World Cities Average Internet Prices EDA with R

Zarina U.

Last edited October 30, 2021

Introduction

In this report, we're going to find out:
1. Top 10 country of the highest and the lowest average internet prices during 2010-2020
2. Description of average internet prices each continent in every year
3. The Outlier of internet prices in every year
4. Comparison of Indonesia's internet prices among other Southeast Asian Country

Dataset

Dataset that will be used in this notebook is World cities average internet prices on 2010 - 2020. You can go to this link to download the dataset -> https://www.kaggle.com/cityapiio/world-cities-average-internet-prices-2010-2020

Load Package

```
library(readr)
library(dplyr)
library(tidyr)
library(ggplot2)
library(forcats)
library(stringr)
```

Data Preparation

Input Data

```
internet_prices <- read_csv("cities_internet_prices_historical.24-10-2021.csv")

# fix variable name
internet_prices <- internet_prices %>%
  select(city = City,
         region = Region,
         country = Country,
         price_2010 = `Internet Price, 2010`,
         price_2011 = `Internet Price, 2011`,
```

```
         price_2012 = `Internet Price, 2012`,
         price_2013 = `Internet Price, 2013`,
         price_2014 = `Internet Price, 2014`,
         price_2015 = `Internet Price, 2015`,
         price_2016 = `Internet Price, 2016`,
         price_2017 = `Internet Price, 2017`,
         price_2018 = `Internet Price, 2018`,
         price_2019 = `Internet Price, 2019`,
         price_2020 = `Internet Price, 2020`,)
```

Checking data type

Based on the structure of `internet_prices`, we can see that all variables have corresponding data type. Column city, region and country are character, and column prices are numeric.

```
str(internet_prices)
```

```
## tibble [695 x 14] (S3: tbl_df/tbl/data.frame)
##  $ city      : chr [1:695] "New York City" "Washington, D.C." "San Francisco" "Berlin" ...
##  $ region    : chr [1:695] "New York" "District of Columbia" "California" NA ...
##  $ country   : chr [1:695] "United States of America" "United States of America" "United States of Am
##  $ price_2010: num [1:695] 43 35 41.7 30.9 35.8 ...
##  $ price_2011: num [1:695] 50 48 42.5 38.6 43.6 ...
##  $ price_2012: num [1:695] 42.2 49.3 36.5 20.5 41.5 ...
##  $ price_2013: num [1:695] 55.9 45.8 48.5 30.1 37.5 ...
##  $ price_2014: num [1:695] 50.3 47 47 32.6 40.3 ...
##  $ price_2015: num [1:695] 53.5 56.3 55.3 27.1 47.6 ...
##  $ price_2016: num [1:695] 54.6 59.2 53.1 26.5 50.8 ...
##  $ price_2017: num [1:695] 59.2 64.2 58.8 32.5 56.3 ...
##  $ price_2018: num [1:695] 61.5 64.9 63.2 36.4 57.5 ...
##  $ price_2019: num [1:695] 61.7 69.1 67.4 33.9 63 ...
##  $ price_2020: num [1:695] 66.4 60.6 69 35.8 66.7 ...
```

Checking missing value

There are 92 missing values on region column. But, in this case, we will ignore this column because we just want to see the difference based on country instead of region.

```
missing_values <- function(df) {
    missing_val <- c()
    for (c in colnames(df)) {
        missing <- sum(is.na(df[[c]]))
        missing_val <- c(missing_val, missing)
    }
    tibble(colnames(df), missing_val)
}
missing_values(internet_prices)
```

```
## # A tibble: 14 x 2
##    `colnames(df)` missing_val
##    <chr>                <int>
```

```
##  1 city                    0
##  2 region                 92
##  3 country                 0
##  4 price_2010              0
##  5 price_2011              0
##  6 price_2012              0
##  7 price_2013              0
##  8 price_2014              0
##  9 price_2015              0
## 10 price_2016              0
## 11 price_2017              0
## 12 price_2018              0
## 13 price_2019              0
## 14 price_2020              0
```

Add continent variable

We want to know the description of average internet prices for each continent in every year. So, this dataset will be useful later.

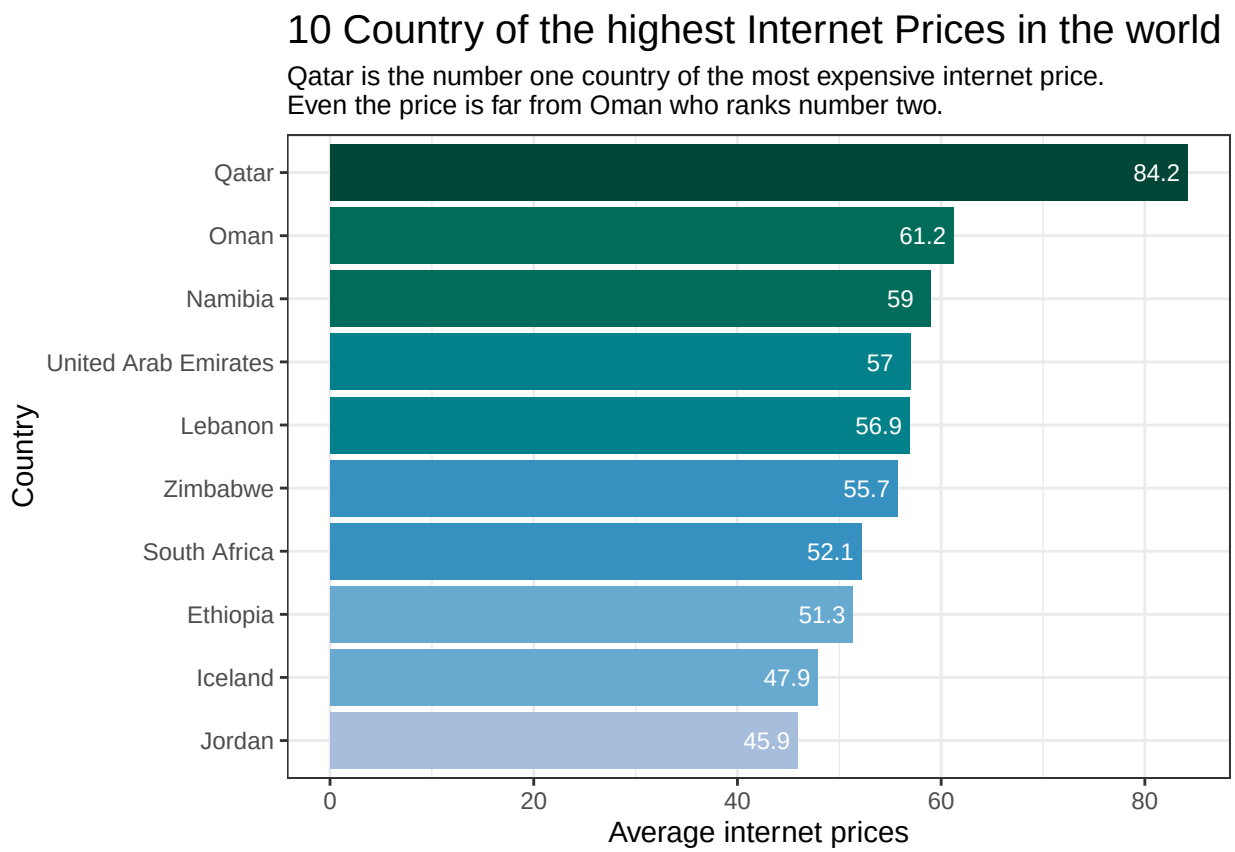
```
# load data continent
data_continent <- read_csv("continents2.csv")
data_continent <- data_continent %>%
  select(country = name, continent = region)

# add several countries that haven't in data_continent
add <- data.frame(country = c("United States of America",
"North Macedonia", "People's Republic of China", "The Bahamas", "Brunei", "Bosnia and Herzegovina", "An

# combine the data
data_continent <- bind_rows(data_continent, add)
```

THE QUESTION

Which country of the highest and the lowest average internet prices during 2010-2020?

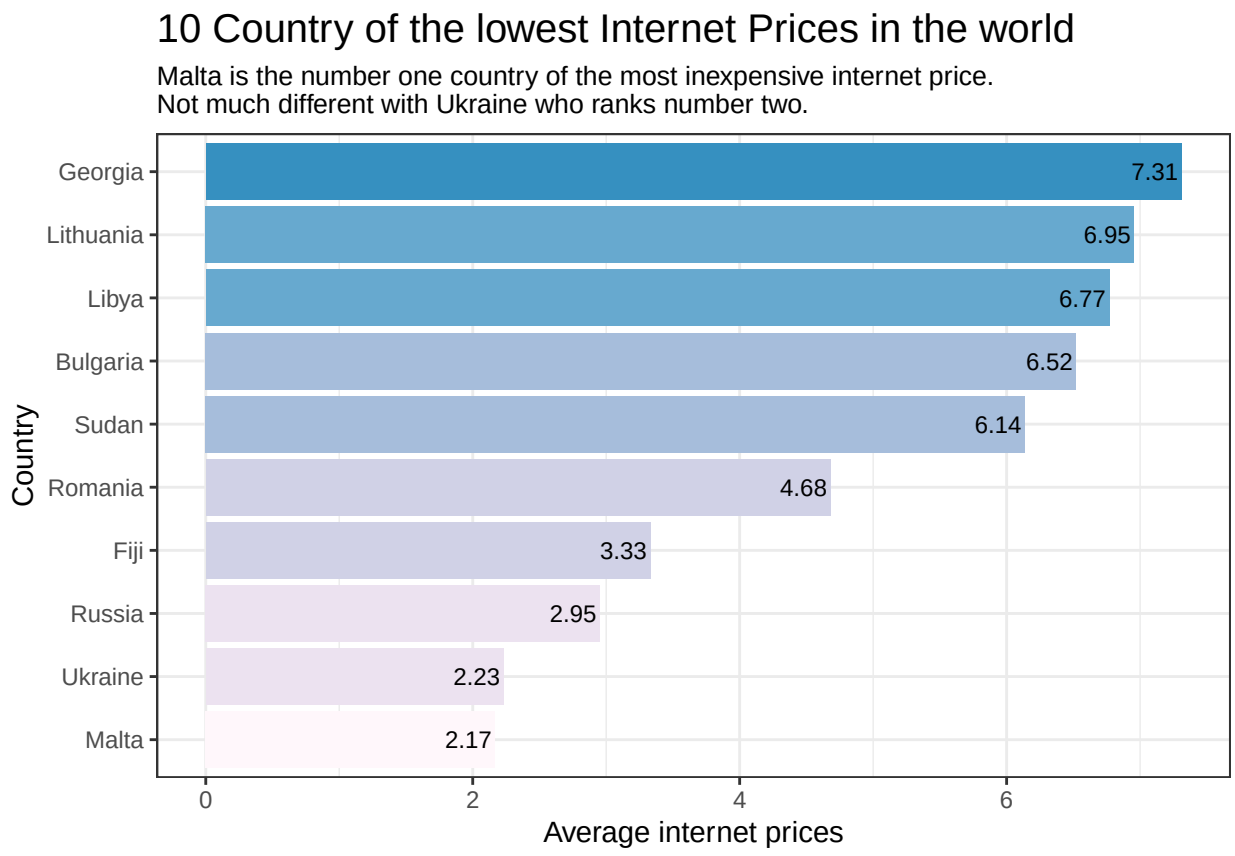
```
gather_year <- internet_prices %>%
  gather(year, price, price_2010:price_2020)

gather_year$year <- str_remove_all(gather_year$year, "price_")

# the highest
top10 <-  gather_year %>%
  group_by(country) %>%
  summarize(mean = mean(price)) %>%
  arrange(desc(mean))
top10 <- top10[1:10,]

ggplot(top10, aes(x = fct_reorder(country, mean), y = mean)) +
```

```
  geom_col(fill = c("#014636", "#016C59", "#016C59", "#02818A", "#02818A", "#3690C0", "#3690C0",  "#67A
  geom_text(aes(label = round(mean, digits = 1)), color = "white", size = 3, nudge_y = -3) +
  coord_flip() +
  labs(
    title = "10 Country of the highest Internet Prices in the world",
    subtitle = "Qatar is the number one country of the most expensive internet price.\nEven the price is
    x = "Country",
    y = "Average internet prices"
  ) +
  theme_bw() +
  theme(
    plot.title = element_text(size = 15),
    plot.subtitle = element_text(size = 10)
    )
```




```
# the lowest
below10 <- gather_year %>%
  group_by(country) %>%
  summarize(mean = mean(price)) %>%
  arrange(mean)
below10 <- below10[1:10,]

ggplot(below10, aes(x = fct_reorder(country, mean), y = mean)) +
  geom_col(fill = c("#FFF7FB", "#ECE2F0", "#ECE2F0", "#D0D1E6", "#D0D1E6", "#A6BDDB", "#A6BDDB",  "#67A
  coord_flip() +
```

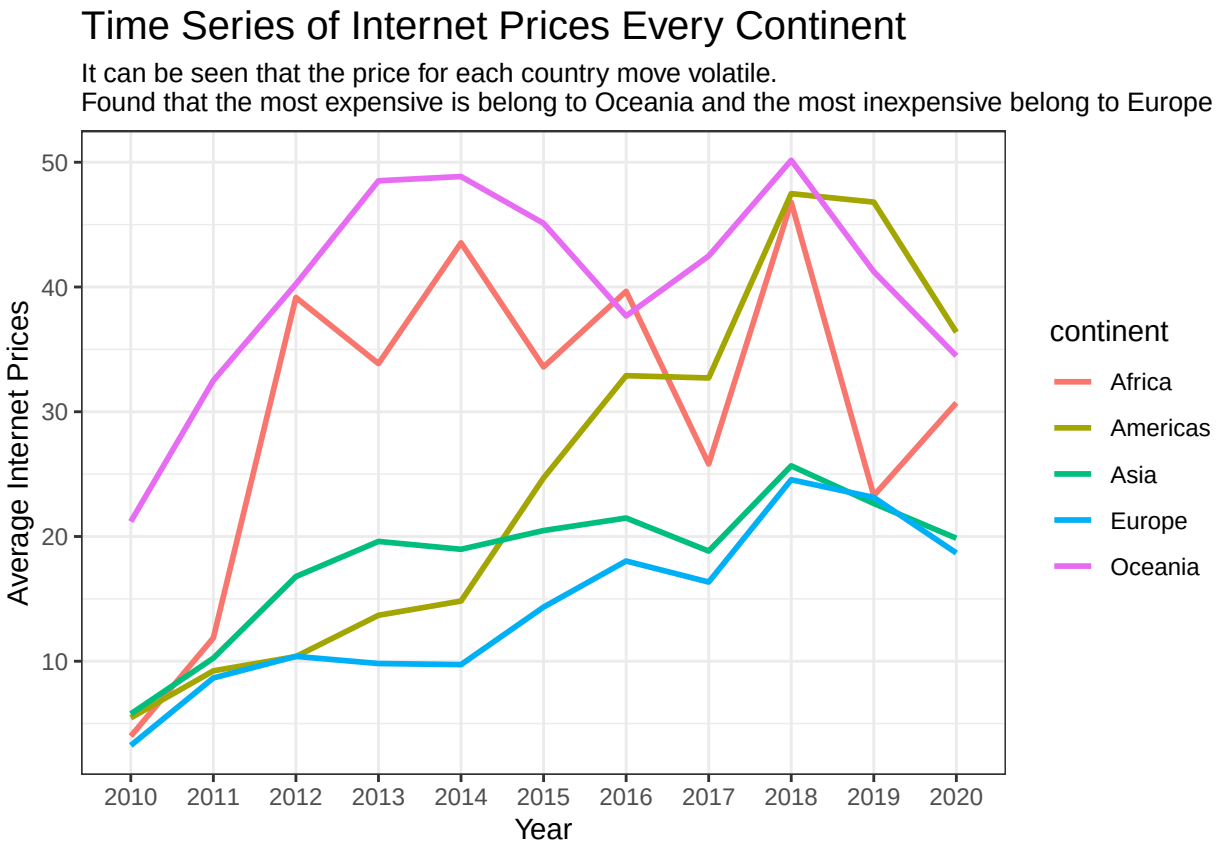
```
  geom_text(aes(label = round(mean, digits = 2)), color = "black", size = 3, nudge_y = -0.2) +
  coord_flip() +
  labs(
    title = "10 Country of the lowest Internet Prices in the world",
    subtitle = "Malta is the number one country of the most inexpensive internet price.\nNot much diffe
    x = "Country",
    y = "Average internet prices"
  ) +
  theme_bw() +
  theme(
    plot.title = element_text(size = 15),
    plot.subtitle = element_text(size = 10)
    )
```



The description of average internet prices each continent every year

```
# join with continent data
join_continent <- gather_year %>%
  inner_join(data_continent, by = c("country" = "country"))

by_continent_year <- join_continent %>%
  group_by(year, continent) %>%
  summarise(mean = mean(price), .groups = "keep")
```

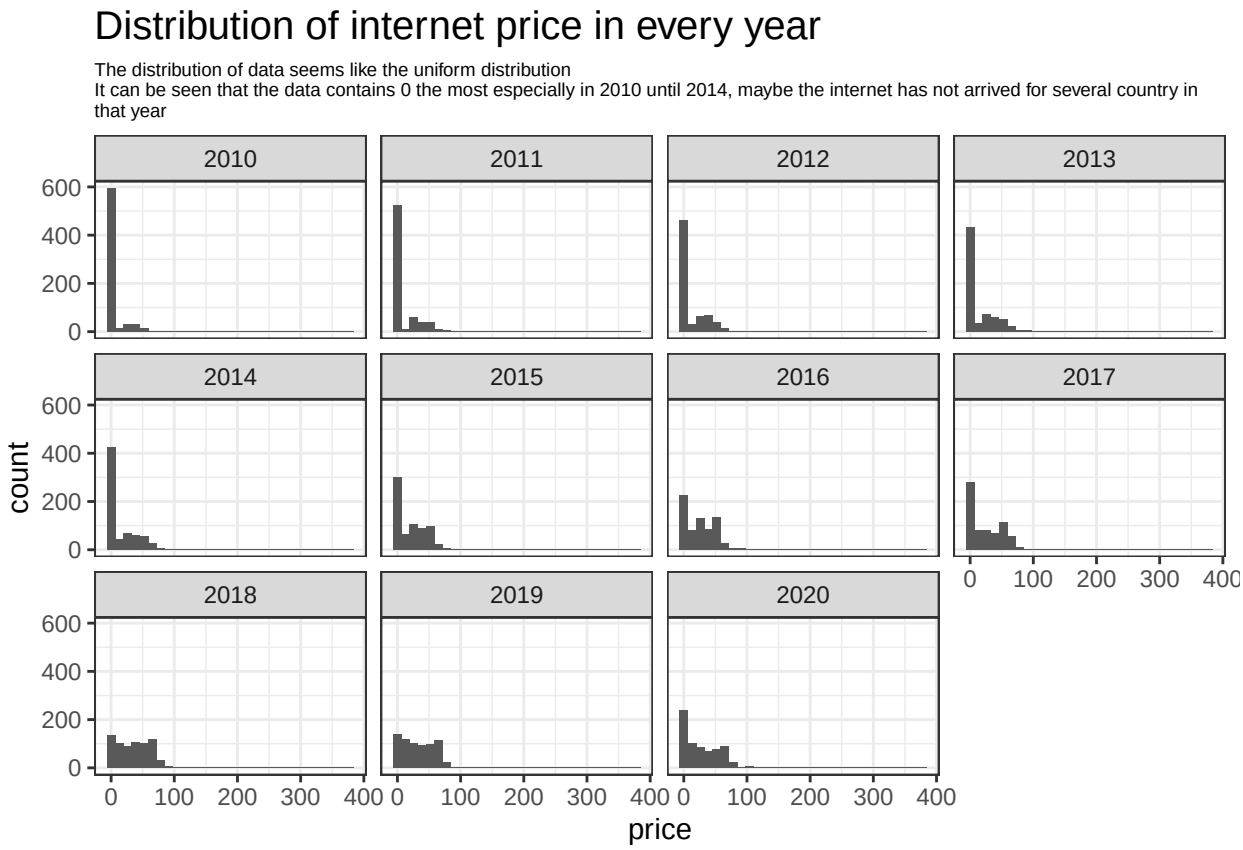
```r
#plot
ggplot(by_continent_year, aes(x = year, y = mean, group = continent)) +
  geom_line(aes(colour = continent), size = 1) +

  labs(
    title = "Time Series of Internet Prices Every Continent",
    subtitle = "It can be seen that the price for each country move volatile.\nFound that the most expe
    x = "Year",
    y = "Average Internet Prices"
  ) +
  theme_bw() +
  theme(
    plot.title = element_text(size = 15),
    plot.subtitle = element_text(size = 10)
    )
```



Is there any outlier of internet prices in each year?

```r
# distribution of internet price in every year
gather_year %>%
  ggplot(aes(price)) +
  geom_histogram(bins = 30) +
  facet_wrap(~ year) +
```

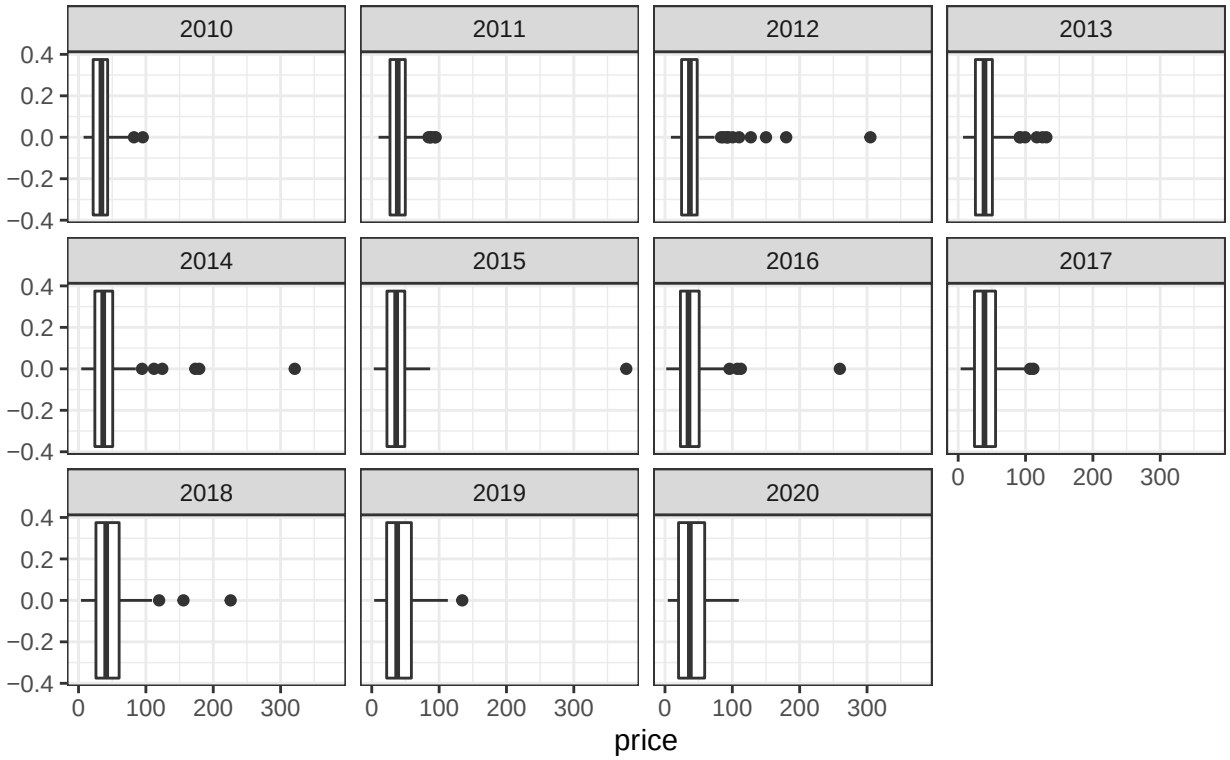
```
  labs(
    title = "Distribution of internet price in every year",
    subtitle = "The distribution of data seems like the uniform distribution\nIt can be seen that the da
  ) +
  theme_bw() +
  theme(
    plot.title = element_text(size = 15),
    plot.subtitle = element_text(size = 7)
    )
```




```
# check the outlier
gather_year %>%
  filter(price != 0.000000) %>%
  ggplot(aes(price)) +
  geom_boxplot() +
  facet_wrap(~ year) +
  labs(
    title = "The Outlier",
    subtitle = "Based on boxplot, the outlier is shown by the dot. As we can see that every year has the
  ) +
  theme_bw() +
  theme(
    plot.title = element_text(size = 15),
    plot.subtitle = element_text(size = 7)
  )
```

The Outlier

Based on boxplot, the outlier is shown by the dot. As we can see that every year has the outlier except 2020.




```
outlier <- gather_year %>%
  filter(price > 120) %>%
  arrange(desc(price))
outlier[,3:5]
```

```
## # A tibble: 15 x 3
##    country                   year  price
##    <chr>                     <chr> <dbl>
##  1 Rwanda                    2015   378.
##  2 Namibia                   2014   321.
##  3 Ethiopia                  2012   305
##  4 Ethiopia                  2016   260.
##  5 Mozambique                2018   226.
##  6 Angola                    2012   180
##  7 Uzbekistan                2014   179.
##  8 Angola                    2014   173.
##  9 Afghanistan               2018   156.
## 10 United States of America 2012   150
## 11 United States of America 2019   134.
## 12 Ghana                    2013   131.
## 13 South Africa             2012   127.
## 14 Namibia                  2013   125.
## 15 Tanzania                 2014   124.
```

Comparison of Indonesia's internet prices among other Southeast Asian Country

```
sea_country <- c("Cambodia", "Myanmar", "Thailand", "Vietnam", "Brunei", "Philippines", "Indonesia", "Ma

by_sea_country <- gather_year %>%
  filter(country %in% sea_country) %>%
  group_by(country) %>%
  summarise(mean = mean(price))

mean_sea_country <- mean(by_sea_country$mean)

ggplot(by_sea_country, aes(x = fct_reorder(country, mean, .desc = TRUE), y = mean)) +
  geom_col(fill = rep(c("grey","#D35151","grey"), times = c(2,1,6))) +
  geom_hline(yintercept = mean_sea_country) +
  geom_text(aes(label = round(mean, digits = 2)), color = "black", size = 3, nudge_y = -1) +
  annotate(
    "text",
    x = 9, y = mean_sea_country+5,
    label = c(round(mean_sea_country, digits = 2), "\naverage\nprice"),
    vjust = 0.95, size = 3, color = "black"
  ) +
  labs(
    title = "Indonesia internet prices among other Southeast Asian Country",
    subtitle = "Internet prices in Indonesia belong to the cheapest among the Southeast Asian Country.\n
    x = "Country",
    y = "Average internet prices"
  ) +
  theme_bw() +
  theme(
    plot.title = element_text(size = 15),
    plot.subtitle = element_text(size = 10)
    )
```

Indonesia internet prices among other Southeast Asian Country

Internet prices in Indonesia belong to the cheapest among the Southeast Asian Country.
Indonesia ranks #3 for the cheapest internet prices among its neighbor