Business Statistics Mid-Term Assessment IB94X0 2023-2024 #1

5582755

2023-10-21

This is to certify that the work I am submitting is my own. All external references and sources are clearly acknowledged and identified within the contents. I am aware of the University of Warwick regulation concerning plagiarism and collusion.

No substantial part(s) of the work submitted here has also been submitted by me in other assessments for accredited courses of study, and I acknowledge that if this has been done an appropriate reduction in the mark I might otherwise have received will be made.

Tasks

This report fulfills the request of the company, performing the specific analyses requested.

1. Children in low income homes by borough Provide a table that shows for each borough: the average number of children in low income homes per ward across all years in the dataset, the standard deviation of children in low income homes per ward across all years in the dataset, the lowest number of children in low income homes recorded for a ward in any year, and the highest number of children in low income homes recorded for a ward in any year.
2. Exclude unusual boroughs Excluding a small number of unusual boroughs. The boroughs to be excluded are: City of London; Kensington and Chelsea; Kingston upon Thames; Richmond upon Thames; Westminster.
3. Visualise data for different years Create a visualisation using either violin/Jitter plot that shows the distribution of children in low income households in wards for each year. The plot should also clearly show a visual representation of the mean value and standard deviation for each year.
4. A t-test comparing earliest and latest year Perform a t-test to compare the children in low income households in wards in the earliest year in the dataset (2014) and the latest year in the dataset (2021) and report the results using Null Hypothesis Significance Testing, and the estimation approach.

Section 1: R code

```
knitr::opts_chunk$set(echo = TRUE)

library(tidyverse)
options(width=100)
library(lubridate)
```

Data Dictionary

We have been given the data on children living in low income households in London. London is split into boroughs (large areas of London which usually have their own local council or local government) and then further sub-divided into wards. This data is provided across 8 years. It shows the counts for the number of children living in low income housing within each ward in London.

Variable	Description
Ward code	Unique code for each Ward
Wards (2018)	Name of the Ward as per year 2018 within each Borough
Borough	Name of the Boroughs within London
year	Year from which data for each ward is given (2014-2021)
children	children living in low income households

Read Data

```
# First read in the data and check if the format for any variable needs to be conv
erted.

children_data <- read_csv("children_low_income_data.csv")
```

```
## Rows: 5120 Columns: 5
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (3): Ward code, Wards (2018), Borough
## dbl (2): year, children
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this messag
e.
```

```
str(children_data)
```

```
## spc_tbl_ [5,120 × 5] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Ward code   : chr [1:5120] "E05000026" "E05000026" "E05000026" "E05000026"
...
##  $ Wards (2018): chr [1:5120] "Abbey" "Abbey" "Abbey" "Abbey" ...
##  $ Borough     : chr [1:5120] "Barking and Dagenham" "Barking and Dagenham" "Ba
rking and Dagenham" "Barking and Dagenham" ...
##  $ year        : num [1:5120] 2014 2015 2016 2017 2018 ...
##  $ children    : num [1:5120] 1090 1119 1306 1417 1466 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   `Ward code` = col_character(),
##   ..   `Wards (2018)` = col_character(),
##   ..   Borough = col_character(),
##   ..   year = col_double(),
##   ..   children = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
# We convert variable year from from a numeric to a factor variable.
children_data$year <- as.factor(children_data$year)
```

Checking the Integrity of Data

Cleaning data by removing for NA values.

```
summary(children_data)
```

```
##   Ward code         Wards (2018)         Borough               year         chi
ldren
##  Length:5120        Length:5120        Length:5120        2014   : 640   Min.
:   5.0
##  Class :character   Class :character   Class :character   2015   : 640   1st Q
u.: 270.0
##  Mode  :character   Mode  :character   Mode  :character   2016   : 640   Median
: 511.0
##                                                          2017   : 640   Mean
: 570.1
##                                                          2018   : 640   3rd Q
u.: 794.0
##                                                          2019   : 640   Max.
:2094.0
##                                                          (Other):1280   NA's
:33
```

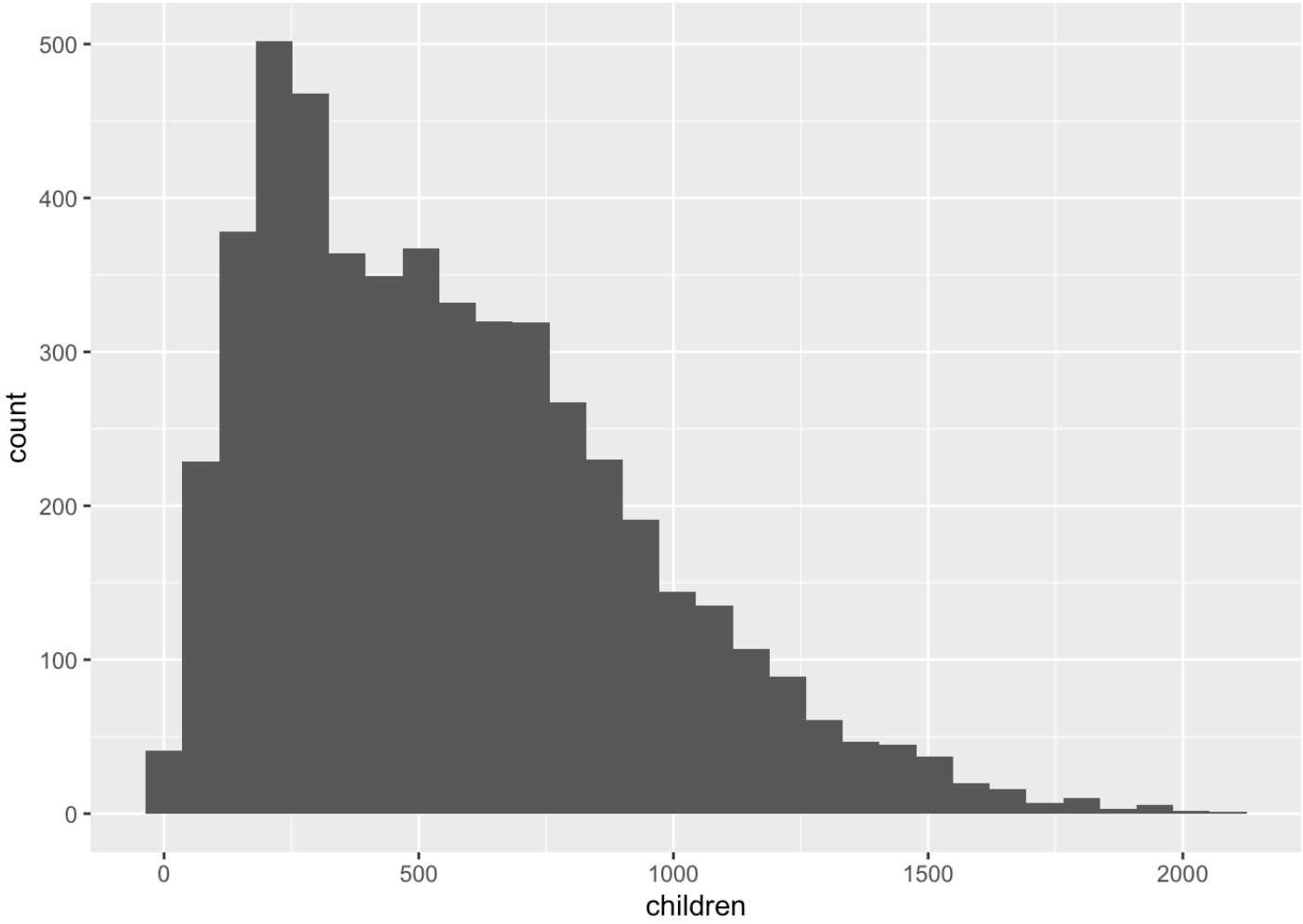
```
# We remove the row for which the data for children is missing.
children_data_clean <- na.omit(children_data)
```

Visualisation of Number of Children in Low Income Household

To check for any Outliers in data.

```
ggplot(children_data_clean)+
  geom_histogram(aes(children))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# There is a strong positive skew relation in the number of children in low income households. We will continue to apply the two sample t-tests as they are specifica lly requested. but t test is not appropriate for the data that is strongly non-nor mally distributed.
```

We do not find any extreme outliers. But we did find that the data is positively skewed.

1) Children in low income homes by borough

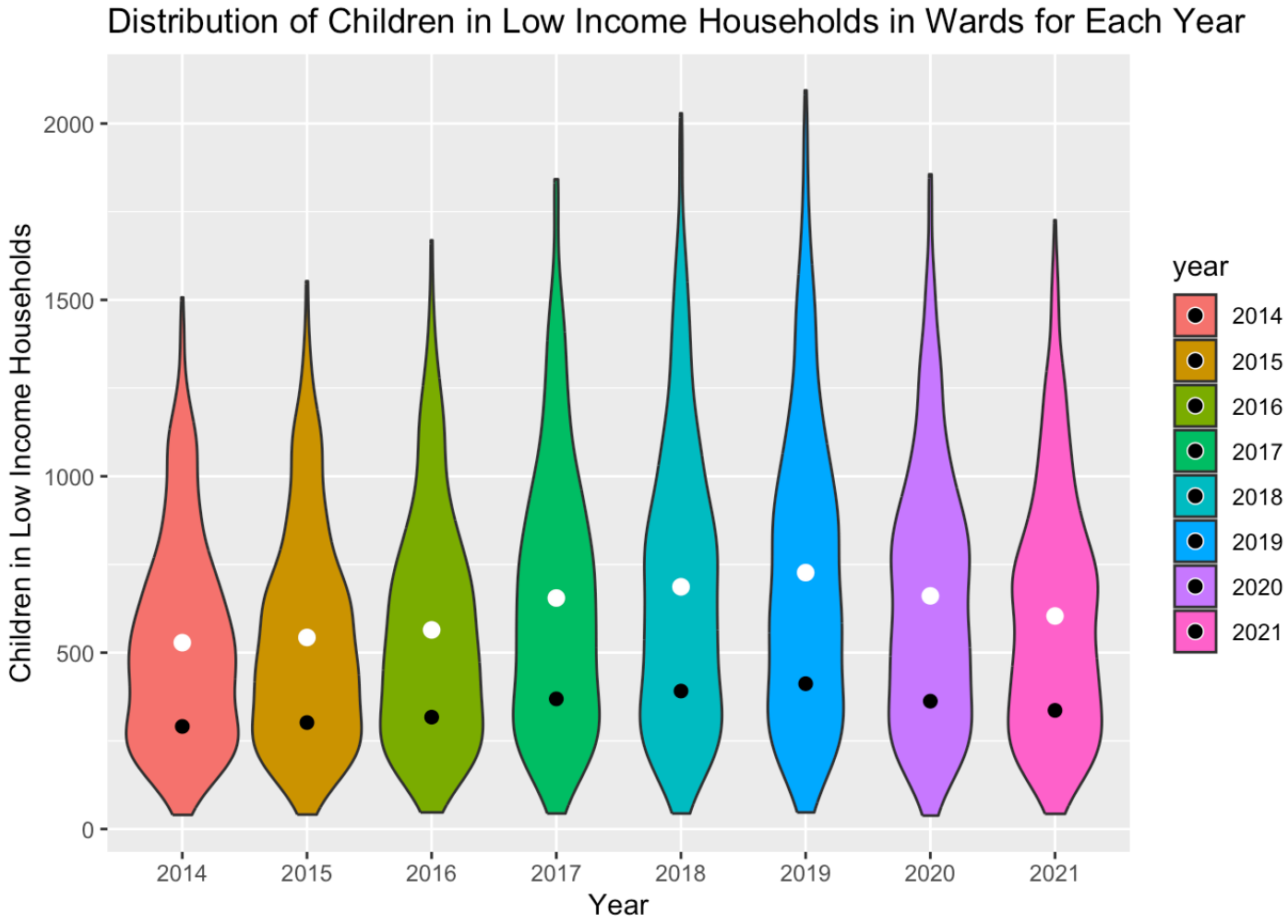
```
boroughs_data <- children_data_clean %>%
  group_by (Borough) %>%
  summarise(frequency = n() ,mean = mean(children, na.rm = TRUE),
            standard_Deviation = sd(children, na.rm=TRUE),
            minimum_children = min(children, `Wards (2018)`,na.rm = TRUE),
            maximum_children = max(children,na.rm = TRUE),
            ward_with_min_children = `Wards (2018)`[which.min(children)],
            ward_with_max_children = `Wards (2018)`[which.max(children)])
```

2) Exclude unusual boroughs

```
# We exclude the unusual boroughs from data as requested and make a new data set n
aming children_excl_borough.

unusual_borough <- c("City of London","Kensington and Chelsea","Kingston upon Tham
es","Richmond upon Thames","Westminster")
children_excl_borough <- children_data_clean %>%
  filter(!(Borough %in% unusual_borough))
```

3) Visualise data for different years

```
ggplot(children_excl_borough, aes(x = year, y = children, fill = year)) +
  geom_violin()+
  stat_summary(fun = mean, geom ="point", color = "white", size=2.5 ) +
  stat_summary(fun = sd, geom ="point", size=2, color="black")+
  labs(title = "Distribution of Children in Low Income Households in Wards for Eac
h Year", x="Year", y="Children in Low Income Households")
```




```
# Black point in the plot shows the standard deviation for each year and the white points shows the mean for each year.
```

The plot shows variation in number of children in low income households over the year 2014 to 2021 in different Wards accros London. White points show the mean number of children in low income households. We can see a slight upward movement in the mean value till year 2019 and then downward movement from 2020. Black point show the standard deviation for each year and we can observe a similar pattern in it as in the mean.

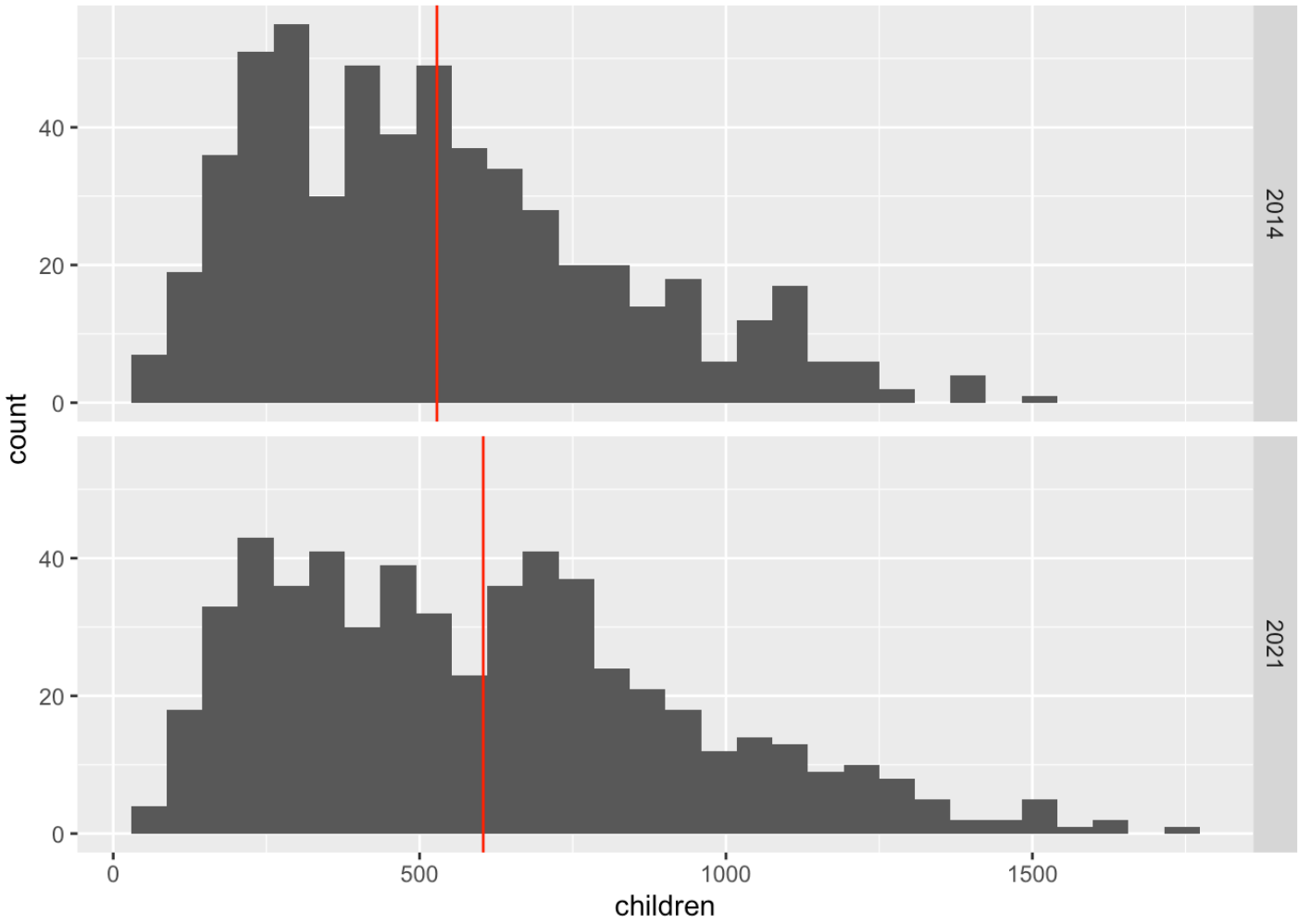
```
# This table show the numeric value of the mean and standard deviation of number c
hildren living in low income households for every year.

yearly_data <- children_excl_borough %>%
  group_by(year) %>%
  summarise( mean = mean(children),
             standard_deviation = sd(children))
```

4) A t-test comparing earliest and latest year

Data quality checks and visualisation of number of children in low income households in every year before performing the t-test.

```
children_2014_2021 <- filter(children_excl_borough,
                             year %in% c("2014", "2021"))
year_14_21 <- filter(yearly_data, year %in% c("2014","2021"))
ggplot(children_2014_2021)+
  geom_histogram(aes(children))+
  facet_grid(year~.)+
  geom_vline(data = year_14_21, mapping = aes(xintercept = mean), col = "red")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



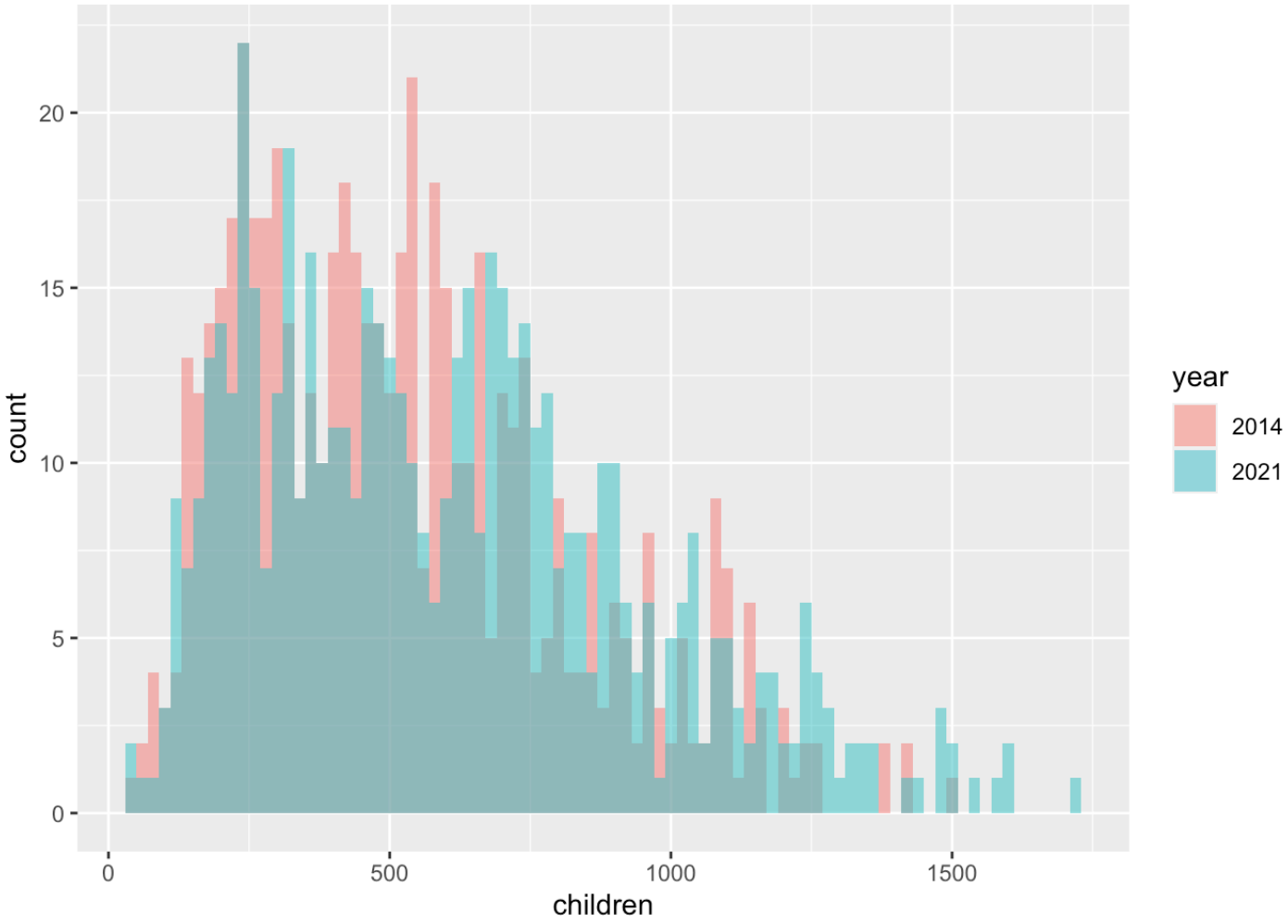

```
# There is a positive skewed relation in the number of children for both the year
s. We will continue to apply the two sample t-tests as they are specifically reque
sted. but t test is not appropriate for the data that is non-normally distributed.
```

Performing the two sample t-test for year 2014 and 2021

```
# p value is < 0.05, ie it is statistically significant and we accept the null hyp
othesis that there is no significant difference in means between year 2014 and 202
1. However, we note that this is a non-normal distribution.

t.test(children ~ year, data = filter(children_2014_2021))
```

```
##
##  Welch Two Sample t-test
##
## data:  children by year
## t = -4.0215, df = 1095.2, p-value = 6.18e-05
## alternative hypothesis: true difference in means between group 2014 and group 2
021 is not equal to 0
## 95 percent confidence interval:
##  -112.4091  -38.6873
## sample estimates:
## mean in group 2014 mean in group 2021
##           528.3214           603.8696
```

```
ggplot(children_2014_2021, aes(x=children, fill=year)) +
  geom_histogram(binwidth=20, position="identity", alpha=0.5)
```




```
# test results are: t = -4.0215, df = 1095.2, p-value = 6.18e-05
```

Estimation Approach

```
library(emmeans) # for emmeans() and pairs()
```

```
# Estimation Question: How much is the difference from year 2014 to 2021?
# effect size between wards over the years can be seen using mean of children in t
he wards in the year.

children <- lm(children~year, data=children_2014_2021)
(  children.emm <- emmeans(children, ~year)  )
```

```
##  year emmean   SE   df lower.CL upper.CL
##  2014    528 13.3 1118      502      554
##  2021    604 13.3 1118      578      630
##
## Confidence level used: 0.95
```

```
(  children.contrast <- confint(pairs(children.emm)))
```

```
##  contrast            estimate   SE   df lower.CL upper.CL
##  year2014 - year2021    -75.5 18.8 1118     -112    -38.7
##
## Confidence level used: 0.95
```

Section 2: Report

Analysing Data on Children Living in Low Income Households in London

This report presents the results of the analyses requested by the company using data provided for the number of children living in the low income households in London. London is divided in Boroughs and then further each Borough is divided into Wards. We were given the number of children living in low income Households in these Wards from 2014 to 2021 with a total of 5120 observations.There was a small amount of missing data and has been removed for those wards prior to the analyses reported below, leaving 5087 observations.

I begin by grouping the boroughs and finding the mean, standard deviation, maximum and minimum number of children in low income households and making a new table (boroughs_data) for this data. I observed that City of London has minimum number of children among all wards which could be due to missing data for wards in that Borough.

Then the company asked to exclude a few Boroughs which were - City of London; Kensington and Chelsea; Kingston upon Thames; Richmond upon Thames; Westminster for which I made a new data table (children_excl_borough).

Then I used a Violin plot (useful for visualizing the distribution and summary statistics of a data set, especially when you want to compare multiple categories or groups) to show the the distribution of children in low income households in wards for each year along with the mean and standard deviation for each year. We saw an slight inverted U- shaped curve for both mean and standard deviation which showed that the number of children have been increasing from year 2014 to 2019 and then they started to decrease.

T-test Results showed that there is no significant difference in mean of number of children in low income households between year 2014 and 2021, but the the results of t-test are not appropriate for this data as it is non-normal distribution.Using Null Hypothesis Significant Testing approach we infered that the mean number of children in the low income households for year 2014 is 528 children. The mean number of children in the low income households for year 2021 is 604 children.The mean number of children in the low income households for year 2021 are 76 more than 2014 and similar results were found when we performed Estimation Approach to analyse the number of children in low income households in year 2014 and 2021.