Project 1

Balancing Urban Life : Varsha Narayanan, Catherine Nader, Julia Joseph

1. Introduction:

We chose three datasets from Kaggle. One dataset (https://www.kaggle.com/datasets/mvieira101/global-cost-of-living) is information about the cost of living in almost 5000 cities across the world. Each row represents a city and each column represents a variable that contributes to the city's cost of living (gasoline, 1 bedroom apartment, electricity, heating, cooling, water, garbage, average monthly net salary, mortgage interest rate in percentages, etc.). All variables are numeric, except for city and country, which are categorical.

Our second dataset (https://www.kaggle.com/datasets/orhankaramancode/city-quality-of-life-dataset) contains information on housing, cost of living, startups, safety, healthcare, education, economy, leisure and culture. These variables are related to a city's quality of life. The values are based on a rating system. Each row represents a city. All variables are numeric, except for city and country, which are categorical variables.

Our third dataset (https://www.kaggle.com/datasets/prasertk/cities-with-the-best-worklife-balance-2022)) contains information on cities around the world with the best work-life balance in order from highest to lowest based on variables, such as overworked population, unemployment, access to mental health resources, affordability, inflation, etc.). Each row represents a city. All variables are numeric, except for city and country, which are categorical variables.

These datasets are interesting to us because as we progress in our professional careers, it would be helpful to know which areas would best support a healthy lifestyle.

We would join all three datasets by city.

One trend that we expect to see is that cities high education score will have a higher cost of living because we expect education to be better quality in more expensive areas due to more funding. Another trend we expect to see is with more remote jobs, there is a greater environmental quality in a city because of the lack of commuting and using less fossil fuels. Finally, we expect to see that with a lower unemployment score (high unemployment rate), there will be more access to mental health resources, as losing a job can negatively affect an individual's mental health.

Research Questions:

1. What is the relationship between a city's education score and cost of a 3-bedroom apartment?
2. Does the number of remote jobs affect the environmental quality in each city?
3. Does unemployment have an effect on the access to mental healthcare resources in US cities?

Let's first load the `tidyverse` package that contains `tidyr`, `dplyr` and `ggplot2`, and `tidytext`, and `textdata`, which we will use in this report.

```
# Load packages
library(tidyverse)
library(tidytext)
library(textdata)
```

Let's take a look at the three datasets:

```
# Loading in datasets from csv files from Kaggle
costoflivingdata <- read_csv("cost-of-living.csv")
head(costoflivingdata)
```

```
## # A tibble: 6 x 59
##    ...1 city     country      x1    x2    x3    x4    x5    x6    x7    x8    x9
##   <dbl> <chr>    <chr>     <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1     0 Delhi    India      4.9   22.0  4.28  1.84  3.67  1.78  0.48  0.19  0.73
## 2     1 Shanghai China      5.59  40.5  5.59  1.12  4.19  3.96  0.52  0.32  2.68
## 3     2 Jakarta  Indonesia  2.54  22.2  3.5   2.02  3.18  2.19  0.59  0.27  1.28
## 4     3 Manila   Philippi~  3.54  27.4  3.54  1.24  1.9   2.91  0.93  0.51  1.63
## 5     4 Seoul    South Ko~  7.16  52.8  6.03  3.02  4.52  3.86  1.46  0.78  2.13
## 6     5 Bangkok  Thailand   2.6   28.1  5.62  2.25  4.21  2.06  0.5   0.26  1.61
## # ... with 47 more variables: x10 <dbl>, x11 <dbl>, x12 <dbl>, x13 <dbl>,
## #   x14 <dbl>, x15 <dbl>, x16 <dbl>, x17 <dbl>, x18 <dbl>, x19 <dbl>,
## #   x20 <dbl>, x21 <dbl>, x22 <dbl>, x23 <dbl>, x24 <dbl>, x25 <dbl>,
## #   x26 <dbl>, x27 <dbl>, x28 <dbl>, x29 <dbl>, x30 <dbl>, x31 <dbl>,
## #   x32 <dbl>, x33 <dbl>, x34 <dbl>, x35 <dbl>, x36 <dbl>, x37 <dbl>,
## #   x38 <dbl>, x39 <dbl>, x40 <dbl>, x41 <dbl>, x42 <dbl>, x43 <dbl>,
## #   x44 <dbl>, x45 <dbl>, x46 <dbl>, x47 <dbl>, x48 <dbl>, x49 <dbl>, ...
```

```
worklifebaldata <- read_csv("Cities with the Best Work-Life Balance 2022.csv")
head(worklifebaldata)
```

```
## # A tibble: 6 x 24
##   `2022` `2021` City     Country Remot~1 Overw~2 Minim~3 Vacat~4 Unemp~5 Multi~6
##    <dbl> <chr>  <chr>    <chr>   <chr>   <chr>     <dbl> <chr>     <dbl> <chr>
## 1      1 2      Oslo     Norway  41.72%  11.20%       25 25         94.7 9.10%
## 2      2 -      Bern     Switze~ 44.86%  11.40%       20 25         99.8 7.60%
## 3      3 1      Helsinki Finland 38.92%  12.70%       25 30         89.3 6.30%
## 4      4 3      Zurich   Switze~ 44.86%  11.90%       20 25         99.2 7.60%
## 5      5 5      Copenha~ Denmark 41.42%  10.50%       25 28         94.8 7.60%
## 6      6 -      Geneva   Switze~ 44.86%  11.90%       20 25         95.2 7.60%
## # ... with 14 more variables: Inflation <chr>,
## #   `Paid Parental Leave (Days)` <dbl>, `Covid Impact` <dbl>,
## #   `Covid Support` <dbl>, Healthcare <dbl>,
## #   `Access to Mental Healthcare` <dbl>, `Inclusivity & Tolerance` <dbl>,
## #   Affordability <dbl>, `Happiness, Culture & Leisure` <dbl>,
## #   `City Safety` <dbl>, `Outdoor Spaces` <dbl>, `Air Quality` <dbl>,
## #   `Wellness and Fitness` <dbl>, `TOTAL SCORE` <dbl>, and abbreviated ...
```

```
cityqualitydata <- read_csv("cityquality.csv")
head(cityqualitydata)
```

```
## # A tibble: 6 x 21
##    ...1 UA_Name  UA_Co~1 UA_Co~2 Housing Cost ~3 Start~4 Ventu~5 Trave~6 Commute
##   <dbl> <chr>    <chr>   <chr>     <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1     0 Aarhus   Denmark Europe     6.13    4.02    2.83    2.51    3.54    6.31
## 2     1 Adelaide Austra~ Oceania    6.31    4.69    3.14    2.64    1.78    5.34
## 3     2 Albuque~ New Me~ North ~    7.26    6.06    3.77    1.49    1.46    5.06
## 4     3 Almaty   Kazakh~ Asia       9.28    9.33    2.46    0       4.59    5.87
## 5     4 Amsterd~ Nether~ Europe     3.05    3.82    7.97    6.11    8.32    6.12
## 6     5 Anchora~ Alaska  North ~    5.43    3.14    2.79    0       1.74    4.72
## # ... with 11 more variables: `Business Freedom` <dbl>, Safety <dbl>,
## #   Healthcare <dbl>, Education <dbl>, `Environmental Quality` <dbl>,
## #   Economy <dbl>, Taxation <dbl>, `Internet Access` <dbl>,
## #   `Leisure & Culture` <dbl>, Tolerance <dbl>, Outdoors <dbl>, and abbreviated
## #   variable names 1: UA_Country, 2: UA_Continent, 3: `Cost of Living`,
## #   4: Startups, 5: `Venture Capital`, 6: `Travel Connectivity`
```

2. Tidying

Our datasets were tidy. Each variable has its own column, each observation (city) has its own row, and each value has its own cell. We selected column variables we thought were the most important and interesting. We renamed the column names to be easy-to-understand and uniform with all other column names.

```
# Global Cost of Living
costofliving <- costoflivingdata %>%
  # select city, utility costs, cost of 1 bed apartment, cost of 3 bed apartment, average month net sal
  select(city,x36, x48, x50, x54, x55, x33) %>%
  # rename columns to meaningful names
  rename(gasoline=x33, utility_cost = x36, apt_1_bed = x48, apt_3_bed = x50, avg_month_net_sal = x54, a
costofliving
```

```
## # A tibble: 4,874 x 7
##    city      utility_cost apt_1_bed apt_3_bed avg_month_net_sal avg_mo~1 gasol~2
##    <chr>            <dbl>     <dbl>     <dbl>             <dbl>    <dbl>   <dbl>
##  1 Delhi             57.4      224.      596.              586.     7.96    1.25
##  2 Shanghai          64.8     1080.     2973.             1383.     5.01    1.17
##  3 Jakarta           80.1      483.     1118.              483.     9.15    0.79
##  4 Manila            97.4      560.     1754.              419.     7.8     1.38
##  5 Seoul            176.       810.     2621.             2672.     3.47    1.41
##  6 Bangkok           77.6      585.     1995.              615.     5.64    1.18
##  7 Kolkata           31.4      149.      398.              536.     8.28    1.33
##  8 Guangzhou         56.2      523.     1219.             1189.     5.11    1.17
##  9 Mumbai            44        514.     1389.              628.     7.87    1.34
## 10 Beijing           53.7     1177.     2732.             1515.     5       1.18
## # ... with 4,864 more rows, and abbreviated variable names 1: avg_mortgage_int,
## #   2: gasoline
```

```
states <- c('Alabama','Alaska','Arizona','Arkansas','California','Colorado','Connecticut','Delaware','F

# in cityquality, country_name is the name of state if city in US, so changed country values for US cit
cityqualitydata$country_name <- ifelse(cityqualitydata$UA_Country %in% states, "USA", cityqualitydata$U

# Quality of Life in Cities
cityquality <- cityqualitydata %>%
  # rename columns to meaningful names, to be uniform and match other datasets' variables format
  rename(city = UA_Name,housing_score = Housing, commute_score = Commute, education_score = Education, 
  # select column variables
  select(city, housing_score, commute_score, education_score, env_quality_score, economy_score, taxatio
cityquality
```

```
## # A tibble: 266 x 7
##    city        housing_score commute_score education_s~1 env_q~2 econo~3 taxat~4
##    <chr>               <dbl>         <dbl>         <dbl>   <dbl>   <dbl>   <dbl>
##  1 Aarhus               6.13          6.31          5.37    7.63    4.89    5.07
##  2 Adelaide             6.31          5.34          5.14    8.33    6.07    4.59
##  3 Albuquerque          7.26          5.06          4.15    7.32    6.51    4.35
##  4 Almaty               9.28          5.87          2.28    3.86    5.27    8.52
##  5 Amsterdam            3.05          6.12          6.18    7.60    5.05    4.95
##  6 Anchorage            5.43          4.72          3.62    9.27    6.51    4.77
##  7 Andorra              3.97          0             0       7.26    0       4.47
##  8 Ankara               9.93          5.29          2.03    2.94    4.09    4.32
```

```
##  9 Asheville            5.86          1.36          3.62    8.49    6.51    4.06
## 10 Asuncion             9.23          4.97          0       3.97    4.11    8.42
## # ... with 256 more rows, and abbreviated variable names 1: education_score,
## #   2: env_quality_score, 3: economy_score, 4: taxation_score
```

```
# Work Life Balance
worklifebal <- worklifebaldata %>%
  # select column variables
  select(City, Country, `TOTAL SCORE`,`Remote Jobs`, `Overworked Population`, Unemployment, Healthcare,
  # rename column names to be uniform and match other datasets' variables format
  rename(city = City, country = Country, wb_score = `TOTAL SCORE`,remote_jobs_perc = `Remote Jobs`, ove
worklifebal
```

```
## # A tibble: 100 x 9
##    city       country    wb_sc~1 remot~2 overw~3 unemp~4 healt~5 acces~6 city_~7
##    <chr>      <chr>        <dbl> <chr>   <chr>     <dbl>   <dbl>   <dbl>   <dbl>
##  1 Oslo       Norway       100   41.72%  11.20%     94.7   100      85      86.5
##  2 Bern       Switzerla~    99.5 44.86%  11.40%     99.8    99.6    78.6    91.8
##  3 Helsinki   Finland       99.2 38.92%  12.70%     89.3    96.7    73      94.9
##  4 Zurich     Switzerla~    96.3 44.86%  11.90%     99.2    99.2    78.6    92.8
##  5 Copenhagen Denmark       96.2 41.42%  10.50%     94.8    94.8    77.6    95.7
##  6 Geneva     Switzerla~    95.8 44.86%  11.90%     95.2    99.1    78.6    85.4
##  7 Ottawa     Canada        95.5 37.81%  10.10%     95.8    96.7    92.4    84.8
##  8 Sydney     Australia     94.0 38.79%  9.70%      95.9    99      67.4    77.6
##  9 Stuttgart  Germany       93.8 36.73%  11.70%     95.2    93.8    82      78.7
## 10 Munich     Germany       93.6 36.73%  11.90%     95.6    95.4    82      88.5
## # ... with 90 more rows, and abbreviated variable names 1: wb_score,
## #   2: remote_jobs_perc, 3: overwrk_pop_perc, 4: unemply_score,
## #   5: healthcare_score, 6: access_ment_health_score, 7: city_safety_score
```

---

3. Joining/Merging

We joined the three datasets using left join and by city. However, first, we used anti_join to see if there were cities not joined because of difference in formatting ("Frankfurt (am Main)" and "Frankfurt"). We used recode and mutate to change these city names to match. Then, we joined all three datasets by city.

```
# using antijoin to see if we need to recode any city names
anti_join(worklifebal, cityquality, costofliving, by ="city")
```

```
## # A tibble: 12 x 9
##    city          country wb_sc~1 remot~2 overw~3 unemp~4 healt~5 acces~6 city_~7
##    <chr>         <chr>     <dbl> <chr>   <chr>     <dbl>   <dbl>   <dbl>   <dbl>
##  1 Frankfurt (a~ Germany    91.1 36.73%  11.60%     92.3    94.7    82      70
##  2 Bremen        Germany    90.7 36.73%  10.80%     85.1    94.8    82      72.9
##  3 Graz          Austria    89.4 36.69%  12.50%     88.4    93.4    77.3    84.8
##  4 Minneapolis   USA        88.1 41.14%  13.00%     97.2    91.6    66.6    59.3
##  5 San Francisco USA        87.3 44.76%  12.90%     95.9    86.5    70.9    55.9
##  6 Virginia Bea~ USA        85.3 34.64%  11.70%     95.8    87.9    67.1    60.3
##  7 Sacramento    USA        84.9 40.34%  12.20%     93.9    88      66.6    58.8
##  8 Tampa         USA        84.3 39.94%  12.50%     96.4    87.2    67.2    60.8
##  9 Tucson        USA        81.8 37.29%  11.30%     95.5    84.2    67.7    55.7
## 10 Tulsa         USA        81.4 35.13%  13.20%     96.9    85.7    65.6    53.6
## 11 Wichita       USA        80.9 33.54%  12.70%     96.4    84.1    66.1    54.4
## 12 El Paso       USA        79.9 34.83%  11.50%     93.3    85.1    67.5    68.4
## # ... with abbreviated variable names 1: wb_score, 2: remote_jobs_perc,
```

```
## #   3: overwrk_pop_perc, 4: unemply_score, 5: healthcare_score,
## #   6: access_ment_health_score, 7: city_safety_score
```

```
# recode Frankfurt (am Main) to Frankfurt
worklifebal2 <- worklifebal %>%
  mutate(city = recode(city,
                       "Frankfurt (am Main)" = "Frankfurt"))
# left join all three datasets and drop na values
jointdata <- worklifebal2 %>%
  left_join(costofliving, by='city') %>%
  left_join(cityquality, by='city') %>%
  drop_na()
jointdata
```

```
## # A tibble: 89 x 21
##    city  country wb_sc~1 remot~2 overw~3 unemp~4 healt~5 acces~6 city_~7 utili~8
##    <chr> <chr>     <dbl> <chr>   <chr>     <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
##  1 Oslo  Norway    100   41.72%  11.20%     94.7   100      85      86.5    239.
##  2 Bern  Switze~    99.5 44.86%  11.40%     99.8    99.6    78.6    91.8    203.
##  3 Hels~ Finland    99.2 38.92%  12.70%     89.3    96.7    73      94.9    106.
##  4 Zuri~ Switze~    96.3 44.86%  11.90%     99.2    99.2    78.6    92.8    249.
##  5 Cope~ Denmark    96.2 41.42%  10.50%     94.8    94.8    77.6    95.7    201.
##  6 Gene~ Switze~    95.8 44.86%  11.90%     95.2    99.1    78.6    85.4    197.
##  7 Otta~ Canada     95.5 37.81%  10.10%     95.8    96.7    92.4    84.8    122.
##  8 Sydn~ Austra~    94.0 38.79%  9.70%      95.9    99      67.4    77.6    129.
##  9 Stut~ Germany    93.8 36.73%  11.70%     95.2    93.8    82      78.7    236.
## 10 Muni~ Germany    93.6 36.73%  11.90%     95.6    95.4    82      88.5    302.
## # ... with 79 more rows, 11 more variables: apt_1_bed <dbl>, apt_3_bed <dbl>,
## #   avg_month_net_sal <dbl>, avg_mortgage_int <dbl>, gasoline <dbl>,
## #   housing_score <dbl>, commute_score <dbl>, education_score <dbl>,
## #   env_quality_score <dbl>, economy_score <dbl>, taxation_score <dbl>, and
## #   abbreviated variable names 1: wb_score, 2: remote_jobs_perc,
## #   3: overwrk_pop_perc, 4: unemply_score, 5: healthcare_score,
## #   6: access_ment_health_score, 7: city_safety_score, 8: utility_cost
```

---

4. Wrangling

We explored our data using all six core `dplyr` functions (`filter`, `select`, `arrange`, `mutate`, `summarize`). Earlier in the report, we used `select` to only show certain columns.

Once we dropped all the rows that has na values and joined the 3 datasets by city, we were left with a dataset that 89 rows and 21 columns. To note, all three datasets had the the city variable in common with no other variables in common amongst the variables we selected. The cityquality dataset initially also had country as a variable however we decided to drop it because we decided to compare amongst cities and concluded that if we needed to compare by countries we could utilize the country variable in the worklifebal dataset. Thus the number of variables that were unique to each dataset is the number of columns subtracted by 1.

In this section, we used `filter` to see all the cities with a city safety score higher than 90. These cities were Bern, Helsinki, Zurich, Copenhagen, Tokyo, Singapore, and Dubai. Using `mutate`, we created a categorical variable based on a numerical variable (unemply_score) called "high_unemp_score" that provides information on whether a city has a high unemployment score, which is above 98. Also, we used `filter` to see the data on cities only in the United States, which we used in a visualization for our third research question later in the report.

We used `arrange` to see the cities sorted based on education score from highest to lowest. Hong Kong has

the highest education score with 9.7110. We removed percent signs from values in remote_jobs_perc and overwrk_pop_perc, and made the values numeric. We used `group_by` and `summarize` to compute summary statistics for 3 numeric variables and 2 categorical variables. We grouped by country and found the mean unemployment score, mean access to mental health care score, max city safety score, and number of records for each country. The countries' means and max safety score did not differ much from the cities' individual scores/values. We joined the summary data with the joined data to compare cities' individual values with the country average. Also, we computed the number of records with high_unemply_score being equal to true and saw that there are 7 cities with high_unemply_score equal to true.

```
# filters city safety score to be above 90
high_safety_scoredata <- jointdata %>% filter(city_safety_score > 90)
# first to be displayed in ouput
high_safety_scoredata
```

```
## # A tibble: 7 x 21
##   city   country wb_sc~1 remot~2 overw~3 unemp~4 healt~5 acces~6 city_~7 utili~8
##   <chr>  <chr>     <dbl> <chr>   <chr>     <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 Bern   Switze~    99.5 44.86%  11.40%     99.8    99.6    78.6    91.8    203.
## 2 Helsi~ Finland    99.2 38.92%  12.70%     89.3    96.7    73      94.9    106.
## 3 Zurich Switze~    96.3 44.86%  11.90%     99.2    99.2    78.6    92.8    249.
## 4 Copen~ Denmark    96.2 41.42%  10.50%     94.8    94.8    77.6    95.7    201.
## 5 Tokyo  Japan      92.5 36.52%  15.40%     96.2    99.3    84      92.5    166.
## 6 Singa~ Singap~    85.7 52.06%  16.90%     98.2    93.6    62.1   100      143.
## 7 Dubai  UAE        61.2 28.89%  23.40%    100      69.4    52.2    97.9    193.
## # ... with 11 more variables: apt_1_bed <dbl>, apt_3_bed <dbl>,
## #   avg_month_net_sal <dbl>, avg_mortgage_int <dbl>, gasoline <dbl>,
## #   housing_score <dbl>, commute_score <dbl>, education_score <dbl>,
## #   env_quality_score <dbl>, economy_score <dbl>, taxation_score <dbl>, and
## #   abbreviated variable names 1: wb_score, 2: remote_jobs_perc,
## #   3: overwrk_pop_perc, 4: unemply_score, 5: healthcare_score,
## #   6: access_ment_health_score, 7: city_safety_score, 8: utility_cost
```

```
# mutates the unemployment score to be categorial
jointdata <- jointdata %>% mutate(high_unemp_score = unemply_score > 98)

# filtering by USA
americadata <- jointdata %>%
  filter(country == "USA")
# second to be displayed in ouput
americadata
```

```
## # A tibble: 41 x 22
##    city  country wb_sc~1 remot~2 overw~3 unemp~4 healt~5 acces~6 city_~7 utili~8
##    <chr> <chr>     <dbl> <chr>   <chr>     <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
##  1 Seat~ USA        88.4 42.28%  13.10%     94.8    89.9    65.7    59.7    222.
##  2 Port~ USA        88.2 38.49%  12.50%     94.8    89.3    66.8    62.5    207.
##  3 Port~ USA        88.2 38.49%  12.50%     94.8    89.3    66.8    62.5    207.
##  4 Salt~ USA        87.8 43.34%  13.30%     98.2    88      66      65.1    123.
##  5 Bost~ USA        87.6 44.35%  12.30%     95      90.7    68.2    66.1    176.
##  6 Wash~ USA        87.1 49.77%  12.90%     95.4    86.6    67.8    56      157.
##  7 Omaha USA        86.0 38.90%  12.20%     97.3    89.5    66.4    60.6    235.
##  8 San ~ USA        85.8 39.52%  12.50%     94.5    88.6    66.6    66.2    190.
##  9 Denv~ USA        85.6 42.58%  12.40%     95.6    88.1    66.7    60      154.
## 10 Rale~ USA        85.3 41.30%  12.50%     96.7    88.7    67.5    65.9    170.
## # ... with 31 more rows, 12 more variables: apt_1_bed <dbl>, apt_3_bed <dbl>,
## #   avg_month_net_sal <dbl>, avg_mortgage_int <dbl>, gasoline <dbl>,
```

```
## #   housing_score <dbl>, commute_score <dbl>, education_score <dbl>,
## #   env_quality_score <dbl>, economy_score <dbl>, taxation_score <dbl>,
## #   high_unemp_score <lgl>, and abbreviated variable names 1: wb_score,
## #   2: remote_jobs_perc, 3: overwrk_pop_perc, 4: unemply_score,
## #   5: healthcare_score, 6: access_ment_health_score, 7: city_safety_score, ...
```

```
# arrange in descending order by education score (highest to lowest)
# third to be displayed in output
jointdata  %>% arrange(desc(education_score))
```

```
## # A tibble: 89 x 22
##    city  country wb_sc~1 remot~2 overw~3 unemp~4 healt~5 acces~6 city_~7 utili~8
##    <chr> <chr>     <dbl> <chr>   <chr>     <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
##  1 Hong~ Hong K~    77.1 40.27%  17.90%     95.8    78.1   100      85.8    199.
##  2 Lond~ UK         89.7 43.50%  13.80%     93.2    94.6    83.8    66.3    288.
##  3 Lond~ UK         89.7 43.50%  13.80%     93.2    94.6    83.8    66.3    130.
##  4 Bost~ USA        87.6 44.35%  12.30%     95      90.7    68.2    66.1    176.
##  5 Los ~ USA        81.6 39.09%  12.80%     92.6    85.6    66.5    58.9    141.
##  6 Sing~ Singap~    85.7 52.06%  16.90%     98.2    93.6    62.1   100      143.
##  7 New ~ USA        84.2 41.96%  11.90%     93.3    85.5    68.1    62.4    171.
##  8 Chic~ USA        82.9 39.18%  12.50%     93.3    86.4    68      47.6    164.
##  9 Toro~ Canada     91.1 37.81%  10.10%     90.4    97.2    92.4    77.3    129.
## 10 Muni~ Germany    93.6 36.73%  11.90%     95.6    95.4    82      88.5    302.
## # ... with 79 more rows, 12 more variables: apt_1_bed <dbl>, apt_3_bed <dbl>,
## #   avg_month_net_sal <dbl>, avg_mortgage_int <dbl>, gasoline <dbl>,
## #   housing_score <dbl>, commute_score <dbl>, education_score <dbl>,
## #   env_quality_score <dbl>, economy_score <dbl>, taxation_score <dbl>,
## #   high_unemp_score <lgl>, and abbreviated variable names 1: wb_score,
## #   2: remote_jobs_perc, 3: overwrk_pop_perc, 4: unemply_score,
## #   5: healthcare_score, 6: access_ment_health_score, 7: city_safety_score, ...
```

```
# source: https://www.tutorialspoint.com/how-to-remove-percentage-sign-at-last-position-from-every-valu
# remove percent signs in remote jobs variable
jointdata$remote_jobs_perc <- gsub("%", "", jointdata$remote_jobs_perc)
# remove percent signs in overworked population variable
jointdata$overwrk_pop_perc <- gsub("%", "", jointdata$overwrk_pop_perc)

jointdata <- jointdata %>%
  # make remote jobs, and overworked population variables numeric
  mutate(remote_jobs_perc = as.numeric(remote_jobs_perc), overwrk_pop_perc = as.numeric(overwrk_pop_per


# creates new variable holding country average for mental health access and unemployment
summary_cntry_data <- jointdata %>%
  # group by country
  group_by(country) %>%
  # summary stats:
  # mean of unemployment (1st summary stat for numeric variable)
  # mean access to mental health (2nd summary stat for numeric variable)
  # max of city safety in each country (3rd summary stat for numeric variable)
  # number of records per country (1st summary stat for categorical variable)
  summarize(ctry_unemployment = mean(unemply_score), ctry_mentalhealth = mean(access_ment_health_score)

# joins summary data to jointdata
# fourth to be displayed in output
```

```
jointdata %>%
  left_join(summary_cntry_data, by = "country") # combining rows by country
```

```
## # A tibble: 89 x 26
##    city  country wb_sc~1 remot~2 overw~3 unemp~4 healt~5 acces~6 city_~7 utili~8
##    <chr> <chr>     <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
##  1 Oslo  Norway    100      41.7    11.2    94.7   100      85      86.5    239.
##  2 Bern  Switze~   99.5    44.9    11.4    99.8    99.6    78.6    91.8    203.
##  3 Hels~ Finland   99.2    38.9    12.7    89.3    96.7    73      94.9    106.
##  4 Zuri~ Switze~   96.3    44.9    11.9    99.2    99.2    78.6    92.8    249.
##  5 Cope~ Denmark   96.2    41.4    10.5    94.8    94.8    77.6    95.7    201.
##  6 Gene~ Switze~   95.8    44.9    11.9    95.2    99.1    78.6    85.4    197.
##  7 Otta~ Canada    95.5    37.8    10.1    95.8    96.7    92.4    84.8    122.
##  8 Sydn~ Austra~   94.0    38.8     9.7    95.9    99      67.4    77.6    129.
##  9 Stut~ Germany   93.8    36.7    11.7    95.2    93.8    82      78.7    236.
## 10 Muni~ Germany   93.6    36.7    11.9    95.6    95.4    82      88.5    302.
## # ... with 79 more rows, 16 more variables: apt_1_bed <dbl>, apt_3_bed <dbl>,
## #   avg_month_net_sal <dbl>, avg_mortgage_int <dbl>, gasoline <dbl>,
## #   housing_score <dbl>, commute_score <dbl>, education_score <dbl>,
## #   env_quality_score <dbl>, economy_score <dbl>, taxation_score <dbl>,
## #   high_unemp_score <lgl>, ctry_unemployment <dbl>, ctry_mentalhealth <dbl>,
## #   max_safety <dbl>, num_records_cntry <int>, and abbreviated variable names
## #   1: wb_score, 2: remote_jobs_perc, 3: overwrk_pop_perc, ...
```

```
# number of cities in each country that have high unemployment score (above 98)
# (2nd summary stat for categorical variable)
summary_high_unemp <- jointdata %>%
  # filter to show only countries with high_unemp_score as true
  filter(high_unemp_score == TRUE) %>%
  # number cities with true value
  summarize(num_high_unemp = n())
# fifth to be displayed in output
summary_high_unemp
```

```
## # A tibble: 1 x 1
##   num_high_unemp
##            <int>
## 1              7
```

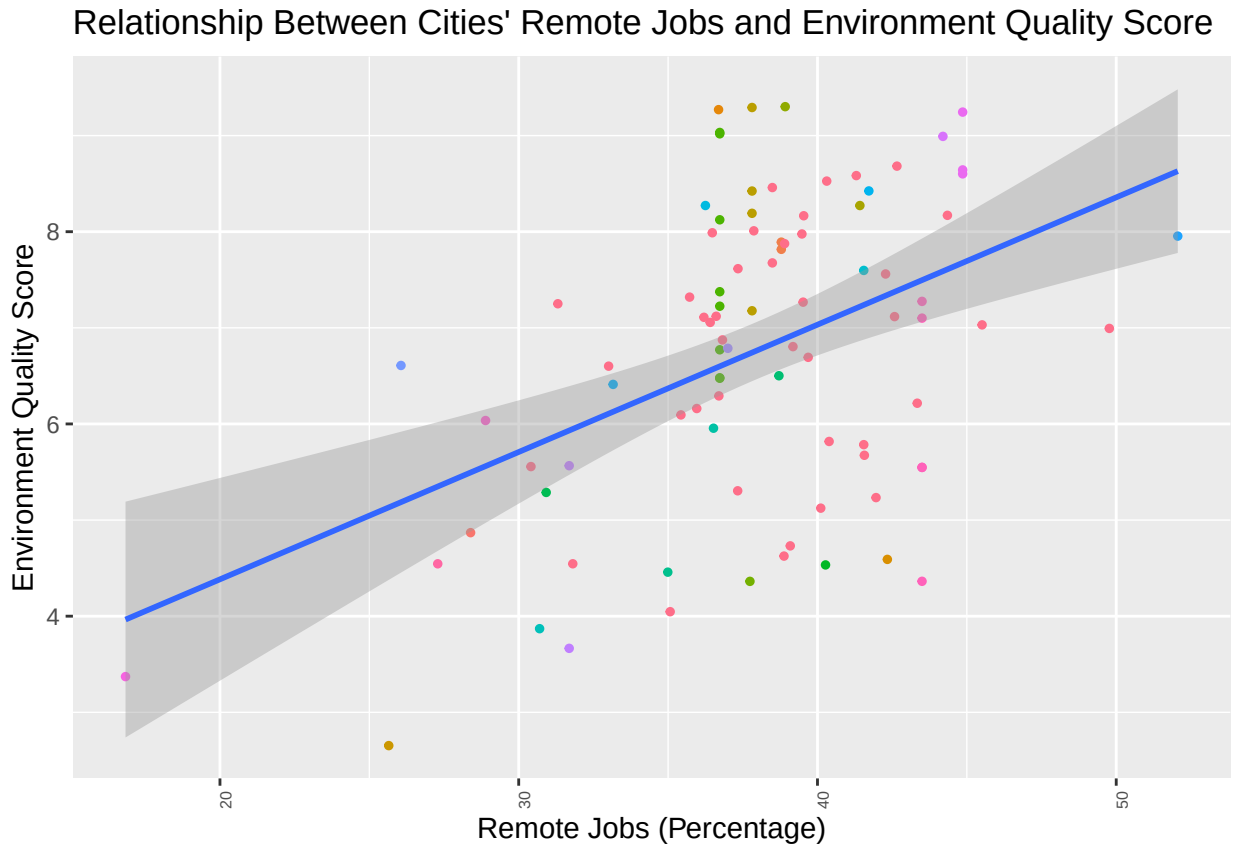
---

5. Visualizing

We created two scatterplots involving 3 variables. The first scatterplot shows the relationship between percentage of remote jobs and the environmental quality score in cities and is colored by country. The second scatterplot shows the relationship between education score and the cost of a 3-bedroom apartment in cities and is colored by country. We created a bar graph and a scatter plot that each involve two variables. The bar graph shows the unemployment score for each US city. The scatter plot shows the relationship between access to mental health care and unemployment score in US cities. We created two density plots that involve one variable each. The first density plot shows the distribution of gasoline cost. The second density plot shows the distribution of education score.

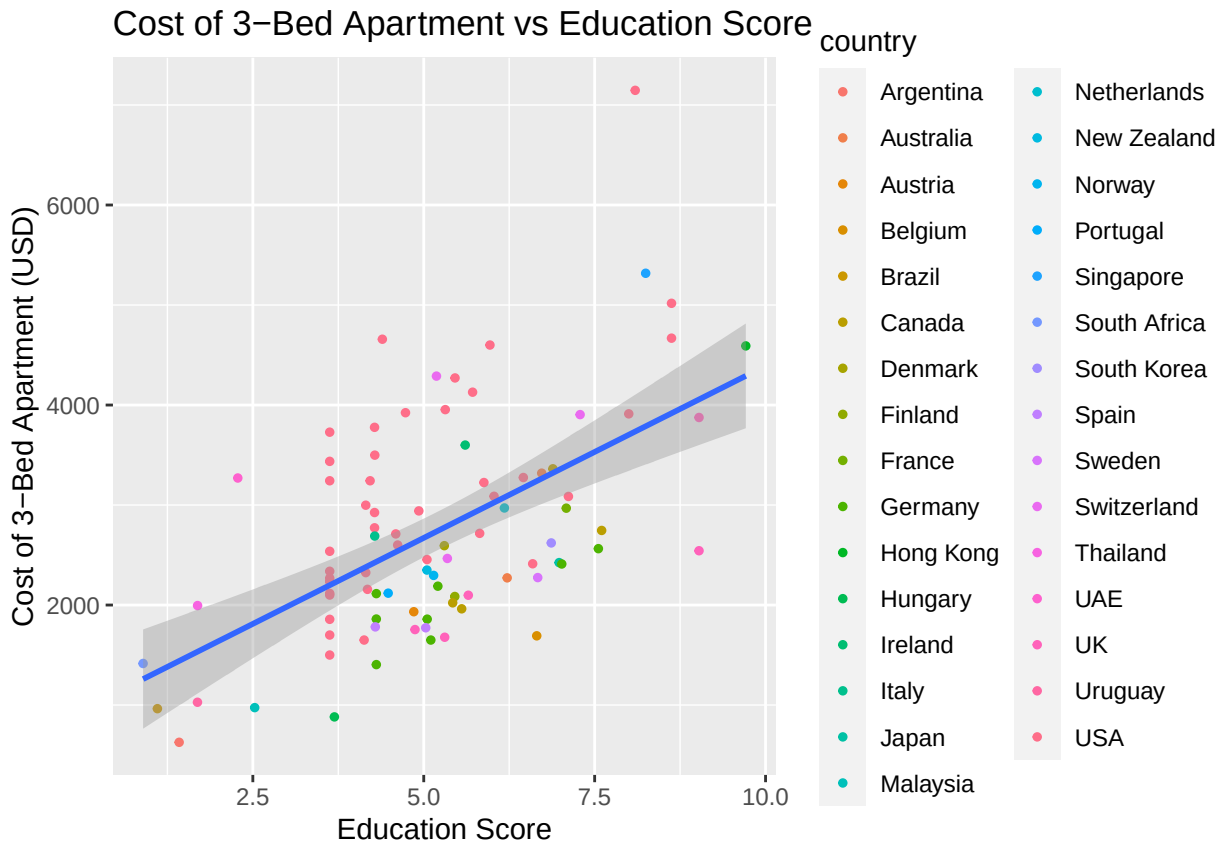
```
#1
ggplot(jointdata, aes(x = remote_jobs_perc, y = env_quality_score)) +
  geom_point(size = 1, aes(color=country)) +
  # changes the angle and size of the x-axis text
```

```
  theme(axis.text.x=element_text(angle = 90, size = 6)) +
  # remove legend
  theme(legend.position = "none") +
  # creates title, x-axis, and y-axis labels
  labs(x = "Remote Jobs (Percentage)", y = "Environment Quality Score", title = "Relationship Between C
  # add regression trend line
  geom_smooth(method = 'lm', aes(x= remote_jobs_perc,y=env_quality_score))
```



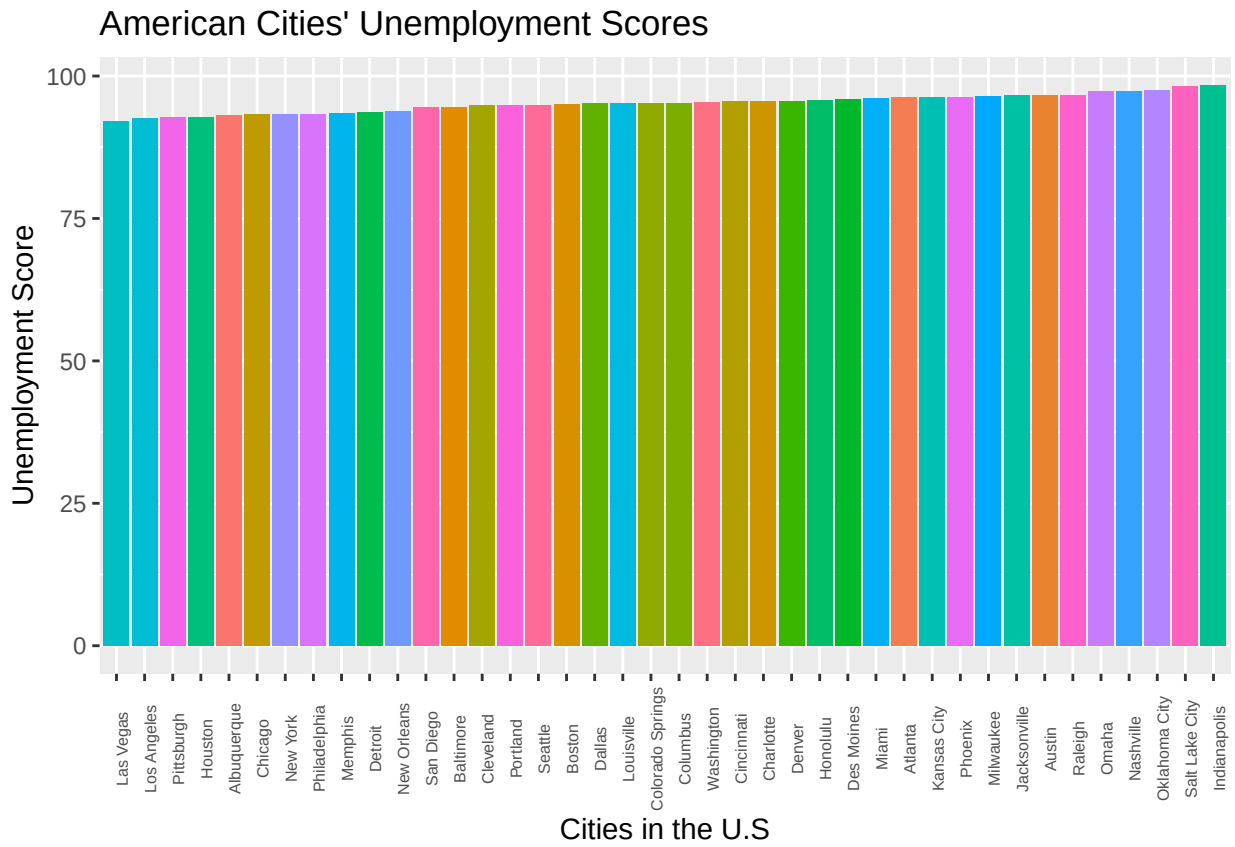

```
#This creates scatter plot with percentage of remote jobs and environmental quality score points are co


#2
ggplot(jointdata, aes(x = education_score, y = apt_3_bed)) +
  geom_point(size = 1, aes(color=country)) +   labs(title = "Cost of 3-Bed Apartment vs Education Score
  # adds regression trend line
  geom_smooth(method = 'lm', aes(x= education_score,y=apt_3_bed))
```



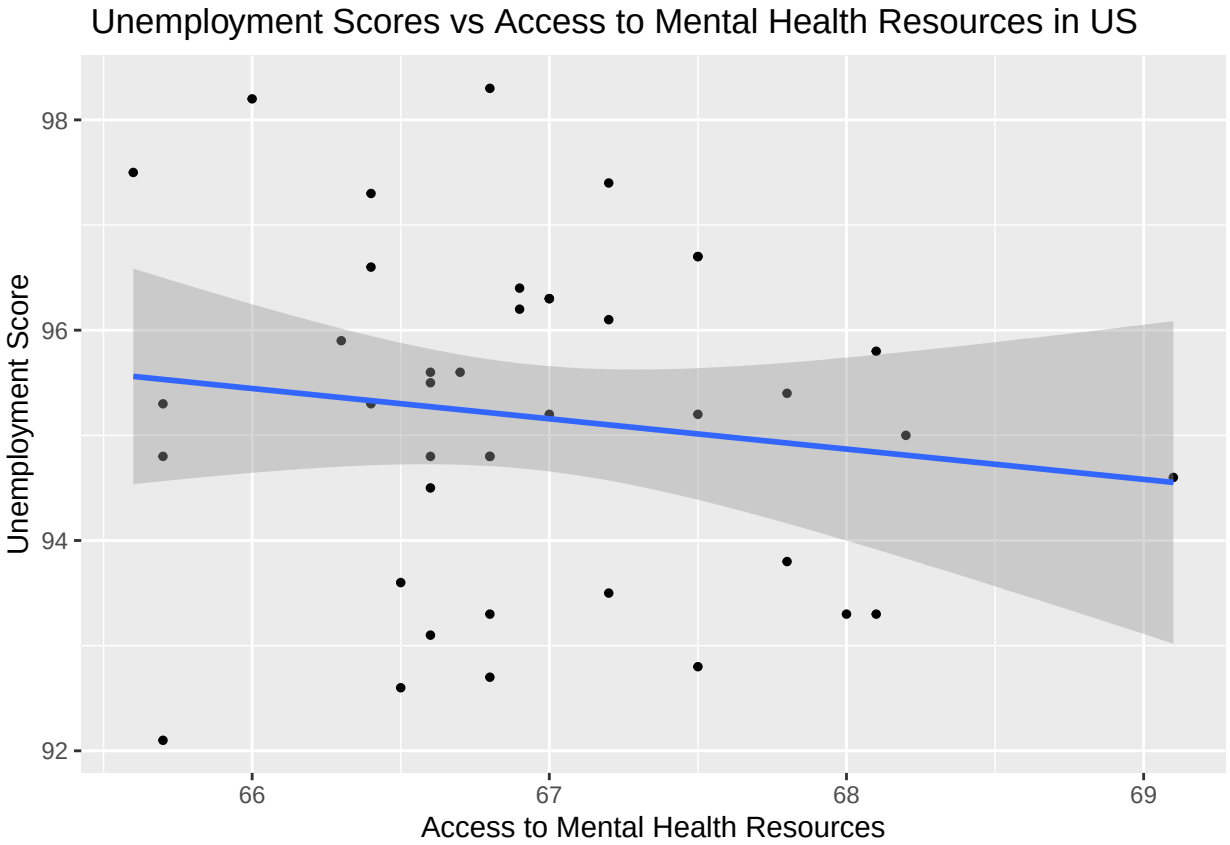

```
# This creates scatter plot with education score and cost of 3-bed apartment for each city and the poin


#3
ggplot(americadata, aes(x = reorder(city, unemply_score), y = unemply_score, fill = city)) +
  geom_bar(stat = "summary") +
  # changes the angle and size of the x-axis text
  theme(axis.text.x=element_text(angle = 90, size = 6)) +
  # removes legend
  theme(legend.position = "none") +
  # creates title, x-axis, and y-axis labels
  labs(x = "Cities in the U.S", y = "Unemployment Score", title = "American Cities' Unemployment Scores
```



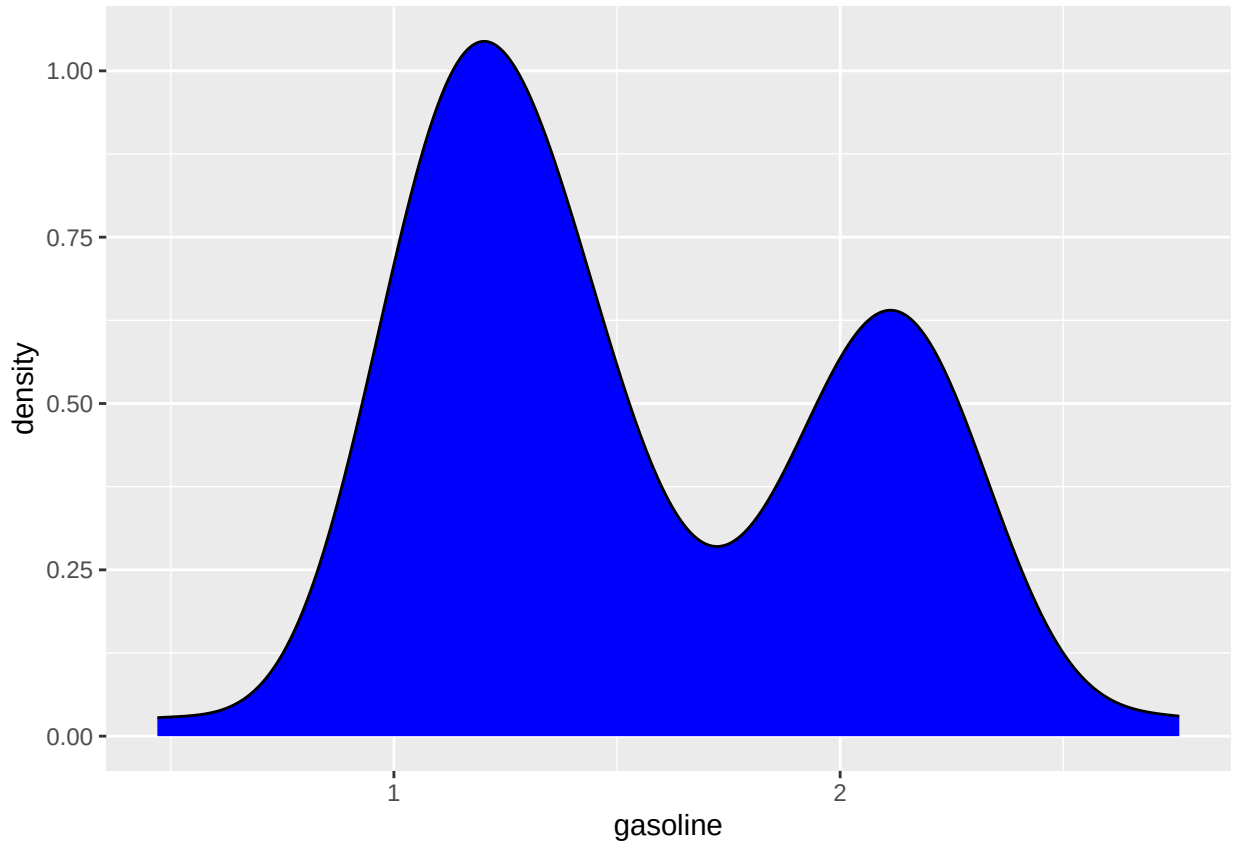

```
# This creates bar plot between city and unemployment score in US.


#4
ggplot(americadata, aes(x = access_ment_health_score, y = unemply_score)) +
  geom_point(size = 1) +
  # removes legend
   theme(legend.position = "none") +
  # adds regression trend line
  geom_smooth(method = 'lm', aes(x= access_ment_health_score,y=unemply_score)) +
  # creates title, x-axis, and y-axis labels
   labs(x = "Access to Mental Health Resources", y = "Unemployment Score", title = " Unemployment Scores
```



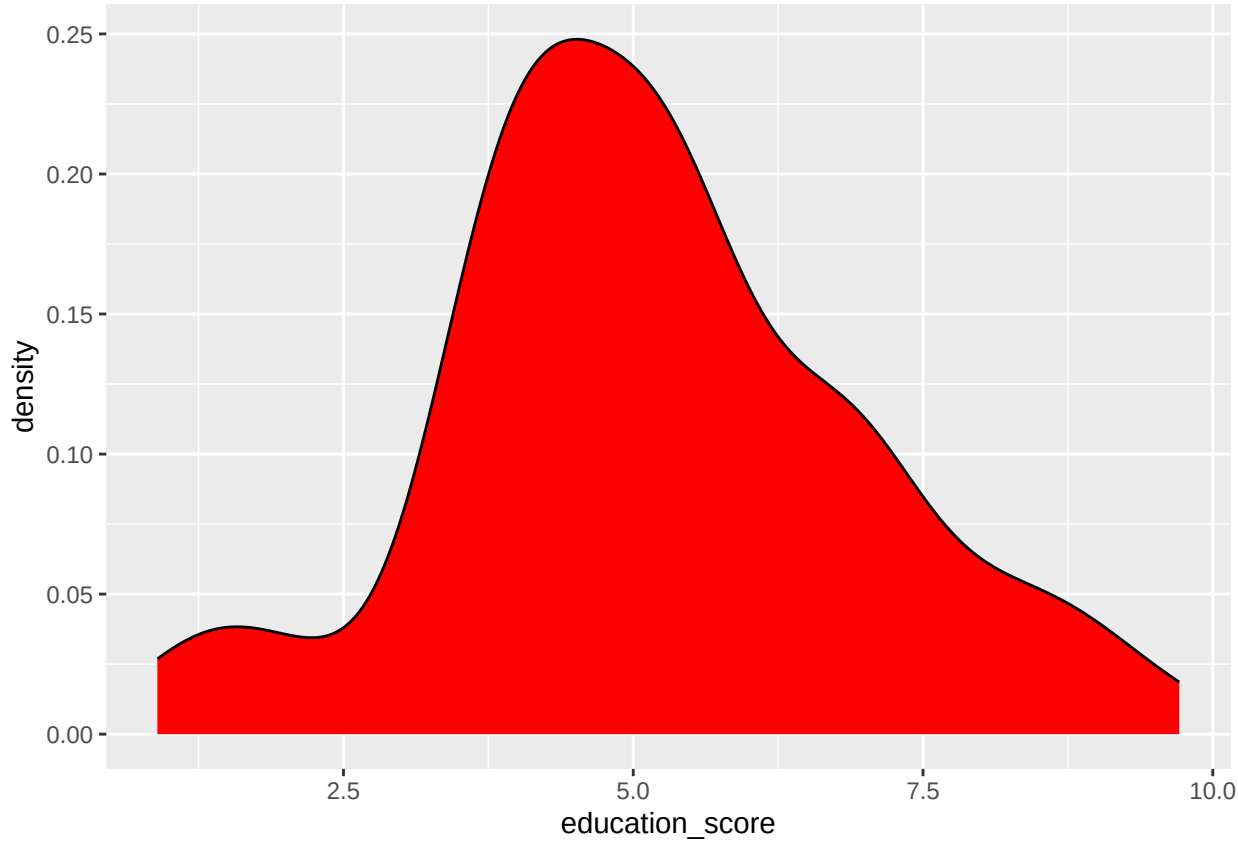

```
# This creates scatter plot with access to mental health and unemployment score

#5
jointdata %>% ggplot(aes(x = gasoline)) +
  # color blue
  geom_density(alpha = 1, fill = "blue")
```




```
# This is a density plot for cost of gasoline


#6
jointdata %>% ggplot(aes(x = education_score)) +
  # color red
  geom_density(alpha = 1, fill = "red")
```




```
# This is a density plot for education score
```

---

6. Discussion

Our first research question was "What is the relationship between a city's education score and cost of a 3-bedroom apartment?" and we expected there to be a positive relationship between education score and the cost of a 3-bedroom apartment in a city because more expensive areas will have more funds for better education. According to our scatterplot in the visualization section (graph 2), there is a mostly positive relationship across countries, which can be seen through the regression line.

Our second research question was "Does the number of remote jobs affect the environmental quality in each city?". From what is seen from the visualization (graph 1), the more percentage of remote jobs there are in each city, the better an environmental score. Although we cannot confirm causation, there is a positive correlation with the percentage of remote jobs and environmental quality score across cities globally which supports the hypothesis.

For the research question 'Does access to mental healthcare resources affect unemployment levels?', we hypothesized that as access to mental health resources increase, unemployment levels will decrease (higher unemployment scores) in American cities. Graph 3 shows in ascending order which American cities have the highest unemployment scores(low unemployment rate) through a bar graph. This graph shows Las Vegas as having the lowest unemployment scores while Indianapolis has the highest. Graph 4 shows a slight negative correlation between access to mental health resources and unemployment scores across American cities through a scatter plot. This was surprising as we expected better access to mental health resources to increase unemployment scores however our visualization rejects our hypothesis.

Graph 5 depicts how common individual price points were for gas prices across cities via a density graph.

There are two main peaks around $1.25 and $2.25 with $1.25 having a significantly higher peak indicating that most cities sold gas at these price points. Graph 6 depicts how common education scores were for cities globally across cities via a density graph. The graph shows that few cities has an education score below 2.5 or above 7.5. The peak was around 4.5. This result was surprising as we expected more countries to be on either end of the spectrum with a low education or high education score but most countries ended up around the middle.

Two out of the three research questions supported our underlining hypothesis - the more expensive it costs living in a city, the better quality life one will likely have. Our data supported that more expensive rent was correlated with higher education levels. Additionally traditionally first world countries that have better infrastructure can support remote jobs which reduces pollution emitted by commuting by car or public transportation. This showed in our graph that showed a positive correlation between remote jobs and environmental score. For the third research question, it is posisble that there are confounding variables that could explain why more access to mental health resources is related to higher levels of employment.

Our density graphs show that most cities hover around an average education score and most cities fall within the range of two peak price points.Thus when considering future areas of residence it would be wise to choose cities that fall within the lower peak price of graph 6 and average or higher education levels. Data points in the bottom right half of graph 2 would represent the ideal cities.

From start to finish, the process of pre-processing and processing data proved more time-consuming than initially thought. It was challenging to find datasets that fit the numerical variable requirements and that had one variable in common. We tried exploring other areas of interest however due to lack of common variables, we decided to explore datasets on cities and countries of the world since most political regions are required to report data on their population. Once we collected the data, it was challenging to only choose 3 research variables from the 21 variables we decided to analyze, however once we decided on our research questions, it was just a matter of looking at functions we learned to class to tidy and wrangle the data. Although tidying and wrangling the data itself wasn't too challenging, it was time-comsuming and took trial and error.

This project led us to gain insight on how 80% of data science is tidying and wrangling data and led us to gain new found respect for the process. Every group member was involved in the data selection. We mostly worked together remotely or in-person. Individually, we each chose a research question, tidied the data and did the data wrangling and visualization appropriate for our respective questions. Additionally, we jointly collaborated to write the introduction and conclusion.

Formatting: (2 pts)

Comment your code, write full sentences, and knit your file!