Data Science Capstone Project

Sarote Tongra-ar

November 5, 2021

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion
- Appendix

EXECUTIVE SUMMARY

Summary of Methodologies
Data collection
Data Wrangling
EDA with Data Visualization
EDA with SQL
Building and interactive map with Folium
Predict Analysis Classification

Summary of results
Exploratory data analysis results
Interactive analytics demo
Predictive analysis

INTRODUCTION

- Project background
- SpaceX advertises Falcon 9 rocket launches with cost of 62 million USD. However, other providers cost up to 165 million USD. The cost saving from SpaceX because of reuse the part of missile. So, if we can determine if the missile can land , we can save cost of launching cost.

- Investigate information
  - What are the key factors rocket will land successfully
  - Effect of parameters that impact the success of landing
  - What condition Space X to achieves the best result on landing

METHODOLOGY

- Data collection methodology

  SpaceX Rest API

  Web scraping

- Performed Data Wrangling (For Machine Learning readiness)

  One Hot coding for Machine leaning and dropping irrelevant info

- Perform exploratory data analysis (DEA) using SQL and web visualization

  Plotting: Bar Graphs, Scatter plot to show relationships between parameters

- Performed interactive visual analytics using Folium (Map) and Plotly Dash

- Performed predictive analysis using classification models

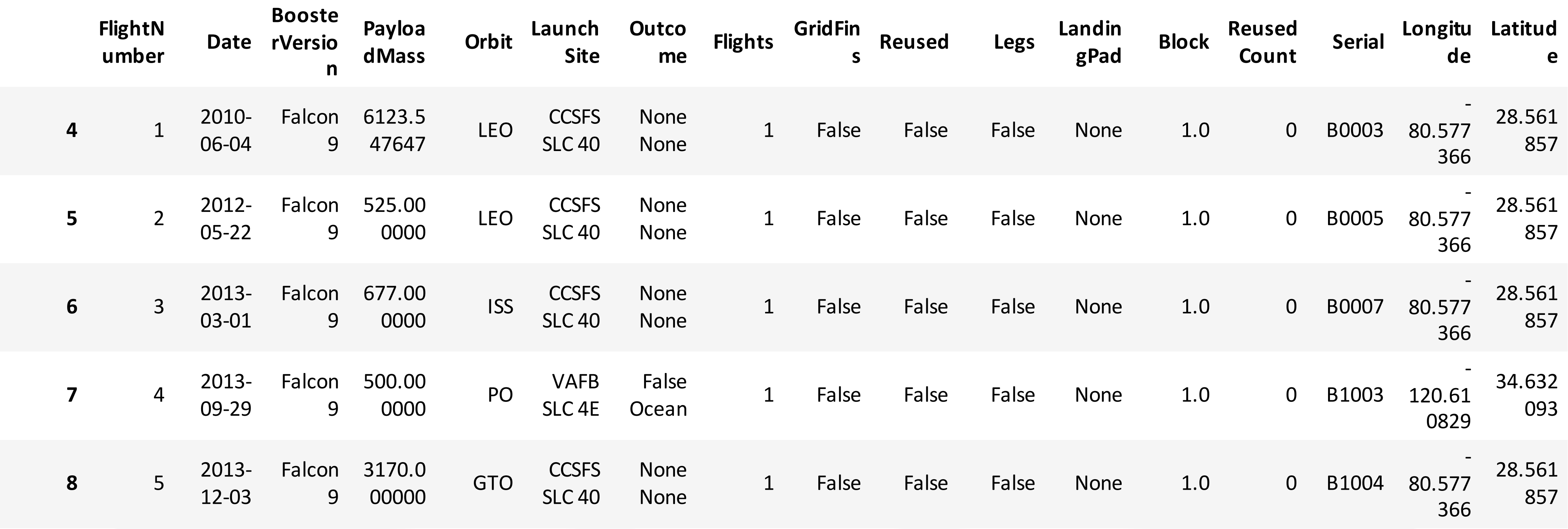
  How to find the best parameter for the classification models

METHODOLOGY

- Data Collecting Methodology
- With Space X Rest API
  - API will give us data detail about launches, including rocket revision, location, payload and etc with landing outcome.
  - The SpaceX Rest API stat with api.spacexdata.com/v4/
- With Web scraping with BeautifulSoup Module from Wikipedia

RESULTS

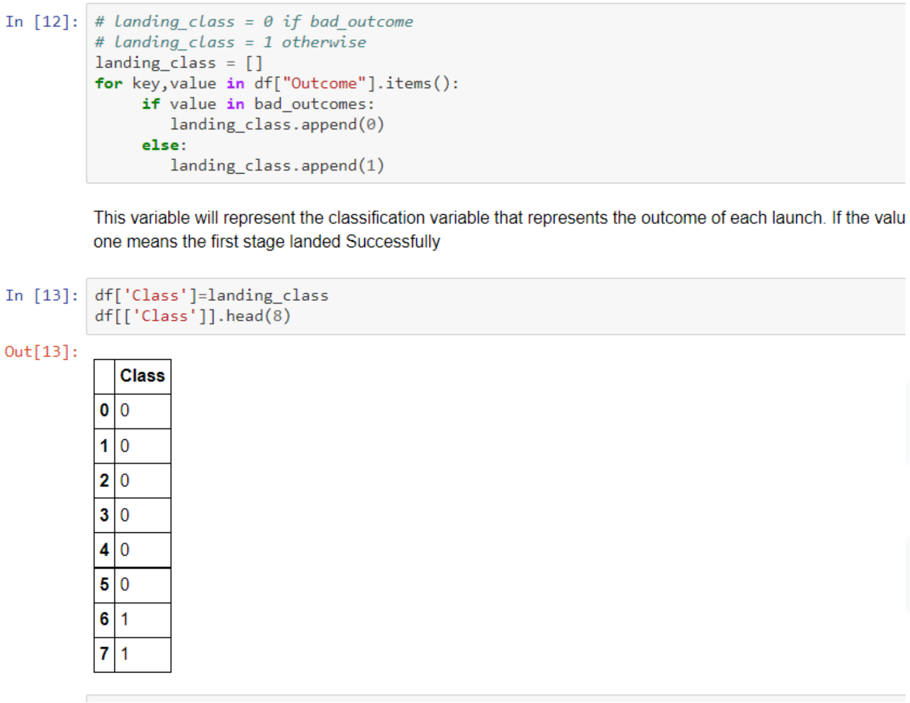
Data Collecting Methodology



	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	1	2010-06-04	Falcon 9	6123.547647	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	2	2012-05-22	Falcon 9	525.000000	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
6	3	2013-03-01	Falcon 9	677.000000	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
7	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
8	5	2013-12-03	Falcon 9	3170.000000	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857

METHODOLOGY Data Wrangling

- Data Wrangling
- To mapping success with classification to 1 and fail to 0 with new column



```
In [12]: # landing_class = 0 if bad_outcome
         # landing_class = 1 otherwise
         landing_class = []
         for key,value in df["Outcome"].items():
               if value in bad_outcomes:
                   landing_class.append(0)
               else:
                   landing_class.append(1)
```

This variable will represent the classification variable that represents the outcome of each launch. If the valu
one means the first stage landed Successfully

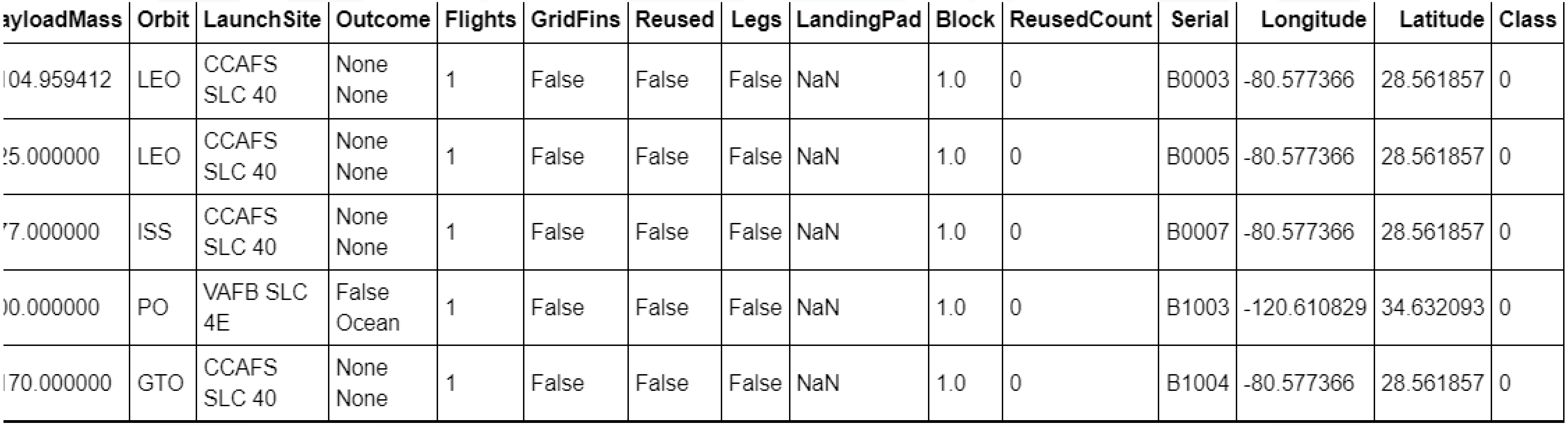
```
In [13]: df['Class']=landing_class
         df[['Class']].head(8)
```

Out[13]:

	Class
0	0
1	0
2	0
3	0
4	0
5	0
6	1
7	1

RESULTS from Datawrangling

Data Wragling



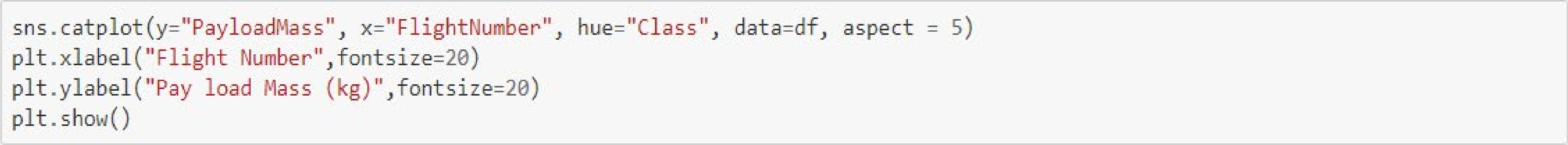
ayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude	Class
04.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.561857	0
5.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.561857	0
7.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.561857	0
0.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610829	34.632093	0
70.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28.561857	0

METHODOLOGY EDA with Data Visualization

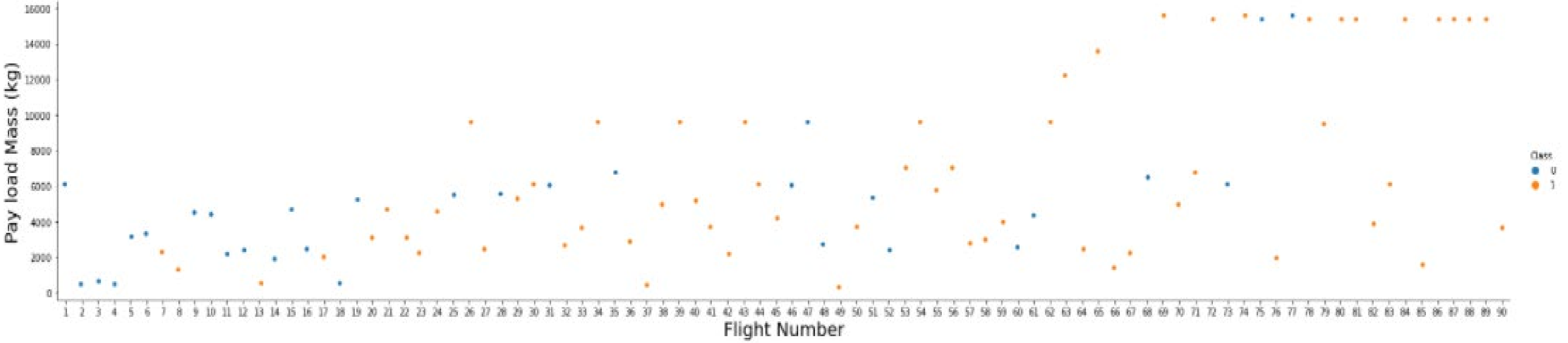
- Visualization for the data with graph been provided
  - Flight Number VS Payload Mass
  - Payload VS Launch Site
  - Orbit Vs Class
  - Orbit with Flight Number
  - Orbit Vs Payload Mass
  - Line Graph Success Rate s Year

RESULTS from EDA Visualization

Flight Number with Play Load Mass

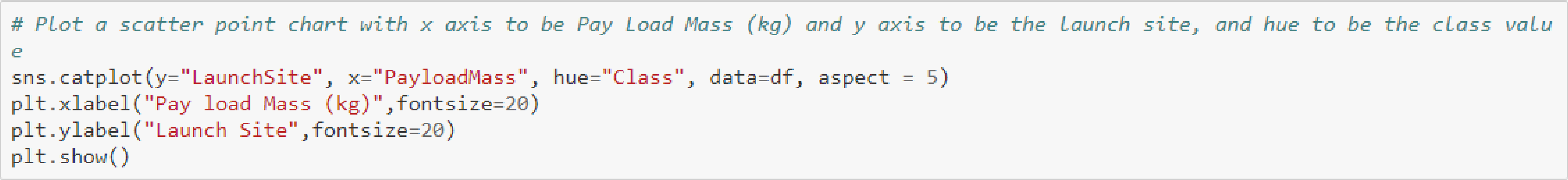


```
sns.catplot(y="PayloadMass", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Pay load Mass (kg)",fontsize=20)
plt.show()
```

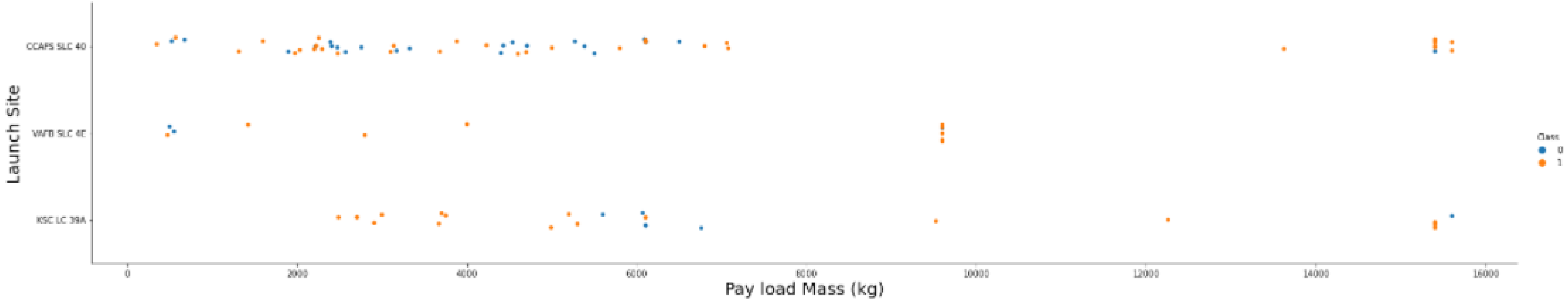


RESULTS from EDA Visualization

Launch Site with Play Load Mass



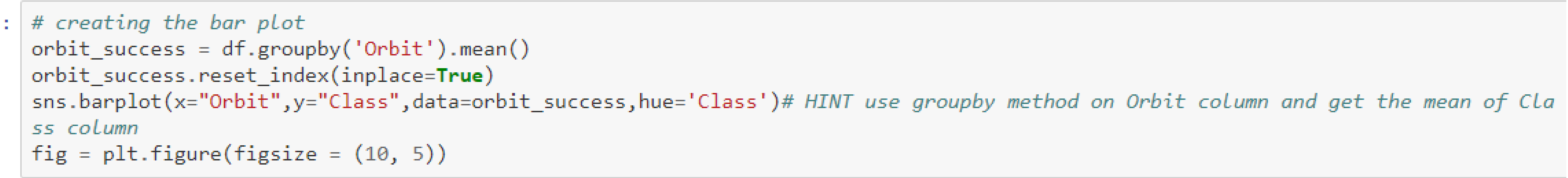
```
# Plot a scatter point chart with x axis to be Pay Load Mass (kg) and y axis to be the launch site, and hue to be the class valu
e
sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("Pay load Mass (kg)",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```



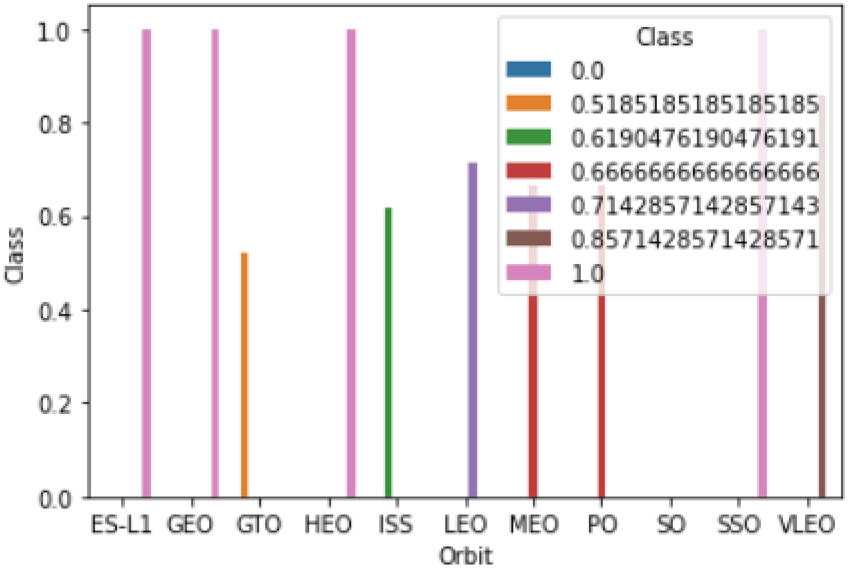
Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

RESULTS from EDA Visualization

Orbit with Class

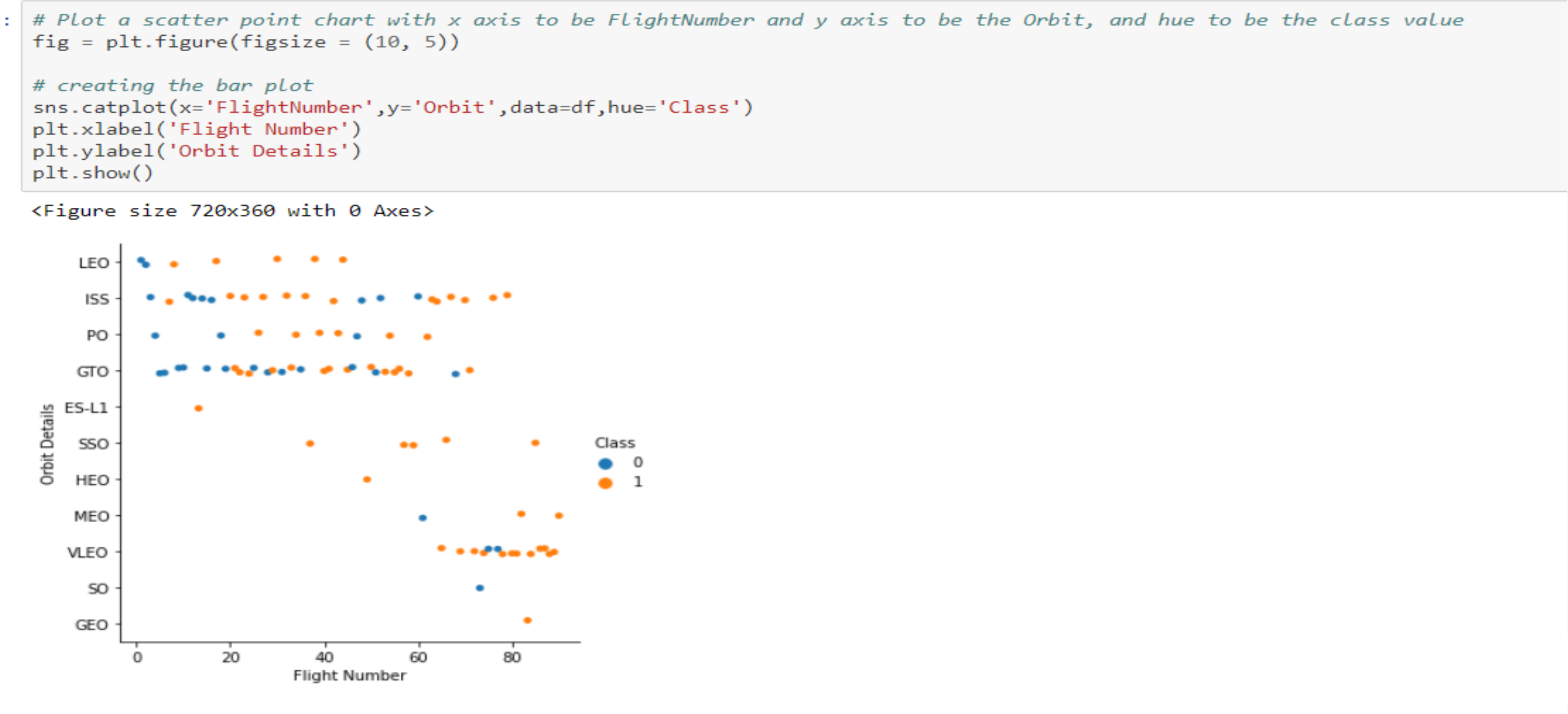


```
# creating the bar plot
orbit_success = df.groupby('Orbit').mean()
orbit_success.reset_index(inplace=True)
sns.barplot(x="Orbit",y="Class",data=orbit_success,hue='Class')# HINT use groupby method on Orbit column and get the mean of Cla
ss column
fig = plt.figure(figsize = (10, 5))
```



RESULTS from EDA Visualization

Orbit with Flight Number



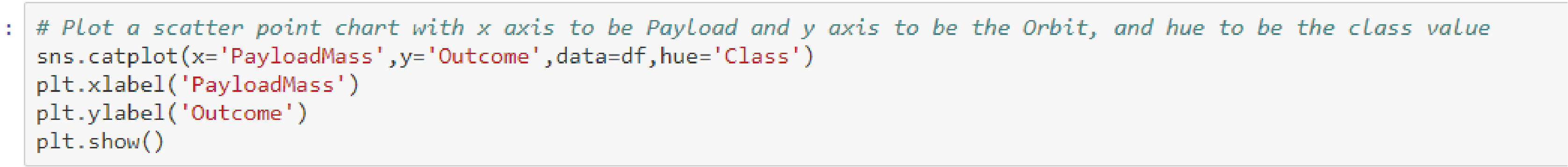
```
# Plot a scatter point chart with x axis to be FlightNumber and y axis to be the Orbit, and hue to be the class value
fig = plt.figure(figsize = (10, 5))

# creating the bar plot
sns.catplot(x='FlightNumber',y='Orbit',data=df,hue='Class')
plt.xlabel('Flight Number')
plt.ylabel('Orbit Details')
plt.show()
```

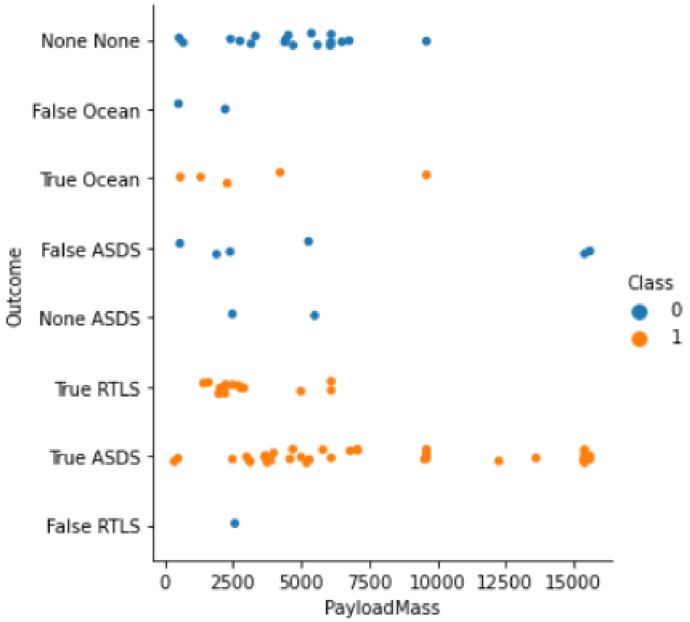
```
<Figure size 720x360 with 0 Axes>
```

RESULTS from EDA Visualization

Play Load Mass with Outcome

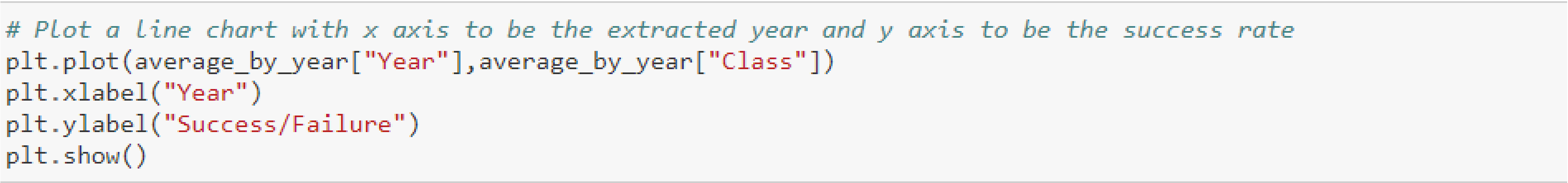


```
# Plot a scatter point chart with x axis to be Payload and y axis to be the Orbit, and hue to be the class value
sns.catplot(x='PayloadMass',y='Outcome',data=df,hue='Class')
plt.xlabel('PayloadMass')
plt.ylabel('Outcome')
plt.show()
```

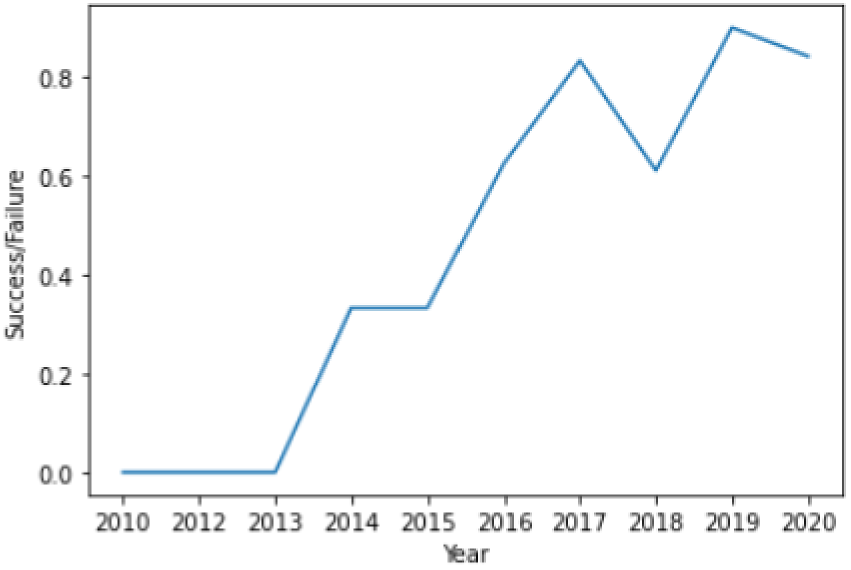


RESULTS from EDA Visualization

Success with Yearly Trend



```
# Plot a line chart with x axis to be the extracted year and y axis to be the success rate
plt.plot(average_by_year["Year"],average_by_year["Class"])
plt.xlabel("Year")
plt.ylabel("Success/Failure")
plt.show()
```

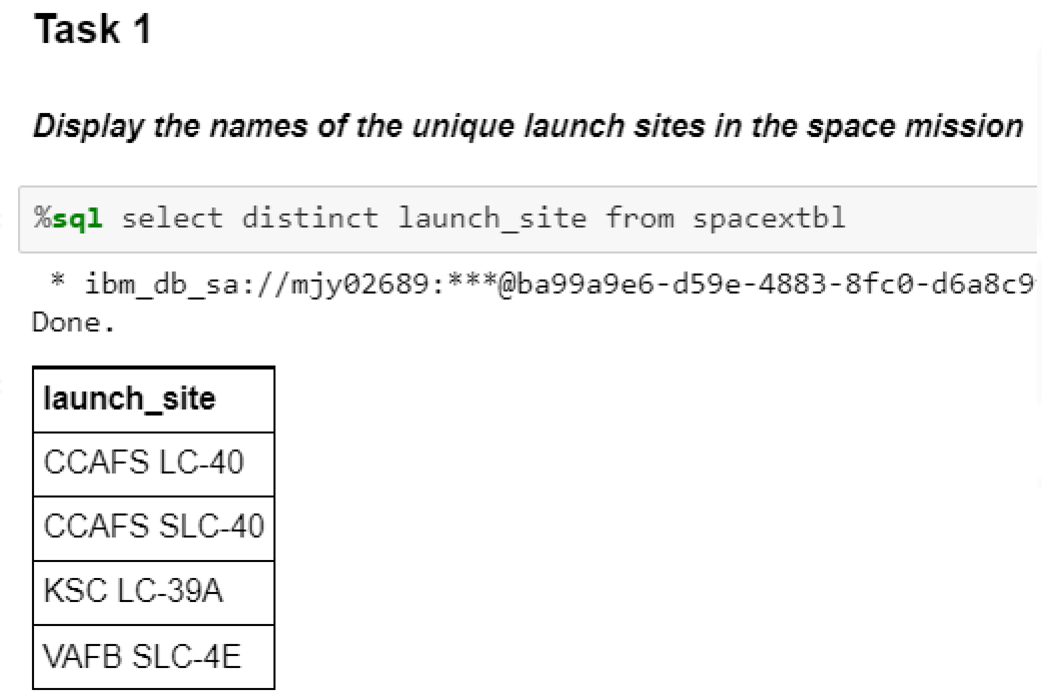


METHODOLOGY EDA SQL

- Visualization for the data with SQL
  - Display data from many query string

RESULTS from SQL Visualization

Unique Launch Site SQL



Task 1

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

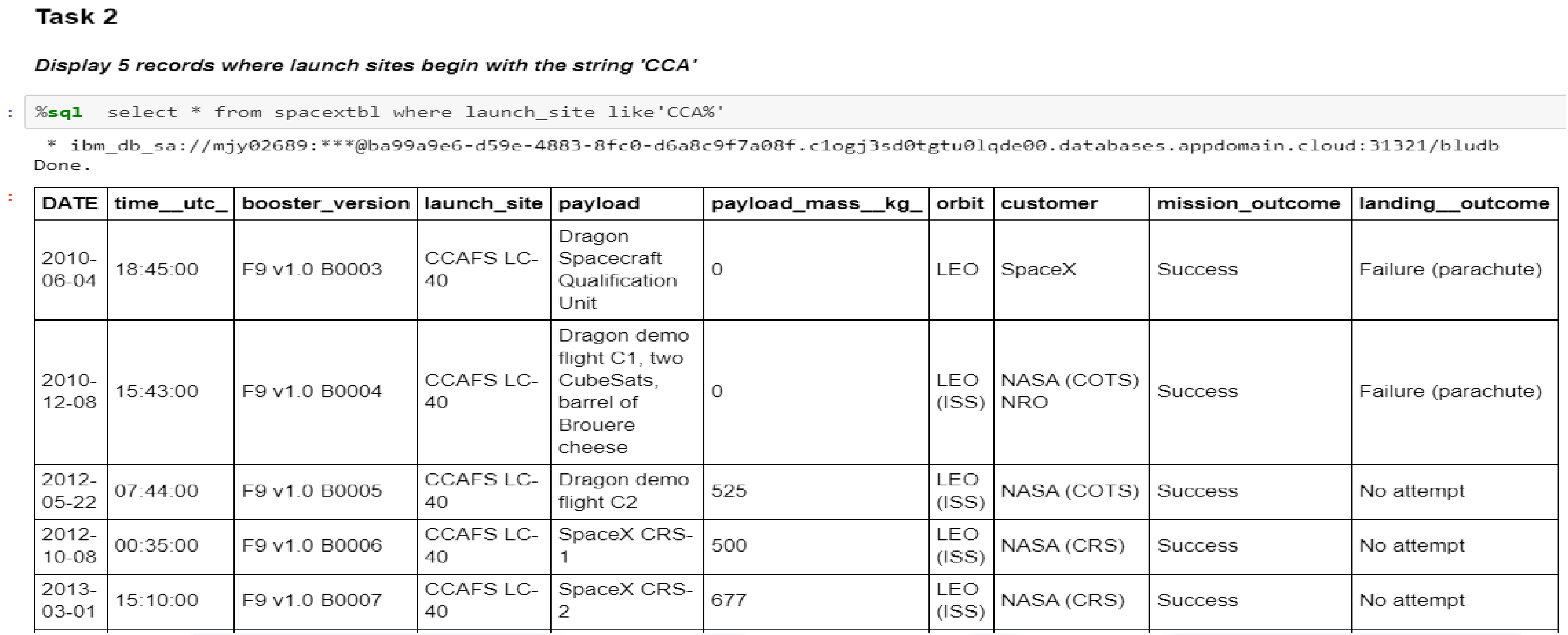
```
%sql select distinct launch_site from spacextbl
```

```
 * ibm_db_sa://mjy02689:***@ba99a9e6-d59e-4883-8fc0-d6a8c9
Done.
```

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

RESULTS from SQL Visualization

Site start name with CCA



Task 2

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

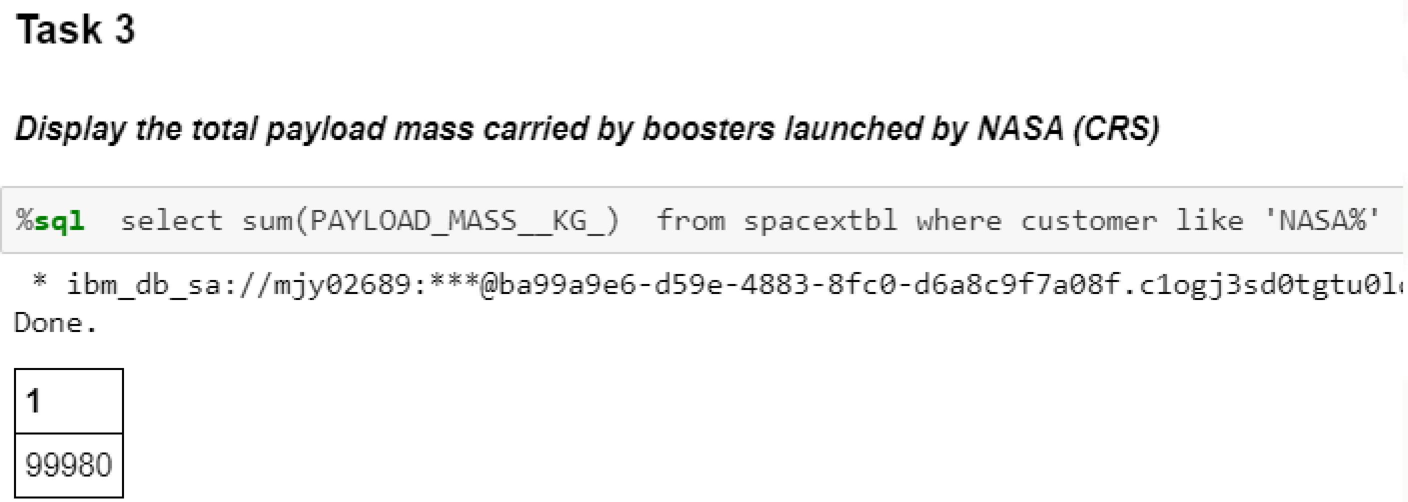
```
%sql  select * from spacextbl where launch_site like'CCA%'
```

```
 * ibm_db_sa://mjy02689:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
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	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

RESULTS from SQL Visualization

Total Pay load Launched by NASA



Task 3

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

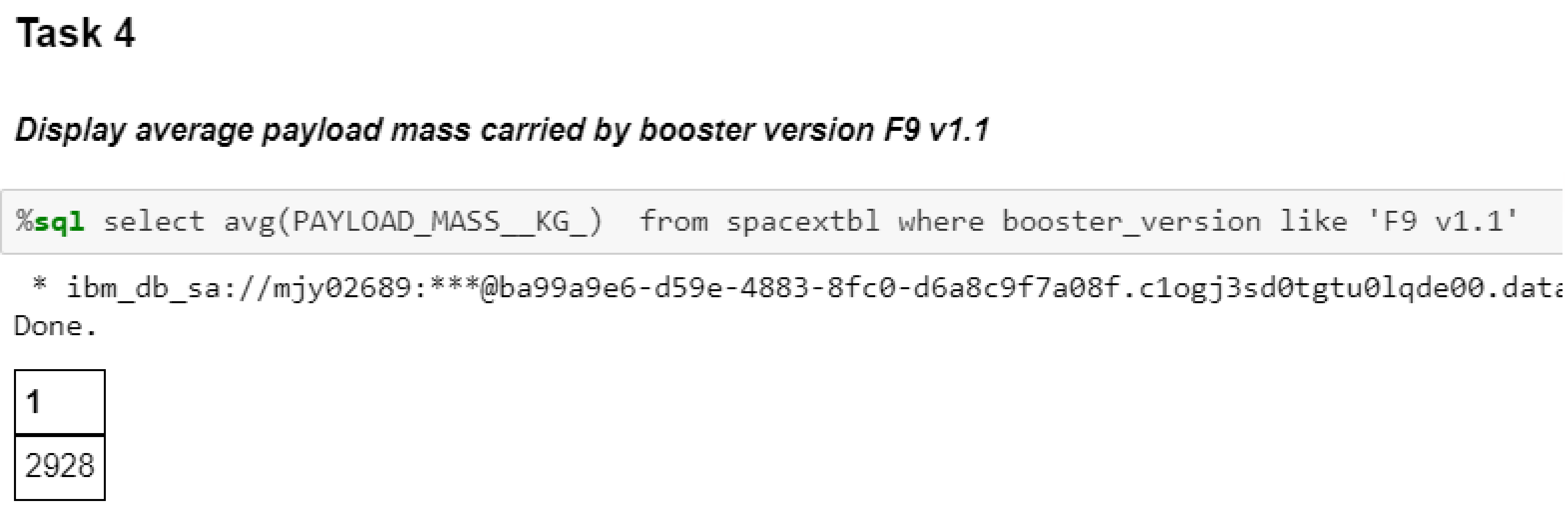
```
%sql  select sum(PAYLOAD_MASS__KG_)  from spacextbl where customer like 'NASA%'
```

```
 * ibm_db_sa://mjy02689:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0l
Done.
```

1
99980

RESULTS from SQL Visualization

Average Pay Load by Booster F9 v1.1



Task 4

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

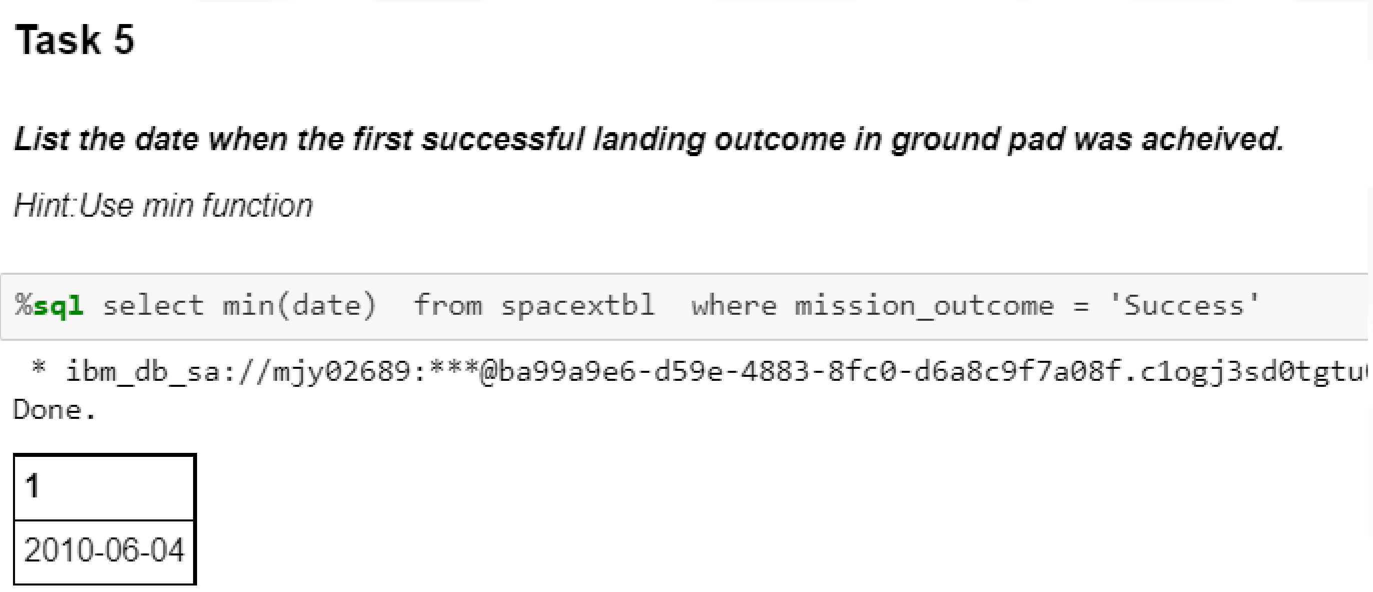
```
%sql select avg(PAYLOAD_MASS__KG_)  from spacextbl where booster_version like 'F9 v1.1'
```

```
 * ibm_db_sa://mjy02689:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.data
Done.
```

1
2928

RESULTS from SQL Visualization

Date with successful landing outcome



Task 5

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

Hint:Use min function

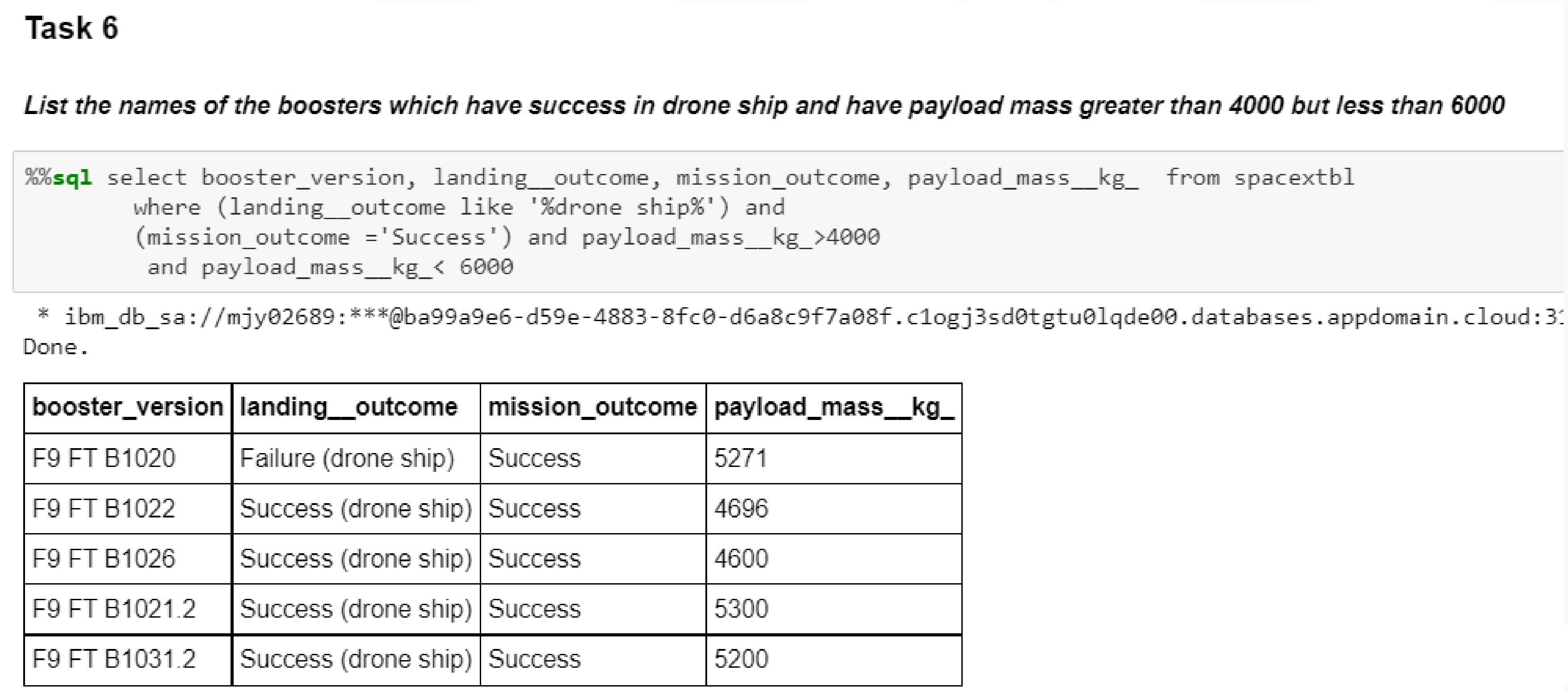
```
%sql select min(date)  from spacextbl  where mission_outcome = 'Success'
```

```
 * ibm_db_sa://mjy02689:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu
Done.
```

1
2010-06-04

RESULTS from SQL Visualization

Name of the boosters which have success in drone ship and pay load >400 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

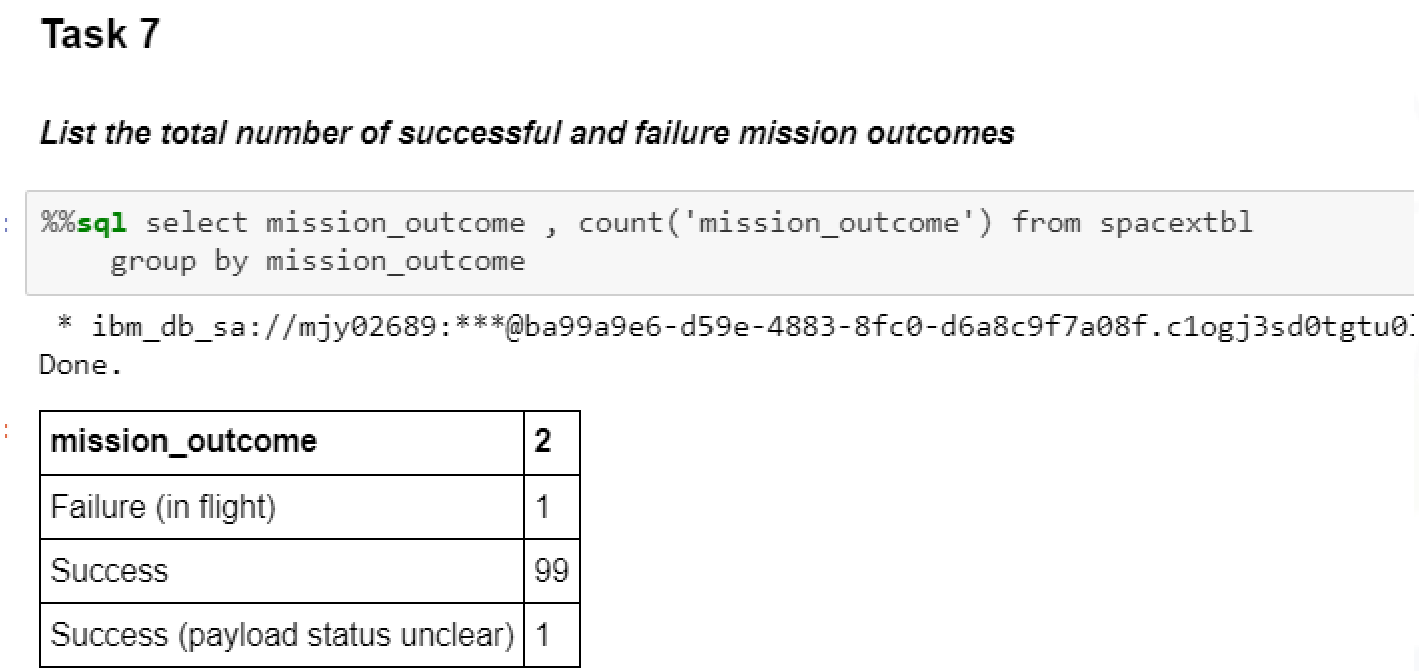
```
%%sql select booster_version, landing__outcome, mission_outcome, payload_mass__kg_  from spacextbl
    where (landing__outcome like '%drone ship%') and
    (mission_outcome ='Success') and payload_mass__kg_>4000
     and payload_mass__kg_< 6000
```

```
 * ibm_db_sa://mjy02689:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:3
Done.
```

booster_version	landing__outcome	mission_outcome	payload_mass__kg_
F9 FT B1020	Failure (drone ship)	Success	5271
F9 FT B1022	Success (drone ship)	Success	4696
F9 FT B1026	Success (drone ship)	Success	4600
F9 FT B1021.2	Success (drone ship)	Success	5300
F9 FT B1031.2	Success (drone ship)	Success	5200

RESULTS from SQL Visualization

Number of success and Failure



Task 7

List the total number of successful and failure mission outcomes

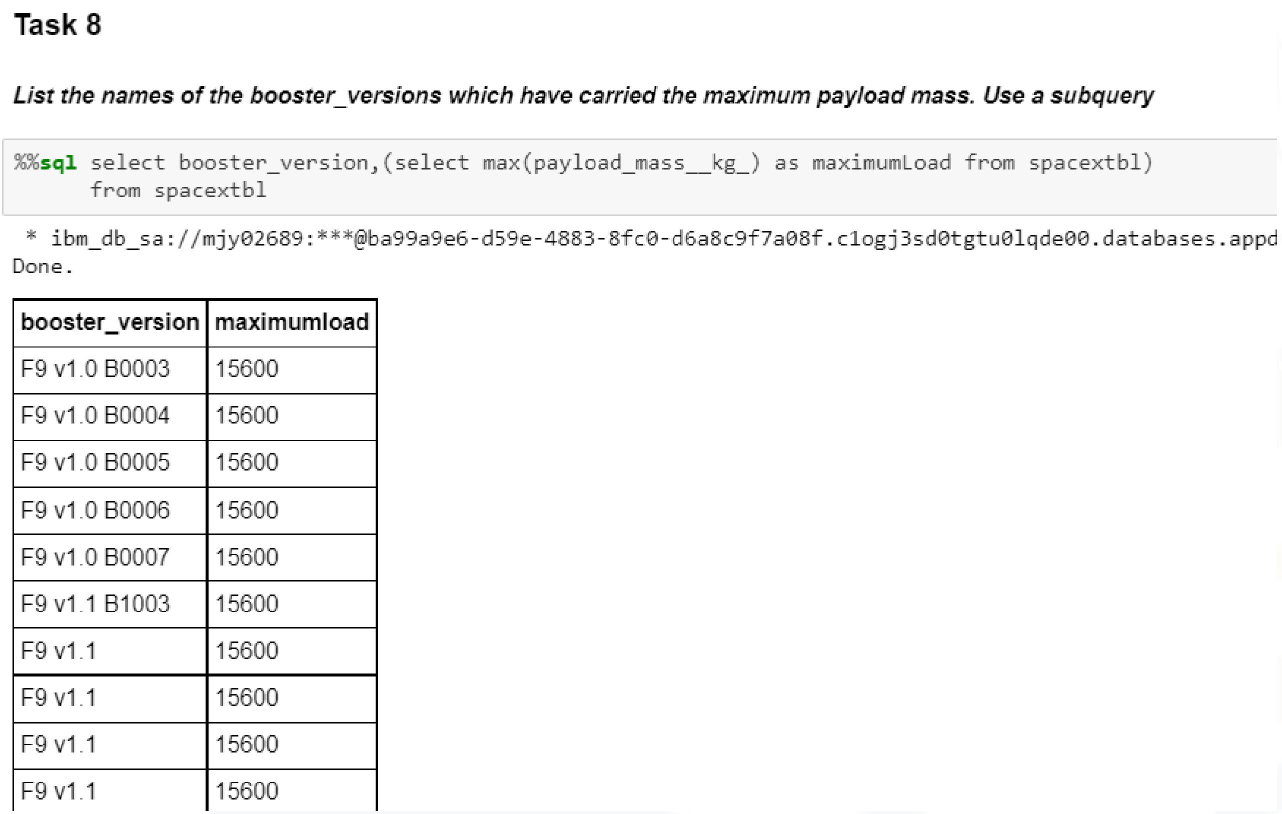
```
%%sql select mission_outcome , count('mission_outcome') from spacextbl
    group by mission_outcome
```

```
 * ibm_db_sa://mjy02689:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0l
Done.
```

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

RESULTS from SQL Visualization

List the name of booster version with maximum payload



Task 8

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

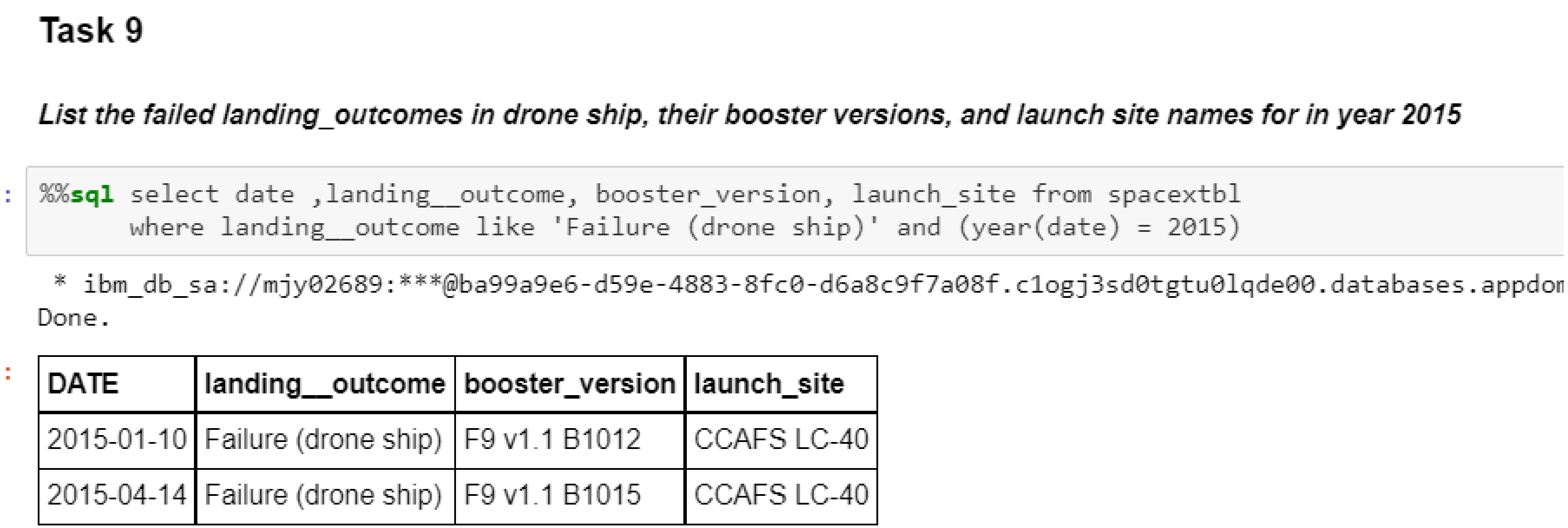
```
%%sql select booster_version,(select max(payload_mass__kg_) as maximumLoad from spacextbl)
      from spacextbl
```

```
 * ibm_db_sa://mjy02689:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appd
Done.
```

booster_version	maximumload
F9 v1.0 B0003	15600
F9 v1.0 B0004	15600
F9 v1.0 B0005	15600
F9 v1.0 B0006	15600
F9 v1.0 B0007	15600
F9 v1.1 B1003	15600
F9 v1.1	15600
F9 v1.1	15600
F9 v1.1	15600
F9 v1.1	15600

RESULTS from SQL Visualization

List the failed landing outcome in drone ship in year 2015



Task 9

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

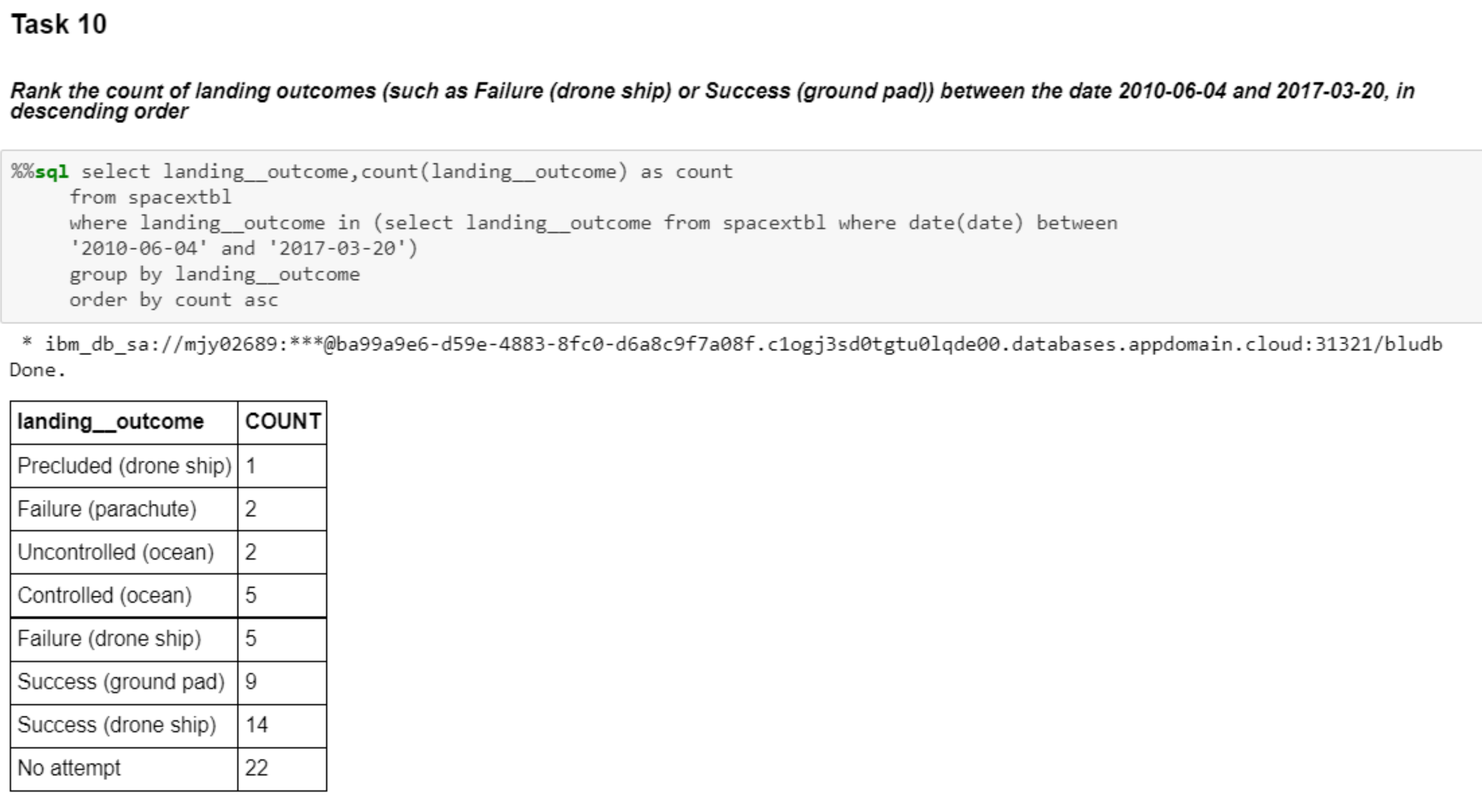
```
%%sql select date ,landing__outcome, booster_version, launch_site from spacextbl
      where landing__outcome like 'Failure (drone ship)' and (year(date) = 2015)
```

```
 * ibm_db_sa://mjy02689:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdom
Done.
```

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

RESULTS from SQL Visualization

Rank the count of landing outcome between date 2010-06-04 till 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

```
%%sql select landing__outcome,count(landing__outcome) as count
     from spacextbl
     where landing__outcome in (select landing__outcome from spacextbl where date(date) between
     '2010-06-04' and '2017-03-20')
     group by landing__outcome
     order by count asc
```

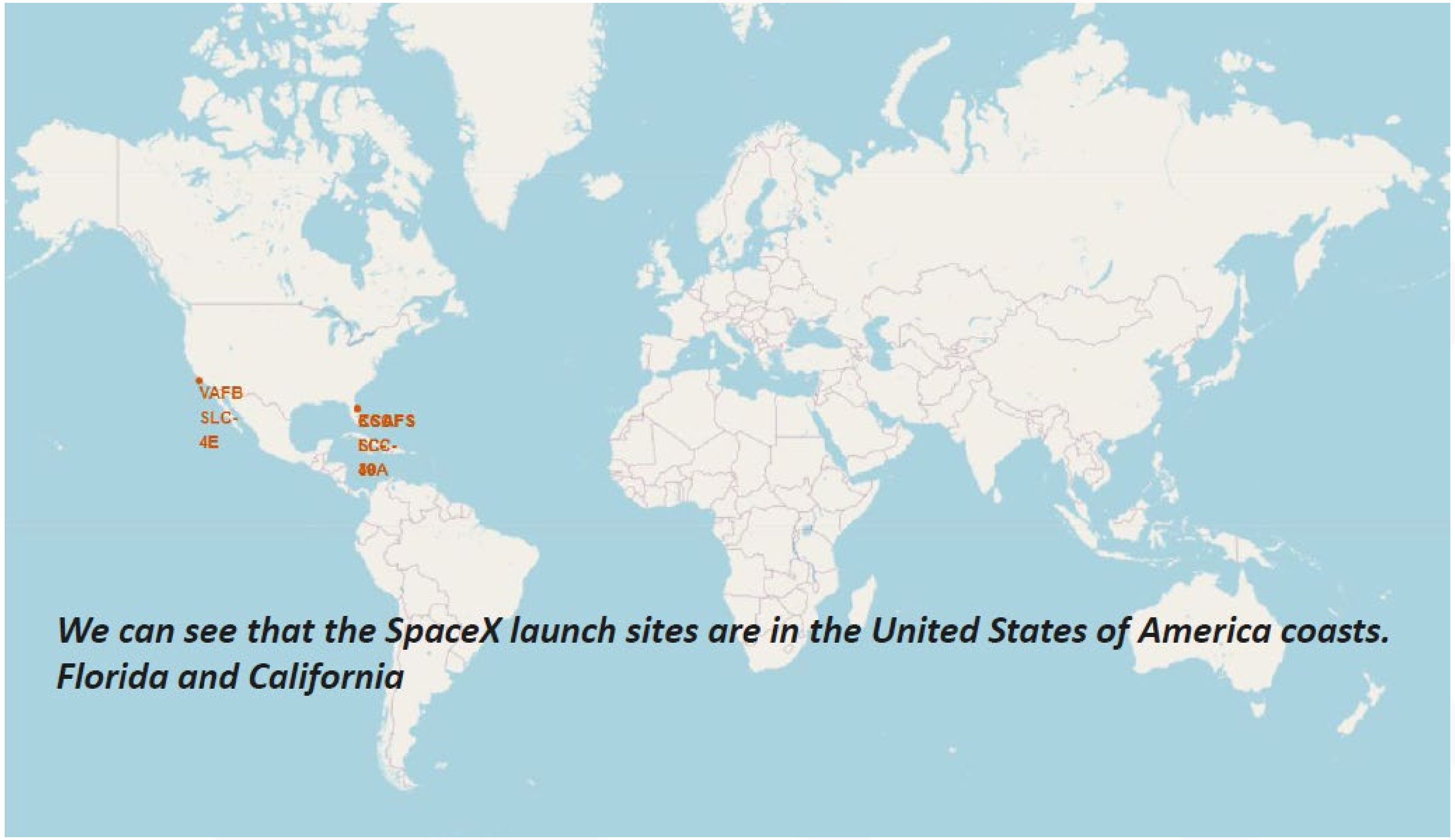
```
 * ibm_db_sa://mjy02689:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

landing__outcome	COUNT
Precluded (drone ship)	1
Failure (parachute)	2
Uncontrolled (ocean)	2
Controlled (ocean)	5
Failure (drone ship)	5
Success (ground pad)	9
Success (drone ship)	14
No attempt	22

METHODOLOGY Visualize with Map

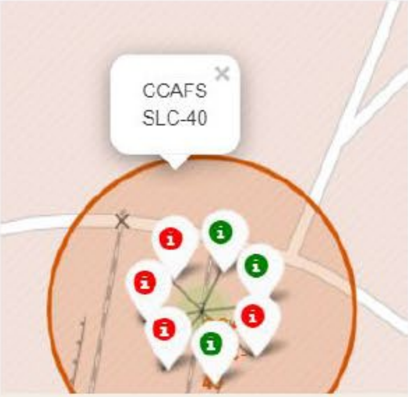
- Visualization for the data with Folium
- Visualize data with map

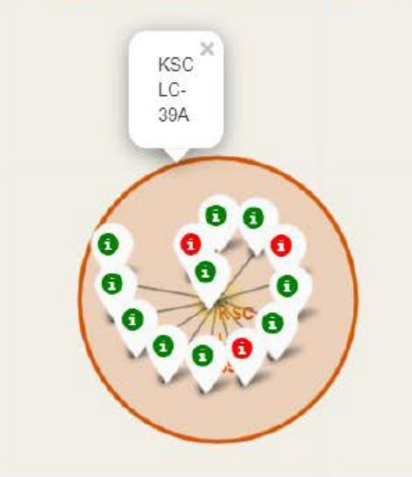
RESULTS from Map

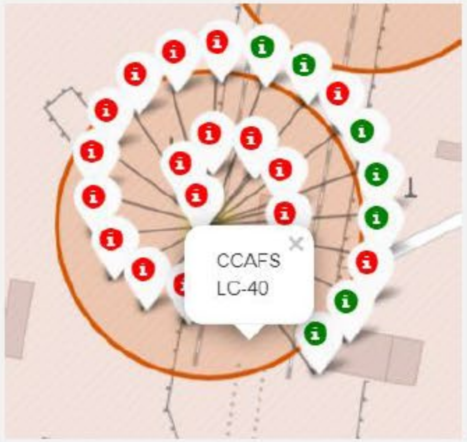


We can see that the SpaceX launch sites are in the United States of America coasts. Florida and California

RESULTS from Map and marker



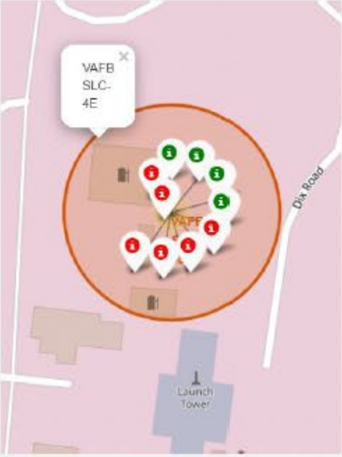




7
CCAFS SLC-40
26
CCAFS LC-40
Centaur Road

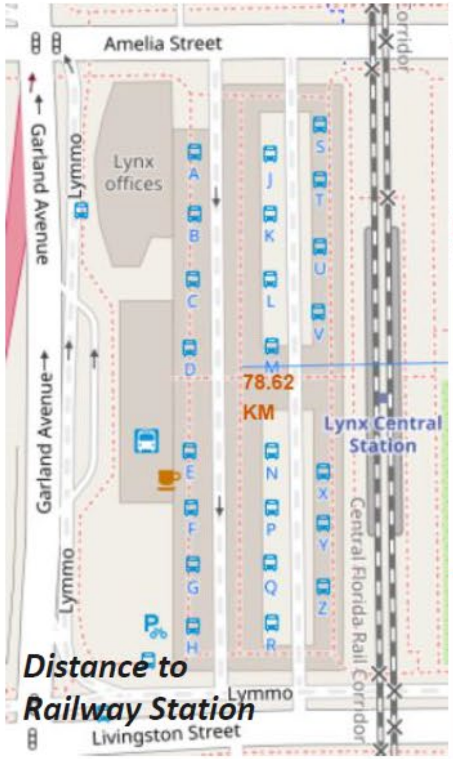
Florida Launch Sites

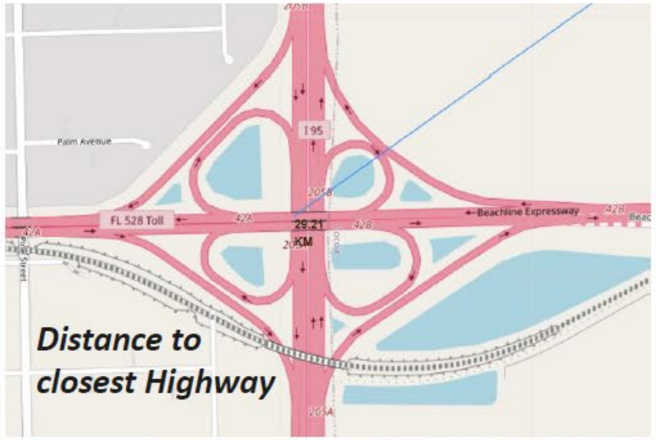
Green Marker shows successful Launches and Red Marker shows Failures

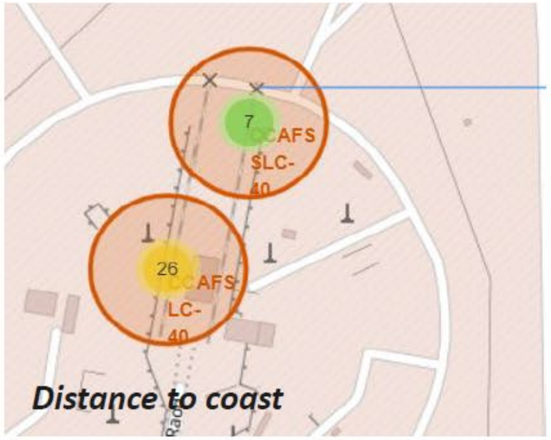


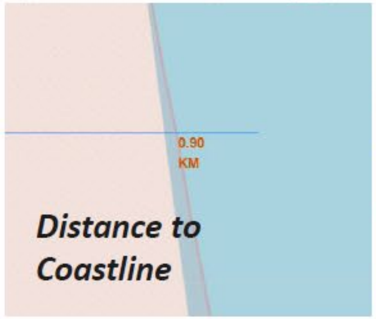
California Launch Site

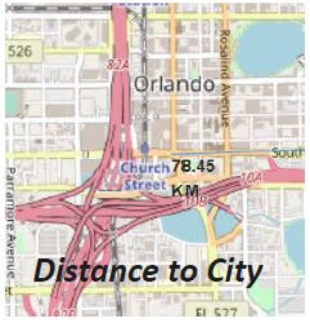
RESULTS from Launched site with land mark









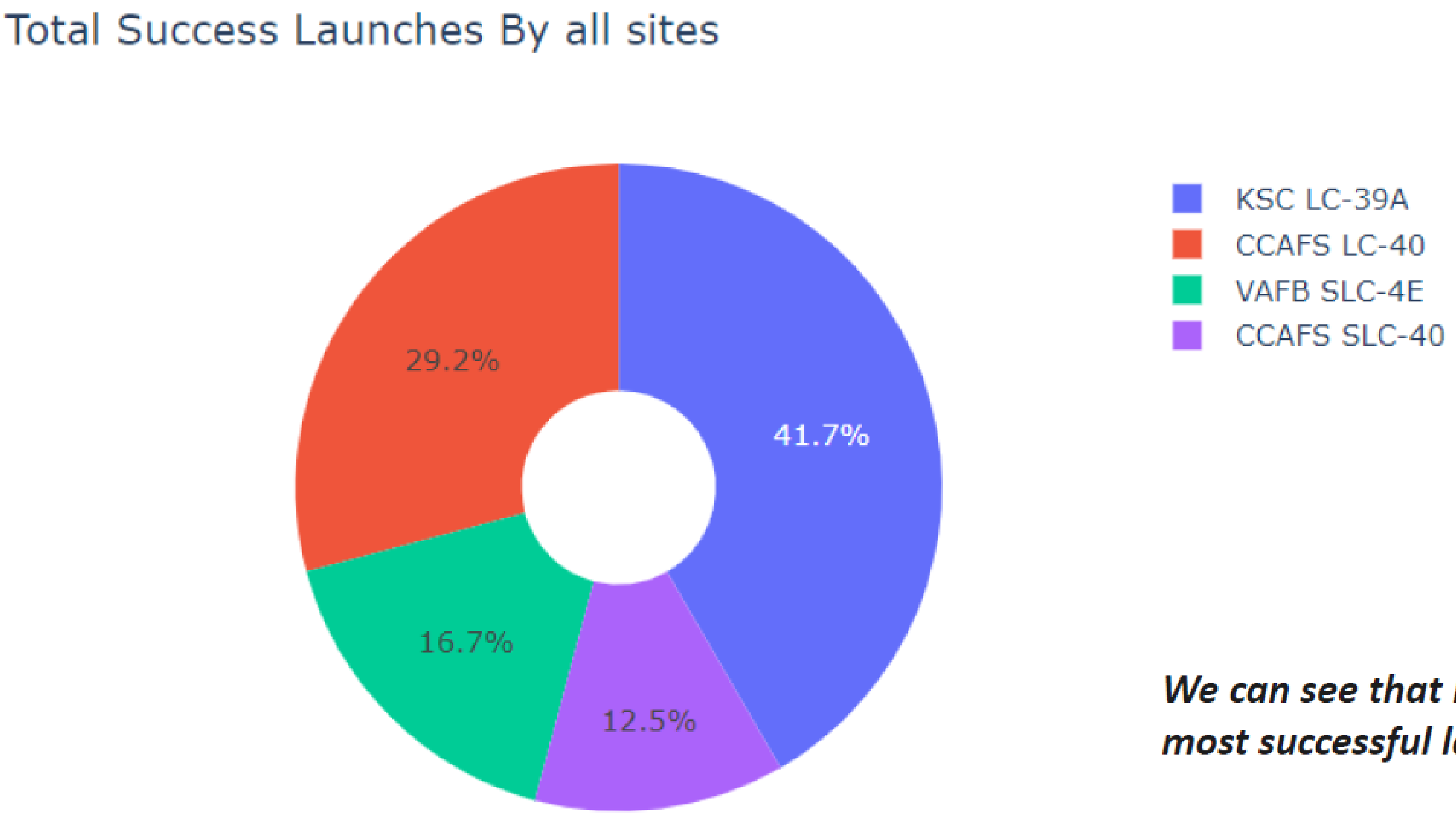


- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes

METHODOLOGY Visualize Dashboard

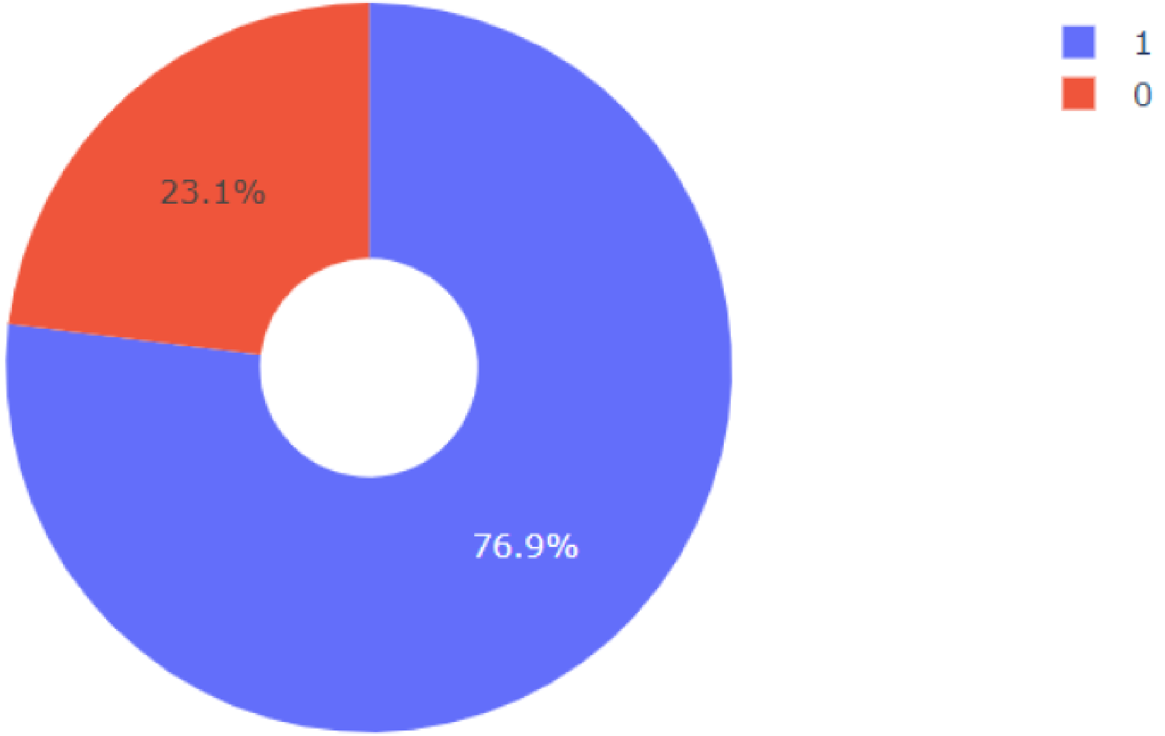
- Visualization for the data with Dashboard

RESULTS from Dasboard



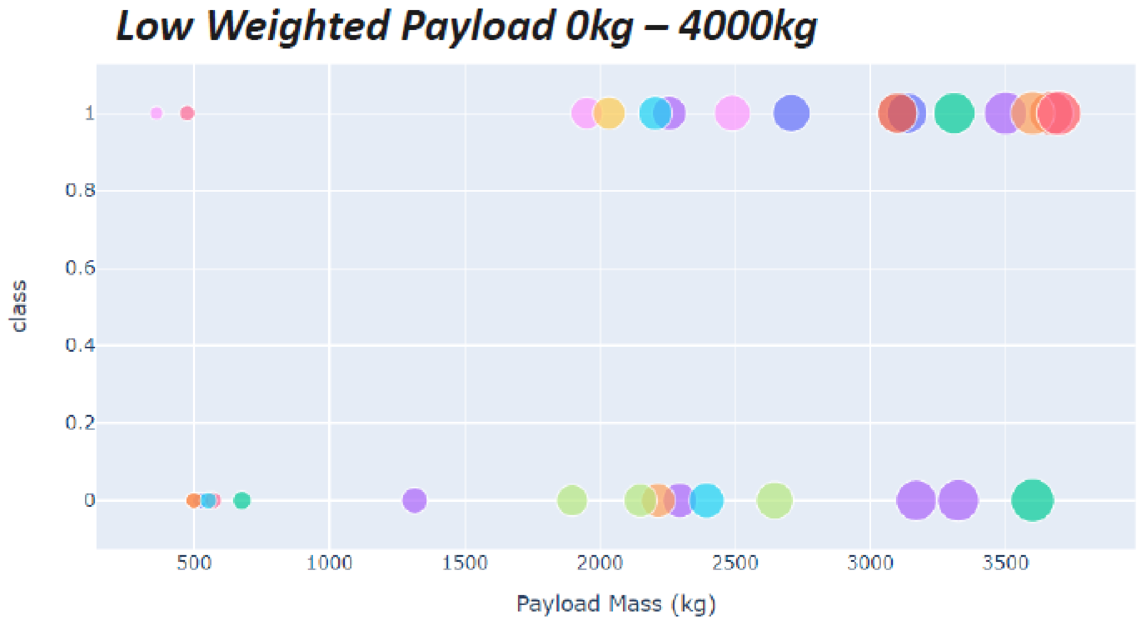
We can see that KSC LC-39A had the most successful launches from all the sites

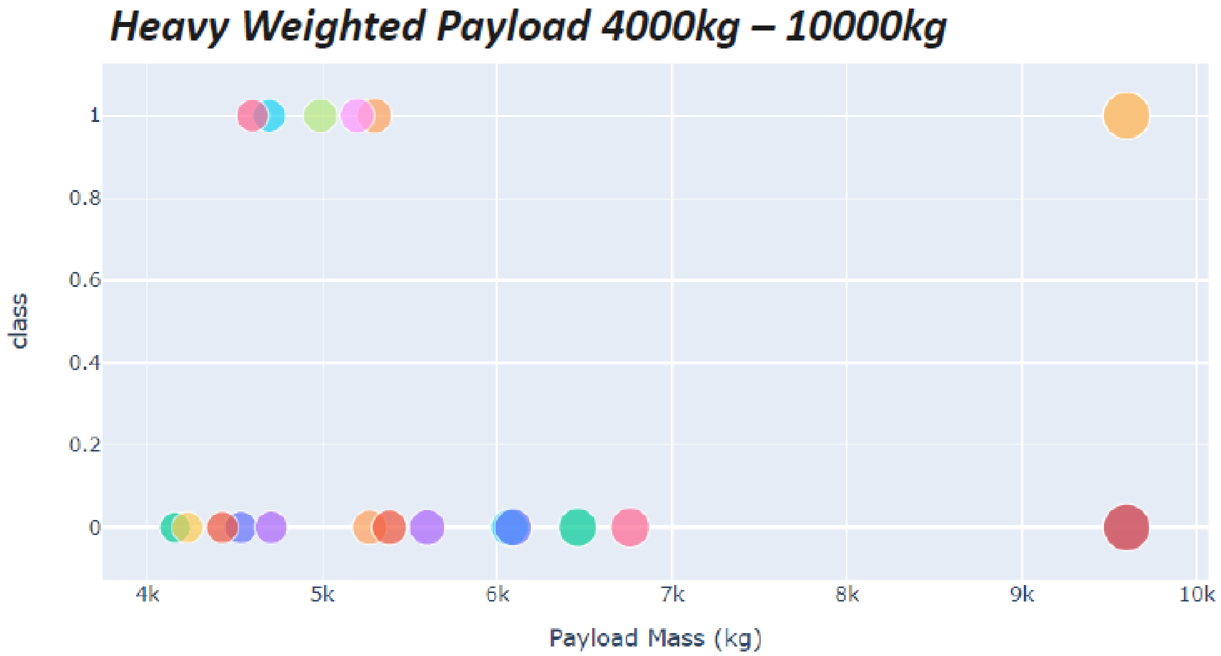
RESULTS from Dasboard



KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

RESULTS from Dasboard



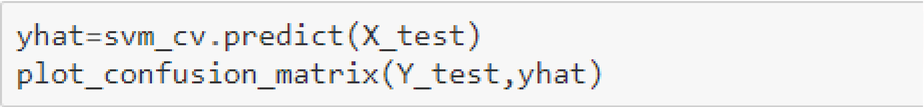


We can see the success rates for low weighted payloads is higher than the heavy weighted payloads

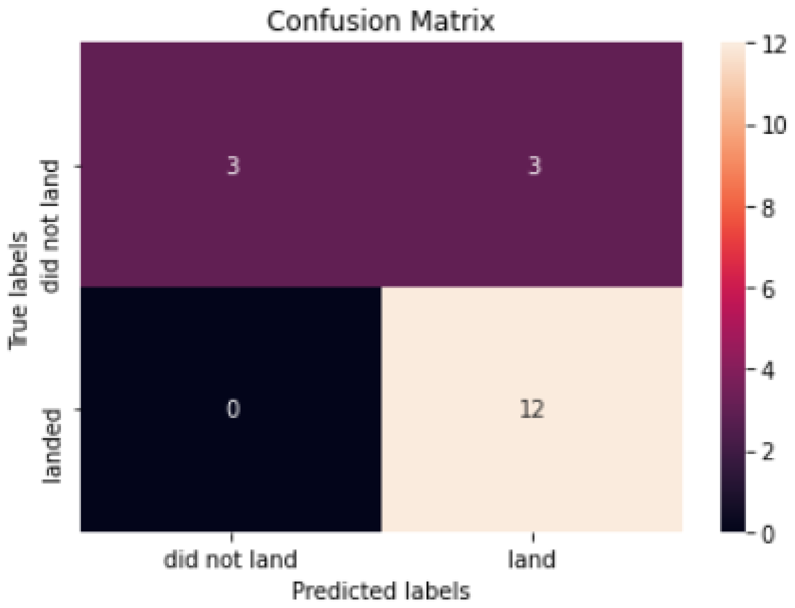
METHODOLOGY Modeling and Classification

- Establish model with best parameters to and classification to success

RESULTS from Classification



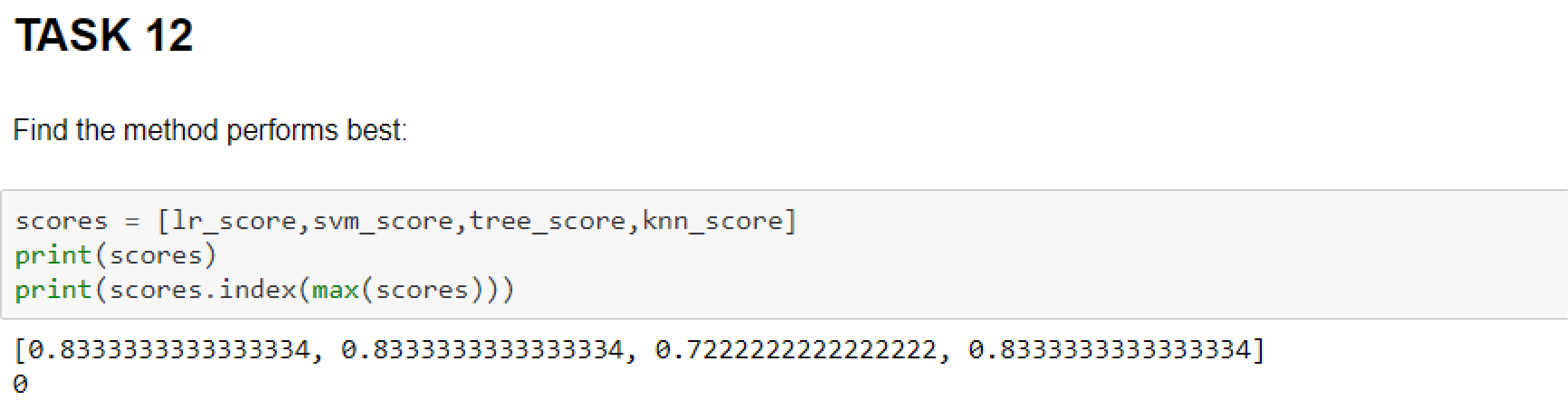
```
yhat=svm_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



- The confusion matrix could predict with high accuracy (83%)

RESULTS from Classification

- The Linear Regression has higher score 83% compares to peers.



TASK 12

Find the method performs best:

```
scores = [lr_score,svm_score,tree_score,knn_score]
print(scores)
print(scores.index(max(scores)))
```

```
[0.8333333333333334, 0.8333333333333334, 0.7222222222222222, 0.8333333333333334]
0
```

CONCLUSION

- Linear Regression has the best model for the dataset
- Successful rate increase vs year
- High successful rate with ESL1 and GEO orbit

APPENDIX