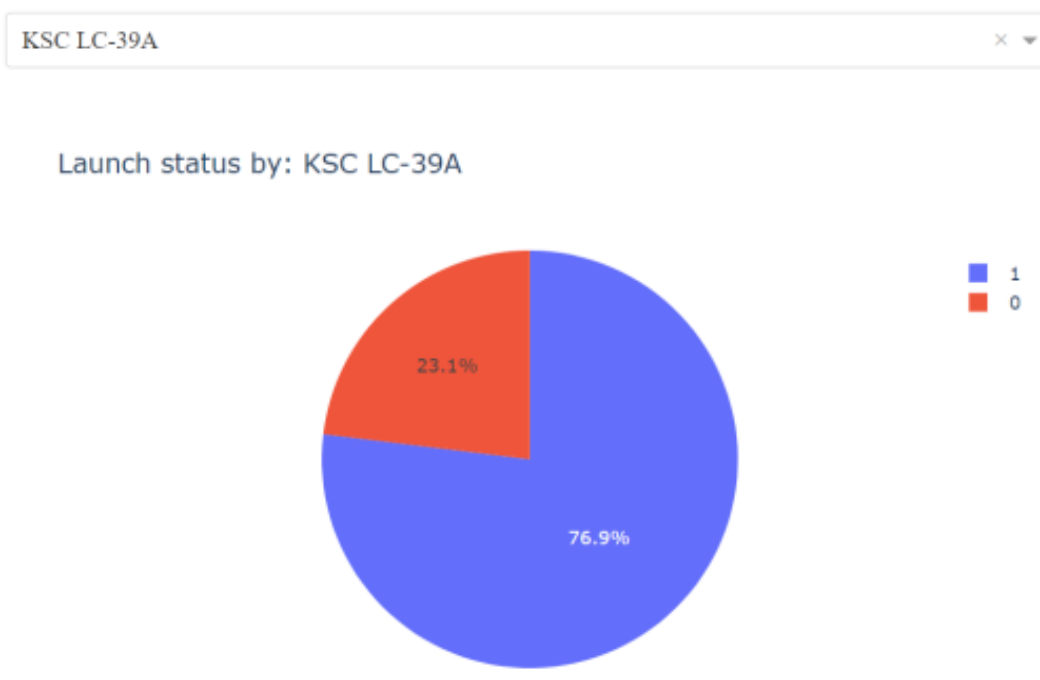
Launch Success Counts For All Sites



- Explain the important elements and findings on the screenshot

- Launch Site ‘KSC LC-39A’ has the highest launch success rate
- Launch Site ‘CCAFS SLC-40’ has the lowest launch success rate

Launch Site with Highest Launch Success Ratio



- Explain the important elements and findings on the screenshot

- KSC LC-39A Launch Site has the highest launch success rate and count

- Launch success rate is 76.9%, Launch success failure rate is 23.1%

Payload vs. Launch Outcome Scatter Plot for All Sites

- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.

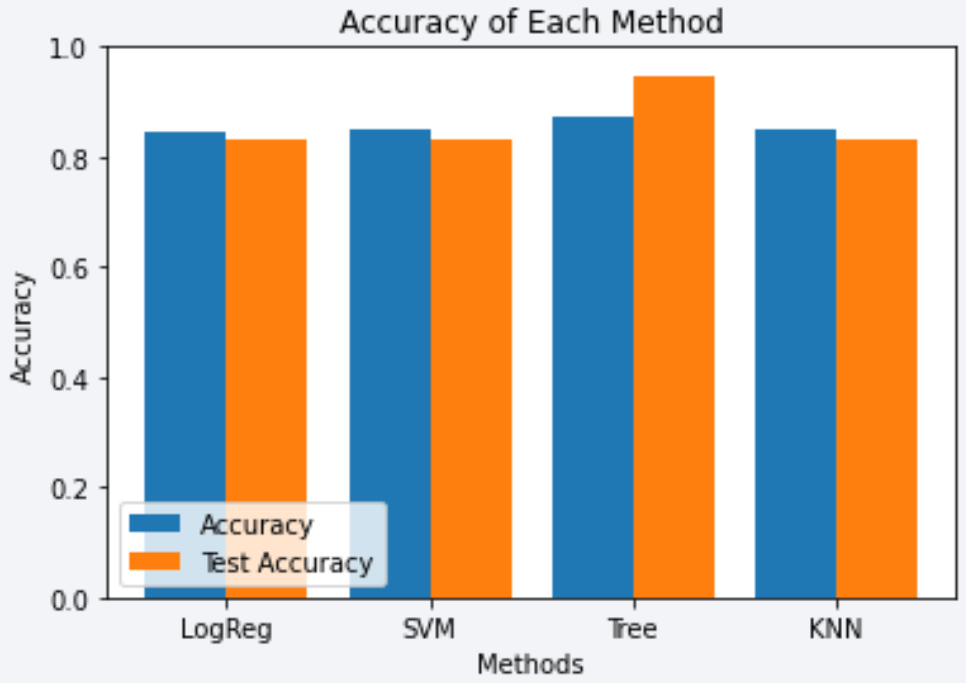
- Most successful launches are in the payload range from 2000 to about 5500
- Booster version category ‘FT’ has the most successful launches
- Only booster with a success launch when payload is greater than 6k is ‘B4’

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Visualize the built model accuracy for all built classification models, in a bar chart



- Find which model has the highest classification accuracy

- 4 classification models were tested, and their accuracies are plotted beside
- The model with the highest classification accuracy is Decision Tree Classifier, which has accuracies over than 87%

Confusion Matrix

- Show the confusion matrix of the best performing model with an explanation

• The confusion matrix is same for all the models (LR, SVM, Decision Tree, KNN)
• Per the confusion matrix, the classifier made 18 predictions
• 12 scenarios were predicted Yes for landing, and they did land successfully (True positive)
• 3 scenarios (top left) were predicted No for landing, and they did not land (True negative)
• 3 scenarios (top right) were predicted Yes for landing, but they did not land successfully (False positive)
• Overall, the classifier is correct about 83% of the time ((TP + TN) / Total) with a misclassification or error rate ((FP + FN) / Total) of about 16.5%

Conclusions

- Different data sources were analyzed, refining conclusions along the process
- Launches above 7,000kg are less risky
- Although most of mission outcomes are successful, successful landing outcomes seem to improve over time, according the evolution of processes and rockets
- Launch Site ‘KSC LC-39A’ has the highest launch success rate and Launch Site ‘CCAFS SLC-40’ has the lowest launch success rate
- Orbits ES-L1, GEO, HEO, and SSO have the highest launch success rates and orbit GTO the lowest
- Lunch sites are located strategically away from cities and closer to coastline, railroads & highways
- The best performing Machine Learning Classfication Model is the Decision Tree with an accuracy of about 87.5%.

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!