**Data and Visualization**

[Optional: Add an image/logo]

Team 10: Jeevan Rai & Abhilash Narayanan

**Executive Summary:** [¼ page description of the project highlights. The summary needs to answer the following questions: What is the problem you are working on? Why is it important? What are the key results and how are they useful to the reader?]

**Table of Contents**

[1. Problem Description 2](#_gjdgxs)

[2. Data Collection and Data Quality 2](#_30j0zll)

[3. Data Exploration 2](#_1fob9te)

[4. Modeling and Evaluation 2](#_3znysh7)

[5. Conclusion 3](#_2et92p0)

[6. List of References 3](#_tyjcwt)

[7. Appendix 3](#_3dy6vkm)

# Problem Description

[**Business Understanding:** Frame the Problem. Add a short description of the problem area. Who is your stakeholder and what does the **stakeholder** want? What are the **questions** you need to answer? Why are they important? What kind of data do you need? Do some research and add references to the Reference section.]

A widespread disease COVID-19 started at the end of 2019, and it quickly spread throughout the entire world impacting every aspect of human society. Since then, there have been several kinds of studies conducted on the impact of this pandemic. The data on those studies are available for the general public. Among those datasets, we will be looking at three different datasets. The primary focus of the analysis is based on understanding the impact of this pandemic on different aspects of human society all around the world since 2019. For an instance, the dataset “Global\_Mobility\_Report.csv” provides insights into the change in public visits to different locations (like parks, grocery stores, recreational locations, etc.) in response to the policies that the governments of different countries implemented to combat against COVID-19 between 2020 and 2021. Additionally, data mining is performed on these datasets to understand any existing relationships between entities. This can be used to predict recurrence of similar pandemic in the near future and the consequences of such situations.

The insights generated from these analyses can be utilized by the leaders of various government functions to establish necessary preventative measures. For example, organizations such as Centers for Disease Control and Prevention (CDC) could allocate resources in necessary preventative measures and contingency plans if a similar pandemic occurs in the future. Such measures not only save the lives of the people, they could save the billions or trillions of government spending that come with such pandemics. For example, the total spending by the US Department of Health and Human Services was $4.7 trillion as of October 2024 (see reference). The dataset “Global\_Mobility\_Report.csv “ can be used by public health officials to understand the impact of public visits to different places due to the pandemic. Such observations can help them to further understand the effectiveness of the policies that they implemented to combat against the pandemic. If we go into a more granular level, this sort of analysis can be very helpful for business owners of all sizes. These business institutions could implement necessary preventative and contingency measures to protect the business and their employees.

As there is a large amount of information embedded within these datasets, several meaningful and important questions could be answered. For example, using the dataset “Global\_Mobility\_Report.csv” the health officials of a specific country could answer if their policies were effective based on the public visits to different places. We could compare the public movement between 2020 and 2021. The government officials not only can understand the effectiveness of their policies, they can also compare themselves to other countries’ policies to combat against the pandemic.

# Data Collection and Data Quality

[**Data Understanding part 1**: Describe data source, expected data quality and reliability. If you have several sources, how can you combine the data? At this point, you want to clean the variable names so they are better readable later on.]

The dataset “Global\_Mobility\_Report.csv” is from the report collected by COVID-19 Community Mobility Reports (see source link below). Although there are some missing values in many feature columns, the overall data quality is high as it was collected by a reliable platform. Depending on the nature of the data, the missing values could be replaced or removed while maintaining the quality of the data for data mining. All 14 variables in the dataset have been labeled appropriately for a user to understand the nature of the dataset.

Placeholder for “COVID-19\_cases\_TX.csv”

Placeholder for “COVID-19\_cases\_plus\_census.csv”

# Data Exploration

[**Data Understanding part 2:** Inspect **data quality** and clean data. Present descriptive **statistics** (e.g., in the table shown below) including how much data you have before and after cleaning. Use appropriate visualization (**histograms, bar charts, scatter plots**, etc.) and methods like **correlations**. Discuss what we can learn from exploring the data and what **recommendations** you can give based on your findings.

Do not use screenshots. Use copy paste to copy tables and charts from your R Notebook rendered as HTML or a Word Document into this document and format them in Word appropriately. Add captions in Word. You will need to adjust the axis label font sizes to make them readable. Labels in figures should be roughly the same size as regular text in the document.]

**COVID-19\_global\_mobility**

'data.frame': 3991405 obs. of 14 variables:

$ **country\_region\_code**  : chr "AE" "AE" "AE" "AE" ...

$ **country\_region**  : chr "United Arab Emirates" "United Arab Emirates" "United Arab Emirates" "United Arab Emirates" ...

$ **sub\_region\_1** : chr "" "" "" "" ...

$ **sub\_region\_2** : chr "" "" "" "" ...

$ **metro\_area** : chr "" "" "" "" ...

$ **iso\_3166\_2\_code** : chr "" "" "" "" ...

$ **census\_fips\_code** : int NA NA NA NA NA NA NA NA NA NA ...

$ **date** : chr "2020-02-15" "2020-02-16" "2020-02-17" "2020-02-18" ...

$ **retail\_and\_recreation\_percent\_change\_from\_baseline**: int 0 1 -1 -2 -2 -2 -3 -2 -1 -3 ...

$ **grocery\_and\_pharmacy\_percent\_change\_from\_baseline** : int 4 4 1 1 0 1 2 2 3 0 ...

$ **parks\_percent\_change\_from\_baseline** : int 5 4 5 5 4 6 6 4 3 5 ...

$ **transit\_stations\_percent\_change\_from\_baseline** : int 0 1 1 0 -1 1 0 -2 -1 -1 ...

$ **workplaces\_percent\_change\_from\_baseline** : int 2 2 2 2 2 1 -1 3 4 3 ...

$ **residential\_percent\_change\_from\_baseline** : int 1 1 1 1 1 1 1 1 1 1 …

Internal structure of dataset “COVID\_19\_global\_mobility”

**country\_region\_code** **country\_region** **sub\_region\_1** **sub\_region\_2**

135 135 1861 9916

**metro\_area** **iso\_3166\_2\_code** **census\_fips\_code**

66 2225 2838

Count of unique values in the columns

Shown above are the internal structures of the dataset “COVID-19\_global\_mobility” and the counts of unique values in each of the 14 feature columns or variables. The dataset has 3991405 observations. Following are brief details on the variables:

1. **country\_region\_code**: Two letters country code, character type, 135 unique codes
2. **country\_region**: Country name, character type, 135 unique country names
3. **sub\_region\_1**: character type, 1861 unique sub regions of different countries, contains the name of a primary administrative subdivision within the country, such as a state, province, or region.
4. **sub\_region\_2**: character type, 9916 unique sub regions of different countries, contains the name of a secondary administrative subdivision within the primary subdivision, such as a county, district, or municipality
5. **metro\_area**: character type, 66 unique metropolitan areas of different countries, areas typically encompass a central city and its surrounding suburbs and exurbs.
6. **iso\_3166\_2\_code**: character type, 2225 unique codes (two letter country code followed by two letter province code)
7. **census\_fips\_code**: character type, 2838 unique codes (two letters code for US state followed by two letters code for its county)
8. **date**: character type which represents date of observations, observations from 2020 to 2021
9. **retail\_and\_recreation\_percent\_change\_from\_baseline**: integer data type which simply indicate the percentage changes in visits to retail and recreation sectors to the baseline (i.e. before COVID-19 pandemic)
10. **grocery\_and\_pharmacy\_percent\_change\_from\_baseline**: integer data type which indicates the percentage changes in visits to grocery and pharmacy sectors to the baseline
11. **parks\_percent\_change\_from\_baseline**: integer data type which indicates the percentage change in the visits to parks to the baseline
12. **transit\_stations\_percent\_change\_from\_baseline**: integer data type which indicates the percentage change in the visits to transit stations (like public bus stations, train stations, etc.) to the baseline
13. **workplaces\_percent\_change\_from\_baseline**: integer data type which indicates the percentage change in the visits to workplaces to the baseline
14. **residential\_percent\_change\_from\_baseline**: integer data type which indicates the percentage change in amount of time people spent in residential locations compared to the baseline

**census\_fips\_code**  **retail\_and\_recreation\_percent\_change\_from\_baseline**

Min. : 1001 Min. :-100.0

1st Qu.:18105 1st Qu.: -41.0

Median :29115 Median : -19.0

Mean :30356 Mean : -23.2

3rd Qu.:45051 3rd Qu.: -4.0

Max. :56045 Max. : 545.0

NA's :3139208 NA's :1478424

**grocery\_and\_pharmacy\_percent\_change\_from\_baseline**

Min. :-100

1st Qu.: -14

Median : -2

Mean : -3

3rd Qu.: 9

Max. : 615

NA's :1564666

**parks\_percent\_change\_from\_baseline**

Min. :-100.0

1st Qu.: -44.0

Median : -17.0

Mean : -9.5

3rd Qu.: 11.0

Max. :1206.0

NA's :2080860

**transit\_stations\_percent\_change\_from\_baseline**

Min. :-100.0

1st Qu.: -48.0

Median : -28.0

Mean : -27.2

3rd Qu.: -7.0

Max. : 554.0

NA's :1973496

**workplaces\_percent\_change\_from\_baseline**

Min. :-100.00

1st Qu.: -32.00

Median : -19.00

Mean : -20.07

3rd Qu.: -5.00

Max. : 260.00

NA's :189760

**residential\_percent\_change\_from\_baseline**

Min. :-46.0

1st Qu.: 4.0

Median : 8.0

Mean : 9.4

3rd Qu.: 14.0

Max. : 65.0

NA's :1678955

Shown above is basic statistical information on all the feature columns that contain values of numerical data type. For example, we can see that there are census fips code 1001 to 56045. The values in the column **“**residential\_percent\_change\_from\_baseline” range from -46 to 65. We can see that there are missing values in all feature columns. See below for the percentage of missing values in each feature column.

country\_region 0.00

iso\_3166\_2\_code 0.00

census\_fips\_code 78.65

retail\_and\_recreation\_percent\_change\_from\_baseline 37.04

grocery\_and\_pharmacy\_percent\_change\_from\_baseline 39.20

parks\_percent\_change\_from\_baseline 52.13

transit\_stations\_percent\_change\_from\_baseline 49.44

workplaces\_percent\_change\_from\_baseline 4.75

residential\_percent\_change\_from\_baseline 42.06

year\_month 0.00

Here is a quick summary of the variables that have non-character data type. Let’s take the variable “census\_fips\_code” where we can see that the code for “census\_fips\_code” goes from 1001 to 56045. The mean of the values is 30356. We can also see it has 3139208 (out of 3991405)observations as “NA”. This means it has about 78% of the values not available. Similar interpretations can be made for the remaining variables.

**Feature Processing**

Since the variable “country\_region” already has names of the countries, we can ignore the variable “country\_region\_code”. The metropolitan areas in the “metro” column represent only the areas of certain countries in the “country\_region” column. Additionally, the focus of the analysis will be limited to the province of certain countries. Thus, the variable “metro” can be ignored. Similarly, the variables “sub\_region\_1” and “sub\_region\_1” can be ignored from the dataset.

*# Convert the date column to Date type and extract year and month*

dataset\_global\_mobility **<-** dataset\_global\_mobility **%>%**

mutate(date **=** as.Date(date, format **=** "%Y-%m-%d")) **%>%**

mutate(year\_month **=** format(date, "%Y-%m"))

For the variable “date”, the data type is of “char” which means they will need to be converted to date type. Since the focus on the trend analysis month-month and year-year, the converted values will be extracted to keep only the year and month. See above.

*# Drop irrelevant columns*

dataset\_global\_mobility **<-** dataset\_global\_mobility **%>%**

select(**-**country\_region\_code, **-**metro\_area, **-**sub\_region\_1, **-**sub\_region\_2, **-**date)

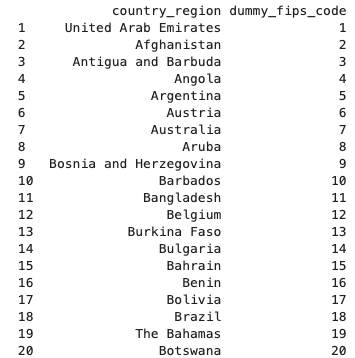
*# Remaining feature columns*

names(dataset\_global\_mobility)

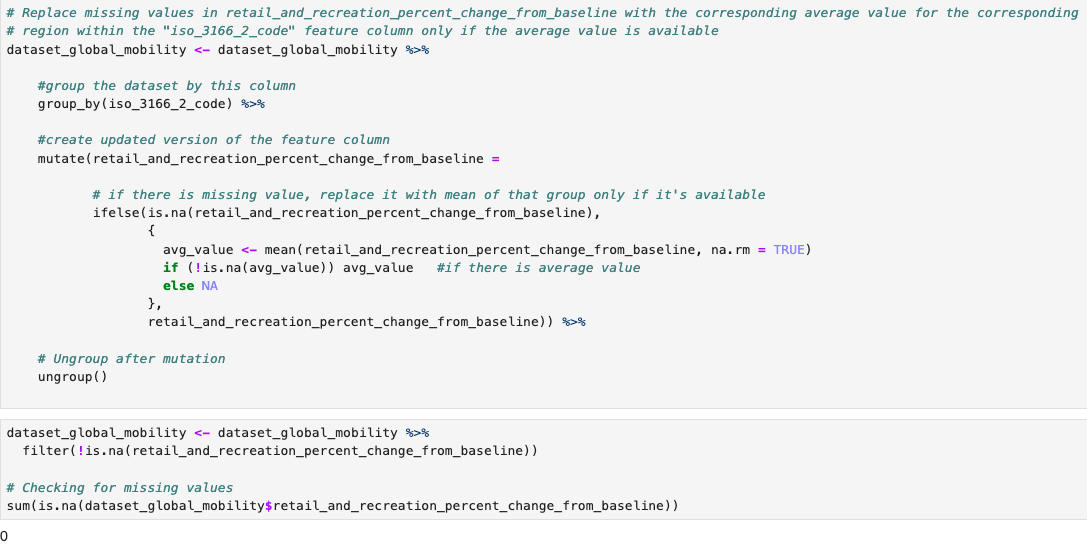
1. 'country\_region'
2. 'iso\_3166\_2\_code'
3. 'census\_fips\_code'
4. 'retail\_and\_recreation\_percent\_change\_from\_baseline'
5. 'grocery\_and\_pharmacy\_percent\_change\_from\_baseline'
6. 'parks\_percent\_change\_from\_baseline'
7. 'transit\_stations\_percent\_change\_from\_baseline'
8. 'workplaces\_percent\_change\_from\_baseline'
9. 'residential\_percent\_change\_from\_baseline'
10. 'year\_month'

After removing irrelevant features from the dataset, here are the feature columns that will be used to make analysis on the dataset. We can see that we removed 4 features from the initial dataset.

Since there are about 78% missing values in the column “census\_fips\_code”. These values are for countries other than the United States since other countries do not have states. This has been validated by looking at the missing values in this column for other countries as shown below. This feature column will not be utilized when analyzing any statistical trends related to other countries than the United States. Since the majority of these missing values are for countries other than the United States, dummy codes can be assigned to these remaining 134 countries in the “census\_fips\_code” column. Shown below are 20 dummy fips codes assigned to the 20 countries. Now the remaining missing values are the United States that are not accounted for in the dataset. This is only about 0.43% of the dataset which is a very small portion of the dataset. So these missing values are removed from the dataset.



The feature column “retail\_and\_recreation\_percent\_change\_from\_baseline” has about 37% missing values which is a significant portion of the observations in the dataset. These missing values for specific regions based on the “iso\_3166\_2\_code” column are replaced with the corresponding region’s average value. There are certain regions that have only missing values (about 5000 observations) within this feature column. Since these remaining missing values are not that significantly big, they are removed completely from the dataset. Shown below shows this feature processing technique implemented.



Similar feature processing technique is applied to “grocery\_and\_pharmacy\_percent\_change\_from\_baseline”, “parks\_percent\_change\_from\_baseline”, “transit\_stations\_percent\_change\_from\_baseline”, and “residential\_percent\_change\_from\_baseline” since they contain about 39%, 52%, 49%, and 42% of missing values respectively. There are only about 4% missing values in the feature column “workplaces\_percent\_change\_from\_baseline”. So, the observations with these missing values are removed from the dataset.

tibble [3,639,804 × 10] (S3: tbl\_df/tbl/data.frame)

$ **country\_region** : chr [1:3639804] "United Arab Emirates" "United Arab Emirates" "United Arab Emirates" "United Arab Emirates" ...

$ **iso\_3166\_2\_code** : chr [1:3639804] "" "" "" "" ...

$ **census\_fips\_code** : int [1:3639804] 1 1 1 1 1 1 1 1 1 1 ...

$ **retail\_and\_recreation\_percent\_change\_from\_baseline**: num [1:3639804] 0 1 -1 -2 -2 -2 -3 -2 -1 -3 ...

$ **grocery\_and\_pharmacy\_percent\_change\_from\_baseline** : num [1:3639804] 4 4 1 1 0 1 2 2 3 0 ...

$ **parks\_percent\_change\_from\_baseline** : num [1:3639804] 5 4 5 5 4 6 6 4 3 5 ...

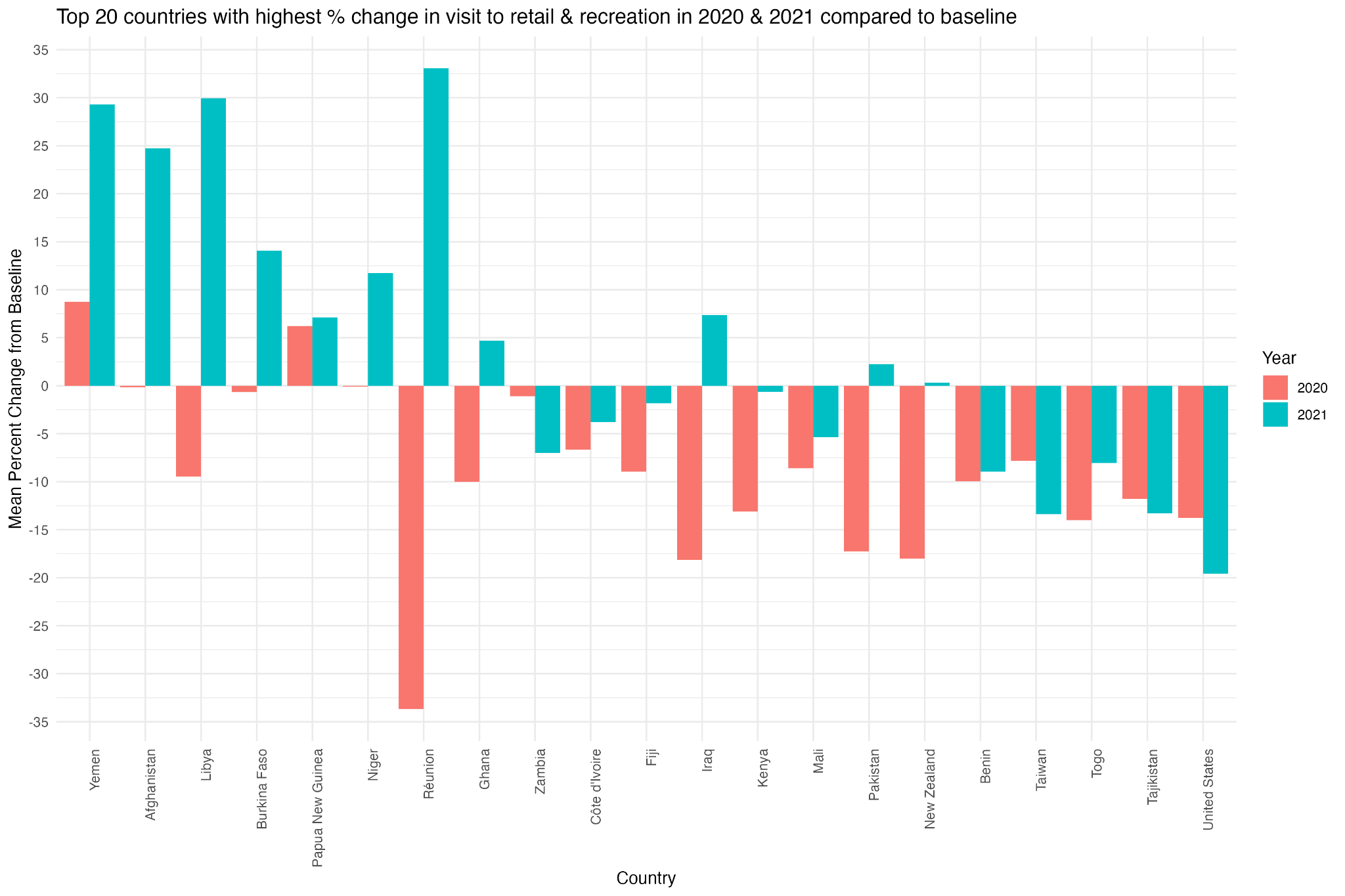
$ **transit\_stations\_percent\_change\_from\_baseline** : num [1:3639804] 0 1 1 0 -1 1 0 -2 -1 -1 ...

$ **workplaces\_percent\_change\_from\_baseline** : int [1:3639804] 2 2 2 2 2 1 -1 3 4 3 ...

$ **residential\_percent\_change\_from\_baseline** : num [1:3639804] 1 1 1 1 1 1 1 1 1 1 ...

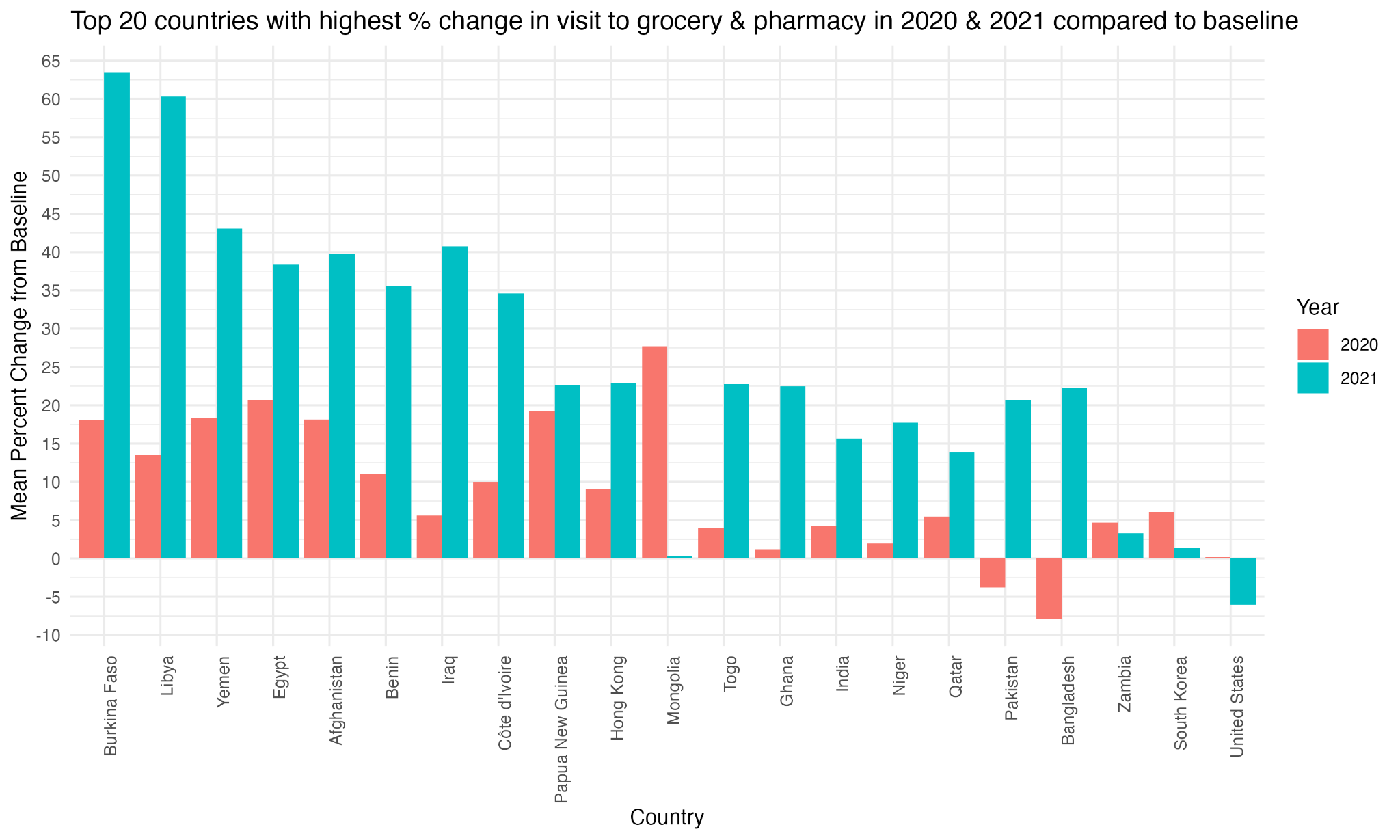
$ **year\_month** : Date[1:3639804], format: "2020-02-01" "2020-02-01" ...

After all the feature processing, here is the quick overview of the updated dataset. There are now 3639804 observations and 10 variables as compared to 3991405 observations and 14 variables in the original dataset. This means that about 8% of the observations were removed from the original dataset as they were missing values. However, there is still a sufficient amount of data left to perform various kinds of data mining.

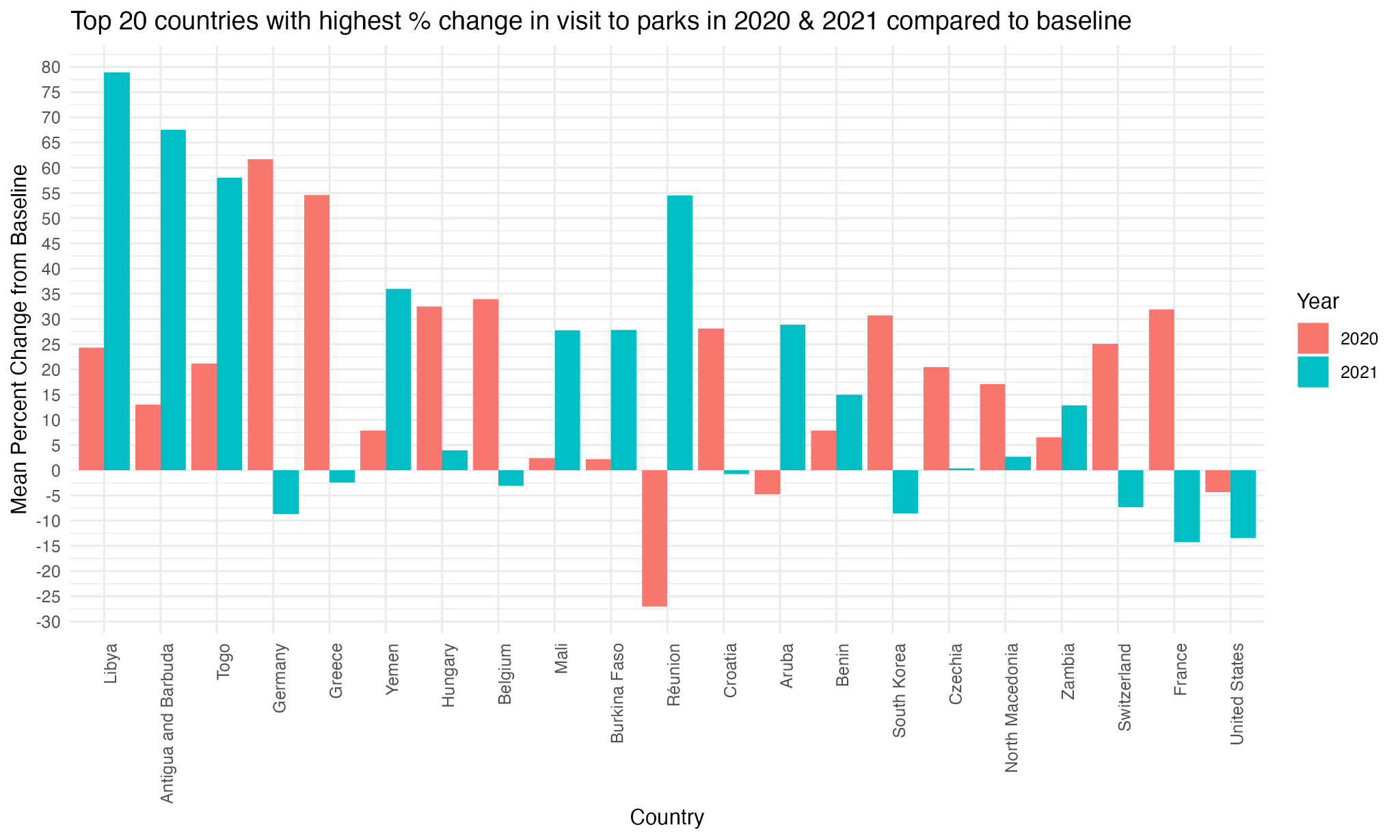
****

The above plot shows change (in percentage) in people's visits to retail and recreation locations for various countries between 2020 and 2021 when compared to baseline. The baseline is a value (normal) given to a day of the week. This day is the median value from the 5-week period (i.e. Jan 3 to Feb 6, 2020). As the analysis for an entire year for a specific country, the baseline is the aggregate of the individual observations recorded for the given year. The chart only focuses on the 20 countries (and the United States) with the highest change. We can clearly see that the visits to these locations in 2020 were significantly lower as compared to the year before COVID-19. For example, Reunion and New Zealand saw the slowdown by about 34% and 18% respectively. But only after a year, there were many countries that saw an increase in the number of visits to these locations again. For example, Reunion and Libya saw an increase of these visits by about 33% and 30% respectively.

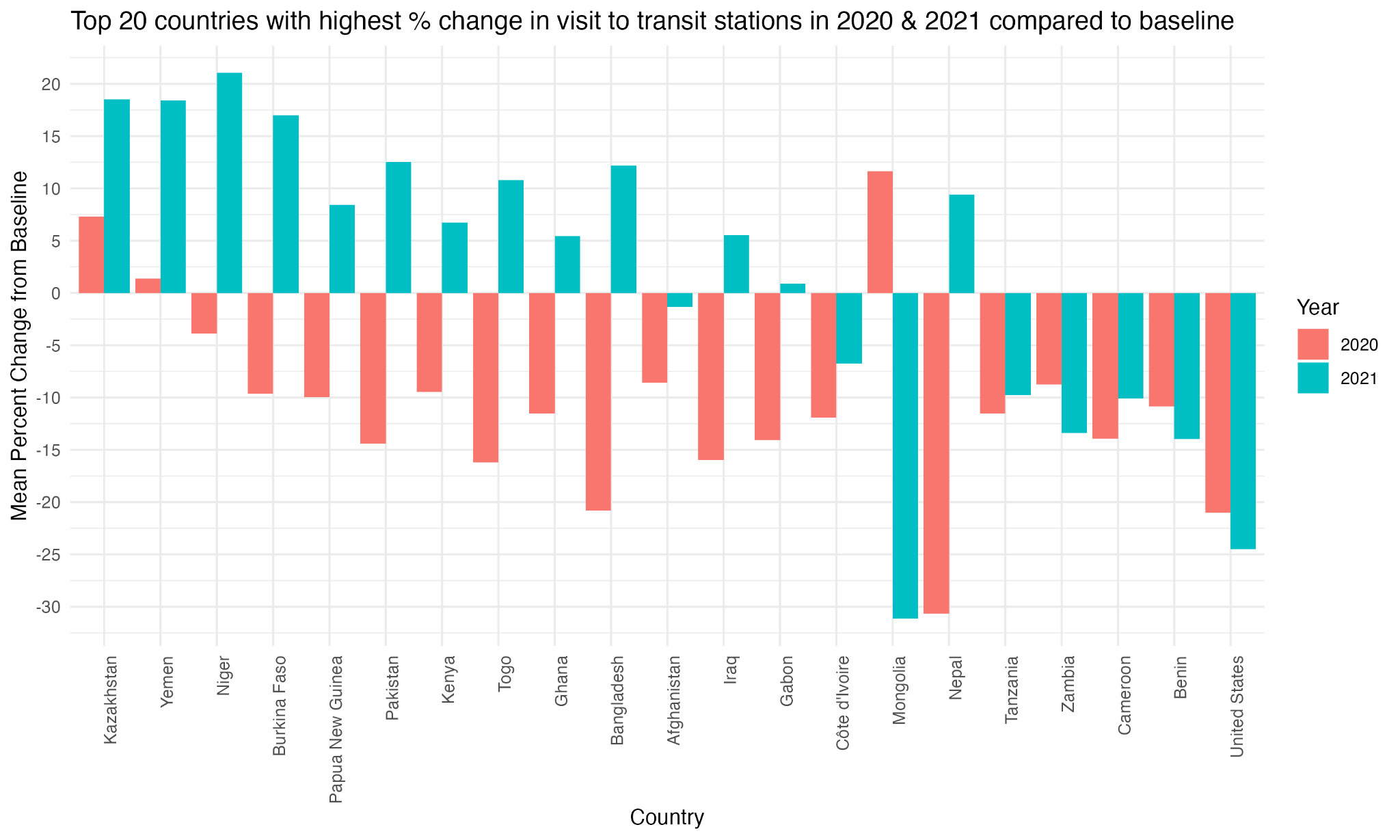
**Data Analysis**



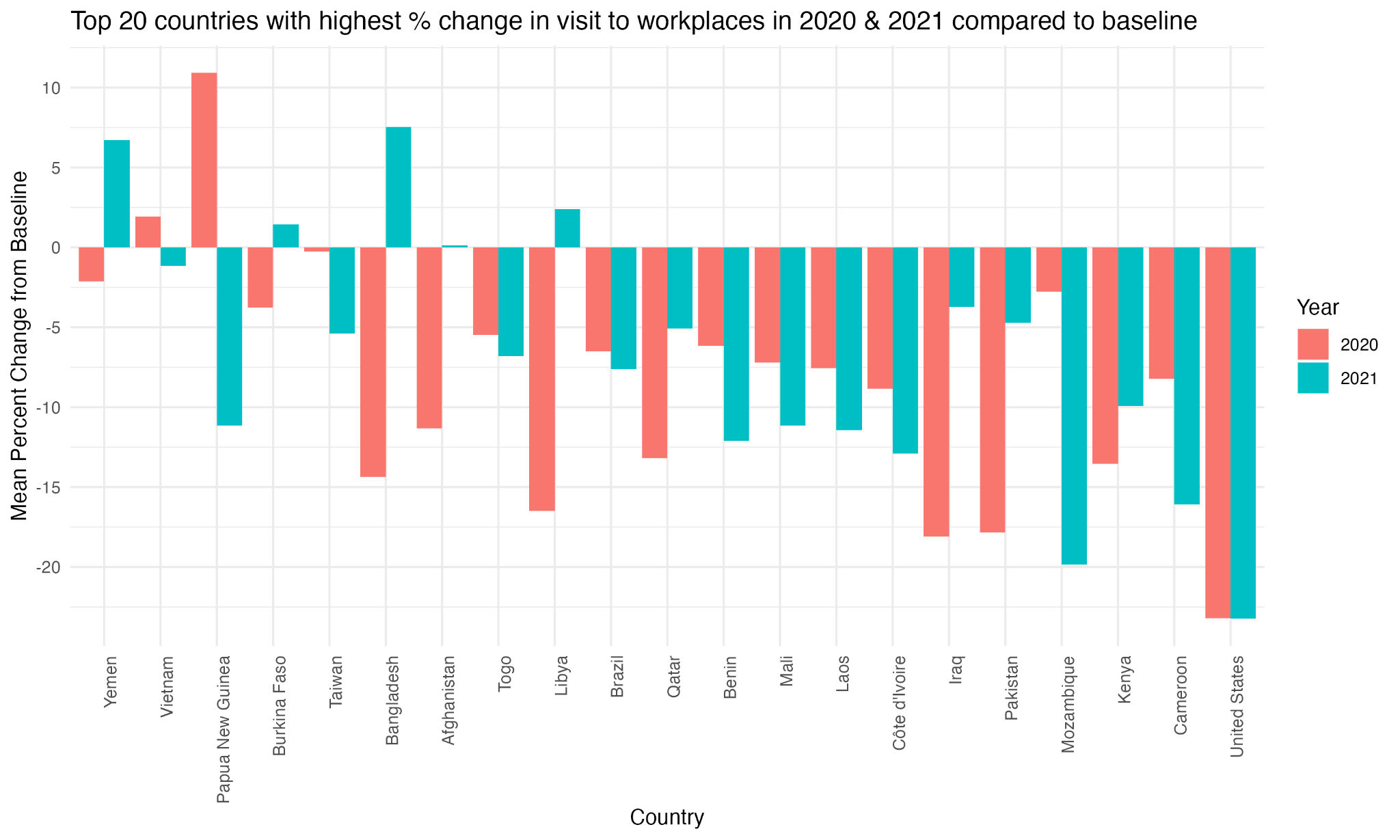
The above chart shows 20 countries (with addition of the United States) that have the highest change (percentage) in the public visits to grocery and pharmacy locations for 2020 and 2021 from the baseline. Compared to the public visits to retail and recreational locations, we can see that the people visited more grocery and pharmacy locations in these countries for both years. This does make sense since the locations such as grocery stores are a more essential part of people's lives than the retail or recreational locations. It also makes sense that the people visited pharmacies more than the retail or recreational locations. It’s clear that the people visited these locations more in 2021 compared to 2020 which indicates that the policies were more relaxed in 2021. It’s interesting to see that the people in the United States made fewer visits to these locations in 2021 than in 2020. It could be due to late response to the pandemic compared to other countries. It could also mean that the United States policies did not relax or made more stringent in 2021. It could also mean that the people were being treated more quickly than other countries which reduced the number of visits to pharmacies. On the other hand, the visits to these locations in Bangladesh and Pakistan were lower in 2020 than in 2021 indicating that the people there perhaps reacted slower to the pandemic than the other countries. Additionally, we can see that the people in Mongolia visited these locations the most amongst all other 20 countries right after the pandemic started (i.e. 2020). This number decreased significantly, from 27% to 1%, for Mongolia in 2021. Overall, this chart gives us a sense of people's visits to grocery stores and pharmacies in these countries between 2020 and 2021 in response to the policies imposed by the respective government officials.



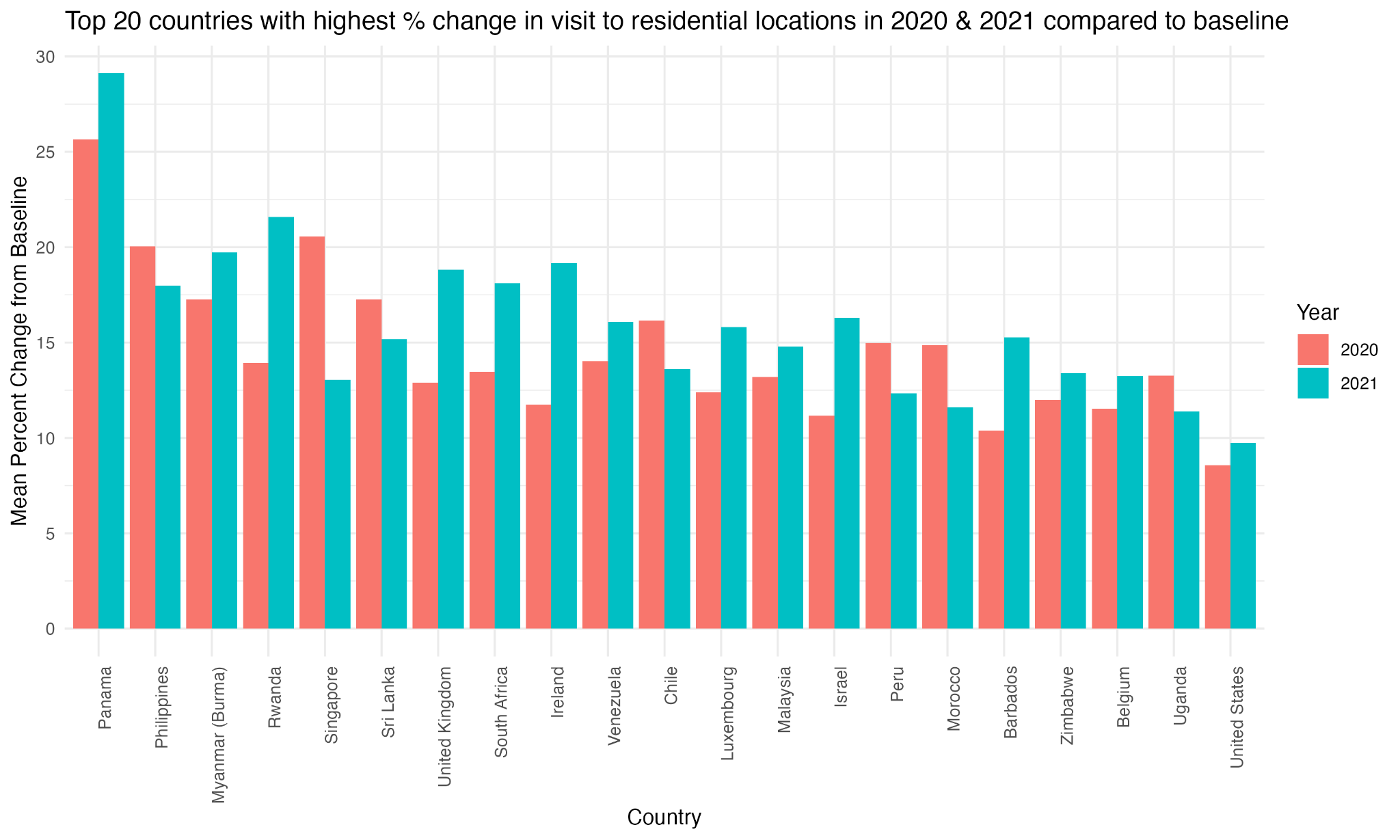
The chart above shows the percentage change in the number of public visits to the parks in the top 20 countries (in addition to the United States) between 2020 and 2021. Similar to the observations made on the previous chart (visits to grocery stores and pharmacies), we can see that the people in most of these countries were still making high visits to the parks in both years. In the year 2020, people in Reunion were making the least number of visits to the parks. But that number increased dramatically in 2021 indicating that either the pandemic policies were relaxed or the public started to cope with the pandemic. It is insightful to see that the people in Germany, France, Switzerland, South Korea, Belgium, and Greece visited the parks more in 2020 than in 2021. There could be several reasons for such behavior. Interestingly, the people in the United States visited the parks less in both years even though they visited the parks lesser in 2021.



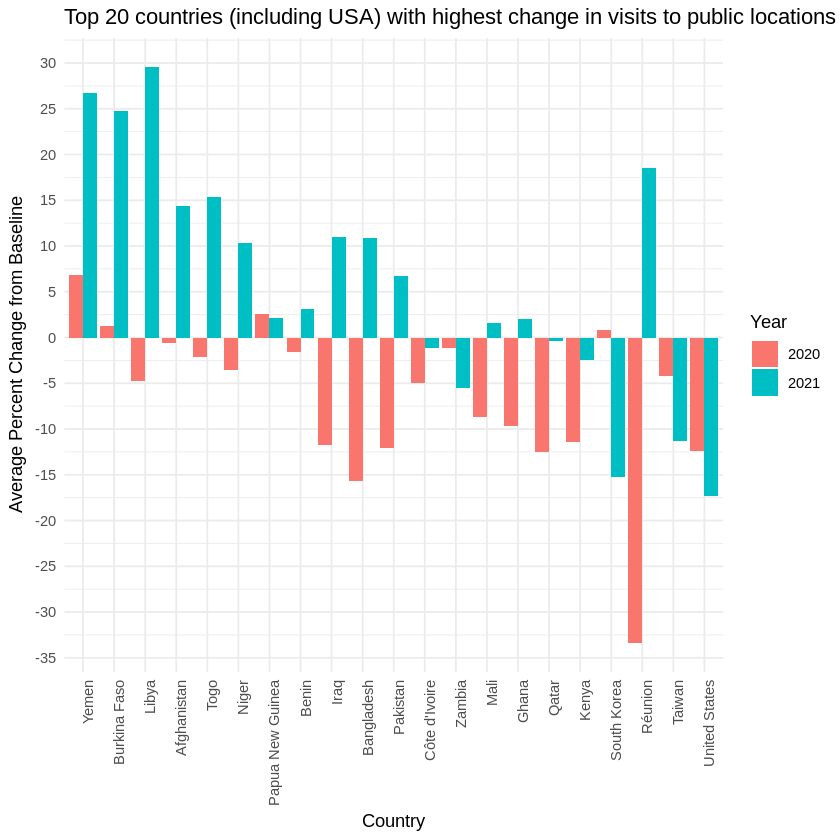
The graph above shows the average change in the number of public visits to public transit stations for the top 20 countries (in addition to the United States) in 2020 and 2021 when compared to the baseline. Similar to what we saw in the graph showing the public visits to the retail and recreational locations, we can see that the people in the majority of the countries visited the transit stations less in 2020. Only the people in Mongolia, Kazakhstan, and Yemen still visited these locations more than the baseline. Similar to previous reasoning, this behavior could be due the respective government policies not being strict enough or the people could have simply ignored those policies. Again, it is interesting to see that the people in Mongolia visited these locations more in 2021 than in 2020. On the contrary, the people in Nepal made significantly less visits to these locations in 2020, and there was slight recovery seen in 2021. On the other hand, the people in the USA still made less visits to these locations in both years. In fact, they visited these locations even less in 2021 as compared to 2020. This indicates that the government restriction policies made the people make less visits to these public transit locations.



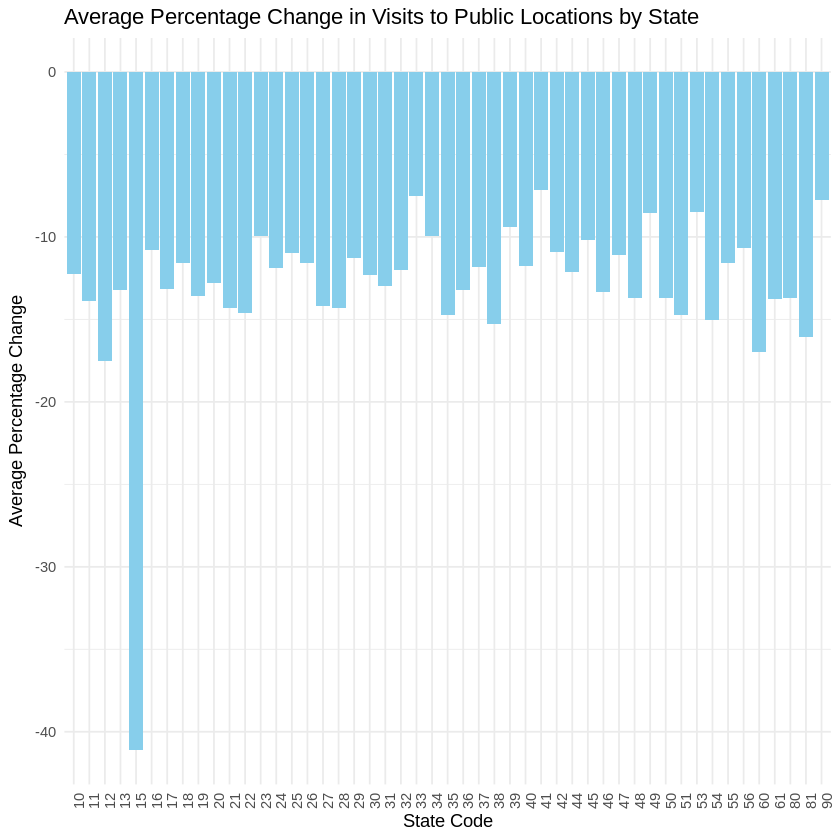
The above graph shows the percentage change of public visits to workplaces in top 20 countries (including USA) for 2020 and 2021 when compared to the baseline. Except for Papua New Guinea, the people in the rest of the other countries made fewer visits to workplaces in both years. It seems like the restriction policies in Papua New Guinea only started to show impact in 2021 as there is negative change in the visits to the workplaces in 2021. It is interesting to see that the people in Bangladesh started going to workplaces more in 2021 than in 2020. It is also interesting to see that the trend remained the same for the USA as the percentage change is almost the same for both years for the USA. This could indicate that the restriction policies remained unchanged from 2020 to 2021 in the USA. All the remaining 20 countries had the people went to workplaces more or less in 2021 compared to 2020.



The above chart shows the percentage change in people staying home for the top 20 countries (in addition to the USA) in 2020 and 2021 when compared to the baseline. As expected, people were staying home more after the pandemic started all around the world. Out of these 21 countries, the people in 13 countries are staying home more in 2021 than in 2020. This indicates that the restrictions were not relaxed or remained effective to keep the people at their residential locations in 2021 too. This observation applies to the people in the USA too.

​​

The above chart shows the average change of public visits to the public locations described in the previous charts in top 20 countries (including USA) for 2020 and 2021 when compared to the baseline. As predicted, we can see that the majority of these countries saw less visits to these locations in 2020, and saw some recovery in 2021. We can see that there were still fewer visits to these locations even in 2021 in the USA , supporting the previous observations on individual public locations previously.



This chart shows the % average change in public visits to public locations (2020 and 2021 combined) for all US states. We can see that the state 15 (i.e. Hawaii) saw the highest change in such visits compared to the rest of the states.

## Modeling and Evaluation

[**Modeling and Evaluation**: What type of model do we apply to the data?

Describe why you chose the particular model, model assumption and limitations, what variable you use for the model, and how well the model works. ]

# Recommendations

[Deployment: Describe how to interpret the model and what **recommendations** you can make based on the findings. How would the stakeholder use the findings and why is the recommendation useful to the stakeholder.]

# Conclusion

[Does the project answer the initial questions? Repeat the key findings and why they are important.]

# List of References

[Cite sources of information in your document and but complete references here. You may use any citation style as long as you are consistent. You can find the basics about how to properly cite references using APA style [here](https://owl.purdue.edu/owl/research_and_citation/apa_style/apa_formatting_and_style_guide/in_text_citations_the_basics.html). ]

Google. (n.d.). *Global mobility report*. Retrieved October 1, 2024, from <https://www.google.com/covid19/mobility/index.html>

U.S. Department of Health and Human Services. (2024). *U.S. spending on COVID-19 as of October 2024*. Retrieved October 1, 2024, from USASPENDING.gov: The Federal Response to COVID-19.

# Appendix

[Put code and long tables that you do not need inside your report here. ]

## Student Contributions

Add a list with who contributed to what part of this report.