DS4A / EMPOWERMENT

MARCH 2022

SOLO ROUTE

SOLO TRAVEL SAFETY RECOMMENDER USING PREDICTIVE MODELING

PREPARED BY:

SHEMAINE **BENSON**

LYNNETTA **BONSU**, MPH

DAISY **OKPA**, MPH

SHANTEL **WILLIAMS**

PROJECT **TEAM 16**

TABLE OF CONTENTS

INTRODUCTION

- Americans forecasted to spend $1.06T in total travel expenditures in 2022. $131B will be for international travel. (Tourism Economics and US Travel Assoc.)
- Aside from the travel advisories published on the US Dept. of State website only provides travel advisories for COVID-19 and terrorism alerts
- Safety precautions for specific groups of travelers are often neglected

BUSINESS **PROBLEM**

Women are not given the same freedoms when solo traveling and are at a greater risk of gender-based violence when solo traveling locally and internationally. We want to provide a resource that all women can utilize that will share the safest destinations for travel based on various safety factors. These factors were modeled and ranked into 6 indexes (social, LGBTQ+, Race, Women, Resource Access, Safety) per country. This will allow women to feel more secure in their travels.

BUSINESS **SOLUTION**

SOLO ROUTE is a company that seeks to offer solutions to help travelers feel empowered, especially for solo experiences. Travelers who identify as women, LGBTQ+, and/or BIPOC face dangers and difficulties while traveling, which can ultimately impact one's physical and/or mental health. Solo Route's safe trip recommender predicts the safest countries to travel to in 2022, in addition to destinations which offer essential resources like connectivity.

IMPACT

SAFETY FOR ALL

- This will impact our target because we want women to feel safe while traveling abroad and we are providing a service that allows them to assess the safety of a variety of areas based on country level reports and user-submitted data.
- It will also provide an economic increase to tourist companies inside of that region because it would highlight that area for safer travel.
- It will help women become more aware of their surroundings and potential danger.

DATA TOOLS

TOOLS USED TO COMPLETE CAPSTONE PROJECT INCLUDE:

TOOL	USED FOR:
PYTHON MATPLOTLIB, NUMPY, PANDAS, SCIKIT-CRITERIA, SEABORN, SKLEARN	DATA CLEANING, ANALYSIS, MODELING
JUPYTER NOTEBOOK	IDE
FIGMA	DASHBOARD MOCKUP
TABLEAU PUBLIC	DASHBOARD APPLICATION
CANVA	DATAFOLIO, PRESENTATION & FINAL REPORT

DATASET DETAILS

DATA SOURCES USED:

The first dataset we used was the ***Social Progress Index*** (SPI). This tool was constructed to measure country performance on social and environmental aspects that impact economic development, in hopes to improve social outcomes. SPI consists of 12 components and 53 indicators. This dataset contains 107,000 rows of data collected and vetted from 2011-2021 for 168 countries. There are 2256 rows and 72 columns in this dataset.

The second dataset used was the ***Legatum Prosperity Index*** (LPI). This tool was designed to identify processes to improve prosperity and reduce poverty in various countries. This dataset contains 50,000 rows of data for 167 countries from 2007-2021. The indicators sheet contained the variables that we were interested in pursuing, this had ~49900 rows and 51 columns.

DATA CLEANING

The multi -step data cleaning process took place individually for each dataset prior to merging. Column names were altered to shorten descriptions, duplicates removed only from SPI dataset and null values were filled as 'NA'. We also searched for any outliers that would affect our analysis, however no apparent outliers were found. All unnecessary columns were removed. We observed consistency among data types and did not find any inconsistencies or formatting issues with the exception of converting 'year' columns to datetime format.

DATA MERGING

The two datasets were merged on 'country' and 'year' with an 'outer' join. Several countries were not merged into the final dataset, and upon further investigation, 8 countries were found to be similar among the two datasets, but with slight inconsistencies in the formatting of the names. These countries were renamed and merged into the final dataset. Data was then normalized between 0 and 1 in order for comparison purposes, as SPI data was originally between 0 and 11, and LPI data between 0 and 1. The following countries were not merged into the final dataset:

SPI: 'Barbados', 'Bhutan', 'Fiji', 'Maldives','Solomon Islands', 'Timor-Leste', 'West Bank and Gaza', 'Brunei Darussalam','Vanuatu', 'American Samoa','Andorra', 'Antigua and Barbuda', 'Bahamas, The', 'Bermuda','Cook Islands', 'Dominica', 'Greenland', 'Grenada', 'Guam','Kiribati', 'Kosovo', 'Marshall Islands', 'Micronesia', 'Monaco','Nauru', 'Niue', 'Northern Mariana Islands', 'Palau','Puerto Rico', 'Samoa', 'San Marino', 'St Kitts and Nevis','St Lucia', 'St Vincent and the Grenadines', 'Tokelau','Tonga', 'Tuvalu', 'Virgin Islands (US)'

LPI: 'Afghanistan', 'Venezuela' (note: SPI did not include these two countries in the dataset "due to the uncertain situation in recent years which might not be fully reflected in all indicators and thus might lead to misleading conclusions":Libya, Syria, Venezuela, and Yemen)

FEATURE **SELECTION &** FEATURE **ENGINEERING**

In order to select which features to consider for our safety indices, our 4-member team served as the decision makers to evaluate each variable. Based on the majority of tallies, we identified the features below to consider. New features were also created after transformation of LPI dataset after pivoting and stacking the dataframe, which are indicated which are also indicated below:

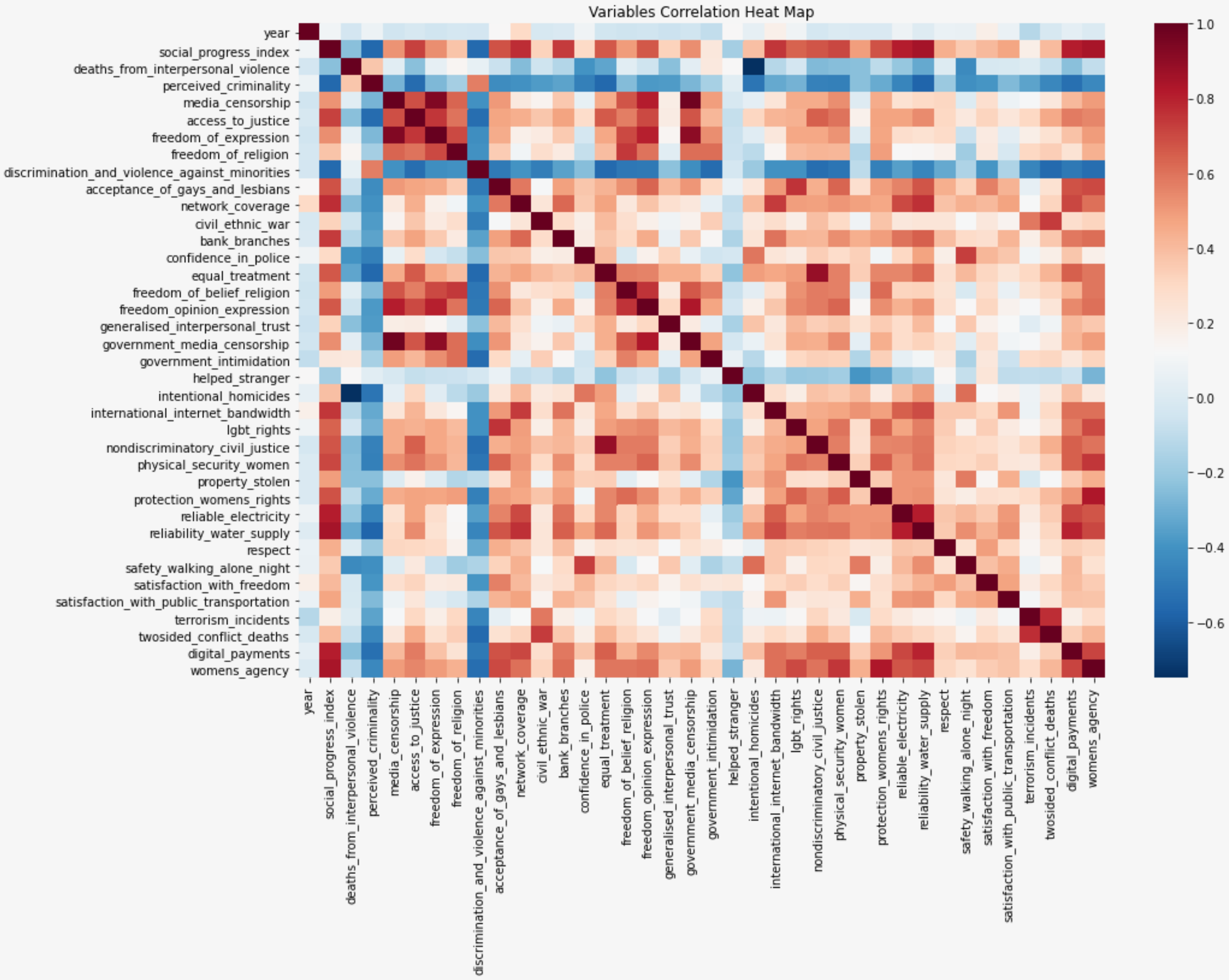
Features *(newly created columns after pivot & stacking dataframe)	**Dataset Source**	**Index to be used:**
acceptance_gays_lesbians	SPI	lgbtq+
access_to_justice	SPI	policy/justice
area_group	LPI	
area_name	LPI	
bank_branches*	LPI	connectivity
civil_ethnic_war*	LPI	overall safety
commercial bank branches*	LPI	connectivity
confidence_in_local_police*	LPI	overall safety
country	SPI	
deaths_interpersonal_violence	SPI	overall safety
digital_payments	LPI	connectivity
discrimination_violence_minorities	SPI	BIPOC
equal_treatment*	LPI	lgbtq+, BIPOC
freedom_of_belief_religion*	LPI	policy/justice
freedome_of_expression*	SPI	policy/justice
freedom_of_opinion_expression*	LPI	policy/justice
freedom_of_religion*	SPI	policy/justice
generalised_interpersonal_trust*	LPI	overall safety
government_media_censorship*	LPI	policy/justice
government_religious_intimidation_hostility*	LPI	policy/justice
helped_stranger*	LPI	overall safety
intentional_homicides*	LPI	overall safety
international_internet_bandwidth*	LPI	connectivity
lgbt_rights*	LPI	lgbtq+
media_censorship	SPI	policy/justice
network_coverage*	LPI	connectivity
non-discriminatory_civil_justice*	LPI	lgbtq+, BIPOC
perceived_criminality	SPI	overall safety
physical_security_women*	LPI	women
property_stolen*	LPI	overall safety
protection_women_rights*	LPI	women
reliability_of_electricity*	LPI	connectivity
reliability_water_supply*	LPI	connectivity
respect*	LPI	policy/justice
safety_walking_alone_night*	LPI	overall safety
satisfaction_with_freedom*	LPI	connectivity
satisfaction_with_public_transportation*	LPI	connectivity
terrorism_incidents*	LPI	overall safety
twosided_conflict_deaths*	LPI	overall safety
use_digital_payments*	LPI	connectivity
womens_agency*	LPI	women
year	LPI	
year	SPI	

EXPLORATORY **DATA ANALYSIS**

Before moving on to any potential modeling, we visualized the data we had to obtain a more accurate assessment of what we are working with. For our numerical variables, histograms were used to visualize the distribution. A great number of variables are left-skewed, but we do have several that are right-skewed. The variables 'equal_treatment' and 'respect' seem to be the only variables that are closest to showing a normal distribution. 'Government_censorship' seems to be bimodal and 'womens_agency' multimodal. 'Reliability_electricity', 'perceived_criminality', 'physical_security_women', and 'lgbt_rights' are variables that are discontinuous.

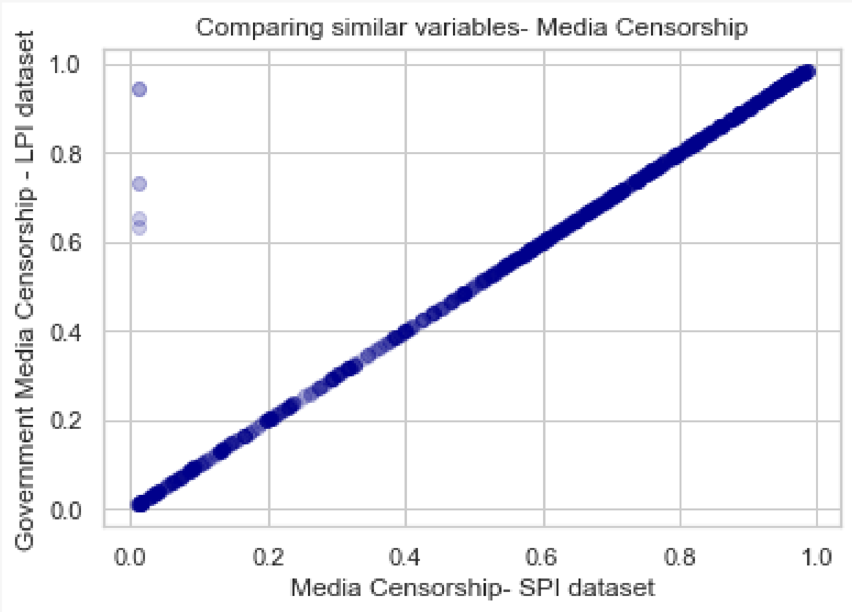
EDA

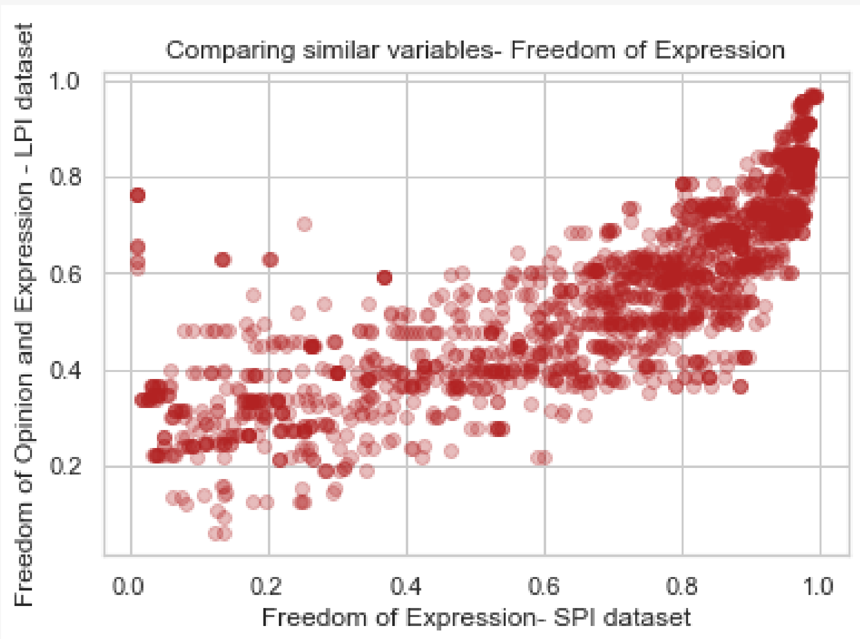
The picture below shows the correlation matrix of our selected variables. The closer to the red color, the more correlated the variables are and the closer to the blue color, the less correlated the variables are to each other. This was completed to show any similarities in the variables that were chosen.



EDA

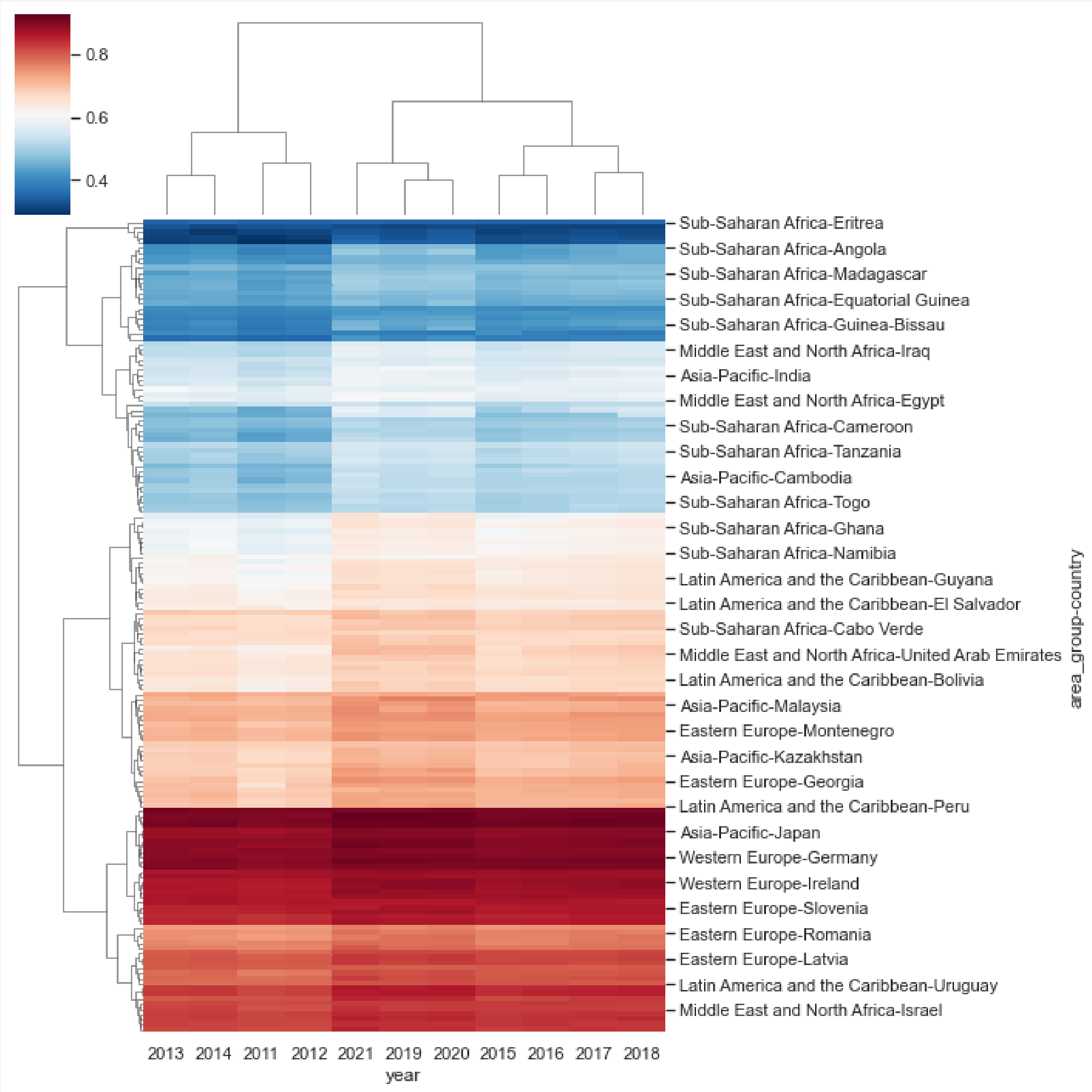
When the datasets were merged, there were a few variables that had similar names in each dataset. We graphed the similar variables using scatterplots to see how similar they were to each other. Below are graphs for comparing government media censorship from the LPI dataset with media censorship from the SPI dataset, despite a few outliers there is a strong positive linear relationship. These variables are practically identical. Another graph shows the relationship between freedom of opinion and expression (LPI) with freedom of expression (SPI). While there is a positive linear relationship, it is not as strong as the previous graph.





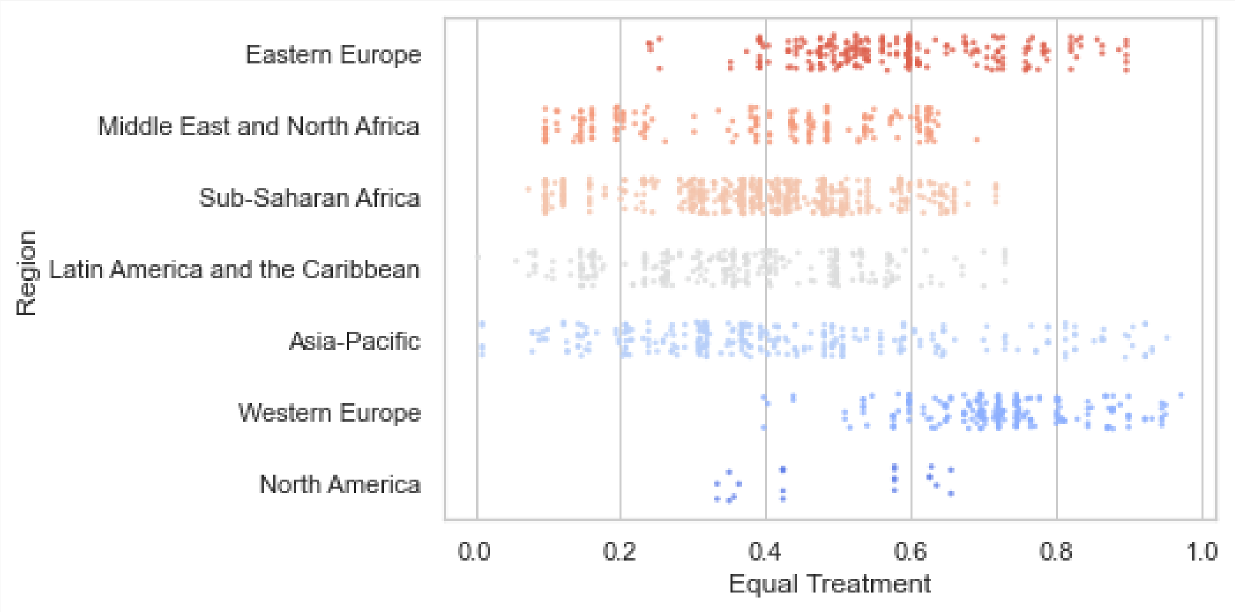
EDA

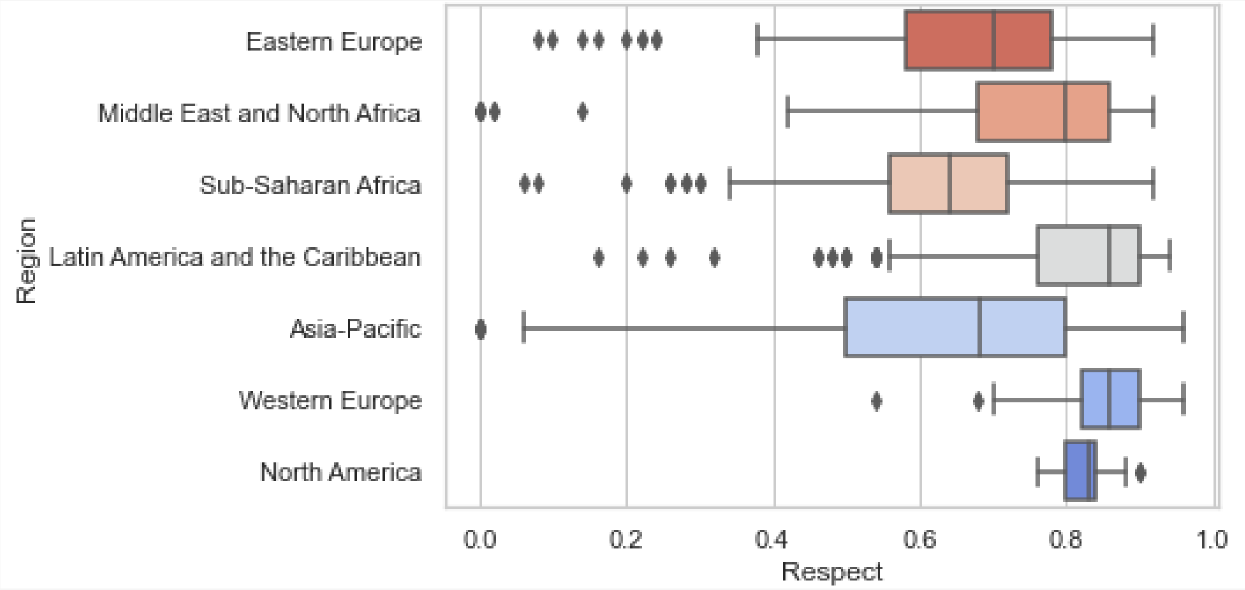
This cluster map shows two-dimensional matrix data among regions and countries, in relation to the social progress index score, across 2010 to 2021. It clustered the countries with the highest scores which are Japan, Germany, Ireland, and Slovenia. 2019-2020 are the years which are clustered together and show highest social progress scores.



EDA

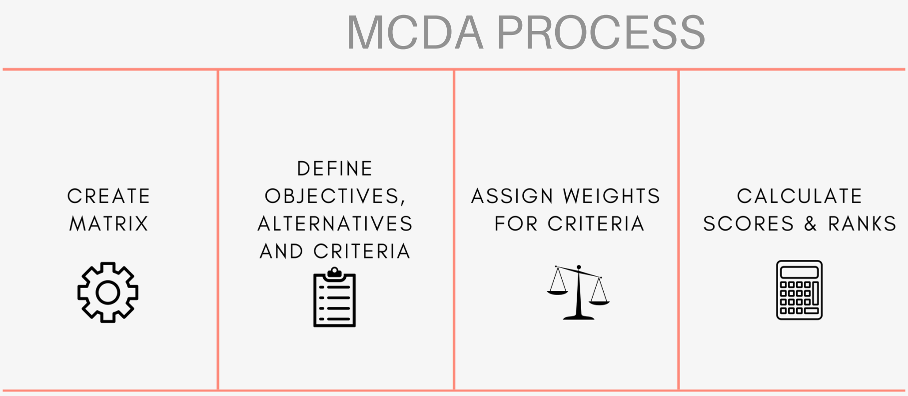
Below, the strip plot on the top shows the distribution of equal treatment scores across 7 regions of the world. North America shows the least amount of data points, which may be due to the small number of countries in this region. The box plot on the right shows a similar result in the North America region. There are also gaps and outliers among Latin America, Sub-Sahara Africa, Middle East/North Africa and Eastern Europe.

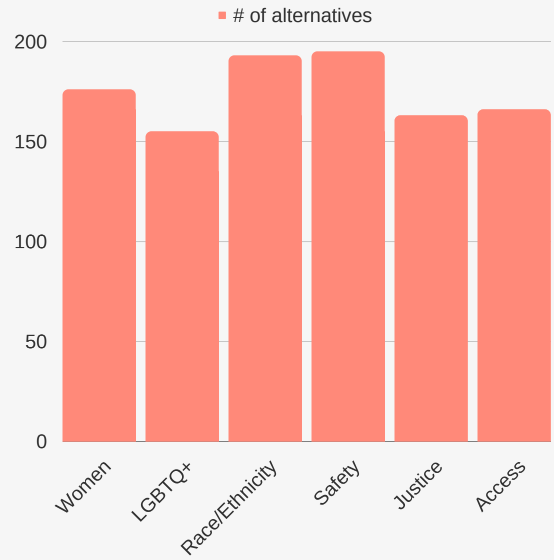




MULTI-CRITERIA DECISION ANALYSIS & SAFETY INDEX

After our data exploratory analysis, we took additional steps in our data preprocessing to transform our data so that it may be more useful for our modeling phase. We needed to formulate our own safety index from the merged datasets in an effort to construct our own safety index (including safety ranks and safety scores) for each country contained in the merged dataset. As a result, we decided to approach this dilemma by using the multi-criteria decision analysis method to summarize the data into scores and ranks to feed into our model. Multi-criteria decision analysis, or MCDA is a decision-making tool that assesses multiple criteria to make an educated decision.





We imported the **scikit-criteria** package and imported the module **skcriteria.madm** into the jupyter notebook then proceeded to filter the dataframe by each year (2011-2021). Data was filtered in order to assign weights based on relevancy of year (which will be useful in subsequent steps). All null values were dropped and data needed to be transformed to avoid errors from using minimizing objects (scikit criteria prefers maximizing objects) and errors in the sum scaler to normalize values after inverting the minimizer values into maximizers. **Matrices** were created for each subset of features (women, lgbtqplus, etc.) for each year, which equated to 66 matrices. **Alternatives** were defined for each safety index type. There were initially 167 alternatives (countries) however many were lost after null values were dropped (see image above on left).

MCDA

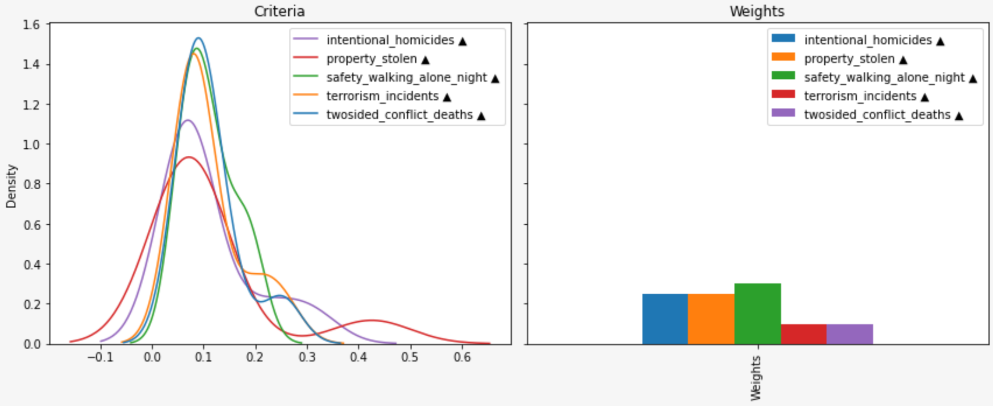


Image on left shows weights of criteria of overall safety index. Criteria and years were weighted based on level of importance/relevancy.

To define **objectives**, we maximized and minimized the features based on their benefit -- maximize benefits and to minimize non-benefits. This process also made the direction of goodness similar for all features. Components of the objective vectors were also scaled between 0 and 1, to normalize the values. **Criteria** were assigned depending on the variables that made up each type of index. Finally, each criterion was assigned a **weight** based on level of importance to the team. An example of how each matrix was formed can be seen here:

```
#apply built-in decision matrix function to create decision matrix object for safety 2021 index
#167 alternatives x 9 criteria
#apply weights if applicable
dm = skc.mkdm(
    matrix =safety21_matrix,
    objectives = objectives_safety, #direction of goodness for each column
    weights = [0.15, 0.15, .20, .05, .05, 0.10, 0.10, 0.05, 0.15], #apply weights
    alternatives = alt_safety, #safety index alternatives
    criteria=['intentional_homicides', 'property_stolen', 'safety_walking_alone_night', 'terrorism_incidents', 'twosi
              'perceived_criminality', 'civil_ethnic_war', 'generalised_interpersonal_trust'] #feature/column name
)
dm

#Invert minimizing objects
#create transformer and store in inverter variable
#apply transformation by calling transform method of the transformer and pass as parameter in dm
inverter = invert_objectives.MinimizeToMaximize()
safety21_dmt=inverter.transform(dm)

#scale values to normalize
#create transformer and store in inverter variable--scalers support target paramter- use both to normalize both matri
#apply transormation by calling transform method of transformer and pass as dmt
#save and overwrite variable dmt
scaler = scalers.SumScaler(target='both')
safety21_dmt=scaler.transform(safety21_dmt)


#convert decision matrix to dataframe
safety21_copy = safety_2021.copy()


#WSM method
safety21_wsm_dec = simple.WeightedSumModel()
safety21_wsm_rank = safety21_wsm_dec.evaluate(safety21_dmt)
safety21_copy.loc[:, '2021_rank_weightedSum']=safety21_wsm_rank.rank_
safety21_copy.loc[:,'2021_score_weightedSum']= safety21_wsm_rank.e_.score

safety21_df = safety21_copy.reset_index(drop=True)
safety21 = safety21_df[['country','area_group','2021_rank_weightedSum','2021_score_weightedSum']]
safety21
```

MCDA

We tested 3 different MCDA algorithms- the weighted sum model, the weighted product model and the technique for order of preference by similarity to ideal solution, also known as TOPSIS. Ultimately, we decided to select the weighted sum model to produce our final rankings/scores for each index. This model served our needs the best and provided logically sound scores that made the most sense for our modeling phase. The rankings were nearly the same for the weighted sum and TOPSIS algorithms, and fairly identical to the weighted product solver.

MCDA methods

	Botswana	Israel	Suriname	Mexico	Haiti	Malawi	United States	Mongolia	Zimbabwe
Rank	3	1	5	2	8	6	4	9	7

Method: WeightedSumModel

	Botswana	Israel	Suriname	Mexico	Haiti	Malawi	United States	Mongolia	Zimbabwe
Rank	3	1	5	2	7	9	4	8	6

Method: WeightedProductModel

	Botswana	Israel	Suriname	Mexico	Haiti	Malawi	United States	Mongolia	Zimbabwe
Rank	3	1	6	2	8	5	4	9	7

Method: TOPSIS

After applying the weighted sum decision solver 66 times, we applied the decision solver again to each of the 6 types of safety indexes by using the scores of each year, for each index, as the new criteria and assigning weights to the criteria, based on currency of the year. The more current the year, the higher the weight. After running the decision solver 6 additional times, we created 6 safety indexes which consisted of scores and ranks for each country, for each index.

WOMEN
3 indicators

LGBT+
4 indicators

BIPOC
3 indicators

POLICY
11 indicators

CONNECTIVITY
7 indicators

SAFETY
11 indicators

MODELING

After performing the MCDA, we decided to use classification modeling as our approach to predict the countries that would be safe to travel to in future years based on which index is being considered. We created a three-tier system for this. Tier 1 being the **safest**, Tier 2 being f**airly safe**, and Tier 3 being **unsafe**.

TIER 1 | **Safest**
TIER 2 | **Fairly Safe**
TIER 3 | **Unsafe**

We sorted the tables that were created from the MCDA in order of rank which consequently was the order of highest weighted score. Once each table was sorted, we then created three bins where the highest, medium, and lowest scores would fall into using the code below.

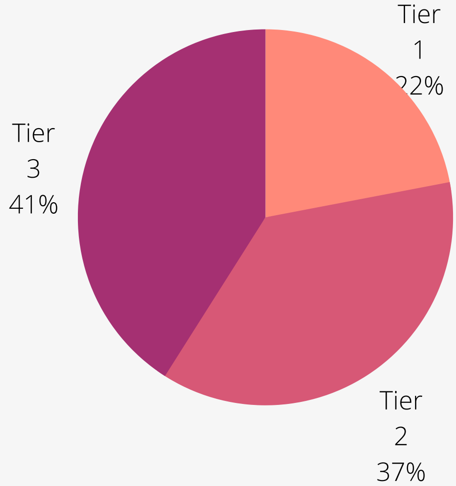
```
bins = [3, 47, 91, 136]
labels=['Tier 3','Tier 2','Tier 1']
women_index['Tier'] = pd.cut(women_index['Whole'], bins=bins, labels=labels)
women_index.head(20)
```

In short, the range of the scores (max score - min score) was divided by 3 and then this amount was used to determine the range of scores for each bin. This would, in turn, create a new column for the "Tiers" that certain countries and their respective score would fall into. Once the Tier column was established, we then found the percentage of the amount of countries that belonged to each tier. This was useful information to better grasp the amount of safe countries there are vs. unsafe countries.

```
#Calculate the % of countries in each tier
t1 = len(women_index[women_index['Tier'] == 'Tier 1'])
t2 = len(women_index[women_index['Tier'] == 'Tier 2'])
t3 = len(women_index[women_index['Tier'] == 'Tier 3'])
p_t1 = round((t1/len(women_index['Tier'])) * 100)
p_t2 = round((t2/len(women_index['Tier'])) * 100)
p_t3 = round((t3/len(women_index['Tier'])) * 100)

print('Tier 1 = ' + str(p_t1) + '%')
print('Tier 2 = ' + str(p_t2)+ '%')
print('Tier 3 = ' + str(p_t3)+ '%')
```

```
Tier 1 = 22%
Tier 2 = 37%
Tier 3 = 41%
```



MODELING

Once we had our classification set up, it was time to perform the modeling. We determined the X and y variables needed for the formulas and started to apply some classification modeling algorithms to see which one would yield the best results. The algorithms we tried were: K-nearest Neighbors, Linear Discriminant Analysis, and Support Vector Machines.

Using the women's safety index as an example, we selected the **K-nearest Neighbors** algorithm as the best one to model this data. All the other indexes (overall safety, race, access, social) were also modeled best by K-nearest Neighbors, except for the LGBTQ+ index which was best modeled by the Support Vector Machines algorithm.

```
#KNN
from sklearn.neighbors import KNeighborsClassifier

knn = KNeighborsClassifier()
knn.fit(X_train, y_train)
print('Accuracy of K-NN classifier on training set: {:.2f}'
     .format(knn.score(X_train, y_train)))
print('Accuracy of K-NN classifier on test set: {:.2f}'
     .format(knn.score(X_test, y_test)))
```

```
Accuracy of K-NN classifier on training set: 1.00
Accuracy of K-NN classifier on test set: 1.00
```

```
#Support Vector
from sklearn.svm import SVC

svm = SVC()
svm.fit(X_train, y_train)
print('Accuracy of SVM classifier on training set: {:.2f}'
     .format(svm.score(X_train, y_train)))
print('Accuracy of SVM classifier on test set: {:.2f}'
     .format(svm.score(X_test, y_test)))
```

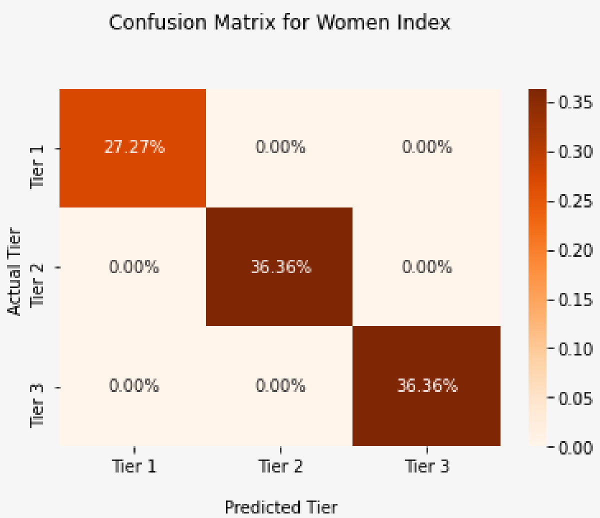
```
Accuracy of SVM classifier on training set: 0.98
Accuracy of SVM classifier on test set: 0.98
```

```
#Linear Discriminant Analysis
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis

lda = LinearDiscriminantAnalysis()
lda.fit(X_train, y_train)
print('Accuracy of LDA classifier on training set: {:.2f}'
     .format(lda.score(X_train, y_train)))
print('Accuracy of LDA classifier on test set: {:.2f}'
     .format(lda.score(X_test, y_test)))
```

```
Accuracy of LDA classifier on training set: 0.95
Accuracy of LDA classifier on test set: 0.93
```

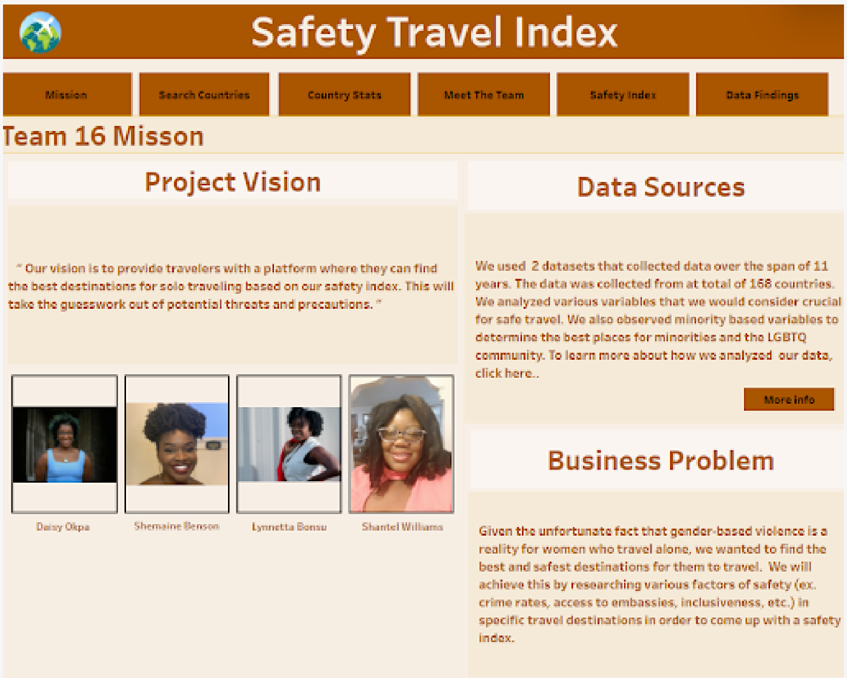
Sticking with using the example of the women's safety index, once we figured out the model we wanted to use, we then applied it to the data get our predicted outcomes. We were able to create a confusion matrix with percentages to visualize what the predicted vs. actual countries that fall into a certain tier would be.



We performed the same steps to the rest of the safety indexes to figure out the predicted outcomes for those as well. Moving forward, we hope to cater this predictive modeling to offer a more granular view of the countries and even cities that would be safe, fairly safe, and unsafe for future travel depending on the safety index being considered or even all of them.

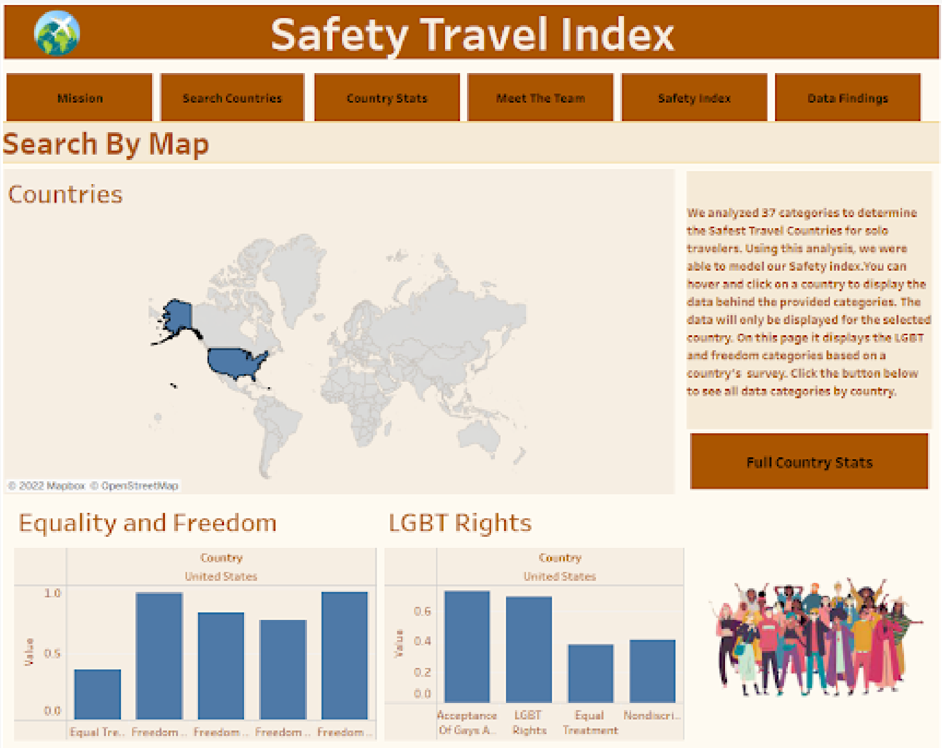
DASHBOARD **FEATURES**

We created an interactive dashboard using Tableau with intentions of the user to interact with our data and view the safety indexes based on select countries. There are a total of 6 pages and each page is accessible through our navigation bar. We will describe each page starting with the mission page.



MISSION

The mission page is the first button on the navigation bar and our home page. This page gives you the idea behind our project. It displays our project vision, business problem and the introduction to our data sources.



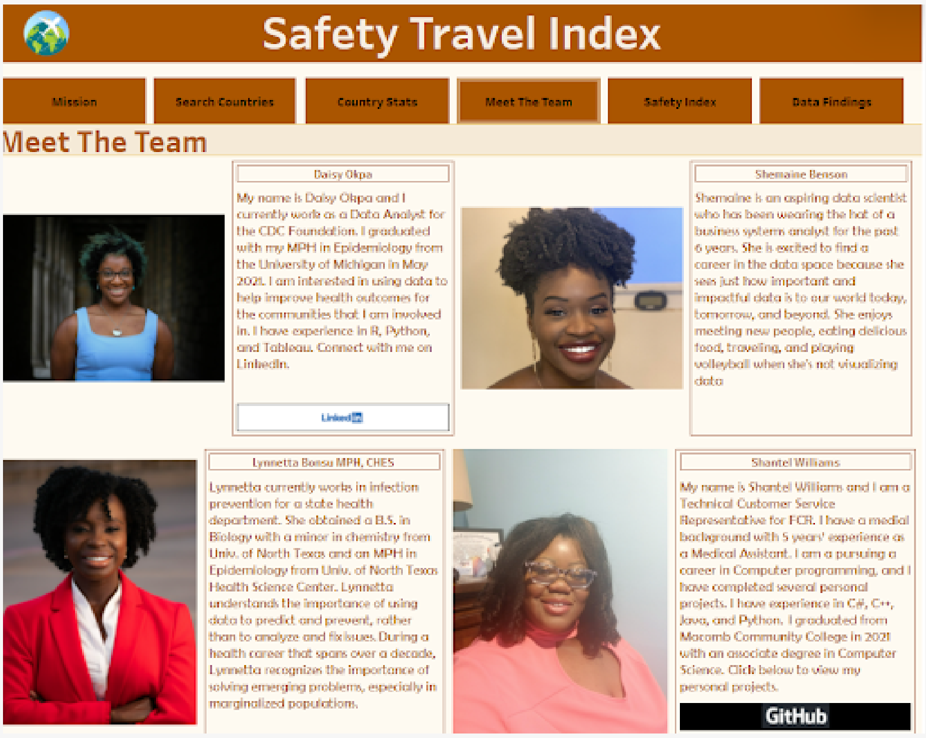
SEARCH BY COUNTRIES

The 'Search Countries' tab of our dashboard displays a map of the world where one can select a country of interest and the two bar graphs on the bottom will update with the countries values for Equality and Freedom and LGBT Rights.

DASHBOARD FEATURES

COUNTRY STATS

We created a Country Stats page to display all variables for a specified country. You can display this page by selecting it from the Navigation bar. The user can then select their desired country from the drop down menu. Once the country is selected, it will display all of the data. Some countries may have missing data if 1) they existed in one dataset, but not the other, or 2) did not have information for those variables. We still wanted to show these countries because they may be popular travel destinations and those values can be filled out over time.



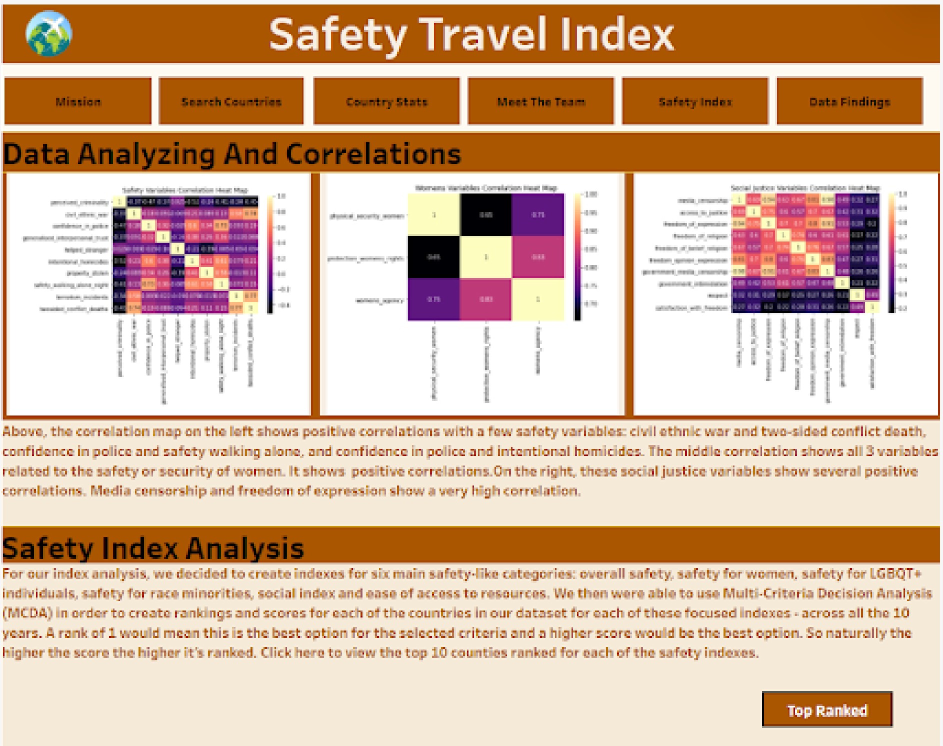
MEET THE TEAM

You can find this page by navigating to the Meet The Team button on the navigation bar. We included a page dedicated to introducing Team 16. This page includes a photo and short biography for each member. We wanted to highlight each member for their dedication and hard work.

DASHBOARD **FEATURES**

SAFETY INDEX

The 'Safety Index' tab is where you can view the overall safety, social, LBGTQ+, race, women, and resource access indexes. The top 10 countries for each index will be on display.



DATA FINDINGS

Our data findings page displays some of our data analysis findings. We have 3 correlation maps that show the relationships between different variables. It also explains how we came up with our safety indexes and how the countries were ranked. It includes a link to the safety index page.

CONCLUSIONS

Overall, working on this project has been very fulfilling. Traveling is something that should be an enlightening experience that everyone can enjoy, so creating a platform that highlights areas where one can travel to while feeling safe and comfortable being their true self is so important. From our analysis and modeling, we found that for overall safety, the top 10 safest countries were Norway, Iceland, Denmark, Singapore, Japan, Switzerland, Netherlands, Sweden, UK, and New Zealand. Many of these countries remained in the top 10 for our other indexes (Social, LGBTQ+, Race, Women, and Resource Access).

As mentioned before, we hope to further refine our modeling to provide more specific information down to the city level for where individuals can travel to safely. Solo travel continues to be on the rise and the hope is that minorities and underrepresented groups represent a large amount of that growing rate. The world is a beautiful place that we all want to explore, but being educated and aware of places to avoid is crucial.

LIMITATIONS

As mentioned in the data merging section, after merging our two datasets, there were several countries that where included in one dataset but not the other. Therefore, this produced null values, which were subsequently dropped during our data pre-processing. Also, due to the absence of a target variable we were not able to select features in a statistically relevant manner by applying various analyses in the feature selection process.

FUTURE **WORK**

If this project were to continue, we would love for users using the platform to be able to provide feedback and recommendations on the places they have traveled on a city/region level as our values are based on the country overall. User feedback can help provide more accurate information moving forward as well as recommendations for places to see and things to do while in the country. As of now, the dashboard currently displays rankings of top ten countries for each index. We plan to expand our dashboard to display all rankings /scores for accessibility purposes. Finally, we hope to incorporate more models and deeper analysis for calculating our indices and predicting scores. We truly hope that our platform can be used on a global scale to help everyone be confident when booking travel to foreign spaces.

REFERENCES

Dashboard | Social Progress Index | Legatum Prosperity Index | Datafolio