IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Snehit Dua
11-09-22

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - Exploratory Data Analysis with Data Visualization
  - Exploratory Data Analysis with SQL
  - Build an Interactive Map with Folium
  - Build a Dashboard with Plotly Dash
  - Predictive Analysis (Classification)
- Summary of all results
  - Exploratory Data Analysis Results
  - Interactive Analytics Demo in Screenshots
  - Predictive Analysis Results

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. We will predict if the Falcon 9 first stage will land successfully.
- Problems to be answered:
  - What variables/features affect the successful landing of the first stage?
  - What is the best classification model for prediction of landing of first stage?
  - Does the rate of success increase over the years?
  - Does the current SpaceX's launch sites have the ideal geographical locations?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data was extracted from:
    - SpaceX API and Wikipedia (using web scrapping)
- Perform data wrangling
  - The landing outcomes were classified as successful (1) and unsuccessful (0).
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - All the four Classification Models were trained on 80% of the data using the best parameters using GridSearchCV and the model with the best accuracy was selected.

Data Collection

- The final dataset is comprised of data tables from SpaceX API (an API publicly available with records of the launches of SpaceX) and SpaceX's Wikipedia page.
- Following are the data columns derived from the above two sources:
  - SpaceX API:

    'FlightNumber', 'Date', 'BoosterVersion', 'PayloadMass', 'Orbit', 'LaunchSite', 'Outcome', 'Flights', 'GridFins', 'Reused', 'Legs', 'LandingPad', 'Block', 'ReusedCount', 'Serial', 'Longitude', 'Latitude'

  - Wikipedia:

    'Flight No.', 'Launch site', 'Payload mass', 'Orbit', 'Customer', 'Payload', 'Launch outcome', 'Version Booster', 'Booster landing', 'Date', 'Time'

Data Collection – SpaceX API

Notebook

- Request made to the API ("https://api.spacexdata.com/v4/launches/past").
- Response decoded as JSON to a pandas DataFrame.
- Columns dropped to get resultant DataFrame with columns:
  - 'rocket', 'payloads', 'launchpad', 'cores', 'flight_number', 'date_utc'.
- Rows dropped with multiple cores and payloads (rockets with 2 extra boosters).
- Several requests made to get details needed from the ids in the DataFrame.
- Data filtered to include only Falcon 9 launches.
- Replaced missing values of Payload Mass with mean value.
- Resultant DataFrame:

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
0	1	2010-06-04	Falcon 9	6123.547647	LEO	CCSFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.561857
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCSFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.561857
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCSFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.561857

Data Collection - Scraping

Notebook

- Request made to Wikipedia's SpaceX page using HTTP GET method.
- Response converted to BeautifulSoup object.
- Required table from the html content saved into dictionary with Column Names as keys and its corresponding data as values.
- Dictionary converted to pandas DataFrame.
- Exported in CSV format.

Data Wrangling

Notebook

- Landing Outcomes:
  - “True Ocean”- Successfully landed to a specific region of ocean.
  - “False Ocean”- Unsuccessfully landed to a specific region of ocean.
  - “True RTLS”- Successfully landed to ground pad.
  - “False RTLS”- Unsuccessfully landed to ground pad
  - “True ASDS”- Successfully landed to drone ship.
  - “False ASDS”- Unsuccessfully landed to drop ship.
  - “None ASDS”- Failure to land.
  - “None None”- Failure to land.

  mapped into Classes: Successful (1) and Unsuccessful (0).
- “True Ocean”, “True RTLS” and “True ASDS” considered Successful Outcomes.

EDA with Data Visualization

Notebook

- Charts Plotted:
  - **Scatter Chart:** Flight Number Vs Payload Mass, Flight Number Vs Launch Site, Payload Mass vs Launch Site, Flight Number Vs Orbit Type and Payload Mass Vs Orbit Type.
  - **Bar Chart:** Orbit Type Vs Success Rate.
  - **Line Chart:** Success Rate Yearly Trend.
- **Scatter Chart** can be used to determine the relationships b/w variables which can be further used in Machine Learning Model.
- **Bar Chart** show comparison among discrete categories.
- **Line Chart** show trends in data over time.

EDA with SQL

Notebook

- SQL Queries performed to:
  - Display the names of the unique launch sites in the space mission.
  - Display 5 records where launch sites begin with the string '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 and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster versions which have carried the maximum payload mass.
  - 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.
  - Rank the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

Build an Interactive Map with Folium

Notebook

- All Launch Sites marked on an interactive map using circles to show proximity to equator and coast.
- Marked successful/failed launches for each site on map using marker cluster, where:
  - Green marker -> success
  - Red marker -> fail
- Marked distances using polylines between a Launch Site and its proximities (Coast Line, Highway and City) to answer following questions:
  - Are launch sites in close proximity to railways?
  - Are launch sites in close proximity to highways?
  - Are launch sites in close proximity to coastline?
  - Do launch sites keep certain distance away from cities?

Build a Dashboard with Plotly Dash

Code File

- Dropdown to select a Launch Site (default: All Sites).
- Pie Chart to show success launches of each sites (All Sites) or success vs fail launches (Specific Site).
- Slider to select Payload Mass (kg) range.
- Scatter Chart to show Payload Mass Vs Success Rate for different Booster Versions for the selected Launch Site within the selected Payload Mass range.

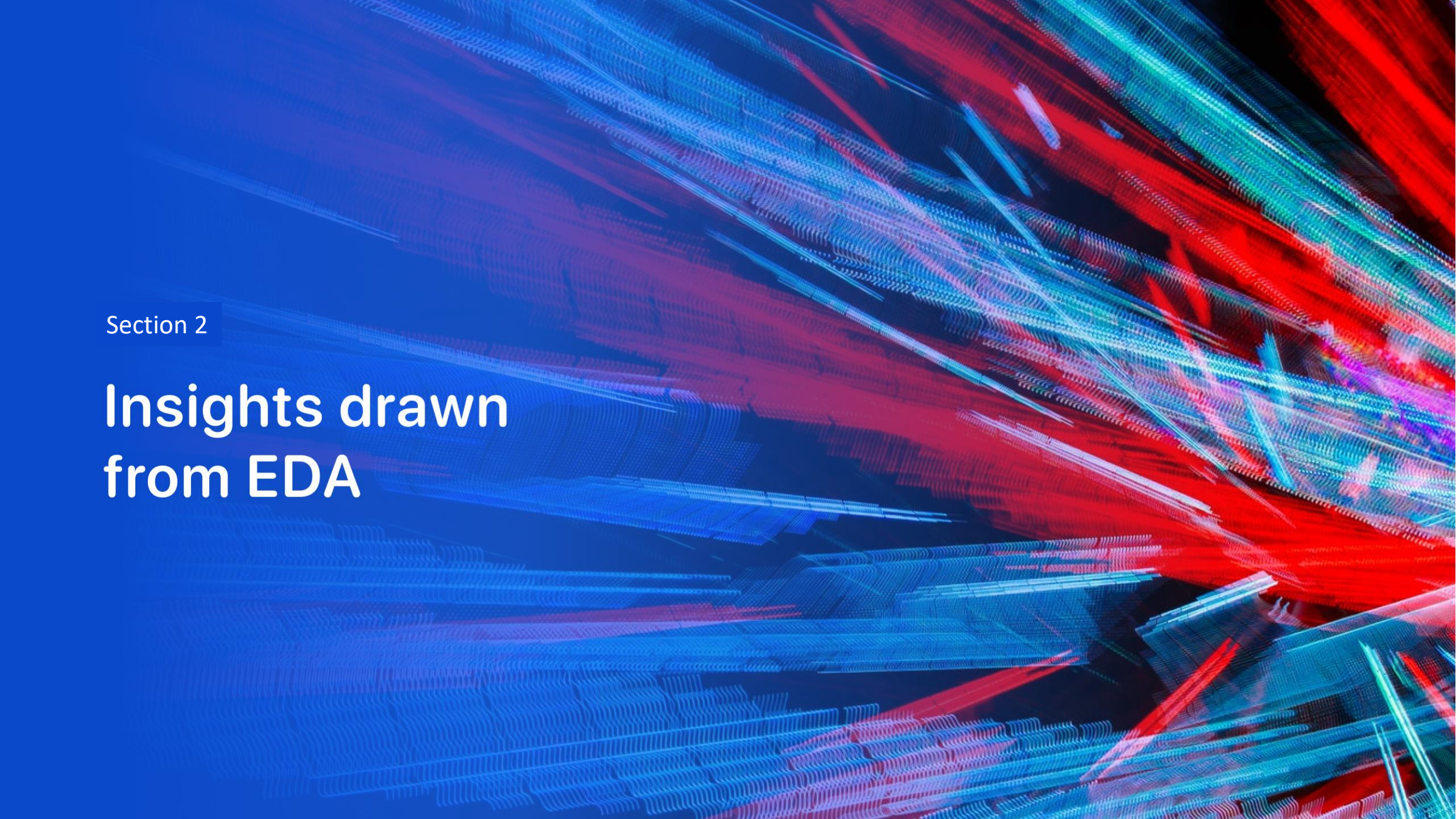
Predictive Analysis (Classification)

Notebook

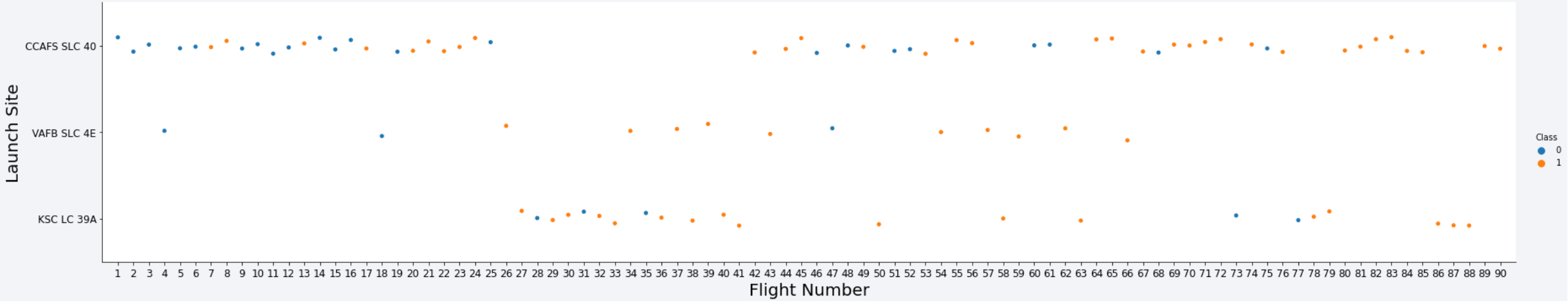
- Independent Variables (X) and Dependent Variables (Y) separated (Y being 'Class').
- Standardized X using Standard Scaler.
- Data divided into 2 parts: Train and Test Data using Train Test Split (test size being 20%).
- GridSearchCV object created with cv = 10 to find best parameters for all the Classification Models.
- Accuracies (.score, Jaccard Score and F1 Score) on test and whole data separately calculated and compared for all the models.
- Model with the best Accuracy chosen.

Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn from EDA

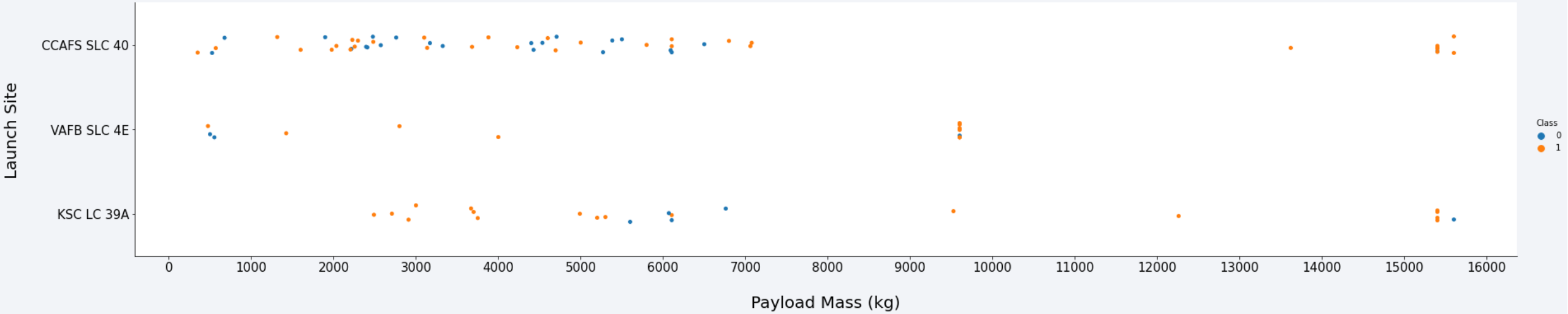
Flight Number vs. Launch Site



Explanation:

- CCAFS SLC 40 is having the majority of launches while VAFB SLC 4E is having the least no. of launches.
- KSC LC 39A is having the highest success rate followed by VAFB SLC 4E.
- Newly launched flights have higher success rate as compared to the earliest ones.

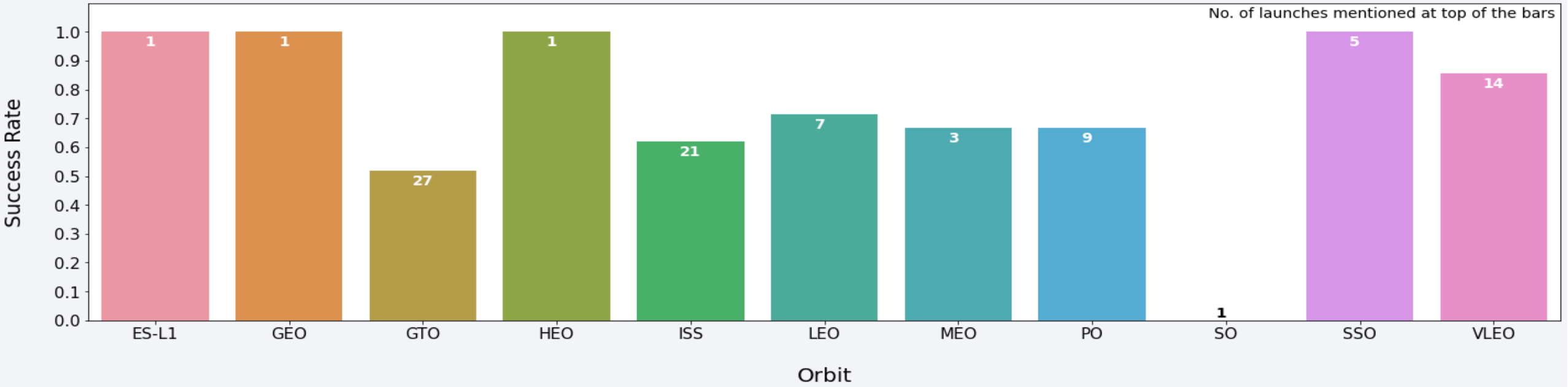
Payload vs. Launch Site



Explanation:

- Higher the Payload Mass higher the success rate.
- Almost every launch with Payload Mass greater than 7000 is successful.
- KSC LC 39A is the only site with 100% success rate for launches with Payload Mass less than 5500.

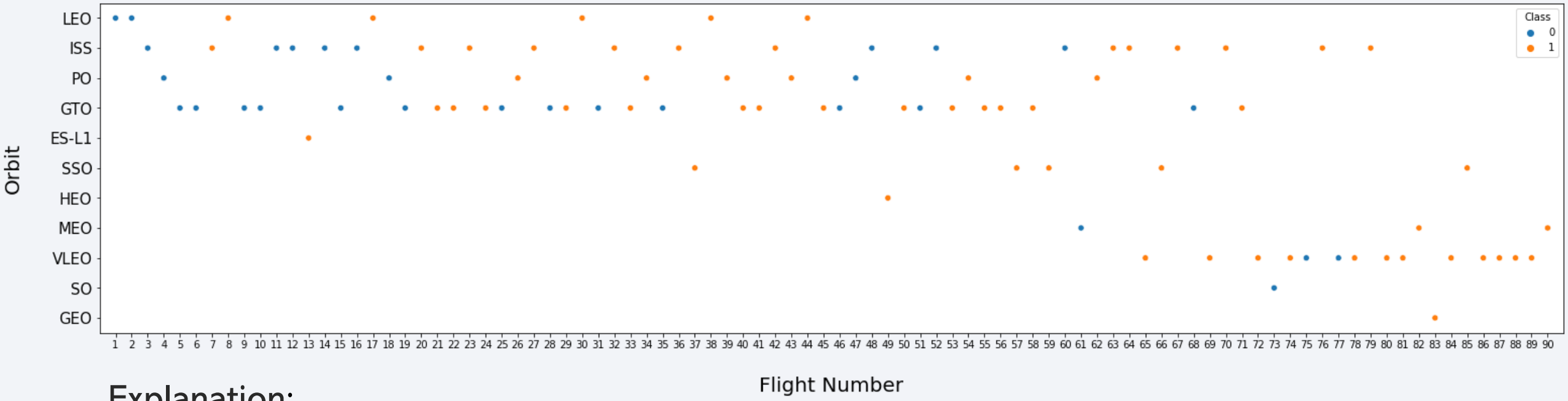
Success Rate vs. Orbit Type



Explanation:

- Orbits with 100 % success rate: ES-L1, GEO, HEO and SSO.
- Only 1 flight is launched per above mentioned orbit types except SSO.
- SO is the only orbit type with 0% success rate but only 1 flight is launched with it.
- GTO (50%) and ISS (60%) are only orbit types with highest launches.

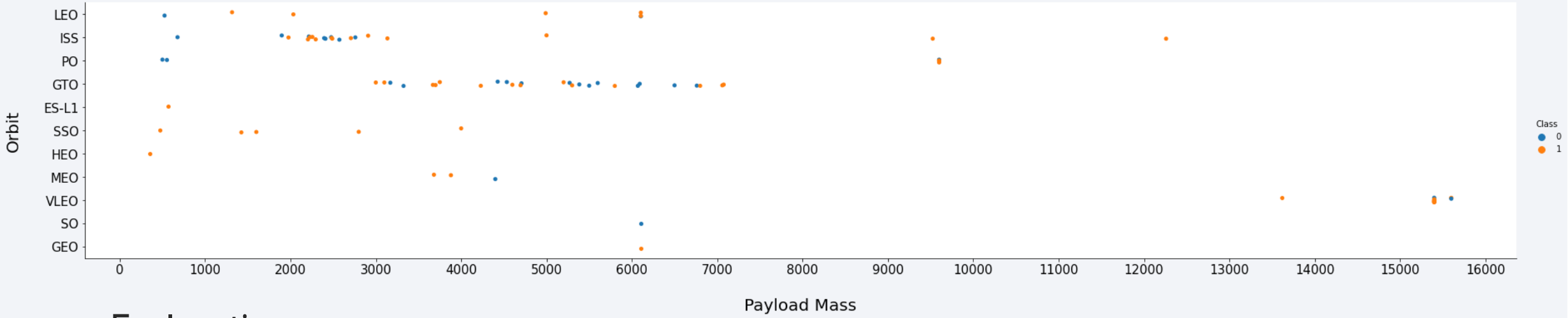
Flight Number vs. Orbit Type



Explanation:

- Majority of new launches (60-90) are with orbit type VLEO with 85% success rate.
- Majority of old launches (1-60) are with orbit type GTO.
- In every orbit as the no. of flights increases; success rate also increases except in GTO.

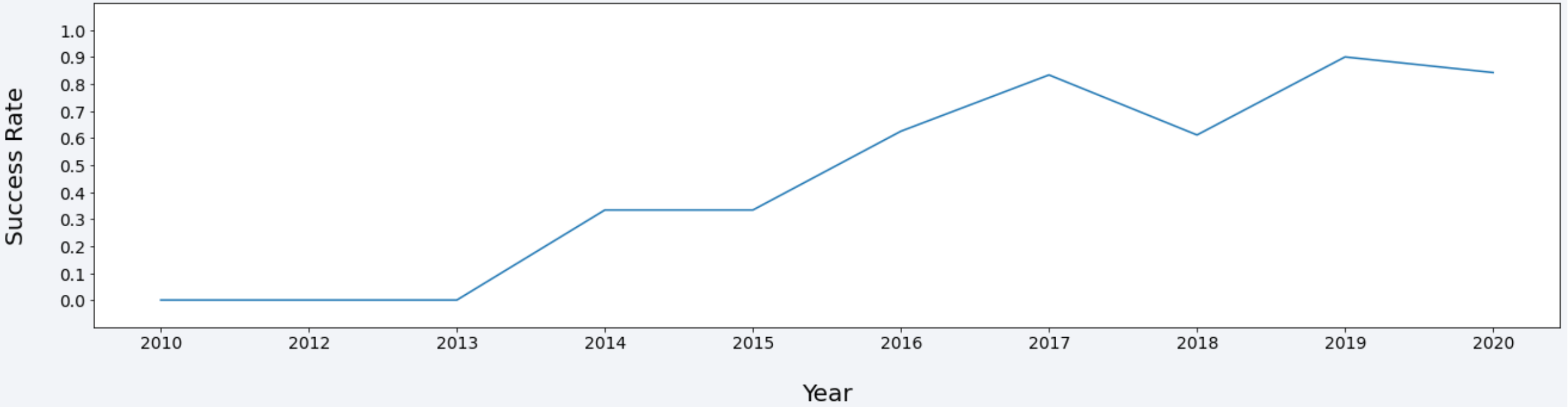
Payload vs. Orbit Type



Explanation:

- LEO, PO and ISS have higher success rates with higher Payload Masses.
- But GTO is not related with Payload Mass as it is having both success and fail landings.
- SSO is having 100 % success rate in between low payload range of 0-4000.
- VLEO appears to work better with higher payload ranges with 85% success.
- We can't say anything about other orbits as there is insufficient evidence.

Launch Success Yearly Trend



Explanation:

- The Success Rate in the initial years (2010-2013) remained same i.e. 0%.
- The Success Rate kept increasing from 2013 till 2020 non uniformly.
- Highest Success Rate in 2019 i.e. 90%.

All Launch Site Names

Query:

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

Output:

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

Explanation:

- Displaying the names of unique Launch Sites in the space Mission.

Launch Site Names Begin with 'CCA'

Query:

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

Output:

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

Explanation:

- Displaying 5 records where Launch Sites begin with ‘CA’.

Total Payload Mass

Query:

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)'
```

Output:

SUM(PAYLOAD_MASS__KG_)
45596

Explanation:

- Displaying the total payload carried by boosters from NASA.

Average Payload Mass by F9 v1.1

Query:

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

Output:

AVG(PAYLOAD_MASS__KG_)
2928.4

Explanation:

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

First Successful Ground Landing Date

Query:

```
%%sql SELECT MIN(DATE) FROM SPACEXTBL WHERE "Landing _Outcome" = 'Success (ground pad)'
```

Output:

MIN(DATE)
01-05-2017

Explanation:

- Displaying the dates of the first successful landing outcome on ground pad.

Successful Drone Ship Landing with Payload between 4000 and 6000

Query:

```
%%sql
SELECT BOOSTER_VERSION, PAYLOAD_MASS__KG_ FROM (SELECT * FROM SPACEXTBL WHERE "Landing _Outcome" = 'Success (drone ship)')
WHERE PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000
```

Output:

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

Explanation:

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

Total Number of Successful and Failure Mission Outcomes

Query:

```
%sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS COUNT FROM SPACEXTBL GROUP BY MISSION_OUTCOME
```

Output:

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

Explanation:

- Displaying the total number of successful and failure mission outcomes.

Boosters Carried Maximum Payload

Query:

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

Output:

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

Explanation:

- Displaying the names of the booster which have carried the maximum payload mass.

2015 Launch Records

Query:

```
%%sql
SELECT SUBSTR(DATE,4,2) AS MONTH, "Landing _Outcome", Booster_Version,  Launch_Site FROM
(SELECT * FROM SPACEXTBL WHERE "Landing _Outcome" = 'Failure (drone ship)' AND DATE LIKE '%2015')
```

Output:

MONTH	Landing _Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Explanation:

- Displaying the records which will display the months, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015.

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

Query:

```
%%sql
SELECT * FROM (SELECT "Landing _Outcome", COUNT("Landing _Outcome") AS COUNT FROM SPACEXTBL WHERE
(substr(DATE, 7, 4) || '-' || substr(DATE, 4, 2) || '-' || substr(DATE, 1, 2)) BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY "Landing _Outcome")
WHERE "Landing _Outcome" LIKE '%Success%' ORDER BY COUNT DESC
```

Output:

Landing _Outcome	COUNT
Success (drone ship)	5
Success (ground pad)	3

Explanation:

- Displaying the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

Section 3

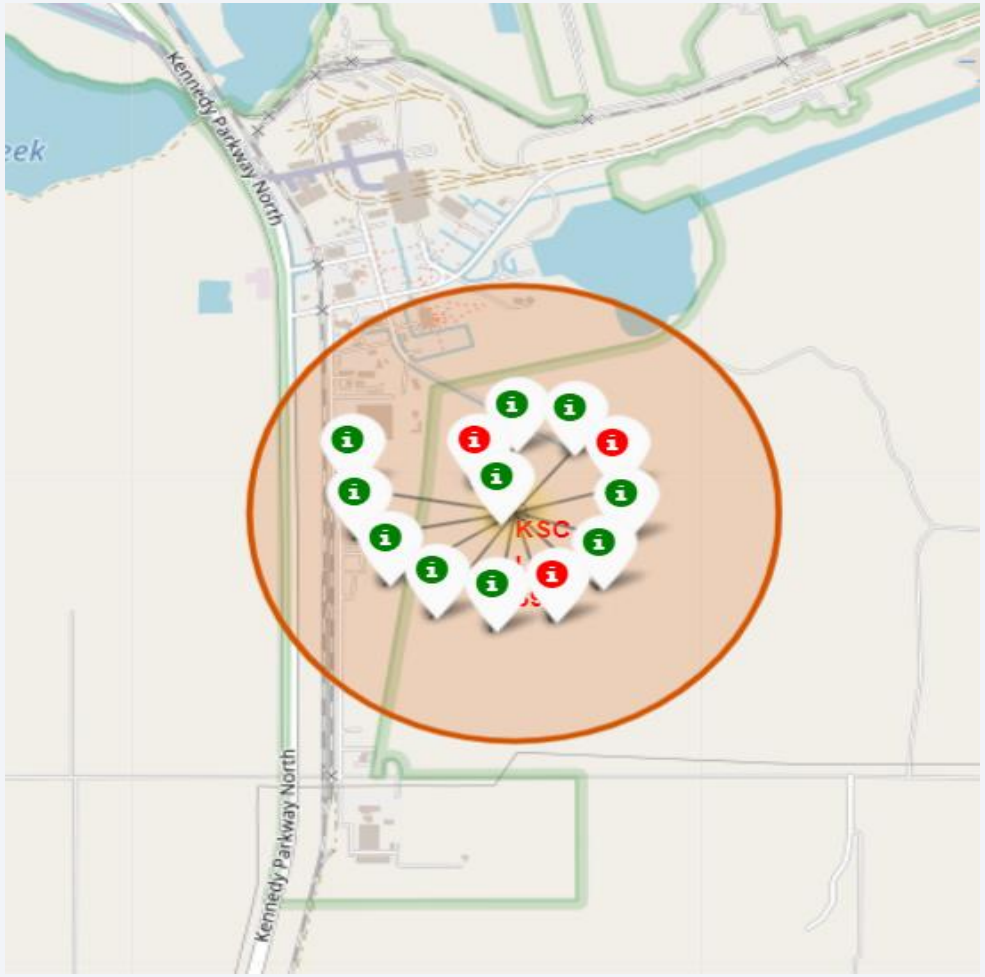
Launch Sites Proximities Analysis

Launch Sites Locations

- All the Launch Sites are in very close proximity to Coast as it reduces the risk of nearby people getting injured by the explosion.
- All the Launch Sites are also in close proximity to Equator line as rockets get an additional natural boost that helps save the cost of putting in extra fuel and boosters.

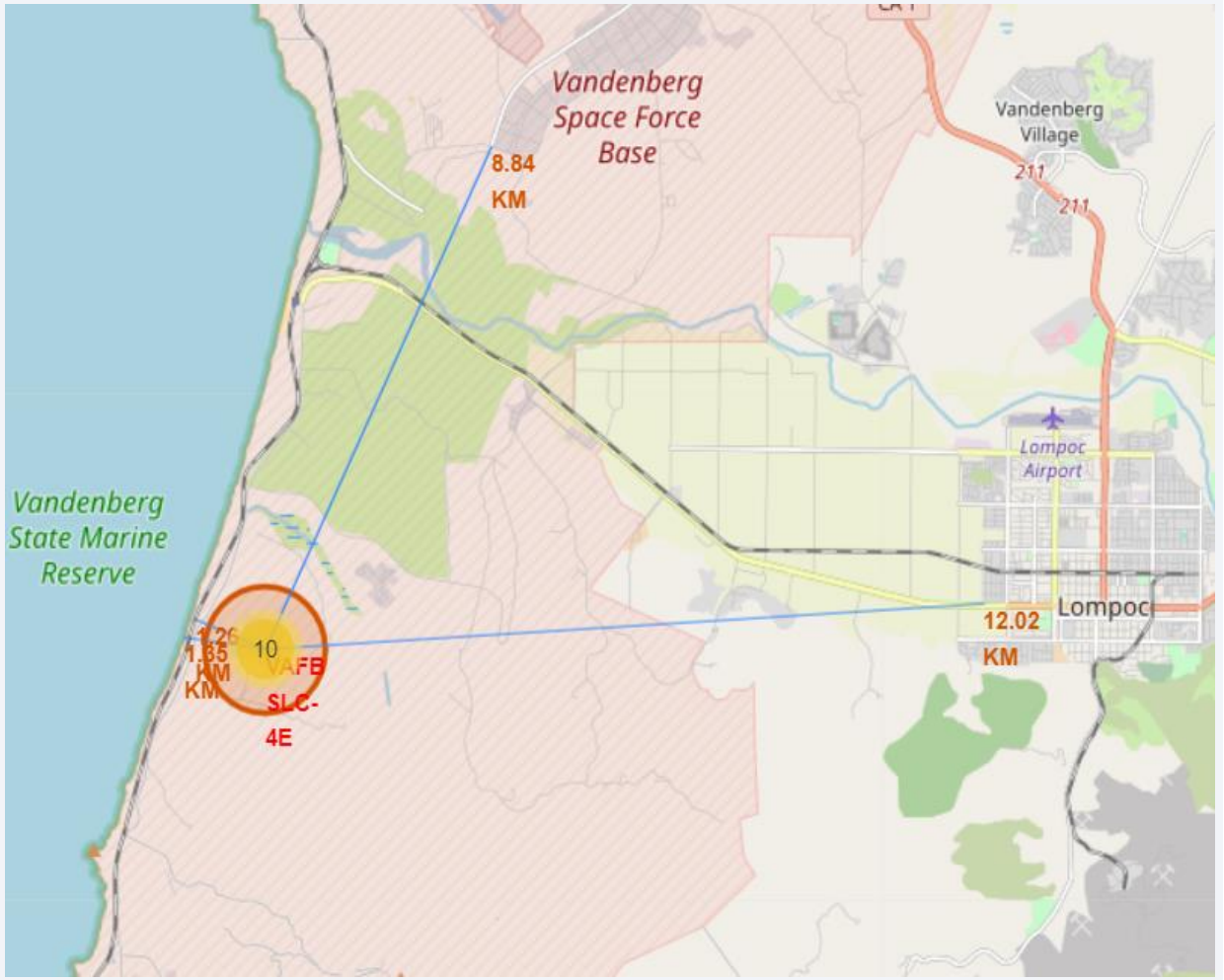
Color Labeled Launch Records

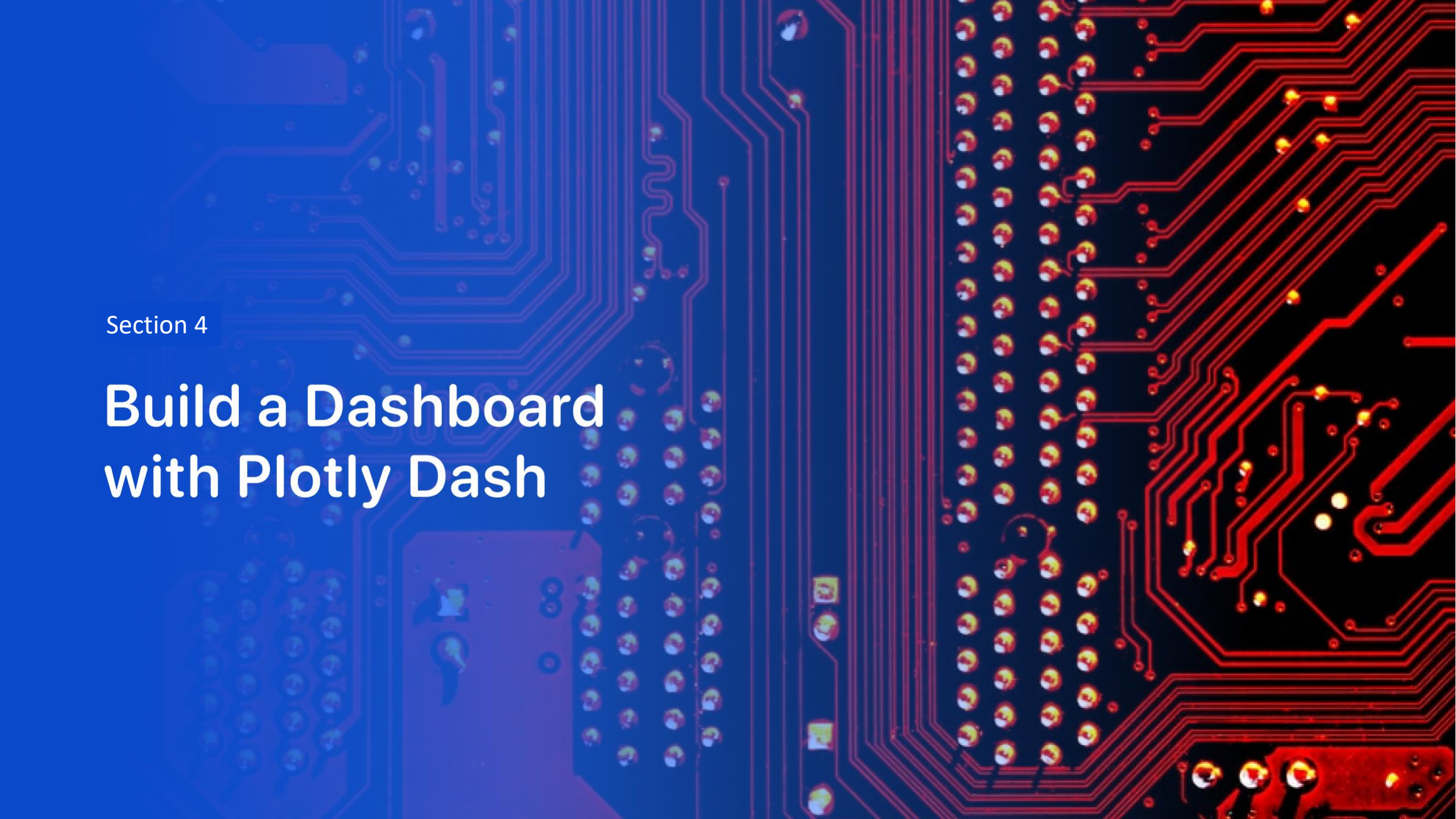
- Color labeled markers shows the Launch Site's launch records classified as:
  - Successful (Green Color)
  - Unsuccessful (Red Color)
- KSC LC-39A is the Launch Site with Highest Success Rate.



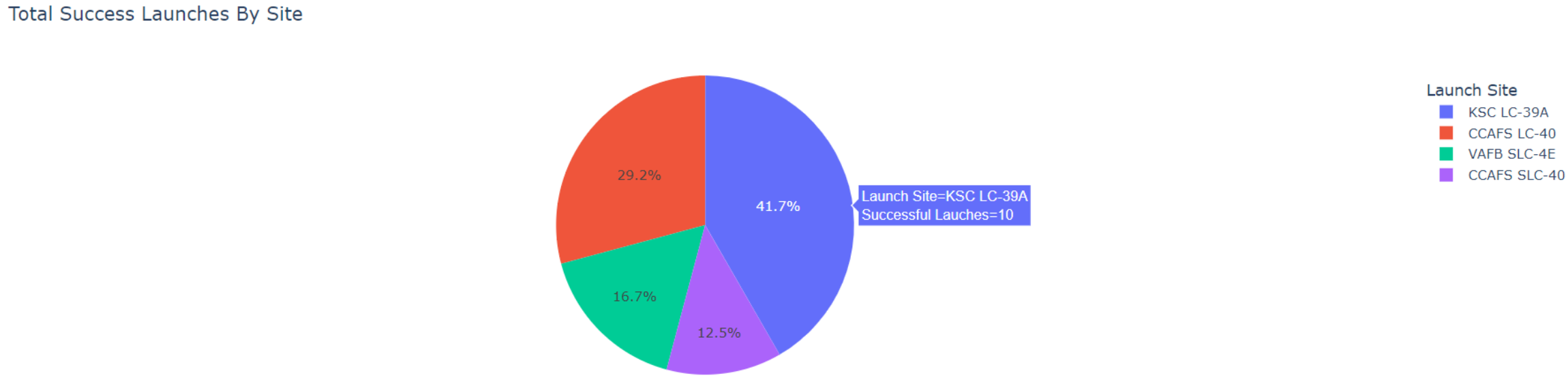
Distance From VAFB SLC-4E to its Proximities

- From the map we can say that:
  - Launch Site is very close to Coast Line (1.35 KM) which is a good thing.
  - It is very close to Railway (1.25 KM) which is a very bad thing.
  - Highway and the closest city is little far i.e. 8.84 KM and 12.02 KM respectively but still failed rockets can cover a distance of 15-20 KM in few seconds.



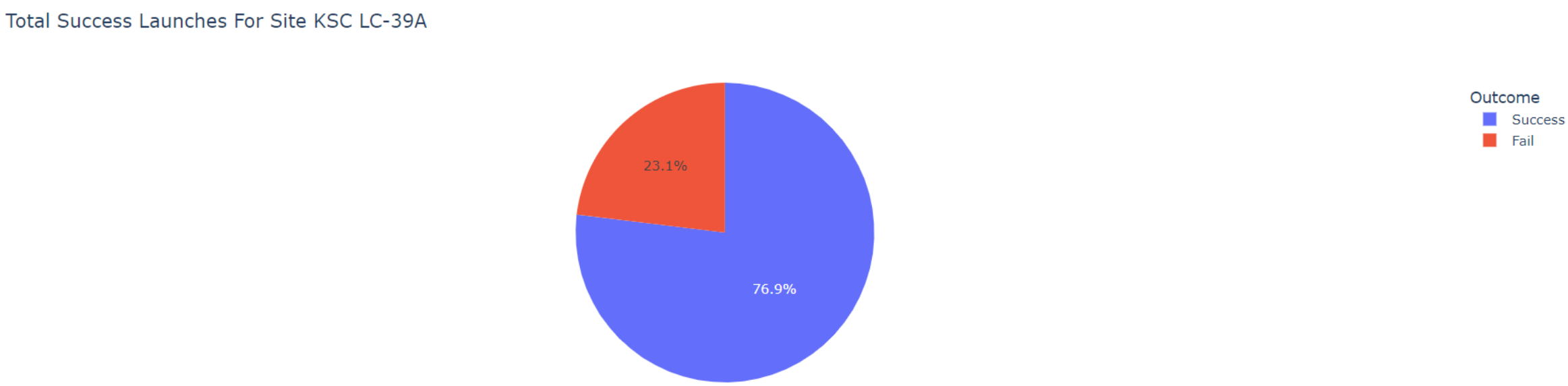

Section 4
Build a Dashboard with Plotly Dash

Total Successful Launches Per Launch Site



- Explanation:
  - KSC LC-39A is having the highest no. of successful launches i.e. 10.
  - CCAFS SLC-40 is having the least no. of successful launches i.e. 3.

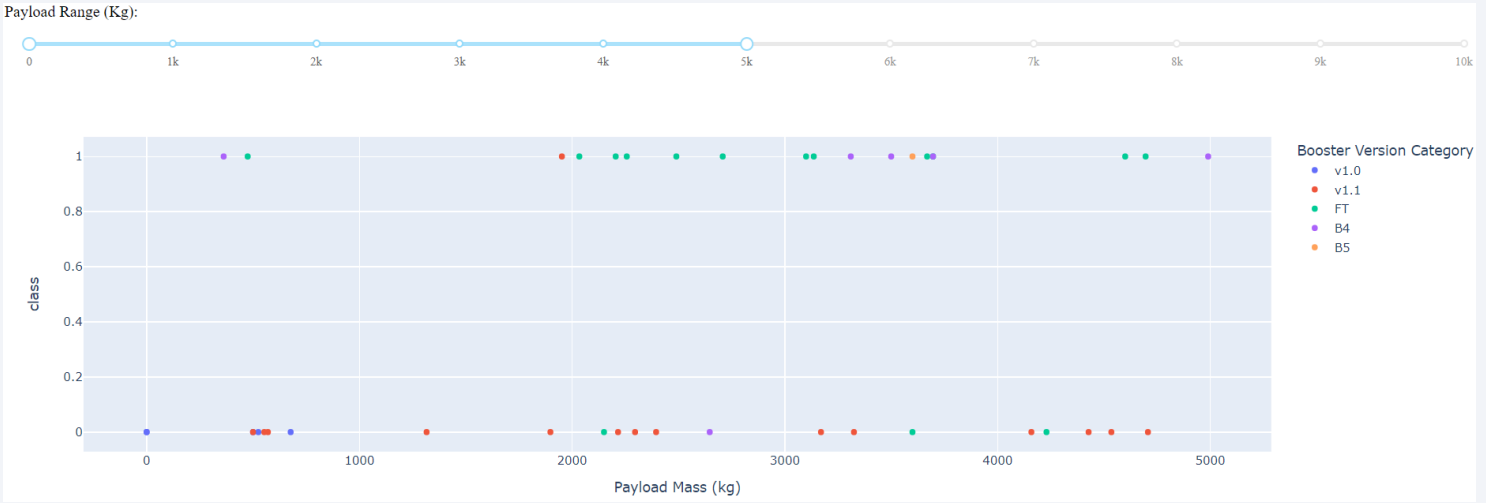
Launch Site with Highest Launch Success Ratio

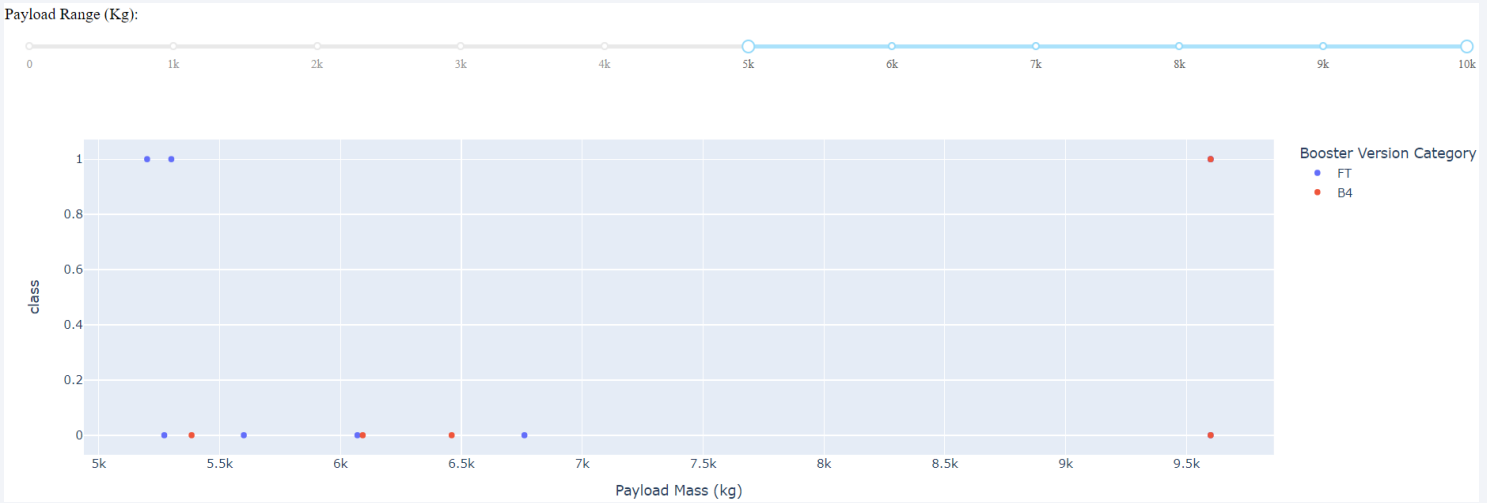


- Explanation:
  - KSC LC-39A is the Launch Site with Highest Launch Success Ratio with 10 successful and 3 unsuccessful landings.

Payload Mass Vs Launch Outcome

- Payload Mass in the range of 2000-5000 have the highest success rate.
- Booster Version B5 have 100% success rate but it is used only in one flight.
- We can say Booster Version v1.0 and v1.1 are the worst performing with 0% and 6% success rates for 4 and 15 flights respectively.
- FT and B4 are the only boosters used with payload above 5000 but these works better in below 5000 with 80% and 83% success rate as compared to 44% and 20% (in higher range).





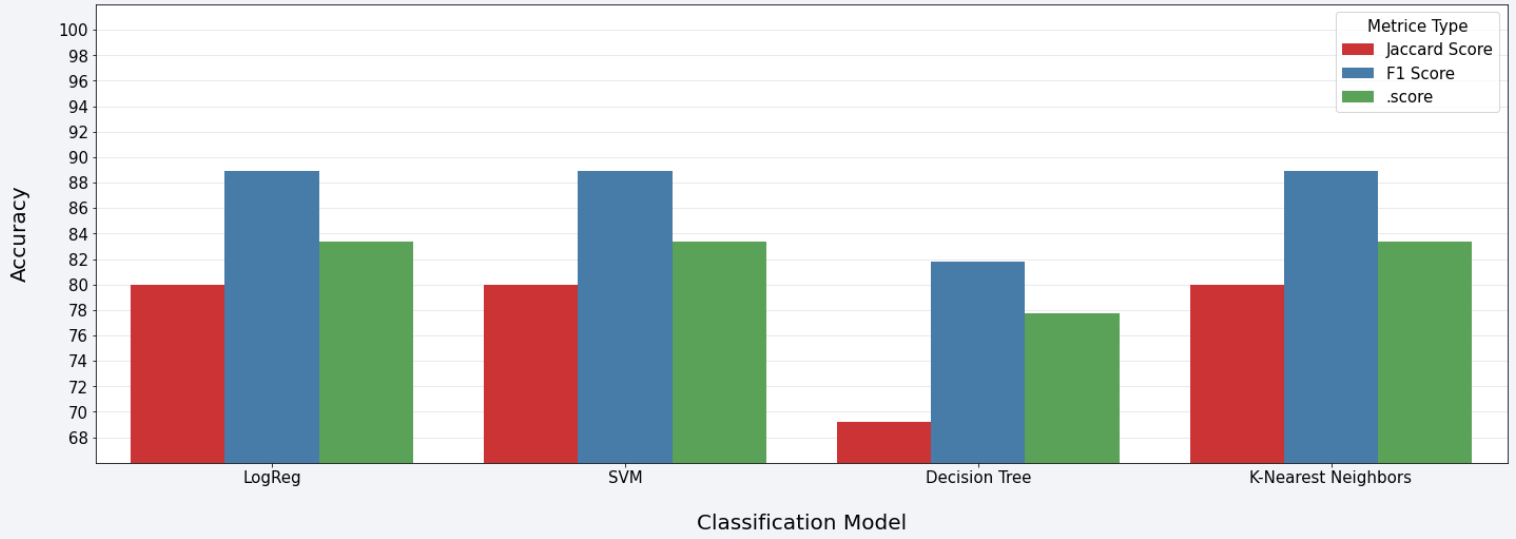
Section 5

Predictive Analysis (Classification)

Classification Accuracy

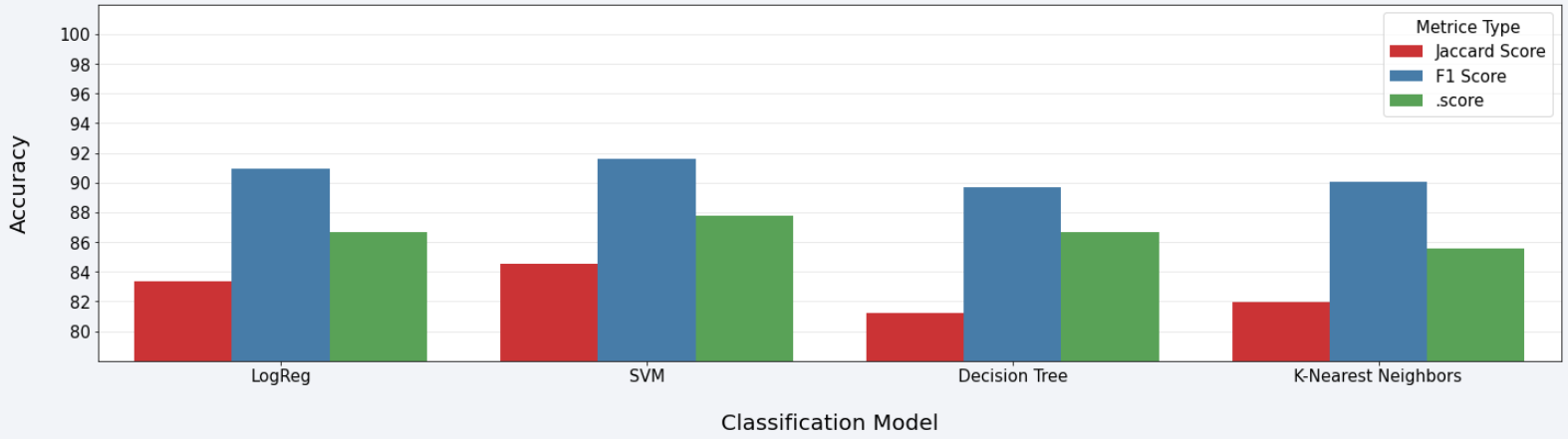
On Test Data:

	LogReg	SVM	Tree	KNN
Jaccard_Score	80.0%	80.0%	69.23%	80.0%
F1_Score	88.89%	88.89%	81.82%	88.89%
Accuracy	83.33%	83.33%	77.78%	83.33%



On Entire Dataset:

	LogReg	SVM	Tree	KNN
Jaccard_Score	83.33%	84.51%	81.25%	81.94%
F1_Score	90.91%	91.6%	89.66%	90.08%
Accuracy	86.67%	87.78%	86.67%	85.56%

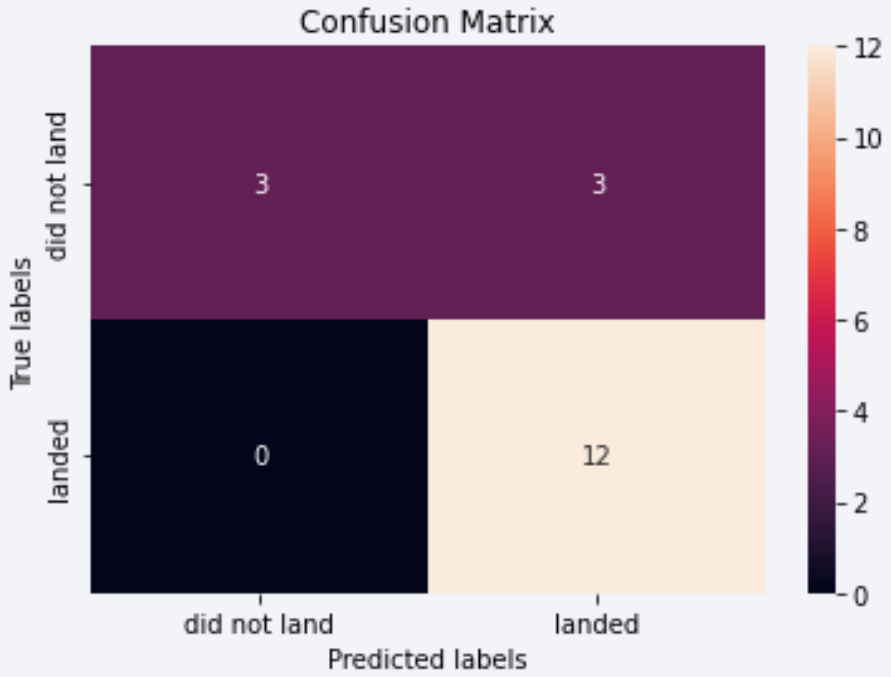


Classification Accuracy

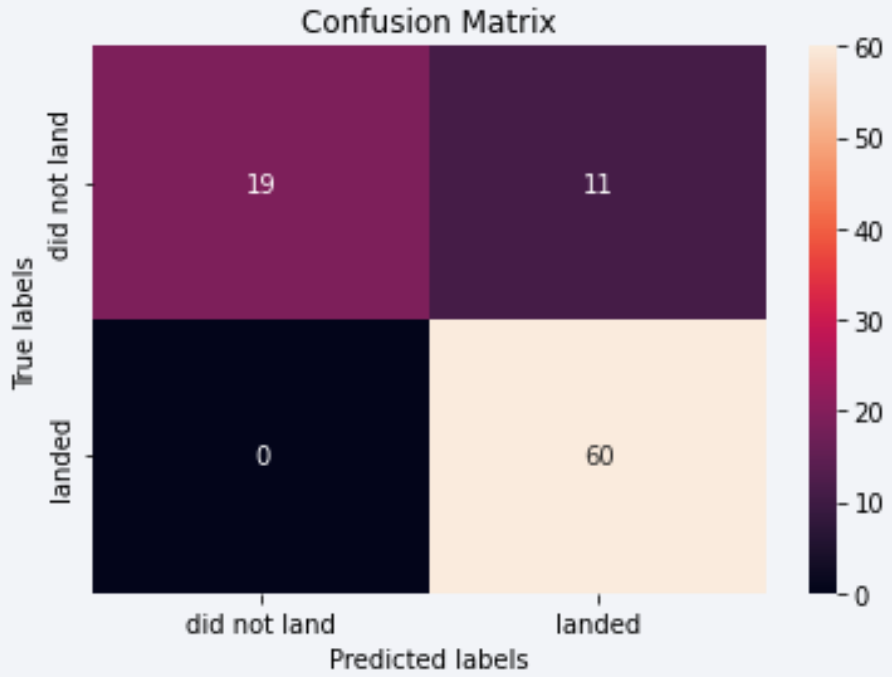
- Explanation:
  - All the Classification Models when tested on the test data (20% of entire dataset) produced the same accuracy except Decision Tree; so, it is not clear which one is the best one.
  - Testing upon the entire dataset, we can say that Support Vector Machine is the best Classification Model as it is the one with highest accuracy tested with all the 3 metrices.

Confusion Matrix

Test Data



Whole Data



- Explanation:
  - On examining we can say that our model is pretty accurate but the main problem is false positives.

Conclusions

- Newly launched flights have higher success rate as compared to the earliest ones.
- Success Rate is low for higher Payload Mass as compared to lower.
- Orbit Types: SSO and VLEO are the best orbits with 100% and 85% success rate with 5 and 15 total launches respectively.
- Orbit Type VLEO works better with higher payload ranges (>13000) with 85% success.
- Success Rate keeps increasing as the years progresses.
- Launch Sites are in close proximity to Coast and Equator Line (pro).
- Launch Sites are in close proximity to Railway, Highway and City (con).
- KSC LC-39A is the Launch Site with Highest Success Rate.
- Payload Mass in the range of 2000-5000 have the highest success rate.
- Worst performing Booster Versions: v1.0 and v1.1.
- Booster Versions: FT and B4 works better with lower payload range (<5000) as compared to higher.
- Support Vector Machine is the best Classification Model in our case.

Appendix

- Created Jupyter Notebooks:
  - API Data Collection
  - Web Scraping
  - Data Wrangling
  - EDA With SQL
  - Data Visualization
  - Launch Site Location Analysis
  - Machine Learning Model

Thank you!