Winning Space Race with Data Science

Krishankant Sharma
03.05.2022

Outline

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

Executive Summary

- Summary of methodologies
  - Collecting the data using web scraping and SpaceX API
  - Data Wrangling and Exploratory Data Analysis (EDA) to find pattern and determine label for training supervision models,
  - Visual Analytics include using folium and Dashboard using Ploty Dash
  - Prediction using Machine Learning.
- Summary of all results
  - Collecting data to showcase valuable result and Insights is possible using public website even like Wikipedia
  - Exploring the data and visualizing technique helps to find key insights and predict the success of launching
  - Machine Learning model showed the best model to preict the best outcome by chosing which characterstics are important.

Introduction

- Objective of this project is to evaluate the price of each launch. Gathering information about Space X and create dashboard and all evaluation require for company Space Y to compete with Space X.

- Problems want to find answers
  - The best way to find estimated cost for launches by predicting successful landing of the first stage of rockets
  - Find the best place to make launch.

Section 1

Methodology

Methodology

Executive Summary

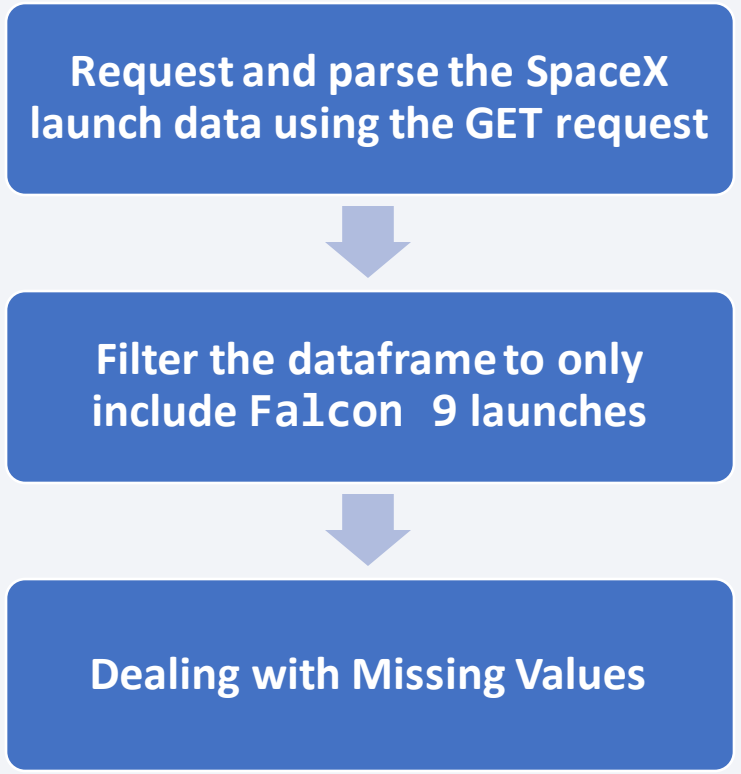
- Data collection methodology:
  - Data is collected from 2 sources:
    - Space X API https://api.spacexdata.com/v4/
    - Web scrapping Wikipedia source https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Perform data wrangling
  - by defining auxiliary function created a landing outcome label from Outcome Column.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - tune, evaluate classification models splitting the data into test and train,.

Data Collection

- Description how data sets were collected.
  - Data is collected from 2 sources:
    - Space X API https://api.spacexdata.com/v4/
    - Web scraping Wikipedia source https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches

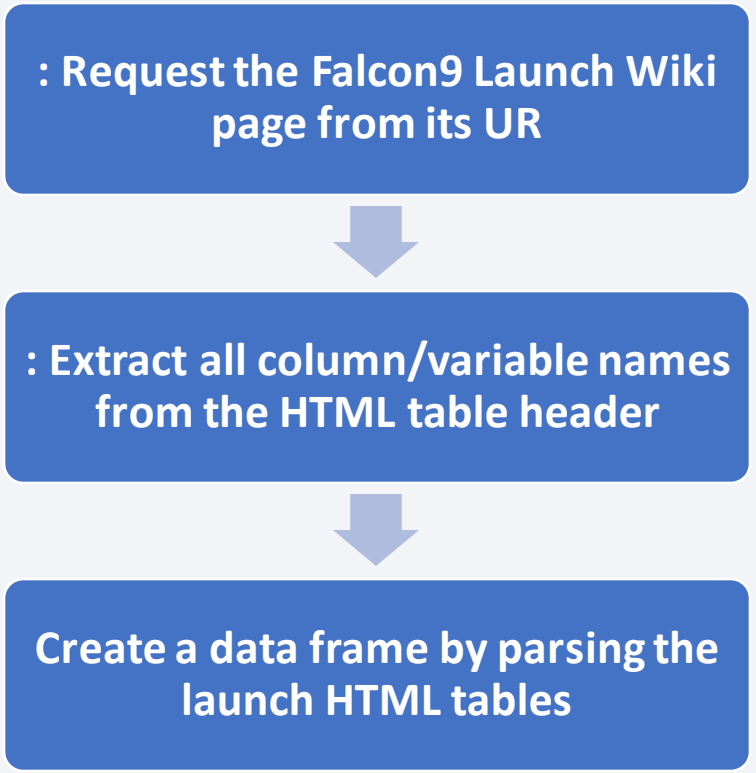
Data Collection – SpaceX API

- Request rocket launch data from SpaceX API. Filter dataframe only for falcon 9. Missing value are replaced with mean value.
- GitHub URL [Applied-datascience-Capstone-IBM/jupyter-labs-spacex-data-collection-api (1).ipynb at main · krishankant1996/Applied-datascience-Capstone-IBM (github.com)]



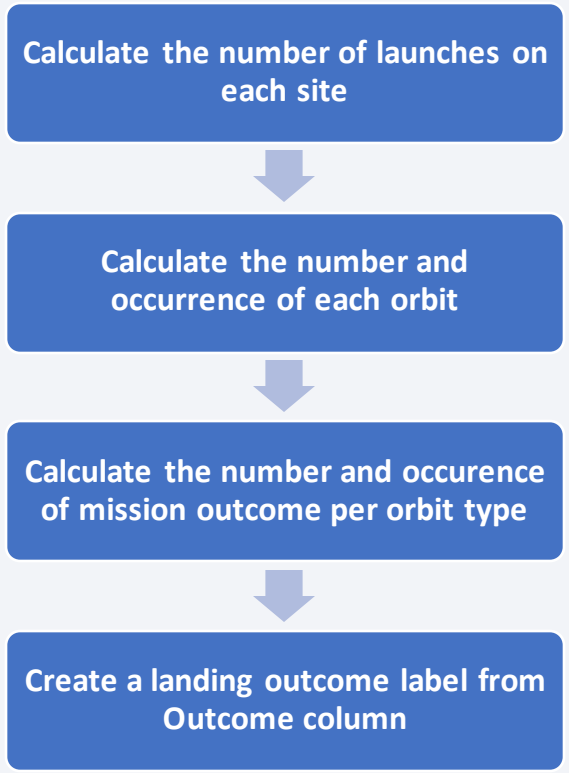
Data Collection - Scraping

- "HTTP GET" method to request the falcon 9 Launch for source
- Using "BeautifulSoup" extract all column and variable name for HTML table Header.
- Extract the column name using an empty dictionary and convert it into pandas dataframe later on.
- GitHub URLApplied-datascience-Capstone-IBM/jupyter-labs-webscraping (1) (1).ipynb at main · krishankant1996/Applied-datascience-Capstone-IBM (github.com)



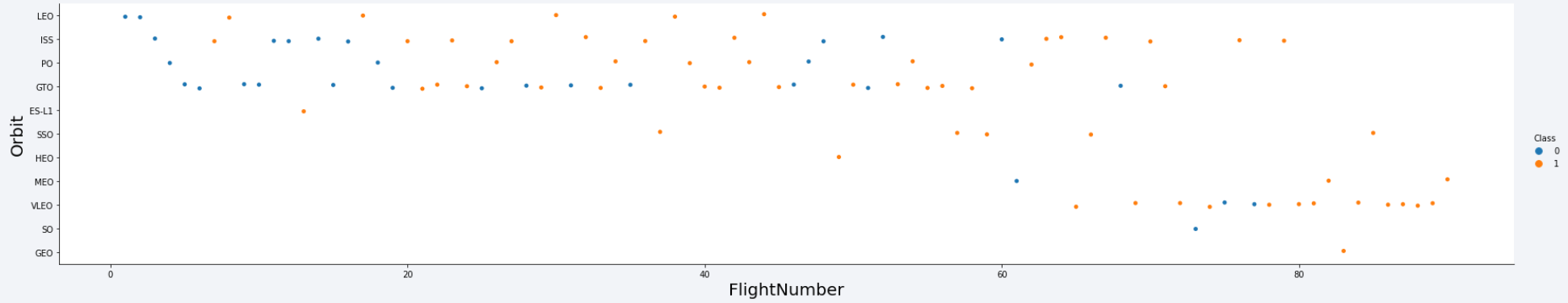
Data Wrangling

- Use "value _counts() on column Launch site to determine the number of launches on each site also same function for column Orbit
- Successful and Unsuccessful outcome count using for loop.
- create a list where the element is zero if the corresponding row in `Outcome` is in the set `bad outcome`; otherwise, it's one. Then assign it to the variable `landing_class`
- Use df.to_csv("dataset_part_2.csv", index=False) to save csv
- GitHub URL [Applied-datascience-Capstone-IBM/labs-jupyter-spacex-Data wrangling (1) (1).ipynb at main · krishankant1996/Applied-datascience-Capstone-IBM (github.com)]



EDA with Data Visualization

- Seabron scatter plot is used to visulaize compression between diffent intity like reltionsip between Flight number and Luanch Site, relationship between Payload and Lauch site, relation between flight number and Orbit Type, between paylaod and Orbit type. A line chart also use to visualie the laouch success yearly trend



Applied-datascience-Capstone-IBM/jupyter-labs-eda-dataviz(2).ipynb at main · krishankant1996/Applied-datascience-Capstone-IBM (github.com)

EDA with SQL

- Names of the unique launch sites in the space mission.
- Top 5 launch site whose name begin with the string "CCA"
- Total Payload mass carried by booster launched by NASA(CRS)
- Average Payload mass carried by booster version F9 v 1.1
- Date when the first successful landing outcome in ground pad was achieved
- Names of the booster which have success in drone ship and have payload mass between 4000 and 6000 kgs
- Total number of successful and failure mission outcomes
- Names of booster versions which have carried the maximum payload mass
- Failed landing outcome in drone ship, their booster version and launch site names for year 2015
- Rank of the count of landing outcomes
- GitHub URL [Applied-datascience-Capstone-IBM/jupyter-labs-eda-sql-coursera (1).ipynb at main · krishankant1996/Applied-datascience-Capstone-IBM (github.com)]

Build an Interactive Map with Folium

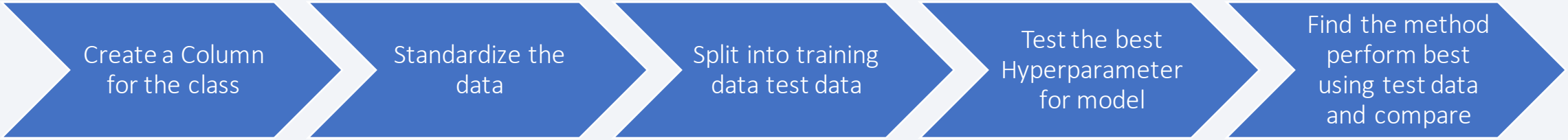
- markers, circles, lines, and marker cluster were used in folium map
- Marker is used to indicate the launch sites.
- Circle are used to specify specific area like NASA johnson space center
- Marker cluster is used to group event like failure and success on launch site
- Lines are used to visualize and say distance between two coordinate.
- GitHub URL Applied-datascience-Capstone-IBM/lab_jupyter_launch_site_location (1).ipynb at main · krishankant1996/Applied-datascience-Capstone-IBM (github.com)

Build a Dashboard with Plotly Dash

- Graphs and plots that are used in Dashboard to visualize the data are
  - Percentage of launches by sties
  - Payload Range
- This allow us to analyze the relation between payload and launch sites in a Intractive way.

Predictive Analysis (Classification)

- By using of four classification model comparison logistic regression, support vector machine (SVM), decision tree and K-nearest Neighbors best model is chosen from evaluation of predicted result from each using train test split

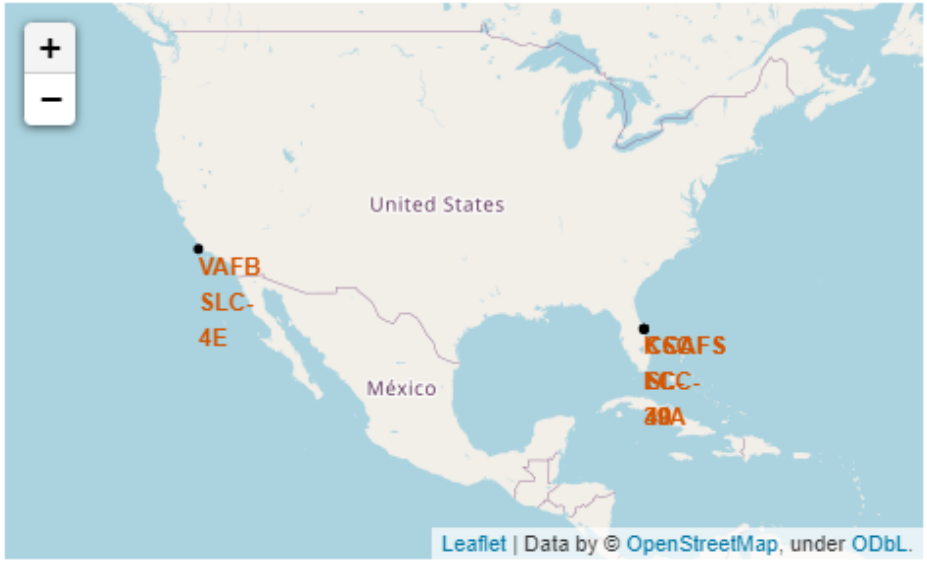


Results

- Exploratory data analysis results
  - Space X have CCAF SLC 40 has most launches and occurrence of obit in GTO most
  - First successful landing outcome happen in 2015
  - F9 v1.1 B1012 and F9 v1.1 B1015 were 2 booster failed landing in drone ship
  - Average payload of F9 v1.1 booster is 2928kgs
  - The number of landing outcome get better year by year.

Results

- Interactive analytics demo in screenshots

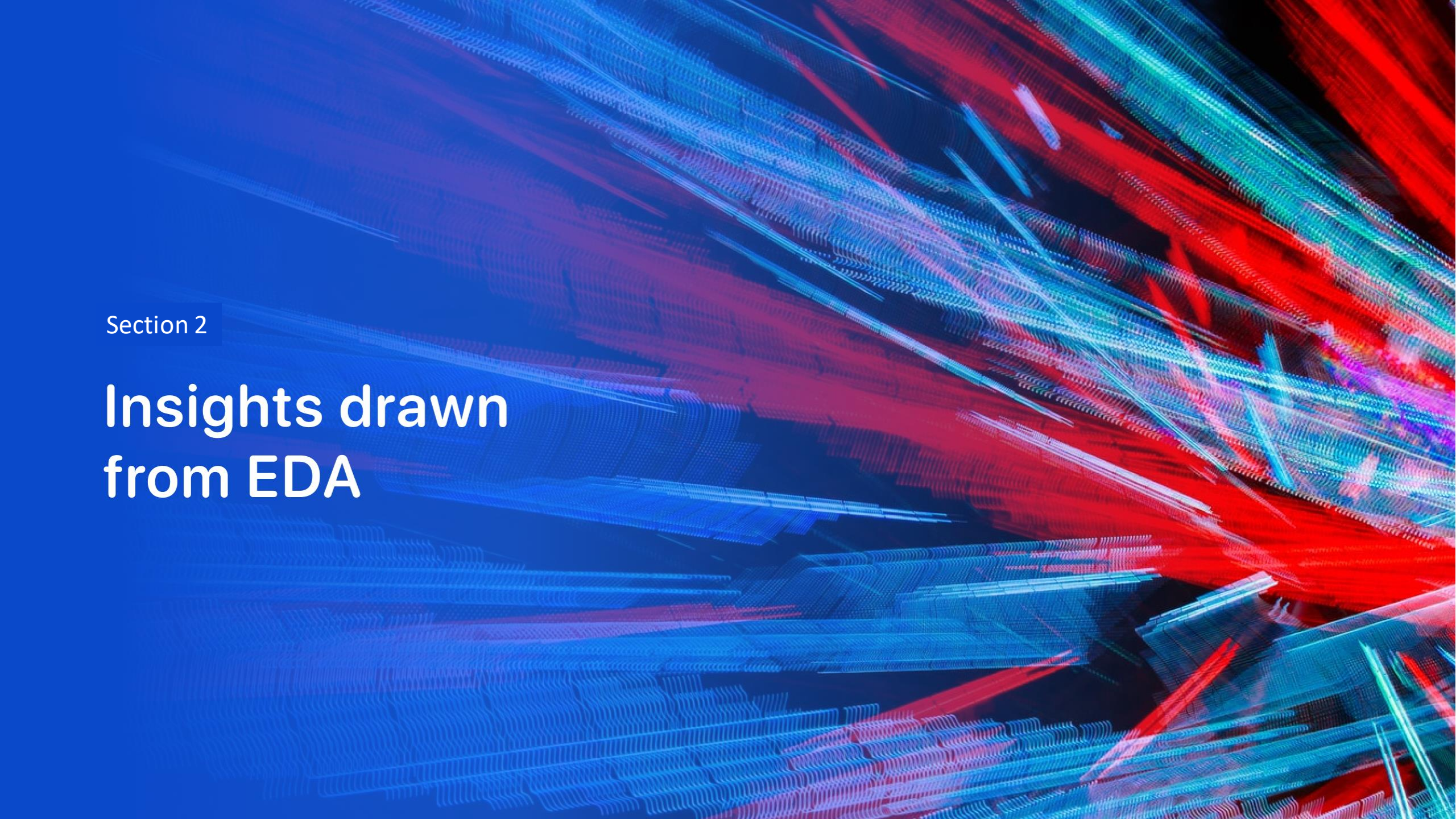


Results

- Predictive analysis results shows that the best model is Decision Tee Classifier. It predict the accuracy of 87% and accuracy for test data over 94%.

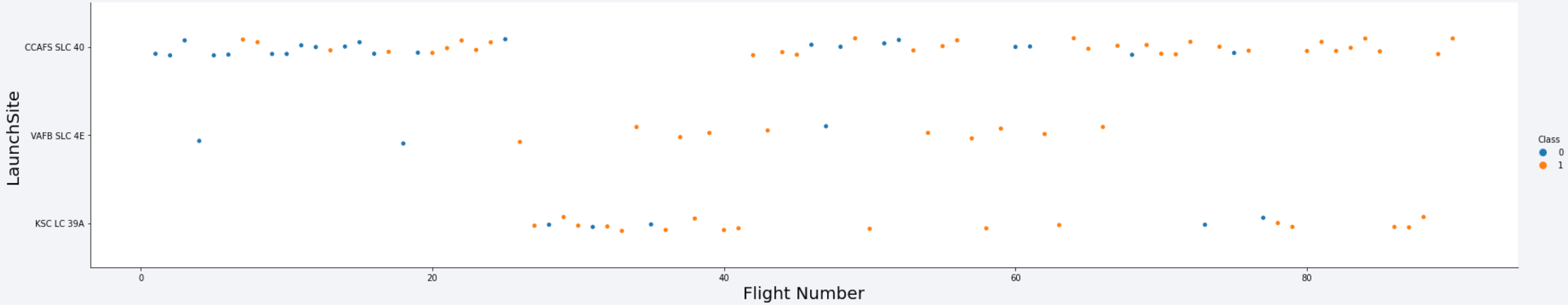
Section 2

Insights drawn from EDA



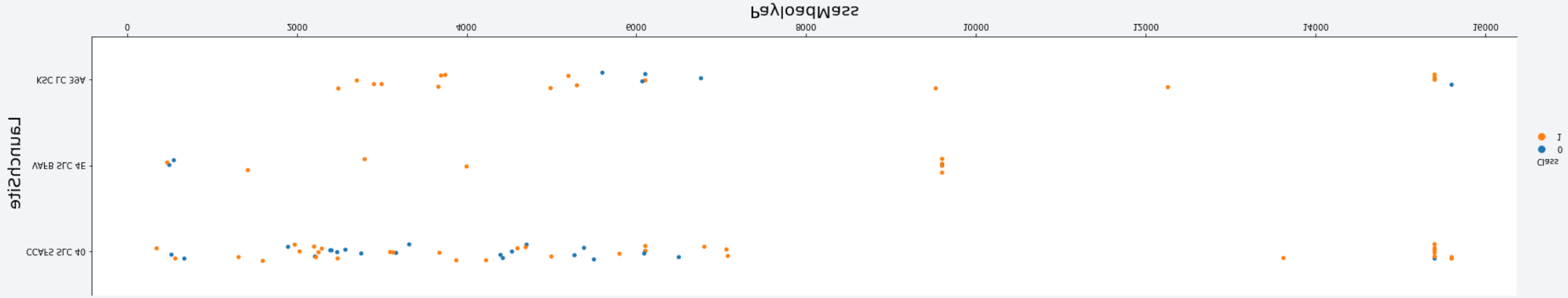
Flight Number vs. Launch Site

- Shown graph show best site for lauch is CCAF5 SLC 40, after which VAFB SLC 4E and then KSC LC 39A.
- Success rate increase over tiem



Payload vs. Launch Site

- Paylaod over 9000kg have excellent success rate



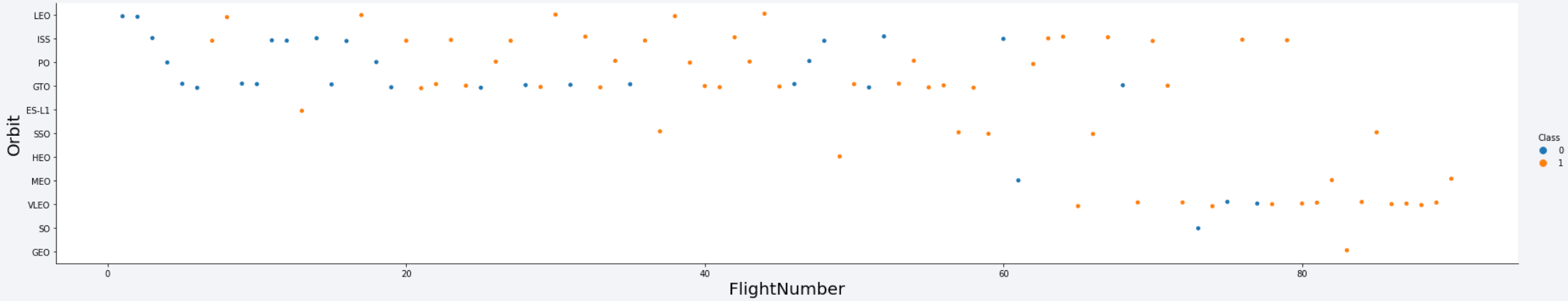
Success Rate vs. Orbit Type

- The biggest success rates hppens to orbits:
  - ES-L1
  - GEo
  - HEO and
  - SSO

```
Orbit
ES-L1    1.000000
GEO      1.000000
GTO      0.518519
HEO      1.000000
ISS      0.619048
LEO      0.714286
MEO      0.666667
PO       0.666667
SO       0.000000
SSO      1.000000
VLEO     0.857143
Name: Class, dtype: float64
```

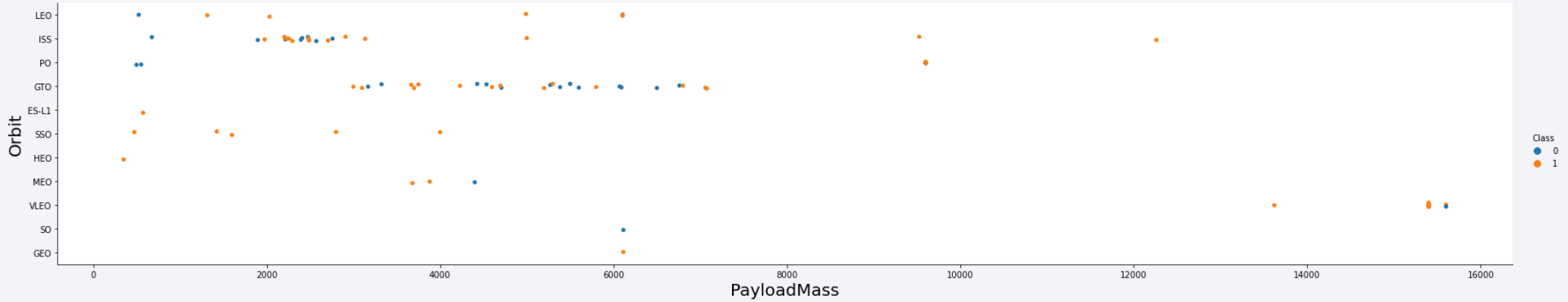
Flight Number vs. Orbit Type

- Succeess rate improve over time.
- VELO orbit is used more in recent time.



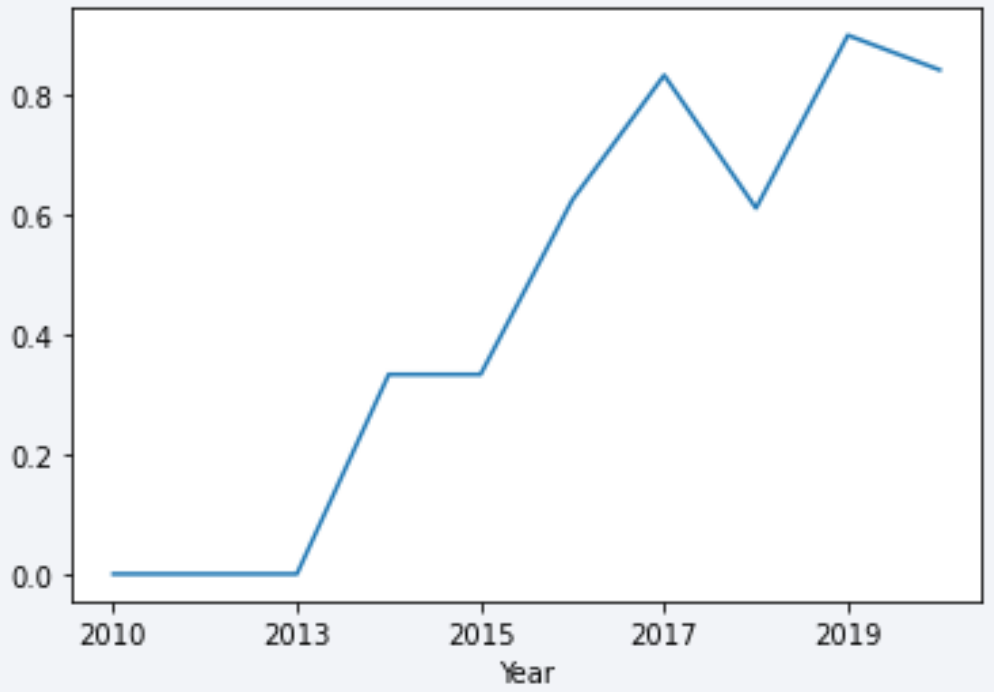
Payload vs. Orbit Type

- There is less launches to the orbits SO and GEO
- ISS orbit has wide range of payload and good success rate



Launch Success Yearly Trend

- Success rate start increasing after 2013
- First three year were R&D and development stages



All Launch Site Names

- There is four launch site
- Selecting Lauch site column with unique occurrences we get this value.

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

Launch Site Names Begin with 'CCA'

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

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

- Calculate the total payload carried by boosters from NASA
- using magic % CRS is contained payload were extracted.

total_payload
111268

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Using avg and selecting F9 v1.1 using where we get following query

```
sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD_f9 FROM SPAC
EXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
```

```
 * ibm_db_sa://mdl23702:***@764264db-9824-4b7c-82df-40d1b138
97c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bl
udb
Done.
```

avg_payload_f9
2928

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Used min to get value of date

```
sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL WHERE
LANDING__OUTCOME = 'Success (ground pad)';
```

```
 * ibm_db_sa://md123702:***@764264db-9824-4b7c-82df-40d1b138
97c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bl
udb
Done.
```

first_success_gp
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
- Using distinct booster version according to filter we get result as follow.

```
sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYL
OAD_MASS__KG_ BETWEEN 4000 AND 6000 AND LANDING__OUTCOME = 'S
uccess (drone ship)';
```

```
 * ibm_db_sa://mdl23702:***@764264db-9824-4b7c-82df-40d1b138
97c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bl
udb
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Grouping mission outcome and count each group item give following result.

```
sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL GR
OUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;
```

```
 * ibm_db_sa://mdl23702:***@764264db-9824-4b7c-82df-40d1b138
97c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bl
udb
Done.
```

mission_outcome	qty
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- List the names of the booster 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) ORDER BY BOOSTER
_VERSION;
```

```
 * ibm_db_sa://mdl23702:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io
90l08kqb1od8lcg.databases.appdomain.cloud:32536/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

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 FROM SPACEXTBL WHERE LANDING__
OUTCOME = 'Failure (drone ship)' AND DATE_PART('YEAR', DATE) = 2015;
```

```
 * ibm_db_sa://md123702:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io
90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
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
- No attempt must be taken in account.

```
sql SELECT LANDING__OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL WHERE DATE
BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING__OUTCOME ORDER
BY QTY DESC;
```

```
 * ibm_db_sa://mdl23702:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io
90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

landing__outcome	qty
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


Section 3
Launch Sites
Proximities Analysis

Launch Sites

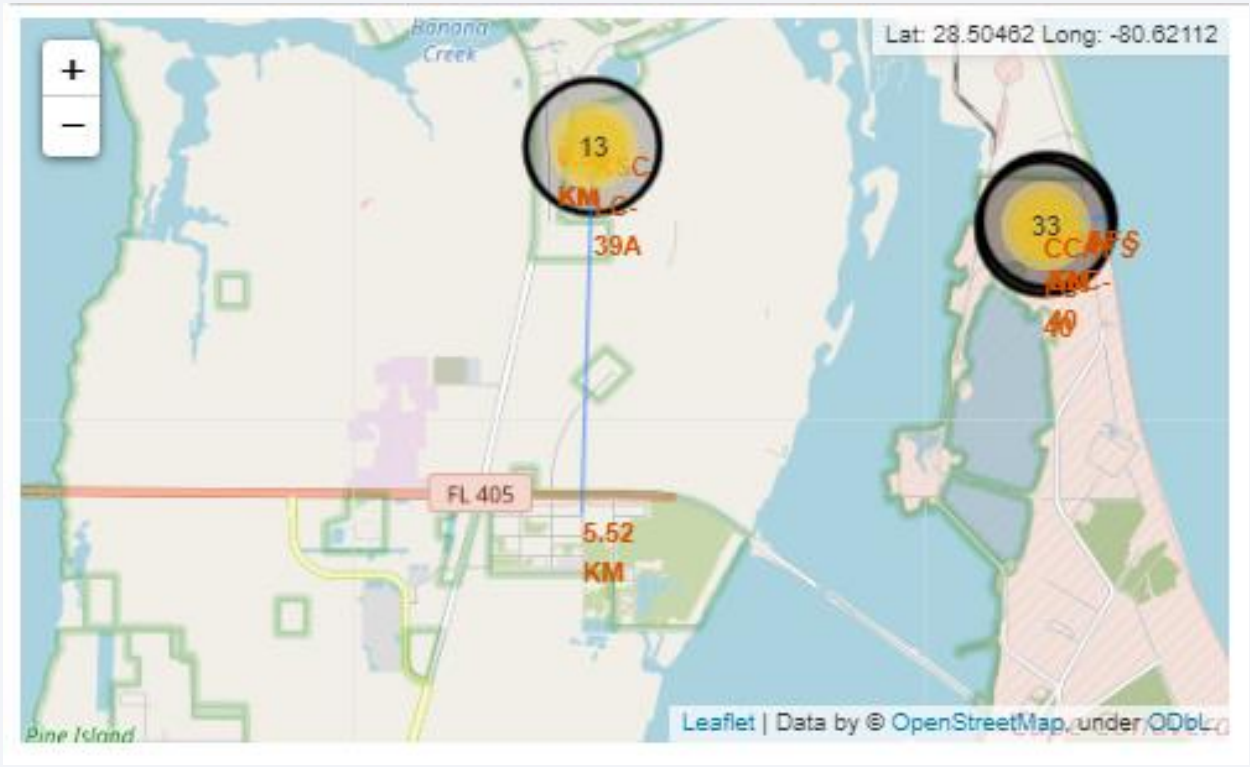
- Explore the generated folium map and make a proper screenshot to include all launch sites' location markers on a global map
- All launch sites are near sea sore

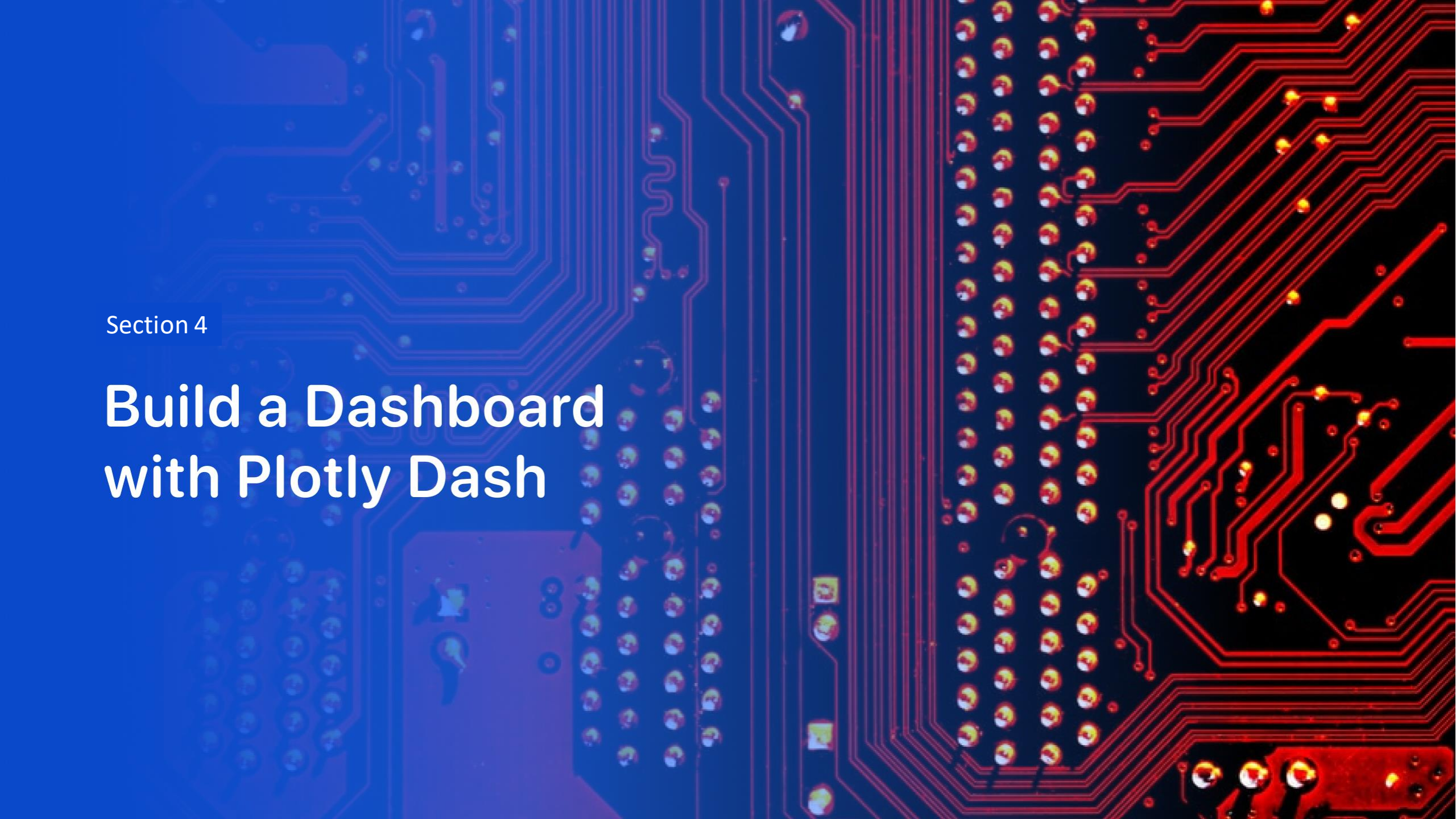
Launch Outcomes by Sites

- Launch side KSC LC-39A
- Green market indicate success and red indicate failure(click on cluster)

Distance from Launch Sites

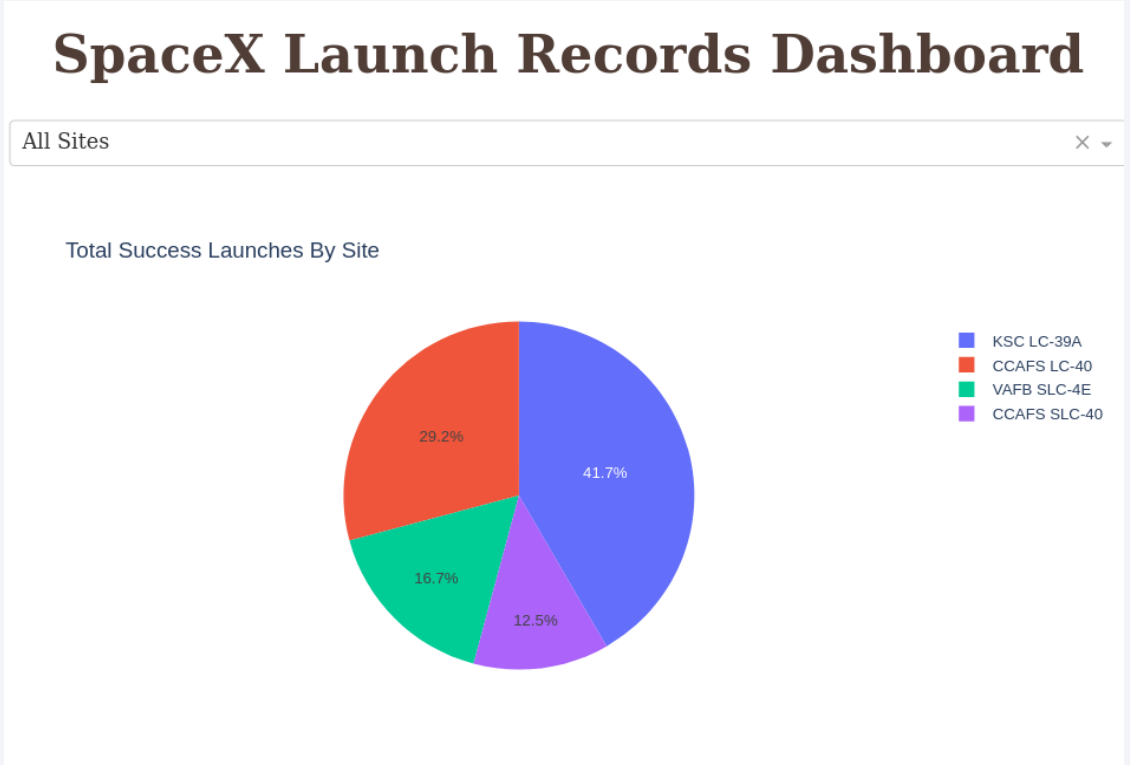
- Shows that KSC LC-39A has good logistics aspects, being near railroad aroad.




Section 4
Build a Dashboard with Plotly Dash

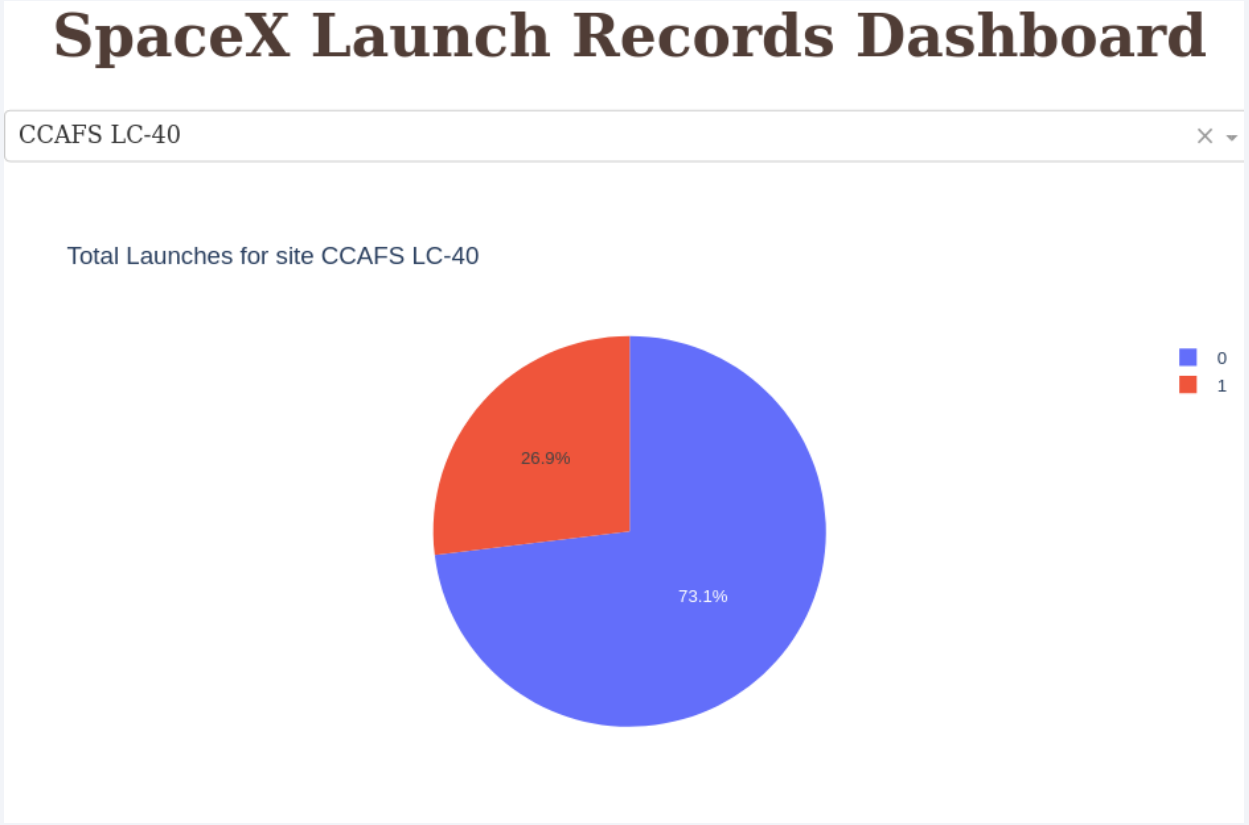
Pie chart Dashboard

- Show the KSC IC-39A is most used launch site second CCAFS LC-40



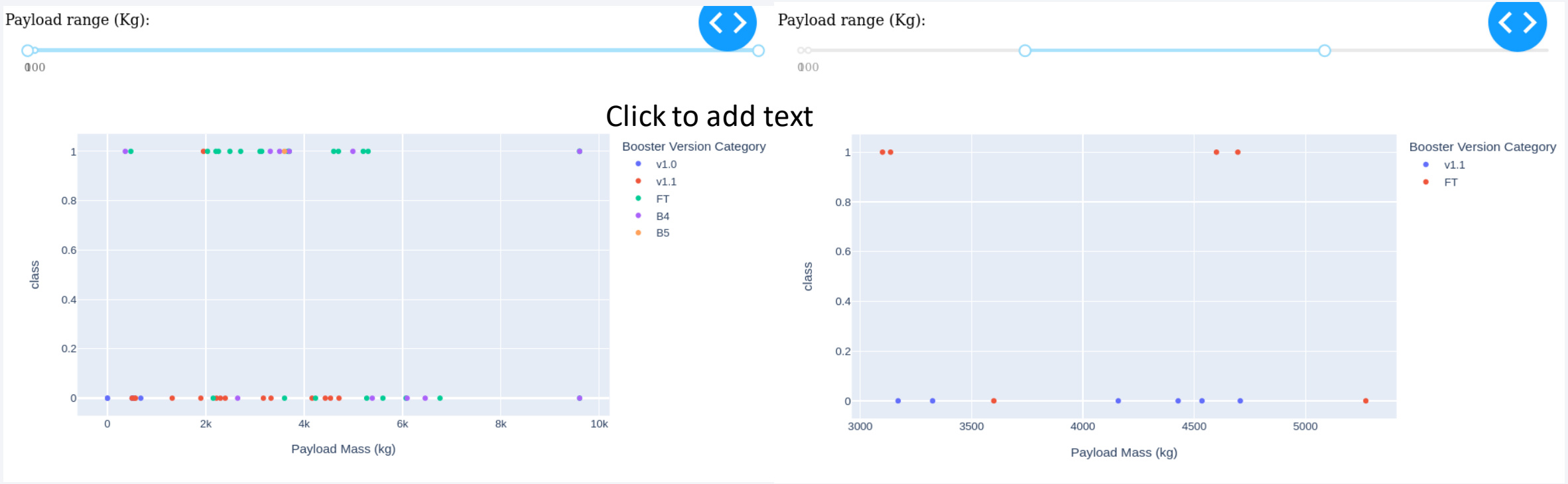
Pie Chart Dashboard (Success Percentile by Launch Site)

- 76.9% of success percentile for KSC LC-39A



Dashboard Launch Outcome scatter Plot

- Payload under 6000 and FT booster are most successful combo.



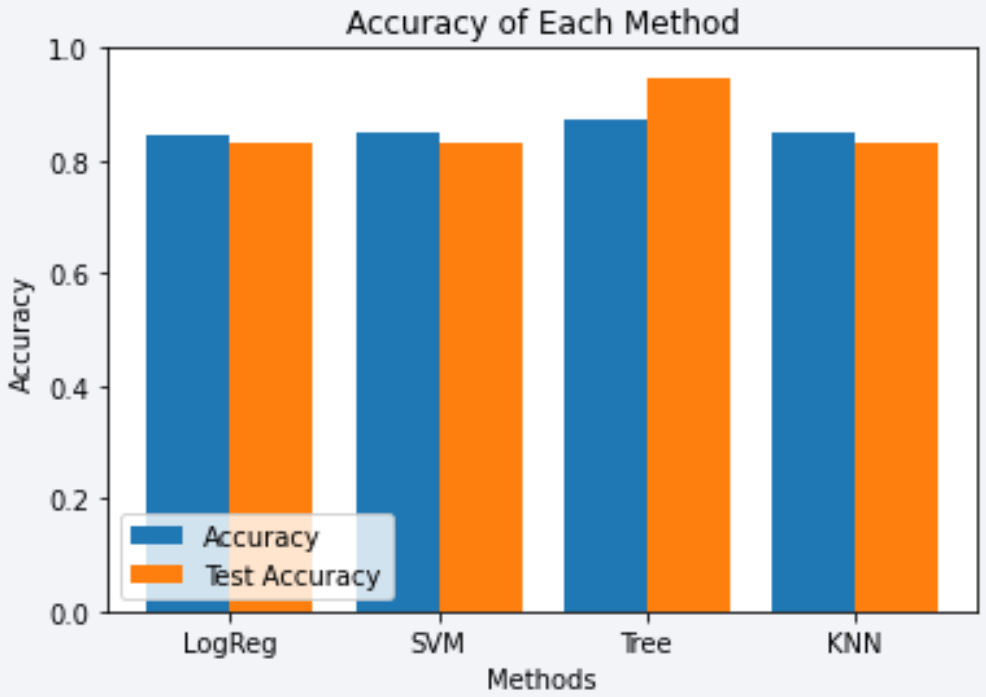
Click to add text

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Bat graph of 4 model used
- Decision tree Classifier has accuracy of 87% which is highest among all.



Confusion Matrix

- Confusion matrix of Decision Tree Classifier proves its accuracy by showing the big numbers of true positive and true negative as compared to the false ones.

Conclusions

- The best launch site is KSC LC-39A
- Lauch location should be chosen wisely
- Lauchner above 7000kg are less risky
- Successful landing improve over time. 2013 is year form which success rate increase
- Decision tree classifier can be used to predict the successful landing outcome to increase landing profit

Appendix

- Choosing the best value of KNN and Decision Tree Classifier is depend of value of k and depth of tree cluster. Choosing a wise value of K and Depth of tree is important
- To get the same outcome please choose Radom. Seed same as given or every time you will get different value.
- Folium is not showing results on Github

Thank you!