IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Pongpisut Kongdan>
<02-December-2022>

Outline

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

Executive Summary

We create a prediction model for the SpaceX Falcon 9 first stage's successful or unsuccessful landing in this capstone project. If we are aware of whether the first stage will land, we can estimate the cost of a launch. To do this, various machine-learning categorization strategies will be employed.

Data collection, data wrangling and preprocessing, exploratory data analysis, data visualization, and machine learning prediction will all be part of the methodology used.
The findings of our investigation and analysis suggest that there are some characteristics of rocket launches that are correlated with successful or unsuccessful launches.

The result shows that The Decision Tree may be the best machine-learning algorithm for this task.

Introduction

This capstone project's major objective is to foretell if the Falcon 9 first stage will successfully land. SpaceX advertises on their website that their rocket launches cost 62 million while other providers charge upwards of 165 million because they take great pride in being able to reuse the first stage of a rocket launch. The reuse of the first stage is largely responsible for these cost savings. The price of a launch can be calculated if we can know if the first stage will land. If a different business wants to compete with SpaceX for a rocket launch, it may use the information provided here.

This gets us to the main query we are attempting to address: For a specific set of characteristics of a Falcon 9 rocket launch to predict the landing outcomes

Section 1

Methodology

Methodology

Data was gathered using two techniques: web scraping launch information from a Wikipedia article and requesting information from the SpaceX API. The data was then transformed and cleaned using the pandas module in Python.

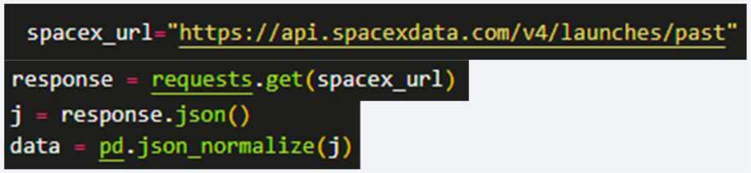
Exploratory data analysis (EDA) was done on the clean data utilizing visualization tools including Python's matplotlib and seaborn packages, as well as SQL queries to provide answers. In order to respond to some analytical queries, interactive visualization packages in Python were employed. Maps were produced using Folium, and interactive data visualizations with Plotly Dash. Perform interactive visual analytics using Folium and Plotly Dash

For the prediction study, four alternative machine learning classification models were employed. The models utilized were decision tree classifier, logistic regression, support vector machines, and k-nearest neighbor. To choose the best model, each one was trained, adjusted, and tested, using suitable evaluators.

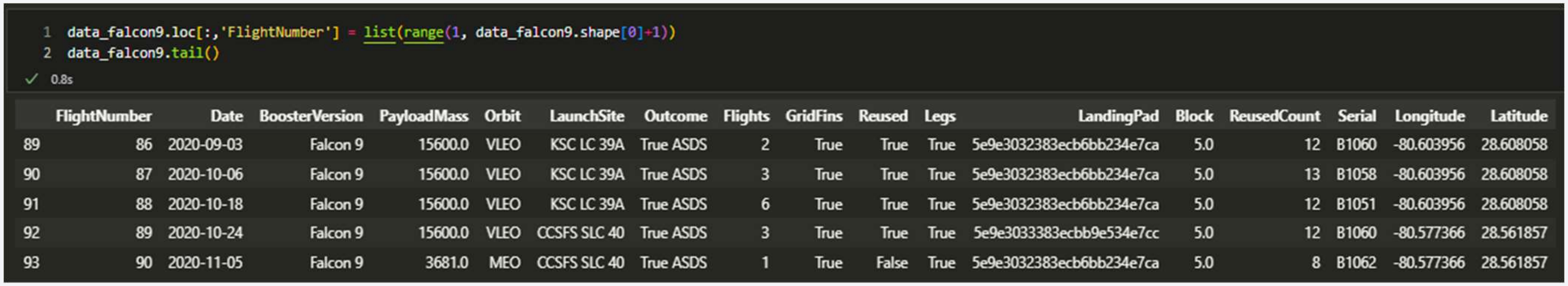
Data Collection – SpaceX API

SpaceX API

1. Request and parse SpaceX laugh data using request.get
2. Normalize JSON to data frame using panda.json_normalize
3. Take a subset of the data frame keeping only useful features
4. Filter data frame to include only Falcon 9 launches
5. Handling missing values



```
spacex_url="https://api.spacexdata.com/v4/launches/past"
response = requests.get(spacex_url)
j = response.json()
data = pd.json_normalize(j)
```



```
data_falcon9.loc[:,'FlightNumber'] = list(range(1, data_falcon9.shape[0]+1))
data_falcon9.tail()
```

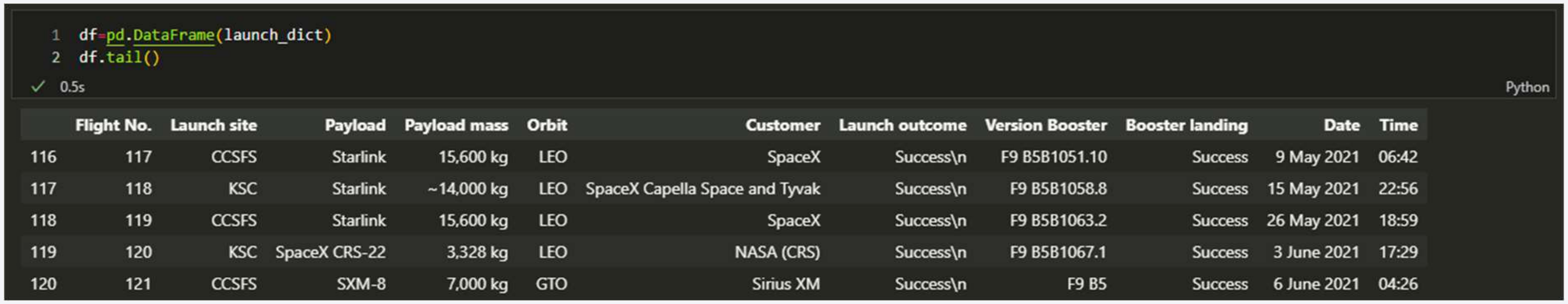
✓ 0.8s

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
89	86	2020-09-03	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	2	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	12	B1060	-80.603956	28.608058
90	87	2020-10-06	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	3	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	13	B1058	-80.603956	28.608058
91	88	2020-10-18	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	6	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	12	B1051	-80.603956	28.608058
92	89	2020-10-24	Falcon 9	15600.0	VLEO	CCSFS SLC 40	True ASDS	3	True	True	True	5e9e3033383ecbb9e534e7cc	5.0	12	B1060	-80.577366	28.561857
93	90	2020-11-05	Falcon 9	3681.0	MEO	CCSFS SLC 40	True ASDS	1	True	False	True	5e9e3032383ecb6bb234e7ca	5.0	8	B1062	-80.577366	28.561857

GitHub Link (Ctrl + Click): Data Collection

Data Collection - Scraping

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



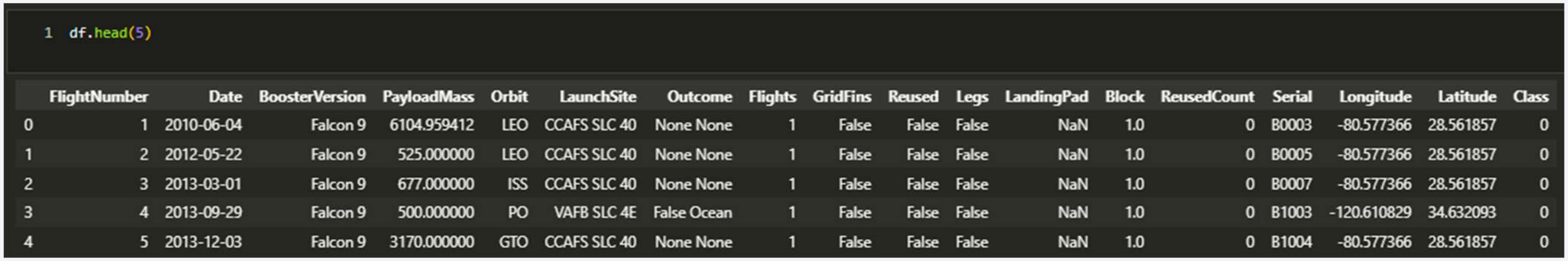
```python
df=pd.DataFrame(launch_dict)
df.tail()
```

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
116	117	CCSFS	Starlink	15,600 kg	LEO	SpaceX	Success\n	F9 B5B1051.10	Success	9 May 2021	06:42
117	118	KSC	Starlink	~14,000 kg	LEO	SpaceX Capella Space and Tyvak	Success\n	F9 B5B1058.8	Success	15 May 2021	22:56
118	119	CCSFS	Starlink	15,600 kg	LEO	SpaceX	Success\n	F9 B5B1063.2	Success	26 May 2021	18:59
119	120	KSC	SpaceX CRS-22	3,328 kg	LEO	NASA (CRS)	Success\n	F9 B5B1067.1	Success	3 June 2021	17:29
120	121	CCSFS	SXM-8	7,000 kg	GTO	Sirius XM	Success\n	F9 B5	Success	6 June 2021	04:26

GitHub Link (Ctrl + Click): Data Collection Scraping

Data Wrangling

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



```
df.head(5)
```

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude	Class
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.561857	0
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.561857	0
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.561857	0
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610829	34.632093	0
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28.561857	0

GitHub Link (Ctrl + Click): Data Wrangling

EDA with Data Visualization

- Scatter plots were used to show how two variables related to one another. There were comparisons between various feature sets, including Flight Number vs. Launch Site, Payload vs. Launch Site, Flight Number vs. Orbit Type, and Payload vs. Orbit Type.
- By using bar charts, it is simple to quickly compare numbers between several groupings. A discrete value is represented by the y-axis, while a category is represented by the x-axis. To compare the Success Rate for various Orbit Types, bar charts were employed.
- A line chart was employed. For displaying data trends across time, line charts are helpful. To display the success rate over a certain number of years.

GitHub Link (Ctrl + Click): EDA with Viz

EDA with SQL

- The following is a list of some of the SQL queries that were run on the dataset:naming the distinct launch sites in the space mission and displaying their names displaying 5 entries for launch sites where the first letter is "CCA"showing the total payload weight carried by NASA-launched rockets (CRS) showing the average payload weight carried by booster F9 v1.1
- The names of the boosters with successful landing outcomes in drone ships and payload masses greater than 4000 but less than 6000, the total number of successful and unsuccessful mission outcomes, the names of the booster versions that carried the maximum payload mass, the failed landing outcomes in drone ships, their booster versions, and the launch site na decreasing order of the number of landing outcomes between 2010-06-04 and 2017-03-20.

GitHub Link (Ctrl + Click): EDA with SQL

Build an Interactive Map with Folium

- A Folium map was updated with new objects. All launch sites and successful/failed launches for each site were shown on a map using marker objects. The distances between a launch location and its close by features, such as the coastline, railroads, highways, and cities, were calculated using line objects.

GitHub Link (Ctrl + Click): Map with Folium

*** If you can not see the folium map, please download the file and run it on your local machine. This problem is probably due to the execution of Javascript is being blocked for some reason, try changing browsers might solve the problem.

Build a Dashboard with Plotly Dash

- A pie chart displaying each site's successful launch. This graph is helpful since it allows you to display the success rate of launches on specific sites or visualize the distribution of landing

- A scatter diagram illustrating the relationship between landing success and the mass of various boosters. The site(s) and payload mass are the dashboard's two inputs. This chart is helpful since it allows you to see how different factors influence the results of the landing.

GitHub Link (Ctrl + Click): Dash

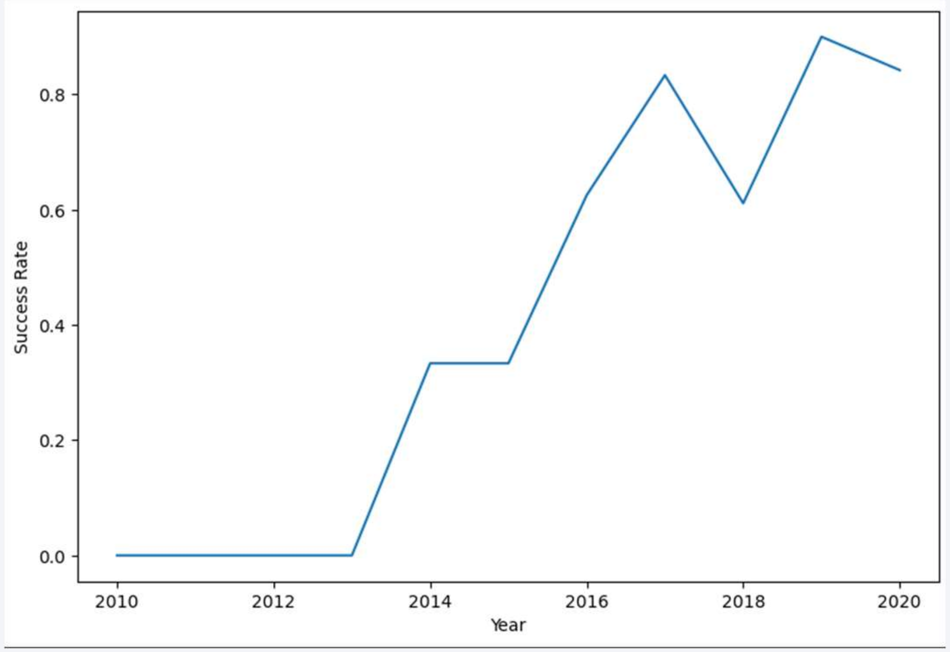
Predictive Analysis (Classification)

1. Create a NumPy array from the column ‘Class’ as ‘Y’
2. Standardizing the features ‘X’
3. Split data into training and test set
4. Create the Logistic regression model, perform prediction, and evaluate the model
5. Create the SVM model, perform prediction, and evaluate the model
6. Create the Decision Trees model, perform prediction, and evaluate the model
7. Create the K-Nearest Neighbours model, perform prediction, and evaluate the model
8. Compare All models based on their accuracy scores and confusion matrix

GitHub Link (Ctrl + Click): Predictive Analysis

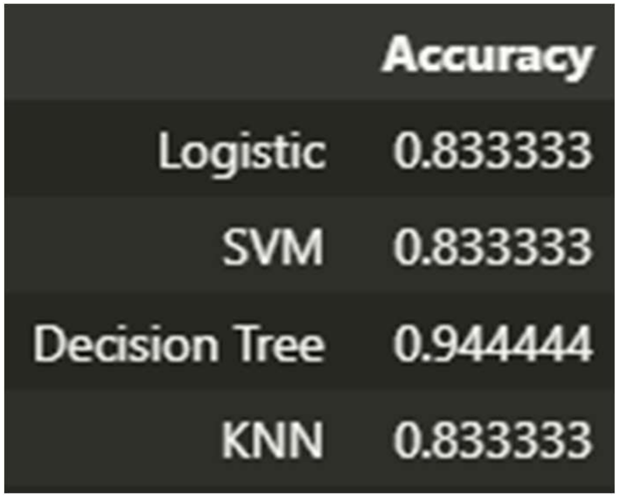
Results: Exploratory Analysis

According to the findings of the exploratory data analysis, the success percentage of the Falcon 9 landings was 66.66%, and it increased over the years.



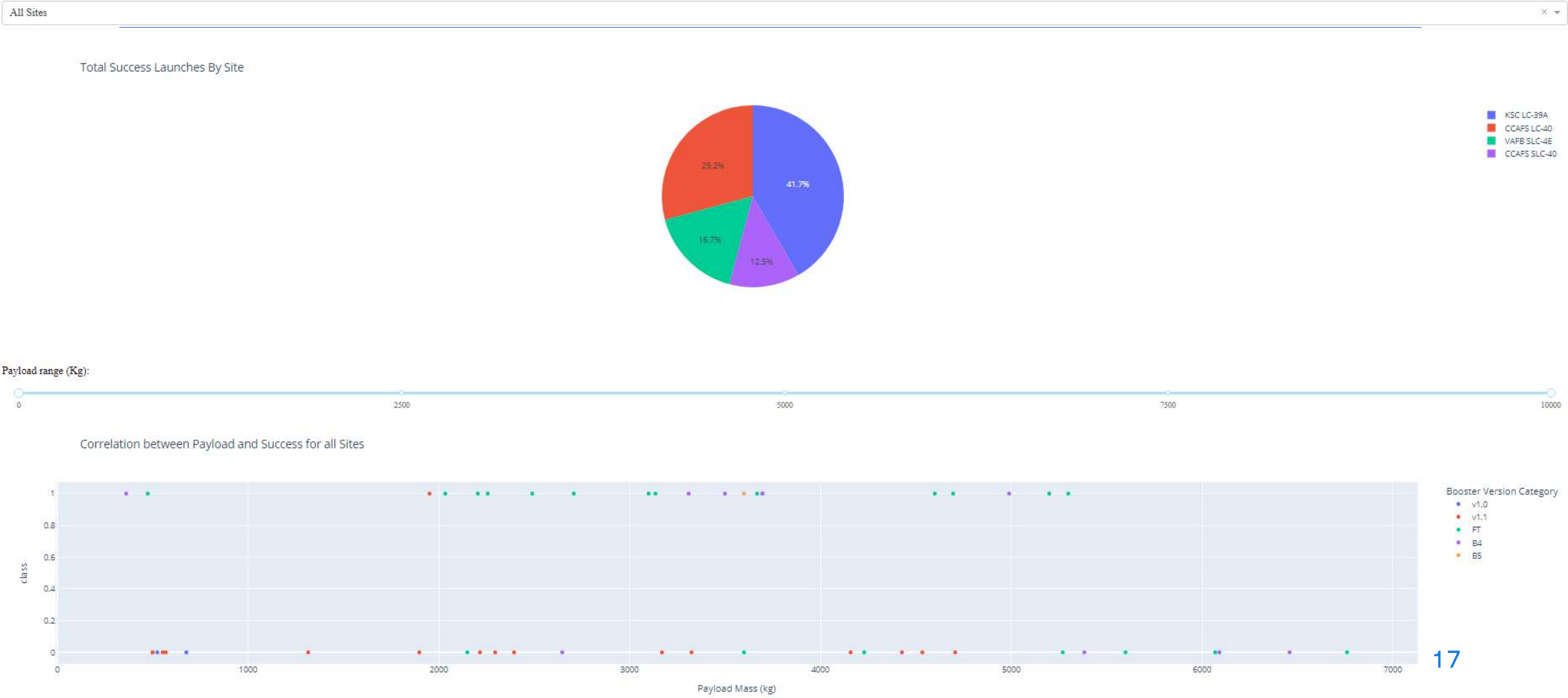
Results: Predictive Analysis

According to the results of the predictive analysis, the Decision Tree algorithm was the best classification technique, which had a 94% accuracy rate,



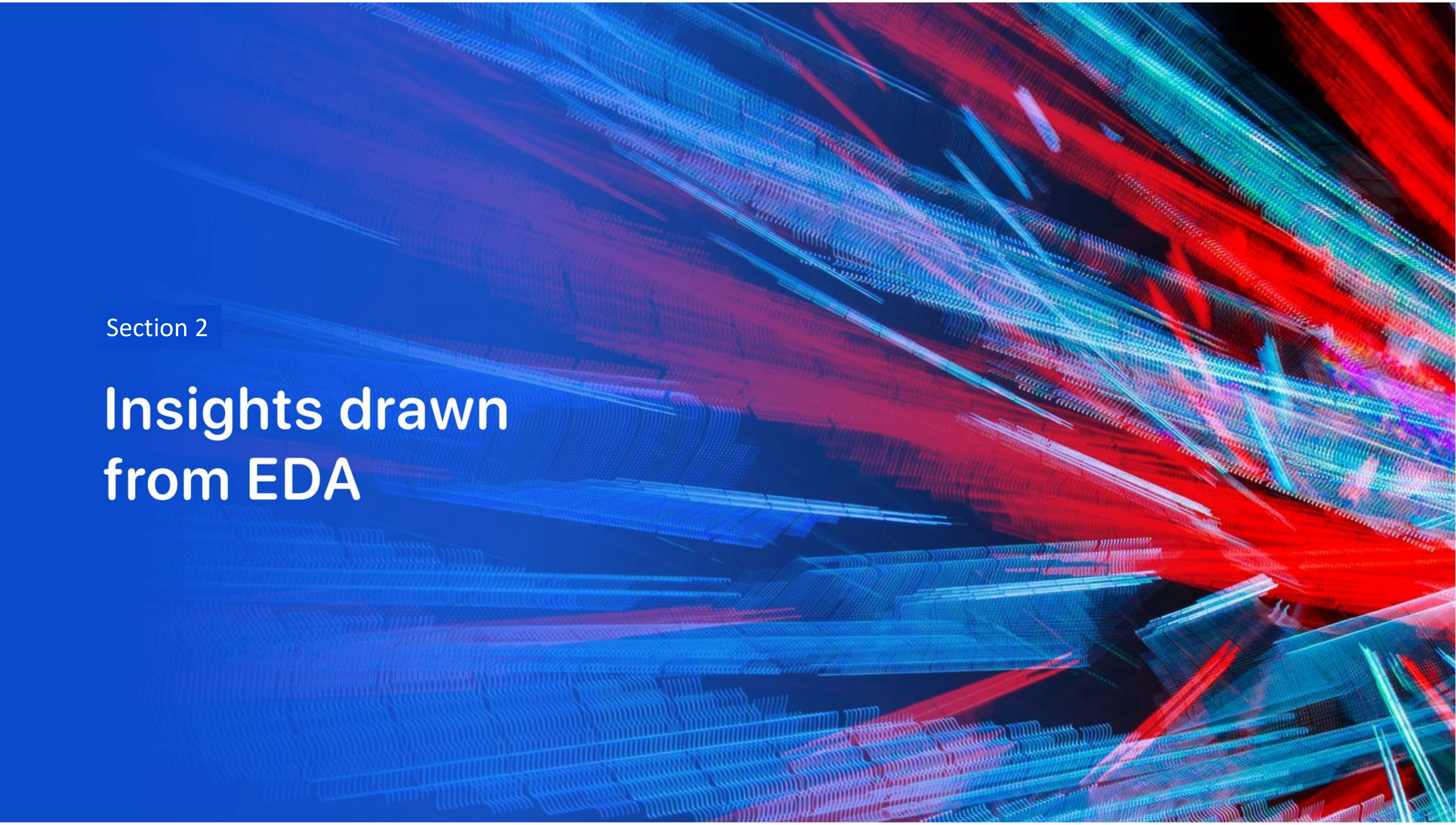
	Accuracy
Logistic	0.833333
SVM	0.833333
Decision Tree	0.944444
KNN	0.833333

Results: SpaceX Launch Records Dashboard



All Sites

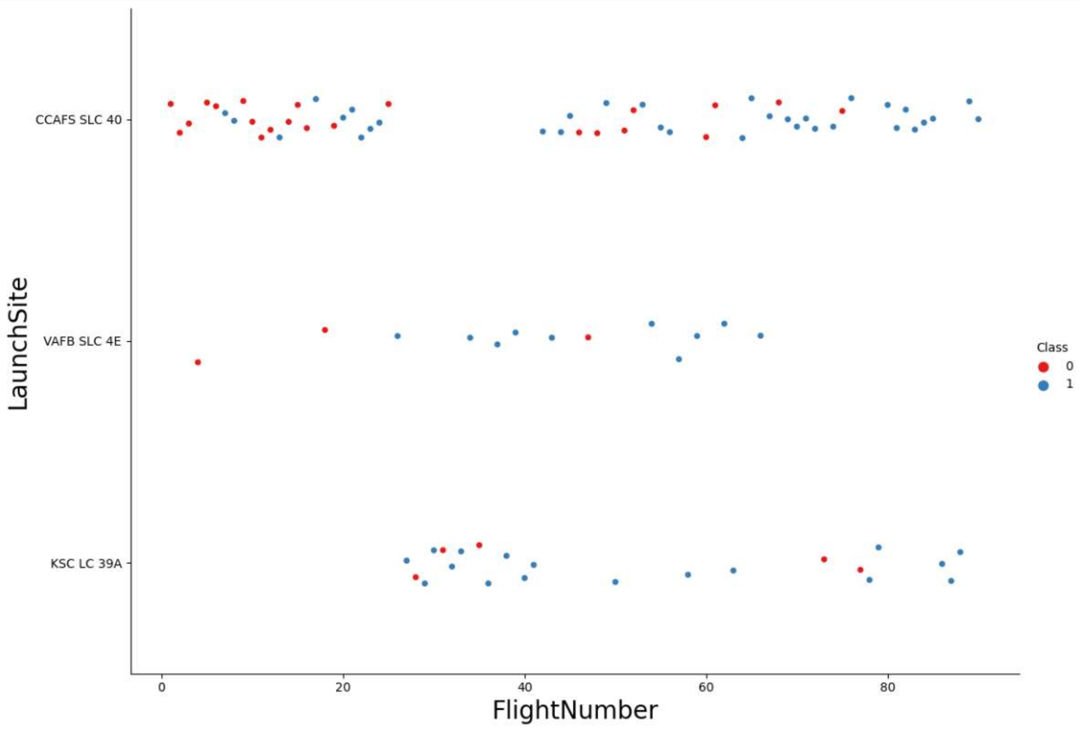
Payload range (Kg):



Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

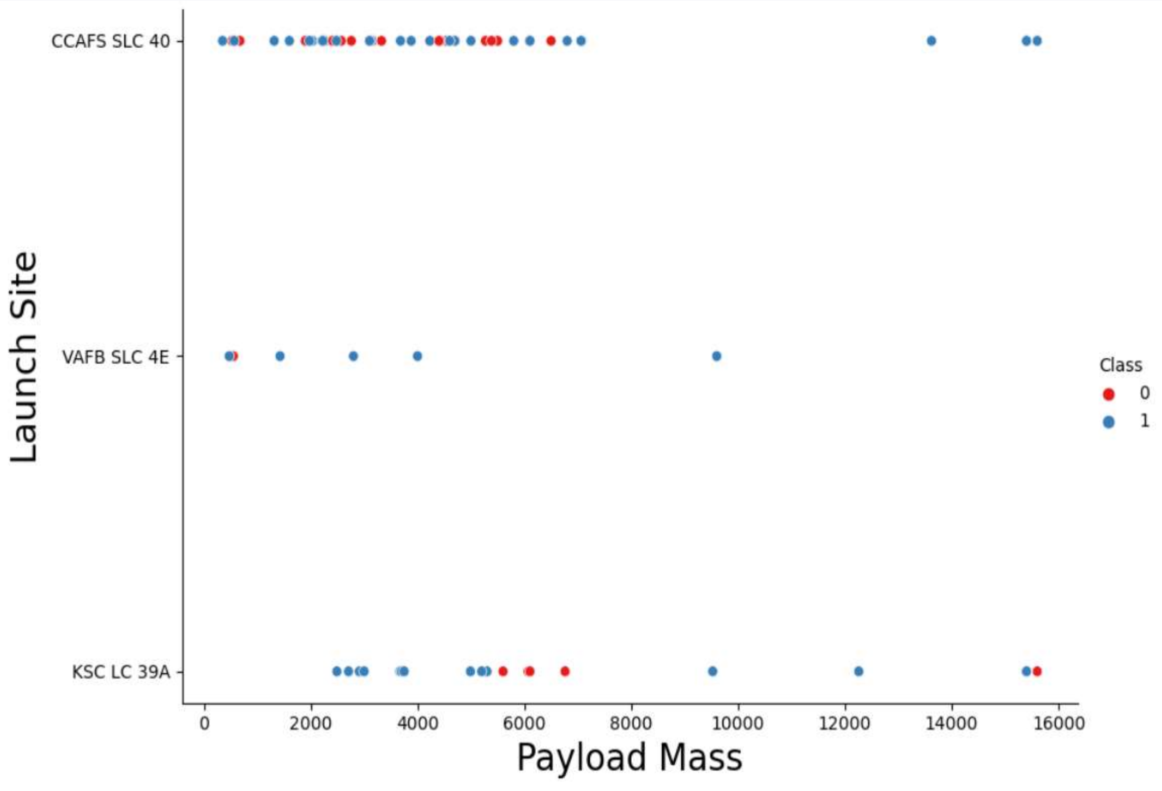


- The successful launches are shown by **blue** dots, whereas the failed launches are represented by **red** dots.

- This graph demonstrates that as the number of flights increased, so did the success rate.

- As the number of flights increases, so does the success rate.

Payload vs. Launch Site

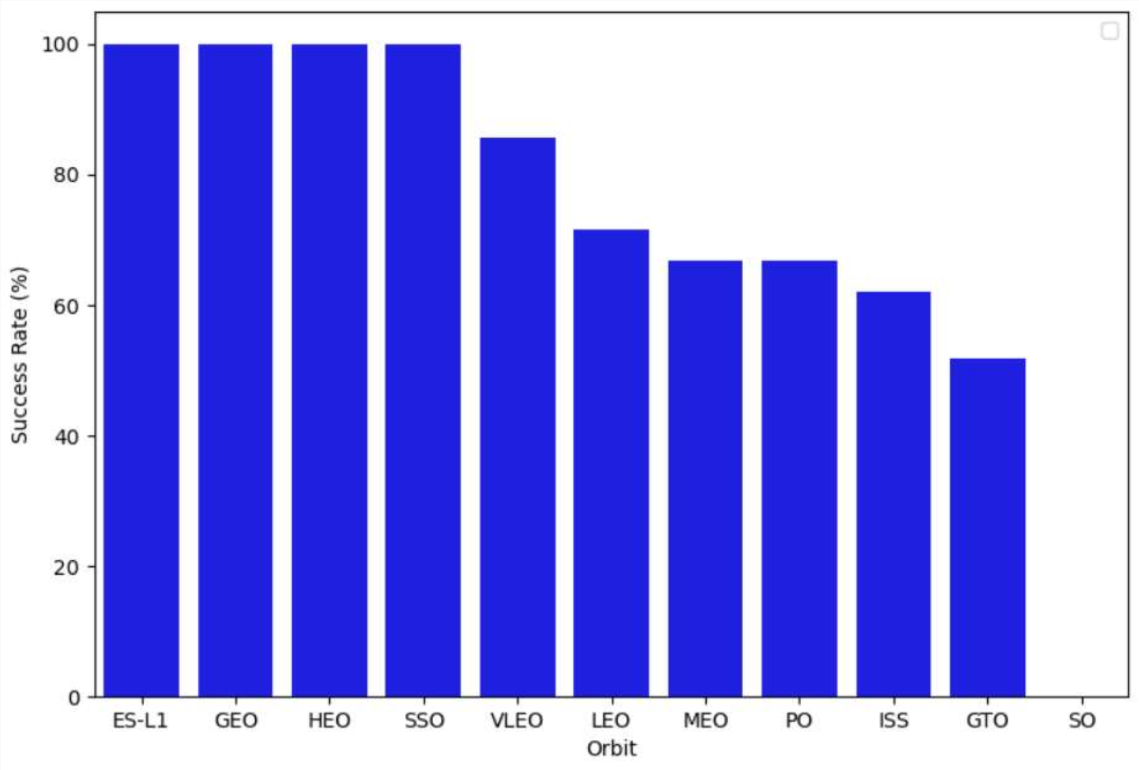


- The successful launches are shown by **blue** dots, whereas the failed launches are represented by **red** dots.

- There are no rockets launched for heavy payload mass from the VAFB-SLC launch site.

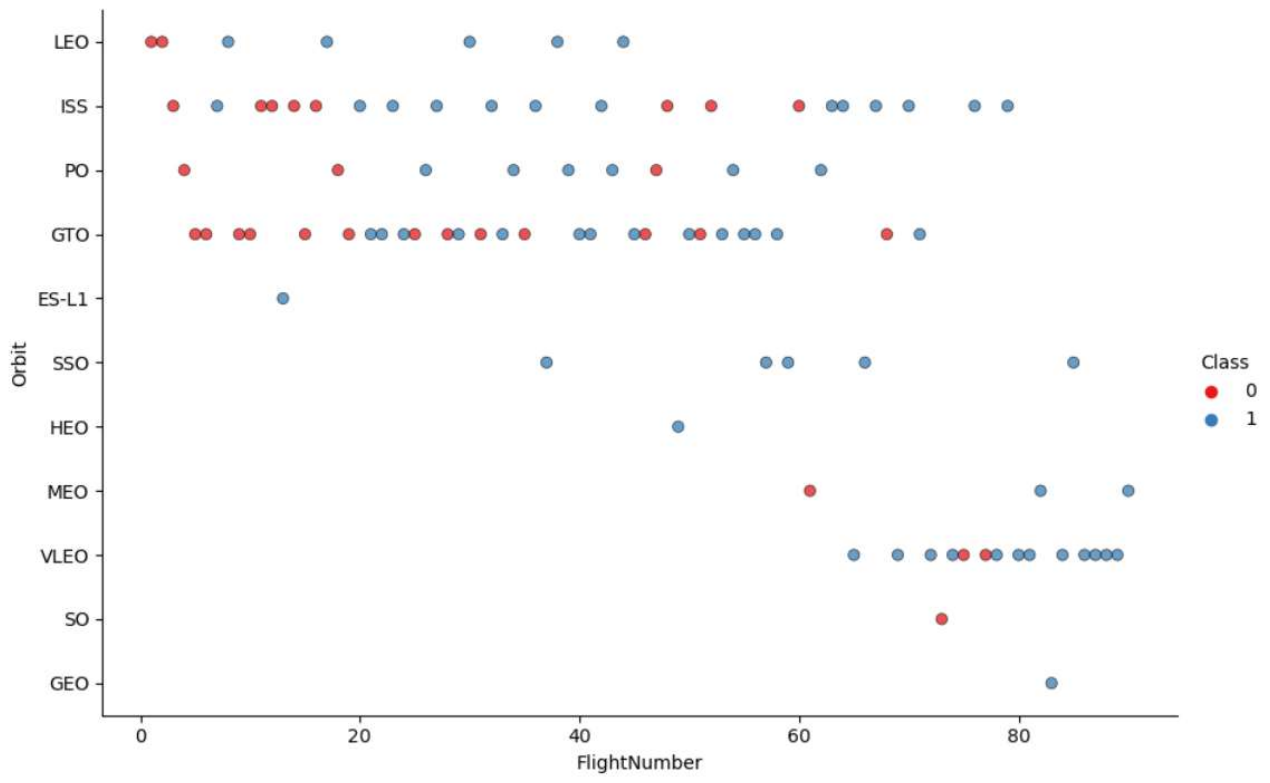
- Decisions cannot be made using this metric because there appears to be a poor association between Payload and Launch Site.

Success Rate vs. Orbit Type



- The launch of SSO, HEO, GEO, and ES-L1 orbits has never failed.
- No rockets from SO orbit were successful.

Flight Number vs. Orbit Type



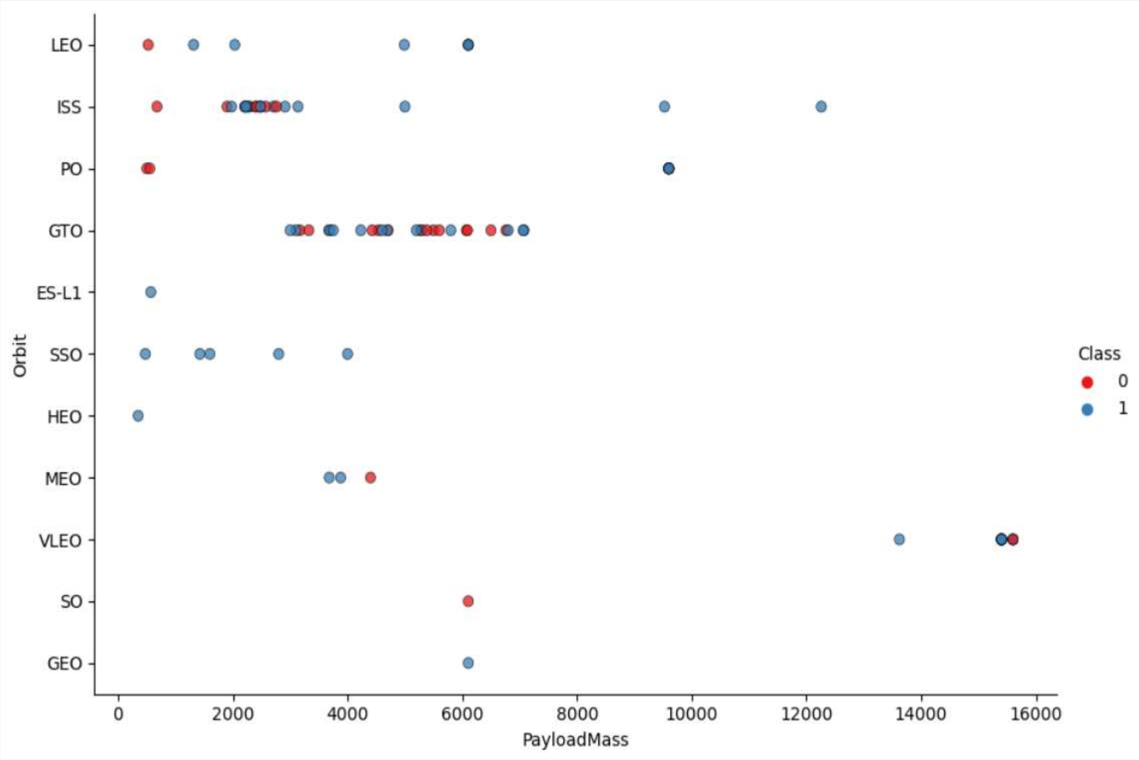
- In the GTO orbit, there doesn't appear to be a correlation between flight numbers.

- In most of the Orbit, the number of flights is positively connected with success.

- Flights with numbers over 40 have a higher success percentage than flights with numbers ranging from 0 to 40.

- Despite having fewer flights than the other orbits, the SSO orbit has a success rate of 100%.

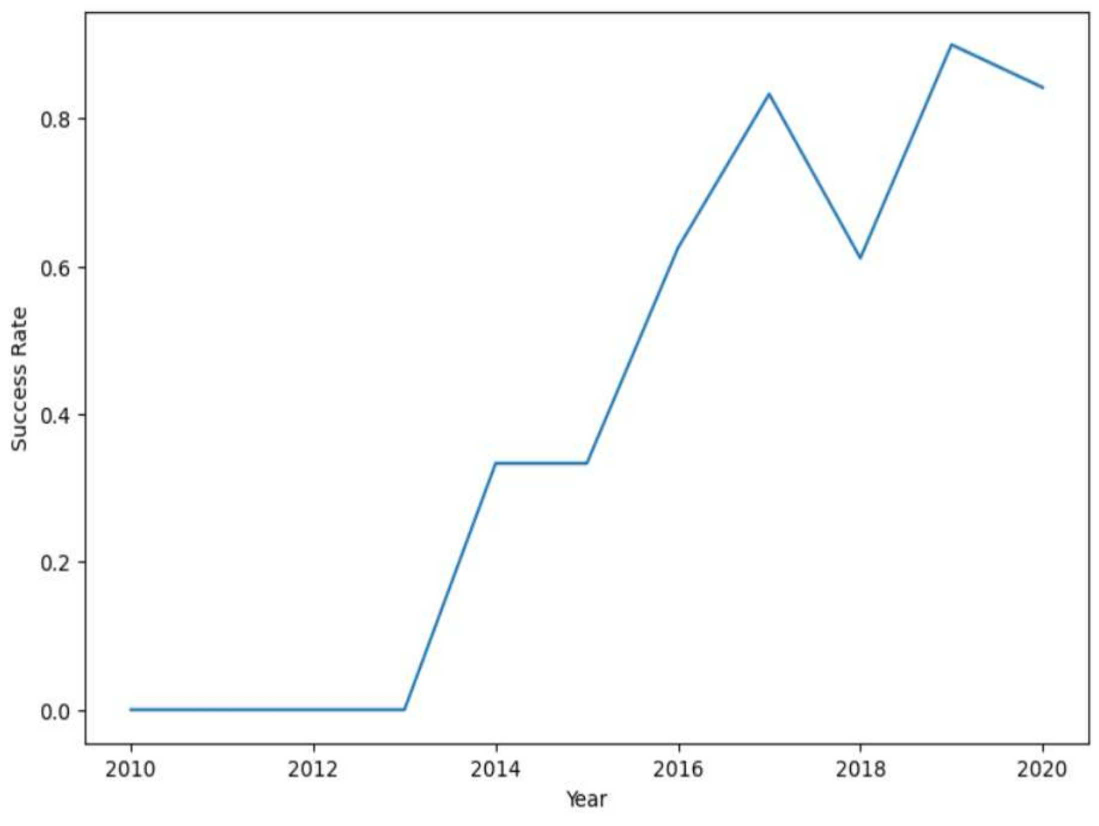
Payload vs. Orbit Type



- Given that both successful and unsuccessful launches are frequently observed, there does not appear to be a direct association between orbit type and payload mass for GTO orbit.

- As payload weight grows, most orbits' success rates increase, however, there is one outlier that failed in VLEO orbit.

Launch Success Yearly Trend

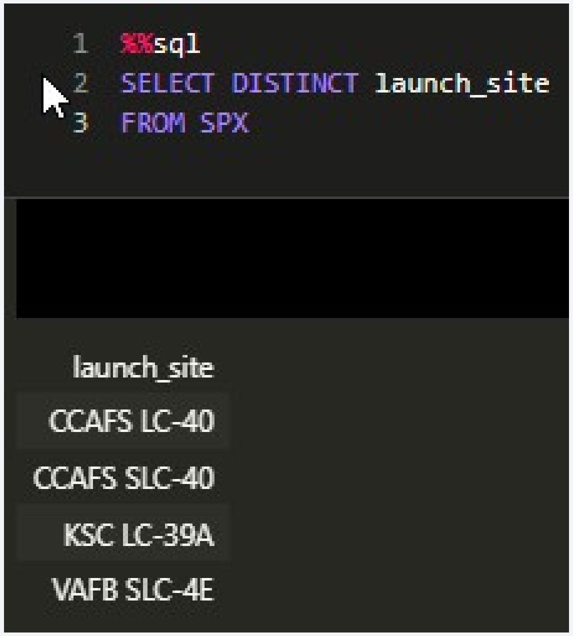


- As the years go by, the chart's overall trend indicates an increase in landing success rate. However, there is a decline in both 2018 and 2020, and more research must be done to determine the reason for the decline.

All Launch Site Names

The DISTINCT clause was used to retrieve the distinct rows from the launch site column. The distinct launch sites are identified as follows:

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



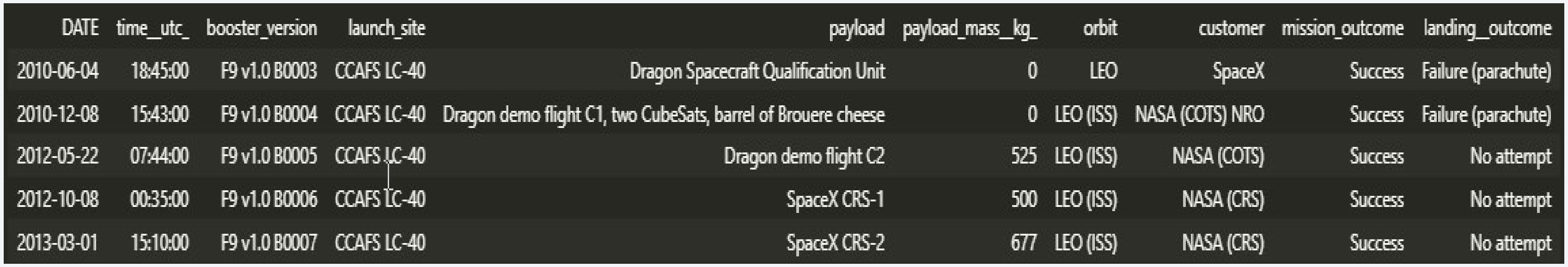
```
%%sql
SELECT DISTINCT launch_site
FROM SPX
```

[redacted]

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

Launch Site Names Begin with 'CCA'

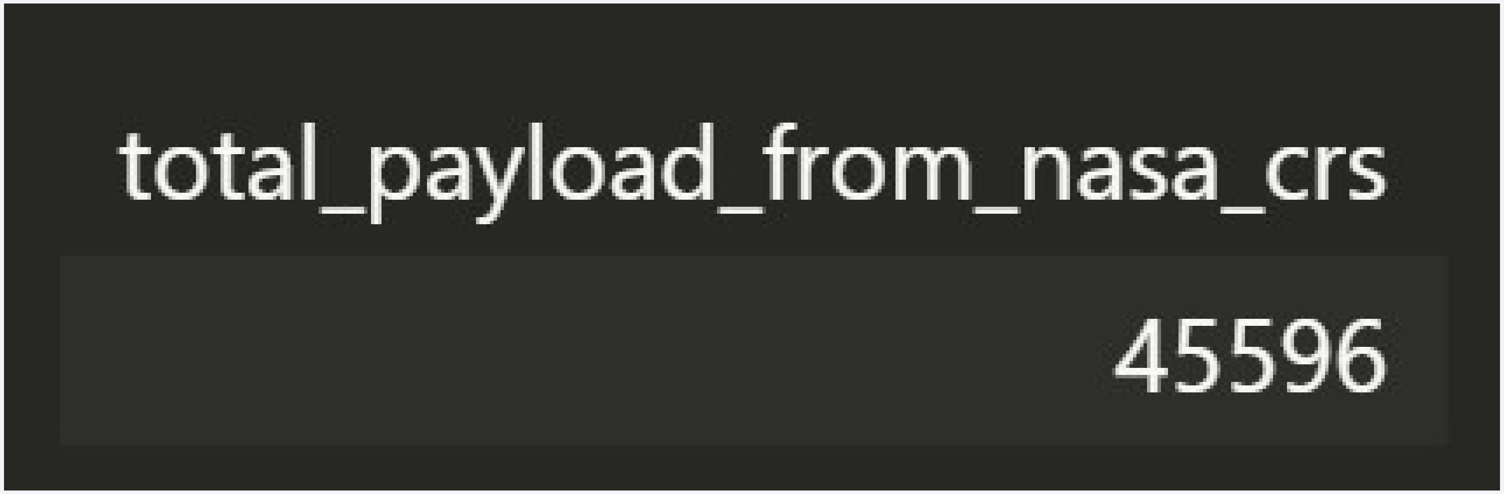
- We select all the columns from the database, then specify where the launch site starts with CCA, then limit the record to 5.



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

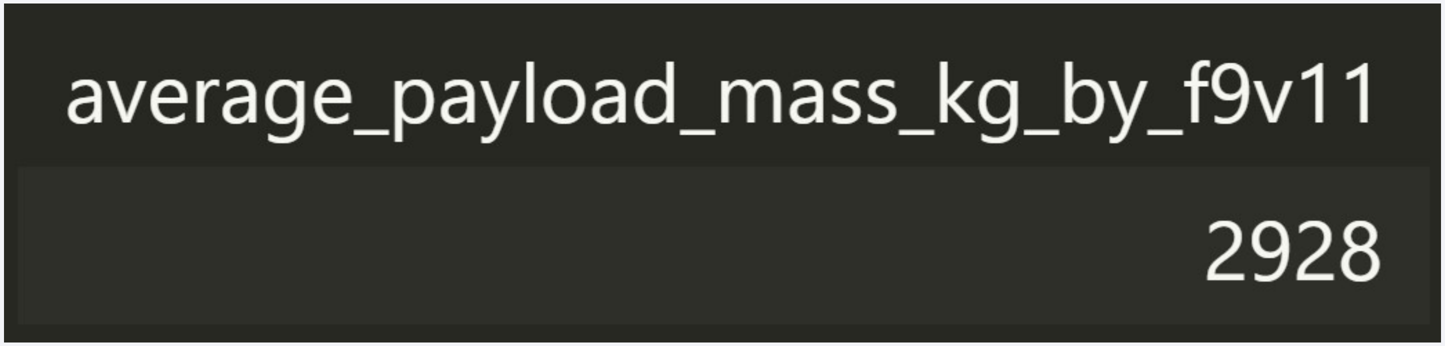
- We use SUM() function to calculate to total payload carried by boosters launched by NASA (CRS)



total_payload_from_nasa_crs
45596

Average Payload Mass by F9 v1.1

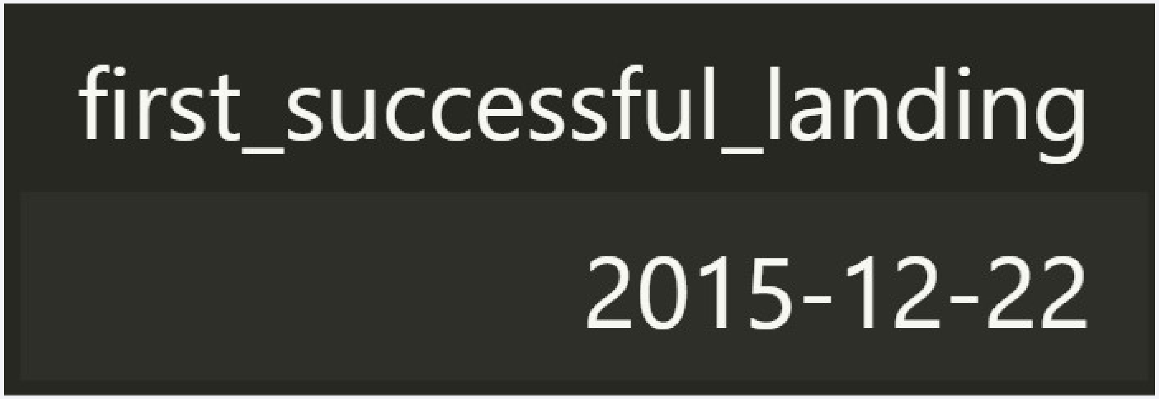
- We use AVG() function to calculate the average payload mass carried by booster version F9 v1.1, which specified by WHERE statement.



average_payload_mass_kg_by_f9v11
2928

First Successful Ground Landing Date

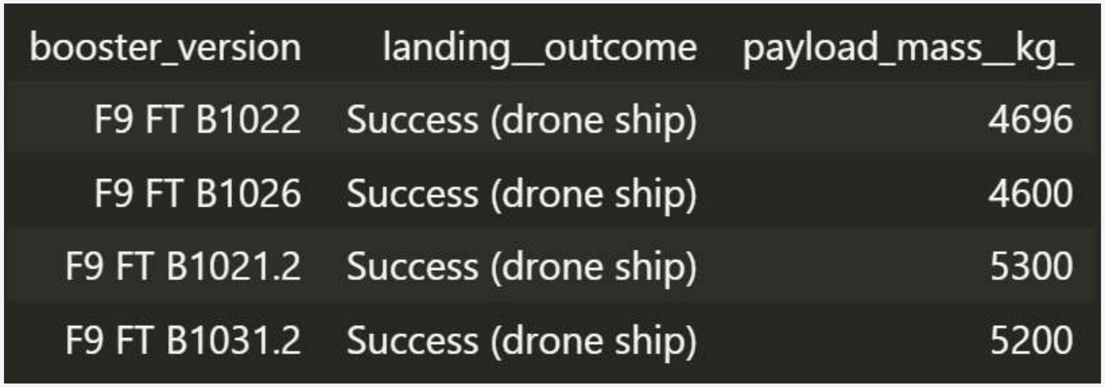
- We use MIN() function to find the first successful ground landing, then we use WHERE statement to filter only the landing which landed on the ground pad.



first_successful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- We select the Successful Drone Ship Landing with a Payload between 4000 and 6000 by using AND operator in WHERE statement:
    - payload_mass__kg_ > 4000
    - landing__outcome = 'Success (drone ship)’
    - payload_mass__kg_ < 6000
- Alternatively, you can use BETWEEN clause combined with WHERE statement.

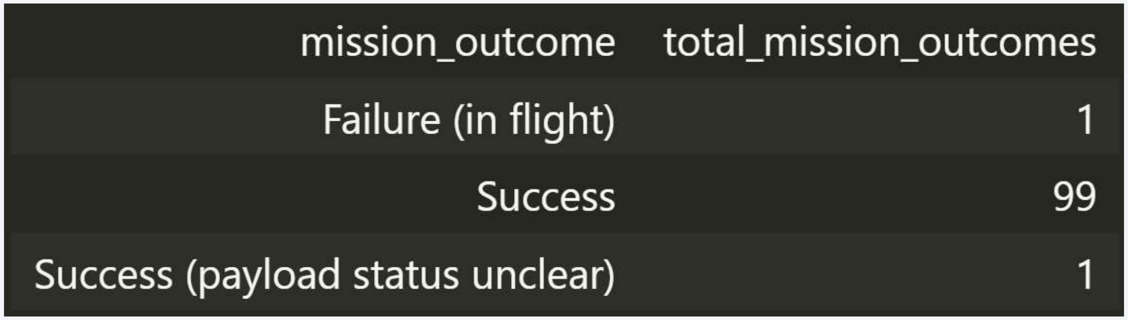


booster_version	landing__outcome	payload_mass__kg_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

Total Number of Successful and Failure Mission Outcomes

- We use COUNT() function to count the number of mission outcomes, and then we grouped them by mission outcome in GROUP BY statement.

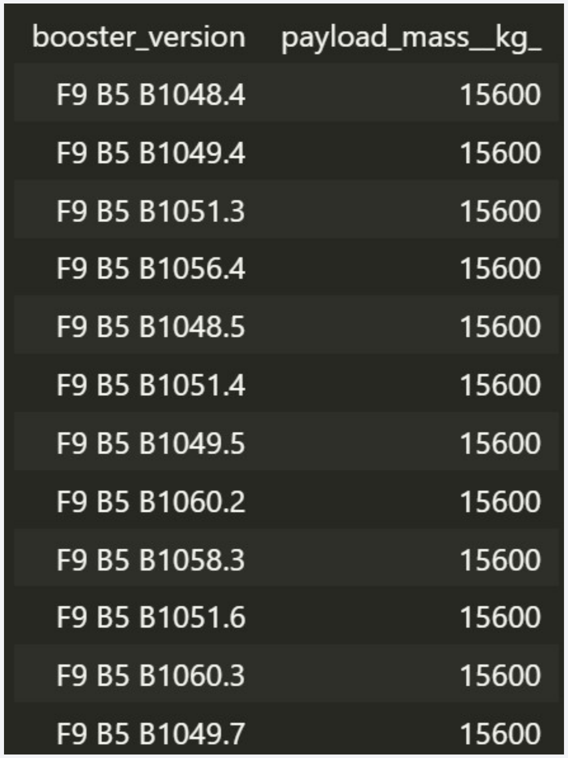
- There are 100 successful mission outcomes, with 1 mission having the status ‘payload status unclear'. Total missions of 101 were conducted.



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

Boosters Carried Maximum Payload

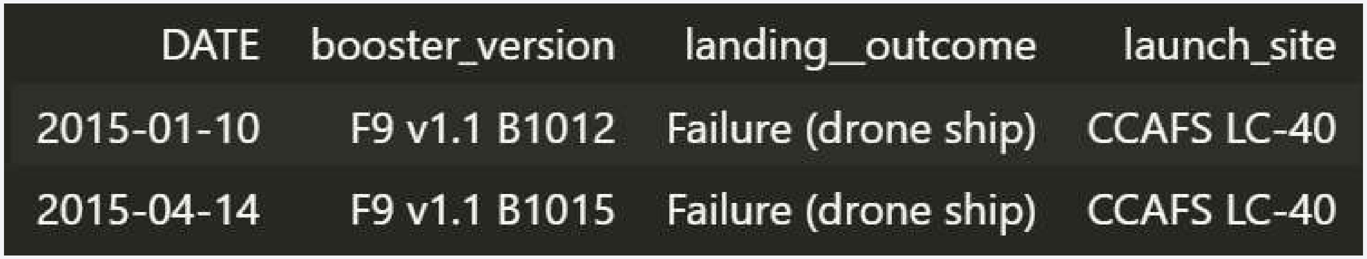
- We use MAX() function to select the maximum payload and then use this maximum payload as a subquery in WHERE statement.



booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

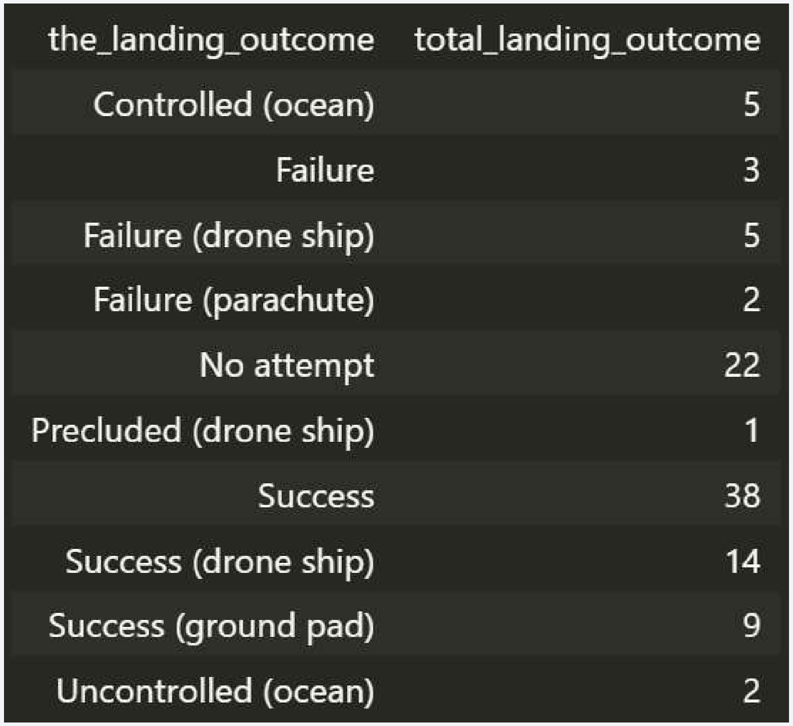
- List the failed landing outcomes in drone ship, their booster versions, and launch site names for the year 2015 was retrieved by specifying landing outcome to failure at drone ship in WHERE statement, then using AND operator the filter the year.



DATE	booster_version	landing__outcome	launch_site
2015-01-10	F9 v1.1 B1012	Failure (drone ship)	CCAFS LC-40
2015-04-14	F9 v1.1 B1015	Failure (drone ship)	CCAFS LC-40

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

- In the SELECT statement, we select the landing outcome and also use the COUNT() function to count the landing outcome, then in WHERE clause we specific the range of dates we want using OR operator, finally in GROUP BY statement we group by landing outcome.

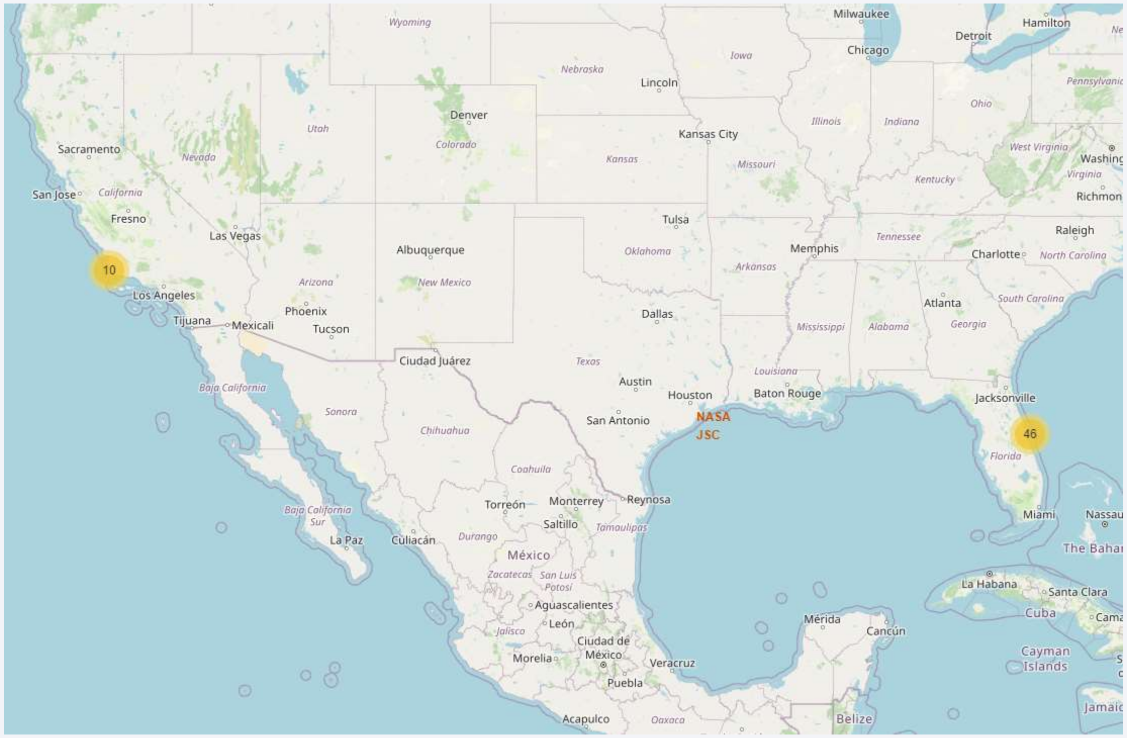


the_landing_outcome	total_landing_outcome
Controlled (ocean)	5
Failure	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	22
Precluded (drone ship)	1
Success	38
Success (drone ship)	14
Success (ground pad)	9
Uncontrolled (ocean)	2

Section 3

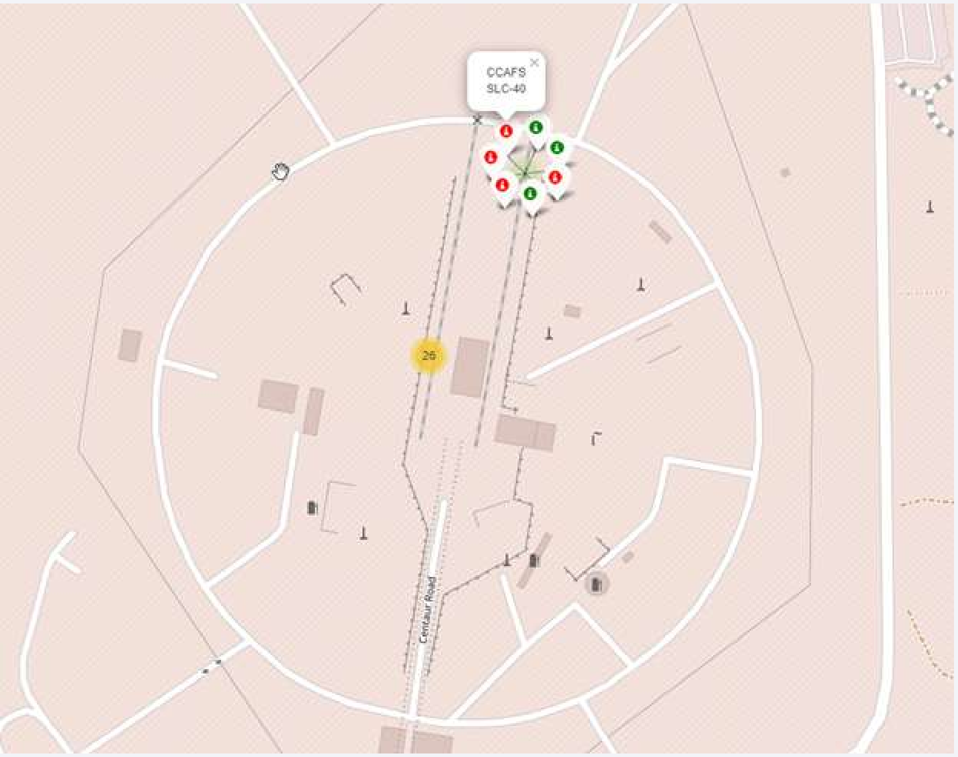
Launch Sites Proximities Analysis

Folium Map: Launch Site Locations



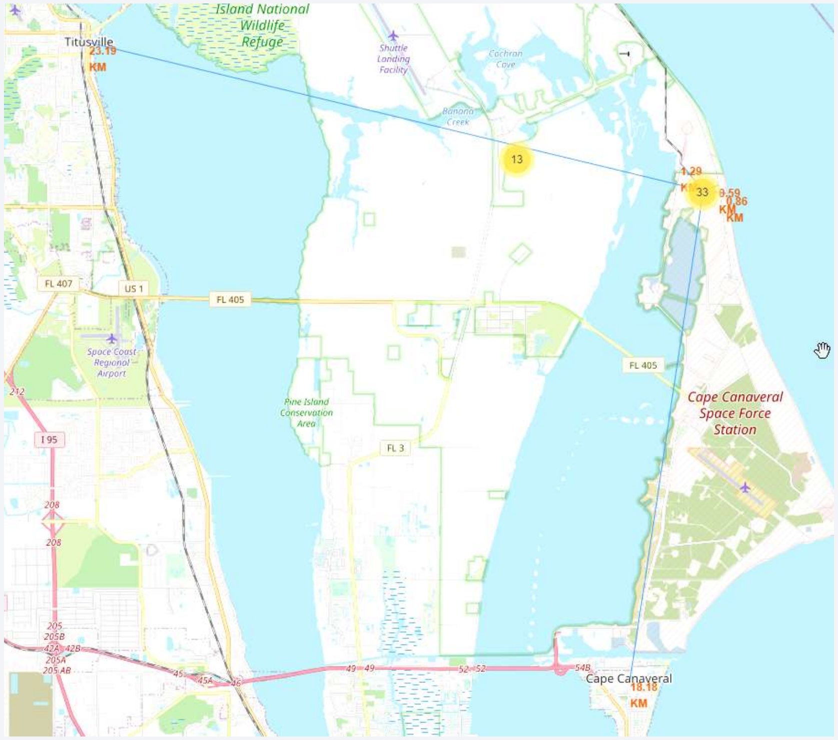
- The locations of all the SpaceX launch sites in the US are indicated by the yellow markers.
- The launch pads have been situated in a suitable location close to the coast.

Folium Map: Color-Labeled Launch Outcomes



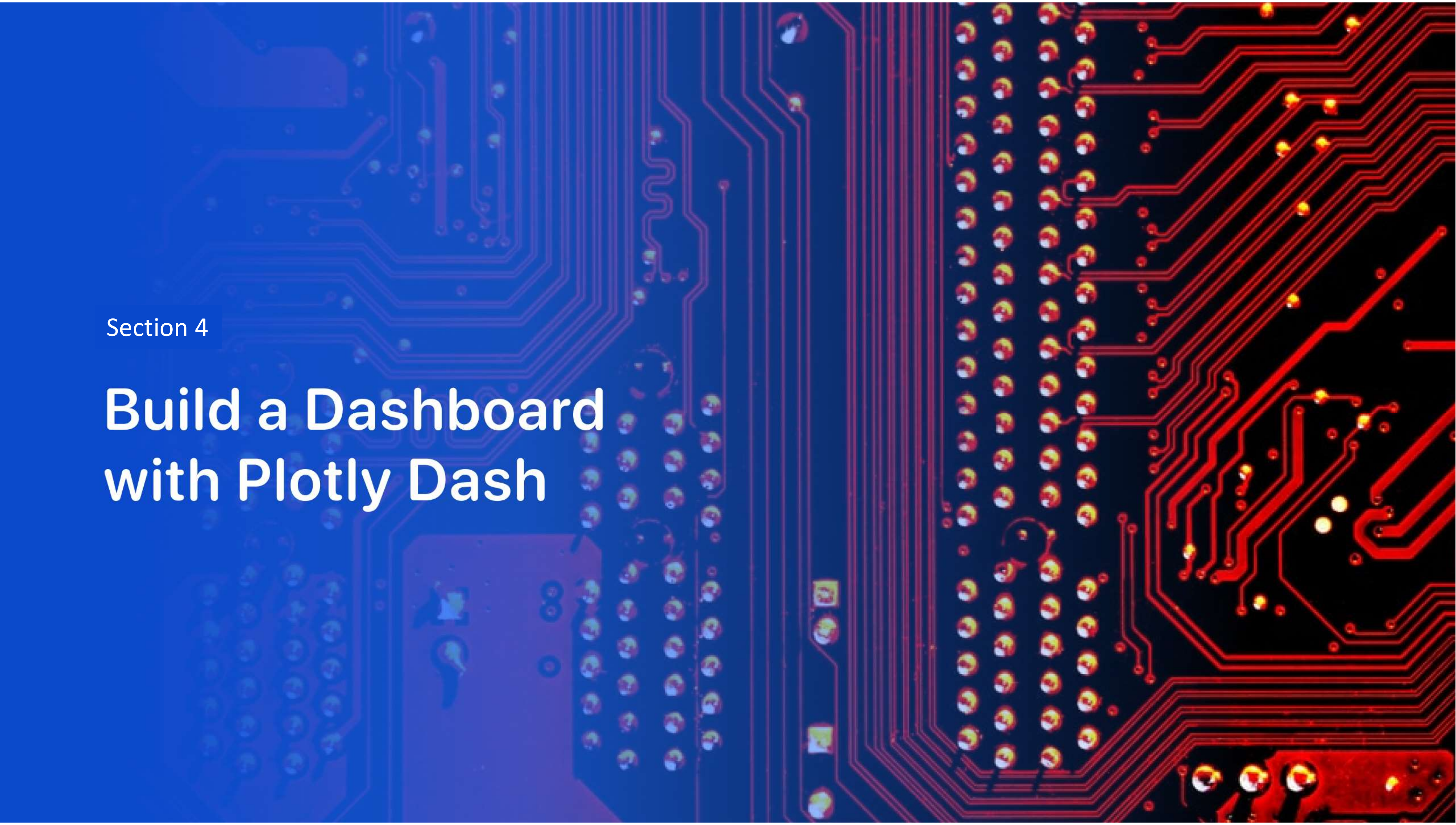
- The launch site will show marker clusters of successful landings (green) or failed landings as we zoom in on a launch location (red).

Folium Map: Proximities



The shown in the Folium map we can conclude that the launch site was proximities to the railroad, highway, coastline, and cities.

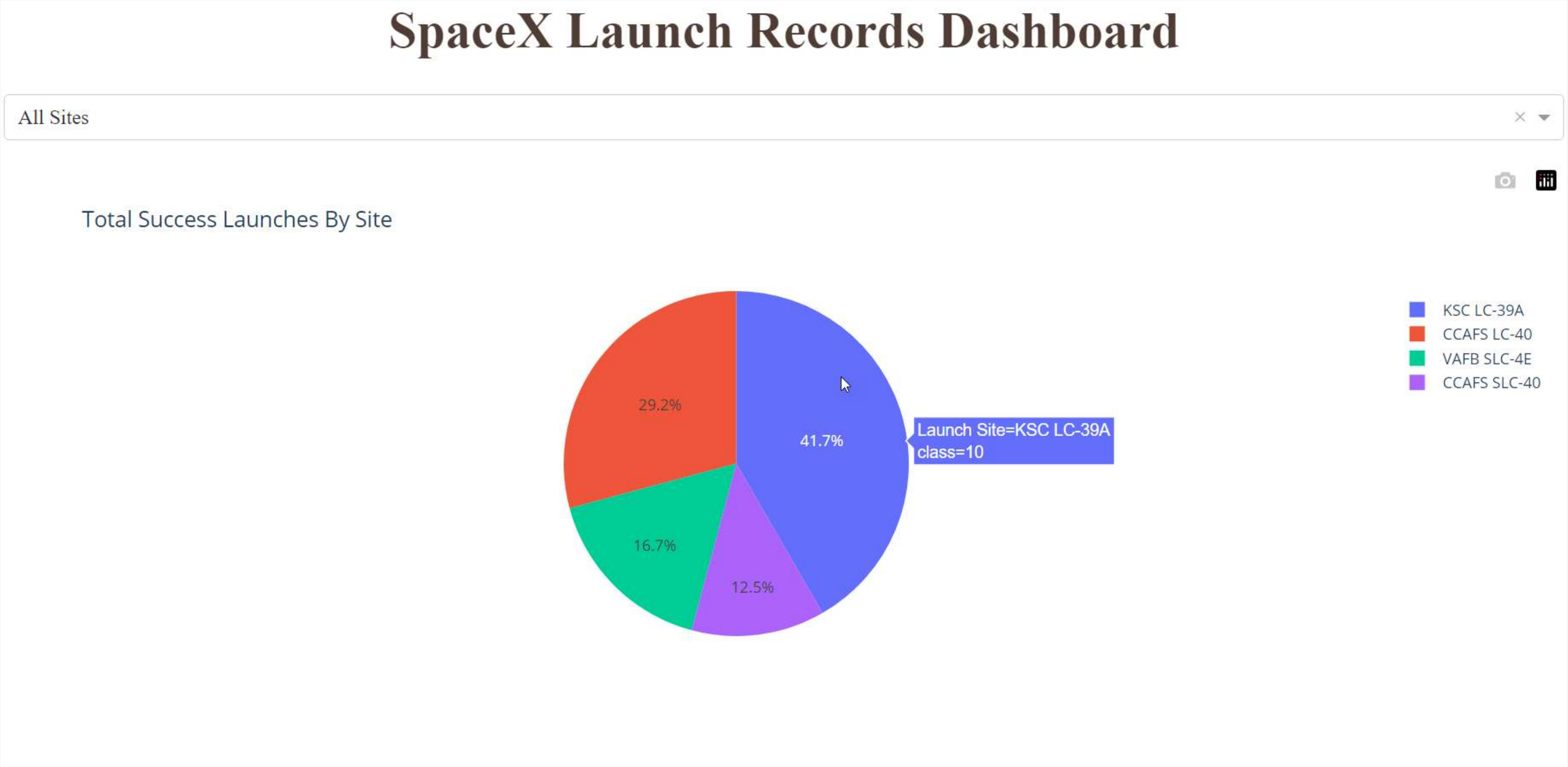
- 1.29 from Railroad
- 0.59 km from Highway
- 0.86 km from the coastline
- 18.18 km from Cape Canaveral City
- 23.19 km from Titusville City



Section 4

Build a Dashboard with Plotly Dash

Interactive Dashboard: Pie Chart



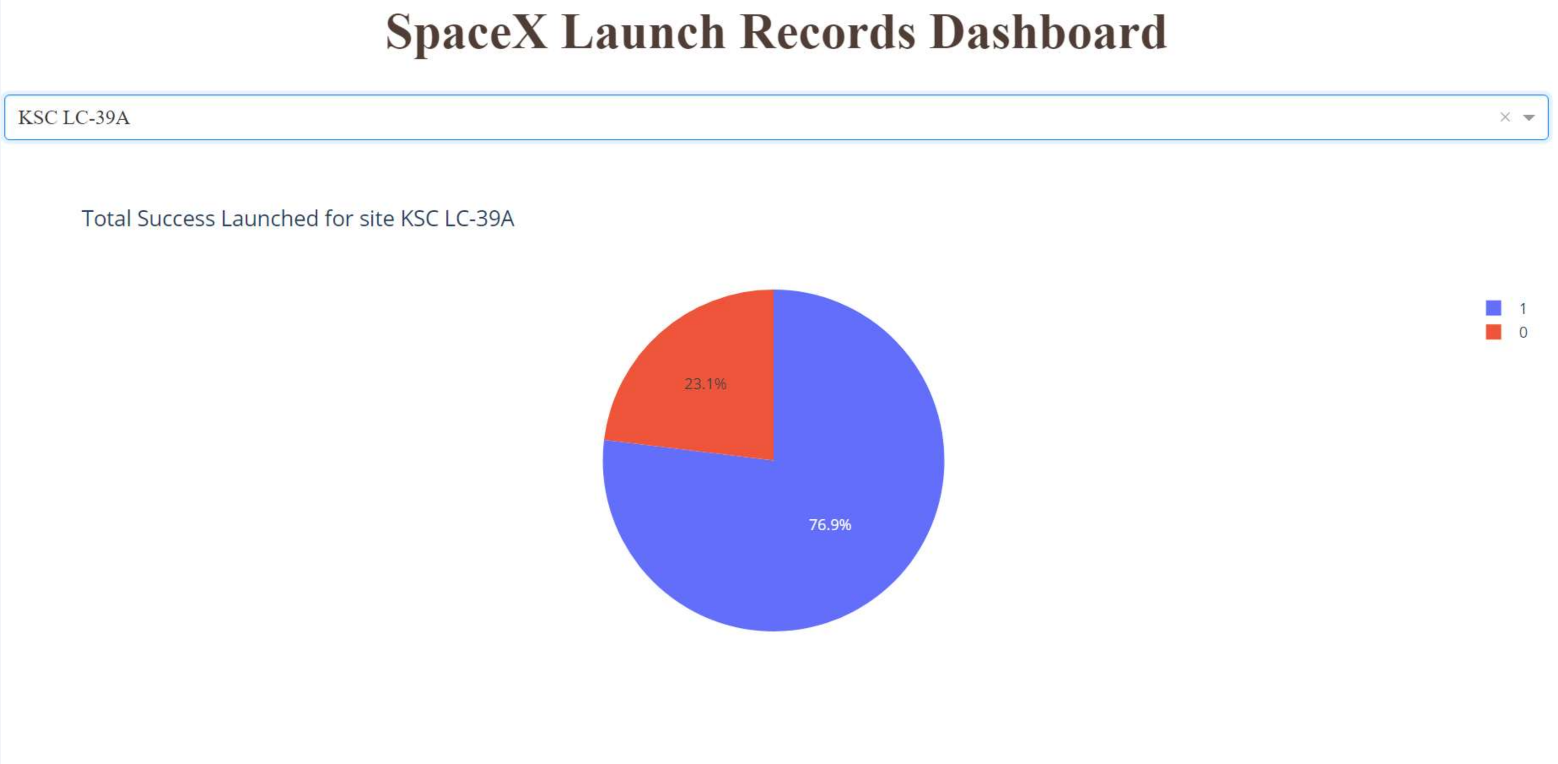
SpaceX Launch Records Dashboard

All Sites

Total Success Launches By Site

The graph show launch mission between launch sites, you can select the launch site to see the success rate and failure rate.

Interactive Dashboard: Pie Chart (KSLC-39A)



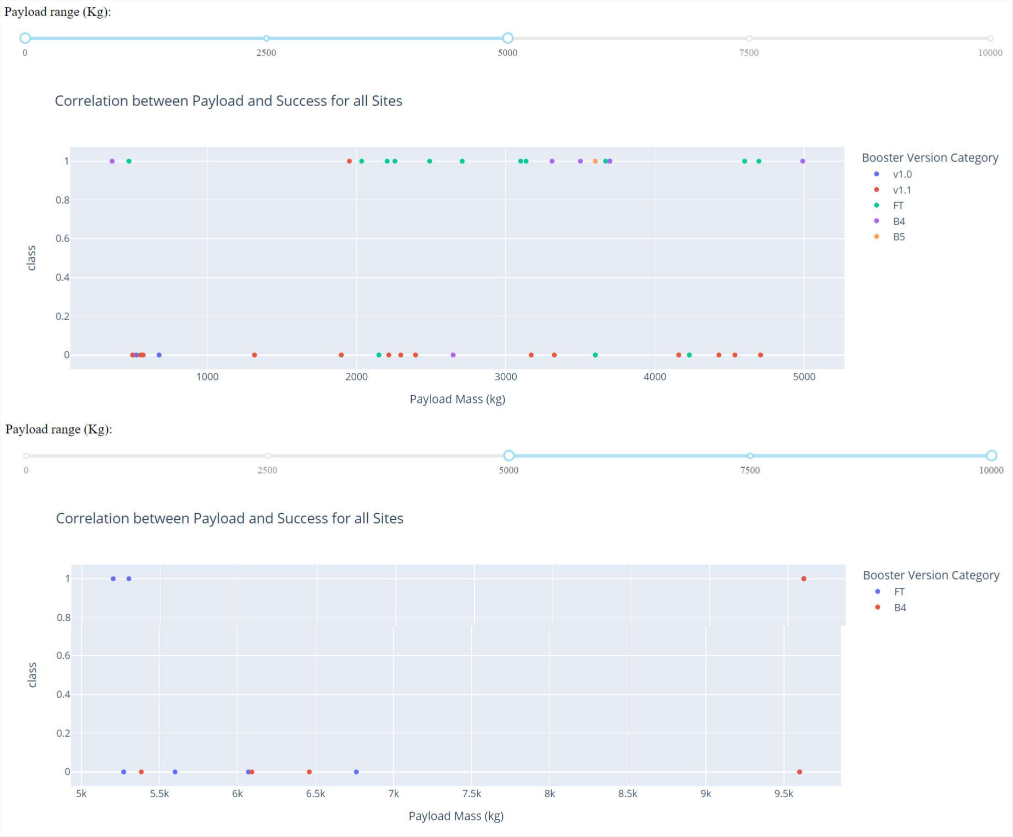
SpaceX Launch Records Dashboard

KSC LC-39A

Total Success Launched for site KSC LC-39A

The graph shows that KSLC-39A site has the highest success rate of 76.9%

Interactive Dashboard: Payloads vs Launch Outcome



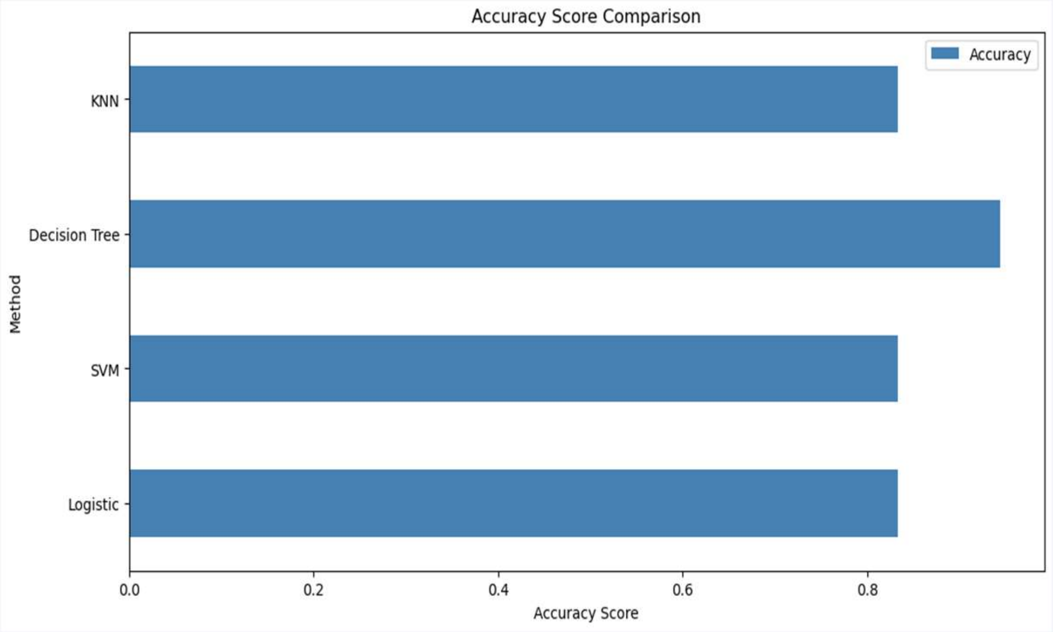
- The graph shows the success rate with various payload mass (kg), range from 0 to 5000kg, and from 5000 to 10000kg.

- As the payload mass increased, the success rate decreased

Section 5

Predictive Analysis (Classification)

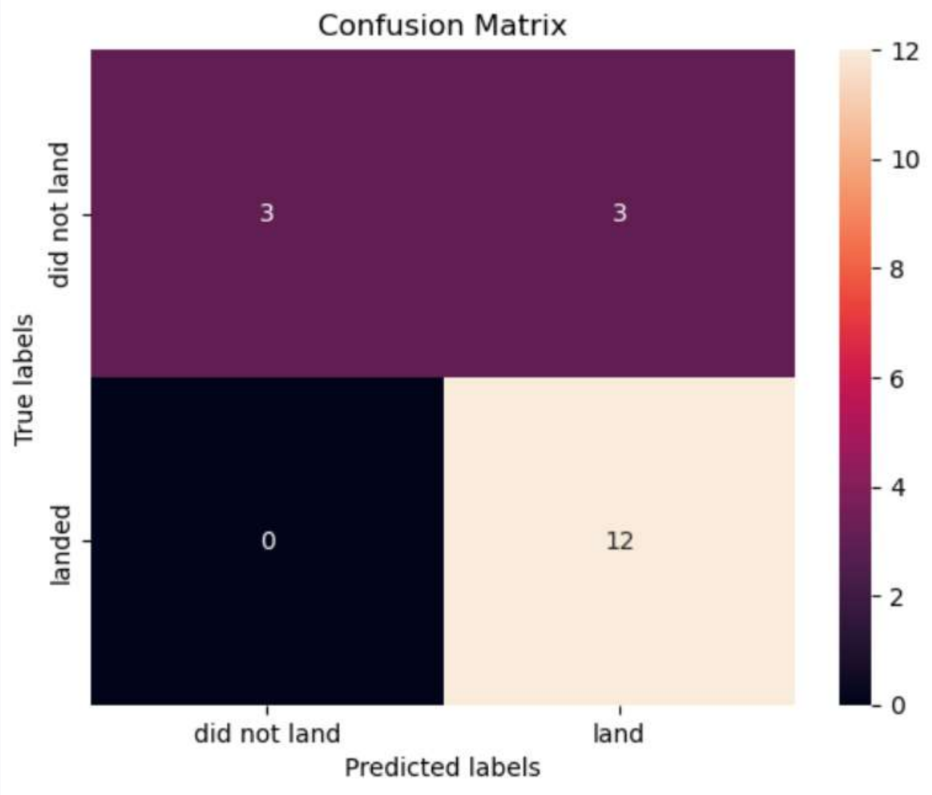
Classification Accuracy



- Accuracy Score for each method are as follows:
  - Logistic regression = 83%
  - SVM = 83%
  - Decision Tree = 94%
  - KNN = 83%

- The Decision Tree Model has the highest accuracy score of 94%, while the other model accuracy score is 83%

Confusion Matrix



- When the True label was a success (True Positive), the model predicted 12 successful landings

- When it was a failure, it predicted 3 unsuccessful landings (True Negative).

- When the True label was an unsuccessful landing, the algorithm also predicted 3 successful landings (False Positive).

- Successful landings were frequently predicted by the model.

Conclusions

- The Decision Tree Classifier is the most accurate predictive model for this dataset, having a 94% accuracy rate.

- The analysis revealed that as the success rate has increased over time, there is a positive correlation between the number of flights and the success rate.

- The launch sites are placed at a safe distance from cities but strategically close to roads and railroads for the transit of people and goods.

- Payload mass can be related to success rate since lighter payloads have typically had better results than bigger payloads.

- The SSO orbit has a success rate of 100%.

Appendix

- GitHub Repository: https://github.com/pkong001/Primary-Repositoty/tree/main/Temp/For-Capstone-Project

Thank you!