IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Kay Ng, Siu Kit>
<Jan 20, 2024>

Outline

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

Executive Summary

- Data was requested from SpaceX API requests and web scrapping from SpaceX Wikipedia.
- After gathering data was collected and being preprocessed, exploratory data analysis (EDA) and interactive visual analytics were conduct using different data ana analytic tools.
- By analyzing different aspects such as orbit types and launch sites. We found that Orbit SSO had 100% success landing rate and Launch Site KSC LC-39A has the most success landing rate among all launching sites of 76.9%.
- Among the four classification methods, Decision Tree has the highest accuracy score reaching 90% which can be used to predict whether the next launch can be success.

Introduction

- Nowadays, with increasing human population, space exploration becomes a hot topic again.
- SpaceX is a US company focuses in designs, manufactures and launches rockets and spacecraft. It was founded in 2002 and is now one of the leading space technology company in the world.
- One famous produce of SpaceX is the reusable rocket Falcon 9, which costs only $62 million in comparing to rivals $165 million in each launch.
- The huge cost-saving is due to the reuse of rocket in the first stage, therefore, if we can predict whether the first stage landing will be successful, we can briefly predict the cost of a launch.
- This report is focusing on determining the factors which affect Falcon 9's first stage landing and predicting the landing outcomes.

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

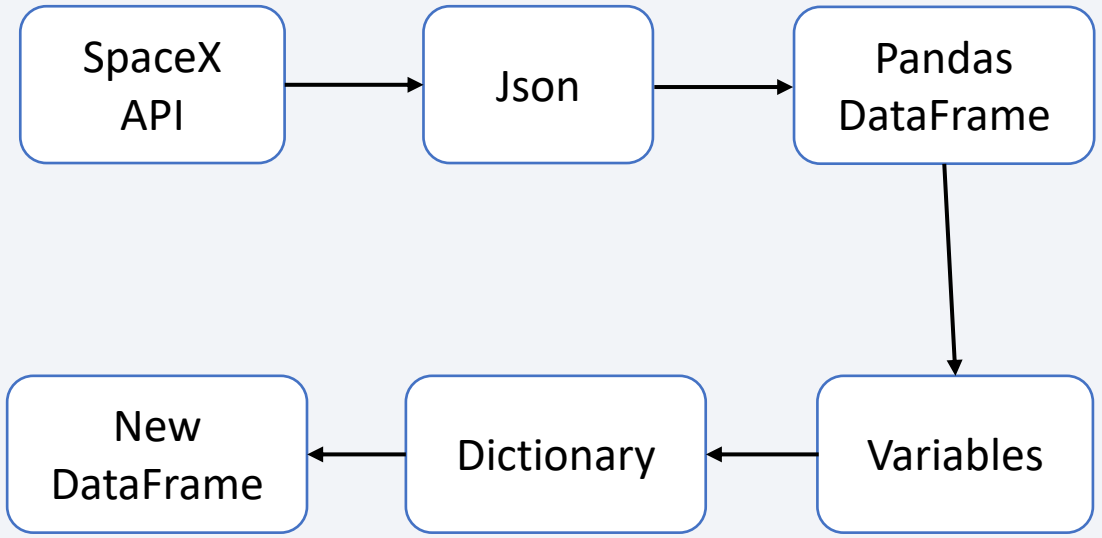
Data Collection

In order to conduct analysis and construct prediction model, we collect relevant data from

- requesting the historical rocket launch data from SpaceX API and
- preforming web scraping to extract Falcon 9 launch records from Wikipedia.

Data Collection – SpaceX API

1. Request the rocket launch data through SpaceX API.
2. The response content was being decoded as a Json and turned into a Pandas data frame.
3. From the data frame we extract several meaningful columns
4. Create variables and combine the columns into a dictionary
5. Create a new data frame from dictionary.



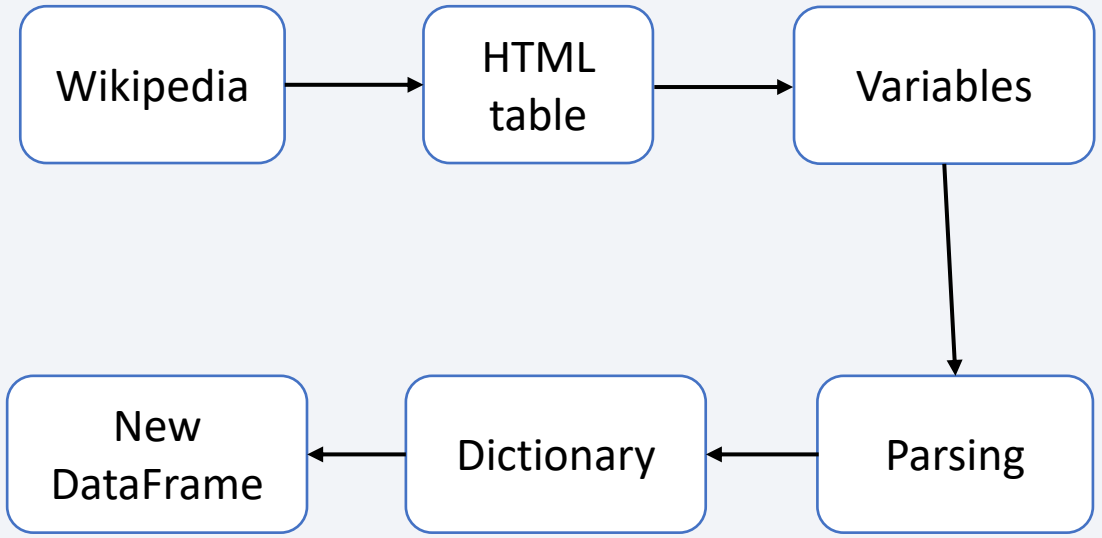
- Please kindly refers to the notebooks below:
- https://github.com/ngsiukit/Applied-Data-Science-Capstone---SpaceX/blob/main/Week1%20SpaceX%20API.ipynb

Data Collection – Web Scraping

Web scrap Falcon 9 launch records with BeautifulSoup:

1. Use HTTP GET method to request the Falcon9 Launch HTML table from Wikipedia.
2. Extract all column/variable names from the HTML table.
3. Parsing the HTML table.
4. Fill in the parsed launch record values into a dictionary.
5. Create a new Pandas data frame from dictionary.

- Please kindly refers to the notebooks below:
- https://github.com/ngsiukit/Applied-Data-Science-Capstone---SpaceX/blob/main/Week1%20Web%20Scrapping.ipynb



Data Wrangling

1. Filter the data frame to only include Falcon 9 launches
2. Identify and exclude the missing values in each attribute
3. Identify which columns are numerical and categorical
4. Create a landing outcome label of 0 and 1
5. Convert the categories in outcome column to landing outcome label

- Please kindly refers to the notebooks below:
- https://github.com/ngsiukit/Applied-Data-Science-Capstone---SpaceX/blob/main/Week1%20Data%20wrangling.ipynb

EDA with Data Visualization

- Scatter charts were plotted to find whether there were correlation between different aspects of the historical launches.
- Bar chart was plotted to show the relation between orbit types and success rate.
- Line chart was plotted to show the trend of overall success rate.

- Please kindly refers to the notebooks below:
- https://github.com/ngsiukit/Applied-Data-Science-Capstone---SpaceX/blob/main/Week2%20EDA%20Visualization.ipynb

EDA with SQL

- Several EDA were conducted, including
  - Launch site information
  - Payload mass information
  - The first successful landing date
  - Successful drone ship landing with Payload between 4000 and 6000
  - Figures of overall mission outcomes
  - Booster information
  - Landing outcomes within specific period

- Please kindly refers to the notebooks below:
- https://github.com/ngsiukit/Applied-Data-Science-Capstone---SpaceX/blob/main/Week2%20sqllite%20EDA.ipynb

Build an Interactive Map with Folium

- In the interactive maps using Folium, circles and markers were created to show the locations of launch site as well as the launch outcome information.
- Lines and distance were created to show the location preference of launching sites.

- Please kindly refers to the notebooks below:
- https://github.com/ngsiukit/Applied-Data-Science-Capstone---SpaceX/blob/main/Week3%20Folium.ipynb

Build a Dashboard with Plotly Dash

- In the Dashboard using Plotly Dash, a pie chart was created to show the launch outcomes of different launch sites.
- A slider was created to show the launch outcomes for different payload range in different launch sites.
- A scatter plot chart was created to show the relationship of payload and success rate for different boosters.

- Please kindly refers to the notebooks below:
- https://github.com/ngsiukit/Applied-Data-Science-Capstone---SpaceX/blob/main/Week4%20Plotly%20Dash

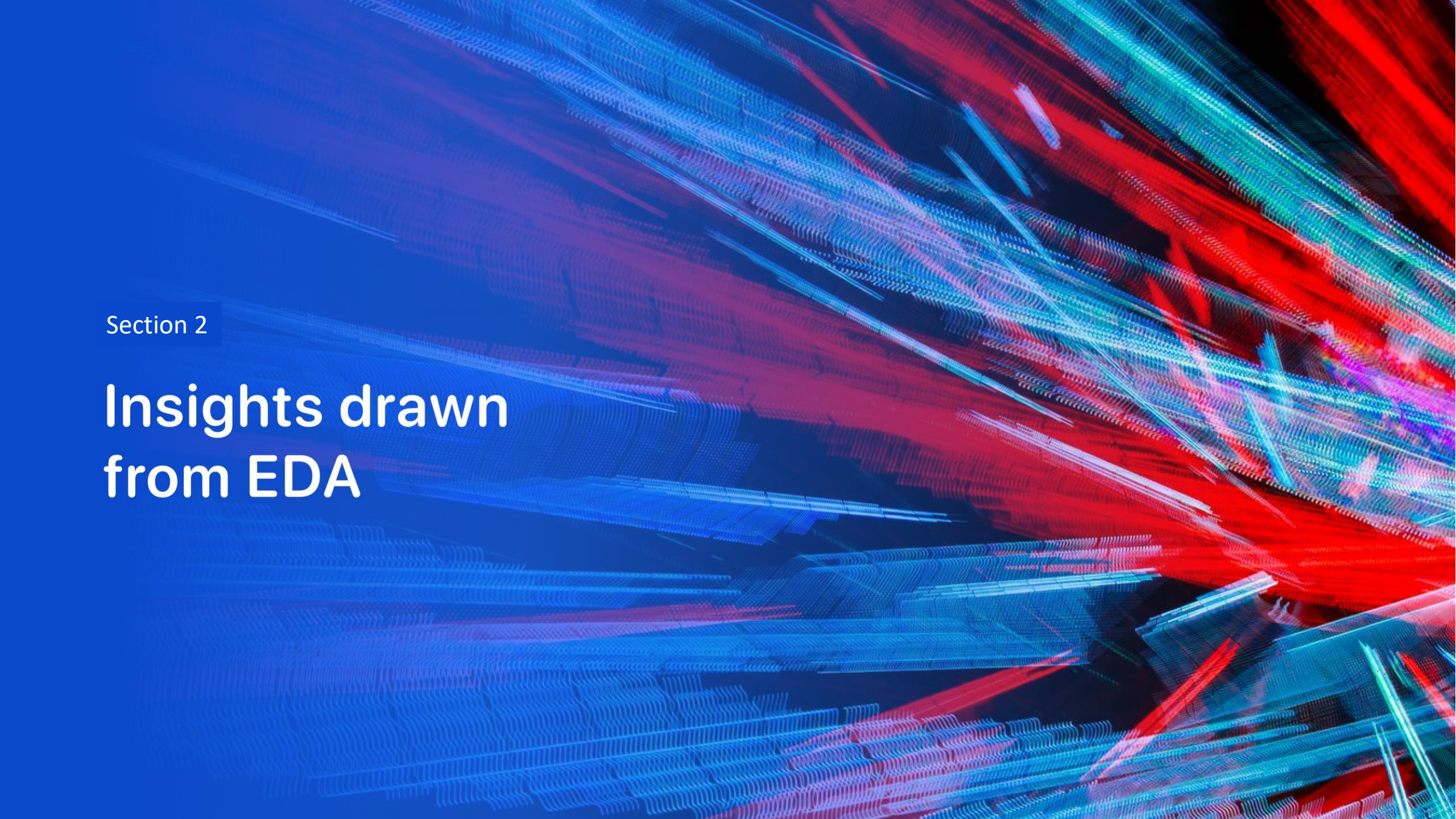
Predictive Analysis (Classification)

- Four classification models were performed to find which model can better predict the launch outcome.
- Four classification models were Logistic Regression, Support Vector Machine, Decision Tree and K Nearest Neighbors

- Please kindly refers to the notebooks below:
- https://github.com/ngsiukit/Applied-Data-Science-Capstone---SpaceX/blob/main/Week4%20Machine%20Learning.ipynb

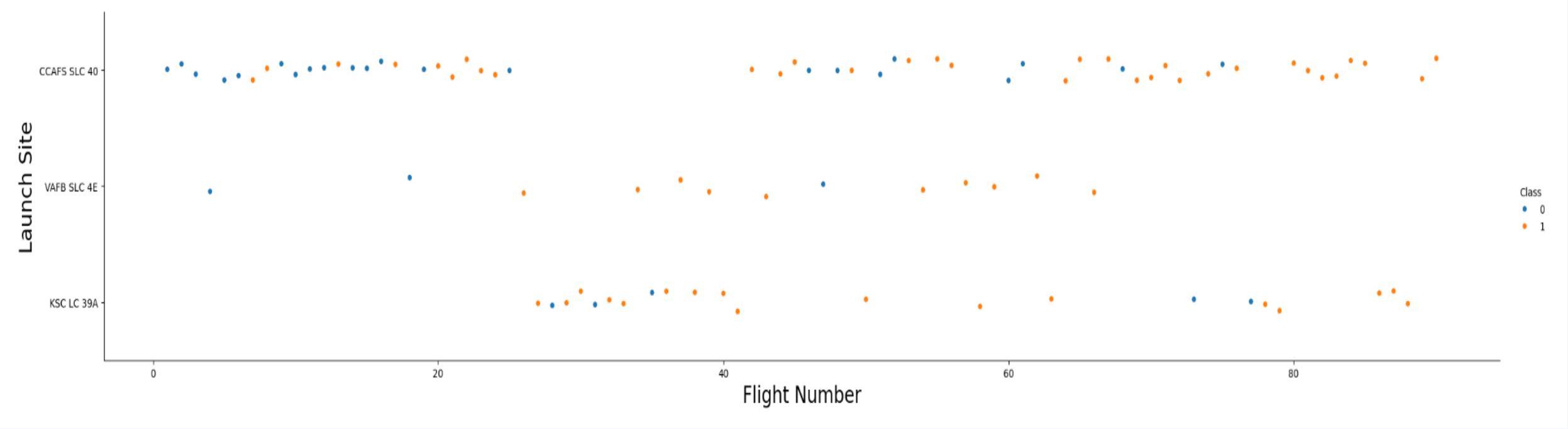
Results

- Different analytic results will be presented in later slides, including
  - Exploratory data analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results using classification models


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

- For Flight number between 25 to 40, there is no launching for Launch Site CCAFS SLC-40.
- The launch switched to Launch Site KSC LC-39A at that time.

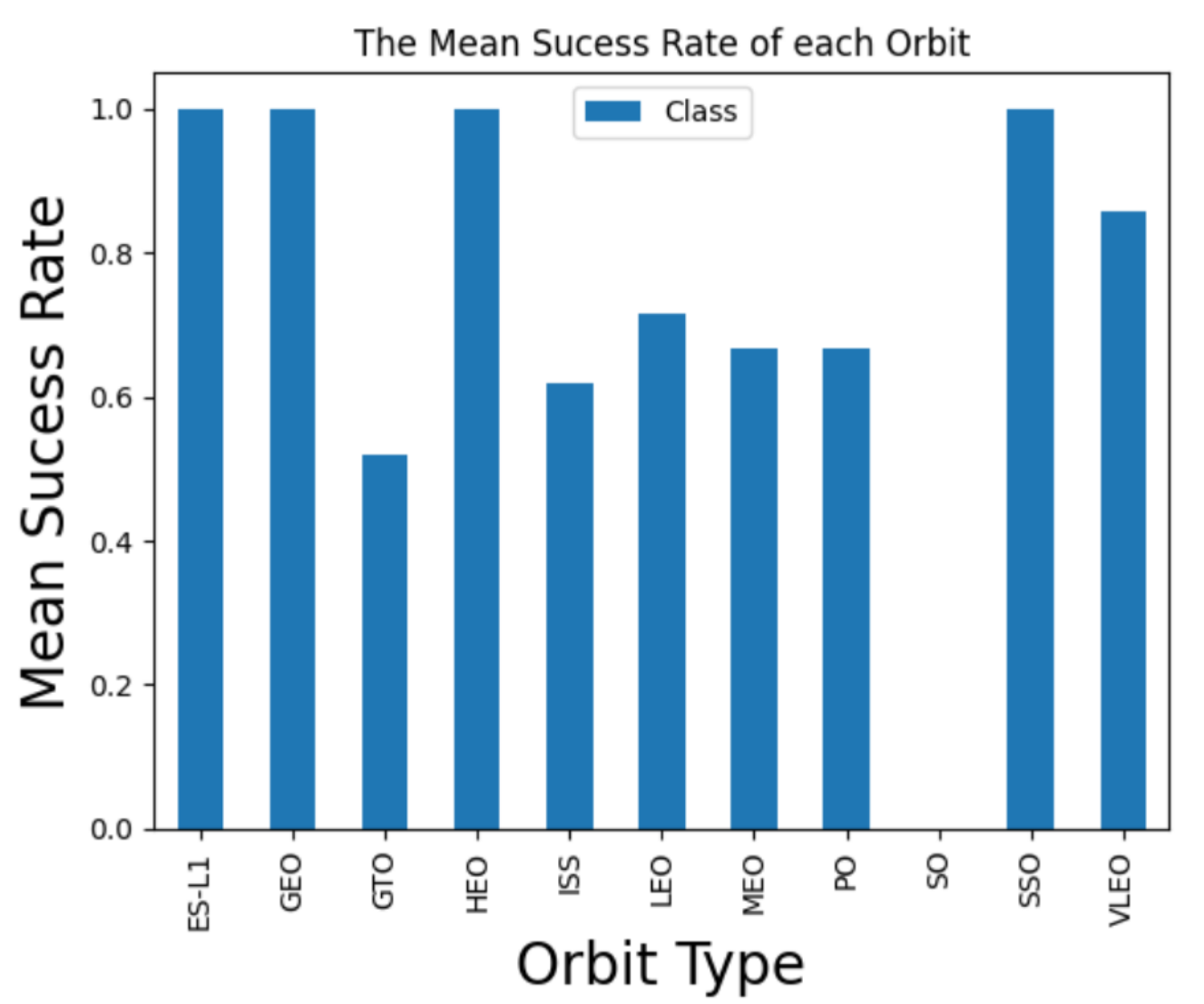


Payload vs. Launch Site

- Most of the rocket launched had payload between 2000kg to 7000kg.
- Launch Site CCAFS SLC-40 had launched several rockets with payload higher than 15000kg which had a high success rate.

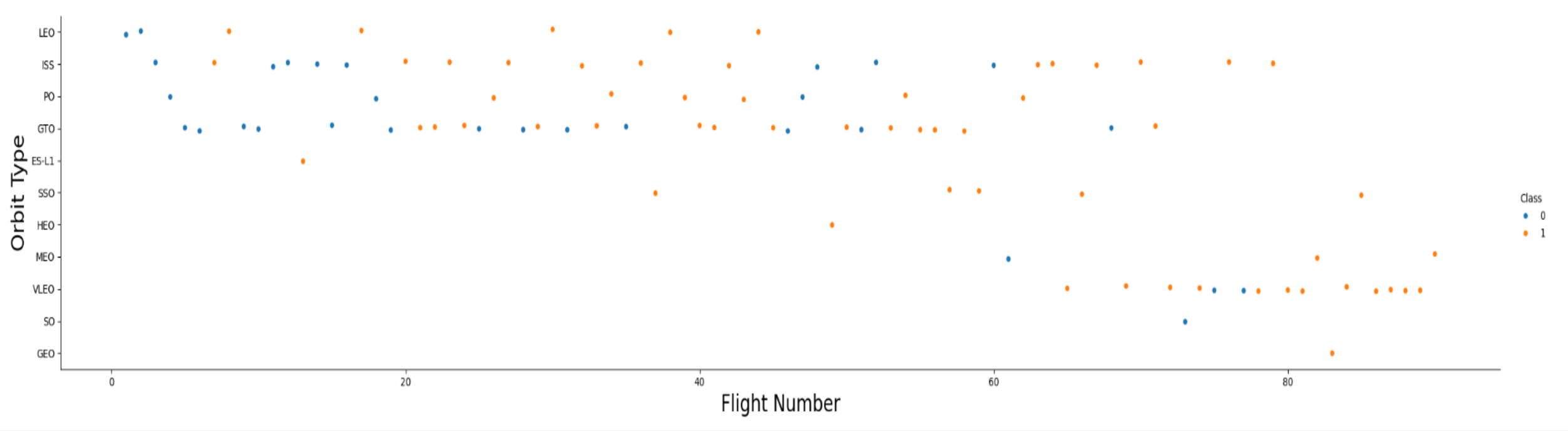
Success Rate vs. Orbit Type

- Orbit ES-L1, GEO, HEO, SSO had 100% success rate.
- However, all of the Orbit ES-L1, GEO, HEO actually had only launched once. The success rate may not reflect the true situation.
- Orbit SSO had 100% success rate with a total of five launches.



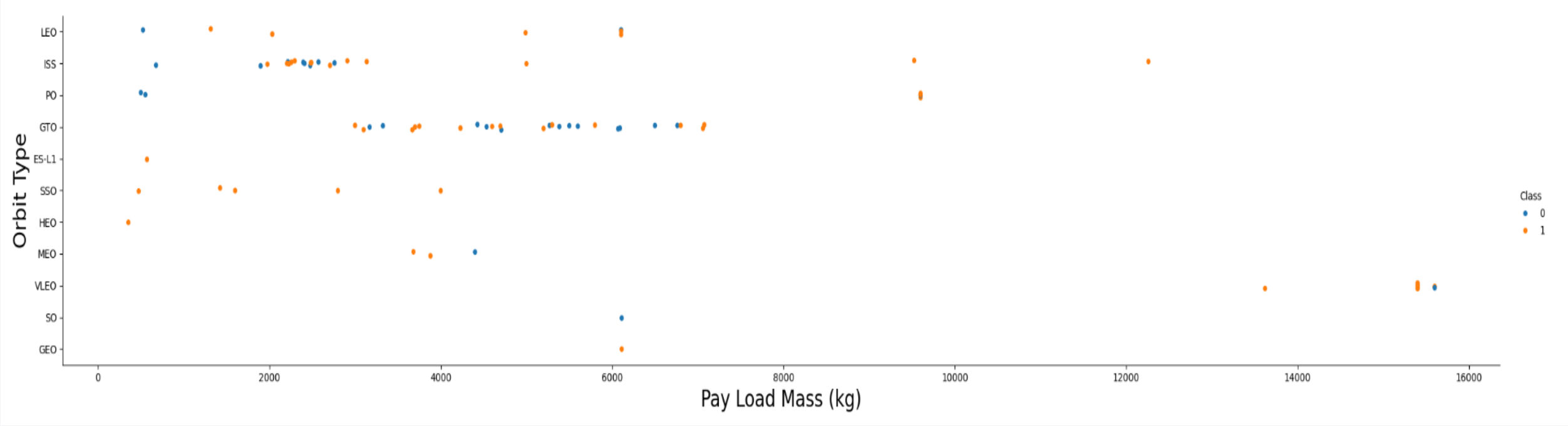
Flight Number vs. Orbit Type

- As mentioned before, Orbit SSO had 100% success rate with a total of five launches.
- Orbit VLEO, although not 100% success rate, had 12 success out of 14 total launches.
- And Orbit ISS although had low total success rate, is doing well in recent launches.

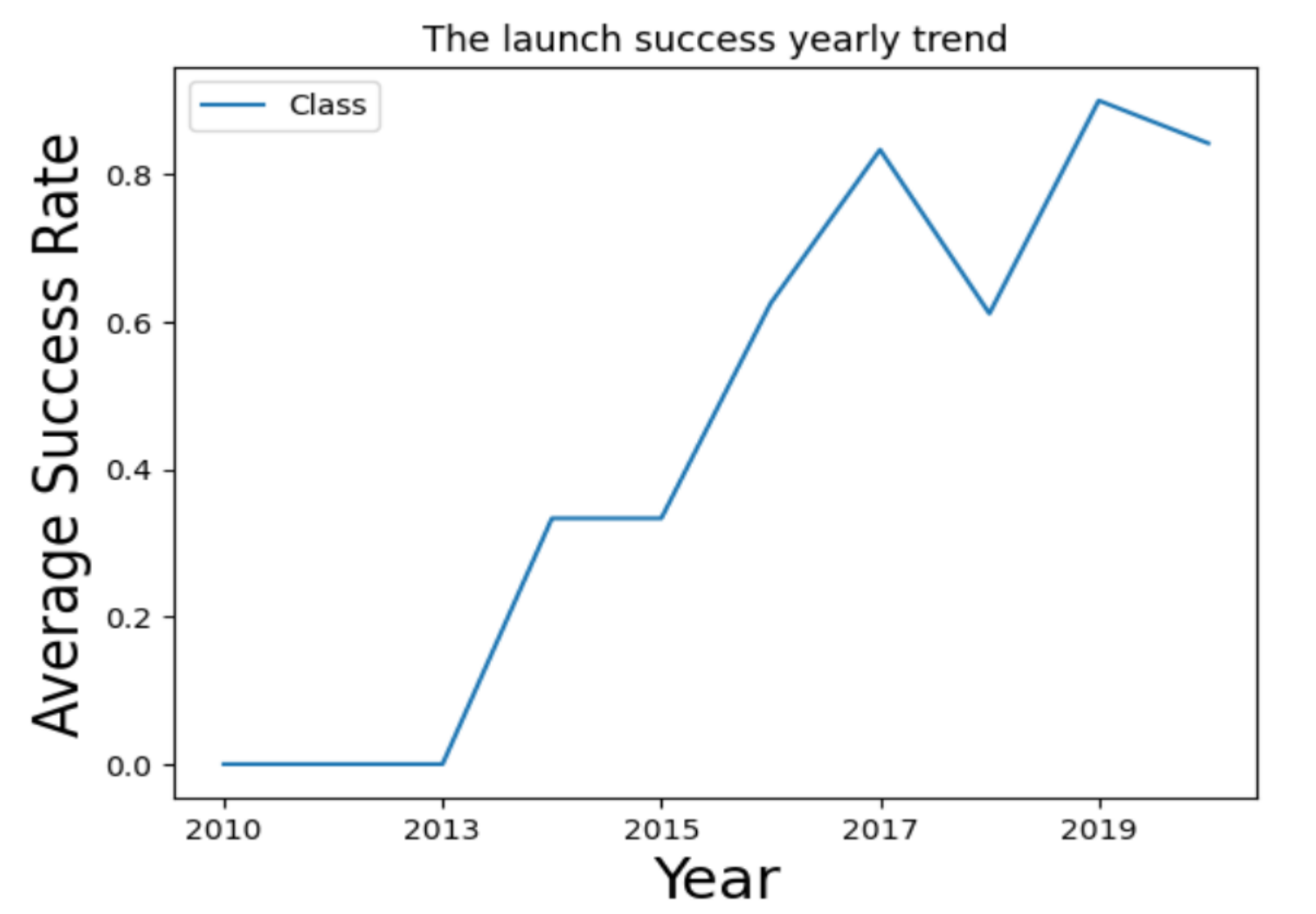


Payload vs. Orbit Type

- As mentioned before, most of the Orbit Types had payload between 2000kg to 7000kg.
- Orbit VLEO had highest payload which were all over 13000kg.
- All Orbit VLEO SSO had payload below 4000kg.



The overall Launch Success Rate of Falcon 9 increases yearly



There are four Launch Site of Falcon 9

Task 1

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

[8]:
```
%sql select distinct Launch_Site from SPACEXTABLE
```

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

[8]:

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

Launch Site Names Begin with 'CCA'

Task 2

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

[9]:
```
%sql select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5
```

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

[9]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outc
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Su
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	Su
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Su
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Su
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Su

Total Payload Mass produced by launching for NASA (CRS) was 45596kg

Task 3

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

```
[10]: %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = 'NASA (CRS)'
```

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

[10]:

sum(PAYLOAD_MASS__KG_)
45596

The Average Payload Mass produced by Booster version F9 v1.1 was 2535kg

Task 4

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

[11]:
```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version like 'F9 v1.1%'
```

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

[11]:

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

The First Successful Landing date was at Dec 22, 2015

Task 5

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

Hint:Use min function

[12]:
```
%sql select min(Date) from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)'
```

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

[12]:

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

[13]:
```
%sql select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ >=4000 and PAYLOAD_MASS__KG_ <=6000 and Landing_Outcome = 'Success (drone ship)'
```

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

[13]:

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

[14]:
```
%sql select Mission_Outcome, count(Mission_Outcome) from SPACEXTABLE group by Mission_Outcome
```

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

[14]:

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

List of Boosters which carried Maximum Payload

Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

[15]:
```
%sql select Booster_Version from SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (select MAX(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

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

[15]:

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

Two Failure landing in Drone Ship in 2015

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, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

[16]:
```
#%sql SELECT "Booster_Version", "Launch_Site" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Failure (drone ship)' AND substr(Date,1,4) = '2015'
%sql SELECT substr(Date,6,2) as month, Landing_Outcome, Booster_Version from SPACEXTABLE where Landing_Outcome = 'Failure (drone ship)' and substr(Date,0,5) = '2015'
```

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

[16]:

month	Landing_Outcome	Booster_Version
01	Failure (drone ship)	F9 v1.1 B1012
04	Failure (drone ship)	F9 v1.1 B1015

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

Task 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.

[17]:
```
#%sql SELECT "Landing_Outcome", COUNT(*) as 'COUNT' FROM SPACEXTABLE WHERE substr(Date,1,4) || substr(Date,6,2) || substr(Date,9,2) between '20100604' and '20
%sql SELECT Landing_Outcome, COUNT(*) AS QTY FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY QTY DESC
```

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

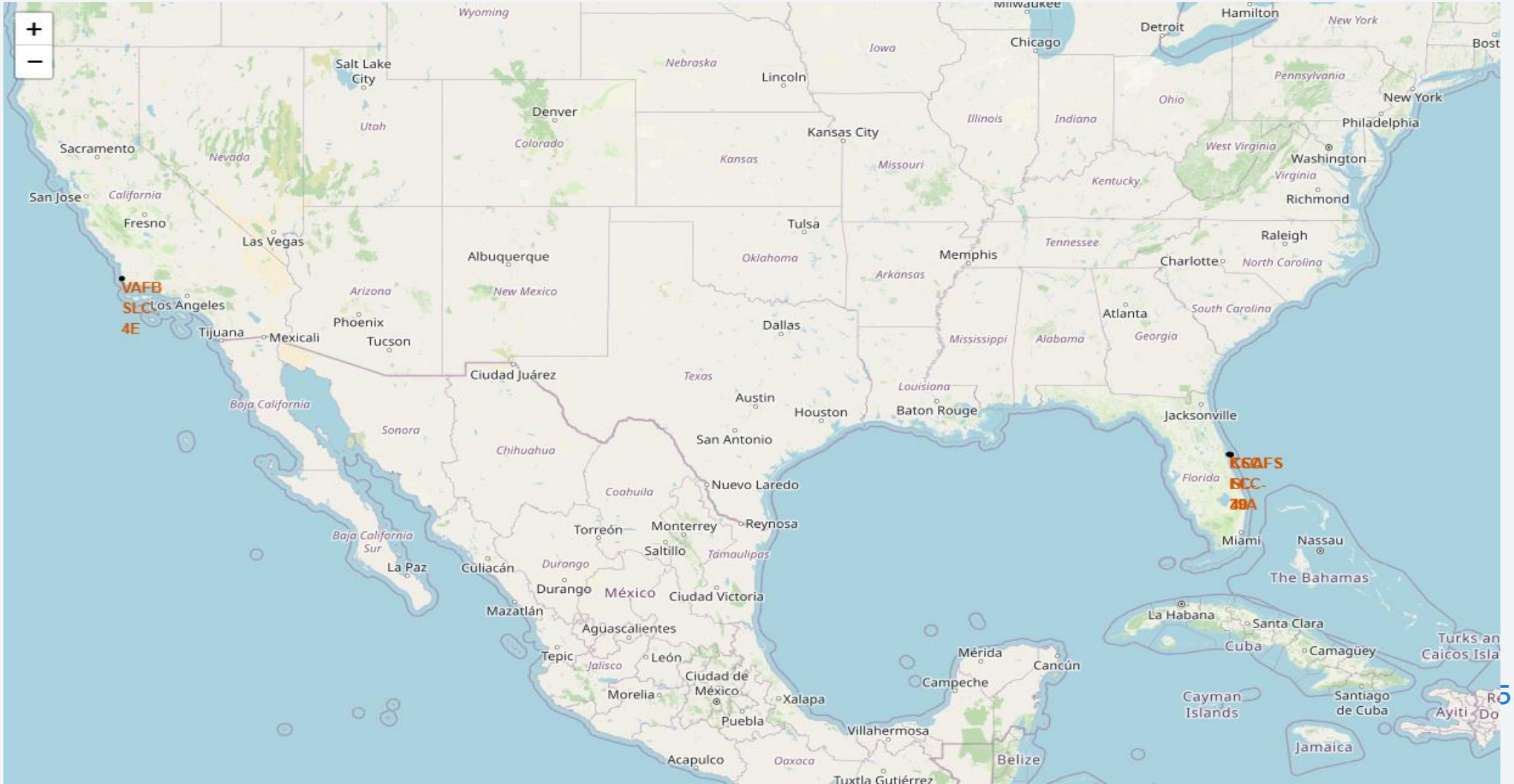
[17]:

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

Section 3

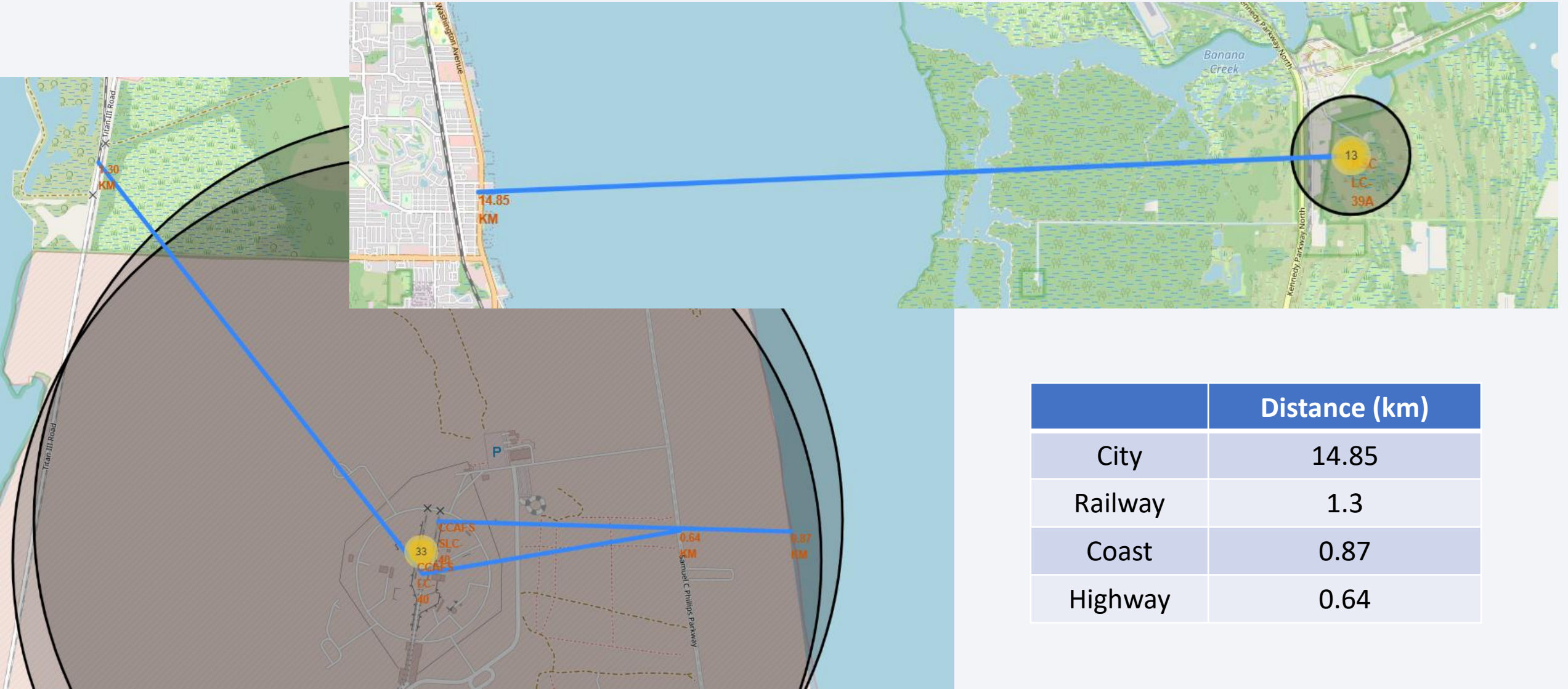
Launch Sites Proximities Analysis

All of the four Falcon 9 launch sites located near the coast

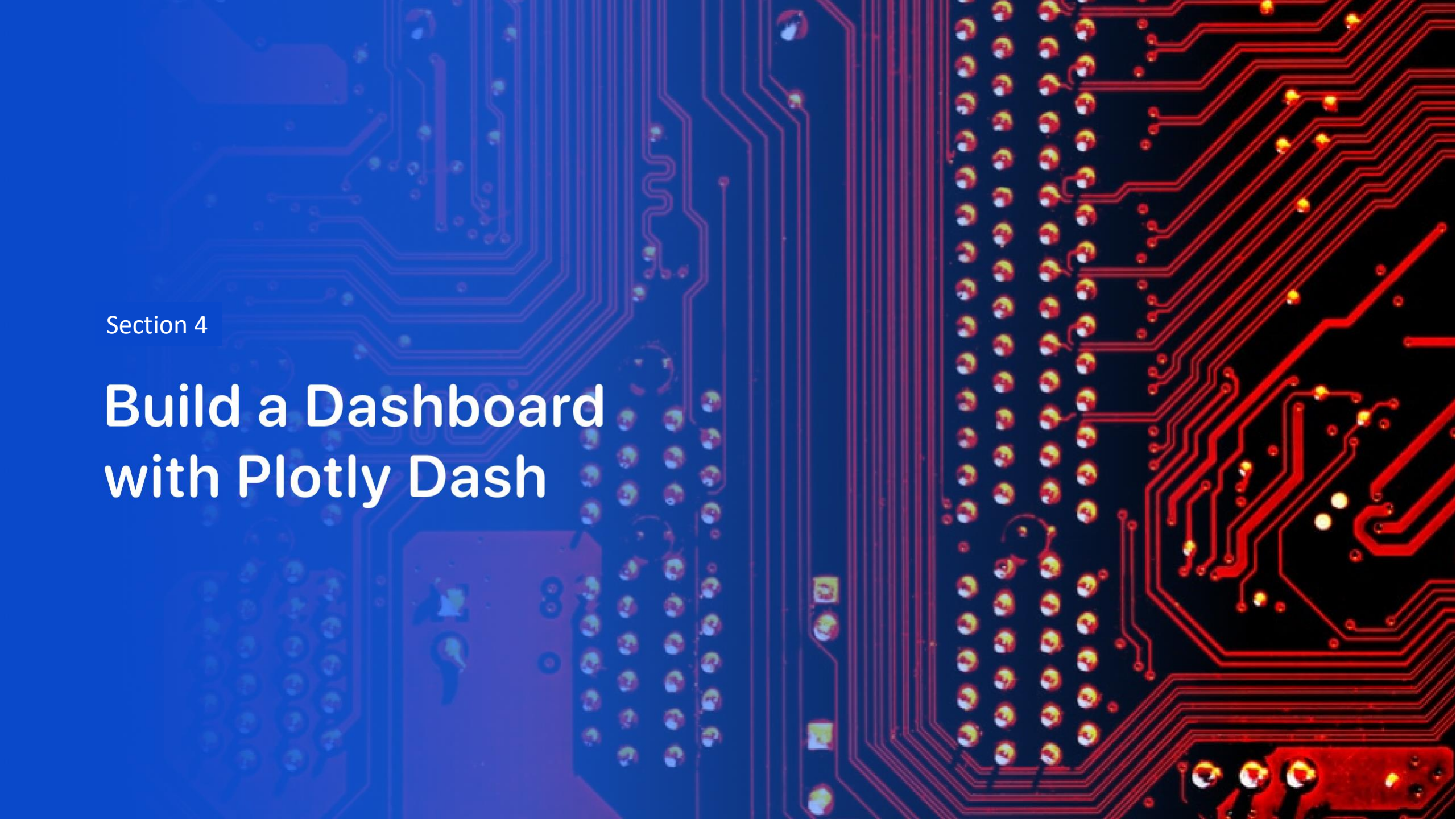


Launch Site KSC LC-39A has the most successful landing cases (Green Markers)

Launch sites located far away from cities.
Distance to railway, coast and highway are similar



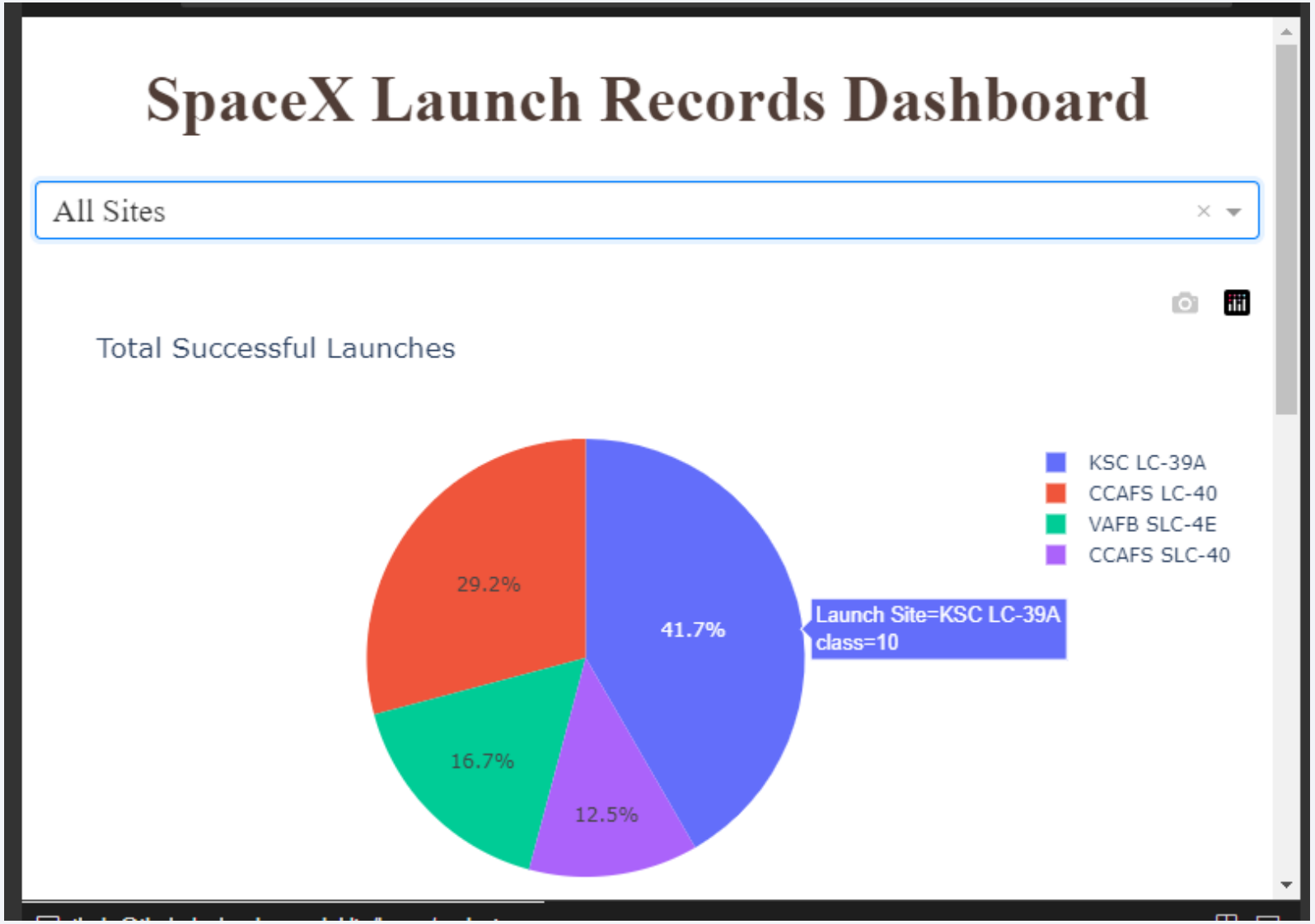
	Distance (km)
City	14.85
Railway	1.3
Coast	0.87
Highway	0.64


Section 4
Build a Dashboard with Plotly Dash

41% of the successful landing was from Launch Site KSC LC-39A

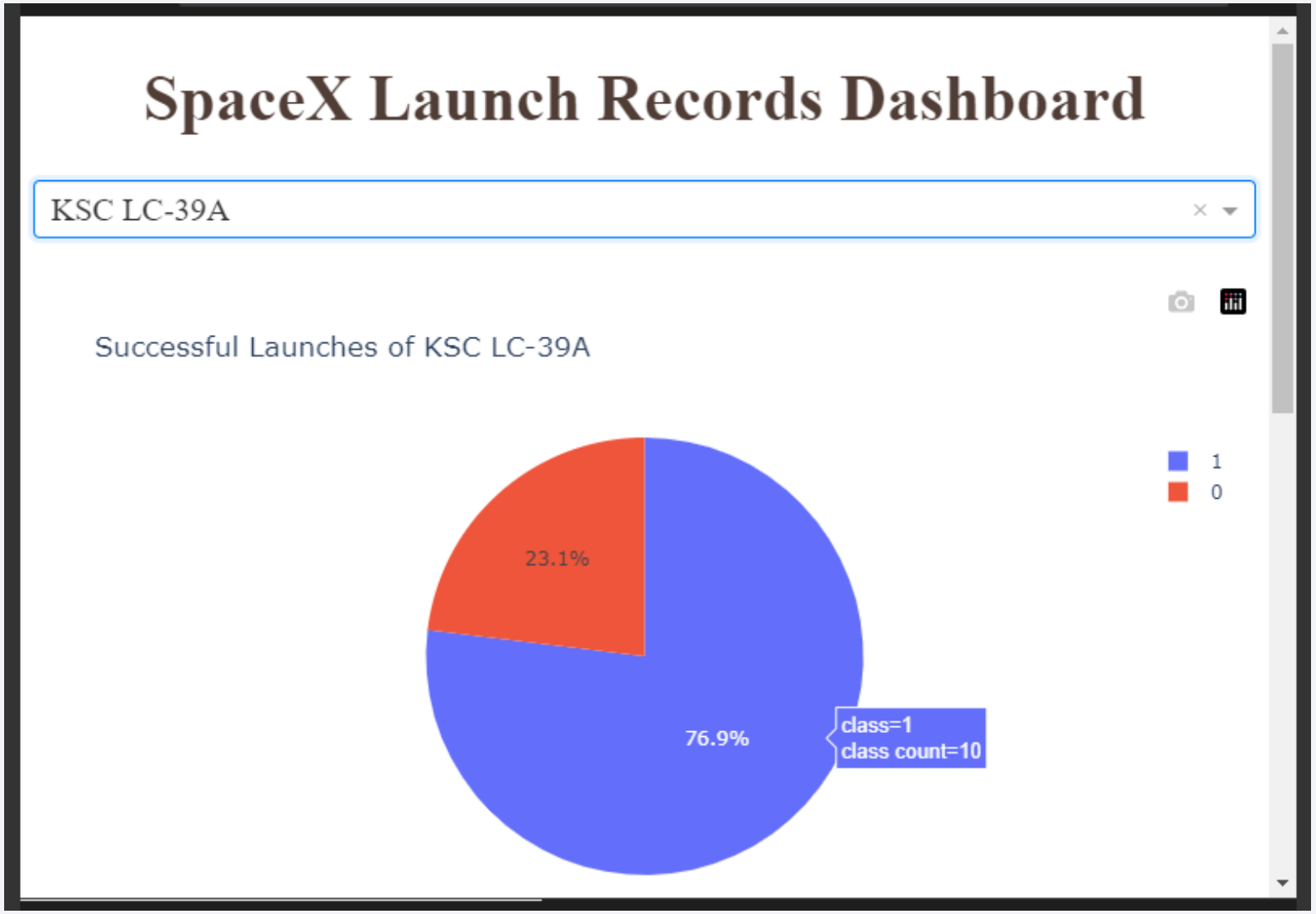
Launch Site	Successful Landing
KSC LC-39A	10
CCAFS LC-40	7
VAFB SLC-4E	4
CCAFS SLC-40	3

Launch Site	Successful Landing
KSC LC-39A	76.9%
CCAFS LC-40	26.9%
VAFB SLC-4E	40%
CCAFS SLC-40	42.9%

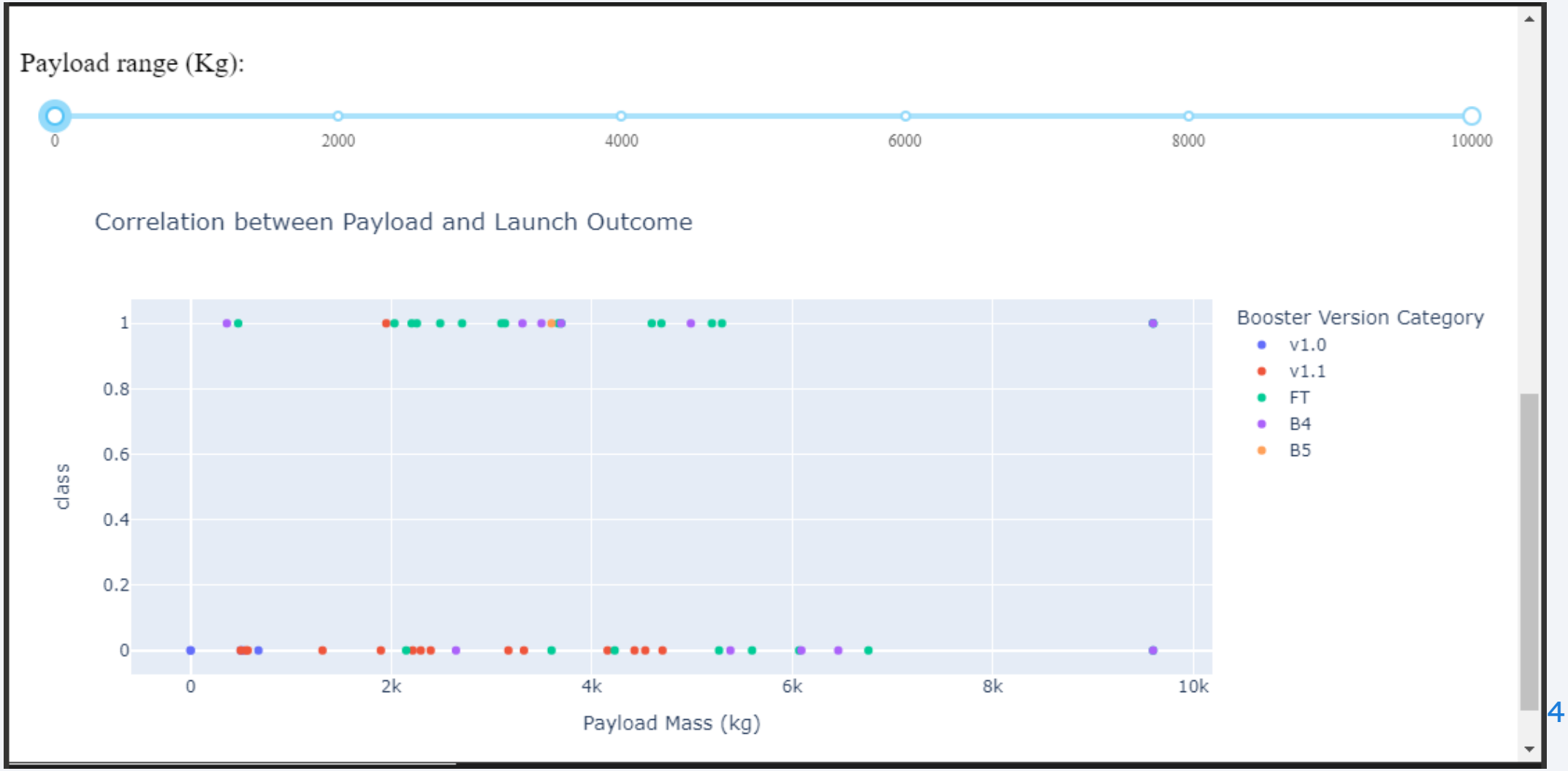


Launch Site KSC LC-39A has a successful launching rate of 76.9%

- 10 out of 13 launches was successful at Launch Site KSC LC-39A (76.9%)
- Compared to Launch Site CCAFS LC-40, only 7 out of 26 launches was successful (26.9%)



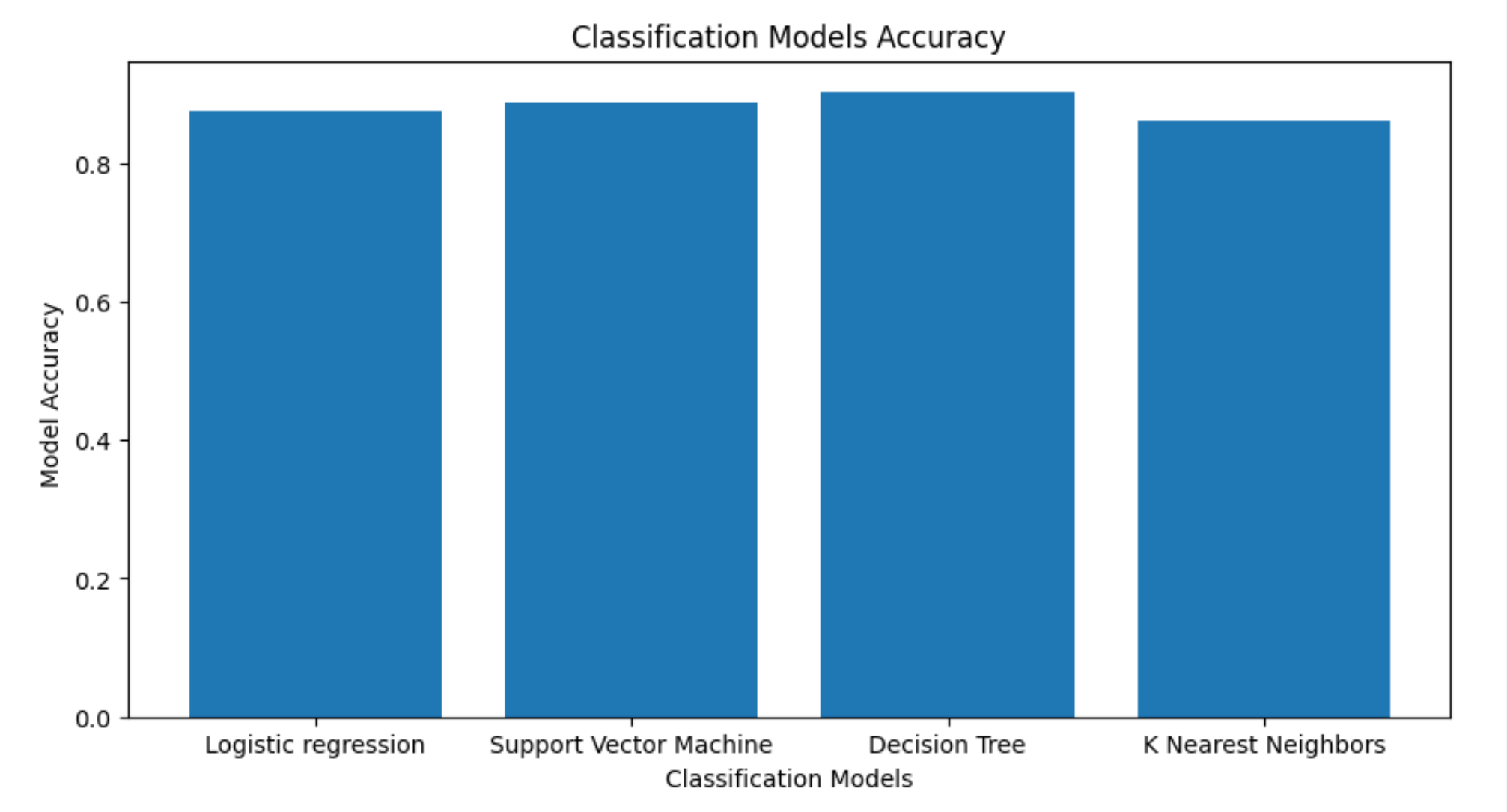
Successful landing mainly from Booster version FT and Payload Mass between 2000kg and 6000kg



Section 5

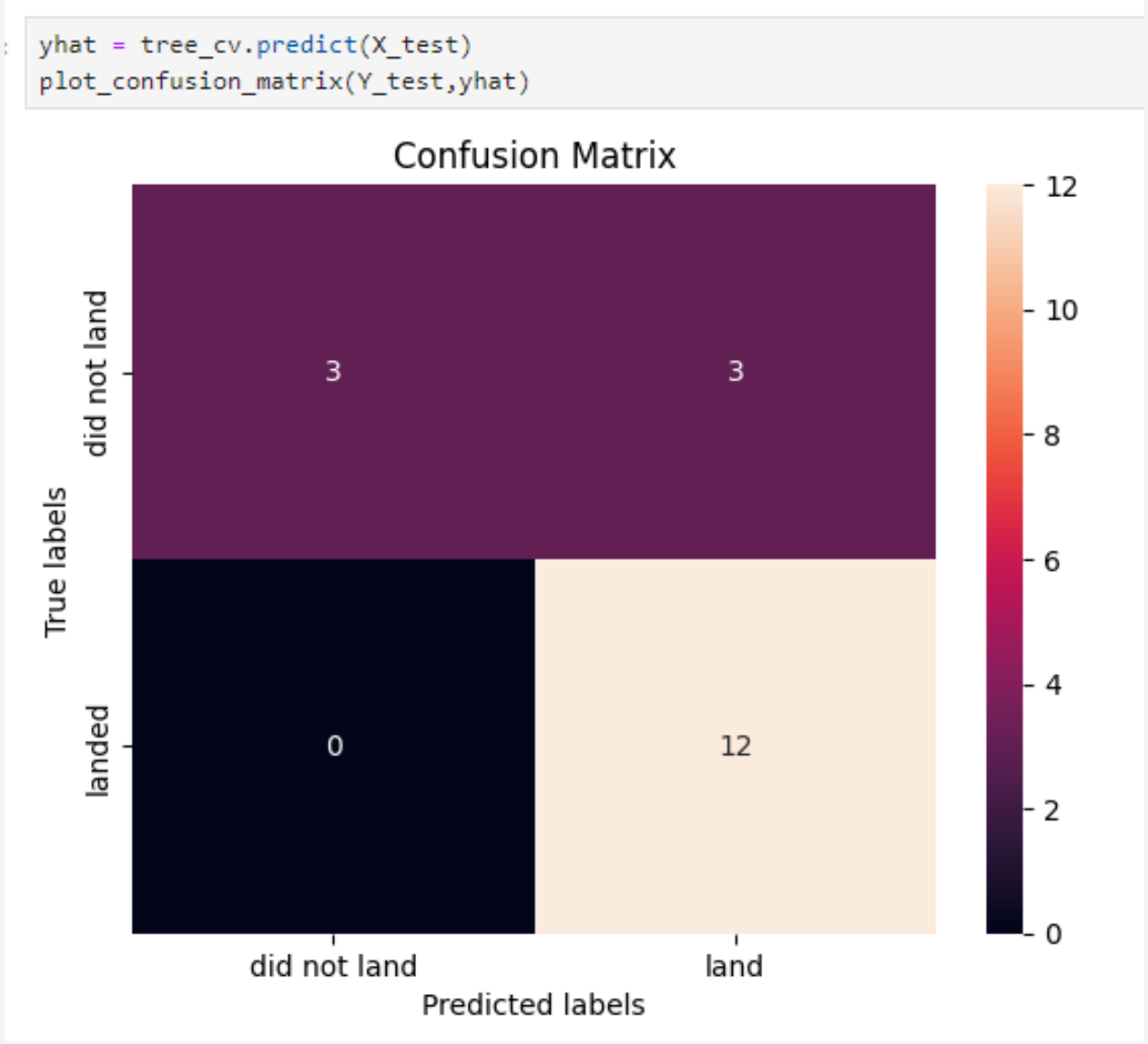
Predictive Analysis (Classification)

Decision Tree has highest Classification Model Accuracy with the four tests



Confusion Matrix of Decision Tree
The only problem is False Positive

```
yhat = tree_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



Conclusions

- Among the four classification methods, Decision Tree has a slightly higher accuracy score reaching 90%.
- Decision Tree has the problem of False Positive, which may predict a welcome result but actually fails.
- If we use 90% accuracy of prediction to calculate the expected cost, the expected cost of Launching will be 62*0.9+162*0.1=$72 mil, which is still a worthy bet.

Appendix

- All completed notebooks and Python files were provided through GitHub repository.

- Please refer to the link below:
- https://github.com/ngsiukit/Applied-Data-Science-Capstone---SpaceX/tree/main

Thank you!