THE DIFFERENCE BETWEEN ACTIVE TRAVEL TO SCHOOL BETWEEN ABERDEEN CITY AND ABERDEENSHIRE

Awaneesh Kumar Tiwari

2024-02-27

Document version: 1.0

Aim of the report

The aim of the report is to find the difference between active travel to school between Aberdeen City and Aberdeenshire and explain every step in the health data science pipeline, from reading in the data to producing the visualization.

Load packages

```
library(tidyverse) # This library is a collection of open-source packages in R programming
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.1     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
                  # language. It is used for data prepration, wrangling and visualization.
library(viridis)  # It provides the base functions for generating the color maps in base R.
```

```
## Loading required package: viridisLite
```

```
library(plotly)    # interactive visualisations
```

```
##
## Attaching package: 'plotly'
```

```
## The following object is masked from 'package:ggplot2':
##
##     last_plot

## The following object is masked from 'package:stats':
##
##     filter

## The following object is masked from 'package:graphics':
##
##     layout
```

```
library(ggplot2)  # Plotting Multiple Charts and changing
                  # secondary axis to percentage
```

Read in the data

```
# Here read_csv function is used to read the scotpho_active_travel.csv file.
active_travel <- read_csv("scotpho_active_travel.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   indicator = col_character(),
##   area_name = col_character(),
##   area_code = col_character(),
##   area_type = col_character(),
##   year = col_double(),
##   period = col_character(),
##   numerator = col_double(),
##   measure = col_double(),
##   lower_confidence_interval = col_double(),
##   upper_confidence_interval = col_double(),
##   definition = col_character(),
##   data_source = col_character()
## )
```

```
# To review the column specification and number of rows and cols.
glimpse(active_travel)
```

```
## Rows: 1,081
## Columns: 12
## $ indicator                <chr> "Active travel to work", "Active travel to w~
## $ area_name                <chr> "Scotland", "NHS Ayrshire & Arran", "NHS Bor~
## $ area_code                <chr> "S00000001", "S08000015", "S08000016", "S080~
## $ area_type                <chr> "Scotland", "Health board", "Health board", ~
## $ year                     <dbl> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 20~
## $ period                   <chr> "2007/2008 survey years", "2007/2008 survey ~
## $ numerator                <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ measure                  <dbl> 14.2, 9.2, 16.8, 16.3, 9.6, 10.5, 15.8, 13.0~
```

```
## $ lower_confidence_interval <dbl> 13.4, 6.9, 10.9, 12.0, 7.2, 7.9, 13.6, 11.4,~
## $ upper_confidence_interval <dbl> 15.0, 11.5, 22.8, 20.6, 12.0, 13.2, 18.0, 14~
## $ definition                <chr> "Percentage", "Percentage", "Percentage", "P~
## $ data_source               <chr> "Scottish Household Survey (SHS)", "Scottish~
```

```
head(active_travel)
```

```
## # A tibble: 6 x 12
##   indicator    area_name    area_code area_type  year period   numerator measure
##   <chr>        <chr>        <chr>     <chr>     <dbl> <chr>        <dbl>   <dbl>
## 1 Active trav~ Scotland     S00000001 Scotland   2007 2007/20~        NA    14.2
## 2 Active trav~ NHS Ayrshir~ S08000015 Health b~  2007 2007/20~        NA     9.2
## 3 Active trav~ NHS Borders  S08000016 Health b~  2007 2007/20~        NA    16.8
## 4 Active trav~ NHS Dumfrie~ S08000017 Health b~  2007 2007/20~        NA    16.3
## 5 Active trav~ NHS Fife     S08000029 Health b~  2007 2007/20~        NA     9.6
## 6 Active trav~ NHS Forth V~ S08000019 Health b~  2007 2007/20~        NA    10.5
## # ... with 4 more variables: lower_confidence_interval <dbl>,
## #   upper_confidence_interval <dbl>, definition <chr>, data_source <chr>
```

Inspect the data

```
#Review the values in the variables
group_by_all(active_travel)
```

```
## # A tibble: 1,081 x 12
## # Groups:   indicator, area_name, area_code, area_type, year, period,
## #   numerator, measure, lower_confidence_interval, upper_confidence_interval,
## #   definition, data_source [1,081]
##    indicator   area_name    area_code area_type  year period   numerator measure
##    <chr>       <chr>        <chr>     <chr>     <dbl> <chr>        <dbl>   <dbl>
##  1 Active tra~ Scotland     S00000001 Scotland   2007 2007/20~        NA    14.2
##  2 Active tra~ NHS Ayrshir~ S08000015 Health b~  2007 2007/20~        NA     9.2
##  3 Active tra~ NHS Borders  S08000016 Health b~  2007 2007/20~        NA    16.8
##  4 Active tra~ NHS Dumfrie~ S08000017 Health b~  2007 2007/20~        NA    16.3
##  5 Active tra~ NHS Fife     S08000029 Health b~  2007 2007/20~        NA     9.6
##  6 Active tra~ NHS Forth V~ S08000019 Health b~  2007 2007/20~        NA    10.5
##  7 Active tra~ NHS Grampian S08000020 Health b~  2007 2007/20~        NA    15.8
##  8 Active tra~ NHS Greater~ S08000031 Health b~  2007 2007/20~        NA    13
##  9 Active tra~ NHS Highland S08000022 Health b~  2007 2007/20~        NA    20.8
## 10 Active tra~ NHS Lanarks~ S08000032 Health b~  2007 2007/20~        NA     7.7
## # ... with 1,071 more rows, and 4 more variables:
## #   lower_confidence_interval <dbl>, upper_confidence_interval <dbl>,
## #   definition <chr>, data_source <chr>
```

```
# Column name 'indicator' has two values 'Active travel to work' and 'Active travel to school' values.
# We will have to filter out 'Active travel to school'.
active_travel_school <- active_travel %>%
filter(indicator == "Active travel to school")

group_by_all(active_travel_school)
```

```
## # A tibble: 658 x 12
## # Groups:   indicator, area_name, area_code, area_type, year, period,
## #   numerator, measure, lower_confidence_interval, upper_confidence_interval,
## #   definition, data_source [658]
##    indicator     area_name area_code area_type  year period    numerator measure
##    <chr>         <chr>     <chr>     <chr>     <dbl> <chr>         <dbl>   <dbl>
##  1 Active trave~ Scotland  S00000001 Scotland   2008 2008/09 ~    203360    52.2
##  2 Active trave~ Scotland  S00000001 Scotland   2009 2009/10 ~    206758    50.1
##  3 Active trave~ Scotland  S00000001 Scotland   2010 2010/11 ~    215884    49.8
##  4 Active trave~ Scotland  S00000001 Scotland   2011 2011/12 ~    212713    50.2
##  5 Active trave~ Scotland  S00000001 Scotland   2012 2012/13 ~    225982    50.4
##  6 Active trave~ Scotland  S00000001 Scotland   2013 2013/14 ~    234004    50.8
##  7 Active trave~ Scotland  S00000001 Scotland   2014 2014/15 ~    244390    51.1
##  8 Active trave~ Scotland  S00000001 Scotland   2015 2015/16 ~    237687    50.4
##  9 Active trave~ Scotland  S00000001 Scotland   2016 2016/17 ~    225427    49.8
## 10 Active trave~ Scotland  S00000001 Scotland   2017 2017/18 ~    229645    49.4
## # ... with 648 more rows, and 4 more variables:
## #   lower_confidence_interval <dbl>, upper_confidence_interval <dbl>,
## #   definition <chr>, data_source <chr>
```

```
# Now, we have a total of 658 rows for 'Active travel to school' out of 1081 rows.


#Similarly, we will have to filter out area_name (Aberdeen city and Aberdeenshire)

active_travel_school_Aberdeencity_Aberdeenshire <- active_travel_school %>%
filter(area_name %in% c("Aberdeen City","Aberdeenshire"))

group_by_all(active_travel_school_Aberdeencity_Aberdeenshire)
```

```
## # A tibble: 28 x 12
## # Groups:   indicator, area_name, area_code, area_type, year, period,
## #   numerator, measure, lower_confidence_interval, upper_confidence_interval,
## #   definition, data_source [28]
##    indicator    area_name  area_code area_type   year period   numerator measure
##    <chr>        <chr>      <chr>     <chr>      <dbl> <chr>        <dbl>   <dbl>
##  1 Active trav~ Aberdeen ~ S12000033 Council a~  2008 2008/09~     10748    65.5
##  2 Active trav~ Aberdeen ~ S12000033 Council a~  2009 2009/10~     10515    64.6
##  3 Active trav~ Aberdeen ~ S12000033 Council a~  2010 2010/11~      7284    61.8
##  4 Active trav~ Aberdeen ~ S12000033 Council a~  2011 2011/12~      8938    62.0
##  5 Active trav~ Aberdeen ~ S12000033 Council a~  2012 2012/13~      9279    62.6
##  6 Active trav~ Aberdeen ~ S12000033 Council a~  2013 2013/14~      7914    59.3
##  7 Active trav~ Aberdeen ~ S12000033 Council a~  2014 2014/15~      9680    59.2
##  8 Active trav~ Aberdeen ~ S12000033 Council a~  2015 2015/16~      9840    57.6
##  9 Active trav~ Aberdeen ~ S12000033 Council a~  2016 2016/17~      9946    59.6
## 10 Active trav~ Aberdeen ~ S12000033 Council a~  2017 2017/18~     10448    59.4
## # ... with 18 more rows, and 4 more variables: lower_confidence_interval <dbl>,
## #   upper_confidence_interval <dbl>, definition <chr>, data_source <chr>
```

```
#28 rows have been filtered out of 658 rows. 14 rows for both Aberdeen City and Aberdeenshire.
```

Prepare the data

```
#filter out column: area_name (Aberdeen city)
travel_school_Aberdeencity <- active_travel_school %>%
 filter(area_name == "Aberdeen City")
group_by_all(travel_school_Aberdeencity)
```

```
## # A tibble: 14 x 12
## # Groups:   indicator, area_name, area_code, area_type, year, period,
## #   numerator, measure, lower_confidence_interval, upper_confidence_interval,
## #   definition, data_source [14]
##    indicator    area_name  area_code area_type   year period   numerator measure
##    <chr>        <chr>      <chr>     <chr>      <dbl> <chr>        <dbl>   <dbl>
##  1 Active trav~ Aberdeen ~ S12000033 Council a~  2008 2008/09~     10748    65.5
##  2 Active trav~ Aberdeen ~ S12000033 Council a~  2009 2009/10~     10515    64.6
##  3 Active trav~ Aberdeen ~ S12000033 Council a~  2010 2010/11~      7284    61.8
##  4 Active trav~ Aberdeen ~ S12000033 Council a~  2011 2011/12~      8938    62.0
##  5 Active trav~ Aberdeen ~ S12000033 Council a~  2012 2012/13~      9279    62.6
##  6 Active trav~ Aberdeen ~ S12000033 Council a~  2013 2013/14~      7914    59.3
##  7 Active trav~ Aberdeen ~ S12000033 Council a~  2014 2014/15~      9680    59.2
##  8 Active trav~ Aberdeen ~ S12000033 Council a~  2015 2015/16~      9840    57.6
##  9 Active trav~ Aberdeen ~ S12000033 Council a~  2016 2016/17~      9946    59.6
## 10 Active trav~ Aberdeen ~ S12000033 Council a~  2017 2017/18~     10448    59.4
## 11 Active trav~ Aberdeen ~ S12000033 Council a~  2018 2018/19~     10051    58.5
## 12 Active trav~ Aberdeen ~ S12000033 Council a~  2019 2019/20~     10046    57.2
## 13 Active trav~ Aberdeen ~ S12000033 Council a~  2020 2020/21~      9812    61.2
## 14 Active trav~ Aberdeen ~ S12000033 Council a~  2021 2021/22~     10239    60.7
## # ... with 4 more variables: lower_confidence_interval <dbl>,
## #   upper_confidence_interval <dbl>, definition <chr>, data_source <chr>
```

```
#14 rows of Aberdeen City have been filtered out of 658 rows.

#filter out column: area_name (Aberdeenshire)
travel_school_Aberdeenshire <- active_travel_school %>%
 filter(area_name == "Aberdeenshire")
group_by_all(travel_school_Aberdeenshire)
```

```
## # A tibble: 14 x 12
## # Groups:   indicator, area_name, area_code, area_type, year, period,
## #   numerator, measure, lower_confidence_interval, upper_confidence_interval,
## #   definition, data_source [14]
##    indicator    area_name  area_code area_type   year period   numerator measure
##    <chr>        <chr>      <chr>     <chr>      <dbl> <chr>        <dbl>   <dbl>
##  1 Active trav~ Aberdeens~ S12000034 Council a~  2008 2008/09~     13357    46.3
##  2 Active trav~ Aberdeens~ S12000034 Council a~  2009 2009/10~     12165    45.0
##  3 Active trav~ Aberdeens~ S12000034 Council a~  2010 2010/11~     13370    45.9
##  4 Active trav~ Aberdeens~ S12000034 Council a~  2011 2011/12~     13581    46.7
##  5 Active trav~ Aberdeens~ S12000034 Council a~  2012 2012/13~     12373    47.7
##  6 Active trav~ Aberdeens~ S12000034 Council a~  2013 2013/14~     14269    47.9
##  7 Active trav~ Aberdeens~ S12000034 Council a~  2014 2014/15~     14601    46.6
##  8 Active trav~ Aberdeens~ S12000034 Council a~  2015 2015/16~     13993    45.0
##  9 Active trav~ Aberdeens~ S12000034 Council a~  2016 2016/17~     14244    47.3
```

```
## 10 Active trav~ Aberdeens~ S12000034 Council a~  2017 2017/18~     13627    43.8
## 11 Active trav~ Aberdeens~ S12000034 Council a~  2018 2018/19~     14172    45.8
## 12 Active trav~ Aberdeens~ S12000034 Council a~  2019 2019/20~     14001    46.2
## 13 Active trav~ Aberdeens~ S12000034 Council a~  2020 2020/21~     14682    49.7
## 14 Active trav~ Aberdeens~ S12000034 Council a~  2021 2021/22~     13681    48.2
## # ... with 4 more variables: lower_confidence_interval <dbl>,
## #   upper_confidence_interval <dbl>, definition <chr>, data_source <chr>
```

```
#14 rows of Aberdeenshire have been filtered out of 658 rows.

#join both dataset
Aberdeencity_aberdeenshire_dataset<-full_join(travel_school_Aberdeencity, travel_school_Aberdeenshire,
by = "year")

# Here by default suffix are added in the new variables
# for example: .x added with Aberdeen City and .y added with Aberdeenshire variables
select_dataset= Aberdeencity_aberdeenshire_dataset %>%
select('indicator.x', 'year', 'numerator.x', 'numerator.y', 'measure.x', 'measure.y') %>%
mutate(numerator_diff=numerator.x-numerator.y, measure_diff=measure.x-measure.y )
# mutate function has been used to create two new variables 'numberator_diff' and 'measure_diff'

head(select_dataset)
```

```
## # A tibble: 6 x 8
##   indicator.x    year numerator.x numerator.y measure.x measure.y numerator_diff
##   <chr>         <dbl>       <dbl>       <dbl>     <dbl>     <dbl>          <dbl>
## 1 Active trave~  2008       10748       13357      65.5      46.3          -2609
## 2 Active trave~  2009       10515       12165      64.6      45.0          -1650
## 3 Active trave~  2010        7284       13370      61.8      45.9          -6086
## 4 Active trave~  2011        8938       13581      62.0      46.7          -4643
## 5 Active trave~  2012        9279       12373      62.6      47.7          -3094
## 6 Active trave~  2013        7914       14269      59.3      47.9          -6355
## # ... with 1 more variable: measure_diff <dbl>
```

```
group_by_all(select_dataset)
```

```
## # A tibble: 14 x 8
## # Groups:   indicator.x, year, numerator.x, numerator.y, measure.x, measure.y,
## #   numerator_diff, measure_diff [14]
##    indicator.x   year numerator.x numerator.y measure.x measure.y numerator_diff
##    <chr>        <dbl>       <dbl>       <dbl>     <dbl>     <dbl>          <dbl>
##  1 Active trav~  2008       10748       13357      65.5      46.3          -2609
##  2 Active trav~  2009       10515       12165      64.6      45.0          -1650
##  3 Active trav~  2010        7284       13370      61.8      45.9          -6086
##  4 Active trav~  2011        8938       13581      62.0      46.7          -4643
##  5 Active trav~  2012        9279       12373      62.6      47.7          -3094
##  6 Active trav~  2013        7914       14269      59.3      47.9          -6355
##  7 Active trav~  2014        9680       14601      59.2      46.6          -4921
##  8 Active trav~  2015        9840       13993      57.6      45.0          -4153
##  9 Active trav~  2016        9946       14244      59.6      47.3          -4298
## 10 Active trav~  2017       10448       13627      59.4      43.8          -3179
## 11 Active trav~  2018       10051       14172      58.5      45.8          -4121
## 12 Active trav~  2019       10046       14001      57.2      46.2          -3955
```

```
## 13 Active trav~  2020        9812       14682      61.2      49.7          -4870
## 14 Active trav~  2021       10239       13681      60.7      48.2          -3442
## # ... with 1 more variable: measure_diff <dbl>
```

```
# 'select_dataset' has multiple variables so to verify data more precisely
# we have to again select only three variables.
# select year, numerator_diff and measure_diff
travel_diff <- select_dataset %>%
select(year, numerator_diff, measure_diff) %>%
group_by(year)

group_by_all(travel_diff)
```

```
## # A tibble: 14 x 3
## # Groups:   year, numerator_diff, measure_diff [14]
##     year numerator_diff measure_diff
##    <dbl>          <dbl>        <dbl>
##  1  2008          -2609         19.3
##  2  2009          -1650         19.6
##  3  2010          -6086         15.9
##  4  2011          -4643         15.3
##  5  2012          -3094         14.9
##  6  2013          -6355         11.4
##  7  2014          -4921         12.6
##  8  2015          -4153         12.6
##  9  2016          -4298         12.3
## 10  2017          -3179         15.6
## 11  2018          -4121         12.6
## 12  2019          -3955         11.0
## 13  2020          -4870         11.5
## 14  2021          -3442         12.6
```

Build data visualisation

```
# Scatter plot - The difference between active travel to school between
# Aberdeen city and Aberdeenshire
p<- travel_diff %>%
  ggplot(aes(x = year, y = measure_diff)) +
  geom_point()+
  geom_smooth(se = TRUE) +
  scale_x_continuous(breaks = seq(2008, 2021, by=2)) +
  ylab("Measure difference (%)") +    # label y-axis
  ggtitle("Diff. between Aberdeen city and Aberdeenshire") +
  theme_bw()+
  theme(plot.background = element_rect(fill = "green"))

ggplotly(p, tooltip = c("x","y"))
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

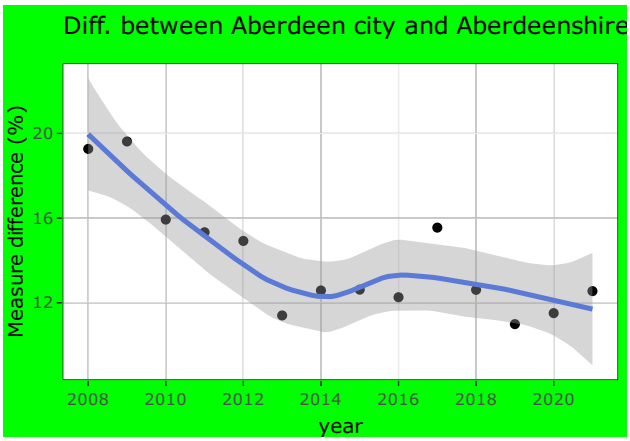

Fig. 1: The difference between active travel to school between Aberdeen City and Aberdeenshire

Explaination of the above graph

Process:

In this graph, scatter plot has selected to display the difference between active travel to school between Aberdeen City and Aberdeenshire. The duration of data presented in the above graph is from 2008 to 2021 (14 years). The first step was to filter out the 'Active travel to school' from 'Indicator' column and then 'Aberdeen city' and 'Aberdeenshire' from 'area_name' column. As a result, there were 28 rows returned for both 'Aberdeen city' and 'Aberdeenshire'.

Afterthat, two new datasets were created; One for Aberdeen City and other for Aberdeenshire to perform the full join and align the columns horizontally. Then the difference between Aberdeen City and Aberdeenshire was checked. To know the difference, there were two columns namely; numerator and measure were selected. Numerator is the total number of people actively traveled to Aberdeen City and Aberdeenshire and measure is the calculation given in the percentage form. Then, there were two new columns created; numerator_diff (numerator.AberdeenCity - numerator.Aberdeenshire) and measure_diff (measure.AberdeenCity - measure.Aberdeenshire). It was found the 'numerator_diff' has negative values (like: -2609, -1650, -6086, -4643, -3094, -6355, -4921, -4153, -4298, -3179, -4121, -3955, -4870 and -3442). The numerators of Aberdeenshire were higher than Aberdeen City but 'measure_diff' (%) was calculated well like: (19.27, 19.62, 15.93, 15.33, 14.92, 11.4, 12.58, 12.62, 12.26, 15.55, 12.61, 10.99, 11.51 and 12.55). Due to the negative values in 'numerator_diff' column, only 'measure_diff' column has been taken to visualize the difference between active travel to school between Aberdeen City and Aberdeenshire.

Output:

In the above graph, x-axis has year column from 2008 to 2011 and y-axis have 'measure_diff' column in percentage. The difference in the 'measure' column shows the interesting pattern through out the year. Initially, the 'measure' difference was 19.3% aprox. in 2008 and since then the downfall pattern was noted until year 2013 (11.4% aprox). After that, it shows the stable pattern until 2021 (12.5% aprox.) with the fluctuation of 1.5% except in 2017 (15.5% aprox) high value. The graph shows there is still a difference of 12% aprox. between active travel to school between Aberdeen City and Aberdeenshire. In term of numerator, there is a difference of 3442 active travels to school between Aberdeen City and Aberdeenshire. There is a slight increase noted in year 2017 (15.5%) which needs to be further investigation whether it is due to the backlog entries, data error or due to any specific event.

What is the data source? What are the data limitations?

The given dataset (scotpho_active_travel.csv), with data from the Scottish Public Health Observatory on active travel to school and work. The source of data is 'Hands Up Scotland Survey (HUSS), Sustrans (Official statistic)'.

There are some limitations in this dataset. It has limited variables to go through in more details. The variable 'measure' given in the percentage and the denominator for this calculation is not mentioned. Numerator is given but the definition for the same is missing. Confidence intervals (lower/upper) are missing.

What are the strengths and limitations of the approach you took for your visualisation?

Strengths -

1. The data is precise, consistent and easy to analyse. It has few variables and most of them are self explanatory. The column 'measure' is given in the percentage.
2. The finding of this study is generalized and a representative of the sample population. This data has 14 years of measures of active travel to school between Aberdeen city and Aberdeenshire. So, it is easier to find the difference between Aberdeen city and Aberdeenshire[i].
3. I tried a best approach to visualize the data, put numerator difference in bar chart and measure difference in line graph to display on secondary axis but the numerators of Aberdeen City are greater than Aberdeenshire and difference calculated (numerator_diff - new variable) in negative values. Whereas, the difference of measure variable (measure_diff - new variable) is calculated in positive. So, It doesn't look good to display both new variables (numerator_diff (-Ve values) and measure_diff(+ve values) in one graph. One indicator will show above the zero line and other below zero. Therefore, I decided to keep only 'measure_diff' variable in a scatter chart.

Limitation-

1.Although, the column (measure) is given in the percentage but the method of the calculation is not known. It could have produced more useful information if denominator was given. 2. The data has limited number of variables so it is difficult to produce complex analysis. 3. The difference between Aberdeen city and Aberdeenshire is calculated in percentage (column name: measure). The other related columns such as: denominator or gender are not available which limits the detailed analysis.

What would you do to ensure your analysis is reproducible?

I would like to mention some important points to ensure the analysis is reproducible. 1. In this analysis, the entire process has documented including data sources, steps taken and choices made. 2. All the codes and narration are given in a R Markdown file so that it can be easily accessible and available for next level of analysis. 3. The date and version control are in place to track the changes and ensure consistency over time. 4. Provided clear instruction on the data science steps such as: import the data, inspect the data, data preparation and data visualization. 5. Followed best practices and standards to maintain readability and consistency in the database[ii]. 6. The R Markdown file has been shared on Github to safe storage and accessible for others to reproduce the analysis. All the codes are well labeled. 7. I have been using R 3.6.3 version for this data analysis and R Markdown file to store the data analysis steps and necessary documentation.

In summary, there are three steps taken to ensure the data analysis is reproducible[iii]:

#1. Before data analysis- Data safely stored in multiple locations and can be taken in portable format, data formatted appropriately for analysis.

#2. During data analysis- The code is clean and thoroughly commented. Software version and computing environments been documented.

#3. After data analysis - The instruction will be given to locate the data file, meta data and codes on Github. Here is Github link to access the data and R Markdown file - https://github.com/awan90/Intro_HDS_project

References: [i]https://betterthesis.dk/research-methods/lesson-1different-approaches-to-research/strengths-and-limitations

[ii]https://www.linkedin.com/advice/3/how-do-you-make-your-data-analysis-transparent-reproducible

[iii] https://esajournals.onlinelibrary.wiley.com/doi/full/10.1002/bes2.1801