IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Varun Sharma
13-Nov-2021

Outline

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

Executive Summary

.

- **Summary of methodologies:**
  - ‘Web scraping’ using API calls.
  - Python packages like Pandas,Numpy,Matplotlib, plotly, dash, folium, seaborn.
  - Machine learning packages like sklearn with sub packages of tree,svm,neighbors(knn)model selection.

- **Summary of all results’:**
  - On increasing the payload mass too much, the chances are the rocket will crash.
  - ‘KSC LC-39A’ Launch site is more suited to launch spaceships as compared to other sites.
  - GTO orbit type should be avoided as it has least success rate.

Introduction

SpaceX is the only private company ever to return a spacecraft from low-earth orbit, which it first accomplished in December 2010. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars wheras 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.

Our Goal is to use this data to predict whether SpaceX will attempt to land a rocket or not?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Details on further slides.
- Perform data wrangling
  - Details on further slides.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Details on further slides.

Data Collection

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

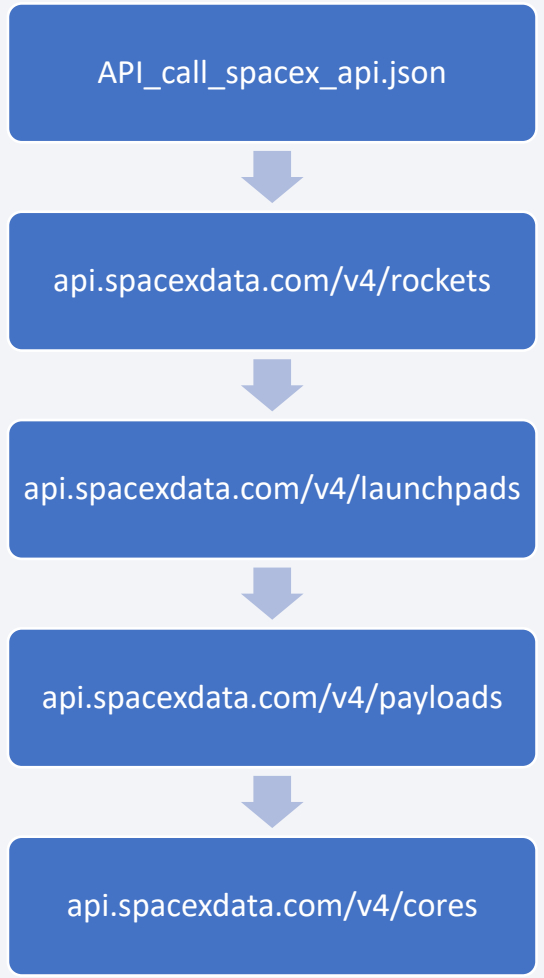
Data Collection – SpaceX API

- 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 Link

Data is collected by making a 'get request' to Web APIs.

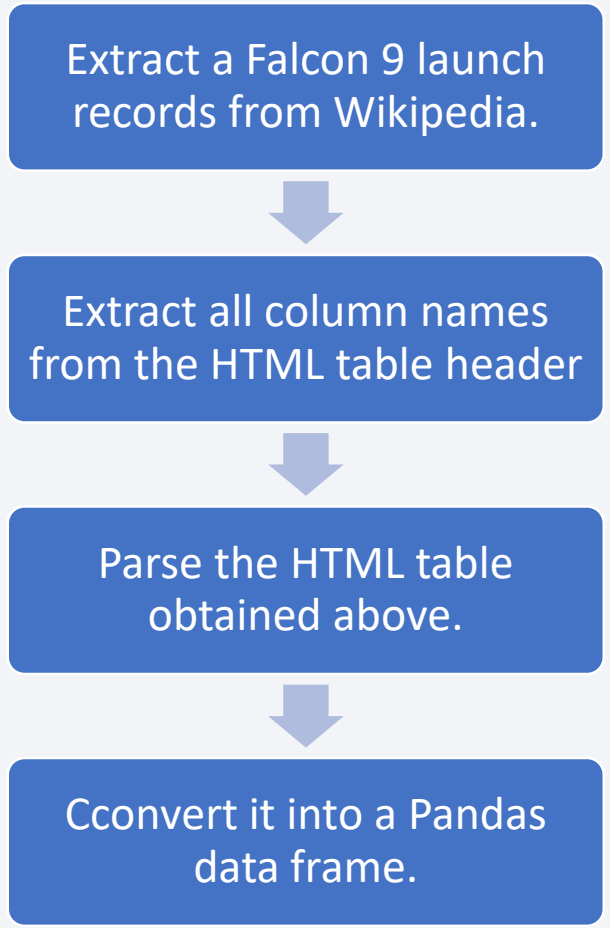


Data Collection - Scraping

- 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

Falcon 9 launch records was Web Scrapped' with 'BeautifulSoup' function and below steps were followed:



Github Link

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

As part of this stage, objective is to find some patterns in the data and determine what would be the label for training supervised models.

Specifically, following steps were performed:

1. Calculate the number of launches on each site.
2. Calculate the number and occurrence of each orbit.
3. Calculate the number and occurrence of mission outcome per orbit type.
4. Create a landing outcome label from Outcome column.

Github Link

EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
- Add the GitHub URL of your completed EDA with data visualization notebook, as an external reference and peer-review purpose

Following charts were plotted as part of this exercise:

1. Scatter point chart to show how 'FlightNumber' & 'Payload' affects 'Launch outcome'.
2. Scatter point chart to show how 'FlightNumber' & 'Launch Site' affects 'Launch outcome'.
3. Scatter point chart to show how 'Launch Site' & 'Payload' affects 'Launch outcome'.
4. Bar chart to check for relationship between 'Success rate & Orbit type'.
5. Scatter point chart to show how 'FlightNumber' & 'Orbit Type Site' affects 'Launch outcome'.
6. Scatter point chart to show how 'Payload' & 'Orbit Type' affects 'Launch outcome'.
7. Line plot to show the launch success yearly trend.

Github Link

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
- Add the GitHub URL of your completed EDA with SQL notebook, as an external reference and peer-review purpose

Following SQL Queries were performed on the Dataset:

1. *Display the names of the unique launch sites in the space mission.*
2. *Display 5 records where launch sites begin with the string 'CCA'.*
3. *Display the total payload mass carried by boosters launched by NASA (CRS)'.*
4. *Display average payload mass carried by booster version F9 v1.1.*
5. *List the date when the first successful landing outcome in ground pad was achieved.*
6. *List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.*
7. *List the total number of successful and failure mission outcomes.*
8. *List the names of the booster versions which have carried the maximum payload mass. Use a subquery.*
9. *List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015.*
10. *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.*

Github Link

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
- Explain why you added those objects
- Add the GitHub URL of your completed interactive map with Folium map, as an external reference and peer-review purpose

Github Link

Objective is to find some geographical patterns about launch sites.

TASK 1: Mark all launch sites on a map using 'Folium Map' object with map of United States centred on the location of the NASA space centre.
TASK 2: Mark the success/failed launches for each site on the map using 'Marker cluster' object.
TASK 3: Calculate the distances between a launch site to its proximities and draw a line using 'PolyLine' object.

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
- Explain why you added those plots and interactions
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

Github Link

1. **Added a Launch Site Drop-down Input Component:**

We have four different launch sites and we would like to first see which one has the largest success count.

2. **Added a Pie chart :**

Based on the site selected from the above added ‘dropdown’, a pie chart is rendered.

3. **Added a Range Slider to Select Payload:**

To find if variable payload is correlated to mission outcome.

4. **Added a scatter plot :**

To observe how payload may be correlated with mission outcomes for selected site(s).

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

Github Link

Following steps were performed at a high level:

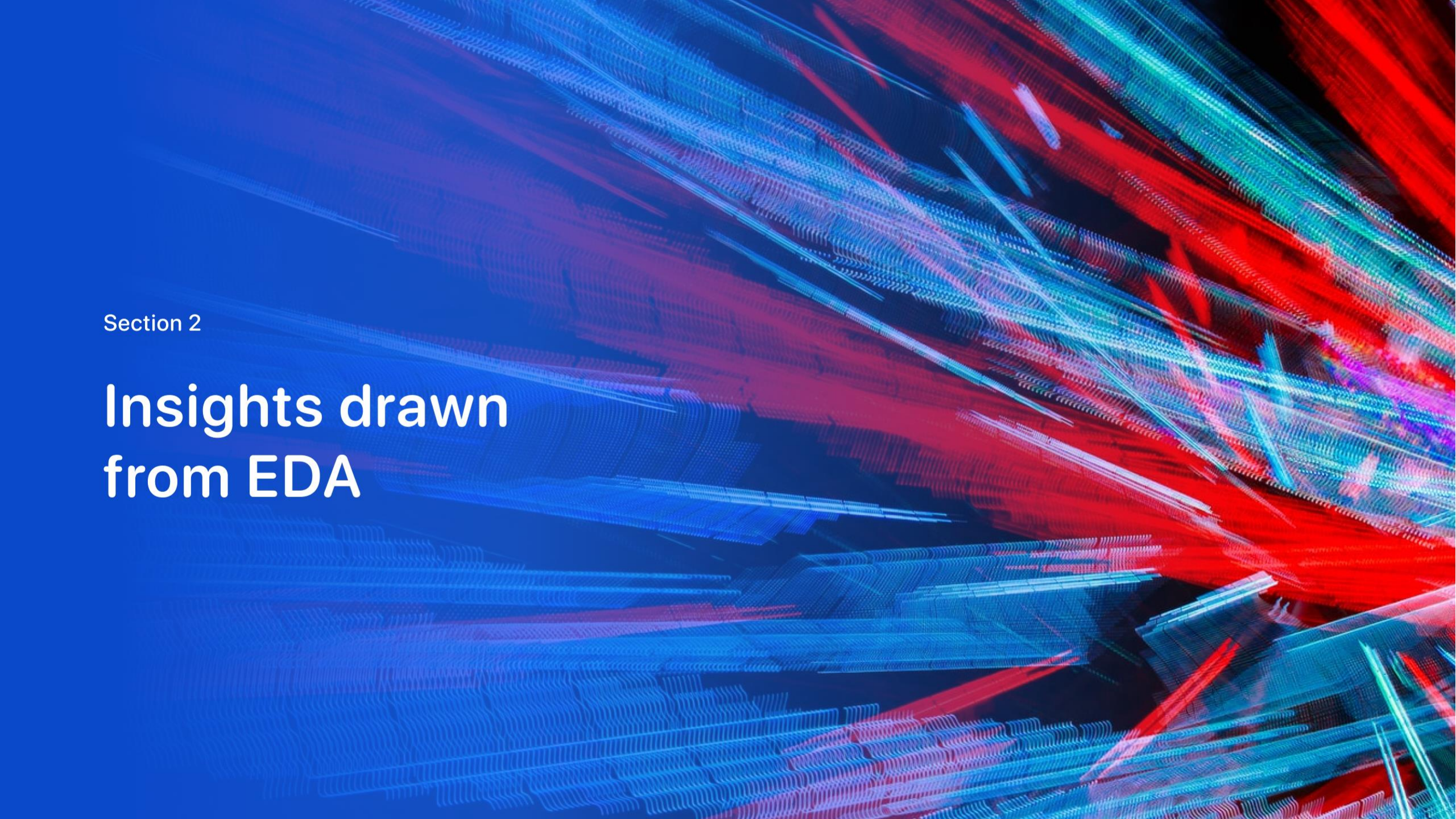
1. Loaded the data into DataFrame.
2. Created a 'Numpy' array from the column 'class' in the DataFrame and assign it variable 'Y'.
3. Standardized the data in X then reassign it to the variable X using the transform **'fit_transform'**.
4. Split data X and Y into training and test data using the function **'train_test_split'** .
5. Found the the best parameters by fitting the 'GridSearchCV for a classifier (logistic regression, SVM etc)'
6. Calculated the accuracy on the test data using the method score and confusion matrix.

Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

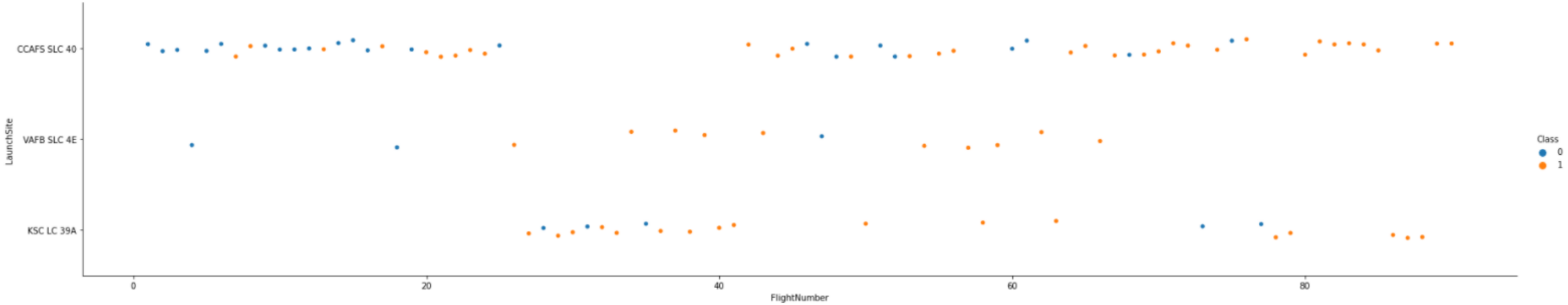
Insights drawn from EDA



Flight Number vs. Launch Site

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

- Show the screenshot of the scatter plot with explanations

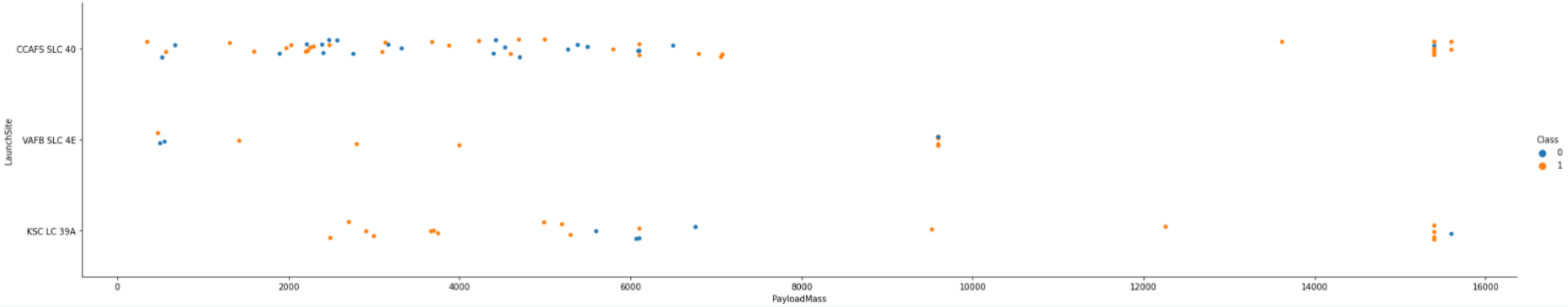


We see that launch site 'VAFB SLC-4E' does not have any 'Flight Number >80. Generally, it can be inferred that higher flight numbers corresponds to 'successful outcome' of launch for all the Launch sites.

Payload vs. Launch Site

- Show a scatter plot of Payload vs. Launch Site

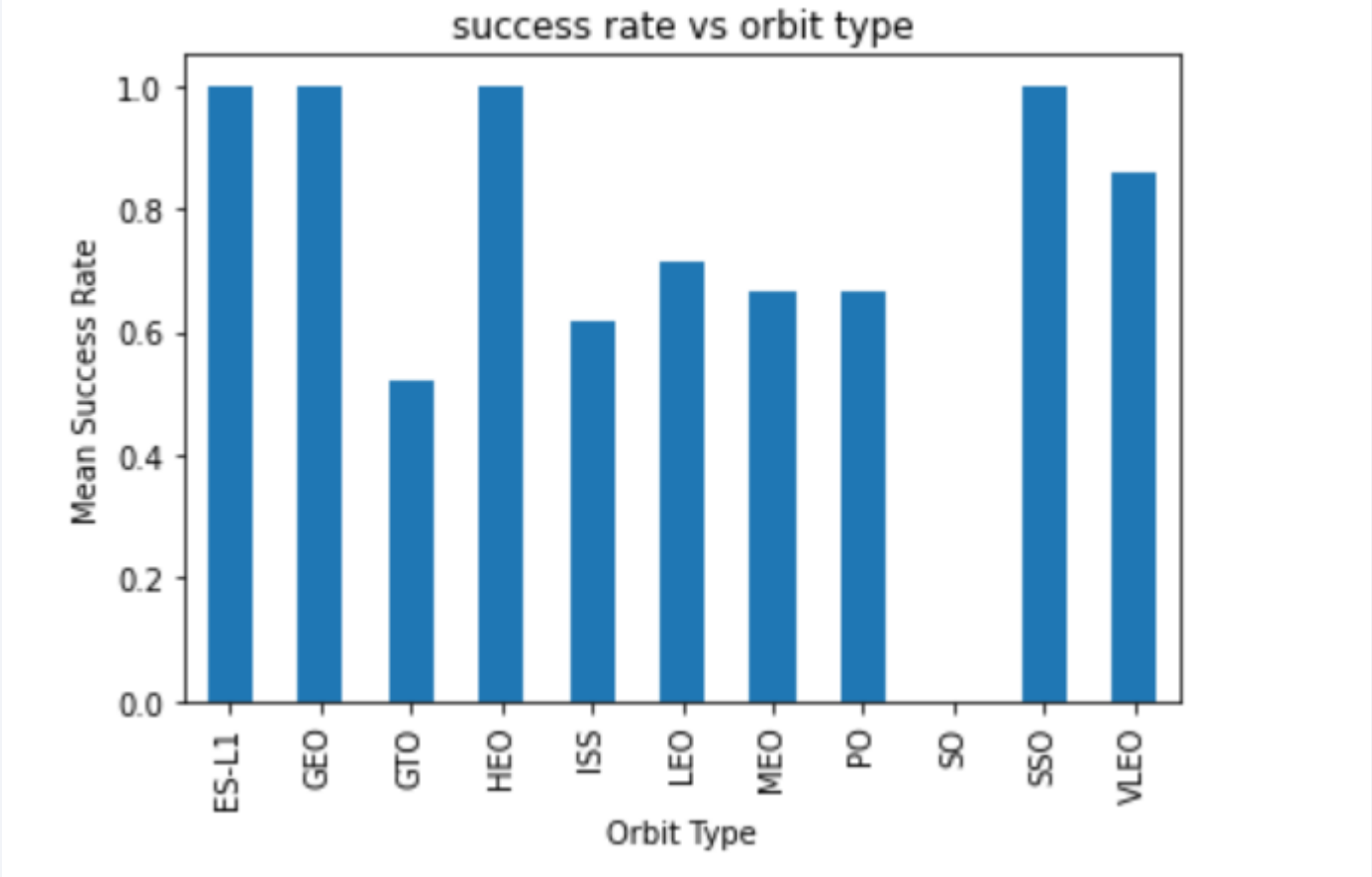
- Show the screenshot of the scatter plot with explanations



We see that for the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).

Success Rate vs. Orbit Type

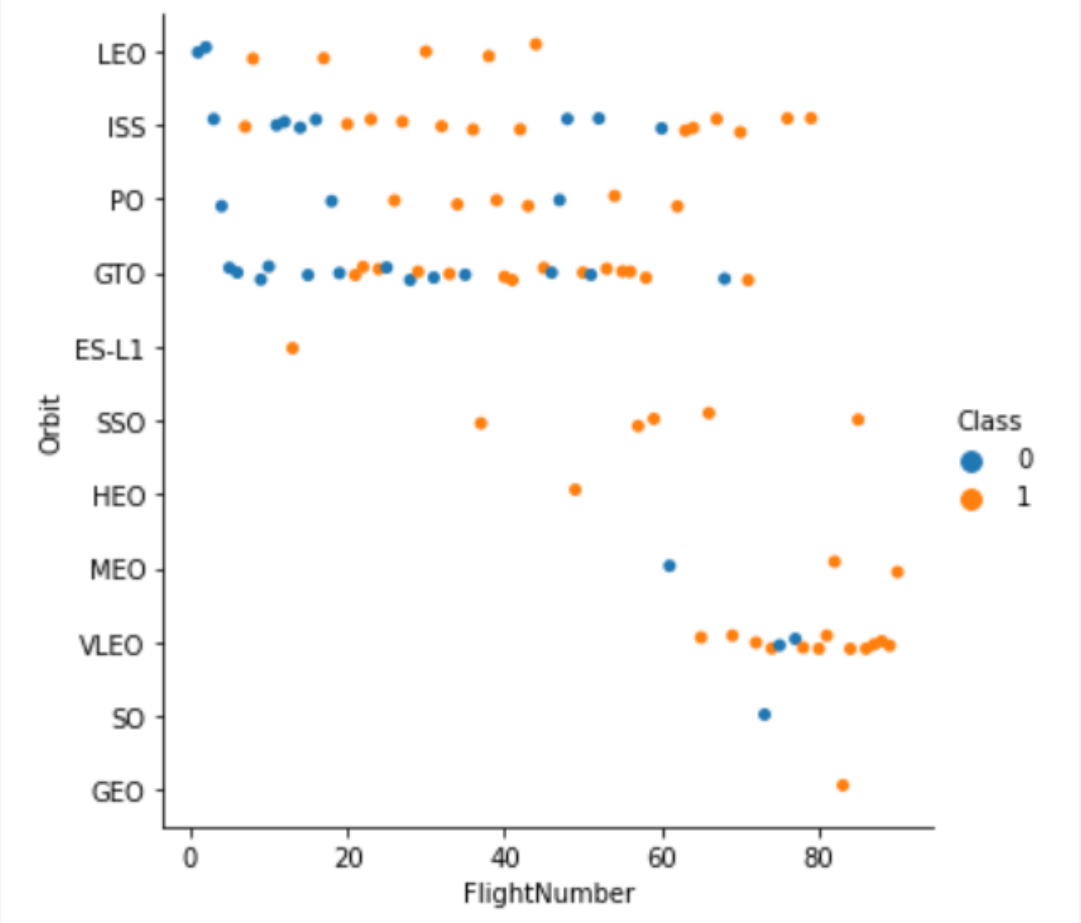
- Show a bar chart for the success rate of each orbit type
- Show the screenshot of the scatter plot with explanations



We see that 'Orbit Type' 'ES-L1', 'GEO','HEO','SSO' all have high success rate (mean rate of 1). Whereas, 'GTO' has the lowest success rate of around 0.52

Flight Number vs. Orbit Type

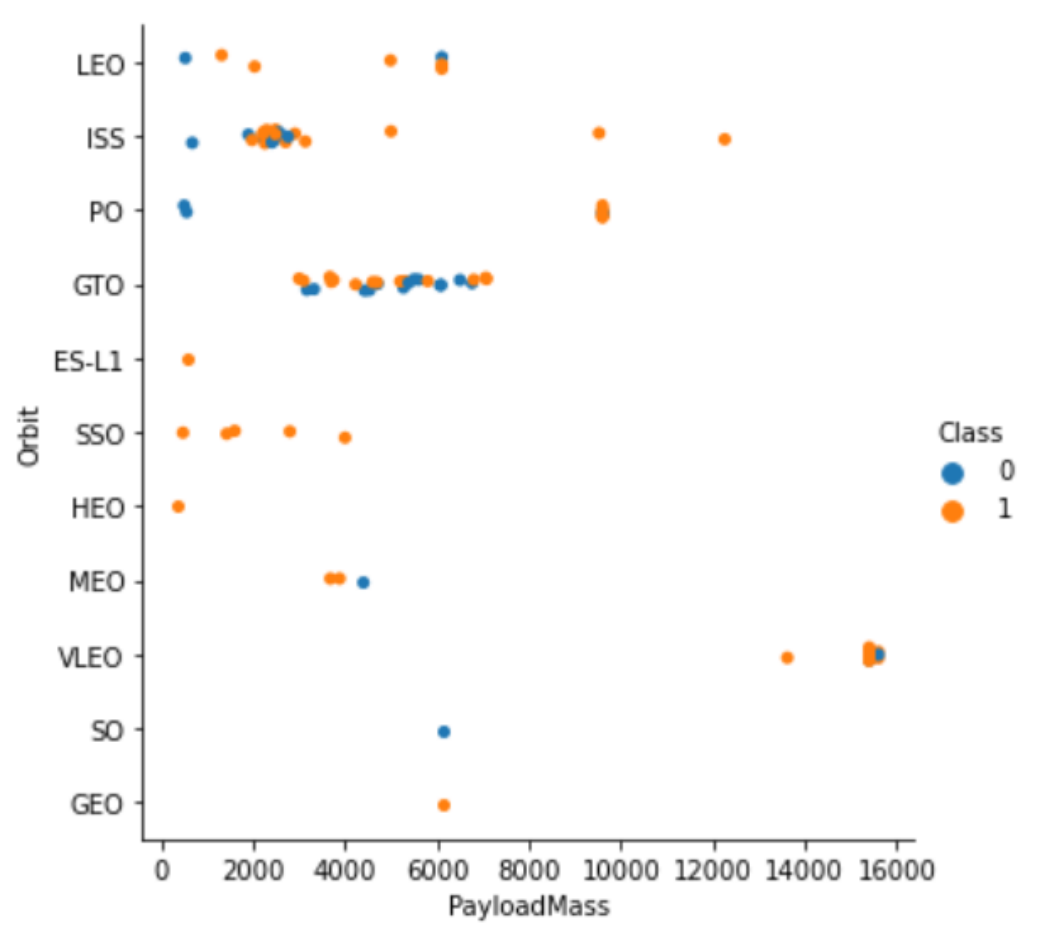
- Show a scatter point of Flight number vs. Orbit type
- Show the screenshot of the scatter plot with explanations



We see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type

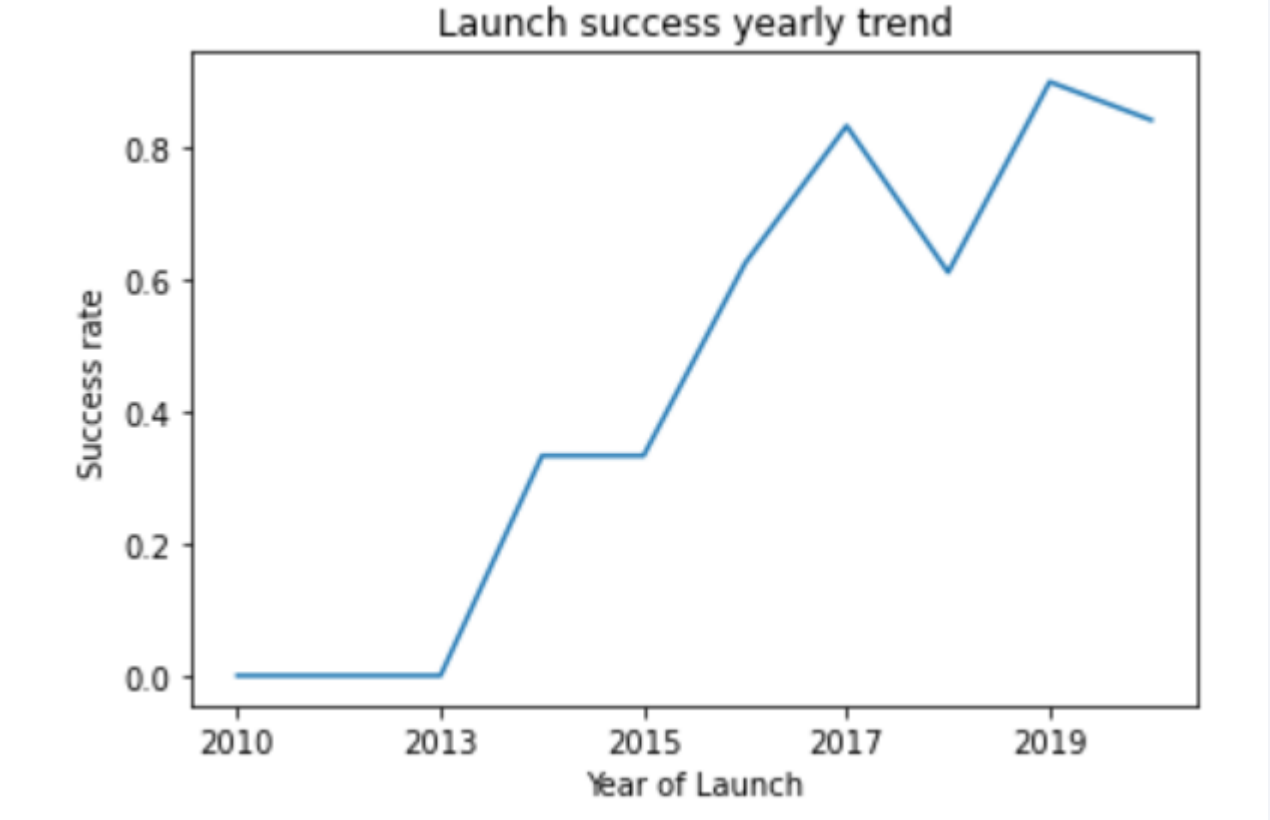
- Show a scatter point of payload vs. orbit type
- Show the screenshot of the scatter plot with explanations



We see that in the with heavy payloads the successful landing rate are more for Polar, LEO and ISS. However for GTO we cannot distinguish this well.

Launch Success Yearly Trend

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



We can observe that the 'success rate' since 2013 kept increasing till 2020, with an exception of 2018 where a dip was observed in the rate.

All Launch Site Names

- Find the names of the unique launch sites
- Present your query result with a short explanation here

```
%sql select distinct LAUNCH_SITE from KLX66977.SPACEXTBL
```

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

These are the unique 'Launch sites' from which the SpaceX rockets are launched!

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 KLX66977.SPACEXTBL WHERE LAUNCH_SITE like 'CCA%' FETCH FIRST 5 ROWS ONLY
```

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

Five records where Launch Site name begins with 'CCA'.

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- Present your query result with a short explanation here

```
%sql select sum(PAYLOAD_MASS__KG_)as "Total PayLoad Mass" from KLX66977.SPACEXTBL where CUSTOMER = 'NASA (CRS)'
```

Total PayLoad Mass
45596

Total Payload Mass in 'Kg' that is carried out by boosters of customer 'NASA'.

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

```
%sql select avg(PAYLOAD_MASS__KG_) as "Avg PayLoad Mass" from KLX66977.SPACEXTBL where BOOSTER_VERSION = 'F9 v1.1'
```

Avg PayLoad Mass
2928

'Average Payload Mass' carried out by F9 v1.1, obtained by querying 'SPACEXTBL' with a where condition on 'Booster version'.

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

```
%sql select min(DATE) as "First Succ Date" from KLX66977.SPACEXTBL where LANDING__OUTCOME like 'Success%' and LANDING__OUTCOME = 'Success (ground pad)'
```

First Succ Date
2015-12-22

Fetching the earliest date when the successful landing was achieved on ‘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

- Present your query result with a short explanation here

```
%sql select CUSTOMER,BOOSTER_VERSION from KLX66977.SPACEXTBL where LANDING__OUTCOME = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000
```

customer	booster_version
SKY Perfect JSAT Group	F9 FT B1022
SKY Perfect JSAT Group	F9 FT B1026
SES	F9 FT B1021.2
SES EchoStar	F9 FT B1031.2

Used ‘Select clause’ along with ‘Where clause’ and ‘between clause’.

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 count(*) as "COUNT",MISSION_OUTCOME from KLX66977.SPACEXTBL group by MISSION_OUTCOME
```

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

We see that almost all mission outcome point to 'success'.

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 BOOSTER_VERSION from KLX66977.SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from KLX66977.SPACEXTBL)
```

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

Used a subquery to determine the boosters which have carried maximum payload mass

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

```
%sql  select landing__outcome,LAUNCH_SITE,BOOSTER_VERSION from KLX66977.SPACEXTBL where LANDING__OUTCOME = 'Failure (drone ship)' and EXTRACT(YEAR from DATE)=2015
```

landing__outcome	launch_site	booster_version
Failure (drone ship)	CCAFS LC-40	F9 v1.1 B1012
Failure (drone ship)	CCAFS LC-40	F9 v1.1 B1015

2 Booster versions retrieved whose landing outcomes was failure in drone ship.

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 select count(LANDING__OUTCOME) as "COUNT",LANDING__OUTCOME from KLX66977.SPACEXTBL WHERE DATE between '2010-06-04' and '2017-03-20' group by LANDING__OUTCOME  order by COUNT desc
```

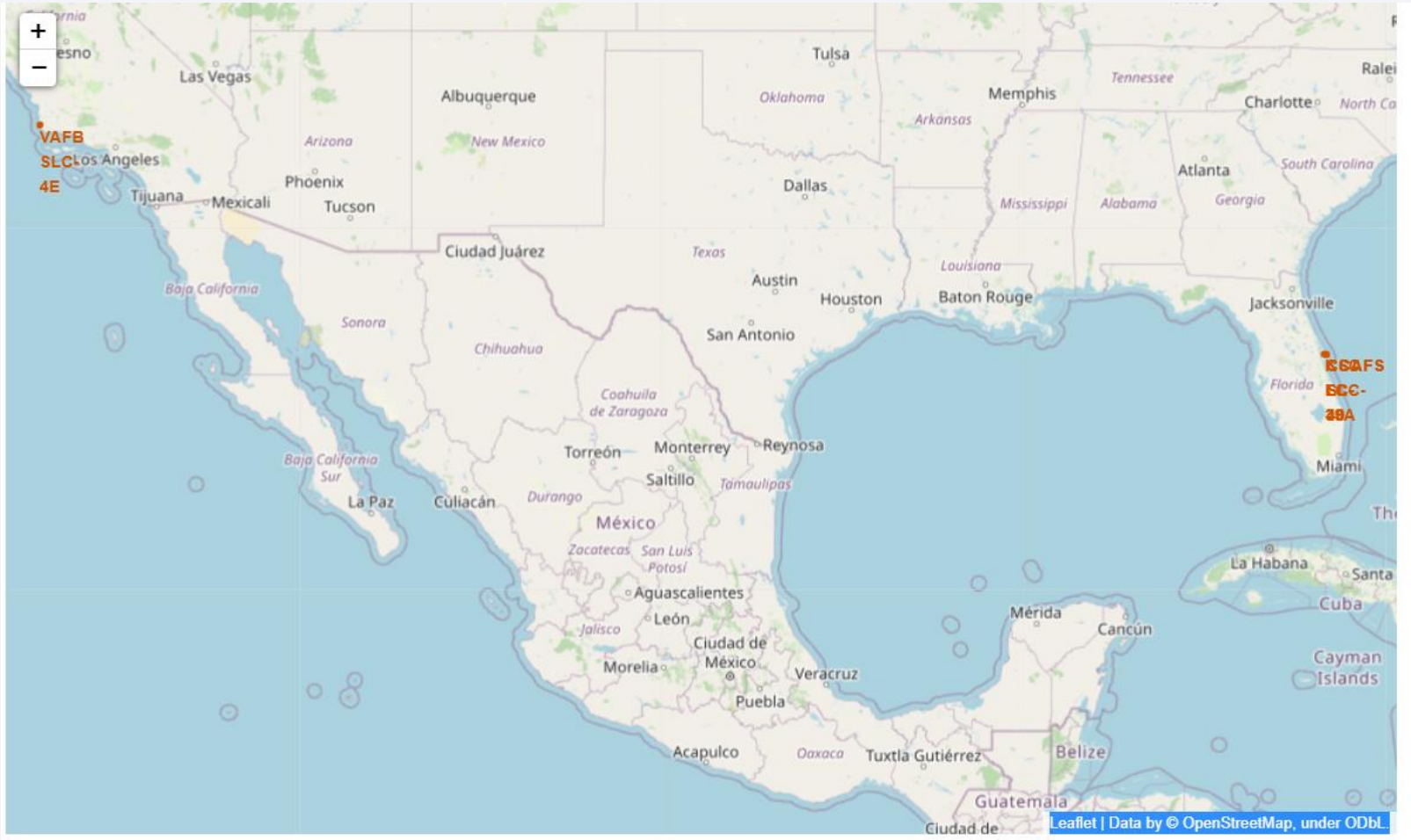
COUNT	landing__outcome
10	No attempt
5	Failure (drone ship)
5	Success (drone ship)
3	Controlled (ocean)
3	Success (ground pad)
2	Failure (parachute)
2	Uncontrolled (ocean)
1	Precluded (drone ship)

We can see that landing outcome of ‘No attempt’ has the highest count.

Section 4

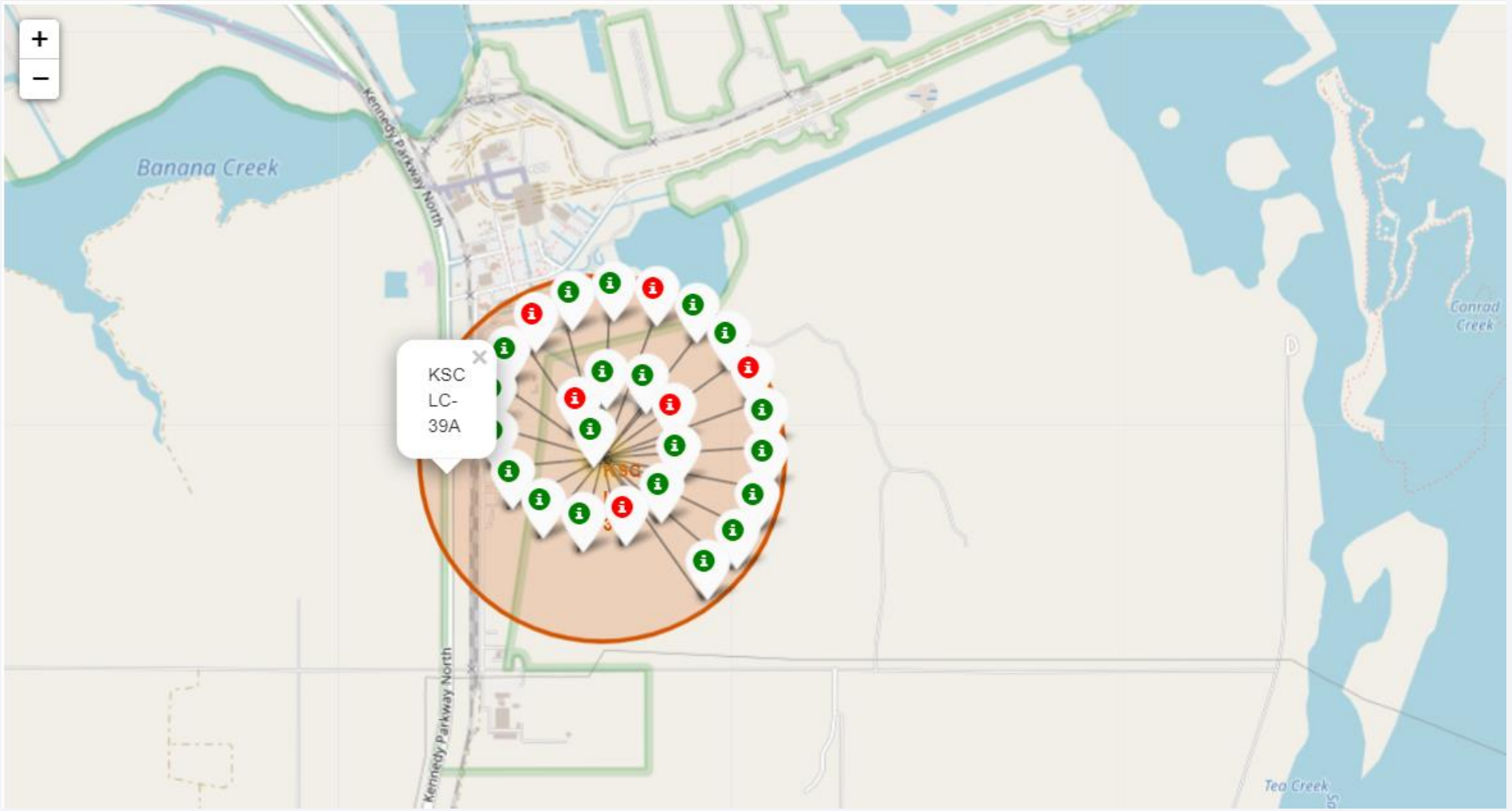
Launch Sites Proximities Analysis

US map with Spacex Launch sites markings



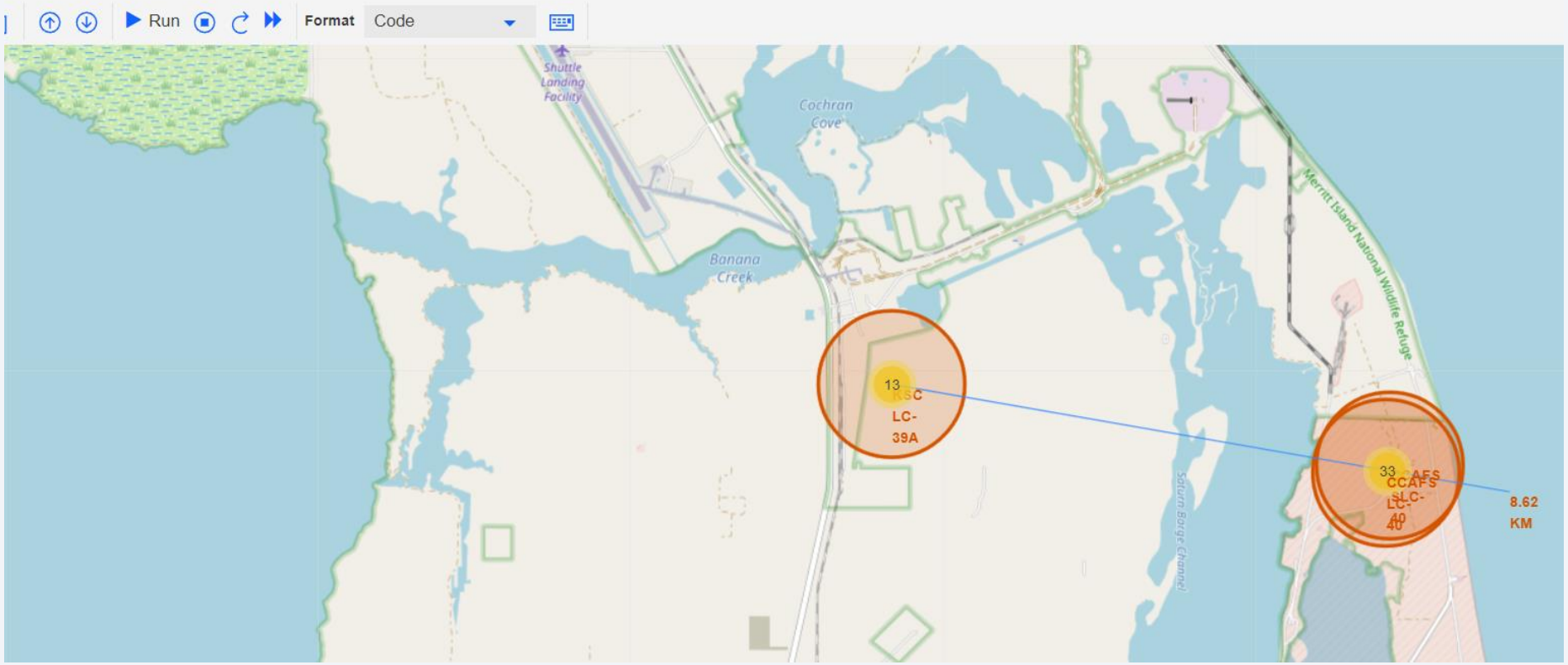
1. It can be observed that Launch sites are in proximity to the Equator line.
2. Also, they are very close to the coast.

Map showing launch outcomes for each site.



It can be seen that launch site ‘KSC LC-39A’ has high success rate depicted by green markers.

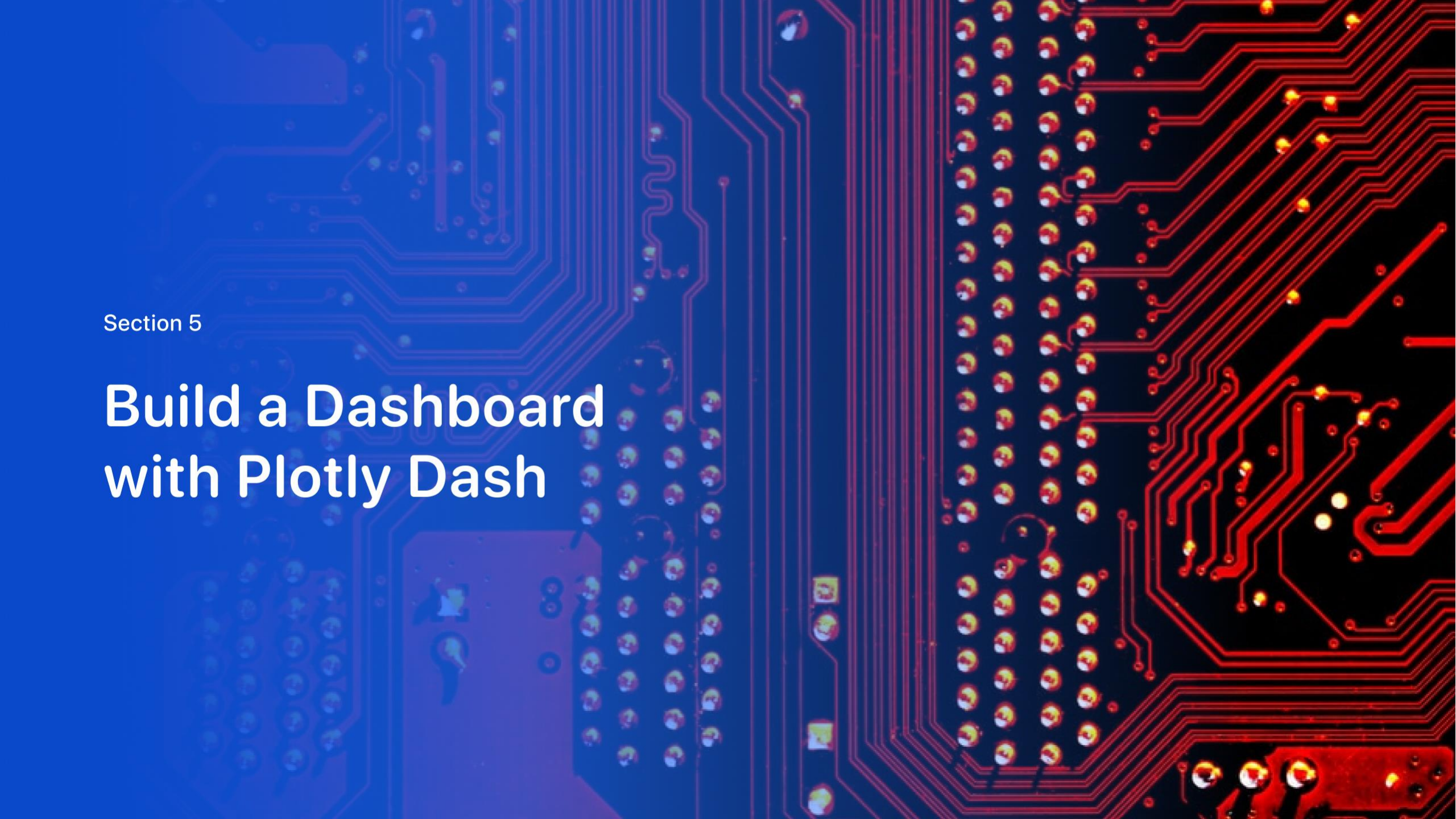
Map depicting the Launch site and its proximity to landmarks such as ‘coast line, railway, highway’.



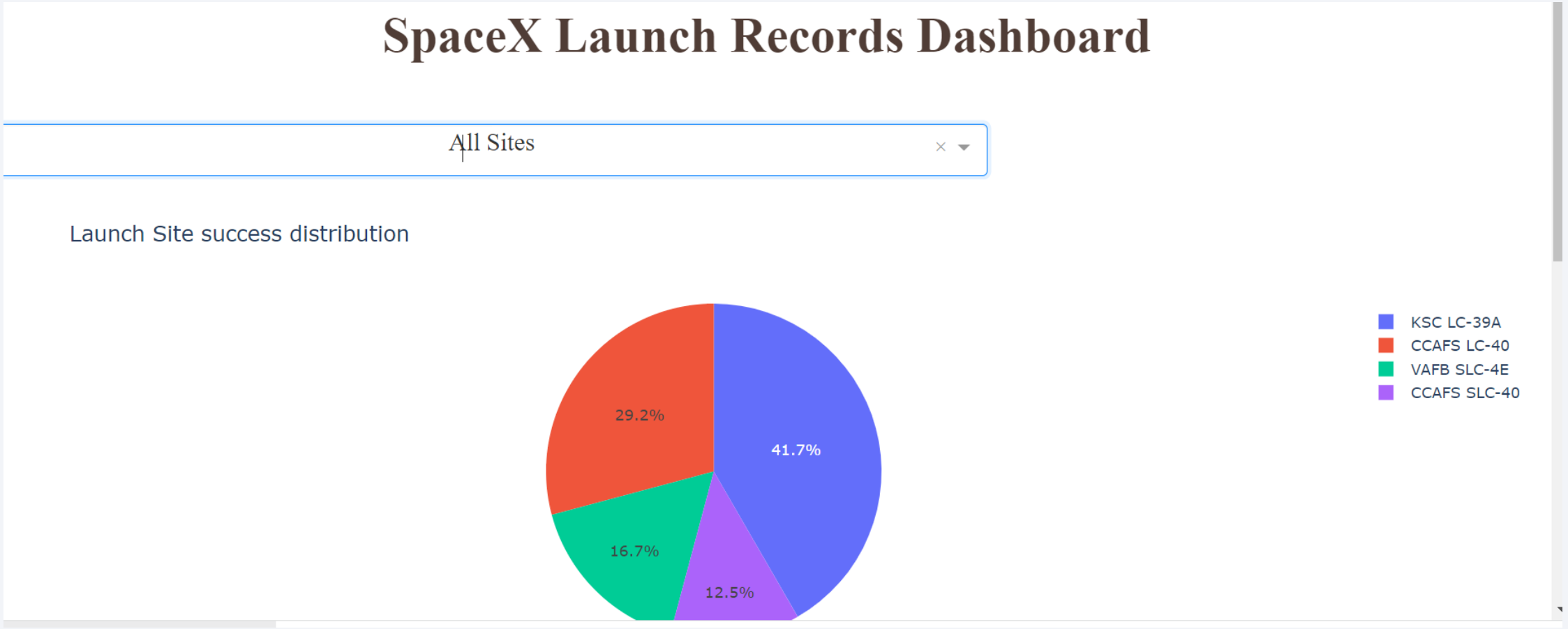
- As seen earlier, launch sites are quite close to coast line (seen here in the figure with a line connecting the two).
- Also, railway lines, highways seem to be close by.
- Another important finding is that the sites are generally away from the cities.

Section 5

Build a Dashboard with Plotly Dash

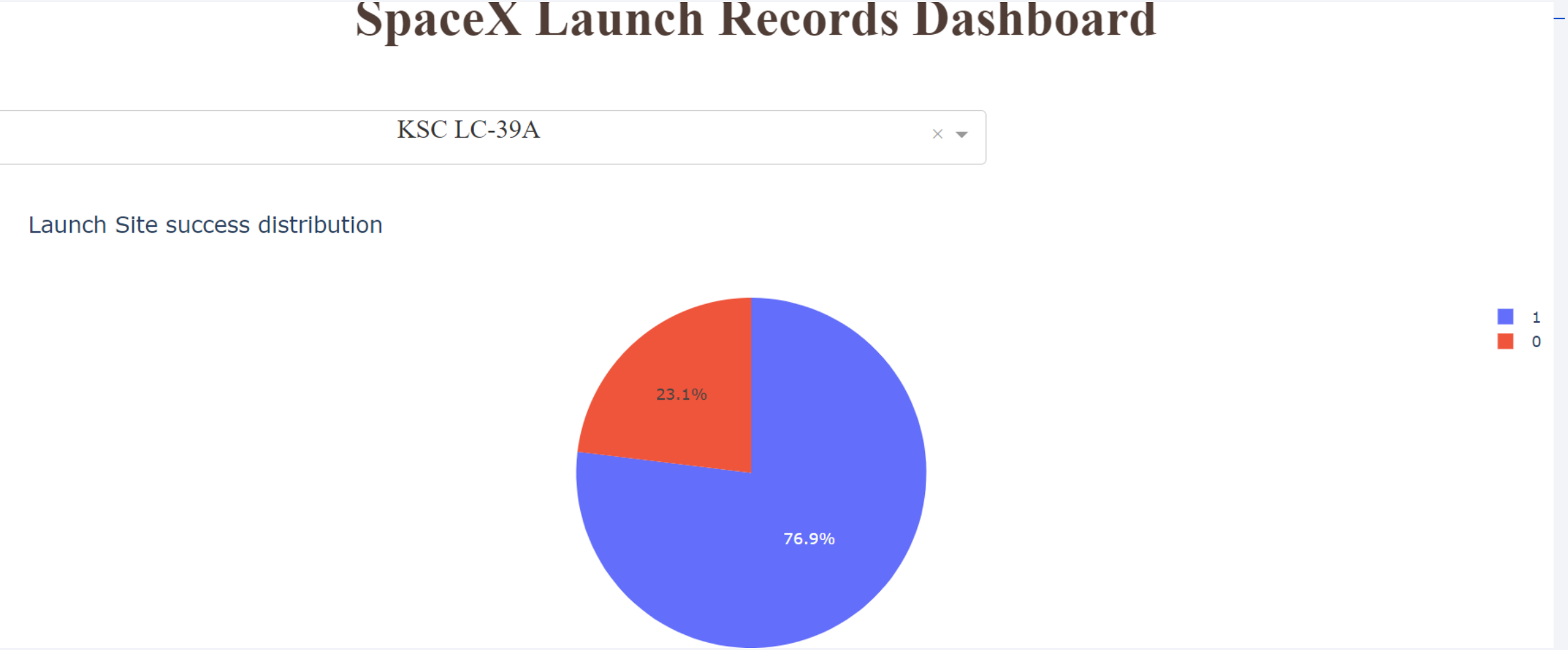


Launch site success distribution



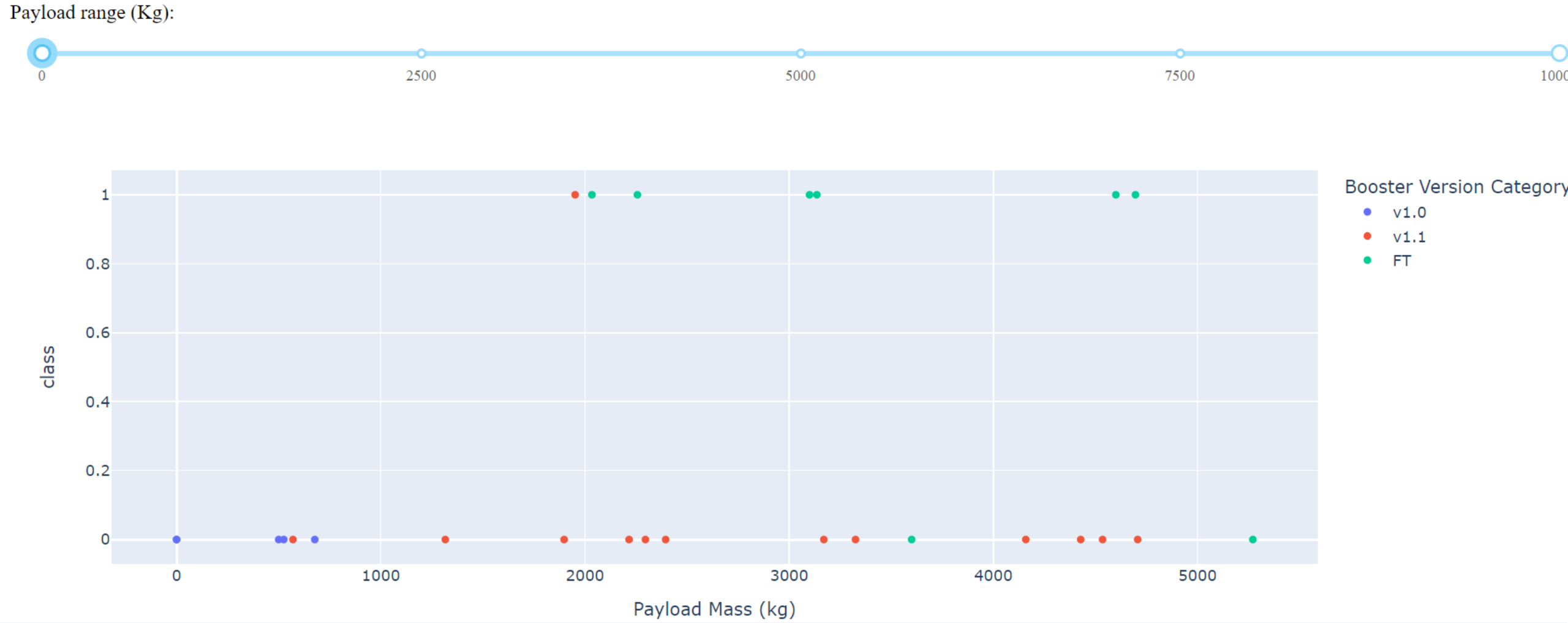
It can be seen that 'KSC LC-39A' has the maximum share of successful launches, while 'CCAFS SLC-40' has lowest percentage of success launches.

Pie chart depicting success ratio for KSC LC-39A



It can be seen that more than 75% of launches done from ‘KSC LC-39A’ are successful.

Success payload scatter plot

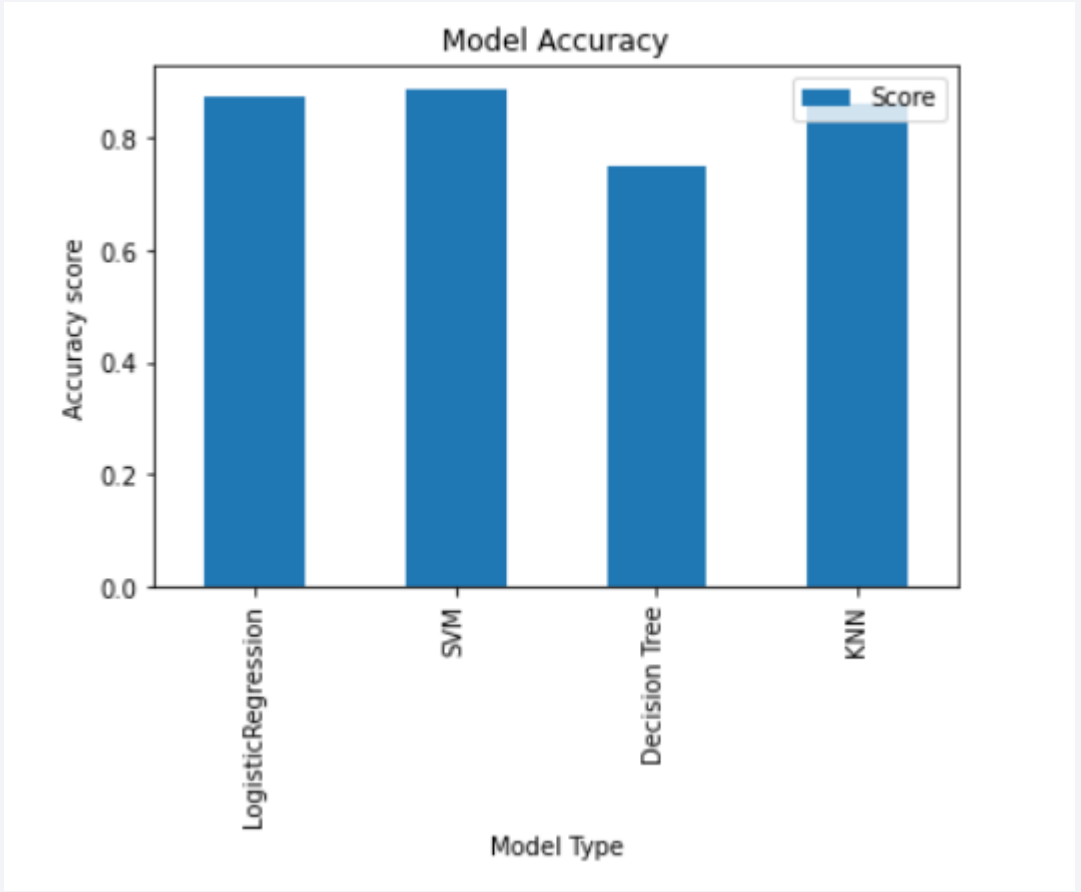


Section 6

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



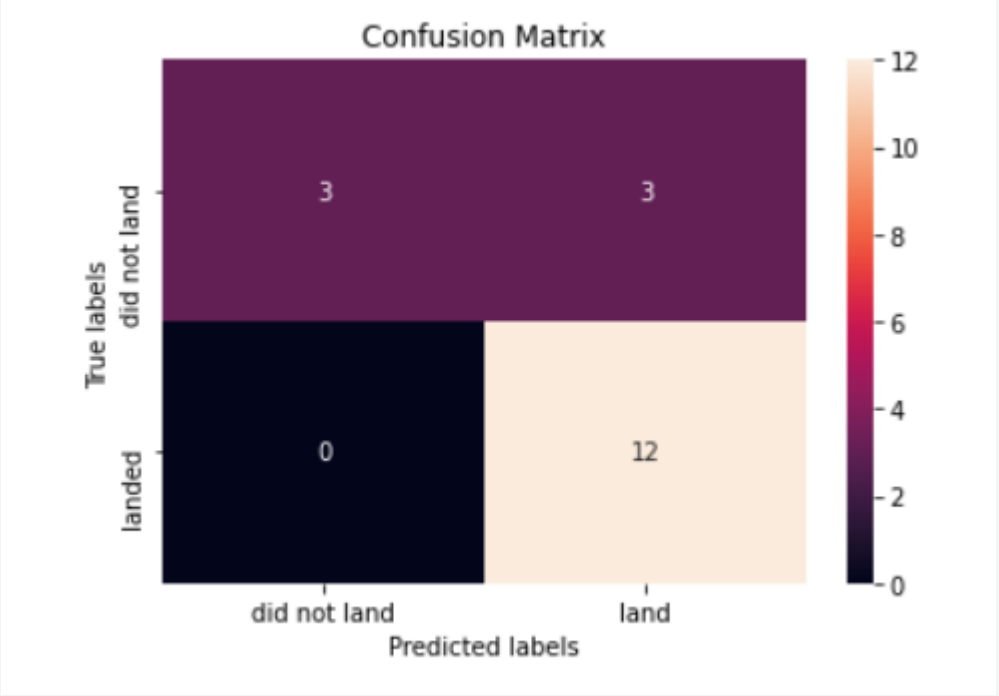
From the bar graph, it is apparent that 'SVM' model has the highest accuracy.

Confusion Matrix

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

Although, this is the highest performing model, We see that the major problem is **false positives.**

False Positive Rate : FP/'Actual No' ➔ 3/6 = 0.5



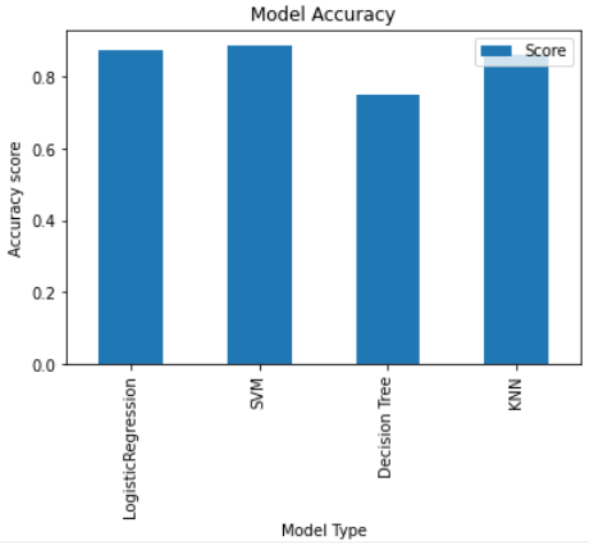
Conclusions

- On increasing the payload mass too much, the chances are the rocket will crash.
- ‘KSC LC-39A’ Launch site is more suited to launch spaceships as compared to other sites.
- GTO orbit type should be avoided as it has least success rate.

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

```
# Bar chart to compare different model's accuracies:
score_frame = pd.DataFrame({'Score':[0.875,0.888,0.75,0.861]},index=['LogisticRegression','SVM','Decision Tree','KNN']);
score_frame.plot(kind = 'bar')
plt.title('Model Accuracy')
plt.xlabel('Model Type')
plt.ylabel('Accuracy score')
plt.show()
score_frame
```



Thank you!