IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name> Christian Gollee
<Date> 07/04/2025

Outline

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

Executive Summary

- **Summary of methodologies:**

Data Collection: Retrieved SpaceX launch data through API/web scraping and CSV datasets.Extracted key features such as launch site, payload mass,etc.

Data Wrangling: Cleaned and structured the data for analysis.Handled missing values, formatted columns and created new features like class.

Exploratory Data Analysis(EDA): Used visualizations (pie charts,scatter plots) to identify patterns and trends.

Visualization: Created an interactive dashboard using Dash and Plotly to explore successful launch sites.

Feature Engineering: Converted categorical variables using One-Hot Encoding.

Model Development: Built 4 classification models: Logistic Regression, Decision Tree, KNN, and SVM.

Used GridSearchCV to tune hyperparameters .Split data into training and test sets for evaluation

Model Evaluation: Compared models based on accuracy scores. selected the best performing model.

- **Summary of all results:**

Plotted pie chart and bar chart to analyze success vs. failure rates.

KSC LC-39A had a higher proportion of successful landing compared to others.

Built four ML models and Best model was **Decision Tree classification** with Accuracy Score of **89%** on test data.

Introduction

❖Project background and context:

- **Predicting Falcon 9 First Stage Landing Success**
- Successful landings = cost savings, efficiency, and competitive advantage
- SpaceX offers Falcon 9 launches at a competitive cost of $62 million, largely due to the reusability of the first stage.
- Competing rocket providers charge up to $165 million for a launch.

❖Problems you want to find answers:

- To predict whether the Falcon 9 first stage will land successfully using machine learning models
- This prediction could help competitors estimate costs, assess SpaceX's reliability, and make informed bids for satellite launches.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

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

Data Collection – SpaceX API

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

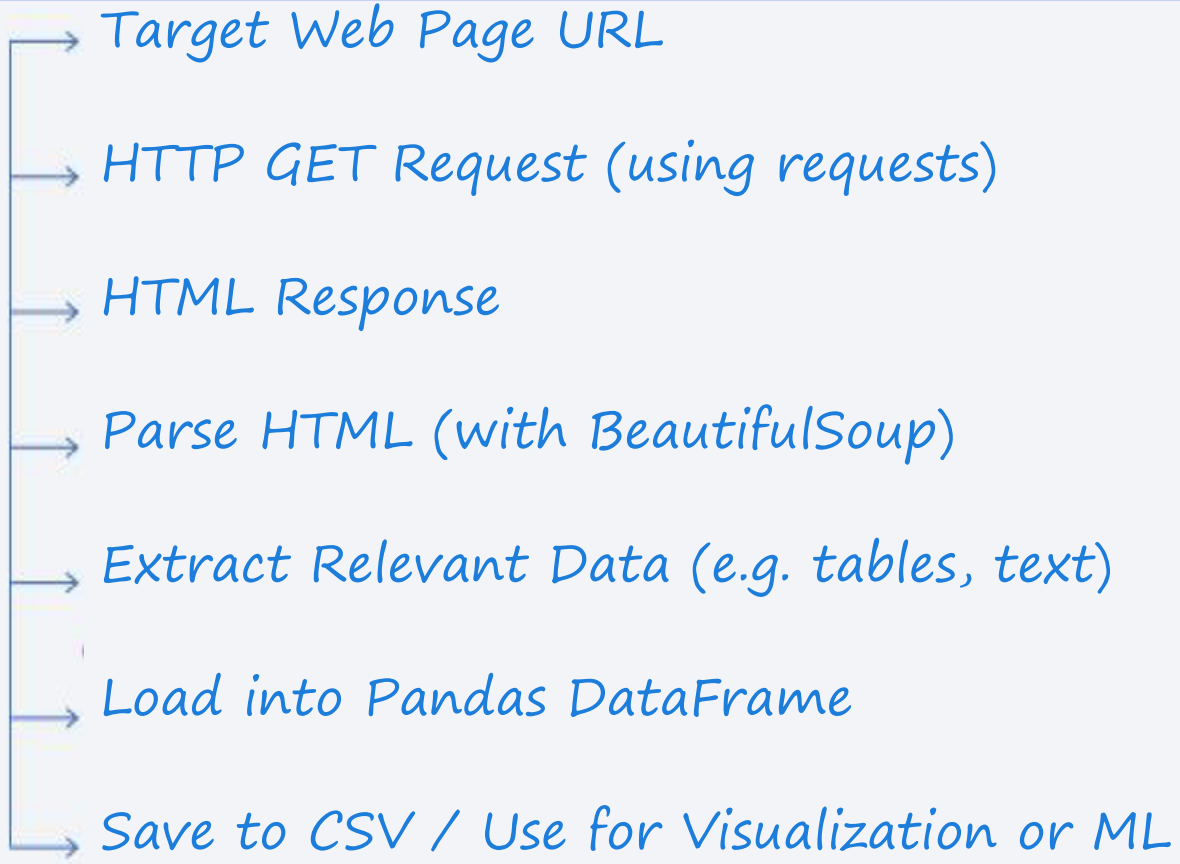
- Add the GitHub URL of the completed SpaceX API calls notebook (must include completed code cell and outcome cell), as an external reference and peer-review purpose

GitHub:
URL="https://github.com/Ross1911/DataScience_capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb"

Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts

- Add the GitHub URL of the completed web scraping notebook, as an external reference and peer-review purpose



GitHub:
URL="https://github.com/Ross1911/DataScience_capstone/blob/main/jupyter-labs-webscraping.ipynb"

Data Wrangling

Describe how data were processed

You need to present your data wrangling process using key phrases and flowcharts

Add the GitHub URL of your completed data wrangling related notebooks, as an external reference and peer-review purpose

GitHub:
URL="https://github.com/Ross1911/DataScience_capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb"

EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
- Flight Number vs. Launch Site Scatter Plot

Why: To observe the success/failure trend across different launch sites over time.

- Payload vs. Launch Site Scatter Plot

Why: To explore if payload mass impacted success rates at various sites.

- Success Rate by Orbit (Bar Chart)
- Why:Why: To analyze which orbit types were most associated with successful landings.
- Average Success Rate by Year (Line Plot)
- Why: To track SpaceX's progress in landing success over the years.
- Add the GitHub URL of your completed EDA with data visualization notebook, as an external reference and peer-review purpose:

- "**https://github.com/Ross1911/DataScience_capstone/blob/main/edadataviz.ipynb**"

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed

---

- First SQL query used to get unique launch sites
- Next query used to get name of launch sites begin with `CCA`
- Then to get total payload carried by boosters from NASA
- Then to get average payload mass carried by booster version F9 v1.1
- Then to get dates of the first successful landing outcome on ground pad
- Then to get names of boosters successed on drone ship and payload mass between 4000 and 6000
- Then to get total number of successful and failure mission outcomes
- Then to get names of the booster which have carried the maximum payload mass
- Then to get failed landing_outcomes in drone ship
- Then to get Ranking Landing Outcomes Between 2010-06-04 and 2017-03-20
- Add the GitHub URL of your completed EDA with SQL notebook, as an external reference and peer-review purpose:

- **"https://github.com/Ross1911/DataScience_capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb"**

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map:
- First i mark launch sites using markers
- Then i created circle for each launch sites
- Then i created distance line for each launch sites connected to highway,railway and coastline of each launch sites.
- Explain why you added those objects:
- Added markers to know their coordinates(latitude,longitude)
- Added a circle to highlight the launch sites
- Added distance line to calculate distance between highway,railway,coastline and launch sites
- Add the GitHub URL of your completed interactive map with Folium map, as an external reference and peer-review purpose:
- **"https://github.com/Ross1911/DataScience_capstone/blob/main/lab_jupyter_launch_site_location(folium).ipynb"**

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard:
- First i add pie chart
- Then a Scatter plot
- Then a Payload Range slider
- Created a drop-down button for launch sites
- Explain why you added those plots and interactions:
- Added pie chart to get successfull launch sites
- Added Scatter plot to known is there any relation between payload mass and success rate
- Added Payload Range slider so user can select spacific payload mass
- Added drop-down button so user can select spacific launch site.
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose:
- **"https://github.com/Ross1911/DataScience_capstone/blob/main/spacex-dash-app.py"**

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
- You need present your model development process using key phrases and flowchart
- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose
- **“https://github.com/Ross1911/DataScience_capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb”**

Feature Scaling (Standardization)

↓

Model Selection (e.g., Logistic Regression, Decision Tree, etc.)

↓

Model Training

↓

Model Evaluation (Accuracy, ConfusionMatrix, Precision_score,etc)

↓

Hyperparameter Tuning (e.g., GridSearchCV)

↓

Model Comparison (Bar chart)

↓

Best Model Selection

Results

- Exploratory data analysis results
- Plotted pie chart and bar chart to analyze success vs. failure rates.
- Observed that majority of launches were successful (~75%+ success rate).
- Pie charts showed variation in success rates across different launch sites.
- For example, KSC LC-39A had a higher proportion of successful landing compared to others.
- Scatter plots showed a slight correlation between No. of flights, payload mass and launch success.
- Added color to scatter plots using booster version category to understand its impact
- Certain Booster version category(e.g,FT) with low payload range have higher chance of successful landing
- Interactive analytics demo in screenshots
- Predictive analysis results
- Build four Machine Learning models Logestic Regression, SVM, Decision Tree classification and KNN.
- Models were elvaluated on Accuracy Score, Confusion Matrix, Precesion Score etc.
- Best model was **Decision Tree classification** with Accuracy Score of **89%** on test data.
- Other models Logestic Regression, SVM, KNN have Accuracy score of 83%,83%,83% respectively.

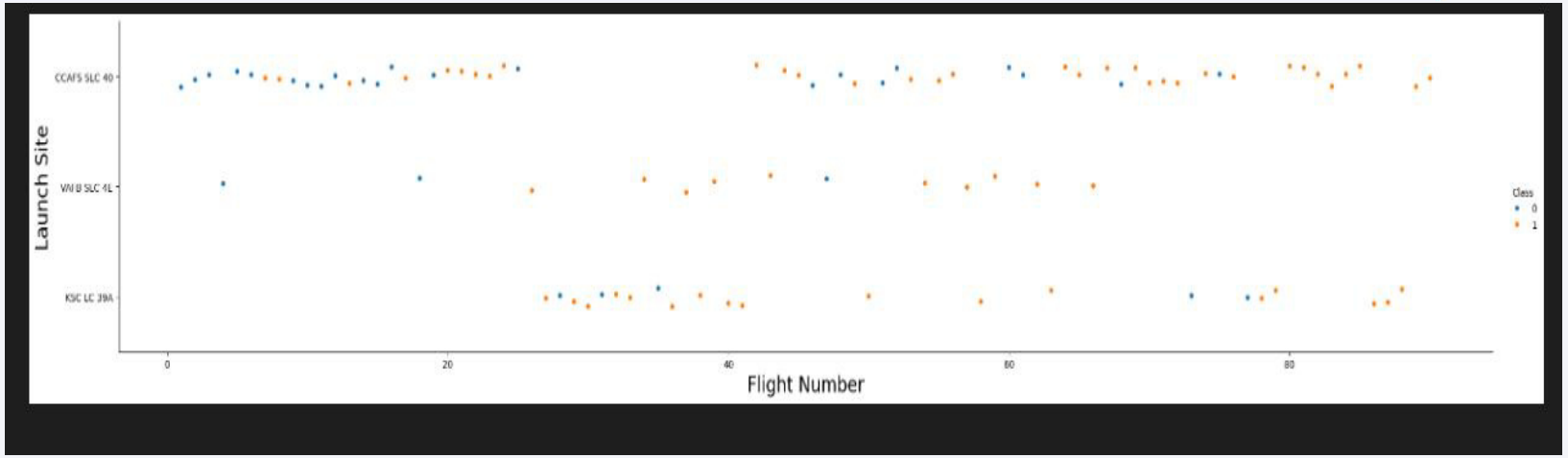
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

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

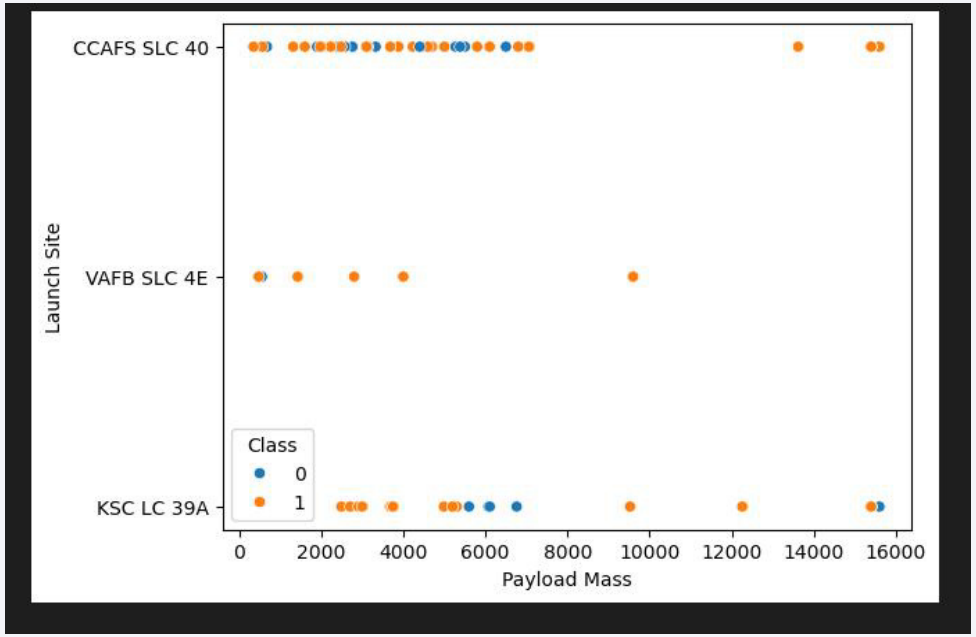
- Show the screenshot of the scatter plot with explanations



- The site CCAFS SLC 40 has the highest number of launches.
- KSC LC 39A and VAFB SLC 4E have comparatively fewer launches.
- Success rates (class = 1) appear to increase over flight number, especially for launch site(CCAFS SLC 40).
- KSC LC 39A and CCAFS SLC 40 have more consistent success rates in later flights.
- VAFB SLC 4E shows fewer launches, making it harder to conclude a strong pattern there.
- In earlier flight numbers, blue dots (class = 0) are more common, showing that initial launches were less successful.
- Recent launches are mostly successful across all sites, especially after flight number ~50.

Payload vs. Launch Site

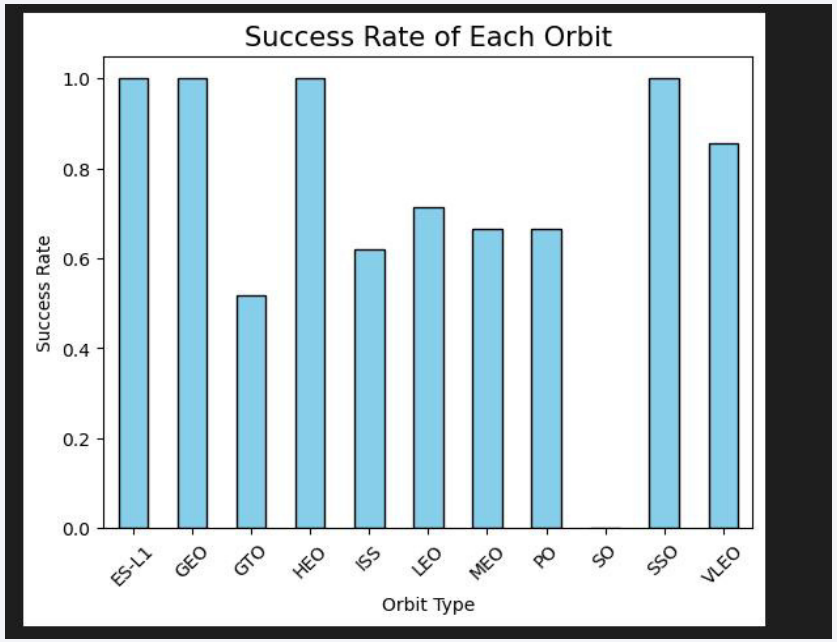
- Show a scatter plot of Payload vs. Launch Site
- Show the screenshot of the scatter plot with explanations



- No Clear Correlation Between Payload Mass and Success.
- Launches with a wide range of payload masses (from ~0 to 16000 kg) have had both successes (orange dots) and failures (blue dots).
- This suggests that payload mass alone is not a strong predictor of launch success.
- The majority of launches occurred at CCAFS SLC 40, spanning almost the entire payload range.
- Launches at VAFB SLC 4E are fewer and generally involve moderate payloads (around 2000–10000 kg).
- Site KSC LC 39A is also used for almost entire payload range.

Success Rate vs. Orbit Type

- Show a bar chart for the success rate of each orbit type
- Show the screenshot of the bar plot with explanations



- **Highest Success Rates (100%)**

GEO (Geostationary Orbit)
HEO (Highly Elliptical Orbit)
SSO (Sun-Synchronous Orbit)
ES-L1 (Earth-Sun Lagrange Point 1)
These orbits had no failed missions, showing complete reliability in launches targeting them.

- **Lowest Success Rate**

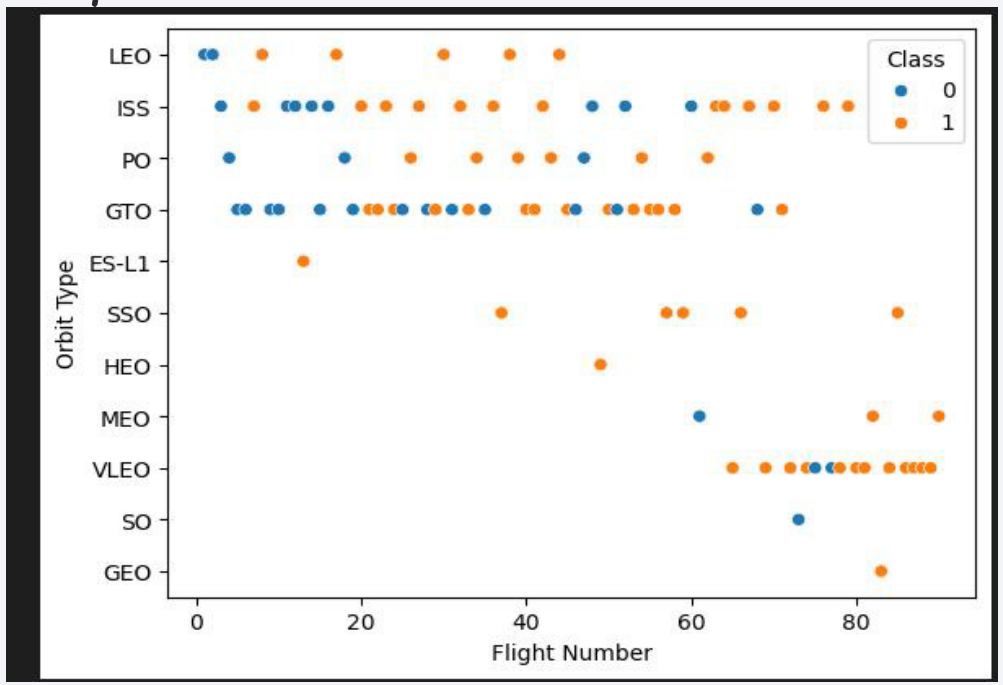
GTO (Geostationary Transfer Orbit) has the lowest success rate (~52%). This suggests it might be more challenging or risk-prone, possibly due to the complexity of reaching this orbit.

- **Moderate Success Rates**

LEO (Low Earth Orbit), ISS (International Space Station), MEO (Medium Earth Orbit), and PO (Polar Orbit) all have moderate success rates (60–70%).
These orbits are frequently used, but show room for improvement.

Flight Number vs. Orbit Type

- Show a scatter point of Flight number vs. Orbit type
- Show the screenshot of the scatter plot with explanations



- **Improvement Over Time**

Across most orbit types (LEO, ISS, GTO, etc.), orange points (success) increase in frequency as the flight number increases.
This implies that SpaceX's landing success improved over time, likely due to tech advancements and experience.

- **GTO Orbit Challenges**

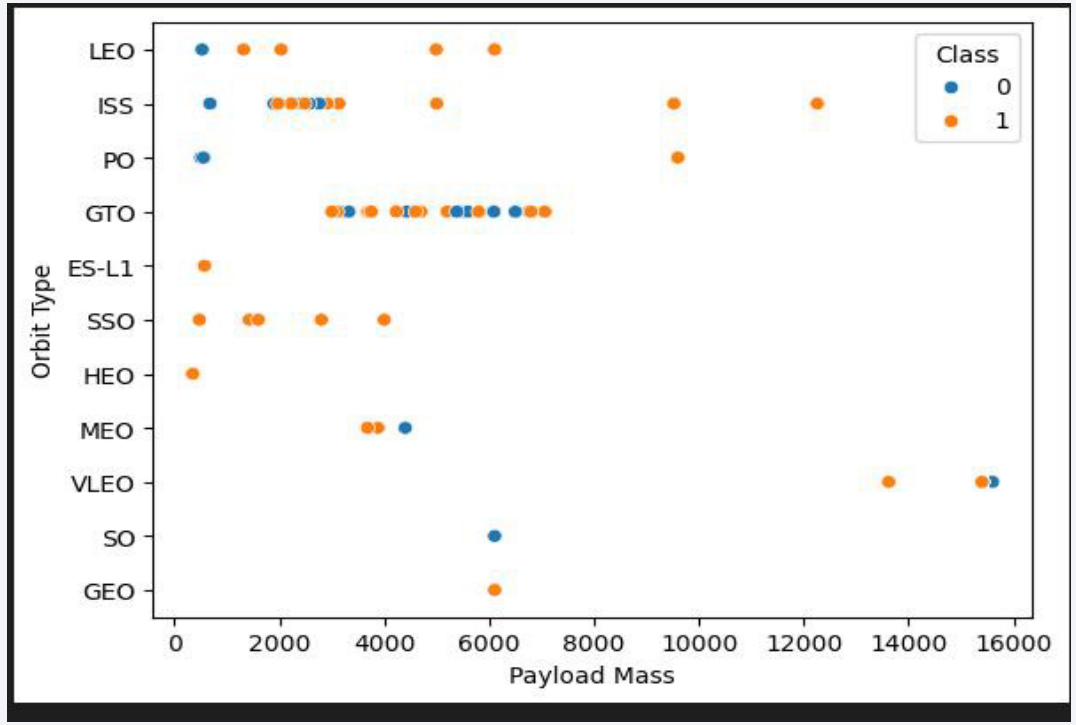
GTO shows a mix of successes and failures, especially in mid-range flight numbers.
Confirms earlier insight: GTO is more technically complex, impacting consistency.

- **High Success in Certain Orbits**

SSO, HEO, VLEO, and ES-L1 mostly show successful landings (orange).
These may be lower risk or better optimized missions.

Payload vs. Orbit Type

- Show a scatter point of payload vs. orbit type
- Show the screenshot of the scatter plot with explanations



- **GTO Orbit Carries Heavier Payloads**

GTO (Geostationary Transfer Orbit) missions often carry payloads between 4000–7000 kg, and both successes and failures are visible. This orbit appears to be more demanding in terms of payload mass and precision.

- **High Payload Doesn't Always Mean Failure**

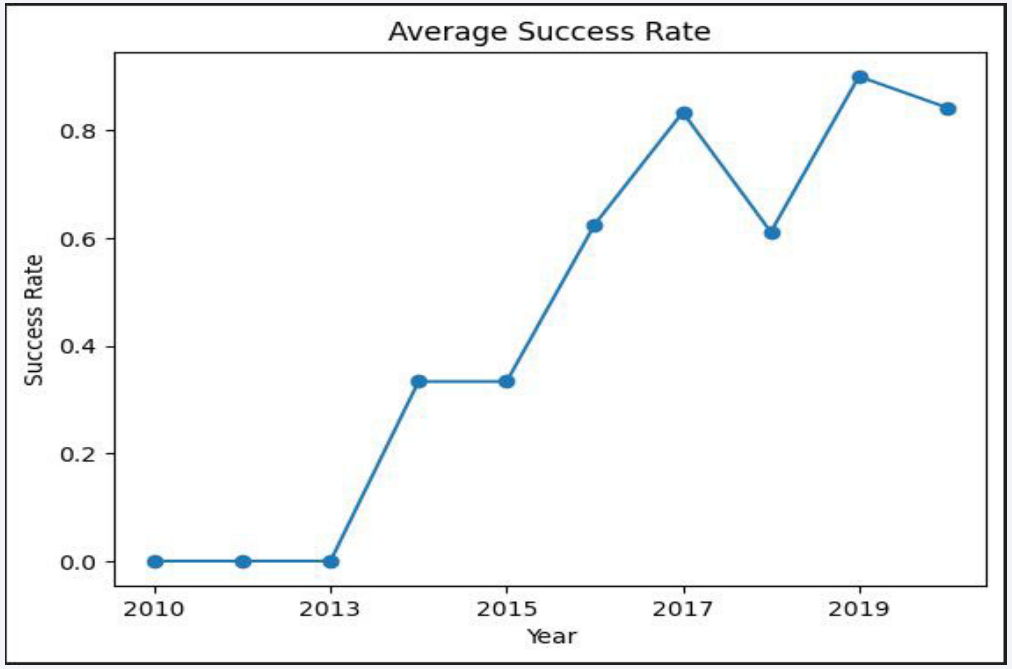
There are successful landings (orange) even with payloads exceeding 10,000 kg, especially in GTO and SO orbits.
Shows that SpaceX technology can handle heavy payloads and still land boosters successfully.

- **Smaller Payloads Mostly Succeed**

In LEO, SSO, and ISS orbits, the payloads are generally lighter (mostly < 5000 kg), and most missions are successful.
Indicates more reliable landings with smaller missions.

Launch Success Yearly Trend

- Show a line chart of yearly average success rate
- Show the screenshot of the line plot with explanations



- **Initial Struggles (2010–2013)**

From 2010 to 2013, the success rate was 0% — indicating no successful landings during the early launch attempts.

- **Breakthrough in 2014**

In 2014, there was a significant jump to a success rate of ~33%, showing the first successful landing attempts began this year.

- **Steady Improvement (2015–2017)**

The success rate held steady in 2015, then rapidly increased in 2016 and peaked in 2017 at over 80%. Indicates rapid learning and technology improvement at SpaceX.

- **Highest Success in 2019**

2019 marked the highest average success rate, nearly 90%, reflecting strong operational efficiency.

All Launch Site Names

- Find the names of the unique launch sites
- CCAFS LC-40
- VAFB SLC-4E
- KSC LC-39A
- CCAFS SLC-40
- Present your query result with a short explanation here:

%sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE;

This query gives us names of unique launch sites

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Present your query result with a short explanation here:
- %sql select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5;

[13]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Ou
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (para
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 (para
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No a
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No a
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No a

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- Total Payload Mass= **45596**
- Present your query result with a short explanation here:
- **%sql select SUM(PAYLOAD_MASS__KG_) as Total_Payload_Mass from SPACEXTABLE where Customer ='NASA (CRS)';**

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Average_Payload_Mass= **2928.4**
- Present your query result with a short explanation here:
- **%sql select AVG(PAYLOAD_MASS__KG_) as Average_Payload_Mass from SPACEXTABLE where Booster_Version ='F9 v1.1';**

First Successful Ground Landing Date

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

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
- F9 FT B1022
- F9 FT B1026
- F9 FT B1021.2
- F9 FT B1031.2
- Present your query result with a short explanation here:
- %sql select distinct Booster_Version from SPACEXTABLE where Landing_Outcome='Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000;

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:
- **%sql select Mission_Outcome, count(*) as Total_Count from SPACEXTABLE group by Mission_Outcome;**

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Present your query result with a short explanation here:
- %sql select distinct 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

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

month_names	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

- Present your query result with a short explanation here:
- %sql select substr(Date,6,2) as month_names, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTABLE where Landing_Outcome like 'Failure (drone ship)' and substr(Date,0,5)='2015';

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

- Present your query result with a short explanation here:

```
%sql with landing_counts as (select Landing_Outcome, count(*) as total_count from SPACEXTABLE where Date between '2010-06-04' and '2017-03-20' group by Landing_Outcome) select Landing_Outcome,total_count,(select count(*) from landing_counts lc2 where lc2.total_count > lc1.total_count) + 1 as rank from landing_counts lc1 order by rank;
```

Landing_Outcome	total_count	rank
No attempt	10	1
Failure (drone ship)	5	2
Success (drone ship)	5	2
Controlled (ocean)	3	4
Success (ground pad)	3	4
Failure (parachute)	2	6
Uncontrolled (ocean)	2	6
Precluded (drone ship)	1	8

Section 3

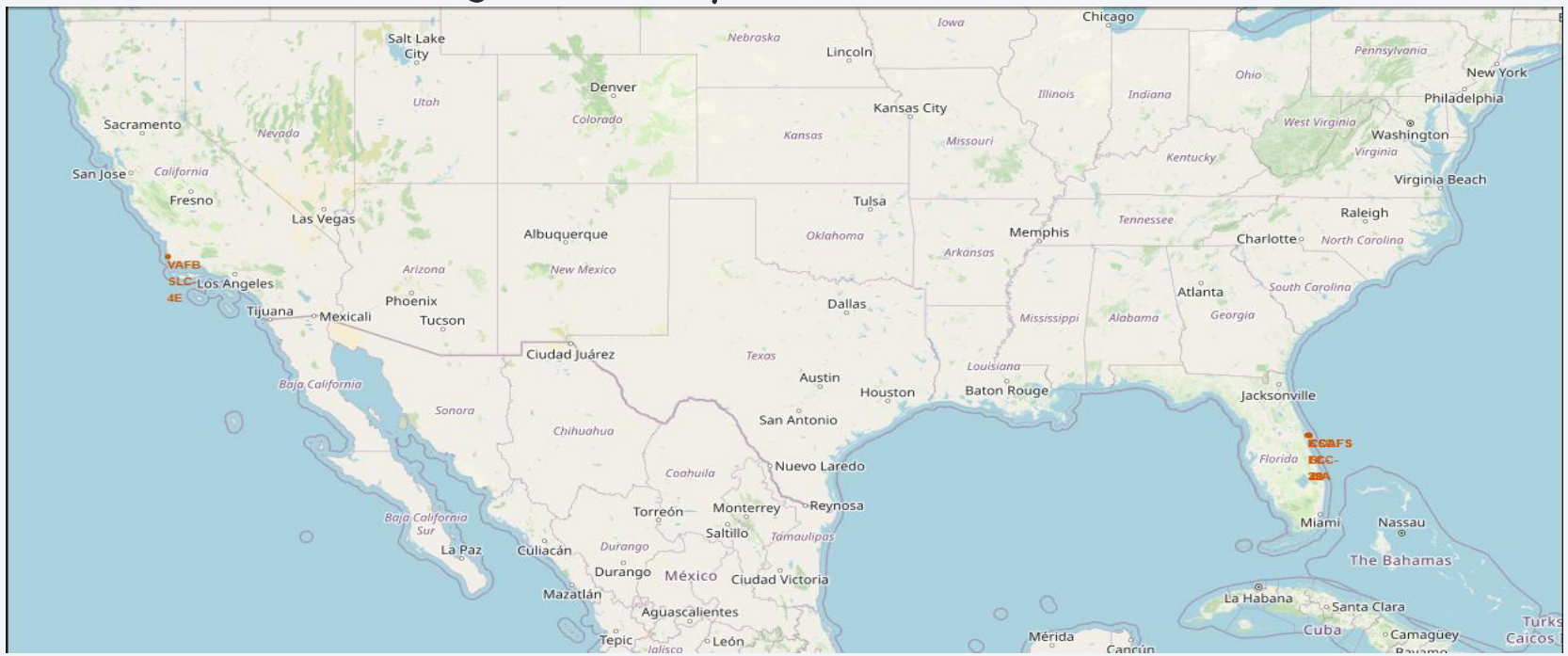
Launch Sites Proximities Analysis

<Folium Map Screenshot 1>

- Replace <Folium map with marked launch sites> title with an appropriate title

---

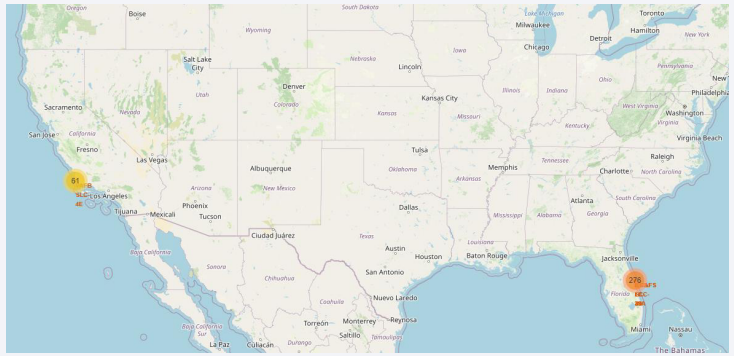
- Explore the generated folium map and make a proper screenshot to include all launch sites' location markers on a global map

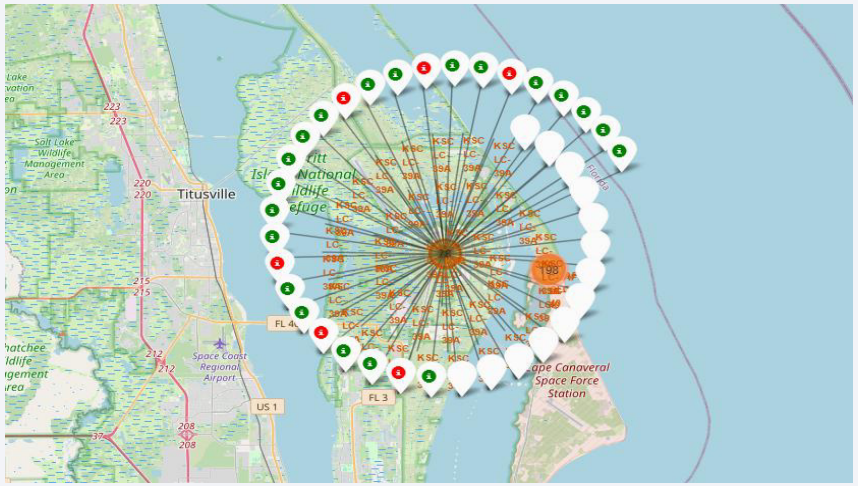


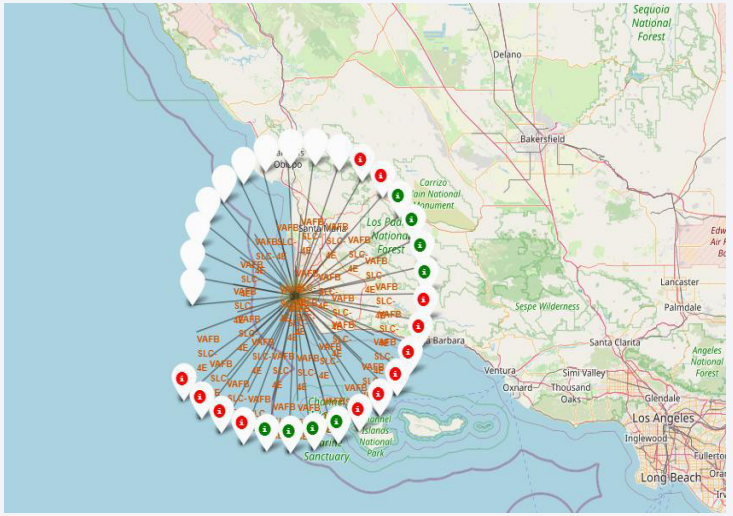
- Explain the important elements and findings on the screenshot:
- From the image we can understand that not all launch sites are in proximity to the Equator line.
- All launch sites are in very close proximity to the coast.
- I have marked and labeled all launch sites through their latitude and longitude coordinates.

<Folium Map Screenshot 2>

- Replace <Folium map with color-labeled launch outcomes> title with an appropriate title
- Explore the folium map and make a proper screenshot to show the color-labeled launch outcomes on the map



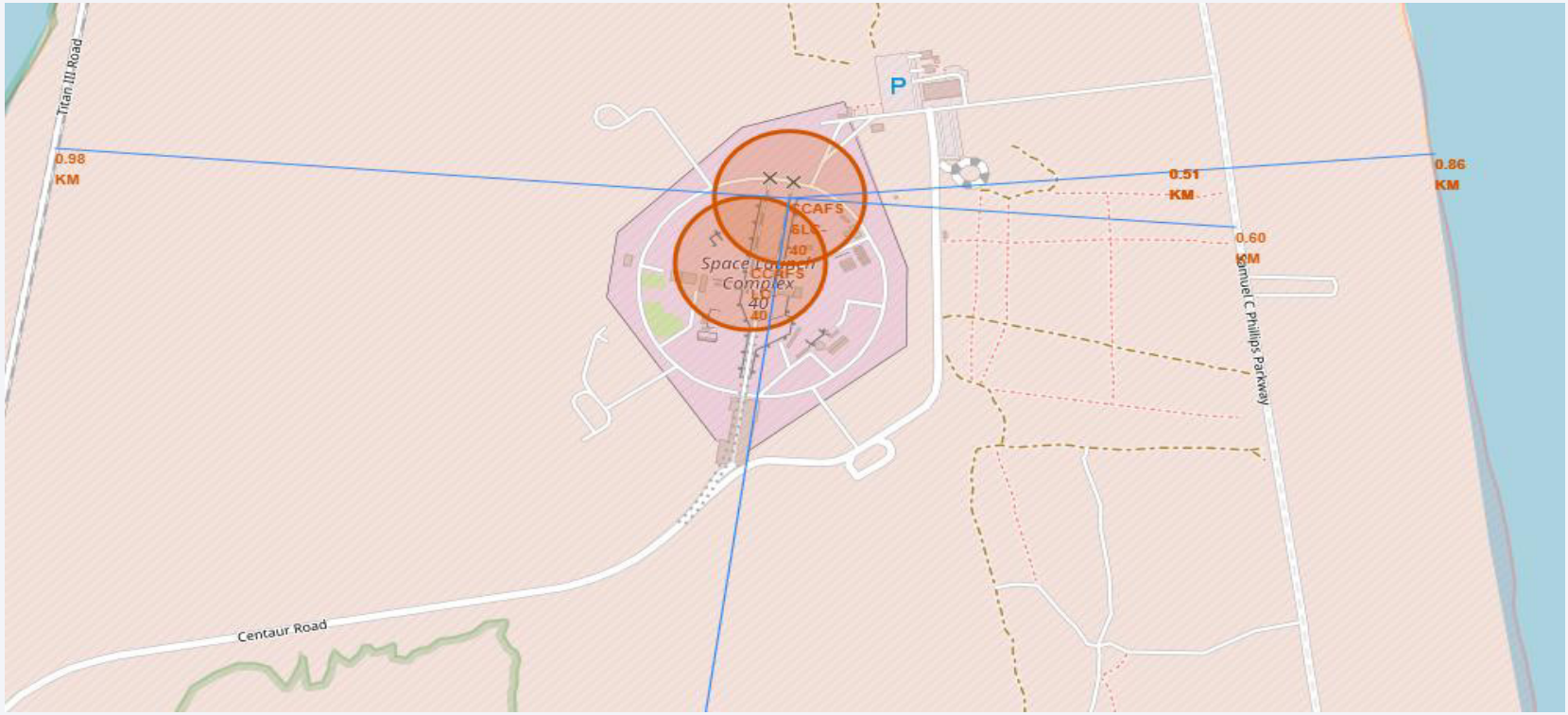




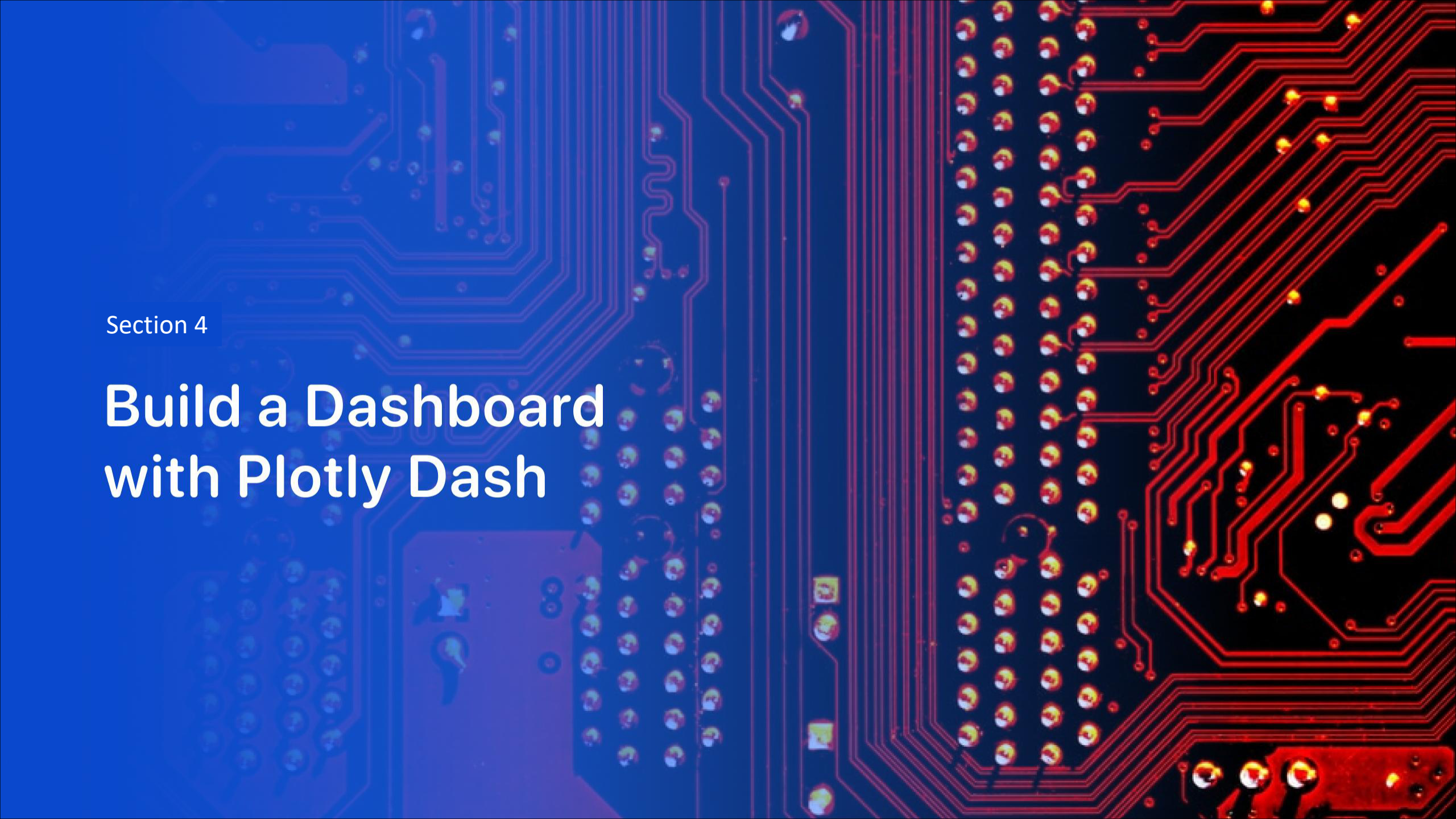
- Explain the important elements and findings on the screenshot
- Marked all sites with color-labeled launch outcomes where launch outcome=successfull=Green and launch outcome=unsuccessfull=Red
- From the map i find that launch site KSC LC-39A has the highest success rate.

<Folium Map Screenshot 3>

- Replace <Folium map of CCAFS SLC-40 launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed> title with an appropriate title
- Explore the generated folium map and show the screenshot of a selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed

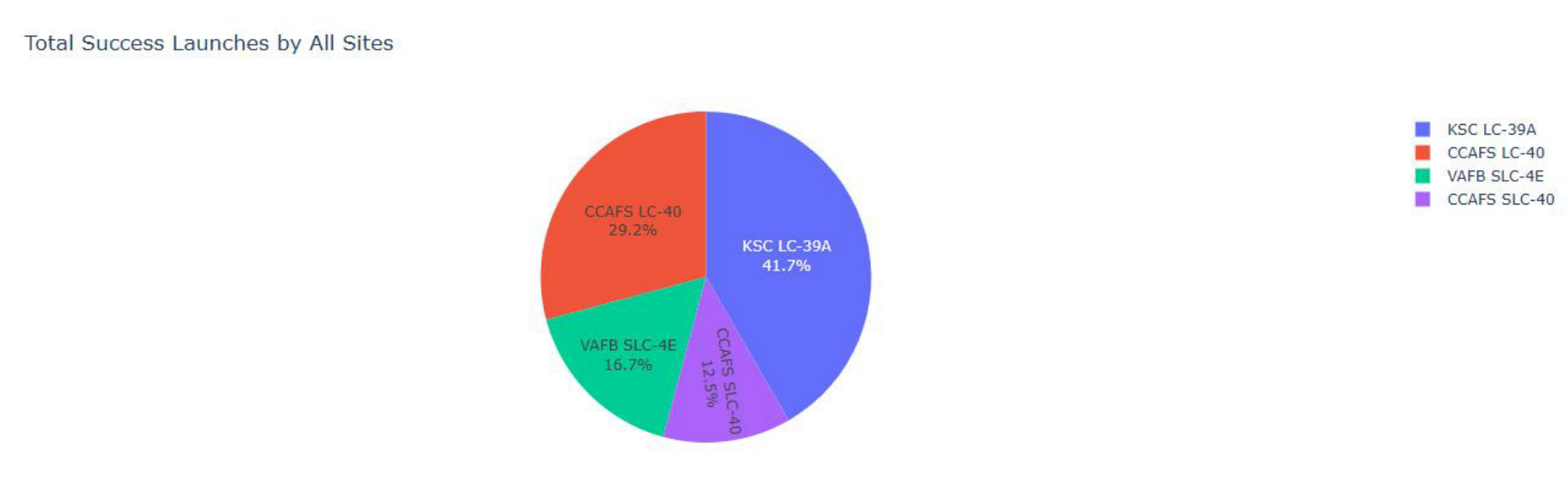


- Explain the important elements and findings on the screenshot
- From the image we can see launch site CCAFS SLC-40 is closest to highway with 0.60KM, then coastline with 0.86 KM and then railway with 0.98 KM


Section 4
Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>

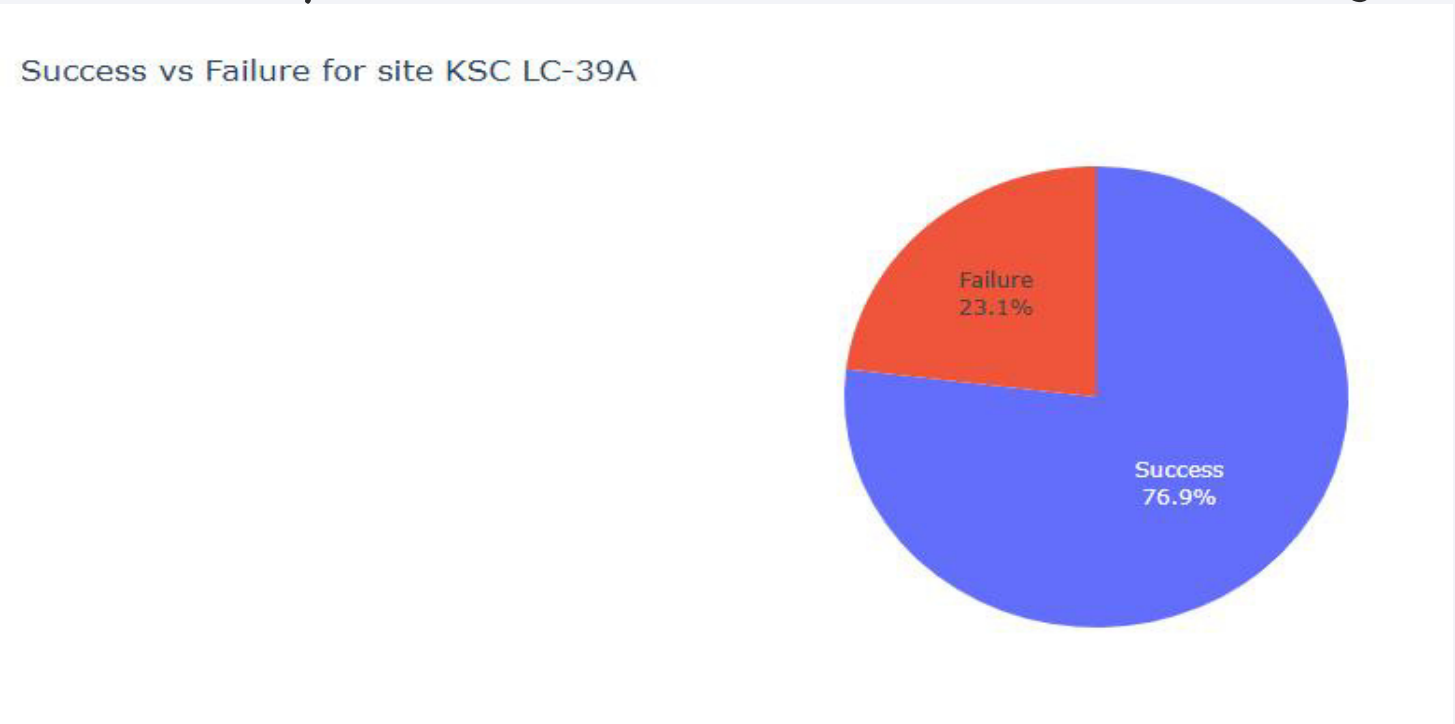
- Replace <SpaceX Launch Records Dashboard> title with an appropriate title
- Show the screenshot of launch success count for all sites, in a piechart



- Explain the important elements and findings on the screenshot
- From the image we can see that launch site KSC LC-39A has the highest success rate with 41.7%,followed with CCAFS LC-40,VAFB SLC-4E and CCAFS SLC-40 with 29.2%,16.7%,12.5% respectively.

<Dashboard Screenshot 2>

- Replace <**Piechart for the launch site with highest success ratio**> title with an appropriate title
- Show the screenshot of the piechart for the launch site with highest launch success ratio



- Explain the important elements and findings on the screenshot
- From the image we can conclude that launch site KSC LC-39A has success ratio of 76.9% which is highest compared to other launch sites

<Dashboard Screenshot 3>

- Replace <Payload vs. Launch Outcome scatter plot> title with an appropriate title

- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider

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

- from the image we conclude that booster version FT with payload range 2000 to 5500 kg have the highest success rate.

- And booster version V1.1 have the lowest success rate in payload range 500 to 4500 kg.

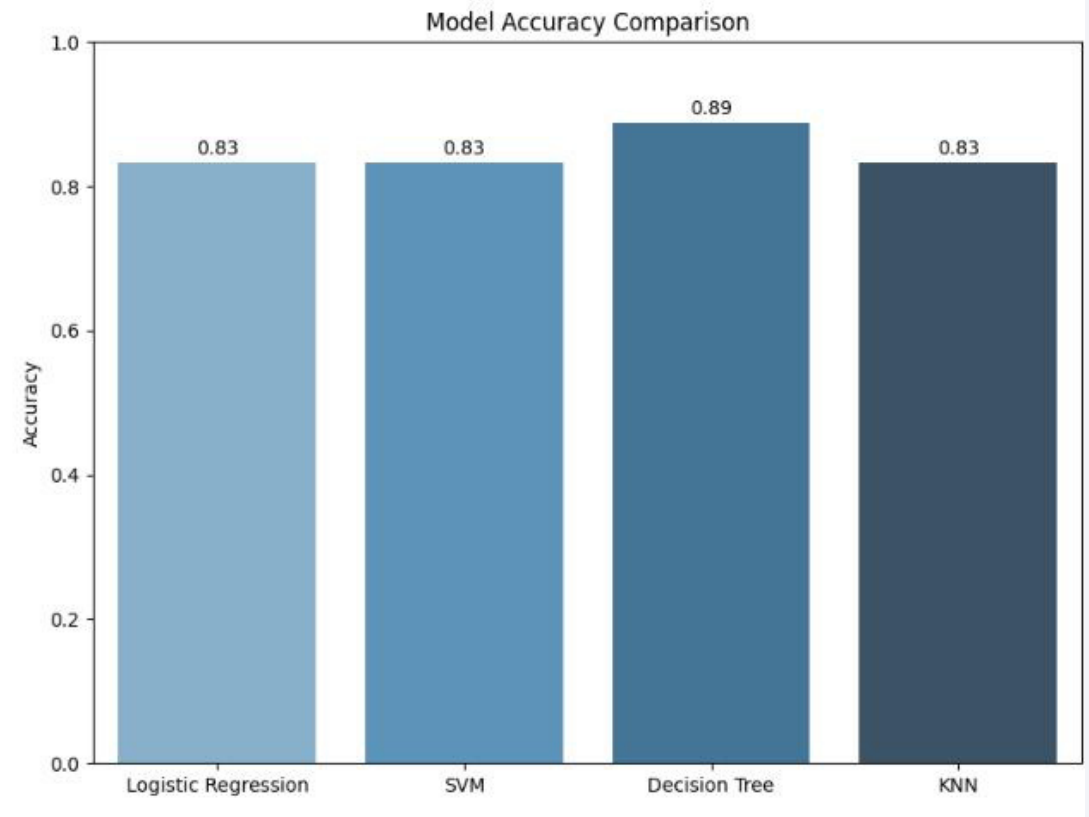
Section 5

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
- From the image we can see that Decision Tree model has the highest accuracy of 0.89(89%) on test data.



Confusion Matrix

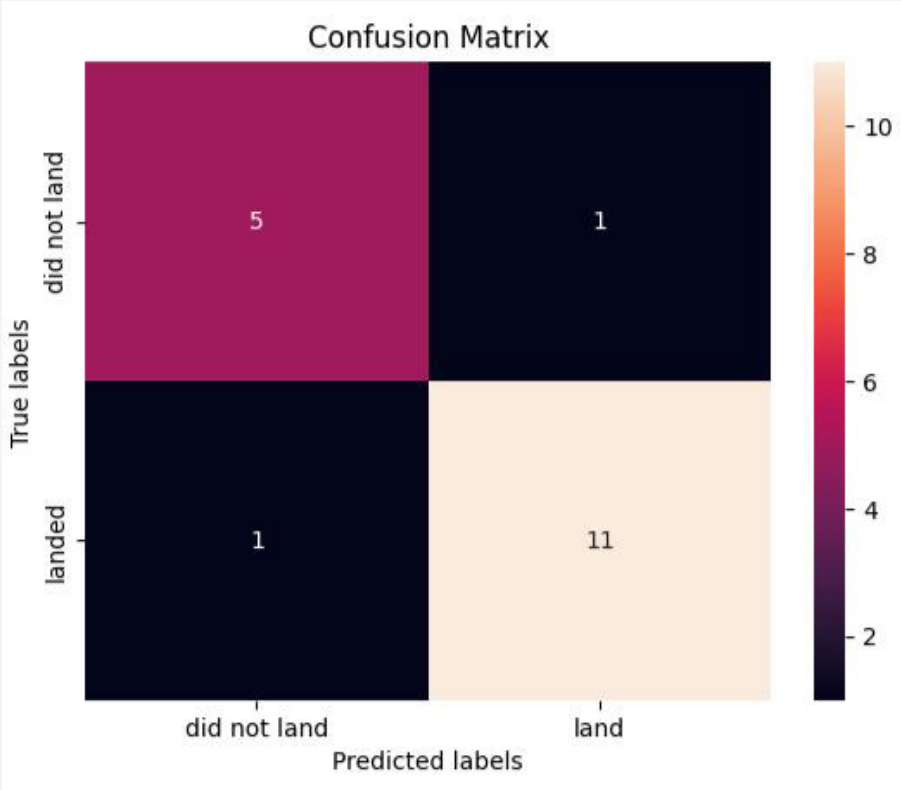
- Show the confusion matrix of the best performing model with an explanation
- This confusion matrix is of Decision Tree classifier model.
- From the matrix we can see model's Prediction:
- True Positives (TP): 11

  Predicted "land" and it actually landed.
- True Negatives (TN): 5

  Predicted "did not land" and it actually did not land.
- False Positives (FP): 1

  Predicted "land" but it did not land.
- False Negatives (FN): 1

  Predicted "did not land" but it did land.



Conclusions

- Point 1: Successfully predicted Falcon 9 first stage landing outcomes using historical launch data and machine learning models.
- Point 2: Tested 4 classification models.

  The best performing model achieved ~89% accuracy.

  Model evaluation using confusion matrices, GridSearchCV, and cross-validation.
- Point 3: Success rates improved significantly over the years.
- Certain launch sites and orbits had higher landing success.
- No. of Flights, Payload mass and orbit type were strong predictors of landing success.
- Point 4:Accurate landing predictions can help estimate launch costs and offer competitive insights for new aerospace companies.

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!