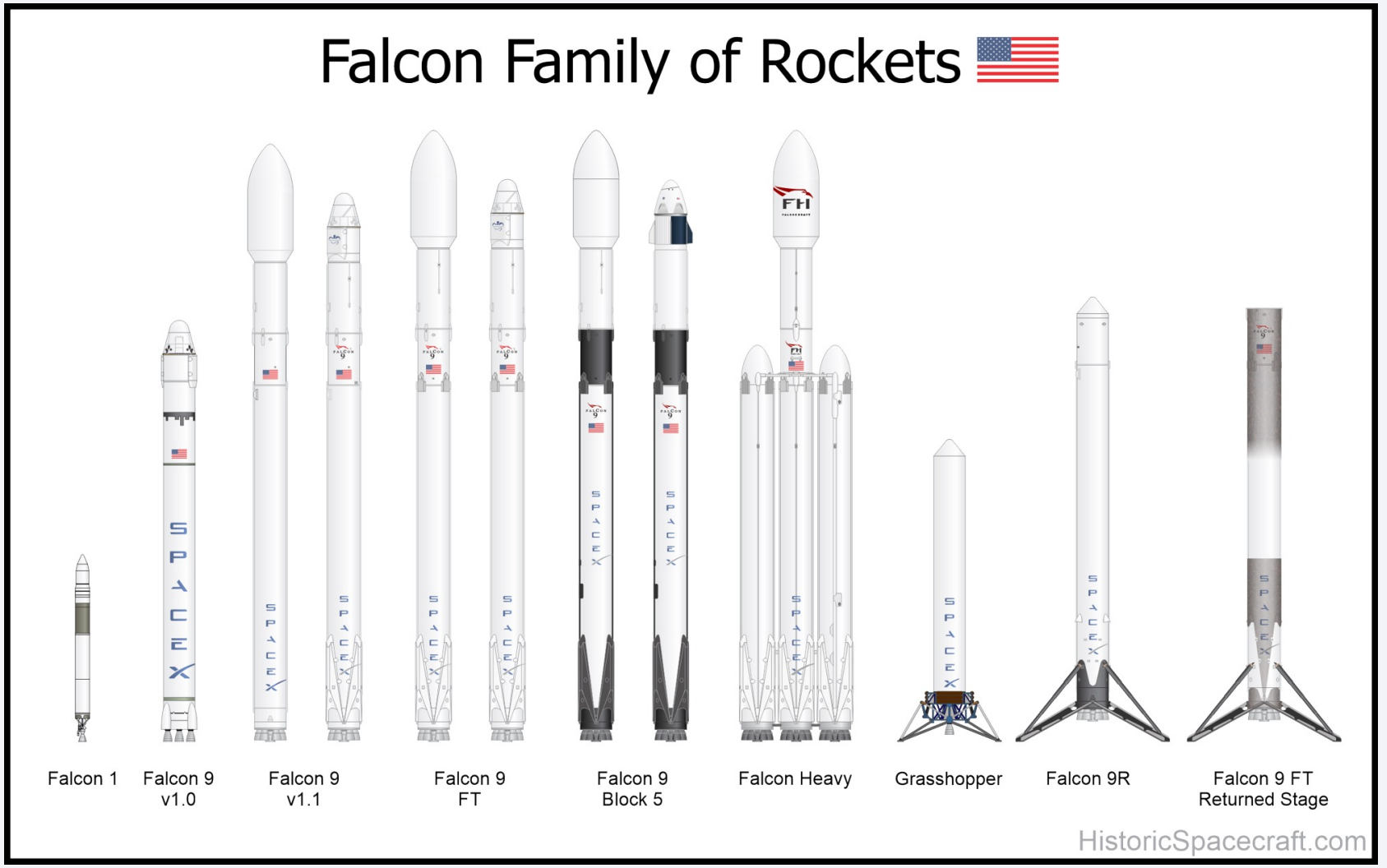
Winning Space Race with Data Science

Paul Hegedus
August 30th, 2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

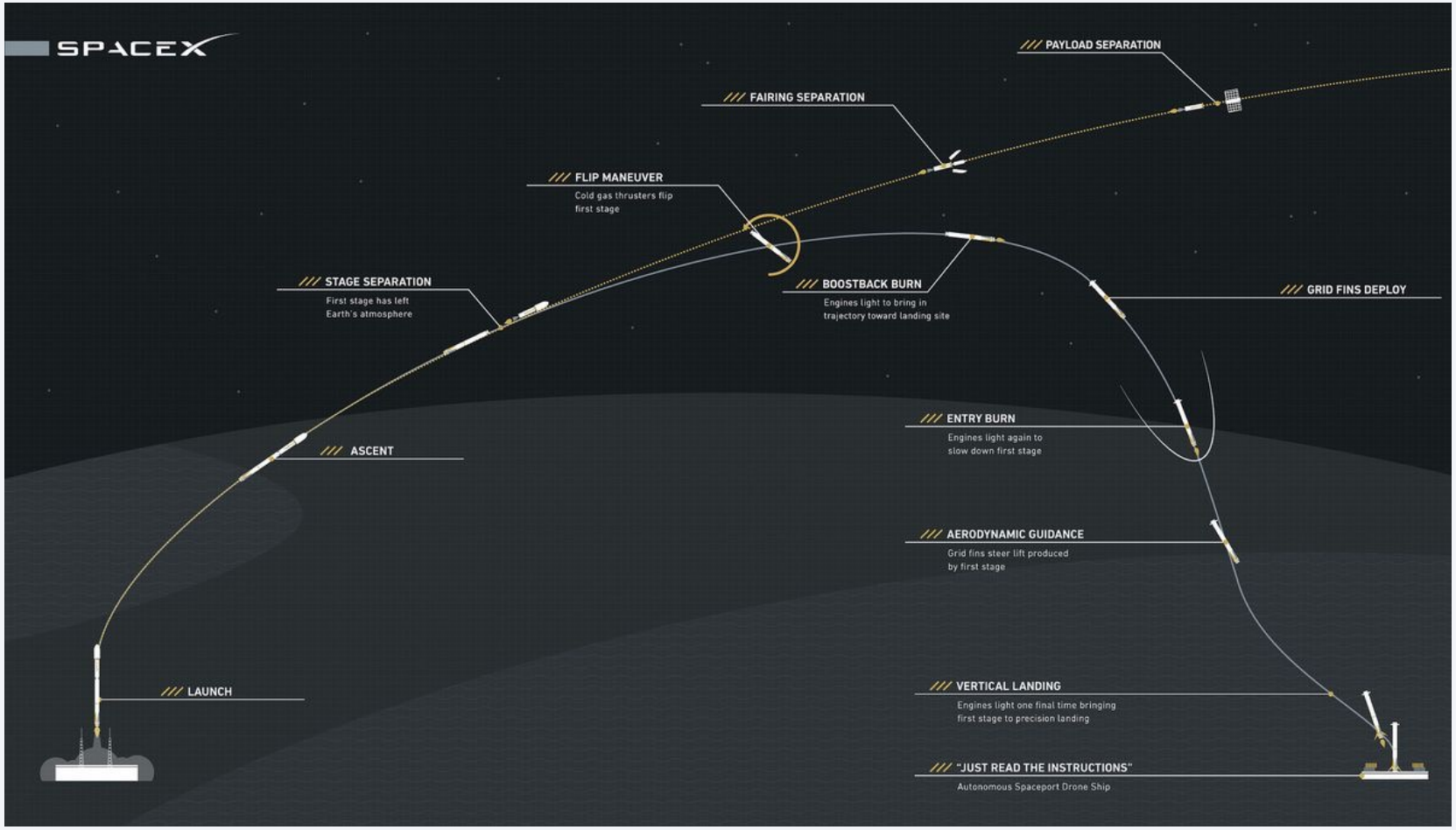


Executive Summary

- Methodology
  - 01 Data Collection API
  - 02 Web Scraping
  - 03 Data Wrangling
  - 04 Exploratory Data Analysis
- Results
  - 05 Data Visualization
  - 06 Launch Site Locations
  - 07 Machine Learning Predictions

Introduction

- Project background and context
  - The commercial space age is here, companies are making space travel affordable for everyone. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards 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.
- Problems you want to find answers
  - What influences a successful rocket landing?
  - What conditions are required for a successful reuse of first stage rocket?



Section 1

Methodology

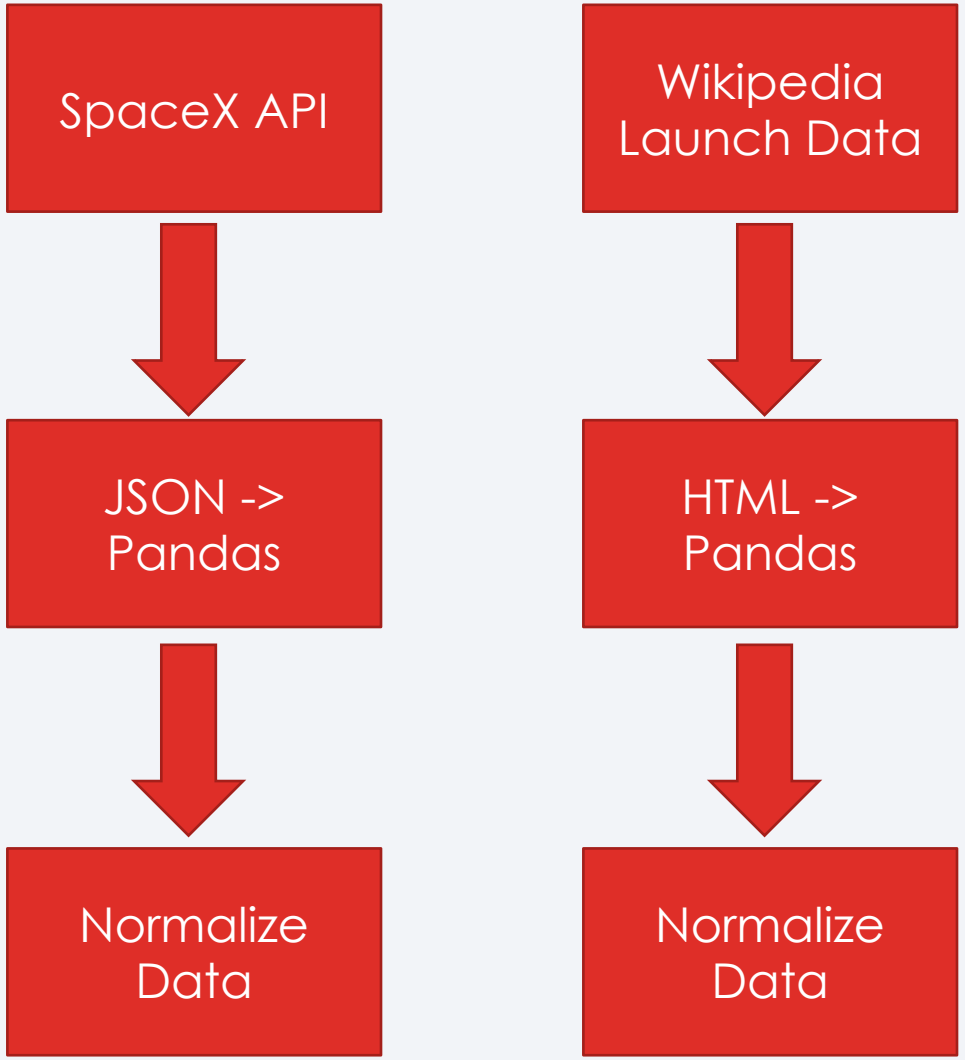
Methodology

Executive Summary

- Data collection methodology:
  - Collect data from SpaceX API and web scraping from Wikipedia
- Perform data wrangling
  - One hot encoding for categorical variables for machine learning models
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Viewed patterns in data with scatter and bar plots
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- SpaceX launch data gathered from SpaceX REST API
- Json files turned into pandas dataframes
- Data cleaned, missing data fixed
- Falcon launch data scraped from Wikipedia
- Falcon 9 data isolated
- Launch data from Wikipedia converted to pandas dataframes



Data Collection – SpaceX API

- SpaceX API data collected, converted to pandas dataframe, normalized, missing values filled in

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

```
# Use json_normalize meethod to convert the json result into a dataframe
response = requests.get(static_json_url).json()
data = pd.json_normalize(response)
```

- Add the GitHub URL of the completed SpaceX API calls notebook
  - https://github.com/paulhegedus/IBM-DataScienceCapstone/blob/c9380f7e05450ee0c9df2e7e411804dccb255911/01_DataCollectionAPI.ipynb

```
# Create a data from launch_dict
launch_df = pd.DataFrame.from_dict([launch_dict])
```

```
# Calculate the mean value of PayloadMass column
PayloadMass = pd.DataFrame(data_falcon9['PayloadMass'].values.tolist()).mean(1)
# Replace the np.nan values with its mean value
rows = data_falcon9['PayloadMass'].values.tolist()[0]

df_rows = pd.DataFrame(rows)
df_rows = df_rows.replace(np.nan, PayloadMass)

data_falcon9['PayloadMass'][0] = df_rows.values
```

Data Collection - Scraping

- Scraped HTML data from Wikipedia, parsed using BeautifulSoup, converted to pandas dataframe

```
static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid
```

```
html_data = requests.get(static_url)
```

```
soup = BeautifulSoup(html_data.text, 'html.parser')
```

```
df=pd.DataFrame(launch_dict)
```

- Add the GitHub URL of the completed web scraping notebook
  - https://github.com/paulhegedus/IBM-DataScienceCapstone/blob/c9380f7e05450ee0c9df2e7e411804dccb255911/02_Webscraping.ipynb

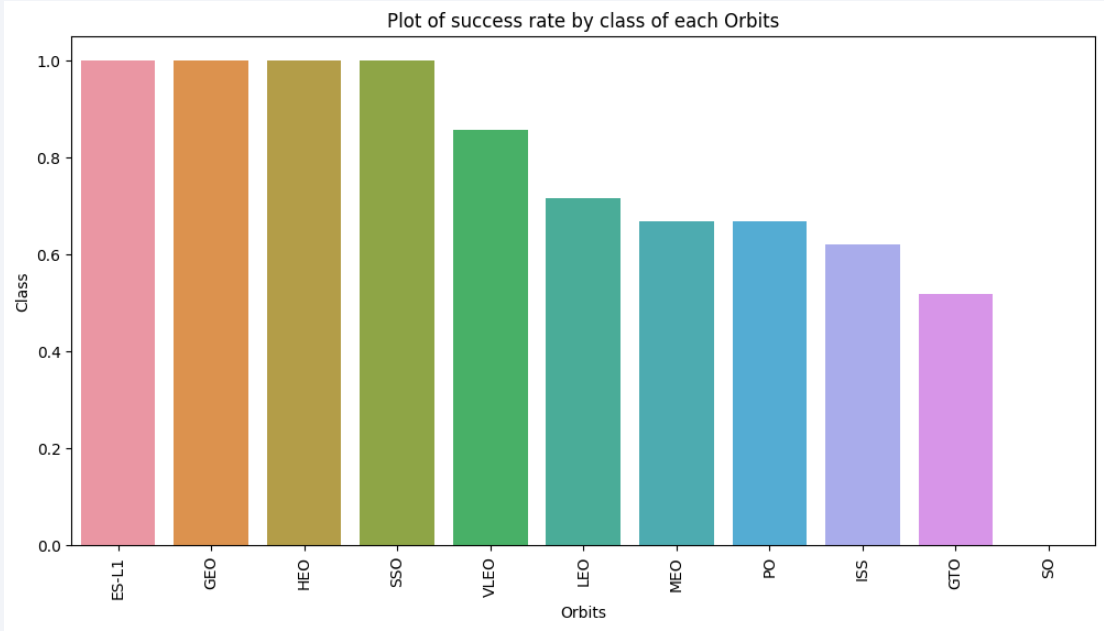
Data Wrangling

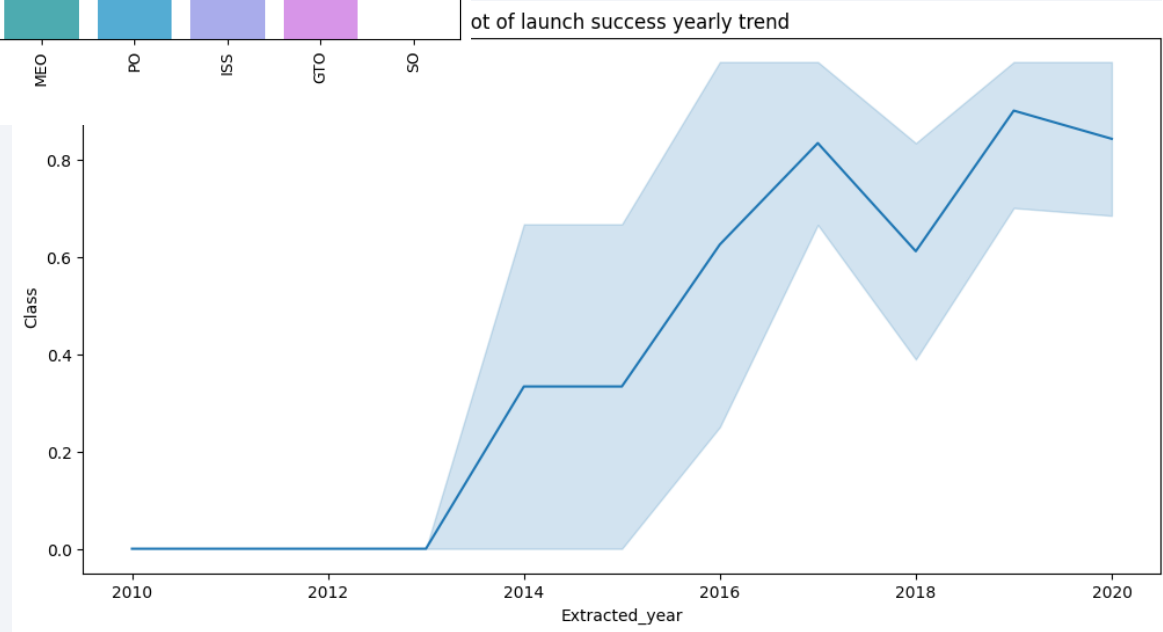
- Load data, assess missing data, assess launches at each site, calculate number and occurrence of mission outcome per orbit type
- Result = create landing outcome column
- Add the GitHub URL of your completed data wrangling related notebooks
  - https://github.com/paulhegedus/IBM-DataScienceCapstone/blob/c9380f7e05450ee0c9df2e7e411804dccb255911/03_DataWrangling.ipynb

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPa
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	Nal
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	Nal
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	Nal
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	Nal
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	Nal

EDA with Data Visualization

- Charts
  - Flight Number vs Payload Mass
  - Flight Number vs Launch Site
  - Payload vs Launch Site
  - Orbit vs Flight Number
  - Payload vs Orbit Type
  - Orbit vs Payload Mass
- Add the GitHub URL of your completed EDA with data visualization notebook
  - https://github.com/paulhegedus/IBM-DataScienceCapstone/blob/c9380f7e05450ee0c9df2e7e411804dccb255911/05_EDA_DataVisualization.ipynb





EDA with SQL

- SQL queries performed
  - Names of unique launch sites
  - Total payload mass carried by NASA (CRS) boosters
  - Average payload mass by Falcon 9 boosters
  - Total number of successful and failed missions
  - Failed landing outcomes with booster version and launch site
- Add the GitHub URL of your completed EDA with SQL notebook
  - https://github.com/paulhegedus/IBM-DataScienceCapstone/blob/c9380f7e05450ee0c9df2e7e411804dccb255911/04_EDAwSQL.ipynb

Build an Interactive Map with Folium

- Visualizing Launch Data
  - Lat and long coordinates implemented with circle marker and label of launch site
  - Successful launches labeled in green, failed launches labeled in red with markerCluster
  - Distance to coastline mapped with lines to indicate proximity to ocean
  - Distance to highways mapped with lines to indicate proximity to transportation
  - Distance to cities mapped with lines to indicate proximity to human populations
  - Distance to railways mapped with lines to indicate proximity to transportation
- Add the GitHub URL of your completed interactive map with Folium map
  - https://github.com/paulhegedus/IBM-DataScienceCapstone/blob/c9380f7e05450ee0c9df2e7e411804dccb255911/06_LaunchSiteLocation_Folium.ipynb

Build a Dashboard with Plotly Dash

- Plotly Dash dashboard created
  - Pie charts showing total launches by all sites
  - Scatter plots showing outcome and payload mass for booster version

- Add the GitHub URL of your completed Plotly Dash lab
  - https://github.com/paulhegedus/IBM-DataScienceCapstone/blob/c9380f7e05450ee0c9df2e7e411804dccb255911/05_EDA_DataVisualization.ipynb

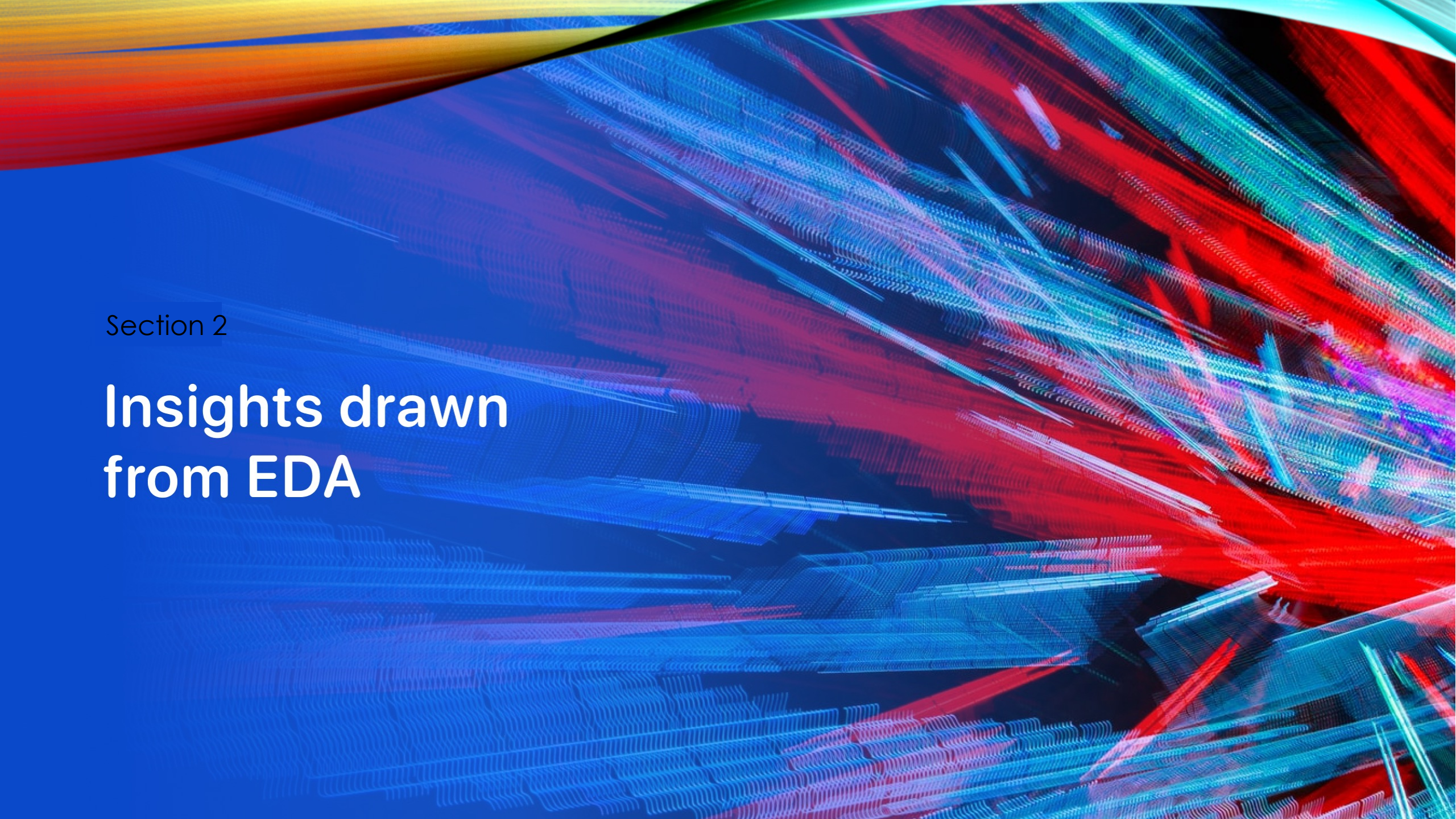
Predictive Analysis (Classification)

- Logistic regression, Support vector machine, decision tree, and KNN algorithms used
- Models fit and grid search used to tune hyperparameters
- Models evaluated with confusion matrix
- Accuracy score calculated on test datasets
- Add the GitHub URL of your completed predictive analysis lab
  - https://github.com/paulhegedus/IBM-DataScienceCapstone/blob/c9380f7e05450ee0c9df2e7e411804dccb255911/07_MachineLearningPrediction.ipynb

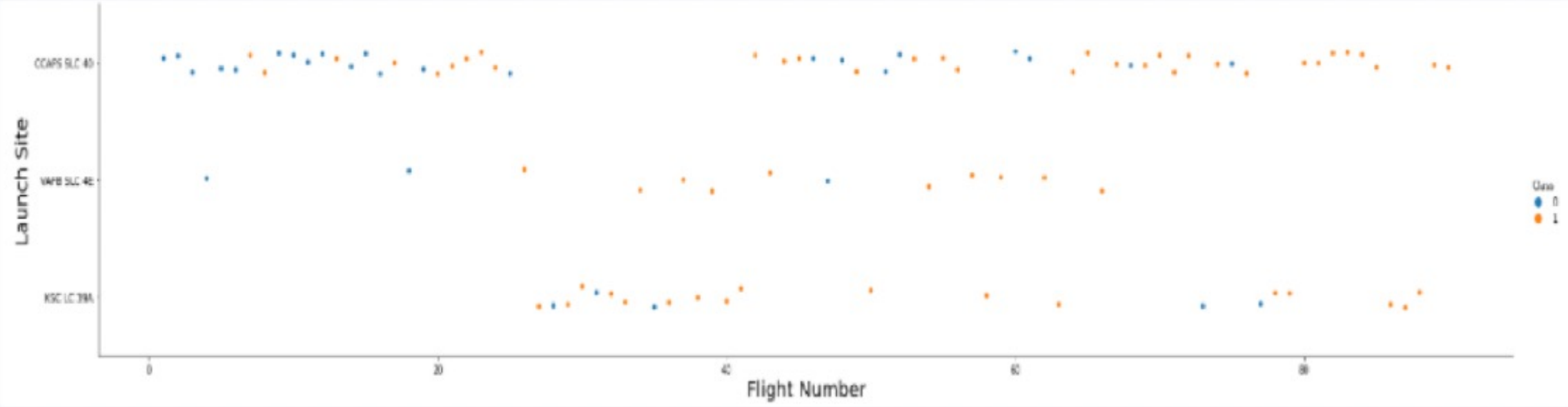
	Accuracy Score
Logistic Regression	0.8464
SVM	0.8482
Decision Tree	0.8732
KNN	0.8482

Results

- Exploratory data analysis results
  - 80 Falcon 9 launch observations
- Interactive analytics demo in screenshots
  - Not close proximity to transportation or human populations. Close proximity to coast for safety
- Predictive analysis results
  - Logistic regression best model

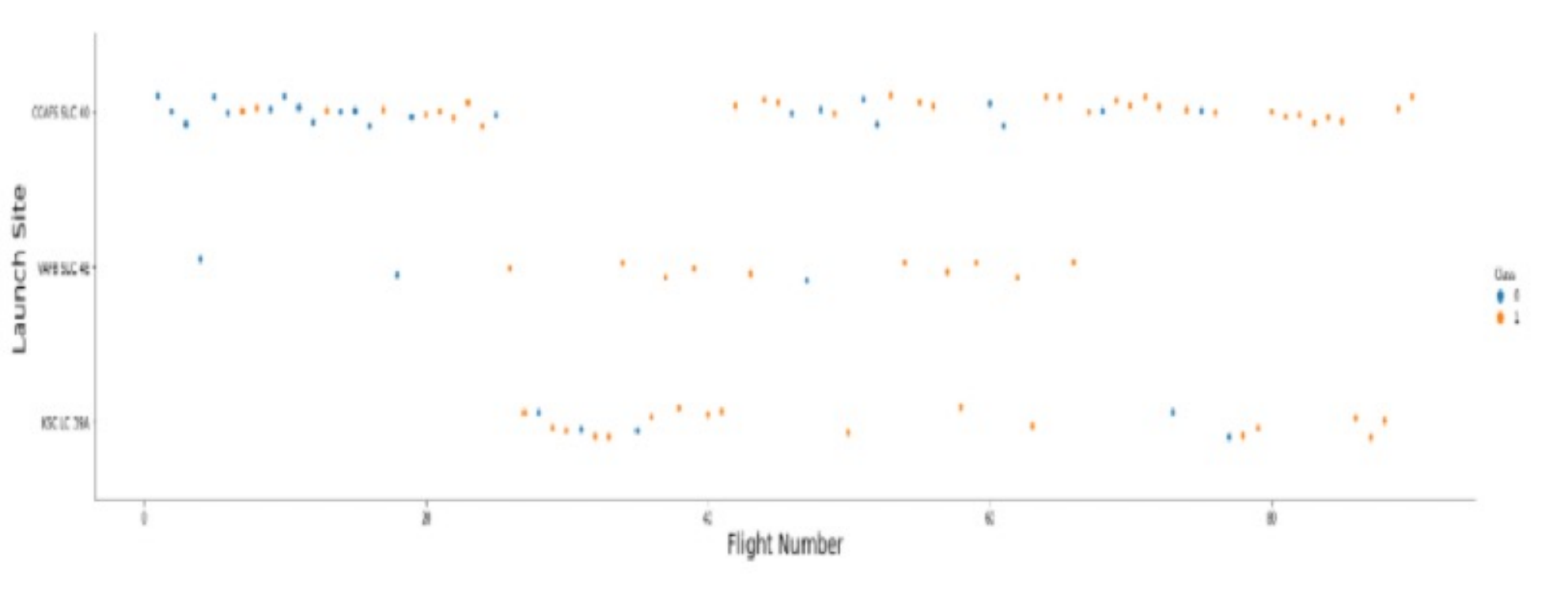

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



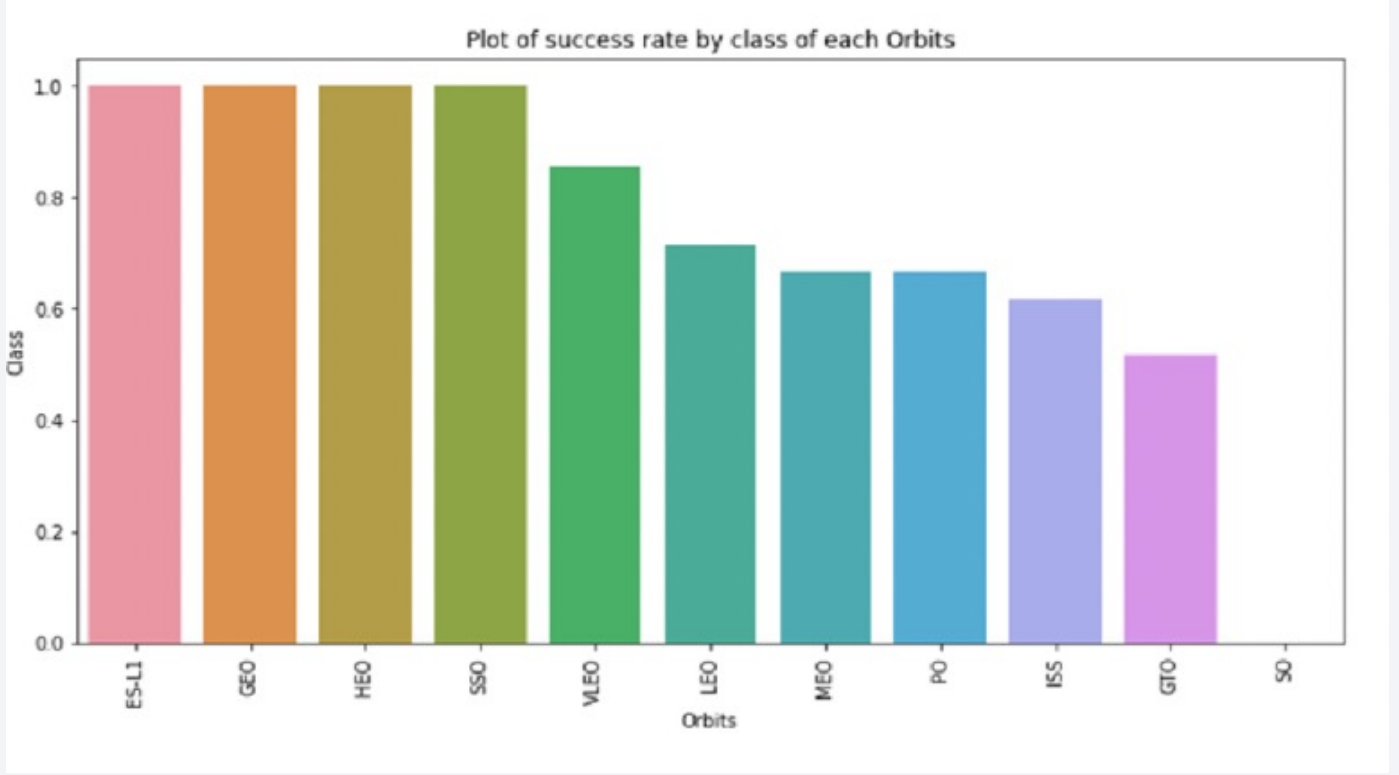
- More amount of flights at launch site = more success

Payload vs. Launch Site



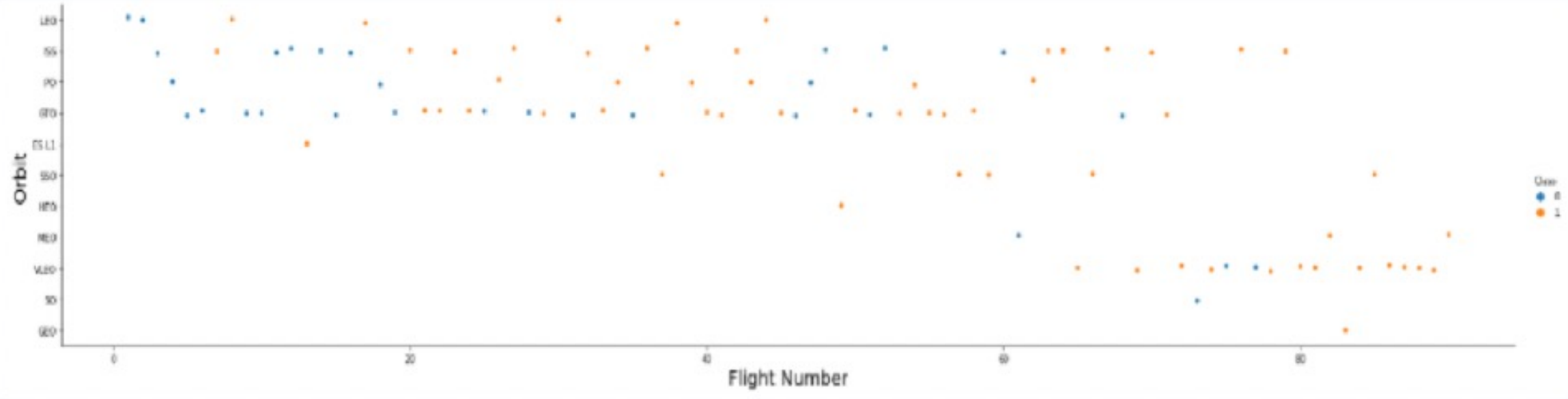
- Larger payload = higher success

Success Rate vs. Orbit Type



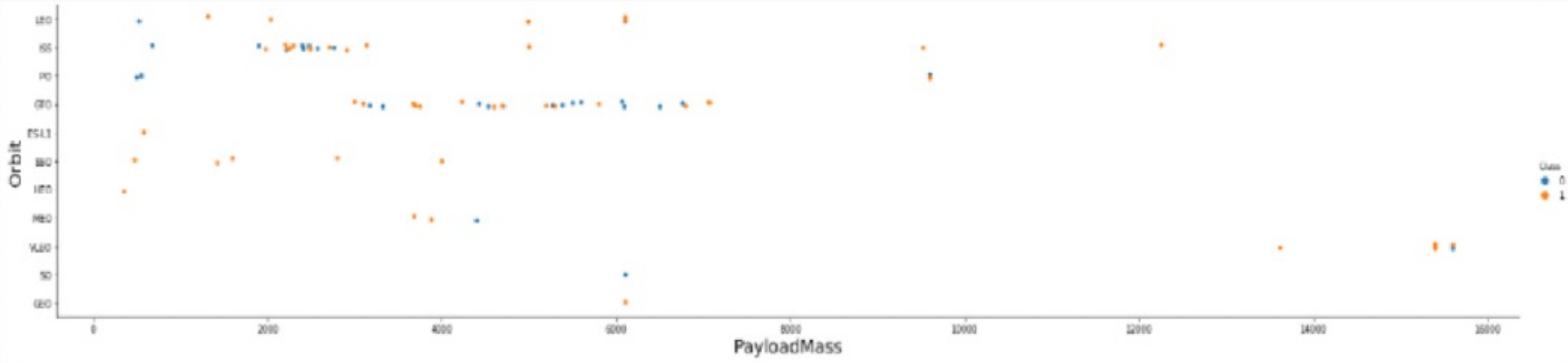
- ES-L1, GEO, HEO, SSO had the highest success rate

Flight Number vs. Orbit Type



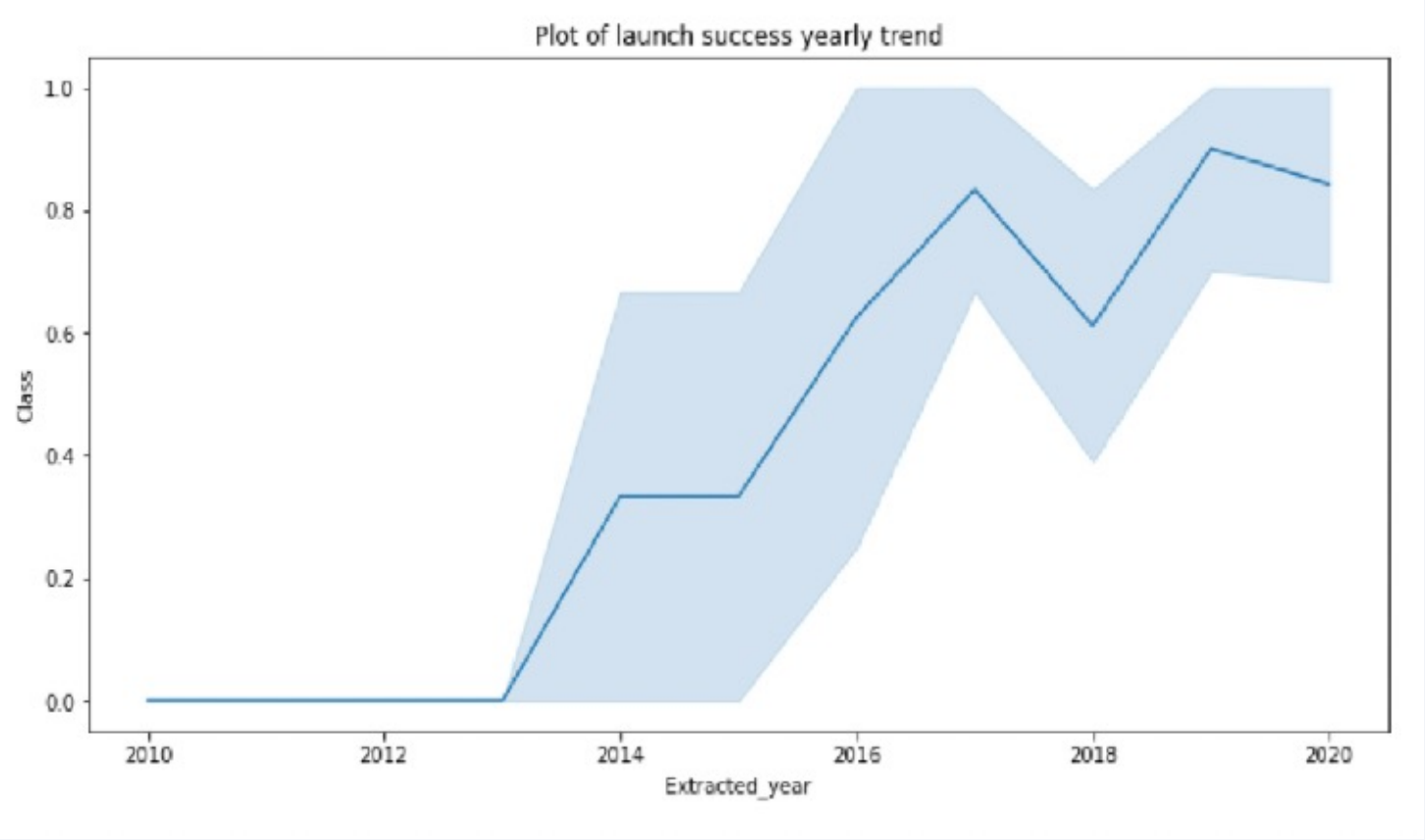
- LEO had more success with more flights
- GEO has no relationship

Payload vs. Orbit Type



- Heavy payload = successful landing for all orbits

Launch Success Yearly Trend



- Success rate increases over time

All Launch Site Names


```
%%sql
SELECT DISTINCT LAUNCH_SITE
FROM SPACEXTBL
```

- Unique launch site names derived from Space X data

0	KSC LC-39A
1	CCAFS LC-40
2	CCAFS SLC-40
3	VAFB SLC-4E

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

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

	date	time	boosterversion	launchsite	payload	payloadmasskg	orbit	customer	missionoutcome	landingoutcome
0	2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	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
3	2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS Total_PayloadMass
FROM SPACEXTBL
WHERE CUSTOMER LIKE 'NASA (CRS)'
```

Total_PayloadMass
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS Avg_PayloadMass
FROM SPACEXTBL
WHERE BOOSTER_VERSION = 'F9 v1.1'
```

Avg_PayloadMass
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

```
%%sql
SELECT MIN(DATE) AS FirstSuccessfull_landing_date
FROM SPACEXTBL
WHERE LANDING__OUTCOME LIKE 'Success (ground pad)'
```

FirstSuccessfull_landing_date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
%%sql
SELECT BOOSTER_VERSION
FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Success (drone ship)'
AND PAYLOAD_MASS__KG_ > 4000
AND PAYLOAD_MASS__KG_ < 6000
```

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

```
%%sql
SELECT COUNT(MISSION_OUTCOME) AS SuccessOutcome
FROM SPACEXTBL
WHERE MISSION_OUTCOME LIKE 'Success%'

SELECT COUNT(MISSION_OUTCOME) AS FailureOutcome
FROM SPACEXTBL
WHERE MISSION_OUTCOME LIKE 'Failure%'
```

100

1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

```
%%sql
SELECT BOOSTER_VERSION, PAYLOAD_MASS__KG_
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_)
                           FROM SPACEXTBL)
ORDER BY BOOSTER_VERSION
```

F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

2015 Launch Records

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

```
%%sql
SELECT BOOSTER_VERSION, LAUNCH_SITE, LANDING__OUTCOME
FROM SPACEXTBL
WHERE LANDING__OUTCOME LIKE 'Failure (drone ship)'
AND DATE BETWEEN '2015-01-01' AND '2015-12-31'
```

F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- 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)
FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING__OUTCOME
ORDER BY COUNT(LANDING__OUTCOME) DESC
```

No attempt	10
Success (drone ship)	6
Failure (drone ship)	5
Success (ground pad)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

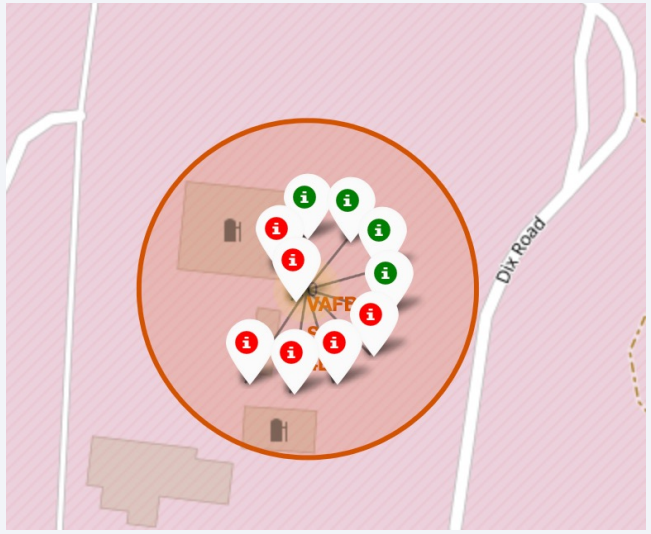
All Launch Sites

- Launch sites in CA and FL only

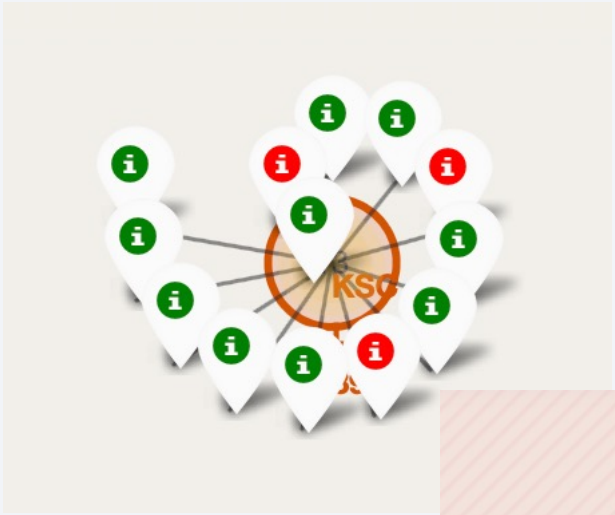
Launch Sites with Colors

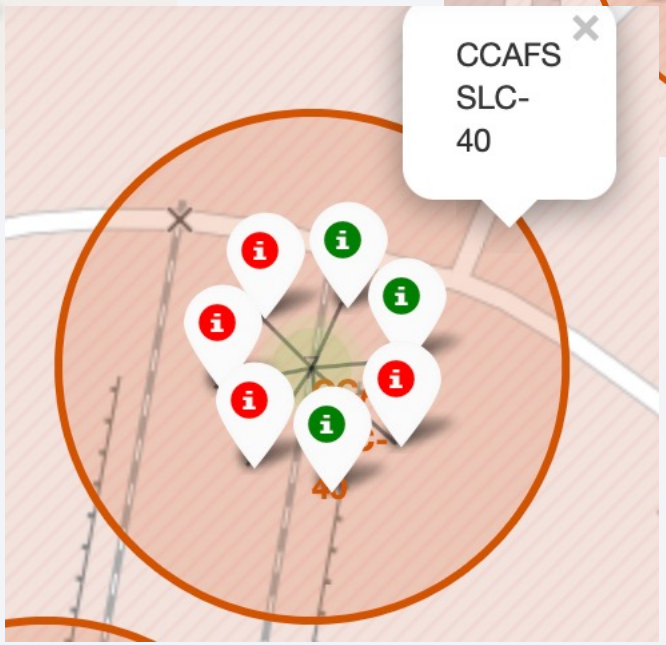
- Green = Success
- Red = Failure

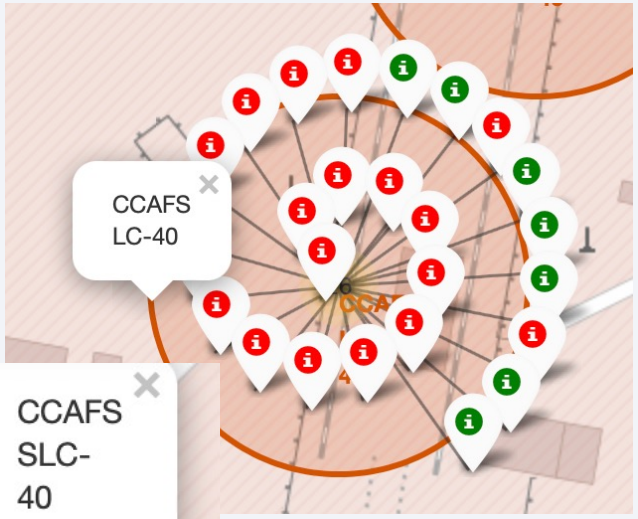
California



Florida

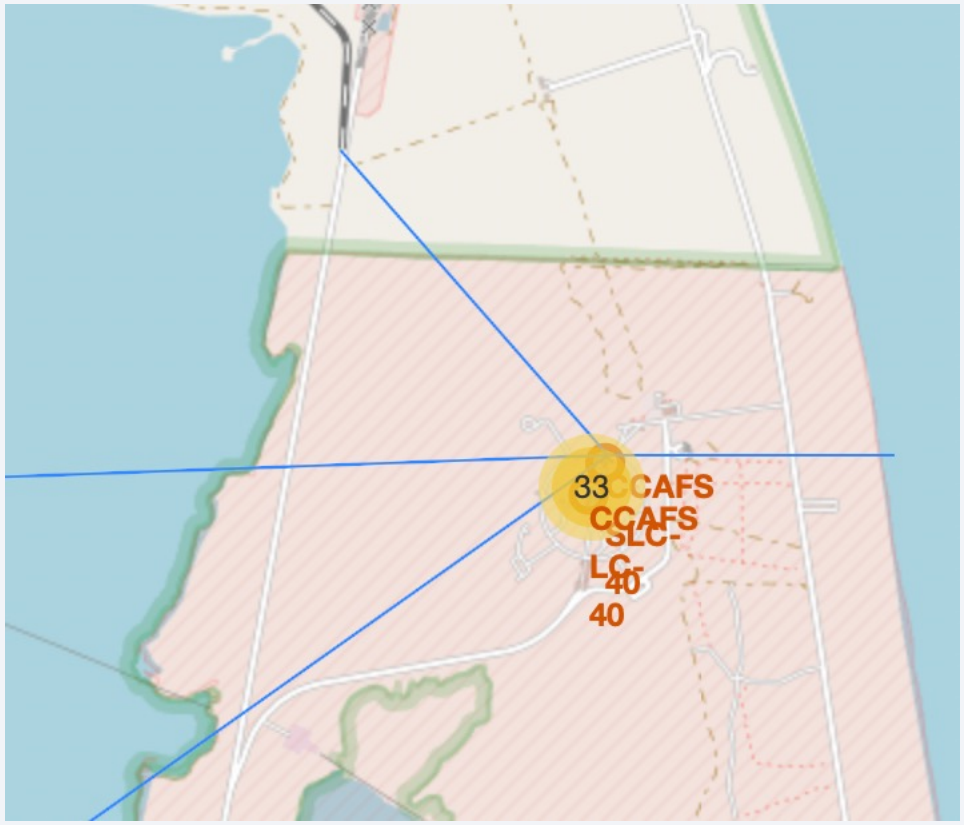






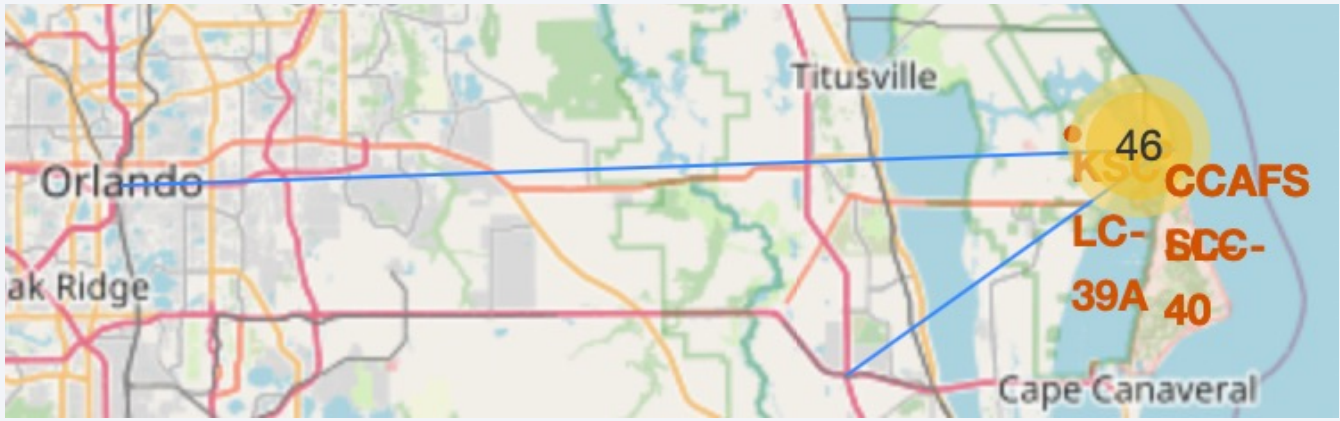
Launch Site Proximities

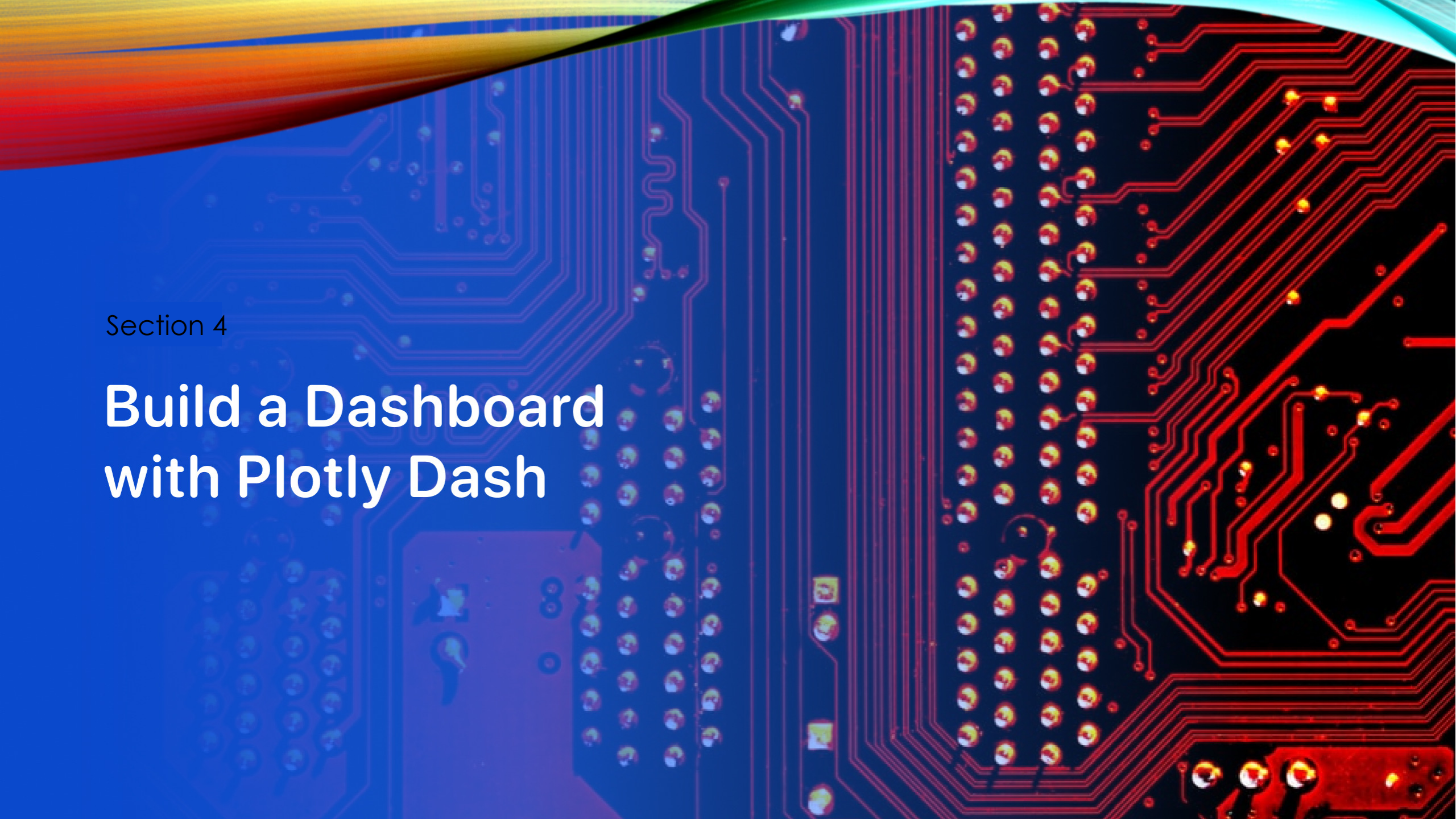
- Proximity to Coast and Railway



- **CCAFS-LC40 close to coast and railway**
- **CCAFS-LC40 not close to major city and highway**

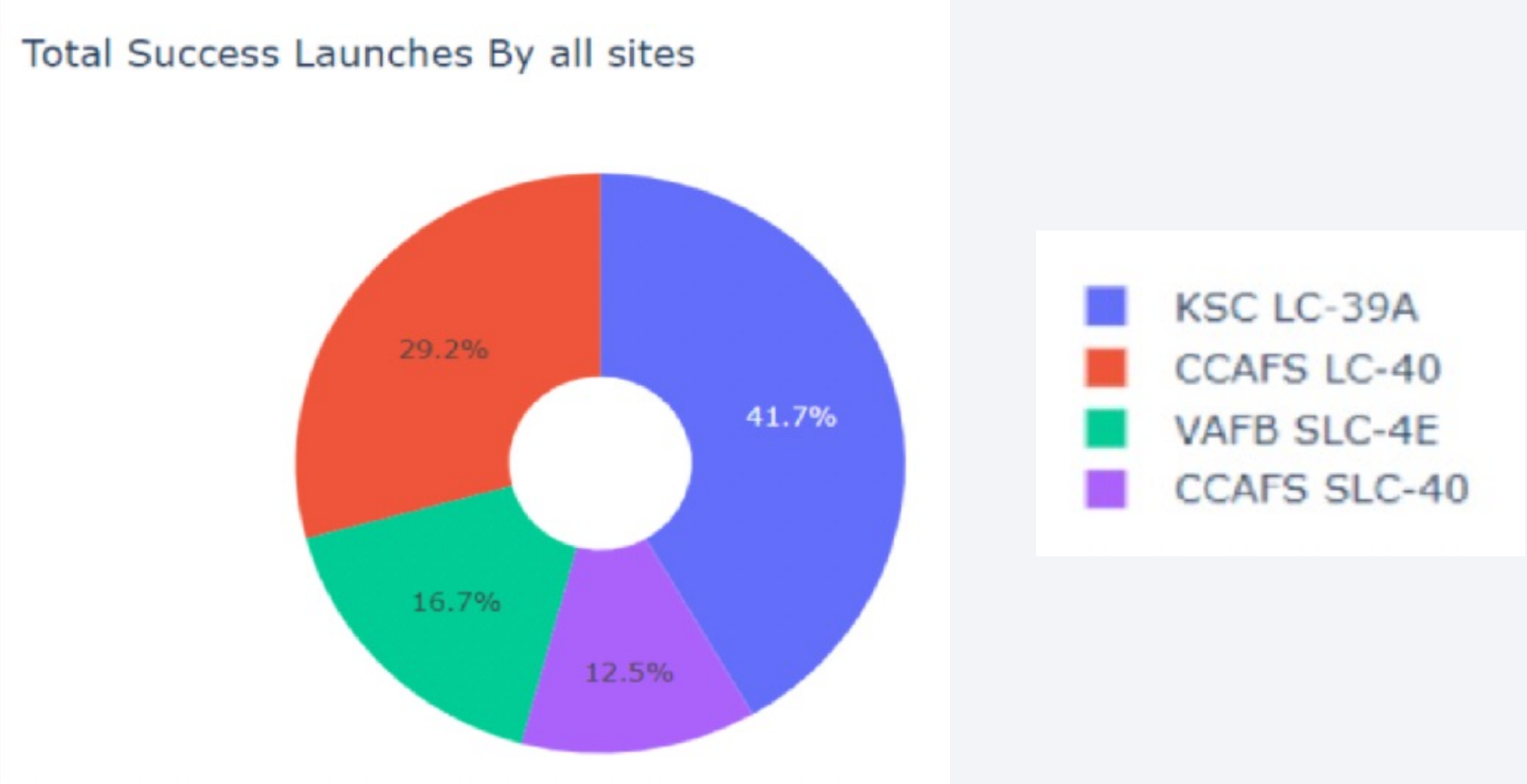
- Proximity to Major City and Highway




Section 4
Build a Dashboard with Plotly Dash

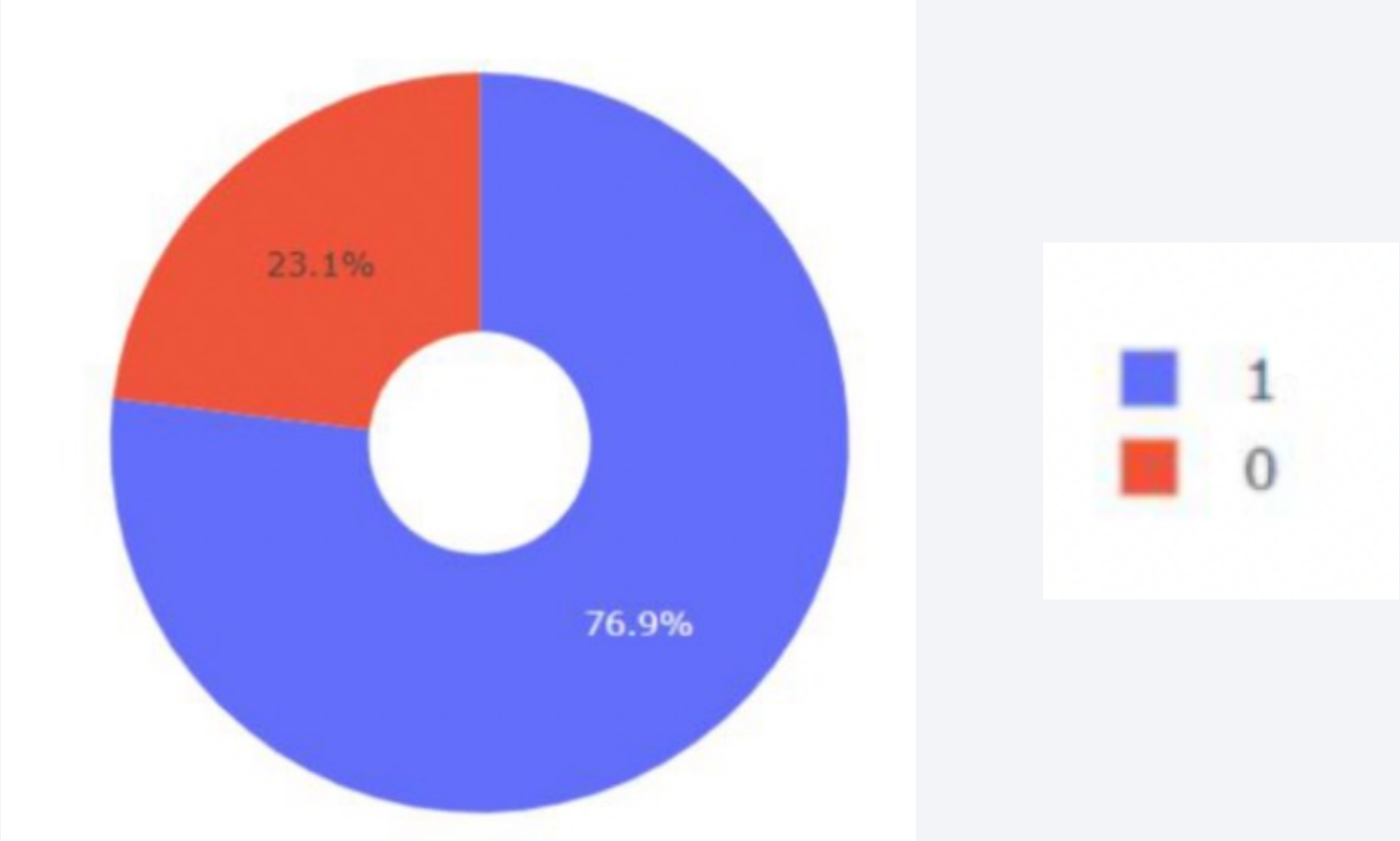
Success % by Launch Site

- KSC LC-39A most successful site



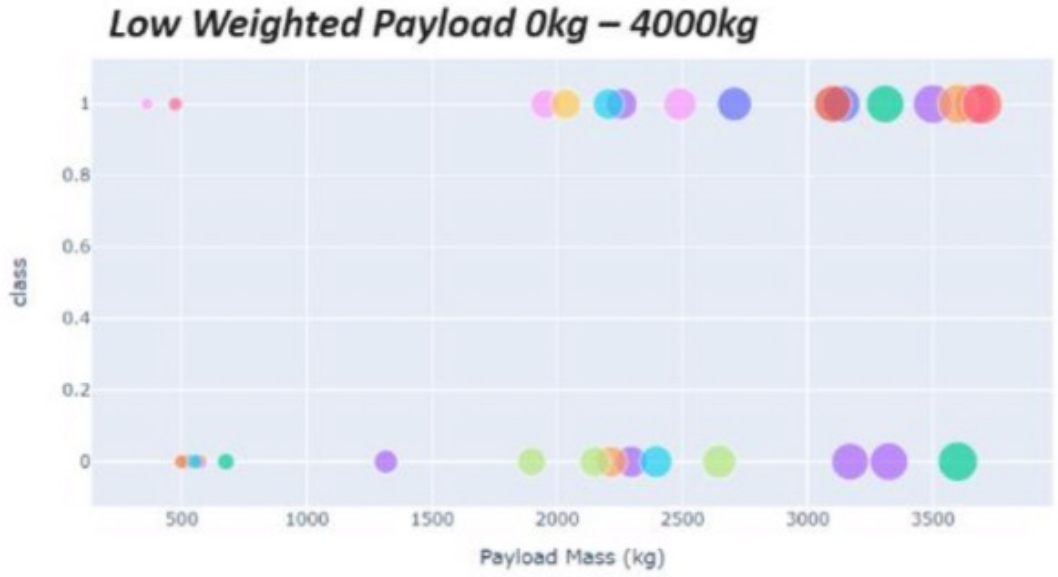
KSC LC-39A Launch Success

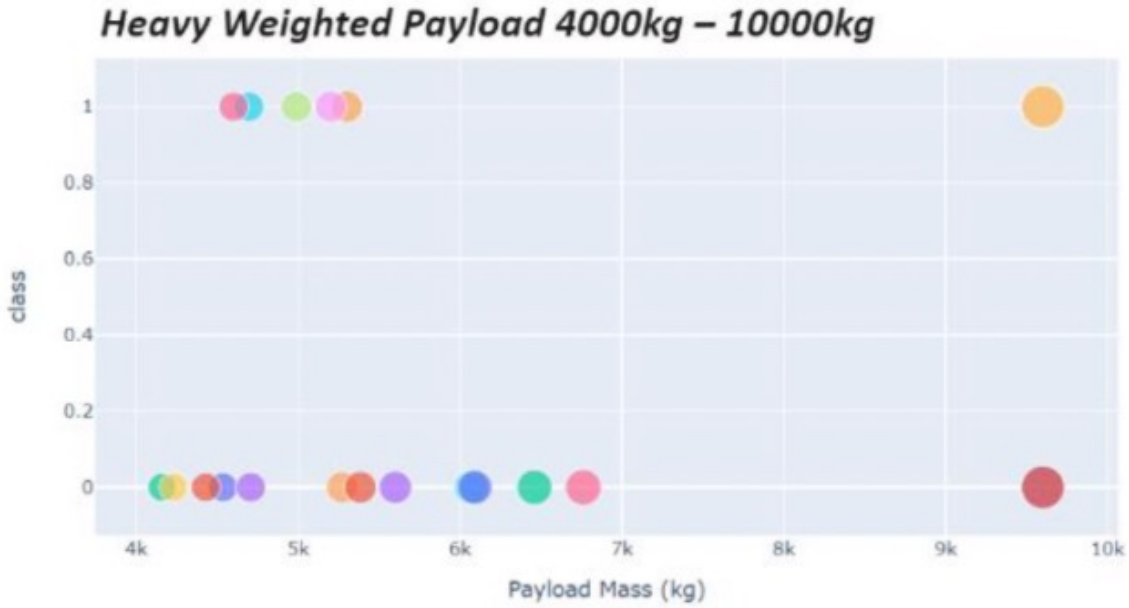
- 76.9% Successful Launches (1), 23.1% Failed Launches (0)



Payload vs Launch Outcome

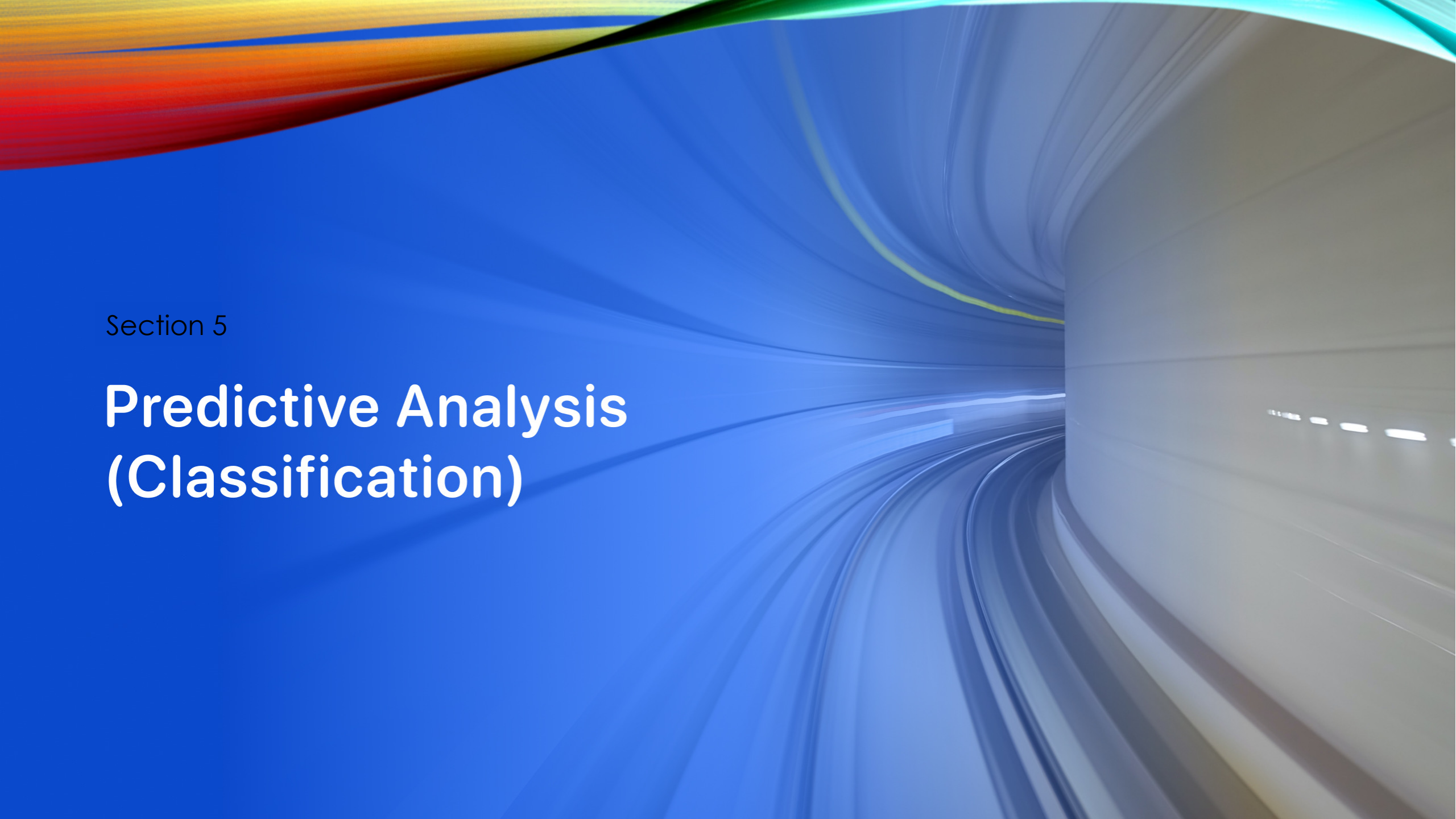
Low weight payloads = more success
Heavy weight payloads = less success





Section 5

Predictive Analysis (Classification)



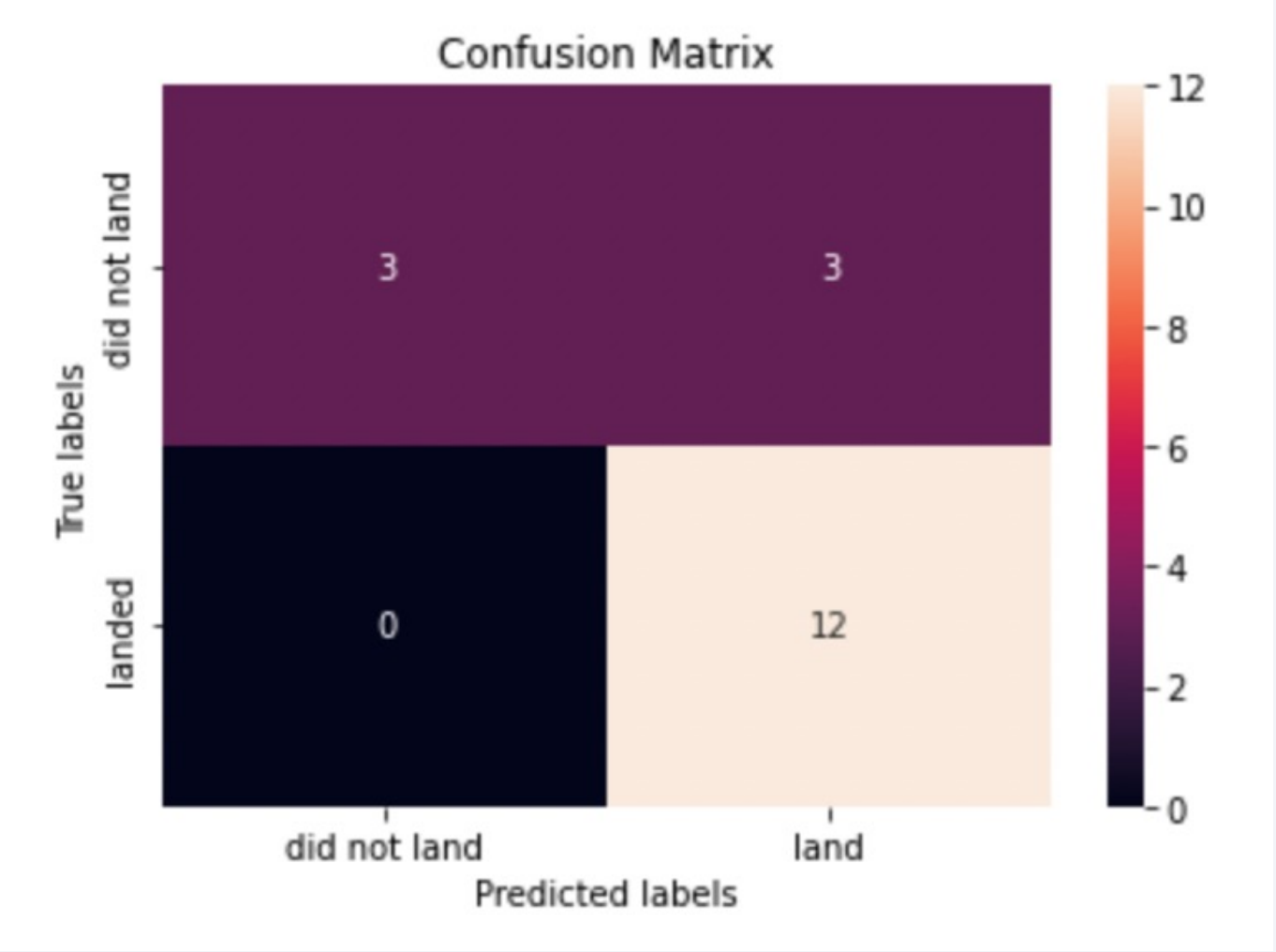
Classification Accuracy

- Logistic Regression = highest accuracy

	Accuracy Score
Logistic Regression	0.8464
SVM	0.8482
Decision Tree	0.8732
KNN	0.8482

Confusion Matrix

- False positives = major problem
  - Predicted landing but actually did not land
- No False Negatives!



Conclusions

- ES-L1, GEO, HEO, SSO were most successful orbits with 100% success
- Success increases over time
- KSC LC-39A most successful launch site
- Low weight payloads have more success than heavy payloads
- Logistic Regression was the best machine learning classifier

Thank you!