THE GREAT (NEW) SPACE RACE

EXAMINING SPACEX ROCKET LAUNCH DATA

PREPARED BY: CAMERON COOK

FOR: IBM APPLIED DATA SCIENCE CAPSTONE

LAST UPDATED: 7.15.22

OUTLINE

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

EXECUTIVE SUMMARY

- Understanding the success probability percentage of SpaceX rocket launches can be used to generate financial cost analysis models for competing companies
- Using the following data science methods gives us a robust understanding of the data:
  - Data collection via API & Web Scraping
  - Exploratory Data Analysis & Visualization
  - Performing Predictive Analysis
- We found the following trends:
  - As the # of flights increases, success rate (safely landing 1st stage) also increases
  - As payload increases, the average success rate decreases
  - Overall success rate has increased since the beginning of the program

INTRODUCTION

We will predict if the Falcon 9 first stage will land successfully. 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. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

METHODOLOGY

DATA COLLECTION

- Data Collection API
  - Calculate the number of launches on each site
  - Calculate the number and occurrence of each orbit
  - Calculate the number and occurrence of mission outcome per orbit type
  - Create a landing outcome label from Outcome column

- Web Scraping
  - Request the Falcon 9 Launch Wiki Page from URL
  - Extract all column/variable names from the HTML table header
  - Create a data frame by parsing the launch HTML tables

DATA ANALYSIS & VISUALIZATION (STATIC)

- Scatter Plots
  - Flight Number vs. Payload Mass
  - Flight Number vs. Launch Site
  - Flight Number vs. Orbit
  - Payload vs. Orbit

- Bar Chart
  - Success Rate by Launch Site

- Line Chart
  - Success Rate by Year

DATA ANALYSIS & VISUALIZATION (INTERACTIVE)

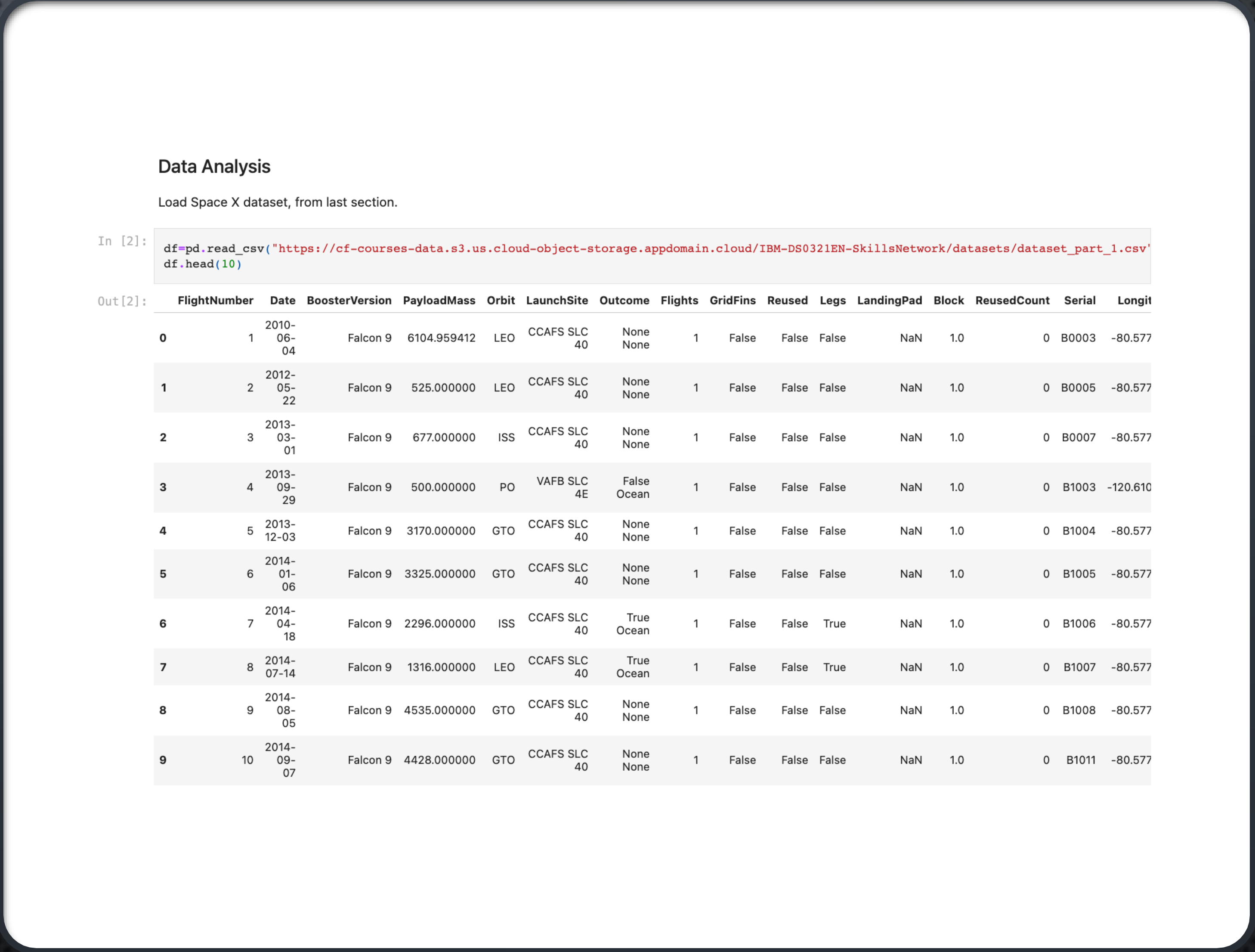
- Interactive Map w/ Folium
  - Showing launch sites
  - Showing relationship between launch sites and significant geographical landmarks (I.e. railroads, highways and coastlines)

- Interactive Dashboard w/ Plotly
  - Launch site dropdown input
  - Success pie chart
  - Range slider to select payload
  - Success payload scatter plot

RESULTS PART 1: DATA COLLECTION

DATA COLLECTION

A sample of the data set imported using the SpaceX .csv data file



Data Analysis

Load Space X dataset, from last section.

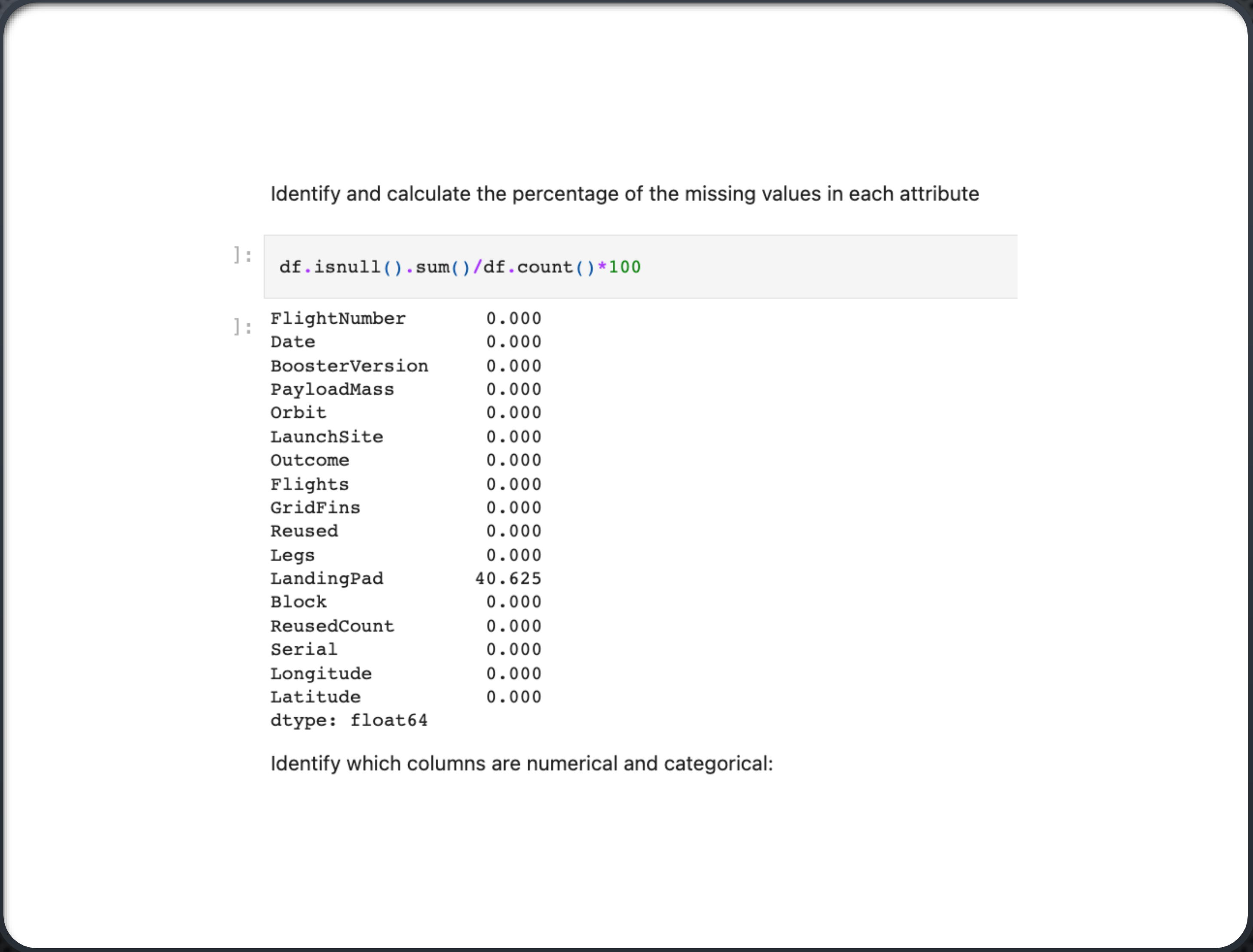
```
In [2]: df=pd.read_csv("https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/dataset_part_1.csv'
        df.head(10)
```

Out[2]:

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longit
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577
5	6	2014-01-06	Falcon 9	3325.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1005	-80.577
6	7	2014-04-18	Falcon 9	2296.000000	ISS	CCAFS SLC 40	True Ocean	1	False	False	True	NaN	1.0	0	B1006	-80.577
7	8	2014-07-14	Falcon 9	1316.000000	LEO	CCAFS SLC 40	True Ocean	1	False	False	True	NaN	1.0	0	B1007	-80.577
8	9	2014-08-05	Falcon 9	4535.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1008	-80.577
9	10	2014-09-07	Falcon 9	4428.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1011	-80.577

DATA COLLECTION

Identify & Calculating missing values in each attribute



Identify and calculate the percentage of the missing values in each attribute

```
]:  df.isnull().sum()/df.count()*100
```

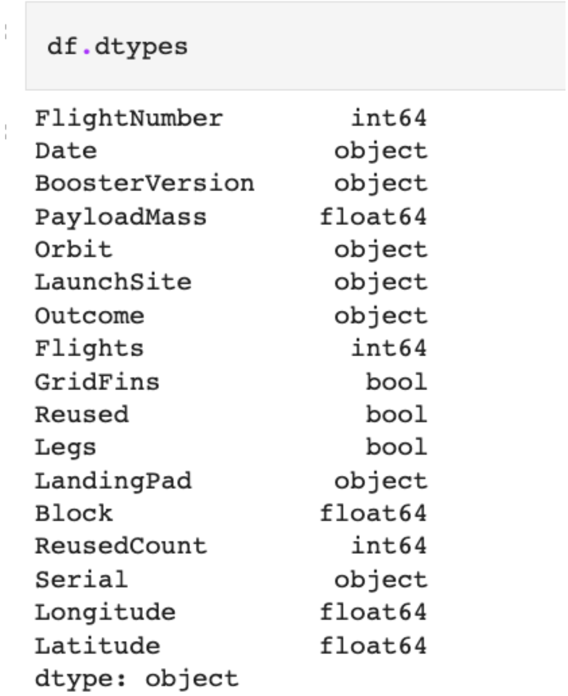
```
]:  FlightNumber       0.000
    Date               0.000
    BoosterVersion     0.000
    PayloadMass        0.000
    Orbit              0.000
    LaunchSite         0.000
    Outcome            0.000
    Flights            0.000
    GridFins           0.000
    Reused             0.000
    Legs               0.000
    LandingPad        40.625
    Block              0.000
    ReusedCount        0.000
    Serial             0.000
    Longitude          0.000
    Latitude           0.000
    dtype: float64
```

Identify which columns are numerical and categorical:

DATA COLLECTION

Identifying

Df.types

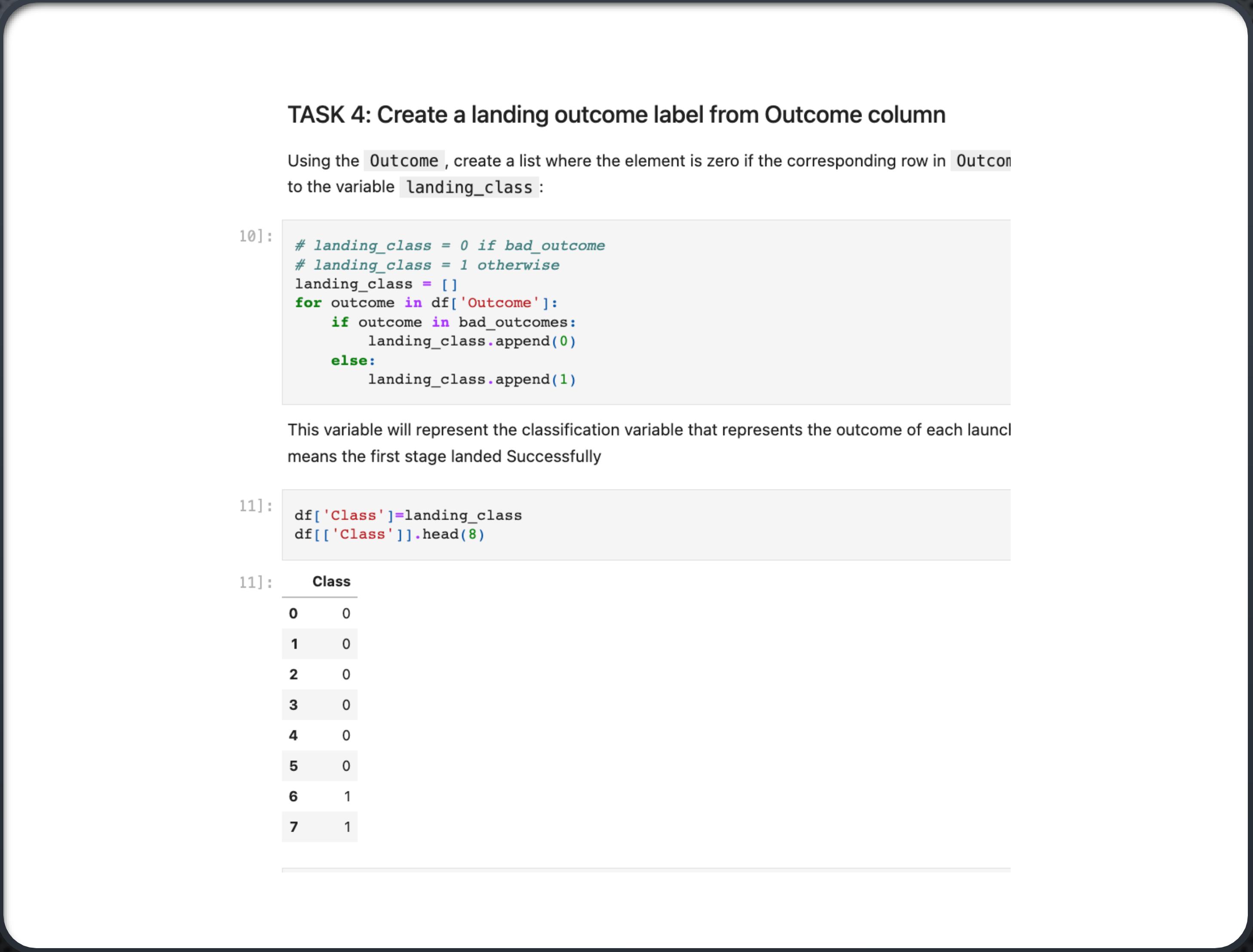


```
df.dtypes
```

```
FlightNumber        int64
Date               object
BoosterVersion     object
PayloadMass       float64
Orbit              object
LaunchSite         object
Outcome            object
Flights             int64
GridFins             bool
Reused               bool
Legs                 bool
LandingPad         object
Block             float64
ReusedCount         int64
Serial             object
Longitude         float64
Latitude          float64
dtype: object
```

DATA COLLECTION

CREATE A LANDING OUTCOME LABEL FROM OUTCOME COLUMN



TASK 4: Create a landing outcome label from Outcome column

Using the `Outcome` , create a list where the element is zero if the corresponding row in `Outcom`
to the variable `landing_class` :

10]:

```
# landing_class = 0 if bad_outcome
# landing_class = 1 otherwise
landing_class = []
for outcome in df['Outcome']:
    if outcome 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 launcl
means the first stage landed Successfully

11]:

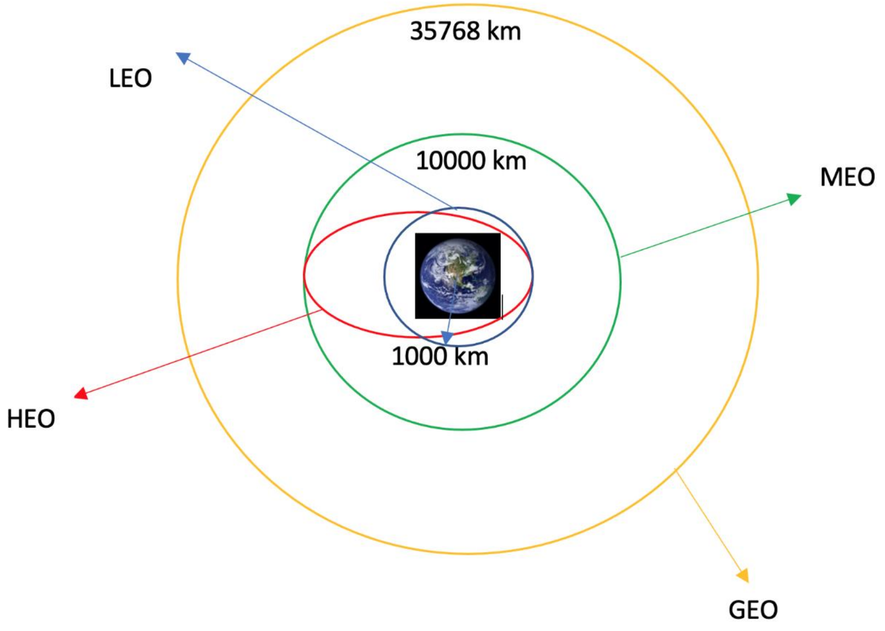
```
df['Class']=landing_class
df[['Class']].head(8)
```

11]:

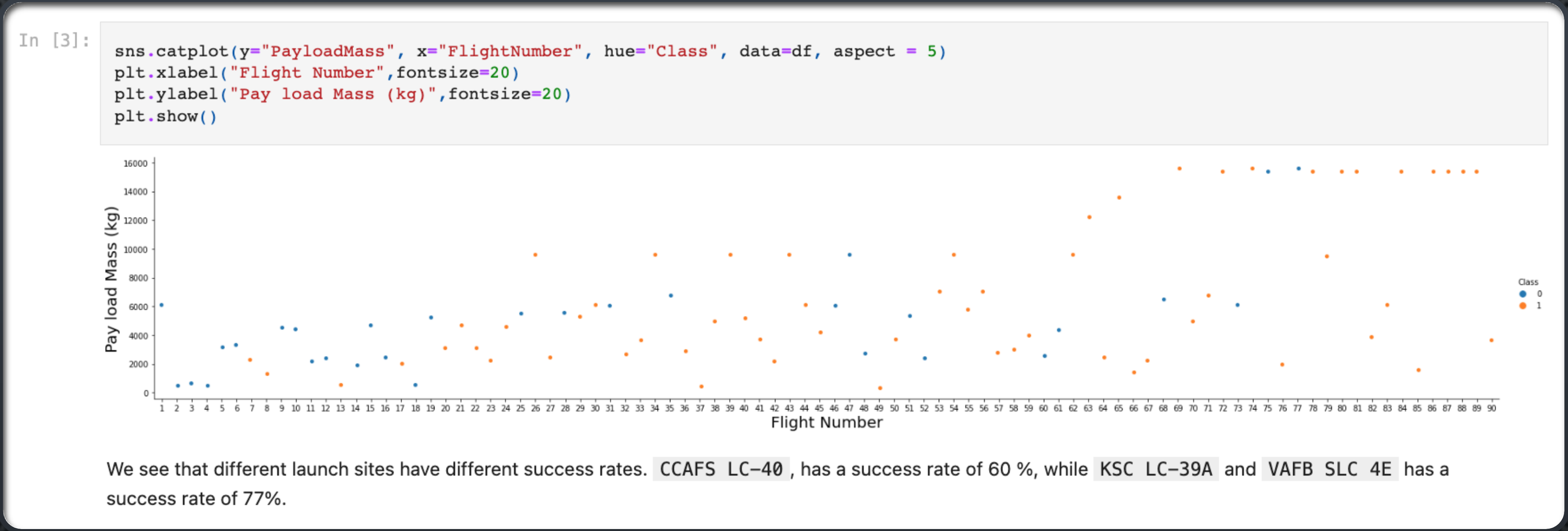
	Class
0	0
1	0
2	0
3	0
4	0
5	0
6	1
7	1

Data Collection

Understanding the different orbit types



RESULTS PART 2: DATA VISUALIZATION

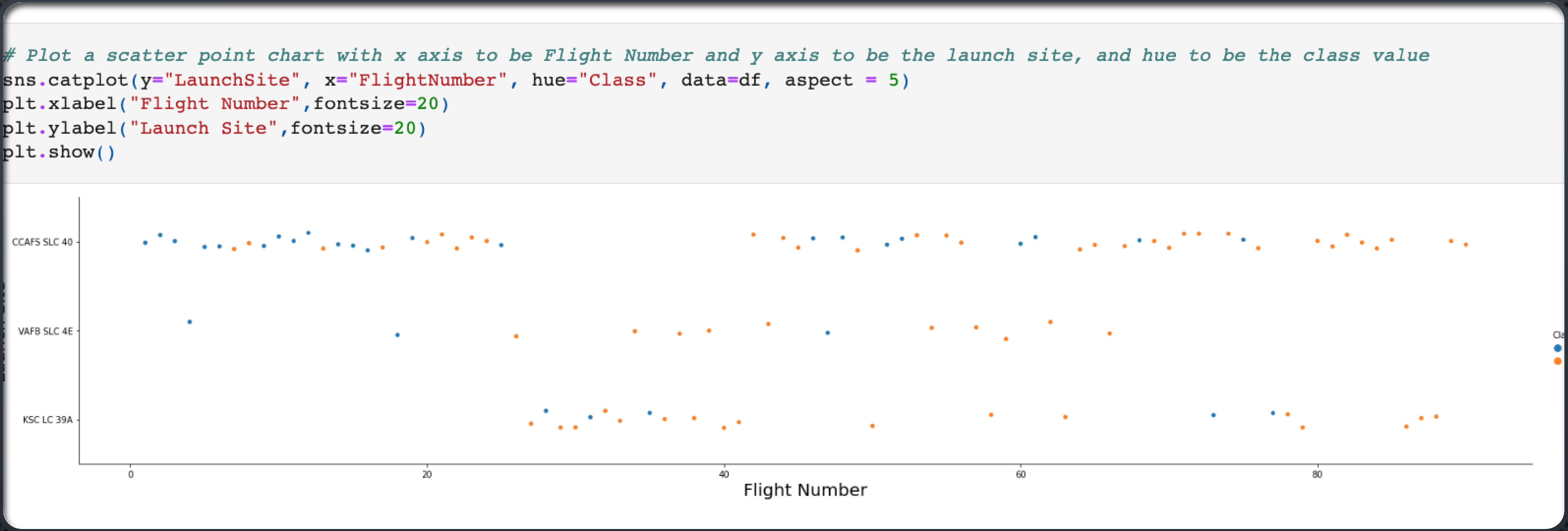


In [3]:

```
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()
```

We see that different launch sites have different success rates. `CCAFS LC-40`, has a success rate of 60 %, while `KSC LC-39A` and `VAFB SLC 4E` has a success rate of 77%.

DATA VISUALIZATION
FLIGHT NUMBER VS. PAYLOAD



```
# Plot a scatter point chart with x axis to be Flight Number and y axis to be the launch site, and hue to be the class value
sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```

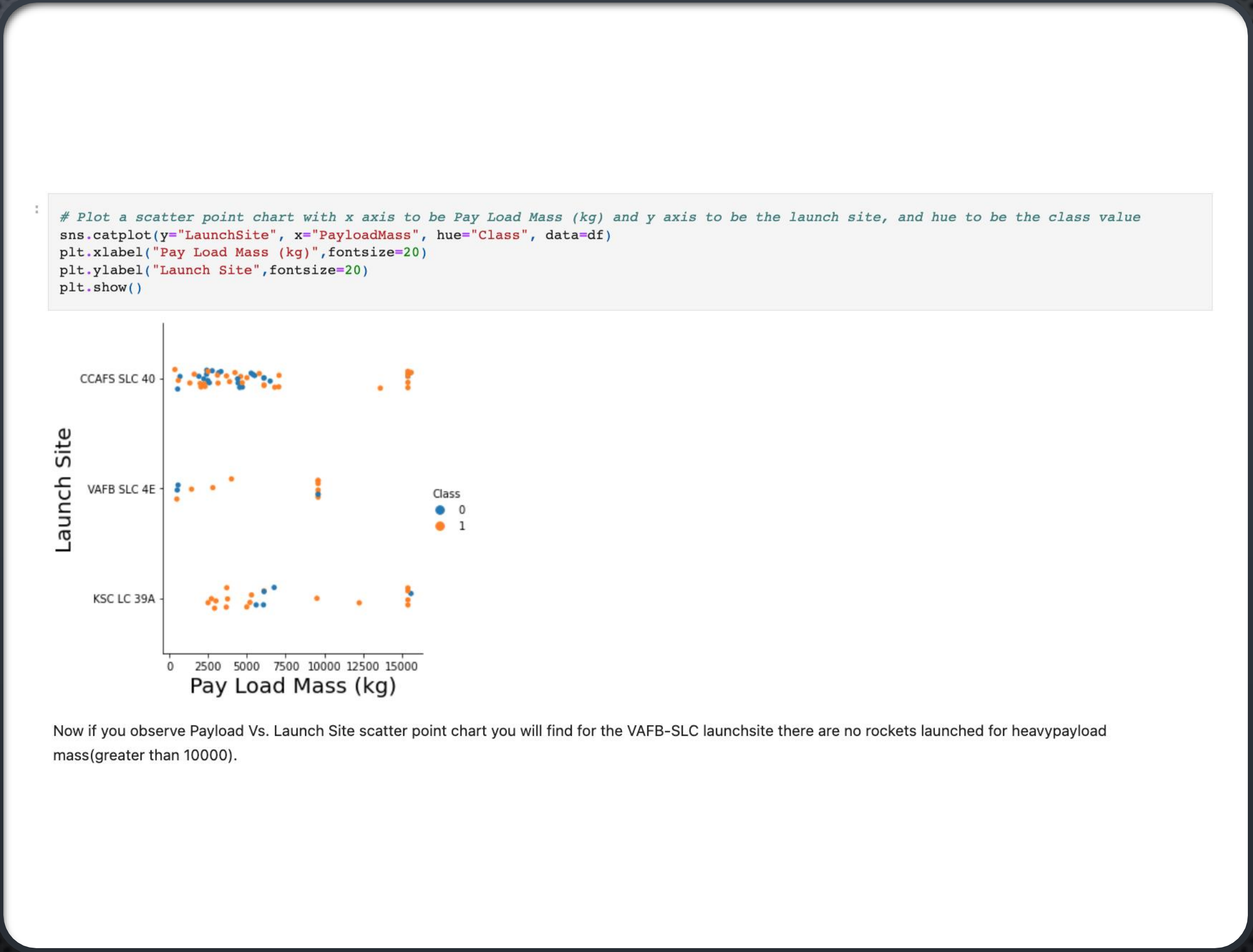
DATA VISUALIZATION
FLIGHT NUMBER VS. LAUNCH SITE

DATA VISUALIZATION

PAYLOAD MASS (KG)

VS.

LAUNCH SITE



```
# 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 value
sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df)
plt.xlabel("Pay Load Mass (kg)",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```

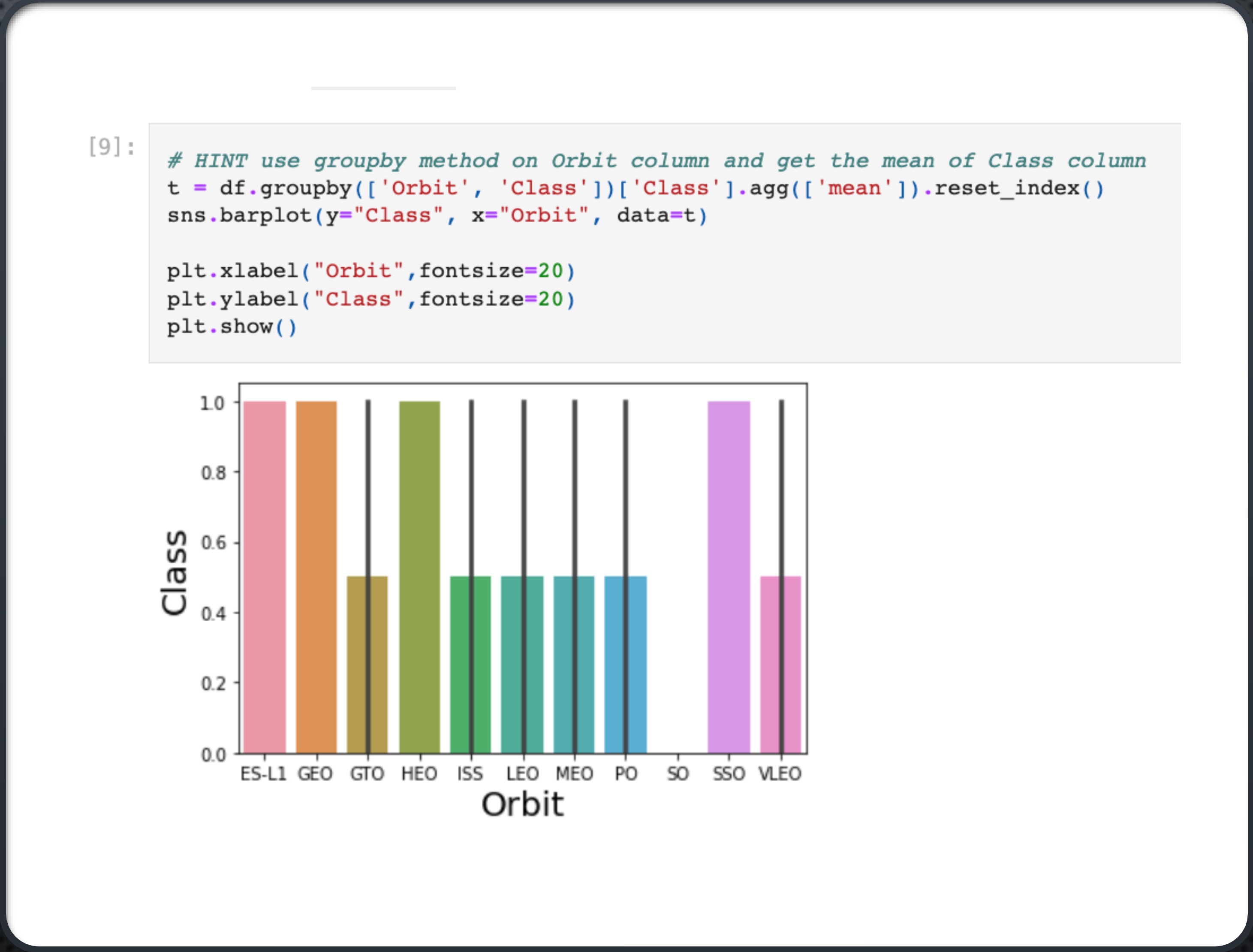
Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

DATA VISUALIZATION

SUCCESS RATE

BY

LAUNCH SITE



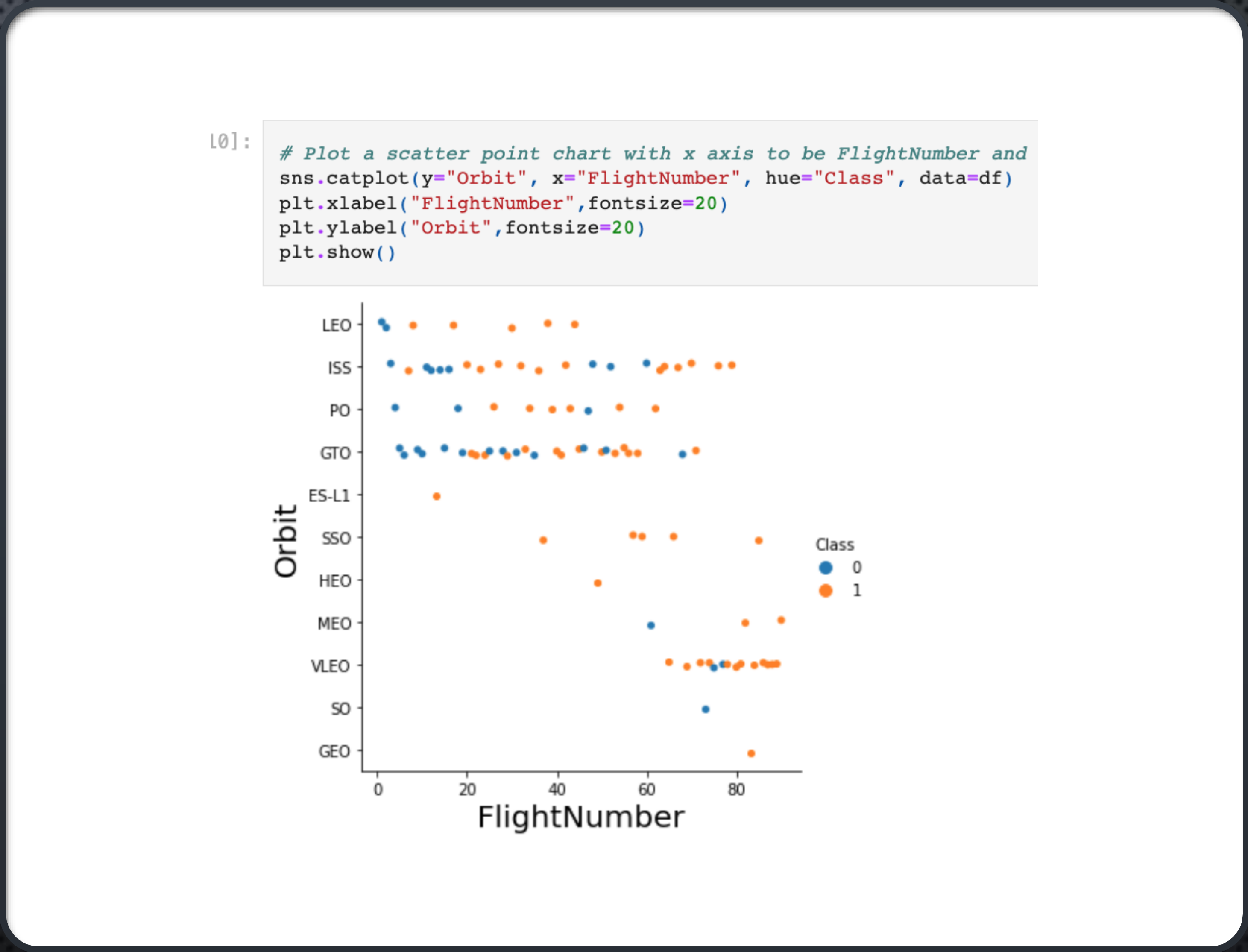
[9]:

```
# HINT use groupby method on Orbit column and get the mean of Class column
t = df.groupby(['Orbit', 'Class'])['Class'].agg(['mean']).reset_index()
sns.barplot(y="Class", x="Orbit", data=t)

plt.xlabel("Orbit",fontsize=20)
plt.ylabel("Class",fontsize=20)
plt.show()
```

DATA VISUALIZATION

FLIGHT NUMBER VS. ORBIT



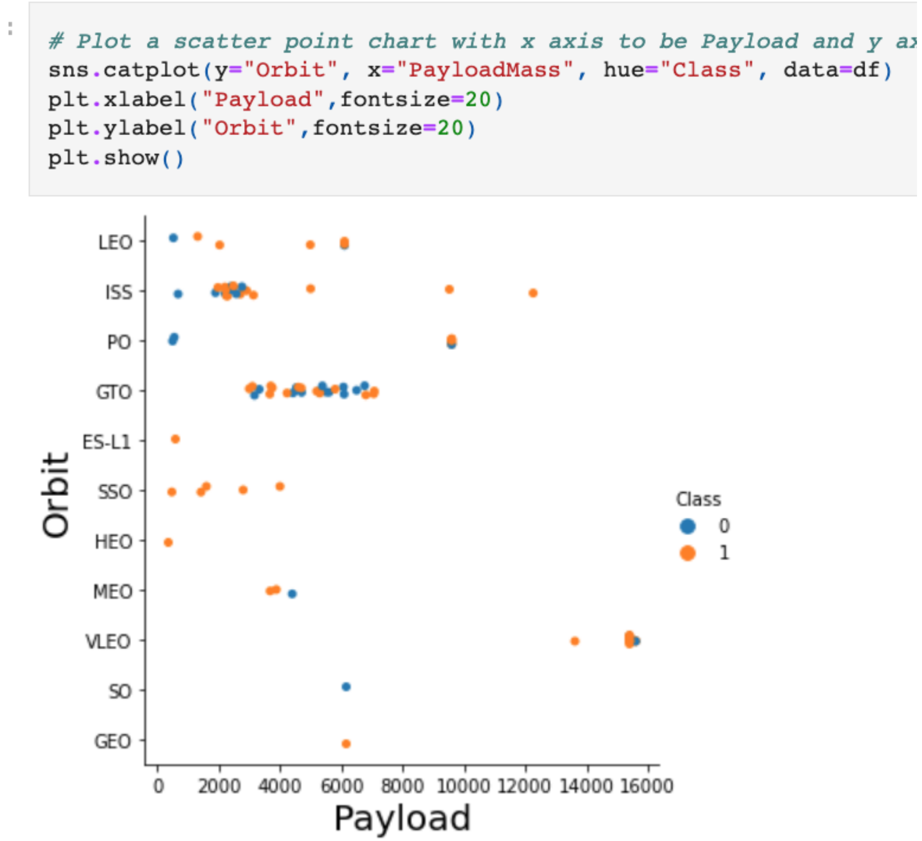
```
L0]:
# Plot a scatter point chart with x axis to be FlightNumber and
sns.catplot(y="Orbit", x="FlightNumber", hue="Class", data=df)
plt.xlabel("FlightNumber",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```

DATA VISUALIZATION

PAYLOAD

VS.

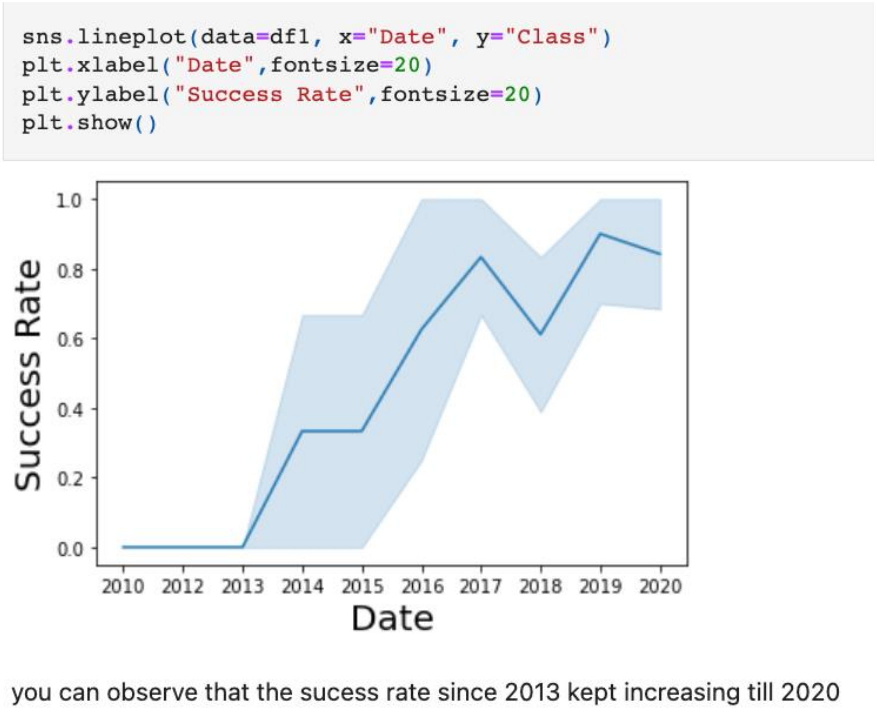
ORBIT



```
# Plot a scatter point chart with x axis to be Payload and y ax
sns.catplot(y="Orbit", x="PayloadMass", hue="Class", data=df)
plt.xlabel("Payload",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```

DATA VISUALIZATION

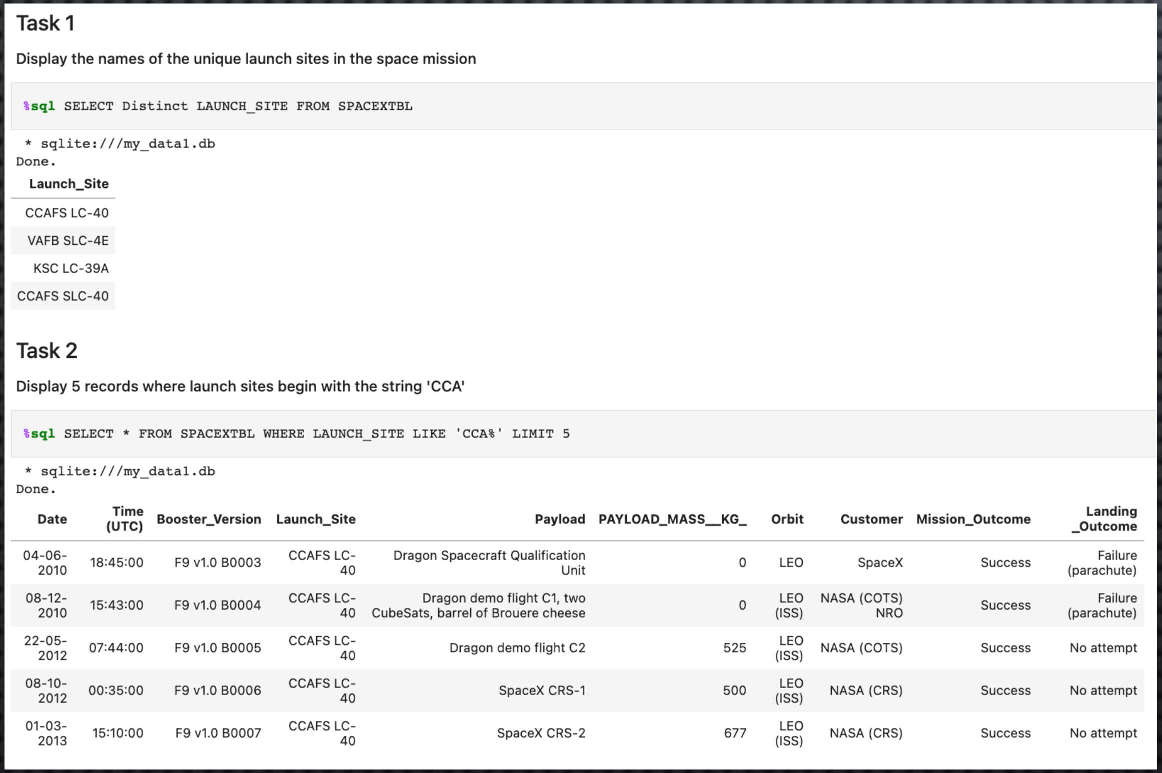
ANNUAL SUCCESS RATE



```
sns.lineplot(data=df1, x="Date", y="Class")
plt.xlabel("Date",fontsize=20)
plt.ylabel("Success Rate",fontsize=20)
plt.show()
```

you can observe that the sucess rate since 2013 kept increasing till 2020

EDA WITH SQL



Task 1

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

```
%sql SELECT Distinct LAUNCH_SITE FROM SPACEXTBL
```

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

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

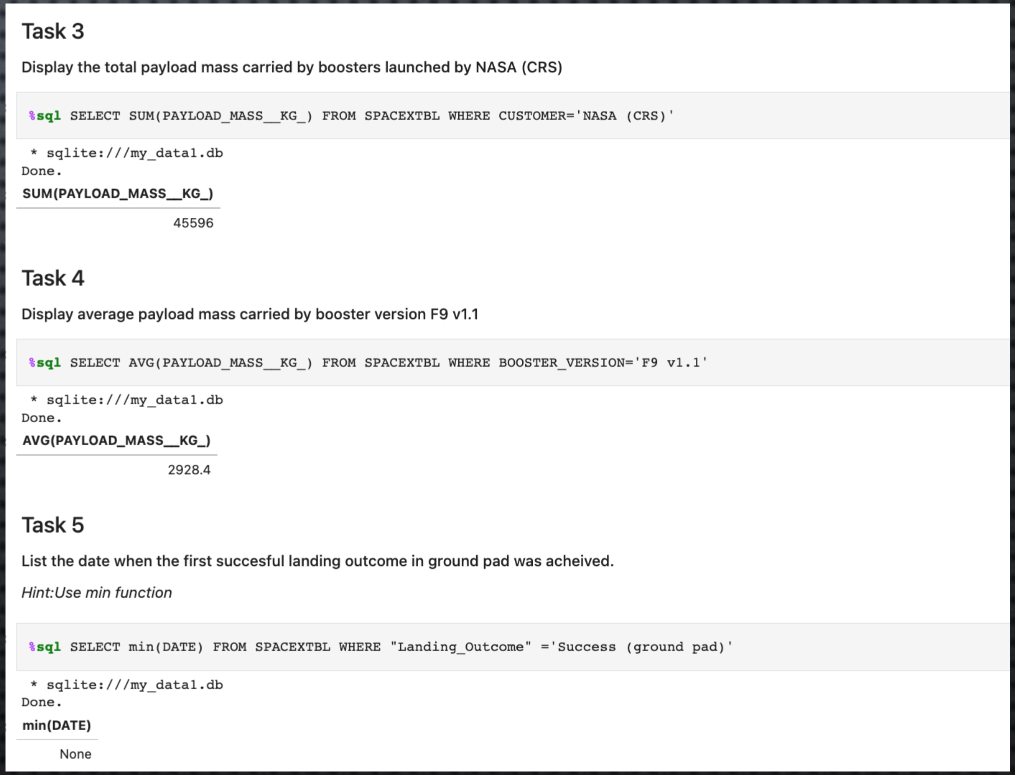
Task 2

Display 5 records where launch sites begin with the string '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
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt



Task 3

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

Task 4

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

Task 5

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

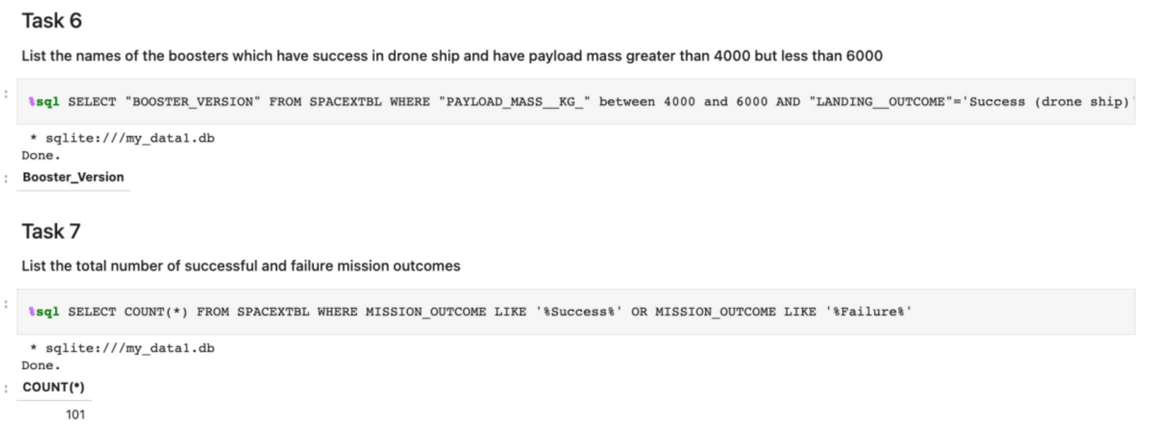
Hint:Use min function

```
%sql SELECT min(DATE) FROM SPACEXTBL WHERE "Landing_Outcome" ='Success (ground pad)'
```

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

min(DATE)
None

EDA WITH SQL



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" FROM SPACEXTBL WHERE "PAYLOAD_MASS__KG_" between 4000 and 6000 AND "LANDING__OUTCOME"='Success (drone ship)
```

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

Booster_Version

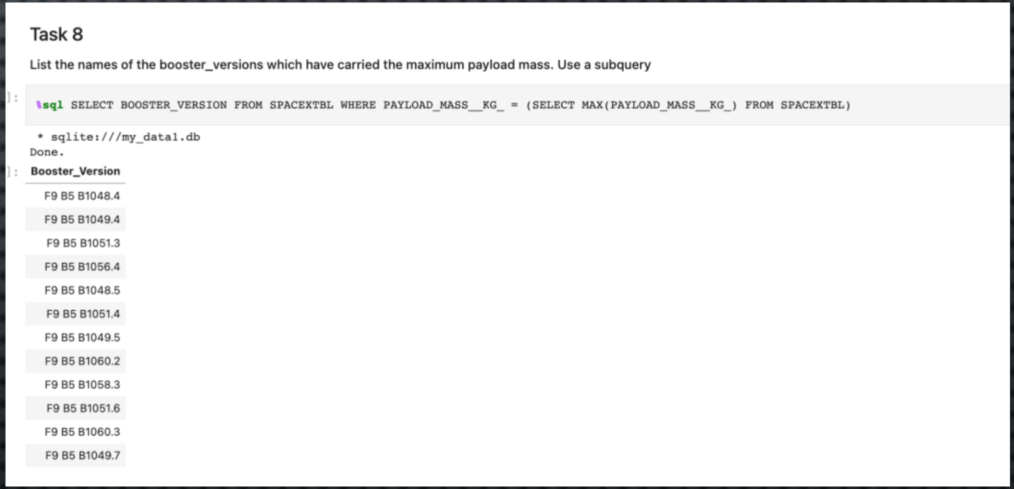
Task 7

List the total number of successful and failure mission outcomes

```
%sql SELECT COUNT(*) FROM SPACEXTBL WHERE MISSION_OUTCOME LIKE '%Success%' OR MISSION_OUTCOME LIKE '%Failure%'
```

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

COUNT(*)
101



Task 8

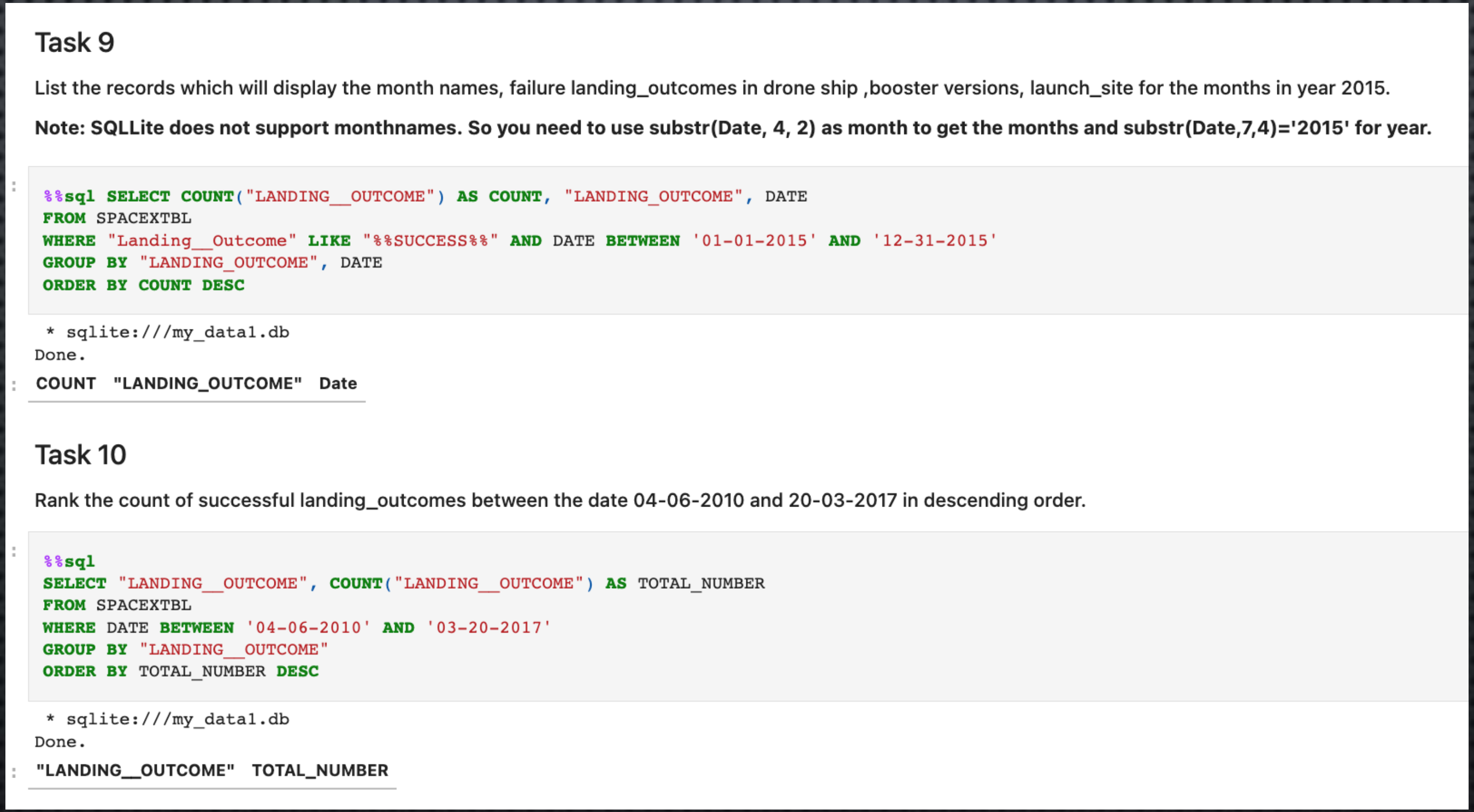
List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%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

EDA WITH SQL



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, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

```
%%sql SELECT COUNT("LANDING__OUTCOME") AS COUNT, "LANDING_OUTCOME", DATE
FROM SPACEXTBL
WHERE "Landing__Outcome" LIKE "%%SUCCESS%%" AND DATE BETWEEN '01-01-2015' AND '12-31-2015'
GROUP BY "LANDING_OUTCOME", DATE
ORDER BY COUNT DESC
```

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

COUNT	"LANDING_OUTCOME"	Date

Task 10

Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

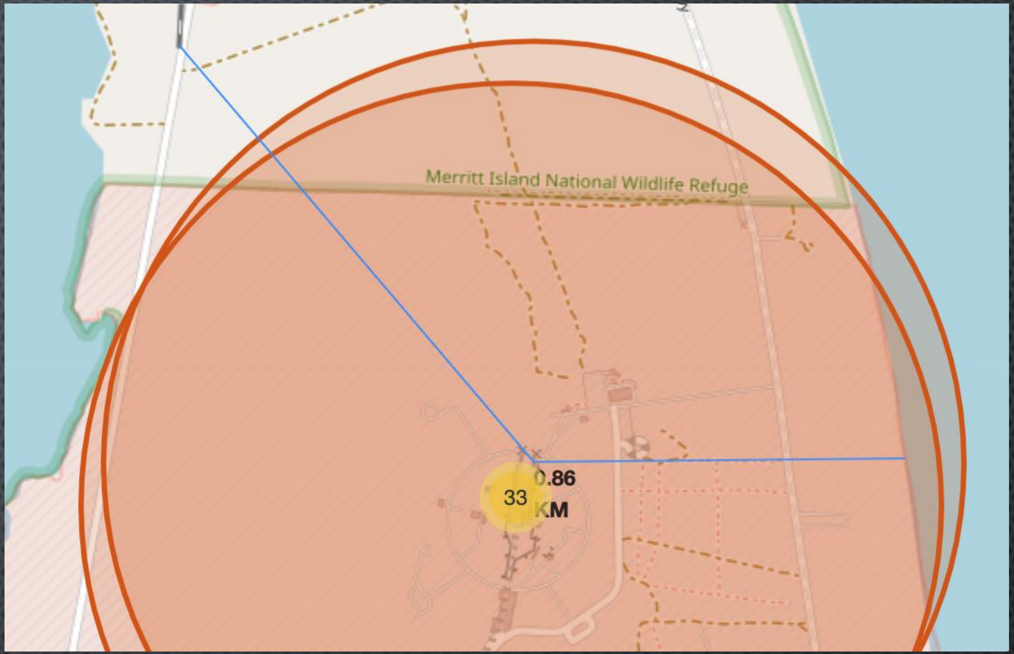
```
%%sql
SELECT "LANDING__OUTCOME", COUNT("LANDING__OUTCOME") AS TOTAL_NUMBER
FROM SPACEXTBL
WHERE DATE BETWEEN '04-06-2010' AND '03-20-2017'
GROUP BY "LANDING__OUTCOME"
ORDER BY TOTAL_NUMBER DESC
```

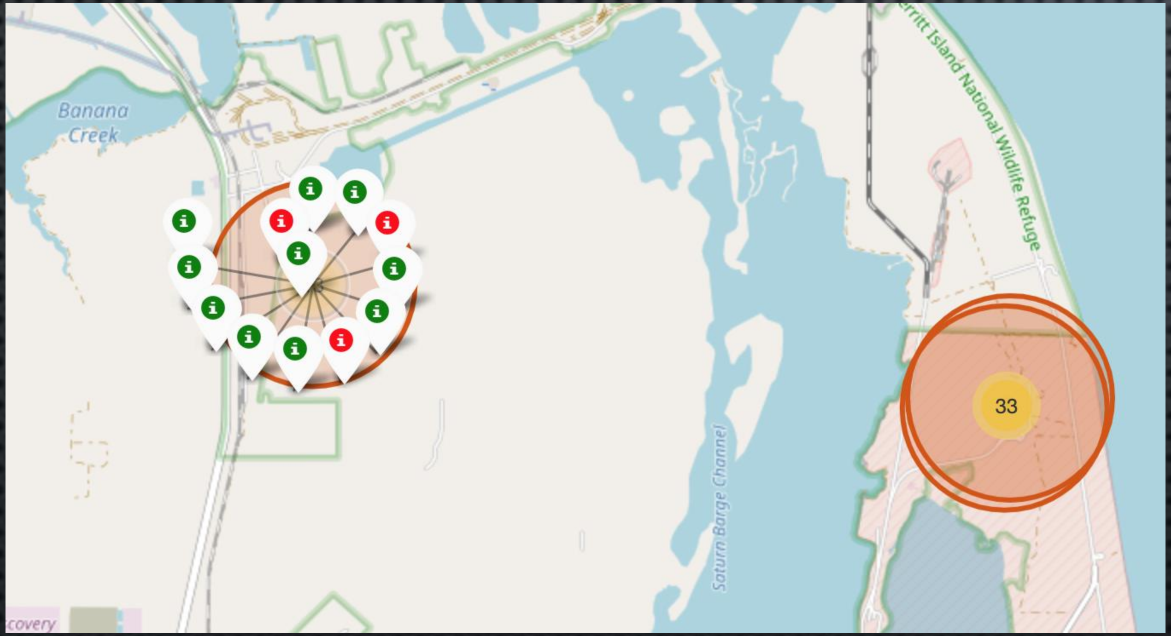
```
 * sqlite:///my_data1.db
Done.
```

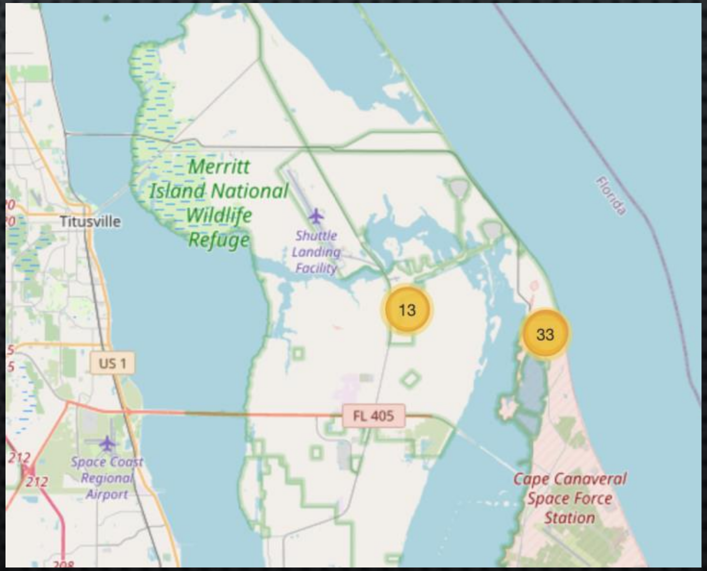
"LANDING__OUTCOME"	TOTAL_NUMBER

RESULTS PART 3: INTERACTIVE VISUALIZATION

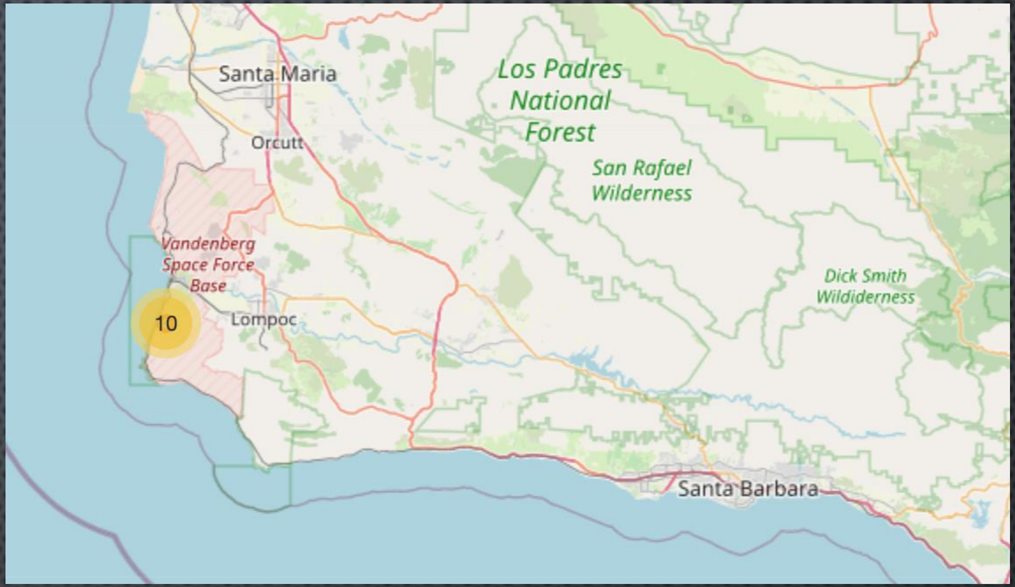
FLORIDA LAUNCH SITES

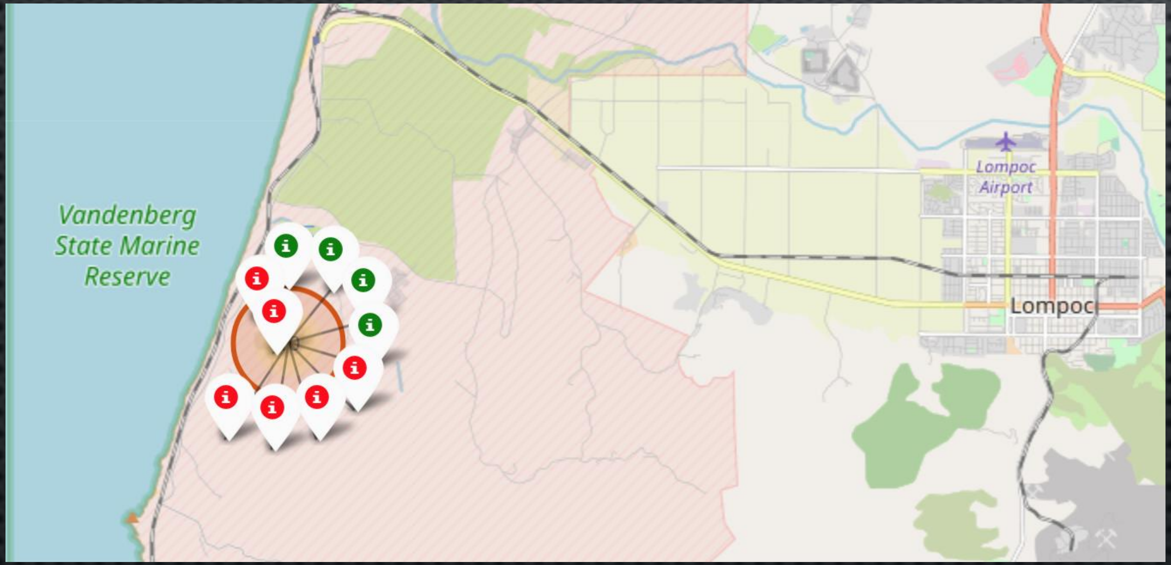


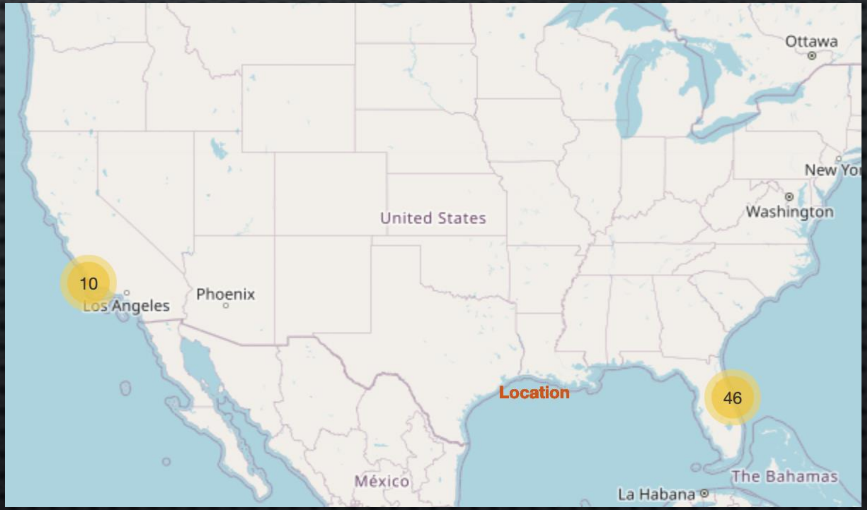



Merritt Island National Wildlife Refuge
Titusville
Shuttle Landing Facility
13
33
Florida
US 1
FL 405
Space Coast Regional Airport
Cape Canaveral Space Force Station
212

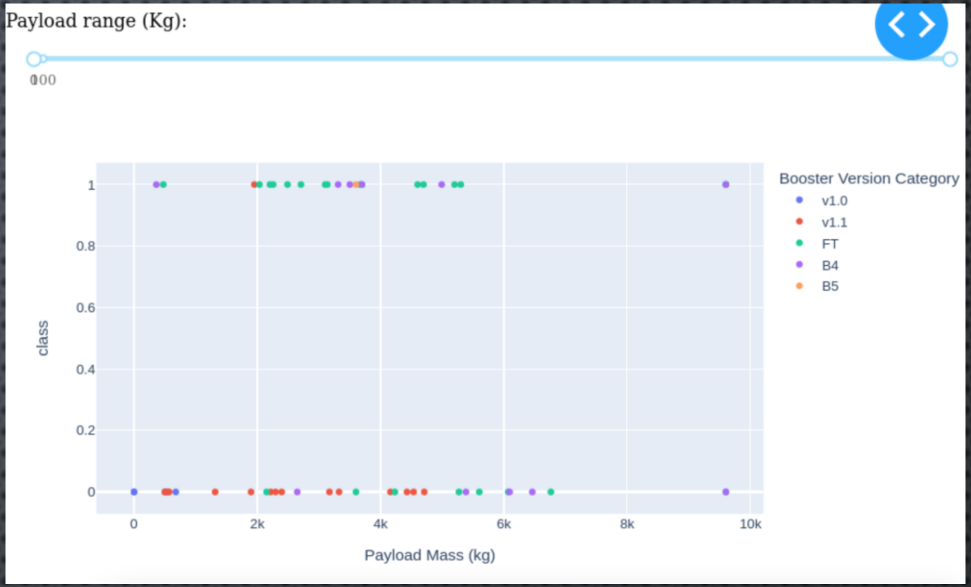
CALIFORNIA LAUNCH SITES



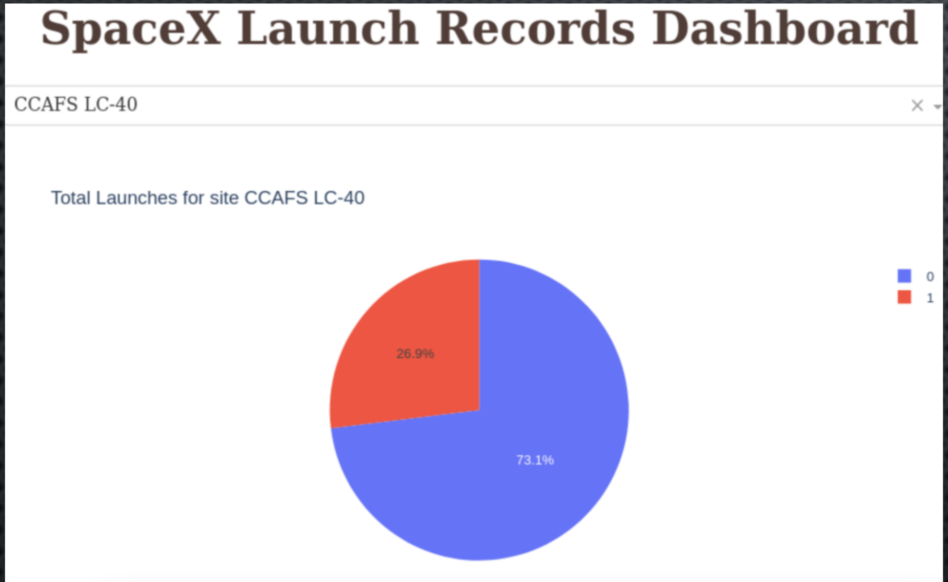




PLOTLY DASHBOARD

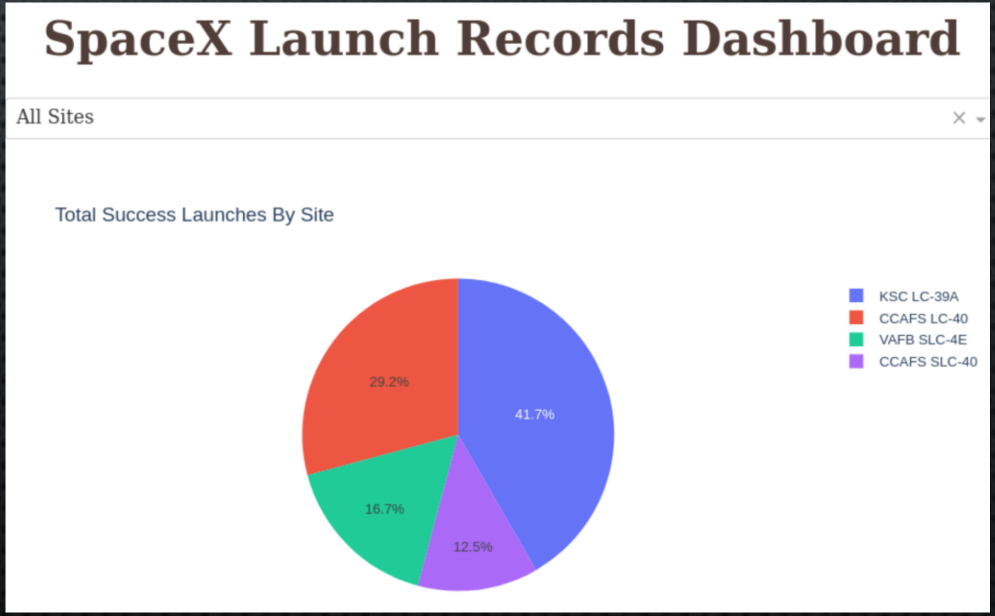


Payload range (Kg):



SpaceX Launch Records Dashboard

CCAFS LC-40



SpaceX Launch Records Dashboard

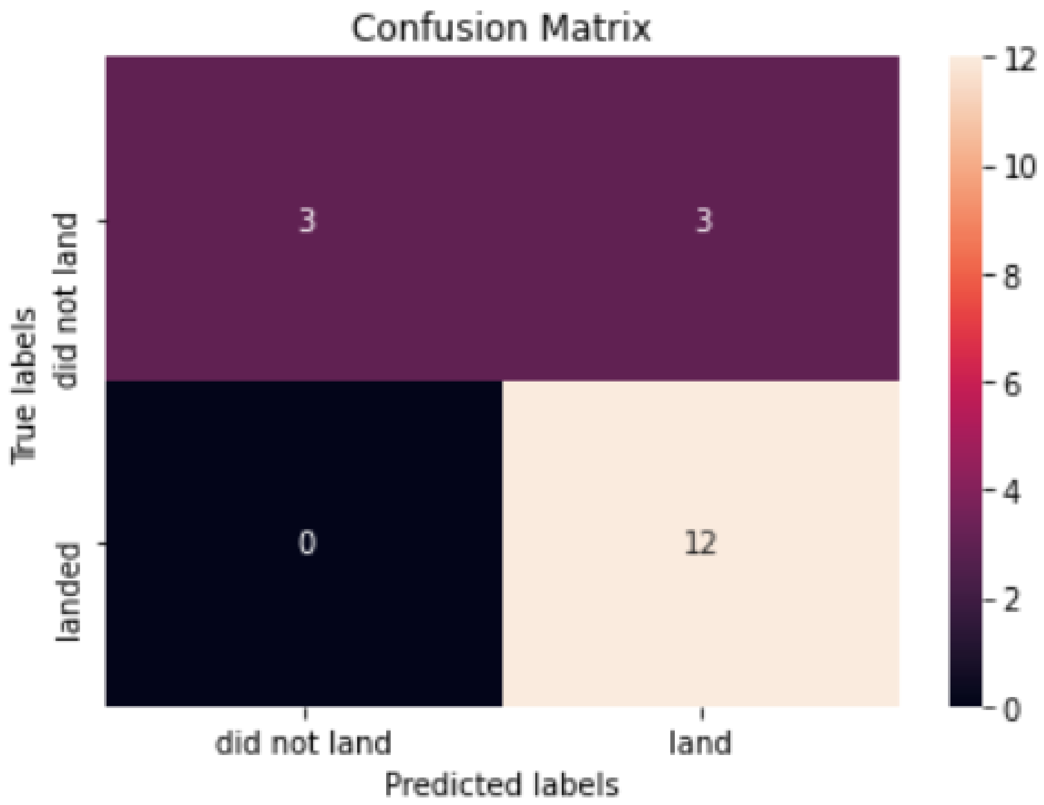
All Sites

RESULTS PART 4: PREDICTIVE ANALYSIS

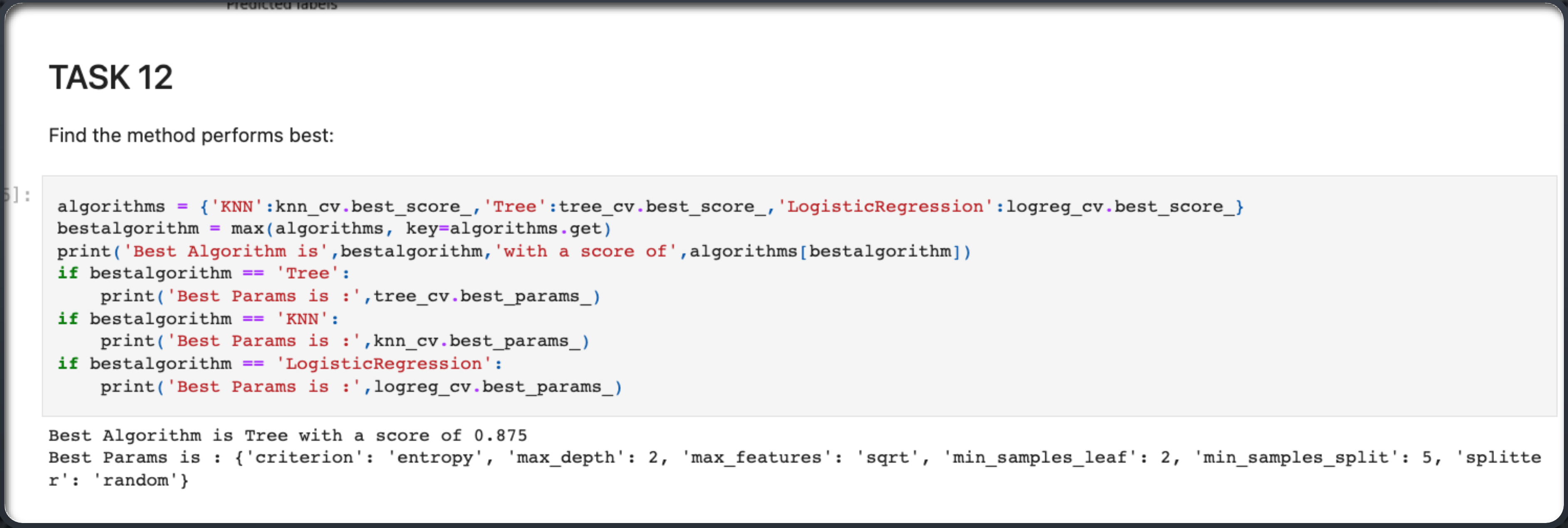
PREDICTIVE ANALYSIS

MACHINE LEARNING CONFUSION MATRIX SHOWING:

- 3 TRUE NEGATIVES
- 3 FALS NEGATIVES
- 0 FALSE POSITIVES
- 12 TRUE POSITIVES



PREDICTIVE ANALYSIS



Predicted labels

TASK 12

Find the method performs best:

```
algorithms = {'KNN':knn_cv.best_score_,'Tree':tree_cv.best_score_,'LogisticRegression':logreg_cv.best_score_}
bestalgorithm = max(algorithms, key=algorithms.get)
print('Best Algorithm is',bestalgorithm,'with a score of',algorithms[bestalgorithm])
if bestalgorithm == 'Tree':
    print('Best Params is :',tree_cv.best_params_)
if bestalgorithm == 'KNN':
    print('Best Params is :',knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best Params is :',logreg_cv.best_params_)
```

```
Best Algorithm is Tree with a score of 0.875
Best Params is : {'criterion': 'entropy', 'max_depth': 2, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 5, 'splitte
r': 'random'}
```

It was determined that the best machine learning model was the decision tree classifier

CONCLUSION

Final Thoughts

- Currently, SpaceX is capable of launching 2.6 rockets for everyone 1 rocket of competing companies.
- It can be assumed that the success rate of SpaceX rockets will continue to increase over time, widening the potential gap between them and their competitors
- There may be a correlation between payload and success rate the warrants further examination. However there are still too many variables not included in this study including:
  - Environmental stats including wind, temperature, barometric pressure, season, etc.
  - Enhancements and modifications to the operating systems of the rocket in between launches
- Final Conclusion: We have built a useful tool to predict the success probability of the falcon9 rocket returning safely to be reused. This information could be used to build a financial forecast for a startup company. However, it should be assumed that they will have to catch up with SpaceX first and that can only come with experience and failed launches of their own.