Applied Data Science Capstone - SpaceX Historical Data Study

Yiwen Zhou

2021 Dec 26

OUTLINE

- Completed the required Executive Summary slide (1 pt)
- Completed the required Introduction slide (1 pt)
- Completed the required data collection and data wrangling methodology related slides (1 pt)
- Completed the required EDA and interactive visual analytics methodology related slides (3 pts)
- Completed the required predictive analysis methodology related slides (1 pt)
- Completed the required EDA with visualization results slides (6 pts)
- Completed the required EDA with SQL results slides (10 pts)
- Completed the required interactive map with Folium results slides (3 pts)
- Completed the required Plotly Dash dashboard results slides (3 pts)
- Completed the required predictive analysis (classification) results slides (6 pts)
- Completed the required Conclusion slide (1 pts)
- Applied your creativity to improve the presentation beyond the template (1 pts)
- Displayed any innovative insights (1 pts)

Completed the required Executive Summary slide (1 pt)
Completed the required Introduction slide (1 pt)

EXECUTIVE SUMMARY and INTRODUCTION

- Overall historical data is not enough to have too many statistically significant conclusions.
- But we do have a good view of historical trend SpaceX launches and landings:
  - Over the time success rate of landing got much improved;
  - Over the time SpaceX has been exploring more new orbits;
  - Over the time SpaceX is capable for more payload mass.
- We have various nice visualizations of historical data:
  - Various static plots and tables using plotly or seaborn;
  - Interactive statistical plots using dash and geographical plots using folium.
- Predictions of landing outcome has accuracy of 85%.

Data collection and data wrangling methodology

- Use "requests.get(spacex_url)" to retrive data and tables from website;
- Turn the above data into a pandas data frame to manipulate data. For example, the "groupby" method is frequently used;
- Can also use "BeautifulSoup" to parse the web information;
- "ipython-sql" makes it easier to do SQL operations in python jupyter notebook;
- Use "seaborn", "plotly.express" or simply "matplotlib.pyplot" to plot figures in order to visualize data;
- Use "dash" to make interactive data visualizations.

Completed the required data collection and data wrangling methodology related slides (1 pt)

Completed the required data collection and data wrangling methodology related slides (1 pt)

SpaceX Data collection and data wrangling results

The following are the occurrence of each orbit lauched for SpaceX:

GTO	27
ISS	21
VLEO	14
PO	9
LEO	7
SSO	5
MEO	3
SO	1
HEO	1
GEO	1
ES-L1	1

The following are the landing outcomes of SpaceX:

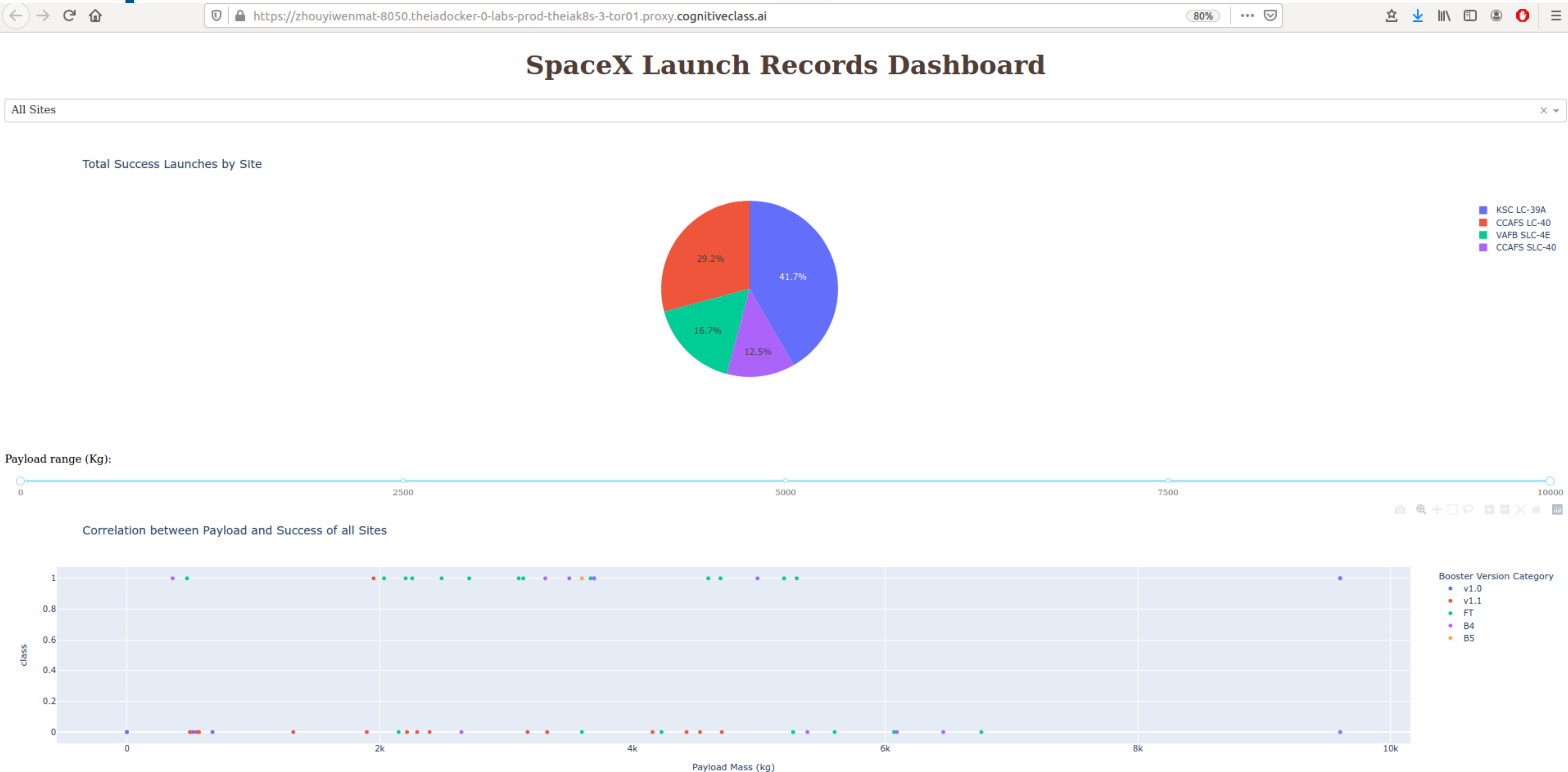
True ASDS	41
None None	19
True RTLS	14
False ASDS	6
True Ocean	5
None ASDS	2
False Ocean	2
False RTLS	1

The following are the Launch Sites of SpaceX:

CCAFS SLC 40	55
KSC LC 39A	22
VAFB SLC 4E	13

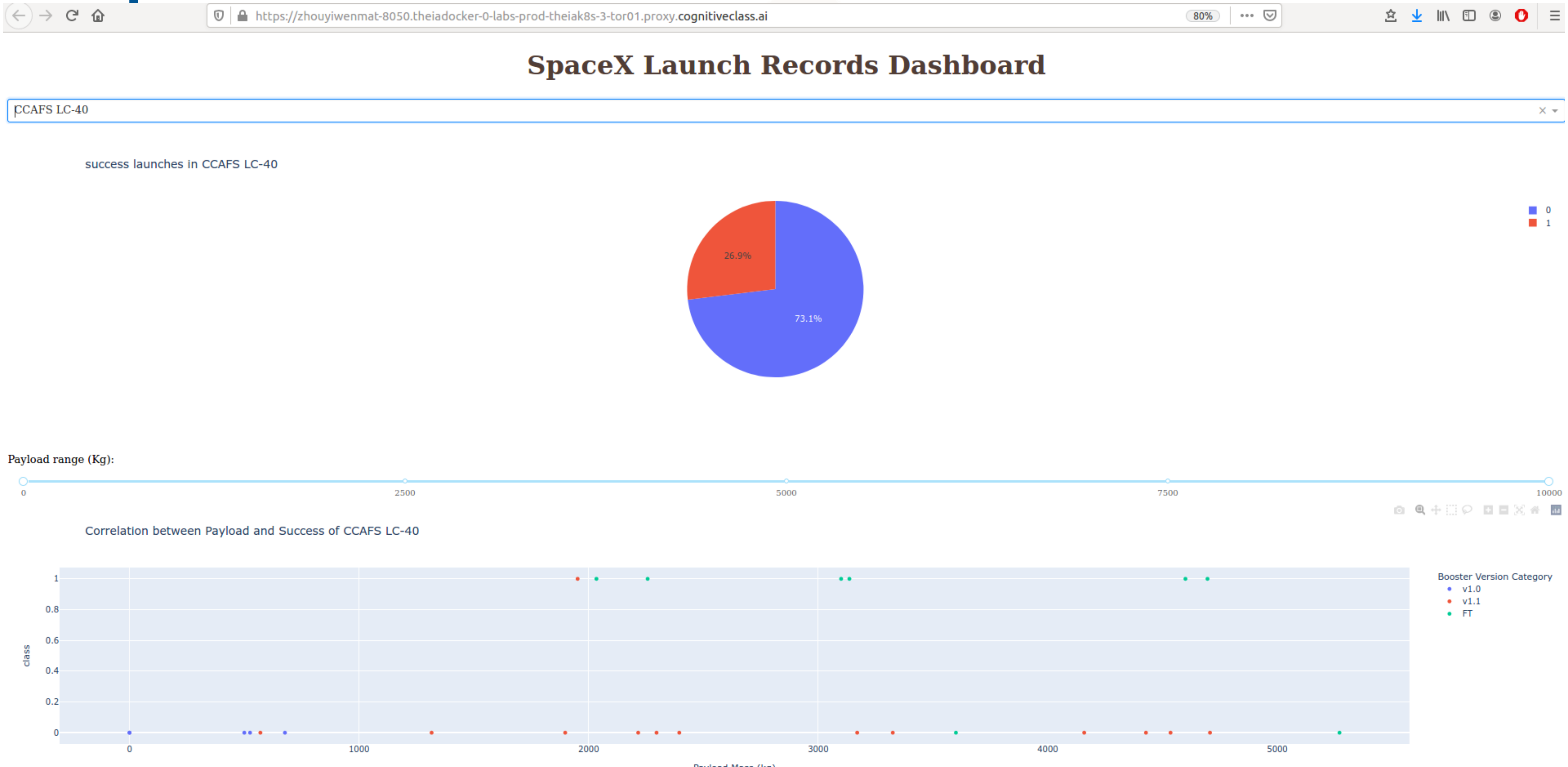
Completed the required EDA and interactive visual analytics methodology related slides (3 pts)
Completed the required Plotly Dash dashboard results slides (3 pts)

SpaceX Interactive Visualizations



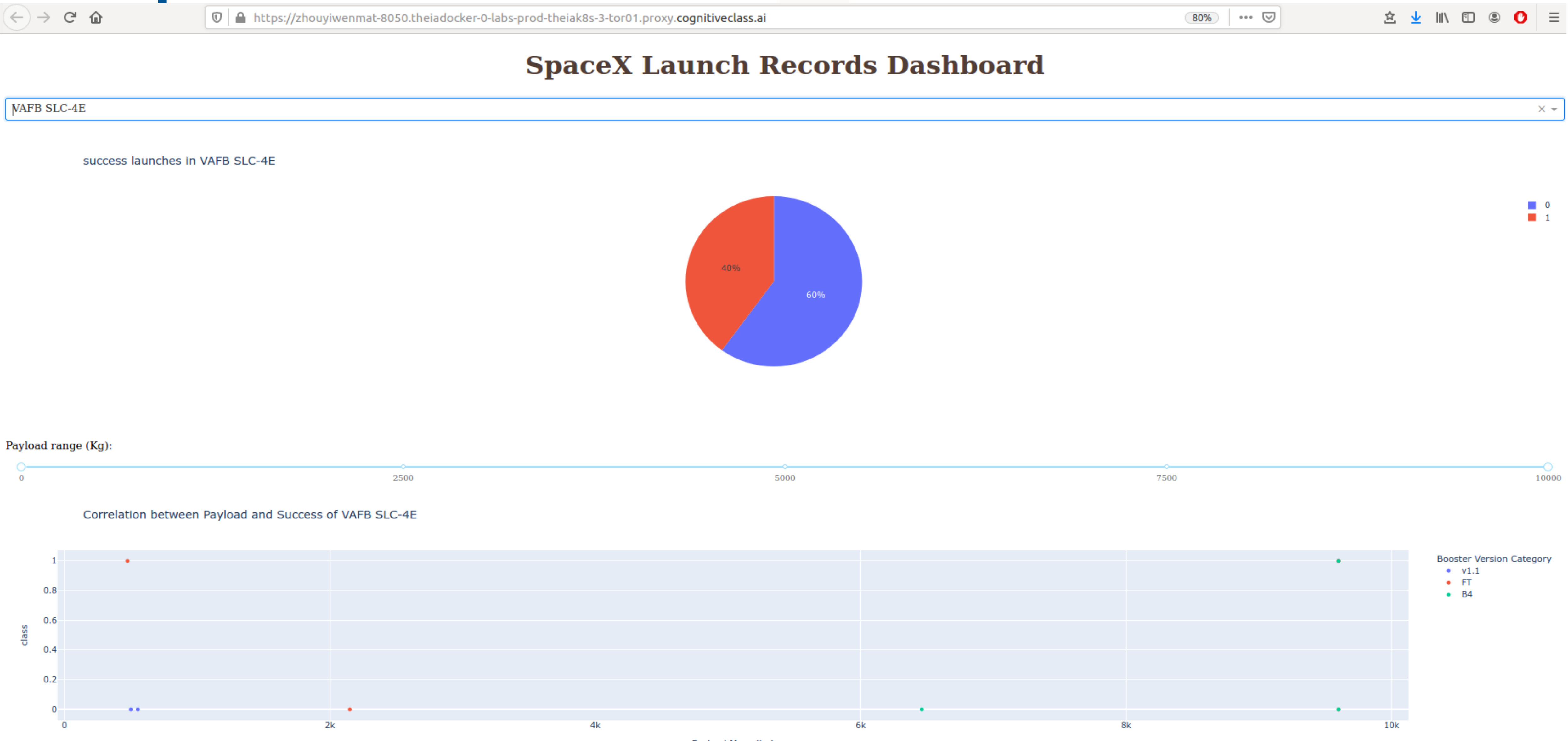
Completed the required EDA and interactive visual analytics methodology related slides (3 pts)
Completed the required Plotly Dash dashboard results slides (3 pts)

SpaceX Interactive Visualizations



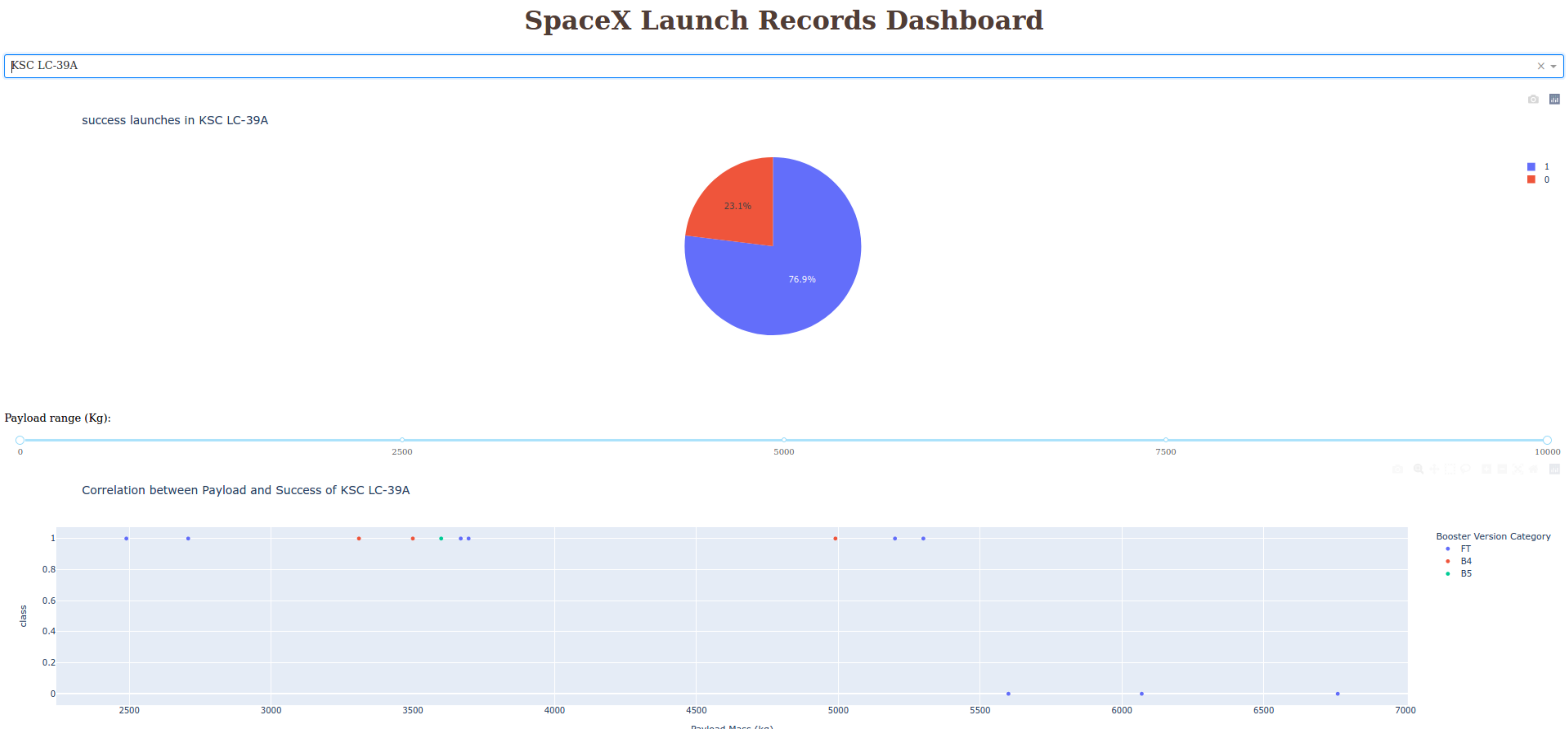
Completed the required EDA and interactive visual analytics methodology related slides (3 pts)
Completed the required Plotly Dash dashboard results slides (3 pts)

SpaceX Interactive Visualizations



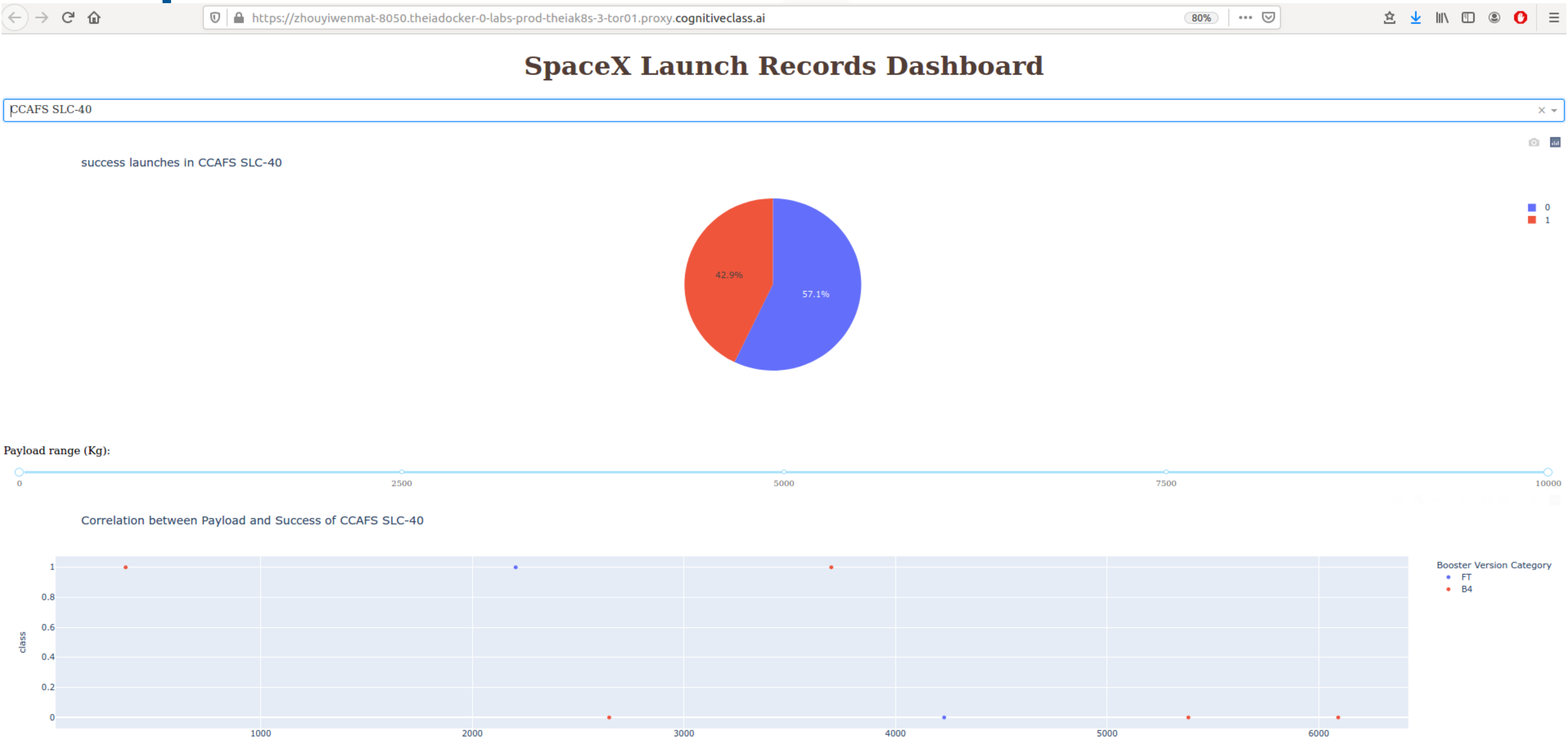
Completed the required EDA and interactive visual analytics methodology related slides (3 pts)
Completed the required Plotly Dash dashboard results slides (3 pts)

SpaceX Interactive Visualizations



Completed the required EDA and interactive visual analytics methodology related slides (3 pts)
Completed the required Plotly Dash dashboard results slides (3 pts)

SpaceX Interactive Visualizations



Completed the required predictive analysis methodology related slides (1 pt)
Completed the required predictive analysis (classification) results slides (6 pts)

Predictive Analysis Methodology and results

Goal: Predict whether a landing of Space X will be successful or not.

Method	Logistic Regression	Support Vector Machine	Decision Tree	K-Means
Best hyper-parameters	C = 0.01, penalty = l2, solver = lbfgs	C=1.0, gamma = 0.0316, kernel = sigmoid	criterion = gini, max_depth = 6, max_features = sqrt, min_samples_leaf = 4	algorithm = auto, n_neighbors = 10, p=1
Train set accuracy	0.846	0.848	0.891	0.848
Test set Score	0.833	0.833	0.889	0.833
False positive rate	0.167	0.167	0.111	0.167
False negative rate	0.0	0.0	0.0	0.0
Conclusion:			Best performance	

Common problems: False positive rate is all too high.

Completed the required EDA with visualization results slides (6 pts)

EDA with visualization results

Exploratory Data Analysis results:

- More recent flights (with larger flight number) have been carrying more and more pay load mass, and success rate is also higher;

- Most of the unsuccessful landings were earlier flight numbers, which were mostly launched from CCAFS SLC 40;
- From VAFB SLC 4E never launched more than 10,000 pay load mass;

- SSO has highest success landing rate (100%); VLEO also has high success rate (85.7%); GTO / ISS / LEO / MEO / PO have similar success rate (50.0% ~ 70.0%);
- Latest (more recent) flight numbers have been trying new variety of orbits;
- Some orbits (like LEO / GTO / SSO) seem to have relatively consistent pay load masses;

- Success rate has been improved a lot in recent years. Current success rate is about 85%.

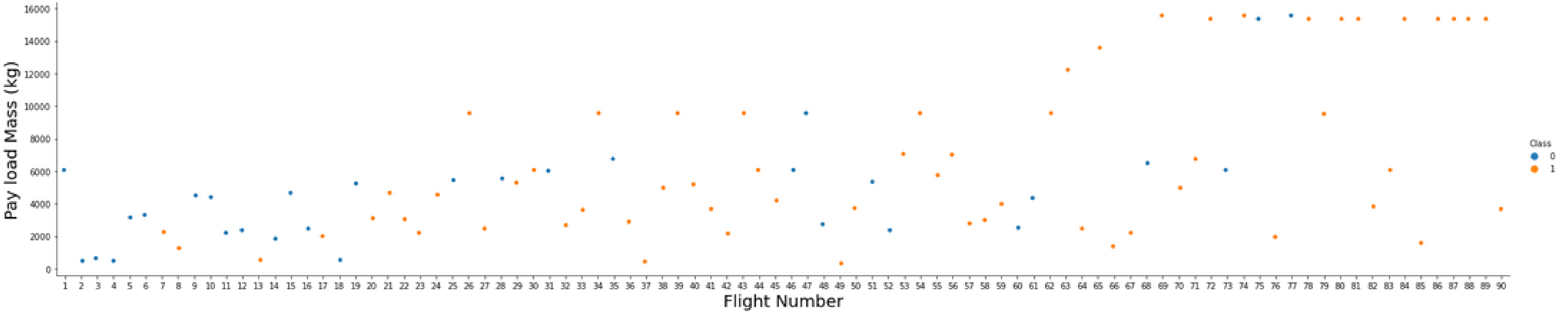
In the next few pages we show some plots supporting the above conclusions:

Completed the required EDA with visualization results slides (6 pts)

EDA with visualization results

1. More recent flights (with larger flight number) have been carrying more and more pay load mass, and success rate gets higher:

```
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()
```

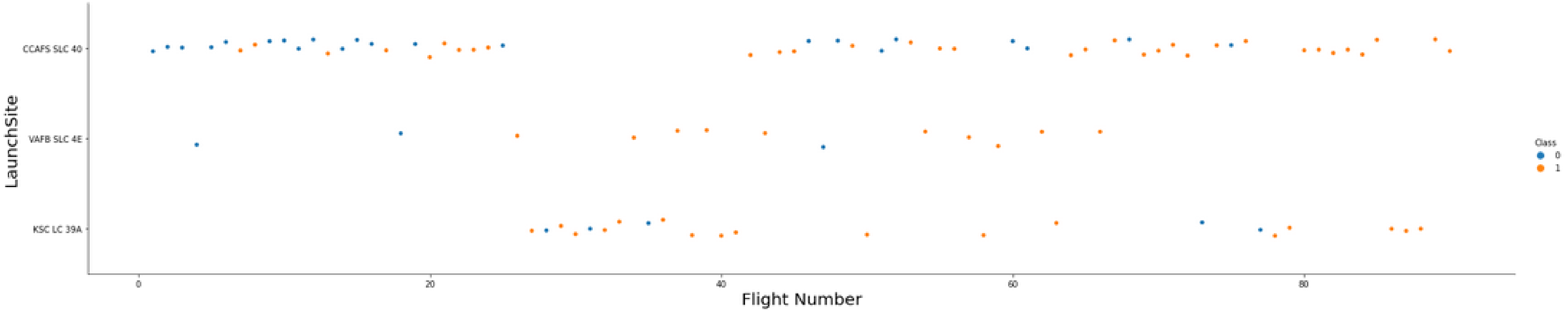


Completed the required EDA with visualization results slides (6 pts)

EDA with visualization results

2. Most of the unsuccessful landings were earlier flight numbers, which were mostly launched from CCAFS SLC 40:

```
# 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("LaunchSite",fontsize=20)
plt.show()
```



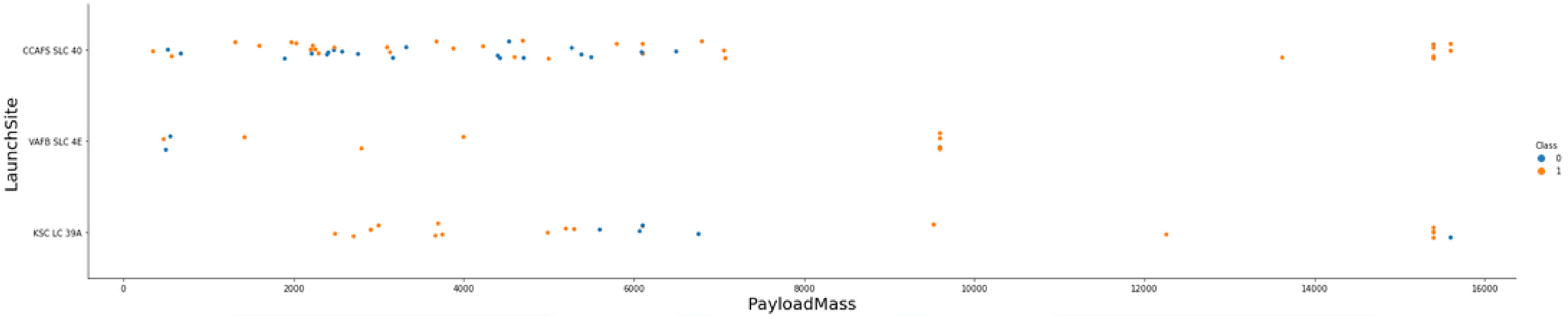
IBM Developer SKILLS NETWORK

Completed the required EDA with visualization results slides (6 pts)

EDA with visualization results

3. Most of the unsuccessful landings were earlier smaller pay load mass launched from CCAFS SLC 40;
4. From VAFB SLC 4E never launched more than 10,000 pay load mass:

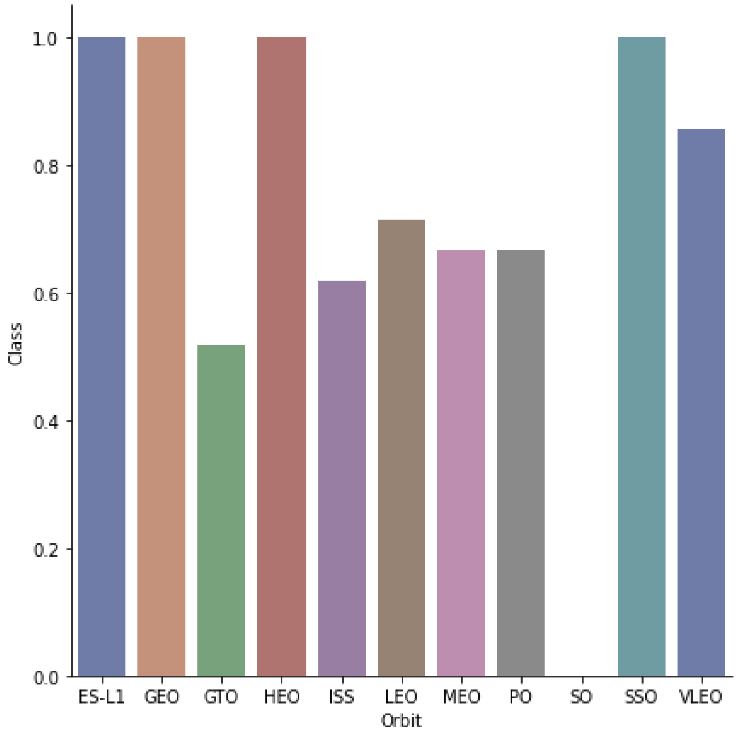
```
# Plot a scatter point chart with x axis to be Pay Load Mass (kg) and y axis to be the launch site, and hue to be the class value

sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("PayloadMass",fontsize=20)
plt.ylabel("LaunchSite",fontsize=20)
plt.show()
```



EDA with visualization results

```
sns.catplot(
    data=sucrate_by_orbit, kind="bar",
    x="Orbit", y="Class",
    ci="sd", palette="dark", alpha=.6, height=6
)
```

<seaborn.axisgrid.FacetGrid at 0x7fc3f83ea9d0>



Analyze the ploted bar chart try to find which orbits have high sucess rate.

Counts of launches for each orbit:

	Orbit	Class
0	ES-L1	1
1	GEO	1
2	GTO	27
3	HEO	1
4	ISS	21
5	LEO	7
6	MEO	3
7	PO	9
8	SO	1
9	SSO	5
10	VLEO	14

1) ES-L1 / GEO / HEO / SO each have only 1 sample point, so the success rate of them is statistical insignificant;

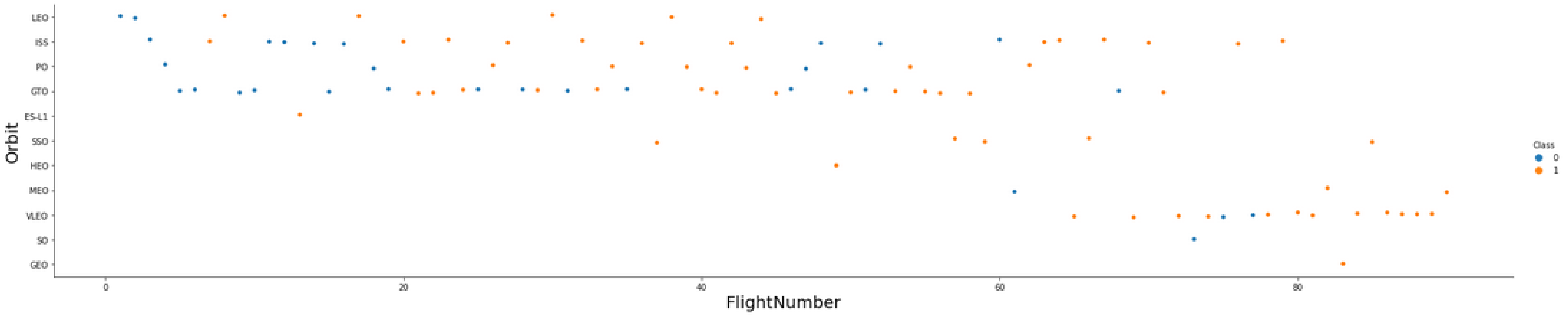
2) Among GTO / ISS / LEO / MEO / PO / SSO / VLEO:

- SSO has highest success landing rate (100%)
- VLEO also has high success rate (85.7%)
- The rest have similar success rate (50.0% ~ 70.0%)

EDA with visualization results

6. We can see that latest (more recent) flight numbers have been trying different orbits:

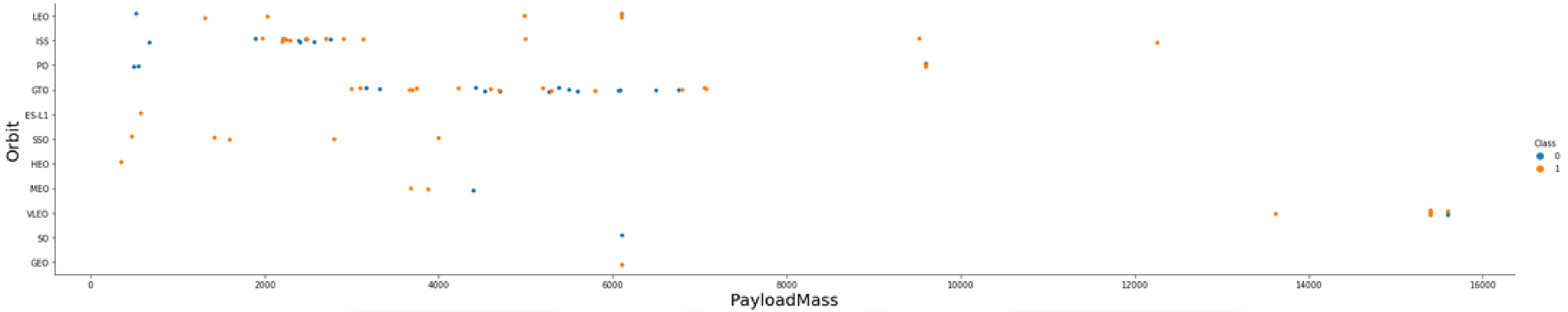
```
# Plot a scatter point chart with x axis to be FlightNumber and y axis to be the Orbit, and hue to be the class value

sns.catplot(y="Orbit", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("FlightNumber",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```



EDA with visualization results

7. Some orbits (like LEO / GTO / SSO) seem to have relatively consistent pay load masses:

```
# 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(y="Orbit", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("PayloadMass",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```



Completed the required EDA with visualization results slides (6 pts)

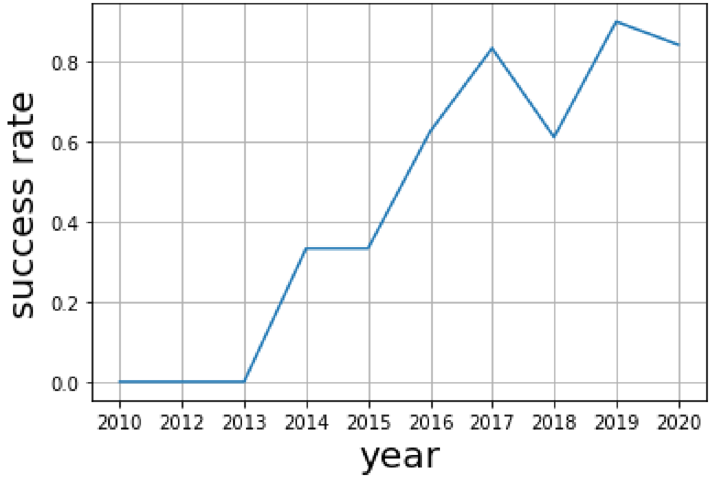
EDA with visualization results

8. Success rate has been improved a lot in recent years.

```
sucrate_by_year
```

	year	Class
0	2010	0.000000
1	2012	0.000000
2	2013	0.000000
3	2014	0.333333
4	2015	0.333333
5	2016	0.625000
6	2017	0.833333
7	2018	0.611111
8	2019	0.900000
9	2020	0.842105

```
# Plot a line chart with x axis to be the extracted year

sns.lineplot(data=sucrate_by_year, x="year", y="Class")
plt.xlabel("year",fontsize=20)
plt.ylabel("success rate",fontsize=20)
plt.grid()
plt.show()
```



Completed the required EDA with SQL results slides (10 pts)

EDA with SQL results

Task 1

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

```
%%sql

SELECT DISTINCT launch_site FROM SPACEXTBL
```

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

Task 2

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

```
%%sql

SELECT * FROM SPACEXTBL
WHERE launch_site LIKE 'CCA%'
LIMIT 5
```

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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)
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-12	22:41:00	F9 v1.1	CCAFS LC-40	SES-8	3170	GTO	SES	Success	No attempt

Completed the required EDA with SQL results slides (10 pts)

EDA with SQL results

Task 3

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

```
%%sql

SELECT SUM(payload_mass__kg_) as total_payload_mass_NASA
FROM SPACEXTBL
WHERE customer LIKE 'NASA%'
```

total_payload_mass_nasa
36679

Task 4

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

```
%%sql

SELECT AVG(payload_mass__kg_) as avg_payload_mass_F9_v1_1
FROM SPACEXTBL
WHERE booster_version LIKE 'F9 v1.1%'
```

avg_payload_mass_f9_v1_1
3226

Task 5

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

Hint:Use min function

```
%%sql

SELECT DISTINCT landing__outcome FROM SPACEXTBL
```

landing__outcome
Controlled (ocean)
Failure
Failure (drone ship)
Failure (parachute)
No attempt
Success
Success (drone ship)
Success (ground pad)

Completed the required EDA with SQL results slides (10 pts)

EDA with SQL results

Task 5

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

```
%%sql

SELECT MIN(DATE) FROM SPACEXTBL GROUP WHERE landing__outcome LIKE 'Success%'
```

1
2016-06-05

```
%%sql

SELECT * FROM SPACEXTBL
WHERE DATE IN (SELECT MIN(DATE) FROM SPACEXTBL WHERE landing__outcome LIKE 'Success%' )
AND landing__outcome LIKE 'Success%'
```

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2016-06-05	05:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)

Completed the required EDA with SQL results slides (10 pts)

EDA with SQL results

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 DISTINCT booster_version FROM SPACEXTBL
WHERE landing__outcome LIKE 'Success (drone ship)%'
AND payload_mass__kg_ BETWEEN 4000 AND 6000
```

booster_version
F9 FT B1031.2
F9 FT B1022

Task 7

List the total number of successful and failure mission outcomes

```
%%sql

SELECT landing__outcome, COUNT(*) AS count
FROM SPACEXTBL
GROUP BY landing__outcome
```

landing__outcome	COUNT
Controlled (ocean)	1
Failure	1
Failure (drone ship)	2
Failure (parachute)	2
No attempt	12
Success	18
Success (drone ship)	5
Success (ground pad)	4

Completed the required EDA with SQL results slides (10 pts)

EDA with SQL results

Task 8

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

```
%%sql

SELECT * FROM SPACEXTBL
WHERE payload_mass__kg_ IN (SELECT MAX(payload_mass__kg_) FROM SPACEXTBL)
```

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2019-11-11	14:56:00	F9 B5 B1048.4	CCAFS SLC-40	Starlink 1 v1.0, SpaceX CRS-19	15600	LEO	SpaceX	Success	Success
2020-07-01	02:33:00	F9 B5 B1049.4	CCAFS SLC-40	Starlink 2 v1.0, Crew Dragon in-flight abort test	15600	LEO	SpaceX	Success	Success
2020-04-06	01:25:00	F9 B5 B1049.5	CCAFS SLC-40	Starlink 7 v1.0, Starlink 8 v1.0	15600	LEO	SpaceX, Planet Labs	Success	Success
2020-03-09	12:46:14	F9 B5 B1060.2	KSC LC-39A	Starlink 11 v1.0, Starlink 12 v1.0	15600	LEO	SpaceX	Success	Success
2020-06-10	11:29:34	F9 B5 B1058.3	KSC LC-39A	Starlink 12 v1.0, Starlink 13 v1.0	15600	LEO	SpaceX	Success	Success

```
%%sql

SELECT DISTINCT booster_version FROM SPACEXTBL
WHERE payload_mass__kg_ IN (SELECT MAX(payload_mass__kg_) FROM SPACEXTBL)
```

booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1058.3
F9 B5 B1060.2

Completed the required EDA with SQL results slides (10 pts)

EDA with SQL results

Task 9

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

```
%%sql

SELECT landing__outcome, booster_version, launch_site FROM SPACEXTBL
WHERE landing__outcome LIKE 'Failure (drone ship)%'
AND DATE LIKE '2015%'
```

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40

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

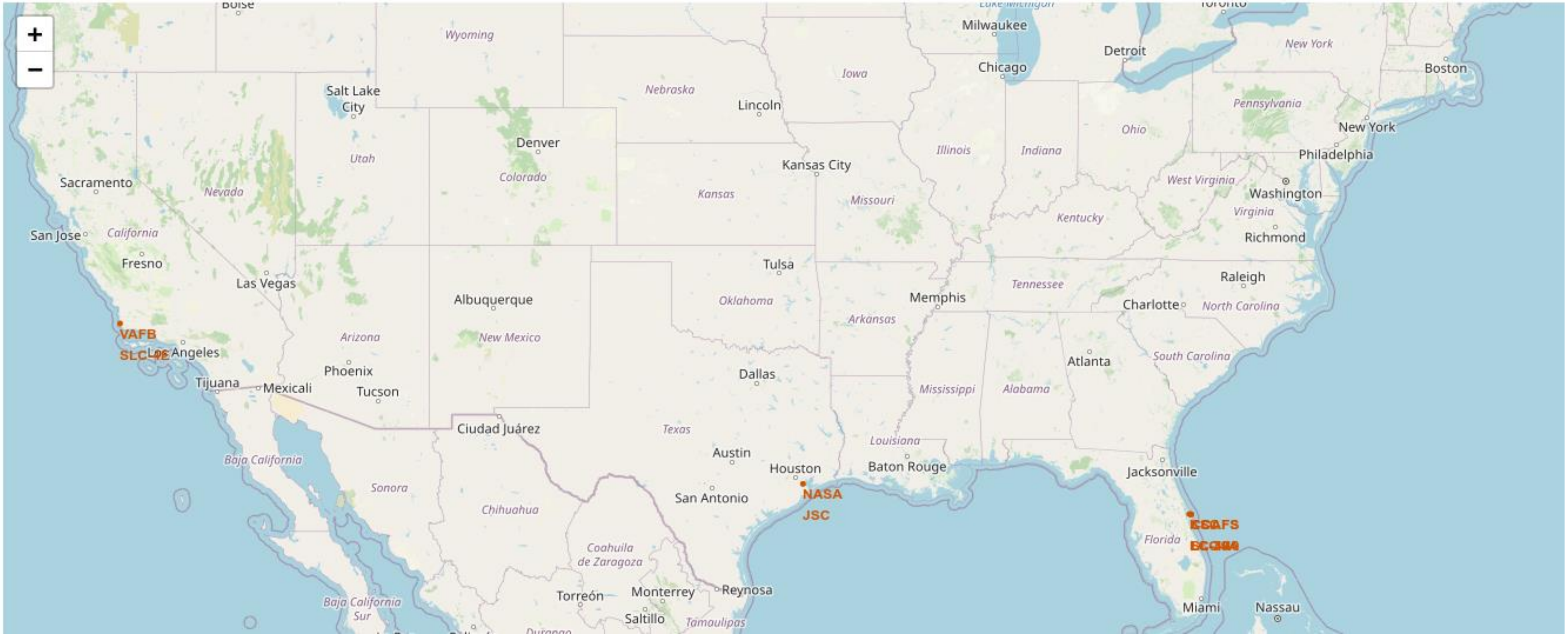
```
%%sql

SELECT landing__outcome, COUNT(*) AS count
FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY landing__outcome
ORDER BY count DESC
```

landing__outcome	COUNT
No attempt	7
Failure (drone ship)	2
Success (drone ship)	2
Success (ground pad)	2
Controlled (ocean)	1
Failure (parachute)	1

Completed the required interactive map with Folium results slides (3 pts)

Interactive Map With Folium results

Task 1: Mark all launch sites on a map

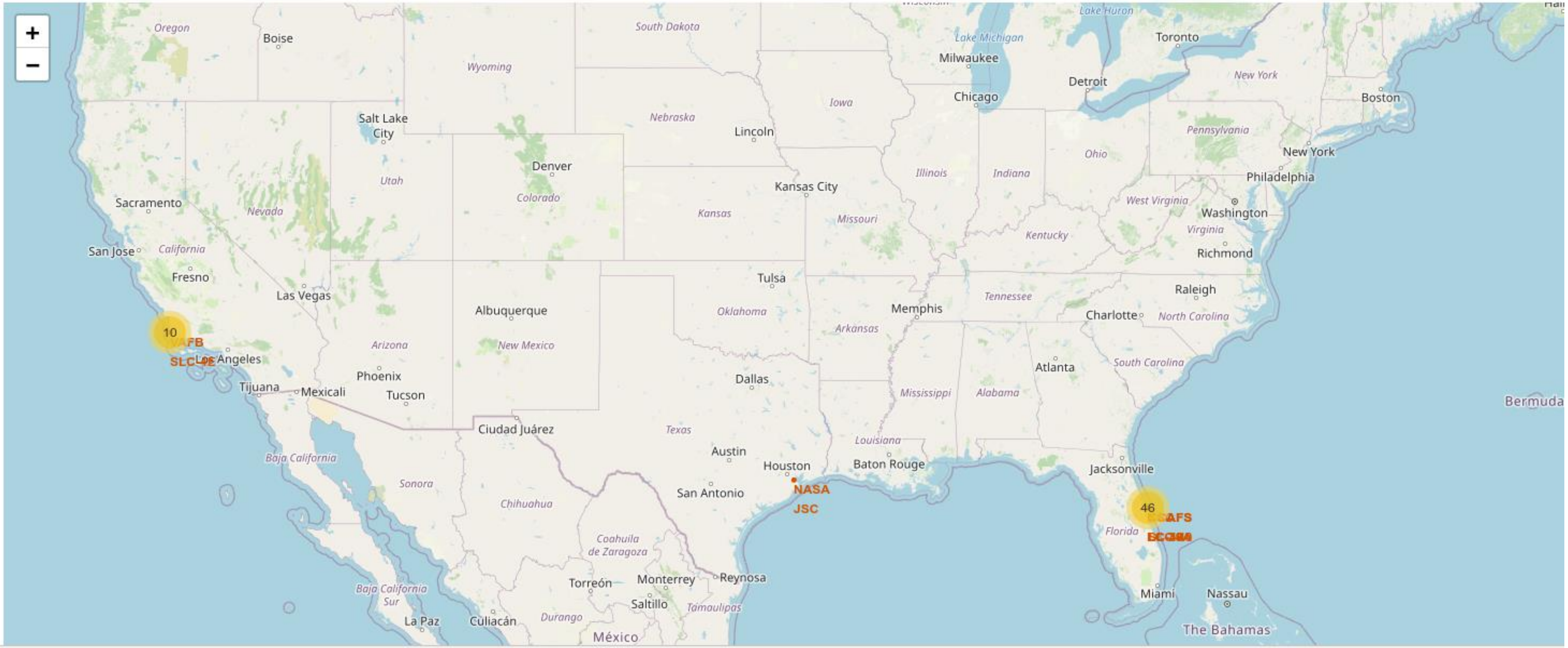


Completed the required interactive map with Folium results slides (3 pts)

Interactive Map With Folium results

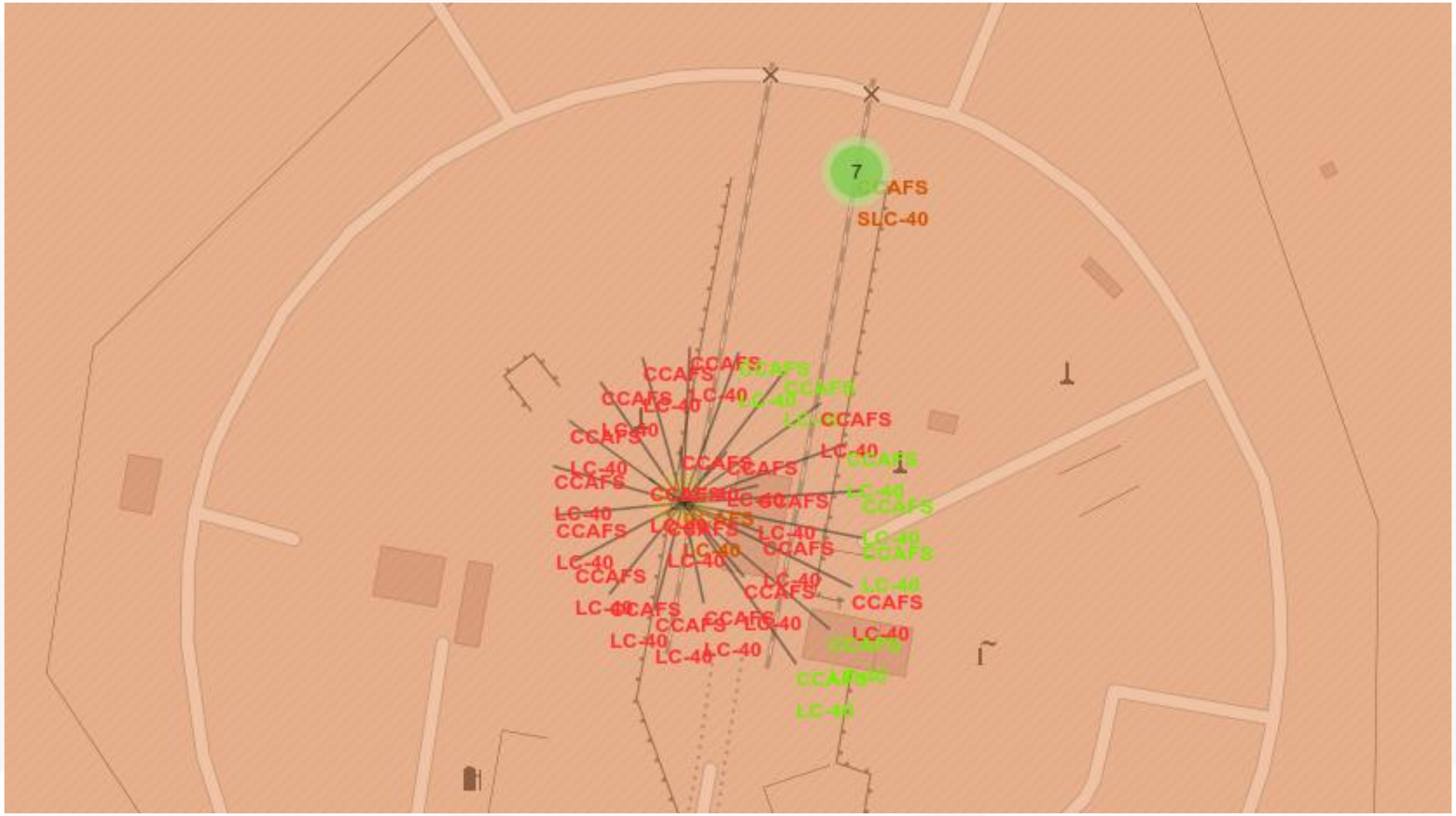
Task 2: Mark the success/failed launches for each site on the map

Out[18]:



Completed the required interactive map with Folium results slides (3 pts)

Interactive Map With Folium results



Completed the required Conclusion slide (1 pts)

CONCLUSION

- Overall historical data is not enough to have too many statistically significant conclusions.

- We have a good view of historical trend SpaceX launches and landings:

  - Over the time success rate of landing got much improved;
  - Over the time SpaceX has been exploring more new orbits;
  - Over the time SpaceX is capable for more payload mass.

- We have various nice visualizations of historical data:

  - Various static plots and tables using plotly or seaborn;
  - Interactive statistical plots using dash and geographical plots using folium.

- Predictions of landing outcome has accuracy of 85%.