IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Rocco Beltran
07-07-2023

Outline

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

Executive Summary

Summary of methodologies

Data collection via API, Web scraping.

Data Wrangling.

Exploratory Data Analysis with SQL.

Exploratory Data Analysis with Visualization.

Interactive visual analytics with Folium.

Interactive dashboard using Dash.

Predictive Analysis(Classification).

Summary of all results

Exploratory Data Analysis results

Interactive visual analytics and dashboard

Predictive Analysis(Classification) results

Introduction

- Project background and context
- In this project, we will try to predict if the Falcon 9 first stage will land successfully. SpaceX advertises its 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 are due to that SpaceX can reuse the first stage. Therefore if we are able to determine if the first stage will land, we can determine the cost of a launch. This information can be used if a different company would like to bid against SpaceX for a rocket launch.
- Problems you want to find answers
    - How does different factors such as number of flights, payload mass, launch site, orbit type affect whether the first stage will land successfully?
    - Does the launch success yearly trend increase or decrease?
    - In proximity to what are the launch sites?
    - How close are they to that proximity?
    - What is the success rate for each site?
    - What model would predict best the successful landing?

Section 1

Methodology

Methodology

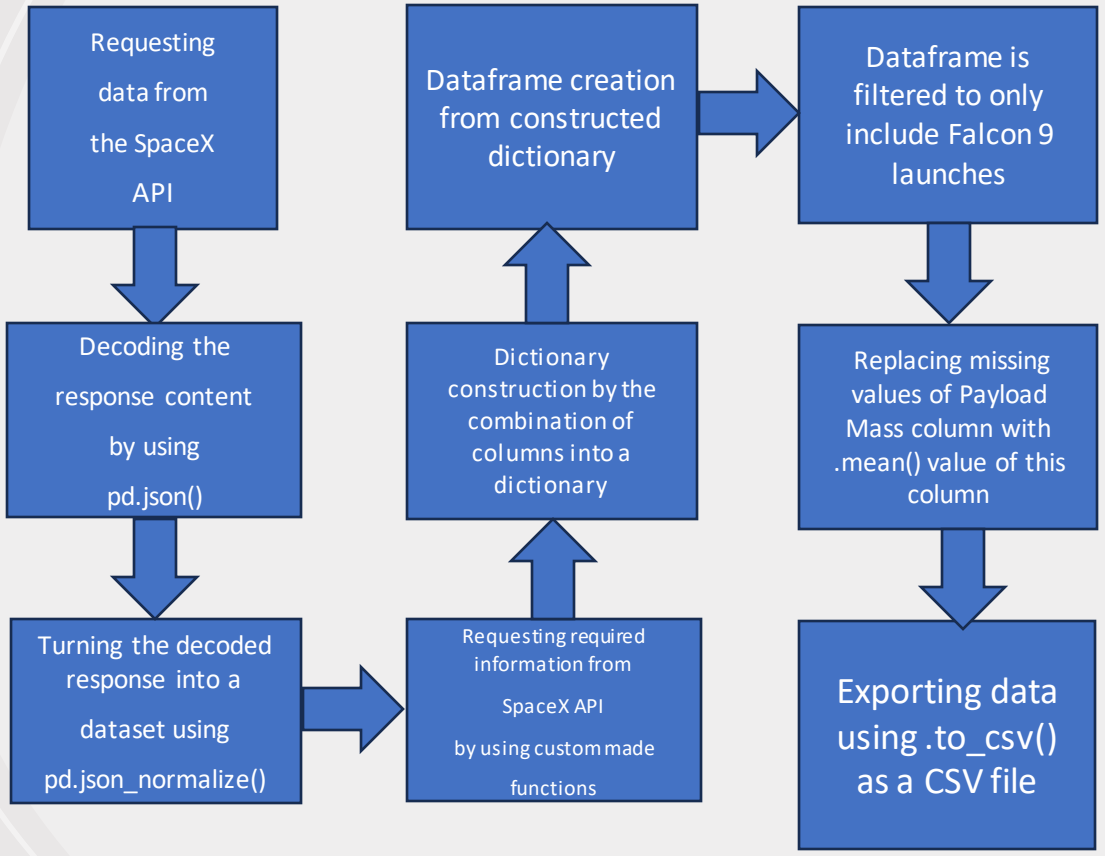
- **Executive Summary:**

- **Data collection methodology:**
  - Data was collected from the SpaceX REST API, and with Web Scraping from Wikipedia.
- **Performed data wrangling by:**
  - Removed unnecessary columns
  - Filtered the dataframe to only include Falcon 9 launches
  - Dealt with missing values
  - Performed exploratory data analysis
  - Determined training labels
  - Created a landing outcome label from Outcome column for classification models
- **Performed exploratory data analysis (EDA) using visualization and SQL**
- **Performed interactive visual analytics using Folium and Plotly Dash**
- **Performed predictive analysis using classification models:**
  - Built, tuned and evaluated classification models to determine which is the best fit to predict the first stage landing.

Data Collection

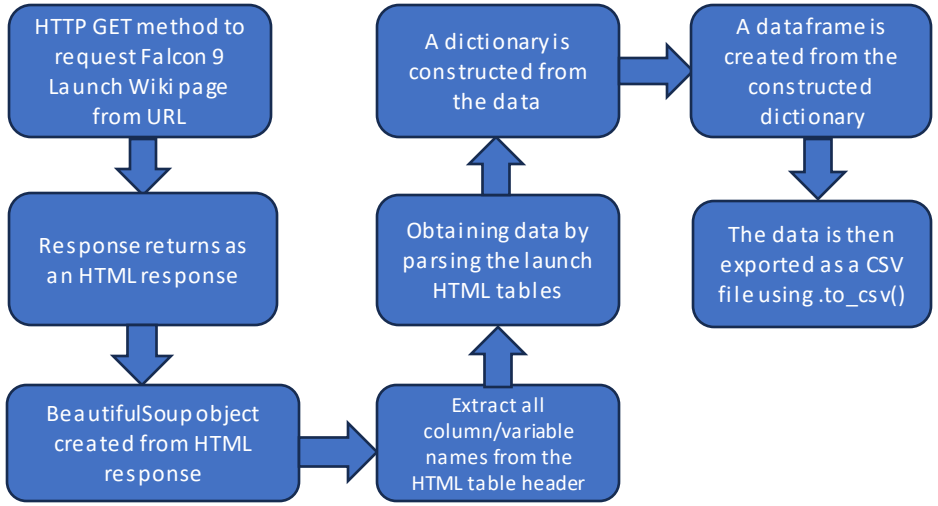
- In the Data Collection process, data sets were collected in two ways which are using a REST API from SpaceX and Web Scraping of data from a table in its Wikipedia page.
  - Data was obtained from this two different sources in order to limit the bias that could arise if obtained from just one source.
- **Data Set Collection using the REST API:**
  - In this method of data collection a data set was constructed from the data obtained via this method.
  - The resultant dataset columns are:
  - FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude and Latitude.
- **Data Set Collection using Web Scraping:**
  - In this method of data collection a table present in Wikipedia that contained relevant information was extracted and parsed using BeautifulSoup, a response body is ceated. The columns were then extracted and then their contents from the response body. A dataset was then created.
  - The resultant dataset columns are:
  - FlightNo, Launch site, Payload, Payload mass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date and Time.

Data Collection – SpaceX API



Link Here

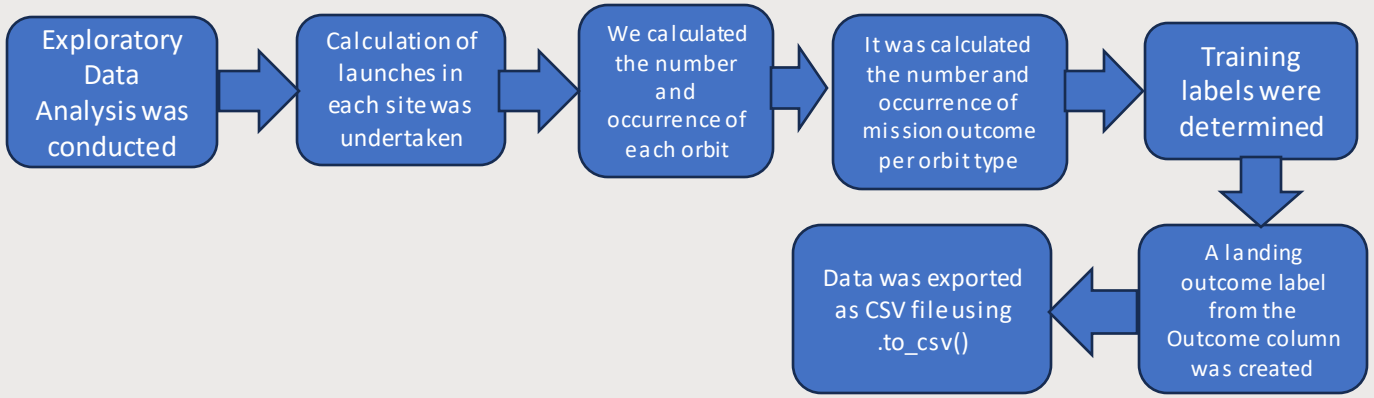
Data Collection - Scraping



Link Here

Data Wrangling

- In this section, we performed some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.
- In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, `True Ocean` means the mission outcome was successfully landed to a specific region of the ocean while `False Ocean` means the mission outcome was unsuccessfully landed to a specific region of the ocean. `True RTLS` means the mission outcome was successfully landed to a ground pad `False RTLS` means the mission outcome was unsuccessfully landed to a ground pad. `True ASDS` means the mission outcome was successfully landed on a drone ship `False ASDS` means the mission outcome was unsuccessfully landed on a drone ship.
- In this section we converted mainly those outcomes into Training Labels where `1` means the booster successfully landed and `0` means it was unsuccessful.

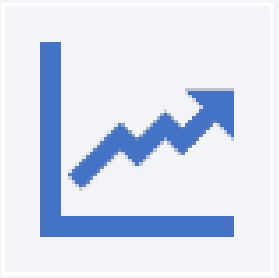


Link Here

EDA with Data Visualization

Various scatter point charts were plotted in order to better visualize the relationship between FlightNumber(indicating the continuous launch attempts) and PayloadMass, between FlightNumber and LaunchSite, between FlightNumber and Orbit and Payload Mass and Orbit. A scatter plot allows as to visualize the relationship between one independent variable and one dependent variable.

A bar chart was plotted in order to visualize the different success rates for orbit type. A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent, in this case the frequency of launch success rates for each orbit type.



A line chart was visualized with x axis to be year and y axis to be average success rate in order to envision the launch success yearly trend. The line chart is a simple, two-dimensional chart with an X and Y axis, each point representing a single value. The data points are joined by a line to depict a trend, usually over time. In this case average launch success rate yearly trend.

Link Here

EDA with SQL

- SQL queries were perfomed in order to:
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where the launch sites begin with the string 'KSC'
  - 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 where the successful landing outcome in drone ship was achieved
  - List the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass List the records which will display the month names, successful landing_outcomes in ground pad, booster versions, launch_site for the months in year 2017
  - Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order

Link Here

Build an Interactive Map with Folium

- A map was created as site_map with an initial center location to be NASA Johnson Space Center using folium.Map(). A red circle at NASA Johnson Space center was then created with folium.Circle() and then labelled as NASA Johnson Space Center using folium.Popup() and a marker is then created also using folium.map.Marker() to all of this be then added using site_map.add_child(Circle) and site_map.add_child(Marker).
- This same map was then re-initialed with all the above procedures done for each launch site with all the resulting labels and circles being added to the map. A map with center at NASA Johnson Space Center, all launch sites labelled, and their corresponding circles were created. Using MousePosition() we created a box in the top-right corner of the map which would facilitate us the coordinates where the mouse is, it will come in handy when calculating the distance from a site to a proximity. A cluster of markers were created using the color assigned to the launch site outcome. Green for success and red for failure. This is then added to the map. Finally the distance of these sites to the railway, coastline, closest highways, cities, wildlife protected areas and the equator was calculated, a line representing the distance was added along with a label stating the given distance using marker.add_to(site_map) and folium.Polyline().add_to(site_map).
- I have added the circles in order to mark the places where the launch sites are, I have labelled them so that the circles are representative of the site they are marking. The cluster of markers can give us an idea of the ratio of success vs failure of each site. The line stating the distance provides a good idea of close to what things a site has to be.

Build a Dashboard with Plotly Dash

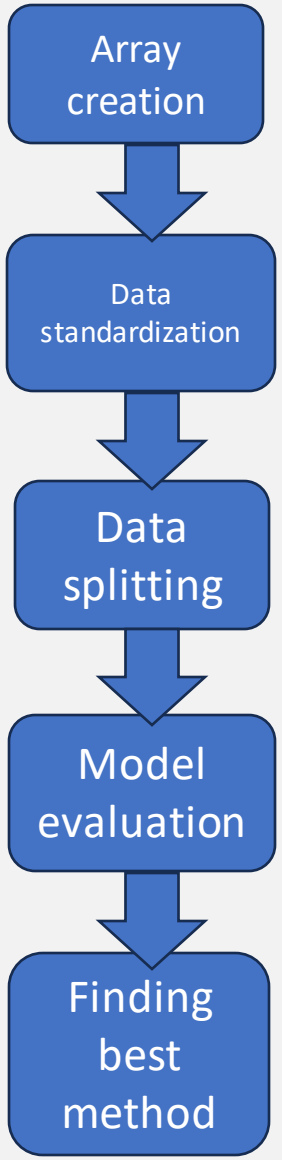
- We have added a dropdown list with all sites selected by default.
- We have added a pie chart that represents the total successful launches per site. If a specific site is selected then it will display the success vs failed counts for the site.
- A slider was added to select payload range
- We have added a scatter chart that shows the correlation between payload and launch success.
- The dropdown list allows us to select the different sites that we have or even re-select all sites as all sites become selected by default
- A pie chart is a circular statistical graphic which is divided into slices to illustrate numerical proportion. In this case the numerical proportion is the total successful launches per site when all sites are selected or the success vs failed counts for a specific site if a site is selected.
- A slider was added to be able to select min and max payload mass in kg in order to better understand between which masses had the most launch success , and also to see which shall be the for example min payload for a launch success to occur.
- A scatter chart, also called a scatter plot, is a chart that shows the relationship between two variables. They are an incredibly powerful chart type, allowing viewers to immediately understand a relationship or trend, which would be impossible to see in almost any other form. In this scenario it would allow us to understand the relationship between payload and launch success.

Link Here

Predictive Analysis (Classification)

- We have:
- Created a NumPy array from the column class in data, by applying the method .to_numpy() and then assigned it to the variable Y, the output was a Pandas series.
- Standardized the data in a variable X which contains the training labels we determined earlier then reassigned it to the same variable X. The result is an array.
- We split the data into training and testing data using the function train_test_split. Then the models will be trained and hyperparameters will be selected using the function GridSearchCV.
- Created a GridSearchCV object with cv = 10 and various classification algorithms. Then fitted the object with the training data. The best parameters along with the best accuracy on the training data for this algorithms were then determined and then evaluated on their accuracy with the test data. Confusion matrices were then plotted in order to gain more insight about each models true positive, false positive, true negative and false negative.
- Found the method which performed best.

Link Here



Results

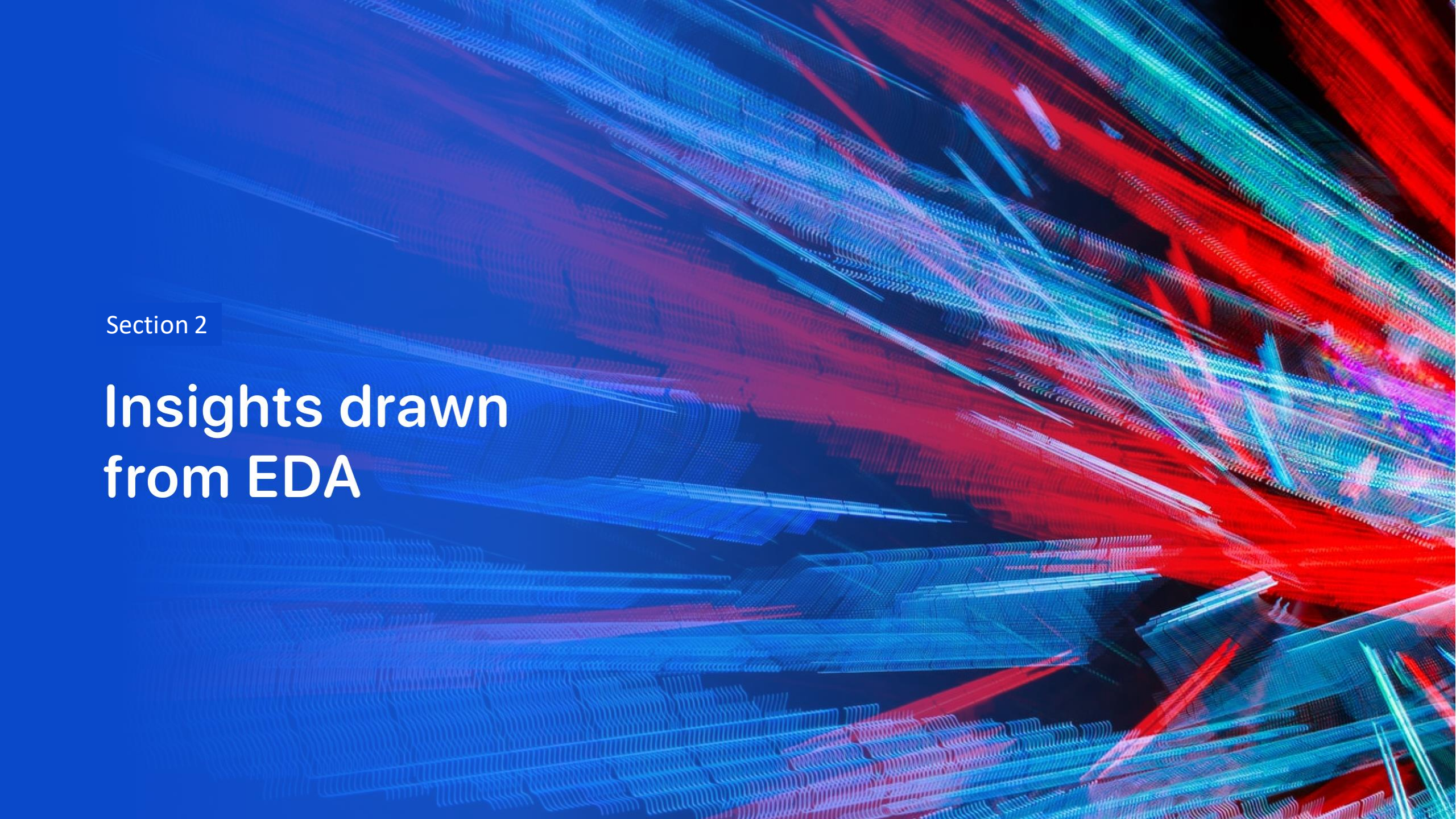
Exploratory data analysis results

Interactive analytics demo in screenshots

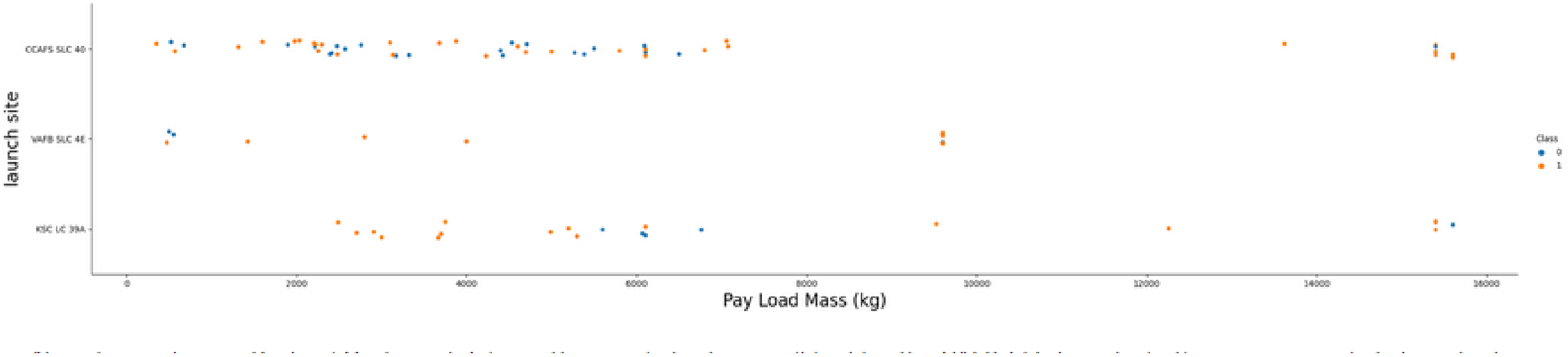
Predictive analysis results

Section 2

Insights drawn from EDA



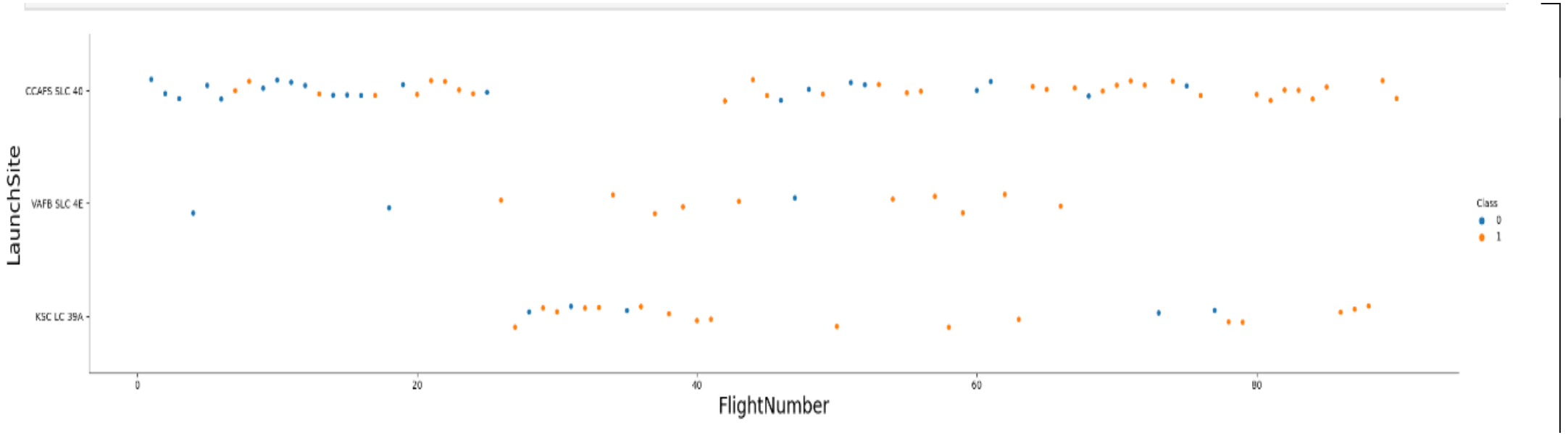
```
plt.show()
```



Payload vs. Launch Site

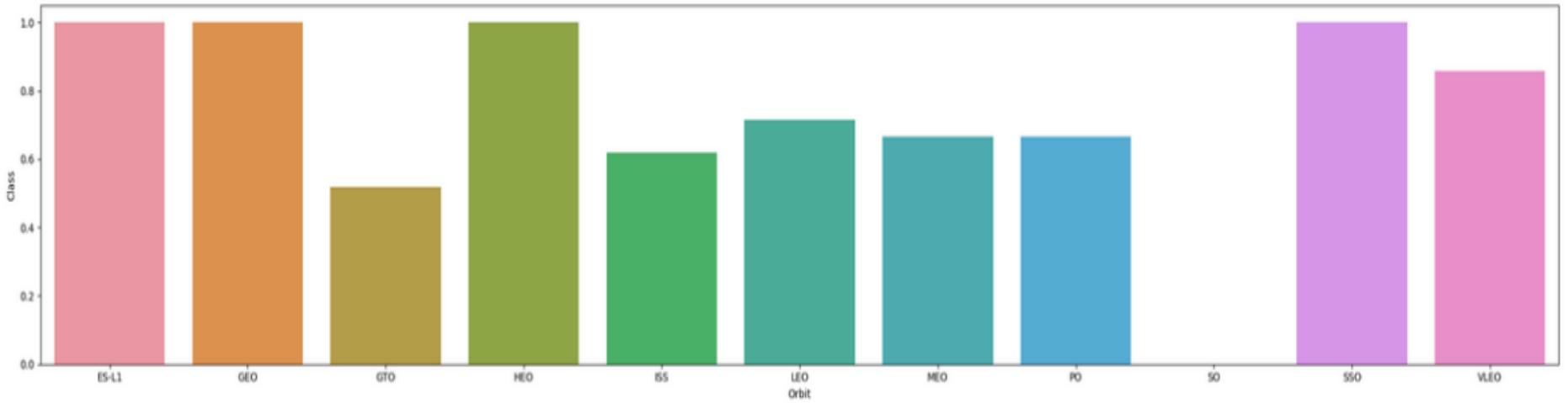
- In here 94% of launches for CCAFS SLC 40 are between a payload mass in kilograms of 0 and approximately 7000.
- For VAFB SLC 4E 67% of launches happened between masses 0 and 4000.
- At KSC LC 39A 55% of launches were focused between specific masses of 2000 to approximately 6300.
- 63% of launches belong to CCAFS SLC 40 with a success rate of 55%.
- Sites VAFB SLC 4E and KSC LC 39A have success rates of 78% and 75% respectively.
- As masses increased fewer launches were made but the more likely that they turned out a success.

Flight Number vs. Launch Site



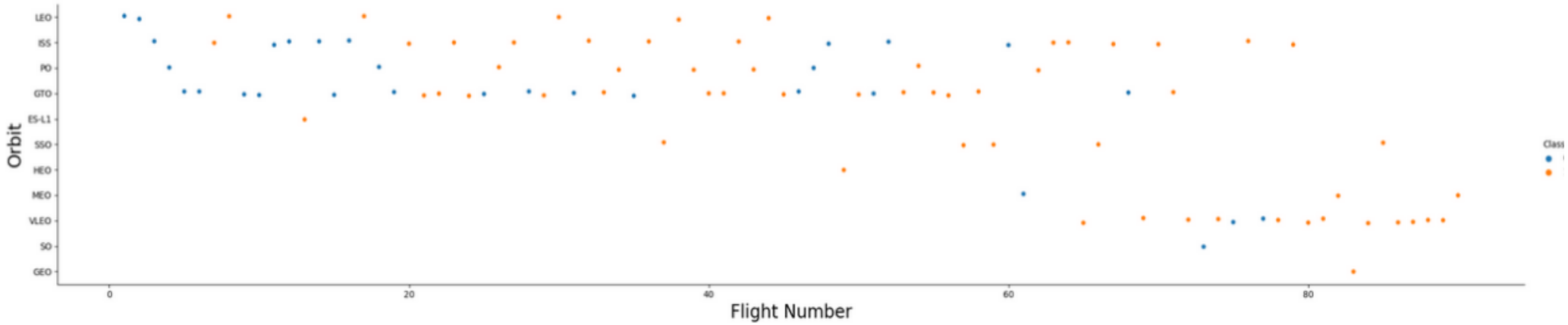
- We see that 61% of launches were made by CCAFS SLC-40.
- Nevertheless CCAFS SLC-40 has the lowest success rate .
- VAFB SLC 4E and KSC LC 39A share a success rate of approximately 77%.
- We see that as the launches in each site increase the more likely that the launch may be a success.

Success Rate vs. Orbit Type



- In this bar chart four orbits, ES-L1, GEO, HEO and SSO have a success rate of around 100%.
- We have an orbit with the lowest success rate other than SO(0%) of approximately 50% which is GTO.
- Four orbits; ISS, LEO, HEO, and PO have success rates of around 60%.
- The VLEO orbit has a success rate of approximately 80%.

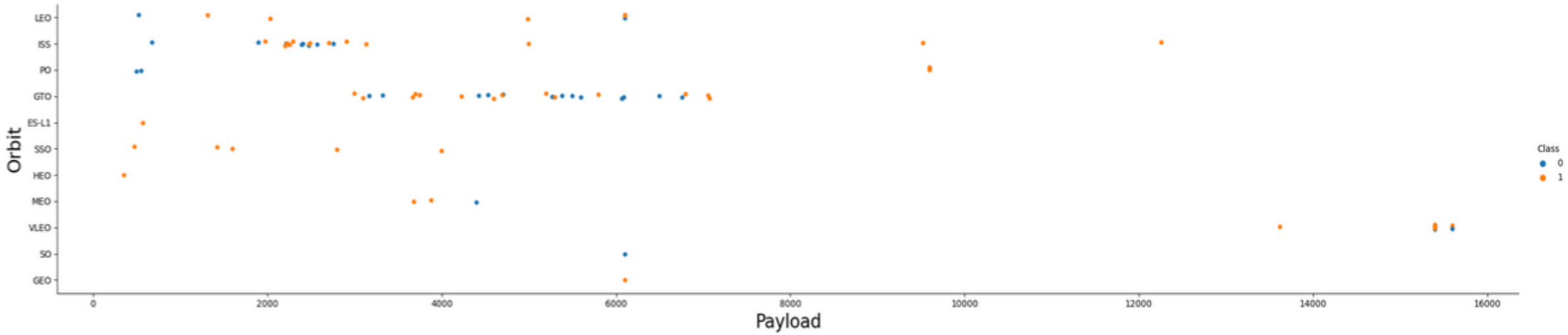
Flight Number vs. Orbit Type



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

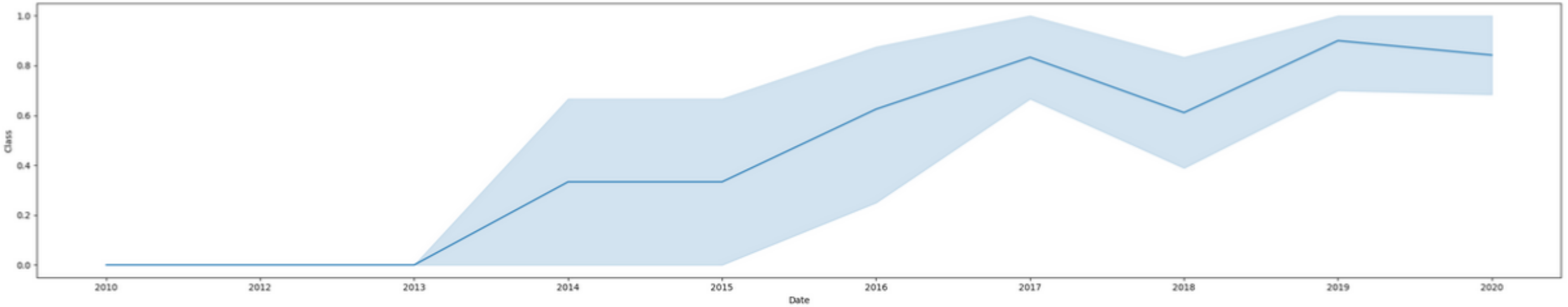
In the LEO orbit we see a positive relationship between number of flights and success outcomes. It seems that launch success outweighs launch failure in the ISS, PO and GTO orbit as the number of flights are over 20. Nevertheless the relationship throughout the GTO orbit appears to not be clear. We see a single launch that is a success for orbits ES-L1, HEO and GEO with number of flights over 10, 40 and 80 respectively. For the SSO orbit we have a total of five successful launches made with a number of flights over 30. We have two orbits MEO and SO that have a unique launch each with a failure of 100% for a number of flights between 60 and 80, on the other the VLEO orbit has two launch failure in between 70 and 80 flights though the launch success of this site greatly outweighs its failure rate. Also all VLEO launches were made over 60 flights.

Payload vs. Orbit Type



With heavy payloads the successful landing or positive landing rate are more for PO, LEO and ISS. However for GTO we cannot distinguish this well as both positive landing and negative landing(unsuccessful mission) are both there here. For ES-L1, SSO, HEO and MEO for the launches made under a payload mass of maximum 4000 but over 0 we see that they hold a success rate of 100% based on a total of nine launches made between that range to those four orbits.

Launch Success Yearly Trend



We can observe that the success rate since 2013 kept increasing till 2020.


SQLite

All Launch Site Names

```
[15]: %sql select distinct(Launch_Site) from SPACEXTBL
 * sqlite:///my_data1.db
Done.
```

t[15]:

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

The select distinct statement allows us to choose unique names or values of a dataframe with the name of the specific column specified in the parentheses. The from command specifies from what dataframe, and SPACEXTBL is simply the table from which the information was obtained. This query displays all unique launch site names.

Launch Site Names Begin with 'KSC'

In [27]:
```
%sql select * from SPACEXTBL where Launch_Site like "KSC%" limit 5
```

```
 * sqlite:///my_data1.db
Done.
```

Out[27]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
19-02-2017	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
16-03-2017	06:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
30-03-2017	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
01-05-2017	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO		Success
15-05-2017	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5	6070	GTO	Inmarsat		

The select * from SPACEXTBL picks all elements of a table, and that is why a where clause is used to specify that we want all elements of the table that in column of launch sites fulfills a given condition. The like command is used in a where clause to search for a specified pattern in a column, a start of a name is passed between quotation marks with a percentage sign after the last letter to specify that it is the start of a name and we do not know in what it ends. Limit 5 is passed down to show that we want only to display 5 records. This query displays 5 records where launch site name begins with the string 'KSC'.

Total Payload Mass

NASA (CRS)

```
n [43]: %sql select sum(PAYLOAD_MASS__KG_) as sum from SPACEXTBL where Customer like "NASA (CRS)"
 * sqlite:///my_data1.db
Done.
ut[43]:
```

sum
45596

The select statement followed by the sum() function displays the total amount of a numeric column specified in the parentheses. I decided to use the as command to rename a column or a table with an alias of sum, this is to make the column more readable, an alias exists only for the duration of the query. As always the table from where we want it is determined, a where clause is used, we make clear that from the customer column where the client is NASA (CRS) we want the total amount of payload carried by boosters launched by NASA (CRS). This query displays the total payload mass carried by boosters launched by NASA (CRS).

Average Payload Mass by F9 v1.1

```
[48]: %sql select avg(PAYLOAD_MASS__KG_) as "average payload" from SPACEXTBL where Booster_Version like "%F9 v1.1%"
 * sqlite:///my_data1.db
Done.
t[48]:  average payload
        2534.6666666666665
```

The avg() function returns the average value of the numeric column passed. Again I decided to use the as command, nevertheless I had to use quotation marks as if I didn't use it for two separated words it would have resulted in error. Table is specified where in the Booster_Version column the booster name was like F9 v1.1. I decided to use one percentage sign a the start and one at the end although this would mean that that particular thing would be in the middle of something, because if that is the name alone it wouldn't make any difference and it would simply retrieve the name of the booster version, in this case we knew that it was a name alone and that no other names had it in the middle which means that it did not retrieve any other booster. This query displays the average payload mass carried by booster version F9 v1.1.

First Successful Ground Landing Date

[79]:
```
%sql select min(Date) as DATE from spacextbl where "Landing _Outcome" like "Success (drone ship)"
```

```
 * sqlite:///my_data1.db
Done.
```

t[79]:

DATE
06-05-2016

We selected the minimum date as date where the column of landing outcome which in this case quotation marks had to be used to avoid error was like success for drone ship. Quotation marks that contain column names make no difference in the picking of the column, in this case it had to be used. This query lists the date where the successful landing outcome in drone ship was achieved.

Successful Drone Ship Landing with Payload between 4000 and 6000

In [90]:
```
%sql select Booster_Version as Name, PAYLOAD_MASS__KG_ as "Payload Mass(kg)" from spacextbl \
where (PAYLOAD_MASS__KG_ between 4000 and 6000) and ("Landing _Outcome" like "% (ground pad)")
```

```
 * sqlite:///my_data1.db
Done.
```

Out[90]:

Name	Payload Mass(kg)
F9 FT B1032.1	5300
F9 B4 B1040.1	4990
F9 B4 B1043.1	5000

This is a subquery, it basically works by using a first subquery that selects the booster version which I thought was appropriate to select as Name, the payload mass also with a name that I saw fit, then from the table from which I wanted all of this, and now is when we state a where clause and then is when another 2nd subquery which contains the payload masses that are between 4000 and 6000kg and also a landing outcome which ended in something like (ground pad) is written. It is basically a query that takes information of another query. This subquery lists the names of the boosters which have success in ground pad and have a payload mass greater than 4000 but less than 6000 in kilograms. I saw fit to include the listing of the payload mass next to the booster version name.

Total Number of Successful and Failure Mission Outcomes

In [91]:
```
%sql select mission_outcome, count(*) from spacextbl \
group by mission_outcome \
order by mission_outcome
```

```
* sqlite:///my_data1.db
Done.
```

Out[91]:

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

In this query we selected all the outcomes of the mission, we selected the total count of all the mission outcomes selected by passing mission outcome before count(*) from the table that we want and we grouped by mission outcome and then ordered all by them. This query lists the total number of successful and failure mission outcomes.

Boosters Carried Maximum Payload

```
In [82]: %%sql select Booster_Version as Name from spacextbl where PAYLOAD_MASS__KG_ in \
         (select max(PAYLOAD_MASS__KG_) from spacextbl)

 * sqlite:///my_data1.db
Done.
Out[82]:
```

Name
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

We select the booster version as name from the spacextbl where the payload mass in the maximum payload mass was selected from the spacextbl. This is another subquery. This subquery lists the names of the booster versions which have carried the maximum payload mass.

2017 Launch Records

In [112]:
```
%sql select substr(Date,4,2) as month, "Landing _Outcome", booster_version, launch_site from spacextbl \
where ("Landing _Outcome" like 'Success (ground pad)') and (substr(Date,7,4)='2017')
```

```
 * sqlite:///my_data1.db
Done.
```

Out[112]:

month	Landing _Outcome	Booster_Version	Launch_Site
02	Success (ground pad)	F9 FT B1031.1	KSC LC-39A
05	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
06	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
08	Success (ground pad)	F9 B4 B1039.1	KSC LC-39A
09	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A
12	Success (ground pad)	F9 FT B1035.2	CCAFS SLC-40

We select month as month, the landing outcome, the booster version and the launch site from spacextbl where the landing outcome is like Success (ground pad) and the year is equal to 2017. Substr(Date,4,2) and substr(Date,7,4) represent month and year respectively as Sqlite does not support monthnames. This subquery lists the records which display the month names, successful landing outcomes in ground pad, booster versions and launch site for the months in year 2017.

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

```
In [116]: %sql select "landing _outcome", count(*) as count from spacextbl where date between "04-06-2010" and "20-03-2017" \
          group by "landing _outcome" order by count(*) desc

 * sqlite:///my_data1.db
Done.
```

Out[116]:

Landing _Outcome	count
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

We select landing outcome and the count for all landing outcomes from the spacextbl where the date is between 04-06-2010 and 20-03-2017 and the we group by landing outcome and then order by count of each in descending order. This query ranks the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order.


Section 3
Launch Sites
Proximities Analysis

Launch Site Location

- It is evident in the image that launch sites are situated in the coastline. It is good to remember that they are testing rockets, therefore in a hypothetical fall or error which would require the rocket to be blown up, being near water would come in handy.

- SpaceX attempts rocket landings on regions of the ocean as well as on drone ships which are located on water, the reason for this is that in the future if ships weren't to have enough fuel to attempt a landing on a ground pad, then it could be done on water be it directly or on a drone ship. Also ground pads are located close to the coast.

Launch Outcomes

- First site: In here we see that KSC LC-39A has a total of ten successful landings and three failed. Below.

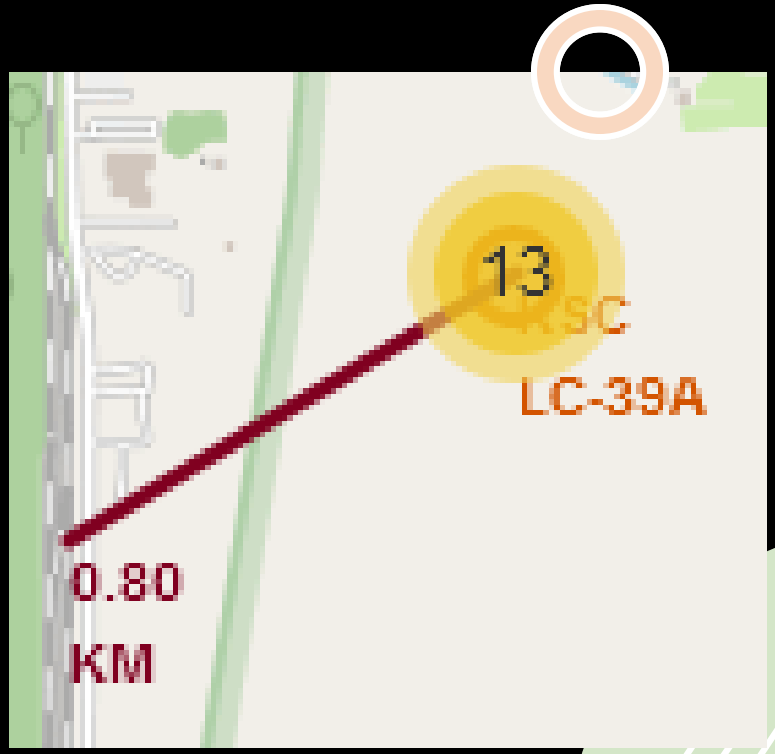
- Second site: At the top right we see two pictures of two sites(CCAFS SLC-40 and CCAFS LC-40) that have three successful outcomes versus four failed outcomes and seven successful outcomes versus nineteen failed outcomes.

- Third site: VAFB SLC-4E located close to the Vanderberg Space Force Base has four success outcomes versus six failed outcomes.

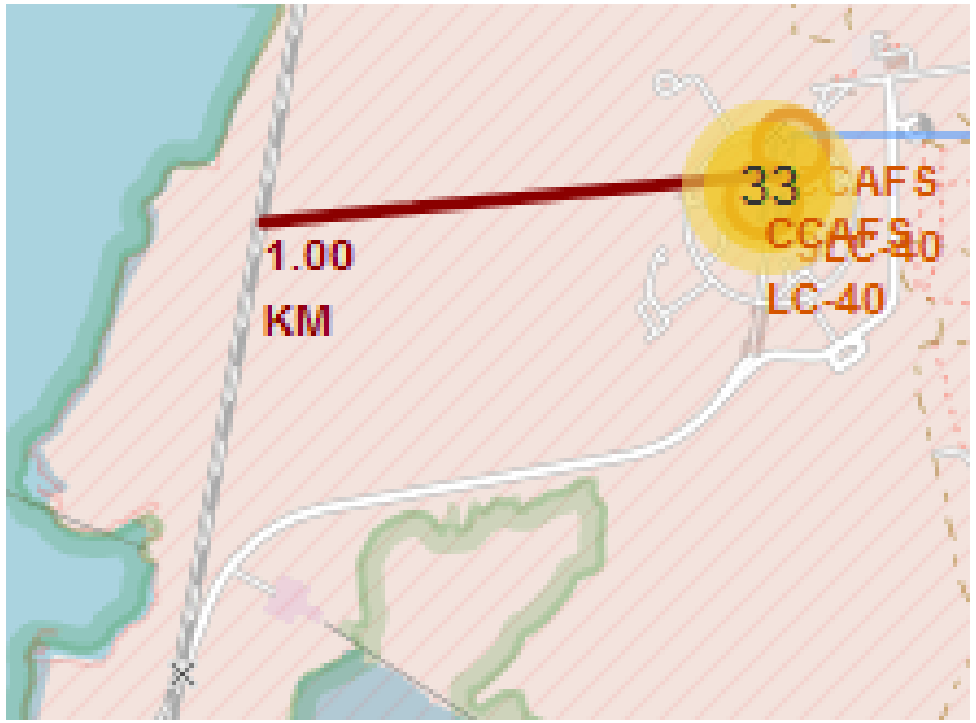
Distance to Railways

VAFB SLC-4E
Distance to Railway

KSC LC-39A Distance to Railway



CCAFS SLC-40 and CCAFS LC-40 Distance to Railway



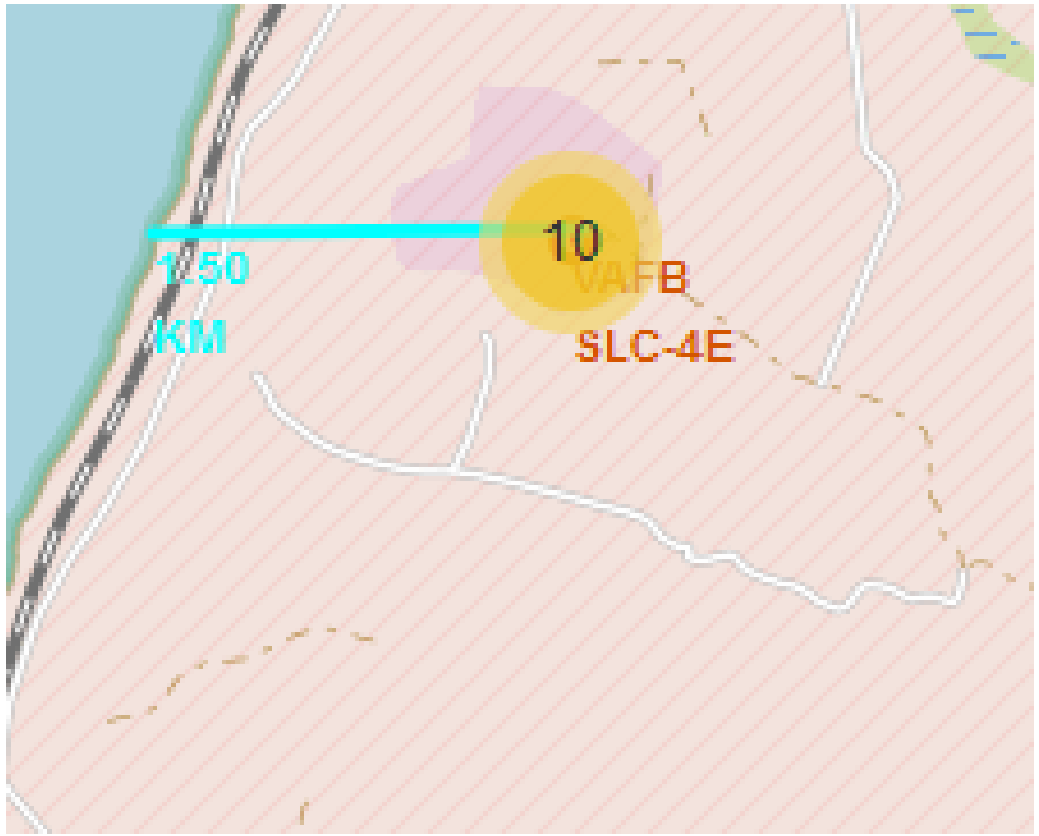
Summary

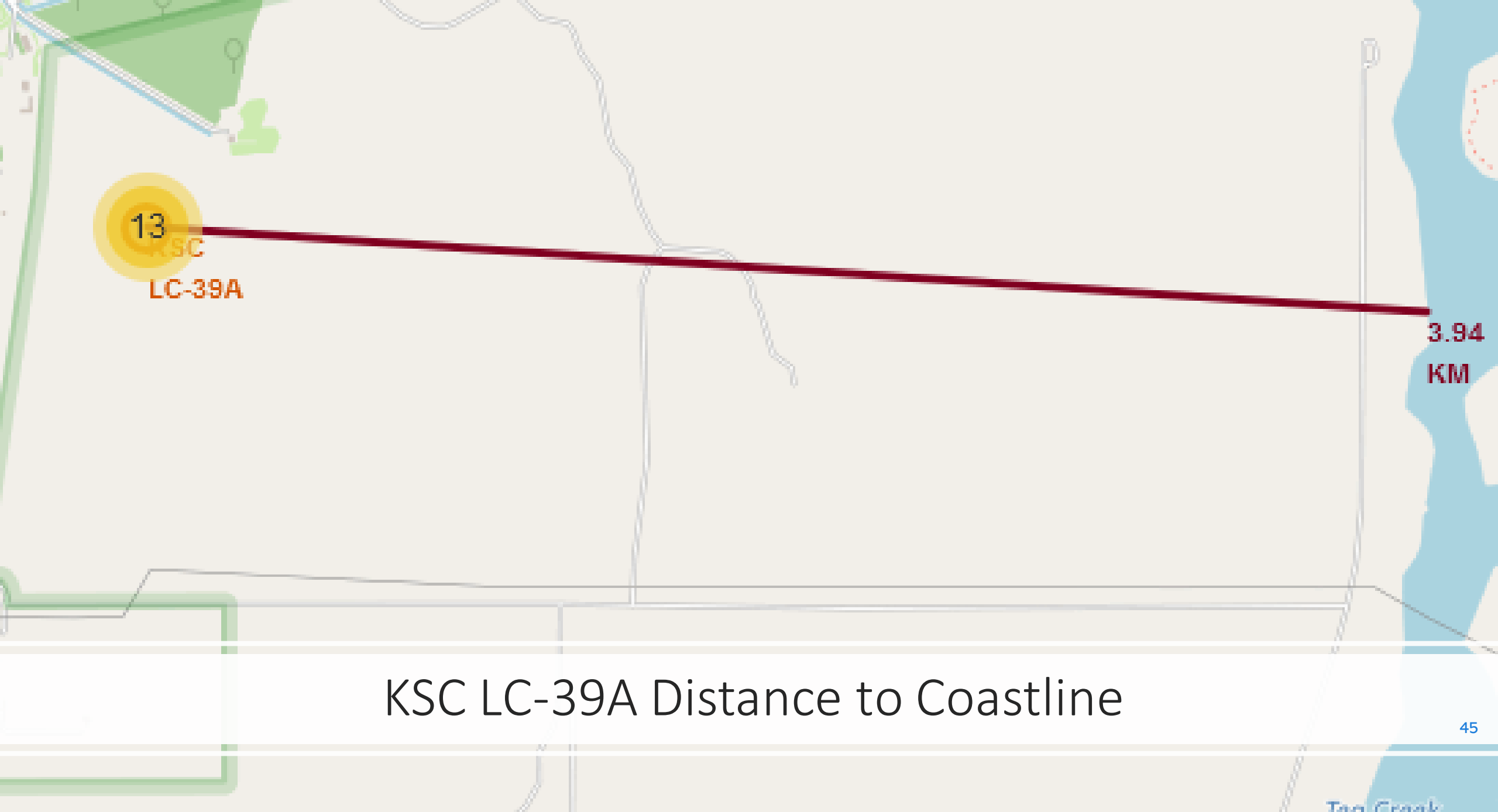
- VAFB SLC-4E is at 1.37 km from the closest point along the railway.
- CCAFS SLC and LC-40 are closer to the railway at specifically 1 km.
- KSC LC-39A is at 0.80 km being the most proximate to its respective railway.

Distance to

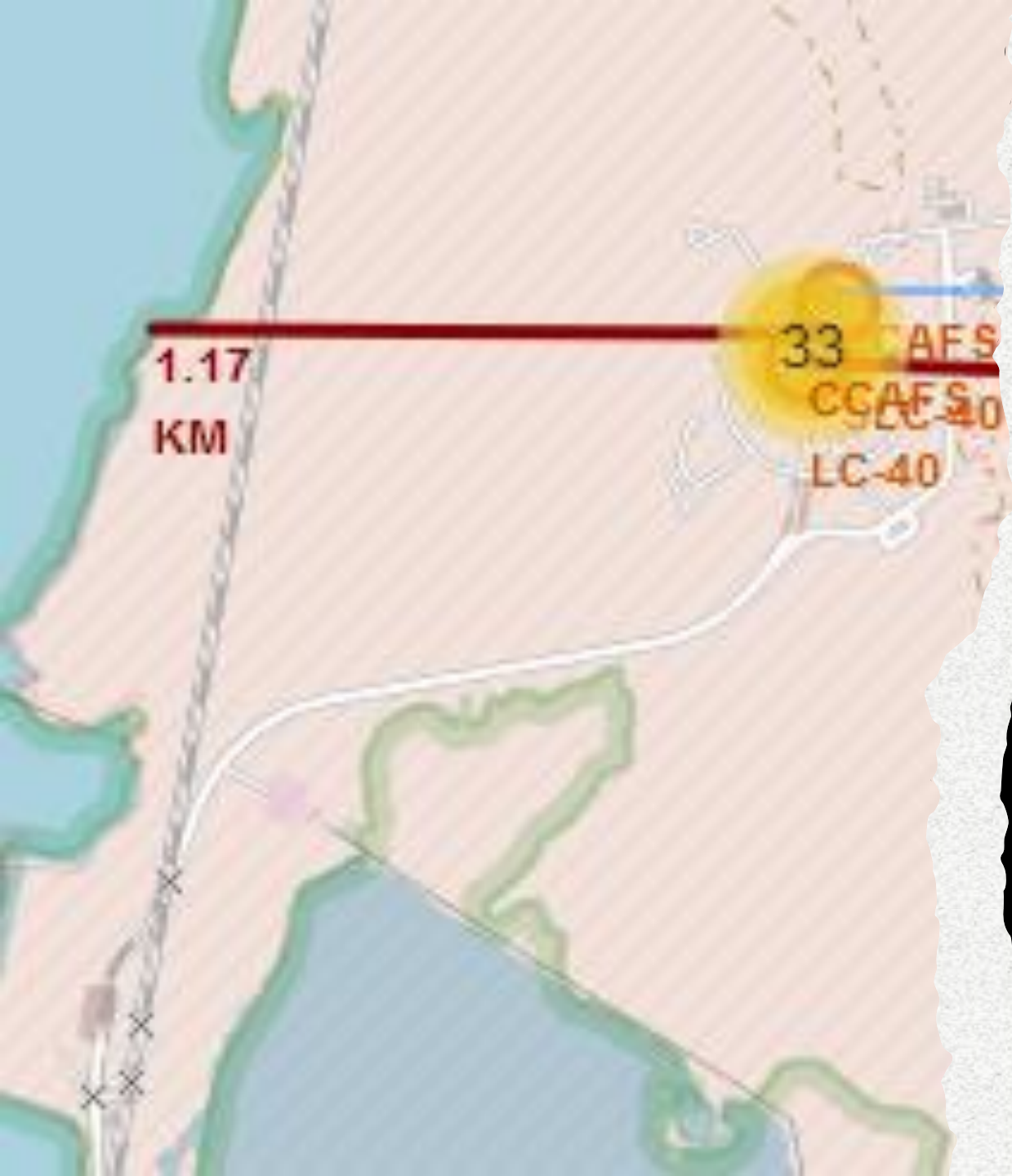
Coastlines

VAFB SLC-4E Distance to Coastline

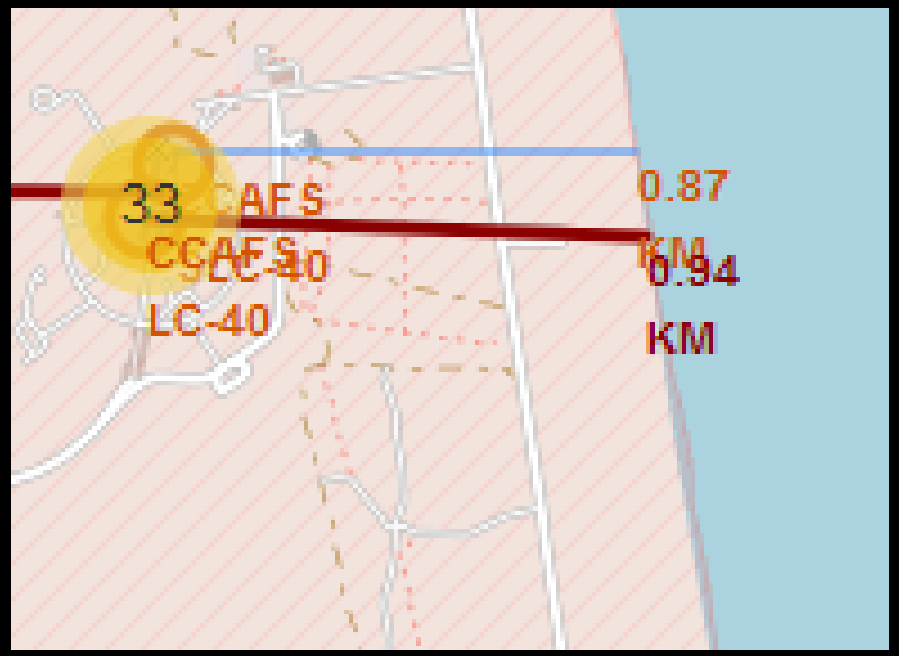




KSC LC-39A Distance to Coastline



CCAFS SLC and LC-40 Distance to Coastline



CCAFS SLC-40 Distance to Coastline

against

CCAFS LC-40

Summary

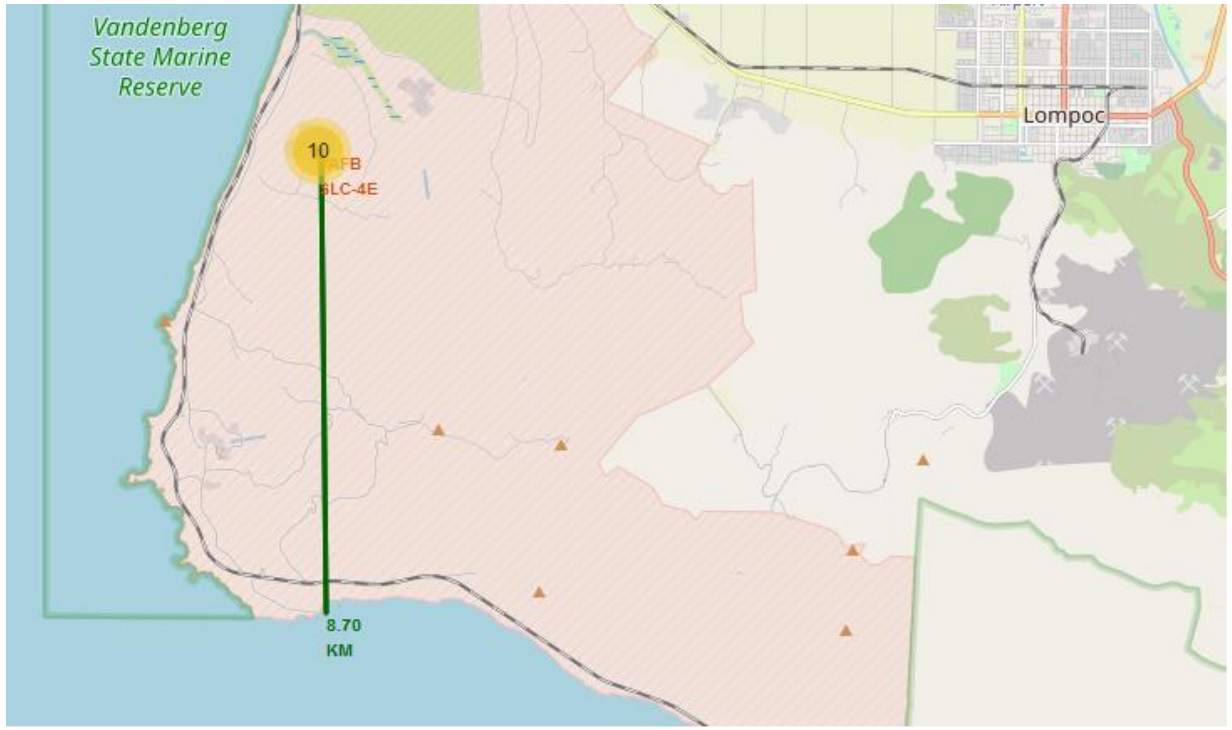
- VAFB SLC-4E a site associated with 67% success rate, is at 1.50 km. Actually CCAFS SLC and LC-40 have a closer distance to both of their coastlines, they are also associated with about 61% of all launches.

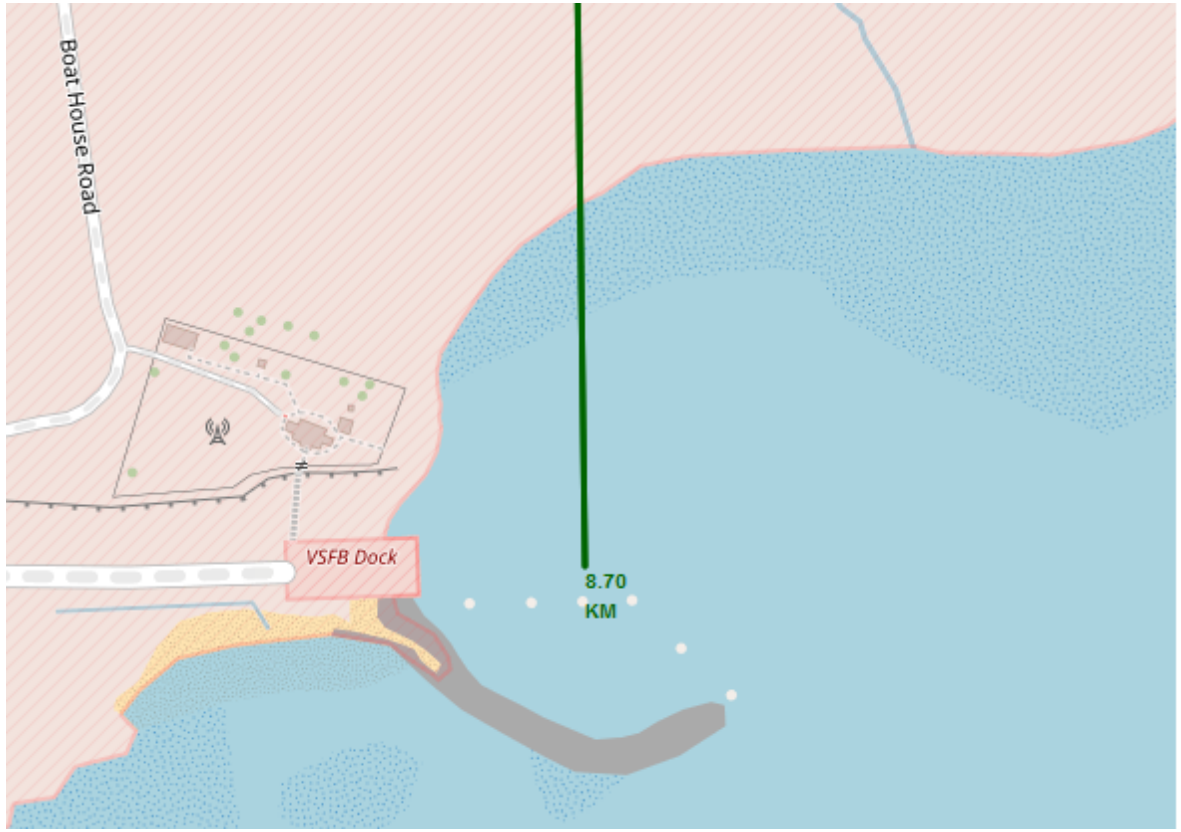
- KSC LC-39A is the furthest to the coastline, it has the highest success rate.

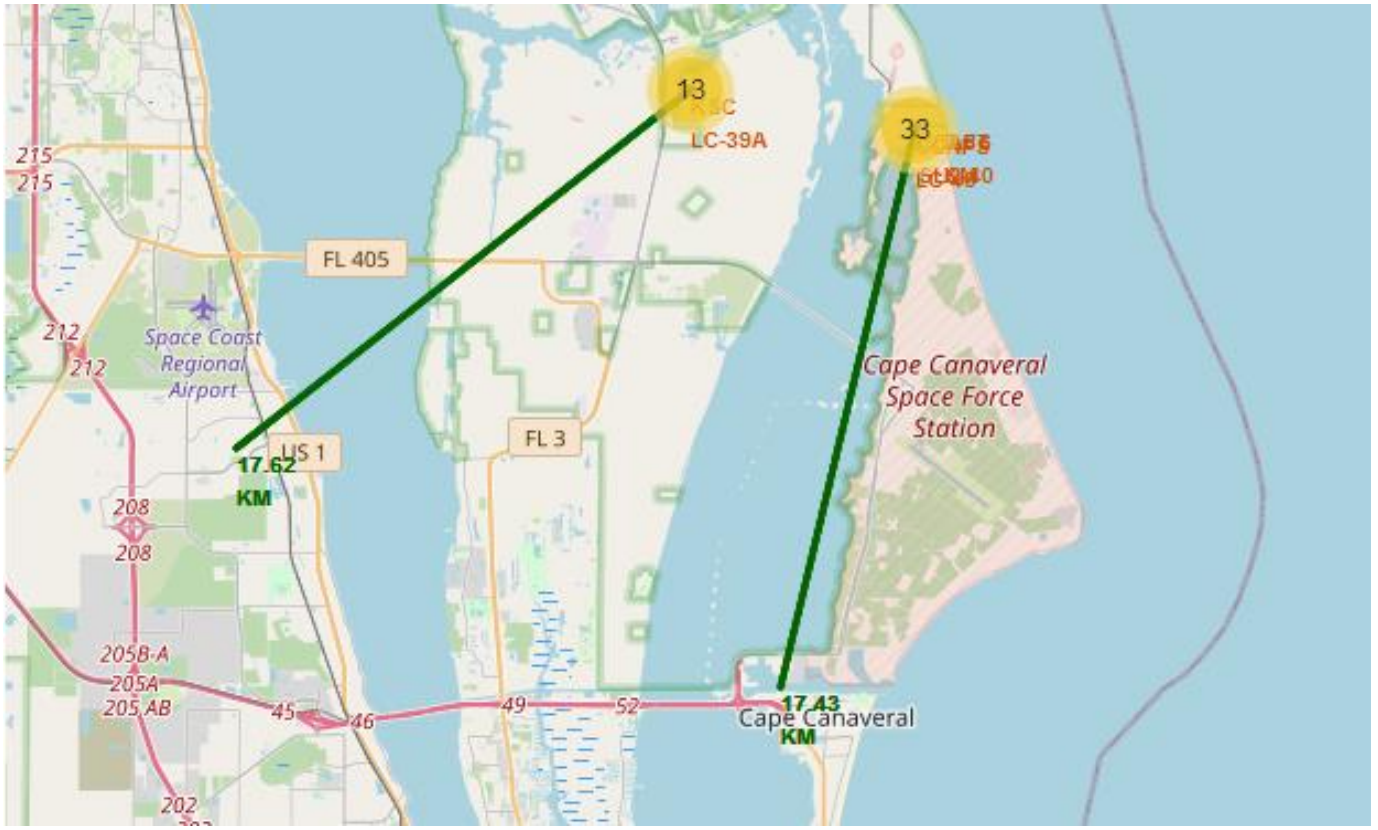
- It is likely that they have used the sites closer to the coastline for trial and error and left this last site for actual testing.

Distance to Ports and Dock

VAFB SLC-4E Distance to Dock







- KSC LC-39A, CCAFS SLC and LC-40 Distance to Ports

- Summary

  - Being close to a port or dock can be advantageous as sites could benefit from acquiring resources that could travel on ships being cheaper than road.

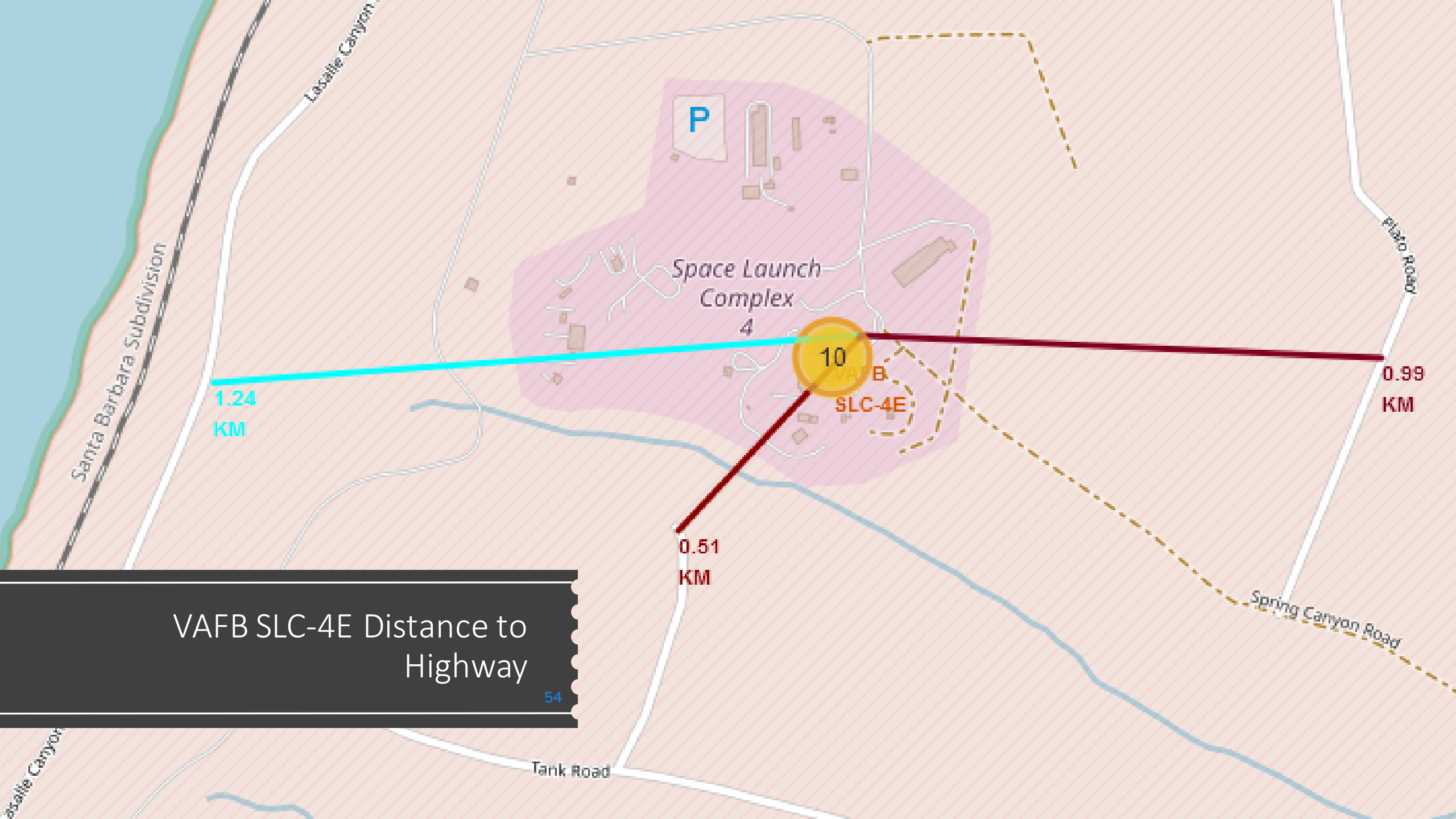
  - VAFB doesn't have a port close only one dock. It is located at 8.70 km being the more close to a place where a ship can berth.

  - KSC is furthest to such a place, its distance to Port St John is 17.62 km.

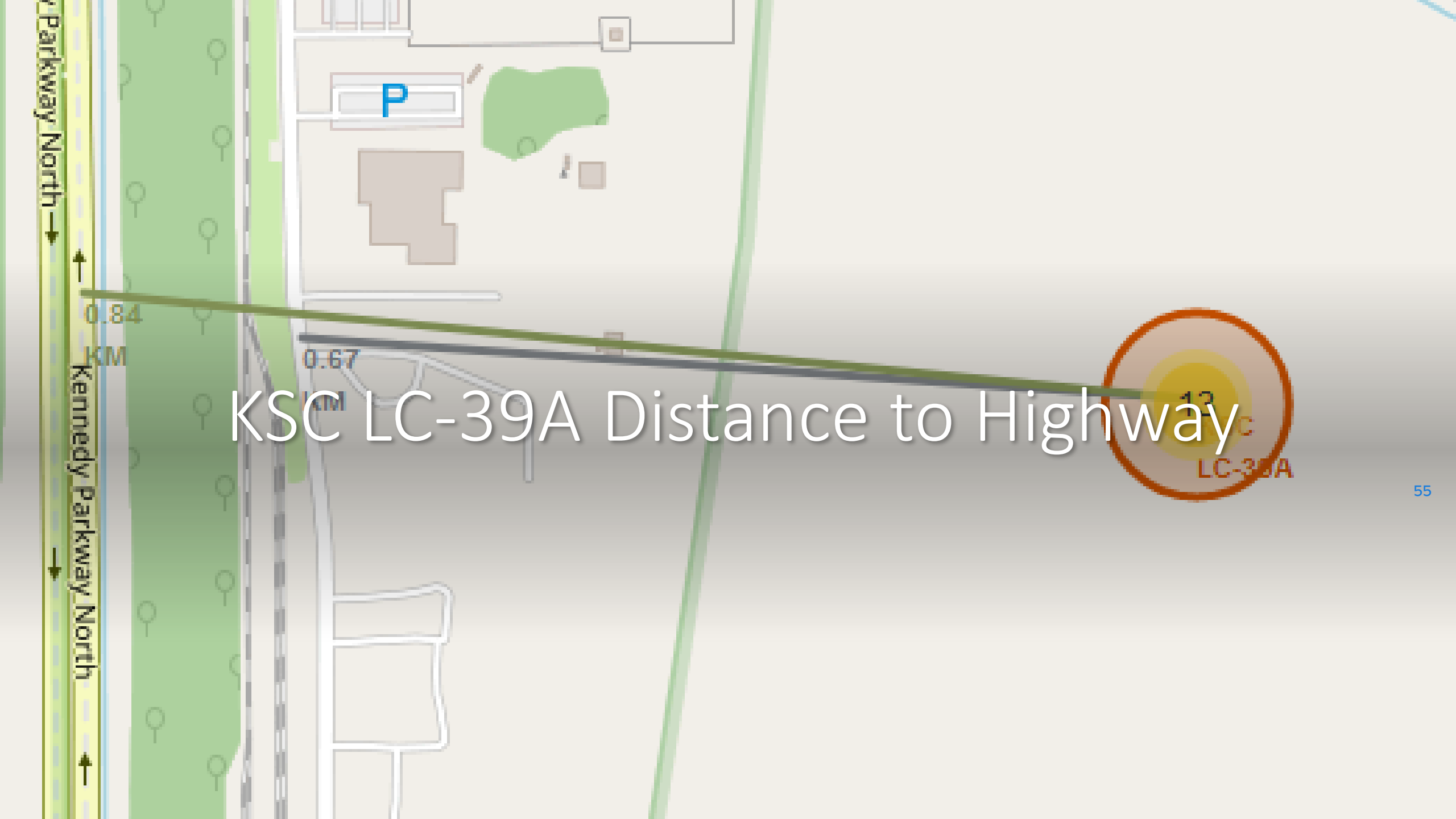
  - The distance of CCAFS is 17.43 km.

  - The Port Canaveral district is located just north of the city of Cape Canaveral. It is one of the busiest cruise ports in the world.

Distance to Highways



VAFB SLC-4E Distance to Highway



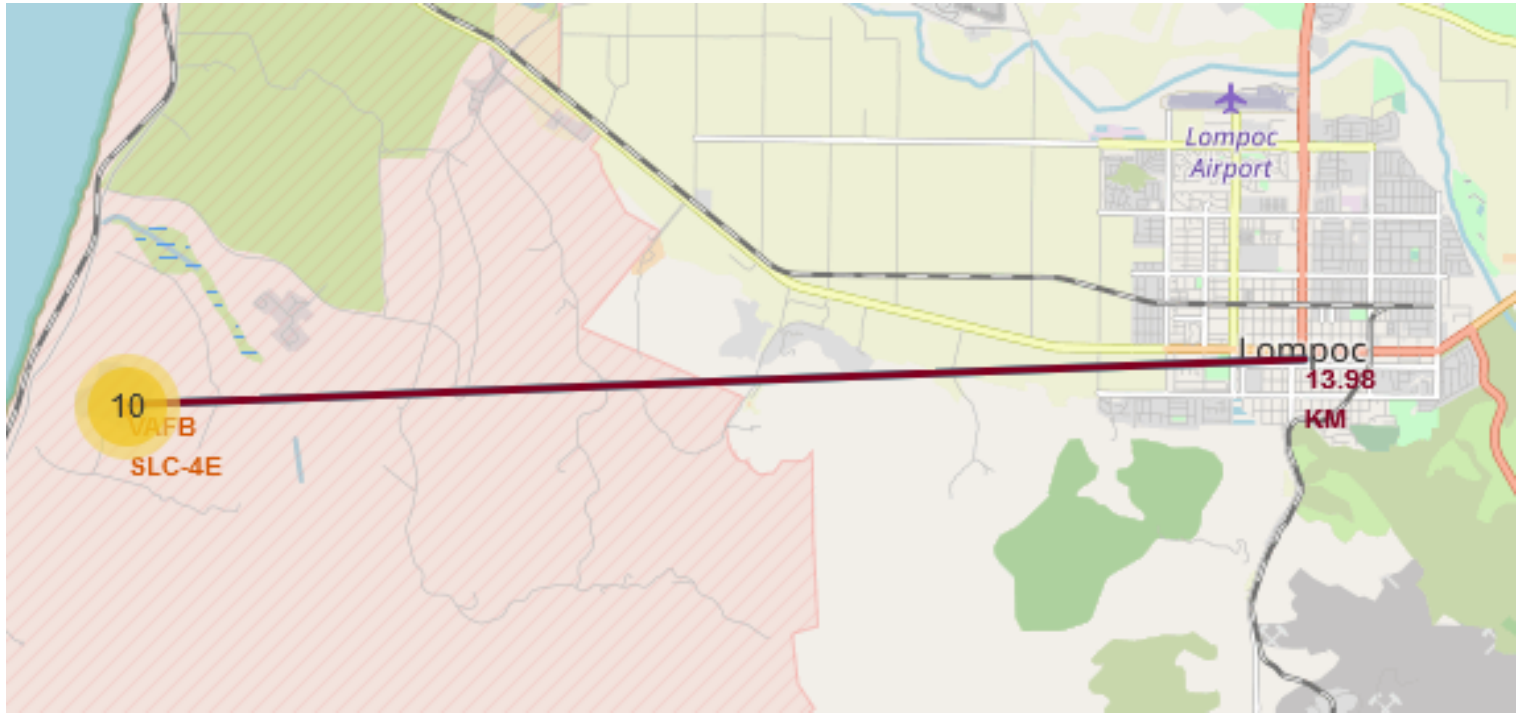
KSC LC-39A Distance to Highway

CCAFS SLC-40 and CCAFS LC-40 Distance to Highway

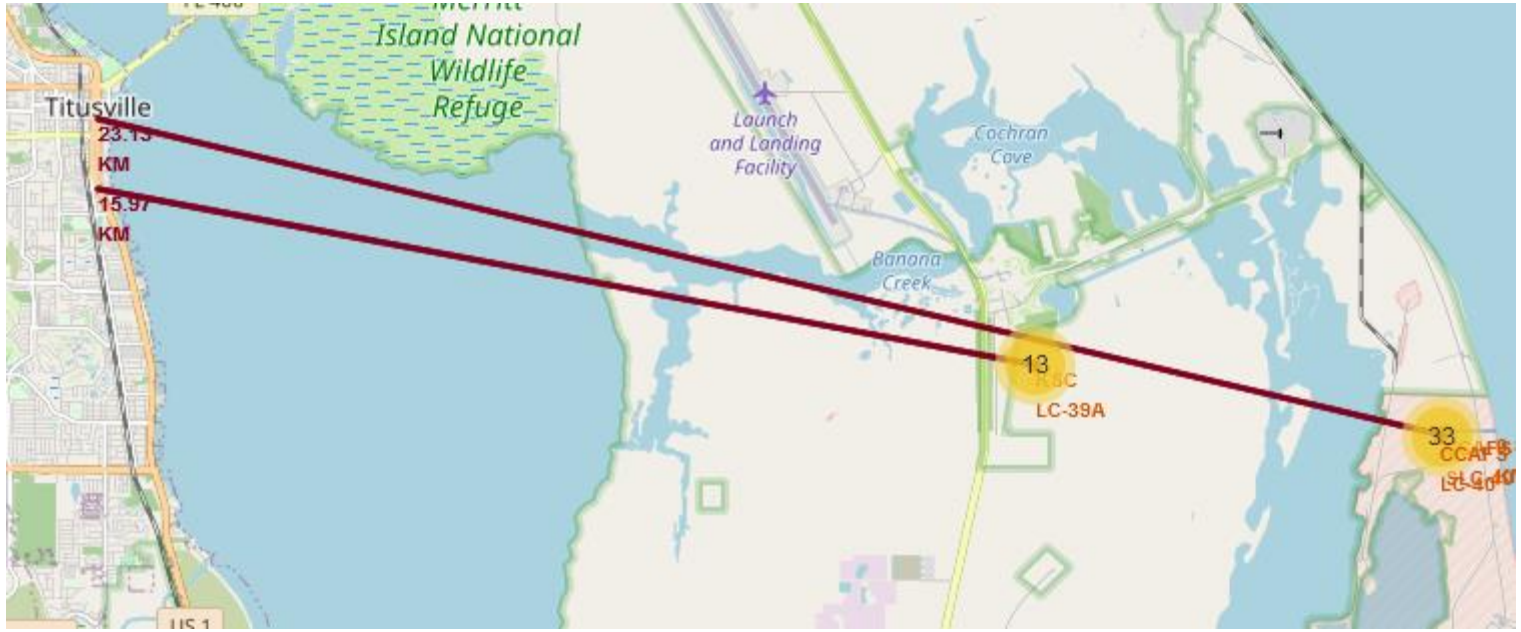
- Summary
- Most sites have at least three highways in close proximity.
- Highways usually go in and out of cities.
- Many of the distances are very close to a kilometer but don't quite make it to it.
- VAFB SLC-4E has more distance to most of its highways overall with two out of three having 1.24 km And one having 0.99 km.
- CCAFS SLC and LC-40 are closer to their highways, they both have better launch rates.
- KSC LC-39A is very close to its highways to a maximum of 0.84 km and a minimum of 0.67 km, also it has the highest success rate of the four.
- Sites have to be around 1 km.

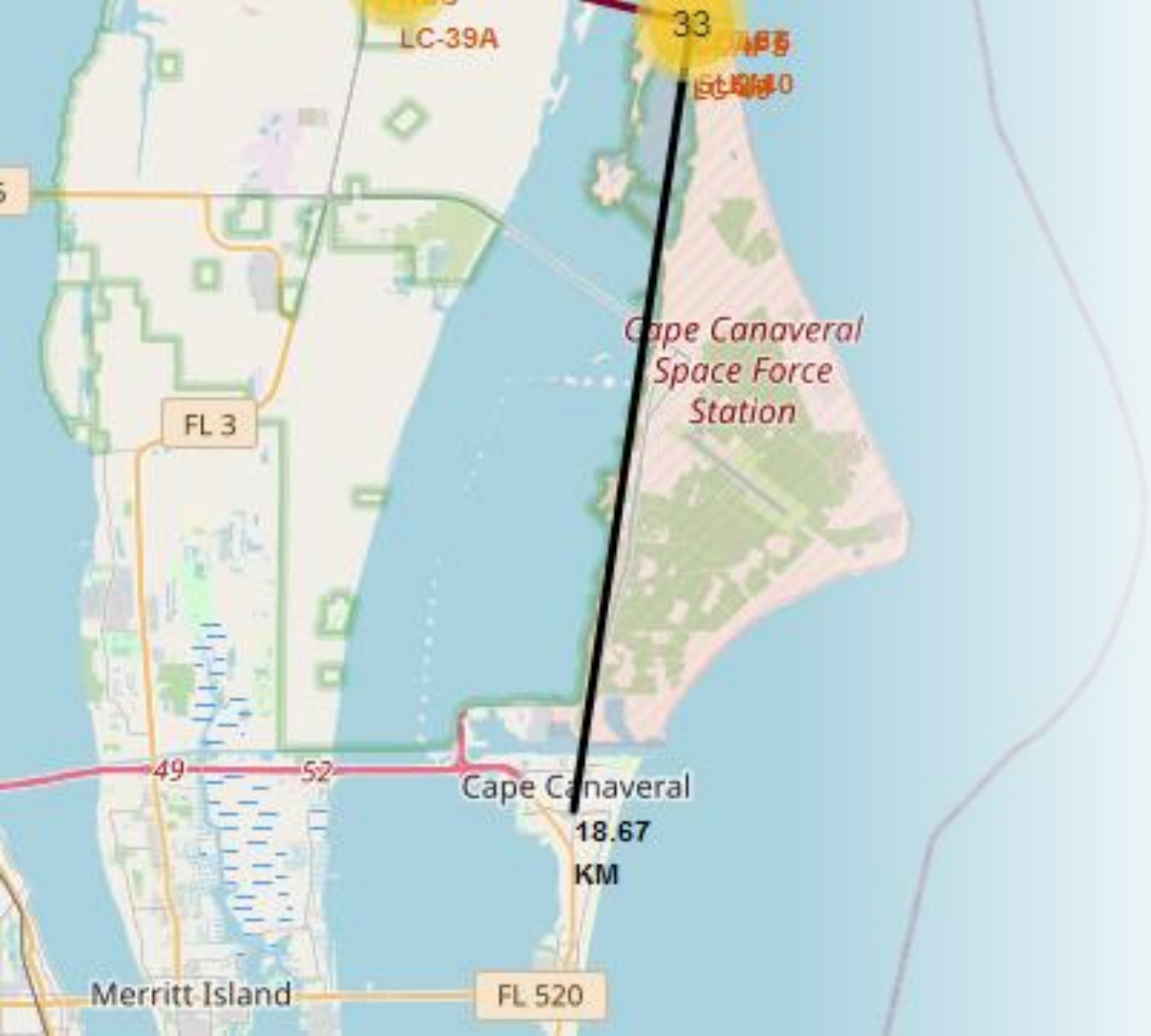
Distance to Cities

VAFB SLC-4E Distance to City



KSC LC-39A, CCAFS SLC-40 and CCAFS LC-40 Distance to City



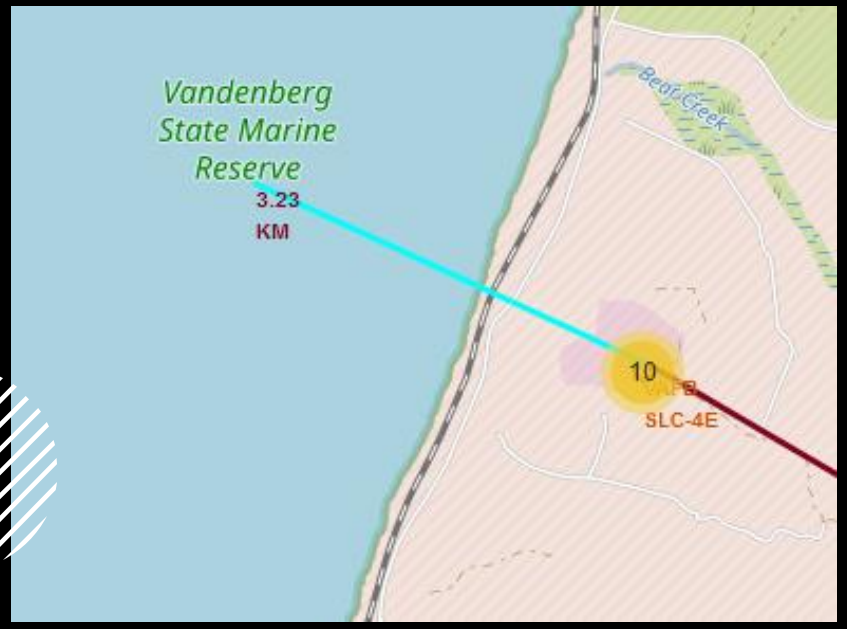


- CCAFS SLC and LC-40 Distance to Cape Canaveral (city)

Summary

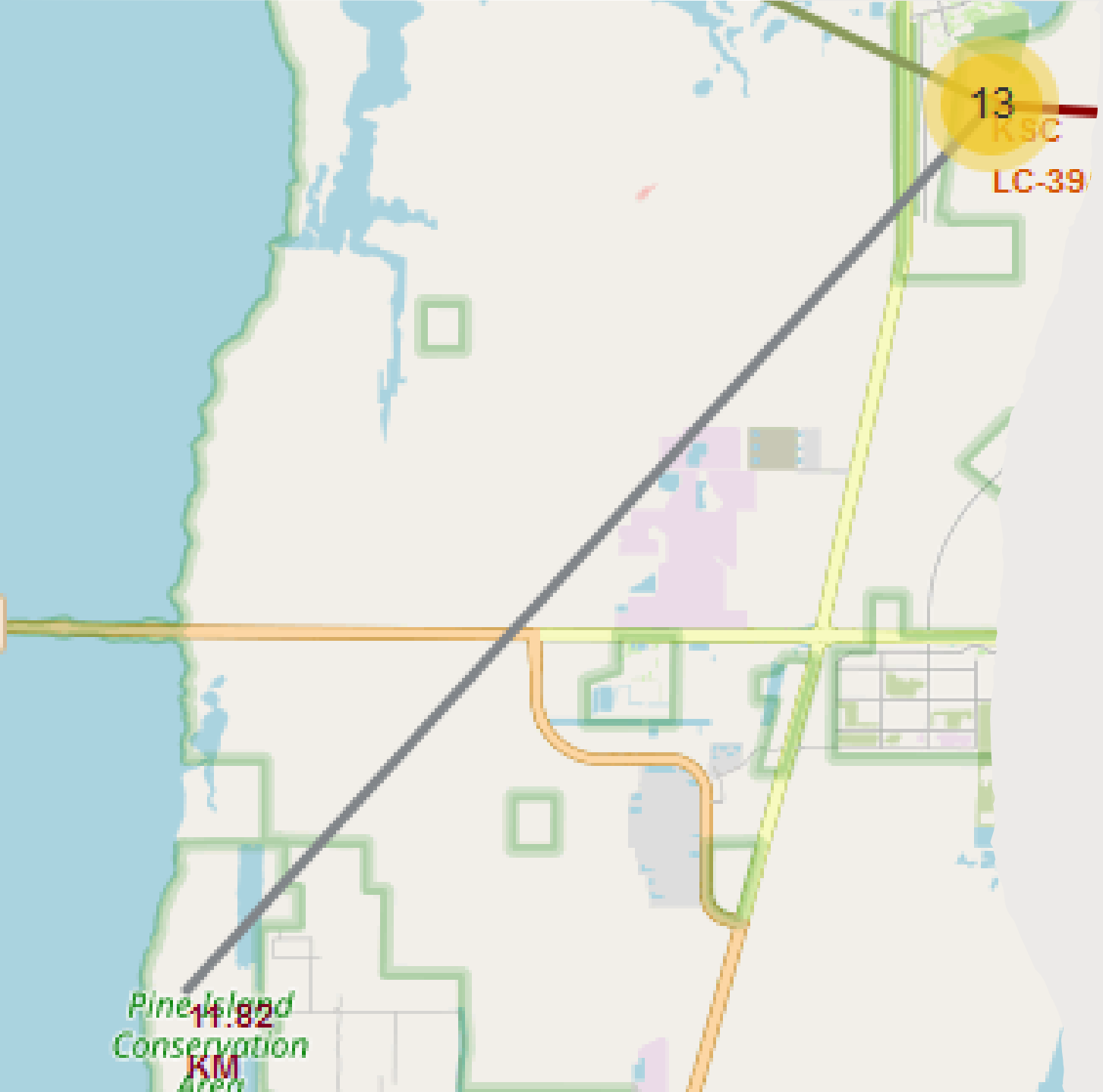
- Half of the sites did not reach 16 km, but the other half did not exceed 23.12 km.
- All sites were under 30 km to the city.
- A city is a densely populated area, keeping distance from it would be wise as to not endanger the population when testing a rocket though not too much distance as you still need some of the services a city might have.
- The United States Census Bureau has designated the Vandenberg Space Force Base as a separate-census designated place (CDP) for statistical purposes, spanning the base's residential population. As of the 2020 census the population was of 3,559.
- The Kennedy Space Center is located on Meritt Island which has a population of approximately 34,518 from the 2020 census.
- The city of Cape Canaveral lies just south of the Port Canaveral District. As of 2020 it has a population of 9,972.
    - Port Canaveral is one of the busiest cruise ports in the world.
- The management of KSC and CCAFS work very closely together, share resources and operate facilities on each other's property.

Wildlife Protected Areas

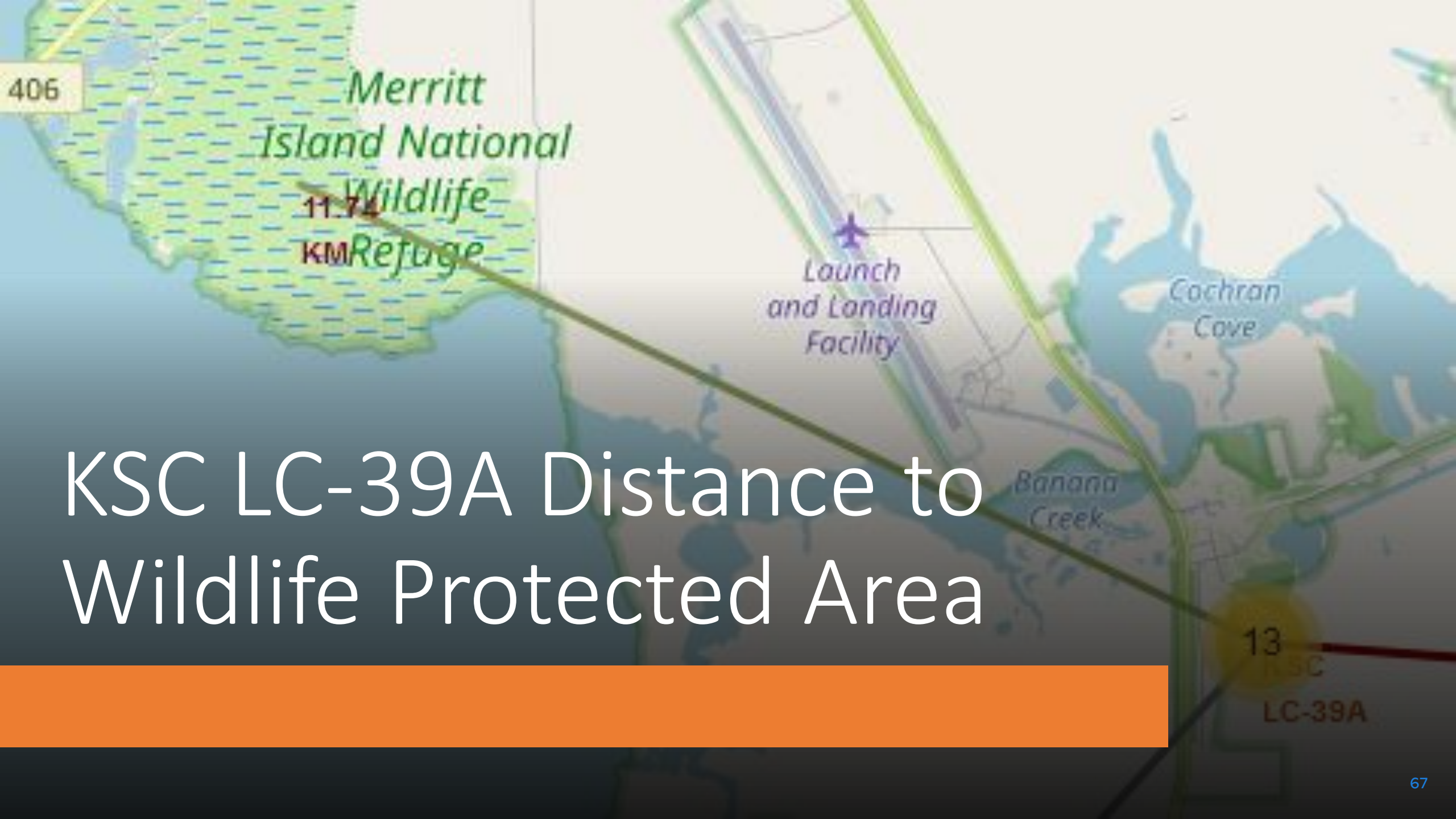


- VAFB SLC-4E Distance to Wildlife Protected Area

- KSC LC-39A Distance to Wildlife Protected Areas



- KSC LC-39A Distance to Wildlife Protected Area



KSC LC-39A Distance to Wildlife Protected Area

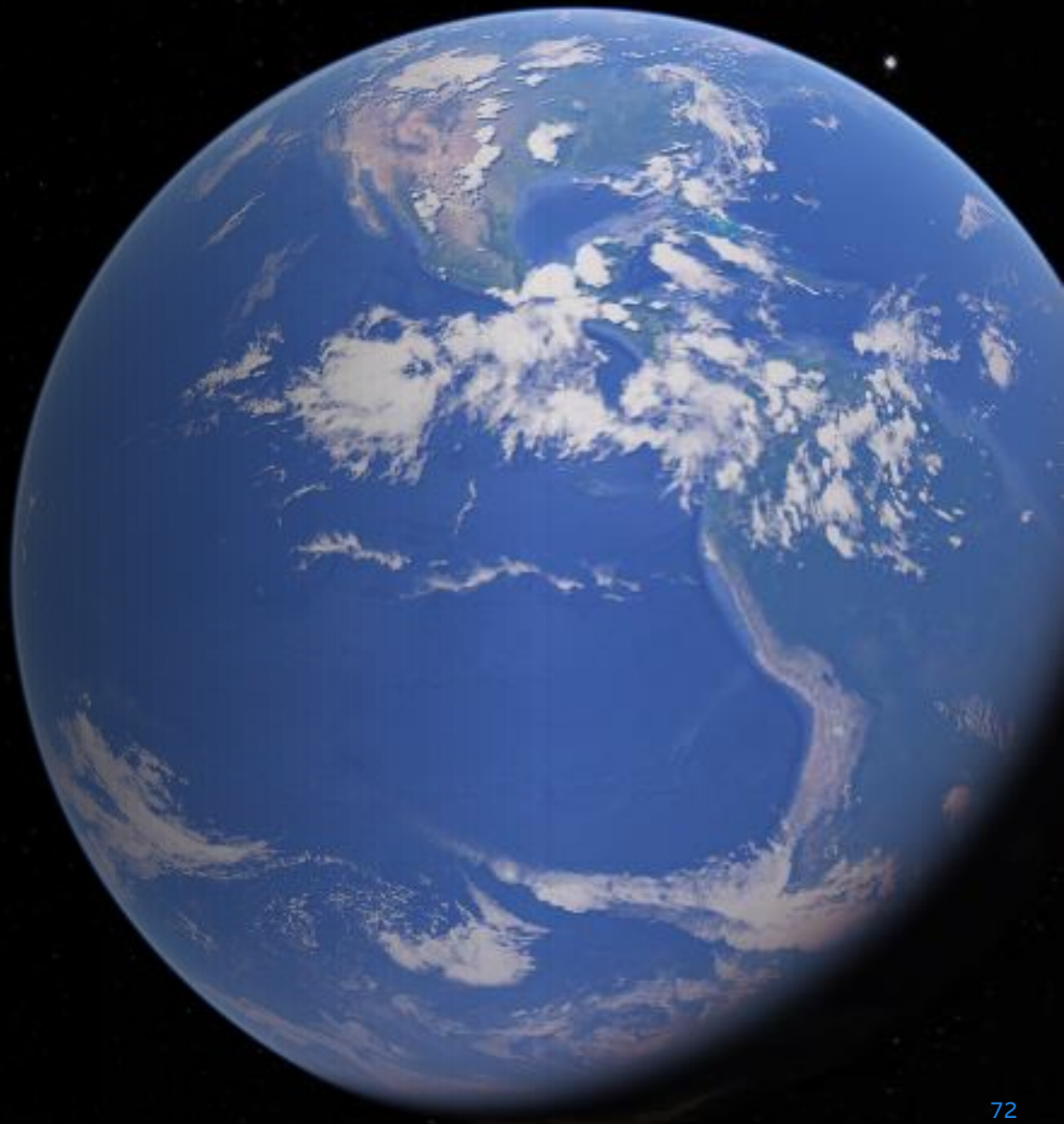
KSC LC-39A Distance to Wildlife Protected Area

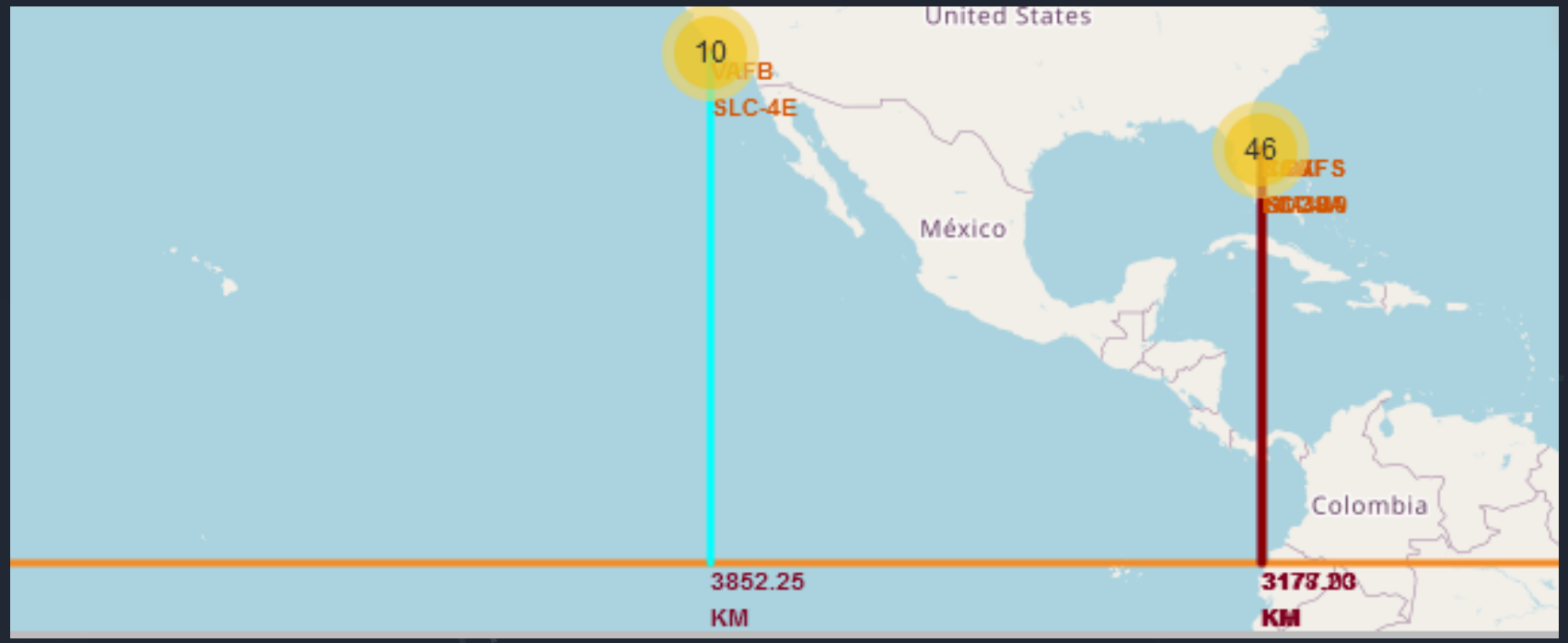
CCAFS SLC-40 and CCAFS LC-40 Distance to Wildlife Protected Area

• Summary

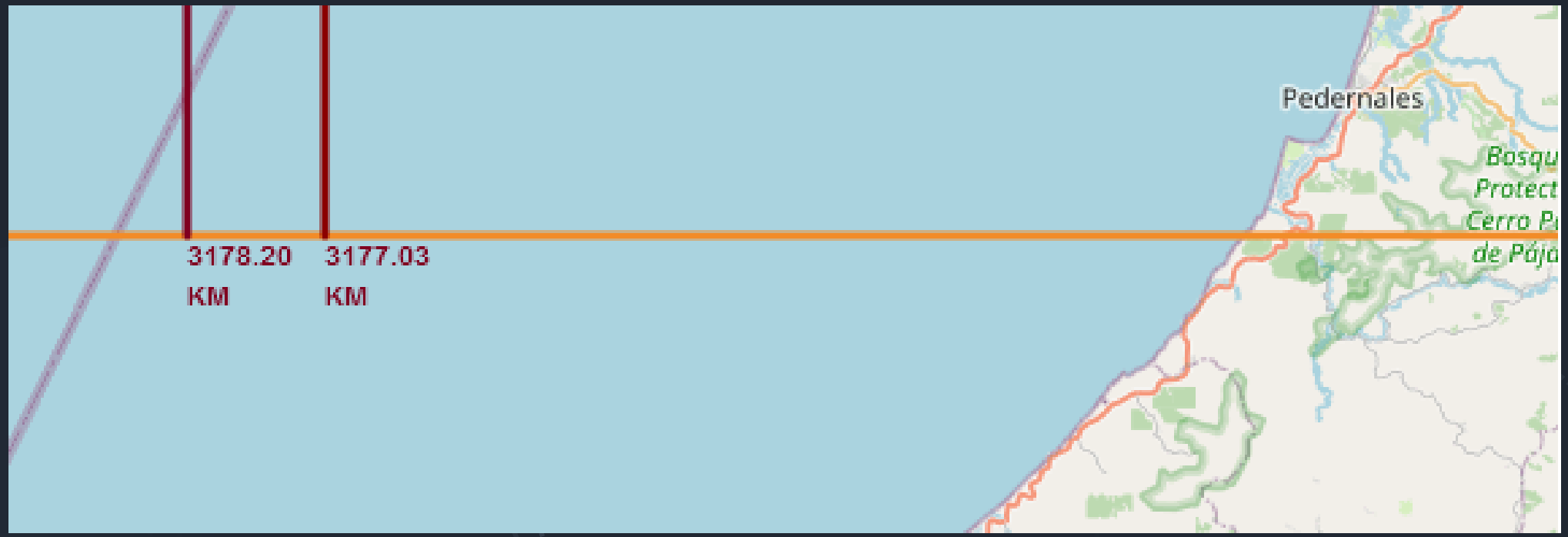
- All sites are within 2 km up to 13 km of distance to a wildlife protected area.
- Basically, launch sites have an area where only authorized people can enter, there aren't cars everywhere and etc... This are safety measures for the population. Sometimes it is possible to visit this. By themselves have restrictions on visitors.
- Protected areas of wildlife work similar in that only a given few can enter, usually this is to carry out a role. Sometimes it is possible to visit this also. By themselves have restrictions on visitors.
- The piece of land is the same, both need an area that they can control. Therefore it is not rare to find one or more of this areas close to a launch site independently of who was there before as they both require a similar place.
- The United States Census Bureau has designated the Vandenberg Space Force Base as a separate-census designated place (CDP) for statistical purposes, spanning the base's residential population. As of the 2020 census the population was of 3,559.
- VAFB can restrict access to sections of beaches during the nesting season of the Western Snowy Plover. Fail to respect the closing a number of times can lead to the entirety of the beaches getting closed.
- Only one beach of three beaches is open to the public while the other two are restricted to people with regular access to the base.
- The impact between a rocket launch and the wildlife close to such a site is not clear.

Distance to the Equator





- The Distance of four sites to the equator, we see that VAFB SLC-4E is at a distance of 3852.25 km of the equator



- The distance of three sites to the equator. We see that in the first line which corresponds to KSC LC-39A site, it is at a distance of 3178.20 km to the equator. The second line corresponds to sites CCAFS SLC-40 and CAFFS LC-40, they both are at 3177.03 km to the equator.

Summary

- Most launch sites are located near the equator. This has the advantage of the earths spin that is maximum at the equator. This means that it is cheaper to launch an object and put it into orbit as less fuel is needed which signifies cheaper rockets and more heavier payloads.
- Cape Canaveral is closer to the equator, all sites except VAFB SLC-4E are present at Cape Canaveral.
- A rocket launched from VAFB SLC-4E from east to west is at retrograde orbit and would need additional velocity to counteract earth's rotation. This perfectly translates in more fuel consumption and more expensive rockets, which means more expensive launches.

Launch Sites Proximities Analysis Summary

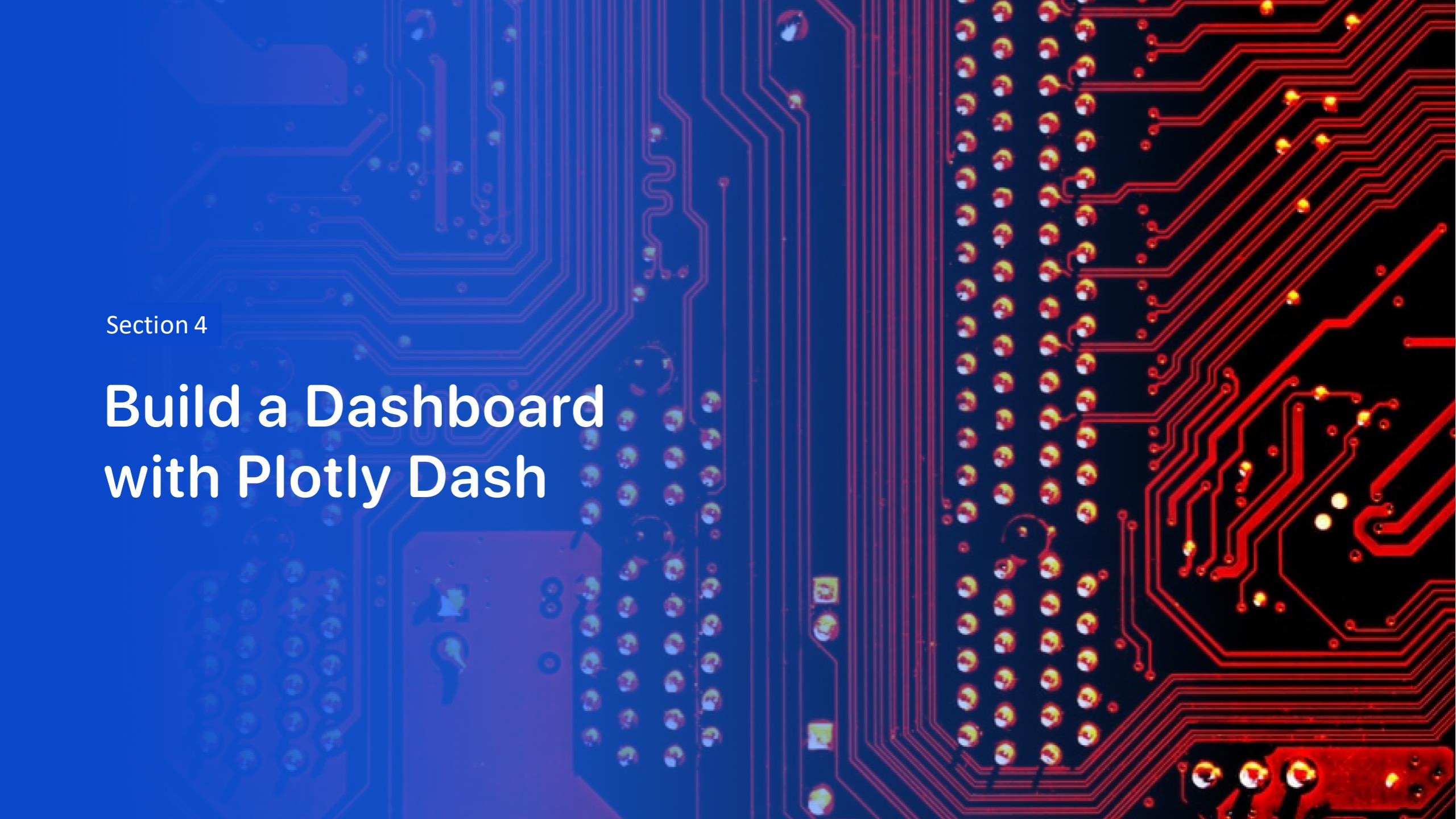
- KSC LC-39A is the closest to its respective railway.
- KSC LC-39A is the furthest to the coastline, it has the highest success rate.
- It is likely that they have used the sites closer to the coastline for trial and error and left this last site for actual testing.
- Most sites are present at the eastern coast.
- Launching a rocket to the east gives an additional boost to the rocket as well as minimizes fuel consumption.
- All sites are within 2 km up to 13 km of distance to a wildlife protected area.
- Launch sites and some wildlife protected areas require similar places that are considerably far from population but not too much.
- Also they both have in common that they are usually restricted to the general public. Nevertheless they are both able to have visitors during visiting hours, they are given a tour of specific areas that are open to the public. And similarly some permit visitors while others don't.
- Being close to a port or dock be advantageous as sites could benefit from acquiring resources that could travel on ships being cheaper than road.
- VAFB doesn't have a port close only one dock. It is located at 8.70 km being the more close to a place where a ship can berth

- All sites are present at Cape Canaveral except VAFB SLC-4E
- Cape Canaveral is located at the eastern coast.
- A rocket launched from VAFB SLC-4E from east to west is at retrograde orbit and would need additional velocity to counteract earth's rotation. This perfectly translates in more fuel consumption and more expensive rockets, which means more expensive launches.
- Most sites have at least three highways in close proximity.
- KSC LC-39A is the closest to highways, it has the better success rate of the four.

- Most launch sites are located near the equator. This has the advantage of the earths spin that is maximum at the equator. This means that it is cheaper to launch an object and put it into orbit as less fuel is needed which signifies cheaper rockets and more heavier payloads.
- All sites are present at Cape Canaveral except VAFB SLC-4E.
- Cape Canaveral is closer to the equator. This could perfectly explain the better success rate for the sites overall versus the west coast launch site.

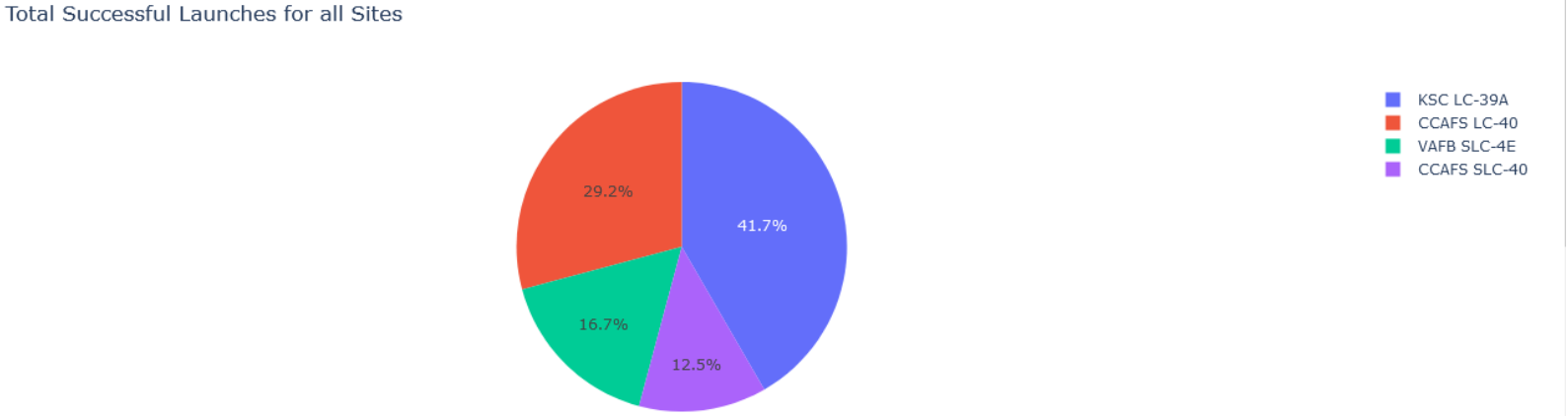
- Sites have to be around 1 km from a highway.
- Half of the sites did not reach 16 km, but the other half did not exceed 23.12 km.
- All sites were under 30 km to the city.
- KSC LC-39A is at a distance that is between VAFB SLC-4E and CCAFS SLC and LC-40 to the nearest city.
- A city is a densely populated area, keeping distance from it would be wise as to not endanger the population when testing a rocket though not too much distance as you still need some of the services a city might have.
- The United States Census Bureau has designated the Vandenberg Space Force Base as a separate-census designated place (CDP) for statistical purposes, spanning the base's residential population. As of the 2020 census the population was of 3,559.
- The Kennedy Space Center is located on Meritt Island which has a population of approximately 34,518 from the 2020 census.
- The city of Cape Canaveral lies just south of the Port Canaveral District. As of 2020 it has a population of 9,972.
    - Port Canaveral is one of the busiest cruise ports in the world.
- The management of KSC and CCAFS work very closely together, share resources and operate facilities on each other's property.

- Did you know most launch sites in the world are present to the east of countries.

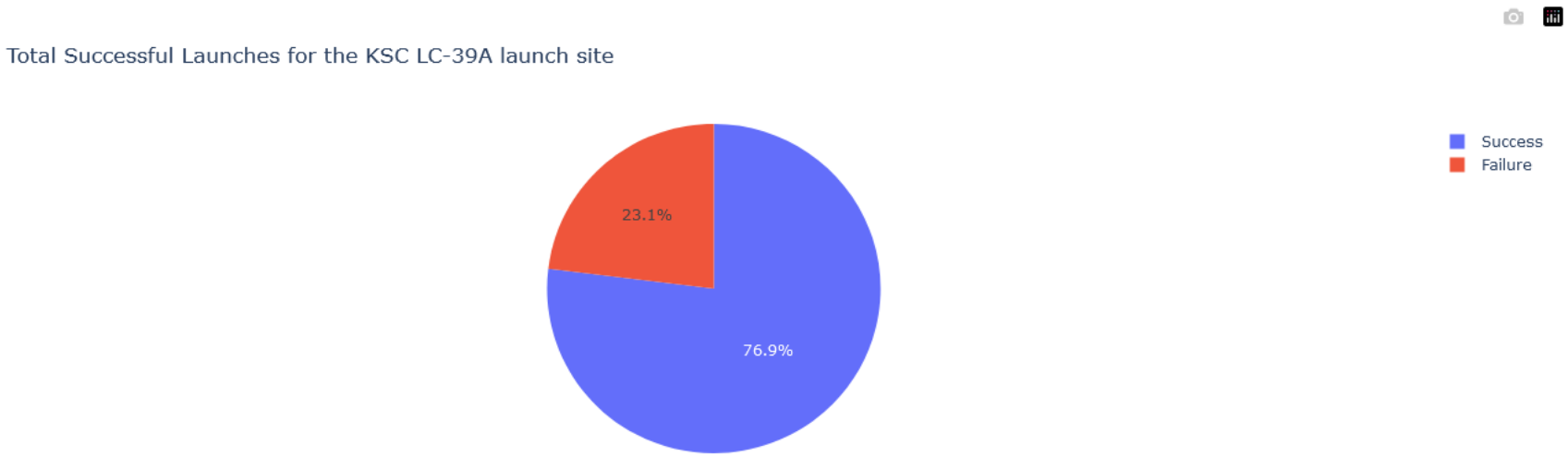

Section 4
Build a Dashboard with Plotly Dash

Launch Success Count for all Sites

- As we suspected from the previous section, we see that the launch site KSC LC-39A does have the biggest success rate.



Highest Launch Success Ratio Site



More specifically, this launch site has a 76.9% success against a 23.1% failure.

Payload vs. Launch Outcome(all sites)

We see a positive relationship between a booster version category of FT and success. We also see that most successful outcomes happened between 2000 and 4000 kilograms of payload mass.

When analyzing more in depth the outcome for a launch in between 2000 and 4000 kilograms we can see that the other category of booster version that has a high success rate second to FT is B4. Also from all the launches that were made inside the payload interval for a category v1.1 we see that they all failed.

Section 5

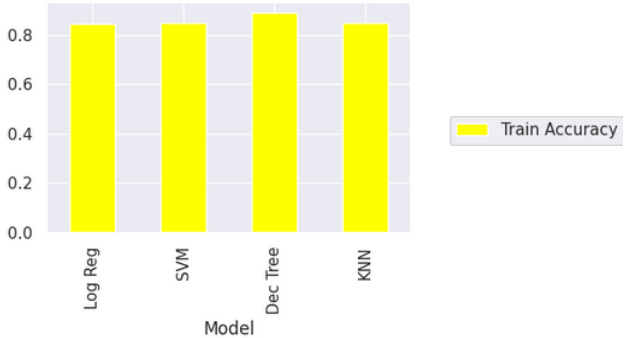
Predictive Analysis (Classification)

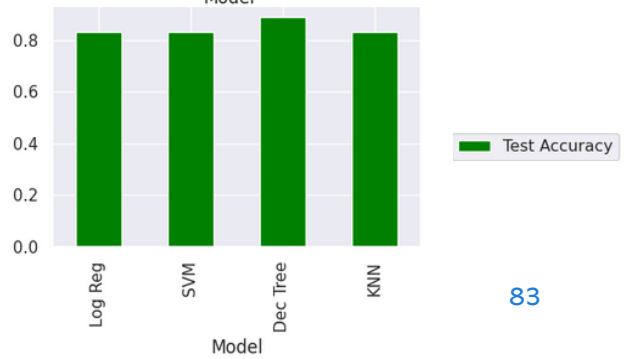
Classification Accuracy

Model	Train Accuracy	Test Accuracy
Log Reg	0.846429	0.833333
SVM	0.848214	0.833333
Dec Tree	0.889286	0.888889
KNN	0.848214	0.833333

Looking at the table, train accuracy bar chart, and the test accuracy bar chart we can see that the decision tree model has a better accuracy. The tuned hyperparameters were:

tuned hpyerparameters :(best parameters) {'criterion': 'entropy', 'max_depth': 4, 'max_features': 'auto', 'min_samples_leaf': 1, 'min_samples_split': 2, 'splitter': 'random'}

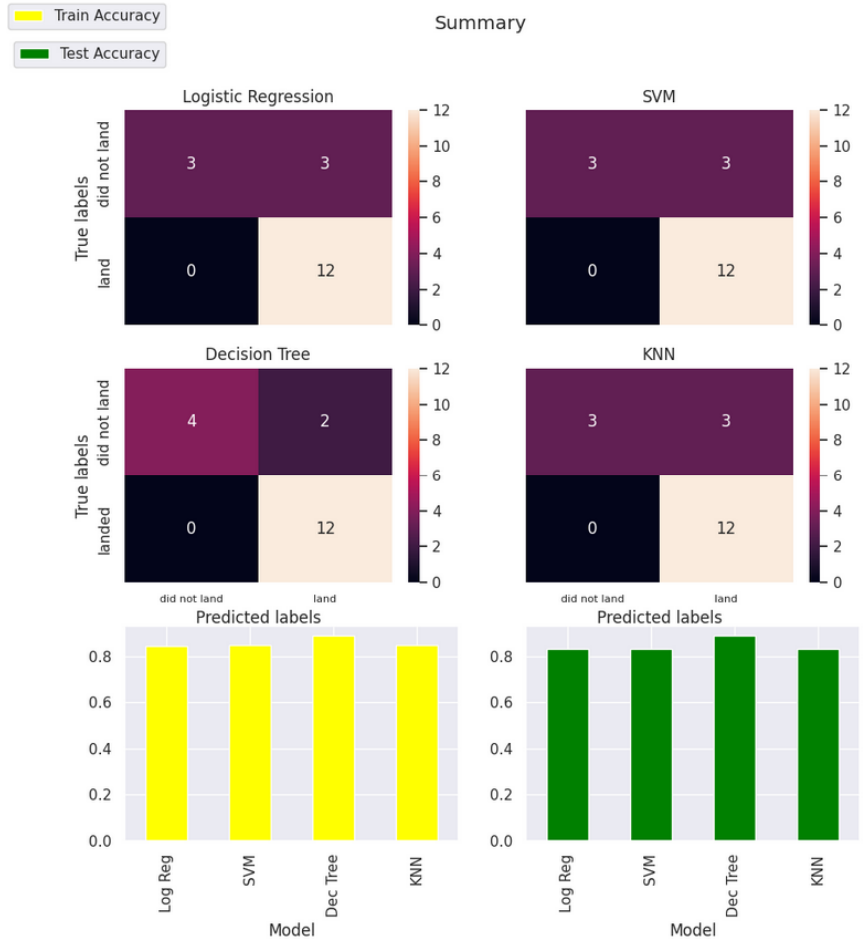




Confusion Matrix

- The decision tree model has a total of 4 true negatives, 2 false positives, 0 false negatives and 12 true positives.

This confusion matrix reveals that the model was able to successfully predict a total of 16 labels against 2 unsuccessful predictions. Other models could only successfully predict 15 labels versus 3 unsuccessful.



Model	Train Accuracy	Test Accuracy
Log Reg	0.846429	0.833333
SVM	0.848214	0.833333
Dec Tree	0.889286	0.888889
KNN	0.848214	0.833333

- Classification Accuracy and all Confusion Matrices

Conclusions Part 1

- As the launch rate in each site increases, the more likely it is for the next launch to be a success.
- Fewer launches were made with more heavy payloads, but high probability of the launches being a success.
- ES-L1, GEO, HEO and SSO orbit types have a success rate of around 100%.
- Launch success greatly outweighs failure outcome for all of the launches made to an orbit type of VLEO over 60 flights.
- ES-L1, SSO, HEO and MEO hold a success rate of a 100% for a total 9 launches made to those four orbits between a payload mass of 0 and 4000 kilograms.
- The launch success rate has been increasing from 2013 till 2020.

Conclusions Part 2

- Launch sites are:
  - At very close proximity from railways.
  - Near coastlines, although KSC LC-39A is the furthest with a 3.94 km distance.
  - Launch sites have been at a shorter distance to a ship berthing place (Ports and Dock) than what they were to their closest cities.
  - Close to highways, they were around 1 km of distance to their highways and that is why I say that they have to be around 1 km to a highway. Also the closest was KSC LC-39A. Highways usually go in and out of cities.
  - At a distance to a city under 30 km. A city is a densely populated area in which it would be wise to maintain a distance as not to endanger the population when testing a rocket that could perhaps fall but not too much as cities might have services that a launch site might need such us firefighters for example. The US has designated the Vandenberg Air Force Base as a separate census-designated-place (CDP). It has a population of for statistical purposes. It has a population of 3,559 as of 2020. KSC LC-39A is located at Merrit Island which has a population as of 2020 of 34,518. Also it is located the closest to Titusville. CCAFS SLC and LC-40 are located at Cape Canaveral which has the city of Cape Canaveral that although closer to the launch sites than Titus ville, it is further. The city of Cape Canaveral has a population of 9,972 as of 2020. Located just north of the city of Cape Canaveral is Port Canaveral which is one of the busiest cruise ports of the world. The management of KSC and CCAFS work very closely together, share resources and operate facilities on each other's property.
  - Close to wildlife refuges. Launch sites require places that can be monitorable as for people's safety they cannot go around every place as wandering for them could be fatal when perhaps testing a rocket and them being in an area were debris might fall. Sites are relatively far from cities and therefore from population. You can visit some of the launch sites. A wildlife refuge needs to be far from population as it contains wildlife threatened by specific human activities, you can visit some of this places, they also need an area that they can restrict, they both need a remote area. An example is the case of the Vandenberg Air Force Base which has three beaches in which one is open to the public in general, two are restricted to people who regularly access the base, well during the nesting season of the Western Snowy Plover sections of these beaches are closed and if a given number of trespasses have been met they close the beaches entirely. KSC LC-39A is not the closest to a wildlife refuge as it is CCAFS the one that is though it carries on the fact that it is in close proximity to two wildlife refuges as is VAFB.
  - Most launch sites are located near the equator. This has the advantage of the earths spin that is maximum at the equator. This means that it is cheaper to launch an object and put it into orbit as less fuel is needed which signifies cheaper rockets and more heavier payloads. Cape Canaveral is closer to the equator, all sites except VAFB SLC-4E are present at Cape Canaveral. A rocket launched from VAFB SLC-4E from east to west is at retrograde orbit and would need additional velocity to counteract earth's rotation. This perfectly translates in more fuel consumption and more expensive rockets, which means more expensive launches.

Conclusions Part 3

- Most Launch sites are to the east of countries.
- KSC LC-39A has the biggest success rate, more specifically it holds a 76.9% success.
- Strong relationship between success and booster version category FT, most successful outcomes happened between a payload mass of 2000 and 4000 kilograms, for a launch in this interval the other category of booster version that is second in high success rate to FT is B4.
- The decision tree model performed better than the other models in train and test accuracy, it has the highest accuracy in the train and test accuracy bar plots, the confusion matrix indicates a total of 16 successful predictions against 2 failed. Other models had 15 successful predictions versus 3 failed, in my opinion decision tree fits better for this problem.

Appendix

- Skills Network labs
- Github
- IBM
- my_data1.db
- Db2
- Wikipedia
- Google Earth

Thank you!