Happiness, Internet Use, & Human Development Index Around the World.Rmd

2023-03-03

Happiness, Internet Use, and Human Development Index Around the World

Group members: Avery Zuckerman, Kaitlyn Rouse, and Bella Crain

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.0     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.1     ✔ tibble    3.1.8
## ✔ lubridate 1.9.2     ✔ tidyr     1.3.0
## ✔ purrr     1.0.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the □]8;;http://conflicted.r-lib.org/□conflicted package□]8;;□ to force all
conflicts to become errors
```

1.) Introduction

a. Description of Datasets

The “Happiness and Corruption Globally” dataset measures happiness on a global scale by country but was renamed when imported as “Happy”. For this dataset, we intend to use the variable relating to happiness score (numeric), freedom (numeric), and the common variable of country (categorical). The “Global Human Development Index” dataset uses multiple ways to describe the human development index by country but was renamed when imported as “HDI”. For this dataset, we intend to use the variables of the human development group (categorical), human development index (2020) (numeric), and the common variable of country (categorical). The “Global Internet Usage” dataset measures the use of the internet globally by country but was renamed when imported as “Internet”. For this dataset, we intend to use the variables of internet use rate (numeric), urban rate (numeric), and the common

variable of country (categorical). The datasets being utilized for the project were acquired using Kaggle. Each row in the datasets represents a distinct county in the world. We will join the datasets by country. An expected trend from joining the three datasets would be that a higher internet usage would be correlated with a higher human development index score and a higher happiness score. These datasets were interesting to our group because internet and technological development impacts our everyday lives, so we wanted to investigate how such factors may impact the happiness of individuals around the world.

```
## [1] "/stor/home/az7885/Project"
```

```
## Rows: 213 Columns: 4
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (1): country
## dbl (3): incomeperperson, internetuserate, urbanrate
## 
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 792 Columns: 13
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (2): Country, continent
## dbl (11): happiness_score, gdp_per_capita, family, health, freedom, generosi...
## 
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 195 Columns: 880
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr   (4): ISO3, Country, Human Development Groups, UNDP Developing Regions
## dbl (876): HDI Rank (2021), Human Development Index (1990), Human Developmen...
## 
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

b. Defining Research Questions

The first research question aims to discover “What is the relationship between internet usage, happiness score, and HDI score?”. We might expect to see that increased internet usage and HDI score would be correlated with a increase in happiness score. The second research question tackles the question of “What is the relationship between happiness score and the human development groups?”. We might expect to see that increased human development groups would be related to higher

happiness scores. Finally, the third research question examines “How does freedom correlate with the urban rate?” We might expect to see that increased freedom is associated with an increased urban rate.

2.) Tidying and Wrangling Datasets

```
# Tidying the Data for the Human Development Index Dataset by Human Development Index
Year
HDI_tidy <-  pivot_longer(HDI,
                              cols = c(6:37),
                              names_to = "HDI_year",
                              values_to = "HDI_score")
# Wrangling the 'HDI' Dataset to Select for Desired Variables and Filter by Year 2020
new_HDI <- HDI_tidy %>%
  dplyr::select("Country", "HDI_year", "Human Development Groups", "HDI_score")%>%
  filter(HDI_year =="Human Development Index (2020)")
new_HDI
```

```
## # A tibble: 195 × 4
##    Country              HDI_year                       Human Developme…¹ HDI_s…²
##    <chr>                <chr>                          <chr>               <dbl>
##  1 Afghanistan          Human Development Index (2020) Low                 0.483
##  2 Angola               Human Development Index (2020) Medium              0.59 
##  3 Albania              Human Development Index (2020) High                0.794
##  4 Andorra              Human Development Index (2020) Very High           0.848
##  5 United Arab Emirates Human Development Index (2020) Very High           0.912
##  6 Argentina            Human Development Index (2020) Very High           0.84 
##  7 Armenia              Human Development Index (2020) High                0.757
##  8 Antigua and Barbuda  Human Development Index (2020) High                0.788
##  9 Australia            Human Development Index (2020) Very High           0.947
## 10 Austria              Human Development Index (2020) Very High           0.913
## # … with 185 more rows, and abbreviated variable names
## #   ¹`Human Development Groups`, ²HDI_score
```

```
#Determining how many Observations are in 'new_HDI'
nrow(new_HDI)
```

```
## [1] 195
```

```
# 'Happiness' Dataset is Already Tidy
# Wrangling 'Happiness' Dataset to Select for Desired Variables and Filter by Year 20
20
new_happiness <- Happy %>%
  dplyr::select("Country", "freedom", "happiness_score", "Year") %>%
  filter(Year == 2020)
new_happiness
```

```
## # A tibble: 132 × 4
##    Country     freedom happiness_score  Year
##    <chr>         <dbl>           <dbl> <dbl>
##  1 Finland       0.662            7.81  2020
##  2 Denmark       0.665            7.65  2020
##  3 Switzerland   0.629            7.56  2020
##  4 Iceland       0.662            7.50  2020
##  5 Norway        0.670            7.49  2020
##  6 Netherlands   0.614            7.45  2020
##  7 Sweden        0.650            7.35  2020
##  8 New Zealand   0.647            7.30  2020
##  9 Austria       0.603            7.29  2020
## 10 Luxembourg    0.610            7.24  2020
## # … with 122 more rows
```

```
# Determining how many Observations are in 'new_happiness'
nrow(new_happiness)
```

```
## [1] 132
```

```
# 'Internet' Dataset is Already Tidy
# Wrangling 'Internet' Dataset to Select for Desired Variables
new_internet <- Internet %>%
  dplyr::select("country", "internetuserate", "urbanrate")
new_internet
```

```
## # A tibble: 213 × 3
##    country             internetuserate urbanrate
##    <chr>                         <dbl>     <dbl>
##  1 Afghanistan                    3.65      24.0
##  2 Albania                       45.0       46.7
##  3 Algeria                       12.5       65.2
##  4 Andorra                       81         88.9
##  5 Angola                        10.0       56.7
##  6 Antigua and Barbuda           80.6       30.5
##  7 Argentina                     36.0       92  
##  8 Armenia                       44.0       63.9
##  9 Aruba                         41.8       46.8
## 10 Australia                     75.9       88.7
## # … with 203 more rows
```

```
# Determining how many Observations are in 'new_internet'
nrow(new_internet)
```

```
## [1] 213
```

The number of observations in ‘new_HDI’ is 195, in ‘new_happiness’ there are 132 observations, and in ‘new_internet’ there are 213 observations. For the ‘new_happiness’ dataset, we intend to use the variable relating to happiness score (numeric) and freedom (numeric). In the ‘new_HDI’ dataset, we intend to use the variables of the human development groups (categorical) and human development index (2020) (numeric). In the ‘new_internet’ dataset, we intend to use the variables of internet use rate (numeric) and urban rate (numeric). The common variable between the three datasets is the ID variable of country. There were no IDs that were left out after joining, as we filtered prior to joining.

3.) Joining the Datasets

```
# Joining the 'Happiness' Dataset with the 'Internet' Dataset
happy_int <- left_join(new_happiness, new_internet, by = c("Country" = "country"))
happy_int
```

```
## # A tibble: 132 × 6
##    Country     freedom happiness_score  Year internetuserate urbanrate
##    <chr>         <dbl>           <dbl> <dbl>           <dbl>     <dbl>
##  1 Finland       0.662            7.81  2020            86.9      63.3
##  2 Denmark       0.665            7.65  2020            88.8      86.7
##  3 Switzerland   0.629            7.56  2020            82.2      73.5
##  4 Iceland       0.662            7.50  2020            95.6      92.3
##  5 Norway        0.670            7.49  2020            93.3      77.5
##  6 Netherlands   0.614            7.45  2020            90.7      81.8
##  7 Sweden        0.650            7.35  2020            90.0      84.5
##  8 New Zealand   0.647            7.30  2020            83.0      86.6
##  9 Austria       0.603            7.29  2020            72.7      67.2
## 10 Luxembourg    0.610            7.24  2020            90.1      82.4
## # … with 122 more rows
```

```
# Joining the 'HDI' Dataset with the Merged 'happy_int' Dataset
complete_data <- left_join(happy_int, new_HDI, by = "Country")
complete_data
```

```
## # A tibble: 132 × 9
##    Country     freedom happiness…¹  Year inter…² urban…³ HDI_y…⁴ Human…⁵ HDI_s…⁶
##    <chr>         <dbl>       <dbl> <dbl>   <dbl>   <dbl> <chr>   <chr>     <dbl>
##  1 Finland       0.662        7.81  2020    86.9    63.3 Human … Very H…   0.938
##  2 Denmark       0.665        7.65  2020    88.8    86.7 Human … Very H…   0.947
##  3 Switzerland   0.629        7.56  2020    82.2    73.5 Human … Very H…   0.956
##  4 Iceland       0.662        7.50  2020    95.6    92.3 Human … Very H…   0.957
##  5 Norway        0.670        7.49  2020    93.3    77.5 Human … Very H…   0.959
##  6 Netherlands   0.614        7.45  2020    90.7    81.8 Human … Very H…   0.939
##  7 Sweden        0.650        7.35  2020    90.0    84.5 Human … Very H…   0.942
##  8 New Zealand   0.647        7.30  2020    83.0    86.6 Human … Very H…   0.936
##  9 Austria       0.603        7.29  2020    72.7    67.2 Human … Very H…   0.913
## 10 Luxembourg    0.610        7.24  2020    90.1    82.4 Human … Very H…   0.924
## # … with 122 more rows, and abbreviated variable names ¹happiness_score,
## #   ²internetuserate, ³urbanrate, ⁴HDI_year, ⁵`Human Development Groups`,
## #   ⁶HDI_score
```

```
# Determining how many Observations are in Completely Merged 'complete_data' Dataset
nrow(complete_data)
```

```
## [1] 132
```

```
# Selecting for Desired Variables within the Combined Dataset
complete_data <- complete_data %>%
  dplyr::select("Country", "freedom", "happiness_score", "internetuserate", "urbanrat
e", "Human Development Groups", "HDI_score") %>%
  na.omit()
complete_data
```

```
## # A tibble: 124 × 7
##    Country     freedom happiness_score internetuserate urbanrate Human…¹ HDI_s…²
##    <chr>         <dbl>           <dbl>           <dbl>     <dbl> <chr>     <dbl>
##  1 Finland       0.662            7.81            86.9      63.3 Very H…   0.938
##  2 Denmark       0.665            7.65            88.8      86.7 Very H…   0.947
##  3 Switzerland   0.629            7.56            82.2      73.5 Very H…   0.956
##  4 Iceland       0.662            7.50            95.6      92.3 Very H…   0.957
##  5 Norway        0.670            7.49            93.3      77.5 Very H…   0.959
##  6 Netherlands   0.614            7.45            90.7      81.8 Very H…   0.939
##  7 Sweden        0.650            7.35            90.0      84.5 Very H…   0.942
##  8 New Zealand   0.647            7.30            83.0      86.6 Very H…   0.936
##  9 Austria       0.603            7.29            72.7      67.2 Very H…   0.913
## 10 Luxembourg    0.610            7.24            90.1      82.4 Very H…   0.924
## # … with 114 more rows, and abbreviated variable names
## #   ¹`Human Development Groups`, ²HDI_score
```

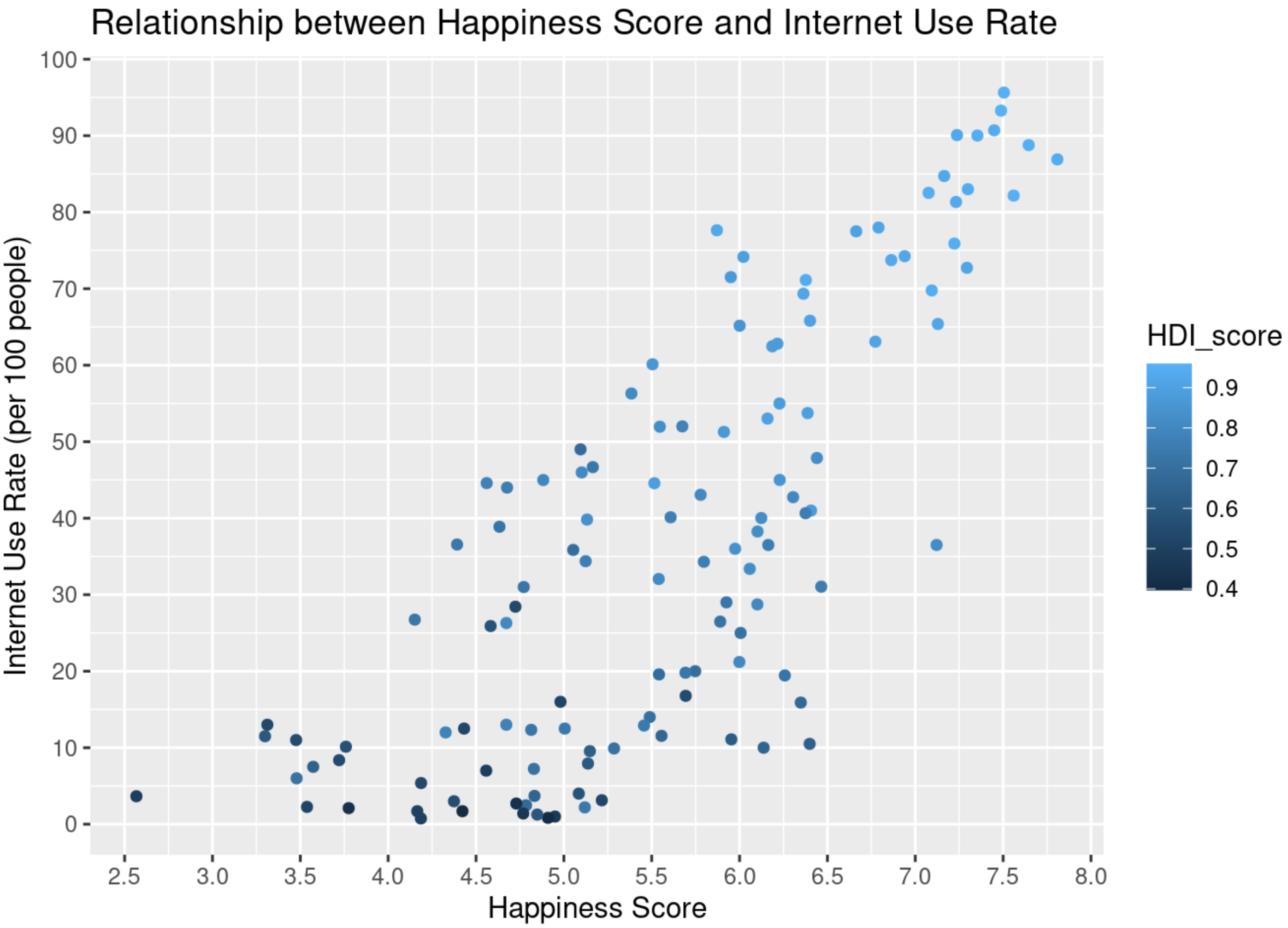
There are 132 rows in the joined dataset ‘complete_data’. The only ID variable in common throughout all 3 datasets was the country variable. For the tidied ‘new_HDI’ dataset the ID variables included ‘HDI_year’, ‘HDI_score’, and ‘Human Development Groups’. For the tidied ‘new_happiness’ dataset the ID variables included ‘freedom’, ‘happiness_score’, and ‘Year’. For the tidied ‘new_internet’ dataset the ID variables included ‘internetuserate’ and ‘urbanrate’. There are 81 observations that were removed considering ‘new_internet’ had 213 observations prior to joining. This could potentially leave out some data points and could possibly present less accurate or misleading data.

4.) Research Question 1: “What is the relationship between internet usage, happiness score, and HDI Score?”

```
# Exploring the Relationship Between Internet Usage and Happiness Score
complete_data%>%
  dplyr::select("internetuserate","happiness_score", "Country") %>%
  arrange(desc("happiness_score"))
```

```
## # A tibble: 124 × 3
##    internetuserate happiness_score Country    
##              <dbl>           <dbl> <chr>      
##  1            86.9            7.81 Finland    
##  2            88.8            7.65 Denmark    
##  3            82.2            7.56 Switzerland
##  4            95.6            7.50 Iceland    
##  5            93.3            7.49 Norway     
##  6            90.7            7.45 Netherlands
##  7            90.0            7.35 Sweden     
##  8            83.0            7.30 New Zealand
##  9            72.7            7.29 Austria    
## 10            90.1            7.24 Luxembourg 
## # … with 114 more rows
```

```
# Visualization 1 for Research Question 1
complete_data%>%
  ggplot(aes(x= happiness_score, y= internetuserate, color = HDI_score))+
  geom_point()+
  theme_gray()+
  labs(title = "Relationship between Happiness Score and Internet Use Rate",
       x = "Happiness Score", y = "Internet Use Rate (per 100 people)")+
  scale_x_continuous(breaks = seq(0,10,0.5))+
   scale_y_continuous(breaks = seq(0,100,10))
```



```
# Summary Statistics for the Visualization
cor(complete_data$internetuserate, complete_data$happiness_score, use = "pairwise.com
plete.obs")
```

```
## [1] 0.7778816
```

```
cor(complete_data$internetuserate, complete_data$HDI_score, use = "pairwise.complete.
obs")
```

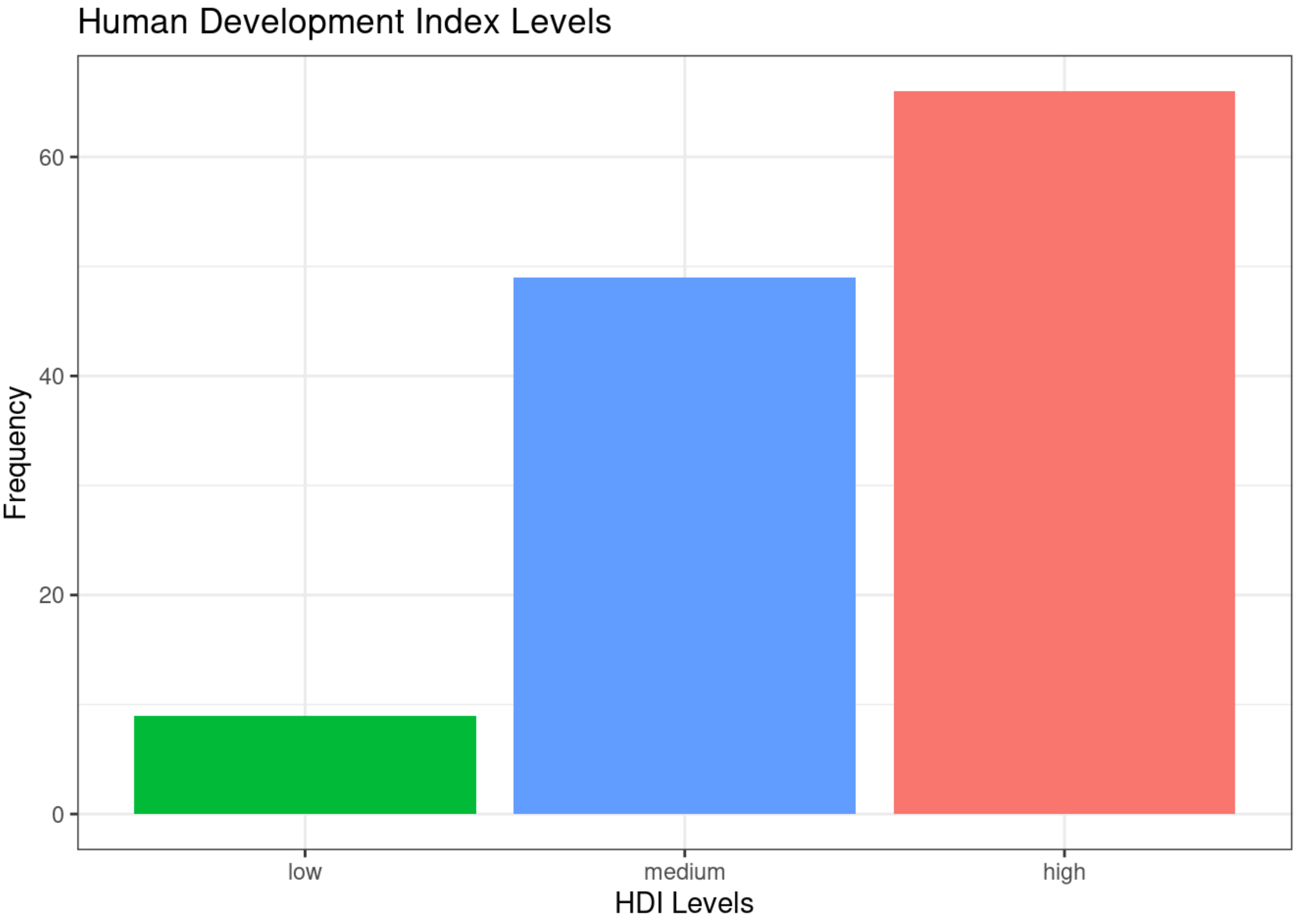
```
## [1] 0.87325
```

```
cor(complete_data$HDI_score, complete_data$happiness_score, use = "pairwise.complete.
obs")
```

```
## [1] 0.7742036
```

```
# Visualization 2 for Research Question 1
complete_data_m <- complete_data%>%
  dplyr::select("Country","HDI_score")%>%
  mutate(HDI_Level = case_when(HDI_score < 0.5 ~ 'low',
                            HDI_score < 0.75 ~ 'medium', HDI_score < 1 ~ 'high'))%>%
  arrange(desc(HDI_Level))

 complete_data_m %>%
   ggplot(aes(x= reorder(HDI_Level, HDI_score), fill=HDI_Level))+
  geom_bar()+
  theme_bw()+
  theme(legend.position= "none") +
  labs(title = "Human Development Index Levels", x = "HDI Levels", y = "Frequency")+
  scale_y_continuous(breaks = seq(0,100,20))
```



```
# Summary Statistics for the Visualization
table(complete_data_m$HDI_Level)
```

```
##
##   high    low medium
##     66      9     49
```

Examining Visualization 1: The scatterplot demonstrates that the highest happiness scores are associated with both higher HDI score and internet use rate. The ggplot demonstrated that there is a relatively positive relationship between internet use rate, happiness score, and HDI score. The correlation values between the three variables were computed. The correlation between internet use rate and happiness score was 0.7778816, the correlation between internet use rate and HDI score was 0.87325, and the correlation between HDI score and happiness score was 0.7742036.Therefore, all three variables are highly correlated with each other. Examining Visualization 2: The bar plot demonstrates that as the HDI levels increase from ‘low’ to ‘high’, the frequency of countries in those respective levels increases as well. In other words, the ‘low’ HDI level group denoted as a HDI_score < 0.5 has the lowest occurrences throughout the countries, the ‘medium’ HDI level group denoted as a HDI_score < 0.75 has a frequency larger than the ‘low’ group but smaller than the ‘high group’, and the ‘high’ HDI level group denoted as a HDI_score < 1 has the highest occurrences throughout the countries. A frequency table determined that there are 66 countries with a ‘high’ HDI level, 49 with a ‘medium’ level, and 9 with a ‘low’ level.

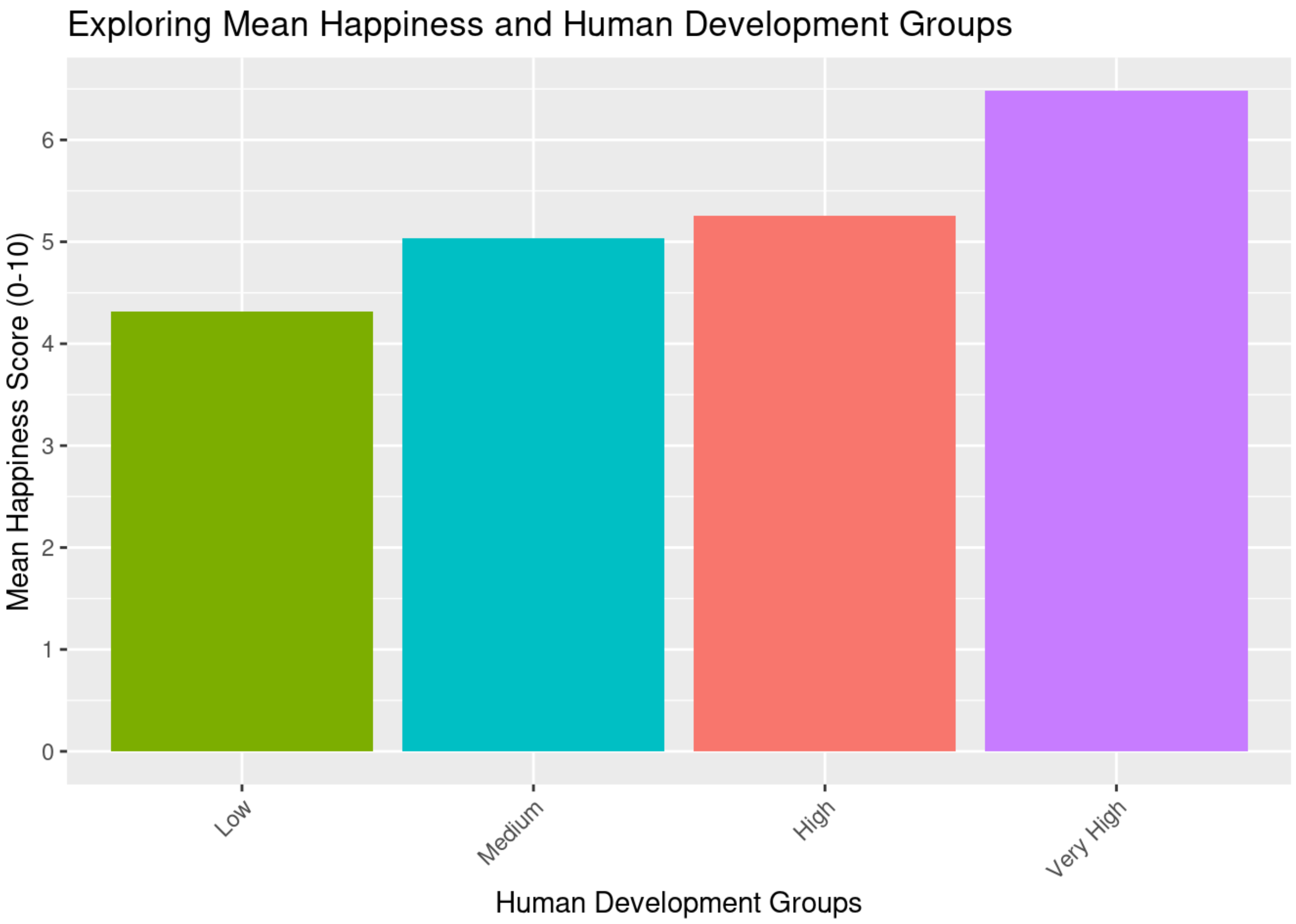
5.) Research Question 2: “What is the relationship between happiness score and the human development groups?”

```
# Exploring the Relationship between Human Development Groups and the Happiness Score
complete_data %>%
  group_by(`Human Development Groups`) %>%
  dplyr::select("Human Development Groups","happiness_score", "Country") %>%
  summarize(count = n())
```

```
## # A tibble: 4 × 2
##   `Human Development Groups` count
##   <chr>                      <int>
## 1 High                          29
## 2 Low                           20
## 3 Medium                        22
## 4 Very High                     53
```

```
# Visualization 1 for Research Question 2
complete_data_a <- complete_data%>%
  group_by(`Human Development Groups`)%>%
  dplyr::select("Country","Human Development Groups","happiness_score")%>%
  mutate(avg_happy= mean(happiness_score))
 complete_data_a %>%
  ggplot(aes(x= reorder(`Human Development Groups`, avg_happy), y= avg_happy, fill=`H
uman Development Groups`))+
    geom_histogram(stat = 'summary', fun = 'mean')+
  scale_y_continuous(breaks = seq(0,10,1)) +
  theme_grey()+
  theme(legend.position= "none")+
  theme(axis.text.x=element_text(angle=45,hjust=1))+
  labs(title = "Exploring Mean Happiness and Human Development Groups",
       x = "Human Development Groups", y = "Mean Happiness Score (0-10)")
```

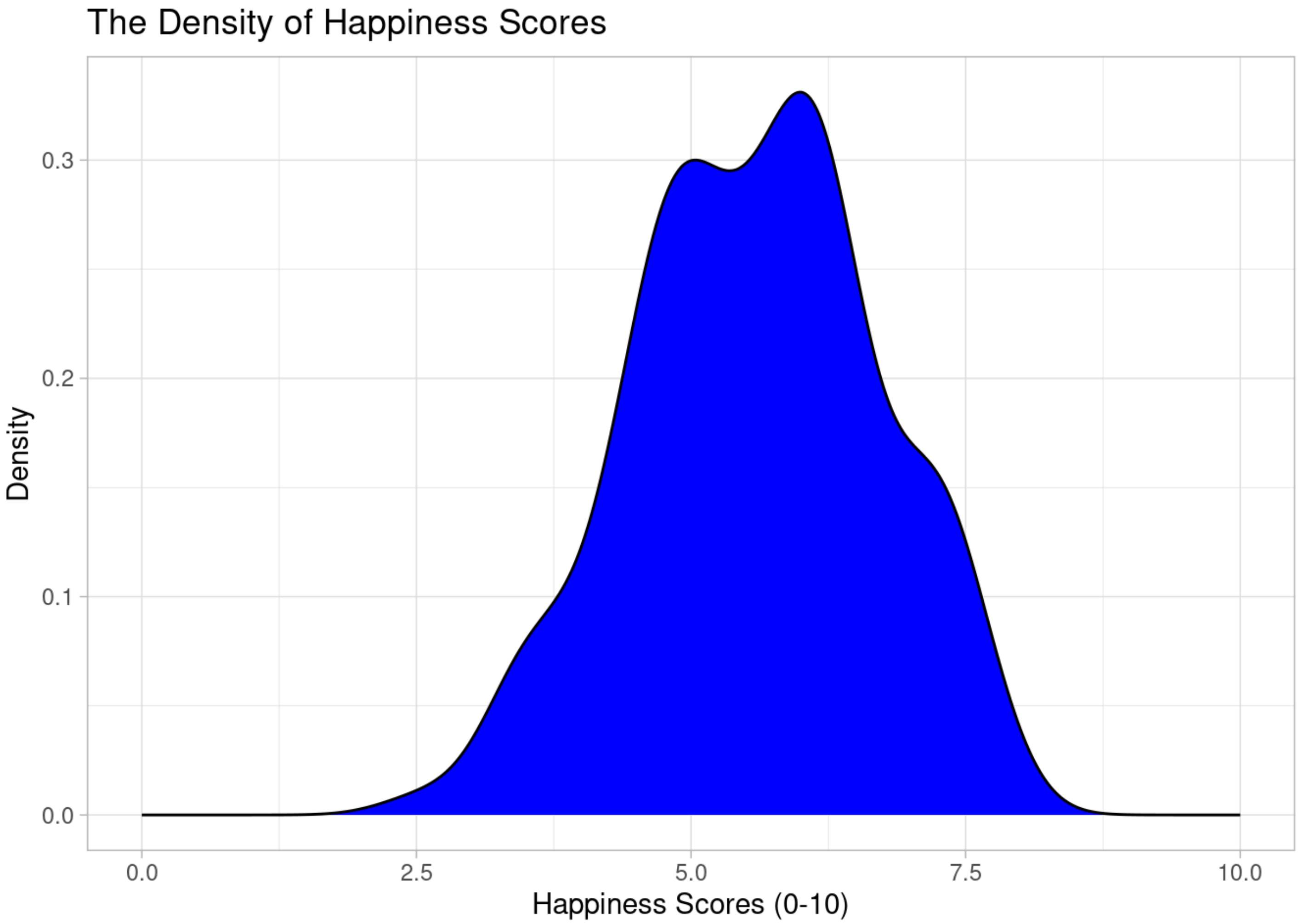
```
## Warning in geom_histogram(stat = "summary", fun = "mean"): Ignoring unknown
## parameters: `binwidth`, `bins`, and `pad`
```



```
# Summary Statistics for the Visualization
complete_data_a %>%
  group_by(`Human Development Groups`) %>%
  summarize(mean(avg_happy))
```

```
## # A tibble: 4 × 2
##   `Human Development Groups` `mean(avg_happy)`
##   <chr>                                  <dbl>
## 1 High                                    5.25
## 2 Low                                     4.32
## 3 Medium                                  5.03
## 4 Very High                               6.48
```

```
# Visualization 2 for Research Question 2
ggplot(complete_data, aes(x = happiness_score)) +
  geom_density(fill = "blue") +
  scale_x_continuous(limits = c(0,10)) +
  theme_light()+
  labs(title = "The Density of Happiness Scores", y = "Density", x = "Happiness Score
s (0-10)")
```



```
# Summary Statistics for the Visualization
summary(complete_data$happiness_score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   2.567   4.807   5.641   5.588   6.352   7.809 
```

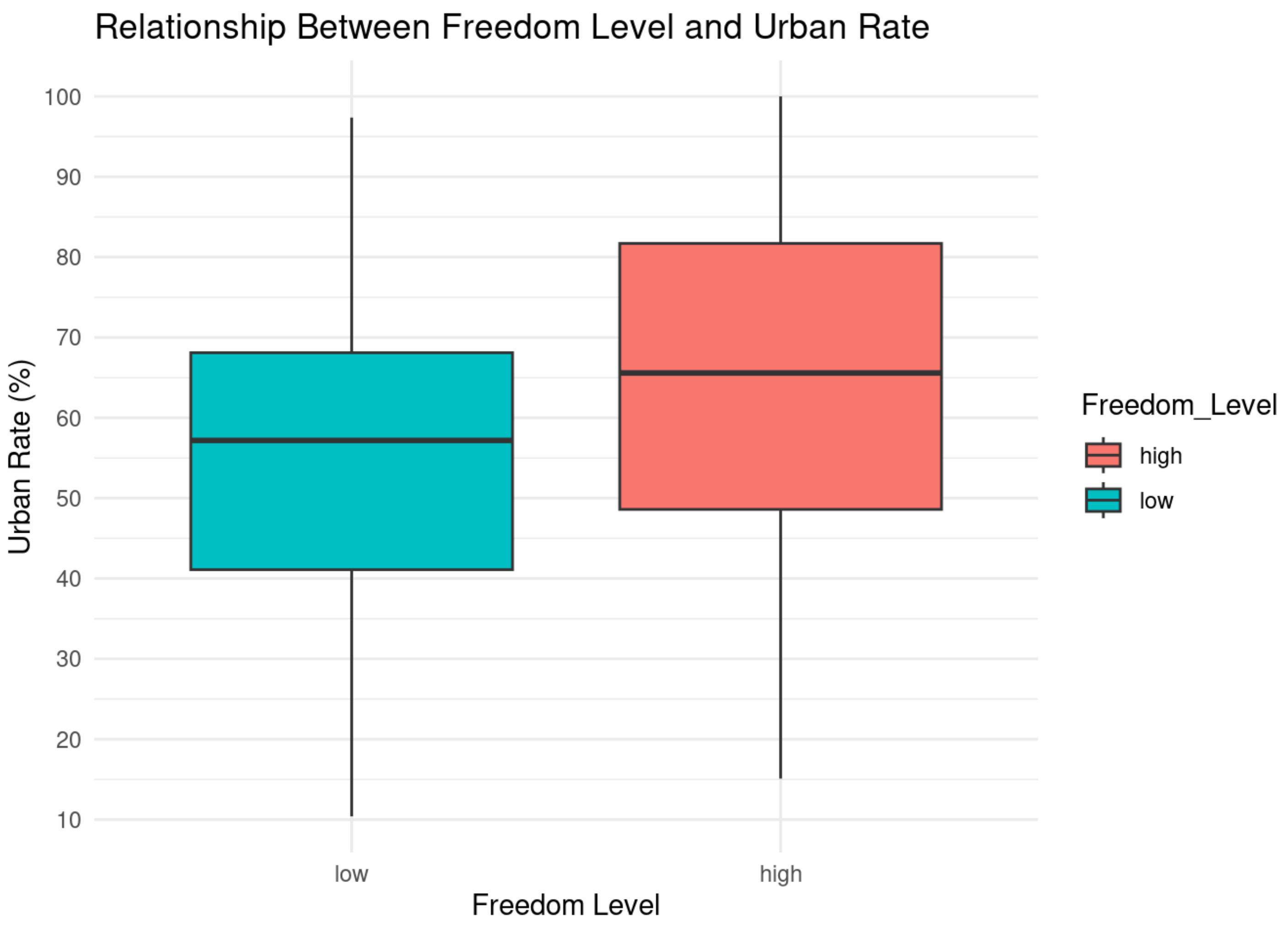
Examining Visualization 1: It was determined that there are 20 with ‘Low’ human development groups, 22 with ‘Medium’, 29 countries with ‘High’, and 53 with ‘Very High’. From the histogram, it can be concluded that human development groups are directly correlated with average happiness scores. As the ‘Mean Happiness Scores’ increased, the human development groups increased from ‘Low’ to ‘Very High’. The statistics computed resulted in there being a mean happiness score of 6.482860 for the ‘Very High’ human development group, 5.251848 for ‘High’, 5.033668 for ‘Medium’, and 4.315610 for ‘Low’ on a happiness score scale of 0-1. This shows that average happiness scores are higher for more developed countries. Examining Visualization 2: The second visualization is a density plot which demonstrates that the majority of the happiness scores are within the ranges of around 4-7. The density plot exhibits a normal distribution, so the mean was computed to be 5.588 on a scale of 0-10.

6.) Research Question 3 : “How does freedom correlate with the urban rate?”

```
# Exploring the Relationship Between Freedom and Urban Rate
complete_data%>%
  dplyr::select("Country","freedom","urbanrate")%>%
  mutate(Freedom_Level = case_when(freedom < 0.5 ~ 'low',
                         freedom >= 0.5 ~ 'high')) %>%
  arrange(desc(urbanrate))
```

```
## # A tibble: 124 × 4
##    Country        freedom urbanrate Freedom_Level
##    <chr>            <dbl>     <dbl> <chr>
##  1 Singapore        0.635     100   high
##  2 Kuwait           0.570      98.4 high
##  3 Belgium          0.500      97.4 low
##  4 Malta            0.633      94.3 high
##  5 Venezuela        0.272      93.3 low
##  6 Uruguay          0.594      92.3 high
##  7 Iceland          0.662      92.3 high
##  8 Argentina        0.521      92   high
##  9 Israel           0.421      91.7 low
## 10 United Kingdom   0.525      89.9 high
## # … with 114 more rows
```

```
# Visualization 1 for Research Question 3
complete_data_3 <- complete_data%>%
  dplyr::select("Country","freedom","urbanrate")%>%
  mutate(Freedom_Level = case_when(freedom < 0.5 ~ 'low',
                         freedom >= 0.5 ~ 'high'))%>%
  arrange(desc(urbanrate))
complete_data_3 %>%
  ggplot(aes(x= reorder(Freedom_Level, freedom), y=urbanrate, fill=Freedom_Level))+
  geom_boxplot()+
  scale_y_continuous(breaks = seq(0,100,10)) +
  theme(legend.position= "none") +
  theme_minimal()+
  labs(title = "Relationship Between Freedom Level and Urban Rate", x = "Freedom Leve
l", y = "Urban Rate (%)")
```



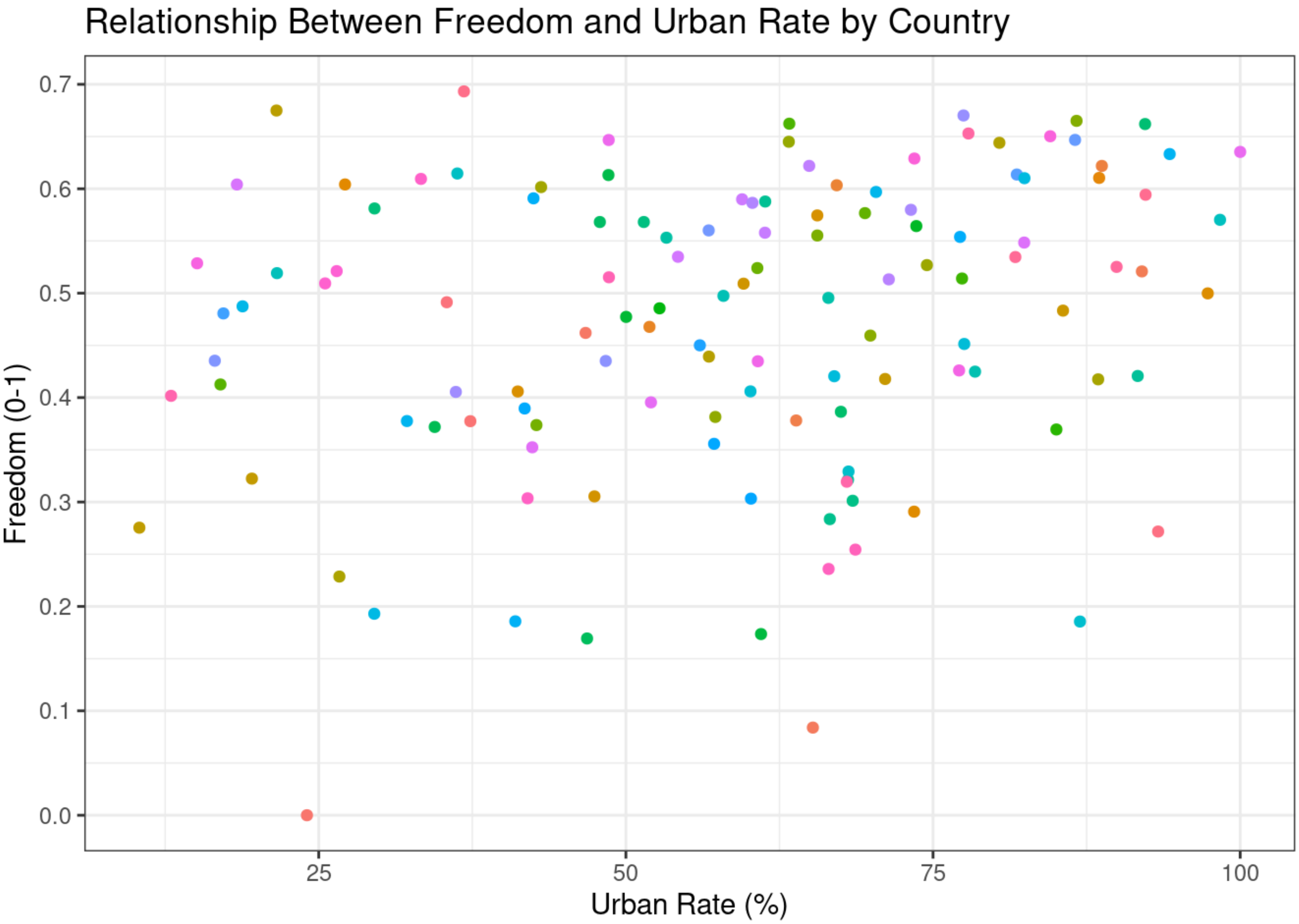
```
#Summary Statistics for the Visualization
complete_data_3 %>%
  filter(Freedom_Level == 'low') %>%
  summary()
```

```
##    Country             freedom         urbanrate     Freedom_Level
##  Length:63          Min.   :0.0000   Min.   :10.40   Length:63
##  Class :character   1st Qu.:0.3034   1st Qu.:41.10   Class :character
##  Mode  :character   Median :0.3864   Median :57.18   Mode  :character
##                     Mean   :0.3625   Mean   :54.33
##                     3rd Qu.:0.4352   3rd Qu.:68.10
##                     Max.   :0.4998   Max.   :97.36
```

```
complete_data_3 %>%
  filter(Freedom_Level == 'high') %>%
  summary()
```

```
##    Country             freedom          urbanrate       Freedom_Level     
##  Length:61          Min.   :0.5091   Min.   : 15.10   Length:61         
##  Class :character   1st Qu.:0.5531   1st Qu.: 48.60   Class :character  
##  Mode  :character   Median :0.5899   Median : 65.58   Mode  :character  
##                     Mean   :0.5884   Mean   : 63.25                     
##                     3rd Qu.:0.6219   3rd Qu.: 81.70                     
##                     Max.   :0.6933   Max.   :100.00                     
```

```
# Visualization 2 for Research Question 3
complete_data%>%
  dplyr::select("urbanrate","freedom", "Country") %>%
  ggplot(aes(x = urbanrate, y = freedom, color = Country)) +
  geom_point() +
  scale_y_continuous(breaks = seq(0,1,0.1)) +
  labs(title = "Relationship Between Freedom and Urban Rate by Country", x = "Urban R
ate (%)", y = "Freedom (0-1)") +
  theme_bw() +
  theme(legend.position = "none")
```



```
#Summary Statistics for the Visualization
cor(complete_data$freedom, complete_data$urbanrate, use = "pairwise.complete.obs")
```

```
## [1] 0.2153138
```

Examining Visualization 1: The box plot demonstrates that there is not a direct established relationship between urban rate and freedom level as there are countries with high urban rates but low freedom levels and countries with low urban rates and high freedom levels. Additionally, the box plot demonstrates that there is significant overlap of urban rates between the two freedom levels (‘low’ and ‘high’). Because the boxplots have normal distribution, mean was used as a summary statistic. The mean urban rate of ‘low’ freedom level was 54.33% while the mean urban rate of ‘high’ freedom level was 63.25%. The means are not exactly the same, however, the IQRs of both boxplots overlap, so the two different freedom levels do not have significantly different urban rates.

Examining Visualization 2: The scatterplot additionally shows no direct correlation between urban rate and freedom by country as the data points are scattered throughout the plot in a random distribution. The correlation between freedom and urban rate is 0.2153138, which further demonstrates that they are not highly correlated

7.) Discussion

There were three research questions that were investigated including “What is the relationship between internet usage, happiness score, and HDI score?”, “What is the relationship between happiness score and the human development groups?”, and “How does freedom correlate with the urban rate?” Through tidying, wrangling, and visualizing the intended variables of each research question, it was determined that happiness scores are associated with both higher HDI score and internet use rate. The summary statistics of correlation values between the three variables in Research Question 1 Visualization 1 were computed. The correlation between internet use rate and happiness score was 0.7778816, the correlation between internet use rate and HDI score was 0.87325, and the correlation between HDI score and happiness score was 0.7742036. This shows that all three variables had strong correlations with each other. For Research Question 1 Visualization 2, a frequency table was constructed and it computed that there are 66 countries with a ‘high’ HDI level, 49 with a ‘medium’ HDI level, and 9 with a ‘low’ HDI level. Additionally, this demonstrates that human development groups are directly correlated with happiness scores because it was determined that as the happiness scores increased, the human development groups increased from ‘Low’ to ‘Very High’. Through the histogram constructed to analyze Research Question 2 Visualization 1, it was also determined that human development groups are directly correlated with average happiness scores. The statistics computed resulted in there being a mean happiness score of 6.482860 for the ‘Very High’ human development group, 5.251848 for the ‘High’ group, 5.033668 for the ‘Medium’ group, and 4.315610 for the ‘Low’ group on a happiness score scale of 0-10. These statistics demonstrate that the mean happiness scores increase as the countries are more developed. For Research Question 2 Visualization 2, a density plot was also made representing the happiness scores of the countries. The density plot showed a normal

distribution, so the summary statistic of mean was computed to be 5.588. This shows the average happiness score is 5.588 on a scale of 1-10 throughout the countries. On the other hand, the visualizations for Research Question 3 both showcased that there was not a significant relationship between urban rate and freedom level. For Research Question 3 Visualization 1, a boxplot showing the potential relationship between freedom level and urban rate was made. Since both the box plots showed a normal distribution, the mean urban rate for 'low' and 'high' freedom levels were 54.33% and 63.25%, respectively. Additionally, the interquartile range (IQR) was 41.10-68.10% and 48.60-81.70% for the 'low' and 'high' freedom level, respectively. Therefore, since the interquartile ranges for the different freedom levels overlap, it means there is not a significant difference between the two distributions. For Research Question 3 Visualization 2, a scatterplot showing the relationship between freedom rate and urban rate was created. This visualization did not seem to show a significant relationship between the two variables since the points showed no pattern. However, we discovered the correlation between freedom rate and urban rate to be 0.2153138. Since the correlation value was closer to 0 than 1, it means that the freedom and urban rate are not highly correlated and confirms the conclusions from the visualization. All of our expectations matched the outcomes of the data. Finally, we should be careful about interpreting our data as concrete because, as stated in the introduction, there were 81 observations that were removed from the 'new_internet' dataset since it had 213 observations prior to joining. Because of these missing observations, we cannot be certain that there is no established relationship between urban rate and freedom as well as not being able to be 100% certain about any of the conclusions made from data included from the 'new-internet' dataset. The additional 81 data points could have influenced the outcomes and therefore must be considered to fully accept the conclusions. Something that could be done better for next time would be to include multiple years and find datasets that have the same number of observations so that more data points are included and so that no observations are left out. This would ensure that the results are 100% representative of the utilized datasets. When reflecting on conducting this project, it was challenging to find datasets with variables in common. Additionally, we learned how to use dplyr functions along with tidying functions in order to merge and filter the datasets to the variables of interest.

8.) Formatting

Acknowledgements: Thanks to Kaggle for providing the datasets. Additionally, Avery Zuckerman, Bella Crain, and Kaitlyn Rouse for collaborating together on all stages and questions of this project. Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.