IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Tan Yi Rong>
<2022/2/25>

Outline

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

Executive Summary

- Summary of methodologies

  This project predict the landing success rate of the first stage of the SpaceX Falcon 9 rocket with data science methodology involving data collection, data wrangling, exploratory data analysis, interactive visual analytics, and predictive analysis.

- Summary of all results

  - Out of the 90 landings of Falcon 9 rocket between the year of 2010 and 2020, the overall success rate is ≈ 66.67%.
  - We observed that the success rates started to increase since 2013.
  - There are various relationships between the features including flight number, launch site, payload, orbit, etc.
  - By applying machine learning method, we built classification models for the prediction and all classifiers have the same prediction accuracy of ≈ 83.33%.

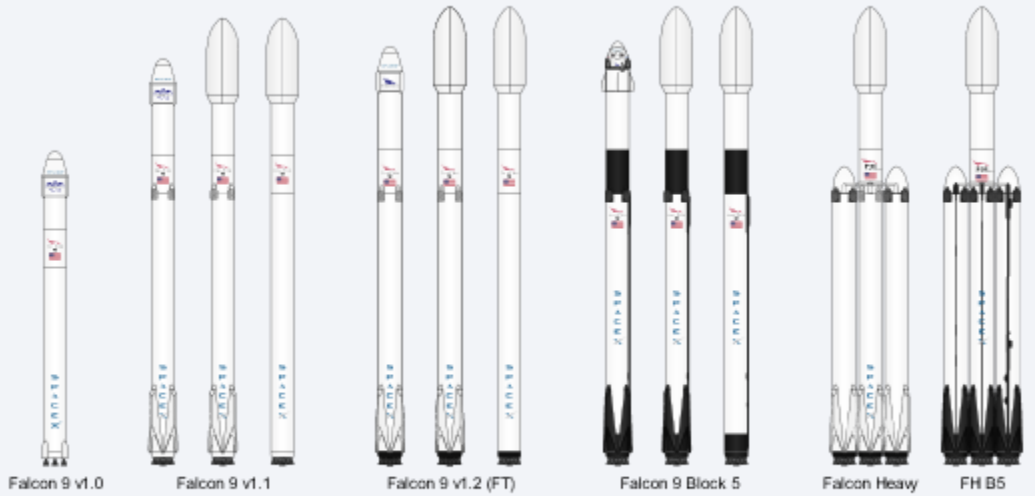
Introduction

- Project background and context

  Space X 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 Space X can reuse the first stage. Therefore, if we can determine if the first stage will land successfully, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch.

- Problems you want to find answers

  Whether or not the Falcon 9 first stage will land successfully?



Section 1

Methodology

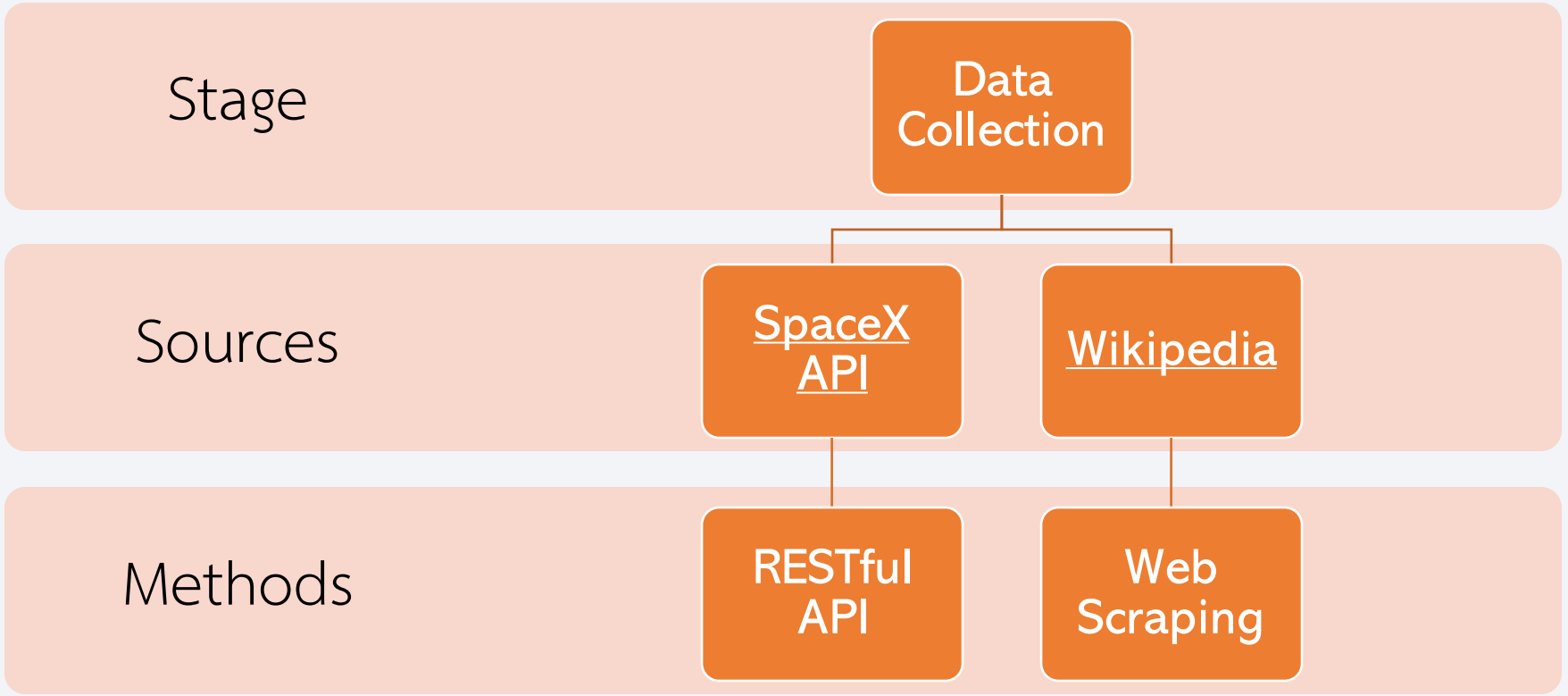
Methodology

Executive Summary

- Data collection methodology:
  - RESTful API and Web Scraping
- Perform data wrangling
  - Exploratory Data Analysis and Training Labels.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build, tune and evaluate classification models using Scikit-learn involving Train-Test Split and Grid Search.

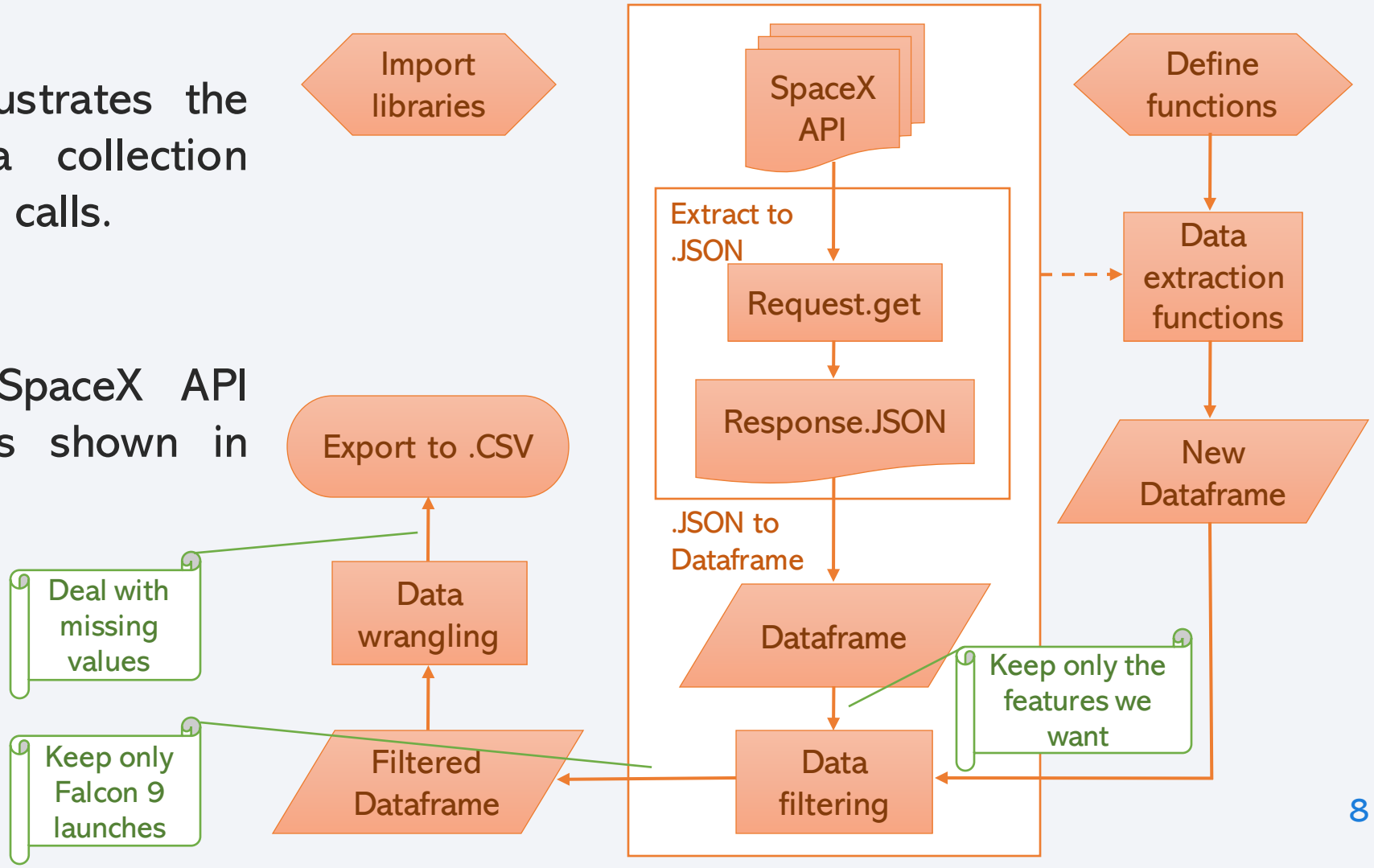
Data Collection

- Data sets were collected from two sources:



Data Collection – SpaceX API

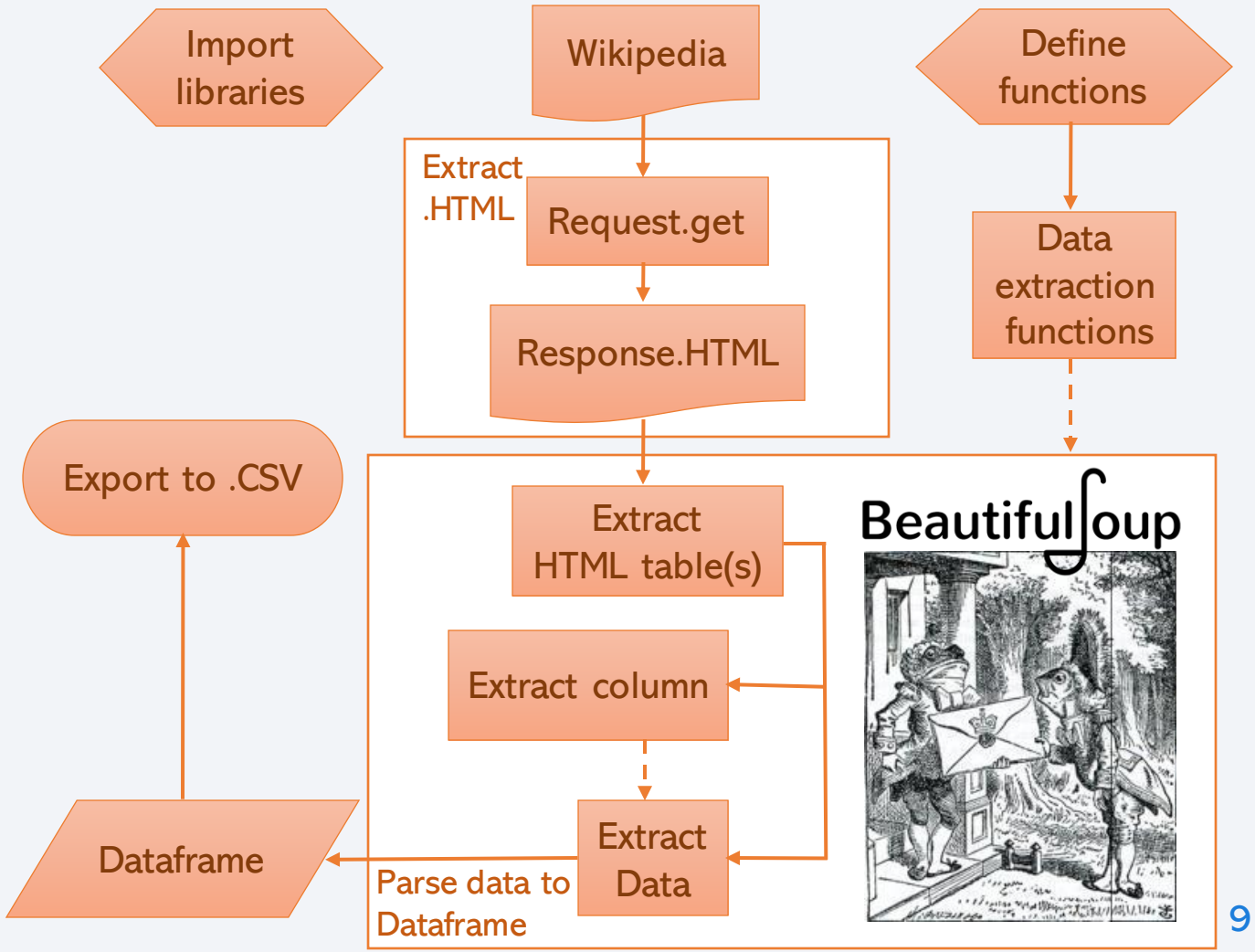
- The flowchart illustrates the process of data collection with SpaceX REST calls.
- The completed SpaceX API calls notebook is shown in this LINK.



Data Collection - Scraping

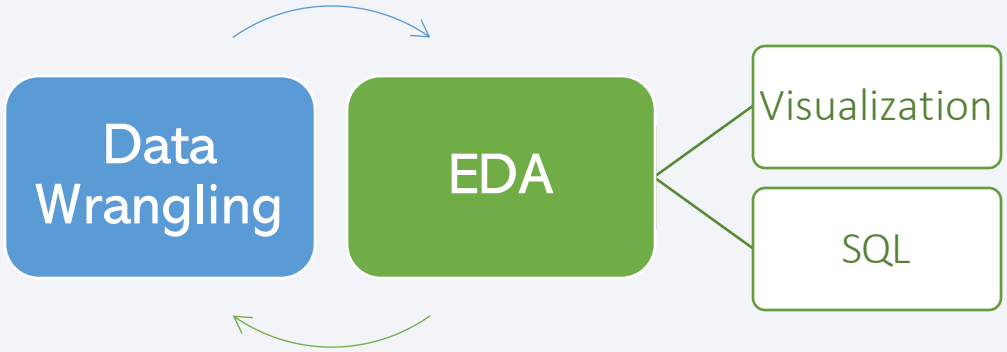
- Web scraping process is illustrated in the flowchart.

- Click HERE for the completed web scraping notebook.

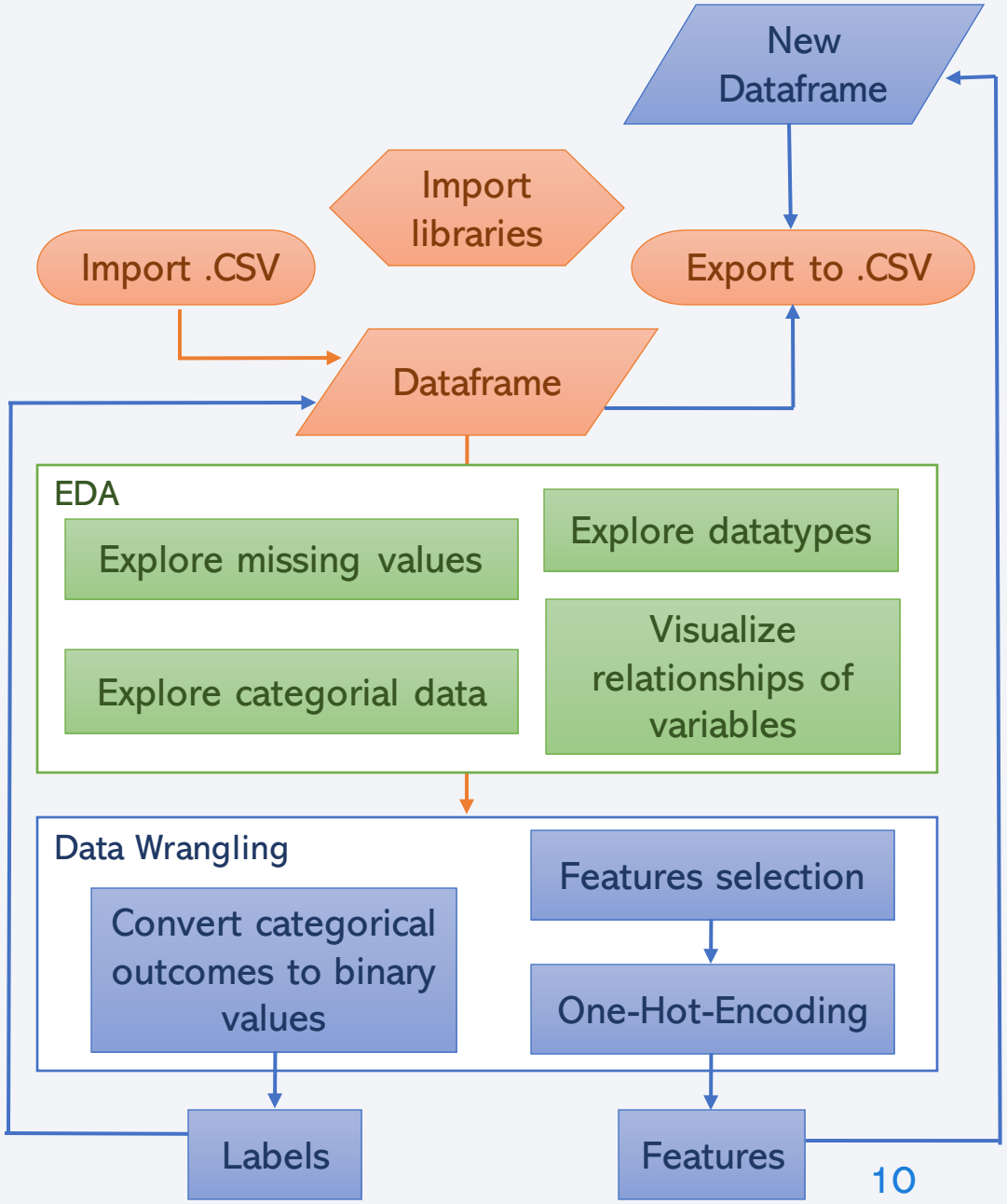


Data Wrangling

- Exploratory Data Analysis (EDA) was performed to find some patterns in the data and determine the label for training supervised models and prepare data feature engineering.



- Click this LINK for the completed data wrangling related notebooks.



EDA with Data Visualization

- **Scatterplot charts** (`seaborn.catplot`) were plotted to see how relationships between variables will affect if the landings success or not. **Bar chart** was plotted to find the relationship between success rate of each variable. **Line chart** was plotted to determine the yearly trend of success rate.

Scatterplot	Bar chart	Line chart
• FlightNumber vs. PayloadMass • FlightNumber vs LaunchSite • Payload and Launch Site • FlightNumber and Orbit type • Payload and Orbit type	• Success rate of each orbit type	• Landing success yearly trend

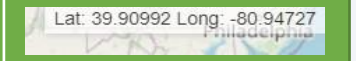
- The completed EDA with data visualization notebook is shown in this LINK.

EDA with SQL

- The following SQL queries were performed:
  - 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. Use a subquery.
  - List the failed landing_outcomes in drone ship, their booster versions, and launch site names for year 2015.
  - 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.
- HERE shows the completed EDA with SQL notebook.

Build an Interactive Map with Folium

- In the folium map, map objects such as `circles`, `markers` and `polylines` were created and added. `MarkerClusters` were also used to contain many markers having the same coordinate. A `MousePosition` was also added to get coordinate for a mouse over a point on the map.
- The purposes of those objects added on the map are stated as table below:

Objectives \ Map objects	Circle	Marker	MarkerCluster	Polyline	MousePosition
					
Task 1: Mark all launch sites on a map.	✓	✓			
Task 2: Mark the success/failed launches for each site.		✓	✓		
Task 3: Distances between a launch site to its proximities.		✓		✓	✓

- This NOTEBOOK displays the completed interactive map with Folium map.

Build a Dashboard with Plotly Dash

- The dashboard consists of two sections:

1

Firstly, a **dropdown menu** of the four launch sites was added and the selected launch site will render a **pie chart** visualizing the launch success rate (Success vs Failure).

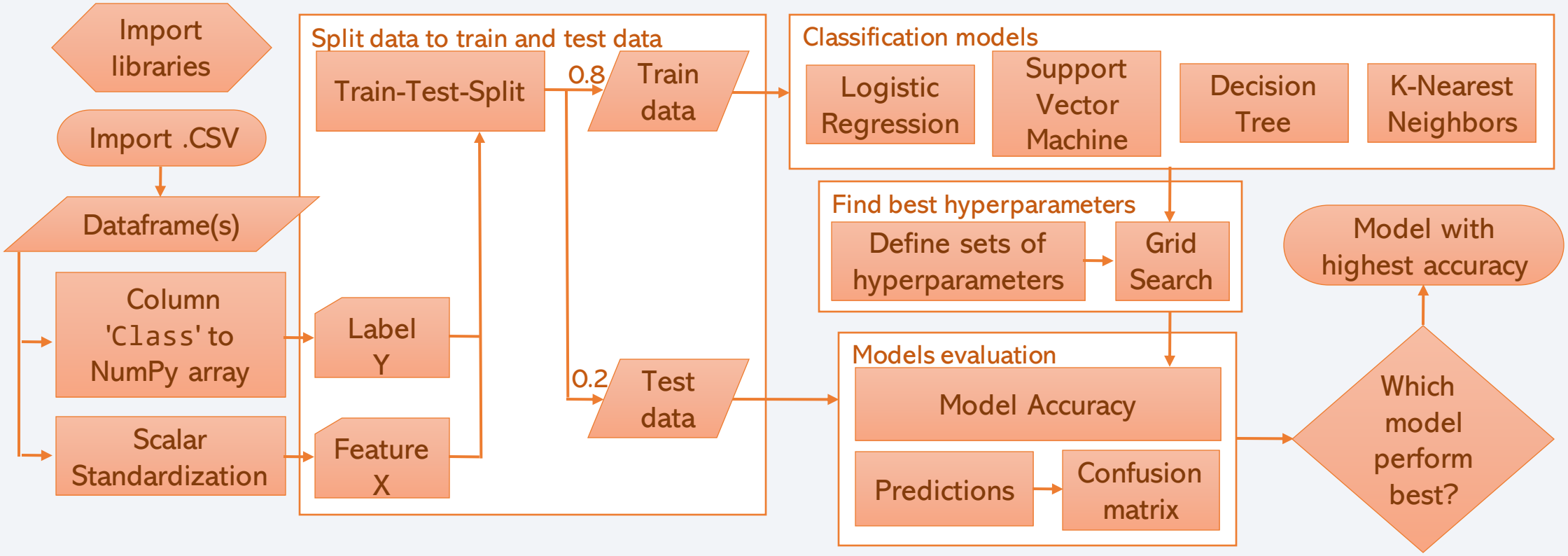
2

Then, a **range slider** was added to select a specific range of payload mass and a **scatterplot** will displays allowing us to observe how payload may be correlated with the landing outcomes for the selected launch site(s).

- The completed Plotly Dash lab was displayed HERE.

Predictive Analysis (Classification)

- The best performing classification model was built, evaluated, improved, and found as flowchart below:



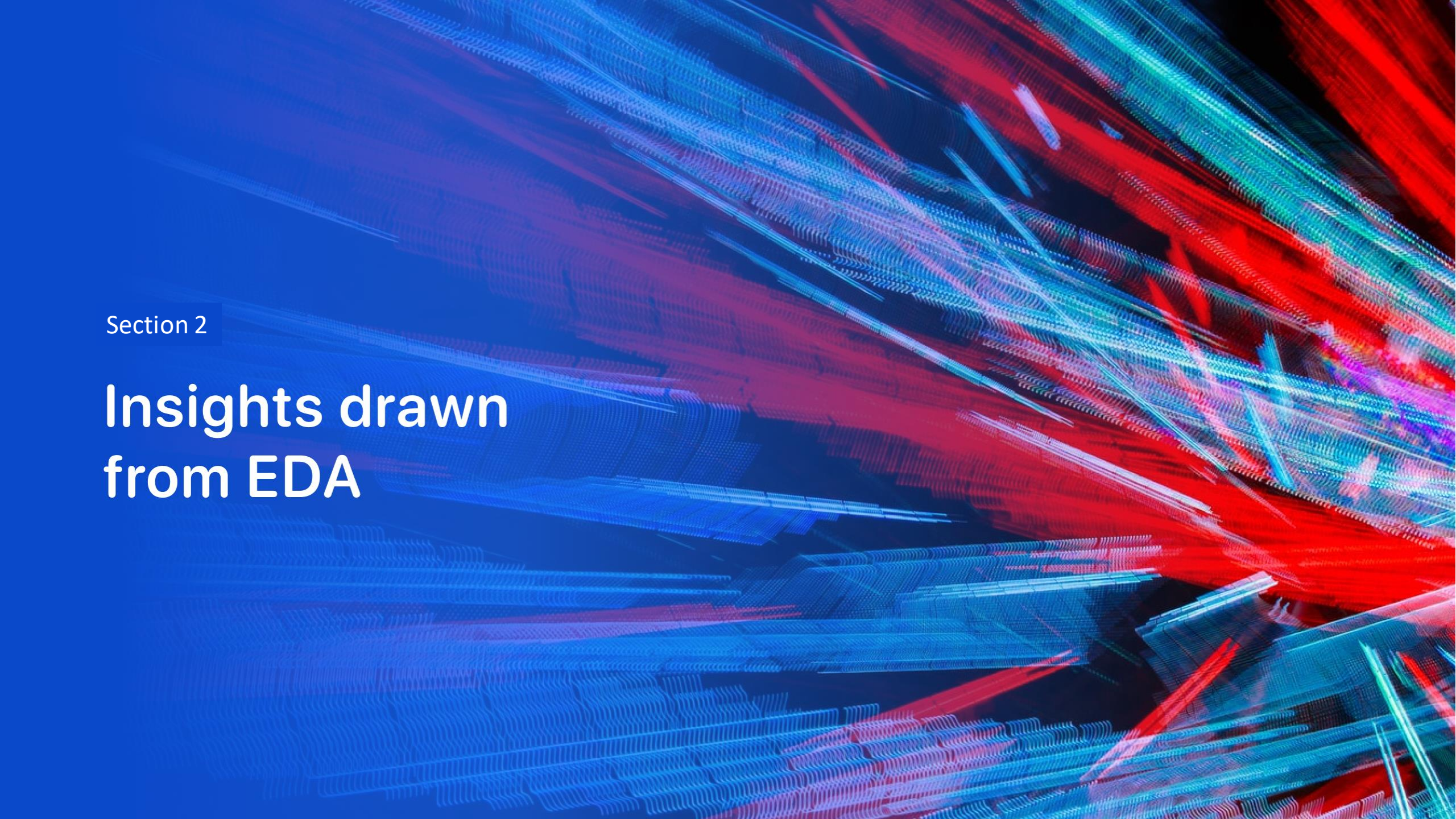
- This URL link you to the completed predictive analysis lab.

Results

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

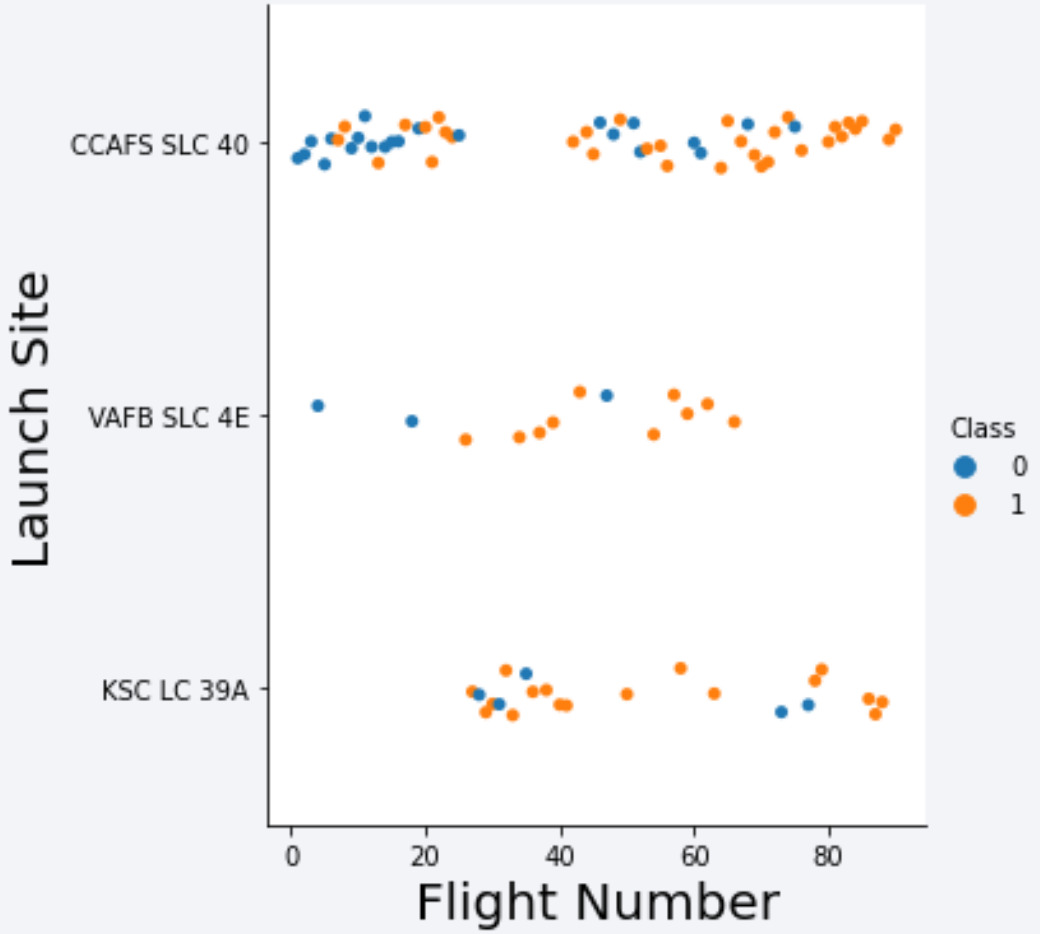
Section 2

Insights drawn from EDA

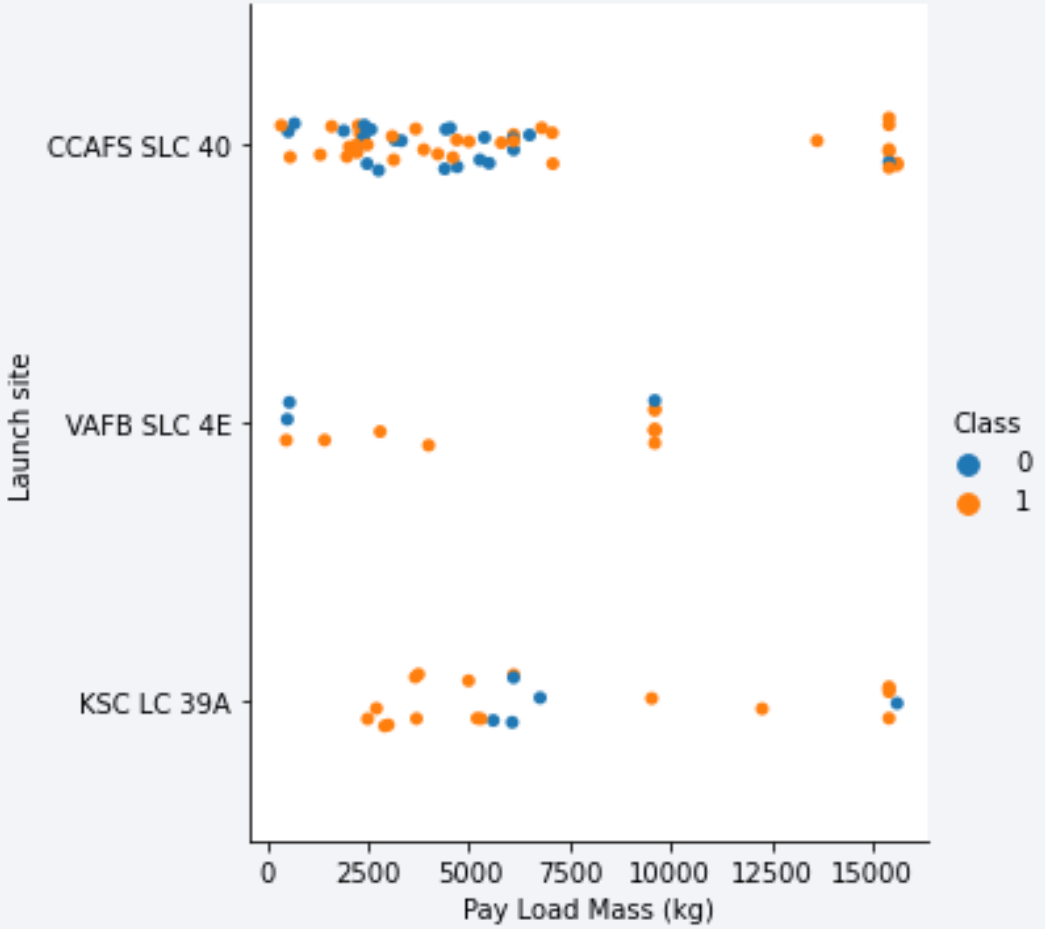


Flight Number vs. Launch Site

- Figure right shows a scatterplot of Flight Number vs. Launch Site.
- Overall, as the flight number increases, the first stage is more likely to land successfully, especially in launch site CCAFS SLC 40. We also observe that no rockets launched for VAFB SLC 4E after 70th flight and before 20th flight for KSC LC 39A.



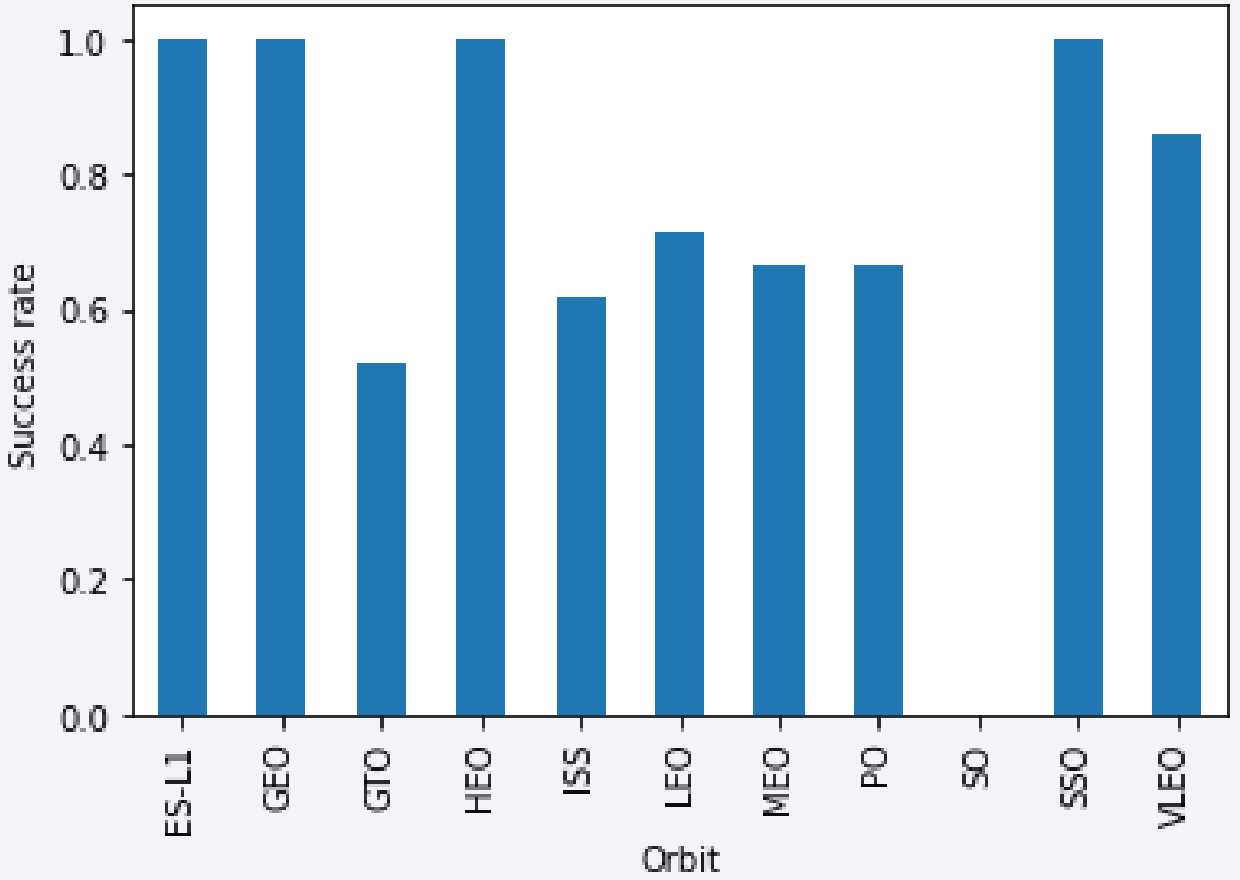
Payload vs. Launch Site



- Figure on left shows a scatterplot of Payload vs. Launch Site.
- For the VAFB SLC 4E launch site, there are no rockets launched for heavy payload mass (greater than 10000).

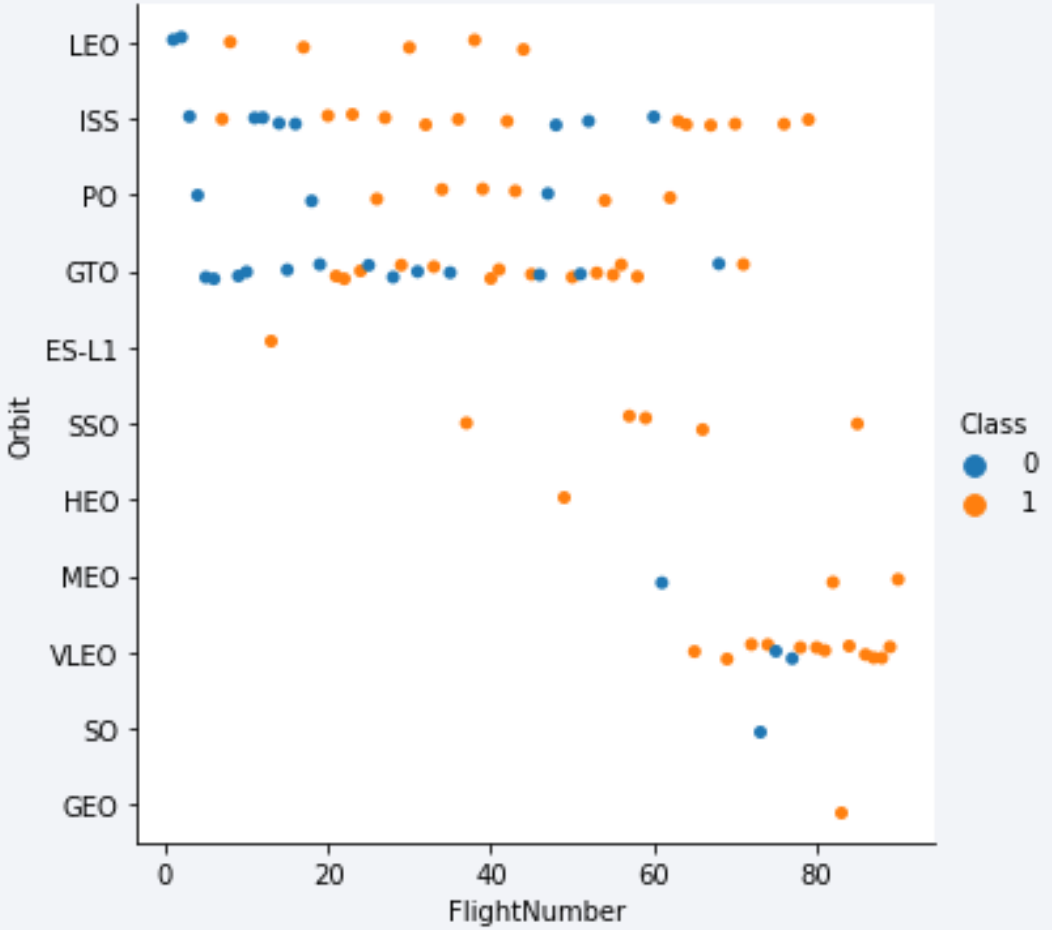
Success Rate vs. Orbit Type

- Bar chart on the right shows the success rate of each orbit type.
- We can observe that orbit ES-L1, GEO, HEO, and SSO have 100% success rate; but no rockets were successfully landing for orbit SO.

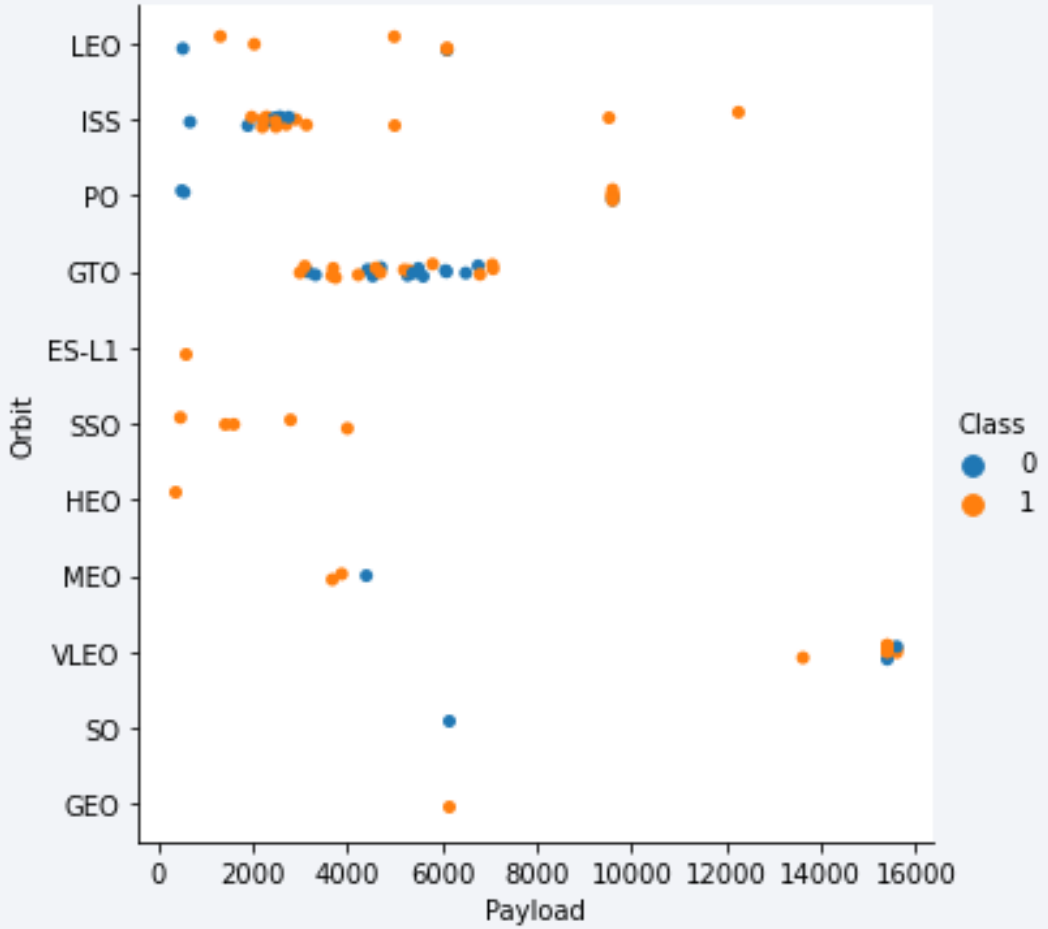


Flight Number vs. Orbit Type

- Scatter point on the right illustrates the 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.

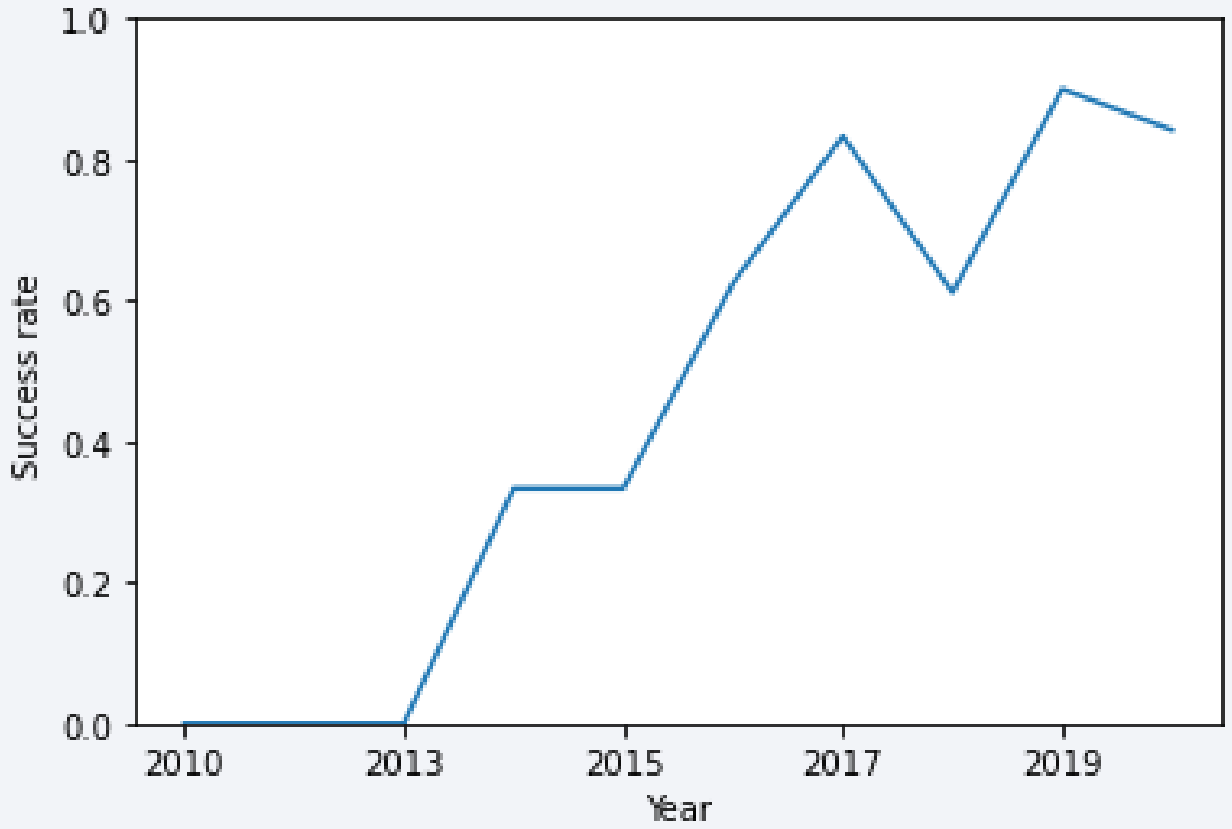


Payload vs. Orbit Type



- Figure on left illustrates a scatter point of payload vs. orbit type.
- With heavy payloads, the successful landing are more for Polar, LEO and ISS. However, we cannot distinguish this well for GTO as both positive landing rate and negative landing are both occurred.

Launch Success Yearly Trend



- Line chart of yearly average success rate is as shown on the figure left.
- In general, the sucess rate since 2013 kept increasing till 2020.

All Launch Site Names

- Find the names of the unique launch sites

```
%sql SELECT UNIQUE(LAUNCH_SITE) FROM SPACEXDATASET
```

```
 * ibm_db_sa://nzc67634:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/bludb
Done.
```

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

- There are 4 launch sites for SpaceX Falcon 9 rocket within the past 10 years.

Launch Site Names Begin with 'CCA'

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

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

```
 * ibm_db_sa://nzc67634:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/bludb
Done.
```

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

- The query displays a list of records and information with launch site names begin with CCA

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

```
%sql SELECT SUM(payload_mass__kg_) AS total_payload_mass FROM SPACEXDATASET \
    WHERE customer = 'NASA (CRS)'
```

```
 * ibm_db_sa://nzc67634:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/bludb
Done.
```

total_payload_mass
45596

- The total payload mass carried by boosters from NASA is 45596 kg.

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 average_payload_mass FROM SPACEXDATASET \
    WHERE booster_version = 'F9 v1.1'
```

```
 * ibm_db_sa://nzc67634:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/bludb
Done.
```

average_payload_mass
2928

- The average payload mass carried by booster version F9 v1.1 is 2928 kg.

First Successful Ground Landing Date

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

```
%sql SELECT MIN(DATE) FROM SPACEXDATASET \
    WHERE landing__outcome = 'Success (ground pad)'
```

```
 * ibm_db_sa://nzc67634:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/bludb
Done.
```

1
2015-12-22

- The first successful landing outcome on ground pad happened on 2015 December 22nd.

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 SPACEXDATASET \
    WHERE (landing__outcome = 'Success (drone ship)') \
            AND (payload_mass__kg_ BETWEEN 4001 AND 5999)
```

```
 * ibm_db_sa://nzc67634:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/bludb
Done.
```

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

- There are 4 boosters that satisfied the queries, namely F9 FT B1022, F9 FT B1026, F9 FT B1021.2, and F9 FT B1031.2.

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

```
%sql SELECT mission_outcome, COUNT(mission_outcome) AS total_number FROM SPACEXDATASET \
    GROUP BY mission_outcome
```

```
 * ibm_db_sa://nzc67634:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/bludb
Done.
```

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

- There is only 1 failure mission outcome and 100 success mission outcomes, with one uncleared payload status.

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 SPACEXDATASET \
    WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM SPACEXDATASET)
```

```
 * ibm_db_sa://nzc67634:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/bludb
Done.
```

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

- The result shows a list of 12 booster versions that carried the maximum payload mass of 15600 kg.

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 SPACEXDATASET \
    WHERE (landing__outcome = 'Failure (drone ship)') AND (DATE BETWEEN '2015-01-01' AND '2015-12-31')
```

```
 * ibm_db_sa://nzc67634:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/bludb
Done.
```

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

- In 2015, booster version F9 v1.1 B1012 and B1015 have failed landing in drone ship both on launch site 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(landing__outcome) AS count_outcome FROM SPACEXDATASET \
    WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' \
    GROUP BY landing__outcome \
    ORDER BY count_outcome DESC
```

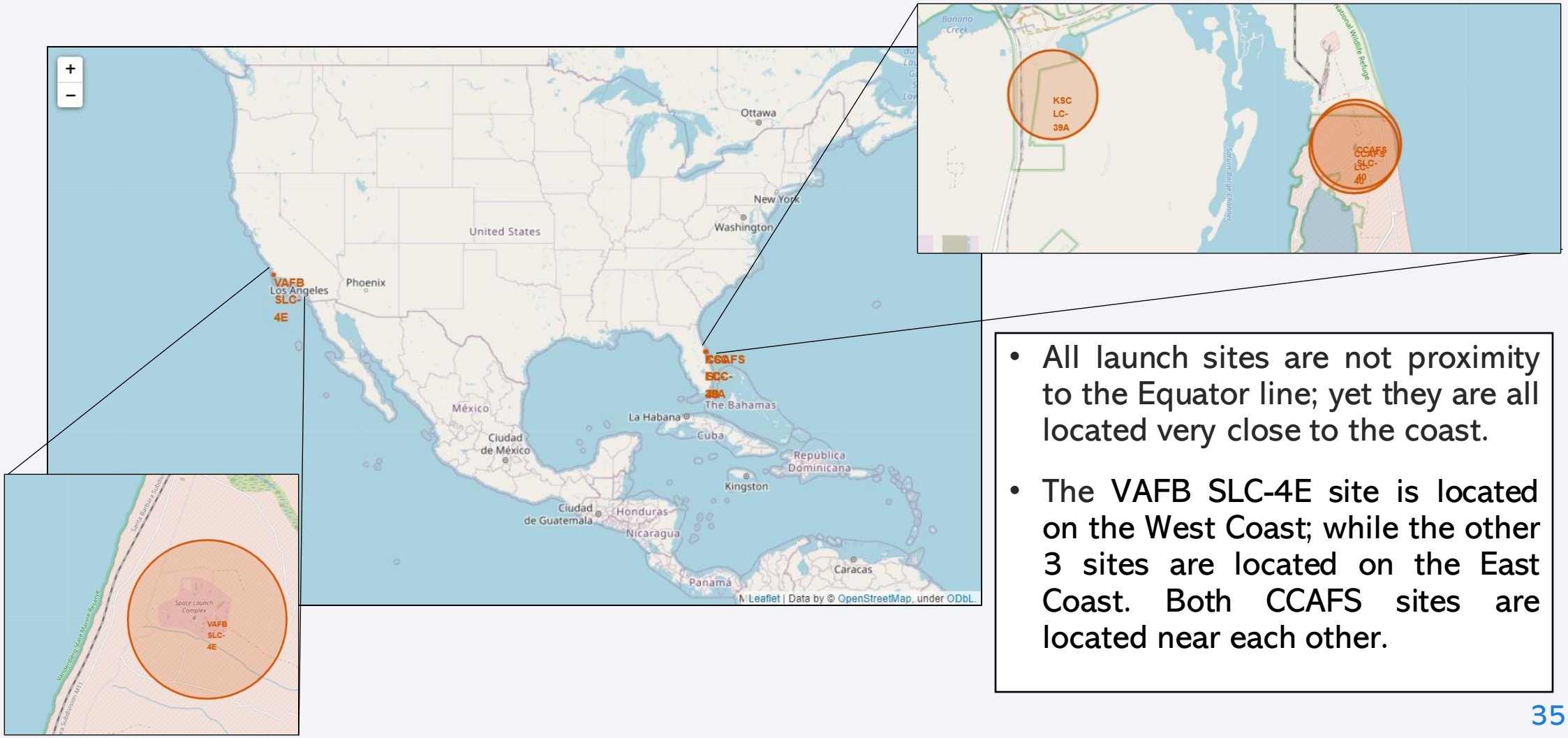
```
 * ibm_db_sa://nzc67634:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/bludb
Done.
```

landing__outcome	count_outcome
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

- Within the specified dates, most of the landing outcomes were no attempt, followed by both success and failure in drone ship.

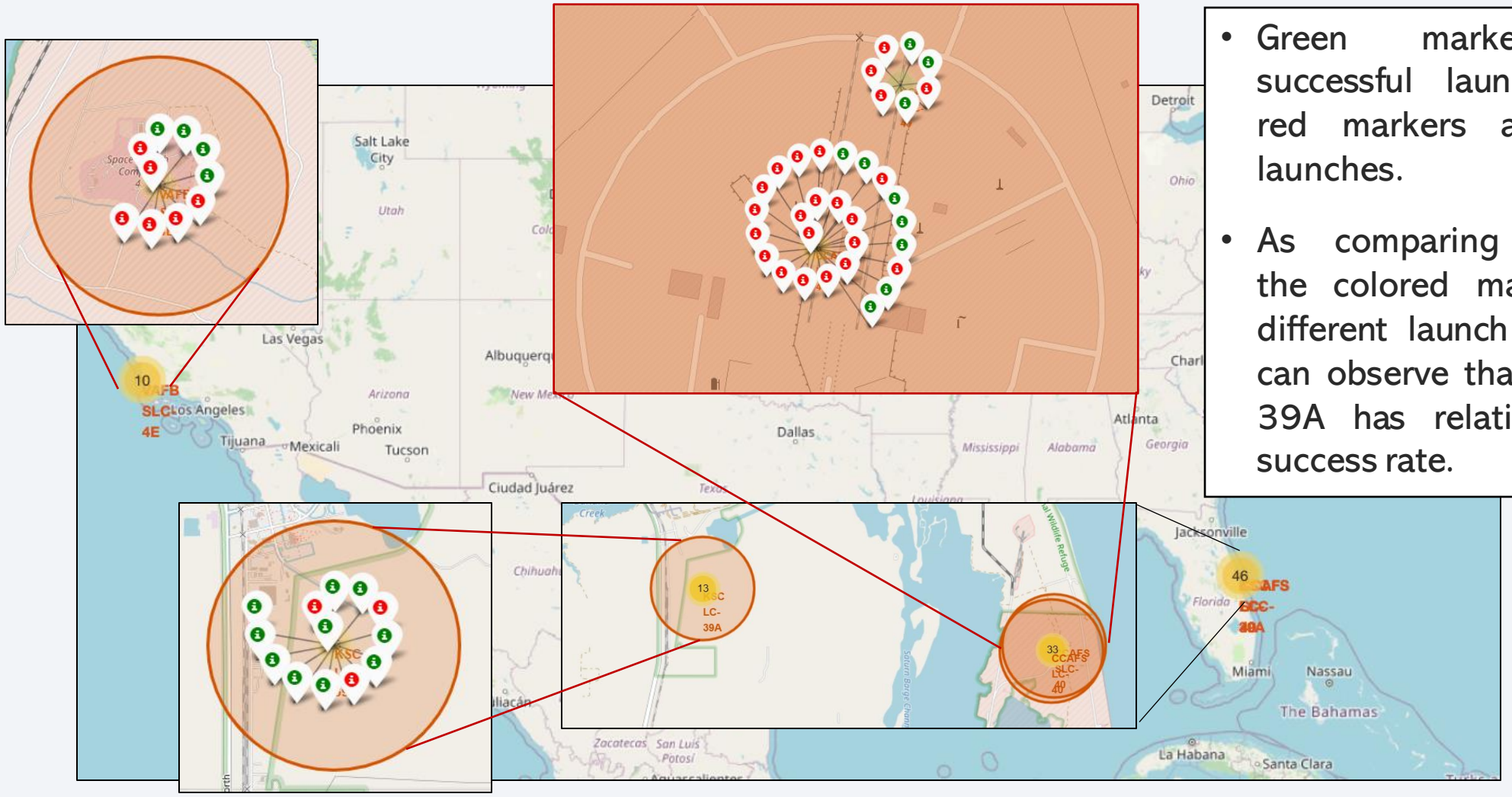

Section 3
Launch Sites
Proximities Analysis

Locations of All Launch Sites



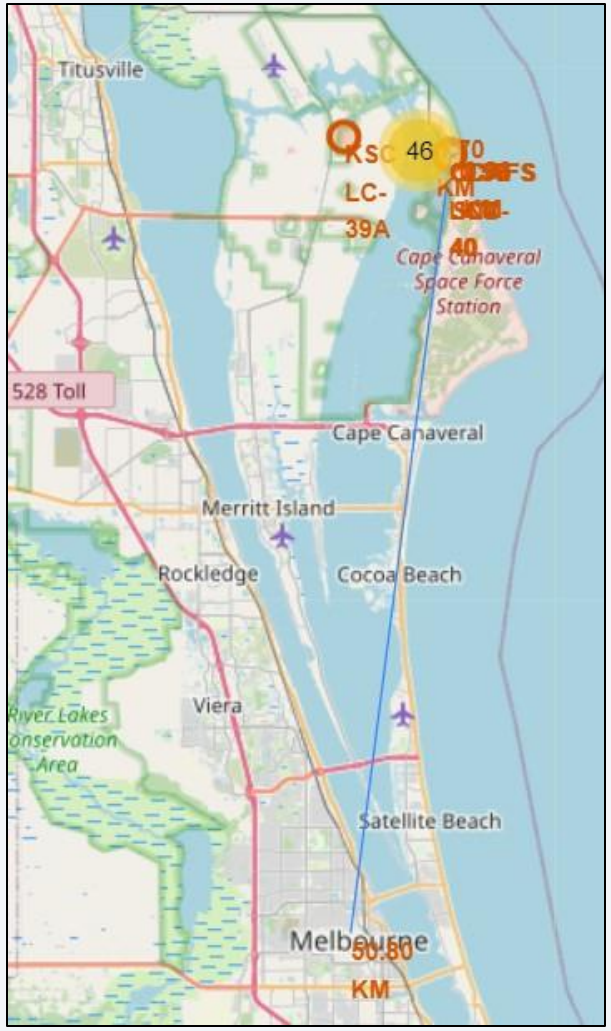
- All launch sites are not proximity to the Equator line; yet they are all located very close to the coast.
- The VAFB SLC-4E site is located on the West Coast; while the other 3 sites are located on the East Coast. Both CCAFS sites are located near each other.

Success or Failed Launches for each Site



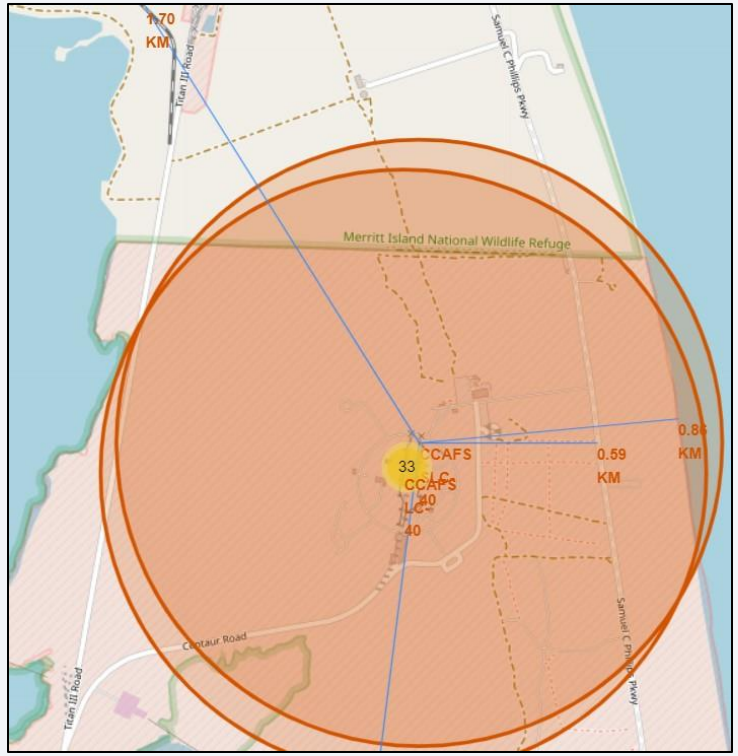
- Green markers are successful launches and red markers are failed launches.
- As comparing between the colored markers for different launch sites, we can observe that KSC LC-39A has relatively high success rate.

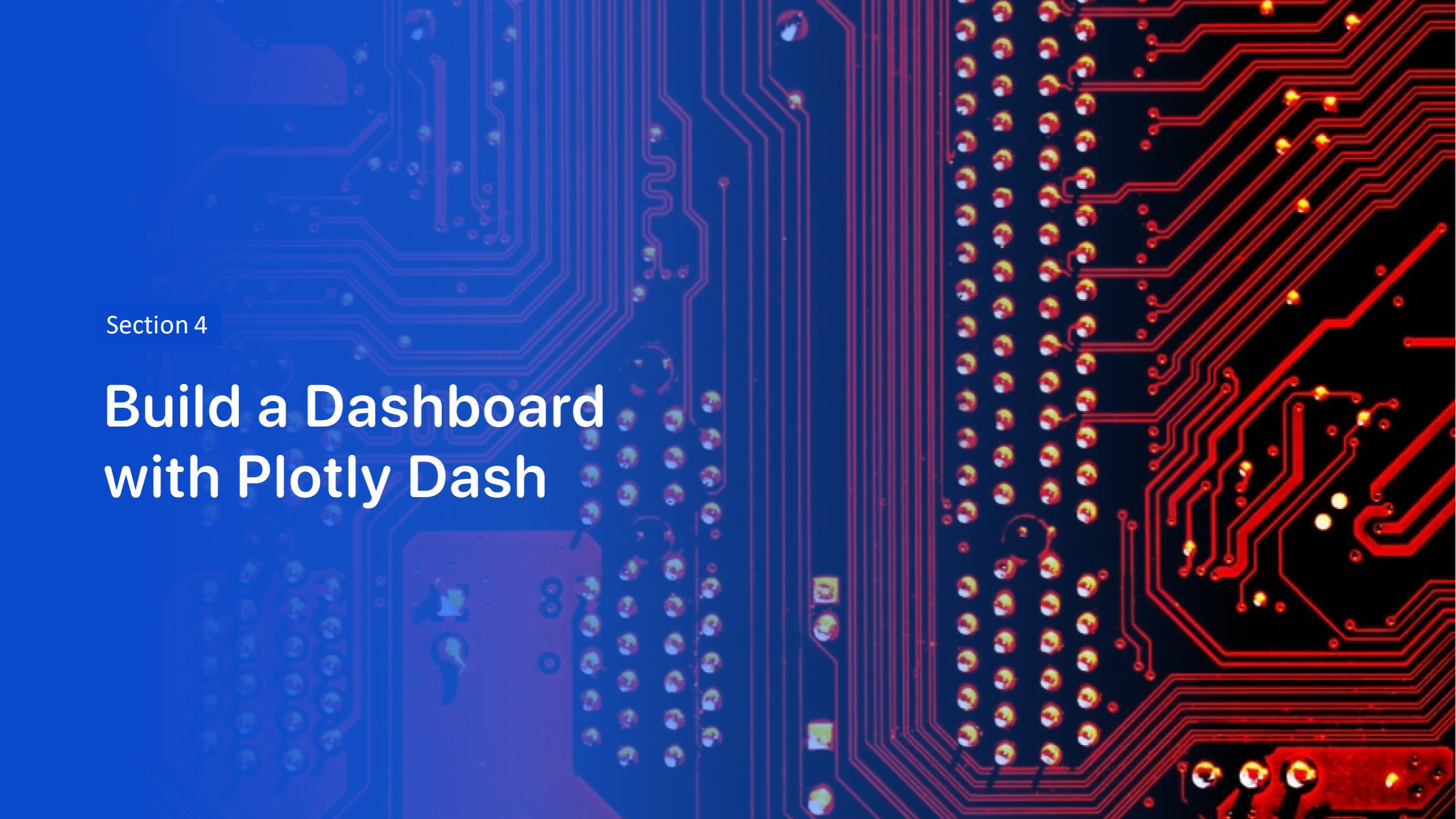
Distances between a Launch Site to its Proximities



- The maps display the distances between site CCAFS SLC-40 to its proximities such as nearest city, railway, highway, and coastline.
- The results are summarized as table below:

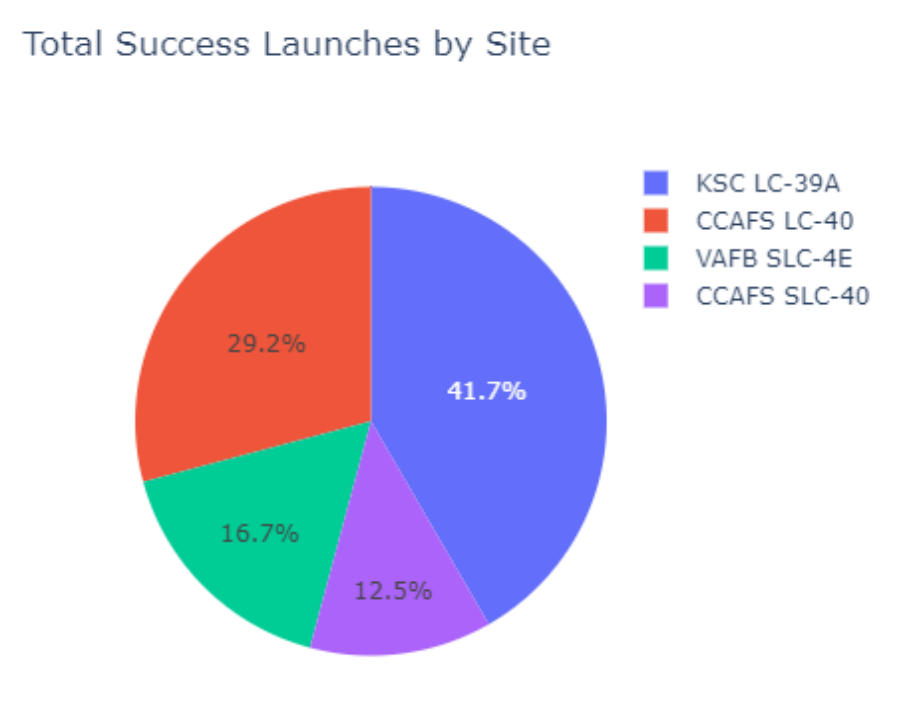
Proximities	Distances (KM)
Nearest city	50.80
Railway	1.70
Highway	0.59
Coastline	0.86




Section 4
Build a Dashboard with Plotly Dash

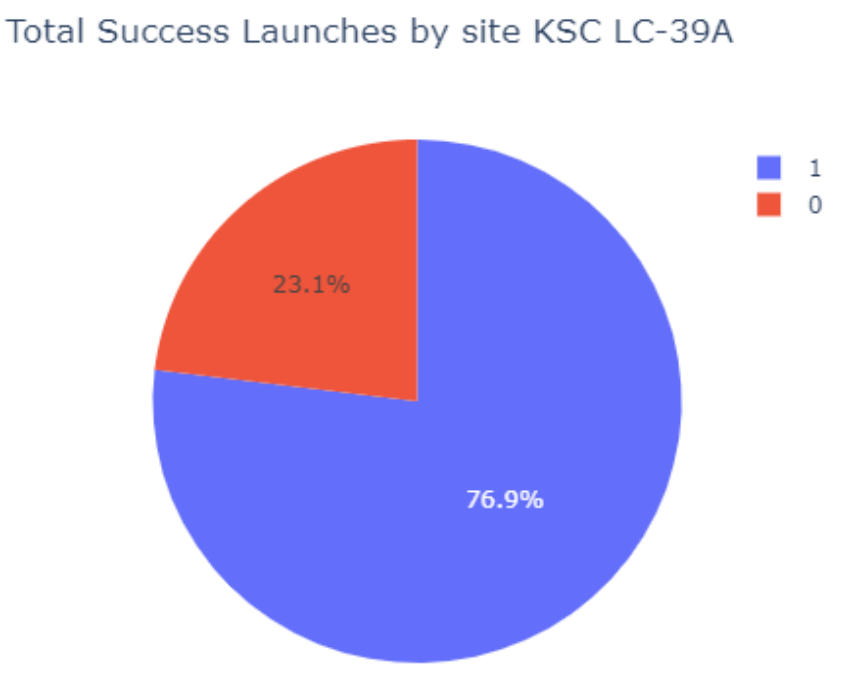
Launch Success Counts for All Sites

- Based on the SpaceX Launch Record Dashboard, pie chart on the right displays the total counts of success for all sites.
- Launches on KSC LC-39A site has the highest success counts, with 41.7% success launches among all.



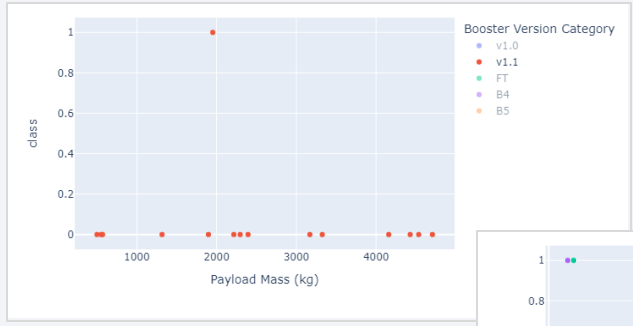
Launch Success Ratio for Launch Site

- In the SpaceX Launch Record Dashboard, pie charts on the right show the success ratio (`1 = success vs 0 = failed`).
- KSC LC-39A site has the highest success ratio of 76.9%; while the other sites have < 50% of success ratio (Refer to Appendix 3).

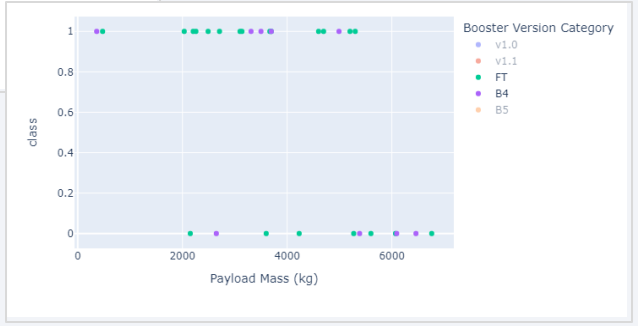


Payload VS. Launch Outcome for All Sites

- The scatterplots illustrate the relationships between payload and launch outcomes with different booster versions. Overall, greater payload mass will decrease the success rate. (above)



- However, we can also see some outlier in payload range after 7000 kg. Hence zooming between the range from 0 to 7000 kg, we found that booster v1.1 has a very low success rate; as boosters FT and B4 have higher success rate with lower payload mass. (right)

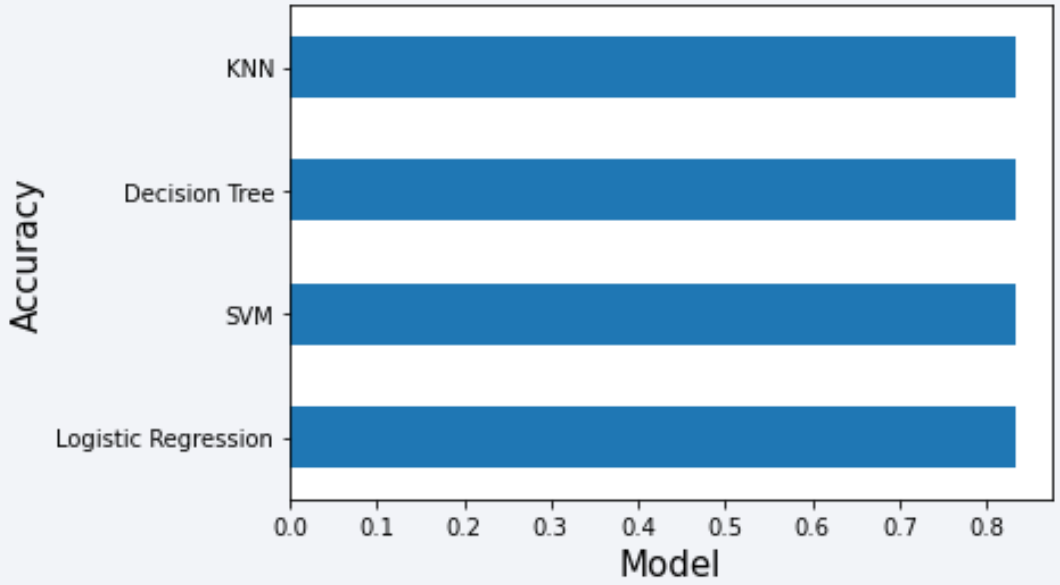


Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Bar chart below visualizes the built model accuracy for all built classification models.
- It seems that all four models have the similar classification accuracy of ≈83.33%.



Confusion Matrix

- Since the accuracy for all models perform the same, their confusion matrix are also similar as shown in the figure.
- With the confusion matrix having a 80% precision and 100% sensitivity, we can observe that the models predicted all the landed outcomes correctly though there are few "did not land" outcomes mis-predicted as land by the models.

Conclusions

- This project was conducted to predict if the first stage of SpaceX Falcon 9 rockets will land successfully. Hence, classification models were built and evaluated for this purpose.
- All four models including Logistic Regression, SVM, Decision Tree, and K-Nearest Neighbors are suitable models, having high model accuracy (>80%).
- Our results also found that the successful landings are highly related to where the rockets launch and which booster versions are used. The lighter the payload, the higher the success rate.

Appendix

Appendix 1: Example of Python code and outputs of Machine Learning Prediction

Create a logistic regression object then create a GridSearchCV object `logreg_cv` with cv = 10. Fit the object to find the best parameters from the dictionary `parameters`.

In [10]:
```
parameters ={'C':[0.01,0.1,1],
             'penalty':['l2'],
             'solver':['lbfgs']}
```

In [11]:
```
parameters = {"C":[0.01,0.1,1],'penalty':['l2'], 'solver':['lbfgs']}# l1 lasso l2 ridge
lr = LogisticRegression()
logreg_cv = GridSearchCV(lr, parameters, cv=10)
logreg_cv.fit(X_train, Y_train)
```

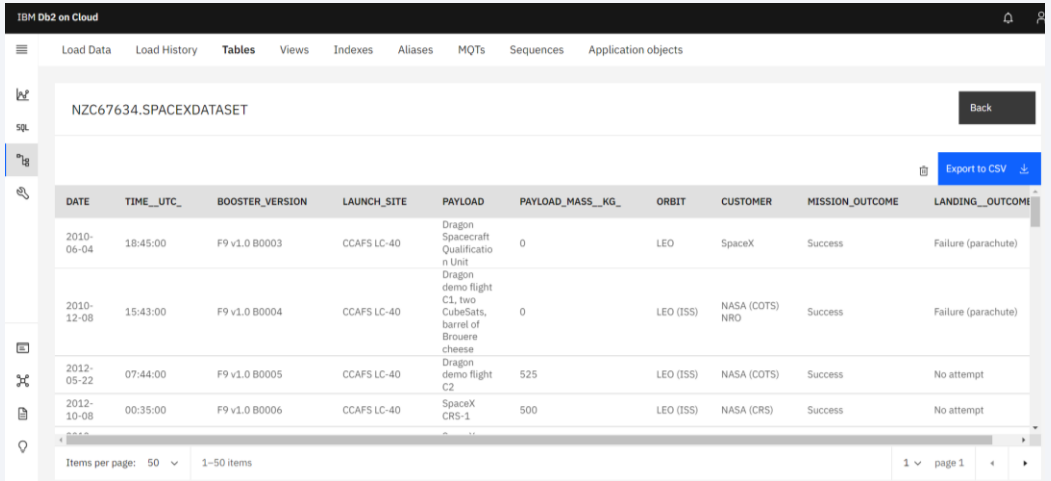
Out[11]:
```
GridSearchCV(cv=10, estimator=LogisticRegression(),
             param_grid={'C': [0.01, 0.1, 1], 'penalty': ['l2'],
                         'solver': ['lbfgs']})
```

We output the `GridSearchCV` object for logistic regression. We display the best parameters using the data attribute `best_params\_` and the accuracy on the validation data using the data attribute `best_score\_`.

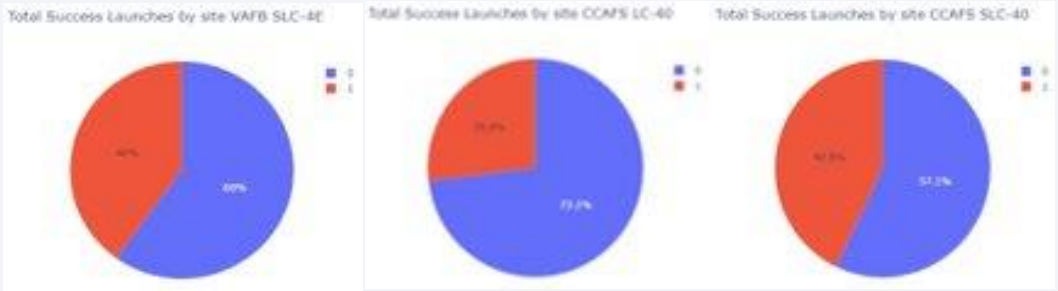
In [12]:
```
print("tuned hpyerparameters :(best parameters) ",logreg_cv.best_params_)
print("accuracy :",logreg_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}
accuracy : 0.8464285714285713
```

Appendix 2: Dataset stored in IBM DB2 Database



Appendix 3: Pie Charts of Success Ratio for Other Sites



Appendix 4: Example of SQL query

```
%sql SELECT landing__outcome, COUNT(landing__outcome) AS count_outcome FROM SPACEXDATASET \
    WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' \
    GROUP BY landing__outcome \
    ORDER BY count_outcome DESC
```

Thank you!