Six Steps of In-Depth Exploratory Data Analysis on Merged Datasets

Selecting columns of interest and target feature(s)

1. Which columns in your data sets will help you answer the questions posed by Your problem statement?

```
country
area_group
year
Social_progress_index
Deaths_from_interpersonal_violence
Perceived_criminality
Media_censorship
Access_to_justice
freedom_of_expression
Freedom_of_religion
discrimination_and_violence_against_minorities
Acceptance_of_gays_and_lesbians
network_coverage
Civil_ethnic_war
Bank_branches
Confidence_in_police
Equal_treatment
freedom_of_belief_religion
Freedom_opinion_expression
generalised_interpersonal_trust
Government_media_censorship
Government_intimidation
Helped_stranger
intentional_homicides
International_internet_bandwidth
lgbt_rights
Nondiscriminatory_civil_justice
physical_security_women
Property_stolen
Protection_womens_rights
reliable_electricity
Reliability_water_supply
Respect
Safety_walking_alone_night
Satisfaction_with_freedom
Satisfaction_with_public_transportation
Terrorism_incidents
Twosided_conflict_deaths
```

digital_payments
womens_agency

2. Which columns represent the key pieces of information you want to examine (i.e. your target variables)?

We plan to create a Safety Index for travel, therefore most of our columns would prove to be key pieces in determining ratings for each region and country contained in our index.
safety (10)= perceived_criminality, civil_ethnic_war, confidence_in_police, generalised_interpersonal_trust, helped_stranger, intentional_homicides, property_stolen, safety_walking_alone_night, terrorism_incidents,twosided_conflict_deaths
Women (3) = physical_security_women, protection_womens_rights, womens_agency
race (3)= discrimination_and_violence_against_minorities, equal_treatment, nondiscriminatory_civil_justice
social (9)= media_censorship, access_to_justice, freedom_of_expression, freedom_of_religion, freedom_of_belief_religion, freedom_opinion_expression, government_media_censorship, government_intimidation, respect, satisfaction_with_freedom
access (5)= network_coverage, bank_branches, international_internet_bandwidth, reliable_electricity, reliability_water_supply, satisfaction_with_public_transportation, digital_payments
Lgbt+ (4) = acceptance_of_gays_and_lesbians, lgbt_rights, equal_treatment, nondiscriminatory_civil_justice

3. How many numerical, textual, datetime etc. columns are in your dataset?

dtypes: **float64**(37), **int64**(1), **object**(3)

4. Pick out any similar columns among your disparate data sets for potential linking later on on the EDA process.

We merged our datasets using area_name from LPI and country from SPI. Below are all of the similar columns from each dataset. Scatter plots have been created to compare similar variables in 2-dimensional plots section.

LPI Dataset	**SPI Dataset**
area_name (now merged)	country (merged)
year (now merged)	year (now merged)
freedom_opinion_expression (similar)	freedom_of_expression (similar)
nondiscriminatory_civil_justice	access_to_ justice (similar)

(similar)	
lgbt_rights (similar)	acceptance_gays_lesbians (similar)
generalised_interpersonal_trust (similar)	perceived_criminality (similar)
government_media_censorship (similar)	media_censorship (similar)
freedom_of_belief_religion (similar)	freedom_of_religion (similar)

Explore Individual columns for preliminary insights

5. How many null values are present in your data (what percentage)?

The whole dataset has **6.83%** of null values.

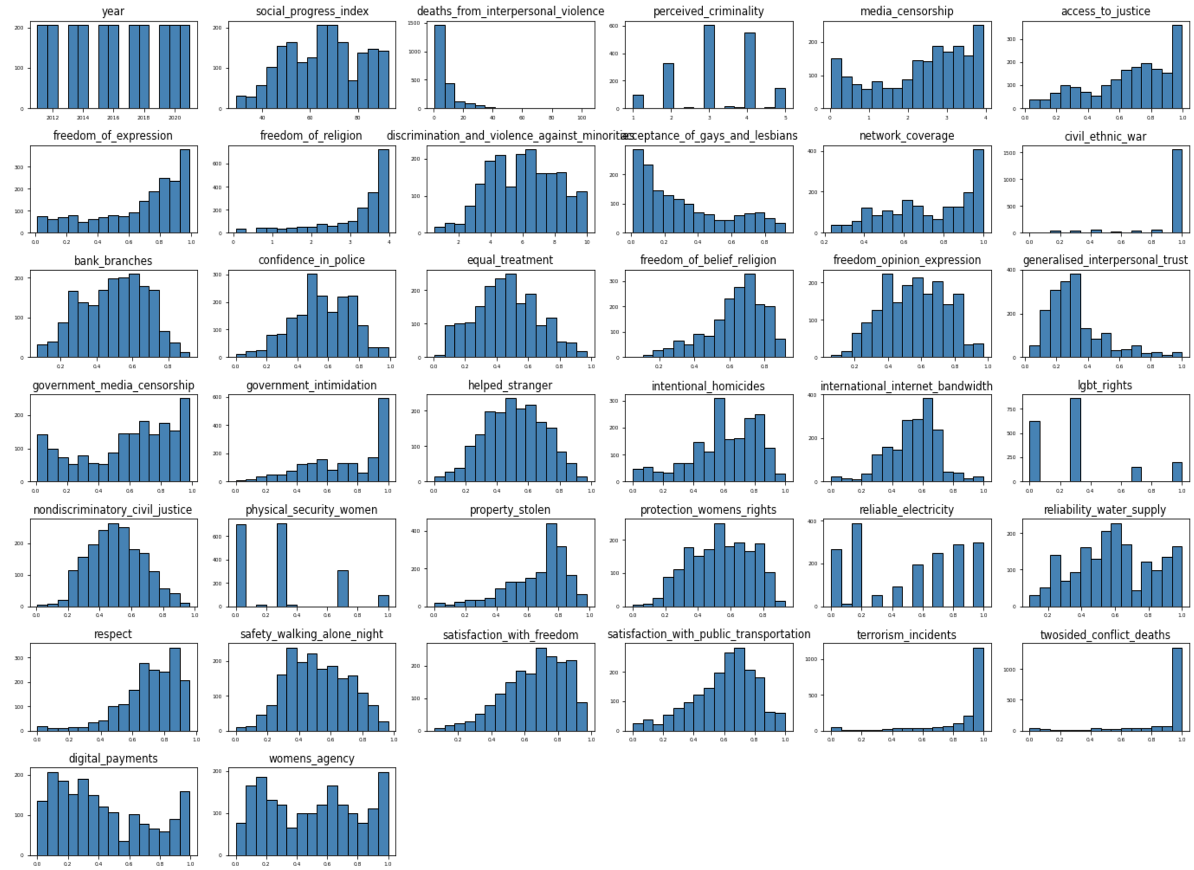
Here is the break Down of the null values per column:

country	0.00
year	0.00
area_group	19.41
social_progress_index	18.45
deaths_from_interpersonal_violence	1.94
perceived_criminality	22.11
media_censorship	15.53
access_to_justice	15.53
freedom_of_expression	15.53
freedom_of_religion	15.53
discrimination_and_violence_against_minorities	14.52
acceptance_of_gays_and_lesbians	33.54
network_coverage	18.93
civil_ethnic_war	18.93
bank_branches	18.93
confidence_in_police	18.93
equal_treatment	18.93
freedom_of_belief_religion	18.93
freedom_opinion_expression	18.93
generalised_interpersonal_trust	18.93
government_media_censorship	18.93
government_intimidation	18.93
helped_stranger	18.93

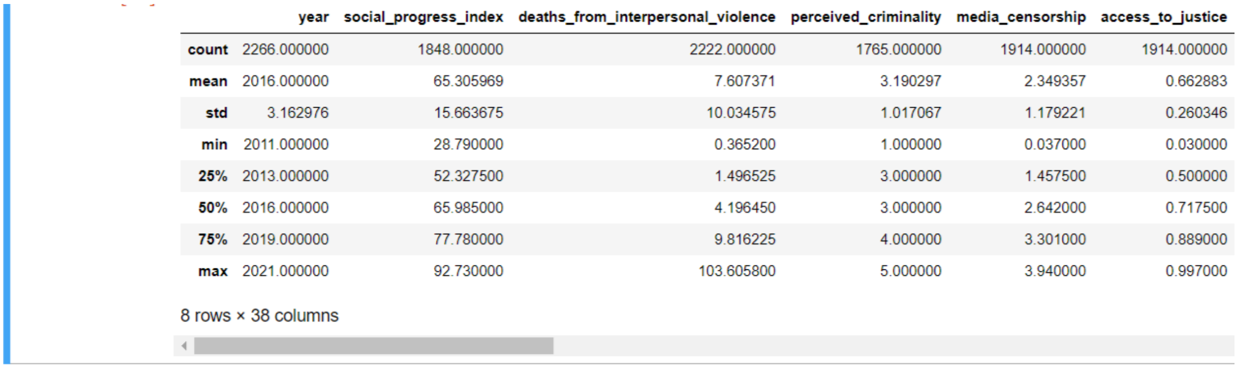
intentional_homicides	18.93
international_internet_bandwidth	18.93
lgbt_rights	18.93
nondiscriminatory_civil_justice	18.93
physical_security_women	18.93
property_stolen	18.93
protection_womens_rights	18.93
reliable_electricity	18.93
reliability_water_supply	18.93
respect	18.93
safety_walking_alone_night	18.93
satisfaction_with_freedom	18.93
satisfaction_with_public_transportation	18.93
terrorism_incidents	18.93
twosided_conflict_deaths	18.93
digital_payments	18.93
womens_agency	18.93
_merge	0.00

6. Plot one-dimensional distributions of numerical columns (ex. histograms) and observe the overall shape of the data (i.e. normal distribution, skewed, multimodal, discontinuous)

It would appear that we have a mixture of distributions. 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.

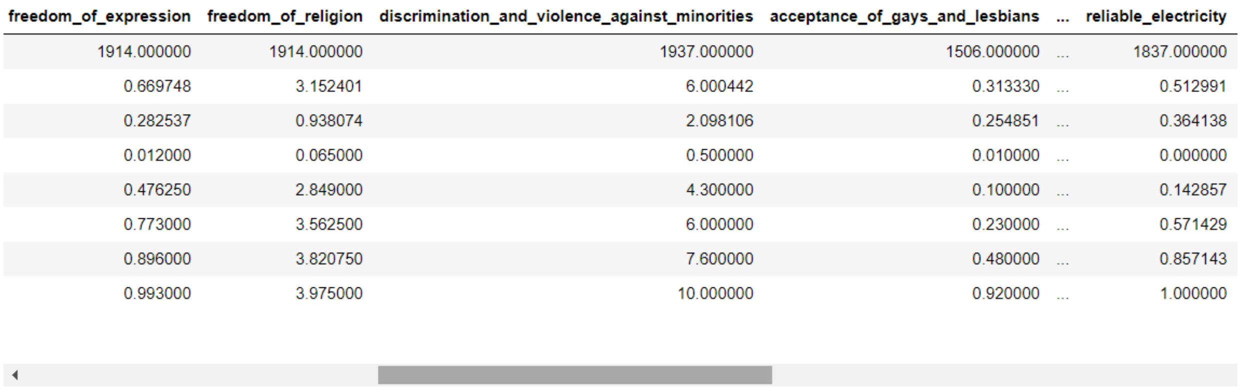


7. Compute basic statistics of numerical columns

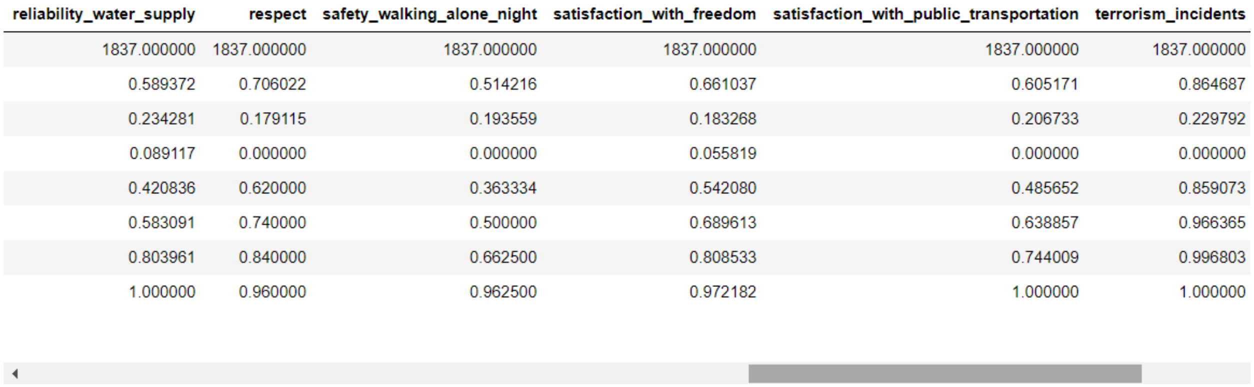


	year	social_progress_index	deaths_from_interpersonal_violence	perceived_criminality	media_censorship	access_to_justice
count	2266.000000	1848.000000	2222.000000	1765.000000	1914.000000	1914.000000
mean	2016.000000	65.305969	7.607371	3.190297	2.349357	0.662883
std	3.162976	15.663675	10.034575	1.017067	1.179221	0.260346
min	2011.000000	28.790000	0.365200	1.000000	0.037000	0.030000
25%	2013.000000	52.327500	1.496525	3.000000	1.457500	0.500000
50%	2016.000000	65.985000	4.196450	3.000000	2.642000	0.717500
75%	2019.000000	77.780000	9.816225	4.000000	3.301000	0.889000
max	2021.000000	92.730000	103.605800	5.000000	3.940000	0.997000

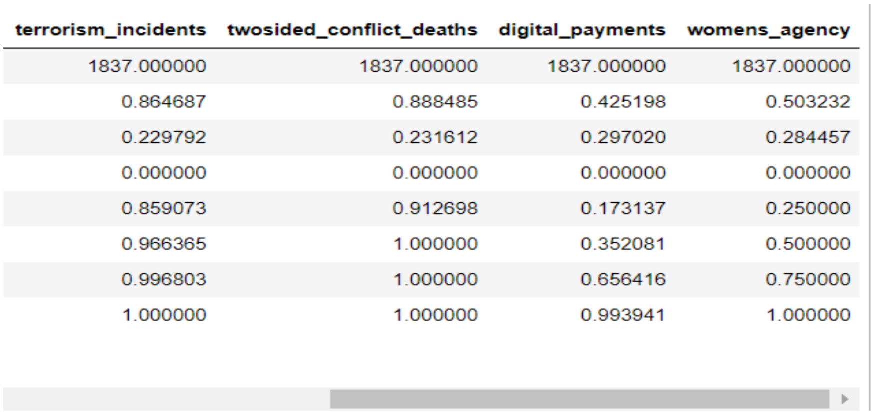
8 rows × 38 columns



freedom_of_expression	freedom_of_religion	discrimination_and_violence_against_minorities	acceptance_of_gays_and_lesbians	...	reliable_electricity
1914.000000	1914.000000	1937.000000	1506.000000	...	1837.000000
0.669748	3.152401	6.000442	0.313330	...	0.512991
0.282537	0.938074	2.098106	0.254851	...	0.364138
0.012000	0.065000	0.500000	0.010000	...	0.000000
0.476250	2.849000	4.300000	0.100000	...	0.142857
0.773000	3.562500	6.000000	0.230000	...	0.571429
0.896000	3.820750	7.600000	0.480000	...	0.857143
0.993000	3.975000	10.000000	0.920000	...	1.000000



reliability_water_supply	respect	safety_walking_alone_night	satisfaction_with_freedom	satisfaction_with_public_transportation	terrorism_incidents
1837.000000	1837.000000	1837.000000	1837.000000	1837.000000	1837.000000
0.589372	0.706022	0.514216	0.661037	0.605171	0.864687
0.234281	0.179115	0.193559	0.183268	0.206733	0.229792
0.089117	0.000000	0.000000	0.055819	0.000000	0.000000
0.420836	0.620000	0.363334	0.542080	0.485652	0.859073
0.583091	0.740000	0.500000	0.689613	0.638857	0.966365
0.803961	0.840000	0.662500	0.808533	0.744009	0.996803
1.000000	0.960000	0.962500	0.972182	1.000000	1.000000



terrorism_incidents	twosided_conflict_deaths	digital_payments	womens_agency
1837.000000	1837.000000	1837.000000	1837.000000
0.864687	0.888485	0.425198	0.503232
0.229792	0.231612	0.297020	0.284457
0.000000	0.000000	0.000000	0.000000
0.859073	0.912698	0.173137	0.250000
0.966365	1.000000	0.352081	0.500000
0.996803	1.000000	0.656416	0.750000
1.000000	1.000000	0.993941	1.000000

8. Calculate subgroup size of text/categorical data (such as the pd.value_counts() method)

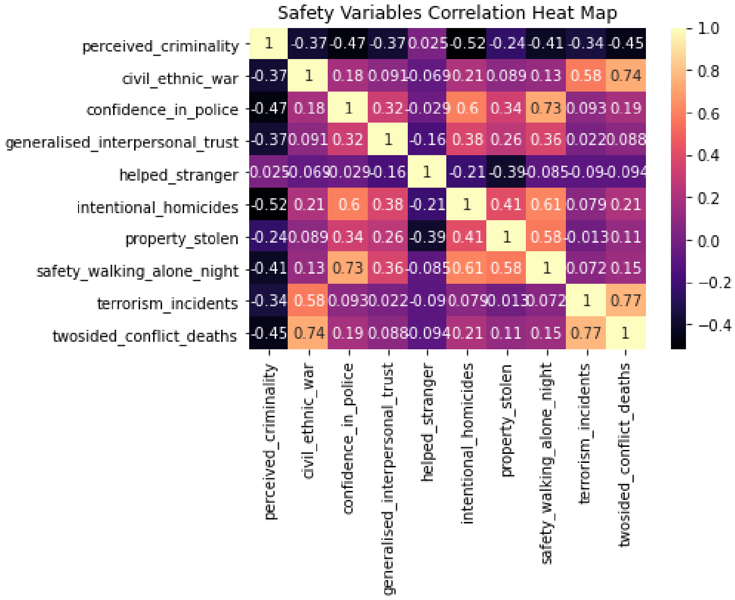
We ran **df['country'].value_counts()**. This displayed that each country had a count of 11 because we have data for 11 years total. We have a total of 206 countries.
Name: country, Length: 206, dtype: int64

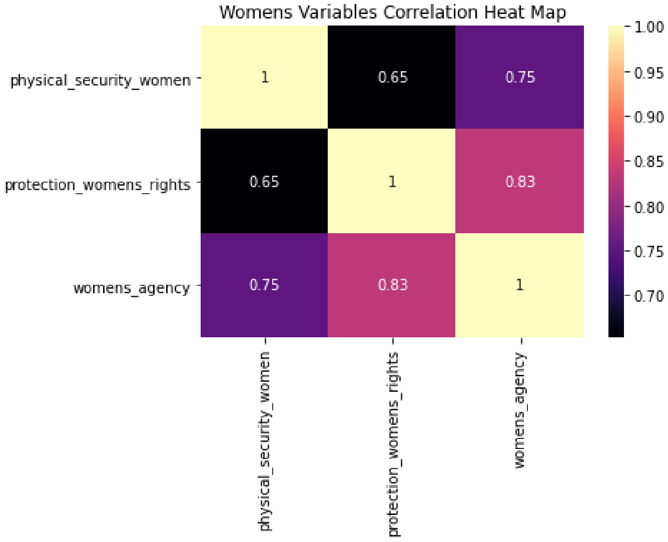
9. Explore any date/datetime columns for basic trends. How long is the period of time covered by the dataset? Do any seasonality trends immediately become Apparent?
The only datetime variable within our dataset is ‘year’, from 2010 to 2021.

Plot two-dimensional distributions of your variables of interest against your target variable(s).

Correlation heatmaps

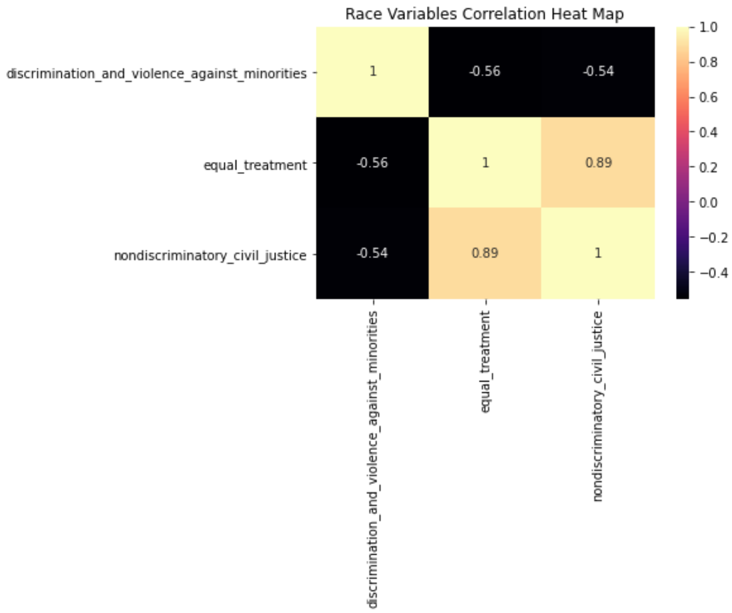
The first 6 correlation heatmaps are created to show the correlations between possible supportive indexes for subpopulation and to show possible intersectionality.

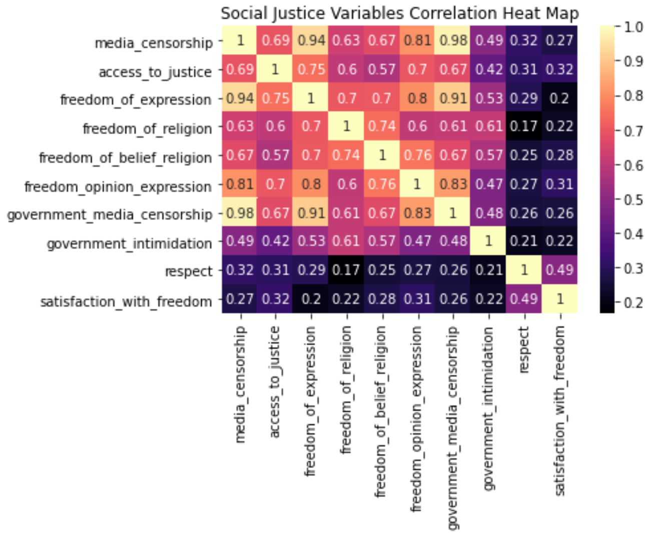




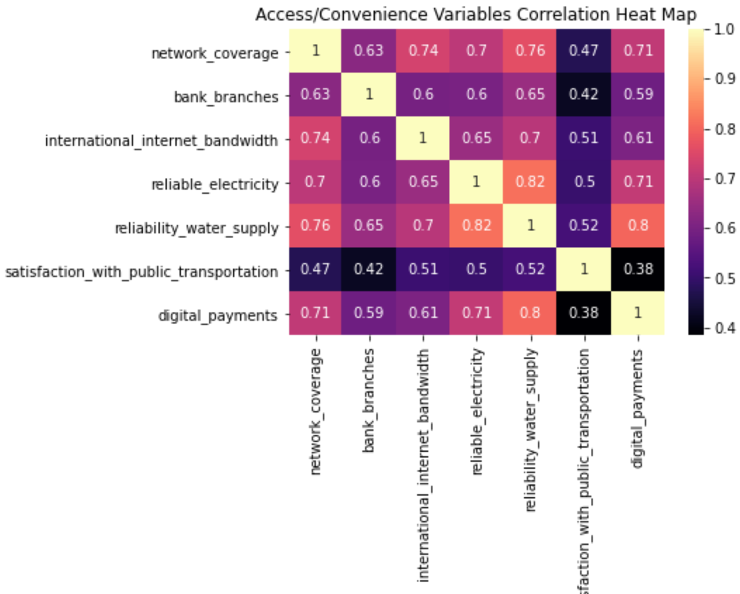
Above, the correlation map on the left shows positive correlations with a few safety variables: civil ethnic war and two-sided conflict death (0.74), confidence in police and safety walking alone (0.73), and confidence in police and intentional homicides (0.60). On the right, all 3 variables related to the safety or security of women show positive correlations(0.83, 0.75 and 0.65).

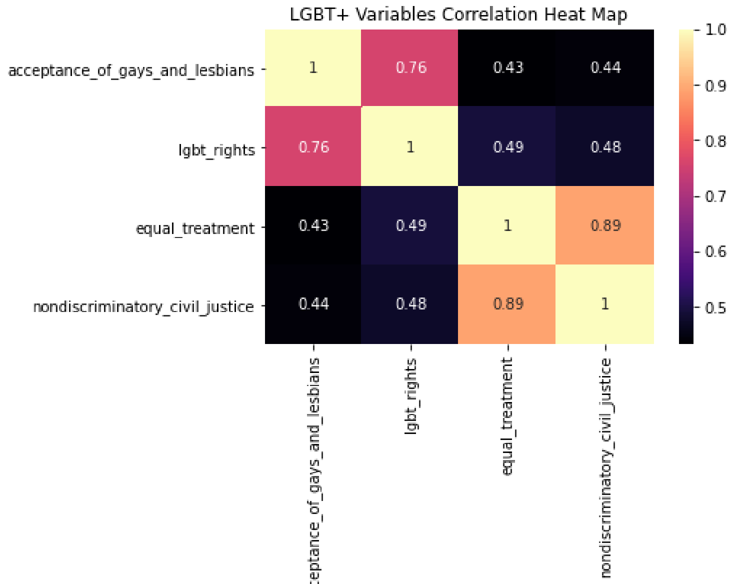
Below, the correlation heat map on the left shows a very high positive correlation between equal treatment and non-discriminatory civil justice variables (0.89), but a negative correlation between equal treatment and discrimination and violence against minorities (-0.56) as well as non-dsicriminatory civil justice and discrimination and violence against minorities (-0.54). These variables are all related to discrimination against minorities. On the right, these social justice variables show several positive correlations. Media censorship and freedom of expression show a very high correlation at 0.94.



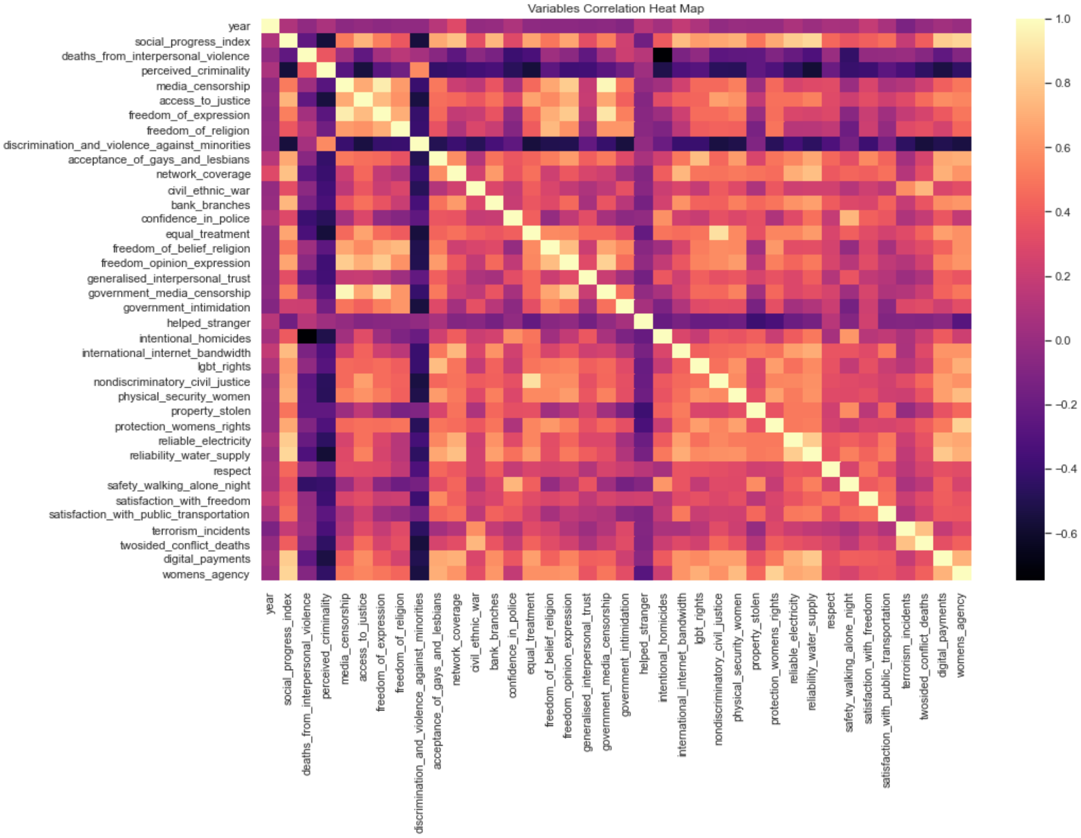


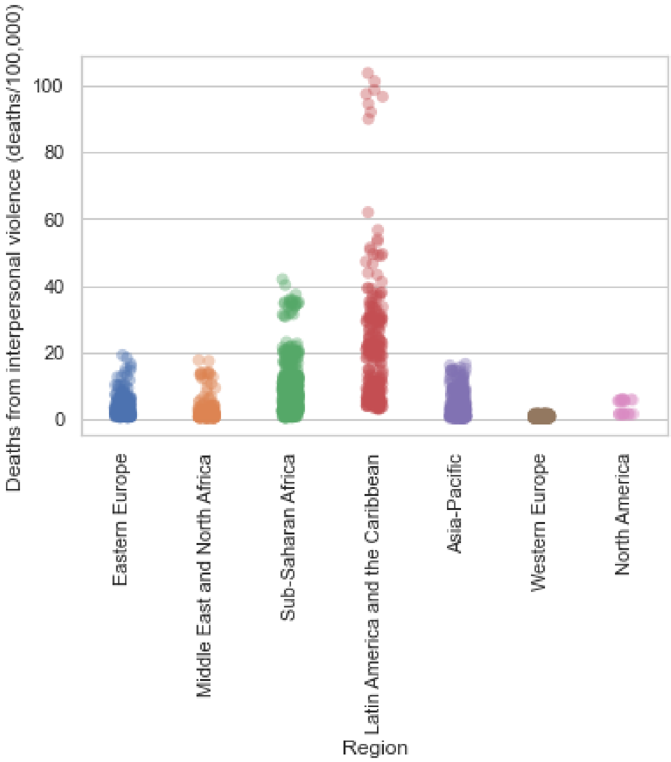
These next two correlation heat maps show the groupings of variables for access and convenience while traveling and acceptance of LGBT+ communities. The left map shows the positive correlation between reliability of electricity and reliability of water supply (0.82), reliability of water supply and access to digital payments (0.80), reliability of water supply and network coverage (0.76), international internet bandwidth and network coverage (0.74), international internet bandwidth and reliability of water supply (0.70). On the right, lgbt rights and acceptance of gays and lesbians show a correlation of (0.76).





This heatmap shows the correlation among all variables in our dataset.

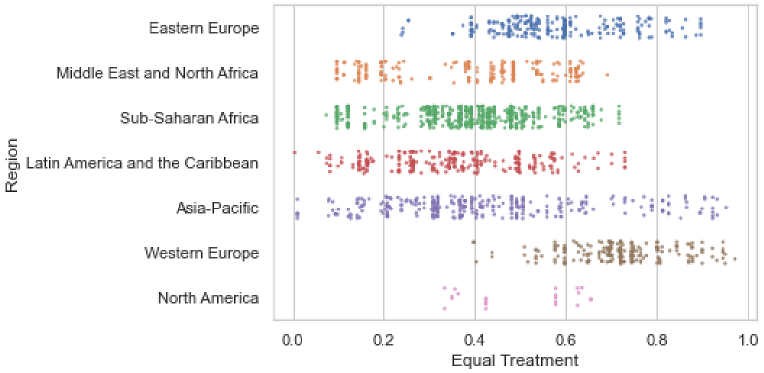


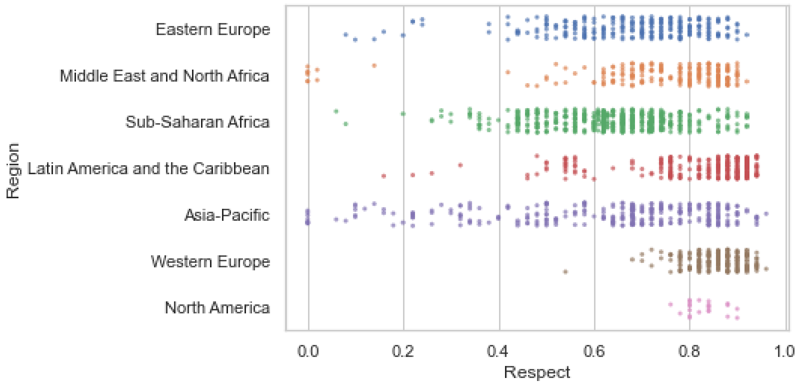


Strip plots

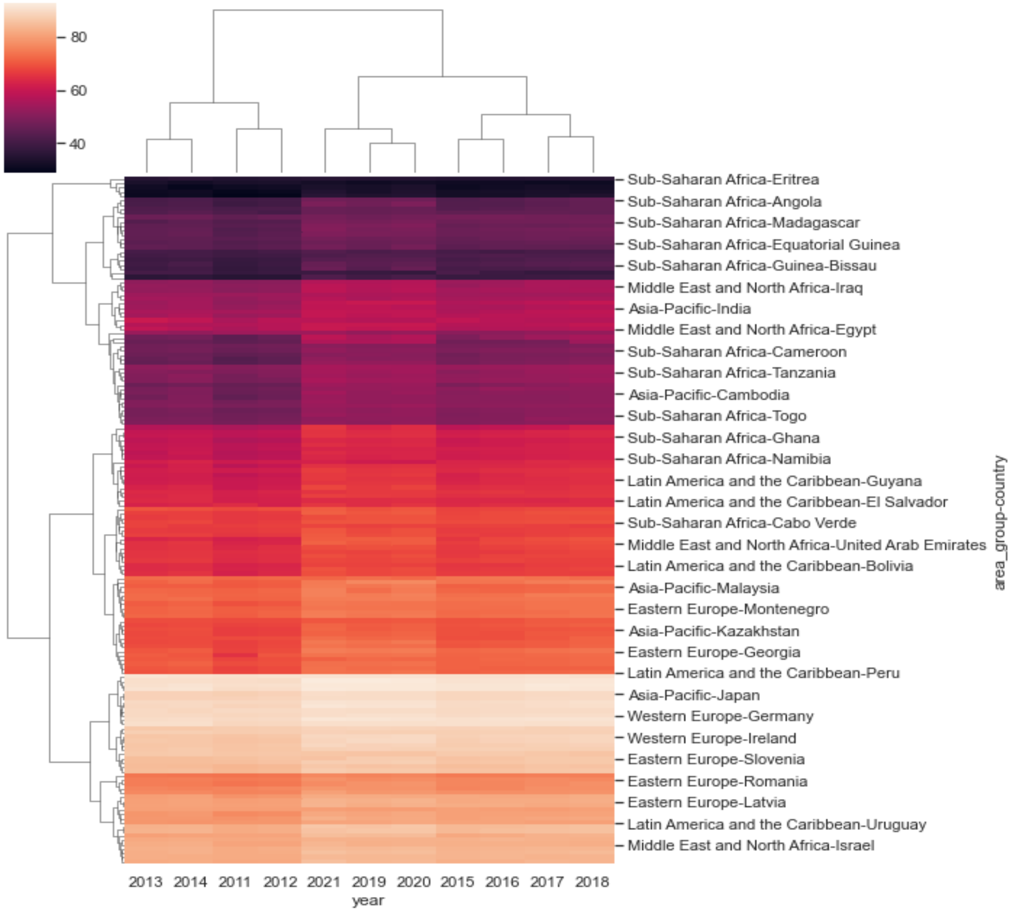
The first strip plot fuses the distribution of deaths from interpersonal violence among 7 regions of the world. This plot reveals a gap, missing values or outliers in the Latin America and the Caribbean region.

Below, the strip plot on the left 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 strip 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.





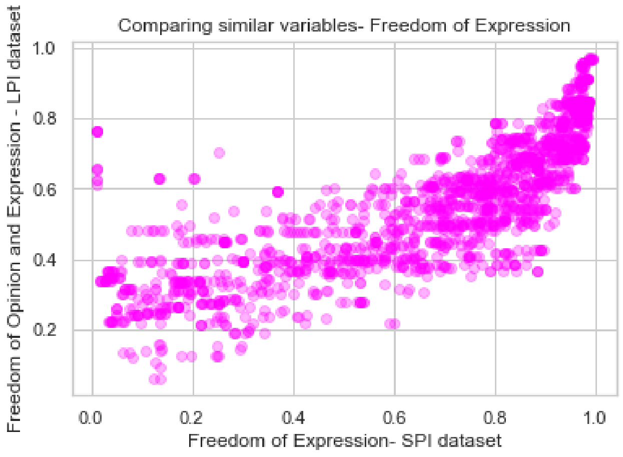
Cluster Map

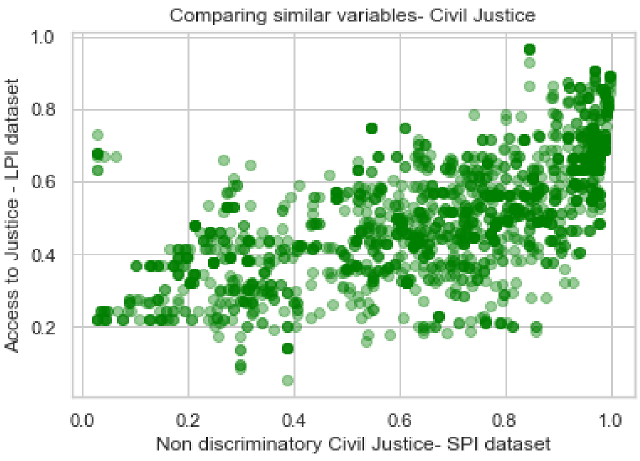


This cluster map shows two-dimensional matrix data among regions and countries, in relation to the social progress index score, across 2010 to 2021. It clustered the countries with the highest scores which are Japan, Germany, Ireland, and Slovenia. 2019-2020 are the years which are clustered together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

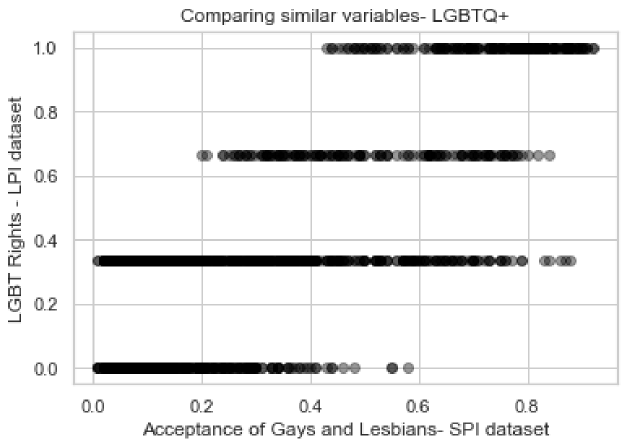
Scatter Plots

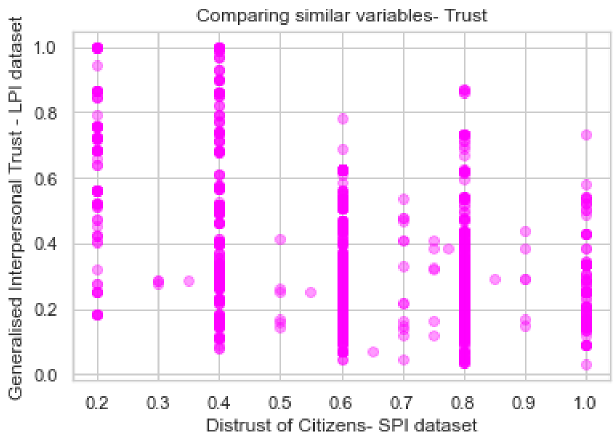
These 6 scatter plots below are used to show the relationships among variables that we identified as similar, between the two disparate datasets we have merged.





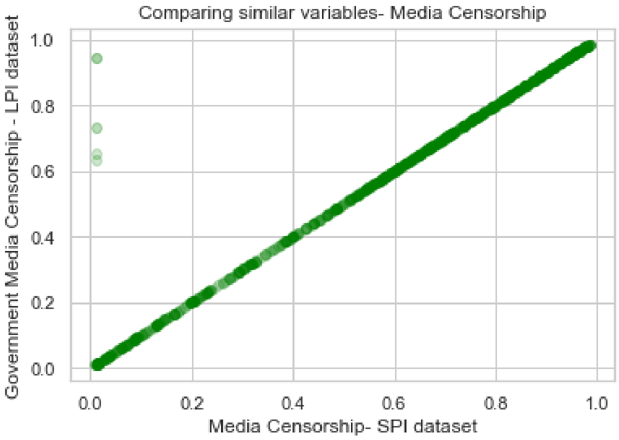
On the right above, the similar freedom of expression variables show a positive trend between these variables. On the left, the two variables concerning civil justice also show a positive relationship between the two similar variables.

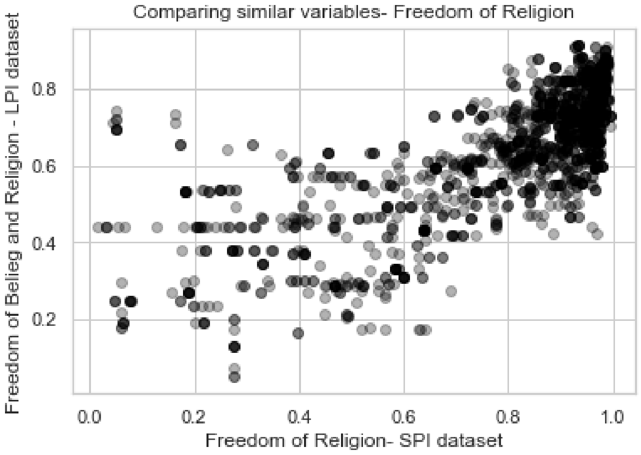




Above, the scatter plot on the left reveals at least one of the numerical variables are categorical. There does seem to be somewhat of a positive correlation shown between the two similar variables concerning LGBT+ acceptance. The plot on the right also shows the categorical numerical properties of one of the variables and shows a possible negative relationship between these two variables. These variables measure similar values on opposite scales.

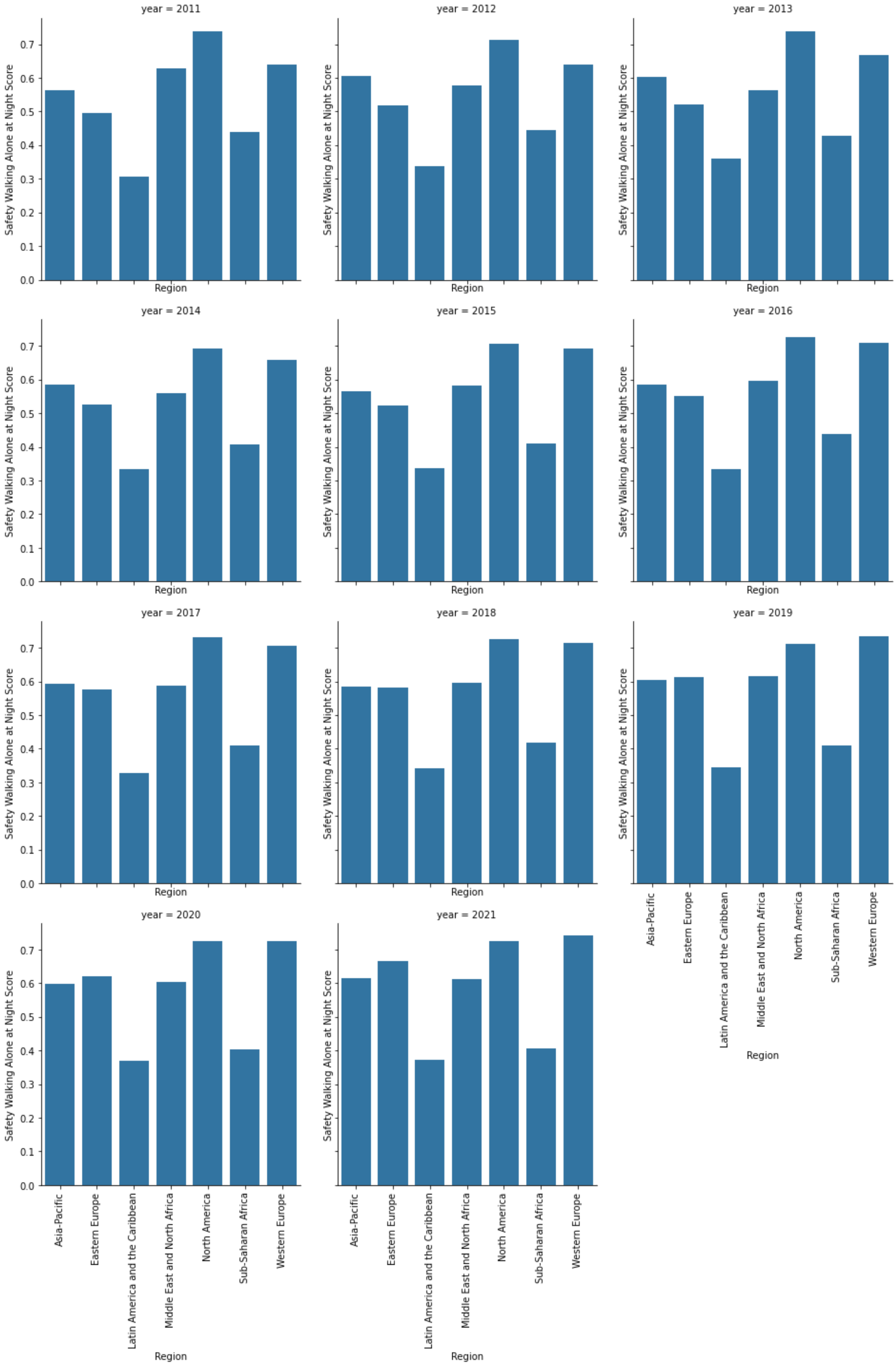
Below, these last two scatterplots both show positive trends between similar variables. The plot on the right shows a linear and nearly perfect positive trend between the two variables concerning media censorship. There are a few outliers that are shows. On the left, there is a generally positive relationship between the two freedom of religion variables.





Facet Grid Bar plot

The facet grid bar plot on the next page shows safety walking alone each night for each year per region. Throughout the 11 years, Latin America and the Caribbean region consistently had the lowest scores compared to all other regions in the world. North America and Western Europe also show consistency, however it is in trend of having the highest scores for safe walking alone at night.


year = 2011
year = 2012
year = 2013
year = 2014
year = 2015
year = 2016
year = 2017
year = 2018
year = 2019
year = 2020
year = 2021
Safety Walking Alone at Night Score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Region
Asia-Pacific
Eastern Europe
Latin America and the Caribbean
Middle East and North Africa
North America
Sub-Saharan Africa
Western Europe

10. Across different values of your independent variable, how does the dependent variable change?

Looking at our Safety Variables Correlation Map we compared multiple factors of our index. Some things to be noted was that twosided_conflict_deaths is highly correlated with terrorism incidents and civil ethnic war. This makes sense because the disagreements between areas or nations would cause more hostility between the areas. This wouldn't be surprising or out of the ordinary. Also, our clustermap revealed that in recent years (2019-2021) the social progress scores showed the highest scores.

11. Which interactions of variables provide the most interesting insights?

The interaction among the variables we identified as similar between the two datasets uncovered a few interesting insights. Of the 6 scatterplots created, 5 showed positive relationships between variables that we plan to link. The 6th plot showed a negative relationship, but this is due to the variables measuring similar values on opposite scales, as stated previously. Another interesting insight is that perceived criminality has a weak correlation with confidence in police. One would assume that the higher the perceived criminality is the less there is confidence in the police. Lastly, the variable of respect seems to score high in every region of the world. This was surprising, but reassuring to discover.

12. What trends do you see in the data? Do they support or contradict the hypothesis of your problem statement?

Overall, we feel that the various trends found in our data support our hypothesis, which is that our constructed safety index would predict the best countries to visit for solo travelers . In addition, we are factoring in subpopulations such as women, minorities and the LGBT+ community. So far, our clustermap revealed that several eastern and western European countries may be safest to travel, which are usually the regions that are most traveled. These are also regions that showed the most data points for higher equal treatment scores. Also, when viewing trends through the past 11 years for safety walking alone at night for each region. Latin America and the Caribbean region consistently had the lowest scores compared to all other regions in the world. This information is helpful because it shows that the data is consistent for all regions throughout the last decade and has little variability.

Analyze any correlations between your independent and dependent variables

13. Understand and resolve surprising correlations between these variables, and use this information to validate your initial hypothesis.

Our correlations will help us to see how variables are related to each other. Our project is designed to take the guesswork out of safe traveling and what to expect in specific locations. Our safety index will compare specific variables per country and rate accordingly. The variables we have explored thus far are the features, or inputs for our model. Therefore the dependent variable will be the output of our constructed safety index

Craft a compelling story from the work you've done in the previous steps

14. Which charts, graphs, and tables provide the most compelling evidence in support of your project idea?

Because we are making our own rating system for a safety index, we will be using all aspects of data to determine scores. Each chart, graph and table is useful for telling a story and understanding our data. We have successfully identified variables that will help us explore sub populations and intersectionality, in addition to variables that could be useful in constructing other indices that one might find useful for safe travel, such as access and convenience features. Based on the results we receive for each country based on our inputs and features, these scores will be instrumental in determining if a country is safe to travel solo. We hope to provide insight and support for those who are seeking to travel solo and require assistance in determining which destinations would be best for them to explore.

15. If your data analysis has largely disproved your initial hypothesis, can you craft a narrative for this alternative?

So far, our analysis has not disproved our initial hypothesis. We plan to approach our project with an open mind and no predetermined notions about a specific area or country. At this time, we do not believe there is a need to craft a narrative for an alternative. There may be some variables that we may reconsider implementing in our regression model, however most of the variables that we have identified and explored are supportive of our hypothesis.