IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Matt Kirsling
8/12/2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- Summary of methodologies
  - Data collection from API and web scraping
  - Data wrangling
  - Exploratory data analysis using SQL, Pandas, Matplotlib
  - Interactive visual analytics and dashboards with Folium and Dash
  - Predictive analysis
- Summary of all results
  - The best hyperparameters for Logistic Regression, SVM, Decision Tree and KNN classifiers.
  - The best performing method

Introduction

- Project background and context

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. The goal of this project is to create a machine learning model to predict if the first stage will land successfully.

- Problems you want to find answers
  - What features contribute to a successful landing?
  - Which machine learning model provides the most accurate prediction?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data was gathered via a REST API and via web scraping wiki pages.
- Perform data wrangling
  - The data was searched for null values and outcomes were reformatted to a binary success field.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

Sources:

SpaceX Rest API

Wikipedia

Transformations:

Unnecessary Columns Removed

Nulls Removed

Outcomes mapped to 1 and 0

Result:

Dataframe

Data Collection – SpaceX API

- Data was collected from the SpaceX API using request.get. The JSON result was reformatted into a dataframe like below. Nulls were dealt with.

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857

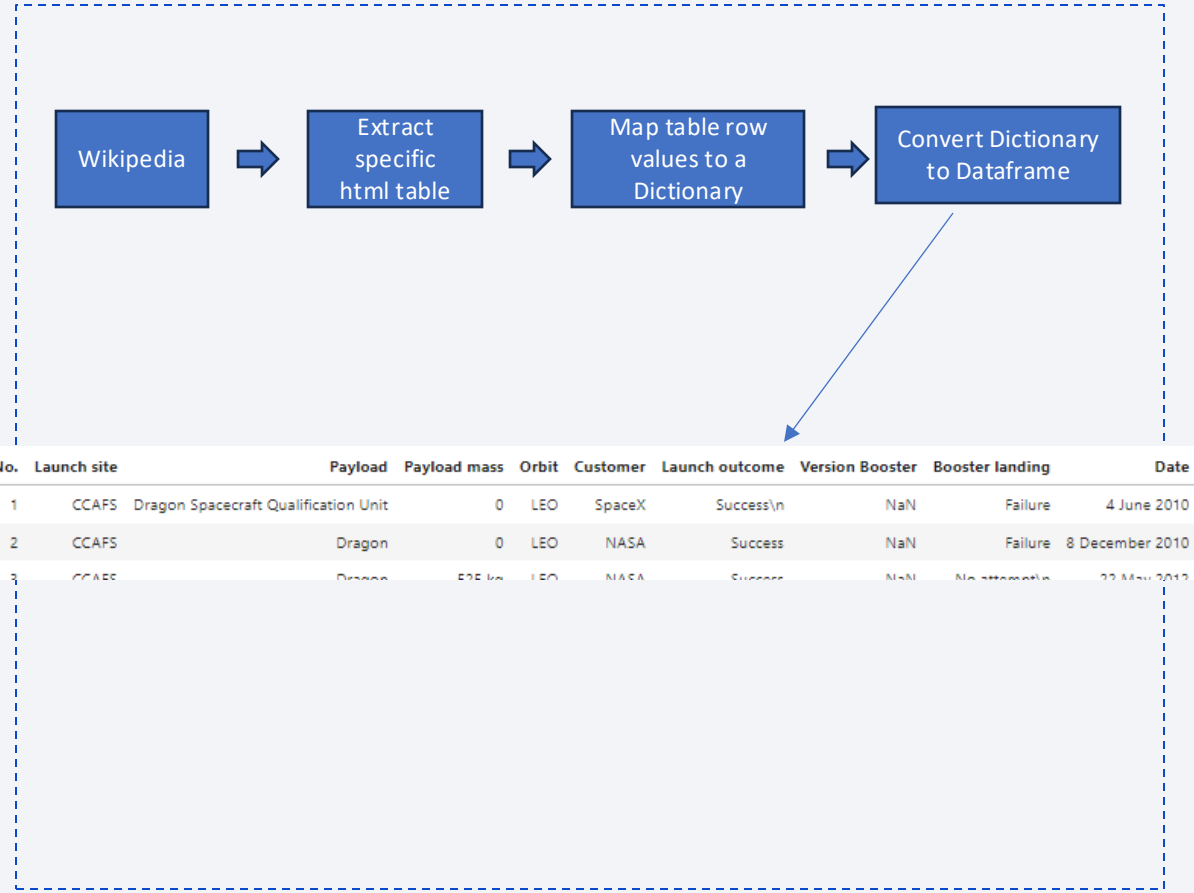
- The link to the notebook is [Data collection notebook]


```
# Calculate the mean value of PayloadMass column
payload_mean = data_falcon9['PayloadMass'].mean()

# Replace the np.nan values with its mean value
data_falcon9['PayloadMass'] = data_falcon9['PayloadMass'].replace(np.nan, payload_mean)

data_falcon9.isnull().sum()
```

Data Collection - Scraping

- Gather historical Falcon 9 launch records from wikipedia. Reformat HTML into a Pandas Dataframe.

- The link to the notebook is [Webscraping notebook]

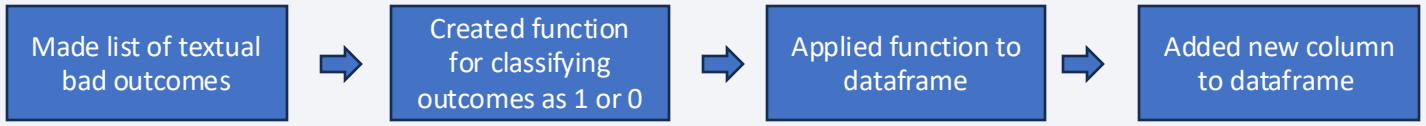


28]:

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	NaN	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	NaN	Failure	8 December 2010	15:43

Data Wrangling

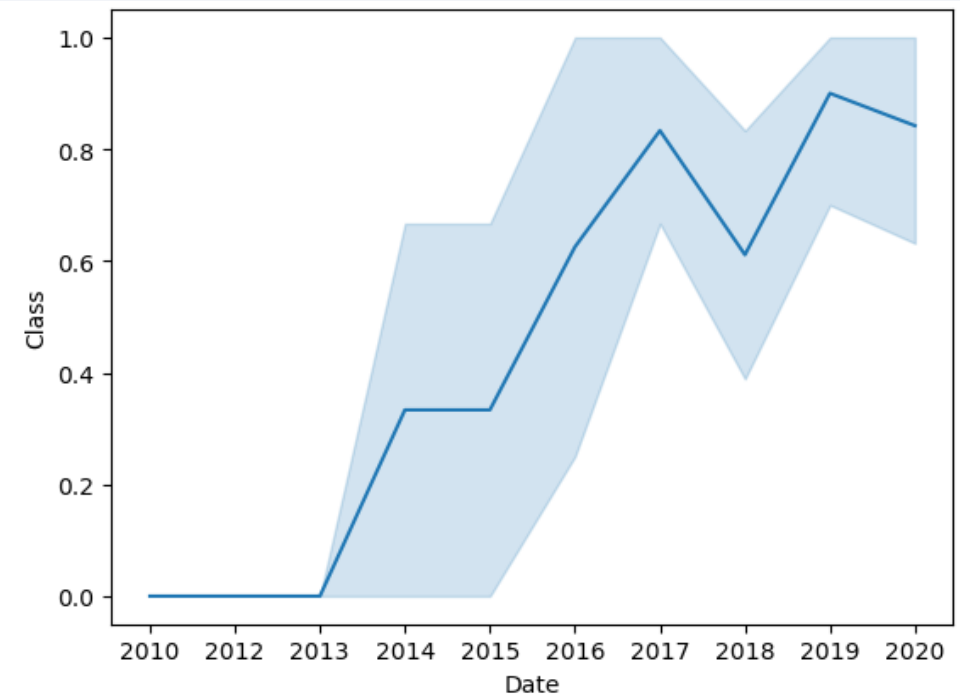
- Reviewed value counts for orbit and launch site.
- Created outcome column for model training.



- Found the mean success rate to be 67%
- The link to the notebook is: Data wrangling notebook

EDA with Data Visualization

- We used a scatterplots to compare the success rates by payload mass, orbit type and launch site. Overall, the success rate increased over time per the trend below.



- The link to the notebook is: EDA notebook

EDA with SQL

- Verified the number of launch sites: 4
- Calculated total payload mass for NASA (CRS): 45,596
- Calculated average payload for booster 'F9 v1.1': 2,928.4
- Found earliest success date: 22/12/2015
- Found boosters with drone ship success and a payload mass > 4000 and < 6000
- Listed counts for each mission outcome
- Found boosters using the maximum payload mass
- Found outcomes by month for 2015
- Ranked landing outcomes between 2010-6-4 and 2017-3-20
- The link to the notebook is: EDA with SQL

Build an Interactive Map with Folium

- We marked the launch sites and added features to each site to indicate the success or failure of launches at the site.
- We looked for proximity to features such as: railways, highways and coastlines.
- The link to the notebook is: Folium notebook

Build a Dashboard with Plotly Dash

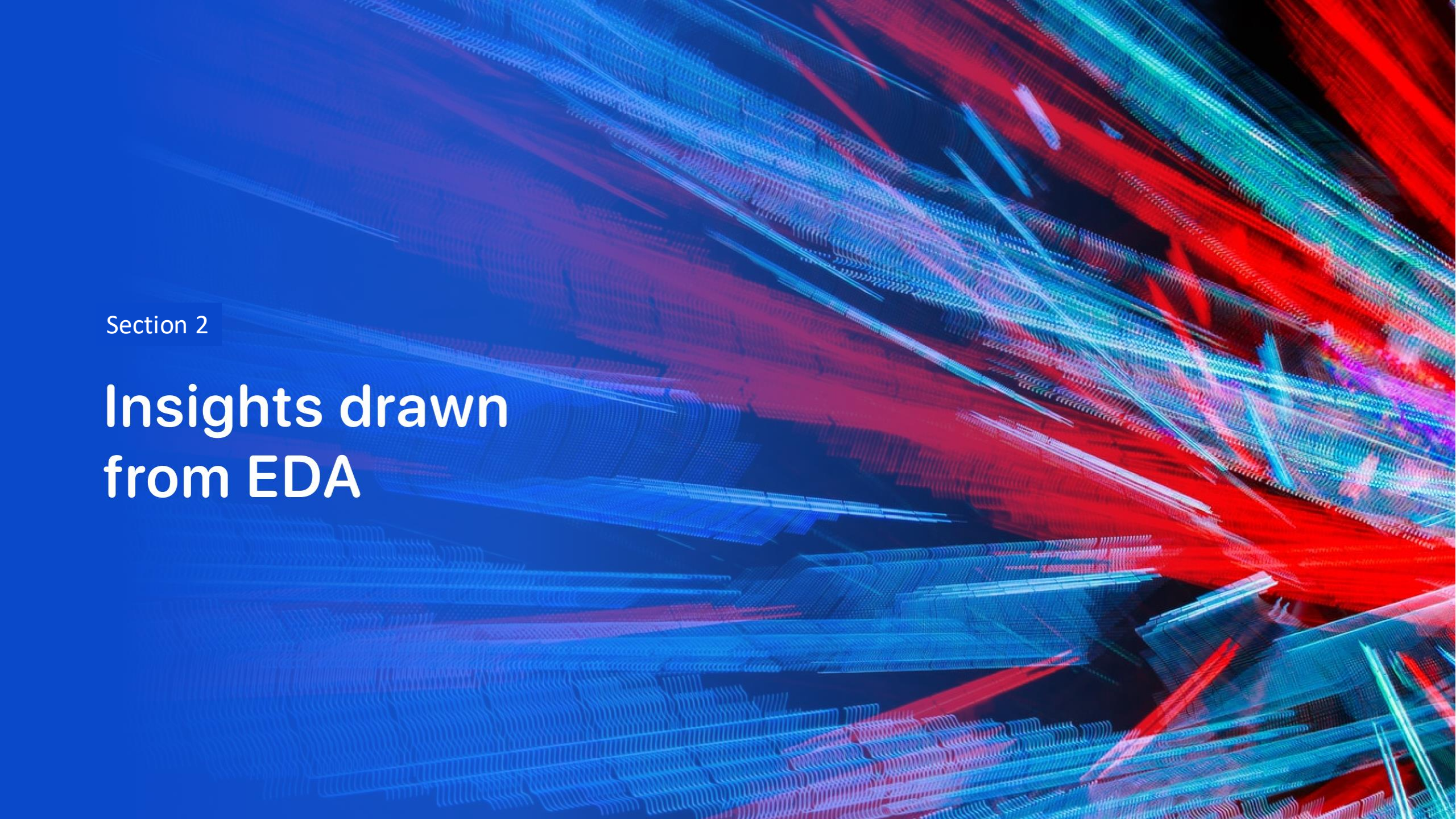
- We used a pie chart to summarize total launches by site.
- We used scatterplot to compare Outcome and Payload Mass by booster version.
- The link to the notebook is: Dash notebook

Predictive Analysis (Classification)

- We standarized X by using a StandardScaler
- We split the data into a training set and a test set (20% of data set and a random state of 2)
- We evaluated the following models using GridSearchCV: Logistic Regression, SVM, Decision Trees, KNN.
- The link to the notebook is: Model notebook

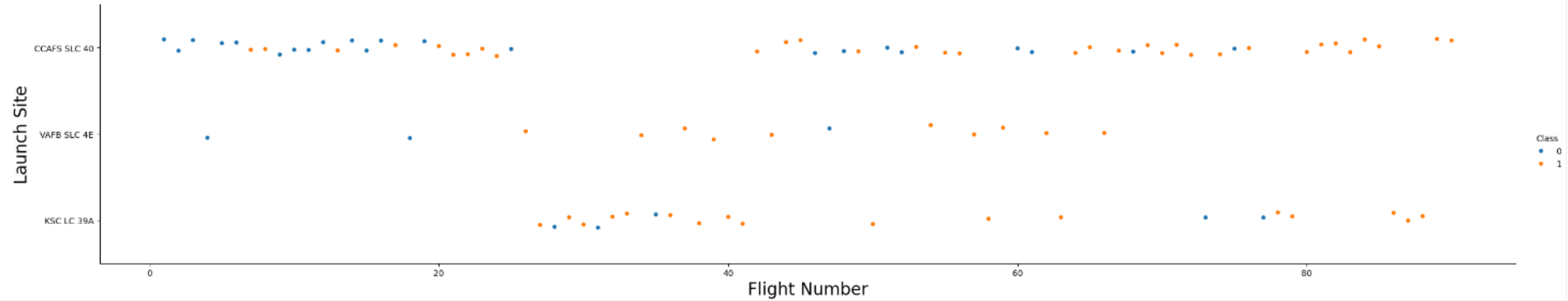
Results

- Exploratory data analysis results
  - Launch success improved over time
  - KSC LC-39A has highest success rate
  - These orbits had 100% success rate: ES-L1, GEO, HEO and SSO.
- Interactive analytics demo in screenshots
  - All launch sites are near coastlines and distanced from cities.
- Predictive analysis results
  - The Decision Tree model performed the best


Section 2
Insights drawn from EDA

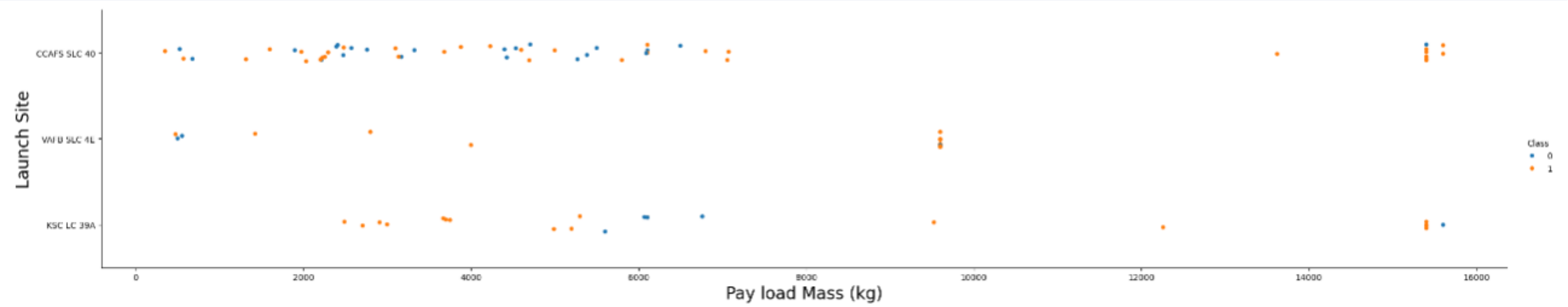
Flight Number vs. Launch Site

- From the plot we saw success increase over time. The CCAFS site had a higher failure rate with early launches.



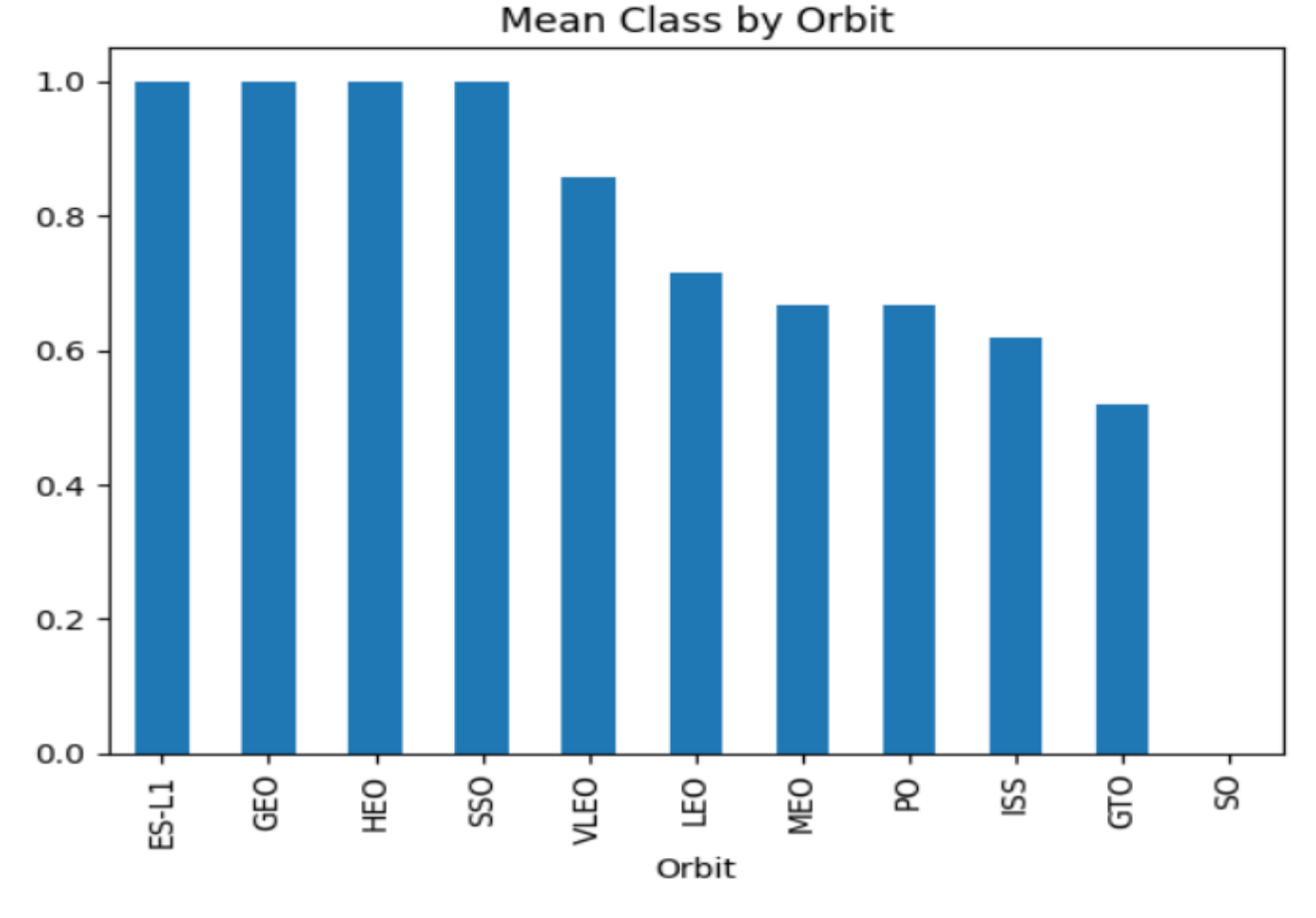
Payload vs. Launch Site

- The VAFB did not launch any rockets > 10,000 kg and had a high success rate across all launches.
- CCAFS had a higher success rate with higher payloads



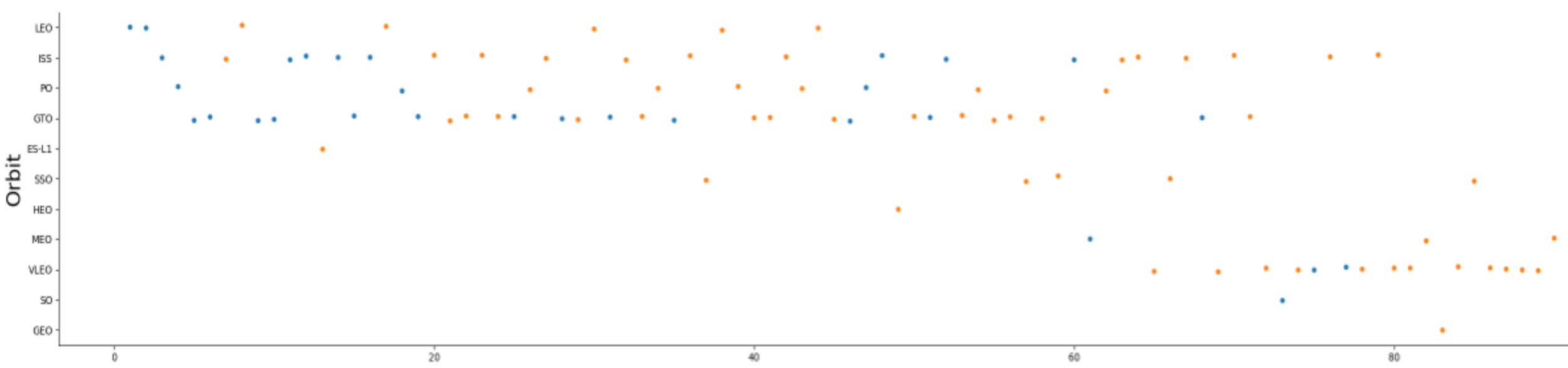
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO and SSO had the highest success rates



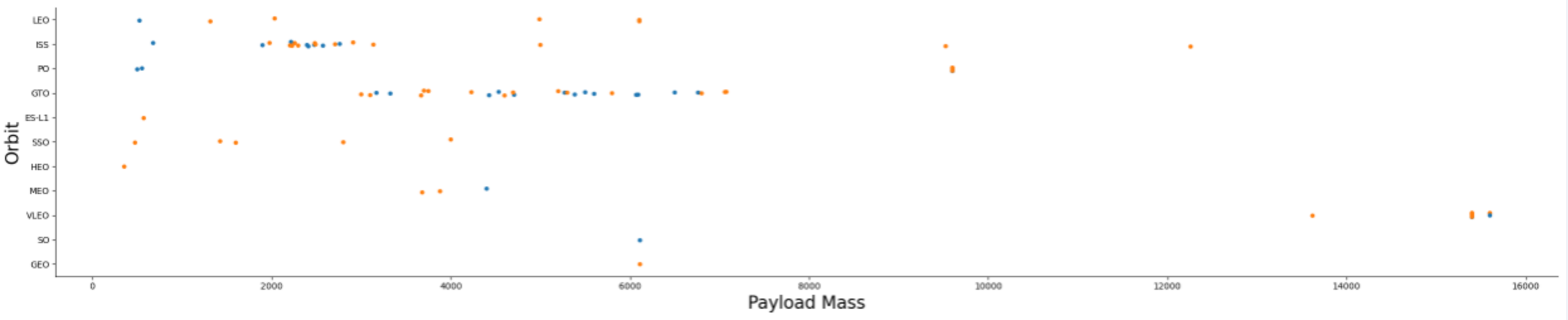
Flight Number vs. Orbit Type

- LEO and VLEO had higher success over time
- GTO did not improve significantly over time



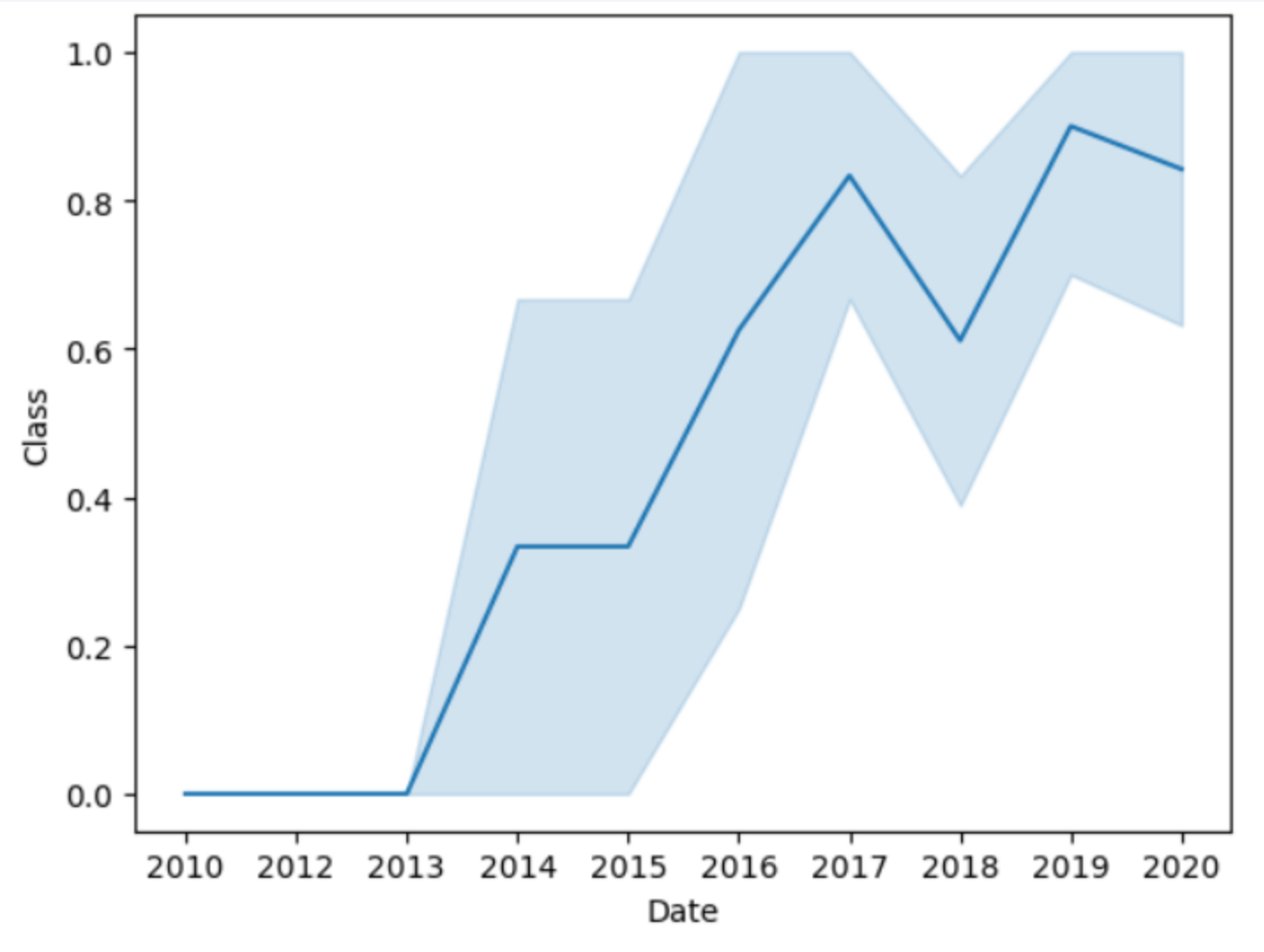
Payload vs. Orbit Type

- Heavier paylods had greater success for PO, LEO and ISS



Launch Success Yearly Trend

- The success rate has increased from 2013 to 2020



All Launch Site Names

- Used distinct to get list of names
- select distinct Launch_Site from SPACEXTABLE

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

Launch Site Names Begin with 'CCA'

- Used LIMIT and wildcard '%' to filter launch sites
- select * from SPACEXTABLE where Launch_Site like 'CCA%' LIMIT 5

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

- Used SUM aggregate function and WHERE clause to calculate total
- select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer ='NASA (CRS)'

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Used AVG aggregate function and WHERE clause to calculate average.
- select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version ='F9 v1.1'

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Used MIN aggregate function and where clause with wildcard: '%' to find earliest successful launch date.

- n(Date) from SPACEXTABLE where Landing_Outcome like 'Success%'

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Used DISTINCT to limit list and used WHERE clause to filter records
- select distinct Booster_Version from SPACEXTABLE where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000

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

Total Number of Successful and Failure Mission Outcomes

- Used GROUP BY to calculate record count by Mission Outcome.
- select Mission_Outcome,count(*) from SPACEXTABLE group by Mission_Outcome

Mission_Outcome	count(*)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Used sub query to get maximum payload mass and filtered records using this value. Returned booster versions using DISTINCT.
- select distinct Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)

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

2015 Launch Records

- Added columns to results for year and month using SUBSTR function on the date field.
- select substr(Date, 6,2) dmonth, substr(Date,0,5) dyear,* from SPACEXTABLE where substr(Date,0,5)='2015'

dmonth	dyear	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
01	2015	2015-01-10	9:47:00	F9 v1.1 B1012	CCAFS LC-40	SpaceX CRS-5	2395	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)
02	2015	2015-02-11	23:03:00	F9 v1.1 B1013	CCAFS LC-40	DSCOVR	570	HEO	U.S. Air Force NASA NOAA	Success	Controlled (ocean)

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

- Displayed record counts by Landing Outcome using GROUP BY for specific date range.
- select Landing_Outcome,count(*) from SPACEXTABLE where Date between '2010-06-04' and '2017-03-20' group by Landing_Outcome order by count(*) desc

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

Section 3

Launch Sites Proximities Analysis

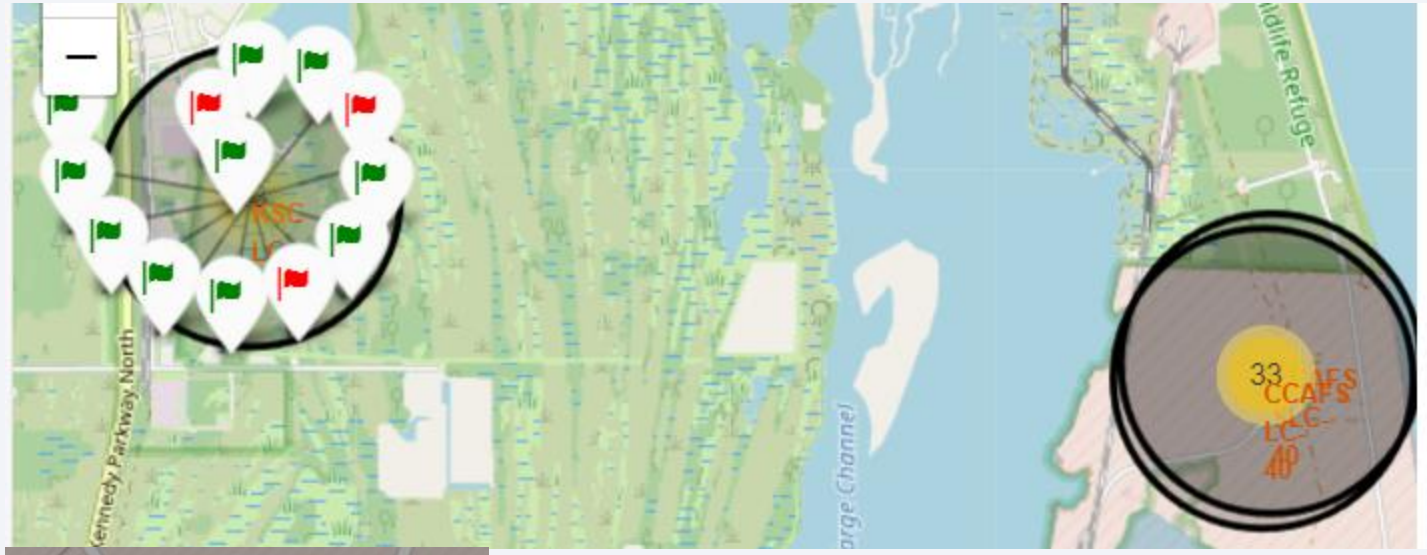
All launch sites

- Launchs were done from the East and West coasts of the United States of America

Launch site success rates

California Success Rate

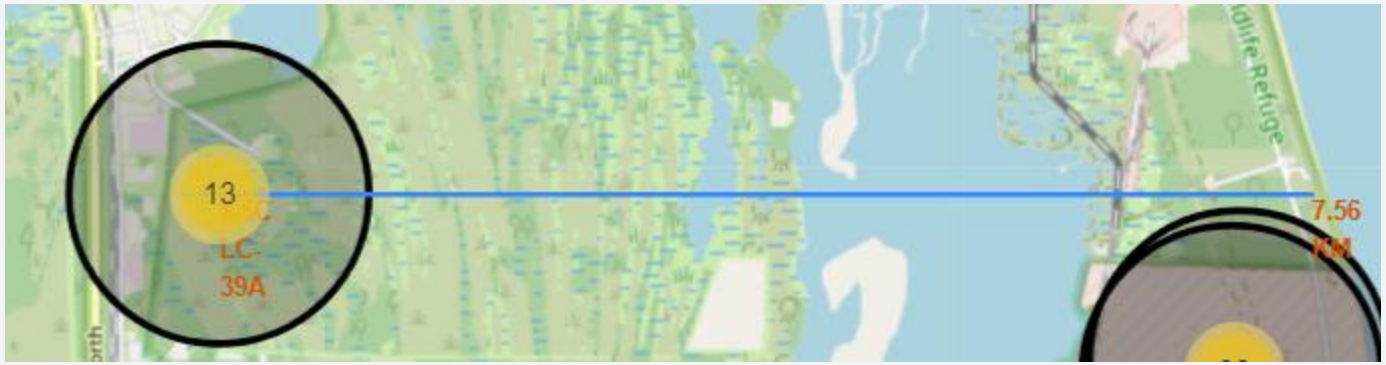
Florida Success Rate



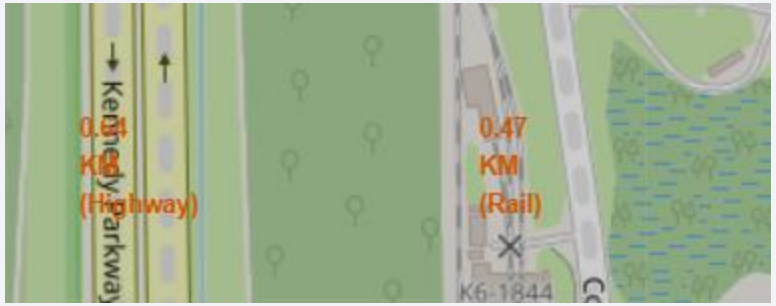
Green markers indicate success

Launch site KSCLC-39A distance to landmarks

- Distance to coast 7.56KM

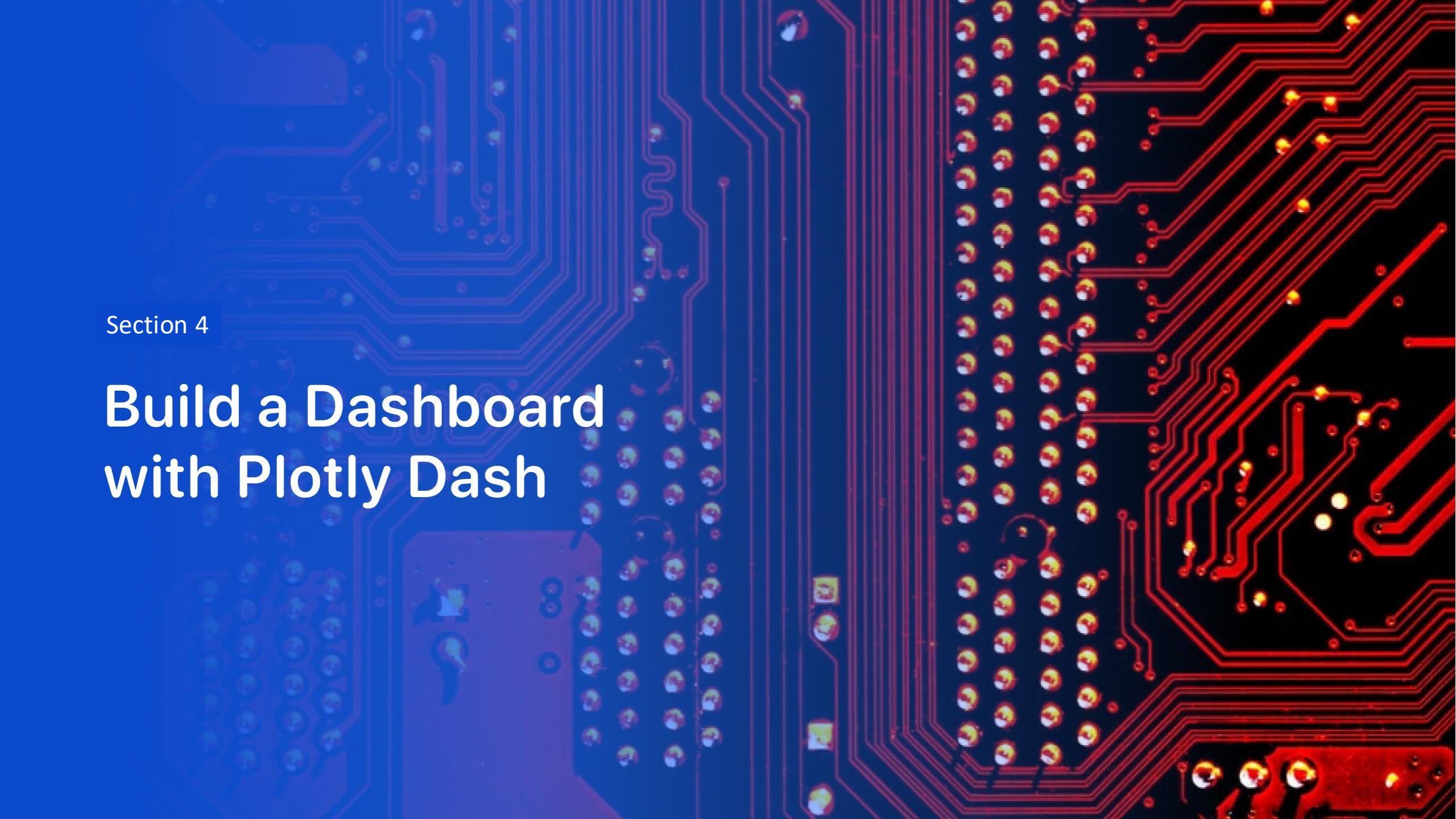


- Distance to highway and railroad <1KM



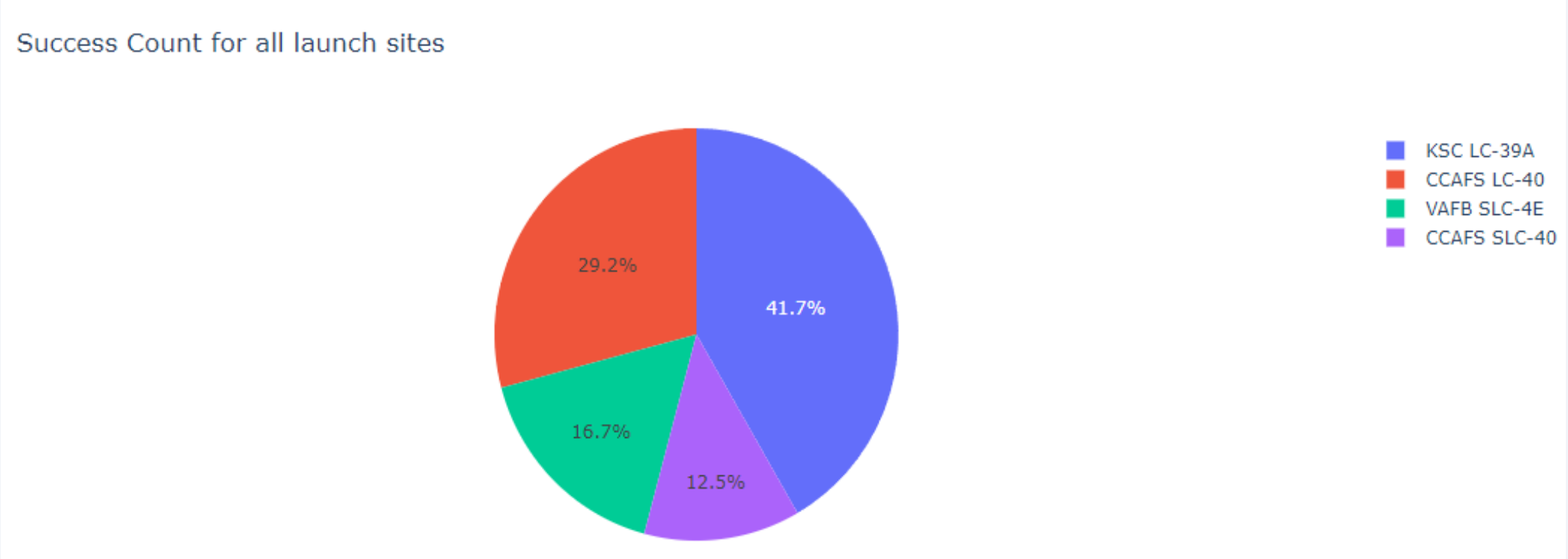
Distance to nearest

This site was further from the coast but extremly close to a railroad and highway.


Section 4
Build a Dashboard with Plotly Dash

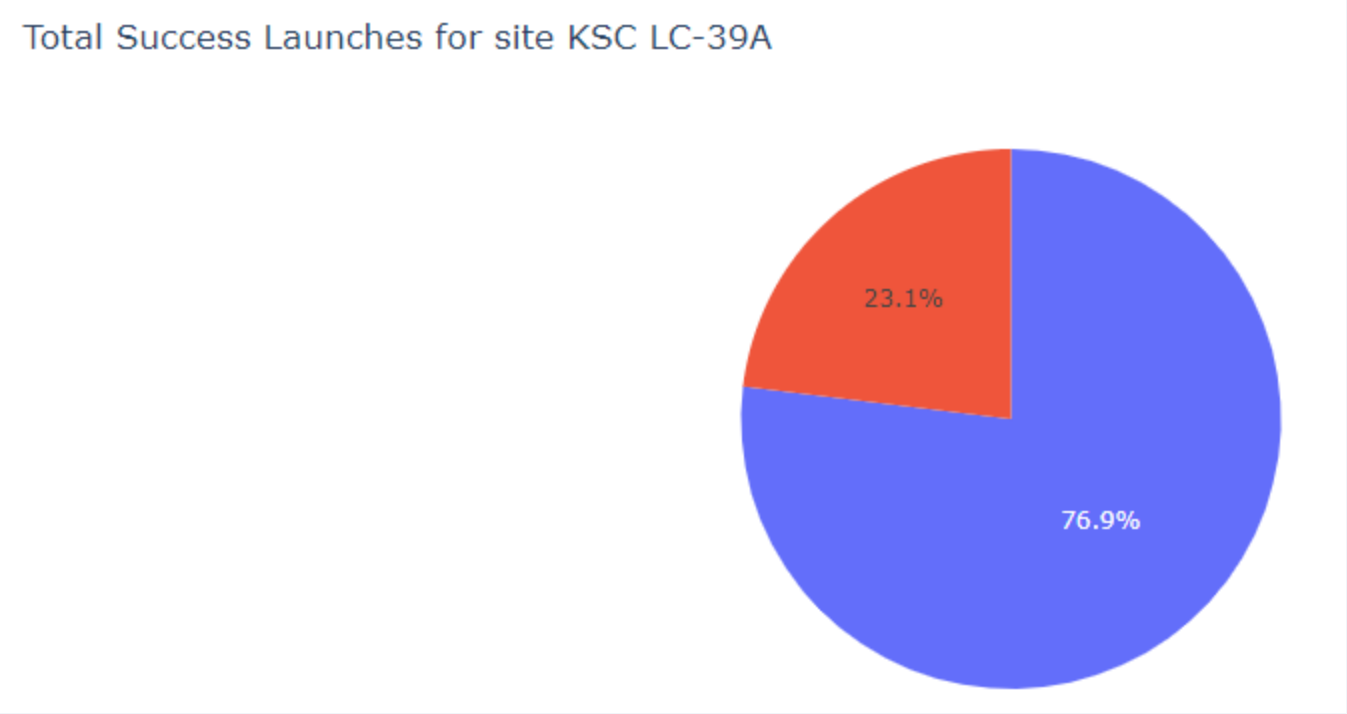
Pie chart for success % of all sites

- KSC LC-39A was the most successful site



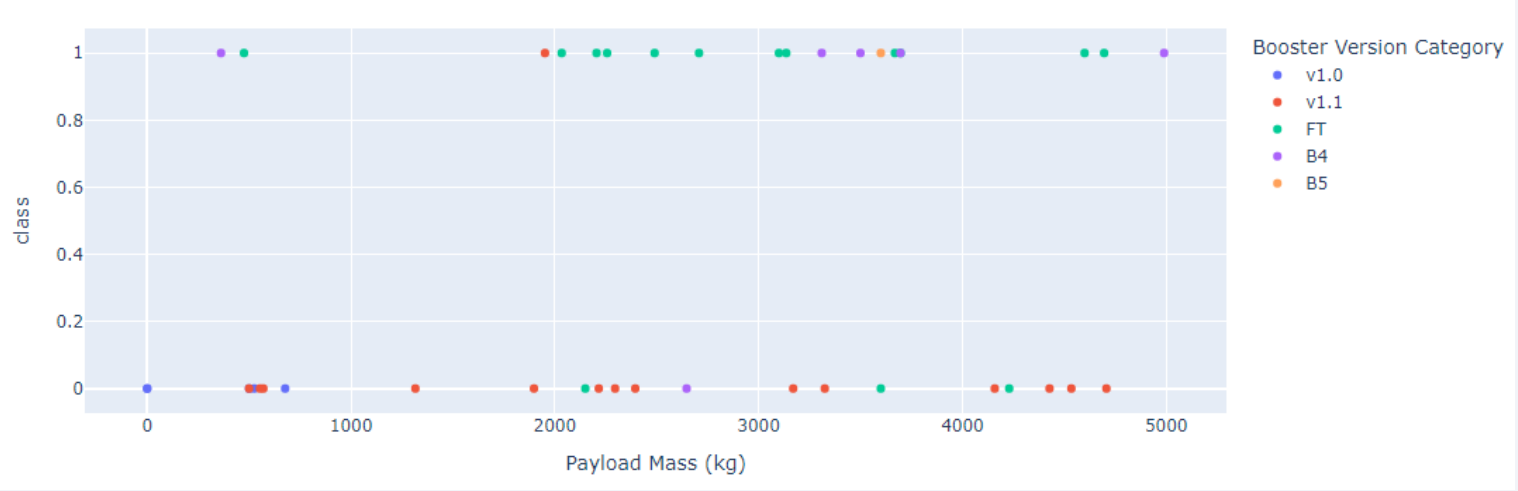
Pie chart ratio for most succesful site

- 76.9% launches were successful at KSC L-39A

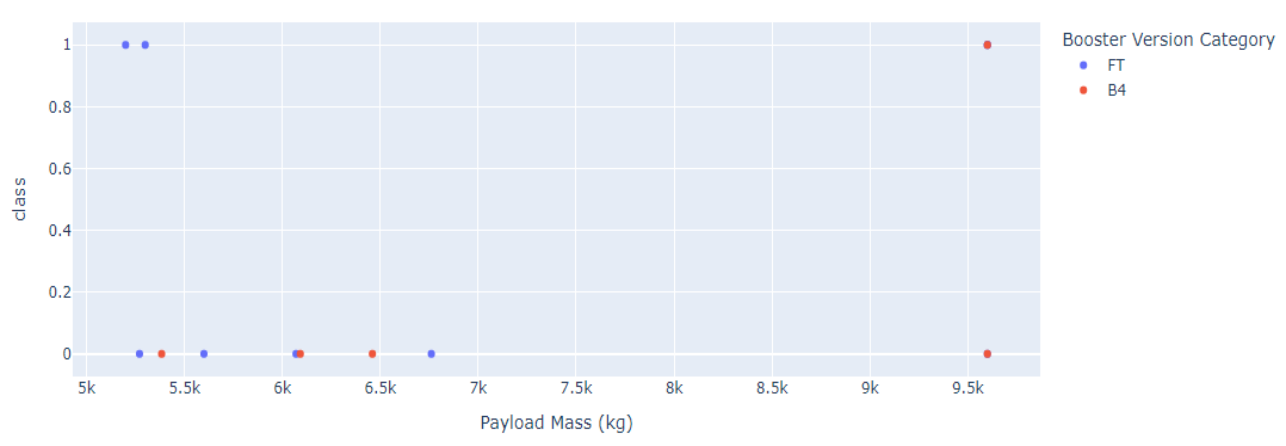


Payload vs. Launch Outcome scatter plot for all sites

- Payload 0-5000. FT had high success rate and v1.1 had a low success rate



- Payload 5000-10000. FT had higher success rate than B4 with higher payloads



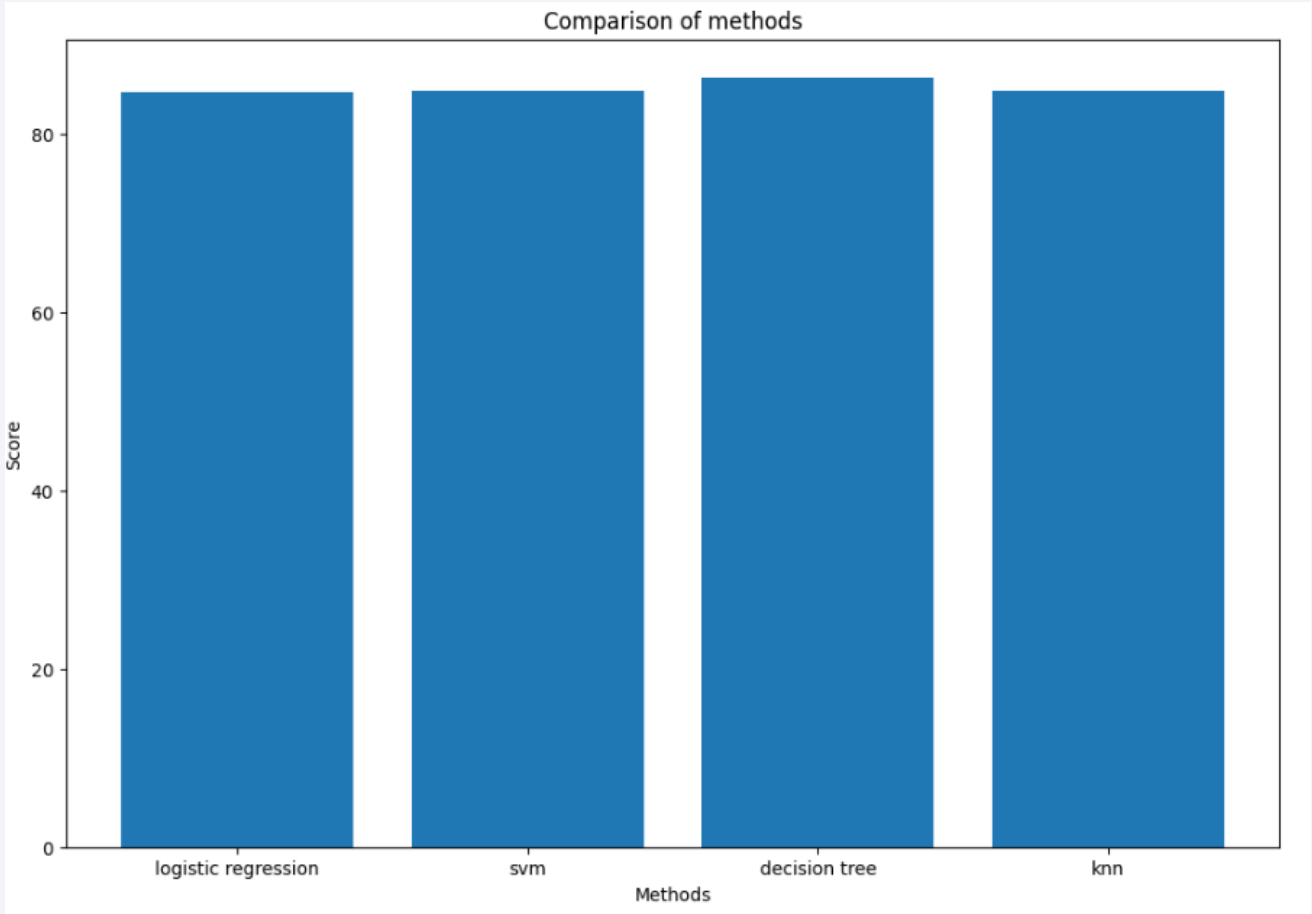
Section 5

Predictive Analysis (Classification)

Classification Accuracy

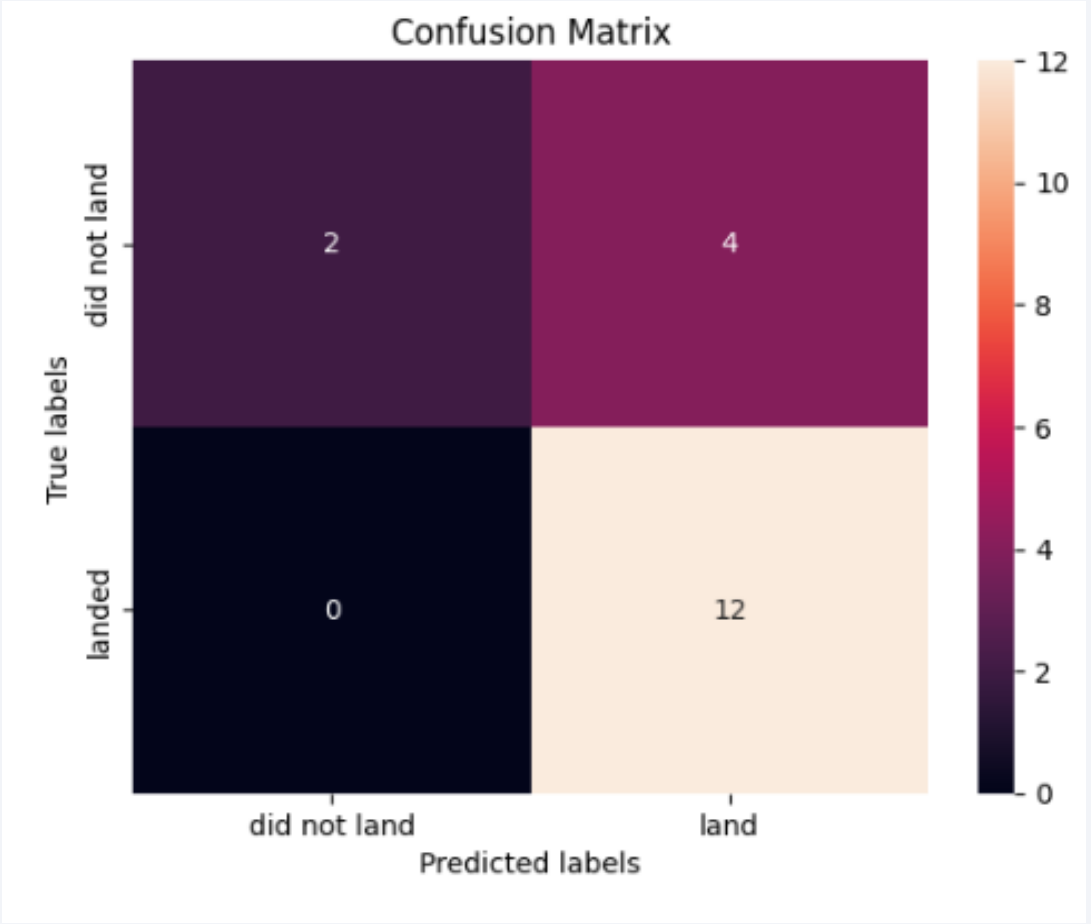
- Decision Trees had slightly greater accuracy than the other models

Method	Scores
decision tree	86.250000
knn	84.821429
svm	84.821429
logistic regression	84.642857



Confusion Matrix

- The main issue is false positives, the decision tree classifier classified 4 unsuccessful landings as landed.



Conclusions

- Launch sites were more successful as the number of flights increased.
- KSC LC-39A was the most successful launch site.
- These orbits had 100% success rate: ES-L1, GEO, HEO and SSO.
- The higher the payload mass the greater the success rate.
- The decision tree model slightly outperformed the other models on the test data set

Thank you!