IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sukanya Guha Roy
12th May 2024

Outline

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

Executive Summary

- **Summary of methodologies**
- **Data Handling:** Combined data collection, cleaning, and transformation for robust dataset preparation.
- **Exploratory Analysis:** Utilized statistical techniques and visual analytics for in-depth data exploration.
- **Predictive Modeling:** Employed advanced algorithms for predictive analysis, ensuring thorough model evaluation.
- **Interactive Tools:** Implemented Folium for geographic insights and Plotly Dash for dynamic visual dashboards.

Executive Summary

- **Summary of all results**
- **Visual Discoveries:** Uncovered key insights and patterns through comprehensive visualizations.
- **SQL Insights:** Derived critical findings and trends from complex SQL queries.
- **Geospatial Trends:** Identified significant geographical patterns using interactive Folium maps.
- **Dashboard Interactivity:** Implemented user interactions via Plotly Dash.
- **Predictive Accuracy:** Achieved substantial predictive performance, highlighting significant classifications.
- **Conclusion:** Presented findings, implications for future exploration.

Introduction

Falcon 9: Revolutionizing Space Launches

- Innovation in Focus: SpaceX's Falcon 9 rocket, leading the charge in reducing space launch costs.
- The Power of Reusability: Emphasizing the cost-saving impact of reusing the first stage.

Predictive Challenge: Landing Success

- **Crucial Question:** Can we accurately predict the successful landing of Falcon 9's first stage?
- **Why It Matters:** Insights into financial implications and competitive dynamics in the space industry. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

Section 1

Methodology

Methodology

Executive Summary

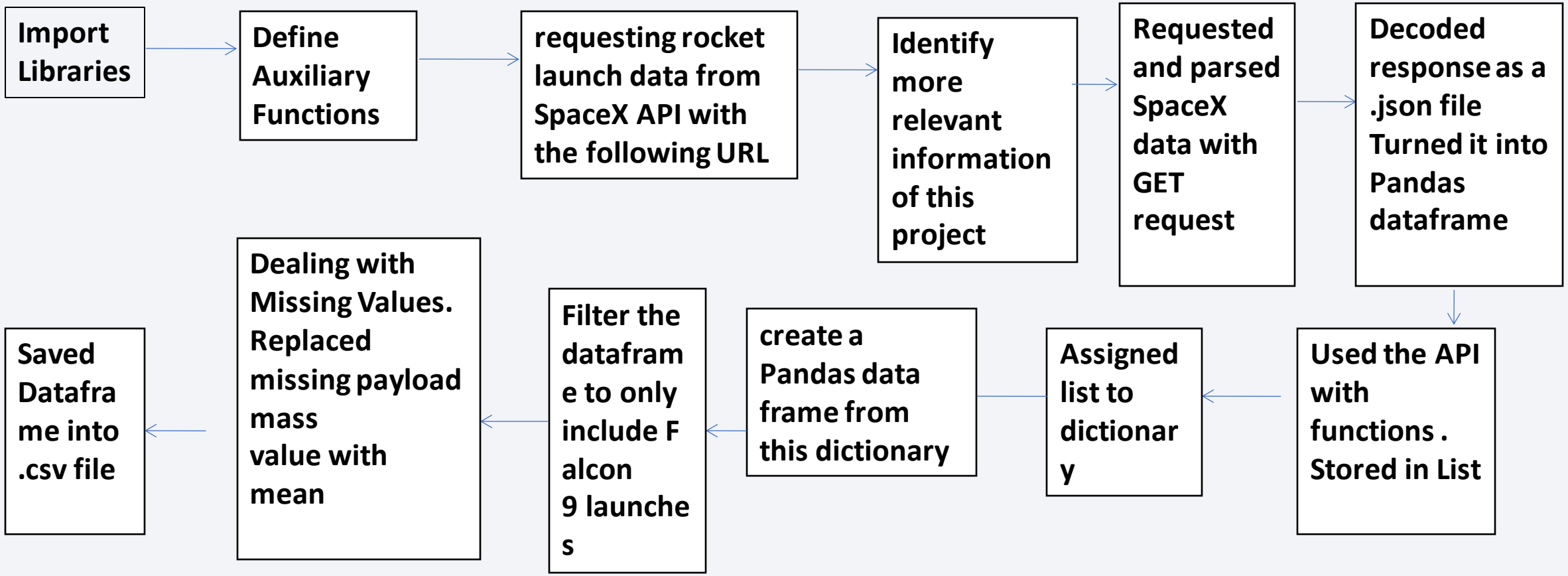
- Data collection methodology:
  - **Data was collected from open source SpaceX REST API**
  - **Webscraping Falcon9 launch data in Wikipedia**
- Data wrangling
  - **Normalization and Parsing**
  - **Filtering and Cleaning**
  - **Handling Nulls**
- Exploratory data analysis (EDA) using visualization and SQL Query
- Interactive visual analytics: Launch sites using Folium and Success rates with Plotly Dash.
- Perform predictive analysis using classification models
- **Model Building**
- **Tuning and Evaluation**

Data Collection

- **API Utilisation:** The SpaceX launch data is collected using the SpaceX REST API. This API gives us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.

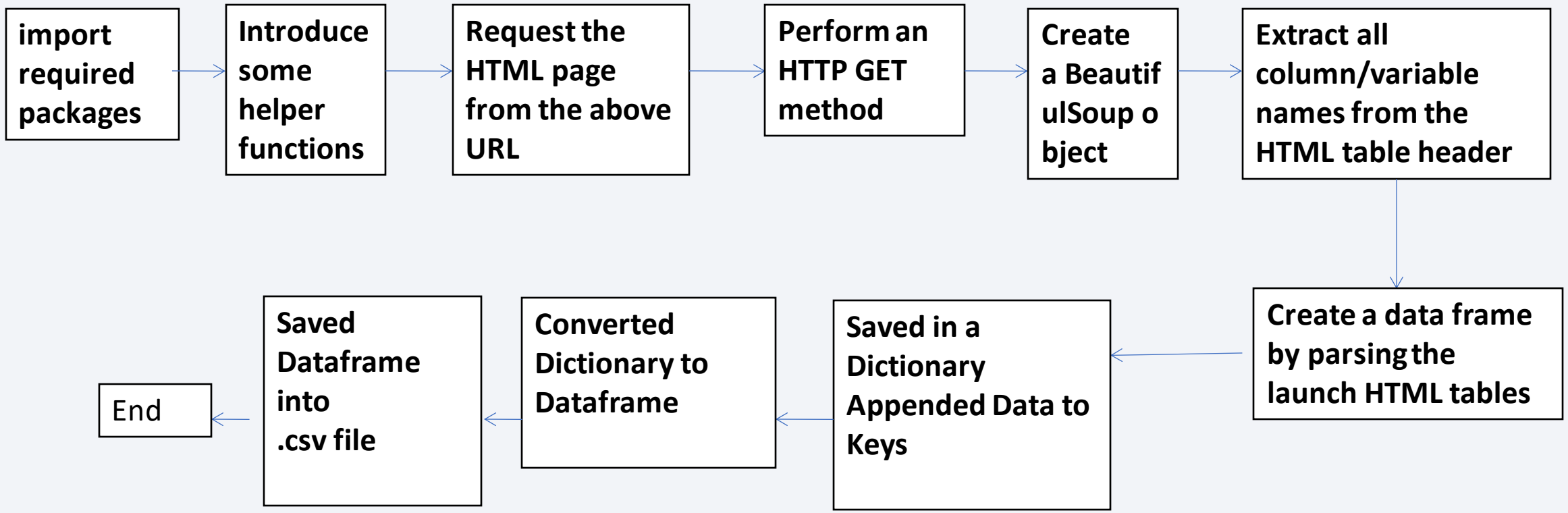
- **Web Scraping:** Implemented Python BeautifulSoup package to scrape HTML tables containing Falcon 9 launch records from relevant Wiki pages.

Data Collection – SpaceX API



GitHub URL : https://github.com/SukanyaGuhaRoy/applied-data-science-capstone-project-spacex-ibm/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection – Web scraping



GitHub URL: https://github.com/SukanyaGuhaRoy/applied-data-science-capstone-project-spacex-ibm/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

Overview of Data Wrangling Process:

From SpaceX API:

• Identified missing values.

Replaced missing "PayloadMass" value with mean value.

SpaceX dataset Further Data wrangling

Import Libraries and Define Auxiliary Functions

• Load Space X dataset

• Identify and calculate the percentage of the missing values in each attribute using isNull()

• Identify which columns are numerical and categorical

•Create a landing outcome label from Outcome column

	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 URL: https://github.com/SukanyaGuhaRoy/applied-data-science-capstone-project-spacex-ibm/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

Perform Exploratory Data Analysis and Feature Engineering using Pandas and Matplotlib.

1) Scatter plot of Flight Number vs. Launch Site

Findings: Different launch sites have varying success rates.

2) Scatter plot of Payload vs. Launch Site

Findings: For the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000)

3) Bar chart for the success rate of each orbit type

Findings: Identification of orbits with higher success rates.

4) Scatter point of Flight number vs. Orbit type

Findings: 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.

5) Scatter point of payload vs. orbit type

Findings:

With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) areboth there here.

6) Line chart of yearly average success rate

Findings: We can observe that the success rate since 2013 kept increasing till 2020.

Github URL: https://github.com/SukanyaGuhaRoy/applied-data-science-capstone-project-spacex-ibm/blob/main/edadataviz.ipynb

EDA with SQL

Github URL https://github.com/SukanyaGuhaRoy/applied-data-science-capstone-project-spacex-ibm/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

- **Analyzing Launch Sites: Identified the names of the unique launch sites in the space mission.**

  Key Finding: Multiple launch sites with varying characteristics

- **Extracting Launch Site Records: Display records where launch sites begin with the string 'CCA'.**
- **Payload Analysis: Display the total payload mass carried by boosters launched by NASA (CRS)**

  Insights: Helps in understanding payload capacity for specific customers.

- **Booster version payload Analysis: Display average payload mass carried by booster version F9 v1.1**

  Insights: Provides an average benchmark for payload capabilities of the booster.

- **First successful landing outcome: List the date when the first successful landing outcome in ground pad was achieved.**

  Key findings: Marks an important achievement in SpaceX's journey

- **List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.**
- **List the total number of successful and failure mission outcomes**
- **List the names of the booster_versions which have carried the maximum payload mass. Use a subquery**

  Key Insights: Understanding the maximum payload capacity achieved.

- **Detail Drone Ship Failures Analysis in year 2015:** Displayed specific records with failure landing outcomes in drone ships for the year 2015.

  Findings: Allows a focused analysis on a particular year and failure type.

- **Landing Outcomes Landing(date 2010-06-04 and 2017-03-20):** Ranked Landing Outcomes: Ranked the count of different landing outcomes between specific dates.

  Analysis: Provides a descending order view of landing outcomes

Build an Interactive Map with Folium

Summary of Map Objects Added to the Folium Map:

- **Markers:** Placed at various locations to represent specific points of interest or data points.
- **Circles:** Used to visualize a certain radius or range around specific locations.
- **Polylines (Lines):** Connects different markers, representing pathways, routes, or boundaries.

Reason for Adding These Objects:

- **Markers:** To highlight and provide information about specific locations, such as schools, restaurants, or landmarks.
- **Circles:** To illustrate proximity or coverage areas from a particular point, which can be useful in understanding accessibility or reach.
- **Polylines:** To show routes, connections, or boundaries between different locations, aiding in visualizing relationships or paths

GitHub URL: https://github.com/SukanyaGuhaRoy/applied-data-science-capstone-project-spacex-ibm/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

SpaceX Launch Dashboard Summary

1. Interactive Elements:

- **Launch Site Dropdown: Filter data by launch sites.**
- **Payload Range Slider: Select payload mass range.**

2. Visualisations:

- **Success Pie Chart: Visualize launch success rates.**
- **Payload Scatter Chart: Explore payload mass vs. launch success correlation.**

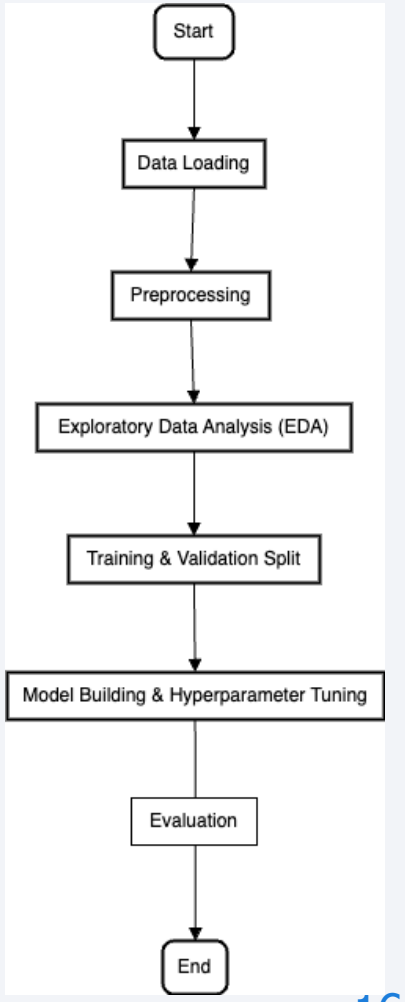
Purpose: • **Enhance user engagement and enable in-depth analysis of SpaceX launch data through interactive filters and informative visualisations.**

GitHub URL: https://github.com/SukanyaGuhaRoy/applied-data-science-capstone-project-spacex-ibm/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

Key Phrases:

- Data Loading & Preprocessing
- Exploratory Data Analysis (EDA)
- Training & Validation Split
- Model Building & Hyperparameter Tuning
- Evaluation using Confusion Matrix & Accuracy Score
- Selection of Best Performing Model



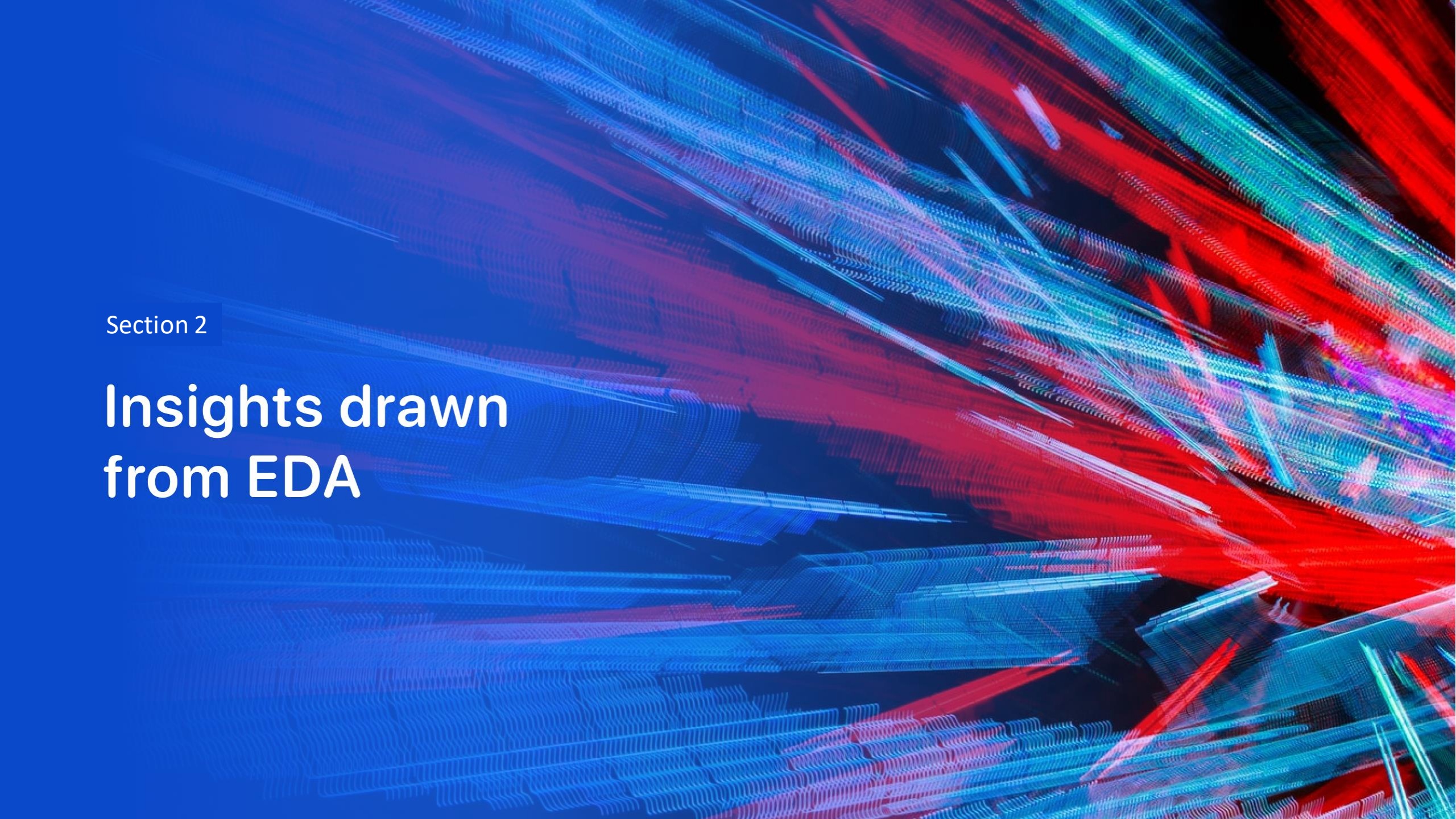
GitHub URL: https://github.com/SukanyaGuhaRoy/applied-data-science-capstone-project-spacex-ibm/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

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

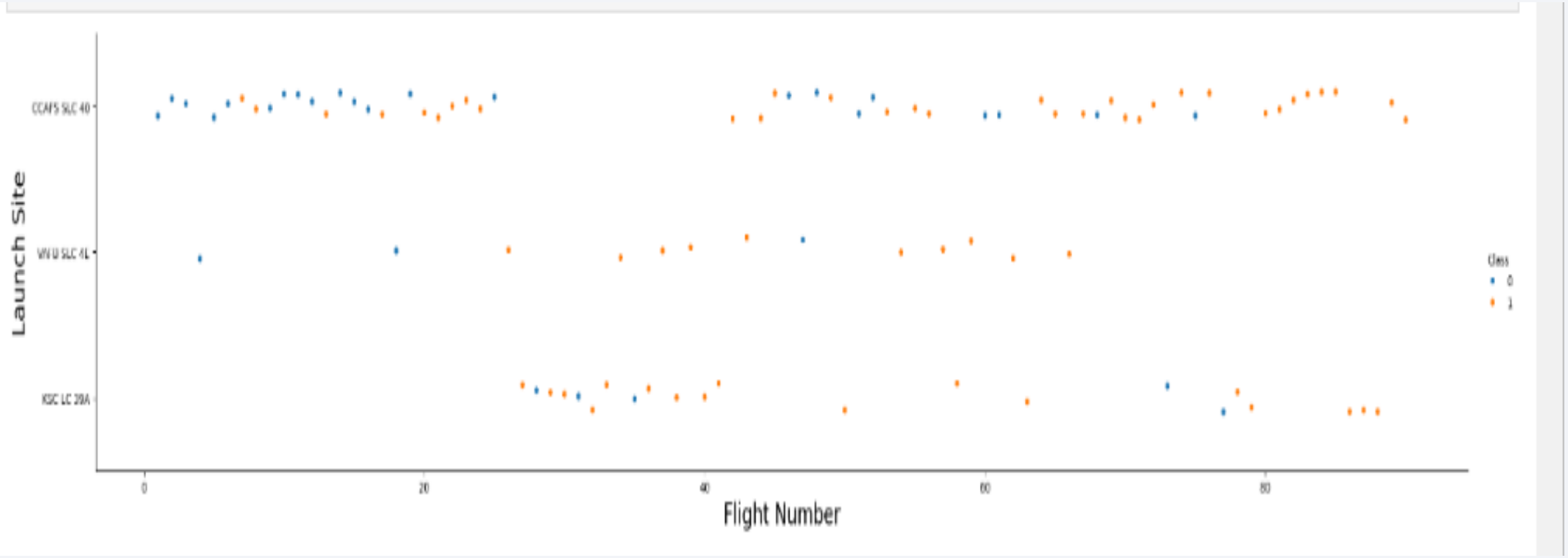
Section 2

Insights drawn from EDA



Flight Number vs. Launch Site

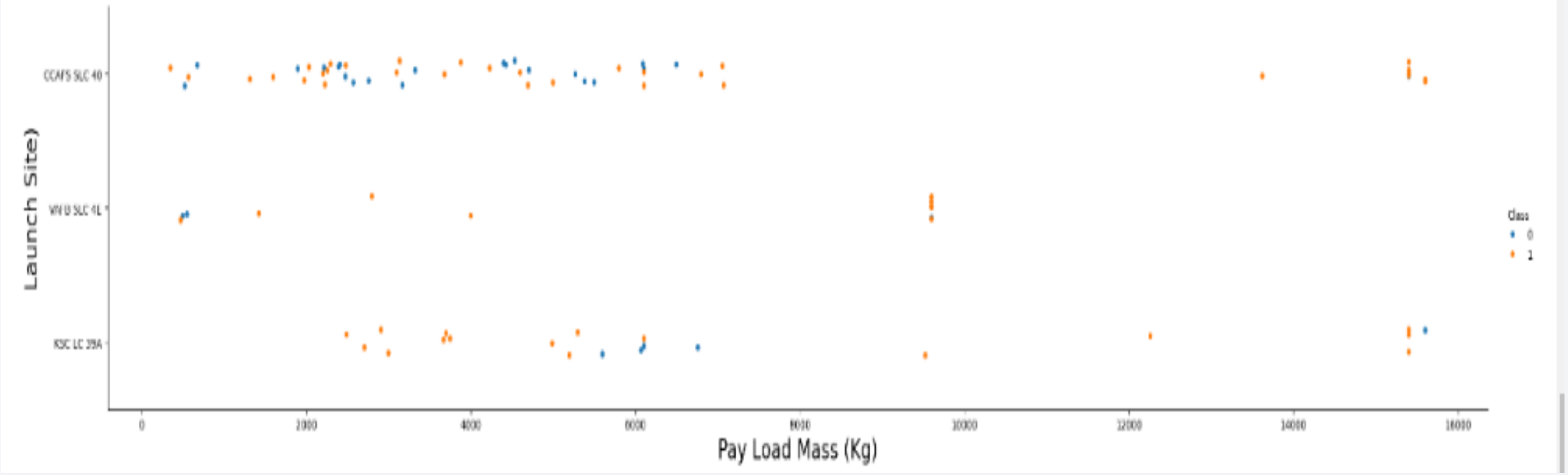
- Show a scatter plot of Flight Number vs. Launch Site



- **Findings: Different launch sites have varying success rates.**

Payload vs. Launch Site

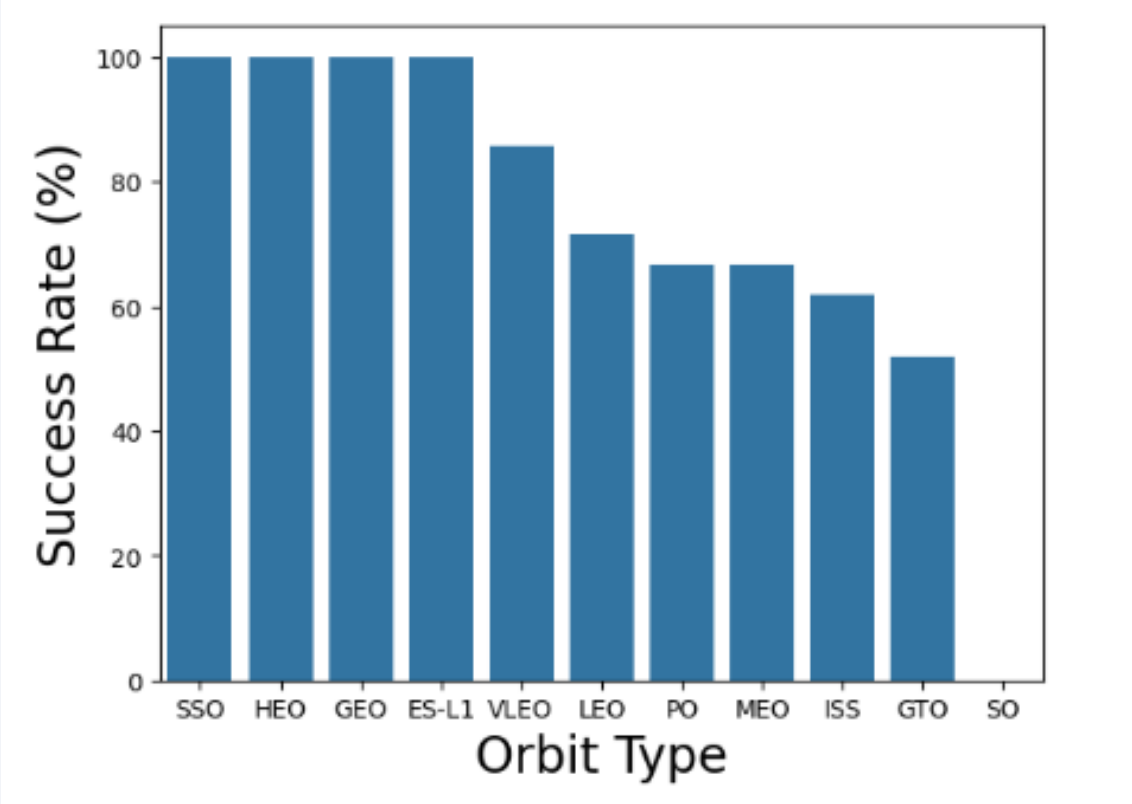
- Show a scatter plot of Payload vs. Launch Site



- **Findings: For the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000)**

Success Rate vs. Orbit Type

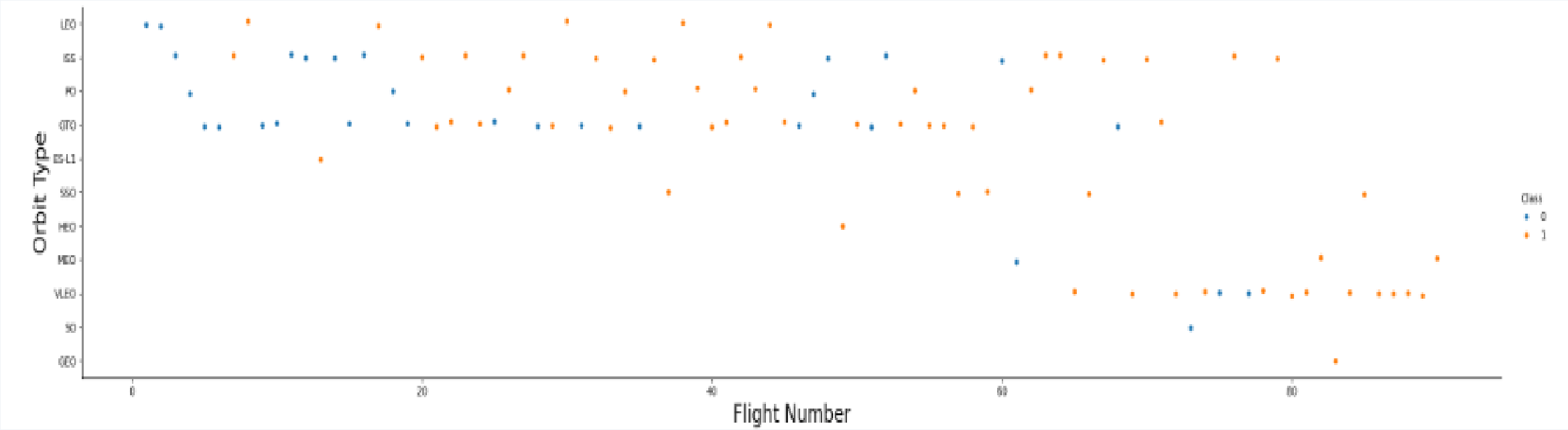
- Show a bar chart for the success rate of each orbit type



- **Findings: Identification of orbits with higher success rates.**

Flight Number vs. Orbit Type

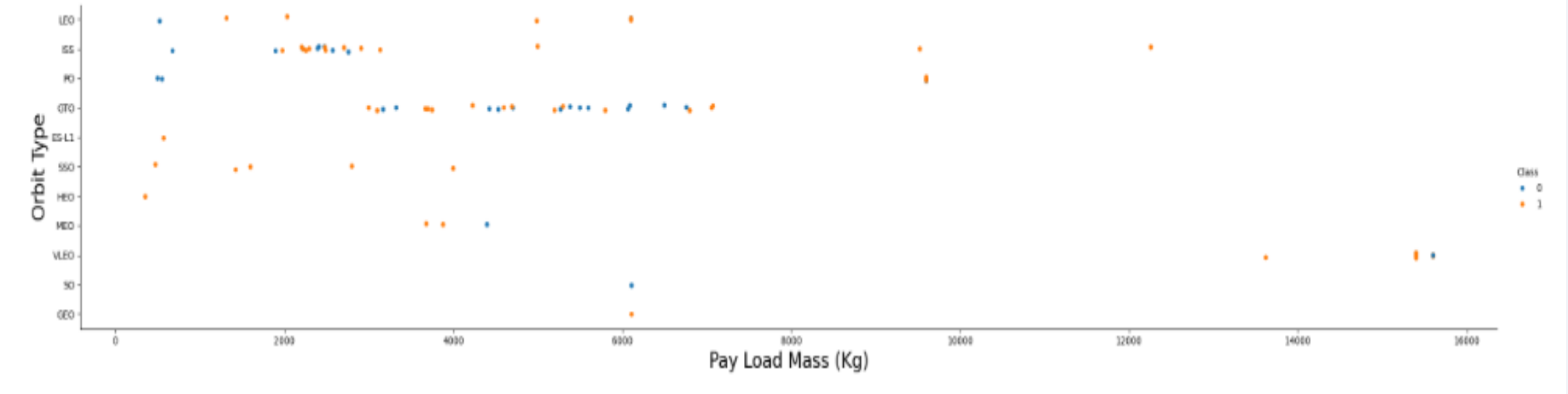
- Show a scatter point of Flight number vs. Orbit type



- **Findings: 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

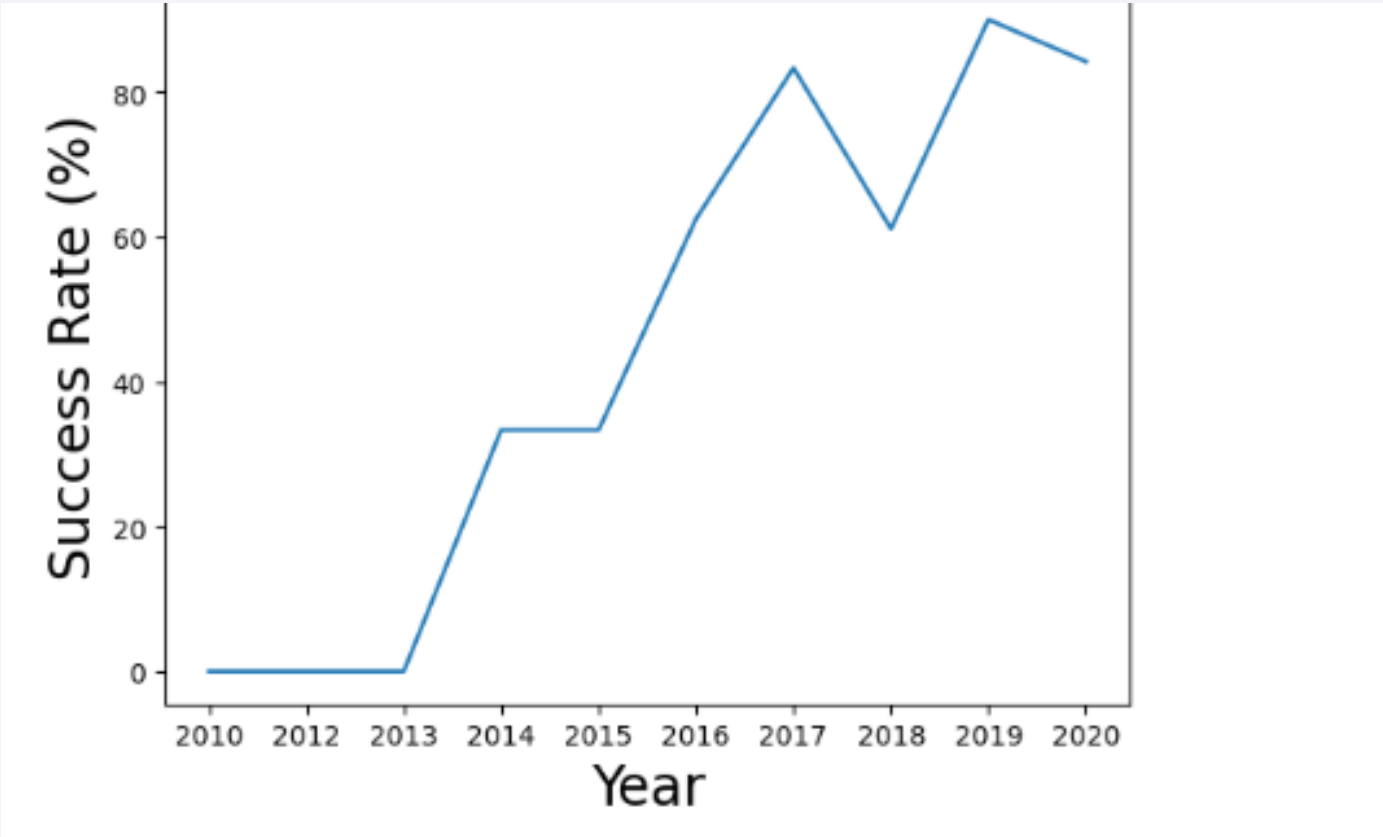
- Show a scatter point of payload vs. orbit type



- **Findings with explanations: With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.**
- **However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.**

Launch Success Yearly Trend

- Show a line chart of yearly average success rate



- **Explanations: We can observe that the sucess rate since 2013 kept increasing till 2020.**

All Launch Site Names

Find the names of the unique launch sites

```
[9]: %sql select distinct(LAUNCH_SITE) from SPACEXTBL
 * sqlite:///my_data1.db
Done.
```

[9]:

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

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' limit 5
```

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

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

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

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

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

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

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

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

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

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

```
%sql select min(DATE) from SPACEXTBL where Landing_Outcome = 'Success (ground pad)'
```

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

min(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%sql select BOOSTER_VERSION from SPACEXTBL where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000
```

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

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

Total Number of Successful and Failure Mission Outcomes

```
%sql select count(MISSION_OUTCOME) from SPACEXTBL where MISSION_OUTCOME = 'Success' or MISSION_OUTCOME = 'Failure (in flight)'
```

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

count(MISSION_OUTCOME)
99

Boosters Carried Maximum Payload

```
%sql select BOOSTER_VERSION from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

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

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

```
[40]: %sql select substr(Date, 6,2) as Month, Landing_Outcome, Booster_Version, Launch_Site   from SPACEXTABLE where Landing_Outcome = 'Failure (drone ship)' and substr(Date,0,5) = '2015'
```

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

[40]:

Month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

```
#%sql select * from SPACEXTBL where Landing_Outcome like 'Success%' and (DATE between '2010-06-04' and '2017-03-20') order by date desc
%sql select Date, Landing_Outcome, count(*) as 'Count' from SPACEXTABLE where Date between '2010-06-04' and '2017-03-20' group by Landing_Outcome order by Count desc
```

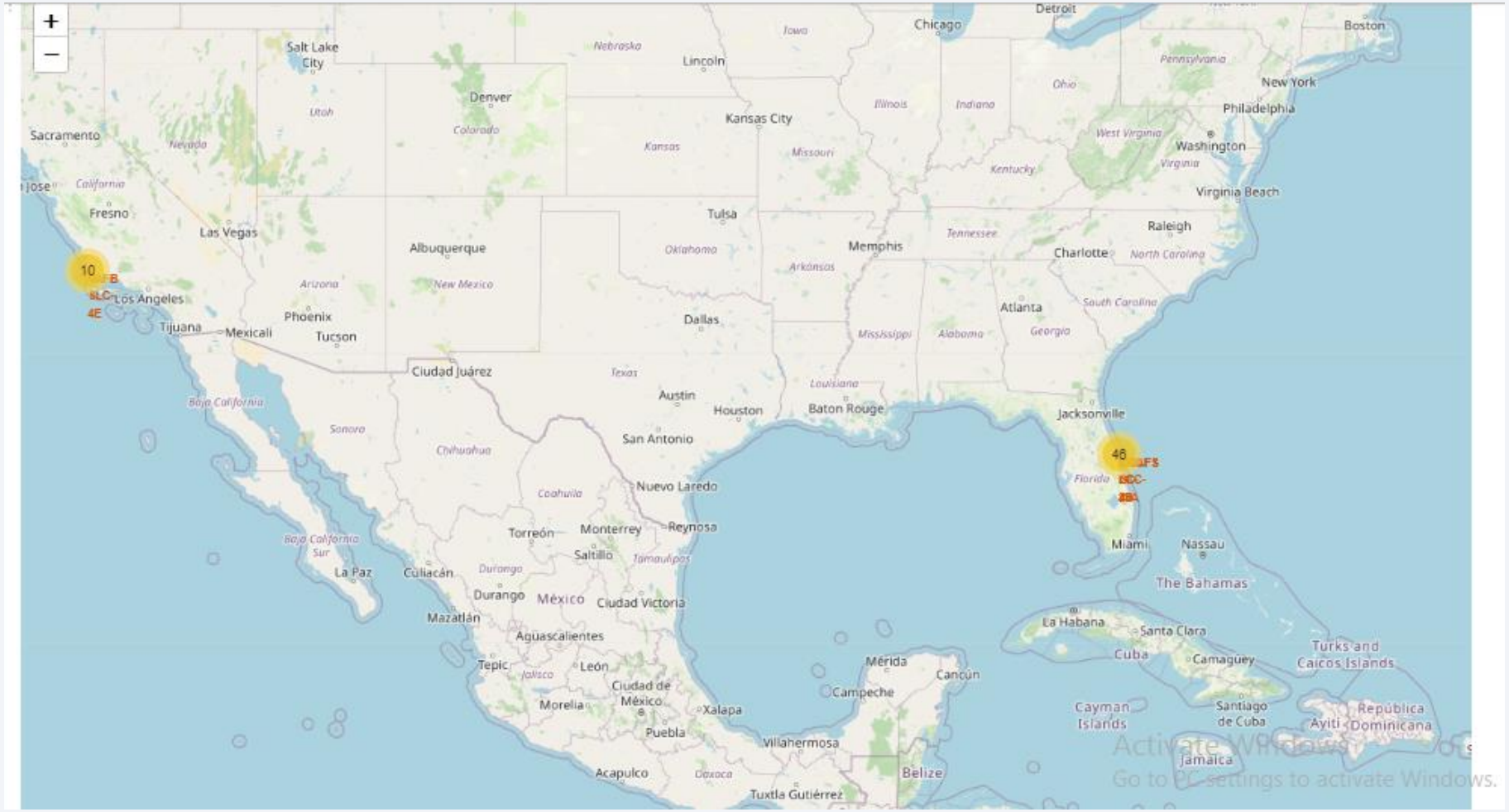
```
 * sqlite:///my_data1.db
Done.
```

Date	Landing_Outcome	Count
2012-05-22	No attempt	10
2016-04-08	Success (drone ship)	5
2015-01-10	Failure (drone ship)	5
2015-12-22	Success (ground pad)	3
2014-04-18	Controlled (ocean)	3
2013-09-29	Uncontrolled (ocean)	2
2010-06-04	Failure (parachute)	2
2015-06-28	Precluded (drone ship)	1

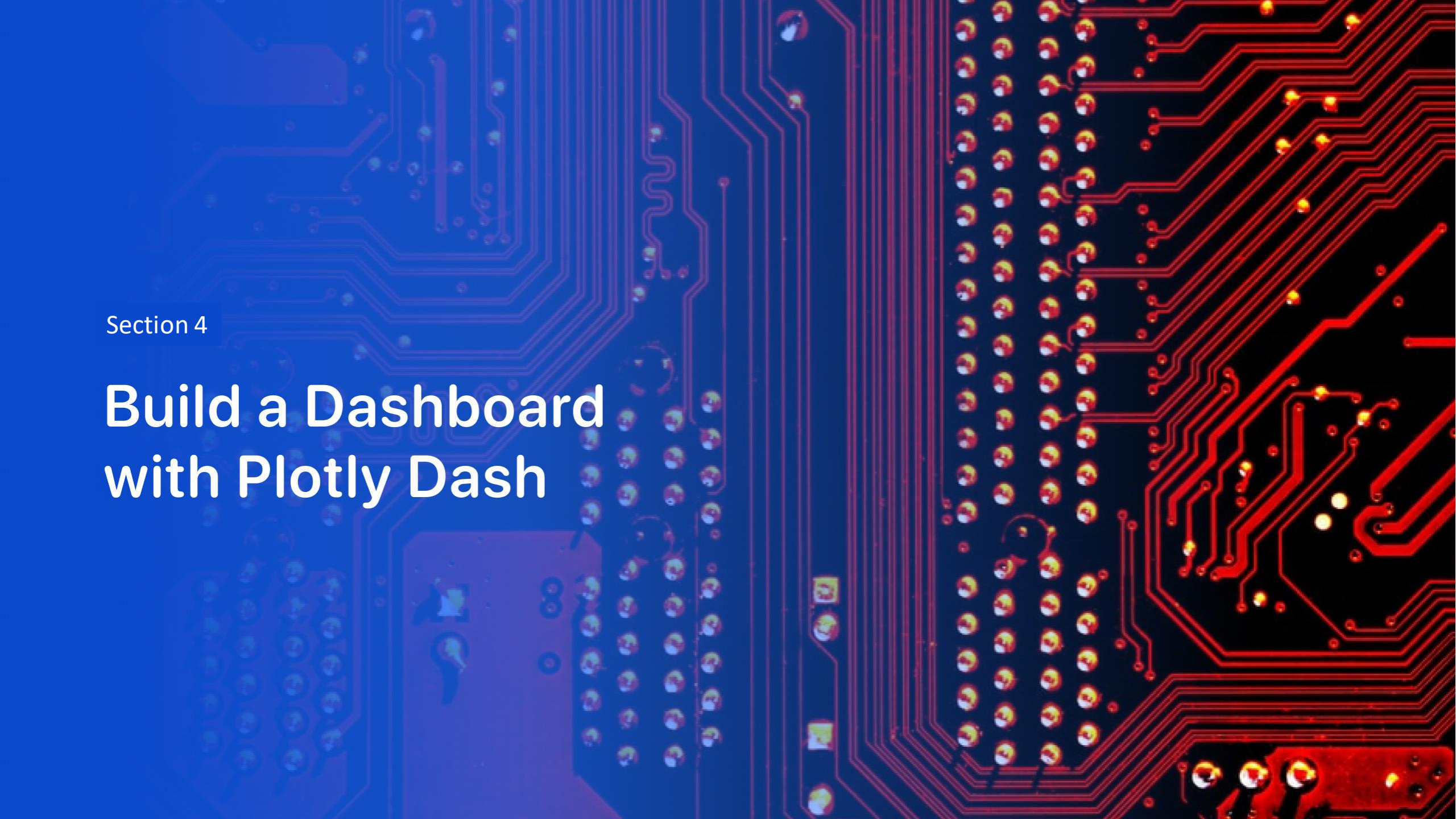

Section 3
Launch Sites
Proximities Analysis

Mark all launch sites on a map

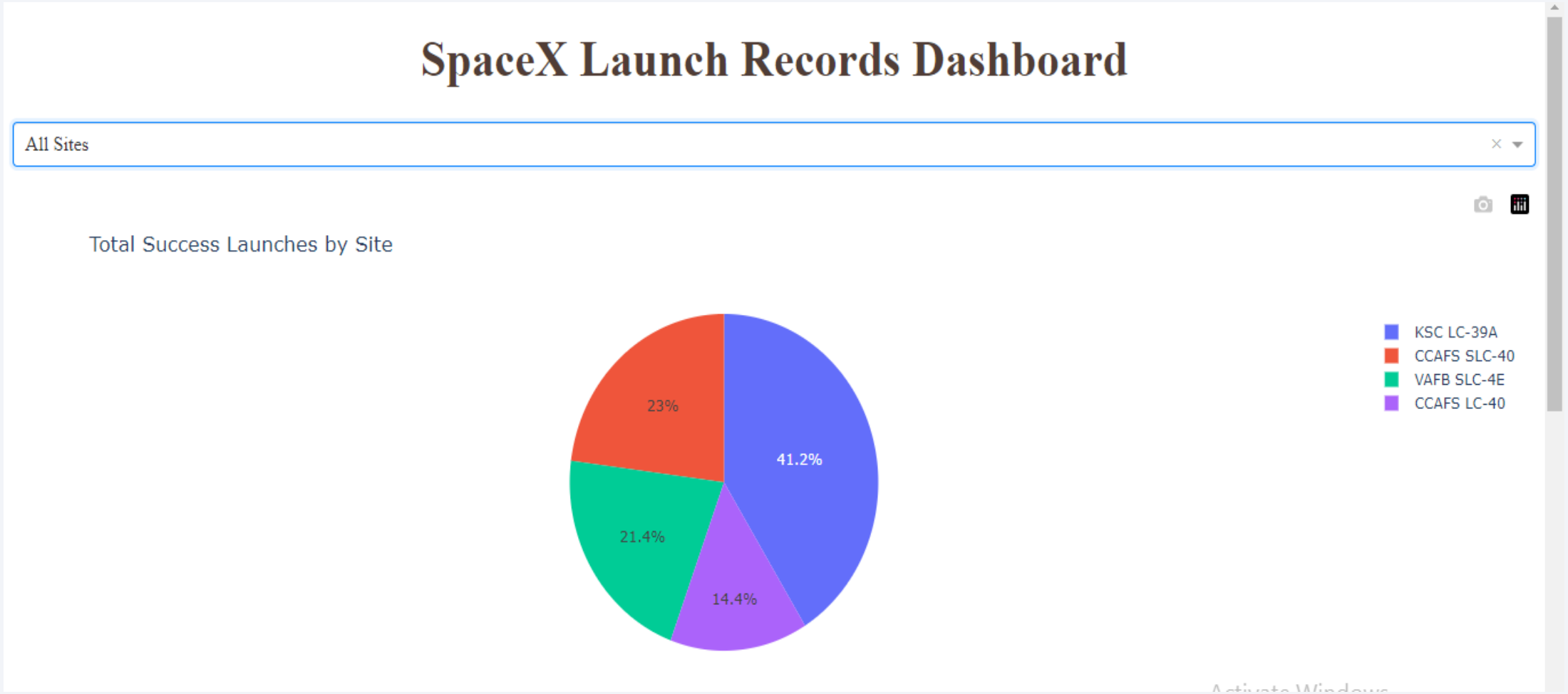
color-labeled launch outcomes



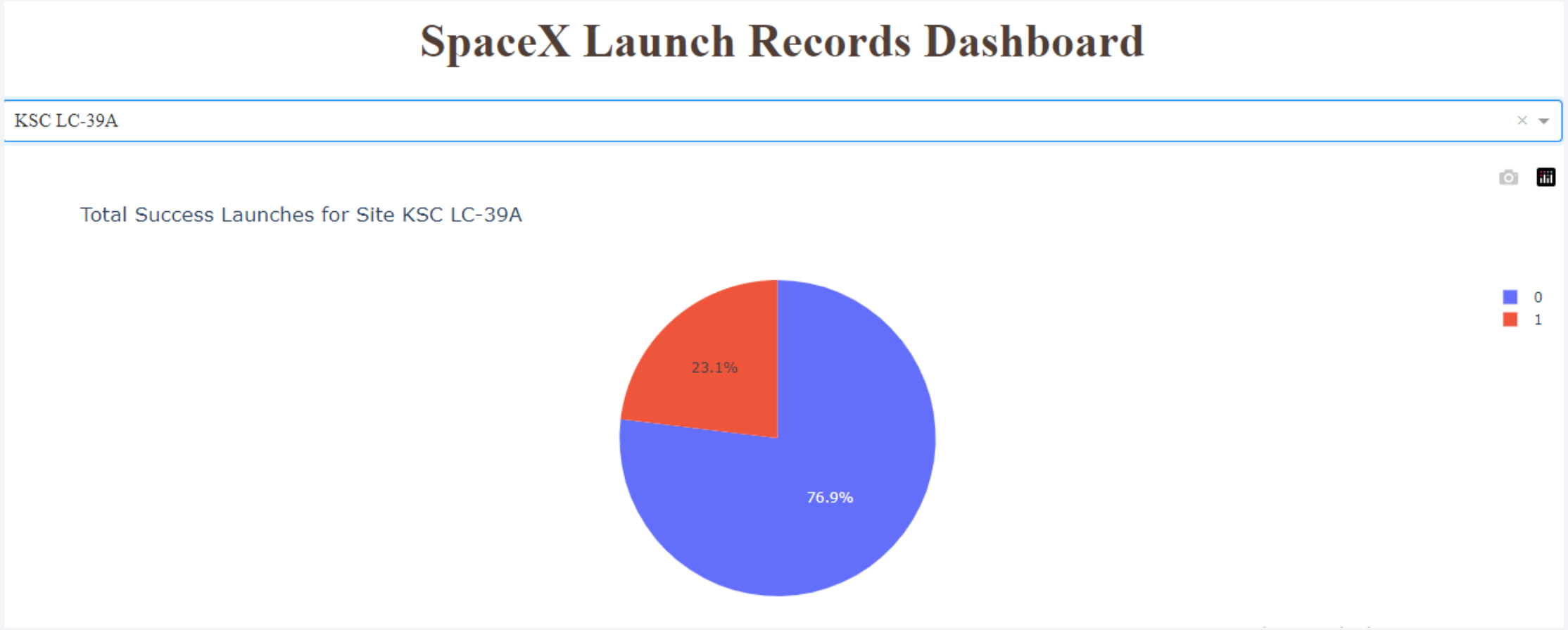
selected launch site to its proximities


Section 4
Build a Dashboard with Plotly Dash

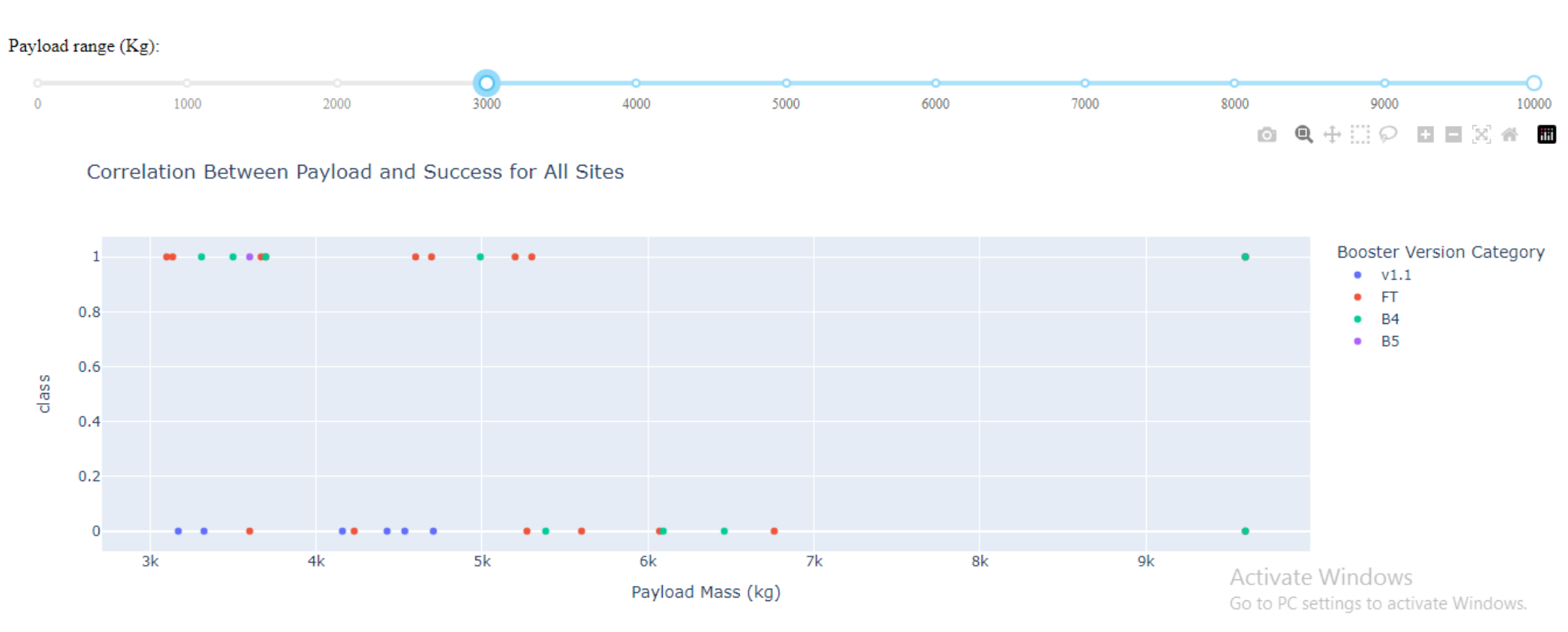
launch success count for all sites



launch site with highest launch success



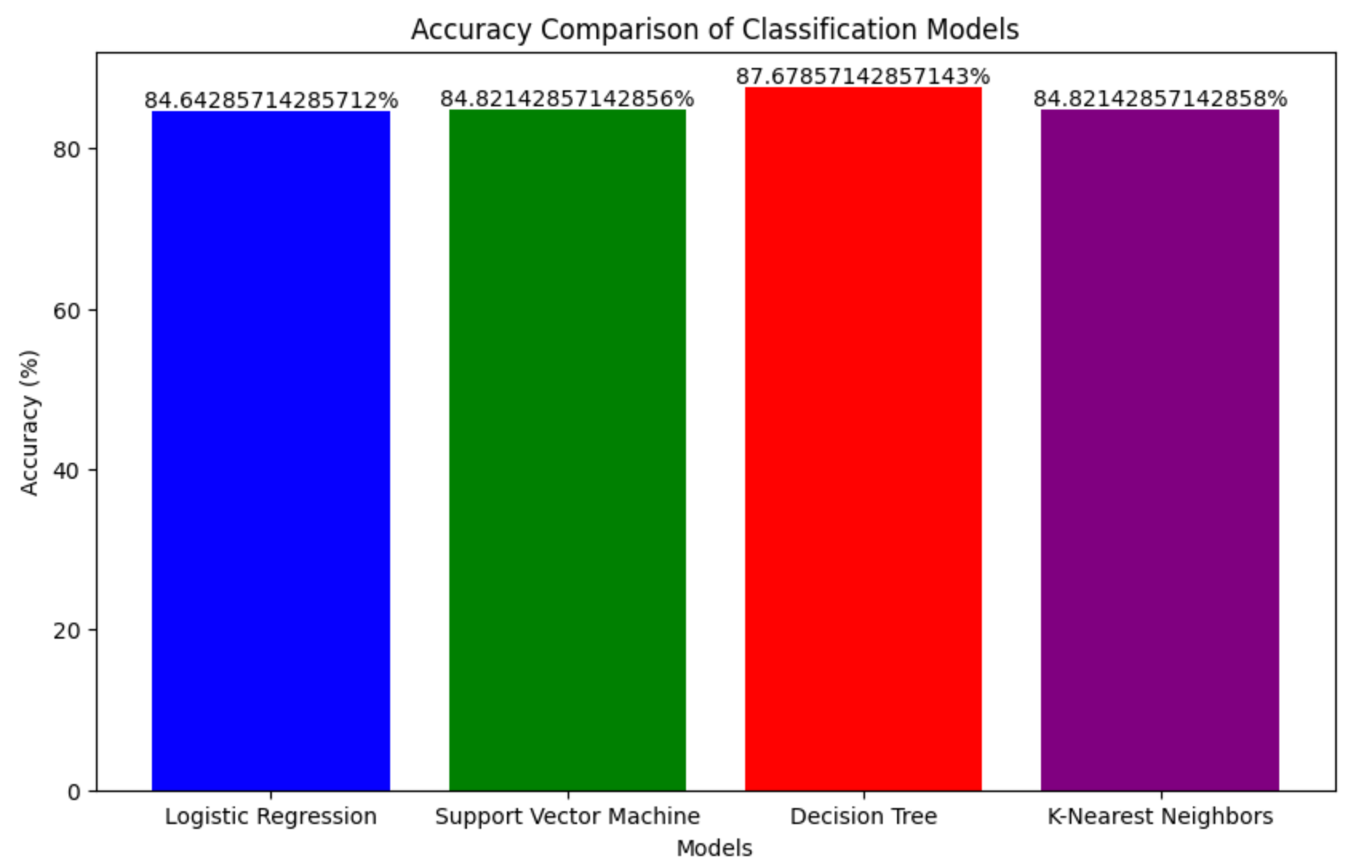
Payload vs. Launch Outcome



Section 5

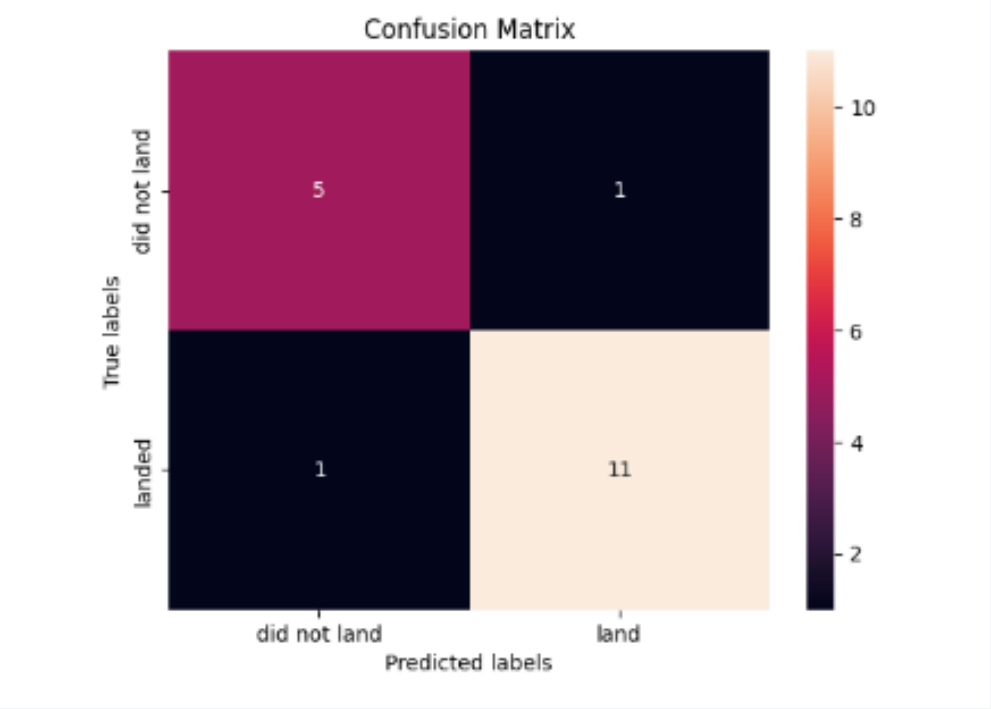
Predictive Analysis (Classification)

Classification Accuracy



Confusion Matrix

- Show the confusion matrix of the best performing model which is the Decision Tree with the highest correlation between predicted and the true label with an explanation



Conclusions

- Success rate since 2013 kept increasing till 2020.
- A supervised classification model capable of predicting launch outcome with accuracy.
- Different launch sites have varying success rates.
- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS. However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Appendix

```
# Create an app layout
app.layout = html.Div(children=[html.H1('SpaceX Launch Records Dashboard',
                                        style={'textAlign': 'center', 'color': '#503D36',
                                               'font-size': 40}),
                                # TASK 1: Add a dropdown list to enable Launch Site selection
                                # The default select value is for ALL sites
                                # dcc.Dropdown(id='site-dropdown',...)

                                dcc.Dropdown(id='site-dropdown',
                                options=[
                                    {'label': 'All Sites', 'value': 'All Sites'},
                                    {'label': 'CCAFS LC-40', 'value': 'CCAFS LC-40'},
                                    {'label': 'VAFB SLC-4E', 'value': 'VAFB SLC-4E'},
                                    {'label': 'KSC LC-39A', 'value': 'KSC LC-39A'},
                                    {'label': 'CCAFS SLC-40', 'value': 'CCAFS SLC-40'}
                                ],
                                placeholder='Select a Launch Site Here',
                                value='All Sites',
                                searchable=True
                                ),
                                html.Br(),

                                 # TASK 2: Add a pie chart to show the total successful launches count for all sites
                                # If a specific launch site was selected, show the Success vs. Failed counts for the site
                                html.Div(dcc.Graph(id='success-pie-chart')),
                                html.Br(),

                                html.P("Payload range (Kg):"),
```

Appendix

```
                                # TASK 3: Add a slider to select payload range
                                #dcc.RangeSlider(id='payload-slider',...)
                                dcc.RangeSlider(id='payload-slider',
                                min=0,
                                max=10000,
                                step=1000,
                                marks={i: '{}'.format(i) for i in range(0, 10001, 1000)},
                                value=[min_payload, max_payload]),


                                # TASK 4: Add a scatter chart to show the correlation between payload and launch success
                                html.Div(dcc.Graph(id='success-payload-scatter-chart')),
                                ])



# TASK 2:
# Add a callback function for `site-dropdown` as input, `success-pie-chart` as output
@app.callback( Output(component_id='success-pie-chart', component_property='figure'),
               Input(component_id='site-dropdown', component_property='value'))
def get_pie_chart(launch_site):
    if launch_site == 'All Sites':
        fig = px.pie(values=spacex_df.groupby('Launch Site')['class'].mean(),
                     names=spacex_df.groupby('Launch Site')['Launch Site'].first(),
                     title='Total Success Launches by Site')
    else:
        fig = px.pie(values=spacex_df[spacex_df['Launch Site']==str(launch_site)]['class'].value_counts(normalize=True),
                     names=spacex_df['class'].unique(),
                     title='Total Success Launches for Site {}'.format(launch_site))
    return(fig)
```

Appendix

```
# TASK 4:
# Add a callback function for `site-dropdown` and `payload-slider` as inputs, `success-payload-scatter-chart` as output
@app.callback( Output(component_id='success-payload-scatter-chart', component_property='figure'),
               [Input(component_id='site-dropdown', component_property='value'),
                Input(component_id='payload-slider',component_property='value')])
def get_payload_chart(launch_site, payload_mass):
    if launch_site == 'All Sites':
        fig = px.scatter(spacex_df[spacex_df['Payload Mass (kg)'].between(payload_mass[0], payload_mass[1])],
                 x="Payload Mass (kg)",
                 y="class",
                 color="Booster Version Category",
                 hover_data=['Launch Site'],
                 title='Correlation Between Payload and Success for All Sites')
    else:
        df = spacex_df[spacex_df['Launch Site']==str(launch_site)]
        fig = px.scatter(df[df['Payload Mass (kg)'].between(payload_mass[0], payload_mass[1])],
                 x="Payload Mass (kg)",
                 y="class",
                 color="Booster Version Category",
                 hover_data=['Launch Site'],
                 title='Correlation Between Payload and Success for Site {}'.format(launch_site))
    return(fig)



# Run the app
if __name__ == '__main__':
    app.run_server(port=8051)
```

GitHub URLs

- https://github.com/SukanyaGuhaRoy/applied-data-science-capstone-project-spacex-ibm.git
- https://github.com/SukanyaGuhaRoy/applied-data-science-capstone-project-spacex-ibm/blob/main/jupyter-labs-spacex-data-collection-api.ipynb
- https://github.com/SukanyaGuhaRoy/applied-data-science-capstone-project-spacex-ibm/blob/main/jupyter-labs-webscraping.ipynb
- https://github.com/SukanyaGuhaRoy/applied-data-science-capstone-project-spacex-ibm/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb
- https://github.com/SukanyaGuhaRoy/applied-data-science-capstone-project-spacex-ibm/blob/main/edadataviz.ipynb
- https://github.com/SukanyaGuhaRoy/applied-data-science-capstone-project-spacex-ibm/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb
- https://github.com/SukanyaGuhaRoy/applied-data-science-capstone-project-spacex-ibm/blob/main/lab_jupyter_launch_site_location.ipynb
- https://github.com/SukanyaGuhaRoy/applied-data-science-capstone-project-spacex-ibm/blob/main/spacex_dash_app.py
- https://github.com/SukanyaGuhaRoy/applied-data-science-capstone-project-spacex-ibm/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Thank you!