Winning Space Race with Data Science

Santhosh Suryadevara
Apr 14, 2023

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

- Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch. We will go over different ways to analyse the data we have & come define a model that predicts the landing outcome.

Section 1

Methodology

Methodology

Executive Summary

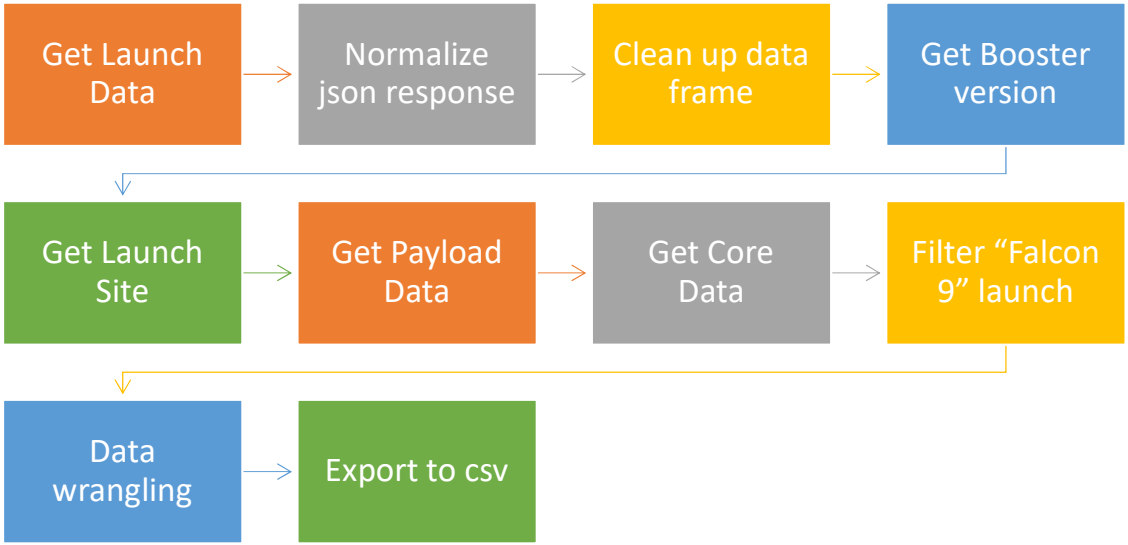
- Data collection methodology:
  - We used Space-X API's to get the details about launches. We then consolidated the above data with respective API's to get details about booster version, launch site, payload data, core data.
  - After collecting the above data, we created a data frame.
  - We also did web scrapping on Wikipedia page to get the details of space x launch records.
- Perform data wrangling
  - We calculated the mean of respective columns & replaced them with nan values.
  - We analyzed the data & determined the training labels for the outcomes.
- Perform exploratory data analysis (EDA) using visualization and SQL
  - We loaded the csv file into DB2 & ran some queries to analyze the data. Ex: Distinct launch sites, total payload mass carried by certain boosters, first date of successful landing, total number of success & failure outcomes etc.
  - We have also done EDA using scatter plots to show the relationship between different attributes like FlightNumber vs PayloadMass, Orbit vs PayloadMass, launch success rate depending on different attributes, etc.

Methodology

Executive Summary

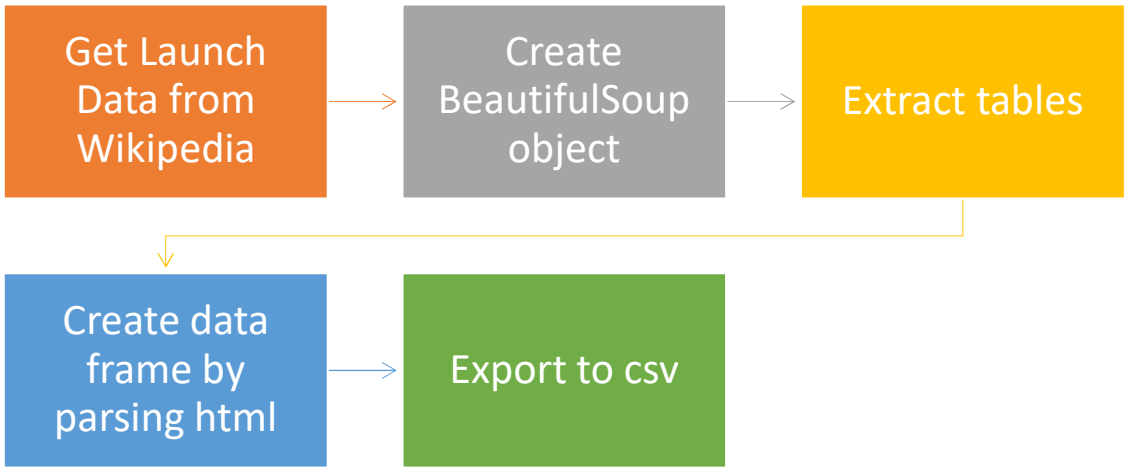
- We used visual analytics on a map to visualize the launch sites, success & failed launches for each site and calculated the distances between a launch site to its proximities to find out the nearest rail, road & air transports availability.
- We developed interactive visual analytics using dash, to display the percentage of successful launches & the booster version categories used at a particular site.
- Perform predictive analysis using classification models
    - Once the data is fetched, we take the Class column into an array using numpy.
    - We then standardize the data.
    - Then we split the data into train & test data.
    - We then ran the into different models like Logistic Regression, SVM, Decision Tree, k-nearest neighbors model.

Data Collection - API



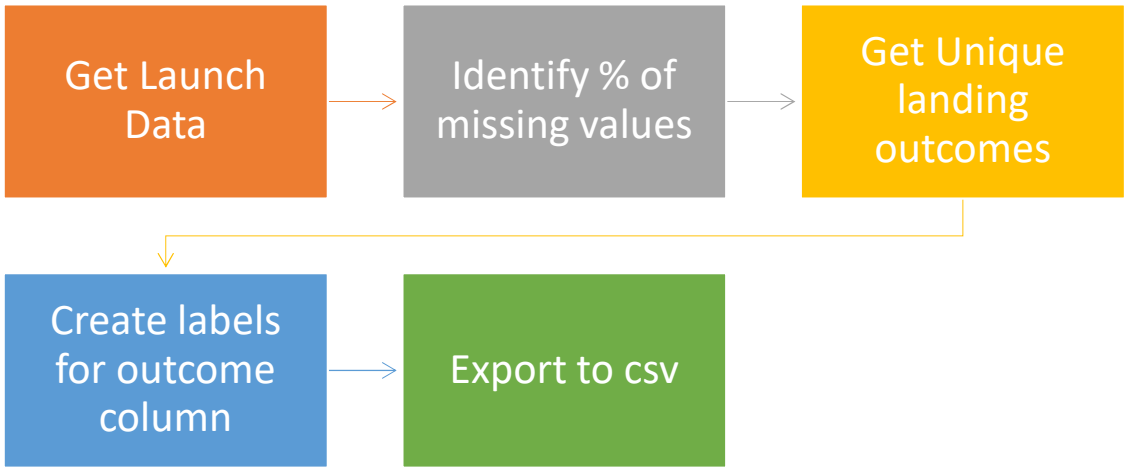
- Git Hub: https://github.com/santhosh00724/capstone/blob/main/week1/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping



- Git Hub: https://github.com/santhosh00724/capstone/blob/main/week1/WebScrapping.ipynb

Data Wrangling



- Git Hub: https://github.com/santhosh00724/capstone/blob/main/week1/DataWrangling.ipynb

EDA with Data Visualization

- Summary of charts plotted:
  - Categorical plots for FlightNumber vs PayloadMass overlay launch outcome. We see as flight number increases the success rate increases & last 20 flights have high payload success rate.
  - Categorical plots for FlightNumber vs Launch Site overlay launch outcome. We see that CCAFS is highly used launch site compared to others & success rate is also high.
  - Categorical plots for PayloadMass vs Launch Site overlay launch outcome. We can see VAFB-SLC launch site is not used for heavy payload mass > 10K.
  - Bar chart for Orbit vs Class to find out which Orbit has highest success outcome. We can see ES-L1, GEO, HEO, SSO has highest success rate.
  - Categorical plots for FlightNumber vs Orbit overlay launch outcome. As the flight number increases the success rate also increases for Orbit VLEO.
  - Categorical plots for PayloadMass vs Orbit overlay launch outcome. For heavy payloads we see orbit Polar, LEO, ISS has good success rate.
  - Line plot for Year vs Class to see the success rate increased as we increase the year from 2013 till 2020.
- GitHub URL : https://github.com/santhosh00724/capstone/blob/main/week2/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- Finding out unique launch sites used in the mission.
- Listing first 5 records where launch sites starts with “CCA”.
- To find the total payload mass carried by boosters launched by customer NASA (CRS).
- Calculated average payload mass carried by booster version F9 v1.1
- To find the first date of successful landing.
- Listing names of boosters which have success in drone ship & have payload between 4000 to 6000.
- Listing names of booster versions which have carried maximum payload mass.
- List month names in 2015 for failure landing outcomes & it’s booster versions, launch sites.
- Ranking the count of successful landing outcomes between 04-06-2010 and 20-03-2017 in descending order.
- GitHub URL : https://github.com/santhosh00724/capstone/blob/main/week2/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
  - Map : To display USA map.
  - Marker : To display the launch site
  - Circle : To circle the selected launch site
  - MarkerCluster : To make a cluster of markers to differentiate the
  - PolyLine : To visualize the distance between the selected location to the launch site.

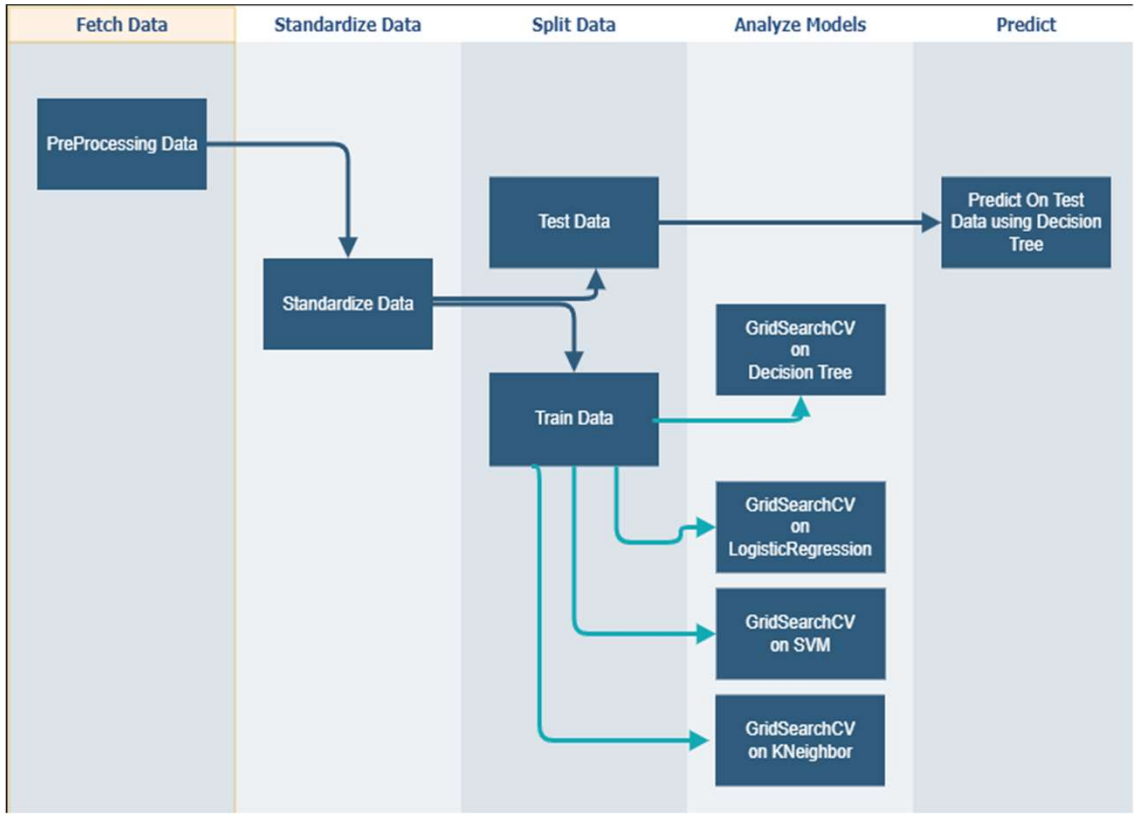
- GitHub URL : https://github.com/santhosh00724/capstone/blob/main/week3/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
  - Used Pie Chart to display percentages of successful launches of different sites.
  - Used scatter plot to display the relationship between launch outcomes vs payload mass for different booster versions at different sites.
- Explain why you added those plots and interactions
  - User can select a launch site from drop down to get more information about the success rates, based on the payload for different booster versions.
- GitHub URL ：https://github.com/santhosh00724/capstone/blob/main/week3/spacex_dash_app.py

Predictive Analysis (Classification)

- GitHub URL : https://github.com/santhosh00724/capstone/blob/main/week4/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

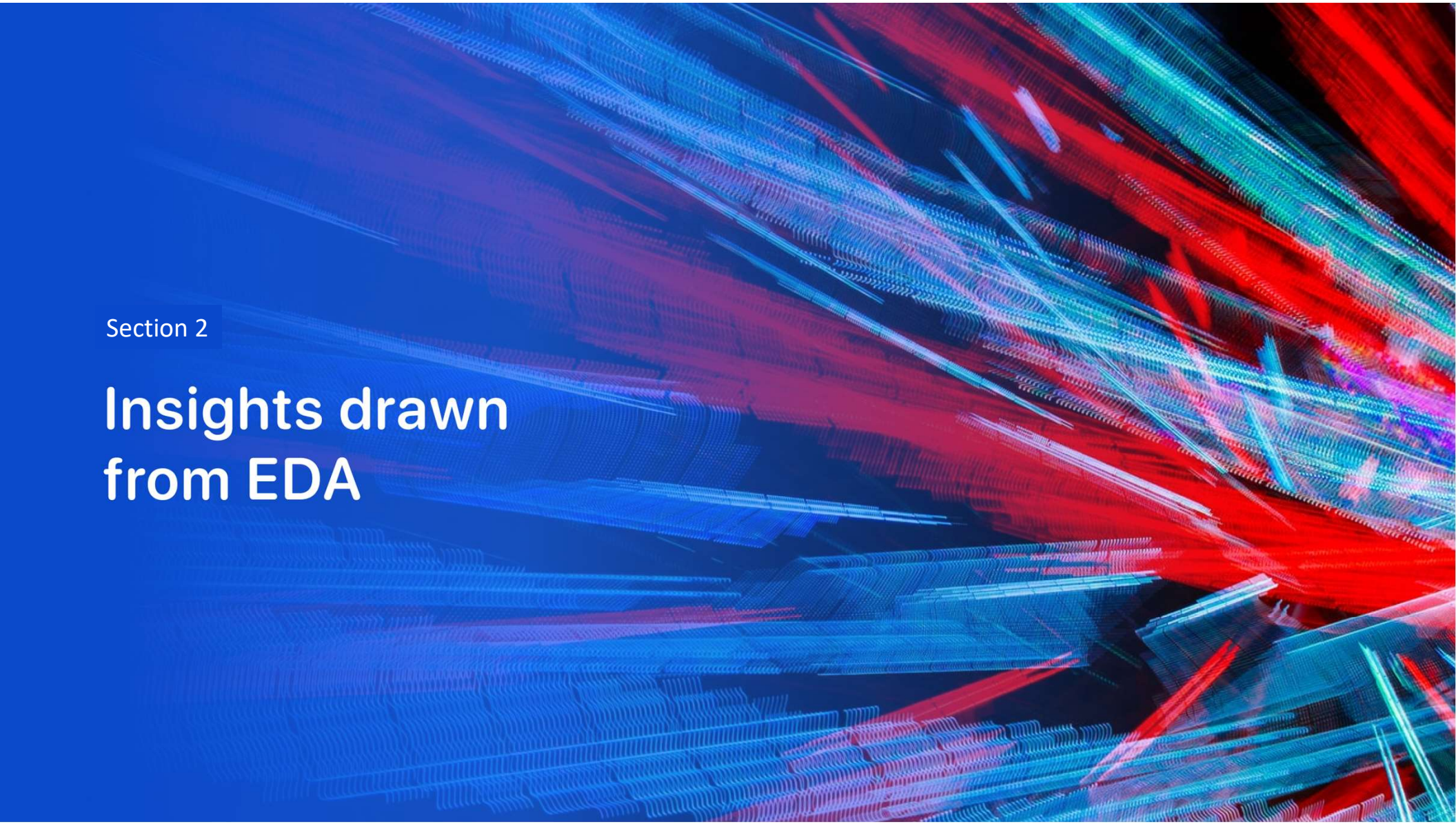


Results

Exploratory data analysis results

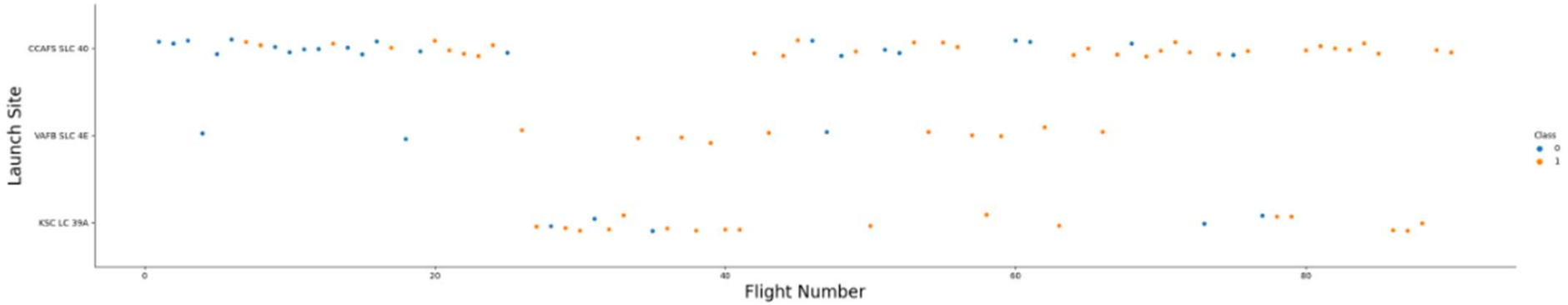
Interactive analytics demo in screenshots

Predictive analysis results


Section 2
Insights drawn
from EDA

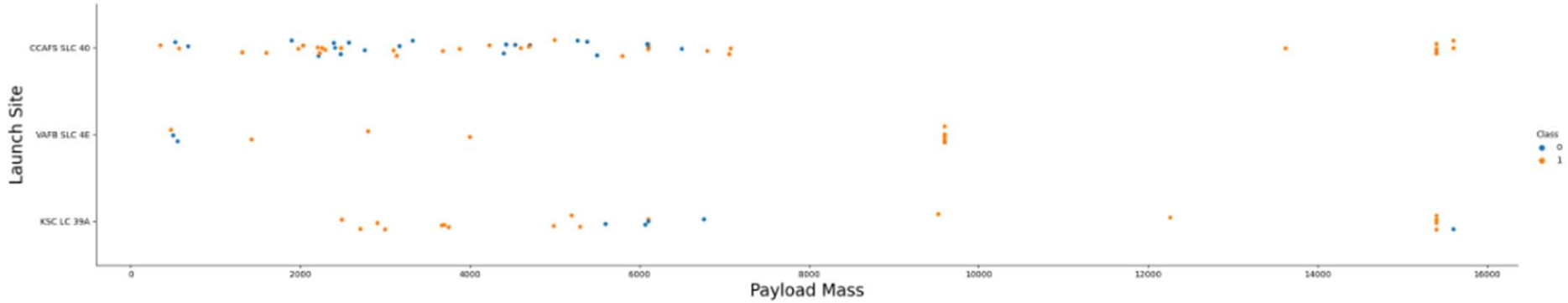
Flight Number vs. Launch Site

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



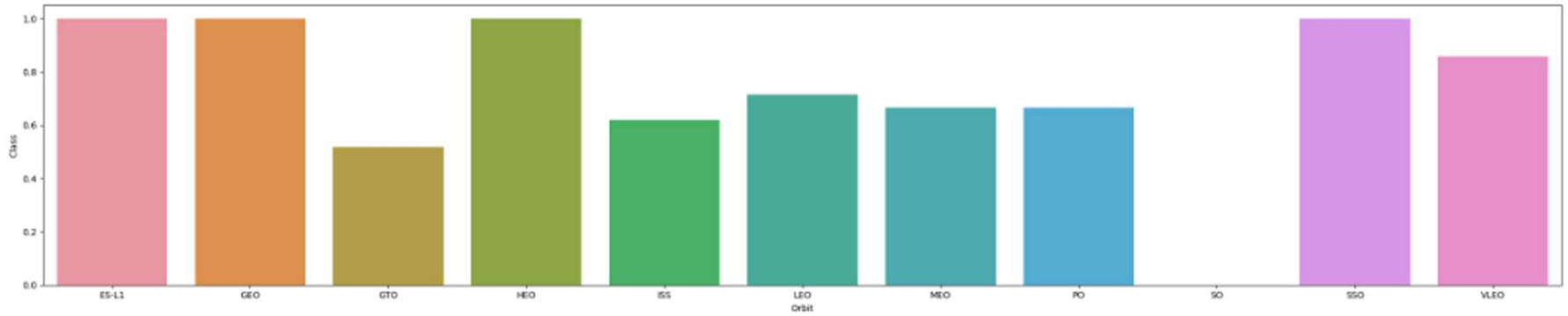
Payload vs. Launch Site

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



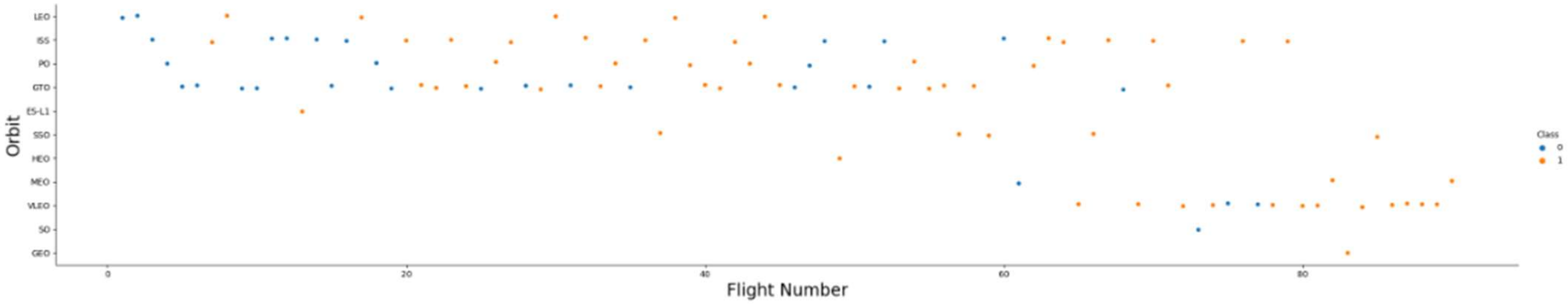
Success Rate vs. Orbit Type

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



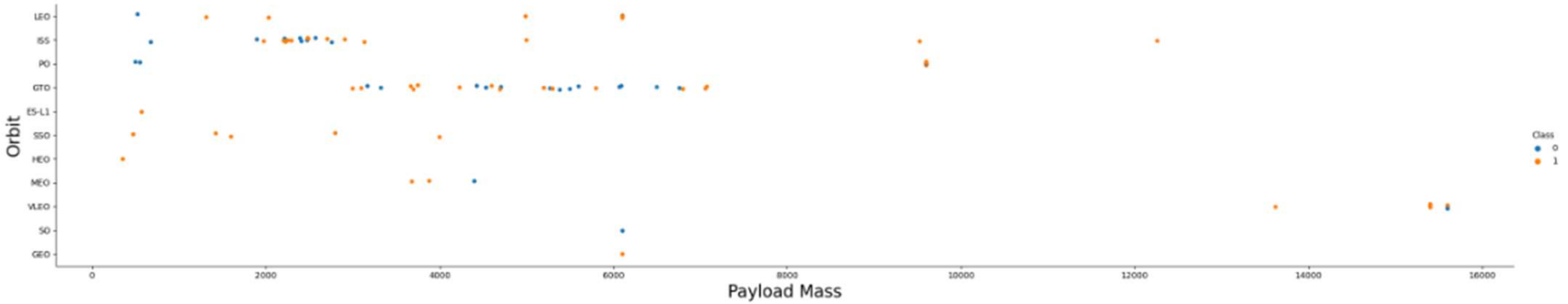
Flight Number vs. Orbit Type

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



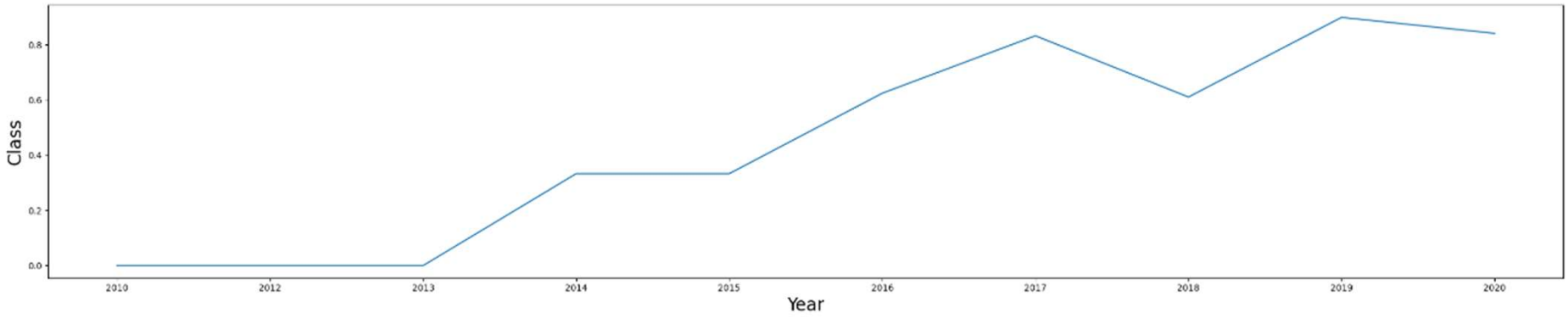
Payload vs. Orbit Type

- 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
- Present your query result with a short explanation here

Task 1

Display the names of the unique launch sites in the space mission

```
%sql select DISTINCT(LAUNCH_SITE) from SPACEXTBL
```

```
 * sqlite:///my_data1.db
Done.
```

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

Launch Site Names Begin with 'CCA'

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

Task 2

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

```
%sql select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' and rowid <=5
```

```
 * sqlite:///my_data1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	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

Task 3

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

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where CUSTOMER = 'NASA (CRS)'
```

```
 * sqlite:///my_data1.db
Done.
```

sum(PAYLOAD_MASS__KG_)
45596

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

Task 4

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

```
%sql select AVG(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version = 'F9 v1.1'
```

```
 * sqlite:///my_data1.db
Done.
```

AVG(PAYLOAD_MASS__KG_)
2928.4

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

Task 5

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

Hint:Use min function

```
%sql select min(date) from SPACEXTBL where Mission_Outcome = 'Success'
```

```
 * sqlite:///my_data1.db
Done.
```

min(date)
01-03-2013

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

Task 6

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 Mission_Outcome = 'Success' and PAYLOAD_MASS__KG_ between 4000 and 6000
```

```
 * sqlite:///my_data1.db
Done.
```

Booster_Version
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 FT B1032.2

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

Task 7

List the total number of successful and failure mission outcomes

```
%sql select Mission_Outcome, count(*) from SPACEXTBL group by Mission_Outcome
```

```
 * sqlite:///my_data1.db
Done.
```

Mission_Outcome	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

Task 8

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)
```

```
 * sqlite:///my_data1.db
Done.
```

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
- Present your query result with a short explanation here

Task 9

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

```
%sql select substr(Date, 4,2) as month, "Landing _Outcome", booster_version, launch_site from SPACEXTBL where substr(Date, 7,4) = '2015' and "Landing
```

```
 * sqlite:///my_data1.db
Done.
```

month	Landing _Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

Task 10

Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

```
%sql select "Landing _Outcome", count(*) as rank from spacextbl where "Landing _Outcome" not like 'Failure%' and date between '04-06-2010' and '20-03-
```

```
 * sqlite:///my_data1.db
Done.
```

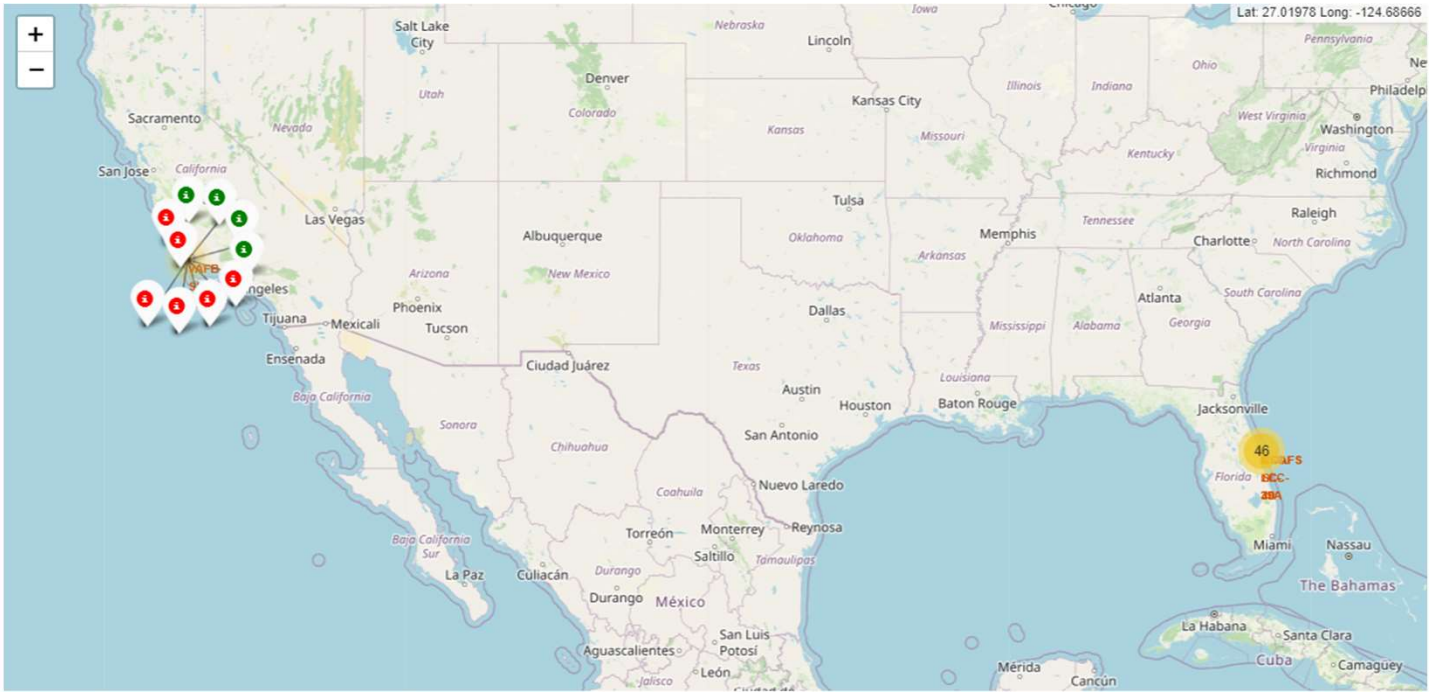
Landing _Outcome	rank
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Controlled (ocean)	3
No attempt	1

Section 3

Launch Sites Proximities Analysis

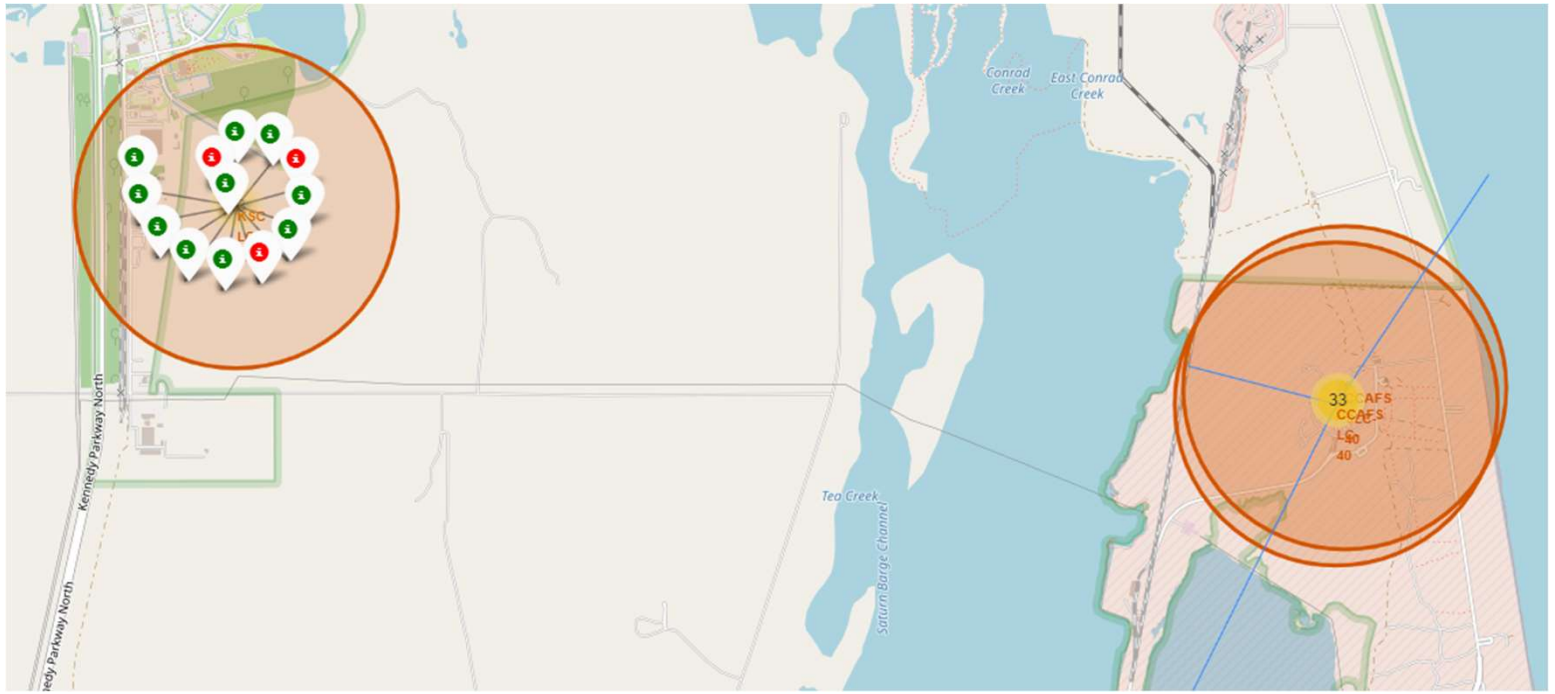
Launch sites

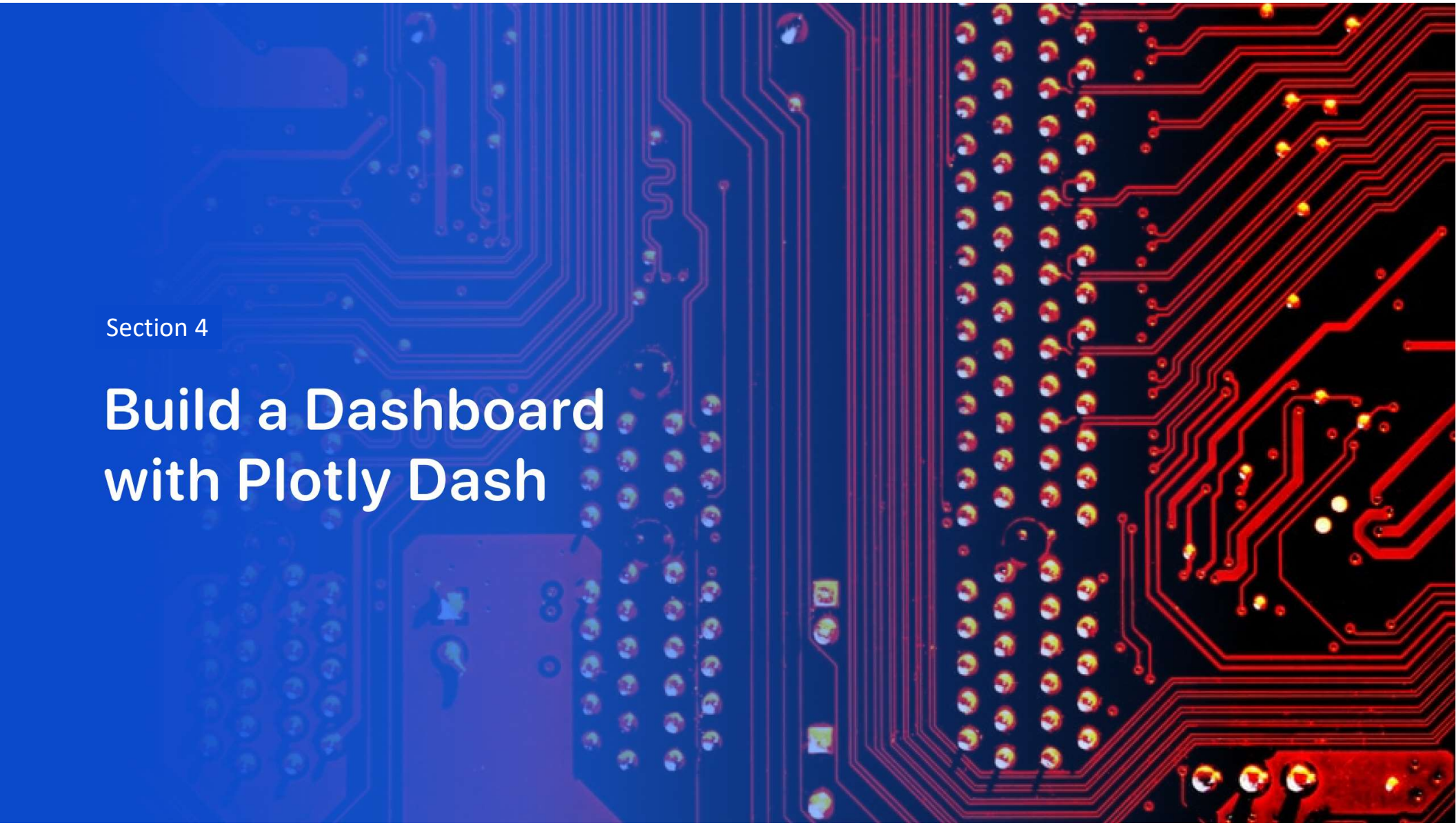
- Location of each site is displayed with a marker. Red marker indicates a failure & green marker indicates a success.
- All the launch sites are clustered based on the location using MarkerCluster.



Launch Site

- Green marker shows successful outcome & red marker shows failure. The launch sites are clustered with a circle.
- Nearest locations & distances of sea, road has been displayed in blue lines.



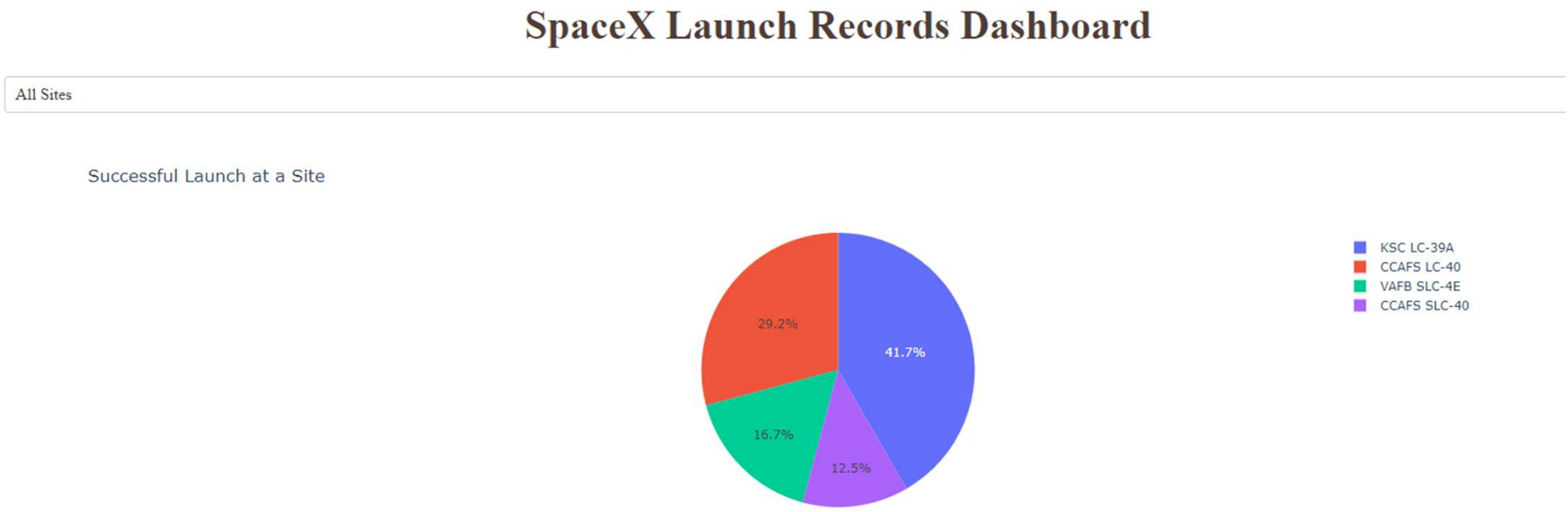


Section 4

Build a Dashboard with Plotly Dash

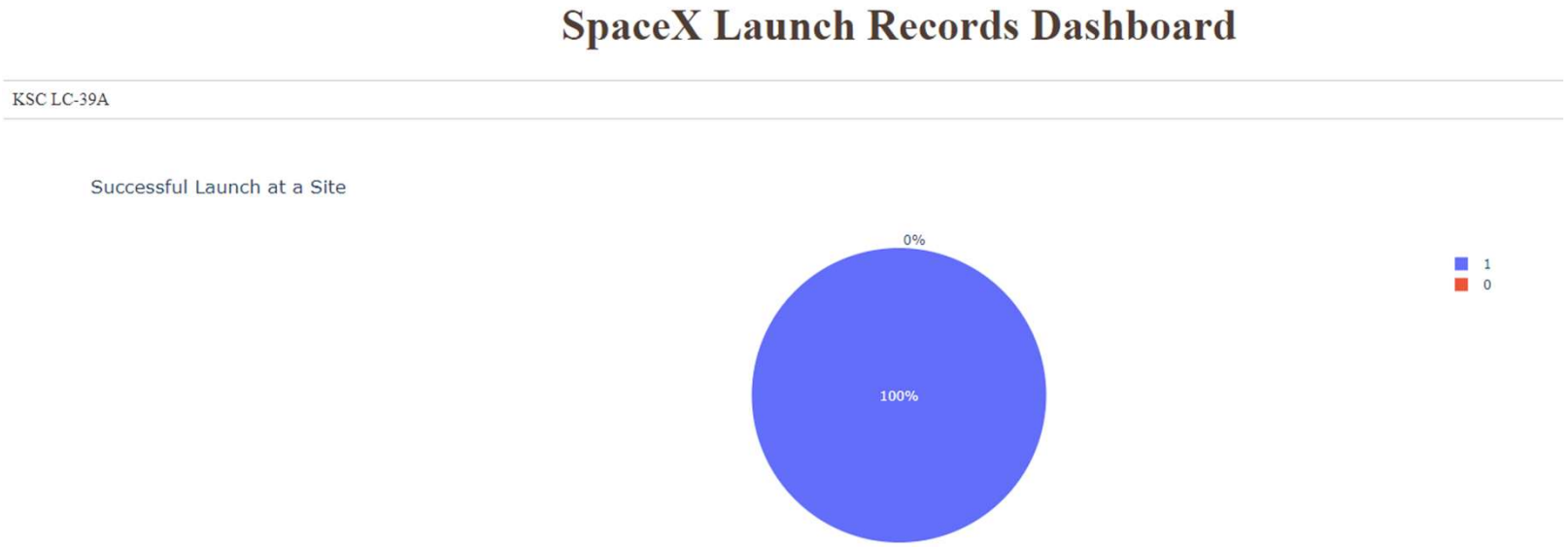
Interactive analysis

- Used pie chart to display the successful launch at particular site.
- Site KSC LC-39A has 41.7% of successful launches.



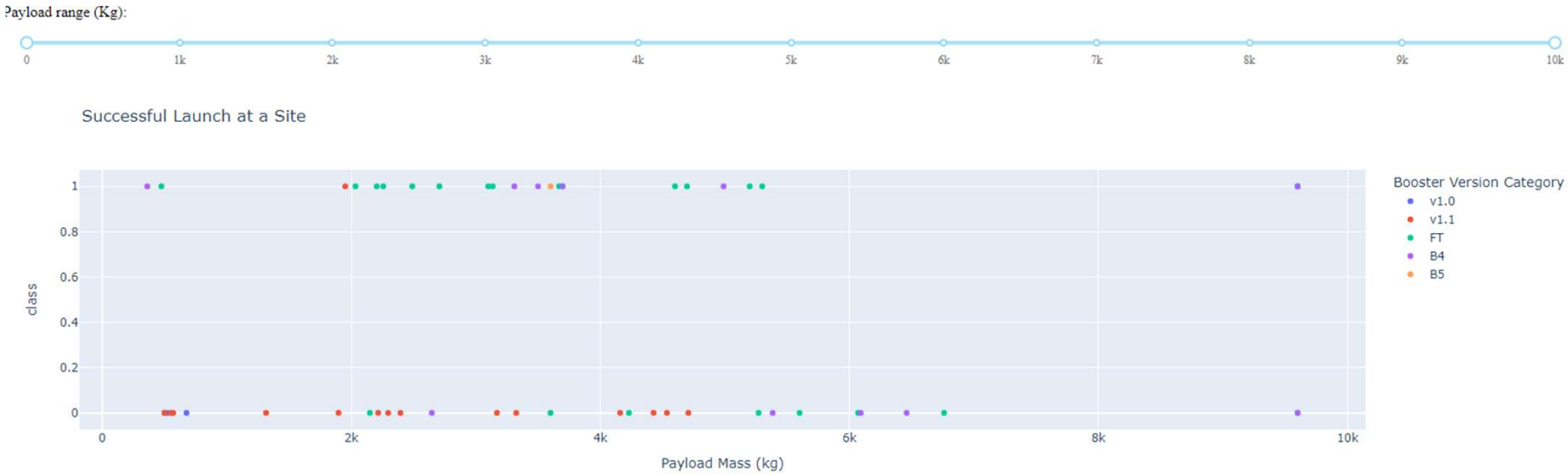
Interactive analysis

- Show the screenshot of the piechart for the launch site with highest launch success ratio
- Explain the important elements and findings on the screenshot



Interactive analysis

- 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.

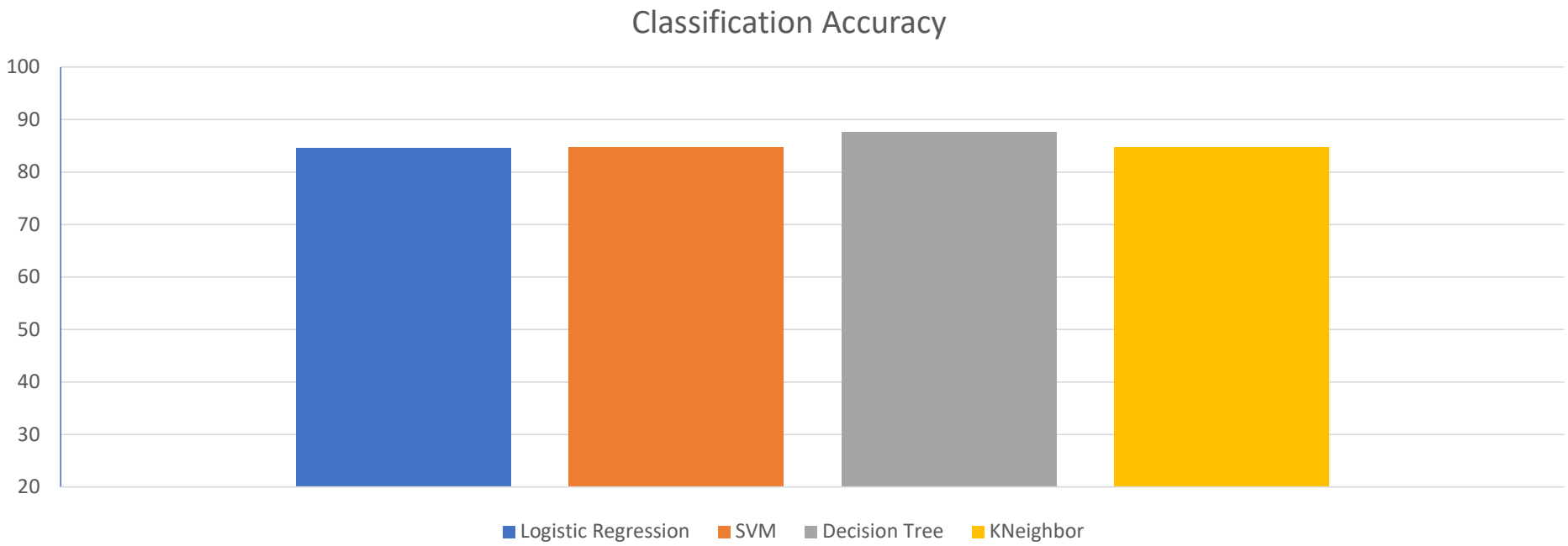


Section 5

Predictive Analysis (Classification)

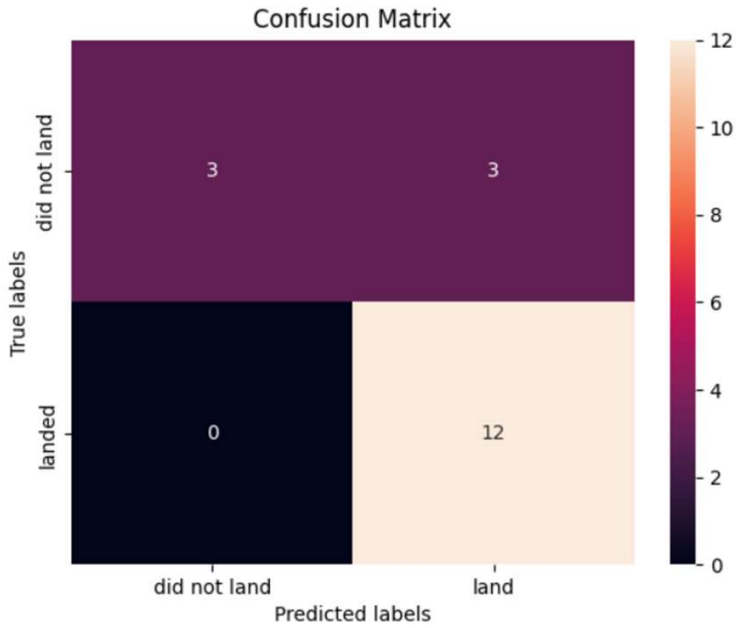
Classification Accuracy

- Decision Tree has 87.67% accuracy



Confusion Matrix

- The model has predicted highest true positives of 12.
- Model didn’t predict any false negatives.



Conclusions

Decision Tree Classifier has the best accuracy score of 84.82%.

After evaluating using different models we can see all the models have attanined similar accuracy score of 83.33% on the test data.

Thank you!