Winning Space Race with Data Science

Azamat Kuzdibay
20. 04. 2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection with SpaceX API and web scraping
  - Data wrangling in order to extract launch sites, orbits and corresponding outcomes
  - Exploratory data analysis (EDA) using visualization and SQL
  - Interactive visual analytics using Folium and Plotly Dash
  - Predictive analysis using classification models
- Summary of all results
  - Effect of launch sites, payload and booster version was shown
  - Prediction model with over 80% accuracy was established

Introduction

- How to be successful as a new contender in space travel industry?
  - Current market leader: Space X
  - Advantage: reuse of the first stage rocket  wing 62M USD cost vs. 165M+ from others
  - Potential market gap: what if the first stage cannot be reused? This would allow competitors to outbid the price
- Problems you want to find answers
  - Will the first stage of the SpaceX rocket land?
  - What factors influence the success of first stage rocket landing?

Section 1

Methodology

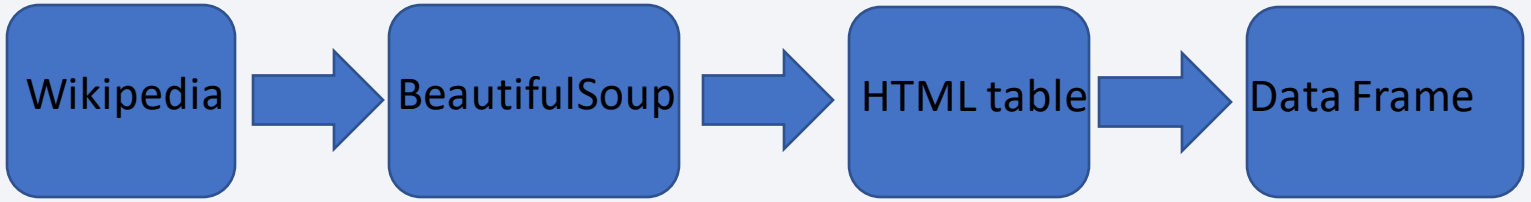
Methodology

Executive Summary

- Data collection methodology:
  - Web scraping: Data was imported from Wikipedia
  - Data was directly accessed through SpaceX API
- Perform data wrangling
  - Data was organized in order to see the different launch sites, target orbits and corresponding outcomes
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Tuning models (linear regression, decision trees, support vector machine and k Nearest Neighbour) using training set of data on a parameter space and then applying the best parameter set on the test data

Data Collection

- Web scraping: Data was imported from Wikipedia



- Data was directly accessed through SpaceX API

Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts:
- SpaceX API call resulted in JSON file that was converted into a DataFrame
- Add the GitHub URL for reference:

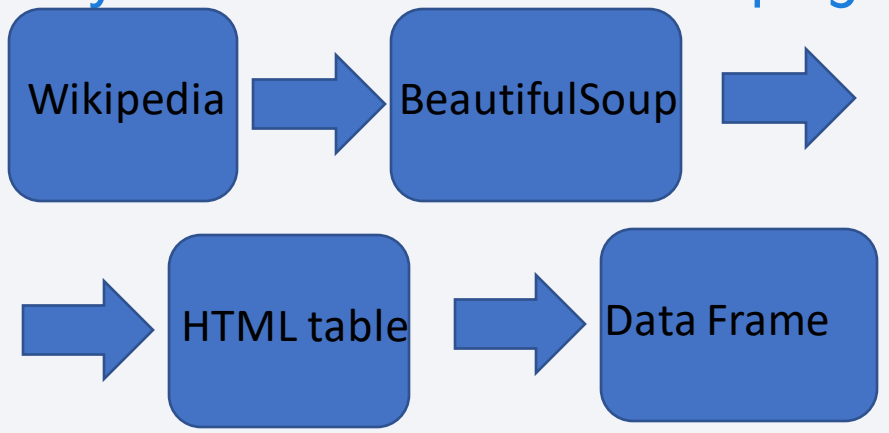
https://github.com/GhostOfGreg/applied_data_science_capstone/blob/1e0f10039c477ac59feff0e2c3c62fefd223d807/Data%20Collection%20API%20Lab.ipynb

Place your flowchart of SpaceX API calls here

Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts:
  - Wikipedia site was called in with BeautifulSoup.
  - Then HTML table were exctracted and transferred into a DataFrame
- Add the GitHub URL for reference:
  - https://github.com/GhostOfGreg/applied_data_science_capstone/blob/1e0f10039c477ac59feff0e2c3c62fefd223d807/Data%20Collection%20with%20Web%20Scraping%20lab.ipynb

Place your flowchart of web scraping here



Data Wrangling

- Describe how data were processed
  - Different launch sites were extracted
  - Target orbits were investigated
  - Corresponding outcomes were listed
- Add the GitHub URL for reference:
  - https://github.com/GhostOfGreg/applied_data_science_capstone/blob/1e0f10039c477ac59feff0e2c3c62fefd223d807/Part%203:%20Data%20Wrangling.ipynb

EDA with Data Visualization

- Charts plotted: (Aim: find correspondences to successful outcome of stage 1 return)
  - Flight Number and Launch Site
  - Payload and Launch Site
  - Success rate of each orbit type
  - Flight Number and Orbit type
  - Payload and Orbit type
  - Launch success yearly trend
- Add the GitHub URL of your completed EDA with data visualization notebook, as an external reference and peer-review purpose:
  - https://github.com/GhostOfGreg/applied_data_science_capstone/blob/1e0f10039c477ac59fef f0e2c3c62fefd223d807/Part%205:%20Expl.%20Data%20Analysis%20-%20Visualization.ipynb

EDA with SQL

- Using bullet point format, summarize the SQL queries you 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 acheived.
  - 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 in 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
- Add the GitHub URL:
  - https://github.com/GhostOfGreg/applied_data_science_capstone/blob/1e0f10039c477ac59feff0e2c3c62fefd223d807/Part%204:%20Expl.%20Data%20Analysis%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map:
  - Mark all launch sites on the Map
  - Mark the success/failed launches for each site on the map
  - Calculate the distances between a launch site to its proximities (e.g. open water)
- Explain why you added those objects: to explore if the location or the surroundings of the location has any effect on the stage 1 rocket return outcome
- Add the GitHub URL:
  - https://github.com/GhostOfGreg/applied_data_science_capstone/blob/1e0f10039c477ac59feff0e2c3c62fefd223d807/Part%206:%20Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard:
  - 2 inputs: launch site as dropdown menu and payload range as slider was implemented
  - Outputs: Pie chart with success rate regarding the launch site and scatter plot for success rate of chosen launch site and payloads
- Explain why you added those plots and interactions: To have a more clear data extracted for the same question that was analyzed with Folium maps
- Add the GitHub URL:
  - https://github.com/GhostOfGreg/applied_data_science_capstone/blob/1e0f10039c477ac59fef f0e2c3c62fefd223d807/Part%207:%20Spacex_dash_app.py

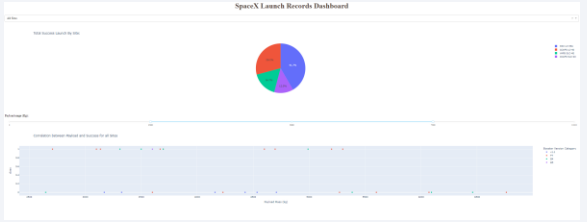
Predictive Analysis (Classification)

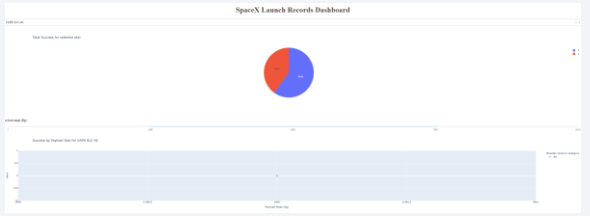
- Summarize how you built, evaluated, improved, and found the best performing classification model
  - Models: linear regression, decision trees, support vector machine and k Nearest Neighbour
  - Training set to explore parameter space and finding best parameters
  - Then applying the best parameter set on the test data
- Add the GitHub URL:

  https://github.com/GhostOfGreg/applied_data_science_capstone/blob/1e0f10039c477ac59feff0e2c3c62fefd223d807/Part%208:%20Machine%20Learning%20Prediction.ipynb

Results

- Exploratory data analysis results
  - Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavy payload mass(greater than 10000)
  - 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.
  - 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.
  - The sucess rate since 2013 kept increasing till 2020
- Interactive analytics demo in screenshots



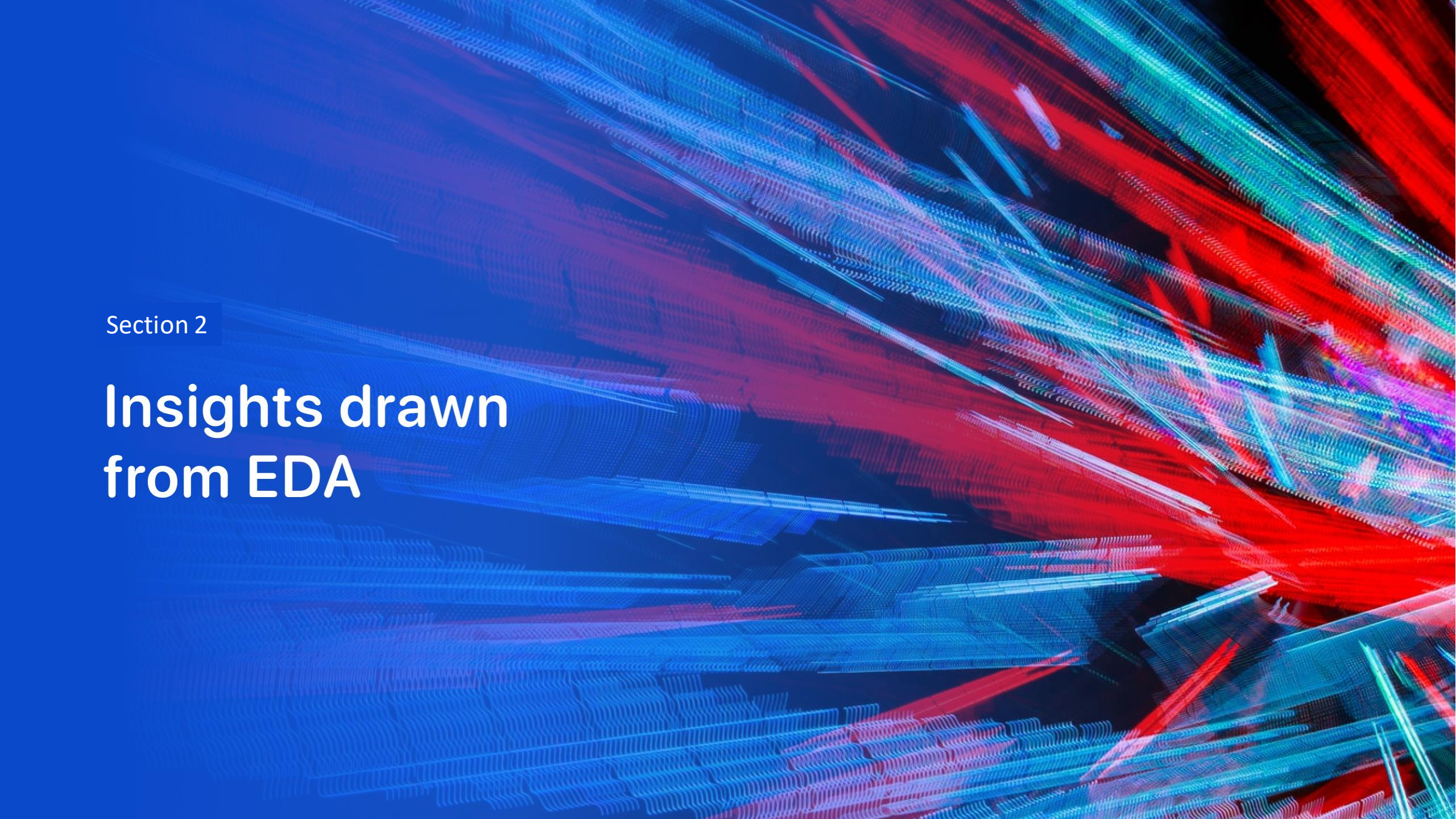


- Predictive analysis results:

	Algorithm	Jaccard	F1-score	LogLoss	Accuracy
0	KNN	0.800000	0.888889	NA	0.833333
1	Decision Tree	0.769231	0.869565	NA	0.833333
2	SVM	0.800000	0.888889	NA	0.833333
3	LogisticRegression	0.800000	0.888889	0.478667	0.833333

Section 2

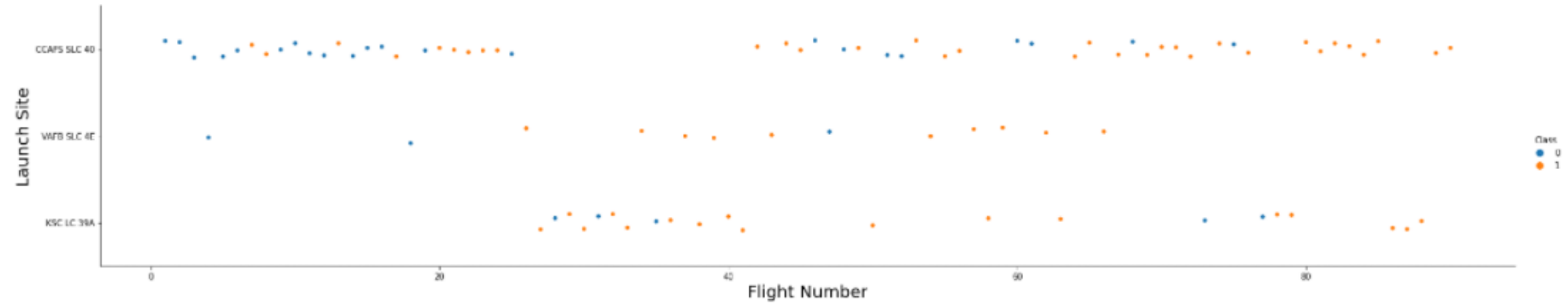
Insights drawn from EDA



Flight Number vs. Launch Site

- Show a scatter plot of Flight Number vs. Launch Site

- Show the screenshot of the scatter plot with explanations

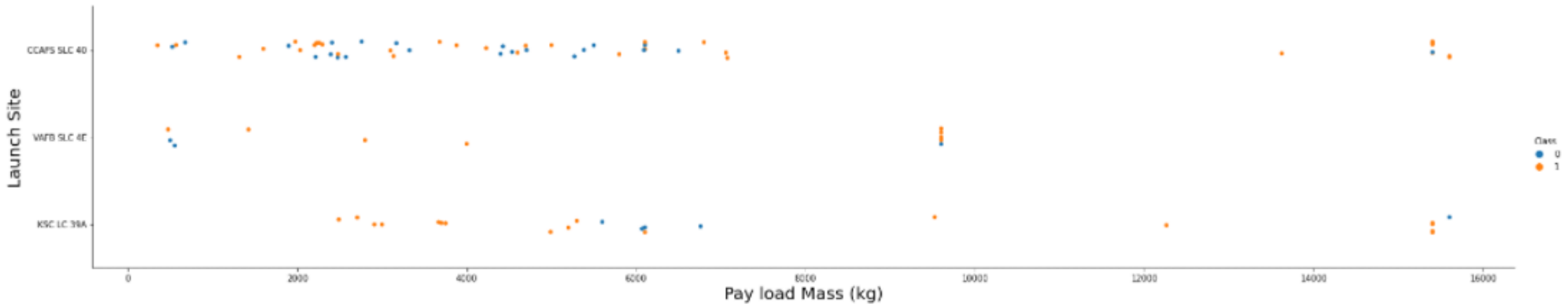


Now try to explain the patterns you found in the Flight Number vs. Launch Site scatter point plots.

Payload vs. Launch Site

- Show a scatter plot of Payload vs. Launch Site

- Show the screenshot of the scatter plot with explanations

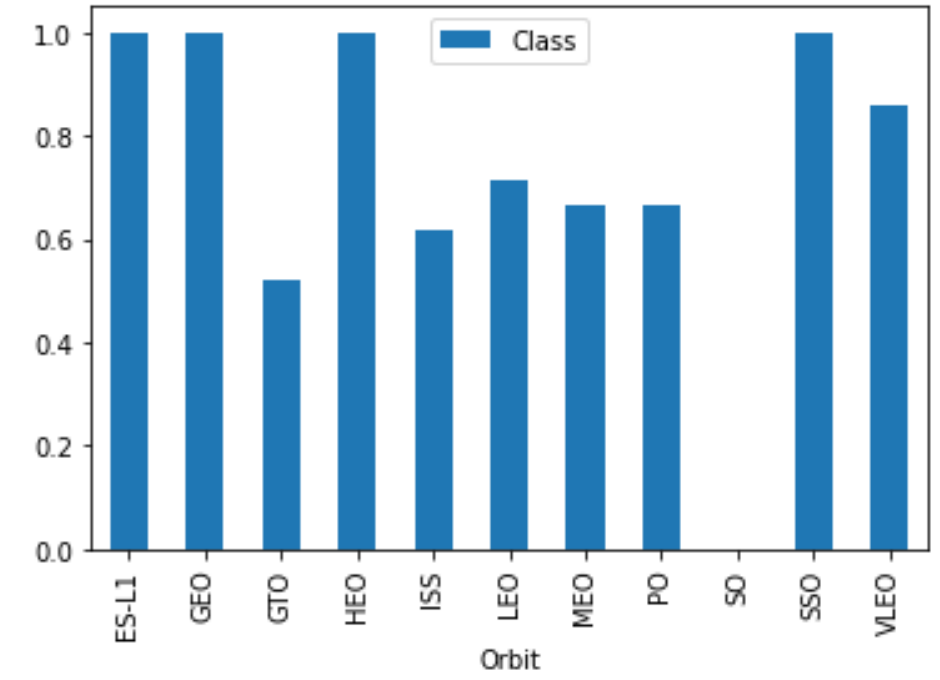


Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type

- Show a bar chart for the success rate of each orbit type

- Show the screenshot of the scatter plot with explanations

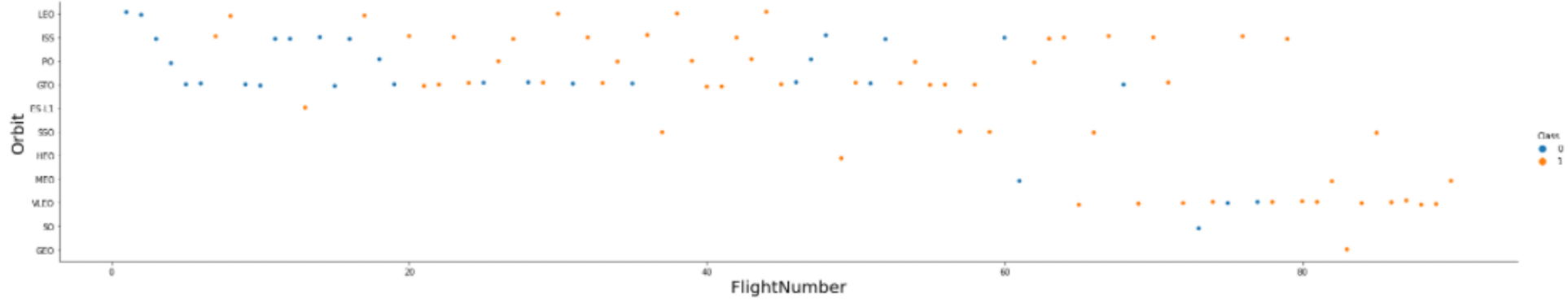


Analyze the ploted bar chart try to find which orbits have high sucess rate.

Flight Number vs. Orbit Type

- Show a scatter point of Flight number vs. Orbit type

- Show the screenshot of the scatter plot with explanations

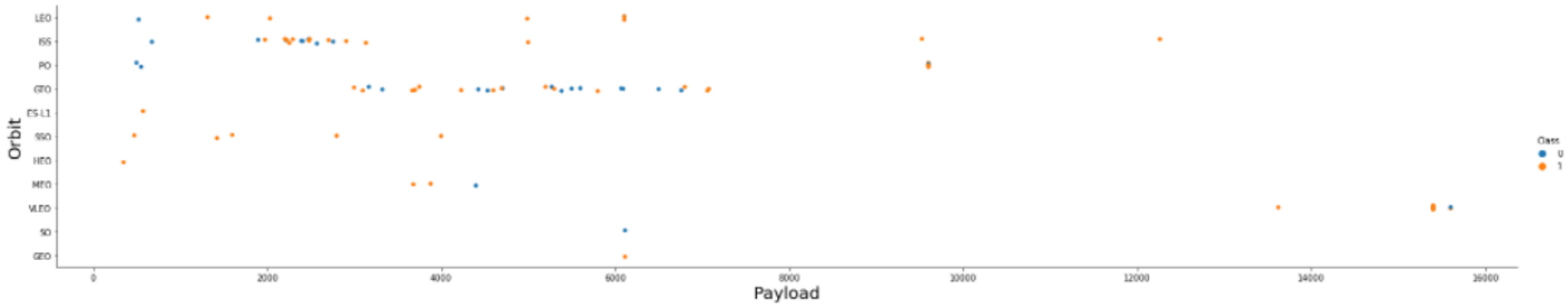


You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type

- Show a scatter point of payload vs. orbit type

- Show the screenshot of the scatter plot with explanations



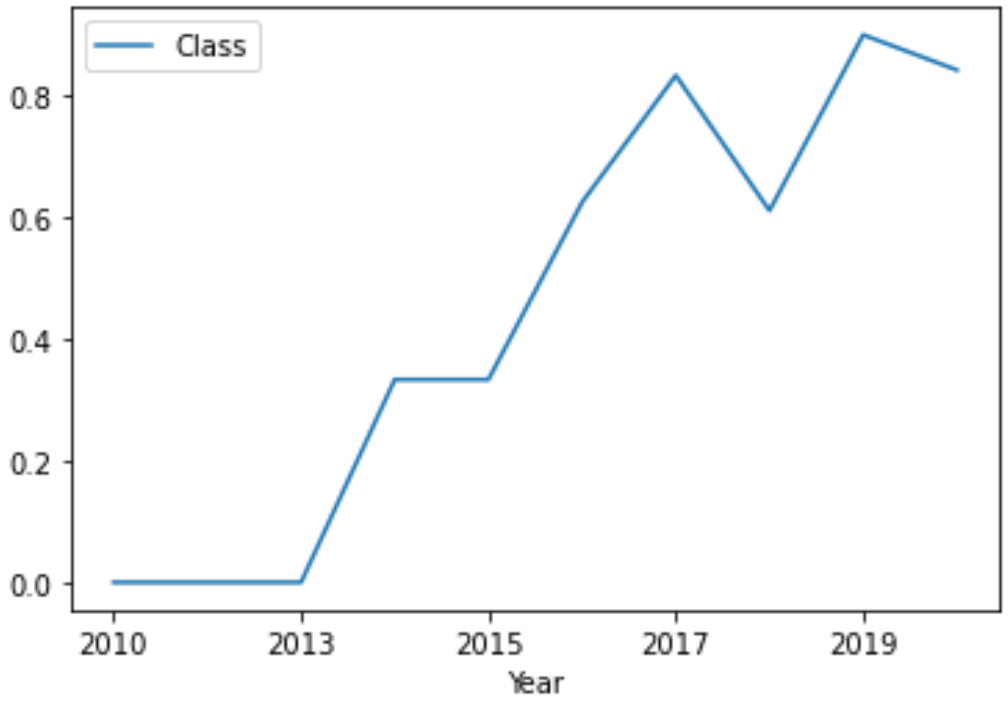
With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

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

- Show a line chart of yearly average success rate

- Show the screenshot of the scatter plot with explanations



you can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

- Find the names of the unique launch sites
- Present your query result with a short explanation here: there are 4 launch sites

Display the names of the unique launch sites in the space mission

```
%sql select LAUNCH_SITE from SPACEXTBL group by LAUNCH_SITE
```

```
 * ibm_db_sa://tfb03092:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

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`
- Present your query result with a short explanation here: 5 records were selected

Display 5 records where launch sites begin with the string 'CCA'

```
%sql select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' limit 5
```

```
 * ibm_db_sa://tfb03092:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/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

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- Present your query result with a short explanation here: payloads for a specific customer were summed

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql select sum(payload_mass__kg_) from spacextbl where customer like '%CRS%'
```

```
 * ibm_db_sa://tfb03092:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

1
48213

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Present your query result with a short explanation here: the average payload mass was calculated

Display average payload mass carried by booster version F9 v1.1

```
%sql select avg(payload_mass__kg_) from spacextbl where booster_version like '%F9 v1.1%'
```

```
 * ibm_db_sa://tfb03092:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

1
2534

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Present your query result with a short explanation here: date with a specific landing outcome criteria was selected

List the date when the first successful landing outcome in ground pad was acheived.

Hint:Use min function

```
%sql select min(date) from spacextbl where landing__outcome = 'Success (ground pad)'
```

```
 * ibm_db_sa://tfb03092:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

1
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

- Present your query result with a short explanation here: the corresponding booster versions are listed

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%sql select booster_version from spacextbl where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000
```

```
 * ibm_db_sa://tfb03092:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Present your query result with a short explanation here: in this table not many failures were present

List the total number of successful and failure mission outcomes

```
%sql select mission_outcome, count(*) from spacextbl group by mission_outcome
```

```
 * ibm_db_sa://tfb03092:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

mission_outcome	2
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
- Present your query result with a short explanation here: F9 B5 booster seems successful

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%sql select booster_version from spacextbl where payload_mass__kg_ = (select max(payload_mass__kg_) from spacextbl)
```

```
 * ibm_db_sa://tfb03092:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

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

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

- Present your query result with a short explanation here: there were 2 failures in 2015

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 like 'Failure%' and year(date) = 2015
```

```
 * ibm_db_sa://tfb03092:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/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

- Present your query result with a short explanation here: In the selected time frame, success was dominant

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 mission_outcome, count(*) from spacextbl where date between '2010-06-04' and '2017-03-20' group by mission_outcome order by count(*) desc
```

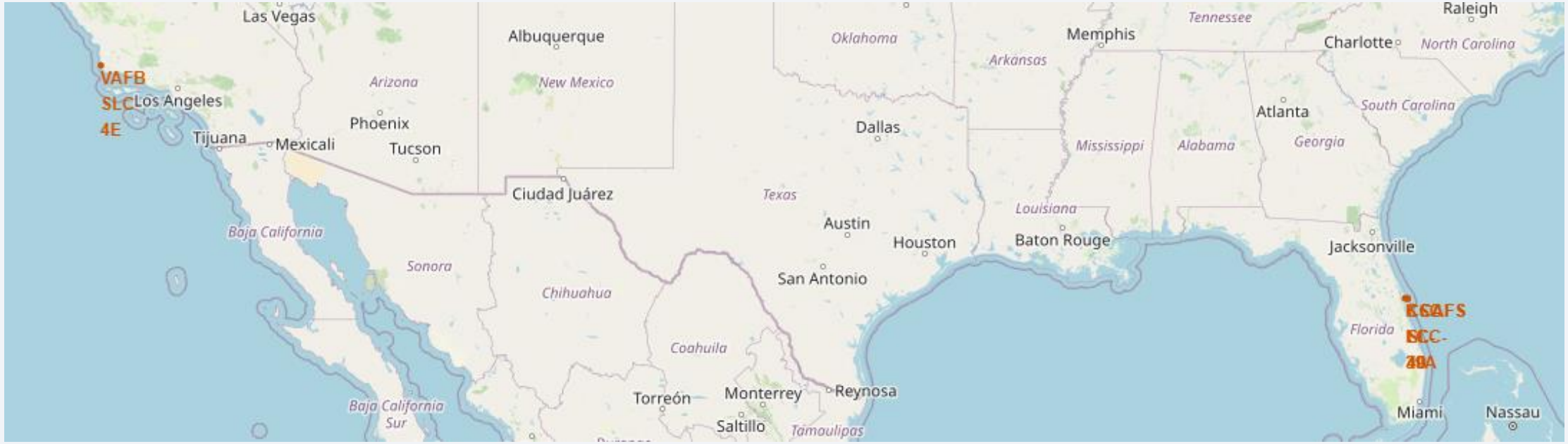
```
 * ibm_db_sa://tfb03092:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

mission_outcome	2
Success	30
Failure (in flight)	1


Section 3
Launch Sites
Proximities Analysis

Launch site locations

- Replace <Folium map screenshot 1> title with an appropriate title
- Explore the generated folium map and make a proper screenshot to include all launch sites' location markers on a global map
- Explain the important elements and findings on the screenshot: There are launch sites both on the east and west coast



Launch outcomes

- Replace <Folium map screenshot 2> title with an appropriate title

- Explore the folium map and make a proper screenshot to show the color-labeled launch outcomes on the map

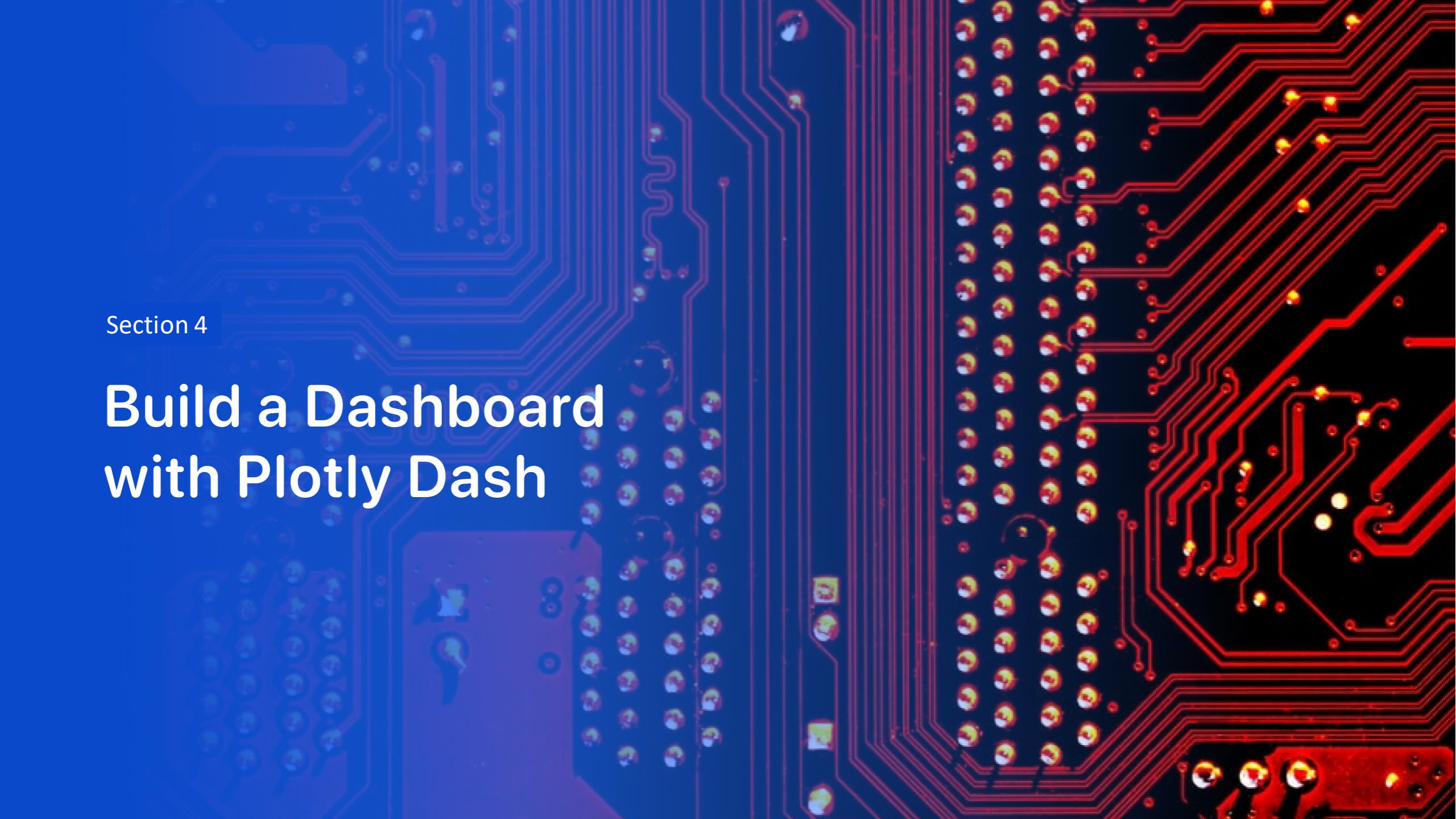
- Explain the important elements and findings on the screenshot: it is possible to interactively explore the successful launches for each site

Distance from Features

- Replace <Folium map screenshot 3> title with an appropriate title

- Explore the generated folium map and show the screenshot of a selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed

- Explain the important elements and findings on the screenshot: we see that all launch sites are close to open water


Section 4
Build a Dashboard with Plotly Dash

Success count for all sites

- Replace <Dashboard screenshot 1> title with an appropriate title
- Show the screenshot of launch success count for all sites, in a piechart
- Explain the important elements and findings on the screenshot: KSC LC-39A has the biggest success rate

Succes ratio of KSC -L39A

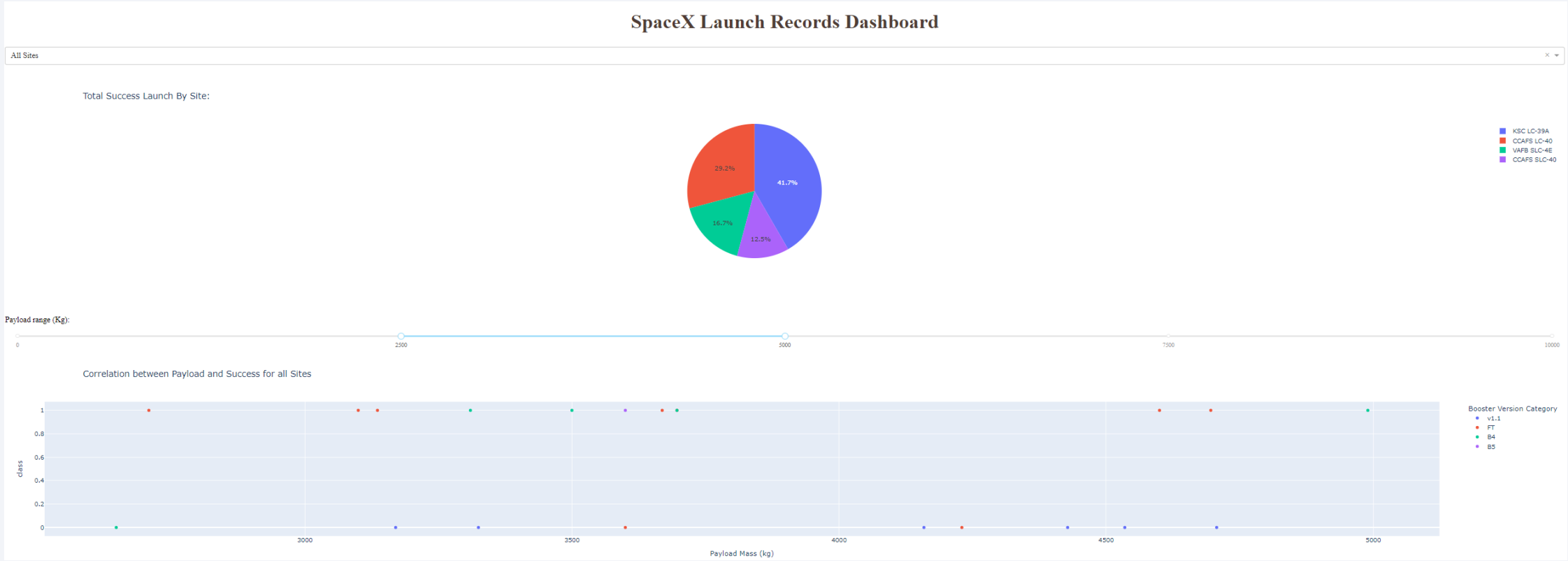
- Replace <Dashboard screenshot 2> title with an appropriate title

- Show the screenshot of the piechart for the launch site with highest launch success ratio

- Explain the important elements and findings on the screenshot: 76.9% success in outcome

Success rate for all sites and payloads limit

- Replace <Dashboard screenshot 3> title with an appropriate title
- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider
- Explain the important elements : we see that payloads between 2500 and 5000 kg had success with 4 different booster



Section 5

Predictive Analysis (Classification)

Classification Accuracy

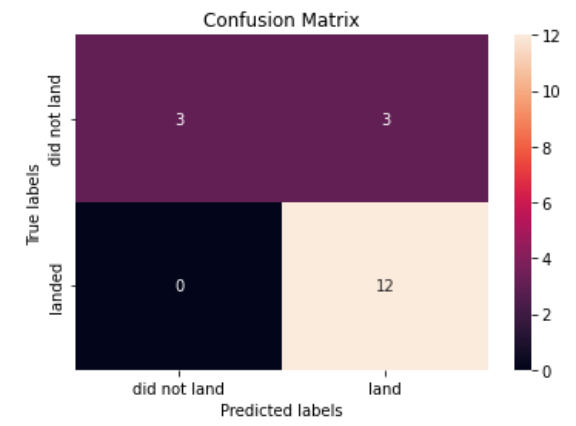
- Visualize the built model accuracy for all built classification models, in a bar chart

- Find which model has the highest classification accuracy:
  - All model performed similarly

	Algorithm	Jaccard	F1-score	LogLoss	Accuracy
0	KNN	0.800000	0.888889	NA	0.833333
1	Decision Tree	0.769231	0.869565	NA	0.833333
2	SVM	0.800000	0.888889	NA	0.833333
3	LogisticRegression	0.800000	0.888889	0.478667	0.833333

Confusion Matrix

- Show the confusion matrix of the best performing model with an explanation:
  - False positive cases are present (3)
  - Otherwise the prediction is accurate for 12 True positive and 3 true negative cases



Conclusions

- We saw that launch site, payload and booster version both can have influence on success
- We did a lot of SQL queries in order to practice sql
- Knowing that the success year increases with time (I.e. Year) can we really conclude anything except that with newer technology, spacex performs better?
- There is really no point to make a 47 slide powerpoint, except to show everything you did. No meeting member would appreciate 47 slides and with 1.5 minutes talk per slide, **70 minutes** presentation!! Even for a PhD defense you can talk typically 45-60 minutes!!
- Specially after the reading part of week 5 exercise, where it was pointed out how to summarize your findings and tell **only the important things in a compact way**!

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project
- Random snippet to practice to make an appendix:

```
# TASK 4:
# Add a callback function for `site-dropdown` and `payload-slider` as inputs, `success-payload-scatter-chart` as output
@app.callback(Output(component_id='success-payload-scatter-chart', component_property='figure'),
              [Input(component_id='site-dropdown', component_property='value'), Input(component_id="payload-slider", component_property="value")])
def get_scatter_chart(entered_site, slider_value):
    low, high = slider_value
    weight_df = spacex_df[(spacex_df['Payload Mass (kg)'] > low) & (spacex_df['Payload Mass (kg)'] < high)]

    if entered_site == 'ALL':
        fig = px.scatter(weight_df, x='Payload Mass (kg)', y='class', color='Booster Version Category', title='Correlation between Payload and Success for all
        return fig
    else:
        filtered_df = weight_df[weight_df['Launch Site'] == entered_site]
        fig = px.scatter(filtered_df, x='Payload Mass (kg)', y='class', color='Booster Version Category', title=f'Success by Payload Size for {entered_site}')
        return fig
```

Thank you!