Data Science Capstone Final Presentation

Michael Eisenberg

July 28, 2022

IBM Developer

SKILLS NETWORK

EXECUTIVE SUMMARY

- Point1
- Point2
  - Sub Point 1
  - Sub Point 2
  - Sub Point 3
- Point3
- Point4
- Point5

INTRODUCTION

- Rocket launches are expensive
  - Typically costs more the $165 million per launch
  - SpaceX launch typically costs around $62 million
- How can we reduce costs of rocket launches?
  - SpaceX claims it can launch Falcon 9 rockets and reuse the first stage
  - SpaceX claims it can land the first stage to maintain reusability
- Does SpaceX have the formula to landing the first stage?
  - What datapoints are important in a successful landing?

METHODOLOGY – Data Collection

- Using the Python requests API, collected data from the Falcon 9 Launches Wikipedia page and the SpaceX API
- Used BeatifulSoup to scrape the data from the Wikipedia page
- Converted the data obtain from both sources into Pandas Dataframes
  - Performed necessary cleaning of the data to transform into a state that is ready for exploratory analysis and building machine learning models
  - Information obtained from both sources had data regarding launch sites, payload, orbit type, and more

Methodology – Data Collection Web Scraping

- Web scraped the Falcon 9 Wikipedia page using the requests and BeautifulSoup Libraries and converted in a Pandas dataframe
- Notebook Link: Web Scraping Notebook

```
# use requests.get() method with the provided static_url
# assign the response to a object
response = requests.get(static_url)
```

Create a `BeautifulSoup` object from the HTML `response`

```
# Use BeautifulSoup() to create a BeautifulSoup object from a response text content
soup = BeautifulSoup(response.text)
```

Methodology – Data Collection SpaceX API

- Obtained information on SpaceX rocket launches using the requests library on the SpaceX API
- Converted the scraped data into a Pandas dataframe using pd.json_normalize and performed necessary cleanup
- Notebook Link: SpaceX API Notebook

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

```
response = requests.get(spacex_url)
```

```
# Lets take a subset of our dataframe keeping only the features we want and the flight number, and date_utc.
data = data[['rocket', 'payloads', 'launchpad', 'cores', 'flight_number', 'date_utc']]

# We will remove rows with multiple cores because those are falcon rockets with 2 extra rocket boosters and rows that have multiple payloads in a sing
data = data[data['cores'].map(len)==1]
data = data[data['payloads'].map(len)==1]

# Since payloads and cores are lists of size 1 we will also extract the single value in the list and replace the feature.
data['cores'] = data['cores'].map(lambda x : x[0])
data['payloads'] = data['payloads'].map(lambda x : x[0])

# We also want to convert the date_utc to a datetime datatype and then extracting the date leaving the time
data['date'] = pd.to_datetime(data['date_utc']).dt.date

# Using the date we will restrict the dates of the launches
data = data[data['date'] <= datetime.date(2020, 11, 13)]
```

Methodology – Data Wrangling

- Performed exploratory data analysis to discover patterns as well as best approaches for further analysis
- Used a Pandas Dataframe and relevant methods
  - pd.value_counts() to determine how often certain categories such as orbit type and landing outcome occurred in the data
  - Applied mean method to a column to determine the overall landing success rate
- Notebook: Data Wrangling Notebook

```
# landing_outcomes = values on Outcome column
landing_outcomes = df['Outcome'].value_counts()
landing_outcomes
```

```
True ASDS      41
None None      19
True RTLS      14
False ASDS      6
True Ocean      5
False Ocean     2
None ASDS       2
False RTLS      1
Name: Outcome, dtype: int64
```

Methodology – EDA with Data Visualization

- Created scatter charts to visualize the relationship between the following variables and their impact on a successful launch
  - Launch site vs. Flight Number
  - Launch site vs. Payload
  - Orbit type vs. Flight Number
  - Orbit type vs. Payload
  - Each scatter chart displayed if each point represented a successful or unsuccessful launch
- Created a bar chart to furhter visualize the impact of orbit type on launch success
- Created a line chart to evaluate how launch success rate has varied over the years
- Notebook Link: EDA with Data Visualization Notebook

```
# 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("Pay load Mass (kg)",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```

Methodology – EDA with SQL

- Loaded data from csv file into IBM DB2 for use in SQL magic queries from jupyter notebook
- Queries performed provided insight into content of data and relationship between attributes
  - Display the names of the launch sites in the space mission
  - Display 5 records where launch sites begin with 'CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in ground pad was achieved
  - List the names of the boosters which have success in drone ship launches and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failed missions
  - List the names of the booster versions which have carried the maximum payload mass
  - List the failed landings in drone ship, their booster versions, and launch site names in 2015
  - Rank the count of landing outcomes between 2010-06-04 and 2017-03-20
- Notebook Link: EDA with SQL Notebook

Methodology – EDA with Interactive Folium Map

- Used the Python Folium library, which is designed for creating interactive maps
- Loaded SpaceX launch data into a Pandas Dataframe to be used in creating the maps
- Maps created
  - Maps where launch sites are marked
  - Maps marking (and differentiating) successful and failed launches
  - Maps marking distance between launch sites and important locations
- Link to Notebook: Interactive Folium Map Notebook

```
for index, record in spacex_df.iterrows():
    # TODO: Create and add a Marker cluster to the site map
    marker = folium.Marker(
        [record['Lat'], record['Long']],
        # Create an icon as a text label
        icon=folium.Icon(color='white', icon_color=record['marker_color'])
    )
    marker_cluster.add_child(marker)

site_map
```

Methodology – EDA with Interactive Dashboard

- Used the Python Dash library to create the dashboard
- Used plotly express to create the charts
- Dashboard contained two charts
  - Pie chart representing the total success launches by all sites or the percent of successful launches in a given site
  - Scatter chart representing the correleation between payload and launch success
  - Charts relied on callback functions and user input for the interaction and changes within the charts
- Link to Python file for the Dash App: Dash App Python File

```
@app.callback(
    Output(component_id="success-pie-chart", component_property="figure"),
    Input(component_id="site-dropdown", component_property="value"),
)
def get_pie_chart(entered_site):
```

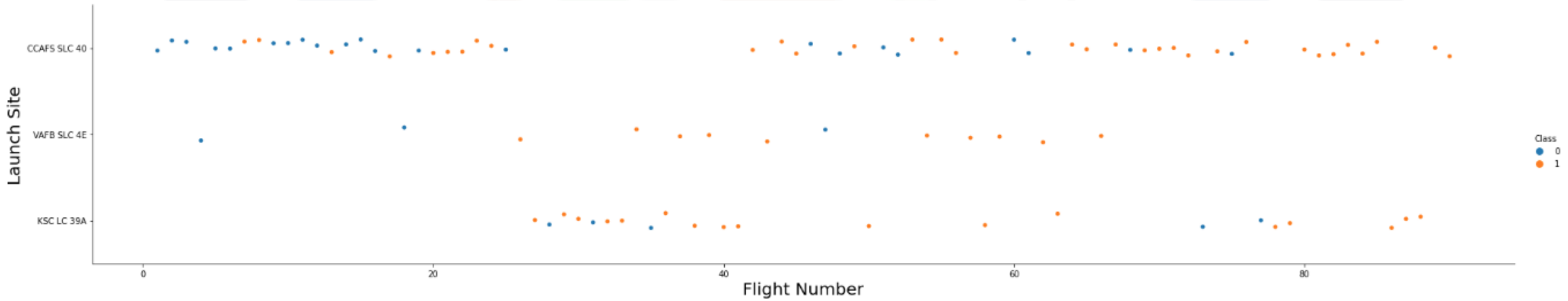
Methodology – Predictive Analysis

- Loaded into Pandas Dataframe for used in classification algorithms
- Performed K-Nearest Neighbors, Logistic Regression, Support Vector Machine, and Decision Tree analysis on the SpaceX data
- Performed necessary standardization of data prior to splitting into training and test sets
- Used GridSearch to determine best hyperparameters for each model type
- Evaluated each of the best models and selected the one with the highest accuracy for use in making predictions
- Notebook Link: Predictive Analysis Notebook

EDA With Data Visualization Results

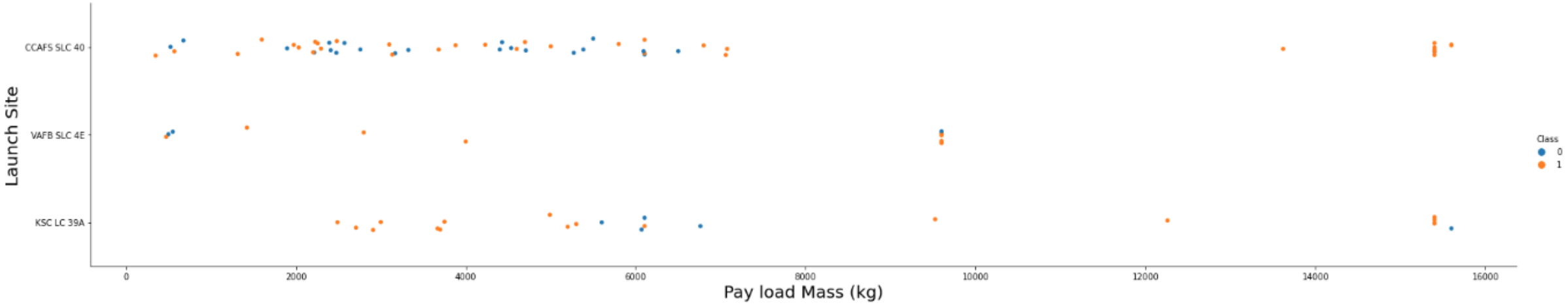
Flight Number vs. Launch Site

- Flights from site KSC LC 39A were primarily successful
- Earlier flights from CCAFS SLC 40 and VAFB SLC 4E were more likely to fail while larger flight numbers meant the launch was most likely to succeed
- Flights from KSC LC 39A had failures on both the lower and upper end of flight numbers



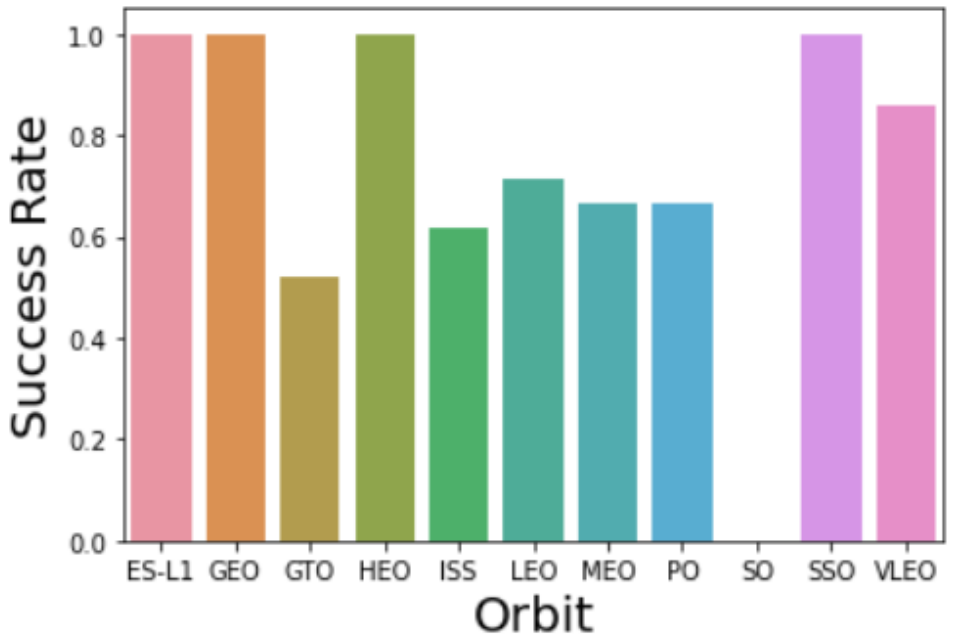
Launch Site vs. Payload

- Flights with a larger payload mass had a higher probability of success
- Flights from KSC LC 39A with a payload less than 6000 were most likely to be successful
- Flights from VAFB SKC 4E did not have a significantly higher probability of success or failure regardless of payload mass
- Flights from CCAFS SLC 40 had approximately the same probability of success and failure for flights with low payloads



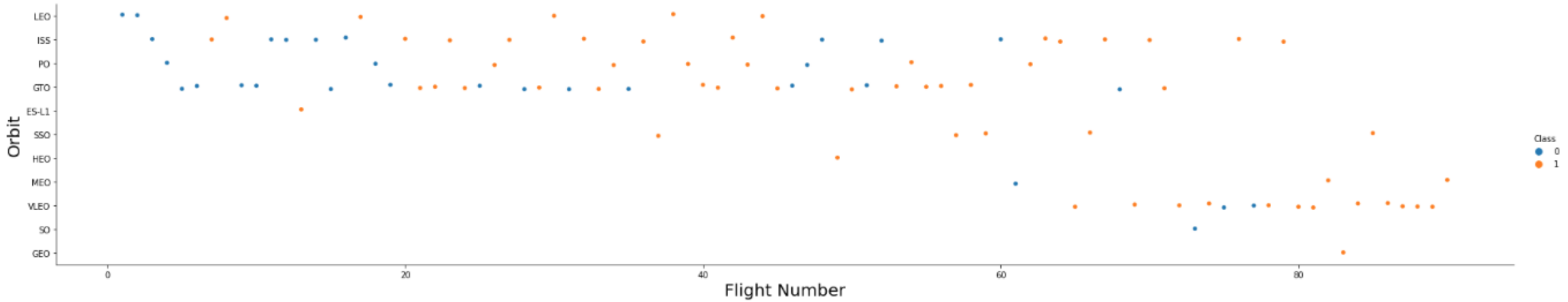
Success Rate vs. Orbit Type

- Flights sent to orbits ES-L1, GEO, HEO, and SSO had a 100 percent success rate
- Flights into VLEO orbit had a high success rate of over 80 percent
- Flights sent to ISS, LEO, MEO, and PO orbits had lower success rates, with at least a 25% probability of failure
- Flights into GTO or SO orbit had the lowest probability of success, with a success rate of only approximately 50 percent for GTO orbit and zero percent for SO orbit



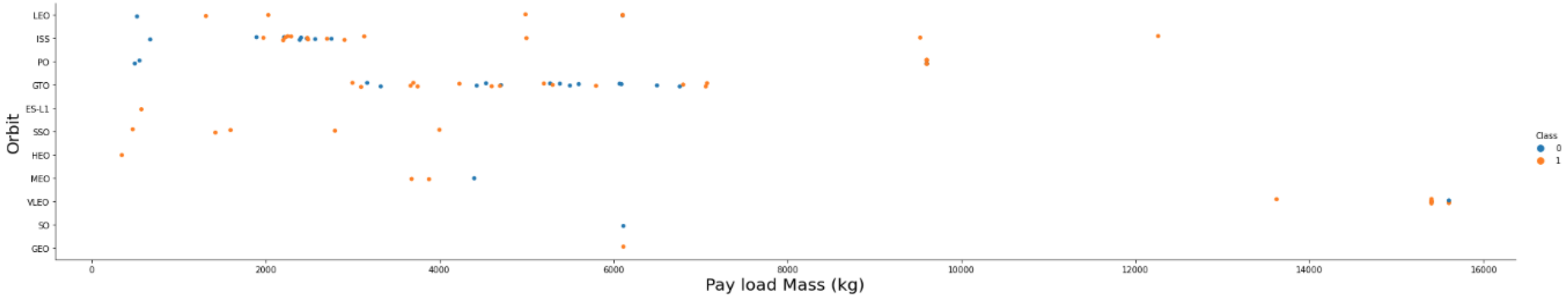
Orbit Type vs. Flight Number

- ISS, and LEO flights were more likely to fail at lower flight numbers
- GTO and PO flights does not have a relationship with flight number with respect to success
- SSO and VLEO orbits were highly successful



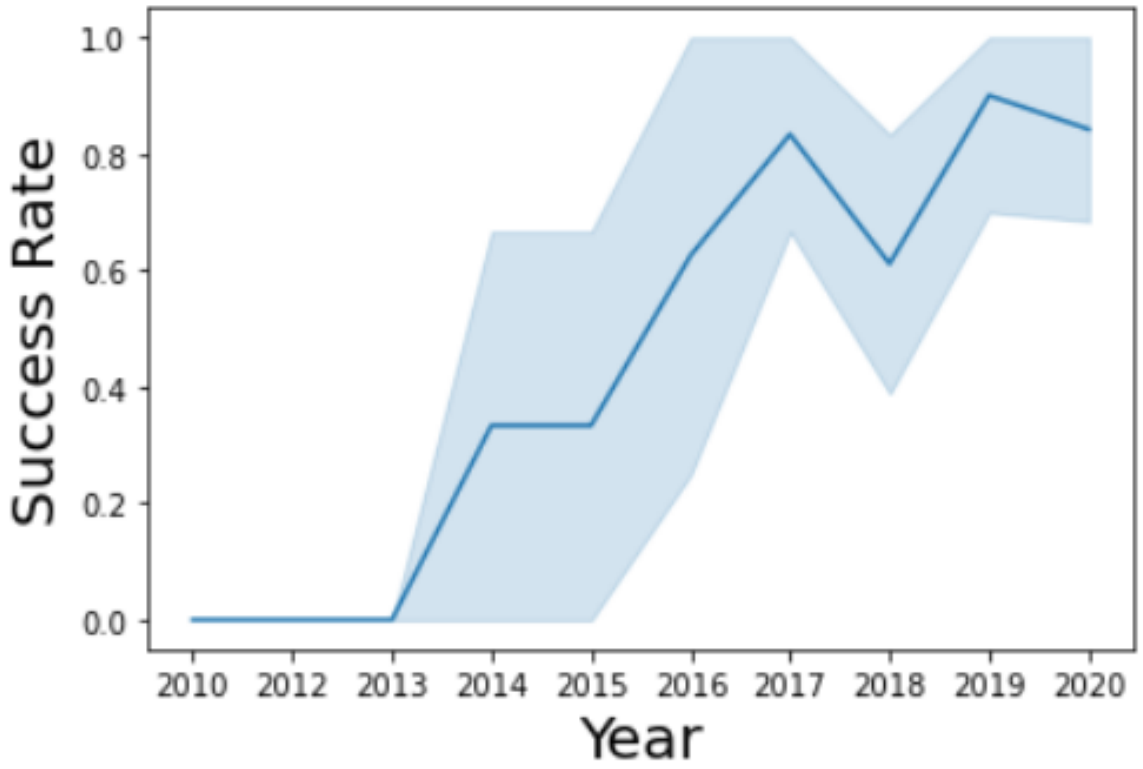
Orbit Type vs. Payload

- GTO flights does not have a relationship with Payload mass with respect to succesful flights
- ISS, PO, and LEO flights were more successful as payload mass increased
- VLEO flights were all at high payloads, with the only failure coming at approximately the highest payload (where there is also a success) for a VLEO flight



Success Rate vs. Year

- Flight success rate has increased over time, with the only significant drop being from years 2017 to 2018



EDA With SQL Results

Names of the Unique Launch Sites

- Obtaining the names of all launch sites used
- There were only four launch sites used in the SpaceX flights

```
%sql select distinct(LAUNCH_SITE) from SPACEXTBL
```

```
 * ibm_db_sa://zfh74831:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/BLUDB
Done.
```

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

5 records where launch sites begin with 'CCA'

- Obtaining 5 records where the Launch Site name begins with 'CCA'

```
%sql select * from SPACEXTBL where LAUNCH_SITE LIKE 'CCA%' LIMIT 5
```

```
 * ibm_db_sa://zfh74831:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/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

Total Payload Mass of NASA (CRS) boosters

- Obtaining the sum of the payloads across all flights from the NASA CRS launch site, which is 45596 kilograms

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

```
 * ibm_db_sa://zfh74831:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/BLUDB
Done.
```

1
45596

Average Payload Mass of F9 v1.1 boosters

- Obtaining the average payload mass of F9 v1.1 boosters, which is 2928 kilograms

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

```
 * ibm_db_sa://zfh74831:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/BLUDB
Done.
```

1
2928

First Successful Ground Pad Landing

- Obtaining the date of the first successful ground pad landing, which was on December 22, 2015

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

```
 * ibm_db_sa://zfh74831:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/BLUDB
Done.
```

1
2015-12-22

Names of boosters with successful drone ship landings and a payload between 4,000 and 6,000 kilograms

- Obtaining the names of boosters with at least one successful drone ship landing where the payload was between 4,000 and 6,000 kilograms
- There were four boosters with a successful landing of this type

```
%sql select distinct(BOOSTER_VERSION) from SPACEXTBL where LANDING__OUTCOME = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000
```

```
 * ibm_db_sa://zfh74831:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/BLUDB
Done.
```

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

Count of successful and failed missions

- Obtaining the number of successful and failed missions
  - If a landing outcome begins with 'Failure' then the outcome was a failure and a success otherwise

```
%sql select count(LANDING__OUTCOME) as failure from SPACEXTBL where LANDING__OUTCOME LIKE 'Failure%'
```

```
 * ibm_db_sa://zfh74831:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/BLUDB
Done.
```

1]:

failure
10

```
%sql select count(LANDING__OUTCOME) as failure from SPACEXTBL where LANDING__OUTCOME LIKE 'Success%'
```

```
 * ibm_db_sa://zfh74831:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/BLUDB
Done.
```

2]:

failure
61

Boosters carrying the maximum payload

- Obtaining the list of boosters which have carried the maximum payload mass

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

```
 * ibm_db_sa://zfh74831:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/BLUDB
Done.
```

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

Failed Drone Ship Landings in 2015

- Obtaining the failed drone ship landings in 2015, their launch site, and their booster version
- There were only two such landings

```
%sql select BOOSTER_VERSION, LAUNCH_SITE, DATE from SPACEXTBL
    where LANDING__OUTCOME = 'Failure (drone ship)'
    and DATE between '2015-01-01' and '2016-01-01'
```

```
 * ibm_db_sa://zfh74831:***@b70af05b-76e4-4bca-a1f5-23dbb4c6
Done.
```

];

booster_version	launch_site	DATE
F9 v1.1 B1012	CCAFS LC-40	2015-01-10
F9 v1.1 B1015	CCAFS LC-40	2015-04-14

Count landing outcomes between June 4, 2010 and March 20, 2017

- Obtaining the landing outcomes and how often they occurred for all flights between June 4, 2010 and March 20, 2017

```
%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 2 desc
```

```
 * ibm_db_sa://zfh74831:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/BLUDB
Done.
```

7]:

landing__outcome	COUNT
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Interactive Folium Map Results

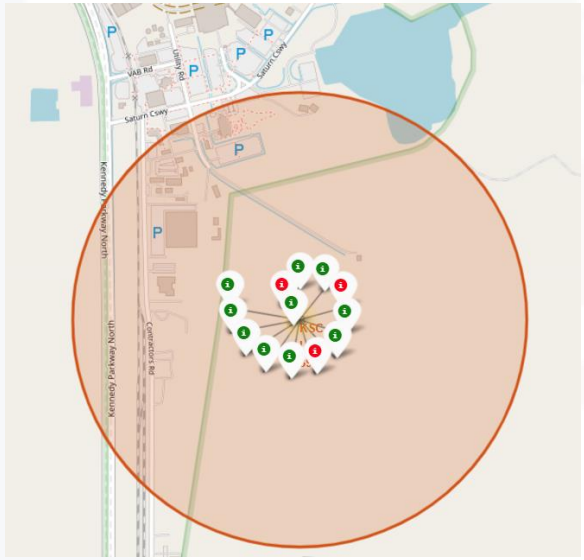
Mark all launch sites on a map

- Displaying the launch sites on a map using markers
- All launch sites are in the United States

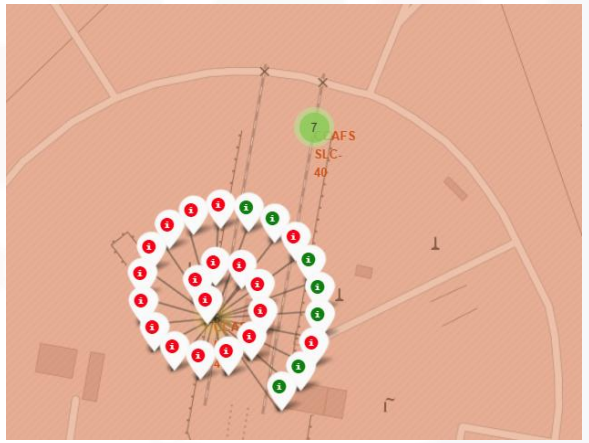
Mark all success/failed launches for each site

- Displaying each launch on their launch site and whether the launch was a success or failure

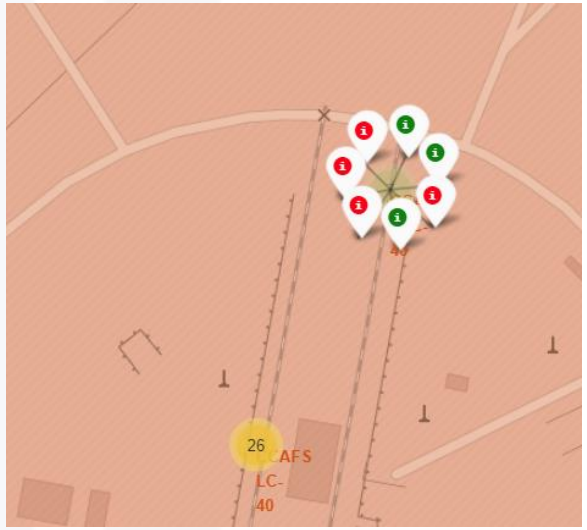
VAFB SLC-4E



KSC LC-39A



CCAFS LC-40

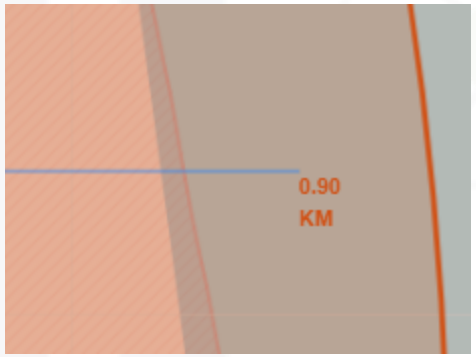


CCAFS SLC-40

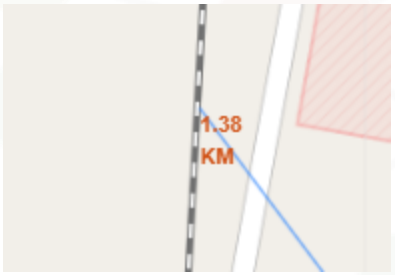
Calculate distance from a launch site to its proximities

- Calculating the distance from CCAFS SLC-40 to its proximities (coastline, highway, railway, and city)

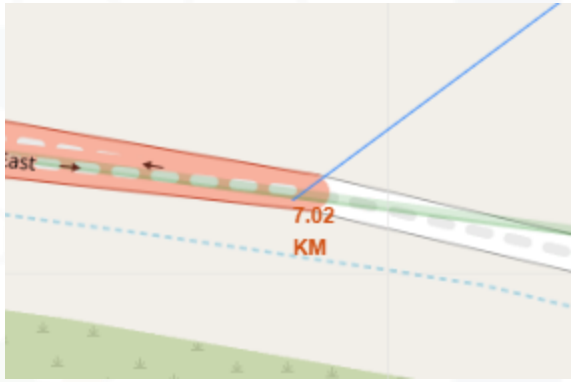
CCAFS SLC-40



Closest Coastline



Closest Railway



Closest Highway

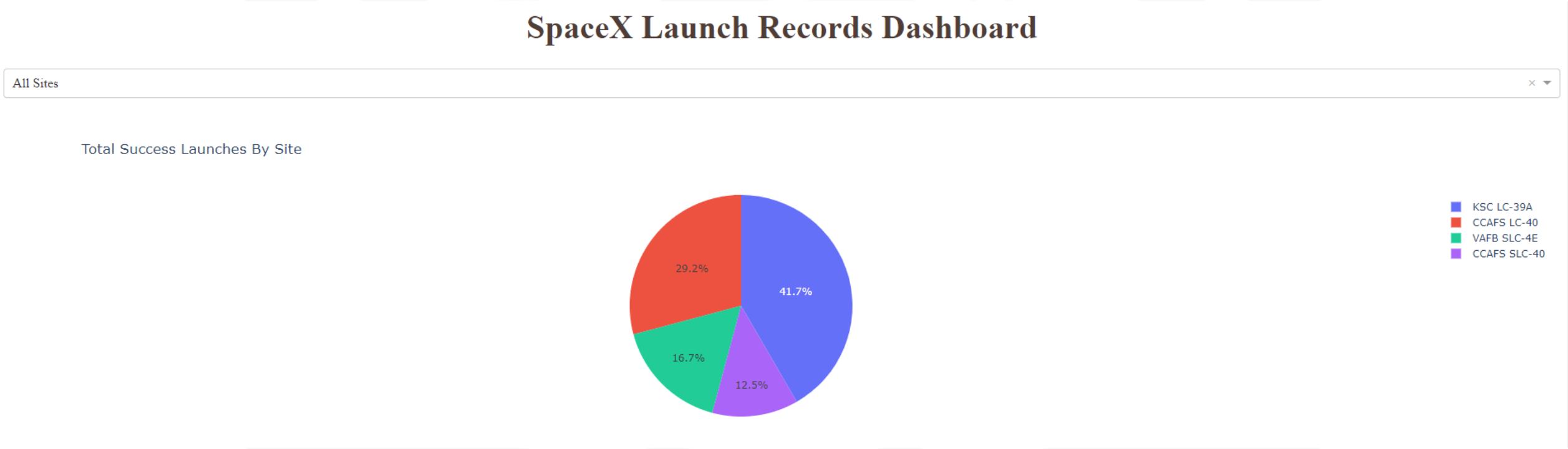
Closest Major City

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

Interactive Dashboard Results

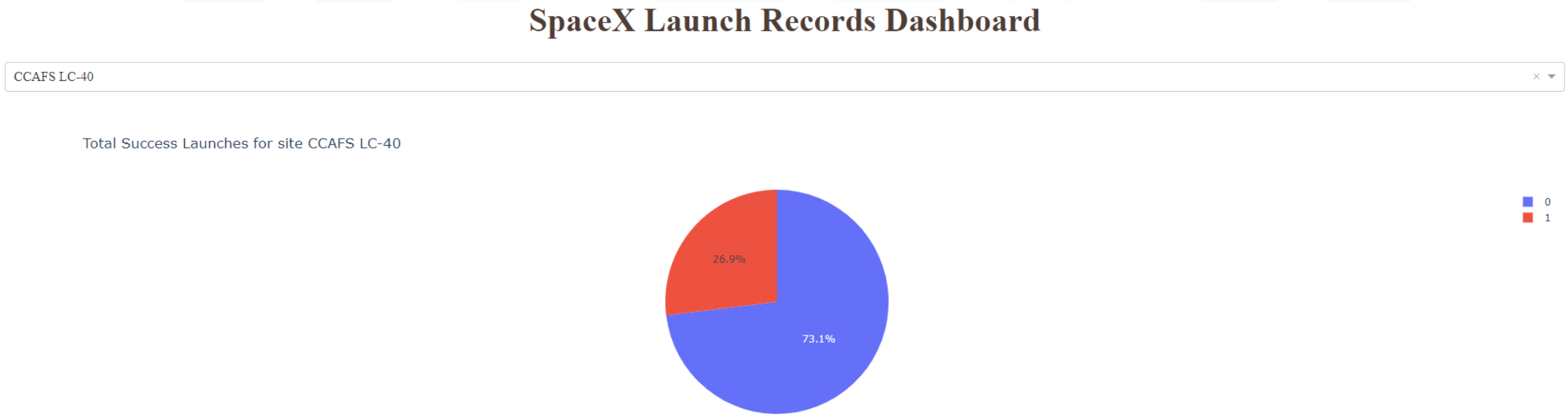
Percentage of successful launches for each launch site

- Displaying each launch site in a pie chart and their percentage of all successful launches



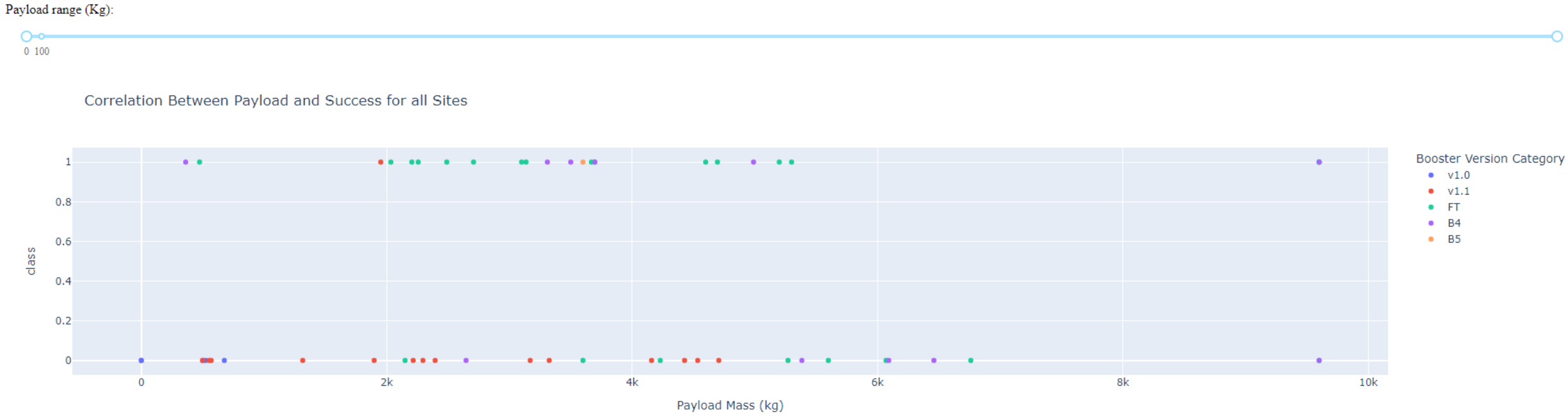
Success vs. Failure percentage for CCAFS LC-40

- Displaying a pie chart showing the percentage of launches that were a success and failure for launch site CCAFS LC-40



Correlation between payload and success for all sites

- Displaying a scatter plot representing the correlation between payload mass and success rate for all launch sites, differentiating between booster versions



Predictive Analysis (Classification) Results

Classification Accuracy

- Accuracy score for each model on the test set
  - KNN: 0.8333333333333334
  - Logistic Regression: 0.8333333333333334
  - SVM: 0.8333333333333334
  - Decision Tree: 0.94444444444444
- Vales for best_score_
  - KNN: 0.8482142857142858
  - Logisitic Regression: 0.8464285714285713
  - SVM: 0.8482142857142856
  - Decision Tree: 0.8892857142857142
- Since the Decision Tree model has the highest test score and best_score_ value, we can say it is the best model for making predictions on the success of a launch

Confusion Matrix

- Displaying the Confusion matrix for the decision tree model
- Precision: 0.9231
- Recall: 1.0

Conclusion

- The decision tree model is the best model to use in predicting whether a launch will be successful
- Launch site KSC LC-39A had the highest amount of successful launches
- Launch sites have easy access to transportation but are far from major cities
- Launches with a higher payload mass or a higher flight number tend to be more successful than those with lower payloads or flight numbers
- Launches into ES-L1, GEO, HEO, SSO, or VLEO orbits were successful while launches into GTO or SO orbits did not find much success
- The percentage of successful launches has generally increased with only one significant drop that occurred from 2017 to 2018