IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Unnikrishnan S
25/06/2023

Outline

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

Executive Summary

- Performed data visualization to gather data
- Used logistic regression and classification to find if a launch will be successful

Introduction

- To determine SpaceY (similar to spaceX) launch characteristics:
- To determine if a launch will be a success or not
- To determine price of each launch
- To determine if SpaceX will reuse first stage

Section 1

Methodology

Methodology

Executive Summary

- Data sets were downloaded from IBM cloud. It included details such as flight no, Date, BoosterVersion etc
- Perform data wrangling
  - Missing values were identified and the percentage calculated
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

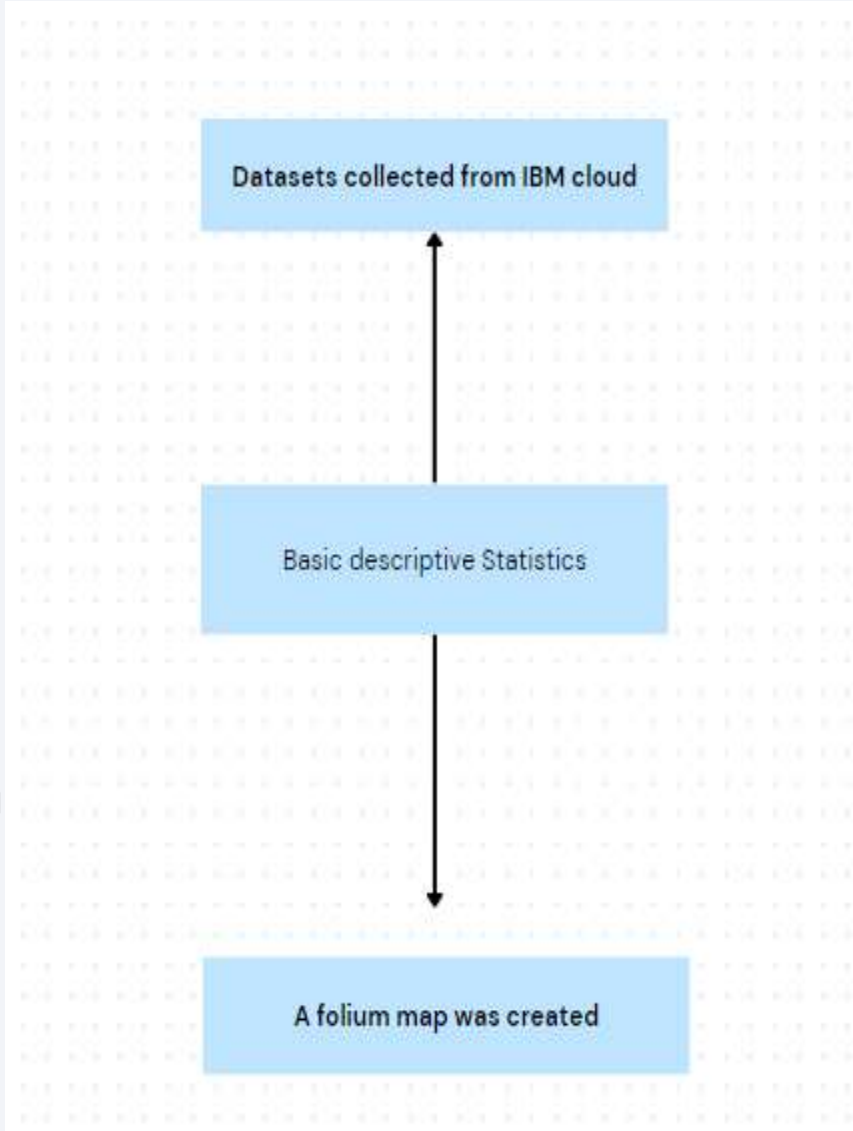
Data Collection

- The datasets were downloaded from IBM cloud

  eg: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/spacex_launch_geo.csv

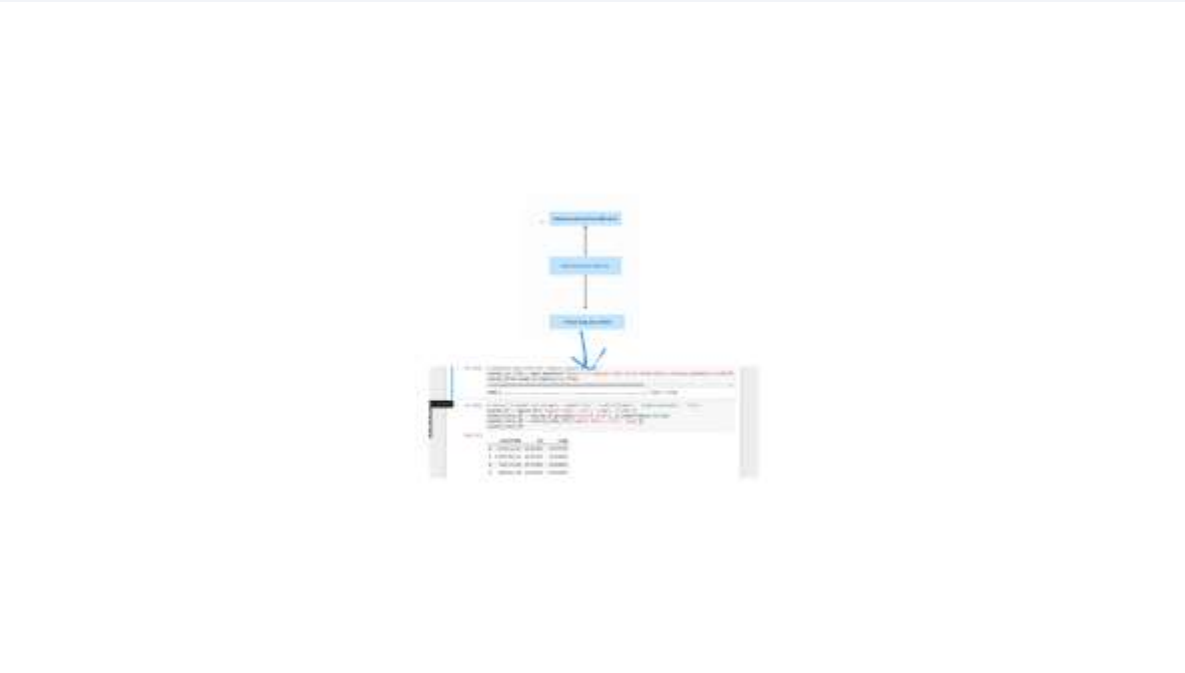
- Basic descriptive analytics

- A folium Map object(Maps) with an initial center location to be NASA Johnson Space Center at Houston Texas



Data Collection – SpaceX API

- Downloaded the required dataset using wget.download
- wget.download('https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/spacex_launch_geo.csv')
- **https://github.com/kunni9279/Coursera-IBM_Data_scientist/commit/d461c1a43323da66ef95188646405cfb2a273b64**



Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts
- Created a wikipedia URL link, downoaded all content in text format and created a beautiful soup object
- https://github.com/kunni9279/Coursera-IBM_Data_scientist/commit/d461c1a43323da66ef95188646405cfb2a273b64

Data Wrangling

- A new column in launch_sites dataframe called marker_color to store the marker colors based on the class value was created
- A color was assigned to launch outcome
- A landing class was created

- https://github.com/kunni9279/Coursera-IBM_Data_scientist/commit/d461c1a43323da66ef95188646405cfb2a273b64

EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
- Scatter plots were used to show pay load mass, flight no etc against launch sites
- Success rate of each orbit type was visualized using bar plots
- Scatter plots were used to show flight number and y axis to be the Orbit, and hue to be the class value flight no etc against launch sites
- https://github.com/kunni9279/Coursera-IBM_Data_scientist/commit/d461c1a43323da66ef95188646405cfb2a273b64

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
- Select * from df;
- Select count(flights) as count from df group by BoosterVersion;
- https://github.com/kunni9279/Coursera-IBM_Data_scientist/commit/d461c1a43323da66ef95188646405cfb2a273b64

Build an Interactive Map with Folium

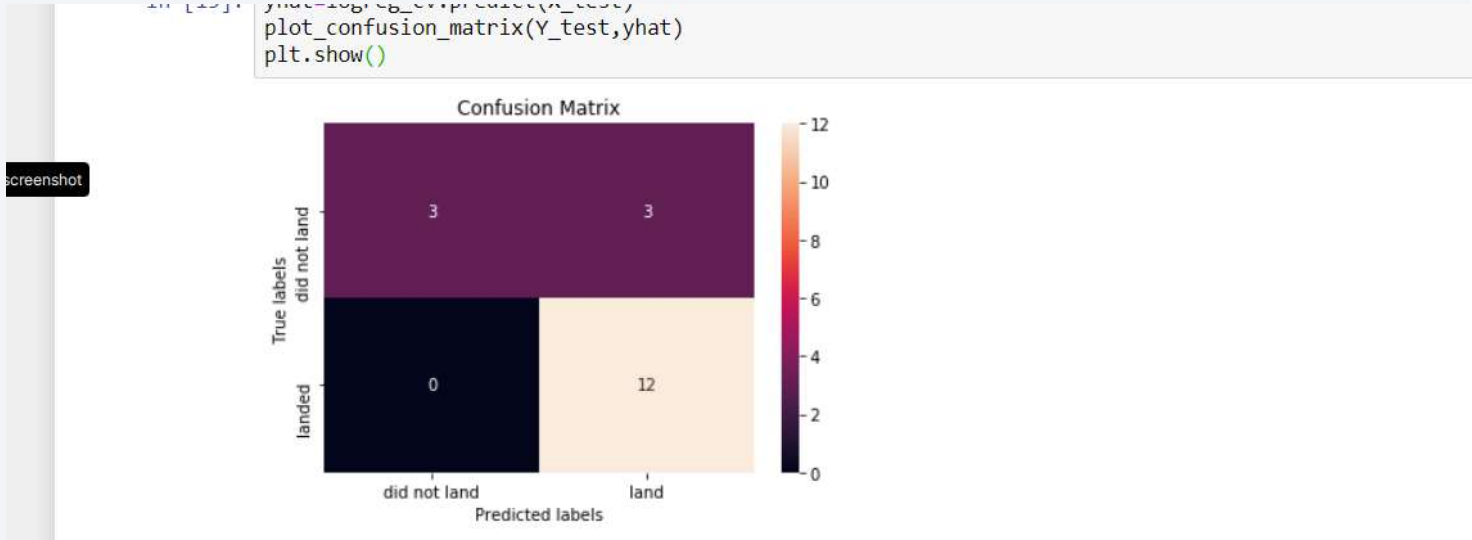
- Markers were created using folium.marker to show the distance of the coordinate
- Lines were drawn from launch sites to nearest cities
- A blue circle was created at NASA johnson space centre
- https://github.com/kunni9279/Coursera-IBM_Data_scientist/commit/d461c1a43323da66ef95188646405cfb2a273b64

Build a Dashboard with Plotly Dash

- Scatter point charts, barplots, line charts were added
- Scatter points were used to visualize relationship between Payload and LaunchSite
- Success rate of each type of orbit was visualized with barplots
- https://github.com/kunni9279/Coursera-IBM_Data_scientist/commit/d461c1a43323da66ef95188646405cfb2a273b64

Predictive Analysis (Classification)

- Logistic regression was used to predict if a landing would be successful or not
- A confusion matrix with labels landed and did not land was created
- https://github.com/kunni9279/Coursera-IBM_Data_scientist/commit/d461c1a43323da66ef95188646405cfb2a273b64



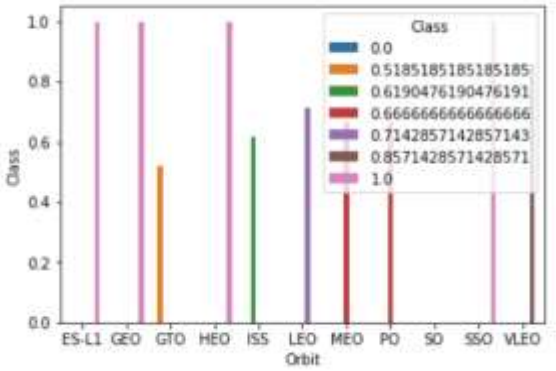
Results

- Predictive analysis results

```
In [14]: features = df[['FlightNumber', 'PayloadMass', 'Orbit', 'LaunchSite', 'Flights', 'GridFins', 'Reused',
         features.head()
```

Out[14]:

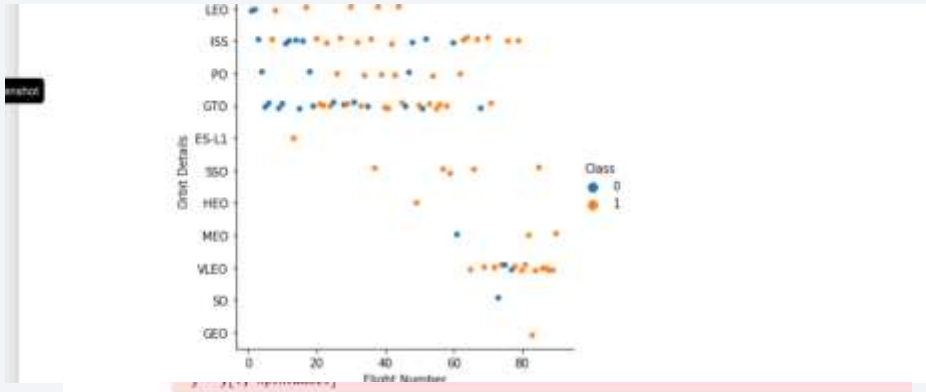
	FlightNumber	PayloadMass	Orbit	LaunchSite	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial
0	1	6104.959412	LEO	CCAFS SLC 40	1	False	False	False	NaN	1.0	0	B0003
1	2	525.000000	LEO	CCAFS SLC 40	1	False	False	False	NaN	1.0	0	B0005
2	3	677.000000	ISS	CCAFS SLC 40	1	False	False	False	NaN	1.0	0	B0007
3	4	500.000000	PO	VAFB SLC 4E	1	False	False	False	NaN	1.0	0	B1003
4	5	3170.000000	GTO	CCAFS SLC 40	1	False	False	False	NaN	1.0	0	B1004




```
SVM_accuracy=svm_cv.score(X_test, Y_test)
Logistic_Regression=logreg_cv.score(X_test, Y_test)

Report['Logistic_Reg'] = [Logistic_Regression]
Report['SVM'] = [SVM_accuracy]
Report['Decision Tree'] = [Decision_tree_accuracy]
Report['KNN'] = [knn_accuracy]



Report.transpose()
```

Out[27]:

	0
Method	Test Data Accuracy
Logistic_Reg	0.833333
SVM	0.833333
Decision Tree	0.833333
KNN	0.833333

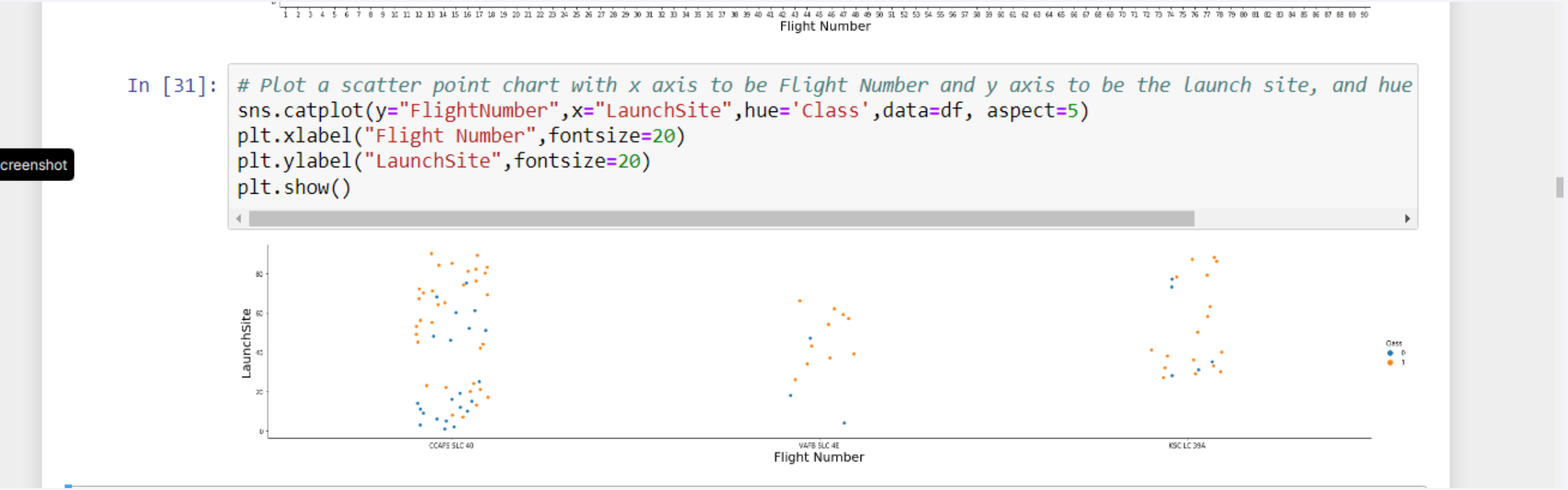


Section 2

Insights drawn from EDA

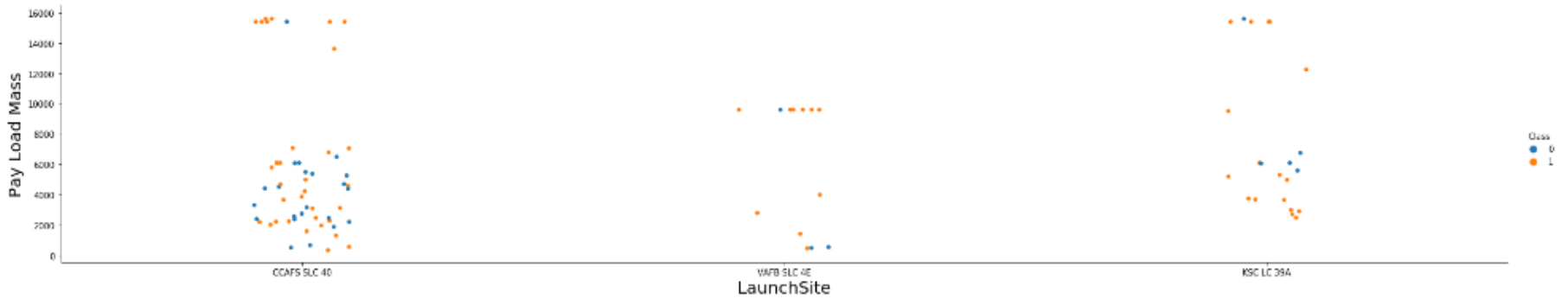
Flight Number vs. Launch Site

- There are just 3 launch sites



Payload vs. Launch Site

```
In [39]: # Plot a scatter point chart with x axis to be LaunchSite and y axis to be the Payload, and hue to be t
sns.catplot(y="PayloadMass",x="LaunchSite",hue='Class',data=df, aspect=5)
plt.xlabel("LaunchSite",fontsize=20)
plt.ylabel("Pay Load Mass",fontsize=20)
plt.show()
```

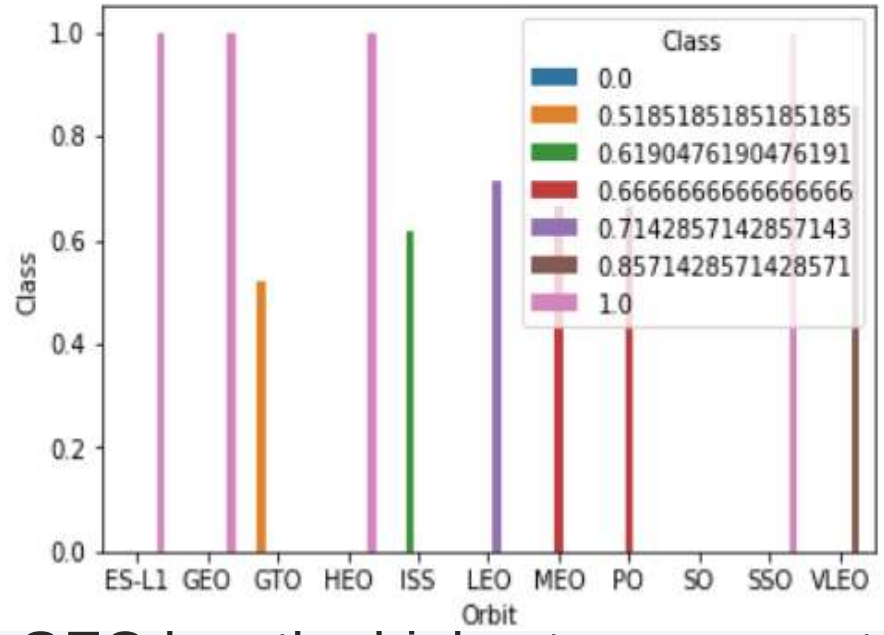


- Most launches happen in the first site

Success Rate vs. Orbit Type

```
orbit_success = df.groupby('Orbit').mean()
orbit_success.reset_index(inplace=True)
sns.barplot(x="Orbit",y="Class",data=orbit_success,hue='Class')
```

Out[8]: <matplotlib.axes._subplots.AxesSubplot at 0xf319978>

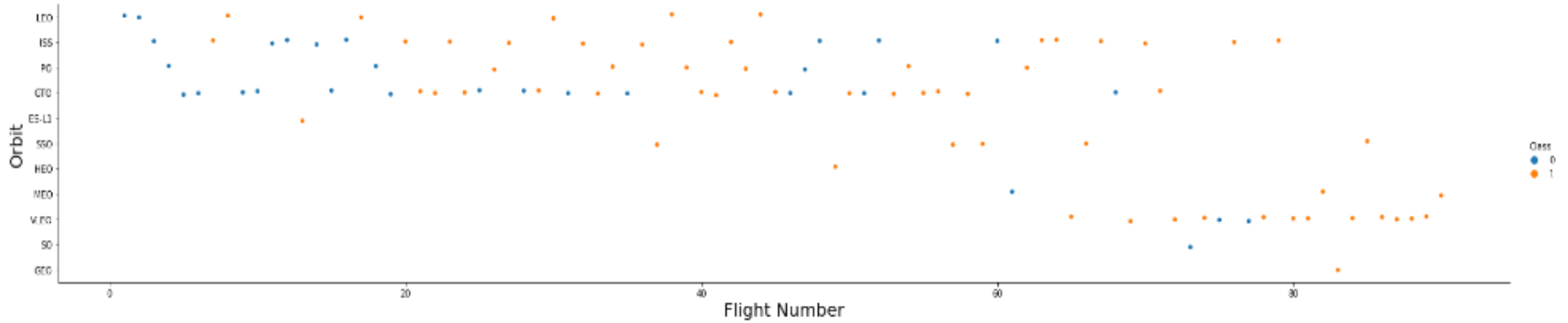


- HEO,GEO has the highest success rate

Flight Number vs. Orbit Type

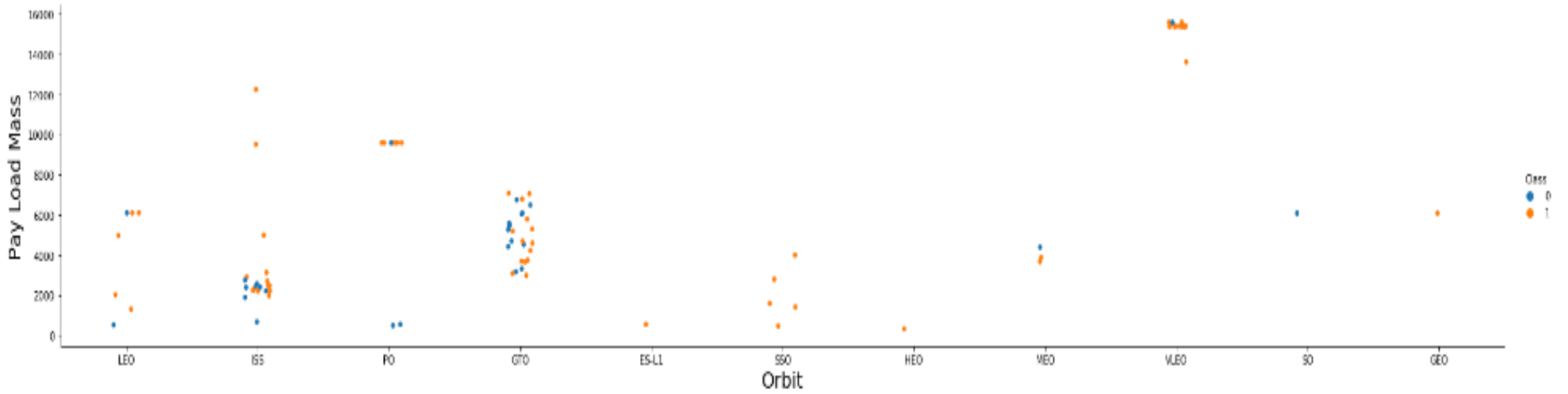
```
#We can plot out the FlightNumber vs. PayloadMassand overlay the outcome of the launch. We see that as
```

In [40]:
```
sns.catplot(y="Orbit", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```

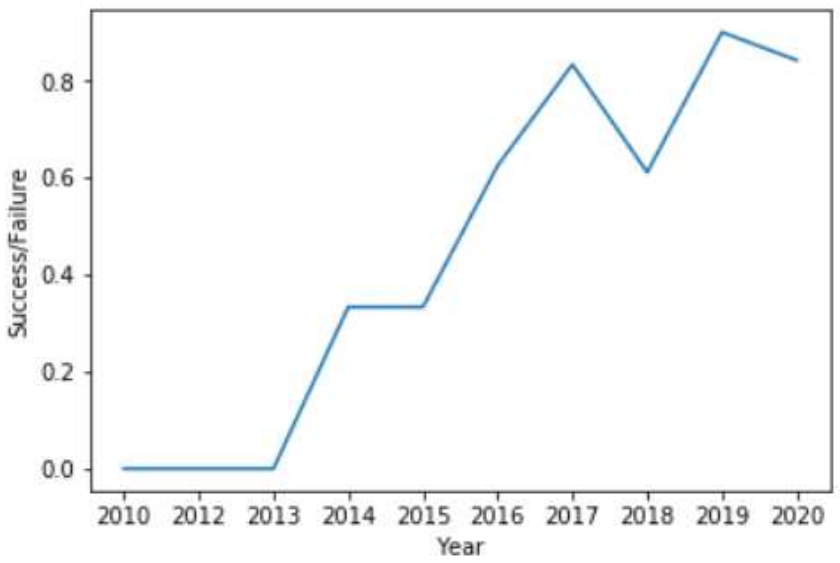


Payload vs. Orbit Type

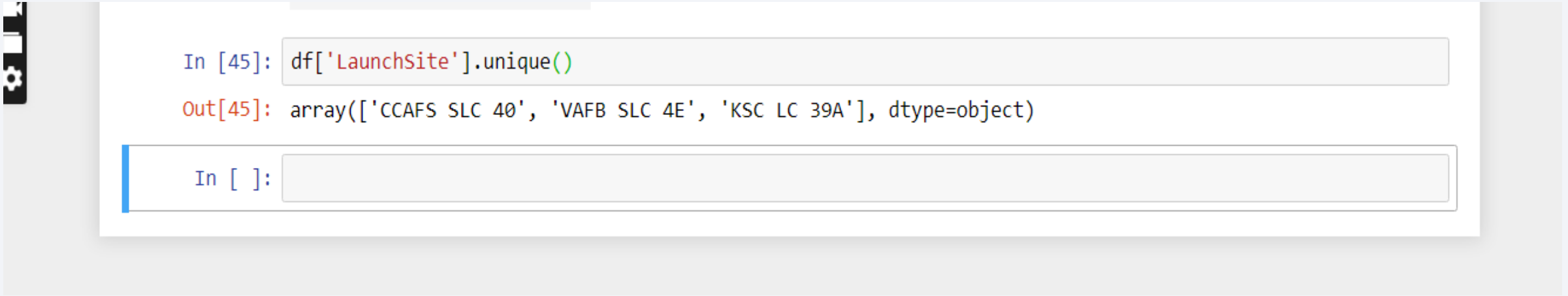
```
In [41]: sns.catplot(y="PayloadMass",x="Orbit",hue='Class',data=df, aspect=5)
         plt.xlabel("Orbit",fontsize=20)
         plt.ylabel("Pay Load Mass",fontsize=20)
         plt.show()
```



Launch Success Yearly Trend



All Launch Site Names



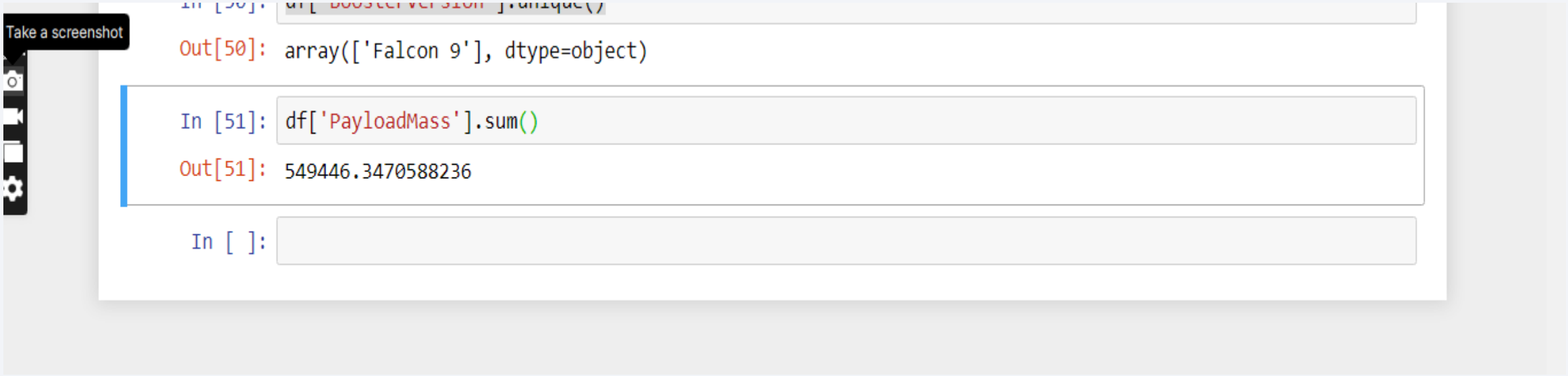
- There are just three unique launch sites

Launch Site Names Begin with 'CCA'

- SELECT * FROM db WHERE LAUNCH_SITE LIKE'CCA%' LIMIT 5;
- We can use SQL to query the necessary result
- Db is our database

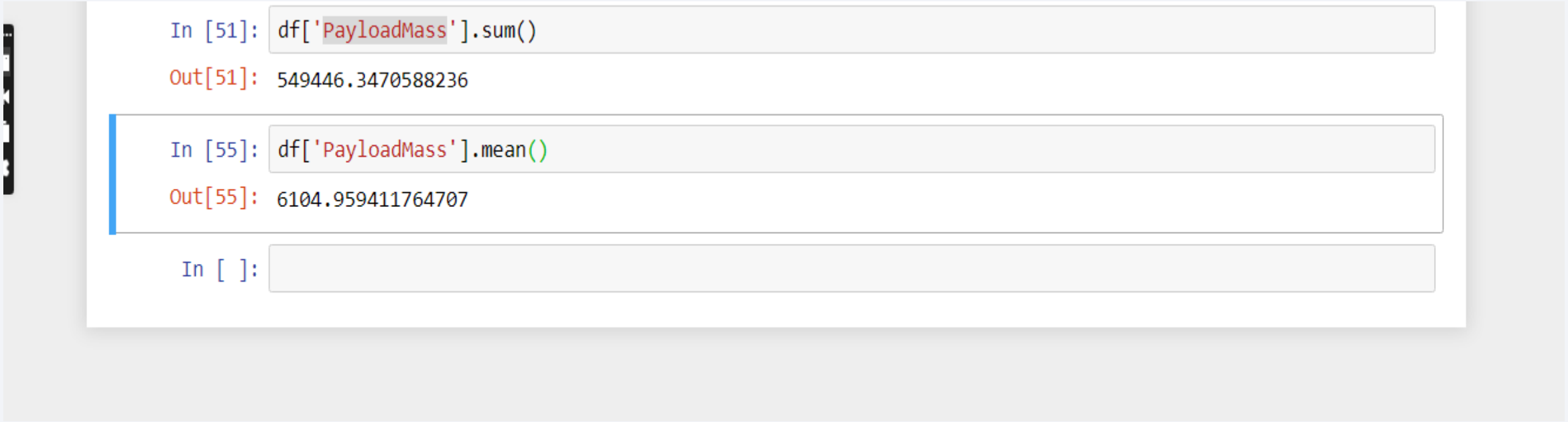
Total Payload Mass

- The total payload mass is 549446



Average Payload Mass by F9 v1.1

- The mean is 6104.959411764707



First Successful Ground Landing Date

```
In [62]: new=df[df['Outcome']=='True RTLS']

In [64]: new['Date']

Out[64]: 16    2015-12-22
         22    2016-07-18
         26    2017-02-19
         29    2017-05-01
         31    2017-06-03
         35    2017-08-14
         37    2017-09-07
         41    2017-12-15
         43    2018-01-08
         56    2018-10-08
         65    2019-06-12
         66    2019-07-25
```

Successful Drone Ship Landing with Payload between 4000 and 6000

```
Name: Date, dtype: object
```

```
In [78]: df[(df['PayloadMass']>4000) & (df['PayloadMass']<6000) & (df['Outcome']=='True Ocean')]
```

Out[78]:

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	Landing
44	45	2018-01-31	Falcon 9	4230.0	GTO	CCAFS SLC 40	True Ocean	2	True	True	True	

```
In [ ]:
```

- Only one booster successfully did so

Total Number of Successful and Failure Mission Outcomes

- Total 60 successful missions with 9 failed missions

```
Name: Outcome, dtype: object

In [47]: df['Outcome'].value_counts()

Out[47]: True ASDS      41
         None None      19
         True RTLS      14
         False ASDS      6
         True Ocean      5
         None ASDS       2
         False Ocean     2
         False RTLS      1
         Name: Outcome, dtype: int64

In [ ]:
```

Take a screenshot

Boosters Carried Maximum Payload

- Falcon 9 has the highest payload

2015 Launch Records

```
In [60]: include = df[df['Date'].dt.year == 2015]

In [61]: include.head()
```

Out[61]:

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	
11	12	2015-01-10	Falcon 9	2395.0	ISS	CCAFS SLC 40	False ASDS	1	True	False	True	5e9e303
12	13	2015-02-11	Falcon 9	570.0	ES-L1	CCAFS SLC 40	True Ocean	1	True	False	True	
13	14	2015-04-14	Falcon 9	1898.0	ISS	CCAFS SLC 40	False ASDS	1	True	False	True	5e9e303
14	15	2015-04-27	Falcon 9	4707.0	GTO	CCAFS SLC 40	None None	1	False	False	False	
15	16	2015-06-28	Falcon 9	2477.0	ISS	CCAFS SLC 40	None ASDS	1	True	False	True	5e9e303

- Launch records for the year 2015

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

- Successful landings are 11

```
In [77]: #& (df['Outcome']=='True%')]
         y=df.loc[mask]
         y['Outcome'].value_counts()

Out[77]: None None      8
         True ASDS      5
         False ASDS     4
         True Ocean     3
         True RTLS      3
         None ASDS      2
         False Ocean    2
         Name: Outcome, dtype: int64

In [ ]:
```

Section 3

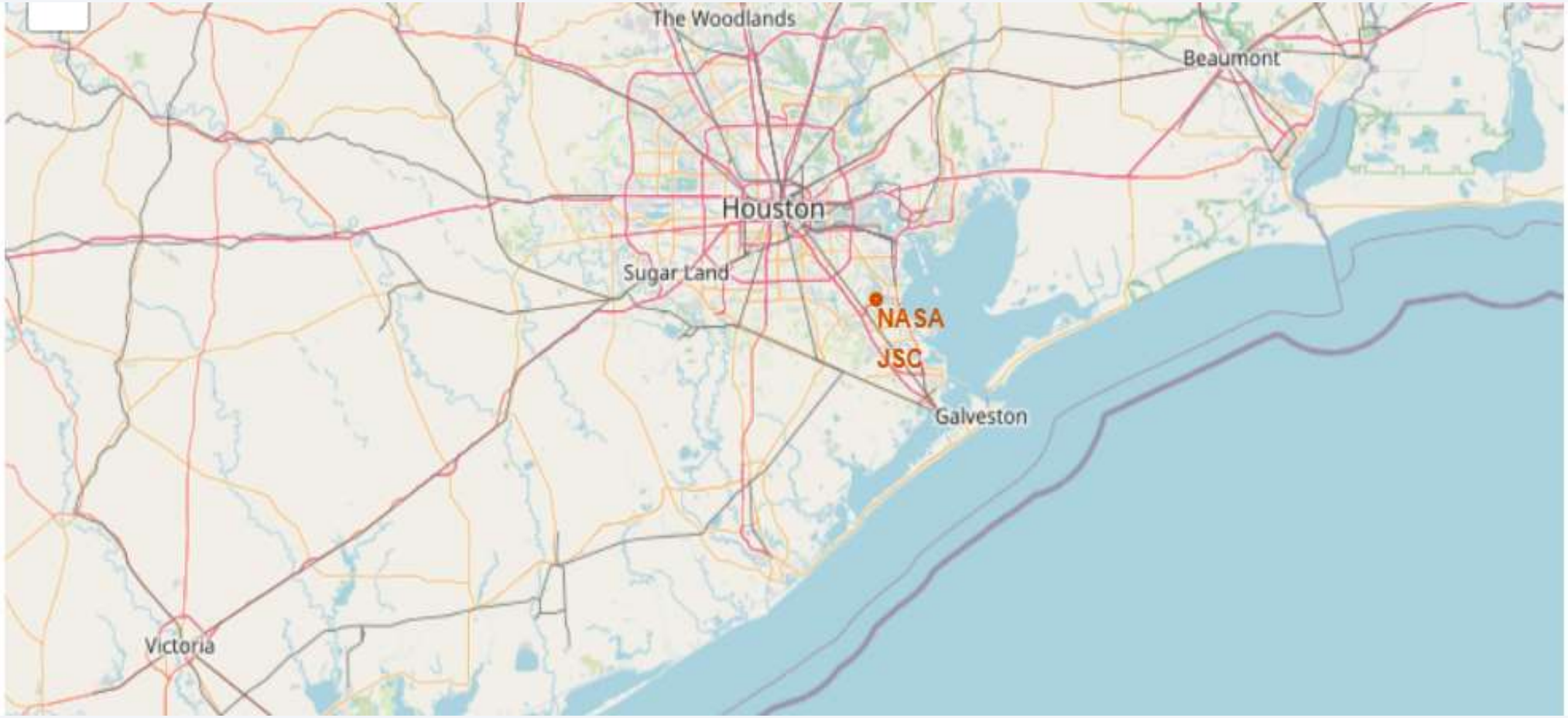
Launch Sites Proximities Analysis

Folium Map


United States
México
10
VAFB
SLC-
4E
46
AFS
SCC-
Maroc /
المغرب
موريتانيا

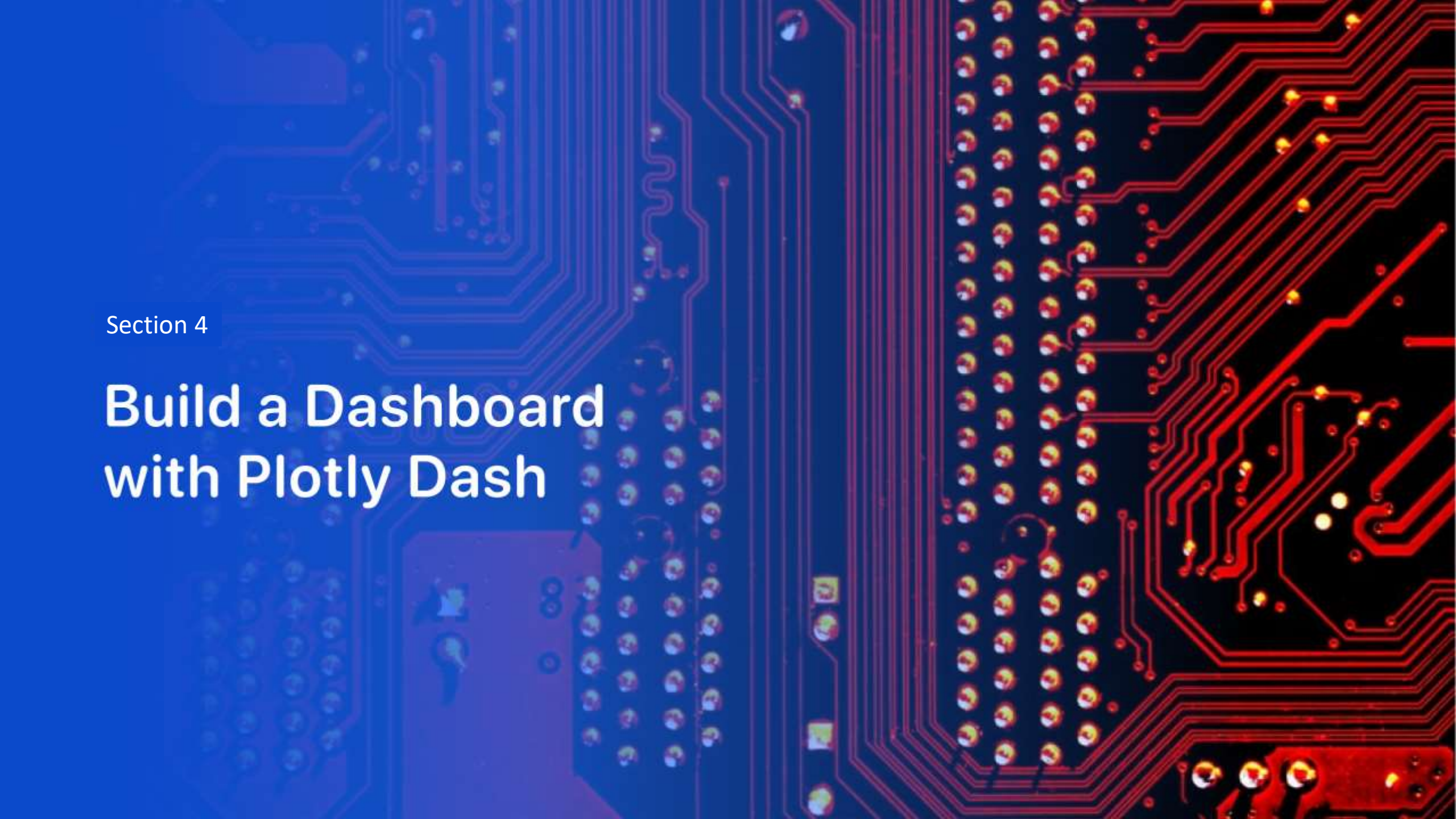
Folium Map Screenshot 2

Folium Map Screenshot 3



Section 4

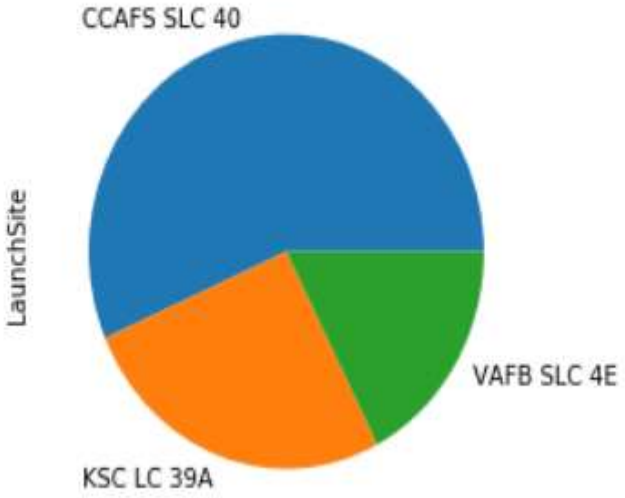
Build a Dashboard with Plotly Dash



Dashboard Screenshot 1

- CCAFS SLC 40 has the highest number of successful launches

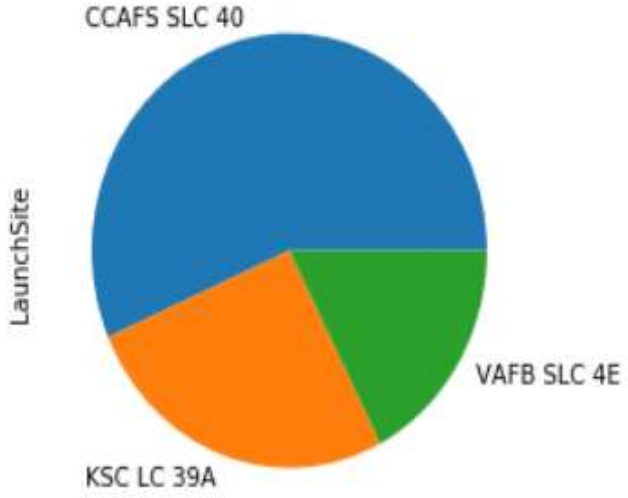
```
In [125]: df3['LaunchSite'].value_counts().plot(kind='pie')
Out[125]: <matplotlib.axes._subplots.AxesSubplot at 0x18212240>
```



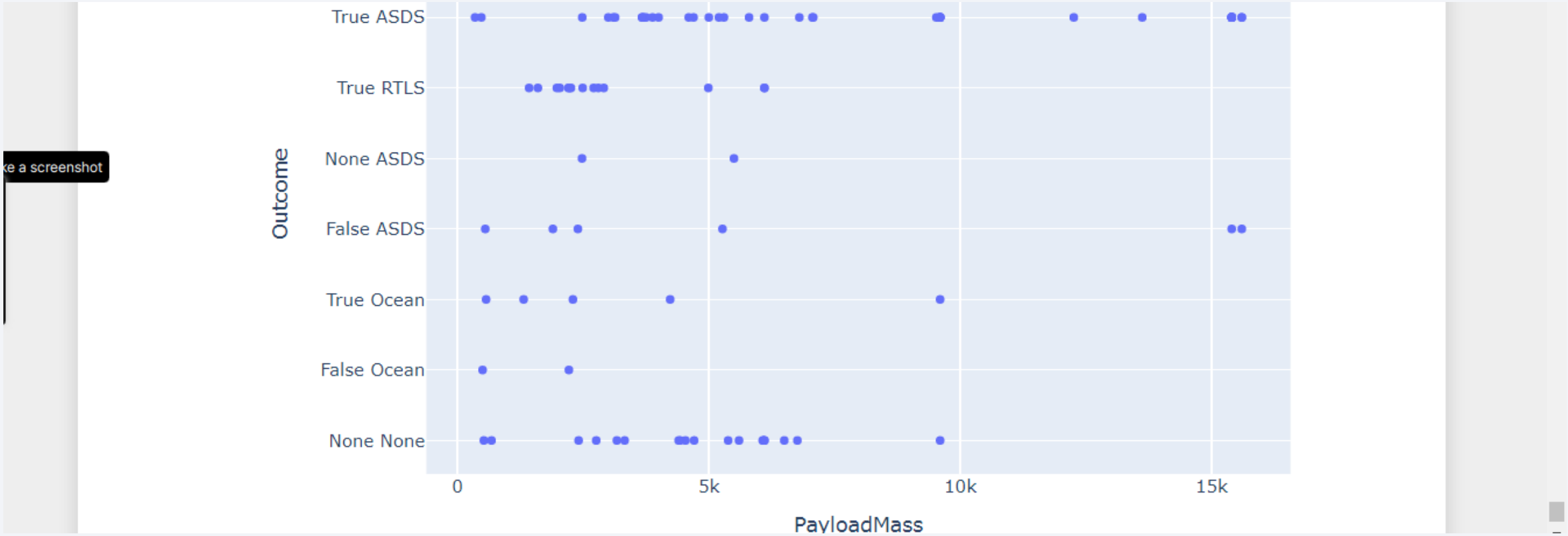
Dashboard Screenshot 2

```
In [125]: df3['LaunchSite'].value_counts().plot(kind='pie')
```

```
Out[125]: <matplotlib.axes._subplots.AxesSubplot at 0x18212240>
```



<Dashboard Screenshot 3>



Section 5

Predictive Analysis (Classification)

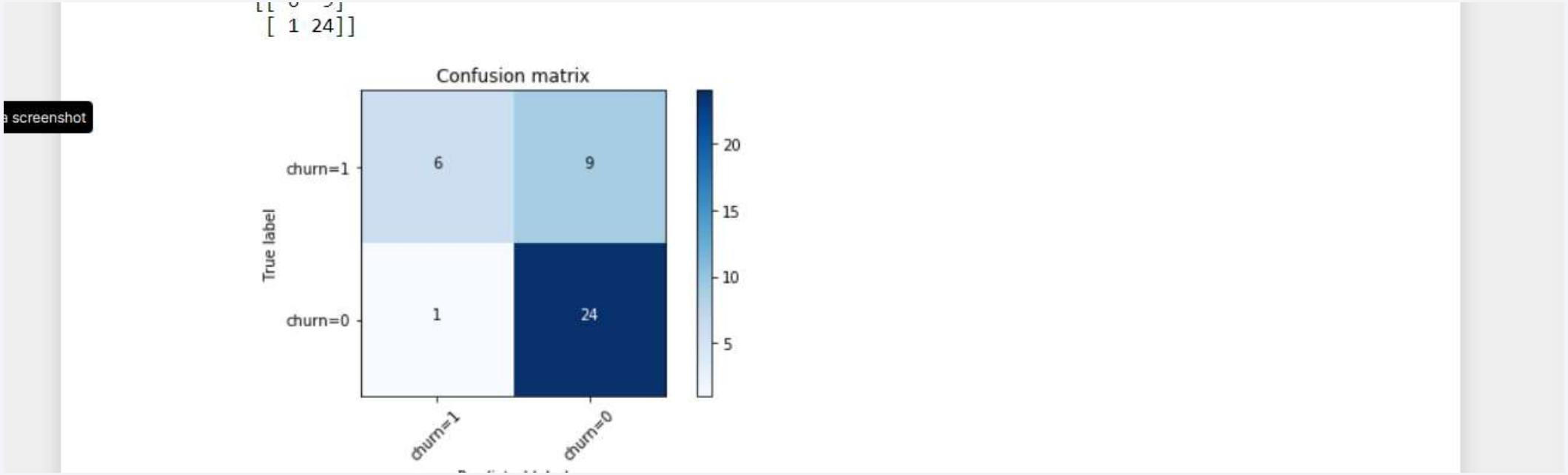
Classification Accuracy

```
SVM_accuracy=svm_cv.score(X_test, Y_test)
Logistic_Regression=logreg_cv.score(X_test, Y_test)

Report['Logistic_Reg'] = [Logistic_Regression]
Report['SVM'] = [SVM_accuracy]
Report['Decision Tree'] = [Decision_tree_accuracy]
Report['KNN'] = [knn_accuracy]



Report.transpose()
```

Out[27]:

	0
Method	Test Data Accuracy
Logistic_Reg	0.833333
SVM	0.833333
Decision Tree	0.833333
KNN	0.833333

Confusion Matrix



Conclusions

. CCAFS SLC 40 has the highest number of successful launches

A successful regression and classification model was developed to predict if a launch will be successful

Thank you!