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

- 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 missing values, formatted colums 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 successfull 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 artd 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 – SpaceX API

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

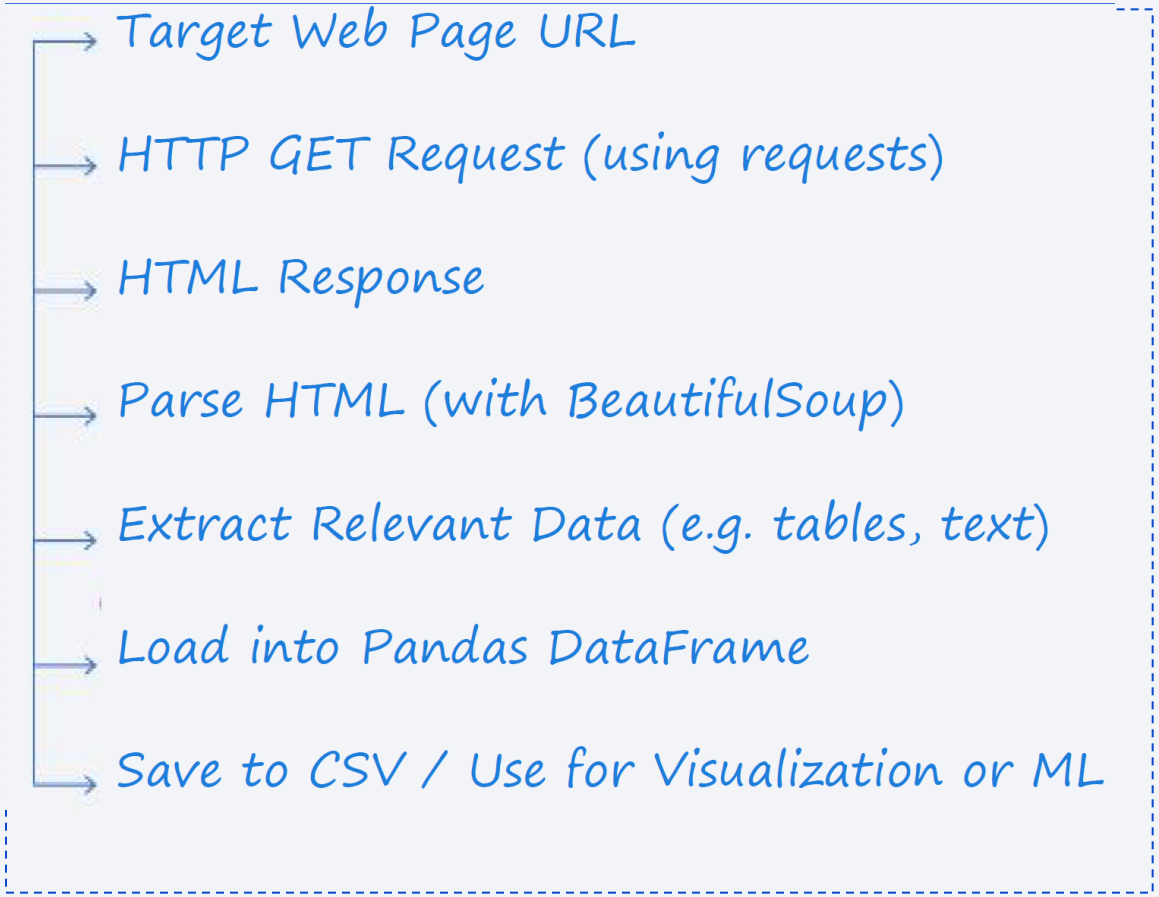
- Add the GitHub URL of the completed SpaceX API calls notebook (must include completed code cell and outcome cell), as an external reference and peer-review purpose

GitHub URL:
Coursera-Data-Science-Specialisation/Capstone/jupyter-labs-spacex-data-collection-api.ipynb at main · golleech/Coursera-Data-Science-Specialisation

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:
Coursera-Data-Science-Specialisation/Capstone/jupyter-labs-webscraping.ipynb at main · golleech/Coursera-Data-Science-Specialisation

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:
Coursera-Data-Science-Specialisation/Capstone/labs-jupyter-spacex-Data wrangling.ipynb at main · golleech/Coursera-Data-Science-Specialisation

EDA with Data Visualization

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: 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 (anding success over the years.

GitHub URL:
Coursera-Data-Science-Specialisation/Capstone/edadataviz.ipynb at main · golleech/Coursera-Data-Science-Specialisation

EDA with SQL

- %sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
- %sql select * from spacextable where Launch_Site like "CCA%" limit 16;
- %sql SELECT SUM(Payload) AS TotalPayloadMass FROM SPACEXTABLE WHERE Customer LIKE '%NASA (CRS)%';
- %sql select avg(payload_mass__kg_) as meanpayload from spacextable where booster_version like "F9%“
- %sql select min(date) from spacextable where mission_outcome = 'Success' and landing_outcome like '%ground pad%';
- %sql select * from spacextable where landing_outcome like '%drone ship%' and payload_mass__kg_ between 4000 and 6000;
- %sql select count(*) from spacextable where mission_outcome = "Success";
- %sql select booster_version from spacextable where payload_mass__kg_ = (select max(payload_mass__kg_) from spacextable);
- %sql select substr(date, 6, 2) as month_name, landing_outcome, booster_version, launch_site from spacextable where landing_outcome like "%failure%drone ship%" and substr(date, 0, 5) = "2015";
- %sql select landing_outcome, count(landing_outcome) as outcome_count from spacextable where date between "2010-06-04" and "2017-03-20" group by landing_outcome order by date desc;

GitHub URL:
Coursera-Data-Science-Specialisation/Capstone/jupyter-labs-eda-sql-coursera_sqllite.ipynb at main · golleech/Coursera-Data-Science-Specialisation

Build an Interactive Map with Folium

- 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 and coastline of each launch Sites.
- 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

GitHub URL:
Coursera-Data-Science-Specialisation/Capstone/lab_jupyter_launch_site_location.ipynb at main · golleech/Coursera-Data-Science-Specialisation

Build a Dashboard with Plotly Dash

- First I added a pie chart
- Then a Scatter plot
- Then a Payload Range slider
- I Created a drop -down button for launch sites
- I added pie chart to get successful launch sites
- I added Scatter plot to see if there is any relation between payload mass and success rate
- I added Payload Range slider so user can select spacific payload mass
- I Added drop -down button so user can select spacific launch Site.

GitHub URL:
Coursera-Data-Science-Specialisation/Capstone/spacex-dash-app.py at main · golleech/Coursera-Data-Science-Specialisation

Predictive Analysis (Classification)

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

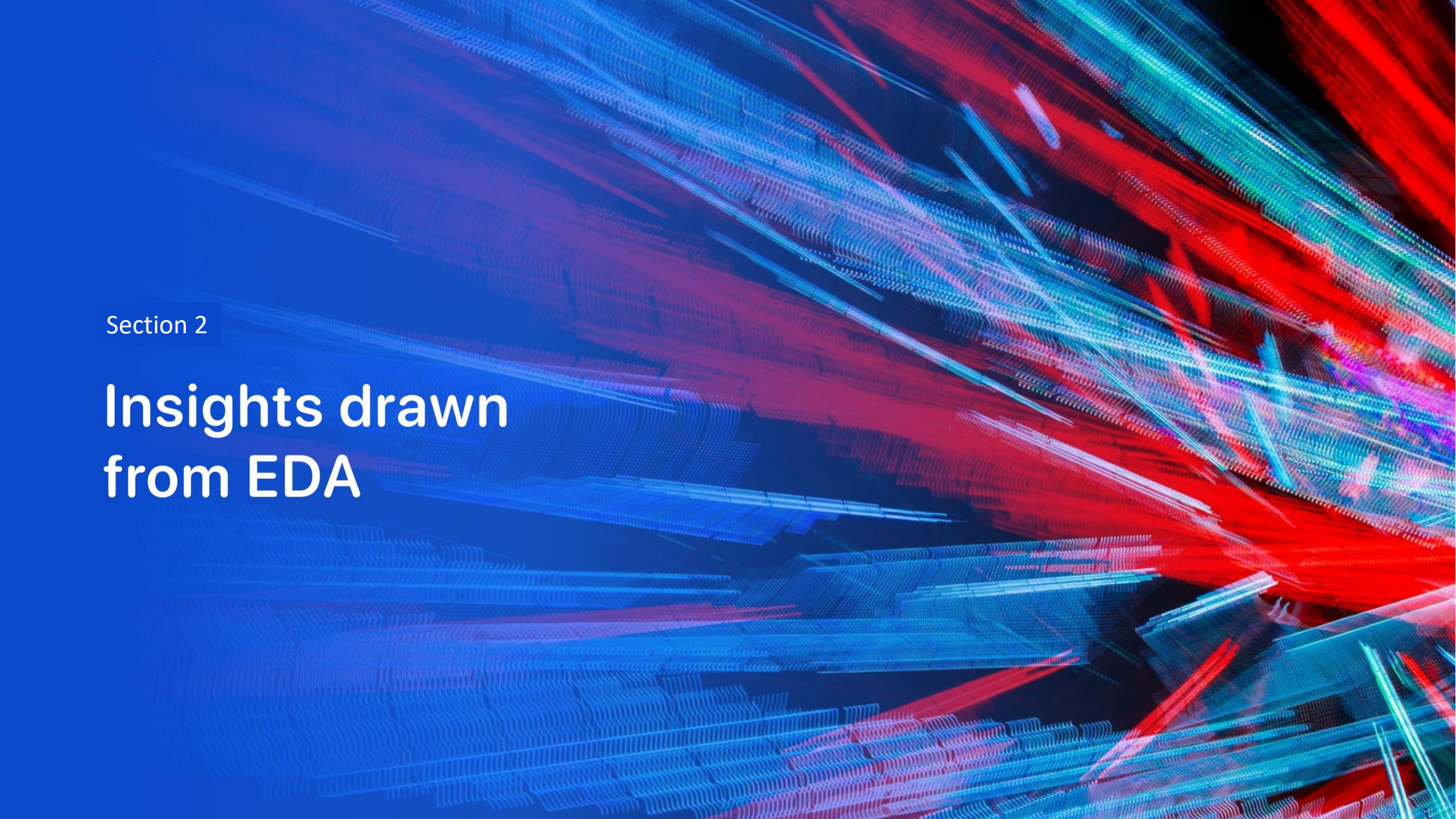
GitHub URL:
Coursera-Data-Science-Specialisation/Capstone/SpaceX_Machine Learning Prediction_Part_5.ipynb at main · golleech/Coursera-Data-Science-Specialisation

Results

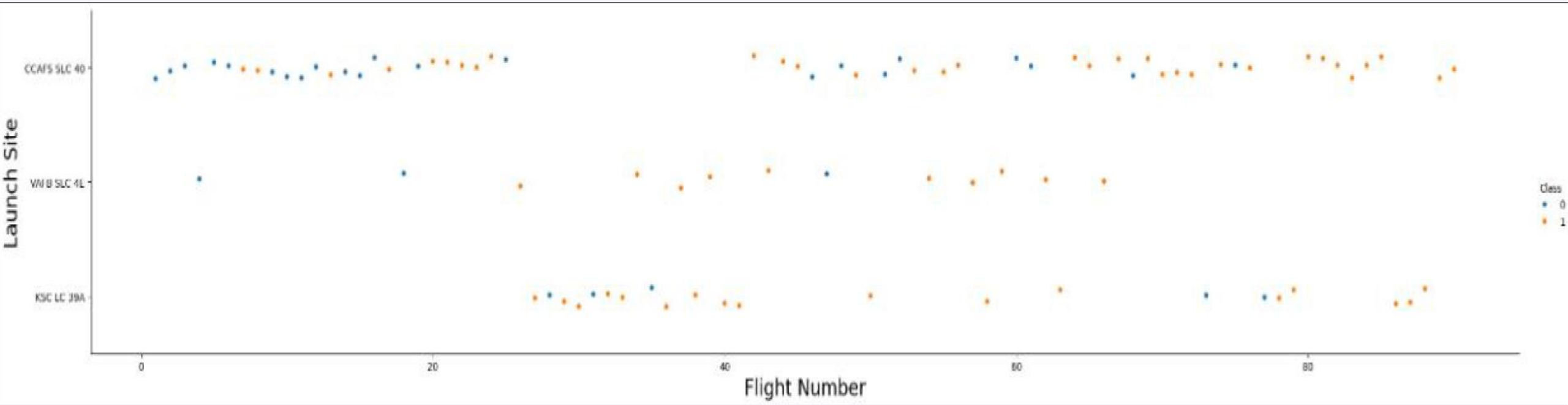
- Plotted pie chart and bar chart to analyze success vs. failure rates
- Observed that majority of launches were successful (~ 75% success rate)
- Pie charts showed variation in success rates across different launch Sites
  - For example, KSC LC-39A had a higher proportion of successful landing compared 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 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 evaluated on Accuracy Score, Confusion Matrix, Precision Score etc.
- Best model was Decision Tree classification With Accuracy Score of on test data
- Other models Logestic Regression, SVM, KNN have Accuracy score of 83% respectively

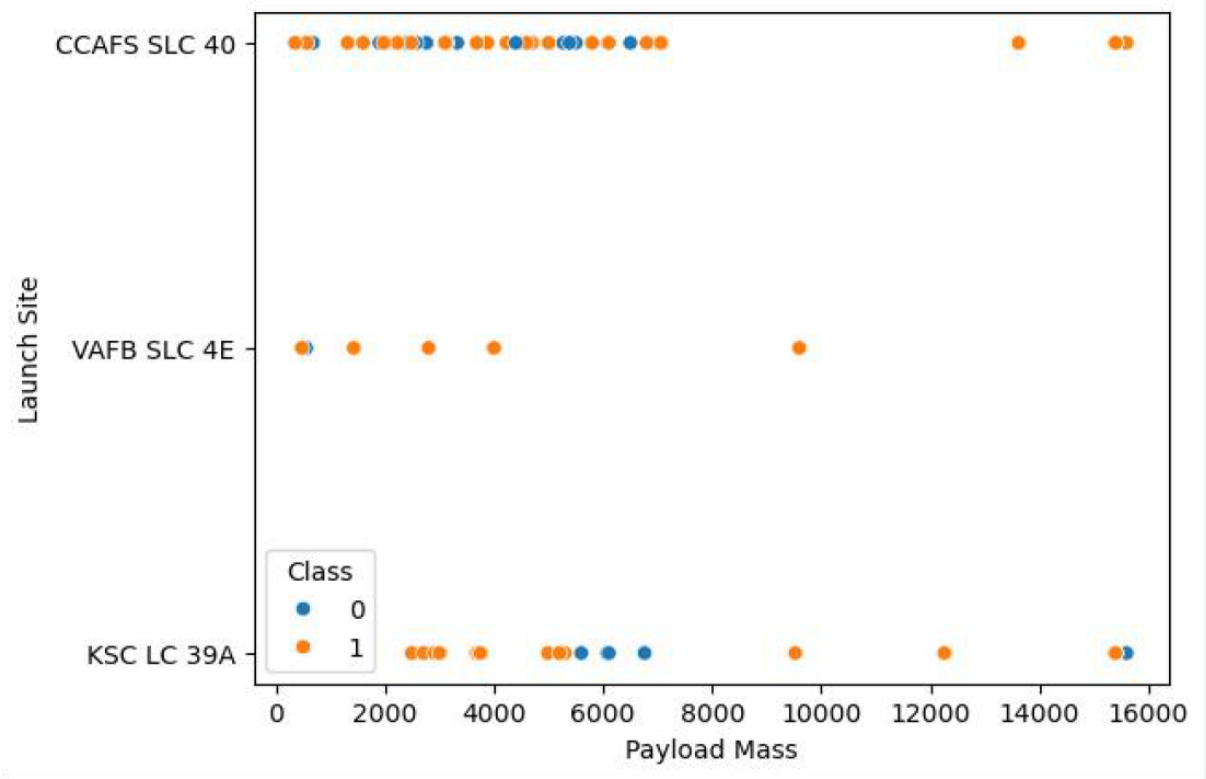

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



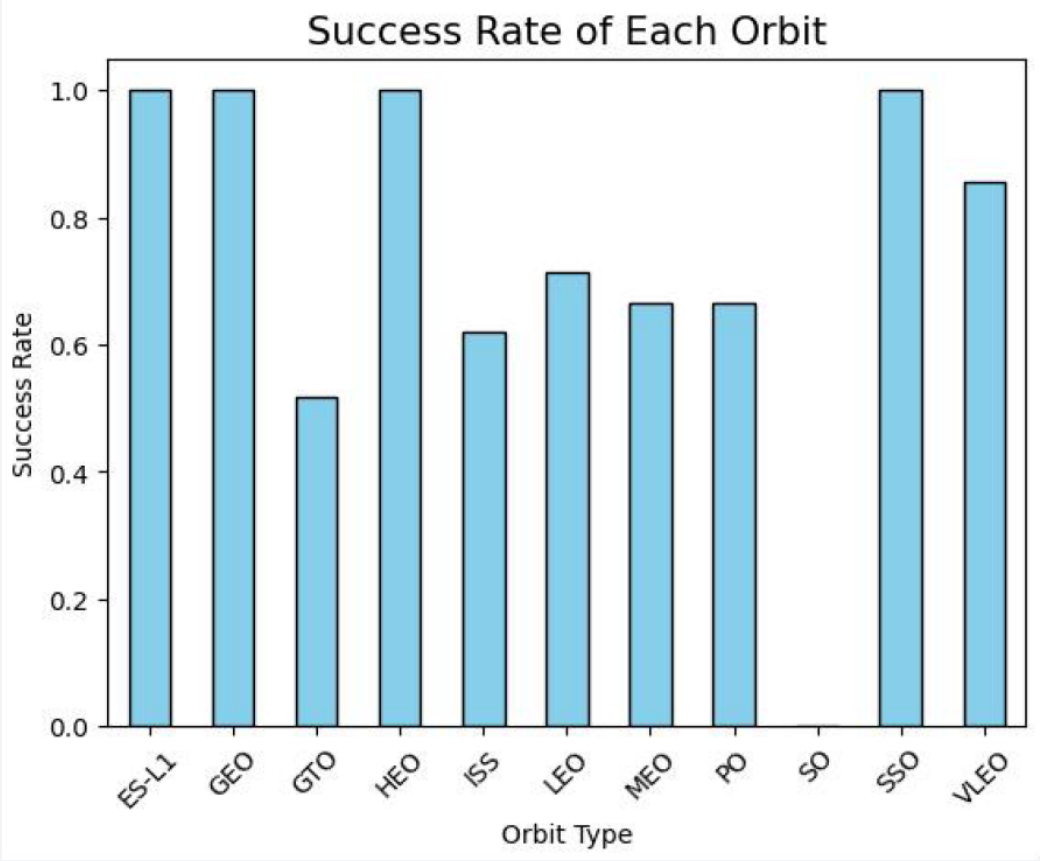
Payload vs. Launch Site

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



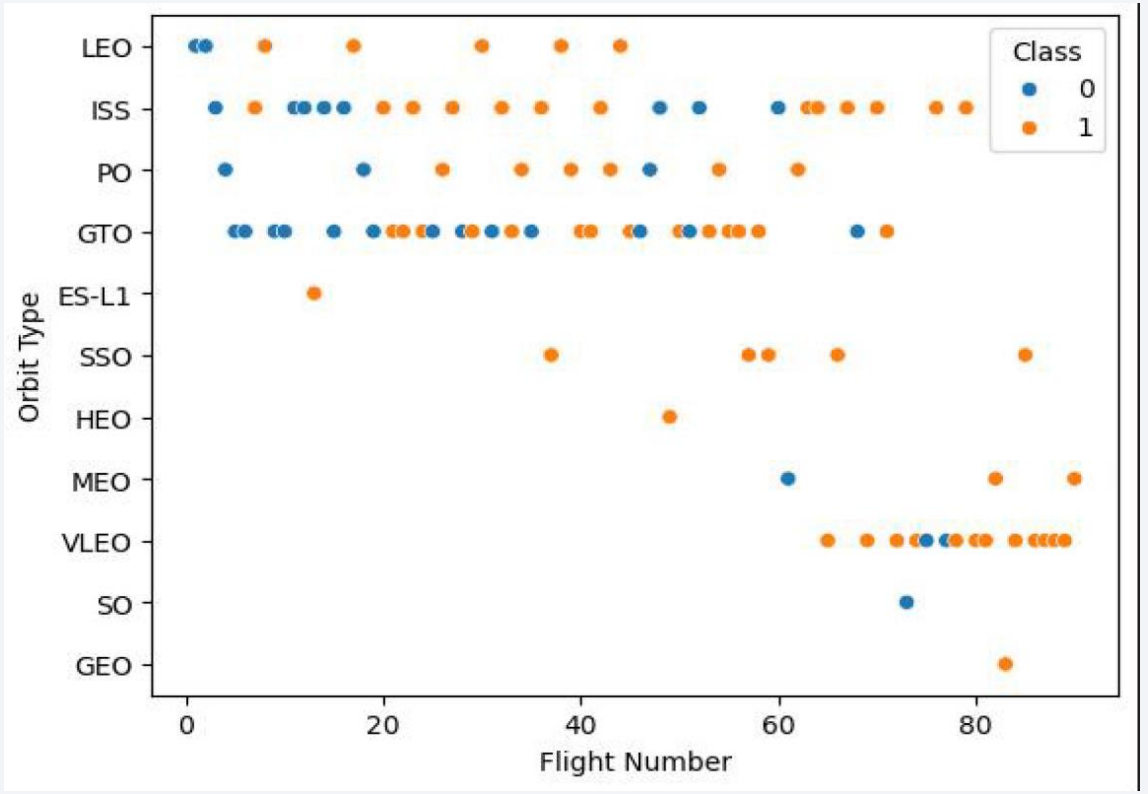
Success Rate vs. Orbit Type

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



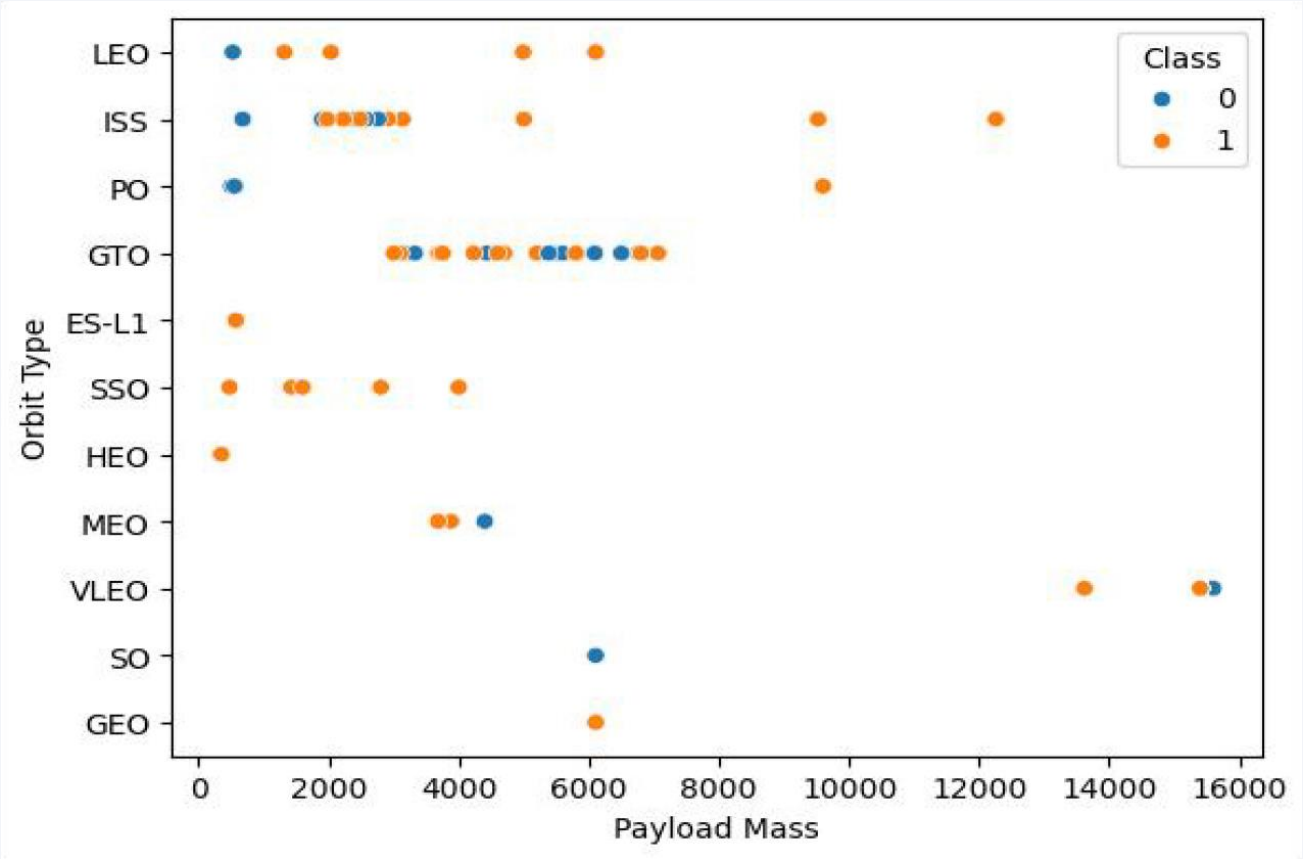
Flight Number vs. Orbit Type

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



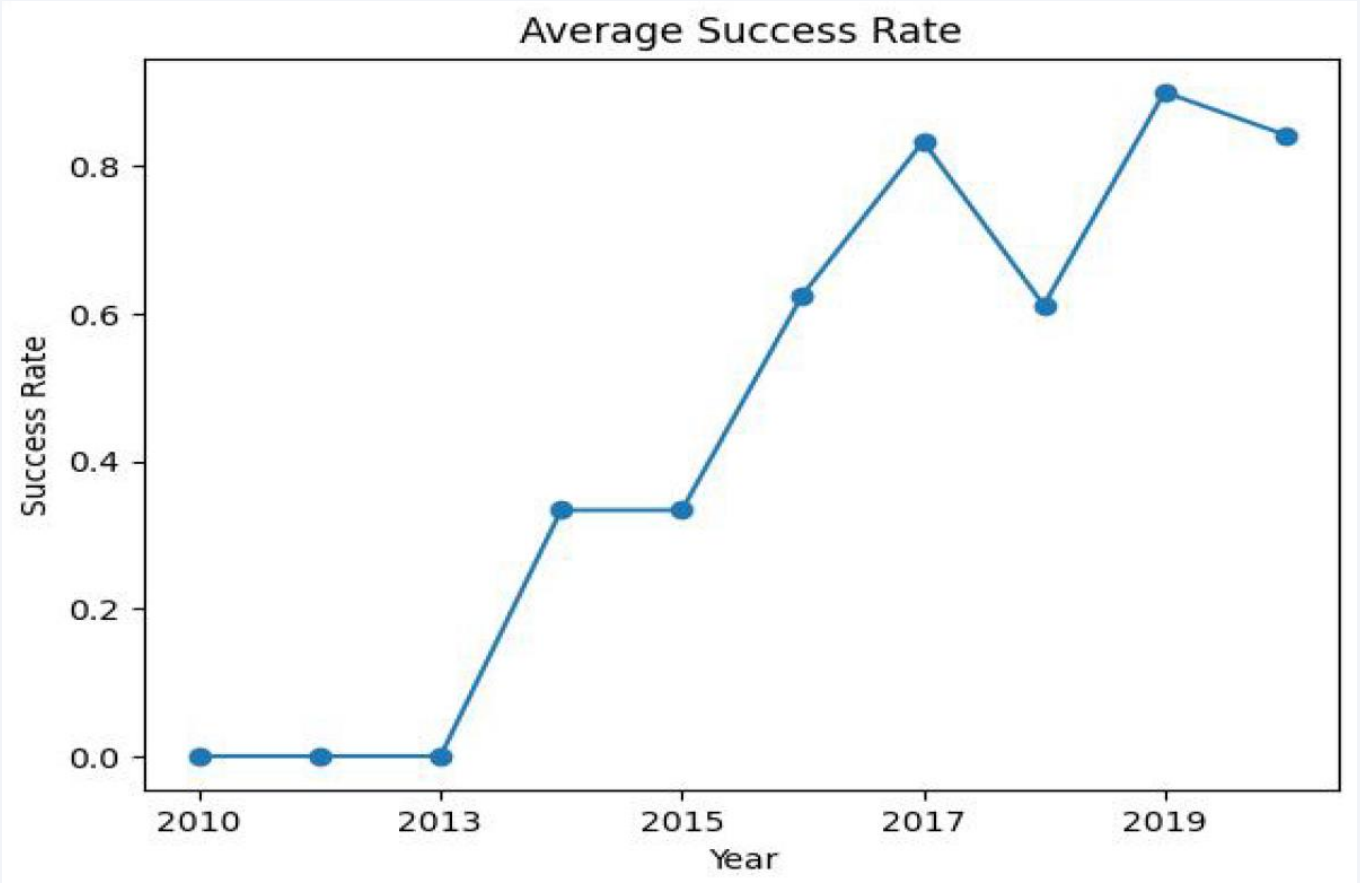
Payload vs. Orbit Type

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



Launch Success Yearly Trend

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



All Launch Site Names

- Find the names of the unique launch Sites
- CCAFS LC - 40
- VAFB SLC - 4 E
- KSC LC - 39A
- CCAFS SLC - 40
- %sql SELECT DISTINCT Launch_ site FROM SPACEXTABLE
- This querg gives us names of unique launch Sites

Launch Site Names Begin with 'CCA'

- %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
- %sql select SUM(payload_mass_kg_) as 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
- %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
- %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

- F9 FT B1022
- F9 FT B1026
- F9 FT B1021.2
- F9 FT B1031.2
- %sql select distinct Booster_Version from SPACEXTABLE where Landing_ Outcome="Success (drone ship)" and PAYLOAD_ MASS_ KG_ > 4000 and PAYLOAD_ MASS_ < 6000;

Total Number of Successful and Failure Mission Outcomes

- %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

- %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

- %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”;

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

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

- %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, total_count, (select count(*) from landing_counts lc2 where > lcl2.total_count > lc1.total_count) + 1 as rank from landing_counts lc1 order by rank;

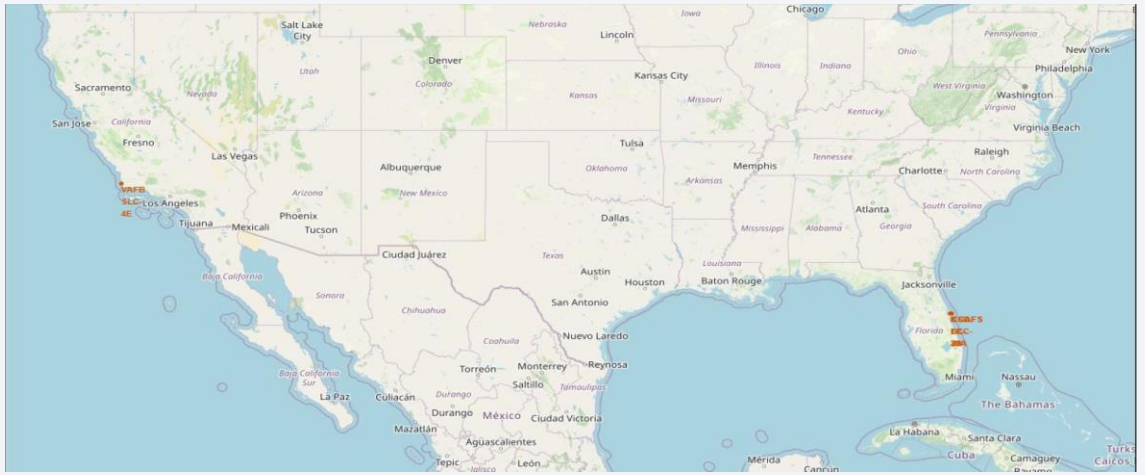
Landing_Outcome	total_count	rank
No attempt	10	1
Failure (drone ship)	5	2
Success (drone ship)	5	2
Controlled (ocean)	3	4
Success (ground pad)	3	4
Failure (parachute)	2	6
Uncontrolled (ocean)	2	6
Precluded (drone ship)	1	8

Section 3

Launch Sites Proximities Analysis

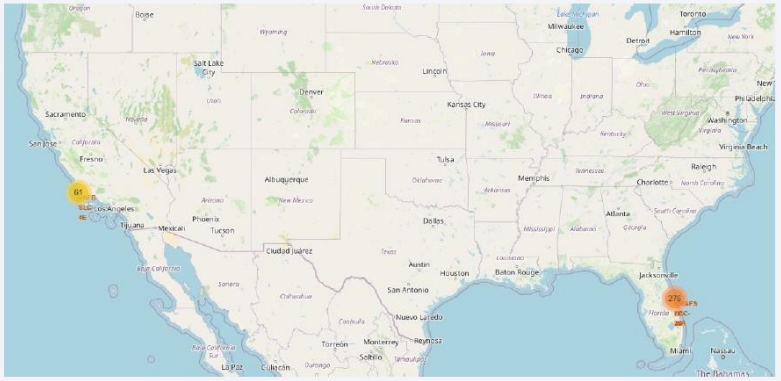
<Folium Map Screenshot 1>

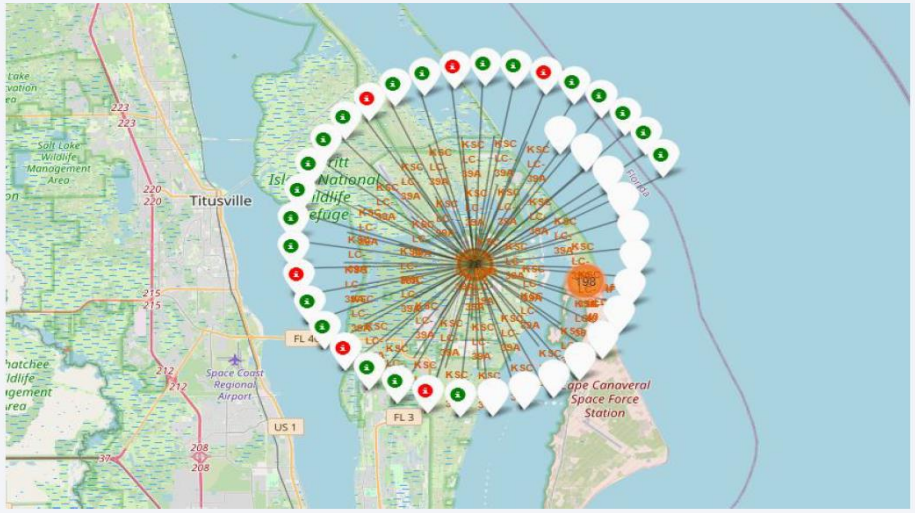
- From the image we can understand that not all launch sites are in proximity to the Equator line.
- All launch sites are in very Close proximity to the coast.
- I have marked and labeled all launch sites through their latitude and longitude coordinates.

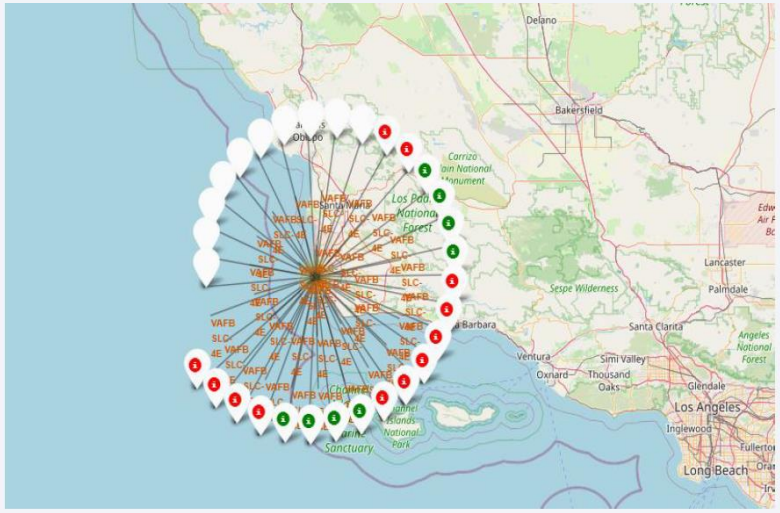


<Folium Map Screenshot 2>

- Marked all sites With color-labeled launch outcomes where launch outcome=successful = GREEN and launch launch outcome=unsuccessful=Red
- From the map I find that launch Site KSC LC-39A has the highest success rate.

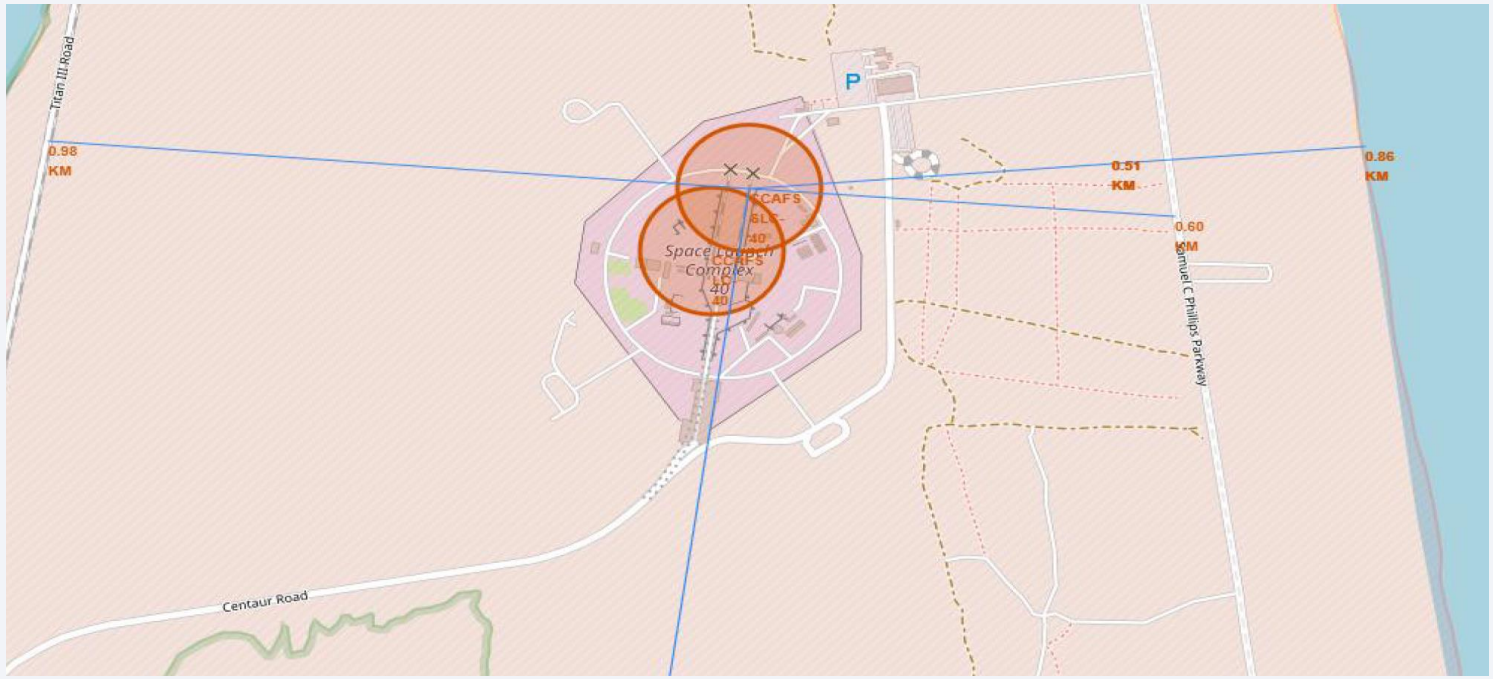


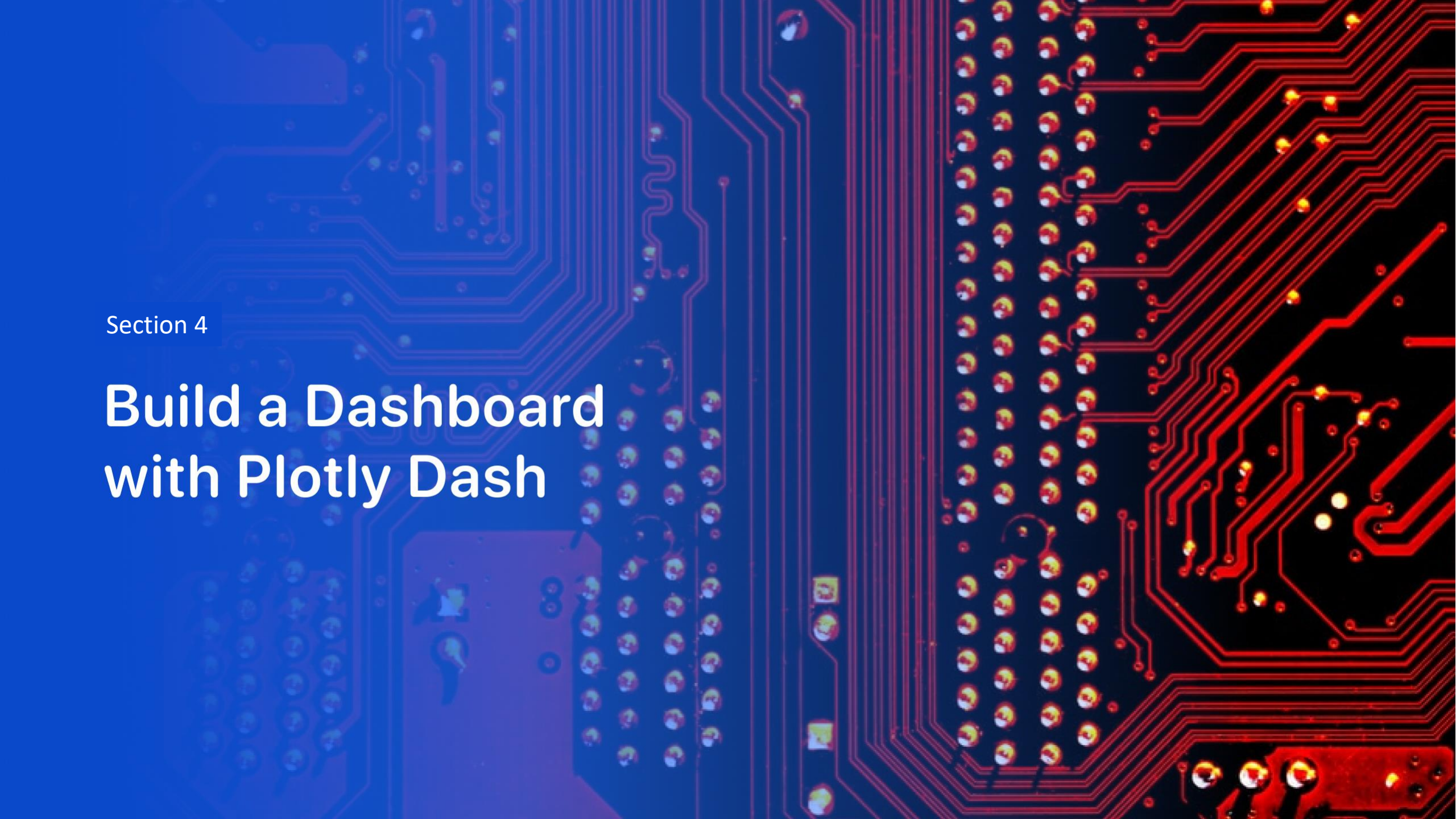




<Folium Map Screenshot 3>

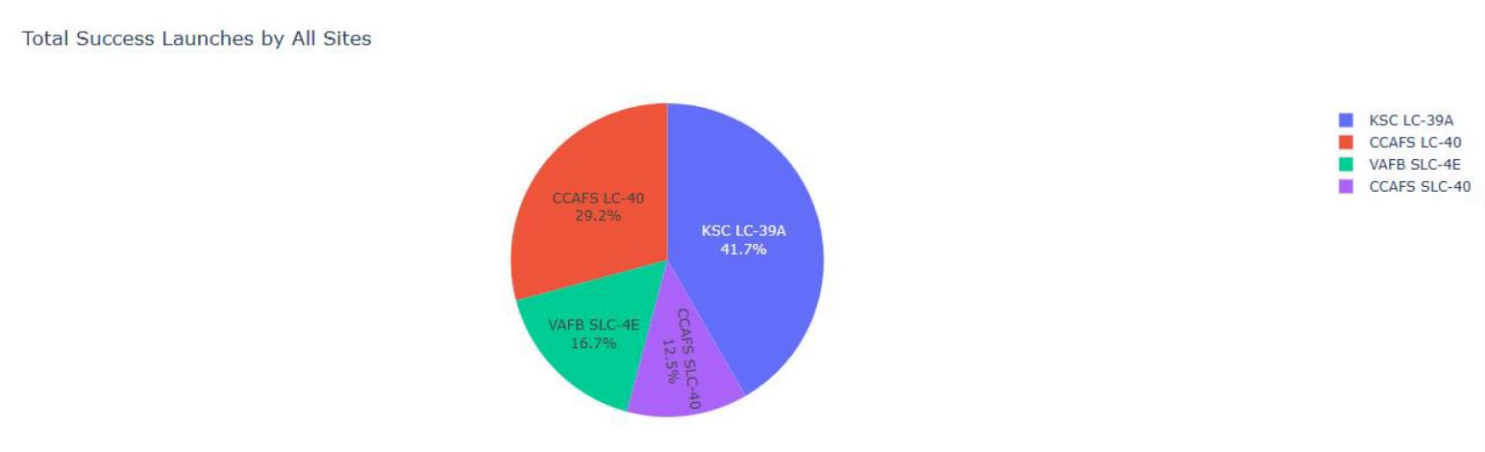
- from the image we can see launch Site CCAFS SLC-40 is closest to highway With 0,60KM, then coastline With 0,86 KM and then railway With 0,98 KM




Section 4
Build a Dashboard with Plotly Dash

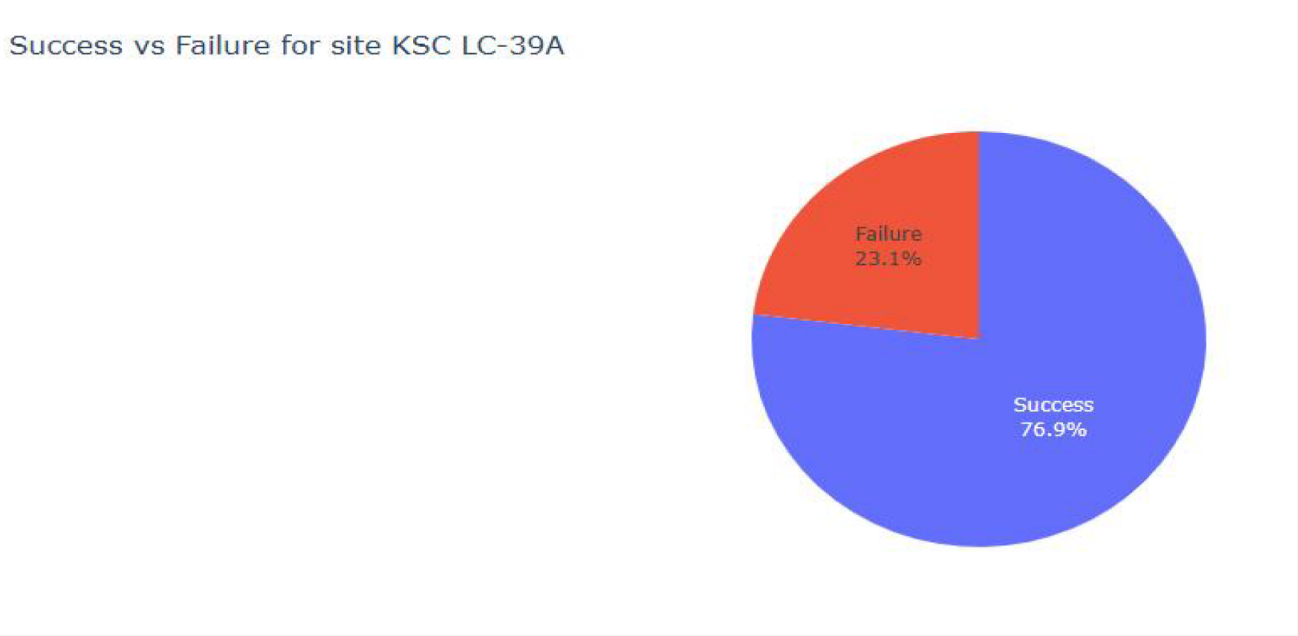
<Dashboard Screenshot 1>

- From the image we cart see that launch site KSC LC-39A has the highest success rate with 41.7% followed by CCAFS LC-40, VAFB SLC-4E and CCAFS SLC-40 with 29.2%, 16.7% and 12.5% respectively



<Dashboard Screenshot 2>

- From the image we can conclude that launch site KSC LC-39A has a success ratio of 76.9% which is highest compared to other launch sites



<Dashboard Screenshot 3>

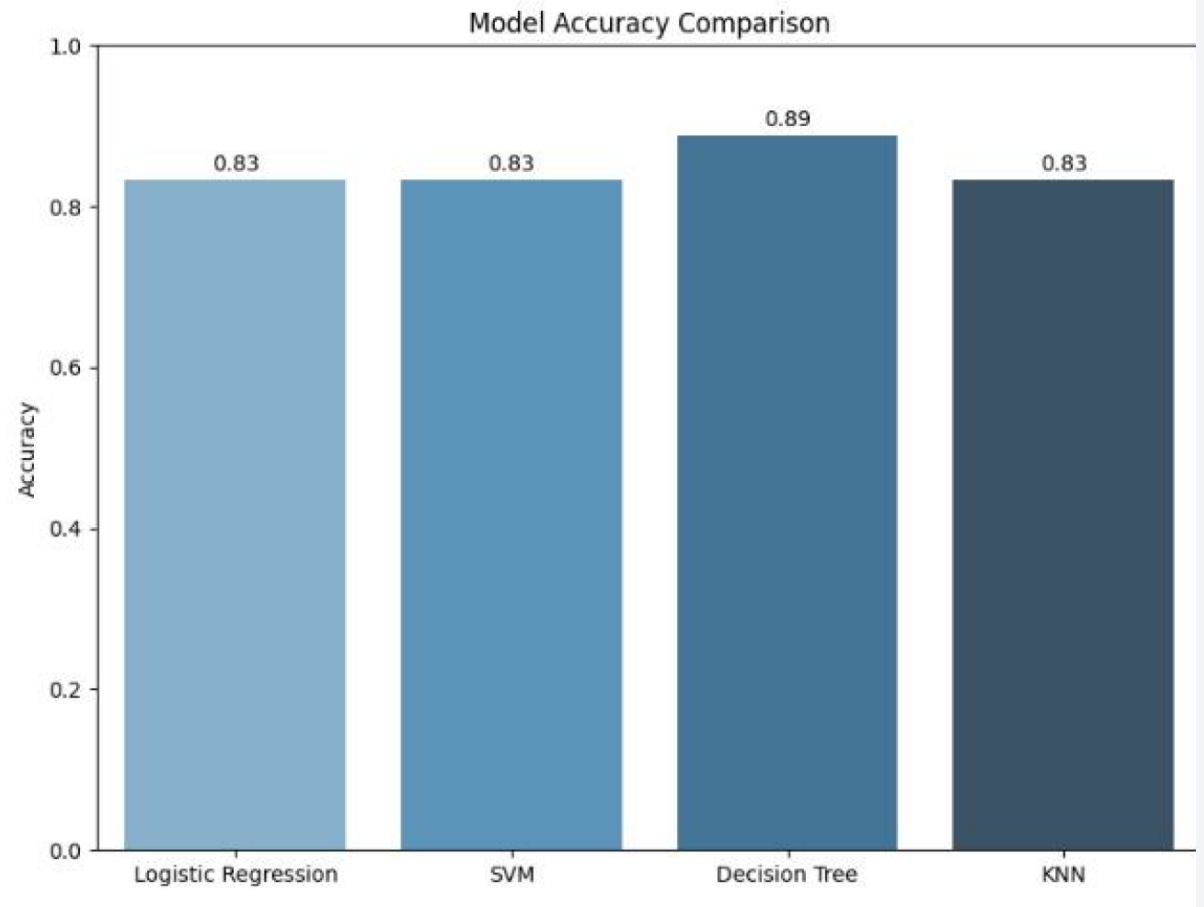
- from the image we conclude that booster version FT With payload range 2000 to 5500 kg have the highest success rate
- And booster version V1.1 have the lowest success rate in payload range 500 to 4500 kg.

Section 5

Predictive Analysis (Classification)

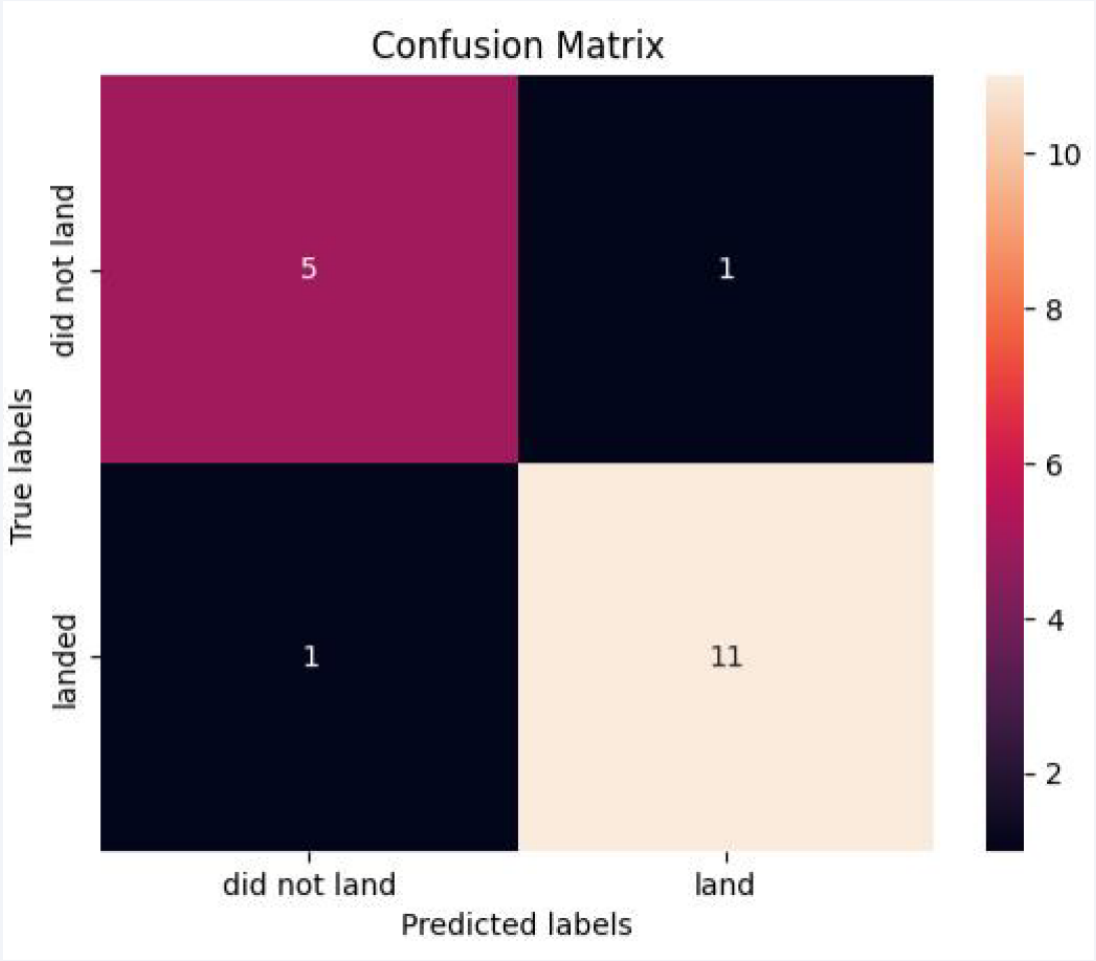
Classification Accuracy

- From the image we can see that Decision Tree model has the high accuracy of 89% on test data



Confusion Matrix

- This confusion matrix is of Decision Tree classifier model
- From the matrix we cart see 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
- redicted "did not land" but it did land.



Conclusions

- Successfully predicted Falcon 9 first stage landing outcomes using historical launch data and machine learning models
- Tested 4 classification models, The best performing model achieved 89% accuracy, Model evaluation using confusion matrices, GridSearchCV, and cross-validation
- 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
- 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!