Winning Space Race with Data Science

Lee Kun Wei
2024.09.15

Outline

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

Executive Summary

- Summary of methodologies
  - collecting data from various sources
  - improve the quality by performing data wrangling
  - exploring the processed data to gather insights
  - to build, evaluate, and refine predictive models
- Summary of all results
  - EDA results
  - Interactive analytics
  - Predictive analysis

Introduction

- Project background and context
  - SpaceX 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 SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch.
- Problems you want to find answers
  - to predict if the Falcon 9 first stage will land successfully.

Section 1

Methodology

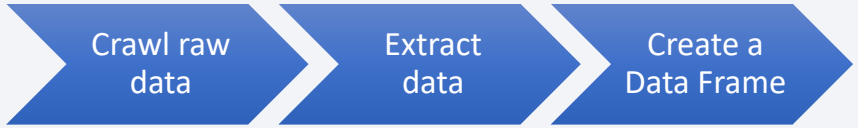
Methodology

Executive Summary

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

Data Collection

- There are 2 methods to collect datasets, SpaceX API and scraping from Wikipedia.
- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts
  - Web Scraping from Wikipedia
  - SpaceX API



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

  - https://github.com/kwlee1201/Coursera_IBM/blob/main/Ch10_jupyter-labs-spacex-data-collection-api.ipynb

- SpaceX API

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	
...	...	...	...	...	...	...	...	...	...	...	...	
89	86	2020-09-03	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	2	True	True	True	5e9e3032383
90	87	2020-10-06	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	3	True	True	True	5e9e3032383
91	88	2020-10-18	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	6	True	True	True	5e9e3032383
92	89	2020-10-24	Falcon 9	15600.0	VLEO	CCSFS SLC 40	True ASDS	3	True	True	True	5e9e3033383
93	90	2020-11-05	Falcon 9	3681.0	MEO	CCSFS SLC 40	True ASDS	1	True	False	True	5e9e3032383

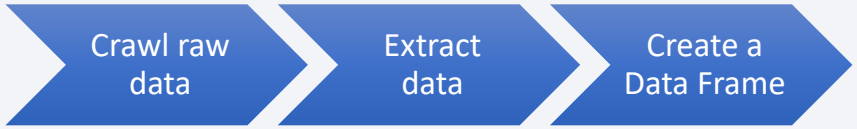
90 rows × 17 columns

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
  - https://github.com/kwlee1201/Coursera_IBM/blob/main/Ch10_jupyter-labs-webscraping.ipynb

- Web Scraping from Wikipedia



	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Launch Site	Version Booster	Booster landing	Date	Time
0	1	NaN	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	CCAFS	F9 v1.0B0003.18	Failure	4 June 2010	18:45
1	2	NaN	Dragon	0	LEO	NASA	Success	CCAFS	F9 v1.0B0004.18	Failure	8 December 2010	15:43
2	3	NaN	Dragon	525 kg	LEO	NASA	Success	CCAFS	F9 v1.0B0005.18	No attempt\n	22 May 2012	07:44
3	4	NaN	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	CCAFS	F9 v1.0B0006.18	No attempt	8 October 2012	00:35
4	5	NaN	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	CCAFS	F9 v1.0B0007.18	No attempt\n	1 March 2013	15:10

Data Wrangling

- Describe how data were processed
  - Exploratory Data Analysis ：exploratory of data types, distribution, percentage of missing values
  - Determine Training Labels
- 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
  - https://github.com/kwlee1201/Coursera_IBM/blob/main/Ch10_labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
  - scatter point chart：Showing the relationship between two attributes
  - bar chart ：Comparison of numerical values across categories
  - line chart ：Changes in values over different points in time"
- Add the GitHub URL of your completed EDA with data visualization notebook, as an external reference and peer-review purpose
  - https://github.com/kwlee1201/Coursera_IBM/blob/main/Ch10_edadataviz.ipynb

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
  - Counting the number of records for different data types
  - exploratory of data distribution
- Add the GitHub URL of your completed EDA with SQL notebook, as an external reference and peer-review purpose
  - https://github.com/kwlee1201/Coursera_IBM/blob/main/Ch10_jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

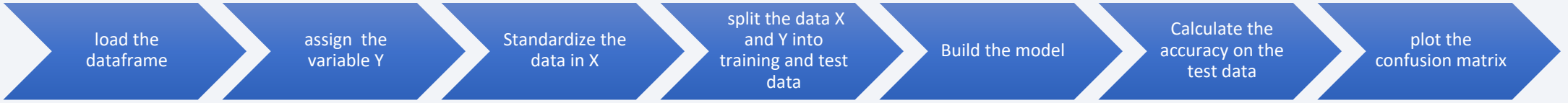
- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
- Explain why you added those objects
  - markers：Mark the success/failed launches for each site on the map
  - lines ： Calculate the distances between a launch site to its proximities
- Add the GitHub URL of your completed interactive map with Folium map, as an external reference and peer-review purpose
  - https://github.com/kwlee1201/Coursera_IBM/blob/main/Ch10_lab_jupyter_launch_site_location.ipynb

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
  - Piechart: Show the launch success count and ratio for all sites
  - scatter plot : Show of Payload vs. Launch Outcome scatter plot for all sites
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose
  - https://github.com/kwlee1201/Coursera_IBM/blob/main/Ch10_spacex_dash_app.py

Predictive Analysis (Classification)

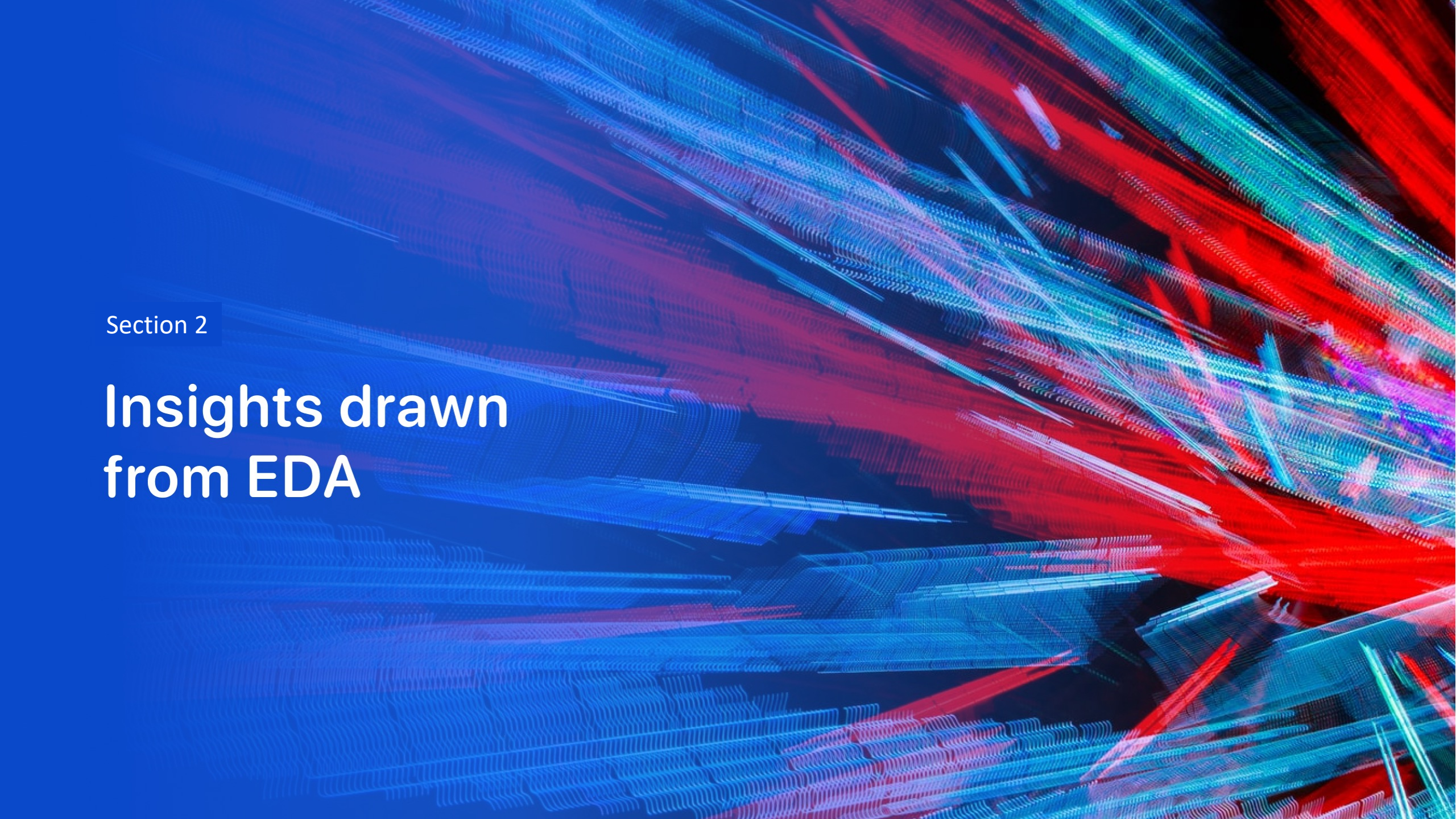
- Summarize how you built, evaluated, improved, and found the best performing classification model
- You need present your model development process using key phrases and flowchart



- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose
  - https://github.com/kwlee1201/Coursera_IBM/blob/main/Ch10_SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

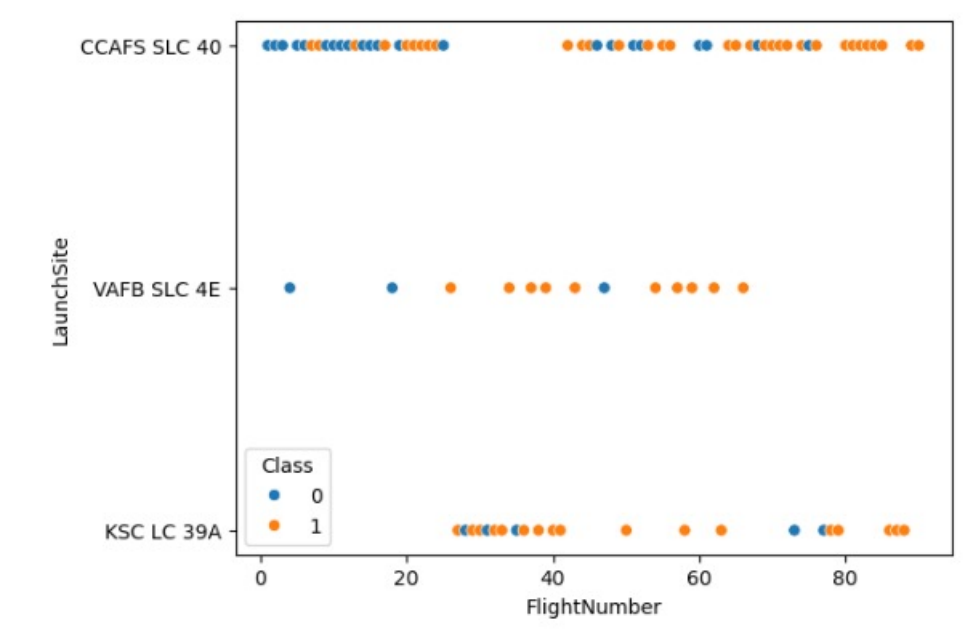
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

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

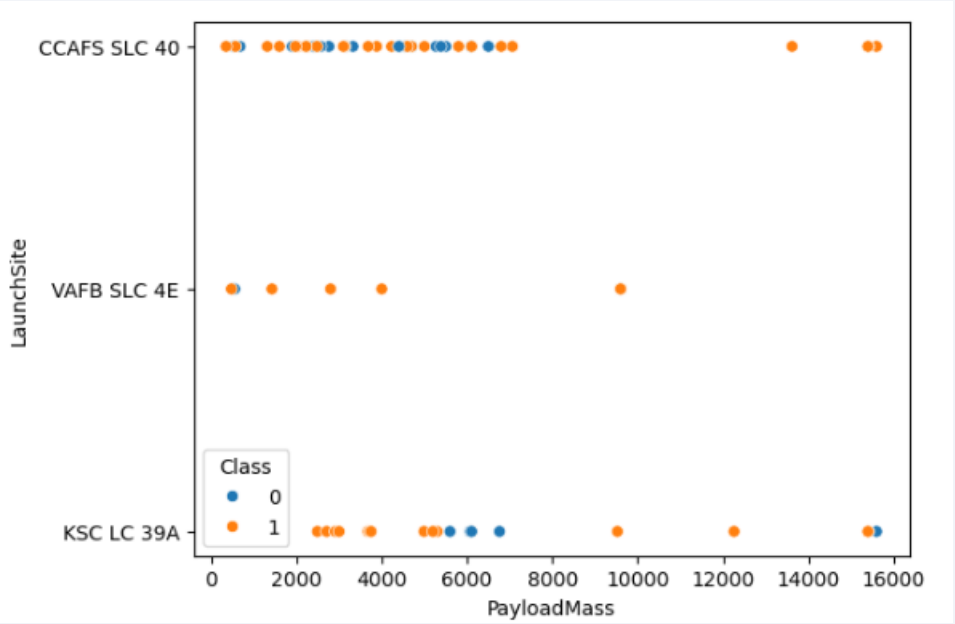
- Show the screenshot of the scatter plot with explanations
  - The success rates of CCAFS SLC 40 and KSC LC 39A are almost100% when FightNumber is over 80.



Payload vs. Launch Site

- Show a scatter plot of Payload vs. Launch Site

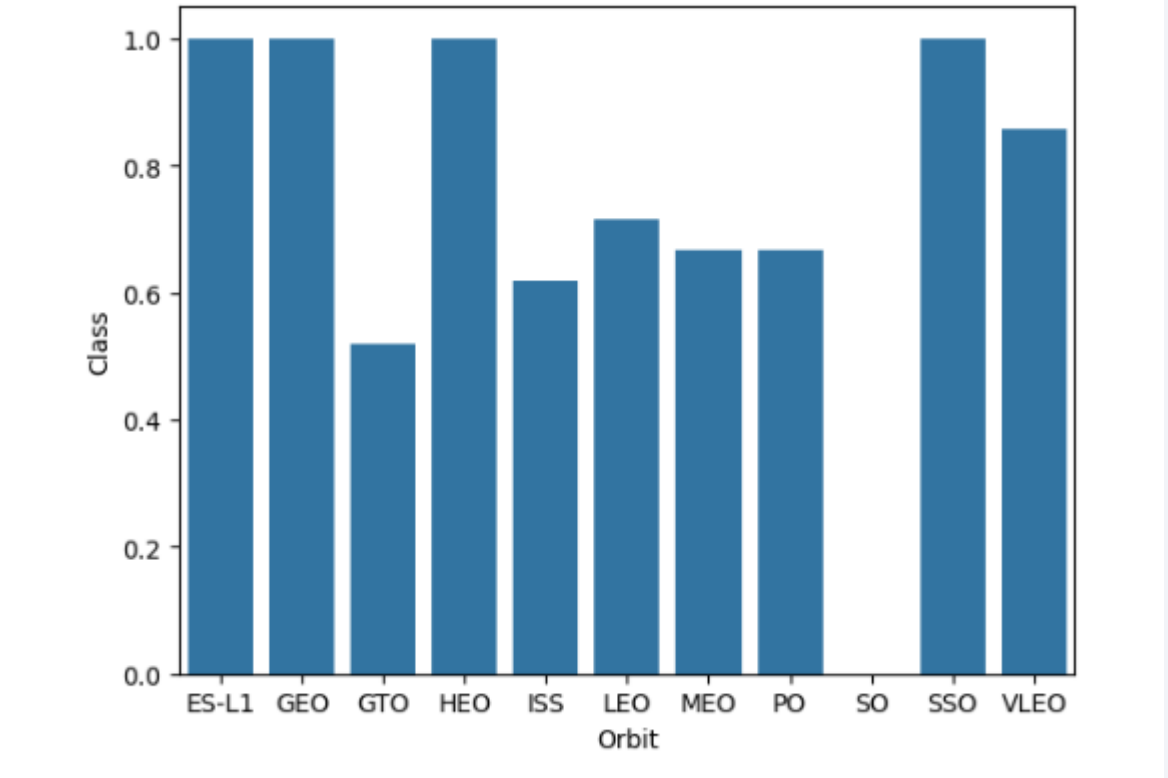
- Show the screenshot of the scatter plot with explanations
  - The failure rates is higher when PayloadMass is between 4000 and 7000.
  - AFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000)



Success Rate vs. Orbit Type

- Show a bar chart for the success rate of each orbit type

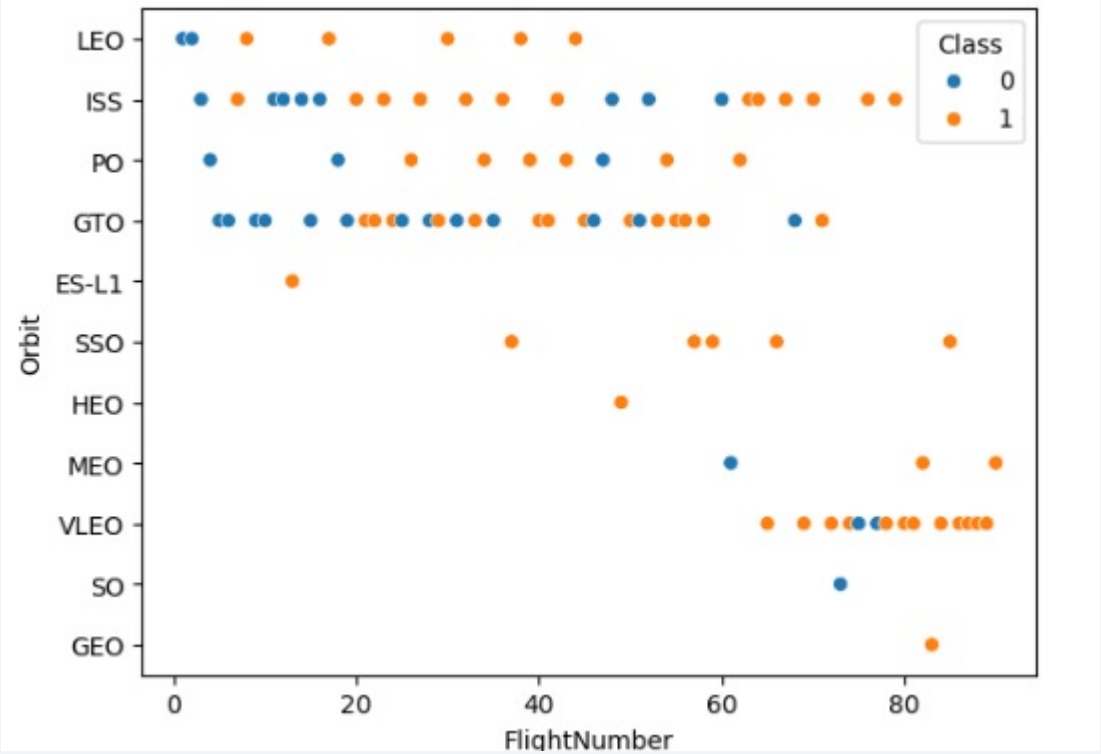
- Show the screenshot of the scatter plot with explanations
  - The success rates of ES-L1, GEO, HEO and SSO are 100%.



Flight Number vs. Orbit Type

- Show a scatter point of Flight number vs. Orbit type

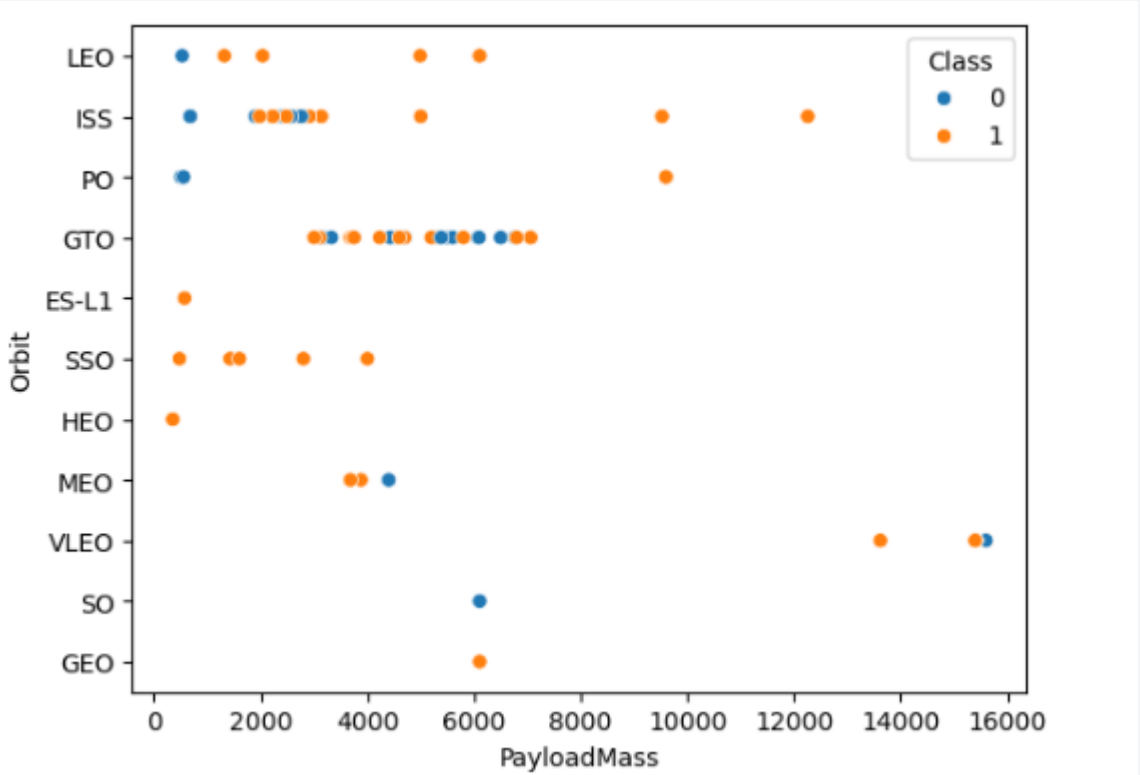
- Show the screenshot of the scatter plot with explanations
  - In the LEO orbit, success seems to be related to the number of flights.



Payload vs. Orbit Type

- Show a scatter point of payload vs. orbit type

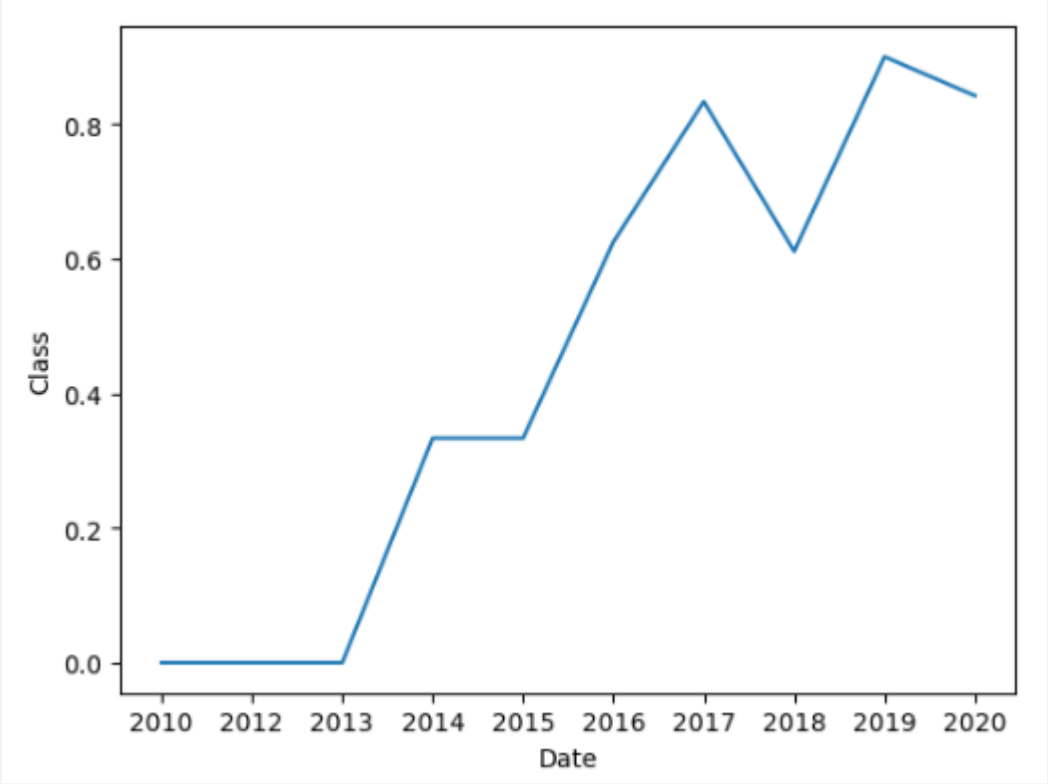
- Show the screenshot of the scatter plot with explanations
  - With heavy payloads the successful landing rate are more for LEO and ISS.



Launch Success Yearly Trend

- Show a line chart of yearly average success rate

- Show the screenshot of the scatter plot with explanations
  - success rate since 2013 kept increasing till 2020



All Launch Site Names

- Find the names of the unique launch sites
- Present your query result with a short explanation here
  - There are four unique launch sites, including CCAFS LC-40, CCAFS SLC-40, KSC LC-39A and VAFB SLC-4E.

Task 1

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

In [10]:
```
%sql SELECT LAUNCH_SITE, COUNT(1) FROM SPACEXTBL GROUP BY LAUNCH_SITE
```

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

Out[10]:

Launch_Site	COUNT(1)
CCAFS LC-40	26
CCAFS SLC-40	34
KSC LC-39A	25
VAFB SLC-4E	16

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Present your query result with a short explanation here
  - The first 5 records begin with ‘CCA’ which are all CCAFS LC-40

Task 2

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

In [11]:
```
%sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
```

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

Out[11]:

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

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- Present your query result with a short explanation here
  - There are 45,596 payload carried by boosters from NASA.

Task 3

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

In [12]:
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER='NASA (CRS)'
```

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

Out[12]:

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
  - The average payload mass carried by booster is 2928.4.

Task 4

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

In [13]:
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE BOOSTER_VERSION='F9 v1.1'
```

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

Out[13]:

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
  - It's 2010/6/4.

Task 5

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

Hint:Use min function

In [14]:
```
%sql SELECT MIN(DATE) FROM SPACEXTBL WHERE MISSION_OUTCOME='Success'
```

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

Out[14]:

MIN(DATE)
2010-06-04

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
  - As shown in the table below.

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

In [15…

```
%sql SELECT Booster_Version, COUNT(1) FROM SPACEXTBL WHERE MISSION_OUTCOME='Success' AND PAYLOAD_MASS__KG_ BETWEE
```

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

Out[15…

Booster_Version	COUNT(1)
F9 B4 B1040.2	1
F9 B4 B1040.1	1
F9 B5 B1046.2	1
F9 B5 B1046.3	1
F9 B5 B1047.2	1
F9 B5 B1048.3	1
F9 B5 B1051.2	1
F9 B5 B1058.2	1
F9 B5B1060.1	1
F9 B5B1062.1	1
F9 FT B1021.2	1
F9 FT B1031.2	1
F9 FT B1032.2	1
F9 FT B1020	1
F9 FT B1022	1
F9 FT B1026	1
F9 FT B1030	1
F9 FT B1032.1	1
F9 v1.1	1
F9 v1.1 B1011	1
F9 v1.1 B1014	1
F9 v1.1 B1016	1

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
  - Successful outcomes:100
  - Failure outcomes: 1

Task 7

List the total number of successful and failure mission outcomes

In [16]:
```
%sql SELECT MISSION_OUTCOME, COUNT(1) FROM SPACEXTBL GROUP BY MISSION_OUTCOME
```

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

Out[16]:

Mission_Outcome	COUNT(1)
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
  - As shown in the table below.
  - Payload mass are all 15600kg.

Task 8

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

In [17]:
```
%sql SELECT Booster_Version, PAYLOAD_MASS__KG_ , COUNT() FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ =(SELECT MAX(PAYLOA
```

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

Out[17]:

Booster_Version	PAYLOAD_MASS__KG_	COUNT()
F9 B5 B1048.4	15600	1
F9 B5 B1048.5	15600	1
F9 B5 B1049.4	15600	1
F9 B5 B1049.5	15600	1
F9 B5 B1049.7	15600	1
F9 B5 B1051.3	15600	1
F9 B5 B1051.4	15600	1
F9 B5 B1051.6	15600	1
F9 B5 B1056.4	15600	1
F9 B5 B1058.3	15600	1
F9 B5 B1060.2	15600	1
F9 B5 B1060.3	15600	1

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
  - Launch site are all CCAFS LC-40.

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.

In [18]:
```
%sql SELECT substr(Date, 6,2) as month, Booster_Version, Launch_Site FROM SPACEXTBL WHERE Landing_Outcome LIKE 'F
```

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

Out[18]:

month	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	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
  - The most frequent outcome is no attempt.

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.

In [19]:
```
%sql SELECT Landing_Outcome, COUNT(1) FROM SPACEXTBL WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY La
```

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

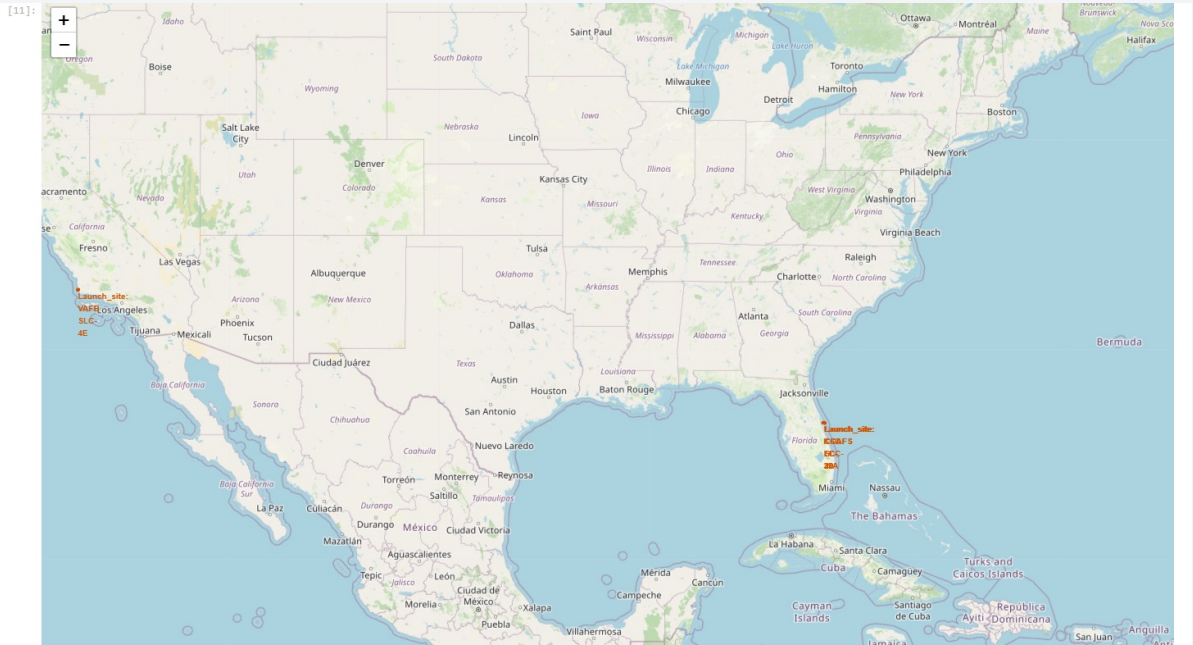
Out[19]:

Landing_Outcome	COUNT(1)
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

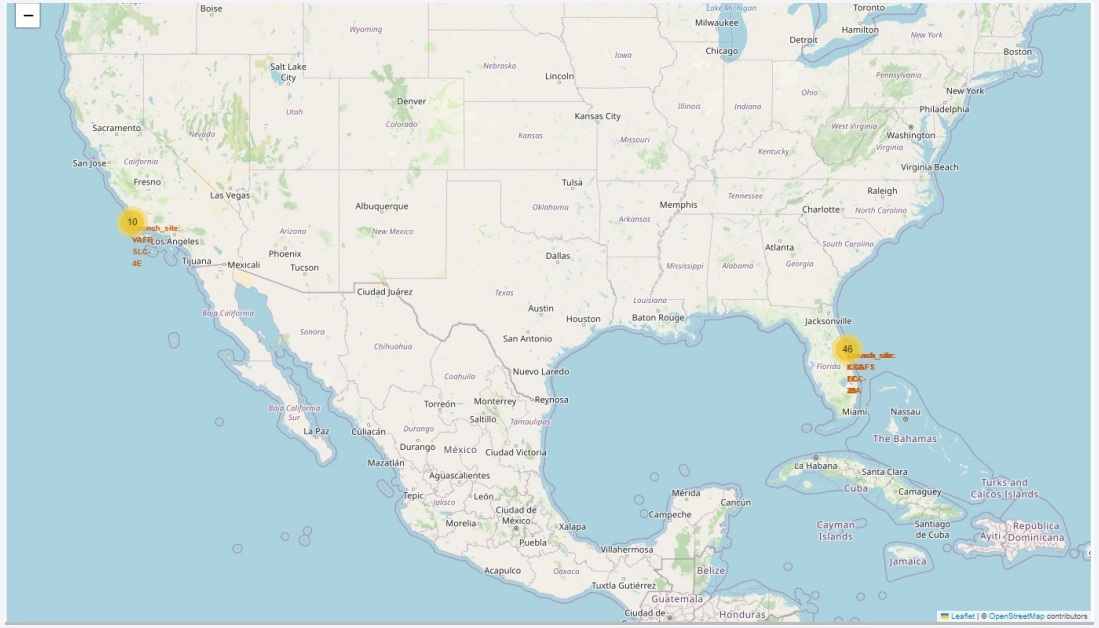
<Folium Map Screenshot 1>

- Replace <Folium map screenshot 1> title with an appropriate title
- Explore the generated folium map and make a proper screenshot to include all launch sites' location markers on a global map
- Explain the important elements and findings on the screenshot
  - All the launch sites are in very close proximity to the coast.



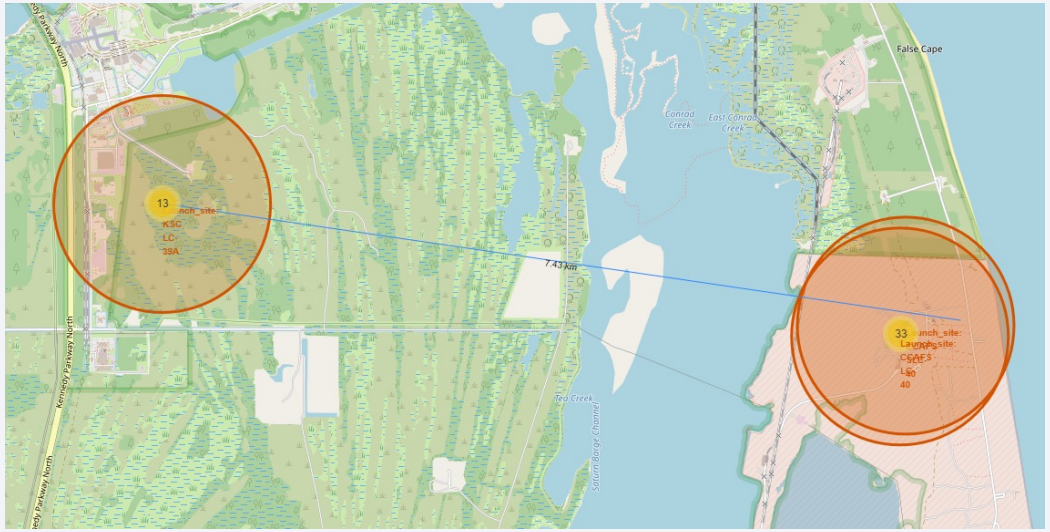
<Folium Map Screenshot 2>

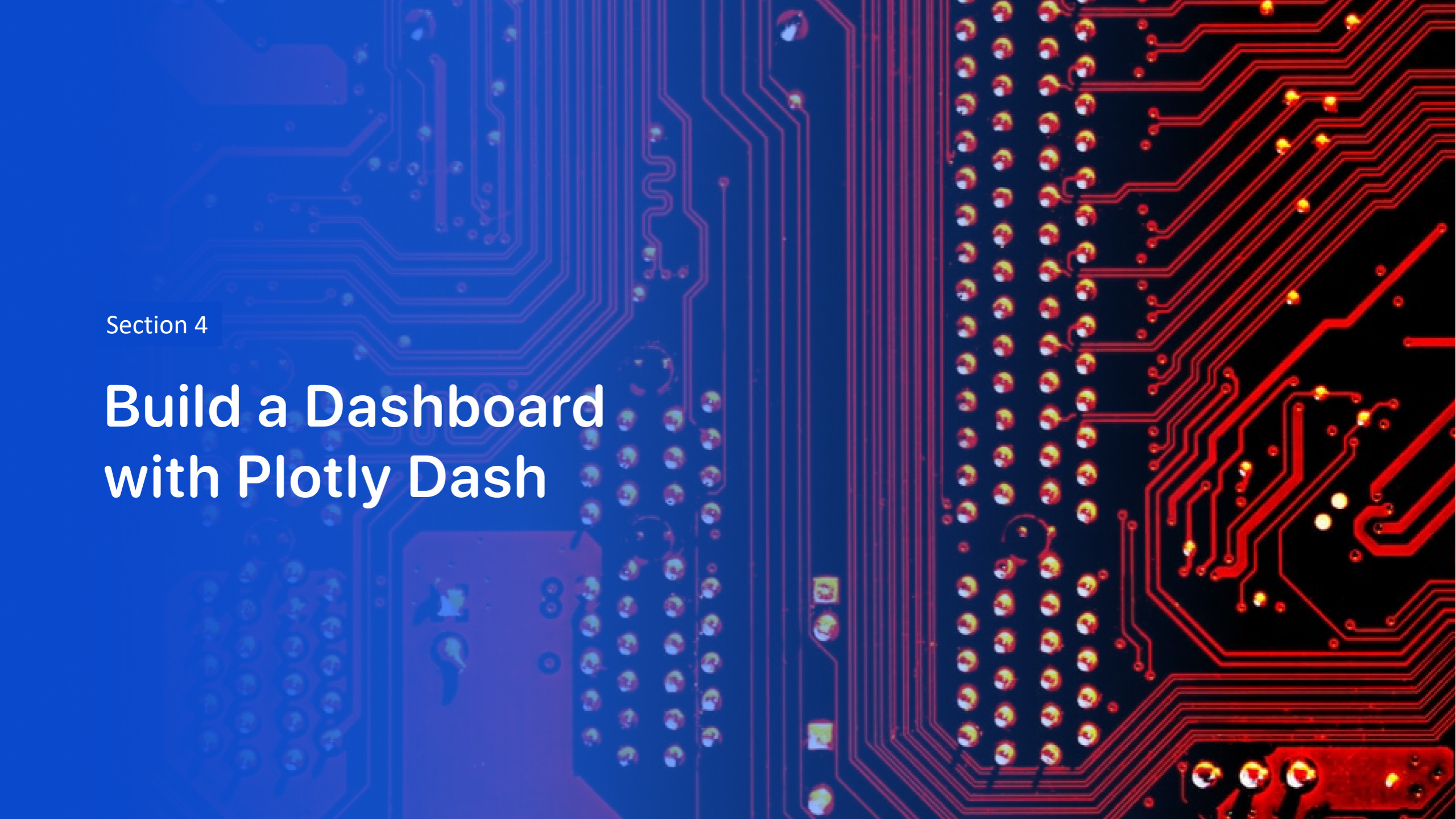
- Replace <Folium map screenshot 2> title with an appropriate title
- Explore the folium map and make a proper screenshot to show the color-labeled launch outcomes on the map
- Explain the important elements and findings on the screenshot
    - From the color-labeled markers in marker clusters, can easily identify which launch sites have relatively high success rates.



<Folium Map Screenshot 3>

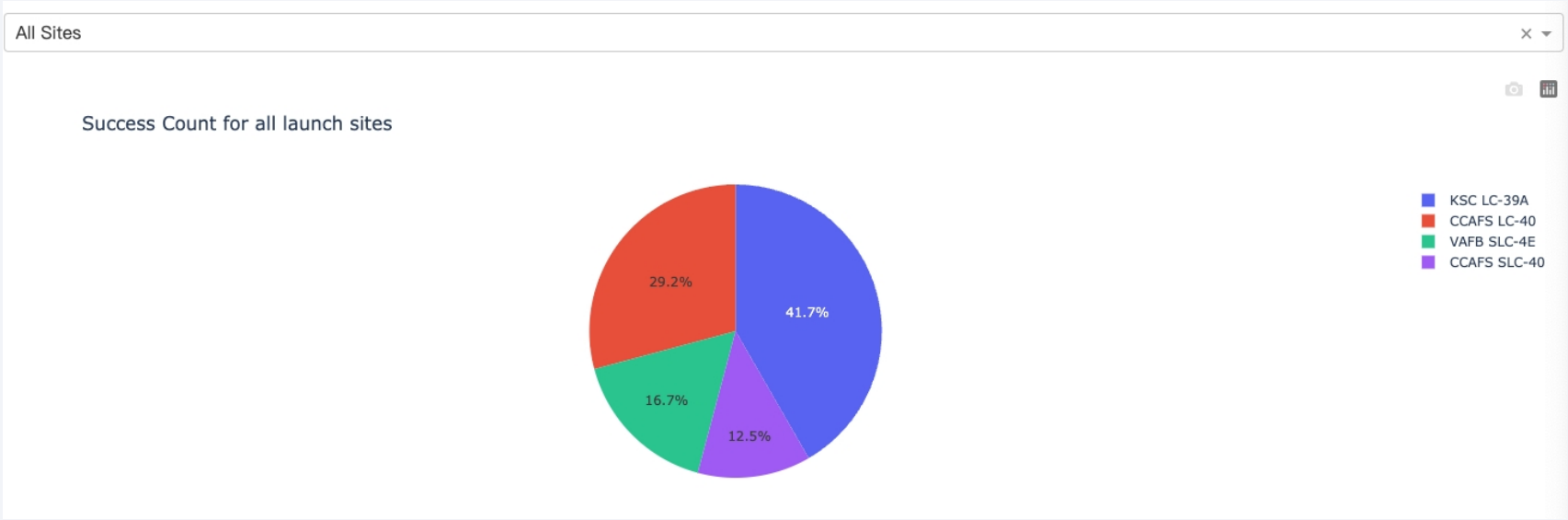
- Replace <Folium map screenshot 3> title with an appropriate title
- Explore the generated folium map and show the screenshot of a selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed
- Explain the important elements and findings on the screenshot
  - All launch sites are in close proximity to coastline.




Section 4
Build a Dashboard with Plotly Dash

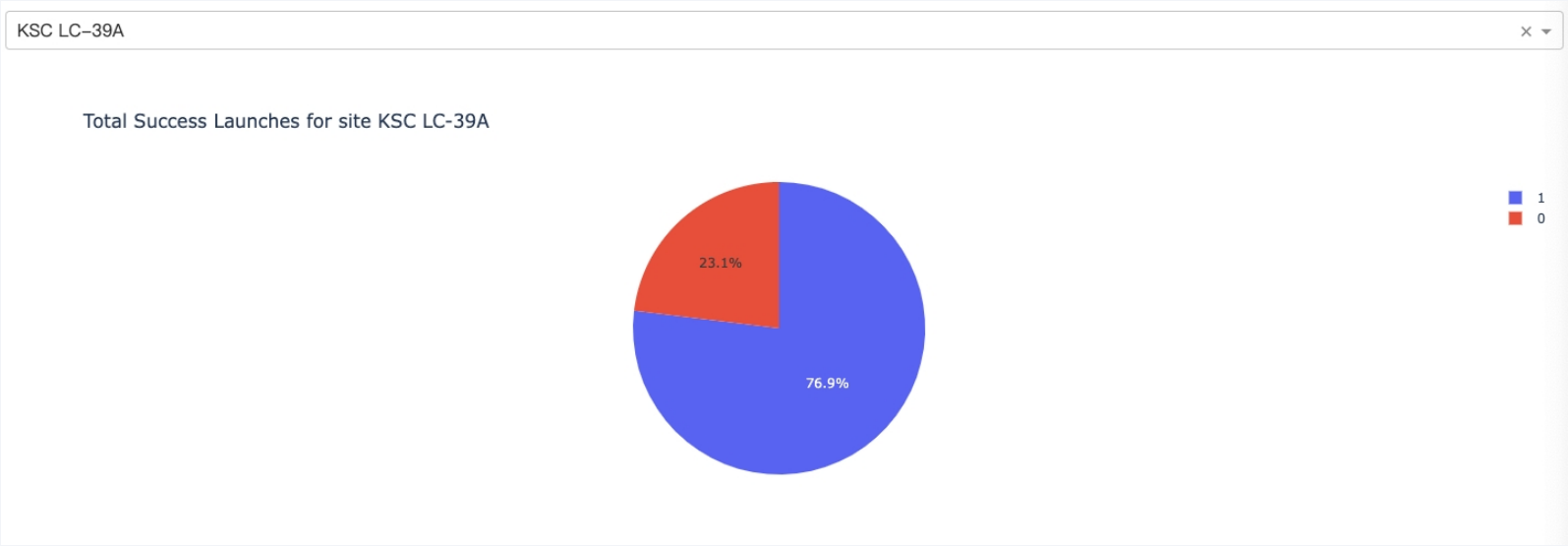
<Dashboard Screenshot 1>

- Replace <Dashboard screenshot 1> title with an appropriate title
- Show the screenshot of launch success count for all sites, in a piechart
- Explain the important elements and findings on the screenshot
  - KSC LC-39A has the most success counts in all of the launch sites.



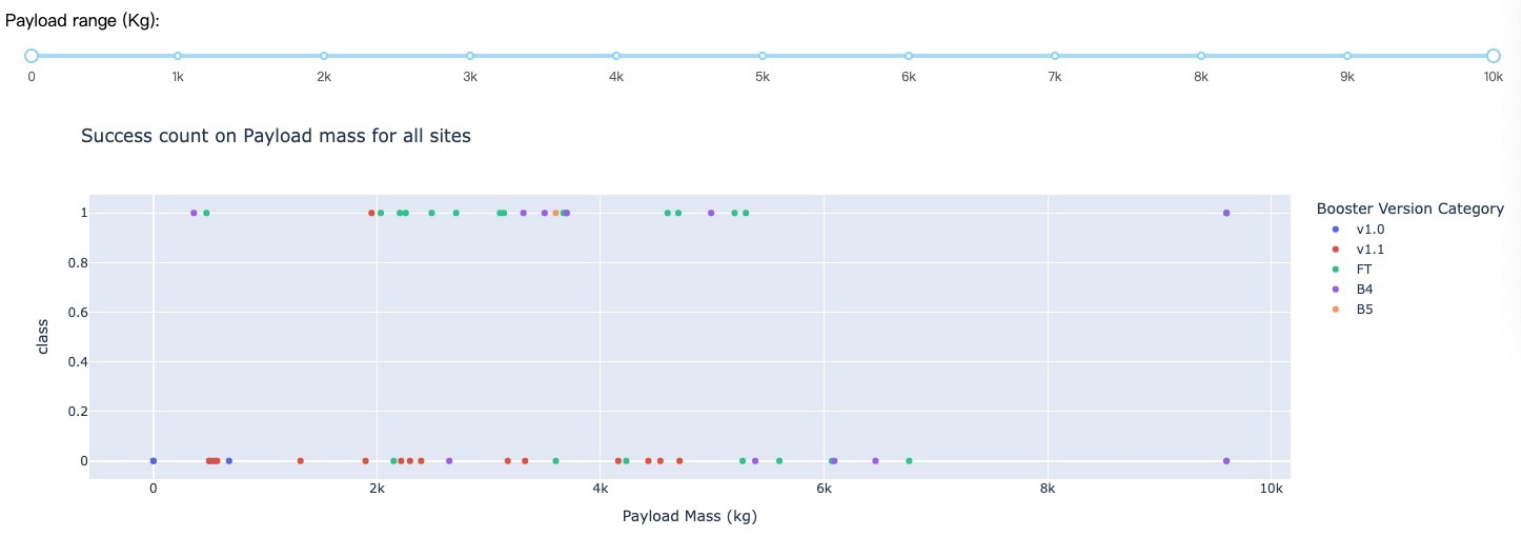
<Dashboard Screenshot 2>

- Replace <Dashboard screenshot 2> title with an appropriate title
- Show the screenshot of the piechart for the launch site with highest launch success ratio
- Explain the important elements and findings on the screenshot
  - The success rate is 76.9% for site KSC LC-39A.



<Dashboard Screenshot 3>

- Replace <Dashboard screenshot 3> title with an appropriate title
- 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.
  - The booster version of FT with payload mass between 2k to 6k has the largest success rate.

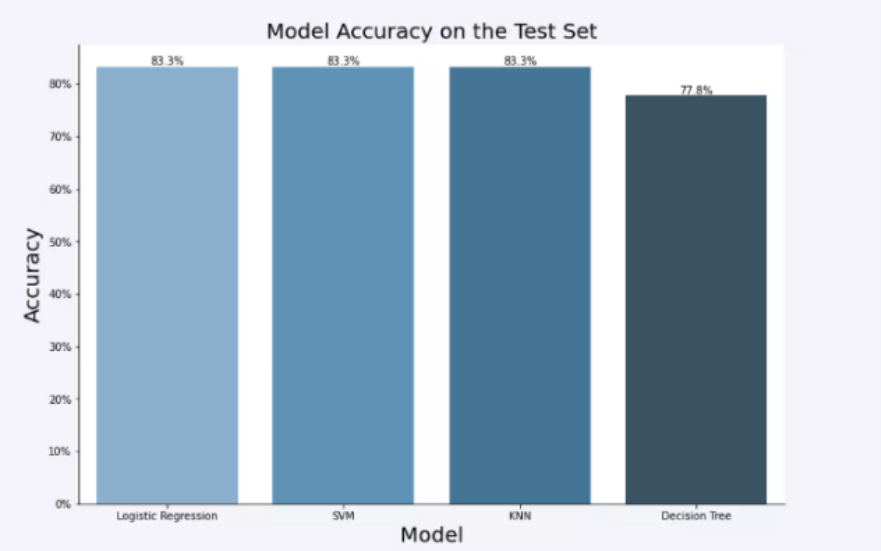


Section 5

Predictive Analysis (Classification)

Classification Accuracy

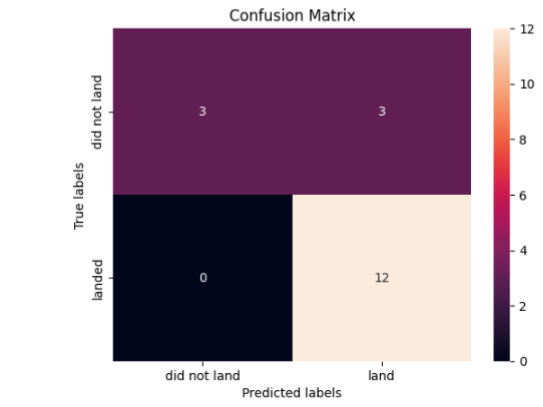
- Visualize the built model accuracy for all built classification models, in a bar chart

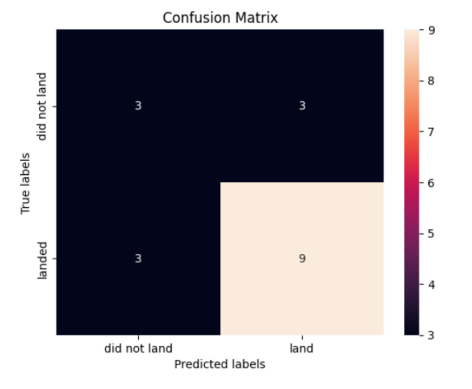


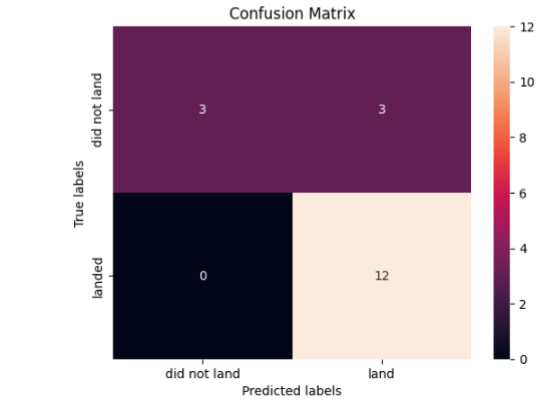
- Find which model has the highest classification accuracy
  - There are 3 models, logistic regression, svm, and knn, are the best model in this case.

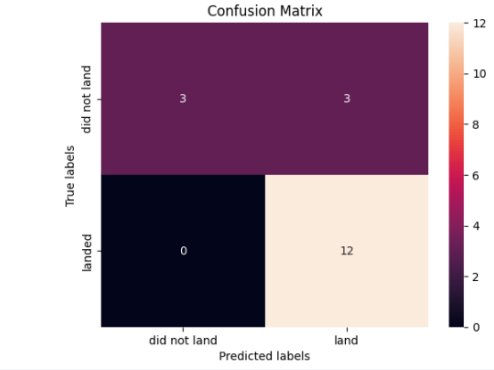
Confusion Matrix

- Show the confusion matrix of the best performing model with an explanation
  - the accuracy rate=(3+12)/(3+3+12)=0. 8333333333333334









Conclusions

- There are 3 models, logistic regression, svm, and knn, are the best model in this case.
- Low weighted payloads perform better than the heavier payloads.
- KSC LC 39A had the most successful launches from all the sites.

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
  - https://github.com/kwlee1201/Coursera_IBM

Thank you!