IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Nitin Jain
26 Feb 2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection via API and Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis
  - Interactivity using Dash and Folium
  - Test with test data; and Predictive Analysis
- Summary of all results
  - There are chances of successful landing of first stage of SpaceX
  - Due to low cost, it may be difficult for competitors to bid unless they do some extraordinary advancements.

Introduction

- Project background and context
  - We want to predict if the Falcon 9 first stage will land successfully. 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.
- Problems you want to find answers
  - Whether Competitors can bid against SpaceX or not.
  - How successfully first stage will land?
  - To determine the cost of a launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- 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

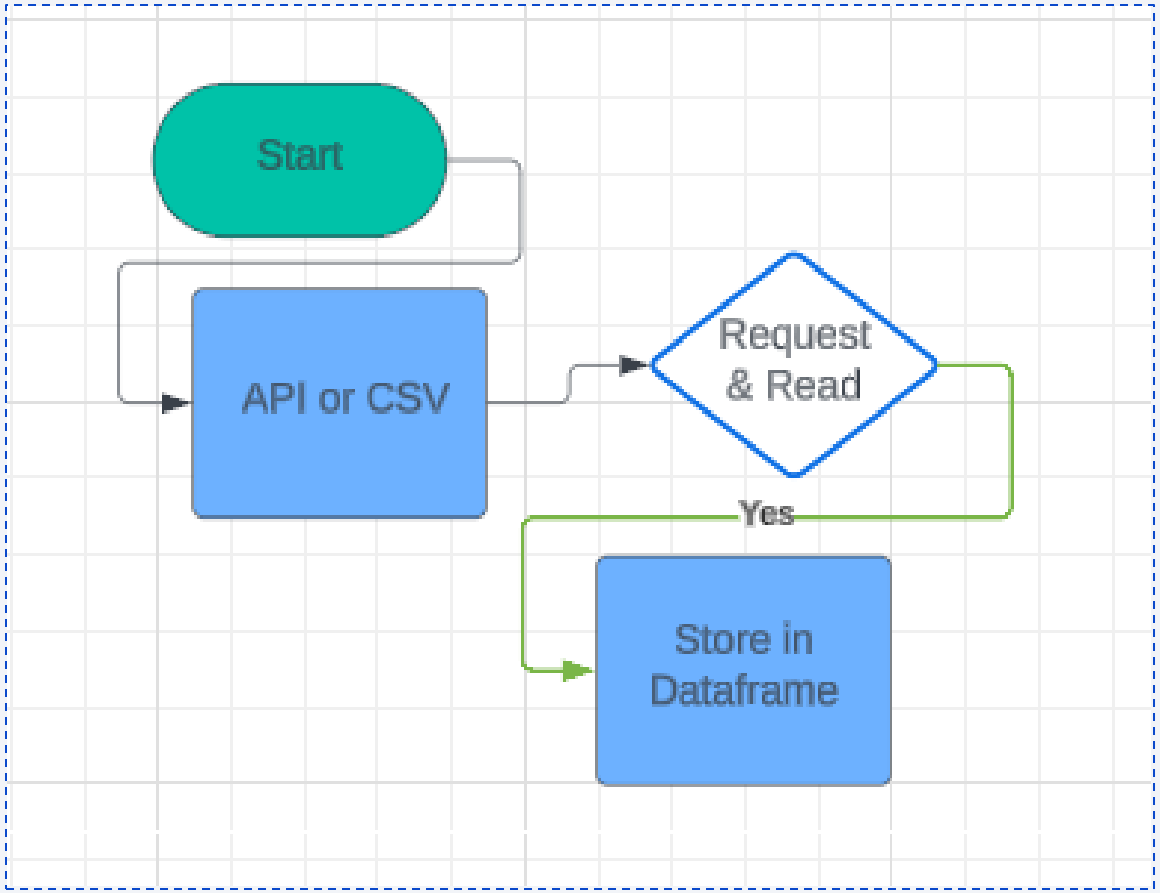
Data Collection

- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts

- CSV/API ----> Dataframe
- Webpage ---> WebScraping--> Dataframe

Data Collection – SpaceX API

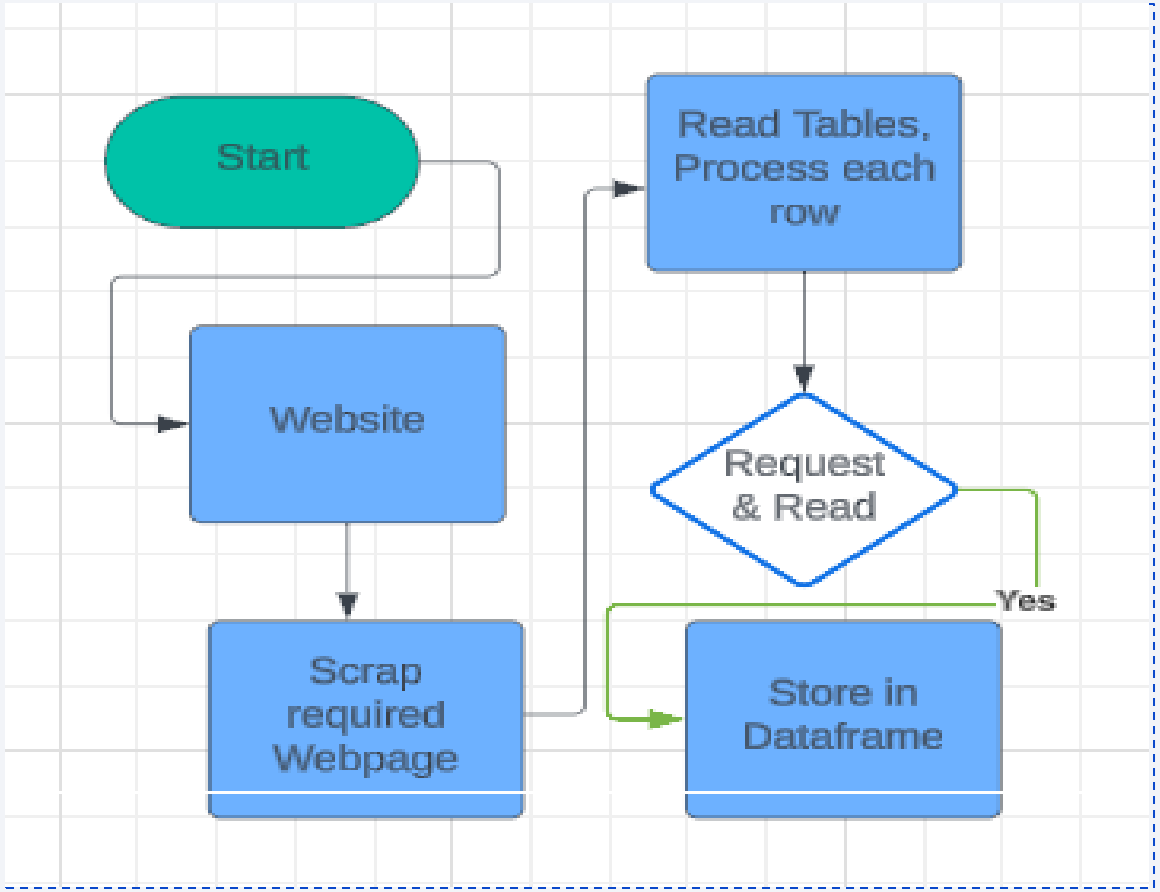
- Data is read either from an API link or a CSV file and stored in the dataframe
- For JSON data, it has to be parsed before storing in dataframe.
- Add the GitHub URL of the completed SpaceX API calls notebook (must include completed code cell and outcome cell), as an external reference and peer-review purpose
- https://github.com/jaindynike/IBMDSCapstoneProject



Data Collection - Scraping

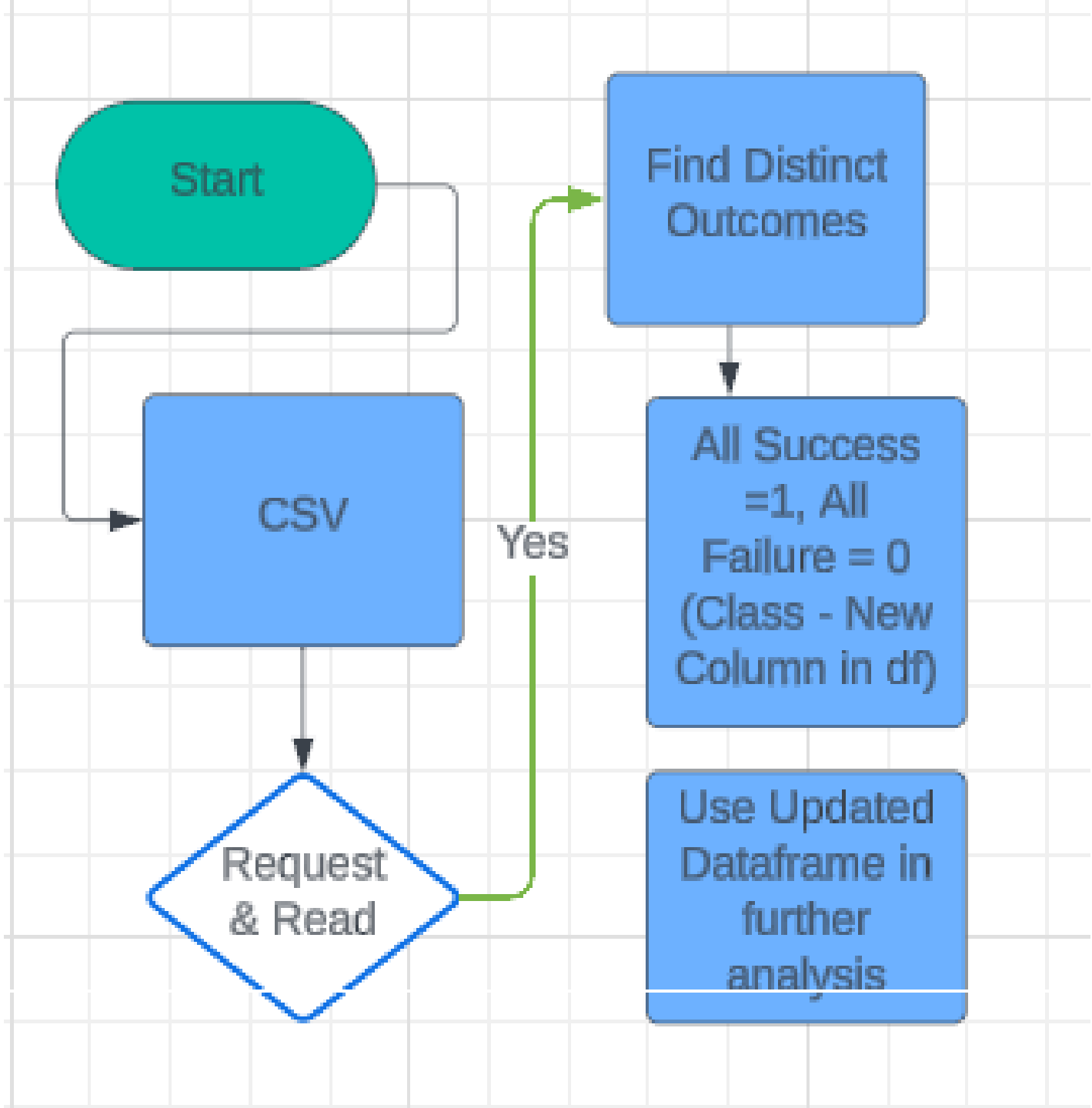
- Present your web scraping process using key phrases and flowcharts

- Add the GitHub URL of the completed web scraping notebook, as an external reference and peer-review purpose



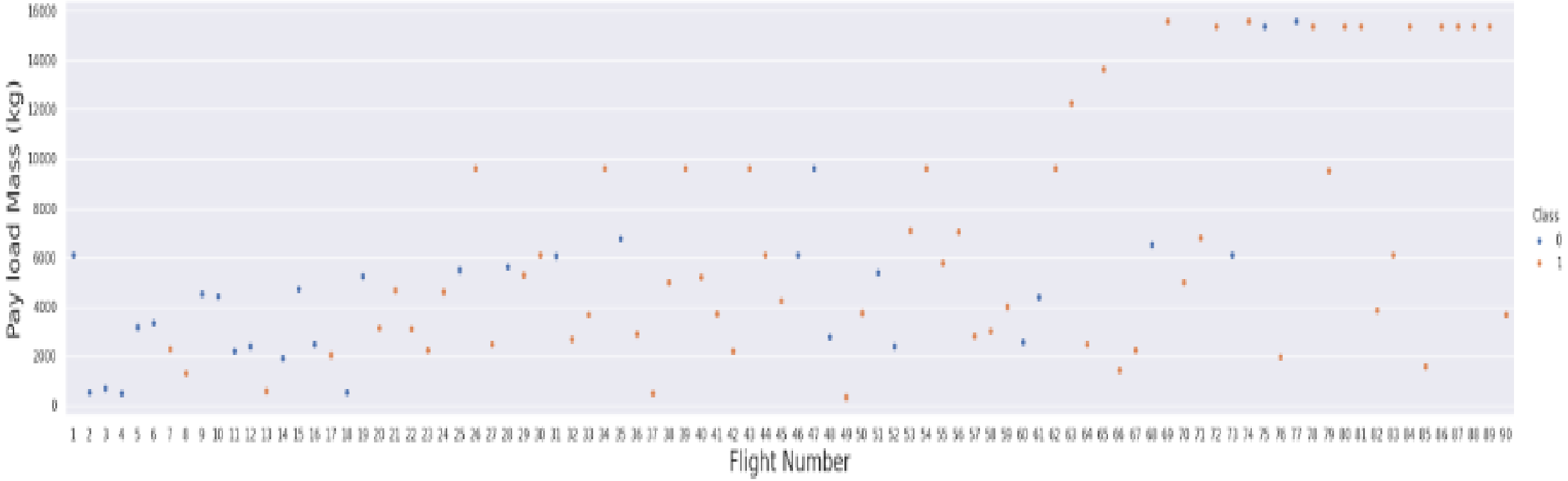
Data Wrangling

- A new column, Class, is created in the dataframe for converting the Outcome columns into simple form of (0 or 1) for success or failure. Currently Outcome column contains:
- `True ASDS` 41
- `None None` 19
- `True RTLS` 14
- `False ASDS` 6
- `True Ocean` 5
- `False Ocean` 2
- `None ASDS` 2
- `False RTLS` 1
- It is a process of simplifying data and making it suitable for further analysis.



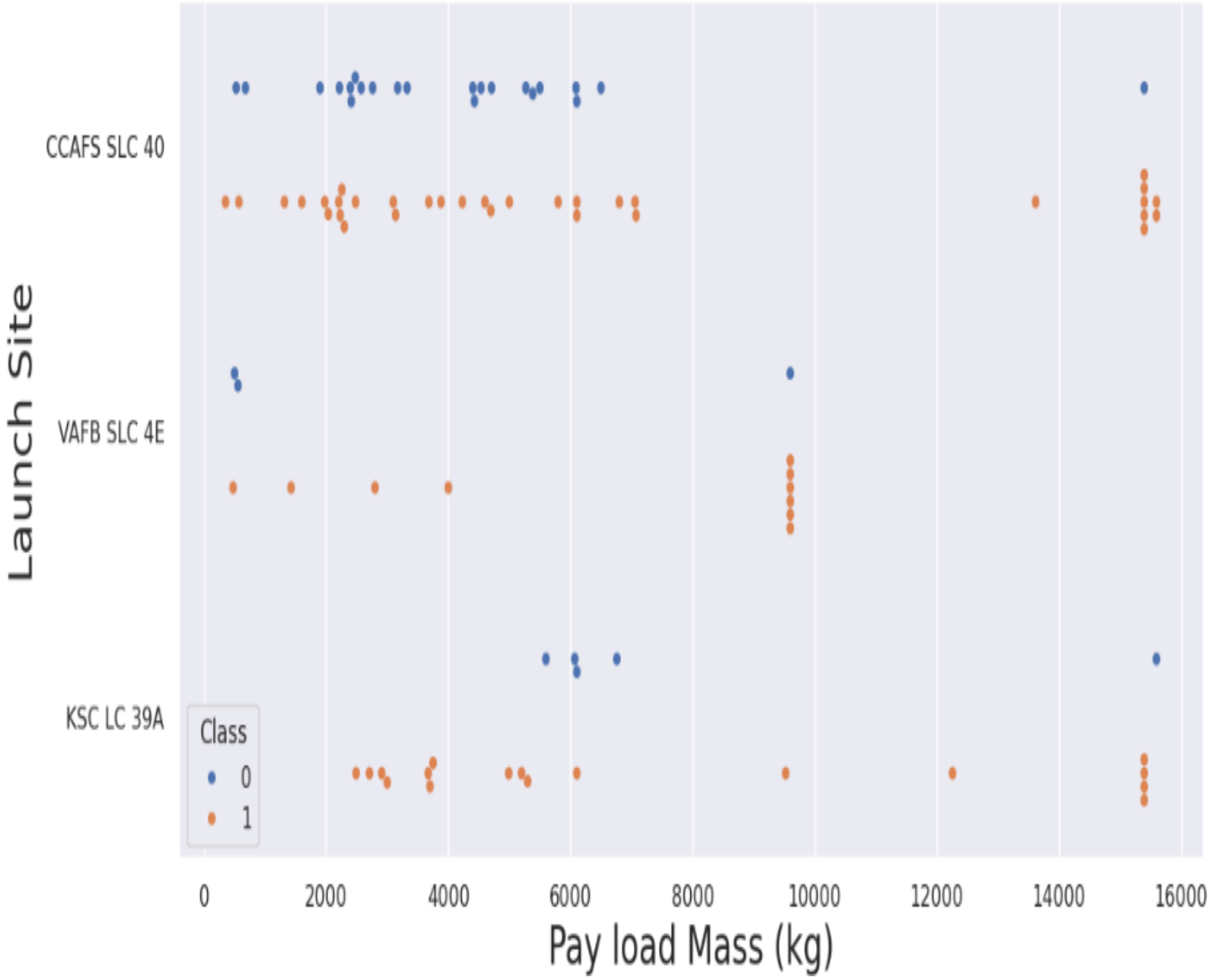
EDA with Data Visualization

- We see that as the flight number increases, the first stage is more likely to land successfully. The payload mass is also important; it seems the more massive the payload, the less likely the first stage will return.
- We see that different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.



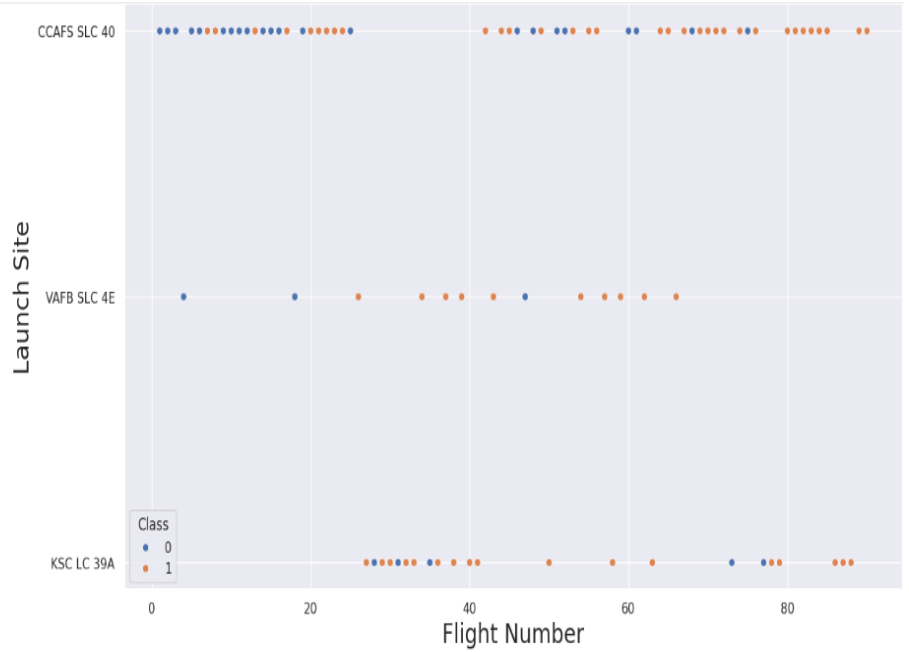
EDA with Data Visualization

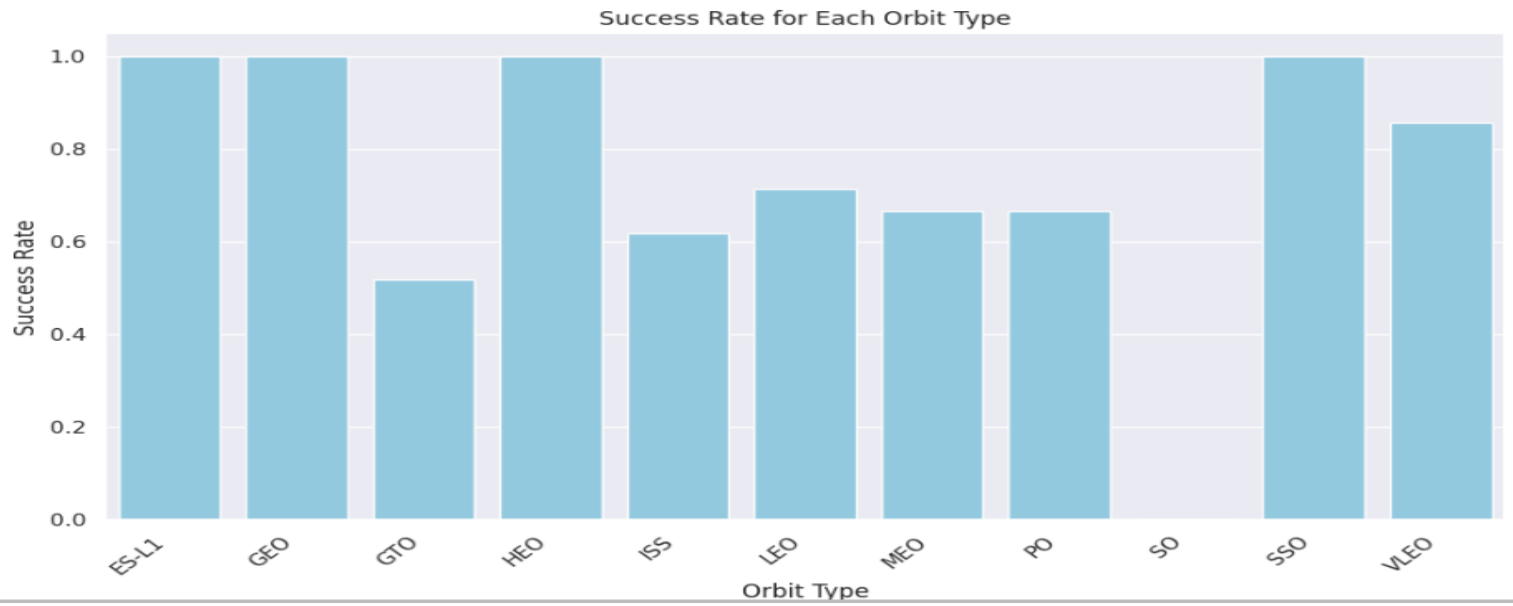
- Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000)



EDA with Data Visualization

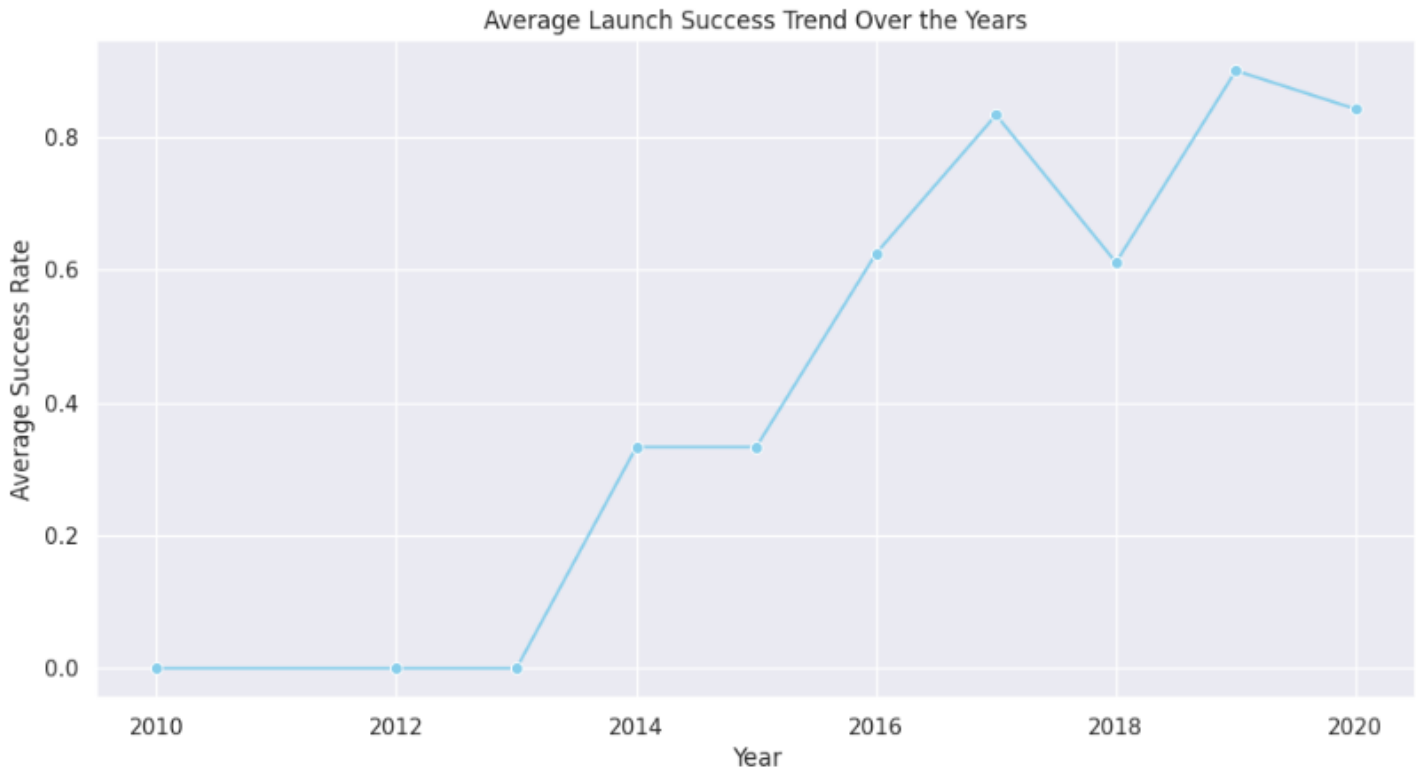
- Other Charts





EDA with Data Visualization

- You can observe that the success rate since 2013 kept increasing till 2020



EDA with SQL

- Loads database table SPACEXTBL from dataframe df using connection con:
  - df.to_sql("SPACEXTBL", con, if_exists='replace', index=False,method="multi")
- Creates Table SPACEXTABLE from SPACEXTBL using code:
  - %sql create table SPACEXTABLE as select * from SPACEXTBL where Date is not null
- Displays the names of the unique launch sites in the space mission
  - %sql SELECT distinct Launch_Site FROM SPACEXTABLE;
    - Result: CCAFS LC-40 VAFB SLC-4E KSC LC-39A CCAFS SLC-40
- Displays 5 records where launch sites begin with the string 'CCA'
  - %sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
- Displays the total payload mass carried by boosters launched by NASA (CRS)
  - %sql Select SUM(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = "NASA (CRS)"
    - Result: SUM(PAYLOAD_MASS__KG_) 45596

EDA with SQL

- Displays average payload mass carried by booster version F9 v1.1:
  - %sql Select AVG(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version = "F9 v1.1"
    - Result: AVG(PAYLOAD_MASS__KG_) 2928.4
- Lists the date when the first succesful landing outcome in ground pad was acheived.
  - %sql Select MIN(Date) from SPACEXTABLE where Landing_Outcome LIKE "Success%"
    - Result: MIN(Date) 2015-12-22
- Lists 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 SPACEXTABLE WHERE "Landing_Outcome" LIKE 'Success (drone ship)%' AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000;
    - Result: Booster_Version F9 FT B1022 F9 FT B1026 F9 FT B1021.2 F9 FT B1031.2
- Lists the total number of successful and failure mission outcomes
  - %sql SELECT CASE WHEN "Mission_Outcome" = 'Success' THEN 'Success' ELSE 'Failure' END as "Categorized_Outcome", COUNT(*) as "Total" FROM SPACEXTABLE GROUP BY "Categorized_Outcome";
    - Result: Categorized_Outcome
      - Total Failure 3 Success 98

EDA with SQL

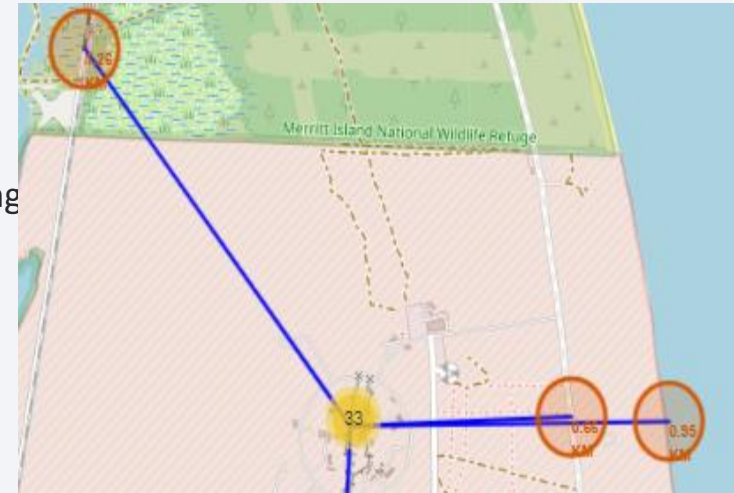
- Displays average payload mass carried by booster version F9 v1.1:
  - %sql Select AVG(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version = "F9 v1.1"
    - Result: AVG(PAYLOAD_MASS__KG_) 2928.4
- Lists the date when the first succesful landing outcome in ground pad was acheived.
  - %sql Select MIN(Date) from SPACEXTABLE where Landing_Outcome LIKE "Success%"
    - Result: MIN(Date) 2015-12-22
- Lists the names of the booster_versions which have carried the maximum payload mass. Use a subquery
  - %sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTABLE);

EDA with SQL

- Lists the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
    - %sql SELECT SUBSTR("Date", 6, 2) AS "Month", "Landing_Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTABLE WHERE SUBSTR("Date", 0, 5) = '2015' AND "Landing_Outcome" = 'Failure (drone ship)'
- Ranks 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 "Outcome_Count" FROM SPACEXTABLE WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20' AND "Landing_Outcome" IN ('Success (ground pad)', 'Failure (drone ship)') GROUP BY "Landing_Outcome" ORDER BY "Outcome_Count" DESC;

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
- Markers help to pin the location on to the map. Locations are like these:
  - closest_highway = [28.5625, -80.57063]
  - closest_railway = [28.57117, -80.58541]
  - closest_city = [28.07681, -80.60621] * Note all digits after decimal
- While Markers just let the map know which point to mark, the actual circle is made by the circle command
  - circle = folium.Circle(coordinate, radius=100, color='#d35400', fill=True)
    .add_child(folium.Popup('Coastline'))
  - site_map.add_child(circle)
- Line are used to connect two markers
  - lines = folium.PolyLine(locations=[[launch_site_lat, launch_site_long], [coastline_lat, coastline_long
    , color='blue', weight=2.5)
  - site_map.add_child(lines)



Build an Interactive Map with Folium

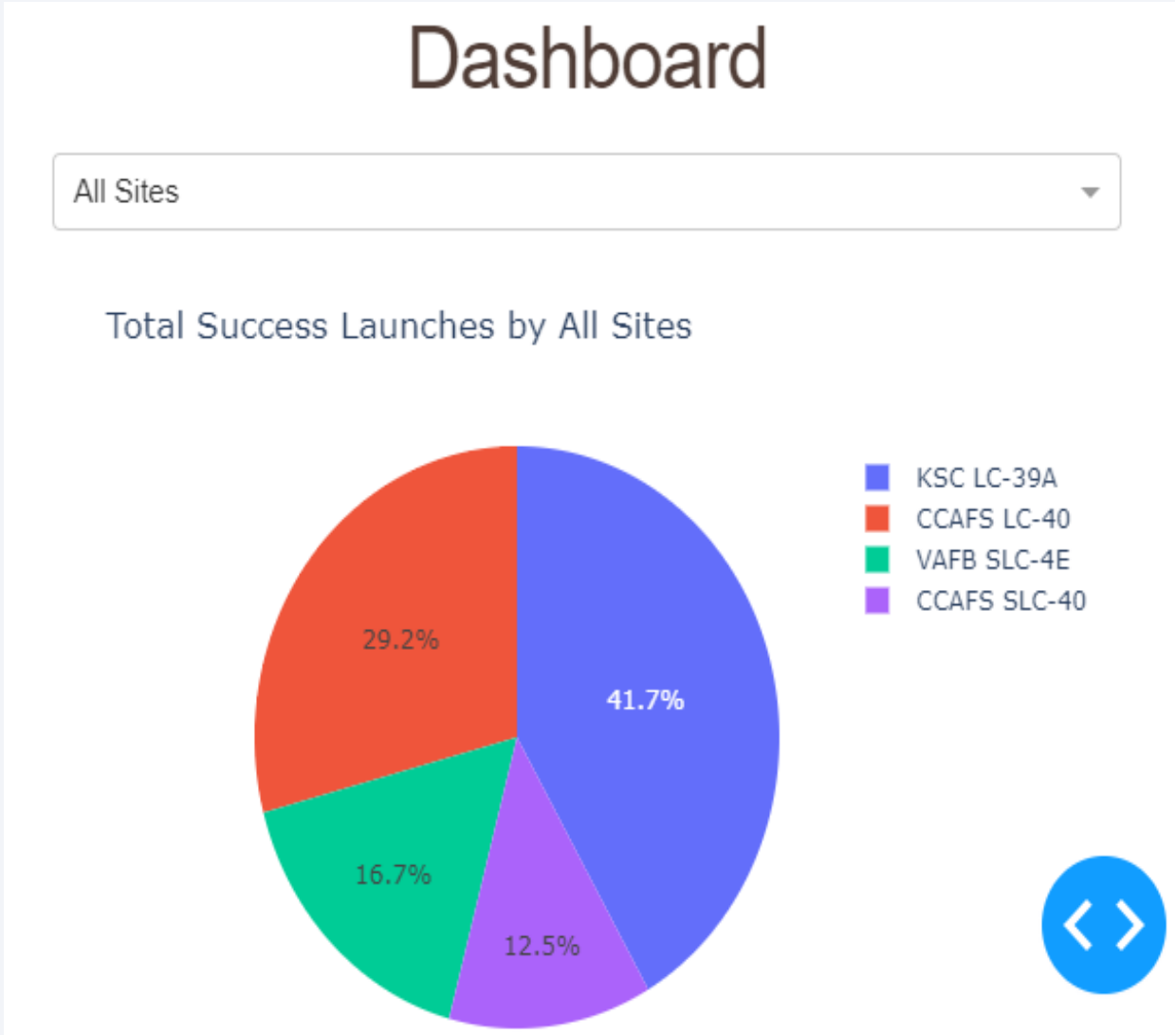
- Sample code for marker:

```
coordinate = [coastline_lat, coastline_long]
distance_marker = folium.Marker(
  coordinate,
  icon=DivIcon(
    icon_size=(20,20),
    icon_anchor=(0,0),
    html='<div style="font-size: 12; color:#d35400;"><b>%s</b></div>'  % "{:10.2f} KM".format(distance_coastline),
    )

  )

site_map.add_child(distance_marker)
```

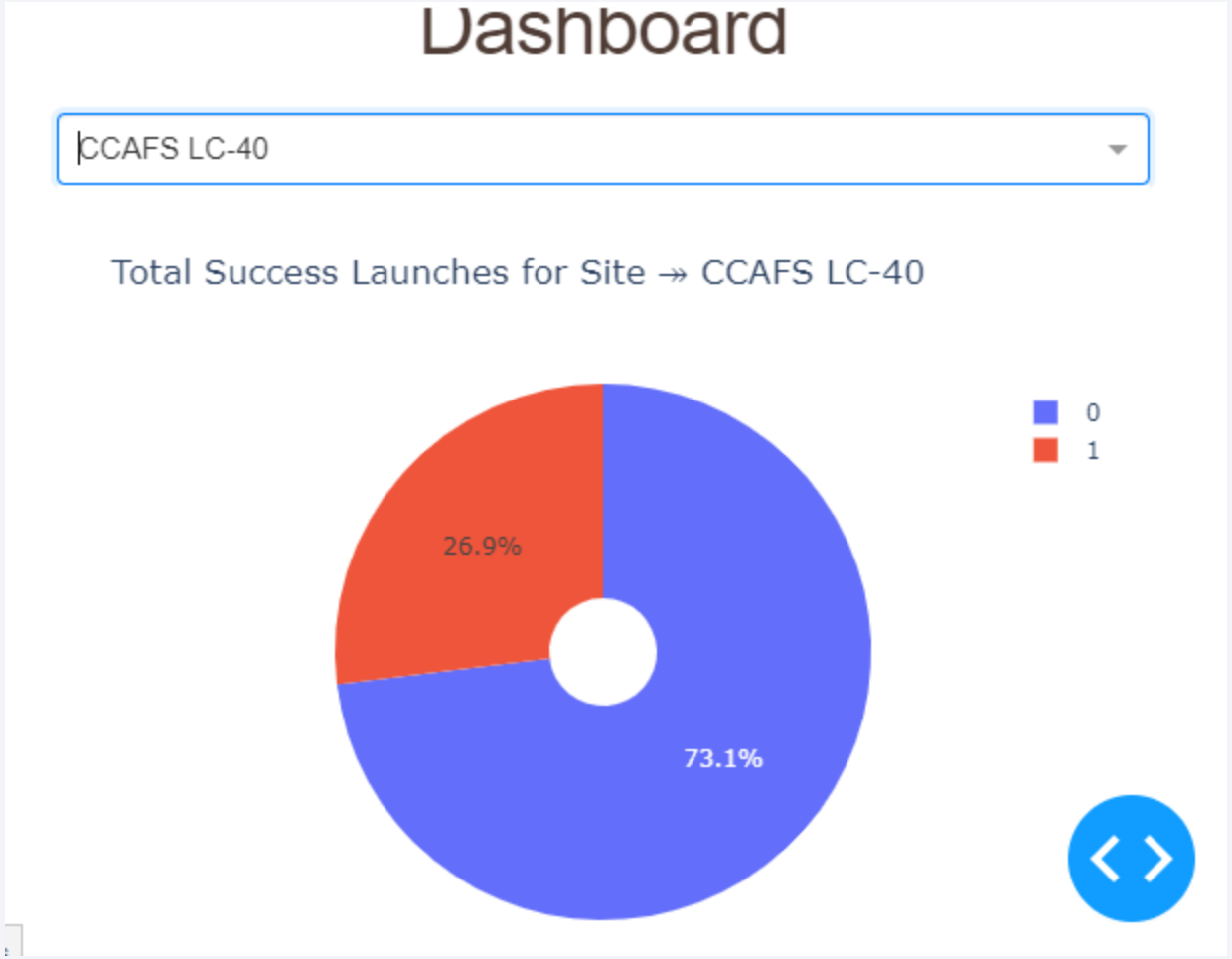
- MousePosition(position='topleft', prefix='Lat:', separator=' Long:').add_to(site_map) can be used to display live coordinates on Mouse Over.
- While folium.Map() gives whole word map,
  - folium.Map(location=nasa_coordinate, zoom_start=5) gives map centering on nasa coordinates and zoom level of 5.

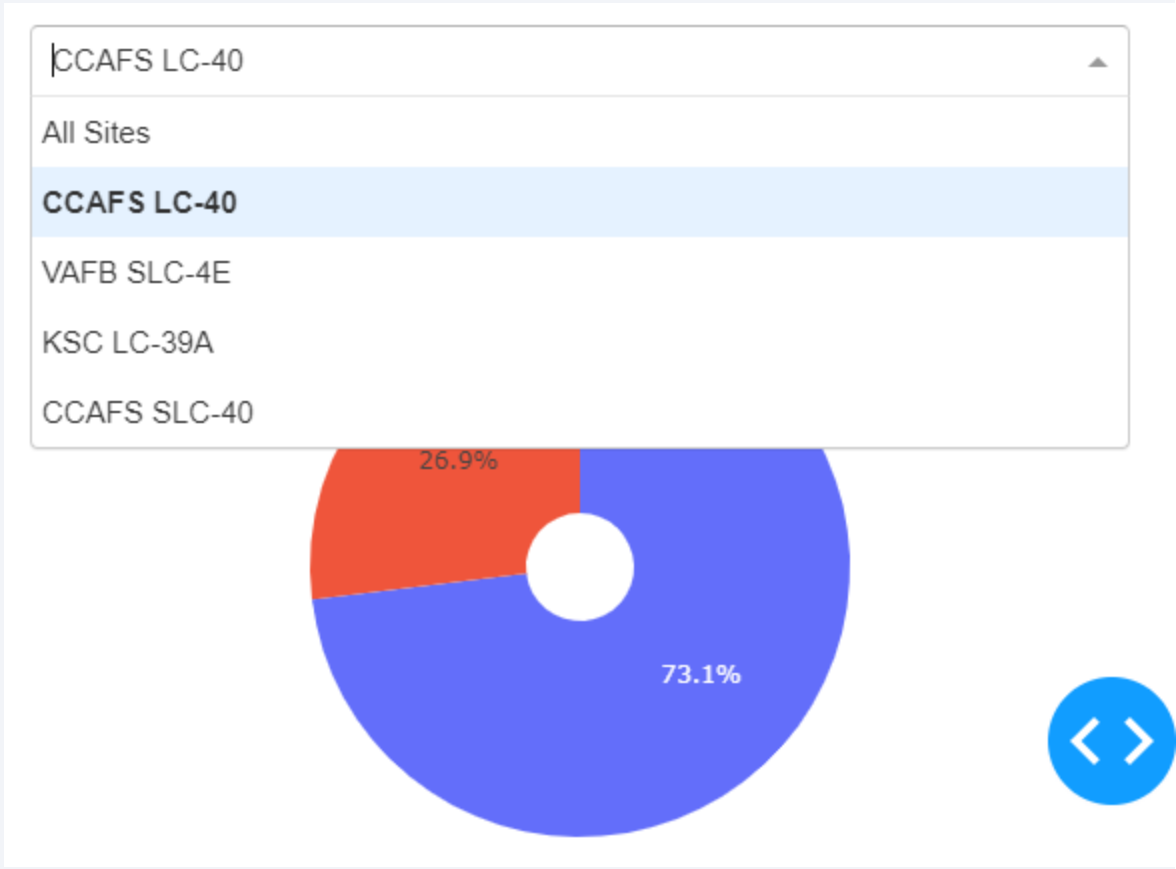
Build a Dashboard with Plotly Dash

- Dashboard gives drop down for either one site or all sites. Please refer next slide for more images.
- In case of all sites, only success cases are taken
- In case of single sites, both success and failure cases are taken.



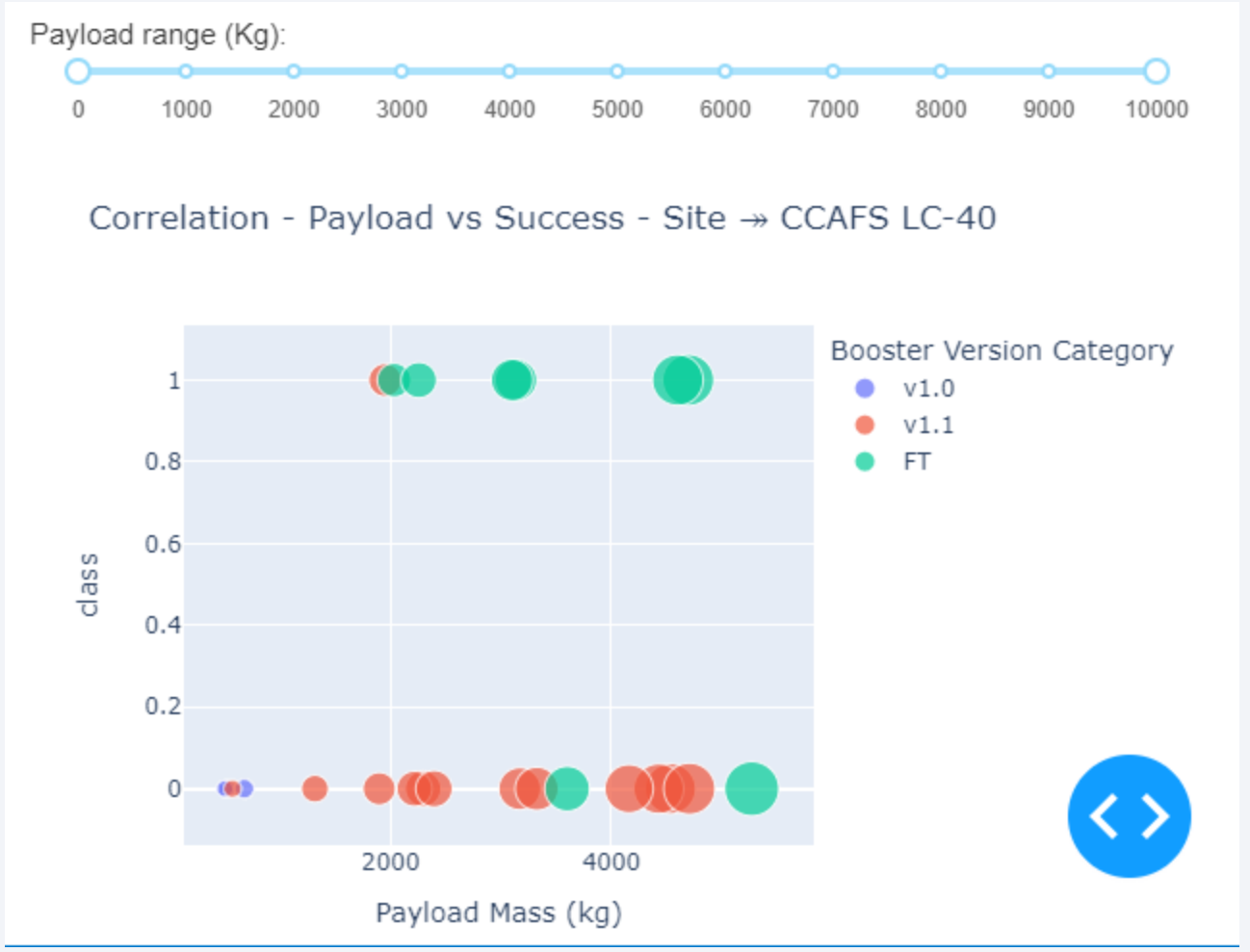
Build a Dashboard with Plotly Dash





Build a Dashboard with Plotly Dash

- This is slider. Slider acts as an input to the graph. The low and high values received from the slider filters the data for the graph within that load range.
- The color is based on Booster version category.
- This shows coorelation between successful launches for various payload masses.

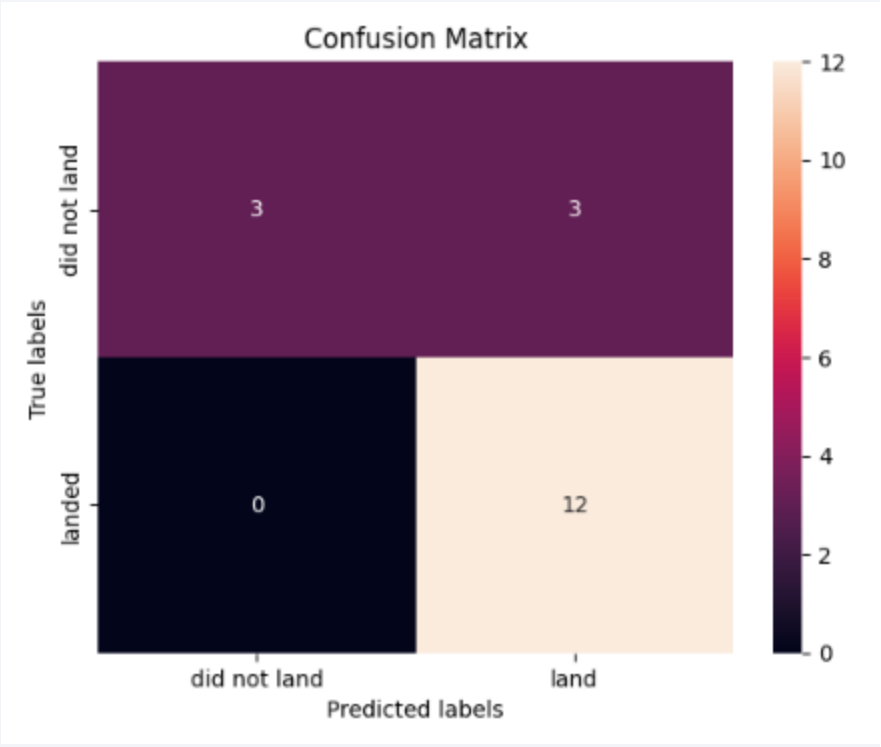


Predictive Analysis (Classification)

- All four Logistic Regression, SVM, Decision Tree and KNN have the same confusion matrix.
- To understand confusion, matrix, follow this link:

https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5

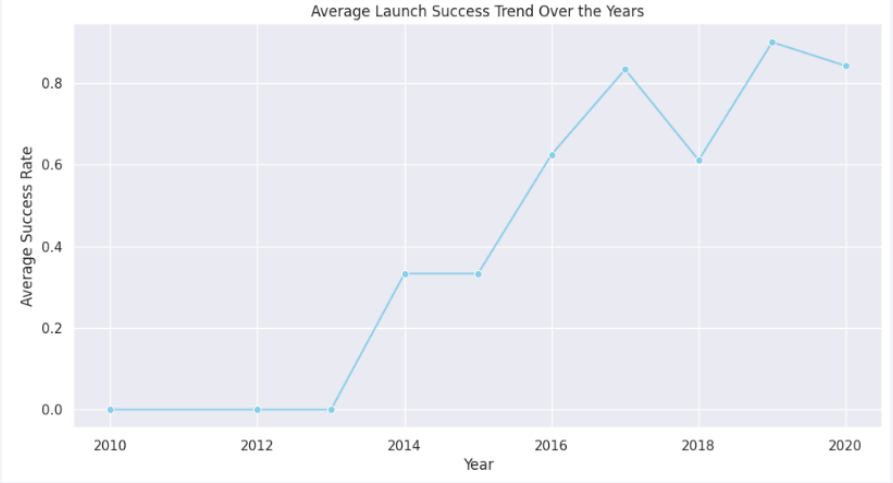
- Accuracy for logistic regression using the method score: 0.8333333333333334
-
- Suitable kernel was sigmoid.
- We had 18 test samples.

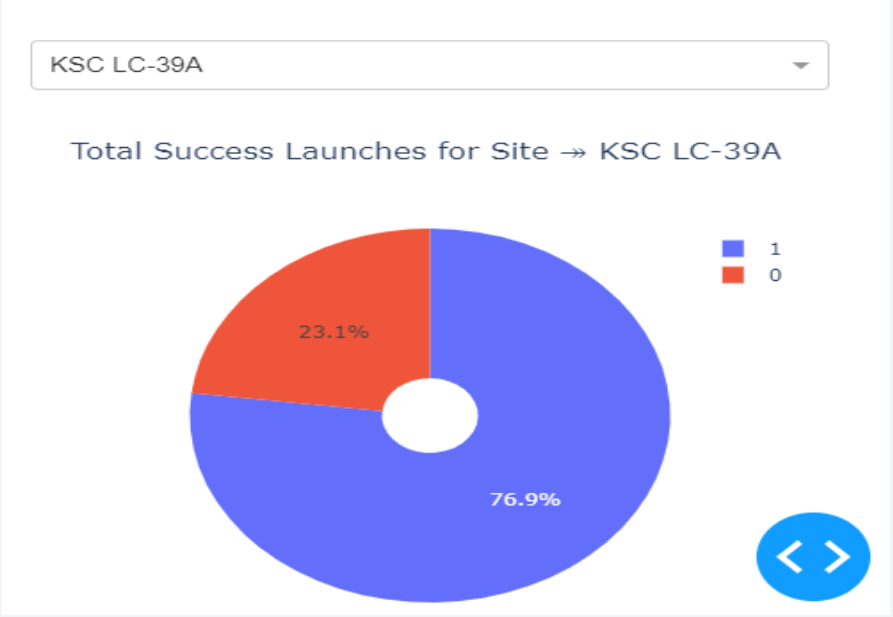


Results

- Exploratory data analysis results
  - You can observe that the sucess rate since 2013 kept increasing till 2020
- Interactive analytics demo in screenshots
  - KSC LC-39A has the most success rate.
- Predictive analysis results

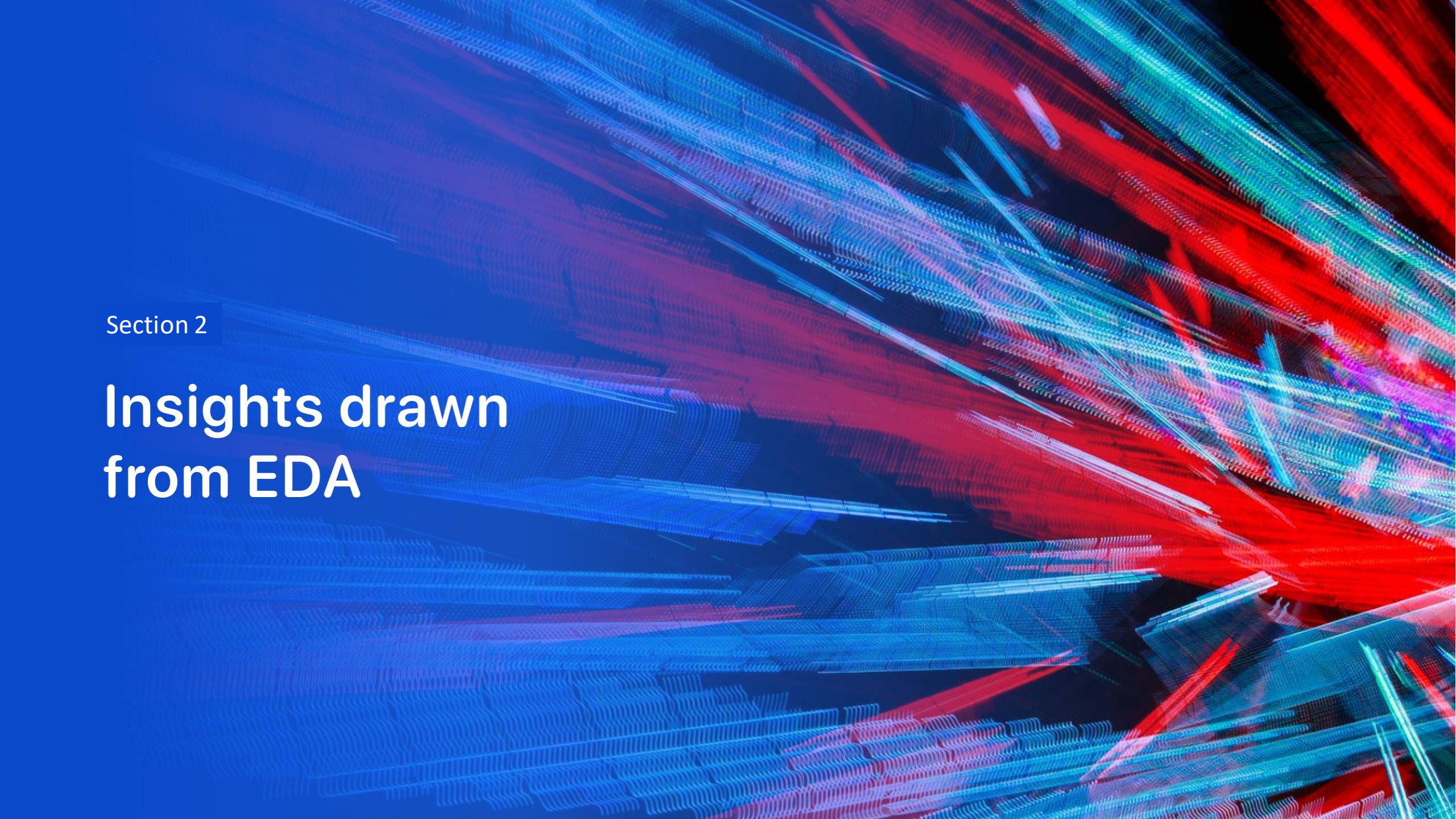
  Sigmoid Kernel has the best result of the validation dataset.





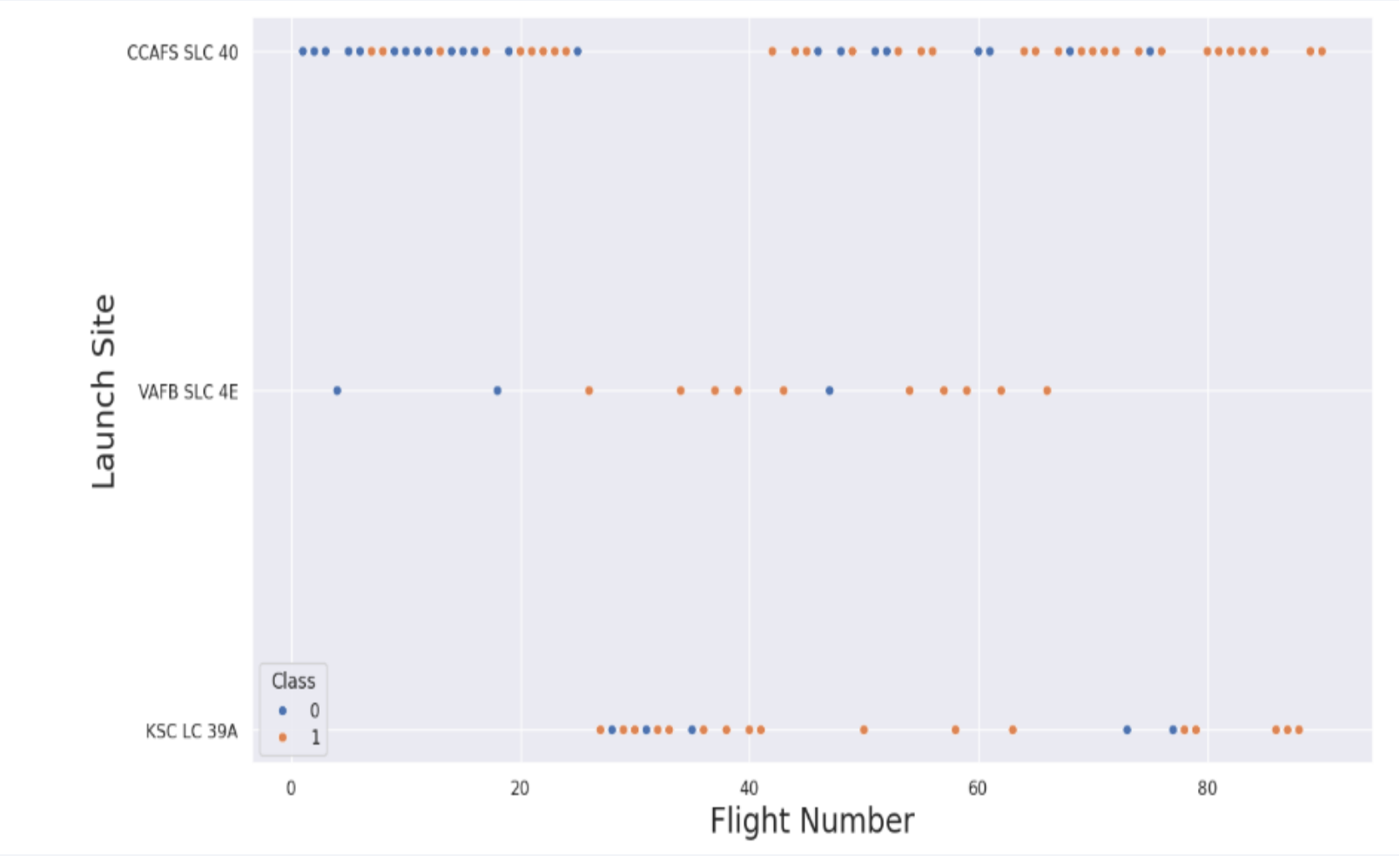
Section 2

Insights drawn from EDA



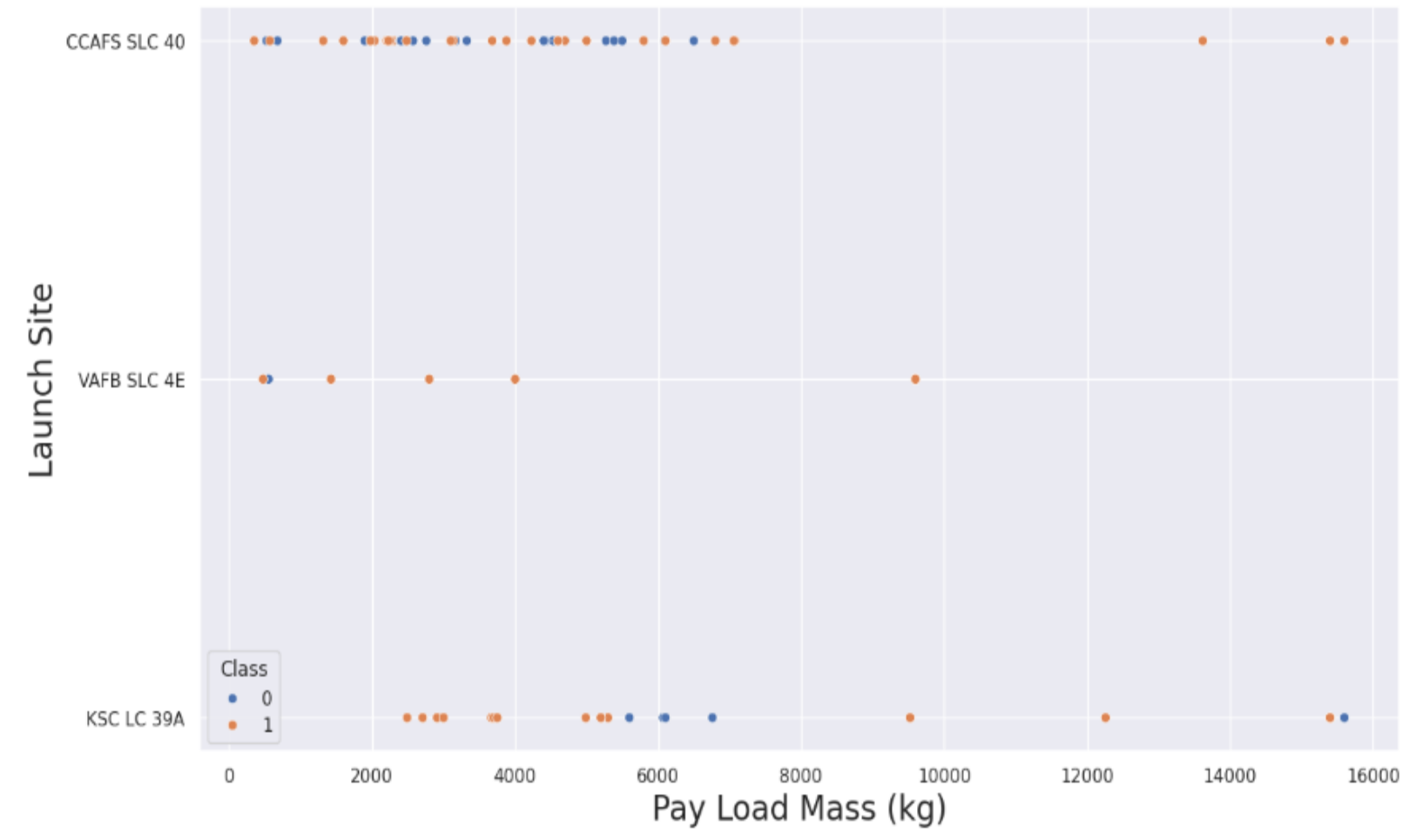
Flight Number vs. Launch Site

- Success rate increasing with higher flight numbers.



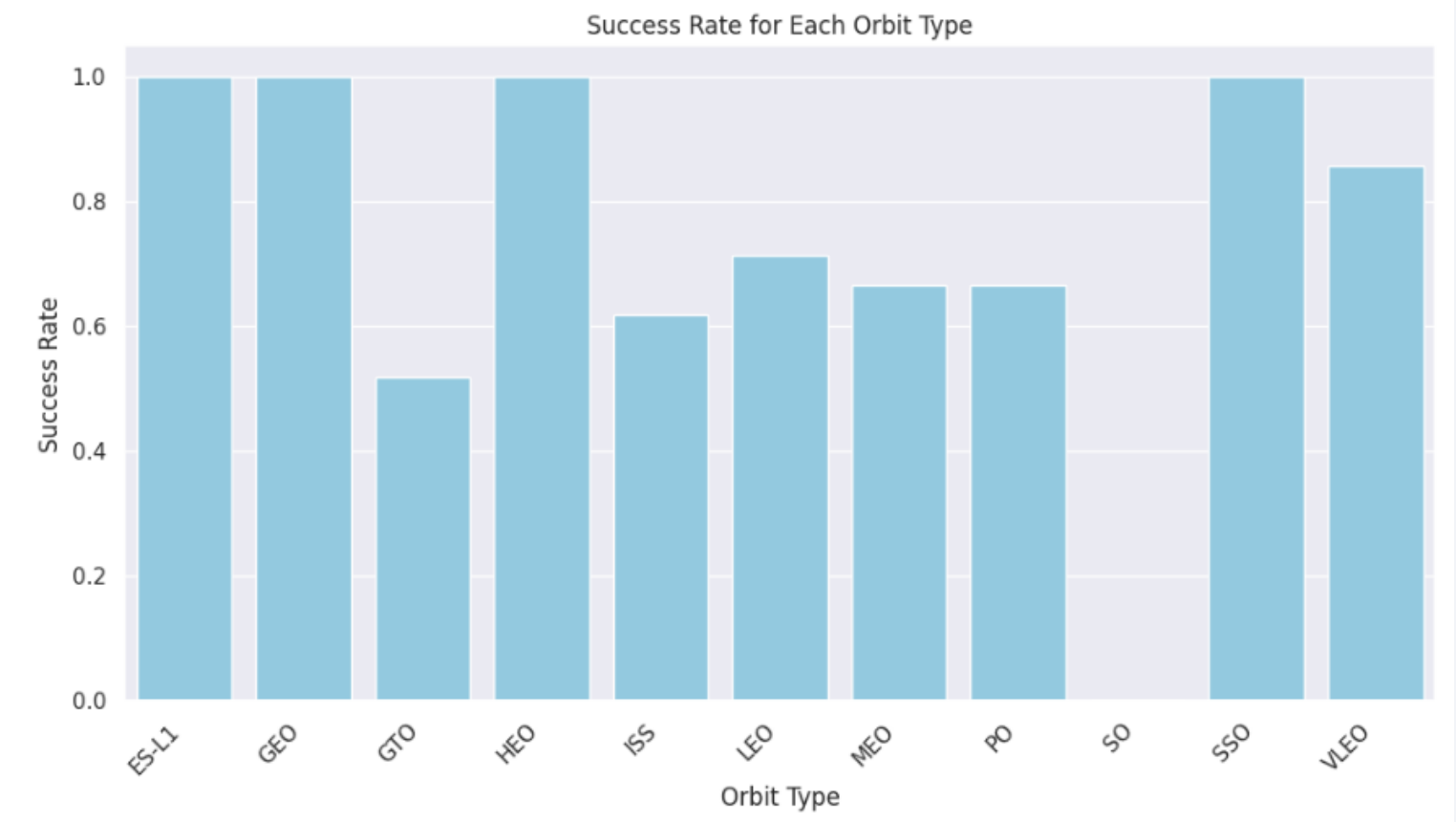
Payload vs. Launch Site

- You will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).



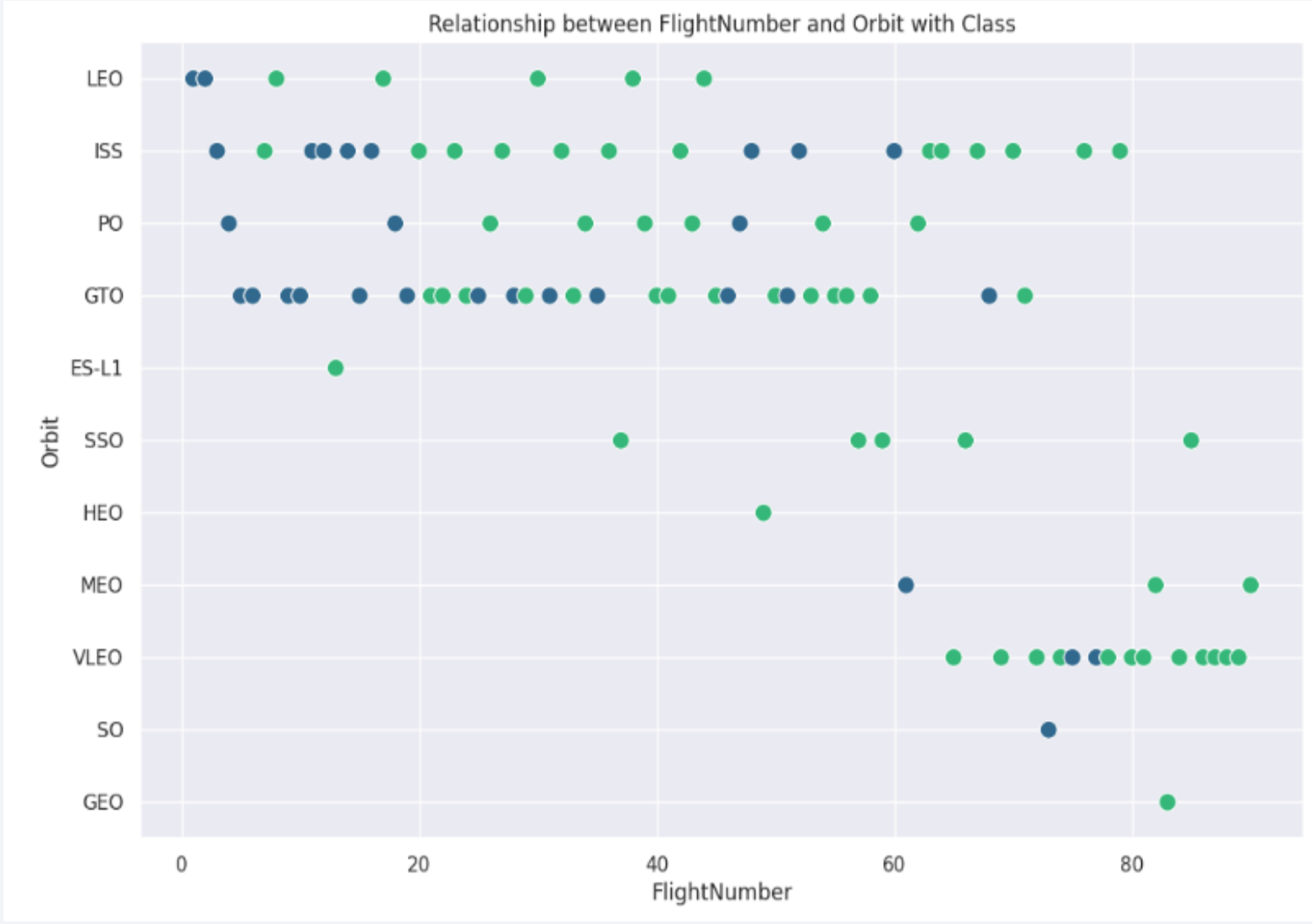
Success Rate vs. Orbit Type

- These orbits have high success rates:
  - ES-L1
  - GEO
  - HEO
  - SSO



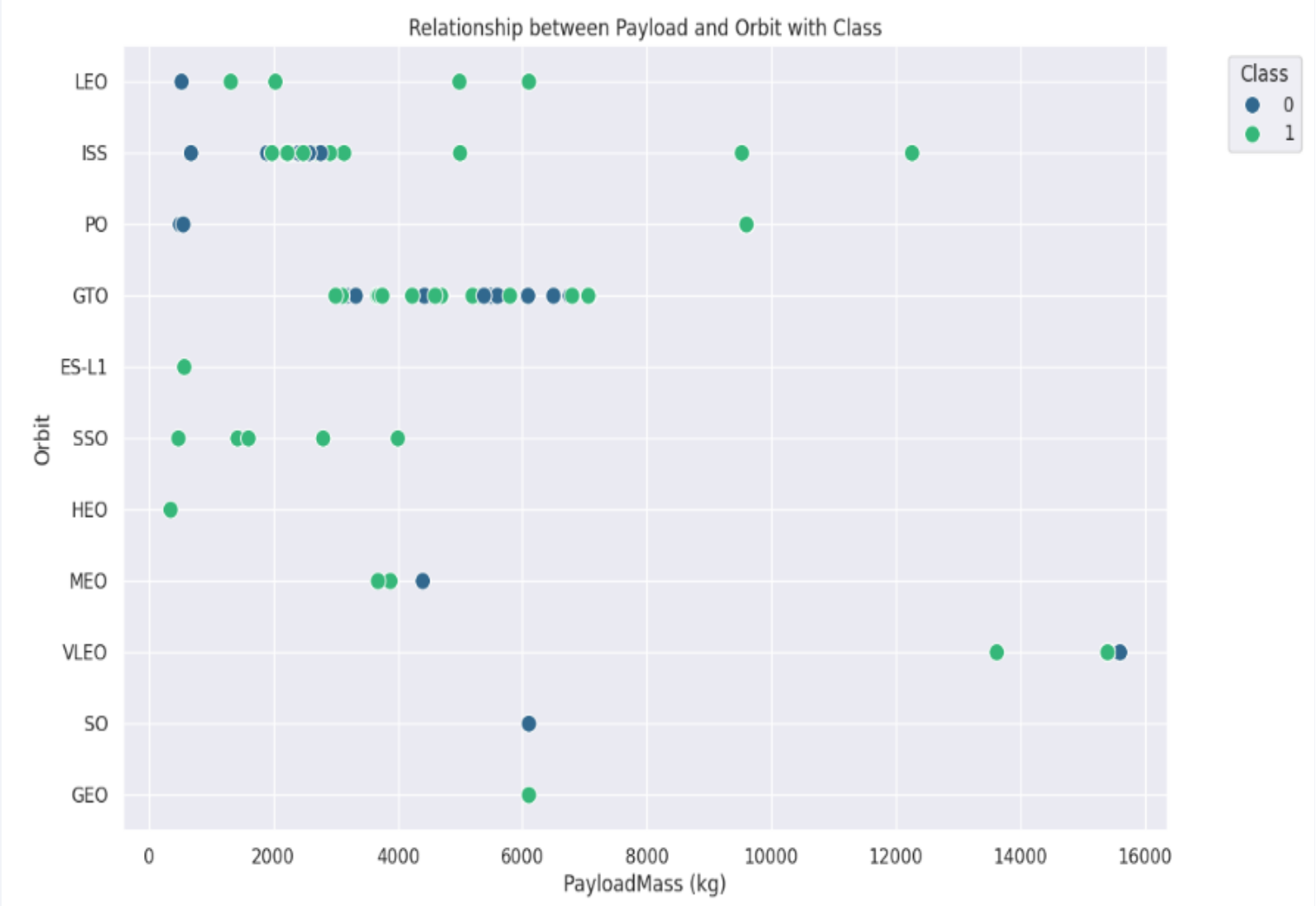
Flight Number vs. Orbit Type

- You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit



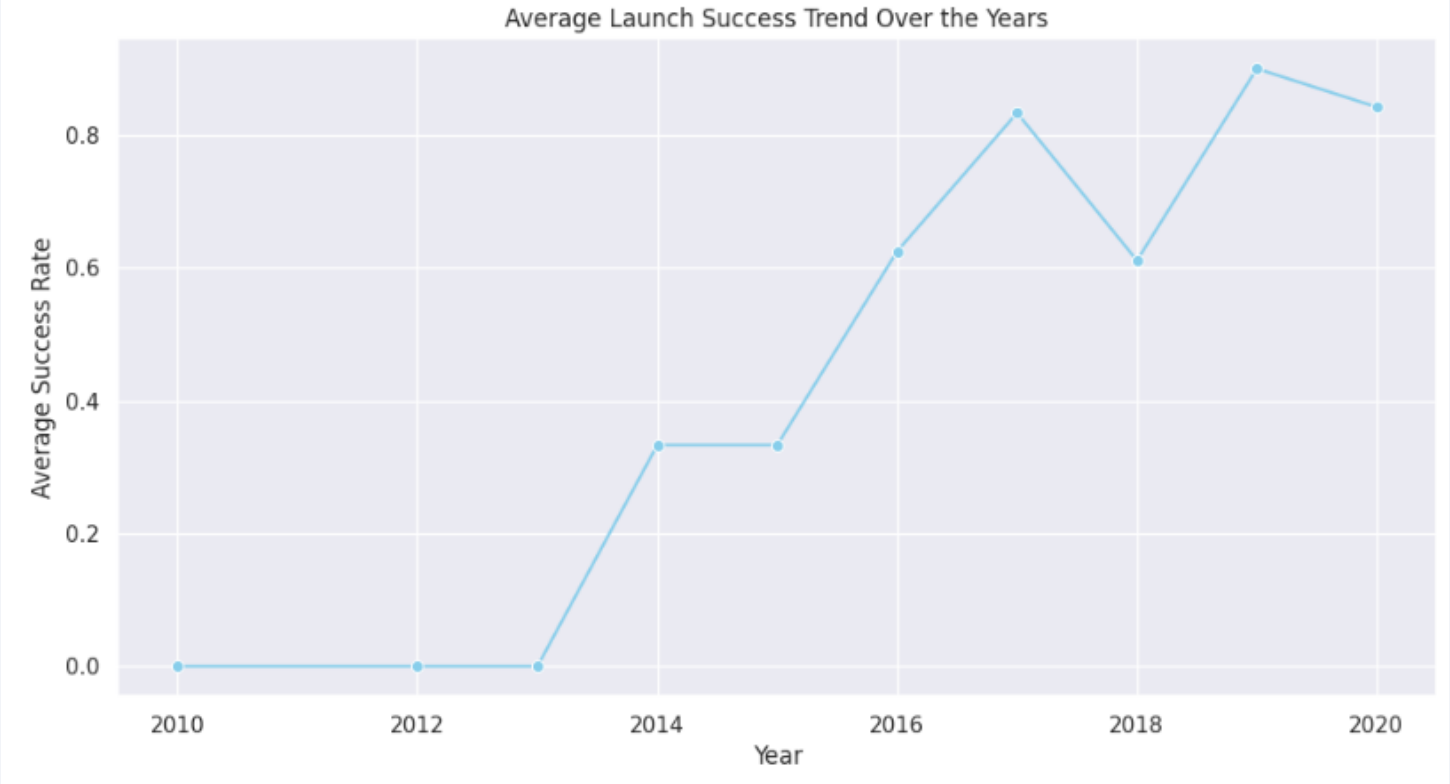
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.



Launch Success Yearly Trend

- You can observe that the success rate since 2013 kept increasing till 2020



All Launch Site Names

- Displays the names of the unique launch sites in the space mission
  - %sql SELECT distinct Launch_Site FROM SPACEXTABLE;
    - Result:

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

    - These are unique list of names of the launch sites.

Launch Site Names Begin with 'CCA'

- Displays 5 records where launch sites begin with the string 'CCA'
  - %sql 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_Outc
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parach
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 (parach
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No atte
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No atte
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No atte

Total Payload Mass

- Displays the total payload mass carried by boosters launched by NASA (CRS)
  - %sql Select SUM(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = "NASA (CRS)"
    - Result: SUM(PAYLOAD_MASS__KG_) 45596

SUM(PAYLOAD_MASS__KG_)
45596

Su

Average Payload Mass by F9 v1.1

- Displays average payload mass carried by booster version F9 v1.1:
  - %sql Select AVG(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version = "F9 v1.1"
    - Result: AVG(PAYLOAD_MASS__KG_) 2928.4

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Lists the date when the first succesful landing outcome in ground pad was acheived.
  - %sql Select MIN(Date) from SPACEXTABLE where Landing_Outcome LIKE "Success%"
    - Result: MIN(Date) 2015-12-22

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Lists 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 SPACEXTABLE WHERE "Landing_Outcome" LIKE 'Success (drone ship)%' AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000;
    - Result: Booster_Version F9 FT B1022 F9 FT B1026 F9 FT B1021.2 F9 FT B1031.2

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

Total Number of Successful and Failure Mission Outcomes

- Lists the total number of successful and failure mission outcomes
  - %sql SELECT CASE WHEN "Mission_Outcome" = 'Success' THEN 'Success' ELSE 'Failure' END as "Categorized_Outcome", COUNT(*) as "Total" FROM SPACEXTABLE GROUP BY "Categorized_Outcome";
    - Result: Categorized_Outcome
      - Total Failure 3 Success 98

Categorized_Outcome	Total
Failure	3
Success	98

Boosters Carried Maximum Payload

- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
- `%sql SELECT "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

- Lists the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
  - `%sql SELECT SUBSTR("Date", 6, 2) AS "Month", "Landing_Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTABLE WHERE SUBSTR("Date", 0, 5) = '2015' AND "Landing_Outcome" = 'Failure (drone ship)'`

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

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(*) AS "Outcome_Count" FROM SPACEXTABLE WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20' AND "Landing_Outcome" IN ('Success (ground pad)', 'Failure (drone ship)') GROUP BY "Landing_Outcome" ORDER BY "Outcome_Count" DESC;`

Landing_Outcome	Outcome_Count
Failure (drone ship)	5
Success (ground pad)	3


Section 3
Launch Sites
Proximities Analysis

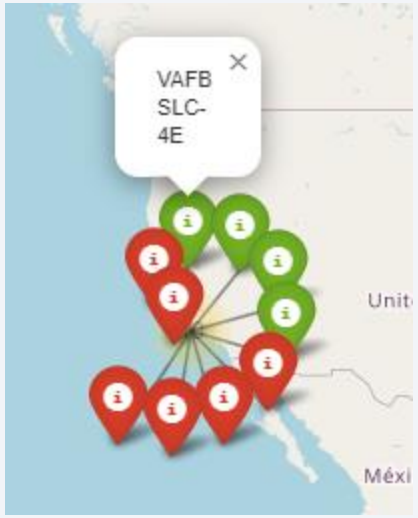
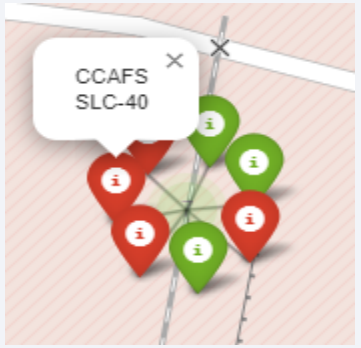
All Launch Sites at a glance

--VAFB SLC 4E is on left side directly on the global map

--KSC LC-39A, CCAFS SLC-40 and CCAFS LC-40 are on the right side, those are zoomed out to show them all on a single global map.

Color Labeled Launch Site Details

- Maximum Green(Success) is for KSC LC-39A

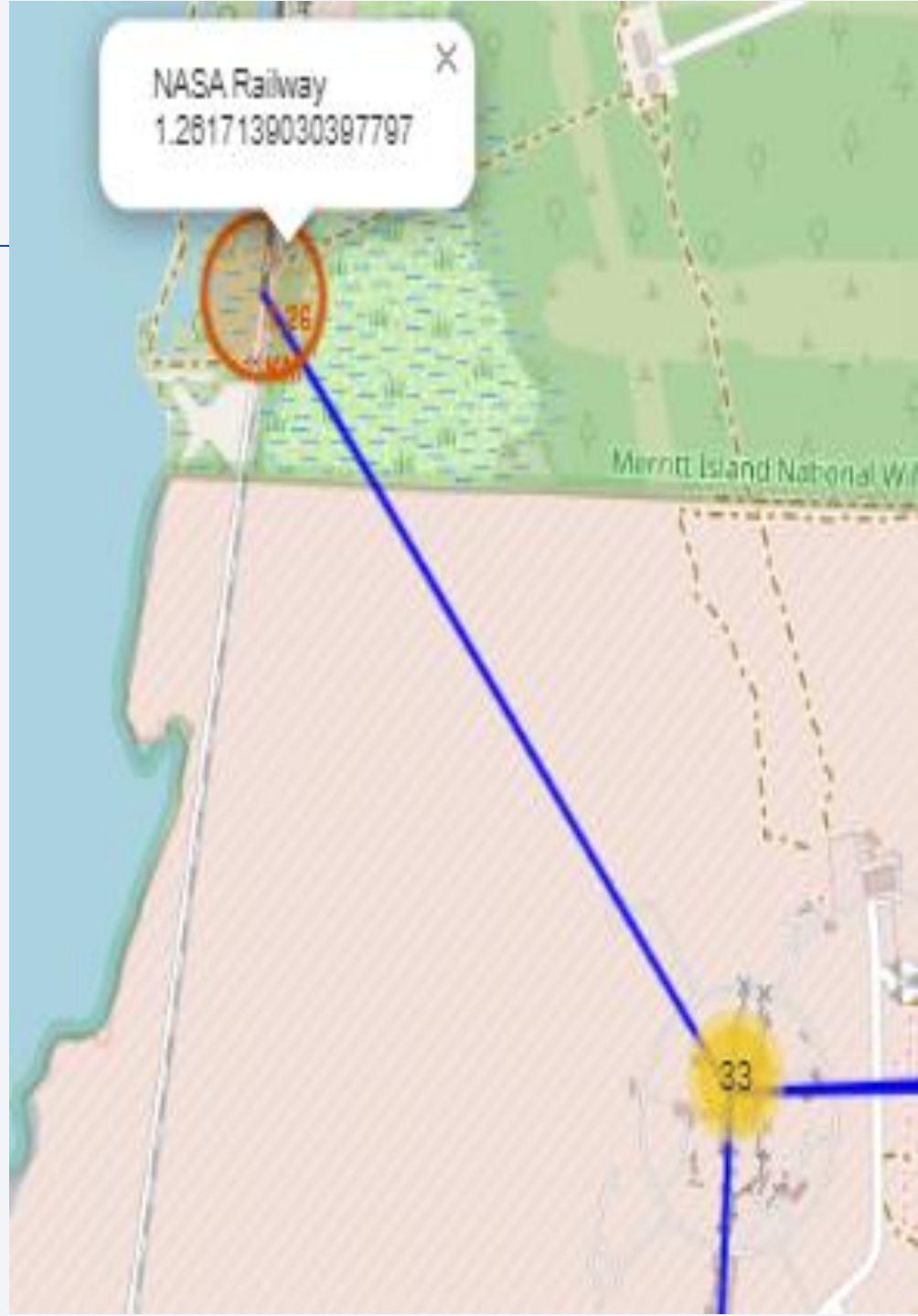
KSC LC-39A	VAFB SLC-40	CCAFS LC-40	CCAFS SLC-40
			

Launch Site Distances – Highway and Coastline

- 0.66 KM from CCAFS LC-40 – Highway
- 0.95 KM from CCAFS LC-40 – Coastline
- Note: Zoom and see the KM

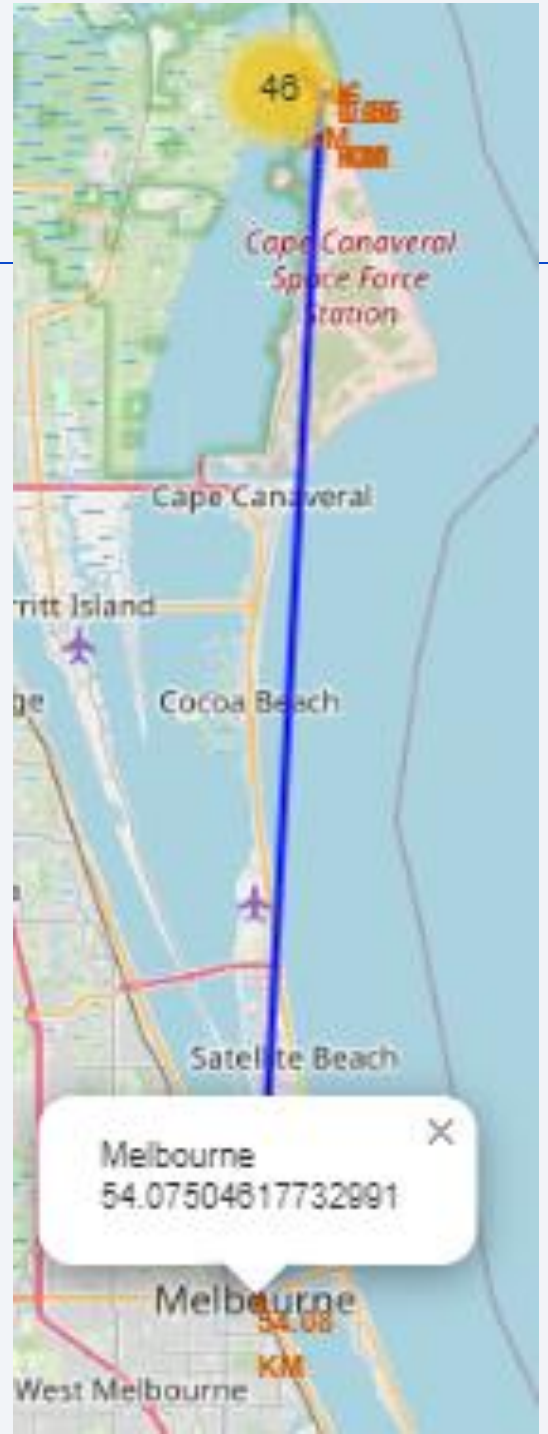
Launch Site Distances – Railway

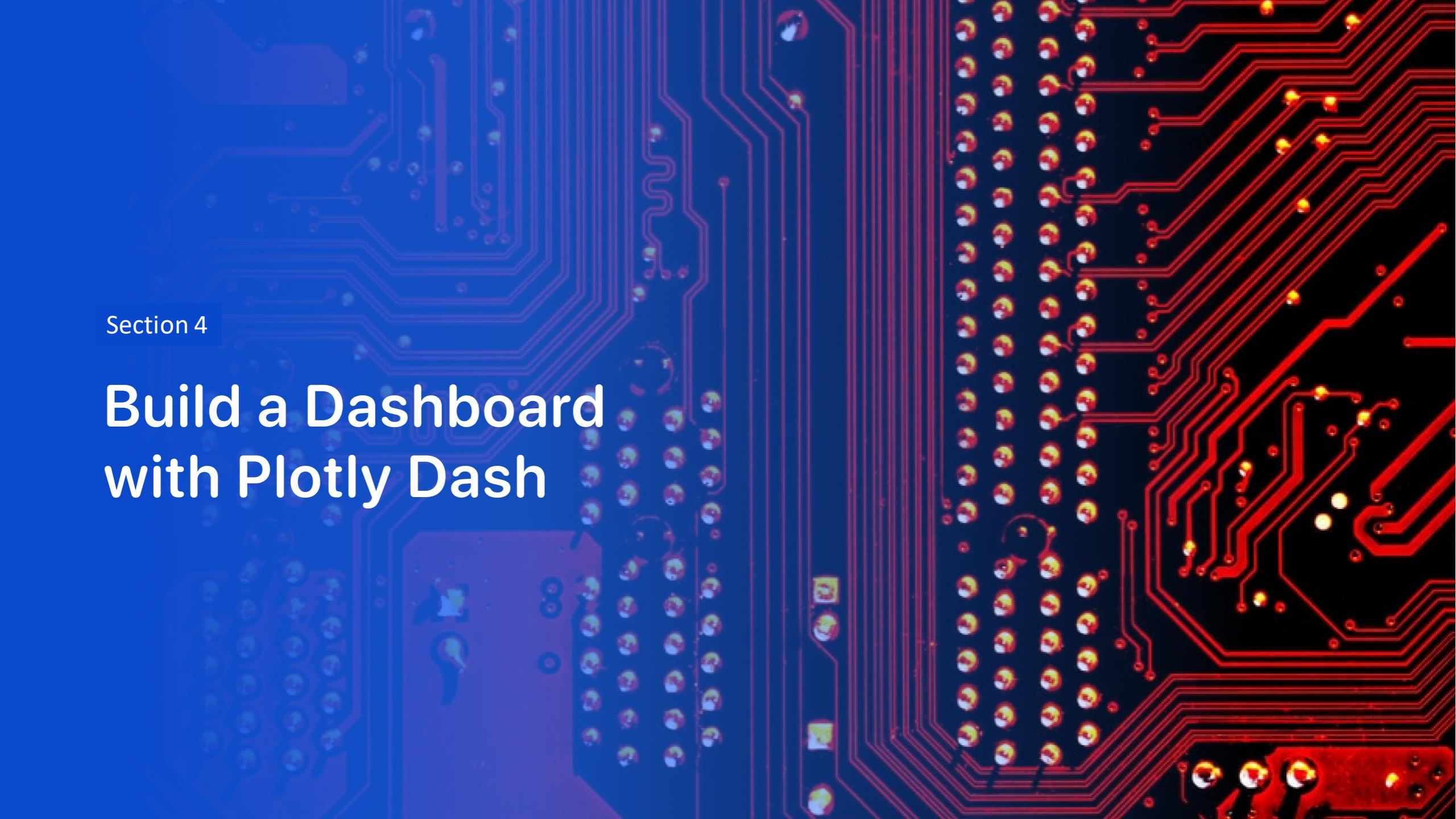
- 1.26 KM from CCAFS LC-40
- Note: Zoom and see the KM



Launch Site Distances – City

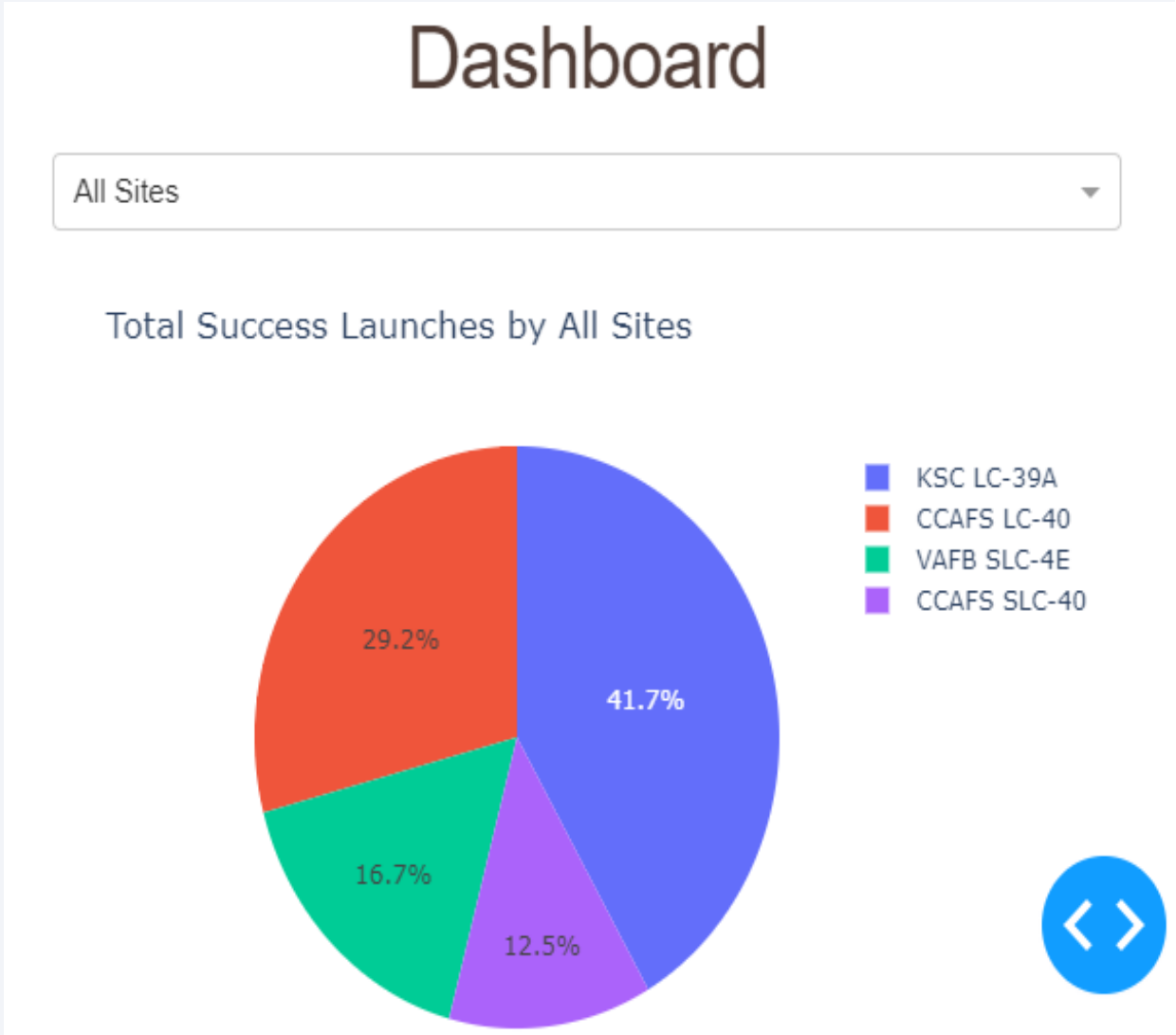
- 54.07 KM from CCAFS LC-40
- Note: Zoom and see the KM




Section 4
Build a Dashboard with Plotly Dash

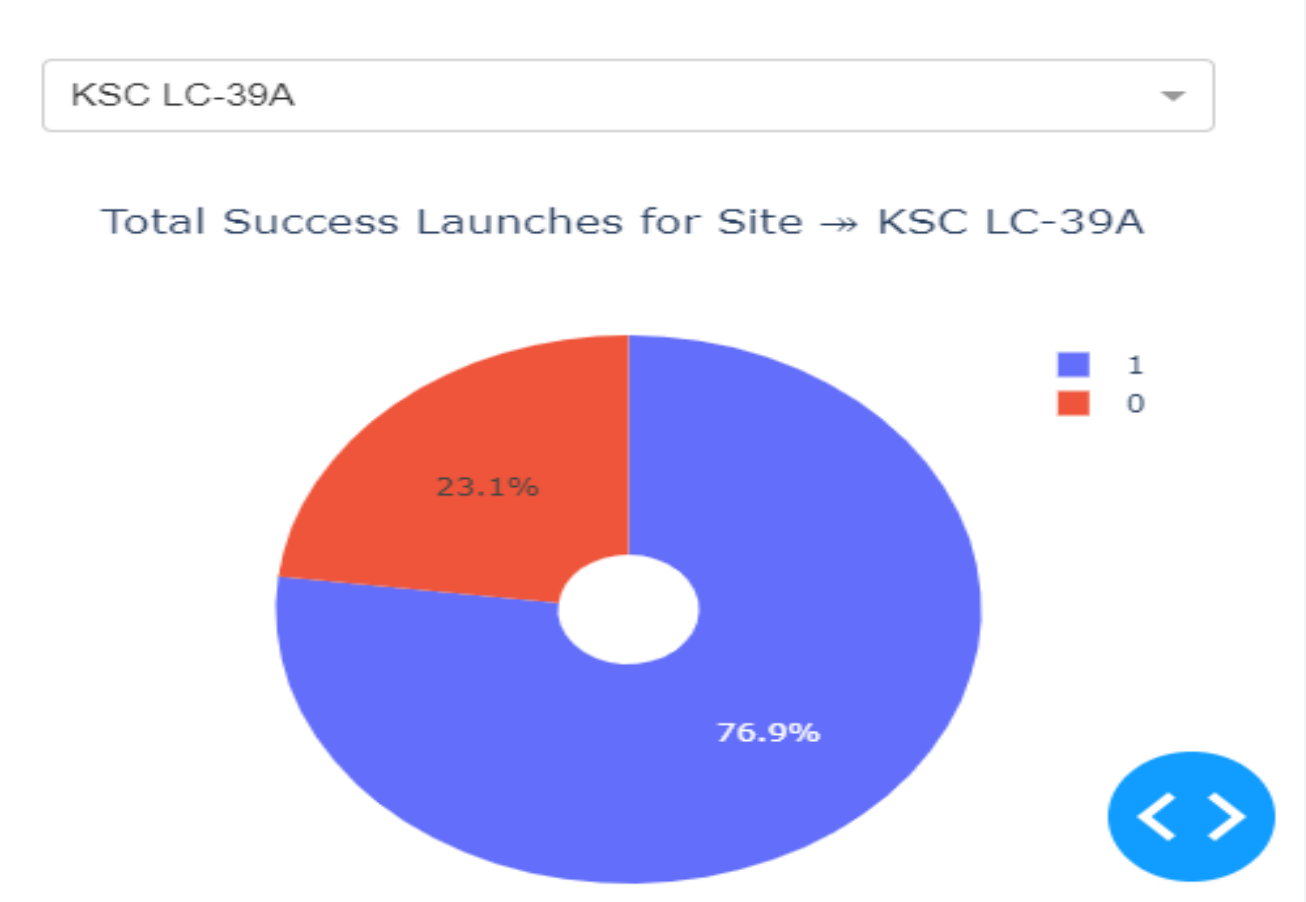
Launch success count for all sites, in a piechart

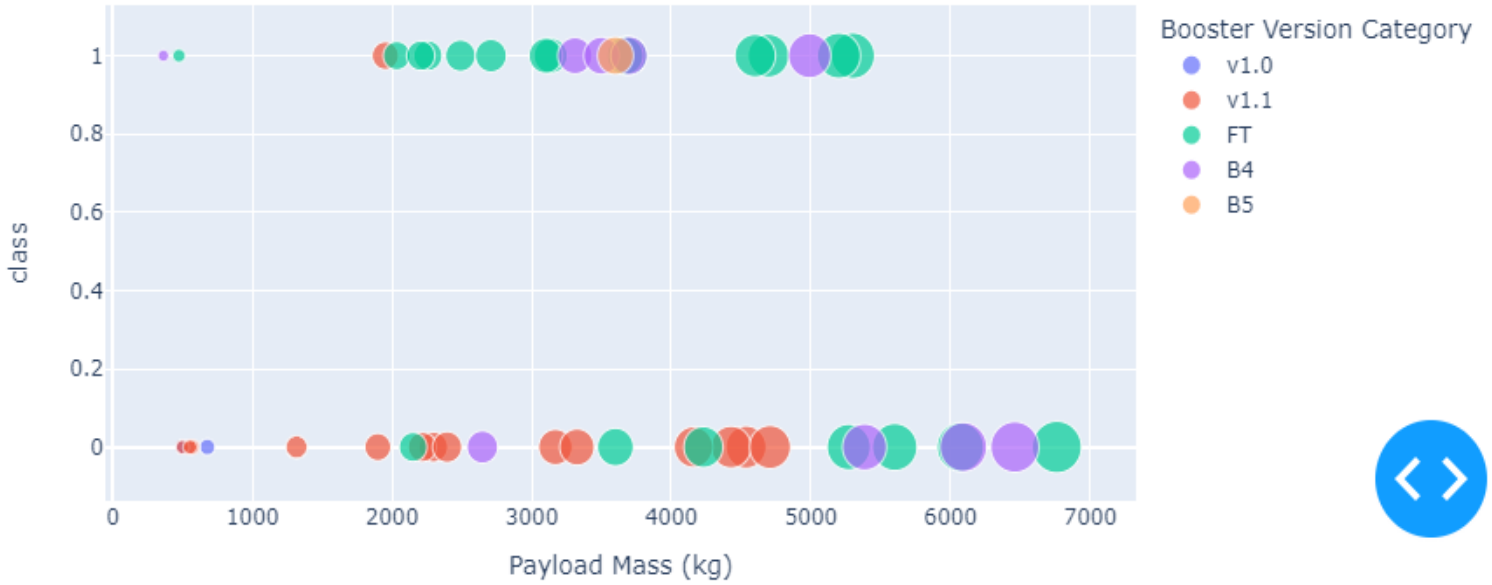
- KS LC-39A has the maximum success rate.



Piechart for the launch site with highest launch success ratio

- KSC LC-39A has the most success rate.





- Shows screenshot of Payload vs. Launch Outcome scatter plot for all sites, with range from 0 to 10000, pls see next slide for another screenshot

- Shows screenshot of Payload vs. Launch Outcome scatter plot for all sites, with range from 2000 to 8000, pls see next slide for another screenshot

Payload vs. Launch Outcome scatter plot for all sites

Section 5

Predictive Analysis (Classification)

Classification Accuracy

ALGO	ACCURACY	ACCURACY ON TEST DATA	TUNE HYPERPARAMETERS
Logistic Regresion	0.846428571	0.833333333	{'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}
SVM	0.848214286	0.833333333	{'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}
KNN	0.848214286	0.833333333	{'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}
Decision Tree	0.876785714	0.833333333	{'criterion': 'gini', 'max_depth': 14, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 10, 'splitter': 'best'}

Decision Tree is the winner in the accuracy.

```
The method which performs best is " Decision Tree " with a score of 0.8767857142857143
```

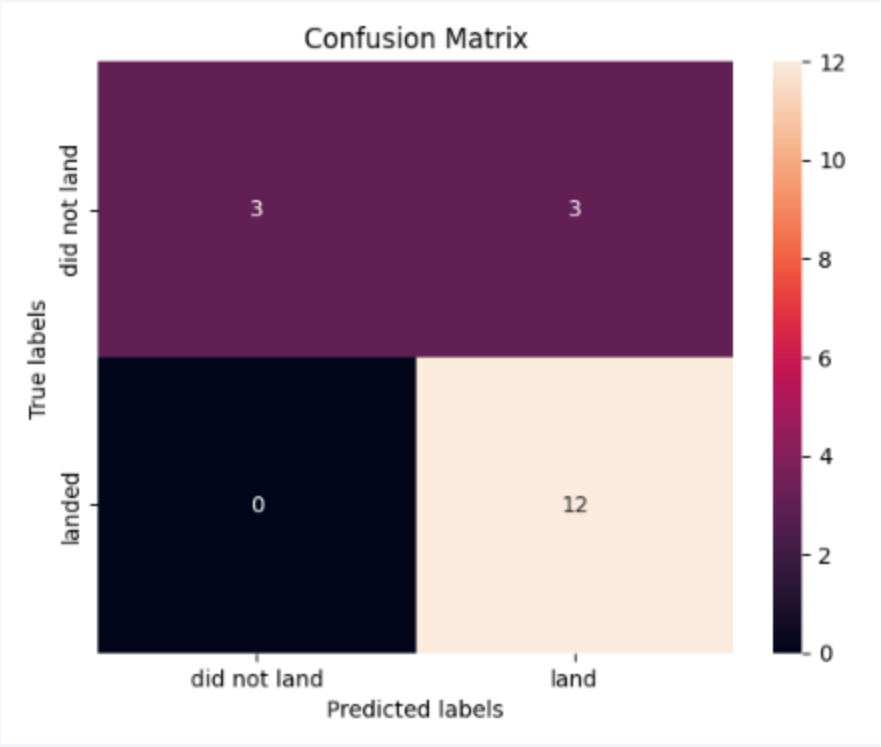
Confusion Matrix

All four Logistic Regression, SVM, Decision Tree and KNN have the same confusion matrix.

To understand confusion, matrix, follow this link:

https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5

```
Accuracy for logistic regression using the method score:
0.8333333333333334
```



Conclusions

- Decision Tree is the winner in the best Algorithm
- SpaceX launch success rates has been increasing since 2013 and kept increasing till 2020
- These orbits have high success rates:
  - ES-L1
  - GEO
  - HEO
  - SSO
- KSC LC-39A has the most success rate.
- Competitors may find difficult to bid against SPACEX unless very extraordinary efforts are made.

Appendix

- https://www.youtube.com/watch?v=8tNm3OjlVkc Basics of Plotly Dash
- https://www.youtube.com/watch?v=GW0cNAnngFk&t=734s Some Idea about Folium in Hindi
- PyCharm to try all this at local computer, It takes a lot of setup.
- My GIT HUB Link: https://github.com/jaindynike/IBMDSCapstoneProject/tree/main

Thank you!