Final Project

Anna MacFarlane, Edgar Salas, Jerry Fu, Taalin RaoShah

10/27/2020

Introduction

CodeStone chose to explore the data from the World Happiness Reports of 2018 and 2019, published by the Sustainable Development Network. According to the WHR, this data is derived from the Gallup World Poll, a systematic telephone survey and in person interviews in over 160 countries whose surveys claim to represent 80%+ of the population (2014). The calls are made via random phone number generation and randomly selecting households (GWP, 2014).

Generally, we want to explore the following question: which factors correlate most strongly to happiness across the globe? We chose to focus on four variables as well as the overall happiness score in our analysis. The four variables chosen were GDP per capita, Perceptions of Corruption, Social Support, and Life Expectancy, selected because we thought they represented economic, political, social, and medical aspects of well being respectively. Furthermore, the data set included 312 cases, each representing the data for a particular country. Guided by our general research question, we also examined the variation of their distribution by region, and which of these factors have the strongest correlations.

Considering the concerning state of our world in 2020, including the worsening effects of climate change, threats to democracy, and much more, we found it topical and insightful to evaluate what contributes to happiness within each nation and across the globe. The first World Happiness Report, published in 2012, presents the report as a means of grappling with the countless contradictions that exist in modern society such as the balance between pursuing economic success versus protecting the environment or the tradeoffs between personal profit and community trust (Helliwell et al., 2012). Eight years later, these paradoxes persit, and the potential solutions are closely linked to definitions of morality, heightening their controversy. Considering the continued debate over such questions, we believe there are grounds for further investigation into trends of happiness over time and the factors that contribute to it.

```r
library(tidyverse)
library(plyr) #package for join command
library(maps) #package for world map
library(broom)
```

```r
report_2018 <- read_csv("data/2018.csv") %>%
  mutate(year = "2018",
         `Perceptions of corruption` = as.numeric(`Perceptions of corruption`))
report_2019 <- read_csv("data/2019.csv") %>%
  mutate(year = "2019")
country_region <- read_csv("data/2020.csv") %>%
  mutate(year = "2020") %>%
  select(`Country name`, `Regional indicator`)

names(country_region) <- c("Country", "Region")
names(report_2018) <- str_replace_all(names(report_2018), c(" " = "_"))
names(report_2019) <- str_replace_all(names(report_2018), c(" " = "_"))
```

```
worldhappiness <- full_join(report_2018, report_2019)
worldhappiness <- worldhappiness %>%
  rename(c("Country_or_region" = "Country")) %>%
  left_join(country_region)

glimpse(worldhappiness)
```

```
## Rows: 312
## Columns: 11
## $ Overall_rank                 <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12,...
## $ Country                      <chr> "Finland", "Norway", "Denmark", "Icela...
## $ Score                        <dbl> 7.632, 7.594, 7.555, 7.495, 7.487, 7.4...
## $ GDP_per_capita               <dbl> 1.305, 1.456, 1.351, 1.343, 1.420, 1.3...
## $ Social_support               <dbl> 1.592, 1.582, 1.590, 1.644, 1.549, 1.4...
## $ Healthy_life_expectancy      <dbl> 0.874, 0.861, 0.868, 0.914, 0.927, 0.8...
## $ Freedom_to_make_life_choices <dbl> 0.681, 0.686, 0.683, 0.677, 0.660, 0.6...
## $ Generosity                   <dbl> 0.202, 0.286, 0.284, 0.353, 0.256, 0.3...
## $ Perceptions_of_corruption    <dbl> 0.393, 0.340, 0.408, 0.138, 0.357, 0.2...
## $ year                         <chr> "2018", "2018", "2018", "2018", "2018"...
## $ Region                       <chr> "Western Europe", "Western Europe", "W...
```

```
worldhappiness %>%
  filter(is.na(Region))
```

```
## # A tibble: 20 x 11
##    Overall_rank Country Score GDP_per_capita Social_support Healthy_life_ex~
##           <dbl> <chr>   <dbl>          <dbl>          <dbl>            <dbl>
##  1           26 Taiwan   6.44          1.36           1.44             0.857
##  2           32 Qatar    6.37          1.65           1.30             0.748
##  3           38 Trinid~  6.19          1.22           1.49             0.564
##  4           49 Belize   5.96          0.807          1.10             0.474
##  5           58 Northe~  5.84          1.23           1.21             0.909
##  6           76 Hong K~  5.43          1.40           1.29             1.03
##  7           97 Bhutan   5.08          0.796          1.34             0.527
##  8           98 Somalia  4.98          0              0.712            0.115
##  9          137 Sudan    4.14          0.605          1.24             0.312
## 10          142 Angola   3.80          0.73           1.12             0.269
## 11          150 Syria    3.46          0.689          0.382            0.539
## 12           25 Taiwan   6.45          1.37           1.43             0.914
## 13           29 Qatar    6.37          1.68           1.31             0.871
## 14           39 Trinid~  6.19          1.23           1.48             0.713
## 15           64 Northe~  5.72          1.26           1.25             1.04
## 16           76 Hong K~  5.43          1.44           1.28             1.12
## 17           84 North ~  5.27          0.983          1.29             0.838
## 18           95 Bhutan   5.08          0.813          1.32             0.604
## 19          112 Somalia  4.67          0              0.698            0.268
## 20          149 Syria    3.46          0.619          0.378            0.44
## # ... with 5 more variables: Freedom_to_make_life_choices <dbl>,
## #   Generosity <dbl>, Perceptions_of_corruption <dbl>, year <chr>, Region <chr>
```

```
worldhappiness <- worldhappiness %>%
  mutate(Region = ifelse(Country == "Taiwan", "East Asia", Region)) %>%
  mutate(Region = ifelse(Country == "Qatar", "Middle East and North Africa", Region)) %>%
  mutate(Region = ifelse(Country == "Trinidad & Tobago", "Latin America and Caribbean", Region)) %>%
  mutate(Region = ifelse(Country == "Belize", "Latin America and Caribbean", Region)) %>%
  mutate(Region = ifelse(Country == "Northern Cyprus", "Middle East and North Africa", Region)) %>%
  mutate(Region = ifelse(Country == "Hong Kong", "East Asia", Region)) %>%
  mutate(Region = ifelse(Country == "Bhutan", "South Asia", Region)) %>%
  mutate(Region = ifelse(Country == "Somalia", "Middle East and North Africa", Region)) %>%
  mutate(Region = ifelse(Country == "Sudan", "Middle East and North Africa", Region)) %>%
  mutate(Region = ifelse(Country == "Angola", "Middle East and North Africa", Region)) %>%
  mutate(Region = ifelse(Country == "Syria", "Middle East and North Africa", Region)) %>%
  mutate(Region = ifelse(Country == "North Macedonia", "Central and Eastern Europe", Region))
```

Methodology

First, we did a general overview of the distribution of Happiness Scores across different countries and regions. In the map colored by Happiness Score, we noticed that countries in different regions of the world tended to have different happiness scores. In order to see if what we were visualizing was statistically significant, a chi-squared test was used to test if there was an association between Happiness Score and Region. A chi-squared test statistic is computed by adding up the squares of the difference between the observed and expected values in a table, divided by the expected values. The expected values are equal to the row total multiplied by the column total, and divided by the overall total. Finally, p-values are derived from degrees of freedom and the chi-square test statistic. A very small chi-square test statistic means that your observed data fits your expected data well and that a relationship exists, while a large chi-square test statistic means the opposite. In our case, our null hypothesis was that there isn't an association between region and happiness, and our alternative hypothesis was that there was an association between region and happiness. With a p-value of 0.1253, we did not have sufficient evidence to reject the null hypothesis at an alpha of 0.05. However, the Pearson chi-square test might not be appropriate for our dataset, which we will discuss more in our discussion section. Disregarding this for now, we moved on to looking at the individual metrics which went into determining happiness score. In trying to answer the extent to which factors are most important in affecting Happiness, we looked at GDP per capita, Perceptions of Corruption, Social Support, and Life Expectancy.

Variable Sources and Definitions

```
Happiness Score: Happiness score is a self-reported measure of overall current life satisfaction. This w

GDP per capita: PPP (purchasing power parity) is a rate of conversion which attempts to equalize the pu

Social Support: Social support is a self-reported measure of whether or not the respondent feels they ca

Life Expectancy: Life expectancy data was extrapolated from the WHOs health observation data up to 2016

Perceptions of Corruption: Perception of Corruption is a self-reported measure of whether or not respond
```

Helliwell, J., Layard, R., & Sachs, J. (2018). World Happiness Report 2018, New York: Sustainable Development Solutions Network.

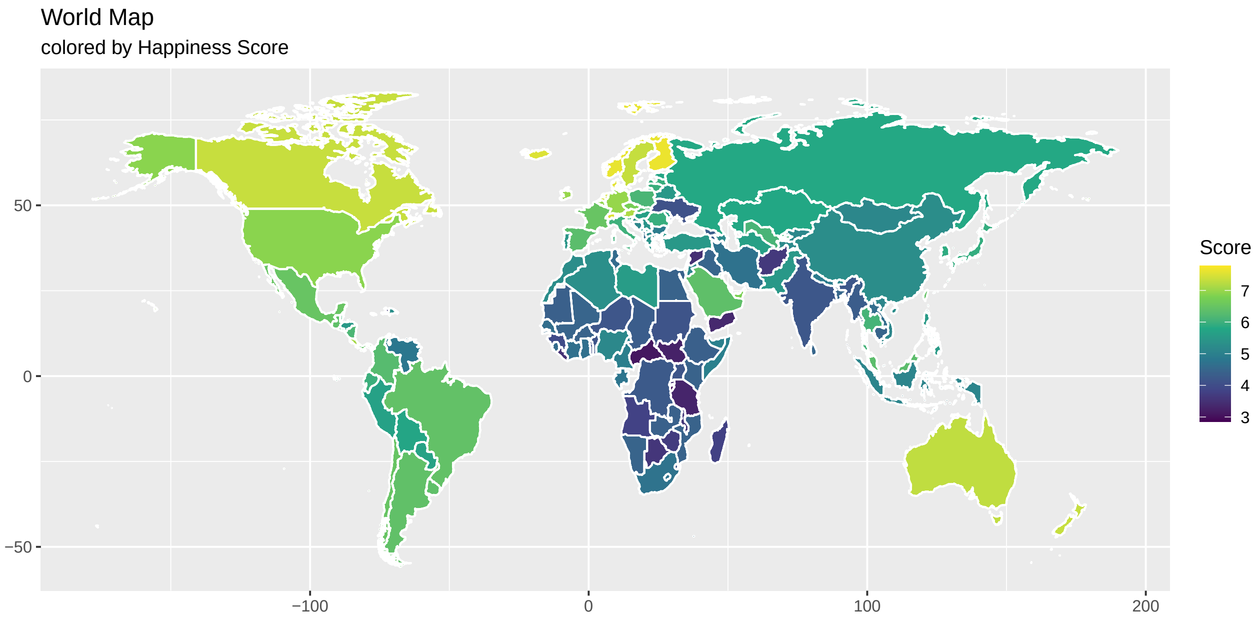
Helliwell, J., Layard, R., & Sachs, J. (2019). World Happiness Report 2019, New York: Sustainable Development Solutions Network.

Results

Is there a relationship between region of the world and happiness score?

```
world_map <- map_data("world") %>%
  mutate(region = ifelse(region == "USA", "United States", region)) %>%
  mutate(region = ifelse(region == "Democratic Republic of the Congo", "Congo (Kinshasa)", region))

happiness_score_map <- left_join(worldhappiness, world_map, by = c("Country" = "region"))
```

```
ggplot(happiness_score_map, aes(x = long, y = lat, group = group)) +
  geom_polygon(aes(fill = Score), color = "white") +
  labs(title = "World Map", subtitle = "colored by Happiness Score", x = "", y = "") +
  scale_fill_viridis_c(option = "D")
```



https://www.datanovia.com/en/blog/how-to-create-a-map-using-ggplot2/

```
chisq.test(worldhappiness$Score, worldhappiness$Region)
```

```
## Warning in chisq.test(worldhappiness$Score, worldhappiness$Region): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  worldhappiness$Score and worldhappiness$Region
## X-squared = 2729.8, df = 2646, p-value = 0.1253
```

Multiple regression for all considered factors

```
m_happy <- lm(Score ~ Perceptions_of_corruption + Social_support +
                GDP_per_capita + Healthy_life_expectancy, data = worldhappiness)

m_happy %>%
  tidy() %>%
  select(term, estimate)
```

```
## # A tibble: 5 x 2
##   term                      estimate
##   <chr>                        <dbl>
## 1 (Intercept)                  2.10
## 2 Perceptions_of_corruption    1.95
## 3 Social_support               1.41
## 4 GDP_per_capita               0.827
## 5 Healthy_life_expectancy      0.941
```

```
glance(m_happy)$r.squared
```

```
## [1] 0.7526699
```

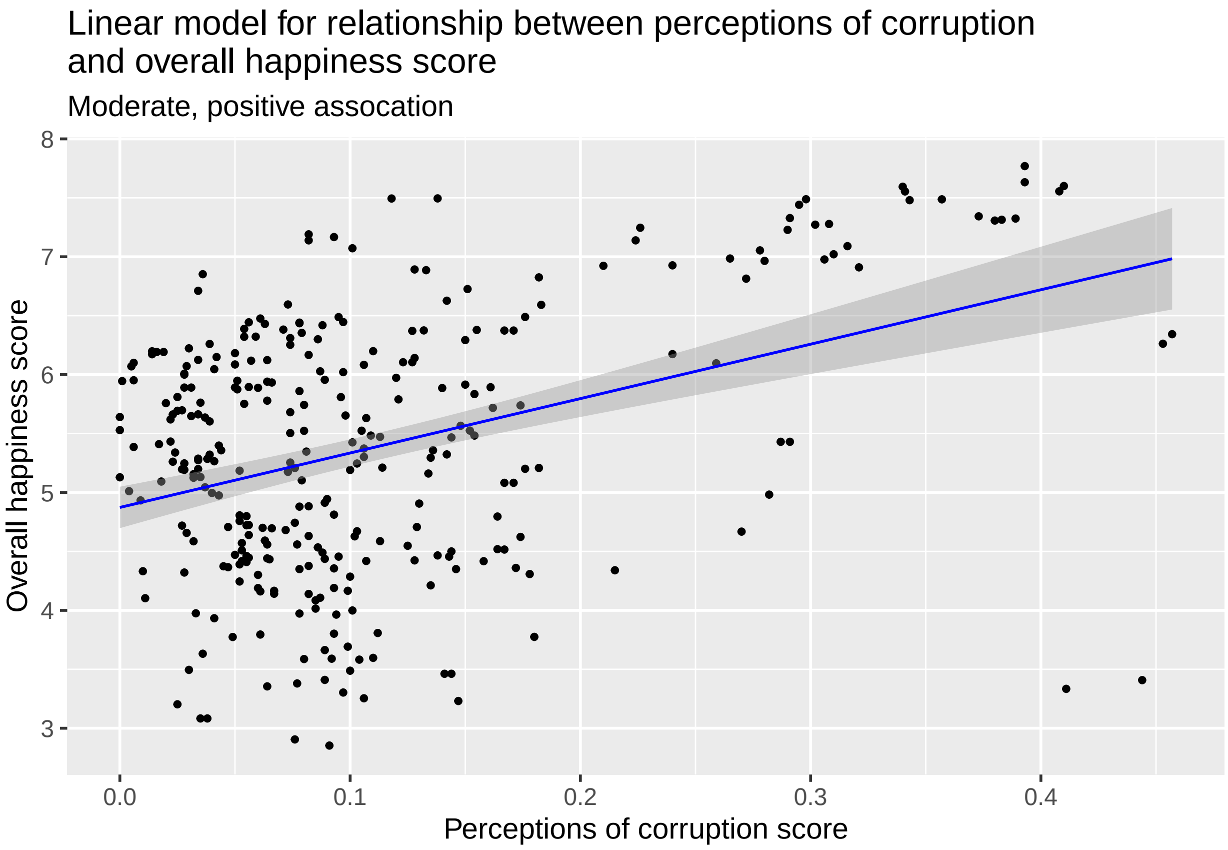
Perceptions of Corruption

«««< HEAD

```
# Linear model for relationship between perceptions of corruption score and overall happiness score
m_corrupt <- lm(Score ~ Perceptions_of_corruption, data = worldhappiness)

m_corrupt %>%
  tidy() %>%
  select(term, estimate)
```

```
## # A tibble: 2 x 2
##   term                      estimate
##   <chr>                        <dbl>
## 1 (Intercept)                   4.87
## 2 Perceptions_of_corruption     4.62
```

```
# Visualization of linear model

#ggplot(data = worldhappiness, aes(x = Perceptions_of_corruption,

ggplot(data = m_corrupt, aes(x = Perceptions_of_corruption,
                              y = Score)) +
  geom_point(size = 0.8) +
  geom_smooth(method = lm, color = "blue", size = 0.5) +
  labs(title = "Linear model for relationship between perceptions of corruption
and overall happiness score",
       subtitle = "Moderate, positive assocation",
       x = "Perceptions of corruption score",
       y = "Overall happiness score")
```



```
#Confidence interval for linear model above
confint_tidy(m_corrupt, conf.level = 0.95) %>%
  slice(2)
```

```
## # A tibble: 1 x 2
##   conf.low conf.high
##      <dbl>     <dbl>
## 1     3.42      5.82
```

«««< HEAD

======= ### Life Expectancy

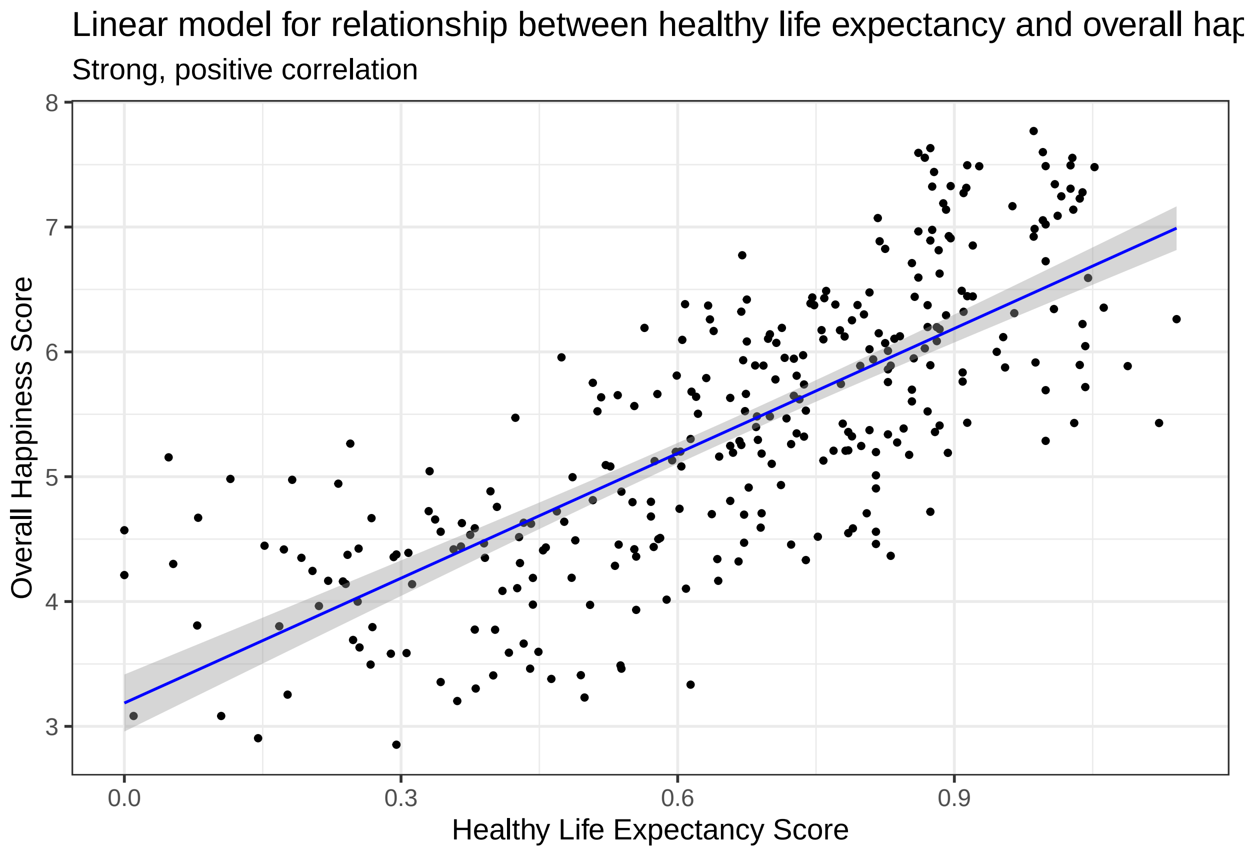
```
life_expectancy <- worldhappiness %>%
  filter(!is.na(Healthy_life_expectancy)) %>%
  select(Country, Score, Healthy_life_expectancy, year)

life_expectancy %>%
  summarize(r_overall = cor(Score, Healthy_life_expectancy))
```

```
##   r_overall
## 1 0.7558749
```

```
m_life <- lm(Score ~ Healthy_life_expectancy, data = worldhappiness)

m_life %>%
  tidy() %>%
  select(term, estimate)
```

```
## # A tibble: 2 x 2
##   term                      estimate
##   <chr>                        <dbl>
## 1 (Intercept)                   3.19
## 2 Healthy_life_expectancy       3.33
```

```
ggplot(data = life_expectancy, aes(x = Healthy_life_expectancy, y = Score)) +
  geom_point(size = 0.8) +
  geom_smooth(method = lm, color = "blue", size = 0.5) +
  theme_bw() +
  labs(title = "Linear model for relationship between healthy life expectancy and overall happiness scor
       subtitle = "Strong, positive correlation", x = "Healthy Life Expectancy Score",
        y = "Overall Happiness Score")
```



Social Support

```
#Sample stat.
socsup <- worldhappiness %>%
  filter(!is.na(Social_support)) %>%
  select(Country, Score, Social_support, year)

socsup %>%
  summarize(r_overall = cor(Score, Social_support))
```

```
##   r_overall
## 1 0.7610805
```

```
#Create tidy for linear model
m_score_socsup <- lm(Score ~ Social_support, data = worldhappiness)
```

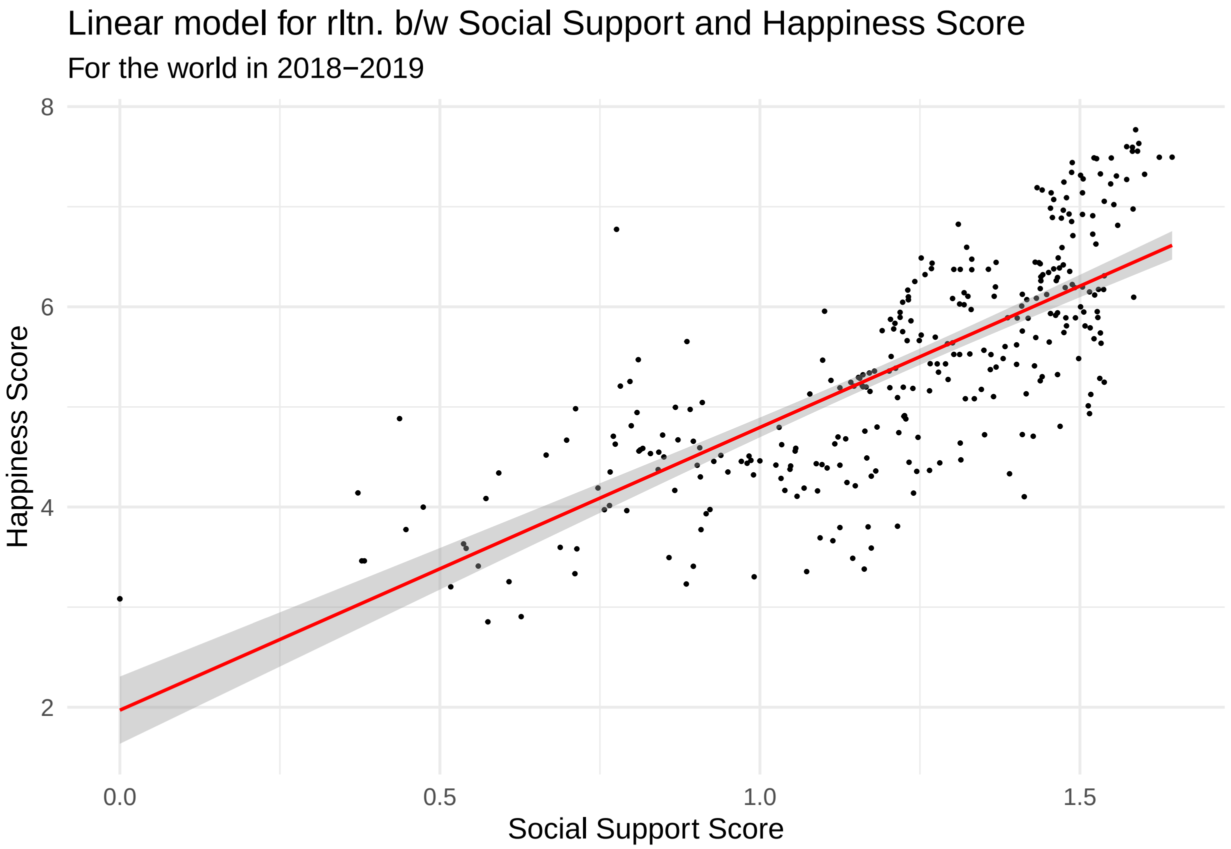
```r
m_score_socsup %>%
  tidy() %>%
  select(term, estimate)
```

```
## # A tibble: 2 x 2
##   term           estimate
##   <chr>             <dbl>
## 1 (Intercept)        1.97
## 2 Social_support     2.82
```

```r
#Conf. Interval
broom::confint_tidy(m_score_socsup, conf.level = 0.95)
```

```
## # A tibble: 2 x 2
##   conf.low conf.high
##      <dbl>     <dbl>
## 1     1.63      2.31
## 2     2.56      3.09
```

```r
#Model
ggplot(data = socsup, mapping = aes(x = Social_support, y = Score)) +
  geom_point(size = 0.3) +
  geom_smooth(method = "lm", color = "red", size = 0.6) +
  labs(title = "Linear model for rltn. b/w Social Support and Happiness Score",
       subtitle = "For the world in 2018-2019", x = "Social Support Score",
       y = "Happiness Score") +
  theme_minimal()
```



Discussion

Based on data from the World Happiness Report, we aimed to explore how some factors of interest affected overall happiness in a country. We found that GDP per capita, Social Support, and Life Expectancy were all strongly correlated with the overall happiness score, and Perceptions of Corruption (reverse-scored) was moderately correlated. Furthermore, we wanted to know whether happiness was independent of country. Essentially, does the average individual's happiness depend on the country that they live in? We found that happiness varies significantly from country to country. On top of that, some regions have high concentrations of High-, mid-, and low-happiness, suggesting that region also plays a role. For example, while North America, Nordic countries, and Oceania have very high levels of overall happiness, regions like Central Africa and South Asia have lower levels of overall happiness.

While thought was put into the variables selected for investigation, the scope of our analysis was limited by only looking at GDP per capita, Perceptions of Corruption, Social Support, and Life Expectancy with respect to the overall Score. Additionally, evaluating each category independently when considering the linear models does not account for the likelihood that one variable may not be entirely independent of the other. While the conditions appeared to be met for the use of the linear model, some of the spreads did not appear to be completely normal which may impact the validity of the conclusions drawn.

Considering that the data was collected through Gallup, a group well-respected for proper data collection, it can be evaluated as reliable. However, the translation of particular interviews and surveys into quantitative numbers leaves room for subjectivity to affect the data. While a variable such as GDP per capita is more concrete, something such as Perceptions of Corruption may be dependent on other factors such as access to education or level of freedom of the media. Consequently, the results of data analysis should be interpreted with this in mind.

We were unable to reject the null hypothesis for our chi-square test for happiness and region, but there is a very likely possibility that we made a type II error. In order to perform a Pearson's chi-squared test, four conditions have to be satisfied: simple random sample, sample size, expected cell count, and independence. Based on the description of the dataset, the survey was random, enough people were sampled, and the people sampled were independent of each other. Due to the nature of our data, however, expected counts for several cells were smaller than 5, and so the test loses a lot of accuracy. A solution to this would be to add the field simulate.p.value = TRUE to the chi-squared test, which uses the Monte Carlo method to obtain a p-value from random sampling.

Going forward, we would hope to study what specific factors lead to happiness among a general population within a country. In order to do this, we would need more variables that affect people's lives: e.g. level of infrastructure, quality of healthcare, quality of education, safety. We might also seek to answer questions like the following: do the factors that predict happiness vary by region? Why are some regions happier than others? What factor is the most important for happiness across the board?