SC1 - Assessed Coursework 1 - UK National Census 2021

Emma Tarmey

Semester 1, 2022/23

Contents

1 Outline of Purpose

For this coursework piece, the author has chosen to take on a case study - analyzing the results of the 2021 United Kingdom Census, detailed here. We will do so using publicly available census summary data obtained from the Office for National Statistics (ONS) available here. We do so towards the goal of performing our own explorations, and attempting to verify summaries given by news outlets about the UK population. As a result, this coursework broadly covers the topic of "Reproducibility" from the module, but also seeks to function as a standalone exploration of the census results. We will then take this approach to confirming two ideas generally taken as facts of life in the UK - the idea that the UK population is ageing with time, and the idea that the length of commutes in the UK are increasing with time.

For the purposes of our exploration, in particular comparing demographics over time, we consider three central data sources:

- Data for the 2021 UK Census is available directly from the ONS here
- Data for the 2011 UK Census is available from nomis here
- Data for the 2001 UK Census is available from nomis here

2 The Data

Broadly, the UK 2021 Census is a nationwide survey that took place in March 2022, conducted by the Office for National Statistics. The purpose of the census is to track various UK population demographics, such as employment status, disability status, housing status, education status and so on. This then enables UK population parameters to be calculated directly without needing to circumvent the uncertainty which would necessarily follow a sample-based estimate.

The census was conducted as a series of questions, some mandatory and some voluntary, to be conducted by every member of every household in the UK. The responses are recorded individually and then be submitted jointly by household. Questions can support both multiple choice and write-in answers. Additionally, though the census was primarily conduced online, respondents could also respond via the post if needed. Approximately 16,000,000 of the 59,597,300 respondants for England and Wales responded to the Census using filled-in paper copies of the questionnaire.

The ONS provides summary sheets detailing aggregate totals of various demographic factors, as well as break-down summary tables of these totals. At time of writing, the "roll-out" of the census results is ongoing, and as a result not all summaries are currently available, with the "Phase one topic summaries" set to be fully released by January 19th 2023. Despite this inconvenience, there is enough presently available for a satisfactory analysis of various UK demographic questions.

Towards the goal of keeping the results reproducible, we have left as much of the data preparation and manipulation as possible explicit within this document. The only data pre-processing tasks which took place outside of this document are:

- Providing more meaningful file names. In particular 2011 census results always download as "bulk.csv", and as such were re-named.
- Converting the 2021 excel sheets into csv format for ease of file manipulation
- Isolating square data-tables in the csv files, making a point of excluding editorial comments and guides

3 Data Exploration

3.1 Exploration 1 - Does the UK have an ageing population?

It is considered common knowledge that the UK has an ageing population. Additionally, the Centre for Ageing Better claims directly in their 2022 Annual State of Ageing Report that the population is "ageing rapidly", and more specifically that ...

We can examine this claim for ourselves by comparing the UK Census data for 2021, 2011 and 2001 directly. In particular, with the topic summaries for "Demography and Migration" (referred to also as "Age Structure") we may test this claim ourselves. For data compatability reasons, we may two key observations:

- The different census' summary tables do not present results over the same age ranges. As a result, we must manually aggregate the groups in order to compare the distributions
- For similar data access reasons, we restrict our search from the whole UK to just England and Wales.

We begin by preparing our data as follows:

```
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 4.1.3
```

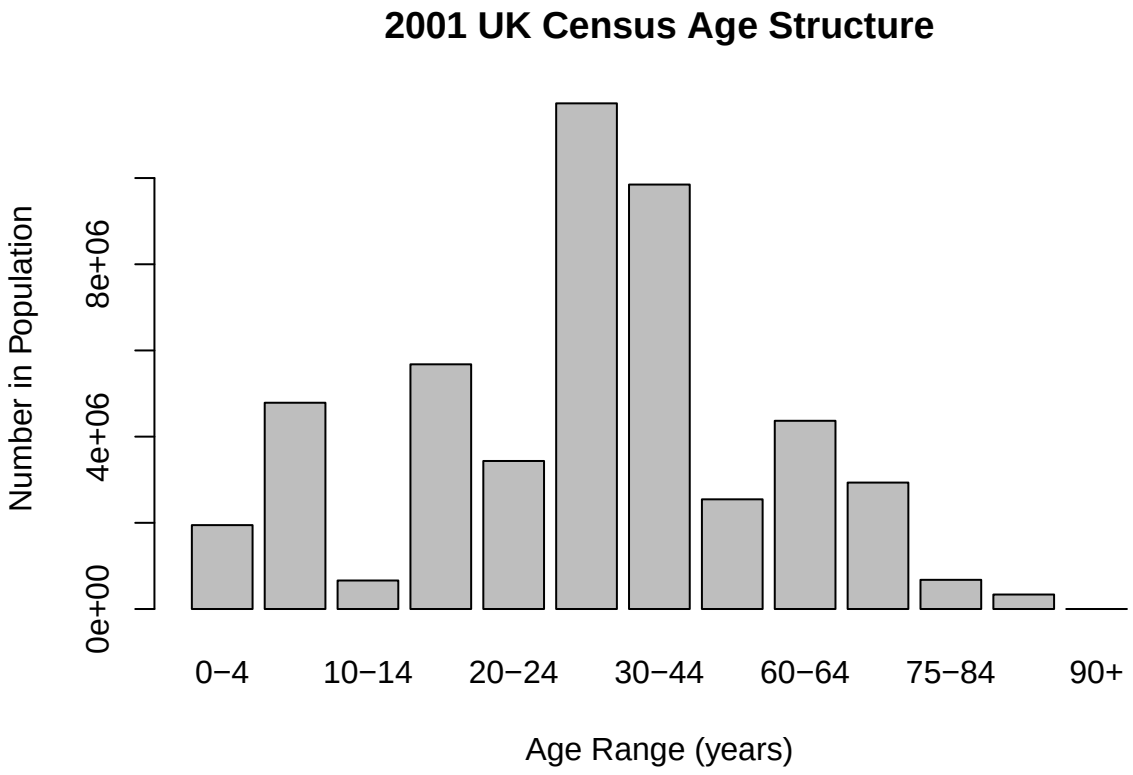
```
# empty R memory
rm(list=ls())

# census results from different years are in different formats
# as a result, data preparation is required

census.2021.data <- transpose( read.csv(
                    "uk_census_data/2021/2021_population_and_household_estimates.csv",
                    header = FALSE) )[-c(1:3), c(1:2)]
census.2011.data <- transpose( read.csv(
                    "uk_census_data/2011/2011_age_structure.csv",
                    header = FALSE) )[-c(1:5, 22, 23), ]
census.2001.data <- transpose( read.csv(
                    "uk_census_data/2001/2001_age_structure.csv",
                    header = FALSE) )[-c(1:5, 22, 23), ]

age.ranges.2001         <- c("0-4", "5-7", "8-9", "10-14", "15", "16-17", "18-19",
                             "20-24", "25-29", "30-44", "45-59", "60-64", "65-74",
                             "75-84", "85-89", "90+")
age.ranges.2011         <- c("0-4", "5-7", "8-9", "10-14", "15", "16-17", "18-19",
                             "20-24", "25-29", "30-44", "45-59", "60-64", "65-74",
                             "75-84", "85-89", "90+")
age.ranges.2021         <- c("0-4", "5-9", "10-14", "15-19", "20-24", "25-29",
                             "30-34", "35-39", "40-44", "45-49", "50-54", "55-59",
                             "60-64", "65-69", "70-74", "75-79", "80-84", "85-89",
                             "90+")
common.aggregate.ranges <- c("0-4", "5-9", "10-14", "15-19", "20-24", "25-29",
                             "30-44", "45-59", "60-64", "65-74", "75-84", "85-89",
                             "90+")

census.2021.data[, 2] <- as.numeric(census.2021.data[, 2])
census.2011.data[, 2] <- as.numeric(census.2011.data[, 2])
census.2001.data[, 2] <- as.numeric(census.2001.data[, 2])

census.2021.data[, 1] <- age.ranges.2021
```

```
census.2011.data[, 1] <- age.ranges.2011
census.2001.data[, 1] <- age.ranges.2011

aggregated.2021 <- c(census.2021.data[1, 2],
                     census.2021.data[2, 2],
                     census.2021.data[3, 2],
                     census.2021.data[4, 2],
                     census.2021.data[5, 2],
                     census.2021.data[6, 2],
                     census.2021.data[7, 2] + census.2021.data[8, 2] + census.2021.data[9, 2],
                     census.2021.data[10, 2] + census.2021.data[11, 2] + census.2021.data[12, 2],
                     census.2021.data[13, 2],
                     census.2021.data[14, 2] + census.2021.data[15, 2],
                     census.2021.data[16, 2] + census.2021.data[17, 2],
                     census.2021.data[18, 2],
                     census.2021.data[19, 2])


aggregated.2011 <- c(census.2011.data[1, 2],
                     census.2011.data[2, 2] + census.2011.data[3, 2],
                     census.2011.data[4, 2],
                     census.2011.data[5, 2] + census.2011.data[6, 2] + census.2011.data[7, 2],
                     census.2011.data[8, 2],
                     census.2011.data[9, 2],
                     census.2011.data[10, 2],
                     census.2011.data[11, 2],
                     census.2011.data[12, 2],
                     census.2011.data[13, 2],
                     census.2011.data[14, 2],
                     census.2011.data[15, 2],
                     census.2011.data[16, 2])

aggregated.2001 <- c(census.2001.data[1, 2],
                     census.2001.data[2, 2] + census.2001.data[3, 2],
                     census.2001.data[4, 2],
                     census.2001.data[5, 2] + census.2001.data[6, 2] + census.2001.data[7, 2],
                     census.2001.data[8, 2],
                     census.2001.data[9, 2],
                     census.2001.data[10, 2],
                     census.2001.data[11, 2],
                     census.2001.data[12, 2],
                     census.2001.data[13, 2],
                     census.2001.data[14, 2],
                     census.2001.data[15, 2],
                     census.2001.data[16, 2])

census.2021.aggregated <- data.frame(common.aggregate.ranges, aggregated.2021)
census.2011.aggregated <- data.frame(common.aggregate.ranges, aggregated.2011)
census.2001.aggregated <- data.frame(common.aggregate.ranges, aggregated.2001)
```

With the above data preparations completed, we can begin to plot out our age range distributions and look for our pattern.

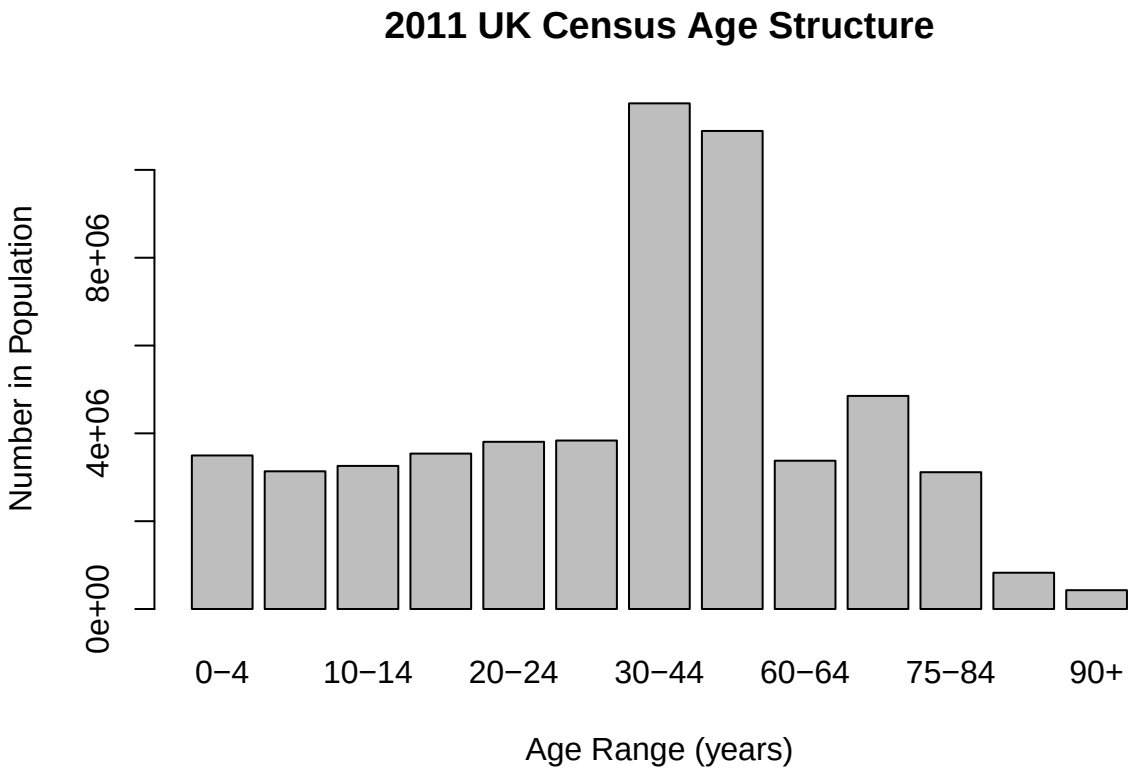
Our first distribution, of ages across the population of England and Wales in 2001, looks as follows:

```
barplot(as.numeric(census.2001.aggregated[, 2]),
        names.arg = common.aggregate.ranges,
        xlab = "Age Range (years)",
        ylab = "Number in Population",
        main = "2001 UK Census Age Structure")
```



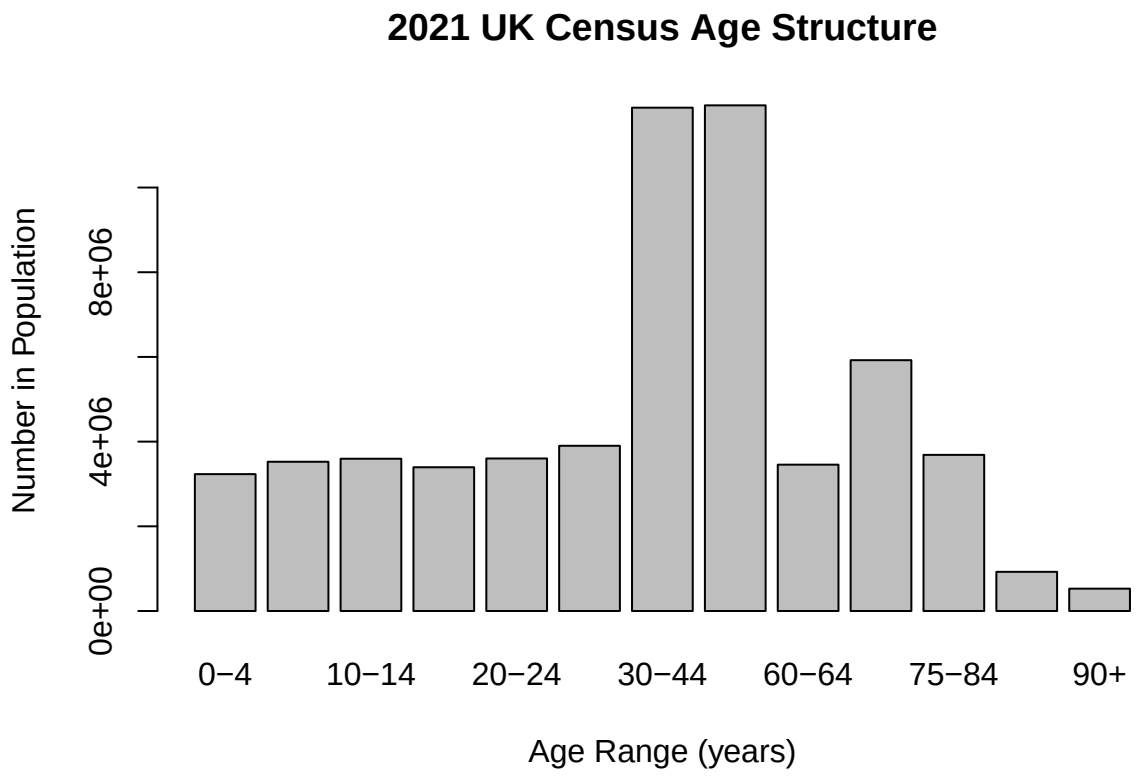
Our second distribution, of ages across the population of England and Wales in 2011, looks as follows:

```
barplot(as.numeric(census.2011.aggregated[, 2]),
        names.arg = common.aggregate.ranges,
        xlab = "Age Range (years)",
        ylab = "Number in Population",
        main = "2011 UK Census Age Structure")
```



Our third distribution, of ages across the population of England and Wales in 2021, looks as follows:

```
barplot(as.numeric(census.2021.aggregated[, 2]),
        names.arg = common.aggregate.ranges,
        xlab = "Age Range (years)",
        ylab = "Number in Population",
        main = "2021 UK Census Age Structure")
```



As we can see from the above, the most populous bulge in the data is shifting slowly to the right over the 20 year interval. Moreover, looking at the median values of each year we see:

- 2001 Median Age: 37
- 2011 Median Age: 39
- 2021 Median Age: 40

Without access to the original data, we cannot accurately hypothesis-test whether the population shows a statistically significant increase in mean age over time. Despite this, the pattern seems clear. As a result, graphically and numerically, we can indeed confirm that England and Wales have an ageing population, and that the population has aged consistently over the last 20 years.

3.2 Exploration 2 - Travel to Work Distance in the UK is increasing

It is considered common knowledge that when cities are built around "cars as default" for travel, people will tend to travel further for work. Moreover, the Trades Union Congress claimed in a 2019 Analysis that commuting is increasing across the UK. But is the average length of commute actually increasing over time? We can examine this question by means of checking ONS data in a similar fashion to the above.

The data-sets, as provided by the ONS, are standardised across commute ranges. But, the table rows are in different orders, and in 2001 in particular the "Other" category is left as multiple seperate subgroups with specified reasons for their exclusion. As a result, for comparison with other censuses, we aggregate all non-commuter and non-home categories in the 2001 data as "Other."

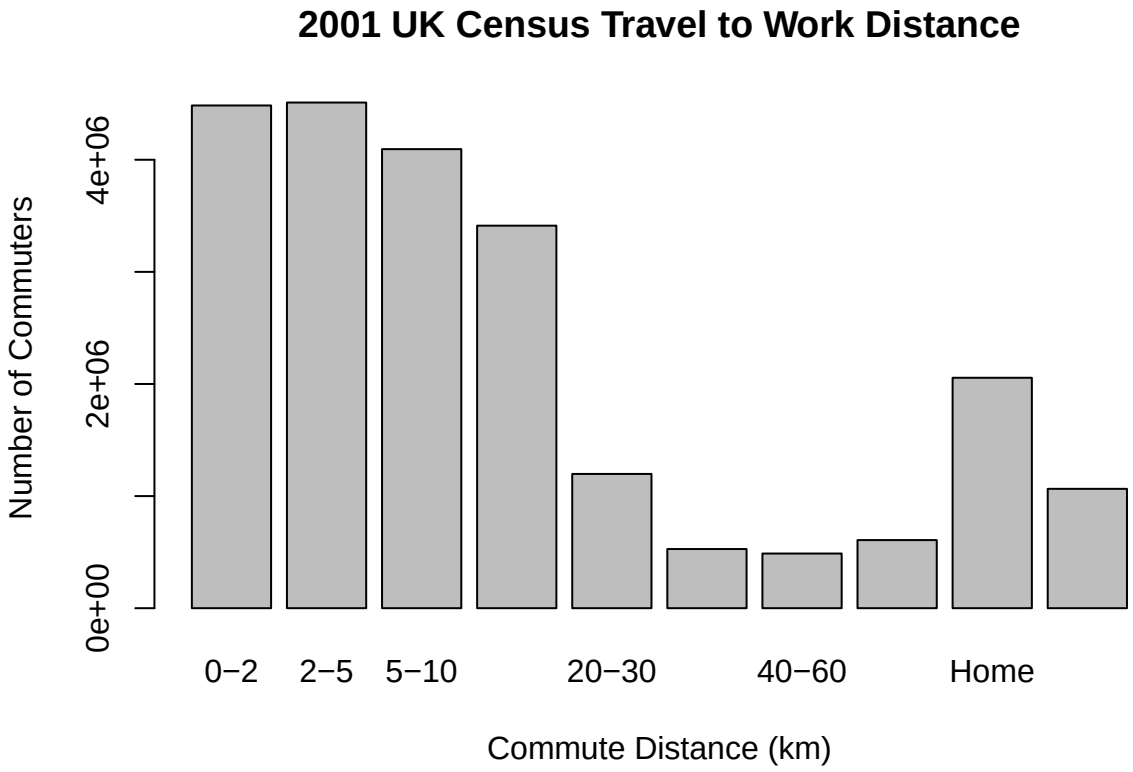
We begin by preparing our data as follows:

```
# empty R memory
rm(list=ls())

census.2021.data <- read.csv(
                    "uk_census_data/2021/2021_commute_distance.csv",
                    header = FALSE)[-c(1), -c(3)]
census.2011.data <- transpose( read.csv(
                    "uk_census_data/2011/2011_commute_distance.csv",
                    header = FALSE) )[-c(1:5, 16, 17), c(1:2)]
census.2001.data <- transpose( read.csv(
                    "uk_census_data/2001/2001_commute_distance_trim.csv",
                    header = FALSE) )[-c(1:3, 16, 17), c(1:2)]

distance.ranges <- c("0-2", "2-5", "5-10", "10-20", "20-30", "30-40", "40-60",
                     "60+", "Home", "Other")

census.2021.data[, 2] <- trimws(census.2021.data[, 2])
census.2021.data[, 2] <- as.numeric( gsub(",","", census.2021.data[, 2]) )
census.2011.data[, 2] <- as.numeric(census.2011.data[, 2])
census.2001.data[, 2] <- as.numeric(census.2001.data[, 2])

census.2001.travels <- c(census.2001.data[2, 2],
                         census.2001.data[3, 2],
                         census.2001.data[4, 2],
                         census.2001.data[5, 2],
                         census.2001.data[6, 2],
                         census.2001.data[7, 2],
                         census.2001.data[8, 2],
                         census.2001.data[9, 2],
                         census.2001.data[1, 2],
                         census.2001.data[10, 2] + census.2001.data[11, 2] + census.2001.data[12, 2]
                         )

census.2021.data[, 1] <- distance.ranges
census.2011.data[, 1] <- distance.ranges
census.2001.std       <- data.frame(distance.ranges, census.2001.travels)

#View(census.2021.data)
#View(census.2011.data)
#View(census.2001.std)
```

With the above data preparations completed, we may now plot our distributions as previously.

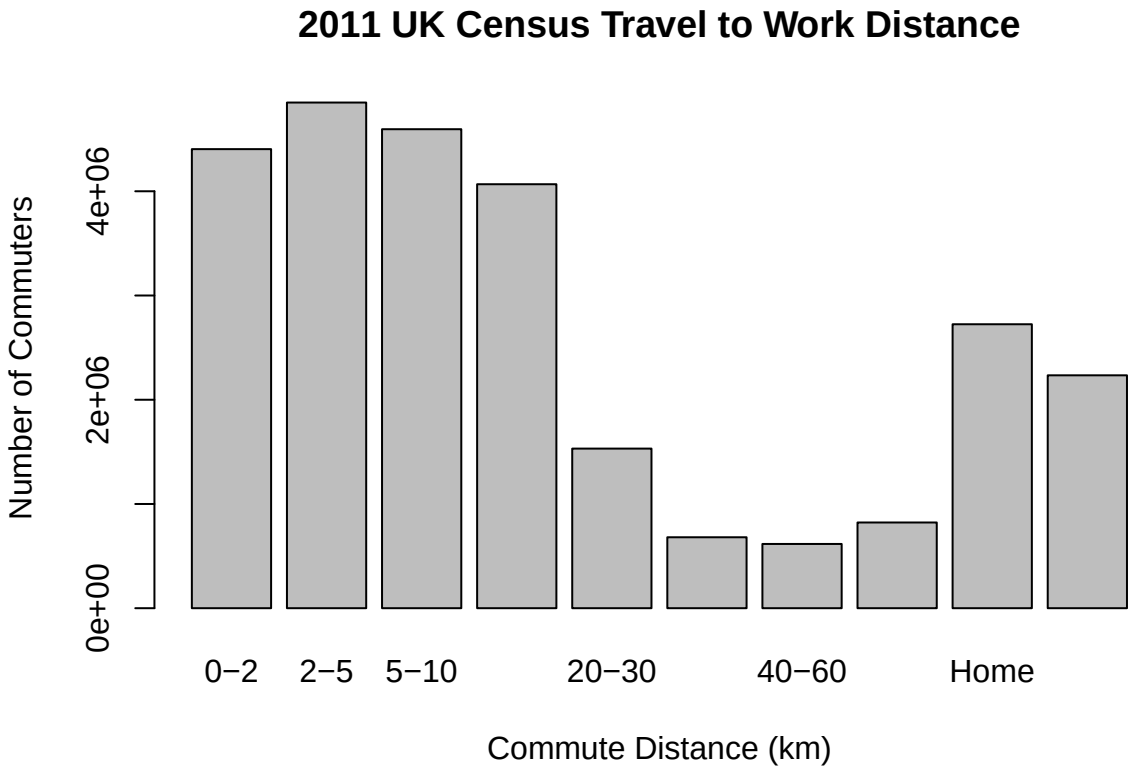
Our first distribution, of travel to work distances across the population of England and Wales in 2001, looks as follows:

```
barplot(as.numeric(census.2001.std[, 2]),
        names.arg = distance.ranges,
        xlab = "Commute Distance (km)",
        ylab = "Number of Commuters",
        main = "2001 UK Census Travel to Work Distance")
```



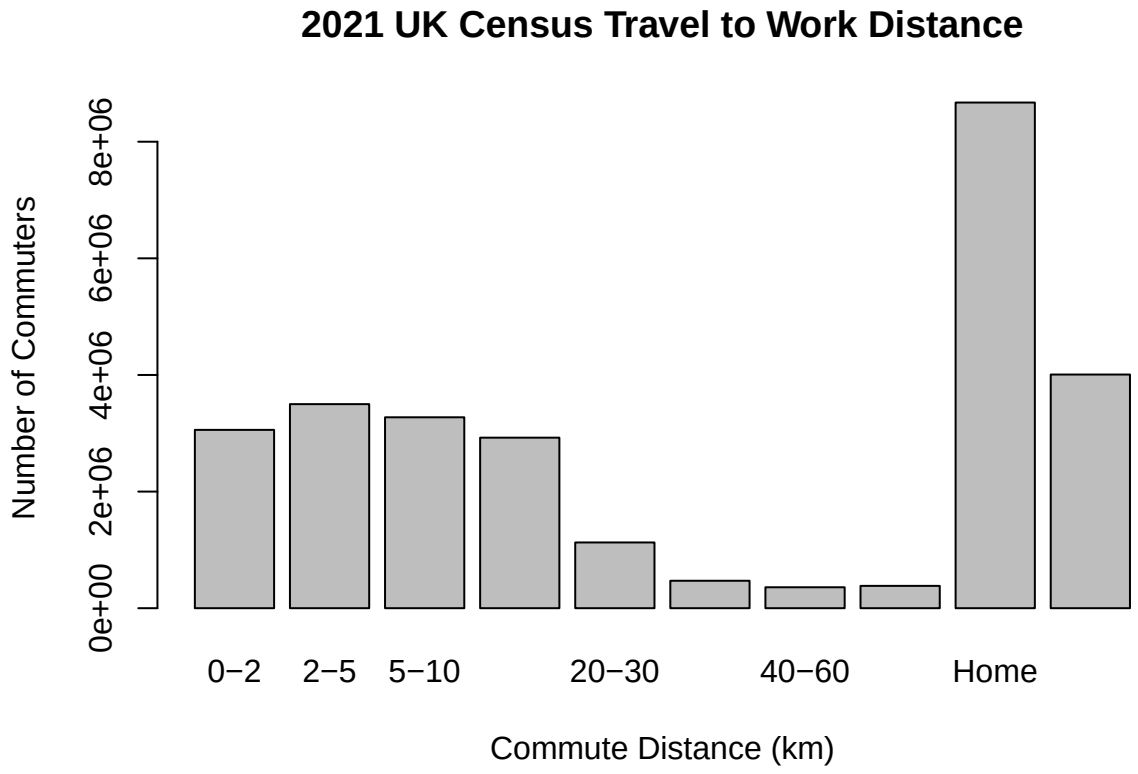
Our second distribution, of travel to work distances across the population of England and Wales in 2011, looks as follows:

```
barplot(as.numeric(census.2011.data[, 2]),
        names.arg = distance.ranges,
        xlab = "Commute Distance (km)",
        ylab = "Number of Commuters",
        main = "2011 UK Census Travel to Work Distance")
```



Our third distribution, of travel to work distances across the population of England and Wales in 2021, looks as follows:

```
barplot(as.numeric(census.2021.data[, 2]),
        names.arg = distance.ranges,
        xlab = "Commute Distance (km)",
        ylab = "Number of Commuters",
        main = "2021 UK Census Travel to Work Distance")
```



As we can see from the above, the most populous bulge in the data does shift to the right as we move from 2001 to 2011. However, in 2021, working from home is massively more popular than in previous years. This in turn throws off our pattern a little, but the broad increase in popularity of longer commutes still marked by the flattening of the left hand side of the graph, as commute distances in the ranges of 5-10 kilometers and 10-20 kilometers become more popular relatively to other non-home ranges.

4 Conclusions

In the above, we have used publicly available data-sets to re-examine and confirm our existing beliefs about life in the UK. In particular, we have confirmed that over the last 20 years, the UK has an ageing population, the average distance travelled to work is increasing, and working from home has become much more popular. Our hope is that by keeping all data preparations explicitly local to this document, we both maintain the reproducibility of our results and demonstrate some difficulties that come with working with real-world data-sets.

The full code used to produce this report, as well as copies of all data-sheets used, are available on my GitHub page here.